Assignment 2

33 In a survey of 400 likely voters, 215 responded that they would vote for the incumbent, and 185 responded that they would vote for the challenger. Let p denote the fraction of all likely voters who preferred the incumbent at the time of the survey, and let $\hat{p}$ be the fraction of survey respondents who preferred the incumbent. $\hat{p} = 215/400$

a. Use the survey results to estimate p.

b. Use the estimator of the variance of $\hat{p}$, $\hat{p}(1-\hat{p})/n$, to calculate the standard error of your estimator.

c. What is the p-value for the test $H_0: p = 0.5$ vs. $H_1: p \neq 0.5$?

d. What is the p-value for the test $H_0: p = 0.5$ vs. $H_1: p > 0.5$?

e. Why do the results from (c) and (d) differ?

f. Did the survey contain statistically significant evidence that the incumbent was ahead of the challenger at the time of the survey? Explain.

a) $\hat{p}$ is unbiased estimator of p

$\hat{p} = \frac{215}{400} = 0.5375$

b) $\text{Var}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n} = \frac{0.5375(0.4625)}{400}$

Std dev:- $\sqrt{\text{Var}(\hat{p})}$

$= \sqrt{\frac{0.5375(0.4625)}{400}}$

$= \frac{\sqrt{0.5375 \times 0.4625}}{20}$

$= 0.0249$

c) $t^{act} = \frac{\hat{p} - \mu_{p,0}}{\sigma_{\hat{p}}} = \frac{0.5375 - 0.5}{0.0249} =$ **1.506**

Since sample size $> 30 (= 400)$ we use formula

$$p\text{-val} = 2\,\Phi(-|t^{act}|) = 2\,\Phi(-1.506)$$

NCD: $t^{act}(-10^{10}, -1.506) + t^{act}(1.506, 10^{10})$

$2 \times 0.066 =$ **0.132**

d) $p\text{-value} = 1 - \Phi(t^{act})$

NCD → lower CV : 1.506

Upper CV : 10^{10}

p-val = **0.066**

e) (c) is a **two-tailed** test while (d) is a **one tailed** test to the **right**.

f) The one tailed test showed us that p val. was larger than p_{CV} ($0.066 > 0.05$). And $t^{act} = 1.506$ was smaller than $t_{CV} = 1.644$, which suggests that we cannot reject the null hypothesis. However the study **doesn't** display significant evidence of portraying that the incumbent was ahead of the challanger at the time of the survey.

Explain.

3.4 Using the data in Exercise 3.3:

a. Construct a 95% confidence interval for p.

b. Construct a 99% confidence interval for p.

c. Why is the interval in (b) wider than the interval in (a)?

a. Without doing any additional calculations, test the hypothesis $H_0: p = 0.50$ vs. $H_1: p \neq 0.50$ at the 5% significance level.

a) 95% CFI = $-1.96, 1.96$

$\therefore \hat{p}$ vals = $\hat{p} \pm CV * SE(\hat{p})$

$= 0.5375 \pm 1.96 * 0.0249 = (0.4887, 0.5863)$

b) $\hat{p}$ vals $= 0.5375 \pm 2.576 \times 0.0249 = (0.4735, 0.6015)$

c) (b) has larger CI, $\therefore$ lowest level of significance hence wider interval.

d) Do not reject null hypothesis as $\hat{p}$ value is greater than significance level. $0.066 > 0.063$.

3.10 Suppose a new standardized test is given to 100 randomly selected third-grade students in New Jersey. The sample average score $\overline{Y}$ on the test is 58 points, and the sample standard deviation, s_Y, is 8 points.

a. The authors plan to administer the test to all third-grade students in New Jersey. Construct a 95% confidence interval for the mean score of all New Jersey third graders.

b. Suppose the same test is given to 200 randomly selected third graders from Iowa, producing a sample average of 62 points and sample standard deviation of 11 points. Construct a 90% confidence interval for the difference in mean scores between Iowa and New Jersey.

