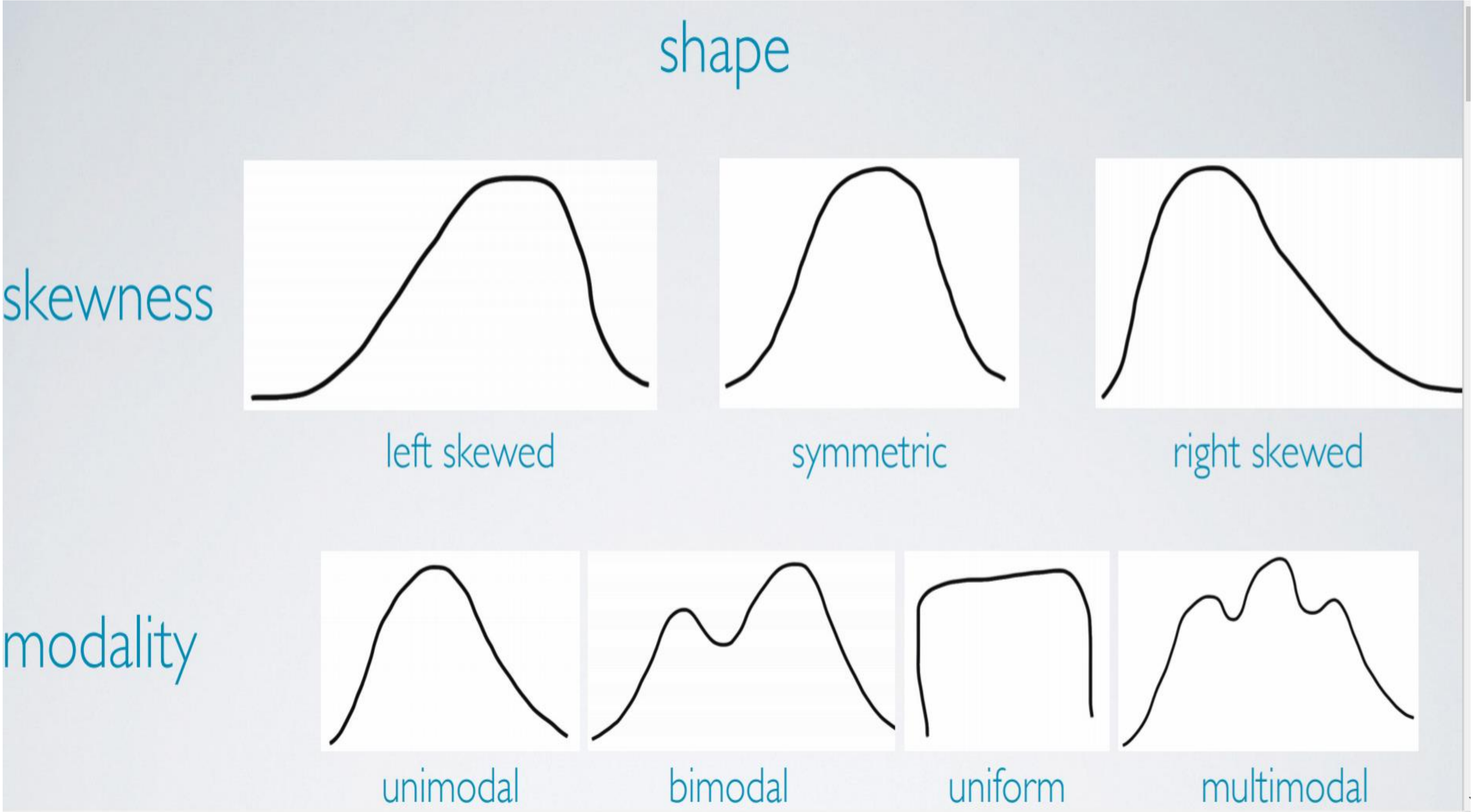
Data Science
Unit 2



shape

skewness

left skewed

symmetric

right skewed

modality

unimodal

bimodal

uniform

multimodal

Measures of Centre

- Measures of centre – mean mode median ; relation between mean, median in the context of skew.
- For left skewed data, generally mean < median
- For right skewed data, generally mean > median

Example

Which of the following is true?

a. In a symmetric distribution, more than 50% data are below and less than 50% are above the mean.

b. In a left skewed distribution, roughly 50% of data are below and 50% are above the mean

c. In a right skewed distribution, less than 50% of the data are below the mean

d. In a left skewed distribution, less than 50% of the data are below the mean

- Ans. ‘d’ (Since median is the 50th percentile, and in a left skewed distribution mean < median, less than 50% of the data will be smaller than the mean)

Measures of Spread and Robustness

- Measures of spread – Range, SD

- Robust stats – Centre : median (R), Mean (NR) ;
- Spread: IQR (R), SD, Range (NR)

- Robust stats are useful for describing skewed distributions or extreme observations
- Non Robust statistics are useful for symmetric distributions

Example

- You have collected data from a large Company with many employees having a salary of less than 1 L per month, a fewer managers who have salaries between 1 L to 1.5 L and a few high level executives whose salaries are beyond 1.5 L. determine the shape, these salaries would be expected to follow and decide whether median or mean would better represent a typical salary of an employee of the company.

A. Right skewed, mean is a better measure of salary

B. Right skewed, median is a better measure of salary

C. Symmetric, mean is a better measure of salary

D. Symmetric, median is a better measure of salary

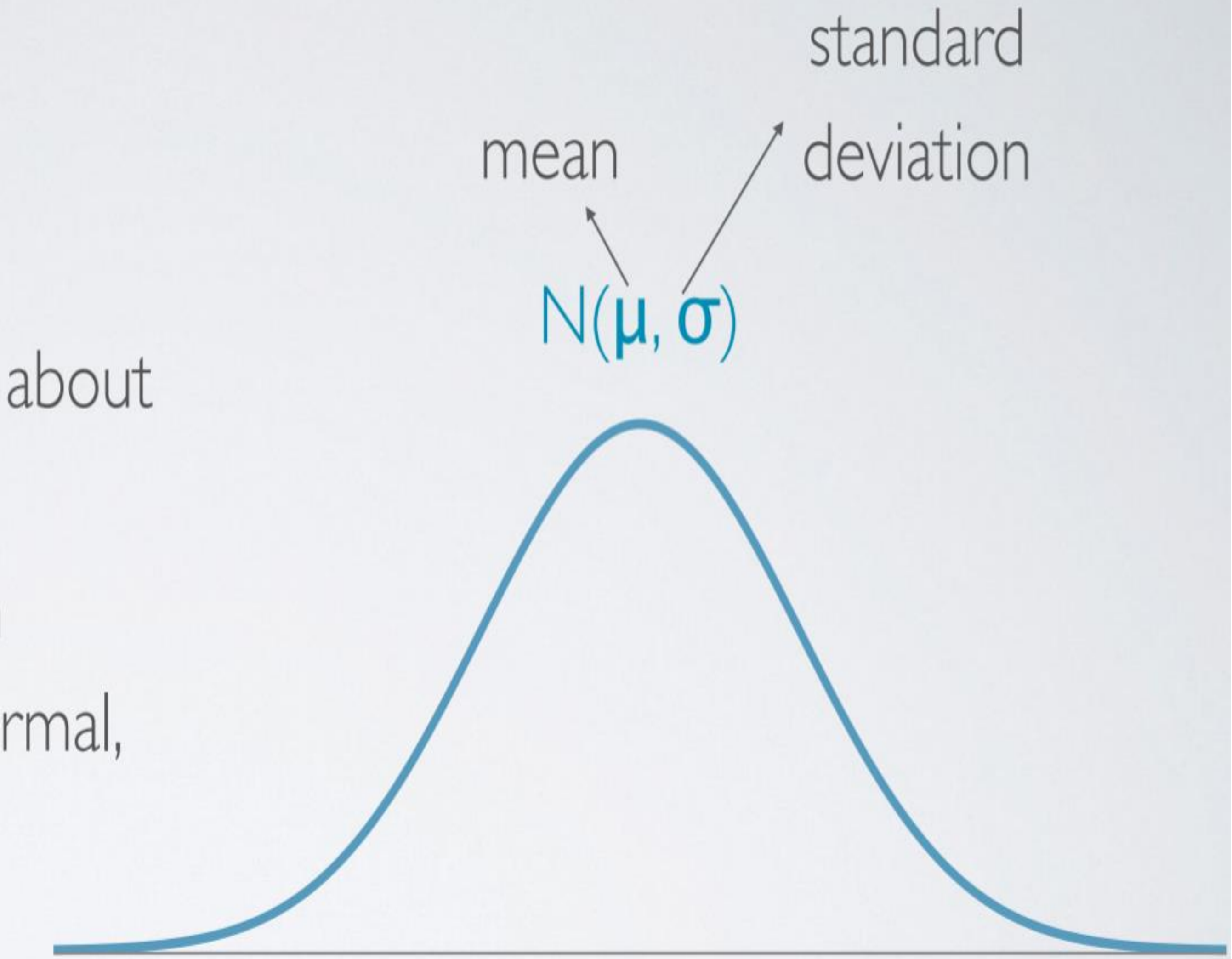
E. Left skewed, mean is a better measure of salary

