Data Distribution & Hypothesis Testing

Shah Ayub Quadri

Ayub.quadri89@gmail.com

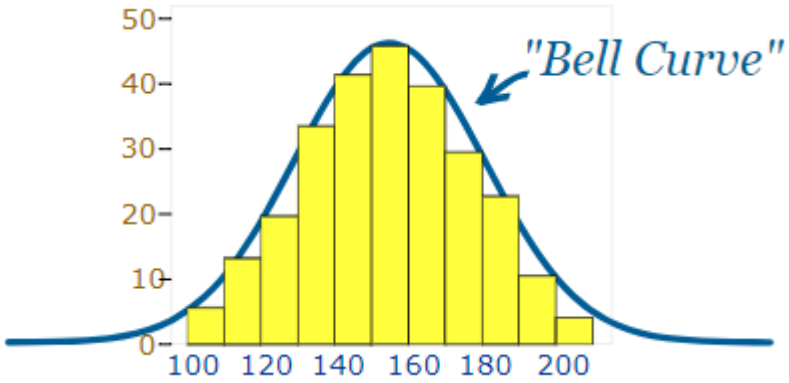
Content

- Data Distribution
  - Normal Distribution
  - Standard Deviations rule
  - Z-scores
- Hypothesis
  - Null Hypothesis
  - Alternate Hypothesis
- Critical region Close up
  - One tailed critical region
  - Two tailed critical region
- Types of errors
  - Type I
  - Type II
  - Power of Hypothesis

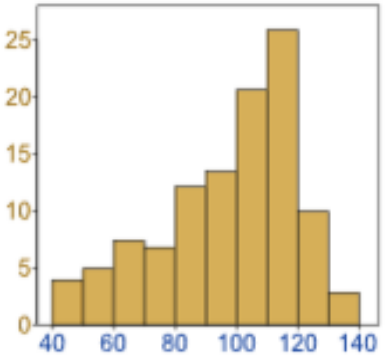
Data Distribution

- Data can spread either towards
  - left or
  - right or
  - Jumbled across

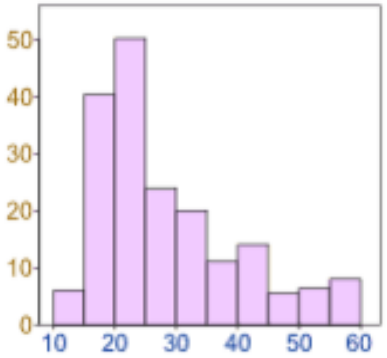
- When data tends to be around a central value with no left and right biases then such spread is know as ***Normal distribution***



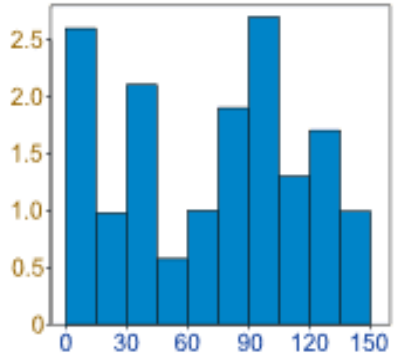
It can be spread out more on the left



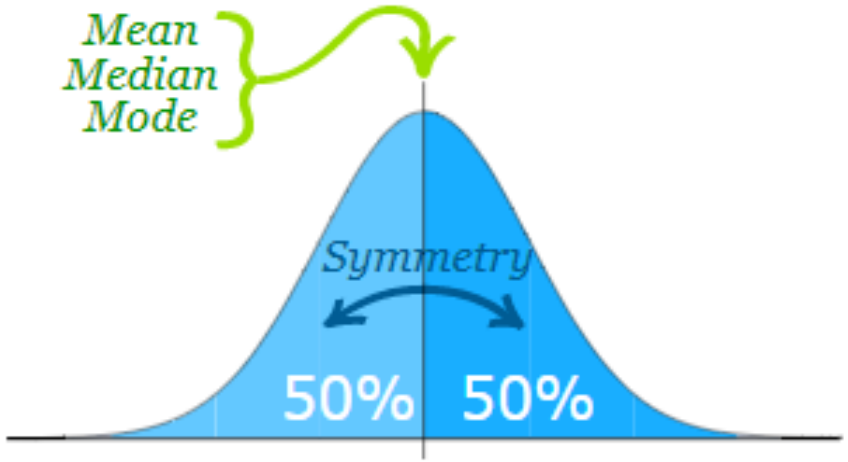
Or more on the right



Or it can be all jumbled up



We say the data is "normally distributed":

