ECOM20001 Econometrics 1

Tutorial 4 Semester 1, 2022

Chin Quek

Department of Economics

Introduction

- Hypothesis Testing of Sample Means, p-values
- Confidence Intervals
- Testing Differences of Means Between Independent Samples
- Introduction to Simple Linear Regression

File downloaded from Canvas

- `tute4.R`
- `tute4_cps.csv`
- `consumption.csv`

Also install the `stargazer` package which we will use to generate a table of summary statistic.

```
install.packages("stargazer", dependencies=TRUE)
```

- Note: need to call (`load`) the package each time you wish to use it when running a R script.

```
## Set the working directory for the tutorial file
setwd("your working directory")
```

In-tutorial Question 1

Suppose you have a random sample of data with a mean μ, and you conduct the following hypothesis test:

$$H_0 : \mu = 10 \qquad \text{vs} \qquad H_1 : \mu \neq 10$$

Having performed the test, you obtained a p-value of 0.07.

a. Does the 90% CI for the population mean contain $\mu = 10$? Explain.

b. With only the information provided in the question, can you determine if $\mu = 8$ is contained in the 90% CI? Explain.

Dataset

The dataset (tute4_cps.csv) contains Current Population Survey information for 15,052 individuals in the U.S and has the following 5 variables:

- year: year individual was randomly surveyed; either 1992 or 2012
- ahe: individual's average hourly earnings (in real terms, 2012=100)
- bachelor: equals 1 if individual has a bachelor degree, 0 otherwise
- female: equals 1 if individual is female, 0 otherwise
- age: age of the individual at time of survey

```r
# Loading the dataset
data=read.csv("tute4_cps.csv")

head(data)
```

```
##   year       ahe bachelor female age
## 1 1992 18.310410        1      0  29
## 2 1992 16.364930        1      0  33
## 3 1992  9.441307        0      0  30
## 4 1992  2.557021        0      0  32
## 5 1992 24.477460        1      0  31
## 6 1992 14.172190        1      1  26
```

Review of indexing data

```
data$female[1]
```

```
## [1] 0
```

```
# first 10 observations
data[1:10,]
```

```
##    year       ahe bachelor female age
## 1  1992 18.310410        1      0  29
## 2  1992 16.364930        1      0  33
## 3  1992  9.441307        0      0  30
## 4  1992  2.557021        0      0  32
## 5  1992 24.477460        1      0  31
## 6  1992 14.172190        1      1  26
## 7  1992 12.745760        0      1  31
## 8  1992 29.110700        0      0  33
## 9  1992 18.095840        0      0  29
## 10 1992 19.767740        0      0  30
```

How would you

- create a subset of data related to females?
- obtain the mean of earnings for females using the subset?

How would you

- create a subset of data related to females?
- obtain the mean of earnings for females using the subset?

```
# Create a subset of females
fem <- data[data$female==1,]
# first 5 observations of the bach dataset
fem[1:5,]
```

```
##    year      ahe bachelor female age
## 6  1992 14.17219        1      1  26
## 7  1992 12.74576        0      1  31
## 12 1992 12.98180        1      1  30
## 13 1992 22.47930        1      1  34
## 14 1992 11.80163        0      1  26
```

```
# Mean of earnings for females using the subset
mean(fem$ahe)
```

```
## [1] 17.80898
```

- Instead of first creating a subset, using just one line of code, how would you obtain the mean of earnings for females?
- What is the earnings of males without bachelor degrees?
- What is the earnings of males without bachelor degrees in 2012?

- Instead of first creating a subset, using just one line of code, how would you obtain the mean of earnings for females?
- What is the earnings of males without bachelor degrees?
- What is the earnings of males without bachelor degrees in 2012?

```
# Mean of earnings for females
mean(data$ahe[data$female==1])
```

```
## [1] 17.80898
```

```
# What is the earnings of males without bachelor degrees?
mean(data$ahe[data$female==0])
```

```
## [1] 20.57906
```

```
mean(data$ahe[data$female==0 & data$bachelor==0])
```

```
## [1] 17.36584
```

```
# What about the earnings of males without bachelor degrees in 2012?
mean(data$ahe[data$female==0 & data$bachelor==0 & data$year==2012])
```

```
## [1] 17.04357
```

In-tutorial Question 2

Using R, what is the sample mean and standard deviation of `ahe` for males and females?

