

Advanced Data Science

Topic 11b – Part 5

1. What We'll Cover

This topic will introduce...

- **What is data science.**
- **Key concepts – the scientific method.**
- **Useful terminology.**
- **Important tools - Statistics.**
- **Data collection & Experiment Design.**
- **Probability basics.**
- **Data distributions.**
- **Hypothesis testing.**

} Part 5

The aim: to help you understand what it means to be a data scientist and to get you familiar with data science tools.

2. Hypothesis Testing

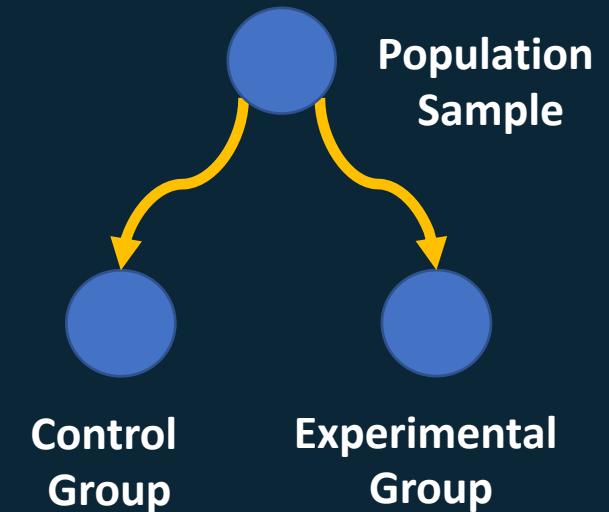
- Hypothesis testing is a statistical approach used to find the optimal answer to the questions we pose about the world around us.
- It uses available knowledge captured in data, to reach conclusions regarding hypotheses in a rigorous way.
- The method is useful when undertaking experimental studies.
- Suppose we are tasked with determining if a medicine works.
- We form hypotheses and split a sample population into control and experimental groups.
- We can use hypothesis testing to determine which of the hypotheses holds over the groups.
- That is, which has the most evidence in its favour.
- Here we are introducing the foundations of statistical inference central to data science and machine learning.

Null Hypothesis

$$H_0 = \text{No effect}$$

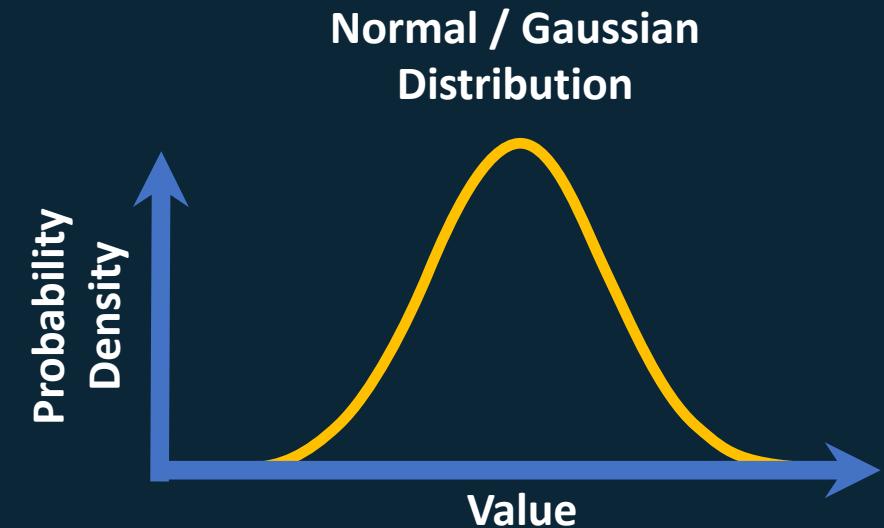
Alternative Hypothesis

$$H_a \text{ or } H_1 = \text{Effect}$$



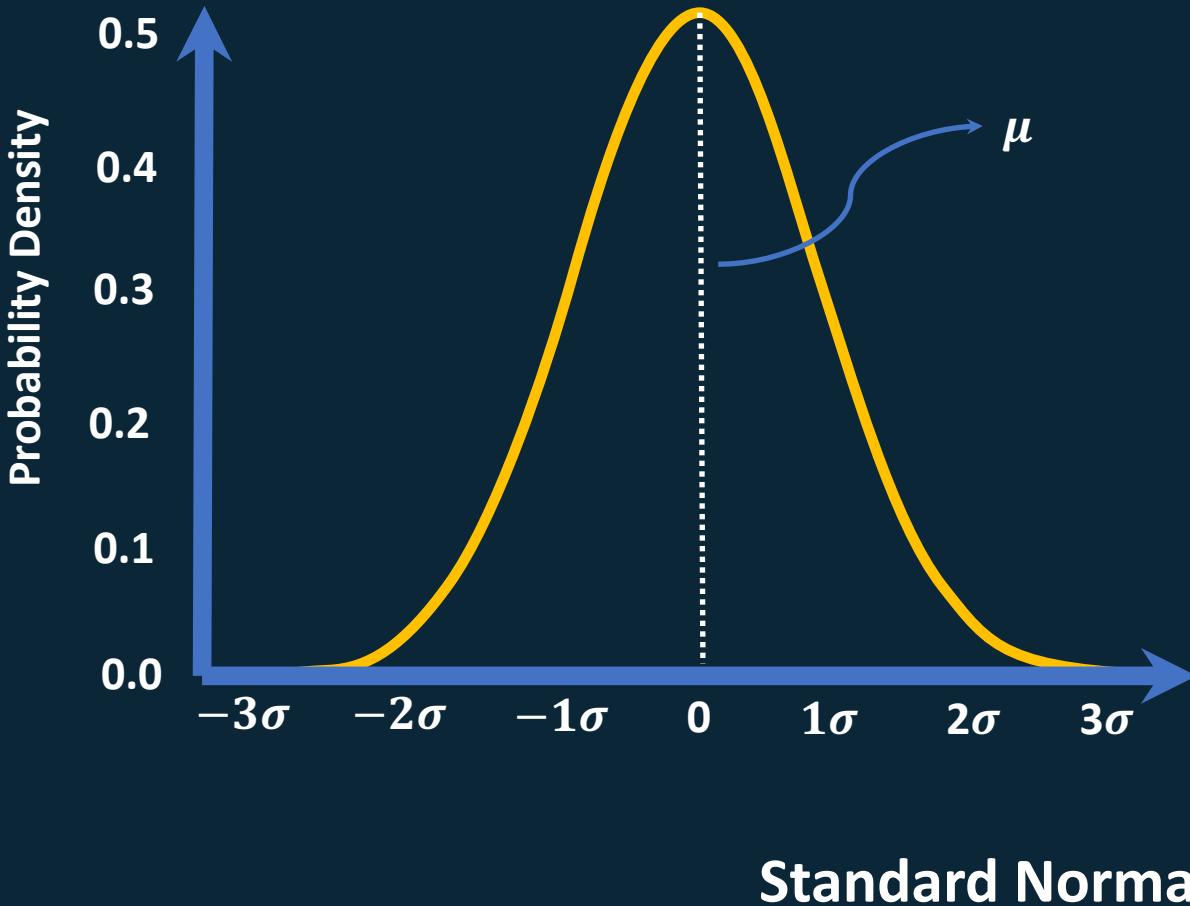
3. How this fits in?

- So far you've come across concepts from lot's of different areas.
- You've learned about probability theory, the different types of data distribution (unimodal, bi-modal, multi-modal), the law of large numbers, and how to compute summary statistics over samples of data, and entire populations.
- We covered this material to help prepare you for the concepts I'll very shortly introduce related to hypothesis testing. I'm sure you're relieved that none of this time was wasted!
- So with that in mind, lets return to thinking about a distribution I've mentioned a few times during this course – the normal distribution.

