TECHNISCHE UNIVERSITÄT WIEN

VIENNA UNIVERSITY OF TECHNOLOGY

Biostatistics

Achim Zeileis

http://www.ci.tuwien.ac.at/~zeileis/

Outline syllabus

- Basic concepts and designs, terminology,
- Descriptive statistics, graphical methods,
- Classical and nonparametric inference,
- (Generalized) Linear models,
- Contingency tables,
- Survival analysis.

Basic concepts

- observational study vs. experiment,
- study unit, cohort,
- clinical trials,
- laboratory and comparative experiment,
- placebo,
- controlled vs. uncontrolled designs,
- randomization (simple, restricted, stratified),

Basic concepts

- parallel, in series, cross-over design,
- blind and double blind trials,
- longitudinal vs. cross-sectional study,
- prospective vs. retrospective study.

Basic concepts

History:

- prior to 1950: haphazard development of medicine. Medical literature emphasized individual case studies ⇒ unscientific and inefficient.
- 1948: UK Medical Research Concil used randomized control in a Stroptomycin trial for Tuberculosis.
- 1954: field trial of Salk Polio vaccine in certain areas of the U.S.
  2nd grade children were offered treatment, 1st and 3rd grade control (total 1 million children), but *volunteer bias* and

Basic concepts

lack of blindness ⇒ valid control still difficult. Children *not* agreeing to vaccination had a lower incident rate of incidence/paralysis/death.

- Further 0.8 million took part in a randomised double-blind trial. Every child received injection but half these did not contain vaccine and child/parent/evaluating physician did not know which.
  - incidence reduced by 50%
  - paralysis from those getting polio 70% less
  - no deaths in vaccine group (4 in placebo group)

Basic concepts

John Stuart Mill established three criteria for inferring causality:

- covariation,
- temporal precedence,
- elimination of alternative explanations.

How are these accomplished in clinical trials?

- clinical treatment and assessment
- randomization, blindness of trials

Basic concepts

The 4 stages of a clinical trial programme:

1. Clinical pharmacology & toxicity concerned with drug safety, performed on non-patients ($n = 10 - 50$),
2. Initial clinical investigation for saftey & efficacy ($n = 50 - 100$),
3. Full-scale evaluation of drug vs. control ($n = 100 - 1000$),
4. Post-marketing surveillance of side effects etc.

1 in 10,000 drugs get to clinical stage, of these 1 in 5 reach marketing.

Basic concepts

The protocol contains all details of the trial conduct and is needed to gain permission to conduct the trial. It should contain items on

- Purpose
  - motivation
  - aims
- Design & conduct
  - patient selection critieria (inclusion / exclusion)
  - number of patiens
  - schedule: assignment, design, randomization, evaluation
  - principal response
  - forms: "informed consent", monitoring, record
  - analysis methods

Basic concepts

Protocol deviations:

- Things always go wrong: patients drop out, do not meet the inclusion/exclusion criteria, forget to take medicine, take too much, take other medicine, ...,
- Analysis: per protocol vs. intention to treat,
- Example: surgery vs. radiotherapy for cancer,
- Record protocol deviations.

Basic concepts

Be careful with:

- Multiplicity: multiple responses or tests,
- Interim analysis,
- Bonferroni correction: $\alpha = 1 - (1 - \varepsilon)^k \approx k\varepsilon$,
- Combination of trials,
- Simpson's paradox,
- Publication bias.

Descriptive Statistics

A *variable* is a quantity that may vary from object to object. A *sample* (data set) is a collection of values of one or more variables. A member of the sample is called an *element*.

Taxonomy of variables

- qualitative vs. quantitative
- discrete vs. continous

In R qualitative variables are coded as *factors*:

```
factor(x, levels = sort(unique(x)), labels, ordered = FALSE,
       exclude = NA)
```

Descriptive Statistics

Scale of variables

- nominal (diseases, marital status),
- ordinal (quality of teaching),
- interval (temperatures, dates),
- scale (distance, age, height).

Descriptive Statistics

A *statistic* is a numerical characteristic of a sample.

Statistics derived from counts

- contingency tables,
- empirical frequency distribution,
- empirical cumulative distribution function,
- mode.

Descriptive Statistics

Statistics derived from ranks

- quantiles, percentiles, median,
- min, max, range, IQR, five number summary.

Statistics derived from moments

- (sample) mean, variance,
- skewness, kurtosis.

Descriptive Statistics

Plots

- boxplot,
- histogram,
- mosaic display,
- stem and leaf plot,
- cleveland dotplot,
- *no* pie charts.

Exploratory Analysis

Univariate: 1 numerical variable

numeric description: univariate statistics like mean, variance, standard deviation, five number summary.

visualization: histogram, boxplot.

Univariate: 1 categorical variable

numeric description: contigency tables (absolute and relative frequencies).

visualization: barplot, (pie chart).

Exploratory Analysis

Bivariate: 2 numerical variables

numeric description: correlation coefficients.

visualization: scatter plot.

Bivariate: 2 categorical variables

numeric description: contigency tables, odds ratio.

visualization: mosaic plot.

Exploratory Analysis

Bivariate: 1 dependent numerical and 1 explanatory categorical variable.

numeric description: groupwise statistics (e.g., means).

visualization: parallel boxplots.

Bivariate: 1 dependent categorical and 1 explanatory numerical variable.

Idea: transform the numerical variable into a categorical and then proceed as before.

numeric description: discretized contingency tables

visualization: discretized mosaic plots.

Inference

A *parameter* is a numerical characteristic of a population.
Correspondence: statistic::sample, parameter::population.

Common approaches to making statements about population parameters are *estimation* and *hypothesis testing*.

A *random variable* is associated with a random sample. If a statistic is computed from a random sample, it is a random variable. Its distribution is called the *sampling distribution*.

Inference

- **null hypothesis**: a collection of hypothesized values for a parameter (H_0).
- **alternative hypothesis**: a collection of values for a parameter which will be considered if H_0 is rejected (H_A or H_1).
- **rejection region**: set of values of a statistic for which H_0 is rejected. The boundaries are called *critical values*.
- **type I error**: error if H_0 is rejected when it is true.
- **type II error**: error if H_0 is not rejected when it is false.

Inference

- **significance level**: the probability of a type I error, usually denoted α. $1-\alpha$ is called the *confidence level*.
- **power**: probability to reject H_0 when it is false, usually denoted $1-\beta$. Hence, β is the probability of a type II error.
- **p value**: the value p ($0 < p < 1$), such that for $\alpha > p$ the test rejects H_0 and for $\alpha < p$ it does not. More intuitively, p is the probability under H_0 of observing a value at least as unlikely as the value of the test statistic.

Inference

Notation:

Let $\xi_1, \ldots, \xi_n$ be independently identically distributed (i.i.d.) random variables with distribution $F(\theta)$ and with observations $x_1, \ldots, x_n$.

$$\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i.$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\xi_i - \bar{\xi})^2$$

Inference

Inference about the mean and variance of a population (exact for $F(\theta) = \mathcal{N}(\mu, \sigma^2)$):

$$Z = \frac{\bar{\xi} - \mu}{\sigma/\sqrt{n}}, \qquad Z \sim \mathcal{N}(0,1).$$

$1-\alpha$ confidence interval for μ:

$$\bar{\xi} \pm z_{1-\alpha/2}\sigma/\sqrt{n}.$$

$$X^2 = \frac{(n-1)s^2}{\sigma^2}, \qquad X^2 \sim \chi^2_{n-1}.$$

Inference

One-sample t test: variance unknown

$$t = \frac{\bar{\xi} - \mu}{s/\sqrt{n}}, \qquad t \sim t_{n-1}.$$

Two-sample t test: two independent samples with variances unknown, but *equal*

$$t = \frac{(\bar{\xi}_1 - \bar{\xi}_2) - (\mu_1 - \mu_2)}{s_p \sqrt{1/n_1 + 1/n_2}}, \qquad t \sim t_{n_1+n_2-2},$$

with

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Inference

If the variances are unknown and not equal an approximate t can be used (Welch approximation).

A special case is that of *paired* data, e.g. gain of sleep after usage of drug. Convert to the one-sample case by taking differences $\Rightarrow$ independence within the pairs is not needed.

In R:

```
t.test(x, y = NULL, alternative = "two.sided", mu = 0, paired = FALSE,
       var.equal = FALSE, conf.level = 0.95)
```

The argument `alternative` can also be `"less"` or `"greater"`.

Inference

If the observations are normal the t distribution is *exact*, otherwise it is the *unconditional asymptotic* distribution.

Alternative approach for optaining the sampling distribution: use the *conditional permutation* distribution (conditional on observations) and *approximate* this by simulation.

Motivation: if ξ_1 and ξ_2 are from the same distribution the labels 1 and 2 can be permuted $\Rightarrow$ *re-randomization test*.

Inference

F test: two independent samples

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}, \qquad F \sim F_{n_1-1,n_2-1}$$

In R:

```
var.test(x, y, ratio = 1, alternative = "two.sided", conf.level = 0.95)
```

The argument `alternative` can again be `"less"` or `"greater"`. Both test functions return an object of class `"htest"` which has its own `print` method.

Inference

Distributions in R: `dnorm`, `pnorm`, `qnorm`, `rnorm`, etc.

