Statistics Applied to Economics

Degree in Economics

F.Tusell

Dpto. Métodos Cuantitativos

Academic Year 2021–2022

eman ta zabal zazu

Universidad
del País Vasco

Euskal Herriko
Unibertsitatea

Content I

Poisson probability function

- Defined on non-negative integers, $x = 0, 1, 2, \ldots$ with $P_X(x)$:
$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Poisson probability function

- Defined on non-negative integers, $x = 0, 1, 2, \ldots$ with $P_X(x)$:

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

- Well defined; obviously non-negative, and:

Poisson probability function

- Defined on non-negative integers, $x = 0, 1, 2, \ldots$ with $P_X(x)$:

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

- Well defined; obviously non-negative, and:

$$\sum_{x=0}^{\infty} P_X(x) \quad = \quad \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!}$$

Poisson probability function

- Defined on non-negative integers, $x = 0, 1, 2, \ldots$ with $P_X(x)$:

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

- Well defined; obviously non-negative, and:

$$\begin{aligned}\sum_{x=0}^{\infty} P_X(x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \ldots\right)\end{aligned}$$

Poisson probability function

- Defined on non-negative integers, $x = 0, 1, 2, \dots$ with $P_X(x)$:
$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$
- Well defined; obviously non-negative, and:

$$\begin{aligned}\sum_{x=0}^{\infty} P_X(x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \dots\right) \\ &= e^{-\lambda}e^{\lambda} = 1\end{aligned}$$

Poisson probability function

- Defined on non-negative integers, $x = 0, 1, 2, \ldots$ with $P_X(x)$:
$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$
- Well defined; obviously non-negative, and:

$$\begin{aligned}\sum_{x=0}^{\infty} P_X(x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \ldots\right) \\ &= e^{-\lambda}e^{\lambda} = 1\end{aligned}$$

How do we get last expression from the previous one?

Poisson probability function

- Defined on non-negative integers, $x = 0, 1, 2, \ldots$ with $P_X(x)$:

$$P_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

- Well defined; obviously non-negative, and:

$$\begin{aligned} \sum_{x=0}^{\infty} P_X(x) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \ldots\right) \\ &= e^{-\lambda}e^{\lambda} = 1 \end{aligned}$$

Using a Taylor series expansion $e^t = 1 + t + t^2/2! + t^3/3! + \ldots$

Historical notes

- Named after Siméon Denis Poisson (1781-1840)

Historical notes

- Named after Siméon Denis Poisson (1781-1840)
- French mathematician, contemporaneous of Lagrange, Laplace and Fourier.

Historical notes

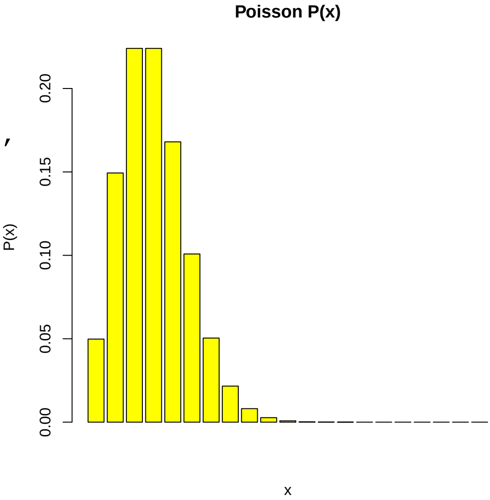
- Named after Siméon Denis Poisson (1781-1840)
- French mathematician, contemporaneous of Lagrange, Laplace and Fourier.
- Did important work in many areas of Mathematics.

Historical notes

- Named after Siméon Denis Poisson (1781-1840)
- French mathematician, contemporaneous of Lagrange, Laplace and Fourier.
- Did important work in many areas of Mathematics.
- See `http://en.wikipedia.org/wiki/Siméon_Denis_Poisson`.

What does it look like?

```
> x <- 0:1
> dpois(x,lambda=3)
[1] 0.04978707 0.14936121
> x <- 0:20
> barplot(dpois(x,lambda=3),
          col="yellow",
          xlab="x",
          ylab="P(x)",
          main="Poisson P(x)")
```



Moment generating function

$$\varphi_X(u) \overset{\text{def}}{=} E[e^{uX}] \quad = \quad \sum_{x=0}^{\infty} e^{ux} P_X(x) = \sum_{x=0}^{\infty} e^{ux} \frac{e^{-\lambda}\lambda^x}{x!} \tag{1}$$

Moment generating function

$$
\begin{aligned}
\varphi_X(u) \overset{\text{def}}{=} E[e^{uX}] &= \sum_{x=0}^{\infty} e^{ux} P_X(x) = \sum_{x=0}^{\infty} e^{ux} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^u)^x}{x!}
\end{aligned}
\tag{1}
$$

Moment generating function

$$
\begin{aligned}
\varphi_X(u) \overset{\text{def}}{=} E[e^{uX}] &= \sum_{x=0}^{\infty} e^{ux} P_X(x) = \sum_{x=0}^{\infty} e^{ux} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^u)^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^u)^x}{x!} \qquad (1)
\end{aligned}
$$

$$(2)$$

Moment generating function

$$
\begin{align}
\varphi_X(u) \overset{\text{def}}{=} E[e^{uX}] &= \sum_{x=0}^{\infty} e^{ux} P_X(x) = \sum_{x=0}^{\infty} e^{ux} \frac{e^{-\lambda}\lambda^{x}}{x!} \nonumber \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^{u})^{x}}{x!} \nonumber \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^{u})^{x}}{x!} \tag{1} \\
&= e^{-\lambda} e^{\lambda e^{u}} \tag{2} \\
& \tag{3}
\end{align}
$$

Moment generating function

$$
\begin{aligned}
\varphi_X(u) \overset{\text{def}}{=} E[e^{uX}] &= \sum_{x=0}^{\infty} e^{ux} P_X(x) = \sum_{x=0}^{\infty} e^{ux} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^u)^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^u)^x}{x!} && (1) \\
&= e^{-\lambda} e^{\lambda e^u} && (2) \\
&= e^{\lambda(e^u - 1)} && (3)
\end{aligned}
$$

Moment generating function

$$
\begin{aligned}
\varphi_X(u) \overset{\text{def}}{=} E[e^{uX}] &= \sum_{x=0}^{\infty} e^{ux} P_X(x) = \sum_{x=0}^{\infty} e^{ux} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^u)^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^u)^x}{x!} \qquad (1) \\
&= e^{-\lambda} e^{\lambda e^u} \qquad (2) \\
&= e^{\lambda(e^u - 1)} \qquad (3)
\end{aligned}
$$

How do we get (2) from (1) above?

Moment generating function

$$
\begin{aligned}
\varphi_X(u) \overset{\text{def}}{=} E[e^{uX}] &= \sum_{x=0}^{\infty} e^{ux} P_X(x) = \sum_{x=0}^{\infty} e^{ux} \frac{e^{-\lambda}\lambda^x}{x!} \\
&= \sum_{x=0}^{\infty} \frac{e^{-\lambda}(\lambda e^u)^x}{x!} \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^u)^x}{x!} && (1) \\
&= e^{-\lambda} e^{\lambda e^u} && (2) \\
&= e^{\lambda(e^u - 1)} && (3)
\end{aligned}
$$

Yet another use of $e^t = 1 + t + t^2/2! + t^3/3! + \ldots$.

Mean and variance

- Remember:

$$\alpha_1 = \left[\frac{\partial \varphi_X(u)}{\partial u} \right]_{u=0} \qquad \alpha_2 = \left[\frac{\partial^2 \varphi_X(u)}{\partial u^2} \right]_{u=0}$$

Mean and variance

- Remember:

$$\alpha_1 = \left[\frac{\partial \varphi_X(u)}{\partial u}\right]_{u=0} \qquad \alpha_2 = \left[\frac{\partial^2 \varphi_X(u)}{\partial u^2}\right]_{u=0}$$

- Hence,

$$\alpha_1 = \left[\frac{\partial}{\partial u} e^{\lambda(e^u - 1)}\right]_{u=0}$$

Mean and variance

- Remember:

$$\alpha_1 = \left[\frac{\partial \varphi_X(u)}{\partial u}\right]_{u=0} \qquad \alpha_2 = \left[\frac{\partial^2 \varphi_X(u)}{\partial u^2}\right]_{u=0}$$

- Hence,

$$\begin{aligned} \alpha_1 &= \left[\frac{\partial}{\partial u} e^{\lambda(e^u - 1)}\right]_{u=0} \\ &= \left[\frac{\partial\left(\lambda(e^u - 1)\right)}{\partial u} \times e^{\lambda(e^u - 1)}\right]_{u=0} \end{aligned}$$

Mean and variance

- Remember:

$$\alpha_1 = \left[\frac{\partial \varphi_X(u)}{\partial u}\right]_{u=0} \qquad \alpha_2 = \left[\frac{\partial^2 \varphi_X(u)}{\partial u^2}\right]_{u=0}$$

- Hence,

$$\begin{aligned}\alpha_1 &= \left[\frac{\partial}{\partial u} e^{\lambda(e^u-1)}\right]_{u=0} \\ &= \left[\frac{\partial\left(\lambda(e^u-1)\right)}{\partial u} \times e^{\lambda(e^u-1)}\right]_{u=0} \\ &= \left[\lambda e^u e^{\lambda(e^u-1)}\right]_{u=0} = \lambda\end{aligned}$$

Mean and variance

- Remember:

$$\alpha_1 = \left[\frac{\partial \varphi_X(u)}{\partial u}\right]_{u=0} \qquad \alpha_2 = \left[\frac{\partial^2 \varphi_X(u)}{\partial u^2}\right]_{u=0}$$

- Hence,

$$\begin{aligned}
\alpha_1 &= \left[\frac{\partial}{\partial u} e^{\lambda(e^u-1)}\right]_{u=0} \\
&= \left[\frac{\partial\left(\lambda(e^u-1)\right)}{\partial u} \times e^{\lambda(e^u-1)}\right]_{u=0} \\
&= \left[\lambda e^u e^{\lambda(e^u-1)}\right]_{u=0} = \lambda \\
\alpha_2 &= \left[\frac{\partial^2}{\partial u^2} e^{\lambda(e^u-1)}\right]_{u=0} = \lambda + \lambda^2
\end{aligned}$$

Mean and variance

- Remember:

$$\alpha_1 = \left[\frac{\partial \varphi_X(u)}{\partial u}\right]_{u=0} \qquad \alpha_2 = \left[\frac{\partial^2 \varphi_X(u)}{\partial u^2}\right]_{u=0}$$

- Hence,

$$\begin{aligned}
\alpha_1 &= \left[\frac{\partial}{\partial u} e^{\lambda(e^u-1)}\right]_{u=0} \\
&= \left[\frac{\partial\left(\lambda(e^u-1)\right)}{\partial u} \times e^{\lambda(e^u-1)}\right]_{u=0} \\
&= \left[\lambda e^u e^{\lambda(e^u-1)}\right]_{u=0} = \lambda \\
\alpha_2 &= \left[\frac{\partial^2}{\partial u^2} e^{\lambda(e^u-1)}\right]_{u=0} = \lambda + \lambda^2
\end{aligned}$$

How to obtain the variance from $\alpha_1 = E[X]$ and $\alpha_2 = E[X^2]$?

Mean and variance

- Remember:

$$\alpha_1 = \left[\frac{\partial \varphi_X(u)}{\partial u}\right]_{u=0} \qquad \alpha_2 = \left[\frac{\partial^2 \varphi_X(u)}{\partial u^2}\right]_{u=0}$$

- Hence,

$$\begin{aligned}
\alpha_1 &= \left[\frac{\partial}{\partial u} e^{\lambda(e^u-1)}\right]_{u=0} \\
&= \left[\frac{\partial\left(\lambda(e^u-1)\right)}{\partial u} \times e^{\lambda(e^u-1)}\right]_{u=0} \\
&= \left[\lambda e^u e^{\lambda(e^u-1)}\right]_{u=0} = \lambda \\
\alpha_2 &= \left[\frac{\partial^2}{\partial u^2} e^{\lambda(e^u-1)}\right]_{u=0} = \lambda + \lambda^2
\end{aligned}$$

$m = \alpha_1 = \lambda$ and $\sigma^2 = \alpha_2 - (\alpha_1)^2 = \lambda$.

Sum of independent Poisson variables

- Let $X_i \sim \mathcal{P}(\lambda_i)$ for $i = 1, \ldots, n$, independent of each other.

Sum of independent Poisson variables

- Let $X_i \sim \mathcal{P}(\lambda_i)$ for $i = 1, \ldots, n$, independent of each other.
- Let $X = X_1 + \ldots + X_n$. Then, $X \sim \mathcal{P}(\lambda_1 + \ldots + \lambda_n)$.

Sum of independent Poisson variables

- Let $X_i \sim \mathcal{P}(\lambda_i)$ for $i = 1, \ldots, n$, independent of each other.
- Let $X = X_1 + \ldots + X_n$. Then, $X \sim \mathcal{P}(\lambda_1 + \ldots + \lambda_n)$.
- Proof is easy:

$$\begin{aligned} \varphi_X(u) &= \varphi_{X_1}(u) \times \cdots \times \varphi_{X_n}(u) \\ &= e^{\lambda_1(e^u - 1)} \times \cdots \times e^{\lambda_n(e^u - 1)} \\ &= e^{(\lambda_1 + \ldots + \lambda_n)(e^u - 1)} \end{aligned}$$

and we recognize in the last expression the mgf of a Poisson random variable with $\lambda = \lambda_1 + \ldots + \lambda_n$.

Sum of independent Poisson variables

- Let $X_i \sim \mathcal{P}(\lambda_i)$ for $i = 1, \ldots, n$, independent of each other.
- Let $X = X_1 + \ldots + X_n$. Then, $X \sim \mathcal{P}(\lambda_1 + \ldots + \lambda_n)$.
- Proof is easy:

$$\begin{aligned} \varphi_X(u) &= \varphi_{X_1}(u) \times \cdots \times \varphi_{X_n}(u) \\ &= e^{\lambda_1(e^u - 1)} \times \cdots \times e^{\lambda_n(e^u - 1)} \\ &= e^{(\lambda_1 + \ldots + \lambda_n)(e^u - 1)} \end{aligned}$$

and we recognize in the last expression the mgf of a Poisson random variable with $\lambda = \lambda_1 + \ldots + \lambda_n$.

Would *the average* of $X_1, \ldots, X_n$ be Poisson-distributed?

Sum of independent Poisson variables

- Let $X_i \sim \mathcal{P}(\lambda_i)$ for $i = 1, \ldots, n$, independent of each other.
- Let $X = X_1 + \ldots + X_n$. Then, $X \sim \mathcal{P}(\lambda_1 + \ldots + \lambda_n)$.
- Proof is easy:

$$\begin{aligned} \varphi_X(u) &= \varphi_{X_1}(u) \times \cdots \times \varphi_{X_n}(u) \\ &= e^{\lambda_1(e^u-1)} \times \cdots \times e^{\lambda_n(e^u-1)} \\ &= e^{(\lambda_1+\ldots+\lambda_n)(e^u-1)} \end{aligned}$$

and we recognize in the last expression the mgf of a Poisson random variable with $\lambda = \lambda_1 + \ldots + \lambda_n$.

No, $\varphi_{\overline{X}}(u) = e^{(\lambda_1+\ldots+\lambda_n)(e^{u/n}-1)}$ which is **not** the mgf of a Poisson.

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and

  $$\lim_{n \to \infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$

  then the distribution of Z_n approaches the distribution of Z

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and
$$\lim_{n\to\infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$
then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and

$$\lim_{n \to \infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$

  then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

$$\varphi_{Z_n}(u) \quad = \quad [q + pe^u]^n = [(1 - p) + pe^u]^n$$

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and

  $$\lim_{n\to\infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$

  then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

$$\begin{aligned} \varphi_{Z_n}(u) &= [q + pe^u]^n = [(1-p) + pe^u]^n \\ &= [1 + p(e^u - 1)]^n \end{aligned}$$

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and

  $$\lim_{n \to \infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$

  then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

$$\begin{aligned} \varphi_{Z_n}(u) &= [q + pe^u]^n = [(1-p) + pe^u]^n \\ &= [1 + p(e^u - 1)]^n \\ &= \left[1 + \frac{\lambda}{n}(e^u - 1)\right]^n \end{aligned}$$

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and

  $$\lim_{n\to\infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$

  then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

$$\begin{aligned}
\varphi_{Z_n}(u) &= [q + pe^u]^n = [(1-p) + pe^u]^n \\
&= [1 + p(e^u - 1)]^n \\
&= \left[1 + \frac{\lambda}{n}(e^u - 1)\right]^n \\
\lim_{n\to\infty} \varphi_{Z_n}(u) &= \lim_{n\to\infty} \left[1 + \frac{\lambda(e^u - 1)}{n}\right]^n = e^{\lambda(e^u - 1)}
\end{aligned}$$

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and

  $$\lim_{n\to\infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$

  then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

$$\begin{aligned}
\varphi_{Z_n}(u) &= [q + pe^u]^n = [(1-p) + pe^u]^n \\
&= [1 + p(e^u - 1)]^n \\
&= \left[1 + \frac{\lambda}{n}(e^u - 1)\right]^n \\
\lim_{n\to\infty} \varphi_{Z_n}(u) &= \lim_{n\to\infty} \left[1 + \frac{\lambda(e^u - 1)}{n}\right]^n = e^{\lambda(e^u - 1)}
\end{aligned}$$

- Last expression is $\varphi_Z(u)$ of a Poisson distribution with parameter λ.

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and
$$\lim_{n\to\infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$
then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

$$\begin{aligned}
\varphi_{Z_n}(u) &= [q + pe^u]^n = [(1-p) + pe^u]^n \\
&= [1 + p(e^u - 1)]^n \\
&= \left[1 + \frac{\lambda}{n}(e^u - 1)\right]^n \\
\lim_{n\to\infty} \varphi_{Z_n}(u) &= \lim_{n\to\infty} \left[1 + \frac{\lambda(e^u - 1)}{n}\right]^n = e^{\lambda(e^u - 1)}
\end{aligned}$$

- Last expression is $\varphi_Z(u)$ of a Poisson distribution with parameter λ.

Remember what additional condition was required on $\varphi_Z(u)$?

Poisson as a limit of the binomial

- Remember: if we have a sequence of random variables Z_n and
$$\lim_{n\to\infty} \varphi_{Z_n}(u) \to \varphi_Z(u)$$
then the distribution of Z_n approaches the distribution of Z
- Now, consider $Z_n \sim b(p = \lambda/n, n)$, We have,

$$\begin{aligned}
\varphi_{Z_n}(u) &= [q + pe^u]^n = [(1-p) + pe^u]^n \\
&= [1 + p(e^u - 1)]^n \\
&= \left[1 + \frac{\lambda}{n}(e^u - 1)\right]^n \\
\lim_{n\to\infty} \varphi_{Z_n}(u) &= \lim_{n\to\infty} \left[1 + \frac{\lambda(e^u - 1)}{n}\right]^n = e^{\lambda(e^u - 1)}
\end{aligned}$$

- Last expression is $\varphi_Z(u)$ of a Poisson distribution with parameter λ.

It has to be continuous $u = 0$.

Practical use of the limiting distribution (I)

- Whenever $np \to \infty$, normal approximation better.

Practical use of the limiting distribution (I)

- Whenever $np \to \infty$, normal approximation better.
- Poisson approximation best for $\lambda = np < 18$.

Practical use of the limiting distribution (I)

- Whenever $np \to \infty$, normal approximation better.
- Poisson approximation best for $\lambda = np < 18$.
- Particularly useful when np very small (in which case normal approximation is quite poor).

Practical use of the limiting distribution (I)

- Whenever $np \to \infty$, normal approximation better.
- Poisson approximation best for $\lambda = np < 18$.
- Particularly useful when np very small (in which case normal approximation is quite poor).
- Discrete approximation with a discrete distribution: no continuity corrections, no nothing.

Practical use of the limiting distribution (I)

- Whenever *np* $\to \infty$, normal approximation better.
- Poisson approximation best for $\lambda = np < 18$.
- Particularly useful when *np* very small (in which case normal approximation is quite poor).
- Discrete approximation with a discrete distribution: no continuity corrections, no nothing.
- Poisson probabilities $P_X(x) = e^{-\lambda}\lambda^x/x!$ quite easy to compute, even on a pocket calculator.

Practical use of the limiting distribution (I)

- Whenever $np \to \infty$, normal approximation better.
- Poisson approximation best for $\lambda = np < 18$.
- Particularly useful when np very small (in which case normal approximation is quite poor).
- Discrete approximation with a discrete distribution: no continuity corrections, no nothing.
- Poisson probabilities $P_X(x) = e^{-\lambda}\lambda^x/x!$ quite easy to compute, even on a pocket calculator.

What problems would you anticipate calculating $P_X(x)$?

Practical use of the limiting distribution (I)

- Whenever $np \to \infty$, normal approximation better.
- Poisson approximation best for $\lambda = np < 18$.
- Particularly useful when np very small (in which case normal approximation is quite poor).
- Discrete approximation with a discrete distribution: no continuity corrections, no nothing.
- Poisson probabilities $P_X(x) = e^{-\lambda}\lambda^x/x!$ quite easy to compute, even on a pocket calculator.

Large factorials might be the only problem ($69! = 1.711225 \times 10^{98}$).

Practical use of the limiting distribution (II)

- Tables do exist.

Practical use of the limiting distribution (II)

- Tables do exist.
- We have the usual assortment of `{d,p,q,r}pois` functions in R, to assist with any computations.

Practical use of the limiting distribution (II)

- Tables do exist.
- We have the usual assortment of `{d,p,q,r}pois` functions in R, to assist with any computations.
- A useful recurrence:

$$P_X(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \underbrace{\frac{e^{-\lambda}\lambda^{(x-1)}}{(x-1)!}}_{P_X(x-1;\lambda)} \times \frac{\lambda}{x}$$

  so each probability can be obtained from the previous multiplying by $\frac{\lambda}{x}$. (First one, $P_X(0;\lambda) = e^{-\lambda}$.)

Practical use of the limiting distribution (II)

- Tables do exist.
- We have the usual assortment of `{d,p,q,r}pois` functions in R, to assist with any computations.
- A useful recurrence:

$$P_X(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} = \underbrace{\frac{e^{-\lambda}\lambda^{(x-1)}}{(x-1)!}}_{P_X(x-1;\lambda)} \times \frac{\lambda}{x}$$

  so each probability can be obtained from the previous multiplying by $\frac{\lambda}{x}$. (First one, $P_X(0;\lambda) = e^{-\lambda}$.)
- Avoids large factorials.

Practical use of the limiting distribution (III)

```
> dbinom(x=2,size=50,prob=0.1)          # Exact binomial
[1] 0.0779429
> dpois(x=2,lambda=50*0.1)              # Poisson approxi
[1] 0.08422434
> pnorm((2.5-5)/sqrt(50*0.1*.9)) - pnorm((1.5-5)/sqrt(
[1] 0.06981634
> dbinom(x=2,size=500,prob=0.01)        # Exact binomial
[1] 0.08363103
> dpois(x=2,lambda=500*0.01)            # Poisson approxi
[1] 0.08422434
> pnorm((2.5-5)/sqrt(500*0.01*.99)) - pnorm((1.5-5)/sq
[1] 0.07273327
```

The “rare events” model

- Many units, n, with small probability p of failure, and $np < 18$ give a Poisson-distributed number of units failing.

The “rare events” model

- Many units, n, with small probability p of failure, and $np < 18$ give a Poisson-distributed number of units failing.
- Examples:
  - Many soldiers, small probability of dying by horse kick $\Rightarrow$ number of soldiers dead approximately Poisson-distributed.

The “rare events” model

- Many units, n, with small probability p of failure, and $np < 18$ give a Poisson-distributed number of units failing.
- Examples:
    - Many soldiers, small probability of dying by horse kick $\Rightarrow$ number of soldiers dead approximately Poisson-distributed.
    - Many phone lines, small probability of one of them being in use $\Rightarrow$ simultaneaous calls placed at any one moment Poisson-distributed.

The “rare events” model

- Many units, n, with small probability p of failure, and $np < 18$ give a Poisson-distributed number of units failing.
- Examples:
  - Many soldiers, small probability of dying by horse kick $\Rightarrow$ number of soldiers dead approximately Poisson-distributed.
  - Many phone lines, small probability of one of them being in use $\Rightarrow$ simultaneaous calls placed at any one moment Poisson-distributed.
  - Many houses insured against fire, small probability of any of them catching fire in the insurance period $\Rightarrow$ total number of claims in that period Poisson-distributed.

The “rare events” model

- Many units, n, with small probability p of failure, and $np < 18$ give a Poisson-distributed number of units failing.
- Examples:
  - Many soldiers, small probability of dying by horse kick $\Rightarrow$ number of soldiers dead approximately Poisson-distributed.
  - Many phone lines, small probability of one of them being in use $\Rightarrow$ simultaneaous calls placed at any one moment Poisson-distributed.
  - Many houses insured against fire, small probability of any of them catching fire in the insurance period $\Rightarrow$ total number of claims in that period Poisson-distributed.
  - Arrival intervals i.i.d. exponentially distributed, $f_X(x) = \theta e^{-\theta x} \Rightarrow$ total number of arrivals in $(T, T+t)$ Poisson-distributed with $(T, T+t)$ Poisson-distributed with $\lambda = \theta t$.

Example 1 (I)

Consider a company with 120 workers. On average, they spend 10% of their time calling to the outside. They place calls independently of each other.

- What is the mean value of the number of people simultaneously calling outside?

Example 1 (I)

Consider a company with 120 workers. On average, they spend 10% of their time calling to the outside. They place calls independently of each other.

- What is the mean value of the number of people simultaneously calling outside?
- With 16 outgoing phone lines, what is the probability of being able to service all calls?

Example 1 (I)

Consider a company with 120 workers. On average, they spend 10% of their time calling to the outside. They place calls independently of each other.

- What is the mean value of the number of people simultaneously calling outside?
- With 16 outgoing phone lines, what is the probability of being able to service all calls?
- If the company is split in two divisions, with respectively 80 and 40 people and 10 and 6 phone lines, what's the probability of being able to service all calls?

Example 1 (I)

Consider a company with 120 workers. On average, they spend 10% of their time calling to the outside. They place calls independently of each other.

- What is the mean value of the number of people simultaneously calling outside?
- With 16 outgoing phone lines, what is the probability of being able to service all calls?
- If the company is split in two divisions, with respectively 80 and 40 people and 10 and 6 phone lines, what's the probability of being able to service all calls?
- What are your conclusions? Is it better to provide a centralized service or not?

Example 1 (II)

- What is the mean value of the number of people simultaneously calling outside?
- If there are 16 outgoing phone lines, what is the probability of being able to service all calls?
- Two divisions, respectively 80 and 40 people and 10 and 6 lines. Probability of being able to service all calls?

```
> 120 * 0.1
[1] 12
> #
> ppois(16,lambda=12)
[1] 0.898709
> #
> ppois(10,80*0.1) *
   ppois(6,40*0.1)
[1] 0.7255885
```

Example 2

Your are auditing a company. They claim high quality of their records, with a proportion of 0.1% at most containing errors. You screen 4000 records, uncovering 6 mistakes (i.e., a proportion of 0.15%, or 50% larger than their alleged error rate). What would you conclude about the veracity of their claims?

- Assuming their claims are right, total number of errors in 4000 records Poisson distributed, with $\lambda = 4000 \times 0.001 = 4$ in the worst case.
- If $\lambda = 4$, the probability of over 5 errors is

  ```
  > 1 - ppois(5,lambda=4)
  [1] 0.2148696
  ```

  which is by no means small.
- There is no conclusive evidence to challenge their claim: with $\lambda = 4$, 6 errors out of 4000 records is by no means abnormal.

Example 3

Five hundred school children enjoy recreation. The probability that any of them injures himself and comes to the infirmary of the school to have a wound bandaged is $p = 0.01$. How many bandages must the infirmary stock at the beginning of the day so that the probability of running out is less than 0.001?

- The number of children injured is distributed as $\mathcal{P}(\lambda = 5)$.
- Bandages required are less than or equal

  ```
  > qpois(0.999,lambda=5)
  [1] 13
  ```

  with probability 0.999, so enough to stock 13.
- Let's check:

  ```
  > 1 - ppois(12:13,lambda=5)
  [1] 0.002018852 0.000697990
  ```

  We see indeed that 12 would not be enough and 13 is.

Example 4

The probability of a type of cancer in children of school age is 0.001 per children-year (=1 out of 1000 children on the average). You are suspicious of the mobile phone antennas erected in the vicinity of your district public shool, and find out that out of 400 children, 3 have contracted the disease. Is that an abnormal incidence rate?

- The number of cancer cases is distributed as $\mathcal{P}(\lambda = 0.4)$.
- The probability of less than or equal to 0, 1, 2, 3, 4 cases is:

  ```
  > ppois(0:4,lambda=0.4)
  [1] 0.6703200 0.9384481 0.9920737 0.9992237 0.
  ```

  so 3 cases is fairly rare, happening by pure chance less than 1% of the time.

Example 4 (continued)

Setup like of the previous example. You collect data on all 1300 schools with 400 children each within 200m of mobile phone antennas. Have 540 cases of cancer in all, worst one alone had 4 cases. What would you say?

- Total number of cases is $\mathcal{P}(\lambda = 0.4 \times 1300)$. Then,

  ```
  > 1 - ppois(539,lambda=1300*0.4)
  [1] 0.1956853
  ```

  doesn't look abnormal; expected about 19% of the time.
- The school with 4 cases does look abnormal *in isolation:*

  ```
  > 1 - ppois(3,lambda=0.4)
  [1] 0.0007762514
  ```

- As the worst case among the 1300 schools examined, it can no longer be considered abnormal:

  ```
  > 1 - ( ppois(3,lambda=0.4) )^1300
  [1] 0.6356057
  ```

Reminder of some useful relationships

- $b(p,n) \xrightarrow{d} \mathcal{P}(\lambda = np)$ as $n \to \infty$ with $np < 18$.

Reminder of some useful relationships

- $b(p, n) \xrightarrow{d} \mathcal{P}(\lambda = np)$ as $n \to \infty$ with $np < 18$.
- $b(p, n) \xrightarrow{d} N(np, npq)$ as $n \to \infty$ with $np > 18$.

Reminder of some useful relationships

- $b(p, n) \overset{d}{\to} \mathcal{P}(\lambda = np)$ as $n \to \infty$ with $np < 18$.
- $b(p, n) \overset{d}{\to} N(np, npq)$ as $n \to \infty$ with $np > 18$.
- $\mathcal{P}(\lambda) \overset{d}{\to} N(\lambda, \lambda)$ as $\lambda \to \infty$.

Reminder of some useful relationships

- $b(p, n) \xrightarrow{d} \mathcal{P}(\lambda = np)$ as $n \to \infty$ with $np < 18$.
- $b(p, n) \xrightarrow{d} N(np, npq)$ as $n \to \infty$ with $np > 18$.
- $\mathcal{P}(\lambda) \xrightarrow{d} N(\lambda, \lambda)$ as $\lambda \to \infty$.
- If $Z_n \sim b(p, n)$, then $Z_n/n \xrightarrow{p} p$ as $n \to \infty$.

What is ahead of us

- We need to introduce quite a few distributions.

What is ahead of us

- We need to introduce quite a few distributions.
- The fastest presentation requires that for a while we don't try to motivate each one.

What is ahead of us

- We need to introduce quite a few distributions.
- The fastest presentation requires that for a while we don't try to motivate each one.
- You may have a feeling of lack of purpose. . .

What is ahead of us

- We need to introduce quite a few distributions.
- The fastest presentation requires that for a while we don't try to motivate each one.
- You may have a feeling of lack of purpose. . .
- . . .but trust me:

What is ahead of us

- We need to introduce quite a few distributions.
- The fastest presentation requires that for a while we don't try to motivate each one.
- You may have a feeling of lack of purpose. . .
- . . .but trust me:
- It is much the same as having to learn the periodic table of elements before any serious work in Chemistry. . .

What is ahead of us

- We need to introduce quite a few distributions.
- The fastest presentation requires that for a while we don't try to motivate each one.
- You may have a feeling of lack of purpose. . .
- . . .but trust me:
- It is much the same as having to learn the periodic table of elements before any serious work in Chemistry. . .
- . . .or the rudiments of music before you can play piano.

What is ahead of us

- We need to introduce quite a few distributions.
- The fastest presentation requires that for a while we don't try to motivate each one.
- You may have a feeling of lack of purpose. . .
- . . .but trust me:
- It is much the same as having to learn the periodic table of elements before any serious work in Chemistry. . .
- . . .or the rudiments of music before you can play piano.
- *The roots of knowledge are bitter, but the fruit is very sweet* (Rabindranath Tagore)

The gamma function $\Gamma(r)$

- Defined as:

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$$

The gamma function $\Gamma(r)$

- Defined as:

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$$

- Defined for all r, although only for $r > 0$ it will be of interest to us.

The gamma function $\Gamma(r)$

- Defined as:
$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$$
- Defined for all r, although only for $r > 0$ it will be of interest to us.
- Sometimes called Euler integral of the second kind.

The gamma function $\Gamma(r)$

- Defined as:

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$$

- Defined for all r, although only for $r > 0$ it will be of interest to us.
- Sometimes called Euler integral of the second kind.
- Does not have closed form; value can be computed analytically for certain values of r, numerically for others.

The gamma function $\Gamma(r)$

- Defined as:
$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$$
- Defined for all r, although only for $r > 0$ it will be of interest to us.
- Sometimes called Euler integral of the second kind.
- Does not have closed form; value can be computed analytically for certain values of r, numerically for others.
- Interestingly, $\Gamma(r) = (r-1)!$ for natural r.

How do you think $\Gamma(r)$ changes with r?

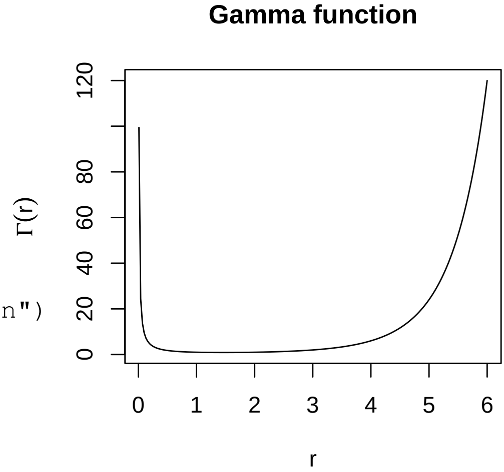
The gamma function $\Gamma(r)$

- Defined as:
$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt$$
- Defined for all r, although only for $r > 0$ it will be of interest to us.
- Sometimes called Euler integral of the second kind.
- Does not have closed form; value can be computed analytically for certain values of r, numerically for others.
- Interestingly, $\Gamma(r) = (r-1)!$ for natural r.

Clearly, $\Gamma(r) \to \infty$ as $r \to \infty$, but also as $r \to 0$.

$\Gamma(r)$ in R

```
> gamma(5)
[1] 24
> factorial(4)
[1] 24
> curve(gamma,from=0.01,
       to=6,n=200,
       ylab=expression(
             Gamma(r)
            ),
       xlab="r",
       main="Gamma function")
```



The gamma distribution $\gamma(a, r)$ (I).

- It is clear that

$$F_X(x) = \frac{1}{\Gamma(r)} \int_0^x t^{r-1} e^{-t} dt$$

is a well defined distribution on $[0, \infty)$.

The gamma distribution $\gamma(a, r)$ (I).

- It is clear that
$$F_X(x) = \frac{1}{\Gamma(r)} \int_0^x t^{r-1} e^{-t} dt$$
is a well defined distribution on $[0, \infty)$.
- If we make the change $t \to at$ for $a > 0$ right hand side still defines the $\gamma(a, r)$ distribution function:
$$\frac{a^r}{\Gamma(r)} \int_0^x t^{r-1} e^{-at} dt$$

The gamma distribution $\gamma(a, r)$ (I).

- It is clear that

$$F_X(x) = \frac{1}{\Gamma(r)} \int_0^x t^{r-1} e^{-t} dt$$

  is a well defined distribution on $[0, \infty)$.
- If we make the change $t \to at$ for $a > 0$ right hand side still defines the $\gamma(a, r)$ distribution function:

$$\frac{a^r}{\Gamma(r)} \int_0^x t^{r-1} e^{-at} dt$$

- Density function therefore is:

$$f_X(x) = \frac{a^r}{\Gamma(r)} t^{r-1} e^{-at}$$

The gamma distribution $\gamma(a, r)$ (II).

- Alternative parameterizations:

$$\begin{aligned} f_X(x) &= \frac{a^r}{\Gamma(r)} t^{r-1} e^{-at} \\ f_X(x) &= \frac{1}{\Gamma(r) s^r} t^{r-1} e^{-t/s} \end{aligned}$$

The gamma distribution $\gamma(a, r)$ (II).

- Alternative parameterizations:

$$\begin{aligned} f_X(x) &= \frac{a^r}{\Gamma(r)} t^{r-1} e^{-at} \\ f_X(x) &= \frac{1}{\Gamma(r) s^r} t^{r-1} e^{-t/s} \end{aligned}$$

- In either case, r is the “shape” parameter and a (or s) the “scale” or “rate” parameter.

The gamma distribution $\gamma(a, r)$ (II).

- Alternative parameterizations:

$$\begin{aligned} f_X(x) &= \frac{a^r}{\Gamma(r)} t^{r-1} e^{-at} \\ f_X(x) &= \frac{1}{\Gamma(r) s^r} t^{r-1} e^{-t/s} \end{aligned}$$

- In either case, r is the “shape” parameter and a (or s) the “scale” or “rate” parameter.
- Important to check definition when using tables…

The gamma distribution $\gamma(a, r)$ (II).

- Alternative parameterizations:

$$\begin{aligned} f_X(x) &= \frac{a^r}{\Gamma(r)} t^{r-1} e^{-at} \\ f_X(x) &= \frac{1}{\Gamma(r) s^r} t^{r-1} e^{-t/s} \end{aligned}$$

- In either case, r is the "shape" parameter and a (or s) the "scale" or "rate" parameter.
- Important to check definition when using tables…
- …although you will rarely use the $\gamma(a, r)$ directly.

The gamma distribution $\gamma(a, r)$ in R

- Usual assortment of `[d,p,q,r]gamma` functions.

The gamma distribution $\gamma(a, r)$ in R

- Usual assortment of `[d,p,q,r]gamma` functions.
- Sintax is, e.g. `dgamma(x, shape, rate, scale)`

$$
\begin{aligned}
f_X(x) &= \frac{a^r}{\Gamma(r)} t^{r-1} e^{-at} \\
f_X(x) &= \frac{1}{\Gamma(r) s^r} t^{r-1} e^{-t/s}
\end{aligned}
$$

- In either case, r is the “shape” parameter and a the “rate” (or s is the “scale”) parameter.

