Data Visualisation

CMP020L013A

Week 11: Statistical Methods in Data Visualisation

Mohammad Javaheri

(Dr Mohammad Ali Javaheri Javid)

Agenda

- the importance of statistical methods
- the use and types of descriptive statistics in visualisation
- the use and types of inferential statistics in visualisation

Statistical Methods

- ►Statistical methods enable us to analyse quantitative data, specifically

  (1) to inspect data quality and characteristics and

  (2) to discover relationships (e.g., causal) among experimental variables or to estimate population characteristics.

1. **Descriptive** statistics
2. **Inferential** statistics

Descriptive vs. Inferential Statistics

- A **descriptive statistic** is a summary statistic that quantitatively describes or summarises features of collected data
- **Descriptive** statistics is the process of using and analysing those statistics.
- **Inferential** statistics (or statistical inference or modeling) is the process of making propositions about a population using data drawn from it through sampling.
- Using *descriptive statistics*, we summarise a sample of data;
- Using *inferential statistics*, we make propositions about the population.

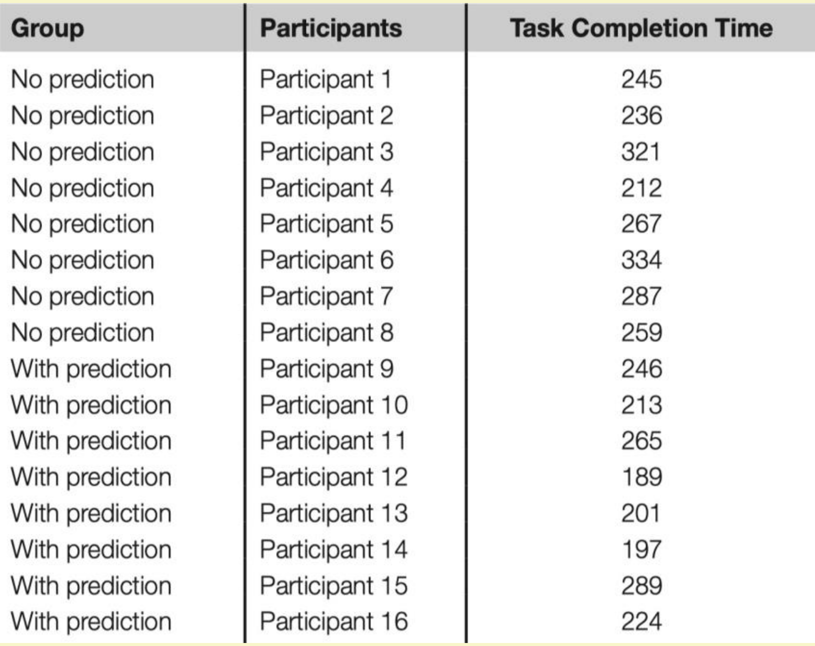
Descriptive vs. Inferential Statistics

- When do we use descriptive and inferential statistics?
- Applications of Descriptive statistic
  - To assess data quality and structure
  - To describe population characteristics
  - To assess dependence among variables
- Applications of Inferential statistics
  - To test hypotheses
  - To estimate parameters
  - To perform clustering or classification

Descriptive Statistics

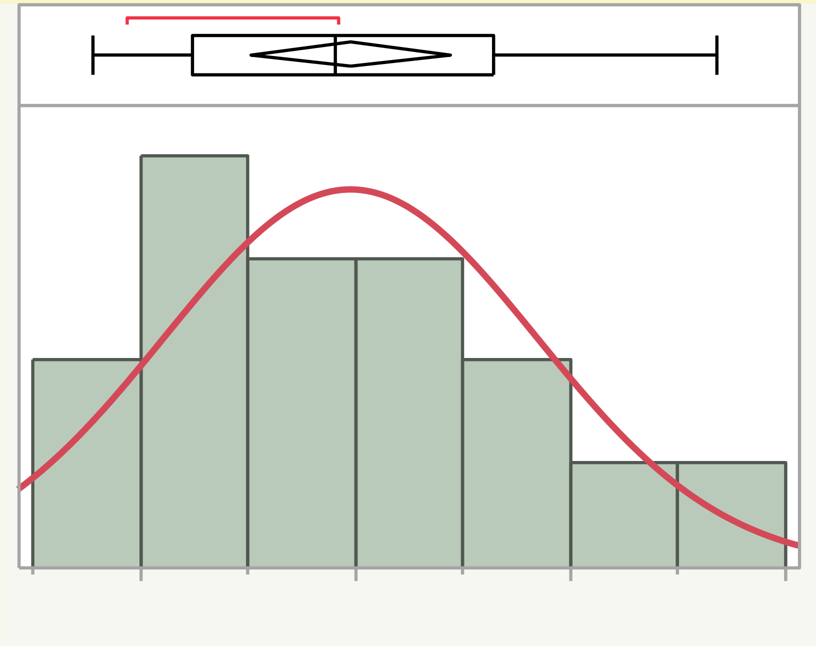
▸ How to perform descriptive statistics?

