Swarthmore Honors Examination in Statistics

May 2024

Instructions:

1. You will have three hours to complete this exam.
2. This is a closed-materials exam. You may not refer to any books, notes, online sources, software, or other resources, except for the simple calculator provided to do arithmetic.
3. Show your work and explain your reasoning with sufficient justification.
4. Write your solutions neatly on separate sheets of paper. Label your responses clearly (1a, 1b, 1c, 2a, etc.).
5. If you cannot answer part of a problem, but need its answer for a later part, just make up a reasonable value and use it. For example "I can't do part (a), so I am going to assume that the answer is 0.2 so I can do part (b)." This applies to multiple steps within a single part and non-numerical answers as well. ("I'm not sure about the analysis but I'll assume we reject the null hypothesis to make my conclusion.") If you get stuck, do whatever you need to do to keep making progress (e.g., make stuff up).
6. If there are parts that contain computations that you are unable to do by hand or with a simple calculator, you can answer these parts by explaining how you would perform the computations. For example, you can provide commands you would use if you had access to software (e.g, "I would find the value of z using *qnorm(0.975, 0, 1))* in *R* and then I would do $100 + 10z$"). You can also draw well labeled pictures representing how would you perform the computations.
7. I recommend that you read through the entire exam first and then start with questions you feel most confident on to make the best use of your time.
8. Good luck!

I have not committed any form of academic dishonesty in preparing for and/or taking this exam.

Signature: _____

1. It can be hard to get people to respond to surveys, especially when they ask sensitive questions. Randomized response is an approach that can help participants feel comfortable answering questions while also ensuring that researchers can get quality estimates of what they are studying.

   Participants are asked to flip a coin before responding to the question: "Have you ever cheated on an exam?". If the coin lands on tails, they answer "yes" to the question. If the coin lands on heads, they answer truthfully. Because the researcher does not see the coin flip result, they do not know which individuals are telling the truth.

   - **(a.)** Suppose z is the proportion of people in the survey that answer "yes" under this approach. What is your best estimate of the proportion of people in the survey who have actually cheated on an exam? Show your work.
   - **(b.)** Are there any advantages to using a weighted coin rather than one that has an equal probability of landing on heads and tails? Consider a coin that comes up Heads 10% of the time and one that comes up Heads 90% of the time as you explain your reasoning.

2. Explain how you would estimate the *sampling* distribution for the sample standard deviation of a sample of size 50 from an unknown distribution. You can assume that you are given a sample of size 50 from this distribution, and you can use simulation. How would you estimate this sampling distribution's mean and standard error? You can write pseudo-code or describe in words, just make sure you have given me enough information to translate your words into code.

3. Given two estimators, how do you decide which one is best? Write a paragraph in response. You should describe at least two metrics of evaluation.

4. Suppose we are in a hypothesis test framework where the distribution under the null hypothesis is equal to $\frac{1}{2}$ and the distribution under the alternative hypothesis is equal to $-\frac{1}{2}x + 1$. Note: Our domain is the interval between 0 and 2.

   - **(a.)** Graph both of these distributions on the same plot.
   - **(b.)** How do you know that these are appropriate probability distributions?
   - **(c.)** We will use a rejection of the form $\{X < k\}$. On your graph from (a.), indicate the regions that represent the Type I and Type II probabilities of error.
   - **(d.)** Find k such that the test has level 0.02.
   - **(e.)** For the rejection region found in (d.), find the probability of Type II error and interpret this value in the context of the problem.

5. Please base your responses to each question on the R code provided below:

```
counter <- 0
n <- 100

for(i in 1:1000){

  x <- rpois(n, 7)
  L <- mean(x) - qnorm(.95, 0, 1)*(sqrt(7)/sqrt(n))
  U <- mean(x) + qnorm(.95, 0, 1)*(sqrt(7)/sqrt(n))

    if (L < 7 && U > 7){
      counter <- counter + 1
     }

}
```

Note that the *rpois* function has two parameters: the first is the sample size, and the second is the mean. The function generates random numbers from the Poisson distribution. The *qnorm* function has three parameters: the first is a probability (calculated as area to the left of a value), the second is the mean, and the third is the standard deviation. The function calculates a quantile from a normal distribution.

- **(a.)** What are L and U calculating? Be specific.
- **(b.)** What is the counter measuring?
- **(c.)** What value do you expect to see from counter/1000 and why?
- **(d.)** Would your answer in (c.) change if $n = 10$? Why or why not?
- **(e.)** Would your answer in (c.) change if all of the 7s were changed to 75s (using the original $n = 100$)? Why or why not?

6. We have data on avocado producers in two parts of the United States: California and the Central South Region (of the US) and we want to be able to predict the total bags (in thousands) produced using other information about the producers.

   Data Description

   Each data point is a producer of avocados in either California or the Central South Region.

