# Rose's Statistics

Rose Enos

2025

Adapted from

- *OpenIntro Statistics*, 4th edition by David Diez, Mine Çetinkaya-Rundel, and Christopher D Barr
- *Boolean Logic and Discrete Structures*, September 2024 Edition by Sandy Irani
- Lectures by Professor Mine Dogucu at the University of California, Irvine for STATS 67
- Lectures by Professor Irene Gassko at the University of California, Irvine for I&C SCI 6B and I&C SCI 6D

## Contents

# 1 Data

## 1.1 Collection

**Data** are observations. **Statistics** is the study of collecting, analyzing, and drawing conclusions from data. A **summary statistic** is a value describing a data set.

A **case**, or **observational unit**, is the set of data from a single situation. A **variable** describes a certain datum of every case. A **data matrix** represents cases as rows and variables as columns.

A **numerical** variable holds an arithmetically meaningful value. A **discrete** numerical variable holds one of finitely many or countably infinitely many values. A **continuous** numerical variable holds one of uncountably infinitely many values.

A **level** is a category that contains a case. A **categorical** variable holds a level. An **ordinal** categorical variable holds an ordered level. A **nominal** categorical variable holds an unordered level.

Two variables are **associated**, or **dependent**, if they appear connected. They are **negatively** associated if their directions are opposite. They are **positively** associated if their directions are the same. Two variables are **independent** if they do not appear connected.

In a hypothetically causal association, the **explanatory**, or **independent** or **predictor**, variable influences the **response**, or **dependent**, variable.

An **observational study** collects data by only observation to establish simple association. A **cohort** is a set of similar individuals in a study.

An **experiment** collects data by treatment and observation to establish causation. A **randomized** experiment randomly assigns treatment to cases. A **placebo** is a fake treatment.

A **population** is a set of all similar individuals. A **sample** is a subset of a population.

**Anecdotal evidence** is a data set from a very small sample whose elements may or may not be **representative** of the population.

Samples should be **randomly** selected to reduce **bias** toward or away from certain parts of the population.

The **non-response rate** of a survey is the rate of non-response within the sample. A nonzero non-response rate introduces **non-response bias**. Then the survey results may or may not be representative of the population.

In a **convenience sample**, more accessible cases have higher chances of being selected. Then the sample is disproportionately more representative of more accessible cases.

**Observational data** is data where no treatment is applied. A **confounding variable**, or **lurking variable** or **confounding factor** or **confounder**, is a variable associated with both the explanatory and response variables. Where there are confounding variables, observational data cannot establish causation.

A **prospective study** selects a sample before the target data exists. A **retrospective study** selects a sample after the target data exists.

In **simple random sampling**, each case in the population has the same, independent chance of being selected.

A **stratum** of the population is a subset of similar cases. The strata partition the population. In **stratified sampling**, another sampling method is applied to each stratum, and the sample contains the selected cases from each stratum.

A **cluster** of the population is a subset of cases. The clusters partition the population and are similar. In **cluster sampling**, another sampling method is applied to the clusters, and the sample contains all cases from each selected cluster. In **multistage sampling**, another sampling method is applied to each selected cluster, and the sample contains the selected cases from each selected cluster.

Randomized experiments employ four principles:

- **Controlling**: confounding is reduced by equalizing procedures.
- **Randomization**: confounding and bias are reduced by random assignment.
- **Replication**: reliability is increased by large sampling and study replication.
- **Blocking**: results may be specialized by assigning similar cases to **blocks** before random assignment.

The **treatment group** receives treatment. The **control group** receives no treatment.

A study is **blind** if patients do not know which group they are in. A **placebo** is a fake treatment given to control patients to make the study blind. The **placebo effect** is a slight, real change due to placebo.

A study is **double blind** if neither patients nor researchers know which group each patient is in.

## 1.2 Summarization

A **scatterplot** plots a numerical variable against another numerical variable and places one point for each case.

A **dot plot** plots a numerical variable and places one point for each case.