1. Prepare data table



Group	Participants	Task Completion Time
No prediction	Participant 1	245
No prediction	Participant 2	236
No prediction	Participant 3	321
No prediction	Participant 4	212
No prediction	Participant 5	267
No prediction	Participant 6	334
No prediction	Participant 7	287
No prediction	Participant 8	259
With prediction	Participant 9	246
With prediction	Participant 10	213
With prediction	Participant 11	265
With prediction	Participant 12	189
With prediction	Participant 13	201
With prediction	Participant 14	197
With prediction	Participant 15	289
With prediction	Participant 16	224

2. Inspect data distribution



Source: Jonathan Lazar et al., Research Methods in Human-Computer Interaction, 2nd Edition, 2017

Descriptive Statistics

- Types of analyses in descriptive statistics
  - **Univariate analysis**
    - It involves describing the distribution of a single variable, including the type/form of distribution, central tendency, and dispersion.
  - **Bivariate or multivariate**
    - analysis involves describing the relationships between pairs of variables in terms of correlation, covariance, and slope.

Descriptive Statistics

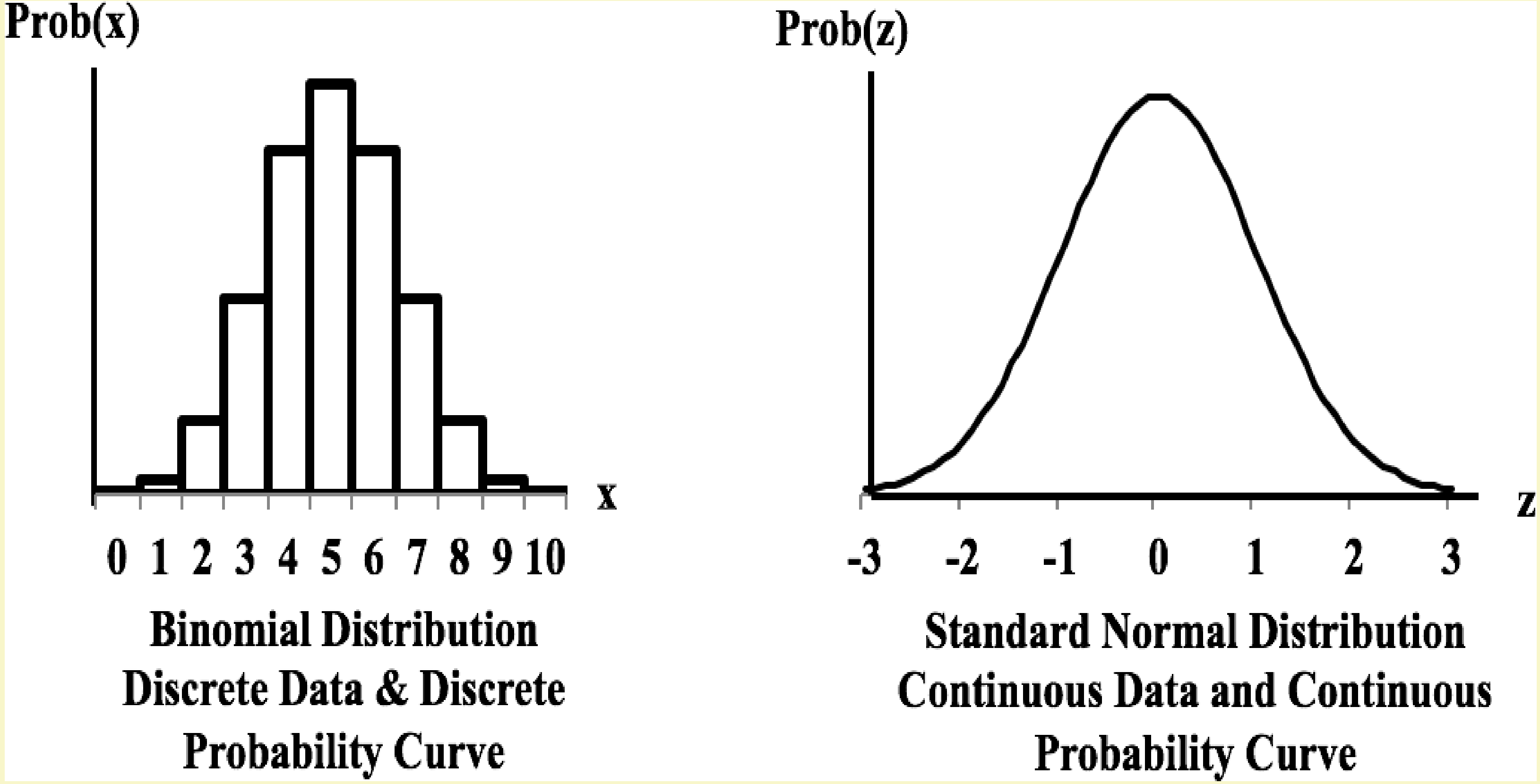
1. **Univariate Analysis**

- What do we look at in univariate analysis?
  - **Distribution**
    - what does our distribution look like? (For discrete, ordinal, or continuous data types)
  - **Central tendency**
    - where is most of our data? (For continuous data types only)
  - **Dispersion**
    - how much does the deviate from the centre?

Descriptive Statistics

1. Univariate Analysis

- ► **Distribution** — discrete, ordinal, or continuous data types



Binomial Distribution
Discrete Data & Discrete
