BIOSTATISTICS

① Introduction

1) derive whole populat. properties from random sample ← Point / interval, Stat tests, Anova, Survival
2) Rules for sample determination < study design, Power analysis
3) Interprete test results

② Probability triple (space)

1) sample space Ω : set of outcomes
2) set of events $\mathcal{F}$: collection of all, closed events E (each: subset of outcomes) $+\Omega, \emptyset$ $\mathcal{F}: \{\emptyset, \Omega, \{1,2\}, \Omega/\{1,2\}\}$
3) probability measure P: each event E of set $\mathcal{F}$: probability $P(E)$ $P: \mathcal{F} \to [0,1]$

Kolmogorov axioms for P

1) $0 \leq P(E) \leq 1$
2) $P(\Omega) = 1$
3) $P(\bigcup_i E_i) = \sum_i P(E_i)$ · E_i disjoint

Events - σ-Algebra $\mathcal{F}$
1) $\Omega \in \mathcal{F}$
2) $A \in \mathcal{F} \to \bar{A} \in \mathcal{F}$
3) $A_1, \dots \in \mathcal{F} \to \bigcup_n A_n \in \mathcal{F}$
smallest $\mathcal{F} = \{\emptyset, \Omega\}$, biggest: potency

Consequences

1) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (sum rule)
2) $P(\Omega \setminus A) = 1 - P(A)$ (compl. prob.)
3) $P(B|A) = \dfrac{P(B \cap A)}{P(A)}$ (cond. prob.) ... normieren, subset selection then count

Terms

i) Independence A, B if $P(A \cap B) = P(A)P(B)$
ii) Disjoint A, B if $P(A \cap B) = 0$ (mutually exclusive) $\equiv$ dependent

Bayes theorem

N disjoint events $\Omega = A_1 \cup \dots \cup A_N$, $B \notin \Omega$

$$P(A_i | B) = \frac{P(A_i) \cdot P(B|A_i)}{\sum_k P(A_k) \cdot P(B|A_k)} \equiv \langle P(A_i) P(B|A_i) \rangle$$

disease 1/2, enzyme test

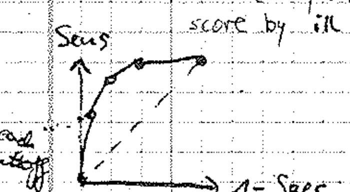
ROC curves (receiver operating characteristics) → categorial biomarker

$$\text{Sens} = \frac{TP}{TP + FN} \qquad \text{Spec} = \frac{TN}{TN + FP}$$

$\text{Sens} = P(B = \text{pos} | A_1)$ score by ill

$\text{Spec} = P(B = \text{neg} | A_2)$ score free by healthy

$A \setminus B$ → Marker 1 2 3 4 5 ORDERED
true A_1, A_2
move cutoff and recount



$A = 0{,}5$ → no correla
$A \to 1$ → strong pos corr
$A \to 0$ → " neg corr

Random variable

- function assigning real numbers to results of experim
- Elements of Ω ⟶ Real numbers $\mathbb{R}$
  Random variable X ($\in d_1 + A_2$)

discrete: finite / countably infinite $P(X \leq x)$
continuous: $P(X \leq x)$, $\forall x \in \mathbb{R}$ repr. by int dens func $f \geq 0$; $P(X \leq x) = \int_{-\infty}^{x} f(x')\,dx'$

→ Probability mass (disc) (density (cont)) function ; cumulative distribution function

Binomial distribution

$f(k;n,p) = P(X=k) = \binom{n}{k} p^k (1-p)^k$ $\qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$

Prob. of k successes in n trials, prob p success

$E(x) = np \quad = E(X) = \sum k \cdot f(k)$

$V(x) = np(1-p) = E((X-E(x))^2) = \sum f(k)(k-np)^2$

Poisson distribution

$P(X=k) = \frac{\mu^k e^{-\mu}}{k!} \qquad k = 0,1,2,\ldots$

μ: expected nr of occurr. in interval

k : nr of actual occurr of event

$E(x) = \mu$

$V(x) = \mu$

FIXED INTERVAL

Gaussian distribution

$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

μ: mean value

σ: standard deviat

symmetric

$E(x) = \mu = \int x f(x)\,dx$ $\qquad \mu \pm \sigma$: 68%

$V(X) = \sigma^2 = \int (x-\mu)^2 f(x)\,dx$ $\qquad \mu \pm 2\sigma$: 95%

→ pattern when $\left\{\begin{array}{l}\text{large number events}\\ \text{independent}\\ \text{random}\\ \text{small effect}\end{array}\right.$

Standard normal distribution

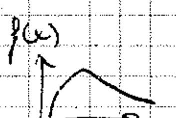
$\mu = 0, \sigma = 1$

$f(x_{st}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_{st}^2}{2}}$

Logarithmic Gaussian distribution

$y = \log x \rightarrow x = e^y$

$y \downarrow N(\mu, \sigma^2)$, $x \downarrow$ log-normal



Uniform distribution

$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$

$E(x) = \frac{b+a}{2}$

$V(x) = \frac{1}{12}(b-a)^2$

Exponential distribution

$f(t) = \lambda e^{-\lambda t}$ → INTERVAL BETWEEN TWO EVENTS

$E(t) = \frac{1}{\lambda}$

$V(t) = \frac{1}{\lambda^2}$

χ^2 distribution

$X_1, \ldots X_n \sim N(0,1)$ $\qquad$ n: degrees of freedom

$X_1^2 + \ldots + X_n^2 \sim \chi_n^2 = \Gamma\left(\frac{n}{2}, 2\right)$ $\qquad \rightarrow n \gg 1 \Rightarrow N(n, \sqrt{2n})$