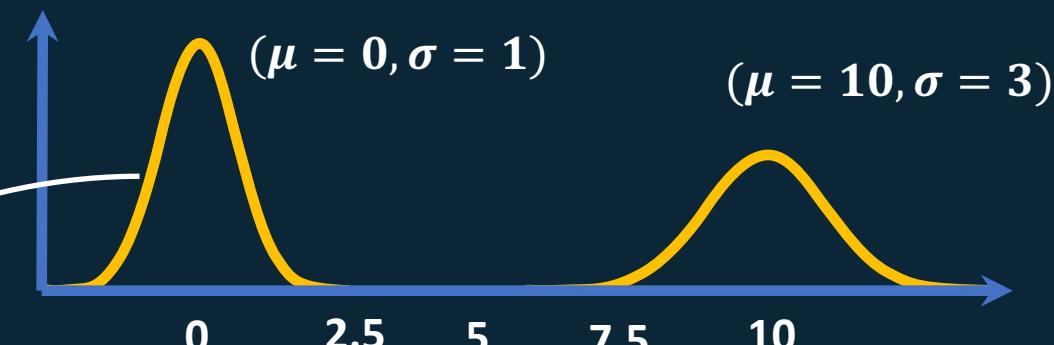


4. Back to Normal

Normal / Gaussian
Distribution

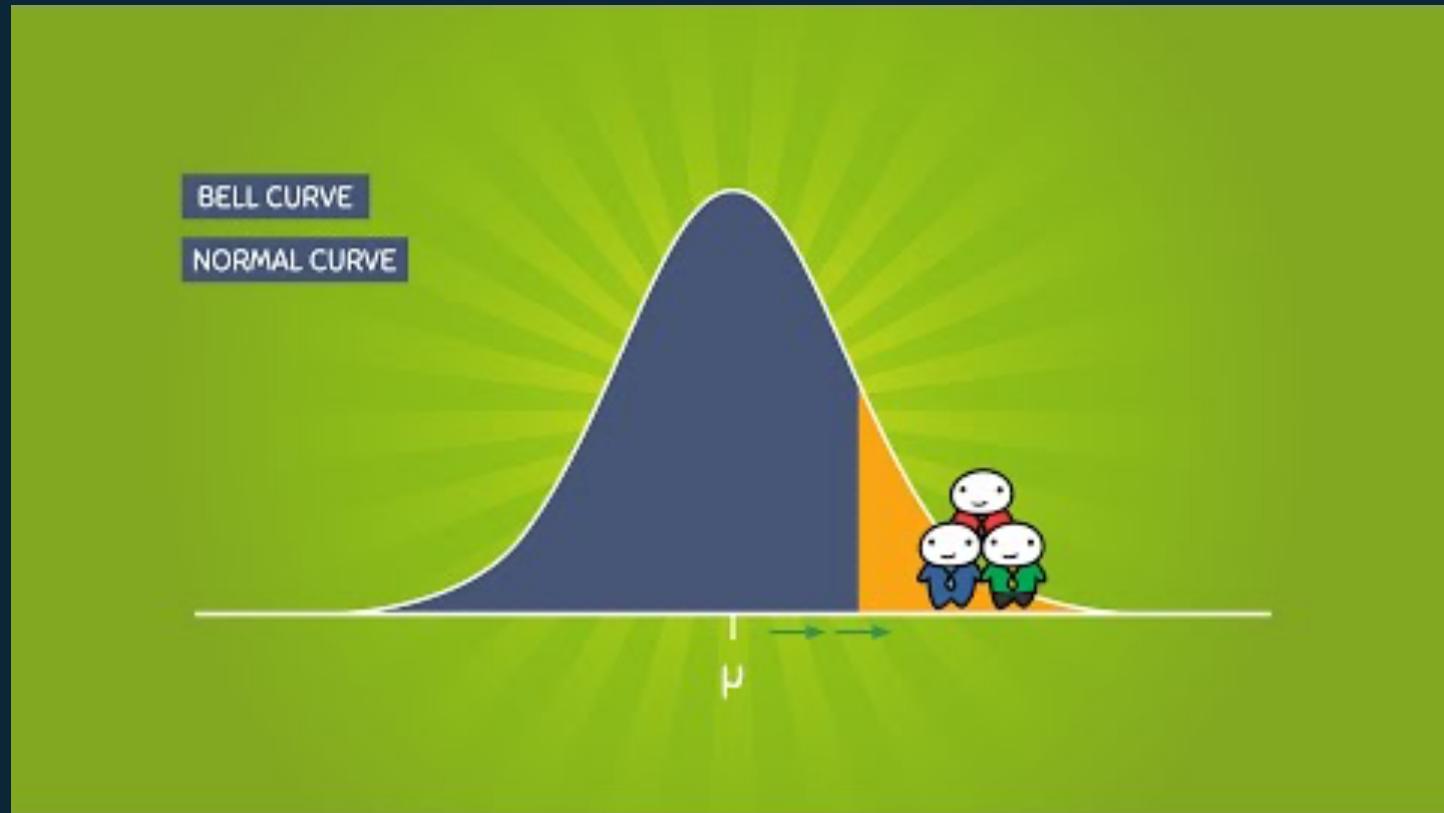


- The normal curve is always a symmetric, unimodal, bell-shaped curve.
- The shape of the curve is determined by two parameters.
 - The mean, μ .
 - The standard deviation, σ .
- We can describe any normal curve via a pair, e.g. $(\mu = 0, \sigma = 1)$.



Standard Normal Distribution

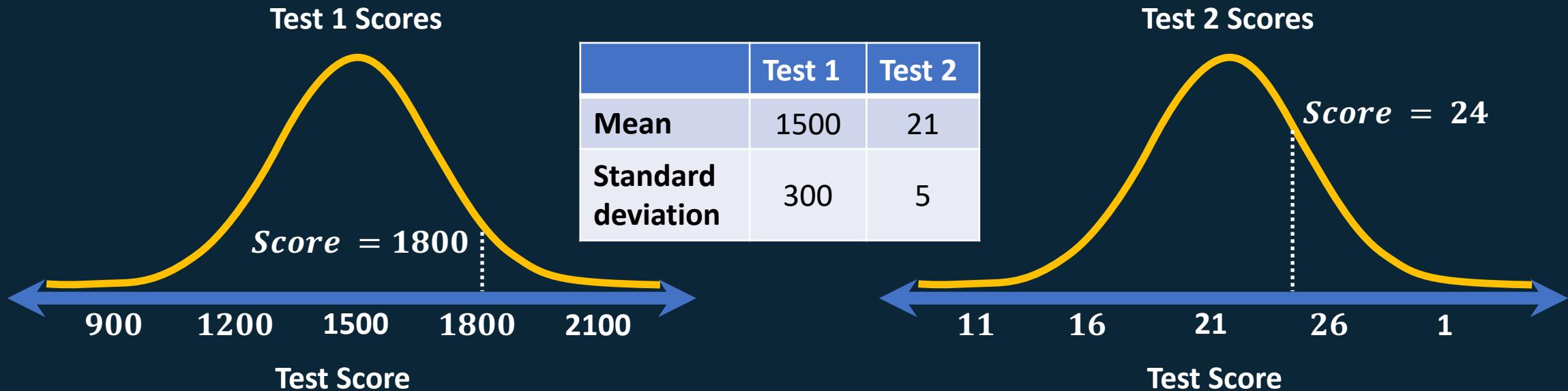
5. Back to Normal



Credit: Simple Learning Pro

6. Z-Score

- Suppose you're given two normal distributions. These represent the test scores of a collection of students on two different tests.
- We then get scores for an individual student.
- They score 1800 on test 1, and 24 on test 2.
- We then collect details about the mean and standard deviation of the data for each test.
- The question is, did the student do better on test 1 or test 2?



7. Z-Score

- One way to answer this question, is to determine how many standard deviations from the mean each test result is.
- We're assuming here the better result to be the one further from the mean in the positive direction.
- We can use the Z-score to determine how many standard deviations an observation x falls above or below the mean.

Score Test 1 = 1800

Score Test 2 = 24

	Test 1	Test 2
Mean μ	1500	21
Standard deviation σ	300	5

$$z = \frac{x - \mu}{\sigma} \quad \text{Z - Score}$$

Test 1 – Z Score

$$\begin{aligned} z &= \frac{1800 - 1500}{300} \\ &= \frac{300}{300} \\ &= 1 \sigma \end{aligned}$$

Better
result!

Test 2 – Z Score

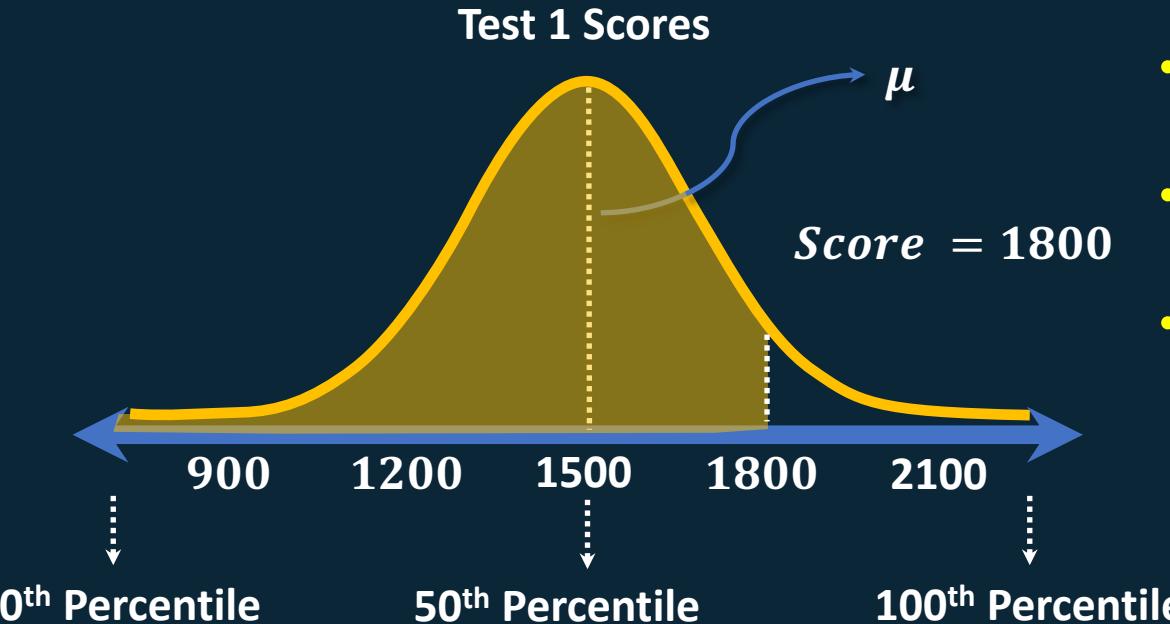
$$\begin{aligned} z &= \frac{24 - 21}{5} \\ &= \frac{3}{5} \\ &= 0.6 \sigma \end{aligned}$$

8. Z-Score



Credit: Simple Learning Pro

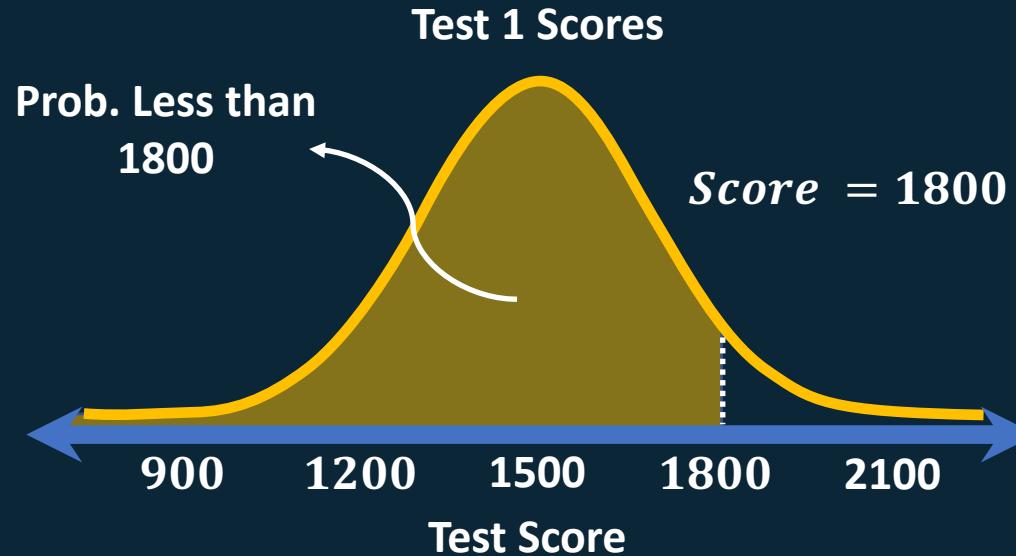
9. Percentile



- Suppose we wanted to know the percentile of the result for the student on test 1.
- This can be represented by the area below the score achieved by the student.
- The total area under the curve is equal to one. Think back to probability – there are lots of potential scores a student can get, but the probability of all those added together is 1.

$$P(\text{Score} = 0) + P(\text{Score} = 1) + \dots + P(\text{Score} = 2100) + P(\text{Score} = 2101) \dots = 1$$

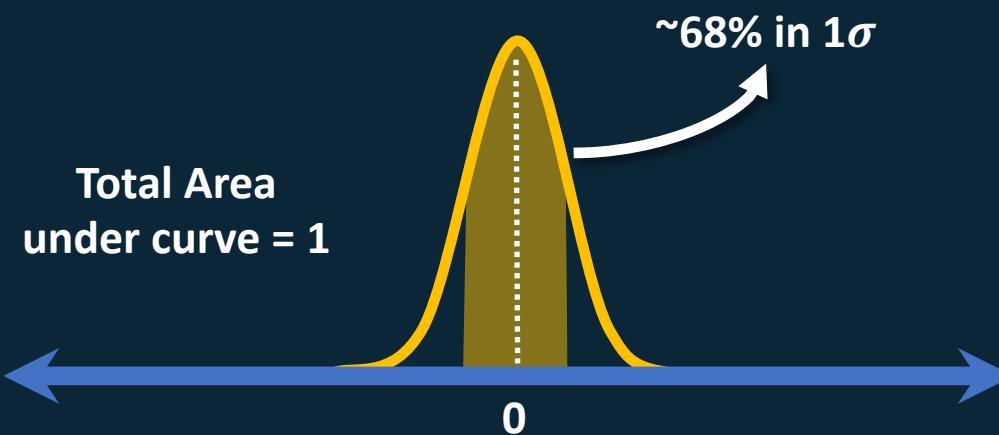
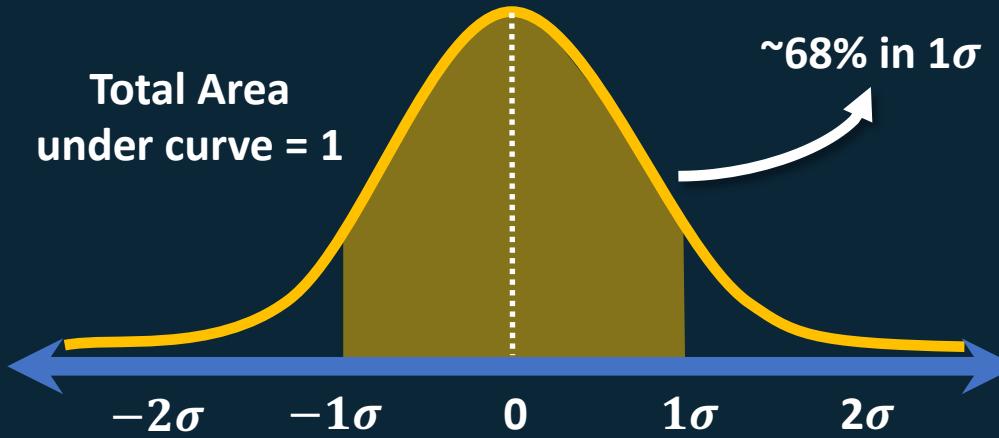
10. Percentile



