GOOD
OOD

Recitation 1

Exercise 1

Given the following table

Student	Gender	Grade	City	Faculty	Family Income
1	F	28	Milan	Economics	18,000
2	F	25	Florence	Statistics	53,000
3	M	23	Naples	Political Sciences	100,000
4	F	30	Rome	Economics	13,000
5	M	25	Rome	Statistics	89,000
6	M	28	Milan	Statistics	89,000

1. Which are the subjects?
2. Which are the variables?
3. Which type of variables are they?

Exercise 2

The following table reports the number of sick leaves for a sample of employees

Value	Frequency		
0	12	32	32%
1	8	.22	22%
2	5	.14	14%
3	8	.22	22%
4	1	03	03%
5+	3	.08	08%
	37		

1. Add proportions and percentages to the above table
2. Draw a pie chart and a bar chart to represent these data

Exercise 3

The following data represent the sizes of 20 families that live in a small town in Guatemala:

$$5, 7, 10, 8, 6, 6, 10, 7, 5, 10, 8, 11, 8, 6, 9, 5, 10, 7, 11, 5 \mid 104$$

1. Build the frequency table with counts, proportions and percentages

Exercise 4

15 people attended the launch of a new book and the time required for reading the book (in months) for each of them is as follows

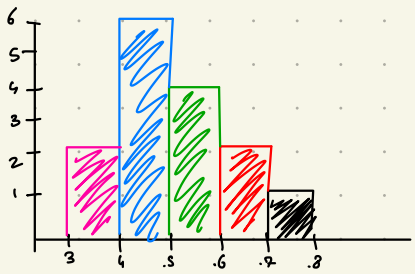
0.[redacted], 0.61, [redacted], 0.64, 0.[redacted], 0.80

1. Build the frequency table with counts, proportions and percentages by considering the following classes:

$$[0.30, 0.40), [0.40, 0.50), [0.50, 0.60), [0.60, 0.70), [0.70, 80]$$

2. Draw the histogram representing the distribution of the above variable
3. What can be said about the variable by looking at this graph?

	FREQ	PROP	PERC
.30 ≤ x < .40	2	.13	13
.40 ≤ x < .50	6	.4	40
.50 ≤ x < .60	4	.27	27
.60 ≤ x < .70	2	.13	13
.70 ≤ x < .80	1	.07	07
	[redacted]		
	15		



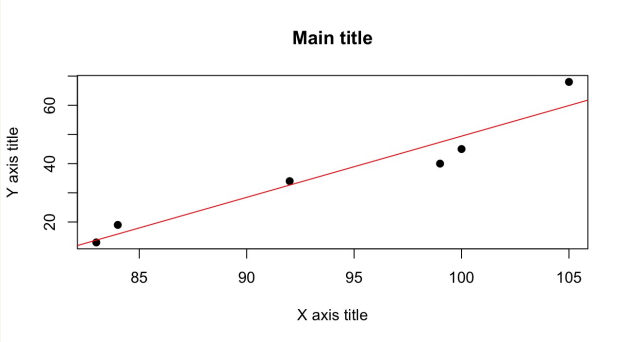
Exercise 5

The following table reports the IQ scores on an intelligence test and the annual salaries (in thousands of euro) of a sample of employees

Subject	IQ	Salary
1	105	68
2	84	19
3	83	13
4	92	34
5	99	40
6	100	45

1. Draw a scatterplot representing the Salary as a function of the IQ scores

```
x <- c(105,84,83,92,99,100)
y <- c(68,19,13,34,40,45)

plot(x, y, main = "Main title",
     xlab = "X axis title", ylab = "Y axis title",
     pch = 19, frame = TRUE)

abline(lm(y ~ x, data = mtcars), col = "red")
```



Exercise 6

The number of rainy days that occurred each month are:

$$10, 14, 14, 8, 5, 6, 1, 3, 7, 10, 14, 9$$

1. Compute the mean, the median and the mode
2. Comparing the mean and the median, what can be said about the variable distribution?
3. If the number of rainy days doubles each month, what will be the mean? And the median? And the mode?

mean = 8.42 mode = 14 ①

median = 8.5

②

③ mean_2 = 2 · mean

mode_2 = 2 · mode

Exercise 7

Consider a sample of 10 MBA graduates, whose first salary (in thousands of dollars per year) after graduation are as follows:

$$200, 70, 88, 130, 175, 89, 95, 120, 400, 55$$

1. Compute the mean and the median salary
2. Comparing the mean and the median, what can be said about the salary distribution?
3. What happens to the mean if the second year after graduation all salaries increase by 10 %?
4. What happens to the mean if all MBA graduates receive an extra amount of 10 thousand dollars?
5. What happens to the mean if all MBA graduates receive an extra amount of 10 thousand dollars in addition to an increase of 10%?
6. What happens to the mean and to the median if we accidentally report 4000 in place of 400?

① mean = 142.2

median = 107.5

②

③ mean_2 = 1.1 · mean

④ $\frac{1}{n}\sum_{i=1}^{n} x_i + 10$

⑤ $\frac{1.1}{n}\sum_{i=1}^{n} x_i + 10$

⑥ To the median nothing... instead the mean become 502.2

Recitation 2

Exercise 2

7 people were selected at random in a restaurant and their age was recoded. These are the data

Subject	Age
1	35
2	32
3	16
4	66
5	22
6	46
7	32

1. Compute the sample standard deviation and the range for the variable Age
2. If we observe a further subject 82 years old, what happens to the mean? And to the standard deviation?

① $\bar{x} = 33.57$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 16.47$$

② $\bar{x} = 41.375$

$\sigma = 24.20$

Exercise 3

The following table gives the number of dentists per 100,000 inhabitants for 12 U.S. states in 2000

State	# of dentists
Ohio	57
Indiana	35
Illinois	61
Michigan	64
Wisconsin	90
Minnesota	78
Iowa	60
Missouri	53
North Dakota	55
South Dakota	57
Nebraska	71
Kansas	30

1. Compute the mean and the sample standard deviation.
2. Find the the 25th percentile (1st quartile), the 50th percentile (2nd quartile) and the 75th percentile (3rd quartile).
3. If in 2001 the number of dentists per 100,000 inhabitants reduces by 5% in all states, which are the new values for the mean and the sample standard deviation? And for the quartiles?
4. If the following transformation holds $y = 2x + 3$, which are the new values for the mean and the sample standard deviation? And for the quartiles?
5. Compute the range and the interquartile range.
6. Draw the boxplot. What can we conclude by looking at this graph? Should any of the observed values be classified as potential outliers?

① $\bar{x} = 59.25$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 15.71$$

② $(sx[3] + sx[4])/2 = 54$ ※ 1st quartile

$(sx[6] + sx[7])/2 = 58.5$ ※ 2nd QUARTILE

$(sx[9] + sx[10])/2 = 67.5$ ※ 3rd QUARTILE

③ $\bar{x} = 56.25$

$\sigma = 14.92$

$q1 = 51.3$

$q2 = 55.58$

$q3 = 64.125$

④ $\bar{x} = 121.5$

$\sigma = 63.84$

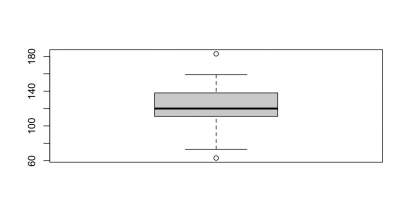
$q1 = 111$

$q2 = 120$

$q3 = 138$

⑤ range $= sx[12] - sx[1] = 120$

⑥




```
# 1-2

x <-
c(57,35,61,64,90,78,60,53,55,57,71,30)
sx = sort(x)
m = mean(x) #mean
len = length(x) # x's lenght
dx = x-m
dx2=dx**2
v = (1/len-1)*sum(dx2) #variance
ssd = sqrt(sum(dx2)/len-1) #sample
standard deviation

q1 = (sx[3]+sx[4])/2 #1st quartile
q2 = (sx[6]+sx[7])/2 #2nd quartile
q3 = (sx[9]+sx[10])/2 #3rd quartile
iqr = q3-q1 #interquartile range
Po1= q1 -1.5*iqr #Potential Outliers
Po2 = q3 + 1.5*iqr #Potential Outliers
boxplot(x)

# 3

x2=x*0.95
m2 = m*0.95
ssd2=ssd * 0.95
v2 = v *(0.95**2)
q1_02= q1*0.95
q2_02= q2*0.95
q3_02 = q3*0.95
iqr2= iqr * 0.95
PO1_02=Po1 *0.95
PO2_02=Po2*0.95
boxplot(x2)

# 4-5

x3 = x*2 +3
m3 = mean(x3)
sx3=sort(x3)
dx_03 = x3- m
dx2_03 = dx_03**2
ssd3 = sqrt(sum(dx2_03)/len-1) #sample
standard deviation
q1_03 = (sx3[3]+sx3[4])/2 #1st quartile
q2_03 = (sx3[6]+sx3[7])/2 #2nd quartile
q3_03 = (sx3[9]+sx3[10])/2 #3rd quartile
iqr3 = q3_03- q1_03
range3 = sx3[12]-sx3[1]
boxplot(x3)

# 6
boxplot(x3)
```

Exercise 4

The following table has been observed on a sample of 400 subjects

	Political Orientation		
Area	Right	Left	Independent
North	68	56	32
Center	52	72	20
South	33	32	35

1. Compute the marginal percentage distribution for Area and Political Orientation.
2. Compute the conditional percentage distribution for the variable Political Orientation for each Area.
3. Is there any association between these two variables?
4. If yes, which are the categories that are in correspondence?

P(AREA | POLIT) =

	R	L	I	
N	68 48.6%	56 35.9%	32 20.5%	156
C	52 39.1%	72 44.9%	20 12.5%	160
S	33 32.0%	32 32.0%	35 35.0%	100
	156 38.7%	160 39.7%	87 21.6%	403

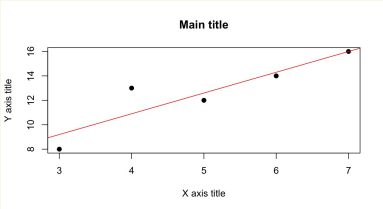
Exercise 5

Consider the data

x	y
3	8
4	13
5	12
6	14
7	16

1. Sketch a scatterplot.
2. Would you expect a positive association, a negative association or no association between x and y?
3. Compute the correlation coefficient, r.

r = 0.906

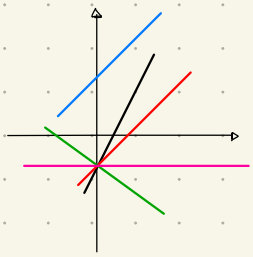



```
x <- c(3,4,5,6,7)
y <- c(8,13,12,14,16)

mx=mean(x)
my=mean(y)


xp=(x-mx)
xp2=xp**2

yp=(y-my)
yp2=yp**2


r=(sum(xp*yp))/(sqrt(sum(xp2)*sum(yp2)))


plot(x, y, main = "Main title",
    xlab = "X axis title", ylab = "Y axis title",
    pch = 19, frame = TRUE)
abline(lm(y ~ x, data = mtcars), col = "red")
```

Exercise 6

Identify the values of the y-intercept a and the slope b. For each regression line, say if you expect to have a positive, negative or zero correlation.

1. $\hat{y} = -3 + 1.5x$.
2. $\hat{y} = -3 + x$.
3. $\hat{y} = -3 - x$.
4. $\hat{y} = 3 + x$.
5. $\hat{y} = -3$.