To check if a sample is normally distributed a **Q-Q plot** can be used, which plots the quantiles of a normal distribution against the ordered sample.

```
qqplot(x)
qqline(x, col = 4)
```

This does mainly

```
plot(qnorm(1:n - 0.5)/n, sort(x))
```

This is similar to the "normal probability paper", which plot $x_{(i)}$ against `qnorm(i/n)`.

Inference

Count data and **binomial random variables** (i.e., $F(\theta) = Bin(n, \pi)$):

Binary responses are observed in mutually independent "Bernoulli trials" with identical outcome probabilities.

Outcome is usually labelled as "success" or "failure". A binomial random variable is the count of the number of successes in n Bernoulli trials with probability of success π. Density:

$$b(x; n, \pi) = \binom{n}{x} \pi^x (1-\pi)^{n-x},$$

with $\mathrm{E}[\xi] = n\pi$ and $\mathrm{VAR}[\xi] = n\pi(1-\pi)$.

Inference

Exact testing: use the number of successes x as test statistic.

Problem: *discrete* distribution $\Rightarrow$ exact level α cannot be obtained.

One possibility: reject H_0 if the probability of observing a value no less or no greater than x is no greater than $\alpha/2$. This corresponds to a p value of $2 \cdot \min(F(x), 1 - F(x-1))$.

In R:

```
binom.test(x, n, p = 0.5, alternative = "two.sided")
```

Inference

Approximate (large-sample) testing: use normal approximation.

$$Z = \frac{\xi - n\pi}{\sqrt{n\pi(1-\pi)}}$$

is approximately standard normal. Rule of thumb: n is large if $\mathrm{VAR}[\xi] = n\pi(1-\pi) \geq 10$). In fact, if variance ≤ 100 better use "continuity corrected" statistic:

$$Z_c = \frac{\xi - n\pi - \mathrm{sign}(\xi - n\pi - 1/2)/2}{\sqrt{n\pi(1-\pi)}}$$

Inference

Association and prediction: if two (quantitative) variables are collected for each data item they can be visualized in a *scatterplot*. In R:

```
plot(x, y, xlim = range(x), ylim = range(y), type = "p",
     main, xlab, ylab, ...)
```

Pairwise scatterplots can be plotted for data frames:

```
pairs(x, labels = dimnames(x)[[2]], panel = points, ...)
```

The *sample Pearson product moment correlation coefficient* measures the *linear* association between two vector x and y:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Inference

This is corresponding to the theoretical $\rho = \sigma_{XY}/(\sigma_X \sigma_Y)$. In R:

```
cor(x, y = x, use = "all.obs")
```

If (X, Y) are jointly normally distributed independence is equivalent to zero correlation. One can show that

$$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right) \approx \mathcal{N}\left(\frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

This can be used for testing hypotheses about ρ. In R:

```
cor.test(x, y, alternative = "two.sided", method = "pearson")
```

tests the null of no correlation.

Inference

The correlation coefficient is quite sensitive to outliers. One robust nonparametric measure of association is the Spearman rank correlation coefficient, the Pearson correlation applied to the ranks of x and y.

A second one is Kendall's τ which is given by

$$\tau = \frac{\kappa}{n(n-1)/2},$$

where κ basically counts the number of concordant pairs and is the sum of $\text{sign}((x_i - x_j)(y_i - y_j))$ over all distinct pairs $i < j$. (Note that the same is obtained when using the ranks.)

Both Spearman's and Kendall's rank correlation coefficients have an asymptotic zero mean normal distribution under the null of independence.

Terminology

- A statistical procedure is *robust* if the procedure performs well
  - when the needed assumptions are not violated "too badly",
  - for a large family of probability distributions.
- A family of probability distributions is *nonparametric* if the distributions of the family cannot conveniently characterized by a few parameters.
- Statistical procedures that are valid for a nonparametric family of distributions are called *nonparametric statistical procedures*.

Terminology

- A statistical procedure is *distribution-free* over a specific family of distributions if the statistical properties of the procedure do not depend on the underlying distribution being sampled.

- A statistical test is distribution-free if under the null it has the same distribution for all members of the family.

- To compare two procedures, one could use the *relative efficiency* defined as the ratio of the sample sizes needed to have the same statistical power.

Terminology

Usual assumption in classical inference about distribution in population: normal (due to CLT).
Implicitly: continuous, symmetric, support $= \mathbb{R}$, all moments exist, short-tailed.
Advandtages: flexible properties, well-known theory.

Usual assumptions in nonparametric inference: continuous (and symmetric).
Advantages: few assumptions needed, simple to understand.
Disadvantages: lower efficiency under normality, few applications for regression, time series and multivariate models.

Terminology

- $\xi \sim F \quad \Rightarrow \quad F \circ \xi \sim U[0,1]$,

- ECDF: $n \cdot F_n(x) \sim Bin(n, F(x))$,

- $||F_n - F||_\infty \quad \xrightarrow{\mathsf{p}} \quad 0$,

- ordered sample: $\mathsf{P}[x_{(1)} < \ldots < x_{(n)}] = 1$,

- ranks: $x_{(R_i)} = x_i$, distribution of $R_i = r(x_i)$ does not depend on F, $\mathsf{E}[R_i] = (n+1)/2$, usually $\mathsf{CORR}[\xi_i, R_i]$ "high".

Terminology

Problem: ties

- omit tied values (only if fraction very small!),

- randomize (does not affect distribution of R_i),

- average ranks (distribution changes!), this is most common and also implemented in R: `rank(x)`,

- for test statistics consider most extreme values.

Tests for the mean

Sign test:

Let $\xi_{0.5}$ be the median of F (continuous), then:

$$T = \sum_{i=1}^{n} \mathbb{1}_{(0,\infty)}(\xi_i - \xi_{0.5}) \quad \sim \quad Bin(n, 0.5)$$

This can be used for a binomial test of hypotheses about the median (or other quantiles):

$$H_0: \quad \xi_{0.5} = \xi_0.$$

Asymptotic Relative Efficiency (ARE) compared to t test: 0.64.

Tests for the mean

If F additionally symmetric: **Wilcoxon signed rank test**.

Use a weighted version of the statistic T:

$$W^+ = \sum_{i=1}^{n} \mathbb{1}_{(0,\infty)}(\xi_i - \xi_{0.5}) \cdot r(|\xi_i - \xi_{0.5}|).$$

Under H_0 the distribution of W^+ is symmetric about $\mathrm{E}[W^+] = n(n+1)/4$.

$$\mathrm{P}[W^+ = w] = \frac{a_n(w)}{2^n} \qquad \left(w \in 0, \ldots, \frac{n(n+1)}{2}\right),$$

where $a_n(w)$ is the number of subsets of $\{1, \ldots, n\}$ with sum w. The following recursion holds: $a_n(w) = a_{n-1}(w-n) + a_{n-1}(w)$.

Tests for the mean

2-sample case: **Wilcoxon rank sum test**.

Assumption: two independent samples of size m and n from F_1 and F_2 with $F_2(x) = F_1(x - \Delta)$.

For a test of $H_0 : \Delta = 0$ use the ranks of ξ_2 in the pooled sample of size $N = n + m$; the test statistic is the sum of these ranks:

$$W = \sum_{j=1}^{n} R_j.$$

This test is equivalent to the Mann-Whitney test:

$$U = \sum_{i,j} \mathbb{1}_{(0,\infty)}(\xi_{2j} - \xi_{1i}) = W - \frac{n(n+1)}{2}.$$

Tests for the mean

The distribution (without ties) can again be computed by a simple recursion. The ARE of the Wilcoxon test relative to the 2-sample t test is 0.955.

In R:

```
wilcox.test(x, y = NULL, alternative = "two.sided", mu = 0,
            paired = FALSE, exact = FALSE, correct = TRUE)
```

This function cannot handle ties, if there are any use the function `wilcox.exact()` in the package `exactRankTests`.

Tests for the mean

k-sample case: **Kruskal Wallis test**:

Let R_{ij} be the rank of the observation j in group i with $\mathrm{E}[R_{i\cdot}] = n_i \cdot (N+1)/2$, then the test statistic is:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{\left(R_{i\cdot} - n_i \frac{N+1}{2}\right)^2}{n_i}.$$

In R:

```
kruskal.test(x, g)
```

Rank tests

The idea of the Wilcoxon rank sum test can be generalized to linear rank statistics

$$S = \sum_{i=1}^{N} c(i) a(R_i^*),$$

where $c(\cdot)$ is a regression function (typically, $c(i) = \mathbb{1}_{m+1,\ldots,N}(i)$ in the 2-sample case) and $a(\cdot)$ is a score function (or influence function) that depends on the ranks R_i^* from the pooled sample.

Special cases of S are the Wilcoxon test, the Median test, the Von Der Waerden test etc. all with suitable scores. Included in this class are also tests for scale.

Rank tests

To compute the distribution of S, there are again two strategies:

The *asymptotic* distribution is normal (under suitable assumptions).

The the *exact permutation* distribution can be also computed. One possibility is the Mehta & Patel network algorithm, another way (which is implemented in `exactRankTests`) is the Streitberg & Röhmel shift algorithm. The latter relies on the fact that the scores are integer-valued, the former is essentially only available in commercial software packages.

Tests for scale

Assumption: $F_1(x - \xi_{0.5}) = F_2((x - \xi_{0.5})/\eta)$, i.e., equal medians, different scales.

Mood test:

$$S_M = \sum_{j=1}^{n} \left(R_j - \frac{N+1}{2}\right)^2$$

If $\eta > 1$ then ξ_2 tends to have more extreme values and hence S_M will be "large".

