Curtin College

DIPLOMA OF ENGINEERING

EMTH1019 WEEK 3

Curtin College

Sampling Distribution & Estimation

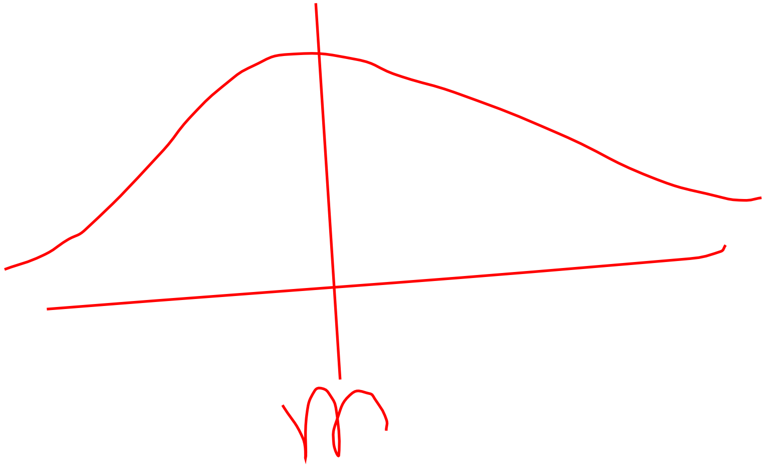
Checking data for normality

Obtain the sampling distribution of the mean

Use a confidence interval to estimate a population parameter

WEEKS 3 AND 4

- The next two weeks builds on the Normal Distribution content from week 2.
- It is essential that you are competent in this introductory normal distribution content.



WEEK 3 KEY POINTS

In week 2 we determined z when μ and σ were **known** $\boxed{z = \frac{x-\mu}{\sigma}\ .}$

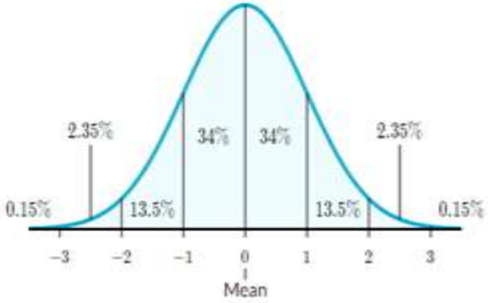
Week 3

- **Bias and incorrect assumptions** can make calculations **meaningless**
- If $\boldsymbol{\sigma}$ **(*Pop sd*) is known** but μ (*Pop mean*) **is unknown** can still use z tables (conditions apply)
- Modified z equation $\boxed{z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}$
  - $\bar{x}$ is the sample mean and n is the sample size.

	sample	population
mean	$\bar{x}$ $\bar{X}$	μ
std. dev	$\sigma/\sqrt{n}$	σ

- Central Limit Theorem (CLT) if you take multiple SRS samples of size n from a population then the mean of the samples should be μ.
  - **Mean of the means is μ**
- CLT does **not** tell you anything about $\boldsymbol{\sigma}$
- **Error**: rearrange $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ to get $\boldsymbol{z\frac{\sigma}{\sqrt{n}} = |\bar{x} - u|}$= error $(\text{eg.} = \pm 1\ psi)$

NORMALITY

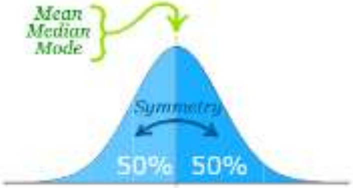


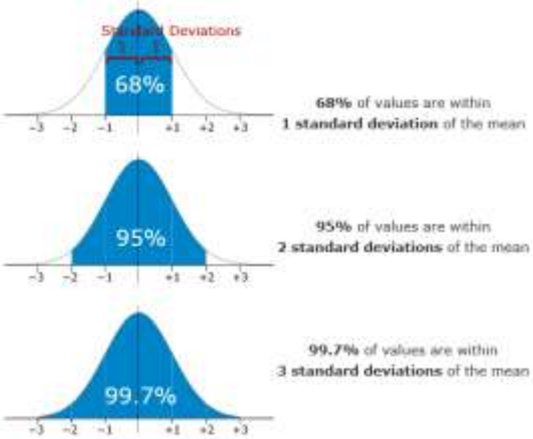
There are lots of things that are normally distributed:

- Blood pressure, mass or length of “things” produced

Normally distributed data has ALL of these and more:

- Mean=mode=median
- 50% of values < mean, 50% of values > mean
- Symmetrical
- 68% of data lies within ±1 standard d
- 95% of data within ±2 SD
- 99.7% of data within ±3 SD
- A complex formula




Standard Deviations
68%
68% of values are within
1 standard deviation of the mean
95%
95% of values are within
2 standard deviations of the mean
99.7%
99.7% of values are within
3 standard deviations of the mean
-3
-2
-1
+1
+2
+3

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

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

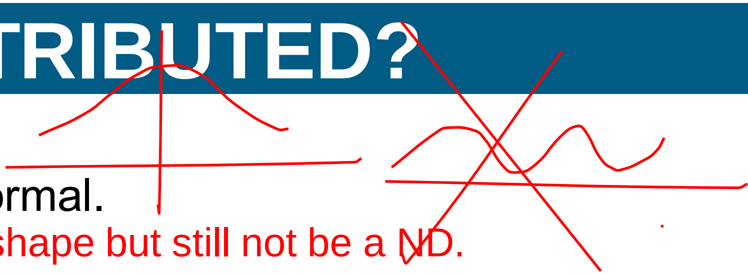
TESTING FOR NORMALITY

Rainfall data - Is my data normal enough that I can use z?

1.88	2.23	2.58	2.07	2.94	2.29	3.14
2.15	1.95	2.51	2.86	1.48	1.12	2.76
3.10	2.05	2.23	1.70	1.57	2.81	1.24
3.29	1.87	1.50	2.99	3.48	2.12	4.69
2.29	2.12					

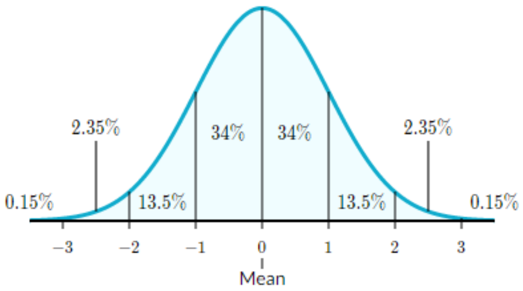
Curtin College

IS MY DATA NORMALLY DISTRIBUTED?

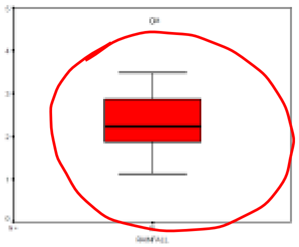


You could:

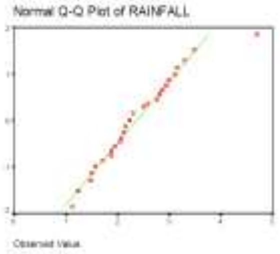
1. Plot the data in a histogram and see if it “looks” normal.
   - The histogram can be symmetrical and have a bell shape but still not be a ND.
2. Calculate the mean, mode, median. Are they equal (enough)?
   - This may show symmetry but not all symmetrical distributions are normal distributions.
3. Calculate the standard deviation and then see what percentage of values lie in each interval.
   - This is still quite rough and no guarantee of a ND.



4. Create a boxplot
   - Median is in middle of box
   - Whiskers should be slightly longer than half the box length
   - Whiskers should be ~ equal
   - It should look symmetrical
   - Still no guarantee that the distribution is ND.



5. Normal Quantile Plot
   - Plots observed values against predicted value.
   - Better test.



SERIOUS TESTING FOR NORMALITY

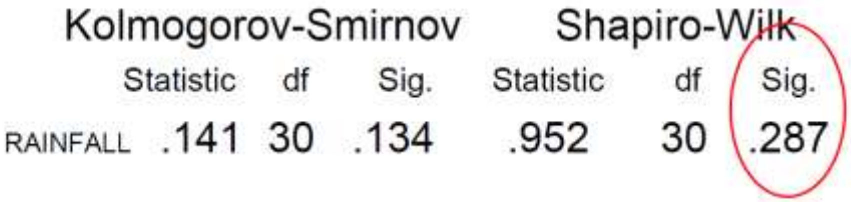
The previous methods are all somewhat subjective

2 tests that are often used to test for Normality. *They both have limitations*.

- Kolmogorov – Smirnov
- Shapiro-Wilk

If we put the rainfall data through a statistical package such as “R” or even Excel we can calculate these statistics (difficult calculations).

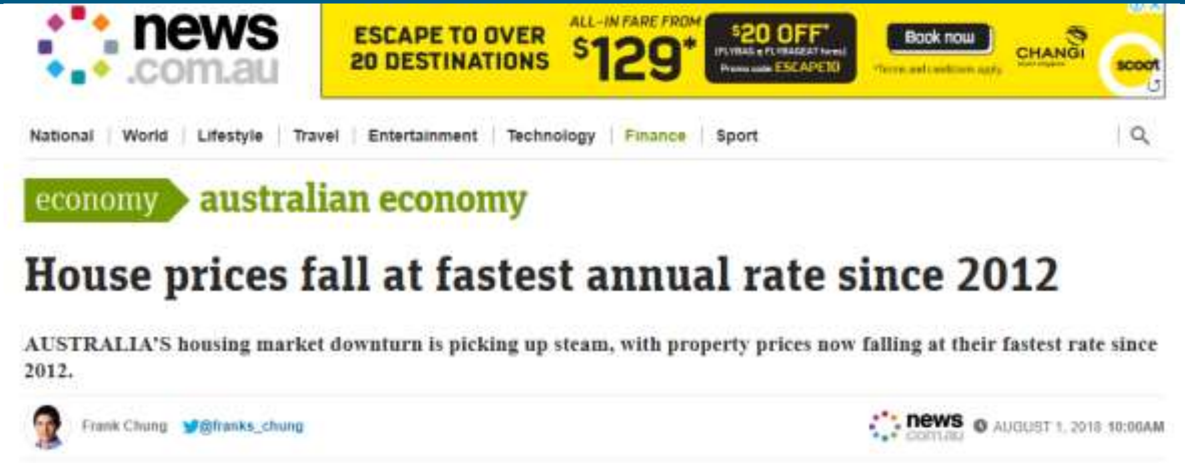
From “R”



	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
RAINFALL	.141	30	.134	.952	30	.287

IF… the ***Shapiro-Wilk Sig*** $\geq \mathbf{0.05}$ then we can say that the evidence ***suggests*** it is OK to treat the parent population as normally distributed.