   Response: TotalBags1000 represents the total bags produced (in thousands)

   Possible covariates:

   - regionSouthCentral: 1 if region is Central South of the US, 0 if region is California
   - year - growing year
   - AveragePrice - average price ($) for an avocado
   - TotalVolume1000 - total volume produced (in 1000s of cubic units)
   - AveragePrice:regionSouthCentral - interaction of AveragePrice and regionSouthCentral
   - regionSouthCentral:year - interaction of regionSouthCentral and year
   - smallSold1000 - bags of small avocados sold (in 1000s)

   Note that the model labels are *below* the model output and there are 7 total models.

```
Call:
lm(formula = TotalBags1000 ~ TotalVolume1000, data = avocados)

Residuals:
    Min      1Q  Median      3Q     Max
-1879.7  -468.9   104.0   370.9  1542.0

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -755.40051  154.48244   -4.89 1.57e-06 ***
TotalVolume1000    0.35819    0.02572   13.93  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 511.1 on 336 degrees of freedom
Multiple R-squared:  0.3659,    Adjusted R-squared:  0.364
F-statistic: 193.9 on 1 and 336 DF,  p-value: < 2.2e-16
```

Model 1

```
Call:
lm(formula = TotalBags1000 ~ AveragePrice + region + TotalVolume1000,
    data = avocados)

Residuals:
     Min       1Q   Median       3Q      Max
-1997.69  -355.97    67.61   358.38  1398.57

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2.587e+03  3.439e+02  -7.522 5.05e-13 ***
AveragePrice     1.071e+03  1.770e+02   6.050 3.89e-09 ***
regionSouthCentral  1.886e+02  6.810e+01   2.769  0.00593 **
TotalVolume1000     4.732e-01  3.107e-02  15.234  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 485.5 on 334 degrees of freedom
Multiple R-squared:  0.4313,    Adjusted R-squared:  0.4262
F-statistic: 84.45 on 3 and 334 DF,  p-value: < 2.2e-16
```

Model 2

```
Call:
lm(formula = TotalBags1000 ~ AveragePrice * region + TotalVolume1000,
    data = avocados)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.26  -312.42    28.52   323.80  1530.67

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.292e+03  3.450e+02  -6.645 1.24e-10 ***
AveragePrice                     7.678e+02  1.897e+02   4.047 6.45e-05 ***
regionSouthCentral              -9.339e+02  2.931e+02  -3.186 0.001577 **
TotalVolume1000                  4.801e-01  3.046e-02  15.759  < 2e-16 ***
AveragePrice:regionSouthCentral  1.209e+03  3.074e+02   3.933 0.000102 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 475.3 on 333 degrees of freedom
Multiple R-squared:  0.4566,    Adjusted R-squared:  0.4501
F-statistic: 69.95 on 4 and 333 DF,  p-value: < 2.2e-16
```

Model 3

```
Call:
lm(formula = TotalBags1000 ~ region * year, data = avocados)

Residuals:
    Min      1Q  Median      3Q     Max
-942.50 -326.06   -9.88  242.89 1713.70

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -806833.20   71962.89 -11.212  < 2e-16 ***
regionSouthCentral      -360552.35  101770.89  -3.543 0.000452 ***
year                        400.88      35.69  11.231  < 2e-16 ***
regionSouthCentral:year     178.79      50.48   3.542 0.000454 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 436.1 on 334 degrees of freedom
Multiple R-squared:  0.5411,    Adjusted R-squared:  0.5369
F-statistic: 131.3 on 3 and 334 DF,  p-value: < 2.2e-16
```

Model 4

```
Call:
lm(formula = TotalBags1000 ~ region * year + AveragePrice, data = avocados)

Residuals:
    Min      1Q  Median      3Q     Max
-825.71 -210.20   -2.57  213.32 1293.86

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1.022e+06  6.130e+04 -16.668  < 2e-16 ***
regionSouthCentral     -2.520e+05  8.382e+04  -3.007  0.00284 **
year                    5.083e+02  3.042e+01  16.707  < 2e-16 ***
AveragePrice           -1.381e+03  1.077e+02 -12.823  < 2e-16 ***
regionSouthCentral:year  1.248e+02  4.158e+01   3.002  0.00289 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 357.4 on 333 degrees of freedom
Multiple R-squared:  0.6928,    Adjusted R-squared:  0.6891
F-statistic: 187.7 on 4 and 333 DF,  p-value: < 2.2e-16
```

