# Homework 2 solution

*Joshua Loftus*

### Question 0

Suppose we are calculating a 95% confidence interval for a mean based on a sample of size $n$.

**a.**

At first, the interval is just too wide to be useful. We decide to increase the sample size to get a narrower interval. What sample size would we need to get an interval that will probably be about half as wide as the original one?

**Solution**: The width of the interval depends on $\frac{1}{\sqrt{n}}$, so if we increase the sample size to $4n$ then the width will be changed by a factor of $1/2$: $\frac{1}{\sqrt{4n}} = \frac{1}{2}\frac{1}{\sqrt{n}}$

**b.**

Suppose $n = 100$, and the confidence interval has the form $\bar{X} \pm cSE$ where $SE$ is the sample standard error of the mean (note: this is not the same as a the sample standard deviation), and the constant $c$ is obtained from the $t$-distribution with $n - 1$ degrees of freedom. For example, for a 95% confidence level, $c$ is

```
qt(.975, 99)
```

```
## [1] 1.984217
```

What is the length of the interval? The answer should be some constant times $SE$. What is the length if we increase the confidence level to 99%? (You will need to use `R` to answer this second part, but can copy the answer to this page by hand).

**Solution**: The width is $(\bar{X} + cSE) - (\bar{X} - cSE) = 2cSE = 2(1.984)SE$

If we increase the confidence level to 99%, the constant $c$ changes to

```
qt(.995, 99)
```

```
## [1] 2.626405
```

So the interval will be a factor of

```
qt(.995, 99)/qt(.975, 99)
```

```
## [1] 1.323648
```

(about 32%) wider.

### Question 1

This question is about whether or not a certain coin is fair. We model the coin as a Bernoulli random variable $C \sim \text{Ber}(p)$, with $P(C = \text{heads}) = p$ and $P(C = \text{tails}) = 1 - p$. We toss the coin $n$ times and record the outcomes as $C_1, C_2, \ldots, C_n$.

**a.**

We want to know if the coin is fair or not. What are the null and alternative hypotheses?

**Solution**: The null hypothesis is $H_0 : p = \frac{1}{2}$ and the alternative is $H_1 : p \neq \frac{1}{2}$.

**b.**

If we toss the coin $n$ times and record how many times it lands on heads, we can use that number as a test statistic. What is the distribution of this test statistic under the null hypothesis?

**Solution**: The number of heads in $n$ coin tosses is binomial, so the distribution under the null is $\text{Bin}\left(n, \frac{1}{2}\right)$.

**c.**

Suppose the observed value of this test statistic is $x$. How could you calculate a $p$-value using the distribution under the null hypothesis?

**Solution**: By calculating the probability a $\text{Bin}\left(n, \frac{1}{2}\right)$ random variable has an outcome at least as extreme as $x$.

**d.**

Suppose $n = 10$ and the number of heads is 9. For the *two-sided alternative*, which outcomes are at least as extreme as 9? Use `pbinom` to calculate the probability, under the null distribution, of seeing an outcome at least as extreme as 9, and write the answer below. Based on this, would you reject the null hypothesis at the 5% significance level? What about at the 1% significance level?

**Solution**: The outcomes that are at least as extreme as 9 out of 10, for the two-sided alternative, are 0, 1, 9, and 10.

```
pbinom(1, 10, 1/2) + (1-pbinom(8, 10, 1/2))
```

```
## [1] 0.02148438
```

We would reject the null hypothesis at the 5% significance level but not at the 1% significance level.

**e.**

Suppose $n = 2$. In this case, what is the smallest possible $p$-value for a one-sided test?

**Solution**: The most extreme outcome for a one-sided test is either 0 successes (less) or 2 successes (greater), and either case has probability 1/4 under the null hypothesis. So the smallest possible $p$-value is 1/4.