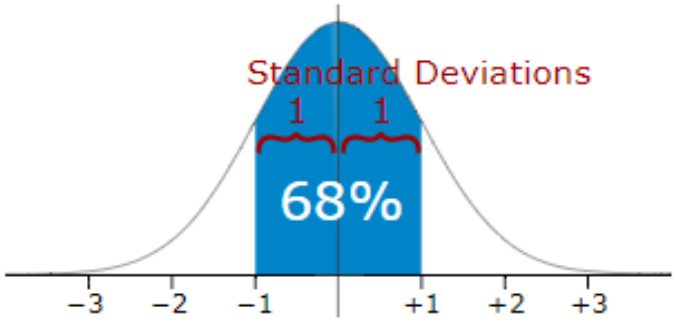


The **Normal Distribution** has:

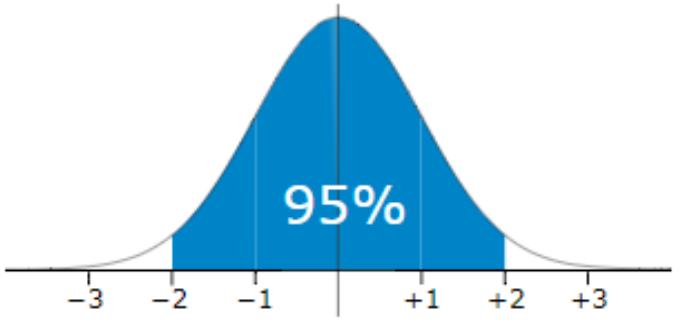
- mean = median = mode
- symmetry about the center
- 50% of values less than the mean and 50% greater than the mean

Standard Deviation: it's a measure of how spread out the numbers are.

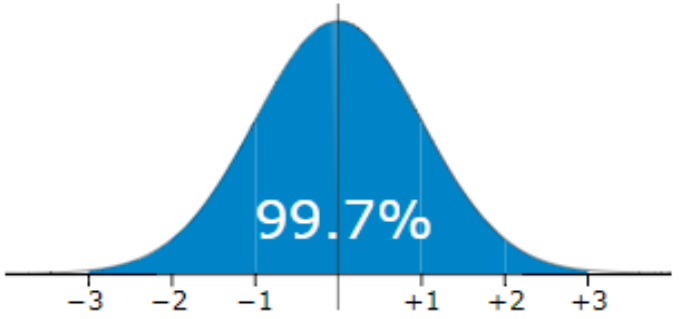
- SD is denoted by (σ)
- $\sigma = \sqrt{variance}$
- Population Variance$(\sigma^2) = \frac{\sum(x-\mu)^2}{N}$
- Sample variance$(s) = \frac{\sum(x-\bar{x})^2}{N-1}$
- 1st SD $= (\mu \pm \sigma)$
- 2nd SD $= (\mu \pm 2\sigma)$
- 3rd SD $= (\mu \pm 3\sigma)$



