Stat 330 Online: Homework 6 (Module 5) Solutions

Show all of your work, and upload this homework to Canvas.

1. Suppose you take a random sample of 30 individuals from a large population and record a numeric value for each. For this sample, the sample mean is 4.2 and sample variance is 49. You wish to estimate the unknown population mean μ.

   (a) Calculate a 90% confidence interval for μ.

   (b) Calculate a 95% confidence interval for μ.

   (c) Based on (a) and (b), comment on what happens to the width of a confidence interval (increase/decrease) when you increase your confidence level.

   (d) Suppose your sample size is 100 instead of 30. Keep the sample mean and variance at 4.2 and 49 respectively. Calculate a new 90% confidence interval for μ.

   (e) Based on (a) and (d), comment on what happens to the width of a confidence interval (increase/decrease) when you increase your sample size keeping everything else the same.

   Answer:

   (a) For a 90% confidence interval, we use $z_{\alpha/2} = z_{0.05} = 1.65$ in the calculation.

$$\begin{aligned}&\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\\ &= 4.2 \pm 1.65\frac{7}{\sqrt{30}}\\ &= 4.2 \pm 2.1087\\ &= (2.0913,\ 6.3087)\end{aligned}$$

   (b) For a 95% confidence interval, we use $z_{\alpha/2} = z_{0.025} = 1.96$ in the calculation.

$$\begin{aligned}&\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\\ &= 4.2 \pm 1.96\frac{7}{\sqrt{30}}\\ &= 4.2 \pm 2.5049\\ &= (1.6951,\ 6.7049)\end{aligned}$$

   (c) When we increase confidence (and everything else remains the same), the width of the confidence interval increases.

   (d)

$$\begin{aligned}&\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}\\ &= 4.2 \pm 1.65\frac{7}{\sqrt{100}}\\ &= 4.2 \pm 1.155\\ &= (3.045,\ 5.355)\end{aligned}$$

   (e) When sample size increases (and everything else remains the same), the width of the confidence interval decreases.

2. Four randomly selected entry-level computer engineers have salaries (in \$1000s):

```
50,  68,  85,  76
```

(a) Construct a 90% confidence interval for the average salary of an entry-level computer engineer based on a normal distribution.

(b) Construct a 90% confidence interval for the average salary of an entry-level computer engineer based on a t-distribution. The formula for a $(1-\alpha)100\%$ t-based confidence interval is $\bar{X} \pm t_{\alpha/2} s/\sqrt{n}$ where s is the sample standard deviation, and $t_{\alpha/2}$ is the t-value with $n-1$ degrees of freedom. You can use the fact that here $t_{0.05} = 2.35$ and $t_{0.1} = 1.64$.

(c) Compare the previous two confidence intervals you obtain. Which one is more conservative (wider)?

(d) Using a normal-based 90% confidence interval, is there evidence that the average salary of all entry-level computer engineers is different from $80,000? Briefly explain.

Answer: First we have, $\bar{x} = 69.75, s = 14.88$, and $n = 4$.

(a)

$$\begin{aligned} &\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \\ &= 69.75 \pm 1.65\frac{14.88}{\sqrt{4}} \\ &= 69.75 \pm 12.28 \\ &= (57.47,\ 82.03) \end{aligned}$$

(b) The only thing that changes is that we use $t_{.05}$.

$$\begin{aligned} &\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} \\ &= 69.75 \pm 2.35\frac{14.88}{\sqrt{4}} \\ &= (52.27,\ 87.23) \end{aligned}$$

(c) The interval using the t value is more conservative

(d) No, $80 is in the interval meaning it is a plausible value.

3. In order to ensure efficient usage of a server, it is necessary to estimate the mean number of concurrent users. According to records, the average number of concurrent users at 100 randomly selected times is 37.7. The sample standard deviation is $s = 9.2$.

(a) Construct a 90% confidence interval for the expectation of the number of concurrent users.

(b) Conduct a hypothesis test to test whether the true mean number of concurrent users is *greater* than 35. Based on your hypothesis test, do you have evidence that the true mean number of concurrent users is *greater* than 35?

Answer:

(a) $\bar{x} \pm z_{.05}\frac{s}{\sqrt{n}} = 37.7 \pm 1.645\frac{9.2}{10} = 37.7 \pm 1.5$ or $(36.2, 39.2)$

(b) $H_0 : \mu = 35$
$H_A : \mu > 35$

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{37.7 - 35}{9.2/\sqrt{100}} = 2.9348$$

The p-value is $P(Z > 2.93) = 0.0017$. Since the $p - value$ is very small, we have strong evidence against the null hypothesis. We have strong evidence that the true mean number of concurrent users is greater than 35.

4. We have to accept or reject a large shipment of items. For quality control purposes, we collect a random sample of 200 items and find 24 defective items in it.

   (a) Construct a 95% confidence interval for the proportion of defective items in the whole shipment.

   (b) The manufacturer claims that only 10% of the items in the shipment are defective. Does the 95% confidence interval support this claim?

