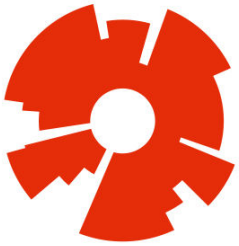


HÁSKÓLINN Í REYKJAVÍK
REYKJAVIK UNIVERSITY

Computer Science Department

Stats 1

Tölfræði 1

Workbook

Lilja Ýr Guðmundsdóttir
Email: liljag18@ru.is

Contents

Introduction

This is my workbook for Statistics class containing any notes and work I've done over the semester. There are 118 problems and examples solved here (even though the workbook turn in guideline said there were only 117, might have made a counting error). There are also some that aren't correct but I learned from them.

1 Week 1

1.1 Concepts

We calculate the mean, median and standard deviation, easiest to just use R for this.
Example: The Statistics course has 220 students: 88 girls and 132 boys.
The problem solving classes are Tuesday and Wednesday.
On Tuesday 44 students are expected: 11 girls and 33 boys
On Wednesday 176 are expected: 77 girls and 99 boys.
A student name is randomly picked from the list of all registered students.
What is the probability that the randomly selected student is a girl? -> 0.4
What is the probability that the randomly selected student is expected Wednesday? -> 0.8
What is the probability that the student is a girl who is expected on Wednesday? -> 0.35
What is the probability that the student is a girl, or a student expected on Tuesday? -> 0.55

1.2 R Code

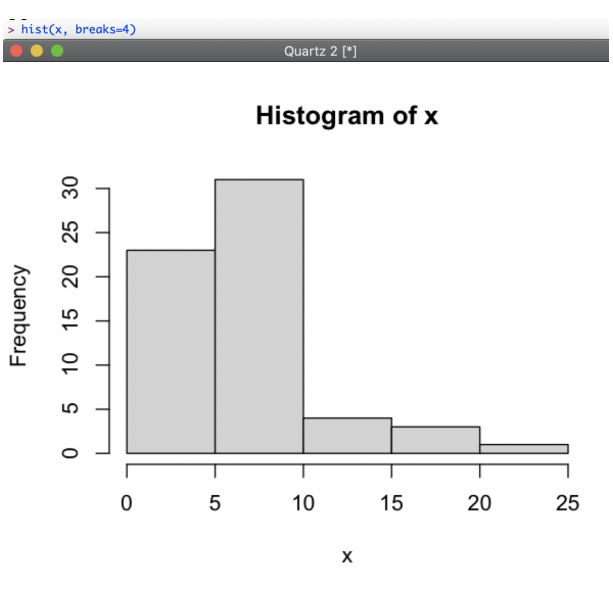
```
getwd()
etwd("/Users/liljayrgudmundsdottir/Documents/HR_Files/5onn/Tolfraedi/datas
data <- read.csv(file = 'table1-2.csv')
x = data$PM
median(x)
mean(x)
sd(x)
```

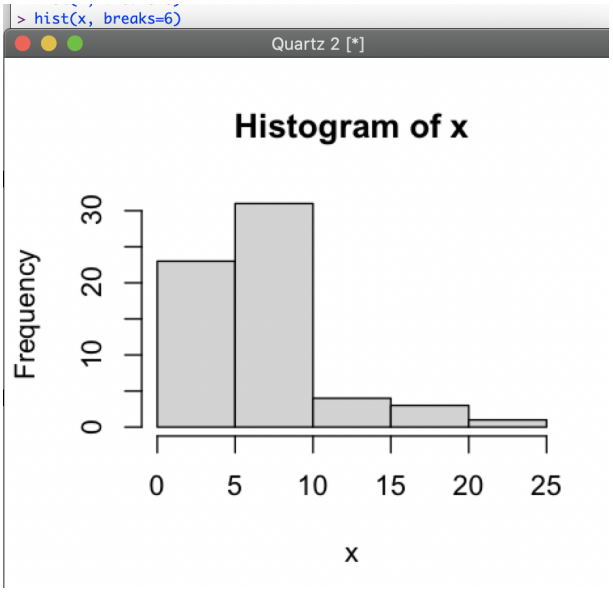
1.3 Example Problem Set

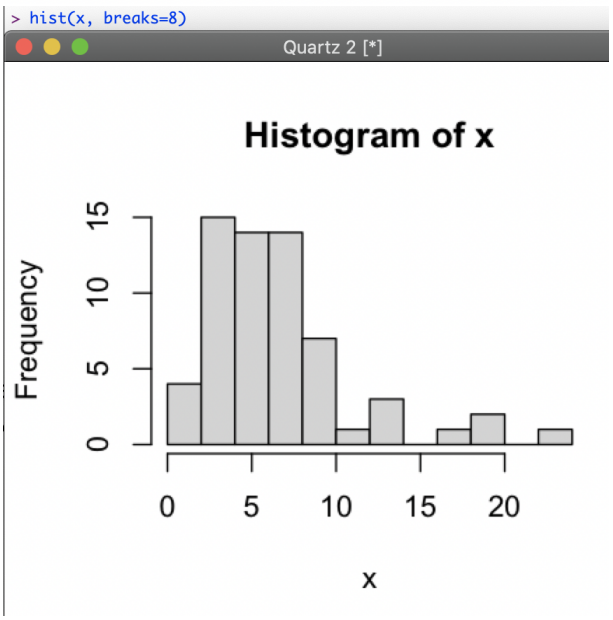
Datasets and software: Download the datasets from Canvas the online version of the book and use Ch1\table1-2.csv to calculate the median, mean value, and standard deviation. Make a histogram and a boxplot. Use Excel or R. Use 4, 6, and 8 bins and compare.

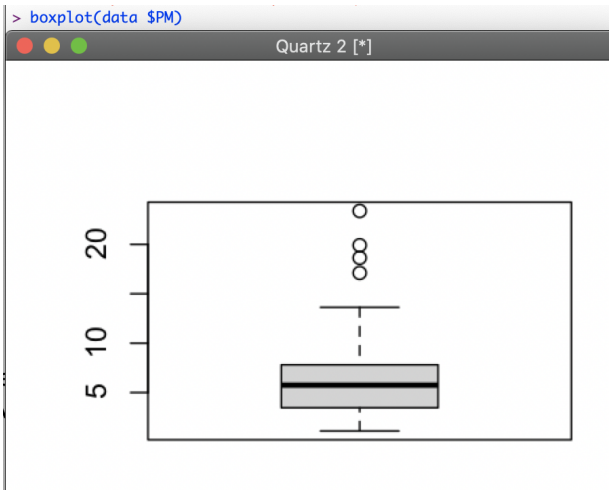
```
> getwd()
[1] "/Users/liljayrgudmundsdottir"
> setwd("/Users/liljayrgudmundsdottir/Documents/HR_Files/5onn/Tolfraedi")
> getwd()
[1] "/Users/liljayrgudmundsdottir/Documents/HR_Files/5onn/Tolfraedi"
> setwd("/Users/liljayrgudmundsdottir/Documents/HR_Files/5onn/Tolfraedi/datasets/Ch1")
> getwd()
[1] "/Users/liljayrgudmundsdottir/Documents/HR_Files/5onn/Tolfraedi/datasets/Ch1"
> data <- read.csv(file =table1-2.csv)
Error: unexpected symbol in "data <- read.csv(file =table1-2.csv"
> data <- read.csv(file ='table1-2.csv')
```

```
> x = data$PM
> x
 [1]  7.59  6.28  6.07  5.23  5.54  3.46  2.44  3.01 13.63 13.02 23.38  9.24  3.22  2.06
[15]  4.04 17.11 12.26 19.91  8.50  7.81  7.18  6.95 18.64  7.10  6.04  5.66  8.86  4.40
[29]  3.57  4.35  3.84  2.37  3.81  5.32  5.84  2.89  4.68  1.85  9.14  8.67  9.52  2.68
[43] 10.14  9.20  7.31  2.09  6.32  6.53  6.32  2.01  5.91  5.60  5.61  1.50  6.46  5.29
[57]  5.64  2.07  1.11  3.32  1.83  7.56
> median(x)
[1] 5.75
> mean(x)
[1] 6.596452
> sd(x)
[1] 4.518998
```









Data summary:

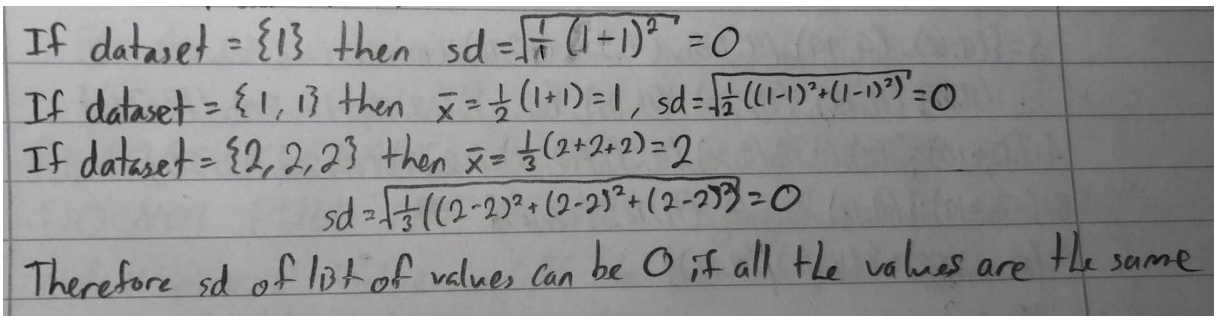
5. Find a sample size for which the median will always equal one of the values in the sample.

When n = 1 then the median will always be equal to one value in the dataset. This is true for all odd number amounts of data points.

6. For a list of positive numbers, is it possible for the standard deviation to be greater than the mean? If so, give an example. If not, explain why not.

Yes, depending on the data set, i.e (1,2,3,16) gives a mean of 5.5 and a standard deviation of 7.04. Also works for spread values.

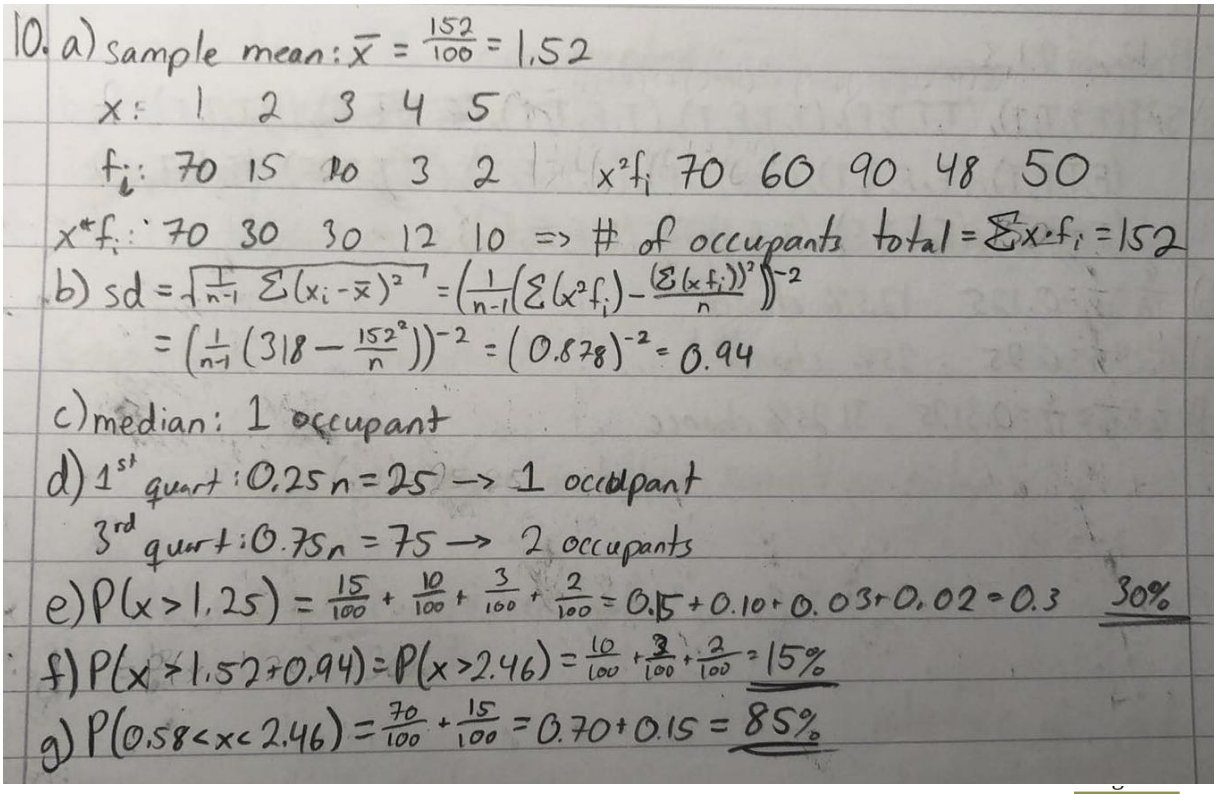
7. Is it possible for the standard deviation of a list of numbers to equal 0? If so, give an example. If not, explain why not.



Problem 1.2.10 page 24 Edition 5: A sample of 100 cars driving on a freeway during a morning commute was drawn, and the number of occupants in each car was recorded. The results were as follows:

Occupants	1	2	3	4	5
Number of Cars	70	15	10	3	2

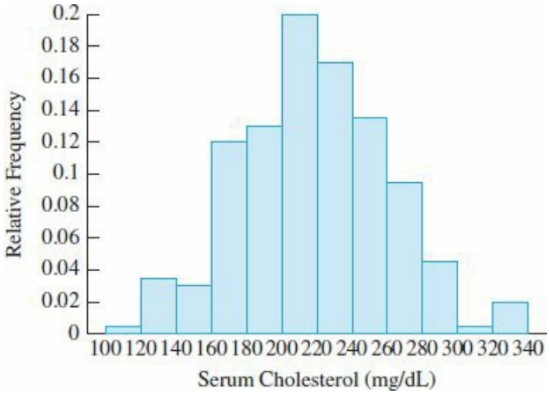
a. Find the sample mean number of occupants.
b. Find the sample standard deviation of the number of occupants.
c. Find the sample median number of occupants.
d. Compute the first and third quartiles of the number of occupants.
e. What proportion of cars had more than the mean number of occupants?
f. For what proportion of cars was the number of occupants more than one standard deviation greater than the mean?
g. For what proportion of cars was the number of occupants within one standard deviation of the mean?

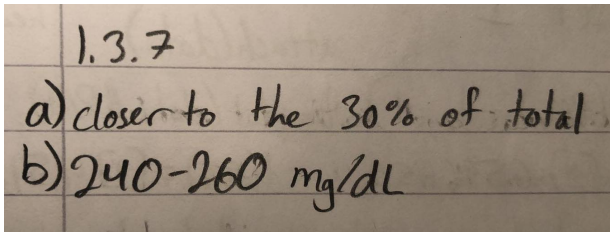


7. The figure below is a histogram showing the distribution of serum cholesterol level for a sample of men. Use the histogram to answer the following questions:

a. Is the percentage of men with cholesterol levels above 240 mg/dL closest to 30%, 50%, or 70%?

b. In which interval are there more men: 240–260 mg/dL or 280–340 mg/dL?





Standard deviation:

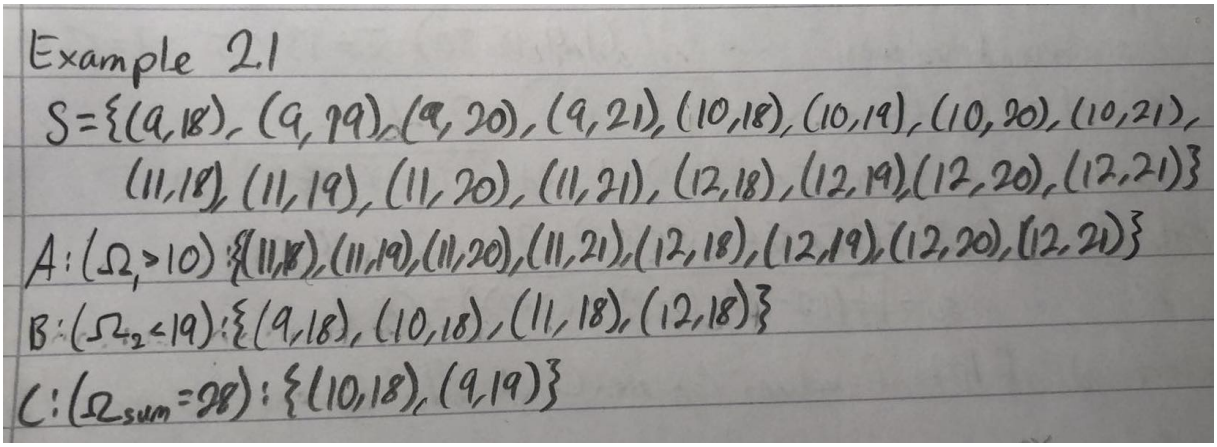
Prove Eq.1.3 using Eq.1.2 at page 15. What do you obtain if n » 1 ?
Couldn't quite get this one, didn't make sense.

Sample space:

Example 2.1

An electrical engineer has on hand two boxes of resistors, with four resistors in each box. The resistors in the first box are labeled 10 Ω (ohms), but in fact their resistances are 9, 10, 11, and 12 Ω. The resistors in the second box are labeled 20 Ω, but in fact their resistances are 18, 19, 20, and 21 Ω. The engineer chooses one resistor from each box and determines the resistance of each.

Let *A* be the event that the first resistor has a resistance greater than 10, let *B* be the event that the second resistor has a resistance less than 19, and let *C* be the event that the sum of the resistances is equal to 28. Find a sample space for this experiment, and specify the subsets corresponding to the events *A*, *B*, and *C*.



3. A section of an exam contains four True-False questions. A completed exam paper is selected at random, and the four answers are recorded.

a. List all 16 outcomes in the sample space.

b. Assuming the outcomes to be equally likely, find the probability that all the answers are the same.

c. Assuming the outcomes to be equally likely, find the probability that exactly one of the four answers is “True.”

d. Assuming the outcomes to be equally likely, find the probability that at most one of the four answers is “True.”

Problem 2.1.3.

a) $S = \{(T,T,T,T), (T,T,T,F), (T,T,F,T), (T,F,T,T), (F,T,T,T), (F,T,T,F),$
$(F,T,F,T), (F,F,T,T), (F,F,T,F), (F,F,F,T), (F,T,F,F), (T,F,F,F)$
$(T,T,F,F), (T,F,T,F), (T,F,F,T), (F,F,F,F)\}$

b) $\frac{2}{16} = \frac{1}{8} = 0.125$ 12.5% chance

c) $\frac{4}{16} = \frac{1}{4} = 0.25$ 25% chance

d) $\frac{4}{16} + \frac{1}{16} = \frac{5}{16} = 0.3125$ 31.25% chance

1.4 Problem Set

Lilja Ýr Guðmundsdóttir
23 August, 2020

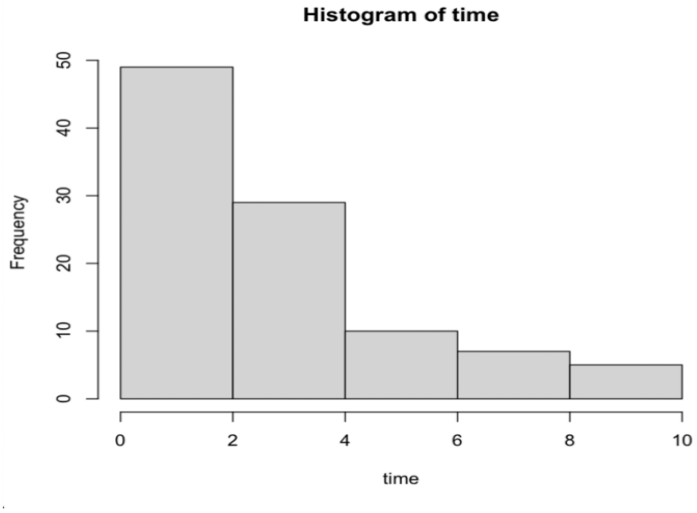
Verkefni 1

I used R to find the correct values (see code on the next page for how).

Results:

	Result
Mean	2.809726
Standard Deviation	2.35351

Histogram with 5 bins:



Lilja Ýr Guðmundsdóttir
23 August, 2020

Code (after navigating to right directory):

```
> data= read.table('dataC.csv', header=T)
> attach(data)
> mean(time)
[1] 2.809726
> sd(time)
[1] 2.35351
> hist(time, breaks=5)
```

2 Week 2

2.1 Concepts

- $P(\overline{A}) = 1 - P(A)$
- $P(A)P(B) = P(A \cup B)$
- $P(A \cap B) = P(A) + P(B) - P(A \cup B)$
- $P(A \cap B) = P(A) + P(B)$ if A and B are independent
- $P(A)P(B) = P(A \cap B)$ if A and B are independent
- Baye's rule: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- $P(A|B)$ means A given B
- Suppose P(A|B)=P(A) ($\neq 0 or 1$) then In this case P(B|A) = P(B)
- If P(A|B)=P(A) then the events A and B are independent
- A disease test can be 99% accurate, but still a large proportion of people tested positive are actually disease-free. This is possible when the disease is rare.
- A parent and an offspring (daughter or son) have 50% DNA in common
- Suppose both parents have genotype (a,b) at a specific location on the genome. The probability for their child to have (b,b) is 0.25

2.2 Example Problem Set

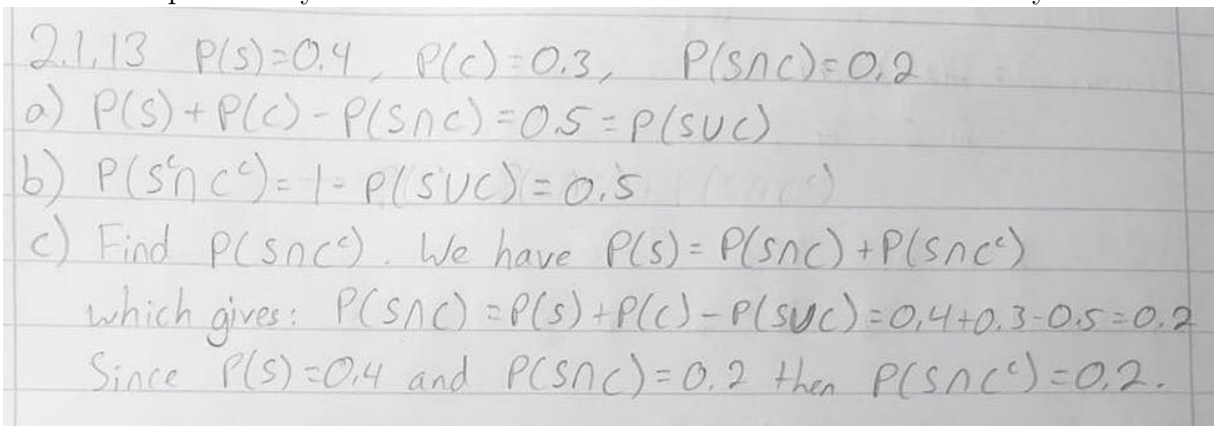
Combined events

2.1.13 Let S be the event that a randomly selected college student has taken a statistics course, and let C be the event that the same student has taken a chemistry course. Suppose $P(S) = 0.4$, $P(C) = 0.3$, and $P(S \cap C) = 0.2$.

a. Find the probability that a student has taken statistics, chemistry, or both.

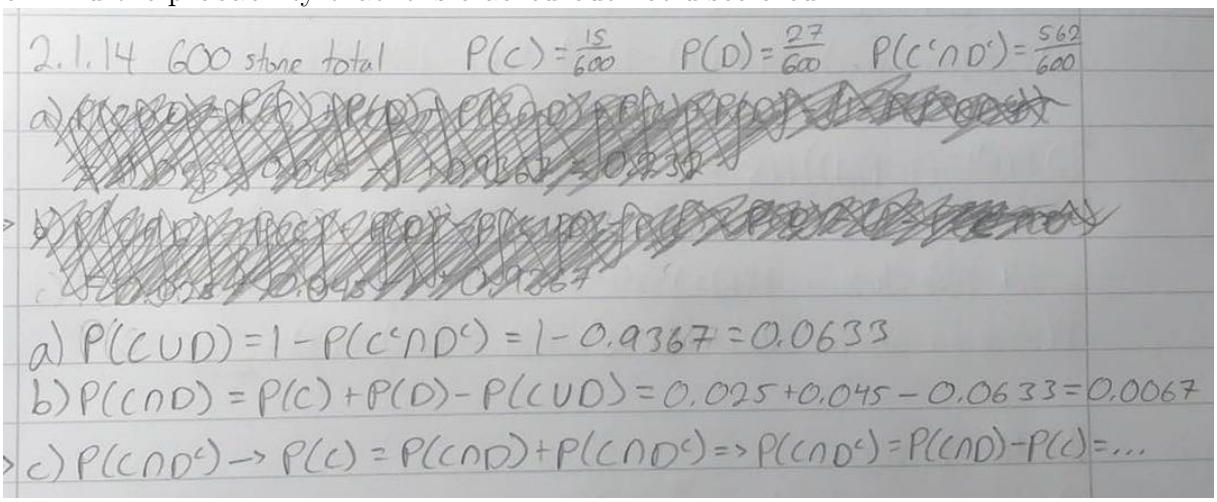
b. Find the probability that a student has taken neither statistics nor chemistry.

c. Find the probability that a student has taken statistics but not chemistry.

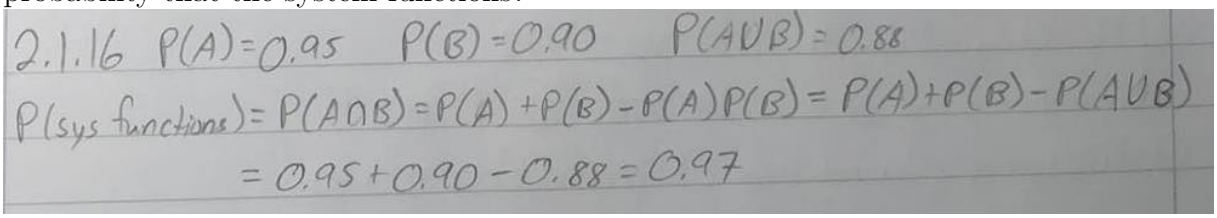


2.1.14 Six hundred paving stones were examined for cracks, and 15 were found to be cracked. The same 600 stones were then examined for discoloration, and 27 were found to be discolored. A total of 562 stones were neither cracked nor discolored. One of the

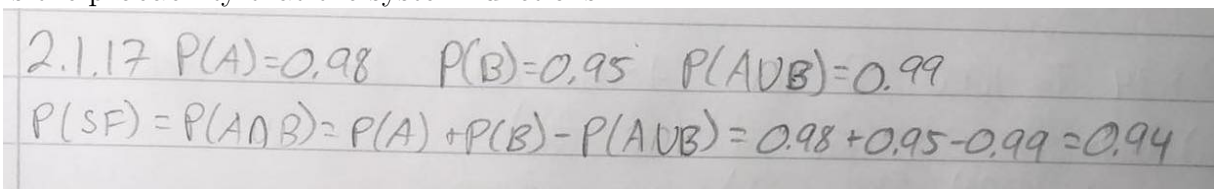
600 stones is selected at random.
a. Find the probability that it is cracked, discolored, or both.
b. Find the probability that it is both cracked and discolored.
c. Find the probability that it is cracked but not discolored.



2.1.16 A system contains two components, A and B. The system will function so long as either A or B functions. The probability that A functions is 0.95, the probability that B functions is 0.90, and the probability that both function is 0.88. What is the probability that the system functions?



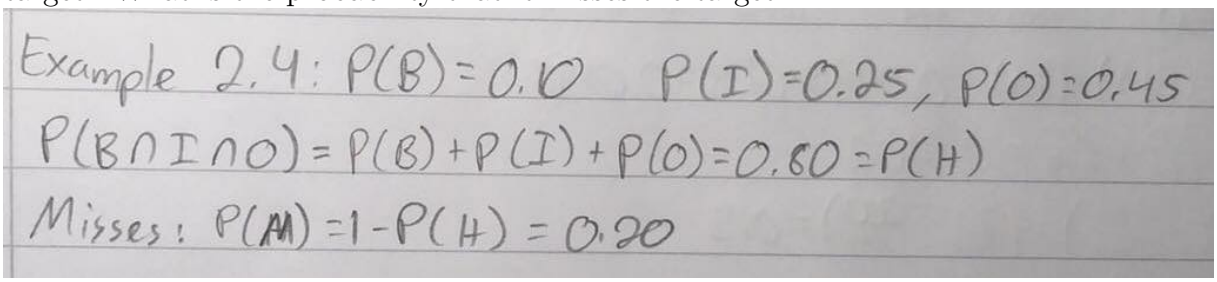
2.1.17 A system contains two components, A and B. The system will function only if both components function. The probability that A functions is 0.98, the probability that B functions is 0.95, and the probability that either A or B functions is 0.99. What is the probability that the system functions?



Combined events and conditional probabilities

Example 2.4 A target on a test firing range consists of a bull's-eye with two concentric rings around it. A projectile is fired at the target. The probability that it hits the bull's-eye is 0.10, the probability that it hits the inner ring is 0.25, and the probability

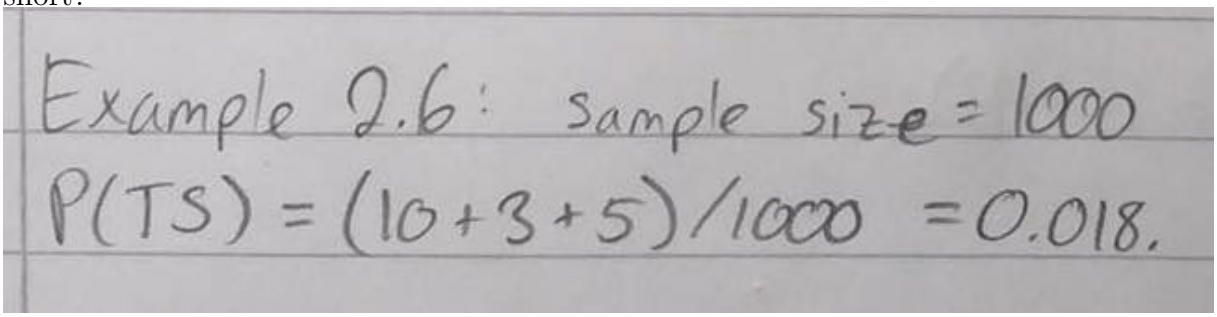
that it hits the outer ring is 0.45. What is the probability that the projectile hits the target? What is the probability that it misses the target?



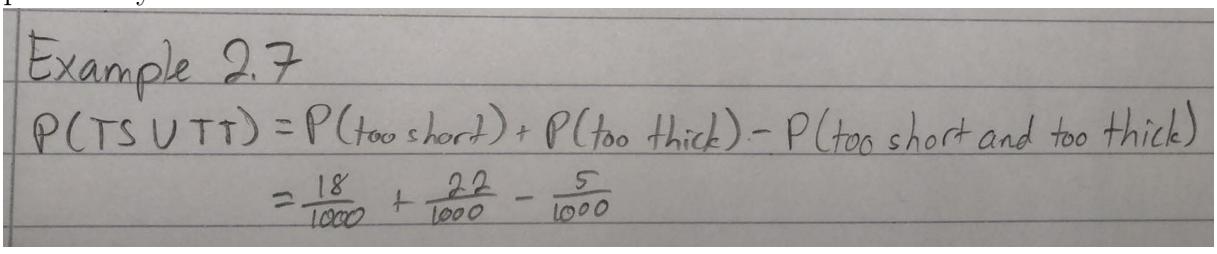
Example 2.6 An extrusion die is used to produce aluminum rods. Specifications are given for the length and the diameter of the rods. For each rod, the length is classified as too short, too long, or OK, and the diameter is classified as too thin, too thick, or OK. In a population of 1000 rods, the number of rods in each class is as follows:

	Diameter		
Length	Too thin	OK	Too thick
Too Short	10	3	5
OK	38	900	4
Too long	2	25	13

A rod is sampled at random from this population. What is the probability that it is too short?

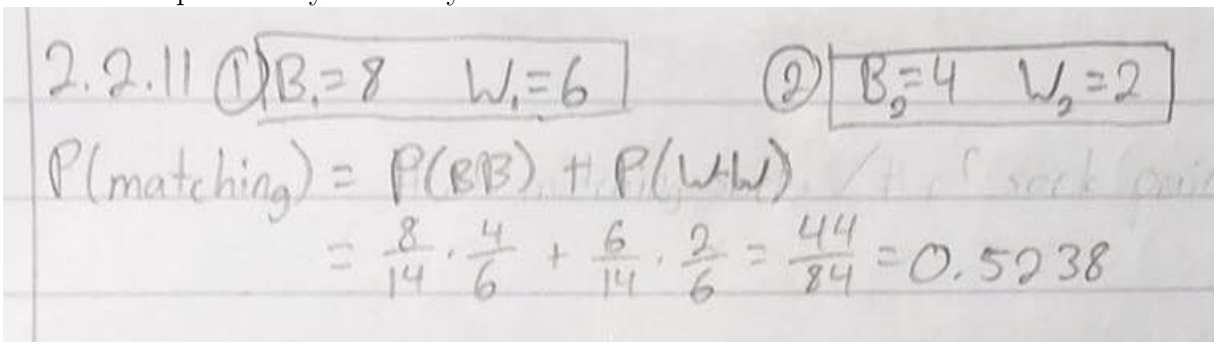


Example 2.7 Refer to Example 2.6. If a rod is sampled at random, what is the probability that it is either too short or too thick?

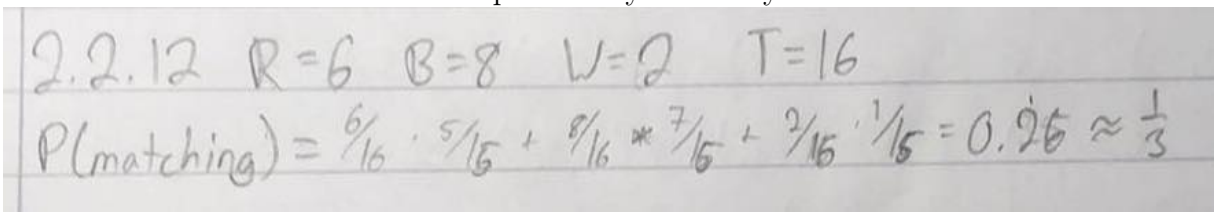


2.2.11 One drawer in a dresser contains 8 blue socks and 6 white socks. A second drawer contains 4 blue socks and 2 white socks. One sock is chosen from each drawer.

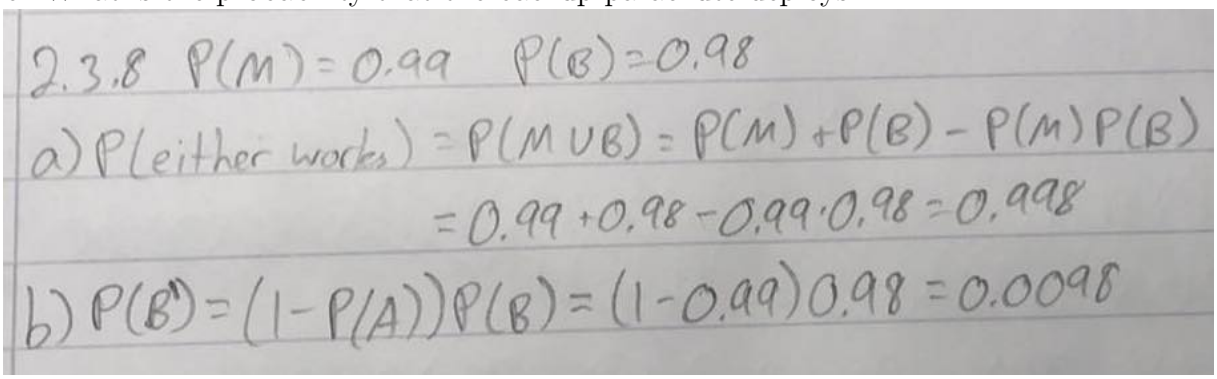
What is the probability that they match?



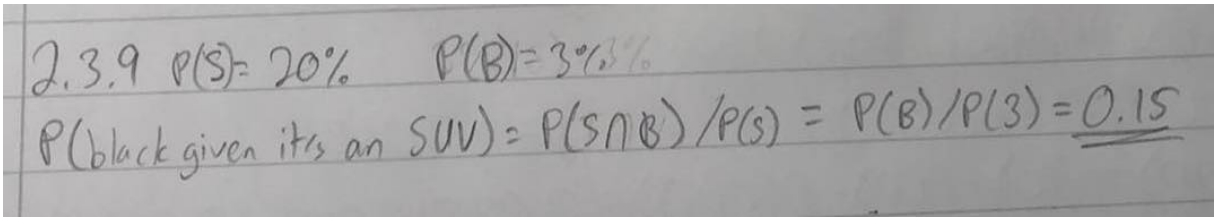
2.2.12 A drawer contains 6 red socks, 4 green socks, and 2 black socks. Two socks are chosen at random. What is the probability that they match?



2.3.8 A drag racer has two parachutes, a main and a backup, that are designed to bring the vehicle to a stop after the end of a run. Suppose that the main chute deploys with probability 0.99, and that if the main fails to deploy, the backup deploys with probability 0.98.
a. What is the probability that one of the two parachutes deploys?
b. What is the probability that the backup parachute deploys?



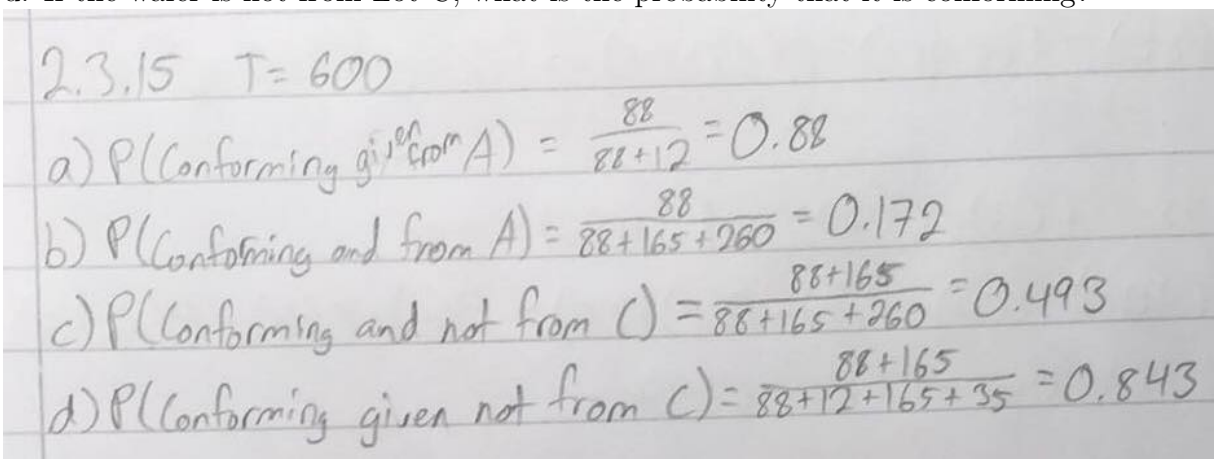
2.3.9 At a certain car dealership, 20% of customers who bought a new vehicle bought an SUV, and 3% of them bought a black SUV. Given that a customer bought an SUV, what is the probability that it was black?



2.3.15 A population of 600 semiconductor wafers contains wafers from three lots. The wafers are categorized by lot and by whether they conform to a thickness specification. The following table presents the number of wafers in each category. A wafer is chosen at random from the population.

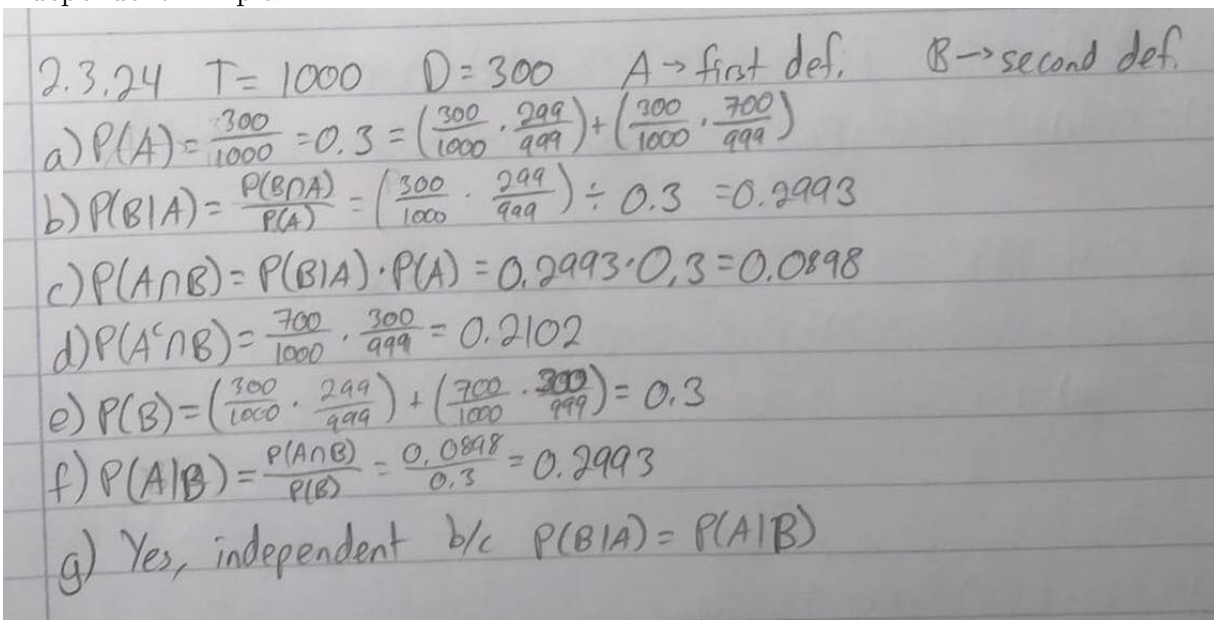
Lot	Conforming	Nonconforming
A	88	12
B	165	35
C	260	40

a. If the wafer is from Lot A, what is the probability that it is conforming?
b. If the wafer is conforming, what is the probability that it is from Lot A?
c. If the wafer is conforming, what is the probability that it is not from Lot C?
d. If the wafer is not from Lot C, what is the probability that it is conforming?



2.3.24 A lot of 1000 components contains 300 that are defective. Two components are drawn at random and tested. Let A be the event that the first component drawn is defective, and let B be the event that the second component drawn is defective.
a. Find $P(A)$.
b. Find $P(B|A)$.
c. Find $P(A \cap B)$.
d. Find $P(A^c \cap B)$.
e. Find $P(B)$.
f. Find $P(A|B)$.

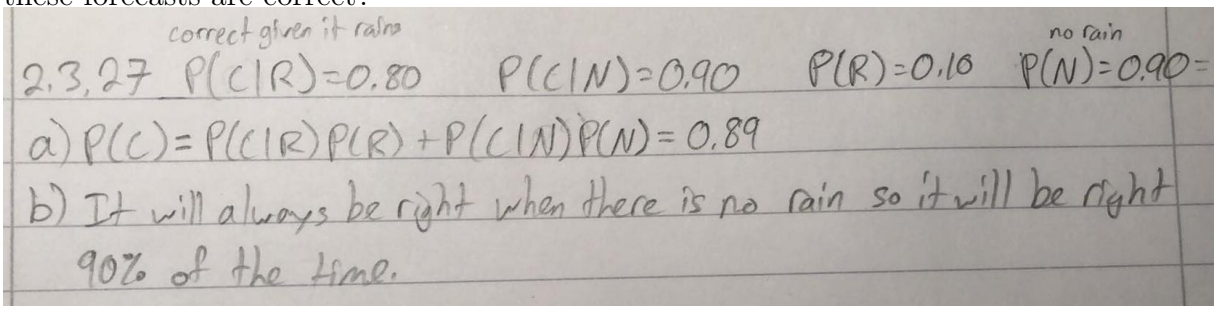
g. Are A and B independent? Is it reasonable to treat A and B as though they were independent? Explain.



2.3.27 Each day, a weather forecaster predicts whether or not it will rain. For 80% of rainy days, she correctly predicts that it will rain. For 90% of non-rainy days, she correctly predicts that it will not rain. Suppose that 10% of days are rainy and 90% are non-rainy.

a. What proportion of the forecasts are correct?

b. Another forecaster always predicts that there will be no rain. What proportion of these forecasts are correct?



2.3 Problem Set

Lilja Yr Guðmundsdóttir

1. 50 light bulbs (5 bulbs bad) $0.1 = P(B)$ $\therefore P(G) = (1 - P(B)) = 0.9$ (45 bulbs)

a) $P(2 \text{ good bulbs}) = P(G_1) \cdot P(G_2) = \frac{45}{50} \cdot \frac{44}{49} = \frac{1980}{2450} \approx 0.809$ or $\sim 81\%$

b) $P(1^{st} \text{ bad}, 2^{nd} \text{ good}) = P(B_1) \cdot P(G_2) = \frac{5}{50} \times \frac{45}{49} = \frac{45}{490} \approx 0.092$ or $\sim 9.2\%$

c) $P(1 \text{ bulb}, 2^{nd} \text{ good}) = P(B \cap G) \cdot P(G_2) = (P(B) \cdot P(G_2)) + (P(G) + P(G_2))$
$= \left(\frac{5}{50} \cdot \frac{45}{49}\right) + \left(\frac{45}{50} \cdot \frac{44}{49}\right) = \frac{1980}{2450} + \frac{45}{490} \approx 0.901$ or $\sim 90\%$ chance

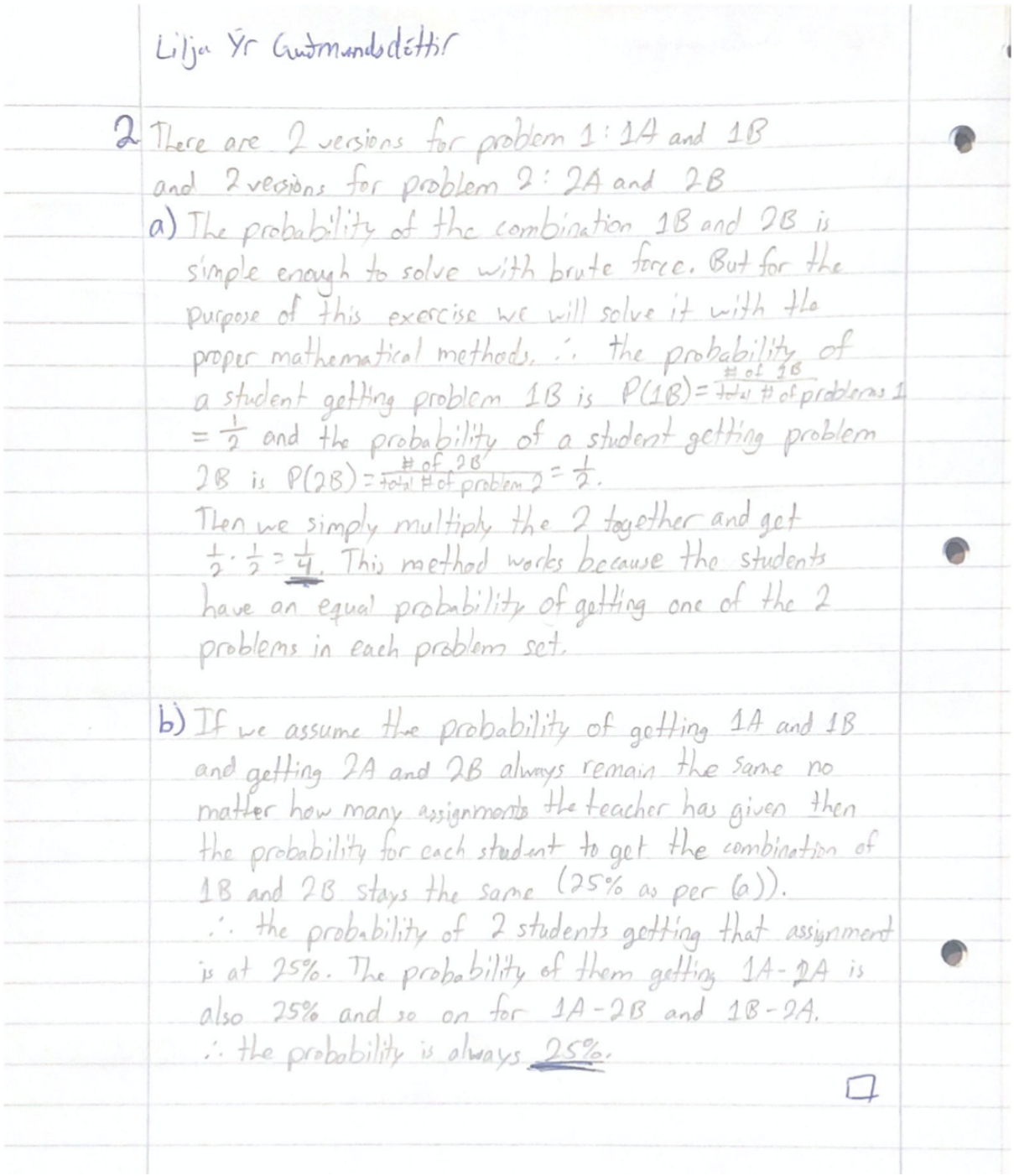
Explanations for problem 1

a) You first pick one light bulb out of all 50 so the probability of that one being good is $P(G_1) = \frac{\#\text{ good bulbs}}{\text{total }\#\text{ of bulbs}}$.
Then when the second bulb is being picked then there is one less good bulb to pick and one less bulb in total assuming we don't replace. Therefore ($\therefore$) The probability of picking another good bulb is $P(G_2) = \frac{\#\text{ good bulbs remaining}}{\text{total }\#\text{ of remaining bulbs}} = \frac{\#\text{good bulbs total} - 1}{\text{total }\#\text{ of bulbs} - 1}$.
Then just multiply the two values together to get total probability.

b) Same principle as in (a) except now we have to factor in the # of bad bulbs. First we pick one light bulb and the probability of that one being bad is $P(B_1) = \frac{\#\text{ of bad bulbs}}{\text{total }\#\text{ of bulbs}}$.
Then we pick another bulb from a pile of bulbs one fewer than before in total but, assuming that the first bulb that was picked was bad, the amount of good bulbs hasn't changed. $\therefore$ the probability of a good bulb is $P(G_2) = \frac{\#\text{ of good bulbs}}{\text{total }\#\text{ of remaining bulbs}} = \frac{\#\text{ of good bulbs}}{\text{total }\#\text{ of bulbs} - 1}$.
Then just multiply the two together like before.

c) This splits into 2 parts:
i) Probability of drawing 2 good bulbs: this was accomplished in (a).
ii) Probability of drawing first 1 bad bulb and then 1 good bulb: this was accomplished in (b).
To get the final probability of drawing a good bulb irrespective of whether the first was good or bad simply add parts (i) and (ii) together. □

BRUNNER 14



2.4 Problem Set solutions

1. A box contains 50 light bulbs, some of them bad and some of them good. The probability to pick a bad one is p. Two bulbs are drawn at random. Calculate:
a) The probability that both are good.We have N bulbs in total. Of them $B = pN$ are

bad and $G = (1-p)N$ are good, and $G + B = N$. $P(both1and2good) = P(1good \cap 2good) = P(2good \cap 1good) = P(2good|1good)P(1good) = \frac{G-1}{N-1}\frac{G}{N}$.
b) The probability that if the first is bad the second is good.
If we picked one bad, there are still G good inside, but N-1 in total. $P(2good|1bad) = \frac{G}{N-1}$.
c) The probability that the second is good irrespective of whether the first was good or bad.
Use the law of total probability: $P(2good) = P(2good \cap 1good) + P(2good \cap 1bad) = P(2good|1good)P(1good) + P(2good|1bad)P(1bad) = \frac{G-1}{N-1}\frac{G}{N} + \frac{G}{N-1}\frac{B}{N} = \frac{G}{N}$. So we see that if we do not condition on the quality of bulb 1, then the probability that the 2-nd is good is the same as the probability that the 1-st is good.

