The exam consists of three exercises with several sub-questions. Answer all parts. Make sure to use Stata only when explicitly asked and otherwise perform all calculations by hand, showing all your work. Full credit will be given only to correct answers that are fully justified. All sub-questions have similar weight, unless otherwise stated.

Exercise 1 (65 points)

The American Community Survey (ACS) is a demographics survey program conducted by the U.S. Census Bureau. It regularly gathers information such as income, employment, ethnicity and other socio-economic characteristics for a random sample of the US population. These data are used by many public-sector, private-sector, and not-for-profit stakeholders to allocate funding, track shifting demographics, plan for emergencies, and learn about local communities. The dataset called exercise1.dta presents data from the 2012 ACS.

(a) (5 points) Using an appropriate command in Stata, find out mean and standard deviation for all variables subsetting the dataset to male only. Next, compare the same statistics with those for female workers. What are the key patterns that you observe? (note: for males the variable `male` takes value of 1, zero otherwise).

Solutions. This is done in Stata with the command `summarize if`. There 1,031 males and 969 females. Not all information is available for all variables, so the sample size varies across variables. Notwithstanding that, female on average earn $14335.99 against $32627.3 for males. The standard deviation of income is also smaller for females (26215.57 vs 58729.79), indicating also a tighter distribution around the mean for this group. On average, females commute about 24 minutes to work, against 27 for males, with a standard deviation of 21 and 22 respectively. The employment variable is categorical and takes three value (employed, not in labor force, unemployed). It's average for females is 1.588903 vs 1.493842 for males (st dev of .6009067 and .6291023), indicating that females are more likely to be not in the labor force or unemployed. The race variable takes relatively similar values of 3.591331 and 3.583899 (st dev .8394749 vs .8456013) for females and males respectively. Given the categories of race, this indicates that there is a larger probability of being white compare to other races.

Grading info. Roughly check that they can use basic summary statistics in the context of the problem, as we do when writing papers. I am usually not strict here.

(b) (10 points) What is the probability that a randomly selected female individual is employed? How does this compare to the probability that a randomly selected male is employed?

Solution. This means calculating $P(employment = 1|male = 0)$ and $P(employment = 1|male = 1)$

$$P(employment = 1|male = 0) = \frac{P(employment = 1, male = 0)}{P(male = 0)} = \frac{373/1,605}{793/1,605} = 47.03657\%$$

$$P(employment = 1|male = 1) = \frac{P(employment = 1, male = 1)}{P(male = 1)} = \frac{470/1,605}{812/1,605} = 57.881773\%$$

Females are 10 percentage point less likely than male to be employed.

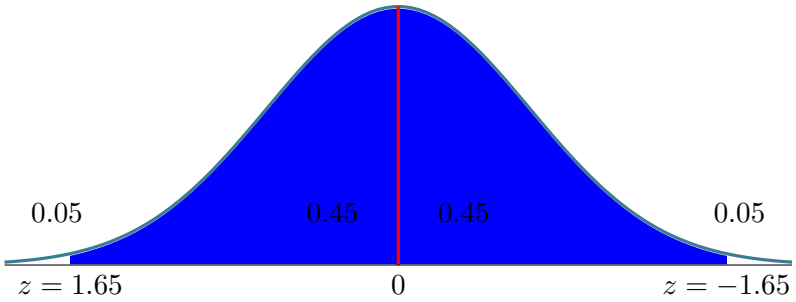
Grading info. Roughly 4 points each for correct calculations, plus 2 points for some conclusion on how the two compare.

(c) (10 points) By hand and without using Stata (except for retrieving any essential information), calculate a 90% confidence interval for the proportion of males in the population. Show all your work.

