2.1. Problem Statement: STATISTICS 1

You survey households in your area to find the average rent they are paying. Find

the standard deviation from the following data:

$1550, $1700, $900, $850, $1000, $950.

Variance: $s^2 = \frac{\sum(\overline{X} - X_i)^2}{N}$ Standard Deviation: $s = \sqrt{\frac{\sum(\overline{X} - X_i)^2}{N}}$

Step 1: Find the mean.

N=no of samples:6

(x1+x2+x3+x4+x5+x6) / n

(1550+1700+900+850+1000+950)/6

Mean=6950/6 =**1158.33**

Step 2: For each data point, find the square of its distance to the mean.

1550-1158.33 = 392^2 =153664

1700-1158.33 = 542^2=293764

900-1158.33 = - 258^2=66564

850-1158.33 = - 308^2=94864

1000-1158.33 =- 158^2=24964

950-1158.33 =- 208^2 =43264

Step 3: Sum the values from Step 2.

153664 + 293764 + 66564 + 94864 + 24964 + 43264 = 677084

Step 4: Divide by the number of data points.

Variance : 677084 / 6 - 1 = **135416.8**

Step 5: Take the square root

Standard Deviation : √ 135416.8 = **367.99**

Find the variance for the following set of data representing trees in California (heights in feet):

3, 21, 98, 203, 17, 9

Variance:

$$s^2 = \frac{\sum(\overline{X} - X_i)^2}{N}$$

Step 1: Find the mean.

N=no of samples:6

(x1+x2+x3+x4+x5+x6) / n

(3 + 21 + 98 + 203 + 17 + 9)/6

Mean=351/6 =**58.5**

Step 2: For each data point, find the square of its distance to the mean.

3 - 58.5 = -56^2 = 3136

21 - 58.5 = - 38^2 = 1444

98 - 58.5 = - 40^2 = 1600

203 - 58.5 = 145^2 = 21025

17 - 58.5 = -42^2 = 1764

9 - 58.5 = -50^2 = 2500

Step 3: Sum the values from Step 2.

3136 + 1444 + 1600 + 21025 + 1764 + 2500 = 31469

Step 4: Divide by the number of data points.

Variance : 31469 / 6 - 1 = **6293.8**

3. In a class on 100 students, 80 students passed in all subjects, 10 failed in one subject, 7 failed in two subjects and 3 failed in three subjects. Find the probability distribution of the variable for number of subjects a student from the given class has failed in.

Probability distribution

Probability of no failures in any subject : 80 / 100 **: 0.8**

Probability of failures in one subject : 10 / 100 **: 0.1**

Probability of failures in two subjects : 7 / 100 : **0.07**

Probability of failiures in three subjects : 3 / 100 :**0.03**

2.2. Problem Statement: STATISTICS 2

1. A test is conducted which is consisting of 20 MCQs (multiple choices questions)

with every MCQ having its four options out of which only one is correct.

Determine the probability that a person undertaking that test has answered

exactly 5 questions wrong.

$$P(X) = \frac{n!}{(n-X)!\,X!} \cdot (p)^X \cdot (q)^{n-X}$$

n = 20

Probability of the person giving the correct answer : 1 / 4

Probability of the person giving the wrong answer : 1 –(1 / 4) : 3 / 4

Propability of the person giving 5 questions wrong

P(5 out of 20 are wrong): 20 ! % ((20 – 5) ! * 5 !) = 15504 * (3 / 4) ^ 5 * (1 / 4) ^ 15

: 3.42E-6

: **0.00000342**

2. A die marked A to E is rolled 50 times. Find the probability of getting a

“D” exactly 5 times.

$$P(X) = \frac{n!}{(n-X)!\,X!} \cdot (p)^X \cdot (q)^{n-X}$$

n =50

Probability of getting ‘D’ : 1 / 5

Probability of not getting ‘D’ : 1- (1/5) = 4 / 5

P(Getting ‘D’ exactly 5 times) : 50 ! % (45 ! * 5 !) * (1 / 5) ^ 5 * (4 / 5) ^ 45

: **0.029**

3. Two balls are drawn at random in succession without replacement from an urn

containing 4 red balls and 6 black balls.

Find the probabilities of all the possible outcomes.

Number of Red Balls : 4

Number of Black Balls :6

Total Number of Balls: 10

Number of Balls Drawn Randomly : 2

Possible Events:

P(Red , Red) = (4/10) * (3/9) = 0.13

P(Red , Black) = (4/10) * (6/9) = 0.26

P(Black , Red) = (6/10) * (4/9) = 0.26

P(Black , Black) =(6/10) * (5/9) = 0.33

Probability of no Red Ball : **0.33**

Probability of 1 red Ball or Black Ball : 0.26 + 0.26 =**0.52**

Probability of 2 red Balls **: 0.13**

Probability of no Black Ball **: 0.13**

Probability of 2 Black Balls **: 0.33**

2.3. Problem Statement: STATISTICS 3

Blood glucose levels for obese patients have a mean of 100 with a standard deviation

of 15. A researcher thinks that a diet high in raw cornstarch will have a positive

effect on blood glucose levels. A sample of 36 patients who have tried the raw

cornstarch diet have a mean glucose level of 108. Test the hypothesis that the raw

cornstarch had an effect or not.