2. An assignment for the statistics class consists of two problems. Problem 1 has two versions, 1A and 1B, and and problem 2 has also two versions, 2A and 2B.The two versions of each problem are randomized.Meaning that each student receives at random 1A or 1B (not both) with equal probabilities, and 2A or 2B (not both) with equal probabilities.Calculate:
a) The probability that a student obtains a specific combination.
All possible combinations for a single student: 1A-2A, 1A-2B, 2A-1B, 2A-2B The probability of any combination is 1/4.
b) The probability that two students obtain the same assignment:
All combinations for student 1: 1A-2A, 1A-2B, 2A-1B, 2A-2B
All combinations for student 2: 1A-2A, 1A-2B, 2A-1B, 2A-2B
We observe4 possibilities for the same assignment out of16 possibilities in total, so the result is 4/16=1/4.

3 Week 3

3.1 Concepts

Example:The table shows the values of a discrete random variable n and their probabilities p(n).

n	0	1	2	3
p(n)	0.4	0.5	0.2	

What is the mean value of n? -> Mean value of n is 1.

Bayes rule:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)}$$

- The cumulative distribution function F(a) is the probability that the random variable is less or equal to a
- The variance of a random variable x is not $[E(x)]^2 - E(x^2)$
- Considering a continuous random variable x with values between 0 and 1 the probability that x=0.5 is 0
- Consider a continuous random variable x. The area enclosed between the curve representing the probability density function f(x) and the x axis depends on f(x)
- Suppose R and S are two independent random variables, both with the same variance v. The variance of R-S is 2v
- Suppose Z is a random variable with standard deviation s. The standard deviation is $\sqrt{2} * x$
- We want to measure the average mass of coffee beans of a certain type. We measure the mass of 100 beans, one at a time, and take the average. We also calculate the sample standard deviation s. The standard deviation of the average mass is s/10
- In order to reduce the std. of the average mass 10 times we should increase the number of measured coffee beans $\sqrt{10}$ times
- Two independent random variables have values between 0 and 1. Their covariance is 0

3.2 Example Problem Set

Examples of statistics in health sciences

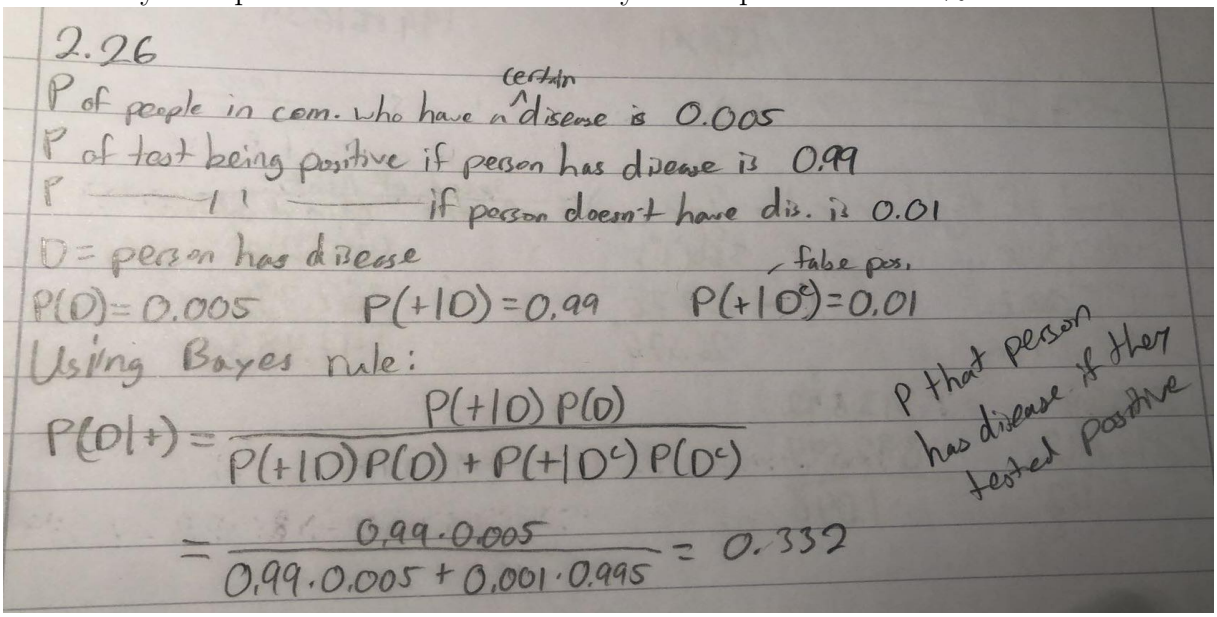
Example 2.26 The proportion of people in a given community who have a certain disease is 0.005. A test is available to diagnose the disease. If a person has the disease, the probability that the test will produce a positive signal is 0.99. If a person does not have the disease, the probability that the test will produce a positive signal is 0.01. If a person tests positive, what is the probability that the person actually has the disease?

D = person has the disease
P(D) = 0.005, P(+|D) = 0.99, P(+|$\hat{\text{DC}}$) = 0.01 (false positive)

Using Bayes rule:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|D^c)P(D^c)} = \frac{0.99 * 0.005}{0.99 * 0.005 + 0.001 * 0.995} = 0.332$$

Probability that person has the disease if they tested positive is 33.2%

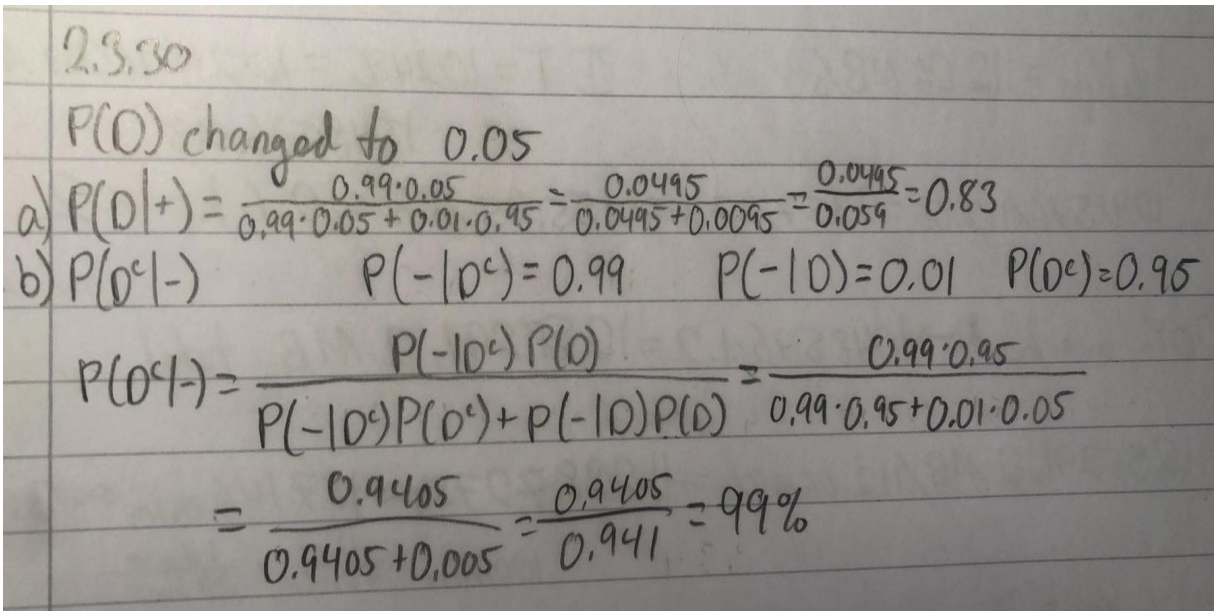


Problem 2.3.30 A drag racer has two parachutes, a main and a backup, that are designed to bring the vehicle to a stop after the end of a run. Suppose that the main chute deploys with probability 0.99, and that if the main fails to deploy, the backup deploys with probability 0.98.
a. What is the probability that one of the two parachutes deploys?

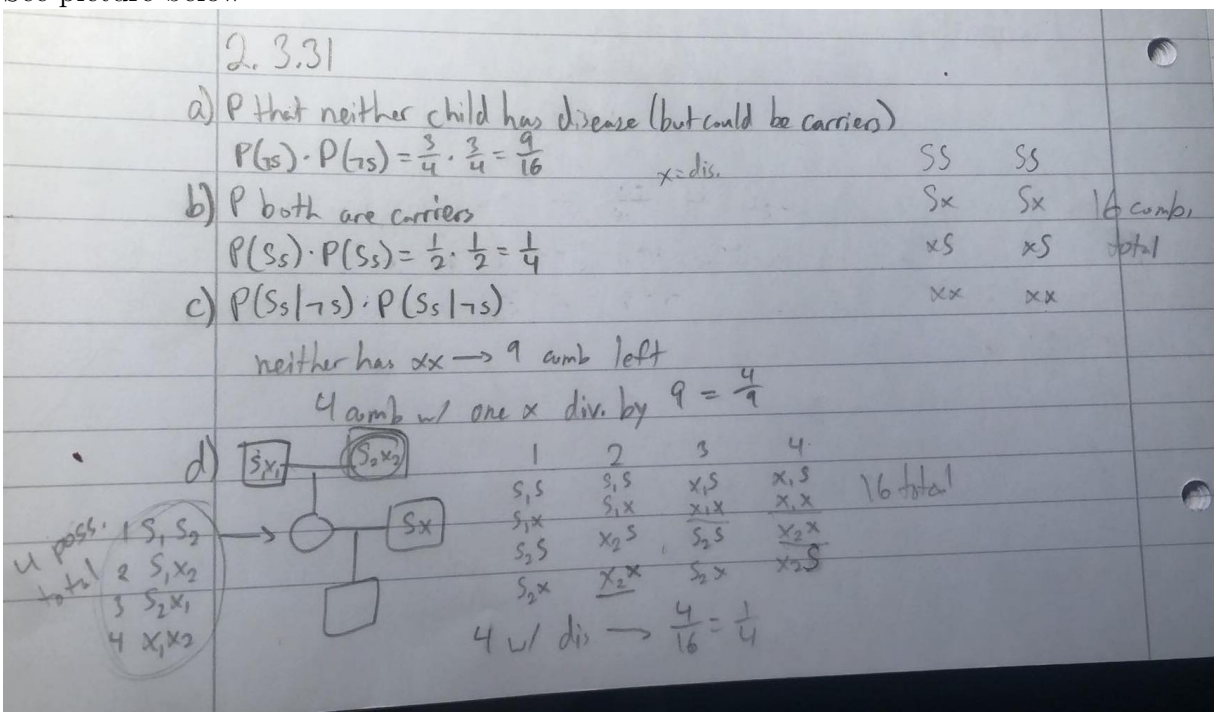
$$P(D|+) = \frac{0.99 * 0.05}{0.99 * 0.005 + 0.01 * 0.95} + \frac{0.0495}{0.059} = 0.83$$

b. What is the probability that the backup parachute deploys?
$P(-|D^c) = 0.99, P(-|D) = 0.01, P(D^c) = 0.95$ which gives:

$$P(D^c|-) = \frac{P(-|D^c)P(D)}{P(-|D^c)P(D^c) + P(-|D)P(D)} = \frac{0.99 * 0.95}{0.99 * 0.95 + 0.01 * 0.05} = 0.99$$



2.3.31 Sickle-cell anemia is an inherited disease in which red blood cells are misshapen and sticky. Sickle cells tend to form clumps in blood vessels, inhibiting the flow of blood. Humans have two genes for sickle-cell anemia, either of which may be S for normal cells or s for sickle cells. A person with two copies of the s gene will have sickle-cell anemia. A person with one s gene and one S gene will not have the disease, but will be a carrier, which means that the s gene may be transmitted to the person's offspring. If two carriers have a child, the probability is 0.25 that the child will have the disease and 0.5 that the child will be a carrier. Outcomes among children are independent.
a. A mother and father who are both carriers have two children. What is the probability that neither child has the disease?
P(¬s)*P(¬s) = $\frac{3}{4}*\frac{3}{4} = 9/16$
b. What is the probability that both children are carriers?
P(Ss)*P(Ss) = 1/2 * 1/2 = 1/4
c. If neither child has the disease, what is the probability that both are carriers?
P(Ss|¬s)*P(Ss|¬s) = neither has ss => 9 combos left, 4 combos with one s divided by 9 gives 4/9
d. A woman who is the child of two carriers has a child by a man who is a carrier. What is the probability that this child has the disease?
See picture below



Reliability analysis

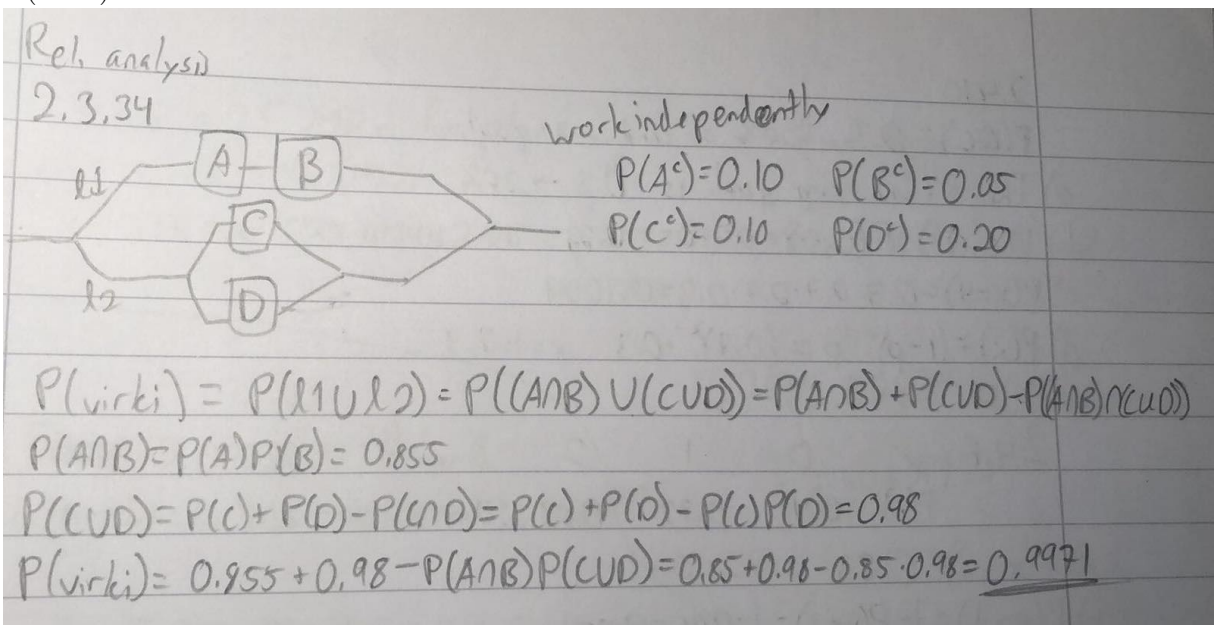
2.3.34 A system consists of four components connected as shown in the diagram in the image below. Assume A, B, C, and D function independently. If the probabilities that A, B, C, and D fail are 0.10, 0.05, 0.10, and 0.20, respectively, what is the probability that the system functions?

P(virki) = P(l1∪l2) = P((A∩B)∪(C∪D)) = P(A∩B) + P(C∪D) - P((A∩B)(C∪D))

P(A∩B) = P(A)P(B) = 0.855

P(C∪D) = P(C) + P(D) - P(C∩D) = P(C) + P(D) - P(C)P(D) = 0.98

P(virki) = 0.855 + 0.98 - 0.855*0.98 = 0.9971

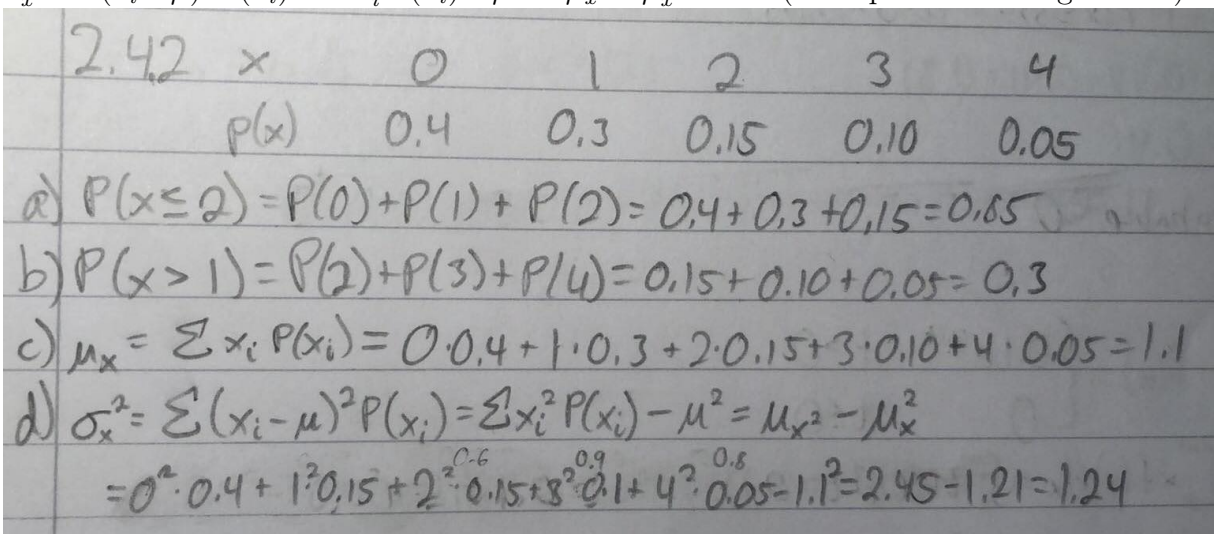


Discrete random variables

2.4.2 Computer chips often contain surface imperfections. For a certain type of computer chip, the probability mass function of the number of defects X is presented in the following table.

x	0	1	2	3	4
p(x)	0.4	0.3	0.15	0.10	0.05

a. Find $P(X \leq 2)$.
$P(X \leq 2) = P(0)+P(1)+P(2) = 0.4 + 0.3 + 0.15 = 0.85$
b. Find $P(X > 1)$.
$P(X > 1) = P(2)+P(3)+P(4) = 0.15 + 0.10 + 0.05 = 0.3$
c. Find μ_x.
$\mu_x = \Sigma x_i P(x_i) = 0*0.4 + 1*0.3 + 2*0.15 + 3*0.10 + 4*0.05 = 1.1$
d. Find σ_x^2.
$\sigma_x^2 = \Sigma(x_i - \mu)^2 P(x_i) = \Sigma x_i^2 P(x_i) - \mu^2 = \mu_{x^2} - \mu_x^2 = 1.24$ (full equation in image below).



2.4.7 A computer sends a packet of information along a channel and waits for a return signal acknowledging that the packet has been received. If no acknowledgment is received within a certain time, the packet is re-sent. Let X represent the number of times the packet is sent. Assume that the probability mass function of X is given by (see image below) where c is a constant.

SEE IMAGE BELOW FOR ANSWERS.

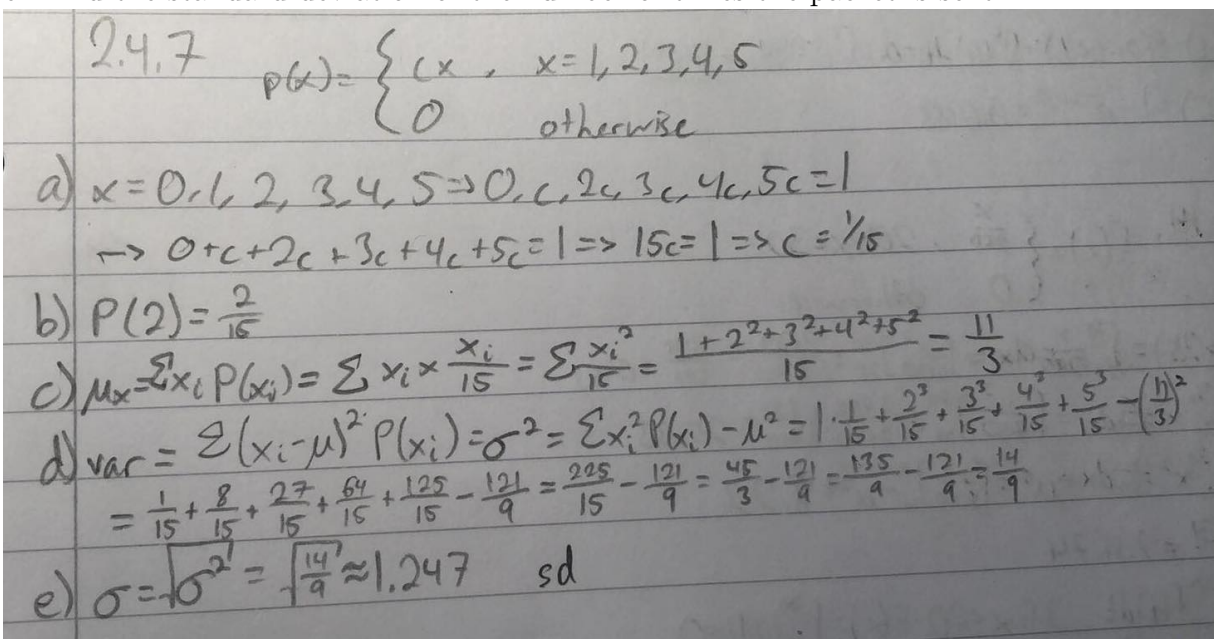
a. Find the value of the constant c so that p(x) is a probability mass function.

b. Find P(X = 2).

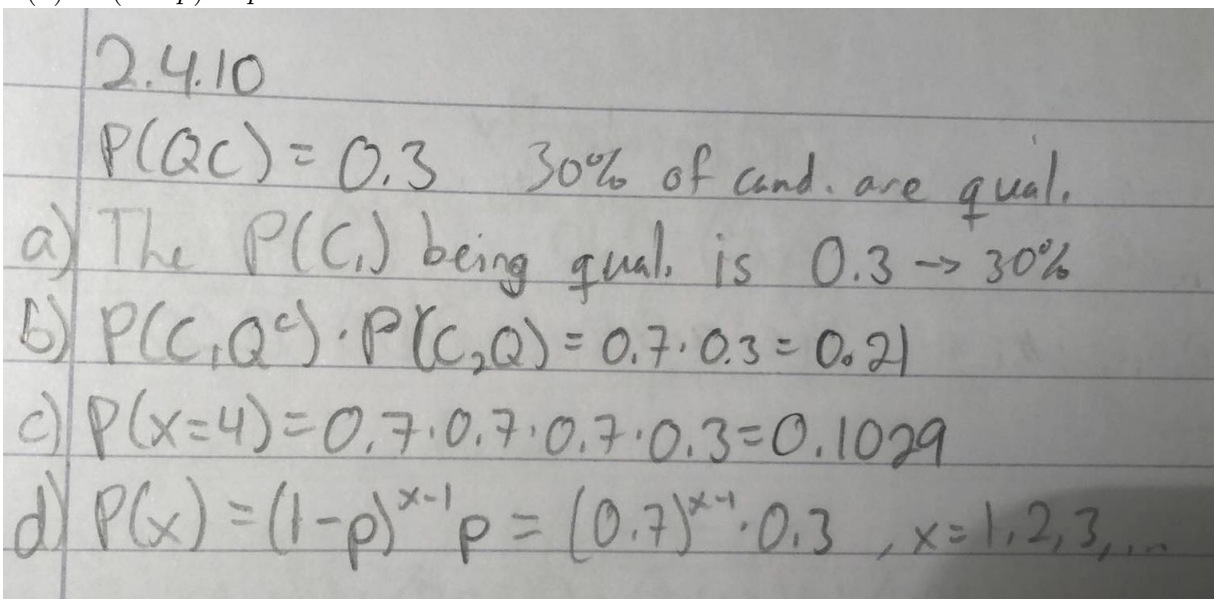
c. Find the mean number of times the packet is sent.

d. Find the variance of the number of times the packet is sent.

e. Find the standard deviation of the number of times the packet is sent.



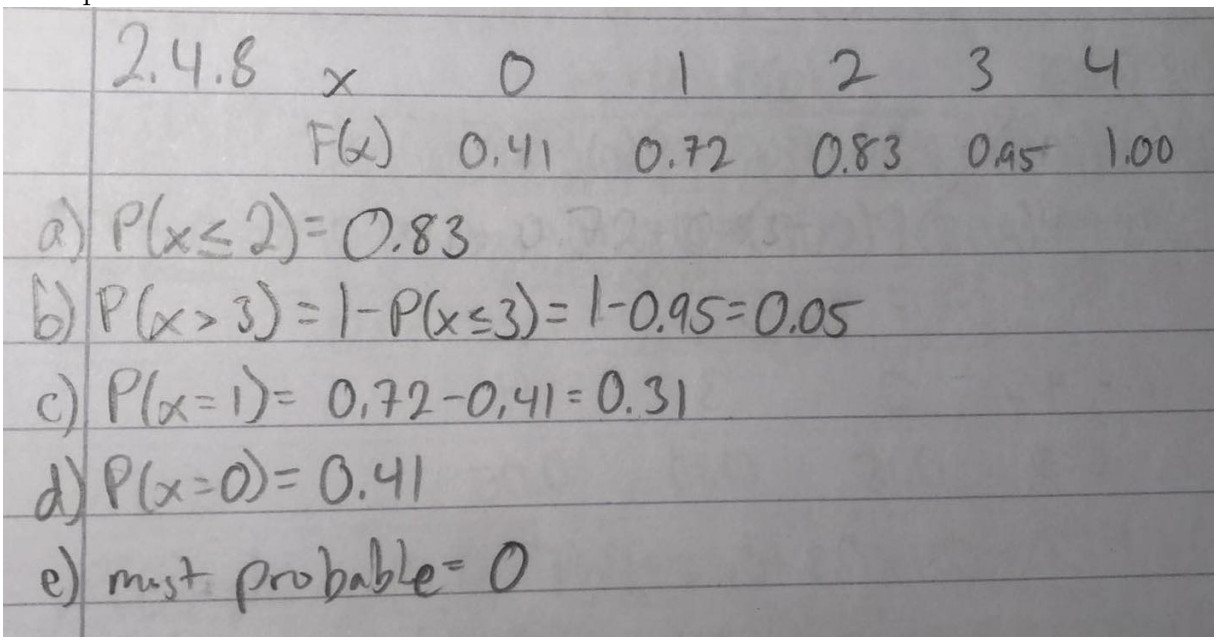
2.4.10 Candidates for a job are interviewed one by one until a qualified candidate is found. Thirty percent of the candidates are qualified.
a. What is the probability that the first candidate is qualified?
The P(c_1) being qualified is 0.3 -> 30%
b. What is the probability that the first candidate is unqualified and the second candidate is qualified?
P(C_1Q^c)*P(C_2Q) = 0.7*0.3 = 0.21
c. Let X represent the number of candidates interviewed up to and including the first qualified candidate. Find P(X = 4).
P(X=4) = 0.7*0.7*0.7*0.3 = 0.1029
d. Find the probability mass function of X.
P(x) = $(1-p)^{x-1}p$...



2.4.8 After manufacture, computer disks are tested for errors. Let X be the number of errors detected on a randomly chosen disk. The following table presents values of the cumulative distribution function F(x) of X.

x	0	1	2	3	4
F(x)	0.41	0.72	0.83	0.95	1.00

a. What is the probability that two or fewer errors are detected?
$P(x \leq 2) = 0.83$
b. What is the probability that more than three errors are detected?
$P(x>3) = 1 - P(X \leq 3) = 1\text{-}0.95 = 0.05$
c. What is the probability that exactly one error is detected?
$P(x=1) = 0.72\text{-}0.41 = 0.31$
d. What is the probability that no errors are detected?
$P(x=0) = 0.41$
e. What is the most probable number of errors to be detected?
Most probable = 0



Continuous random variables

2.4.15 The lifetime of a transistor in a certain application is random with probability density function in the image below.

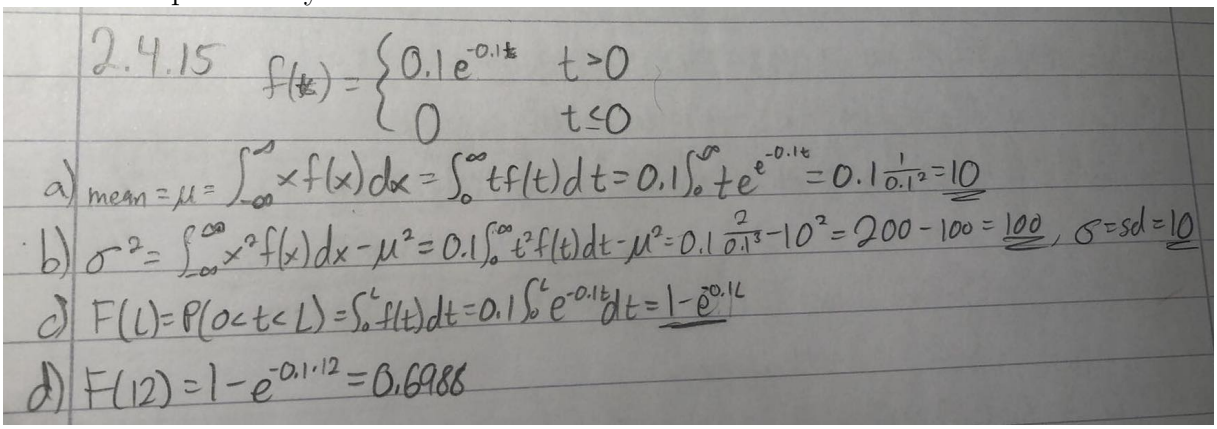
SEE IMAGE FOR CALCULATIONS.

a. Find the mean lifetime.

b. Find the standard deviation of the lifetimes.

c. Find the cumulative distribution function of the lifetime.

d. Find the probability that the lifetime will be less than 12 months.



2.4.14 Elongation (in percent) of steel plates treated with aluminum are random with probability density function

a. What proportion of steel plates have elongations greater than 25%?

SEE IMAGE FOR CALCULATIONS.

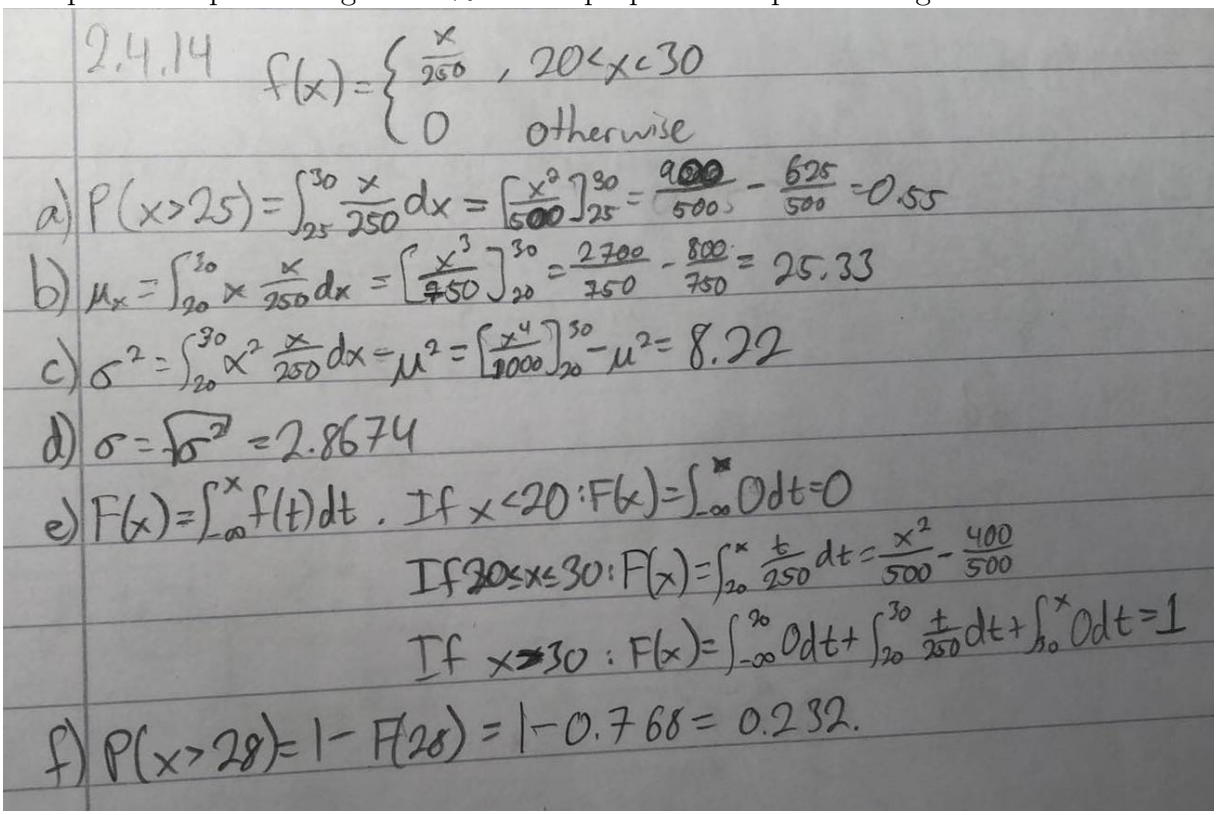
b. Find the mean elongation.

c. Find the variance of the elongations.

d. Find the standard deviation of the elongations.

e. Find the cumulative distribution function of the elongations.

f. A particular plate elongates 28%. What proportion of plates elongate more than this?



2.4.24 Particles are a major component of air pollution in many areas. It is of interest to study the sizes of contaminating particles. Let X represent the diameter, in micrometers, of a randomly chosen particle. Assume that in a certain area, the probability density function of X is inversely proportional to the volume of the particle; that is, assume that (in function in image below) c is a constant.

a. Find the value of c so that f(x) is a probability density function.

b. Find the mean particle diameter.

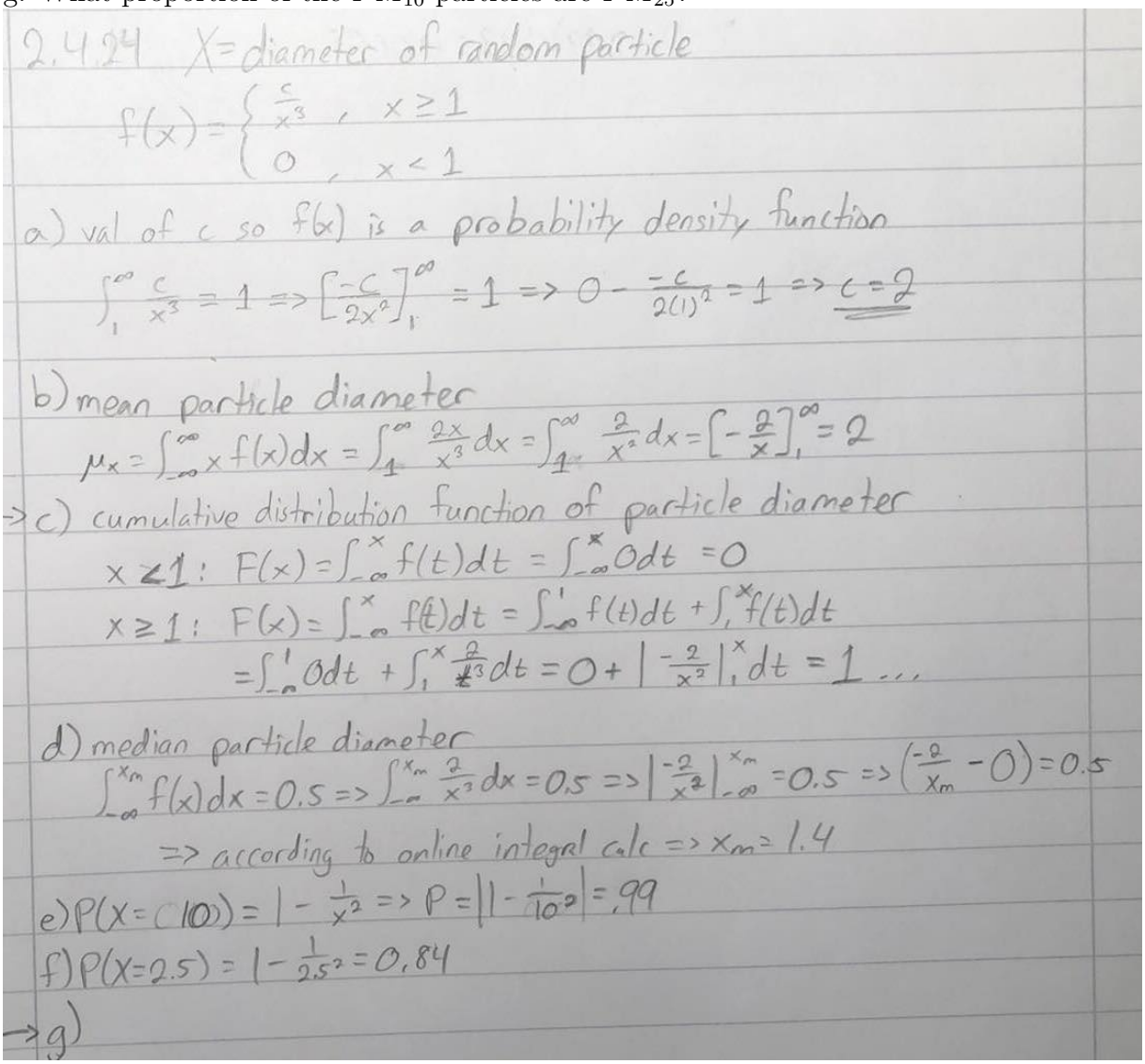
c. Find the cumulative distribution function of the particle diameter.

d. Find the median particle diameter.

e. The term PM_{10} refers to particles 10 μm or less in diameter. What proportion of the contaminating particles are PM10?

f. The term PM_{25} refers to particles 2.5 μm or less in diameter. What proportion of the contaminating particles are PM_{25}?

g. What proportion of the PM_{10} particles are PM_{25}?



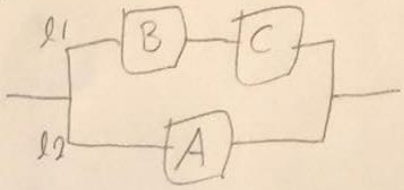
3.3 Problem Set

1. The system shown in the figure consists of three elements A, B, C, which work independently. Each of them has a probability of failing p=0.3. Calculate:
a) The probability that all elements work.
b) The probability that the system works.
2. Calculate the standard deviation of the random variable n represented in the table.

n	0	1	2	3
p(n)	0.4	0.3	0.2	0.1

3. Consider a random variable $x \in (0,\infty)$ with the probability density function $f(x) = e$.
a) Find the cumulative distribution function
b) Find the probability that $x>1$.

1.



P(each failure) = 0.3

a) $P(B) = P(C) = P(A) = 1 - P(\text{Fail}) = \underline{0.7}$

b) $P(\text{vicki}) = P(1 \cup 2) = P((C \cap B) \cup (A))$

$= P(C \cap B) + P(A) - P((C \cap B) \cap (A))$

$P(C \cap B) = P(C)P(B) = 0.7 \cdot 0.7 = 0.49$

$P(A) = 0.7$

$P(\text{vicki}) = 0.49 + 0.7 - P(C \cap B)P(A) = 0.49 + 0.7 - 0.49 \cdot 0.7$

$= \underline{0,847}.$ □

2.

n	0	1	2	3
P(n)	0.4	0.3	0.2	0.1

i) $\mu_x = \sum x_i P(x_i) = 0 \cdot 0.4 + 1 \cdot 0.3 + 2 \cdot 0.2 + 3 \cdot 0.1 = 0.3 + 0.4 + 0.3 = 1$

ii) $\sigma_x^2 = \sum x_i^2 P(x_i) - \mu_x^2 = 0^2 \cdot 0.4 + 1^2 \cdot 0.3 + 2^2 \cdot 0.2 + 3^2 \cdot 0.1 - 1^2$

$= 0 + 0.3 + 0.8 + 0.9 - 1 = 2 - 1 = 1$

$\therefore$ the sd $= \sigma_x = \sqrt{1} = \underline{1}.$ □

3. $x \in (0; \infty)$; probability density function $f(x) = e^{-x}$.

a) $F(x) = \int_0^x f(t)dt = \int_0^x e^{-t} = [-e^{-t}]_0^x = \underline{1 - e^{-x}}$

b) ~~$P(x>1) = \int_1^\infty t f(t)dt = \int_1^\infty t e^t dt = -(x+1)e^{-x}$~~

$= 1 - \int_0^1 t f(t)dt = 1 - \int_0^1 t e^{-t} dt = 1 - (1 - 2e^{-1}) = 1 - 1 + 2e^{-1} = \underline{2e^{-1}}.$ □

3.4 Problem set Solution

1.The system shown in the figure consists of three elements A, B, C, which work independently. Each of them has a probability of failing p=0.3. Calculate:
a) The probability that all elements work.
P(one element works)=1-p. P(all elements work)= P(A∩B∩C) = P(A)P(B)P(C)=$(1-p)^3$=0.343 Where P(A) means P(A works)
b) The probability that the system works.
P(A∪(B∩C)) = P(A) + P(B∩C) –P(A∩B∩C) = (1-p) + $(1-p)^2$–$(1-p)^3$= 0.847. The three versions differ only by a permutation of elements, but the numerical results are identical.

2. Calculate the standard deviation of the random variable n represented in the table.

n	0	1	2	3
Version A p(n)	0.4	0.3	0.2	0.1
Version B p(n)	0.1	0.2	0.3	0.4
Version C p(n)	0.2	0.3	0.1	0.4

$$\sigma^2 = \sum_i n^2 p(n) - \left[\sum_i np(n)\right]^2 = 2 - 1\ (A) \text{ or } 5 - 2^2\ (B) \text{ or } 4.3 - 1.7^2\ (C) = 1\ (A, B) \text{ or } 1.41\ (C)$$

$$\sigma = 1\ (A, B) \text{ or } 1.187\ (C)$$

3. Consider a random variable x∈(0,∞) with the probability density function f(x)=e - x.
a) Find the cumulative distribution function

$$F(x) = \int_0^x f(x')dx' = 1 - e^{-x}$$

b) Find the probability that x>1.

$$P(x > 1) = 1 - P(x < 1) = 1 - F(1) = e^{-1} \approx 0.368$$

4 Week 2

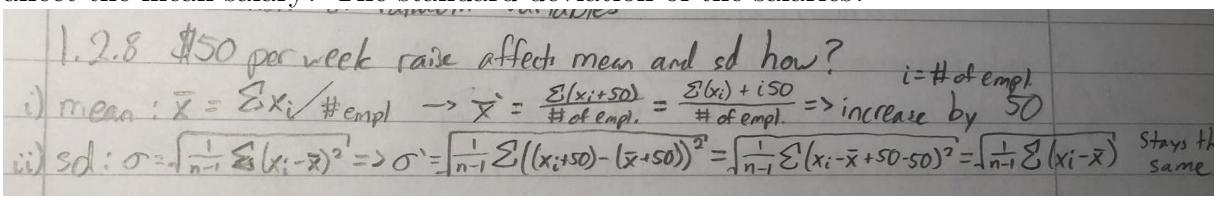
4.1 Concepts

- You measure a certain physical quantity several times and you calculate the mean value. The uncertainty of the mean value is smaller than the uncertainty of each measured value
- The bias of measurements can be reduced by reducing the systematic errors.
- Consider a cube with side equal to L. If the uncertainty of L is 0.3% the uncertainty of the volume is 0.9%
- Consider a measured a variable x and the function $U(x) = c\sqrt{x}$ where c is a constant number. The relative uncertainty of U is two times smaller than the relative uncertainty of x.
- The uncertainty of a function U of two measured variables x and y can be calculated with the formula based on partial derivatives if x and y are independent.

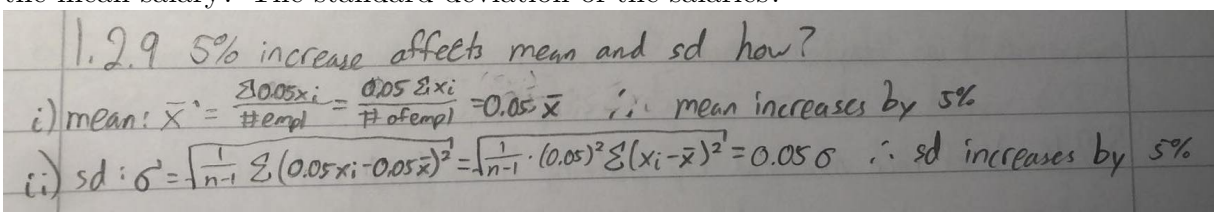
4.2 Example Problem Set

Linear functions of random variables

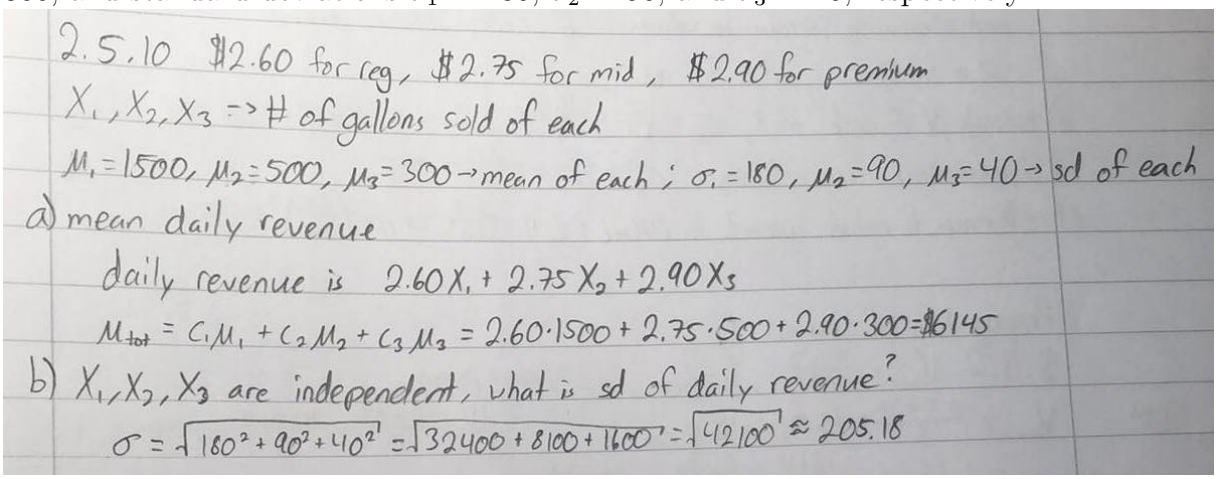
1.2.8 In a certain company, every worker received a $50-per-week raise. How does this affect the mean salary? The standard deviation of the salaries?



