Human-Computer Interaction

Statistics I:

Descriptive Statistics +

Introduction to Inferential Statistics

Professor Bilge Mutlu

Today's Agenda

- Statistical data analysis
- Descriptive statistics
- Correlation analysis
- Introduction to inferential statistics

Statistical Data Analysis

Why do we need to use statistics?

Statistical methods enable us to analyze 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

*What is the difference between **descriptive** and **inferential** statistics?*

A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features of collected data, while **descriptive statistics** is the process of using and analyzing those statistics.[1]

Inferential statistics, or statistical inference (or modeling), is the process making propositions about a population using data drawn from the population through sampling.[2]

Simply put, using descriptive statistics, we summarize a sample of data; using inferential statistics, we make propositions about the population.

---

[1] Wikipedia: Desciptive Statistics

[2] Wikipedia: Inferential Statistics

When do we use descriptive and inferential statistics?

Usually, descriptive and inferential statistics are used together.

Descriptive statistics:

» To assess data quality and structure

» To describe population characteristics

» To assess dependence among variables

Inferential statistics:

» To test hypotheses

» To estimate parameters

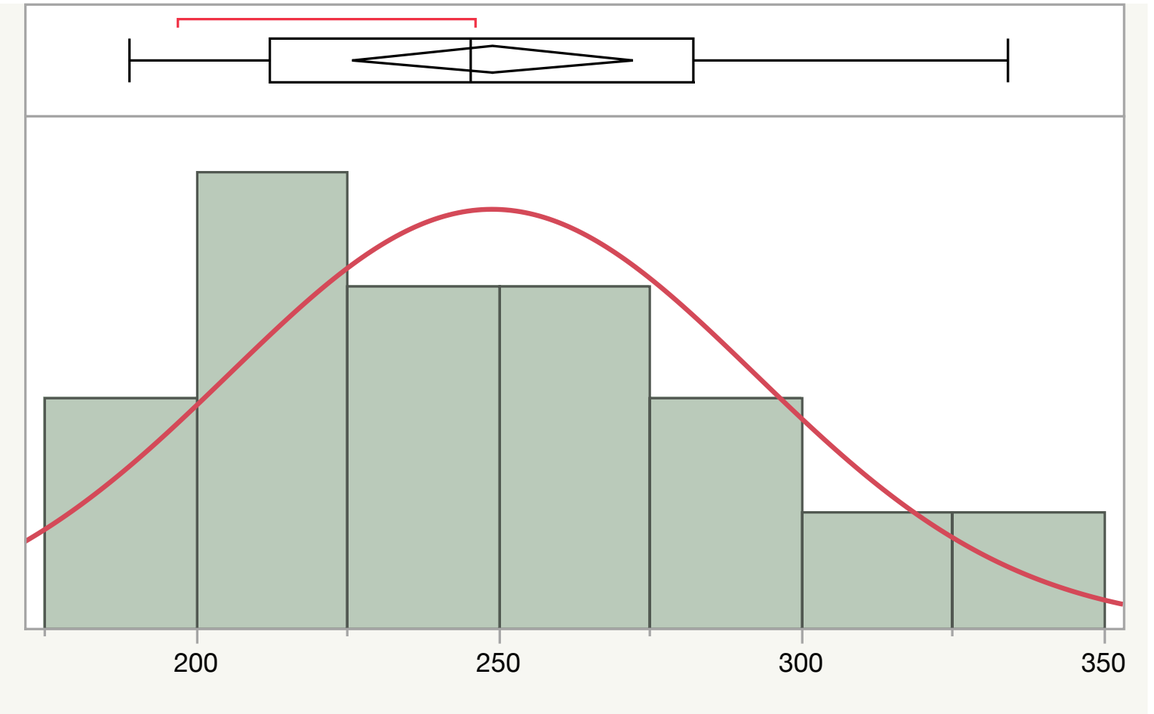
» To perform clustering or classification

Descriptive Statistics

How do we perform descriptive statistics?

First, by preparing our data table and inspecting our data distribution.[3]

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



[3] Lazar et al., 2017, Chapter 4

What are the types of analyses in descriptive statistics?

Univariate analysis involves describing the distribution of a single variable, including *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*.

What do we look at in univariate analysis?

