Basic Statistics 2

What Is a Z-Score?

- A Z-score is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean.
- If a Z-score is 0, it indicates that the data point's score is identical to the mean score. A Z-score of 1.0 would indicate a value that is one standard deviation from the mean.
- Z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean.

The Difference Between Z-Scores and Standard Deviation

- Standard deviation is essentially a reflection of the amount of variability within a given data set.
  - What is Variability?
  - Variability is the extent to which data points in a statistical distribution or data set diverge from the average value as well as the extent to which these data points differ from each other.
  - standard deviation is a statistic that measures the dispersion of a dataset relative to its mean
- The Z-score, by contrast, is the number of standard deviations a given data point lies from the mean.

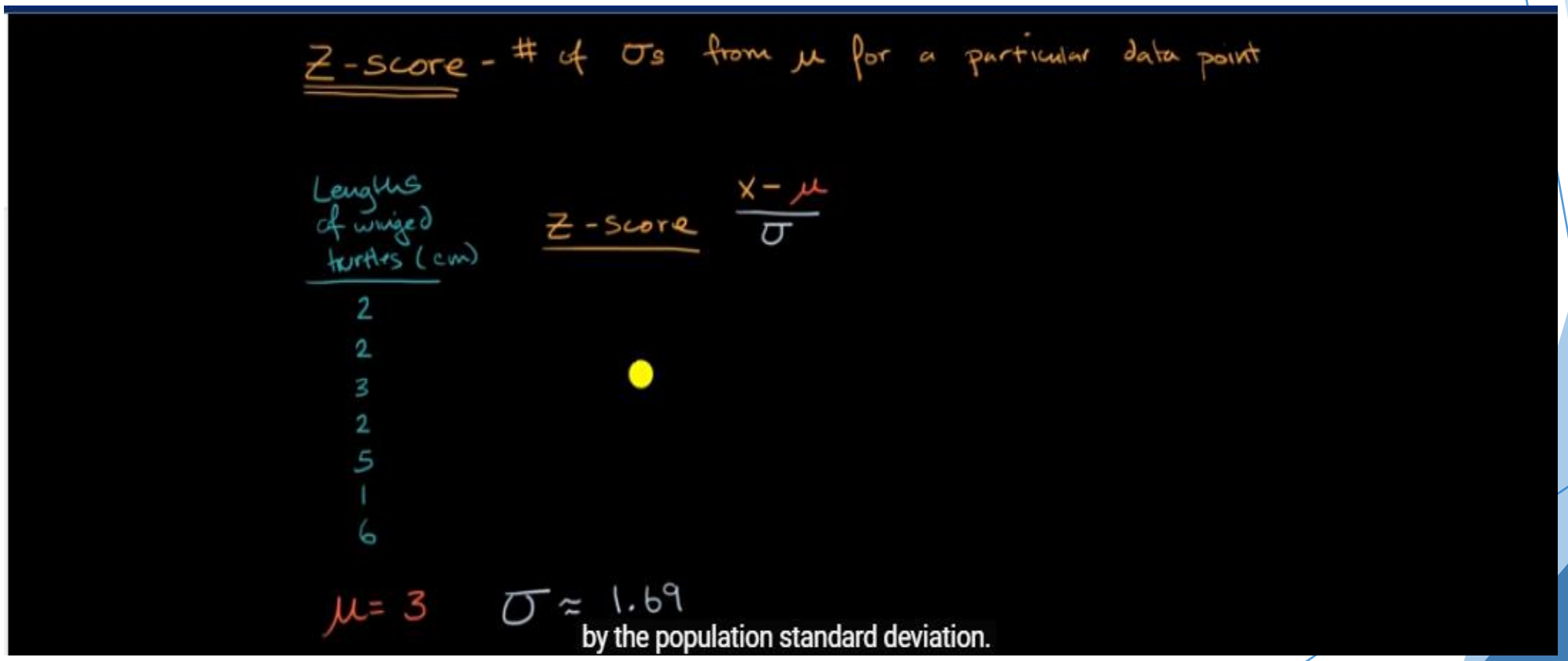
- For data points that are below the mean, the Z-score is negative. In most large data sets, 99% of values have a Z-score between **-3 and 3,** meaning they lie within three standard deviations above and below the mean.

