Lab 5 - Hypothesis testing and linear regression.

Due: Wednesday May 3rd 11.59PM in PDF form via Websubmit.

Note that the whole coding part of this lab must be done in R.

1 What to submit (MANDATORY)

1. PDF file ((lab5_fisrtname_lastname.pdf))including plots and a snapshot of the code used to answer the questions.
   - Names of the collaborators.
   - Number of late days for this assignment.
   - Number of late days so far.
   - References used
2. R script (lab5_firstname_lastname.R) with the code used.

Failing to meet any of the above requirements will cause a decrease of your grade.

2 Background

2.1 Confidence interval for the population mean

In class we learned that according to the central limit theorem, the distribution of the sample mean $\overline{X}_n$ is approximately a normal distribution with a mean of μ (the population mean) and standard deviation of $\frac{\sigma}{\sqrt{n}}$ (where σ is the population standard deviation). For a random variable with a normal distribution, the probability that its value is within 2 standard deviations of its mean is about 0.95. Obviously, if there is a certain distance between the sample mean (recall that the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$) and the population mean, we can describe that distance by starting at either value. So, if the sample mean $\overline{X}_n$ falls within a certain distance of the population mean μ, then the population mean μ falls within the same distance of the sample mean. Therefore, the statement, "There is a 95% chance that the sample mean $\overline{X}_n$ falls within 2 standard deviations of μ" can be rephrased as: "We are 95% confident that the population mean μ falls within 2 standard deviations units of $\overline{X}_n$". This second statement is exactly the interpretation of the confidence interval. Similarly, if our hypothesis is that the population mean is equal to μ_0, and μ_0 is within 2 standard deviations units of $\overline{X}_n$, we say that the hypothesis is not rejected at a significance level of $\alpha = 0.05$.

Definition:
Given a sample of size n, under the assumption that we know the population standard deviation σ, the **two sided confidence interval** of our sample is computed as follows:

$$\overline{X}_n \pm z \times \frac{\sigma}{\sqrt{n}} \tag{1}$$

where $\overline{X}_n$ is the sample mean and z is a multiplier that depends on the level of significance α.

Some important values of z are :

- For $\alpha = 0.1$ (90% confidence interval), $z_{\alpha/2} = 1.645$
- For $\alpha = 0.05$ (95% confidence interval), $z_{\alpha/2} = 1.96$
- For $\alpha = 0.01$ (99% confidence interval), $z_{\alpha/2} = 2.576$

Note that if we want to compute **one sided confidence interval** then we have to use z_α. This is because in the case of one sided intervals we are interested only in the lower value (when the alternative hypothesis is "greater than") or the upper value (when the alternative hypothesis is "less than").

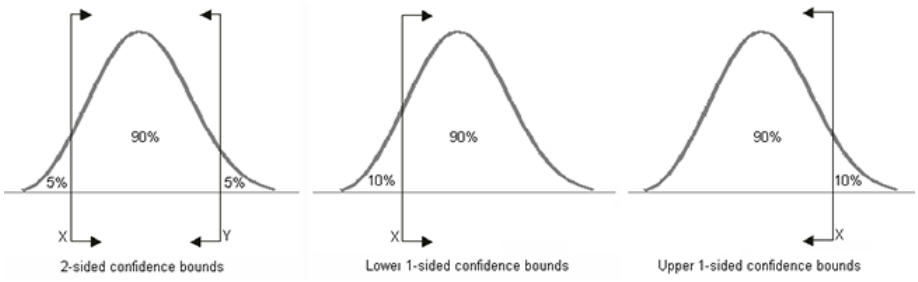

Figure 1: Different types of confidence intervals, for all three figures $\alpha = 0.10$.

2.2 Hypothesis testing for population mean

Recall that there are basically 4 steps in the process of hypothesis testing:

1. State the null (H_0) and alternative hypotheses (H_1).
2. Collect relevant data from a random sample and summarize them (using a test statistic).
3. Find the p-value, the probability of observing data like those observed assuming that H_0 is true.
4. Based on the p-value, decide whether we have enough evidence to reject H_0 (and accept H_1) , and draw our conclusions in context. To make a decision we have to chose a significance level. In this lab, unless explicitly stated, we will use 0.05 significance level.

Assume that μ is our population mean. Note that the null hypothesis always takes the form: $H_0 : \mu = \mu_0$ (where μ_0 is some value). The test statistic can take one of the following three forms, depending on what is our alternative hypothesis:

1. $H_1 : \mu > \mu_0$ (right-tailed test)
2. $H_1 : \mu < \mu_0$ (left-tailed test)
3. $H_1 : \mu \neq \mu_0$ (double-tailed test)

In hypothesis testing we have to distinguish between two cases: 1) the case where the population standard deviation (σ) is known, and 2) the case where σ is unknown. In the first case the test we will use is called the **z-test for the population mean** μ. In the second case, the test is called the **t-test for the population mean** μ.

In the first case, the test statistic will have a standard normal (z) distribution (when H_0 is true), and in the second case, the test statistic will have a t-distribution (when H_0 is true).