Probability Curve

Standard Normal Distribution
Continuous Data and Continuous
Probability Curve

Descriptive Statistics

1. **Univariate Analysis**
   - ► Distribution — discrete, ordinal, or continuous data types
   - ► For the **Discrete data Distribution**,
   - ► the values of the Variable X can only be non-negative integers, because they are Counts.
   - ► The Probabilities for Discrete data Distribution are shown as separate columns.
   - ► There is nothing between the columns, because there are no values on the horizontal axis between the individual integers.

Descriptive Statistics

1. **Univariate Analysis**

- ► Distribution — discrete, ordinal, or continuous data types
- ► For **Continuous Distributions**,
- ► values of horizontal-axis Variable are real numbers, and there are an infinite number of them between any two integers.
- ► Continuous data are also called Measurement data; examples are length, weight, pressure, etc.
- ► The Probabilities for Continuous Distributions are infinitesimal points on smooth curves.

Descriptive Statistics

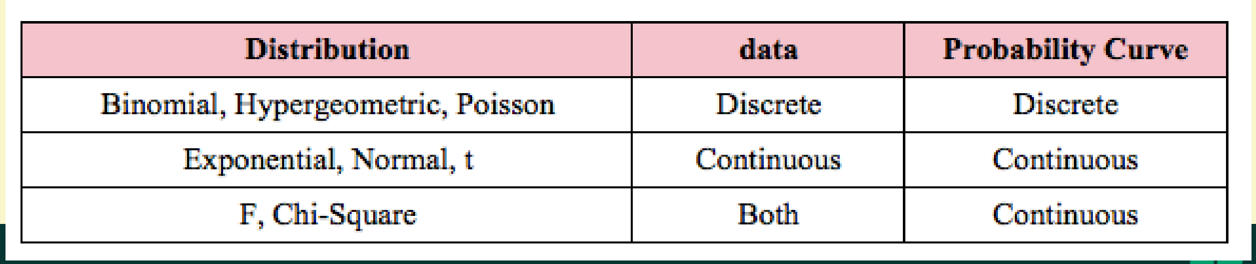
1. **Univariate Analysis**
   - ► Distribution — discrete, ordinal, or continuous data types

►For the first six Distributions described in the table below, the data used to create the values on the horizontal axis come from a single Sample or Population or Process.

► The F and Chi-Square ($\chi 2$) Distributions are hybrids. Their horizontal axis Variable is calculated from a ratio of two numbers, and the source data don’t have to be one type or another.

►Being a ratio, the horizontal axis Variable (F or $\chi 2$) is Continuous.

►The Probability curve is smooth and Continuous.

