Statistics and data analysis 2020

Final Exam (Alef)

Guidelines

- There are **4** (**FOUR**) questions in the exam. You need to answer **all** of them (no choice).
- You can respond in English and/or Hebrew.
- Write the answers to the questions in exam notebooks. Don't use the exam printout.
- Justify all your answers. Even though many of the questions are not purely mathematical, you should mathematically explain your answers. You may assume results proven (or stated as a fact) in class or in the homework (unless the question instructs otherwise).
- Make sure you write in a clear and legible way. Grading will also depend on the clarity and not only on correctness.
- You can use the reference and formulae sheet as provided, including the standard normal table.
- Use normal approximation when appropriate and needed.
- You can use hand held calculators.
- No other auxiliary material can be used during the exam.
- The total time of the exam is 3 (three) hours.
- Good luck!

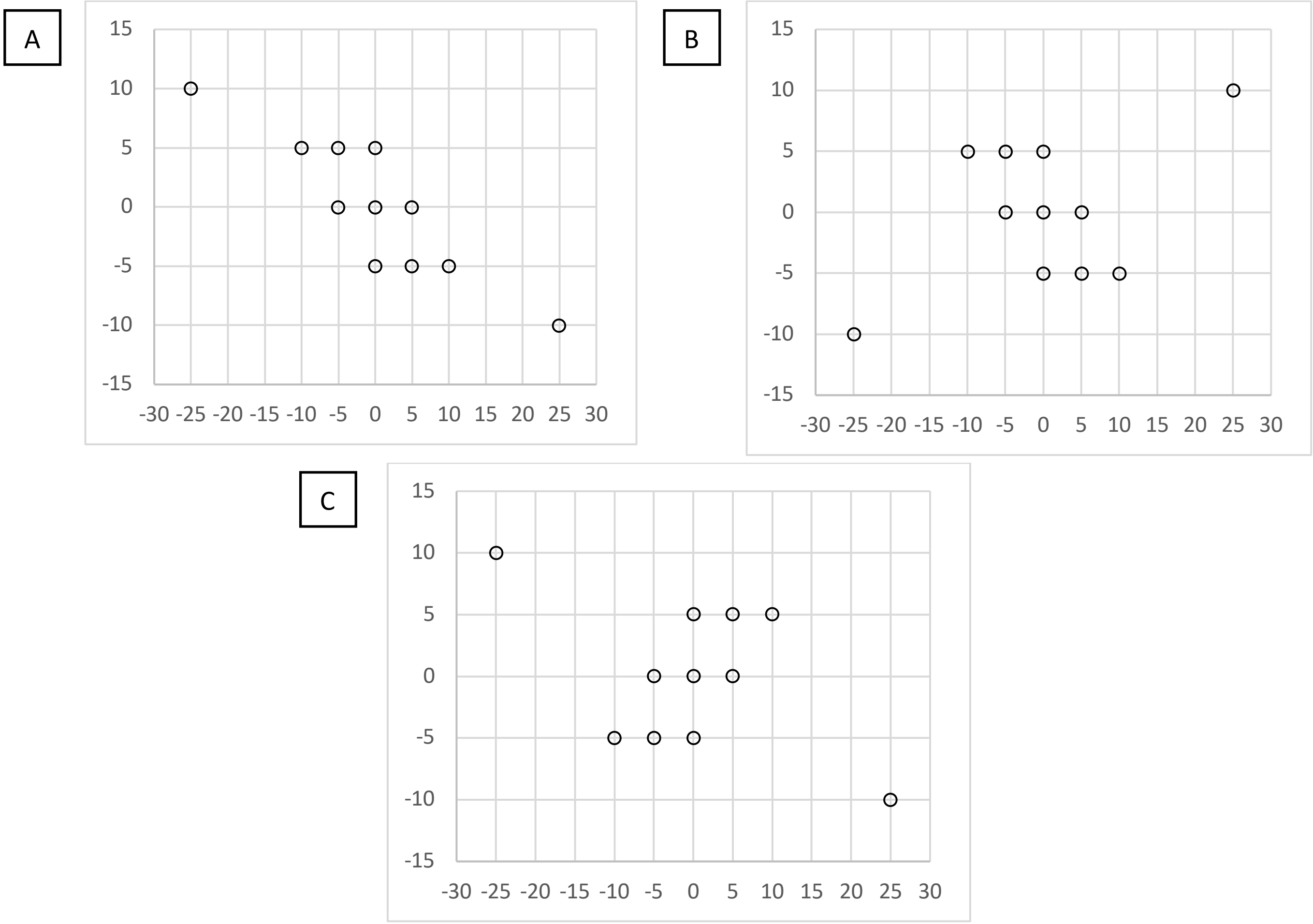
Question 1 (25 pts)