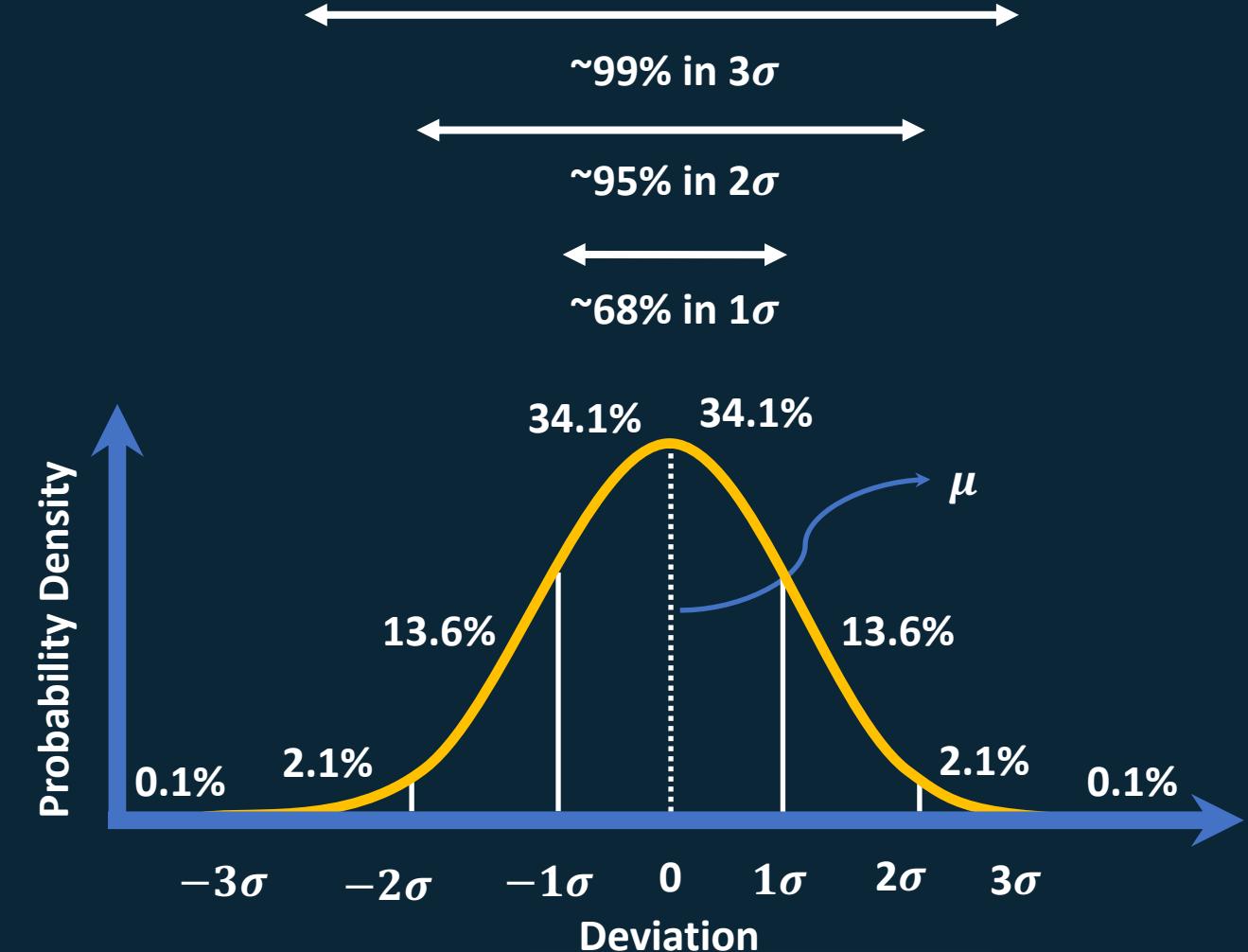
- So what percentile does the student's performance correspond to?
- We're looking for the probability that a score is less than 1800 in this case.
- This would be quite tricky to compute – there's a lot of numbers to add!
- There's a trick we can use to easily compute the percentile for normally distributed data.

$$P(\text{Score} = 0) + P(\text{Score} = 1) + \dots + P(\text{Score} = 1799) = ?$$

11. Percentile & Normal Probability

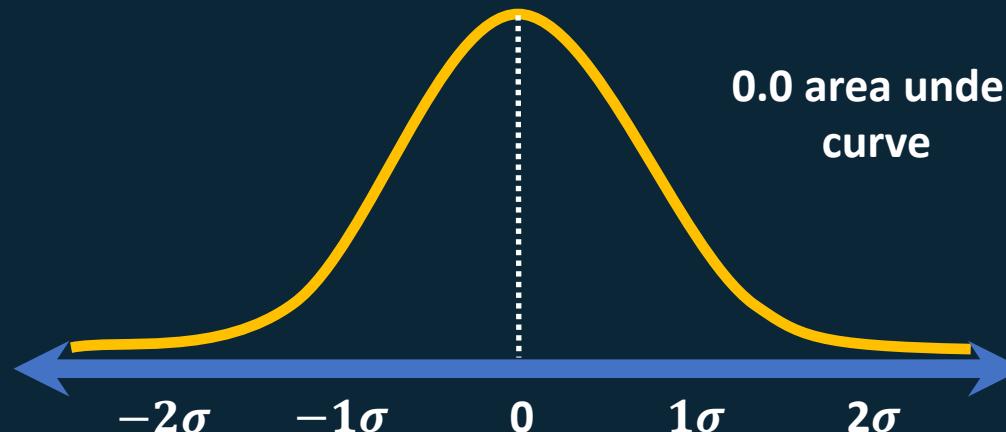


Then an area of 0.682 falls within 1σ

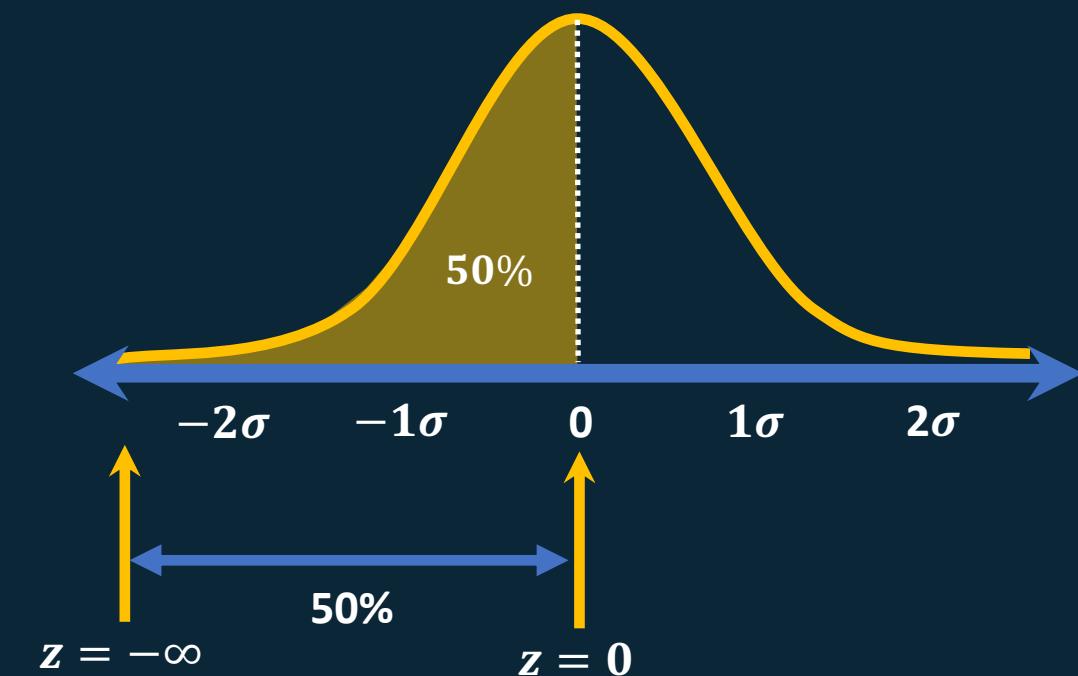
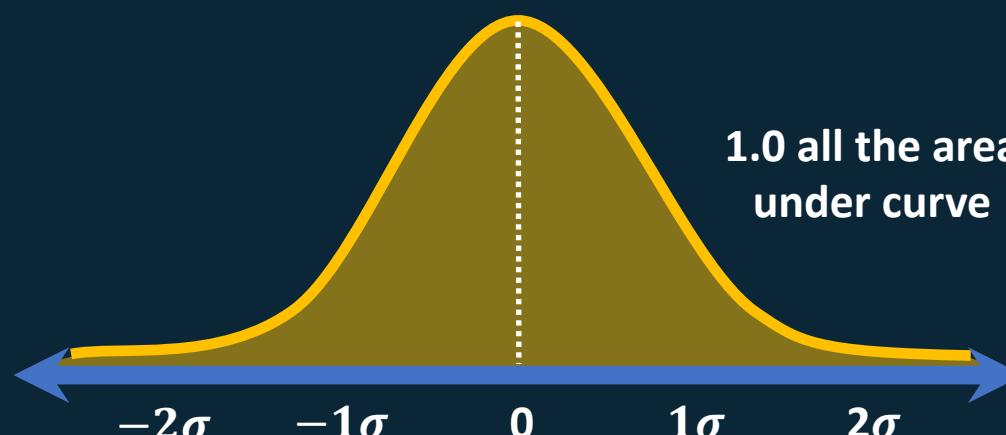