F. Left skewed, median is a better measure of salary

- Ans. B

normal distribution

- unimodal and symmetric
  - bell curve
- follows very strict guidelines about how variably the data are distributed around the mean
- many variables are nearly normal, but none are exactly normal

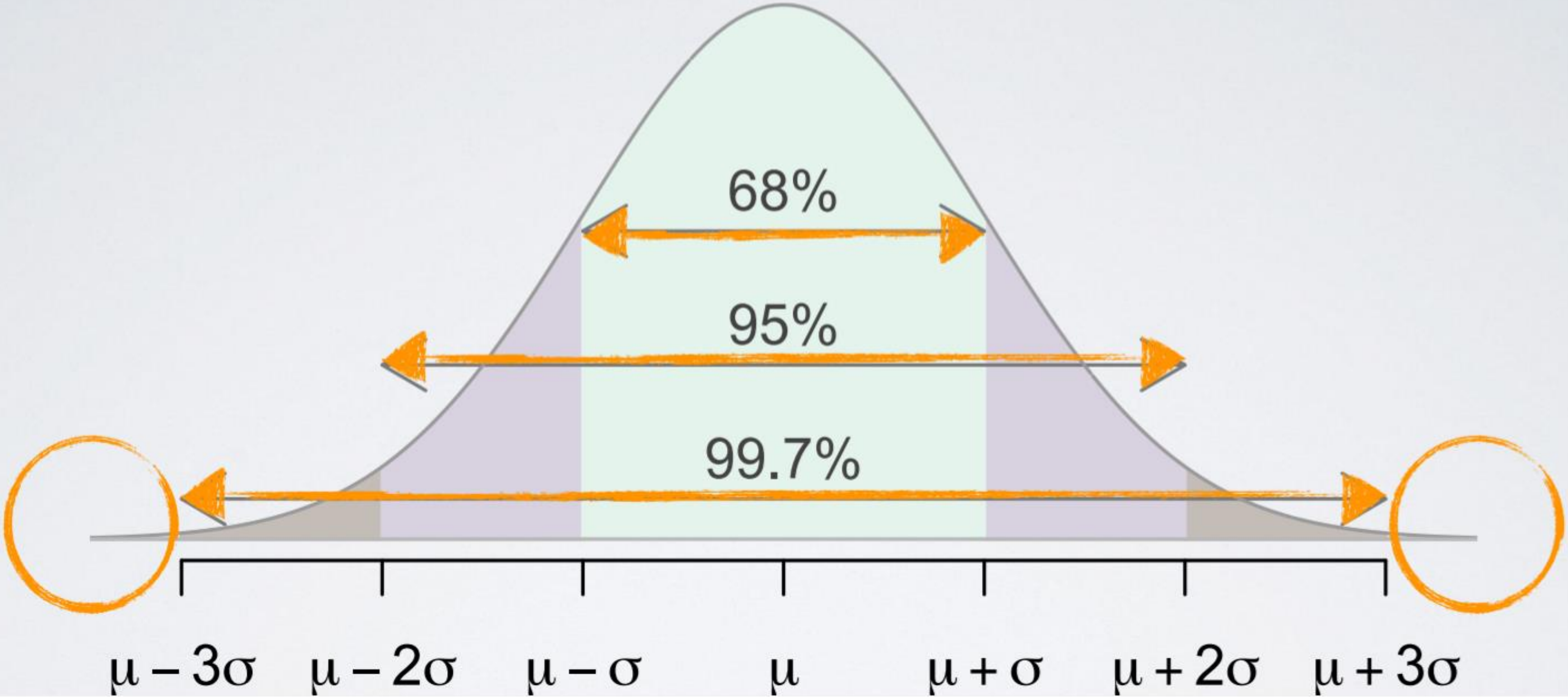


mean

standard deviation

$N(\mu, \sigma)$

68 - 95 - 99.7% rule



Examples

Scholastic Aptitude Test (SAT) scores are distributed nearly normally with mean 1500 and Standard Deviation 300, which of the following is false?

a. Roughly 68% of students score between 1200 to 1800
b. Roughly 95% of students score between 900 to 2100
c. Roughly 99.7% of students score between 600 to 2400
d. No students can score below 600

Ans. d

Example

- A doctor collects a large set of heart rate measurements, that approximately follow a normal distribution. The Doctor reports only 3 statistics mean = 110 beats per minute; minimum = 65 beats per minute and the maximum = 155 beats per minute. Which of the following is most likely to be the Standard deviation?

a. 5

b. 15

c. 35

d. 90

Ans. b

Example

- A college Admissions officer wants to determine which of the 2 applicants scored better in their test. Student Pam scored 1800 in Scholastic Aptitude Test (SAT) and Student Jim who scored 24 in American College Testing Examination.

SAT N(mu=1500, sigma=300)

ACT N(mu=21, sigma=5)

