Applied Statistics – Week 1

Definition :-

Terms Involved

- **Population : -** It Is A Representative Sample Of A Larger Group Of People (Or Even Things) With One Or More Characteristics In Common.
- **Sample : -** It Is An Analytic Subset Of A Larger Population
- **Parameter : -** It Is A Number Describing A Whole Population (E.G., Population Mean), While A **Statistic** Is A Number Describing A Sample (E.G., Sample Mean).
- **Statistics : -**Branch Of Mathematics Dealing With The Collection, Analysis, Interpretation, And Presentation Of Masses Of Numerical Data

Measures of central tendancy – 3M

Property	Formula	Major Points
Mean	x = Σx / n	Sensitive to Outliers. More useful when data is symmetrical.
Median	For n = odd; Median = (n + 1) / 2 For n = even; Median = Avg. of n/2 and (n + 1) / 2	Robust to outliers, useful when the data is skewed
Mode	Highest repeating value in dataset	Finds central value. Helpful in categorical features

Measures of Variation

Property	Formula	Major Points
Sample Variance	$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$	
Sample Standard Deviation	$s = \sqrt{\frac{\Sigma_{i=1}^{N}(x_i-\bar{x})^2}{N-1}}$	Square root of variance. Robust to outliers. Commonly used
IQR	IQR = Q3 – Q1	Less sensitive to outliers
Range	Highest Value - Lowest Value	Highly sensitive to unusual values. Not used often

Applied Statistics – Week 1

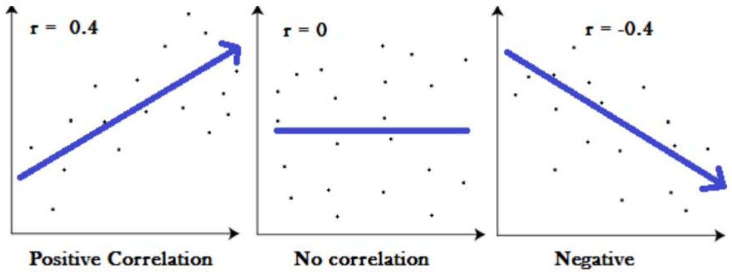
Measures of Relative Position

Property	Formula	Major Points
Percentile	Data is divided into 100 equal parts by increasing order	For applying normal distributions
Quartile	Data is divided into 4 equal parts. For ex. Q3(third is the value greater than ¾ of others.	Used to compute IQR
Z-Score	Data is divided into 4 equal parts. For ex. Q3(third is the value greater than ¾ of others.	Measures distance from mean in terms of standard deviation

Correlation Coefficient - Correlation coefficient formulas are used to find

how strong a relationship is between data. The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all.



Person Correlation –

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

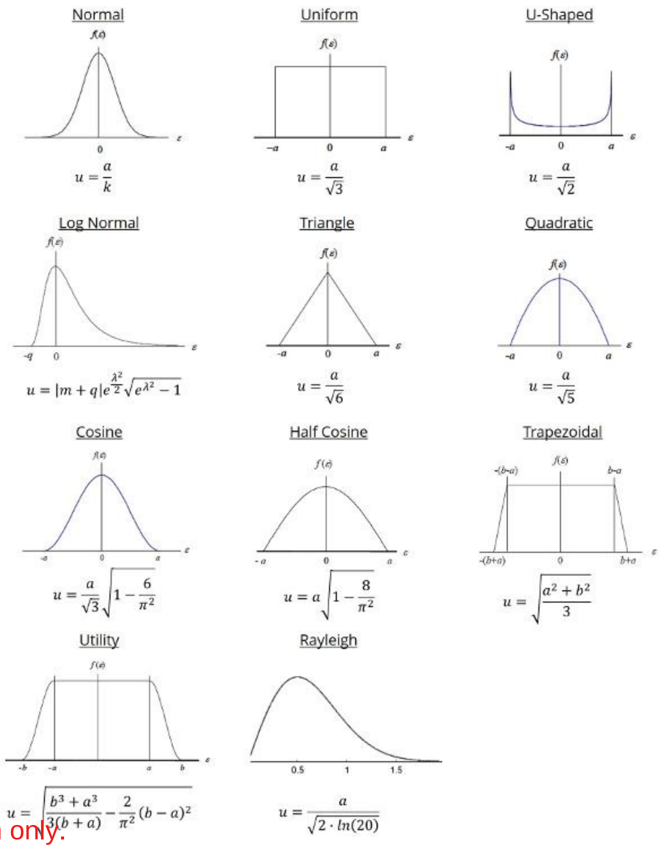
$\bar{y}$ = mean of the values of the y-variable