BEWARE OF HEADLINES AND PERCENTAGE CHANGES


news.com.au
ESCAPE TO OVER 20 DESTINATIONS
ALL-IN FARE FROM $129*
$20 OFF*
Promo code ESCAPE10
Book now
CHANGI
scoot
National | World | Lifestyle | Travel | Entertainment | Technology | Finance | Sport
economy
australian economy

House prices fall at fastest annual rate since 2012

AUSTRALIA'S housing market downturn is picking up steam, with property prices now falling at their fastest rate since 2012.

Frank Chung @franks_chung

news.com.au AUGUST 1, 2018 10:00AM

- House prices are 1.9% lower than their September 2017 peak
  - Do house prices have to always ↑?
- Median house price decline in 5 of 8 capital cities
- Melbourne prices
  - Top quartile ↓4.1%, lower quartile ↑ 7.5%
  - Median house price fell 1.8% to $709,568 in last quarter
  - What was the old median? Does this mean anything?
- House prices are still 31% higher than 5 years ago
  - So is the 1.8% ↓ significant?

https://www.news.com.au/finance/economy/australian-economy/house-prices-fall-at-fastest-annual-rate-since-2012/news-story/c0312b566721d373a9d32d67e19e480b

SIMPLE RANDOM SAMPLE - SRS

There is nothing simple about getting a SRS

Take a sample of 5 numbers and calculate the mean.

5	12	14	6	18
15	0	13	10	15
19	16	13	7	19
12	18	15	11	9
7	0	19	17	10

Definition

A **SRS** of size ***n*** consist of ***n*** units from a population chosen in such a way that every possible group of ***n*** units has the same probability of being chosen as the sample.

This can be difficult to achieve

BIAS

The design of an experiment is biased if it systematically favors certain outcomes

Why did you choose the numbers you did?

Was there a pattern to the way you chose the numbers?

Did you purposely choose or not choose 18,18,19,19,19?

Were you trying to choose 5 numbers that might add up to the mean?

How could we ensure that the 5 numbers we select are a SRS?

Good data

- Controlled, randomized, planned, does not only use the easy to collect data, removes selection bias.

Good design of experiment

- You don’t go to basketball courts to find the average uni student height!

US PRESIDENTIAL ELECTIONS 1936

In 1936 the incumbent United States President, Franklin D. Roosevelt, was up for re-election and faced Republican Alf Landon. The Great Depression was 7 years old.

The Literary Digest, an influential weekly magazine of the time, had begun political polling and had correctly predicted the outcome of the previous 5 presidential elections.

biased ?

For this election *The Literary Digest* polled a sample of over **2 million** people based upon telephone and car registrations. Based on this data they predicted Landon would win with over **57%** of the popular vote.

REALITY CHECK

Roosevelt won **60.8%** of the popular vote and all but 2 states with the largest landslide in US election history.

How did the Literary Digest get it so wrong?

https://www.qualtrics.com/blog/the-1936-election-a-polling-catastrophe/

Curtin College

HOW DID THEY GET IT SO WRONG?

The sample was biased

For this election *The Literary Digest* polled a sample of over 2 million people based upon telephone and car registrations. ***n*** was very large.

- During the Great Depression only the wealthy owned phones and cars.
- The wealthy were generally supporting Landon and his policies
- *The Literary Digest* was sampling the "wealthy" phone and car owners
- The "not wealthy" still voted
- People don't always tell the truth when asked how they will vote…

An incorrect sample, no matter how large, is still INCORRECT

random sample

BREXIT – UK VOTING ON LEAVING THE EU

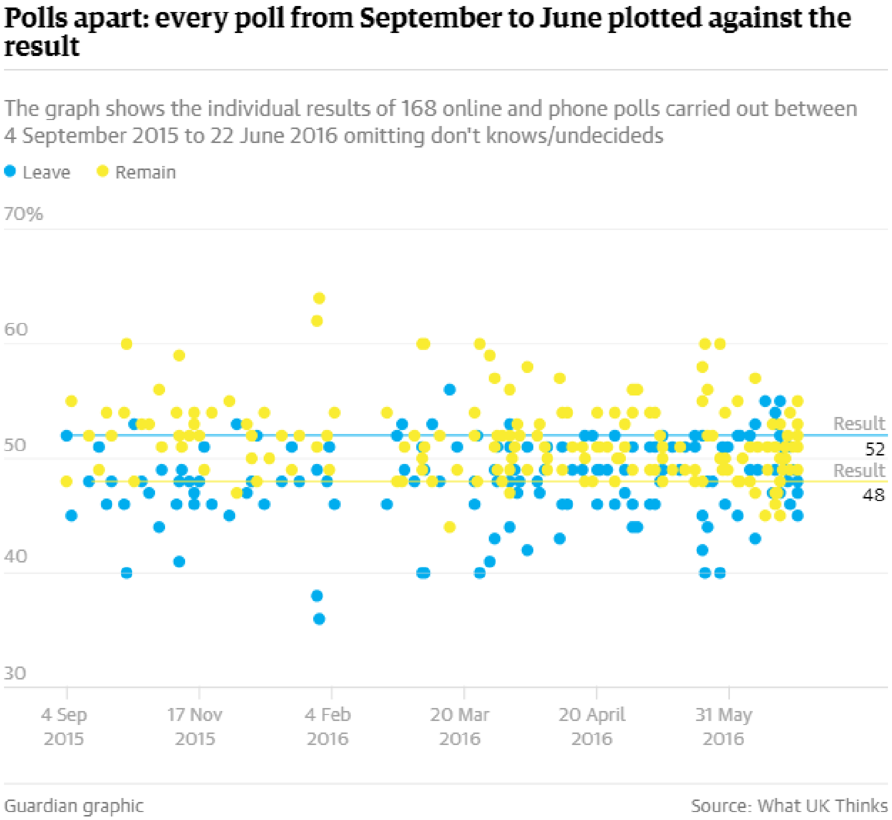
Did the pollsters get it very wrong on the EU referendum?

- 168 polls carried out before vote
- Only 16 predicted 52:48 leave EU

Awful graph but…

- Look at blue line and blue dots
- Blue=leave EU
- Blue line is actual result
- Dots are polls
- The average of blue dots ~48%
- Actual result ~52%

- https://www.theguardian.com/politics/2016/jun/24/how-eu-referendum-pollsters-wrong-opinion-predict-close



Polls apart: every poll from September to June plotted against the result

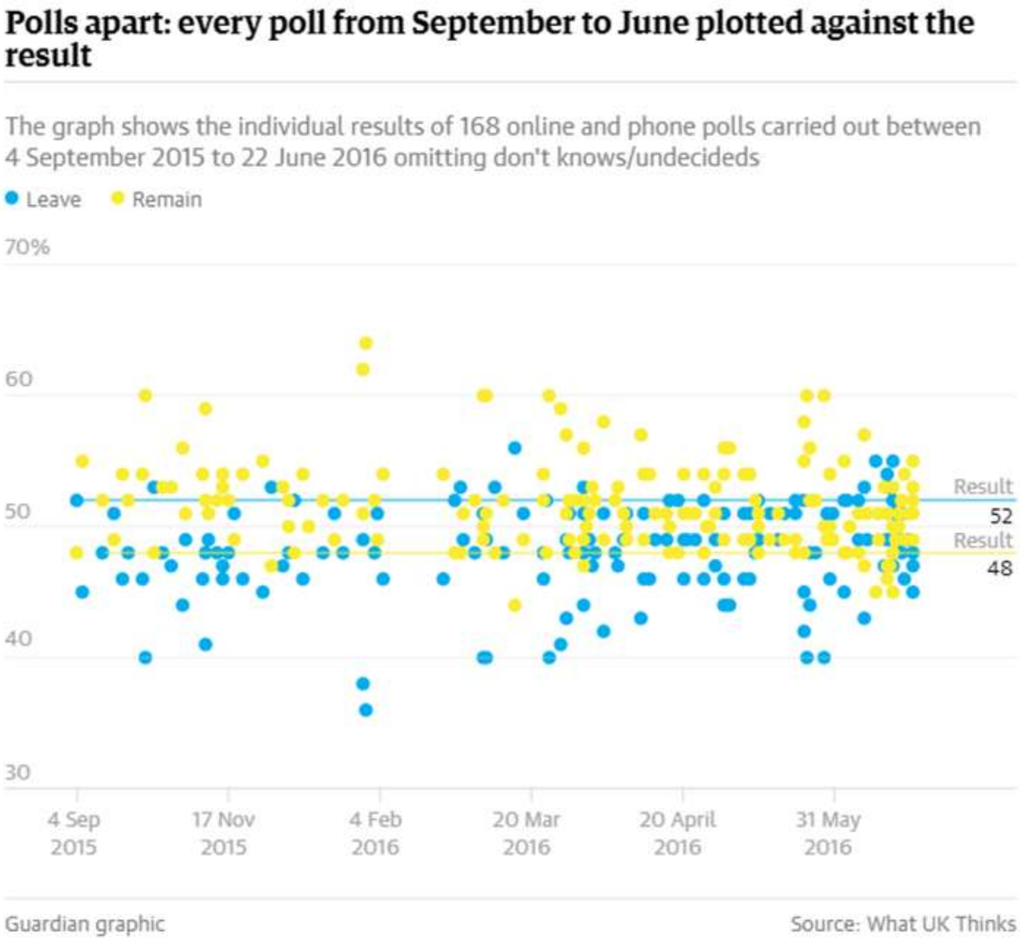
The graph shows the individual results of 168 online and phone polls carried out between 4 September 2015 to 22 June 2016 omitting don't knows/undecideds

Guardian graphic

Source: What UK Thinks

DATA VISUALIZATION REVISION

Do you understand this graph? What is wrong with it?