R code chunk below provided in the html document

```
## Mean and standard deviation of earnings for females
mean(data$ahe[data$female==1])
```

```
## [1] 17.80898
```

```
sd(data$ahe[data$female==1])
```

```
## [1] 8.873493
```

```
## Mean and standard deviation of earnings for males
mean(data$ahe[data$female==0])
```

```
## [1] 20.57906
```

```
sd(data$ahe[data$female==0])
```

```
## [1] 10.5533
```

Discuss these numbers and the density plots produced for `ahe` for males and females (reproduced below), which reveals what is known as the gender wage gap.

- Provide **economic explanations** for your results. (Recall from tutorial 2 an economic explanation focuses on the costs and benefits of a behaviour for explaining empirical patterns).
- In this example, what are the different economic costs and benefits among males and females in generating household earnings?

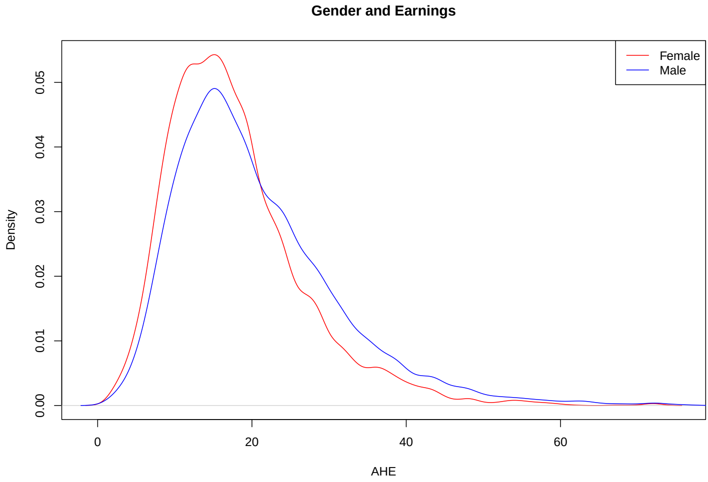
Discuss these numbers and the density plots produced for `ahe` for males and females (reproduced below), which reveals what is known as the gender wage gap.

- Provide **economic explanations** for your results. (Recall from tutorial 2 an economic explanation focuses on the costs and benefits of a behaviour for explaining empirical patterns).
- In this example, what are the different economic costs and benefits among males and females in generating household earnings?

	Mean of AHE	Standard deviation of AHE
Males	\$20.58	\$10.55
Females	\$17.81	\$8.87

- Difference in AHE = \$20.58 - \$17.81 = \$2.77 average earnings per hour

Density plots of `ahe` for females and males



Potential economic explanations for this gender earnings gap???

In-tutorial Question 3

Using R, what is the sample mean and standard deviation of ahe for individuals with and without bachelor degrees?

```
## Mean and standard deviation of earnings for individuals
## with bachelor degree
mean(data$ahe[data$bachelor==1])
```

```
## [1] 23.34672
```

```
sd(data$ahe[data$bachelor==1])
```

```
## [1] 10.71684
```

```
## Mean and standard deviation of earnings for individuals
## without bachelor degree
mean(data$ahe[data$bachelor==0])
```

```
## [1] 16.04614
```

```
sd(data$ahe[data$bachelor==0])
```

```
## [1] 7.855756
```

	Mean of AHE	Standard deviation of AHE
Bachelor degree	\$23.35	\$10.72
No Bachelor degree	\$16.05	\$7.86