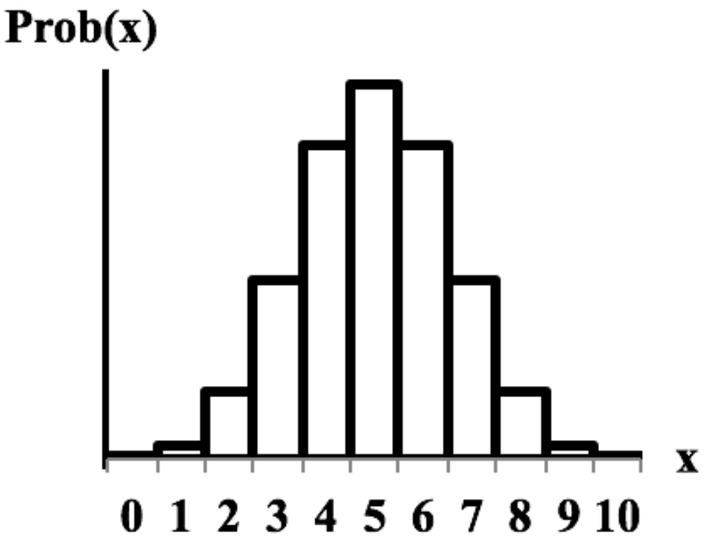
1. Distribution — what does our distribution look like?[4]
2. Central tendency — where is the majority of our data?[5]
3. Dispersion — how much does the deviate from the center?[5]

[4] Data from "Mutlu et al. (2009). Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. *HRI 2009.*"

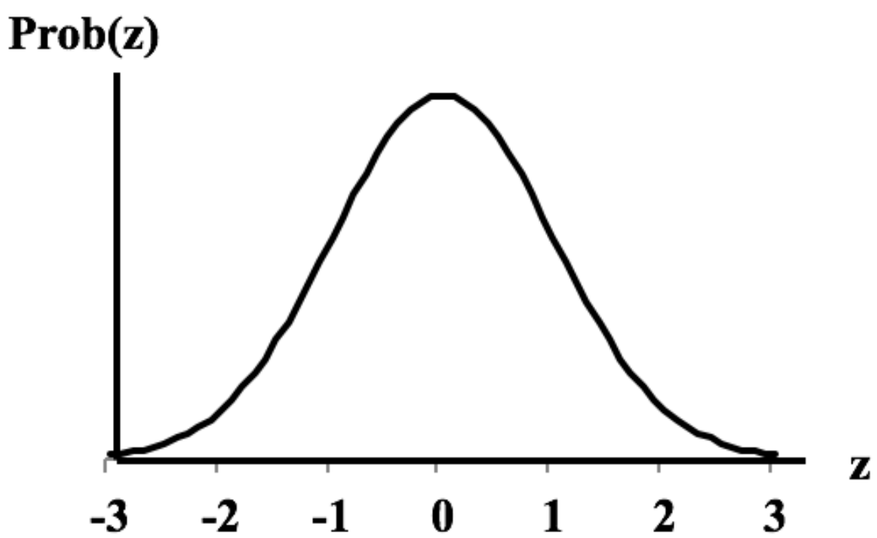
[5] For continuous data types only

Distribution[6]

Distributions can be **discrete** or **continuous**.



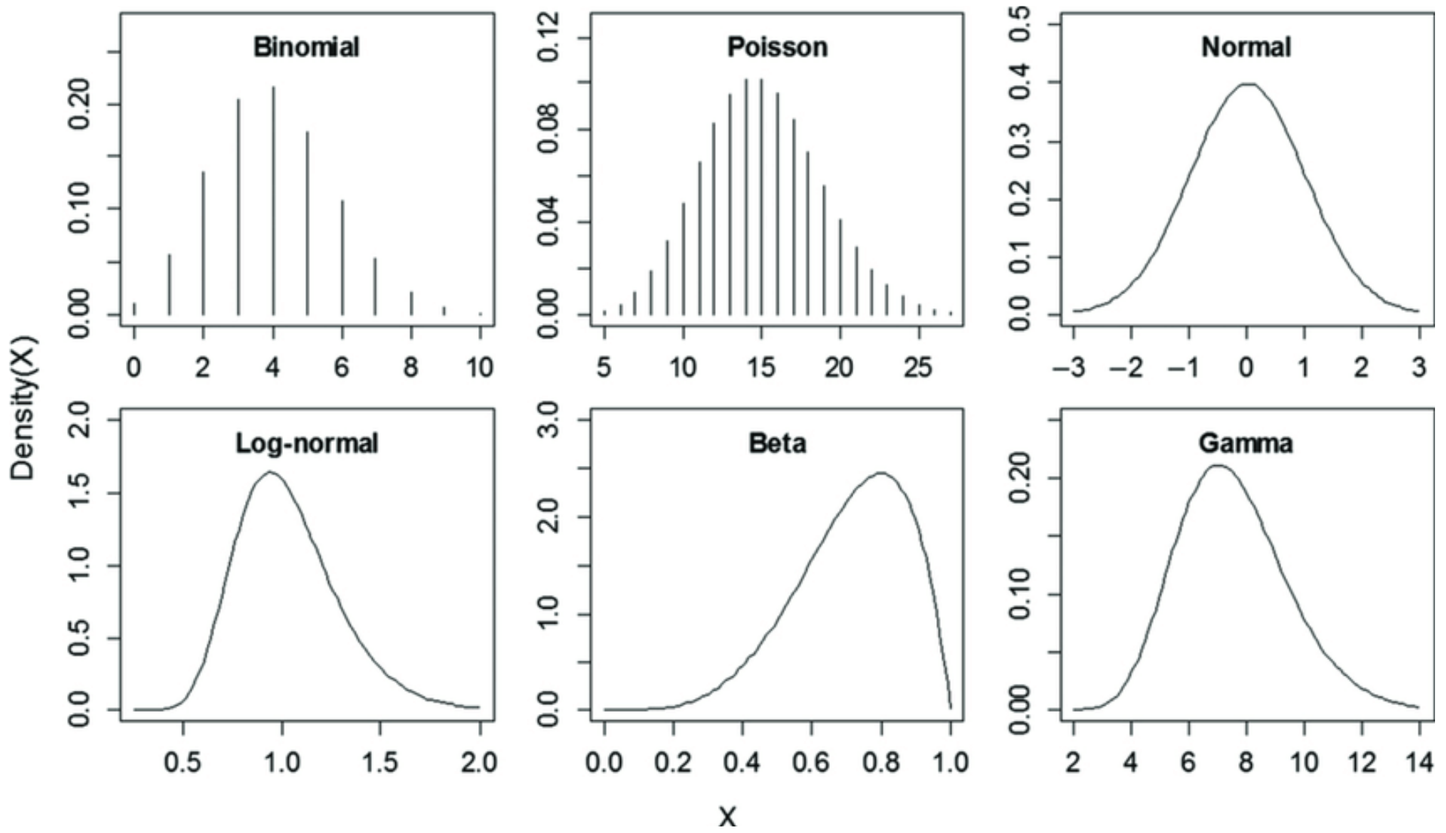
Binomial Distribution
Discrete Data & Discrete Probability Curve