Tests for scale

Alternatively : **Ansari-Bradley test**.

Scores:

$$a_{AB}(i) \quad = \quad \min\{i, N+1-i\},$$

thus the largest and smallest observation have the same weight etc.

In R:

```
mood.test(x, y, alternative = c("two.sided", "less", "greater"))
ansari.test(x, y, alternative = c("two.sided", "less", "greater"),
            exact = NULL, conf.int = FALSE, conf.level = 0.95)
```

Conditional inference

Recently, there was an increased interest in more general permutation tests than only rank-based tests. The corresponding theory is also referred to as *conditional inference*.

Strasser & Weber (1999) established a unified framework for the asymptotic permutation distribution of linear statistics

$$S \quad = \quad \text{vec}\left(\sum_{i=1}^{N} c(X_i)a(Y_i)\right),$$

where X_i is the explanatory variable (e.g., factor with two categories) and Y_i is the dependent variable (corresponding to ξ_i in the previous notation). Furthermore, $c(\cdot)$ is again a regression function and $a(\cdot)$ again a score function or influence function.

Conditional inference

The influence function $a(\cdot)$ may depend on the full vector $(Y_1, \ldots, Y_N)^\top$ but only in a permutation symmetric way, i.e., it can depend on statistics that do not exploit the order of the observations.

Examples for the influence function:

- observations: $a(Y_i) = Y_i$,
- ranks: $a(Y_i) = r(Y_i)$,
- other transformations based on moments or counts.

Conditional inference

Examples for the regression function:

- for qualitative X: dummy coding $c(X_i) = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$,
- for quantitative X: as above, observations or ranks, etc.

The conditional distribution of the statistic S conditional on the observations $(X_i, Y_i)^\top$ $(i = 1, \ldots, n)$ is derived under the null hypothesis of independence

$$F(y) \quad = \quad F(y \mid x)$$

given all permutations from $\sigma(X, Y)$.

Conditional inference

Given the first two moments of the scores

$$\begin{aligned} \mathsf{E}_\sigma[a] &= \frac{1}{N}\sum_{i=1}^{N} a(Y_i) \\ \mathsf{VAR}_\sigma[a] &= \frac{1}{N}\sum_{i=1}^{N} (a(Y_i) - \mathsf{E}_\sigma[a])(a(Y_i) - \mathsf{E}_\sigma[a])^\top \end{aligned}$$

the first two moments of the conditional distribution of S can be computed:

$$\mu = \mathsf{E}_\sigma[S] = \mathrm{vec}\left(\left(\sum_{i=1}^{N} c(X_i)\right)\mathsf{E}_\sigma[a]^\top\right)$$

and similarly for $\Sigma = \mathsf{VAR}_\sigma[S]$.

Conditional inference

Based on S (which might be a vector) different types of scalar test statistics can be derived, e.g., a maximum type statistic or a quadratic type of test statistic:

$$\begin{aligned} t_{\mathsf{max}} &= \max_k \left|\frac{S_k - \mu_k}{\sqrt{(\Sigma_{kk})}}\right| \\ t_{\mathsf{quad}} &= (S-\mu)\Sigma^{-1}(S-\mu) \end{aligned}$$

If S is scalar, the $t_{\mathsf{quad}} = t_{\mathsf{max}}^2$ and both statistics are equivalent.

Conditional inference

To compute the conditional permutation distribution, there are again several possibilities:

- The *asymptotic* permutation distribution is again normal.
- The *exact* permutation distribution can be computed by various algorithms including the Streitberg & Röhmel shift algorithm (for integer-valued scores) and the van de Wiel split algorithm (which is memory-intensive).
- The *exact* permutation distribution can also be approximated by resampling (or re-randomization) techniques.

This general framework is implemented in the package `coin`.

Conditional inference

Special cases of this framework include 2-sample tests (`y` numeric and `x` a factor):

- Choosing $c(\cdot) =$ 2-sample regression function and $a(Y_i) = Y_i$, yields a statistic similar to the t test statistic (with slightly different standard deviation).
  In R: `independence_test(y ~ x)`
- Choosing $c(\cdot) =$ 2-sample regression function and $a(Y_i) = r(Y_i)$, yields a statistic similar to the Wilcoxon test statistic.
  In R: `wilcox_test(y ~ x)`

Conditional inference

Other special cases are correlation tests (`y` and `x` both numeric):

- Choosing $c(X_i) = X_i$ and $a(Y_i) = Y_i$ yields a correlation test similar to the Pearson correlation test in `cor.test(x, y)`.
  In R: `independence_test(y ~ x)`

- Choosing ranks $c(X_i) = r(X_i)$ and $a(Y_i) = r(Y_i)$ yields a correlation test similar to the Spearman correlation test in `cor.test(x, y, method = "spearman")`.
  In R: `spearman_test(y ~ x)`

Inference procedures

Classical parametric inference: early 1900s

Idea:

1. Impose a distribution (typically normal) on all variables.
2. Compute a test statistic that highlights deviation of interest (e.g., difference in means or ratio of variances).
3. Compute exact unconditional distribution based on the assumption.

Problem: What if distribution assumptions are violated?

Inference procedures

Solution:

1. Rely on the central limit theorem. The distribution is still valid asymptotically, it is the unconditional asymptotic distribution.
2. Use the conditional permutation distribution.

Problem with 2.: exact conditional distribution could only be computed for small samples, only limited asymptotic theory available → approach 2. not used for a long time.

Inference procedures

Rank-based inference: mid 1900s

Idea:

1. Transform the observations by taking ranks.
2. Compute a test statistic that highlights deviation of interest (e.g., by choosing a suitable score function).
3. Compute exact conditional permutation distribution by simple recursion or the asymptotic permutation distribution (normal).

Problem: Not (directly) applicable to original observations (instead of ranks).

Justification: Ranks introduce a certain robustness in the procedures. *Be careful:* The procedures are only robust in certain directions and they do also make assumptions!

Inference procedures

Conditional inference: increased interest again since late 1900s

Idea:

1. Take an arbitrary transformation of the data (including identity) that is appropriate for the problem (i.e., choice of regression and score function).
2. Compute the independence test statistic.
3. Compute exact conditional permutation distribution (via different algorithms), asymptotic conditional permutation distribution (normal), or approximate the exact conditional permutation distribution by resampling.

Problem: Although conditional inference is applicable in much more general situations, it is not (yet) well-known in many statistical communities.

Goodness-of-fit tests

Tests for the hypothesis

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0$$

Kolmogorov-Smirnov test

Test statistic: $D_n = ||F_0 - F_n||_\infty$.

The test can be used to construct confidence bounds for F. Tests are also available for one-sided alternatives. If F_0 not fully specified parameters can be estimated by Maximum Likelihood (ML), but the test becomes conservative.

Goodness-of-fit tests

There is also a Kolmogorov-Smirnov test in the 2-sample case:

$$D_{m,n} \quad = \quad ||F_{1m} - F_{2n}||_\infty.$$

In R:

```
ks.test(x, y, ..., alternative = "two.sided")
```

where `y` can be a string specifying a distribution.

Goodness-of-fit tests

Alternative: χ^2 **test**,
especially for F discrete, otherwise use class probabilities.

Use χ^2 statistic: (observed - expected)^2/expected.

Let N_j be the empirical class frequencies with $n = \sum_{j=1}^k N_j$ and p_j the expected class probabilities:

$$X^2 \quad = \quad \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j},$$

which is asymptotically χ^2_{k-1} distributed.

Goodness-of-fit tests

If F_0 not fully specified and depends on an r-dimensional parameter vector which has to be estimated:

1. (grouped) Maximum Likelihood, usual ML is anti-conservative and X^2 is not χ^2 distributed. Likelihood:
$$L = \prod_{j=1}^{k} p_j(\theta)^{N_j}.$$
2. Minimum χ^2, choose θ by minimizing $X^2(\theta)$.

In both cases X^2 is asymptotically χ^2_{k-r-1} distributed.

Types of test statistics

This chapter includes a large collection statistics in a 2-sample setting, so one could ask: Why so many? What are the differences?

Interestingly, the null hypothesis is typically *equivalent*. It is essentially that the distribution F_1 in the first sample is the same as the distribution F_2 in the second sample.

Types of test statistics

The main difference between the tests is that the test statistics are derived for different potential differences between F_1 and F_2:

- test statistics for the mean highlight deviations in the location of the distributions,
- test statistics for the scale highlight deviations in the scale of the distributions,
- goodness-of-fit test statistics try to highlight in any direction (omnibus tests).

The consequence is that different types of test statistics have different power for detecting certain alternatives.

Finally, some tests differ in the way they compute the sampling distribution: exact vs. asymptotic, and conditional vs. unconditional.

Robust mean

- The α-trimmed mean
$$\frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)} \qquad (k = \lfloor \alpha n \rfloor).$$
- The α-winsorized mean
$$\frac{1}{n}\left((k+1)(x_{(k+1)} + x_{(n-k)}) + \sum_{i=k+2}^{n-k-1} x_{(i)}\right), \qquad (k = \lfloor \alpha n \rfloor).$$
- The weighted mean
$$\frac{\sum_i w_i x_{(i)}}{\sum_i w_i}, \qquad w_i > 0.$$
In R, `weighted.mean(x, w, na.rm = FALSE)`.

