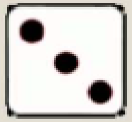
DATA SCIENCE FOR ENGINEERS

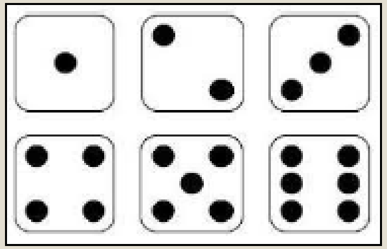
WEEK 3

Random Phenomenon

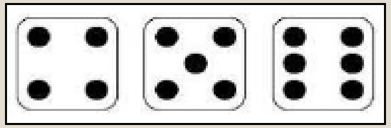
- The phenomenon/experiment whose outcomes are not predictable with certainty are called **random phenomenon.**



- **Sample space:** The set of all possible outcome of an experiment.
  Ex → S={1,2,3,4,5,6}



- **Event:** Any subset of sample space
  Ex → outcome is greater than 3. E={4,5,6}



Probability

- Probability is a measure that assign real value to every outcome of a random phenomenon
- The probability is ratio of number of ways an event can happen to the number of ways sample space is present

$$P(E) = \frac{n(E)}{n(S)}$$

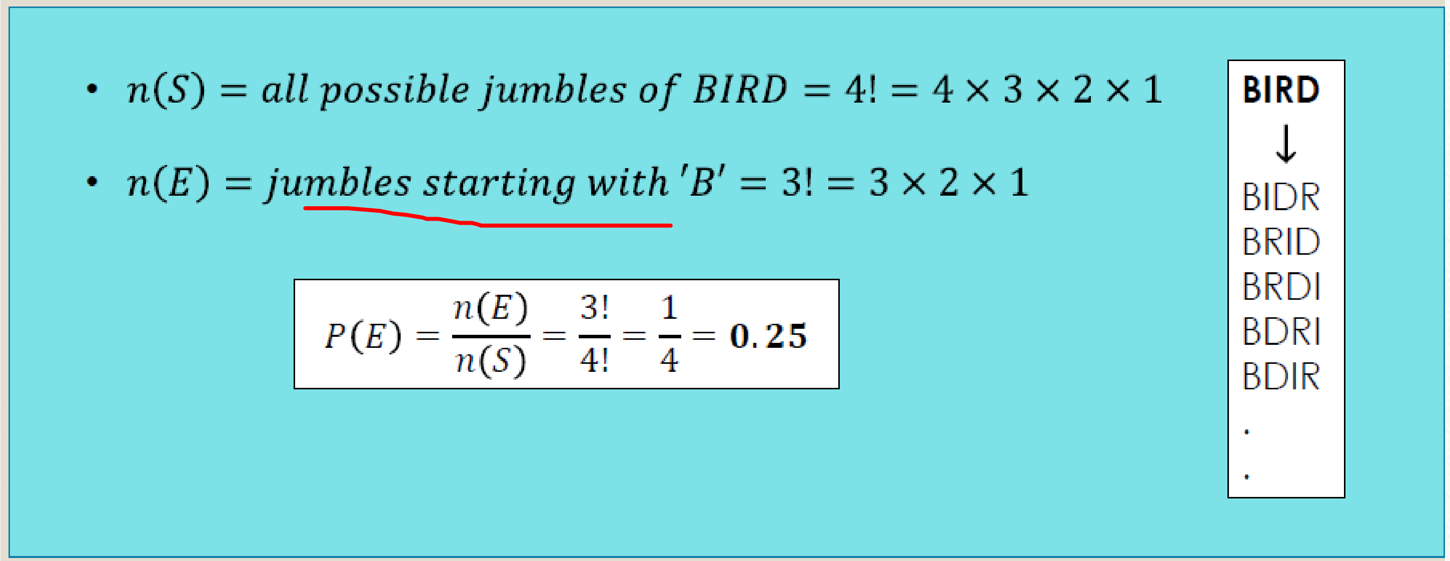
$E = \{1, 3, 5\}$ $n(E) = 3$

$S = \{1, 2, 3, 4, 5, 6\}$ $n(S) = 6$ $P(E) = \frac{3}{6} = \frac{1}{2}$

- Axioms of probability

- $0 \leq P(E) \leq 1$ (Probability is non-negative and less than one)
- $P(S) = 1$ (Probability of entire sample space in1)
- $P(A \cup B) = P(A) + P(B)$ (For two mutually exclusive events)

Q. If out of all possible jumbles of the ‘BIRD’, a random word is picked, what is the probability, that this will start with a ‘B’.



- $n(S) = all\ possible\ jumbles\ of\ BIRD = 4! = 4 \times 3 \times 2 \times 1$
- $n(E) = \underline{jumbles\ starting\ with}\ 'B' = 3! = 3 \times 2 \times 1$

$$\boxed{P(E) = \frac{n(E)}{n(S)} = \frac{3!}{4!} = \frac{1}{4} = \mathbf{0.25}}$$

BIRD
↓
BIDR
BRID
BRDI
BDRI
BDIR
.
.

BIRD

$n(S) = 4!$

B _ _ _ $\quad n(E) = 3!$

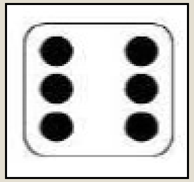
$$= \frac{3 \times 2}{4 \times 3 \times 2} = \frac{1}{4}$$

Events

- **Mutually Exclusive Event:** The occurrence of one event implies that other event does not occur

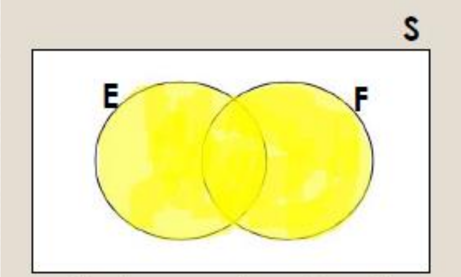
Ex→ Coin toss gives either Head or Tail not both

- **Independent Event:** two events are independent if occurrence of one has no influence on other

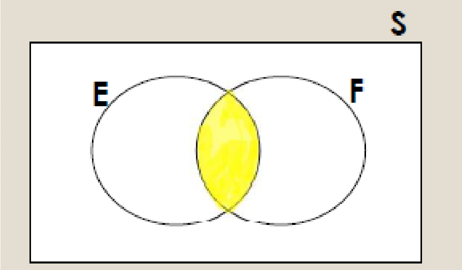


Ex→ Landing on heads after tossing a coin AND rolling a 5 on a single 6-sided die

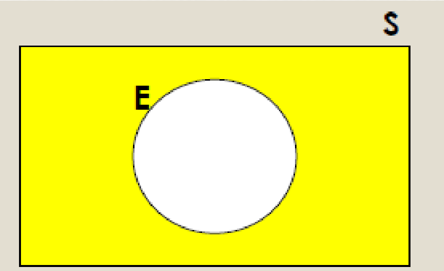
Venn Diagrams



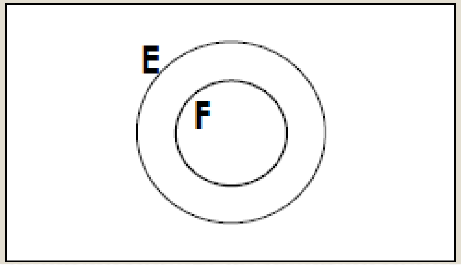
Colour region: $E \cup F$



Colour region: $E \cap F$



Colour region: E^c



Conditional Probability

- If two event A & B are not independent, then information available about the outcome of event A can influence the predictability of event B

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

$P(A) = \frac{2}{4}$

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

- Example → Two fair coins are toss

Event A: First toss is Head = {HT,HH}

Event B: Two successive head = {HH}

P(A) = n(A)/n(S) = 2/4 = 0.5

P(B) = n(B)/n(S) = 1/4 = 0.25

HH
HT
TH
TT

If the event A is given then probability of event B is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{0.25}{0.5} = \mathbf{0.5}$$

Random Variable

- ➢ A random variable (RV) is a map from sample space to a real line such that there is a unique real number corresponding to every outcome of sample space.
- ➢ Ex → coin toss sample space [H T] map to [0 1]

➢ Discrete random variable:

A variable that take one value from a discrete set of values

Ex → dice rolling [1 2 3 4 5 6]

➢ Continuous random variable:

The variable that can take continuous range of values

Ex → Temperature of the week [28.3, 21.0, 25.9, 26.1, 32.6, 30.0, 29.8]

Probability Mass/Density Function

➢ Probability mass/density function help to assign the probability to every outcome of a sample space

Probability Mass Function (PMF)

- PMF use for discrete random variable
- $\boldsymbol{p(x) = P[X = x]}$
- Ex → In coin toss

$P[X = 0] = 0.5, P[X = 1] = 0.5$

Probability Density Function (PDF)

- PDF use for continuous random variable
- $\boldsymbol{p(a < x < b) = \int_a^b F(x)\, dx}$

Combination

total = n

r items

$$^{n}C_r = \frac{n!}{(n-r)!\,r!}$$

Q. The box contain 20 defective items and 80 non-defective items. If two items are selected at random without replacement, what will be the probability that both items are defective ?

$$= \frac{20!}{(20-2)! \times 2!}$$

$$\cancel{^{100}C_2}$$

$$= \frac{20 \times 19 \times \cancel{18!}}{\cancel{18!} \times 2!}$$

$$= \frac{20 \times 19}{2}$$

- P(both items are defective)= ?