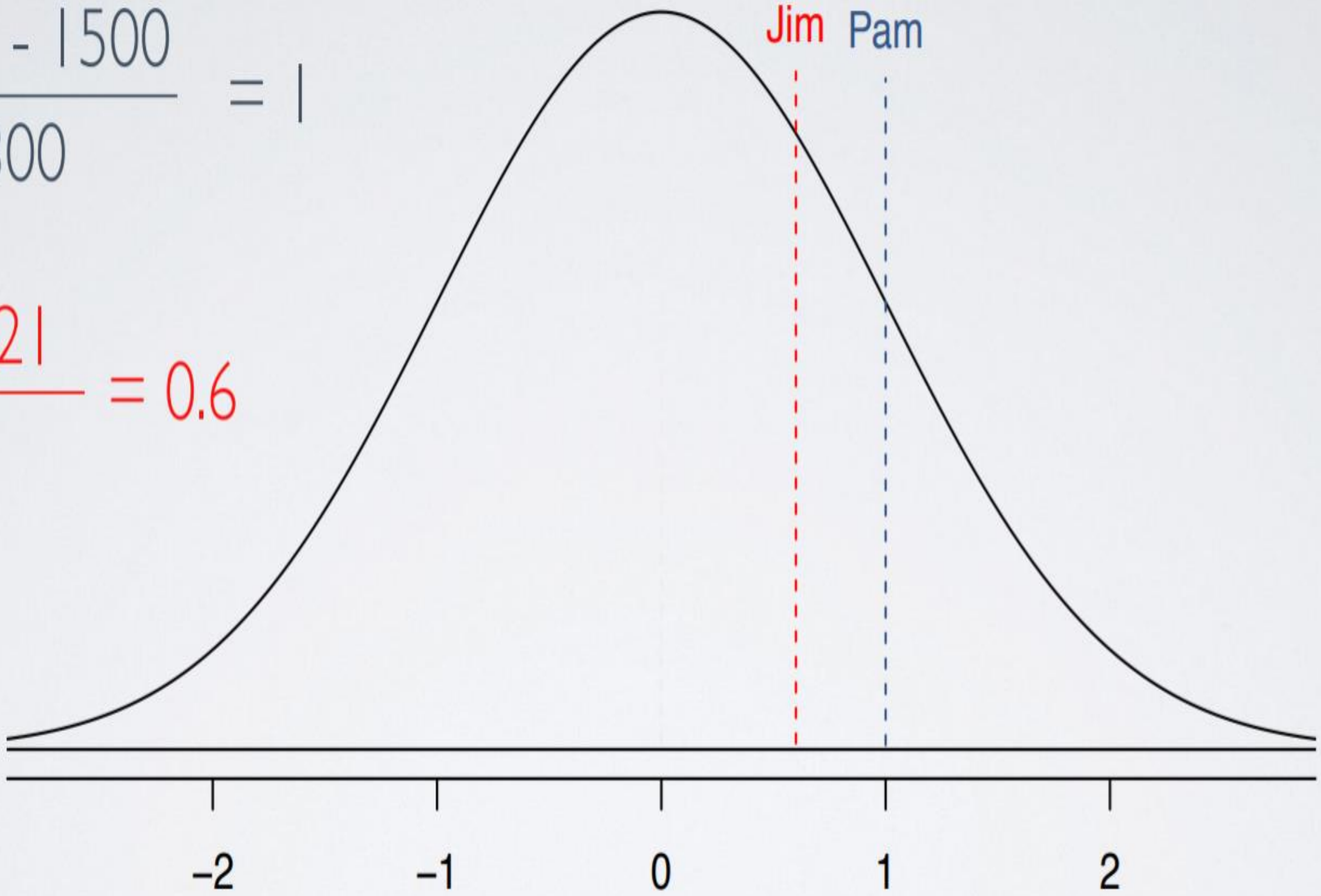
standardizing with Z scores

- standardized (Z) score of an observation is the number of standard deviations it falls above or below the mean
- Z score of mean = 0
- unusual observation: $|Z| > 2$
- defined for distributions of any shape

$$\boxed{Z = \frac{\text{observation - mean}}{\text{SD}}}$$

$$\text{Pam: } \frac{1800 - 1500}{300} = 1$$

$$\text{Jim: } \frac{24 - 21}{5} = 0.6$$



Example

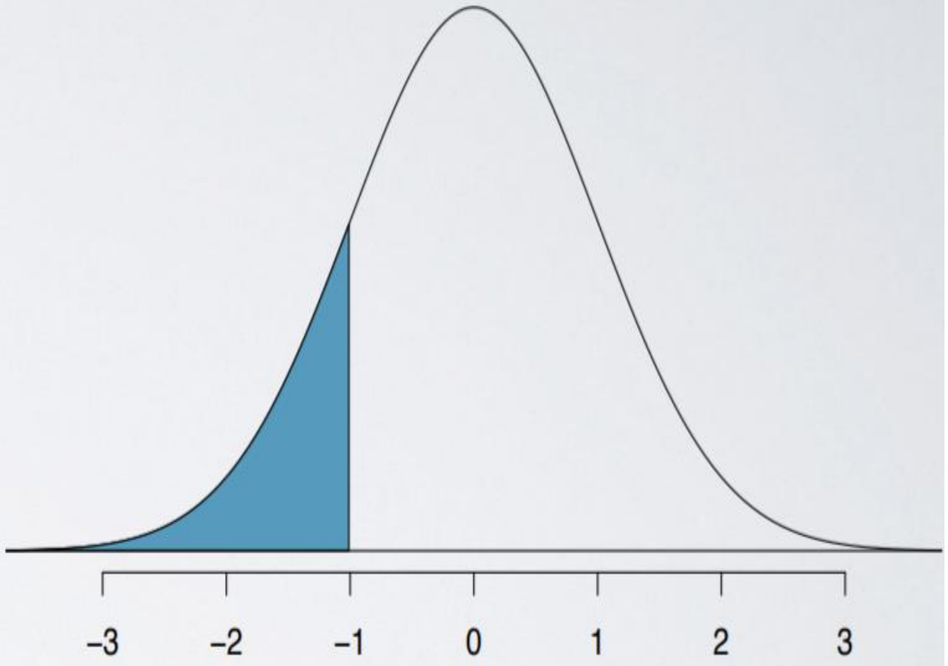
Scores on a standardized test are nearly normally distributed with a mean of 100 and a standard deviation of 20. If these scores are converted to standard normal Z-scores, which of the following statements will be correct?

a. The mean will be 0 and the median should be roughly 0 as well
b. The mean will equal 0, but the median can not be determined
c. The mean of the standardized scores will be 100
d. The mean of the standardized scores will be 5

Ans. ‘a’

percentiles

- when the distribution is normal, Z scores can be used to calculate percentiles
- percentile is the percentage of observations that fall below a given data point
- graphically, percentile is the area below the probability distribution curve to the left of that observation.

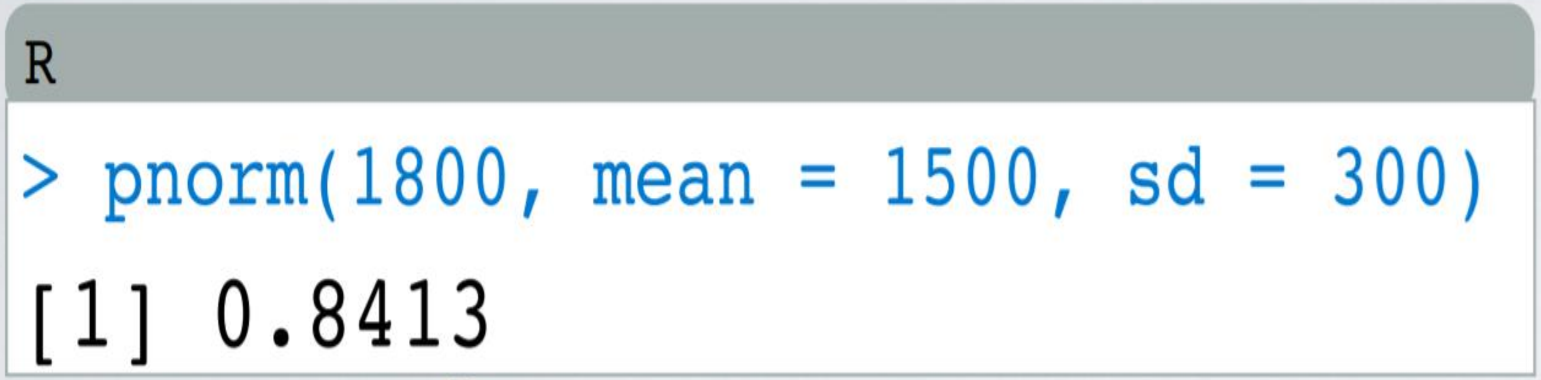


Example

SAT scores are distributed normally, with mean 1500 and SD 300. John scored 1800 in his SAT. What is his percentile score?

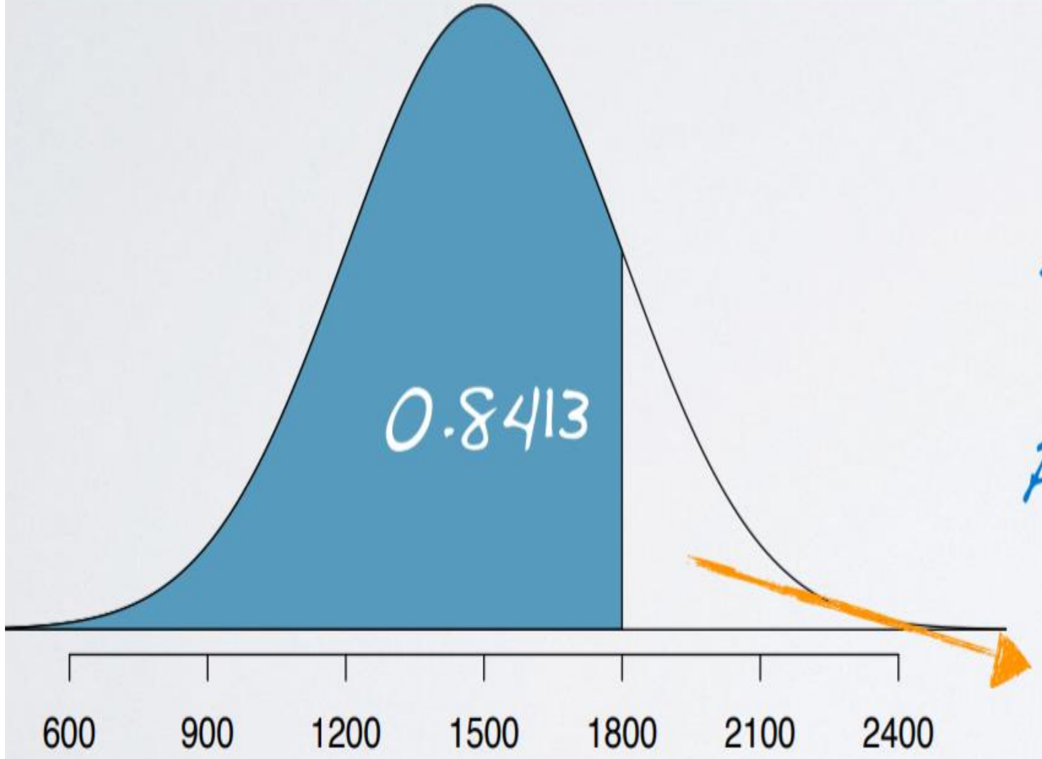
```
 > pnorm (1800, mean = 1500, SD = 300)
[1] 0.8413
```

This means John scored better than 84.13% of the SAT students.