- Difference in AHE = \$23.35 - \$16.05 = \$7.30 average earnings per hour

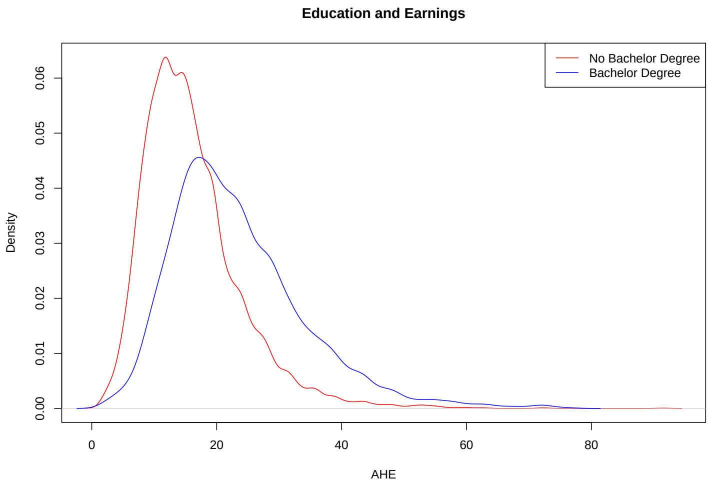
	Mean of AHE	Standard deviation of AHE
Bachelor degree	\$23.35	\$10.72
No Bachelor degree	\$16.05	\$7.86

- Difference in AHE = \$23.35 - \$16.05 = \$7.30 average earnings per hour

Discuss these numbers and the density plots produced for `ahe` for individuals with and without bachelor's degrees. Provide economic explanation(s) for your results.

```
plot(density(data$ahe[data$bachelor == 0]), col = "red",
     lty = 1, xlab = "AHE",
     main = "Education and Earnings")
lines(density(data$ahe[data$bachelor == 1]), col = "blue", lty = 1)
legend("topright", legend = c("No Bachelor Degree", "Bachelor Degree"),
       col = c("red", "blue"), lty = c(1,1))
```

Density plots of `ahe` for bachelor and non-bachelor



Potential economic explanations for this education earnings gap???

In-tutorial Question 4

There does seem to be a difference in the average `ahe` between males and females who have degrees, and without degrees.

Run the following codes in R.

In-tutorial Question 4

There does seem to be a difference in the average ahe between males and females who have degrees, and without degrees.

Run the following codes in R.

WITHOUT bachelor degrees in 2012

```
diff1=mean(data$ahe[data$female==1 & data$year==2012 &
                      data$bachelor==0])-
  mean(data$ahe[data$female==0 & data$year==2012 &
                  data$bachelor==0])
print(diff1)
```

```
## [1] -3.924525
```

Hypothesis test of Difference in Means

$$H_0 : \mu_{\text{AHE\_Female\_2012\_NoBach}} = \mu_{\text{AHE\_Male\_2012\_NoBach}} \qquad \text{vs}$$

$$H_1 : \mu_{\text{AHE\_Female\_2012\_NoBach}} \neq \mu_{\text{AHE\_Male\_2012\_NoBach}}$$

```
t.test(data$ahe[data$female==1 & data$year==2012 & data$bachelor==0],
       data$ahe[data$female==0 & data$year==2012 & data$bachelor==0])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$ahe[data$female == 1 & data$year == 2012 & data$bachelor ==
## t = -15.361, df = 3269.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.425451 -3.423600
## sample estimates:
## mean of x mean of y
##  13.11905  17.04357
```

WITH bachelor degrees in 2012

```
diff2=mean(data$ahe[data$female==1 & data$year==2012 &
                      data$bachelor==1])-
  mean(data$ahe[data$female==0 & data$year==2012 &
                  data$bachelor==1])
print(diff2)
```

```
## [1] -3.796481
```

Hypothesis test of Difference in Means

$$H_0 : \mu_{\text{AHE\_Female\_2012\_Bach}} = \mu_{\text{AHE\_Male\_2012\_Bach}} \qquad \text{vs}$$

$$H_1 : \mu_{\text{AHE\_Female\_2012\_Bach}} \neq \mu_{\text{AHE\_Male\_2012\_Bach}}$$

```
t.test(data$ahe[data$female==1 & data$year==2012 & data$bachelor==1],
       data$ahe[data$female==0 & data$year==2012 & data$bachelor==1])
```

```
##
##  Welch Two Sample t-test
##
## data:  data$ahe[data$female == 1 & data$year == 2012 & data$bachelor ==
## t = -10.778, df = 3851.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.487066 -3.105896
## sample estimates:
## mean of x mean of y
##  21.50238  25.29886
```