68% of values are within
1 standard deviation of the mean



95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the mean

Z-score

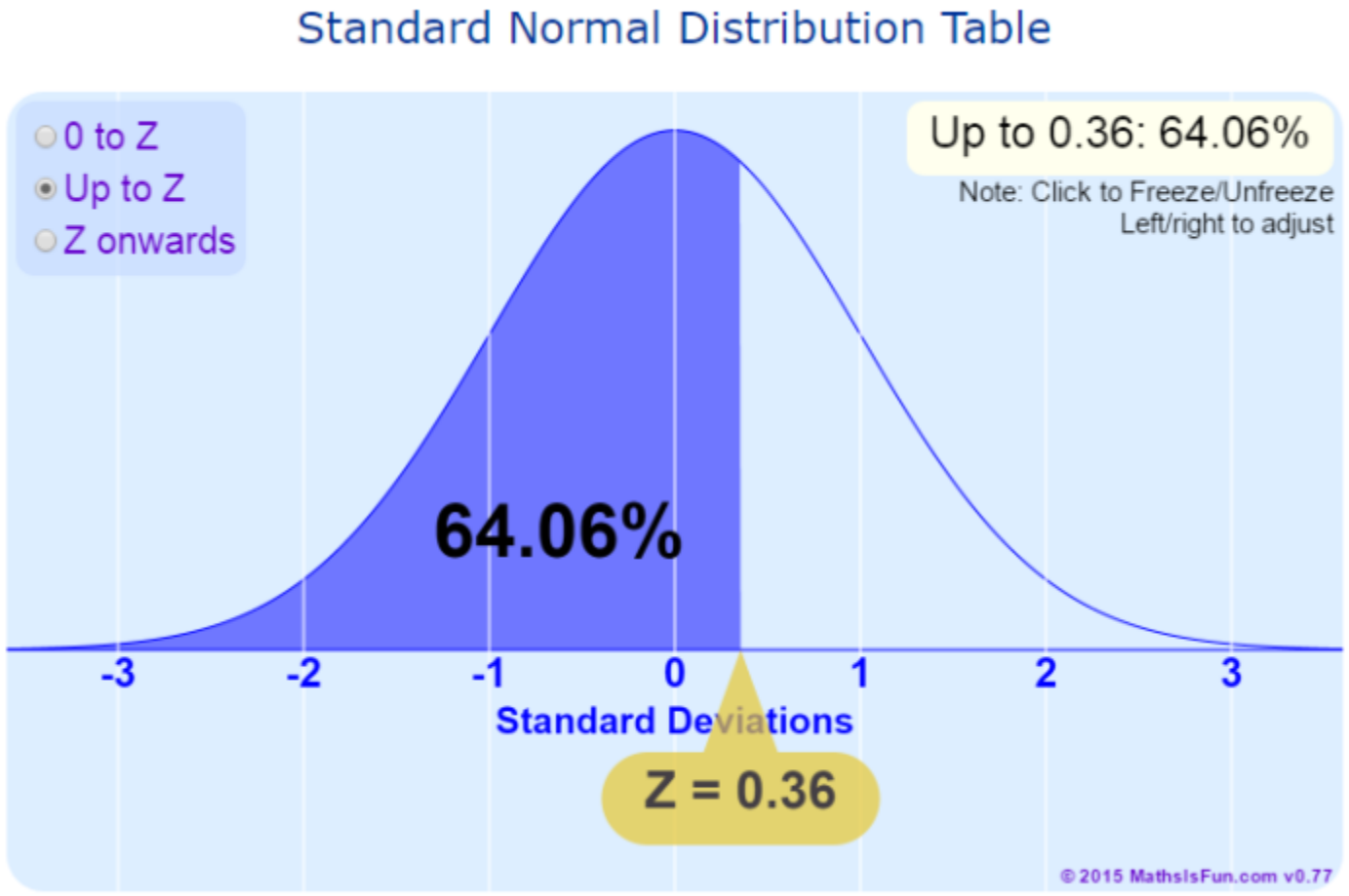
it's a measure of how many standard deviations below or above the population mean a raw data is.

- z-score is also known as a Standard Score
- Z-score $= \frac{(x-\mu)}{\sigma}$

Eg: let's say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be

$$Z = \frac{190-150}{25} = 1.6$$

Which means z score is 1.6 SD away form mean



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

Hypothesis

Hypothesis test gives a way of using samples to test whether or not statistical climes are likely to be true or not.

Examples of Hypothesis test

- Two hypotheses in competition:
  - H_0: The NULL hypothesis, usually the most conservative.
  - H_1 or H_A: The ALTERNATIVE hypothesis, the one we are actually interested in.
- Examples of NULL Hypothesis:
  - The coin is fair
  - The new drug is no better (or worse) than the placebo
- Examples of ALTERNATIVE hypothesis:
  - The coin is biased (either towards heads or tails)
  - The coin is biased towards heads
  - The coin has a probability 0.6 of landing on tails
  - The drug is better than the placebo

Problems on Hypothesis testing

- A school principal claims that the students from her school have an average score of 7/10 in a English Proficiency test.
- You doubt that claim and take a random sample of 40 students and you find a mean score of 5.5/10, with a sample standard deviation of 1. Can you reject the principal's claim?

Step 1: Decide on the hypothesis

Average score on the test is 7/10.

This is called Null Hypothesis and is represented by H_0.