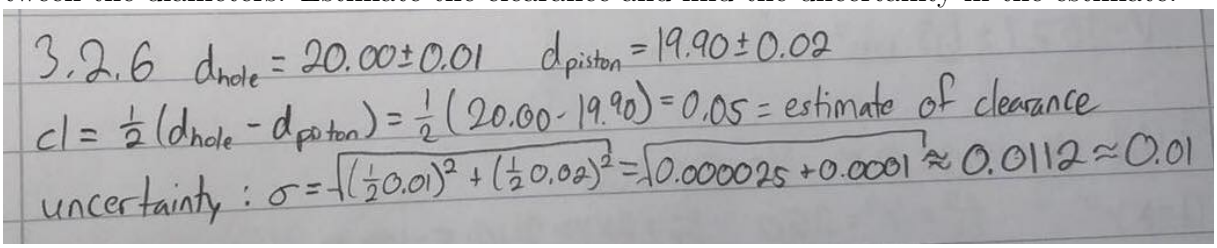
1.2.9 In another company, every worker received a 5% raise. How does this affect the mean salary? The standard deviation of the salaries?



2.5.10 A gas station earns \$2.60 in revenue for each gallon of regular gas it sells, \$2.75 for each gallon of midgrade gas, and \$2.90 for each gallon of premium gas. Let X1, X2, and X3 denote the numbers of gallons of regular, midgrade, and premium gasoline sold in a day. Assume that X1, X2, and X3 have means $\mu_1 = 1500$, $\mu_2 = 500$, and $\mu_3 = 300$, and standard deviations $\sigma_1 = 180$, $\sigma_2 = 90$, and $\sigma_3 = 40$, respectively.

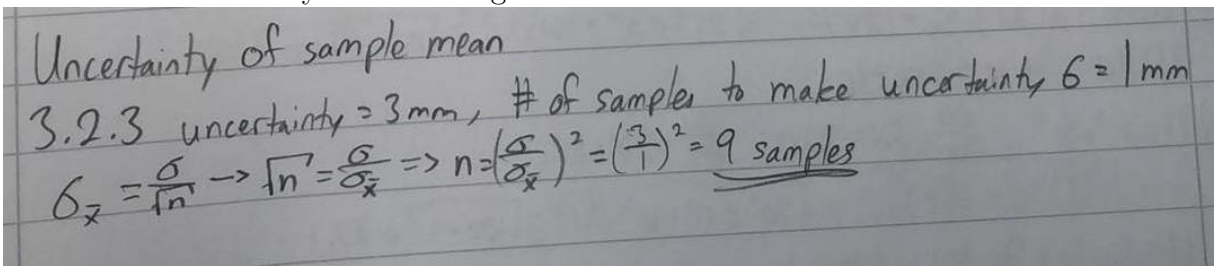


3.2.6 A cylindrical hole is bored through a steel block, and a cylindrical piston is machined to fit into the hole. The diameter of the hole is 20.00 ± 0.01 cm, and the diameter of the piston is 19.90 ± 0.02 cm. The clearance is one-half the difference between the diameters. Estimate the clearance and find the uncertainty in the estimate.



Uncertainty of sample mean

3.2.3 The length of a rod is to be measured by a process whose uncertainty is 3 mm. Several independent measurements will be taken, and the average of these measurements will be used to estimate the length of the rod. How many measurements must be made so that the uncertainty in the average will be 1 mm?



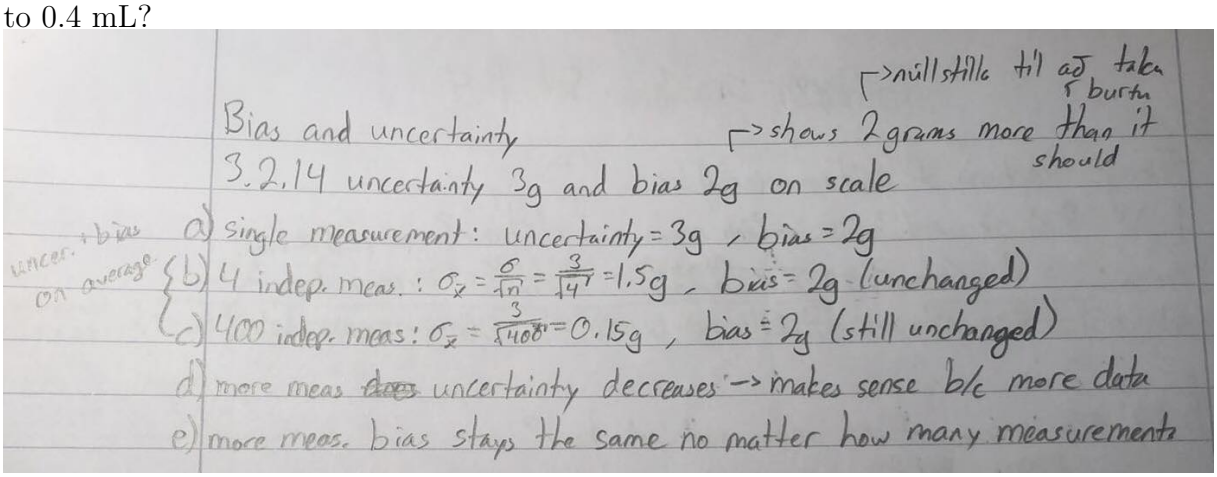
Bias and uncertainty

3.2.14 The volume of a rock is measured by placing the rock in a graduated cylinder partially filled with water and measuring the increase in volume. Eight independent measurements are made. The average of the measurements is 87.0 mL, and the standard deviation is 2.0 mL.

a. Estimate the volume of the rock, and find the uncertainty in the estimate.

b. Eight additional measurements are made, for a total of 16. What is the uncertainty, approximately, in the average of the 16 measurements?

c. Approximately how many measurements would be needed to reduce the uncertainty to 0.4 mL?



3.2.15 A certain scale has an uncertainty of 3 g and a bias of 2 g.

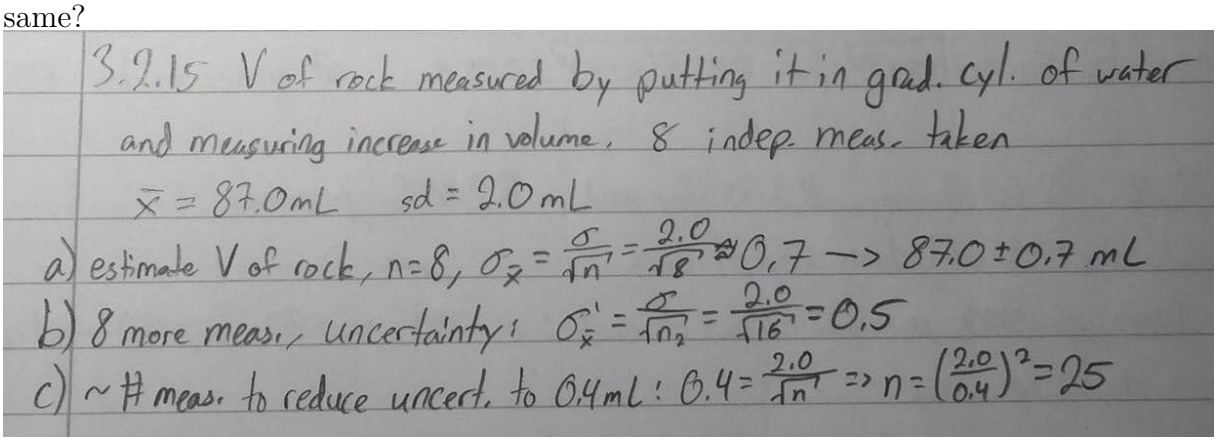
a. A single measurement is made on this scale. What are the bias and uncertainty in this measurement?

b. Four independent measurements are made on this scale. What are the bias and uncertainty in the average of these measurements?

c. Four hundred independent measurements are made on this scale. What are the bias and uncertainty in the average of these measurements?

d. As more measurements are made, does the uncertainty get smaller, get larger, or stay the same?

e. As more measurements are made, does the bias get smaller, get larger, or stay the same?



Propagation of errors

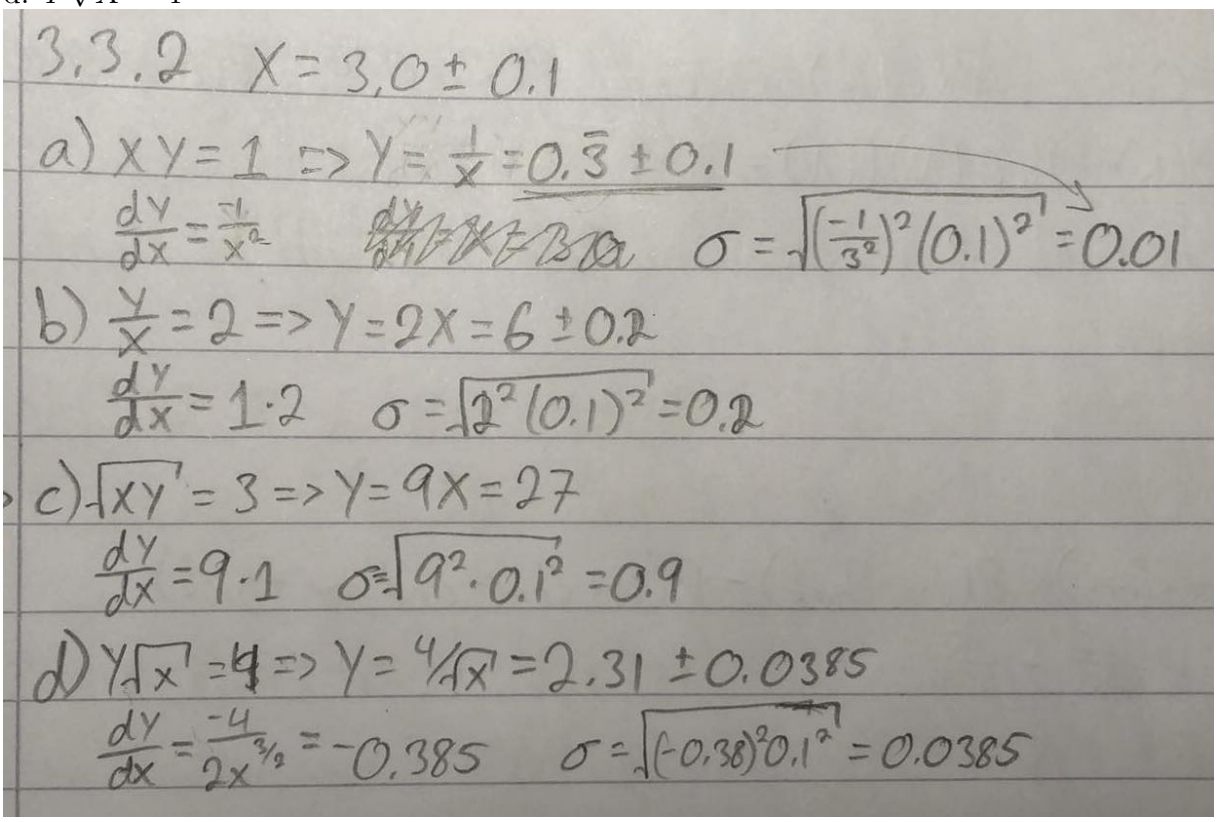
3.3.2 Given that X and Y are related by the given equation, and that $X = 3.0 \pm 0.1$, estimate Y and its uncertainty.

a. $XY = 1$

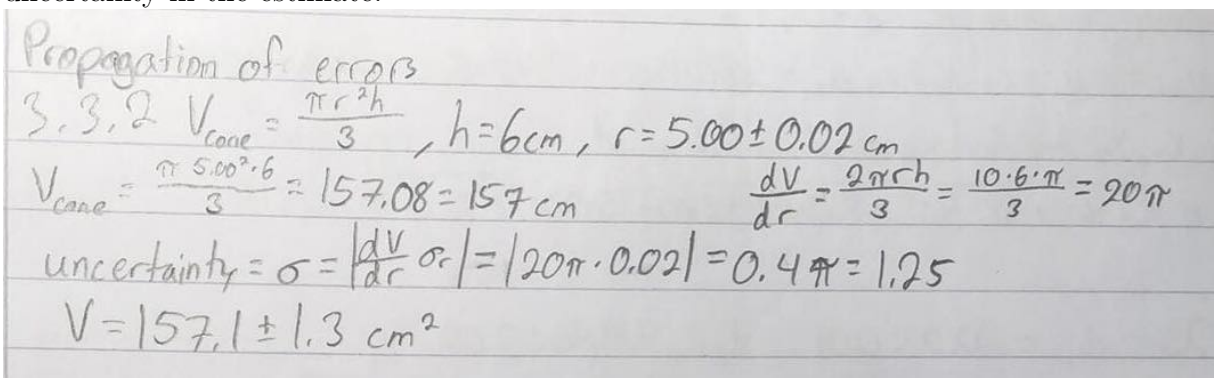
b. $Y/x = 2$

c. $\sqrt{XY} = 3$

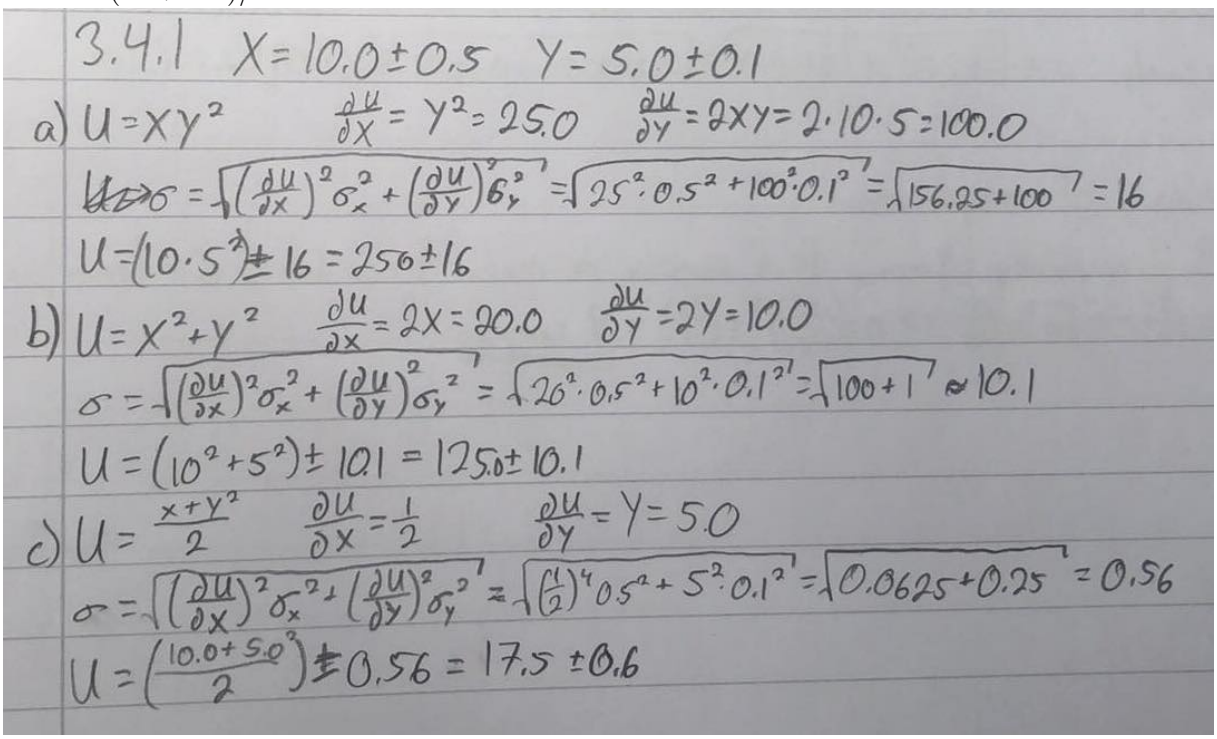
d. $Y\sqrt{X} = 4$



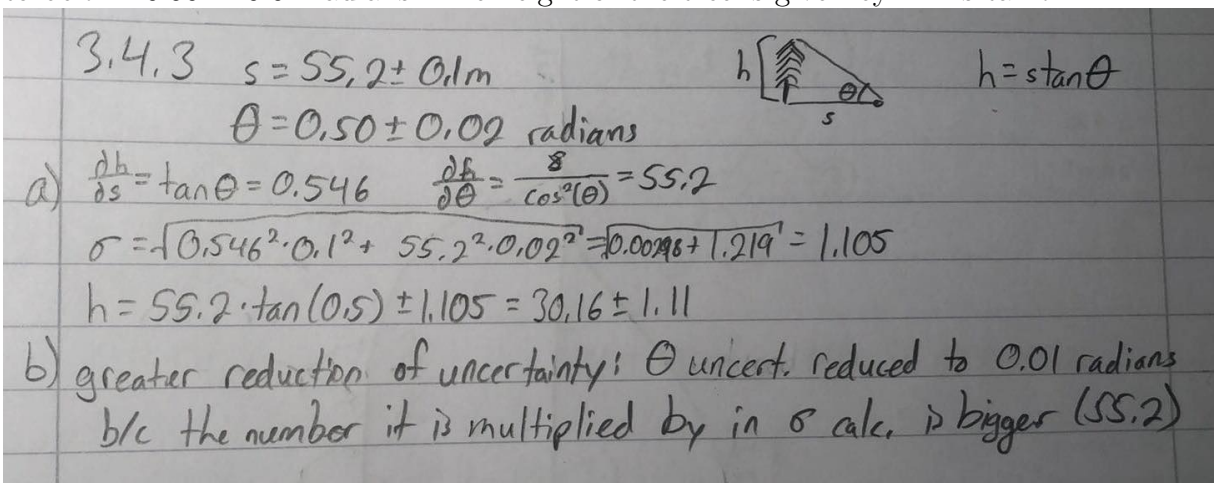
3.3.3 The volume of a cone is given by $V = \pi r_r h/3$, where r is the radius of the base and h is the height. Assume the height is 6 cm, measured with negligible uncertainty, and the radius is r = 5.00 ± 0.02 cm. Estimate the volume of the cone, and find the uncertainty in the estimate.



3.4.1 Find the uncertainty in U, assuming that X = 10.0 ± 0.5, Y = 5.0 ± 0.1, and
a. U = XY^2
b. U = $X^2 + Y^2$
c. U = (X + Y^2)/2

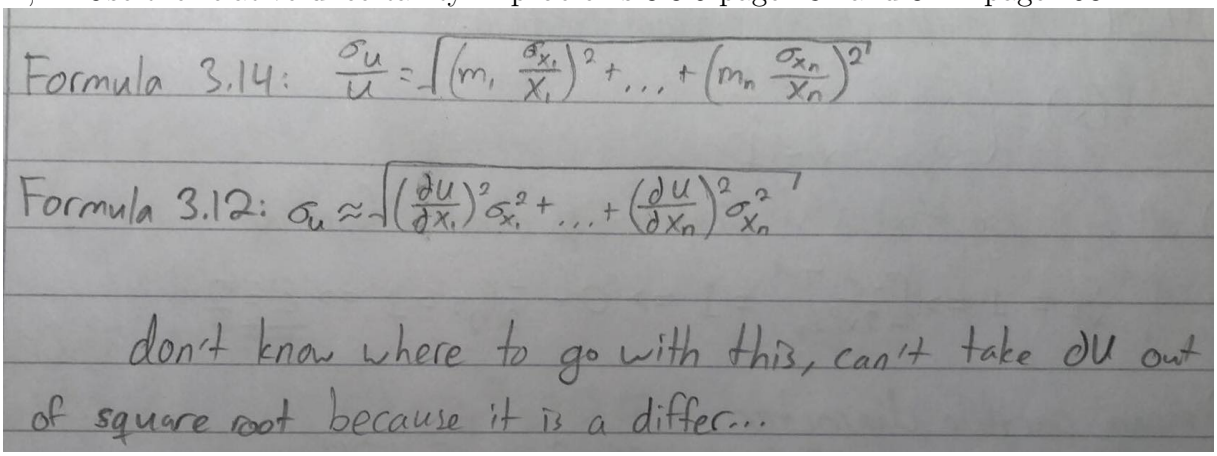


3.4.3 From a fixed point on the ground, the distance to a certain tree is measured to be s = 55.2 ± 0.1 m and the angle from the point to the top of the tree is measured to be θ = 0.50 ± 0.02 radians. The height of the tree is given by h = s tan θ.



Relative uncertainty

Prove formula 3.14 at page 192 using the formula 3.12 at page 186 with 3 variables X, Y, Z. Use the relative uncertainty in problems 3.3.5 page 184 and 3.4.2 page 193.



4.3 Problem Set

Verkefni 4

1. $x \in (0,1)$ (RV) $\mu = 0.5$ $\sigma_x = \frac{1}{\sqrt{12}}$ $y \in (a,b) \rightarrow y = a + (b-a)x$

a) mean of y: we multiply and get $\mu_y = (b-a)\mu_x + a = a + (b-a)\mu_x$

$\Rightarrow \underline{\mu_y = a + 0.5(b-a)}$ (given by rules defined by Linear Comb. of RVs).

b) sd of y: we get by those same rules used in (a) that σ_y is:

$\sigma_y^2 = (b-a)^2\sigma_x^2 \Rightarrow \sigma_y = (b-a)\sigma_x \Rightarrow \sigma_y = (b-a)\frac{1}{\sqrt{12}} \Rightarrow \underline{\sigma_y = \frac{b-a}{\sqrt{12}}}$. □

2. Same RV as in (1) but a sample of it of size n. How large does n have to be for the uncertainty to be ± 0.01?

We obtain uncertainty of sample size by using the equation $\sigma_x' = \frac{\sigma_x}{\sqrt{n}}$.

In order to get $\sigma_x' = \pm 0.01$ then: $n = \left(\frac{\sigma_x}{\sigma_x'}\right) = \left(\frac{\frac{1}{\sqrt{12}}}{0.01}\right) = \frac{50\sqrt{3}}{3} \approx 28.86$

Therefore the sample size n has to be at least $\underline{29}$ (to be sure the uncertainty is around ± 0.01). □

3. Cyl. silver wire w/ $D = 4.0 \pm 0.1$ mm

resistance: $R = \rho\frac{L}{A}$, $\rho = 1.59 \times 10^{-8}\,\Omega m$, $L = 200 \pm 1$ → length of wire

A is cross sectional area.

a) relative uncertainty of A

To calculate A we do $A = \frac{D}{2}\pi = \frac{4.0}{2}\pi = 2\pi$. Now lets find the uncertainty.

We find that $\sigma_A = \left|\frac{\partial A}{\partial D}\right|\sigma_D = \left|\frac{1}{2}\pi\right|\sigma_D = \frac{1}{2}\pi \cdot 0.1 = 0.157 \approx \underline{0.2}$.

b) relative uncertainty of resistance R.

To calculate R we use the given equation $R = \rho\frac{L}{A}$.

We find that $\sigma_R = \sqrt{\left(\left|\frac{\partial R}{\partial A}\right|\sigma_A\right)^2 + \left(\left|\frac{\partial R}{\partial L}\right|\sigma_L\right)^2} = \sqrt{\left(-\frac{\rho L}{A^2}\cdot\sigma_A\right)^2 + \left(\frac{\rho}{A}\sigma_L\right)^2}$

$\Rightarrow \sigma_R = \sqrt{\left((-1.59\times10^{8})\cdot\frac{200}{(2\pi)^2}\cdot\frac{1}{20}\pi\right)^2 + \left(\frac{1.59\times10^{-8}}{2\pi}\cdot 1\right)^2}$

$\Rightarrow \sigma_R = \sqrt{1.6\times10^{-16} + 6.4\times10^{-18}} = 1.289\times10^{-8} \approx \underline{1.3\times10^{-8}}$. □

4.4 Problem Set Solutions

1. A random variable x∈(0,1)with uniform distribution has mean value μ=0.5 and standard deviation σ=1/$\sqrt{}$12.If we need a random variable y with uniform distribution in another interval, y∈(a,b), we can obtain it with the linear transformation y=a+(b-a)x. Use the properties of the linear combinations of random variables to calculate
a) the mean of y: E(y)=a+(b-a)E(x)=a+(b-a)0.5=(b+a)/2
b) the standard deviation of y: Var(y)=$(b-a)^2$ Var(x)=$(b-a)^2/12$ => σ_y=(b-a)/$\sqrt{}$12
NOTE: Do not use integrals.

2. Consider a sample of the previous random variable x of size n. How large should be n such that the sample mean has an uncertainty ±0.01 ?
The uncertainty of the sample mean is $\sigma_{\bar{x}} = \frac{\sigma}{n} = 0.01 => n = (\frac{\sigma}{0.01})^2 = \frac{1}{12*10^{-4}} = 833$

3. Consider a cylindrical Silver wire with diameter D=4.0±0.1mm. The electrical resistance of the wire is R=Ω where ϱ=1.59×10^8Ωm (ohm meter) is the resistivity of the material, L=200±1mm is the length of the wire, and A is the cross sectional area.
a) relative uncertainty of A: $A = \frac{\pi D^2}{4} => \frac{\sigma_A}{A} = 2\frac{\sigma_D}{D} = 2\frac{0.1}{4} = 0.05$

b) relative uncertainty of R: $R = \rho\frac{L}{A} \Rightarrow \frac{\sigma_R}{R} = \sqrt{\left(\frac{\sigma_L}{L}\right)^2 + \left(\frac{\sigma_A}{A}\right)^2} = \sqrt{\left(\frac{1}{200}\right)^2 + (0.05)^2} = 0.05025$

Acceptable result for any given L (200, 250, or 300 nm): $\frac{\sigma_R}{R} \approx 0.05$

COMMENTS: The reason to ask for relative uncertainties, $\frac{\sigma_A}{A}$ and $\frac{\sigma_R}{R}$, is that in this example they are much easier to calculate than the absolute uncertainties σ_A and σ_R. We see that only a couple of ratios are needed, which can be obtined in a couple of minutes. However, many students still tried to do the full math, calculated all derivatives from scratch, or did other unnecessary calculations, and wasted their time. Other students did not show the relative uncertainties, like they did not know what they are. This concept has been very thoroughly covered in videos 4 and 5 of the Lecture Package 1 and in the last group of problems of week 4, but also in the data analyses of the physics labs already done by most of the students. That is why the course material must be studied before going over the assignment, and if so the 60 min time is more than enough for solutions and for the upload.
(Problems similar to 1 and 2 were also covered over the week 4.)

5 Week 2

5.1 Concepts

- 56 teams of 5 can be formed out of 8 individuals
- A random variable x has a binomial distribution with n=8 and p=0.8. The probability that x=5 is 0.147.
- The probability that x is less or equal to 5 if $x \sim Bin(8, 0.8)$ is 0.203
- The mean value of $x \sim Bin(n, p)$ is np
- The variance of $x \sim Bin(n, p)$ is np(1-p)
- The temperature in an office space is expected to be T = 22 Celsius degrees with an (absolute) uncertainty S = 1 Celsius degree, such that the temperature is reported as T ± S. The relative uncertainty is S/T
- In the Poisson distribution P(k, λ) the meaning of the parameter lambda (λ) is the mean value of k
- If the mean value of a Poisson random variable k is 15 then the probability that k=15 is 0.102
- Hvað veldur stórum stökkum á rafmagnsfjármálamörkuðum? Misræmi í framboði og eftirspurn, og að það að erfitt er að geyma rafmagn sem er búið að framleiða.
- Hvernig þarf að breyta Poisson ferlinu til þess að það geti orðið gott líkan fyrir þyrpingamyndun á rafmagnsmörkuðum? Gera tíðnina slembna, þannig að þegar stórt stökk kemur þá eykst tíðnin tímabundið, en er þó miðsækin til lengri tíma litið.

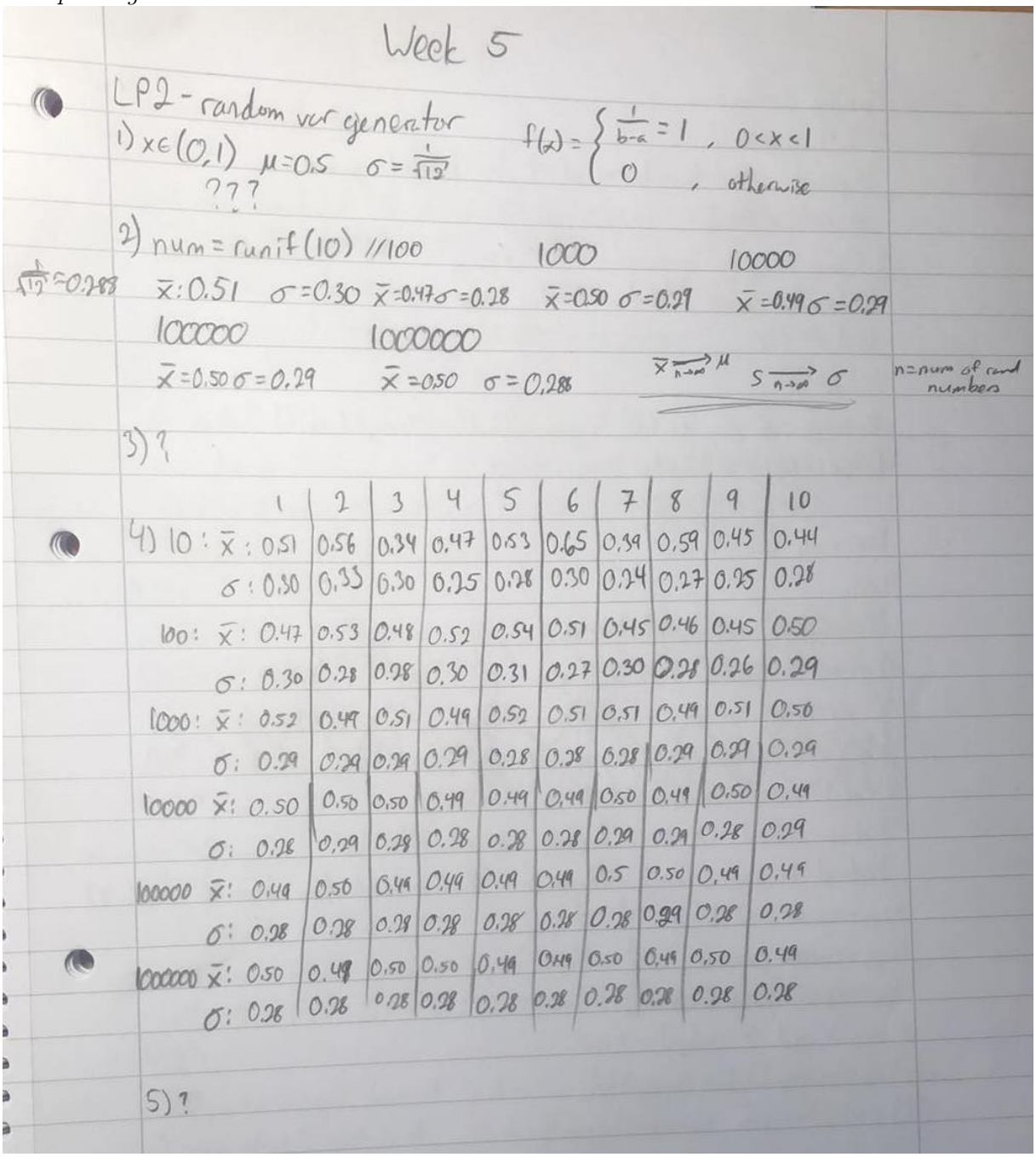
Estimate proportion and uncertainty:
$\sigma_{p_A} = \sqrt{\frac{p_A(1-p_A}{n}}$ where $p_A = \frac{x}{n}$ and n is the total number of things in the sample.

5.2 R code

```
dbinom(2, 4, 0.6)              # P(X=2), Bin(4, 0.6)
1-pbinom(2, 8, 0.2)            # P(X>2), Bin(8, 0.2)
pbinom(2, 5, 0.4)              # P(X<=2), Bin(5, 0.4)
sum(dbinom(3:5, 6, 0.7))       # P(3<=X<=5), Bin(6, 0.7)
num = runif(10)      #generate 10 random numbers
```

5.3 Example Problem Set

Computer generated random numbers



Week 5

LP2 - random var generator

1) $x \in (0,1)$ $\mu = 0.5$ $\sigma = \frac{1}{\sqrt{12}}$??? $f(x) = \begin{cases} \frac{1}{b-a} = 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$

2) num = runif(10) //100 1000 10000

$\frac{1}{\sqrt{12}} = 0.288$ $\bar{x}$: 0.51 $\sigma = 0.30$ $\bar{x} = 0.47$ $\sigma = 0.28$ $\bar{x} = 0.50$ $\sigma = 0.29$ $\bar{x} = 0.49$ $\sigma = 0.29$

100000 1000000

$\bar{x} = 0.50$ $\sigma = 0.29$ $\bar{x} = 0.50$ $\sigma = 0.288$ $\bar{x} \xrightarrow[n \to \infty]{} \mu$ $s \xrightarrow[n \to \infty]{} \sigma$ n = num of rand numbers

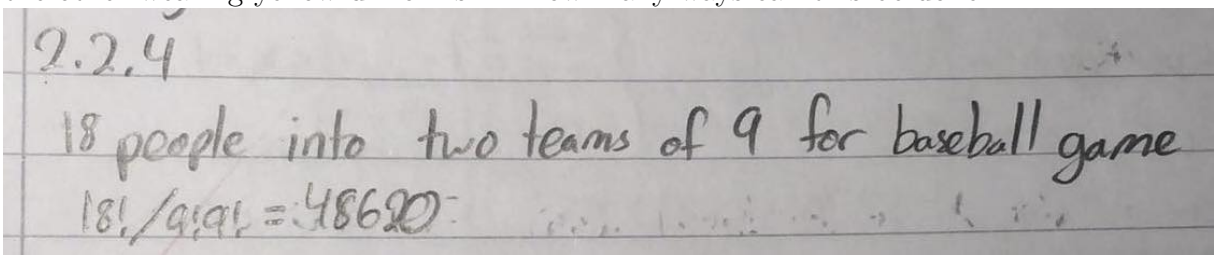
3) ?

	1	2	3	4	5	6	7	8	9	10
4) 10: $\bar{x}$:	0.51	0.56	0.34	0.47	0.53	0.65	0.39	0.59	0.45	0.44
σ:	0.30	0.33	0.30	0.25	0.28	0.30	0.24	0.27	0.25	0.28
100: $\bar{x}$:	0.47	0.53	0.48	0.52	0.54	0.51	0.45	0.46	0.45	0.50
σ:	0.30	0.28	0.28	0.30	0.31	0.27	0.30	0.28	0.26	0.29
1000: $\bar{x}$:	0.52	0.49	0.51	0.49	0.52	0.51	0.51	0.49	0.51	0.50
σ:	0.29	0.29	0.29	0.29	0.28	0.28	0.28	0.29	0.29	0.29
10000 $\bar{x}$:	0.50	0.50	0.50	0.49	0.49	0.49	0.50	0.49	0.50	0.49
σ:	0.28	0.29	0.28	0.28	0.28	0.28	0.29	0.29	0.28	0.29
100000 $\bar{x}$:	0.49	0.50	0.49	0.49	0.49	0.49	0.5	0.50	0.49	0.49
σ:	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.29	0.28	0.28
100000 $\bar{x}$:	0.50	0.49	0.50	0.50	0.49	0.49	0.50	0.49	0.50	0.49
σ:	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28	0.28

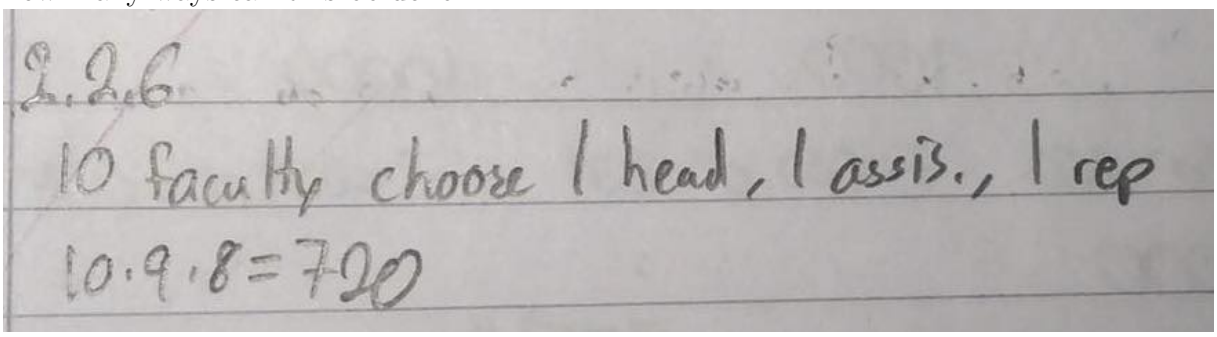
5) ?

Counting combinations

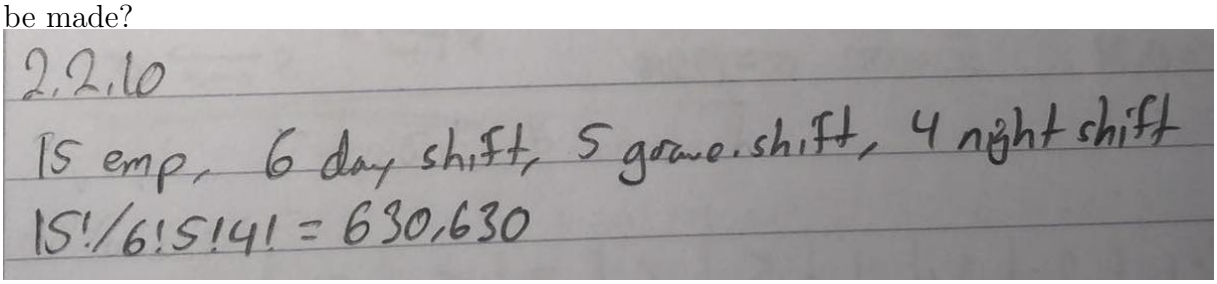
2.2.4 A group of 18 people have gotten together to play baseball. They will divide themselves into two teams of 9 players each, with one team wearing green uniforms and the other wearing yellow uniforms. In how many ways can this be done?



2.2.6 A college math department consisting of 10 faculty members must choose a department head, an assistant department head, and a faculty senate representative. In how many ways can this be done?



2.2.10 A company has hired 15 new employees, and must assign 6 to the day shift, 5 to the graveyard shift, and 4 to the night shift. In how many ways can the assignment be made?



Bernoulli (or binary) variables

4.1.6 Two dice are rolled. Let X = 1 if the dice come up doubles and let X = 0 otherwise. Let Y = 1 if the sum is 6, and let Y = 0 otherwise. Let Z = 1 if the dice come up both doubles and with a sum of 6 (that is, double 3), and let Z = 0 otherwise.

a. Let pX denote the success probability for X. Find p_X.

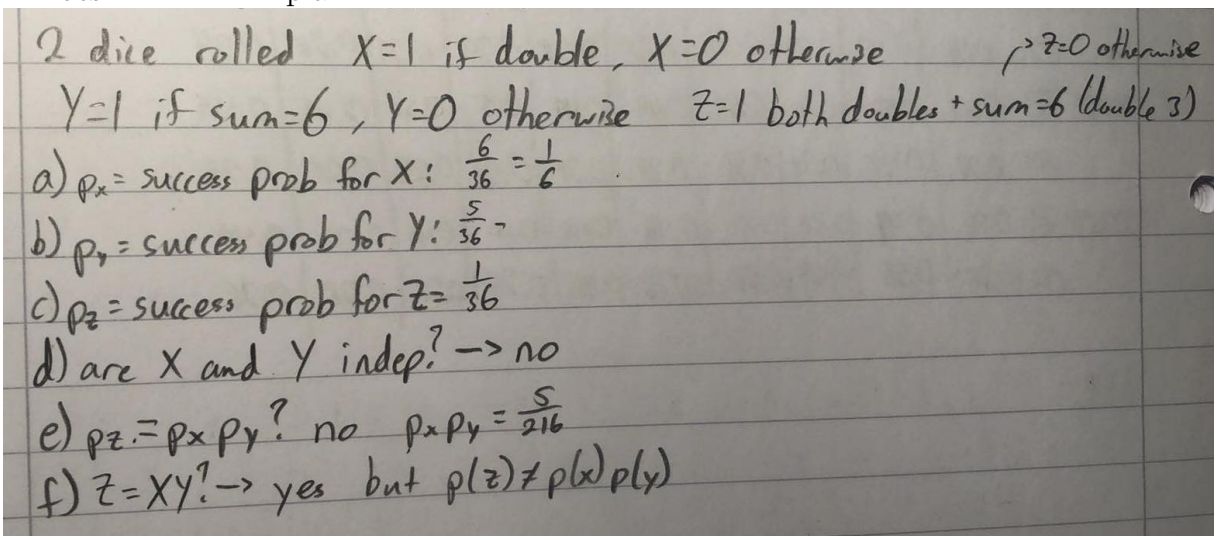
b. Let pY denote the success probability for Y. Find p_Y.

c. Let pZ denote the success probability for Z. Find p_Z.

d. Are X and Y independent?

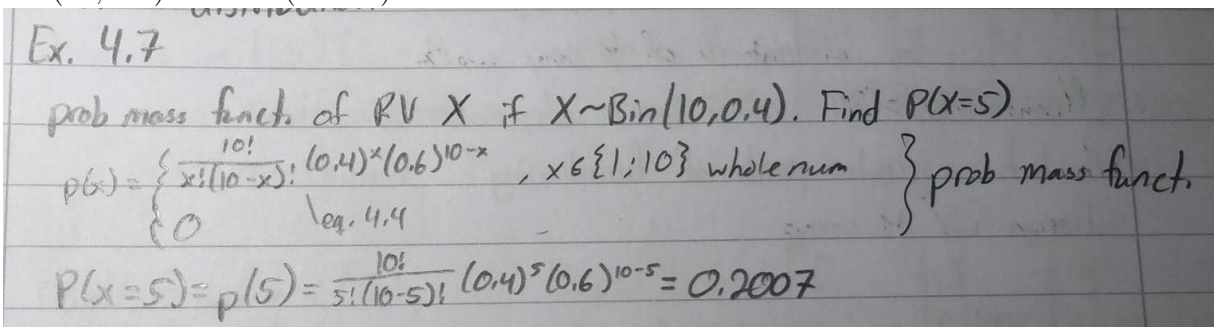
e. Does $p_Z = p_X\ p_Y$?

f. Does Z = XY? Explain.

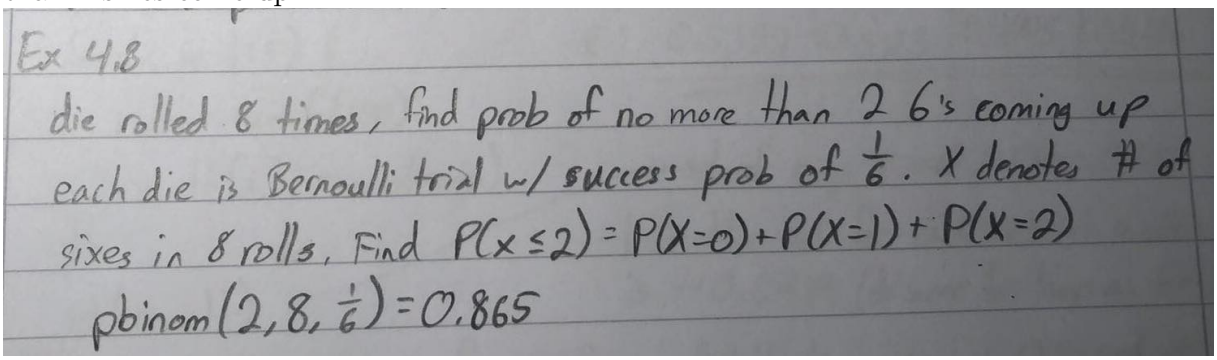


Binomial distribution:

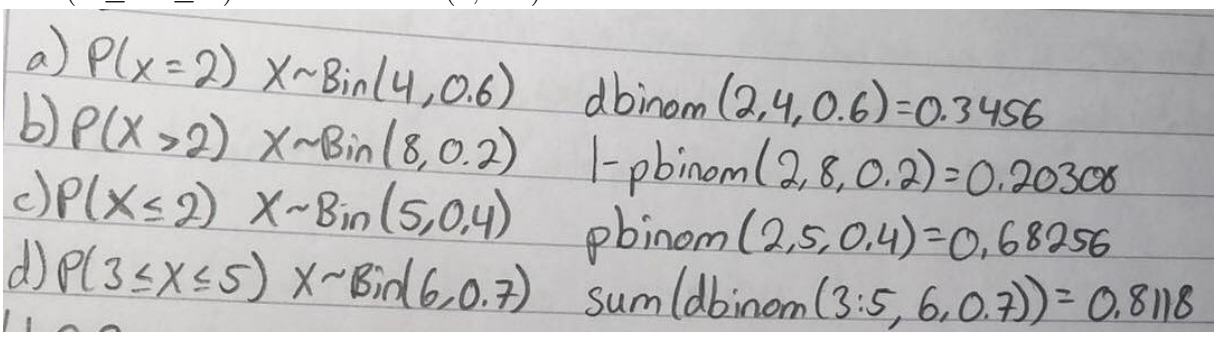
Example 4.7 Find the probability mass function of the random variable X if X ∼ Bin(10, 0.4). Find P(X = 5).



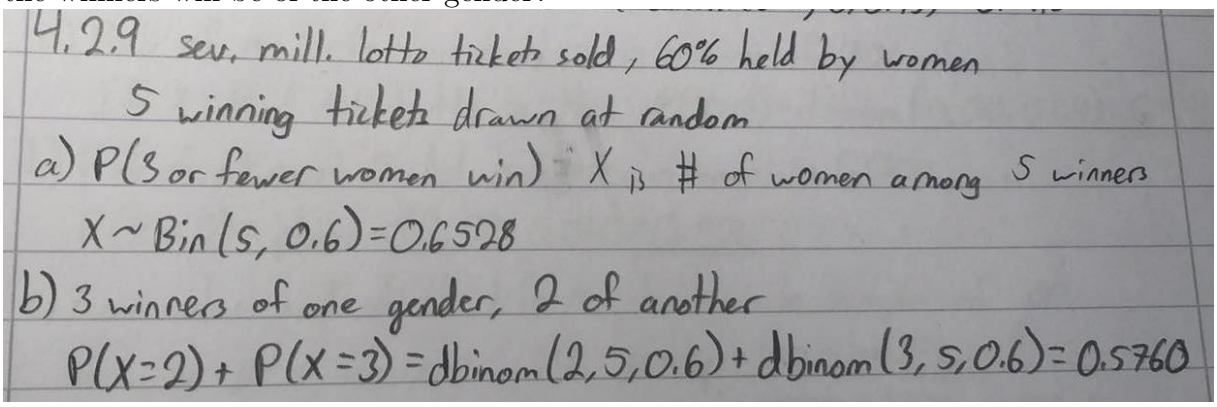
Example 4.8 A fair die is rolled eight times. Find the probability that no more than 2 sixes come up.



4.2.3 Find the following probabilities:
a. P(X = 2) when X ∼ Bin(4, 0.6)
b. P(X > 2) when X ∼ Bin(8, 0.2)
c. P(X ≤ 2) when X ∼ Bin(5, 0.4)
d. P(3 ≤ X ≤ 5) when X ∼ Bin(6, 0.7)



4.2.9 Several million lottery tickets are sold, and 60% of the tickets are held by women. Five winning tickets will be drawn at random.
a. What is the probability that three or fewer of the winners will be women?
b. What is the probability that three of the winners will be of one gender and two of the winners will be of the other gender?