The **mean**, or **average**, measures the center of a data **distribution** by

$$\bar{x} = \frac{\sum_i x_i}{n}$$

for a sample and $\mu$ for a population. A **weighted mean** is

$$\bar{x} = \frac{\sum_S (\bar{x}_S |S|)}{\sum_S |S|}$$

where $S$ are partitioning subsets of the data set.

A **bin** is a set of numerical data within a certain range. A **histogram** plots a numerical variable against ordered bins of another numerical variable. A histogram shows **data density** and shape.

A **long tail** is an area of low density. A distribution is **right skewed** if it has a right long tail. A distribution is **left skewed** if it has a left long tail. A distribution is **symmetric** if it is neither right nor left skewed.

A **mode** is an area of high density. A distribution may be **unimodal** (with one mode), **bimodal** (with two modes), or **multimodal** (with more than two modes).

The **deviation** of a datum is its displacement from the mean. The **variance** is

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

for a sample and

$$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$$

for a population. The **standard deviation** is $s$ for a sample and $\sigma$ for a population.

The **median** is the midpoint of the data set. If there are an even number of data, the median is the mean of the two middlemost data. If there are an odd number of data, the median is the middlemost datum.

The **first quartile** $Q_1$ is the median of the lesser half of the data. The **third quartile** $Q_3$ is the median of the greater half of the data. The **interquartile range** (IQR) is

$$IQR = Q_3 - Q_1$$

The **whiskers** are the ranges

$$[Q_1 - 1.5 \times IQR, Q_1)$$

$$(Q_3, Q_3 + 1.5 \times IQR]$$

An **outlier** is a point not in the IQR or the whiskers. Outliers may indicate skew, data collection errors, or interesting properties of the data.

A **box plot** plots a numerical variable by summary statistics:

- A vertical line represents the median.
- Two vertical lines represent the first and third quartiles.
- Two horizontal lines between the ends of the first and third quartile lines represent the interquartile range.
- Two horizontal lines outward from the box represent the whiskers.
- Additional points represent the outliers.

**Robust statistics** are not or slightly affected by outliers. The median and IQR are robust statistics. The mean and standard deviation are not robust statistics.

A **transformation** is a rescaling of data using a function. Transformations commonly use the logarithm, the square root, and the inverse. Transformations may reveal associations, reduce skew, show shape, or make associations linear.

An **intensity map** plots a numerical variable on a geographic map and represents the value in each area by color on a scale.

A **contingency table** shows a categorical variable as rows and another categorical variable as columns. The cells show the number of data in the intersecting categories. The table shows **row totals** and **column totals**, the sum of the cells in each row or column, respectively.

A contingency table may alternatively show the proportions of matching data, only one categorical variable, **row proportions**, or **column proportions**.

A **bar plot** plots the frequencies or relative frequencies of matching data as vertical bars against a categorical variable. A **stacked bar plot** divides each bar by another categorical variable. A **side-by-side bar plot** places each part of a divided bar as its own bar.

A **mosaic plot** breaks a square into columns with widths according to relative frequency in a categorical variable, and breaks each column into rectangles with heights according to relative frequency in another categorical variable.

A **pie chart** shows relative frequencies of a categorical variable as slices of a circle.

A **side-by-side box plot** plots a box plot of a numerical variable for each category of a categorical variable.

A **hollow histogram** plots a histogram of a numerical variable for each category of a categorical variable.

**Random noise** is variation in a sample but that is not representative of the population.

In an experiment, the **independence model** is the assumption that two variables are independent. The **alternative model** is the assumption of a certain dependence between the variables. **Statistical inference** is the study of model selection.

A **simulation** implements the independence model to determine the hypothetical distribution of data from chance alone.

# 2 Probability

## 2.1 Axioms

An **experiment** is a procedure that gives a random **outcome** $s$. A **process** is a sequence of experiments. The **sample space** $S$ is the set of possible outcomes. An **event** $E$ is a subset of the sample space. An **elementary event**, or **atomic event**, is an event that contains one outcome.