The general linear model

Consider the linear regression model:

$$y_i \quad = \quad x_i^\top \beta + \varepsilon_i, \qquad (i = 1, \ldots, n),$$

where for observation i:

- y_i — dependent variable,
- x_i — vector of k regressors,
- β — vector of k unknown regression coefficients,
- ε_i — a disturbance term.

The general linear model

In matrix notation:

$$y \quad = \quad X\beta + \varepsilon.$$

Assumptions:

(A.1) X is nonstochastic with $\text{rank}(X) = k$,

(A.2) $\text{E}[\varepsilon_i] = 0$, thus $\mu_i = \text{E}[y_i] = x_i^\top \beta$,

(A.3) The disturbance terms are iid. with variance σ^2,

(A.4) The disturbances are normal, thus $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ independently.

The general linear model

- OLS estimate: $\hat{\beta} \; = \; (X^\top X)^{-1} X^\top y$.
- Fitted values: $\hat{y} \; = \; X\hat{\beta}$.
- Residuals: $e \; = \; y - \hat{y}$.
- Residual Sum of Squares: $RSS \; = \; e^\top e = (y - \hat{y})^\top (y - \hat{y})$.
- Variance estimate: $\hat{\sigma} \; = \; \frac{RSS}{n-k}$.
- $R^2 \; = \; 1 - RSS / \sum_{i=1}^{n} (y_i - \bar{y})^2$.

The general linear model

Tests of the regression coefficients:

- $\hat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 (X^\top X)^{-1}\right)$.
- Under $H_0 : \beta_j = 0$

$$\frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X^\top X)^{-1}_{jj}}} \sim t_{n-k}.$$

- To test q restrictions $R\beta = r$ fit a full model with RSS_1 and a restricted model with RSS_0 and compare these in an F test:

$$\frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-k)} \sim F_{q,n-k}.$$

The general linear model

Modelchecking:

- plots: residuals vs. fitted, residuals vs. regressors
- F test: actual vs. trivial model

Model selection:

- best subsets regression
- stepwise procedures, e.g., based on Akaike Information Criterion (AIC).

The general linear model

Analysis of variance (ANOVA): 1-way design, i.e., observations fall into k groups

$$y_{ij} \quad = \quad \mu + \alpha_i + \varepsilon_{ij}.$$

- overparamterized: $k+1$ paramters but only k groups,
- fit using some constraint, e.g.,
  $\mu = 0$, $\alpha_1 = 0$, $\sum_i \alpha_i = 0$
  the latter two correspond to *contrasts*.

This can be interpreted as a linear regression with a qualitative regressor.

The general linear model

The model matrix can be written as

$$X \quad = \quad [1\ X_{\mathrm{a}}]$$

where X_{a} is a binary incidence or dummy matrix. To remove over-parametrization consider

$$X^* \quad = \quad [1\ X_{\mathrm{a}} C_{\mathrm{a}}]$$

where C_{a} is a $k \times (k-1)$ *contrast matrix*. This defines a model with parameters α^* which is equivalent to estimating the original parameters α subject to the *identification constraint*

$$c_{\mathrm{a}}^\top \alpha = 0$$

where c_{a} is such that $c_{\mathrm{a}}^\top C_{\mathrm{a}} = 0$.

The general linear model

2-way ANOVA: two explanatory factors (qualitative variables) with interaction:

$$y_{ijk} \quad = \quad \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}.$$

Analysis of covariance (ANCOVA): qualitative and quantitative regressors, i.e., fit separate regressions for several groups—potentially with restrictions about the regression coefficients.

The general linear model

The R function `lm()` is used to fit *all* such linear models.

```
lm(formula, data, subset, weights, na.action, ...)
```

where

- `formula` — a symbolic description of the model to be fit,
- `data` — an optional data frame containing the variables,
- `subset` — an optional vector specifying a subset of observations to be used in the fitting process.

The general linear model

`y ~ x` `y ~ 1 + x`	Simple linear regression model of y on x. The intercept is implicit in the first and explicit in the second formula.
`y ~ x - 1` `y ~ x + 0`	Simple linear regression of y on x through the origin (without an intercept term).
`log(y) ~ x1 + x2`	Multiple regression of the transformed variable $\log(y)$ on x_1 and x_2 (with an implicit intercept term).
`y ~ 1 + x + I(x^2)` `y ~ poly(x, 2)`	Polynomial regression of y on x of degree 2.
`y ~ X`	Multiple linear regression of y on the variables (i.e., the columns) of X.

The general linear model

`y ~ a`	Single classification (one-way) analysis of variance of y with classes determined by a.
`y ~ a + x`	Single classification (one-way) analysis of covariance of y with classes determined by a and covariate x.
`y ~ a + b`	Two factor (two-way) analysis of variance of y, without interaction terms.

`y ~ a * b`
`y ~ a + b + a : b`
`y ~ b %in% a`
`y ~ a / b`
	Two factor (two-way) analyses of variance of y on a and b. The first two specify the same crossed classification and the second the same nested one. In abstract terms all four specify the same model subspace.

The general linear model

`y ~ (a + b + c)^2` `y ~ a*b*c - a:b:c`	Three factor experiment with a model containing main effects and two factor interactions only.
`y ~ a * x` `y ~ a / x`	Separate linear regression models of y on x within the levels of a, with different codings. The last form produces explicit estimates of as many different intercepts and slopes as there are levels in a.

The general linear model

The model operators are as follows.

Y ~ M	Y is modelled as M
M_1 + M_2	Include M_1 and M_2
M_1 - M_2	Include M_1 leaving out terms of M_2
M_1 : M_2	The tensor product of M_1 and M_2
M_1 %in% M_2	Similar to $M_1 : M_2$, but with a different coding
M_1 * M_2	$M_1 + M_2 + M_1 : M_2$
M_1 / M_2	$M_1 + M_2$ %in% M_1
M ^ n	All terms in M together with "interactions" up to order n.
I(M)	Insulate M. Inside M all operators have their normal arithmetic meaning.

The general linear model

coefficients(lmobj)	Extract the regression coefficient (matrix). Short: coef(lmobj).
deviance(lmobj)	Residual sum of squares.
fitted.values(lmobj)	Extract the fitted values. Short: fitted(lmobj).
formula(lmobj)	Extract the model formula.
plot(lmobj)	Produce useful diagnostics plots.
predict(lmobj)	For extracting the fitted values or making new predictions.
print(lmobj)	Print call and coefficients.
residuals(lmobj)	Extract the (matrix of) residuals. Short: resid(lmobj).
summary(lmobj)	Print a comprehensive summary of the results of the regression analysis.

The GLM

Motivation: The general linear model (glm) is appropriate in many situations, but not always, e.g.,

- does not fit well,
- needs too many parameters,
- non-linear relationship between response and regressors,
- a change in mean is accompanied by a change in variance,
- response is not normal: binary, counts, survival time etc.

This leads to the Generalized Linear Model (GLM).

The GLM

As in the general linear model the influence of the explanatory variables is linear:

$$\eta_i \quad = \quad x_i^\top \beta,$$

where η_i is called *linear predictor*. The relationship between the linear predictor and the modelled mean is generalized from $\mu_i = \eta_i$ to

$$g(\mu_i) \quad = \quad \eta_i,$$

where g is called the *link function* and $h = g^{-1}$ the inverse link function. The assumption that y_i is normal is generalized to the assumption that y_i has a specified exponential family distribution:

$$f_Y(y|\theta,\phi) \quad = \quad \exp\left(\frac{y\theta - \gamma(\theta)}{\phi} + \tau(y,\phi)\right),$$

where ϕ is a (possibly known) scale parameter and θ controls the distribution of y.

The GLM

The assumptions imply that

$$\mathrm{E}[y] = \mu = \gamma'(\theta), \qquad \mathrm{VAR}[y] = \phi \cdot \gamma''(\theta).$$

That is, up to a scale parameter, the distribution of μ is determined by its mean, and the variance of y is proportional to $V(\mu) = \gamma''(\theta(\mu))$ which is called the *variance function*.

Three important classes of models can be described in this framework:

- Gaussian,
- Poisson,
- Binomial.

The GLM

Gaussian: The classical Gaussian case has

$$\begin{aligned} &\log(f_Y(y, \theta, \phi)) \\ &= -(y-\mu)^2/(2\sigma^2) - \log(2\pi\sigma^2)/2 \\ &= (y\mu - \mu^2/2)/\sigma^2 - (y^2/\sigma^2 + \log(2\pi\sigma^2))/2, \end{aligned}$$

so $\theta = \mu$, $\gamma(\theta) = \theta^2/2$, and $\phi = \sigma^2$.

Poisson: For the Poisson distribution with mean μ we have

$$\log f(y) = y \log \mu - \mu - \log(y!),$$

so $\theta = \log(\mu)$, $\phi = 1$, and $\gamma(\theta) = \mu = e^\theta$.

The GLM

Binomial: In the binomial case the response y can be the number of successes in n trials with probability of success π or a single Bernoulli variable coding success or failure. The latter view is also called *logistic regression*. The density is

$$\begin{aligned} &\log f(y) \\ &= \log\left(\pi^y (1-\pi)^{1-y}\right) \\ &= y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi), \end{aligned}$$

so we take $\phi = 1$, θ as the logit transform of π, and $\gamma(\theta) = -\log(1-\pi) = \log(1+e^\theta)$.