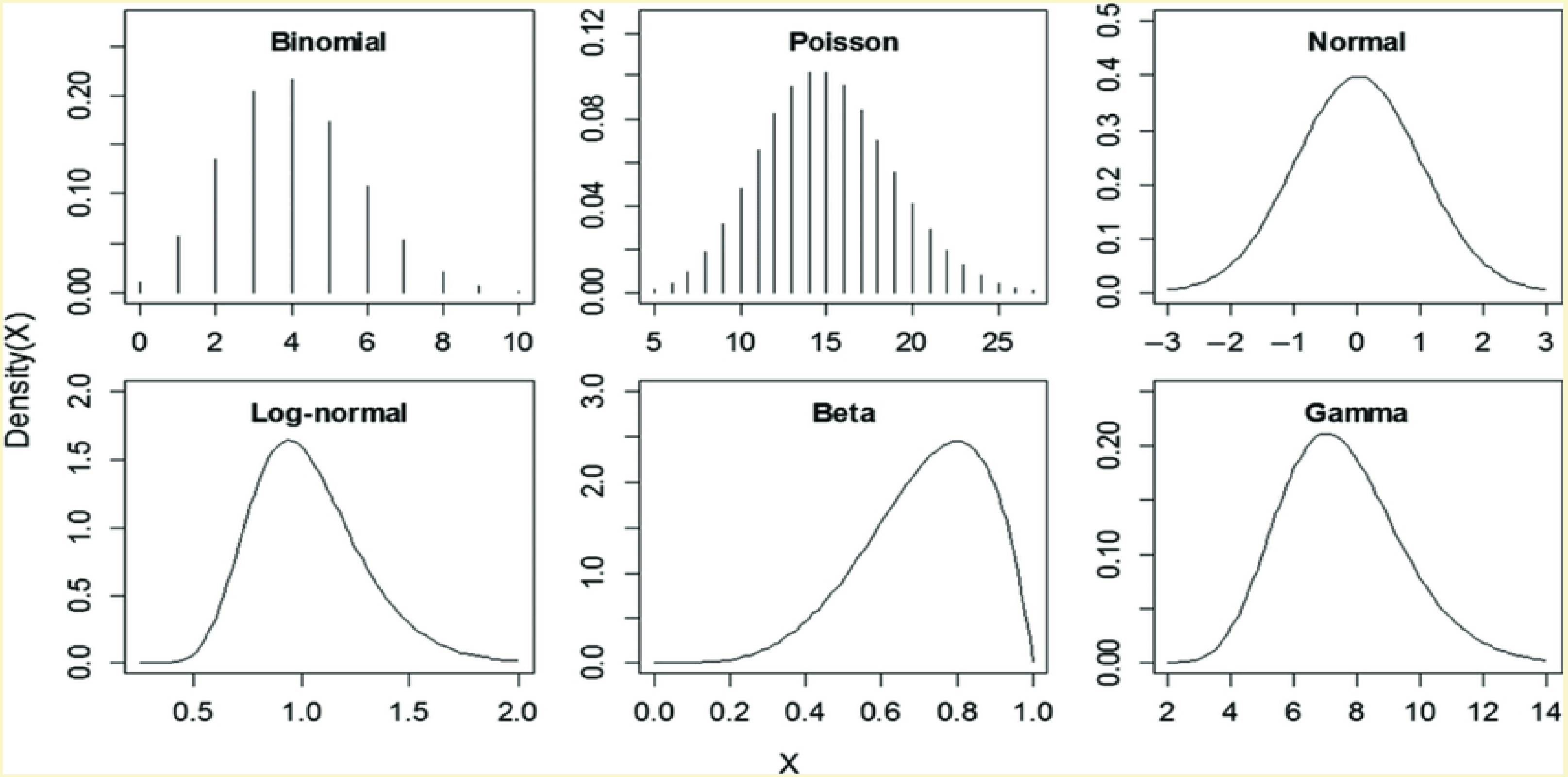


Distribution	data	Probability Curve
Binomial, Hypergeometric, Poisson	Discrete	Discrete
Exponential, Normal, t	Continuous	Continuous
F, Chi-Square	Both	Continuous

Descriptive Statistics

1. Univariate Analysis

- Distribution — discrete, ordinal, or continuous data types
- Data from discrete or continuous variables can take different forms and follow different probability distributions.



Source: by Daniel Wolcott

Descriptive Statistics

1. Univariate Analysis

Central tendency: the tendency for values of a variable to gather around the middle of the distribution

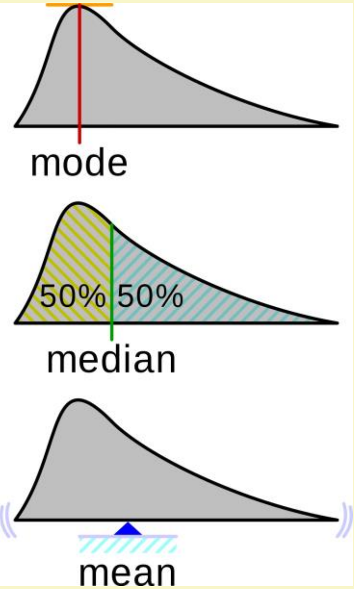
- Data from discrete or continuous variables can take different forms and follow different probability distributions.

- **Mean** is the arithmetic average of all the values in the distribution.

$$\sum \frac{x}{n}$$, where ***x*** is the value the variable can take, and ***n*** is the set size

- **Median** is the middle value when all the values are ordered.

- **Mode** is the value that occurs most frequently in the data.