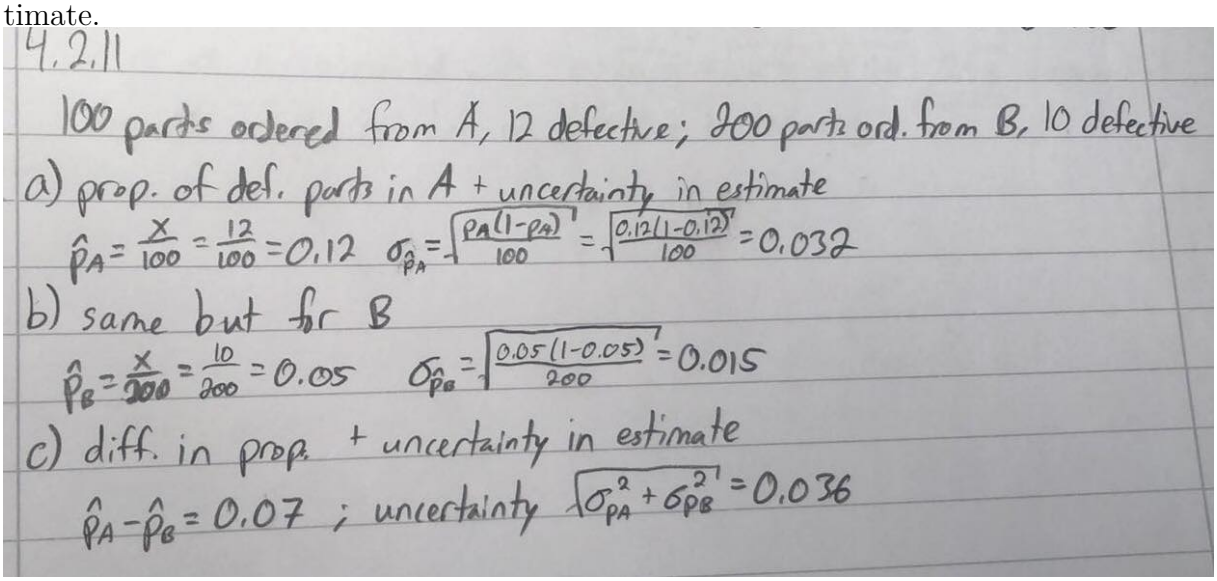


4.2.11 A quality engineer samples 100 steel rods made on mill A and 150 rods made on mill B. Of the rods from mill A, 88 meet specifications, and of the rods from mill B, 135 meet specifications.

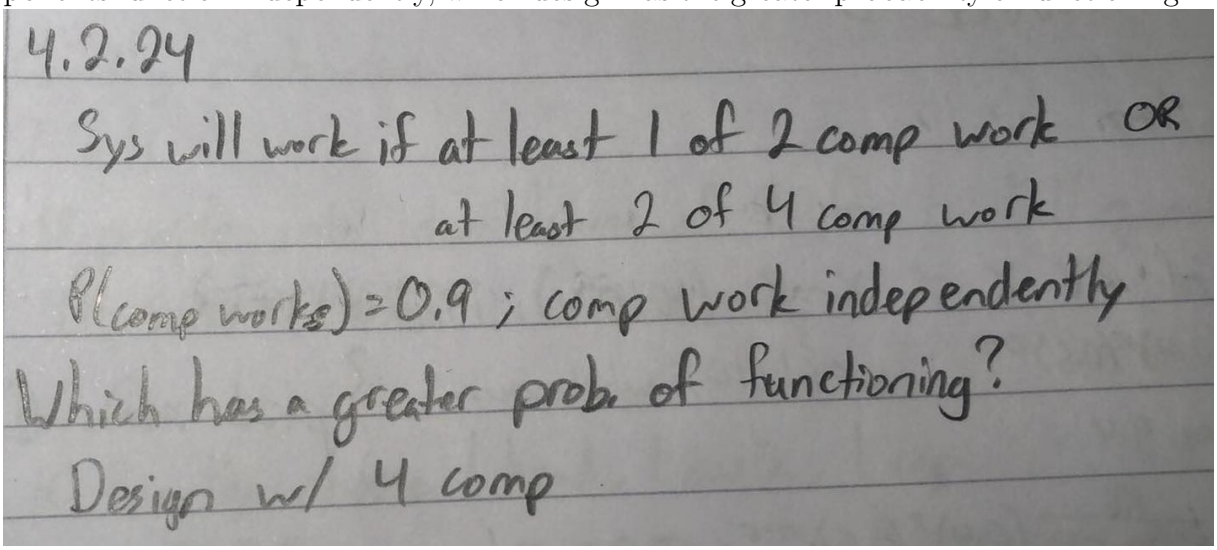
a. Estimate the proportion of rods from mill A that meet specifications, and find the uncertainty in the estimate.

b. Estimate the proportion of rods from mill B that meet specifications, and find the uncertainty in the estimate.

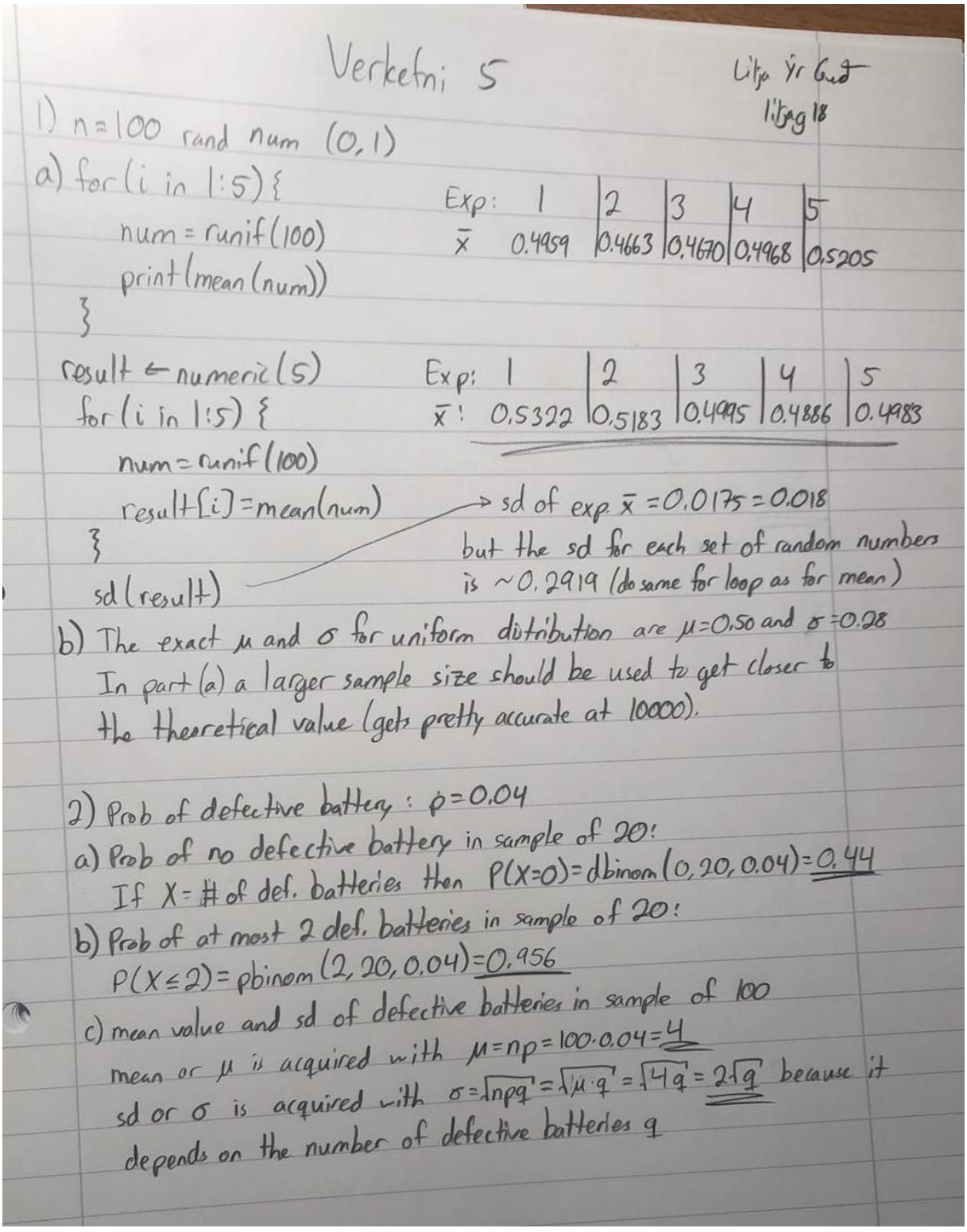
c. Estimate the difference between the proportions, and find the uncertainty in the estimate.



4.2.24 One design for a system requires the installation of two identical components. The system will work if at least one of the components works. An alternative design requires four of these components, and the system will work if at least two of the four components work. If the probability that a component works is 0.9, and if the components function independently, which design has the greater probability of functioning?



5.4 Problem Set



Verkefni 5

Lilja Ýr Guð
líbag 18

1) n=100 rand num (0,1)

a)
```
for(i in 1:5){
    num = runif(100)
    print(mean(num))
}
```

Exp:	1	2	3	4	5
$\bar{x}$	0.4959	0.4663	0.4670	0.4968	0.5205

```
result ← numeric(5)
for(i in 1:5){
    num = runif(100)
    result[i] = mean(num)
}
sd(result)
```

Exp:	1	2	3	4	5
$\bar{x}$:	0.5322	0.5183	0.4995	0.4886	0.4983

→ sd of exp. $\bar{x}$ = 0.0175 = 0.018
but the sd for each set of random numbers is ~0.2919 (do same for loop as for mean)

b) The exact μ and σ for uniform distribution are $\mu = 0.50$ and $\sigma = 0.28$
In part (a) a larger sample size should be used to get closer to the theoretical value (gets pretty accurate at 10000).

2) Prob of defective battery: $p = 0.04$

a) Prob of no defective battery in sample of 20:
If X = # of def. batteries then $P(X=0) = \text{dbinom}(0, 20, 0.04) = 0.44$

b) Prob of at most 2 def. batteries in sample of 20:
$P(X \leq 2) = \text{pbinom}(2, 20, 0.04) = 0.956$

c) mean value and sd of defective batteries in sample of 100
mean or μ is acquired with $\mu = np = 100 \cdot 0.04 = 4$
sd or σ is acquired with $\sigma = \sqrt{npq} = \sqrt{\mu \cdot q} = \sqrt{4q} = 2\sqrt{q}$ because it depends on the number of defective batteries q

5.5 Problem Set Solution

1.a) Do this computer experiment: Generate with your favourite software a sample of n=100 random numbers with uniform distribution in the interval (0,1) and obtain their sample mean $\bar{X}$. Call this Experiment 1 and put the value of $\bar{X}$ in a small table,as shown below. Do five more such experiments and complete the table. Use at least three significant digits for each number to be able to compare them properly. Finally, calculate the standard deviation of these six values for $\bar{X}$ seen as a set of six random numbers. Mention briefly the computer functions or commands used.

Experiment	1	2	3	4	5	6	st.dev of $\bar{x}$
$\bar{x}$	0.5010	0.5056	0.5291	0.4746	0.4387	0.5188	0.03306

The table shows an example of results which can be obtained with R or Excel using the models posted in Canvas.
b) What is the theoretically expected standard deviation (or uncertainty) of the sample mean of n=100 random numbers x∈(0,1)with uniform distribution?(Remember the exact μ and σ for the uniform distribution.) What should be done at point a) to obtain the st.dev.of $\bar{x}$ as close as possible to the theoretically expected value
The theoretical standard deviation: $\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{12}}\frac{1}{\sqrt{100}} = 0.0289$. See also Verkefni 4 problem 2 and Miniproject (Phase 1) point 5.
It is assumed that the student practiced such exercises before the assignment and understood the idea of randomness and the effects of the sample size.It is also assumed that the student distinguishes two types of samples in this problem: 6 samples of n=100 numbers with uniform distribution, and one sample of 6 numbers $\bar{X}$ representing their sample means.In order to obtain the st.dev. of the $\bar{x}$'s as close as possible to 0.0289 one has to increase the sample of $\bar{x}$'s, or the number of "experiments", from 6 to a lot more. To how many more? The answer will be given in Week 7.

2. Consider a brand of electric batteries which are defective with probability p = 0.04. Calculate:
a) The probability that there is no defective battery in a sample of 20. dbinom(0,20,p)=(1-p)**20=0.442 (or BINOM.DIST in Excel with FALSE option)
b)The probability that at most two batteries are defectivein a sample of 20.pbinom(2,20,p)=0.956 (or BINOM.DIST in Excel with TRUE option)
c) The mean value and the standard deviation of the number of defective batteries in a sample of 100.mean=np=4 std=$\sqrt{}$np(1-p)=1.96

6 Week 2

6.1 Concepts

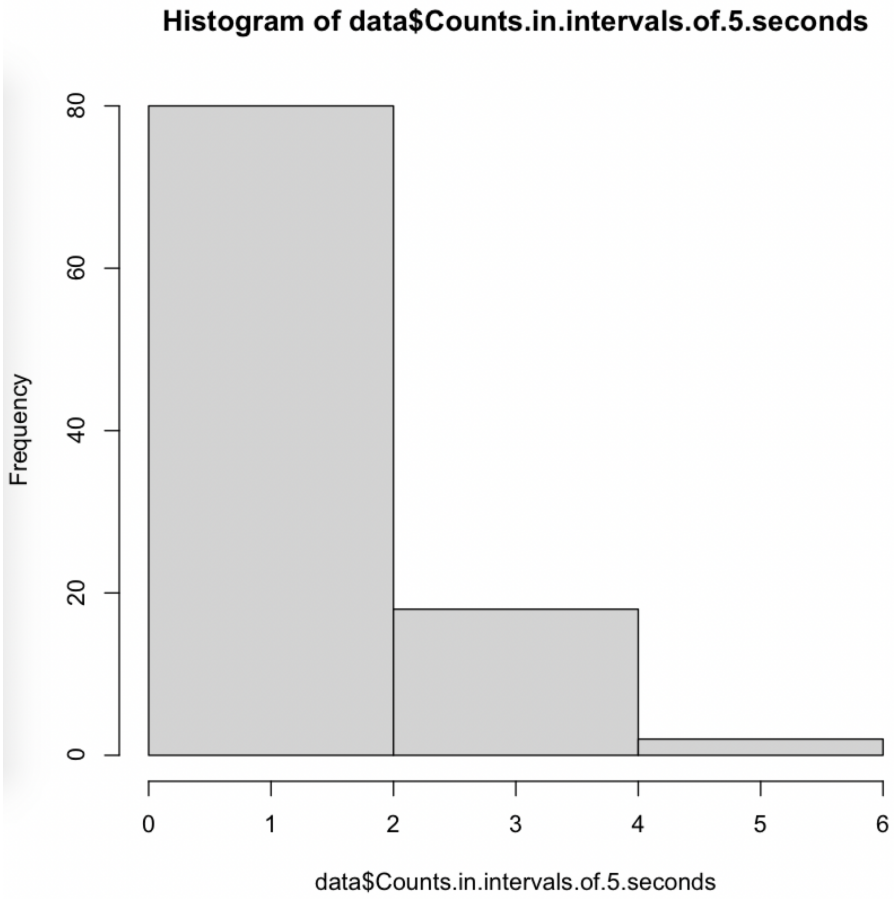
- The area under the normal probability distribution function with parameters mu and sigma is equal to 1
- When sigma increases the normal probability distribution function becomes wider
- Consider $x \sim N(0,1)$. The probability that $x < -1$ is 16%
- Consider $x \sim N(0,1)$. The probability that $-1 < x < 1$ is 68%
- Consider a sample of n random numbers $x_1, x_2, ..., x_n$ with a distribution D. If n is large enough their sample mean follows a normal distribution irrespective of the distribution D
- A Poisson distribution with a large parameter lambda can be approximated with a normal distribution having the standard deviation sqrt (lambda)
- The mean value of a random variable having a lognormal distribution with parameters $\mu = 0$ and sigma=1 is 1.65
- The distribution of the body mass index is lognormal
- Consider x a random variable having an exponential distribution with parameter lambda=3. The probability that x is less or equal to the mean value is 0.63
- Hvaða líkindadreifingar eru oftast notaðar í jarðvarmamati með Monte-Carlo hermun? Jöfn dreifing og þríhyrningslaga dreifing

6.2 R code

```
hist(data\$Counts.in.intervals.of.5.seconds, 4)w = table(data\$Counts.in.i
for (i in w) print(ppois(i, 1.47))
dpois(1,6) + dpois(0,6)  #P(X<=1)
dbinom(1, 0.2)  #same as dpois(1, 0.2)
1 - pnorm(-0.85,0,1)     #area under normal curve right of z--0.85
1-ppois(5,5)     #poisson distribution
n=10
z=vector()
z=0
k=100
for(i in 1:k){z[i]=mean(runif(n,0,1))}
curve(dnorm(x, mean=mean(z), sd=sd(z)), col="red", add=T)

qnorm(0.25,mean,sd)       #first quartile
```

6.3 Example Problem Set

Poisson distribution: Poisson distribution of radiation:Put the data on background radiation uploaded in Canvas (Modules/Other_Materials/Background_radiation.csv) on a histogram, calculate the frequencies predicted by the Poisson distribution with λ equal to the observed mean value, create another histogram. Compare the experimental and the expected histograms.See video

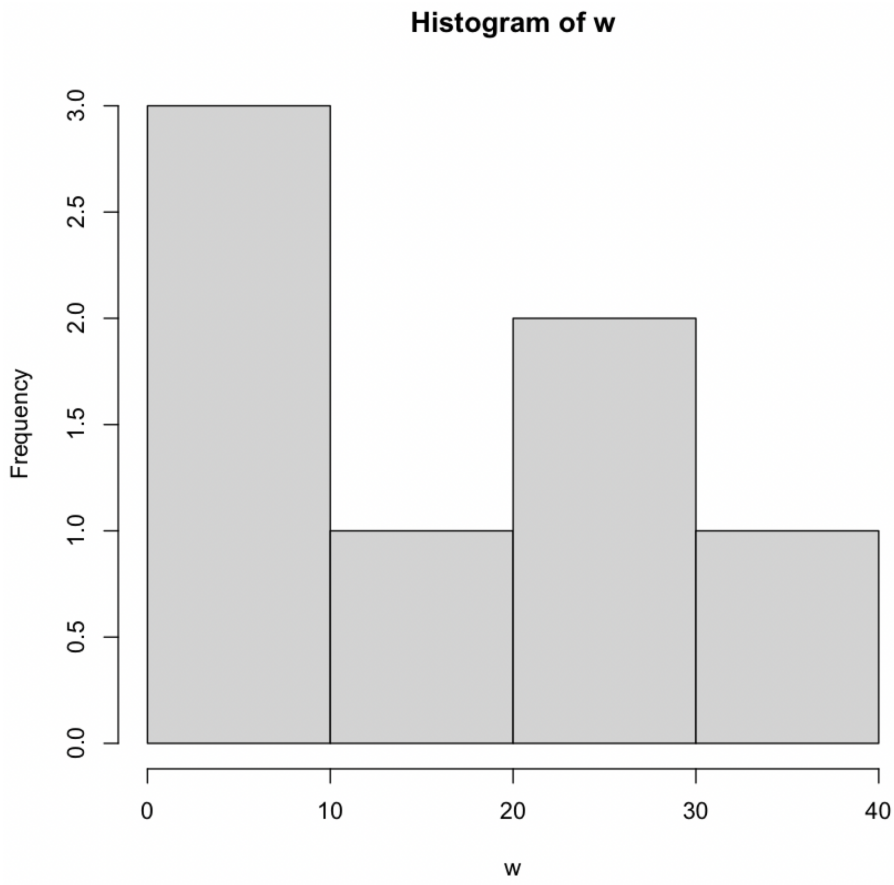


hist(data$Counts.in.intervals.of.5.seconds, 4)
w = table(data$Counts.in.intervals.of.5.seconds)

```
> hist(data$Counts.in.intervals.of.5.seconds, 4)
> w = table(data$Counts.in.intervals.of.5.seconds)
> w

 0  1  2  3  4  5  6
24 33 23 15  3  1  1
```

$\lambda = mean = 1.47$
for (i in w) print(ppois(i, 1.47))

```
> for (i in w){ print(ppois(i, 1.47)) }
[1] 1
[1] 1
[1] 1
[1] 1
[1] 0.9380662
[1] 0.5679159
[1] 0.5679159
```



4.3.3 In a certain city, the number of potholes on a major street follows a Poisson distribution with a rate of 3 per mile. Let X represent the number of potholes in a two-mile stretch of road. Find:

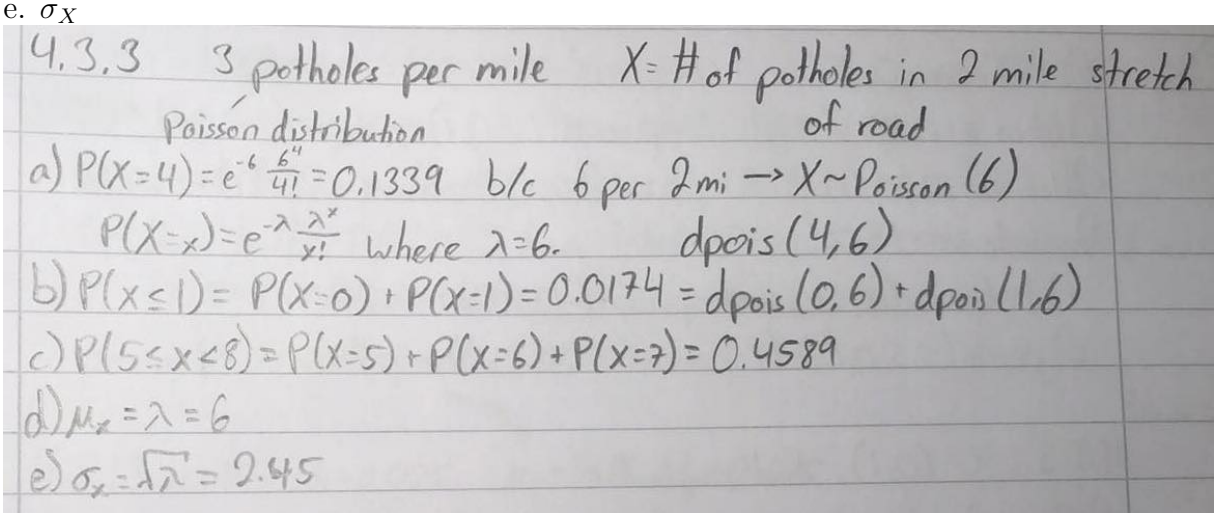
a. $P(X = 4)$

b. $P(X \leq 1)$

c. $P(5 \leq X < 8)$

d. μ_X

e. σ_X



4.3.6 One out of every 5000 individuals in a population carries a certain defective gene. A random sample of 1000 individuals is studied.

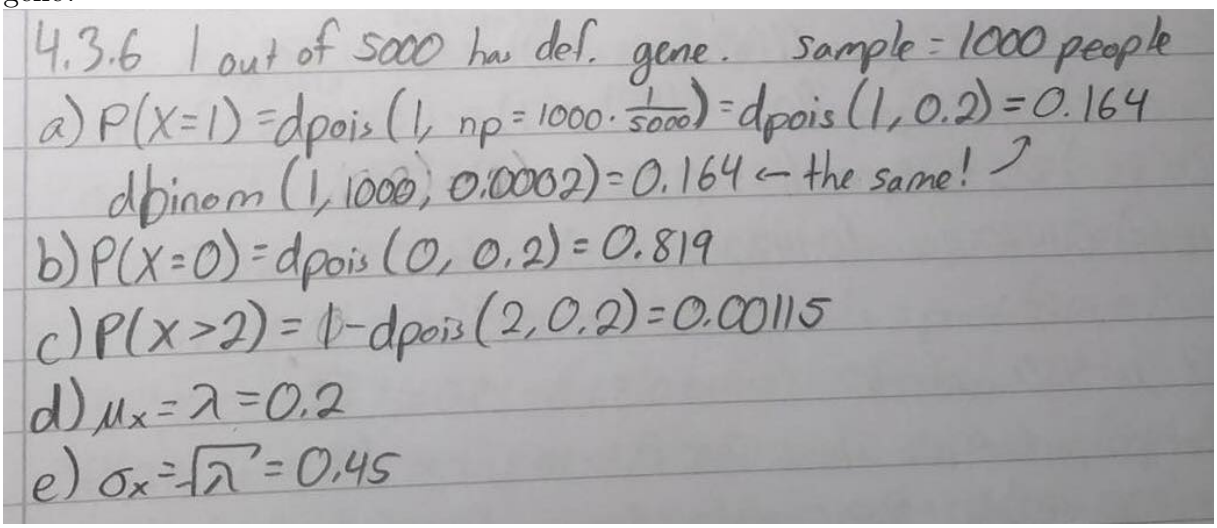
a. What is the probability that exactly one of the sample individuals carries the gene?

b. What is the probability that none of the sample individuals carries the gene?

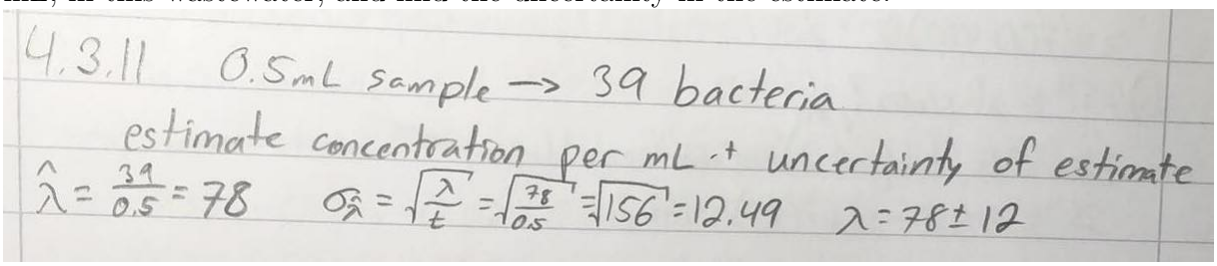
c. What is the probability that more than two of the sample individuals carry the gene?

d. What is the mean of the number of sample individuals that carry the gene?

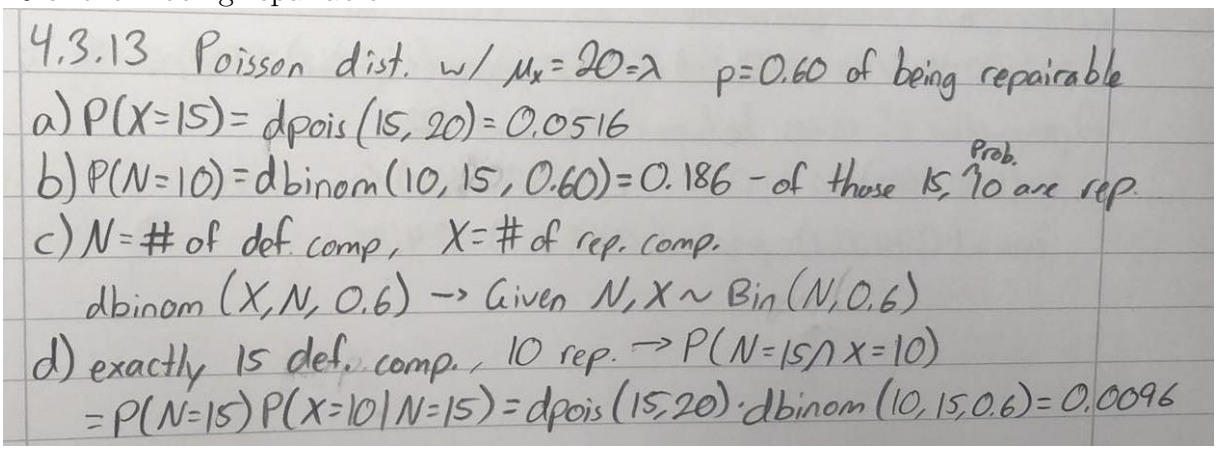
e. What is the standard deviation of the number of sample individuals that carry the gene?



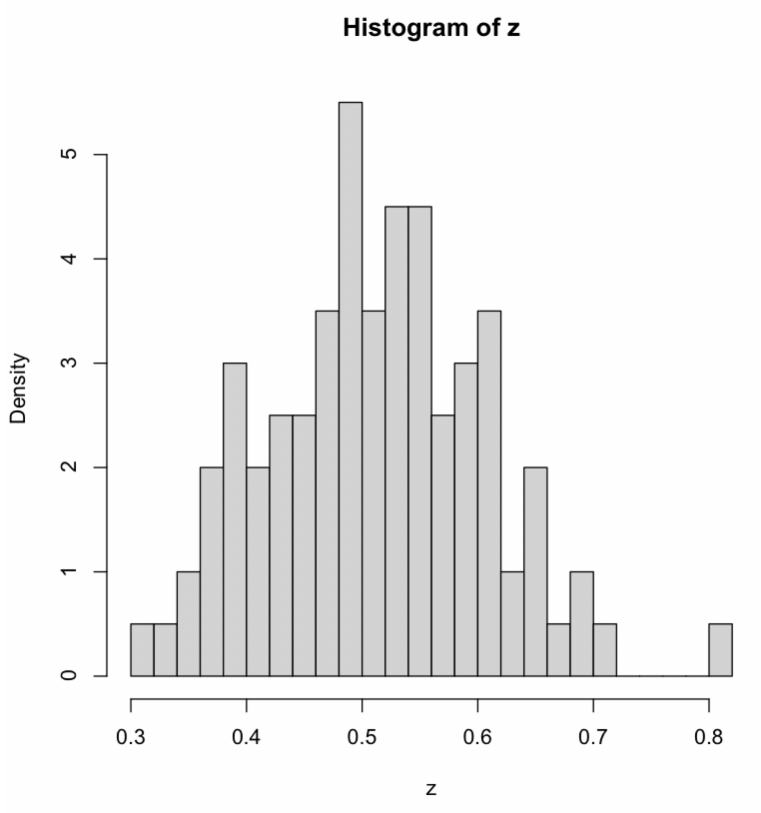
4.3.11 A microbiologist wants to estimate the concentration of a certain type of bacterium in a waste-water sample. She puts a 0.5 mL sample of the wastewater on a microscope slide and counts 39 bacteria. Estimate the concentration of bacteria, per mL, in this wastewater, and find the uncertainty in the estimate.



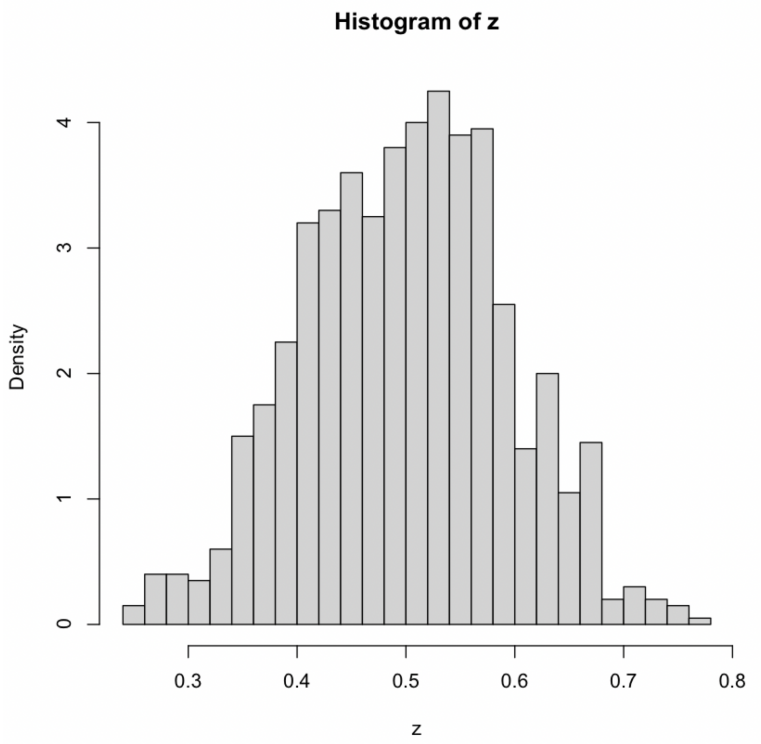
4.3.13 The number of defective components produced by a certain process in one day has a Poisson distribution with mean 20. Each defective component has probability 0.60 of being repairable.
a. Find the probability that exactly 15 defective components are produced.
b. Given that exactly 15 defective components are produced, find the probability that exactly 10 of them are repairable.
c. Let N be the number of defective components produced, and let X be the number of them that are repairable. Given the value of N, what is the distribution of X?
d. Find the probability that exactly 15 defective components are produced, with exactly 10 of them being repairable.



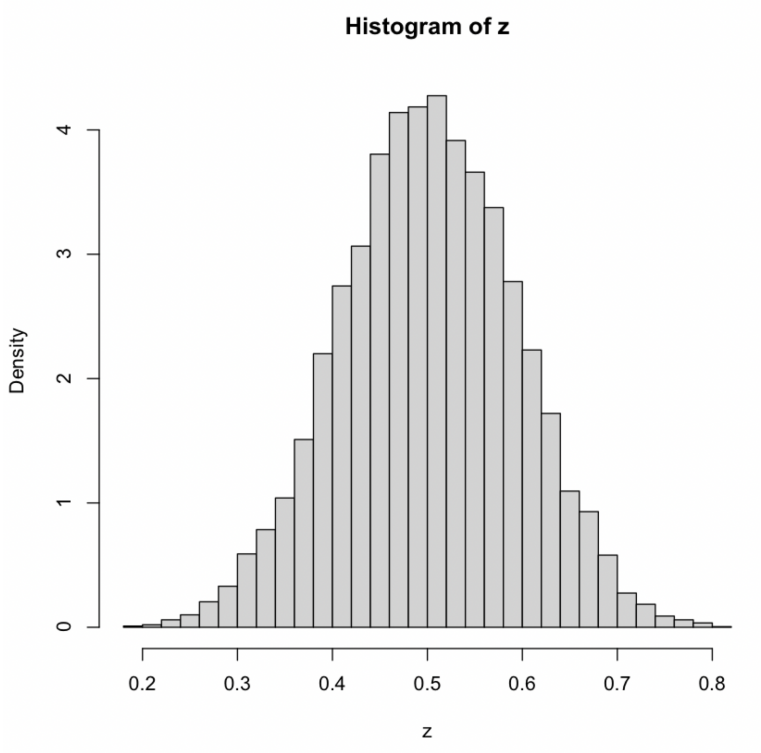
Normal distribution:
Miniproject Phase 2. Prove the central limit theorem with a numerical experiment.


Histogram of z
Density
0
1
2
3
4
5
0.3
0.4
0.5
0.6
0.7
0.8
z

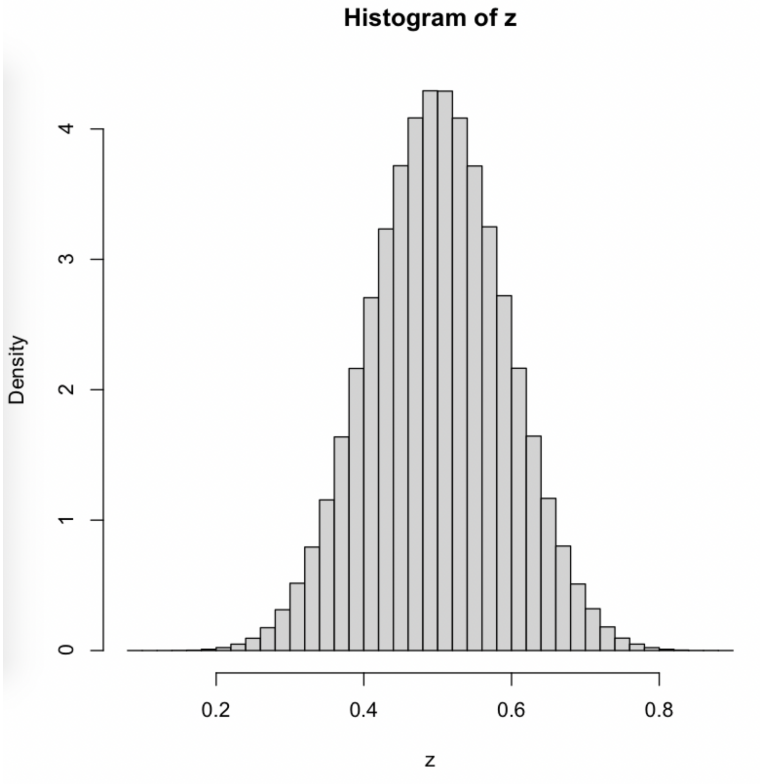
k=100


Histogram of z
Density
0
1
2
3
4
0.3
0.4
0.5
0.6
0.7
0.8
z

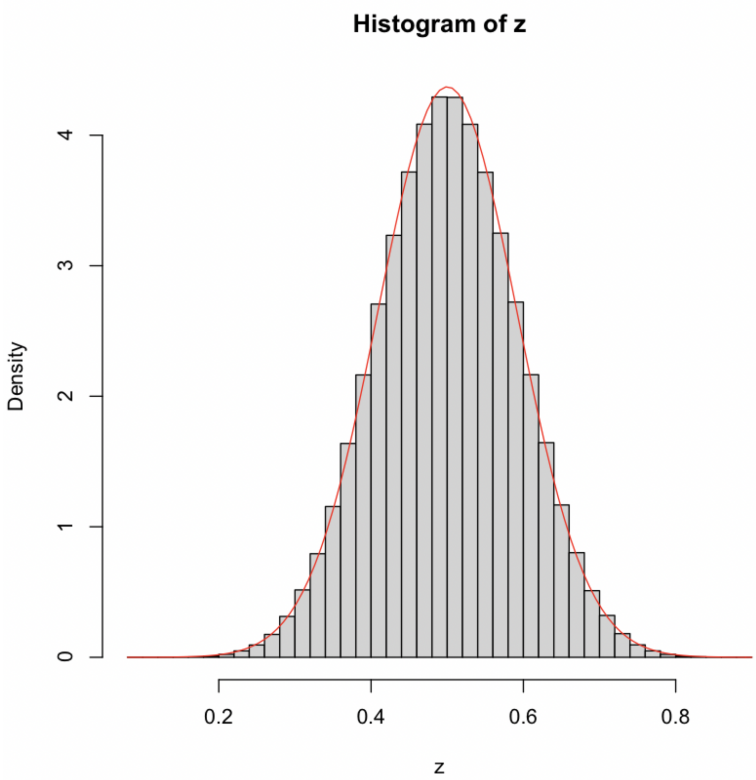
k=1000



k=10000



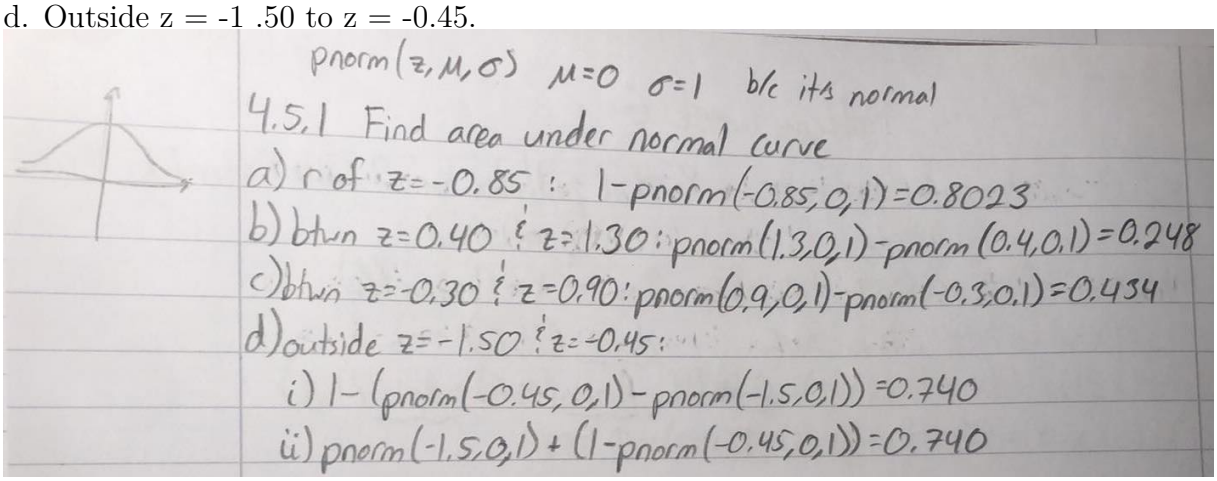
k=1000000



k=1000000 with red curve - curve(dnorm(x, mean=mean(z), sd=sd(z)), col="red", add=T) This shows teh central limit theorem because with larger sample sizes (k) the more normalized the data looks.

4.5.1 Find the area under the normal curve

a. To the right of z = -0.85.
b. Between z = 0.40 and z = 1.30.
c. Between z = -0.30 and z = 0.90.
d. Outside z = -1 .50 to z = -0.45.



4.5.3 Let $Z \sim N(0, 1)$. Find a constant c for which

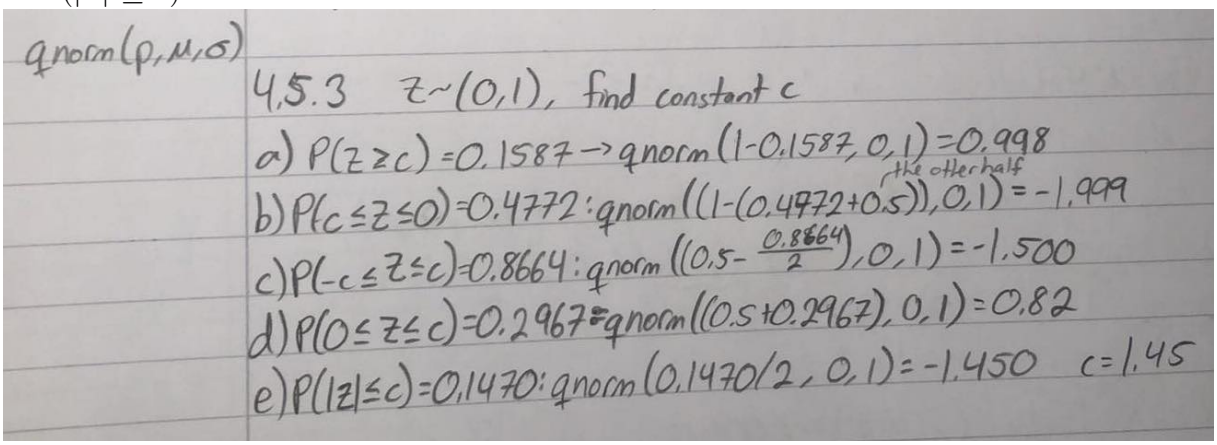
a. $P(Z \geq c) = 0.1587$

b. $P(c \leq Z \leq 0) = 0.4772$

c. $P(-c \leq Z \leq c) = 0.8664$

d. $P(0 \leq Z \leq c) = 0.2967$

e. $P(|Z| \leq c) = 0.1470$



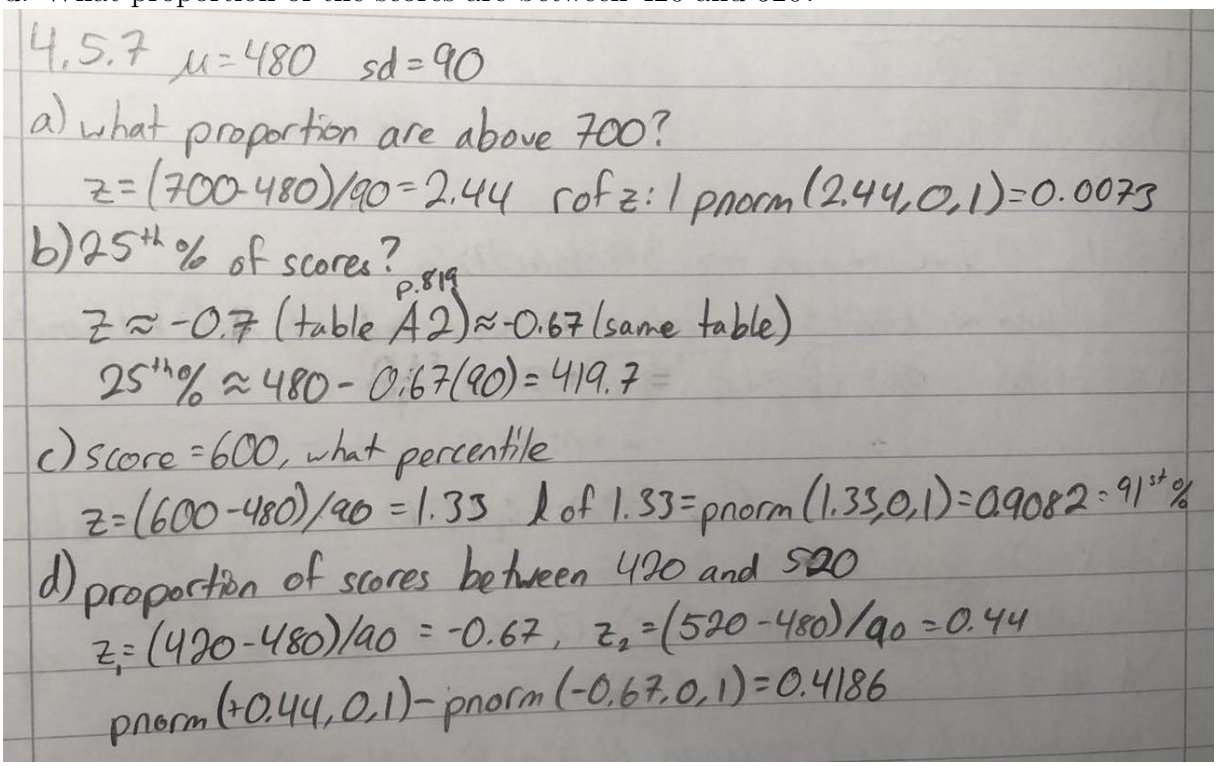
4.5.7 Scores on a standardized test are approximately normally distributed with a mean of 480 and a standard deviation of 90.

a. What proportion of the scores are above 700?

b. What is the 25th percentile of the scores?

c. If someone's score is 600, what percentile is she on?

d. What proportion of the scores are between 420 and 520?



4.5.9 The lifetime of a lightbulb in a certain application is normally distributed with mean $\mu = 1400$ hours and standard deviation $\sigma = 200$ hours.

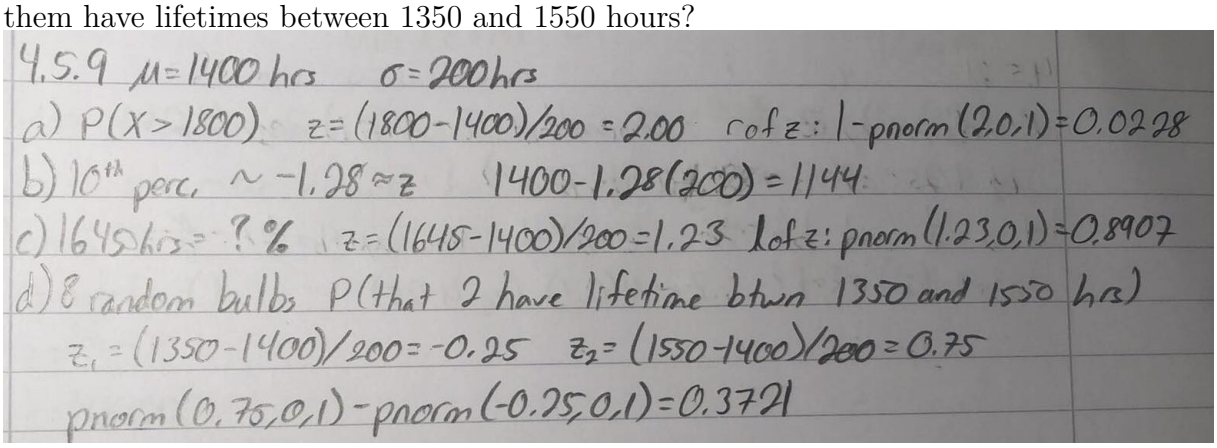
a. What is the probability that a lightbulb will last more than 1800 hours?

b. Find the 10th percentile of the lifetimes.

c. A particular lightbulb lasts 1645 hours. What percentile is its lifetime on?

d. What is the probability that the lifetime of a light-bulb is between 1350 and 1550 hours?

e. Eight lightbulbs are chosen at random. What is the probability that exactly two of them have lifetimes between 1350 and 1550 hours?



4.5.13 A cylindrical hole is drilled in a block, and a cylindrical piston is placed in the hole. The clearance is equal to one-half the difference between the diameters of the hole and the piston. The diameter of the hole is normally distributed with mean 15 cm and standard deviation 0.025 cm, and the diameter of the piston is normally distributed with mean 14.88 cm and standard deviation 0.015 cm. The diameters of hole and piston are independent.

a. Find the mean clearance.