## Question 2

In 403 episodes of The Joy of Painting, Bob Ross often painted landscapes that involved "happy trees" and "majestic mountains." See below for some output from `R` of a "cross tabulation" counting how many of the

paintings had no trees or mountains, no trees but mountains, no mountains but trees, and both mountains and trees.

```
counts <- table(bob_ross$trees, bob_ross$mountains)
```

```
##
##            no mountains mountains
##   no trees           61         5
##   trees             243        94
```

```
chisq.test(counts)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  counts
## X-squared = 11.222, df = 1, p-value = 0.0008081
```

What is the null hypothesis for this test? What do you conclude based on the test output?

**Solution**: The null hypothesis is that whether or not a painting has trees and whether or not it has mountains are *independent*. Since the $p$-value is very small, we reject the null hypothesis and conclude that trees and mountains are dependent. (This is true at the 5% significance level, the 1% significance level, and any significance level larger than the $p$-value).

### Question 3

#### a.

The numbers below are from a highly cited paper about neonatal sex differences in "social perception." In that study, the researchers presented neonatal infants with two stimuli to look at: a picture of a face and an object. They recorded how much time each infant spent looking at either of the two stimuli, and recorded preferences as percent of total time spent looking at each. These researchers were interested in seeing if there was a difference in the average preferences between female and male infants. The numbers below are for the percent of time spent looking at the object (not the face). The study included 58 female and 44 male subjects.

```
x1 <- 51.9
s1 <- 23.3
n1 <- 44
x2 <- 40.6
s2 <- 25.0
n2 <- 58
# This next line is just degrees of freedom
# it's a complicated formula, ignore it
df <- (s1^2/n1 + s2^2/n2)^2/((s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1))
c <- qt(.975, df)
SE <- sqrt(s1^2/n1 + s2^2/n2)
c((x1-x2) - c*SE, (x1-x2) + c*SE)
```

```
## [1]  1.756463 20.843537
```

What did this code calculate, what is the meaning of the output? Based on this output, do we reject the null hypothesis that there is no difference in percent time looking at an object at the 5% significance level? Do we reject the null hypothesis at the 1% significance level? For both of these, answer yes, no, or not enough information.

**Solution**: This output is a 95% confidence interval for the difference average male and average female percent time spent looking at an object.

Since this interval does not contain 0, we would reject the null hypothesis at the 5% significance level. For the 1% significance level, we would need a 99% confidence interval, so there is not enough information (unless we went back and calculated it by changing `c` to `qt(.995, df)`).

**b.**

Below is a 95% confidence interval for the difference (male - female) in percent of time spent looking at a face.

```
## [1] -12.692188   5.092188
```

Based on this interval, do we reject the null hypothesis that male babies spent 5% more time looking at a face than female babies at the 5% significance level? (Yes, no, not enough information)

**Solution**: Since this interval contains 5, we cannot reject that null hypothesis.

**c.**

If we have two hypothesis tests computed at the 5% significance level, each one individually has a 5% probability of type 1 error (false rejection). Assuming the tests are independent, and both null hypotheses are true, what is the probability of at least one false rejection? The answer should be a number, but show your work.

**Solution**: The probability of at least one false rejection is $1 - P(\text{no false rejections})$. By independence, $P(\text{no false rejection}) = P(\text{test 1 not a false rejection})P(\text{test 2 not a false rejection}) = 0.95^2$. So the probability of at least one false rejection is

```
1 - (0.95)^2
```

```
## [1] 0.0975
```

or 9.75%. (Note that this is higher than 5%!)

### Question 4

What is $p$-hacking? Give one example of how it might be done in practice. What are some recommendations on how to avoid it?

**Solution**: $p$-hacking means cheating/using tricks to try to get a small $p$-value. But it doesn't have to be malicious- it can be done by accident!

It might be done in practice by measuring and testing many different variables, or testing for differences in lots of different subgroups of the sample, and then only reporting the tests that were significant. Or it could be done by selectively changing the data–increasing the sample size until the $p$-value is significant, removing outliers, and so on.

Recommendations to avoid $p$-hacking

- Have a plan: decide in advance what tests will be done and report the results of all of them
- Be open: keep a notebook (for example in R Markdown) of everything that you try and sharing it to make your analysis reproducible