Midterm practice problem solutions

Joshua Loftus

Study recommendations

- Read the lecture notes posted on the course page. You can mostly skip over `R` code, since there will be no coding on the exam. You should look at the output from the code, like test and confidence interval output, or plots. These notes should be sufficient to answer any exam problems, but you can also review the textbook references linked on the course page.
- Review homework solutions.
- Any topics covered in lectures may appear on the exam. This includes, for example:
  - Confidence intervals and hypothesis tests for a mean, for a difference in means, and a hypothesis test for independence
  - Relationship between sample size, confidence level, and width of confidence interval
  - Null and alternative hypotheses
  - Relationship between intervals and tests
  - Significance level and confidence level
  - Type 1 and type 2 errors
  - Definition, interpretation, and calculation of p-values
  - Use and misuse of p-values
  - Relationship between effect size, sample size, and power
  - Basic idea of how to calculate a large enough sample size
  - Covariance, correlation, and linear relationships
  - Regression lines and relationship with correlation
  - Interpretation and estimation of regression model coefficients

Practice problems

1.

This question is about the `candy_rankings` data in the `fivethirtyeight` package. 85 different candies were presented in matchups to many people, and the percent of time that each candy won the matchups it appeared in was recorded. Several other characteristics of each candy are also included in the data, such as whether or not it contained chocolate, is fruit flavored, is a candy bar, and the percentile of sugar content it has relative to the other candies.

```
# A few rows of data
candy_rankings[1:3, c(1,2,3,9,11,13)]
```

```
## # A tibble: 3 x 6
##   competitorname chocolate fruity bar   sugarpercent winpercent
##   <chr>          <lgl>     <lgl>  <lgl>        <dbl>      <dbl>
## 1 100 Grand      T         F      T            0.732       67.0
## 2 3 Musketeers   T         F      T            0.604       67.6
## 3 One dime       F         F      F            0.0110      32.3
```

a.

Interpret the output of each test below. What is the parameter being tested? What are the null and alternative hypotheses? Do you reject the null hypothesis at the 5% significance level? If you do not reject the null, give an example of a different significance value at which you would reject. If you do reject the null,

give an example of a different null hypothesis that you would not reject. If the test is for a difference, explain why it would or would not make sense to do the paired version of the test.

```
t.test(candy_rankings$winpercent, mu = 50, alternative = "less")
```

```
##
##  One Sample t-test
##
## data:  candy_rankings$winpercent
## t = 0.19847, df = 84, p-value = 0.5784
## alternative hypothesis: true mean is less than 50
## 95 percent confidence interval:
##      -Inf 52.97122
## sample estimates:
## mean of x
##  50.31676
```

Solution: The parameter is the average μ of the win percent variable. The null hypothesis is $H_0 : \mu = 50$ and the alternative is $H_1 : \mu < 50$. Since the p-value is greater than 0.05, we do not reject the null at the 5% significance level. If we used a significance level of 60% we would reject the null.