Standard Normal Distribution
Continuous Data and Continuous Probability Curve

[6] Image source

Data from discrete or continuous variables can take different forms and follow different probability distributions.[7]



[7] Image source: Daniel Wolcott

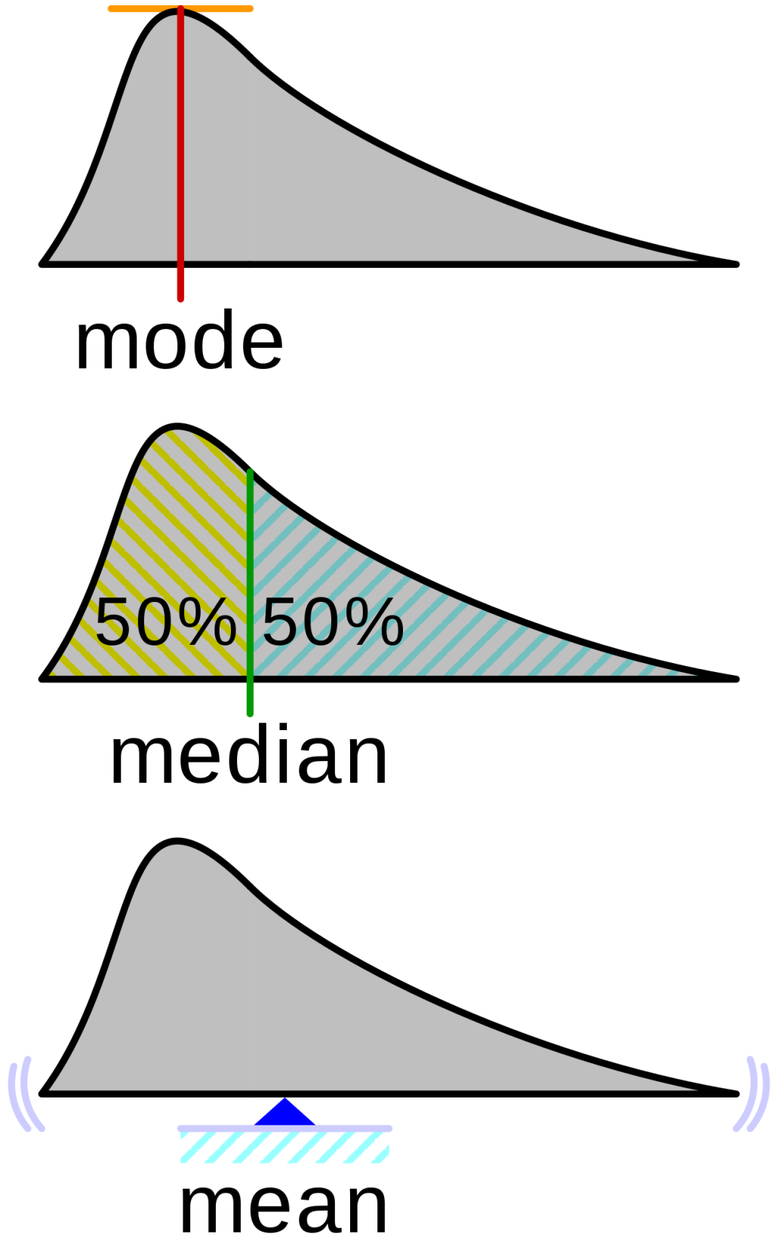
Central tendency[8]

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

Mean is the arithmetic average of all the values in the distribution. $\sum \frac{x}{n}$ where x is the values the variable can take and n is the set size.