Model 5

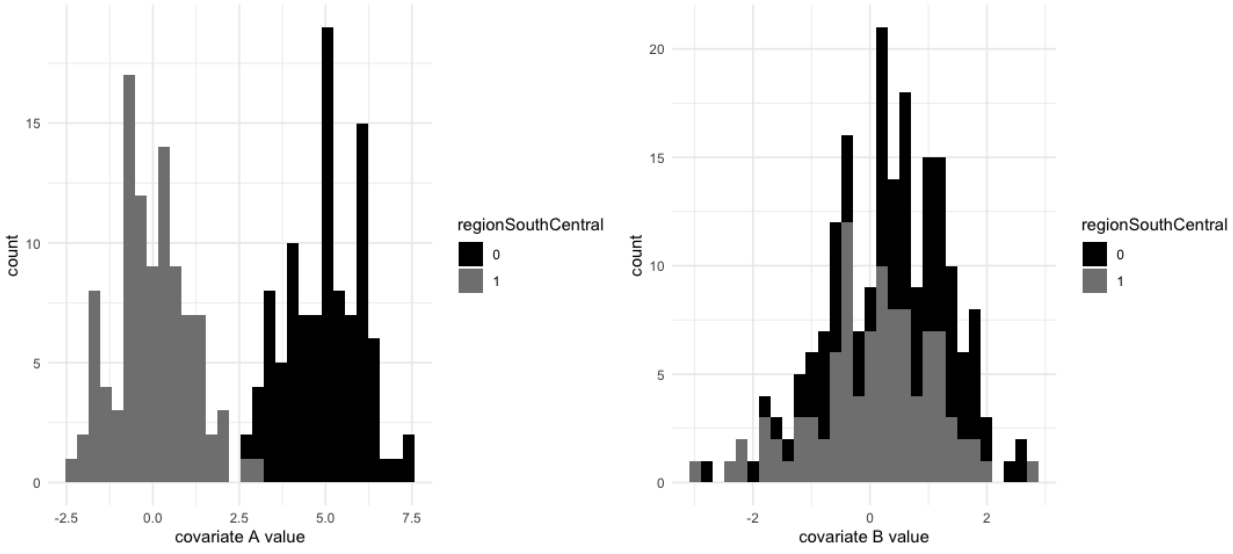
```
Call:
lm(formula = TotalBags1000 ~ smallSold1000 + year, data = avocados)

Residuals:
   Min     1Q Median     3Q    Max
-919.4 -250.4 -117.4  251.0 1758.1

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -9.913e+05  5.193e+04 -19.090   <2e-16 ***
smallSold1000  5.088e-02  3.276e-02   1.553    0.121
year           4.923e+02  2.575e+01  19.115   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 444.5 on 335 degrees of freedom
Multiple R-squared:  0.5219,    Adjusted R-squared:  0.5191
F-statistic: 182.9 on 2 and 335 DF,  p-value: < 2.2e-16
```

Model 6

```
Call:
lm(formula = TotalBags1000 ~ AveragePrice * region + Total.Volume,
    data = avocados)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.26  -312.42    28.52   323.80  1530.67

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2.292e+03  3.450e+02  -6.645 1.24e-10 ***
AveragePrice                   7.678e+02  1.897e+02   4.047 6.45e-05 ***
regionSouthCentral            -9.339e+02  2.931e+02  -3.186 0.001577 **
Total.Volume                   4.801e-04  3.046e-05  15.759  < 2e-16 ***
AveragePrice:regionSouthCentral  1.209e+03  3.074e+02   3.933 0.000102 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 475.3 on 333 degrees of freedom
Multiple R-squared:  0.4566,    Adjusted R-squared:  0.4501
F-statistic: 69.95 on 4 and 333 DF,  p-value: < 2.2e-16
```

Model 7

- **(a.)** Provide an interpretation (in the context of the problem) of the highlighted number (0.48) in Model 3.
- **(b.)** Write the hypotheses that are being tested to obtain the p-value highlighted in Model 6 (<2e-16). Be sure to define any parameters you use in the context of the problem.
- **(c.)** In any of the model outputs, circle a coefficient that is statistically significant but not practically significant. Explain your reasoning in a sentence or two.
- **(d.)** Consider all of the models presented. Which of them would you choose as the "best" model of total bags produced? Convince me using at least two pieces of evidence.
- **(e.)** Now suppose we are interested in modeling the response *regionSouthCentral*. Which covariate, A or B, do you think would be a stronger predictor of this response? Explain your reasoning in a sentence or two.



7. We have data on the number of divorce records (across the US in 1987) for a variety of age groups in 32 of the 50 states. The following table provides summary statistics for each group.

Age group	Husband (sample mean # records)	Wife (sample mean # records)	Husband (sample std dev. of # records)	Wife (sample std dev. of # records)
<20	27.28	120.41	39.02	165.21
[20, 24]	715.88	1229.31	1016.46	1816.37
[25, 29]	1692.91	1911.75	2687.88	3119.73
[30, 34]	1787.94	1730.00	3029.49	2929.72
[35, 39]	1491.84	1366.59	2538.23	2311.37
[40, 44]	1104.25	957.50	1873.20	1665.19
[45, 49]	701.78	532.28	1211.89	964.88
50+	973.28	607.81	1740.91	1105.07

- **(a.)** Use the data provided to answer the following question: Does there seem to be a difference in average number of records of divorce between men and women in the [50+] age category? Justify any decisions you make about your approach. If you get stuck, feel free to make assumptions that simplify the approach but be clear that you are doing this and why.
- **(b.)** Use the data provided to answer the following question: does age appear to be associated with the difference in average number of records of divorce among wives overall? You may pick a subset of the data to analyze, but again justify any decisions you make about your approach. Again, if you get stuck, feel free to make assumptions that simplify the approach but be clear that you are doing this and why.
- **(c.)** What would you need to know in order to feel comfortable generalizing your findings from parts (a.) and (b.) to all 50 states?

8. What is something you did to prepare for this exam that helped you with one or more of these questions? Your answer can just be a sentence long.