A. (6 pts)
Consider the pairs of observed measurements below. There are three of them. Determine a matching between Pearson and Spearman correlation values in the rows of Table 1 below and the letter enumeration (A to C in Fig 1) of the depicted cases. Indicate the matching clearly in your notebook.

Table 1:

Number (to be matched to the figures)	Pearson correlation	Spearman correlation
1	0.47	0.1
2	-0.88	-0.84
3	-0.47	-0.1

Fig 1 (A-C):



B. (7 pts)

Consider:

$$(x_1, y_1), \dots, (x_{3n}, y_{3n})$$

Where $n \geq 3$.

Let $\tau(v, u)$ = the Kendall correlation on the vectors v and u.

TRUE or FALSE:

If:

$\tau\big((x_1, \dots, x_n), (y_1, \dots, y_n)\big) \geq 0.5$ and

$\tau\big((x_{n+1}, \dots, x_{2n}), (y_{n+1}, \dots, y_{2n})\big) \geq 0.5$ and

$\tau\big((x_{2n+1}, \dots, x_{3n}), (y_{2n+1}, \dots, y_{3n})\big) \geq 0.5$

Then:

$$\tau\big((x_1, \dots, x_{3n}), (y_1, \dots, y_{3n})\big) \geq 0.5$$

Prove your answer.

C. (12 pts)

Safety tests were conducted for cars made in the Randomistan Opel factory and in the Germany Opel factory. Higher safety scores are better. The house statistician in the company had decided to declare the Randomistan cars safer if the WRS p-value of the observed data is better than 0.15.

For each one of the situations described below state whether Randomistan cars are declared safer.

In #3 also state what would happen if one were to use Student t-test rather than a WRS test.

Explain your calculations and answers.

1. (3 pts)

Randomistan	9.3	8.8	8.5		
Germany	9.1	8.2	8.1	8	7.9

2. (3 pts)

Randomistan	8.8	8.6	8.1		
Germany	9.2	9.1	9	8.9	5

3. (6 pts)

Randomistan	10	0.06	0.05					
Germany	9.95	9.9	9.85	0.04	0.03	0.02	0.01	0

Question 2 (25 pts)

A. (Total 10pts)

1. (5 pts) Define three random variables W, X and Y that assume values on the non-negative integers so that:
    a. W, X and Y assume at least two values with non-zero probability (they are not constant)
    b. Let Z = W+X+Y. Then Z is uniformly distributed over the numbers {0, 1, 2, … , 128, 129}.
2. (5 pts) TRUE or FALSE: H(W) < 4 and H(X) < 4 and H(Y) < 4. Explain.

B. (7 pts) $X \sim NegBinom(r, p)$ where $0 < p < 1$.

Given that:

$P(X = 1) = 0$

$P(X = 2) = \frac{1}{9}$

Compute the values of $E(X)$ and $V(X)$.

C. (Total 8pts)

A research team preforms blood test to evaluate differences, in molecular blood components (glucose, cholesterol, triglycerides, etc.), comparing people who practice a certain diet to people who do not.

They considered 100 different features and computed one sided WRS p-values between the two groups.

These p-values are 0<p1<=p2<=…<=p100<=1.

Assume: p5=0.006, p10=0.011, p20=0.05.

1. (4 pts) Is it possible that the team can report a set of features from this study with FDR better than 0.1? Explain your answer.
2. (4 pts) Is it possible that there is no set of features from this study with FDR better than 0.1? Explain your answer.