The gamma distribution $\gamma(a, r)$ in R

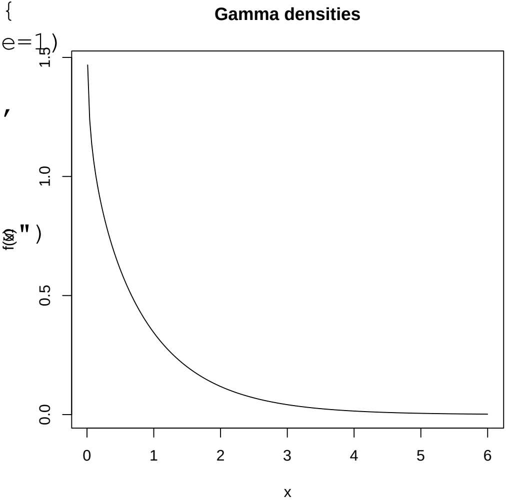
- Usual assortment of `[d,p,q,r]gamma` functions.
- Sintax is, e.g. `dgamma(x, shape, rate, scale)`

$$\begin{aligned} f_X(x) &= \frac{a^r}{\Gamma(r)} t^{r-1} e^{-at} \\ f_X(x) &= \frac{1}{\Gamma(r) s^r} t^{r-1} e^{-t/s} \end{aligned}$$

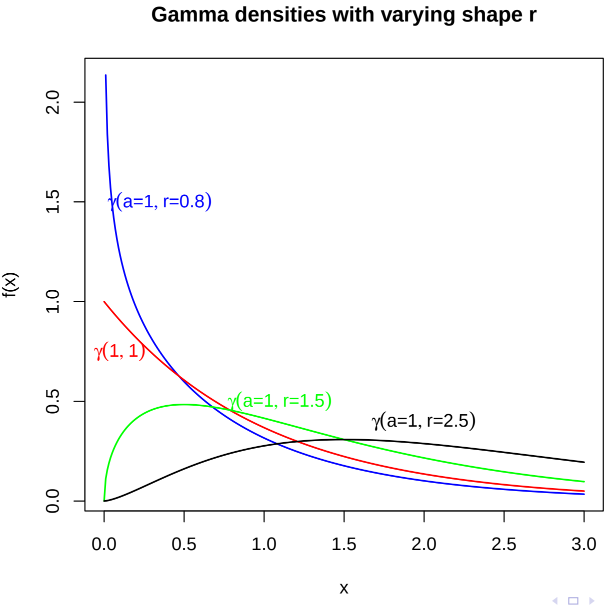
- In either case, r is the “shape” parameter and a the “rate” (or s is the “scale”) parameter.
- Only one of rate or scale needs to be specified.

What does the $\gamma(a, r)$ look like? (I)

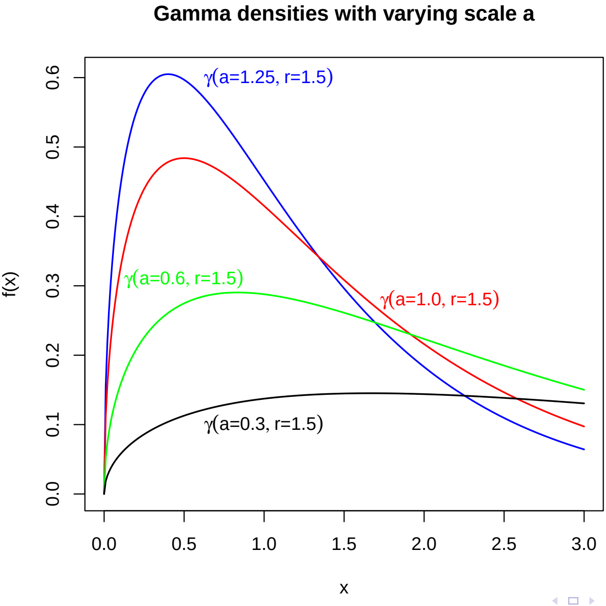
```
> gammar0.9 <- function(x) {
   dgamma(x,shape=0.9,scale=1)
   }
> curve(gammar0.9,from=0.01,
      to=6,n=200,
      ylab="f(x)",
      xlab="x",
      main="Gamma densities")
```



What does the $\gamma(a, r)$ look like? (II)



What does the $\gamma(a, r)$ look like? (III)



Moment generating function of the $\gamma(a, r)$ (I).

$$\varphi_X(u) \quad = \quad E[e^{uX}] = \int_o^\infty \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} e^{ux} dx$$

Moment generating function of the $\gamma(a, r)$ (I).

$$\begin{aligned} \varphi_X(u) &= E[e^{uX}] = \int_o^{\infty} \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} e^{ux} dx \\ &= \frac{a^r}{\Gamma(r)} \int_o^{\infty} x^{r-1} e^{-(a-u)x} dx \end{aligned}$$

Moment generating function of the $\gamma(a, r)$ (I).

$$\begin{aligned}
\varphi_X(u) &= E[e^{uX}] = \int_o^\infty \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} e^{ux} dx \\
&= \frac{a^r}{\Gamma(r)} \int_o^\infty x^{r-1} e^{-(a-u)x} dx \\
&= \frac{a^r}{\Gamma(r)} \left[\frac{(a-u)^r}{\Gamma(r)} \right]^{-1}
\end{aligned}$$

Moment generating function of the $\gamma(a, r)$ (I).

$$\begin{aligned}
\varphi_X(u) &= E[e^{uX}] = \int_o^\infty \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} e^{ux} dx \\
&= \frac{a^r}{\Gamma(r)} \int_o^\infty x^{r-1} e^{-(a-u)x} dx \\
&= \frac{a^r}{\Gamma(r)} \left[\frac{(a-u)^r}{\Gamma(r)} \right]^{-1} \\
&= \left(1 - \frac{u}{a} \right)^{-r}
\end{aligned}$$

Moment generating function of the $\gamma(a, r)$ (I).

$$\begin{aligned}
\varphi_X(u) &= E[e^{uX}] = \int_o^\infty \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} e^{ux} dx \\
&= \frac{a^r}{\Gamma(r)} \int_o^\infty x^{r-1} e^{-(a-u)x} dx \\
&= \frac{a^r}{\Gamma(r)} \left[\frac{(a-u)^r}{\Gamma(r)} \right]^{-1} \\
&= \left(1 - \frac{u}{a}\right)^{-r}
\end{aligned}$$

See how the integral went away?

Moment generating function of the $\gamma(a, r)$ (I).

$$\begin{aligned}
\varphi_X(u) &= E[e^{uX}] = \int_o^\infty \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} e^{ux} dx \\
&= \frac{a^r}{\Gamma(r)} \int_o^\infty x^{r-1} e^{-(a-u)x} dx \\
&= \frac{a^r}{\Gamma(r)} \left[\frac{(a-u)^r}{\Gamma(r)} \right]^{-1} \\
&= \left(1 - \frac{u}{a}\right)^{-r}
\end{aligned}$$

It is equal to content within brackets in next-to-last expression.

Moment generating function of the $\gamma(a, r)$ (II).

- Let $X = X_1 + \ldots + X_n$ independent gamma random variables with equal scale parameter and respectively $r_1, \ldots, r_n$ as shape parameter. Then:

  $$\begin{aligned} \varphi_X(u) &= \left(1 - \frac{u}{a}\right)^{-r_1} \cdots \left(1 - \frac{u}{a}\right)^{-r_n} \\ &= \left(1 - \frac{u}{a}\right)^{-(r_1 + \ldots + r_n)} \end{aligned}$$

  so X is $\gamma(a, r_1 + \ldots + r_n)$ distributed.

Moment generating function of the $\gamma(a, r)$ (II).

- Let $X = X_1 + \ldots + X_n$ independent gamma random variables with equal scale parameter and respectively $r_1, \ldots, r_n$ as shape parameter. Then:

$$\begin{aligned} \varphi_X(u) &= \left(1 - \frac{u}{a}\right)^{-r_1} \cdots \left(1 - \frac{u}{a}\right)^{-r_n} \\ &= \left(1 - \frac{u}{a}\right)^{-(r_1 + \ldots + r_n)} \end{aligned}$$

  so X is $\gamma(a, r_1 + \ldots + r_n)$ distributed.
- The same does not hold if the scale parameters are not equal.

Moment generating function of the $\gamma(a, r)$ (III).

- Let $Y = cX$ with $X \sim \gamma(a, r)$. Then $Y \sim \gamma(a/c, r)$.

Moment generating function of the $\gamma(a, r)$ (III).

- Let $Y = cX$ with $X \sim \gamma(a, r)$. Then $Y \sim \gamma(a/c, r)$.
- Simple rescaling of a gamma gives again a gamma with arbitrary first parameter.

Moment generating function of the $\gamma(a, r)$ (III).

- Let $Y = cX$ with $X \sim \gamma(a, r)$. Then $Y \sim \gamma(a/c, r)$.
- Simple rescaling of a gamma gives again a gamma with arbitrary first parameter.
- Proof is trivial:

$$\begin{aligned}\varphi_{cX}(u) &= E[e^{ucX}] = \varphi_X(cu) \\ &= \left(1 - \frac{cu}{a}\right)^{-r} \\ &= \left(1 - \frac{u}{a/c}\right)^{-r}\end{aligned}$$

so cX is $\gamma(a/c, r)$ distributed.

Mean and variance of $\gamma(a, r)$.

- Mean and variance are now easy to compute:

$$\begin{aligned}
\left[\varphi'_X(u)\right]_{u=0} &= \left[-r\left(1-\frac{u}{a}\right)^{-r-1}\left(-\frac{1}{a}\right)\right]_{u=0} \\
&= \frac{r}{a} \\
\left[\varphi''_X(u)\right]_{u=0} &= \left[r(r+1)\left(1-\frac{u}{a}\right)^{-r-2}\left(-\frac{1}{a}\right)^2\right]_{u=0} \\
&= \frac{r^2}{a^2}+\frac{r}{a^2}
\end{aligned}$$

Hence, $m = r/a$ and $\sigma^2 = \alpha_2 - (\alpha_1)^2 = r/a^2$.

Mean and variance of $\gamma(a, r)$.

- Mean and variance are now easy to compute:

$$\begin{aligned}
\left[\varphi_X'(u)\right]_{u=0} &= \left[-r\left(1-\frac{u}{a}\right)^{-r-1}\left(-\frac{1}{a}\right)\right]_{u=0} \\
&= \frac{r}{a} \\
\left[\varphi_X''(u)\right]_{u=0} &= \left[r(r+1)\left(1-\frac{u}{a}\right)^{-r-2}\left(-\frac{1}{a}\right)^2\right]_{u=0} \\
&= \frac{r^2}{a^2}+\frac{r}{a^2}
\end{aligned}$$

  Hence, $m = r/a$ and $\sigma^2 = \alpha_2 - (\alpha_1)^2 = r/a^2$.
- It can also be checked that the mode is at $\frac{r-1}{a}$ (or zero, in case $r < 1$ and monotone decreasing density).

Mean and variance of $\gamma(a, r)$.

- Mean and variance are now easy to compute:

$$\begin{aligned}
\left[\varphi_X'(u)\right]_{u=0} &= \left[-r\left(1-\frac{u}{a}\right)^{-r-1}\left(-\frac{1}{a}\right)\right]_{u=0} \\
&= \frac{r}{a} \\
\left[\varphi_X''(u)\right]_{u=0} &= \left[r(r+1)\left(1-\frac{u}{a}\right)^{-r-2}\left(-\frac{1}{a}\right)^2\right]_{u=0} \\
&= \frac{r^2}{a^2} + \frac{r}{a^2}
\end{aligned}$$

  Hence, $m = r/a$ and $\sigma^2 = \alpha_2 - (\alpha_1)^2 = r/a^2$.
- It can also be checked that the mode is at $\frac{r-1}{a}$ (or zero, in case $r < 1$ and monotone decreasing density).

Mean and variance of $\gamma(a, r)$.

- Mean and variance are now easy to compute:

$$\begin{aligned}
\left[\varphi_X'(u)\right]_{u=0} &= \left[-r\left(1-\frac{u}{a}\right)^{-r-1}\left(-\frac{1}{a}\right)\right]_{u=0} \\
&= \frac{r}{a} \\
\left[\varphi_X''(u)\right]_{u=0} &= \left[r(r+1)\left(1-\frac{u}{a}\right)^{-r-2}\left(-\frac{1}{a}\right)^2\right]_{u=0} \\
&= \frac{r^2}{a^2}+\frac{r}{a^2}
\end{aligned}$$

  Hence, $m = r/a$ and $\sigma^2 = \alpha_2 - (\alpha_1)^2 = r/a^2$.
- It can also be checked that the mode is at $\frac{r-1}{a}$ (or zero, in case $r < 1$ and monotone decreasing density).

How would you choose $\gamma(a, r)$ with mean 2 and variance 5?

Mean and variance of $\gamma(a, r)$.

- Mean and variance are now easy to compute:

$$\begin{aligned}
\left[\varphi_X'(u)\right]_{u=0} &= \left[-r\left(1-\frac{u}{a}\right)^{-r-1}\left(-\frac{1}{a}\right)\right]_{u=0} \\
&= \frac{r}{a} \\
\left[\varphi_X''(u)\right]_{u=0} &= \left[r(r+1)\left(1-\frac{u}{a}\right)^{-r-2}\left(-\frac{1}{a}\right)^2\right]_{u=0} \\
&= \frac{r^2}{a^2}+\frac{r}{a^2}
\end{aligned}$$

  Hence, $m = r/a$ and $\sigma^2 = \alpha_2 - (\alpha_1)^2 = r/a^2$.
- It can also be checked that the mode is at $\frac{r-1}{a}$ (or zero, in case $r < 1$ and monotone decreasing density).

Matching moments.

Exponential distribution $\exp(\lambda)$ (I)

- A very important particular case occurs when $r = 1$. Then,

$$\gamma(a, r = 1) \quad = \quad \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} = a e^{-ax}$$

Exponential distribution $\exp(\lambda)$ (I)

- A very important particular case occurs when $r = 1$. Then,

$$\gamma(a, r = 1) \quad = \quad \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} = a e^{-ax}$$

- Conventionally, a denoted by λ. Distribution called exponential, $\exp(\lambda)$.

Exponential distribution $\exp(\lambda)$ (I)

- A very important particular case occurs when $r = 1$. Then,

$$\gamma(a, r = 1) \quad = \quad \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} = a e^{-ax}$$

- Conventionally, a denoted by λ. Distribution called exponential, $\exp(\lambda)$.
- Alternative in terms of $\theta = 1/\lambda$:

$$f_X(x) = \lambda e^{-\lambda x} = \frac{1}{\theta} e^{-x/\theta}$$

Exponential distribution $\exp(\lambda)$ (I)

- A very important particular case occurs when $r = 1$. Then,

$$\gamma(a, r = 1) \quad = \quad \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} = ae^{-ax}$$

- Conventionally, a denoted by λ. Distribution called exponential, $\exp(\lambda)$.
- Alternative in terms of $\theta = 1/\lambda$:

$$f_X(x) = \lambda e^{-\lambda x} = \frac{1}{\theta} e^{-x/\theta}$$

- If we stick with the λ-parameterization, $m = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$.

Exponential distribution $\exp(\lambda)$ (I)

- A very important particular case occurs when $r = 1$. Then,

$$\gamma(a, r = 1) \quad = \quad \frac{a^r}{\Gamma(r)} x^{r-1} e^{-ax} = ae^{-ax}$$

- Conventionally, a denoted by λ. Distribution called exponential, $\exp(\lambda)$.
- Alternative in terms of $\theta = 1/\lambda$:

$$f_X(x) = \lambda e^{-\lambda x} = \frac{1}{\theta} e^{-x/\theta}$$

- If we stick with the λ-parameterization, $m = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$.
- Clearly, $F_X(x) = 1 - e^{-\lambda x}$.

Exponential distribution $\exp(\lambda)$ (II)

- The moment generating function comes straight from the $\gamma(a = \lambda, r = 1)$ general case:

$$\varphi_X(u) = \left(1 - \frac{u}{\lambda}\right)^{-1}$$

Exponential distribution $\exp(\lambda)$ (II)

- The moment generating function comes straight from the $\gamma(a = \lambda, r = 1)$ general case:

$$\varphi_X(u) = \left(1 - \frac{u}{\lambda}\right)^{-1}$$

- With $f_X(x) = \lambda e^{-\lambda x}$ and $F_X(x) = 1 - e^{-\lambda x}$ no need of tables; however, still the usual R functions `{d,p,q,r}exp`.

Exponential distribution $\exp(\lambda)$ (II)

- The moment generating function comes straight from the $\gamma(a = \lambda, r = 1)$ general case:

$$\varphi_X(u) = \left(1 - \frac{u}{\lambda}\right)^{-1}$$

- With $f_X(x) = \lambda e^{-\lambda x}$ and $F_X(x) = 1 - e^{-\lambda x}$ no need of tables; however, still the usual R functions `{d,p,q,r}exp`.
- Sintax: `dexp(x,rate)` where `rate` is λ.

Exponential distribution $\exp(\lambda)$ (II)

- The moment generating function comes straight from the $\gamma(a = \lambda, r = 1)$ general case:

$$\varphi_X(u) = \left(1 - \frac{u}{\lambda}\right)^{-1}$$

- With $f_X(x) = \lambda e^{-\lambda x}$ and $F_X(x) = 1 - e^{-\lambda x}$ no need of tables; however, still the usual R functions `{d,p,q,r}exp`.
- Sintax: `dexp(x,rate)` where `rate` is λ.

What if we sum n independent exponential variables with the same λ?

Exponential distribution $\exp(\lambda)$ (II)

- The moment generating function comes straight from the $\gamma(a = \lambda, r = 1)$ general case:

$$\varphi_X(u) = \left(1 - \frac{u}{\lambda}\right)^{-1}$$

- With $f_X(x) = \lambda e^{-\lambda x}$ and $F_X(x) = 1 - e^{-\lambda x}$ no need of tables; however, still the usual R functions `{d,p,q,r}exp`.
- Sintax: `dexp(x,rate)` where `rate` is λ.

We get a variable distributed as $\gamma(\lambda, n)$.

Square-normal distribution

- If $X \sim N(0,1)$, what is the distribution of $Y = X^2$?

Square-normal distribution

- If $X \sim N(0,1)$, what is the distribution of $Y = X^2$?
- $F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y})$.

Square-normal distribution

- If $X \sim N(0,1)$, what is the distribution of $Y = X^2$?
- $F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y})$.
- Therefore $F_Y(y) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$, and

$$\begin{aligned} f_Y(y) &= \phi(\sqrt{y}) \times \frac{1}{2\sqrt{y}} - \phi(-\sqrt{y}) \times \left(-\frac{1}{2\sqrt{y}}\right) \\ &= \phi(\sqrt{y})\frac{1}{\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-\frac{y}{2}} \qquad\qquad (y > 0) \end{aligned}$$

Square-normal distribution

- If $X \sim N(0,1)$, what is the distribution of $Y = X^2$?
- $F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y})$.
- Therefore $F_Y(y) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$, and

$$\begin{aligned} f_Y(y) &= \phi(\sqrt{y}) \times \frac{1}{2\sqrt{y}} - \phi(-\sqrt{y}) \times \left(-\frac{1}{2\sqrt{y}}\right) \\ &= \phi(\sqrt{y})\frac{1}{\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-\frac{y}{2}} \qquad\qquad (y > 0) \end{aligned}$$

What density is this a particular case of?

Square-normal distribution

- If $X \sim N(0,1)$, what is the distribution of $Y = X^2$?
- $F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y})$.
- Therefore $F_Y(y) = \Phi(\sqrt{y}) - \Phi(-\sqrt{y})$, and

$$\begin{aligned} f_Y(y) &= \phi(\sqrt{y}) \times \frac{1}{2\sqrt{y}} - \phi(-\sqrt{y}) \times \left(-\frac{1}{2\sqrt{y}}\right) \\ &= \phi(\sqrt{y})\frac{1}{\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-\frac{y}{2}} \qquad\qquad (y > 0) \end{aligned}$$

It is clearly a $\gamma(a = \frac{1}{2}, r = \frac{1}{2})$.

Things you can easily check:

- If Y is square-normal, $E[Y] = 1$.
  (Try it both ways, using the "gamma ancestry" of Y and the direct approach: remember $Y = X^2$ and $X \sim N(0, 1)$).

Things you can easily check:

- If Y is square-normal, $E[Y] = 1$.
  (Try it both ways, using the “gamma ancestry” of Y and the direct approach: remember $Y = X^2$ and $X \sim N(0,1)$).
- If Y is square-normal, its variance is 2.

Things you can easily check:

- If Y is square-normal, $E[Y] = 1$.
  (Try it both ways, using the "gamma ancestry" of Y and the direct approach: remember $Y = X^2$ and $X \sim N(0,1)$).
- If Y is square-normal, its variance is 2.
- If $X_1, \ldots, X_n$ are i.i.d $N(0,1)$, then $Y = X_1^2 + \ldots + X_n^2$ is distributed as $\gamma(\frac{1}{2}, \frac{n}{2})$.

Things you can easily check:

- If Y is square-normal, $E[Y] = 1$.
  (Try it both ways, using the “gamma ancestry” of Y and the direct approach: remember $Y = X^2$ and $X \sim N(0, 1)$).
- If Y is square-normal, its variance is 2.
- If $X_1, \ldots, X_n$ are i.i.d $N(0, 1)$, then $Y = X_1^2 + \ldots + X_n^2$ is distributed as $\gamma(\frac{1}{2}, \frac{n}{2})$.
- If X is exponential(λ), $2\lambda X$ is $\gamma(\frac{1}{2}, 1)$.

Things you can easily check:

- If Y is square-normal, $E[Y] = 1$.
  (Try it both ways, using the "gamma ancestry" of Y and the direct approach: remember $Y = X^2$ and $X \sim N(0, 1)$).
- If Y is square-normal, its variance is 2.
- If $X_1, \ldots, X_n$ are i.i.d $N(0, 1)$, then $Y = X_1^2 + \ldots + X_n^2$ is distributed as $\gamma(\frac{1}{2}, \frac{n}{2})$.
- If X is exponential(λ), $2\lambda X$ is $\gamma(\frac{1}{2}, 1)$.
- Mimic the method used to derive the square-normal density to find the log-normal density, i.e., the density of Y such that $\log_e(Y)$ is normal.

The χ_n^2 distribution

- It is just the $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$ obtained in last slide. . .

The χ_n^2 distribution

- It is just the $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$ obtained in last slide…
- …or, if you prefer, the distribution of the sum of n independent $N(0, 1)$ squared, each of which is $\gamma(a = \frac{1}{2}, r = \frac{1}{2})$

The χ^2_n distribution

- It is just the $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$ obtained in last slide…
- …or, if you prefer, the distribution of the sum of n independent $N(0, 1)$ squared, each of which is $\gamma(a = \frac{1}{2}, r = \frac{1}{2})$
- As particular case of a $\gamma(a, r)$ we know:

The χ^2_n distribution

- It is just the $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$ obtained in last slide…
- …or, if you prefer, the distribution of the sum of n independent $N(0,1)$ squared, each of which is $\gamma(a = \frac{1}{2}, r = \frac{1}{2})$
- As particular case of a $\gamma(a, r)$ we know:

$$m = n \quad \sigma^2 = 2n \quad \varphi_Y(u) = (1 - 2u)^{-\frac{n}{2}}$$

The χ^2_n distribution

- It is just the $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$ obtained in last slide…
- …or, if you prefer, the distribution of the sum of n independent $N(0, 1)$ squared, each of which is $\gamma(a = \frac{1}{2}, r = \frac{1}{2})$
- As particular case of a $\gamma(a, r)$ we know:

$$\begin{array}{ccc} m = n & \sigma^2 = 2n & \varphi_Y(u) = (1 - 2u)^{-\frac{n}{2}} \\ m = r/a & \sigma^2 = r/a^2 & \varphi_Y(u) = \left(1 - \frac{u}{a}\right)^{-r} \end{array}$$

The χ_n^2 distribution

- It is just the $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$ obtained in last slide…
- …or, if you prefer, the distribution of the sum of n independent $N(0, 1)$ squared, each of which is $\gamma(a = \frac{1}{2}, r = \frac{1}{2})$
- As particular case of a $\gamma(a, r)$ we know:

$$\begin{array}{ccc} m = n & \sigma^2 = 2n & \varphi_Y(u) = (1 - 2u)^{-\frac{n}{2}} \\ m = r/a & \sigma^2 = r/a^2 & \varphi_Y(u) = \left(1 - \frac{u}{a}\right)^{-r} \end{array}$$

- n usually called “degrees of freedom”.

What does it look like? (I)

- The density is,

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-x/2}$$

What does it look like? (I)

- The density is,

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2}$$

- As it is a $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$, will be monotone decreasing for $r \leq 1$ ($\Rightarrow n \leq 2$).

What does it look like? (I)

- The density is,

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2}$$

- As it is a $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$, will be monotone decreasing for $r \leq 1$ ($\Rightarrow n \leq 2$).
- For $n > 2$ a single maximum and a long right tail (right-skewed).

What does it look like? (I)

- The density is,

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-x/2}$$

- As it is a $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$, will be monotone decreasing for $r \leq 1$ ($\Rightarrow n \leq 2$).
- For $n > 2$ a single maximum and a long right tail (right-skewed).
- Becomes closer to symmetric as n grows.

What does it look like? (I)

- The density is,

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-x/2}$$

- As it is a $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$, will be monotone decreasing for $r \leq 1$ ($\Rightarrow n \leq 2$).
- For $n > 2$ a single maximum and a long right tail (right-skewed).
- Becomes closer to symmetric as n grows.

What do you think the χ^2_n converges to as $n \to \infty$?

What does it look like? (I)

- The density is,

$$f_X(x) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-x/2}$$

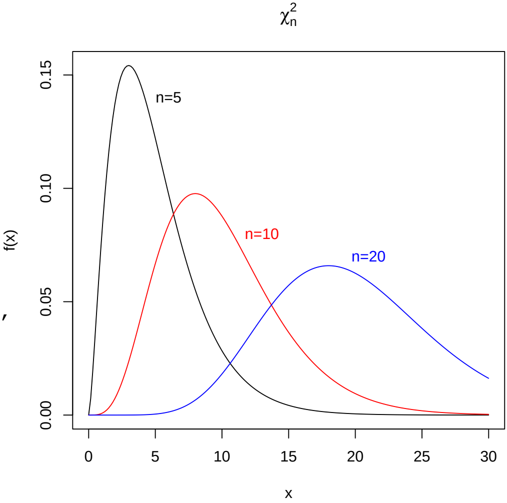
- As it is a $\gamma(a = \frac{1}{2}, r = \frac{n}{2})$, will be monotone decreasing for $r \leq 1$ ($\Rightarrow n \leq 2$).
- For $n > 2$ a single maximum and a long right tail (right-skewed).
- Becomes closer to symmetric as n grows.

What do you think the χ^2_n converges to as $n \to \infty$?

$\chi^2_n \xrightarrow{\mathrm{d}} N(n, 2n)$ by the CLT.

What does it look like? (II)

```
> chisqn <- function(x) {
     dchisq(x,df=n)
   }
> n <- 5
> curve(chisqn,
   from=0.0,to=30,n=200,
   ylab="f(x)",xlab="x",
   main=expression(chi[n]^2))
> n <- 10
> curve(chisqn,from=0.0,col="red",
       to=30,n=200,add=TRUE)
> n <- 20
> curve(chisqn,from=0.0,col="blue",
       to=30,n=200,add=TRUE)
> text(6,0.14,"n=5")
> text(13,0.08,"n=10",col="red")
> text(21,0.07,"n=20",col="blue")
```



Non-central χ_n^2 variables

- The ordinary or “central” χ_n^2 is the sum of n independent $N(0,1)$ squared.

Non-central χ^2_n variables

- The ordinary or “central” χ^2_n is the sum of n independent $N(0, 1)$ squared.
- If the squared normal variables have non-zero mean, we have instead the “non central” chi square.

Non-central χ^2_n variables

- The ordinary or “central” χ^2_n is the sum of n independent $N(0,1)$ squared.
- If the squared normal variables have non-zero mean, we have instead the “non central” chi square.
- If $Y = X_1^2 + \ldots + X_n^2$ with $X_i \sim N(m_i, 1)$, then $Y \sim \chi^2_n(\delta)$ (the “non central” chi square).

Non-central χ^2_n variables

- The ordinary or "central" χ^2_n is the sum of n independent $N(0,1)$ squared.
- If the squared normal variables have non-zero mean, we have instead the "non central" chi square.
- If $Y = X_1^2 + \ldots + X_n^2$ with $X_i \sim N(m_i, 1)$, then $Y \sim \chi^2_n(\delta)$ (the "non central" chi square).
- $\delta = m_1^2 + \ldots + m_n^2$ is the so-called "non-centrality parameter".

Non-central χ^2_n variables

- The ordinary or “central” χ^2_n is the sum of n independent $N(0,1)$ squared.
- If the squared normal variables have non-zero mean, we have instead the “non central” chi square.
- If $Y = X_1^2 + \ldots + X_n^2$ with $X_i \sim N(m_i, 1)$, then $Y \sim \chi^2_n(\delta)$ (the “non central” chi square).
- $\delta = m_1^2 + \ldots + m_n^2$ is the so-called “non-centrality parameter”.
- Some tables/books define the non-centrality parameter as $\delta = \frac{1}{2}(m_1^2 + \ldots + m_n^2)$, so check.

χ^2_n in R

Usual set of functions: `{d,p,q,r}chisq`.

```
>     dchisq(15.3,12)
[1] 0.05196885
>     pchisq(15.3,12)
[1] 0.7745611
>     qchisq(0.99,12)
[1] 26.21697
>     qchisq(0.99,12,ncp=15)
[1] 52.15618
```

Snedecor's $\mathcal{F}_{m,n}$

- The ratio of two χ^2_m and χ^2_n independent of each other each divided by their degrees of freedom,

  $$\frac{\chi^2_m/m}{\chi^2_n/n}$$

  follows a distribution named "Snedecor's $\mathcal{F}_{m,n}$" (after George W. Snedecor (1882 -1974)).

Snedecor's $\mathcal{F}_{m,n}$

- The ratio of two χ^2_m and χ^2_n independent of each other each divided by their degrees of freedom,

  $$\frac{\chi^2_m/m}{\chi^2_n/n}$$

  follows a distribution named "Snedecor's $\mathcal{F}_{m,n}$" (after George W. Snedecor (1882 -1974)).
- Fairly complex density,

  $$f_X(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}} \Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{n}{2}\right)} x^{m/2-1} (n + mx)^{-(n+n)/2}$$

Snedecor's $\mathcal{F}_{m,n}$

- The ratio of two χ^2_m and χ^2_n independent of each other each divided by their degrees of freedom,

  $$\frac{\chi^2_m/m}{\chi^2_n/n}$$

  follows a distribution named "Snedecor's $\mathcal{F}_{m,n}$" (after George W. Snedecor (1882 -1974)).
- Fairly complex density,

  $$f_X(x) = \frac{m^{\frac{m}{2}} n^{\frac{n}{2}} \Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} x^{m/2-1}(n+mx)^{-(n+n)/2}$$

- For $n > 2$, $m = n/(n-2)$ if $n > 2$ and for $n > 4$,

  $$\sigma^2 = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$$

Use of tables for $\mathcal{F}_{m,n}$

- Same as we did not need tables of $b(p, n)$ for $p > 0.5$, we can do with tables for the $\mathcal{F}_{m,n}$ for $\alpha < 0.5$ and obtain the rest indirectly.

Use of tables for $\mathcal{F}_{m,n}$

- Same as we did not need tables of $b(p, n)$ for $p > 0.5$, we can do with tables for the $\mathcal{F}_{m,n}$ for $\alpha < 0.5$ and obtain the rest indirectly.
- If $X \sim \mathcal{F}_{m,n}$, trick is to use

$$\begin{aligned} 1 - \alpha = P(X < \mathcal{F}^{\alpha}_{m,n}) &= P\left(\frac{\chi^2_m/m}{\chi^2_n/n} < \mathcal{F}^{\alpha}_{m,n}\right) \\ &= P\left(\frac{\chi^2_n/n}{\chi^2_m/m} > \frac{1}{\mathcal{F}^{\alpha}_{m,n}}\right) \end{aligned}$$

This shows,

$$\frac{1}{\mathcal{F}^{\alpha}_{m,n}} = \mathcal{F}^{1-\alpha}_{n,m}$$

Non-central versions of $\mathcal{F}_{m,n}$

- If the χ^2 in the numerator has non-centrality parameter δ, the resulting $\mathcal{F}_{m,n}$ is called non-central with the same non-centrality parameter.

Non-central versions of $\mathcal{F}_{m,n}$

- If the χ^2 in the numerator has non-centrality parameter δ, the resulting $\mathcal{F}_{m,n}$ is called non-central with the same non-centrality parameter.
- If both numerator and denominator are non-central χ^2, the ratio is a doubly non-central $\mathcal{F}_{m,n}$.

Non-central versions of $\mathcal{F}_{m,n}$

- If the χ^2 in the numerator has non-centrality parameter δ, the resulting $\mathcal{F}_{m,n}$ is called non-central with the same non-centrality parameter.
- If both numerator and denominator are non-central χ^2, the ratio is a doubly non-central $\mathcal{F}_{m,n}$.
- Tables in general for only the ordinary or central case.

$\mathcal{F}_{m,n}$ in R

As usual, `{d,p,q,r}f` functions.

```
> pf(3.23, 5, 12)      # Prob left 3.23 in F(5;12)
[1] 0.9554027
> qf(0.95, 5, 12)      # Value leaving a tail of 0.
[1] 3.105875
> qf(0.99, 5, 12)      # Id. for tail of 0.01
[1] 5.064343
> qf(0.99, 5, 12, 8)   # Id. for a non-central F
[1] 11.62582
>                      # with ncp=8
```

What does the $\mathcal{F}_{m,n}$ look like? (I)

- If n not too small, shape close to scaled χ^2_m.

What does the $\mathcal{F}_{m,n}$ look like? (I)

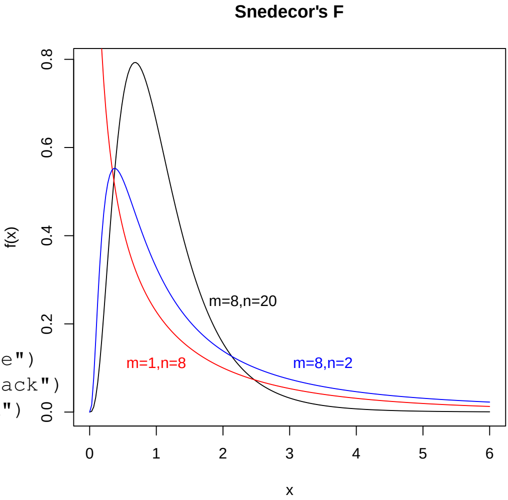
- If n not too small, shape close to scaled χ^2_m.
- If both m and n large, closely concentrated around 1.

What does the $\mathcal{F}_{m,n}$ look like? (I)

- If n not too small, shape close to scaled χ^2_m.
- If both m and n large, closely concentrated around 1.
- Right-skewed.

What does the $\mathcal{F}_{m,n}$ look like? (II)

```
> sned <- function(x) {
     df(x,m,n)
   }
> m <- 8 ; n <- 20
> curve(sned,
   from=0.0,to=6,n=200,
   ylab="f(x)",xlab="x",
   main="Snedecor's F")
> m <- 1 ; n <- 8
> curve(sned,from=0.0,col="red",
       to=6,n=200,add=TRUE)
> m <- 8 ; n <- 2
> curve(sned,from=0.0,col="blue",
       to=6,n=200,add=TRUE)
> text(3.5,0.11,"m=8,n=2",col="blue")
> text(2.3,0.25,"m=8,n=20",col="black")
> text(1.0,0.11,"m=1,n=8",col="red")
```



Student's t_n distribution

- Distribution of the ratio of independent $N(0,1)$ and $\sqrt{\chi^2_n/n}$ random variables:
$$t_n = \frac{N(0,1)}{\sqrt{\chi^2_n/n}}$$

Student's t_n distribution

- Distribution of the ratio of independent $N(0,1)$ and $\sqrt{\chi^2_n/n}$ random variables:
$$t_n = \frac{N(0,1)}{\sqrt{\chi^2_n/n}}$$
- Named after W. Gosset (1876-1937), who usually signed his work as "Student".

Student's t_n distribution

- Distribution of the ratio of independent $N(0,1)$ and $\sqrt{\chi^2_n/n}$ random variables:

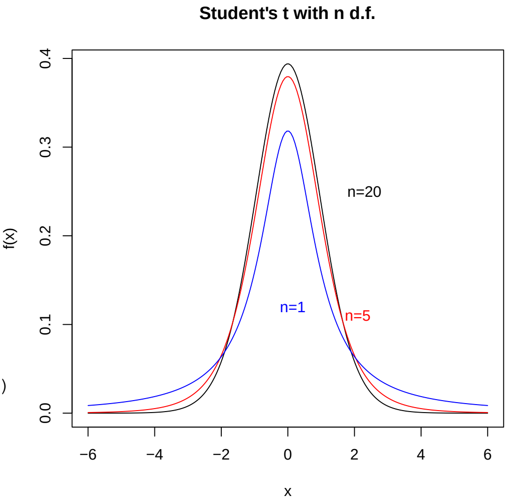
$$t_n = \frac{N(0,1)}{\sqrt{\chi^2_n/n}}$$

- Named after W. Gosset (1876-1937), who usually signed his work as "Student".
- Has density,

$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)}$$

What does Student's t_n look like?

```
> tx <- function(x) {
     dt(x,n)
   }
> n <- 20
> curve(tx,
   from=-6,to=6,n=200,
   ylab="f(x)",xlab="x",
   main="Student's t with n d.f.")
> n <- 5
> curve(tx,from=-6,col="red",
       to=6,n=200,add=TRUE)
> n <- 1
> curve(tx,from=-6,col="blue",
       to=6,n=200,add=TRUE)
> text(0.15,0.12,"n=1",col="blue")
> text(2.3,0.25,"n=20",col="black")
> text(2.1,0.11,"n=5",col="red")
```



Moments of the t_n distribution

- Not all moments exist for all n.

Moments of the t_n distribution

- Not all moments exist for all n.
- As an striking example, when $n = 1$,

$$\begin{aligned} f_X(x) &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)} \\ &= \frac{1}{\pi}\frac{1}{1+x^2} \end{aligned}$$

is the Cauchy distribution, and has no mean!

Moments of the t_n distribution

- Not all moments exist for all n.
- As an striking example, when $n = 1$,

$$\begin{aligned} f_X(x) &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{1}{2}(n+1)} \\ &= \frac{1}{\pi}\frac{1}{1+x^2} \end{aligned}$$

  is the Cauchy distribution, and has no mean!
- For greater n, higher order moments are non existent.

Reminder of some useful relationships

- $t_n^2 = \mathcal{F}_{1,n}$

Reminder of some useful relationships

- $t_n^2 = \mathcal{F}_{1,n}$
- t_n approaches a $N(0,1)$ as $n \to \infty$.

Reminder of some useful relationships

- $t_n^2 = \mathcal{F}_{1,n}$
- t_n approaches a $N(0,1)$ as $n \to \infty$.
- $\mathcal{F}_{m,n}$ approaches a χ_m^2/m as $n \to \infty$.

Reminder of some useful relationships

- $t_n^2 = \mathcal{F}_{1,n}$
- t_n approaches a $N(0,1)$ as $n \to \infty$.
- $\mathcal{F}_{m,n}$ approaches a χ_m^2/m as $n \to \infty$.
- If $X \sim \gamma(a, r)$ then $cX \sim \gamma(a/c, r)$.

Reminder of some useful relationships

- $t_n^2 = \mathcal{F}_{1,n}$
- t_n approaches a $N(0,1)$ as $n \to \infty$.
- $\mathcal{F}_{m,n}$ approaches a χ^2_m/m as $n \to \infty$.
- If $X \sim \gamma(a,r)$ then $cX \sim \gamma(a/c,r)$.
- In particular, sum of exponentials, $= \gamma(\lambda,1)$, can be turned into a χ^2_2 multiplying by a constant.

Reminder of some useful relationships

- $t_n^2 = \mathcal{F}_{1,n}$
- t_n approaches a $N(0,1)$ as $n \to \infty$.
- $\mathcal{F}_{m,n}$ approaches a χ_m^2/m as $n \to \infty$.
- If $X \sim \gamma(a, r)$ then $cX \sim \gamma(a/c, r)$.
- In particular, sum of exponentials, $= \gamma(\lambda, 1)$, can be turned into a χ_2^2 multiplying by a constant.
- $\mathcal{F}_{m,n}^{1-\alpha} = \frac{1}{\mathcal{F}_{n,m}^{\alpha}}$

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.
- Hydrogen, H, boils at $-253C$, helium, He, at $-269C$.

