Name: Y. Lamontagne

Test 3

This test is graded out of 38 marks. No books, notes, unauthorised electronic devices are allowed. You must show all your work, the correct answer is worth 1 mark the remaining marks are given for the work. If you need more space for your answer use the back of the page.

Question 1.[1] Is there strong evidence of global warming? Let's consider a small scale example, comparing how temperatures have changed in the US from 1968 to 2008. The daily high temperature reading on January 1 was collected in 1968 and 2008 for 51 randomly selected locations in the continental US. Then the difference between the two readings (temperature in 2008 - temperature in 1968) was calculated for each of the 51 different locations. The average of these 51 values was 1.1 degrees with a standard deviation of 4.9 degrees. We are interested in determining whether these data provide strong evidence of temperature warming in the continental US.

a. *(4 marks)* Calculate a 90% confidence interval for the average difference between the temperature measurements between 1968 and 2008.

The samples from 1968 and 2008 are dependent samples since each observation from 1968 has a corresponding observation 2008. Observations are paired by cities.

$\bar{d} = 1.1$

$\sigma_{\bar{d}} = SE = \frac{\sigma_d}{\sqrt{n}} \approx \frac{s_d}{\sqrt{n}} = \frac{4.9}{\sqrt{51}}$

$\alpha = 10\%$, $n = 51$

$df = n - 1 = 51 - 1 = 50$

$t_{\alpha/2} = 1.68$

Since σ_d is not known and n is large (≥ 30) we use the t-distribution. The sample is considered independence since 51 is less than 10% of all locations in the US.

$E = t_{\alpha/2} SE = 1.68\left(\frac{4.9}{\sqrt{51}}\right)$

$(\bar{d} - t_{\alpha/2} SE, \bar{d} + t_{\alpha/2} SE) = \left(1.1 - 1.68\left(\frac{4.9}{\sqrt{51}}\right), 1.1 + 1.68\left(\frac{4.9}{\sqrt{51}}\right)\right)$

$= (-0.1, 2.3)$

b. *(1 mark)* Does the confidence interval provide convincing evidence that the temperature was higher in 2008 than in 1968 in the continental US? Explain.

Since 0 is part of the confidence interval the confidence interval does not provide convincing evidence that the temperature was higher in 2008.

[1] OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

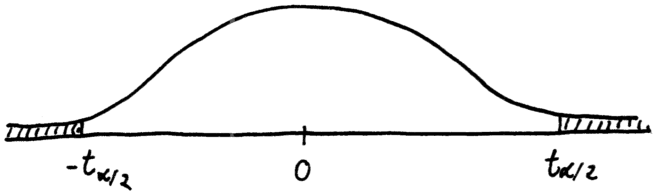
Question 2. [2] The National Assessment of Educational Progress tested a simple random sample of 1,000 thirteen year old students in both 2004 and 2008 (two separate simple random samples). The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively provides data on the average math scores from tests conducted by the National Assessment of Educational Progress in 2004 and 2008. Two separate simple random samples, each of size 1,000, were taken in each of these years. The average and standard deviation in 2004 were 257 and 39, respectively. In 2008, the average and standard deviation were 260 and 38, respectively.

a. *(4 marks)* Do these data provide evidence that the average math score for 13 year old students has changed from 2004 to 2008? Use a 10% significance level. *(Write the hypotheses, check conditions, write a conclusion supported by the test statistic)*

$H_0: \mu_{\bar{x}_{04} - \bar{x}_{08}} = \mu_{x_{04}} - \mu_{x_{08}} = 0$

$H_a: \mu_{\bar{x}_{04} - \bar{x}_{08}} = \mu_{x_{04}} - \mu_{x_{08}} \neq 0$

$\alpha = 10\%$



$\bar{X}_{04} = 257$

$s_{04} = 39$

$\bar{X}_{08} = 260$

$s_{08} = 38$

$n_{04} = 1000$

$n_{08} = 1000$

$$df = \frac{\left(\frac{39^2}{1000} + \frac{38^2}{1000}\right)^2}{\frac{39^4}{1000^2 \cdot 999} + \frac{38^2}{1000^2 \cdot 999}} = 1996.6 = 1996$$

Check conditions:

- Sample is less than 10% of 13 year old studying in the US
- n is large, $n \geq 30$ for both sample.