Question 3 (25 pts)

A. (Total 15pts)
Consider the Markov Chain $X_0, X_1, X_2, \dots, X_n, \dots$ defined over the states 1, 2, 3, 4 given by:

$$T = \begin{pmatrix} 0.1 & 0.7 & 0.1 & 0.1 \\ 0.05 & 0.95 & 0 & 0 \\ 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0.9 & 0 & 0.1 \end{pmatrix}$$

1. (5 pts) What is $P(X_2 = 2|X_0 = 1)$?
2. (5 pts) What is $P(X_1 = 2)$ assuming that $X_0 \sim Unif(1,2,3,4)$?
3. (5 pts) Assuming that $X_0 \sim Unif(1,2,3,4)$.
   TRUE or FALSE:
   $H(X_{21}) < 2.$
   Explain your answer.

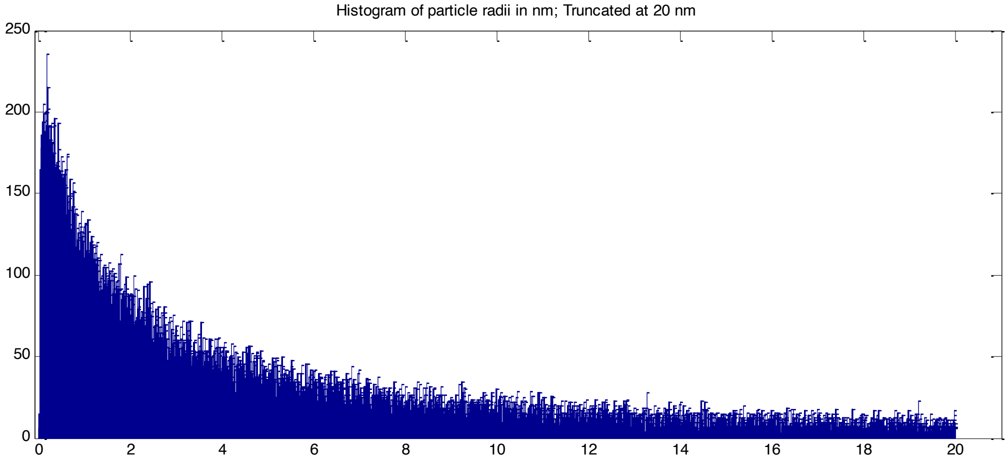
B. (10 pts) Give an example of a transition matrix T of a Markov Chain defined over the states 1, 2, 3 for which the uniform distribution is not stationary. Furthermore, your matrix should satisfy $\forall i, j\ T(i, j) \neq 0$.

Question 4 (25 pts)

<u>This question has 5 parts numbered A-E</u>.

Parts A-C (Total 15pts)

A scientist is generating nanoparticles for an experiment. She observes the following distribution of particle radii, in nms (nano-meters):



This histogram representation of the distribution is calculated from 100K particles. The x-axis units are nms. The histogram is truncated at 20 nm. 30687 particles of the 100K measured had radius ≥ 20 nm.

For the above data representing 100K particles, the scientist calculated empirical statistics.

The empirical mean of the data is $\hat{\mu} = e^4$ nm

The empirical standard deviation is $\hat{\sigma} = \sqrt{e^{12} - e^8}$ nm.

The empirical median of the distribution is at e^2 nm.

Let R denote the random variable that represents radii of the particles generated by the scientist.

R~LogNorm

A. (5 pts)
According to the model you have developed what is the radius r so that
# of particles with radius < r = 20000? (leave answer in exp notation if necessary)

B. (5 pts)
The experiment requires at most 10% of particles to have a radius larger than e^4 nm. Show, based on your model, that the population generated here is therefore not adequate for the experiment.

C. (5 pts)
The scientist can treat the particles and decrease all particle radii.

A reasonably priced process will lead to all radii decreasing exactly $\sqrt{e}$ fold (a particle with radius r will have radius $r\ \cdot 1/\sqrt{e}$ after the treatment).
A more expensive process will lead to all radii decreasing exactly e fold (a particle with radius r will have radius $r\ \cdot 1/e$ after the treatment).
She consulted with her statistician colleague as to whether either of the treatments will solve the problem and specifically as to whether the less expensive one will do it.
What advice would you give in this case? Show all your calculations.