12. Normal Probability & Z-score

0% of the data, 0% Prob. Of a value falling in this range

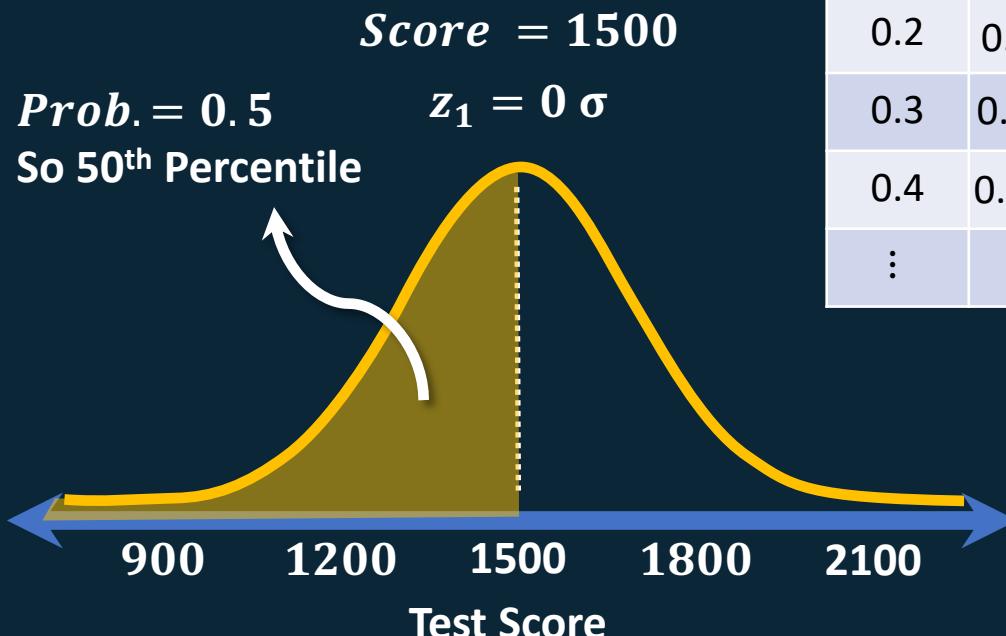


100% of the data, 100% Prob. of a value falling in this range



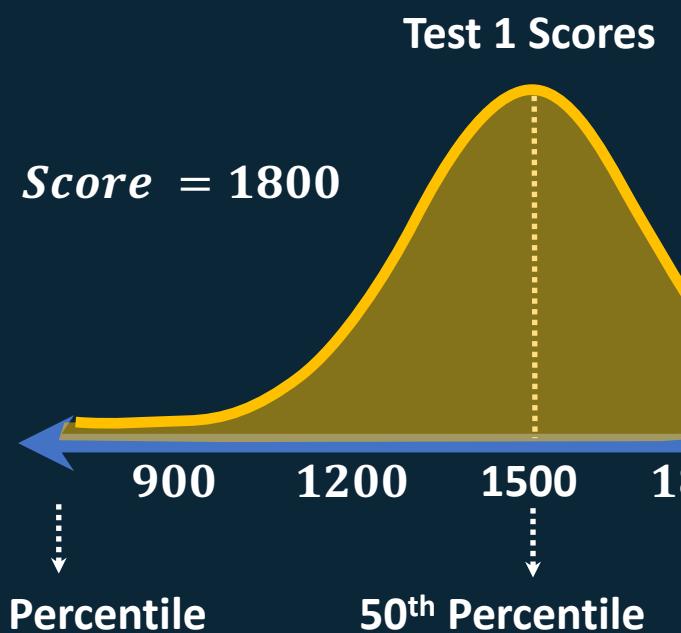
$Z = 0$ corresponds to the 50th percentile.

13. Percentile & Normal Probability Table



Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
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	0.55557	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
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

14. Percentile & Normal Probability Table

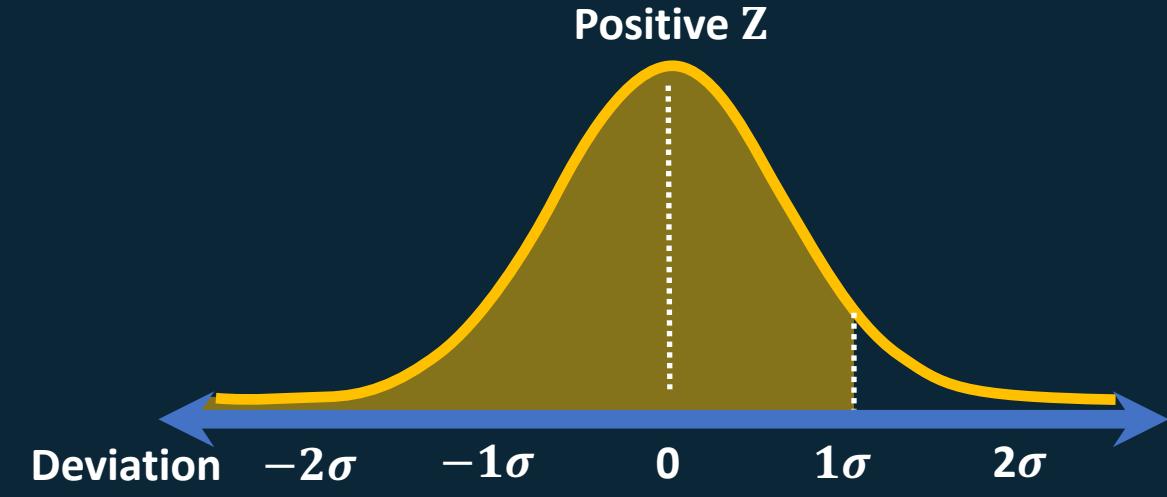
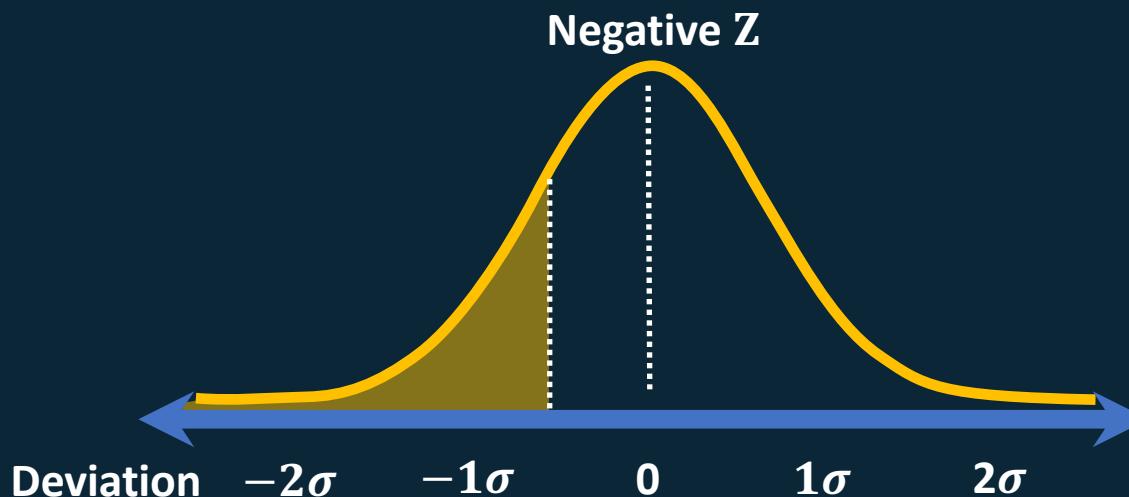


Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.55557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	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
:	:	:	:	:	:	:	:	:	:	:
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.851	0.8554	0.8577	0.8599	0.8621

Prob. = 0.8413 = 84.13% = 84th Percentile

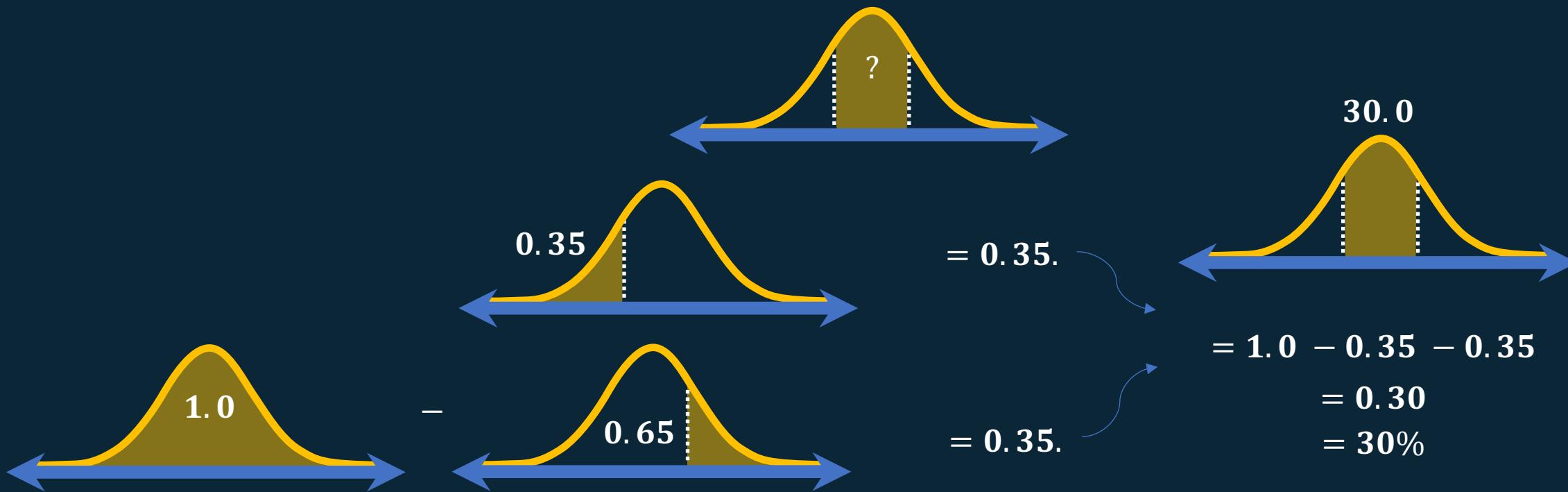
15. Percentile & Normal Probability Table

- There two normal probability tables: for when Z is negative, when z is positive.
- You don't need to remember normal probability tables.
- We can create them in code.
- What matters is that you understand that:
 - normal probability tables exist.
 - they can be used to determine what percentile an observation is in.
 - you must usually compute the Z -score to make use of them.

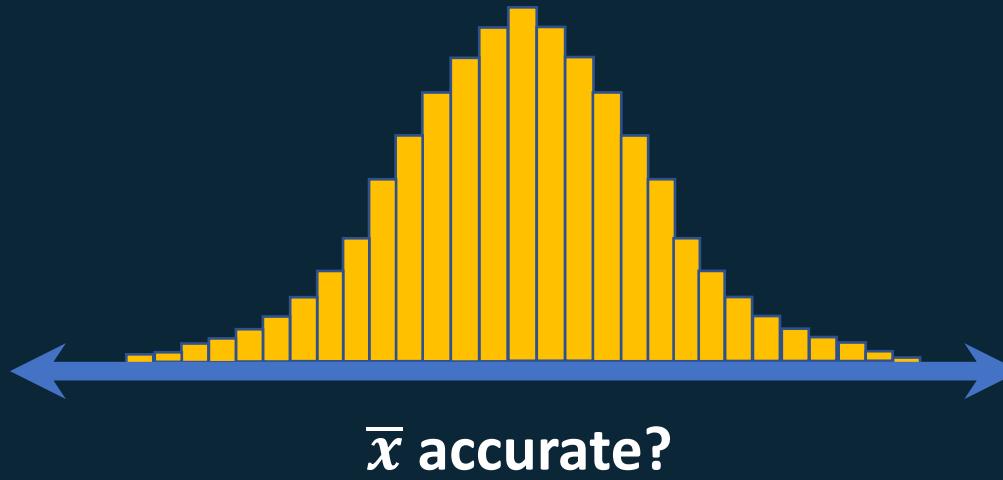


16. Percentile & Normal Probability Table

- Sometimes we may not be looking for simple percentiles for our data.
- We may wish to know what proportion of our data sits between two specific positions.
- We can use the concepts we've already learned to answer some questions.
- We can do this by first calculating percentiles and then subtracting them from 1.
- Once we determine the remainder, we can use this in further calculations.