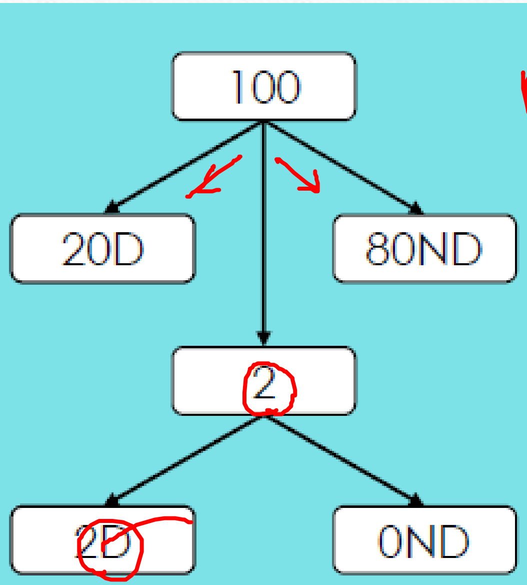
$$P = \frac{20C_2 \times 80C_0}{100C_2}$$

$$= \frac{190 \times 1}{4950} = \frac{19}{495}$$

$$P = \frac{^{20}C_2 \times {}^{80}C_0}{^{100}C_2}$$

$$\frac{20}{100} \times \frac{19}{99}$$

$$=$$



100

20D

80ND

2

2D

0ND

$$\frac{\dfrac{20 \times 19 \times \cancel{18!}}{(\cancel{20-2})! \times 2!} \times \dfrac{80!}{(80-0)! \times 0!}}{\dfrac{100 \times 99 \times \cancel{98!}}{(\cancel{100-2})! \times 2!}}$$

3 R   8 B

11 balls

Distributions

Binomial

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

where

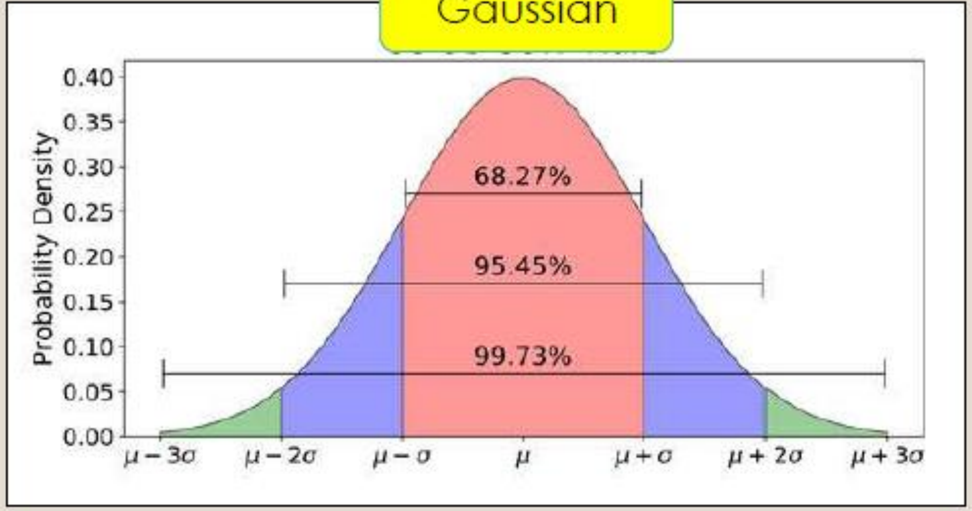
n = the number of trials (or the number being sampled)

x = the number of successes desired

p = probability of getting a success in one trial

$q = 1 - p$ = the probability of getting a failure in one trial

Gaussian

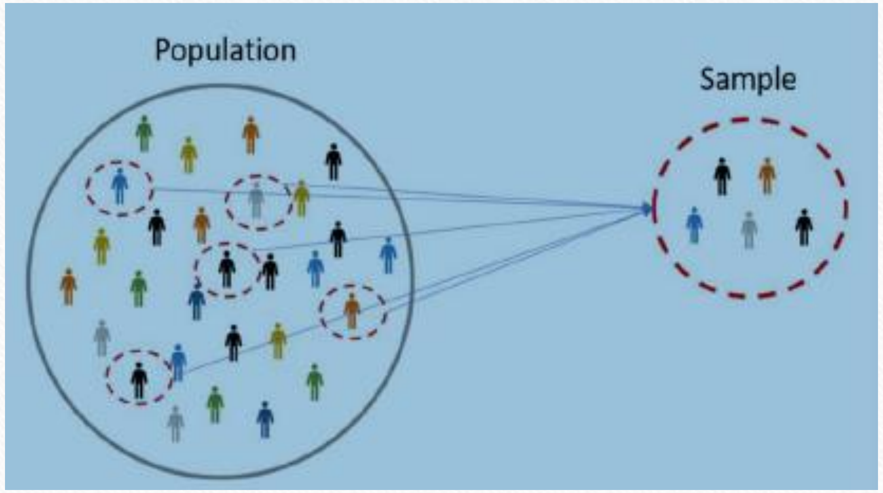


$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

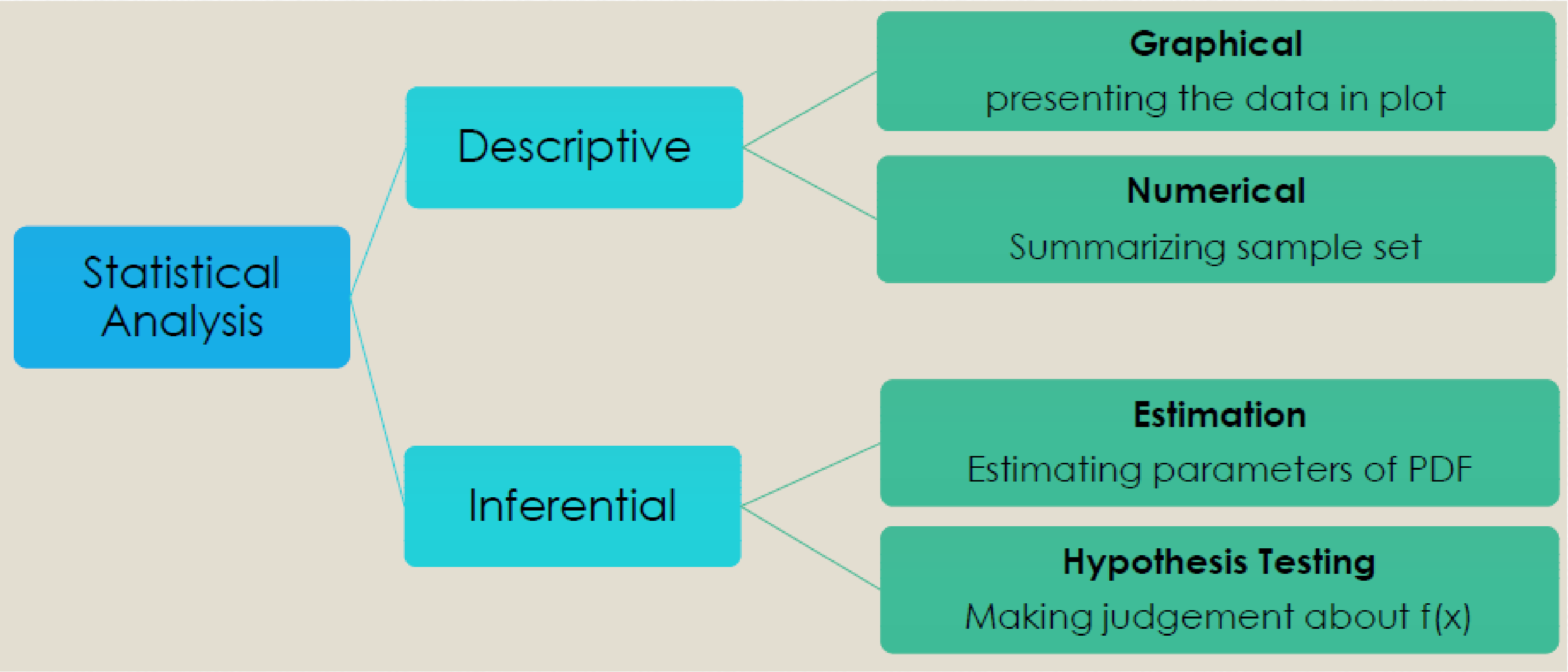
where σ is the standard deviation and μ the mean