The GLM

The R function `glm()` is used for fitting GLMs:

```
glm(formula, family = gaussian, data, weights, subset, na.action, ...)
```

Family	Canonical Link	Name	Variance	Name
`gaussian`	μ	`identity`	1	`constant`
`binomial`	$\log(\mu/(1-\mu))$	`logit`	$\mu(1-\mu)$	`mu(1-mu)`
`poisson`	$\log(\mu)$	`log`	μ	`mu`
`Gamma`	$-1/\mu$	`inverse`	μ^2	`mu^2`
`inverse.gaussian`	$-2/\mu^2$	`1/mu^2`	μ^3	`mu^3`
`quasi`	$g(\mu)$		$V(\mu)$	

The GLM

Each response distribution allows for a variety of link functions. The combination of a response distribution and a link function is called the *family* of the model. The canonical link function is $g = (\gamma')^{-1}$, i.e., the one for which $\theta = \eta$. The gaussian and inverse gaussian families only admit the canonical link. For the others, the situation is as follows:

Family	Possible Links
binomial	logit, probit, cloglog (complementary log-log)
poisson	identity, log, sqrt
Gamma	identity, inverse, log
quasi	logit, probit, cloglog, identity, inverse, log, 1/mu^2, sqrt

The GLM

Logistic regression:
Logistic modelling is very popular for binary data, it assumes a *log-linear* relation between the regressors and the *odds*:

$$\frac{\pi}{1-\pi} = \frac{\mathsf{P}[Y=1|x_i]}{\mathsf{P}[Y=0|x_i]} = \exp\left(x_i^\top \beta\right).$$

A useful way of describing the importance of a factor (e.g., treatment) is the *odds ratio*.

$$\frac{\mathsf{P}[Y=1|x_1=1]}{\mathsf{P}[Y=0|x_1=1]} \Big/ \frac{\mathsf{P}[Y=1|x_1=0]}{\mathsf{P}[Y=0|x_1=0]} = \exp(\beta_1),$$

so if for example $\exp(\beta_1) = 1.3$ the odds for success are 30% higher on treatment.

The GLM

Probit link:

$$g(\mu) \quad = \quad \Phi^{-1}(\mu).$$

Motivation:
Suppose π is the probability to kill an insect with a certain poison dose and each insect has a random normal tolerance T, then:

$$\pi = \mathsf{P}[T \leq \text{dose}] = \Phi\left(\frac{\text{dose} - \mu_T}{\sigma_T}\right) \quad \Rightarrow \quad \Phi^{-1}(\pi) = \alpha + \beta \cdot \text{dose}.$$

Complementary log-log link:

$$g(\mu) \quad = \quad \log\left(-\log(1-\mu)\right),$$

which is *not* symmetric, i.e., gives different results, if modelling successes or failures.

The GLM

The parameters in GLMs are usually estimated by Maximum Likelihood, using the IWLS (Iterative Weighted Least Squares) algorithm.

Discrepancy of a fit:
Having estimated a model, usually the discrepancy of the fit is of interest. Commonly this is measured by the logarithm of a ratio of likelihoods, called the *deviance*.

Given n observations, the simplest model just has one parameter (a common μ for all y) and the full model has n parameters yielding a perfet match of the y and the μ.

Log likelihood:

$$\log L(\mu|y) \quad = \quad \sum_{i=1}^{n} \log f(y_i|\theta_i).$$

The (scaled) deviance is defined as

$$D^*(y,\mu) \quad = \quad 2\log L(y|y) - 2\log L(\mu|y) = \phi D(y,\mu),$$

which is just the residual sum of squares for the normal distribution. In the other models:

Family	Deviance
Gaussian	$\sum(y-\widehat{\mu})^2$
Poisson	$2\sum\{y\log(y/\widehat{\mu}) - (y-\widehat{\mu})\}$
binomial	$2\sum\{y\log(y/\widehat{\mu}) + (1-y)\log((1-y)/(1-\widehat{\mu}))\}$
Gamma	$2\sum\{-\log(y/\widehat{\mu}) + (y-\widehat{\mu})/\widehat{\mu}\}$
inverse Gaussian	$\sum(y-\widehat{\mu})^2/(\widehat{\mu}^2 y)$

Another possibility to measure the discrepancy of a fit is the generalized Pearson X^2 statistic

$$X^2 \quad = \quad \sum \frac{(y-\widehat{\mu})^2}{V(\widehat{\mu})},$$

which is again the residual sum of squares for the normal distribution and the original Pearson X^2 statistic for the Poisson and binomial distribution.

Both D and X^2 have asymptotic χ^2 distributions, but the approximation may be poor even for large n. The deviance has the advantage of being additive for nested sets of models (with ML estimates).

Analysis of deviance: To compare a full model with estimate $\widehat{\mu}_1$ with a restricted model with q restrictions and an estimate $\widehat{\mu}_0$ the excess deviance can be used

$$D(y,\widehat{\mu}_0) - D(y,\widehat{\mu}_1),$$

which has an asymptotic χ^2_q distribution.

Residuals: There are various definitions of residuals in GLMs, the most common ones are:

- Pearson residuals: $r_P = (y-\widehat{\mu})/\sqrt{V(\widehat{\mu})}$,
- deviance residuals: $r_D = \text{sign}(y-\widehat{\mu})\sqrt{d_i}$,

where d_i is the contribution of each unit to the deviance such that $\sum_i d_i = D$. In both cases the sum of squared residuals yields the respective statistic for the discrepancy of a fit.

In R: Mostly the same commands can be used for objects of class `"lm"` and `"glm"`. Thus the same extractor functions like `residuals()`, `fitted()` or `coef()` can be used as well as `anova()`. Possibly an argument specifying the type of residual/test/etc. has to be supplied.

Multiple testing

In the general linear model $Y = X\beta + \varepsilon$ mulitple comparisons of parameters of the form $c_i^\top \beta$ or of interest which can be tested using

$$T_i = \frac{c_i^\top \hat{\beta} - c_i^\top \beta}{\hat{\sigma}\sqrt{c_i^\top (X^\top X)^{-1} c_i}}.$$

For p such contrasts $\{c_1^\top \beta, \ldots, c_p^\top \beta\}$ the joint distribution for $T = \{T_1, \ldots, T_p\}$ is multivariate t.

In R this procedure is available in the package `multcomp`.

Contingency tables

Data on categorical variables is usually represented in contingency tables—the cross-classification of the variables.

First consider the following special case: comparing two binary variables.

	Success	Failure
Sample 1	n_{11}	n_{12}
Sample 2	n_{21}	n_{22}

We want to test the hypothesis that the probabilities of success are the same in both samples.

Contingency tables

Fisher's exact test: Conditional on the row and column totals n_{11} has a hypergeometric distribution with parameters $n_{\cdot 1}$, $n_{\cdot 2}$, and $n_{1\cdot}$. In R:

```
fisher.test(x, y = NULL, alternative = "two.sided")
```

Large sample test: If we have two independent Bernoulli trials,

$$Z = \frac{p_1 - p_2 - (\pi_1 - \pi_2)}{\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}}$$

is approximately standard normal, and can be used for approximate confidence intervals and significance tests for the difference of proportions.

Contingency tables

χ^2 **test**: Under H_0 the success probabilities are equal $\pi_1 = \pi_2$ and the expected counts are:

	Success	Failure
Sample 1	$n_{1\cdot}\pi$	$n_{1\cdot}(1-\pi)$
Sample 2	$n_{2\cdot}\pi$	$n_{2\cdot}(1-\pi)$

Estimate π by the pooled sample proportion $p = (n_{11} + n_{21})/n_{\cdot\cdot}$ and calculate the X^2 statistic that simplifies to

$$X^2 = \frac{n_{\cdot\cdot}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}$$

which is asymptotically χ^2 with one degree of freedom. X^2 equals Z^2 if the pooled estimate $p(1-p)(1/n_1 + 1/n_2)$ is used for the variance.

Contingency tables

In R

```
prop.test(x, n, p = NULL, alternative = "two.sided", conf.level = 0.95,
          correct = TRUE)
```

can be used for both testing whether proportions (probabilities of success) in several groups are the same, or that they equal certain given values.