Effect of education on earnings

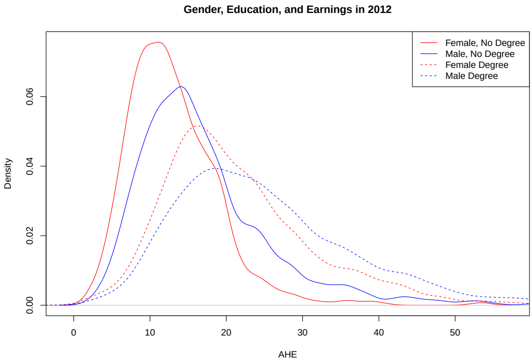
Estimated gender earnings gap

- Without bachelor degrees: \$3.92
- With bachelor degrees: \$3.80

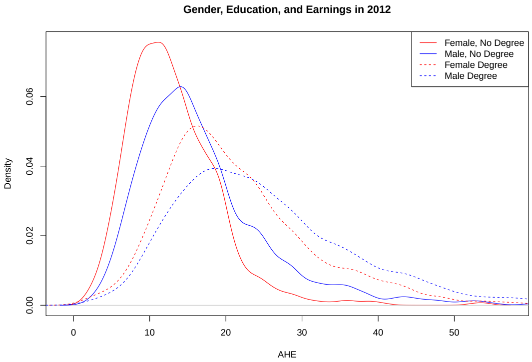
What can we tell about the effect of education from the differences in means?

Why do you think the gender earnings gap differs among males and females with and without bachelor's degrees?

Density plots of `ahe` for gender and education



Density plots of `ahe` for gender and education



Smaller gap in the means among males and females with bachelor degrees

Possible economic explanation:

- Among women with bachelor's degree, there is a smaller propensity to have as many children, and hence less disruption in their careers due to children, which would mean a smaller gender earnings gap among people with bachelors degrees.

R code - Density plots of ahe for gender and education

```r
plot(density(data$ahe[data$female==1 & data$year==2012 & data$bachelor==0])
     col="red",lty=1,main="Gender, Education, and Earnings in 2012",
     xlab="AHE")
lines(density(data$ahe[data$female==0 & data$year==2012 & data$bachelor==0]
      col="blue",lty=1)
lines(density(data$ahe[data$female==1 & data$year==2012 & data$bachelor==1]
      col="red",lty=2)
lines(density(data$ahe[data$female==0 & data$year==2012 & data$bachelor==1]
      col="blue",lty=2)
legend("topright", legend=c("Female, No Degree", "Male, No Degree",
                            "Female Degree", "Male Degree"),
       col=c("red","blue","red","blue"), lty=c(1,1,2,2))
```

In-tutorial Question 5

The dataset `consumption.csv` contains a **population** of 60 families.

The variables are:

- `consumption`: family consumption in $/week
- `income`: family disposable income in $/week

Load `stargazer` package and read in dataset

```
library(stargazer)

## Load dataset on income and consumption
data1=read.csv("consumption.csv")
```

Question 5a, 5b: Compute unconditional and conditional means

a. What is the population mean of consumption, i.e. $E(consumption)$
b. What is the conditional mean $E(consumption|income \leq 100)$?

```
ymean = mean(data1$Consumption)
ycondmean=mean(data1[data1$Income <= 100, "Consumption"], na.rm = TRUE)

print(ymean)
```

```
## [1] 121.2
```

```
print(ycondmean)
```

```
## [1] 71.54545
```

Question 5c: Run OLS on the population

- c. Run the following in the population and confirm the Population Regression Line (PRL) is Consumption $= 17 + 0.6\,\text{Income}$

```
reg1 = lm(Consumption ~ Income, data = data1)
summary(reg1)
```

```
##
## Call:
## lm(formula = Consumption ~ Income, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.00  -8.25   2.50   8.00  28.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.00000    4.66197   3.647  0.00057 ***
## Income       0.60000    0.02549  23.537  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.32 on 58 degrees of freedom
## Multiple R-squared:  0.9052, Adjusted R-squared:  0.9036
## F-statistic:   554 on 1 and 58 DF,  p-value: < 2.2e-16
```

```
stargazer(reg1, type = "text",
          dep.var.labels = c("Consumption"))
```