**Discrete probability** is the study of experiments with finite or countably infinite sample spaces. The **probability** $P(s)$ of $s$ is the proportion of occurrences of $s$ in infinite repetitions of an experiment. The

A **probability distribution** over $S$ is

$$P : S \to \{x \in \mathbb{R} : 0 \leq x \leq 1\}$$

such that

$$\sum_{s \in S} P(s) = 1$$

The probability of $E$ is

$$P(E) = \sum_{s \in E} P(s)$$

An **impossible event** has probability 0. A **certain event** has probability 1.

The **law of large numbers** states that the proportion $\hat{p}_n$ of occurrences with a particular outcome $x$ converges to the probability $p = P(x)$ of that outcome:

$$\lim_{n \to \infty} p_n = p$$

The **uniform distribution** gives

$$P(s) = \frac{1}{|S|}$$

$$P(E) = \frac{|E|}{|S|}$$

By the inclusion-exclusion principle, the **joint probability** of events is

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Two events are **mutually exclusive** if they are disjoint. Then

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

The **complementary event** of $E$ is

$$\bar{E} = S - E$$

It follows that

$$P(\bar{E}) = 1 - P(E)$$

The **conditional probability** of an **event of interest** $E_1$ given a **condition** $E_2$ is

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

For a uniform distribution,

$$P(E_1 \mid E_2) = \frac{|E_1 \cap E_2|}{|E_2|}$$

It follows that

$$P(E_1 \mid E_2) + P(\bar{E}_1 \mid E_2) = 1$$

Two events are **independent** if

$$\begin{aligned} & P(E_1 \mid E_2) = P(E_1) \\ \iff & P(E_1 \cap E_2) = P(E_1)P(E_2) \\ \iff & P(E_2 \mid E_1) = P(E_2) \end{aligned}$$

$n$ events are **mutually independent** if

$$P(E_1 \cap \cdots \cap E_n) = P(E_1) \cdots P(E_n)$$

**Bayes' theorem** states

$$P(E_1 \mid E_2) = \frac{P(E_2 \mid E_1)P(E_1)}{P(E_2 \mid E_1)P(E_1) + P(E_2 \mid \bar{E}_1)P(\bar{E}_1)}$$

where $P(E_1) \neq 0 \neq P(E_2)$. **Bayesian statistics** is the study of belief revision by Bayes' theorem.

Sampling **with replacement** maintains selected cases in selection consideration. Sampling **without replacement** removes selected cases from selection consideration. For a sample that is large relative to the population, sampling without replacement makes selections dependent.

A **Venn diagram** shows two distinct, intersecting circles representing events $A, B$. The intersection holds the elements of $A \cap B$. The non-intersecting part of each circle holds the rest of its respective elements.

A **tree diagram** is a subset diagram of $S$ by disjoint events and shows conditional probabilities for each event. The first level is the **primary branch**. The second level is the **secondary branch**.

## 2.2 Discrete Distributions

A **random variable** on a sample space $S$ is

$$X : S \to \mathbb{R}$$

$$(X = r) = \{s \in S : X(s) = r\}$$

$$P(X = r) = \sum_{s \in (X=r)} P(s)$$

The **distribution** over $X$ is

$$\{(r, P(X = r)) : r \in X(S)\}$$

The **expected value** of $X$

$$E[X] = \sum_{s \in S} X(s)P(s) = \sum_{r \in X(S)} rP(X = r)$$

represents the expected average value from infinitely many repetitions. Expected values are combine linearly.

The **variance** of $X$ is

$$\sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i)$$

Then the **standard deviation** is $\sigma$. Independent variances combine as

$$X = \sum_i c_i X_i \implies \sigma^2 = \sum_i c_i^2 \sigma_i^2$$

A **Bernoulli trial** is an experiment with possible outcomes **success**, with value 1 and probability $p$, and **failure**, with value 0 and probability $q$. Trials in a Bernoulli process are mutually independent with

$$\sigma = \sqrt{p(1-p)}$$

A **sample proportion** $\hat{p}$ is the proportion of successes in a sample of trials.