In this case, H_0: $\mu = 0.7$

If Null Hypothesis is rejected based on evidence, an Alternate Hypothesis, H_1, needs to be accepted. **We always start with the assumption that Null Hypothesis is true.**

In this case, H_1: $\mu < 0.7$

Step 2: Choose your statistic

Sample size = 40

Normal distribution is a good approximation

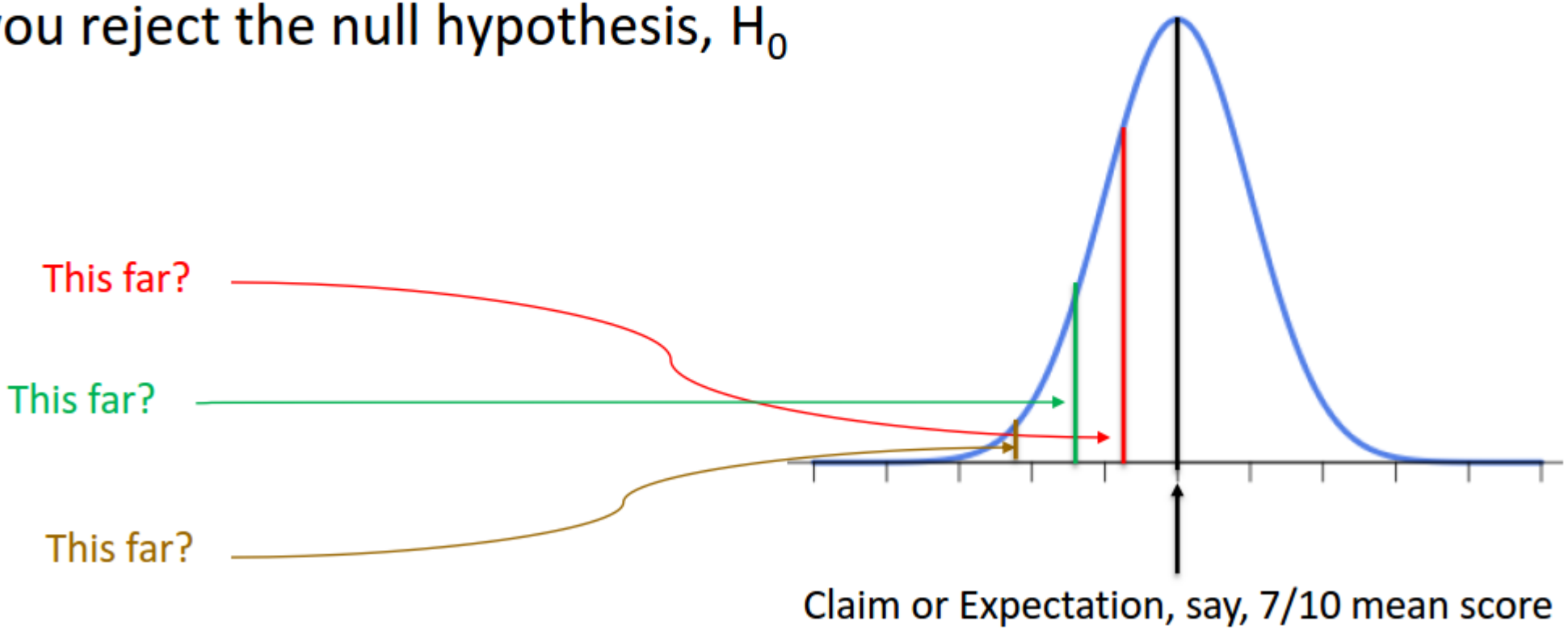
Std Err = $\frac{s}{\sqrt{n}} = \frac{1.0}{\sqrt{40}} = 0.158$

X ~ N(0.7, 0.158^2) = N(0.7, 0.025)

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{0.55 - 0.7}{0.158} = -0.94$$

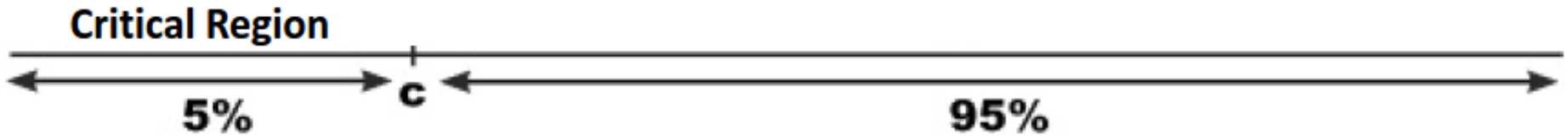
Step 3: Specify the Significance Level

First, we must decide on the Significance Level, α. It is a measure of how unlikely you want the results of the sample to be before you reject the null hypothesis, H_0



Step 4: Determine the critical region

If X represents the sample mean score, the critical region is defined as $P(X < c) < \alpha$ where $\alpha = 5\%$.