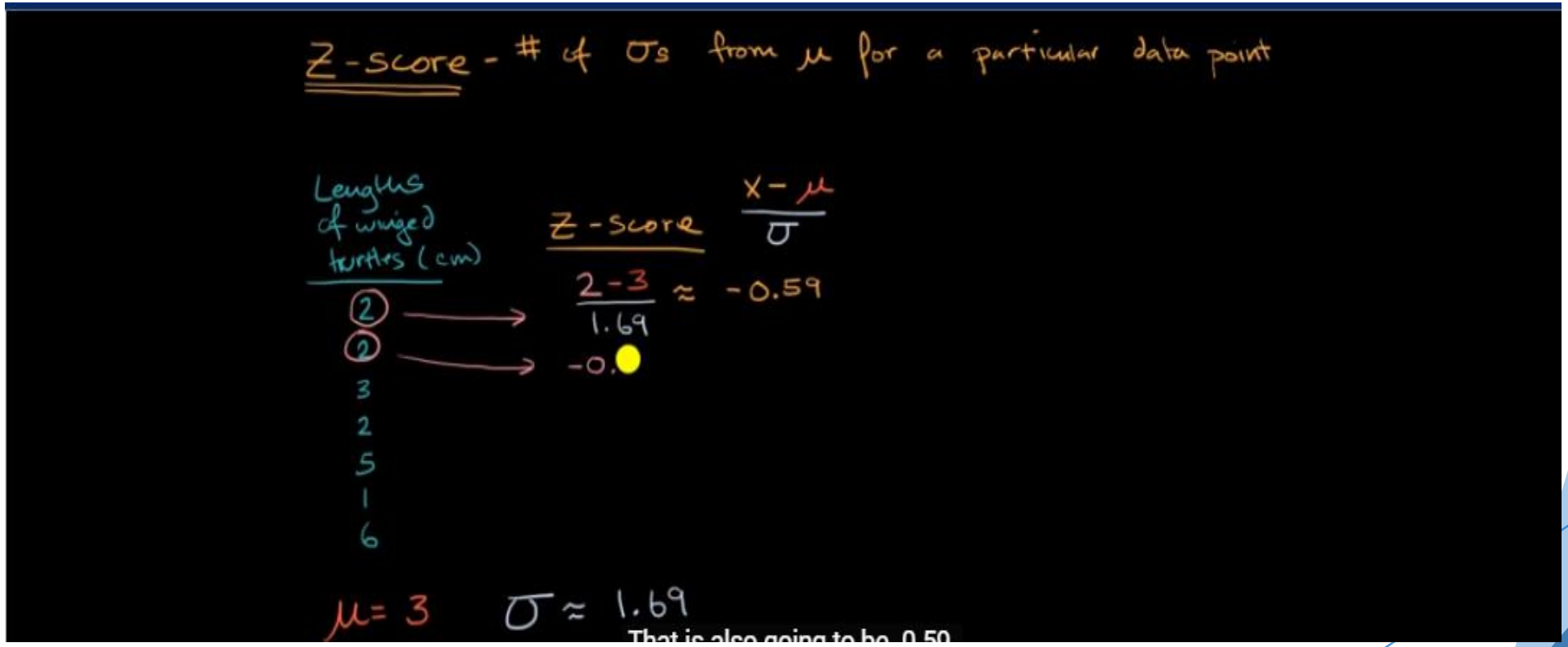
z-score

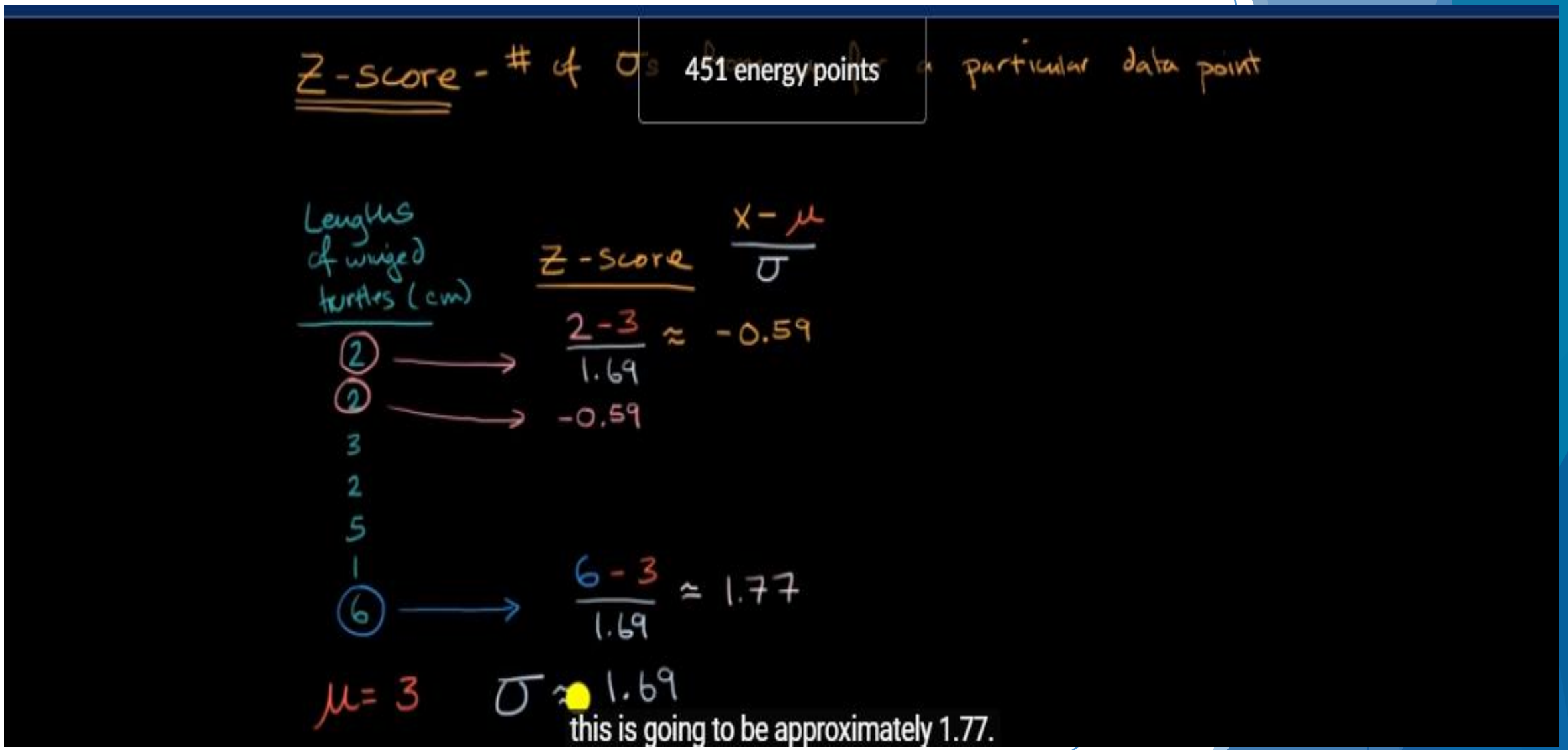
- A z-score, in simple terms, is a score that expresses the value of a distribution in standard deviation with respect to the mean. Let's take a look at the following formula that calculates the z-score:

$$z = (X - \mu) / \sigma$$

Calculating z score of each point




Z-score - # of σs from μ for a particular data point
Lengths of winged feathers (cm)
Z-score
$\frac{x-\mu}{\sigma}$
$\frac{2-3}{1.69} \approx -0.59$
2
2
3
2
5
1
6
$\mu = 3$
$\sigma \approx 1.69$
That is also going to be -0.59


Z-score - # of σ
451 energy points
particular data point
Lengths of winged turtles (cm)
Z-score
$\frac{x-\mu}{\sigma}$
2
2
3
2
5
1
6
$\frac{2-3}{1.69} \approx -0.59$
-0.59
$\frac{6-3}{1.69} \approx 1.77$
$\mu = 3$
$\sigma \approx 1.69$
this is going to be approximately 1.77.

- All the z score is above 1 and below 2 in all above ,mean.

Before applying to law school in the US, students need to take an exam called the LSAT. Before applying to medical school, students need to take an exam called the MCAT. Here are some summary statistics for each exam:

Exam	Mean	Standard deviation
LSAT	$\mu = 151$	$\sigma = 10$
MCAT	$\mu = 25.1$	$\sigma = 6.4$

Juwan took both exams. He scored 172 on the LSAT and 37 on the MCAT.

Which exam did he do relatively better on?

because they are on different scales

LCAT z score

Before applying to law school in the US, students need to take an exam called the LSAT. Before applying to medical school, students need to take an exam called the MCAT. Here are some summary statistics for each exam:

Exam	Mean	Standard deviation
LSAT	$\mu = 151$	$\sigma = 10$
MCAT	$\mu = 25.1$	$\sigma = 6.4$

LSAT

$\frac{|172 - 151|}{10}$

2.1 sto. dev. above mean

Juwan took both exams. He scored 172 on the LSAT and 37 on the MCAT.

Which exam did he do relatively better on?

the mean.

