

PRINCIPLES OF STATISTICAL INFERENCE (PSI)

The required exercise for submission is **Exercise 5**.

Module 3 Practical Exercises: Reduction in HIV viral load

These practical exercises illustrate, primarily via simulation of repeated studies, some of the basic hypothesis testing concepts discussed throughout the module content. Statistical software will be used for simulation to investigate the performance of statistical tests in repeated studies.

Stata and R code for these exercises are provided on Canvas.

Reduction of HIV Viral Load

In the extended example we discussed a setting in which antiretroviral drugs were used to reduce HIV viral load, with the binary outcome being whether or not patients achieved adequate suppression of the virus. In these exercises we look at the *magnitude of reduction* in viral load as a continuous outcome, that is, the change in viral load from just prior to therapy to that achieved after six months of therapy. For simplicity, we will just examine the reduction in one of the treatment groups, the combination therapy group, and consider hypothesis testing of whether there is statistical evidence of a reduction on this therapy (at a specified α level). In practice, it would be important to compare this reduction with that of a control group in a randomized trial. Testing of this type, which is analogous to the two-sample comparisons of proportions in the extended example, will be considered as one of the illustrations in subsequent modules.

Our sample $y = (y_1, \dots, y_n)$ corresponds to observations of the change in viral load over a six-month period for n individuals on combination therapy. We will begin by assuming that this change has a $N(\mu, \sigma^2)$ distribution, although in later exercises we will sample from a non-normal distribution by incorporating outliers. We will study two statistical tests. The first is the **one-sample t-test**, which is based on the rule:

Reject H_0 if $|t| \geq c$, where the test statistic:

$$t = \frac{\bar{y}}{s_{n-1}/\sqrt{n}}$$

is an observation from a t_{n-1} distribution when H_0 is **true**.

The second test is the **signed rank test**, which is a non-parametric test based on taking the ranks of all the observations after ignoring their sign, and then comparing the ranks of negative observations with the ranks of positive observations. We will not define this test in detail here, but will use software to carry out the hypothesis tests. Rank-based tests such as this will be discussed further in Module 6.

Exercise 1: We begin by considering only the **one-sample t-test**.

- Write down the null-hypothesis and two-sided alternative hypothesis that would be of interest.
- Describe how to determine the critical value c , based on a significance level α , and thus calculate the critical value c for $n = 50$ and $\alpha = 0.05$.
- Describe how to calculate the p-value for the sample, and then calculate it supposing that the mean reduction in viral load was $\bar{y} = 0.2$, $n = 50$ and the observed variance was $s_{n-1}^2 = 1.04$.
- What would you conclude regarding the hypotheses from part (a)?
- Using the confidence interval based on the t-distribution described by

$$\bar{y} \pm t_{n-1,0.05} \frac{s_{n-1}}{\sqrt{n}}$$

- (where $t_{n-1,0.05}$ is the point on the t-distribution that excludes 0.025 in each tail), calculate a confidence interval for the mean reduction in viral load. Explain how it is consistent with the result in part (d).
- Suppose that the reduction in viral load is only of practical benefit if it exceeds 0.3. Interpret your answers to parts (d) and (e) in the context of this additional information.

Exercise 2: This question concerns the *significance level* (or type I error) of the two tests (i.e., one-sample t-test and the signed rank test). Consider 1000 studies repeated on the same population in which there is no average reduction in viral load from the treatment (i.e., $\bar{y} = 0$). Suppose that the variance of the reduction between individuals is 1.0 (on the scale used to measure viral load). This question involves simulating the results from these studies based on normally distributed reductions, using a sample size of 50.

- Write down the interpretation of the significance level of a hypothesis test, in the context of conducting repeated studies.
- Using statistical software, simulate 1000 repeated studies based on the assumptions above (Stata and R code is available on Canvas). For each study, decide whether the null hypothesis would be rejected or not at the 5% significance level based on the results of each of the two types of hypothesis tests. Hence, obtain a simulation-based estimate of the significance level for each of the two tests.
- Based on the results in part (b), do the two tests seem to have the significance level that you would expect?

Exercise 3: This question concerns the *power* of the two tests (i.e., one-sample t-test and the signed rank test). Use the same assumptions as in Exercise 2, except we will now assume that the population does indeed have a reduction in viral load when treated with combination therapy.