Recall that in a 95% CI, there is a 5% chance that the sample will not contain the population mean. Hence if the sample falls in the critical region, the null hypothesis that 0.7 is the mean score is rejected.

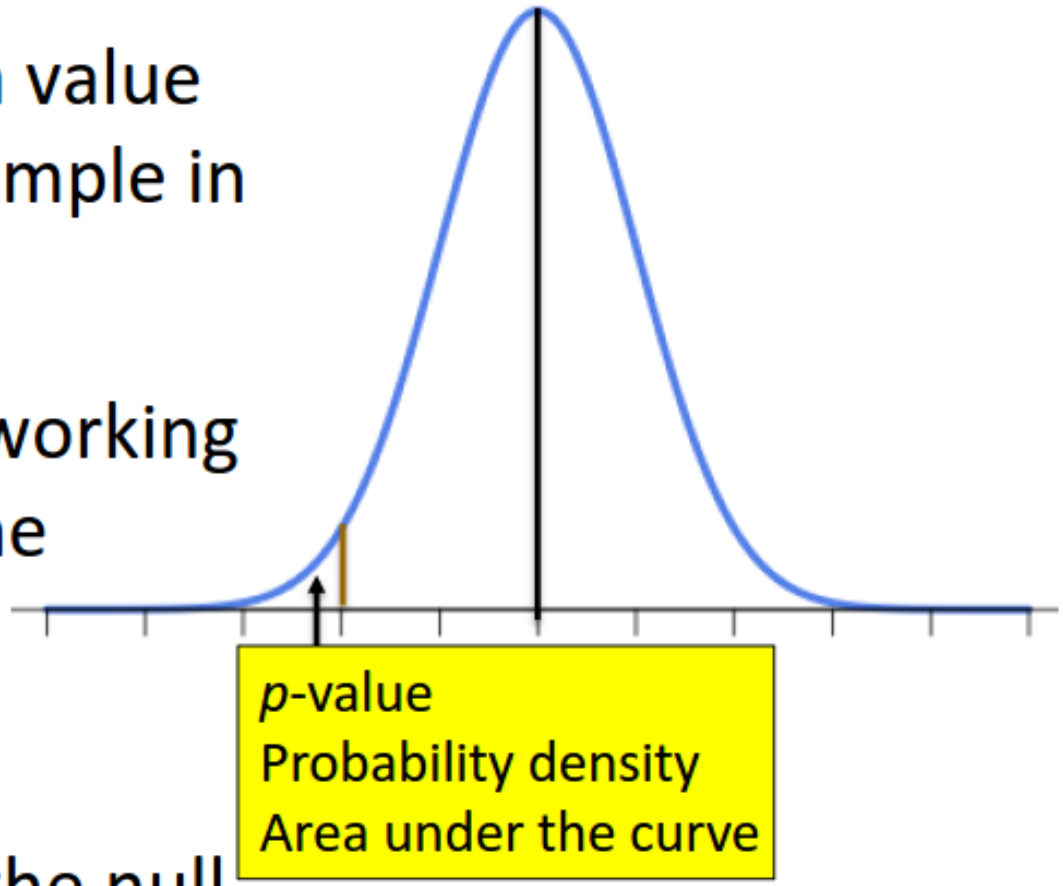
That is the reason 5% or 0.05 is called the Significance Level. In a 99% CI, 0.01 is the Significance Level.

Step 5: Find the *p*-value

p-value is the probability of getting a value up to and including the one in the sample in the direction of the critical region.

It is a way of taking the sample and working out whether the result falls within the critical region of the hypothesis test.

Essentially, this is the value used to determine whether or not to reject the null hypothesis.



Step 5: Find the *p*-value

In our sample, we found a mean score of 5.5/10. This means our *p*-value is $P(X \leq 0.55)$, where X is the distribution of the mean scores in the sample.