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.
- Hydrogen, H, boils at $-253C$, helium, He, at $-269C$.
- **Any** experimenter,

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.
- Hydrogen, H, boils at $-253C$, helium, He, at $-269C$.
- **Any** experimenter, **anywhere**,

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.
- Hydrogen, H, boils at $-253C$, helium, He, at $-269C$.
- **Any** experimenter, **anywhere**, **anytime**, can reproduce that result.

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.
- Hydrogen, *H*, boils at $-253C$, helium, *He*, at $-269C$.
- **Any** experimenter, **anywhere**, **anytime**, can reproduce that result.
- *He* will *always* boil at lower temperature than *H*.

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.
- Hydrogen, H, boils at $-253C$, helium, He, at $-269C$.
- **Any** experimenter, **anywhere**, **anytime**, can reproduce that result.
- He will *always* boil at lower temperature than H.
- An inequivocal statement can be made to this effect.

Scientific knowledge is objective, reproducible

- In many branches of science, reproducibility is easy.
- Hydrogen, H, boils at $-253C$, helium, He, at $-269C$.
- **Any** experimenter, **anywhere**, **anytime**, can reproduce that result.
- He will *always* boil at lower temperature than H.
- An inequivocal statement can be made to this effect.
- This is a *deterministic* phenomenon.

Not all phenomena are deterministic

- The same coin, in exactly the same circumstances, as far as it is feasible to check, sometimes will fall heads, sometimes will fall tails.

Not all phenomena are deterministic

- The same coin, in exactly the same circumstances, as far as it is feasible to check, sometimes will fall heads, sometimes will fall tails.
- We call phenomena such as this “random”.

Not all phenomena are deterministic

- The same coin, in exactly the same circumstances, as far as it is feasible to check, sometimes will fall heads, sometimes will fall tails.
- We call phenomena such as this “random”.
- Randomness is very tightly woven in the fabric of Nature.

Not all phenomena are deterministic

- The same coin, in exactly the same circumstances, as far as it is feasible to check, sometimes will fall heads, sometimes will fall tails.
- We call phenomena such as this “random”.
- Randomness is very tightly woven in the fabric of Nature.
- Even at the elementary particle level, of two exactly looking radioactive atoms, one may disintegrate in $(t, t + \Delta t)$ while the other may not.

Not all phenomena are deterministic

- The same coin, in exactly the same circumstances, as far as it is feasible to check, sometimes will fall heads, sometimes will fall tails.
- We call phenomena such as this “random”.
- Randomness is very tightly woven in the fabric of Nature.
- Even at the elementary particle level, of two exactly looking radioactive atoms, one may disintegrate in $(t, t + \Delta t)$ while the other may not.
- Individual outcomes of random events are impossible to predict, which seems to break all chances of reproducibility.

Can we make meaningful statements with random events?

- Two dice, *A* and *B*. *A* is “loaded” so that it will tend to give **6** more easily than *B*.

Can we make meaningful statements with random events?

- Two dice, *A* and *B*. *A* is “loaded” so that it will tend to give **6** more easily than *B*.
- Can we say that *A* produces more **6** results than *B*?

Can we make meaningful statements with random events?

- Two dice, *A* and *B*. *A* is “loaded” so that it will tend to give **6** more easily than *B*.
- Can we say that *A* produces more **6** results than *B*?
- Certainly, not in a reproducible way. In 10 throws, *B* may produce more **6**’s. Or in 20 throws. Or in 100 throws.

Can we make meaningful statements with random events?

- Two dice, *A* and *B*. *A* is “loaded” so that it will tend to give **6** more easily than *B*.
- Can we say that *A* produces more **6** results than *B*?
- Certainly, not in a reproducible way. In 10 throws, *B* may produce more **6**’s. Or in 20 throws. Or in 100 throws.
- However, we **imagine** that the random mechanism underlying *A* will *in the long run* produce more **6**’s than will the case with *B*. The relative frequency of **6** with *A* will be larger that with *B* in the long run.

Can we make meaningful statements with random events?

- Two dice, *A* and *B*. *A* is “loaded” so that it will tend to give **6** more easily than *B*.
- Can we say that *A* produces more **6** results than *B*?
- Certainly, not in a reproducible way. In 10 throws, *B* may produce more **6**’s. Or in 20 throws. Or in 100 throws.
- However, we **imagine** that the random mechanism underlying *A* will *in the long run* produce more **6**’s than will the case with *B*. The relative frequency of **6** with *A* will be larger that with *B* in the long run.
- Probabilities, the idealized limits of these long run frequencies, are **our model.**

Can we make meaningful statements with random events?

- Two dice, *A* and *B*. *A* is “loaded” so that it will tend to give **6** more easily than *B*.
- Can we say that *A* produces more **6** results than *B*?
- Certainly, not in a reproducible way. In 10 throws, *B* may produce more **6**’s. Or in 20 throws. Or in 100 throws.
- However, we **imagine** that the random mechanism underlying *A* will *in the long run* produce more **6**’s than will the case with *B*. The relative frequency of **6** with *A* will be larger that with *B* in the long run.
- Probabilities, the idealized limits of these long run frequencies, are **our model.**
- Therefore we state our suspicion as: “We believe $P_A(6) > P_B(6)$”, **in terms of a model.**

Two very important points:

1. All statements we make are phrased in terms of the parameters in our model (in this case, the probabilities $P_A(6)$ and $P_B(6)$ of getting a **6** with either die).

Two very important points:

1. All statements we make are phrased in terms of the parameters in our model (in this case, the probabilities $P_A(6)$ and $P_B(6)$ of getting a **6** with either die).
2. No way of checking for sure whether $P_A(6) > P_B(6)$. But, if that were true, sampling evidence would tend to show more **6**'s with die A than with B in the long run.

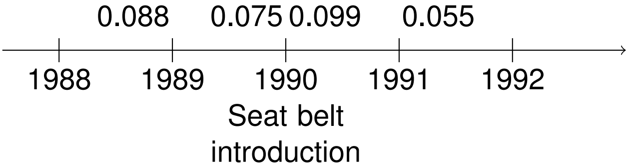
Two very important points:

1. All statements we make are phrased in terms of the parameters in our model (in this case, the probabilities $P_A(6)$ and $P_B(6)$ of getting a **6** with either die).
2. No way of checking for sure whether $P_A(6) > P_B(6)$. But, if that were true, sampling evidence would tend to show more **6**'s with die *A* than with *B* in the long run.

A statistician can never establish a fact for sure, only gather evidence which supports (or conflicts with) a hypothesis.

Looking only at data does not support meaningful answers

- Consider this (fictional) data on proportion of deaths in traffic accidents:



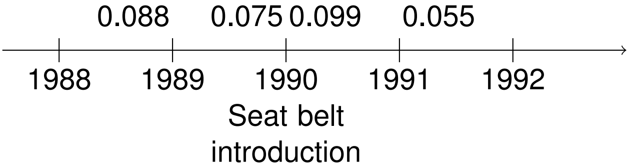
0.088 0.075 0.099 0.055

1988 1989 1990 1991 1992

Seat belt introduction

Looking only at data does not support meaningful answers

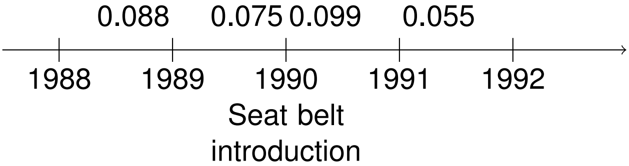
- Consider this (fictional) data on proportion of deaths in traffic accidents:



- Did the introduction of seat belts diminish deaths?

Looking only at data does not support meaningful answers

- Consider this (fictional) data on proportion of deaths in traffic accidents:



0.088 0.075 0.099 0.055

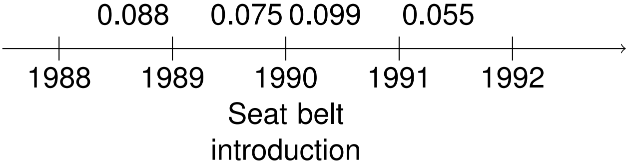
1988 1989 1990 1991 1992

Seat belt introduction

- Did the introduction of seat belts diminish deaths?
- $0.099 > 0.075$

Looking only at data does not support meaningful answers

- Consider this (fictional) data on proportion of deaths in traffic accidents:



0.088 0.075 0.099 0.055

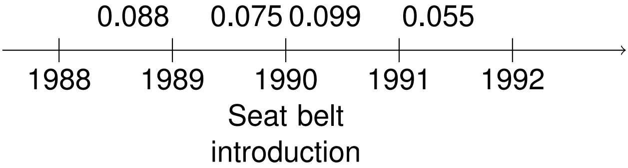
1988 1989 1990 1991 1992

Seat belt introduction

- Did the introduction of seat belts diminish deaths?
- $0.099 > 0.075$
- On the other hand, $0.055 < 0.088 \ldots$

Looking only at data does not support meaningful answers

- Consider this (fictional) data on proportion of deaths in traffic accidents:



- Did the introduction of seat belts diminish deaths?
- $0.099 > 0.075$
- On the other hand, $0.055 < 0.088\ldots$
- ...and $0.088 + 0.075 > 0.099 + 0.055$

We formulate problems in terms of models!

- Proportions of deaths keep changing; they are “fluid”.

We formulate problems in terms of models!

- Proportions of deaths keep changing; they are “fluid”.
- We need something “fix” to hold on.

We formulate problems in terms of models!

- Proportions of deaths keep changing; they are “fluid”.
- We need something “fix” to hold on.
- We **imagine** deaths before and after seat belt introduction are generated by (possibly different) random mechanisms. **That’s our model.**

We formulate problems in terms of models!

- Proportions of deaths keep changing; they are “fluid”.
- We need something “fix” to hold on.
- We **imagine** deaths before and after seat belt introduction are generated by (possibly different) random mechanisms. **That’s our model.**
- Proportions tend to stabilize around fixed, “solid” probabilities.

We formulate problems in terms of models!

- Proportions of deaths keep changing; they are “fluid”.
- We need something “fix” to hold on.
- We **imagine** deaths before and after seat belt introduction are generated by (possibly different) random mechanisms. **That’s our model.**
- Proportions tend to stabilize around fixed, “solid” probabilities.
- Simple model: $P[\mathrm{Death}] = p_0$ before, $P[\mathrm{Death}] = p_1$ after.

We formulate problems in terms of models!

- Proportions of deaths keep changing; they are “fluid”.
- We need something “fix” to hold on.
- We **imagine** deaths before and after seat belt introduction are generated by (possibly different) random mechanisms. **That’s our model.**
- Proportions tend to stabilize around fixed, “solid” probabilities.
- Simple model: $P[\mathrm{Death}] = p_0$ before, $P[\mathrm{Death}] = p_1$ after.
- *Now* we can ask ourselves: is $p_0 \neq p_1$? (Or $>$, or $<$ as the case may be.)

To summarize so far…

- Certain phenomena are “random”; in situations apparently identical, the outcome changes.

To summarize so far…

- Certain phenomena are “random”; in situations apparently identical, the outcome changes.
- However, in long series of repetitions relative frequencies seem to settle around fixed numbers.

To summarize so far…

- Certain phenomena are “random”; in situations apparently identical, the outcome changes.
- However, in long series of repetitions relative frequencies seem to settle around fixed numbers.
- Probabilities are a model for this.

To summarize so far. . .

- Certain phenomena are “random”; in situations apparently identical, the outcome changes.
- However, in long series of repetitions relative frequencies seem to settle around fixed numbers.
- Probabilities are a model for this.
- We phrase our questions and problems in terms of what is permanent: the probabilities or the **parameters of the model.**

The beauty of the whole thing

- Without a model, we would have to fill a tank to measure capacity.

The beauty of the whole thing

- Without a model, we would have to fill a tank to measure capacity.
- Geometry tells us $V = \frac{4}{3}\pi r^3$. Much easier!

The beauty of the whole thing

- Without a model, we would have to fill a tank to measure capacity.
- Geometry tells us $V = \frac{4}{3}\pi r^3$. Much easier!
- We do *not* say that the tank is an sphere.

The beauty of the whole thing

- Without a model, we would have to fill a tank to measure capacity.
- Geometry tells us $V = \frac{4}{3}\pi r^3$. Much easier!
- We do *not* say that the tank is an sphere.
- We *do* say that Euclidean geometry (and the formulae developed for spheres) work to a good approximation for “spherical” real objects.

The beauty of the whole thing

- Without a model, we would have to fill a tank to measure capacity.
- Geometry tells us $V = \frac{4}{3}\pi r^3$. Much easier!
- We do *not* say that the tank is an sphere.
- We *do* say that Euclidean geometry (and the formulae developed for spheres) work to a good approximation for "spherical" real objects.
- Statistical models are similar.

The beauty of the whole thing

- Without a model, we would have to fill a tank to measure capacity.
- Geometry tells us $V = \frac{4}{3}\pi r^3$. Much easier!
- We do *not* say that the tank is an sphere.
- We *do* say that Euclidean geometry (and the formulae developed for spheres) work to a good approximation for “spherical” real objects.
- Statistical models are similar.

Come to think of it, that purely intelectual constructions tell so much about the real world is a wonder!

Turning questions into statistical inference problems

- Our models will usually be distributions, some of whose parameters are unknown.

Turning questions into statistical inference problems

- Our models will usually be distributions, some of whose parameters are unknown.
- Our questions can usually be phrased in terms of values of those parameters.

Turning questions into statistical inference problems

- Our models will usually be distributions, some of whose parameters are unknown.
- Our questions can usually be phrased in terms of values of those parameters.
  - What is the average mortality for seat belt users? $\Leftrightarrow$ What is p_{Users}? *(estimation problem)*

Turning questions into statistical inference problems

- Our models will usually be distributions, some of whose parameters are unknown.
- Our questions can usually be phrased in terms of values of those parameters.
  - What is the average mortality for seat belt users?
    $\Leftrightarrow$ What is p_{Users}? *(estimation problem)*
  - Do seat belts reduce mortality in traffic accidents?
    $\Leftrightarrow$ Is $p_{Users} < p_{NonUsers}$? *(hypothesis test problem)*

Turning questions into statistical inference problems

- Our models will usually be distributions, some of whose parameters are unknown.
- Our questions can usually be phrased in terms of values of those parameters.
  - What is the average mortality for seat belt users?
    $\Leftrightarrow$ What is p_{Users}? *(estimation problem)*
  - Do seat belts reduce mortality in traffic accidents?
    $\Leftrightarrow$ Is $p_{Users} < p_{NonUsers}$? *(hypothesis test problem)*
- Other problems not quite fitting in either category (e.g., serialization)

Turning questions into statistical inference problems

- Our models will usually be distributions, some of whose parameters are unknown.
- Our questions can usually be phrased in terms of values of those parameters.
    - What is the average mortality for seat belt users?
      $\Leftrightarrow$ What is p_{Users}? *(estimation problem)*
    - Do seat belts reduce mortality in traffic accidents?
      $\Leftrightarrow$ Is $p_{Users} < p_{NonUsers}$? *(hypothesis test problem)*
- Other problems not quite fitting in either category (e.g., serialization)
- If model is “good”, answering questions about the model will enlighten us about the real world.

Is all this *that* new?

- No, it isn’t. All along we have introduced in problems these ideas without making them explicit.

Is all this *that* new?

- No, it isn’t. All along we have introduced in problems these ideas without making them explicit.
- If you think for a moment, many previous examples were phrased in a manner suggesting inferential problems.

Is all this *that* new?

- No, it isn’t. All along we have introduced in problems these ideas without making them explicit.
- If you think for a moment, many previous examples were phrased in a manner suggesting inferential problems.

Can you think of some instances?

Is all this *that* new?

- No, it isn't. All along we have introduced in problems these ideas without making them explicit.
- If you think for a moment, many previous examples were phrased in a manner suggesting inferential problems.

Problem regarding cancer incidence in a school was abnormally high.

Is all this *that* new?

- No, it isn't. All along we have introduced in problems these ideas without making them explicit.
- If you think for a moment, many previous examples were phrased in a manner suggesting inferential problems.

Problem regarding cancer incidence in a school was abnormally high.

Estimating the proportion of people who would vote for a candidate.

Is all this *that* new?

- No, it isn't. All along we have introduced in problems these ideas without making them explicit.
- If you think for a moment, many previous examples were phrased in a manner suggesting inferential problems.

Problem regarding cancer incidence in a school was abnormally high.

Estimating the proportion of people who would vote for a candidate.

Problem in which we were asked to check if mean service time in a car repair shop was $m = 1/\lambda = 65$ minutes.

To summarize. . .

- Parameters pertain to the *population* (= the model)

To summarize. . .

- Parameters pertain to the *population* (= the model)
- What we observe is the empirical evidence available: samples.

To summarize. . .

- Parameters pertain to the *population* (= the model)
- What we observe is the empirical evidence available: samples.
- A *sample* is a collection of elements generated by the population, usually through random sampling.

To summarize. . .

- Parameters pertain to the *population* (= the model)
- What we observe is the empirical evidence available: samples.
- A *sample* is a collection of elements generated by the population, usually through random sampling.
- From what we observe in the sample, we infer properties of the population, the model.

Point and interval estimation

- Sometimes we are content with a value “close” to the (unattainable) value of the true parameter. Then we have a problem of *point estimation*.

Point and interval estimation

- Sometimes we are content with a value “close” to the (unattainable) value of the true parameter. Then we have a problem of *point estimation*.
- Sometimes we want an interval that most of the time (with given *confidence*) will cover the true value of the parameter. This is an *interval estimation* problem.

Point and interval estimation

- Sometimes we are content with a value “close” to the (unattainable) value of the true parameter. Then we have a problem of *point estimation*.
- Sometimes we want an interval that most of the time (with given *confidence*) will cover the true value of the parameter. This is an *interval estimation* problem.
- Common sense will sometimes guide us in choosing an estimator. . .

Point and interval estimation

- Sometimes we are content with a value “close” to the (unattainable) value of the true parameter. Then we have a problem of *point estimation*.
- Sometimes we want an interval that most of the time (with given *confidence*) will cover the true value of the parameter. This is an *interval estimation* problem.
- Common sense will sometimes guide us in choosing an estimator. . .
- . . .but a more principled approach is desirable.

Samples and statistics

- A sample is a randomly chosen set from the population.

Samples and statistics

- A sample is a randomly chosen set from the population.
- Capital letters denote random values the members of the sample can yield: $\vec{X} = (X_1, X_2, \ldots, X_n)$.

Samples and statistics

- A sample is a randomly chosen set from the population.
- Capital letters denote random values the members of the sample can yield: $\vec{X} = (X_1, X_2, \ldots, X_n)$.
- Lower case letters, $\vec{x} = (x_1, x_2, \ldots, x_n)$, denote the actual, fixed values obtained in a concrete sample taken.

Samples and statistics

- A sample is a randomly chosen set from the population.
- Capital letters denote random values the members of the sample can yield: $\vec{X} = (X_1, X_2, \ldots, X_n)$.
- Lower case letters, $\vec{x} = (x_1, x_2, \ldots, x_n)$, denote the actual, fixed values obtained in a concrete sample taken.
- A *statistic* is a function of the sample: $S = S(\vec{X})$ or $s = s(\vec{x})$. Before the sample is taken, it is a random variable; after the sample is taken, it becomes a number (or vector of numbers)

Estimators and estimates

- An statistic designed to be “close” to the value of a parameter is an *estimator*.

Estimators and estimates

- An statistic designed to be “close” to the value of a parameter is an *estimator*.
- The value it takes is an *estimate.*

Estimators and estimates

- An statistic designed to be “close” to the value of a parameter is an *estimator*.
- The value it takes is an *estimate.*
- Example: $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$ is a (usually good) estimator of the mean of a distribution. *Given* a concrete sample $x_1, \ldots, x_n$, $\overline{x} = 5.8$ is an estimate.

Estimators and estimates

- An statistic designed to be “close” to the value of a parameter is an *estimator*.
- The value it takes is an *estimate.*
- Example: $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$ is a (usually good) estimator of the mean of a distribution. *Given* a concrete sample $x_1, \ldots, x_n$, $\overline{x} = 5.8$ is an estimate.
- With different samples, the same estimator will produce different estimates each time.

Methods for choosing point estimators

- What we choose as an estimator depends on our goals and *loss function* (= how much cost errors).

Methods for choosing point estimators

- What we choose as an estimator depends on our goals and *loss function* (= how much cost errors).
- For didactical reasons, we will look first at some recipes, then study their properties.

Methods for choosing point estimators

- What we choose as an estimator depends on our goals and *loss function* (= how much cost errors).
- For didactical reasons, we will look first at some recipes, then study their properties.
- Two important estimators:

Methods for choosing point estimators

- What we choose as an estimator depends on our goals and *loss function* (= how much cost errors).
- For didactical reasons, we will look first at some recipes, then study their properties.
- Two important estimators:
  - Method of moments.

Methods for choosing point estimators

- What we choose as an estimator depends on our goals and *loss function* (= how much cost errors).
- For didactical reasons, we will look first at some recipes, then study their properties.
- Two important estimators:
  - Method of moments.
  - Method of maximum likelihood.

Methods for choosing point estimators

- What we choose as an estimator depends on our goals and *loss function* (= how much cost errors).
- For didactical reasons, we will look first at some recipes, then study their properties.
- Two important estimators:
  - Method of moments.
  - Method of maximum likelihood.
- Least squares method is a particular case of the method of moments.

Method of moments: motivation (I)

- We think that a sample comes from a certain family of distributions.

Method of moments: motivation (I)

- We think that a sample comes from a certain family of distributions.
- We have to choose one member of that family (for instance, one particular Poisson from the family of all Poisson distributions)

Method of moments: motivation (I)

- We think that a sample comes from a certain family of distributions.
- We have to choose one member of that family (for instance, one particular Poisson from the family of all Poisson distributions)
- Want the one that is “closest” in some sense to observed data.

Method of moments: motivation (I)

- We think that a sample comes from a certain family of distributions.
- We have to choose one member of that family (for instance, one particular Poisson from the family of all Poisson distributions)
- Want the one that is “closest” in some sense to observed data.
- Makes sense to match theoretical moments to empirical moments. After all, moments determine the distribution.

Method of moments: motivation (I)

- We think that a sample comes from a certain family of distributions.
- We have to choose one member of that family (for instance, one particular Poisson from the family of all Poisson distributions)
- Want the one that is “closest” in some sense to observed data.
- Makes sense to match theoretical moments to empirical moments. After all, moments determine the distribution.

Method of moments: motivation (I)

- We think that a sample comes from a certain family of distributions.
- We have to choose one member of that family (for instance, one particular Poisson from the family of all Poisson distributions)
- Want the one that is “closest” in some sense to observed data.
- Makes sense to match theoretical moments to empirical moments. After all, moments determine the distribution.
- Usually lower order moments best (and simpler).

Method of moments (II)

- Equate moments of the distribution (usually function of parameters) to sample moments.

Method of moments (II)

- Equate moments of the distribution (usually function of parameters) to sample moments.
- Solve for the parameters.

Method of moments (II)

- Equate moments of the distribution (usually function of parameters) to sample moments.
- Solve for the parameters.
- Need as many equations as there are parameters.

Method of moments (II)

- Equate moments of the distribution (usually function of parameters) to sample moments.
- Solve for the parameters.
- Need as many equations as there are parameters.
- Example: $\mathcal{P}(\lambda)$, sample of n observations.

$$m = \lambda = \frac{X_1 + X_2 + \ldots + X_n}{n} = \overline{X}$$

Method of moments (II)

- Equate moments of the distribution (usually function of parameters) to sample moments.
- Solve for the parameters.
- Need as many equations as there are parameters.
- Example: $\mathcal{P}(\lambda)$, sample of n observations.

$$m = \lambda = \frac{X_1 + X_2 + \ldots + X_n}{n} = \overline{X}$$

Method of moments (II)

- Equate moments of the distribution (usually function of parameters) to sample moments.
- Solve for the parameters.
- Need as many equations as there are parameters.
- Example: $\mathcal{P}(\lambda)$, sample of n observations.

$$m = \lambda = \frac{X_1 + X_2 + \ldots + X_n}{n} = \overline{X}$$

- Could also use,

$$\lambda + \lambda^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2$$

  Usually lower order moments best (and simpler).

Method of moments (III)

- Example: estimate m and σ^2 of $N(m, \sigma^2)$.

Method of moments (III)

- Example: estimate m and σ^2 of $N(m, \sigma^2)$.
- Now we need two equations:

$$\begin{aligned} m &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \sigma^2 + m^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

from which

$$\begin{aligned} \hat{m} &= \overline{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{aligned}$$

Method of moments (IV)

- Example: estimate θ in a $U(0,\theta)$.

Method of moments (IV)

- Example: estimate θ in a $U(0, \theta)$.
- The mean is $m = \theta/2$. Therefore,

$$\begin{aligned} \frac{\theta}{2} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \hat{\theta} &= 2\overline{X} \end{aligned}$$

Method of moments (IV)

- Example: estimate θ in a $U(0, \theta)$.
- The mean is $m = \theta/2$. Therefore,

$$\begin{aligned} \frac{\theta}{2} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \hat{\theta} &= 2\overline{X} \end{aligned}$$

- Not a particularly good estimator, as we will see.

Method of moments (V)

- Example: estimate λ in a $\exp(\lambda)$.

Method of moments (V)

- Example: estimate λ in a $\exp(\lambda)$.
- The mean is $m = 1/\lambda$. Therefore,

$$\begin{aligned} \frac{1}{\lambda} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \hat{\lambda} &= \frac{1}{\overline{X}} \end{aligned}$$

Method of moments (VI)

- Example: estimate a and r in a $\gamma(a, r)$.

Method of moments (VI)

- Example: estimate a and r in a $\gamma(a, r)$.
- Remember that $m = r/a$ and $\sigma^2 = r/a^2$. Therefore,

$$\begin{aligned} \frac{r}{a} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \frac{r}{a^2} + \frac{r^2}{a^2} &= \frac{1}{n} \sum_{i=1}^{n} X_i^2 \end{aligned}$$

Method of moments (VI)

- Example: estimate a and r in a $\gamma(a, r)$.
- Remember that $m = r/a$ and $\sigma^2 = r/a^2$. Therefore,

$$\begin{aligned} \frac{r}{a} &= \frac{X_1 + X_2 + \ldots + X_n}{n} \\ \frac{r}{a^2} + \frac{r^2}{a^2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{aligned}$$

- We can solve for a and r to obtain:

$$\begin{aligned} \hat{a} &= r/\overline{X} \\ \hat{r} &= \frac{\overline{X}^2}{n^{-1}\sum_{i=1}^{n} X_i^2 - \overline{X}^2} \end{aligned}$$

Method of moments (VII)

- Sometimes, method just don’t work!

Method of moments (VII)

- Sometimes, method just don't work!
- For instance, if we try to estimate c from a density

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+(x-c)^2}$$

we will get nowhere.

Method of moments (VII)

- Sometimes, method just don't work!
- For instance, if we try to estimate c from a density

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+(x-c)^2}$$

we will get nowhere.

- Some distributions have no moments, so nothing to match.

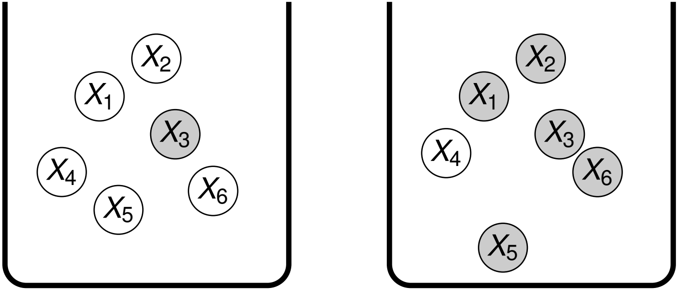
Method of moments (VII)

- Sometimes, method just don't work!
- For instance, if we try to estimate *c* from a density

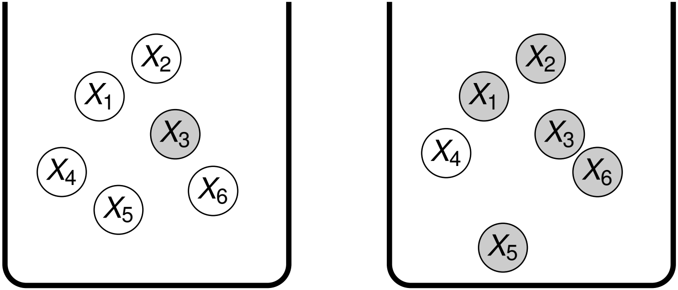
$$f_X(x) = \frac{1}{\pi} \frac{1}{1 + (x - c)^2}$$

  we will get nowhere.
- Some distributions have no moments, so nothing to match.
- In case at hand, we could use a *censored* (or *trimmed* mean), like the median.

Method of maximum likelihood (I)

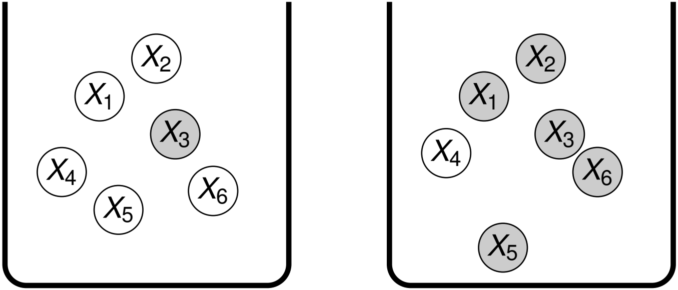


Method of maximum likelihood (I)



- We are allowed to sample one of the two urns, but we are not told which one it is. We pick one ball which happens to be grey

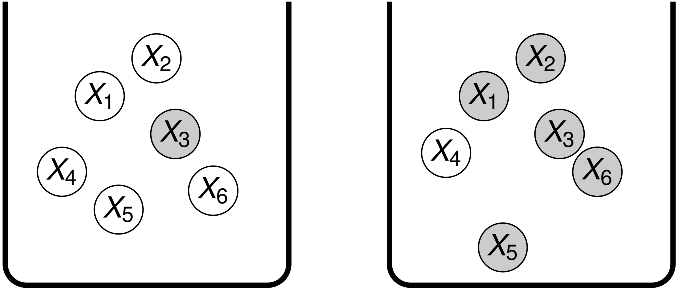
Method of maximum likelihood (I)



- We are allowed to sample one of the two urns, but we are not told which one it is. We pick one ball which happens to be grey

What would be your guess?

Method of maximum likelihood (I)



- We are allowed to sample one of the two urns, but we are not told which one it is. We pick one ball which happens to be grey

Right urn, as it can generate grey balls more easily.

Method of maximum likelihood (II)

- Logic underlying previous choice is maximum likelihood logic.

  When confronted to two or more states of nature which may have produced a given evidence, we choose the one(s) with optimal capability to generate such evidence.

Method of maximum likelihood (II)

- Logic underlying previous choice is maximum likelihood logic.

  When confronted to two or more states of nature which may have produced a given evidence, we choose the one(s) with optimal capability to generate such evidence.

- Both urns could generate a grey ball, but the second one does so much more easily.

Method of maximum likelihood (II)

- Logic underlying previous choice is maximum likelihood logic.

  When confronted to two or more states of nature which may have produced a given evidence, we choose the one(s) with optimal capability to generate such evidence.

- Both urns could generate a grey ball, but the second one does so much more easily.
- Why assume that something “strange” has happened if we can see the evidence as the outcome of something “common”?

Method of maximum likelihood (III)

- If joint density of a given sample is $f(\vec{x};\theta)$, $\theta \in \Theta$, we call *likelihood function* $f(\vec{x};\theta)$ **seen as a funcion of $\boldsymbol{\theta}$** for given $\vec{x}$.

Method of maximum likelihood (III)

- If joint density of a given sample is $f(\vec{x};\theta)$, $\theta \in \Theta$, we call *likelihood function* $f(\vec{x};\theta)$ **seen as a funcion of $\boldsymbol{\theta}$** for given $\vec{x}$.
- To maximize the likelihood is tantamount to choosing the $\boldsymbol{\theta}$ which gives maximum density to the observed sample.

Method of maximum likelihood (III)

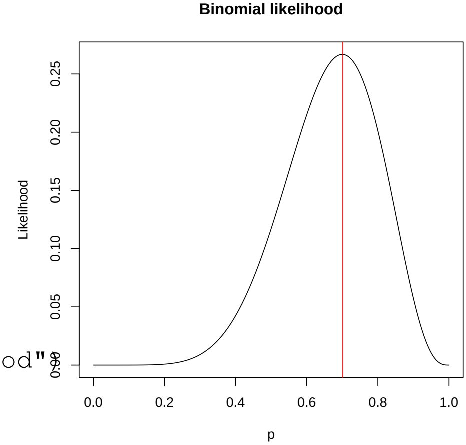
- If joint density of a given sample is $f(\vec{x};\theta)$, $\theta \in \Theta$, we call *likelihood function* $f(\vec{x};\theta)$ **seen as a funcion of $\boldsymbol{\theta}$** for given $\vec{x}$.
- To maximize the likelihood is tantamount to choosing the $\boldsymbol{\theta}$ which gives maximum density to the observed sample.
- Maximizing θ is *maximum likelihood estimate*, $\hat{\theta}_{MLE}$.

Method of maximum likelihood (III)

- If joint density of a given sample is $f(\vec{x};\theta)$, $\theta \in \Theta$, we call *likelihood function* $f(\vec{x};\theta)$ **seen as a funcion of $\boldsymbol{\theta}$** for given $\vec{x}$.
- To maximize the likelihood is tantamount to choosing the $\boldsymbol{\theta}$ which gives maximum density to the observed sample.
- Maximizing θ is *maximum likelihood estimate*, $\hat{\theta}_{MLE}$.
- $f(\vec{x};\theta)$ and $\log f(\vec{x};\theta)$ both achieve their maximum for the same value of θ. Usually easier to maximize the second.

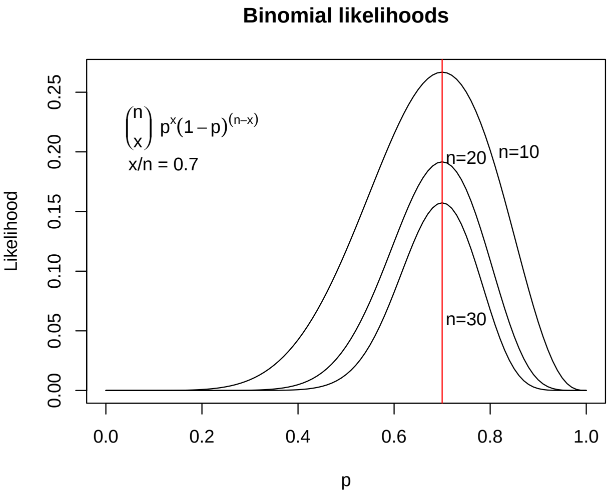
Likelihood example: binomial distribution (I)

```
> n <- 10 ; x <- 7
> binom <- function(p) {
  l <- choose(n,x) * p^x *
       (1-p)^(n-x)
  return(l)
 }
> curve(binom,from=0.00,
   to=1,n=200,
   ylab="Likelihood",
   xlab="p",
   main="Binomial likelihood")
> abline(v=x/n,col="red")
```



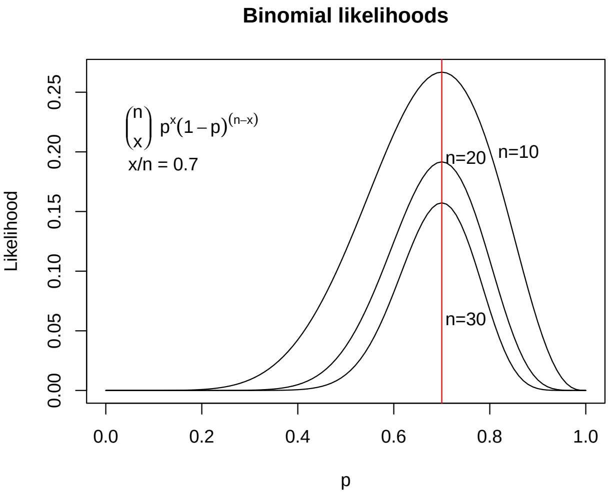
What would happen with different values of *x* and *n*?

Likelihood example: binomial (II)



Are the likelihood functions like density functions?

Likelihood example: binomial (II)



Clearly not; areas below change, not always 1.

Example: MLE of p with $x_1, x_2, \ldots, x_n$ i.i.d. $b(p)$

$$f(\vec{x}; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Example: MLE of p with $x_1, x_2, \ldots, x_n$ i.i.d. $b(p)$

$$
\begin{aligned}
f(\vec{x}; p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
\log f(\vec{x}; p) &= \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p)
\end{aligned}
$$

Example: MLE of p with $x_1, x_2, \ldots, x_n$ i.i.d. $b(p)$

$$\begin{aligned}
f(\vec{x}; p) &= \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i} \\
\log f(\vec{x}; p) &= \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i \right) \log(1-p) \\
\frac{\partial \log f(\vec{x}; p)}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0
\end{aligned}$$

Example: MLE of p with $x_1, x_2, \ldots, x_n$ i.i.d. $b(p)$

$$\begin{aligned}
f(\vec{x};p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
\log f(\vec{x};p) &= \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-p) \\
\frac{\partial \log f(\vec{x};p)}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \\
\hat{p}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

Example: MLE of p with $x_1, x_2, \ldots, x_n$ i.i.d. $b(p)$

$$\begin{aligned}
f(\vec{x}; p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
\log f(\vec{x}; p) &= \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p) \\
\frac{\partial \log f(\vec{x}; p)}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \\
\hat{p}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Example: MLE of p with $x_1, x_2, \ldots, x_n$ i.i.d. $b(p)$

$$\begin{aligned}
f(\vec{x}; p) &= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\
\log f(\vec{x}; p) &= \sum_{i=1}^{n} x_i \log(p) + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-p) \\
\frac{\partial \log f(\vec{x}; p)}{\partial p} &= \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0 \\
\hat{p}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

Only $\sum_{i=1}^{n} x_i$ is necessary to compute the MLE.

Example: MLE of λ with $x_1, x_2, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$

$$f(\vec{x};\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

Example: MLE of λ with $x_1, x_2, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$

$$\begin{aligned} f(\vec{x};\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\ \log f(\vec{x};\lambda) &= -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!) \end{aligned}$$

Example: MLE of λ with $x_1, x_2, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$

$$\begin{aligned}
f(\vec{x};\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
\log f(\vec{x};\lambda) &= -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!) \\
\frac{\partial \log f(\vec{x};\lambda)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0
\end{aligned}$$

Example: MLE of λ with $x_1, x_2, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$

$$
\begin{aligned}
f(\vec{x}; \lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
\log f(\vec{x}; \lambda) &= -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!) \\
\frac{\partial \log f(\vec{x}; \lambda)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0 \\
\hat{\lambda}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}
\end{aligned}
$$

Example: MLE of λ with $x_1, x_2, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$

$$
\begin{aligned}
f(\vec{x};\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
\log f(\vec{x};\lambda) &= -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!) \\
\frac{\partial \log f(\vec{x};\lambda)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0 \\
\hat{\lambda}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}
\end{aligned}
$$

Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Example: MLE of λ with $x_1, x_2, \ldots, x_n$ i.i.d. $\mathcal{P}(\lambda)$

$$
\begin{aligned}
f(\vec{x};\lambda) &= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
\log f(\vec{x};\lambda) &= -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!) \\
\frac{\partial \log f(\vec{x};\lambda)}{\partial \lambda} &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0 \\
\hat{\lambda}_{MLE} &= \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}
\end{aligned}
$$

Only $\sum_{i=1}^{n} x_i$ is necessary to compute the MLE.