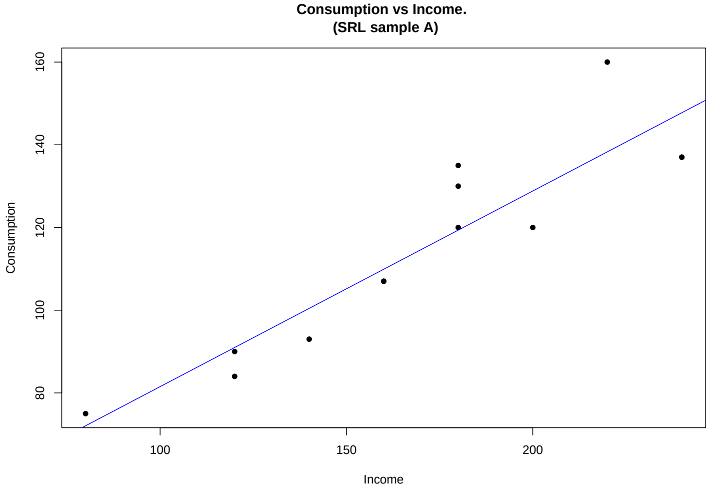
Question 5d: Construct sample A from the population, run OLS, do scatter plot

- (d.) Using the R code provided, construct a random sample of 13 families for the population. Call it Sample A. Run the following regression and also create a scatterplot.

```r
set.seed(2904)

# Construct a random sample of 13 families from the population
data1a <- data1[sample(nrow(data1), 13, replace=TRUE),]

# Estimate a linear regression model for sample A
reg1a = lm(Consumption ~ Income, data = data1a)

# Produce a summary of results
summary(reg1a)
```

```
##
## Call:
## lm(formula = Consumption ~ Income, data = data1a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.750  -7.444  -2.906   2.939  21.711
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.21667   11.41745   2.997   0.0121 *
## Income       0.47306    0.06454   7.329 1.49e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.74 on 11 degrees of freedom
## Multiple R-squared:   0.83,  Adjusted R-squared:  0.8146
## F-statistic: 53.72 on 1 and 11 DF,  p-value: 1.487e-05
```

```r
plot(data1a$Income,data1a$Consumption,
     main="Consumption vs Income. \n (SRL sample A)",
     xlab="Income", ylab="Consumption", col="black", pch=16)
abline(reg1a, col="blue")
```



Question 5e: Construct predicted values, residuals, etc.

- e. Using the R code below, construct the following variables:

  - pred: predicted consumption
  - resid: residual
  - resid2: squared residual

Then compute the sum of the residual and the sum of the squared residuals. What do you find?

```
data1a$pred = predict(reg1a, data=data1a)
data1a$resid = data1a$Consumption-data1a$pred
data1a$resid2 = data1a$resid^2
sumresid = sum(data1a$resid)
sumresid2 = sum(data1a$resid2)

print(sumresid)
```

```
## [1] -5.684342e-14
```

```
print(sumresid2)
```

```
## [1] 1268.972
```

Question 5f: Construct Sample B and estimate SRL

f. using the R code provided, construct another random sample of 13 families for the population.
Call it Sample B. Then construct a scatterplot using the *population* with the PRL, SRL of sample A, and SRL of sample B included.

Briefly interpret the results.

```
# Construct another random sample of 13 families from the population
data1b <- data1[sample(nrow(data1),13,replace=TRUE),]

# Estimate a linear regression model for Sample B
reg1b = lm(Consumption ~ Income, data = data1b)

# Produce a summary of results
summary(reg1b)
```

```
##
## Call:
## lm(formula = Consumption ~ Income, data = data1b)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3257  -4.5665  -0.4862   7.5539  12.5539
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 28.80734   10.22168   2.818   0.0167 *
## Income       0.54799    0.05736   9.553 1.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.396 on 11 degrees of freedom
## Multiple R-squared:  0.8924, Adjusted R-squared:  0.8827
## F-statistic: 91.26 on 1 and 11 DF,  p-value: 1.165e-06
```

Scatter plot of population with PRL and SRLs

```
plot(data1$Income,data1$Consumption, main="Consumption vs Income.
     \n(black: PRL; blue: SRL sample A; red: SRL sample B)",
     xlab="Income", ylab="Consumption", col="black", pch=16)
abline(reg1, col="black")
abline(reg1a, col="blue")
abline(reg1b, col="red")
legend("bottomright", legend=c("PRL", "SRL sample A", "SRL sample B"),
       col=c("black", "blue", "red"), lty=1, cex=0.8,
       box.lty=0)
```