17. Standard Error



Standard Error of the Sample mean \bar{x}

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Samples in observation

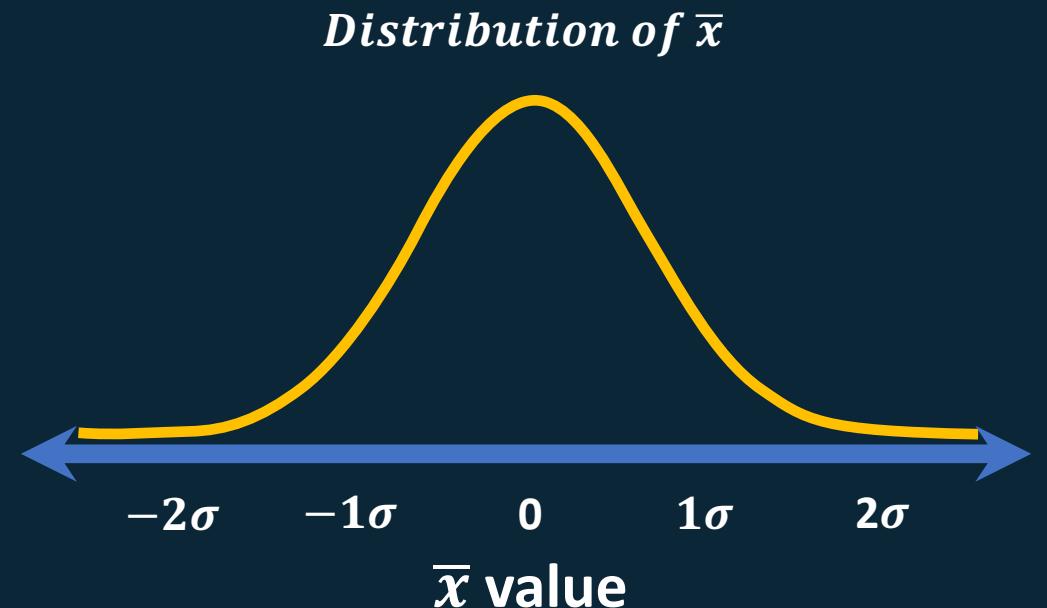
Standard deviation

A yellow curved arrow points from the term "Samples in observation" to the denominator \sqrt{n} in the equation.

- When we collect data, it usually represents a sample from a much larger population.
- Are our summary statistics accurate?
- The sample mean \bar{x} won't be exactly equal to the population mean μ . It might vary from the true population quite a lot, if the sample is small.
- The standard deviation associated with an estimate is called the Standard error of an estimate.
- The standard error for \bar{x} is an important statistic – provides an indication of how uncertain we are in \bar{x} .

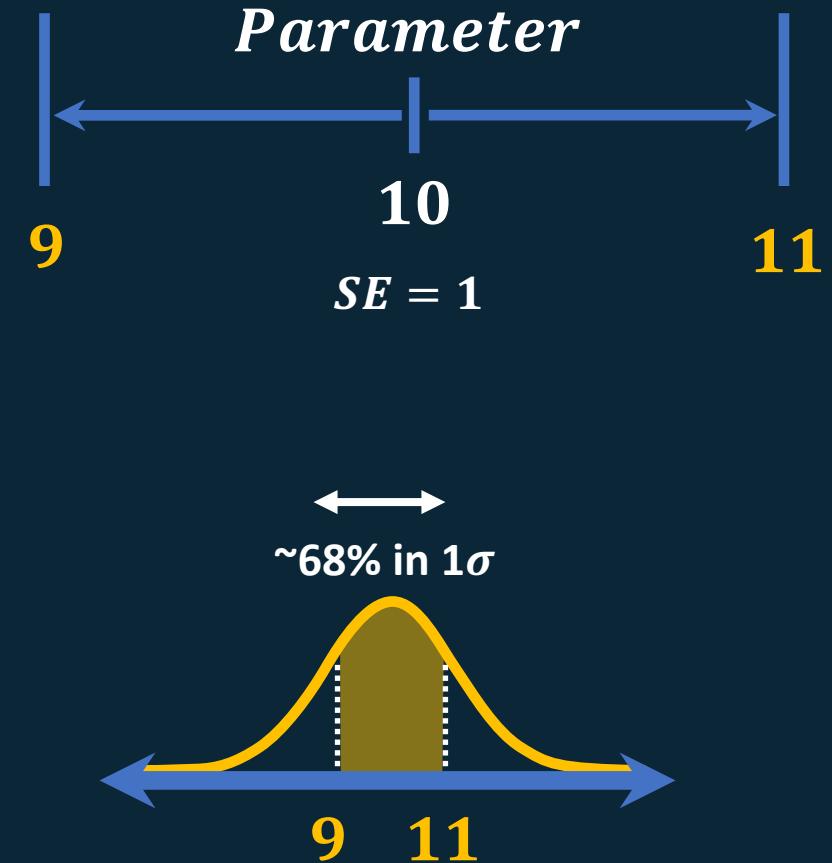
18. Confidence Intervals

- The sample mean for a collection of observations, represents an estimate of μ .
- If we were to make another random sampling, we'll get a slightly different mean estimate.
- If we were to take many random samples from the population, and compute the sample mean for each
 - we would obtain a distribution for the sample mean.
- The average of his distribution is going to be very close to the true mean.
- But how confident are we in our sample mean estimate?



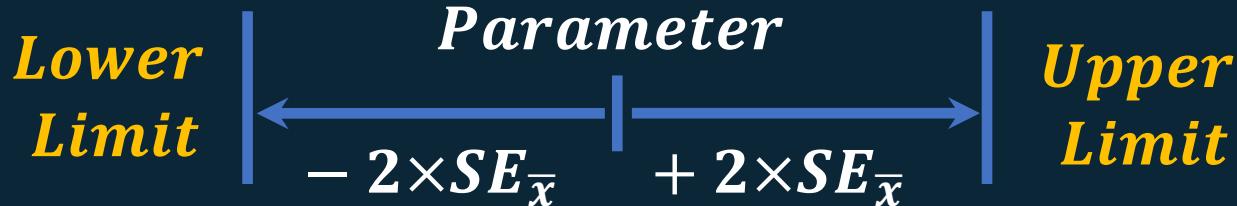
19. Confidence Intervals

- We can apply what we call “confidence intervals” to our estimates, to quantify our confidence level.
- A confidence interval contains the plausible range of values for an estimated parameter, when taking uncertainty into account using the standard error.
- For example, suppose have an estimate for some parameter equal to 10.
- Suppose we also know the estimated parameter has a standard error of 1.
- This means it can plausibly deviate by 1.
- We can take this into account by creating an interval, that takes this deviation into account.
- The plausible range is given by the parameter plus 1, and minus 1 (\pm).
- This is a confidence interval.



20. 95% Confidence Interval

- We can construct a 95% confidence interval over the parameter we wish to estimate, in this case the sample mean, via the following simple formula:



Parameter being estimated

Estimate $\pm 2 \times SE_{\bar{x}}$

Plus-minus Symbol

Standard Error of the Sample mean \bar{x}

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Samples in observation

Standard deviation

21. 95% Confidence Interval - Example

Standard Error:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

~ 2σ Confidence Interval:

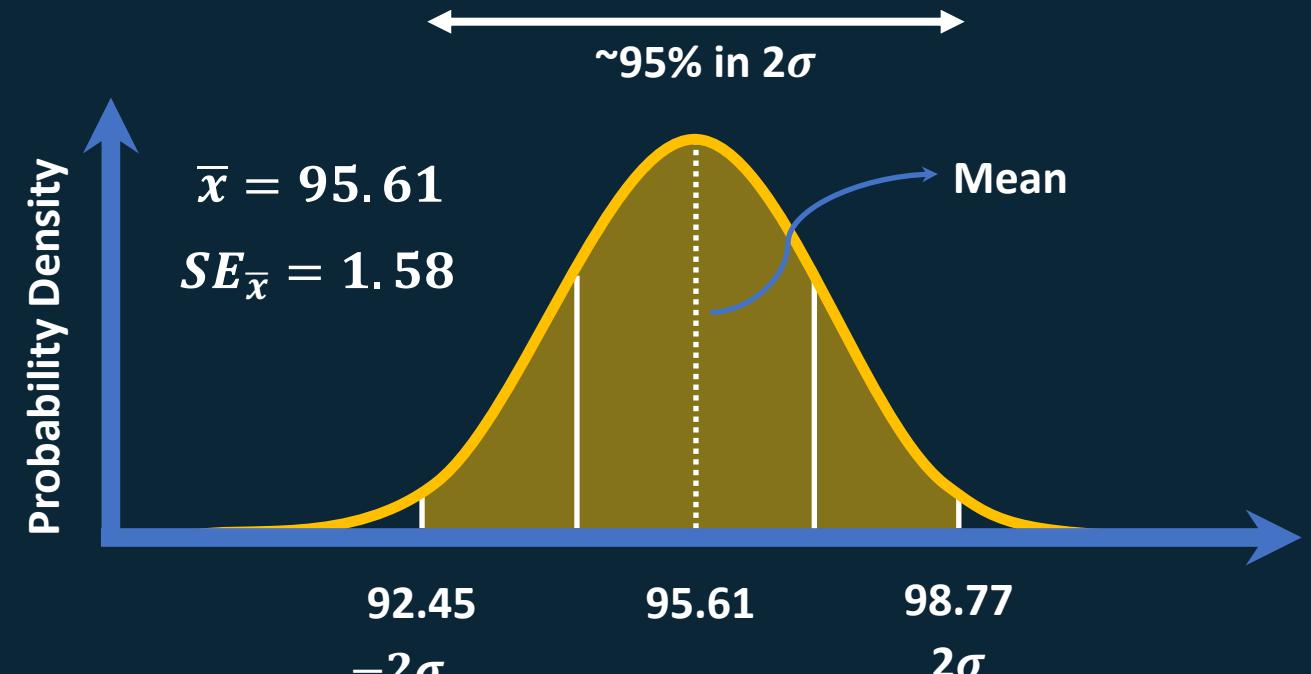
$$Estimate \pm 2 \times SE$$

$\overbrace{95.61 \pm 2 \times 1.58}^{\sim 95\% \text{ Interval}} = (92.45, 98.77)$

$95.61 - (2 \times 1.58) = 92.45$

$95.61 + (2 \times 1.58) = 98.77$

True mean somewhere in this interval, with ~95% confidence



22. 95% Confidence Interval



Credit: Zed statistics

23. Confidence Intervals

Standard Error:

$$SE = \frac{\sigma}{\sqrt{n}}$$

95% Confidence Interval:

$$Estimate \pm 1.96 \times SE$$

99% Confidence Interval:

$$Estimate \pm 2.58 \times SE$$

Margin of Error

$$z^* \times SE$$

- Perhaps 95% confidence isn't good enough for you – well you can compute a 99% confidence interval using the formula shown.
- These intervals will apply to normal data only.

$$Estimate \pm 2 \times SE_{\bar{x}}$$

↓
This value!

