

### Statistics

the branch of mathematics in which data are used descriptively or inferentially to find or support answers for scientific and other quantifiable questions.

It encompasses various techniques and procedures for recording, organizing, analyzing, and reporting quantitative information.

### Difference - parametric test & non-parametric test

PROPERTIES	PARAMETRIC	NON-PARAMETRIC
assumptions	YES	NO
value for central tendency	mean	median/mode
probability distribution	normally distributed	user specific
population knowledge	required	not required
used for	interval data	nominal, ordinal data
correlation	pearson	spearman
tests	t test, z test, f test, ANOVA	Kruskal Wallis H test, Mann-whitney U, Chi-square

### Correlation Coefficient

a statistical measure of the strength of the relationship between the relative movements of two variables

value ranges from **-1 to +1**

-1 = perfect negative or inverse correlation

+1 = perfect positive correlation or direct relationship

0 = no linear relationship

### Alternatives

PARAMETRIC	NON-PARAMETRIC
one sample z test, one sample t test	one sample sign test
one sample z test, one sample t test	one sample Wilcoxon signed rank test
two way ANOVA	Friedman test
one way ANOVA	Kruskal wallis test
independent sample t test	mann-whitney U test
one way ANOVA	mood's median test
pearson correlation	spearman correlation

### Paired t-test

*to compare means of two related groups*

ex. compare weight of 20 mice before and after treatment

*two conditions:*

- pre post treatment
- two diff conditions ex two drugs

#### ASSUMPTIONS

- random selection
- normally distributed
- no extreme outliers

#### FORMULA

$t = \frac{m}{s/\sqrt{n}}$

m = sample mean of differences

df = n - 1

### t-distribution

*aka Student's t-distribution* = probability distribution similar to normal distribution but **has heavier tails**

used to estimate pop parameters for small samples

*Tail heaviness is determined by degrees of freedom* = gives lower probability to centre, higher to tails than normal distribution, also have higher kurtosis, symmetrical, unimodal, centred at 0, larger spread around 0

df = n - 1

above 30df, use z-distribution

**t-score** = no of SD from mean in a t-distribution

we find:

- upper and lower boundaries
- p value

#### TO BE USED WHEN:

- small sample
  - SD is unknown
- #### ASSUMPTIONS
- cont or ordinal scale
  - random selection
  - NPC
  - equal SD for indep two-sample t-test

### Two-sample z-test

*to determine if means of two independent populations are equal or different*

*to find out if there is significant diff bet two pop by comparing sample mean*

knowledge of:

**SD and sample >30 in each group**

eg. compare performance of 2 students, average salaries, employee performance, compare IQ, etc

FORMULA:

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

s= SD

formula:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

$(\mu_1 - \mu_2)$  = hypothesized difference bet pop means

### Point Biserial correlation

*measures relationship between two variables*

**rpbi** = correlation coefficient

**one continuous variable (ratio/interval scale)**

**one naturally binary variable**

FORMULA:

$$rpb = \frac{M1 - M0}{S_n} \cdot \sqrt{pq}$$

$S_n$  = SD

### Two-sample z-test

*to determine if means of two independent populations are equal or different*

*to find out if there is significant diff bet two pop by comparing sample mean*

knowledge of:

**SD and sample >30 in each group**

eg. compare performance of 2 students, average salaries, employee performance, compare IQ, etc

FORMULA:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

### z-test

#### for hypothesis testing

to check whether means of two populations are equal to each other when pop variance is known

we have knowledge of:

- SD/population variance and/or sample  $n=30$  or more

if both unknown -> t-test

**left-tailed**

**right-tailed**

**two-tailed**

### z-test (cont)

**REJECT NULL HYPOTHESIS IF Z STATISTIC IS STATISTICALLY SIGNIFICANT WHEN COMPARED WITH CRITICAL VALUE**

z-statistic/ z-score = no representing result from z-test

z critical value divides graph into acceptance and rejection regions

if z stat falls in rejection region ->  $H_0$  can be rejected

**TYPES**

One-sample z-test

Two-sample z-test

### ANOVA

#### Analysis of Variance

*comparing several sets of scores*

*to test if means of 3 or more groups are equal*

*comparison of variance between and within groups*

*to check if sample groups are affected by same factors and to same degree*

*compare differences in means and variance of distribution*

**ONE-WAY ANOVA**=no of IVs

single IV with different (2) levels/variations have measurable effect on DV

*compare means of 2 or more indep groups*

aka:

- one-factor ANOVA

- one-way analysis of variance

- between subjects ANOVA

#### Assumptions

- independent samples

- equal sample sizes in groups/levels

- normally distributed

- equal variance

**F test is used to check statistical significance**

higher F value --> higher likelihood that difference observed is real and not due to chance

*used in field studies, experiments, quasi-exp*

**CONDITIONS:**

- min 6 subjects

- sample no of samples in each group

$H_0: \mu_1 = \mu_2 = \mu_3 \dots \mu_k$  i.e. all pop means are equal

$H_a$ : at least one  $\mu_i$  is different i.e atleast one of the k pop means is not equal to the others

**$\mu_i$  is the pop mean of group**

### Spearman Correlation

*non-parametric version of Pearson correlation coefficient*  
 named after Charles Spearman  
 denoted by  $\rho$  (rho)  
 determine the strength and direction of monotonic variables bet two variables measured at ordinal, interval or ratio levels & whether they are correlated or not

**monotonic function**=one variable never increases or never decreases as its IV changes

- monotonically increasing= as X increases, Y never decreases
- monotonically decreasing= as X increases, Y never increases
- not monotonic= as X increases, Y sometimes dec and sometimes inc

for analysis with: ordinal data, continuous data

**uses ranks instead of assumptions of normality**  
 aka Spearman Rank order test

**FORMULA:**  
 $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$   
 $d_i$  = difference between two ranks of each observation  
 -1 to +1  
 +1 = perfect association of ranks  
 0 = no association  
 -1 = perfect negative association of ranks

**closer the value to 0, weaker the association**

**Value Ranges**  
 0 to 0.3 = weak monotonic relationship  
 0.4 to 0.6 = moderate strength monotonic relationship  
 0.7 to 1 = strong monotonic relationship

### Parametric and Non-parametric test

Fixed set of parameters, certain assumptions about **distribution of population**

**PARAMETRIC** - *prior knowledge of pop distribution i.e NORMAL DISTRIBUTION*

**NON-PARAMETRIC** - *no assumptions, do not depend on population, DISTRIBUTION FREE tests, values found on nominal or ordinal level*

easy to apply, understand, low complexity

*decision based on* - distribution of population, size of sample

parametric - mean & <30 sample

non-parametric - median/mode & >30 sample or regardless of size

### Advantages & Disadvantages - NON-PARAMETRIC TESTS

ADVANTAGES	DISADVANTAGES
simple, easy to understand	less powerful than parametrics
no assumptions	counterpart parametric if exists, is more powerful
more versatile	not as efficient as parametric tests
easier to calculate	may waste information
hypothesis tested may be more accurate	requires larger sample to be as powerful as parametric test
small sample sizes are okay	difficult to compute large samples by hand
can be used for all types of data (nominal, ordinal, interval)	tabular format of data required that may not be readily available
can be used with data having outliers	

### Application

PARAMETRIC TESTS	NON-PARAMETRIC TESTS
- quantitative & continuous data	- mixed data
- normally distributed	- unknown distribution of population
- data is estimated on <b>ratio</b> or <b>interval</b> scales	- different kinds of measurement scales

### degrees of freedom

independent values in the data sample that have freedom to vary

**FORMULA:**  
 no of values in a data set minus 1  
 $df = N - 1$

### t-test

*statistical test to determine if significant difference between avg scores of two groups*

1908-William Sealy Gosset-**student t-test and t-distribution for hypothesis testing**

knowledge of:

- distribution** - normally distributed
- no knowledge of SD**

### t-test (cont)

#### TYPES:

**one-sample t-test** - single group

#### FORMULA:

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

#### SD FORMULA:

$$\sigma = \sqrt{\sum (X - \mu)^2 / N}$$

$$s = \sqrt{\sum (X - \mu)^2 / n - 1}$$

**independent two-sample t-test** - two groups

**paired/dependent samples t-test** - sig diff in paired measurements, compares means from same group at diff times (test-retest sample)

H0: no effective difference = **measured diff is due to chance**

Ha: two-tailed/ one-tailed *nonequivalent means/smaller or larger than hypothesized mean*

PERFORM **two-tailed test**: to find out difference bet two populations

**one-tailed**: one pop mean is > or < other

### Independent two-sample t-test

*aka unpaired t-test*

**to compare mean of two independent groups**

ex. avg weight of males and females

two forms:

- **student's t-test**: assumes SD is equal

- **welch's t-test**: less restrictive, no assumption of equal SD

*both provide more/less similar results*

#### ASSUMPTIONS:

- normally distributed
- SD is same
- independent groups
- randomly selected
- independent observations
- measured on **interval** or **ratio** scale