R

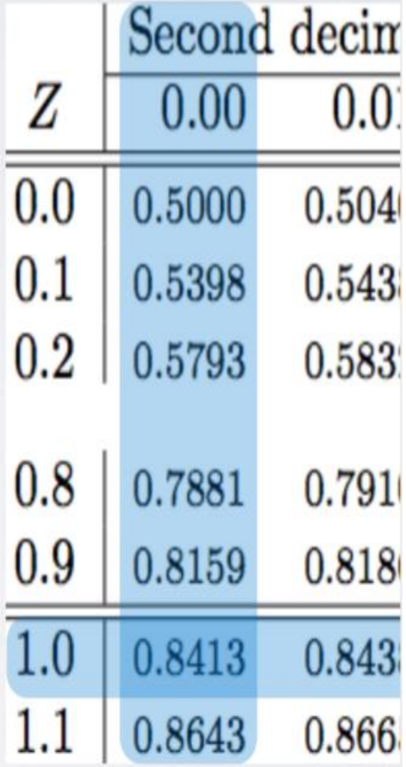
```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413
```



$$Z = \frac{1800 - 1500}{300} = 1$$

$P(Z < 1) = 0.8413$

$1 - 0.8413 = 0.1587$



	Second decim	
Z	0.00	0.0
0.0	0.5000	0.504
0.1	0.5398	0.543
0.2	0.5793	0.583
0.8	0.7881	0.791
0.9	0.8159	0.818
1.0	0.8413	0.843
1.1	0.8643	0.866

Example

A friend tells you that she scored in the top 10% of the SAT (mean = 1500, SD = 300). What is the lowest score she could have gotten?

The total area under the curve is 1, the percentile score associated with cut off value for top 10% is 1 - 0.1 = 0.9

Using the table, the value associated with 90th percentile i.e. 0.9 is 0.8997 and corresponding Z-score is 1.28.

Hence, 1.28 = (x – 1500)/ 300 i.e. x = 1884.

In R ,

```
>qnorm(0.90,1500,300)
[1]1884.465
```

Example

ACT scores are distributed nearly normally, with mean 21 and SD 5. A friend of yours tells you that he scored in the bottom 10% in the exam. What is the highest possible score, he would have got?

a. 14.6

b. 27.4

c. 12.75

d. 29.25

Ans. ‘a’

Example

Suppose weights of checked baggage of airlines passengers follow a nearly normal distribution with mean 45 pounds and SD 3.2 pounds. Most airlines charge a fee for baggage that weigh in excess of 50 pounds. What % of airline passengers are expected to incur this fee?

If in a month, 64000 passengers visited the airport, with Rs. 2500 as excess charges, how much revenue was generated at the airport?

Example

Z = (50-45)/3.2 = 1.56

For a Z-score of 1.56, we get 0.9406 as area below the curve.

Hence complementing it, 1-0.9406 = 0.0594 i.e. 5.94% is the expected answer

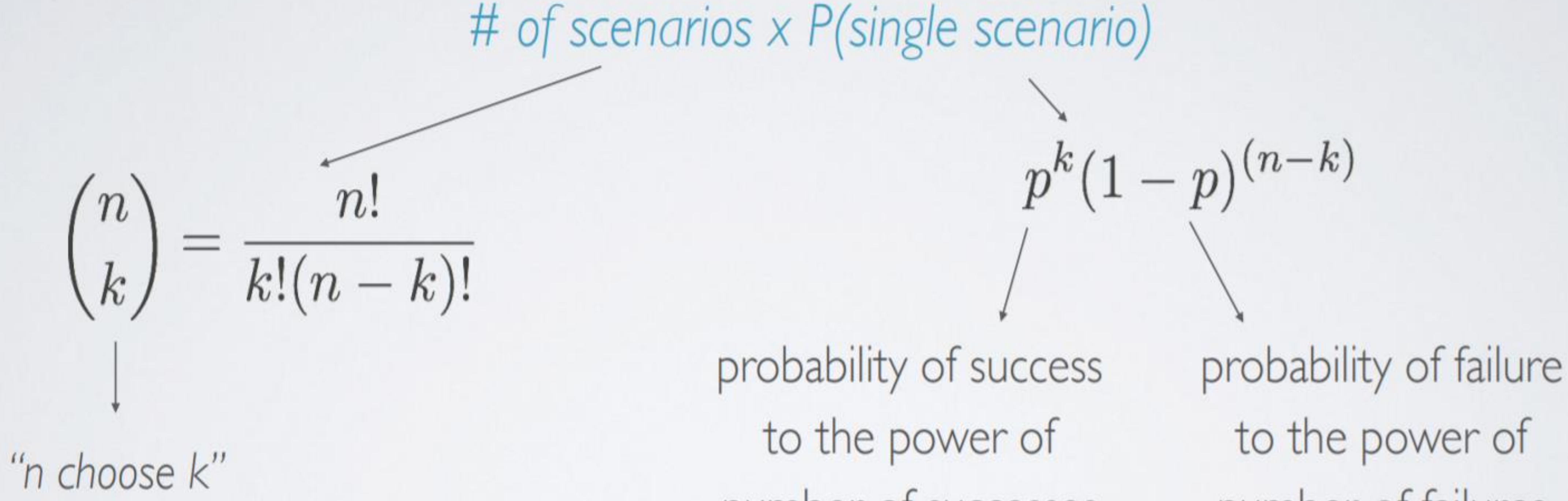
The revenue generated would be –

5.94 of 64000 = 3802 passengers

Revenue generated = Rs. 95,05,000

binomial distribution

the binomial distribution describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success p



# of scenarios x P(single scenario)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

"n choose k"

$$p^k(1-p)^{(n-k)}$$

probability of success to the power of number of successes

probability of failure to the power of number of failures

Binomial distribution:

If *p* represents probability of success, *(l -p)* represents probability of failure, *n* represents number of independent trials, and *k* represents number of successes

$$\text{P(k successes in n trials)} = \binom{n}{k} p^k \, (1-p)^{(n-k)}$$

where $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$

Binomial Conditions

- For a variable to follow a binomial distribution the conditions to follow are –

1. The trials must be independent
2. The number of trials ‘n’ must be fixed
3. Every trial outcome must be classified as a success or failure
4. The probability of success, p, must be same for each trial

Example

According to a 2013 Gallup poll, worldwide only 13% of employees are engaged at work (psychologically committed to their jobs and likely to be making positive contributions to their organizations). Among a random sample of 10 employees, what is the probability that 8 of them are engaged at work?

n = 10; p (success = engaged) = 0.13; 1-p = 0.87 ; k = 8

P(k=8) = (10 choose 8) 0.13^{8} x 0.87^{2}
= 0.00000278

Mean & Standard Deviation

Among a random sample of 100 employees, how many would be expected to be engaged at work? p = 0.13

$\mu = 100 \times 0.13 = 13$

In mathematical terms,

Expected value of Binomial distribution = $\mu = np$

Mean & Standard Deviation

In every random sample of 100 employees, exactly 13 will NOT be engaged at work. How much do we expect the value to vary?

This variability around the mean can be anticipated using Standard Deviation.

$\sigma = (np(1-p))^{1/2}$

In the given case, $\sigma = (100 \text{ x } 0.13 \text{ x } 0.87)^{1/2}$
$= 3.36$

This means that 13 out of 100 employees are expected to be engaged at work, give or take approximately 3.36 employees.