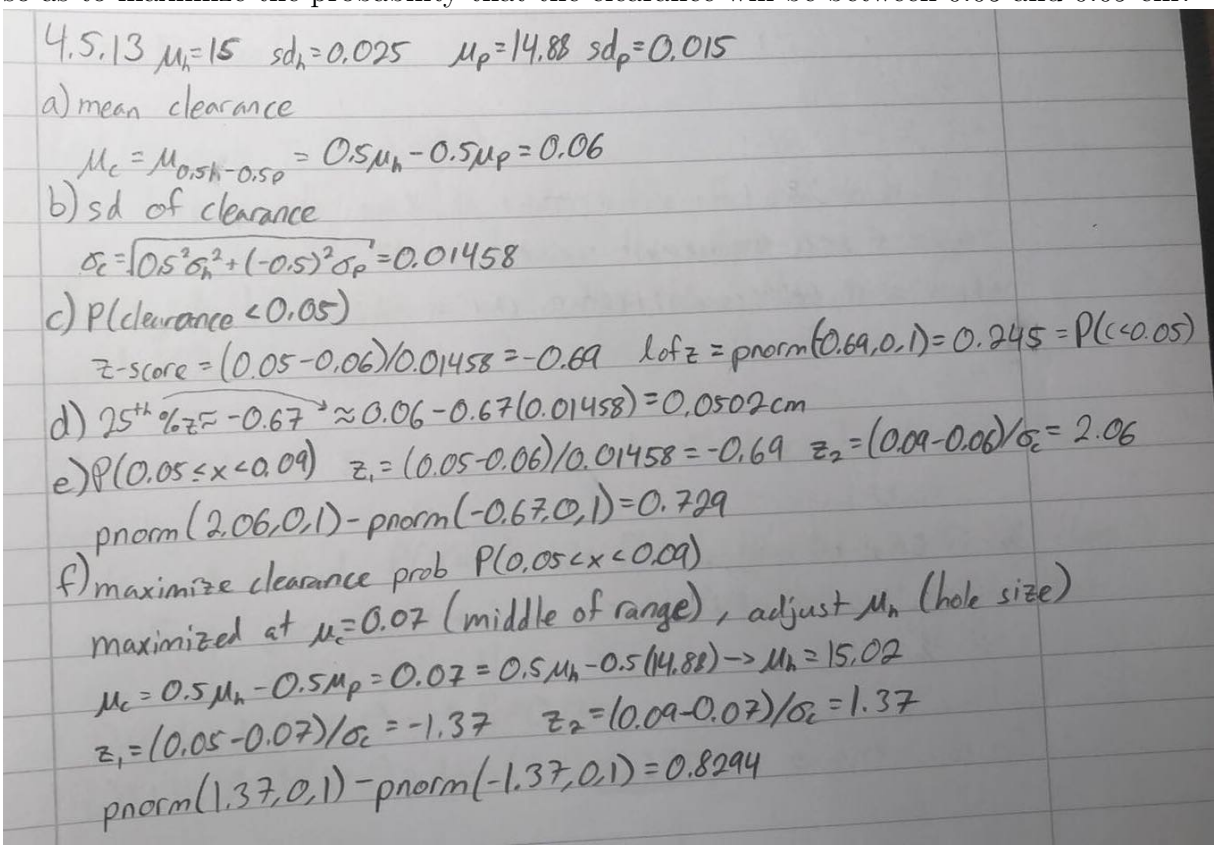
b. Find the standard deviation of the clearance.

c. What is the probability that the clearance is less than 0.05 cm?

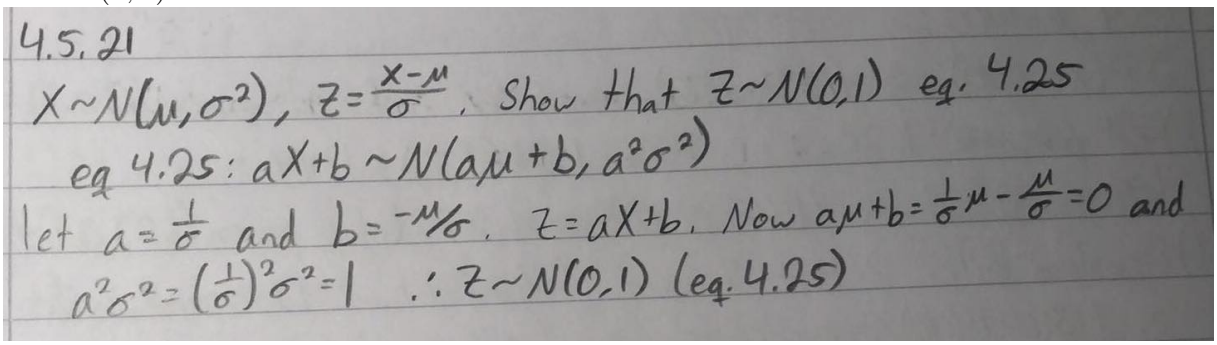
d. Find the 25th percentile of the clearance.

e. Specifications call for the clearance to be between 0.05 and 0.09 cm. What is the probability that the clearance meets the specification?

f. It is possible to adjust the mean hole diameter. To what value should it be adjusted so as to maximize the probability that the clearance will be between 0.05 and 0.09 cm?



4.5.21 Let X $\sim N(\mu, \sigma^2)$, and let $Z = (X - \mu)/\sigma$.Use Equation (4.25) to show that $Z \sim N(0, 1)$.

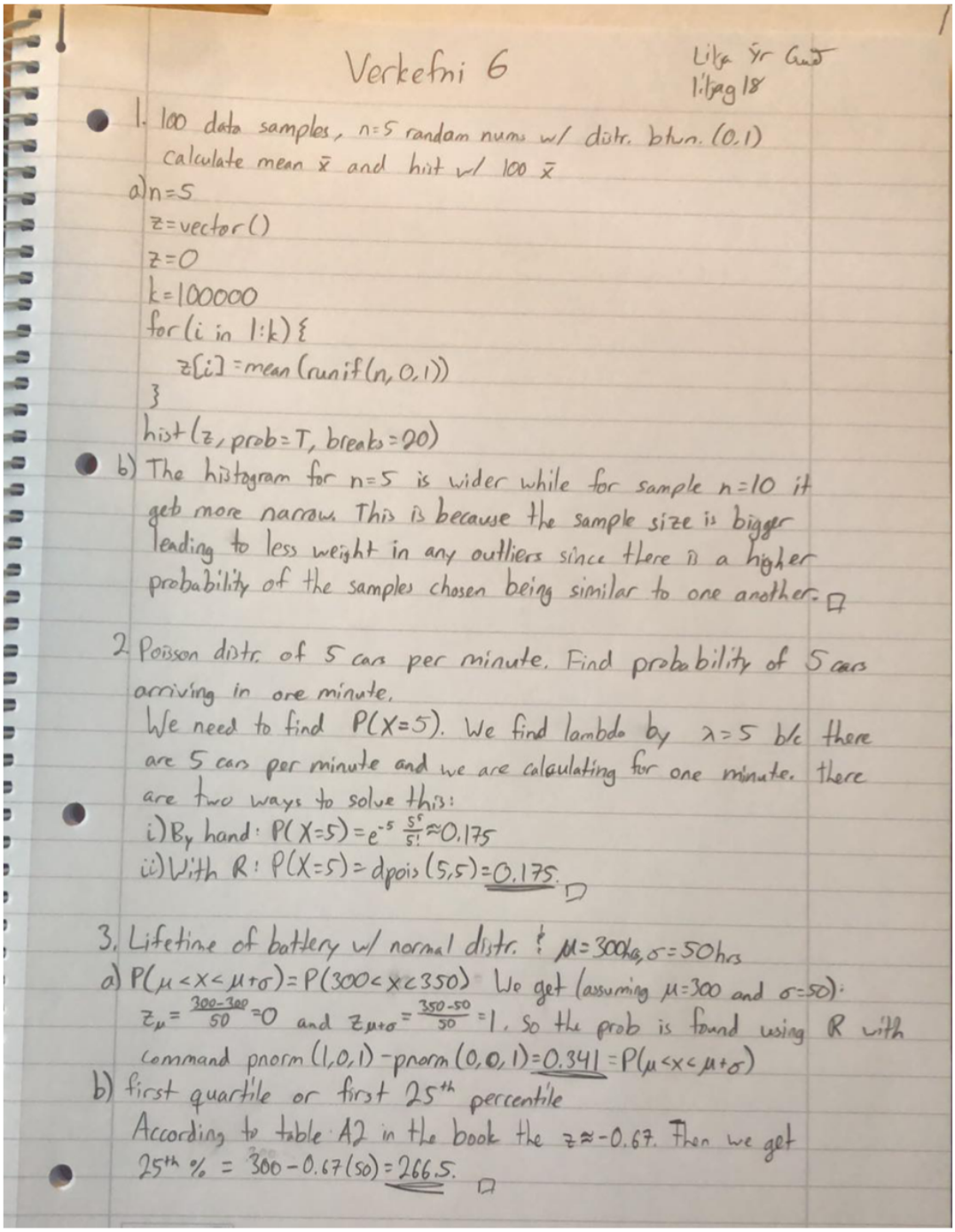


6.4 Problem Set

Lilja Ýr Guðmundsdóttir - liljag18

Verkefni 6

Here are the solutions, the histograms for question 1 are on the next page.



Verkefni 6 — Lilja Ýr Guð — liljag18

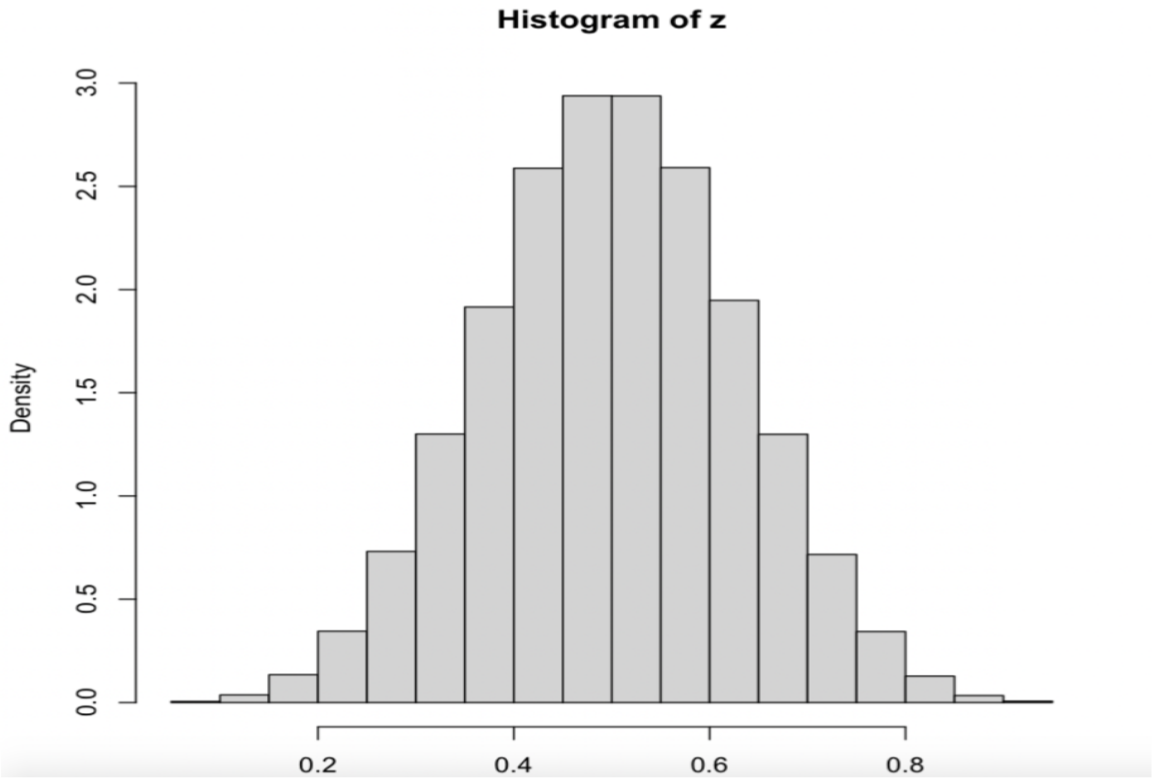
1. 100 data samples, n=5 random nums w/ distr. btwn. (0,1) calculate mean $\bar{x}$ and hist w/ 100 $\bar{x}$

a) n=5
```
z=vector()
z=0
k=100000
for(i in 1:k){
  z[i]=mean(runif(n,0,1))
}
hist(z,prob=T,breaks=20)
```

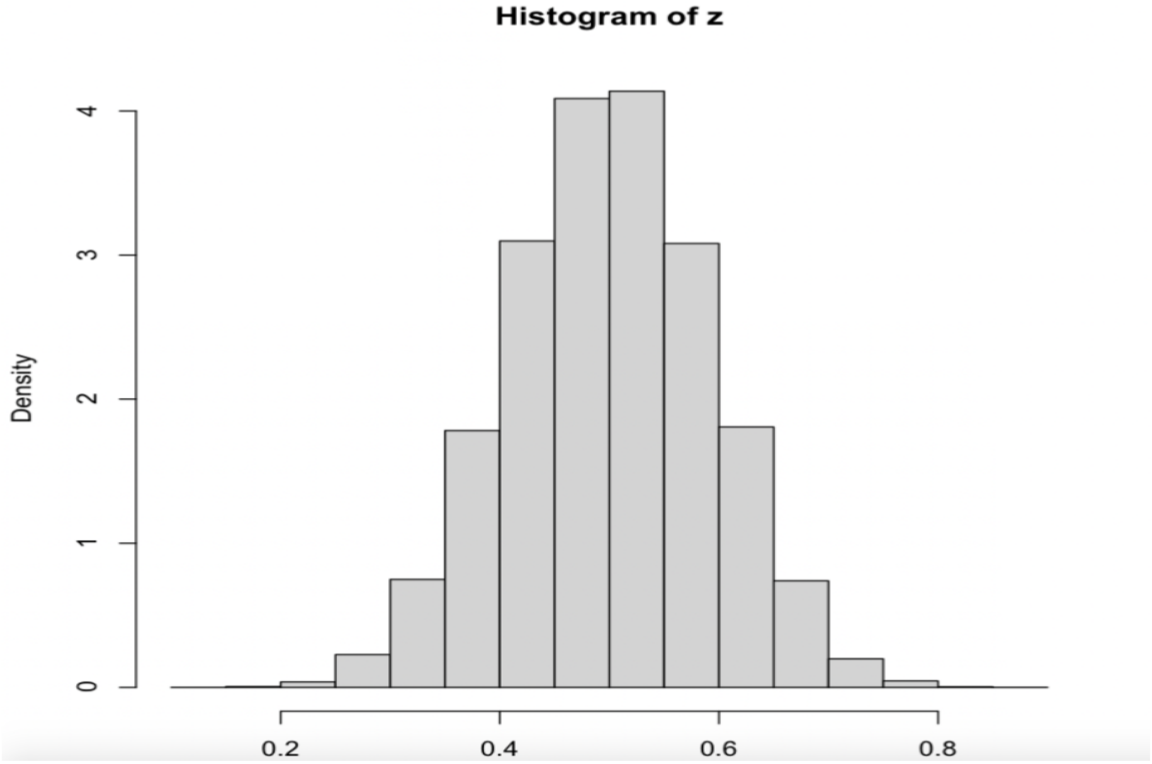
b) The histogram for n=5 is wider while for sample n=10 it gets more narrow. This is because the sample size is bigger leading to less weight in any outliers since there is a higher probability of the samples chosen being similar to one another. □

2. Poisson distr. of 5 cars per minute. Find probability of 5 cars arriving in one minute.
We need to find $P(X=5)$. We find lambda by $\lambda=5$ b/c there are 5 cars per minute and we are calculating for one minute. there are two ways to solve this:
i) By hand: $P(X=5)=e^{-5}\frac{5^5}{5!}\approx 0.175$
ii) With R: $P(X=5)=$ dpois(5,5) $=0.175$. □

3. Lifetime of battery w/ normal distr. ! $\mu=300$hrs, $\sigma=50$hrs
a) $P(\mu<X<\mu+\sigma)=P(300<X<350)$ We get (assuming $\mu=300$ and $\sigma=50$):
$z_\mu=\frac{300-300}{50}=0$ and $z_{\mu+\sigma}=\frac{350-50}{50}=1$. So the prob is found using R with command pnorm(1,0,1) − pnorm(0,0,1) $=0.341=P(\mu<X<\mu+\sigma)$
b) first quartile or first 25th percentile
According to table A2 in the book the $z\approx -0.67$. Then we get
25th % = 300 − 0.67(50) = 266.5. □



Here is the histogram for n=5.



Here is the histogram for n=10.

6.5 Problem Set Solutions

1. a) Do the following computer experiment: -Use your favourite computer tools to generate 100 data samples, each of n=5 random numbers with uniform distribution between (0,1). For each sample calculate the sample mean $\bar{x}$. Make a histogram with these 100 calculated $\bar{x}$'s.
-Repeat this experiment using n=10 and make a new histogram with the new 100 sample means.
-Include both histograms in the uploaded solution using at least 5 bins and mention briefly the computer functions or commands used.
a) Example of solution: Here I will denote x_bar with zn=5;

```
k=100; z1=0
for(i in 1:k) {z1[i]=mean(runif(n,0,1))}
hist(z1,xlim=c(0,1))

n=10; k=100; z2=0
for(i in 1:k) {z2[i]=mean(runif(n,0,1))}
hist(z2,xlim=c(0,1))
```

b) Which histogram is wider and which one is narrower? Explain why.
The width of the histogram depends on the standard deviation of the 100 numbers.

The larger the std. the wider the histogram, which can be seen as an approximation of a normal distribution.
For n=5 I got sd(z1)=0.118and the "true" value is1/sqrt(12)/sqrt(5)=0.129
For n=10 I got sd(z2)=0.0893 and the "true" value is1/sqrt(12)/(10)=0.0913.
So the histogram for z1is wider than the histogram for z2, because sd(z1) > sd(z2).
Comment 1: Here 2 points are given only if the explanation includes some numbers describing the std. in each case, or the argument that the true std decreases like 1/sqrt(12)/sqrt(n).
Comment2: We know that for larger k, like 1000000,sd(z) calculated with R will be very close to the "true" or "exact" std., both for n=5 and n=10. Try this exercise if you did not do it yet, to consolidate your understanding.

2. The number of cars arriving at an intersection follows a Poisson distribution with a mean rate of 5per minute. Find the probability that more than 5 cars arrive in one minute.

```
1-ppois(5,5)=0.384
```

3. The lifetime of a battery is normally distributed with mean μ=300hours and standard deviation σ=50hours. Calculate:
a) The probability that a battery lasts between μ and $\mu + \sigma$ hours.

```
m=300;  s=50
pnorm(m+s,m,s)-pnorm(m,m,s) = 0.3413
```

b) The first quartile of the lifetimes.

```
qnorm(0.25,m,s) = 266.28
```

7 Week 7

7.1 Concepts

- An estimator is biased if its mean value is different from the exact value
- An estimated value of the probability parameter p of a binomial distribution is the observed number of successes divided by the number of trials n. The mean squared error of this estimate is p(1-p)/n
- The likelihood function is the probability of the observed data as function of a parameter.
- The maximum likelihood estimator of the parameter lambda of the Poisson distribution is equal to the sample mean of the observed data.
- Consider x a random variable with normal distribution. The maximum likelihood estimator of the mean value is equal to the sample mean of the observed data.
- A confidence interval is reported as plus or minus two standard deviations. Considering that the random variable has a normal distribution, this interval corresponds to confidence level of approximately 95%
- Decrease the estimated standard deviation to narrow down a 95% confidence interval in the case of a normal distribution
- A random variable with a chi square distribution is always positive
- A random variable has a Chi square distribution with 3 degrees of freedom. Its mean value is 3
- Consider two independent random variables z1 and z2 with standard normal distribution. The random variable equal to z1 squared plus z2 squared has Chi square distribution with 2 degrees of freedom
-

Mean: $E(Y) = e^{\mu + \frac{\sigma^2}{2}}$
SD: $variance = V(Y) = \frac{1}{\lambda^2} = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = sd^2$
Median: $m = e^{\mu}$

7.2 R code

```
1-pexp(1,lambda)      # Returns the cumulative probability distribution
                        #for value x
qexp(p, r)            #Returns the inverse cumulative probability
                      #distribution for probability p
plnorm(mean,6,6)      #used to compute the log normal value of the
                      #cumulative probability density function
```

7.3 Example Problem Set

Lognormal distribution:

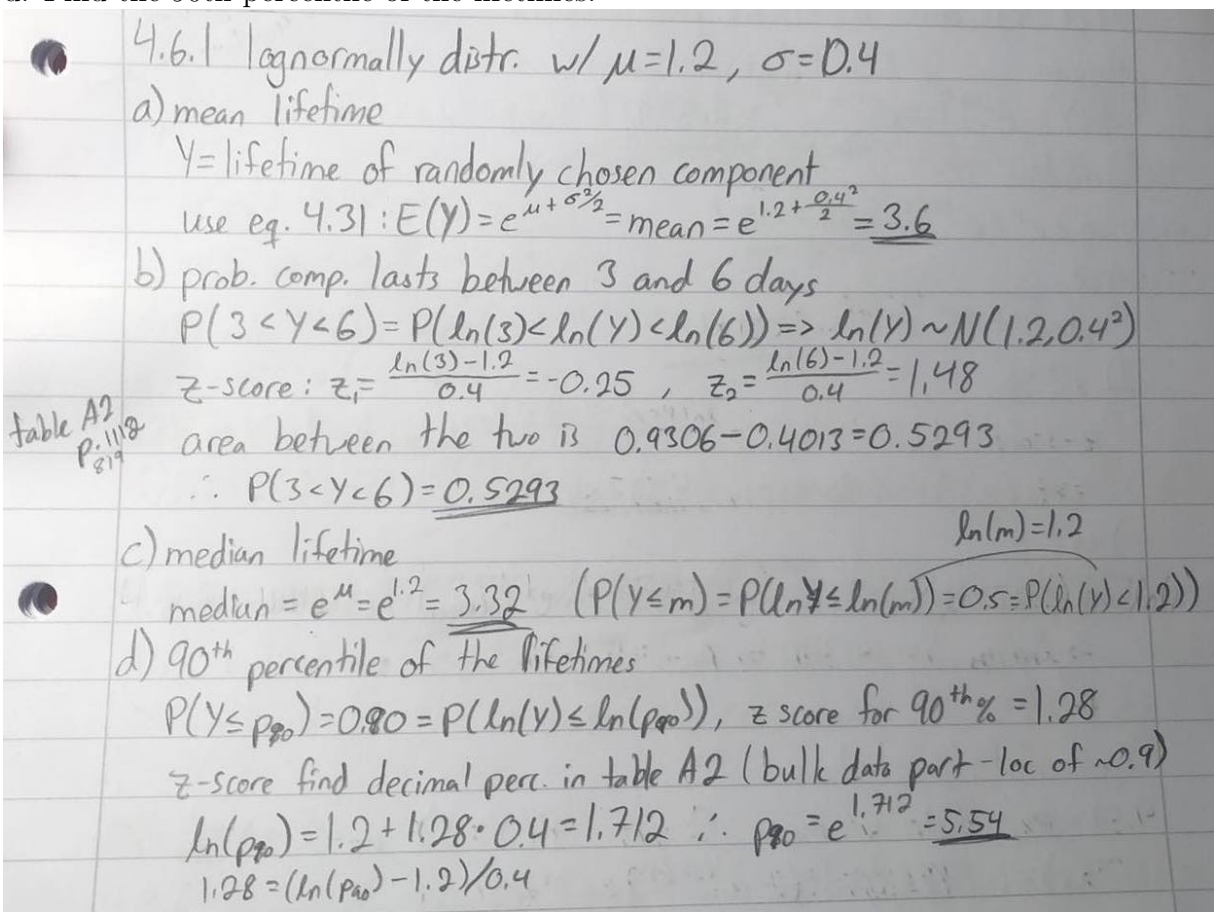
4.6.1 The lifetime (in days) of a certain electronic component that operates in a high-temperature environment is lognormally distributed with $\mu = 1.2$ and $\sigma = 0.4$.

a. Find the mean lifetime.

b. Find the probability that a component lasts between three and six days.

c. Find the median lifetime.

d. Find the 90th percentile of the lifetimes.



4.6.3 The body mass index (BMI) of a person is defined to be the person's body mass divided by the square of the person's height. The article "Influences of Parameter Uncertainties within the ICRP 66 Respiratory Tract Model: Particle Deposition" (W. Bolch, E. Farfan, et al., Health Physics, 2001:378–394) states that body mass index (in kg/m^2) in men aged 25–34 is lognormally distributed with parameters $\mu = 3.215$ and $\sigma = 0.157$.

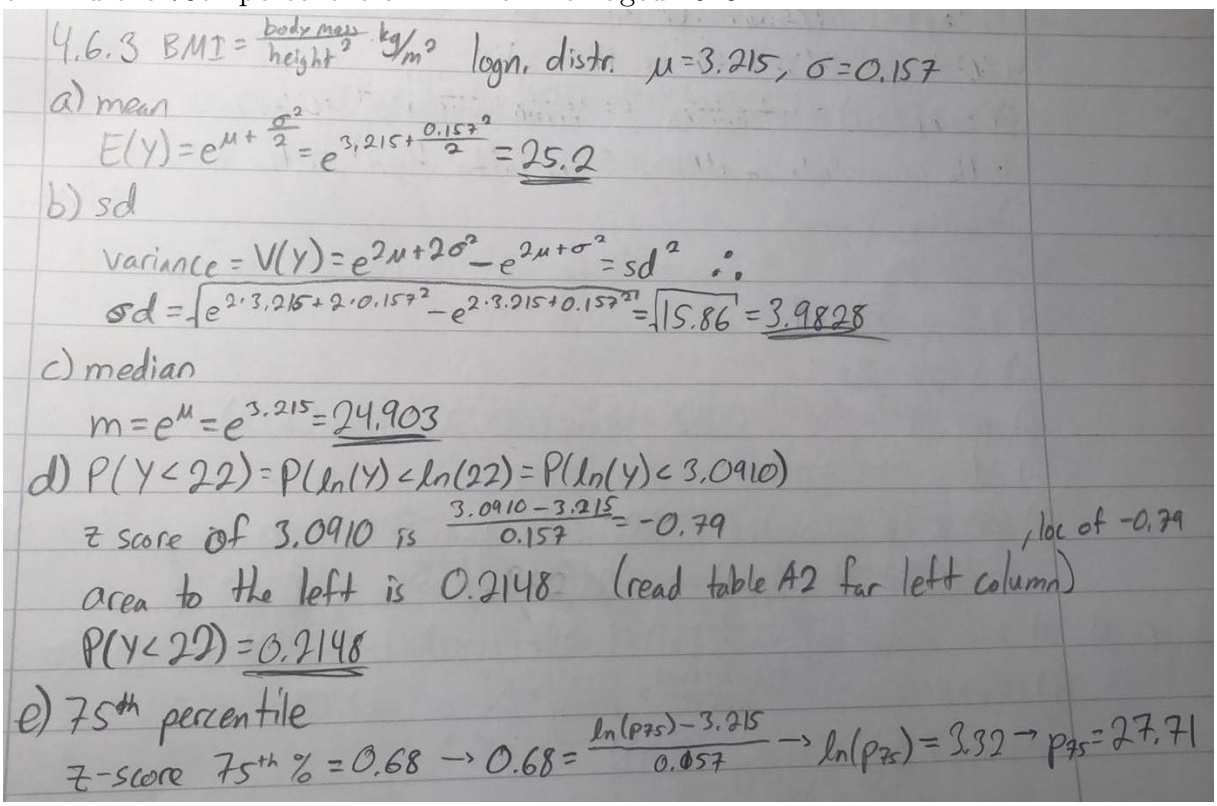
a. Find the mean BMI for men aged 25–34.

b. Find the standard deviation of BMI for men aged 25–34.

c. Find the median BMI for men aged 25–34.

d. What proportion of men aged 25–34 have a BMI less than 22?

e. Find the 75th percentile of BMI for men aged 25–34.



4.6.4 The article "Stochastic Estimates of Exposure and Cancer Risk from Carbon Tetrachloride Released to the Air from the Rocky Flats Plant" (A. Rood, P. McGavran, et al., Risk Analysis, 2001:675-695) models the increase in the risk of cancer due to exposure to carbon tetrachloride as lognormal with μ = -15.65 and σ = 0.79.

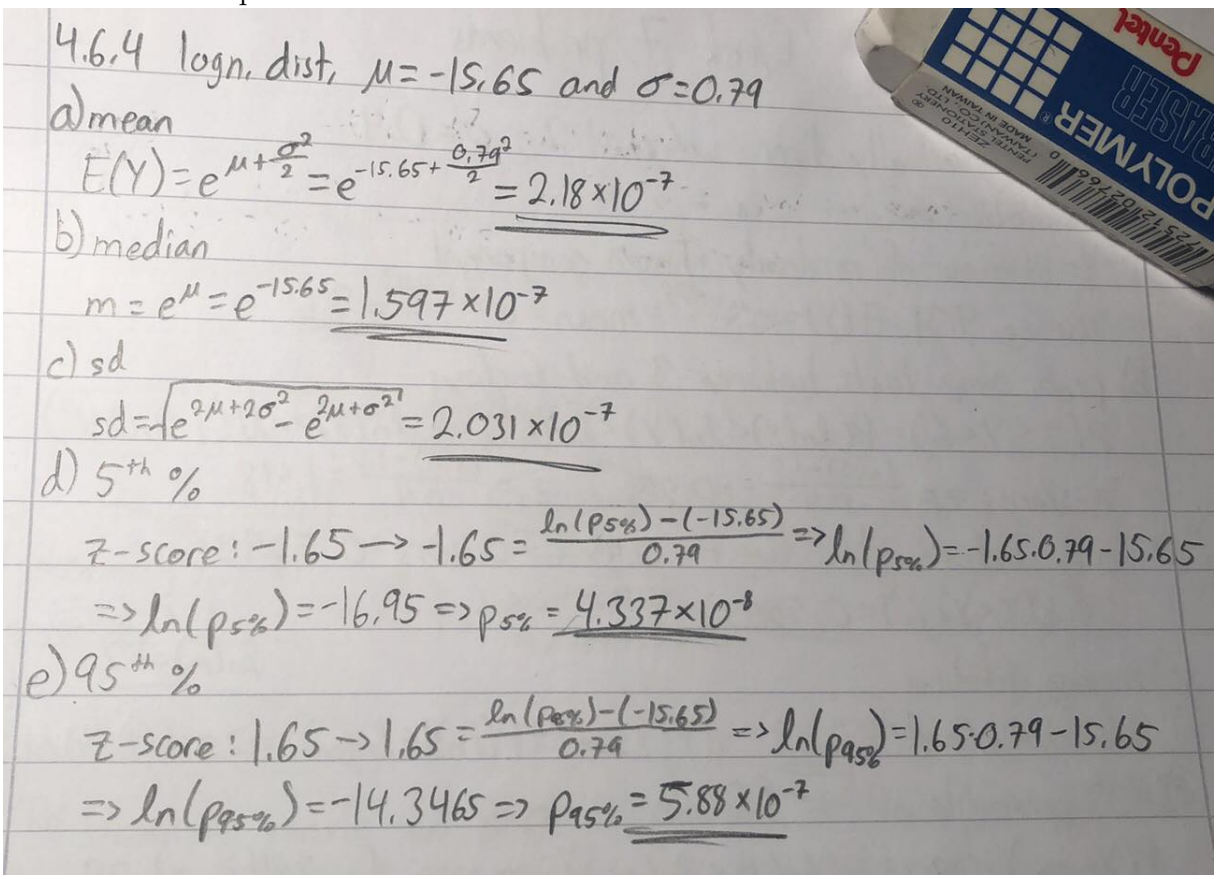
a. Find the mean risk.

b. Find the median risk.

c. Find the standard deviation of the risk.

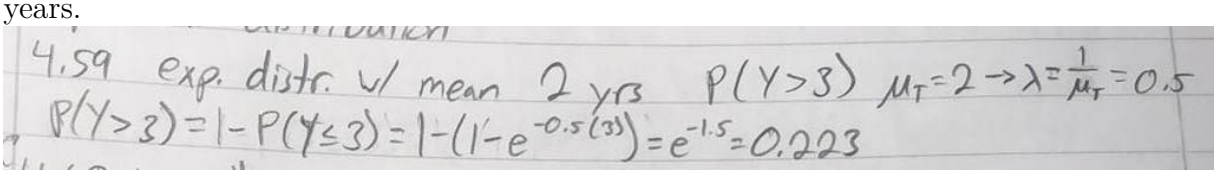
d. Find the 5th percentile.

e. Find the 95th percentile.

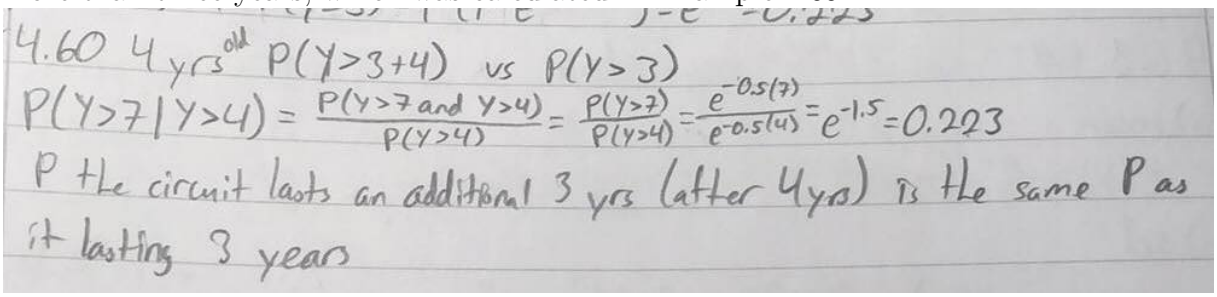


Exponential distribution:

Example 4.59 The lifetime of a particular integrated circuit has an exponential distribution with mean 2 years. Find the probability that the circuit lasts longer than three years.



Example 4.60 Refer to Example 4.59. Assume the circuit is now four years old and is still functioning. Find the probability that it functions for more than three additional years. Compare this probability with the probability that a new circuit functions for more than three years, which was calculated in Example 4.59.



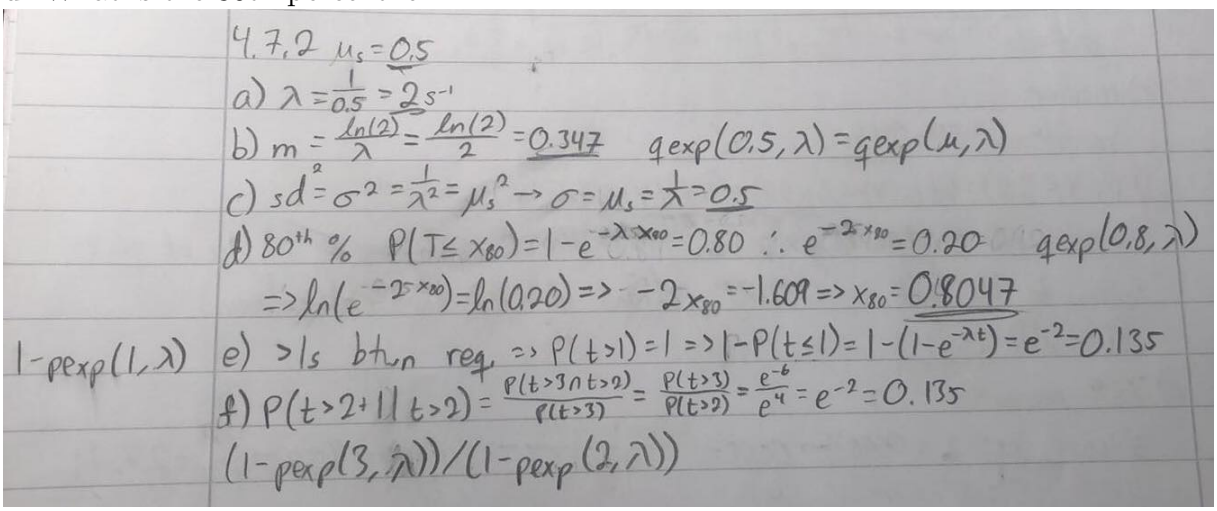
4.7.2 The time between requests to a web server is exponentially distributed with mean 0.5 seconds.

a. What is the value of the parameter λ?

b. What is the median time between requests?

c. What is the standard deviation?

d. What is the 80th percentile?



4.7.6 Refer to Exercise 2.

a. Find the probability that there will be exactly 5 requests in a 2-second time interval.

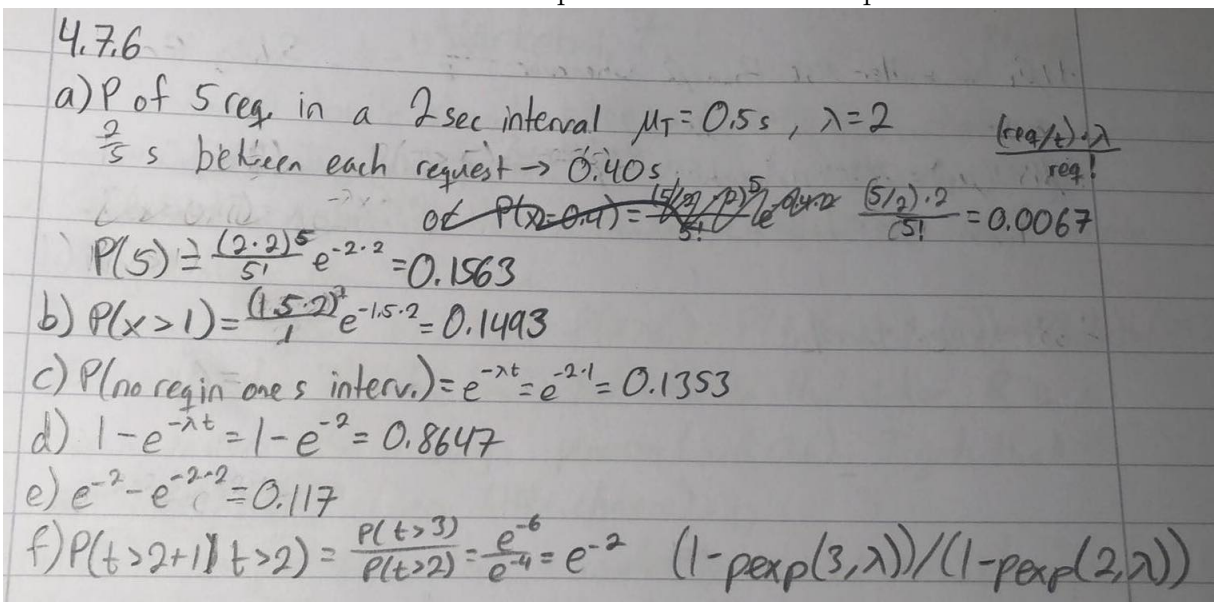
b. Find the probability that there will be more than 1 request in a 1.5-second time interval.

c. Find the probability that there will be no requests in a 1-second time interval.

d. Find the probability that the time between requests is greater than 1 second.

e. Find the probability that the time between requests is between 1 and 2 seconds.

f. If there have been no requests for the past two seconds, what is the probability that more than one additional second will elapse before the next request?

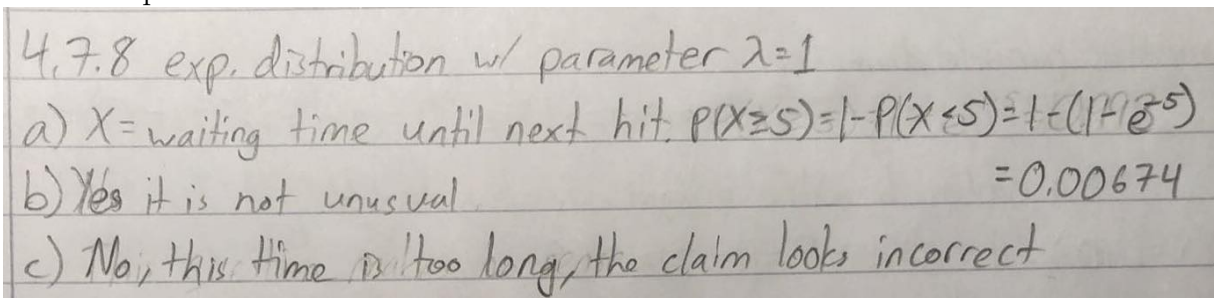


4.7.8 Someone claims that the waiting time, in minutes, between hits at a certain website has the exponential distribution with parameter $\lambda = 1$.

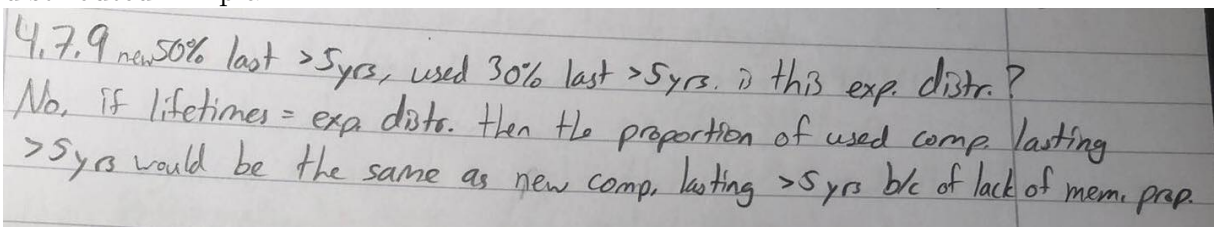
a. Let X be the waiting time until the next hit. If the claim is true, what is $P(X \geq 5)$?

b. Based on the answer to part (a), if the claim is true, is five minutes an unusually long time to wait?

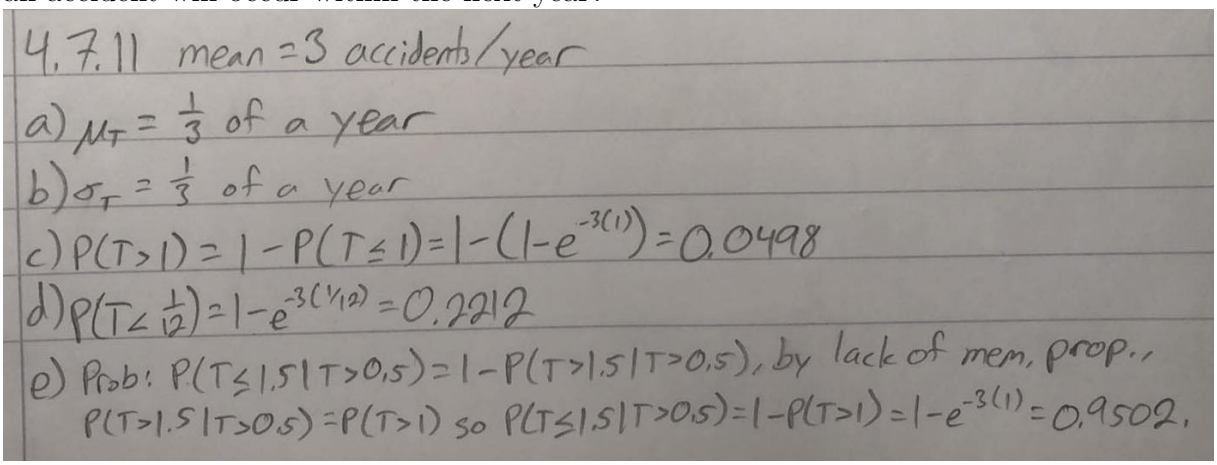
c. If you waited five minutes until the next hit occurred, would you still believe the claim? Explain.



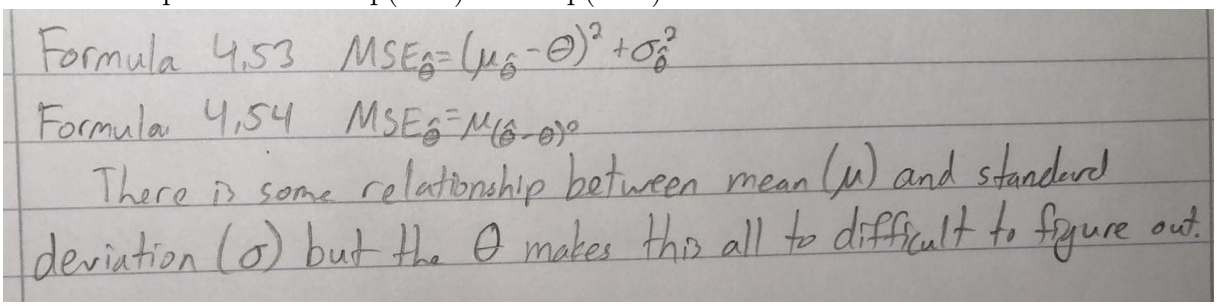
4.7.9 A certain type of component can be purchased new or used. Fifty percent of all new components last more than five years, but only 30% of used components last more than five years. Is it possible that the lifetimes of new components are exponentially distributed? Explain.



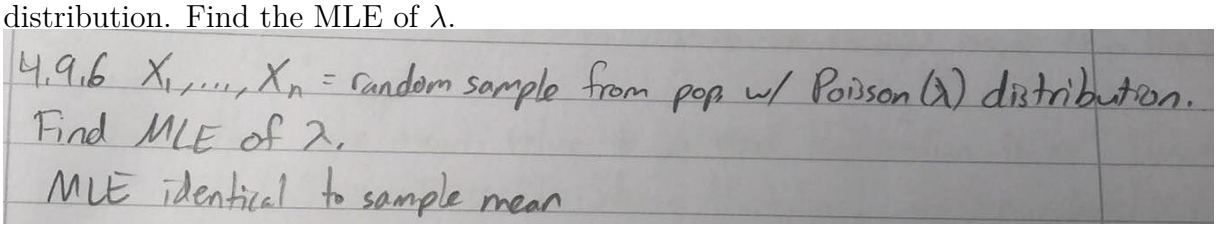
4.7.11 The number of traffic accidents at a certain intersection is thought to be well modeled by a Poisson process with a mean of 3 accidents per year.
a. Find the mean waiting time between accidents.
b. Find the standard deviation of the waiting times between accidents.
c. Find the probability that more than one year elapses between accidents.
d. Find the probability that less than one month elapses between accidents.
e. If no accidents have occurred within the last six months, what is the probability that an accident will occur within the next year?



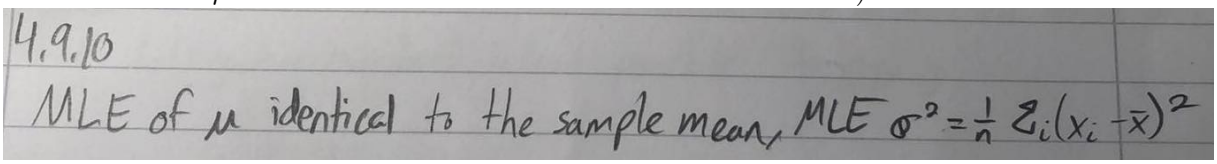
Estimates
Show the equivalence of Eq.(4.53) and Eq.(4.54) from the textbook.



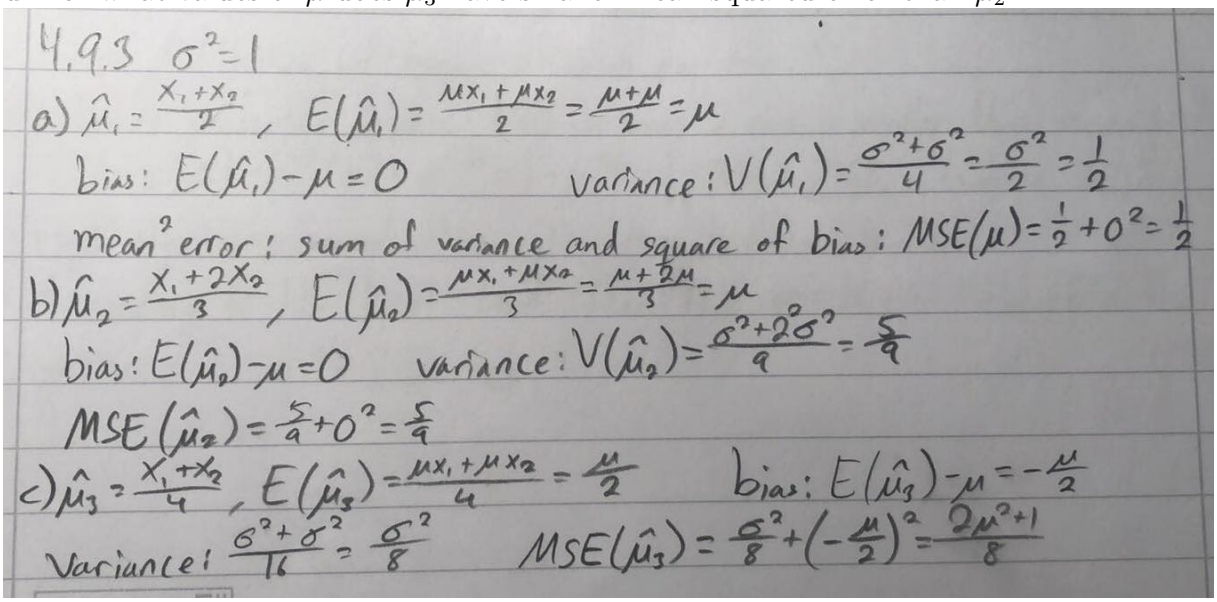
4.9.6 Let $X_1, \ldots, X_n$ be a random sample from a population with the Poisson(λ) distribution. Find the MLE of λ.