Statistical sampling

- Population: Set of all possible outcomes of a random experiment
- Sample set: Finite set of observation obtained from experiment
- Sampling help to make inferences about the population
- The inference may be uncertain because samples might be uncertain



Statistical Analysis



Statistical Analysis

Descriptive

Graphical

presenting the data in plot

Numerical

Summarizing sample set

Inferential

Estimation

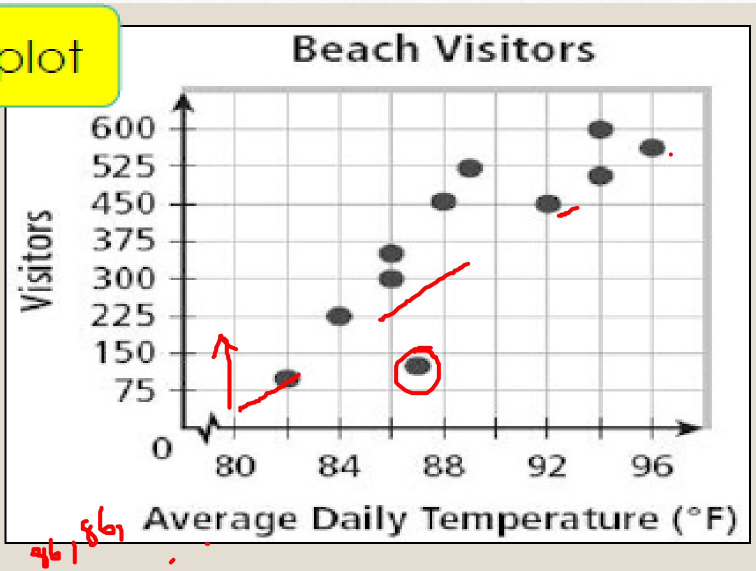
Estimating parameters of PDF

Hypothesis Testing

Making judgement about f(x)

Graphical Statistics

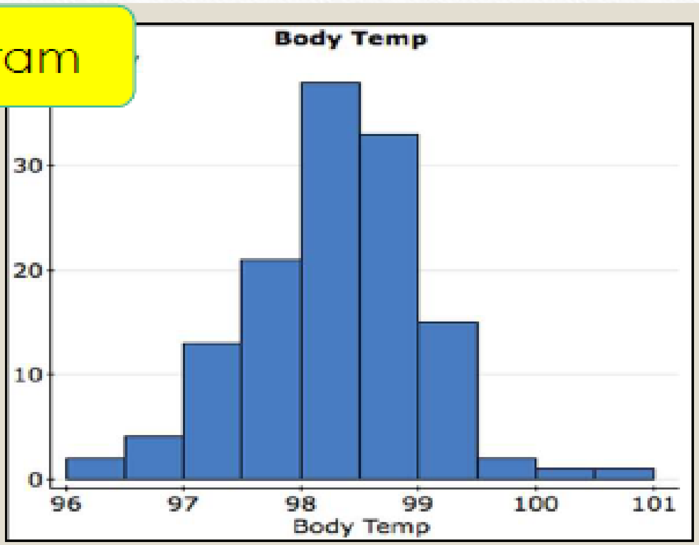
Scatter plot



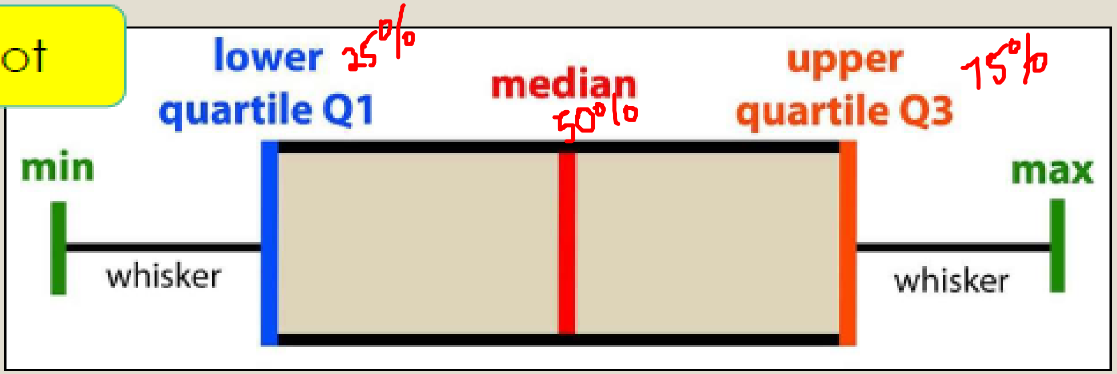
temp / x = 80 82, 84, 86, 86,

Visitors / y = 80, 225, 300, ...

Histogram



Box plot



Numerical Statistics

Mean: Mean is average or norm ➔ $\frac{1+3+4+6+6+7+8}{7} = 5$

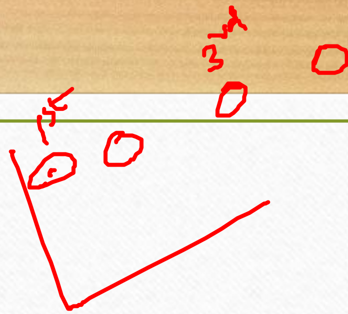
Median: Median is middle value ➔ 1 3 4 **6** 6 7 8

Mode: Mode is most frequent value ➔ 1 3 4 6 6 7 8

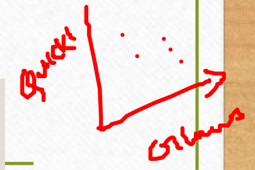
Range: Difference between lowest and highest value ➔ 8 -1= 7

Goals scored in seven matched

1 3 4 6 6 7 8



QUICKI = inverse



Covariance and Correlation

Covariance

- Covariance indicates the direction of the linear relationship between variables
- Covariance values are not standardized.
- Value can be anything

$$\operatorname{cov}(X, Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}.$$

Correlation

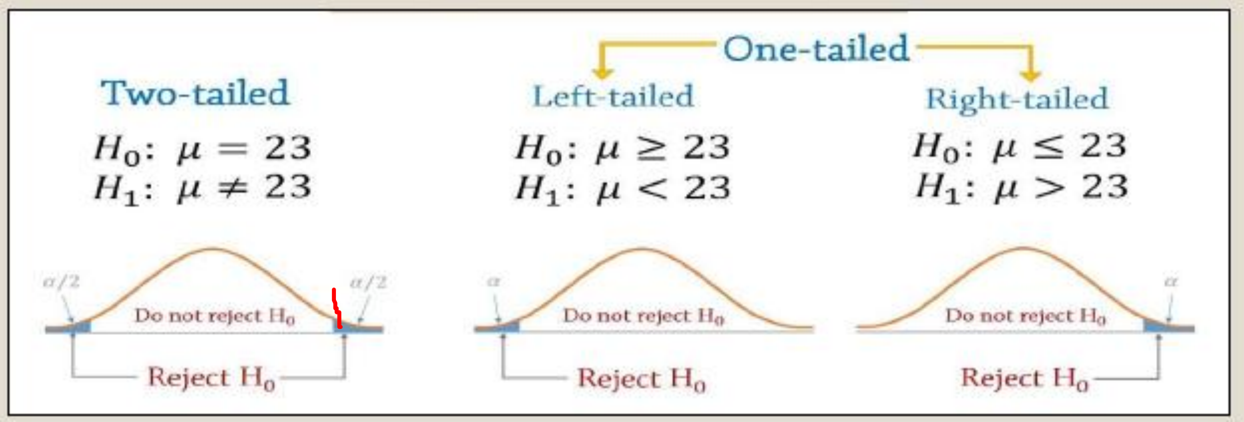
- Correlation measures both the strength and direction of the linear relationship between two variables
- Correlation values are standardized
- Value lie between -1 and +1