If $X$ describes the number of successes in a process of length $n$, then the **binomial distribution** is the distribution over $X$ with

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

$k$ is the number of successes. A distribution is binomial if

1. The trials are independent.
2. The number of trials is fixed.
3. Each trial outcome is a success or a failure.
4. The probability of success is the same for each trial.

The **geometric distribution** is an exponentially decaying distribution that describes how many failures are observed before a success for an independent and identically distributed Bernoulli random variable.

$$f(k) = (1-p)^k p$$

$$\mu = \frac{1}{p}$$

$$\sigma = \sqrt{\frac{1-p}{p^2}}$$

$k$ is the number of failures.

The **negative binomial distribution** describes how many failures are observed before a specified number of successes is reached. A distribution is negative binomial if

1. The trials are independent.
2. Each trial outcome is a success or a failure.
3. The probability of success is the same for each trial.
4. The last trial is the last trial before the specified number of successes is reached.

$$f(k) = \binom{r+k-1}{k} p^r (1-p)^k$$

$k$ is the number of failures and $r$ is the number of successes.

The **Poisson distribution** describes the number of successes in a large population over a unit of time. The **rate** $\lambda$ is the average number of successes per unit time. A distribution may be Poisson if

1. The population is large.
2. The rate is constant across units of time.

$$f(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$k$ is the number of successes.

The **chi-square distribution** is positive and right skewed. The **degrees of freedom** $\nu$ determines the distribution and is positively related to symmetry, center value, and variability.

## 2.3 Continuous Distributions

A **probability density function**, or **density** or **distribution**, of a continuous random variable is a smooth curve from a histogram with infinitesimally small bins. An area under the curve between two points is the probability of observing a value between those points.

The **normal curve**, or **normal distribution**, is a symmetric, unimodal, bell-shaped distribution described by

$$N(\mu, \sigma)$$

where $\mu, \sigma$ are **parameters**. The **standard normal distribution** is

$$N(\mu = 0, \sigma = 1)$$

The **z-score** of $x$ is

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

z-scores help compare probabilities of points in symmetric distributions.

The **percentile** of a point is the percentage of points that lie below it. A percentile can be found by a computer or a probability table.

The **68-95-99.7 rule**, or **empirical rule**, states that the probabilities of falling within 1, 2, or 3 standard deviations of the mean are 68%, 95%, and 99.7%, respectively.

The normal distribution

$$N\left(\mu = np, \sigma = \sqrt{np(1-p)}\right)$$

approximates the binomial distribution if

1. $np \geq 10$
2. $n(1-p) \geq 10$

Then the area under the normal distribution in domain $[a-0.5, b+0.5]$ approximates the area under the binomial distribution in domain $[a, b]$.

The **t distribution** has similar shape to the standard normal distribution, but has thicker tails. $\nu$ determines the distribution and is positively related to resemblance to the normal distribution.

The **f distribution** has a positive, decaying shape. $\mu_1$ and $\mu_2$ determine the distribution.

# 3 Inference

## 3.1 Methods

A **parameter** is a value that describes a population. A **point estimate** is an estimate of the parameter from a sample. The **error** is the displacement of the point estimate from the parameter.

The **sampling error**, or **sampling uncertainty**, describes the tendency of errors. The **sample size** is the number of observations in the sample. **Bias** is systemic incorrect estimation of the parameter.

The **sampling distribution** is the distribution of point estimates of infinitely many random samples.

- Center: The mean is the parameter.
- Spread: The **standard error** is the standard deviation.
- Shape: The distribution is approximately normal.

An **unbiased estimator** is a point estimate whose sampling distribution mean is the parameter.

A **confidence interval** is a range of plausible values for the parameter. The **confidence level** is the approximate proportion of possible samples whose confidence intervals would contain the parameter. The **critical value** is the number of standard deviations that captures the confidence level in an appropriate distribution for the type of inference. The **margin of error** is the product of the critical value and the standard error. Then the confidence interval includes the margin of error on either side of the point estimate. We can be the confidence level percent confident that the parameter lies in the confidence interval.