3 z-test for the population mean (σ is known)

3.1 Learning example

The SAT is constructed so that scores in each portion have a national average of 500 and standard deviation of 100. The distribution is close to normal. The dean of students of Ross College suspects that in recent years the college attracts students who are more quantitatively inclined. A random sample of 4 students from a recent entering class at Ross College had an average math SAT (SAT-M) score of 550. Does this provide enough evidence for the dean to conclude that the mean SAT-M of all Ross college students is higher than the national mean of 500? Assume that the scores of all Ross College students are also normally distributed with a standard deviation of 100.

1. State null and alternative hypothesis.

When we discussed probability models based on sampling distributions, we concluded that sample mean, $\overline{X}_n$, is a random variable with the following properties:

- The mean is the same as the population mean, μ.
- The standard deviation is $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the population.
- The sample means are normally distributed if the underlying variable being sampled is normally distributed in the population or the sample size is large enough to guarantee approximate normality. Recall that this last statement is the Central Limit Theorem. As a general guideline, if $n > 30$, the Central Limit Theorem applies and we can use the normal distribution to model the distribution of $\overline{X}_n$

Based on this description of the sampling distribution of the sample mean $\overline{X}_n$, we can define a test statistic that measures the distance between the hypothesized value of μ (denoted μ_0) and the sample mean (determined by the data) in standard deviation units. The test statistic is:

$$Z_n = \frac{\overline{X}_n - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{2}$$

Comments

- Note that our test statistic (because it is a z-score), tells us how far $\overline{X}_n$ is from the null value μ_0 measured in standard deviations. Since $\overline{X}_n$ represents the data and μ_0 represents the null hypothesis, the test statistic is a measure of how different our data are from what is claimed in the null hypothesis. The larger the test statistic, the more evidence we have against H_0, since what we saw in our data is very different from what H_0 claims.
- All inference procedures are based on probability. We are trying to determine if our sample results are likely or unlikely based on our assumptions about the population. This requires that we have a probability model that describes the long-term behavior of sample results that are randomly collected from a population that fits our hypothesis. For this reason, the Central Limit Theorem gives us criteria for deciding if the z-test for the population mean can be used. We need to verify:
  1. The sample is random (or at least can be considered as random in context).
  2. We are in one of the three situations marked with yes in the following table:

Conditions: z-test for a population mean	Small sample size ($n \leq 30$)	Large sample size ($n > 30$)
Variable x_i in the population from normal distribution	YES	YES
Variable x_i in the population not from normal distribution	NO	YES

- 3. If the conditions are met, then values of $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ vary normally, or at least close enough to normally to use a normal model to calculate probabilities. When $\overline{X}_n$ values are normal, then the z-scores will be normally distributed with a mean of 0 and a standard deviation of 1.

Now let's get back to our SAT example.

2. Can we use the z-test to do our analysis? Hint: recall the condition we have to check
3. What is the value of the sample mean $\overline{X}_n$?
4. What is the value of population standard deviation σ?

5. What is the value of sample size ?
6. Compute the z-statistics and explain how one should interpret the result.
7. Find the p-value of the test using the normal table (`http://www.normaltable.com/`). Hint: Recall that the p-value when H_1 is ''greater than'' (right tailed z-test) is $Pr(Z \geq z)$. The normal table shows $Pr(Z < z)$
8. Suppose we reject the null hypothesis if our results are significant at 5% level. Can we reject the null hypothesis given the p-value we obtained?
9. What would be the minimum sample size we need to reject the null hypothesis with a significance level of 95%? Hint: first you have to find the z value for which $p - value \leq 0.05$. Then you can compute the n needed.
10. Now let's verify our results with code. For this we are going to use R. Create a function called significance that on input: the sample size n, the population standard deviation σ, the population mean μ and the sample mean $\overline{X}_n$, computes z and the p-value. Hint: in R the function pnorm computes $Pr(Z < z)$. To create a function in R you do the following function_name = function(parameters)**Submit your code.**
11. Execute the significance function for increasing values of the sample size (starting with $n = 4$ increment every time by 1) until the results are statistically significant, i.e, p-value≤ 0.05. Provide a results table with the following 4 columns: n, z (test statistic), p-value and significant (yes/no). Which is the minimum sample size for which we can reject the null hypothesis? Using R you can test all the values of n from 4 to 14 by entering significance(5:14). **Submit your code and table.**

3.2 Problem

Every year, the Environmental Protection Agency (EPA) collects data on fuel economy (randomly sampling from the entire population). With rising gasoline prices, consumers are using these figures as they decide which automobile to purchase. We will look at two-seater automobiles, many of which are sporty vehicles. Based upon the latest 2017 EPA sample, we wish to test the hypothesis that the combined city and highway miles per gallon (mpg) of two-seater automobiles is greater than 20. The standard deviation for all vehicles is 4.7 mpg. The dataset containing the data is *epa.csv* and the column you are interested in is COMB.MPG.