∴ the t-distribution is "very normal"

$t_{\alpha/2} = 1.65$, $SE \approx \sqrt{\frac{s_{04}^2}{1000} + \frac{s_{08}^2}{1000}}$

b. *(1 mark)* It is possible that your conclusion in part a. is incorrect. What type of error is possible for this conclusion? Explain.

What is the probability of committing that error?

It is possible that we rejected H_0 even though H_0 is true. This type of error is called a Type II error.

$\alpha = P(\text{rejecting } H_0 \mid H_0 \text{ is true})$

$= 10\%$

$$T = \frac{(\bar{X}_{04} - \bar{X}_{08}) - (\mu_{04} - \mu_{08})}{SE}$$

$$= \frac{257 - 260}{\sqrt{\frac{39^2}{1000} + \frac{38^2}{1000}}} = -1.74$$

The test statistic lies in the rejection region. ∴ we reject H_0 in favour of H_a. The data provides evidence that the average math score has changed.

[2] OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

Question 3.[3] *(4 marks)* A 90% confidence interval for a population mean is $(65, 77)$. The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation. Justify.

$(\bar{x} - E,\ \bar{x} + E) = (65, 77)$

So ① $\bar{x} - E = 65$

② $\bar{x} + E = 77$

① + ②

$2\bar{x} = 65 + 77$

$\bar{x} = 71$

$\therefore\ \bar{x} + E = 77$

$71 + E = 77$

$E = 6$

Use t-distribution since we do not know σ

$df = n - 1 = 25 - 1 = 24$

$\alpha = 10\%$

$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$

$t_{\alpha/2} = 1.71$

$6 = 1.71 \frac{s}{\sqrt{25}}$

$s = \frac{6\sqrt{25}}{1.71} = 17.5$

Question 4.[4] In a 2010 Survey USA poll, 70% of the 119 respondents between the ages of 18 and 34 said they would vote in the 2010 general election for Prop 19, which would change California law to legalize marijuana and allow it to be regulated and taxed. At a 95% confidence level, this sample has an 8% margin of error. Based on this information, determine if the following statements are true or false.

1. *(1 mark)* There is a 95% probability that between 62% and 78% of the California voters in this sample support Prop 19.
   True or **False**
2. *(1 mark)* We are 95% confident that between 62% and 78% of all California voters between the ages of 18 and 34 support Prop 19.
   True or **False**
3. *(1 mark)* If we considered many random samples of 119 California voters between the ages of 18 and 34, and we calculated 95% confidence intervals for each, 95% of them will include the true population proportion of 18-34 year old Californians who support Prop 19.
   True or **False**
4. *(1 mark)* In order to decrease the margin of error to 4%, we would need to quadruple *(multiply by 4)* the sample size.
   True or **False**
5. *(1 mark)* Based on this confidence interval, there is sufficient evidence to conclude that a majority of California voters between the ages of 18 and 34 support Prop 19.
   True or **False**

---

[3] OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

[4] modified OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

Question 5.[5] Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.

A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

a. *(1 mark)* Write hypotheses for this research in symbols and in words.

$H_0: p = 0.5$ ($\geq$) Majority of Americans who decide not to go to college do so because they cannot afford it

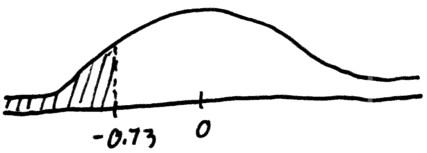
$H_a: p < 0.5$ Minority of Americans who decide not to go to college do so because they cannot afford it.

b. *(1 mark)* Check the conditions required to complete this test.

- sample is assumed to be independent since 331 is less than 10% of the population of American adults.
- $np_0 = 331(0.5) = 165.5 \geq 10$, $n(1-p_0) = 331(0.5) = 165.5 \geq 0$ we can use the normal distribution to approximate the binomial distribution.

c. *(2 mark)* Calculate the test statistic. and p-value

$$Z = \frac{\hat{p} - p_0}{SE} = \frac{0.48 - 0.50}{\sqrt{\frac{0.5(1-0.5)}{331}}} = -0.73$$



p-value $= P(Z < -0.73) = 0.2358$

d. *(1 mark)* What do you conclude at 5% significance? Interpret your conclusion in context.

Since the p-value > 0.05 we fail to reject the null hypothesis.