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$f(\vec{x}; m, \sigma^2) \quad = \quad \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - m)^2/2\sigma^2}$$

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$
\begin{aligned}
f(\vec{x}; m, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - m)^2/2\sigma^2} \\
\log f(\vec{x}; m, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2\sigma^2}
\end{aligned}
$$

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$\begin{aligned}
f(\vec{x}; m, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - m)^2/2\sigma^2} \\
\log f(\vec{x}; m, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2\sigma^2} \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial m} &= \frac{\sum_{i=1}^{n}(x_i - m)}{\sigma^2} = 0
\end{aligned}$$

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$\begin{aligned}
f(\vec{x}; m, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-m)^2/2\sigma^2} \\
\log f(\vec{x}; m, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2\sigma^2} \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial m} &= \frac{\sum_{i=1}^{n}(x_i - m)}{\sigma^2} = 0 \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - m)^2}{2\sigma^4} = 0
\end{aligned}$$

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$
\begin{aligned}
f(\vec{x}; m, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-m)^2/2\sigma^2} \\
\log f(\vec{x}; m, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2\sigma^2} \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial m} &= \frac{\sum_{i=1}^{n}(x_i - m)}{\sigma^2} = 0 \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - m)^2}{2\sigma^4} = 0
\end{aligned}
$$

whence

$$
\hat{m}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \qquad \hat{\sigma^2}_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}
$$

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$
\begin{aligned}
f(\vec{x}; m, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i - m)^2/2\sigma^2} \\
\log f(\vec{x}; m, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2\sigma^2} \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial m} &= \frac{\sum_{i=1}^{n}(x_i - m)}{\sigma^2} = 0 \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - m)^2}{2\sigma^4} = 0
\end{aligned}
$$

whence

$$
\hat{m}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \qquad \hat{\sigma^2}_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}
$$

Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Example: MLE of m, σ^2 with $x_1, \ldots, x_n$ i.i.d. $N(m, \sigma^2)$

$$\begin{aligned}
f(\vec{x}; m, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-m)^2/2\sigma^2} \\
\log f(\vec{x}; m, \sigma^2) &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \frac{(x_i - m)^2}{2\sigma^2} \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial m} &= \frac{\sum_{i=1}^{n}(x_i - m)}{\sigma^2} = 0 \\
\frac{\partial \log f(\vec{x}; m, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - m)^2}{2\sigma^4} = 0
\end{aligned}$$

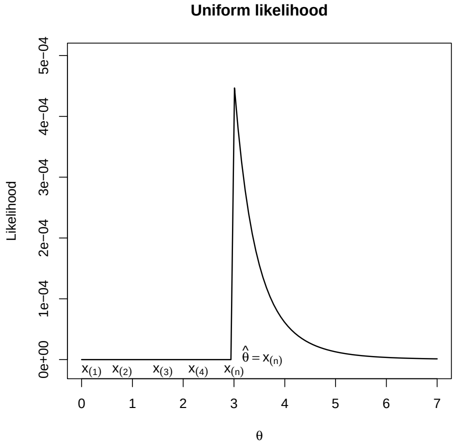
whence

$$\hat{m}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x} \qquad \hat{\sigma^2}_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

Only $\sum_{i=1}^{n} x_i$ and $\sum_{i=1}^{n}(x_i - \overline{x})^2$ necessary.

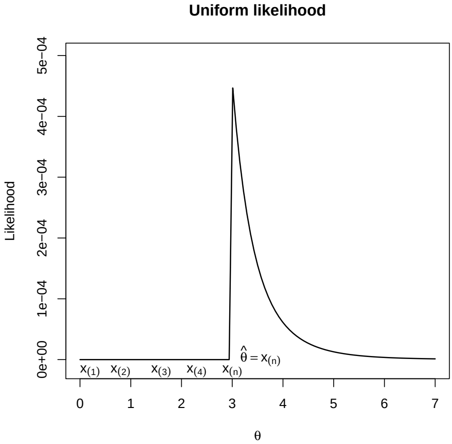
Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

- Likelihood function not as usual.



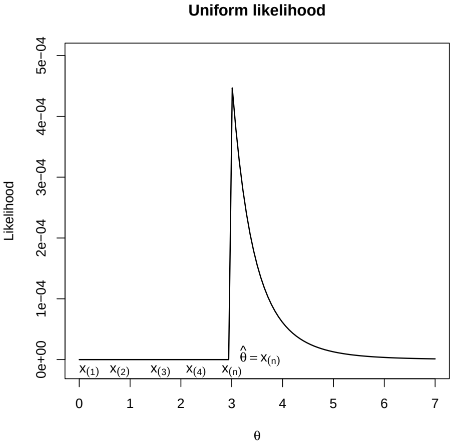
Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

- Likelihood function not as usual.
- Not differentiable.



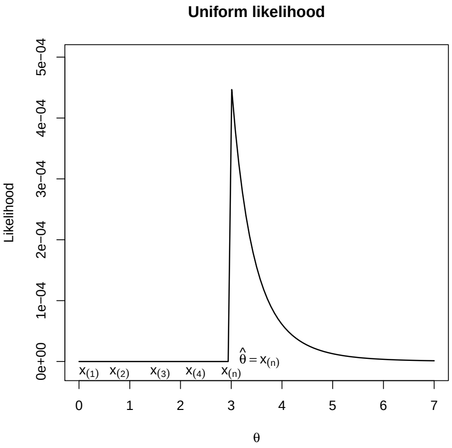
Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

- Likelihood function not as usual.
- Not differentiable.
- Pick maximum by inspection.



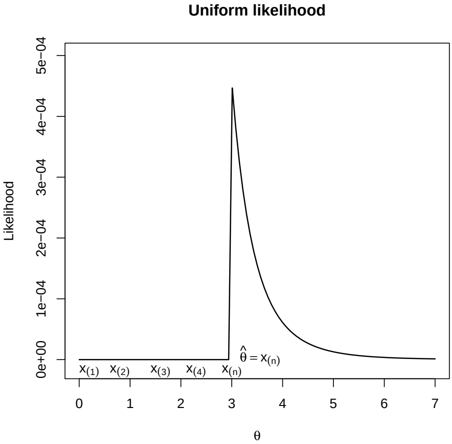
Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

- Likelihood function not as usual.
- Not differentiable.
- Pick maximum by inspection.
- $x_{(1)}, \ldots, x_{(n)}$ called “order statistics”.



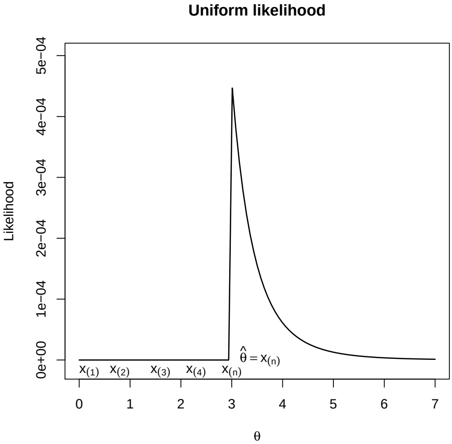
Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

- Likelihood function not as usual.
- Not differentiable.
- Pick maximum by inspection.
- $x_{(1)}, \ldots, x_{(n)}$ called "order statistics".



Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

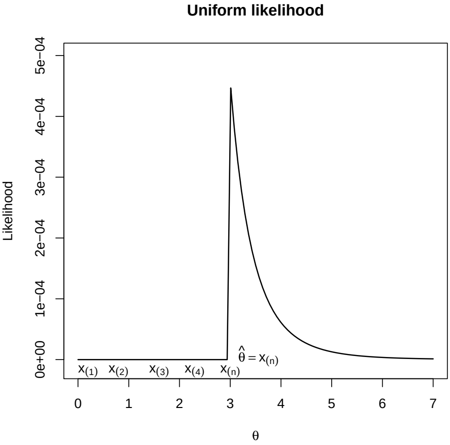
- Likelihood function not as usual.
- Not differentiable.
- Pick maximum by inspection.
- $x_{(1)}, \ldots, x_{(n)}$ called "order statistics".



Do we need to know all $x_1, \ldots, x_n$ in order to compute the MLE?

Example: MLE of θ with $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$

- Likelihood function not as usual.
- Not differentiable.
- Pick maximum by inspection.
- $x_{(1)}, \ldots, x_{(n)}$ called “order statistics”.



Only one ($x_{(n)}$, the largest) is necessary!

Example: MLE of a in a $\gamma(a, r)$

$$f(\vec{x}; a, r) = \prod_{i=1}^{n} \frac{a^r}{\Gamma(r)} x_i^{r-1} e^{-x_i/a} = \frac{a^{rn}}{(\Gamma(r))^n} \left(\prod_{i=1}^{n} x_i \right)^{r-1} e^{-\sum_{i=1}^{n} x_i}$$

Example: MLE of a in a $\gamma(a, r)$

$$\begin{aligned} f(\vec{x}; a, r) &= \prod_{i=1}^{n} \frac{a^{r}}{\Gamma(r)} x_i^{r-1} e^{-x_i/a} = \frac{a^{rn}}{(\Gamma(r))^{n}} \left(\prod_{i=1}^{n} x_i \right)^{r-1} e^{-\sum_{i=1}^{n} x_i} \\ \log f(\vec{x}; a, r) &= rn \log(a) - n \log \Gamma(r) + (r-1) \sum_{i=1}^{n} \log(x_i) - \frac{\sum_{i=1}^{n} x}{a} \end{aligned}$$

Example: MLE of a in a $\gamma(a, r)$

$$
\begin{aligned}
f(\vec{x}; a, r) &= \prod_{i=1}^{n} \frac{a^r}{\Gamma(r)} x_i^{r-1} e^{-x_i/a} = \frac{a^{rn}}{(\Gamma(r))^n} \left(\prod_{i=1}^{n} x_i \right)^{r-1} e^{-\sum_{i=1}^{n} x_i} \\
\log f(\vec{x}; a, r) &= rn \log(a) - n \log \Gamma(r) + (r-1) \sum_{i=1}^{n} \log(x_i) - \frac{\sum_{i=1}^{n} x}{a} \\
\frac{\partial \log f(\vec{x}; a, r)}{\partial a} &= \frac{rn}{a} - \frac{\sum_{i=1}^{n} x_i}{a^2} = 0
\end{aligned}
$$

Example: MLE of a in a $\gamma(a, r)$

$$\begin{aligned}
f(\vec{x}; a, r) &= \prod_{i=1}^{n} \frac{a^r}{\Gamma(r)} x_i^{r-1} e^{-x_i/a} = \frac{a^{rn}}{(\Gamma(r))^n} \left(\prod_{i=1}^{n} x_i \right)^{r-1} e^{-\sum_{i=1}^{n} x_i/} \\
\log f(\vec{x}; a, r) &= rn \log(a) - n \log \Gamma(r) + (r-1) \sum_{i=1}^{n} \log(x_i) - \frac{\sum_{i=1}^{n} x}{a} \\
\frac{\partial \log f(\vec{x}; a, r)}{\partial a} &= \frac{rn}{a} - \frac{\sum_{i=1}^{n} x_i}{a^2} = 0 \\
\frac{\partial \log f(\vec{x}; a, r)}{\partial r} &= n \log(a) - n \frac{\partial \log \Gamma(r)}{\partial r} + \sum_{i=1}^{n} \log(x_i) = 0
\end{aligned}$$

Example: MLE of a in a $\gamma(a, r)$

$$
\begin{aligned}
f(\vec{x}; a, r) &= \prod_{i=1}^{n} \frac{a^r}{\Gamma(r)} x_i^{r-1} e^{-x_i/a} = \frac{a^{rn}}{(\Gamma(r))^n} \left(\prod_{i=1}^{n} x_i \right)^{r-1} e^{-\sum_{i=1}^{n} x_i} \\
\log f(\vec{x}; a, r) &= rn \log(a) - n \log \Gamma(r) + (r-1) \sum_{i=1}^{n} \log(x_i) - \frac{\sum_{i=1}^{n} x}{a} \\
\frac{\partial \log f(\vec{x}; a, r)}{\partial a} &= \frac{rn}{a} - \frac{\sum_{i=1}^{n} x_i}{a^2} = 0 \\
\frac{\partial \log f(\vec{x}; a, r)}{\partial r} &= n \log(a) - n \frac{\partial \log \Gamma(r)}{\partial r} + \sum_{i=1}^{n} \log(x_i) = 0
\end{aligned}
$$

Given r, $\hat{a}_{MLE} = \frac{\sum_{i=1}^{n} X_i}{rn} = \frac{\overline{X}}{r}$, different from moment estimator.

Example: MLE of a in a $\gamma(a, r)$

$$
\begin{aligned}
f(\vec{x}; a, r) &= \prod_{i=1}^{n} \frac{a^r}{\Gamma(r)} x_i^{r-1} e^{-x_i/a} = \frac{a^{rn}}{(\Gamma(r))^n} \left(\prod_{i=1}^{n} x_i \right)^{r-1} e^{-\sum_{i=1}^{n} x_i} \\
\log f(\vec{x}; a, r) &= rn \log(a) - n \log\Gamma(r) + (r-1) \sum_{i=1}^{n} \log(x_i) - \frac{\sum_{i=1}^{n} x}{a} \\
\frac{\partial \log f(\vec{x}; a, r)}{\partial a} &= \frac{rn}{a} - \frac{\sum_{i=1}^{n} x_i}{a^2} = 0 \\
\frac{\partial \log f(\vec{x}; a, r)}{\partial r} &= n \log(a) - n \frac{\partial \log\Gamma(r)}{\partial r} + \sum_{i=1}^{n} \log(x_i) = 0
\end{aligned}
$$

Given r, $\hat{a}_{MLE} = \frac{\sum_{i=1}^{n} X_i}{rn} = \frac{\overline{X}}{r}$, different from moment estimator.
($\hat{r}_{MLE}$ is harder, as it involves the digamma function
$\partial \log\Gamma(r)/\partial r$.)

MLE and parameter transformations

- We may have different choices of parameters. For instance, in a binomial of given size n, we may choose as the parameter p, but also the *odds* $\theta = \frac{p}{1-p} = \theta(p)$.

MLE and parameter transformations

- We may have different choices of parameters. For instance, in a binomial of given size n, we may choose as the parameter p, but also the *odds* $\theta = \frac{p}{1-p} = \theta(p)$.
- **Theorem.** If the function relating θ and p is 1-1 (and has therefore has inverse), $\hat{\theta}_{MLE} = \theta(\hat{p}_{MLE})$.

MLE and parameter transformations

- We may have different choices of parameters. For instance, in a binomial of given size n, we may choose as the parameter p, but also the *odds* $\theta = \frac{p}{1-p} = \theta(p)$.
- **Theorem.** If the function relating θ and p is 1-1 (and has therefore has inverse), $\hat{\theta}_{MLE} = \theta(\hat{p}_{MLE})$.
- Indeed, if we denote by $\ell(p; \vec{x})$ the likelihood and replace p by $\theta^{-1}(\theta)$, clearly the maximum of $\ell(\theta^{-1}(\theta); \vec{x})$ will still be attained for a value of θ making $\theta^{-1}(\theta) = \hat{p}_{MLE}$, and therefore $\hat{\theta}_{MLE} = \theta(\hat{p}_{MLE})$.

MLE and parameter transformations

- We may have different choices of parameters. For instance, in a binomial of given size n, we may choose as the parameter p, but also the *odds* $\theta = \frac{p}{1-p} = \theta(p)$.
- **Theorem.** If the function relating θ and p is 1-1 (and has therefore has inverse), $\hat{\theta}_{MLE} = \theta(\hat{p}_{MLE})$.
- Indeed, if we denote by $\ell(p; \vec{x})$ the likelihood and replace p by $\theta^{-1}(\theta)$, clearly the maximum of $\ell(\theta^{-1}(\theta); \vec{x})$ will still be attained for a value of θ making $\theta^{-1}(\theta) = \hat{p}_{MLE}$, and therefore $\hat{\theta}_{MLE} = \theta(\hat{p}_{MLE})$.
- Nice property. In the binomial example, $\hat{p}_{MLE} = \overline{X}/n$, and the MLE of the odds is just: $\hat{\theta}_{MLE} = \frac{\hat{p}_{MLE}}{1-\hat{p}_{MLE}}$.

Can we always use the MLE estimator?

- In principle, yes. In practice, it might be too complex.

Can we always use the MLE estimator?

- In principle, yes. In practice, it might be too complex.
- As an striking example, consider again the Cauchy distribution with location θ for which the moment estimator failed.

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

Can we always use the MLE estimator?

- In principle, yes. In practice, it might be too complex.
- As an striking example, consider again the Cauchy distribution with location θ for which the moment estimator failed.

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

- If we try to write,

$$\ell(\theta, \vec{x}) = \prod_{i=1}^{n}\left(\frac{1}{\pi}\frac{1}{1+(x_i-\theta)^2}\right)$$

even for a fairly small *n* we will get a terribly complex expression.

Can we always use the MLE estimator?

- In principle, yes. In practice, it might be too complex.
- As an striking example, consider again the Cauchy distribution with location θ for which the moment estimator failed.

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

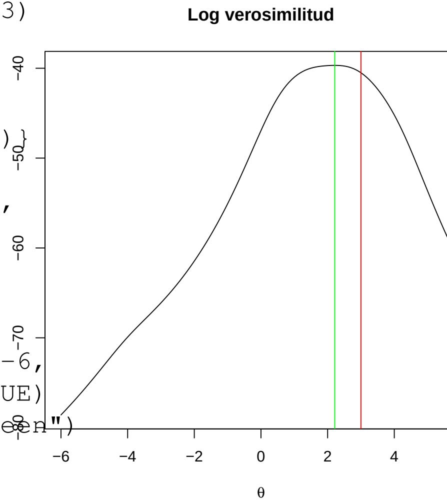
- If we try to write,

$$\ell(\theta, \vec{x}) = \prod_{i=1}^{n}\left(\frac{1}{\pi}\frac{1}{1+(x_i-\theta)^2}\right)$$

  even for a fairly small *n* we will get a terribly complex expression.
- No hope to maximize that analitycally.

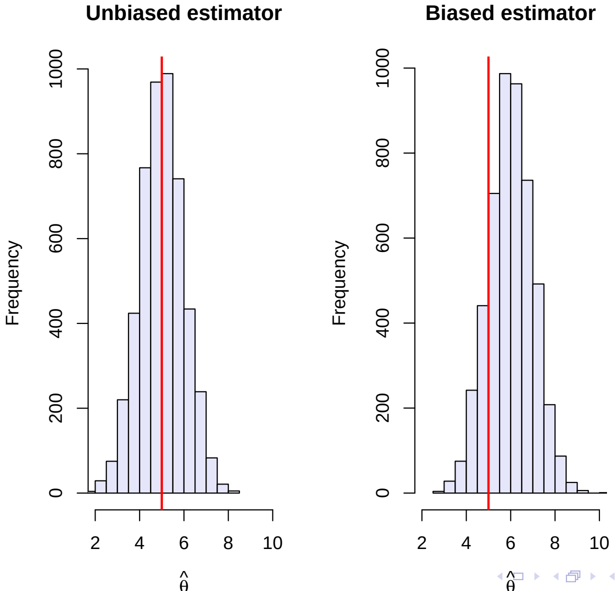
Numerical optimization

```
> x  <- rcauchy(n=15, location=3)
> thetas <- seq(from=-6,to=6,
                length.out=150)
> loglik <- function(theta) {
   sum(-log(pi*(1+(x-theta)^2)))}
> plot(thetas,
      sapply(thetas,FUN=loglik),
      type="l",ylab="",
      main="Log verosimilitud",
      xlab=expression(theta))
> abline(v=3,col="red")
> MLE <- optimize(loglik,lower=-6,
             upper=6, maximum=TRUE)
> abline(v=MLE$maximum, col="green")
```



Unbiasedness (I)

- $\hat{\theta}$ unbiased for θ means that $E[\hat{\theta}] = \theta$.

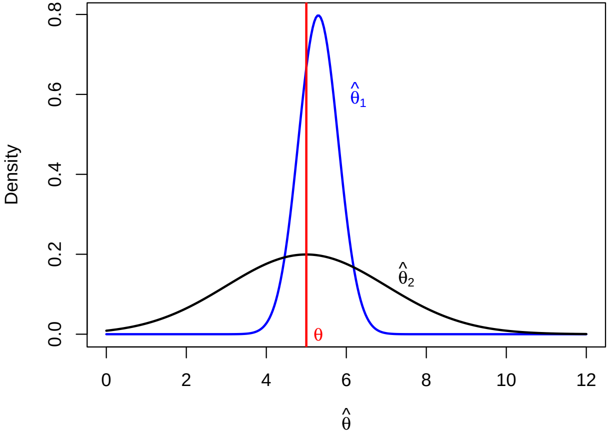


Unbiasedness (II)

- In principle a desirable property…

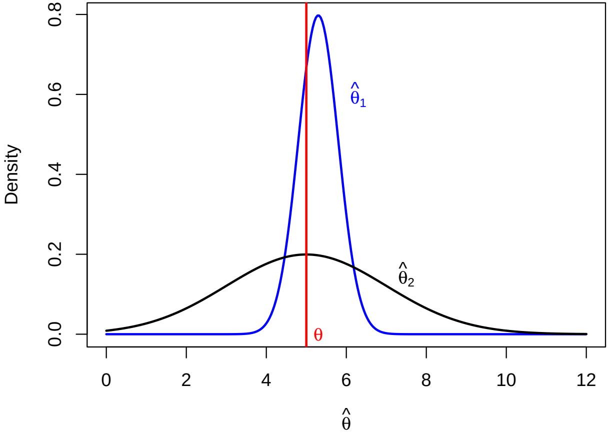
Unbiasedness (II)

- In principle a desirable property. . .
- . . .but sometimes we may prefer a biased estimator.



Unbiasedness (II)

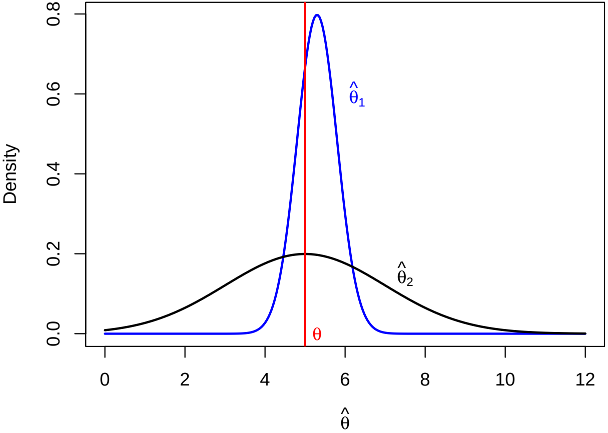
- In principle a desirable property. . .
- . . .but sometimes we may prefer a biased estimator.



Which one would you prefer?

Unbiasedness (II)

- In principle a desirable property. . .
- . . .but sometimes we may prefer a biased estimator.



If squared error loss, $E|\hat{\theta} - \theta|^2$, we might prefer $\hat{\theta}_1$, even if

Unbiasedness (III)

- In a $\mathcal{P}(\lambda)$, $\hat{\lambda} = \overline{X}$ is unbiased.

Unbiasedness (III)

- In a $\mathcal{P}(\lambda)$, $\hat{\lambda} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{m} = \overline{X}$ is unbiased.

Unbiasedness (III)

- In a $\mathcal{P}(\lambda)$, $\hat{\lambda} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{m} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$ is biased.

Unbiasedness (III)

- In a $\mathcal{P}(\lambda)$, $\hat{\lambda} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{m} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$ is biased.
- $\hat{\sigma}^2_{*} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ is unbiased; $\hat{\sigma}^2_{MLE}$ is *asymptotically unbiased*.

Unbiasedness (III)

- In a $\mathcal{P}(\lambda)$, $\hat{\lambda} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{m} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is biased.
- $\hat{\sigma}^2_{*} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is unbiased; $\hat{\sigma}^2_{MLE}$ is *asymptotically unbiased*.
- In a $\exp(\lambda)$, $\hat{\lambda} = \frac{1}{\overline{X}}$ is biased;

$$E[\hat{\lambda}] = E\left[\frac{1}{\overline{X}}\right] \neq \frac{1}{E[\overline{X}]}$$

Unbiasedness (III)

- In a $\mathcal{P}(\lambda)$, $\hat{\lambda} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{m} = \overline{X}$ is unbiased.
- In a $N(m, \sigma^2)$, $\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is biased.
- $\hat{\sigma}^2_{*} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is unbiased; $\hat{\sigma}^2_{MLE}$ is *asymptotically unbiased*.
- In a $\exp(\lambda)$, $\hat{\lambda} = \frac{1}{\overline{X}}$ is biased;

$$E[\hat{\lambda}] = E\left[\frac{1}{\overline{X}}\right] \neq \frac{1}{E[\overline{X}]}$$

- In general, if g is a non-linear function,

$$E[\hat{\theta}] = \theta \quad \not\Rightarrow \quad E[g(\hat{\theta})] = g(\theta)$$

A digression: Jensen’s inequality

- In some cases the sign of the bias is predictable.

A digression: Jensen's inequality

- In some cases the sign of the bias is predictable.
- **Jensen's inequality:** If $g(x)$ is a convex function and X is *any* random variable,

$$E[g(X)] \geq g\left(E[X]\right)$$

A digression: Jensen's inequality

- In some cases the sign of the bias is predictable.
- **Jensen's inequality:** If $g(x)$ is a convex function and X is *any* random variable,

$$E[g(X)] \geq g\left(E[X]\right)$$

- The inequality is strict for stric convexity and reversed for concave functions.

A digression: Jensen's inequality

- In some cases the sign of the bias is predictable.
- **Jensen's inequality:** If $g(x)$ is a convex function and X is *any* random variable,

$$E[g(X)] \geq g\left(E[X]\right)$$

- The inequality is strict for stric convexity and reversed for concave functions.
- **Example:** $g(x) = 1/x$ is convex, so

$$E[g(\overline{X})] = E[1/\overline{X}] \geq 1/E[\overline{X}] = g\left(E[\overline{X}]\right)$$

Unbiasedness (IV)

- Among two unbiased estimators, we would prefer the one with smaller variance.

Unbiasedness (IV)

- Among two unbiased estimators, we would prefer the one with smaller variance.
- If any of both are biased, we have to take this into account.

Unbiasedness (IV)

- Among two unbiased estimators, we would prefer the one with smaller variance.
- If any of both are biased, we have to take this into account.
- One way is to select the one with minimum mean squared error (MSE).

$$\begin{aligned} \mathrm{MSE}(\hat{c}) &= E[(\hat{c} - c)^2] \\ &= E[(\hat{c} - E(\hat{c}) + E(\hat{c}) - c)^2] \\ &= E[\hat{c} - E(\hat{c})]^2 + E[E(\hat{c}) - c)]^2 \\ &= \sigma_{\hat{c}}^2 + (\mathrm{bias}(\hat{c}))^2 \end{aligned}$$

Unbiasedness (IV)

- Among two unbiased estimators, we would prefer the one with smaller variance.
- If any of both are biased, we have to take this into account.
- One way is to select the one with minimum mean squared error (MSE).

$$\begin{aligned}
\mathrm{MSE}(\hat{c}) &= E[(\hat{c} - c)^2] \\
&= E[(\hat{c} - E(\hat{c}) + E(\hat{c}) - c)^2] \\
&= E[\hat{c} - E(\hat{c})]^2 + E[E(\hat{c}) - c)]^2 \\
&= \sigma_{\hat{c}}^2 + (\mathrm{bias}(\hat{c}))^2
\end{aligned}$$

What implicit assumption does MSE about gravity of estimation error?

Unbiasedness (IV)

- Among two unbiased estimators, we would prefer the one with smaller variance.
- If any of both are biased, we have to take this into account.
- One way is to select the one with minimum mean squared error (MSE).

$$\begin{aligned}\mathrm{MSE}(\hat{c}) &= E[(\hat{c}-c)^2] \\ &= E[(\hat{c}-E(\hat{c})+E(\hat{c})-c)^2] \\ &= E[\hat{c}-E(\hat{c})]^2 + E[E(\hat{c})-c)]^2 \\ &= \sigma^2_{\hat{c}} + (\mathrm{bias}(\hat{c}))^2\end{aligned}$$

"Twice as large, four times as bad." Arbitrary, mathematically convenient.

Even “good” estimators may be biased

- Estimators generally enjoying good properties may nonetheless be biased.

Even “good” estimators may be biased

- Estimators generally enjoying good properties may nonetheless be biased.
- An example is $\hat{\theta}_{MLE}$ in a $U(0, \theta)$. We have seen that $\hat{\theta}_{MLE} = X_{(n)}$. Since *always* $X_{(n)} \leq \theta$, it is clear that $E[X_{(n)}] < \theta$.

Even "good" estimators may be biased

- Estimators generally enjoying good properties may nonetheless be biased.
- An example is $\hat{\theta}_{MLE}$ in a $U(0, \theta)$. We have seen that $\hat{\theta}_{MLE} = X_{(n)}$. Since *always* $X_{(n)} \leq \theta$, it is clear that $E[X_{(n)}] < \theta$.
- Quite often MLE is biased, although in general it is asymptotically unbiased.

Consistency (I) (reminder: probability limits)

- We say that the limit in probability of a sequence or random variables $\{Z_n\}$ is Z if for any $\epsilon > 0$, $\eta > 0$ there is N such that for $n > N$:

$$P(|Z_n - Z| < \epsilon) \geq 1 - \eta$$

Consistency (I) (reminder: probability limits)

- We say that the limit in probability of a sequence or random variables $\{Z_n\}$ is Z if for any $\epsilon > 0$, $\eta > 0$ there is N such that for $n > N$:

$$P(|Z_n - Z| < \epsilon) \geq 1 - \eta$$

- In plain English: if taking sufficiently advanced terms of $\{Z_n\}$ we can be within ϵ of Z with probability as close to 1 as we wish.

Consistency (I) (reminder: probability limits)

- We say that the limit in probability of a sequence or random variables $\{Z_n\}$ is Z if for any $\epsilon > 0$, $\eta > 0$ there is N such that for $n > N$:

$$P(|Z_n - Z| < \epsilon) \geq 1 - \eta$$

- In plain English: if taking sufficiently advanced terms of $\{Z_n\}$ we can be within ϵ of Z with probability as close to 1 as we wish.
- Compare with usual notion of limit in mathematical analysis.

Consistency (I) (reminder: probability limits)

- We say that the limit in probability of a sequence or random variables $\{Z_n\}$ is Z if for any $\epsilon > 0$, $\eta > 0$ there is N such that for $n > N$:

$$P(|Z_n - Z| < \epsilon) \geq 1 - \eta$$

- In plain English: if taking sufficiently advanced terms of $\{Z_n\}$ we can be within ϵ of Z with probability as close to 1 as we wish.
- Compare with usual notion of limit in mathematical analysis.
- Usual notation is $Z_n \overset{p}{\to} Z$ or $\text{plim}(Z_n) = Z$.

Consistency (II)

- $\hat{\theta}_n$ denotes an estimator of θ based on a sample of size n. For instance, we might have

$$\hat{\theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Consistency (II)

- $\hat{\theta}_n$ denotes an estimator of θ based on a sample of size n. For instance, we might have

$$\hat{\theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- $\hat{\theta}_n$ is *consistent* if $\hat{\theta}_n \xrightarrow{p} \theta$

Consistency (II)

- $\hat{\theta}_n$ denotes an estimator of θ based on a sample of size n. For instance, we might have

$$\hat{\theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- $\hat{\theta}_n$ is *consistent* if $\hat{\theta}_n \xrightarrow{p} \theta$
- In plain English: if by increasing the sample size we can obtain arbitrary precision with as close to 1 confidence as we choose.

Consistency (II)

- $\hat{\theta}_n$ denotes an estimator of θ based on a sample of size n. For instance, we might have

$$\hat{\theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- $\hat{\theta}_n$ is *consistent* if $\hat{\theta}_n \xrightarrow{p} \theta$
- In plain English: if by increasing the sample size we can obtain arbitrary precision with as close to 1 confidence as we choose.
- In general, consistency is the very least we ask for. (We want to be rewarded for our effort in sampling!)

Consistency (III)

- We can usually show consistency by using:

Consistency (III)

- We can usually show consistency by using:
  1. The laws of large numbers

Consistency (III)

- We can usually show consistency by using:
  1. The laws of large numbers
  2. Tchebychev inequality.

How can we have consistency and not unbiasedness?

Consistency (III)

- We can usually show consistency by using:
  1. The laws of large numbers
  2. Tchebychev inequality.
- Consistency does not imply unbiasedness.

Think of $\hat{\theta}_n$ taking the true value θ with probability $1 - \frac{1}{n}$ and the value n with probability $\frac{1}{n}$.

Consistency vía Tchebychev inequality

Example: consistency of $\hat{\lambda} = \overline{X}$ as estimator of λ of a $\mathcal{P}(\lambda)$.

- We know $E[\hat{\lambda}] = \lambda$ and $\mathrm{Var}(\hat{\lambda}) = \lambda/n$.

Consistency vía Tchebychev inequality

Example: consistency of $\hat{\lambda} = \overline{X}$ as estimator of λ of a $\mathcal{P}(\lambda)$.

- We know $E[\hat{\lambda}] = \lambda$ and $\mathrm{Var}(\hat{\lambda}) = \lambda/n$.
- Then (Tchebycheff),

$$P(|\hat{\lambda} - \lambda| < \underbrace{k\sqrt{\lambda/n}}_{\epsilon}) \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

Consistency vía Tchebychev inequality

Example: consistency of $\hat{\lambda} = \overline{X}$ as estimator of λ of a $\mathcal{P}(\lambda)$.

- We know $E[\hat{\lambda}] = \lambda$ and $\mathrm{Var}(\hat{\lambda}) = \lambda/n$.
- Then (Tchebycheff),

$$P(|\hat{\lambda} - \lambda| < \underbrace{k\sqrt{\lambda/n}}_{\epsilon}) \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

- Make your pick of $1 - \eta$ as close to 1 as desired; whatever the implied k, we only have to choose n large enough to make ϵ as small as we wish.

Sometimes consistency can be checked directly

- Consider the case $X \sim U(0, \theta)$ where we showed $\hat{\theta}_{MLE} = X_{(n)}$.

Sometimes consistency can be checked directly

- Consider the case $X \sim U(0, \theta)$ where we showed $\hat{\theta}_{MLE} = X_{(n)}$.
- Now, the probability that $\hat{\theta}_{MLE}$ is **not** within $\epsilon > 0$ distance of true θ is the probability that $X_{(n)}$ (and hence `all` sample values) are below $(\theta - \epsilon)$:

$$P(|\hat{\theta}_{MLE} - \theta| > \epsilon) = \left(\frac{\theta - \epsilon}{\theta}\right)^n$$

and the last term clearly goes to zero as $n \to \infty$.

Sometimes consistency can be checked directly

- Consider the case $X \sim U(0, \theta)$ where we showed $\hat{\theta}_{MLE} = X_{(n)}$.
- Now, the probability that $\hat{\theta}_{MLE}$ is **not** within $\epsilon > 0$ distance of true θ is the probability that $X_{(n)}$ (and hence `all` sample values) are below $(\theta - \epsilon)$:

  $$P(|\hat{\theta}_{MLE} - \theta| > \epsilon) = \left(\frac{\theta - \epsilon}{\theta}\right)^n$$

  and the last term clearly goes to zero as $n \to \infty$.
- Therefore, it is clear that $\hat{\theta}_{MLE} \xrightarrow{p} \theta$.

Unbiasedness + variance$\to 0 \implies$ consistency

- Again, simple application of Tchebychev’s inequality.

Unbiasedness + variance$\rightarrow$ 0 $\Longrightarrow$ consistency

- Again, simple application of Tchebychev's inequality.
- Unbiasedness implies $E(\hat{\theta}_n) = \theta$.

$$P(|\hat{\theta}_n - \theta| < \underbrace{k\sigma_n}_{\epsilon} \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

Unbiasedness + variance$\to$ 0 $\Longrightarrow$ consistency

- Again, simple application of Tchebychev's inequality.
- Unbiasedness implies $E(\hat{\theta}_n) = \theta$.

$$P(|\hat{\theta}_n - \theta| < \underbrace{k\sigma_n}_{\epsilon} \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

- Let $1 - \eta$ be as close to 1 as desired; whatever the implied k, ϵ can be made small for large n, as $\sigma_n \to 0$.

Unbiasedness + variance$\to$ 0 $\Longrightarrow$ consistency

- Again, simple application of Tchebychev's inequality.
- Unbiasedness implies $E(\hat{\theta}_n) = \theta$.

$$P(|\hat{\theta}_n - \theta| < \underbrace{k\sigma_n}_{\epsilon} \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

- Let $1 - \eta$ be as close to 1 as desired; whatever the implied k, ϵ can be made small for large n, as $\sigma_n \to 0$.
- If both variance and bias decrease to zero, we also have consistency.

Consistency of moment estimators (I)

- Moment estimators are usually consistent.

Consistency of moment estimators (I)

- Moment estimators are usually consistent.
- Sketch of argument for a particular case:

$$\alpha_1(\theta) = m = \overline{X}$$

Consistency of moment estimators (I)

- Moment estimators are usually consistent.
- Sketch of argument for a particular case:

$$\alpha_1(\theta) = m = \overline{X}$$

- **If** $m = \alpha_1(\theta)$ has a continuous inverse function, $\hat{\theta} = \alpha_1^{-1}(\overline{X})$.

Consistency of moment estimators (I)

- Moment estimators are usually consistent.
- Sketch of argument for a particular case:

$$\alpha_1(\theta) = m = \overline{X}$$

- **If** $m = \alpha_1(\theta)$ has a continuous inverse function, $\hat{\theta} = \alpha_1^{-1}(\overline{X})$.
- Now, convergence of $\overline{X}$ to m (law of large numbers) entails convergence of $\hat{\theta}$ to θ:

$$\mathrm{plim}(\hat{\theta}) = \alpha_1^{-1}(\mathrm{plim}(\overline{X})) = \alpha_1^{-1}(m) = \theta$$

Consistency of moment estimators (I)

- Moment estimators are usually consistent.
- Sketch of argument for a particular case:

$$\alpha_1(\theta) = m = \overline{X}$$

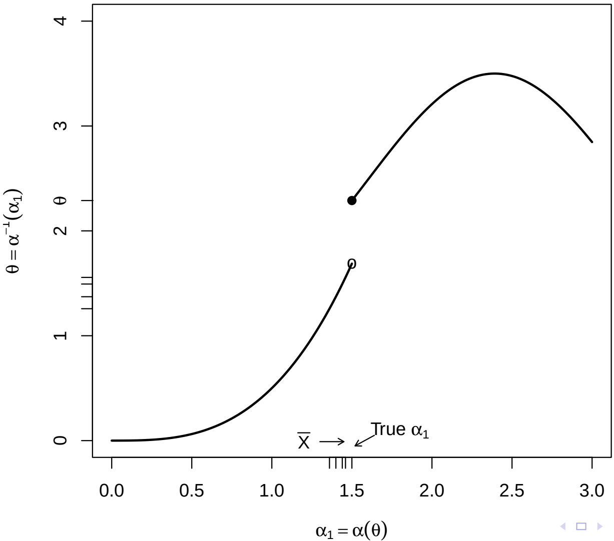
- **If** $m = \alpha_1(\theta)$ has a continuous inverse function, $\hat{\theta} = \alpha_1^{-1}(\overline{X})$.
- Now, convergence of $\overline{X}$ to m (law of large numbers) entails convergence of $\hat{\theta}$ to θ:

$$\mathrm{plim}(\hat{\theta}) = \alpha_1^{-1}(\mathrm{plim}(\overline{X})) = \alpha_1^{-1}(m) = \theta$$

- Notice: if $\alpha_q^{-1}()$ were not continuous, $\overline{X}$ could be very close of m and $\alpha_q^{-1}(\overline{X})$ **not** close to $\alpha_q^{-1}(m) = \theta$.

Consistency of moment estimators (II)

An illustration of what happens:



Consistency is not everything!

- Consistency is an asymptotic property. It tells us what happen when the sample size goes to infinity.

Consistency is not everything!

- Consistency is an asymptotic property. It tells us what happen when the sample size goes to infinity.
- In practice, we may be limited to small samples, and then the consistency property offers little comfort.

Consistency is not everything!