Polls apart: every poll from September to June plotted against the result

The graph shows the individual results of 168 online and phone polls carried out between 4 September 2015 to 22 June 2016 omitting don't knows/undecideds

Guardian graphic

Source: What UK Thinks

Curtin College

PAST EXAM QUESTION

A recent study investigated the relationship between energy drink consumption and academic performance. The researchers concluded that increased energy drink consumption is associated with a lower Grade Point Average (GPA is a measure of academic performance) amongst undergraduate students at a significance of 1%. sample

Data was collected by a voluntary online survey on social media. 848 of the survey participants claimed they were undergraduate students and provided data on their: age, gender, GPA and energy drink consumption in the last month.

The data was analysed using powerful statistical software. sample

- Use your knowledge of statistics to critique the data collection method used.
- Recommend one way in which the data collection could be improved.

1. online survey by social media: which is not a reliable source. You can't believe what people tell you. -> The data is not reliable.
2. dont use social media. if we need the GPA --> need to envolve the university.
if we need the drink consumption --> drink diary.

SAMPLING DISTRIBUTION

The distribution of a statistic is the Sampling Distribution

How would you ensure that you take an SRS of n=10 to determine the sample mean $\bar{x}$?

33	84	70	21	0	23	46	1	82	67
45	53	22	84	27	9	97	50	66	32
63	98	99	71	87	49	100	31	24	46
18	54	11	16	18	50	53	37	49	34
22	16	79	93	28	11	42	94	94	37

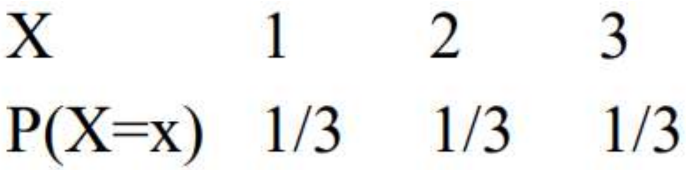
- Should your sample mean = the population mean?
- If you took 25 SRS of n=10 and then took the average what would the average of all those samples look like?

The average of the sample means =

mean of the (*sample*) means = $\frac{\sum_{i=1}^{i=25} \bar{x}_i}{25} \approx \mu$ = population mean.

WHAT IF YOU HAVE PROPORTIONS INSTEAD OF RAW DATA?

Assume a population consists of a large & equal number of 3 possible outcomes. You do not know the raw data. Calculate the mean and variance of the parent population.



X	1	2	3
P(X=x)	1/3	1/3	1/3

The mean = $E(X)$ = 1*1/3 + 2* 1/3 +3* 1/3 = 2

The variance= $E(X^2)$-$(E(X))^2$ = 1^2 * 1/3 + 2^2 * 1/3 + 3^2 * 1/3 – 2^2

= 4 – (1/3 + 4/3 + 9/3) = 4 – 14/3 = 2/3

This is a **symmetric** distribution with mean =2 and variance = 2/3

WHAT IF I TAKE SAMPLE SIZE OF 2?

You need to create a new probability table and redo the calcualtions.

The distribution of these means, $\bar{x}$, is the sampling distribution of the means when n=2.

	1 & 1	1 & 2, 2 &1	1 & 3, 3 &1, 2 & 2	2 & 3, 3 & 2	3 & 3
Average of sample =$\bar{X}$	1	1.5	2	2.5	3
$\boldsymbol{P(\bar{X})}$	$\frac{1}{3}*\frac{1}{3}=\frac{1}{9}$	$\frac{1}{3}*\frac{1}{3}+\frac{1}{3}*\frac{1}{3}=\frac{2}{9}$	$\frac{1}{3}*\frac{1}{3}+\frac{1}{3}*\frac{1}{3}+\frac{1}{3}*\frac{1}{3}=\frac{3}{9}$	$\frac{1}{3}*\frac{1}{3}+\frac{1}{3}*\frac{1}{3}=\frac{2}{9}$	$\frac{1}{3}*\frac{1}{3}=\frac{1}{9}$

$$E(\bar{X}) = 1*\frac{1}{9}+1.5*\frac{2}{9}+2*\frac{3}{9}+2.5*\frac{2}{9}+3*\frac{1}{9} = 2$$

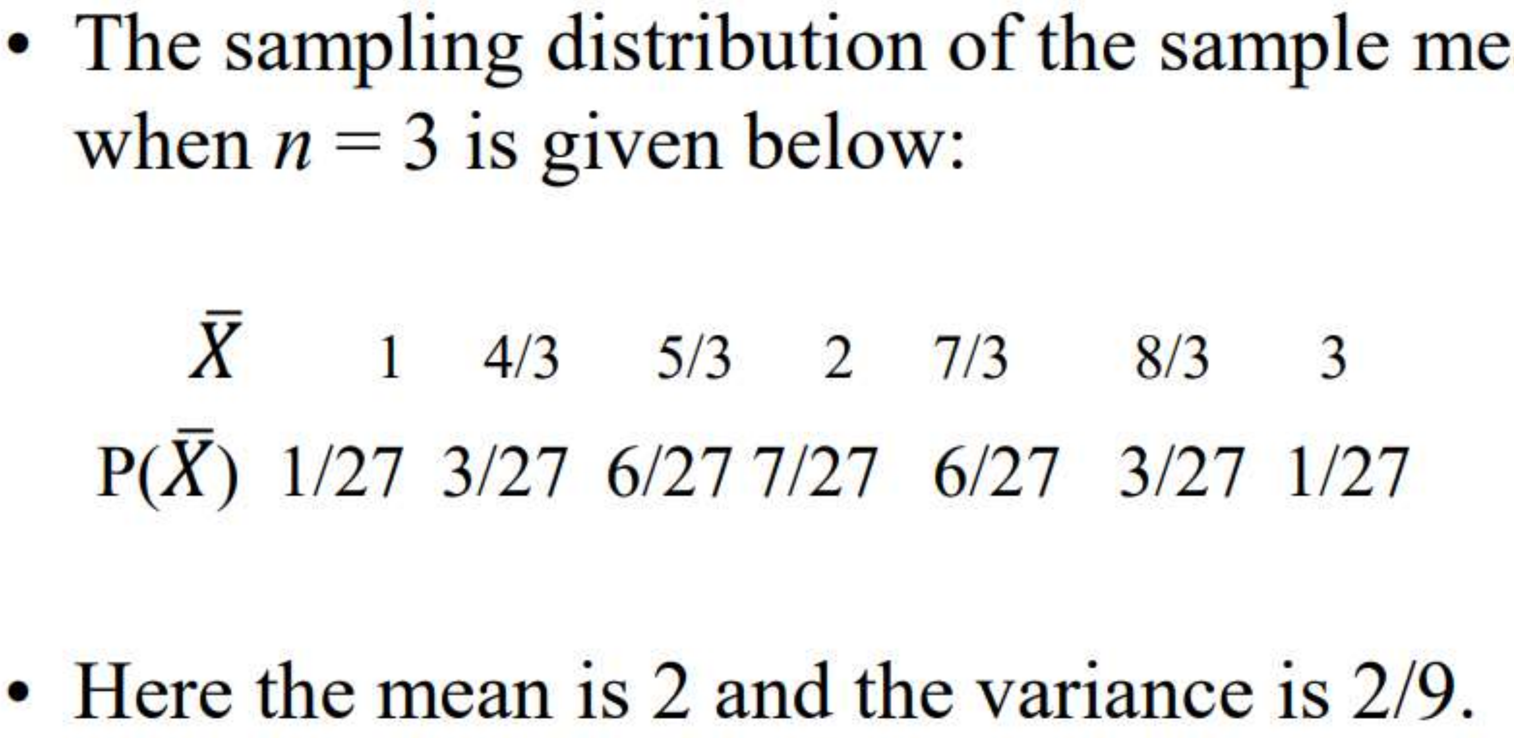
$$E(\bar{X}^2) = 1*1*\frac{1}{9}+1.5*1.5*\frac{2}{9}+2*2*\frac{3}{9}+2.5*2.5*\frac{2}{9}+3*3*\frac{1}{9} = \frac{13}{3}$$

$$Var(\bar{X}) = E(X^2)-(E(X))^2=\frac{13}{3}-4=\frac{1}{3}$$

E(X) is still 2 but E(X^2) changed, the variance of the sample means is halved.

- If n=3, expected value=2, variance of the sample means=2/9

CALCULATION FOR N=3



- The sampling distribution of the sample mean when *n* = 3 is given below:

$\bar{X}$	1	4/3	5/3	2	7/3	8/3	3
$P(\bar{X})$	1/27	3/27	6/27	7/27	6/27	3/27	1/27

- Here the mean is 2 and the variance is 2/9.

WARNING: SAMPLING DISTRIBUTION OF SAMPLE MEAN

The distribution of the sample means **does not** have to equal the distribution of the parent population.

In the example we saw

- The Var(X) depended on the sample size n.
- Var(X) of the sample means was less than the original variance

What does this mean for us?

