Tutorial on common statistical traps (2019)

Exercise 1: The trouble with p-values

In the lectures, we encountered p-values, which have a meaningful mathematical definition, but are much better known for being highly confusing and notoriously difficult to handle in hypothesis-testing. Here, we analyze the issue.
Let there be a summary statistic s, which is a function of the data to be analyzed, $s = f(x)$. Depending on the realization x of the data, the numerical value of s will change. The definition of a p-value is then

$$p = \mathcal{P}(s \geq s_c | H_0), \tag{1}$$

which is the frequentist probability $\mathcal{P}$ that the statistic s is as extreme, or more extreme, than some threshold s_c, given that our modelling of noise and all other assumptions in H_0 are correct.

We begin by working out correct implications of the p-value's definition. Let us use the statistical process

$$x \sim \mathcal{G}(0, 1). \tag{2}$$

Part I: p-values and their frequentist sense

1. Derive the p-value $p = \mathcal{P}(x > 3)$ by analytically working out the integral.
2. Calculate the p-value numerically as follows:
   a) draw N Gaussian random samples with unit variance and zero mean
   b) count how many of these have values greater than 3
   c) divide by the total number of samples N to get the fraction F of samples which have $x > 3$
   d) for $N \to \infty$ you should observe $F \to \mathcal{P}(x > 3)$.
3. If you have done everything correctly, you will see that the p-value $\mathcal{P}(x > 3)$ correctly predicts the frequency of samples for which $x > 3$ is true. By design, the null-hypothesis in this exercise was true and read $H_0 : x \sim \mathcal{G}(0, 1)$.

Part II: p-values in hypothesis testing

Until now, the p-value simply predicted how often we will observe an $x > 3$, if we draw Gaussian random samples from $\mathcal{G}(0, 1)$. Numerically, the p-value for this setup is ≈ 0.0013.

Now we go a step further and use the p-value for hypthesis testing. The idea behind p-values in hypothesis testing is:

I have calculated how many samples should fall into the extreme right-wing tail of my Gaussian, such that they have $x > 3$. The associated p-value is $p \approx 0.0013$, i.e. about 1 in a 1000 samples will fall into the right-wing tail beyond $x > 3$. If I therefore only draw one single sample and declare it 'my observation', then it would be extremely unlikely that it falls so far away from the mean. Therefore, if it DOES fall into the region with $x > 3$, I rather think my null-hypothesis is wrong, because why should I be the one unlucky guy out of 1000 people who observes that extreme a realization of my stochastic process?

Here, we therefore agree to reject a null-hypothesis if our observation has a p-value of $p < 0.0013$, which here implies observing an $x > 3$. Therefore:

1. draw a single datum x_i from your distribution
2. if $x_i > 3$, reject the null-hypothesis $H_0 : x \sim \mathcal{G}(0, 1)$
3. repeat n times and count how often you reject the true null-hypothesis.

Part III: p-values and how they accumulate our own mistakes

Now we lie to ourselves; but for planned and pedagogical reasons only. Of course, real data analyses are highly complex, and at many intermediate steps we are forced to work with idealizations, approximations or sometimes even guesstimates based on low S/N. Considered individually, these idealizations are usually unproblematic. However, the p-value clandestinely sweeps up *all* of them: it compresses an entire data analysis into a single scalar number.
We shall now see how devastating this can be for hypothesis testing.

1. In your code, change the variance of the Gaussian distribution that produces your samples from 1 to 2. Convince your brain to forget this 'order-of-magnitude approximation'.
2. Now we wish to test the hypothesis $H_0 : \mu = 0$. To do so,
   a) draw a single datum x_i from your Gaussian distribution
   b) if $x_i > 3$ reject the null-hypothesis
   c) repeat n times and calculate how often you rejected the true null-hypothesis

You will see that you reject the true-null hypothesis too often, because you have made a mistake elsewhere in the analysis.
For the next round of this exercise

1. Reset the variance of your Gaussian distribution to 1.
2. Set the mean to $\mu = -0.5$ and forget about this little blunder.
3. Now draw again one sample x_i after the other, and again reject the null-hypothesis if $x_i > 3$.
4. repeat n times and calculate how often you now accepted the false null-hypothesis $H_0 : \mu = 0$.

You will see that summarizing an entire analysis into a single p-value is such a drastic compression that in the end, you don't know what precisely your p-value measured: Your own mistakes? Or a problem with the null-hypothesis?
In the upcoming lectures we will therefore learn about how to disentangle complex analyses, and troubleshoot/optimize their individual steps.