$E(x) = n$

$V(x) = 2n$

F - distribution (Fischer)

homogeneity tests

$X \sim \chi_m^2, \; Y \sim \chi_n^2$

$F_{mn} = \frac{X/m}{Y/n}$; $\frac{1}{F_{mn}} = F_{nm}$

$E(F) = \frac{n}{n-2} \qquad n > 2$

$V(F) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad n > 4$

Student's or t- distribution

$f_n(x) = C_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \overset{n \gg 1}{\approx} \mathcal{N}(0,1)$ n: degrees of freedom

$E(X) = \begin{cases} 0 & n > 1 \\ / & \text{oth} \end{cases}$

$X \sim \mathcal{N}(0,1)$, $Q \sim \chi_n^2$: $T = \frac{X}{\sqrt{Q/n}} \sim t$

$V(X) = \begin{cases} \frac{n}{n-2} & n > 2 \\ \infty & 1 < n \leq 2 \\ / & \text{oth} \end{cases}$

Moments

$\mu'_{n}(c) = \int_{-\infty}^{+\infty} (x-c)^n f(x)\,dx$ // standardized: divide σ : $\left(\frac{x-c}{\sigma}\right)^n$

- Mean $\mu = \mu'_1(0) = \int_{-\infty}^{+\infty} x\, f(x)\,dx$
- Variance $\sigma^2 = \mu'_2(\mu) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx$
- Skewness $\gamma_1 = \mu'_3(\mu)/\sigma^3 = \int \left(\frac{x-\mu}{\sigma}\right)^3 f(x)\,dx = \frac{\mu_3(\mu)}{\sigma^3}$
- Kurtosis $\gamma_2 = \mu'_4(\mu)/\sigma^4 = \int \left(\frac{x-\mu}{\sigma}\right)^4 f(x)\,dx$

Levels measure

- Nominal - label $= \neq$
- Ordinal - ordered $< >$
- Interval - difference $+ -$
- Ratio - zero $* /$

③ Fundamentals

do ex. 3.1

- Arithmetic mean $\bar{x} = \frac{1}{n}\sum_i x_i$ → minimizes $(x-\bar{x})^2$
  - ↳ linearity: $y_i = c_1 x_i + c_2 \rightarrow \bar{y} = c_1\bar{x} + c_2$
- Geometric mean: $\bar{x}_g = \sqrt[n]{\prod x_i}$ ... growth processes

  $\log(\bar{x}_g) = \frac{1}{n}(\log x_1 + \ldots + \log x_n)$

- Median of ordered sample (n) → robust against outliers → minimizes $|x-\tilde{x}|$

  $\tilde{x} = \begin{cases} x_{(n+1)/2} & n \text{ odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & n \text{ even} \end{cases}$

  $\bar{x} = \tilde{x}$ symmetric

  $\bar{x} > \tilde{x}$ pos. skewed ... paar verd. viel

  $\bar{x} < \tilde{x}$ neg skewed

- Mode : most occurring value → not useful

3-2

Spread

- **Range**

$r = x_{max} - x_{min}$ → sensitive outliers → depends on n

- Percentiles p (sample n) / Quantile

$k = np/100$

$$\begin{cases} V_p = X_{ceil(k)} & \text{if } k \text{ not integer} \\ V_p = \frac{1}{2}(X_k + X_{k+1}) & \text{if } k \text{ is integer} \end{cases}$$

p-th percentile is a value V_p such that (at least) $p\%$ sample points $\leq V_p$

$V_p = \Phi^{-1}(p)$

V_{25}, V_{50}, V_{75} → 1st, 2nd, 3rd quartile

Quantil distance $QD = V_{1-p} - V_p$

- Variance

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$ sample variance $= \frac{1}{n-1}\left(\sum x_i^2 - n\bar{x}^2\right)$ $n-1$ → converge to true value (n fast)

$s = \sqrt{s^2}$ sample standard devix

(i) $y_i = x_i + c$ ⟶ $S_x^2 = S_y^2$ $V(X) = E(X^2) - (EX)^2$

(ii) $y_i = cx_i$ ⟶ $S_y^2 = c^2 S_x^2$; $S_y = cS_x$

- Coefficient of variation

$C_v = \frac{s}{\bar{x}} \cdot 100\%$

Sample means ...assuming equal size, otherwise weighting

$\bar{X}_{\bar{x}} = \frac{1}{m}\sum_i \bar{x}_i$; $\bar{x}_{\bar{x}} = \bar{x}_{glob}$ mean of sample means

$S_{\bar{x}}^2 = \frac{s^2}{n}$ variance of sample means

$S_{\bar{x}} = \sqrt{S_{\bar{x}}^2} = \frac{s}{\sqrt{n}}$ Standard error of the mean (standard devix of sample means)

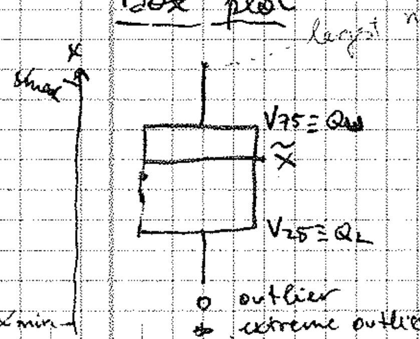
Covariance

$Cov(X_1, X_2) = E[(X_1 - EX_1)(X_2 - EX_2)] = \frac{1}{n-1}\sum_i (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$

$\rho(X_1, X_2) = Cov(X_1, X_2) / (\sigma_1 \sigma_2)$ correlation coefficient

If $\rho(X_1, X_2) = 0$ → X_1 and X_2 uncorrelated (independent)

