1.1.8 Hypothesis tests: a review of basic concepts

Some of the very basic concepts and ideas about hypothesis tests can be reviewed by means of simple examples, without dwelling into likelihood theory. This is the purpose of this section. In the next section, we will present the theoretical details of likelihood inference along with detailed examples of biological relevance. These lectures will be the founding blocks of the rest of our course.

Fisher's tea lady:

In R.A. Fisher's experimental designs book there is a ten pages account of an experiment where he basically laid out the most important principles of experimentation. The experiment is known as "Fisher's tea lady experiment". This experiment was also later described by Fisher's daughter, who wrote his biography. A pdf file of this textbook fragment is posted in the course web page. This account tells the story of a lady that claimed to be able to distinguish between a tea cup which was prepared by pouring the tea first and then the milk and another tea cup where the milk was poured first. Fisher then wonders if there is there a good experiment that could be devised in order to formally test the lady's claim. The null hypothesis of this purported experiment would then be that the lady has no selection ability whatsoever. A logical experiment would consist of offering the lady a set of "tea-first" cups and another set of "milk-first" cups and let her guess the tea cup type (milk-first or tea-first) of each one. The question is -Fisher noted- that it is not evident how many of each type and in what order shall this be done in order to carry a convincing experiment. Fisher begins by noting that, the more cups are offered to the lady, the harder it is to achieve a perfect classification of all the tea cups. Also, note that by giving her the same number of tea-first cups than milk-first cups we would allow each of the 2 types to get the same simultaneous presentation (*i.e.* opportunity to be chosen). Suppose that we ask the lady to select 4 milk-first from a total of 8 cups (That is, we offer her 4 milk-first and 4 tea-first cups). In how many ways can she make the 4 choices? Fisher noted that for the first cup there are 8 choices, for the second there are 7 choices, 6 choices for the third and finally, 5 choices for the fourth milk-first cup. Therefore, this succession of choices can be made in $8\times7\times6\times5 = 1680$ number of ways. But this takes into account not only every possible set of 4, but also every possible set in every possible order. Now, 4 objects can be arranged in order in $4\times3\times2\times1 = 24$ ways and therefore, since the 4 cups are assumed to be identical in every respect and we do not care about the order in which these 4 cups were given, then the number of ways of picking 4 cups out of 8 is

$$\begin{aligned}\frac{\#\text{ number of ways of assigning 4 cups as milk-first among the 8 cups}}{\#\text{ of ways that 4 cups can be ordered}} &= \frac{8\times7\times6\times5}{4\times3\times2\times1}\\ &= \frac{8!}{4!(8-4)!},\end{aligned}$$

which is $\binom{8}{4} = 70$. So if the lady was picking purely at random and didn't have any distinguishing ability whatsoever, she would have a probability of 1/70 of picking

up a particular sequence of cups assigned by her as milk-first that happens to be the correct one. What if we set 3 milk-first cups and 3 tea-first cups? Then, since $\binom{6}{3} = 20$ the lady would have a 1/20 probability of picking the correct sequence just by chance. Fisher decided to go for the harder test and decided to give her 4 milk-first and 4 tea-first cups. Now that we have decided on the number of cups, we can compute the probability of each possible outcome if the lady was picking purely at random (that is, if she had no ability to distinguish between a tea-first and a milk-first cup). The possible outcomes of the experiment are the following: the lady could pick 4 right out of the 4 of one type and therefore get 0 wrong out of the other type. We will denote this event $4R/0W$. She could also get three right of the first type and one wrong of the second type. This event will be denoted by $3R/1W$. According to this notation scheme, the other possible events are $2R/2W$, $1R/3W$ and $0R/4W$. Computing the probabilities of each of these events is a straightforward counting exercise. Considering the event $3R/1W$ for instance, we note that there are $\binom{4}{3}$ number of ways of picking 3 right out of 4 of the first type and independently of that, there are $\binom{4}{1}$ ways of choosing 1 wrong out of the other 4 cups of the second type. Iterating this argument for the other events we get that,

$$P(3R/1W) = \frac{\binom{4}{3} \times \binom{4}{1}}{\binom{8}{4}} = \frac{16}{70}.$$

Likewise,

$$P(4R/0W) = \frac{\binom{4}{4} \times \binom{4}{0}}{\binom{8}{4}} = \frac{1}{70},$$

$$P(2R/2W) = \frac{\binom{4}{2} \times \binom{4}{2}}{\binom{8}{4}} = \frac{36}{70},$$

$$P(1R/3W) = \frac{\binom{4}{1} \times \binom{4}{3}}{\binom{8}{4}} = \frac{16}{70},$$

and

$$P(0R/4W) = \frac{\binom{4}{0} \times \binom{4}{4}}{\binom{8}{4}} = \frac{1}{70}.$$