Solution. Because the variable is dichotomous, and not continuous, it follows a Bernoulli distribution. Therefore, if we want to study the proportion of students who are absent in this particular sample, we are mainly studying $S_n = X_1 + X_2 + \ldots + X_n$, and therefore S_n is the number of successes (here being a male) in n trials (here, in a sample of n people). The Central Limit Theorem guarantees that, as long as the sample size is large, $\bar{p} \sim N(p, p(1-p)/n)$, where I indicate here with $\bar{p}$ the share of males in the sample and with p same share in the population. Although we do not know what p is (in fact, the point here is to find a plausible range for it), we know that, as $\bar{p}$ is normally distributed, we could find:

$$\begin{aligned}
& P\Big(-c < \frac{\bar{p}-p}{\sqrt{p(1-p)/n}} < c\Big) = 90\% \\
\rightarrow \quad & P\Big(-c \times \sqrt{\frac{p(1-p)}{n}} < \bar{p} - p < c \times \sqrt{\frac{p(1-p)}{n}}\Big) = 90\% \\
\rightarrow \quad & P\Big(\bar{p} - c \times \sqrt{\frac{p(1-p)}{n}} < \mu < \bar{p} + c \times \sqrt{\frac{p(1-p)}{n}}\Big) = 90\%
\end{aligned}$$

I need three elements then: $\bar{p}$, n, and c. c is found in the tables. From Stata, $n = 2,000$, $\bar{p} = .5155$. I find c in the tables. The picture below shows the probability I will need to look up. Since the tables report $P(0 < Z < c)$, and I want a probability in each tail of 0.05, I look for c such that the probability between 0 and c is 0.50-0.05=0.45. Such number is between 1.64 and 1.65. To be more conservative, I take 1.65 (but an alternative could be 1.64 or 1.645).



Putting all this information together:

$$\mu \in \Big[.5155 - 1.65\sqrt{\frac{.5155(1-.5155)}{2000}}; .5155 + 1.65\sqrt{\frac{.5155(1-.5155)}{2000}}\Big] =$$
$$= [.49706131, .53393869].$$

So a likely range (90% of the times) for the share of males is between 49.7 and 53.3%.

Grading info. A likely mistake here is not to realize that this is a Bernoulli. Take 4 points off for that, and then a couple if there are calculation mistakes. They should show a bit how they derive the confidence interval.

(d) (10 points) A young statistician notices that from the full Census of the US population the proportion of males is 49%. Knowing that in your sample the proportion is males 51.55%, the young statistician tells you: "The average number of males in the sample is biased!". Do you agree or disagree? Explain.

Solutions. This is a very incorrect statement. Sampling variability implies that a sample is unlikely to give us exactly p. After all, we are only taking a slice of the population, not all of it. Unbiasedness does not say that $\bar{p} = p$! It says that if we were to take many many samples, on average we would get the right p, $E(\bar{p}) = p$. So it is totally normal that in a single sample $\bar{p} \neq p$. Of course in practice we will always have only one sample. This is why inference requires us to discuss not only point estimates, but also their variability. 49% is within the confidence interval found at the previous point so we would actually fail to reject that the two numbers are statistically different.

Grading info. Be strict here. It should come across if they have really understood the difference between an estimate and an estimator.

(e) (10 points) From the full Census of the US population in 2010, you know that average income for females was \$21,000 with standard deviation of \$50,000. In 2011 the government implements a series of labor market policies aiming at increasing female labor force productivity. Using the data of the 2012 ACS test whether the implemented policies did increase earnings of females or not.Use the 5% significance level.

Solution. The text suggests testing the following hypothesis:

$H_0 : \mu = 21,000$

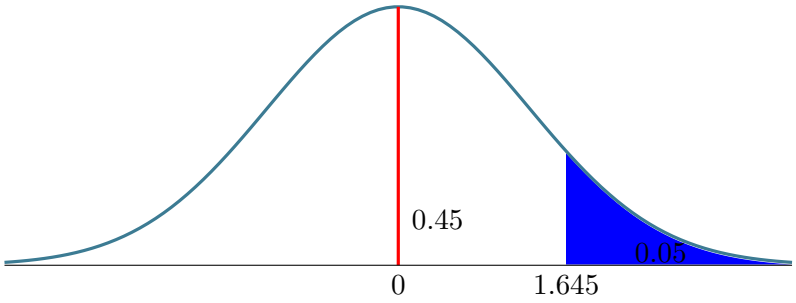
$H_1 : \mu > 21,000$

where μ indicates average income for females and 21,000 is the initial level we want to compare the sample with.