A **hypothesis test** determines whether sample data conforms to a hypothesis about the population. The **null hypothesis** $H_0$ is a claim to be tested. The **alternative hypothesis** $H_a$ is a claim that implies that the null hypothesis is false. With sufficient evidence, we can reject the null hypothesis. However, we can only accept a hypothesis with complete evidence. The null hypothesis may equate the parameter to a **null value**.

A **type 1 error** is incorrectly rejecting the null hypothesis. A **type 2 error** is incorrectly failing to reject the null hypothesis. The **significance level** $\alpha$ is the target proportion of type 1 errors committed. The significance level is the complement of the confidence level.

The **test statistic** is a relevant statistic describing the testable attributes of the data. The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as the point estimate, given that the null hypothesis is true. The **null distribution** is the sampling distribution of the null value. Then the p-value is the area outside the confidence level. If the p-value is less than the significance level, then the error is **statistically significant** and we reject the null hypothesis.

## 3.2 Single Proportion

For a single proportion, the **central limit theorem** states that, if

- observations are independent by random assignment or random selection;
- $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ (the **success-failure condition**);
- $n$ is at most the population size

then the sampling distribution is approximately normal with

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The critical value $z^*$ is from the normal distribution. Then the confidence interval is

$$\hat{p} \pm z^* \sigma_{\hat{p}}$$

The hypotheses of a hypothesis test are

- $H_0$: $p = p_0$
- $H_a$: $p \neq p_0$

The test statistic is the **z-score**

$$Z = \frac{\hat{p} - p_0}{\sigma_{\hat{p}}}$$

and represents the distance of the point estimate from the null value in standard deviations. The p-value the area of the tails past $Z$ on the normal distribution.

## 3.3 Double Proportions

For two proportions, if

- observations are dependent within and between two samples;
- the success-failure condition holds for each sample

then

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If $H_0$ states $\hat{p}_1 = \hat{p}_2$, then it is sufficient that the success-failure condition holds for the **pooled proportion**

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Then

$$\sigma_{\hat{p}_1-\hat{p}_2} \approx \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}$$

The hypotheses of a hypothesis test are

- $H_0$: $p_1 = p_2$
- $H_a$: $p_1 \neq p_2$

The z-score is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sigma_{\hat{p}_1-\hat{p}_2}}$$

## 3.4 Chi-Square

For data in $k > 2$ bins, given a null distribution, the standardized difference in a bin is

$$Z_i = \frac{n\hat{p}_i - np_i}{\sqrt{np_i}}$$

The test statistic

$$X^2 = \sum_{i=1}^{k} Z_i^2$$

represents the tendency of the observed distribution to deviate from the null distribution. If

- observations are independent;
- the null distribution places at least five observations in each bin

then the distribution approximates the chi-square distribution with

$$\nu = k - 1$$

The hypotheses are

- $H_0$: The data are equally distributed into bins.
- $H_a$: the data are not equally distributed into bins.

The p-value is the area of the tail past $X^2$.

We can convert data over discrete time by stating that the null distribution is geometric, with $p$ calculated from the dataset disregarding time. Then the expected counts are given by $f$.

If the null hypothesis for a two-way table with $k$ rows and $l$ columns is that the variables are independent, then

$$np_{i,j} = \frac{np_i np_j}{n}$$

and

$$\nu = (k-1)(l-1)$$

## 3.5 Single Mean

For a single mean, the central limit theorem states additionally that if

- observations are independent;
- the sample distribution is approximately normal:
  - If $n < 30$ and there are no clear outliers, then the population distribution must be approximately normal.
  - If $n \geq 30$, then it is sufficient that there are no clear outliers.

then the sampling distribution approximates the t distribution with

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$
$$\nu = n - 1$$

The confidence interval is

$$\bar{x} \pm t^*_\nu \sigma_{\bar{x}}$$

The hypotheses of a hypothesis test are

- $H_0$: $\mu = \mu_0$
- $H_a$: $\mu \neq \mu_0$

The test statistic is the **t-score**

$$T = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

The p-value is from $T$ on the t distribution.