- Write down the interpretation of the power of a hypothesis test, in the context of conducting repeated studies.
- Repeat the simulations carried out in Exercise 2, except now assume that the population mean reduction is 0.2, 0.4, 0.6, 0.8 and 1.0 (i.e. carry out 5 additional sets of simulations for each of these possible mean reductions by appropriately modifying the simulation parameters). For each study decide whether the null hypothesis would be rejected or not at the 5% significance level, based on the results of each of the two types of hypothesis tests. For each value of the mean reduction, obtain a simulation-based estimate of the power for each test.
- For the mean reduction ranging from 0 to 1, plot a simulation-based estimate of the power function for each test (i.e., plot both functions on the same graph).
- What advantage does the t-test have? Based on your simulation results, is it an important advantage and why do you think it arises?

Exercise 4: Suppose we have a population in which there is no reduction in viral load on average (as in Exercise 2). Suppose we conduct k studies on this population and test whether there is a statistically significant reduction in each case. The test statistic that we use is unimportant for this exercise.

- Write down an expression for the probability that *at least one* of these studies yields a result that is statistically significant, using a significance level of α for each test? (Hint: the probability that at least one is significant is 1 minus the probability that none are significant).
- Calculate this for $\alpha = 0.05$ and $k = 1, 2, 3, 5, 10, 15, 20$.
- By referring to your results in part (b), summarise the problem with conducting multiple tests and concluding statistical significance if one of them rejects the null hypothesis.
- Supposes we let our significance level depend on the number of tests we are conducting. In particular, if we are conducting k tests, suppose we use a significance level of α/k for each test, where $\alpha = 0.05$. Based on this, repeat the calculation in part (b), to determine the probability that at least one of the studies will yield a statistically significant result.
- Using the results of part (d), explain how this correction to the significance level rectifies the problem identified in part (c).
(This is called *Bonferroni correction for multiple comparisons*, which is one of several different approaches to deal with the problem.)

The assessable exercise changes the focus from viral load to the effect of medication on blood pressure.

Exercise 5:

Aim: We want to compare the performance of the two independent sample t-test and the Wilcoxon rank-sum test in this exercise. We fix the (nominal) level of these tests at $\alpha = 5\%$. This will again be achieved through simulations using the following scenario:

We want to assess the effect of a blood-pressure lowering medication in a specific population. Imagine a study where a total of 200 patients are assigned to either a placebo (control) or the test drug (treated). Each arm of the study has 100 patients. We focus on the diastolic blood pressure and assume that its distribution is normal with mean $\mu_0 = 80$ and a standard deviation of $\sigma = 10$ in the placebo group. We will also assume that the distribution of diastolic blood pressure in the treated group is normal with mean $\mu = 80 - \delta$ where δ represents the reduction in diastolic blood pressure, or the treatment effect. The same standard deviation as in the placebo group, $\sigma = 10$, can be used.

- a. Carry out simulations similar to those conducted in Exercise 3b, with $\delta = 0, 2, 4, 6, 8$, and 10 . The main difference is that we now have tests based on two different (independent) samples. Stata and R code for these simulations are available on Canvas. Tabulate and plot the same simulation-based estimates of the power functions (as functions of δ) for each test, as was carried out in Exercise 3c.
- b. Which test seems to have an advantage when the distribution is indeed Normal in both groups?

We now consider that the distribution of diastolic blood pressure in the treated group is a “contaminated” normal. Approximately 97% of observations come from the same normal distribution as before and the remainder are sampled from a distribution with a larger mean and variance, $N(130, \gamma^2)$ where $\gamma = 2\sigma = 20$. Outliers are likely to arise in the treated group since about 3% of the observations are sampled from a different population.

- c. Repeat part (a) using the data described above (i.e., where 3% of the observations in the treated group are “contaminated”). Stata and R code to perform these simulations can be found on Canvas.
- d. Does the t-test maintain its nominal significance level of 5%? Which test seems to have an advantage now?

A note on the Wilcoxon two sample (rank sum) test: This test is commonly used as an alternative to the two (independent) sample t-test, when it is thought that the assumption of normality does not hold. It is based on the ranks of the observations.

Exercise 6: Suppose you can approximate the t-distribution by the normal distribution. Analogous to the expression given in the extended example, the expression for the sample size required for a **one-sided** hypothesis test is:

$$n = \frac{\sigma^2}{d^2} (z_{1-\alpha} + z_{1-\beta})^2$$

where d is the difference in the mean viral load reduction to be detected and σ^2 is the variance of the reduction. How does the required sample size change as the following quantities increase (provide brief reasons):

- a. The variance of the reduction in viral load?
- b. The detectable difference d ?
- c. The power $1-\beta$?
- d. (*optional*) Justify the above formula.