Applied Statistics – Week 2 - Probability

- Population - It is a representative sample of a larger group of people (or even things) with one or more characteristics in common.
- Sample - It is an analytic subset of a larger population.
- Event - A probability event can be defined as a set of outcomes of an experiment.

Rules for Probability Distribution	
Rule	**Formula**
Complement Rule	P(A) = 1 – P(A’)
Multiplication Rule (Dependant)	P (A ∩ B) = P(A) * P (B | A)
Multiplication Rule (Independent)	P (A ∩ B) = P(A) * P(B)
Addition Rule – Not Mutally Exclusive	P (A ∪ B) = P(A) + P(B) – P ((A ∩ B)
Mutually Exclusive	P (A ∪ B) = P (A) + P (B)
Conditional Probability	P (A | B) = P (A ∩ B) / P (B)
Bayes Theorem	P (A | B) = P (B | A) * P(A) / P (B)

Rules for Probability Distribution



Poisson Distribution

The Poisson distribution is a probability distribution that represents the number of times an event occurs in a fixed time and/or space interval and is defined by parameter λ (lambda).

Examples of events that can be described by the Poisson distribution include the number of bikes crossing an intersection in a specific hour and the number of meteors seen in a minute of a meteor shower.

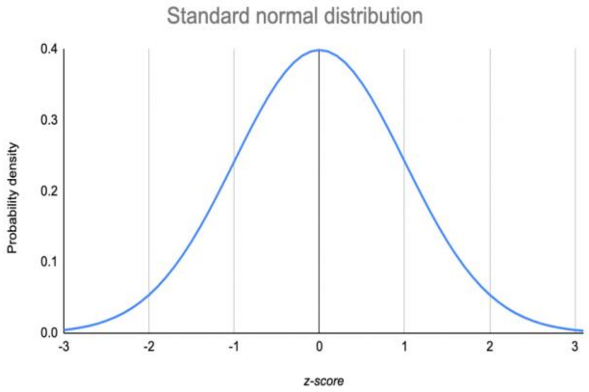
Mathematically:

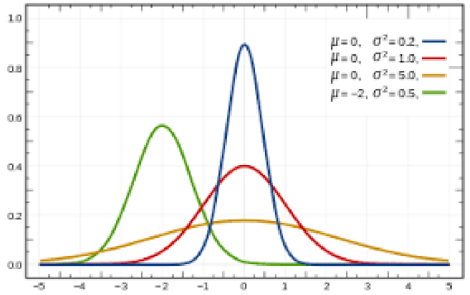
$$X \sim \text{Binomial}\ (n, p), E\ (X) = n \times p$$

$$Y \sim \text{Poisson}\ (\lambda), E(Y) = \lambda$$

Normal Distribution

It is a probability density function that looks like this where σ = 0, μ = 1





Mean = Mode = Median

Applied Statistics – Week 3 – Hypothesis Testing

Hypothesis Term	Definition
Significance Level ($\boldsymbol{\alpha}$)	Defines the strength of evidence in probabilistic terms. Specifically, alpha represents the probability that tests will produce statistically significant results when the null hypothesis is correct
Confidence Level (C)	The percentage of all possible samples that can be expected to include the true population parameter.
Critical Value($\boldsymbol{zc}$)	zc is the critical value of a standard normal distribution under $H0$. Critical values divide the rejection and non-rejection regions. Set using p-values or to a threshold value of 0.05 (5%) or 0.01 (1%), but always ≤ 0.10 (10%)
Test Statistic ($\boldsymbol{zdata}$)	$zdata$ is the test value of z of a standard normal distribution under $H0$. If $zdata$ is inside the rejection region, demarked by zc , then we can reject the null hypothesis, $H0$.
p-value	Probability of obtaining a sample “more extreme” than the ones observed in your data, assuming $H0$ is true.
Hypothesis	A premise or claim that we want to test.
Null Hypothesis: $\boldsymbol{H0}$	Currently accepted value for a parameter. (middle of the distribution) Is assumed true for the purpose of carrying out the hypothesis test. • Always contains an “=“ {=, ≤, ≥} • The null value implies a specific sampling distribution for the test statistic • $\boldsymbol{H0}$ is the middle of the normal distribution curve at z = 0. • Can be rejected, or not rejected, but NEVER supported
Alternative Hypotheses: $\boldsymbol{H}$a	Also called Research Hypothesis or $H1$. Is the opposite of $H0$ and involves the claim to be tested. Is supported only by carrying out the test if the null hypothesis can be rejected. • Always contains “>“ (right-tailed), “), “<“ Left tailed or != two tailed • Can be Supported (By rejecting the null), or not supported (by failing or rejecting the null) but never rejected

