Comments on hypothesis testing (13 April 2020)

Last week we introduced some concepts on hypothesis testing. Before we tackle some problems, we need to introduce some new concepts and ideas. We have so far described the problem of hypothesis testing as dealing with the choice between two competing hypothesis, conventionally called H_0 (the "null" or *statu quo* hypothesis: what we are willing to accept if there is no evidence to the contrary) and H_a (the "alternative" or competing hypothesis: what we would be prepared to entertain if H_0 seems to be at variance with the evidence).

Pure significance tests

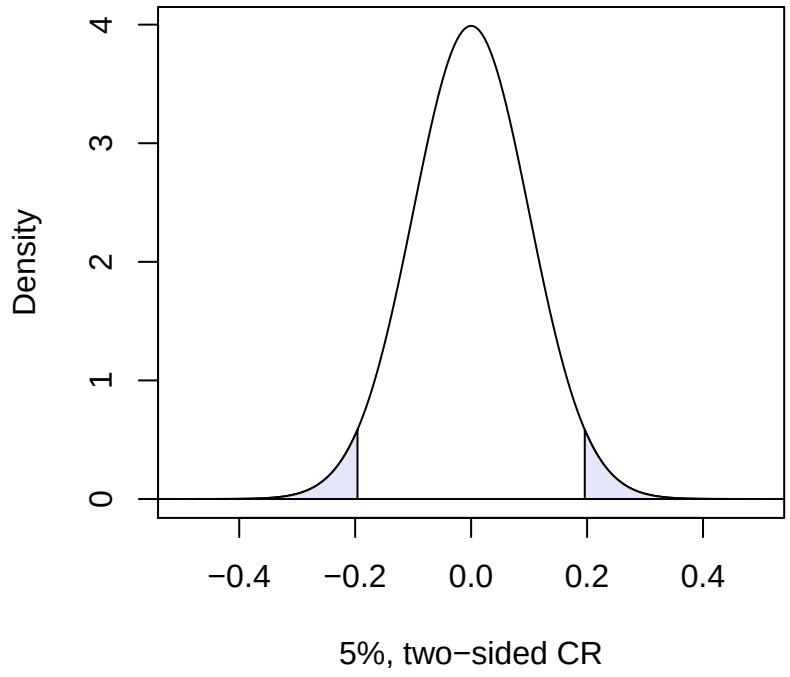
Some times (see slide **Pure significance tests**) we only have a null hypothesis H_0: we want to test whether the evidence is compatible with a certain H_0 with no clear idea of alternatives.

For instance, we would like to test that measurements of an instrument have zero bias (=measurement errors have mean zero). If this is not true, it could be that the instrument underestimates the true values or else that it overestimates; and we have no reason to believe that the failure of the instruments should be in any specific direction.

When this is the case, we compute the likely region for the test statistic and anything outside is made the critical region. Assume for the time being that we *know* that measurements are normally distributed with variance 1. With a sample size of only one observation ($n = 1$) we would reason as follows: "If indeed $X \sim N(0, \sigma^2 = 1)$, 95% of the time we should obtain an observation between -1.96 and 1.96. An observation outside that interval is "rare" (happens only 5% of the time), hence evidence against H_0."

The critical region is shown in lavender in Figure 1. Not having any preconceived ideas about the way our instrument could fail to be exact, we consider errors of measurement too large, whether positive or negative, as evidence against H_0. The critical region CR is made of the two tails and α is the surface in blueish, lavender color.

Figure 1: Pure significance test of $H_0 : X \sim N(0,1)$ with no specified alternative, sample size $n = 1$ and $\alpha = 0.05$



Testing against a pre-defined alternative

Sometimes we know that if H_0 fails to be true, it must be *because* another H_a is true. For instance, installing a certain type of insulation in the walls of a house may or may not lead to a reduction in energy consumption, but certainly will not produce an increase: however effective, the additional layer of insulation can only *reduce* the loss of heat, never increase it.

Suppose we measure energy consumption in two neighbouring houses, the first with insulation, the second without, but otherwise identical. Assume that in the former energy consumption is distributed as $N(m_1, \sigma^2 = 1/2)$ and in the second as $N(m_2, \sigma^2 = 1/2)$. (We are making unwarranted assumptions here, like normality and $\sigma^2 = 1/2$; we shall have occasion to lift such assumptions later.)