MCAT Z score

Before applying to law school in the US, students need to take an exam called the LSAT. Before applying to medical school, students need to take an exam called the MCAT. Here are some summary statistics for each exam:

Exam	Mean	Standard deviation
LSAT	$\mu = 151$	$\sigma = 10$
MCAT	$\mu = 25.1$	$\sigma = 6.4$

LSAT

$\frac{172 - 151}{10}$

2.1 sto. dev. above mean

MCAT

$\frac{37 - 25.1}{6.4}$

$\frac{11.9}{6.4} < 2$

~ 1.86

Juwan took both exams. He scored 172 on the LSAT and 37 on the MCAT.

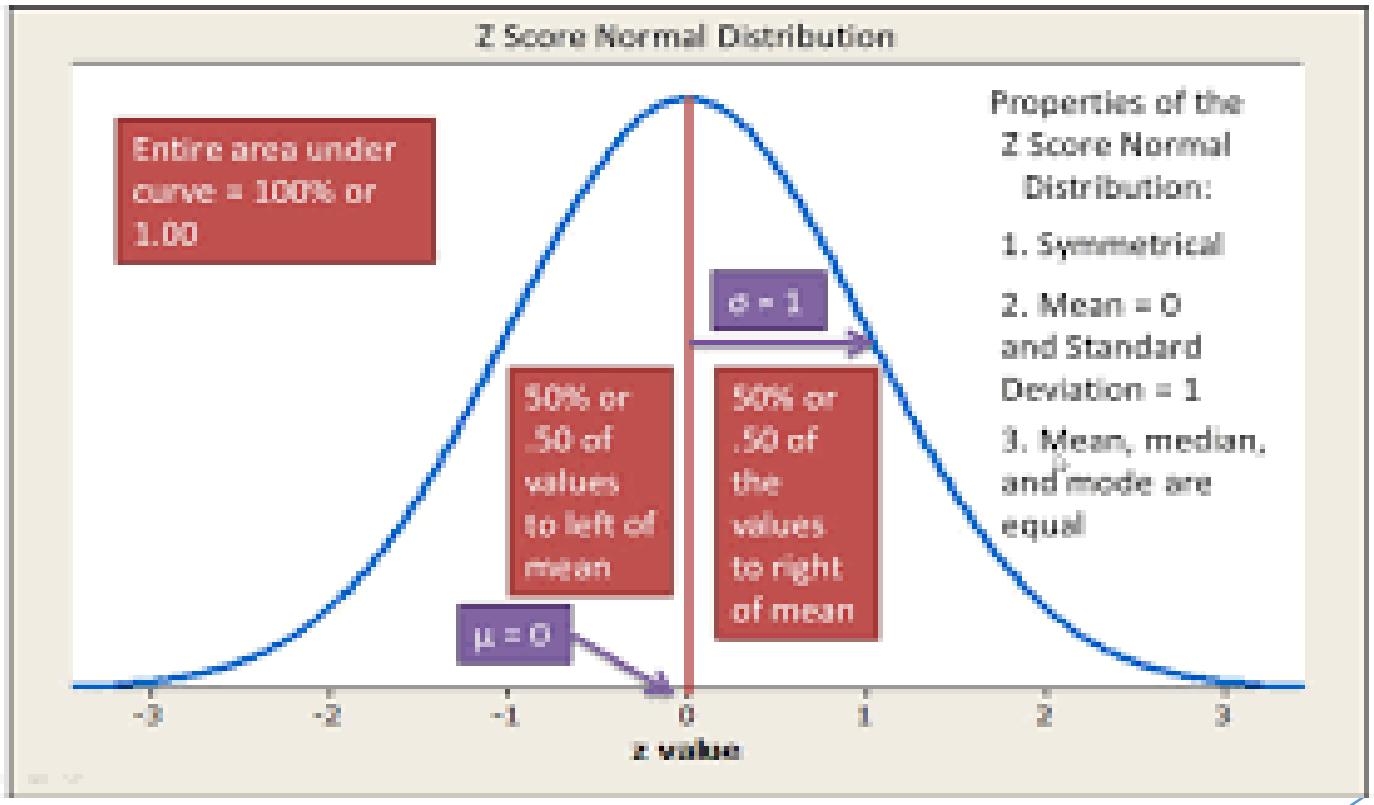
Which exam did he do relatively better on?

So relatively speaking, he did slightly better on the LSAT.

Relatively scoring better from two exams

- We can say the student did relatively well in LSAT where he achieved 2.1
- above the mean.

- Let's understand this concept with an example where the **null hypothesis** is that it is common for students to score 68 marks in mathematics.
- Let's define the significance level at 5%. If the p-value is less than 5%, then the null hypothesis is rejected and it is not common to score 68 marks in mathematics.
- Let's get the z-score of 68 marks:
- >>> zscore = (68 - classscore.mean()) / classscore.std()
- >>> zscore
- 2.283


Z Score Normal Distribution
Entire area under curve = 100% or 1.00
σ = 1
50% or .50 of values to left of mean
50% or .50 of the values to right of mean
μ = 0
Properties of the Z Score Normal Distribution:
1. Symmetrical
2. Mean = 0 and Standard Deviation = 1
3. Mean, median, and mode are equal
-3
-2
-1
0
1
2
3
z value

Central limit theoram

- The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all samples are identical in size), regardless of population distribution shape.
- CLT is a statistical theory that states that given a sufficiently large sample size from a population with a finite level of variance, the mean of all samples from the same population will be approximately equal to the mean of the population.
- all the samples will follow an approximate normal distribution pattern, with all variances being approximately equal to the variance of the population divided by each sample's size.

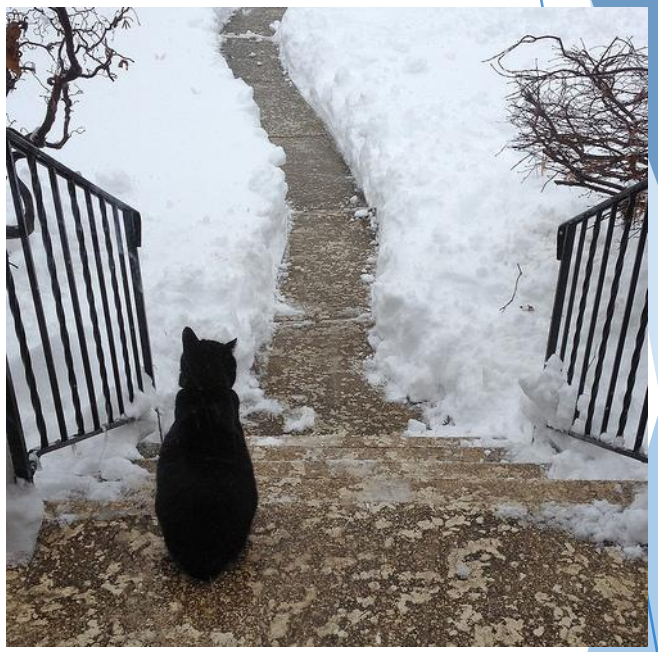
Hypothesis Testing

- An **objective** method of making decisions or **inferences** from sample data (evidence)

Hypothesis testing

- Sample data used to choose between two choices i.e. **hypotheses** or statements about a population

- We typically do this by comparing what we have observed to what we expected if one of the statements (**Null Hypothesis**) was true



Hypothesis testing Framework

- Always two hypotheses:

H_A: Research (Alternative) Hypothesis

- What we aim to gather evidence of
- Typically that there **is** a difference/effect/relationship etc.

