Practice for second midterm exam

Solution

Q1

A startup develops algorithms to determine if an article is "fake news." They do this by defining a parameter θ representing the trustworthiness of the given article. Their algorithms input the text of an article and output estimates $\hat{\theta}$. Engineers develop two candidate algorithms: one using advanced deep learning methods $\hat{\theta}_{\text{DL}}$, and one using a simpler model called logistic regression $\hat{\theta}_{\text{LR}}$. The DL estimator is unbiased, but has a variance equal to 1. The LR estimator has a bias of $-\frac{1}{2}$ and a variance of $\frac{1}{2}$. What is the MSE of the LR estimator? Which method has the lower MSE?

$\text{MSE}(\hat{\theta}_{\text{LR}}) = \text{Bias}(\hat{\theta}_{\text{LR}})^2 + \text{Var}(\hat{\theta}_{\text{LR}}) = (-1/2)^2 + 1/2 = 3/4$

And $\text{MSE}(\hat{\theta}_{\text{DL}}) = 0^2 + 1 = 1$ so the LR estimator has lower MSE.

Q2

Suppose population household income in the US has mean $\mu = \$70,000$ and standard deviation $\sigma = \$30,000$. In this problem we know these true parameters. We survey households randomly and collect a sample of size $n = 100$, and let $\bar{X}$ denote the mean income of the sample.

Part (a)

How would you use the normal distribution to approximate $P(\bar{X} < \$64,000)$?

$P(\bar{X} < 64000) \approx P(Z < 64000)$ for $Z \sim N(70000, 30000^2/100)$.

Since $64000 = 70000 - 2 \cdot 3000$ this is the probability of normal being less than 2 standard deviations below its mean.

We know probability of being outside 2 standard deviations from the mean is 5%, and this is split evenly between upper and lower tails. So $P(\bar{X} < 64000) \approx 2.5\%$.

Part (b)

Why can you do this even though the distribution of incomes is not normal? (We know it is not normal because it is skewed)

We can do this because of the central limit theorem! The sample mean is (approximately) normal even if the original random variables are not.

Q3

The standard deviation of a random variable is σ and the standard error of the mean of an i.i.d. sample of size n (for some $n > 1$) of this random variable is SE. Which of the following are true? Indicate with a check mark.

- $\sigma = SE$

- $\sigma > SE$ ✓
- σ decreases as n increases
- SE decreases as n increases ✓
- If the sample increases from n to $2n$, then SE decreases to $SE/2$
- If the sample increases from n to $2n$, then σ decreases to $\sigma/2$
- Neither σ nor SE decrease as n increases

(Remember: $SE = \sigma/\sqrt{n}$)

Q4

Suppose $U_1, U_2, \ldots, U_n$ are i.i.d., $E[U_1] = \mu$, $\text{Var}(U_1) = \sigma^2$, let $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \bar{U})^2$. What is the T statistic for this sample? Suppose $n = 100$, $\bar{U} = 1.8$, $S = 10$, and $P(T > 2.62) = 0.005$ (so $P(T < -2.62) = 0.005$).

Part (a)

What is the 99% confidence interval for μ based on this sample? (You don't need to simplify expressions with numbers)

$$T = \frac{\bar{U} - \mu}{SE} = \frac{\bar{U} - \mu}{S/\sqrt{n}} = \frac{1.8 - \mu}{10/\sqrt{100}}$$

99% confidence interval: $\bar{U} \pm 2.62SE = \bar{U} \pm 2.62$

Part (b)

True or false, and explain: the probability that the interval computed above contains μ is 99%.

False. Once we observe specific numbers for the random variables, the interval either containts μ or does not. The **frequentist interpretation of confidence intervals** means that if we **repeated** the experiment by collecting another sample, calculating the interval for that sample, and repeating many times, then about 99% of the resulting intervals would contain the true μ.

Q5

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.

```
# First 3 rows of data
candy_rankings[1:3, c(1,2,4,9,11,13)]
```

```
## # A tibble: 3 x 6
##   competitorname chocolate caramel bar   sugarpercent winpercent
##   <chr>          <lgl>     <lgl>   <lgl>        <dbl>      <dbl>
## 1 100 Grand      T         T       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
```

Part (a)

Interpret the output of each test below. (1) What is the parameter being tested (NA if this does not apply)? (2) What are the null and alternative hypotheses? (3) Do you reject the null hypothesis at the 5% significance level? (4) Write one sentence explaining very simply the practical conclusion you make based on the test.

```
tab <- table(candy_rankings$chocolate, candy_rankings$caramel)
rownames(tab) <- c("Chocolate", "No choc.")
colnames(tab) <- c("Caramel", "No cara.")
tab
```

```
##
##             Caramel No cara.
##   Chocolate      44        4
##   No choc.       27       10
```

```
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tab
## X-squared = 4.0354, df = 1, p-value = 0.04456
```

(1) NA, no parameter. (2) H_0 : the chocolate and caramel (categorical) variables are independent, and H_1 : they are dependent. (3) Yes, because the p-value is less than 0.05. (4) Based on this test, we conclude that whether a candy contains chocolate and whether it contains caramel are statistically dependent.