$$Correlation = \frac{Cov\,(x, y)}{\sigma x * \sigma y}$$

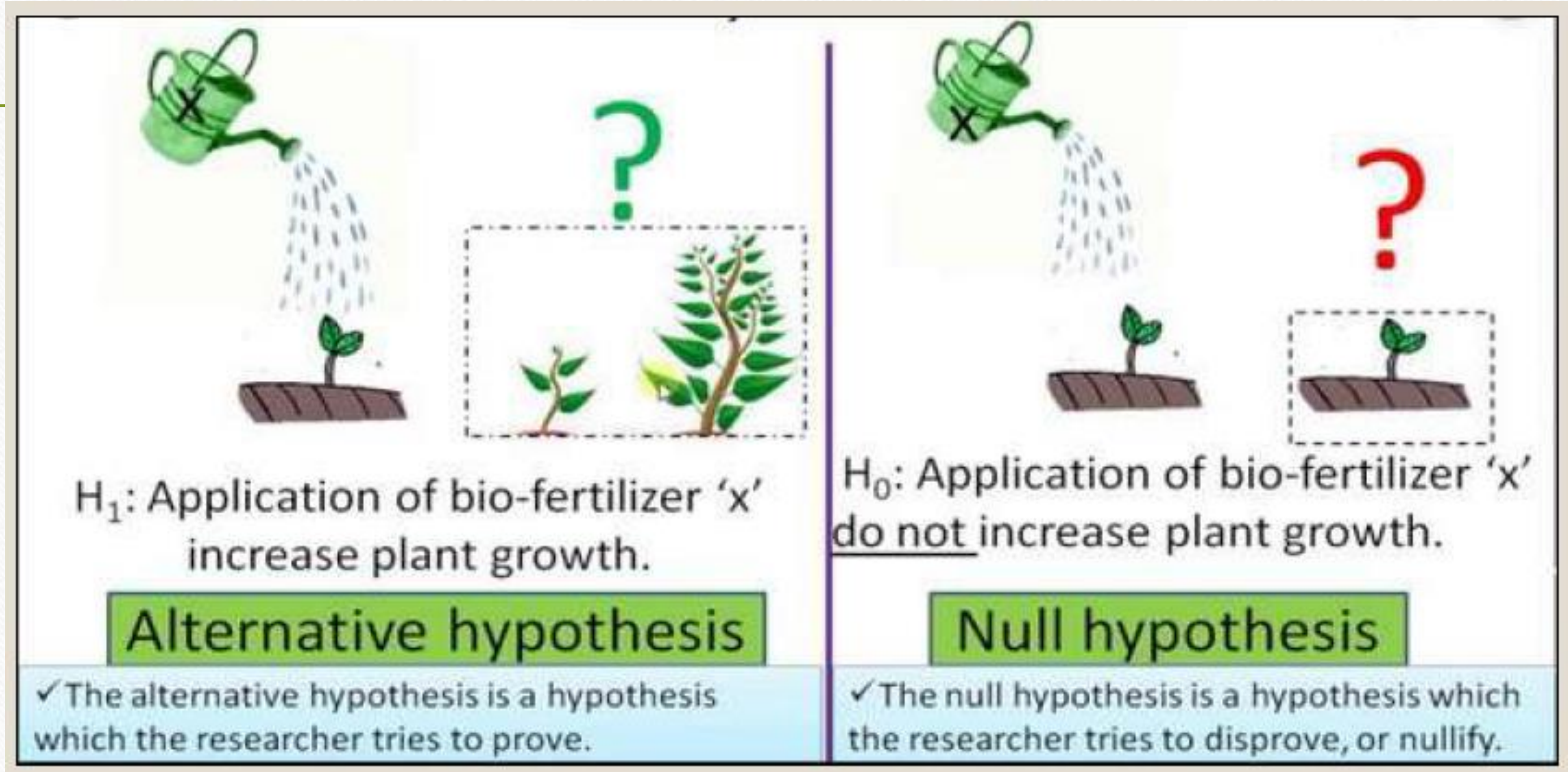
-0.05

Hypothesis Testing

- Hypothesis testing is used when we want to make decisions
- Ex: Whether the effect of drug A similar to drug B
- Hypothesis testing is generally converted to a test of mean and variance parameter of population



Null and Alternate Hypothesis



H_1: Application of bio-fertilizer 'x' increase plant growth.

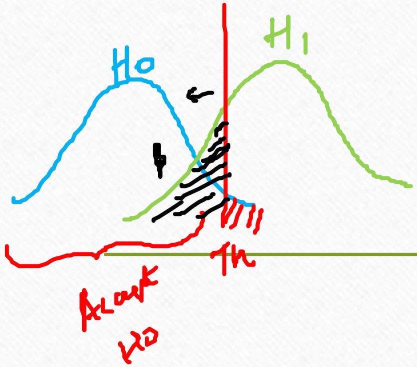
Alternative hypothesis

✓ The alternative hypothesis is a hypothesis which the researcher tries to prove.

H_0: Application of bio-fertilizer 'x' do not increase plant growth.

Null hypothesis

✓ The null hypothesis is a hypothesis which the researcher tries to disprove, or nullify.

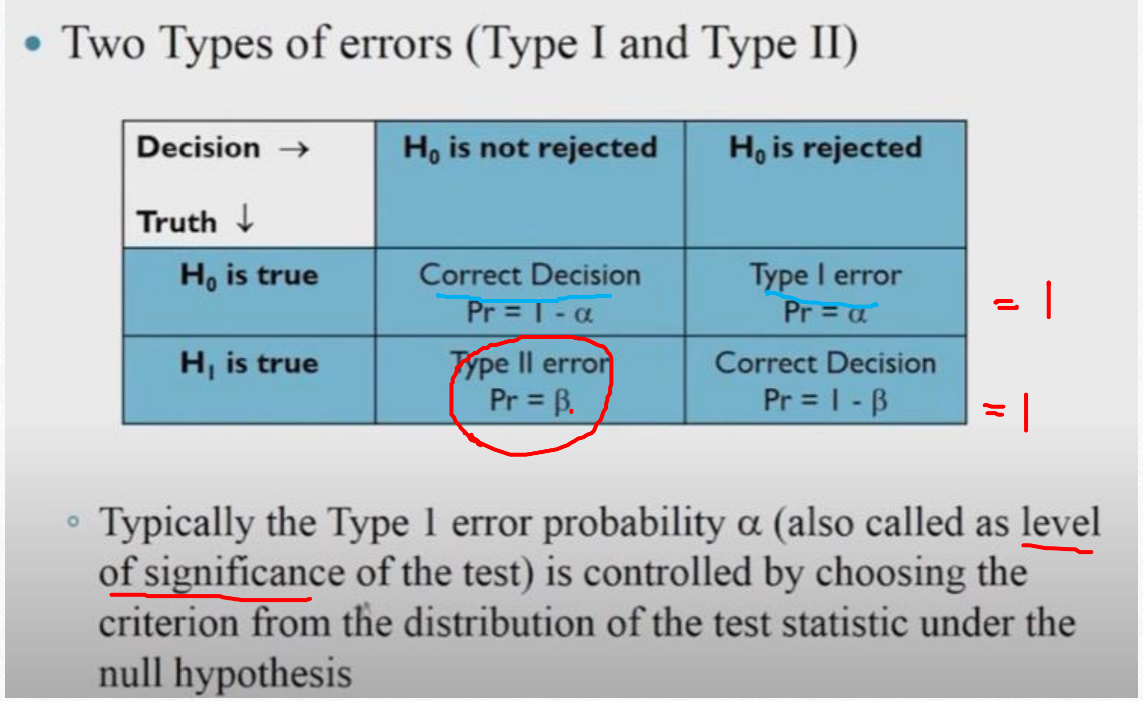


Types of Errors

Type I error

Ho as true

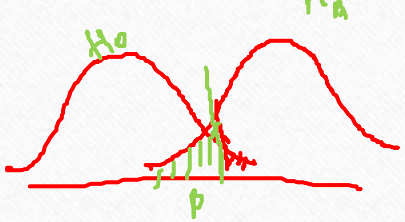
Decision → Ho rejected



- Two Types of errors (Type I and Type II)

Decision → Truth ↓	H_0 is not rejected	H_0 is rejected	
H_0 is true	Correct Decision Pr = 1 - α	Type I error Pr = α	= 1
H_1 is true	Type II error Pr = β.	Correct Decision Pr = 1 - β	= 1

  - Typically the Type 1 error probability α (also called as level of significance of the test) is controlled by choosing the criterion from the distribution of the test statistic under the null hypothesis



1) Sumit wants to contact one of his friends, but he remembers only the first 9 of the 10 digits of the contact number. He is sure that the last digit of the contact number is an odd number. He selects an odd number randomly. If the random variable X denotes the last digit of the contact number, then calculate Var(X).

○ 5

[✓ 8]

○ 33

○ None of the above

$\_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \overset{\text{odd}}{\_}$

Total = 10

odd = 5

$RV = \{1, 3, 5, 7, 9\}$

$\bar{X}$ or $\mu = 5$

$x - \mu = \{-4, -2, 0, 2, 4\}$

$(x-\mu)^2 = \{16, 4, 0, 4, 16\}$

$E[(x-\mu)^2] = \frac{40}{5} = 8$

$\bar{X}$

$x - \bar{x}$

$(x - \bar{x})^2$

$Var(x) = E[(x-\bar{x})^2$

$X \sim N(\text{mean, variance})$

2) Suppose X~Normal(μ,4). For n=20 iid samples of X, the observed sample mean is 5.2. What conclusion would a z-test reach if the null hypothesis assumes μ=5 (against an alternative hypothesis μ≠5) at a significance level of α=0.05?