The hypothesis the there is no improvement in energy consumption can be stated as $H_0 : m_2 = m_1$ or equivalently $H_0 : m_2 - m_1 = 0$. Since we know that

the added layer of insulation cannot possible increase energy consumption, if H_0 fails it *must* be because $H_a : m_2 - m_1 > 0$ is true.

If we take one observation from each of the houses and compute $Z = X_2 - X_1$, then $Z \sim N(0, 1)$ under H_0 (you now see why we took $\sigma^2 = 1/2$ above; so that Z will have variance 1).

Hence, one might be tempted to revert to the situation illustrated by Figure 1 and reject H_0 if Z is in either tail. This would be wrong; for $Z < -1.96$ would be a rare observation under H_0, but even rarer if $H_a : m_2 - m_1 > 0$ is true. It is only Z large and positive that we can take as evidence against H_0 and in favor of H_a, the only alternative we are prepared to entertain.

Otherwise said, we will not consider a two-sided critical region as in Figure 1, but rather a one sided critical region made of the right-hand tail of size $\alpha = 0.05$ (refer to slide **Testing against an alternative** H_a).

You can easily modify the reasoning to cover a case in which only negative values of Z would be taken as evidence against the null and the critical region would be the left-hand tail.

Most powerful tests

Taking the critical region on one side or the other, as required in each case, makes a lot of sense and is intuitively easy to grasp. The question we have to ask ourselves is: "In what direction should the test statistic deviate when the null is not true?". If the answer is "To the right", we will place there the critical region, and likewise if the answer is "To the left".

Now two things should be clear. First, that acting in the way described we are maximizing the power, for we are maximizing the probability the the test statistic falls in the critical region. Second, that his "common-sense" approach to the placement of the critical region may not be enough in complex situations in which the way to proceed may not be obvious. We will introduce for such situations a principled approach, the Neyman-Pearson theorem, which will guide us in the construction of the more powerful test against a given alternative.

The Neyman-Pearson theorem. Motivation

In keeping with our usual method, we will try first to motivate the theorem and provide the intuition behind it, then present the theorem more formally.

Table 1: Distribution of X under two possible hypothesis

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

Let's consider a situation as depicted in slide **The Neyman-Pearson theorem (I)**; we reproduce the small table giving the probabilities of X under two possible hypothesis:

Consider the simple case in which we are required to decide between $H_0 : \theta = \theta_0$ versus $H_a : \theta = \theta_a$. We can take a single observation X and are required to choose between H_0 and H_a. This amounts to choosing a critical region, CR, so that when $X \in CR$ we will reject H_0 and decide for H_a.

We choose first $\alpha = 0.05$; remember that H_0 is always the "default" or *statu quo* hypothesis: what we believe unless given evidence to the contrary. Typically, rejection of H_0 is costly, so we set α, the probability of incorrect rejection (or type I error), small.

If we set $CR = \{4, 5\}$, then $\alpha = 0.05$ as required, for that is the probability of obtaining $X \in CR$ under H_0. But there are other possibilities: $CR = \{3, 5\}$ or $CR = \{2\}$.

Now, with all these CR's having the same $\alpha = 0.05$, we would prefer the one with smallest β (= type II probability error) or, equivalently, largest $1 - \beta$ (= power).

Notice that we want to include in the critical region points which have large probability under H_a. If we include, for instance, 2, the probability of the CR when H_a is true (= power) increases by 0.10. If we consider $CR = \{3, 5\}$, the power will be $0.25 + 0.10 = 0.35$. It is still better if we take the $CR = \{4, 5\}$, for in that case the power is $0.30 + 0.10 = 0.40$.

In this toy problem we have been able to look at all possible critical regions of size[1] $\alpha = 0.05$ and pick the one with the largest power. Clearly, in a real problem (specially with continuous random variables) this is not a feasible way to proceed.

[1] "Size" is the significance level or α.