Median is the middle value when all the values in the distribution are ordered.

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



[8] By Cmglee - Own work, CC BY-SA 3.0

Dispersion[9]

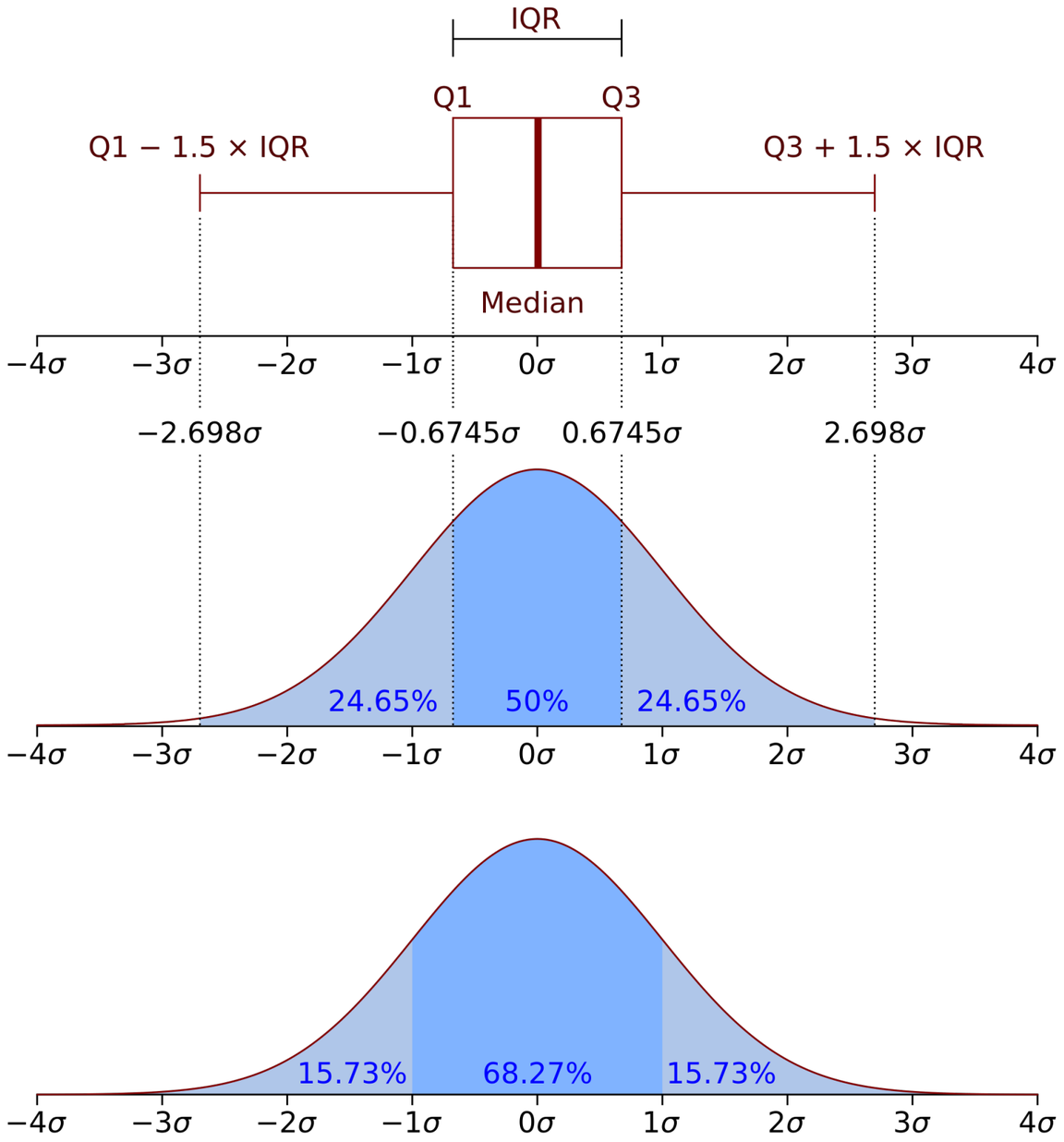
Dispersion captures the *spread* and *shape* of the data distribution.

