# Comments on hypothesis testing (6 April 2020)

## "Hard" logic

Slides **Logically equivalent statements (I)-(II)** introduce one simple fact: $p \Rightarrow q$ (read "$p$ implies $q$", where $p$ and $q$ are any two clauses) is logically equivalent to $\neg q \Rightarrow \neg q$ (read "not $q$ implies not $p$").

The second statement is called the *contra-positive* of the first, so what we are saying is that any statement and its contra-positive are both true or both false. You have examples in the first slide.

To show the truth of an statement, no amount of concordant evidence will do. In other words, if we want to show that whales live in the water, no number of whales that we actually see living in the water will suffice to establish the statement conclusively as true: it could always be the case that some whale, unseen to us, lives in the woods.

However, if we find a single whale not living in the water, *then* we can claim the statement "whales live in the water" to be false. This whale not living in the water is what we call a counterexample to the statement "All whales live in the water."

## "Soft" logic

We are concerned with random phenomena, in which the relationship between things we observe is not fully reproducible. (Remember that lecture, some weeks ago, when we dealt with this at length?)

In the realm of random phenomena things are not so clear-cut as in what we have called "hard logic". That a coin is regular does not imply that in any series of throws it gives about 50% of "heads". We may find (quite rarely) a series of throws where "heads" occur much more frequently than 50% (or much less frequently). But we can say (slides **Statements probabilistically related (I)-(II)**) that **most of the time** the relative frequency of "heads" is close to 50%.

The "contra-positive" of this mild statement is that when the percentage of "heads" is far from 50%, it is unlikely that the coin is regular: it does not preclude that the coin be regular, but makes us doubt it.

A percentage of "heads" far from 50% would be seen as a "soft counterexample" to the regularity of the coin.

## Hypothesis testing: generalities

This is the kind reasoning we apply when testing hypothesis:

1. We work out what should be the expected result of an experiment, if the hypothesis under test were true (for instance, we would expect about 50% of "heads" if the coin is regular).

2. We perform the experiment (throw the coin many times and compute the percentage of "heads").

3. If what we obtain departs substantially from what we would expect to find most of the time (as would be the case if the percentage of "heads" were quite different from 50%), we claim that our hypothesis is false.

This is further elaborated in slides **Hypothesis testing (I)-(III)**.

Notice that we are bound to make mistakes; for, even if our hypothesis is true, every once in a while we will find an abnormal result in our experiment which will make us think that it is not. We call this "type I error" (rejection of a hypothesis which is true), and its probability is $\alpha$, the *significance level*.

This is not the only mistake we can make; sometimes, our hypothesis will be false, and yet the experiment designed will fail to show a result abnormal enough to make us reject the hypothesis. Then we will incur in a "type II error" (failure to reject a hypothesis which should be rejected), the probability of which we call $\beta$.

## Hypothesis testing concepts and notation

Slides **The anatomy of a hypothesis test (I)-(V)** introduce some notation and additional concepts. One of them is *critical region*: the set of values of the test statistic which we consider "rare" and therefore lead to the rejection on the hypothesis under test.

Two comments are worth making here: if we perform our hypothesis test looking at the value of a test statistic, we better do it in such a way that we do not loose information: almost invariably we will want the test statistic to be a sufficient statistic (cf. previous slides on sufficiency).

Another comment is that we can always have the type I probability error $\alpha$ equal to zero. If we decide *never* to reject the hypothesis under test, we will never incur the error of rejecting it incorrectly! But of course this is not satisfactory, what we want is no incorrect rejections, but with the possibility of rejection when it is adequate.

## Hypothesis testing design trade-offs

This leads us to the last two slides for the present lecture: slides **The trade-off between Type I and Type II errors** and **Trade-off between Type I and II errors - Illustration**. We have to possible extreme decisions:

1. Set an empty critical region, i.e. our test statistic will *never* be there and we will *never* reject the hypothesis under test. Then $\alpha = 0$ (there is zero probability of incorrect rejection, for we do not reject at all!) but $\beta$ may be very large.

2. Set a critical region which includes all possible values of the test statistic. Then, the test statistic will *always* be in the critical region, we will always reject and therefore *never* fail to reject the hypothesis under test. Then $\beta$ will be zero (we cannot fail to reject an incorrect hypothesis, because we always reject, whether appropriate or not!); but $\alpha$ may be very large.

Between these two extremes, we have all intermediate choices. As we enlarge the critical region, $\alpha$ will grow and $\beta$ will decrease. Ideally, we would choose among the feasible values of $\alpha$ and $\beta$ the pair which minimizes the total cost of error –whether type I or type II error. This may be very complex, so a simpler approach is used: set $\alpha$ to some value which seems acceptable, then for said $\alpha$ minimize $\beta$ (or equivalently maximize the power $1 - \beta$). This leads to *most powerful tests* and the recipe to construct them, the Neyman-Pearson theorem, which is the topic we will address in the next lecture.