○ [✓ Accept H_0]

○ Reject H_0

$n = 20$

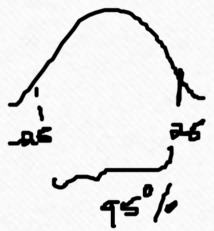
$\bar{x} = 5.2$

$\sigma^2 = 4$

$H_0 : \mu = 5$

$H_1 : \mu \neq 5$

$\alpha = 0.05$ 95% confidence interval



$$= \frac{5.2 - 5}{2/\sqrt{20}} = \frac{0.2}{2/2\sqrt{5}}$$

$$Z = \frac{\bar{x} - \mu}{5/\sqrt{n}}$$

= 0.~~447~~

0.4545

$X \sim \text{Norm}(100, (15)^2)$

Eg: Suppose the IQ in a certain population is normally distributed with a mean of $\mu = 100$ and standard deviation of $\sigma = 15$. A scientist wants to know if a new medication affects IQ levels, so she recruits $\underbrace{\text{20 patients}}_{\text{sample}}$ to use it for one month and records their IQ levels at the end of the month.

Write a code to show how to perform a one sample z-test in R to determine if the new medication causes a significant difference in IQ levels:

IQ levels: **88, 92, 94, 94, 96, 97, 97, 97, 99, 99, 105, 109, 109, 109, 110, 112, 112, 113, 114, 115**

$\mu = 100$

$\sigma = 15$

$n = 20$

$\bar{x} = 103.5$

sample mean

$H_0 : \mu = 100$

$H_1 : \mu \neq 100$

3) A box contains 8 items out of which 2 are defective. A sample of 5 items is to be selected randomly (without replacement) from the box. If the random variable X represents the number of defective items in a selection of 5 items, then find E(X). (Enter the answer correct to 2 decimal places)

○ ✓ 1.25

○ 5

Total = 8

Def = 2 = D

ND = 6

$X = \{0, 1, 2\}$

$P(X=0) = \frac{{}^6C_5 \times {}^2C_0}{{}^8C_5} = \frac{\frac{6\times \cancel{5!}}{(6-5)!\times \cancel{5!}}}{\frac{8\times 7\times 6\times \cancel{5!}}{(8-5)!\times \cancel{5!}}} = \frac{\frac{6}{1}}{\frac{8\times 7\times \cancel{6}}{\cancel{3\times 2}}} = \frac{6}{56}$

$\frac{n(E)}{Total}$

$P(X=1) = P(\text{one defective}) = \frac{{}^2C_1 \times {}^6C_4}{{}^8C_5} = \frac{30}{56}$

$P(X=2) = P(2\ \text{def}) = \frac{{}^2C_2 \times {}^6C_3}{{}^8C_5}$

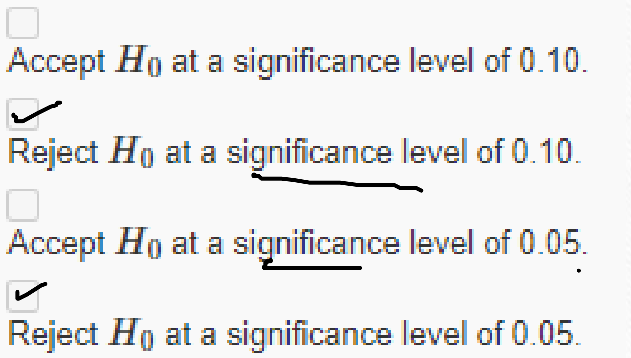
$= \frac{20}{56}$

X	0	1	2
$P(x)$	$\frac{6}{56}$	$\frac{30}{56}$	$\frac{20}{56}$

$E(X) = \sum x\,p(x) = 0\times\frac{6}{56} + 1\times\frac{30}{56} + 2\times\frac{20}{56}$

$= \frac{30+40}{56} = \frac{70}{56} = 1.25$

4) Suppose X~Normal(μ,9). For n=100 iid samples of X, the observed sample mean is 11.8. What conclusion would a z-test reach if the null hypothesis assumes μ=10.5 (against an alternative hypothesis μ≠10.5)?



□ Accept H_0 at a significance level of 0.10.

☑ Reject H_0 at a significance level of 0.10.

□ Accept H_0 at a significance level of 0.05.

☑ Reject H_0 at a significance level of 0.05.

$X \sim N(\mu, 9)$

$n = 100$

$\bar{x} = 11.8$

$H_0 : \mu = 10.5$

$H_1 : \mu \neq 10.5$

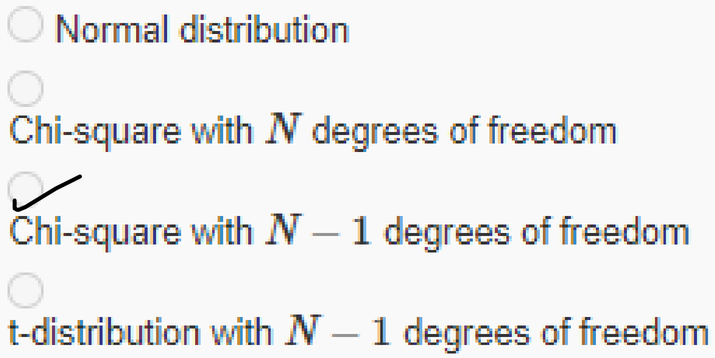
$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \underline{\underline{4.333}}$$

Critical value

1.645 for ~~0.10~~ 0.05

1.2816 0.10

5) A sample of N observations are independently drawn from a normal distribution. The sample variance follows



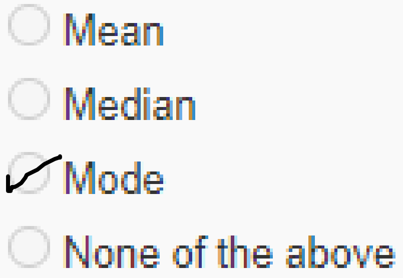
○ Normal distribution

○ Chi-square with N degrees of freedom

○ ✓ Chi-square with $N-1$ degrees of freedom

○ t-distribution with $N-1$ degrees of freedom

6) Which one of the following is best measure of central tendency for categorical data?



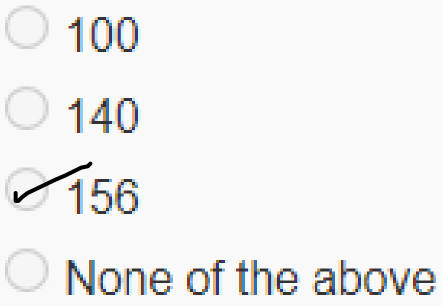
○ Mean

○ Median

○ [✓ Mode]

○ None of the above

7) Let X and Y be two independent random variables with Var(X) = 9 and Var(Y) =3,
 find Var(4X−2Y+6)



○ 100

○ 140

[✓ 156]

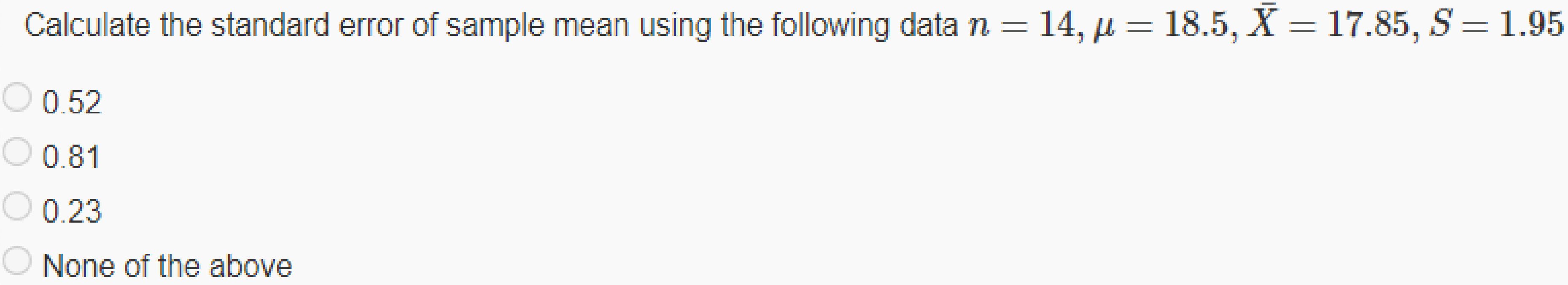
○ None of the above

$$\mathrm{Var}(aX + bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$$

$4^2 Var\, X + 2^2 Var(y)$.

