Hypothesis Tests 2 – Solutions
COR1-GB.1305 – Statistics and Data Analysis

Complete Examples

1. Is the crowd wise? National Public Radio's *Planet Money* podcast performed an experiment to measure the "wisdom of the crowd" with regard to estimating the weight of a cow[1]. After being shown a picture of a cow, respondents were asked to guess its weight, in pounds. The mean of the 17,109 guesses was 1282 pounds, and the standard deviation was 534. The true weight of the cow was 1355 pounds. So, the crowd of 17,109 respondents under-estimated the weight of the cow by 73 pounds. It's possible that a larger crowd could do better. Given the data available, is this plausible? That is, is it plausible that with a large enough crowd, the estimation error could be made arbitrarily small? We will answer this by performing a hypothesis test.

   (a) What is the sample? What is the population? What is the interpretation of the population mean, μ?

   Solution:
   The sample is the reported $n = 17109$ guesses. One possible population is the (hypothetical) guesses of all respondents, if the experiment were allowed to run forever, accumulating more and more respondents. That is, the population is a hypothetical "infinite crowd". The population mean is the average guess of all respondents in the population.

   (b) What are the null and alternative hypotheses, in terms of μ?

   Solution:

   $$H_0 : \mu = 1355 \quad \text{(the ``infinite crowd'' would get the weight exactly right)}$$
   $$H_a : \mu \neq 1355 \quad \text{(the ``infinite crowd'' would err in estimating the weight)}$$

   (c) Compute the test statistic.

   Solution:

   $$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{(1282) - (1355)}{(534)/\sqrt{(17109)}} \\ &= -17.88 \end{aligned}$$

[1] http://www.npr.org/sections/money/2015/08/07/430372183/episode-644-how-much-does-this-cow-weigh

(d) How strong is the evidence against the null hypothesis?

Solution:
An approximate p-value is

$$\begin{aligned} p &\approx P(|Z| > 17.88) \\ &< .0001 \end{aligned}$$

If the "infinite crowd" were completely accurate, then the chance of seeing data like observed would be less than .0001 (less than 0.01%). This is very compelling evidence that the infinite crowd is not completely accurate.

(e) Is it plausible that with a large enough crowd, the estimation error could be made arbitrarily small?

Solution: As part (d) notes, this is very implausible.

Note: it is impossible to assign a probability to the statement. In particular, the p-value does *not* give the probability that the statement (the null hypothesis) is true.

2. Before Facebook's recent redesign, the mean number of ad clicks per day was 100K. In the 49 days after the redesign, the mean number of ad clicks per day was 105K and the standard deviation was 35K. Is there significant evidence that the redesign affected the expected number of ad clicks? Perform a test at the 5% level.

   (a) What is the sample? What is the population?

   Solution: The sample is the number of ad clicks on the measured $n = 49$ days.

   The population is the number of ad clicks on all days after the redesign.

   (b) What are the null and alternative hypotheses?

   Solution: Let μ be the expected clicks per day after the redesign We use "thousands of clicks" as the units for all relevant quantities.

   The null hypothesis is that the redesign had no effect on expected ad clicks. The alternative hypothesis is that μ changed after the redesign:

$$H_0 : \mu = 100$$
$$H_a : \mu \neq 100$$

   (c) What is the test statistic?

   Solution: The test statistic is based on the sample, the observed clicks in the $n = 49$ days after the redesign. Let $\bar{x}$ denote the mean clicks per day in the sample; let s denote the sample standard deviation. Our test statistic is

$$\begin{aligned} t &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{105 - 100}{35/\sqrt{49}} \\ &= 1. \end{aligned}$$

   (d) Approximately what is the p-value?

   Solution:

$$\begin{aligned} p &\approx P(|Z| > 1) \\ &= 0.3137. \end{aligned}$$

   (e) What assumptions are you making?

   Solution: We need to assume that the observed sample is a simple random sample from the population.

Admittedly, the assumption does *not* hold, since there is strong selection bias in the sample: we are sampling days right after the website redesign with higher probability than days far into the future. For example, we have no chance of sampling a day three years into the future.

Since the assumption does not hold, we are in a bit of an awkward position. The hypothesis test may not be valid. One way in which it may not be valid is the following: it usually takes people a few weeks to adjust to a website redesign, so the ad click behavior in our sample of 49 days may not be representative of all future ad click behavior.

(f) What is α? What is the result of the test?

Solution: Despite the caveats mentioned in part (d), we will proceed with the test. For a level-5% test, $\alpha = 0.05$. Since $p \geq 0.05$, we do not reject H_0. If there were no difference in expected ad clicks before and after the redesign, there would be a 31.37% chance of seeing data like we observed. There is no evidence of a difference.

Types of Errors

3. In a hypothesis test, our decision will either be "reject H_0" or "do not reject H_0". Under what situations will each of these decisions be in error?

Solution: Type I error: H_0 is true, but we reject it.
Type II error: H_0 is false, and we fail to reject it.

4. We reject H_0 when the p-value is below α.

(a) If H_0 is true, what is the probability of making a Type I error?

Solution: We reject H_0 when the p-value is less than α. This happens when $|T| > t_{\alpha/2,n-1}$. So, if the null hypothesis is true, then the probability of making a Type I error is

$$P(|T| > t_{\alpha/2,n-1}) = \alpha.$$

(b) If H_0 is false, what is the probability of *not* making a Type II error?

Solution: We cannot give a direct answer to this question, because it depends on the true value of μ. In general, the probability of not making a Type II error is called the "power" of the test; it is given the symbol β or $\beta(\mu)$. If μ is close to μ_0, then β will be small (close to α); if μ is far from μ_0, then β will be large (close to 1).

More p-values

5. Suppose we perform a hypothesis test and we observe a p-value of $p = .02$. True or false: There is a 2% chance that the null hypothesis is true.

Solution: False. The p-value is the probability of getting a test statistic at least as extreme as what was observed. Heuristically, we can think of this as

$$\mathrm{P}(\text{Data} \mid H_0 \text{ is true}) = 2\%.$$

The statement in the problem is

$$\mathrm{P}(H_0 \text{ is true} \mid \text{Data}) = 2\%.$$

Clearly, this is not the same.

6. Suppose we perform a hypothesis test and we observe a p-value of $p = .02$. True or false: If we reject the null hypothesis, then there is a 2% chance of making a type I error.

Solution: False. We can only make a type I error when the null hypothesis is true. Thus, the statement in question 6 is *exactly the same* as the statement in question 5.

7. Suppose we perform a hypothesis test and we observe a T test statistic $t = -2.02$, corresponding to a p-value of $p = .02$. True or false: If we were to repeat the experiment and the null hypothesis were actually true, then there would be a 2% chance of observing a test statistic at least as extreme as $t = -2.02$.

Solution: True. The p-value is the probability of getting a test statistic as least as extreme as the observed value if the null hypothesis were true. Note: for a one-sided less-than alternative, extreme means "less than or equal to."