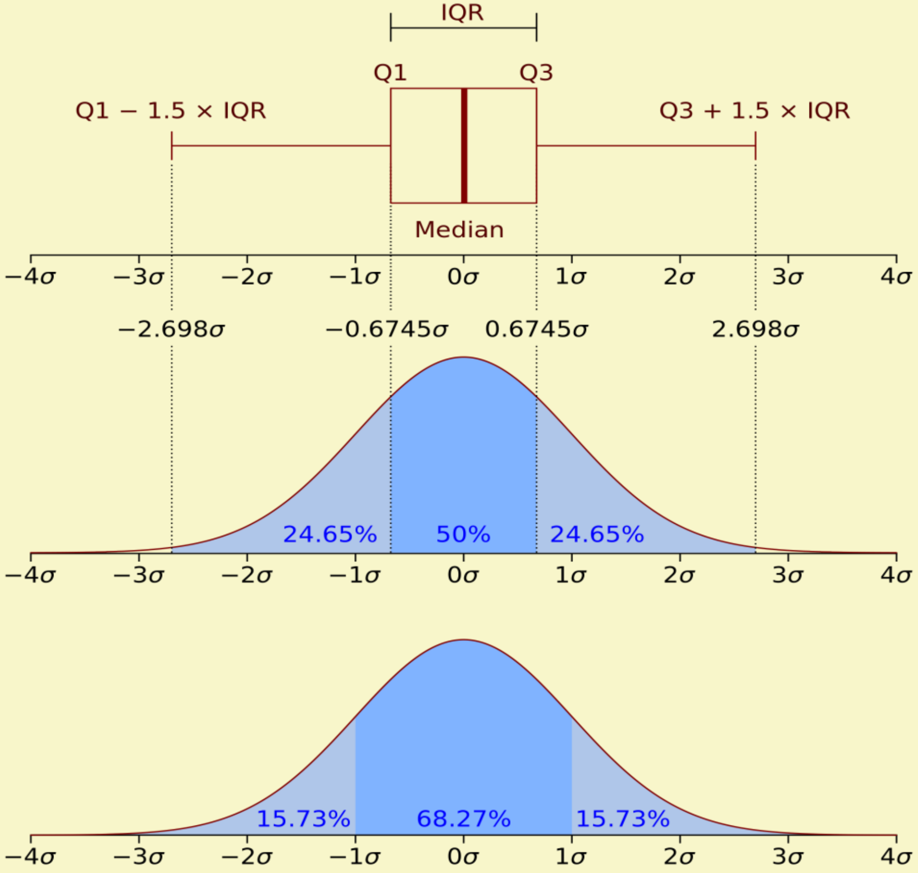


Descriptive Statistics

1. Univariate Analysis

- Dispersion: captures the spread and shape of the data distribution

- **Range** is the difference between the smallest and the largest values.
- **Quartiles** break the distribution into four equally sized parts.
- **Variance** is the squared deviation of the variable from its mean.
- **Standard deviation** measures the amount of variation or dispersion in values.