(Step 2) The test statistic takes into account the potential difference between the (treated) sample and the population, as well as sampling variability. The test statistic is:

$$TS = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}}$$

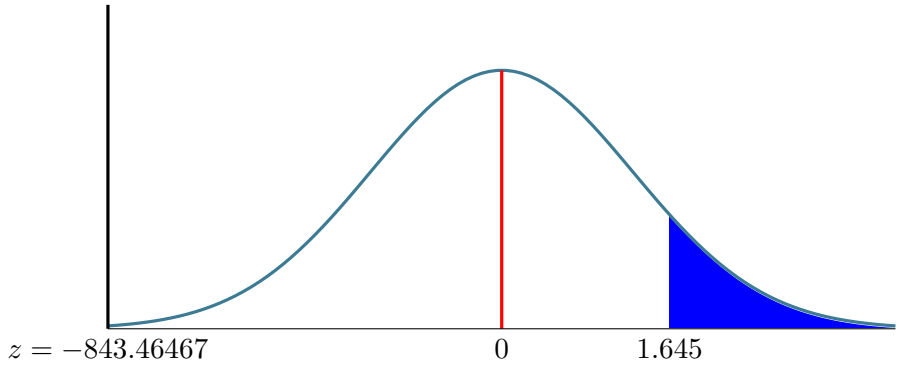
(Step 3) The central limit theorem (CLT) guarantees that for a sample large enough $TS \sim N(0, 1)$. Therefore, intuitively we will reject the null hypothesis if the observed test statistic is greater than a critical value. We define then a rejection region using a $\alpha = 0.05$. Because we want a probability of 0.05 in the right tail (or a probability of 0.5-0.05=0.45 between 0 and the critical value we are looking for), we need to search in the table the value c for which $P(0 < Z < c) = 0.45$. This is shown also in the picture below. Such value is between 1.64 and 1.65. We can use 1.645.



(Step 4.) Using the info in the example, we calculate:

$$TS = \frac{14335.99 - 21000}{\sqrt{50000/801}} = -843.46467$$

In the graph TS is:



(Step 5.) Because $TS = -843.46467 < 1.645$, we fail to reject H_0. This means that we do not have enough evidence to suggest that the program was successful in increasing earnings.

Grading info. They should get this right. Make sure they report all relevant info (hypothesis, test statistics, distribution ecc) and you can be strict on the form and results because this is very standard.

(f) (10 points) With the new labor market policies, the objective was not only to increase income for females but also to reduce its variability. Assume that the income is normally distributed in the population. At a 5% significance level, test whether the program was successful in reducing income dispersion. Note: the critical value for this test is 735.

Solution. We need to test whether the standard deviation after the program is lower, so this is a one-tailed test. The standard deviation before the program was \$50,000. The hypothesis can be written as:

$H_0 : \sigma = 50000$

$H_1 : \sigma < 50000$

By looking at the difference between our sample and 50000 we will know whether the standard deviation is now higher.

(Step 2) The test statistic takes into account this potential difference. The test statistic is:

$$TS = \frac{(n-1)s^2}{\sigma^2}$$

(Step 3) As we are told that returns are normally distributed, $TS \sim \chi^2_{n-1}$ so in this case $TS \sim \chi^2_{800}$

As asked, we define then a rejection region using a $\alpha = 0.05$. This means that we reject as long as TS is far in the left tail and we set the probability of that occurring to 5%. We need the critical value associated to a 5% probability in the left tail.

(Step 4.) From part (a) we know $s = 26215.57$ and $n = 801$. Using the info in the example we calculate:

$$TS = \frac{(801-1)26215.57^2}{50000^2} = 219.92194$$

(Step 5.) Because $TS = 219.92194 < 735$, we reject H_0. This means that we have enough evidence to conclude that the variability in income has decreased after the program.

Grading info. This is a one-tailed (left) test and they might be confused about rejecting or not the null hypothesis. I would take 3-4 points off if the conclusion is incorrect.