Box plot largest not outlying value



- learn width, skewness, median

(1) $Q_U - \tilde{x} \approx \tilde{x} - Q_L$ symm

(2) $>$ pos. skewed (to the right)

(3) $<$ negat. " (to the left)

- Outlier $x \gtrless$ $Q_U + 1.5(Q_U - Q_L)$, $Q_L -$

- Extreme out $\gtrless$ Q_U $+ 3$, Q_L $-$

(4) Point - interval estimation

- sample of population unknown pdf → measure location (μ) and spread (σ)

  $\bar{X} \approx \mu$ (arith mean)

  $s \approx \sigma$ (sample var)

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ Distribution of arithmetic means

  $\Delta\mu = \frac{\sigma}{\sqrt{n}}$; $\Delta\bar{X} = \frac{s}{\sqrt{n}}$

(1) $X \sim N(\mu, \sigma^2)$
(2) $\bar{X} \sim N(\mu, \sigma^2/n)$
(3) $X_{st} = \frac{1}{\sigma}(X - \mu)$

(4) $\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$ (5) $\frac{\bar{X}-\mu}{s}\sqrt{n} \sim t_{n-1}$

$\sigma \Leftrightarrow s$

Normal t - dist

Student distr. $n-1$ d.o.f.

Central limit theorem

$x_1, \dots, x_n$; population μ, σ^2

$X \nsim N(\mu, \sigma^2)$ not normal (single not, but means are)

but for large n (>20) the mean $\bar{X} \sim N(\mu, \sigma^2/n)$

Interval of mean:

$Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \sim N(0,1) \rightarrow -1{,}96 < Z < 1{,}96 \Rightarrow 95\%$ ($1{,}96 = z_{97{,}5}$)

$\mu \in \underbrace{\left(\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}\right)}_{\text{conf. interval}}$ with 95% probability, from repeated samples of size n

$-1{,}96 < z < 1{,}96$

$\Updownarrow$

$t = \frac{\bar{X}-\mu}{s}\sqrt{n} \sim t_{n-1}$

$\mu \in \underbrace{\left(\bar{X} \pm t_{n-1,\,97{,}5\%} \cdot \frac{s}{\sqrt{n}}\right)}_{\text{conf. interval}}$ w 95% prob.

$t_{n-1,\,97{,}5\%}$: 2,78 ($n=5$); 2 ($n=60$); 1,96 ($n>60$)

$-2 < t < 2$

$\rightarrow P\left(t_{n-1,\frac{\alpha}{2}} < T < t_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$

$1-\alpha$: confidence level

α: error probability (that $\bar{x}$ outside μ interval)

- Higher n: smaller (sharper) conf. interval
- $(1-\alpha)\cdot 100\%$ of all conf. int will include true (unknown) mean

Only $\alpha \cdot 100\%$ not " (error)!

- When $n > 200$

  $\mu \in \left(\bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right)$

For σ^2

If $X_i \sim N(\mu, \sigma^2) \rightarrow \frac{n-1}{\sigma^2}s^2 \sim \chi^2_{n-1}$

$P\left(\chi^2_{n-1,\frac{\alpha}{2}} < \frac{(n-1)s^2}{\sigma^2} < \chi_{n-1,1-\frac{\alpha}{2}}\right) = 1-\alpha$

CI: $(n-1)s^2/\chi^2 \dots < \sigma^2 < (n-1)s^2/\chi^2 \dots$

(5) Statistical tests

- Test hypothesis, unique decision making criterion
- H_0 : null hypothesis (of no effect), what you want to reject
- H_1: alternative hypoth.

  × equality distribution parameters (μ, σ)
  × " pdf
  × correl'd random variables

(0) H_0, H_1
(i) Test variable
(ii) Significance level / error probability α (5%)
(iii) Critical region B such that $P(T \in B \mid H_0 \text{ true}) \leq \alpha$

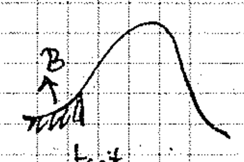
t - test

- Comparison 2 mean values • samples NOT < multimodal / too much skewed
- Normally distributed X, Y ↑, not very sens. to devi'd
- $\sigma_x = \sigma_y$ (reasonably) → F-test

ONE SAMPLE ONE-SIDED [one arm vs popul'd]

- $H_0 : \mu = \mu_0$
- $H_1 : \mu < \mu_0$
- $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ ≡ $\dfrac{\text{Parameter estim - true value}}{\text{Param. error}}$ → devi'd in units of std dev.

  $s/\sqrt{n} \equiv \Delta\bar{x}$



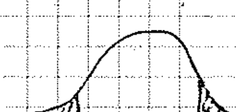
$H_1 : \mu > \mu_0$

- $\alpha = 0{,}05$
- Decision: $t \geq t_{n-1,\alpha}$: accept H_0 — $t \leq t_{n-1,\alpha}$
  $t < t_{n-1,\alpha}$: reject H_0 w.s.l — $t > t_{n-1,\alpha}$

Due to $t \leq$ -- , H_0 is reject w α% s.level, i.e. ..

ONE SAMPLE TWO-SIDED (more conservative)
↳ tcrit ~~larger~~ larger

- $H_0 : \mu = \mu_0$
- $H_1 : \mu \neq \mu_0$
- $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$



- $\alpha = 0{,}05$
- $|t| \leq t_{n-1, 1-\frac{\alpha}{2}}$: accept H_0 / cannot be rejected
  $|t| > t_{n-1, 1-\frac{\alpha}{2}}$: reject H_0 w.s.l

[double arm study]

TWO SAMPLES TWO-SIDED (indep. samp.)