Applied Statistics – Week 3 – Hypothesis Testing

Hypothesis Testing	Steps
Hypothesis Testing	1) Formula H0 And Has 2) Graph: Sketch And Label Critical Value (Left-tailed, Right-tailed, Two-tailed) 3) Decision Rule: Use Significance Level (α), Confidence Level (C), Confidence Interval, Or Critical Value (zc). E.G. We Will Reject $H0$ If $zdata$ > 1.645 4) Critical Value: Determine Critical Values (zc) To Mark The Rejection Regions 5) Test Statistic: Calculate The Test Statistic ($zdata$) From The Sample Data 6) Conclusion: Reject The Null Hypothesis (Supporting The Alternative Hypothesis) Otherwise Fail To Reject The Null Hypothesis, Then State Claim

Hypothesis Formulation

If Claim Consist Of	Then Hypothesis Test Is	Represented By
"Is Equal To", "Is Exactly", "Is The Same As", "Is Between" "Is At Least" "Is At Most"	Two-tailed = Left-tailed ≤ Right-tailed ≥	H0
"Is Not Equal To", "Is Different From", "Has Changed From" "Is Less Than", "Is Below", "Is Lower Or Smaller Than" "Is Greater Than", "Is Above", "Is Longer Or Bigger Than" $Make\ sure\ H0 + Ha = all\ possible\ outcome.$	Two-tailed ≠ Left-tailed < Right-tailed >	Ha

Decision Rule

P-value	Use Probability Value To Determine Z In Normal Distribution Table
Significance Level α	Usually At A Threshold Value Of 5% Or 1% But Always ≤ 10% α = 1 − C
Confidence Level (C)	With A Confidence Of 0.95 (95%) Or 0.99 (99%), But Always ≥ 0.90 (90%). $C = 1 - \alpha$
Examples:	E.G. We Will Reject $H0$ If Significance Level Is Less Than 5% E.G. We Will Reject $H0$ If Confidence Level Is Greater Than 95% E.G. We Will Reject $H0$ If Confidence Interval Is Between 5% And 95% ($e.g.$ ±5%) E.G. We Will Reject $H0$ If $zdata > zc$ In A Right-tailed Test

Determine Critical Value

Critical Values (zc)	Determine zc by looking up α, C, or p-values in a standard normal distribution table. Two-tailed tests have two values for zc .

Applied Statistics – Week 3 – Hypothesis Testing

Type	Description
Test For Independence	Tests For The Independence Of Two Categorical Variables
Homogeneity Of Variance	Test If More Than Two Subgroups Of A Population Share The Same Multivariate Distribution
Goodness Of Fit	Whether A Multinomial Model For The Population Distribution (P1,....Pm) Fits Our Data
Assumptions 1. One Or Two Categorical Variables 2. Independent Observations 3. Outcomes Mutually Exclusive 4. Large N And No More Than 20% Of Expected Counts < 5	
Anova Analysis	Comparing The Means Of Two Or More Continuous Popula-tions
One-way Layout	A Test That Allows One To Make Comparisons Between The Means Of Two Or More Groups Of Data.
Two-way Layout	A Test That Allows One To Make Comparisons Between The Means Of Two Or More Groups Of Data, Where Two Indepe-ndent Variables Are Considered.
Assumptions About Data: 1. Each Data Y Is Normally Distributed 2. The Variance Of Each Treatment Group Is Same 3. All Observations Are Independent	