Example

- A 2012 Gallup poll survey suggests that 26.2% of Americans are obese. Which of the following is false?

a. Among a random sample of 1000 Americans, we can expect 262 to be obese

b. Random samples of 1000 Americans, where there are at most 230 are obese people, would be considered unusual

c. The standard deviation of number of obese Americans in random samples of 1000 is roughly 14

d. Random samples of 1000 Americans, where at least 300 are obese would not be considered unusual

Ans. ‘d’

Central Limit Theorem (CLT): The distribution of sample statistics is nearly normal, centered at the population mean, and with a standard deviation equal to the population standard deviation divided by square root of the sample size.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{s\,\cancel{\sigma}}{\sqrt{n}}\right)$$

shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** Either the population distribution is normal, or if the population distribution is skewed, the sample size is large (rule of thumb: n > 30).

CLT- Example

Suppose you have a slightly right skewed population distribution of annual incomes in a developed nation, with mean 30000$ and SD 20000$. Suppose you take 10000 random samples of size 625 from this population. Which of the following is most likely to be the distribution of the means of these samples?

a. Right skewed, mean = 30000$, SD = 20000$

b. Nearly normal, mean = 30000$, SD = 20000$

c. Nearly normal, mean = 30000$, SD = 20000$ / $(625)^{1/2}$

d. Nearly normal, mean = 30000$, SD = 20000$ / $(10000)^{1/2}$

e. Left skewed, mean = 30000$, SD = 20000$ / $(625)^{1/2}$

Ans. ‘c’

Confidence Interval for a mean

Confidence Interval is defined as a plausible range of values for the population parameter

Using only a sample statistic for estimation of parameter is unreliable. If we report a point estimate, we most probably wont hit the exact population parameter, but if we report a range of plausible values, we have a good shot at capturing the parameter

Sample mean **X** is our best guess for the unknown population mean. Hence any interval we construct is to be around that **X** that we know, to be our best guess.

From CLT we know **X** is distributed nearly normally and the center of the distribution is at the unknown population mean

A plausible range of values for the population parameter is called a confidence interval.

- If we report a point estimate, we probably won't hit the exact population parameter.
- If we report a range of plausible values we have a good shot at capturing the parameter.

Confidence Interval for a mean

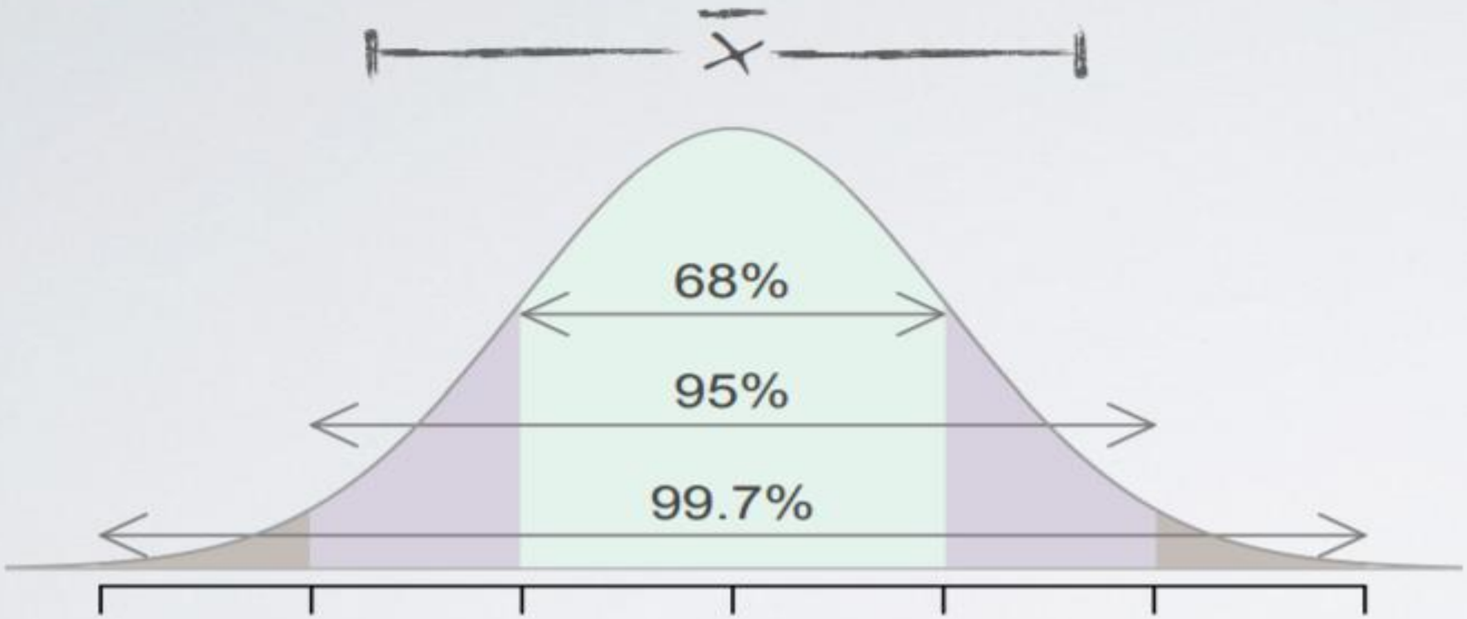
Considering the nearly normally distribution rule, we can state that roughly 95% of random samples will have sample means that are within 2 standard errors of the population mean.

So the 95% Confidence Interval can be constructed approximately as sample mean ± 2 SE i.e. Approx. 95% CI : X ± 2 SE

± 2 SE is called Margin of Error (ME)

Central Limit Theorem (CLT):

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$



approximate 95% CI: $\bar{x} \pm \boxed{2SE}$

margin of error (ME)

Confidence Interval of a Population Mean

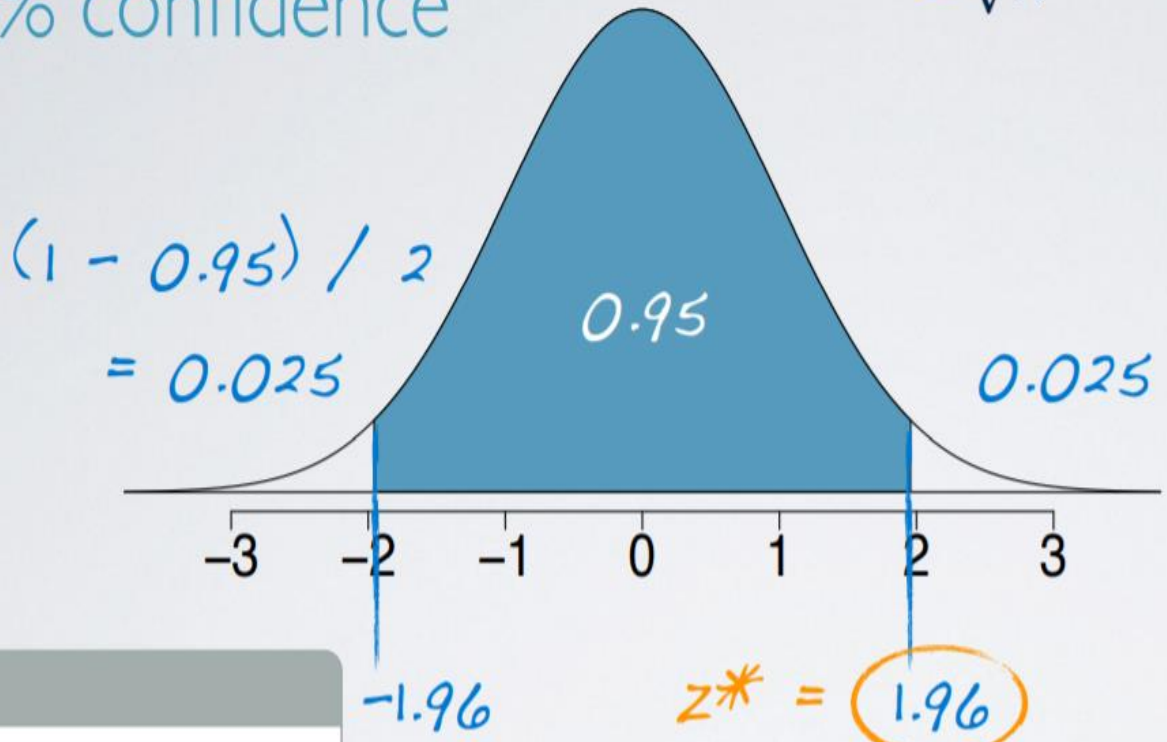
Conditions for this confidence interval are same as for CLT –