Step 1:

As per the hypothesis ,population mean is 100

H0 (in null hypotheses) : μ : 100

H1 (in Alternative hypothesis) : μ > 100

Step 2:

Lets assume the significance level as 5% (ie)

the random chance probability is computed by

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean
$\sigma =$ Standard Deviation

For this set of data: z= (108-100) / (15/√36)=3.20

By looking at z- table , p-value associated with 3.20 is 0.9993 i.e. probability of having value less than 108 is 0.9993 and more than or equals to 108 is (1-0.9993)=0.0007.

Step-3: It is less than 0.05 so we will reject the Null hypothesis i.e. there is raw cornstarch effect

1. In one state, 52% of the voters are Republicans, and 48% are Democrats. In a second state, 47% of the voters are Republicans, and 53% are Democrats. Suppose a simple random sample of 100 voters are surveyed from each state. What is the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state?

let P_1 = the proportion of Republican voters in the first state, P_2 = the proportion of Republican voters in the second state, p_1 = the proportion of Republican voters in the sample from the first state, and p_2 = the proportion of Republican voters in the sample from the second state. The number of voters sampled from the first state (n_1) = 100, and the number of voters sampled from the second state (n_2) = 100.

- We make sure the samples from each population are big enough to model differences with a normal distribution. Because $n_1P_1 = 100 * 0.52 = 52$, $n_1(1 - P_1) = 100 * 0.48 = 48$, $n_2P_2 = 100 * 0.47 = 47$, and $n_2(1 - P_2) = 100 * 0.53 = 53$

- Find the mean of the difference in sample proportions: $E(p_1 - p_2) = P_1 - P_2 = 0.52 - 0.47 = 0.05$.

- Find the standard deviation of the difference.

$$\sigma_d = \text{sqrt}\{ [P_1(1 - P_1) / n_1] + [P_2(1 - P_2) / n_2] \}$$
$$\sigma_d = \text{sqrt}\{ [(0.52)(0.48) / 100] + [(0.47)(0.53) / 100] \}$$
$$\sigma_d = \text{sqrt } (0.002496 + 0.002491) = \text{sqrt}(0.004987) = 0.0706$$

- find the probability that p_1 is less than p_2. This is equivalent to finding the probability that $p_1 - p_2$ is less than zero. transform the random variable $(p_1 - p_2)$ into a z value score. The transformation is

$$z_{p1 - p2} = (x - \mu_{p1 - p2}) / \sigma_d = = (0 - 0.05)/0.0706 = -0.7082$$

Therefore, the probability that the survey will show a greater percentage of Republican voters in the second state than in the first state is 0.24.

2. You take the SAT and score 1100. The mean score for the SAT is 1026 and the

standard deviation is 209. How well did you score on the test compared to the

average test taker?

z-score equation

$$Z = \frac{1100 - \mu}{\sigma}$$

$$Z = \frac{1100 - 1026}{\sigma}$$

$$Z = \frac{1100 - 1026}{209}$$

(1100 – 1026) / 209 = .354.

P (x < 1100) = P (z < 0.354) = 0.6368 = 63.68%

The score is .6368 or 63.68%. have scored below ,

and (1 - 0.6368) = 0.3632 or 36.32% have scored above

2.4. Problem Statement: STATISTICS 4

1. Is gender independent of education level? A random sample of 395 people were

surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

High-School Bachelor Masters Ph.D. Total

Female 60 54 46 41 201

Male 40 44 53 57 194

Total 100 98 99 98 395

Question: Are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

This test is performed by using a Chi-square test of independence.

Two categorical variables within a two-way table,also called a

r × c contingency table,

where r = number of rows, c = number of columns.

Null Hypothesis: The two categorical variables are independent.

Alternative Hypothesis: The two categorical variables are dependent.

The chi-square test statistic is calculated by using the formula:

$\chi 2 = \sum(O - E) \text{ ^ } 2 / E$

where O represents the observed frequency.

E is the expected frequency under the null hypothesis and computed by:

E = (row total * column total) / sample size

Comparing the value of the test statistic to the critical value of $\chi 2\alpha$ with degre e of freedom = (r - 1) (c - 1),and reject the null hypothesis if $\chi 2 > \chi 2\alpha$

	High-School	Bachelor	Masters	Ph.D.	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

Table of expected counts

	High-School	Bachelor	Masters	Ph.D.	Total
Female	50.866	49.868	50.377	49.868	201
Male	49.114	48.132	48.623	48.132	194
Total	100	98	99	98	395

$\chi 2$= (60 - 50.866)^ 2 / 50.866 + (54 – 49.868)^ 2 / 49.868 + (46 – 50.377)^2 / 50.377 +

(41 – 49.868)^2 / 49.868 + (40 -49.114)^2 / 49.114 + (44 – 48.132) ^2 / 48.132

+ (53 – 48.623)^2 / 48.623 + (57 – 48.132)^2 / 48.132 = 8.006

The critical value of $\chi 2$ with 1 degree of freedom is 3.84.