Two data sets are **paired** if they have a natural bijective relationship. Then inference can be done on the set of the differences of each mapped pair.

## 3.6 Double Means

For two means of non-paired data, if

- observations are independent within and between samples;
- each sample distribution is approximately normal

then the sampling distribution approximates the t distribution with

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$
$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\nu \approx (n_1 - 1)(n_2 - 1)$$

If

$$\sigma_1 \approx \sigma_2$$

then

$$\bar{s} = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}$$

$$\sigma_{\bar{x}_1 - \bar{x}_2} \approx \sqrt{\frac{\bar{s}^2}{n_1} + \frac{\bar{s}^2}{n_2}}$$

$$\nu \approx n_1 + n_2 - 2$$

The hypotheses of a hypothesis test are

- $H_0$: $\mu_1 = \mu_2$
- $H_a$: $\mu_1 \neq \mu_2$

The t-score is

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

An **effect size** of an experiment is a difference in output against the control. The **power** $1 - \beta$ is the probability of rejecting the null hypothesis given an actual non-null effect size, and represents the probability of detecting an effect that is both statistically and practically significant. If the minimum practically significant effect size is $(\bar{x}_1 - \bar{x}_2)_a$, then an adequate sample size per experimental group is

$$n = \left\lceil \frac{(z^* + Z_{1-\beta})^2}{(\bar{x}_1 - \bar{x}_2)_a^2}(\sigma_1^2 + \sigma_2^2) \right\rceil$$

## 3.7 ANOVA

**Analysis of variance**, or **ANOVA**, is a hypothesis test on means across $k$ groups, given

- observations are independent within and between groups;
- each sample distribution is approximately normal;
- the variability of each group is approximately equal.

then inference can be done with the f distribution.

The degrees of freedom between groups is

$$\nu_1 = k - 1$$

The **sum of squares between groups** is

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

The **mean square between groups**

$$MSG = \frac{1}{\nu_1} SSG$$

represents the variability between groups.

The degrees of freedom within groups is

$$\nu_2 = n - k$$

The **sum of squares total** is

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The **sum of squared errors** is

$$SSE = SST - SSG$$

The **mean square error**

$$MSE = \frac{1}{\nu_2} SSE$$

represents the variability within each group.

The hypotheses are

- $H_0$: The variances are equal between groups.
- $H_a$: The variances are not equal between groups.

The **f-score** is

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

The p-value is calculated for the upper tail of the distribution.

**Multiple comparisons** is performing difference-of-means tests on each relevant pair of groups. The **Bonferroni correction** states that multiple comparisons tests should be performed with significance level

$$\alpha^* = \frac{\alpha}{K}$$

where $K$ is the number of tests. If all possible pairs are relevant, then

$$K = \frac{k(k-1)}{2}$$

# 4 Regression

## 4.1 Simple

**Simple linear regression** fits a model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

to data where $x$ is the explanatory variable, $y$ is the response variable, and $\varepsilon$ is the error. If

- observations are independent;
- the relationship is approximately linear;
- the distribution of residuals is approximately normal;
- the variability of residuals is approximately constant along the entire line

then the line estimate is

$$\hat{y} = b_0 + b_1 x$$

The **residual** of a point is

$$e_i = y_i - \hat{y}_i$$

The **residual plot** is a scatterplot of the residuals of the data.

The **correlation**

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y} \right)$$

represents the strength of the linear relationship between $x$ and $y$. Perfect negative and positive relationships respectively give $r = -1$ and $r = 1$.

The **least squares criterion** is

$$\sum_{i=1}^{n} e_i^2$$

The **least squares line** minimizes the least squares criterion. The least squares line has

$$b_1 = \frac{s_y}{s_x} r$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$b_1$ is the estimated difference in $y$ upon a unit increase in $x$. If the data set extends to $x = 0$, then $b_0$ is the average value of $y$ at $x = 0$. **Extrapolation** is estimating $y$ beyond the domain of the data set, and is usually unreliable.

