Standard Error: variability of multiple samples of population
Standard Deviation: variability of individual point
Degree of Freedom: Data sizes are large enough it is generally insignificant

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \quad SE(\bar{X}) = \sigma_{\bar{X}} = {}^{\sigma}/_{\sqrt{n}} \quad SE(p) = \sqrt{\frac{p(1-p)}{n}}$$

Confidence Interval = point estimate ±margin of error
Margin of Error↓: sample size↑/ confidence level ↓

$$CI = \bar{X} \pm Z * (\frac{\sigma}{\sqrt{n}}) \quad Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \quad E = t_{\alpha/2}(\frac{S}{\sqrt{n}})$$

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

$$= \bar{X} \pm t_{n-1}(1-\alpha) * (\frac{s}{\sqrt{n}})$$

Confidence Interval	α	Z Statistic
0.9	0.1	1.645
0.95	0.05	1.96
0.99	0.01	2.58

Null Hypothesis: status quo or existing knowledge

	H_0	H_A	H_0	H_A
Left-tailed	$\mu \geq \mu_0$	$\mu < \mu_0$	$p > p_0$	$p < p_0$
Right-tailed	$\mu \leq \mu_0$	$\mu > \mu_0$	$p < p_0$	$p > p_0$
Two-tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	$p = p_0$	$p \neq p_0$

T-statistic / Z-statistic

One-sample

$$t_{data} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$Z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Two-sample (A/B testing)

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

$$\frac{p_1 - p_2}{\sqrt{p_{pooled}\cdot(1-p_{pooled})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

F-statistic $= \frac{MSG}{MSE}$

	Df	Sum Sq	Mean Sq
Group	$df_G = k-1$	$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$MSG = \frac{1}{df_G}SSG$
Residuals	$df_E = n-k$	$SSE = \sum_{i=1}^{k}(n_i - 1)s_i^2$	$MSE = \frac{1}{df_E}SSE$
Total	$df_T = n-1$	$SST = \sum_{i=1}^{n}(x_i - \bar{x})^2$	

Analysis of Variance (ANOVA)
Statistical procedure that tests for statistically significant differences among multiple categories
H_0: Average web stickiness is identical for all pages ($\mu_1 = \mu_2 = \mu_3 = \mu_4$)
H_A:The average web stickiness varies by page

Chi-Square Test
H_0: All proportions are the same

$$R = \frac{observed - expected}{\sqrt{expected}} \quad \chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} R^2$$

$$df = (r-1)*(c-1)$$

Exploratory Factor Analysis (EFA): <u>understand the structure in data.</u> Similar to PCA, creating components and loading, but EFA attempts to find solutions that are maximally interpretable and and interpreted by factor loading and naming factors.

Principle Component Analysis (PCA): <u>Reduce the number of dimensions</u> while retaining as much information as possible. Mitigate <u>multicollinearity</u> by creating uncorrelated variables (PC)
- **Standardize the data** → ensure all variable have the same impact on the results
- Compute the covariance matrix → Heat map
- Create components with loadings and attributes

components loadings attributes

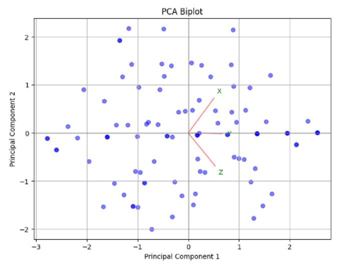
$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots + \phi_{p1}X_p$$
- Choose a number of principal components to retain → Scree plot
- Transform the original data using the projection matrix

Attribute Types	Visualization
category-category	Contingency table, contour plot
measure-measure	Correlation matrix, scatter plot
category-measure	Side-by-side box plots or violin plots

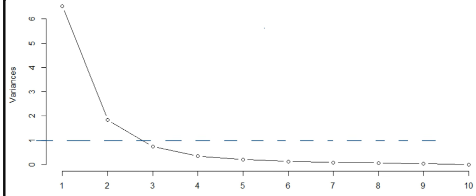
Attribute	Description	Examples
Nominal	"Name" or identifier. Represents some category or state (also referred as categorical attributes) There is no order (rank, position) among values of nominal attribute	Gender, marital status, occupation, ID numbers, zip codes
Ordinal	Similar to nominal except the values have a meaningful order	Street number, grades, ranks,
Interval	Differences between values is meaningful	Temperature in C or F, IQ scores, SAT scores
Ratio	Similar to interval except there is a "true zero" so it is meaningful to talk about ratios between values	Temperature in K, age, monetary value, mass, length

Biplot: displays both the scores and loadings in a PCA on the same plot, showing projections of both original data points (points) and variables (vectors) onto PC

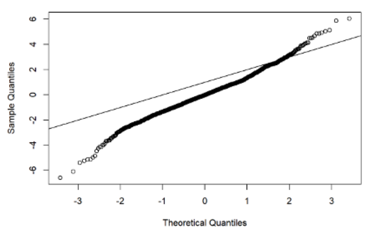
	X	Y	Z
PC1	0.510702	0.676224	0.530947
PC2	0.726843	-0.009732	-0.686734
PC3	-0.459219	0.736632	-0.496479



Scree plot: show the variance contribution of each PC, determining the optimal number of PC to retain. The Elbow points determine the amount of PC