$$Var(4X - 2Y + 6) = \underline{16}\, Var(X) + 4\, Var(Y) = \underline{\underline{156}}$$

8)



Calculate the standard error of sample mean using the following data $n = 14, \mu = 18.5, \bar{X} = 17.85, S = 1.95$

○ 0.52

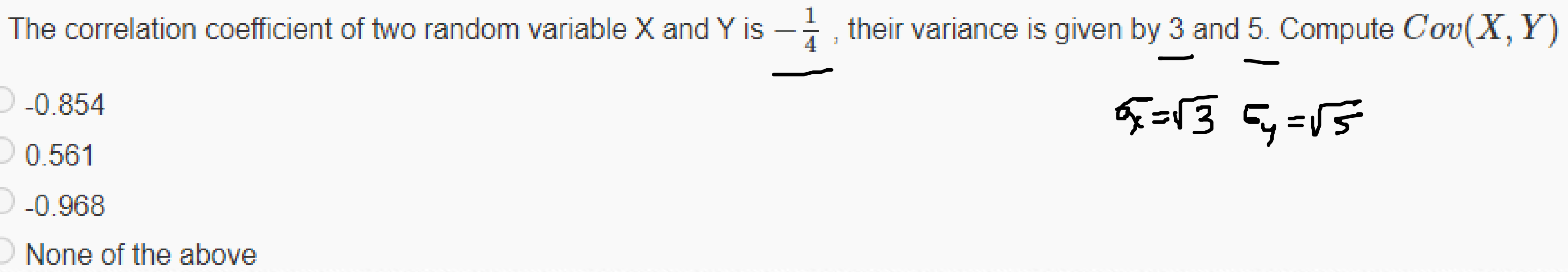
○ 0.81

○ 0.23

○ None of the above

$$SE = \frac{S}{\sqrt{n}} = \frac{1.95}{\sqrt{14}} = \underline{\underline{0.52}}$$

9)



The correlation coefficient of two random variable X and Y is $-\frac{1}{4}$, their variance is given by $\underline{3}$ and $\underline{5}$. Compute $Cov(X,Y)$

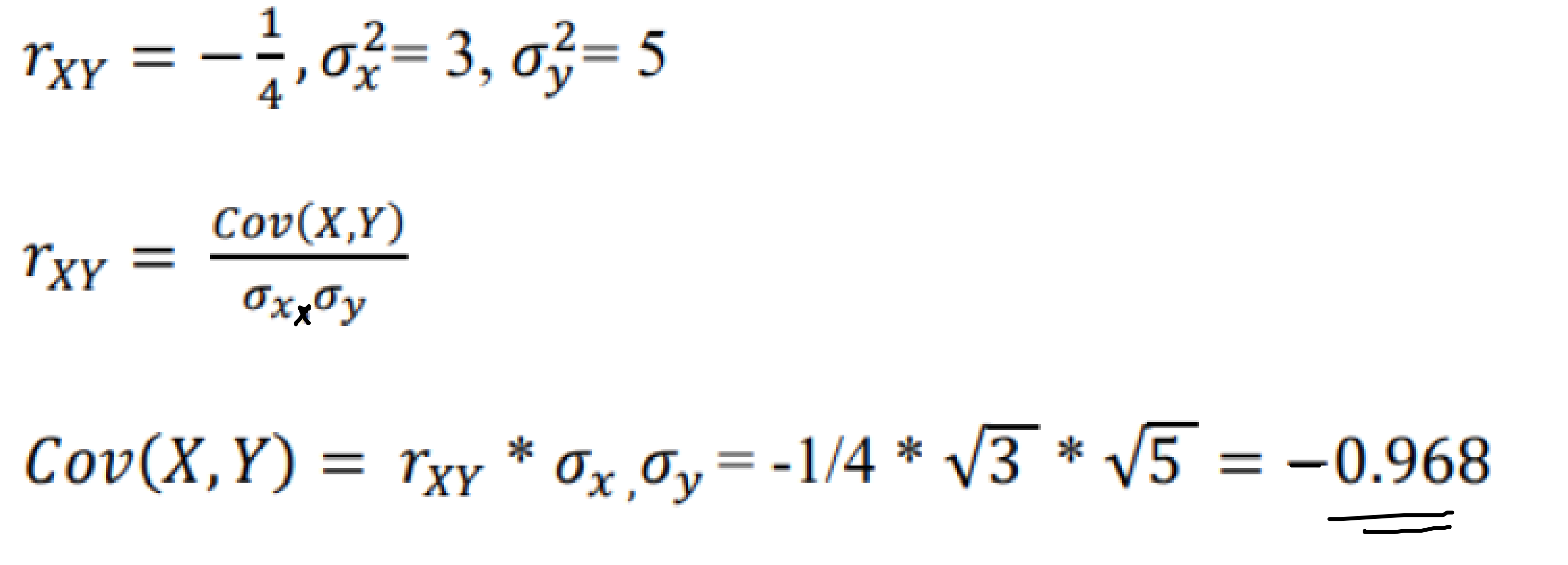
○ -0.854

○ 0.561

○ -0.968

○ None of the above

$\sigma_x = \sqrt{3}$ $\sigma_y = \sqrt{5}$



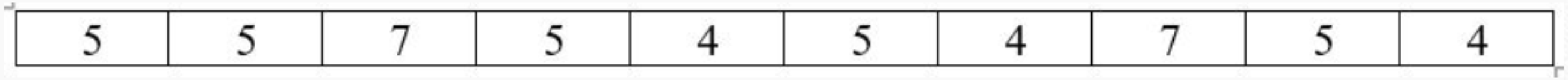
$$r_{XY} = -\frac{1}{4}, \sigma_x^2 = 3, \sigma_y^2 = 5$$

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_x \times \sigma_y}$$

$$Cov(X,Y) = r_{XY} * \sigma_x, \sigma_y = -1/4 * \sqrt{3} * \sqrt{5} = -\underline{\underline{0.968}}$$

10) Find the t-statistic for the sample data given that the population mean of the distribution is 8

Sample Data:



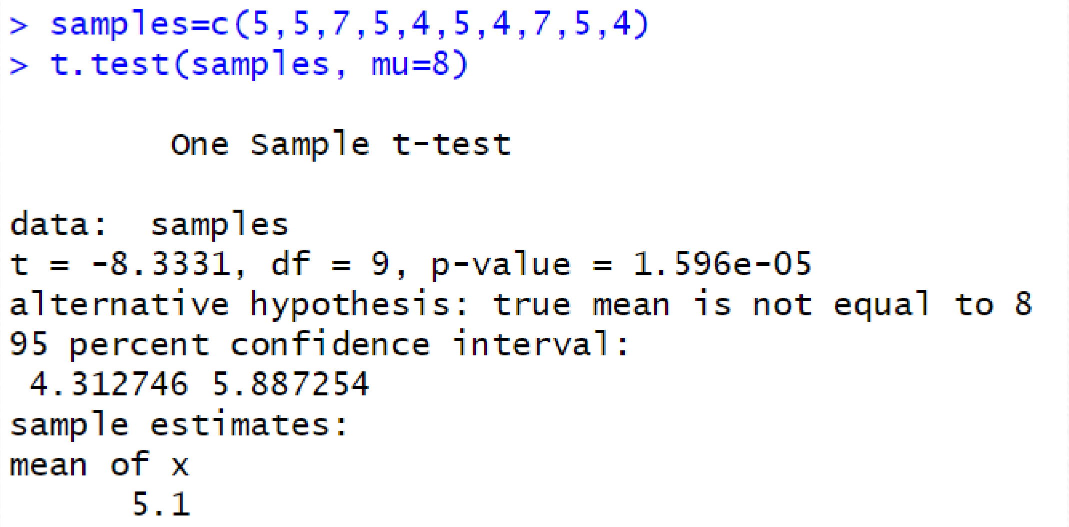
5	5	7	5	4	5	4	7	5	4

○ 3.93

○ 1.44

○ -8.33

○ None of the above



```
> samples=c(5,5,7,5,4,5,4,7,5,4)
> t.test(samples, mu=8)

        One Sample t-test

data:  samples
t = -8.3331, df = 9, p-value = 1.596e-05
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 4.312746 5.887254
sample estimates:
mean of x
      5.1
```

T-test

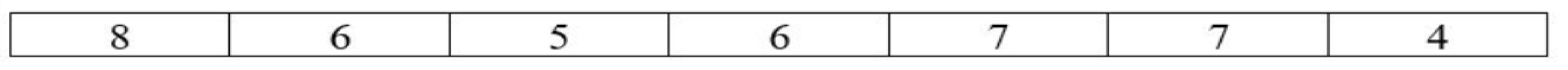
- Used when variance is not known
- Use for sample size less then 30
- Based on student-t distribution