   (c) We would like to collect a large enough sample so that the margin of error in a 95% confidence interval is no greater than 0.02. How many items do we need to randomly inspect?

   Answer: $\hat{p} = \frac{24}{200} = .12, \hat{se}(\hat{p}) = \sqrt{\frac{.12(.88)}{200}} = .023$

   (a)

$$\begin{aligned}&\hat{p} \pm z_{\alpha/2}\hat{se}(\hat{p})\\ &= .12 \pm 1.96(.023)\\ &= .12 \pm .045\\ &= (.075,\ .165)\end{aligned}$$

   (b) Yes, 0.10 falls inside this interval and thus could be a plausible value with 95% confidence.

   (c) $n = \left(\frac{1.96(.5)}{.02}\right)^2 = 2401$ (using the conservative approach)

5. An engineer claims that the probability for a system component to fail under a high-stress environment is at most 0.02, but the quality assurance specialist suspects otherwise. A total of 200 components have been tested under a high-stress environment, and 10 components failed. Perform a hypothesis test to verify the engineer's claim for the quality assurance specialist.

   Answer:

   $H_0 : p = .02$
   $H_A : p > .02$

   $Z = \frac{.05-.02}{\sqrt{\frac{.02(.98)}{200}}} = 3.03$

   The p-value is $\mathbb{P}(Z > 3.03) = .0012$

   We have strong evidence against the engineers claim. We have strong evidence that the probability of a failure is greater than .02.

6. A consumer has used websites A and B for online shopping. To see which website has a faster delivery, the consumer used each website multiple times and collected the following summary statistics for the number of days waited between making the order until having the items delivered. (Assume sample sizes are large enough for normal based inference)

	Website A	Website B
Sample size	20	30
Sample mean	4.5 days	5.2 days
Sample standard deviation	2 days	1 days

   (a) Construct a 99% confidence interval for the difference between the average number of days required for delivery using the two websites.

   (b) Interpret the confidence interval.

   (c) Are you confident that the delivery speeds of the two websites are different?

   (d) Conduct a hypothesis test for whether the delivery speeds of the two websites are different. Show all steps and state your conclusion

Answer:

Let website A be group 1 and website B be group 2

(a)

$$\begin{aligned}&\bar{x}_1 - \bar{x}_2 \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\\ &= -.7 \pm 2.58(.483)\\ &= -.7 \pm 1.25\\ &= (-1.95,\ .55)\end{aligned}$$

(b) We are 99% confident the true difference in mean days required for delivery (website A *minus* website B) is between -1.95 and .55).

(c) No, 0 is inside the interval meaning it is plausible the mean delivery time is the same.

(d) $H_0 : \mu_1 = \mu_2$
$H_A : \mu_1 \neq \mu_2$

$Z = \frac{-.7}{.483} = -1.45$

The p-value in this case is $2\mathbb{P}(Z < -1.45) = .1471$

The p-value is pretty big, indicating no evidence against H_0 :. No evidence of a difference (same conclusion as our interval).

7. An online pollster publishes the results of two recent polls of the approval rates of a candidate in two states. It reports that 35% of the respondents approves the candidate in State A and 45% approves the candidate in State B. A random sample of 600 respondents were polled in State A, and 500 were polled in State B.

Conduct a hypothesis test for whether the approval rates differ in the two states. Show all steps and state your conclusion

Answer: Let group 1 be state A and group 2 be state B
$n_1 = 600, \hat{p}_1 = .35, n_2 = 500, \hat{p}_2 = .45$

$H_0 : p_1 = p_2$
$H_A : p_1 \neq p_2$

$\hat{p}_{pool} = \frac{600(.35)+500(.45)}{600+500} \approx .40$

$Z = \frac{.35-.45}{\sqrt{\frac{.4(.6)}{600}+\frac{.4(.6)}{500}}} = -3.37$

The p-value is $2\mathbb{P}(Z < -3.37) \approx .0007$

There is strong evidence against H_0. Strong evidence of a difference between the two states.

8. Suppose you play fantasy golf and want to come up with a model to predict a golfer's score. Data was gathered on 75 PGATour players. Variables collected were x_1 = Driving Distance, x_2 = Putts per hole, and y = scoring average. We would like to find a good linear regression model to predict a golfer's scoring average based on one of the x variables. We will make two regression models and compare. A training data set of 70 observations was used for model fitting with 5 observations left out as a test set. On the

training data, the following statistics were calculated:

$$\sum_{i=1}^{70} x_1 = 20349.9, \qquad \sum_{i=1}^{70} x_2 = 123.886, \qquad \sum_{i=1}^{70} y = 4969.631$$

$$\sum_{i=1}^{70}(x_2 - \overline{x}_2)^2 = .0368 \qquad \sum_{i=1}^{70}(x_2 - \overline{x}_2)(y - \overline{y}) = .4239$$

(a) The linear regression model using x_1 as a predictor of y is $\hat{y} = 78.48 - .026x_1$. Using the summary statistics, give the regression model for using x_2 as a predictor of y.