Table 2: Quantities of potatoes available from suppliers A to F and their prices

Supplier	A	B	C	D	E	F
Asking price € per Kg.	0.10	0.15	0.10	0.25	0.30	0.12
Kgs. per euro	10	6.66	10	4	3.33	8.33
Kgs. available	3000	2000	1500	1500	1000	1000
Total value	300	300	150	375	300	120

The Neyman-Pearson theorem is about buying potatoes

Suppose you are charged with the duty of buying a large quantity of potatoes for an institution, as much as you can with the available budget of 1000€. Suppose, for the sake of simplicity, that all potatoes are completely homogeneous. When you go to the market, you find suppliers ready to supply small quantities, so you have to buy from several of them to complete your task. Suppose further that the available quantities from each supplier and the asked prices are as in the Table 2 (where some information is redundant).

One possible way of spending your money would be to compute all possible ways to spend 1000€ picking at most "Total value" from each supplier, and then see which of these combinations whose total cost is 1000€ yields you more potatoes.

Of course in real life you would never do that! Rather you would go to the suppliers with the lowest price (A and C) and purchase as much potatoes they can offer: this is 4500 Kgs. worth 450€. With your remaining money, you would go to F and spend 120€ to get you a further 1000 Kg. of potatoes. At his point you would still have 430€ in cash. Your next visit would be B, where you would spend 300€ to get a further 2000 Kgs. and your final 130€ would be spent with D, buying $130/0.25 = 520$ Kgs. at the much heftier price of 0.25€.

In all, you would end up with $4500 + 1000 + 2000 + 520 = 8020$ Kg. of potatoes and zero cash.

It is quite obvious that you have made the best possible use of your money, as you have never bought from a supplier until cheaper sources were exhausted. All you had to do is to visit first suppliers offering lower prices —which is the same as saying that they give more Kgs. per unit of money.

With this in mind, let us revisit Table 1 which we will complete now with an additional line giving the ratio of the second row to the first (see Table 3). Remember that when we include one point in the critical region,

Table 3: Distribution of X under two possible hypothesis

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
(Row 2):(Row 1)	0.166	0.577	2.00	6.25	7.5	10

the probabilities in the second row are the increase in power while those in the first row give the increase in α. The third line can then be interpreted as "the amount of power we get per unit of α 'spent' when we include one point in the critical region."

If we want a critical region of size $\alpha = 0.05$, the metaphor of the potatoes tells us what to do: keep adding points to the critical region until we "spend" our 0.05 units of α; and do so in order, starting by the points that give us more power per unit of α (already visible in the third row).

We would go on to "buy" power including first point 5 (for an expenditure of 0.01 units of α). With the remaining 0.04 units of α we could "buy" for our critical region the power offered by point 4 or point 3; clearly 4 offers a better deal, with 7.5 units of power per unit of α spent. When we "buy" that power, we would have our 0.05 of α spent, and our critical region would be $\{4, 5\}$ with a total power of 0.40; and that's it.

If you pause now for a moment to think you will realize that what you have done is to select for your critical region points verifying:

$$\frac{P(x;\theta_a)}{P(x;\theta_0)} \geq 7.5 \tag{1}$$

i.e. giving at least 7.5 units of power per unit of α "spent". It should be obvious that you can get higher power only by "spending" more α. This nicely illustrate the compromise the statistician has to face when designing a hypothesis test: he has to trade α for β.

If you now go to slide **The Neyman-Pearson theorem (II)** you will realize that it merely states the same than equation (1). The next two slides, **The Neyman-Pearson theorem - Proof (I)-(II)** offer a semi-formal proof which will convince you, if need be, that the rule given by equation (1) is completely general.

Next come three slides with examples, **Neyman-Pearson example (I)-(III)**. Notice that in all cases the Neyman-Pearson theorem gives the *shape* of the critical region, but no indication about the value of k_α: this we have to determine ourselves, and depends on the α we want.

The final three slides in this block, **Neyman-Pearson and sufficiency (I)-(III)** point to a connection to sufficiency. We found that there are in some cases statistics which squeeze all available information from a sample, so we should pick our estimators and test statistics as functions of these sufficient statistics. We saw that the MLE (maximum likelihood estimator) had in some sense "built in" sufficiency. The same happens with the likelihood ratio used in the Neyman-Pearson method of finding most powerful tests: this should be reassuring.