Q-Q (Quantile-quantile) **plots**: test of normality
Expected Z-Score of data element at that percentile
Straight line: normal distribution / Deviation: curves, outliers



Interquartile Range (IQR)
$= Q_3 - Q_1$
Coefficient of Variation (CV)

$$CV = \frac{\sigma}{\mu} * 100$$

Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

<u>Missing Completely at Random (**MCAR**): independent</u>
Data entries are missing due to mishandling or loss of some patient forms regardless of any patient characteristics or outcomes.
<u>Missing at Random (**MAR**): depend on observed data</u>
Patients in a certain age group are less likely to fill out a follow-up survey about treatment outcomes, and age is recorded in the dataset.
<u>Missing Not at Random (**MNAR**): depend on unobserved data</u>
Patients experiencing severe side effects from a treatment are less likely to report their follow-up outcomes due to their health condition.

Outlier: bad data/ different types / good data
- Z-scores: beyond 4-5 standard deviation / IQR rule: <25 or >75 percentile with 1.5 IQR
- Histogram, Boxplots
- Domain-specific rules: based on expert knowledge or business logic (e.g. age > 120 is invalid)

Poisson: Discrete / **Skewed:** Continuous **/ Bimodal:** two distinct peaks (mode)
Exponential: Time between events, *customer arrive time, time between bus arrival*
Lognormal: always positive, *stock prices, population and company sizes*

Bootstrapping: Draw multiple samples *with replacement* from the sample and recalculate the statistic or model result for each response:
- Draw n samples *with replacement* from the original sample
- Record the statistic (e.g., mean) of the n sampled values
- Repeat R times
- Use the R results to: Calculate their standard deviation (to estimate the standard error) / Produce a histogram or boxplot / Find a confidence interval

Obs	X	Y
1	4.3	2.4
2	2.1	1.1
3	5.3	2.8

Original Data

Obs	X	Y
2	2.1	1.1
3	5.3	2.8
1	4.3	2.4

Bootstrap

Obs	X	Y
3	5.3	2.8
1	4.3	2.4
3	5.3	2.8

Bootstrap

Obs	X	Y
2	2.1	1.1
2	2.1	1.1
1	4.3	2.4

Bootstrap

Permutation tests: Draw multiple samples *without replacement*
Randomly select 21 observations from the group of 36
- Calculate their average
- Calculate the average of the other 15
- Calculate the <u>difference</u> between the two averages
- Repeat this procedure a large number of times and observe the results (via a histogram or similar visualization)

P-value: x of n permutations have a difference of their means greater than original means, p-value = x/n. If p-value > alpha, reject H0 and accept HA

Approach	Pros	Cons
Classical statistics	▪ Fast and computationally efficient ▪ Works well for normally distributed data and simple parametric models. ▪ Well-established theoretical foundation	▪ Relies on assumptions (e.g., normality, independence, equal variance). ▪ Can be inaccurate for small samples, skewed distributions, or complex models. ▪ Does not generalize well to non-parametric problems or cases where standard error formulas are difficult or impossible to derive
Computational statistics	▪ No distributional assumptions—works well for skewed or non-normal data. ▪ Works for small sample sizes where normal approximations may fail. ▪ More flexible—can be applied to complex models where standard errors are hard or impossible to derive.	▪ Computationally expensive, especially for large datasets. ▪ Can be sensitive to the number of bootstrap samples—too few can lead to unstable estimates. ▪ May be harder to interpret compared to parametric approaches with clear formulas.

Shirt Size (Large, Medium, Small)	Ordinal
Temperature in degrees Celsius	Interval
Customer satisfactions (1: very dissatisfied, …, 5: very satisfied)	Ordinal
GRE test score	Interval
Types of payment methods (Credit card, cash, check)	Nominal
Day of the week	Nominal
Zip code	Nominal
Student ID number	Nominal
Age of a person	Ratio
Duration between the start and end dates of project	Ratio

Chi-Square Test Example

1. Calculate the totals
   Value W = 2 + 3 + 3 = 8
   Group A = 2+4+5+4 = 15
   Total = 75
2. Compute the expected counts
   Expected W = (8*15)/75 = 1.6
3. Compute the R values
   $$R = \frac{observed - expected}{\sqrt{expected}}$$
4. Compute the Chi-square Statistic
   Sum of total = Chi-Square = 0.66

	Group A	Group B	Group C
Value W	2	3	3
Value X	4	6	8
Value Y	5	8	10
Value Z	4	10	12

	Group A	Group B	Group C
Value W	1.6	2.88	3.52
Value X	3.6	6.48	7.92
Value Y	4.6	8.28	10.12
Value Z	5.2	9.36	11.44

	Group A	Group B	Group C
Value W	0.316	0.071	−0.277
Value X	0.211	−0.189	0.028
Value Y	0.186	−0.097	−0.038
Value Z	−0.526	0.209	0.166

Classical Statistics	Computational Statistics
Small to moderate sample size	Large to very large sample size
Independent, identically distributed data sets	Nonhomogeneous data sets
Mathematically tractable	Numerically tractable
Well focused questions	Imprecise questions
Strong unverifiable assumptions Relationships (linearity, additivity) Error structures (normality)	Weak or no assumptions Relationships (nonlinearity) Error structures (distribution free)
Predominantly closed form algorithms	Predominantly iterative algorithms