24. Testing Hypotheses

- We can start testing competing hypotheses using confidence intervals. Suppose we have a dataset describing the finishing times of runners in a race.
- We want to determine if the runners finished in a faster time this year, compared to last year.
- We form two competing hypotheses for this data. The null hypothesis is that there is no difference in average finishing times. The alternative hypothesis, is that the average runtime was different this year compared to last.
- The average runtime for last year's run was 93.29 minutes. (93 minutes and 17 seconds). We thus reframe our hypotheses given this data.

Null Hypothesis

H_0 = No difference

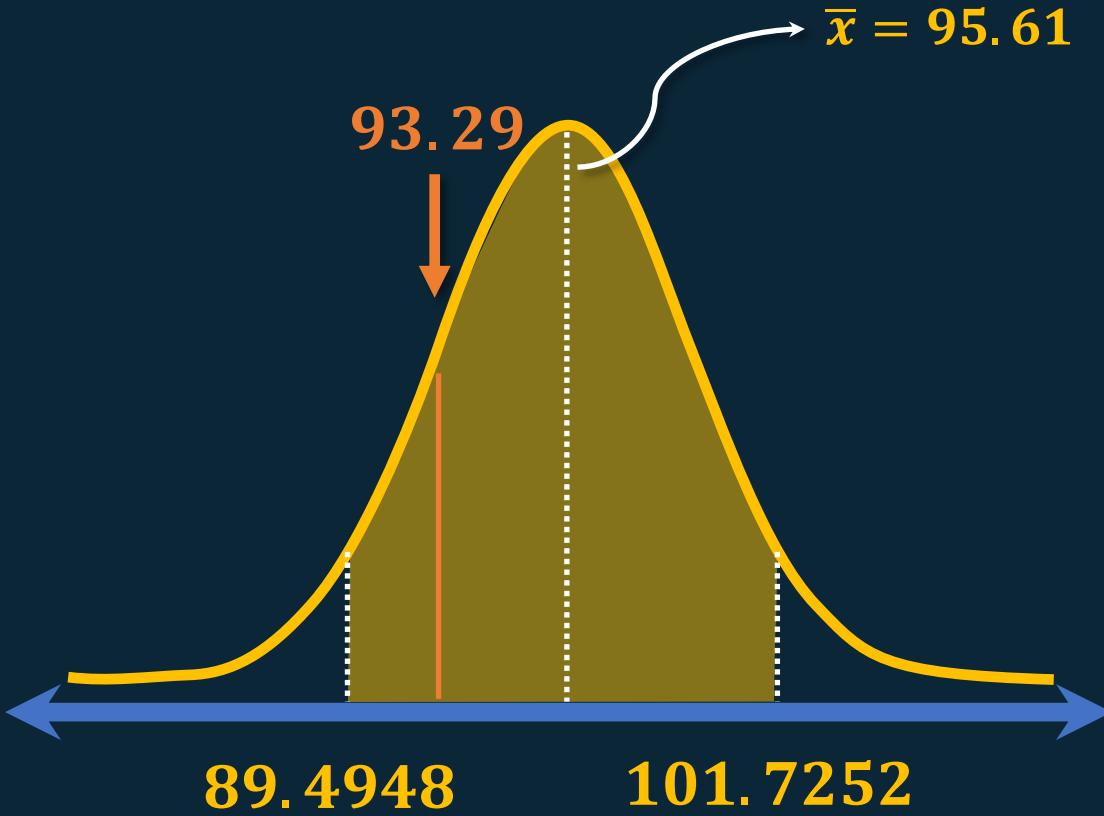
$H_0: \mu_{2019} = 93.29$

Alternative Hypothesis

H_a or H_1 = A difference

$H_1: \mu_{2019} \neq 93.29$

25. Testing H_0 with Confidence Intervals



Null Hypothesis

$$H_0: \mu_{2019} = 93.29$$

$$n = 100$$

Standard Error:

$$SE_{\bar{x}} = \frac{31.2}{\sqrt{100}} = 3.12$$

Alternative Hypothesis

$$H_1: \mu_{2019} \neq 93.29$$

$$s = 31.2$$

95% Confidence Interval:

$$95.61 \pm 1.96 \times 3.12$$

$$1.96 \times 3.12 = 6.1152$$

Lower limit: $95 - 6.1152 = 89.4948$

Upper limit: $95 + 6.1152 = 101.7252$

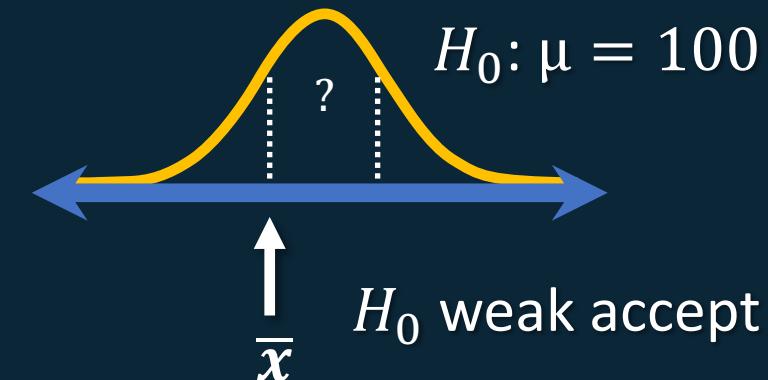
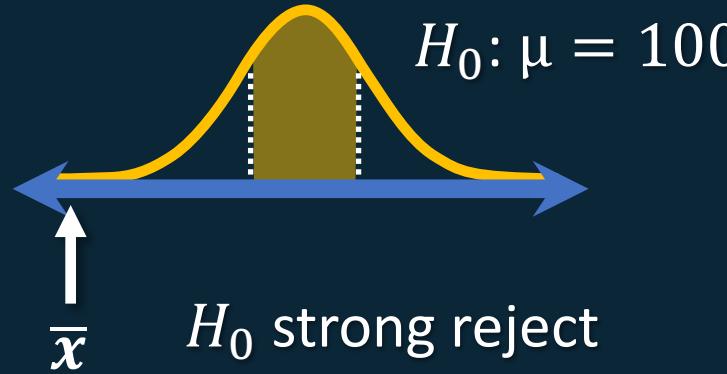
26. Decision Errors

- In general for any hypothesis test, there are four potential test outcomes.
- When running hypotheses tests, we aim to minimise the errors we make. Confidence intervals are great, but alone they don't really help us achieve that. Instead we try to use significance levels to determine how significant a result is, before making a decision.

		Do not reject H_0	$Reject H_0$, Accept H_1
Ground Truth	H_0 True	<i>Success</i>	<i>Type I Error</i>
	H_1 True	<i>Type II Error</i>	<i>Success</i>

27. Decision Errors

- Confidence intervals are simplistic when it comes to hypothesis testing.
- Suppose we use a 95% confidence interval for some sample mean data, where the null hypothesis is accepted if the sample mean falls within 1 standard deviation of the mean.
- Sometimes the evidence against the null hypothesis may be overwhelming, like here.
- But sometimes we may be on the cusp of rejecting the null hypothesis, but don't quite have enough evidence to reject it.
- In these situations it's helpful to be able to quantify our confidence in the decisions we make. We can do this using a tool called, the P-value.



28. Decision Errors



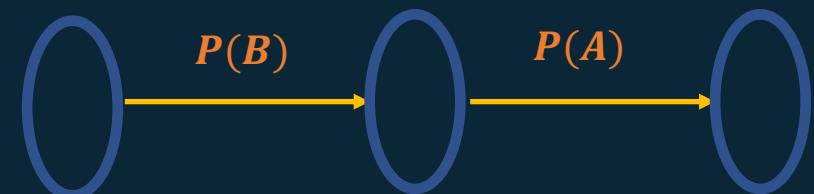
Credit: 365 Data Science

29. P-values

- P-values allow us to test the strength of the evidence against the null hypothesis.
- The P-value is a conditional probability – it is the probability of observing data at least as favourable to the alternative hypothesis as our current dataset is, if the null hypothesis is true.
- It may help to think of this description as a tree diagram. We can see know that the p-value is simply assessing the probability of seeing data this favourable to the alternative hypothesis, given that the null hypothesis is true.
- We usually use a summary statistic such as the sample mean to help compute a P-value.

Conditional Probability

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



If A = Data favourable to H_1

If B = H_0

30. P-values + Significance Level

- Example: A national sleep study suggests students sleep on average 7 hours per night.
- You're a data scientist at a local education authority, and are tasked with determining if student in your area are similar.
- You collect data from a student sample ($n = 110$), and find that students in your area are sleeping on average, over seven hours.
- You want to verify that your students are indeed different from the national sample.
- You form two hypotheses:

Null Hypothesis

H_0 = No difference

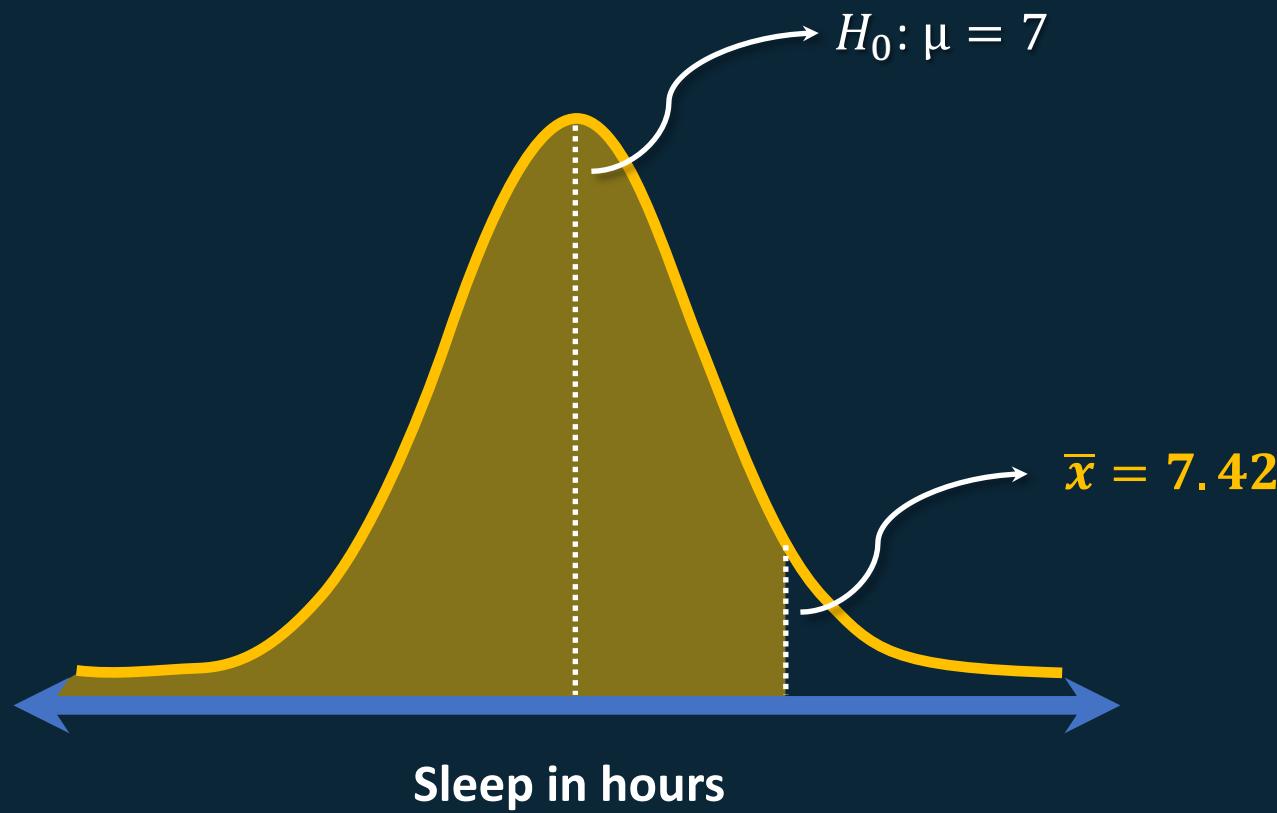
$H_0: \mu = 93.29$

Alternative Hypothesis

H_a or H_1 = A difference

$H_1: \mu > 7$

31. P-values + Significance Level



Null Hypothesis

$$H_0: \mu = 7$$

$$n = 110$$

Alternative Hypothesis

$$H_1: \mu > 7$$

$$s = 1.75 \text{ hours}$$

Z-score:

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

$$z = \frac{7.42 - 7}{0.17} = 2.47$$

Standard error: $\frac{s}{\sqrt{n}} = \frac{1.75}{\sqrt{110}} = 0.17$

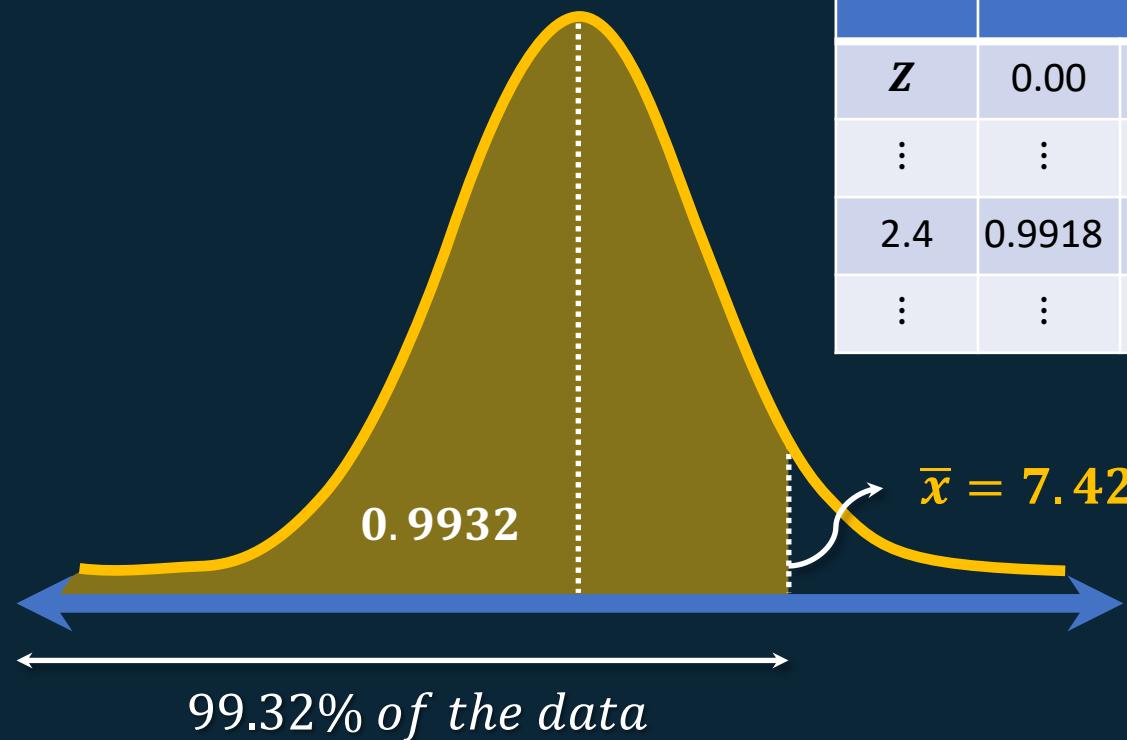
32. P-values + Significance Level

Null Hypothesis

$$H_0: \mu = 7$$

Alternative Hypothesis

$$H_1: \mu > 7$$

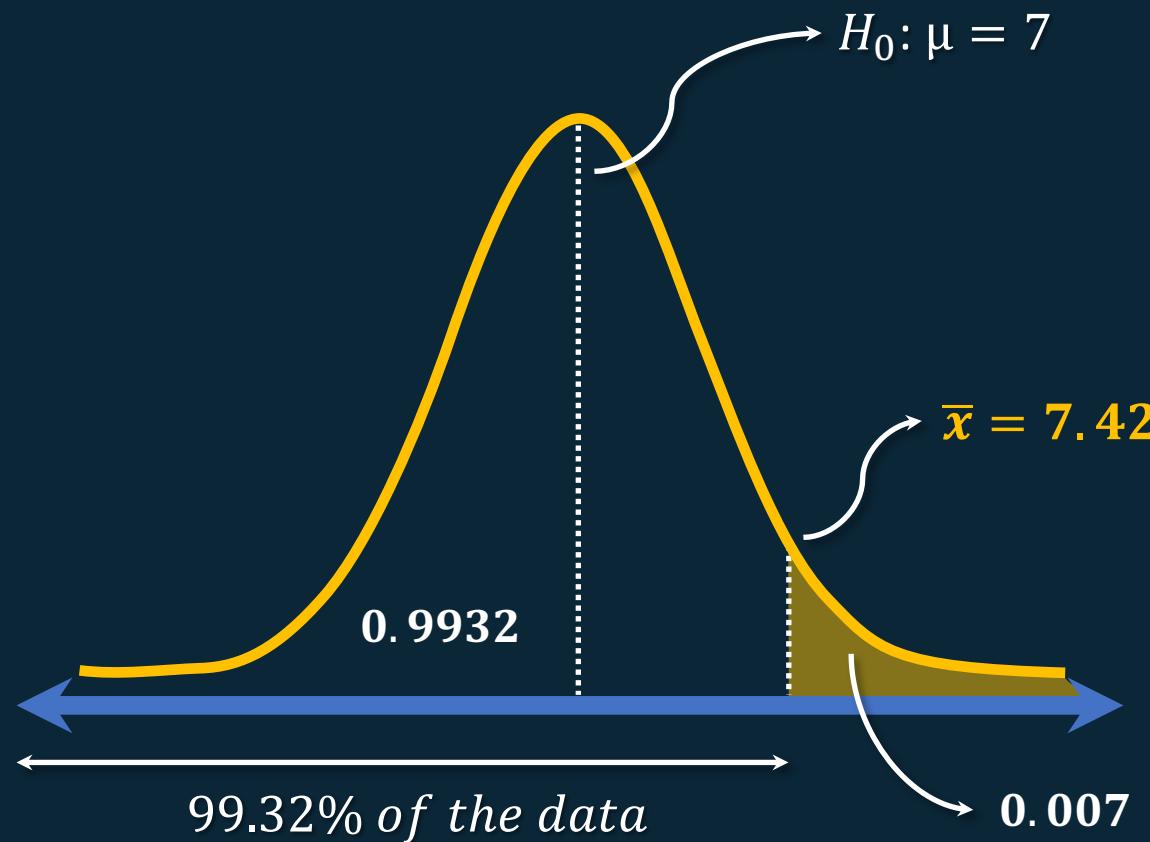


Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$$z = 2.47$$

$$\text{Prob.} = 0.9932$$

33. P-values + Significance Level



Null Hypothesis

$$H_0: \mu = 7$$

Alternative Hypothesis

$$H_1: \mu > 7$$

$$\text{P-value} = 1 - 0.9932 = 0.007$$

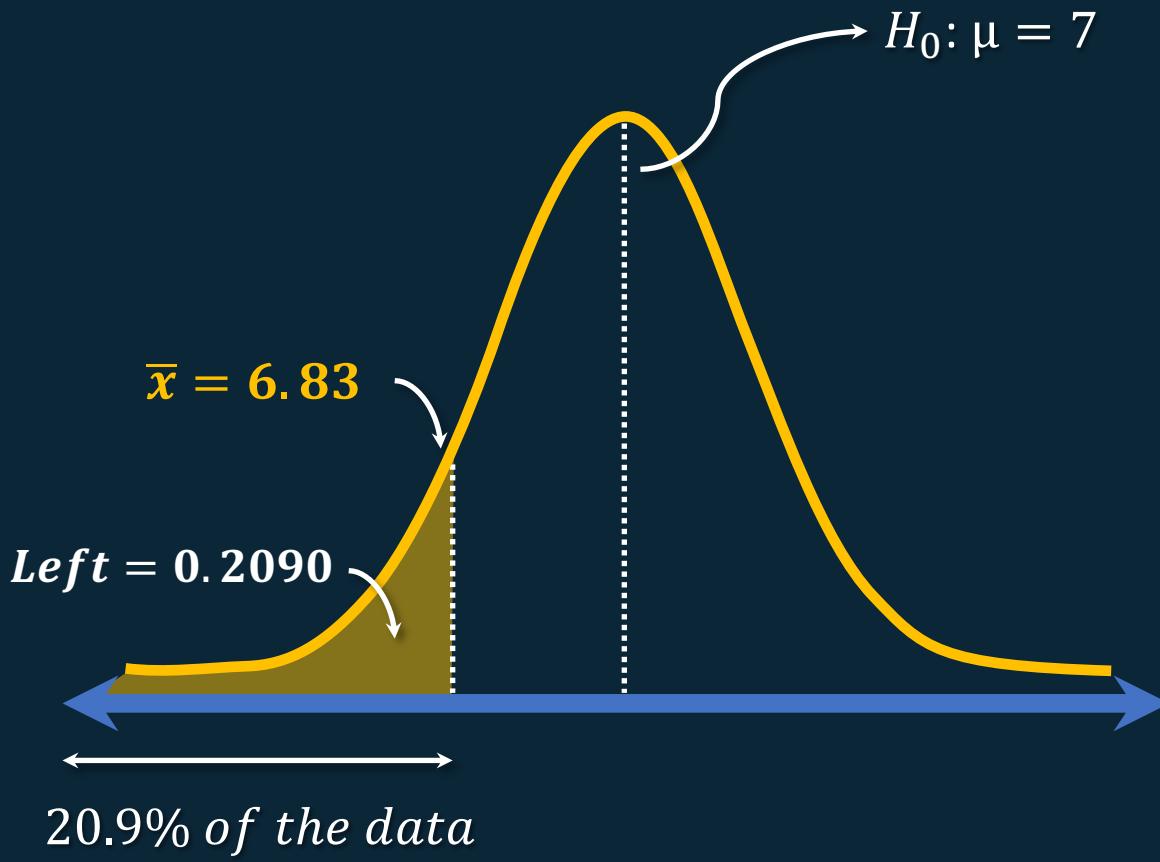
Significance Level

$$\alpha = 0.005$$

$$\text{P-value} = 0.007 < \alpha = 0.005$$

Reject Null Hypothesis

34. Two-sided Hypothesis Test



Null Hypothesis

$$H_0: \mu = 7$$

Alternative Hypothesis

$$H_1: \mu \neq 7$$

$$n = 72$$

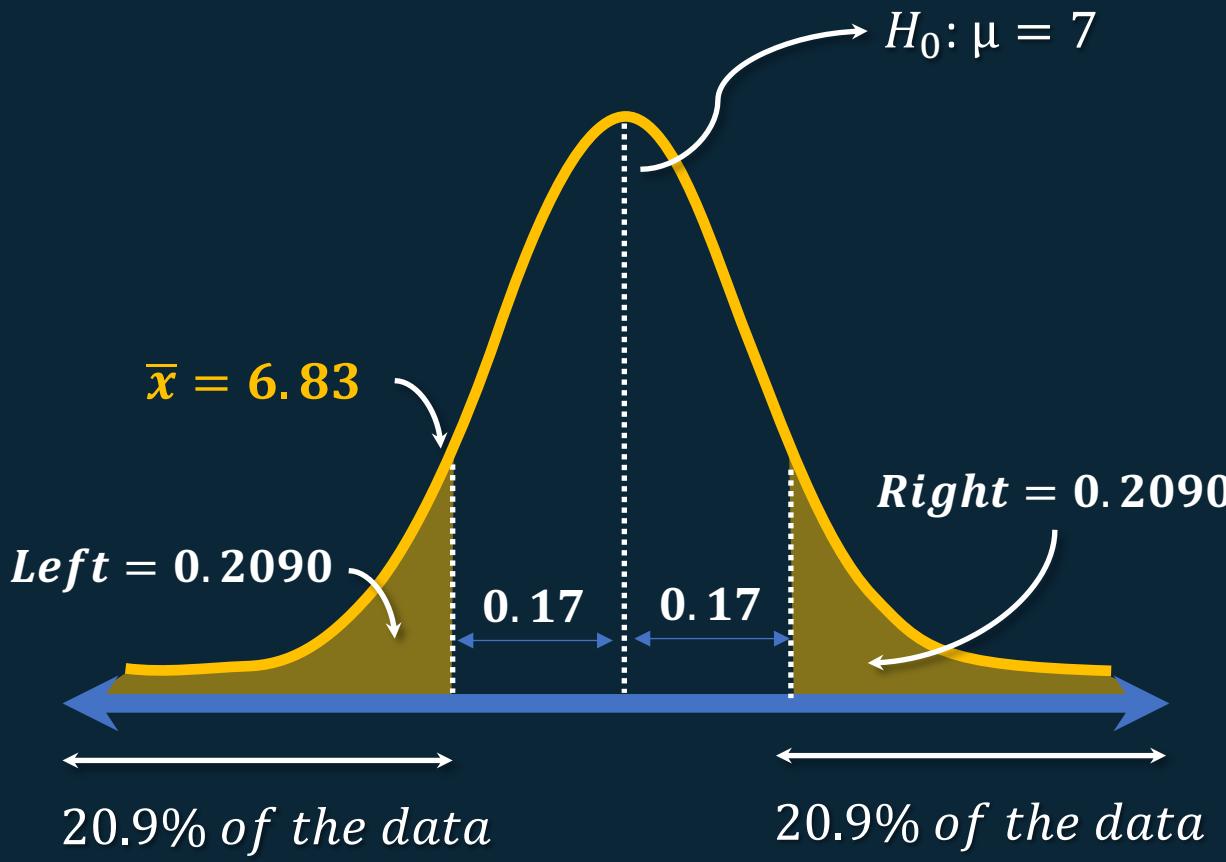
$$s = 1.8 \text{ hours}$$

Z-score:

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

$$z = \frac{6.83 - 7}{0.17} = -0.81 = 0.2090$$

35. Two-sided Hypothesis Test



Null Hypothesis

$$H_0: \mu = 7$$

Alternative Hypothesis

$$H_1: \mu \neq 7$$

$$n = 72$$

$$s = 1.8 \text{ hours}$$

$$z = 0.2090$$

$$\text{P-value} = \text{left tail} + \text{right tail}$$

$$= 0.2090 + 0.2090 = 0.4180$$

$$= 41.80\%$$

36. Two-sided Hypothesis Test

Null Hypothesis

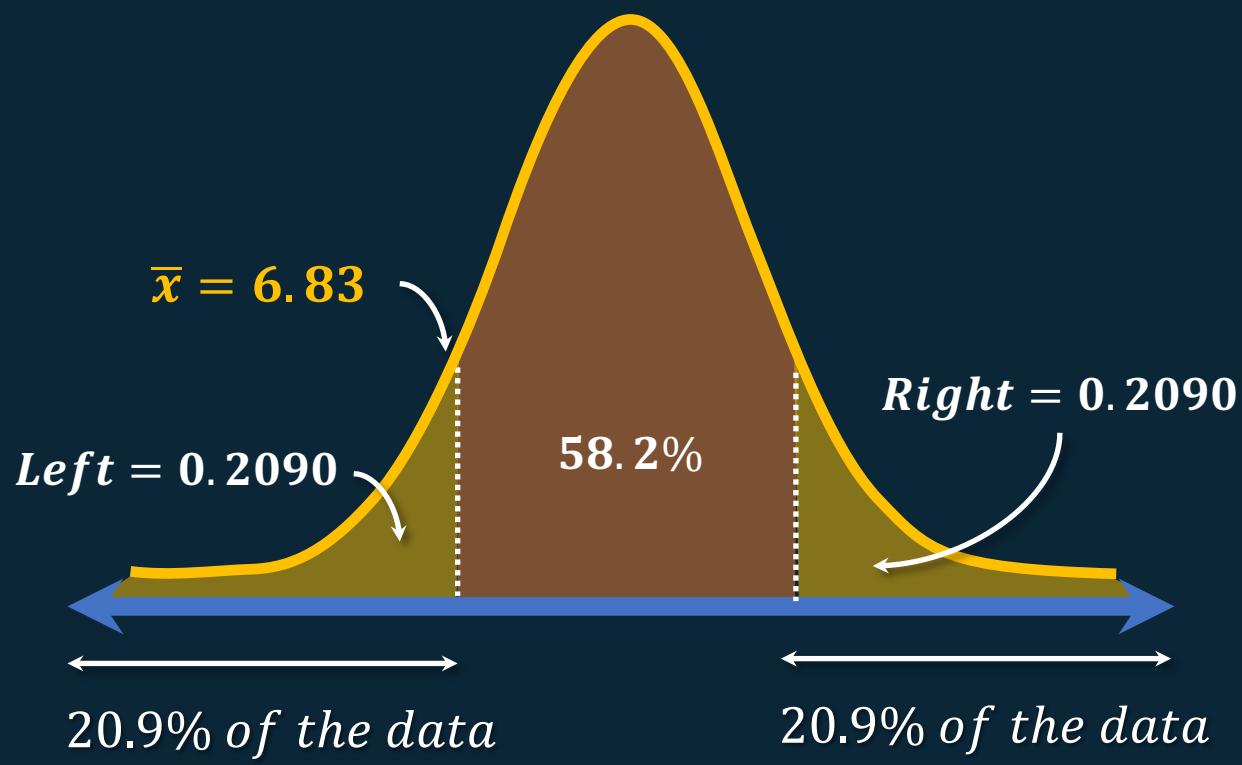
$$H_0: \mu = 7$$

Alternative Hypothesis

$$H_1: \mu \neq 7$$

Significance Level

$$\alpha = 0.005$$



$$P\text{-value} = 0.4180 > \alpha = 0.005$$

Reject Alternative Hypothesis

37. Hypothesis Testing Steps

- Some of what we've covered here may not make sense - yet. That's ok, because nobody becomes a hypothesis testing expert over night!
- What matters is that you appreciate what's happening and why.
 1. We form hypotheses to answer questions about our data.
 2. We collect data samples to test them.
 3. We compute summary statistics over the data sample, such as the sample mean and sample standard deviation.
 4. We compute the Z-score and use this along with normal probability tables to determine the area under the curve.
 5. We use these areas to represent probabilities as p-values, and evaluate them with respect to some significance level, alpha(α).
- A little practice will help make these ideas clearer.



38. Hypothesis testing

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a standard deviation of 0.5 seconds. Do you think that the drug has an effect on response time?

Testing

Hypothesis

p-values

$$H_0: \text{Drug has no effect} \Rightarrow \mu = 1.2 \text{ s (no w/ drug)}$$

$$H_1: \text{Drug has an effect} \Rightarrow \mu \neq 1.2 \text{ s (w/ drug)}$$

$$\text{Assume } H_0: \mu = 1.2 \text{ s}$$

$$Z =$$



Khan Academy

$$\mu_{\bar{x}} = \mu = 1.2 \text{ s}$$



The University of
Nottingham

UNITED KINGDOM · CHINA · MALAYSIA



UNIVERSITY
of York

institute of
CODING