Answer: $\hat{\beta}_1 = \frac{\sum_{i=1}^{70}(x_2-\overline{x}_2)(y-\overline{y})}{\sum_{i=1}^{70}(x_2-\overline{x}_2)^2} = \frac{.4239}{.0368} = 11.519$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{4969.631}{70} - (11.519)\frac{123.886}{70} = 50.61$

So, the equation for predicting y from x_2 is $\hat{y} = 50.61 + 11.519x_2$.

(b) The test data set is:

x_1	x_2	y
291.8	1.699	68.94
292.3	1.755	71.72
294.5	1.740	69.94
296.6	1.769	69.80
305.6	1.774	70.65

Give the 5 predicted values using both models (plug in the x_1's into first model, x_2's into second model) and then calculate the Root Mean Square Error (RMSE) for both models. Recall, RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$. Which model appears to do a better job at predicting y?

Answer: Plugging in the 5 x_1's into the equation with x_1 as the predictor yields the following predicted values:

70.8932, 70.8802, 70.8230, 70.7684, 70.5344

The RMSE is $\sqrt{\frac{6.251}{5}} = 1.118$

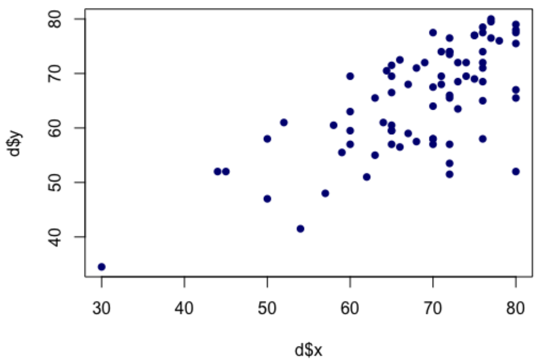
Plugging in the 5 x_2's into the equation with x_2 as the predictor yields the following predicted values:

70.18078, 70.82585, 70.65306, 70.98711, 71.04471

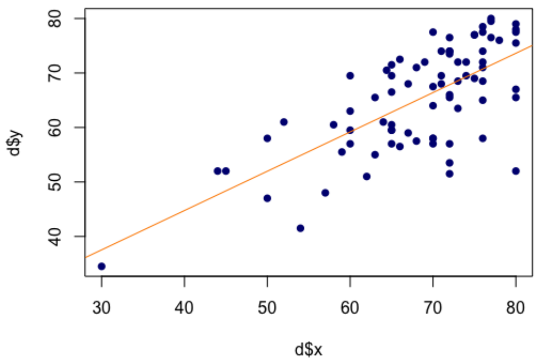
The RMSE is $\sqrt{\frac{4.413}{5}} = .939$

It appears that that x_2 = putts per hole is a better predictor of a player's score. The RMSE is lower (a 16% improvement).

9. *Extra Credit – 4pts*
In Exam 2, we had you predict your exam score. Our goal is to see if our class has "ESP" and can predict their exam scores. I have a data set from exam 2 with two variables. x = a students predicted exam score, and y = their actual score. Data was gathered on 78 pairs from the class. Below is a scatterplot of x vs y



Based on the plot, there appears to be a linear relationship between x and y, so our claim of ESP may have some merit. Thus we propose a regression model as: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
Below is the plot with the fitted least squares regression line added.



(a) Based on the following statistics, give the equation for the least squares regression line.

$$n = 78,\ \bar{x} = 68.35,\ \bar{y} = 65.22,\ s_X^2 = 87.51,\ s_Y^2 = 84.39,$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 4863.78,\ \sum_{i=1}^{n}(y_i - \hat{y})^2 = 3757.657.$$

Answer:
$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x-\overline{x})(y-\overline{y})}{\sum_{i=1}^{n}(x-\overline{x})^2} = \frac{\sum_{i=1}^{n}(x-\overline{x})(y-\overline{y})}{(n-1)s_X^2} = \frac{4863.78}{(77)(87.51} = .722$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 65.22 - (.722)68.35 = 15.87$

So, the least squares regression equation is $\hat{y} = 15.87 + .772x$.

(b) It turns out that the estimator for β_1, $\hat{\beta}_1$, has the following sampling distribution:

$$\hat{\beta}_1 \sim N\left(\beta, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Where an estimate of σ^2 that we can calculate from data is $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2$.

In order to test our claim of “ESP”, we could do the following hypothesis test:

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

In other words, showing that $\beta_1 \neq 0$ would give us evidence that students have some ability to predict their exam scores. Based on all the information provided, we can conduct a normal based Hypothesis test in the same fashion that we have done before. Give the value of the test statistic, calculate the p-value, and draw a conclusion.

Answer:

$H_0 : \beta_1 = 0$
$H_A : \beta_1 \neq 0$

$\hat{\beta}_1 = .722, \hat{se}(\hat{\beta}_1) = \sqrt{\frac{\frac{3757.657}{76}}{(77)87.51}} = .086$

$Z = \frac{.722}{.086} = 8.39$

The p-value is ≈ 0

There is very strong evidence against H_0. Strong evidence that a students predicted score does give some information about what their actual score will be.