#### FORMULA:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

$$df = n_1 + n_2 - 2$$

$$S = \sqrt{\sum (x_1 - \bar{x})^2 + \sum (x_2 - \bar{x})^2 / n_1 + n_2 - 2}$$

### One-sample z-test

*to check if difference between sample mean & population mean when SD is known*

#### FORMULA:

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

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

**z score** is compared to a **z table** (includes % under NPC bet mean and z score), tells us whether the z score is due to chance or not

#### conditions:

knowledge of:

- pop mean
- SD
- simple random sample
- normal distribution

*two approaches to reject H0:*

- **p-value approach** - p-value is the smallest level of significance at which H0 can be rejected... *smaller p-value, stronger evidence*

- **critical value approach** - comparing z stat to critical values... indicate boundary regions where stat is highly improbable to lie = critical regions/rejection regions

if z stat is in critical region -> reject H0

*based on:*

significance level (0.1, 0.05, 0.01), alpha level, Ha

### Biserial correlation

*to measure relationship between **quantitative variables** and **binary variables***

given by Pearson - 1909

biserial correlation coeff varies bet **-1** and **1**

0 = no association

ex. IQ scores and pass/fail correlation

**continuous variable** and **binary variable** (dichotomised to create binary variable)

**rbis** or **rb** = correlation index estimating strength of relationship between artificially dichotomous variable and a true continuous variable

#### ASSUMPTIONS:

- data measured on continuous scale
- one variable to be made dichotomous
- no outliers
- approx normally distributed
- equal variances (SD)

#### FORMULA

$$rb = \frac{M_1 - M_0}{SD_t} \cdot \frac{pq}{y}$$

M1 = mean of grp 1



### Biserial correlation (cont)

M2= mean of grp 2  
p= ratio of grp 1  
q= ratio of grp 2  
SDt= total SD  
y= ordinate

### Pearson Correlation

measures **strength and direction** of a linear relationship between two variables

how two data sets are correlated

gives us info about the slope of the line

r

aka:

- Pearson's r
- bivariate correlation
- Pearson product-moment correlation coefficient (PPMCC)

cannot determine dependence of variables & cannot assess nonlinear associations

r value variation:

-0.1 to -.03 / 0.1 to 0.3 = weak correlation

-0.3 to -0.5 / 0.3 to 0.5 = average/moderate correlation

-0.5 to -1.0 / 0.5 to 1.0 = strong correlation

FORMULA:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

### Mann-Whitney U test

non-parametric test to test the significance of difference two independently drawn groups OR compare outcomes between two independent groups

equi to unpaired t test

CONDITIONS:

No NPC assumption, small sample size >30 with min 5 in each group, continuous data (able to take any no in range), randomly selected samples,

aka:

Mann-Whitney Test

Wilcoxon Rank Sum test

H0: the two pop are equal

Ha: the two pop are not equal

denoted by U

FORMULA:

$$U1 = n1n2 + \frac{n1(n1+1)}{2} - R1$$

$$U2 = n1n2 + \frac{n2(n2+1)}{2} - R2$$

### Mann-Whitney U test (cont)

R= sum of ranks of group

### One-way ANOVA test

Source of Variation	Sums of Squares (SS)	Degrees of Freedom (df)	Mean Squares (MS)	F
Between Treatments	$SSB = \sum_{j=1}^k (\sum_{i=1}^n x_{ij} - \bar{x}_{.j})^2$	dfr = k-1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSE}$
Error (or Residual)	$SSE = \sum_{j=1}^k \sum_{i=1}^n (x_{ij} - \bar{x}_{.j})^2$	dfe = N-k	$MSE = SSE/N-k$	
Total	$SST = SSB + SSE$	dft = N-1		

### One-way ANOVA test

- SSB/SSR = the regression sum of squares
- SSE = the error sum of squares
- SST = the total sum of squares (SST = SSR + SSE)
- dfr = the model degrees of freedom (equal to dfr = k - 1)
- dfe = the error degrees of freedom (equal to dfe = n - k)
- k = the total number of groups (levels of the independent variable)
- n = the total number of observations
- dft = the total degrees of freedom (equal to dft = dfr + dfe = n - 1)

### One-way ANOVA test

- MSR = SSR/dfr = the regression mean square
- MSE = SSE/dfe = the mean square error
- Then the F statistic itself is computed as:  
 $F = MSR/MSE$
- p: The p-value that corresponds to Fdfr, dfe