In addition to the above questions, for the next tests answer the following: (5a) If you do not reject the null, give an example of a different significance value at which you would reject. (5b) If you do reject the null, give an example of a different null hypothesis that you would not reject.

```
t.test(candy_rankings$chocolate, mu = 1/2)
```

```
##
##  One Sample t-test
##
## data:  candy_rankings$chocolate
## t = -1.1961, df = 84, p-value = 0.235
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.3277188 0.5428694
## sample estimates:
## mean of x
## 0.4352941
```

(1) The parameter μ is the population mean of the chocolate variable–i.e. the proportion of candies that contain chocolate, (2) $H_0 : \mu = 1/2$, and $H_1 : \mu \neq 1/2$ (two-sided), (3) we do not reject the null, (4) this sample of data is consistent with the hypothesis that half of the candies in the population have chocolate, (5a) any value outside of the confidence interval would be rejected, so for example $H_0 : \mu = 0.55$ or $H_0 : \mu = 0.3$ would both be rejected.

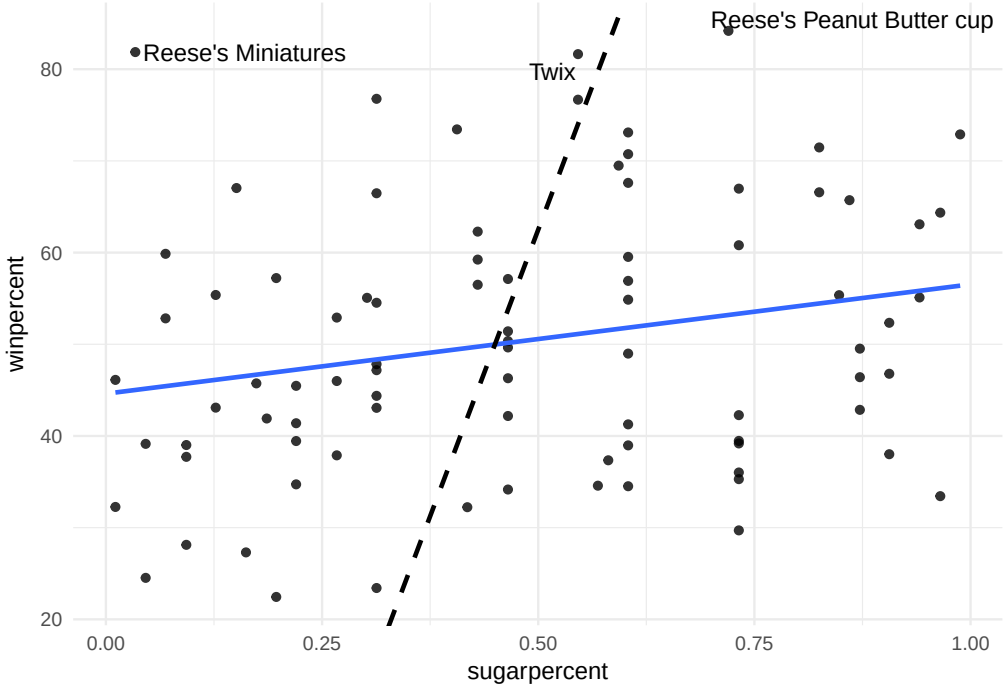
```r
t.test(candy_rankings$winpercent ~ candy_rankings$chocolate)
```

```
##
##  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
```

(1) The parameter $\mu = \mu_1 - \mu_2$ is the difference in population means of the `winpercent` variable between the candies that don't have chocolate μ_1 and the candies that do μ_2, (2) $H_0 : \mu = 0$ i.e. $H_0 : \mu_1 = \mu_2$, and $H_1 : \mu_1 \neq \mu_2$, (3) we do reject the null since the p-value is less than 0.05, (4) we conclude that candies with chocolate and without do not have the same popularity, (5b) any value in the interval would not be rejected, e.g. $H_0 : \mu = -20$ i.e. $\mu_1 + 20 = \mu_2$ would not be rejected.

Part (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**

Part (c)

In the linear model below, let β_0 be the intercept and β_1 the slope of the predictor variable `sugarpercent`.

```
model <- lm(winpercent ~ sugarpercent, data = candy_rankings)
# Model summary
summary(model)
```

```
##
## Call:
## lm(formula = winpercent ~ sugarpercent, data = candy_rankings)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.924 -11.066  -1.168   9.252  36.851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    44.609      3.086  14.455   <2e-16 ***
## sugarpercent   11.924      5.560   2.145   0.0349 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.41 on 83 degrees of freedom
## Multiple R-squared:  0.05251,    Adjusted R-squared:  0.04109
## F-statistic:   4.6 on 1 and 83 DF,  p-value: 0.0349
```

```r
# 95% confidence intervals
confint(model)
```

```
##                   2.5 %   97.5 %
## (Intercept)  38.4713409 50.74754
## sugarpercent  0.8660272 22.98170
```

- What is the value of $\hat{\beta}_0$?

$\hat{\beta}_0$ is the estimated intercept ≈ 44.61

- What is the value of $\hat{\beta}_1$?

$\hat{\beta}_1$ is the estimated slope of `sugarpercent` ≈ 11.92

- What is the null hypothesis for β_1? Do we reject this null hypothesis at the 5% significance level?

$H_0 : \beta_1 = 0$, and yes, we reject this null at the 5% significance level since the p-value $\approx 0.035 < 0.05$.

- True or false, and explain: the 95% confidence interval for β_1 contains 20

True, because $0.866 < 20 < 22.98$

- True or false, and explain: the 99% confidence interval for β_1 contains 0

True, because the p-value for $H_0 : \beta_1 = 0$ is larger than 0.01, so we would fail to reject that hypothesis at the 1% significance level, hence the 99% confidence interval contains 0.

Q6

Part (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:

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

Part (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:

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

Q7

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.

Part (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. 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).

Part (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).