The same statistic (and p value) is computed by

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)))
```

but `prop.test()` is more appropriate for this specific situation.

Contingency tables

Association measures for two binary variables: Suppose we have given the cross-tabulation of true population proportions for *disease* and *exposure*:

	Disease	
Exposure	+ (Yes)	− (No)
+ (Yes)	π_{11}	π_{12}
− (No)	π_{21}	π_{22}

The *relative risk* is defined as

$$\rho = \frac{\text{P[disease +|exposure +]}}{\text{P[disease +|exposure −]}} = \frac{\pi_{11}/\pi_{1\cdot}}{\pi_{21}/\pi_{2\cdot}}$$

Note that any information on the amounts of disease and exposure is missing.

Contingency tables

Given one has the exposure, the *odds* of getting the disease are

$$\frac{\text{P[disease +|exposure +]}}{\text{P[disease −|exposure +]}} = \frac{\pi_{11}}{\pi_{21}}$$

The *odds ratio* is defined as

$$\frac{\text{odds(disease|exposure +)}}{\text{odds(disease|exposure −)}} = \frac{\pi_{11}/\pi_{21}}{\pi_{12}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \omega$$

Note that

$$\frac{\rho}{\omega} = \frac{\pi_{2\cdot}}{\pi_{22}}\frac{\pi_{12}}{\pi_{1\cdot}};$$

hence, if the disease is rare, $\pi_{2\cdot} \approx \pi_{22}$ and $\pi_{1\cdot} \approx \pi_{12}$ and thus $\rho \approx \omega$.

Contingency tables

Data for measuring association between two binary variables typically come in 2×2 contingency tables

	Disease	
Exposure	+ (Yes)	− (No)
+ (Yes)	n_{11}	n_{12}
− (No)	n_{21}	n_{22}

We have to distinguish three sampling patterns.

Contingency tables

Pattern 1: Cross-sectional study. For a sample of size $n_{\cdot\cdot}$, both traits (disease and exposure) are measured on each subject. The expected numbers in the cells are given by

$$\begin{array}{cc} n_{\cdot\cdot}\pi_{11} & n_{\cdot\cdot}\pi_{12} \\ n_{\cdot\cdot}\pi_{21} & n_{\cdot\cdot}\pi_{22} \end{array}$$

Pattern 2: Prospective Study of Exposure. Fixed numbers ($n_{1\cdot}$ and $n_{2\cdot}$) of individuals with and without the exposure are followed. The endpoints are then noted. Expected numbers are obtained as

$$\begin{array}{cc|c} n_{\cdot\cdot}\pi_{11}/\pi_{1\cdot} & n_{\cdot\cdot}\pi_{12}/\pi_{1\cdot} & n_{1\cdot} \\ n_{\cdot\cdot}\pi_{21}/\pi_{2\cdot} & n_{\cdot\cdot}\pi_{22}/\pi_{2\cdot} & n_{2\cdot} \end{array}$$

Note that in this case, one cannot estimate the proportion of exposure.

Contingency tables

Pattern 3: Retrospective Study of Disease. Usually, cases and an appropriate control group are identified. In this case, the size of the disease and control groups, $n_{\cdot 1}$ and $n_{\cdot 2}$, are specified. Expected cell counts are

$$\begin{array}{cc} n_{\cdot\cdot}\pi_{11}/\pi_{\cdot 1} & n_{\cdot\cdot}\pi_{12}/\pi_{\cdot 2} \\ n_{\cdot\cdot}\pi_{21}/\pi_{\cdot 1} & n_{\cdot\cdot}\pi_{22}/\pi_{\cdot 2} \\ \hline n_{\cdot 1} & n_{\cdot 2} \end{array}$$

The probability of getting the disease and hence also the relative risk cannot be estimated (it can only be approximated by the odds ratio if the disease is rare).

Contingency tables

For all three patterns, we find that

$$\frac{\mathsf{E}[n_{11}]\mathsf{E}[n_{22}]}{\mathsf{E}[n_{12}]\mathsf{E}[n_{21}]} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \omega,$$

and one thus estimates the odds ratio by

$$\widehat{\omega} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The hypothesis of no association is equivalent to

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \quad i,j \in \{1,2\}$$

and hence implies (but is not equivalent to!) $\omega = 1$.

Contingency tables

H_0 can be tested with a Fisher or χ^2 test or use that $\log(\widehat{\omega})$ is approximately normal with mean $\log(\omega)$ and variance $\sum_{ij} 1/n_{ij}$.

Often H_0 is to be investigated with stratified data, e.g. multi-center study. This data cannot simply be merged as it can lead to *Simpson's paradox*.

The standard approach to estimating an overall odds ratio is as follows. Compute continuity corrected odds ratios in every stratum

$$\widehat{\omega}_i = \frac{(n_{11}+0.5)(n_{22}+0.5)}{(n_{21}+0.5)(n_{12}+0.5)}, \quad s_i = \sqrt{\sum_{ij} \frac{1}{n_{ij}+0.5}}.$$

Let $a_i = \log(\widehat{\omega}_i)$. Under H_0 $X^2 = \sum_{i=1}^{k}(a_i/s_i)^2$ is approximately χ^2_k.

Contingency tables

One now partitions

$$X^2 = \sum_{i=1}^{k} \frac{(a_i - \bar{a})^2}{s_i^2} + \bar{a}^2 \sum_{i=1}^{k} \frac{1}{s_i^2} = X_H^2 + X_A^2,$$

where

$$\bar{a} = \sum_{i=1}^{k} (a_i/s_i^2) \Big/ \sum_{i=1}^{k} (1/s_i^2)$$

is an appropriate weighted average of the a_i.

X_H^2 is for testing homogeneity of association across strata (asymptotically χ^2_{k-1}). X_A^2 tests whether there is association on the average (approximately χ^2_1).

Contingency tables

Alternatively, use **Mantel-Haenszel test**.

Estimate the odds ratio by

$$\hat{\omega} = \frac{\sum_{i=1}^{k} n_{11}(i) n_{22}(i)/n_{\cdot\cdot}(i)}{\sum_{i=1}^{k} n_{12}(i) n_{21}(i)/n_{\cdot\cdot}(i)}$$

where the $n_{kl}(i)$ are the n_{kl} for the i-table. They also derived a test for conditional independence of two binary traits in several strata (Mantel-Haenszel test). In R,

```
mantelhaen.test(x, y = NULL, z = NULL, correct = TRUE)
```

Contingency tables

Matched or paired data: If the proportions to be compared come from the same sample the preceeding methods are not applicable. For example every patient receives two treatments A & B, thus we observe data of the form: (response to A, response to B), e.g., $(0,1), (1,1), (0,0), (1,1), \ldots$

It is tempting to do a Fisher test in the table

	response	
treatment	yes	no
A	11	37
B	20	28

but that is *invalid*.

Contingency tables

Better summarize as

	B	
A	yes	no
yes	8	3
no	12	25

A suitable test for treatment differences is then a test for symmetry, i.e.

$$H_0: \quad \pi_{ij} = \pi_{ji}$$

in a 2-way contingency table. This is called **McNemar's test**.

Contingency tables

Given data of the form

Case has risk factor?	Control has risk factor? yes	no
yes	a	b
no	c	d

If there is no association between the disease and the risk factor, b has a binomial distribution $Bin(0.5, b+c)$. Hence,

$$X^2 = \frac{(b-c)^2}{b+c}$$

is approximately χ^2_1. In R

```
mcnemar.test(x, y = NULL, correct = TRUE)
```

The odds ratio is estimated by $\hat{\omega} = b/c$.

Contingency tables

Now consider contingency tables more generally. The cross-classification of two categorical variables A and B with r and c levels respectively, which can be build in R by `table(a,b)`, is called two-way table.

		B			
		1	2	$\cdots$	c
A	1	y_{11}	y_{12}	$\cdots$	y_{1c}
	2	y_{21}	y_{22}	$\cdots$	y_{2c}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	r	y_{r1}	y_{r2}	$\cdots$	y_{rc}

As in regression modelling some variables are dependant others explanatory. Some variables are *controlled* others may be *free*, i.e. responses. The margins of controlled variables are *fixed*.

Contingency tables

There are two sampling possibilities for y_{ij}.

1. A and B are responses: 1 sample of size $n = y_{..}$ with a bivariate response.

2. A is controlled, B is response: r samples of size $n_1, \ldots, n_r$ ($n_i = y_{i.}$) and a univariate response in each sample.

In case 1 the question is: "Are the factors *independent*?" Given

$$\{y_{ij}\} \sim MN(n, \{\pi_{ij}\})$$

independence becomes

$$\pi_{ij} = \pi_{i.}\pi_{.j}$$

Contingency tables

In case 2 the question is: "Is the distribution of B *homogeneous* over A?" Now

$$(y_{i1}, \ldots, y_{ic}) \sim MN(n_i, (\pi_{i1}, \ldots, \pi_{ic}))$$

independent for $i = 1, \ldots, r$ and the hypothesis is

$$\pi_{ij} = \bar{\pi}_{.j}$$

In *both* cases the hypothesis can be tested using

$$X^2 = \sum_{ij} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

with $\hat{\mu}_{ij} = (y_{i\cdot} \cdot y_{.j})/y_{..}$ and X^2 is $\chi^2_{(r-1)(c-1)}$.

Contingency tables

In case 1:

$$\hat{\mu}_{ij} \;=\; n\,\hat{\pi}_{ij} \;=\; n\,\hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} \;=\; n\,\frac{y_{i\cdot}}{n}\,\frac{y_{\cdot j}}{n}.$$

In case 2:

$$\hat{\mu}_{ij} \;=\; n_i\,\hat{\pi}_{ij} \;=\; n_i\,\hat{\pi}_{\cdot j} \;=\; n_i\,\frac{y_{\cdot j}}{n} \;=\; \frac{y_{i\cdot}y_{\cdot j}}{n}.$$

Contingency tables

Poisson log-linear model: It can be shown that if

$$X_i \;\sim\; Poi(\mu_i) \quad\Rightarrow\quad \mathrm{P}[\{X_i\}|\sum_j X_j = n] \;\sim\; MN(n,\{\pi_i\}),$$

where $\pi_i = \mu_i/\sum_j \mu_j$.

Based on this the contingency table data can be regarded as $y_{ij} \sim Poi(\mu_{ij})$—in case 1 given $y_{\cdot\cdot} = n$ and in case 2 conditional on $y_{i\cdot} = n_i$.

Maximum Likelihood based on a Poisson model gives the same estimates etc. as the multinomial model.

Contingency tables

The hypotheses of independence and homogeneity that are multiplicative in terms of μ_{ij} become additive in log-linear models. In case 1:

$$\begin{aligned}\log\mu_{ij} &= \log(n\pi_{i\cdot}\pi_{\cdot j})\\ &= \log n \;+\; \log\pi_{i\cdot} \;+\; \log\pi_{\cdot j}\\ &= \mu + \alpha_i + \beta_j,\end{aligned}$$

thus, yielding a 2-way ANOVA without interaction. The residual degrees of freedom are

$$rc \;-\; (1 + (r-1) + (c-1)) \quad=\quad (r-1)\,(c-1).$$

Similarly in case 2.

Contingency tables

In R these models can be fitted like

