Answers for: An introduction to statistical testing

Guillaume Filion

February 3, 2013

Exercise 1 (1 pt)

What is the name of the statistic of Student's t test?

Answer of Exercise 1

The effect size. The t statistic is also a valid answer. Strictly speaking, those statistics are not identical, but the differences are negligible within the scope of that course.

Exercise 2 (1 pt)

What is the power of a test?

Answer of Exercise 2

The probability of rejecting the null as a function of the alternative hypothesis. The complement of (one minus) the false negative rate is also a valid answer, provided it is clear that this is a function of the alternative hypothesis (the power is not a number, it is a function or a curve).

Exercise 3 (1 pt)

What is the null hypothesis of the Wilcoxon test?

Answer of Exercise 3

Sampling is independent, and the two samples are taken from the same distribution. This assumption is often referred to as IID (Independent and Identically Distributed).

The Wilcoxon test is often said to be a test for the median. This is because the *alternative* hypothesis is that the distributions differ only in their median *i.e.* that they are shifted relative to each other. Yet, the null hypothesis is not that the distributions have the same median, as this is *not enough* to know the null distribution of the U statistic.

Exercise 4 (2 pts)

What is the distribution of p-values if the null hypothesis is true?

Answer of Exercise 4

Uniform. We have seen that the probability that the p-value is lower than x is x (for $x \in [0, 1]$), which is the definition of the uniform distribution.

1 Statistics win battles

Here is a passage from an article published in The Guardian.

By 1941-42, the allies knew that US and even British tanks had been technically superior to German Panzer tanks in combat, but they were worried about the capabilities of the new marks IV and V. More troubling, they had really very little idea of how many tanks the enemy was capable of producing in a year. Without this information, they were unsure whether any invasion of the continent on the western front could succeed.

One solution was to ask intelligence to guess the number by secretly observing the output of German factories, or by trying to count tanks on the battlefield. Both the British and the Americans tried this, but they found that the estimates returned by intelligence were contradictory and unreliable. Therefore they asked statistical intelligence to see whether the accuracy of the estimates could be improved.

The statisticians had one key piece of information, which was the serial numbers on captured mark V tanks. The statisticians believed that the Germans, being Germans, had logically numbered their tanks in the order in which they were produced. And this deduction turned out to be right. It was enough to enable them to make an estimate of the total number of tanks that had been produced up to any given moment.

source: http://www.guardian.co.uk/world/2006/jul/20/secondworldwar.tvandradio

Suppose you are a statistician working for the Allies. In the first month of the battle, you have captured tanks with serial number $(257, 303, 247, 12, 23)$. The generals know that they can attack if the number of tanks produced per month is less than 500.

Exercise 5 (1 pt)

Should the Allies attack?

Answer of Exercise 5

Note: this question is actually worth 1 point and not 2 as shown in the exam.

Many answers make sense, because it depends on the risk present in every action. The only answer that did not make sense is the religious respect of the 5% cut-off. If your chances are 92% to win a long and painful war, it makes perfect sense to take that chance. Especially if the alternative is to take no action.

However, it can be that 92% is too risky for you. In that case, not attacking is a valid answer, provided you give a valid estimate of your confidence at the next question, which requires a power analysis.

Exercise 6 (4 pts)

With what confidence?

Answer of Exercise 6

This problem is actually simpler than it looks. It can even be solved in closed form *i.e.* without resampling.

If there were exactly 500 tanks, the chances of observing 5 serial numbers below 304 would be $303/500 \times 302/499 \times \ldots \times 299/496 = 0.08066$. And it is clear that if there are more than 500 tanks, this probability will be lower. So the maximum risk you take by deciding that there are less than 500 tanks is 8%. If you have answered "yes" to the previous question, your confidence is 92%. For the record, here is the `R` code to solve this by resampling.

```
> set.seed(123)
> maxima <- NA
> for (i in 1:100000) {
+    maxima[i] <- max(sample(500, size=5))
+ }
> sum(maxima < 304)

[1] 8115
```

This is the short answer, any of these approaches was granted full points for the question. To this, I would like to add a few comments that will extend the content of the class, and explain how I dealt with alternative ansewrs.

1. One-sided tests. We would like to know whether there are 500 tanks or more, which corresponds to the null hypothesis H_0 that there are 500 or more tanks. This is a composite hypothesis, *i.e.* it does not fully determine the distribution of the serial numbers. So instead, we test the hypothesis $H_0(500)$ that there are exactly 500 tanks and reject the null when the value of the statistic indicates that there are less than 500 tanks. This kind of test is called "one-sided", contrary to the "two-sided" tests that we have seen during the class.

$H_0(500)$ fully determines the distribution of the serial numbers, and rejecting it is equivalent to rejecting H_0.

2. Sufficient statistics. The `max` is the best statistic for this problem, it is actually called a "sufficient statistic". This means that it contains all the information about the real number of tanks, unlike the mean or the median. Intuitively, this makes sense because you could draw the serial numbers $(1, 501, 2, 3, 4)$. With a mean equal to 102.2 you might still wonder whether there are more than 500 tanks, but with a `max` equal to 501 you can immediately accept the null hypothesis.

The definion of sufficient statistics involves some integral calculus, so I won't go there, but I will still show that the `max` gives a test with better power than the `mean`.

For the sake of the example, let us assume that there are 400 tanks (so the null hypothesis is false). For clarity I call this hypothesis $H_1(400)$. To compute the power, we need to fix a level (I remind you that the power depends on the level so we simply cannot compute the power without specifying the level). I will fix the level to 0.05.

First we get the threshold for the `max` with the quantile method. I need to get the quantile of order 0.05, *i.e.* the value such that only 5% of the sampled values are lower.

```
> # Find a threshold using the null distribution.
> maxima <- NA
> for (i in 1:100000) {
+    maxima[i] <- max(sample(500, size=5))
```

```
+ }
> thresh_max <- quantile(maxima, 0.05)
```

With that, let us compute the power for $H_1(400)$ at level 0.05.

```
> maxima400 <- NA
> for (i in 1:100000) {
+    maxima400[i] <- max(sample(400, size=5))
+ }
> sum(maxima400 < thresh_max)

[1] 15335
```

Now let us compute the power when using the `mean`. We follow the same logic: first we find a threshold using the null distribution, and see how often we are below the threshold when $H_1(400)$ is true.

```
> means <- NA
> for (i in 1:100000) {
+    means[i] <- mean(sample(500, size=5))
+ }
> thresh_mean <- quantile(means, 0.05)
> means400 <- NA
> for (i in 1:100000) {
+    means400[i] <- mean(sample(400, size=5))
+ }
> sum(means400 < thresh_mean)

[1] 14190
```

Admittedly, the difference between using the `max` and the `mean` is small.

I granted 1 point for choosing the `max` as a statistic, 2 points for the estimation of the null distribution (whatever statistic was chosen), 1 point for the decision region. Finally, I granted 1 bonus point for providing the random seed so that I could reproduce your sampling.