- You can estimate the population mean from sample data.
- Sample means hide variation in the raw data. Do not assume you can determine the variance of a population from sample data.
- Be very careful about what you assume from calculated data.
- If in doubt pay a statistician!

MEAN AND STANDARD DEVIATION OF A SAMPLE MEAN

If $\bar{X}$ is the mean of a SRS of size n from a population having mean μ and standard deviation σ, then:

	sample	population
mean	$\bar{x}$	μ
std. dev	$\sigma/\sqrt{n}$	σ

- The mean of the means = the population mean

$$\mu_{\bar{X}} = \mu$$

- Variance of the means = the population variance / sample size

$$\sigma^2{}_{\bar{X}} = \frac{\sigma^2}{n}$$

- Standard deviation of the means =square root of variance of the means

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

NOTE: - The validity of these calculations depends on the “If…..

normal distribution

DISTRIBUTION OF $\overline{X}$

If a Normal population has $N(\mu,\sigma)$ distribution then the sample mean of n independent observations has distribution

$$N(\mu,\frac{\sigma^2}{n})$$

$$\boldsymbol{Z=\frac{(\overline{x}-\mu)}{\frac{\sigma}{\sqrt{n}}}\sim N(0,1)}$$

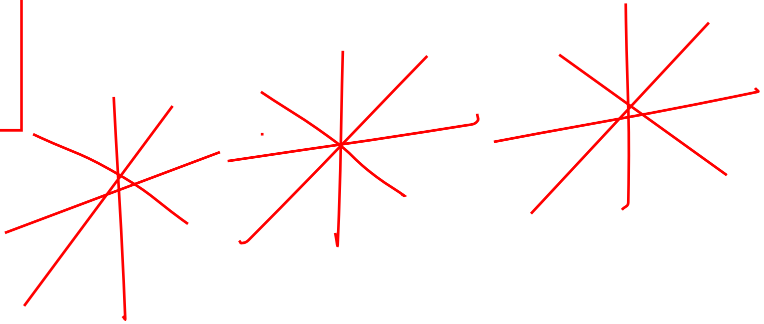
ZZZZZZ

Remember from week 2 that for a **POPULATION**

$$Z = \frac{(x - \mu)}{\sigma}$$

Then for our **SAMPLE MEANS** we can write

$$Z = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}} \sim \textit{and we can work with } N(0, 1)$$



FYI *NORMAL POPULATIONS SLIDE WEEK 2*

Remember

Have tables of values of Standardised Normal Distributions

When we standardise a normal distribution

- We shift the population mean so that the $\mu = 0$
- We create a standardised value for the σ which we call z

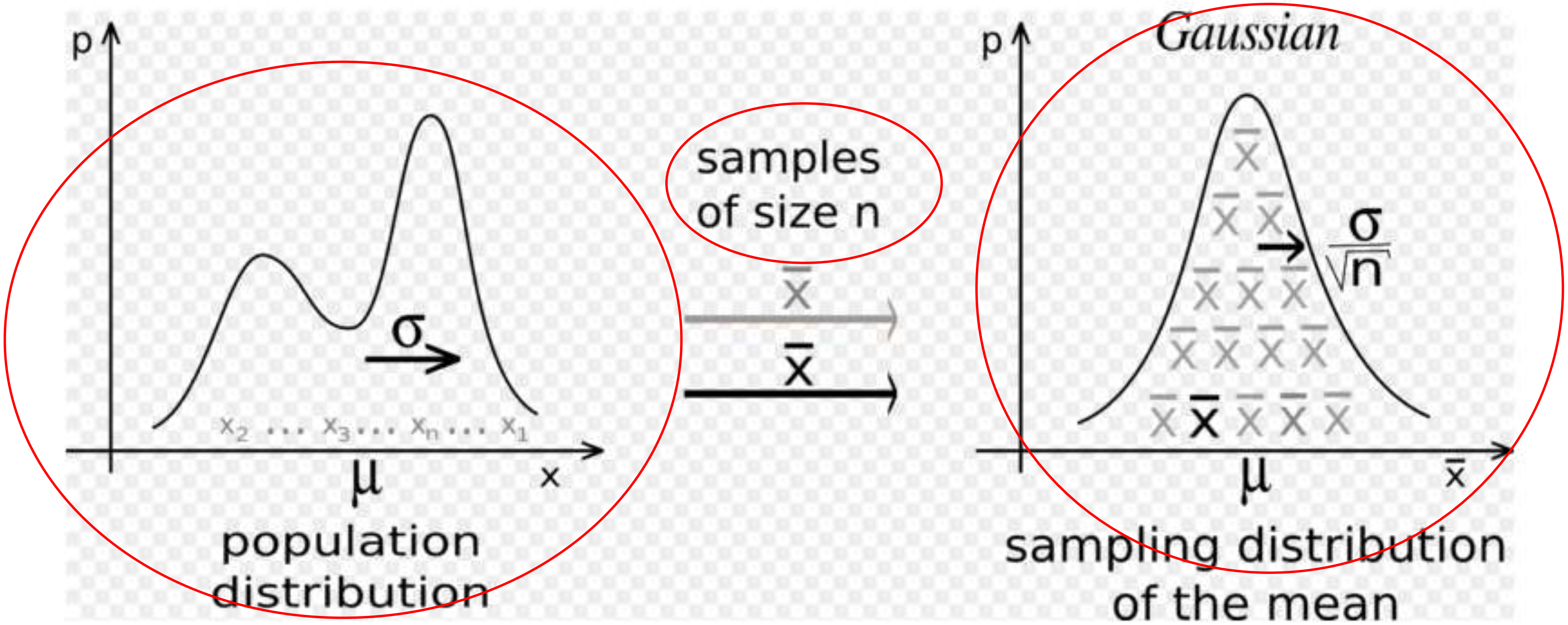
$$\boldsymbol{z = \frac{(X - \mu)}{\sigma}}$$

This is a VERY IMPORTANT EQUATION

- We started with $X \sim N(\mu, \sigma^2)$, and then standardised using $\boldsymbol{z = \frac{(X-\mu)}{\sigma}}$
- We now have $Z \sim N(0,1)$. The mean is zero and the variance= 1

CENTRAL LIMIT THEOREM

The distribution of the sample means approaches a normal distribution as the sample size increases regardless of the distribution of the underlying population (provided it has finite mean and variance)


p
σ
x_2 ... x_3 ... x_n ... x_1
μ
x
population distribution
samples of size n
$\bar{x}$
$\bar{x}$
Gaussian
p
$\frac{\sigma}{\sqrt{n}}$
$\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$ $\bar{x}$
μ
$\bar{x}$
sampling distribution of the mean

https://commons.wikimedia.org/wiki/File:IllustrationCentralTheorem.png#/media/File:IllustrationCentralTheorem.png

CLT IN OTHER WORDS

The Central Limit Theorem (*CLT* for short) basically says that for

- non-normal data,
- the distribution of the **sample means** has an **approximate normal distribution,**
- no matter what the distribution of the original data looks like,
- **as long** as the sample size is large enough and SRS
- and all samples have the same size
- and there is no bias.

Because statisticians know so much about the normal distribution, these analyses are much easier.

- https://www.dummies.com/education/math/statistics/how-the-central-limit-theorem-is-used-in-statistics/

EXAMPLE: COLA

$$\boxed{Z = \frac{X-\mu}{\sigma}}$$

A bottling company uses a filling machine to fill plastic bottles with a popular cola. The bottles are supposed to contain 300ml. In fact the contents vary according to a normal distribution with a mean of 298ml and standard deviation 3ml.

population
$\mu = 298\text{ml}$
$\sigma = 3\text{ml}$

(a) What is the probability that an individual bottle contains less than 295ml? $P(X < 295) = P\left(Z < \frac{295-298}{3}\right)$

$= P(Z < -1) = 0.1587$

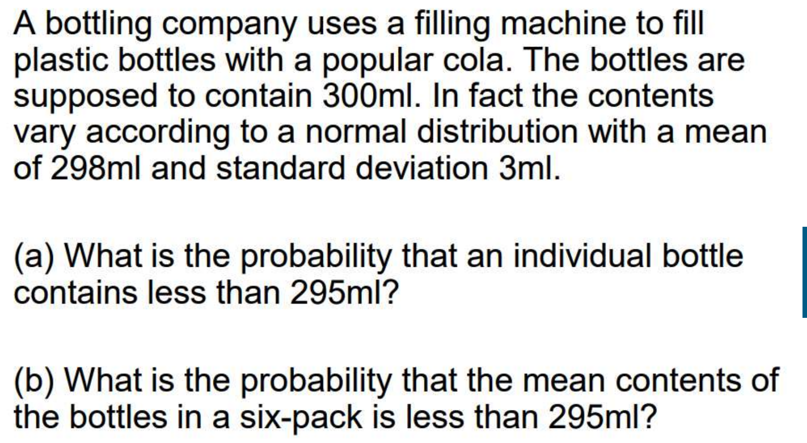
$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

(b) What is the probability that the mean contents of the bottles in a six-pack is less than 295ml? $P(\bar{X} < 295) = P\left(Z < \frac{295-298}{3/\sqrt{6}}\right)$

Sample $n=6$

$= P(Z < -2.45)$

$= 0.0071$



A bottling company uses a filling machine to fill plastic bottles with a popular cola. The bottles are supposed to contain 300ml. In fact the contents vary according to a normal distribution with a mean of 298ml and standard deviation 3ml.

(a) What is the probability that an individual bottle contains less than 295ml?

(b) What is the probability that the mean contents of the bottles in a six-pack is less than 295ml?

What do we know about the population? N(…..)