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$s = sample\ std.dev$$

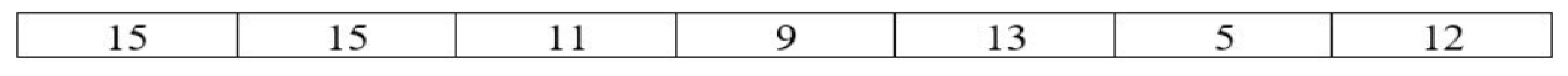
11) Find F-statistic for the following sample sets generated from the same distribution

Sample Data 1:

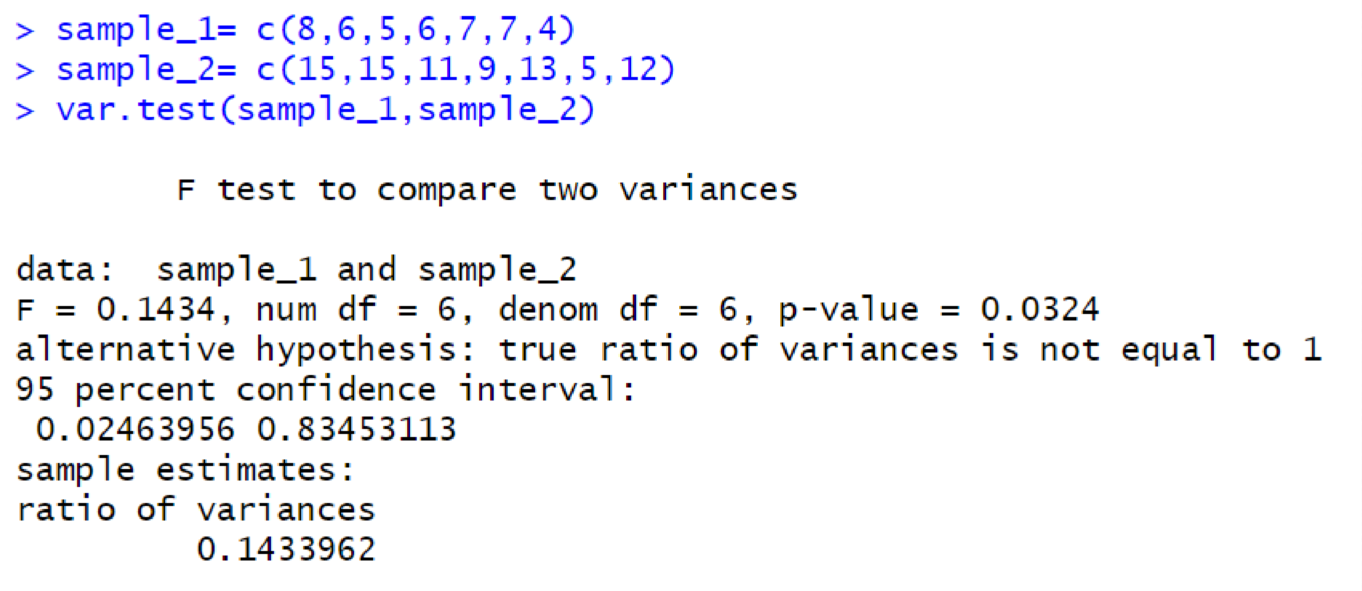


8	6	5	6	7	7	4

Sample Data 2:



15	15	11	9	13	5	12

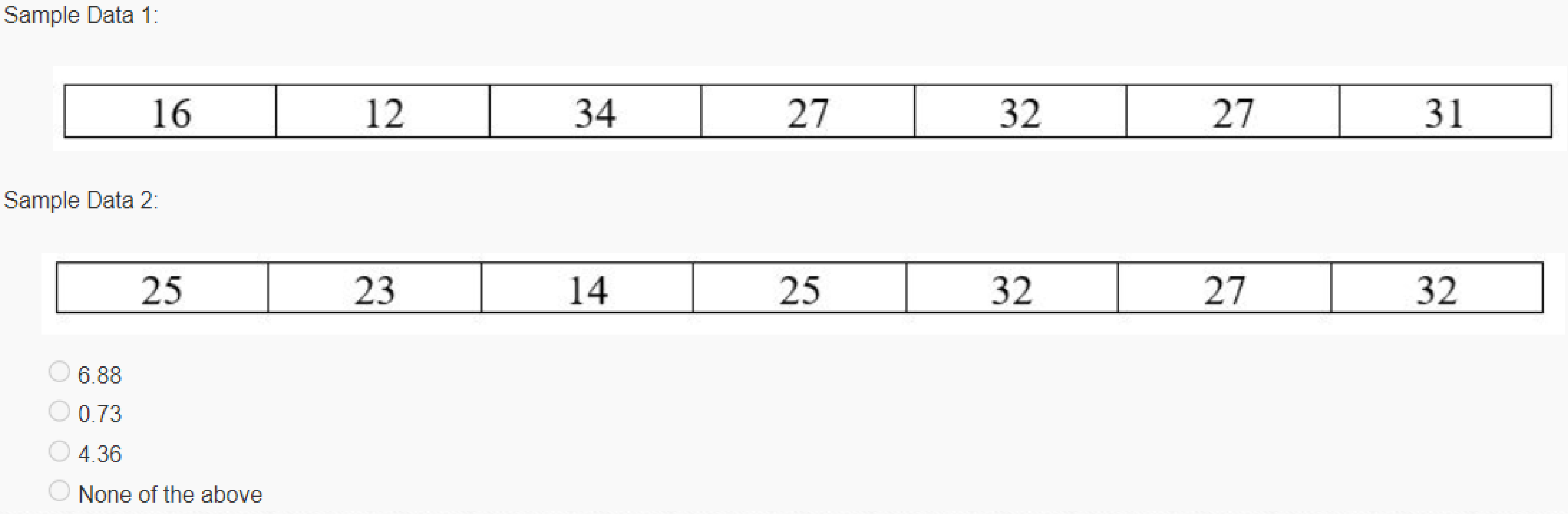


```
> sample_1= c(8,6,5,6,7,7,4)
> sample_2= c(15,15,11,9,13,5,12)
> var.test(sample_1,sample_2)

	F test to compare two variances

data:  sample_1 and sample_2
F = 0.1434, num df = 6, denom df = 6, p-value = 0.0324
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.02463956 0.83453113
sample estimates:
ratio of variances 
         0.1433962 
```

12) Find the F-statistic for the following sample sets generated from the distributions with population variances 3 and 7 respectively



Sample Data 1:

16	12	34	27	32	27	31

Sample Data 2:

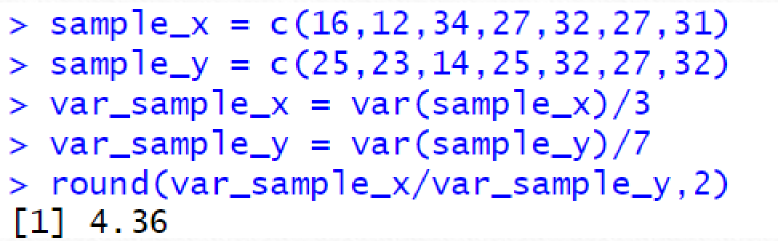
25	23	14	25	32	27	32

○ 6.88

○ 0.73

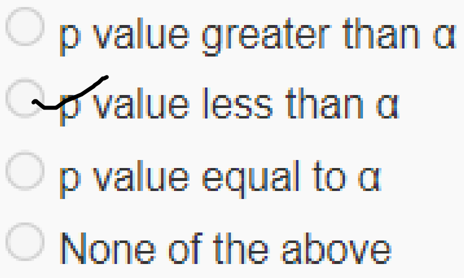
○ 4.36

○ None of the above



```
> sample_x = c(16,12,34,27,32,27,31)
> sample_y = c(25,23,14,25,32,27,32)
> var_sample_x = var(sample_x)/3
> var_sample_y = var(sample_y)/7
> round(var_sample_x/var_sample_y,2)
[1] 4.36
```

13) When will you reject the Null hypothesis?



○ p value greater than α

○ [✓ p value less than α]

○ p value equal to α

○ None of the above

- Since p value is calculated value, we will compare our p value with α (level of significance value)
- If p value is less than α therefore we reject the null hypothesis.

