Handout 3

Reminder

- If $X = X_1 + \ldots + X_n$ with all X_i i.i.d. $b(p)$, $X \sim b(p,n)$, $P_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ and $\varphi_X(u) = [q + pe^u]^n$.
- The sum of independent binomials *with equal* p is binomial.
- The ratio X/n is called the *binomial frequency.*
- If $X \sim N(0,1)$, $\varphi_X(u) = e^{u^2/2}$ (and, equivalently, $\psi_X(u) = e^{-u^2/2}$).
- From the definition of number e:

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a \qquad \lim_{n\to\infty}\left(1+\frac{a}{n}+o(1/n)\right)^n = e^a$$

  ($o(1/n) =$ "negligible as compared to $1/n$").

- Definition and properties of moment generating function:

$$\varphi_X(u) = 1 + \alpha_1(u) + \alpha_2 \frac{1}{2!}u^2 + \alpha_3 \frac{1}{3!}u^3 + \ldots$$

$$\varphi_{aX+b}(u) = E[e^{u(aX+b)}] = E[e^{(ua)X}e^{ub}] = e^{ub}\varphi_X(ua)$$

- To show that $X_n \xrightarrow{d} X$, enough to show that $\lim_{n\to\infty}\varphi_{X_n}(u) = \varphi_X(u)$ (with added requirement that $\varphi_X(u)$ continuous at $u = 0$). Similarly if you prefer to use the characteristic function $\psi_X(u)$.

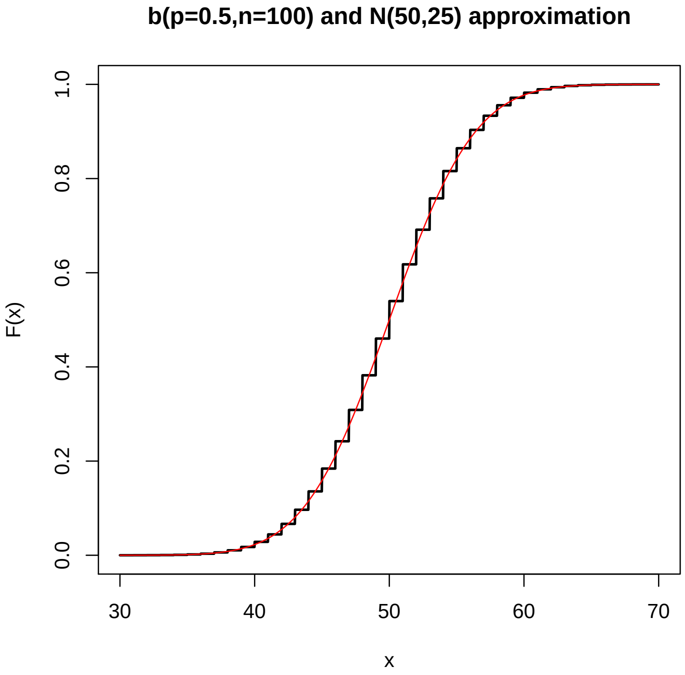
Binomial distribution, $b(p,n)$: normal approximation

- Use of tables or direct computation of $\binom{n}{x}p^x(1-p)^{n-x}$ awkward for large n, x.
- Since $X = X_1 + X_2 + \ldots + X_n$ we can use a normal approximation, invoking the central limit theorem.
- Easy to prove that $X \sim b(p,n) \xrightarrow{d} N(np, \sigma^2 = npq)$ as $n \to \infty$, or equivalently $\frac{X-np}{\sqrt{npq}} \xrightarrow{d} N(0,1)$.

$$\begin{aligned}
Z_n = \frac{X-np}{\sqrt{npq}} &= \underbrace{\frac{1}{\sqrt{n}}\frac{X_1-p}{\sqrt{pq}}}_{V_1} + \ldots + \underbrace{\frac{1}{\sqrt{n}}\frac{X_n-p}{\sqrt{pq}}}_{V_n} \\
\varphi_{Z_n}(u) &= \prod_{i=1}^{n}\varphi_{V_i}(u) \\
&= \prod_{i=1}^{n}\left(1+\frac{1}{2!}\left(\frac{u}{\sqrt{n}}\right)^2 + o(1/n)\right) \\
\lim_{n\to\infty}\varphi_{Z_n}(u) &= \lim_{n\to\infty}\left(1+\frac{1}{2!}\left(\frac{u}{\sqrt{n}}\right)^2 + o(1/n)\right)^n \\
&= e^{\frac{u^2}{2}}
\end{aligned}$$

Notice that the same reasoning applies to iid variables ($m = 0, \sigma^2 = 1$), not necessarily binomial.

- Use of approximation: n moderately large, np "away" from zero. (Why?)



Jump at $x = k$ approximated by increase of continuous distribution from $k - \frac{1}{2}$ to $k + \frac{1}{2}$.

- Continuity correction.

$$P(a \leq Z \leq b) \approx \Phi\left(\frac{b + \frac{1}{2} - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - \frac{1}{2} - np}{\sqrt{npq}}\right)$$

where $\Phi(x)$ is the cumulative distribution function of the $N(0, 1)$.