H_0: Null Hypothesis

- What we assume is true to begin with
- Typically that there is **no** difference/effect/relationship etc.

Could try explaining things in the context of "The Court Case"?

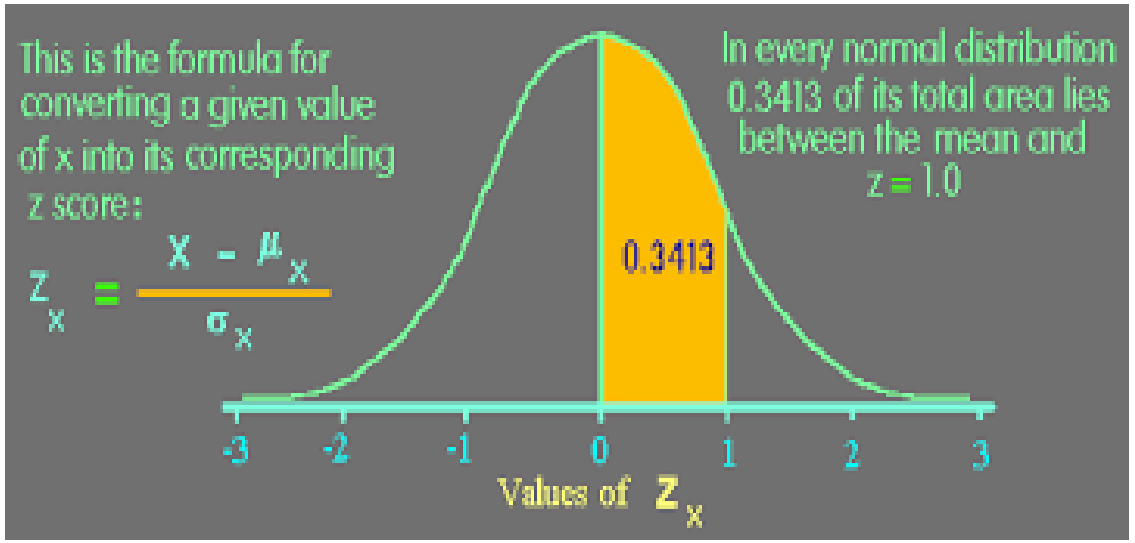
- Members of a jury have to decide whether person
- is guilty or innocent based on evidence

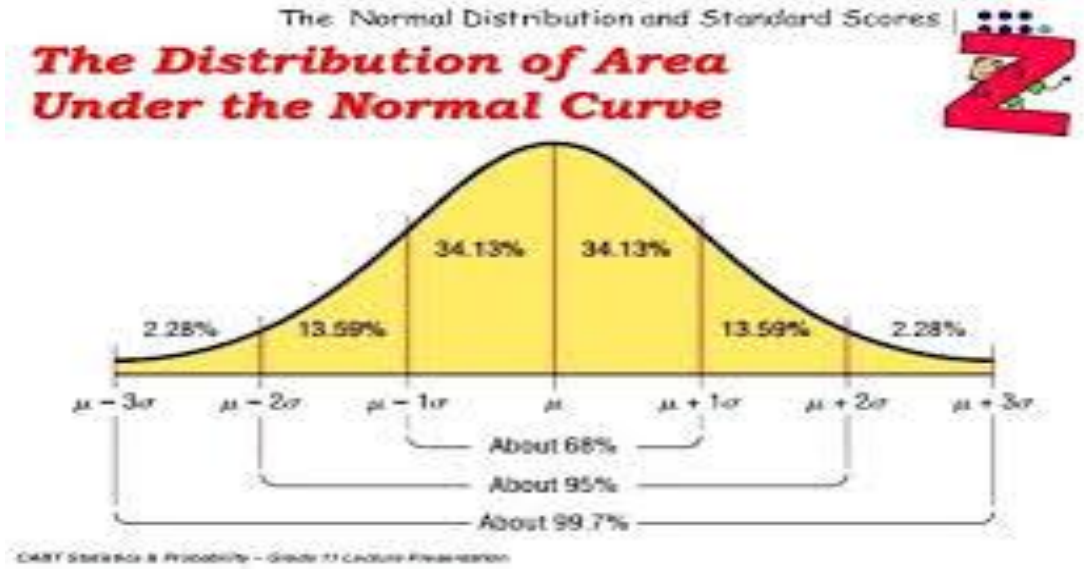
Null: The person is innocent

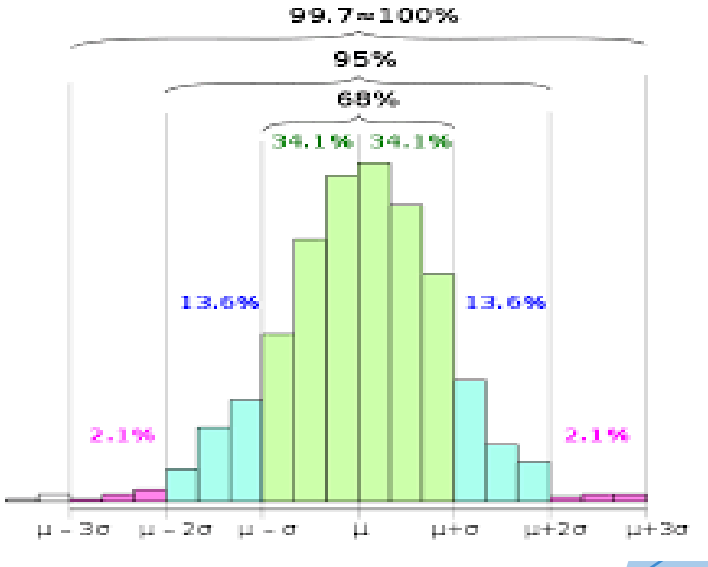
Alternative: The person is not innocent (i.e. guilty)

- The null can only be rejected if there is enough evidence to doubt it
- i.e. the jury can only convict if there is beyond reasonable doubt for the null of innocence
- They do not know whether the person is really guilty or innocent so they may make a mistake

Area under the normal distribution







Types of Errors

Controlled via sample size (=1-Power of test)

Typically restrict to a 5% Risk = level of significance

	Study reports **NO** difference (Do not reject H_0)	Study reports **IS** a difference (Reject H_0)
H_0 is true Difference Does **NOT** exist in population	✓	X Type I Error
H_A is true Difference **DOES** exist in population	X Type II Error	✓

Prob of this = Power of test

TESTS OF HYPOTHESES

- Type I error: H0 is rejected, although it's true ("false alarm").
- Type II error: H0 isn't rejected, although it's false.