Exercise 7

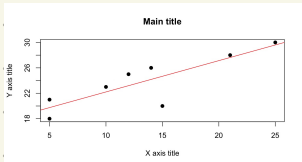
An instructor of Statistics collected data from one of her classes in Spring 2016 to investigate the relationship between Study time per week (number of hours) to predict the final grade. For the 8 students in her class the data were as shown in the table.

Student	Study Time	Grade
1	14	26
2	25	30
3	15	20
4	5	18
5	10	23
6	12	25
7	5	21
8	21	28

1. Identify the response variable and the explanatory variable.
2. Construct a scatterplot.
3. Find and interpret the correlation.

(1) GRADE = response , STUDY TIME = Explanatory variable

(2)




```
x <- c(14,25,15,5,10,12,5,21)
y <- c(26,30,20,18,23,25,21,28)

mx=mean(x)
my=mean(y)


xp=(x-mx)
xp2=xp**2

yp=(y-my)
yp2=yp**2


r=(sum(xp*yp))/(sqrt(sum(xp2)*sum(yp2)))


plot(x, y, main = "Main title",
    xlab = "X axis title", ylab = "Y axis title",
    pch = 19, frame = TRUE)
abline(lm(y ~ x, data = mtcars), col = "red")
```

(3) $$r = \frac{\sum_{i=1}^{8} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{8} (x_i - \bar{x})^2 \sum_{i=1}^{8} (y_i - \bar{y})^2}} = 0.85$$

Exercise 8 - to solve on your own

The following dataset represents the IQ scores of 10 mothers and their eldest daughters.

Mother's IQ	Daughter's IQ
135	121
127	131
124	112
120	115
115	99
112	118
104	106
96	89
94	92
85	90

1. Compute the mean and the sample standard deviation for Mother's and Daughter's IQ.
2. Compute the quartiles for Mother's and Daughter's IQ.
3. Draw the boxplot for Daughter's IQ.
4. Calculate the correlation coefficient between Mother's and Daughter's IQ.
5. Find and interpret the prediction equation.
6. Predict the Daughter's IQ if the observed Mother's IQ is 115.
7. Compute the residual for the 5th observation.
8. Draw the cloud of points and the regression line.

Nella cartella Codici

RECITATION 3

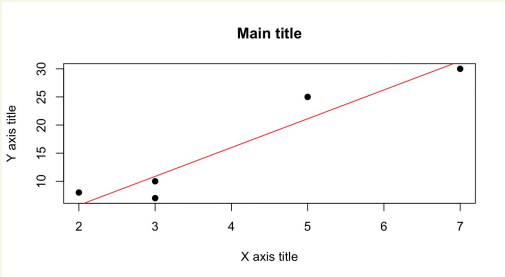
Exercise 1

The table contains size of house in thousands of square feet (X) and selling price in thousands of dollars (Y) for 5 houses.

X	Y
5	25
2	8
3	7
7	30
3	10

1. Compute the regression line of Y on X and interpret the intercept and the slope.
2. Sketch a scatterplot and draw the regression line.
3. Does the regression line fit the data? Motivate your answer.
4. Predict the selling price of a house whose size is 2.5 thousands of square feet.
5. Compute the residual for the last observation ($x = 3$ and $y = 10$).

$r = 0.96$




```
x <- c(5,2,3,7,3)
y <- c(25,8,7,30,10)

mx=mean(x)
my=mean(y)


xp=(x-mx)
xp2=xp**2

yp=(y-my)
yp2=yp**2


r=(sum(xp*yp))/(sqrt(sum(xp2)*sum(yp2)))

slope = sum(xp*yp)/sum(xp2)
q = my-slope*mx

pred =2.5* slope + q

residual = 10 - (3* slope + q)

plot(x, y, main = "Main title",
    xlab = "X axis title", ylab = "Y axis title",
    pch = 19, frame = TRUE)
abline(lm(y ~ x, data = mtcars), col = "red")
```

slope $= m = \dfrac{\sum (x_i - \bar{x})(x_i - \bar{y})}{\sum (x_i - \bar{x})^2} = 5.125$

$f(2,5) = 8.3125$

$q = \bar{y} - m\bar{x} = -4.5$

5 Residual = y observed − y predicted = −0.875

Exercise 2

A company manufactures an electronic device. The company knows that increased temperature shortens the life time of the device. A study is performed in which the life time of the device is determined as a function of temperature. The following data are collected:

Temperature in Celcius (x)	10	20	30	40	50	60	70	80	90
Life time in hours (y)	420	365	285	220	176	117	69	34	5

1. Compute the least squares regression line of Y on X and interpret the intercept and the slope.
2. Draw the cloud of points and the regression line.
3. Provide a measure of goodness of fit and comment it. $r^2 = 0.98$
4. Predict the Life Time of a device if the observed temperature is 45 degrees. → 214.55
5. Compute the residual for the last observation ($x = 90$ and $y = 5$). residual = 29.64

```
x <- c(10,20,30,40,50,60,70,80,90)
y <- c(420,365,285,220,176,117,69,34,5)

mx=mean(x)
my=mean(y)


xp=(x-mx)
xp2=xp**2

yp=(y-my)
yp2=yp**2


r=(sum(xp*yp))/(sqrt(sum(xp2)*sum(yp2)))
r2=r**2

slope = sum(xp*yp)/sum(xp2)
q = my-slope*mx

pred =45* slope + q

residual = 5 - (90* slope + q)

plot(x, y, main = "Main title",
    xlab = "X axis title", ylab = "Y axis title",
    pch = 19, frame = TRUE)
abline(lm(y ~ x, data = mtcars), col = "red")
```

slope = −5.31

q = 453

3 since r^2 is almost 1 it means that it is almost a perfect fit.

Exercise 3

Esercizio 8 Codici

The following dataset represents the IQ scores of 10 mothers and their eldest daughters.

Mother's IQ	Daughter's IQ
135	121
127	131
124	112
120	115
115	99
112	118
104	106
96	89
94	92
85	90

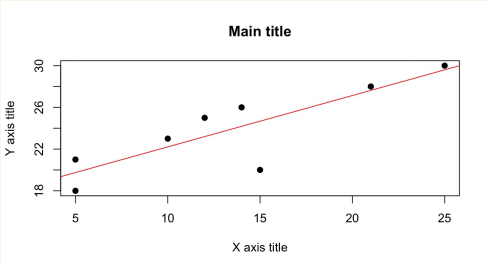
1. Compute the mean and the sample standard deviation for Mother's and Daughter's IQ.
2. Compute the quartiles for Mother's and Daughter's IQ.
3. Draw the boxplot for Daughter's IQ.
4. Calculate the correlation coefficient between Mother's and Daughter's IQ.
5. Find and interpret the prediction equation.
6. Predict the Daughter's IQ if the observed Mother's IQ is 115.
7. Compute the residual for the 5th observation.
8. Draw the cloud of points and the regression line.

Exercise 4

An instructor of Statistics collected data from one of her classes in Spring 2017 to investigate the relationship between Study time per week (number of hours) to predict the final grade. For the 8 students in her class the data were as shown in the table.

Student	Study Time	Grade
1	14	26
2	25	30
3	15	20
4	5	18
5	10	23
6	12	25
7	5	21
8	21	28

1. Identify the response variable and the explanatory variable.
2. Construct a scatterplot.
3. Find and interpret the correlation. $r = 0.85$
4. Compute and interpret r^2. $r^2 = 0.91$
5. Find and interpret the prediction equation. Find and interpret the correlation.
6. Find the predicted score for a student who studies (i) 5 hours per week, (ii) 25 hours per week. $5 \to 19.75$ $25 \to 29.60$
7. Find and interpret the residual for student 4. -1.95




```
x <- c(14,25,15,5,10,12,5,21)
y <- c(26,30,20,18,23,25,21,28)

mx=mean(x)
my=mean(y)

xp=(x-mx)
xp2=xp**2

yp=(y-my)
yp2=yp**2

r=(sum(xp*yp))/(sqrt(sum(xp2)*sum(yp2)))
r2=r**2

slope = sum(xp*yp)/sum(xp2)
q = my-slope*mx

pred1 = 5* slope + q
pred2 = 25* slope + q

residual = 18 - (5* slope + q)

plot(x, y, main = "Main title",
    xlab = "X axis title", ylab = "Y axis title",
    pch = 19, frame = TRUE)
abline(lm(y ~ x, data = mtcars), col = "red")
```

Recitation 4

Exercise 1

A box contains four balls: one red, one blue, one yellow and one pink. Consider an experiment that consists of drawing a ball from the box at random, replacing it, and drawing a second ball.

1. What is the sample space of this experiment?
2. Let A the event that the first ball drawn is yellow. List all the outcomes in A
3. Let B the event that both balls have the same color. List all the outcomes in B
4. Assuming all outcomes have the same probability
   (a) Compute $P(A)$ and $P(A^c)$
   (b) Compute $P(B)$ and $P(B^c)$
   (c) Compute $P(\text{A and B})$
   (d) Compute $P(\text{A or B})$
   (e) Compute $P(A \mid B)$

$\Omega = \{R, B, Y, P\}^2$ 4^2

$A = \{(Y,B), (Y,R), (Y,P), (Y,Y)\}$

$B = \{(Y,Y), (B,B), (R,R), (P,P)\}$

(a) $P(A) = \frac{4}{4^2} = \frac{1}{4} = 0.25$

$P(A^c) = 1 - P(A) = 0.75$

(b) $P(B) = \frac{1}{4} = 0.25$

$P(B^c) = 1 - P(B) = 0.75$

(c) $P(A \cap B) = \frac{1}{16} = 0.06$

(d) $P(A+B) = \frac{7}{16} = 0.44$

(e) $P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.06}{0.22} = 0.\overline{2}$

Exercise 2

Of 12 bottles in a case of wine, 3 are bad. Suppose 2 bottles are randomly chosen from the case. Find the probability that

(a) The first bottle is good

(b) Both bottles are good

(c) Both bottles are bad

(d) The first bottle is good and the second is bad

(e) One bottle is good and one is bad

a) $\frac{9}{12}$

b) $\frac{9}{12} \cdot \frac{8}{11}$

c) $\frac{3}{12} \cdot \frac{2}{11}$

d) $\frac{9}{12} \cdot \frac{3}{11}$

e) $\frac{9}{12} \cdot \frac{3}{11} + \frac{3}{11} \cdot \frac{9}{12}$

Exercise 3

Based on records of automobile accidents, it has been reported the counts who survived (S) and died (D) according to whether they wore a seat belt (Y=yes, N=no). The data are presented in the following contingency table.

Wore Seat Belt	Survived (S)	Died (D)
Yes (Y)	410	5
No(N)	160	15

(a) What is the sample space of possible outcomes for a randomly selected individual involved in an auto accident?

(b) Compute (i) P(D), (ii) P(N).

(c) Compute the probability that an individual did not wear a seat belt and survived.

(d) Based on part 1, what would the answer to part 3 have been if the events N and S were independent? So, are N and S independent, and if not, what does that mean in the context of these data?

(e) Compute the probability that the individual survived, given that the person (i) wore and (ii) did not wear a seat belt. Interpret the results.

(f) Are the events of dying and wearing a seat belt independent? Justify your answer.