These probabilities completely specify the probability mass function of the outcomes of the experiment where the picking was done purely at random, that is, assuming that the lady has no detection ability. Therefore, this is the distribution of outcomes under the null hypothesis. Suppose that the experiment is carried and the lady picks 3 right of the first type and 1 wrong of the second type ($3R/1W$). Is this evidence enough to convince ourselves that she is not picking the cups at random and that she indeed has a detection ability? So we ask ourselves, if the null hypothesis is correct and the lady is picking purely at random, how unlikely it is to get an outcome as extreme or more more than the one we actually observed. This amounts to specify the probability of making only one error or less by pure dumb luck. According to the calculations above, that probability is

$$P(3R/1W) + P(4R/0W) = \frac{17}{70} \approx 0.24.$$

So if the null hypothesis is true, then there is a chance that we would have observed a choice as good or better than the one we saw about 24% (about a fifth) of the time! That chance is way too big to convince our skeptic (Fisher) that his null hypothesis is wrong. 0.24 is in fact, the p-value of the test of the lady's claim. Compare that value to what we are used to think of what a good skeptic's convincing threshold is: 0.05 (or 5% of the time). Hence, here we blatantly failed to reject the null hypothesis! Fisher's account is important in many ways and the most notable is the description of the value of randomization in experimentation (which I explicitly left out in here) as well as his careful elaboration of the logics of hypothesis testing.

Exercise 1.1. Suppose we ask the lady to select 4 milk-first from a total of 8 cups (that is, we offer her 4 milk-first and 4 tea-first cups). In how many ways can she make the four choices? Fisher noted that for the first cup there are 8 choices, for the second there are 7 choices, 6 choices for the third and finally, 5 choices for the fourth milk-first cup. Therefore, this succession of choices can be made in $8\times7\times6\times5 = 1680$ number of ways. But this takes into account not only every possible set of 4, but also every possible set in every possible order. Now, 4 objects can be arranged in order in $4\times3\times2\times1 = 24$ ways and therefore, since the 4 cups are assumed to be identical in every respect and we do not care about the order in which these 4 cups were given, then the number of ways of picking 4 cups out of 8 is

$$\begin{aligned}\frac{\#\text{ number of ways of assigning 4 cups as milk-first among the 8 cups}}{\#\text{ of ways that 4 cups can be ordered}} &= \frac{8\times7\times6\times5}{4\times3\times2\times1}\\ &= \frac{8!}{4!(8-4)!},\end{aligned}$$

which is $\binom{8}{4} = 70$. So if the lady is picking purely at random, she can assign four cups as "milk-first" in 70 different ways.

1. What is the probability that the lady doesn't make any mistake and correctly chooses the 4 milk-first cups?

2. Had Fisher given her 3 milk-first cups and 3 tea-first cups, what would the probability of correctly picking the 3 milk-first cups had been?

3. In fact, when he was thinking how to design the experiment, Fisher chose the number of cups after computing the probability of making no mistakes in two cases: when she is given to select 4 cups out of a total of 8 and when she is given 3 cups out of a total of 6. Given your answers to the two questions above, which number of cups do you think Fisher picked: 4 and 4 or 3 and 3? Why?

4. Enumerate all the possible outcomes of the experiment when a total of 8 cups are given to her (4 of each type). Hint: for instance, one outcome is as follows: she can pick 4 right out of the 4 of one type and therefore get 0 wrong out of the other type. Denote this event as $4R/0W$, where R stands for 'right' and W for 'wrong'. Use the same notation for all the other events.

5. Compute the relative frequency with which every single one of these possible outcomes occurs.

Hypothesis test for the sample mean (known variance):