Range is the difference between the smallest and the largest values.

Quartiles break the distribution to four equally sized parts.

Variance is the squared deviation of the variable from its mean.

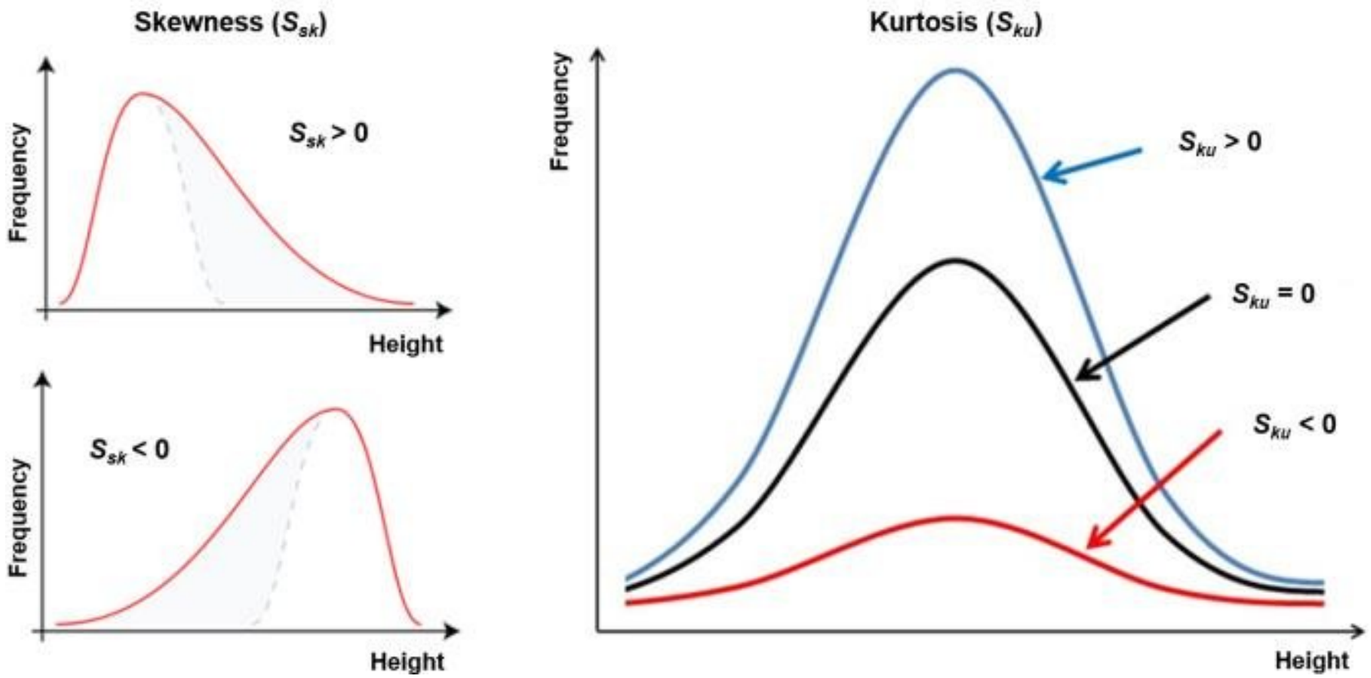
Standard deviation measures the amount of variation or dispersion in values.



[9] By Jhguch at en.wikipedia, CC BY-SA 2.5

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*.[10]



[10] Image source: Attila Bonyár

How do we do descriptive statistics in R?

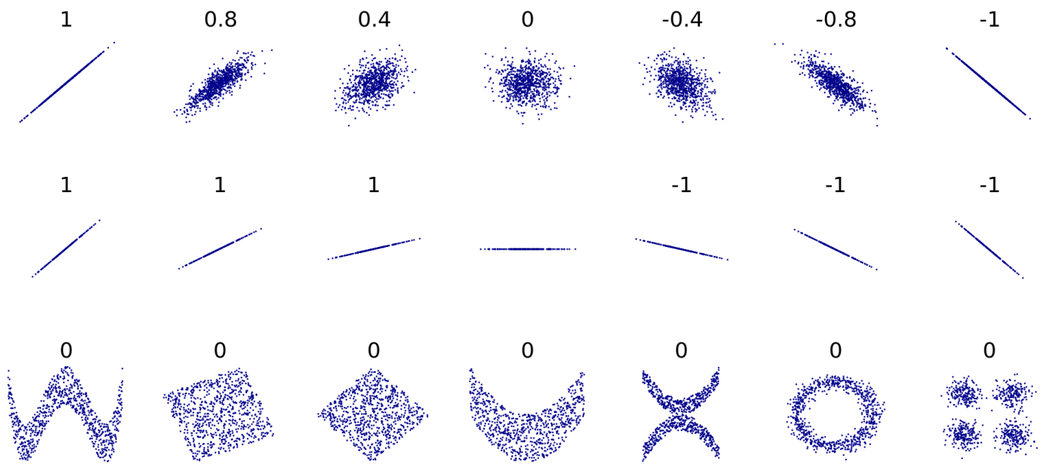
- `describe(var)` calculates all descriptive statistics
- `hist(var)` plots data histogram
- `plot(density(var))` plots the density plot
- `boxplot(var)` plots out a box plot
- `plot(var1, var2)` plots out a scatterplot