	Survived (S)	Died (D)	
Y	410 .99	5 .01	415
N	160 .914	15 .09	175
	570	20	590

a) {YS, YD, NS, ND}

b) $P(D) = \frac{20}{590} = 0.03$

$P(N) = \frac{175}{590} = 0.30$

c) $P(S \cap N) = \frac{160}{590} = 0.27$

d) $P(S \cap N) = P(S) \cdot P(N) = \frac{570}{590} \cdot \frac{175}{590} = 0.27$

e) $P(S|Y) = .99$

$P(S|N) = .914$

f) $P(D \cap Y) = \frac{5}{590} = .0085$

$P(D) \cdot P(Y) = \frac{20}{590} \cdot \frac{415}{590} = 0.024$ NOT INDEP

Exercise 4

The following table classifies US population by age and sex. Values are in units of 1 million.

Age	Sex: Female	Sex: Male
Under 25 (<)	48.8	50.4
At least 25 (≥)	74.5	66.6

Consider a random experiment where a person is chosen at random. Let A be the event *the person is a male* and B the event *the person is under age 25*. Compute

(a) $P(A)$ and $P(A^c)$

(b) $P(B)$ and $P(B^c)$

(c) $P(\text{A and B})$ and $P((\text{A and B})^c)$

(d) $P(\text{A or B})$ and $P((\text{A or B})^c)$

(e) $P(A \mid B)$ and $P(B \mid A)$

(f) Are the two events independent?

	F	M	
<25	48.8 .49	50.4 0.51	99.2
≥25	74.5 .63	66.6 .47	141.2
	123.3	117	240.3

$P(A) = \{\text{the person is a male}\} = P(M)$

$P(B) = \{\text{the person is under age 25}\} = P(<25)$

a) $P(A) = \frac{117}{240.3} = 0.49$ $\quad P(A^c) = 1 - P(A) = 0.51$

b) $P(B) = \frac{99.2}{240.3} = 0.41$ $\quad P(B^c) = 1 - P(B) = 0.59$

c) $P(A \cap B) = P(M \cap <25) = \frac{50.4}{240.3} = 0.21$

$P((A \cap B)^c) = 1 - P(A \cap B) = 0.79$

d) $P(A \text{ or } B) = \frac{66.6 + 48.8 + 50.4}{240.3} = 0.69$

$P((A \text{ or } B)^c) = 1 - P(A \text{ or } B) = 0.31$

e) $P(A|B) = P(M|<25) = \frac{P(M \cap <25)}{P(<25)} = \frac{50.4}{99.2} = 0.50$

$P(B|A) = P(<25|M) = \frac{P(<25 \cap M)}{P(M)} = \frac{50.4}{117} = 0.43$

$P(M) \cdot P(<25) = \frac{117}{240.3} \cdot \frac{99.2}{240.3} = 0.20$ DEPENDENT

Exercise 5

It has been asked to 1000 people whether they were happy in their marriages.The poll reported that 46% were men, while 54% were women. Among men, 40% declared to be very happy, 51% pretty happy, while 9% not too happy. Among women, 43% declared to be very happy, 49% pretty happy, while 8% not too happy.

(a) Set up a contingency table that cross classifies gender by level of happiness.

(b) Compute the probability that a married adult is very happy.

(c) Compute the probability that a married adult is very happy, (i) given that their gender is male and (ii) given that their gender is female.

(d) For these subjects, are the events being very happy and being a male independent?

	M	F	
Ü	184	232.2	416.2
·· ‿	234.6	264.6	499.2
·· ^	41.4	43.2	84.6
	460	540	1000

$P(w \mid Ü) = \frac{232.2}{416.2}$

$P(Ü \mid w) = \frac{232.2}{540}$

(b) $P(Ü) = \frac{416.2}{1000}$

(c) (i) $P(Ü \mid M) = \frac{184}{460} = 0.4$

(ii) $P(Ü \mid F) = \frac{232.2}{540} = 0.43$

(d) $P(Ü) \cdot P(M) = \frac{416.2 \cdot 460}{1000^2} = 0.191$

$P(Ü \cap M) = \frac{184}{1000} = 0.184$

Exercise 6

A population is composed of 2000 American children. It has been asked them questions about their participation in organized and team sports. Among them, 52.7% were boys and 47.3% were girls. Of the boys, 73% currently participate in sports. Of the girls, 63% currently participate in sports. Gender and sports participation are not independent events.

(a) If you randomly select an individual from this population, what is the probability that you would select a girl who currently participates in sports?

(b) What is the probability that a randomly selected individual from this population would be a boy who does not currently participate in sports?

	F	M	
Sport	595,98	769,42	1365,4
NO Sport	350,02	284,58	634.6
	946	1054	2000

$P(F \cap \text{sport}) = \frac{595,98}{2000} = 0.298$

$P(M \cap \text{no sport}) = \frac{284.58}{2000} = 0.142$

Exercise 7

In your pocket, there are 5 keys that all look the same. You need to find the right one to open the front door of your home.

(a) Compute the probability to find the right key on your second attempt assuming that you try them once at time, with no repetition.

—> DOVREBBE ESSERE GIUSTO

5 Keys

right Key = 1

$P(\text{right Key } 1^{st} \text{ chance}) = \frac{1}{5}$

$P(\text{right Key } 2^{nd} \text{ chance}) = \frac{1}{4}$

$\left(\frac{4}{5} \cdot \frac{1}{4}\right) \cdot 4 = \frac{1}{5} \cdot 4 = \frac{4}{5}$

1 2 3 4

$\frac{4}{5} \cdot \frac{1}{4} + \frac{4}{5} \cdot \frac{1}{4} + \frac{4}{5} \cdot \frac{1}{4} + \frac{4}{5} \cdot \frac{1}{4}$

RECITATION 5

Exercise 1

A wheat farmer living in Pennsylvania finds that his annual profit is \$ 80 if the summer weather is typical, \$ 50 if the weather is unusually dry, and \$ 20 if there is a severe storm that destroys much of his crop. Weather bureau records indicate that the probability is 0.70 of typical weather, 0.20 of unusually dry weather, and 0.10 of a severe storm. Let X be the farmer's profit.

1. Construct the probability distribution of X.
2. Find the mean of the probability distribution of X.
3. Find the variance of the probability distribution of X.

$$X \sim \begin{cases} 80 & 0.80 \\ 50 & 0.20 \\ 20 & 0.10 \end{cases}$$

$$E(X) = \sum x \cdot P(x) = 68$$

$$Var(X) = E(X^2) - E(X)^2 = 5020 - 4624 = 396$$

Exercise 2

An instructor always assigns final grades such that 20% are A, 40% are B, 30% are C, and p% are D. The grade point scores are 4 for A, 3 for B, 2 for C, and 1 for D.

1. Find p% and construct the probability distribution for the grade point score of a randomly selected student of this instructor.
2. Find the mean and the variance of this probability distribution.

①

$$X \begin{cases} 4 & 0.2 \\ 3 & 0.4 \\ 2 & 0.3 \\ 1 & 0.1 \end{cases}$$

②

$$E(X) = 4 \cdot 0{,}2 + 3 \cdot 0{,}4 + 2 \cdot 0{,}3 + 0{,}1 = 2.7$$

$$E(X^2) = 8.1$$

$$E(X)^2 = 7{,}29$$

$$Var(X) = 8.1 - 7.29 = 0.81$$

Exercise 3

Let X be a random variable whose possible values are 5, 10, 15, 20. Assuming $P(X = 5) = 0.3$, $P(X = 10) = 0.1$ and $P(X = 15) = 0.2$, compute

1. $P(X = 20)$
2. the expected value of X
3. the variance and the standard deviation of X
4. the probability $P(X > 10)$

① $P(X=20) = 1 - (0.2 + 0.1 + 0.3) = 0.4$

② $E(X) = 13.5$

③ $Var(X) = E(X^2) - E(X)^2 = 222{,}5 - 182{,}25 = 40{,}25$

$$X \begin{cases} 20 & 0.4 \\ 15 & 0.2 \\ 10 & 0.1 \\ 5 & 0.3 \end{cases}$$

$\sigma = \sqrt{40{,}25} = 6{,}34$

④ $P(X > 10) = 1 - P\left(\frac{\bar{x} - \mu}{\sigma} < \frac{10 - 13.5}{6.34}\right) = 1 - \Phi(-0.55) = 1 - .29116 =$

$= 0.70884$

Exercise 4

An estate agency has scheduled two appointments with potential buyers. The seller thinks the probability that the first appointment leads to a sale is 0.3, while this probability is 0.6 for the second appointment. Also, the agency thinks the appointments are independent. Let X be the random variable: n. of sales.

1. Which is the probability distribution of X?
2. Which is the probability of at least one sale?
3. Which is the expected number of sales?
4. Which is the variance of X? And the standard deviation?
5. Let Y be the random variable obtained from the following linear transformation $Y = 2X + 3$. Compute the expected value, the variance and the standard deviation of Y

$Y_1 \sim B_o(0.3)$
$Y_2 \sim B_o(0.6)$

$X = Y_1 + Y_2$

$x=0$: $0.7 \cdot 0.4 = 0.28$
$x=1$: $0.3 \cdot 0.4 + 0.7 \cdot 0.6 = 0.54$
$x=2$: $0.3 \cdot 0.6 = 0.18$

① $X \begin{cases} 0 & 0.28 \\ 1 & 0.54 \\ 2 & 0.18 \end{cases}$

② $P(X \geq 1) = 0.54 + 0.18 = 0.72$

③ $E(X) = 0.9$

④ $Var(X) = 1.26 - 0.81 = 0.45$
$s = \sqrt{.45} = 0.67$

⑤ $Y = (2X+3)$

$Var(cX) = c^2 Var(X)$

$s(cX) = \sqrt{Var(cX)} = \sqrt{c^2 Var X} =$
$= c \cdot \sqrt{Var(X)}$

$E(Y) = 2\,E(X) + 3 = 4.8$
$Var(Y) = 4 \cdot Var(X) = 1.8$
$s(Y) = 2 \cdot s(X) = 1.34$

Exercise 5

In a population the vehicle speed distribution is well approximated by a Normal curve with mean 50 and standard deviation 15. Compute the probability that

1. a randomly selected vehicle speed is greater than 73;
2. a randomly selected vehicle speed is between 40 and 73;
3. a randomly selected vehicle speed is less then 85.

$X \sim N(50, 15^2)$

① $P(X > 73) = 1 - P\left(Z < \frac{73-50}{15}\right) = 1 - \Phi(1{,}53) = 1 - 0.93699 = 0.06301$

② $P(40 < X < 73) = P(Z < 73) - P(Z < 40) = \Phi\left(\frac{73-50}{15}\right) - \Phi\left(\frac{40-50}{15}\right) = 0.93699 - 0.25143 = 0.68556$

③ $P(X < 83) = P\left(Z < \frac{83-50}{15}\right) = \Phi\left(\frac{83-50}{15}\right) = \Phi(2{,}2) = 0.98610$

Exercise 6

At one private college, a minimum IQ score of 125 is necessary to be considered for admission. IQ scores can be well approximated with a Normal distribution with mean of 100 and standard deviation of 15.

1. Compute the probability of being admitted.
2. Compute the probability that a randomly selected IQ score is between 120 and 145.
3. Compute the probability that a randomly selected IQ score is less than 125.
4. Compute the probability that a randomly selected IQ score is less than 90.

