Statistics 324 Homework 11

Svadrut Kukunooru

*Submit your homework to Canvas by the due date and time.

*If an exercise asks you to use R, include a copy of the code and output. Please edit your code and output to be only the relevant portions.

*If a problem does not specify how to compute the answer, you many use any appropriate method. I may ask you to use R or use manual calculations on your exams, so practice accordingly.

*You must include an explanation and/or intermediate calculations for an exercise to be complete.

*Be sure to submit the HWK 11 Auto grade Quiz which will give you ~20 of your 40 accuracy points.

*50 points total: 40 points accuracy, and 10 points completion

Exercise 1 Reconsider the relationship between city air particulate and rates of childhood asthma first discussed in HWK 10. We sampled 15 cities for particulate measured in parts-per-million (ppm) of large particulate matter and for the rate of childhood asthma measured in percents. The data are as follows:

variable															
part (x)	11.6	15.9	15.7	7.9	6.3	13.7	13.1	10.8	6.0	7.6	14.8	7.4	16.2	13.1	11.2
asth (y)	14.5	16.6	16.5	12.6	12.0	15.8	15.1	14.2	12.2	13.1	16.0	12.9	16.4	15.4	14.4

variable:	size	mean	variance
particulate	15	11.42	13.05029
asthma	15	14.51333	2.635524

```r
particulate = c(11.6, 15.9, 15.7, 7.9, 6.3, 13.7, 13.1, 10.8, 6.0, 7.6, 14.8, 7.4, 16.2, 13.1, 11.2)
asthma = c(14.5, 16.6, 16.5, 12.6, 12.0, 15.8, 15.1, 14.2, 12.2, 13.1, 16, 12.9, 16.4, 15.4, 14.4)
model = lm(asthma ~ particulate)
summary(model)
```

```
##
## Call:
## lm(formula = asthma ~ particulate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34226 -0.12842 -0.01514  0.12130  0.29164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.41626    0.17344   54.29  < 2e-16 ***
## particulate  0.44633    0.01452   30.73  1.6e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1963 on 13 degrees of freedom
## Multiple R-squared:  0.9864, Adjusted R-squared:  0.9854
## F-statistic: 944.4 on 1 and 13 DF,  p-value: 1.595e-13
```

a. Suppose we sample a new city whose particulate is **13 ppm**. Create a 95% interval for the predicted rate of childhood asthma in this city.

Point Estimate

$$b_0 = 9.4163, b_1 = 0.4463$$
$$y = 9.4163 + 0.4463 \times 13 = \boxed{15.2182}$$

Critical Value

For a 95% confidence interval, the CV is 1.96.

Standard Error for Estimator

$$0.1963$$

Margin of Error

$$CV \times SE = 1.96 \times 0.1963 = 0.384748$$

Final Interval

$$15.2182 \pm 0.384748$$

b. Create a *95% confidence interval* for the average rate of childhood asthma among cities with **10 ppm** of large particulate. Is this confidence interval wider or narrower than a *95% prediction interval* for the rate of childhood asthma in the next city with **10 ppm** of large particulate? Explain why their relative sizes makes sense.

Point Estimate

$$b_0 = \bar{y} - b_1\bar{x}$$
$$b_1 = 0.3753, b_0 = 10.3515$$
$$\hat{y} = b_0 + b_1 x$$
$$\hat{y} = b_0 + b_1 x$$
$$\hat{y} = 10.3515 + 0.3753 \times 10 = 14.1045$$

Critical Value

For a 95% confidence interval, the CV is 1.96

Standard Error for Estimator

$$SE = \sqrt{\sum_{i=1}^{15} (y_i - \hat{y}_i)^2/(n-2)} = \boxed{1.1493}$$

Margin of Error

$$ME = CV \times SE = 1.96 \times 1.1493 = \boxed{2.2526}$$

Final Interval

$$14.1045 \pm 2.2526$$

The confidence interval is narrower than the prediction interval. This is expected because prediction intervals are wider than confidence intervals. Prediction intervals account for both the variability of the regression line and the variability of individual data points, while confidence intervals only consider the variability of the regression line.

c. Explain why is is not reasonable to construct a 95% interval for the predicted rate of childhood asthma in the next city sampled that has **3 ppm** of large particulate from our model.