(a)

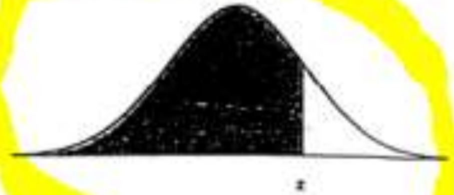
- We are calculating for an individual bottle
- Do a quick sketch. Should your final answer be < or> 0.5?
- Calculate the z statistic
- Use tables to calculate probability

(b)

- What do we know about sample size n=?
- This time we are calculating the probability that the sample mean < 295 for a six pack. What formula do we use for Z?

College

CUMULATIVE PROBABILITIES FOR THE STANDARD NORMAL DISTRIBUTION



$P(Z \le z)$ where $Z \sim N(0,1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879

CALCULATION

a) $P(X < 295) = P\left[Z < \frac{295-298}{3}\right]$

$= P(Z < -1) = .1587$

b) Here we are dealing with $\bar{X}$ where $n = 6$.

$$P(\bar{X} < 295) = P\left[Z < \frac{295-298}{3/\sqrt{6}}\right]$$

$$= P(Z < -2.45)$$

$$= .0071$$

DECREASE IN VARIABILITY AS N INCREASES

- There is a 16% chance that an individual bottle contains <295ml
- There is a 0.7% chance that the average of a six-pack < 295ml
- If we asked the same question about a 12 pack
  - Z~-3.46, P(Z<-3.46) ~ 0.0003 ~.03%

As the sample size n ↑, the variability of the sample means ↓ and the values get closer and closer to the population mean

The probability that the average of the sample is far below or above the mean will decrease as n increases,

COAL DUST

A laboratory weighs filters from a coal mine to measure the amount of dust in the mine atmosphere. Repeated measurements of the weight of the dust on the same filter vary normally with standard deviation of 0.08mg because the weighing is not perfectly precise. The dust on a particular filter actually weighs 123mg. Repeated weightings will then have the normal distribution with mean 123mg and standard deviation 0.08mg.

population

$\sigma = 0.08\text{mg}$

$\mu = 123 \text{ mg}$

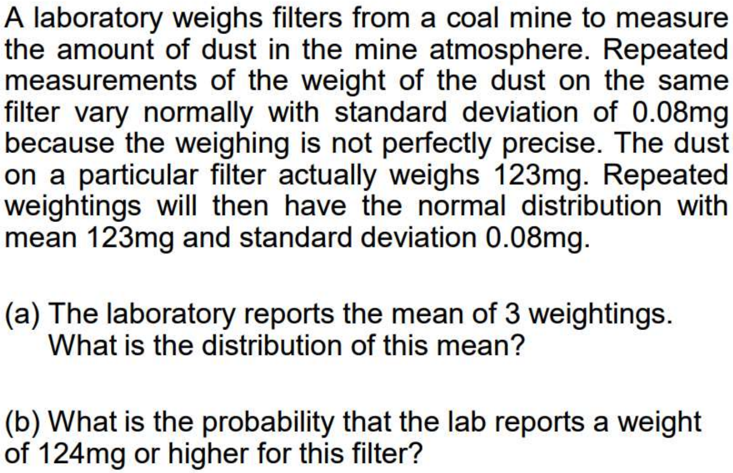
n = 3

(a) The laboratory reports the mean of 3 weightings. What is the distribution of this mean? normal distribution

n = 3

(b) What is the probability that the lab reports a weight of 124mg or higher for this filter?

COAL DUST CONTINUED



A laboratory weighs filters from a coal mine to measure the amount of dust in the mine atmosphere. Repeated measurements of the weight of the dust on the same filter vary normally with standard deviation of 0.08mg because the weighing is not perfectly precise. The dust on a particular filter actually weighs 123mg. Repeated weightings will then have the normal distribution with mean 123mg and standard deviation 0.08mg.

(a) The laboratory reports the mean of 3 weightings.
What is the distribution of this mean?

(b) What is the probability that the lab reports a weight of 124mg or higher for this filter?

(a)

What sort of distribution?

What is the population mean and standard deviation?

What is n?

N(,)

(b)

Sketch

What is 124?

Calculate z

Determine P(Z>)

Solution

normal distribution $\bar{x} = \mu = 123$

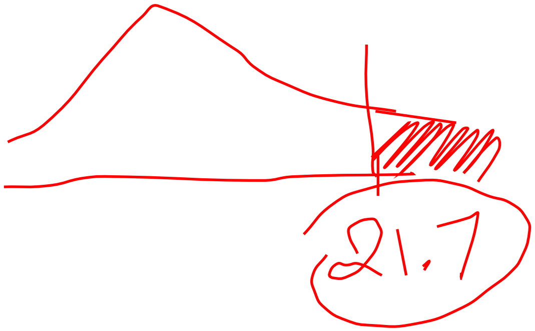
a) $N\left(123, \frac{.08}{\sqrt{3}}\right)$ $\frac{\sigma}{\sqrt{n}} = \frac{0.08}{\sqrt{3}} = 0.04618$

$n = 3$

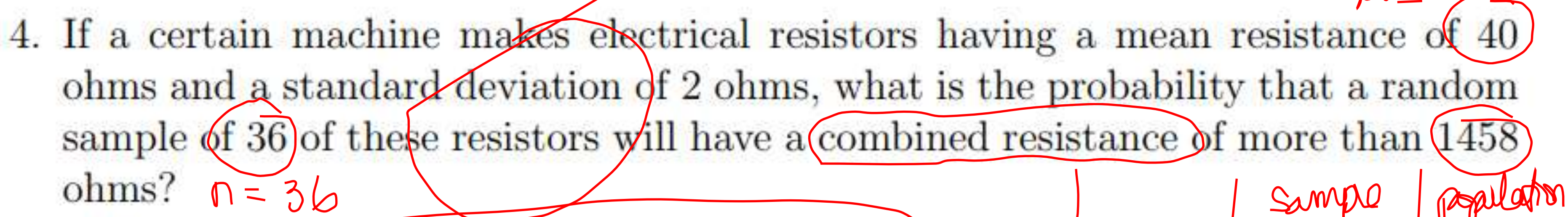
$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

b) $P(\bar{x} > 124) = P\left[Z > \frac{124-123}{.08/\sqrt{3}}\right]$

$= P(Z > 21.7) = 0$



RESISTORS



4. If a certain machine makes electrical resistors having a mean resistance of 40 ohms and a standard deviation of 2 ohms, what is the probability that a random sample of 36 of these resistors will have a combined resistance of more than 1458 ohms?

$\mu =$

$n = 36$

What sort of distribution? Normal distribution

	sample	population
mean	$\bar{x}$ =	μ
std. dev	$\frac{\sigma}{\sqrt{n}}$	σ

Population parameters or sample statistics?

$\mu = 40\ ohms$ and $\sigma = 2\ ohms, n = 36, \bar{x} = ???, \sum x = 1458$

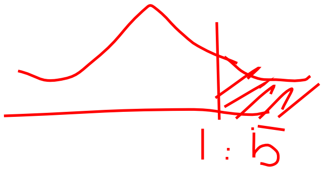
So…

$$\boxed{z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}$$

$$\bar{x} = \frac{1458}{36} = 40.5\ ohms$$

$$P(\bar{x} > 40.5) = P\left(z > \frac{40.5 - 40}{2/\sqrt{36}}\right) = P(z > 1.5)$$

You finish!



$= 1 - P(z < 1.5)$

$= 1 - 0.9332 = 0.0668$

WHICH DISTRIBUTION IS IT?

How to select a distribution for a given problem?

Do NOT

- Assume that as n>30 you can use a Normal Distribution.
- Assume that the data is quantitative and continuous just because there are number.

Do USE

- Theoretical arguments
- Some of the methods and tests mentioned earlier to explore the distribution.
- Ask a statistician ☺

THE PURPOSE OF STATISTICAL INFERENCE

…is to draw conclusions from the data

Most of the time you will find that you are using SAMPLE data to infer something about the POPULATION

The methods you can use are based on sampling distributions and depend on what you know. **For the following section we assume we have:**

- A normal population
- Been given σ of population. If we don't know this we have to use a different distribution.
- μ unknown and is the population parameter of interest.

ESTIMATION

When we use statistics (determined from sample data) to assign a value to a parameter (population) we are estimating the population parameter.

We are not absolutely certain what the population parameter is.

How could we be **100%** certain?

Measuring everything is not always possible or even logical to do.

Estimation method

1. Select a sample
2. Collect the required information from the members of the sample
3. Calculate the value of the sample statistic
4. Assign values to the corresponding population parameter

Sounds easy??

