

Statistical Thinking for Forensic Practitioners

Excel Lab on Part 6: Inference

Fall 2020

We will practice applying the inferential procedures discussed in the Part 6 lecture material to a data set. Open the `framingham_rjc_2003_part5_lab.xlsx` file found on the course website. This file contains a sample of patient data from a hospital in Framingham, MA. The variables collected for each patient include:

- AGE: age in years
- SMOKE: smoking status (0 for non-smoker, 1 for smoker)
- CHD: coronary heart disease status (0 for not diseased, 1 for diseased)
- CHOL: cholesterol level (milligram/deciliter of blood)
- HIGH_CHOL: high cholesterol classification (0 for not high, 1 for high)
- SBP: systolic blood pressure (millimeters of mercury)

Assume that this dataset is representative of all patients in the hospital.

1 Comparison of two means

We are interested in determining whether the systolic blood pressure of smoking patients is *different from* the systolic blood pressure of non-smoking patients in the hospital. We can answer this using the inferential procedures discussed in lecture. Namely, hypothesis tests and confidence intervals. Slides 76-101 will be useful for these exercises.

Recall that the goal of a hypothesis test is to build evidence against the null hypothesis H_0 in favor of an alternative H_a . Let μ_0 and μ_1 be the average systolic blood pressure of non-smoking and smoking patients in the hospital, respectively. Our null hypothesis will be $H_0 : \mu_0 = \mu_1$ or, equivalently, $H_0 : \mu_1 - \mu_0 = 0$ (we will use the latter of these two expressions). We will use a significance level of $\alpha = .01$ for this example.

1. Is this a one-sided or two-sided hypothesis test? (Hint: pay close attention to what we are interested in determining)

2. Express the alternative hypothesis H_a using similar notation as H_0 .

3. We do not know the true values of μ_0 and μ_1 , but can estimate them using the appropriate sample means from our data, \bar{y}_0 and \bar{y}_1 say. Compute the average systolic blood pressure of the two groups in our sample (the `AVERAGEIF` function may prove useful).

4. Use the formula at the bottom of slide 80 and the `VAR.S` function to calculate the standard error SE_{diff} for the difference between the two sample means. Note that the $1/2$ exponent shown on slide 80 can be calculated using the `SQRT` function (the `ROWS` function may be useful for counting the sample sizes of each group).

5. Calculate the t statistic we can use for this hypothesis test.

6. What are the degrees of freedom associated with this t test?

7. Use the `T.INV` function to calculate the critical value associated $\alpha = .01$ significance level and the degrees of freedom you calculated above. (Note: the type of test you identified in question 1 requires that, for an α significance level of .01, we calculate the critical value using `probability` set to $\frac{\alpha}{2} = .005$ in the `T.INV` function as the Type I probability needs to be equally split between the two tails.)

8. Do you reject or fail to reject H_0 ? Explain what your decision means in the context of the problem.

9. Why might we use a p -value instead of simply comparing our t -statistic to a critical t^* value to reach a conclusion in a hypothesis test? (Hint: see slide 71)

10. Use the `T.DIST.2T` function to calculate the p -value associated with this hypothesis test (you may need to use the `ABS` function as the `T.DIST.2T` function does not accept negative numbers, depending on how you constructed your t statistic).

11. Calculate a 99% confidence interval for the difference in mean systolic blood pressure between the smoking and non-smoking patients in the hospital. Why is the conclusion based on this confidence interval the same as the conclusion based on the hypothesis test performed above? (Hint: see slide 85, but replace $2SE_{diff}$ with t^*SE_{diff} where t^* is the critical value determined in question 7).

12. Instead of a 99% confidence interval, suppose we desired a 90% confidence interval for the difference in mean systolic blood pressure between the smoking and non-smoking patients in the hospital. Would this interval be wider or narrower than the interval calculated above? Explain. (Hint: only one quantity changes in the calculation of the 90% interval compared to the 99% interval in this example. Consider how this quantity affects the width of the interval.)

13. Describe what a Type I error would mean in the context of the problem.

