

Practice for second midterm exam

I (name/ID) agree to abide by the Stern Code of Conduct _____

You must work alone and finish this practice exam to receive full homework credit

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?

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 $P(\bar{X} < \$64,000)$?

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)

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

Q4

Suppose U_1, U_2, \dots, 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)

Part (b)

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

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
```

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)
```

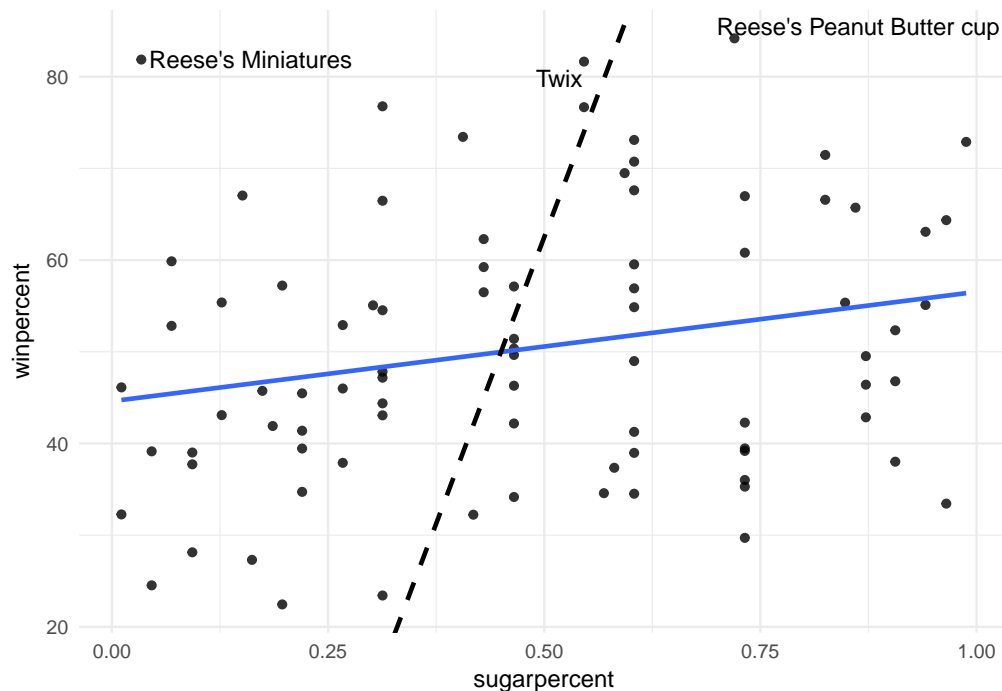
```
##
## One Sample t-test
##
## data:  candy_rankings$chocolate
## t = -1.1961, df = 84, p-value = 0.235
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.3277188 0.5428694
## sample estimates:
## mean of x
## 0.4352941
```

```
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
```

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?
- True/false: The positive slope of the regression line means that the correlation is close to 1.
- 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.
- The vertical distance from Reese's Peanut Butter cup to the solid blue line is about 30.
- True/false: The solid blue line indicates a higher correlation because points are further away from the dashed black line.
- 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 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?

- Which of the two regression models has a lower sum of squared errors, the one predicting win percent or the one predicting sugar percentile?
- True/false: the regression model shows that higher amounts of sugar cause people to like that candy more.

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
# 95% confidence intervals
confint(model)

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

- What is the value of $\hat{\beta}_0$?
- What is the value of $\hat{\beta}_1$?
- What is the null hypothesis for β_1 ? Do we reject this null hypothesis at the 5% significance level?
- True or false, and explain: the 95% confidence interval for β_1 contains 20

- True or false, and explain: the 99% confidence interval for β_1 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: $E[2Y] = 2E[Y]$).
- What is $\text{Cor}(X, -X)$? (hint: $\text{Var}(-X) = \text{Var}(X)$).

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 $1 - \bar{x}$)

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

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?

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?