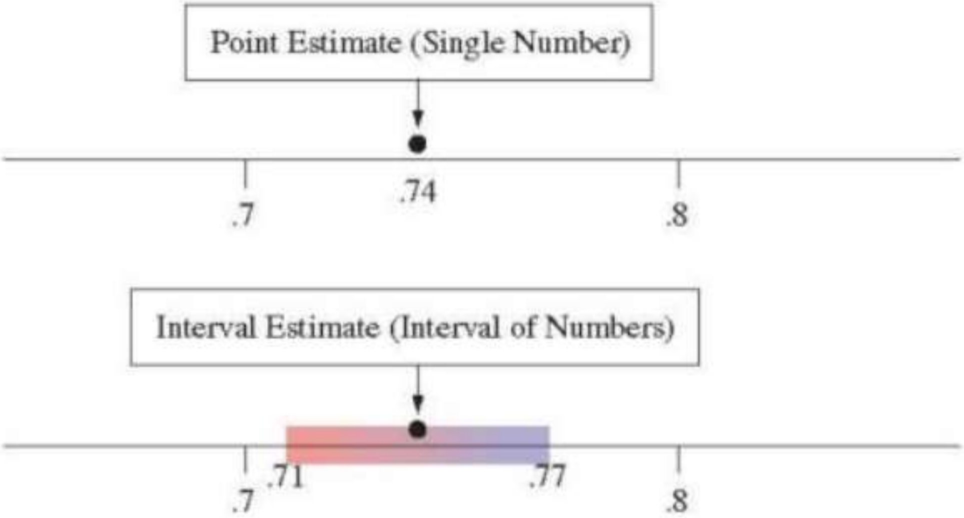
POINT AND INTERVAL ESTIMATES

The Point Estimate

- of a population parameter is a single value used to estimate the population parameter.
- The sample mean $\bar{x}$ is a **point estimate** of the population mean μ

The Interval Estimate

- Is defined by 2 numbers, between which a population parameter could lie
- $a < \mu < b$ is an interval estimate for the population mean μ


Point Estimate (Single Number)
.7
.74
.8
Interval Estimate (Interval of Numbers)
.7
.71
.77
.8

- https://stattrek.com/statistics/dictionary.aspx?definition=point_estimate
- https://stattrek.com/statistics/dictionary.aspx?definition=point_estimate
- https://www.slideserve.com/jermaine-huffman/chapter-7-statistical-inference-confidence-intervals

CONFIDENCE INTERVALS

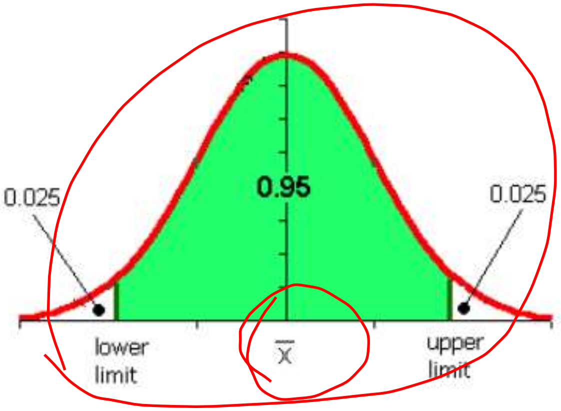
Statisticians use confidence interval to express the degree of uncertainty associated with a sample statistic.

A statistician might state they have used a “95% confidence interval”

What does this mean?

- If the statistician used the same sampling method to select different samples
- computed an interval estimate for each sample,
- they would expect the true population parameter to fall within the interval estimate 95% of the time.

95% C.I : 2.1 → 2.3



CONFIDENCE INTERVALS

Whilst you could have a 3% confidence interval, why would you?

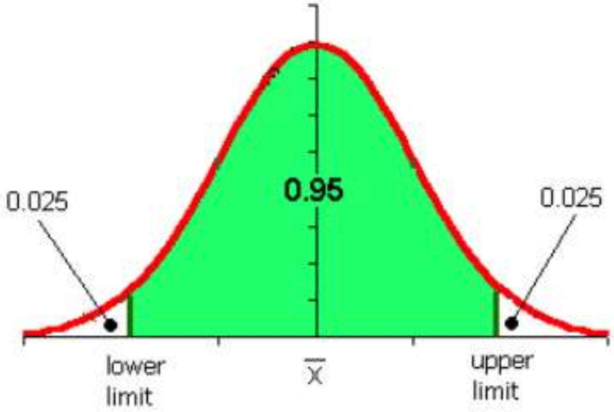
The usual confidence intervals are: 90%, 95% and 99%

Remember we are working with normal distributions

If we use a 95 % confidence interval for a **STANDARD** NORMAL DISTRIBUTION

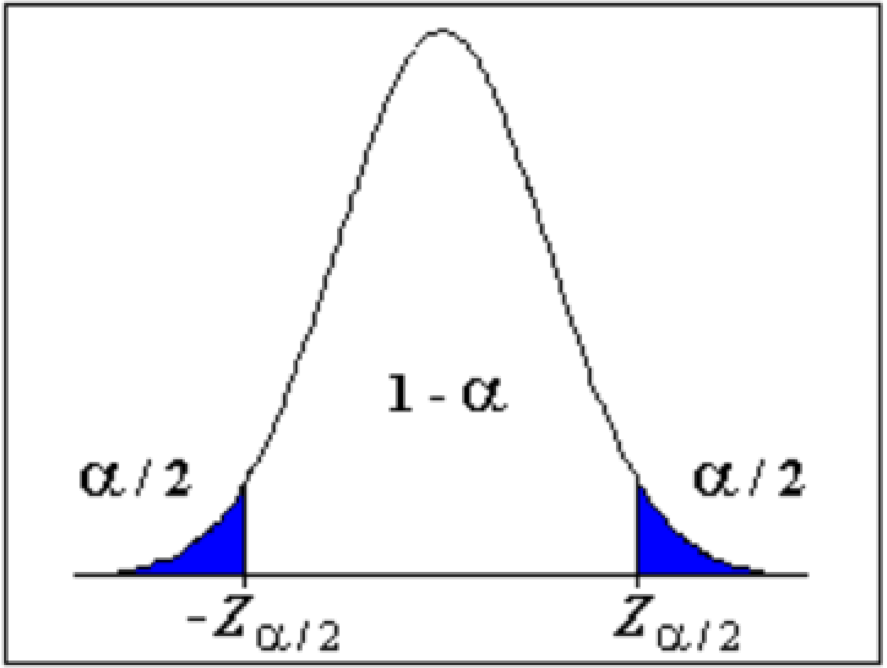
- The green zone is 0.95
- The white tails are 0.025 each
- $0.025 + 0.025 = 0.05 = \alpha$
- You can be told $\boldsymbol{\alpha = 0.05}$ rather than use 95%
- The total area under the curve=1
- If you refer to tables for N(0,1)
- If P(Z< ?)=0.025, then Z=-1.96

- https://onlinecourses.science.psu.edu/stat500/node/30



CONFIDENCE INTERVALS FOR STANDARD NORMAL DISTRIBUTIONS

Confidence	α	$\frac{\alpha}{2}$	z
99%	0.01	0.005	± 2.575
95%	0.05	0.025	± 1.96
90%	0.1	0.05	± 1.645

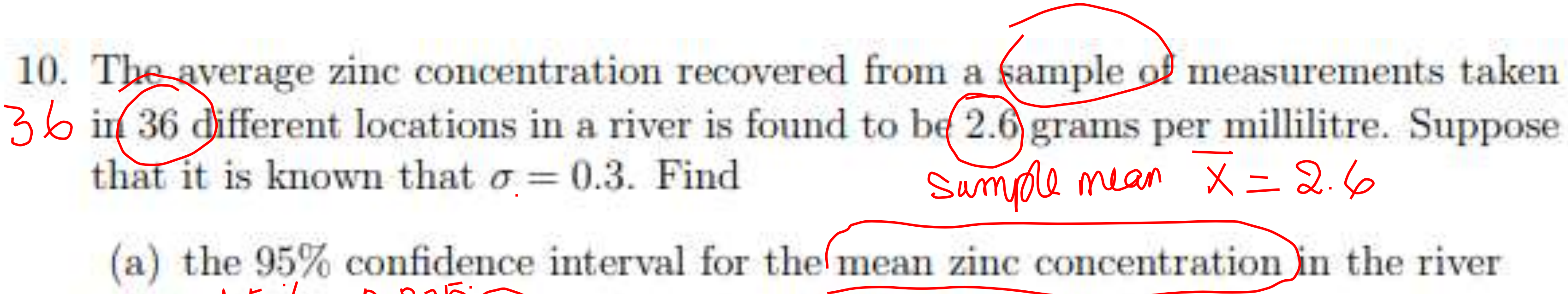


CONFIDENCE INTERVAL OF $\bar{X}$

Remember that the sampling distribution of the mean is

$$N(\mu, \frac{\sigma^2}{n})$$

If 95% of the time $\bar{X}$ will be within $\pm 1.96 \frac{\sigma}{\sqrt{n}}$ of μ

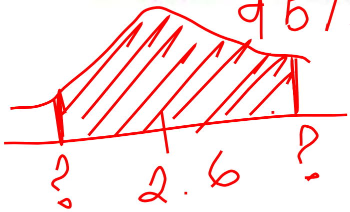


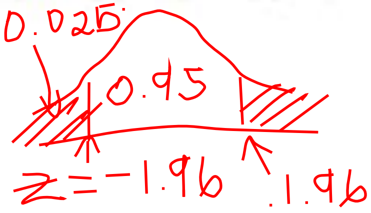
10. The average zinc concentration recovered from a sample of measurements taken in 36 different locations in a river is found to be 2.6 grams per millilitre. Suppose that it is known that $\sigma = 0.3$. Find

(a) the 95% confidence interval for the mean zinc concentration in the river

$n = 36$

sample mean $\bar{X} = 2.6$





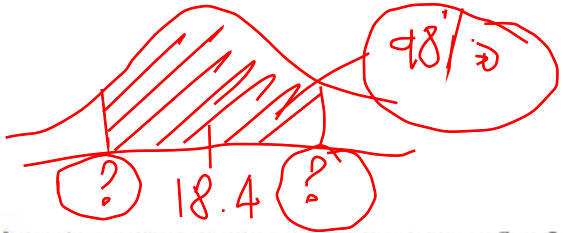
$\frac{1-0.95}{2} = 0.025$

$95\% \text{ CI} = \bar{x} \pm \frac{z\sigma}{\sqrt{n}} = 2.6 \pm \frac{1.96 \times 0.3}{\sqrt{36}}$