$X \sim N(100, 15^2)$

① $P(x \geq 125) = 1 - \Phi\left(\frac{125-100}{15}\right) = 1 - \Phi(1{,}67) = 1 - 0.95254 = 0.04746$

② $P(120 < X < 145) = P(X < 145) - P(X < 120) = \Phi\left(\frac{145-100}{15}\right) - \Phi\left(\frac{120-100}{15}\right) =$

$= \Phi(3) - \Phi(1{,}33) = .99865 - .90824 = 0.09041$

③ $P(X < 125) = \Phi\left(\frac{125-100}{15}\right) = 0.95254$

④ $P(X < 90) = \Phi\left(\frac{90-100}{15}\right) = \Phi(-0.67) = 0.25143$

Exercise 7

Let the random variable Z follow a standard Normal distribution. Compute the probabilities below:

1. P(-2 < Z < -1)
2. P(Z > 1.52)
3. P(-2 < Z < 0.89)
4. P(0 < Z < 2.15)

$N(0,1)$

① $P(Z < -1) - P(Z < -2) = \Phi\left(\frac{-1-0}{1}\right) - \Phi(-2) = 0.15866 - 0.02275 = 0.1359$

② $1 - \Phi(1.52) = 1 - 0.93594 = 0.06426$

③ $\Phi(0.89) - \Phi(-2) = 0.81327 - 0.02275 = 0.7905$

④ $\Phi(2.15) - \Phi(0) = 0.98422 - 0.5 = 0.48422$

Exercise 8

The Mental Development Index (MDI) of the Bayley Scales of Infant Development is a standardized measure used in observing infants over time. It is approximately normal with a mean of 100 and a standard deviation of 16.

1. What proportion of children has an MDI of (i) at least 120? (ii) at least 80?
2. Find the MDI score that is the 99th percentile.
3. Find the MDI score such that only 1% of the population has MDI below it.
4. Find the interquartile range (IQR) of MDI scores.

$X \sim N(100, 16^2)$

① $P(X \geq 120) = 1 - \Phi\left(\frac{120-100}{16}\right) = 1 - \Phi(1{,}25) = 1 - 0.89435 = 0.10565$

② $Z_{(0.99)} \approx 2.33$

				0.03
2.1	.98214	.98257	.98300	.98341
2.2	.98610	.98645	.98679	.98713
2.3	.98928	.98956	.98983	.99010
2.4	.99180	.99202	.99224	.99245

$X_{(99)} = 2.33 \cdot 16 + 100 = 137{,}28$

③ $Z_{(01)} \approx -2{,}325$ $\quad X_{(01)} = -2{,}325 \cdot 16 + 100 = 62{,}8$

④ $Q_{1\,(25)} = -0.67 \cdot 16 + 100 = 89.28$

$Q_{3\,[75]} = 0.68 \cdot 16 + 100 = 110{,}88$ $\qquad$ $IQR = 110.88 - 89.28 = 21{,}6$

Exercise 9

The heights of adult males are normally distributed with a mean of 175 cm and a standard deviation of 11 cm. Let X denote the height of a randomly chosen adult male.

1. Compute $\Pr(X > 165)$
2. Compute $\Pr(157 < X < 183)$
3. Compute $\Pr(X < 152)$
4. Compute the value of x if $\Pr(X > x) = 0.01$
5. Compute the value of x if $\Pr(X > x) = 0.025$

$X \sim N(175, 11^2)$

① $P(X > 165) = 1 - P(X < 165) = 1 - \Phi\left(\frac{165-175}{11}\right) = 1 - \Phi(-0.91) = 1 - 0{,}18141 = 0{,}81859$

② $\Phi\left(\frac{183-175}{11}\right) - \Phi\left(\frac{157-175}{11}\right) = \Phi(0.727) - \Phi(1{,}64) = 0.7673 - 0.0505 = 0.7168$

③ $P(X < 152) = \Phi\left(\frac{152-175}{11}\right) = \Phi(-2{,}09) = 0.01831$

④ $P(X > x) = 0.01$

$1 - \Phi\left(\frac{x-175}{11}\right) = 0.01$ $\qquad$ $1 - 0.01 = \Phi\left(\frac{x-175}{11}\right)$ $\qquad$ $2.33 = \frac{x-175}{11}$ $\qquad$ $x = 200.63$

⑤ $P(X > x) = 0.025$ $\qquad$ $1 - 0.025 = \Phi\left(\frac{x-175}{11}\right)$ $\qquad$ $1.96 = \frac{x-175}{11}$ $\qquad$ $x = 196.56$

Exercise 10

Each newborn baby has a probability of approximately 0.49 of being female and 0.51 of being male. For a family of four children, let X = number of children who are girls.

1. Explain why the three conditions are satisfied for X to have the binomial distribution.
2. Identify n and p for the binomial distribution.
3. Compute the mean and the variance of X.
4. Find the probability that the family has two girls and two boys.

$F \sim Be(0.49)$

$M \sim Be(0.51)$

$X \sim Bin(n, p)$ $\qquad$ $X \sim Bin(4, 0.49)$

$$X \begin{cases} 4 & p^4 \binom{4}{4} \\ 3 & \bar{p} \cdot p^3 \cdot \binom{4}{3} \\ 2 & \bar{p}^2 \cdot p^2 \cdot \binom{4}{2} \\ 1 & \bar{p}^3 \cdot p \quad \binom{4}{1} \\ 0 & \bar{p}^4 \quad \binom{4}{0} \end{cases}$$

$E(X) = n \cdot p = 4 \cdot 0.49 = 1.96$

$Var(X) = n \cdot p(1-p) = 0.9996$

$P(X=2) = \bar{p}^2 \cdot p^2 \binom{4}{2} = 0.49^2 \cdot 0.51^2 \cdot \frac{4!}{2! \cdot 2!} = 0.49^2 \cdot 0.51^2 \cdot 6 = 0.3747$ $\qquad$ ($4! = 24 \cdot 3 \cdot 2!$)

Exercise 11

Consider the experiment of tossing a coin three times. Determine the probability that there are a total of i heads, for $i = 0, 1, 2, 3$.

$X \sim Bin(n, p) \rightarrow X \sim Bin(3, 0.5)$

$P(X=0) = \frac{3!}{0!\,3!} \; 0.5^3 = 0.125$ $\qquad$ $P(X=3) = \frac{3!}{3!\,0!} \; 0.5^3 = 0.125$

$P(X=1) = \frac{3 \cdot 2!}{1!\,2!} \; 0.5^1 \cdot 0.5^2 = 0.375$

$P(X=2) = \frac{3 \cdot 2!}{2!\,1!} \; 0.5^2 \cdot 0.5^2 = 0.375$

Exercise 12

Suppose that each microwave oven produced by a company is defective with probability 0.05. For a sample of 4 microwave ovens

1. determine the probability that all of them have no defects
2. determine the probability of having less than two defective microwaves
3. determine the mean, the variance and the standard deviation of the number of defective microwave ovens

$X \sim Bin(4, 0.05)$

① $P(X=0) = \frac{4!}{0!\ 4!} \overline{0.05}^4 = 0.81$

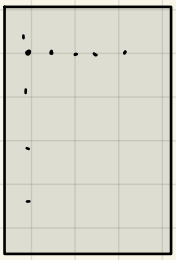
② $P(X<2) = 0.81 + \frac{4 \cdot \cancel{3!}}{1!\ \cancel{3!}} \overline{0.05}^3\ 0.05 = 0.81 + 0.19 = 0.98$

③ $E(X) = 4 \cdot 0{,}05 = 0.2$ $\quad Var(X) = 4\ 0{,}05 \cdot \overline{0.05} = 0.19$ $\quad \sigma = \sqrt{0.19} = 0{,}44$

Exercise 13

A quiz in statistics course has four multiple-choice questions, each with five possible answers. A passing grade is three or more correct answers to the four questions. Allison has not studied for the quiz. She has no idea of the correct answer to any of the questions and decides to guess at random for each.

1. Find the probability she lucks out and answers all four questions correctly.
2. Find the probability that she passes the quiz.



$X \sim Bin(4, \frac{1}{5})$

① $P(X=4) = \frac{\cancel{4!}}{\cancel{4!}\ \cancel{0!}} \cdot \frac{1}{5}^4 = 0.0016$

② $P(X \geq 3) = 0.0016 + \frac{4 \cdot \cancel{3!}}{\cancel{3!}\ 1!} \frac{1}{5}^3 \frac{4}{5} = 0.0272$

Exercise 14

A hotel has 100 rooms and the probability a room is occupied on any given night is 0.6. Assuming independence between rooms, let X be the random variable: n of occupied rooms in the hotel

1. What is the probability of having less than 50% occupancy?
2. What is the probability of at least 75% occupancy on any given night? What is the probability of having between 50% and 75% occupancy?

$X \sim Bin(100, 0.6)$

$X \sim N(\mu, \sigma^2)$ $\quad \mu: n \cdot p > 15$ $\quad \sigma^2: n \cdot p \cdot \bar{p} > 15$

$X \sim N(60, 24)$

we can approximate to a Normal Distribution

$X \sim N(\mu, \sigma^2)$

But we must check first

$n \cdot p > 15$ $\quad \mu = 100 \cdot 0.6 = 60$

$n \cdot p(1-p) > 15$ $\quad \sigma^2 = 100 \cdot 0{,}6 \cdot 0{,}4 = 24$

① $P(X < 50) = \Phi\left(\frac{50-60}{2\sqrt{6}}\right) = \Phi(-2{,}04) = 0.02068$

② $P(X > 75) = 1 - \Phi\left(\frac{75-60}{\sqrt{24}}\right) = 1 - \Phi(3{,}06) = 1 - 0.99889 = 0.00111$

$P(50 < X < 75 = 0.99889 - 0.02068 = 0.97821$

RECITATION 6

Exercise 1

For the population of individuals who own an iPhone, suppose $p = 0.25$ is the proportion that has a given app.

1. For a random sample of size $n = 4$, find the mean and the standard deviation of the sampling distribution of the sample proportion.
2. Find the probability that the proportion of having the app is at least 0.75 when $n = 4$. [Hint: first, find x.].
3. For a random sample of size $n = 150$, find the mean and the standard deviation of the sampling distribution of the sample proportion.
4. Summarize the effect of the sample size on the standard deviation of the sampling distribution of the sample proportion.
5. Describe the shape of the sampling distribution of the sample proportion when $n = 150$.
6. Find the probability that the proportion of having the app is at least 0.75 when $n = 150$.

① $\mu = 0.25$ $\quad \sigma = \sqrt{\frac{0.25(0.75)}{4}} = 0.2165$

② $\hat{p} = \frac{x}{n} \rightarrow \quad x = \hat{p} \cdot n = 0.75 \cdot 4 = 3$

$P(x \geq 3) = P(x=3) + P(x=4) = \frac{4!}{3!\,1!} \cdot 0.25^3 \cdot 0.75 + 0.25^4 = 0.051$

③ $\mu = 0.25$ $\quad \sigma = \sqrt{\frac{0.25(0.75)}{150}} = 0.0354$

④ As n increases the standard deviation decreases. This means that the distribution is more concentrated around the mean, that is the sample proportion.

⑤ $x = 0.75 \cdot 150 = 112,5$

$P(x > 0.75) = \Phi\left(\frac{0.75 - 0.25}{0.0354}\right) = \Phi(14.12) \simeq 0$

Exercise 2

You are studying for taking the final exam in Statistics. Based on your preparation, for any given question you think you have a probability of $p = 0.70$ of getting the correct answer. An exam consists of 60 multiple-choice questions. Consider the sampling distribution of the sample proportion of the 60 questions on which you get the correct answer.