Type I error: H_0 is rejected, although it's true ("false alarm").

Type II error: H_0 isn't rejected, although it's false.

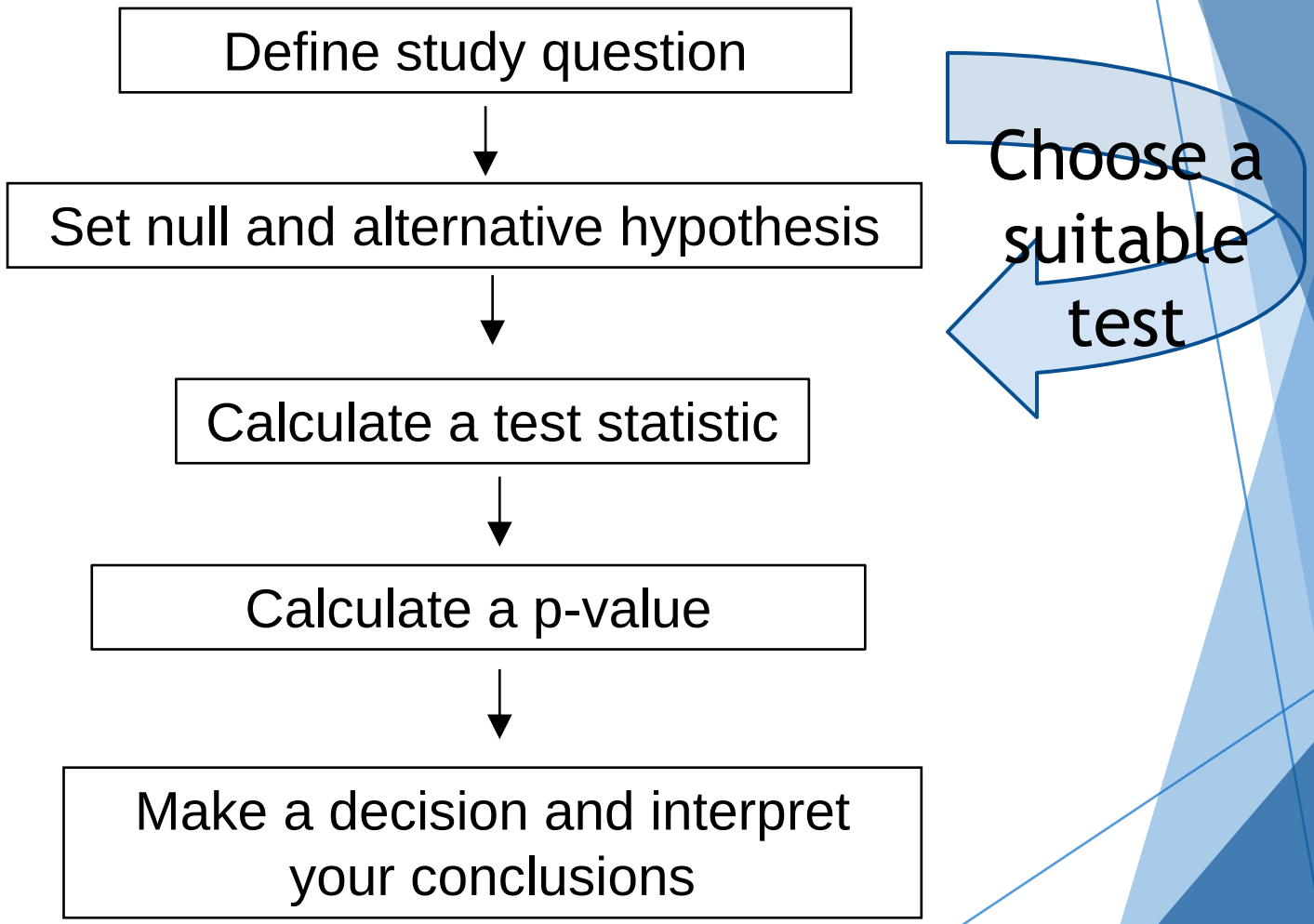
The actual attributes of the population distribution(s) and the error types divide the results to four cases:

	H_0 is true	H_0 is false
H_0 isn't rejected	The right decision	Type II error
H_0 is rejected	Type I error	The right decision

The probability of type I error is called the *risk* or the *level of significance* of the test and it is often denoted by α. The greatest allowed level of significance α is often a starting point of hypothesis testing.

The probability of type II error can't often be calculated, for H_0 may be false in many ways. Often some sort of an (over) estimate is calculated by assuming a typical relatively insignificant way for H_0 to break down. This probability is usually denoted by β. The value $1 - \beta$ is called the *power* of the test. The more powerful a test is, the smaller deviation it notices from H_0.

Steps to undertaking a Hypothesis test



Hypothesis Testing: Decision Rule

- A statistical hypothesis test may return a value called p or the p-value. This is a quantity that we can use to interpret or quantify the result of the test and either reject or fail to reject the null hypothesis. This is done by comparing the p-value to a threshold value chosen beforehand called the significance level.
- The significance level is often referred to by the Greek lower case letter alpha.
- A common value used for alpha is 5% or 0.05. A smaller alpha value suggests a more robust interpretation of the null hypothesis, such as 1% or 0.1%.

A p-value

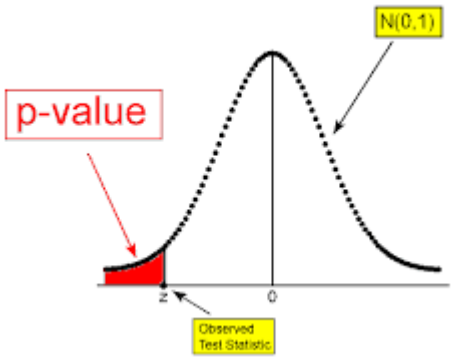
- We can use statistical software to undertake a hypothesis test
- One part of the output is the p-value (P)

