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 presser 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)

   Two-sided. We are interested in determining whether the mean systolic blood pressures is *different* between the two groups, not that one is greater than the other (which would be a one-sided test).

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

   $H_a : \mu_1 - \mu_0 \neq 0$ **(means that the average SBP for the smoking groups is greater than that of the non-smoking group.)**

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).

   $\bar{y}_0 = 130.004$ **and** $\bar{y}_1 = 133.320$ **rounding to 3 decimal places.**

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).

   $SE_{diff} = 1.171$ **rounding to 3 decimal places.**

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

   $t = -2.832$ **rounding to 3 decimal places.**

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

   For a hypothesis test for the difference of two means, $d.f. = n_1 + n_2 - 2 = 1248 + 367 - 2 = 1613$

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.)

   $t^*_{1613} = -2.579$ **(or 2.579 by symmetry of the** t **distribution) rounding to 3 decimal places.**

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

   Reject since the observed t **statistic is less than the critical** t**-value. That is,** $t = -2.832$ **is outside of the "fail-to-reject" range** $[-2.329, 2.329]$**. This means that we reject the assertion that the mean systolic blood pressures are equal between the smoking and non-smoking patients in the hospital (equivalently, difference between the averages is not 0.**

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)

   Using only a critical t^* **value only indicates that our** t**-statistic is within or without the "rejection" region, so we only get a binary classification of "typical" or "atypical." A** p**-value quantifies *how* atypical a** t **statistic is.**

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).

    $p = .005$ **rounding to 3 decimal places.**

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).

    The 99% confidence interval is $[-6.335, -.297]$ **rounding to 3 decimal places. Because 0 is not included in the interval, we conclude that there is a significant difference in average systolic blood pressures between the smoking and non-smoking patients of the hospital. There is *always* a duality between a two-sided test with significance level** α **and a** $(1-\alpha)\,\%$ **confidence interval – if the test rejects** H_0**, then the confidence interval does not include the null hypothesized value and vice versa. Note that this duality does not hold for one-sided tests.**

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.)

    In calculating a 90% confidence interval, we would use a different t^* **critical value, specifically** $t^* = -1.646$ **(or 1.646 by symmetry of the** t **distribution) rounding to 3 decimal places. This is clearly smaller in magnitude from 2.832 meaning the confidence interval would be narrower.**

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

    A Type I error would mean *rejecting* the assertion that the average systolic blood pressure for the smoking and non-smoking groups are equal when, in fact, they *are* equal.

14. High blood pressure is known to be associated with health problems. Before performing the hypothesis test, suppose you were told that a health insurance company would use the outcome of the test to decide whether they would triple the premiums for their smoking customers (i.e., if the null is rejected in favor of the alternative, then they would triple the premiums). Comment on why you might choose a relatively small α significance level (.005, .01, etc.) over a larger significance level (.05, .1, etc.).

    Tripling premiums would likely put a monetary strain on many customers. As such, we might want to make sure to reduce the probability of making a Type I error (which is controlled by the α level).

15. Calculate Cohen's d statistic for the difference between the mean systolic blood pressure (see slide 92).

    $d = -.178$ **round to 3 decimal places.**

16. Using the cutoff values given on slide 92 for low, medium, and large d values, interpret the effect size (as measured by d) in the context of the problem. How does this compare to the results of the hypothesis test (think in terms of practical vs. statistical significance)?

    Since $d = -.178$ is less than .2 in magnitude, it would be classified as "low." This indicates that the effect size of the average systolic blood pressure between the smokers and non-smokers in the hospital is not as "significant" as the hypothesis test indicates. Perhaps the difference is not practically significant (although this would probably best be left up to medical professional to determine in this example).