1. Independence – sampled observations must be independent. We either want a random sample if we have an observational study or a random assignment if we have an experiment. If we are sampling without replacement, we want our sample size to be < 10% of the population.
2. Sample size / skew – $n \geq 30$. Larger, if the population distribution is more skewed

Note: sampling with replacement – when a sampling unit is drawn from a finite population, and is returned to that population, after its characteristics have been recorded, before the next unit is drawn at random, the sampling is said to be with replacement

finding the critical value
95% confidence

$$\bar{x} \pm z^{\star} \frac{s}{\sqrt{n}}$$



```
R
> qnorm(0.025)
[1] -1.96
```

	Second decimal place				
0.07	0.06	0.05	0.04	0.00	Z
0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0004	0.0005	−3.3
0.0005	0.0006	0.0006	0.0006	0.0007	−3.2
0.0008	0.0008	0.0008	0.0008	0.0010	−3.1
0.0011	0.0011	0.0011	0.0012	0.0013	−3.0
0.0015	0.0015	0.0016	0.0016	0.0019	−2.9
0.0021	0.0021	0.0022	0.0023	0.0026	−2.8
0.0028	0.0029	0.0030	0.0031	0.0035	−2.7
0.0038	0.0039	0.0040	0.0041	0.0047	−2.6
0.0051	0.0052	0.0054	0.0055	0.0062	−2.5
0.0068	0.0069	0.0071	0.0073	0.0082	−2.4
0.0089	0.0091	0.0094	0.0096	0.0107	−2.3
0.0116	0.0119	0.0122	0.0125	0.0139	−2.2
0.0150	0.0154	0.0158	0.0162	0.0179	−2.1
0.0192	0.0197	0.0202	0.0207	0.0228	−2.0
0.0244	0.0250	0.0256	0.0262	0.0287	−1.9
0.0307	0.0314	0.0322	0.0329	0.0359	−1.8

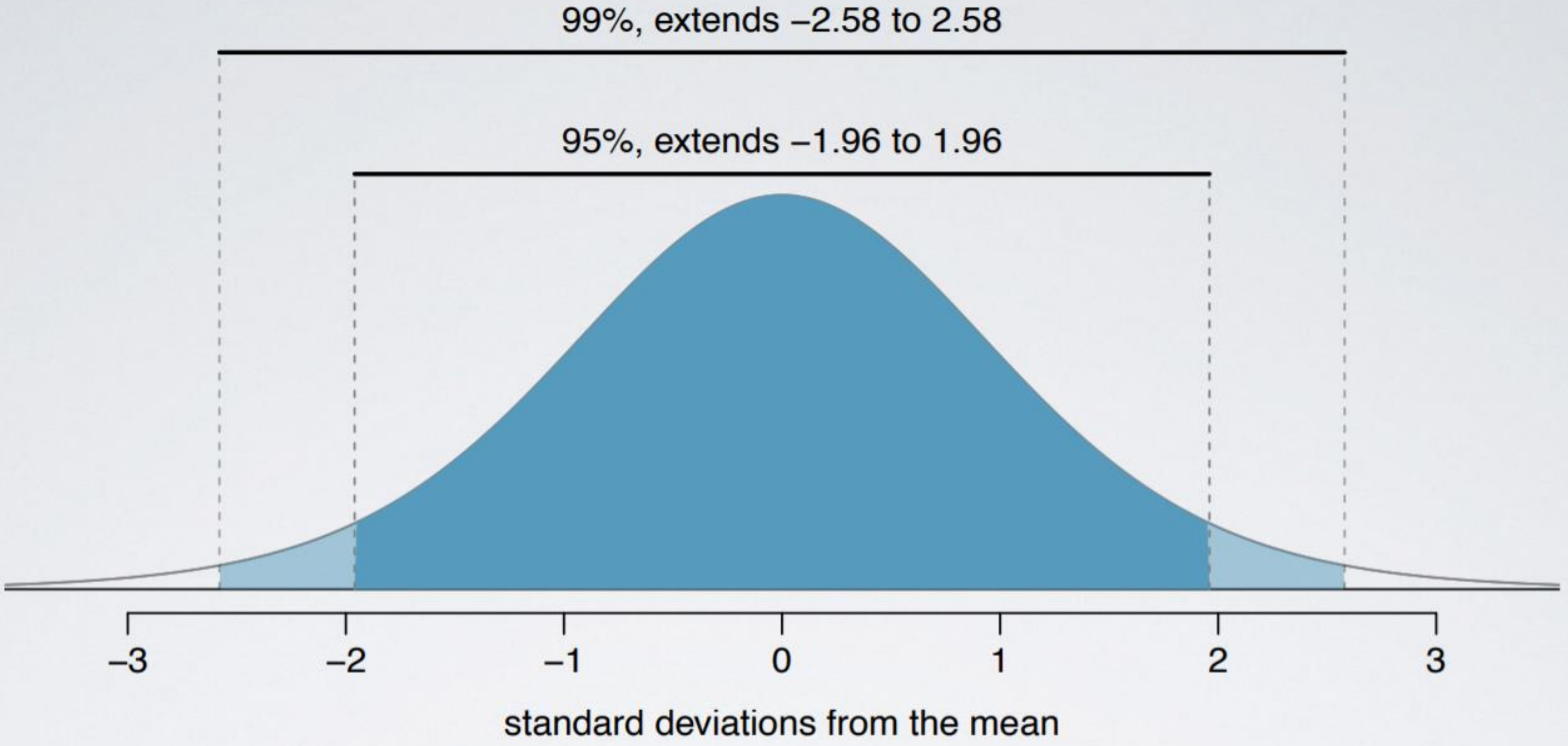
Confidence Interval of a Population Mean

```
> qnorm(0.025)
[1] -1.96
```

What is the critical value for the 98% Confidence Interval?

a. Z = 2.05
b. Z = -1.96
c. Z = 2.33
d. Z = -2.33
e. Z = 1.96

Ans. ‘c’


99%, extends –2.58 to 2.58
95%, extends –1.96 to 1.96
–3
–2
–1
0
1
2
3
standard deviations from the mean

CL ↑ width ↑

Accuracy V/s Precision

- As the confidence level increases, the larger the critical value, hence the larger ME and hence the width of the confidence interval also increases

- Weather forecast….. Next day maximum temperature would be between 5 deg to 45 deg.
- Such weather forecast is not precise. It is not informative. It doesn’t help me decide whether to wear a sweater or light cotton attire.
- As the confidence level increases, the width of CI increases, which increases accuracy. However precision goes down.

Accuracy V/s Precision

- In order to get higher precision as well as higher accuracy, increase sample size. It reduces SE and hence ME. Therefore we can remain at a high confidence level while not needing to increase the Confidence Interval

backtracking to n for a given ME

given a target margin of error, confidence level, and information on the variability of the sample (or the population), we can determine the required sample size to achieve the desired margin of error.

$$\boxed{ME = z^{\star}\frac{s}{\sqrt{n}} \quad \rightarrow n = \left(\frac{z^{\star}s}{ME}\right)^2}$$

A group of researchers want to test the possible effect of an epilepsy medication taken by pregnant mothers on the cognitive development of their children. As evidence, they want to estimate the IQ scores of three-year-old children born to mothers who were on this medication during pregnancy.

Previous studies suggest that the SD of IQ scores of three-year-old children is 18 points.

How many such children should the researchers sample in order to obtain a 90% confidence interval with a margin of error less than or equal to 4 points?

$ME \leq 4$ pts

$CL = 90\%$

$z^* = 1.65$

$\sigma = 18$

$$4 = 1.65 \frac{18}{\sqrt{n}} \longrightarrow n = \left(\frac{1.65 \times 18}{4}\right)^2 = 55.13$$

We need at least 56 such children in the sampl

obtain a maximum margin of error of 4 points.

We found that we needed at least 56 children in the sample to achieve a maximum margin of error of 4 points. How would the required sample size change if we want to further decrease the margin of error to 2 points?