1. Find the mean and the standard deviation of the sampling distribution of this proportion.
2. What do you expect for the shape of the sampling distribution? Motivate your answer. $\rightarrow$ By CLT we can approximate to a Normal distribution
3. If truly $p = 0.70$, would be very surprising if you got correct answers on only 60% of the questions? Justify your answer by using the normal distribution to approximate the probability of a sample proportion of 0.60 or less.

① $\mu = 0.70$ $\quad \sigma = \sqrt{\frac{0.7 \cdot 0.3}{70}} = 0.0591$

② $P(x \leq 0.60) = \Phi\left(\frac{0.6 - 0.70}{0.059}\right) = \Phi(-1,69) = 0.4551$

Exercise 3

Recent studies have shown that about 15% of Italian adults fit the medical definition of being obese. For a random sample of 20 patients, 4 were obese.

1. Which is the proportion of obese individuals in the population?
2. Which is the proportion of obese in the sample?
3. Compute the mean of the sampling proportion
4. Compute the variance and the standard deviation of the sampling proportion

① 0.15

② 0.2

③ .15 ④ $\sigma^2 = \frac{0.15 \cdot 0.75}{20} = 0.006$ $\sigma = 0.077$

③ 0.2 ④ $\sigma^2 = \frac{0.2 \cdot 0.8}{20} = 0.021$ $\sigma = 0.145$

Exercise 4

The 30% of individuals in the population of Android's users downloaded a given app in 2015. For a random sample of 50 users, 18 had the app.

1. Which is the population proportion?
2. Which is the sample proportion?
3. Determine the sampling distribution of the sample proportion
4. What happens to the variance and to the standard deviation of the sample proportion if the sample size is $n = 100$?

① 0.30

② $\frac{18}{50} = 0.36$

③ $\mu = 0.30$ $\sigma = \sqrt{\frac{0.30 \cdot 70}{50}} = 0.065$ $\sigma^2 = 0.0042$

④ $\sigma = \sqrt{\frac{0.30 \cdot 0.70}{100}} = 0.0458$ $\sigma^2 = 0.0021$

Exercise 5

In a city, 70% of the people prefer Candidate A. Suppose 100 people from this city were sampled. Compute the probability that 80% or more of this sample will prefer Candidate A.

① $\mu = 0.7$ $\sigma^2 = \frac{0.7 \cdot 0.3}{100} = 0.0021$

$\sigma = 0.0458$

$P(x > 80) = 1 - \Phi\left(\frac{0.80 - 0.7}{0.0458}\right) = 1 - \Phi(2.18) =$

$= 1 - 0.98537 = 0.01463$

Exercise 6

The distribution of profit per year of a restaurant is approximately normal with a mean of 65 (in thousands of dollars) and a variance of 225.

1. Compute the probability that a randomly selected restaurant has a year profit greater than 70.
2. Compute the probability that a randomly selected restaurant has a year profit less than 50.
3. Find the first quartile of the profit per year.
4. Find the median of the profit per year.
5. Given a sample of 15 restaurants, find the mean and the standard deviation of the sampling distribution for the sample mean.
6. Describe the shape of the sampling distribution of the sample mean.
7. Compute the probability that in a sample of 15 restaurants, the average for the year profit is between 55 and 75.

pop: $X \sim N(65, \overset{225}{15^2})$

① $P(X > 70) = 1 - \Phi\left(\frac{70-65}{15}\right) = 1 - \Phi(0.33) = 1 - 0.6306 = 0.3694$

② $P(X < 50) = 1 - \Phi(-1) = 0.15866$

③ $Q_1 = 25\%$ $\overset{n.\text{table}}{\Rightarrow} -0.67$

$Q_1 = -0.67 \cdot 15 + 65 = 54.95$

④ 65

⑤ $\bar{X} \sim N\left(65, \frac{15}{\sqrt{15}} = 3.87\right) \rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

⑦ $\Phi\left(\frac{75-65}{3.87}\right) - \Phi\left(\frac{55-65}{3.87}\right) = \Phi(2.58) - \Phi(-2.58) = 0.99506 - 0.00494 = 0.99012$

Exercise 7

The GPA of 100 randomly selected students among all those graduating from Tor Vergata in 2015 had a median of 2, mean of 3.22, and standard deviation of 0.26. [that is $2.6/\sqrt{100}$]

1. Compute the probability that a randomly selected GPA from the population is between 2.5 and 3.5. [Note that the population standard deviation is 2.6.]
2. Find the GPA score that is the 82th percentile.
3. Find the interquartile range (IQR) of the GPA.
4. What is the point estimate of the population mean?
5. Is the estimator used in (4) unbiased?
6. Compute the probability that the sample mean of GPA is between 2.5 and 3.5.
7. Find the GPA score of the sampling distribution that is the 82th percentile.
8. Find the interquartile range (IQR) of the sampling distribution of the sample mean.

median $=2$

PoP : $X \sim N(3.22, 2.6^2)$

SAMPLE : $\bar{X} \sim N(3.22, 0.26^2)$

① $P(2.5 < X < 3.5) = \Phi\left(\frac{3.5-3.22}{2.6}\right) - \Phi\left(\frac{2.5-3.22}{2.6}\right) = \Phi(0.11) - \Phi(-0.28) = 0.5438 - 0.38974 = 0.15406$

② $82_{TH} = 0.915$

$0.915 \cdot 2.6 + 3.22 = 5.599$

③ $IQR = Q_3 - Q_1 \rightarrow (0.675 \cdot 2.6 + 3.22) - (-0.67 \cdot 2.6 + 3.22) = 3.497$

④ The sample mean (because we know it)

⑤ The sample mean is an unbiased estimator with a standard error that tends to zero when the sample size increases

⑥ $\Phi\left(\frac{3.5-3.22}{0.26}\right) - \Phi\left(\frac{2.5-3.22}{0.26}\right) = \Phi(1.08) - \Phi(-2.78) = 0.85993 - 0.00272 = 0.85721$

⑦ $0.915 \cdot 0.26 + 3.22 = 3.4579$

⑧ $IQR = 0.3497$

Exercise 8

Duration of incoming calls between 9 and 11 am for a call center has a Normal distribution with mean $\mu = 175$ seconds and variance $\sigma^2 = 16$. For a random sample of 10 incoming calls

1. Determine the sampling distribution of the sampling mean
2. Determine the probability that the these 10 incoming calls have a mean duration higher than 176 seconds
3. Determine the probability that the these 10 incoming calls have a mean duration lower than 173 seconds
4. Determine the probability that the these 10 incoming calls have a mean duration between 173 and 176 seconds

Pop : $X \sim (175, 4^2)$

Sample : $\bar{X} \sim \left(175, \frac{4}{\sqrt{10}}^2\right)$

② $P(\bar{X} > 176) = 1 - \Phi\left(\frac{176-175}{1.265}\right) = 1 - \Phi(0.79) = 1 - 0.78524 = 0.21476$

③ $P(X < 173) = \Phi\left(\frac{173-175}{1.265}\right) = \Phi(-1.58) = 0.05705$

④ $P(176 < X < 176) = \Phi(0.79) - \Phi(-1.58) = 0.78524 - 0.05705 = 0.72819$

Exercise 9

In the population, IQ scores are normally distributed with mean $\mu = 100$ and variance $\sigma^2 = 15$. Suppose to draw a random samples of 25 individuals from the population and measure the IQ score

1. compute the probability of observing a sample mean between 98 and 102 when drawing a sample of 25 individuals
2. compute the probability of observing a sample mean between 98 and 102 when drawing a sample of 5 individuals

POP : $X \sim N(100, 15)$

SAMPLE : $\bar{X} \sim N\left(100, \frac{\sqrt{15}}{5}^2\right)$

① $P(98 < X < 102) = \Phi\left(\frac{102-100}{\frac{\sqrt{15}}{5}}\right) - \Phi\left(\frac{98-100}{\frac{\sqrt{15}}{5}}\right) = \Phi(2{,}58) - \Phi(-2{,}58) = 0.99506 - 0.00494 = 0.99012$

② $\Phi\left(\frac{102-100}{\frac{\sqrt{15}}{\sqrt{5}}}\right) - \Phi\left(\frac{98-100}{\frac{\sqrt{15}}{\sqrt{5}}}\right) = \Phi(1.15) - \Phi(-1.15) = 0.87493 - 0.12507 = 0.74986$

RECITATION 7

Exercise 1

When a General Social Survey asked 1326 subjects, "Do you believe in science?", the proportion who answered yes was 0.82.

1. Compute the standard error of this estimator.
2. Construct the 95% confidence interval. Interpret it in context.
3. How does the result in (2) change if you construct a 99% confidence interval?
4. Describe the effect of the sample size on the confidence interval.

90% 1.645
95% 1.96
99% 2.576

① $se = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.82 \cdot 0{,}18}{1326}} = 0.011$

② $0.82 \pm 1.96 \cdot 0.011 = 0.82 \pm 0021$

③ $0.82 \pm 2.576 \cdot 0{,}011 = 0.82 \pm 0{,}028$

④ If the sample size increases, the margin of error decreases, and thus the C.I. become narrower

Exercise 2

We are interested in estimating the mean IQ of students enrolled at Tor Vergata. It has been drawn a random sample of 25 students. Its mean is 120, while the sample standard deviation s is 10.

1. Construct a 90% confidence interval for the population mean.
2. Suppose to increase the sample size to 125, how does the interval in (1) change?
3. If you want to be more confident, what do you should change? Construct the corresponding confidence interval.
4. Another source claims that the mean IQ of students enrolled at Tor Vergata is 117. Is this a plausible value?

$\bar{X} \sim N(120, 10^2)$ $n = 25$

$\hat{p} = \frac{120}{25} = 4.8$

$s^2 = \frac{10}{\sqrt{24}} = 2$

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208

① $\bar{X} \sim N(120, 10^2)$ $n=25$ C.I. = 90% $se = \frac{10}{\sqrt{25}} = \frac{10}{5} = 2$

$120 \pm t_{24, 0.05} \cdot se \Rightarrow 120 \pm 1.711 \cdot 2 \Rightarrow [116.578, 123.422]$

② $\bar{X} \sim N(120, 10^2)$ $n=125$ C.I. = 90% $se = \frac{10}{\sqrt{125}} = 0.89$

$120 \pm 1.645 \cdot 0.89 \Rightarrow [118.529, 121.471]$

③ If we want to be more confident we increase the confidence interval FROM 90% TO 95%

④ No, since with a sample of 125 and with a 90% of C.I. 117 is not in it

Exercise 3

For the following scenarios, identify the appropriate multiplier, the margin of error, and the corresponding confidence interval.