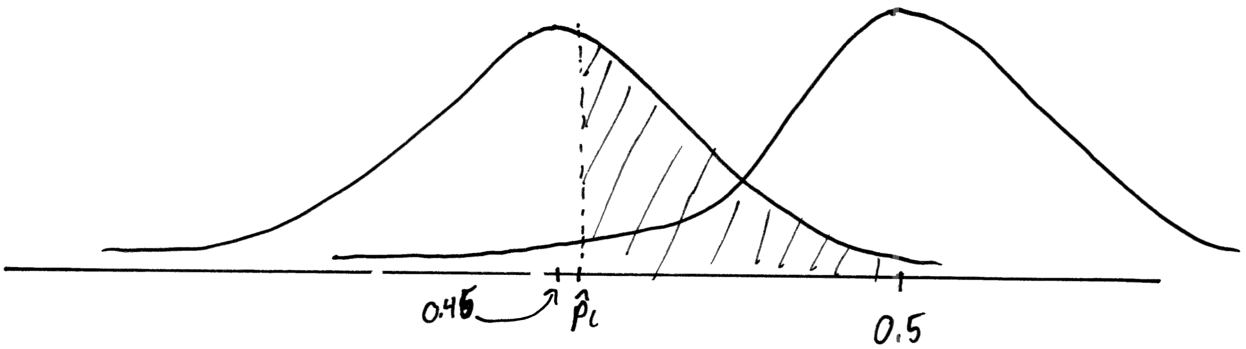
There is very little evidence to conclude that a minority of Americans who decide not to go to college do so because they cannot afford it. Even though the sample proportion is 48%.

[5] modified OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

e. *(1 mark)* What type of error might we have made? Explain in context what the error means.

We might have failed to reject H_0 even though H_0 is false.
This is called a Type II error. $\beta = P(\text{fail to reject } H_0 \mid H_0 \text{ is false})$
In context this implies failing to conclude that a minority of American do not go to school because of financial reasons.
Even though that is the case.

f. *(4 marks)* Find an approximation of the propability of the above error, given the hypothesis test at 5% significance and $p_a = 0.45$.



$$SE = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{331}} \doteq 0.027$$

$$SE_a = \sqrt{\frac{p_a(1-p_a)}{n}} = \sqrt{\frac{0.45(1-0.45)}{331}} \doteq 0.027$$

$$-z_\alpha = \frac{\hat{p}_c - p_0}{SE}$$

$$\hat{p}_c = p_0 - z_\alpha SE$$
$$= 0.5 - 1.645(0.027)$$
$$= 0.46$$

$$\beta = P(\hat{p}_a > \hat{p}_c)$$
$$= P\left(Z > \frac{\hat{p}_c - p_a}{SE_a}\right)$$
$$= P\left(Z > \frac{0.46 - 0.45}{0.027}\right)$$
$$= P(Z > 0.37)$$
$$= 1 - P(Z < 0.37) = 1 - 0.6443 = 0.3557.$$

Question 6.[6] According to a report on sleep deprivation by the Centers for Disease Control and Prevention, the proportion of California residents who reported insufficient rest or sleep during each of the preceding 30 days is 8.0%, while this proportion is 8.8% for Oregon residents. These data are based on simple random samples of 11,545 California and 4,691 Oregon residents.

a. *(5 marks)* Conduct a hypothesis test to determine if these data provide evidence ($\alpha = 0.05$) the rate of sleep deprivation is different for the two states. *(Write the hypotheses, check conditions, write a conclusion supported by the test statistic)*

H_0: $p_c - p_o = 0$ sleep deprivation is the same for both states

H_a: $p_c - p_o \neq 0$ sleep deprivation is not the same for both states

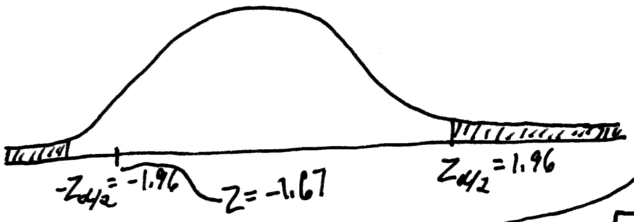
We use pooled proportion since we assume $p_c = p_o$, $\hat{p} = \frac{x_o + x_c}{n_o + n_c} = \frac{\hat{p}_o n_o + \hat{p}_c n_c}{n_o + n_c}$

$= \frac{0.088(4691) + 0.08(11545)}{4691 + 11545}$

$= 0.082$

We assume both sample are independent since both samples are less than 10% of the population of both states. We can use a normal approx. for the distribution since $\hat{p}n_o$, $(1-\hat{p})n_o$, $\hat{p}n_c$, $(1-\hat{p})n_c$ are all greater than 10.