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{n}} \frac{1}{2}$$

$$\frac{1}{2} ME = z^* \frac{s}{\sqrt{4n}}$$

$$4n = 56 \times 4 = 224$$

Example

- A sample of 50 college students were asked, how many course projects they have done on their own so far. The students in the sample had an average of 3.2 projects with a SD of 1.74. Estimate the true average number of projects based on this sample, using a 95% Confidence Interval.
- SE = 1.74 / sqrt (50) = 0.246
- 95% CI = 3.2 ± 1.96 (0.246) = [2.72, 3.68]

- Meaning that we are 95% confident that college students on an average have done 2.72 to 3.68 course projects on their own.

two competing claims

null hypothesis

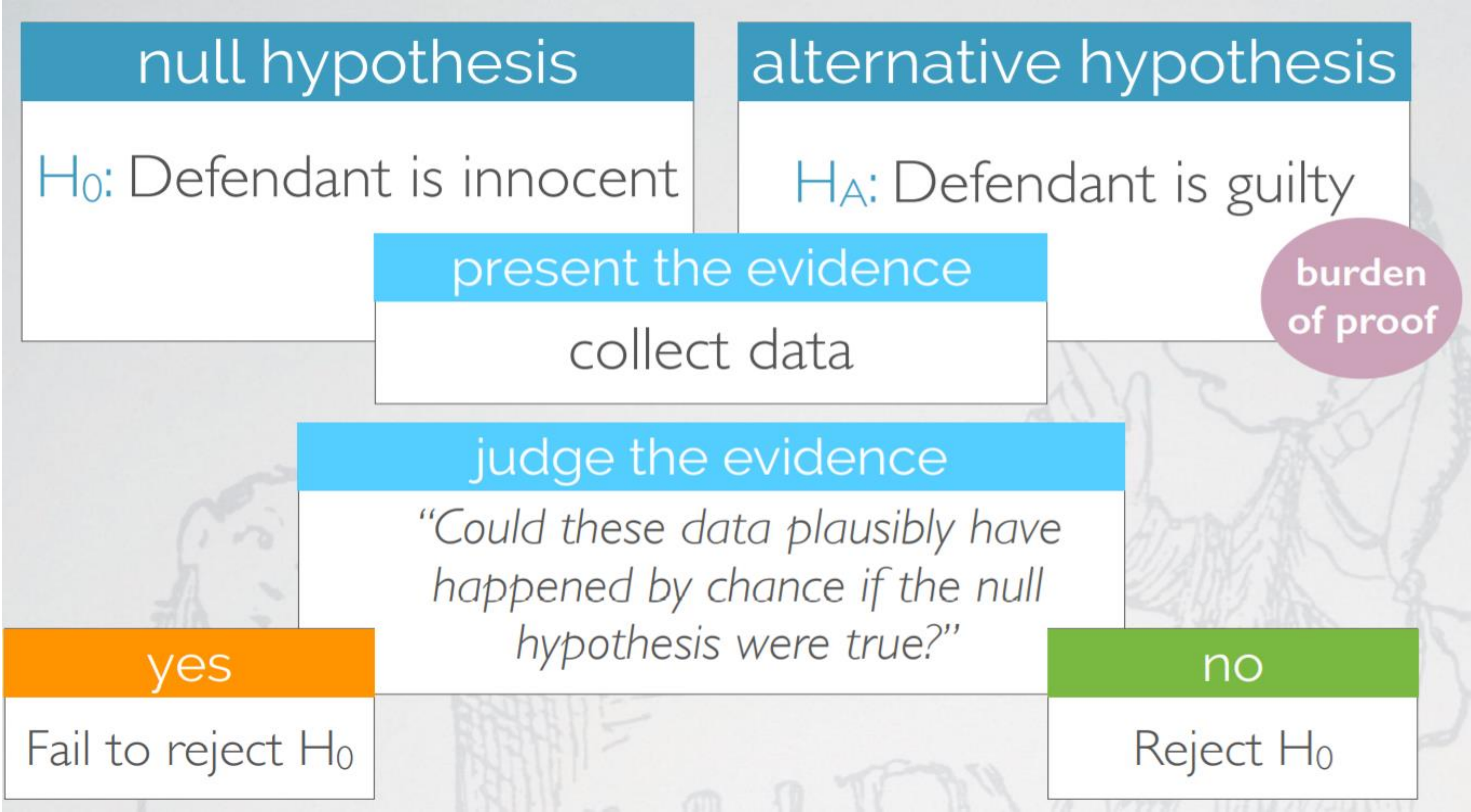
"There is nothing going on"

promotion and gender are independent, no gender discrimination, observed difference in proportions is simply due to chance

alternative hypothesis

"There is something going on"

promotion and gender are dependent, there is gender discrimination, observed difference in proportions is not due to chance.



null hypothesis

H_0: Defendant is innocent

alternative hypothesis

H_A: Defendant is guilty

burden of proof

present the evidence

collect data

judge the evidence

"Could these data plausibly have happened by chance if the null hypothesis were true?"

yes

Fail to reject H_0

no

Reject H_0

Hypothesis Testing - Terminologies

- Null Hypothesis – It is often a utopic, ideal perspective, based on societal norms. Or at times it is a claim to be tested. We set the parameter of interest, equal to some value. It is expressed as Ho
- Alternative Hypothesis – Represents an alternative claim under consideration and is often represented by a range of possible parameter values. We claim that the parameter of interest is either < , >, ≠ the same null value from the null hypothesis. It is expressed as Ha
- The hypothesis is always about the population parameters (µ in this case) and never about sample statistics.

Example

- A study suggests that the average college student spends about 2 hours per day communicating with others online. You believe that this is an underestimate and decide to collect your own sample for the hypothesis test. You randomly sample 60 students from your Institute and find that on average, they spend 3.5 hours a day communicating with others online. What are the appropriate hypothesis?

a. Ho: $X = 2$ Hrs. / day ; Ha : $X > 2$ Hrs. / day

b. Ho: $X = 2$ Hrs. / day ; Ha : $X > 3.5$ Hrs. / day

c. Ho: $\mu = 2$ Hrs. / day ; Ha : $\mu > 2$ Hrs. / day

d. Ho: $\mu = 2$ Hrs. / day

Ans. ‘c’

