**Statistic**

**Population and Sample**

**Population:** collection of all items of interest (N) – are called parameters

**Sample** : A subset of the population (n). fields which we worked called statistic

Less time consuming, Less costly (cheaper)

|  |  |
| --- | --- |
| **Population** | **Sample** |
| Hard to observe |  |
| Hard to contact | Easy to observe |
|  | Easy to contact |

**Sample**

|  |  |
| --- | --- |
| **Randomness** | **Representativeness** |
| A random sample is collected when each member of the sample is chosen from the population strictly by change | A representative sample is a subset of the population that accurately reflects the members of the entire population. |
| Not chosen by change as most student were not even there – Violated – Not Random | It represents only NYU Students who eat in the canteen - Violated |

**Descriptive Statistics**

**Types of Data**

**Categorical and Numerical**

**Categorical:** describes categories for group.

Exp :- Audi, Mercedes or Yes – No Question

**Numerical : Discrete and Continuous**

**Discrete:** # of Children

**Examples of Discrete**: - Grade of a Student A,B,C,D,E,F or 0 to 100%, Number of Objects- Soda botels #1000 and Money

**Continuous :-** Weight of a person (Time on a Clock in Discrete), Time in General is Continuous

A picture containing tennis

Description automatically generated

**Level of Measurement**

**A picture containing text

Description automatically generated**

**Nominal** : Brand of Cars Mercedez, Audi, BMW, Four Season- Winter, Summer, Sprint, Odom

**Ordinal:** strictly order (Disgusting – Unappetizing – Neutral – Tasty – Delicious)

**Interval:** Don’t have a True 0

**Ratio:** has a True 0

**Visualization Techniques**

A screenshot of a cell phone

Description automatically generated

A Pareto diagram is special type of bar chart, where categories are shown in descending order of frequency.

80% of the effect come from 20% of the bugs (80-20 Rule)

Microsoft fixing 20% of its software bugs, they manage to solve 80% of the problems

It shows how subtotals change with each additional category and provide us with a better understanding of our data.

**Numerical Data**

1. **Cross Table (Bar Graph)**
2. **Scatter Plot**

**Measures of Central Tendency**

Means is average. For Population is called Meu and for sample is called x\_bar

By adding up all the components and then dividing by the number of components.

Median is the middle number in an ordered dataset. The median is the number at position (n+1)/2 in the ordered list.

Mode is the value than occurs most often

Which measure is best?

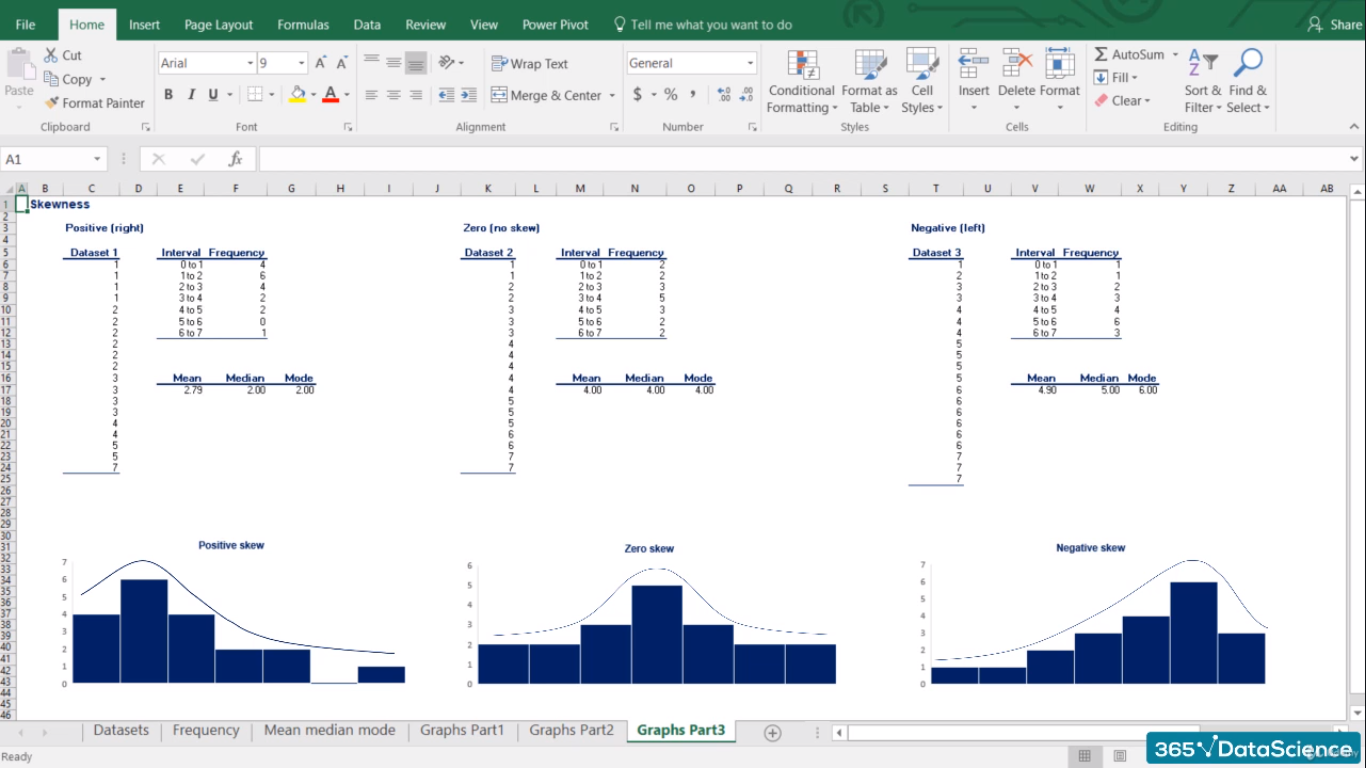
There is no best, but using only one is definitely the worst!