- Will have an alternative approximation with Poisson distribution.

Problems

(At this point, we will use a normal approximation when needed, even it it is not the best option. We will solve again some of these problems using another approximation (Poisson) and compare results.)

1. Assume you have a regular coin ($P(\text{heads}) = P(\text{tails}) = 0.5$).
   (a) What is the probability that you get 6 or more heads in 10 throws?
   (b) What is the probability that you get 60 or more heads in 100 throws?
   (c) What is the probability that you get 600 or more heads in 1000 throws?

2. A common question to anyone doing some sort of statistical consultancy: "Sir, 60% of my patients taking XXX develop cold, as compared to only 50% of my patients not taking XXX. Should I conclude that XXX predisposes to cold?". Explain why this question, as it stands, does not provide enough information for a sensible answer. Relate your answer to the previous problem.

3. There are to candidates A and B running for office in an upcoming election. To estimate the proportion of people who will vote for A, you interview a sample of 1000 randomly chosen individuals, out of which 550 declare for A. What is the probability of getting 550 A-voters or more out of 1000 if, in fact, only 45% of the people are willing to vote for A?

4. You sell tickets for a plane with 340 seats. You know from experience that 15% of the people who buy a ticket never board the plane, because of last minute problems, late connections, etc.
   (a) If you sell 355 tickets, what is the expected number of people showing up at the boarding gate? What is the probability that you will not have enough room?
   (b) How many tickets can you sell if you want that the probability of not being able to accommodate all passengers be less than 0.01?

5. An insurance company specializes in fire risks. They charge a premium of 500€ per year per house. The probability that in a year a house catches fire, is 0.002, in which case the indemnity the insurance company has to pay is 200.000€. They have insured 10000 houses.
   (a) What is the expected gross profit (excess of premiums over the cost of claims) per year? What is the probability that they incur a loss?
   (b) Assume that the company enters a reinsurance agreement with another similar company (same number of houses insured, same premium, same indemnity in case of fire). They agree to share all premiums and claims 50% each. What is now the expected gross profit and probability of loss for the first company?

Reading

[3] § 7.3 and 7.4, or [4], Chapter 25. Many other books cover these topics. Problem 2 is adapted from [1]. problem 3 from [2].

References

[1] R.B. Ash. *Basic Probability Theory.* Dover Pub., 1970.

[2] A. Garín and F. Tusell. *Problemas de Probabilidad e Inferencia Estadística.* Ed. Tébar-Flores, Madrid, 1991.

[3] J. Martín Pliego and L. Ruiz-Maya. *Estadística I : Probabilidad.* Ediciones AC, 2004. In the reserved collection, signature AL-519.2 MAR.

[4] A. Fz. Trocóniz. *Probabilidades. Estadística. Muestreo.* Tebar-Flores, Madrid, 1987.

Pautas docentes

- Estos problemas están resueltos con R en el handout siguiente, sobre el que no suelo trabajar en clase para ahorrar tiempo (lo reparto, simplemente).

- En lugar (o además) de darles reglas misteriosas como que la aproximación normal funciona cuando $np > 18$, creo que es conveniente hacerles ver el motivo. Con np pequeño, los valores de la binomial con máxima probabilidad están en las cercanías de cero y es *muy* asimétrica: no puede ser bien aproximada por una normal, que es simétrica, y que situando su máxima densidad cerca de cero daría probabilidad ≈ 0.5 a valores en el semieje negativo (que no pueden aparecer en la binomial).

- Habiendo tiempo se pueden mostrar con facilidad ejemplos de las funciones de distribución binomial y normal aproximante para diferentes valores de p y n. El código empleado para la gráfica más arriba es

  ```
  > ############## Parametros ajustables ###############
  > p <- 0.01
  > n <- 50
  > ###################################################
  > me  <- n*p
  > s2 <- n*p*(1-p)
  > s  <- sqrt(s2)
  > b <- function(x) { pbinom(floor(x),p=p,size=n) }
  > g <- function(x) { pnorm(x,m=me,sd=s) }
  > curve(b,from=me - 4*s, to=me + 4*s, n=2000, xlab="x",
  +       ylab="F(x)", lwd=2,
  +       main=paste("b(p=",p,",n=",n,") and  N(",me,",",
  +                  round(s,5),") approximation",sep=""))
  > curve(g,from=me-4*s, to=me+4*s, n=2000, add=TRUE, col="red")
  ```

  Puede probarse con diferentes valores de p y n y ver la degradación/mejora

- El ejercicio 4 ha sido ya muy utilizado, resulta motivador y lo entienden bien. La parte (b), quizá la más “real life” y que más les interesa, puede hacerse pesada a mano; quizá aquí conviene anticiparles la solución en R del proximo handout para no invertir mucho tiempo.

- En el problema 5, la parte (b) merece algún comentario: todo lo demás siendo igual (prima, probabilidad de incendio) cuantas más unidades se aseguren mas “determinista” se vuelve el resultado y (en este caso) menor es la probabilidad de entrar en pérdidas. Relacionarlo con el problema 1.