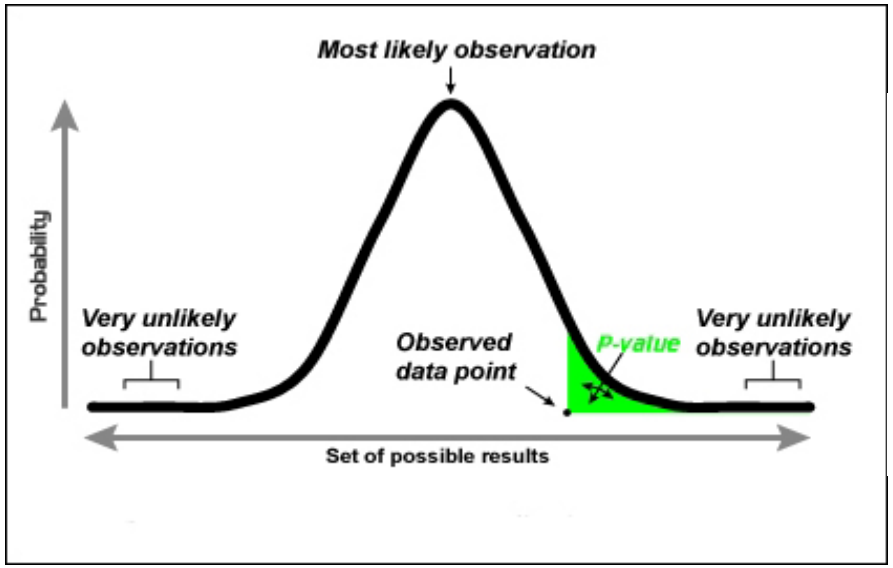
- If P **< 0.05 reject** H_0 => **Evidence** of H_A being true (i.e. **IS** association)

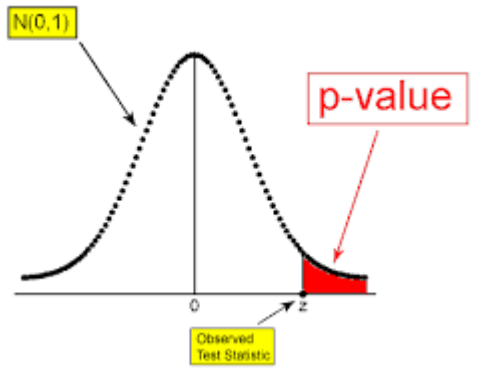
- If P **> 0.05 do not** reject H_0 (i.e. **NO** association)

P-value

- A p-value is the probability of rejecting a null-hypothesis when the hypothesis is proven true. The null hypothesis is a statement that says that there is no difference between two measures. If the hypothesis is that people who clock in 4 hours of study everyday score more that 90 marks out of 100. The null hypothesis here would be that there is no relation between the number of hours clocked in and the marks scored.

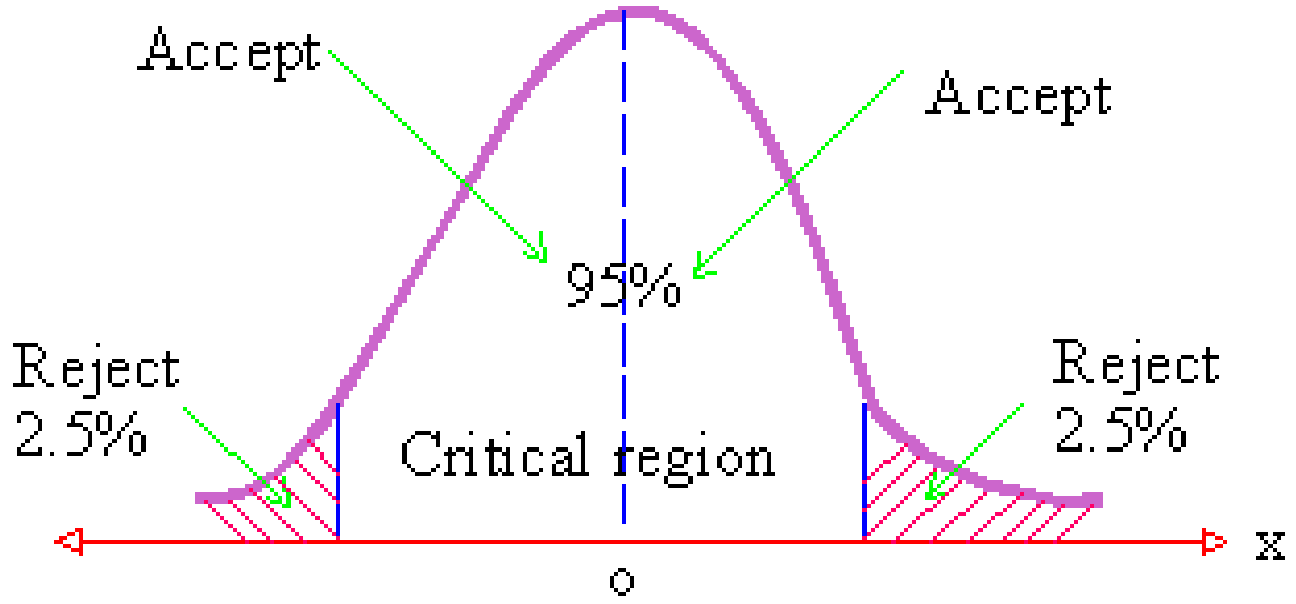






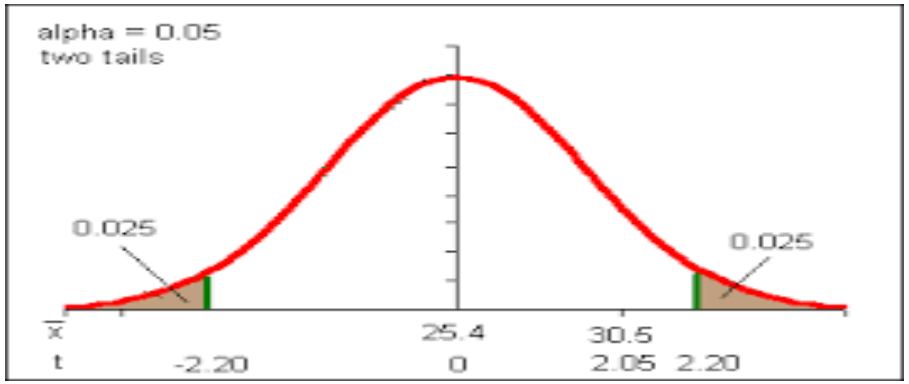
Confidence level =1-Significance level

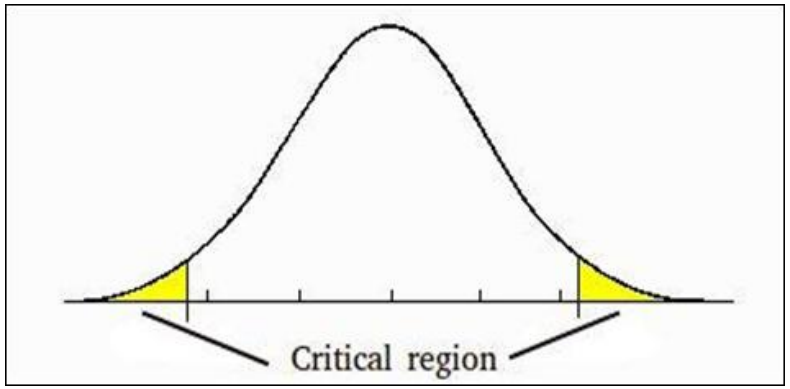
where 1-.05= .95 i.e. 95% (Accept)