Two tailed test:



$Z = \frac{\hat{p}_c - \hat{p}_o - (p_c - p_o)}{SE}$ where $SE \approx \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_c} + \frac{1}{n_o}\right)}$

$= \sqrt{0.082(1-0.082)\left(\frac{1}{4691} + \frac{1}{11545}\right)} = 0.0048$

$= \frac{(0.08 - 0.088) - (0)}{0.0048} = -1.67$

Since the test statistic is not in the rejection region we fail to reject H_0. At 5% significance there is no difference in sleep deprivation of residents of both states.

b. *(~~2~~ marks)* Calculate ~~a 90% confidence interval~~ the margin of error of a 90% confidence interval for the difference between the proportions of Californians and Oregonians who are sleep deprived ~~and interpret it in context of the data~~.

$E = Z_{\alpha/2} SE$ where $SE \approx \sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_c} + \frac{\hat{p}_o(1-\hat{p}_o)}{n_o}}$

We assume both samples are independent since both samples are less than 10% of the population of both states. We can use the normal dist. for the probabilities since $\hat{p}_c n_c$, $(1-\hat{p}_c)n_c$, $\hat{p}_o n_o$, $(1-\hat{p}_o)$ are all greater than 10.

$E = 1.645\sqrt{\frac{0.08(1-0.08)}{11545} + \frac{0.088(1-0.088)}{4691}} = 0.008$

[6] modified OpenIntro Statistics by D.M. Diez, C.D. Barr and M. Çetinkaya-Rundel, OpenIntro LaTeX, code, and PDFs are released under a Creative Commons BY-SA 3.0 license.

Question 7. ~~The table 2010~~ a survey asked 827 randomly sampled registered voters in California "Do you support? Or do you oppose? Drilling for oil and natural gas off the Coast of California? Or do you not know enough to say?" Below is the distribution of responses, separated based on whether or not the respondent graduated from college. Complete a chi-square test for these data to check whether there is a statistically significant ($\alpha = 0.05$) difference in responses from college graduates and non-graduates. *(Write the hypotheses, check conditions, write a conclusion supported by the test statistic)*

	College Grad		Total
	Yes	No	
Support	154	132	286
Oppose	180	126	306
Do not know	104	131	235
Total	438	389	827

H_0: Difference in responses from graduates and non-graduates is independent

H_a: " " " " " " " " dependent.

Lets compute the expected value assuming H_0 is true.

	College Grad	
	Yes	no
Support	151	135
Oppose	162	144
Do not Know	124	111

all the expected values are larger than 5.

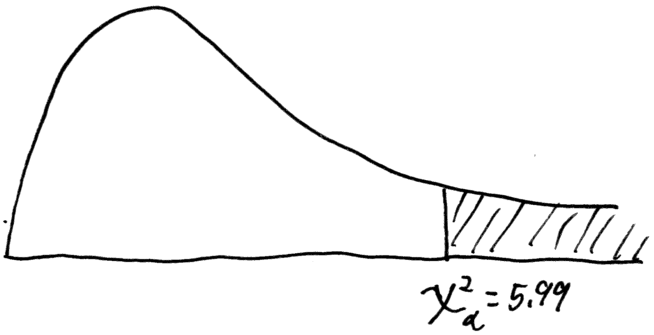
∴ we can use the χ^2 distribution

∴ df = (row − 1)(column − 1) = (3 − 1)(2 − 1) = 2

The test statistic: $\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$

$$= \frac{(154-151)^2}{151} + \frac{(180-162)^2}{162} + \frac{(104-124)^2}{124}$$

$$+ \frac{(132-135)^2}{135} + \frac{(126-144)^2}{144} + \frac{(131-111)^2}{111} = 11.2$$



Test statistic is in the rejection region, hence we reject H_0 in favour of H_a. at 5% significance the difference in response is dependent.

Bonus Question. *(3 marks)* Use the flipping a coin analogy to explain what is meant by a 95% confidence interval of a population parameter.

Once a coin is flipped there is no longer a probability assigned to H or T. It is either H or T. Similarly once a sample is obtained there is no longer a probability assigned to the confidence interval (based on the sample (point estimate)) capturing the population parameter or not.