12. State the null and alternative hypothesis
13. Have the conditions that allow us to safely use the z-test been met?
14. Compute the test statistics and the p-value using the normal table (`http://www.normaltable.com/`).
15. the extend the function you wrote for question 10 such that on input
    - the sample size n
    - the population standard deviation σ
    - the population mean μ
    - the sample mean $\overline{X}_n$
    - alternative: either "less","greater" or "two.sided" indicating the form of the alternative hypothesis.

    computes and outputs sample mean, sample size, z and the p-value. Hint: recall that for the two sided test the $p - value = 2 \times Pr(Z \geq |z|)$**Submit your code.**
16. Provide the output of the function and verify that it matches the theoretical values computed above.
17. Draw conclusions based on the context of the problem.

18. Compute the one sided confidence interval for $\alpha = 0.05$ (95% confidence interval), Provide both upper and lower values.

19. Use R to plot the confidence interval computed above. Recall that we assume that the sample mean $\overline{X}_n$ is normally distributed. Therefore you need to create a normal variable with mean equal to the sample mean and standard deviation equal to the population standard deviation and plot its probability density function. Then to the same figure add two vertical lines corresponding to the lower and upper confidence interval computed above. Hint: the function abline it is used to add vertical or horizontal reference lines to a plot in R. **Submit your code and plot.**

20. What would be the minimum sample size we need to reject the null hypothesis with a significance level of 95%?

3.3 Relating Hypothesis Tests and Confidence Intervals

Suppose we want to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ using a significance level of $\alpha = 0.05$. An alternative way to perform this test is to find a 95% confidence interval for μ and make the following conclusions:

- If μ_0 falls outside the confidence interval, reject H_0.
- If μ_0 falls inside the confidence interval, do not reject H_0.

21. Compute the one sided confidence interval for the SAT problem for $\alpha = 0.05$. Provide both upper and lower value of the confidence interval.

22. Does μ_0 (the population mean) fall outside or inside the confidence interval?

23. Now compute the confidence interval assuming $n = 11$.

24. Does μ_0 (the population mean) fall outside or inside the confidence interval?

4 t-test for the population mean (σ is unknown)

Unfortunately, only in few cases it is reasonable to assume that the population standard deviation (σ) is known. What can we use to replace σ? If you don't know the population standard deviation, the best you can do is find the sample standard deviation, S (which formula is $\sqrt{(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X}_n)^2)}$), and use it instead of σ. In doing so we also have to change the test we use in the hypothesis testing which is now the t-test. The condition under which we can apply the t-test are the same expressed for the z-test (see Table 2). The test statistic for the t-test is defined as:

$$t = \frac{\overline{X}_n - \mu_0}{\frac{S}{\sqrt{n}}} \tag{3}$$

In the denominator we are using S instead of σ. This change has an effect on the distribution of the t-test statistic, which now does not follow a normal distribution. Instead it follows a distribution called t distribution or student distribution.

The t distribution has slightly less area near the expected central value than the normal distribution does, and that the t distribution has correspondingly more area in the "tails" than the normal distribution does. Therefore, the t distribution ends up being the appropriate model in certain cases where there is more variability than would be predicted by the normal distribution.

There are actually many different t distributions. The particular form of the t distribution is determined by its degrees of freedom. The **degrees of freedom** refers to the number of independent observations in a set of data. When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one. This is important when we want to compute the p-values for our hypothesis testing exercise: if the sample size is $n = 10$, we will compute the p-value of $t(n-1) = t(9)$.

25. In order to compare the normal distribution with the t-distribution, plot the density of a normal distribution for 100 values in the range $[-4, 4]$. Then, to the same plot, add the density function of a t-distribution for the following values of degree of freedom: $df = \{1, 3, 8, 30\}$ **Submit your code and plot.**

26. What happens to the t-distribution when we increase the degrees of freedom?

4.1 Problem

We are going to use the SAT problem we analyzed in the previous section but with a little modification. Now we don't know σ. Instead we will use the sample standard deviation S (which we can compute from the sample) as an approximation for σ. This change implies that the z-test is not longer appropriate and we need to use the t-test.

27. Can we use the t-test to do our analysis? Hint: recall the condition we have to check (see Table 2)

28. How many degrees of freedom we have?

29. Given $S = 100$ compute the t-statistic and explain how one should interpret the result.

30. Find the p-value of the test using R. Hint: the function is called pt. Recall that the p-value when H_1 is "greater than" (right tailed z-test), pt by default computes $Pr(T < t)$.

31. Is the p-value for the t-test larger or smaller than the p-value we computed with the z-test? Is it surprising?

32. Suppose we reject the null hypothesis if our results are significant at 5% level. Can we reject the null hypothesis given the p-value we obtained?

33. Compute the 95% one sided confidence interval ($\alpha = 0.05$). In order to compute it you need to find the t statistic value t_α. Provide both lower and upper values of the interval. Is the confidence interval wider that the one computed using the population standard deviation in part 21? Why?Hint: the R function is qt and it computes the t value for a one sided t test. You need to use $S = 100$ in the confidence interval formula since we do not have σ