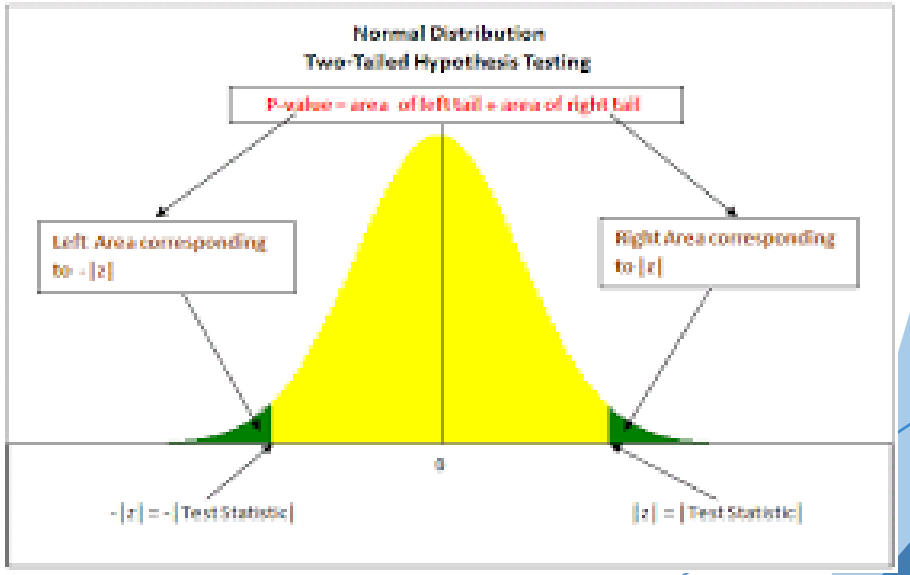


One-tailed and two-tailed tests

- In a two-tailed test, both the tails of the null hypothesis are used to test the hypothesis.





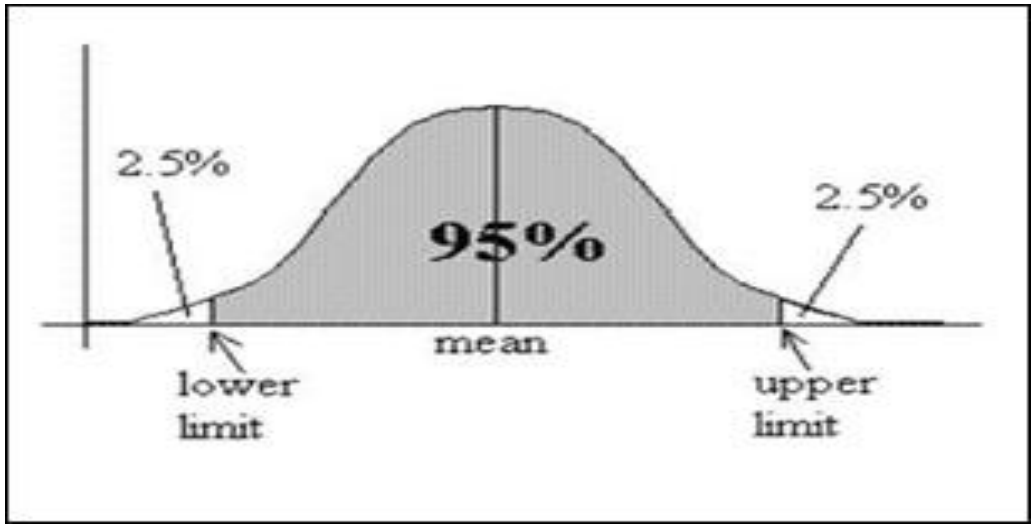


- In a two-tailed test, when a significance level of 5% is used, then it is distributed equally in the both directions, that is, 2.5% of it in one direction and 2.5% in the other direction.

Confidence level

- The significance level can be inverted by subtracting it from 1 to give a confidence level of the hypothesis given the observed sample data.

- confidence level = 1 - significance level



Hypothesis Tests

- ChiSquare test
- Student T test
- Paired T-Test
- Pearson Correlation test
- Spearman correlation test

Chi Square Test

- The Chi-Square test of independence is a statistical test to determine if there is a significant relationship between 2 categorical variables.
- In simple words, the Chi-Square statistic will test whether there is a significant difference in the observed vs the expected frequencies of both variables.

The Chi-Square statistic is calculated as follows:

$$\chi = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The **Null** hypothesis is that there is NO association between both variables.
- The **Alternate** hypothesis says there is evidence to suggest there is an association between the two variables.
- In our case, we will use the Chi-Square test to find which variables have an association with the Survived variable. If we reject the null hypothesis, it's an important variable to use in your model.
- To reject the null hypothesis, the calculated P-Value needs to be below a defined threshold. Say, if we use an alpha of .05, if the p-value < 0.05 we reject the null hypothesis. If that's the case, you should consider using the variable in your model.

Rules to use the Chi-Square Test:

- 1. Variables are Categorical
- 2. Frequency is at least 5
- 3. Variables are sampled independently
- Chi-Square Test in Python
- We will now be implementing this test in an easy to use python class we will call ChiSquare. Our class initialization requires a panda's data frame which will contain the dataset to be used for testing. The Chi-Square test provides important variables such as the P-Value mentioned previously, the Chi-Square statistic and the degrees of freedom.
- `from scipy.stats import chi2_contingency`
- from scipy.stats import chisquare

Parametric Statistical Significance Tests

- Parametric statistical tests assume that a data sample was drawn from a specific population distribution.
- They often refer to statistical tests that assume the Gaussian distribution. Because it is so common for data to fit this distribution, parametric statistical methods are more commonly used.

- In general, each test calculates a test statistic that must be interpreted with some background in statistics and a deeper knowledge of the statistical test itself. Tests also return a p-value that can be used to interpret the result of the test. The p-value can be thought of as the probability of observing the two data samples given the base assumption (null hypothesis) that the two samples were drawn from a population with the same distribution.

- The p-value can be interpreted in the context of a chosen significance level called alpha. A common value for alpha is 5%, or 0.05. If the p-value is below the significance level, then the test says there is enough evidence to reject the null hypothesis and that the samples were likely drawn from populations with differing distributions.
- **p <= alpha**: reject null hypothesis, different distribution.
- **p > alpha**: fail to reject null hypothesis, same distribution.