- $H_0 : \mu_x = \mu_y$ $\quad \bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma^2}{n_x} + \frac{\sigma^2}{n_y}\right) \xrightarrow{\sigma \to s}$ t distributed
- $H_1 : \mu_x \neq \mu_y$
- $t = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$; $s = \sqrt{\dfrac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2}}$ pooled empirical standard deviation
- $\alpha = 5\%$ $\quad$ • $|t| \leq t_{K, 1-\frac{\alpha}{2}}$ accept H_0
- $K = n_x + n_y - 2$ $\quad$ • $|t| > t_{K, 1-\alpha/2}$ reject H_0

TWO SAMPLES PAIRED T-TEST [patient in two arms]

- two samples on same patient (paired), Individuum two diff. states
  - e.g. diuretic / placebo; before/after treatm. → better than two arms: where $\sigma_x \neq \sigma_y$ diff. pat.
- Two states:
  $\{x_1, \dots x_n\}$
  $\{y_1, \dots y_n\}$
- $d_i = y_i - x_i$
  $\{d_1, \dots d_n\}$
- Mean of diff.
  $\bar{d} = \frac{1}{n} \sum_i d_i$
- Sample st. dev. of diff
  $s_d = \sqrt{\frac{1}{n-1} \sum_i (d_i - \bar{d})^2}$

Test ⟹ $\mu_x \neq \mu_y$
$\mu = \mu_x - \mu_y$

- $H_0: \mu = 0$
- $H_1: \mu \neq 0$
- $\alpha = 5\%$
- $t = \frac{\bar{d}}{sd/\sqrt{n}}$
- $K = n-1$
- $|t| \leq t_{K, 1-\frac{\alpha}{2}}$ accept H_0
  $|t| > t_{K, 1-\frac{\alpha}{2}}$ reject H_0

t - test $\sigma_x \neq \sigma_y$

$X \sim \mathcal{N}(\mu_x, \sigma_x^2)$
$Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$

- $H_0: \mu_x = \mu_y$
- $H_1: \mu_x \neq \mu_y$

$\bar{X} - \bar{Y} \sim \mathcal{N}\left(0, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$

- $t \approx \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$ (not t-distrib except if n large)
- $d' = \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2 \Big/ \left[\frac{(s_x^2/n_x)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}\right]$
- $d'' = \text{floor}(d')$
- $|t| > t_{d'', 1-\alpha/2}$: reject H_0

Nonparametric methods

t - test not applicable if
- not normal distr. (μ, σ) not enough
- $\sigma_x \neq \sigma_y$
- central limit theorem not applic.

↓

Assess choice based on:

- F-test eq. of variances
- $\bar{x}, \tilde{x}, Q_U, Q_L$ → box plot
- histogram

Opinions

(1) Parametric if no evidence of no-normal
- more powerful than non-parametric
- use nonp. only if positive evidence no-normal

(2) Nonparametric always, except pos. evidence that par. are applic.
- 95% of power of param.
- should assume as little as possible on data

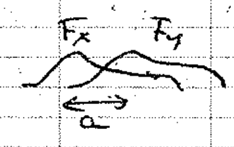
RANK TESTS

- group/rank → ordered serial
- based on relative sizes of observations
- not on the size itself

Parametric		Nonparametric
1) Two samples indep. t-test	→	Mann-Whitney rank sum test
2) Two samples paired t-tes	→	Wilcoxon signed-rank test

Conditions for applicability

- 2 indep. random variables X, Y distributed F_X, F_Y
- $F_Y(x) = F_X(x-a)$ → differ only by a shift
  ↳ $\sigma_x = \sigma_y$



- Independent samples $X_1, \dots X_m$; $Y_1, \dots Y_n$
- $H_0: a = 0$; $H_1: a \neq 0$

Mann-Whitney rank sum test H_0: same popul., drug: no effect

(1) Combine data both samples

(2) Order values lowest to highest

Sam.	A	B	A	B
x_i	100	101	103	105
r	1	2	3	4

(3) Assign ranks to indiv. values

(4) Group same value → $r' = r + \frac{1+g}{2}$ $(x_{r+1}, \dots, x_{r+g})$

(5) Compute rank sum R_1 for first sample (lowest n)

(6) Calculate critical value for $T(R_1)$

- (i) $T = R_1$ → histogram of all rank sum posibilities $n_1, n_2 < 5$
- (ii) $T = R_1$ → lookup table: if $T_{crit}^{l} < T < T_{crit}^{h}$ ↳ accept H_0 $n_1, n_2 < 10$
- (iii) $T(R_1) \sim N(0,1)$ if $T \leq z_{1-\frac{\alpha}{2}}$ $n_1, n_2 > 10$

Wilcoxon signed-rank test   extra condit.: symmetric (often neglected)
$H_0: d = 0$, $H_1: d \neq 0$; $d = \tilde{x} - \tilde{y}$

(1) $d_i = x_i - y_i$ } arrange order abs. values

(2) Ignore $d_i = 0$; rank rest 1 to n with sign!

(3) Group same abs. value $r' = r + \frac{1+g}{2}$ $(|d_{r+1}|, \dots |d_{r+g}|)$

(4) Calculate signed rank sum W

(5) Compute critical value W_{crit}

- (i) histogram of all possib. + − to n differences $n \leq 6$
- (ii) lookup table $|W| \leq W_{crit}$ → accept H_0 $n < 16$
- (iii) $T(W) \sim N(0,1)$ $T \leq z_{1-\frac{\alpha}{2}}$

p-Value

- significance level α at which $t = t_{crit,\alpha}$
- $p = \alpha_{crit}(t = t_{crit,\alpha}) = P(t_{n-1} \leq t)$

$p < 0.001$ very highly significant
$0.001 \leq p < 0.01$ highly significant
$0.01 \leq p < 0.05$ significant
$0.05 \leq p < 0.1$ trend towards significance
$p > 0.1$ not statistically significant

- Two methods for determining stat. significance
  1) Critical value method $t \leftrightarrow t_{crit}$: rejection/acceptance
  2) p-value method p (exact) < 0.05 reject, but gives more info
     $p = 2 \times (1 - \Phi(|t|))$

(6) Analysis of variance

- Means of more than two groups have to be compared
  ↳ pair-wise two-samples cannot be applied directly!