**Skewness**

**Measures of Asymmetry:**

**Almost always, you will use software to compute the Skewness**

Skewness indicates whether the data is concentrated on one side



If mean > median then positive skew and outliers is in right side

If mean< median then negative skew and outliers is in left side

If mean = median then it’s zero skew

**Variance**

Variance (σ2) measure the dispersion of a set of data point around their mean

A close up of a sign

Description automatically generated

* Dispersion is non-negative. Non-negative values don’t cancel out
* Amplifies the effect of large differences

**Standard Deviation**

**A screenshot of a cell phone

Description automatically generated**

**Standard Deviation (σ) is the most common measure of variability of a SINGLE DATASET.**

**Coefficient of Variance (CV) = StdDev/Mean**

**Covariance**

The two variables are correlated and the main statistic to measure this correlation is called covariance.

A screenshot of a cell phone

Description automatically generated

**Correlated Coefficient:-** Correlation adjusts covariance, so that the relationship between the two variable becomes easy and intuitive to interpret

Cov(x,y)/(Stdev(x) \* Stdev(y))

Correlation of 0 🡪 Absolutely independent variables (Coffee in Brazil and Coffee in London)

Correlation between x and y is similar to y and x (Symmetrical with respect to both variables)

**Causality:** Important to understand the direction of causal relationships.

Correlation does not imply causation.

**Practical Exercise**



**Statistics – Inferential Statistics Fundamentals**

A distribution is a function that shows the possible values for a variable and how often then occur.

Rolling a die- Probability of getting 1 – 1/6

Discrete Uniform Distribution- All event has equal chances of occurring

**Normal Distribution:**

* They approximate a wide variety of random variables
* Distributions of sample means with large enough sample size could be approximated to normal
* All computable statics are elegant
* Decisions based on normal distribution insights have a good track goods

A close up of a map

Description automatically generated

A smaller mean would move the group to the left

A bigger mean would move the graph to the right

A close up of a map

Description automatically generated

**Standardization:-**

**Standardized Normal Distribution:-** when mean is zero and standard deviation is one.

Z = (x - )/σ

A close up of a logo

Description automatically generated

**Central Limit Theorem**

A screenshot of a cell phone

Description automatically generated

CLT allows us to performs tests, solve problems and make inferences using the Normal distribution, even when the population is not normally distributed.

**Standard Error:** The Standard Deviation of the distribution formed by sample means

* Used in most statistical tests
* Because it shows how well you approximated the true mean

Standard error decreases when sample size increases.

**ESTIMATORS AND ESTIMATES**

**Two types of Estimates** :- Point Estimates (a single number) and Confidence Intervals

**Properties** :- Efficiency and Bias

**Estimators** :- are judges

**Unbiases Estimator:** expected value = population parameter

**Efficiency:** The most efficient estimator is the unbiased estimator with smallest variance

|  |  |
| --- | --- |
| **Statistics** | **Estimators** |
| Broader term | A type of statistic |

**Statistics – Inferential Statistics Confidence Intervals**

Confidence Level : 0 < α < 1

Point Estimate +- Reliability Factor \* Standard Error

Confidence Intervals Population Variance for Known Audience:

A screenshot of a cell phone

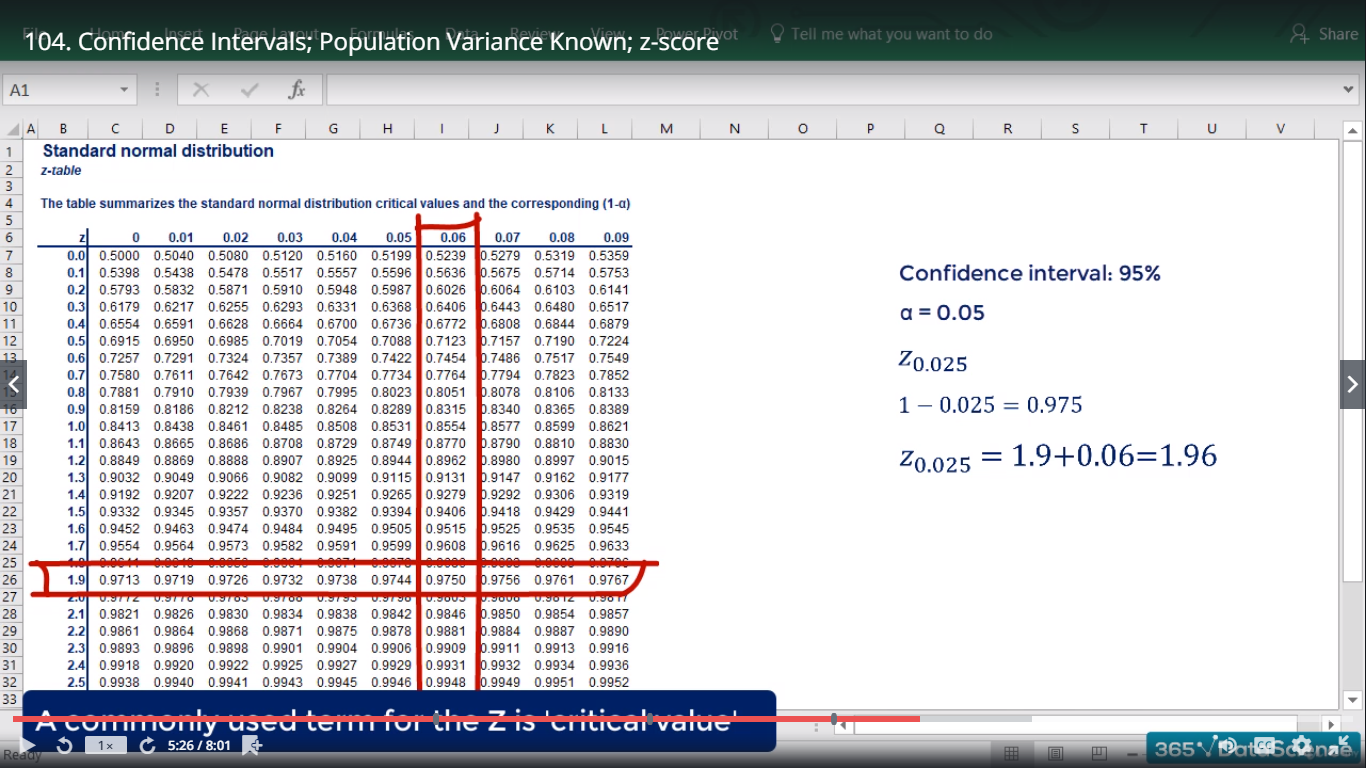
Description automatically generated

A 95% confidence interval means that you are sure that in 95% of the cases, the true population parameter would fall into the specified interval.

Common confidence Level = 90% 95% 99%

α = 0.1 0.05 0.01

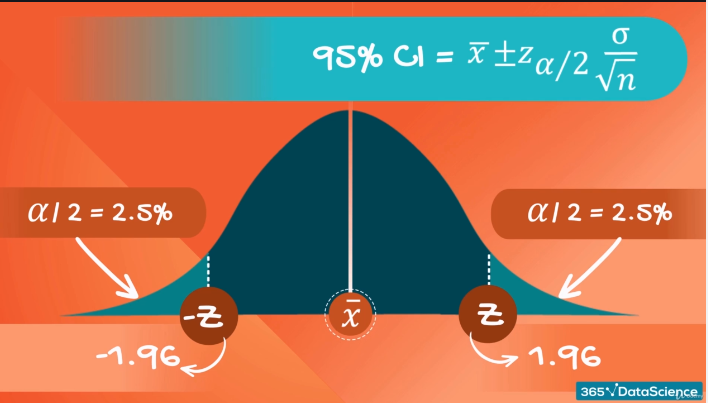




A screenshot of a cell phone

Description automatically generated

**Confidence Interval Clarifications:**

****

**A picture containing person, baseball

Description automatically generated**

**95% is the accepted norm, as we don’t compromise with accuracy too much, but still get a relatively narrow interval.**

**Student t distribution:**

* Inference through small samples
* Unknown population variance
* Hugh real-life application

A close up of text on a white background

Description automatically generated

A close up of a logo

Description automatically generated

Sample size = n

Degree of freedom = n-1

After 30 number of observations the t- statistic table become Z – statistic.