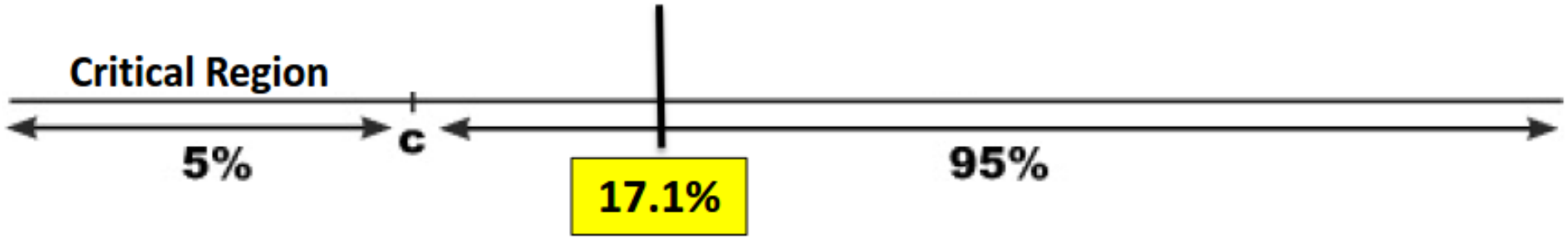
If $P(X \leq 0.55) < 0.05$ (Significance Level), it indicates that 0.55 is inside the critical region, and hence H_0 can be rejected.

Given that Z = -0.94 , $P(X \leq 0.55) = 0.171$

```
> pnorm(0.55,0.7,1/sqrt(40))
[1] 0.1713909
```

So there is a 17% probability of find a mean score of 5.5/10 or less.

Step 6: Is the sample result in the critical region?



Step 7: Make your decision

There isn't sufficient evidence to reject the null hypothesis and so, the claims of the principal are accepted.

Would your conclusion be different if the same average score of 5.5/10 was found from a sample of size 400 ?

What are the null and alternate hypotheses?

$H_0: \mu = 0.7$

$H_1: \mu < 0.7$

What is the test statistic?

$$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{0.55-0.7}{\frac{1}{\sqrt{400}}} = -3$$

p-value = $P(Z < -3.0) = 0.00135$

What is your decision?

Since the *p*-value (0.00135) is less than the Significance Level of 0.05, the null hypothesis can be rejected.

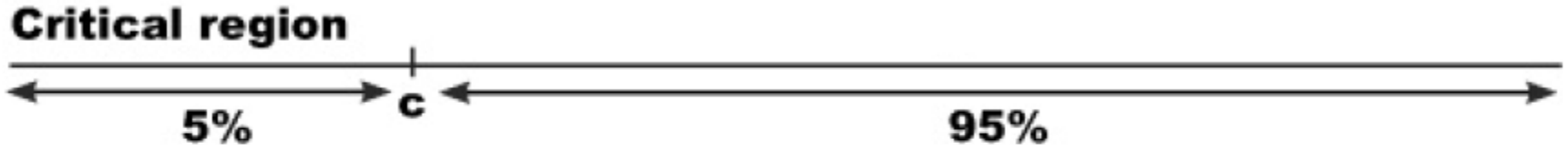
Attention Check

In hypothesis testing, do you assume the null hypothesis to be true or false?

True.

If there is sufficient evidence against the null hypothesis, do you accept it or reject it?

Reject it.



If the *p*-value is less than 0.05 for the above significance level, will you accept or reject the null hypothesis?

Reject it.

Do you need weaker evidence or stronger to reject the null hypothesis if you were testing at the 1% significance level instead of the 5% significance level?

Stronger.

A prisoner is on trial and you are on the jury. The jury's task is to assume that the accused is innocent, but if there is enough evidence, the jury needs to convict him.

In the trial, what is the null hypothesis?

H_0 – The prisoner is innocent (or not guilty).

What is the alternate hypothesis?

H_A – The prisoner is guilty.