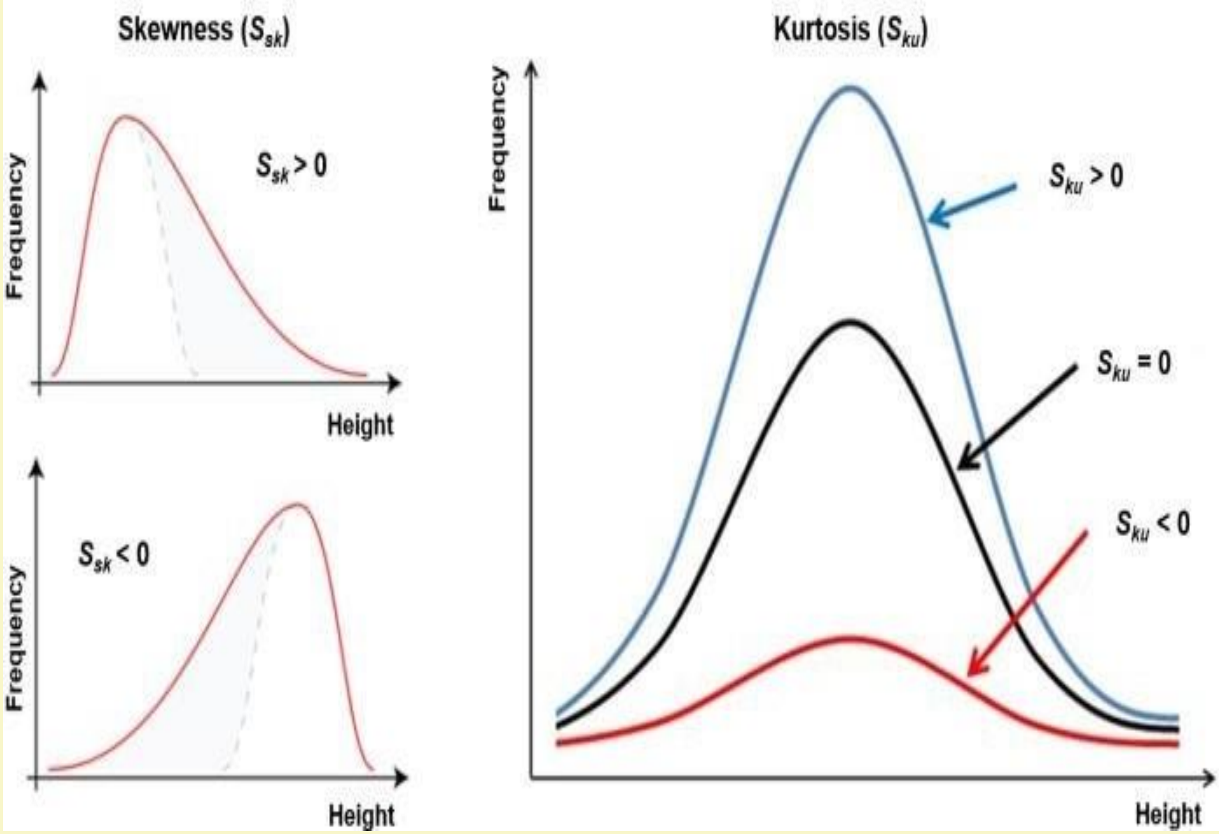
Descriptive Statistics

1. Univariate Analysis

- ► Dispersion: captures the spread and shape of the data distribution

►**Kurtosis** measures how much the values gather in the peak or the tail of the distribution: *leptokurtic, mesokurtic, platykurtic.*

►**Skewness** measures of asymmetry in the distribution: positive, *negative*.



Descriptive Statistics

2. **Bivariate/Multivariate Analysis**

- **Correlation** can be used for *descriptive* or *inferential* statistics.