Extrapolating beyond the range of data introduces uncertainty to the calculation.

```r
father = c(71.3, 65.5, 65.9, 68.6, 71.4, 68.4, 65.0, 66.3, 68.0, 67.3, 67.0, 69.3, 70.1, 66.9)
son = c(68.9, 67.5, 65.4, 68.2, 71.5, 67.6, 65.0, 67.0, 65.3, 65.5, 69.8, 70.9, 68.9, 70.2)
model = lm(son ~ father)
vcov(model)
```

```
##             (Intercept)      father
## (Intercept)  269.521061 -3.96438365
## father        -3.964384  0.05836106
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = son ~ father)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7252 -1.2076 -0.3564  1.2183  2.8928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.6409    16.4171   1.440   0.1754
## father        0.6527     0.2416   2.702   0.0192 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.779 on 12 degrees of freedom
## Multiple R-squared:  0.3782, Adjusted R-squared:  0.3264
## F-statistic:   7.3 on 1 and 12 DF,  p-value: 0.01924
```

Exercise 2 In the paper "Artificial Trees as a Cavity Substrate for Woodpeckers", scientists provided polystyrene cylinders as an alternative roost. The paper related values of x=ambient temperature (C) and y=cavity depth(cm). A scatterplot in the paper showed a strong linear relationship between x and y. The summary for a linear model fit (depth $\sim$ temp) in R is given below.

a. Determine the Pearson's sample correlation (r) from the summaries given.

$$\sqrt{0.7674} = 0.8760137$$

b. Write the linear regression model with least squares estimates for y-intercept: β_0 and slope: β_1 relating ambient temperature (x) and hole depth (y). Interpret the slope and y intercept values in the context of the question.

Linear Model

$$y = 20.12506 - 0.34504x$$

Slope Interpretation

For every unit increase in ambient temperature, hole depth decreases 0.34504

Y-Intercept Interpretation

When ambient temperature is 0, hole depth is 20.12506

c Determine test statistics and p values for the tests in parts c-f. Then draw a conclusion in the context of the question using an $\alpha = 0.05$.

c1. $H_o : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$

t = -5.743 0.000187

c2. $H_o : \beta_1 \geq 0$ vs $H_A : \beta_1 < 0$

t = -5.743 0.000187/2

c3. $H_o : \beta_1 \leq 0$ vs $H_A : \beta_1 > 0$

1 - (0.000187/2) = 0.9999

c4. $H_o : \beta_1 = -0.5$ vs $H_A : \beta_1 \neq -0.5$

t = 2.5792 p-value = 0.02745

d. Compute and interpret a 98% confidence interval for the slope of the regression line β_1.

Point Estimate -0.34504

Critical Value 2.764

Standard Error for Estimator 0.06008

Margin of Error 2.764 * 0.06008 = 0.16606

Final Interval

(-0.5111, -0..17898)

Exercise 3 A player in a gambling game rolls two dice and wins an amount dependent on the number of ones rolled. Let X be the number of ones rolled in two dice, assuming the dice are 6 sided and fair.

a. Show/explain why the following table gives the probability distribution for X:

x	P(X=x)
0	25/36
1	10/36
2	1/36

This shows all possible outcomes and their probabilities for X.

b. Consider conducting a hypothesis test to gather evidence of whether the dice are fair based on 250 games and the proportion of ones that would be rolled in fair die. That is, conduct a hypothesis test of $H_o : \pi_0 = 25/36, \pi_1 = 10/36, \pi_2 = 1/36$ and H_A : at least one proportion of outcomes does not match.

The player plays (rolls two dice) 250 games and records these results:

Number of 1s	Number of Games
0	160
1	70
2	20

bi. Compute the test statistic for the hypothesis given above. Be sure to show your expected counts and your test statistic computation.