$= 2.6 \pm 0.098$

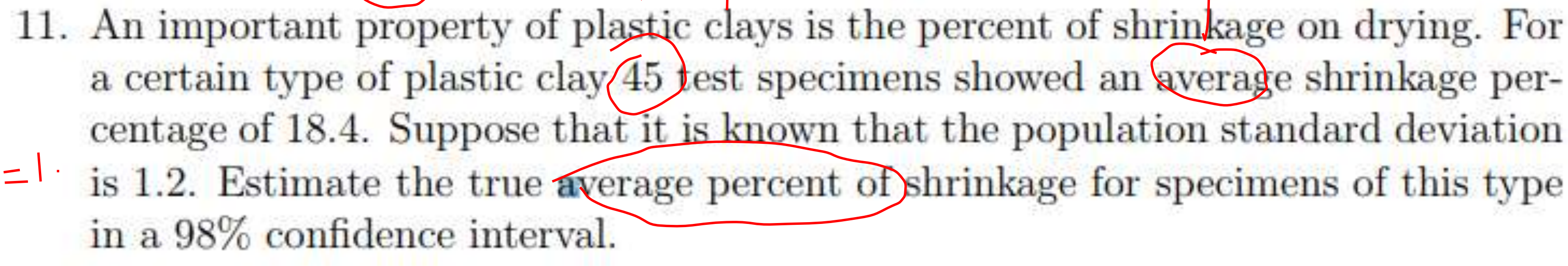
$[2.502, 2.698]$

WORKSHOP QUESTION 11



$n = 45$

sample mean $\bar{x} = 18.4$

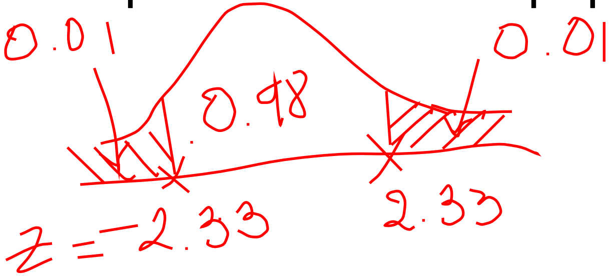


11. An important property of plastic clays is the percent of shrinkage on drying. For a certain type of plastic clay 45 test specimens showed an average shrinkage percentage of 18.4. Suppose that it is known that the population standard deviation is 1.2. Estimate the true average percent of shrinkage for specimens of this type in a 98% confidence interval.

$\sigma = 1.$

How certain are we that the population mean lies in this interval?

Is it possible that the population mean lies outside this interval?



$$\frac{1-0.98}{2} = 0.01$$

$z \to x$?

98% CI

$$= \bar{x} \pm \frac{z\sigma}{\sqrt{n}}$$

$$= 18.4 \pm \frac{2.33 \times 1.2}{\sqrt{45}}$$

$$= 18.4 \pm 0.417$$

98% CI = [18.4 − 0.417 , 18.4 + 0.417]
= [17.983 , 18.817]

GENERAL CONFIDENCE INTERVAL FOR μ

Whilst you might prefer saying “I am 95% confident that…” statisticians often use $\boldsymbol{\alpha}$.

You need to be able to use $\boldsymbol{\alpha}$

- **The $\mathbf{100(1-\alpha)\%}$ confidence interval for $\boldsymbol{\mu}$ when $\boldsymbol{\sigma}$ is known is**

$$\bar{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$$

- The value of $z_{\alpha/2}$ depends on your confidence level
- The interval is in the form

$$\boldsymbol{estimate \pm margin\ of\ error}$$

So if I ask for the margin of error it is the $\pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$

CONFIDENCE INTERVAL σ IS KNOWN

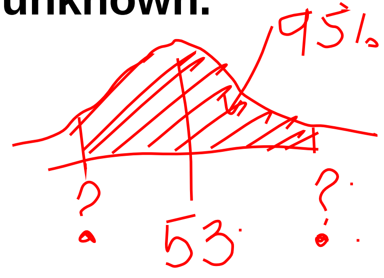
If σ is known can use normal distribution tables, even though μ is unknown.

$n = 10$

A sample of **10** Mars Bars* has been drawn from a large box of Mars Bars. The population standard deviation of the population is 2.7 grams. The sample mean is 53 grams.

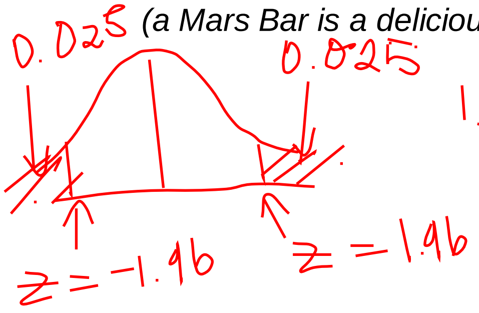
σ

$\bar{x}$



Determine the 95% confidence interval for the population mean.

(a Mars Bar is a delicious chocolate bar first produced in 1932 by Mr Mars)



$$\frac{1-0.95}{2} = 0.025$$

95% CI

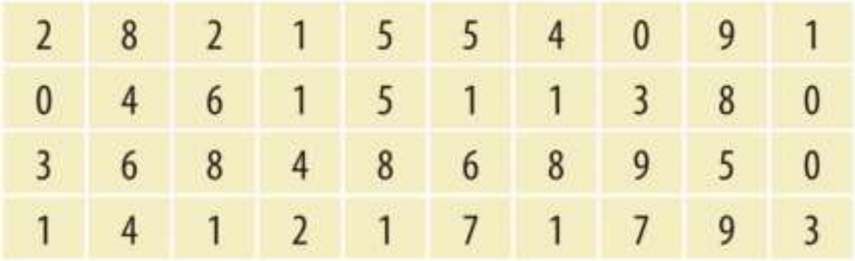
$$= \bar{x} \pm \frac{z\sigma}{\sqrt{n}} = 53 \pm \frac{1.96 \times 2.7}{\sqrt{10}}$$

$$= 53 \pm 1.67 \text{ g}$$

95% CI = [51.33 , 54.67]

RANDOM SAMPLE OF SINGLE DIGIT NUMBERS

Calculate the 90% confidence interval



2	8	2	1	5	5	4	0	9	1
0	4	6	1	5	1	1	3	8	0
3	6	8	4	8	6	8	9	5	0
1	4	1	2	1	7	1	7	9	3

Plug into calculator:

- $n = 40, \sum x = 159, \bar{x} = 3.975\,, \sigma = 2.87$
- This is assuming $2.87 = \ \sigma$ the pop sd.
  - If this assumption is true we can use a normal distribution.
  - If you are not confident that $\sigma = 2.87$ then you would use a different distribution.
- 90% confidence -> $z = \pm 1.65$
- $\boldsymbol{\mu = \overline{x} \pm 1.65 \frac{\sigma}{\sqrt{n}} = 3.975 \pm 1.65 \frac{2.87}{\sqrt{40}} = 3.975 \pm 0.749}$

Answer

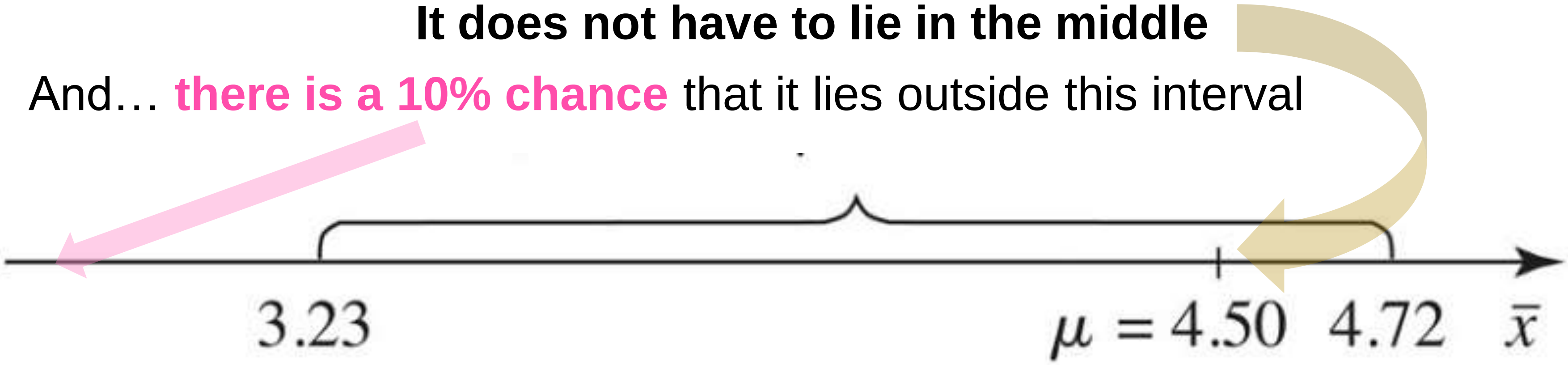
- The 90% confidence interval for μ is (3.23,4.72)

WHERE DOES μ LIE?