c. Can you conclude with a high degree of confidence that the population means for Iowa and New Jersey students are different? (What is the standard error of the difference in the two sample means? What is the *p*-value of the test of no difference in means versus some difference?)

a) $s_y = 8$, $s_{\bar{y}} = \frac{s_y}{\sqrt{n}} = \frac{8}{\sqrt{100}} = \frac{8}{10} = 0.8$

$\hat{\mu} = \bar{y} \pm t_{cv} * (s_{\bar{y}}) = 58 \pm 1.96 \times 0.8 = (56.432, 59.568)$

b) $\mu_1 - \mu_2 = (\bar{y}_1 - \bar{y}_2) \pm \Phi(10\%\,SL) * SE(\bar{y}_1 - \bar{y}_2)$

Critical values $= \pm 1.645$

$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_{\bar{y}_1}^2}{n_1} + \frac{s_{\bar{y}_2}^2}{n_2}} = \sqrt{\frac{64}{100} + \frac{121}{200}} = 1.1158$

$\bar{\mu}_1 - \bar{\mu}_2 = (58 - 62) \pm 1.645\,(1.1158)$

$= (-5.8299, -2.1701)$

c) $H_0: \mu_1 - \mu_2 = 0$; $H_1: \mu_1 - \mu_2 \neq 0$

t-stat value for big sample sizes

$$t^{stat} = \frac{\bar{Y}_1 - \bar{Y}_2}{SE(\bar{Y}_1 - \bar{Y}_2)} = \frac{58 - 62}{1.1158} = -3.5849$$

Pval @ tstat $= \pm 3.5849$ using a 2-tailed test

⇒ Pvalue $(-30^{10}, -3.5849)$ + Pval $(3.5849, 10^{10})$

or $(-10^{6}, -3.5849) * 2 = 2 \times 0.000169$

$= 0.000338$

(3.12) To investigate possible gender discrimination in a firm, a sample of 100 men and 64 women with similar job descriptions are selected at random. A summary of the resulting monthly salaries follows:

	Average Salary ($\bar{Y}$)	Standard Deviation (s_Y)	n
Men	$3100	$200	100
Women	$2900	$320	64

a. What do these data suggest about wage differences in the firm? Do they represent statistically significant evidence that average wages of men and women are different? (To answer this question, first state the null and alternative hypotheses; second, compute the relevant t-statistic; third, compute the p-value associated with the t-statistic; and finally, use the p-value to answer the question.)

b. Do these data suggest that the firm is guilty of gender discrimination in its compensation policies? Explain.

a)

$H_0: \mu_1 - \mu_2 = 0$; $H_1: \mu_1 - \mu_2 \neq 0$

$$t^{stat} \Rightarrow \frac{\bar{Y}_1 - \bar{Y}_2}{SE(\bar{Y}_1 - \bar{Y}_2)} \Rightarrow \frac{3100 - 2900}{44.721} = 4.472$$

$$SE(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{s_{Y_1}^2}{n_1} + \frac{s_{Y_2}^2}{n_2}} = \sqrt{\frac{40000}{100} + \frac{102400}{64}} = 44.722$$

pval. for large sample sizes

$2 \times \text{Pval}(4.4722 \,|\, 10^{10}) = 2 \times 0.00000387$

$= 0.00000774$

Since p value is extremely small, we reject null hypothesis and thus proving that there is a significant difference between avg. wages bet. men & women.

b) We change out test to a right tailed test to determine whether men earn more than women

$H_0: \mu_1 - \mu_2 = 0$; $H_1: \mu_1 - \mu_2 > 0$

p Val for 1sided test is

$1 - \Phi(t^{act}) = 0.00000387$

NCD: lower cv $= -10^{10}$

upper cv $= 4.4722$

$= 0.99999612$

$\therefore 1 - 0.99999612 = 0.00000387$

Since this val. is smaller than confidence interval we reject H_0; thus proving that there is a difference in avg. wages between men & women.