- Consistency is an asymptotic property. It tells us what happen when the sample size goes to infinity.
- In practice, we may be limited to small samples, and then the consistency property offers little comfort.
- Example: (artificial). In a $\mathcal{P}(\lambda)$,

$$\hat{\lambda}_n = \begin{cases} 0 & \text{if } n < 10^5. \\ \overline{X} & \text{if } n \geq 10^5. \end{cases}$$

would be consistent (but pretty bad for sample sizes *n* below 10^5!).

Consistency is not everything!

- Consistency is an asymptotic property. It tells us what happen when the sample size goes to infinity.
- In practice, we may be limited to small samples, and then the consistency property offers little comfort.
- Example: (artificial). In a $\mathcal{P}(\lambda)$,

  $$\hat{\lambda}_n = \begin{cases} 0 & \text{if } n < 10^5. \\ \overline{X} & \text{if } n \geq 10^5. \end{cases}$$

  would be consistent (but pretty bad for sample sizes n below 10^5!).
- Consistency is reassuring, but we need to check for realistic sample sizes (often through simulation).

Efficiency (I)

- Ideally we would want an estimator that is always better than any other: the very best.

Efficiency (I)

- Ideally we would want an estimator that is always better than any other: the very best.
- Such thing does not exist. Consider a Poisson(θ).

Efficiency (I)

- Ideally we would want an estimator that is always better than any other: the very best.
- Such thing does not exist. Consider a Poisson(θ).
- $\overline{X}$ is the (moment, MLE) estimator of θ.

Efficiency (I)

- Ideally we would want an estimator that is always better than any other: the very best.
- Such thing does not exist. Consider a Poisson(θ).
- $\overline{X}$ is the (moment, MLE) estimator of θ.
- The (rather stupid) estimator $\hat{\theta}_* = 2.3$ is better when θ happens to be 2.3!

Efficiency (I)

- Ideally we would want an estimator that is always better than any other: the very best.
- Such thing does not exist. Consider a Poisson(θ).
- $\overline{X}$ is the (moment, MLE) estimator of θ.
- The (rather stupid) estimator $\hat{\theta}_* = 2.3$ is better when θ happens to be 2.3!
- Want to exclude of consideration those estimators which only work in certain circumstances.

Efficiency (I)

- Ideally we would want an estimator that is always better than any other: the very best.
- Such thing does not exist. Consider a Poisson(θ).
- $\overline{X}$ is the (moment, MLE) estimator of θ.
- The (rather stupid) estimator $\hat{\theta}_* = 2.3$ is better when θ happens to be 2.3!
- Want to exclude of consideration those estimators which only work in certain circumstances.
- Requiring unbiasedness is one way of excluding of consideration estimators such as $\hat{\theta}_*$.

Efficiency (II)

- If we restrict our attention to unbiased estimators, it makes sense to chose the one with smallest variance.

Efficiency (II)

- If we restrict our attention to unbiased estimators, it makes sense to chose the one with smallest variance.
- For $\hat{\theta}_1$ and $\hat{\theta}_*$ both unbiased estimators of θ, we define *efficiency of* $\hat{\theta}_1$ *relative to* $\hat{\theta}_*$ as:

$$\frac{\mathrm{Var}(\hat{\theta}_*)}{\mathrm{Var}(\hat{\theta}_1)}$$

Efficiency (II)

- If we restrict our attention to unbiased estimators, it makes sense to chose the one with smallest variance.
- For $\hat{\theta}_1$ and $\hat{\theta}_*$ both unbiased estimators of θ, we define *efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_*$* as:

$$\frac{\mathrm{Var}(\hat{\theta}_*)}{\mathrm{Var}(\hat{\theta}_1)}$$

- Assume Var(θ_*) were the lowest attainable. Then, any estimator with efficiency 1 relative to $\hat{\theta}_*$ will be called **efficient**.

Efficiency: a trivial example

- Let $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, 5$.

Efficiency: a trivial example

- Let $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, 5$.
- Then, $\hat{m}_1 = \frac{X_1 + \ldots + X_5}{5}$ has variance $\frac{1}{5}$.

Efficiency: a trivial example

- Let $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, 5$.
- Then, $\hat{m}_1 = \frac{X_1 + \ldots + X_5}{5}$ has variance $\frac{1}{5}$.
- If we neglect some observations, as in $\hat{m}_2 = \frac{X_1 + \ldots + X_3}{3}$ the variance is $\frac{1}{3}$.

Efficiency: a trivial example

- Let $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, 5$.
- Then, $\hat{m}_1 = \frac{X_1 + \ldots + X_5}{5}$ has variance $\frac{1}{5}$.
- If we neglect some observations, as in $\hat{m}_2 = \frac{X_1 + \ldots + X_3}{3}$ the variance is $\frac{1}{3}$.
- The efficiency of $\hat{m}_2$ relative to $\hat{m}_1$ is:

$$\frac{\mathrm{Var}(\hat{m}_1)}{\mathrm{Var}(\hat{m}_2)} = \frac{3}{5}$$

Efficiency: a less trivial example

- Consider again $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, n = 2k + 1$. $\overline{X}$ is an unbiased estimator of the mean with variance $\frac{1}{n}$

Efficiency: a less trivial example

- Consider again $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, n = 2k + 1$. $\overline{X}$ is an unbiased estimator of the mean with variance $\frac{1}{n}$
- It can be shown that $\hat{m}_* = X_{(k+1)}$ (= the median) has variance $2\pi\sigma^2/4n$ in the normal case.

Efficiency: a less trivial example

- Consider again $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, n = 2k + 1$. $\overline{X}$ is an unbiased estimator of the mean with variance $\frac{1}{n}$
- It can be shown that $\hat{m}_* = X_{(k+1)}$ (= the median) has variance $2\pi\sigma^2/4n$ in the normal case.
- The efficiency of $\hat{m}_*$ relative to $\overline{X}$ is:

$$\frac{\mathrm{Var}(\overline{X})}{\mathrm{Var}(\hat{m}_*)} = \frac{2}{\pi} \approx 0.6366$$

if we indeed are sampling a normal distribution.

Efficiency: a less trivial example

- Consider again $X_i \sim N(m, \sigma = 1)$ for $i = 1, \ldots, n = 2k + 1$. $\overline{X}$ is an unbiased estimator of the mean with variance $\frac{1}{n}$
- It can be shown that $\hat{m}_* = X_{(k+1)}$ (= the median) has variance $2\pi\sigma^2/4n$ in the normal case.
- The efficiency of $\hat{m}_*$ relative to $\overline{X}$ is:

  $$\frac{\mathrm{Var}(\overline{X})}{\mathrm{Var}(\hat{m}_*)} = \frac{2}{\pi} \approx 0.6366$$

  if we indeed are sampling a normal distribution.
- We might tolerate this loss of efficiency to protect ourselves against a heavy tail distribution (like Cauchy).

The Cramer-Rao bound

- But, how do we know $\hat{\theta}$ cannot be improved upon?

The Cramer-Rao bound

- But, how do we know $\hat{\theta}$ cannot be improved upon?
- It turns out that we do have a universal yardstick, under *regularity* conditions (more on that later)

The Cramer-Rao bound

- But, how do we know $\hat{\theta}$ cannot be improved upon?
- It turns out that we do have a universal yardstick, under *regularity* conditions (more on that later)
- For any unbiased $\hat{\theta}$ based on n observations under regularity conditions:

  $$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)};$$

  this is the celebrated Cramer-Rao lower bound.

The Cramer-Rao bound

- But, how do we know $\hat{\theta}$ cannot be improved upon?
- It turns out that we do have a universal yardstick, under *regularity* conditions (more on that later)
- For any unbiased $\hat{\theta}$ based on *n* observations under regularity conditions:

  $$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)};$$

  this is the celebrated Cramer-Rao lower bound.
- $I(\theta)$ is the so-called Fisher information contained in one observation, and is defined as:

  $$I(\theta) = E\left(\frac{\partial \log f(X;\theta)}{\partial \theta}\right)^2$$

Intuition for Fisher information

- Why is $I(\theta)$ a measure of information?

Intuition for Fisher information

- Why is $I(\theta)$ a measure of information?
- Imagine a given (fixed) x;

$$\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2$$

measures how fast $\log f(x;\theta)$ changes in response to changes in θ.

Intuition for Fisher information

- Why is $I(\theta)$ a measure of information?
- Imagine a given (fixed) x;

  $$\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2$$

  measures how fast $\log f(x;\theta)$ changes in response to changes in θ.
- If $\log f(x;\theta)$ were very flat, close values of θ would have similar likelihood, and we would be very uncertain about the “true” θ.

Intuition for Fisher information

- Why is $I(\theta)$ a measure of information?
- Imagine a given (fixed) x;

  $$\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2$$

  measures how fast $\log f(x;\theta)$ changes in response to changes in θ.
- If $\log f(x;\theta)$ were very flat, close values of θ would have similar likelihood, and we would be very uncertain about the “true” θ.
- If $\log f(x;\theta)$ changes fast, it gives much information about θ.

Intuition for Fisher information

- Why is $I(\theta)$ a measure of information?
- Imagine a given (fixed) x;

  $$\left(\frac{\partial \log f(x;\theta)}{\partial\theta}\right)^2$$

  measures how fast $\log f(x;\theta)$ changes in response to changes in θ.
- If $\log f(x;\theta)$ were very flat, close values of θ would have similar likelihood, and we would be very uncertain about the “true” θ.
- If $\log f(x;\theta)$ changes fast, it gives much information about θ.
- If we average the derivative over possible values of X we have Fisher information.

Efficient estimators and the Cramer-Rao bound

- Under regularity conditions, if

  $$\mathrm{Var}(\hat{\theta}) = \frac{1}{nI(\theta)};$$

  the Cramer-Rao lower bound implies the unbiased $\hat{\theta}$ cannot be improved upon by any other unbiased estimator. It is then called **efficient.**

Efficient estimators and the Cramer-Rao bound

- Under regularity conditions, if

  $$\mathrm{Var}(\hat{\theta}) = \frac{1}{nI(\theta)};$$

  the Cramer-Rao lower bound implies the unbiased $\hat{\theta}$ cannot be improved upon by any other unbiased estimator. It is then called **efficient.**
- We *know* what the optimum is before we start.

Efficient estimators and the Cramer-Rao bound

- Under regularity conditions, if

  $$\mathrm{Var}(\hat{\theta}) = \frac{1}{nI(\theta)};$$

  the Cramer-Rao lower bound implies the unbiased $\hat{\theta}$ cannot be improved upon by any other unbiased estimator. It is then called **efficient.**
- We *know* what the optimum is before we start.
- No fear that there is a better estimator that just didn't occur to us!

The Cramer-Rao bound: historical notes

- Harald Crameŕ (1892-1985), swedish statistician, author of the extremely influential *Mathematical Methods of Statistics* (1946), still a good reading.

The Cramer-Rao bound: historical notes

- Harald Crameŕ (1892-1985), swedish statistician, author of the extremely influential *Mathematical Methods of Statistics* (1946), still a good reading.
- C.R.Rao (1920-), a distinguished indian statistician. Aside from the Cramer-Rao bound, other contributions like the celebrated Rao-Blackwell theorem (in the same vein than the Cramer-Rao bound, but more powerful).

The Cramer-Rao bound: historical notes

- Harald Crameŕ (1892-1985), swedish statistician, author of the extremely influential *Mathematical Methods of Statistics* (1946), still a good reading.
- C.R.Rao (1920-), a distinguished indian statistician. Aside from the Cramer-Rao bound, other contributions like the celebrated Rao-Blackwell theorem (in the same vein than the Cramer-Rao bound, but more powerful).
- The original publications date of 1945 (Rao) and 1946 (Cramer).

What are those regularity conditions?

- Basically,

What are those regularity conditions?

- Basically,
  1. The support of the distribution does not depend on the parameter. Example of violation: $U(0, \theta)$.

What are those regularity conditions?

- Basically,
  1. The support of the distribution does not depend on the parameter. Example of violation: $U(0, \theta)$.
  2. The log likelihood function “sufficiently smooth”: differentiable and order of integration and differentiation interchangeable:

$$\frac{\partial}{\partial\theta} E(\log f(X, \theta)) = E\left(\frac{\partial \log f(X, \theta)}{\partial\theta}\right)$$

What are those regularity conditions?

- Basically,
  1. The support of the distribution does not depend on the parameter. Example of violation: $U(0, \theta)$.
  2. The log likelihood function “sufficiently smooth”: differentiable and order of integration and differentiation interchangeable:

$$\frac{\partial}{\partial\theta} E(\log f(X, \theta)) = E\left(\frac{\partial \log f(X, \theta)}{\partial\theta}\right)$$

- Failure of these conditions render unusable the Cramer-Rao bound.

A trick to compute the Cramer-Rao bound.

- It turns out that

$$E\left(\frac{\partial \log f(X,\theta)}{\partial\theta}\right)^2 = -E\left(\frac{\partial^2 \log f(X,\theta)}{\partial\theta^2}\right)$$

A trick to compute the Cramer-Rao bound.

- It turns out that

$$E\left(\frac{\partial \log f(X,\theta)}{\partial \theta}\right)^2 = -E\left(\frac{\partial^2 \log f(X,\theta)}{\partial \theta^2}\right)$$

- Either expression can be used to compute Fisher's information (the denominator of the Cramer-Rao bound).

A trick to compute the Cramer-Rao bound.

- It turns out that

$$E\left(\frac{\partial \log f(X,\theta)}{\partial\theta}\right)^2 = -E\left(\frac{\partial^2 \log f(X,\theta)}{\partial\theta^2}\right)$$

- Either expression can be used to compute Fisher's information (the denominator of the Cramer-Rao bound).
- Usually best the second derivative, but sometimes looking at the first we can easily compute its mean value.

The Cramer-Rao bound: examples (I)

We know $\overline{X}$ is unbiased for λ in a $\mathcal{P}(\lambda)$. Its variance is λ/n. Is there anything better?

The Cramer-Rao bound: examples (I)

We know $\overline{X}$ is unbiased for λ in a $\mathcal{P}(\lambda)$. Its variance is λ/n. Is there anything better?

$$\log f(X, \lambda) \quad = \quad -\lambda + X \log(\lambda) - \log(X!)$$

The Cramer-Rao bound: examples (I)

We know $\overline{X}$ is unbiased for λ in a $\mathcal{P}(\lambda)$. Its variance is λ/n. Is there anything better?

$$\begin{aligned} \log f(X,\lambda) &= -\lambda + X\log(\lambda) - \log(X!) \\ \frac{\partial \log f(X,\lambda)}{\partial \lambda} &= -1 + X/\lambda = \left(\frac{X-\lambda}{\lambda}\right) \end{aligned}$$

The Cramer-Rao bound: examples (I)

We know $\overline{X}$ is unbiased for λ in a $\mathcal{P}(\lambda)$. Its variance is λ/n. Is there anything better?

$$\begin{aligned}
\log f(X,\lambda) &= -\lambda + X\log(\lambda) - \log(X!) \\
\frac{\partial \log f(X,\lambda)}{\partial\lambda} &= -1 + X/\lambda = \left(\frac{X-\lambda}{\lambda}\right) \\
E\left(\frac{X-\lambda}{\lambda}\right)^2 &= \frac{1}{\lambda}
\end{aligned}$$

The Cramer-Rao bound: examples (I)

We know $\overline{X}$ is unbiased for λ in a $\mathcal{P}(\lambda)$. Its variance is λ/n. Is there anything better?

$$\begin{aligned}
\log f(X,\lambda) &= -\lambda + X\log(\lambda) - \log(X!) \\
\frac{\partial \log f(X,\lambda)}{\partial\lambda} &= -1 + X/\lambda = \left(\frac{X-\lambda}{\lambda}\right) \\
E\left(\frac{X-\lambda}{\lambda}\right)^2 &= \frac{1}{\lambda}
\end{aligned}$$

The Cramer-Rao is

$$\mathrm{Var}(\hat{\lambda}) \geq \frac{1}{n\frac{1}{\lambda}} = \frac{\lambda}{n}$$

so $\overline{X}$ is optimal in the unbiased class.

The Cramer-Rao bound: examples (II)

- We might have missed the fact that:

$$E\left(\frac{X-\lambda}{\lambda}\right)^2 = \frac{1}{\lambda};$$

The Cramer-Rao bound: examples (II)

- We might have missed the fact that:

$$E\left(\frac{X-\lambda}{\lambda}\right)^2 = \frac{1}{\lambda};$$

- In that case, taking the derivative of

$$\left(\frac{X-\lambda}{\lambda}\right)$$

  would have readily given us $1/\lambda$.

The Cramer-Rao bound: examples (III)

- Consider estimation of p in a binary distribution.

The Cramer-Rao bound: examples (III)

- Consider estimation of p in a binary distribution.
- Moment and MLE is $\hat{p} = \overline{X}$ with variance $p(1-p)/n$.

The Cramer-Rao bound: examples (III)

- Consider estimation of p in a binary distribution.
- Moment and MLE is $\hat{p} = \overline{X}$ with variance $p(1-p)/n$.
- We have,

$$\log f(X,p) \quad = \quad X\log(p) + (1-X)\log(1-p)$$

The Cramer-Rao bound: examples (III)

- Consider estimation of p in a binary distribution.
- Moment and MLE is $\hat{p} = \overline{X}$ with variance $p(1-p)/n$.
- We have,

$$\begin{aligned} \log f(X,p) &= X\log(p) + (1-X)\log(1-p) \\ \frac{\partial \log f(X,p)}{\partial p} &= \frac{X}{p} - \frac{1-X}{1-p} \end{aligned}$$

The Cramer-Rao bound: examples (III)

- Consider estimation of p in a binary distribution.
- Moment and MLE is $\hat{p} = \overline{X}$ with variance $p(1-p)/n$.
- We have,

$$\begin{aligned}
\log f(X,p) &= X\log(p) + (1-X)\log(1-p) \\
\frac{\partial \log f(X,p)}{\partial p} &= \frac{X}{p} - \frac{1-X}{1-p} \\
E\left(\frac{X}{p} - \frac{1-X}{1-p}\right)^2 &= E\left(\frac{X-p}{p(1-p)}\right)^2 = \frac{1}{p(1-p)}
\end{aligned}$$

The Cramer-Rao bound: examples (III)

- Consider estimation of p in a binary distribution.
- Moment and MLE is $\hat{p} = \overline{X}$ with variance $p(1-p)/n$.
- We have,

$$\begin{aligned} \log f(X,p) &= X\log(p) + (1-X)\log(1-p) \\ \frac{\partial \log f(X,p)}{\partial p} &= \frac{X}{p} - \frac{1-X}{1-p} \\ E\left(\frac{X}{p} - \frac{1-X}{1-p}\right)^2 &= E\left(\frac{X-p}{p(1-p)}\right)^2 = \frac{1}{p(1-p)} \end{aligned}$$

- The CR bound is then,

$$\mathrm{Var}(\hat{p}) \geq \frac{1}{n\frac{1}{p(1-p)}} = \frac{p(1-p)}{n}$$

and $\hat{p} = \overline{X}$ is efficient.

Some facts about the Cramer-Rao bound

- The CR bound may not be attainable.

Some facts about the Cramer-Rao bound

- The CR bound may not be attainable.
- What it says is that we can do no better. . .

Some facts about the Cramer-Rao bound

- The CR bound may not be attainable.
- What it says is that we can do no better. . .
- . . .not that we can do as well.

Some facts about the Cramer-Rao bound

- The CR bound may not be attainable.
- What it says is that we can do no better. . .
- . . .not that we can do as well.
- Hence, estimators with efficiency 1 as defined previously, may not exist.

Some facts about the Cramer-Rao bound

- The CR bound may not be attainable.
- What it says is that we can do no better. . .
- . . .not that we can do as well.
- Hence, estimators with efficiency 1 as defined previously, may not exist.
- In general, the MLE reaches the CR lower bound, at least asymptotically.

The concept of sufficiency (I)

- To obtain estimators, we have made use of a *statistic*, a function of the sample.

The concept of sufficiency (I)

- To obtain estimators, we have made use of a *statistic*, a function of the sample.
- Are we losing something?

The concept of sufficiency (I)

- To obtain estimators, we have made use of a *statistic*, a function of the sample.
- Are we losing something?
- Or, could we do better looking individually at each sample value, rather than to a summarizing function?

The concept of sufficiency (I)

- To obtain estimators, we have made use of a *statistic*, a function of the sample.
- Are we losing something?
- Or, could we do better looking individually at each sample value, rather than to a summarizing function?
- Loose idea: when a statistic “squeezes all the juice” out of a sample, it is sufficient.

The concept of sufficiency (I)

- To obtain estimators, we have made use of a *statistic*, a function of the sample.
- Are we losing something?
- Or, could we do better looking individually at each sample value, rather than to a summarizing function?
- Loose idea: when a statistic “squeezes all the juice” out of a sample, it is sufficient.
- We have to formalize this “squeezing” property.

The concept of sufficiency (II)

- If given a statistic $S = S(\vec{X})$ the conditional density (or probability)
$$f(\vec{X}|S) = \frac{f_{\vec{X}}(\vec{X};\theta)}{f_S(S;\theta)}$$
is independent of θ, $S(\vec{X})$ is said to be **sufficient** for θ.

The concept of sufficiency (II)

- If given a statistic $S = S(\vec{X})$ the conditional density (or probability)
$$f(\vec{X}|S) = \frac{f_{\vec{X}}(\vec{X};\theta)}{f_S(S;\theta)}$$
is independent of θ, $S(\vec{X})$ is said to be **sufficient** for θ.
- Motivation: if once we know $S = S(\vec{X})$ the density (or probability) of the sample values does not depend on θ, *knowing those individual sample values cannot be of help in determining* θ*.*

The concept of sufficiency (II)

- If given a statistic $S = S(\vec{X})$ the conditional density (or probability)
$$f(\vec{X}|S) = \frac{f_{\vec{X}}(\vec{X};\theta)}{f_S(S;\theta)}$$
is independent of θ, $S(\vec{X})$ is said to be **sufficient** for θ.
- Motivation: if once we know $S = S(\vec{X})$ the density (or probability) of the sample values does not depend on θ, *knowing those individual sample values cannot be of help in determining* θ*.*
- All information about θ is then contained in $S = S(\vec{X})$.

The concept of sufficiency (III)

- Let $X_1, \ldots, X_n \sim \mathcal{P}(\lambda)$. Let $S = X_1 + \cdots + X_n$. We know $S \sim \mathcal{P}(n\lambda)$. Then

The concept of sufficiency (III)

- Let $X_1, \dots, X_n \sim \mathcal{P}(\lambda)$. Let $S = X_1 + \cdots + X_n$. We know $S \sim \mathcal{P}(n\lambda)$. Then

$$\begin{aligned} f(\vec{X}|S) &= \frac{f_{\vec{X}}(\vec{X};\lambda)}{f_S(S;\lambda)} \\ &= \frac{\prod_{i=1}^n e^{-\lambda}\lambda^{X_i}/X_1!}{e^{-n\lambda}(n\lambda)^S/S!} \\ &= \frac{S!}{X_1!X_2!\dots X_n!}n^{-S} \end{aligned}$$

The concept of sufficiency (III)

- Let $X_1, \ldots, X_n \sim \mathcal{P}(\lambda)$. Let $S = X_1 + \cdots + X_n$. We know $S \sim \mathcal{P}(n\lambda)$. Then

$$
\begin{aligned}
f(\vec{X}|S) &= \frac{f_{\vec{X}}(\vec{X};\lambda)}{f_S(S;\lambda)} \\
&= \frac{\prod_{i=1}^{n} e^{-\lambda}\lambda^{X_i}/X_1!}{e^{-n\lambda}(n\lambda)^S/S!} \\
&= \frac{S!}{X_1!X_2!\ldots X_n!} n^{-S}
\end{aligned}
$$

- Therefore, S (or any other 1-1 function of S) is sufficient for λ.

The concept of sufficiency (IV)

- As a further example, let's consider the ordered sample $X_{(1)}, \ldots, X_{(n)}$.

The concept of sufficiency (IV)

- As a further example, let's consider the ordered sample $X_{(1)}, \ldots, X_{(n)}$.
- If sampled values are i.i.d., values may arise in any order.

The concept of sufficiency (IV)

- As a further example, let's consider the ordered sample $X_{(1)}, \ldots, X_{(n)}$.
- If sampled values are i.i.d., values may arise in any order.
- *Given* $X_{(1)}, \ldots, X_{(n)}$, any order is equally likely, with probability $1/n!$, whichever the parameter(s) of the distribution may be.

The concept of sufficiency (IV)

- As a further example, let's consider the ordered sample $X_{(1)}, \ldots, X_{(n)}$.
- If sampled values are i.i.d., values may arise in any order.
- *Given* $X_{(1)}, \ldots, X_{(n)}$, any order is equally likely, with probability $1/n!$, whichever the parameter(s) of the distribution may be.
- Therefore, $X_{(1)}, \ldots, X_{(n)}$ is always a sufficient statistic, although of little interest (it doesn't "compact" information).

The factorization theorem (I)

- If we can decompose the joint density (or probability) as a product,
$$f_{\vec{X}}(\vec{X};\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$
where $h(\vec{X})$ does **not** depend on θ, then S is sufficient.

The factorization theorem (I)

- If we can decompose the joint density (or probability) as a product,
  $$f_{\vec{X}}(\vec{X};\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$
  where $h(\vec{X})$ does **not** depend on θ, then S is sufficient.
- Quite easy to prove.

The factorization theorem (I)

- If we can decompose the joint density (or probability) as a product,
$$f_{\vec{X}}(\vec{X};\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$
where $h(\vec{X})$ does **not** depend on θ, then S is sufficient.
- Quite easy to prove.
- Quite practical; we only have to see which function (or functions) of the sample “carry with them” the parameter θ.

The factorization theorem (II)

- Take the Poisson case again. We have,

$$\begin{aligned} f_{\vec{X}}(\vec{X};\lambda) &= \prod_{i=1}^{n} e^{-\lambda}\lambda^{X_i}/X_i! \\ &= \underbrace{e^{-n\lambda}\lambda^{X_1+\ldots+X_n}}_{g(S,\lambda)} \times \underbrace{\prod_{i=1}^{n}(1/X_i!)}_{h(\vec{X})} \end{aligned}$$

The factorization theorem (II)

- Take the Poisson case again. We have,

$$
\begin{aligned}
f_{\vec{X}}(\vec{X};\lambda) &= \prod_{i=1}^{n} e^{-\lambda}\lambda^{X_i}/X_i! \\
&= \underbrace{e^{-n\lambda}\lambda^{X_1+\ldots+X_n}}_{g(S,\lambda)} \times \underbrace{\prod_{i=1}^{n}(1/X_i!)}_{h(\vec{X})}
\end{aligned}
$$

- Clearly, $S = X_1 + \ldots + X_n$ is sufficient.

The factorization theorem (III)

- MLE have “built in” sufficiency.

The factorization theorem (III)

- MLE have “built in” sufficiency.
- Using the factorization theorem, to maximize the left hand side of

  $$f_{\vec{X}}(\vec{X};\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$

  as a function of θ**,** we only need $g(S(\vec{X});\theta)$;

The factorization theorem (III)

- MLE have “built in” sufficiency.
- Using the factorization theorem, to maximize the left hand side of
$$f_{\vec{X}}(\vec{X};\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$
as a function of θ**,** we only need $g(S(\vec{X});\theta)$;
- The term $h(\vec{X})$ is just a constant in the likelihood function.

Some ill-behaved distributions

- Most distributions in common us have sufficient statistics for their parameters.

Some ill-behaved distributions

- Most distributions in common us have sufficient statistics for their parameters.
- This is not always the case. Consider the Cauchy distribution (aka t_1) with location θ:

$$f_X(x : \theta) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$$

Some ill-behaved distributions

- Most distributions in common us have sufficient statistics for their parameters.
- This is not always the case. Consider the Cauchy distribution (aka t_1) with location θ:

$$f_X(x:\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

- If you use the factorization theorem to look for sufficient statistics,

$$f_{\vec{X}}(\vec{X};\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$

  hard as you may try, you will at least need the ordered sample (which is always a sufficient statistic).

Some ill-behaved distributions

- Most distributions in common us have sufficient statistics for their parameters.
- This is not always the case. Consider the Cauchy distribution (aka t_1) with location θ:

$$f_X(x:\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

- If you use the factorization theorem to look for sufficient statistics,

  $$f_{\vec{X}}(\vec{X};\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$

  hard as you may try, you will at least need the ordered sample (which is always a sufficient statistic).
- No further reduction is possible.

Logically equivalent statements (I)

- “If an animal is a whale, it lives in the water.”

Logically equivalent statements (I)

- “If an animal is a whale, it lives in the water.”
- What can be inferred for animals which live in the water?

Logically equivalent statements (I)

- “If an animal is a whale, it lives in the water.”
- What can be inferred for animals which live in the water?
- And for animals which do **not** live in the water?

Logically equivalent statements (I)

- “If an animal is a whale, it lives in the water.”
- What can be inferred for animals which live in the water?
- And for animals which do **not** live in the water?
- $\underbrace{\text{Is a whale}}_{p} \Longrightarrow \underbrace{\text{Lives in the water}}_{q}$

Logically equivalent statements (I)

- “If an animal is a whale, it lives in the water.”
- What can be inferred for animals which live in the water?
- And for animals which do **not** live in the water?
- $\underbrace{\text{Is a whale}}_{p} \Longrightarrow \underbrace{\text{Lives in the water}}_{q}$
- $\underbrace{\text{Does not live in water}}_{\neg q} \Longrightarrow \underbrace{\text{Is not a whale}}_{\neg p}$

Logically equivalent statements (II)

Quite generally,

- $p \Longrightarrow q$ and $\ulcorner q \Longrightarrow \ulcorner p$ are logically equivalent.
  ($\ulcorner$ above stands for negation:)

Logically equivalent statements (II)

Quite generally,

- $p \Longrightarrow q$ and $\neg\, q \Longrightarrow \neg\, p$ are logically equivalent.
  ($\neg$ above stands for negation:)
- Both are true or false.

Logically equivalent statements (II)

Quite generally,

- $p \Longrightarrow q$ and $\neg q \Longrightarrow \neg p$ are logically equivalent. ($\neg$ above stands for negation:)
- Both are true or false.
- When testing hypothesis, we rely on a softened versions of this equivalence.

Statements probabilistically related (I)

- Consider $p \Longrightarrow$ **most of the time** q.

Statements probabilistically related (I)

- Consider $p \implies$ **most of the time** q.
- Then $\neg\, q \implies \quad \neg\, p$ **is likely** (or p is unlikely).

Statements probabilistically related (I)

- Consider $p \Longrightarrow$ **most of the time** q.
- Then $\neg\, q \Longrightarrow \quad \neg\, p$ **is likely** (or p is unlikely).
- Same structure, only now the implications are not required to hold all times.

Statements probabilistically related (I)

- Consider $p \Longrightarrow$ **most of the time** q.
- Then $\neg q \Longrightarrow \quad \neg p$ **is likely** (or p is unlikely).
- Same structure, only now the implications are not required to hold all times.
- $\neg q$ is no longer proof of $\neg p$, *but can be taken as evidence in favour of it.*

Statements probabilistically related (II)

Example:

- $\underbrace{\text{Coin is regular}}_{p} \Longrightarrow$ **most of the time** $\underbrace{\text{about 50\% of heads}}_{q}$.

Statements probabilistically related (II)

Example:

- $\underbrace{\text{Coin is regular}}_{p} \Longrightarrow$ **most of the time** $\underbrace{\text{about 50\% of heads}}_{q}$.
- $\underbrace{\text{Far from 50\% of heads}}_{\neg q} \Longrightarrow \underbrace{\text{Coin not regular}}_{\neg p}$ **is likely**.

Statements probabilistically related (II)

Example:

- $\underbrace{\text{Coin is regular}}_{p} \Longrightarrow$ **most of the time** $\underbrace{\text{about 50\% of heads}}_{q}$.
- $\underbrace{\text{Far from 50\% of heads}}_{\neg q} \Longrightarrow \underbrace{\text{Coin not regular}}_{\neg p}$ **is likely**.
- $\underbrace{\text{Far from 50\% of head}}_{\neg q}$ is taken as evidence in favour of $\neg\, p$ (and therefore against p).

Hypothesis testing (I)

- A **null hypothesis** is an statement which we hold to be true.

Hypothesis testing (I)

- A **null hypothesis** is an statement which we hold to be true.
- If it is indeed true (p), a given experiment should very likely produce a result in a certain range (q).

Hypothesis testing (I)

- A **null hypothesis** is an statement which we hold to be true.
- If it is indeed true (p), a given experiment should very likely produce a result in a certain range (q).
- If it so happens that the result is not observed in the very likely range ($\neg q$), either:

Hypothesis testing (I)

- A **null hypothesis** is an statement which we hold to be true.
- If it is indeed true (p), a given experiment should very likely produce a result in a certain range (q).
- If it so happens that the result is not observed in the very likely range (⌐ q), either:
  1. Something very strange has happened (should not be the case very often)...

Hypothesis testing (I)

- A **null hypothesis** is an statement which we hold to be true.
- If it is indeed true (p), a given experiment should very likely produce a result in a certain range (q).
- If it so happens that the result is not observed in the very likely range ($\neg\, q$), either:
  1. Something very strange has happened (should not be the case very often). . .
  2. . . .or else the null hypothesis is not true to begin with.

Hypothesis testing (I)

- A **null hypothesis** is an statement which we hold to be true.
- If it is indeed true (p), a given experiment should very likely produce a result in a certain range (q).
- If it so happens that the result is not observed in the very likely range (⌐ q), either:
  1. Something very strange has happened (should not be the case very often)...
  2. ...or else the null hypothesis is not true to begin with.
- As statisticians, we go with the second option.

Hypothesis testing (II)

- Empiricism!

Hypothesis testing (II)

- Empiricism!
- If the experiment does not quite agree with the hypothesis, we scrap the hypothesis.

Hypothesis testing (II)

- Empiricism!
- If the experiment does not quite agree with the hypothesis, we scrap the hypothesis.
- *However,* we cannot completely rule out the possibility that something strange happened. We are bound to make errors!

Hypothesis testing (II)

- Empiricism!
- If the experiment does not quite agree with the hypothesis, we scrap the hypothesis.
- *However,* we cannot completely rule out the possibility that something strange happened. We are bound to make errors!
- But we try to keep those to a minimum.

Hypothesis testing (III)

- Notice: we start with an established piece of knowledge (the null hypothesis).

Hypothesis testing (III)

- Notice: we start with an established piece of knowledge (the null hypothesis).
- How we got there, there is no telling.

Hypothesis testing (III)

- Notice: we start with an established piece of knowledge (the null hypothesis).
- How we got there, there is no telling.
- Hypothesis testing does not tell us *how to learn*, but *how we put to test what we have somehow learned.*

Hypothesis testing (III)

- Notice: we start with an established piece of knowledge (the null hypothesis).
- How we got there, there is no telling.
- Hypothesis testing does not tell us *how to learn*, but *how we put to test what we have somehow learned.*
- Quite in keeping with ideas popular in mid XXth century (e.g., Lakatos, *Proofs and refutations*.)

Hypothesis testing (III)

- Notice: we start with an established piece of knowledge (the null hypothesis).
- How we got there, there is no telling.
- Hypothesis testing does not tell us *how to learn*, but *how we put to test what we have somehow learned.*
- Quite in keeping with ideas popular in mid XXth century (e.g., Lakatos, *Proofs and refutations*.)
- Alternative approaches (like the one advocated by the Bayesian school) give more clues on how to learn.

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.
- A **statistical hypothesis** is usually phrased in terms of the values of one or more parameters.

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.
- A **statistical hypothesis** is usually phrased in terms of the values of one or more parameters.
    1. The mean of a distribution is $m = 0$, (one parameter).

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.
- A **statistical hypothesis** is usually phrased in terms of the values of one or more parameters.
  1. The mean of a distribution is $m = 0$, (one parameter).
  2. Two distributions have the same mean: $m_1 = m_2$, (two parameters).

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.
- A **statistical hypothesis** is usually phrased in terms of the values of one or more parameters.
  1. The mean of a distribution is $m = 0$, (one parameter).
  2. Two distributions have the same mean: $m_1 = m_2$, (two parameters).
  3. Two characters are independent: $p_{ij} = p_{i.} \times p_{.j}$.

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.
- A **statistical hypothesis** is usually phrased in terms of the values of one or more parameters.
  1. The mean of a distribution is $m = 0$, (one parameter).
  2. Two distributions have the same mean: $m_1 = m_2$, (two parameters).
  3. Two characters are independent: $p_{ij} = p_{i.} \times p_{.j}$.
- Equivalently, a hypothesis is phrased by stating that a parameter vector belongs to a subset Θ_0 of the entire feasible space Θ.

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.
- A **statistical hypothesis** is usually phrased in terms of the values of one or more parameters.
  1. The mean of a distribution is $m = 0$, (one parameter).
  2. Two distributions have the same mean: $m_1 = m_2$, (two parameters).
  3. Two characters are independent: $p_{ij} = p_{i.} \times p_{.j}$.
- Equivalently, a hypothesis is phrased by stating that a parameter vector belongs to a subset Θ_0 of the entire feasible space Θ.

How would you phrase the hypothesis in items 1 and 2 above?

The anatomy of a hypothesis test (I)

- As already mentioned, a hypothesis is a conjecture.
- A **statistical hypothesis** is usually phrased in terms of the values of one or more parameters.
  1. The mean of a distribution is $m = 0$, (one parameter).
  2. Two distributions have the same mean: $m_1 = m_2$, (two parameters).
  3. Two characters are independent: $p_{ij} = p_{i.} \times p_{.j}$.
- Equivalently, a hypothesis is phrased by stating that a parameter vector belongs to a subset Θ_0 of the entire feasible space Θ.

1) $\Theta_0 = 0$, $\Theta = \mathcal{R}$. 2) $\Theta_0 = \{(x, y) : x = y\}$, $\Theta = \mathcal{R}^2$

The anatomy of a hypothesis test (II)

- In order to test the *null hypothesis* H_0, we use as evidence the information contained in a sample. We usually condense that information using a *test statistic*, $S = S(\vec{X})$.

The anatomy of a hypothesis test (II)

- In order to test the *null hypothesis* H_0, we use as evidence the information contained in a sample. We usually condense that information using a *test statistic*, $S = S(\vec{X})$.
- We better use a sufficient statistic!

The anatomy of a hypothesis test (II)

- In order to test the *null hypothesis* H_0, we use as evidence the information contained in a sample. We usually condense that information using a *test statistic*, $S = S(\vec{X})$.
- We better use a sufficient statistic!
- To be useful, that test statistic must have a known distribution under H_0. This is required, so that we can tell when a sampled value is “rare” under H_0.

The anatomy of a hypothesis test (II)

- In order to test the *null hypothesis* H_0, we use as evidence the information contained in a sample. We usually condense that information using a *test statistic*, $S = S(\vec{X})$.
- We better use a sufficient statistic!
- To be useful, that test statistic must have a known distribution under H_0. This is required, so that we can tell when a sampled value is “rare” under H_0.
- The decision procedure then is:

The anatomy of a hypothesis test (II)

- In order to test the *null hypothesis* H_0, we use as evidence the information contained in a sample. We usually condense that information using a *test statistic*, $S = S(\vec{X})$.
- We better use a sufficient statistic!
- To be useful, that test statistic must have a known distribution under H_0. This is required, so that we can tell when a sampled value is "rare" under H_0.
- The decision procedure then is:

  Reject H_0 ***if the sampled value of*** S ***is "rare", do not reject otherwise.***

The anatomy of a hypothesis test (II)

- In order to test the *null hypothesis* H_0, we use as evidence the information contained in a sample. We usually condense that information using a *test statistic*, $S = S(\vec{X})$.
- We better use a sufficient statistic!
- To be useful, that test statistic must have a known distribution under H_0. This is required, so that we can tell when a sampled value is “rare” under H_0.
- The decision procedure then is:
  Reject H_0 ***if the sampled value of*** S ***is “rare”, do not reject otherwise.***
- What is “rare”? Problem dependent.