```
glm(Freq ~ a + b, data = as.data.frame(tab), family = poisson)
```

or

```
loglin(tab, list("a", "b"))
```

or

```
loglm(Freq ~ a + b, data = as.data.frame(tab))
```

If either r or c is 2 a binomial model might also be appropriate.

Contingency tables

3-way table: use the same ideas as in the two way case. If there are factors A, B, C, some margins can be fixed and others free.

If there are no fixed margins a natural model would be again:

$$\pi_{ijk} \quad = \quad \pi_{i..}\pi_{.j.}\pi_{..k}$$

i.e. a Poisson model `a + b + c`.

Other models could be: `a + b * c`, i.e., homogeneity of A over (B, C) or independence of A and (B, C) respectivley etc.

Important: Always include **all** fixed margins (and their interactions) in every model as there is no point modelling **non-random** aspects.

Contingency tables

It is also convenient to classify the factors as *response* or *stimulus* variables.

For stimulus variables the margins are fixed, i.e., all interactions are included. Interactions between response and stimulus factors indicate structure.

Suppose we have two response variables `a`, `b` and two stimulus variables `c`, `d`, where we find that `a` depends on `c` and `b` on `d`.

Minimal model	`~ a + b + c * d`
Saturated model	`~ a * b * c * d`
True model	`~ a * c + b * d + c * d`
	`~ a + b + c * d + a:c + b:d`

Contingency tables

If there is only a single response `a` with two levels, a binomial model can also be used:

```
glm(a ~ 1, weights = Freq, family = binomial)
glm(a ~ b + c, weights = Freq, family = binomial)
```

If `a` has more levels, then a multinomial model can be fitted using

```
multinom(a ~ 1, weights = Freq)
multinom(a ~ b + c, weights = Freq)
```

from the package `nnet`.

Contingency tables

If the response is ordinal, another natural model would be a proportional-odds model. Under such a model the odds ratio for the cumulative probabilities of the levels of `a` does not depend on the cell to which the probabilities belong.

It can be fitted using the `polr()` function from `MASS`:

```
polr(a ~ 1, weights = Freq)
polr(a ~ b + c, weights = Freq)
```

Summary: Contingency tables

Numerical description: The joint distribution of several categorical variables is typically summarized by contingency tables.

In R:

```
table(a, b)
xtabs(~ a + b)
```

For display, `ftable` is also useful.

Summary: Contingency tables

Visualization: Mosaic plots can be used for visualizing contingency tables. As this is a display for conditional frequencies, it is particularly useful for visualizing conditional independence models. For this, the ordering of the margins is important!

The display can also be enhanced by visualizing the residuals of a log-linear model.

In R:

```
mosaicplot(~ a + b)
```

Summary: Contingency tables

In the previous chapter, (generalized) linear regression models were discussed:

dependent var.	explanatory var.	fitting function
1 numerical	≥ 1 num. & cat.	`lm` (OLS)
.....	.....	.....
1 normal 1 poisson 1 binomial	≥ 1 num. & cat.	`glm` (ML via IWLS)

These are closely related to models used for contingency tables.

Summary: Contingency tables

Models for contingency tables include the following models which all compute ML estimates:

dependent var.	explanatory var.	fitting function
—	> 1 cat.	`loglm`, `loglin` (IPF)
≥ 1 cat.	≥ 1 cat.	`glm` + `poisson` (IWLS)
1 binomial	≥ 1 cat. (& num.)	`glm` + `binomial` (IWLS)
.....	.....	.....
1 ordinal	≥ 1 cat. (& num.)	`polr` (BFGS)
.....	.....	.....
1 multinomial	≥ 1 cat. (& num.)	`multinom` (Neural Network)

Summary: Contingency tables

Goodness-of-fit:

Generalizations of the residual sum-of-squares are the deviance (= deviance residuals sum-of-squares) and generalized χ^2 statistic (= Pearson residuals sum-of-squares). The former corresponds to a likelihood ratio test statistic, the latter to a χ^2 test statistic which both have an asymptotic χ^2 distribution.

Model selection:

- Analysis of deviance
- stepwise AIC/BIC

Survival Analysis

Survival Analysis is the analysis of *lifetime data*, especially in medical statistics, but also in studies of reliability.

Lifetime might refer to

- survival time (i.e. time to death of a patient),
- time to recovery or remission,
- time to failure (e.g. of an electronic component).

Survival Analysis

A distinctive feature of survival data is that observations may be *censored*: often the event of interest (death, failure, recovery) has not occured by the end of the study. Hence all is known for these subjects is that the lifetime is *at least* some value.

These observations can not be ignored as they carry important information. And indeed one hopes that many patients are alive at the end of a medical study!

This most common type of censoring is called *right* censoring. Observations can also be *left* or *interval* censored.

Survival Analysis

Example for left censoring: lifetime $\widehat{=}$ time to recurrence of a tumor. This is only observable during surgery.

Two types of censoring can be distinguished:

- type I censoring: n subjects are observed for a fixed time c. The number of censorings is then random.
- type II censoring: observe n subjects until r events occured.

The typical situation in medical studies is type I right censoring.

Survival Analysis

Basic Concepts: The random variable T measures survival time.

- $T > 0$,
- T has the d.f. $F(t) = \mathsf{P}[T \leq t]$,
- T has p.d.f $f(t) = F'(t)$.

The **survivor function** $S(t)$ measures the probability to survive longer than t:

$$S(t) \quad = \quad \mathsf{P}[T \geq t] \quad = \quad 1 \; - \; F(t).$$

The **hazard function** $h(t)$ measures the *risk* or *proneness* to death at time t given survival up to time t.

Survival Analysis

The hazard function represents the instantaneous death rate for an individual surviving to time t:

$$h(t) \quad = \quad \lim_{\delta \to 0} \frac{\mathsf{P}[t \leq T < t + \delta \,|\, T \geq t]}{\delta}.$$

This is also known as the hazard or failure rate.

The **cumulative hazard function** $H(t)$ is defined as

$$H(t) \quad = \quad \int_0^t h(u)\,\mathrm{d}u.$$

$f(t), S(t), h(t), H(t)$ are equivalent ways of defining or characterizing a specific survival pattern uniquely.

Survival Analysis

The following equations hold:

$$\begin{aligned}
h(t) &= \frac{f(t)}{S(t)}, \\
f(t) &= F'(t) \;=\; -S'(t), \\
h(t) &= -\frac{S'(t)}{S(t)} \;=\; -\frac{\partial \log S(t)}{\partial t}, \\
H(t) &= -\log S(t), \\
S(t) &= \exp(-H(t)).
\end{aligned}$$

Survival Analysis

Kaplan-Meier estimator
This is also called product limit estimate of $S(t)$. Assume there are n observations of survival times without censorings occuring p distinct times

$$t_{(1)} \; < \; \ldots \; < t_{(p)}$$

and let d_i the number of deaths at $t_{(i)}$. Then estimate $S(t)$ as

$$\begin{aligned}
\widehat{S}(t) &= 1 - \widehat{F}(t) \\
&= \frac{n - \sum_{j=1}^{s} d_j}{n} \qquad (t_{(s)} \leq t < t_{(s+1)}) \\
&= \frac{n - d_1}{n} \cdot \frac{n - d_1 - d_2}{n - d_1} \cdots \frac{n - d_1 - \ldots - d_s}{n - d_1 - \ldots - d_{s-1}} \\
&= \left(1 - \frac{d_1}{r_1}\right) \ldots \left(1 - \frac{d_s}{r_s}\right) \\
&= \prod_{j=1}^{s} \left(1 - \frac{d_j}{r_j}\right)
\end{aligned}$$

where r_i is the number at risk (i.e. alive) just before $t_{(i)}$. Then $r_{i+1} = r_i - d_i$. If there are censorings calculate the number at risk correctly, i.e., $r_{i+1} = r_i - d_i - c_{i+1}$ if c_i is the number of censorings in the interval $(t_{(i-1)}, t_{(i)})$.

If there are censorings after the last event $\hat{S}(t) > 0 \; \forall \, t$.

Another approach is to use the Nelson estimator of the cumulative hazard

$$\hat{H}(t) \quad = \quad \sum_{j=1}^{s} \frac{d_j}{r_j}$$

and then use the **Fleming-Harrington estimator**

$$\hat{S}_{FH}(t) \quad = \quad \exp(-\hat{H}(t)).$$

Ties can bias the Nelson estimator: assume 3 nearby times t_1, t_2, t_3 with 7 other subjects at risk. The total increment is $1/10 + 1/9 + 1/8$. If the data were tied the increment would be the lesser $3/10$. This is not a problem with $\hat{S}_{KM}$ which has in both cases a multiplicative step of $7/10$.

Different estimates of the variance of $\hat{H}(t)$ are possible:

$\frac{d_j}{r_j(r_j - d_j)}$	Greenwood
$\frac{d_j}{r_j^2}$	Aalen
$\frac{d_j(r_j - d_j)}{r_j^3}$	Klein

To compute the variance of $\hat{S}(t)$ use a simple Taylor series approximation

$$\text{VAR}[\log f] \approx \text{VAR}[f]/f^2$$

giving

$$\text{VAR}[\hat{S}(t)] \quad = \quad \hat{S}^2(t) \, \text{VAR}[\hat{H}(t)].$$

To compute the variance of $\hat{H}(t)$ the Aalen formula and for $\hat{S}(t)$ the Greenwood formula is preferred.

Confidence intervals for $\hat{S}(t)$ can be computed on the plain scale

$$\hat{S} \; \pm \; z_{1-\alpha/2} \, \text{sd}(\hat{S}),$$

which might give values greater than 1 or less than 0; or on the cumulative hazard or log survival scale

$$\exp[\log(\hat{S}) \; \pm \; z_{1-\alpha/2} \, \text{sd}(\hat{H})],$$

which still might be greater than 1; or on a log hazard scale

$$\exp(-\exp[\log(-\log(\hat{S}) \; \pm \; z_{1-\alpha/2} \, \text{sd}(\log \hat{H}))]),$$

which are always between 0 and 1.

Confidence intervals based on the logit of S are another alternative. However, those based on the cumulative hazard scale have the best performance.

Survival Analysis

All methods presented are available in the package `survival`.