Since 1 < 3.84, we can not reject the null hypothesis. We can conclude that education level depends on gender at 5% level of significance

2. Using the following data, perform a one-way analysis of variance using α=.05.

Write up the results in APA format.

[Group1: 51, 45, 33, 45, 67]

[Group2: 23, 43, 23, 43, 45]

[Group3: 56, 76, 74, 87, 56]

Sample means for the groups: = 48.2, 35.4, 69.8

Intermediate steps in calculating the group variances:

[[1]]

	value	mean	deviations	sq deviations
1	51	48.2	2.8	7.84
2	45	48.2	-3.2	10.24
3	33	48.2	-15.2	231.04
4	45	48.2	-3.2	10.24
5	67	48.2	18.8	353.44

[[2]]

	value	mean	deviations	sq deviations
1	23	35.4	-12.4	153.76
2	43	35.4	7.6	57.76
3	23	35.4	-12.4	153.76
4	43	35.4	7.6	57.76
5	45	35.4	9.6	92.16

[[3]]

value mean deviations sq deviations

1 56 69.8 -13.8 190.44

2 76 69.8 6.2 38.44

3 74 69.8 4.2 17.64

4 87 69.8 17.2 295.84

5 56 69.8 -13.8 190.44

Sum of squared deviations from the mean (SS) for the groups:

[1] 612.8 515.2 732.8

Var1=612.85−1=153.2

Var2=515.25−1=128.8

Var3=732.85−1=183.2

MS error=153.2+128.8+183.23=155.07

Calculating the remaining *error* (or *within*) terms for the ANOVA table:

Df error=15−3=12

SS error=(155.07)(15−3)=1860.8

Intermediate steps in calculating the variance of the sample means:

Grand mean ($\bar{x}_{grand}$x¯grand) = 48.2+35.4+69.83=51.13

group mean grand mean deviations sq deviations

48.2 51.13 -2.93 8.58

35.4 51.13 -15.73 247.43

69.8 51.13 18.67 348.57

(SSmeans)=604.58

Varmeans=604.583−1=302.29

MSbetween=(302.29)(5)=1511.45

Calculating the remaining *between* (or *group*) terms of the ANOVA table:

$df_{groups} = 3 - 1 = 2$

$SS_{group} = (1511.45)(3 - 1) = 3022.9$

Test statistic and critical value

$F = \frac{1511.45}{155.07} = 9.75$

$F_{critical}(2,12) = 3.89$

Decision: reject H0

ANOVA table

source	SS	df	MS
group	3022.9	2	1511.45
error	1860.8	12	155.07
total	4883.7		

Effect size

$\eta^2 = \frac{3022.9}{4883.7} = 0.62$

APA writeup

$F(2, 12) = 9.75$, $p < 0.05$, $\eta^2 = 0.62$.

3. Calculate F Test for given 10, 20, 30, 40, 50 and 5,10,15, 20, 25.

For 10, 20, 30, 40, 50:

F Test is generally defined as ratio of the variances of the given two set of values. First calculate standard deviation and variation of the given set of values. The formula used to calculate SD is,

Standard Deviation Formula

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

The standard deviation is represented by the symbol $^{\delta}$ and variance is square of the standard deviation.

The formula used to calculate F Test is,

Calculating Variance of first set

Total Inputs (N) =(10,20,30,40,50)

Total Inputs (N)=5

Mean (xm)= (x1+x1+x2...xn)/N

Mean (xm)= 150/5

Means(xm)= 30

SD=sqrt(1/(N-1)*((x1-xm)2+(x2-xm)2+..+(xn-xm)2))

=sqrt(1/(5-1)((10-30)2+(20-30)2+(30-30)2+(40-30)2+(50-30)2))

=sqrt(1/4((-20)2+(-10)2+(0)2+(10)2+(20)2))

=sqrt(1/4((400)+(100)+(0)+(100)+(400)))

=sqrt(250)

=15.8114

Variance=SD^2

Variance=15.81142

Variance=25

Calculating Variance of second set

For 5, 10,15,20,25:
Total Inputs(N) =(5,10,15,20,25)
Total Inputs(N)=5
Mean (xm)= (x1+x2+x3...xN)/N
Mean (xm)= 75/5
Means (xm)= 15
$SD=sqrt(1/(N-1)*((x_1-x_m)^2+(x_2-x_m)^2+..+(x_n-x_m)^2))$
$=sqrt(1/(5-1)((5-15)^2+(10-15)^2+(15-15)^2+(20-15)^2+(25-15)^2))$
$=sqrt(1/4((-10)^2+(-5)^2+(0)^2+(5)^2+(10)^2))$
=sqrt(1/4((100)+(25)+(0)+(25)+(100)))
=sqrt(62.5)
=7.9057
$Variance=SD^2$
$Variance=7.9057^2$
Variance=62.5

<u>calculating F Test</u>
F Test = (variance of 10, 20,30,40,50) / (variance of 5, 10, 15, 20, 25)
= 250/62.5
= 4.

The F Test value is 4.