1. The proportion of Tor Vergata students who own a Iphone is 0.65. A random sample of 50 students is drawn. [$\alpha = 0.05$]
2. The mean number of credits completed by students enrolled at Tor Vergata for Fall 16 is 15. A random sample of 100 students is taken. Its standard deviation is 25. [$\alpha = 0.01$]
3. The proportion of registered Italian voters who plan to vote for Candidate A is 0.45. A random sample of 50 registered voters is taken. [$\alpha = 0.10$]
4. The average height of all Tor Vergata students is 175 cm. A random sample of 26 students is taken. Its standard deviation is 5. [$\alpha = 0.05$]

90% 1.645
95% 1.96
99% 2.576

Confidence Level						
	80%	90%	95%	98%	99%	99.8%
Right-Tail Probability						
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208

1. Pop $X \sim Be(0.65)$

$\bar{X} \sim N(0.65, s^2)$ $\hat{p} = 0.65$ $se = \sqrt{\frac{0.65(1-0.65)}{50}} = 0.067$

$0.65 \pm 1.96 \cdot 0.067 \Rightarrow [0.52, 0.78]$

2. Pop $X \sim N(15, \sigma^2)$

Sample $\bar{X} \sim N(15, 25^2)$ $n=100$ $\alpha = 0.01 \Rightarrow 99\%$

$15 \pm 2.576 \cdot \frac{25}{\sqrt{100}} \Rightarrow [8.56, 21.44]$

3. Pop $X \sim Be(0.45)$

Sampl $\bar{X} \sim N(0.45, 0.09^2)$ $se = \sqrt{\frac{.45 \cdot (1-.45)}{50}} = 0.07$ C.I. = 90%

$0.45 \pm 1.645 \cdot (0.09) \rightarrow [0.33, 0.565]$

4. Pop $X \sim N(175, s^2)$ C.I. ~ 95%

Sample $\bar{X} \sim N(175, \frac{5}{\sqrt{26}} = 0.98)$ $n=26$ $175 \pm 2.06 \cdot (0.98) \Rightarrow [172.981, 177.019$

Exercise 4

Describe the effect of standard deviation, sample size and α on the confidence interval.

Exercise 5

At the last referendum, 40% of the Italian population supported the constitutional reform. If a random sample of size $n = 200$ is drawn, which is the probability of observing at least 100 people who voted YES?

Pop $X \sim Be(.40)$

Sample : $n = 200$ $\bar{X} \sim N(.40, 0.035^2)$

$$P(\bar{X} \geq 0.5) = 1 - \Phi\left(\frac{0.5 - 0.4}{0.035}\right) = 1 - \Phi(2.89) = 0.0019$$

Exercise 6

A random sample of 85 students at a large public university revealed that 35 students owned a car that was less than 5 years old.

1. Estimate p, the proportion of all students at the university who own a car less than 5 years old.
2. What is the estimate of the standard error of $\hat{p}$?
3. Compute a 95% confidence interval for p.
4. Would a 90% confidence interval for p be wider or narrower than the previous one?

① $\hat{p} = \frac{35}{85} = 0.412$ $n = 85$

② $se = \sqrt{\frac{0.412 \cdot (1 - 0.412)}{85}} = 0.053$

③ $0.412 \pm 1.96 \cdot (0.053) \rightarrow [0.30812, 0.51588]$

④ $0.412 \pm 1.645 (.053) \rightarrow [0.325, 0.499]$

90% 1.645

95% 1.96

99% 2.576

Exercise 7

A random sample of 1000 construction workers revealed that 122 are presently unemployed.

1. Estimate p, the proportion of all construction workers who are unemployed.
2. Estimate the standard error of $\hat{p}$.
3. Compute a 95% confidence interval for p.

	Confidence Level					
	80%	90%	95%	98%	99%	99.8%
	Right-Tail Probability					
df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.656	318.289
2	1.886	2.920	4.303	6.965	9.925	22.328
3	1.638	2.353	3.182	4.541	5.841	10.214
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.894
6	1.440	1.943	2.447	3.143	3.707	5.208

$n = 1000$ ① $\hat{p} = \frac{122}{1000} = 0.122$ ② $se = 0.01$

③ $0.122 \pm 1.96 \cdot (0.01) \rightarrow [0.1024, 0.1416]$

Exercise 8

The weight given by an electric scale is normally distributed with standard deviation $\sigma = 0.1$ gr. Suppose that the results of five successive weightings of the same object are as follows: 3.142, 3.163, 3.155, 3.150, 3.141.

1. Provide an unbiased estimator for the population mean together with its standard error and its sample distribution.
2. Compute the 95% confidence interval for the mean.
3. Compute the 99% confidence interval for the mean.

Pop $X \sim N(\mu, 0.1^2)$

Sample $\bar{X} \sim N(\mu, \frac{0.1}{\sqrt{5}} = 0.0447^2)$ $n = 5$

$\hat{\mu} = 3.1502$

① 95% $3.1502 \pm 1.96 (0.0447) \rightarrow [3.0626, 3.2378]$

② 99% $3.1502 \pm 2.5758 (0.0447) \rightarrow [3.0351, 3.2653]$

Exercise 9

Suppose a student measures the boiling temperature of a certain liquid 6 times. The observed data are as follows 102.5, 101.7, 103.1, 100.9, 100.5, 102.2 Knowing that the boiling temperature is well approximated by a normal distribution and that the sd for this procedure is 1.2°.

1. Provide an unbiased estimator for the population mean together with an estimate of its standard error.
2. Compute the 95% confidence interval for the mean.
3. Compute the 99% confidence interval for the mean.

$\mu = 101.82$ Pop $X \sim N(\mu, 1.2^2)$

$\sigma = 1.2$

Sample $\bar{X} \sim N(\mu, \frac{1.2}{\sqrt{6}} = 0.49)$

$\bar{X} = 101.82$

② $101.82 \pm 1.96 \cdot (0.49)$ $[100.8596, 102.78]$

③ $101.82 \pm 2.5758 (0.49)$ $[100.538, 103.082]$

Exercise 10

A cigarette producer wants to know the tar level into his cigarettes. On a sample of 300 cigarettes, the mean level is 10.95 mg with a sample standard deviation s = 0.51 mg.

1. Compute a 95% confidence interval for the population mean.
2. Compute a 99% confidence interval for the population mean.

Pop $X \sim N(\mu, \sigma^2)$ S.E. $= \frac{0.51}{\sqrt{300}} = 0.0294$

Sample $\bar{X} - N(10.95, 0.51^2)$ $n = 300$

① $10.95 \pm 1.96 (0.0294) \rightarrow [10.87, 11.03]$

② $10.95 \pm 2.576 (0.0294) \rightarrow [10.83, 11.07]$

RECITATION 8

Exercise 1

In a sample of 402 Tor Vergata first-year students, 174 are enrolled into Statistics course.

1. Find the sample proportion.
2. Is the proportion of students enrolled into Statistics course in the population of all Tor Vergata first-year students different from 0.50 at the significance level $\alpha = 0.05$?

$n = 402$ $\bar{X} \sim N($

$\hat{p} = \frac{174}{402} = 0.433$

Pop $X \sim N(\mu, \sigma^2)$

Sample $\bar{X} \sim (0.433, s^2)$ $n = 402$

STEP 2

$$\begin{cases} H_0 : p = 0.50 \\ H_a : p \neq 0.50 \end{cases}$$

STEP-3

$$Z_{obs} = \frac{0.433 - 0.50}{\sqrt{0.5 \cdot 05}} \cdot \sqrt{402} = -2.69$$

STEP-4

$$2 \cdot P(Z > |Z_{obs}|) = 2 \cdot (1 - \Phi(2.69)) = 2 \cdot (1 - .99643) = 0.00414 < 0.05$$

STEP-5

We reject the null hypothesis

Exercise 2

Therapeutic touch (TT) practitioners claim to improve or heal many medical conditions by using their hands to manipulate a human energy field above the patient's skin. The TT practitioner was asked to identify whether his or her right or left hand was closer to the hand of the researcher. Let p denote the probability of a correct prediction by a TT practitioner. With random guessing, p=1/2. However, the TT practitioners claimed that they could do better than random guessing. They claimed that p is greater than 1/2. Over 130 trials there were 73 correct guesses.

1. State the null and alternative hypotheses for this investigation.
2. Find the sample proportion.
3. Test the null hypothesis H_0 vs the alternative H_1 at the significance level $\alpha = 0.01$.

$$\hat{p} = \frac{73}{130} = 0.561 \qquad \sqrt{\frac{0.561 \cdot \overline{0.561}}{130}} = 0.0435$$

STEP 2

$$\begin{cases} H_0: p = \frac{1}{2} \\ H_a: p > \frac{1}{2} \end{cases}$$

STEP 3

$$Z_{obs} = \frac{0.561 - 0.5}{\sqrt{\frac{0.5 \cdot 0.5}{130}}} = 1.39$$

STEP-4

$$P(Z > 1.39) = 1 - \Phi(1.39) = 0.08226 > 0.01$$

STEP-5

We do not reject the null hypothesis

Exercise 3

It has been asked "What do you think is the ideal number of children for a family to have?". The 50 females who responded had a median of 2, mean of 3.22, and standard deviation of 1.99.

1. Test the null hypothesis $H_0 : \mu = 3$ vs the alternative $H_1 : \mu > 3$ at the significance level $\alpha = 0.05$.

Pop $X \sim N(\mu, \sigma^2)$

Sample $n = 50$ $\quad \overline{X} \sim N(3.22, 1.99^2)$

$$\begin{cases} H_0: \mu = 3 \\ H_1: \mu > 3 \end{cases} \qquad \alpha = 0.05$$

$$Z_{obs} = \frac{3.22 - 3}{\frac{1.99}{\sqrt{50}}} = 0.78$$

$P(Z > 0.78) = 1 - \Phi(0.78) = 0.2177 > 0.05$ $\Rightarrow$ We don't reject the null hypothesis

Exercise 4

Suppose the mean GPA of all students graduating from Penn State University in 2005 was 3.05. The registrar plans to look at records of 100 students graduating in 2015 to see if mean GPA has changed. The sample mean is 2.15, while standard deviation is 1.5.

1. State the null and alternative hypotheses for this investigation.
2. Test the null hypothesis H_0 vs the alternative H_1 at the significance level $\alpha = 0.05$.

Pop $X \sim N(3.05, \sigma^2)$

Sample $\bar{X}(2.15, 1.5^2)$ $n=100$ $\alpha=0.05$

$$\begin{cases} H_0: \mu = 3.05 \\ H_1: \mu \neq 3.05 \end{cases}$$

$$Z = \frac{2.15-3.05}{\frac{1.5}{\sqrt{100}}} = -6 \qquad 2P(Z > |-6|) \approx 0 \Rightarrow \text{we reject the null hypothesis}$$

Exercise 5

In 2006, 25% of births were by mothers aged over than 30. This year, within a sample of 120 children, 34 of them were born by mothers aged more than 30.

1. Compute $\hat{p}$, the estimate of the proportion of births born by mothers aged more than 30 and the corresponding standard error $\hat{se}(\hat{p})$
2. With a significance of 5%, can we say that the proportion of births by mothers aged over than 30 has increased in the last years?

$X \sim Be(0.25)$

Sample $\bar{X} \sim (\mu, s^2)$ $n=120$ $\hat{p}=0.283$ $se(\hat{p}) = 0.041$ $\alpha=0.05$

$$\begin{cases} H_0: \mu = 0.25 \\ H_1: \mu > 0.25 \end{cases}$$

$$Z_{obs} = \frac{0.283 - 0.25}{\sqrt{\frac{0.25 \cdot 0.75}{120}}} = 0.835 \qquad P(Z > 0.84) = 1 - \Phi(0.84) = 0.20045 > 0.05$$

we don't reject the null hypothesis

Exercise 6

It is known that the treatment time for a given disease with a standard therapy has a normal distribution with a mean of 15 days and standard deviation of 3 days. It is claimed that a new therapy can reduce the treatment time. A sample of 70 patients is observed and their average treatment time is calculated to be 14.6 days. Assuming that the standard deviation of the old and the new treatment time is the same, test the null hypothesis $H_0 : \mu = 15$ vs the alternative $H_1 : \mu < 15$ at the significance level α=0.05.