Student's t-Test

- The Student's t-test is a statistical hypothesis test that **two independent data** samples known to have a Gaussian distribution, have the same Gaussian distribution,
- The assumption or null hypothesis of the test is that the means of two populations are equal. A rejection of this hypothesis indicates that there is sufficient evidence that the means of the populations are different, and in turn that the distributions are not equal.
- **Fail to Reject H0**: Sample distributions are equal.
- **Reject H0**: Sample distributions are not equal.
- The Student's t-test is available in Python via the ttest_ind() SciPy function. The function takes two data samples as arguments and returns the calculated statistic and p-value.
- stat, p = ttest_ind(data1, data2)

Paired Student's t-Test

- We may wish to compare the means between two data samples that are related in some way

- The default assumption, or null hypothesis of the test, is that there is no difference in the means between the samples. The rejection of the null hypothesis indicates that there is enough evidence that the sample means are different.
- **Fail to Reject H0**: Paired sample distributions are equal.
- **Reject H0**: Paired sample distributions are not equal.
- The paired Student's t-test can be implemented in Python using the ttest_rel() SciPy function.
- stat, p = ttest_rel(data1, data2)

Measures of Relationship

- **Covariance:** Covariance is a measure of the relationship between the variability of 2 variables i.e It measures the degree of change in the variables, when one variable changes, will there be the same/a similar change in the other variable.

Covariance

A population covariance is

$$Cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

where x_i and y_i are the observed values, μ_x and μ_y are the population means, and N is the population size.

A sample covariance is

$$Cov(x,y) = s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

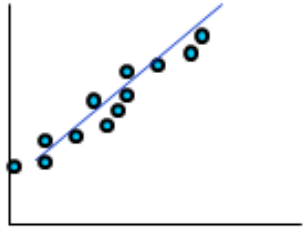
where x_i and y_i are the observed values, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size.

Correlation Coefficient r

- **Measures strength of a relationship between two continuous variables**

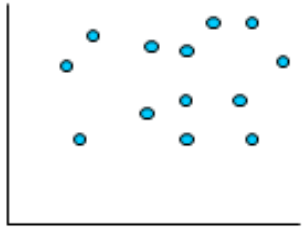
$-1 \leq r \leq 1$

Strong positive linear relationship



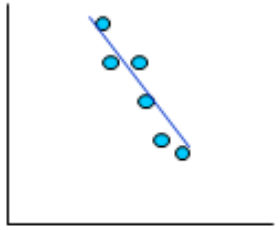
$r = 0.9$

No linear relationship



$r = 0.01$

Strong negative linear relationship



$r = -0.9$

Correlation

- **Correlation:** Correlation gives a better understanding of covariance. It is normalized covariance. Correlation tells us how correlated the variables are to each other. It is also called as Pearson Correlation Coefficient.

$$Correlation = \rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

- The value of correlation ranges from -1 to 1. -1 indicates negative correlation i.e with an increase in 1 variable independent there is a decrease in the other dependent variable.1 indicates positive correlation i.e with an increase in 1 variable independent there is an increase in the other dependent variable.0 indicates that the variables are independent of each other.

Correlation Interpretation

An interpretation of the size of the coefficient has been described by Cohen (1992) as:

Correlation coefficient value			Relationship
-0.3 to +0.3			Weak
-0.5 to -0.3	or	0.3 to 0.5	Moderate
-0.9 to -0.5	or	0.5 to 0.9	Strong
-1.0 to -0.9	or	0.9 to 1.0	Very strong

Cohen, L. (1992). Power Primer. Psychological Bulletin, 112(1) 155-159

Calculating correlation

Height	Weight	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
5	45	−0.14	−5	0.7	0.019	25
5.5	53	−0.36	3	−1.08	0.129	9
6	70	0.86	20	17.2	0.739	400
4.7	42	−0.44	−8	3.52	0.193	64
4.5	40	−0.64	−10	6.4	0.409	100

Sum(Height) = 25.7 Mean(Height) = 5.14

Sum(Weight) = 250 Mean(Weight) = 50

$$\sum(x-\bar{x})(y-\bar{y}) = 26.74$$

$$\sum(x-\bar{x})^2 = 1.489$$

$$\sum(y-\bar{y})^2 = 598$$

$$\text{Correlation} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{26.74}{\sqrt{1.489}\sqrt{598}} = \frac{26.54}{1.220 * 24.454} = 0.889$$

Correlation 0.889 tells us Height and Weight has a positive correlation. It is obvious that as the height of a person increases weight too increases.

- cov(X, Y) = (sum (x - mean(X)) * (y - mean(Y))) * 1/(n-1)
- # calculate covariance matrix
- covariance = cov(data1, data2)
- print(covariance)

Pearson's Correlation

- Pearson's correlation coefficient = covariance(X, Y) / (stdv(X) * stdv(Y))
- from scipy.stats import pearsonr
- from scipy.stats import spearmanr

ANOVA Test

- Analysis of variance (ANOVA) uses F-tests to statistically assess the equality of means when you have three or more groups.
- The term F-test is based on the fact that these tests use the F-statistic to test the hypotheses.
- An F-statistic is the ratio of two variances and it was named after Sir Ronald Fisher.
- Variances measure the dispersal of the data points around the mean. Higher variances occur when the individual data points tend to fall further from the mean

- https://statisticsbyjim.com/hypothesis-testing/t-tests-t-values-t-distributions-probabilities/
- https://statisticsbyjim.com/anova/f-tests-anova/

F Test – (Fisher test)- In One Way ANOVA

An F-statistic is the ratio of two variances, or technically, two mean squares. Mean squares are simply variances that account for the degrees of freedom (DF) used to estimate the variance.

Variances are the sum of the squared deviations from the mean.

$$F = \frac{\text{between-groups variance}}{\text{within-group variance}}$$

11.2 Correlation coefficient

The scatterplot provides a visual impression of the nature of relation between the x and y values in a bivariate data set. In a great many cases the points appear to band around the straight line. Our visual impression of the closeness of the scatter to a linear relation can be quantified by calculating a numerical measure, called the **sample correlation coefficient**

DEFINITION 11.1 (Correlation coefficient). *The sample correlation coefficient, denoted by r (or in some cases r_{xy}), is a measure of the strength of the linear relation between the x and y variables.*

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{10}$$

$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}} \tag{11}$$

$$= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \tag{12}$$

$$= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}, \tag{13}$$

where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = (n-1)s_x^2,$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = (n-1)s_y^2,$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

The quantities S_{xx} and S_{yy} are the sums of squared deviations of the x observed values and the y observed values, respectively. S_{xy} is the sum of cross products of the x deviations with the y deviations.

References

- Khanacademy.com