3.16 Grades on a standardized test are known to have a mean of 1000 for students in the United States. The test is administered to 453 randomly selected students in Florida; in this sample, the mean is 1013, and the standard deviation (s) is 108.

a. Construct a 95% confidence interval for the average test score for Florida students.

b. Is there statistically significant evidence that Florida students perform differently than other students in the United States?

c. Another 503 students are selected at random from Florida. They are given a 3-hour preparation course before the test is administered. Their average test score is 1019, with a standard deviation of 95.

i. Construct a 95% confidence interval for the change in average test score associated with the prep course.

ii. Is there statistically significant evidence that the prep course helped?

d. The original 453 students are given the prep course and then are asked to take the test a second time. The average change in their test scores is 9 points, and the standard deviation of the change is 60 points.

i. Construct a 95% confidence interval for the change in average test scores.

ii. Is there statistically significant evidence that students will perform better on their second attempt, after taking the prep course?

iii. Students may have performed better in their second attempt because of the prep course or because they gained test-taking experience in their first attempt. Describe an experiment that would quantify these two effects.

a) $\bar{y} \pm 1.96 \times SE(\bar{y})$

$SE_y = 108$

$SE_{\bar{y}} = \frac{108}{\sqrt{453}} = 5.0743$

$1013 \pm 1.96 * 5.0743$

$(1022.95, 1003.05)$

b) find t^{stat} C.V. $= \frac{1013 - 1000}{5.0743}$

2.5619

$H_0: \mu_1 - \mu_2 = 0 \; ; H_1: \mu_1 - \mu_2 \neq 0$

since $2.5619 > 1.96$ we can reject null hypothesis can be rejected.

c) i) $H_0 = \mu_1 - \mu_2 = 0 \; ; \; H_1 = \mu_1 - \mu_2 \neq 0$

$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{108^2}{453} + \frac{95^2}{503}}$

$\mu_1 - \mu_2 = (\bar{y}_1 - \bar{y}_2) \pm 1.96 \times 6.6099$

$= (1019 - 1013) \pm (1.96 \times 6.6099)$

$= 6 \pm 12.96$

ii) We change two tailed test to right tailed test.

Test

$H_0: \mu_1 - \mu_2 = 0$; $H_1: \mu_1 - \mu_2 > 0$

t val. at 95% sig. level. is 1.96

$$t_{cv} \Rightarrow \frac{1099 - 1013}{SE(\bar{Y}_1 - \bar{Y}_2)} \Rightarrow \frac{6}{6.8099}$$

$$= 0.9077$$

Since 0.9077 is < 1.96 we do not reject H_0. Hence there is no sig. evidence to prove that the prep course helped

d) i) $9 \pm 1.96 \times \left(\frac{60}{\sqrt{453}}\right) = 9 \pm 5.52$

$= (14.52, 3.48)$

ii) $\frac{\bar{Y}_1 - \bar{Y}_2}{\left(\frac{s_Y}{\sqrt{n}}\right)} = \frac{9}{\frac{60}{\sqrt{453}}} = 3.19257$

Since 3.1926 > 1.96, we reject H_0. ∴ There is sig. evidence that the prep course boosts performance.

iii) New sample size n of first time test takers. Make half do the prep course before giving the test again. Other half will regine the test without doing the course and a difference between their first scores will be compared with the test gain in performance.

4.1 Suppose that a researcher, using data on class size (CS) and average test scores from 100 third-grade classes, estimates the OLS regression:

$$\widehat{TestScore} = 520.4 - 5.82 \times CS, R^2 = 0.08, SER = 11.5.$$

a. A classroom has 22 students. What is the regression's prediction for that classroom's average test score?

b. Last year a classroom had 19 students, and this year it has 23 students. What is the regression's prediction for the change in the classroom average test score?

c. The sample average class size across the 100 classrooms is 21.4. What is the sample average of the test scores across the 100 classrooms? (*Hint:* Review the formulas for the OLS estimators.)

d. What is the sample standard deviation of test scores across the 100 classrooms? (*Hint:* Review the formulas for the R^2 and SER.)