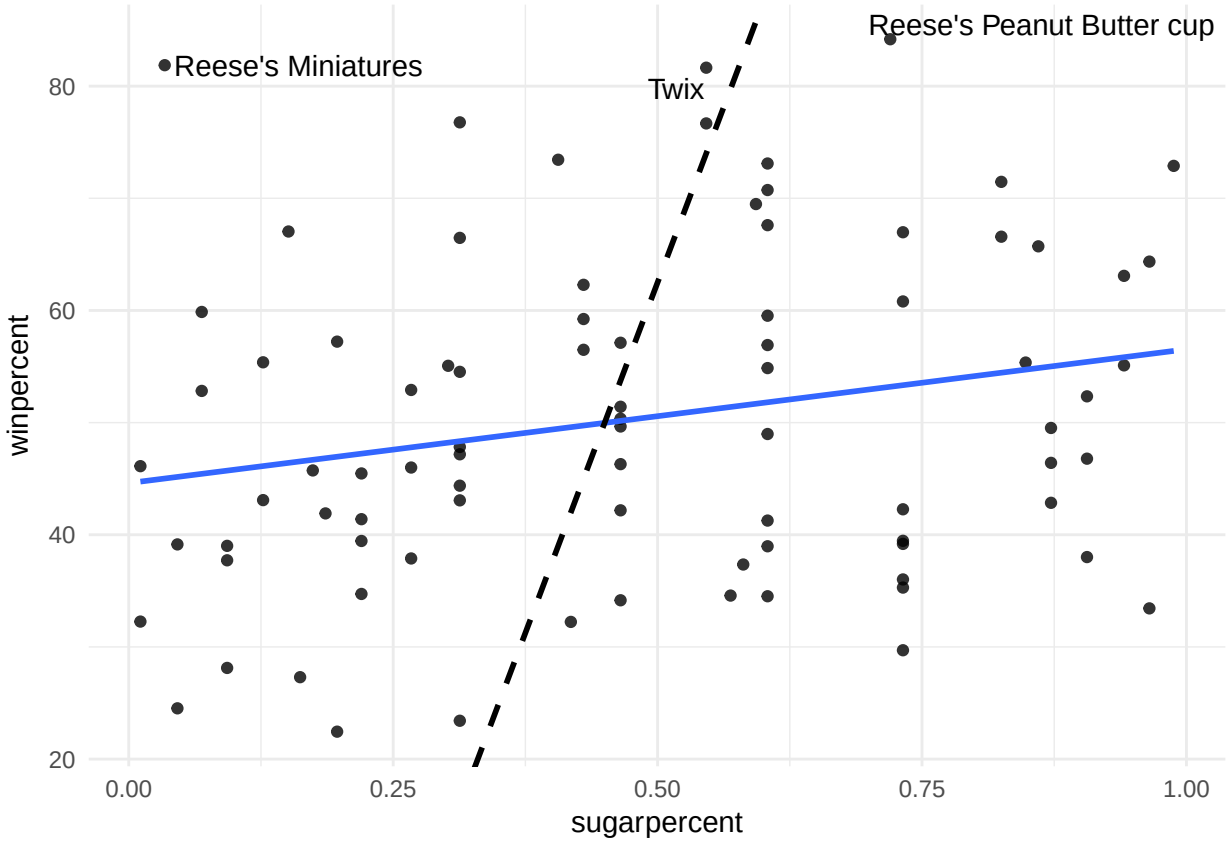
```
t.test(candy_rankings$winpercent ~ candy_rankings$chocolate)
```

```
##
##  Welch Two Sample t-test
##
## data:  candy_rankings$winpercent by candy_rankings$chocolate
## t = -7.3031, df = 67.539, p-value = 4.164e-10
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -23.91110 -13.64744
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            42.14226            60.92153
```

Solution: The parameter is the difference $\mu_1 - \mu_2$ in average win percent between candies without chocolate and candies with chocolate. The null hypothesis is $H_0 : \mu_1 - \mu_2 = 0$ and the alternative hypothesis is $H_1 : \mu_1 - \mu_2 \neq 0$. Since the p-value is less than 0.05 we reject the null at the 5% significance level. A null hypothesis of any value inside the confidence interval (from about -23.9 to about -13.6) would not be rejected, so $H_0 : \mu_1 - \mu_2 = -20$ is an example that we would not reject. It would not make sense to do a paired version because we are comparing different candies to each other, and not two aspects of the same candy. We would not know which chocolate candy to pair with which non-chocolate candy.

b.

Below is a scatterplot with sugar percentile on the x axis and win percent on the y axis. A few of the most popular candies are labeled. The plot also includes the linear regression line predicting win percent from sugar percentile (solid blue line) and the linear regression line predicting sugar percentile from win percent (dashed black line). Answer each of the following questions. For true/false questions, if you answer false, explain why.



- Which of the 3 most popular candies has the most amount of sugar? **Reese's Peanut Butter cup**
- True/false: The positive slope of the regression line means that the correlation is close to 1. **False. It means correlation is bigger than 0**.
- Which of the following is closest to the correlation: -1, -0.7, -0.3, **0.3**, 0.7, 1
- Which of the following is closest to the intercept: 0.50, 20, **45**, 51
- True/false: The regression slope tells us how much the average win percent increases if sugarpercent increases from 0 to 1. **True**
- The vertical distance from Reese's Peanuter Butter cup to the solid blue line is about 30. **True**
- True/false: The solid blue line indicates a higher correlation because points are further away from the dashed black line. **False. Correlation is symmetric**
- Draw a line segment from Reese's Peanut Butter cup indicating its error when predicting sugar percentile using win percent as a predictor. About how large is this error? **(draw a horizontal line from Reese's Peanut Butter cup to the dashed black line) About 0.25**
- Draw a line segment from Reese's Miniatures indicating its error when predicting win percent using sugar percentile as a predictor. About how large is this error? **(draw a vertical line from Reese's Miniatures to the solid blue line) About 38**
- Which of the two regression models has a lower sum of squared errors, the one predicting win percent or the one predicting sugar percentile? **The one predicting sugar percentile**
- True/false: the regression model shows that higher amounts of sugar cause people to like that candy more. **False. Association is not causation**

2.

a.