Expected Counts:

Number of 1s:	0	1	2
Number of Games:	174	69	7

Observed Test Statistic

$$\frac{(160-173)^2}{173} + \frac{1}{69} + \frac{13^2}{7} = 25.134$$

bii. Compute the degrees of freedom and p-value for the hypothesis test. Interpret the results in context.

Degrees of Freedom 249 *P-value*

```
observed = c(160, 70, 20)
expected_probs = c(25/36, 10/36, 1/36)
expected = 250 * expected_probs
result = chisq.test(observed, p=expected_probs)
print(result)
```

```
##
##  Chi-squared test for given probabilities
##
## data:  observed
## X-squared = 25.616, df = 2, p-value = 2.739e-06
```

Interpretation of results in context

Exercise 4 The table below gives data on the 1083 participants in a vaccine trail. Of the 549 patients who received the vaccine, 11 developed type B hepatitis. Of the 534 patients who did not receive the vaccine, 70 developed the virus. Do these data indicate that there is a different distribution of hepatitis between vaccinated and not vaccinated participants?

	Hepatitis	No Hepatitis	Total
Vaccinated	11	538	549
Not Vaccinated	70	464	534
Total	81	1002	1083

a. Use the χ^2 test for homogeneity to gather statistical evidence for your question of interest. Compute the Test Statistic, Degrees of Freedom, and p value, and interpret your findings in the context of the question. (If using chisq.test(), use correct=FALSE)

Hypotheses

$$H_0 : \pi_H = \pi_{NH}$$

$$H_A : \pi_H \neq \pi_{NH}$$

```r
data = matrix(c(11, 538, 70, 464), nrow=2, byrow=TRUE)
chisq.test(data, correct=FALSE)
```

```
##
##  Pearson's Chi-squared test
##
## data:  data
## X-squared = 48.242, df = 1, p-value = 3.768e-12
```

Observed Test Statistic

48.242

P-value 3.768e-12

Conclusion

Since the p-value is very smull, there is very strong evidence against the null hypothesis.

b. Use the appropriate Z test for testing the equality of the proportion of patients who develop type B hepatitis across the vaccinated and unvaccinated populations with a two-sided alternative. Verify the relation $\chi^2 = Z^2$ by comparing the observed test statistics' numeric values to those found in part (a).

Hypotheses

$$H_0 : \pi_H = \pi_{NH}$$

$$H_A : \pi_H \neq \pi_{NH}$$

```r
# Observed counts for vaccinated and unvaccinated populations
observed_vaccinated <- c(11, 538)  # replace with your data
observed_unvaccinated <- c(70, 464)  # replace with your data

# Total counts for each group
total_vaccinated <- sum(observed_vaccinated)
total_unvaccinated <- sum(observed_unvaccinated)

# Sample proportions
p1_hat <- observed_vaccinated[1] / total_vaccinated
p2_hat <- observed_unvaccinated[1] / total_unvaccinated

# Combined sample proportion
p_hat <- (observed_vaccinated[1] + observed_unvaccinated[1]) / (total_vaccinated + total_unvaccinated)
print(p_hat)
```

```
## [1] 0.07479224
```

```r
# Sample sizes
n1 <- total_vaccinated
n2 <- total_unvaccinated

# Z-test for the equality of proportions
Z_test_statistic <- (p1_hat - p2_hat) / sqrt(p_hat * (1 - p_hat) * (1 / n1 + 1 / n2))

2 * pnorm(Z_test_statistic)
```

```
## [1] 3.767558e-12
```

Observed Test Statistic -6.946

P-value

3.767558e-12 *Conclusion*

There is strong evidence against the null hypothesis. > c. If the alternative is that the rate of hepatitis is lower for the vaccinated group, which of the two testing strategies should be used?

You should use the first testing strategy.

THANK YOU FOR ALL OF YOUR HARD WORK THIS SEMESTER!! I HOPE YOU HAVE A RESTFUL, SAFE, AND FUN WINTER BREAK!! See you on the ski trails! :)