14) A standard normal density function is shown. The area of the shaded region is

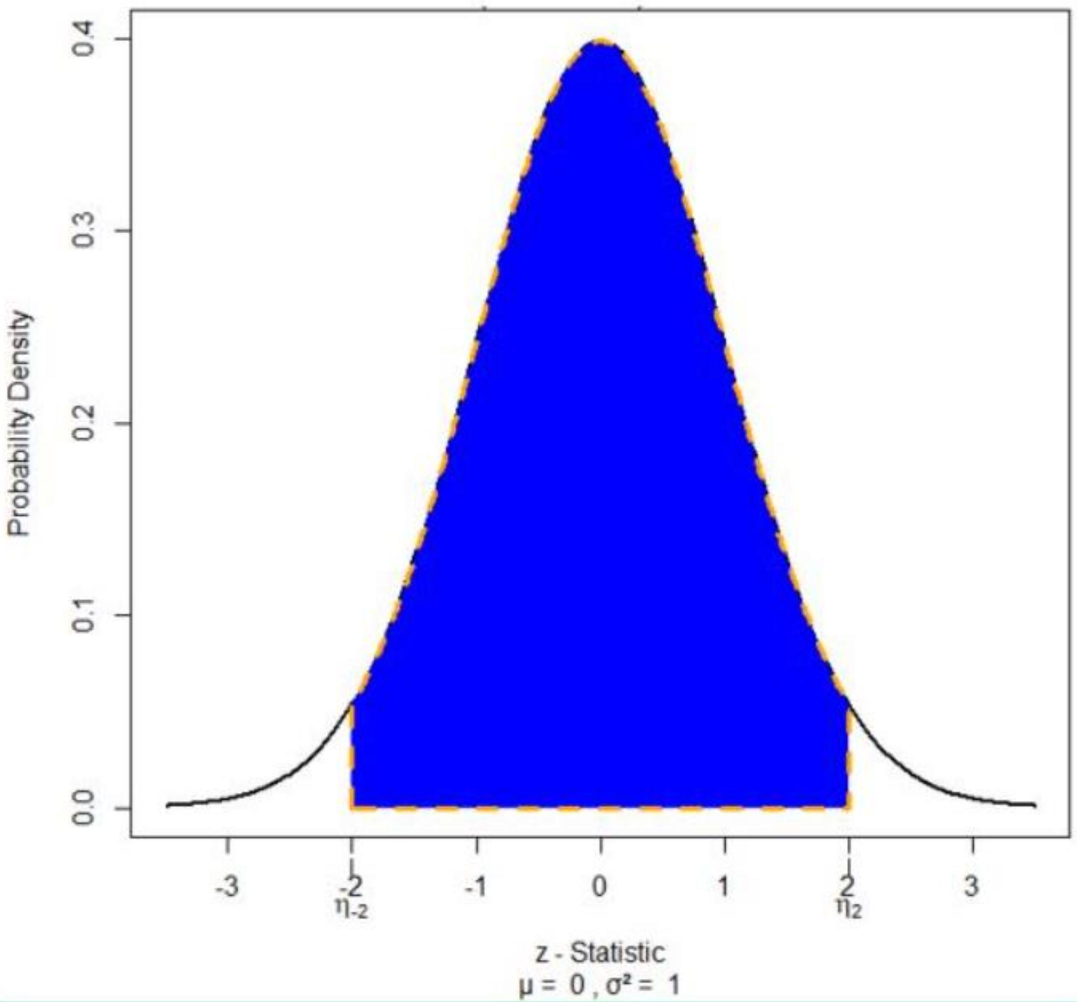
○ 0.046

○ 0.977

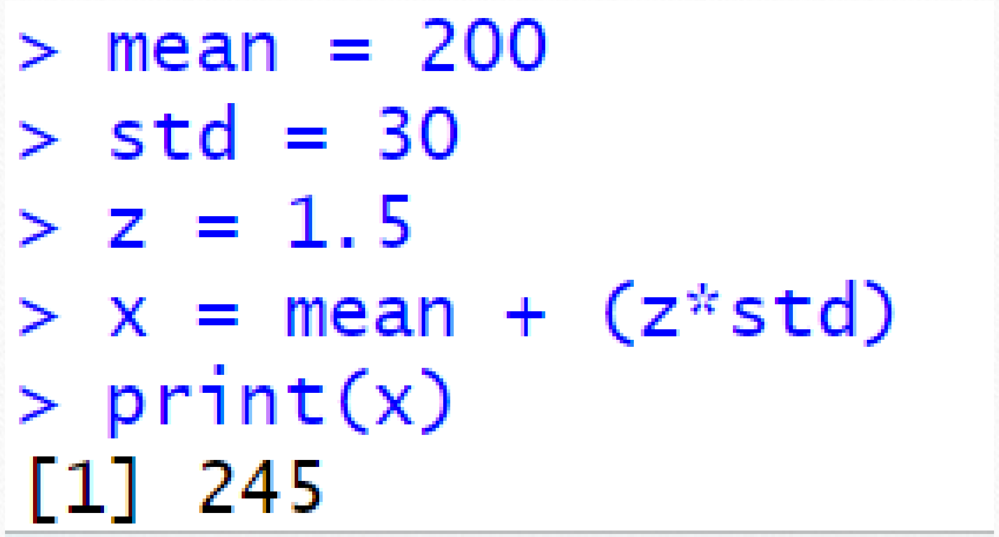
○ 0.954

○ None of the above

https://www.mathsisfun.com/data/standard-normal-distribution-table.html



15) A talent exam is conducted annually which has a mean score of 200 and a standard deviation of 30. If a student's Z-score is 1.50, what was his score in the exam?

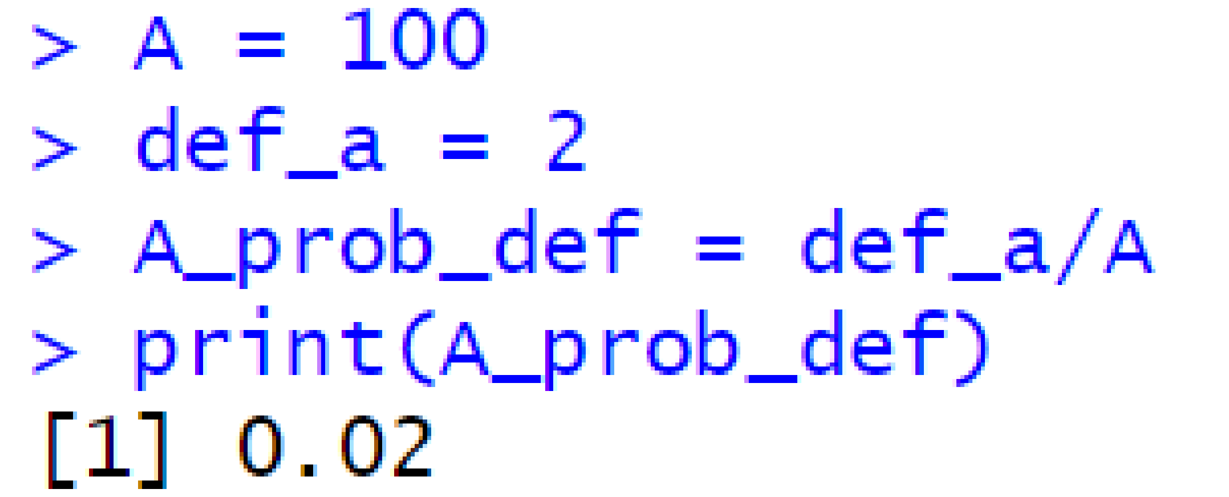


```
> mean = 200
> std = 30
> z = 1.5
> x = mean + (z*std)
> print(x)
[1] 245
```

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

$x = (z * \sigma) + \mu$

16) A car company purchases engine blocks from suppliers A, B, and C. Out of the 100 units supplied by A, two units were found to be defective. Similarly, out of 200 and 300 units supplied by B and C, the number units found to be defective were 10 and 15 respectively. If a quality control person of the car company picks up a block and if the selected block is from A, what is the probability that the block is defective?



```
> A = 100
> def_a = 2
> A_prob_def = def_a/A
> print(A_prob_def)
[1] 0.02
```

A	B	C
100	200	300
2	10	15

$2/100$

17) A car manufacturer purchases car batteries from two different suppliers. Supplier X provides 55% of the batteries and supplier Y provides the rest. 5% of all batteries from supplier X are defective and 4% of all batteries from supplier Y are defective. You select a battery from the bulk and you found it to be defective. What is the probability that it is from Supplier X?

	X	Y
	55%	45%
Def	5%	4%

Bulk = 100%

Def

$P(X/Def)$

$P(\text{selecting defective}) = P(\text{battery from X and it is def})$ or $P(\text{battery from Y and it is def.})$

$= 0.55 \times 0.05 + 0.45 \times 0.04$

$= 0.0455$

$P(X/def)$

$= \frac{P(X \cap def)}{P(def)}$

$= \frac{0.05 \times 0.55}{0.0455}$