⇒ ANOVA method

χ^2 test one sample

- $X \sim \mathcal{N}(\mu, \sigma^2)$; empir. var s^2; n compare with known σ_0^2
- $\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} \rightarrow \chi^2_{n-1}$ distributed, n-1 d.o.f.
- $H_0: \sigma = \sigma_0$; $H_1: \sigma \neq \sigma_0$, α
- If $\chi_0^2 > \chi^2_{n-1,1-\frac{\alpha}{2}}$ or $\chi_0^2 < \chi^2_{n-1,\frac{\alpha}{2}}$ reject H_0

F-test

- tests if $\sigma_x = \sigma_y$ (precondition for t-test)
- Two samples $\{X\}$ and $\{Y\}$, independent, $\{n_x\},\{n_y\}$, normally distributed
- $F = \begin{cases} \frac{S_x^2}{S_y^2} & \text{for } S_x > S_y \\ \frac{S_y^2}{S_x^2} & \text{for } S_y > S_x \end{cases}$ follows F distribution with $\left\{ \begin{array}{l} m_1 = n_x - 1,\ m_2 = n_y - 1 \\ m_1 = n_y - 1,\ m_2 = n_x - 1 \end{array} \right\}$ d.o.f

$\frac{S_{max}^2}{S_{min}^2}$ $m_1 = n_{max} - 1$, $m_2 = n_{min} - 1$

- $H_0: \sigma_x = \sigma_y$, α
- H_1 one-sided: $H_1: \sigma_x < \sigma_y$ or $H_1: \sigma_x > \sigma_y$ → reject H_0 if $F > F_{1-\alpha, m_1, m_2}$
- H_1 two-sided: $H_1: \sigma_x \neq \sigma_y$ $F > F_{1-\frac{\alpha}{2}, m_1, m_2}$

The one way ANOVA

- Comparison of means of arbitrary number of groups
- Each group $\mathcal{N}(?, \sigma^2)$, same σ
- Determine whether variability dominated by < spread within groups / spread between groups $(\mu_1 \neq \mu_2)$

• k groups, n_i, s_i, $n = \sum n_i$

• $s^2_{within} = \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1)\, s_i^2$ pooled variance

• $s^2_{between} = \frac{1}{k-1}\left[\sum_{i=1}^{k} n_i \bar{x}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{k} n_i \bar{x}_i\right)^2\right] = n\, s_{\bar{x}}^2$

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

Recipe

(1) H_0: all groups have same mean ($s_{betw} = s_{within}$)

H_1: at least one group diff mean

(2) $F = \frac{s^2_{between}}{s^2_{within}}$

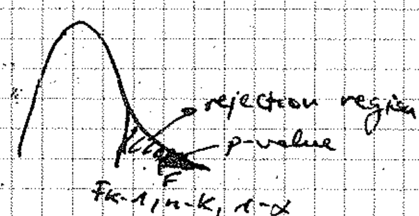
$\rightarrow s_{between} > s_{within}$

$\Rightarrow F > 1$

(3) If $F \leq F_{k-1, n-k, 1-\alpha}$ accept H_0

$F >$ reject H_0

(4) p-value



If difference, compare specific groups

• Two specific (of k) groups

• $H_0: \bar{x}_1 = \bar{x}_2$; $H_1: \bar{x}_1 \neq \bar{x}_2$; α

(1) $s^2 = s^2_{within}$

(2) $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

(3) If $t > t_{n-k, 1-\frac{\alpha}{2}}$ or $t < t_{n-k, \frac{\alpha}{2}}$ reject H_0

otherwise accept

Multiple comparisons: Bonferroni approach

• ensure overall probab. any signif. differences all possible group not $> \alpha^*$.

$\alpha^* = \frac{\alpha}{\binom{k}{2}}$ → tcrit at α^* instead α → more demanding $\alpha^* < \alpha$

• p value stays the same

[1] ANOVA showed at least one group diff μ

[2] Bonferroni mult comparisons specific groups

Bonferroni-Holm

Groups: p_1 p_2 p_3 → stop after 1st non-significant

$\tilde{p}_1 = 4 \cdot p_1$ $\tilde{p}_2 = 3 \cdot p_2$ $\tilde{p}_3 = 2 \cdot p_3$

tests

• significant

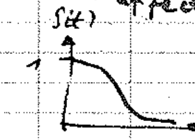
• $p = \alpha_{obs} (t = t_{obs})$

p-value

(7) <u>Survival analysis</u> 8

- final state (result) [binary] of some individuals is unknown < study dura / dif start time / patient not follow
  ↳ clinical studies / survival < 5 y surv / 5 y local tm control / 5 y progr - free

- Censored observation → no follow-up / study end → at least survived X yr
  ↳ X + ↳ affects only d_i
- Endpoints types < death / disease recurrence / explosion



<u>Survival function / probability $S(t)$</u>

$$S(t) = \frac{\text{Nr. indiv. surviving} > t}{\text{Total nr. indiv.}} = P(T > t) \quad T: \text{time until death}$$

t_{50}: median survival time $= S^{-1}(0.5) = Y_{50}$

$\hat{S}(t)$ of sample → observe until all die

<u>Kaplan-Meier estimator</u>

(1) n_i probands being observable at beginning of time interval i → $[t_{i-1}, t_i]$; $t_0 = 0$