Office for
Students



39. Hypothesis testing



Credit: Kahn Academy



40. Activities

IOC Techup 11b - Activities Supporting: Advanced Data Science

Author: Dr. Robert Lyon
Contact: robert.lyon@edgehill.ac.uk (www.scienceguyrob.com)
Institution: Edge Hill University
Version: 1.0

Code & License

The code and the contents of this notebook are released under the GNU GENERAL PUBLIC LICENSE, Version 3, 29 June 2007. The videos are exempt from this, please check the license provided by the video content owners if you would like to use them.

Introduction

This notebook has been written to support the IOC Techup-Women Module, 11b Advanced Data Science.
This resource is supposed to be used in conjunction with the slides made available for the module.

What is Google Colab?

Google Colab provides a software environment you can use to execute code. This means you don't have to setup any complicated software environments for yourself - you can simply load this site and run our activities. You'll need your own Google ID to login and use this resource to its full potential. So please, sign up for a Google account if you **do not** already have one.

Some of the cells below contain videos that you should watch if unfamiliar with the topics covered. If the cell seems empty (you can't see a video), hover your mouse over the cell. A play button should appear. Click that play button, and then the video should fill the cell. The eagle eyes amongst you might realise that inside these cells I'm using Python to embed some HTML code. This loads the video directly from YouTube. But you don't need to worry about those details.

Using This Resource

1. Login to the Colab using a Google account or create one.
2. Next we need to enable your notebook for this resource. To do this, click on the 'Cell' menu at the top of the page. Then click on the 'Run Cell' option in the 'Cell' menu at the top of the page.

Link to the notebook:

<https://colab.research.google.com/drive/1sq5txv7MQy4uXxBKVjI1E9Dai6sHxpmA>



41. Resources

Books:

- “OpenIntro Statistics: Fourth Edition”, D. Diez, M. Çetinkaya-Rundel and C. Barr.
- “Data Science from Scratch: First Principles with Python”, 2nd Edition, J. Grus.
- “Think Stats: Probability and Statistics for Programmers”, A. B. Downey.
- “Statistics in Plain English, Third Edition: Volume 1”, T. C. Urdan.

Tools Websites

- [Kaggle](#) – an online platform where you can tackle data science challenges.
- [Toward data science](#) – a website where data science practitioners share ideas, tutorials and advice.

42. Checkpoint

We've reached another checkpoint. Let's recap what we've introduced so far.

- Normal distributions.
- The Z-score.
- Probability tables.
- Standard error.
- Confidence intervals.
- Hypothesis testing.

From here you can pursue the activities provided in Google Colab, or watch the next set of slides which cover the ethics of data science. It's entirely up to you.