Parts D and E (Total 10pts)

Let $W = -Y$ where $Y \sim LogN(\mu, \sigma^2)$.

D. (5 pts) Express the CDF of W in term of Φ (the CDF of $N(0,1)$).
E. (5 pts) What is the PDF of W?

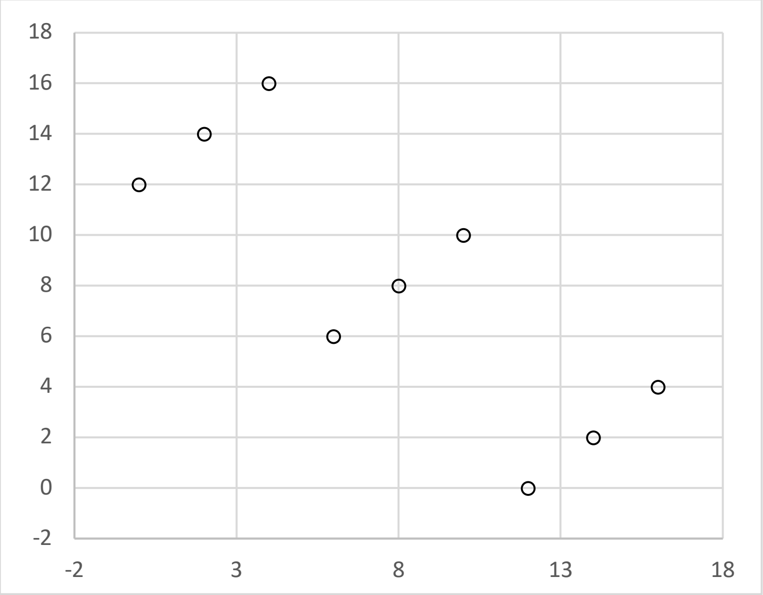
Question 1

A. A – 2
B – 1
C – 3

B. (7 pts)

FALSE

Consider the following data where n=3:



Group	Index	x	y
1	1	0	12
1	2	2	14
1	3	4	16
2	4	6	6
2	5	8	8
2	6	10	10
3	7	12	0
3	8	14	2
3	9	16	4

$$\tau(group\ 1) = 1, \qquad \tau(group\ 2) = 1, \qquad \tau(group\ 3) = 1$$

And

$$\tau(all\ data) = \frac{9 - 27}{\binom{9}{2}} = -\frac{18}{36} = -0.5$$

C.

1.

Randomistan	9.3	8.8	8.5		
Germany	9.1	8.2	8.1	8	7.9

Score	9.3	9.1	8.8	8.5	8.2	8.1	8	7.9
Rank	1	2	3	4	5	6	7	8

N=8

B=3

We observe $T = 1 + 3 + 4 = 8$.

Thus:

$$\mu_T = \frac{3*9}{2} = 13.5$$

$$\sigma_T = \sqrt{\frac{3*5*9}{12}} = \sqrt{11.25}$$

$$Z(T) = \frac{8 - 13.5}{\sqrt{11.25}} = -1.64$$

$$p\_value = 0.0505$$

$\Rightarrow$ Randomistan cars are better.

2.

Randomistan	8.8	8.6	8.1		
Germany	9.2	9.1	9	8.9	5

Score	9.2	9.1	9	8.9	8.8	8.6	8.1	5
Rank	1	2	3	4	5	6	7	8

N=8

B=3

$$\mu_T = \frac{3*9}{2} = 13.5$$

$$\sigma_T = \sqrt{\frac{3*5*9}{12}} = \sqrt{11.25}$$

$$Z(T) = \frac{18 - 13.5}{\sqrt{11.25}} = 1.34$$

$$p\_value \ = 0.9$$

⇒ Randomistan car are NOT better.

3.

WRS:

Randomistan	10	0.06	0.05					
Germany	9.95	9.9	9.85	0.04	0.03	0.02	0.01	0

Score	10	9.95	9.9	9.85	0.06	0.05	0.04	0.03	0.02	0.01	0
Rank	1	2	3	4	5	6	7	8	9	10	11

N=11

B=3

$$\mu_T = \frac{3 * 12}{2} = 18$$

$$\sigma_T = \sqrt{\frac{3 * 8 * 12}{12}} = \sqrt{24}$$

$$Z(T) = \frac{12 - 18}{\sqrt{24}} = -1.22$$

$$p\_value \ = 0.11$$

⇒ Randomistan car are better.