The anatomy of a hypothesis test (III)

Example:

- We believe the mean of a $N(m, \sigma^2 = 1)$ distribution to be zero (H_0). A sample of $n = 100$ observations gives $\overline{X} = 0.20$.

The anatomy of a hypothesis test (III)

Example:

- We believe the mean of a $N(m, \sigma^2 = 1)$ distribution to be zero (H_0). A sample of $n = 100$ observations gives $\overline{X} = 0.20$.
- We are willing to reject the hypothesis if the evidence found is among the 5% “rarest” events that could happen under H_0. What will be our decission?

The anatomy of a hypothesis test (III)

Example:

- We believe the mean of a $N(m, \sigma^2 = 1)$ distribution to be zero (H_0). A sample of $n = 100$ observations gives $\overline{X} = 0.20$.
- We are willing to reject the hypothesis if the evidence found is among the 5% "rarest" events that could happen under H_0. What will be our decission?
- The events that we decide constitute evidence against H_0 is called the critical region.

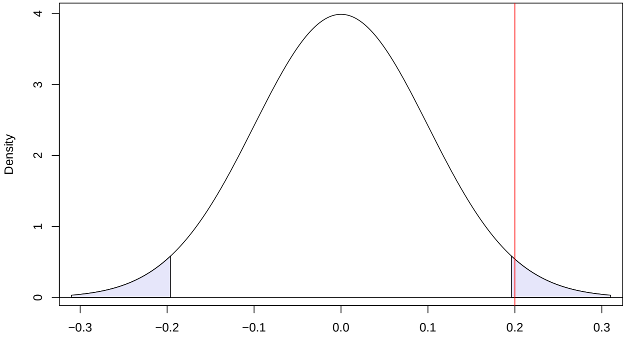
The anatomy of a hypothesis test (III)

Example:

- We believe the mean of a $N(m, \sigma^2 = 1)$ distribution to be zero (H_0). A sample of $n = 100$ observations gives $\overline{X} = 0.20$.
- We are willing to reject the hypothesis if the evidence found is among the 5% “rarest” events that could happen under H_0. What will be our decission?
- The events that we decide constitute evidence against H_0 is called the critical region.
- The probability of the critical region when H_0 is true, is called the significance level.

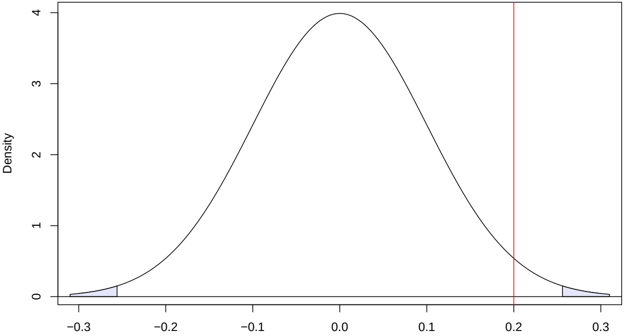
The anatomy of a hypothesis test (IV)

At the stated level of significance (5%), we would reject H_0.



The anatomy of a hypothesis test (V)

With a different level of significance (1%), we would **not** reject H_0.



The trade-off between Type I and Type II errors

- The significance level α is the probability of unduly rejecting H_0.

The trade-off between Type I and Type II errors

- The significance level α is the probability of unduly rejecting H_0.
- We should choose α considering how “grave” or “costly” is such an error, called *Type I error* or *size* of the test.

The trade-off between Type I and Type II errors

- The significance level α is the probability of unduly rejecting H_0.
- We should choose α considering how “grave” or “costly” is such an error, called *Type I error* or *size* of the test.
- If we make α very small (an hence the critical region very small also), we will almost never reject $H_0 \ldots$

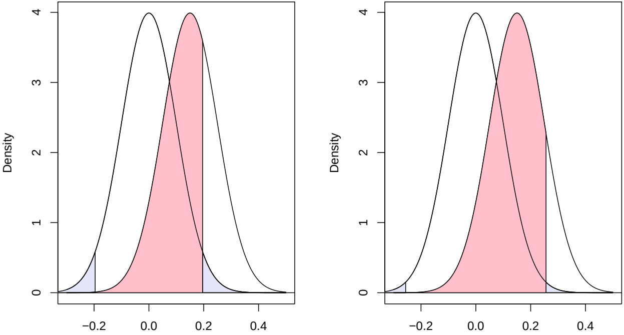
The trade-off between Type I and Type II errors

- The significance level α is the probability of unduly rejecting H_0.
- We should choose α considering how “grave” or “costly” is such an error, called *Type I error* or *size* of the test.
- If we make α very small (an hence the critical region very small also), we will almost never reject H_0 . . .
- . . .even when we would like to, because it is false!

The trade-off between Type I and Type II errors

- The significance level α is the probability of unduly rejecting H_0.
- We should choose α considering how “grave” or “costly” is such an error, called *Type I error* or *size* of the test.
- If we make α very small (an hence the critical region very small also), we will almost never reject H_0 . . .
- . . .even when we would like to, because it is false!
- Not rejecting H_0 when it is false is called *Type II* error, and its probability is denoted by β.

Trade-off between Type I and II errors - Ilustration



Pure significance tests

- We are only considering so far H_0.

Pure significance tests

- We are only considering so far H_0.
- We are looking at empirical evidence to see it it "contradicts" H_0.

Pure significance tests

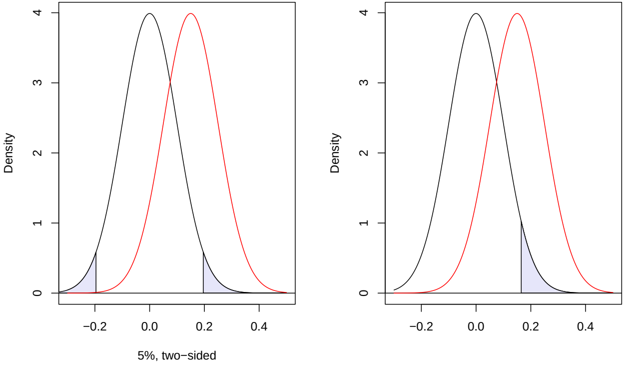
- We are only considering so far H_0.
- We are looking at empirical evidence to see it it "contradicts" H_0.
- When it does, we reject H_0.

Pure significance tests

- We are only considering so far H_0.
- We are looking at empirical evidence to see it it "contradicts" H_0.
- When it does, we reject H_0.
- Sometimes, we have a clear idea of what the "competing" hypothesis is, and in this case we want to use that information.

Testing against an alternative H_a

If we test H_0 against an alternative H_a, a one-sided critical region makes more sense.



Optimal critical regions for H_0 vs. H_a

The usual procedure is:

- Fix α, the probability of unduly rejecting H_0.

Optimal critical regions for H_0 vs. H_a

The usual procedure is:

- Fix α, the probability of unduly rejecting H_0.
- Among all critical regions of size α, find the one which minimizes β (or, equivalently, maximizes $1 - \beta$, the *power*).

Optimal critical regions for H_0 vs. H_a

The usual procedure is:

- Fix α, the probability of unduly rejecting H_0.
- Among all critical regions of size α, find the one which minimizes β (or, equivalently, maximizes $1 - \beta$, the *power*).
- When both H_0 and H_a are *simple* (= fix completely the distribution of the test statistic), a simple procedure exists, base on Neyman-Pearson's theorem.

Optimal critical regions for H_0 vs. H_a

The usual procedure is:

- Fix α, the probability of unduly rejecting H_0.
- Among all critical regions of size α, find the one which minimizes β (or, equivalently, maximizes $1 - \beta$, the *power*).
- When both H_0 and H_a are *simple* (= fix completely the distribution of the test statistic), a simple procedure exists, base on Neyman-Pearson's theorem.
- In other cases, a unique most powerful test may not exist.

The Neyman-Pearson theorem (I)

- After fixing the significance level α, what critical region would give better power against a simple alternative?

The Neyman-Pearson theorem (I)

- After fixing the significance level α, what critical region would give better power against a simple alternative?
- Let's consider testing $H_0 : \theta = \theta_0$ vs. $H_a : \theta = \theta_a$:

x	0	1	2	3	4	5
$P(x; \theta_0)$	0.60	0.26	0.05	0.04	0.04	0.01
$P(x; \theta_a)$	0.10	0.15	0.10	0.25	0.30	0.10

The Neyman-Pearson theorem (I)

- After fixing the significance level α, what critical region would give better power against a simple alternative?
- Let's consider testing $H_0 : \theta = \theta_0$ vs. $H_a : \theta = \theta_a$:

x	0	1	2	3	4	5
$P(x; \theta_0)$	0.60	0.26	0.05	0.04	0.04	0.01
$P(x; \theta_a)$	0.10	0.15	0.10	0.25	0.30	0.10

How would you choose a critical region of size $\alpha = 0.05$ with maximum power?

The Neyman-Pearson theorem (I)

- After fixing the significance level α, what critical region would give better power against a simple alternative?
- Let's consider testing $H_0 : \theta = \theta_0$ vs. $H_a : \theta = \theta_a$:

x	0	1	2	3	4	5
$P(x;\theta_0)$	0.60	0.26	0.05	0.04	0.04	0.01
$P(x;\theta_a)$	0.10	0.15	0.10	0.25	0.30	0.10

Picking $x = 4$ and $x = 5$, for a total power of 0.40.

The Neyman-Pearson theorem (II)

- The intuition is that we want our critical region to be made of points x with high ratio

  $$\frac{f(x;\theta_a)}{f(x;\theta_0)}$$

  where $f(x;\theta_0)$ is the density under the null and $f(x;\theta_a)$ is the density under the alternative.

The Neyman-Pearson theorem (II)

- The intuition is that we want our critical region to be made of points x with high ratio

$$\frac{f(x;\theta_a)}{f(x;\theta_0)}$$

  where $f(x;\theta_0)$ is the density under the null and $f(x;\theta_a)$ is the density under the alternative.
- Neyman-Pearson theorem: *The most powerful test of given size α for $H_0 : \theta = \theta_0$ against the alternative $H_a : \theta = \theta_a$ has critical region of the form:*

$$C_\alpha = \left\{ \vec{x} : \frac{f(\vec{x};\theta_a)}{f(\vec{x};\theta_0)} > k_\alpha \right\}$$

  for a constant k_α which depends on α.

The Neyman-Pearson theorem - Proof (I)

- Consider the critical region

$$C_\alpha = \left\{ \vec{x} : \frac{f(\vec{x};\theta_a)}{f(\vec{x};\theta_0)} > k_\alpha \right\}$$

  and any other α-size region A_α.

The Neyman-Pearson theorem - Proof (I)

- Consider the critical region

  $$C_\alpha = \left\{ \vec{x} : \frac{f(\vec{x};\theta_a)}{f(\vec{x};\theta_0)} > k_\alpha \right\}$$

  and any other α-size region A_α.
- C_α and A_α will in general overlap. Dropping the α subscript:

  $$\int_C f(\vec{x};\theta_0) d\vec{x} = \int_A f(\vec{x};\theta_0) d\vec{x} = \alpha$$

The Neyman-Pearson theorem - Proof (I)

- Consider the critical region

$$C_\alpha = \left\{ \vec{x} : \frac{f(\vec{x};\theta_a)}{f(\vec{x};\theta_0)} > k_\alpha \right\}$$

  and any other α-size region A_α.

- C_α and A_α will in general overlap. Dropping the α subscript:

$$\int_C f(\vec{x};\theta_0)d\vec{x} = \int_A f(\vec{x};\theta_0)d\vec{x} = \alpha$$

- Subtracting $\delta = \int_{C\cap A} f(\vec{x};\theta_0)d\vec{x}$ in both sides:

$$\int_{C\cap A^c} f(\vec{x};\theta_0)d\vec{x} = \int_{A\cap C^c} f(\vec{x};\theta_0)d\vec{x} = \alpha - \delta \geq 0$$

The Neyman-Pearson theorem - Proof (I)

- Consider the critical region

$$C_\alpha = \left\{ \vec{x} : \frac{f(\vec{x};\theta_a)}{f(\vec{x};\theta_0)} > k_\alpha \right\}$$

  and any other α-size region A_α.
- C_α and A_α will in general overlap. Dropping the α subscript:

$$\int_C f(\vec{x};\theta_0) d\vec{x} = \int_A f(\vec{x};\theta_0) d\vec{x} = \alpha$$

- Subtracting $\delta = \int_{C\cap A} f(\vec{x};\theta_0) d\vec{x}$ in both sides:

$$\int_{C\cap A^c} f(\vec{x};\theta_0) d\vec{x} = \int_{A\cap C^c} f(\vec{x};\theta_0) d\vec{x} = \alpha - \delta \geq 0$$

How do we know $\alpha - \delta \geq 0$?

The Neyman-Pearson theorem - Proof (I)

- Consider the critical region

$$C_\alpha = \left\{ \vec{x} : \frac{f(\vec{x};\theta_a)}{f(\vec{x};\theta_0)} > k_\alpha \right\}$$

  and any other α-size region A_α.
- C_α and A_α will in general overlap. Dropping the α subscript:

$$\int_C f(\vec{x};\theta_0) d\vec{x} = \int_A f(\vec{x};\theta_0) d\vec{x} = \alpha$$

- Subtracting $\delta = \int_{C\cap A} f(\vec{x};\theta_0) d\vec{x}$ in both sides:

$$\int_{C\cap A^c} f(\vec{x};\theta_0) d\vec{x} = \int_{A\cap C^c} f(\vec{x};\theta_0) d\vec{x} = \alpha - \delta \geq 0$$

Because $C \cap A \subseteq C$.

The Neyman-Pearson theorem - Proof (II)

- The difference of powers of the two critical regions is:

$$\int_C f(\vec{x};\theta_a)d\vec{x} - \int_A f(\vec{x};\theta_a)d\vec{x}$$

The Neyman-Pearson theorem - Proof (II)

- The difference of powers of the two critical regions is:

$$\int_C f(\vec{x};\theta_a)d\vec{x} - \int_A f(\vec{x};\theta_a)d\vec{x}$$

- Inside C we have $f(\vec{x};\theta_a) > kf(\vec{x};\theta_0)$ and outside $f(\vec{x};\theta_a) \leq kf(\vec{x};\theta_0)$. The difference of powers is:

$$\begin{aligned}
\int_C f(\vec{x};\theta_a)d\vec{x} &- \int_A f(\vec{x};\theta_a)d\vec{x} \\
&= \int_{C\cap A^c} f(\vec{x};\theta_a)d\vec{x} - \int_{A\cap C^c} f(\vec{x};\theta_a)d\vec{x} \\
&\geq k\int_{C\cap A^c} f(\vec{x};\theta_0)d\vec{x} - k\int_{A\cap C^c} f(\vec{x};\theta_0)d\vec{x} \\
&= k(\alpha-\delta) - k(\alpha-\delta) = 0
\end{aligned}$$

Neyman-Pearson example (I)

- In a large company, the number of workers not showing up for work is Poisson-distributed. Workers claim that $\lambda = 1$, while management claims $\lambda = 2$. They check four days and obtain 1, 0, 2, and 2 workers not showing up for work.

Neyman-Pearson example (I)

- In a large company, the number of workers not showing up for work is Poisson-distributed. Workers claim that $\lambda = 1$, while management claims $\lambda = 2$. They check four days and obtain 1, 0, 2, and 2 workers not showing up for work.
  1. Obtain the most powerful critical region to test the workers hypothesis (H_0) against the management's at a 0.05 significance level.

Neyman-Pearson example (I)

- In a large company, the number of workers not showing up for work is Poisson-distributed. Workers claim that $\lambda = 1$, while management claims $\lambda = 2$. They check four days and obtain 1, 0, 2, and 2 workers not showing up for work.
  1. Obtain the most powerful critical region to test the workers hypothesis (H_0) against the management's at a 0.05 significance level.
  2. What is the power of the test?

Neyman-Pearson example (I)

- In a large company, the number of workers not showing up for work is Poisson-distributed. Workers claim that $\lambda = 1$, while management claims $\lambda = 2$. They check four days and obtain 1, 0, 2, and 2 workers not showing up for work.
  1. Obtain the most powerful critical region to test the workers hypothesis (H_0) against the management's at a 0.05 significance level.
  2. What is the power of the test?
- We have:

$$\begin{aligned}
f(\vec{x}; \lambda = 1) &= \prod_{i=1}^{4} \frac{e^{-1} 1^{x_i}}{x_i!} = \frac{e^{-4}}{\prod_{i=1}^{4} x_i!} \\
f(\vec{x}; \lambda = 2) &= \prod_{i=1}^{4} \frac{e^{-2} 2^{x_i}}{x_i!} = \frac{e^{-8} 2^{\sum_{i=1}^{4} x_i}}{\prod_{i=1}^{4} x_i!}
\end{aligned}$$

Neyman-Pearson example (II)

- From Neyman-Pearson, the most powerful critical region of size α is of the form:

$$\begin{aligned} C_\alpha &= \left\{\vec{x} : \frac{f(\vec{x}; \lambda = 2)}{f(\vec{x}; \lambda = 1)} > k_\alpha\right\} \\ &= \left\{\vec{x} : \frac{e^{-8} 2^{\sum_{i=1}^{4} x_i}}{e^{-4}}\right\} \\ &= \left\{\vec{x} : e^{-4} 2^{\sum_{i=1}^{4} x_i} > k_\alpha\right\} \end{aligned}$$

Neyman-Pearson example (II)

- From Neyman-Pearson, the most powerful critical region of size α is of the form:

$$\begin{aligned} C_\alpha &= \left\{ \vec{x} : \frac{f(\vec{x}; \lambda = 2)}{f(\vec{x}; \lambda = 1)} > k_\alpha \right\} \\ &= \left\{ \vec{x} : \frac{e^{-8} 2^{\sum_{i=1}^{4} x_i}}{e^{-4}} \right\} \\ &= \left\{ \vec{x} : e^{-4} 2^{\sum_{i=1}^{4} x_i} > k_\alpha \right\} \end{aligned}$$

- Taking logs and bringing all constants into k'_α:

$$C_\alpha = \left\{ \vec{x} : \sum_{i=1}^{4} x_i > k'_\alpha \right\}$$

Neyman-Pearson example (III)

- We now know **the form** of C_α

$$C_\alpha = \left\{ \vec{x} : \sum_{i=1}^{4} x_i > k'_\alpha \right\}$$

- Have no clue about what the value of k'_α is, but know $\sum_{i=1}^{4} x_i \sim \mathcal{P}(\lambda = 4)$ when H_0 is true.
- For C_α to have size $\alpha = 0.05$, the constant must be a value exceeded with probability no greater than α when sampling a $\mathcal{P}(\lambda = 4)$ distribution. Resorting to tables (or R) gives us:

```
> ppois(0:8,lambda=4)
[1] 0.01832 0.09158 0.23810 0.43347 0.62884
[6] 0.78513 0.88933 0.94887 0.97864
```

- $[8, \infty)$ would be a critical region for $S = \sum_{i=1}^{4} x_i$ quite close to $\alpha = 0.05$; $[9, \infty)$ would have $\alpha = 0.02136$.

Neyman-Pearson and sufficiency (I)

- Do we loose something by using Neyman-Pearson's theorem?
- We decide between H_0 and H_a as if only the likelihood ratio (LR) matters.
- Might be justified if the LR were a sufficient statistic.
- It is! In a sense, it is the "smallest" sufficient statistic.
- We prove a simplified version next.

Neyman-Pearson and sufficiency (II)

- Consider the simple case where $\Theta = \{\theta_0, \theta_1\}$ and both distributions $F_X(x;\theta)$ have commen support.
- The, the likelihood ratio

$$R(\vec{x}) = \frac{f_{\vec{X}}(\vec{x};\theta_0)}{f_{\vec{X}}(\vec{x};\theta_1)}$$

  is a sufficient statistic.
- To prove sufficiency we have to show that

$$f_{\vec{X}}(\vec{x}|R(\vec{x}) = r;\theta_0) = f_{\vec{X}}(\vec{x}|R(\vec{x}) = r;\theta_1)$$

Neyman-Pearson and sufficiency (III)

$$\begin{aligned}
f_{\vec{X}}(\vec{x}|R(\vec{x}) = r; \theta_0) &= \frac{f_{\vec{X}}(\vec{x}; \theta_0)}{\int_{R(\vec{X})=r} f_{\vec{X}}(\vec{x}; \theta_0) d\vec{x}} = \frac{r f_{\vec{X}}(\vec{x}; \theta_1)}{\int_{R(\vec{X})=r} r f_{\vec{X}}(\vec{x}; \theta_1) d\vec{x}} \\
&= \frac{f_{\vec{X}}(\vec{x}; \theta_1)}{\int_{R(\vec{X})=r} f_{\vec{X}}(\vec{x}; \theta_1) d\vec{x}} = f_{\vec{X}}(\vec{x}|R(\vec{x}) = r; \theta_1)
\end{aligned}$$

Some quirks of hypothesis testing (I)

- Very non symmetric role of null and alternative hypothesis.

Some quirks of hypothesis testing (I)

- Very non symmetric role of null and alternative hypothesis.
- Management could have replied the worker's representative: "Why don't we test as null *our hypothesis* and not yours?

Some quirks of hypothesis testing (I)

- Very non symmetric role of null and alternative hypothesis.
- Management could have replied the worker's representative: "Why don't we test as null *our hypothesis* and not yours?
- If evidence is not strong, the null is the surviving hypothesis, whichever it happens to be!

Some quirks of hypothesis testing (I)

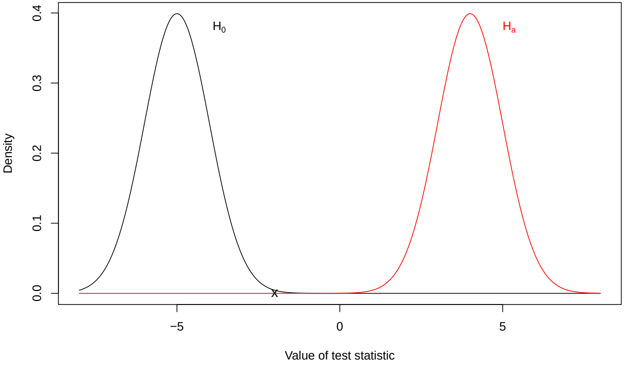
- Very non symmetric role of null and alternative hypothesis.
- Management could have replied the worker's representative: "Why don't we test as null *our hypothesis* and not yours?
- If evidence is not strong, the null is the surviving hypothesis, whichever it happens to be!
- The null should be provisionally established knowledge, put to test. How we arrive to that knowledge, there is no telling.

Some quirks of hypothesis testing (I)

- Very non symmetric role of null and alternative hypothesis.
- Management could have replied the worker's representative: "Why don't we test as null *our hypothesis* and not yours?
- If evidence is not strong, the null is the surviving hypothesis, whichever it happens to be!
- The null should be provisionally established knowledge, put to test. How we arrive to that knowledge, there is no telling.
- Alternative approaches (like bayesian inference) treat conjectures in a more symmetric way.

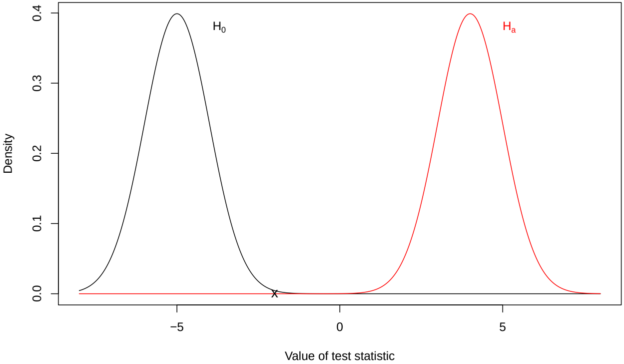
Some quirks of hypothesis testing (II)

- That H_0 is rejected **does not mean that H_a should be accepted.**



Some quirks of hypothesis testing (II)

- That H_0 is rejected **does not mean that H_a should be accepted.**



- An observation at X is evidence against H_0 but much more so against H_a. In such situation, we should revise our hypothesis and admit that other possibilities might exist.

Some quirks of hypothesis testing (III)

- In (classical, frequentist) statistics we start the world anew each morning!

Some quirks of hypothesis testing (III)

- In (classical, frequentist) statistics we start the world anew each morning!
- Doesn't make much sense at times: we have previous information.

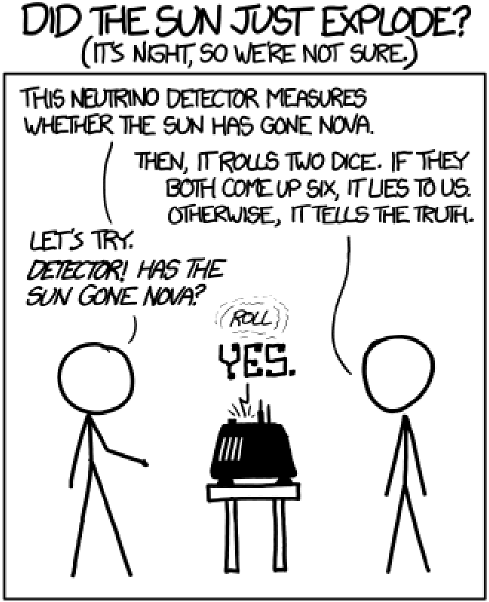
Some quirks of hypothesis testing (III)

- In (classical, frequentist) statistics we start the world anew each morning!
- Doesn't make much sense at times: we have previous information.
- Sometimes, we are so convinced of H_0 that even a very rare result under H_0 will not persuade us to abandon it.

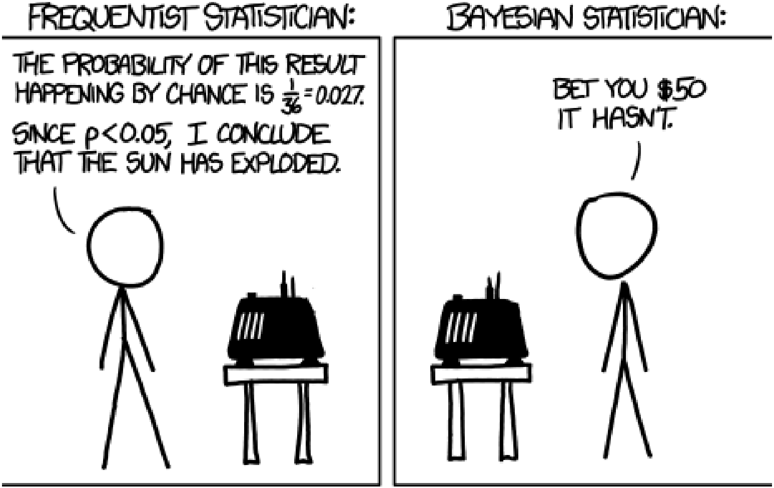
Some quirks of hypothesis testing (III)

- In (classical, frequentist) statistics we start the world anew each morning!
- Doesn't make much sense at times: we have previous information.
- Sometimes, we are so convinced of H_0 that even a very rare result under H_0 will not persuade us to abandon it.
- Under the frequentist approach, no way to deal with this.

Some quirks of hypothesis testing (III)



Some quirks of hypothesis testing (IV)



`https://xkcd.com/1132/`

Goodness-of fit problems

- Quite common hypothesis.

Goodness-of fit problems

- Quite common hypothesis.
  1. Do winning numbers in the Lotería Primitiva appear to come from a discreet uniform distribution over $\{1,2,\ldots,49\}$? *(no parameters estimated, fully especified distribution)*

Goodness-of fit problems

- Quite common hypothesis.
  1. Do winning numbers in the Lotería Primitiva appear to come from a discreet uniform distribution over {1,2,. . . ,49} ? *(no parameters estimated, fully especified distribution)*
  2. Does the number of dead people by horse (or mule) kick in the Prusian army follow a Poisson distribution (plausible; small probability, many people at risk). *(one parameter to be estimated)*

Goodness-of fit problems

- Quite common hypothesis.
  1. Do winning numbers in the Lotería Primitiva appear to come from a discreet uniform distribution over {1,2,. . . ,49} ? *(no parameters estimated, fully especified distribution)*
  2. Does the number of dead people by horse (or mule) kick in the Prusian army follow a Poisson distribution (plausible; small probability, many people at risk). *(one parameter to be estimated)*
  3. Do intervals between accidents at work appear to follow an exponential distribution? *(one parameter to be estimated)*

Goodness-of fit problems

- Quite common hypothesis.
  1. Do winning numbers in the Lotería Primitiva appear to come from a discreet uniform distribution over {1,2,. . . ,49} ? *(no parameters estimated, fully especified distribution)*
  2. Does the number of dead people by horse (or mule) kick in the Prusian army follow a Poisson distribution (plausible; small probability, many people at risk). *(one parameter to be estimated)*
  3. Do intervals between accidents at work appear to follow an exponential distribution? *(one parameter to be estimated)*
- In all these cases, we have data and we want to test adequacy of a given distribution, possibly not fully especified (= some parameter has to be estimated).

Test statistic

- Break down the range of the random variable in k classes. Call O_i the number of observations in class i, $i = 1, 2, \ldots, k$.

Test statistic

- Break down the range of the random variable in k classes. Call O_i the number of observations in class i, $i = 1, 2, \ldots, k$.
- Call E_i the number of expected observations in class i under the null hypothesis (i.e., if the assumed distribution for the data is “true”).

Test statistic

- Break down the range of the random variable in k classes. Call O_i the number of observations in class i, $i = 1, 2, \ldots, k$.
- Call E_i the number of expected observations in class i under the null hypothesis (i.e., if the assumed distribution for the data is “true”).
- Then,

$$Z = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \overset{H_o}{\sim} \quad \chi^2_{k-p-1}$$

Test statistic

- Break down the range of the random variable in k classes. Call O_i the number of observations in class i, $i = 1, 2, \ldots, k$.
- Call E_i the number of expected observations in class i under the null hypothesis (i.e., if the assumed distribution for the data is “true”).
- Then,

$$Z = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \quad \overset{H_o}{\sim} \quad \chi^2_{k-p-1}$$

- k is the number of classes, p the number of parameter estimated, if any.

The gory details

- Where does this come from? Proof not trivial, distribution valid only as an approximation for “large” samples.

The gory details

- Where does this come from? Proof not trivial, distribution valid only as an approximation for “large” samples.
- How large is “large”? No class should have an expected value less than, say, 5. If it does, merge classes.

The gory details

- Where does this come from? Proof not trivial, distribution valid only as an approximation for “large” samples.
- How large is “large”? No class should have an expected value less than, say, 5. If it does, merge classes.
- How to choose k? Reasonably large, but keeping classes “well peopled”.

The gory details

- Where does this come from? Proof not trivial, distribution valid only as an approximation for “large” samples.
- How large is “large”? No class should have an expected value less than, say, 5. If it does, merge classes.
- How to choose k? Reasonably large, but keeping classes “well peopled”.
- Howto choose the class boundaries? Good question.

The gory details

- Where does this come from? Proof not trivial, distribution valid only as an approximation for “large” samples.
- How large is “large”? No class should have an expected value less than, say, 5. If it does, merge classes.
- How to choose k? Reasonably large, but keeping classes “well peopled”.
- Howto choose the class boundaries? Good question.
- Usually no particular alternative: a pure significance test.

The gory details

- Where does this come from? Proof not trivial, distribution valid only as an approximation for “large” samples.
- How large is “large”? No class should have an expected value less than, say, 5. If it does, merge classes.
- How to choose k? Reasonably large, but keeping classes “well peopled”.
- Howto choose the class boundaries? Good question.
- Usually no particular alternative: a pure significance test.
- Critical region: right tail.

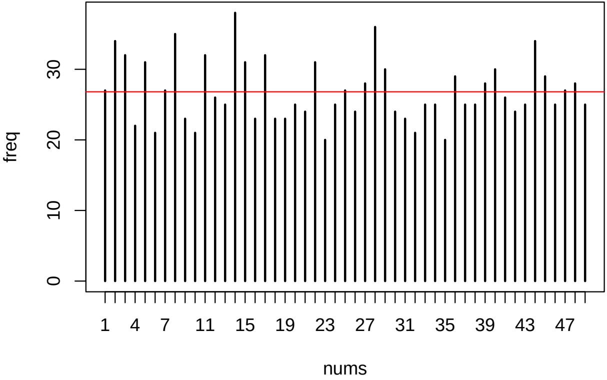
Example - I

```
> primitiva[1:3,1:8]
       Fecha Semana N1 N2 N3 N4 N5 N6
1 01/01/2009      1  4  8 12 25 34 46
2 03/01/2009      1  9 11 21 30 31 44
3 08/01/2009      2  7 17 27 28 29 44
> nums <- as.matrix(primitiva[,3:8])
> freq <- table(nums)
> sum(freq)            # How many numbers seen?
[1] 1314
> e <- sum(freq) / 49  # Expected each under H0
> e
[1] 26.82
```

Example -II

The absolute frequencies of each number are:

```
> plot(freq)
> abline(h=e,col="red")
```



Example -III

- Question is now to decide whether the departures from the expected number of appearances is enough to reject H_0 (=”all numbers equally likely”).

Example - IV

- R has a standard function which does the same at once.

```
>  result <- chisq.test(x=freq,p=rep(1/49,49))
>  result
        Chi-squared test for given probabiliti

data:  freq
X-squared = 34, df = 48, p-value = 0.9
```

Example - V

- If you have to do it manually, your best bet is to arrange computations in a small table.

Example - V

- If you have to do it manually, your best bet is to arrange computations in a small table.
- For instance, you might have in the case shown:

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2/E_i$
27	26.81633	0.183673	0.03373	0.001258
34	26.81633	7.183673	51.60516	1.924394
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
25	26.81633	-1.816327	3.29904	0.123023
			$Z =$	33.6865

Chi square test with estimated parameters (I)

- Data: deaths by horse kick in 200 army corps years.

DEATHS	OBSERVED CASES
0	109
1	65
2	22
3	3
4	1

Chi square test with estimated parameters (I)

- Data: deaths by horse kick in 200 army corps years.

DEATHS	OBSERVED CASES
0	109
1	65
2	22
3	3
4	1

- Is the Poisson distribution a good model for these data?

Chi square test with estimated parameters (I)

- Data: deaths by horse kick in 200 army corps years.

DEATHS	OBSERVED CASES
0	109
1	65
2	22
3	3
4	1

- Is the Poisson distribution a good model for these data?
- The hypothesis does not uniquely fix the distribution.

Chi square test with estimated parameters (I)

- Data: deaths by horse kick in 200 army corps years.

DEATHS	OBSERVED CASES
0	109
1	65
2	22
3	3
4	1

- Is the Poisson distribution a good model for these data?
- The hypothesis does not uniquely fix the distribution.
- The MLE of λ is:

$$\hat{\lambda} = \frac{0 \times 109 + 1 \times 65 + 2 \times 22 + 3 \times 3 + 4 \times 1}{200} = 0.61$$

Chi square test with estimated parameters (II)

- We have estimated one parameter (λ).

Chi square test with estimated parameters (II)

- We have estimated one parameter (λ).
- We do as before, only the E_i are compute from the $\mathcal{P}(\lambda = 0.61)$ distribution.

Chi square test with estimated parameters (II)

- We have estimated one parameter (λ).
- We do as before, only the E_i are compute from the $\mathcal{P}(\lambda = 0.61)$ distribution.
- For instance, since

$$P(x = 0; \lambda = 0.61) = \frac{e^{-0.61}(0.61)^0}{0!} = 0.5433509$$

we would compute E_1 (the expected number of cases with 0 deaths) as: $200 \times 0.5433509 = 108.67$.

Chi square test with estimated parameters (II)

- We have estimated one parameter (λ).
- We do as before, only the E_i are compute from the $\mathcal{P}(\lambda = 0.61)$ distribution.
- For instance, since

  $$P(x = 0; \lambda = 0.61) = \frac{e^{-0.61}(0.61)^0}{0!} = 0.5433509$$

  we would compute E_1 (the expected number of cases with 0 deaths) as: $200 \times 0.5433509 = 108.67$.
- Likewise for the remaining E_i cells.

Chi square test with estimated parameters (III)

- Now we have:

O_i	E_i	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
109	108.67017	0.32983	0.10879	0.00100
65	66.28881	-1.28881	1.66102	0.02505
22	20.21809	1.78191	3.17522	0.15704
3	4.11101	-1.11101	1.23434	0.30025
1	0.62693	0.37307	0.13918	0.22200
			$Z =$	0.70537

Chi square test with estimated parameters (IV)

- We now compare Z with a chi-square with **3** degrees of freedom ($k - p - 1 = 5 - 1 - 1 = 3$):

  ```
  > 1 - pchisq(0.70537,df=3)
  [1] 0.8719
  ```

Contingency table analysis

- A two-dimensional contingency table is an array which classifies observations according to two variables, one occurring in rows, the other in columns.

Contingency table analysis

- A two-dimensional contingency table is an array which classifies observations according to two variables, one occurring in rows, the other in columns.
- Definition can be generalized to different number of dimensions

Contingency table analysis

- A two-dimensional contingency table is an array which classifies observations according to two variables, one occurring in rows, the other in columns.
- Definition can be generalized to different number of dimensions
- For example, we may have:

Gender	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

Contingency table analysis

- A two-dimensional contingency table is an array which classifies observations according to two variables, one occurring in rows, the other in columns.
- Definition can be generalized to different number of dimensions
- For example, we may have:

Gender	Right-handed	Left-handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

- The row and column totals are referred as the *margins.*

Sampling schemes (I)

- We may fix only the total number of cases we cross-tabulate. . .

Sampling schemes (I)

- We may fix only the total number of cases we cross-tabulate. . .
- . . .or we may fix the row margin or the column margin.

Sampling schemes (I)

- We may fix only the total number of cases we cross-tabulate. . .
- . . .or we may fix the row margin or the column margin.
- In the first case we speak of *multinomial sampling*, in the second of *product multinomial sampling.*

Sampling schemes (I)

- We may fix only the total number of cases we cross-tabulate…
- …or we may fix the row margin or the column margin.
- In the first case we speak of *multinomial sampling*, in the second of *product multinomial sampling.*
- Why should we care? Marginal probabilities can only be estimated from “free” margins.

Sampling schemes (II)

- Consider the following case: we pick a sample of 1000 persons and cross-classify them according to ethnic origin and whether they suffered in the last winter from common cold. Want to test relative vulnerability.

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

Sampling schemes (II)

- Consider the following case: we pick a sample of 1000 persons and cross-classify them according to ethnic origin and whether they suffered in the last winter from common cold. Want to test relative vulnerability.

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

- We may estimate the proportion of whites as 905/1000 = 0.905 and the overall prevalence of cold as 0.884

Sampling schemes (III)

- Suppose though we are sampling a population with a tiny proportion of non-whites. We might end up with a table such as:

Race	Had cold	Didn't have cold	Total
Whites	891	108	999
Non-whites	1	0	1
Total	892	108	1000

Sampling schemes (III)

- Suppose though we are sampling a population with a tiny proportion of non-whites. We might end up with a table such as:

Race	Had cold	Didn't have cold	Total
Whites	891	108	999
Non-whites	1	0	1
Total	892	108	1000

- We end up with a table in which non-whites are almost (or totally) absent.

Sampling schemes (III)

- Suppose though we are sampling a population with a tiny proportion of non-whites. We might end up with a table such as:

Race	Had cold	Didn't have cold	Total
Whites	891	108	999
Non-whites	1	0	1
Total	892	108	1000

- We end up with a table in which non-whites are almost (or totally) absent.
- Non-white sample far too small to investigate the matter of interest.

Sampling schemes (IV)

- What we need instead is to sample both races separately, say 500 each:

Race	Had cold	Didn't have cold	Total
Whites	398	102	500
Non-whites	403	97	500
Total	801	199	1000

Sampling schemes (IV)

- What we need instead is to sample both races separately, say 500 each:

Race	Had cold	Didn't have cold	Total
Whites	398	102	500
Non-whites	403	97	500
Total	801	199	1000

- *Then* we are assured to have enough observations in each group.