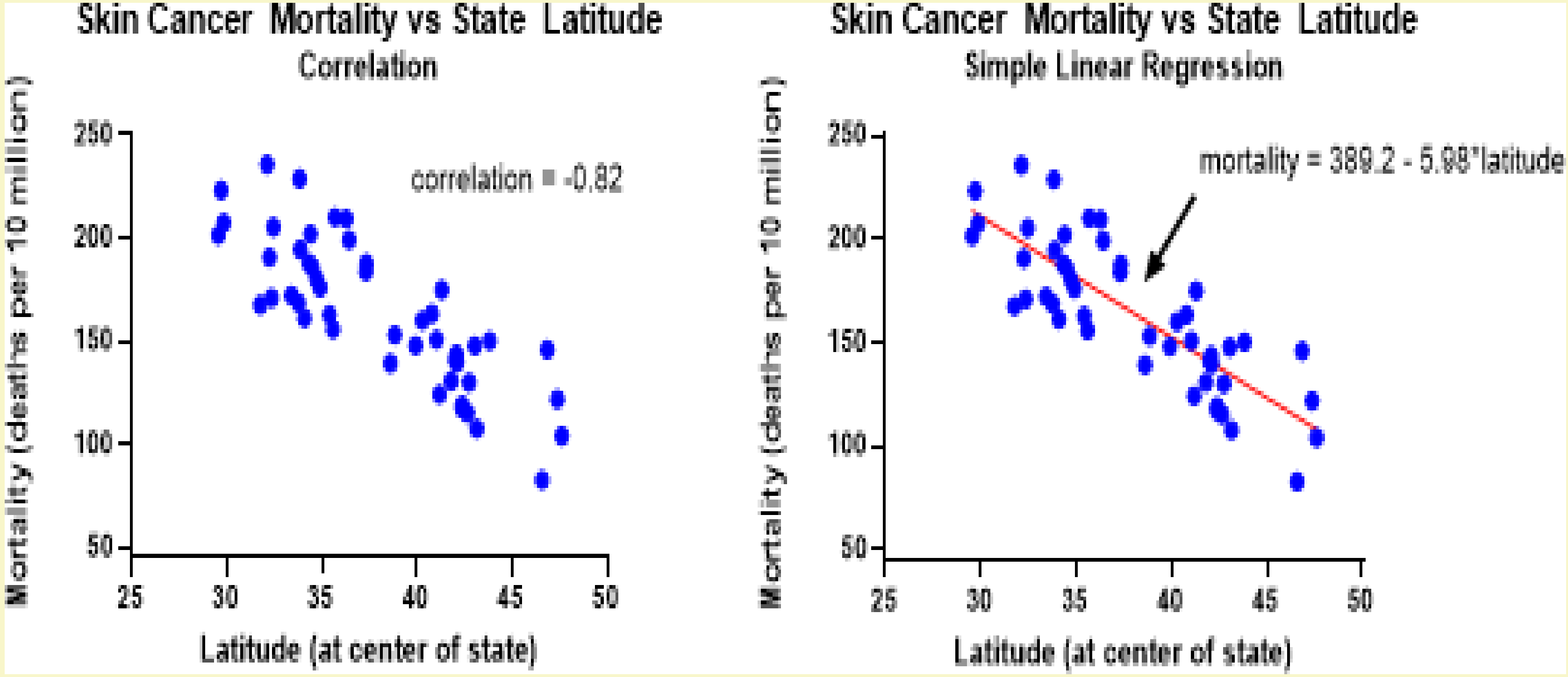


Figure Source

Descriptive Statistics

2. Bivariate/Multivariate Analysis

- What is calculated in Correlation is called a **correlation coefficient**

For a **population**

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

For a **sample**

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

Figure Source

Descriptive Statistics

2. Bivariate/Multivariate Analysis

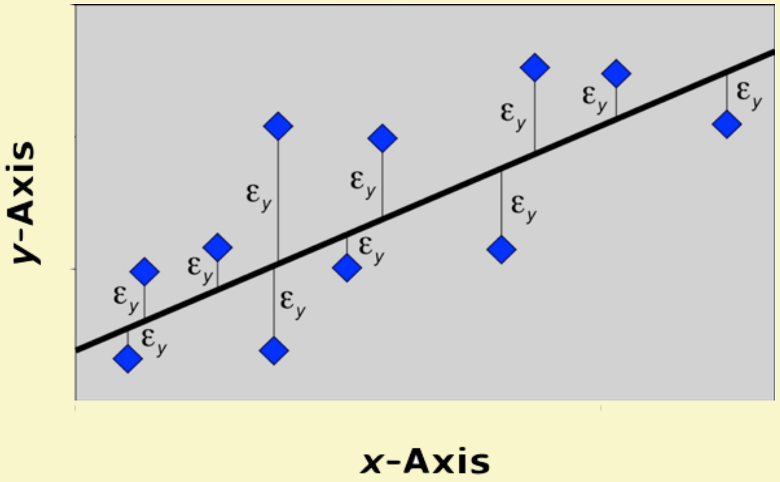
- ► **How can a correlation coefficient be interpreted?**
- ► Correlation coefficient is a measure of relation between two variables that ranges between -1 and 1.
- ▪ **Simple linear correlation**: Pearson's **r** calculates the extent to which the variables are proportional or linearly related to each other.
- ▪ If you square the Pearson r, you get a direct measure of the amount of variance that is shared between the two measures involved
- ► Example:
  - ► Pearson r = .50 means that 25% of the variance between measure X and measure Y is shared.
  - ► Pearson r = .70 means that 49% of the variance is shared.

Descriptive Statistics

2. **Bivariate/Multivariate Analysis**

- ►The proportion can be summarized by a simple line (regression or least squares line)
- ►Determined such that the sum of the squared distances of all the data points from the line is the lowest possible.