For a 90% confidence interval we think μ lies somewhere in this interval

It does not have to lie in the middle

And… **there is a 10% chance** that it lies outside this interval



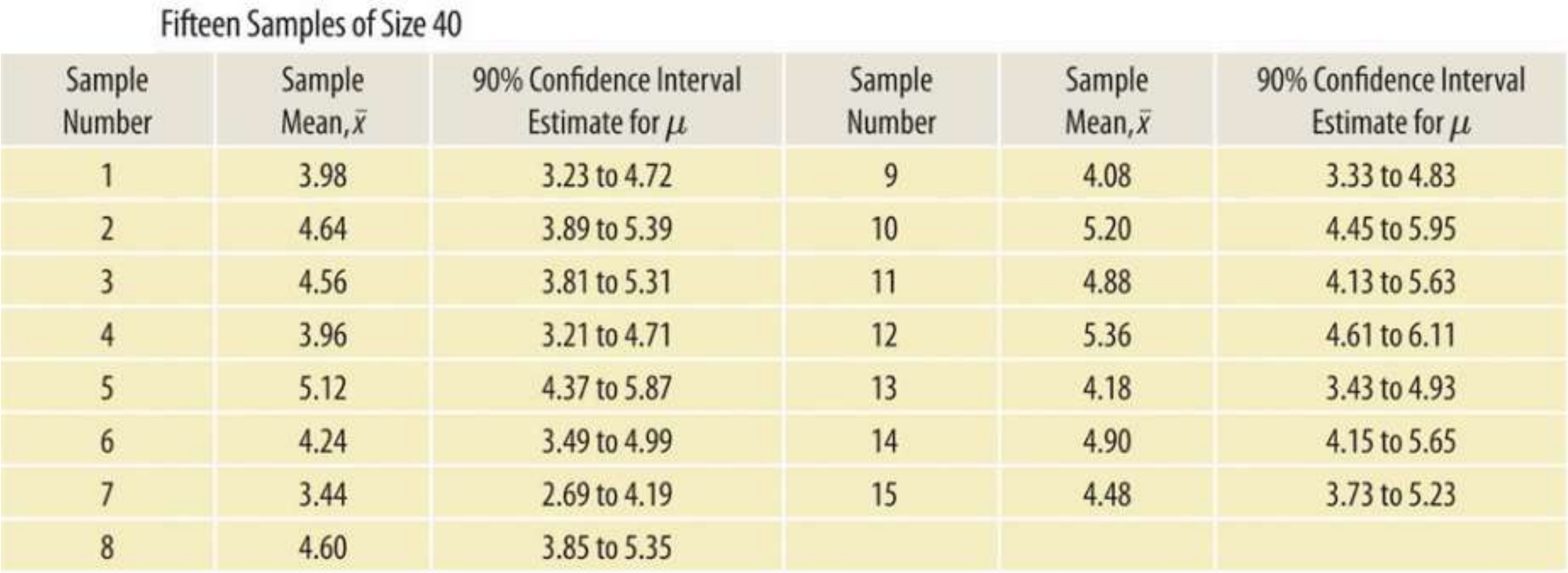
When we use a sampling distribution to estimate a population parameter the answer is an interval. The 90% confidence interval for μ is (3.23,4.72)

WHAT IF I DID THE SRS N=40, 15 TIMES?

The sample mean $\overline{x}_i$ is different each time.

The average of all the sample means $= 67.2/15 = 4.51.$

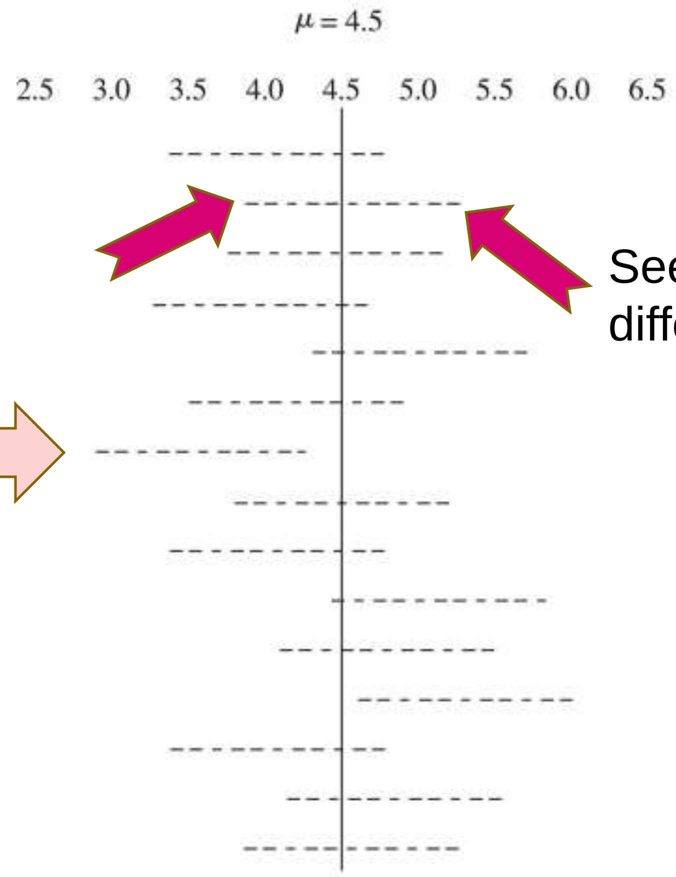
What do you think the population mean is?



Fifteen Samples of Size 40

Sample Number	Sample Mean, $\bar{x}$	90% Confidence Interval Estimate for μ	Sample Number	Sample Mean, $\bar{x}$	90% Confidence Interval Estimate for μ
1	3.98	3.23 to 4.72	9	4.08	3.33 to 4.83
2	4.64	3.89 to 5.39	10	5.20	4.45 to 5.95
3	4.56	3.81 to 5.31	11	4.88	4.13 to 5.63
4	3.96	3.21 to 4.71	12	5.36	4.61 to 6.11
5	5.12	4.37 to 5.87	13	4.18	3.43 to 4.93
6	4.24	3.49 to 4.99	14	4.90	4.15 to 5.65
7	3.44	2.69 to 4.19	15	4.48	3.73 to 5.23
8	4.60	3.85 to 5.35			

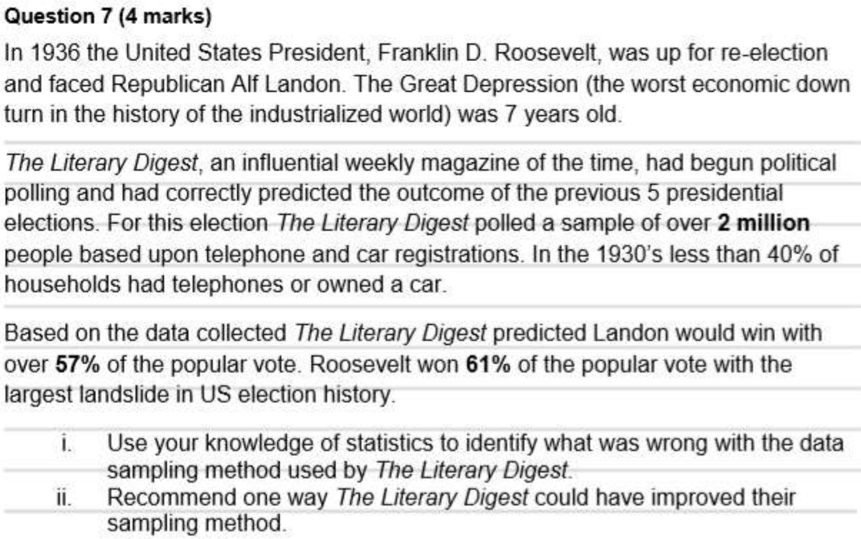
PLOT OF CONFIDENCE INTERVALS FOR EACH SAMPLE



See how each sample has a different upper & lower value

This confidence interval does not even include μ

BIAS PAST EXAM QUESTIONS



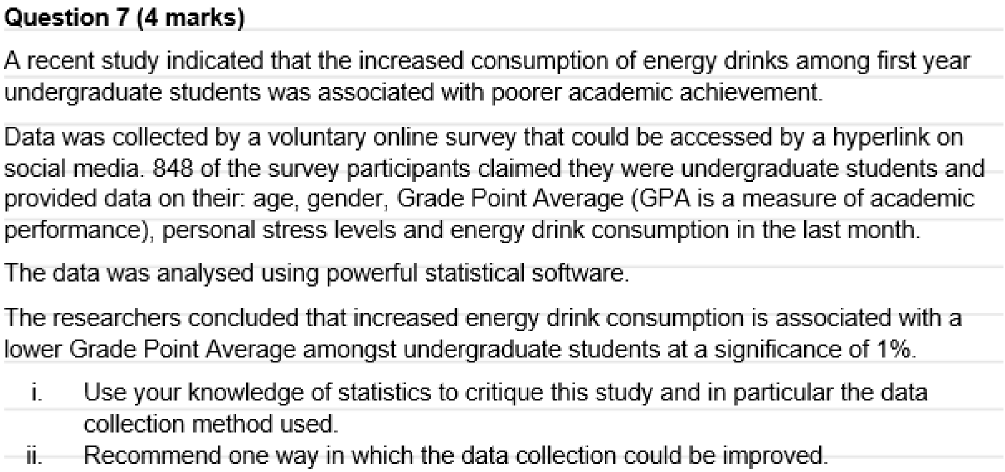
Question 7 (4 marks)

In 1936 the United States President, Franklin D. Roosevelt, was up for re-election and faced Republican Alf Landon. The Great Depression (the worst economic down turn in the history of the industrialized world) was 7 years old.

The Literary Digest, an influential weekly magazine of the time, had begun political polling and had correctly predicted the outcome of the previous 5 presidential elections. For this election *The Literary Digest* polled a sample of over **2 million** people based upon telephone and car registrations. In the 1930's less than 40% of households had telephones or owned a car.

Based on the data collected *The Literary Digest* predicted Landon would win with over **57%** of the popular vote. Roosevelt won **61%** of the popular vote with the largest landslide in US election history.

i. Use your knowledge of statistics to identify what was wrong with the data sampling method used by *The Literary Digest*.
ii. Recommend one way *The Literary Digest* could have improved their sampling method.



Question 7 (4 marks)

A recent study indicated that the increased consumption of energy drinks among first year undergraduate students was associated with poorer academic achievement.

Data was collected by a voluntary online survey that could be accessed by a hyperlink on social media. 848 of the survey participants claimed they were undergraduate students and provided data on their: age, gender, Grade Point Average (GPA is a measure of academic performance), personal stress levels and energy drink consumption in the last month.

The data was analysed using powerful statistical software.

The researchers concluded that increased energy drink consumption is associated with a lower Grade Point Average amongst undergraduate students at a significance of 1%.

i. Use your knowledge of statistics to critique this study and in particular the data collection method used.
ii. Recommend one way in which the data collection could be improved.

IMPORTANT

You must be confident in calculating confidence intervals.

We need them for hypothesis testing in week 4.

This weeks tutorial questions have lots of sample and confidence interval questions.