The **coefficient of determination**

$$r^2 = 1 - \frac{s_e^2}{s_y^2}$$

represents the amount of variation in $y$ that is explained by the least squares line.

**Leverage** is the extent to which a point influences the least squares line. An outlier in $x$ has high leverage. An **influential point** is an outlier in $x$ that would be an outlier in $e$ if it did not influence the least squares line.

The confidence interval of a parameter is

$$b_i \pm t^*_\nu s_{b_i}$$

The hypotheses of a hypothesis test are

- $H_0$: $\beta_1 = 0$
- $H_a$: $\beta_1 \neq 0$

$$T = \frac{b_1 - 0}{s_{b_1}}$$

The p-value is from $T$.

A **diagnostic plot** is a plot of model data that visualizes the conditions for regression. Normality is checked on a histogram of residuals. Variability is checked on a scatterplot of absolute residuals against estimated values. Independence is checked on a scatterplot of residuals against observation time. Variability for predictor variables is checked on box plots of residuals for each level. Linearity and normality for numerical variables is checked on a scatterplot of residuals for each variable.

Data may be improved to fit the conditions for regression by transforming some variables or truncating outliers.

## 4.2 Multiple

**Multiple linear regression** fits a model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

to data where $x_i$ are explanatory variables. If

- observations are independent;
- each explanatory relationship is approximately linear;
- the distribution of the residuals is approximately normal;
- the variability of residuals is approximately constant along the entire line

then the line estimate is

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$

Two explanatory variables are **collinear** if they are correlated. Models usually avoid collinearity to avoid complication.

The **coefficient of multiple determination**, or **adjusted coefficient of determination**

$$r_a^2 = 1 - \frac{s_e^2}{s_y^2}\frac{n-1}{n-k-1}$$

corrects bias in the coefficient of determination.

For numerical $y$ against categorical $x$, an **indicator variable** is a Bernoulli variable associated with each category. The indicator variables of a single categorical variable are mutually exclusive. The **reference level** is the category with respect to which the other categories are measured. Then $b_0$ and $b_0 + b_i$ are the respective estimated value of $y$ upon the reference level and upon the $i$th level.

## 4.3 Logistic

A **generalized linear model** fits a probability distribution to data. The **logit transformation** is

$$\operatorname{logit} p = \log_e \frac{p}{1-p}$$

**Logistic regression** fits a generalized linear model

$$\operatorname{logit} p = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

to data where $p$ is the probability of success in the response variable. If

- observations are independent;
- each explanatory relationship is approximately linear with $\operatorname{logit} p$

then the line estimate is

$$\hat{\operatorname{logit}}\, p = b_0 + b_1 x_1 + \cdots + b_k x_k$$

Multiple linear regression gives a model for $\operatorname{logit} p$. The **Akaike information criterion** is an analog to the coefficient of multiple determination. Linearity is checked on a scatterplot of confidence intervals of sample proportions of buckets against the estimated proportions of the buckets.

# 5 Models

## 5.1 Selection

**Model selection** is eliminating less-important explanatory variables from a model to potentially increase accuracy. A model is **full** if it includes all available explanatory variables, and is otherwise **parsimonious**.

**Stepwise** model selection compares models that differ in inclusion of one explanatory variable. **Backwards elimination** on the full model is repeatedly eliminating the least important variable. **Forward selection** on the empty model is repeatedly adding the most important variable. The selected model is the better of the two reached by backwards elimination and forward selection.

Backwards elimination and forward selection may determine which change would most increase the coefficient of determination or its analog. Then the selected model will be more accurate. Alternatively, backwards elimination and forward selection may determine which variable has the highest or lowest p-value, respectively, and is statistically insignificant or statistically significant, respectively. Then the selected model will be more meaningful.

## 5.2 Validation

**Model validation** is determining the accuracy of a model to a population. A model **overfits** a sample if the model describes the sample well but the population poorly.

A data set may be randomly split into **training data** and **test data**, each of which gives a model with the same parameters. Then the training model is valid if it is similar to the test model.