<u>t-test</u>:

In order to use t-test will first calculate the means:

$$\hat{\mu}_R = \frac{10 + 0.06 + 0.05}{3} = 3.37$$

$$\hat{\mu}_G = \frac{9.95 + 9.9 + 9.85 + 0.04 + 0.03 + 0.02 + 0.01 + 0}{8} = 3.725$$

$$\hat{\mu}_G > \hat{\mu}_R$$

The mean score in Randomistan is smaller than the mean score in Germany and therefore the t-test result won't be significant. We can not reject $\mu_G \geq \mu_R$.

Question 2

A.

1. W = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12} with p=1/13
   X = {0, 13, 26, 39, 52} with p=1/5
   Y = {0, 65} with p=1/2

   Z = W + X + Y = {0, 1, 2, …, 129} with p=1/130

2. TRUE
   All three variables W, X, and Y have less than 16 possible values. For such random variables the maximal possible entropy is less than $\log 16 = 4$. Therefore, the statement is True.

B. $P(X = 1) = 0 \rightarrow r > 1$

$P(X = 2) = \frac{1}{9} \rightarrow r \leq 2$

We got that $r = 2$.

$P(X = 2) = \binom{1}{1}(1 - p)^0 p^2 = p^2 \rightarrow p = \frac{1}{3}$.

(The only way to have two successes in two trails).

$$E(X) = \frac{r}{p} = \frac{2}{\frac{1}{3}} = 6$$

$$V(X) = \frac{r(1-p)}{p^2} = \frac{2 * \frac{2}{3}}{\frac{1}{9}} = 12$$

C.

1. It is possible:
   First option: give p1 a very small value: 10^{-6} for example. Then we have:
   $$FDR(1) = \frac{10^{-6} * 100}{1} = 0.001$$

   Second option: p6=p7=p8=p9=0.006:
   $$FDR(9) = \frac{0.006 * 100}{9} = 0.067$$
   There are infinitely more correct answers to this section ….
2. It is also possible:
   p1=p2=p3=p4=p5=0.006
   p6=p7=p8=p9=p10=0.011

p11=p12=p13=p14=p15=p16=p17=p18=p19=p20=0.05
p21=…=p100=1

$$FDR(5) = \frac{0.006 * 100}{5} = 0.12$$

$$FDR(10) = \frac{0.011 * 100}{10} = 0.11$$

$$FDR(20) = \frac{0.05 * 100}{20} = 0.25$$

$$FDR(100) = 1$$

The rest of the FDRs are greater than 0.1 as well.

Question 3

A.

Consider the Markov Chain $X_0, X_1, X_2, \dots, X_n, \dots$ defined over the states 1, 2, 3, 4 given by:

$$T = \begin{pmatrix} 0.1 & 0.7 & 0.1 & 0.1 \\ 0.05 & 0.95 & 0 & 0 \\ 0.1 & 0.8 & 0.1 & 0 \\ 0 & 0.9 & 0 & 0.1 \end{pmatrix}$$

1. In order to calculate the probability, we need to multiply the first row (which is $X_1 | X_0 = 1$) by the second column:

$$0.1 * 0.7 + 0.7 * 0.95 + 0.1 * 0.8 + 0.1 * 0.9 = 0.905$$

2. $P(X_1 = 2)$ assuming that $X_0 \sim Unif(1,2,3,4)$ is:

$$P(X_1 = 2 \wedge X_0 = 1) + (X_1 = 2 \wedge X_0 = 2) + (X_1 = 2 \wedge X_0 = 3) + (X_1 = 2 \wedge X_0 = 4)$$
$$= (X_1 = 2 | X_0 = 1) * P(X_0 = 1) + (X_1 = 2 | X_0 = 2) * P(X_0 = 2) + (X_1 = 2 | X_0 = 3) * P(X_0 = 3) + (X_1 = 2 | X_0 = 4) * P(X_0 = 4)$$
$$= 0.7 * 0.25 + 0.95 * 0.25 + 0.8 * 0.25 + 0.9 * 0.25$$
$$= 0.8375$$

3. TRUE

Let $\pi = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$

$$\pi \cdot T \neq \pi$$

The uniform distribution is <u>not</u> stationary

Given that $X_0 \sim Unif(1,2,3,4)$, we get that $\forall i \;\; X_i$ is not uniform including X_{21}.