Correlation Analysis

What do we look at in bivariate/multivariate analysis?[11]

Correlation and **covariance** measure the extent to which two variables are linearly related. Correlation is the normalized form of covariance.

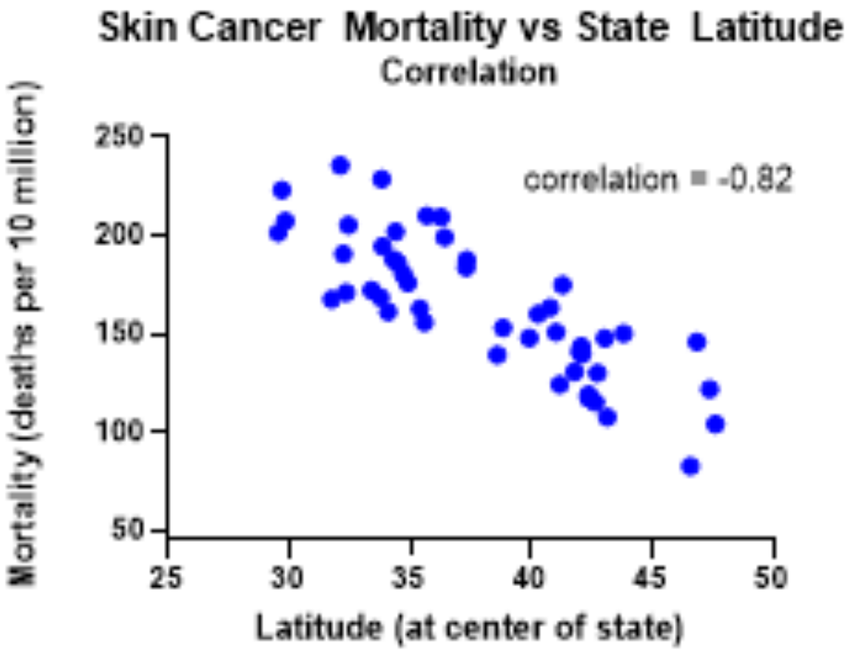
Pearson's r (when the variables are continuous) and *Spearman's* ρ (when one/both are discrete) measure correlation.

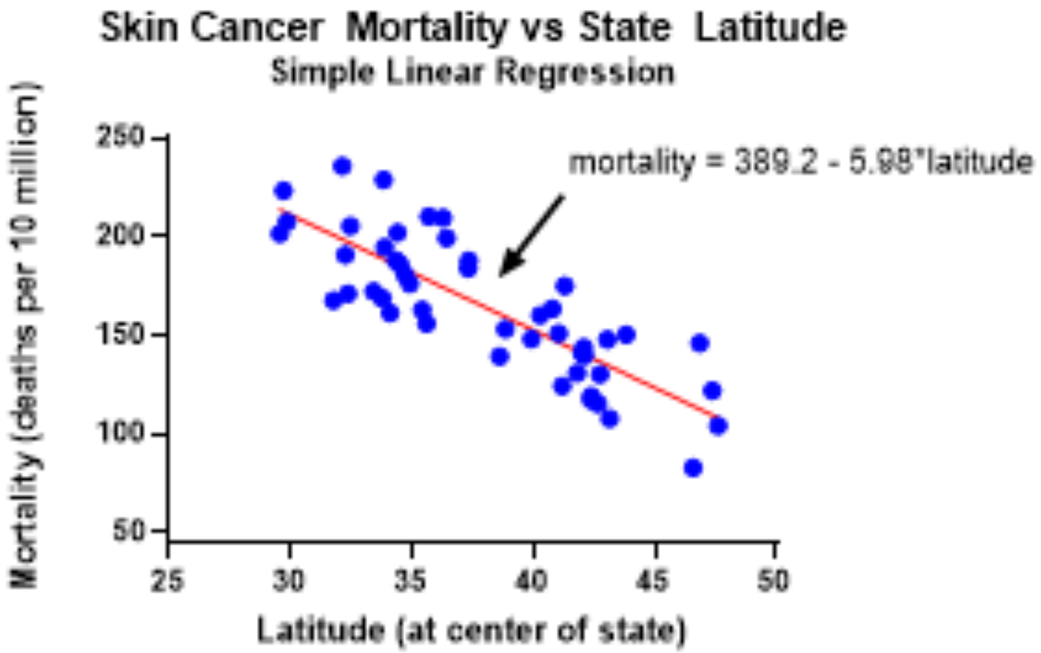


[11] By DenisBoigelot, Imagecreator, CC0

Is correlation descriptive or inferential?[12]

Can be used for *descriptive* or *inferential* statistics.