Pop $X \sim N(15, 3^2)$

Sample $\bar{X} \sim N(14.6, \frac{3}{\sqrt{70}}^2)$

$n = 70$

$$\begin{cases} H_0: \mu = 15 \\ H_1: \mu < 15 \end{cases} \qquad \alpha = 0.05$$

$$Z_{obs} = \frac{14.6 - 15}{\frac{3}{\sqrt{70}}} = -1.1156 \qquad P(Z < -1.12) = 0.13136 > 0.05$$

don't reject

Exercise 7

The total time for manufacturing a certain component is known to have a normal distribution. In a random experiment involving the manufacturing of 50 components, the sample mean was 64.95 seconds and the sample standard deviation was 0.42. The owner of the company thinks the time for manufacturing the component is equal to 65.1 seconds. Is he wrong?

Pop $X \sim N(\mu, \sigma^2)$

Sample $\bar{X} \sim N(64.95, 0.42^2)$ $n=50$

$$\begin{cases} H_0: \mu = 65.1 \\ H_1: \mu \neq 65.1 \end{cases}$$

$$z = \frac{64.95 - 65.1}{\frac{0.42}{\sqrt{50}}} = -2.53$$

$2P(Z > 2.53) = 0.0114 < 0.05$ reject the null hypothesis

Exercise 8

A physician wants to evaluate the efficacy of a new treatment in reducing the body temperature; this variable is well approximated by a Normal distribution. For a random sample of 10 patients, the observed sample mean is $\bar{x} = 2.3$.

1. Compute the 95% confidence interval for the population mean assuming that the sample variance is $s^2 = 0.4$
2. Compute the 99% confidence interval for the population mean assuming that the sample variance is $s^2 = 0.4$
3. Test the null hypothesis $H_0 : \mu = 2.5$ vs the alternative hypothesis $H_1 : \mu < 2.5$ at the significance level $\alpha = 0.05$, assuming that the population variance is known and equal to $\sigma^2 = 0.5$
4. Test the null hypothesis $H_0 : \mu = 2.5$ vs the alternative hypothesis $H_1 : \mu \neq 2.5$ at the significance level $\alpha = 0.05$, assuming that the population variance is known and equal to $\sigma^2 = 0.5$
5. Test the null hypothesis $H_0 : \mu = 2.5$ vs the alternative hypothesis $H_1 : \mu \neq 2.5$ at the significance level $\alpha = 0.05$, assuming that the population variance is known and equal to $\sigma^2 = 0.5$ and the sample size is equal to 100

Pop $X \sim N(\mu, \sigma^2)$

Sample $\bar{X} \sim N(2.3, s^2)$ $n=10$

1. $s=0.2$ $t_{9, 0.025} = 2.262$ 95%

$2.3 \pm 2.262 \cdot 0.2 \rightarrow [2.014, 2.586]$

2. $s=0.4$ 99% $t_{9, 0.005} = 3.25$

$2.3 \pm 3.25 \cdot 0.2 \rightarrow [1.65, 2.95]$

3. $\begin{cases} H_0: \mu = 2.5 \\ H_1: \mu < 2.5 \end{cases}$ $t_{obs} = \frac{2.3 - 2.5}{\sqrt{\frac{0.5}{10}}} = -0.89$ $P(Z < -0.89) = 0.18673 > 0.05$ don't reject

4. $\begin{cases} H_0: \mu = 2.5 \\ H_1: \mu \neq 2.5 \end{cases}$ " " $2P(Z > |-0.89|) = 0.37346 > 0.05$ don't reject

5. $\begin{cases} H_0: \mu = 2.5 \\ H_1: \mu \neq 2.5 \end{cases}$ $z_{obs} = \frac{2.3 - 2.5}{\sqrt{\frac{0.5}{100}}} = -2.82$ $2P(Z > |-2.82|) = 0.0048 < 0.05$ reject

Exercise 9

For a random sample of 100 trees, the observed mean height is $\bar{x} = 11.4$ meters, with a sample standard deviation of $s = 1.3$ meters

1. Compute a 95% confidence interval for the population mean
2. Test the null hypothesis $H_0 : \mu = 11.2$ vs the alternative hypothesis $H_1 : \mu > 11.2$ at the significance level $\alpha = 0.1$
3. Test the null hypothesis $H_0 : \mu = 11.2$ vs the alternative hypothesis $H1 : \mu neq 11.2$ at the significance level $\alpha = 0.1$

Pop $X \sim N(\mu, \sigma^2)$

Sample $\bar{X} \sim N(\mu, 1.3^2)$ $n=100$

$\bar{x} = 11.4$

1. $11.4 \pm 1.96(1.3) \longrightarrow [8.852, 13.948]$

2. $\begin{cases} H_0: \mu = 11.2 \\ H_1: \mu > 11.2 \end{cases}$ $z = \frac{11.4 - 11.2}{\frac{1.3}{10}} = 1.54$ $P(Z > 1.54) = 0.06178 < 0.1$ reject

3. $\begin{cases} H_0: \mu = 11.2 \\ H_1: \mu \neq 11.2 \end{cases}$ " " $2P(Z > 1.54) = 0.12356 > 0.1$ DON'T RESECT

Exercise 10

Let X be the random variable denoting the height of an adult man. This variable is well approximated by a Normal distribution with mean $\mu = 175$ cm and variance $\sigma^2 = 42$

1. Compute the probability that the sample mean is lower than 171 cm for a random sample of size $n = 20$
2. Compute the probability that the sample mean is higher than 180 cm for a random sample of size $n = 20$
3. Compute the probability that the sample mean is in between 171 cm and 180 cm for a random sample of size $n = 20$
4. Compute the probability that in a random sample of size $n = 4$ at most two individuals are less than 173 cm tall
5. Compute the 97.5% quantile of the X variable

Pop $X \sim N(175, 42)$

① $\bar{X} \sim \left(175, \sqrt{\frac{42}{20}}^2\right)$ $n=2$

$P(Z < 171) = \Phi\left(\frac{171-175}{\sqrt{\frac{42}{20}}}\right) = \Phi(-2.76) =$

$= 0.00289$

② $P(Z > 180) = 1 - P\left(Z < \frac{180-175}{1.449}\right) = 0.00028$

③ $P(171 < X < 180) = .99972 - 0.00289 = .99683$

④ $\bar{X} \sim \left(175, \frac{6.48}{2} = 3.24^2\right)$ $n=4$

$P(X < 173) = \Phi\left(\frac{173-175}{\sqrt{42}}\right) = \Phi(-0.31) = 0.37828$

$P(\text{At most 2 with } h < 173) = \frac{4 \cdot 3 \cdot 2!}{2! \; 2!} 0.37828^2 \cdot (1-0.37828)^2 + \frac{4!}{0! \; 4!}(1-0.37828)^4 + \frac{4 \cdot 3!}{1! \; 3!} 0.37828^1 \cdot (1-0.37828)^3$

$= 0.8449$

⑤ $Z_{97.5\%} = 1.96$ $\quad X = 1.96 \cdot 6.48 + 175 = 187.7$

Exercise 11

The Italian police has reported that 44% of murder victims are killed with handguns. If 4 murder victims are randomly selected, find

1. The probability that they were all killed by handguns
2. The probability that none were killed by handguns
3. The probability that at least two were killed by handguns
4. The probability that more than 20 murder victims were killed by handguns in a sample of size $n = 50$

Pop $X \sim Ber(0.44)$

① $P(\bar{X}=4) = 0.44^4 = 0.0375$

② $P(\bar{X}=0) = (1-0.44)^4 = 0.0983$

③ $P(X \geq 2) = \frac{4 \cdot 3 \cdot 2!}{2! \; 2!} 0.44^2 (1-.44)^2 + \frac{4 \cdot 3!}{3!} 0.44^3 (1-.44) + \frac{4!}{4! \; 0!} 0.44^4 = 0.59$

④ $\bar{X} \sim N(22, 3.51^2)$

$\mu = p \cdot n = 22$

$\sigma = \sqrt{p \cdot q \cdot n} = 3.51$

$P(Z > 20) = 1 - \Phi\left(\frac{20-22}{3.51}\right) = 1 - \Phi(-0.59) = 0.71566$

RECITATION 9

1. Suppose that for n = 6 students, x =exam 1 score and y = exam 2 score. The regression line is

$$\hat{y} = 20 + 0.8x$$

a. For each individual observation, calculate the predicted value $\hat{y}_i$ and the residual e_i. Show the results in the chart below.

x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y}_i$	76	80	84	84	88	92
e_i	1	2	-4	2	2	-1

$\longrightarrow \bar{y}$

b. Calculate SSE, the sum of square errors.

c. Calculate the sample estimate of the variance σ^2 for the regression model.

d. Calculate the value that would be given in R for "S = ". Write a sentence that interprets this value.

e. Calculate the value of R^2. To start, you'll have to calculate the value of SST. Write a sentence that interprets the value of R^2.

(b) $SSE = \sum e_i^2 = 30$

(c) $E(MSE) = \sigma^2 = \frac{SSE}{n-2} = \frac{30}{6-2} = 7.5$

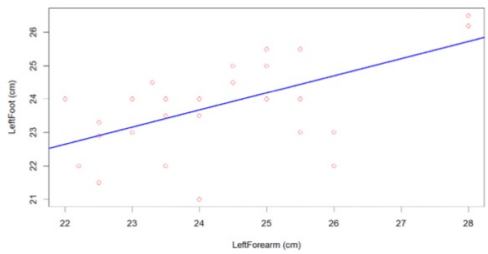
(d) $S = \sqrt{MSE} = 2.74$

(e) $SST = \sum (y_i - \bar{y})^2 = 81 + 4 + 16 + 4 + 36 + 49 = 190$

$\bar{y} = 84$

$R^2 = \frac{SST - SSE}{SST} = 0.84$

3. Data from n = 30 female college students are used to analyze a straight-line relationship between y = left foot length (cm) and x = left forearm length (cm). A plot of the data with the regression line superimposed follows.



a. R output for tests of hypothesis about the intercept and slope is as follows:

```
Call:
lm(formula = LeftFoot ~ LeftArm, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7030 -0.3819  0.3246  0.8315  1.3522

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.3446     3.3560   3.380 0.002149 **
LeftArm       0.5138     0.1377   3.733 0.000857 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.149 on 28 degrees of freedom
Multiple R-squared:  0.3322,    Adjusted R-squared:  0.3084
F-statistic: 13.93 on 1 and 28 DF,  p-value: 0.0008567
```

What is the **p-value** for testing the null hypothesis that the population slope = 0. On the basis of this p-value what can we conclude about the population slope?

b. Refer to the information given in part a. The value of the t-statistic = 3.73 for testing the null hypothesis that the population slope. Show how this t-statistic is calculated using other values given in the output.

$n = 30$

(a) p-value$_{\beta_1}$ = 0.000857 < 0.05 we reject the hypothesis $\beta_1 = 0$

(b) It is given by $\frac{\text{Estimate}}{\text{Std. Error}} = \frac{0.5138}{0.1377}$

(c) $b_1 \pm t_{(1-\alpha)}(n-2) \underbrace{\frac{\sqrt{MSE}}{\sqrt{\sum (x_i - \bar{x})^2}}}_{s(b_1)}$

$0.5138 \pm 2.048 \cdot 0.1377 \longrightarrow [0.2318, 0.7958]$

We reject the null hypothesis because the 95% C.I. doesn't contain 0

c. Refer to the information given in part a. Calculate a 95% confidence interval that estimates the unknown value of the population slope. (Recall that n = 30.)