$$Y = \beta_0 + \sum_{i=1} \beta_1 X_i + \epsilon_i$$

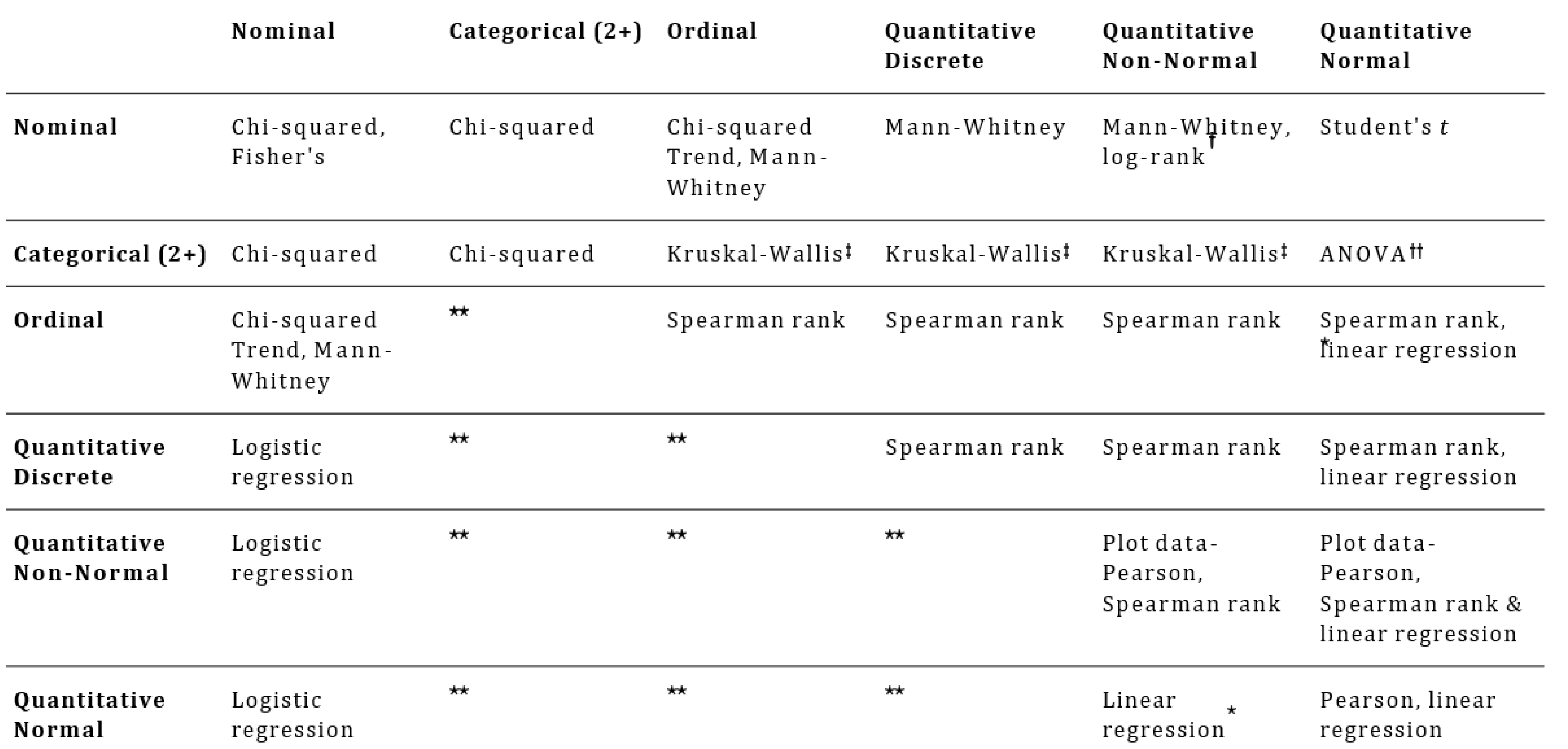


Inferential Statistics

- Inferential statistics involves families of statistical tests that aim to establish statistically significant differences between distributions
- A **statistical test**?
  - It is a mechanism for assessing whether data provides support for particular hypotheses.
- test a **hypothesis**?
  - Hypotheses are provisional statements about relationships among concepts.
  - In hypothesis testing, we seek to determine which statement data is consistent with.

Inferential Statistics

▸How do we determine what test to use?



	Nominal	Categorical (2+)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Nominal	Chi-squared, Fisher's	Chi-squared	Chi-squared Trend, Mann-Whitney	Mann-Whitney	Mann-Whitney, log-rank†	Student's *t*
Categorical (2+)	Chi-squared	Chi-squared	Kruskal-Wallis‡	Kruskal-Wallis‡	Kruskal-Wallis‡	ANOVA††
Ordinal	Chi-squared Trend, Mann-Whitney	**	Spearman rank	Spearman rank	Spearman rank	Spearman rank, *linear regression
Quantitative Discrete	Logistic regression	**	**	Spearman rank	Spearman rank	Spearman rank, linear regression
Quantitative Non-Normal	Logistic regression	**	**	**	Plot data-Pearson, Spearman rank	Plot data-Pearson, Spearman rank & linear regression
Quantitative Normal	Logistic regression	**	**	**	Linear regression*	Pearson, linear regression

Further Reading

- Bilge Mutlu, Human-Computer Interaction: https://wisc-hci-curriculum.github.io/cs770-s20/
- Descriptive statistics: https://en.wikipedia.org/wiki/Descriptive_statistics
- Statistical inference: https://en.wikipedia.org/wiki/Statistical_inference
- https://www.statisticsfromatoz.com/blog/statistics-tip-of-the-week-different-distributions-can-have-discrete-or-continuous-probability-graphs-for-discrete-or-continuous-data
- Visualisation mode median mean: https://commons.wikimedia.org/w/index.php?curid=38969094
- Boxplot vs PDF: https://commons.wikimedia.org/w/index.php?curid=14524285