Sampling schemes (IV)

- What we need instead is to sample both races separately, say 500 each:

Race	Had cold	Didn't have cold	Total
Whites	398	102	500
Non-whites	403	97	500
Total	801	199	1000

- *Then* we are assured to have enough observations in each group.
- Marginal totals do not estimate anything now: the row totals are fixed by design.

Sampling schemes (V)

- If we fix only the total, we are sampling **one** population. The hypothesis of interest is *independence* in that population.

Sampling schemes (V)

- If we fix only the total, we are sampling **one** population. The hypothesis of interest is *independence* in that population.
- If we fix the row totals, we are in effect sampling **two** populations. The hypothesis of interest is *homogeneity* of both populations with respect to the character coded in columns.

Sampling schemes (V)

- If we fix only the total, we are sampling **one** population. The hypothesis of interest is *independence* in that population.
- If we fix the row totals, we are in effect sampling **two** populations. The hypothesis of interest is *homogeneity* of both populations with respect to the character coded in columns.
- Both hypothesis are tested conditional on the margins, and the results are exactly the same for a given table, no matter how it was sampled.

Sampling schemes (V)

- If we fix only the total, we are sampling **one** population. The hypothesis of interest is *independence* in that population.
- If we fix the row totals, we are in effect sampling **two** populations. The hypothesis of interest is *homogeneity* of both populations with respect to the character coded in columns.
- Both hypothesis are tested conditional on the margins, and the results are exactly the same for a given table, no matter how it was sampled.
- Why conditionally on the margins? It is the distribution of counts inside the table what is indicative of independence (or homogeneity), *not* how many people of each race we look at.

Testing independence (I)

- Consider,

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

and assume it was obtained fixing only $N = 1000$.

Testing independence (I)

- Consider,

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

  and assume it was obtained fixing only $N = 1000$.
- The hypothesis of interest is $H_0 : p_{ij} = p_{i.} \times p_{.j}$

Testing independence (I)

- Consider,

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

  and assume it was obtained fixing only $N = 1000$.
- The hypothesis of interest is $H_0 : p_{ij} = p_{i.} \times p_{.j}$
- $\hat{p}_{11} = 0.884 \times 0.905$, and $E_{11} = 1000 \times 0.884 \times 0.905$. Similarly for the rest.

Testing independence (II)

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

- Apparently, we estimate 4 parameters p_{ij} for the 4 cells.

Testing independence (II)

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

- Apparently, we estimate 4 parameters p_{ij} for the 4 cells.
- Conditionally on the margins, only two parameters are free, and need to be counted.

Testing independence (III)

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
801	800.02	0.98	0.9604	0.00120
83	83.98	-0.98	0.9604	0.01144
104	104.98	-0.98	0.9604	0.00915
12	11.02	0.98	0.9604	0.08715
			$Z =$	0.10894

- The expected values are computed as $E_{ij} = Np_{ij} = Np_{i.}p_{.j}$.

Testing independence (III)

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
801	800.02	0.98	0.9604	0.00120
83	83.98	-0.98	0.9604	0.01144
104	104.98	-0.98	0.9604	0.00915
12	11.02	0.98	0.9604	0.08715
			$Z =$	0.10894

- The expected values are computed as $E_{ij} = Np_{ij} = Np_{i.}p_{.j}$.
- For instance, $800.02 = 1000 \times 0.884 \times 0.905$.

Testing independence (III)

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})$	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
801	800.02	0.98	0.9604	0.00120
83	83.98	-0.98	0.9604	0.01144
104	104.98	-0.98	0.9604	0.00915
12	11.02	0.98	0.9604	0.08715
			$Z =$	0.10894

- The expected values are computed as $E_{ij} = Np_{ij} = Np_{i.}p_{.j}$.
- For instance, $800.02 = 1000 \times 0.884 \times 0.905$.
- Degrees of freedom are $k - p - 1 = 4 - 2 - 1 = 1$. So we have to compare 0.10894 with the quantiles of a χ^2_1 distribution.

Testing independence (IV)

- We can easily construct the table:

```
> ColdRace <- matrix(c(801,83,104,12),2,2)
> ColdRace <- as.table(ColdRace)
> colnames(ColdRace) <-  c("Cold","Not-Cold")
> rownames(ColdRace) <-  c("Whites","Non-white
> ColdRace
          Cold Not-Cold
Whites     801      104
Non-whites  83       12
```

Testing independence (V)

Function `loglin` fits, among many other things, the independence model:

```
> result   <- loglin(ColdRace,margin=list(1,2),fit
2 iterations: deviation 0
> result$pearson
[1] 0.1089
> result$df
[1] 1
```

Testing independence (VI)

- The E_{ij} approach quite well O_{ij}:

```
> result$fit
                 Cold Not-Cold
Whites         800.02   104.98
Non-whites      83.98    11.02
```

Testing homogeneity (I)

- Consider again,

Observed counts ($= O_i$)

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

but this time assuming we have fixed the row marginal.

Testing homogeneity (I)

- Consider again,

Observed counts ($= O_i$)

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

but this time assuming we have fixed the row marginal.

- We are testing the hypothesis $H_0 : p_{1j} = p_{2j}$ for all j.

Testing homogeneity (I)

- Consider again,

Observed counts ($= O_i$)

Race	Had cold	Didn't have cold	Total
Whites	801	104	905
Non-whites	83	12	95
Total	884	116	1000

but this time assuming we have fixed the row marginal.

- We are testing the hypothesis $H_0 : p_{1j} = p_{2j}$ for all j.
- Under H_0, $\hat{p}_{.j} = n_{.j}/n_{..}$ is a sensible estimate of $p_{.j}$, common to all i.

Testing homogeneity (II)

- The results are exactly the same, only they are arrived at in a different manner.

Expected counts ($= E_i$)

Race	Had cold	Didn't have cold	Total
Whites	800.02	104.98	905
Non-whites	83.98	11.02	95
Total	884	116	1000

Testing homogeneity (II)

- The results are exactly the same, only they are arrived at in a different manner.

Expected counts ($= E_i$)

Race	Had cold	Didn't have cold	Total
Whites	800.02	104.98	905
Non-whites	83.98	11.02	95
Total	884	116	1000

- The E_{1j} in the first row are computed as $905 \times \hat{p}_{.j}$

Testing homogeneity (II)

- The results are exactly the same, only they are arrived at in a different manner.

Expected counts ($= E_i$)

Race	Had cold	Didn't have cold	Total
Whites	800.02	104.98	905
Non-whites	83.98	11.02	95
Total	884	116	1000

- The E_{1j} in the first row are computed as $905 \times \hat{p}_{.j}$
- The E_{2j} in the second row are computed as $95 \times \hat{p}_{.j}$

Testing homogeneity (III)

- $$Z_1 = \sum_{j=1}^{2} \frac{(O_{1j} - E_{1j})^2}{E_{1j}}$$

  for the cells in the first row would be distributed as $\chi^2_{k-1} = \chi^2_1$ if no parameters were estimated and the $p_{.j}$ used were the correct p_{1j}.

Testing homogeneity (III)

- $$Z_1 = \sum_{j=1}^{2} \frac{(O_{1j} - E_{1j})^2}{E_{1j}}$$

  for the cells in the first row would be distributed as $\chi^2_{k-1} = \chi^2_1$ if no parameters were estimated and the $p_{.j}$ used were the correct p_{1j}.
- Likewise,

  $$Z_2 = \sum_{j=1}^{2} \frac{(O_{2j} - E_{2j})^2}{E_{2j}}$$

  woud be χ^2_1.

Testing homogeneity (III)

- $$Z_1 = \sum_{j=1}^{2} \frac{(O_{1j} - E_{1j})^2}{E_{1j}}$$

  for the cells in the first row would be distributed as $\chi^2_{k-1} = \chi^2_1$ if no parameters were estimated and the $p_{.j}$ used were the correct p_{1j}.
- Likewise,

  $$Z_2 = \sum_{j=1}^{2} \frac{(O_{2j} - E_{2j})^2}{E_{2j}}$$

  woud be χ^2_1.
- $Z = Z_1 + Z_2$ would be distributed as a χ^2_2, but we have to subtract **1** parameter $p_{.1}$ estimated (why not also $p_{.2}$?).

Testing homogeneity (III)

- $$Z_1 = \sum_{j=1}^{2} \frac{(O_{1j} - E_{1j})^2}{E_{1j}}$$

  for the cells in the first row would be distributed as $\chi^2_{k-1} = \chi^2_1$ if no parameters were estimated and the $p_{.j}$ used were the correct p_{1j}.
- Likewise,

  $$Z_2 = \sum_{j=1}^{2} \frac{(O_{2j} - E_{2j})^2}{E_{2j}}$$

  woud be χ^2_1.
- $Z = Z_1 + Z_2$ would be distributed as a χ^2_2, but we have to subtract **1** parameter $p_{.1}$ estimated (why not also $p_{.2}$?).
- **The same** statistic Z follows **the same** distribution under H_0 than in the case of independence.

General rule

- When testing either independence or homogeneity in an $r \times s$ contingency table, in both cases we form

$$Z = \sum_{i,j}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

General rule

- When testing either independence or homogeneity in an $r \times s$ contingency table, in both cases we form

$$Z = \sum_{i,j}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- The resulting value of Z is (under the null hypothesis of independence or homogeneity) distributed as:

$$\chi^2_{(r-1)(s-1)}$$

General rule

- When testing either independence or homogeneity in an $r \times s$ contingency table, in both cases we form

$$Z = \sum_{i,j}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

- The resulting value of Z is (under the null hypothesis of independence or homogeneity) distributed as:

$$\chi^2_{(r-1)(s-1)}$$

- H_0 should be rejected if Z falls in the α right tail of said distribution (alternatively: if the probability to the right of Z in a $\chi^2_{(r-1)(s-1)}$ is "small").

Fisher's exact test (I)

- Consider again our table,

Race	Had cold	Didn't have cold	Total
Whites	n_{11}	n_{12}	$n_{1.}$
Non-whites	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$N = n_{..}$

Fisher's exact test (I)

- Consider again our table,

Race	Had cold	Didn't have cold	Total
Whites	n_{11}	n_{12}	$n_{1.}$
Non-whites	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$\mathrm{N} = n_{..}$

- For given $p_{11}, p_{21}, p_{12}, p_{22}$ its probability would be:

$$\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!} p_{11}^{n_{11}} p_{21}^{n_{21}} p_{12}^{n_{12}} p_{22}^{n_{22}}$$

Fisher's exact test (II)

- The probabilities that N is distributed as it is in the row and column margins are respectively:

$$\frac{N!}{n_{1.}!n_{2.}!}p_{1.}^{n_{1.}}p_{2.}^{n_{2.}} \qquad \frac{N!}{n_{.1}!n_{.1}!}p_{.1}^{n_{.1}}p_{.2}^{n_{.2}}$$

Fisher's exact test (II)

- The probabilities that N is distributed as it is in the row and column margins are respectively:

$$\frac{N!}{n_{1.}!n_{2.}!}p_{1.}^{n_{1.}}p_{2.}^{n_{2.}} \qquad \frac{N!}{n_{.1}!n_{.1}!}p_{.1}^{n_{.1}}p_{.2}^{n_{.2}}$$

- Conditional on the margins, the probability of a given table is:

$$\frac{\left(\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!}p_{11}^{n_{11}}p_{21}^{n_{21}}p_{12}^{n_{12}}p_{22}^{n_{22}}\right)}{\left(\frac{N!}{n_{1.}!n_{2.}!}p_{1.}^{n_{1.}}p_{2.}^{n_{2.}}\right)\left(\frac{N!}{n_{.1}!n_{.2}!}p_{.1}^{n_{.1}}p_{.2}^{n_{.2}}\right)}$$

Fisher's exact test (II)

- The probabilities that N is distributed as it is in the row and column margins are respectively:

$$\frac{N!}{n_{1.}!n_{2.}!}p_{1.}^{n_{1.}}p_{2.}^{n_{2.}} \qquad \frac{N!}{n_{.1}!n_{.1}!}p_{.1}^{n_{.1}}p_{.2}^{n_{.2}}$$

- Conditional on the margins, the probability of a given table is:

$$\frac{\left(\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!}p_{11}^{n_{11}}p_{21}^{n_{21}}p_{12}^{n_{12}}p_{22}^{n_{22}}\right)}{\left(\frac{N!}{n_{1.}!n_{2.}!}p_{1.}^{n_{1.}}p_{2.}^{n_{2.}}\right)\left(\frac{N!}{n_{.1}!n_{.2}!}p_{.1}^{n_{.1}}p_{.2}^{n_{.2}}\right)}$$

- Under the null hypothesis $p_{ij} = p_{i.}p_{.j}$ all nuisance parameters cancel!

Fisher's exact test (III)

- All we are left with for the probability of a given table is:

$$\frac{\left(\dfrac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!}\right)}{\left(\dfrac{N!}{n_{1.}!n_{2.}!}\right)\left(\dfrac{N!}{n_{.1}!n_{.2}!}\right)}$$

Fisher's exact test (III)

- All we are left with for the probability of a given table is:

$$\frac{\left(\dfrac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!}\right)}{\left(\dfrac{N!}{n_{1.}!n_{2.}!}\right)\left(\dfrac{N!}{n_{.1}!n_{.2}!}\right)}$$

- The denominator is always the same.

Fisher's exact test (III)

- All we are left with for the probability of a given table is:

$$\frac{\left(\dfrac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!}\right)}{\left(\dfrac{N!}{n_{1.}!n_{2.}!}\right)\left(\dfrac{N!}{n_{.1}!n_{.2}!}\right)}$$

- The denominator is always the same.
- Can compute the probability of each table under the null $H_0 : p_{ij} = p_{i.}p_{.j}$ and check whether what we have observed is very unlikely.

Fisher's exact test (III)

- All we are left with for the probability of a given table is:

$$\frac{\left(\frac{N!}{n_{11}!n_{12}!n_{21}!n_{22}!}\right)}{\left(\frac{N!}{n_{1.}!n_{2.}!}\right)\left(\frac{N!}{n_{.1}!n_{.2}!}\right)}$$

- The denominator is always the same.
- Can compute the probability of each table under the null $H_0 : p_{ij} = p_{i.}p_{.j}$ and check whether what we have observed is very unlikely.
- Unfeasible for large tables.

Fisher's exact test (IV)

- Function to do it in R. Useful for small tables; no approximations. Will fail for large tables.

```
>  fisher.test(ColdRace)
         Fisher's Exact Test for Count Data

data:  ColdRace
p-value = 0.7
alternative hypothesis: true odds ratio is not
95 percent confidence interval:
 0.5346 2.1400
sample estimates:
odds ratio
     1.113
```

Fisher's exact test (V)

Table: Accidents 1970-2009 of european airlines involving loss of life

Airline	Flights	Accs	Airline	Flights	Accs
Aer Lingus	1200000	0	Icelandair	390000	0
Air France	5900000	8	Lufthansa	7300000	4
Alitalia	3900000	3	KLM	2400000	3
Austrian Airlines	750000	0	Olympic Airways	1800000	3
Braathens	1350000	1	Sabena	1600000	0
British Airways	8270000	3	SAS	5400000	2
British Midland	1030000	1	Swiss/Swissair	3200000	5
easyJet	760000	0	TAP Air Portugal	850000	3
Finnair	1700000	0	Turkish Airlines	2100000	10
Iberia	4500000	4	Virgin Atlantic	150000	0

Fisher’s exact test (VI)

- When counts are very small, the chi square approximation may be bad, and Fisher’s test is required.

```
> chisq.test(accidentes)
        Pearson's Chi-squared test

data:  accidentes
X-squared = 56, df = 19, p-value = 2e-05
```

Fisher’s exact test (VII)

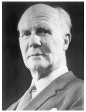
- It is not feasible to compute ALL tables, simulation is performed instead.

```
> fisher.test(accidentes,
              simulate.p.value=TRUE,
              B=10000)
        Fisher's Exact Test for Count Data wit
        simulated p-value (based on 10000 repl

data:  accidentes
p-value = 0.002
alternative hypothesis: two.sided
```

Historical notes

Egon Pearson

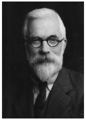


C.R.Rao

K.Pearson

W.S.Gosset

R.A.Fisher



J.Neyman

H.Cramé

1900
Method moments,
χ^2 tests

1908
Student's *t*

1921
Likelihood,
sufficiency,
efficiency

1933
NP lemma,
power,
UMP tests

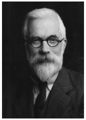
1945
CR
inequality

Historical notes

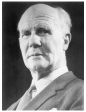
K.Pearson

W.S.Gosset

R.A.Fisher



Egon Pearson



J.Neyman

C.R.Rao

H.Cramér

1900
Method
moments,
χ^2 tests

1908
Student's t

1921
Likelihood,
sufficiency,
efficiency

1933
NP lemma,
power,
UMP tests

1945
CR
inequality

Historical notes

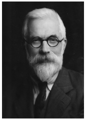
K.Pearson

1900
Method
moments,
χ^2 tests

W.S.Gosset

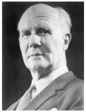
1908
Student's t

R.A.Fisher



1921
Likelihood,
sufficiency,
efficiency

Egon Pearson



J.Neyman

1933
NP lemma,
power,
UMP tests

C.R.Rao

H.Cramér

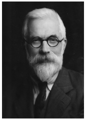
1945
CR
inequality

Historical notes

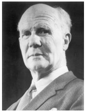
K.Pearson

W.S.Gosset

R.A.Fisher



Egon Pearson



J.Neyman

C.R.Rao

H.Cramér

1900
Method moments,
χ^2 tests

1908
Student's *t*

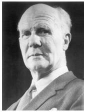
1921
Likelihood,
sufficiency,
efficiency

1933
NP lemma,
power,
UMP tests

1945
CR
inequality

Historical notes

Egon Pearson

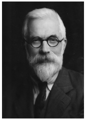


C.R.Rao

K.Pearson

W.S.Gosset

R.A.Fisher



J.Neyman

H.Cramér

1900
Method moments, χ^2 tests

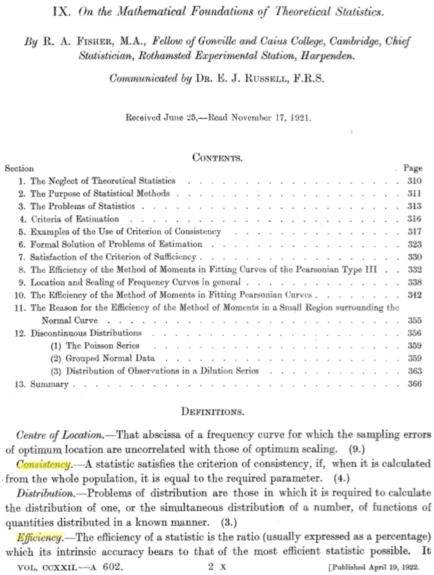
1908
Student's t

1921
Likelihood, sufficiency, efficiency

1933
NP lemma, power, UMP tests

1945
CR inequality

Earlier notions were defined differently



IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by Dr. E. J. Russell, F.R.S.

Received June 25,—Read November 17, 1921.

Contents.

Definitions.

Centre of Location.—That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (9.)

Consistency.—A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

Distribution.—Problems of distribution are those in which it is required to calculate the distribution of one, or the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

Efficiency.—The efficiency of a statistic is the ratio (usually expressed as a percentage) which its intrinsic accuracy bears to that of the most efficient statistic possible. It

VOL. CCXXII.—A 602. 2 X [Published April 19, 1922.

Introduction

- Normal distribution is a useful model in many situations.

Introduction

- Normal distribution is a useful model in many situations.
- Why? Central Limit Theorem.

Introduction

- Normal distribution is a useful model in many situations.
- Why? Central Limit Theorem.
- Even when the the distribution of a random variable is not normal, normal theory based tests are surprisingly adequate.

Introduction

- Normal distribution is a useful model in many situations.
- Why? Central Limit Theorem.
- Even when the the distribution of a random variable is not normal, normal theory based tests are surprisingly adequate.
- By “adequate” is meant that significance levels (α) and power ($1 - \beta$) are close to theoretical values.

Introduction

- Normal distribution is a useful model in many situations.
- Why? Central Limit Theorem.
- Even when the the distribution of a random variable is not normal, normal theory based tests are surprisingly adequate.
- By “adequate” is meant that significance levels (α) and power ($1 - \beta$) are close to theoretical values.
- Several of these tests first introduced by Fisher, put on a firmer ground by the Neyman-Pearson lemma.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (I)

- We have $\overline{X} \sim N(m_0, \sigma^2/n)$ and therefore:

$$T = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (I)

- We have $\overline{X} \sim N(m_0, \sigma^2/n)$ and therefore:

$$T = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- T can be computed because σ^2 is known.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (I)

- We have $\overline{X} \sim N(m_0, \sigma^2/n)$ and therefore:

$$T = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

- T can be computed because σ^2 is known.
- Hence,

$$\text{Prob}\left\{-z_{\alpha/2} \leq T \leq z_{\alpha/2}\right\} = 1 - \alpha$$

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (I)

- We have $\overline{X} \sim N(m_0, \sigma^2/n)$ and therefore:

$$T = \frac{\overline{X} - m_0}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- T can be computed because σ^2 is known.
- Hence,

$$\text{Prob}\left\{-z_{\alpha/2} \leq T \leq z_{\alpha/2}\right\} = 1 - \alpha$$

- We would reject H_0 at the significance level α if $|T| > |z_{\alpha/2}|$.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (II)

- If we expect departures from H_0 to be of the form $m > m_0$ or $m < m_0$ we would adjust the critical region accordingly:

$$\begin{aligned} m > m_0 &\implies \text{Reject if } T > z_\alpha \\ m < m_0 &\implies \text{Reject if } T < -z_\alpha \end{aligned}$$

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (II)

- If we expect departures from H_0 to be of the form $m > m_0$ or $m < m_0$ we would adjust the critical region accordingly:

$$\begin{aligned} m > m_0 &\implies \text{Reject if } T > z_\alpha \\ m < m_0 &\implies \text{Reject if } T < -z_\alpha \end{aligned}$$

- Makes sense when looking at the test statistic $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$; would also be the answer given by the Neyman and Pearson theorem for a simple alternative.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (II)

- If we expect departures from H_0 to be of the form $m > m_0$ or $m < m_0$ we would adjust the critical region accordingly:

$$\begin{aligned} m > m_0 &\implies \text{Reject if } T > z_\alpha \\ m < m_0 &\implies \text{Reject if } T < -z_\alpha \end{aligned}$$

- Makes sense when looking at the test statistic $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$; would also be the answer given by the Neyman and Pearson theorem for a simple alternative.
- "Reject if $|T| > |z_{\alpha/2}|$" is just a compromise when no clear alternative.

A digression: confidence intervals

- When testing H_0 with no given alternative, the “unlikely” region is the critical region.

A digression: confidence intervals

- When testing H_0 with no given alternative, the “unlikely” region is the critical region.
- The “likely” region is the confidence interval.

A digression: confidence intervals

- When testing H_0 with no given alternative, the “unlikely” region is the critical region.
- The “likely” region is the confidence interval.
- This does **not** extend to tests with a prescribed alternative H_a.

A digression: confidence intervals

- When testing H_0 with no given alternative, the “unlikely” region is the critical region.
- The “likely” region is the confidence interval.
- This does **not** extend to tests with a prescribed alternative H_a.
- When we have a H_a, the critical region may be one-sided, not the complement of the confidence interval.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (III)

- What is the payoff of a larger sample size n?

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (III)

- What is the payoff of a larger sample size n?
- The test statistic always is $N(0, 1)$ distributed under the null H_0.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (III)

- What is the payoff of a larger sample size n?
- The test statistic always is $N(0, 1)$ distributed under the null H_0.
- However, under an alternative $m \neq m_0$,

$$T = \frac{\sqrt{n}(\overline{X} - m_0)}{\sigma}$$

has mean $\sqrt{n}(m - m_0)/\sigma$.

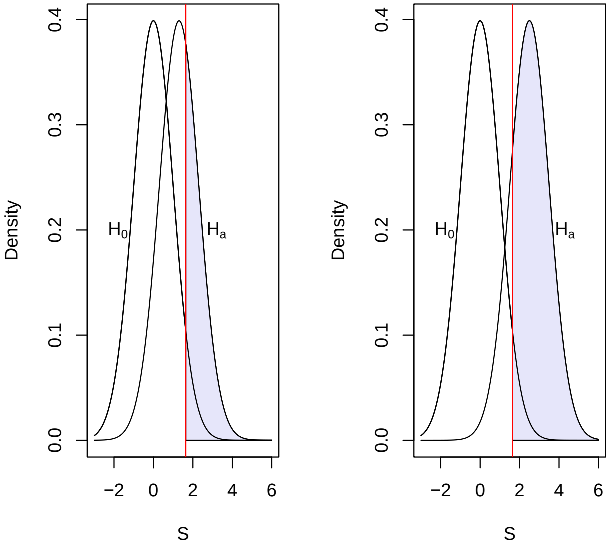
$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (III)

- What is the payoff of a larger sample size n?
- The test statistic always is $N(0,1)$ distributed under the null H_0.
- However, under an alternative $m \neq m_0$,

  $$T = \frac{\sqrt{n}(\overline{X} - m_0)}{\sigma}$$

  has mean $\sqrt{n}(m - m_0)/\sigma$.
- For given m, the greater n, the farther away from 0 is the mean of the test statistic.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 known (IV)



$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (I)

- Now, $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$ cannot be computed, for σ^2 is not known.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (I)

- Now, $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$ cannot be computed, for σ^2 is not known.
- Replacing σ^2 by its estimate $s^2 = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ gives an estimator whose distribution under H_0 is no longer $N(0, 1)$.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (I)

- Now, $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$ cannot be computed, for σ^2 is not known.
- Replacing σ^2 by its estimate $s^2 = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ gives an estimator whose distribution under H_0 is no longer $N(0, 1)$.
- Key fact:

$$\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

  and is **independent** of $\overline{X}$.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (I)

- Now, $T = \dfrac{\overline{X} - m_0}{\sigma/\sqrt{n}}$ cannot be computed, for σ^2 is not known.
- Replacing σ^2 by its estimate $s^2 = n^{-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$ gives an estimator whose distribution under H_0 is no longer $N(0, 1)$.
- Key fact:

  $$\frac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$$

  and is **independent** of $\overline{X}$.
- This paves the way to eliminating the nuisance parameter σ^2 by *studentization*.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (II)

- The ratio,

$$T = \frac{\frac{\sqrt{n}(\overline{X}-m_0)}{\sigma}}{\sqrt{\frac{nS^2/\sigma^2}{n-1}}} \quad = \quad \frac{(\overline{X}-m_0)}{S}\sqrt{n-1} \quad \sim \quad \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

  when $H_0 : m = m_0$ is true.

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (II)

- The ratio,

  $$T = \frac{\frac{\sqrt{n}(\overline{X}-m_0)}{\sigma}}{\sqrt{\frac{nS^2/\sigma^2}{n-1}}} \quad = \quad \frac{(\overline{X}-m_0)}{S}\sqrt{n-1} \quad \sim \quad \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

  when $H_0 : m = m_0$ is true.
- Therefore we can compare the values of the test statistic T to a t_{n-1} (Student's t with $n-1$ degrees of freedom).

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (II)

- The ratio,

$$T = \frac{\frac{\sqrt{n}(\overline{X}-m_0)}{\sigma}}{\sqrt{\frac{nS^2/\sigma^2}{n-1}}} = \frac{(\overline{X}-m_0)}{S}\sqrt{n-1} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

  when $H_0 : m = m_0$ is true.
- Therefore we can compare the values of the test statistic T to a t_{n-1} (Student's t with $n-1$ degrees of freedom).
- Decision rule: "Reject H_0 if $|T| > t_{\alpha/2;n-1}$."

$H_0 : m = m_0$ with $X \sim N(m, \sigma^2)$ and σ^2 unknown (II)

- The ratio,

  $$T = \frac{\frac{\sqrt{n}(\overline{X}-m_0)}{\sigma}}{\sqrt{\frac{nS^2/\sigma^2}{n-1}}} \quad = \quad \frac{(\overline{X}-m_0)}{S}\sqrt{n-1} \quad \sim \quad \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

  when $H_0 : m = m_0$ is true.
- Therefore we can compare the values of the test statistic T to a t_{n-1} (Student's t with $n-1$ degrees of freedom).
- Decision rule: "Reject H_0 if $|T| > t_{\alpha/2;n-1}$."
- Again, we take critical regions of full α size to the right or to the left, if alternative is one-sided.

Example: $H_0 : m_0 = 2, \sigma^2 = 1$ known

- Let the sample be:

```
> x <- c(2.2,3.4,2.9,3,1.6,3,3.1,3.6,1.9)
> length(x)                        #  sample size
[1] 9
> T <- sqrt(9) * ( mean(x) - 2 ) / 1
> T
[1] 2.233
> qnorm(0.975)                     #  leaves tails of
[1] 1.96
```

- In this case, we would reject.

Example: $H_0 : m_0 = 2$, $\sigma^2 = 1$ unknown

- Now, we would compute

```
> T <- sqrt(9-1) * ( mean(x) - 2 ) / sqrt( 8 *
> T
[1] 3.257
> qt(0.975,df=8)          # so reject
[1] 2.306
> var(x)
[1] 0.4703
> sum( (x-mean(x))^2 ) / 9
[1] 0.418
> ( 8 * var(x) / 9 )     # var command uses (n-1
[1] 0.418
```

$H_0 : \sigma^2 = \sigma_0^2$ with $X \sim N(m, \sigma^2)$

- Under the null hypothesis,

$$T = \frac{nS^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

$H_0 : \sigma^2 = \sigma_0^2$ with $X \sim N(m, \sigma^2)$

- Under the null hypothesis,

$$T = \frac{nS^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

- Therefore,

$$\text{Prob}\left\{\chi^2_{n-1;1-\alpha/2} \leq \frac{nS^2}{\sigma_0^2} \leq \chi^2_{n-1;\alpha/2}\right\} = 1 - \alpha$$

$H_0 : \sigma^2 = \sigma_0^2$ with $X \sim N(m, \sigma^2)$

- Under the null hypothesis,

$$T = \frac{nS^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

- Therefore,

$$\text{Prob}\left\{\chi^2_{n-1;1-\alpha/2} \leq \frac{nS^2}{\sigma_0^2} \leq \chi^2_{n-1;\alpha/2}\right\} = 1 - \alpha$$

- Critical region $[0, \chi^2_{n-1;1-\alpha/2}] \cup [\chi^2_{n-1;\alpha/2}, \infty)$, unless we have an alternative $H_a : \sigma^2 < \sigma_0^2$ or $H_a : \sigma^2 > \sigma_0^2$

$H_0 : \sigma^2 = \sigma_0^2$ with $X \sim N(m, \sigma^2)$

- Under the null hypothesis,

$$T = \frac{nS^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

- Therefore,

$$\mathrm{Prob}\left\{\chi^2_{n-1;1-\alpha/2} \leq \frac{nS^2}{\sigma_0^2} \leq \chi^2_{n-1;\alpha/2}\right\} = 1 - \alpha$$

- Critical region $[0, \chi^2_{n-1;1-\alpha/2}] \cup [\chi^2_{n-1;\alpha/2}, \infty)$, unless we have an alternative $H_a : \sigma^2 < \sigma_0^2$ or $H_a : \sigma^2 > \sigma_0^2$
- In the first case the critical region would be $[0, \chi^2_{n-1;1-\alpha}]$, in the second $[\chi^2_{n-1;\alpha}, \infty)$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances known (I)

- The commonest test by far is that of $H_0 : m_1 - m_2 = 0$, but we present the test generally.

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances known (I)

- The commonest test by far is that of $H_0 : m_1 - m_2 = 0$, but we present the test generally.
- We have,

$$\overline{X} - \overline{Y} \quad \sim \quad N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances known (I)

- The commonest test by far is that of $H_0 : m_1 - m_2 = 0$, but we present the test generally.
- We have,

$$\overline{X} - \overline{Y} \quad \sim \quad N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

- Hence, under H_0,

$$\frac{\overline{X} - \overline{Y} - (m_1^* - m_2^*)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n}}} \sim N(0, 1)$$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances known (II)

- Therefore, under H_0,

$$\text{Prob}\left\{-z_{\alpha/2} \leq \frac{\overline{X} - \overline{Y} - (m_1^* - m_2^*)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right\} = 1 - \alpha$$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances known (II)

- Therefore, under H_0,

$$\text{Prob}\left\{ -z_{\alpha/2} \leq \frac{\overline{X} - \overline{Y} - (m_1^* - m_2^*)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \right\} = 1 - \alpha$$

- The critical region for the test statistic is made of the two $\alpha/2$ tails, unless we have reason to expect the deviance to be one-sided.

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances $\sigma_1^2 = \sigma_2^2$ unknown (I)

- We have,

$$\begin{aligned}\overline{X} - \overline{Y} &\sim N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \\ \frac{n_1 S_1^2}{\sigma_1^2} + \frac{n_2 S_2^2}{\sigma_2^2} &\sim \chi^2_{n_1+n_2-2}\end{aligned}$$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances $\sigma_1^2 = \sigma_2^2$ unknown (I)

- We have,

$$\begin{aligned} \overline{X} - \overline{Y} &\sim N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \right) \\ \frac{n_1 S_1^2}{\sigma_1^2} + \frac{n_2 S_2^2}{\sigma_2^2} &\sim \chi^2_{n_1+n_2-2} \end{aligned}$$

- Using the crucial assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ we can construct a test statistic which does not depend on σ^2 .

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances $\sigma_1^2 = \sigma_2^2$ unknown (II)

- Using $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\frac{\overline{X}-\overline{Y}-(m_1-m_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{1}{\sigma}\sqrt{\frac{n_1 S_1^2+n_2 S_2^2}{n_1+n_2-2}}} \quad \sim \quad t_{n_1+n_2-2}$$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances $\sigma_1^2 = \sigma_2^2$ unknown (II)

- Using $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\frac{\overline{X}-\overline{Y}-(m_1-m_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\frac{1}{\sigma}\sqrt{\frac{n_1 S_1^2+n_2 S_2^2}{n_1+n_2-2}}} \quad \sim \quad t_{n_1+n_2-2}$$

- Cancelling the nuisance parameter σ we end up with:

$$\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{n_1 S_1^2+n_2 S_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}$$

$H_0 : m_1 - m_2 = m_1^* - m_2^*$ with X, Y normal, variances $\sigma_1^2 = \sigma_2^2$ unknown (II)

- Using $\sigma_1^2 = \sigma_2^2 = \sigma^2$

$$\frac{\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\frac{1}{\sigma}\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}} \quad \sim \quad t_{n_1+n_2-2}$$

- Cancelling the nuisance parameter σ we end up with:

$$\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}$$

- Assumption $\sigma_1^2 = \sigma_2^2$ crucial, otherwise an open question (so-called Behrens-Fisher problem).

$H_0 : \sigma_1^2/\sigma_2^2 = \sigma_{1*}^2/\sigma_{2*}^2$ with X, Y normal (I)

- With respective sample sizes n_1 and n_2, we have:

$$\frac{n_1 S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \qquad \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

$H_0 : \sigma_1^2/\sigma_2^2 = \sigma_{1*}^2/\sigma_{2*}^2$ with X, Y normal (I)

- With respective sample sizes n_1 and n_2, we have:

$$\frac{n_1 S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \qquad \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

- Clearly both statistics are independent, so

$$\frac{n_1 S_1^2 \sigma_2^2 (n_2 - 1)}{n_2 S_2^2 \sigma_1^2 (n_1 - 1)} \sim \mathcal{F}_{n_1-1, n_2-1}$$

$H_0 : \sigma_1^2/\sigma_2^2 = \sigma_{1*}^2/\sigma_{2*}^2$ with X, Y normal (I)

- With respective sample sizes n_1 and n_2, we have:

$$\frac{n_1 S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1} \qquad \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

- Clearly both statistics are independent, so

$$\frac{n_1 S_1^2 \sigma_2^2 (n_2 - 1)}{n_2 S_2^2 \sigma_1^2 (n_1 - 1)} \sim \mathcal{F}_{n_1-1, n_2-1}$$

- It the hypothesis H_0 is true, replacing σ_1^2, σ_2^2 by their hypothetical values would give a test statistic with the distribution shown.

General ideas

- Tests for a mean or the difference of means are remarkably robust to deviations from normality; however, to play safe we might use tests to be described next.

General ideas

- Tests for a mean or the difference of means are remarkably robust to deviations from normality; however, to play safe we might use tests to be described next.
- Tests for the difference of means are quite sensitive to different variances: the requirement $\sigma_1^2 = \sigma_2^2$ cannot be dispensed with.

Permutation tests (I)

- Easy alternative when distribution cannot be assumed and we can use a computer.

Permutation tests (I)

- Easy alternative when distribution cannot be assumed and we can use a computer.
- Want to test $x_1, \dots, x_{n_1}$ and $y_1, \dots, y_{n_2}$ are indeed samples form the same population, the alternative being that the means are different.

Permutation tests (I)

- Easy alternative when distribution cannot be assumed and we can use a computer.
- Want to test $x_1, \dots, x_{n_1}$ and $y_1, \dots, y_{n_2}$ are indeed samples form the same population, the alternative being that the means are different.
- Our test statistic is $\overline{x} - \overline{y}$. Need something to compare to.

Permutation tests (I)

- Easy alternative when distribution cannot be assumed and we can use a computer.
- Want to test $x_1, \ldots, x_{n_1}$ and $y_1, \ldots, y_{n_2}$ are indeed samples form the same population, the alternative being that the means are different.
- Our test statistic is $\overline{x} - \overline{y}$. Need something to compare to.
- If we arrange the observations as:

  $$x_1, \ldots, x_{n_1}, y_1, \ldots, y_{n_2}$$

  $\overline{x} - \overline{y}$ is just the difference of the averages of the first n_1 and subsequent n_2 observations.

Permutation tests (II)

- If observations come indeed from the same population, the difference between the n_1 and n_2 observations in each group should be of similar magnitude than that among any other set of n_1 and n_2 observations.

Permutation tests (II)

- If observations come indeed from the same population, the difference between the n_1 and n_2 observations in each group should be of similar magnitude than that among any other set of n_1 and n_2 observations.
- Idea: sample repeatedly the whole set of observations in random subsets of n_1 and n_2, and compute each time $(\overline{x} - \overline{y})_j$ $(j = 1, \ldots, N)$.

Permutation tests (II)

- If observations come indeed from the same population, the difference between the n_1 and n_2 observations in each group should be of similar magnitude than that among any other set of n_1 and n_2 observations.
- Idea: sample repeatedly the whole set of observations in random subsets of n_1 and n_2, and compute each time $(\overline{x} - \overline{y})_j$ $(j = 1, \ldots, N)$.
- Compare the observed $\overline{x} - \overline{y}$ to $(\overline{x} - \overline{y})_j$ $(j = 1, \ldots, N)$ and reject H_0 if it is in an extreme position.

Permutation tests (II)

- If observations come indeed from the same population, the difference between the n_1 and n_2 observations in each group should be of similar magnitude than that among any other set of n_1 and n_2 observations.
- Idea: sample repeatedly the whole set of observations in random subsets of n_1 and n_2, and compute each time $(\overline{x} - \overline{y})_j$ $(j = 1, \ldots, N)$.
- Compare the observed $\overline{x} - \overline{y}$ to $(\overline{x} - \overline{y})_j$ $(j = 1, \ldots, N)$ and reject H_0 if it is in an extreme position.
- Sampling is usually done by permuting the original sample, hence the name.

Testing $H_0 : m = m_0$ with no normality (I)

- For "large" n (=sample size), use normal theory tests. "Large" is $n \geq 30$ (if σ^2 is known) and $n \geq 100$ (if it is not).