(2) d_i individuals die ; l_i censored at end of interval $i = t_i$

↳ $n_{i+1} = n_i - d_i - l_i$ at beginning of time int. $i+1$

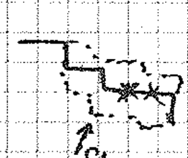
$$\hat{S}(t_i) = \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j}\right)$$
→ only for t_i where death, not censoring
→ do not count l_i as d_i

<u>Greenwood formula</u>

$$S_{\hat{S}(t_i)} = \hat{S}(t_i) \sqrt{\sum_{j=1}^{i} \frac{d_j}{n_j (n_j - d_j)}}$$
→ all previous time intervals
→ standard devia of $\hat{S}$

Confidence interval $\hat{S}(t_i) \pm z_\alpha \, S_{\hat{S}(t_i)}$ + truncate $[0, 1]$

z_α: $N(0,1)$ two-sided
$z_{5\%} \to 1.96$



CI

<u>Log-rank test</u> - Comparison of survival curves (two samples)

- nonparametric test, combine all 1) and suppress censored but reduce n_{i+1} [times two colours]
- t_i : times of death end int. , censored not shown
- d_i : deaths end of interval i
- n_i : persons alive & observ. at beginn t. int. i
- 3) $f_i = \dfrac{d_i^{(tot)}}{n_i^{(tot)}}$: total (conditional) prob. to die at t_i at combined gr.
- 4) $e_i = n_i^{(a)} \cdot f_i$: expected death 2) group (a) [from mort. combined] → to be compared w. combined (tot)
- $U_{Li} = d_i^{(a)} - e_i^{(a)}$: diff exp. - meas (a)
- $S_i(U_L)^2$: contrib. to empirical devia of U_L = $\dfrac{n_i^{(b)} n_i^{(a)} d_i^{(tot)} (n_i^{(tot)} - d_i^{(tot)})}{n_i^{(tot)2} (n_i^{(tot)} - 1)}$
- 5) $U_L = \sum U_{Li}$; 6) $S_{U_L}^2 = \sum S_{U_{Li}}^2$
- 7) $z = \dfrac{|U_L| - \frac{1}{2}}{S_{U_L}} \sim N(0,1)$ distributed
- H_0: survival curves same , H_1: diff ; 8) α
- 9) $z > z_{crit}$ → reject H_0

i	$d_i^{(a)}$	$n_i^{(a)}$	$d_i^{(b)}$	...	(tot)
1					
2					
3					
5					

f_i
e_i
U_L
S_i

* only two groups
* cannot test if other factors like age have influence

(8) Study design

Types < Experimental - Prospective - Longitudinal
Observational < Prospective < Longitudinal / Cross sectional
Retrospective < Longitudinal / Cross sectional

- Observational : DACQ without intervention (e.g. winter study)
- Experimental : (1) set hypothesis
  (2) define intervention (regular sauna)
  (3) measure effect

- Prospective : intervention and later DACQ (exp. alw pros.)
  (1) sauna prev. flu
  (2) establish two groups < no sauna / 1 /week + select sample (age, gender) : problem
  (3) diagnosis flu for each group
- Retrospective : data related to past
  - Interview flu / sauna
  - ⊕ Large samples
  - ⊖ Degree of truth?

- Longitudinal : × consecutive intervention and observation of events
  × multiple observations in time e.g. 5 years survival prosp. sauna study
- Cross sectional : single data taking in sample e.g. screening, surveys retrosp. sauna study

Randomisation

- For prosp. exp. studies w. alternative interventions < phot / prot ther.
  → random selection patients two arms
  → eliminate subject bias : blinded study
  → " inv + " " : double blinded study
  → equal dist of known/unknown bias factors on both arms
- Sample randomisd : select patients via random numbers
  ↳ bad: two arms diff size
- Block randomisd : - equal distrib. of subjects onto the study arms
  AABB - define block size : 4
  ABAB - both arms equal weight each block
  ⋮
  → r. number 1-6 : block allocation sequence
- Stratified randomisd : - balance of imp. features (age) in each arm
  - block randomise for each stratum

→ 3 subgroups per age

(9) Power of statistical tests

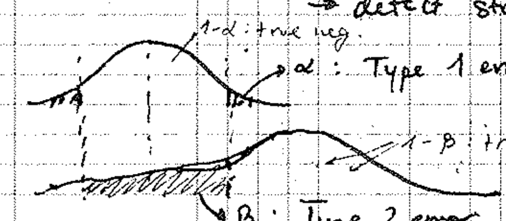
- methods decide whether data compatible with hypothesis H_0
- F, t, z, W, U test statistics
- H_0 rejected if test value out of 95% acceptance region (presuming H_0)

× 2 samples same popul? t-test; Mann-Whitney rank test
× 2 popul? same varia? F-test
× 2 survival curves equal log-rank

- $p < 0.05$ stat significant → test value out of 95% region
- $p \geq 0.05$ not stat sig → H_0 cannot be rejected ≠ H_0 is valid
  could not be proved that is not valid

Reality \ Decision	H_0 rejected (≡ positive)	H_0 accepted (≡ negative)
H_0 false	True positive $p = 1-\beta$	False negative $p = \beta$ Type II error
H_0 true	False positive $p = \alpha$ Type I error	True negative $p = 1-\alpha$

Power of test $\boxed{1-\beta}$ → probability of true positive decision
= correctly rejecting H_0
→ detect stat sign. diff when H_0 really false