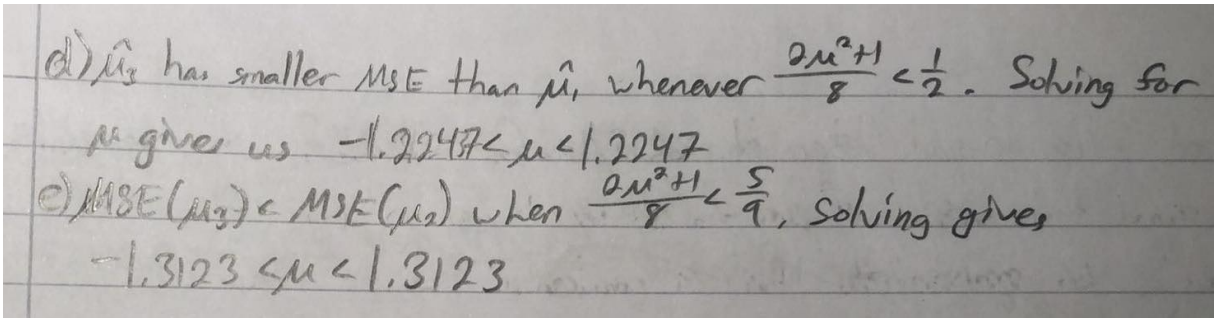


4.9.10 Let $X_1, \ldots, X_n$ be a random sample from a N(μ, σ^2) population. Find the MLEs of μ and of σ. (Hint: The likelihood function is a function of two parameters, μ and σ. Compute partial derivatives with respect to μ and a and set them equal to 0 to find the values $\hat{\mu}$ and $\hat{\sigma}$ that maximize the likelihood function.)

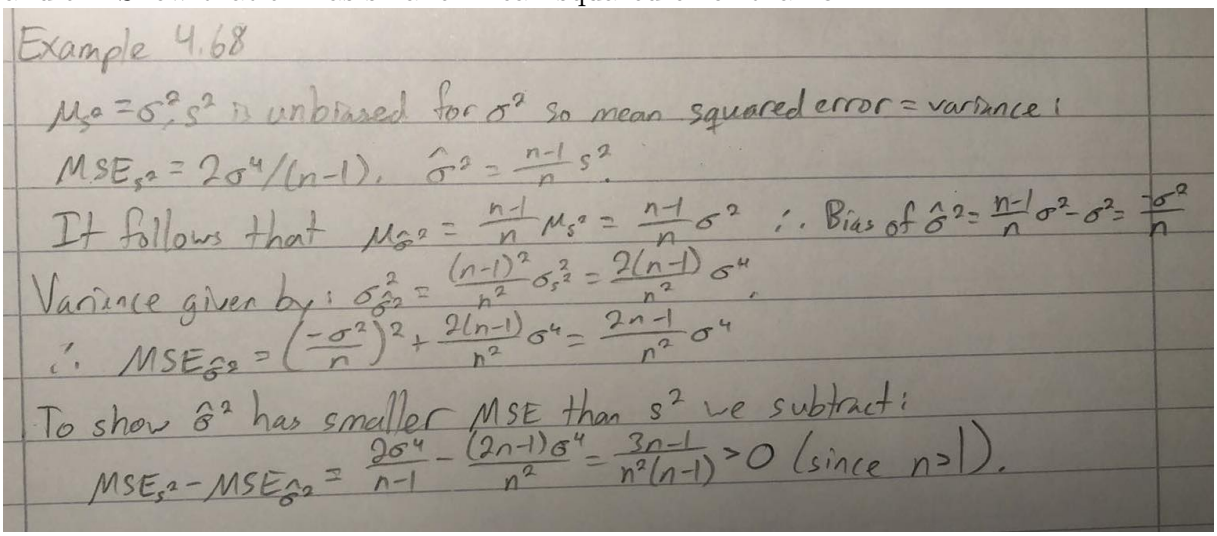


4.9.3 Let X_1 and X_2 be independent, each with unknown mean μ and known variance $\sigma^2 = 1$.
a. Let $\widehat{\mu_1} = \frac{X_1+X_2}{2}$. Find the bias, variance, and mean squared error of $\widehat{\mu_1}$.
b. Let $\widehat{\mu_2} = \frac{X_1+2X_2}{3}$. Find the bias, variance, and mean squared error of $\widehat{\mu_2}$.
c. Let $\widehat{\mu_3} = \frac{X_1+X_2}{4}$. Find the bias, variance, and mean squared error of $\widehat{\mu_3}$.
d. For what values of μ does $\widehat{\mu_3}$ have smaller mean squared error than $\widehat{\mu_1}$?
d. For what values of μ does $\widehat{\mu_3}$ have smaller mean squared error than $\widehat{\mu_2}$?

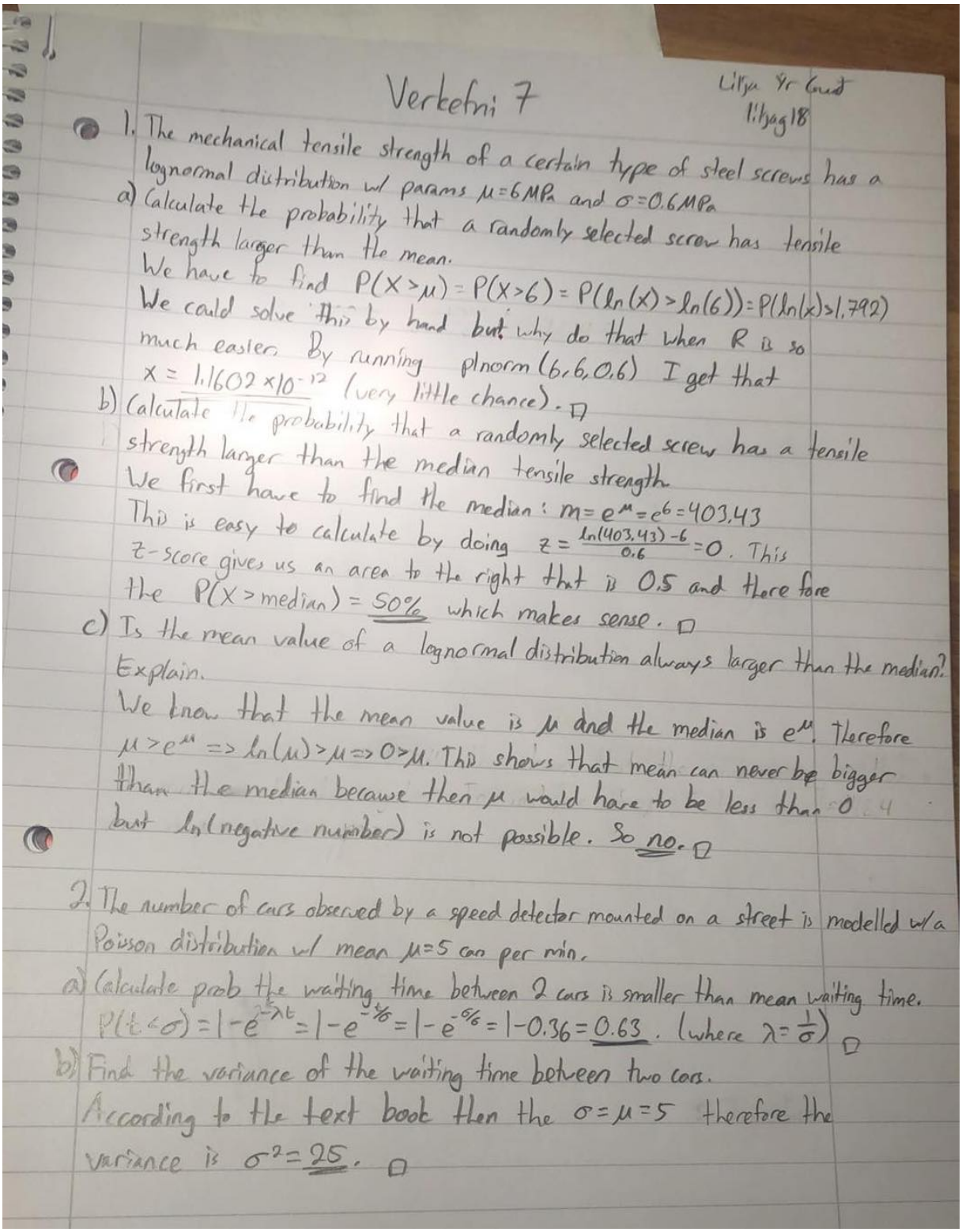




Example 4.68 Let X1, ..., Xn be a random sample from a N((μ, σ^2) population. The sample variance is $s^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$. It can be shown that S^2 has mean $\mu_{s^2} = \sigma^2$ and variance $\sigma^2_{s^2} = 2\sigma^2/(n-1)$. Consider the estimator $\widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)$ in which the sum of the squared deviations is divided by n rather than n - 1. Compute the bias, variance, and mean squared error of both s^2 and $\widehat{\sigma}^2$. Show that $\widehat{\sigma}^2$ has smaller mean squared error than s^2.



7.4 Problem Set



Verkefni 7

Lilja Ýr Gunnarsdóttir
Líftag18

1. The mechanical tensile strength of a certain type of steel screws has a lognormal distribution w/ params $\mu = 6$ MPa and $\sigma = 0.6$ MPa

a) Calculate the probability that a randomly selected screw has tensile strength larger than the mean.

We have to find $P(X > \mu) = P(X > 6) = P(\ln(X) > \ln(6)) = P(\ln(x) > 1.792)$
We could solve this by hand but why do that when R is so much easier. By running plnorm(6,6,0.6) I get that $x = 1.1602 \times 10^{-12}$ (very little chance). □

b) Calculate the probability that a randomly selected screw has a tensile strength larger than the median tensile strength.

We first have to find the median: $m = e^{\mu} = e^{6} = 403.43$
This is easy to calculate by doing $z = \frac{\ln(403.43) - 6}{0.6} = 0$. This z-score gives us an area to the right that is 0.5 and therefore the $P(X > \text{median}) = \underline{50\%}$ which makes sense. □

c) Is the mean value of a lognormal distribution always larger than the median? Explain.

We know that the mean value is μ and the median is e^{μ}. Therefore $\mu > e^{\mu} \Rightarrow \ln(\mu) > \mu \Rightarrow 0 > \mu$. This shows that mean can never be bigger than the median because then μ would have to be less than 0 but ln(negative number) is not possible. So no. □

2. The number of cars observed by a speed detector mounted on a street is modelled w/a Poisson distribution w/ mean $\mu = 5$ cars per min.

a) Calculate prob the waiting time between 2 cars is smaller than mean waiting time.

$P(t < \sigma) = 1 - e^{-\lambda t} = 1 - e^{-5/5} = 1 - e^{-1} = 1 - 0.36 = \underline{0.63}$. (where $\lambda = \frac{1}{\sigma}$) □

b) Find the variance of the waiting time between two cars.

According to the text book then the $\sigma = \mu = 5$ therefore the variance is $\sigma^2 = \underline{25}$. □

7.5 Problem Set Solutions

1. The mechanical tensile strength of a certain type of steel screws has a lognormal distribution with parameters μ=6 MPa (mega Pascal) and σ=0.6 MPa.
a) Calculate the probability that a randomly selected screw has a tensile strength larger than the mean tensile strength.mean=$e^{\mu+\sigma^2/2}$=483

$1-\mathbf{plnorm}(\mathbf{mean}, 6, 6) = 0.382$

b) Calculate the probability that a randomly selected screw has a tensile strength larger than the median tensile strength.
By definition the median is the value of an RV x such thatP(x<median)=P(x > median)= 0.5. Alternatively, obtain the median403.4and check.
c) Is the mean value of a lognormal distribution always larger than the median? Explain.
If x has a lognormal distribution, ln(x) has a normal distribution whose median is μ, and so the median of x is e^{μ}.We see that $e^{\mu+\sigma^2/2} > e^{\mu}$ because $e^{\sigma^2/2} > 1$. So yes, the mean of the lognormal distribution is always larger than the median. This inequality is also expected due to the asymmetry of the distribution, which has a long tail to large values of x, or it is skewed to the right. The 2 points are given either for the proof with exponential functions, or for some explanations of the tail effect indicating that the student understood it.

2. The number of cars observed by a speed detector mounted on a street is modelled with a Poisson distribution with a mean of 5 cars per minute.
a) Calculate the probability that the waiting time between two cars is smaller than the mean waiting time.
The waiting time has an exponential distribution with parameter λ=5whose mean value is T=1/λ.Using the cumulative distribution F(t)=1$-e^{-\lambda t}$ we obtain P(t<T)=F(T)=1$-e^{-1}$=0.632 or

$\mathbf{pexp}(1/5, 5) = 0.632$

b) Find the variance of the waiting time between two cars.
$x \sim Exp(\lambda) \Rightarrow Var(x) = \frac{1}{\lambda^2} = 0.04$

8 Week 8

8.1 Concepts

- If the p-value is less than 0.05 the null hypothesis is rejected
- In a test of a population mean the one sided p-value is one half of the two sided p-value
- The z-score is calculated assuming H0 true
- The p-value is the probability that the disagreement between the null and alternative hypotheses is greater than observed
- Let's think about a case when the alternative hypothesis is true. If we could increase the sample size the p-value should decrease
- The mean value of a random variable with t distribution with n degrees of freedom is 0
- In a t test with 5 degrees of freedom you obtain t=2. Calculate the corresponding 2-sided p-value. Use point (not comma) as decimal symbol. (For example 0.2 not 0,2) -> 0.1019
- If in the previous t test the t number increases, the p-value decreases
- You perform a 1-sided t test with n degrees of freedom and obtain t=0. The p-value is 0.5
- A group of eight individuals with high cholesterol levels were given a new drug that was designed to lower cholesterol levels. Cholesterol levels, in mg/dL, were measured before and after treatment for each individual. The best method to test the effect of the new drug using this data is a t test with paired data

8.2 Example Problem Set

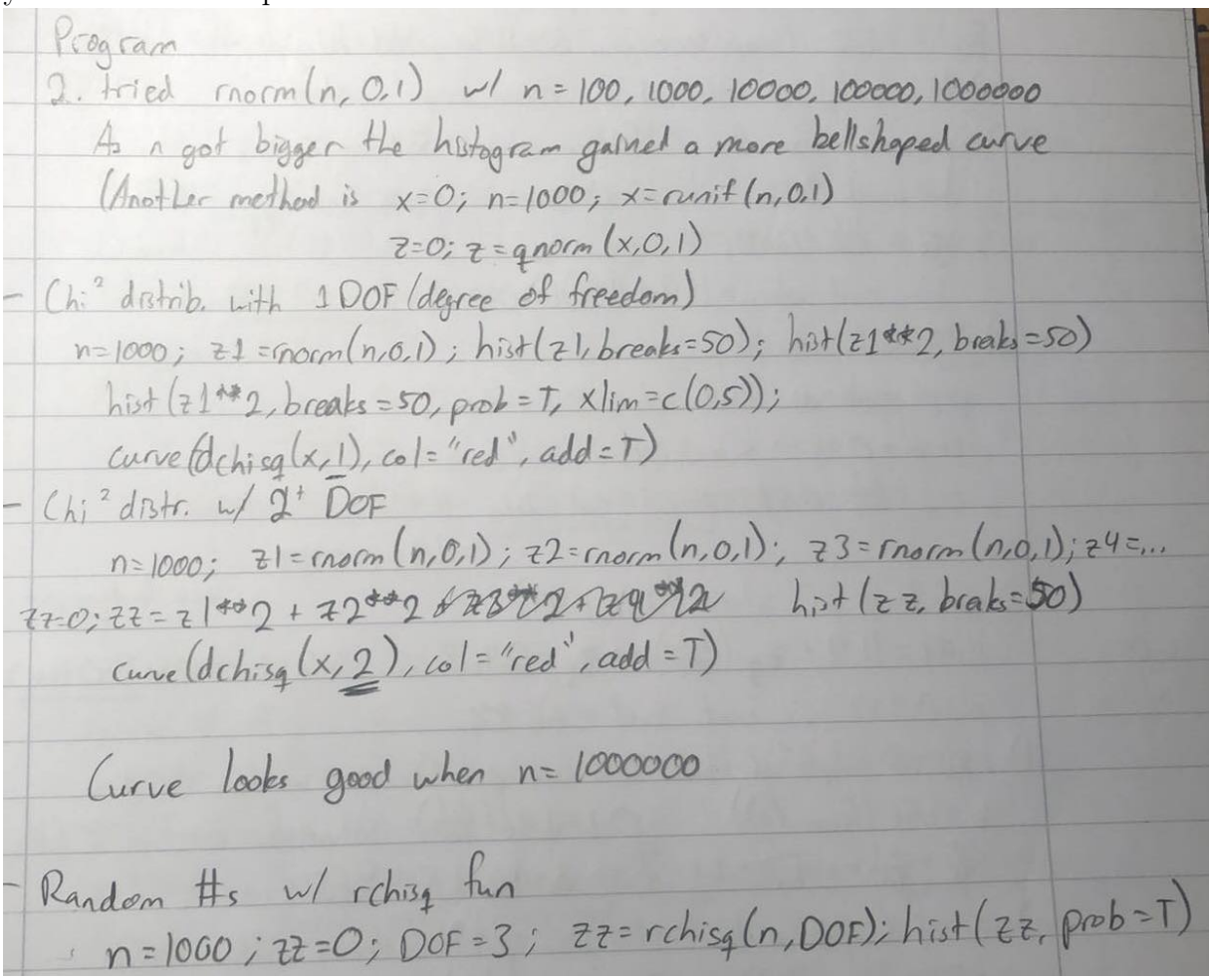
Miniproject phase 3: Computer generated random numbers with normal and chi square distributions

1. See first the recent video on this subject.

2. Use the function rnorm(n,0,1) in R or NORM.INV(RAND(),0,1) in Excel to generate a vector of n=100,1000, ... random numbers with standard normal distribution N(0,1). Observe their histogram and how it evolves with increasing n.

3. Generate now several vectors of n random numbers with N(0,1), z1, z2, z3, ... and make histograms of $x_1 = z_1^2$, $x_2 = z_1^2 + z_2^2$, $x_3 = z_1^2 + z_2^2 + z_3^2$, ... What distribution you obtain? Compare with random numbers generated with the functions rchisq(n,k) or CHISQ.INV(RAND(),k).

4. Consolidate your understanding of the Chi square distribution by practicing such exercises.Make good graphical representations, optimise the number of bins, add on the graphs probability density functions for comparison. Calculate sample mean values using random numbers and compare with the true values. Observe median values.Imagine yourselves other experimental tests.



Confidence intervals for population mean:

5.1.13 The sugar content in a one-cup serving of a certain breakfast cereal was measured for a sample of 140 servings. The average was 11.9 g and the standard deviation was 1.1 g.

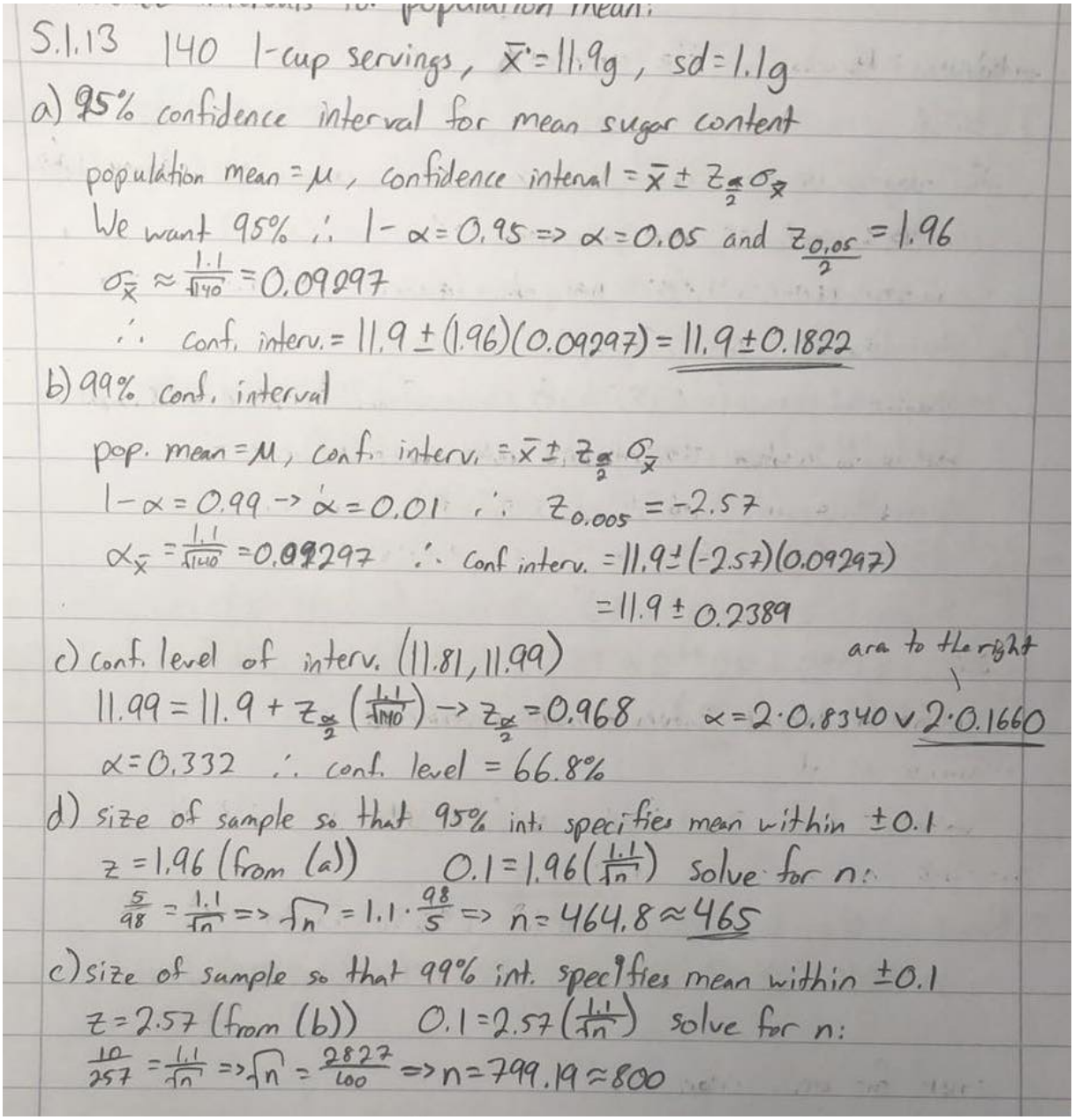
a. Find a 95% confidence interval for the mean sugar content.

b. Find a 99% confidence interval for the mean sugar content.

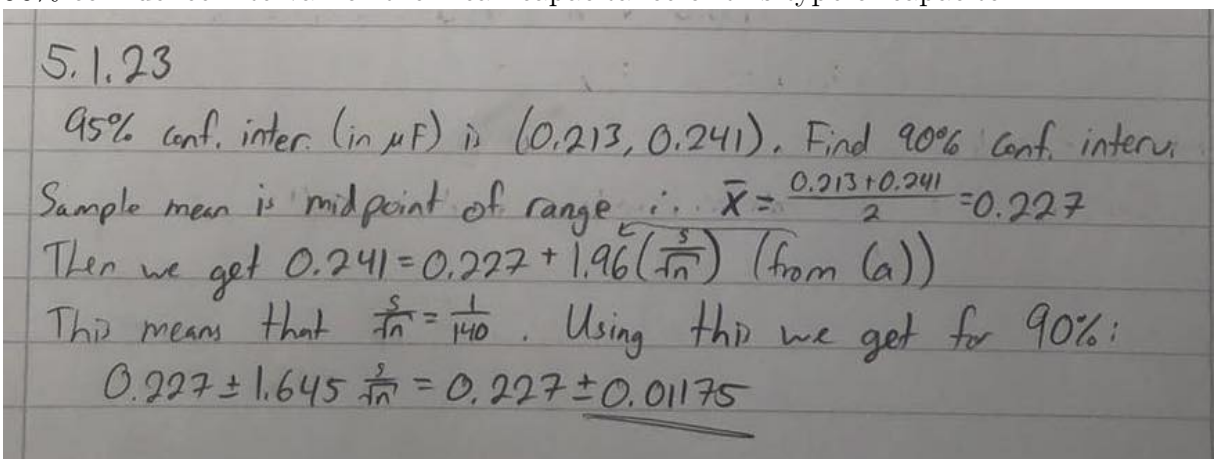
c. What is the confidence level of the interval (11.81, 11.99)?

d. How large a sample is needed so that a 95% confidence interval specifies the mean to within ±0.1?

e. How large a sample is needed so that a 99% confidence interval specifies the mean to within ±0.1?

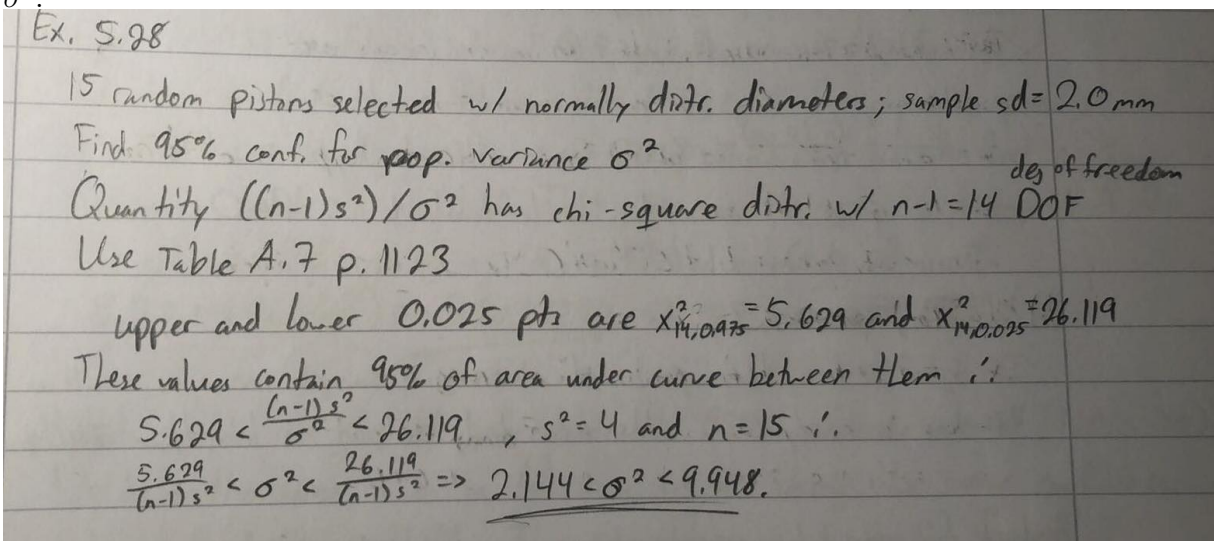


5.1.23 Based on a large sample of capacitors of a certain type, a 95% confidence interval for the mean capacitance, in μF, was computed to be (0.213, 0.241). Find a 90% confidence interval for the mean capacitance of this type of capacitor.



Confidence intervals for the variance:

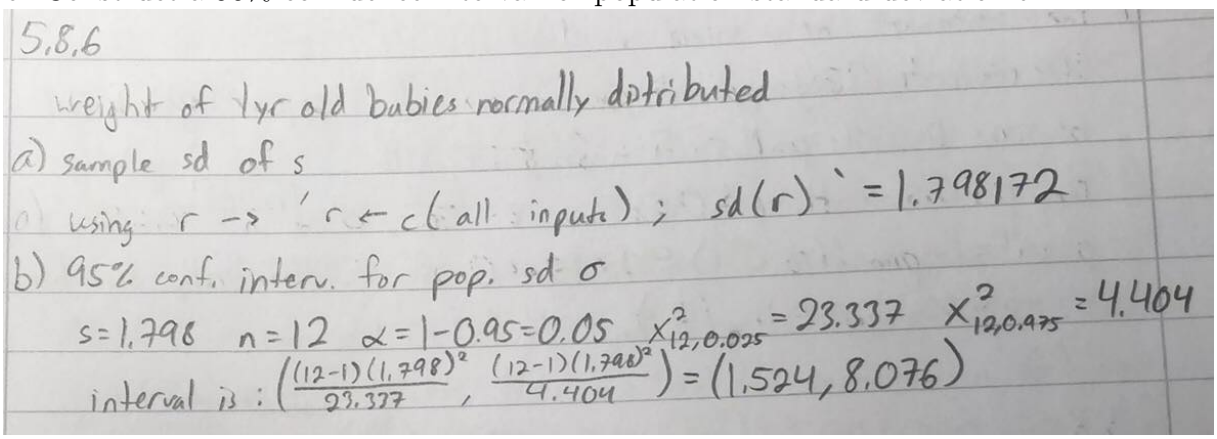
Example 5.28 A simple random sample of 15 pistons is selected from a large population whose diameters are known to be normally distributed. The sample standard deviation of the piston diameters is s = 2.0 mm. Find a 95% confidence for the population variance σ^2.



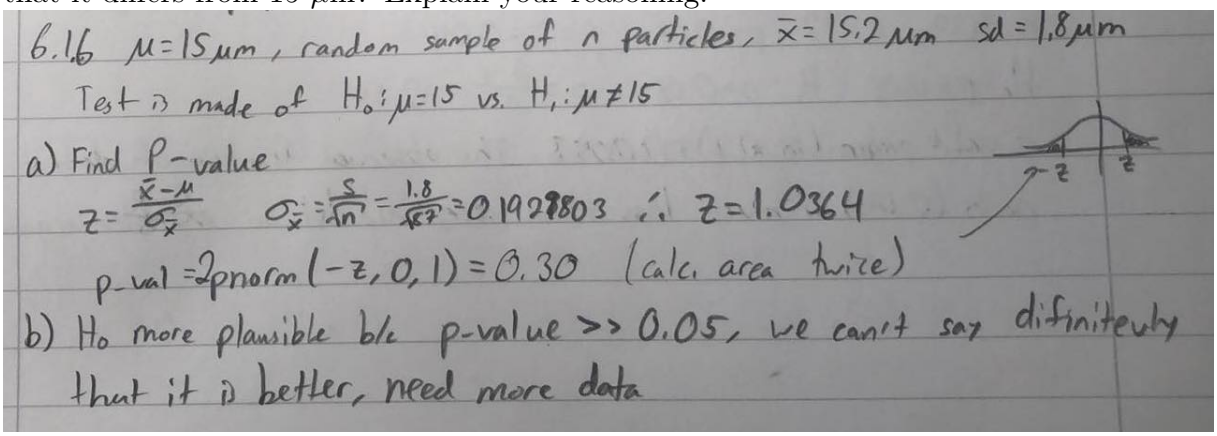
5.8.6 Following are weights, in pounds, of 12 two-month-old baby girls. Assume that the population is normally distributed.

12.23	12.32	11.87	12.34	11.48	12.66
8.51	14.13	12.95	10.30	9.34	8.63

a. Find the sample standard deviation s.
b. Construct a 95% confidence interval for population standard deviation σ.



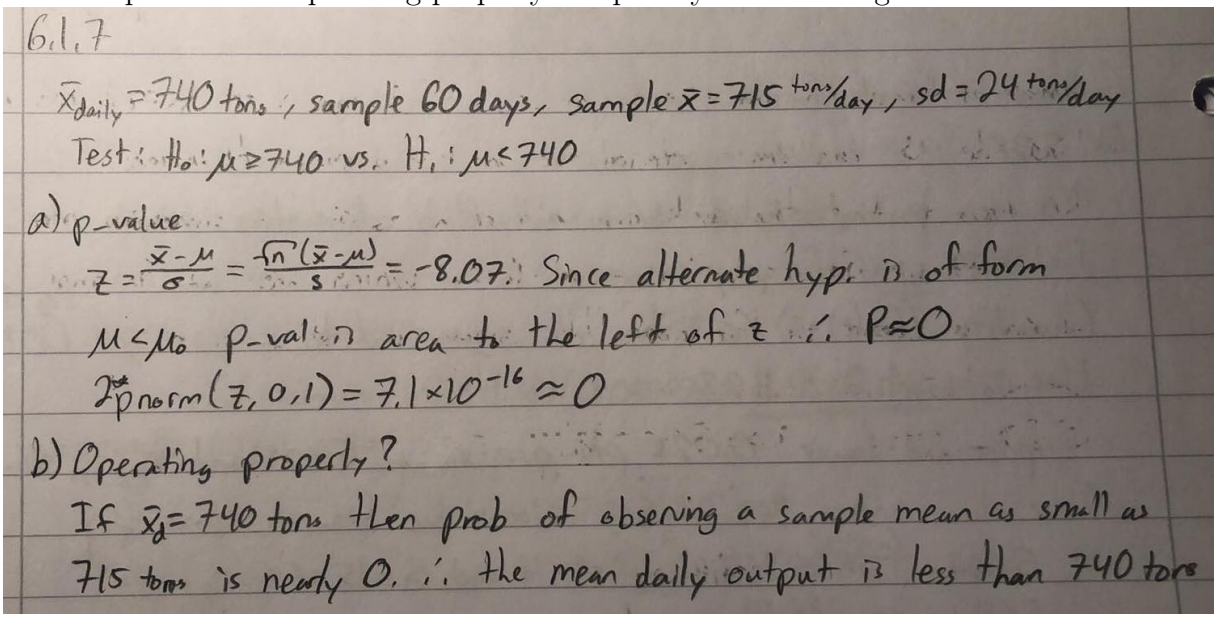
Z-test for population mean:
6.1.6 A certain type of stainless steel powder is supposed to have a mean particle diameter of $\mu = 15$ μm. A random sample of 87 particles had a mean diameter of 15.2 μm, with a standard deviation of 1.8 μm. A test is made of H_0: $\mu = 15$ versus H_1: $\mu \neq 15$
a. Find the P-value.
b. Do you believe it is plausible that the mean diameter is 15 μm, or are you convinced that it differs from 15 μm? Explain your reasoning.



6.1.7 When it is operating properly, a chemical plant has a mean daily production of at least 740 tons. The output is measured on a simple random sample of 60 days. The sample had a mean of 715 tons/day and a standard deviation of 24 tons/day. Let μ represent the mean daily output of the plant. An engineer tests H_0:$\mu \geq 740$ versus H_1:$\mu < 740$.

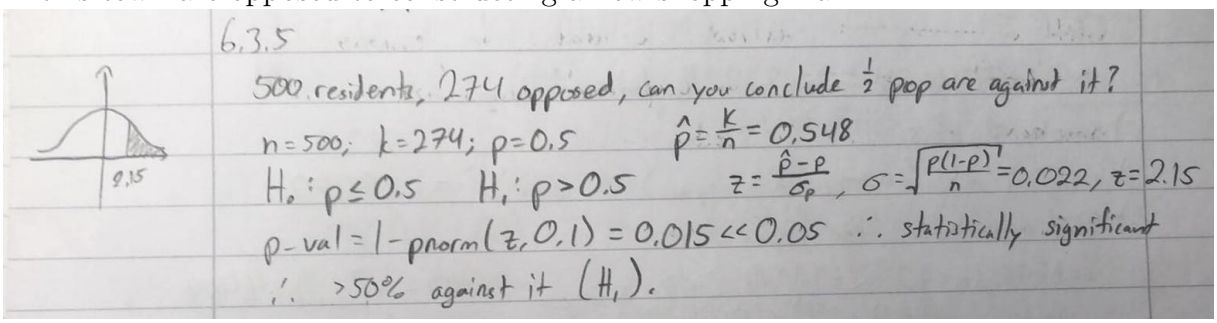
a. Find the P-value.

b. Do you believe it is plausible that the plant is operating properly or are you convinced that the plant is not operating properly? Explain your reasoning.

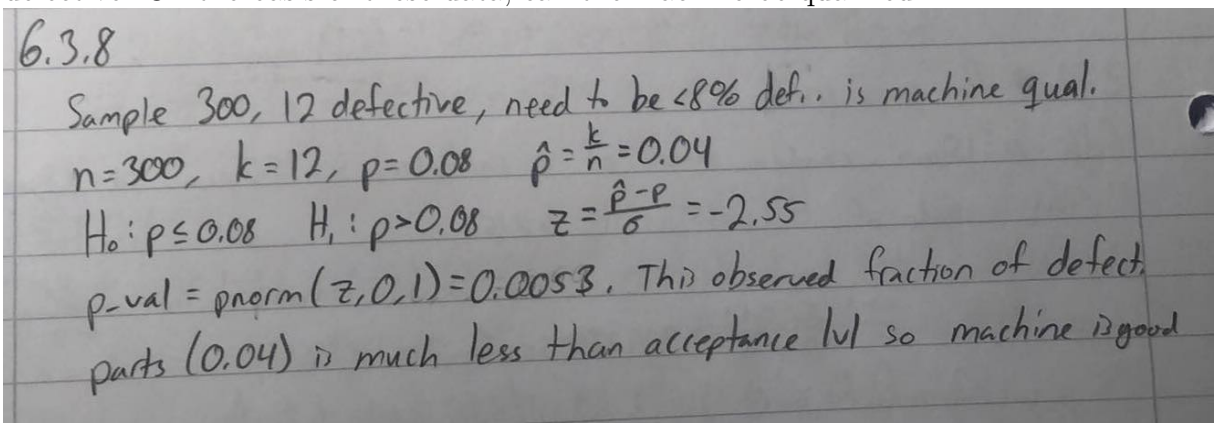


Z-test for population proportions:

6.3.5 In a survey of 500 residents in a certain town, 274 said they were opposed to constructing a new shopping mall. Can you conclude that more than half of the residents in this town are opposed to constructing a new shopping mall?



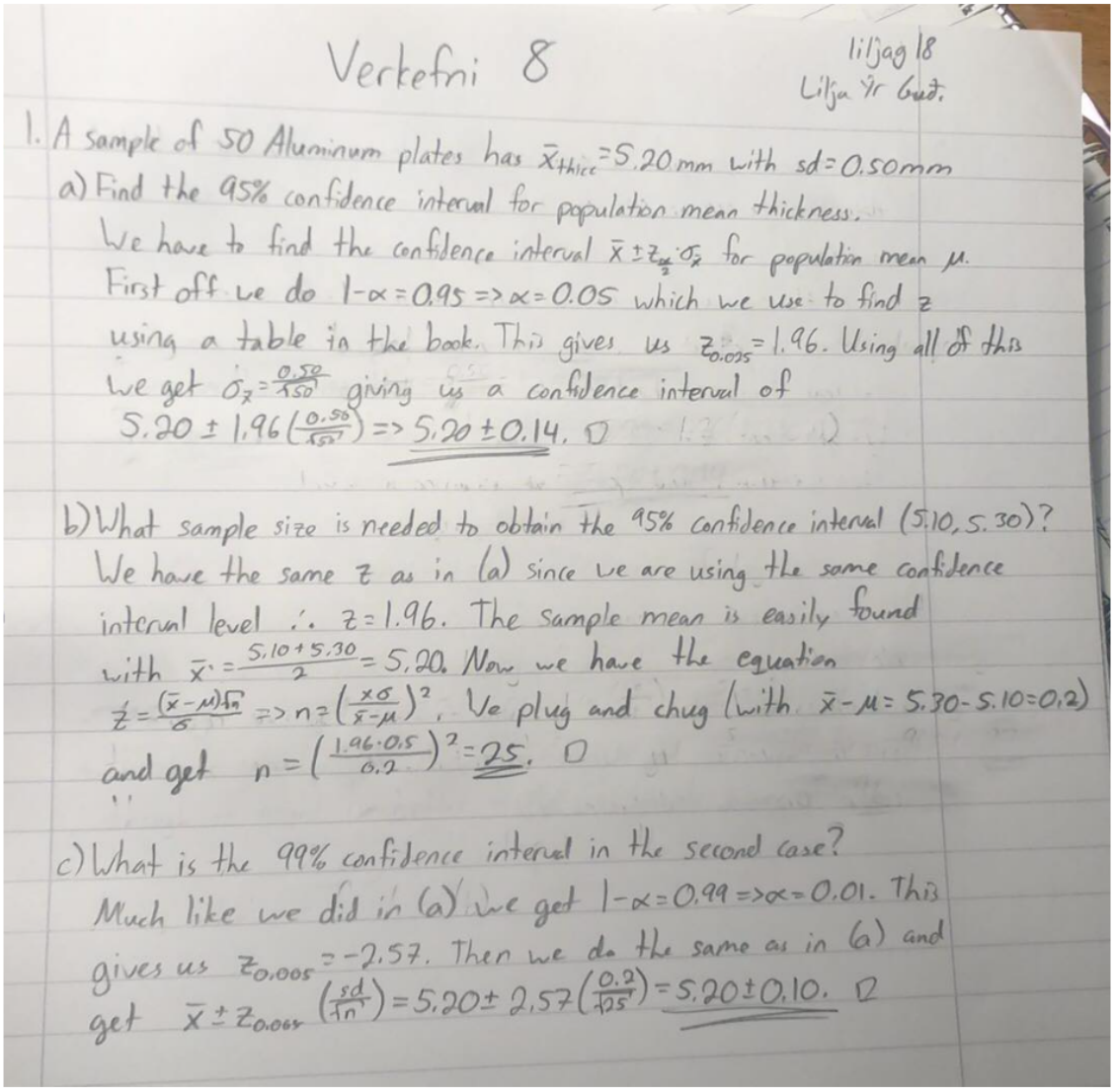
6.3.8 A grinding machine will be qualified for a particular task if it can be shown to produce less than 8% defective parts. In a random sample of 300 parts, 12 were defective. On the basis of these data, can the machine be qualified?



8.3 Problem Set

Verkefni 8

liljag18



Verkefni 8

liljag18
Lilja Ýr Guð.

1. A sample of 50 Aluminum plates has $\bar{x}_{thicc} = 5.20$ mm with $sd = 0.50$mm

a) Find the 95% confidence interval for population mean thickness.

We have to find the confidence interval $\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \sigma_{\bar{x}}$ for population mean μ. First off we do $1-\alpha = 0.95 \Rightarrow \alpha = 0.05$ which we use to find z using a table in the book. This gives us $z_{0.025} = 1.96$. Using all of this we get $\sigma_{\bar{x}} = \frac{0.50}{\sqrt{50}}$ giving us a confidence interval of

$5.20 \pm 1.96\left(\frac{0.50}{\sqrt{50}}\right) \Rightarrow 5.20 \pm 0.14$. □

b) What sample size is needed to obtain the 95% confidence interval (5.10, 5.30)?

We have the same z as in (a) since we are using the same confidence interval level $\therefore z = 1.96$. The sample mean is easily found with $\bar{x} = \frac{5.10+5.30}{2} = 5.20$. Now we have the equation $z = \frac{(\bar{x}-\mu)\sqrt{n}}{\sigma} \Rightarrow n = \left(\frac{z\sigma}{\bar{x}-\mu}\right)^2$. We plug and chug (with $\bar{x} - \mu = 5.30 - 5.10 = 0.2$) and get $n = \left(\frac{1.96 \cdot 0.5}{0.2}\right)^2 = 25$. □

c) What is the 99% confidence interval in the second case?

Much like we did in (a) we get $1-\alpha = 0.99 \Rightarrow \alpha = 0.01$. This gives us $z_{0.005} = -2.57$. Then we do the same as in (a) and get $\bar{x} \pm z_{0.005}\left(\frac{sd}{\sqrt{n}}\right) = 5.20 \pm 2.57\left(\frac{0.2}{\sqrt{25}}\right) = 5.20 \pm 0.10$. □

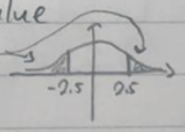
Verkefni 8

liljag18

2. An instrument designed to measure thin layers of paint is tested. To determine the accuracy of the instrument, it was used to take 100 measurements of a layer with a known thickness of 100 micron. The mean of the measurements was 115 micron w/ a sd = 60 micron.

a) Perform a z test to compare the mean of the measurements w/ the real thickness and calculate the 2 sided p-value

We get that $z = \frac{(\bar{x}-\mu)\sqrt{n}}{sd} = \frac{(115-100)\sqrt{100}}{60} = \frac{5}{2} = 2.5$



Then we use r and get p-value = 2pnorm(-z, 0, 1)

$= 0.0124$ □

Two tests are $H_0: \mu = 100$ and $H_1: \mu \neq 100$.

b) Do you believe the instrument is properly callibrated? Explain.

No because the probability of getting these results is 0.012 which is really low (and p-value $<< 0.05$ so we reject H_0 in favor of H_1 which essentially means the instrument is not properly callibrated) □

8.4 Problem Set Solution

1. A sample of 50 Aluminum plates has a mean thickness 5.20 mm with a standard deviation 0.50 mm.
a) Find the 95% confidence interval for the population mean thickness.
x=5.20 σ_x= 0.50/sqrt(50) = 0.0707 95%CI = [x - 1.96σ_x, x + 1.96σ_x]=[5.06, 5.34]
b) What sample size is necessary to obtain the 95% confidence interval [5.10,5.30]?
With the new sample size n' and the new std.σ'_x
x - 5.10 = 1.96$\sigma'_x = \frac{0.5}{\sqrt{n}} \Rightarrow$ n' = $(\frac{1.96*0.5}{0.1})^2 = 96$.
c) What is the 99% confidence interval in the second case?
$\sigma'_x = \frac{0.5}{\sqrt{96}} = 0.051 \Rightarrow$ 95% CI = [x - 2.58σ'_x, x + 2.58σ'_x]=[5.07,5.33].

2. An instrument designed to measure thin layers of paint is tested. To determine the accuracy of the instrument, it was used to take 100 measurements of a layer with a known thickness of 100 micron.The mean of the measurements was 115micron with a standard deviation of 60 micron.
a) Perform a z test to compare the mean of the measurements with the real thickness and calculate the 2-sided p-value.
z-score: z = $\frac{115-100}{60/\sqrt{100}}$ = 2.5p - value = 2*pnorm(-z,0,1) = 0.0124.
b) Do you believe the instrument is properly calibrated? Explain.
The mean measurement is significantly different than the expected value. The instruments needs recalibration.