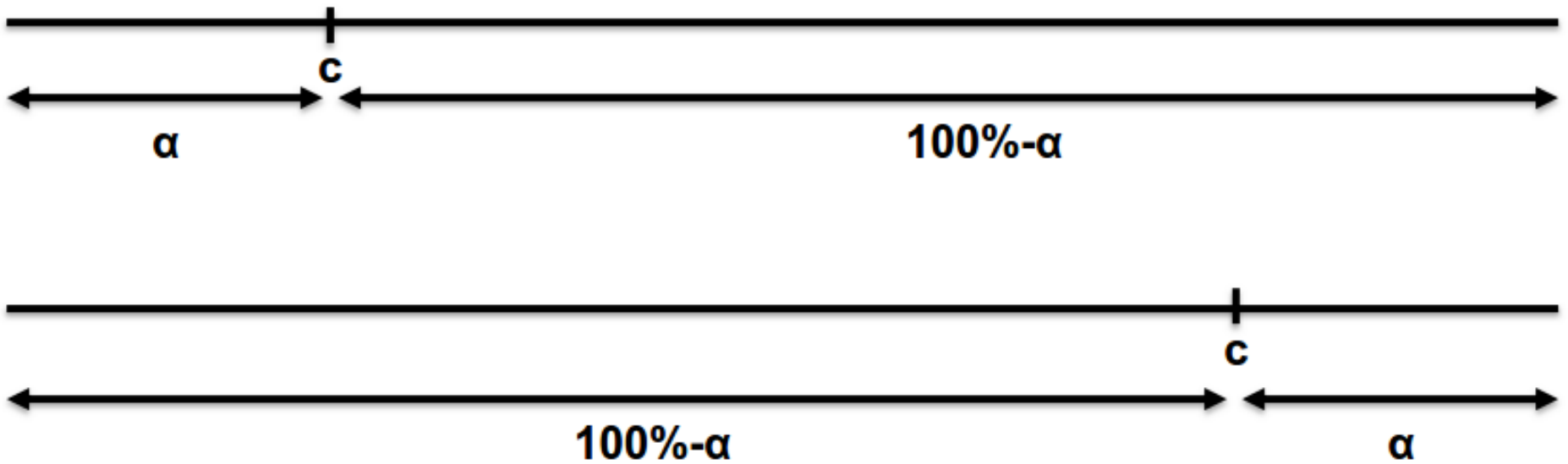
Critical Region Up Close

One-tailed tests

The position of the tail is dependent on H_1.

If H_1 includes a < sign, then the **lower tail** is used.

If H_1 includes a > sign, then the **upper tail** is used.

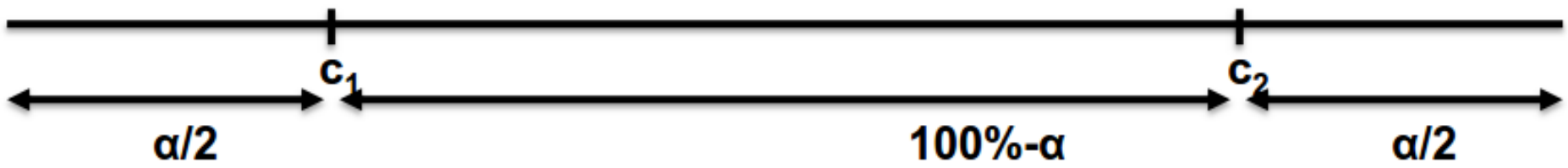


Critical Region Up Close

Two-tailed tests

Critical region is split over both ends. Both ends contain α/2, making a total of α.

If H_1 includes a ≠ sign, then the two-tailed test is used as we then look for a change in parameter, rather than an increase or a decrease.



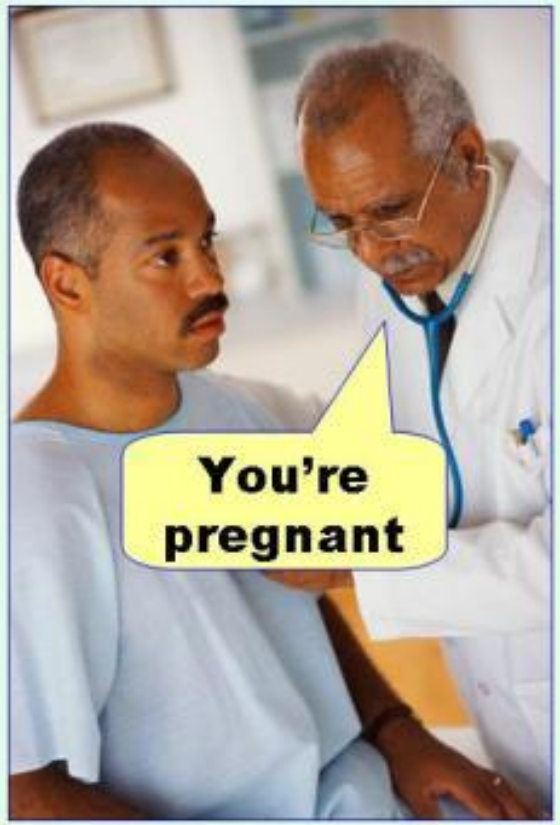
The hypothesis test doesn’t answer the question whether the coin is biased or not; it only states whether the evidence is enough to reject the null hypothesis or not ***at the chosen significance level***.