According to the entropy definition we get that:

$$H(X_{21}) < 2$$

Another solution:

$$P(X_{21} = 2) \geq \text{probability to stay at 2 from } X_0 \text{ to } X_{21}$$
$$= P(X_1 = 2) * 0.95^{20} = 0.8375 * 0.95^{20} = 0.3$$

Therefore, X_{21} is not $Unif(1,2,3,4)$ and again, according to the fact that entropy only maxes at the uniform distribution, we get: $H(X_{21}) < 2$

B.

$$T = \begin{pmatrix} 0.98 & 0.01 & 0.01 \\ 0.98 & 0.01 & 0.01 \\ 0.98 & 0.01 & 0.01 \end{pmatrix}$$

$$\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \begin{pmatrix} 0.98 & 0.01 & 0.01 \\ 0.98 & 0.01 & 0.01 \\ 0.98 & 0.01 & 0.01 \end{pmatrix} = (0.98, 0.01, 0.01) \neq \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

Question 4

A. Given that the median is e^2 we can conclude that the mean of the underlying normal distribution is 2.
According to the LogNormal mean formula:

$$e^{\left(\mu+\frac{\sigma^2}{2}\right)} = e^{2+\frac{\sigma^2}{2}} = e^4$$

$$2+\frac{\sigma^2}{2} = 4$$

$$\sigma = 2$$

Now, we can calculate:

$$\Phi^{-1}(0.2) = \frac{\ln(r)-2}{2}$$

$$-0.84 = \frac{\ln(r)-2}{2}$$

And therefore $\ln(r) = 0.32$ and $r = e^{0.32}$

B. We will calculate $CDF(e^4)$:

$$\Phi\left(\frac{\ln e^4 - 2}{2}\right) = \Phi\left(\frac{4-2}{2}\right) = \Phi(1) = 0.84$$

We have 16% of the particles with radius larger than e^4.

C. A reasonably priced process:
The new median is $e^{1.5}$ and therefore the new underlying mean is 1.5.
The new empirical mean is $e^{3.5}$ and therefore the new underlying std is inferred by:

$$e^{\left(\mu+\frac{\sigma^2}{2}\right)} = e^{1.5+\frac{\sigma^2}{2}} = e^{3.5}$$

$$1.5+\frac{\sigma^2}{2} = 3.5$$

$$\sigma = 2$$

For $CDF(e^4)$ we now have:

$$\Phi\left(\frac{\ln e^4 - 1.5}{2}\right) = \Phi\left(\frac{4-1.5}{2}\right) = \Phi(1.25) = 0.8944$$

A more expensive process:
The new median is e^1 and therefore the new underlying mean is 1.
The new empirical mean is e^3 and therefore:

$$e^{\left(\mu+\frac{\sigma^2}{2}\right)} = e^{1+\frac{\sigma^2}{2}} = e^3$$

$$1+\frac{\sigma^2}{2} = 3$$

$$\sigma = 2$$

For the $CDF(e^4)$ we have:

$$\Phi\left(\frac{\ln e^4 - 1}{2}\right) = \Phi\left(\frac{4-1}{2}\right) = \Phi(1.5) = 0.9332$$

We can use the expensive process and the less expensive one will not work.

D.

$$CDF(w) = P(W \leq w) = P(Y \geq -w) = 1 - CDF(-w) = 1 - \Phi\left(\frac{\ln(-w) - \mu}{\sigma}\right)$$

E.

$$PDF(w) = \frac{d\ CDF(w)}{dw} = -\frac{1}{w\sigma\sqrt{2\pi}} \cdot e^{-\frac{(\ln(-w)-\mu)^2}{2\sigma^2}}$$