Suppose that an education researcher suspects that college students at UF have a higher IQ than the population at large. The average IQ score from the population at large is 100. Because the IQ score can be though as a continuous phenotypic trait, to model its distributional properties this researcher should use a continuous random variable. In particular, here we'll use a Normal distribution, which is symmetric around the mean. Now, suppose that the standard deviation σ of the IQ scores distribution is known and equal to 15. To confront his suspicion with data, the researcher takes a *random sample* of $n = 30$ IQ tests from the population of UF students and obtains a sample mean score $\bar{x}$ equal to 105.3. A colleague of the education researcher is very skeptic of this suspicion and in fact, tells him that an IQ sample mean of 105.3 is not really an unlikely outcome if these 30 samples really came from a population of scores that is normally distributed around a mean $\mu = \mu_0 = 100$. The value $\mu_0 = 100$ embodies the skeptic's point of view, it corresponds to his hypothesized value of the mean of the distribution of IQ scores from which our random sample was drawn. In statistical terms, this is called the *null* hypothesis. Conducting a hypothesis test in this case amounts to convincing the skeptic that the researcher's suspicion (that is, the *alternative* hypothesis) that $\mu > 100$ is indeed supported by the data. In response to his colleague's questioning, the researcher starts by asking himself how unusual a sample mean of 105.3 would be if it really came from the IQ distribution of the population at large. Repeated *independent random sampling* from the population at large of IQ scores generates a series of sample means. Each time a sample of IQ scores is taken, a new sample mean is obtained. Thus, the computed sample mean IQ score can be considered as the outcome of a random variable. Let's denote this random variable $\bar{X}$ (remember that capital letters in this notes denote a random variable, unless otherwise specified). From the Appendix review (and a bit of common sense) we know that if the samples are really random, independent and drawn from a population with mean $\mu_0 = 100$ (the skeptic's hypothesis), the distribution of the sample mean is again Normal, with mean equal to $\mu = \mu_0$ and variance σ^2/n. We write:

$$\bar{X} \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{n}\right).$$

Asking how unusual would a sample mean of 105.3 be if the null hypothesis were to be true then amounts to compute an integral, the area to the right of $\bar{x} = 105.3$ under a normal curve whose mean is $\mu_0 = 100$ and variance is $\frac{\sigma^2}{n} = \frac{15^2}{30}$. This area is in fact a probability. It is the probability that $\bar{X} \geq 105.3$ which is given by

$$P(\bar{X} \geq 105.3) = \int_{105.3}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp -\frac{(\bar{x} - \mu_0)^2}{2\sigma^2/n} d\bar{x}.$$

Fortunately, we can ask `R` to compute that integral for us with the following line:

```
> 1-pnorm(q=105.3, mean=100, sd= 15/sqrt(30))
[1] 0.02647758
```

Alternatively, we could go the old ways and standardize our normal distribution of sample means $\bar{X}$ and get the equivalent quantile value of $\bar{x} = 105.3$ in the standard normal distribution Z. Because $\bar{X}$ can be thought of as the following linear transformation of the standard normal distribution

$$\bar{X} = \frac{\sigma}{\sqrt{n}} Z + \mu_0,$$

solving for Z in this equation allows us to find the standardized value of $\bar{x} = 105.3$. That is, since

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

the standardized value of $\bar{x} = 105.3$ is found to be

$$z_{obs} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{105.3 - 100}{15/\sqrt{30}} = 1.935280.$$

Then, knowing that $P(\bar{X} \geq 105.3) = P(Z \geq 1.935280)$ we just have to do a table look up to find out this probability, or, if we don't have our "Z-table" at hand, just ask `R` again:

```
> 1-pnorm(q=1.935280, mean=0, sd= 1)
[1] 0.02647797
```

which happily corresponds to the previous value found before (besides some numerical round-off error). What does the 0.02647797 means? It simply means that if the skeptic's hypothesis was true and the 30 sampled IQ scores came from a distribution with mean 100, then the probability of observing a sample mean *as big or bigger than* 105.3 is slightly less than 0.03. In other words, if the skeptic's hypothesis was true and we were to repeat the experiment of drawing a sample of size $n = 30$ student's IQ scores and each time compute the sample mean, less than 3% of the time we would actually observe sample means as high or higher than 105.3. So our researcher now has computed a value, 0.02647797, that makes his colleague's hypothesis untenable. Given the evidence against his hypothesis, the skeptic concedes and admits to be convinced. How small has the value of $P(\bar{X} \geq \bar{x})$ to be in order to convince a skeptic? Well, in a typical statistical analysis, the threshold to reject the skeptic's null hypothesis is set to be less than 5%, or 0.05. Very serious scientific experiments set the convincing threshold to 0.01. In any case however, that threshold is what is known as α and the probability of observing a test statistic as extreme (extreme in the direction of the research hypothesis) or more than the value actually observed is known as the *p-value*. So this skeptic vs. researcher argument is really where the famous quasi-robotic "**Decision rule:** Reject H_0 if p-value $< \alpha$" comes from. Also, note that whenever a decision is made, two possible errors arise: first, the null hypothesis could be true, but it is rejected. Since we reject the null hypothesis whenever we observe a p-value less than α, given that the null hypothesis is true, that probability is just given by α. Second, it may be possible that we fail to reject H_0 even if it is false. The probability of that happening is denoted by

β. $1-\beta$ is therefore the probability of making the correct choice and it is known is statistics as the *power* of the test. In future lectures we will deal with trying to compute the power for our ANOVAS. Finally, note that I've written "Failing to reject H_0". Why? Why not simply "accepting H_0"? Well, it can be argued that accepting the null hypothesis may suggest that it has been proved simply because it has not been disproved yet. This is a logical fallacy known as "the argument from ignorance".