$\beta = \Phi(\ldots) - \ldots$
$\delta^* \sqrt{n} + z_{1-\frac{\alpha}{2}}$

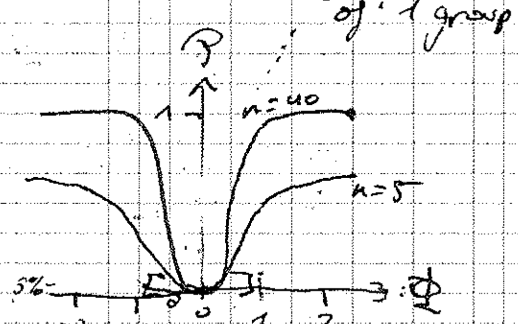
Power depends on: (1) chosen α, (signif. level) $P \propto \alpha$; n
$n_1 = n_2 = n$ (2) ratio between diff to be detected and SEM $P \propto \dfrac{\Delta\mu}{\sigma\sqrt{\frac{2}{n}}}$

$$t' = \frac{\overbrace{\mu_1 - \mu_2}^{\delta}}{\sigma\sqrt{2/n}} \rightarrow \Phi = \frac{\delta}{\sigma} \text{ noncentrality parameter}$$

σ ... of sample, not of mean
... prop to their effect in units of st dev

$P \uparrow$
- $n \uparrow$ (2x)
- $\alpha \uparrow$
- $\Delta\mu \uparrow$, $\sigma \downarrow$: $f = \frac{\delta}{\sigma} = \Phi \uparrow$

Only size of 1 group

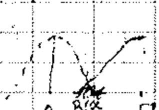


t-test power function
- families of curves [n], fixed α, depending on Φ
- obtain n ($P = 80\%$, $\Phi = 1$, $\alpha = 5\%$) inverting
- study should designed / $P \geq 80\%$

Sample size estimat

$$n \approx \frac{2\sigma^2}{\delta^2}\left(z_{1-\beta} + z_{1-\frac{\alpha}{2}}\right)^2 \quad \text{two-sided}$$

$t_{1-\beta}$ $t_{1-\frac{\alpha}{2}}$



$$\boxed{z_{1-\frac{\alpha}{2}} + z_{1-\beta} = \delta'}$$

$$\delta' = \frac{\delta}{\sigma}\sqrt{\frac{n}{2}} = t'$$

~~Hazard function~~

~~$F(t) = 1 - S(t)$ cdf~~ → $h(t) = -\dfrac{S'(t)}{S(t)}$

~~$f(t) = -S'(t)$~~ pdf

(10) **Regression**

(phenomena in nature)

- stochastic dependency between two random variables X, Y
  - ↳ stoch. factors influencing < either X or Y / both

$X = X(U_1, \dots U_M, V_1, \dots V_j)$
$Y = Y(U_1, \dots U_M, W_1, \dots W_j)$ are stoch. dependent

Regression theory

- predict random variable Y if X is known / fixed
- regression curves in xy plane

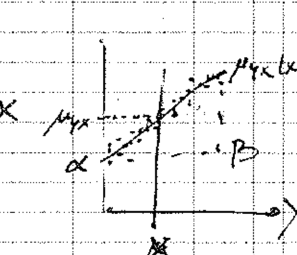
$$\hat{y}(x) = \underbrace{E}_{\text{condit. expected}}(Y \mid X = x), \quad \hat{x}(y) = E(X \mid Y = y)$$

↳ location of most accurate prediction for Y if X has value x
↳ minimising $E[Y - \hat{y}(x)]^2$ mean square error

Linear regression

$\hat{y}(x)$ straight → X and Y linearly correlated

- $\mu_{yx}(x) = \alpha + \beta x = E[Y(x)]$: mean of all y at certain x
- $\sigma_{yx}(x)$: stdev of all y at certain x
- Preconditions:
  - $\mu_{yx} = \alpha + \beta x$
  - For all x, $Y \sim \mathcal{N}(\mu_{yx}, \sigma_{yx})$
  - σ_{yx} constant for all x



population α, β → a, b estimators sample

$\sum_k (y_k - a - b x_k)^2 \rightarrow \frac{\partial}{\partial a} = 0 \; ; \; \frac{\partial}{\partial b} = 0$

$$b = \frac{n \sum XY - \sum X \sum Y}{n(\sum X^2) - (\sum X)^2} \; ; \; a = \bar{Y} - b\bar{X}$$

$$S_{yx} = \sqrt{\frac{\sum[Y - (a + bx)]^2}{n-2}} = \sqrt{\frac{n-1}{n-2}(S_y^2 - b^2 S_x^2)}$$

$$S_a = S_{yx}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}} \; ; \quad S_b = \frac{1}{\sqrt{n-1}} \frac{S_{yx}}{S_x}$$

Confidence intervals

$t = \frac{b - \beta}{S_b} \rightarrow \beta \in (b \pm t_{k,1-\frac{\alpha}{2}} S_b)$ $\quad k = n-2 = $ d.o.f.
↳ $\beta > 0$ sign.