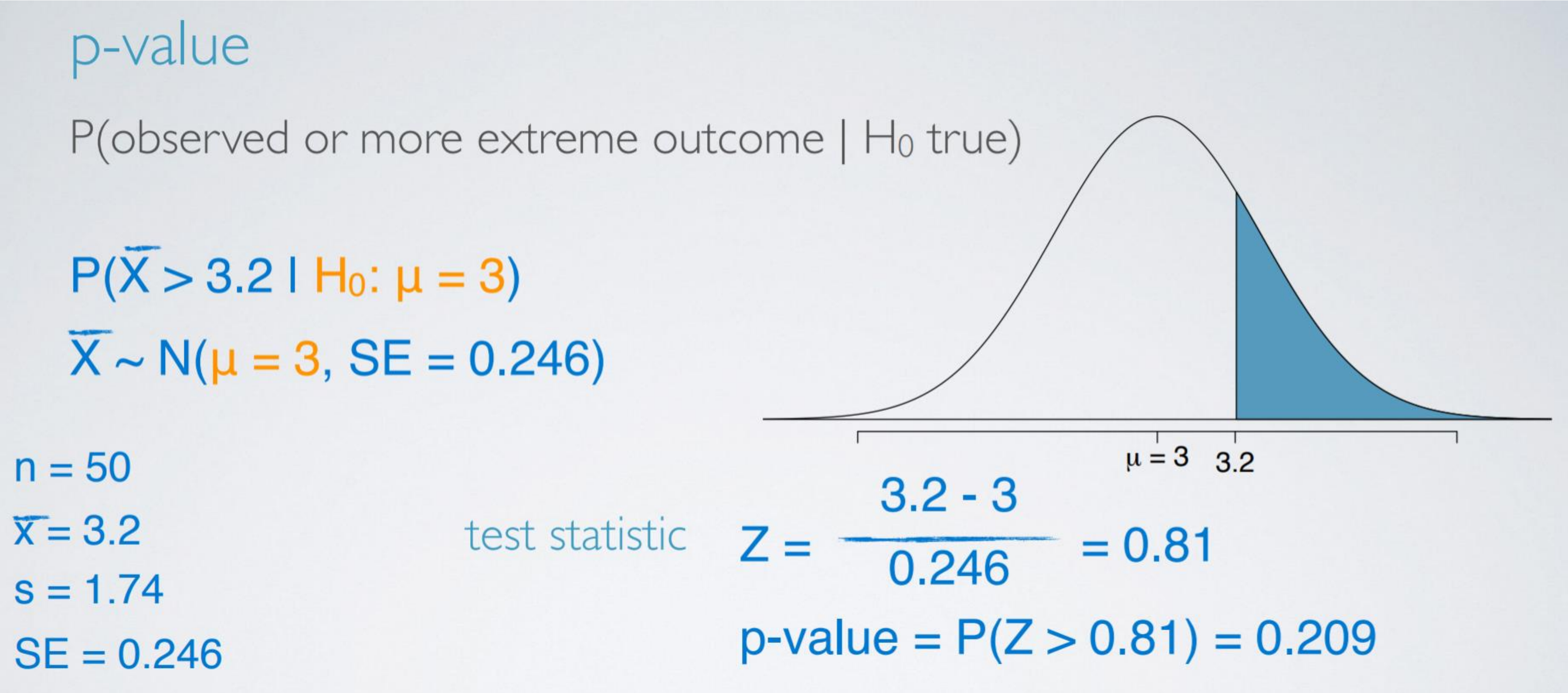
Course Projects example

- P-value: The probability of observing data at least as extreme as the one observed in the original study, under the assumption that the null hypothesis is true, is called the p-value. It is the numerical criteria used for making decisions between competing hypotheses.
- Let us make the null hypothesis that the mean value is $\mu = 3$.
- Hence the p-value is

$P(X > 3.2 \mid Ho: \mu = 3)$

- By default we assume null hypothesis to be true.

Course Projects example



p-value

P(observed or more extreme outcome | H_0 true)

$P(\bar{X} > 3.2 \mid H_0: \mu = 3)$

$\bar{X} \sim N(\mu = 3, SE = 0.246)$

$n = 50$

$\bar{x} = 3.2$

$s = 1.74$

$SE = 0.246$

test statistic

$$Z = \frac{3.2 - 3}{0.246} = 0.81$$

p-value $= P(Z > 0.81) = 0.209$

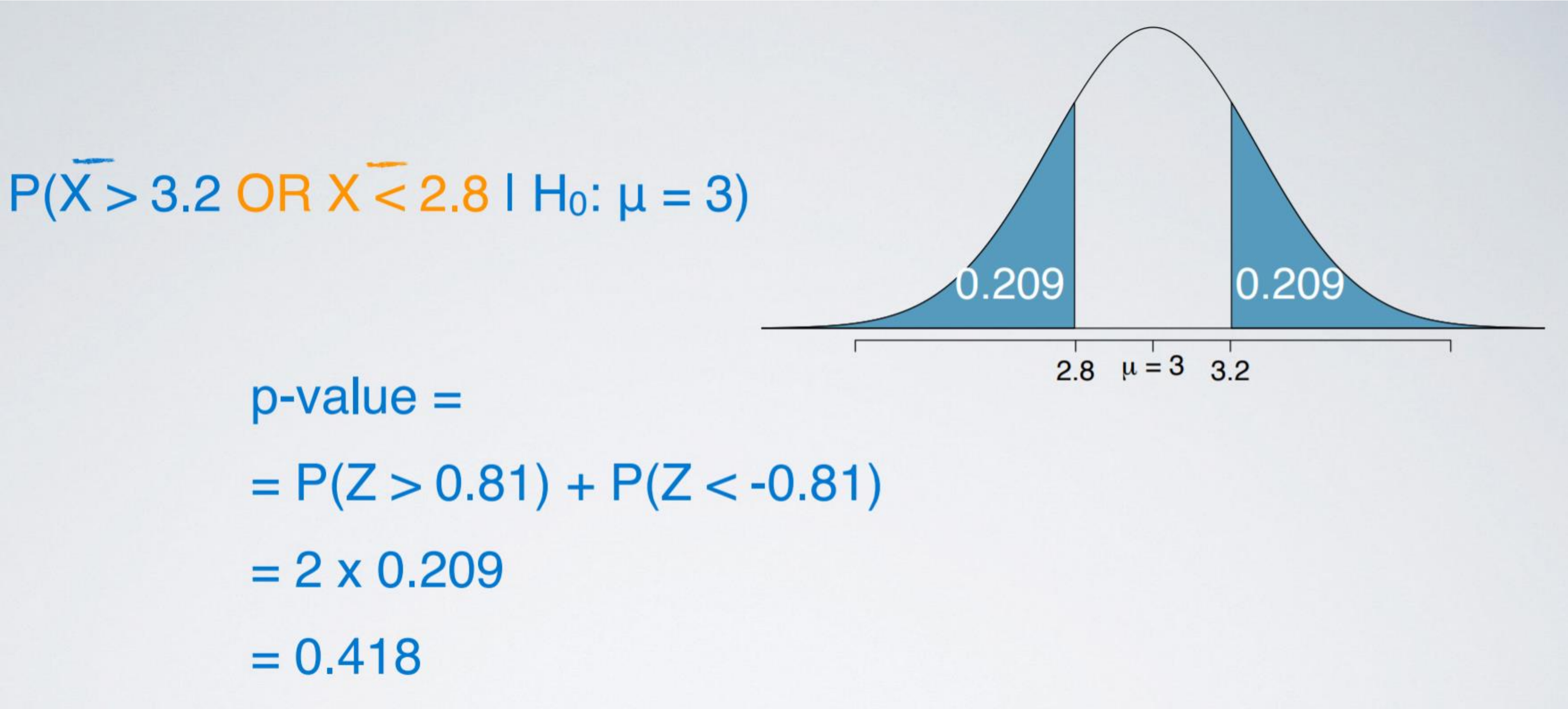
Decision based on the p-value

- We used the test –statistic to calculate the p-value, which is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis was true.
- If the p-value is low – lower than a defined significance level α, which is usually 5%, we say that it would be very unlikely to observe the data, if the null hypothesis were true and hence reject null hypothesis.
- If the p-value is high – (Higher than significance level α), we say that it is likely to observe the data even if the null hypothesis were true, and hence do not reject the null hypothesis.
- Since p-value is 0.209, we do not reject the null hypothesis

Two-sided tests

- Often instead of looking for a divergence from the null in a specific direction, we might be interested in divergence in any direction. Such hypothesis tests are called two-sided or two-tailed tests.
- The definition of a p-value is the same, however the calculation is slightly different, since we need to consider ‘at least as extreme as the observed outcome’ in both the directions, away from the mean.

Two-sided tests



$P(\bar{X} > 3.2 \text{ OR } \bar{X} < 2.8 \mid H_0: \mu = 3)$

p-value =

$= P(Z > 0.81) + P(Z < -0.81)$

$= 2 \text{ x } 0.209$

$= 0.418$

Example

Chain restaurants have been required to display calorie count of each menu item. Prior to displaying calorie counts, the average calorie intake of diners at a restaurant was 1100 calories. After calorie counts started to be displayed on menus, a nutritionist collected data on the number of calories consumed at this restaurant from a random sample of diners. Which of the following would be used to test whether there was evidence of a difference in the average calorie intake of diners at this restaurant?

a. Ho: $\mathbf{x} = 1100$ calories; Ha: $\mathbf{x} \neq 1100$ calories

b. Ho: $\mathbf{x} = 1100$ calories; Ha: $\mathbf{x} < 1100$ calories

c. Ho: $\mu = 1100$ calories; Ha: $\mu \neq 1100$ calories

d. Ho: $\mu < 1100$ calories; Ha: $\mu > 1100$ calories

Ans. ‘c’

Algorithm for Hypothesis testing

1. Set the hypothesis, Ho: μ = null value ; Ha: μ < or > or ≠ null value
2. Calculate the point estimate – sample mean **x** in this case
3. Check for applicability conditions – Similar to CLT i.e. independence (random sample for obs. Study / assignment for experiment) & sample size / skew. We consider this is met with each time.
4. Draw sampling distribution, shade p-value, calculate test statistic Z*
5. Make a decision and interpret it in the context of research question

Example

P-value of approximately 0 provides –

a. Weak evidence against null hypothesis
b. Strong evidence against null hypothesis
c. Strong evidence against alternative hypothesis
d. Weak evidence against alternative hypothesis
e. No evidence against the null hypothesis

Ans. ‘b’