[12] Image source

How is correlation calculated?

We calculate what is called a **correlation coefficient**.

For a population:

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

For a sample:

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

How do we interpret the correlation coefficient?

Correlation coefficient is a measure of relation between two variables that ranges -1 to 1.

» -1 represents a negative correlation

» 0 represents lack of correlation

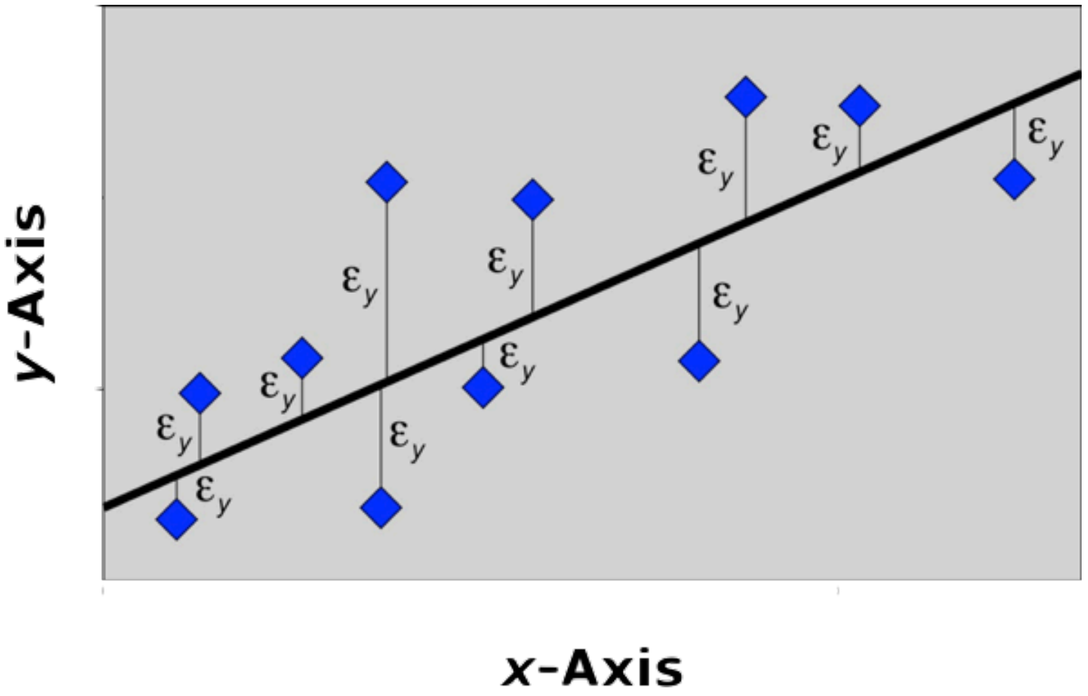
» 1 represents a positive correlation

Simple linear correlation: *Pearson's* r calculates the extent to which the variables are *proportional* or *linearly related* to each other.

r denotes the percent of variation in one variable that is related to the variation in the other. E.g., $r = .70 \twoheadrightarrow$ 49% of the variance is related.

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.

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



How do we do correlation analysis in R?

» `cov(vars)` calculates correlations among all vars

Introduction to Inferential Statistics

*How do we apply **inferential statistics**?*

Inferential statistics involves families of **statistical tests** that aim to establish *statistically significant* differences between distributions.

What is a statistical test?

Definition: A statistical test is a mechanism for assessing whether data provides support for particular hypotheses.

How do we test a hypothesis?

Hypotheses are provisional statements about relationships among concepts. In hypothesis testing, we seek to determine *which* statement data is consistent with.

How many hypotheses do we have consider?

Two mutually exclusive hypotheses/statements about a population:

1. **Null Hypothesis**: Denoted by H_0, it states that a population parameter (e.g., the mean) is equal to a hypothesized value.

2. **Alternative Hypothesis** (or Research Hypothesis): Denoted by H_1 or H_A, it states that the population parameter is smaller, greater, or simply different than the hypothesized value in the null hypothesis.

    - **One-sided hypothesis**: H_1 where the population parameter differs in a particular direction, e.g., higher or lower.

    - **Two-sided hypothesis**: H_1 where the population parameter simply differs in a nondirectional way.

Can you identify what type of hypotheses these are?