9 Week 9

9.1 Concepts

- The p-value returned by a chi square test is always one sided.
- In a chi-square test with 6 DOF the calculated test statistic chi squared is equal to 16. The p_value is between 0.013 and 0.014
- The electroencephalography data discussed in the presentation of today was analysed using t test with paired data
- The data table for a chi square test of independence has 3 rows and 3 columns. It has 4 degrees of freedom
- In a chi-square test the expected values in each category are calculated using the null hypothesis
- In a study on gender balance the proportions of girls and boys are tested in 100 universities with a chi square test. A result can be considered significant if the p-value is lower than 0.0005
- The probability distribution used in a Fisher test is the hypergeometric distribution
- A false positive result corresponds to a p-value < 0.05
- The statistical power of a test increases by increasing the sample size
- What is the odds ratio attributed to the genetic variant associated with myocardial infarction discovered at Decode Genetics in 2007? Between 1.2 and 1.3

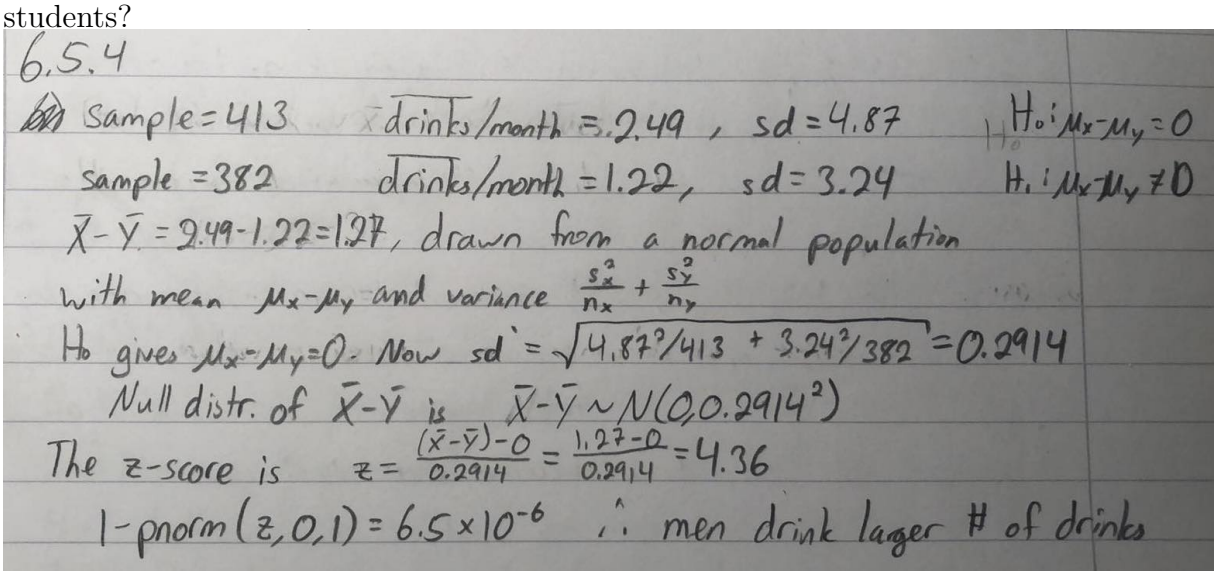
9.2 R code

```
1-pnorm(z,0,1)        #calculate p_value
setwd("/Users/liljayrgudmundsdottir/Documents/HR_Files/5onn/Tolfraedi/
datasets/Ch6")
data <- read.csv('ex6-7-2.csv')
disk_tablets = data$Disk
oval_tablets = data$Oval
t.test(disk_tablets, oval_tablets)        #probem 6.7.2
data_A = data$Brand.A
data_B = data$Brand.B
ttest(data_A, data_B)
t.test(data_A, data_B)      #problem 6.7.14
data_60 = data$X60.degrees
data_61 = data$X61.degrees
t.test(data_60, data_61)          #problem 6.7.15
t.test(FW, SW, var.equal=TRUE)      #problem 6.7.6
t.test(B, G, alt="two.sided", paired=TRUE)          #problem 6.8.1
x=c(467 ,191 ,42); y=c(445 ,171 ,34); z=c(254,129,93)
chisq.test(rbind(x , y, z ))      #6.2.10 chi square
```

9.3 Example Problem Set

Z-test for difference of population means:

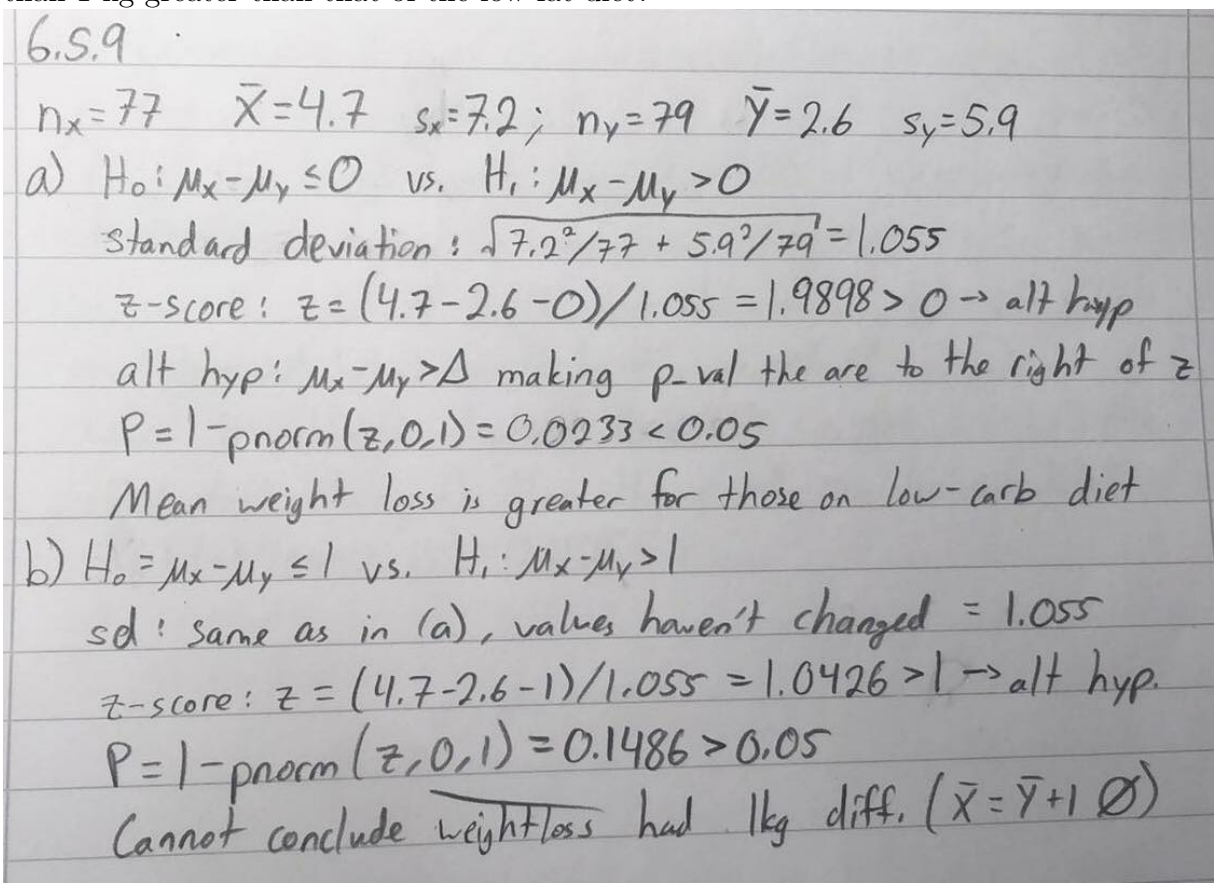
6.5.4 The article "Wired: Energy Drinks, Jock Identity, Masculine Norms, and Risk Taking" (K. Miller, Journal of American College Health, 2008:481–489) reports that in a sample of 413 male college students, the average number of energy drinks consumed per month was 2.49 with a standard deviation of 4.87, and in a sample of 382 female college students, the average was 1.22 with a standard deviation of 3.24. Can you conclude that the mean number of energy drinks is greater for male students than for female students?



6.5.9 Are low-fat diets or low-carb diets more effective for weight loss? This question was addressed in the article "Comparison of the Atkins, Zone, Ornish, and LEARN Diets for Change in Weight and Related Risk Factors Among Overweight Premenopausal Women: The A TO Z Weight Loss Study: A Randomized Trial" (C. Gardner, A. Kiazand, et al., Journal of the American Medical Association, 2007:969–977). A sample of 77 subjects went on a low-carbohydrate diet for six months. At the end of that time the sample mean weight loss was 4.7 kg with a sample standard deviation of 7.2 kg. A second sample of 79 subjects went on a low-fat diet. Their sample mean weight loss was 2.6 kg with a standard deviation of 5.9 kg.

a. Can you conclude that the mean weight loss is greater for those on the low carbohydrate diet?

b. Can you conclude that the mean weight loss on the low-carbohydrate diet is more than 1 kg greater than that of the low-fat diet?



t-test:

6.7.2 In a study of the relationship of the shape of a tablet to its dissolution time, 6 disk-shaped ibuprofen tablets and 8 oval-shaped ibuprofen tablets were dissolved in water. The dissolve times, in seconds, were as follows:
Disk: 269.0 249.3 255.2 252.7 247.0 261.6
Oval: 268.8 260.0 273.5 253.9 278.5 289.4 261.6 280.2
Can you conclude that the mean dissolve times differ between the two shapes?

```
> t.test(disk_tablets, oval_tablets)

    Welch Two Sample t-test

data:  disk_tablets and oval_tablets
t = -2.7757, df = 11.961, p-value = 0.01683
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -26.667185  -3.207815
sample estimates:
mean of x mean of y
 255.8000  270.7375
```

Yes we can conclude that the mean dissolve times differ because p_value is 0.01683 << 0.05. Unequal Variance.

6.7.14 The data file Datasets.zip/datasets/Ch5/ex5-6-11.csv shows measurements of the sodium content (in grams) in samples of two brands of chocolate bar. Can you conclude that the mean sodium content is higher for brand B than for brand A?

```
> t.test(data_A, data_B)

    Welch Two Sample t-test

data:  data_A and data_B
t = -3.1601, df = 19.552, p-value = 0.005024
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.466653 -1.319501
sample estimates:
mean of x mean of y
 33.00000  36.89308
```

Yes we can conclude that because p_value is << 0.05.

6.7.15 The data file Datasets.zip/datasets/Ch5/ex5-6-12.csv shows the permeability coefficient (in $10^{-6} cm^3$(STP)/cm $\cdot$ s $\cdot$ MPa) of CO2 measured polyethylene at both 60°C and 61°C. Can you conclude that the mean permeability coefficient at 60°C differs from that at 61°C?

```
> t.test(data_60, data_61)

    Welch Two Sample t-test

data:  data_60 and data_61
t = -1.0236, df = 14.983, p-value = 0.3223
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.133668  2.505097
sample estimates:
mean of x mean of y
 60.40000  62.71429
```

No there is not a difference because the p_value is >> 0.05.

6.7.6 Two weights, each labeled as weighing 100 g, are each weighed several times on the same scale. The results, in units of μg above 100 g, are as follows:

First weight:	53	88	89	62	39	66
Second weight:	23	39	28	2	49	

Since the same scale was used for both weights, and since both weights are similar, it is reasonable to assume that the variance of the weighing does not depend on the object being weighed. Can you conclude that the weights differ?

```
> t.test(FW, SW, var.equal=TRUE)

    Two Sample t-test

data:  FW and SW
t = 3.3306, df = 9, p-value = 0.008791
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 12.17926 63.75407
sample estimates:
mean of x mean of y
 66.16667  28.20000
```

There they differ because p_value << 0.05. Equal Variance.

6.8.1 The article "Improved Bioequivalence Assessment of Topical Dermatological Drug Products Using Dermatopharmacokinetics" (B. N'Dri-Stempfer, W. Navidi, R. Guy, and A. Bunge, Pharmaceutical Research, 2009:316–328) described a study comparing the amounts of econozole nitrate absorbed into human skin for several formulations of antifungal ointment. Both a brand name and generic drug were applied to the arms of 14 subjects, and the amounts absorbed, in μg/cm2, were measured. Following are the results. Can you conclude that the mean amount absorbed differs between the brand name and the generic drug?

Brand Name	Generic	Difference
2.23	1.42	0.81
1.68	1.95	-0.27
1.96	2.58	-0.62
2.81	2.25	0.56
1.14	1.21	-0.07
3.20	3.01	0.19
2.33	2.76	-0.43
4.06	3.65	0.41
2.92	2.89	0.03
2.92	2.85	0.07
2.83	2.44	0.39
3.45	3.11	0.34
2.72	2.64	0.08
3.74	2.82	0.92

```
#both types of ointment tested on the same person -> paired samples
#For T tests the alternative hypothesis can be:
#One sided, H1: mu > mu0, in R this is denoted as: alternative = "greater"
#One sided, H1: mu < mu0, in R this is denoted as: alternative = "less"
#Two sided, H1: mu not equal to mu0, in R this is denoted as:
#alternative = "two.sided"
#In this problem „can you conclude that the means differ?"
# -> mu not equal to mu0 -> two sided
#Now we will take some examples:
```

```
> t.test(B, G, alt="two.sided", paired=TRUE)

    Paired t-test

data:  B and G
t = 1.4593, df = 13, p-value = 0.1682
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.0826935  0.4269792
sample estimates:
mean of the differences
              0.1721429
```

They don't differ because pvalue >> 0.05. Paired Samples.

6.8.5 Two formulations of a certain coating, designed to inhibit corrosion, are being tested. For each of eight pipes, half the pipe is coated with formulation A and the other half is coated with formulation B. Each pipe is exposed to a salt environment for 500 hours. Afterward, the corrosion loss (in μm) is measured for each formulation on each pipe.

Pipe	A	B
1	197	204
2	161	182
3	144	140
4	162	178
5	185	183
6	154	163
7	136	156
8	130	143

Can you conclude that the mean amount of corrosion differs between the two formulations?

```
> t.test(A, B, alt="two.sided", paired=TRUE)

    Paired t-test

data:  A and B
t = -3.0151, df = 7, p-value = 0.01952
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.842571  -2.157429
sample estimates:
mean of the differences
                    -10
```

They differ because p_value is less than 0.05.

6.8.10 A group of eight individuals with high cholesterol levels were given a new drug that was designed to lower cholesterol levels. Cholesterol levels, in mg/dL, were measured before and after treatment for each individual, with the following results:

Subject	Before	After
1	283	215
2	299	206
3	274	187
4	284	212
5	248	178
6	275	212
7	293	192
8	277	196

a. Can you conclude that the mean cholesterol level after treatment is less than the mean before treatment? b. Can you conclude that the reduction in mean cholesterol level after treatment is greater than 75 mg/dL?

```
> t.test(B, A)

    Welch Two Sample t-test

data:  B and A
t = 10.984, df = 13.789, p-value = 3.357e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 63.85366 94.89634
sample estimates:
mean of x mean of y
  279.125   199.750
```

They differ because p_value is less than 0.05.

Chi-square test:

6.10.2 At an assembly plant for light trucks, routine monitoring of the quality of welds yields the following data:

	Number of Welds		
	High Quality	**Moderate Quality**	**Low Quality**
Day Shift	467	191	42
Evening Shift	445	171	34
Night Shift	254	129	17

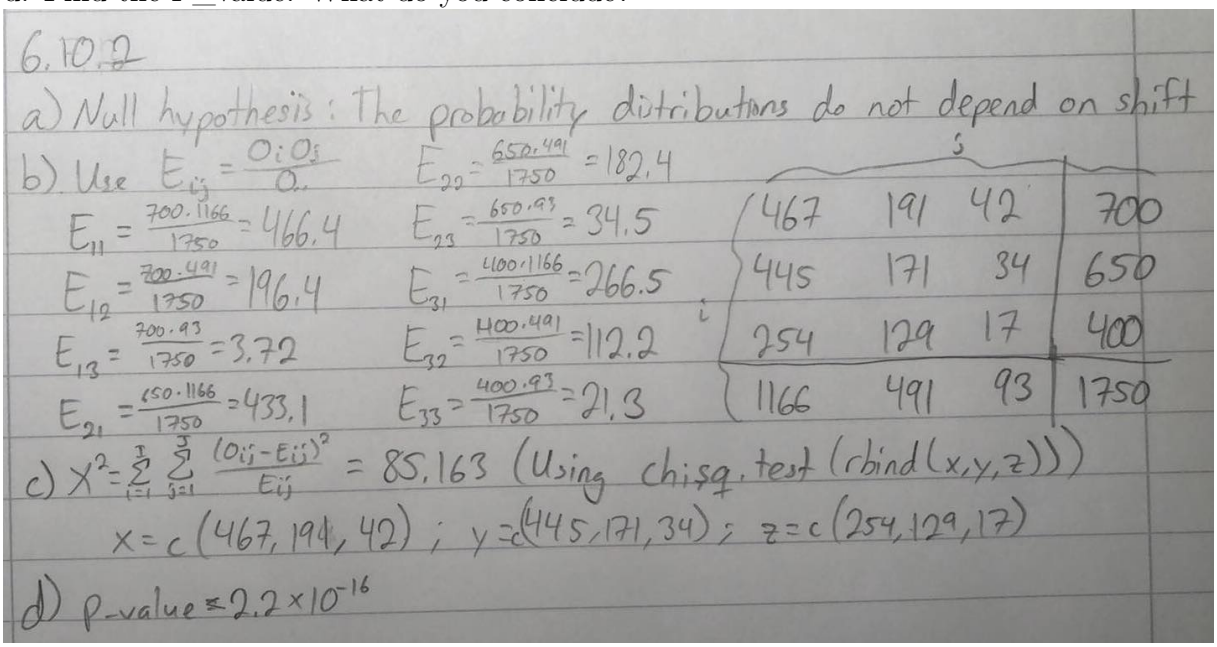
Can you conclude that the quality varies among shifts?

a. State the appropriate null hypothesis.

b. Compute the expected values under the null hypothesis.

c. Compute the value of the chi-square statistic.

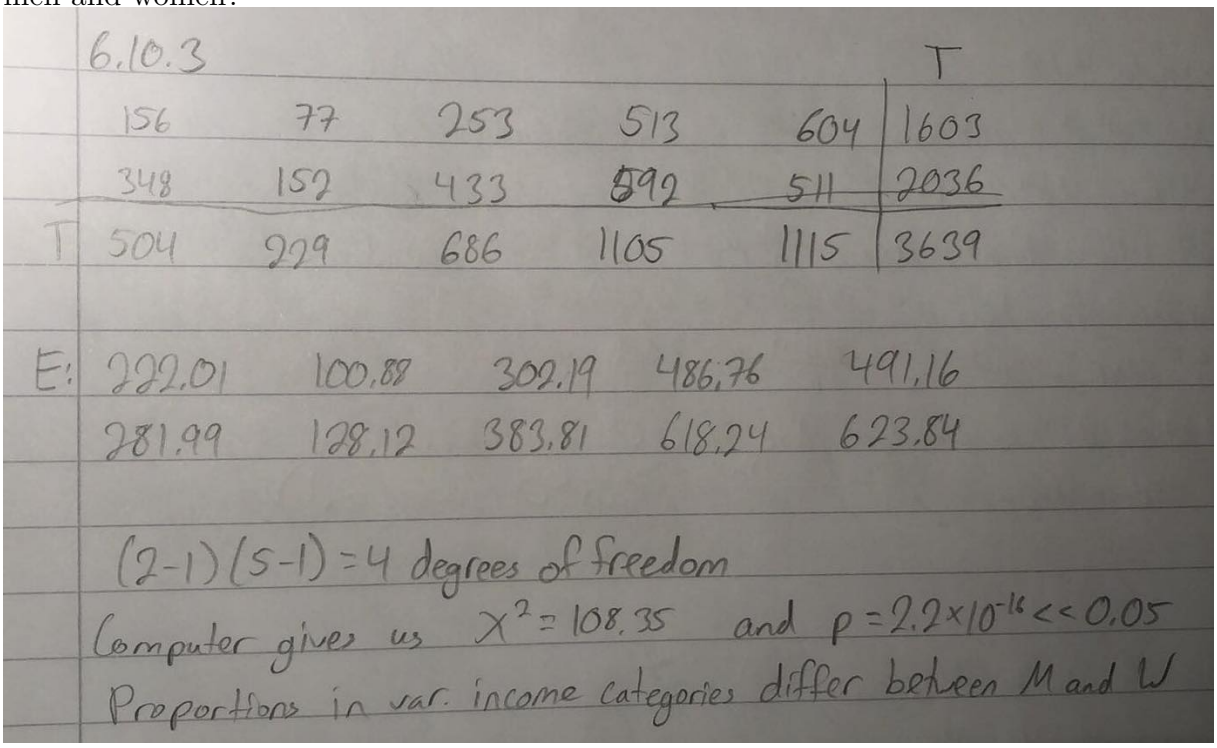
d. Find the P_value. What do you conclude?



6.10.3 The article "Inconsistent Health Perceptions for US Women and Men with Diabetes" (M. McCollum, L. Hansen, et al., Journal of Women's Health, 2007:1421–1428) presents results of a survey of adults with diabetes. Each respondent was categorized by gender and income level. The numbers in each category (calculated from percentages given in the article) are presented in the following table.

Poor	Near Poor	Low Income	Middle Income	High Income	
Men	156	77	253	513	604
Women	348	152	433	592	511

Can you conclude that the proportions in the various income categories differ between men and women?



6.10.4 The article "Analysis of Time Headways on Urban Roads: Case Study from Riyadh" (A. Al-Ghamdi, Journal of Transportation Engineering, 2001:289–294) presents a model for the time elapsed between the arrival of consecutive vehicles on urban roads. Following are 137 arrival times (in seconds) along with the values expected from a theoretical model.

Time	Observed	Expected
0-2	18	23
2-4	28	18
4-6	14	16
6-8	7	13
8-10	11	11
10-12	11	9
12-18	10	20
18-22	8	8
>22	30	19

Can you conclude that the theoretical model does not explain the observed values well?

```
> chisq.test(rbind(O,E))

    Pearson's Chi-squared test

data:  rbind(O, E)
X-squared = 10.72, df = 8, p-value = 0.2181
```

p_val is 0.2181 >> 0.05 so no.

6.10.12 The article "Determination of Carboxyhemoglobin Levels and Health Effects on Officers Working at the Istanbul Bosphorus Bridge" (G. Kocasoy and H. Yalin, Journal of Environmental Science and Health, 2004:1129-1139) presents assessments of health outcomes of people working in an environment with high levels of carbon monoxide (CO). Following are the numbers of workers reporting various symptoms, categorized by work shift. The numbers were read from a graph.

Shift	Morning	Evening	Night
Influenza	16	13	18
Headache	24	33	6
Weakness	11	16	5
Shortness of Breath	7	9	9

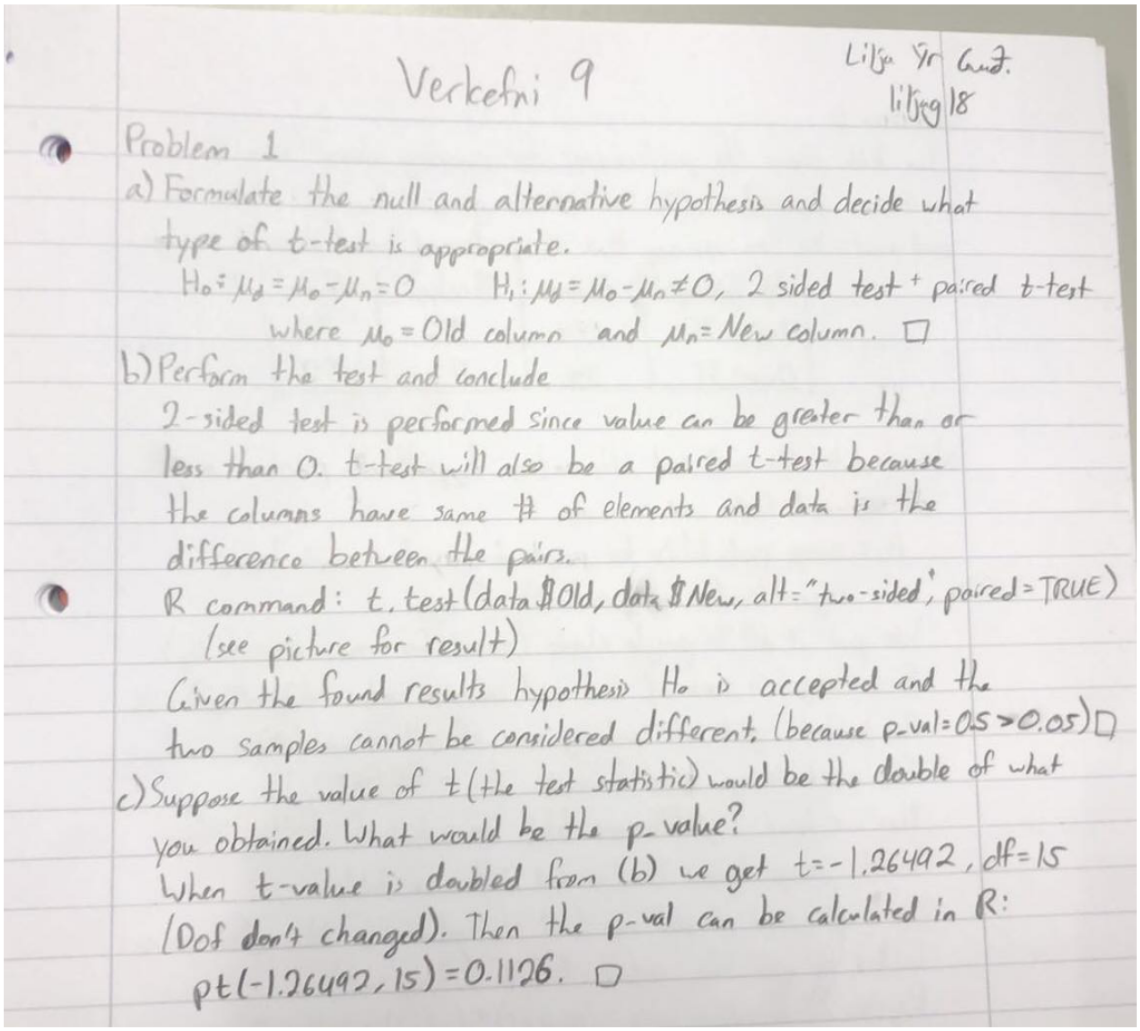
Can you conclude that the proportions of workers with the various symptoms differ among the shifts?

```
> data <- read.csv('ex6-10-12.csv')
> data
                   X Morning Evening Night
1           Influenza      16      13    18
2            Headache      24      33     6
3            Weakness      11      16     5
4 Shortness of Breath       7       9     9
> I = c(16,13,18)
> H = c(24,33,6)
> W = c(11,16,5)
> S = c(7,9,9)
> chisq.test(rbind(W,I,S,H))

    Pearson's Chi-squared test

data:  rbind(W, I, S, H)
X-squared = 17.572, df = 6, p-value = 0.007397
```

There are differences between the shifts.

9.4 Problem Set

Problem 1:



Verkefni 9

Lilja Þr Guð.
liljag18

Problem 1

a) Formulate the null and alternative hypothesis and decide what type of t-test is appropriate.

$H_0: \mu_d = \mu_o - \mu_n = 0$ $H_1: \mu_d = \mu_o - \mu_n \neq 0$, 2 sided test + paired t-test

where μ_o = Old column and μ_n = New column. □

b) Perform the test and conclude

2-sided test is performed since value can be greater than or less than 0. t-test will also be a paired t-test because the columns have same # of elements and data is the difference between the pairs.

R command: t.test(data$Old, data$New, alt="two-sided", paired=TRUE)

(see picture for result)

Given the found results hypothesis H_0 is accepted and the two samples cannot be considered different, (because p-val=0.5 > 0.05) □

c) Suppose the value of t (the test statistic) would be the double of what you obtained. What would be the p-value?

When t-value is doubled from (b) we get t=-1.26492, df=15 (Dof don't changed). Then the p-val can be calculated in R:

pt(-1.26492, 15) = 0.1126. □

```
> data = read.csv("ex6-S-18.csv", header = TRUE)
> data
    Old  New
1  16.3 15.9
2  15.9 16.2
3  15.8 16.0
4  16.2 15.8
5  16.1 16.1
6  16.0 16.1
7  15.7 15.8
8  15.8 16.0
9  15.9 16.2
10 16.1 15.9
11 16.3 15.7
12 16.1 16.2
13 15.8 15.8
14 15.7 15.8
15 15.8 16.2
16 15.7 16.3
> t.test(data$Old,data$New,alt="two.sided",paired=TRUE)

    Paired t-test

data:  data$Old and data$New
t = -0.63246, df = 15, p-value = 0.5366
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2185059  0.1185059
sample estimates:
mean of the differences
                  -0.05

> t2 = 2*(-0.63246)
> pt(t2, 15)
[1] 0.1125996
```

```
data = read.csv("ex6-5-18.csv", header = TRUE)
t.test(data$Old, data$New, alt="two.sided", paired = TRUE)
t2 = 2*(-0.63246)    # t = -0.63246 from the test
pt(t2, 15)  #15 degrees of freedom
```

Problem 2:

Problem 2

The table shows the preferences for running shoes (from 3 unknown brands) of 630 individuals grouped by age: under 35 and over 35. Can you say that the preferred brand depends on age?

Age group	Adidas	Reebok	Nike
Under 35	174	132	90
Over 35	84	72	82

a) Formulate the null and alternative hypothesis, perform a chi squared test, and conclude.

H_0: same probability for people to buy from each brand regardless (or independent) of age.

We put it all in google sheets (excel for mac essentially)
First image is sum of row and column values (O_i, O_j) with the bottom right cell containg sum of all values (O).
The next image uses formula $E_{ij} = \frac{O_i \cdot O_j}{O}$
Then we have a difference table to calculate $\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
in the image below that marked difference table.
Then we sum it all up (all the value in the difference table) to get the value: 8.826 (chisq val)
Degrees of freedom are: (rows - 1) * (col - 1) = 1·2 = 2
Using R (using chisq val) we get

p_val = pchisq(8.826, df = 2, lower.tail = FALSE)
= 0.0121187 < 0.05

This shows that the null hyp is rejected and age does matter when buying shoes from these brands. □

b) What is the probability that the value of x^2 is greater than the obtained value assuming random data?

No idea because I don't know what obtained value is being referred to here. □

	Sum of rows and columns			
				Sum I:
	174	132	90	396
	84	72	78	234
Sum J:	258	204	168	630

	J = 1	J = 2	J = 3
I = 1	162.1714286	128.2285714	105.6
I = 2	95.82857143	75.77142857	62.4

Difference table			
0.8627604953	0.1109243697	2.304545455	
1.460056223	0.1877181642	3.9	
		SUM:	8.826004707

9.5 Problem Set Solution

1. The data file ex6-S-18.csv (from the data collection of the textbook or Canvas/Modules/Other materials/Datasets.zip) contains the diameters of two samples of bearings. One sample is produced with an old technology and the other sample with a new technology. The project manager is asking you for a t-test to find out if the diameters of the two samples can be considered different or not.
a) Formulate the null and the alternative hypotheses and decide what type of t-test is appropriate.
H_0: the diameters of the two samples are equal H_1: the diameters of the two samples are different
t-test with unequal variances.

b) Perform the test and conclude.
p-value = 0.485 For an (incorrect) paired test p-value = 0.536
p-value $> 0.05 \Rightarrow H_0$ is accepted, the diameters cannot be considered different

c) Suppose the value of t (the test statistic) would be the double of what you obtained. What would be the p-value?
For a correct t-test t = -0.70711;2*pt(2*t,29.7) $\approx$ 2*pt(2*t, 30) $\approx$ 0.168

In case of an (incorrect) paired test t = -0.63246 ; 2*pt(2*t,15) $\approx$ 0.225

2. The table shows the preferences for running shoes, from three known brands, of 630 individuals grouped by age: under 35 and over 35. Can you say that the preferred brand depends on age?

Age group	Adidas	Reebok	Nike
Under 35	174	132	90
Over 35	84	72	78

a) Formulate the null and alternative hypotheses, perform a chi square test, and conclude.
H_0: the preferred brand does not depend on the age group H_1: the preferred brand depends on the age group.

```
x=c(174,132,90); y=c(84,72,78)
chisq.test(rbind(x,y))
p-value = 0.01212
```

p-value $< 0.05 \Rightarrow$ The preferred brand and the age are not independent variables. The preferred brand (seems to) depend on the age group.
b) What is the probability that the value of X^2 (the test statistic) is greater than the obtained value, assuming random data?
The obtained value of X^2 is 8.826. According to the definition,the p-value is P($X^2 >$ 8.826) = 0.01212 So no additional calculation is needed. But, optionally, or alternatively, it is a good idea to check:P($X^2 > 8.826$)=1 - pchisq(8.826,2)=0.01212

10 Week 10

10.1 Concepts

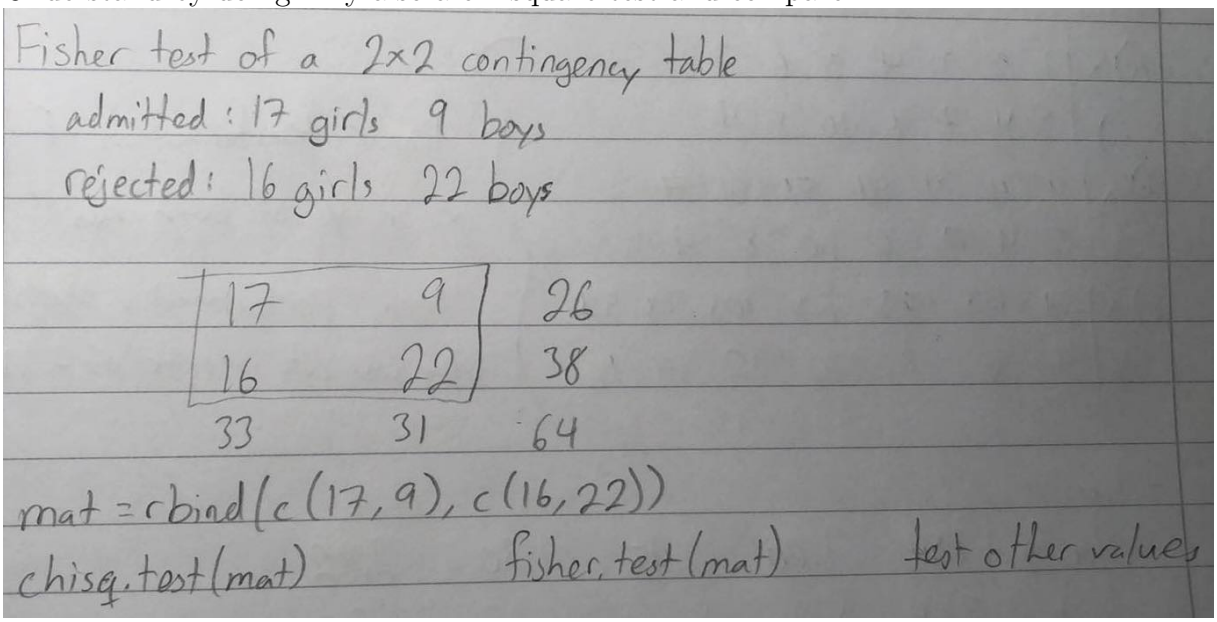
- The covariance of two random variables is always positive, negative, or zero
- The correlation coefficient of two random variables is always between -1 and 1
- The correlation coefficient of two independent variables is expected to be zero
- The 95% confidence interval of the correlation coefficient of two random variables is [-0.15,0.23]. Based on this information we can conclude that the two variables are independent
- The correlation coefficient of two random variables estimated from a data sample is r=-0.6. These two variables can be considered correlated only if the p-value is below 0.05
- The residuals are the differences between the response data y_i and the response predicted by the linear fit
- In a linear regression of the response y to the independent variable x the slope coefficient is 0.5 and the intercept coefficient is 1.5. The predicted response for x=0 is 1.5
- In a regression analysis of the response y to the independent variable x the slope coefficient was 0.5. If now x is considered the response and y the independent variable, the slope will depend on the correlation coefficient
- If the slope coefficient of the response y to the independent variable x is negative then the correlation coefficient of x and y is negative
- The regression coefficients are obtained by minimizing the sum of all squared residuals

10.2 R code

```
mat = rbind(c(17,9), c(16,22))
chisq.test(mat)
fisher.test(mat)
cor.test(x,y)          # testing correlation of x and y
cov(x,y)/(sd(x)*sd(y))          # gives us r used in formula 7.4
cov(dom, fre)
```

10.3 Example Problem Set

Fisher test of a 2x2 contingency table. Experimental approach.
A prestigious university organizes an admission exam for the Master's program. The admitted students are 17 girls and 9 boys. The rejected students are 16girls and 22boys. Do you see a gender bias in this process? Make a 2x2 contingency table and perform a Fisher exact test. Change the numbers and repeat the test several times. Observe the relation between data and p-value. Increase the difference between numbers, change the sample size, etc. What if you have all numbers equal, or two pairs of equal numbers? Understand by doing. Try also a chi square test and compare.

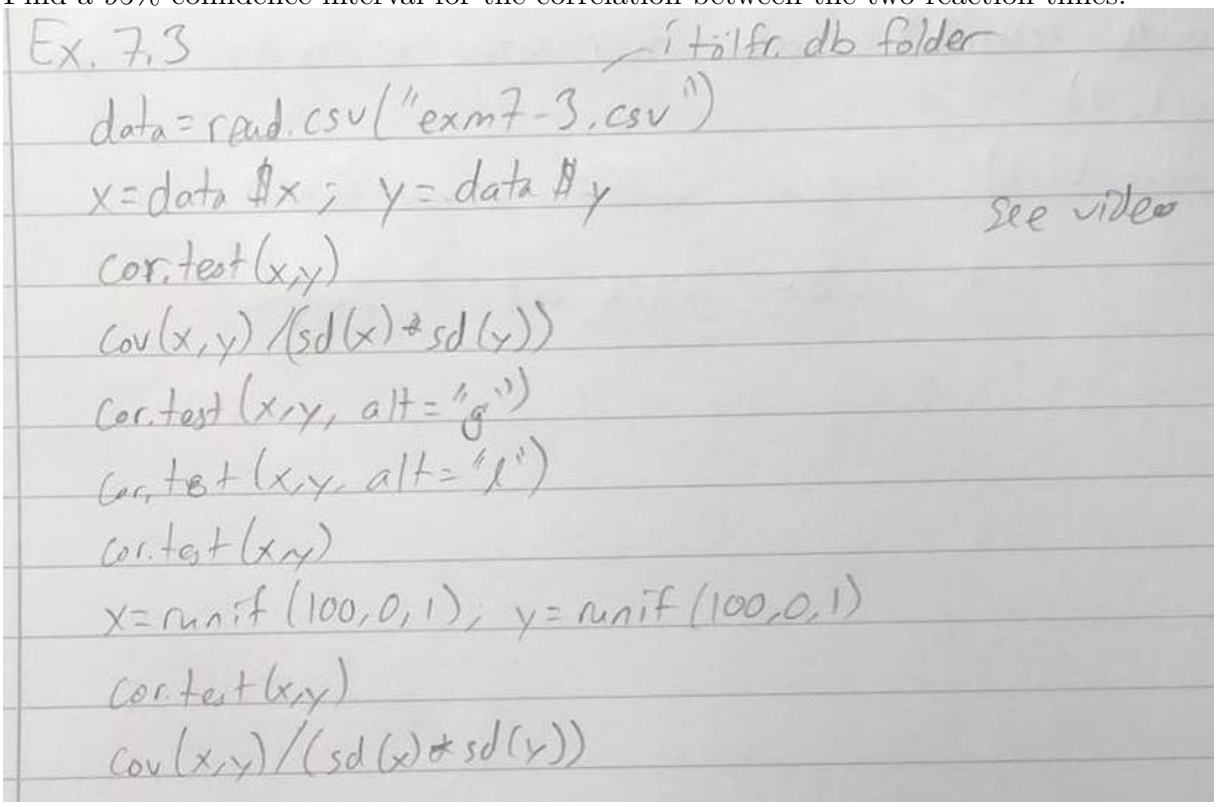


Correlation coefficients and correlation tests:

Example 7.3 In a study of reaction times, the time to respond to a visual stimulus (x) and the time to respond to an auditory stimulus (y) were recorded for each of 10 subjects. Times were measured in ms. The results are presented in the following table.

x	161	203	235	176	201	188	228	211	191	178
y	159	206	241	163	197	193	209	189	169	201

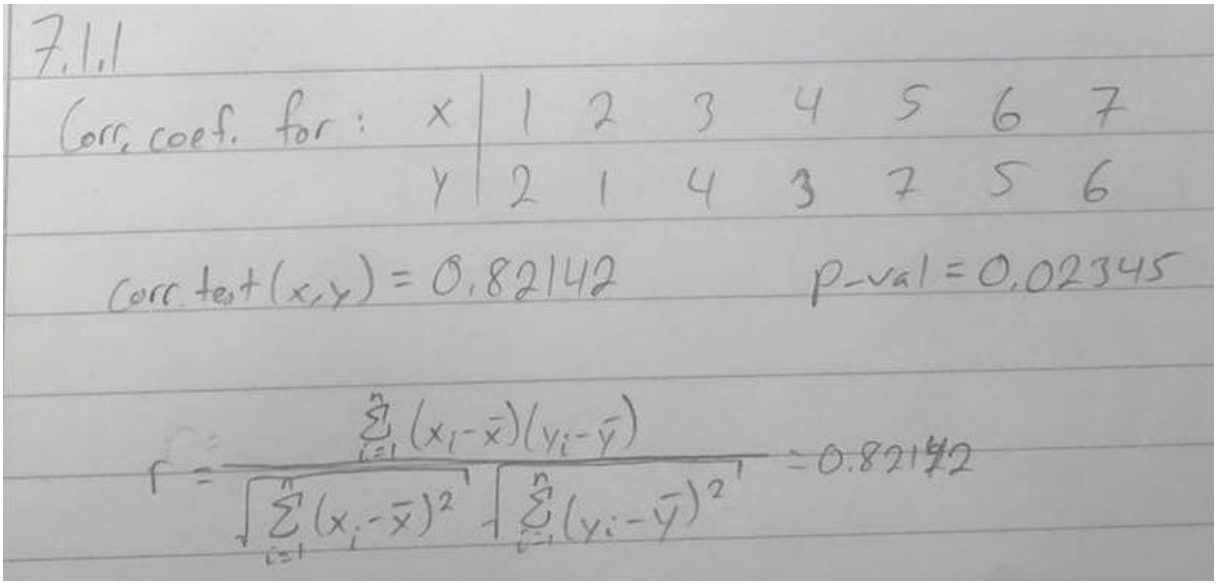
Find a 95% confidence interval for the correlation between the two reaction times.



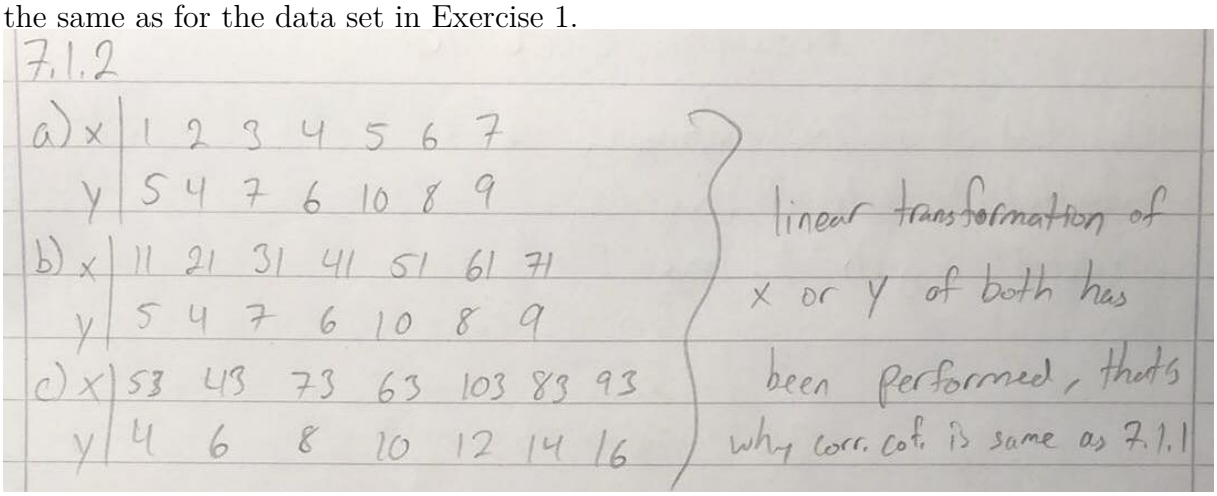
```
> x=data$x
> y=data$y
> cor.test(x,y)

    Pearson's product-moment correlation

data:  x and y
t = 3.991, df = 8, p-value = 0.004
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3830003 0.9549492
sample estimates:
      cor
0.8158796

> cov(x,y)/(sd(x)*sd(y))
[1] 0.8158796
> cor.test(x,y,alt='g')

    Pearson's product-moment correlation

data:  x and y
t = 3.991, df = 8, p-value = 0.002
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.4797593 1.0000000
sample estimates:
      cor
0.8158796

> cor.test(x,y,alt='l')

    Pearson's product-moment correlation

data:  x and y
t = 3.991, df = 8, p-value = 0.998
alternative hypothesis: true correlation is less than 0
95 percent confidence interval:
 -1.0000000  0.9431764
sample estimates:
      cor
0.8158796
```

7.1.1 Compute the correlation coefficient for the following data set

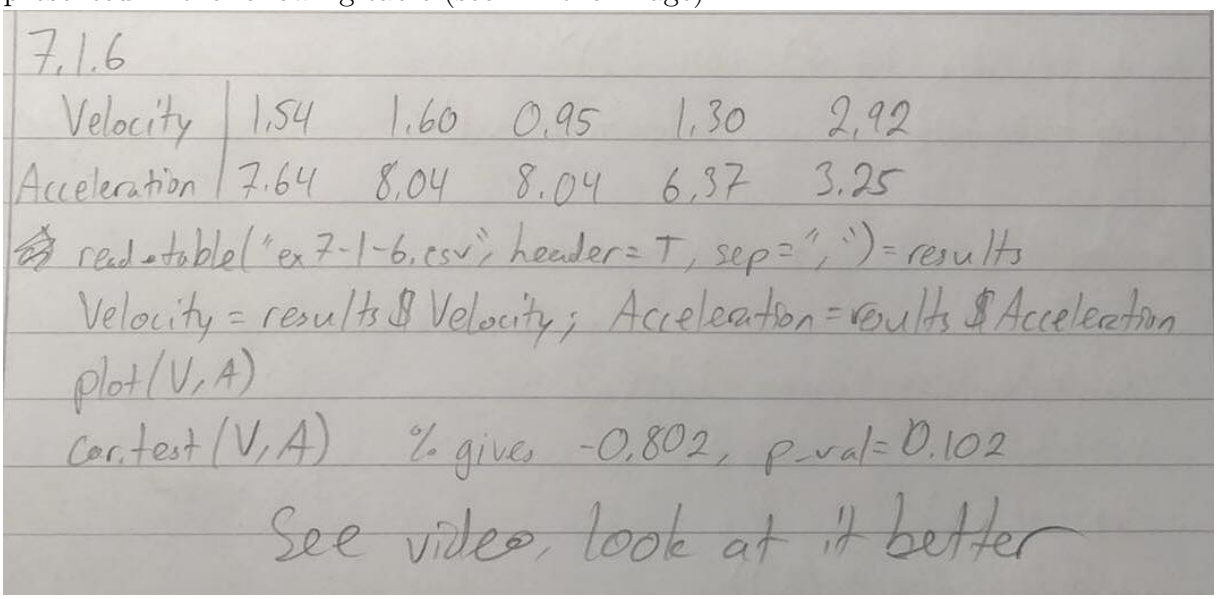
x	1	2	3	4	5	6	7
y	2	1	4	3	7	5	6



7.1.2 For each of the following data sets, explain why the correlation coefficient is the same as for the data set in Exercise 1.



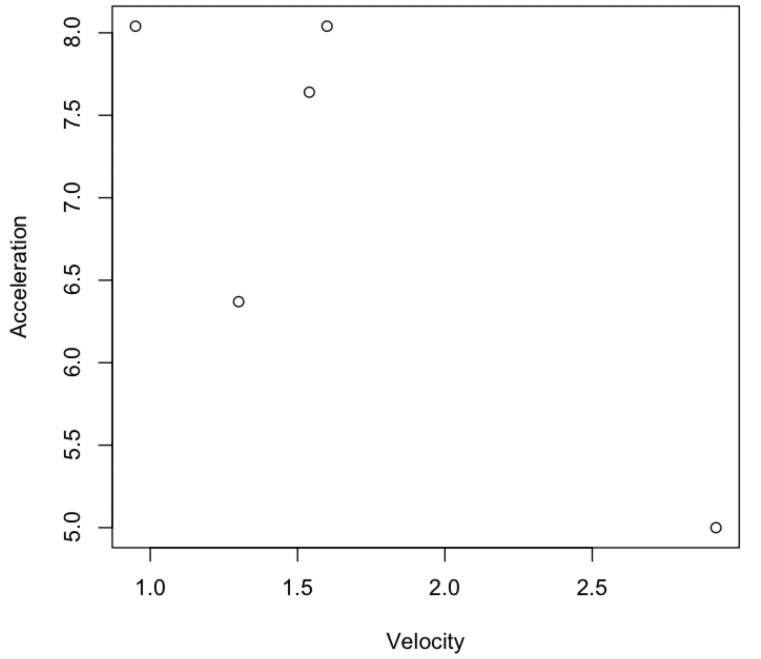
7.1.6 In a study of ground motion caused by earthquakes, the peak velocity (in m/s) and peak acceleration (in m/s^2) were recorded for five earthquakes. The results are presented in the following table (seen in the image).



a. Compute the correlation coefficient between peak velocity and peak acceleration.

```
> cor.test(V,A)

    Pearson's product-moment correlation

data:  V and A
t = -2.3317, df = 3, p-value = 0.102
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9864057  0.2725283
sample estimates:
       cor
-0.8027545
```

b. Construct a scatterplot for these data.



c. Is the correlation coefficient an appropriate summary for these data? Explain why or why not.
I'd say it is because there is a weak correlation between the Velocity and Acceleration as can be seen in the scatterplot.
d. Someone suggests converting the units from meters to centimeters and from seconds to minutes. What effect would this have on the correlation?
This would have no affect because it is just transforming the data a little bit and shouldn't affect the correlation.