Errors

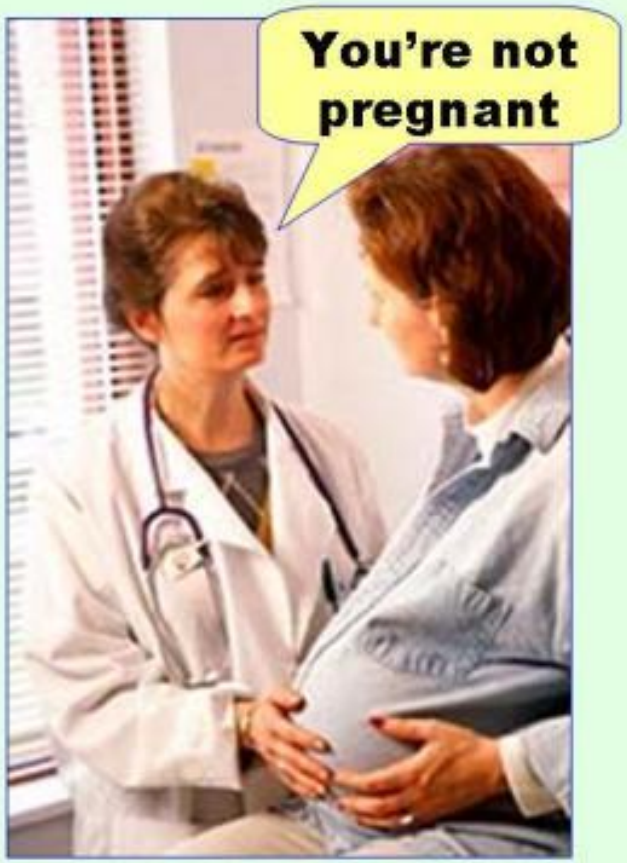
- Type I: We reject the NULL hypothesis incorrectly
- Type II: We "accept" it incorrectly

		State of Nature	
		Null true	Null false
Action	Fail to reject null (negative)	Correct decision True Negative Specificity P(*Accept* H_0 | H_0 True)	Type II error (β) False Negative P(*Accept* H_0 | H_0 False)
	Reject null (positive)	Type I error (α) False Positive P(Reject H_0 | H_0 True)	Correct decision (Power) True Positive Sensitivity/Recall P(Reject H_0 | H_0 False)

Type I error
(false positive)



Type II error
(false negative)



Attention check

What are the possible ways of the jury coming to an incorrect verdict?

If the prisoner is innocent, and the jury gives a ‘guilty’ verdict.

If the prisoner is guilty, and the jury gives an ‘innocent’ verdict.

Which one is Type I and which one Type II?

- First one is Type I because null hypothesis actually was correct but rejected incorrectly.
- Second one is Type II because null hypothesis was false but was accepted incorrectly.

What is the Power of the test?

- Since it is opposite of Type II, it will be finding the prisoner guilty when the prisoner is actually guilty, i.e., rejecting the null hypothesis correctly.

References

- Standard deviation – why use (n-1) instead of n?

  https://www.khanacademy.org/video/review-and-intuition-why-we-divide-by-n-1-for-the-unbiased-sample-variance

  http://nebula.deanza.edu/~bloom/math10/m10divideby_nminus1.pdf

- Z-statistic vs t-statistic:

  https://www.khanacademy.org/video/z-statistics-vs-t-statistics

- Hypothesis testing:

  https://www.khanacademy.org/video/hypothesis-testing-and-p-values