```
Surv(time, time2, event, type, origin = 0)
```

is a *packaging* function and typically used as the LHS of a formula. `event` is a status indicator (normally 0 corresponds to "alive", and 1 to "dead"), and `type` describes the type of censoring.

```
survfit(object, data, weights, subset, na.action, ...)
```

Computes an estimate of a survival curve for censored data or computes the predicted survivor function for a Cox proportional hazards model. E.g.,

```
survfit(Surv(time, status) ~ sex)
```

estimates the survivor functions for males and females.

Survival Analysis

k-sample comparisons
To compare the survival probabilities in k groups graphically the Kaplan-Meier curves can be plotted. To test if there is a difference the Fleming-Harrington G^ρ family of tests can be used. The idea is to compute the (weighted) observed and expected number of events in each group $i = 1, \ldots, k$ are computed and compared with a χ^2 statistic.

The expected number of events is given by

$$E_i \quad = \quad \sum_{j=1}^{p} \hat{S}_j^\rho \, \frac{r_{ij}}{r_{\cdot j}} \, d_{\cdot j},$$

and the observed number of events is computed as

$$O_i \quad = \quad \sum_{j=1}^{p} \hat{S}_j^\rho \, d_{ij}.$$

Survival Analysis

$\hat{S}_j$ is the estimated survivor function (from the pooled sample) evaluated just prior to time $t_{(j)}$.

In the case of $\rho = 0$ this yields the *log rank test* and 0_i is just the number of events in group i.

For $\rho = 1$ it is equivalent to the Peto & Peto modification of the *Gehan-Wilcoxon test*.

In R:

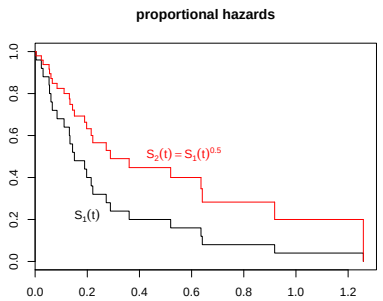
```
survdiff(formula, data,  rho = 0, subset)
```

Survival Analysis

The tests are designed for the alternative of *proportional hazards*

$$h_2(t) \quad = \quad c \cdot h_1(t),$$

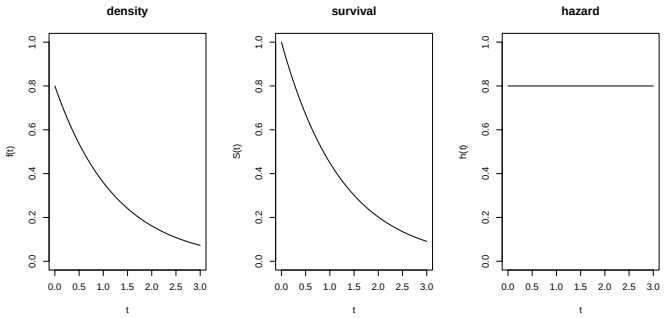
which is equivalent to $S_2(t) \quad = \quad S_1(t)^c$. This implies that the survivor curves do not cross.



Survival Analysis

Parametric models: Exponential distribution:

$$f(t) = \lambda \exp(-\lambda t), \quad S(t) = \exp(-\lambda t), \quad h(t) = \lambda.$$



Survival Analysis

Weibull distribution:

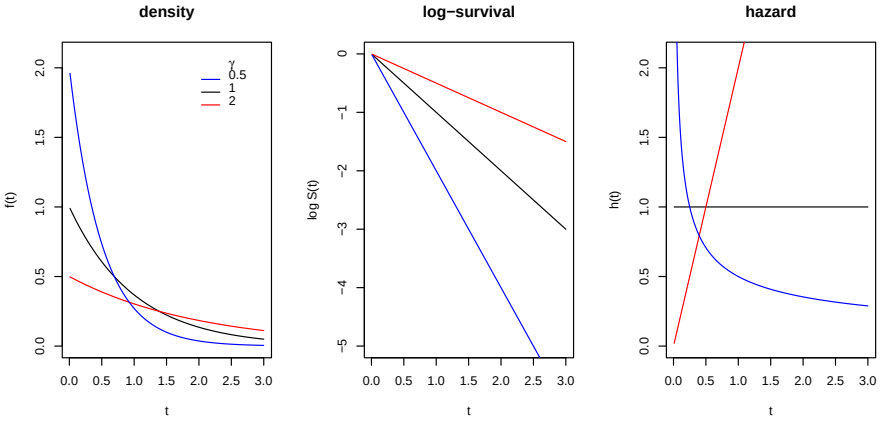
$$\begin{aligned} f(t) &= \lambda\gamma\, t^{\gamma-1} \exp(-\lambda t^{\gamma}), \\ S(t) &= \exp(-\lambda t^{\gamma}), \\ h(t) &= \lambda\gamma\, t^{\gamma-1}. \end{aligned}$$

The hazard function is

- $\gamma > 1$: increasing,
- $\gamma = 1$: constant (exponential),
- $\gamma < 1$: decreasing.

Survival Analysis

With $\lambda = 1$:



Survival Analysis

Regression models: Assume a proportional hazards model for n observations

$$h_i(t) \quad = \quad \lambda_i \cdot h_0(t) \qquad (i = 1, \ldots, n),$$

where $h_0(t)$ is called *baseline hazard*. The factor λ_i is allowed to depend on a set of covariates x_i, usually using a log link

$$\lambda_i \quad = \quad \exp(x_i^{\top}\beta).$$

Thus, the baseline hazard corresponds to an observation with $x = 0$.

To estimate the parameters β there are mainly two approaches:

(i) parametric: parametrize $h_0(t)$,

(ii) semi-parametric: leave $h_0(t)$ unspecified.

(i) Parametric regression models:
The two most common parametric models are

distribution	baseline hazard
Exponential	$h_0(t) = 1$
Weibull	$h_0(t) = \gamma t^{\gamma-1}$

Estimate β by maximum likelihood: e.g. for exponential case

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} [\lambda_i \cdot \exp(-\lambda_i t_i)]^{\delta_i} \exp(-\lambda_i t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{n} \lambda_i^{\delta_i} \exp(-\lambda_i t_i) \\ \log L(\beta) &= \sum_{i=1}^{n} \delta_i \cdot x_i^\top \beta - \sum_{i=1}^{n} \exp(x_i^\top \beta) \cdot t_i, \end{aligned}$$

where δ_i indicates an event or censoring respectively. For the Weibull distribution an additional scale parameter γ is estimated.

In R:

```
survreg(formula, data, subset, na.action, link, dist, ...)
```

where `link` can be `"log"` or `"identity"` and the distribution can be one of `"weibull"`, `"exponential"`, `"logistic"`, `"gaussian"`, `"lognormal"` or `"loglogistic"`. The function returns an object of class `survreg` which inherits from class `glm`.

(ii) Cox proportional hazards model:
Maximum likelihood requires specification of the hazard, hence construct the conditional likelihood. Instead of $\mathrm{P}[T_i = t_i]$ use

$$\mathrm{P}[T_i = t_i \,|\, \text{one individual dies at } t_i].$$

This leads to

$$\frac{h_0(t_i)\exp(x_i^\top\beta)}{\sum_{j:t_j \geq t_i} h_0(t_i)\exp(x_j^\top\beta)} = \frac{\exp(x_i^\top\beta)}{\sum_{j:t_j \geq t_i}\exp(x_j^\top\beta)}$$

and thus, the conditional likelihood is

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(x_i^\top\beta)}{\sum_{j:t_j \geq t_i}\exp(x_j^\top\beta)} \right)^{\delta_i}.$$

If there are censorings this is called *partial likelihood*.

In R:

```
coxph(formula, data, weights, subset, na.action, ...)
```

For both parametric and semi-parametric models asymptotic normality can be shown. Hence, confidence intervals can be constructed and inference about the paramater estimates can be done in the usual way (LR, Wald, Score test).

The estimates can be interpreted as follows: for observations with covariates x_1 and x_2 respectively

$$\frac{h_1(t)}{h_2(t)} = \exp((x_1 - x_2)^\top\beta),$$

independent of t.