7.1.9 Tire pressure (in kPa) was measured for the right and left front tires on a sample of 10 automobiles. Can you conclude that the correlation coefficient is positive? Calculate the p-value corresponding to this hypothesis. Use the data set ex7-1-9.csv. cor.test(r,l,alt="great")

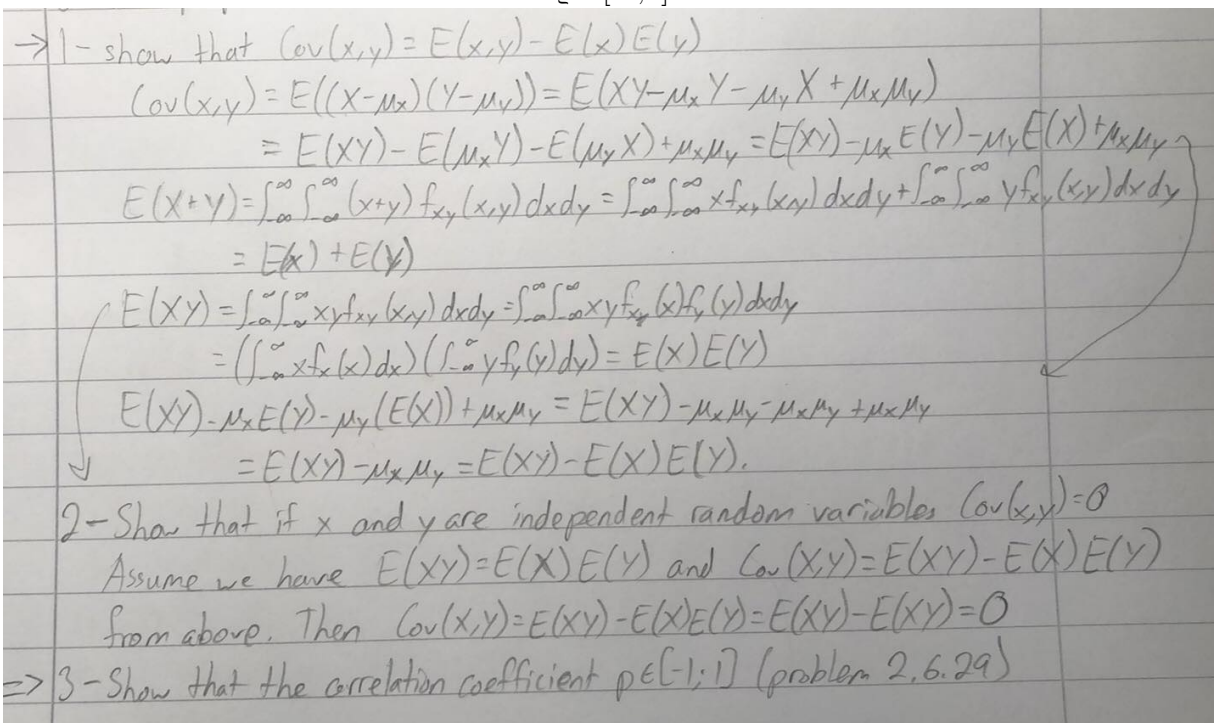
```
> setwd("/Users/liljayrgudmundsdottir/Documents/HR_Files/5onn/Tolfraedi/datasets/Ch7")
> data <- read.csv('ex7-1-9.csv')
> data
   Right Left
1    184  185
2    206  203
3    193  200
4    227  213
5    193  196
6    218  221
7    213  216
8    194  198
9    178  180
10   207  210
> r = data$Right
> l = data$Left
> cor.test(r,l,alt="great")

    Pearson's product-moment correlation

data:  r and l
t = 7.1965, df = 8, p-value = 4.638e-05
alternative hypothesis: true correlation is greater than 0
95 percent confidence interval:
 0.7786318 1.0000000
sample estimates:
      cor
0.9306981
```

General properties of covariance and correlation coefficients:

1 - Show that Cov(x,y)=E(xy)-E(x)E(y)

2 - Show that if x and y are independent random variables Cov(x,y)=0

3 - Show that the correlation coefficient $\varrho \in$[-1,1]. This is Problem 2.6.29



Joint distributions

2.6.2 In a certain community, levels of air pollution may exceed federal standards for ozone or for particulate matter on some days. For a particular summer season, let X be the number of days on which the ozone standard is exceeded and let Y be the number of days on which the particulate matter standard is exceeded. Assume that the joint probability mass function of X and Y is given in the following table:

	y			
x	0	1	2	$p_X(x)$
0	0.10	0.11	0.05	0.26
1	0.17	0.23	0.08	0.48
2	0.06	0.14	0.06	0.26
$p_Y(y)$	0.33	0.48	0.19	

a. Find the marginal probability mass function $p_X(x)$.

See values in the table above for $p_X(x)$. Now when we add them all together we get $\sum_i p_X(x) = 0.26 + 0.48 + 0.26 = 1$ which fulfills the property of the marginal probability mass function where the sum of all possible x and y have to add up to 1.

b. Find the marginal probability mass function $p_Y(y)$.

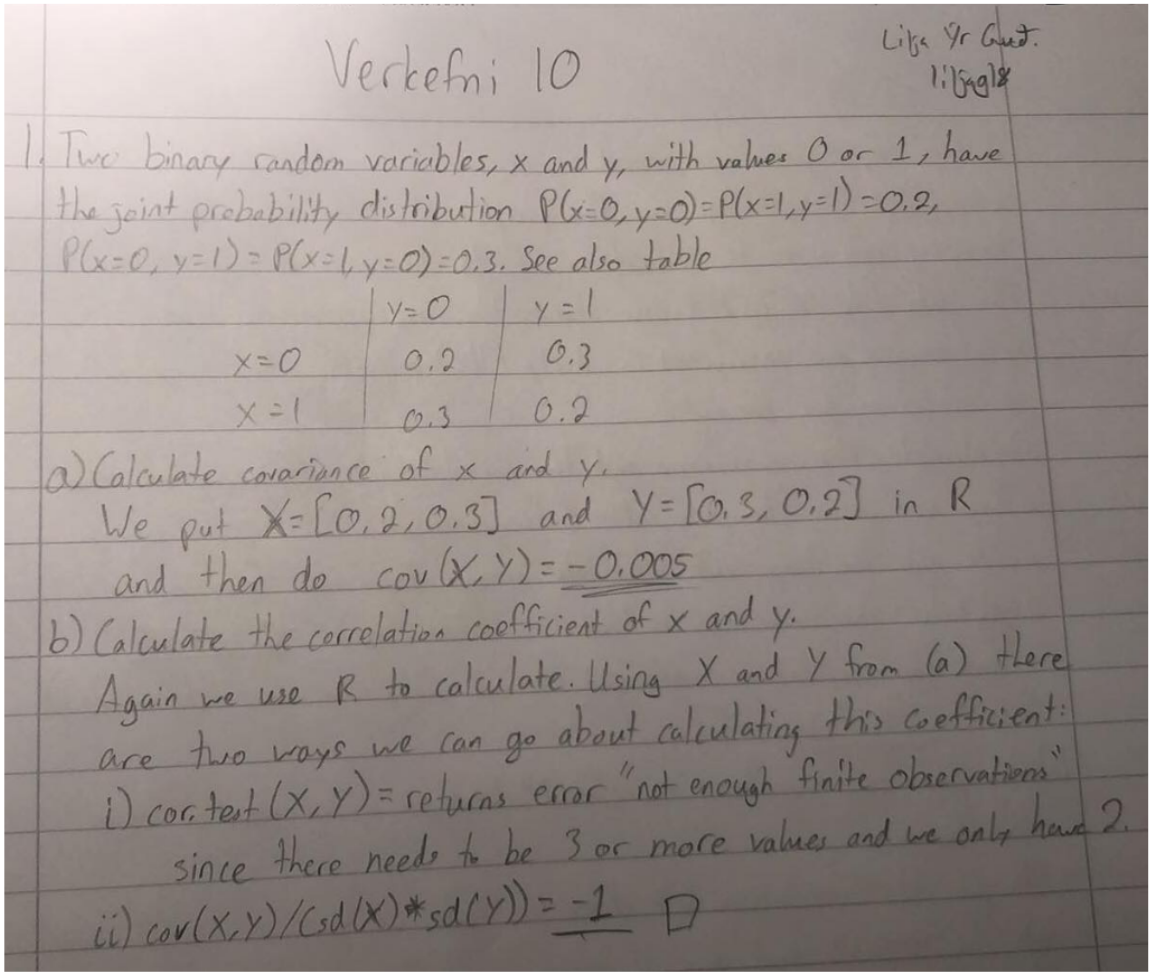
See values in the table above for $p_Y(y)$. Now when we add them all together we get

$\sum_j p_Y(y) = 0.33 + 0.48 + 0.19 = 1$ which fulfills the property of the marginal probability mass function where the sum of all possible x and y have to add up to 1.
c. Find μ_X.
We have $\mu_X = E(x) = \sum_i x_i p(x_i) = 0 * 0.26 + 1 * 0.48 + 2 * 0.26 = 1$.
d. Find μ_Y.
We have $\mu_Y = E(y) = \sum_j y_j p(y_j) = 0 * 0.33 + 1 * 0.48 + 2 * 0.19 = 0.86$.
e. Find σ_X.
We have $E(x^2) = \sum_i x_i^2 p(x_i) = 0^2 * 0.26 + 1^2 * 0.48 + 2^2 * 0.26 = 1.52$. This gives us $\sigma^2 = E(x^2) - (E(x))^2 = 1.52 - 1 = 0.52$
f. Find σ_Y.
We have $E(y^2) = \sum_j y_j^2 p(y_j) = 0^2 * 0.33 + 1^2 * 0.48 + 2^2 * 0.19 = 1.24$. This gives us $\sigma^2 = E(y^2) - (E(y))^2 = 1.24 - 0.86 = 0.38$
g. Find Cov(X, Y).
First we have to find μ_{XY} because $Cov(X, Y) = \mu_{XY} - \mu_X \mu_Y$.
$\mu_{XY} = 0*0*0.10 + 1*0*0.17 + 2*0*0.06 + 0*1*0.11 + 1*1*0.23 + 2*1*0.14 + 0*2* 0.05 + 1*2*0.08 + 2*2*0.06 = 1*1*0.23 + 2*1*0.14 + 1*2*0.08 + 2*2*0.06 = 0.91$
This gives us $Cov(X, Y) = 0.91 - 1 * 0.86 = 0.05$
h. Find $\varrho_{X,Y}$.
We get $\varrho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{0.05}{0.52*0.38} = 0.25$
i. Are X and Y independent? Explain.
For x and y to be considered independent $Cov(X, Y) = 0$ and $\varrho = 0$. As you can see from previous calculations, neither is equal to 0. Correlation coefficient gives info on how much x and y depend on each other.

10.4 Problem Set

Lilja Ýr Guðmundsdóttir - liljag18

Verkefni 10



Verkefni 10

Lilja Ýr Guð.
liljag18

1. Two binary random variables, x and y, with values 0 or 1, have the joint probability distribution $P(x=0,y=0)=P(x=1,y=1)=0.2$, $P(x=0,y=1)=P(x=1,y=0)=0.3$. See also table

	y=0	y=1
x=0	0.2	0.3
x=1	0.3	0.2

a) Calculate covariance of x and y.
We put X=[0.2, 0.3] and Y=[0.3, 0.2] in R and then do cov(X,Y) = -0.005

b) Calculate the correlation coefficient of x and y.
Again we use R to calculate. Using X and Y from (a) there are two ways we can go about calculating this coefficient:
i) cor.test(X,Y) = returns error "not enough finite observations" since there needs to be 3 or more values and we only have 2.
ii) cov(X,Y)/(sd(X)*sd(Y)) = -1 □

Verkefni 10

2. A team of engineers performed an experiment in which accelerometers were placed on a bridge for the purpose of measuring the damping ratios of the vibrations at 18 different frequencies. The data file ex7-2-9.csv shows the measurements. The frequencies are measured in Hertz (Hz) meaning the number of vibrations per second.

a) Calculate the covariance and the correlation coefficient of the two variables (damping ratio & frequency).

We get the data and split into arrays dam and fre as in picture. The covariance is then cov(dam, fre) = -0.13738 and the correlation coefficient (found with cor.test(dam, fre)) is -0.91533. □

b) Are the two variables correlated? Perform a correlation test and explain.

As per the correlation test done in (a) we get a coefficient of -0.91533 and the p-value is 1.016×10^{-7} and the 95 percent confidence interval is -0.968 to -0.783 (as seen in the picture). Since both values on the interval are negative we look at the p-value. Seeing as the value is less than 0.05 ($1.016 \times 10^{-7} << 0.05$) then there is a correlation between the variables. □

c) Suppose we change the units of frequencies from Hz to number of vibrations per minute. Will the correlation coefficient change? Explain.

The correlation coefficient will not change because the coefficient is essentially the relationship between the z-scores of the two variables. So even though the numbers in one variable changes (in this case frequency by a factor of x60) and the standard deviation and mean change, the z-score does not and therefore the coefficient stays the same. □

Verkefni 10

Image for problem 2:

```
> data = read.csv("ex7-2-9.csv", header = TRUE)
> data
   Damping.Ratio Frequency
1            0.3      2.72
2            0.3      2.84
3            0.3      3.77
4            0.4      2.07
5            0.4      2.20
6            0.4      2.34
7            0.4      2.61
8            0.5      1.80
9            0.5      1.93
10           0.5      1.53
11           0.6      0.77
12           0.6      1.26
13           0.6      1.66
14           0.7      0.89
15           0.7      1.00
16           0.7      0.66
17           0.8      1.13
18           0.8      0.37
> dam = data$Damping.Ratio
> dam
 [1] 0.3 0.3 0.3 0.4 0.4 0.4 0.4 0.5 0.5 0.5 0.6 0.6 0.6 0.7 0.7 0.7 0.8 0.8
> fre = data$Frequency
> fre
 [1] 2.72 2.84 3.77 2.07 2.20 2.34 2.61 1.80 1.93 1.53 0.77 1.26 1.66 0.89 1.00 0.66
[17] 1.13 0.37
> cov(dam, fre)
[1] -0.1373758
> cor.test(dam, fre)

    Pearson's product-moment correlation

data:  dam and fre
t = -9.0917, df = 16, p-value = 1.016e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9683741 -0.7831137
sample estimates:
       cor
-0.9153284

>
```

10.5 Problem Set Solution

1. Two binary random variables, x and y, with values 0 or 1, have the joint probability distribution P(x=0, y=0)=P(x=1, y=1)=0.2, P(x=0, y=1)=P(x=1, y=0)=0.3. See also the table

	y=0	y=1
x=0	0.2	0.3
x=1	0.3	0.2

a)Calculate the covariance of x and y.Cov(x,y) = E(x*y) -E(x)*E(y) = 0.2 -(0.5)2= -0.05
b) Calculate the correlation coefficient of x and y.ϱ(x,y) = Cov(x,y) / (σ_x* σ_y) = -0.05 / (0.25) =-0.2

2. A team of engineers performed an experiment in which accelerometers were placed on a bridge for the purpose of measuring the damping ratios of the vibrations at 18 different frequencies.The data file ex7-2-9.csv (from the data collection of the textbook or Canvas / Modules / Other materials / Datasets.zip) shows the measurements. The frequencies are measured in Hertz (Hz), meaning the number of vibrations per second.
a) Calculate the covariance and the correlation coefficient of the two variables(damping ratio and frequency).
cov(dr,f) = -0.1373758, cor(dr,f) = -0.9153284

b) Are the two variables correlated? Perform a correlation test and explain.
cor.test(dr,f)
data: dr and f
t = -9.0917, df = 16, p-value = 1.016e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.9683741 -0.7831137
sample estimates:
cor -0.9153284
The p-value = 1.016e-07 $<$ 0.05 indicates that the two variables are correlated.

c) Suppose we change the units of frequencies from Hz to number of vibrations per minute. Will the correlation coefficient change? Explain.
No, the correlation coefficient does not change by scaling one variable. (Nor by performing other linear transformations of the variables.)

11 Week 11

11.1 Concepts

Under assumptions 1 through 4 (page 549),

- The quantities $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are normally distributed random variables.
- The means of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the true values β_0 and respectively.
- The standard deviations of $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are estimated with

$$s_{\widehat{\beta}_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{7.36}$$

and

$$s_{\widehat{\beta}_1} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{7.37}$$

where $s = \sqrt{\frac{(1-r^2)\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}}$ is an estimate of the error standard deviation σ. Page 552

- The R squared coefficient represents the proportion of the response variance explained by the regression
- In a regression analysis the null hypothesis (H0) about the slope coefficient is the true slope is zero
- In a regression analysis the alternative hypothesis (H1) about the slope coefficient is the true slope is not zero
- If the p-value of the slope coefficient is larger than 0.05 we conclude that the data has no real trend
- The distribution used to calculate the p-value of the regression coefficient is t distribution with the number of DOF equal to the sample size -2

11.2 R code

```
reg=lm(L~W)
summary(reg)
attributes(reg)
reg$coefficients
reg$fitted.values
reg$residuals
confint(reg) #confidence intervals
abline(reg, col="red")
sum=0          # remember to reset sum to 0 every time
for(i in T) {sum = sum + (i-meaan(T))**2}
sum=0
for(i in c(1,2,3,4,5,6,7,8,9)){sum = sum + (T[i]-mean(T))*(C[i]-mean(C))}
```

```
cor.test(T, M, method = "pearson")
mreg=lm(S ~ M + T)          # multiple regression where S is on the y axis
confint(mreg)
```

11.3 Example Problem Set

Simple linear regression:
Examples 7.6-7.8 using table7-1.csv and converting the data in SI units. (Video) In each case verify the numbers obtained for the regression coefficients with the formulas obtained in the lecture material. (equivalent to Eqs.7.14 and 7.15 in the textbook). Flip the variables length and force, repeat the regression,and verify the result that the product of the former and present slopes $= r^2$ Show that the correlation coefficient does not depend on the choice American or European units.
Convert to SI units: L=L*2.54; W=W*4.448 Use European units cm, Newtons (SI)

```
> L= data$Measured.Length..y.
> L
 [1] 5.06 5.01 5.12 5.13 5.14 5.16 5.25 5.19 5.24 5.46 5.40 5.57 5.47 5.53 5.61 5.59 5.61 5.75
[19] 5.68 5.80
> L
 [1] 5.06 5.01 5.12 5.13 5.14 5.16 5.25 5.19 5.24 5.46 5.40 5.57 5.47 5.53 5.61 5.59 5.61 5.75 5.68 5.80
> W = data$Weight..x.
> W
 [1] 0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8
> L=L*2.54; W=W*4.448 #Use European units cm, Netwons (SI)
> L
 [1] 12.8524 12.7254 13.0048 13.0302 13.0556 13.1064 13.3350 13.1826 13.3096 13.8684 13.7160 14.1478
[13] 13.8938 14.0462 14.2494 14.1986 14.2494 14.6050 14.4272 14.7320
> W
 [1]  0.0000  0.8896  1.7792  2.6688  3.5584  4.4480  5.3376  6.2272  7.1168  8.0064  8.8960  9.7856
[13] 10.6752 11.5648 12.4544 13.3440 14.2336 15.1232 16.0128 16.9024
```

Formula 7.14: $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$

Formula 7.15: $\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}$

```
> reg=lm(L~W)
> reg

Call:
lm(formula = L ~ W)

Coefficients:
(Intercept)            W
    12.6993       0.1168
```

Use the formulas to check regression coefficient:

```
> beta1=(mean(W*L)-mean(W)*mean(L))/(mean(W**2)-mean(W)**2)
> beta1
[1] 0.1168492
> beta0=mean(L)-beta1*mean(W)
> beta0
[1] 12.69927
```

Now we flip L and W:

```
> beta1=(mean(W*L)-mean(L)*mean(W))/(mean(L**2)-mean(L)**2)
> beta1
[1] 8.124132
> beta0=mean(W)-beta1*mean(L)
> beta0
[1] -102.7421
> reg=lm(W~L)
> reg

Call:
lm(formula = W ~ L)

Coefficients:
(Intercept)            L
   -102.742        8.124
```

Some regression stuff including r^2:

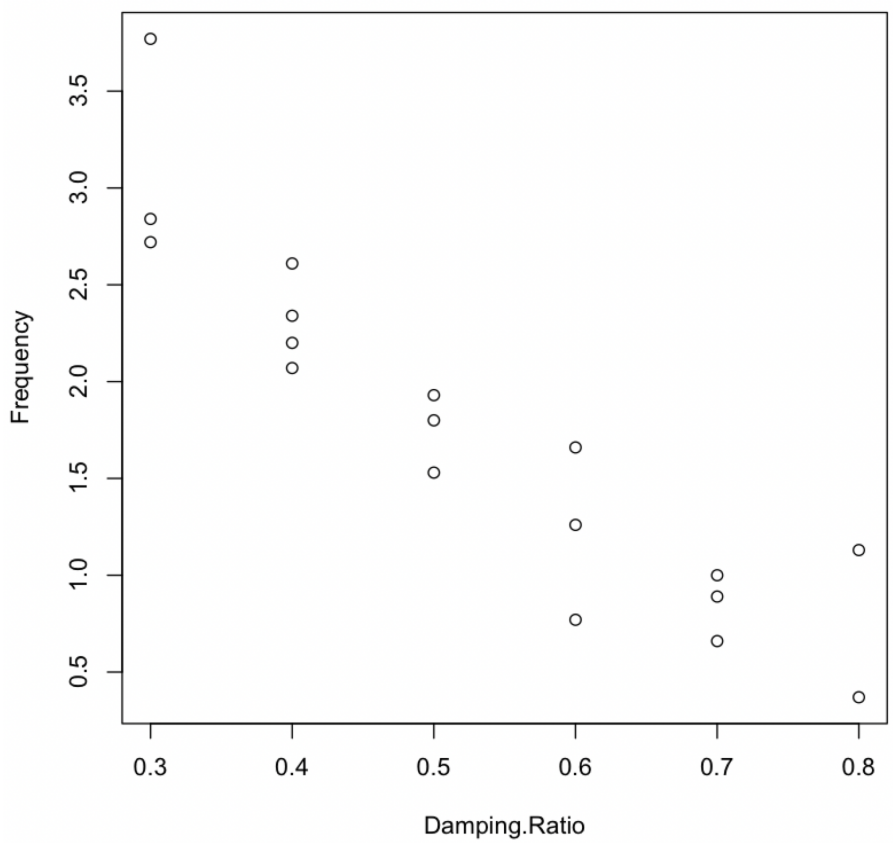
```
> summary(reg)

Call:
lm(formula = L ~ W)

Residuals:
     Min       1Q   Median       3Q      Max
-0.24432 -0.08652 -0.01359  0.09554  0.30509

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.699274   0.062926  201.81  < 2e-16 ***
W            0.116849   0.006365   18.36  4.2e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.146 on 18 degrees of freedom
Multiple R-squared:  0.9493,  Adjusted R-squared:  0.9465
F-statistic:   337 on 1 and 18 DF,  p-value: 4.204e-13

> attributes(reg)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"          "fitted.values" "assign"
 [7] "qr"            "df.residual"   "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"

> reg$coefficients
(Intercept)           W
 12.6992743   0.1168492
> reg$fitted.values
       1        2        3        4        5        6        7        8        9       10       11
12.69927 12.80322 12.90717 13.01112 13.11507 13.21902 13.32297 13.42692 13.53087 13.63482 13.73876
      12       13       14       15       16       17       18       19       20
13.84271 13.94666 14.05061 14.15456 14.25851 14.36246 14.46641 14.57036 14.67431
> reg$residuals
           1            2            3            4            5            6            7            8
 0.153125714 -0.077823308  0.097627669  0.019078647 -0.059470376 -0.112619398  0.012031579 -0.244317444
           9           10           11           12           13           14           15           16
-0.221266466  0.233584511 -0.022764511  0.305086466 -0.052862556 -0.004411579  0.094839398 -0.059909624
          17           18           19           20
-0.113058647  0.138592331 -0.143156692  0.057694286
> confint(reg) #confidence interavls
                2.5 %     97.5 %
(Intercept) 12.5670722 12.8314764
W            0.1034767  0.1302216
```

Not sure how to prove this though.

7.2.11 Structural engineers use wireless sensor networks to monitor the condition of dams and bridges. The article "Statistical Analysis of Vibration Modes of a Suspension Bridge Using Spatially Dense Wireless Sensor Network" (S. Pakzad and G. Fenves, Journal of Structural Engineering, 2009:863–872) describes an experiment in which accelerometers were placed on the Golden Gate Bridge for the purpose of estimating vibration modes. For 18 vertical modes, the system was underdamped (damping ratio < 1). Following are the damping ratios and frequencies for those modes.

Damping Ratio	Frequency (Hz)
0.3	2.72
0.3	2.84
0.3	3.77
0.4	2.07
0.4	2.20
0.4	2.34
0.4	2.61
0.5	1.80
0.5	1.93
0.5	1.53
0.6	0.77
0.6	1.26
0.6	1.66
0.7	0.89
0.7	1.00
0.7	0.66
0.8	1.13
0.8	0.37

a. Construct a scatterplot of frequency (y) versus damping ratio (x). Verify that a linear model is appropriate.

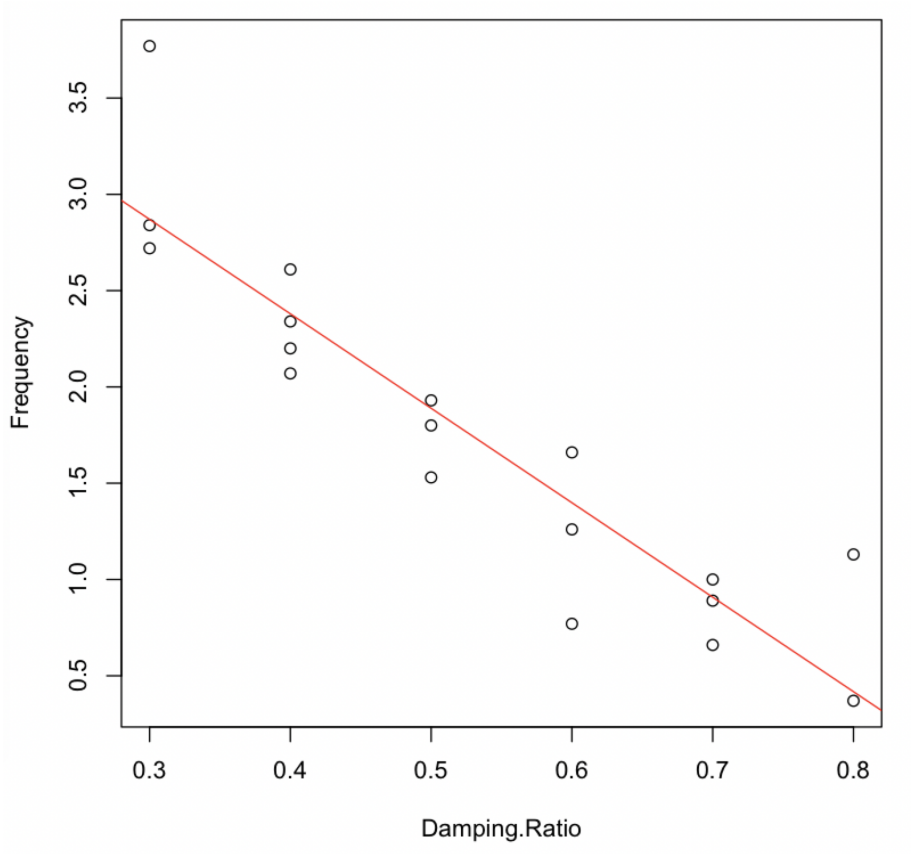


As seen in the image a linear model is appropriate, put in a line using:

```
> D=data$Damping.Ratio
> F=Data$Frequency
Error: object 'Data' not found
> F=data$Frequency
> reg=lm(F~D)
> reg

Call:
lm(formula = F ~ D)

Coefficients:
(Intercept)            D
      4.342       -4.905

> abline(reg, col="red")
```

We also have correlation between the two:

```
> cor(F,D)
> [1] -0.9153284
```

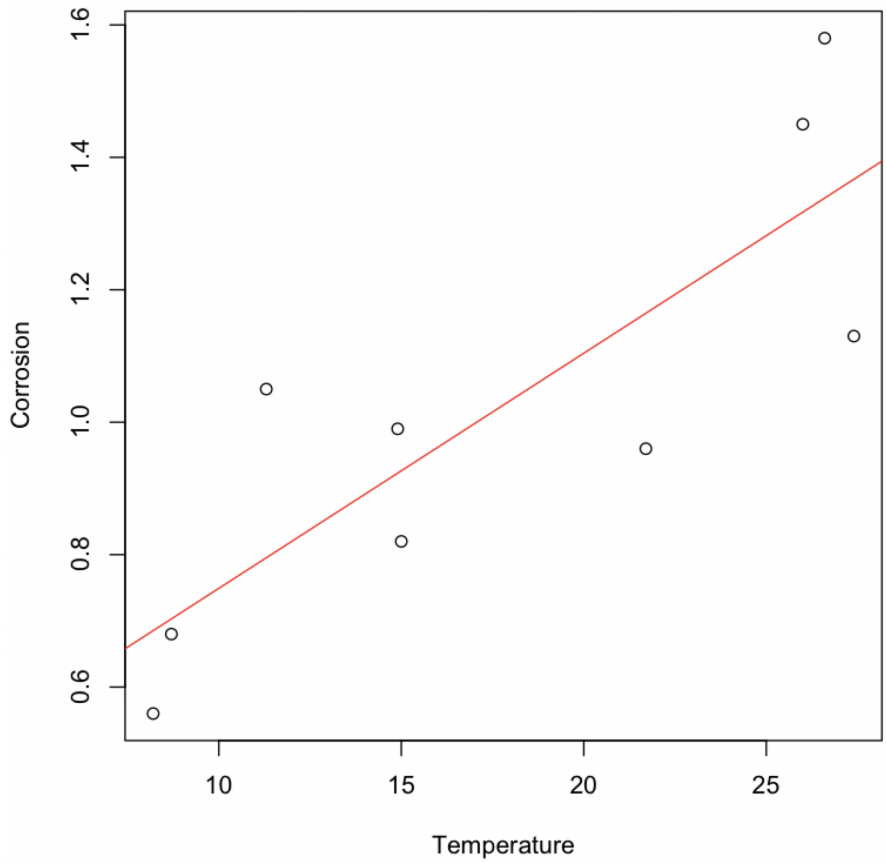


b. Compute the least-squares line for predicting frequency from damping ratio.
We use formula 7.14 and 7.15 where $\overline{x} = 0.527778$ and $\overline{y} = 1.752778$ and $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0.476111$ and $\sum_{i=1}^{n}(y_i - \overline{y})^2 = 13.672761$ and $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = -2.335389$:
$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = -4.905134$ and $\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = 4.341599$
The equation of the least-squares line is $y = 4.341599 - 4.905134x$.
c. If two modes differ in damping ratio by 0.2, by how much would you predict their frequencies to differ?
By 4.905134(0.2) =0.9810 Hz.
d. Predict the frequency for modes with damping ratio 0.75.
We get 4.3415994.905134(0.75) =0.6627.
e. Should the equation be used to predict the frequency for modes that are overdamped (damping ratio > 1)? Explain why or why not.
No it shouldn't because values greater than 1 are outside the range of the data.
f. For what damping ratio would you predict a frequency of 2.0?
If x is the damping ratio then we would get (for frequency=2.0) $x = \frac{2.0 - 4.341599}{-4.905134} = 0.47738$.

7.2.12 The article "Effect of Environmental Factors on Steel Plate Corrosion Under Marine Immersion Conditions" (C. Soares, Y. Garbatov, and A. Zayed, Corrosion Engineering, Science and Technology, 2011:524–541) describes an experiment in which nine steel specimens were submerged in seawater at various temperatures, and the corrosion rates were measured. The results are presented in the following table (obtained by digitizing a graph).

Temperature (°C)	Corrosion (mm/yr)
26.6	1.58
26.0	1.45
27.4	1.13
21.7	0.96
14.9	0.99
11.3	1.05
15.0	0.82
8.7	0.68
8.2	0.56

a. Construct a scatterplot of corrosion (y) versus temperature (x). Verify that a linear model is appropriate.



b. Compute the least-squares line for predicting corrosion from temperature.

We use formula 7.14 and 7.15 where $\overline{x} = 17.75556$ (mean(T)) and $\overline{y} = 1.024444$ and $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 485.5022$ and $\sum_{i=1}^{n}(y_i - \overline{y})^2 = 486.3852$ and $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 503.625$;
$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 1.0373$ and $\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = -17.3939$
These values are wrong and I think I made a calculation error but don't know where so using other answers we get the equation of the least-squares line as $y = 0.393960 - 0.035509x$.
c. Two steel specimens whose temperatures differ by 10°C are submerged in seawater. By how much would you predict their corrosion rates to differ?
We get $0.035509 * 10 = 0.35509mm/yr$.
d. Predict the corrosion rate for steel submerged in seawater at a temperature of 20°C.
Now we use the equation of the least-squares $0.393960 + 0.035509 * 20 = 1.1041mm/yr$.
e. Compute the fitted values.

```
> reg$fitted.values
        1         2         3         4         5         6         7         8         9
1.3385033 1.3171978 1.3669106 1.1645084 0.9230461 0.7952131 0.9265970 0.7028892 0.6851347
```

f. Compute the residuals. Which point has the residual with the largest magnitude?

```
> reg$residuals
          1           2           3           4           5           6           7           8
 0.24149673  0.13280223 -0.23691060 -0.20450837  0.06695394  0.25478693 -0.10659697 -0.02288924
          9
-0.12513466
```

g. Compute the correlation between temperature and corrosion rate.
We use cor(T,C) = 0.8326265.
h. Compute the regression sum of squares, the error sum of squares, and the total sum of squares.

```
> reg

Call:
lm(formula = C ~ T)

Coefficients:
(Intercept)            T
    0.39396      0.03551

> summary(reg)

Call:
lm(formula = C ~ T)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23691 -0.12513 -0.02289  0.13280  0.25479

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.393960   0.171536   2.297  0.05526 .
T           0.035509   0.008927   3.978  0.00534 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1967 on 7 degrees of freedom
Multiple R-squared:  0.6933,  Adjusted R-squared:  0.6494
F-statistic: 15.82 on 1 and 7 DF,  p-value: 0.00534
```

Confidence intervals for regression coefficients:
Obtain the 95% confidence intervals of the spring constant corresponding to table7-1.csv in European units (Video). (See Example 7.11 Edition 4 = Example 7.12 Edition 5 for American units)
First we find the spring constant which was computed before using formula 7.14 and gives us $\widehat{\beta_1} = 0.2046$. Using Hooke's law we get $s_{\widehat{\beta_1}} = \frac{\sqrt{\frac{(1-r^2)\sum_{i=1^n}(y_i-\overline{y})^2}{n-2}}}{\sqrt{\sum_{i=1^n}(x_i-\overline{x})^2}} = \frac{0.0575}{\sqrt{26.6}} = 0.0111$.
The number of degrees of freedom is n-2 = 20=2 = 18 giving us a $t_{18,0.025} = 2.101$. Therefore we get the confidence interval: $0.2046 \pm (2.101) * (0.0111) = (0.181, 0.228)$. This is only valid for the given data and cannot be used as generalized knowledge.

7.3.9 In a study to determine the relationship between ambient outdoor temperature and the rate of evaporation of water from soil, measurements of average daytime temperature in °C and evaporation in mm/day were taken for 40 days. The results are shown in the following table.

Temp.	Evap.
11.8	2.4
21.5	4.4
16.5	5.0
23.6	4.1
19.1	6.0
21.6	5.9
31.0	4.8
18.9	3.0
24.2	7.1
19.1	1.6
11.8	3.8
24.2	5.0
15.8	2.6
26.8	8.0
24.8	5.4
26.2	4.2
14.2	4.4
14.1	2.2
30.3	5.7
15.2	1.2
18.6	3.5
25.4	5.5
22.1	4.8
25.4	4.8
22.6	3.2
24.4	5.1
15.8	3.3
22.3	4.9
23.2	7.4
19.7	3.3
14.0	1.1
13.6	3.5
25.4	5.1
17.7	2.0
24.7	5.7
24.3	4.7
25.8	5.8
28.3	5.8
29.8	7.8
26.5	5.1

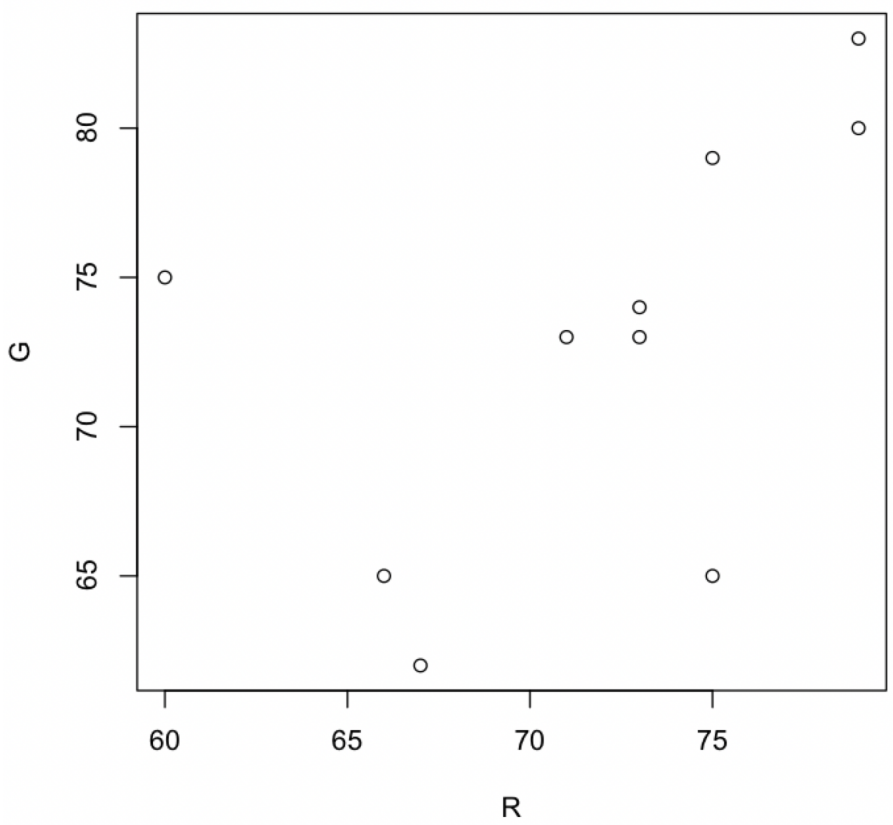
a. Compute the least-squares line for predicting evaporation (y) from temperature (x). We use formula 7.14 and 7.15 where $\overline{x} = 21.5075$ (mean(T)) and $\overline{y} = 4.48$ and $\sum_{i=1}^{n}(x_i -$

$\overline{x})^2 = 1072.528$ and $\sum_{i=1}^n (y_i - \overline{y})^2 = 112.624$ and $\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) = 239.656$:
$\widehat{\beta_1} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = 0.22345$ and $\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = -0.32584$
This gives us least-squares line $y = -0.32584 + 0.22345x$.
b. Compute 95% confidence intervals for β_0 and β_1.
First we must compute $r^2, s, s_{\widehat{\beta_0}}, and s_{\widehat{\beta_1}}$:
$r^2 = \frac{|\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})|^2}{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2} = 0.4755$
$s = \sqrt{\frac{(1-r^2)\sum_{i=1}^n (y_i - \overline{y})^2}{n-2}} = 1.2468$
$s_{\widehat{\beta_0}} = s * \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}} = 0.8422$
$s_{\widehat{\beta_1}} = \frac{s}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} = 0.0381$
Now with 40-2=38 degrees of freedom we get $t_{38,0.025} \approx 2.024$ This gives us a confidence interval for $\widehat{\beta_0} = -0.32584 \pm 2.024 * 0.8422 = (-2.031, 1.379)$ and $\widehat{\beta_1} = 0.22345 \pm 2.024 * 0.0381 = (0.146, 0.301)$
c. Predict the evaporation rate when the temperature is 20°C.
We get $\widehat{\beta_0} + \widehat{\beta_1} * 20 = 4.1432$.

Effect of outliers:
7.4.9 The National Assessment for Educational Progress measured the percentage of eighth grade students who were proficient in reading and the percentage of students who graduated from high school in each state in the U.S. The results for the ten most populous states are as follows:

State	Reading Proficiency	Graduation Rate
California 60 75		
Texas	73	74
New York	75	65
Florida	66	65
Illinois	75	79
Pennsylvania	79	83
Ohio	79	80
Michigan	73	73
Georgia	67	62
North Carolina	71	73

a. Construct a scatterplot of graduation rate (y) versus reading proficiency (x). Which state is an outlier?



The outlier is clearly California, the point to the far right.

b. Compute the least-squares line for predicting graduation rate from reading proficiency, using the data from all ten states.

We use formula 7.14 and 7.15 where $\overline{x} = 71.8$ and $\overline{y} = 72.9$ and $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 323.6$ and $\sum_{i=1}^{n}(y_i - \overline{y})^2 = 438.9$ and $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = 192.8$:

$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = 0.5952$ and $\widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x} = 30.1614$

This gives us the equation $y = 30.122 + 0.596x$.

c. Remove the outlier and compute the least-squares line, using the data from the other nine states.

Using the same calculations as above we get the equation $y = -22.714 + 1.305$.

d. Is the outlier an influential point? Explain.

The outlier is influential because the equation changes quite a bit from when it is part of it and when it is not.

12 Week 11

12.1 Concepts

Continuation of multiple and simple regression using:

$$r^2 = \frac{|\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})|^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}$$

$$s = \sqrt{\frac{(1 - r^2)\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-2}}$$

$$s_{\widehat{\beta_0}} = s * \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

$$s_{\widehat{\beta_1}} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

12.2 R code

```
summary(reg)
cor.test(T, M, method="pearson")    #cor.test(T, M) also works
```

12.3 Example Problem Set

Multiple regression This is an exercise inspired by problems 8.1.1-8.1.2-8.1.3 using the data file ex8-1-3.csv.

a) Do a simple linear regression of Strength using Manganese as independent variable and find the slope coefficient, its 95% confidence interval, and its p-value. Find also the correlation coefficient r. Is the correlation between the Strength and Manganese significant?

We get cor(S,M) = 0.8775644 for correlation.

First we must compute $r^2, s, s_{\widehat{\beta_0}}, ands_{\widehat{\beta_1}}$:

$r^2 = 0.7701$ taken from:

```
> reg

Call:
lm(formula = M ~ S)

Coefficients:
(Intercept)            S
    -3.8513       0.2389

> summary(reg)

Call:
lm(formula = M ~ S)

Residuals:
     Min       1Q   Median       3Q      Max
-0.50569 -0.14668 -0.03403  0.23481  0.49268

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.85127    1.49400  -2.578    0.019 *
S            0.23887    0.03076   7.765 3.73e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2811 on 18 degrees of freedom
Multiple R-squared:  0.7701,  Adjusted R-squared:  0.7573
F-statistic:  60.3 on 1 and 18 DF,  p-value: 3.735e-07
```

$s = \sqrt{\frac{(1-r^2)\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-2}} = 1.3741$

$s_{\widehat{\beta_0}} = s * \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}} = 3.1195$

$s_{\widehat{\beta_1}} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}} = 0.5524$

Now with 20-2=18 degrees of freedom we get $t_{18,0.025} \approx 2.101$ This gives us a confidence interval for $\widehat{\beta_0} = 23.5713 \pm 2.101 * 3.1195$ (M is x) and $\widehat{\beta_1} = 3.223 \pm 2.101 * 0.5524$

b) Repeat for the response of Strength to the independent variable Thickness.

First we must compute $r^2, s, s_{\widehat{\beta_0}}, ands_{\widehat{\beta_1}}$:

$r^2 = 0.05151$ taken from summary(reg).

$s = \sqrt{\frac{(1-r^2)\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-2}} = 2.1512$
$s_{\widehat{\beta_0}} = s * \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}} = 1.095$
$s_{\widehat{\beta_1}} = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}} = 0.3284$
Now with 20-2=18 degrees of freedom we get $t_{18,0.025} \approx 2.101$ This gives us a confidence interval for $\widehat{\beta_0} = 51.3664 \pm 2.101 * 1.095$ (M is x) and $\widehat{\beta_1} = -0.3166 \pm 2.101 * 0.3284$
c) Perform a correlation test for Manganese and Thickness and evaluate if these two variables can be considered independent.
Using R we get cor(M,T) = 0.08590841 and:

```
> cor.test(T, M, method = "pearson")

    Pearson's product-moment correlation

data:  T and M
t = 0.36583, df = 18, p-value = 0.7188
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.3707052  0.5090760
sample estimates:
       cor
0.08590841
```

And since p_value is 0.7188 >> 0.05 so they are not considered independent.
d) Do a multiple regression of the Strength with the variables Manganese and Thickness and find the slope coefficients, their 95% confidence intervals, their p-values, and the R-squared coefficient.
Here is single regression:

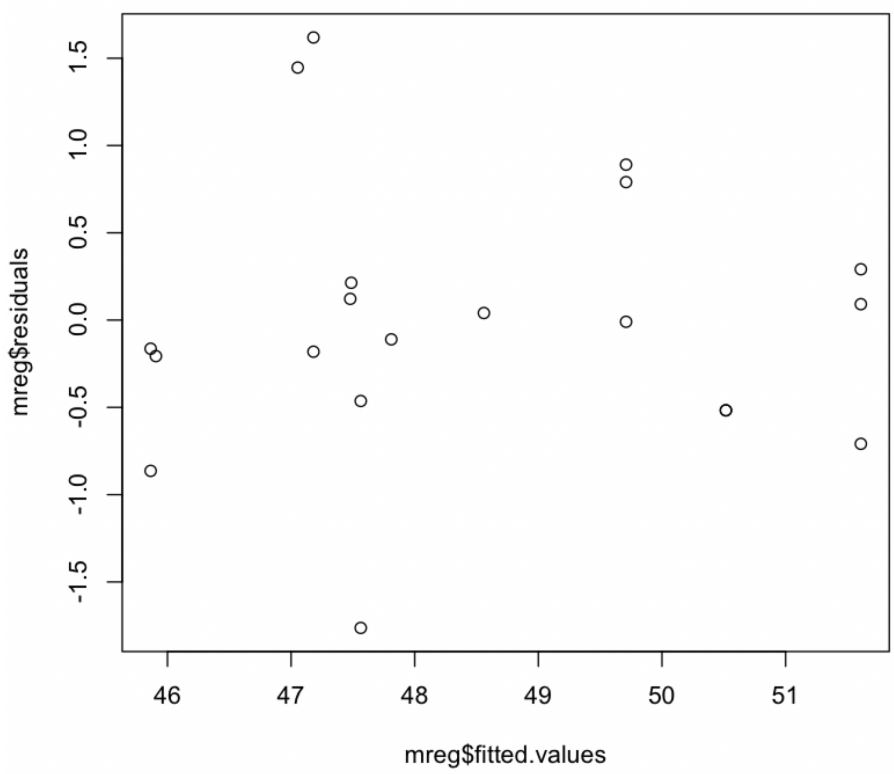
```
> sreg1=lm(S~M)
> summary(sreg1)

Call:
lm(formula = S ~ M)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1065 -0.4424  0.1428  0.5984  1.6936

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  23.5714     3.2217   7.316 8.54e-07 ***
M             3.2240     0.4152   7.765 3.73e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.033 on 18 degrees of freedom
Multiple R-squared:  0.7701,  Adjusted R-squared:  0.7573
F-statistic:  60.3 on 1 and 18 DF,  p-value: 3.735e-07

> sreg2=lm(S ~ T)
> summary(sreg2)

Call:
lm(formula = S ~ T)

Residuals:
   Min     1Q Median     3Q    Max
-3.097 -1.658 -0.157  1.383  3.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.3664     2.9120  17.639 8.32e-13 ***
T            -0.3166     0.3202  -0.989    0.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.098 on 18 degrees of freedom
Multiple R-squared:  0.05151, Adjusted R-squared:  -0.001187
F-statistic: 0.9775 on 1 and 18 DF,  p-value: 0.3359
```

```
> sreg2=lm(S ~ T)
> summary(sreg2)

Call:
lm(formula = S ~ T)

Residuals:
   Min     1Q Median     3Q    Max
-3.097 -1.658 -0.157  1.383  3.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.3664     2.9120  17.639 8.32e-13 ***
T            -0.3166     0.3202  -0.989    0.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.098 on 18 degrees of freedom
Multiple R-squared:  0.05151, Adjusted R-squared:  -0.001187
F-statistic: 0.9775 on 1 and 18 DF,  p-value: 0.3359
```

And here is the multiple regression:

```
> mreg=lm(S ~ M + T)
> summary(mreg)

Call:
lm(formula = S ~ M + T)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76341 -0.47679 -0.06003  0.23305  1.61900

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  26.6408     2.7234   9.782 2.14e-08 ***
M             3.3201     0.3320  10.001 1.55e-08 ***
T            -0.4249     0.1261  -3.371  0.00363 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8228 on 17 degrees of freedom
Multiple R-squared:  0.8622,  Adjusted R-squared:  0.846
F-statistic: 53.19 on 2 and 17 DF,  p-value: 4.823e-08

> confint(mreg)
                2.5 %     97.5 %
(Intercept) 20.8950303 32.3865779
M            2.6197043  4.0205221
T           -0.6908562 -0.1589444
> mreg$residuals
           1            2            3            4            5            6            7            8
-0.110439305 -0.708797177  0.091202823  0.291202823 -0.516928804  0.790381759 -0.516928804 -0.009618241
           9           10           11           12           13           14           15           16
 0.890381759  0.213663096 -0.463408045 -0.863806860  0.121572014 -0.163806860 -0.180997778 -0.206296889
          17           18           19           20
 1.619002222 -1.763408045  1.446472311  0.040557998
> mreg$fitted.values
       1        2        3        4        5        6        7        8        9       10       11
47.81044 51.60880 51.60880 51.60880 50.51693 49.70962 50.51693 49.70962 49.70962 47.48634 47.56341
      12       13       14       15       16       17       18       19       20
45.86381 47.47843 45.86381 47.18100 45.90630 47.18100 47.56341 47.05353 48.55944
```

e) Compare you results of the multiple regression with those shown in the textbook page 606 Edition 4 or page 613 Edition 5.



In the graph above you can see the residual vs the fitted values for the multiple regression. Compared to the example in the book mine is a bit more spread out from side to side and has more points grouped at the top.

f) Is the multiple regression a better approximation of the data than the simple regressions? Explain.

Yes because it is more accurate because it has more than one variable at play.