(g) (10 points) Consider the test performed at point (f). Do you trust your answer? What are its major shortcomings?

Solution. For this test, we need to assume that income is normally distributed. This is an unreasonable assumption as income cannot take negative values. We should be wary about the interpretation of this test.

Grading info. I am not sure they will understand this, they might talk about sample selection or similar. In that case, you can give 5 points or so if the answers are reasonable.

Exercise 2 (15 points)

In this exercise, use **only** the tables provided at the end of this text. Justify your answers briefly indicating how you are solving the exercise.

(a) Let $X \sim N(0, 9)$. Find $P(X > -1)$.

Solution. Since $X \sim N(0, 9)$, it follows that $X/3 \sim N(0, 1)$. Hence

$$P(X > -1) = P(\frac{X - 0}{3} > \frac{-1 - 0}{3}) = P(Z > -\frac{1}{3})$$

By symmetry and keeping in mind that the tables provide $P(0 < Z < c)$, this is:

$$P(Z > -\frac{1}{3}) = 0.5 + P(0 < X < 0.3) = 0.5 + 0.1293 = 0.6293$$

(b) Individual earnings in the US are approximated well by a χ^2_{15}. Consider a randomly selected US worker. What is the probability that his/her earnings are strictly above the federal minimum wage, today set at \$7.25?

Solution. Let X indicate earnings. We are told $X \sim \chi^2_{15}$ and we need to find $P(X > 7.25)$. From the tables the number is a bit higher than 95%.

(c) For an $X \sim F_{30,2}$, find c such that $P(X > c) = 0.05$.

Solution. We read this directly on the last page of the tables: c=19.426

Grading info. Just grade these are right/wrong.

Exercise 3 (20 points)

(a) From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is going to be performed on 5 patients. Let X be the random variable equal to the number of successes out of the 5 attempts. What is the probability for the surgery to fail exactly 3 times?

Solution. This is

This is:

$$p(X = 2) = \binom{5}{2} p^2(1-p)^3 = .0081$$

Grading info. Grade as right/wrong.

(b) A young statistician thinks that rather than using the sample variance, the following estimator $\tilde{x}$

$$\tilde{x} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}.$$

Describe which statistical properties you would like this new estimator $\tilde{x}$ to exhibit.

Solutions. Unbiasedness, efficiency and consistency. A definition should be provided for all.

(c) Explain how you would check *empirically* that the proposed estimator $\tilde{x}$ satisfies the statistical properties you have discussed at point (b).

Solutions. We have performed in class Monte Carlo simulations to study unbiasedness of the sample mean. Something similar could be done here. The process of performing a MC simulation should be briefly discussed.

Grading info. Students are more familiar with showing unbiasedness. Give 8 points or so if they briefly explain how to perform a monte carlo. The two extra points are for those who also think about how to "prove" efficiency (I don't think anybody should know how to do consistency).

(d) Why does a zero independence imply a zero covariance, but a zero covariance does not imply independence? Briefly explain.

Solution. Covariance measures linearity. Two variables might be strongly related, but in a non-linear way. The covariance might therefore be zero, even though a relationship exist. On the other hand, if two variables are independent they will also not be linearly related, hence their covariance will be zero.

Grading info. Grade roughly right/wrong because we have discussed this in class.

```
1
2  ***********************************************
3
4  ** Final Exam Spring 2021
5
6  ***********************************************
7  clear all
8
9  cd "/Users/co/Documents/Teaching/Courses Taught/Statistics/Exam"
10
11 use exercise1_s21.dta
12
13 * (a)
14 summarize if male == 1
15 summarize if male == 0
16
17 * (b)
18 tabulate employment male
19 display   373 / 793
20 display   470/812
21
22
23 * (c)
24
25 summarize male
26 display r(mean)-1.65*sqrt((r(mean)*(1-r(mean)))/2000)
27 display r(mean)+1.65*sqrt((r(mean)*(1-r(mean)))/2000)
28
29
30 * (f)
31 summarize income if male == 0
32 display ((801-1)*r(sd)^2)/50000^2
33
```