a) Test score $= 520.4 - 5.82 \times 22$

$= 392.36$

b) $(520.4 - 5.82 \times 19) - (520.4 - 5.82 \times 23)$

$520.4 - 5.82 \times 19 - 520.4 + 5.82 \times 23$

$= 23.28$

c) $520.4 - 5.82 \times 21.4$

$= 395.85$

d) $S_y = \sqrt{\frac{TSS}{n-1}}$ $\quad TSS = \frac{SSR}{1-R^2} = \frac{(n-2)\,SER^2}{1-R^2}$

$= \sqrt{\frac{98 \times (11.5)^2}{\frac{1 - 0.08^2}{99}}} = 11.4804$

4.5 A professor decides to run an experiment to measure the effect of time pressure on final exam scores. He gives each of the 400 students in his course the same final exam, but some students have 90 minutes to complete the exam, while others have 120 minutes. Each student is randomly assigned one of the examination times, based on the flip of a coin. Let Y_i denote the number of points scored on the exam by the i^{th} student ($0 \leq Y_i \leq 100$), let X_i denote the amount of time that the student has to complete the exam ($X_i = 90$ or 120), and consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Explain what the term u_i represents. Why will different students have different values of u_i?

b. Explain why $E(u_i|X_i) = 0$ for this regression model.

c. Are the other assumptions in Key Concept 4.3 satisfied? Explain.

d. The estimated regression is $\hat{Y}_i = 49 + 0.24\,X_i$.

i. Compute the estimated regression's prediction for the average score of students given 90 minutes to complete the exam. Repeat for 120 minutes and 150 minutes.

ii. Compute the estimated gain in score for a student who is given an additional 10 minutes on the exam.

a) u_i is a factor that affects performance other than time & testing. Could be time spent studying for the exam, could be an error term that resulted in wrong grading.

b) Since u_i is independent of X_i - (Key concept 4.3, 1)

$$P(u_i \cap X_i) = 0 \;\therefore\; E(\mu_i | X_i) = E(\mu_i) = 0$$

c) (2) & (3) both assumptions are satisfied because students are randomly drawn from total enrollment. Outliers are unlikely to be large as $(0 \leq Y_i \leq 100)$ and X_i is only between 90 & 120.

d) i) 90 mins = 70.6, 120 mins = 77.8, 150 mins = 85

ii) $0.24 \times 10 = 2.4$

4.8 Suppose that all of the regression assumptions in Key Concept 4.3 are satisfied except that the first assumption is replaced with $E(u_i|X_i) = 2$. Which parts of Key Concept 4.4 continue to hold? Which change? Why? (Is $\hat{\beta}_1$ normally distributed in large samples with mean and variance given in Key Concept 4.4? What about $\hat{\beta}_0$?)

All key concepts of 4.3 hold.
We can assume an increase in $\hat{\beta}_0$ instead of changing u_i; giving us the regression

$$(\hat{\beta}_0 + 2) + \hat{\beta}_1 X_i + (u_i - 2)$$

4.11 Consider the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$.

a. Suppose you know that $\beta_0 = 0$. Derive a formula for the least squares estimator of β_1.

b. Suppose you know that $\beta_0 = 4$. Derive a formula for the least squares estimator of β_1.

a) $\beta_0 = Y_i - \beta_1 X_i$

$$\beta_0^2 = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2$$

$$0 = \sum_{i=1}^{n} (Y_i^2 - 2 Y_i \beta_1 X_i + \beta_1^2 X_i^2)$$

Differentiating w.r.t. β_1 we get

$$-2(Y_i X_i) + 2\beta_1 X_i^2$$

$$-2 \sum_{i=1}^{n} X_i (Y_i - \beta_1 X_i) = 0$$

$$\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i^2\right) \beta_1 = 0$$

$$\beta_1 = \frac{+\sum_{i=1}^{n} (X_i Y_i)}{\sum_{i=1}^{n} (X_i^2)}$$

b) since $\hat{\beta}_0 = 4$

We can assume regression to follow $\hat{\beta}_0 = 0$ as done in a) but Y_i would be $(Y_i - 4)$ thus giving us

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i (Y_i - 4)}{\sum_{i=1}^{n} (X_i)^2}$$