How do you interpret this confidence interval in the context of these data?

What conclusion can you reach, based on this confidence interval, to test

$$H_0 : \beta_1 = 0 \text{ v.s.} H_1 : \beta_1 \neq 0$$

d. For the left foot and left forearm data, SSTO = 55.34 and SSE = 36.95. Complete the following table with numerical values.

Source	DF (degrees of freedom)	SS (Sum of Squares)	MS (Mean square)	F-value
Regression	1	18.30	18.30	13.86739
Error	28	36.95	1.32	
Total	29	55.34		

e. Calculate the value of R^2. Then, write a sentence that interprets that value in the context of these data.

f. What is the p-value of the F-test (Hint: based on R output in part a)? What conclusion do you reach from this p-value?

ⓓ

$SST = SSE + SSR$

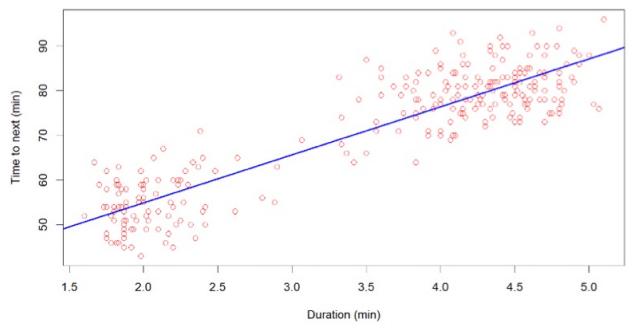
$SSR = SST - SSE$

ⓔ $R^2 = \frac{SSR}{SST} = 0.33$

ⓕ 0.000857 : it is the same as the t-test for the slop (happen when $b_{no-x} = b_1$) and since is very small it denote that there is high relation between Left forearm and Left foot

2. Data from n = 227 eruptions of the Old Faithful geyser in the United States are used to study the straight-line relationship between y = time to next eruption (min.) and x = duration of the present eruption (min.). Following is a plot of the data with the regression line superimposed.

$n = 227$




```
Call:
lm(formula = waiting ~ eruptions, data = faithful)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0796  -4.4831   0.2122   3.9246  15.9719

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
eruptions    10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

a. What is the value of the slope of the regression line? Write a sentence that interprets this value in the context of this situation.

ⓐ 10.7296

For every unit that x is increasing the y increases of 10.7296

b. Following is computer output that gives a 95% confidence interval for the mean value of y and a 95% prediction interval for y when duration of the present eruption = 4.5 minutes.

```
> predict(out, newdata, interval="predict",se.fit = TRUE)
$fit
       fit       lwr       upr
1 81.75778 70.07611 93.43946

$se.fit
[1] 0.4796786

> predict(out, newdata, interval="confidence",se.fit = TRUE)
$fit
       fit     lwr      upr
1 81.75778 80.8134 82.70217

$se.fit
[1] 0.4796786
```

Prediction = $33.4744 + 10.7296 \cdot 4.5 = 81.76$

Show how the value under "Fit" (81.76) was calculated.

c. Refer to the output in part b. Write a sentence (or two) that interprets the interval given under "95% CI." Can you show how the confidence interval is calculated? (Suppose t*= 1.97)

d. Refer to the output in part b. Write a sentence (or two) that interprets the interval given under "95% PI." Can you show the formula to get the prediction interval? (Suppose t*= 1.97, MSE= 34.9754).

e. Is it appropriate to construct a confidence interval for mean value of y or prediction interval for y when duration of the present eruption = 7 minutes? Why?

ⓒ 95 % C.I.

$81.75778 \pm 1.97 \cdot 0.4496786 \longrightarrow [80.3128, 82.7027]$

ⓓ 95% P.I.

$81.75778 \pm 1.97 \sqrt{0.4496786^2 + 34.9754} \longrightarrow [70.06894, 93.44662]$

ⓔ No, since we don't know the behaviour of the regression line outside the model

RECITATION 10

2. Data from n = 113 hospitals are used to evaluate factors related to the risk that patients get an infection while in the hospital. The y-variable = infection risk, the percentage of patients who get an infection while hospitalized. The x-variables are Stay = average length of stay, Age = average patient age, Xrays = measure of how many Xrays are done in the hospital, and Services = a measure of how many different services the hospital offers. Here's the R output.

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.97831 -0.77113 -0.09527  0.67732  2.52121 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.170874   1.285745   0.133 0.894521    
Stay         0.237209   0.060957   3.891 0.000173 ***
Age         -0.014071   0.022708  -0.620 0.536789    
Xrays        0.020383   0.005524   3.690 0.000354 ***
Services     0.022718   0.006970   3.260 0.001493 ** 
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.04 on 108 degrees of freedom
Multiple R-squared:   0.42,	Adjusted R-squared:  0.3986 
F-statistic: 19.56 on 4 and 108 DF,  p-value: 3.96e-12
```

```
Analysis of Variance Table

Response: InfctRsk
           Df  Sum Sq Mean Sq F value    Pr(>F)    
Stay        1  57.305  57.305 52.9915 5.693e-11 ***
Age         1   2.075   2.075  1.9189 0.1688367    
Xrays       1  13.719  13.719 12.6861 0.0005495 ***
Services    1  11.489  11.489 10.6246 0.0014928 ** 
Residuals 108 116.791   1.081                      
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR = 84.588$

$MSR = \frac{SSR}{p-1=4} = 21.147$

$SSE = 116.791$

$MSE = \frac{SSE}{n-p=108} = 1.081$

$SST = 201$

a. What is the value of the sample coefficient of the "Stay" predictor variable? Write a sentence that interprets this value, in terms of how increasing length of stay affects infection risk in a hospital.

b. The *p*-value given for the variable Services is 0.001. What null and alternative hypotheses are being evaluated with this *p*-value? What conclusion should we make about the variable Services?

c. Can we claim that one more day of average stay among patients will be associated with higher change in infection risk than one more X-ray for hospitals?

d. In the analysis of variance table, the value under "P" is 0.000. This is a p-value. What null and alternative hypotheses are being evaluated with this p-value? On the basis of this p-value, what conclusion can we make?

e. Using values given in the analysis of variance table, calculate R^2 for the model. Then, write a sentence that interprets this value in the context of this problem.

f. Can you calculate a **95% Confidence interval** for mean infection risk for hospitals with average length of patients' stay 12, average patient age 50, numbers of Xrays are done in the hospital 80, and 40 different services are offered? (Suppose standard error of the fit is 0.2119, t*=1.98).

g. Can you calculate a **95% Prediction interval** for mean infection risk for a hospital with average length of patients' stay 12, average patient age 50, numbers of Xrays are done in the hospital 80, and 40 different services are offered? (Suppose standard error of the fit is 0.2119, t*=1.98).

(a) 0.237209

When, Age, Xrays, Services are held constant the infection risk(y) increase by 0.237209 every time "Stay" increase by one unit.

(b) the p-value for the variable Services is $0.001 < 0.05$ so we reject the null hypothesis that we can say that Services is significant

(c) ?

(d) $H_0: \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$ / H_a: At least one $\beta_i \neq 0$ for $i = 1, \dots, p-1$

The null hypothesis means that y is not related to none of the x-variables in the model

Since the p-value is under 0.000 -> 0.000... < 0.05 we can reject the null hypothesis

We can conclude that there is a least one of the x-variables, that is significant

(e) $R^2 = \frac{SSR}{SST} = 0.42$

It means that the four x-variables explains 42% of the observed variation of the hospital infection risk

(f) STAY = 12 x_1

AGE = 50 x_2

X RAYS = 80 x_3

SERVICES = 40 x_4

95% C.I.

$\hat{y} = 0.171 + 0.237 \cdot 12 + (-0.0141) \cdot 50 + 0.02 \cdot 80 + 0.0227 \cdot 40 = 4.818$

$4.818 \pm 1.98 \cdot 0.2119 \longrightarrow [4.398, 5.238]$

(g) 95% P.I.

$4.818 \pm 1.98 \sqrt{0.2119^2 + 1.081} \longrightarrow [2.717, 6.919]$

3. Suppose that a multiple regression is done in which a y-variable is predicted using a linear model with ***five predictor variables***. There are n = 76 observations in the dataset. SSTO = 1200 and SSE = 630.

a. Determine the value of MSE, the mean square error, in this situation.

b. Determine the value of MSR, the mean square for the regression, in this situation.

c. Calculate the value of the F statistic that would be used to test the null hypothesis that all coefficients multiplying x-variables in the equation have the value 0.

(a) $MSE = \frac{SSE}{n-p} = \frac{630}{76-6} = 9$

(b) $SSR = 1200 - 630 = 570$

$MSR = \frac{570}{6-1=5} = 114$

(c) $F = \frac{MSR}{MSE} = 12.67$

4. Data from 55 college student is used to estimate a multiple regression model with y = left forearm length, x1= left foot length, and x2 = right foot length. All variables were measured in centimeters. Some R output follows.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.8288 -1.1713 -0.3654  0.8935  5.5785

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7104     2.5179   4.651 2.31e-05 ***
LeftFoot      0.3519     0.2961   1.188    0.240
RtFoot        0.1850     0.2816   0.657    0.514
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.796 on 52 degrees of freedom
Multiple R-squared:  0.3688,    Adjusted R-squared:  0.3445
F-statistic: 15.19 on 2 and 52 DF,  p-value: 6.382e-06
```

```
Analysis of Variance Table

Response: LeftArm
          Df  Sum Sq Mean Sq F value    Pr(>F)
LeftFoot   1  96.627  96.627 29.9466 1.293e-06 ***
RtFoot     1   1.393   1.393  0.4316    0.5141
Residuals 52 167.785   3.227
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> vif(out)
LeftFoot   RtFoot
9.156082 9.156082
```

[A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2}$$

where:

R_i^2 = Unadjusted coefficient of determination for regressing the ith independent variable on the remaining ones

In general terms,

- VIF equal to 1 = variables are not correlated
- VIF between 1 and 5 = variables are moderately correlated
- VIF greater than 5 = variables are highly correlated]

a. What does the *p*-value given in the analysis of variance indicate about the model in this situation? Briefly explain.

b. Interpret the results of the individual t-tests for the two x-variables. Then, explain why these results do not contradict the result of the analysis of variance *F* test.

c. What are VIF for both variables, and what are does this value mean?

d. Give some comments on the following output.

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.0316 -1.1927 -0.3893  0.7362  5.6462

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.76796    2.50285   4.702 1.88e-05 ***
LeftFoot     0.53557    0.09734   5.502 1.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.787 on 53 degrees of freedom
Multiple R-squared:  0.3635,    Adjusted R-squared:  0.3515
F-statistic: 30.27 on 1 and 53 DF,  p-value: 1.109e-06
```

(a) Since the p-value is very low we know that there is at least x-variable that is significant.

(b)

The conclusion for each t-test is that the variable is not a useful predictor given the presence of the other variable in the model. Combined with the result of the F test (part a), this means that we only need one foot length, not both, to adequately predict forearm length.

(c) Since the VIF is greater than 5 for both x-variable it indicates an high level of multicollinearity which can leads to a bad interpretation of the graph

(d) By removing Right foot we can notice the increase of the significant level