Testing $H_0 : m = m_0$ with no normality (I)

- For "large" n (=sample size), use normal theory tests. "Large" is $n \geq 30$ (if σ^2 is known) and $n \geq 100$ (if it is not).
- For smaller n, remember Tchebycheff inequality:

$$\mathrm{Prob}\left\{|X - m| < k\sigma\right\} \geq 1 - \frac{1}{k^2}$$

Testing $H_0 : m = m_0$ with no normality (I)

- For “large” n (=sample size), use normal theory tests. “Large” is $n \geq 30$ (if σ^2 is known) and $n \geq 100$ (if it is not).
- For smaller n, remember Tchebycheff inequality:

$$\text{Prob}\left\{|X - m| < k\sigma\right\} \geq 1 - \frac{1}{k^2}$$

- For the particular case of $\overline{X}$ we have:

$$\text{Prob}\left\{|\overline{X} - m| < \frac{k\sigma}{\sqrt{n}}\right\} \geq 1 - \frac{1}{k^2}$$

Testing $H_0 : m = m_0$ with no normality (II)

- Therefore, replacing k by $1/\sqrt{\alpha}$ we have:

$$\text{Prob}\left\{|\overline{X} - m| < \frac{\sigma}{\sqrt{n\alpha}}\right\} \geq 1 - \alpha$$

Testing $H_0 : m = m_0$ with no normality (II)

- Therefore, replacing k by $1/\sqrt{\alpha}$ we have:

$$\text{Prob}\left\{|\overline{X} - m| < \frac{\sigma}{\sqrt{n\alpha}}\right\} \geq 1 - \alpha$$

- This gives as a basis for a confidence interval for m and a test: "Reject H_0 at the α significance level if $|\overline{X} - m_0| > \sigma/\sqrt{n\alpha}$."

Testing $H_0 : m = m_0$ with no normality (II)

- Therefore, replacing k by $1/\sqrt{\alpha}$ we have:

$$\text{Prob}\left\{|\overline{X} - m| < \frac{\sigma}{\sqrt{n\alpha}}\right\} \geq 1 - \alpha$$

- This gives as a basis for a confidence interval for m and a test: “Reject H_0 at the α significance level if $|\overline{X} - m_0| > \sigma/\sqrt{n\alpha}$.”
- If σ^2 is unknown, replace it by its estimate s^2 to have an approximate test.

Testing $H_0 : m = m_0$ with no normality (II)

- Therefore, replacing k by $1/\sqrt{\alpha}$ we have:

$$\text{Prob}\left\{|\overline{X} - m| < \frac{\sigma}{\sqrt{n\alpha}}\right\} \geq 1 - \alpha$$

- This gives as a basis for a confidence interval for m and a test: “Reject H_0 at the α significance level if $|\overline{X} - m_0| > \sigma/\sqrt{n\alpha}$.”
- If σ^2 is unknown, replace it by its estimate s^2 to have an approximate test.
- This distribution-free method gives tests less powerful (and confidence intervals wider) than the normal theory tests.

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.
- Clearly $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ is an unbiased estimate of p.

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.
- Clearly $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ is an unbiased estimate of p.
- How to test hypothesis on p or estimate it by interval? We know that for large n approximately,

$$\frac{\overline{X} - p}{s/\sqrt{n}} \approx N(0, 1)$$

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.
- Clearly $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ is an unbiased estimate of p.
- How to test hypothesis on p or estimate it by interval? We know that for large n approximately,

$$\frac{\overline{X} - p}{s/\sqrt{n}} \approx N(0, 1)$$

- We can estimate s^2 by $\hat{p}(1 - \hat{p})$ or (conservatively) by 0.25.

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.
- Clearly $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ is an unbiased estimate of p.
- How to test hypothesis on p or estimate it by interval? We know that for large n approximately,

$$\frac{\overline{X} - p}{s/\sqrt{n}} \approx N(0,1)$$

- We can estimate s^2 by $\hat{p}(1 - \hat{p})$ or (conservatively) by 0.25.
- However we estimate p, approximately, for large n,

$$\frac{\overline{X} - m}{s/\sqrt{n}} \approx N(0,1)$$

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.
- Clearly $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ is an unbiased estimate of p.
- How to test hypothesis on p or estimate it by interval? We know that for large n approximately,

$$\frac{\overline{X} - p}{s/\sqrt{n}} \approx N(0,1)$$

- We can estimate s^2 by $\hat{p}(1-\hat{p})$ or (conservatively) by 0.25.
- However we estimate p, approximately, for large n,

$$\frac{\overline{X} - m}{s/\sqrt{n}} \approx N(0,1)$$

How would we construct a confidence interval for p

The case of a proportion (I)

- One case of particular interest is that of a proportion. Variable X the value 0 or 1 (“yes” or “no”, or similar dichotomous values coded to 1/0).
- We are interested in the probability of 1, p.
- Clearly $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$ is an unbiased estimate of p.
- How to test hypothesis on p or estimate it by interval? We know that for large n approximately,

$$\frac{\overline{X} - p}{s/\sqrt{n}} \approx N(0, 1)$$

- We can estimate s^2 by $\hat{p}(1 - \hat{p})$ or (conservatively) by 0.25.
- However we estimate p, approximately, for large n,

$$\frac{\overline{X} - m}{s/\sqrt{n}} \approx N(0, 1)$$

$(\overline{X} \pm z_{\alpha/2} s/\sqrt{n})$

The case of a proportion (II)

Example:

- In a sample of 500 parts from a very large batch, 33 are found to be defective. Would the hypothesis $H_0 : p = 0.04$ be rejected against an alternative $H_a : p > 0.04$? $(\alpha = 0.05)$.

The case of a proportion (II)

Example:

- In a sample of 500 parts from a very large batch, 33 are found to be defective. Would the hypothesis $H_0 : p = 0.04$ be rejected against an alternative $H_a : p > 0.04$? $(\alpha = 0.05)$.
- The estimate of p would be $33/500 = 0.066$ and $s^2 = pq = 0.04 \times 0.96 = 0.0384$. Under H_0,

$$\frac{(\overline{X} - 0.04)}{\sqrt{0.0384}/\sqrt{500}} \approx N(0, 1);$$

the critical region would be to the right.

The case of a proportion (II)

Example:

- In a sample of 500 parts from a very large batch, 33 are found to be defective. Would the hypothesis $H_0 : p = 0.04$ be rejected against an alternative $H_a : p > 0.04$? $(\alpha = 0.05)$.
- The estimate of p would be $33/500 = 0.066$ and $s^2 = pq = 0.04 \times 0.96 = 0.0384$. Under H_0,

  $$\frac{(\overline{X} - 0.04)}{\sqrt{0.0384}/\sqrt{500}} \approx N(0,1);$$

  the critical region would be to the right.
- Replacing $\overline{X}$ by $33/500$ we get a value for the test statistic of 2.97, well inside a critical region of size $\alpha = 0.01$. So we would reject H_0 at said level of significance.

The case of a proportion (III)

Example (continued):

- If we were asked to estimate by interval the true p with confidence $1-\alpha=0.99$, we could use:

$$\frac{(\overline{X}-p)}{\sqrt{\frac{0.0667\times 0.9333}{500}}}\approx N(0,1)$$

The case of a proportion (III)

Example (continued):

- If we were asked to estimate by interval the true p with confidence $1 - \alpha = 0.99$, we could use:

$$\frac{(\overline{X} - p)}{\sqrt{\dfrac{0.0667 \times 0.9333}{500}}} \approx N(0, 1)$$

- Then,

$$\mathrm{Prob}\left\{\overline{X} - 2.5758\sqrt{\frac{0.06222}{500}} \leq p \leq \overline{X} + 2.5758\sqrt{\frac{0.06222}{500}}\right\} \approx 0$$

The case of a proportion (III)

Example (continued):

- If we were asked to estimate by interval the true p with confidence $1-\alpha=0.99$, we could use:

$$\frac{(\overline{X}-p)}{\sqrt{\frac{0.0667\times 0.9333}{500}}}\approx N(0,1)$$

- Then,

$$\text{Prob}\left\{\overline{X}-2.5758\sqrt{\frac{0.06222}{500}}\leq p\leq \overline{X}+2.5758\sqrt{\frac{0.06222}{500}}\right\}\approx 0$$

- The confidence interval would thus be (0.0666 ± 0.0287)

The case of a proportion (III)

Example (continued):

- If we were asked to estimate by interval the true p with confidence $1 - \alpha = 0.99$, we could use:

$$\frac{(\overline{X} - p)}{\sqrt{\dfrac{0.0667 \times 0.9333}{500}}} \approx N(0, 1)$$

- Then,

$$\mathrm{Prob}\left\{\overline{X} - 2.5758\sqrt{\frac{0.06222}{500}} \le p \le \overline{X} + 2.5758\sqrt{\frac{0.06222}{500}}\right\} \approx 0$$

- The confidence interval would thus be (0.0666 ± 0.0287)
- Replacing s^2 by the upper bound of $p(1 - p) = 0.25$ would be *very* conservative here.

Testing differences of means

- We state without proof the following approximate results:

$$\begin{aligned}
\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} &\approx N(0,1) \qquad (n_1 \geq 30, n_2 \geq 30) \\
\frac{\overline{X} - \overline{Y} - (m_1 - m_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} &\approx N(0,1) \qquad (n_1 \geq 100, n_2 \geq 100)
\end{aligned}$$

- Those approximate distributions can be used in the construction of test statistics or confidence intervals.

Testing differences of proportions

- The results in the previous slide can be specialized to the case of two proportions. In that case,

$$\begin{aligned} \overline{X} &= \frac{Z_1}{n_1} \qquad m_1 = p_1 \\ \overline{Y} &= \frac{Z_2}{n_2} \qquad m_2 = p_2 \\ \frac{\frac{Z_1}{n_1} - \frac{Z_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} &\approx N(0,1) \end{aligned}$$

Testing differences of proportions

- The results in the previous slide can be specialized to the case of two proportions. In that case,

$$\begin{aligned} \overline{X} &= \frac{Z_1}{n_1} \qquad m_1 = p_1 \\ \overline{Y} &= \frac{Z_2}{n_2} \qquad m_2 = p_2 \\ \frac{\frac{Z_1}{n_1} - \frac{Z_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} &\approx N(0,1) \end{aligned}$$

- Again, sample sizes should be large.

Testing differences of proportions

- The results in the previous slide can be specialized to the case of two proportions. In that case,

$$\begin{aligned} \overline{X} &= \frac{Z_1}{n_1} \qquad m_1 = p_1 \\ \overline{Y} &= \frac{Z_2}{n_2} \qquad m_2 = p_2 \\ \frac{\frac{Z_1}{n_1} - \frac{Z_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} &\approx N(0,1) \end{aligned}$$

- Again, sample sizes should be large.

How would we construct a confidence interval for $(p_1 - p_2)$?

Testing differences of proportions

- The results in the previous slide can be specialized to the case of two proportions. In that case,

$$\begin{aligned} \overline{X} &= \frac{Z_1}{n_1} \qquad m_1 = p_1 \\ \overline{Y} &= \frac{Z_2}{n_2} \qquad m_2 = p_2 \\ \frac{\frac{Z_1}{n_1} - \frac{Z_2}{n_2} - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} &\approx N(0,1) \end{aligned}$$

- Again, sample sizes should be large.

$$\left(\frac{Z_1}{n_1} - \frac{Z_2}{n_2}\right) \pm z_{\alpha/2} \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

The OC (“operating characteristic”) curve (I)

- The performance of a test of H_0 against a set of alternatives usually described by the OC curve: it gives the probability of non-rejection of H_0 for both the null and a range of alternatives.

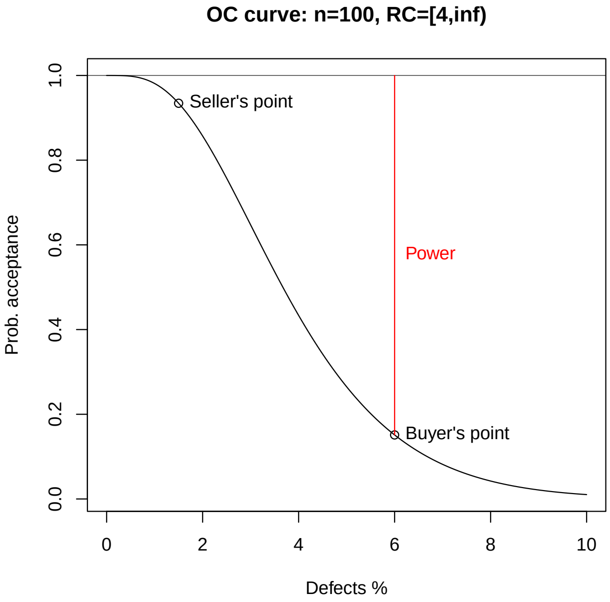
The OC (“operating characteristic”) curve (I)

- The performance of a test of H_0 against a set of alternatives usually described by the OC curve: it gives the probability of non-rejection of H_0 for both the null and a range of alternatives.
- Common in specification of industrial quality sampling protocols.

The OC ("operating characteristic") curve (I)

- The performance of a test of H_0 against a set of alternatives usually described by the OC curve: it gives the probability of non-rejection of H_0 for both the null and a range of alternatives.
- Common in specification of industrial quality sampling protocols.
- The conflicting interests of buyer and seller are specified in two points, through which the the curve is forced to pass.

The OC (“operating characteristic”) curve (II)



Paired comparisons (I)

- When performing the classical t-test we assume *independent* observations.
- Sometimes, this is clearly not the case. A different test would be indicated: one possibility is the **paired comparisons test**.

Paired comparisons (II)

- Consider the following data on weight at birth of male babies:

Mother	First	Second
A	3.800	4.150
B	2.400	2.755
C	2.750	2.900
D	1.800	1.990
Average	2.687	2.949

- It doesn't make much sense to assume independence between babies in the first and secon column.
- We may notice that second babies are always heavier; this has probability $1/16$ of happening under the null hypothesis of equal weights.

Paired comparisons (III)

- If weights had the same mean for babies of the same mother, difference of weight should have mean zero.

Mother	**First**	**Second**	**First−Second**
A	3.800	4.150	−0.350
B	2.400	2.755	−0.355
C	2.750	2.900	−0.150
D	1.800	1.990	−0.190
Average	2.687	2.949	−0.261

- This suggests one way of testing which accounts for dependence.

Paired comparisons (IV)

```
> First  <- c(3.80, 2.40, 2.75, 1.80)
> Second <- c(4.150, 2.755, 2.900, 1.990)
> t.test(x=First,y=Second)
        Welch Two Sample t-test

data:  First and Second
t = -0.43, df = 6, p-value = 0.7
alternative hypothesis: true difference in means i
95 percent confidence interval:
 -1.764  1.241
sample estimates:
mean of x mean of y
    2.688     2.949
```

Paired comparisons (V)

```
> t.test(x=First,y=Second, paired=TRUE)
        Paired t-test

data:  First and Second
t = -4.9, df = 3, p-value = 0.02
alternative hypothesis: true mean difference is no
95 percent confidence interval:
 -0.43094 -0.09156
sample estimates:
mean difference
        -0.2613
```

Sampling of independent observations

- We have been assuming samples

  $$X_1, X_2, \ldots, X_n$$

  made of independent observations.
- This makes sense:
  - When we sample an infinite population: seeing one value does not affect the probability of seeing the same or another value.
  - When we sample with replacement.
- With finite populations without replacement, what we see affects the probability of what is yet to be seen.

Finite versus infinite populations (I)

- With infinite populations, precision depends only on sample size.
- Usually, standard error of estimation is $\sqrt{\frac{\sigma^2}{n}}$ where n is sample size and σ^2 the population variance.
- If estimator is **consistent** we approach (but never quite hit with certainty) the true value of the parameter.

Finite versus infinite populations (II)

- If population is finite of size N, we could inspect all units and estimate anything with certainty:

  $$\hat{m} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

  would verify $m = \hat{m}$ if $n = N$.
- All parameters can, in principle, be known with certainty!
- With $n \neq N$,
  - If $n/N \approx 0$, independent sampling good approximation.
  - If $n/N \gg 0$, we have to take into account that we are looking at a substantial portion of the population.

An overview of things to come

We will see:

- What makes sampling without replacement more complex.
- What relationship there is among independent and non-independent sampling.
- What other types of sampling exist.

The central approximation

- Requirement: replacement or "large" population size N.
- If n is "large" and $X_1, \ldots, X_n$ "near" independent,

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(m, \sigma^2/n)$$

- Then,

$$\text{Prob}\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq m \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

Estimation of the population total

- Since $T = Nm$, we just have multiply by N the extremes of the interval for m.
- Hence,

$$\text{Prob}\left(N\overline{X} - Nz_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq T \leq N\overline{X} + Nz_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

Estimation of a proportion

- If X_i is a binary variable, $\overline{X}$ is the sample proportion.
- We have $\overline{X} \sim N(p, pq/n)$
- Usual estimate of variance is $\hat{p}(1-\hat{p})/n$.
- Sometimes we use a (conservative) estimate: $pq \leq 0.25$, hence a bound for σ^2 is $0.25/n$.

Sampling error with confidence $1 - \alpha$.

- From
$$\mathrm{Prob}\left(\overline{X} - z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}} \leq m \leq \overline{X} + z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$
we see that we will be off the true value m by less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with probability $1 - \alpha$.
- This is called the “$1 - \alpha$ (sampling) error”.
- “Sampling error” also used to mean standard deviation of the estimate.

Finding the required sample size n

- **Example:** What n do we need so that with confidence 0.95 the error in the estimation of a proportion is less than 0.03?
- **Solution:** Error is less than $z_{\alpha/2}\sqrt{\frac{\sigma^2}{n}}$ with confidence $1-\alpha$.
- Confidence 0.95 means $z_{\alpha/2} = 1.96$
- Want $0.03 > 1.96\sqrt{\frac{\sigma^2}{n}}$. Worst case scenario is $\sigma^2 = 0.25$.
- Therefore, $n > \frac{(1.96)^2 \times 0.25}{0.03^2} = 1067.11$ will do. Will take $n = 1068$.

Interesting facts (I)

- Under independent sampling required sample size depends only on variance and precision required.
- Questions like: "Is a sample of 4% enough?" are badly posed.
- $n = 400$ (4% of a population with $N = 10000$) insufficient to give a precision of 0.03 with confidence 0.95.
- … but $n = 3000$ (0.3% of a population with $N = 1000000$) vastly enough!

Interesting facts (II)

- As long as populations are large detail is expensive!
- To estimate a proportion in the CAPV with the precision stated requires about $n = 1068$.
- To estimate the same proportion for each of the three Territories with the same precision, requires three times as large a sample!
- Subpopulation estimates have much lower precision than those for the whole population.

Estimation of the mean (I)

- In independent sampling,

$$\begin{aligned} E[\overline{x}] &= E\left[\frac{X_1 + \ldots + X_n}{n}\right] \\ &= \frac{m + m + \ldots + m}{n} = \frac{nm}{n} = m \end{aligned}$$

- $E[X_i] = m$ irrespective of what other values are in the sample.
- Without replacement, distribution of X_i depends on what other values are already present in the sample.
- The same result as for independent sampling is true!

Estimation of the mean (II)

- **Theorem 1** In a finite population of size N with $m = \sum_{i=1}^{N} y_i/N$, for samples $Y_1, \ldots, Y_n$ without replacement of size $n < N$ we have:

$$E[\overline{Y}] = m$$

- **Proof**
  - $Y_1, Y_2, \ldots, Y_n$ are the elements of the sample.
  - $y_1, y_2, \ldots, y_N$ are the elements of the population.

Estimation of the mean (III)

- There are $\binom{N}{n} = \frac{N!}{(N-n)!n!}$ different samples.
- Of those, $\binom{N-1}{n-1}$ contain each of the values $y_1, y_2, \ldots, y_N$.
- Clearly,

$$\sum (Y_1 + Y_2 + \ldots + Y_n) = \binom{N-1}{n-1}(y_1 + y_2 + \ldots + y_N)$$

where the sum in the left is taken over all $\binom{N}{n}$ different samples. Dividing by $\binom{N}{n}$ finishes the proof.

Estimation of the mean (IV)

- Indeed,

$$\begin{aligned} \frac{\sum(Y_1+Y_2+\ldots+Y_n)}{\binom{N}{n}} &= \frac{\binom{N-1}{n-1}(y_1+y_2+\ldots+y_N)}{\binom{N}{n}} \\ &= \frac{n}{N}(y_1+y_2+\ldots+y_N) \end{aligned}$$

- Therefore,

$$E[\overline{Y}] = \frac{\sum(Y_1+\ldots+Y_n)/n}{\binom{N}{n}} = \frac{(y_1+\ldots+y_N)}{N} = E[\overline{y}] = m$$

The indicator variable method

- We have

  $$(Y_1 + Y_2 + \ldots + Y_n) = (y_1 Z_1 + y_2 Z_2 + \ldots y_N Z_N)$$

  where Z_i is a binary variable which takes value 1 if y_i belongs to a given sample.
- The probability of that happening is n/N. Then,

  $$E[(Y_1 + Y_2 + \ldots + Y_n)] = \frac{n}{N}(y_1 + y_2 + \ldots y_N),$$

  which again gives the previous result $E[\overline{Y}] = \overline{y} = m$.

Population variance an quasi-variance

- They are defined as:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N} \qquad \tilde{\sigma}^2 = \frac{\sum_{i=1}^{N}(y_i - \overline{y})^2}{N-1}$$

- Similarly for sample analogues:

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n} \qquad \tilde{s}^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

- Turns out some formulae are simpler in terms of quasi-variances.

Variance of $\overline{Y}$ (I)

- **Theorem 2** In a finite population of size N, the estimator $\overline{Y}$ of $m = \sum_{i=1}^{N} y_i/N$ based on a sample of size $n < N$ without replacement $Y_1, \ldots, Y_n$ has variance:

$$\mathrm{Var}[\overline{Y}] = \frac{\tilde{\sigma}^2}{n}\left(1 - \frac{n}{N}\right)$$

- Factor

$$\left(1 - \frac{n}{N}\right)$$

  usually called “finite population correction factor” or “correction factor”.

Variance of $\overline{Y}$ (II)

- **Remarks:**
- It is the same expression as in independent random sampling with i) σ^2 replaced by $\tilde{\sigma}^2$, and ii) corrected with the factor $(1 - n/N)$.
- If $n = N$, the variance $\text{Var}(\overline{Y})$ is 0 (why?).
- Formula covers middle ground between infinite populations ($n/N = 0$) and census sampling ($n/N = 1$).

Variance of $\overline{Y}$ (III)

- **Proof**

$$\begin{aligned}\mathrm{Var}(\overline{Y}) &= \mathrm{Var}\left(\frac{y_1 Z_i + \ldots + y_N Z_N}{n}\right)\\ &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \mathrm{Var}(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i} y_i y_j \mathrm{Cov}(Z_i, Z_j)\right]\end{aligned}$$

- We only need expressions for $\mathrm{Var}(Z_i)$ and $\mathrm{Cov}(Z_i, Z_j)$.

Variance of $\overline{Y}$ (IV)

- Since Z_i is binary with probability n/N,

$$\mathrm{Var}(Z_i) = (n/N)(1 - n/N).$$

- But $\mathrm{E}[Z_i Z_j] = P(Z_i = 1, Z_j = 1) = \frac{n(n-1)}{N(N-1)}$, so

$$\mathrm{Cov}(Z_i, Z_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 = -\frac{n(1-n/N)}{N(N-1)}$$

- Replacing in expression for $\mathrm{Var}(\overline{Y})$ will lead to result.

Variance of $\overline{Y}$ (V)

$$\begin{aligned}
\mathrm{Var}(\overline{Y}) &= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 \underbrace{\mathrm{Var}(Z_i)}_{(n/N)(1-n/N)} + \sum_{i=1}^{N}\sum_{j\neq i} y_i y_j \underbrace{\mathrm{Cov}(Z_i, Z_j)}_{-\frac{n(1-n/N)}{N(N-1)}}\right] \\
&= \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i} y_i y_j\right]
\end{aligned}$$

- Will rewrite expression in brackets.

Variance of $\overline{Y}$ (VI)

- Remark that,

$$\begin{aligned}\sum_{i=1}^{N}(y_i - m)^2 &= \sum_{i=1}^{N} y_i^2 - \frac{\left(\sum_{i=1}^{N} y_i\right)^2}{N} \\ &= \frac{N-1}{N}\left[\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}\frac{y_i y_j}{N-1}\right]\end{aligned}$$

- The expression in square brackets in th r.h.s is therefore $\frac{N}{N-1}\sum_{i=1}^{N}(y_i - m)^2$.

Variance of $\overline{Y}$ (VII)

- We are now done!

$$
\begin{aligned}
\mathrm{Var}(\overline{Y}) &= \frac{1}{n^2}\left(\frac{n}{N}\right)\left(1-\frac{n}{N}\right)\underbrace{\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i} y_i y_j\right]}_{\frac{N}{N-1}\sum_{i=1}^{N}(y_i-m)^2} \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{\sum_{i=1}^{N}(y_i-m)^2}{N-1} \\
&= \left(1-\frac{n}{N}\right)\frac{\tilde{\sigma}^2}{n}
\end{aligned}
$$

Sample size for given precision (I)

- The $(1-\alpha)$ error is

$$\delta = z_{\alpha/2}\sqrt{\frac{\tilde{\sigma}^2}{n}(1-n/N)}$$

- Solving for n we obtain

$$n = \frac{N z_{\alpha/2}^2 \tilde{\sigma}^2}{N\delta^2 + \tilde{\sigma}^2 z_{\alpha/2}^2}$$

- In terms of the variance, it can be written as:

$$n = \frac{N z_{\alpha/2}^2 \sigma^2}{(N-1)\delta^2 + \sigma^2 z_{\alpha/2}^2}$$

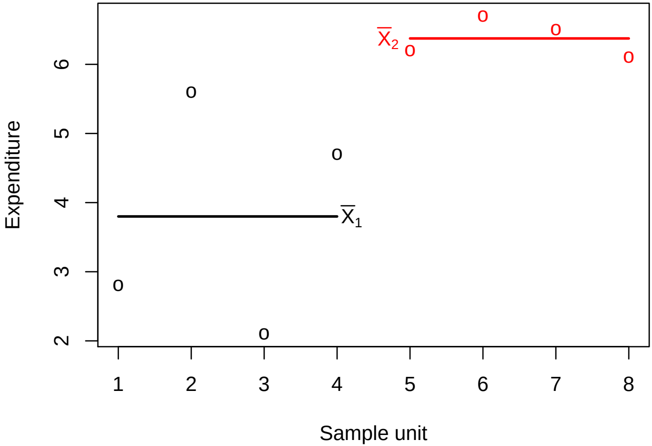
Sample size for given precision (II)

- $\tilde{\sigma}^2$ or σ^2 are required.
- We either replace an upper bound or conservative estimation for σ^2.
- Failing that, we estimate σ^2 or $\tilde{\sigma}^2$.
- Turns out $\tilde{s}^2$ is an unbiased estimate for $\tilde{\sigma}^2 \ldots$
- $\ldots$yet the difference between $\tilde{\sigma}^2$ and σ^2 or $\tilde{s}^2$ and s^2 is so small in practice that they are used interchangeably.

Why strata?

- Sometimes we know something about the composition of the population, knowledge that can be put to use.
- **Example:** We might know that males and females have different spending in e.g. tobacco or cosmetics.
- To estimate average spending, it makes sense to sample males and females, and combine the estimations.
- Sometimes, the target quantity might be similar, but the variance quite different. Also makes sense to differentiate.

Example 1



- Makes sense to estimate mean in each subpopulation

Definitions and notation

- We assume the population is divided in h strata. Total size is $N = N_1 + N_2 + \ldots + N_h$.
- The i-th stratum has a mean $m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$ and variance $\sigma_i^2 = \frac{1}{N_i} \sum_{j=1}^{N_i} (y_{ij} - m_i)^2$.
- Clearly,

$$\begin{aligned} m &= \sum_{i=1}^{h} \left(\frac{N_i}{N} \right) m_i \\ \sigma^2 &= \sum_{i=1}^{h} \frac{N_i}{N} \sigma_i^2 + \sum_{i=1}^{h} \frac{N_i}{N} (m_i - m)^2 \end{aligned}$$

Estimation of the mean

- The estimator of the mean when sampling without replacement the whole population has variance $\frac{\tilde{\sigma}^2}{n}(1 - n/N)$.
- Similarly, the estimation of the mean of each stratum has variance $\sigma_i^2 = \frac{\tilde{\sigma}_i^2}{n}(1 - n_i/N_i)$.
- The variance of the global mean reconstituted from the estimated means of the strata is

$$\sigma_*^2 = \sum_{i=1}^{h} \left(\frac{N_i}{N}\right)^2 \frac{\tilde{\sigma}_i^2}{n_i}(1 - n_i/N_i)$$

Does the estimation of *m* improve?

- Yes. If we sample each stratum in proportion to its size (i.e., $n_i/N_i = n/N$ for all i), then:

$$\frac{\tilde{\sigma}^2}{n}(1 - n/N) - \sigma_*^2 = \left(1 - \frac{n}{N}\right) \sum_{i=1}^{h} \left(\frac{N_i}{N}\right) \left[\frac{N_i - 1}{N - 1} - \frac{N_i}{N}\right] \frac{\tilde{\sigma}_i^2}{n_i} + \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_{i=1}^{h} \frac{N_i}{N - 1}(m_i - m)^2$$

- Marked Improvement when the m_i's very different.

Optimal allocation (I)

- It makes little sense to spend sampling effort for homogeneous strata.
- After all, if a stratum is perfectly homogeneous, looking at a single observation is enough.
- Let $w_i = N_i/N$. If we can spend C, we should minimize

$$F = \sum_{i=1}^{h} w_i^2 \frac{\tilde{\sigma}_i^2}{n_i} + \lambda \left(\sum_{i=1}^{h} c_i n_i - C \right)$$

- First term, variance neglecting finite population correction.
- Second term, restriction on total sampling cost, assuming c_i cost per unit sampled in stratum i.

Optimal allocation (II)

- Taking derivatives w.r.t. n_i ($i = 1, \ldots, h$) and equating to zero, we obtain

$$\frac{\partial F}{\partial n_i} = \frac{w_i^2 \tilde{\sigma}_i^2}{-n_i^2} + \lambda c_i = 0$$

- From that expression we get,

$$n_i \propto \frac{N_i \tilde{\sigma}_i}{N\sqrt{c_i}}$$

- Therefore enough to allocate n_i proportional to right hand side..
- Intuition: sample more **big** strata and **disperse** strata; sample less strata where sampling is relatively more **costly**.

Optimal allocation (III)

- But, how to determine n_i?
- We know

$$n_i = k \frac{N_i \tilde{\sigma}_i}{N \sqrt{c_i}}$$

- Further,

$$C = \sum_{i=1}^{h} c_i n_i = k \sum_{i=1}^{h} \frac{N_i \tilde{\sigma}_i \sqrt{c_i}}{N}$$

- Therefore,

$$n_i = \underbrace{\frac{NC}{\sum_{i=1}^{h} N_i \tilde{\sigma}_i \sqrt{c_i}}}_{\color{red}{k}} \times \frac{N_i \tilde{\sigma}_i}{N \sqrt{c_i}}$$

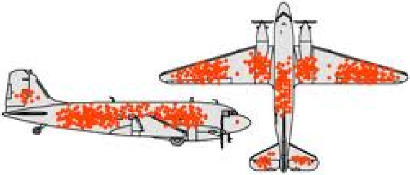
Abraham Wald on sample selection

Abraham Wald (1902-1950)

- Hungarian-born. Graduated (Ph.D. Mathematics) from University of Vienna, 1931.
- Fled to the USA in 1938, as Nazi persecution intensified in Austria.
- Important contributions to the war effort as statistician (notably sequential analysis)
- Was consulted about aircraft armoring.

What Wald saw that the others did not

- *Mark hits in B-29 bombers as they come back.*



- Pretty obvious! Will armor the most beaten areas.
- *I didn't tell you to do that!*
- Do you want us to protect the areas with no hits?
- *That's exactly what I suggest!*

Sample selection is ubiquitous!

- If you ask for volunteers in a field study, no chance you will get a truly random sample.
- Never do!
- Do not let the survey taker choose the units.
- A random sample is not a “grab set”.
- Build a census, randomize properly, address the chosen units and no others.

Methods of sampling that you should be aware of

- Multi-step sampling and conglomerate sampling
- Systematic sampling
- If you use systematic sampling (every n-th unit with random start), make sure no periodicities exist that will destroy randomness.

Convergence in distribution $X_n \xrightarrow{d} X$

- Means that *the distribution* of X_n approaches that of X.

Convergence in distribution $X_n \xrightarrow{d} X$

- Means that *the distribution* of X_n approaches that of X.
- Does not imply in any way that X_n approaches X.

Convergence in distribution $X_n \xrightarrow{d} X$

- Means that *the distribution* of X_n approaches that of X.
- Does not imply in any way that X_n approaches X.
- Central limit theorem (CLT) establishes convergence in distribution to a normal in many circumstances.

Convergence in distribution $X_n \xrightarrow{d} X$

- Means that *the distribution* of X_n approaches that of X.
- Does not imply in any way that X_n approaches X.
- Central limit theorem (CLT) establishes convergence in distribution to a normal in many circumstances.
- If

  $$\varphi_{X_n}(u) \longrightarrow \varphi_X(u)$$

  and $\varphi_X(u)$ is continuous at $u = 0$, then $X_n \xrightarrow{d} X$ (or, equivalently, $F_{X_n}(x) \longrightarrow F_X(x)$ in all continuity points of $F_X(x)$).

Converge in probability $X_n \xrightarrow{p} X$ (I)

- This is a different beast

Converge in probability $X_n \xrightarrow{p} X$ (I)

- This is a different beast
- It *does* imply that X_n approaches X

Converge in probability $X_n \xrightarrow{p} X$ (I)

- This is a different beast
- It *does* imply that X_n approaches X
- . . .but not in the usual “mathematics” sense.

Converge in probability $X_n \xrightarrow{p} X$ (I)

- This is a different beast
- It *does* imply that X_n approaches X
- …but not in the usual “mathematics” sense.
- When we say that e.g. $a_n = 1/n$ converges to 0 we mean that for large enough n, $a_n \approx 0$.

Converge in probability $X_n \xrightarrow{p} X$ (I)

- This is a different beast
- It *does* imply that X_n approaches X
- ...but not in the usual "mathematics" sense.
- When we say that e.g. $a_n = 1/n$ converges to 0 we mean that for large enough n, $a_n \approx 0$.
- No matter how small ϵ, n can always be found such that $|a_n - 0| < \epsilon$.

Converge in probability $X_n \overset{p}{\to} X$ (II)

- When we say that $X_n \overset{p}{\to} X$ we mean that *with arbitrarily large probability* (but not certainty!) $X_n \approx X$.

Converge in probability $X_n \xrightarrow{p} X$ (II)

- When we say that $X_n \xrightarrow{p} X$ we mean that *with arbitrarily large probability* (but not certainty!) $X_n \approx X$.
- Formally: when for all $\epsilon > 0, \eta > 0$ there is N such that when $n > N$ we have:

$$P(|X_n - X| < \epsilon) \geq 1 - \eta$$

Converge in probability $X_n \xrightarrow{p} X$ (II)

- When we say that $X_n \xrightarrow{p} X$ we mean that *with arbitrarily large probability* (but not certainty!) $X_n \approx X$.
- Formally: when for all $\epsilon > 0, \eta > 0$ there is N such that when $n > N$ we have:

$$P(|X_n - X| < \epsilon) \geq 1 - \eta$$

- This does not *guarantee* that $|X_n - X| < \epsilon$ for any n, but it makes it highly probable.

Converge in probability $X_n \xrightarrow{p} X$ (II)

- When we say that $X_n \xrightarrow{p} X$ we mean that *with arbitrarily large probability* (but not certainty!) $X_n \approx X$.
- Formally: when for all $\epsilon > 0, \eta > 0$ there is N such that when $n > N$ we have:

$$P(|X_n - X| < \epsilon) \geq 1 - \eta$$

- This does not *guarantee* that $|X_n - X| < \epsilon$ for any n, but it makes it highly probable.
- Tchebycheff's inequality is a simple way to show convergence in probability in many cases.

Convergence in probability via Tchebychev inequality (I)

- Consider the simple case where $Z_n \sim (m, \sigma^2/n)$. This happens for instance with the binomial frequency $(X_1 + \ldots + X_n)/n$ with $m = p$ and $\sigma^2 = pq$.

Convergence in probability via Tchebychev inequality (I)

- Consider the simple case where $Z_n \sim (m, \sigma^2/n)$. This happens for instance with the binomial frequency $(X_1 + \ldots + X_n)/n$ with $m = p$ and $\sigma^2 = pq$.
- Then (Tchebycheff),

$$P(|Z_n - p| < \underbrace{k\sqrt{pq/n}}_{\epsilon}) \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

Convergence in probability via Tchebychev inequality (I)

- Consider the simple case where $Z_n \sim (m, \sigma^2/n)$. This happens for instance with the binomial frequency $(X_1 + \ldots + X_n)/n$ with $m = p$ and $\sigma^2 = pq$.
- Then (Tchebycheff),

$$P(|Z_n - p| < \underbrace{k\sqrt{pq/n}}_{\epsilon}) \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

- Make your pick of $1 - \eta$ as close to 1 as desired; whatever the implied k, we only have to choose n large enough to make ϵ as small as we wish.

Convergence in probability via Tchebychev inequality (II)

- Clearly, nothing special about the binomial.

Convergence in probability via Tchebychev inequality (II)

- Clearly, nothing special about the binomial.
- Same thing will happen with any $Z_n \sim (m, \sigma_n^2)$ such that $\sigma_n^2 \to 0$.

Convergence in probability via Tchebychev inequality (II)

- Clearly, nothing special about the binomial.
- Same thing will happen with any $Z_n \sim (m, \sigma_n^2)$ such that $\sigma_n^2 \to 0$.
- Then,

$$P(|Z_n - m| < \underbrace{k\sigma_n}_{\epsilon}) \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

ensures $Z_n \xrightarrow{p} m$.

Convergence in probability via Tchebychev inequality (II)

- Clearly, nothing special about the binomial.
- Same thing will happen with any $Z_n \sim (m, \sigma_n^2)$ such that $\sigma_n^2 \to 0$.
- Then,

$$P(|Z_n - m| < \underbrace{k\sigma_n}_{\epsilon}) \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

  ensures $Z_n \xrightarrow{p} m$.
- Common mean and variance going to zero is a sufficient condition for convergence in probability (to the common mean) of a random sequence.

Convergence in mean square and almost surely

- We will hardly use them. Called “strong convergences” (convergence in probability is called “weak convergence”).

Convergence in mean square and almost surely

- We will hardly use them. Called “strong convergences” (convergence in probability is called “weak convergence”).
- Convergence in mean square: $X_n \overset{m.s.}{\to} X$ if

$$\lim_{n\to\infty} E|X_n - X|^2 = 0.$$

Convergence in mean square and almost surely

- We will hardly use them. Called “strong convergences” (convergence in probability is called “weak convergence”).
- Convergence in mean square: $X_n \overset{m.s.}{\to} X$ if

$$\lim_{n\to\infty} E|X_n - X|^2 = 0.$$

- Easy to see this implies converge in probability.

Convergence in mean square and almost surely

- We will hardly use them. Called “strong convergences” (convergence in probability is called “weak convergence”).
- Convergence in mean square: $X_n \overset{m.s.}{\to} X$ if

$$\lim_{n\to\infty} E|X_n - X|^2 = 0.$$

- Easy to see this implies converge in probability.
- Convergence almost surely: $X_n \overset{a.s.}{\to} X$ if

$$P(\omega : \lim_{n\to\infty} X_n(\omega) = X) = 1.$$

Convergence in mean square and almost surely

- We will hardly use them. Called “strong convergences” (convergence in probability is called “weak convergence”).
- Convergence in mean square: $X_n \overset{m.s.}{\to} X$ if

$$\lim_{n\to\infty} E|X_n - X|^2 = 0.$$

- Easy to see this implies converge in probability.
- Convergence almost surely: $X_n \overset{a.s.}{\to} X$ if

$$P(\omega : \lim_{n\to\infty} X_n(\omega) = X) = 1.$$

- It also implies weak convergence.

Books and Monographies I