The duality between Confidence Intervals and Hypothesis Tests:

Suppose we were conducting a hypothesis test of

$$H_0 : \mu = \mu_0 = 500$$
$$H_a : \mu \neq \mu_0.$$

We go out and take a random sample of size $n = 60$ knowing that $\sigma = 100$. Look at the graph below and locate the rejection region and the acceptance region for this example of a two-sided hypothesis test.

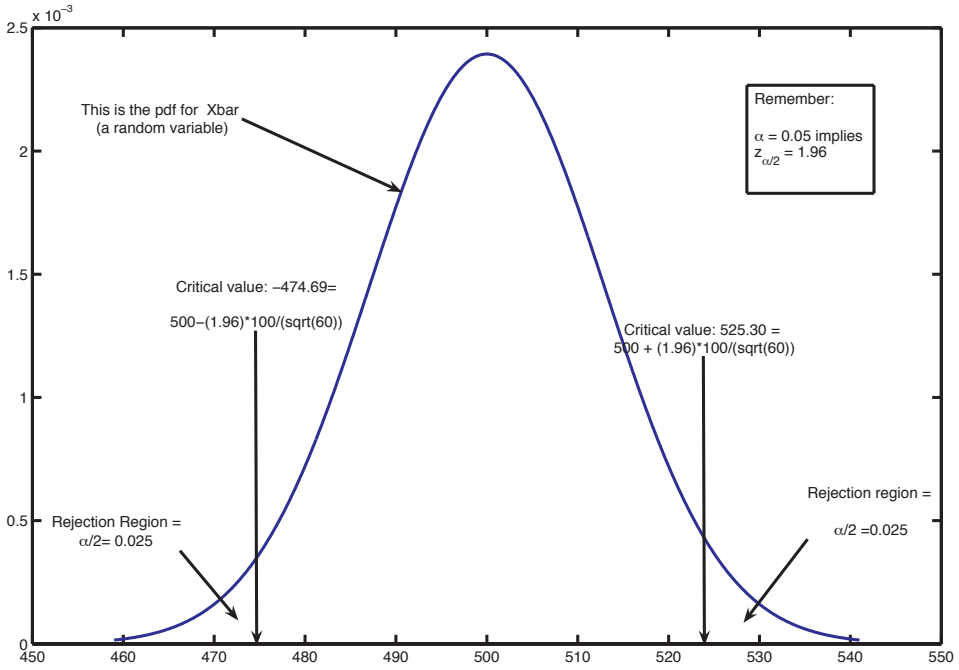

Figure 5: Probability distribution for $\bar{X}$: Depicted are the rejection and the acceptance regions for the two-sided hypothesis test.

What's the size of the acceptance region? That's a probability, it's the area between the two critical quantile values of the distribution of $\bar{X}$. This area is found

to be:

$$\begin{aligned} & P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \\ = \; & P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad \text{(subtracting} \quad \mu \quad \text{everywhere)} \\ = \; & P\left(-\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad \text{(subtracting} \quad \bar{X} \quad \text{everywhere)} \\ = \; & P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \quad (\times -1). \end{aligned} \tag{11}$$

So what this is showing is that, in fact, the confidence interval for μ is just the set of all values of μ_0 for which the null hypothesis $\mu = \mu_0$ would not be rejected in a test against the alternative hypothesis $\mu \neq \mu_0$. Note that because $\bar{x}$ is the *realized* value (that is, one fixed quantity) from the probability distribution of $\bar{X}$, it doesn't make sense to ask

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) =?$$

For a particular random sample, the realized confidence interval

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

either contains or does not contains the true mean μ and we actually do not know which of these two outcomes occurred. If we were to repeat the experiment many many times however, and each time after taking a random sample of size n, we computed the sample mean and its realized confidence interval, then $(1-\alpha)\times 100\%$ of the time the realized confidence interval would contain the true mean μ. For each individual confidence interval, the true mean would either be inside or it would not. Thus, repeating this experiment many many times and computing a confidence interval each time can be though of as a horse shoe game where we have our eyes closed. We shoot the horse shoe many many times (*i.e.* we get the random sample, compute its mean and confidence interval) and each time we either make a stake (*i.e.* the true mean is contained in our realized confidence interval) or we miss the stake (the true mean is not contained in our realized confidence interval), but we do not know for sure what happened (we have our eyes closed!). The only thing we know from the probability calculations above is that $(1-\alpha)\times 100\%$ ($= 95\%$ if $\alpha = 0.05$) of the time the true mean value will be contained in the confidence interval. This is a very hard concept to understand and it is not commonly understood properly.