For random variables X and Y, the definition of covariance in the model/population world is

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

and the definition of correlation is

$$\text{Cor}(X, Y) = \frac{\text{Cov(X,Y)}}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Use these definitions to answer the following, showing your work:

- Simplify $\text{Cov}(X, 2Y)$ (hint: $E[2Y] = 2E[Y]$).

Solution:

$$\text{Cov}(X, 2Y) = E[(X - E[X])(2Y - 2E[Y])] = 2E[(X - E[x])(Y - E[Y])] = 2\text{Cov}(X, Y)$$

- What is $\text{Cor}(X, -X)$? (hint: $\text{Var}(-X) = \text{Var}(X)$).

Solution:

$$\text{Cor}(X, -X) = \frac{\text{Cov(X,-X)}}{\sqrt{\text{Var}(X)\text{Var}(-X)}} = \frac{-\text{Cov(X,X)}}{\sqrt{\text{Var}(X)^2}} = \frac{-\text{Var(X)}}{\sqrt{\text{Var}(X)^2}} = -1$$

b.

The sample/data world definition of covariance is

$$\text{cov}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \text{ where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and similarly for } \bar{y}$$

and the sample correlation is

$$\text{cor}(x, y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right), \text{ where } s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and similarly for } s_y$$

Use these definitions to answer the following, showing your work:

- If we transform x to $1 - x$, what happens to the sample covariance? (Hint: the sample mean becomes $1 - \bar{x}$)

Solution:

$$\text{cov}(1 - x, y) = \frac{1}{n-1}\sum_{i=1}^{n}(1 - x_i - (1 - \bar{x}))(y_i - \bar{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(-x_i + \bar{x}))(y_i - \bar{y}) = -\text{cov}(x, y)$$

- If we transform y to $y + 5$, what happens to $\text{cor}(x, y)$? (hint: s_y does not change)

Solution:

$$\text{cor}(x, y + 5) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i + 5 - (\bar{y} + 5)}{s_y} \right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) = \text{cor}(x, y)$$

3.

You are consulting for a company that wants to assess different versions of its product. The company obtains data by presenting each version to a different set of its customers and asking them to rate the quality. The company has many customers, but a limited budget for this project. Describe what statistical method you would use in each scenario below, giving as much detail as you can, and anticipating any practical aspects of the process that you have learned about.

a.

The company's standard product is yellow, but they are considering introducing a blue version. They want to know if the average quality rating for the blue one is higher, but they don't want to begin production unless they are sure about the quality increase. How can they answer this question?

Solution: The company could do a hypothesis test for the difference in quality between the yellow and blue versions. They null hypothesis will be no difference, and the alternative hypothesis will be that the blue version has higher quality. They should choose a (statistical) significance level depending on how certain they want to be about the decision, and an effect size that is the smallest increase in quality they would be interested in pursuing (practical significance). Then they could determine how large of a sample to collect to have a reasonably high probability of rejecting the null hypothesis if there is a practically significant effect. They should plan the analysis as comprehensively as possible before beginning to collect the data, and keep a detailed record of the analysis (for example in an `R` Markdown document). They should consider plotting the quality scores to see if the average quality is a reasonable summary of the quality distribution (check for skew, outliers, multiple modes).

b.

There are versions of the product with different numbers of decorative buttons, ranging from 2 up to 20 buttons. Before designing a new product inspired by the current one, the company wishes to find out if the quality rating is higher or lower for a larger number of buttons. What could they do to answer this question?

Solution: The company could calculate the covariance, correlation, or regression line (and see if the slope is positive) to determine if there is a positive relationship between number of buttons and quality rating. Since these are all measures of a linear relationship, they should also plot the data to check for non-linearity or other issues (like skew or outliers, for example).