» The SUS scores of Google Maps and Apple Maps will not differ.

» The usability of Google Docs and Microsoft Word will be rated differently by users.

» Users will file their taxes faster using TurboTax 2022 than they will using TurboTax 2023.

» Users will reach targets faster using a mouse than a joystick and fastest using a touchpad.

So how do we determine what test to use?

The appropriate test for a given hypothesis-testing scenario is determined by the *data types* of the **input** and **output** variables.

Recap: Data types include:

» Nominal

» Ordinal

» Interval

» Ratio

The distribution of internal and ratio data can be *normal* or *non-normal*.

	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

Footnotes[13] * ** *** **** *****

Note: Rows are *input* variables, columns are *output* variables.

---

[13] Hinton, 2014, Statistics explained

* If data are censored.

** The Kruskal-Wallis test is used for comparing ordinal or non-Normal variables for more than two groups, and is a generalisation of the Mann-Whitney U test. The technique is beyond the scope of this book, but is described in more advanced books and is available in common software (Epi-Info, Minitab, SPSS).

*** Analysis of variance is a general technique, and one version (one way analysis of variance) is used to compare Normally distributed variables for more than two groups, and is the parametric equivalent of the Kruskal-Wallis test.

**** If the outcome variable is the dependent variable, then provided the residuals (see) are plausibly Normal, then the distribution of the independent variable is not important.

***** There are a number of more advanced techniques, such as Poisson regression, for dealing with these situations. However, they require certain assumptions and it is often easier to either dichotomise the outcome variable or treat it as continuous.

Which methods will we cover in this class?

» X^2 👈

» Student's t

» ANOVA

» Regression

Contingency analysis

Pearson Chi-Square — Compares observed vs. expected counts using squared differences. Approximation that follows a chi-square distribution when sample size is large.

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Works well when all expected counts ≥ 5.

Likelihood Ratio (G^2) — Also an approximation, but derived from maximum likelihood estimation (MLE). Compares how likely the observed data are under the null vs. full (saturated) model.

$$G^2 = 2 \sum O \ln\left(\frac{O}{E}\right)$$

Converges to χ^2 as sample size increases.

Fisher's Exact Test — Not an approximation. Computes the exact probability of observing the data under fixed row and column totals (hypergeometric).

$$P(\text{observed}) = \frac{\#\text{ ways to get observed table}}{\#\text{ possible tables}}$$

Sums probabilities of tables as or more extreme than observed. Used when sample sizes or expected counts are small.

How do we use the test outputs?

» Once χ^2 is computed, we compare it to a chi-square distribution with the appropriate degrees of freedom.

» The **p-value** is the probability of obtaining a χ^2 value *as large or larger* than the observed one *if the null hypothesis were true*.

» A small p-value means the observed deviations are unlikely under the null → evidence against the null hypothesis.

» When the deviations are *statistically significant*, the likelihood of it occurring by change is low, determined by a margin, called α level.

» In HCI research, $\alpha = .05$ is used, thus the probability, p, that the difference is occurring by change has to be $p > .05$ to establish *significance*.

How do we conduct a χ^2 test?

Data is summarized in a **contingency table** that cross-tabulates multivariate frequency distributions of variables in a matrix format.[4]

Robot	Reported Gaze Cue
Robovie	Yes
Geminoid	Yes
Robovie	Yes
Geminoid	No
Robovie	Yes
Geminoid	No
Geminoid	No
Robovie	No
Robovie	Yes
Geminoid	No
Robovie	Yes
Geminoid	No
Robovie	No

```
          Reported.Gaze.Cue
Robot      No Yes
  Geminoid 10   3
  Robovie   3  10
```

[4] Data from "Mutlu et al. (2009). Nonverbal leakage in robots: communication of intentions through seemingly unintentional behavior. *HRI 2009.*"

Chi-squared test in R

```
gaze <- read.table('robot-gaze.csv', sep=",", header=TRUE)
chisq.test(table(gaze))

Pearson's Chi-squared test with Yates' continuity correction

data:  table(gaze)
X-squared = 5.5385, df = 1, p-value = 0.0186
```

Chi-squared test in JMP

Analyze > Fit X by Y

N	DF	-LogLike	RSquare (U)
26	1	3.9765190	0.2207

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	7.953	0.0048*
Pearson	7.538	0.0060*

Fisher's Exact Test	Prob	Alternative Hypothesis
Left	0.9994	Prob(Robot=Robovie) is greater for Reported Gaze Cue=No than Yes
Right	0.0085*	Prob(Robot=Robovie) is greater for Reported Gaze Cue=Yes than No
2-Tail	0.0169*	Prob(Robot=Robovie) is different across Reported Gaze Cue

What's Next

» Inferential statistics, cont'd

» Next assignment: conducting descriptive & inferential statistics on your survey assignment data