A close up of a sign

Description automatically generated

Sample Size (n) increase, ME decrease, Confidence Interval decrease

**Confidence Interval for 2 Populations**

Dependent and Independent

**Dependent:** When we are researching the same subject over time (Weight loss and blood samples)

|  |  |
| --- | --- |
| **Dependence** | **Independence** |
| Before and after situation | Population variance known |
| Cause and effect | Population variance unknown but assumed to be equal |
|  | Population variance unknown but assumed to be different |

**Dependent Samples:**  Effect of Magnesium in new drug - Magnesium level before and after

A screenshot of a cell phone

Description automatically generated

**Independent - Known Population Variance**

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated



**Population variance unknown but assumes to be equal**

**A screenshot of a cell phone

Description automatically generated**

The degree of freedom are equal to the total sample size minus the number of variables. Nx + ny – 2



**Practical Inferential Statistics**

**Hypothesis Testing**

A hypothesis is an idea that can be tested.

|  |  |
| --- | --- |
| **Hypothesis** | **Notation** |
| Null Hypothesis | H0 |
| Alternative hypothesis | H1 or HA |

**Example (Two sided Test)**

H0 : μ0 = 113000

H1 : μ0 <> 11300

Accept if : x\_bar is close enough to the true mean

Reject if : x\_bar is too far from the true mean

**Example (One sided Test)**

H0 : μ0 >= 125000

H1 : μ0 < 12500

The Researcher is trying to REJECT the null hypothesis.

The NULL HYPOTHESIS is the statement we are trying to reject. Therefore the NULL is the present state of affairs while the alternative is our personal opinion.

**Significance level:** The probability of rejecting the null hypothesis, if it is true. Typical values for alpha are: 0.01, 0.05 (most common), 0.1

**Exp:**

**University Dean** : Population mean grade is 70%

H0 : μ0 = 70%

H1 : μ0 <> 70%

A picture containing road

Description automatically generated

A picture containing bus, road

Description automatically generated

**One Sided test**

**A close up of a sign

Description automatically generated**

**If the test statistic is bigger than the cut-off z-score we would reject the null, otherwise we wouldn’t.**

**Errors in Hypothesis Testing**

Type I Error: Reject a true null hypothesis (Alpha) (False Positive)

Type II Error: Accept a false null hypothesis (beta)

Rejecting a false null hypothesis (Probability = 1 – Beta)

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

**We know:** How to test hypothesis, How to reject them, at various levels of significance

**We don’t know**: A level of significance after which we can no longer do it.

**p-value:-** p-value is the smallest level of significance at which we can still reject the null hypothesis, given the observed sample statistic.

**Population Variance known**

**Exp:**

Standard Error = 2739, Population Std = 15000,

N ~ (μ, σ2), n = 30

Z = -4.67

We rejected the null at 0.05 and 0.01

Rule : you should reject the null hypothesis, if

p-value < Alpha

A picture containing screenshot

Description automatically generated

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Where and How are p-values used?

* Most statistical software calculates p-values for each test
* Research decides significance post-factum
* P-values are usually found with 3 digits after the dot x.xxx
* The closed to 0.000 the p-value, the better

A close up of a sign

Description automatically generated

**Population Variance Unknown**

H0 : μ0 <= 40%

H1 : μ0 > 40%

A screenshot of a cell phone

Description automatically generated

**Decision Rule:**

**Accept if: p-value > Significance level**

**Reject if: p-value < Significance level**

**Multiple Populations**

**Dependent Samples**

**A close up of a sign

Description automatically generated**



**Independent Samples, Known Variance**

**Management & Engineering Student Average Grade**

**A screenshot of a cell phone

Description automatically generated**

**Z-value = -.244**

**p-value = 0.015**

**0.015 < 0.05 = We reject the Null Hypothesis**

**There is enough statistical evidence that the mean difference it not 4%**

**Independence Samples, Unknown Variances but assumed to be equal**

****



**Practical Example 3**