$t = \frac{a - \alpha}{S_a} \rightarrow \alpha \in (\alpha \pm t_{k,1-\frac{\alpha}{2}} S_a)$ → contains 0 : trend

$\underset{\text{popul. line}}{\mu_{yx} = \alpha + \beta x} \neq \underset{\text{regr. line}}{\hat{y} = a + bx} \rightarrow \overset{\text{curve}}{S_{\hat{y}(x)}} = S_{yx}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)S_x^2}}$ wider towards end

$\hat{y} \in (y \pm t_{k,1-\frac{\alpha}{2}} S_{\hat{y}})$

[error of mean when $x = \bar{x}$ at regr. line] $= \frac{S_{yx}}{\sqrt{n}}$

- For individual observation (instead of $\hat{y}$)
  1) variability determined by S_{yx}
  2) " given by uncertainty line of means $S_{\hat{y}}$

$S_{yN} = \sqrt{S_{yx}^2 + S_{\hat{y}}^2}$

↳ $\hat{y} \in (y \pm t_{k,1-\frac{\alpha}{2}} S_{yN})$

Comparison two regression lines

1) test differences slopes just
2) test differences intercepts just — based on t-test
3) test equality whole line

SLOPES $\quad K = n_1 + n_2 - 4 \quad$ **INTERCEPTS**

$t = (b_1 - b_2)/_{S_{b_1-b_2}}$ $\qquad$ $t = (a_1 - a_2)/_{S_{a_1-a_2}}$

$H_0: b_1 = b_2$ $\qquad$ $H_0: a_1 = a_2$

$S_{b_1-b_2} = \sqrt{S_{b_1}^2 + S_{b_2}^2}$ $\quad (n_1 = n_2) \quad$ $S_{a_1-a_2} = \sqrt{S_{a_1}^2 + S_{a_2}^2}$

$(n_1 \neq n_2)$: pooled variance estimator var. around reg. lines

$$S_{yx_p}^2 = \frac{(n_1-2)\, S_{yx_1}^2 + (n_2-2)\, S_{yx_2}^2}{K}$$

$$S_{b_1 b_2} = S_{yx_p}\sqrt{\frac{1}{(n_1-1)S_{x_1}^2} + \frac{1}{(n_2-1)S_{x_2}^2}} \qquad S_{a_1 a_2} = S_{yx_p}\sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{\overline{x_1}^2}{(n_1-1)S_{x_1}^2} + \frac{\overline{x_2}^2}{(n_2-1)S_{x_2}^2}}$$

EQUALITY

H_0: Two regression lines are equal

(1) Calculate regr. lines $\{x_1, y_1\}$; $\{x_2, y_2\}$

(2) Pooled variance estimate both r. lines S_{yx_p} (combined)

(3) Calculate common regression line and S_{yx_s} $\quad (n = n_1 + n_2)$

(4) Estimate variance reduction when separate fits $\quad S_{yx_\Delta}^2 = \dfrac{(n_1+n_2-2)S_{yx_s}^2 - (n_1+n_2-4)S_{yx_p}^2}{2}$

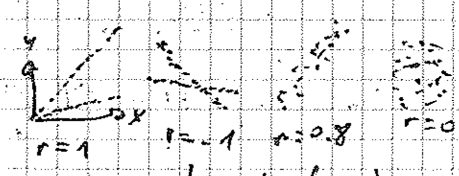
(5) Perform F-test $\quad F = \dfrac{S_{yx_\Delta}^2}{S_{yx_p}^2} \quad m_1 = 2, \; m_2 = n_1 + n_2 - 4$

(6) If $F > F_{m_1, m_2, 1-\alpha}$: reject H_0

Correlation

- Regr. analysis
  - Change of dep. variable following change of indep.
  - Conf. int for predicting dep. var. at fixed value of indep.
- Correlation
  - causality unknown (which dep / indep) not defined
  - describes strength of relationship between two variables
- Pearson product-moment correlation coefficient

$$r = \frac{\Sigma(x-\overline{x})(y-\overline{y})}{\sqrt{\Sigma(x-\overline{x})^2\, \Sigma(y-\overline{y})^2}} \qquad -1 \le r \le 1$$



↳ not slope!

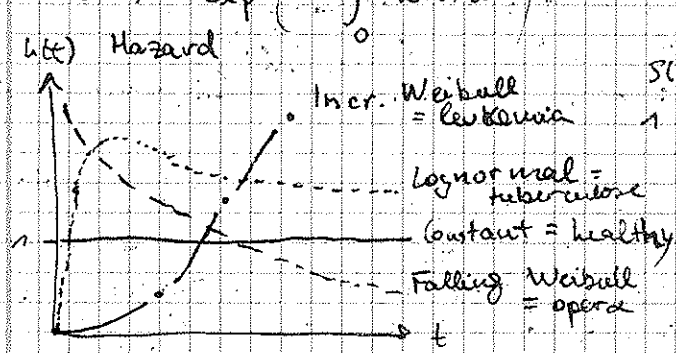
multiv. regr. missing

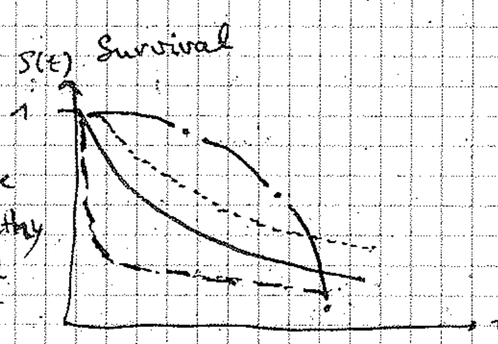
Hazard function $h(t)$; $S(t)$: survival function

$F(t) = 1 - S(t)$ cdf

$f(t) = -S'(t)$ pdf $\quad \rightarrow h(t) = -\dfrac{S'(t)}{S(t)} = \lim_{\Delta t \to 0} P(T \in (t, t+\Delta t) \mid T > t)/\Delta t$ $\quad$ Rate event occurs if not happened until t

$\rightarrow S(t) = \exp\left(-\int_0^t h(t')\,dt'\right) \qquad h(t) = \lambda \rightarrow S(t) = e^{-\lambda t}$





Cox regression

- Estimate $S(t)$ after account for covariates (age, gender, smoker, ...)
  - ↳ not possible Kaplan-Meier

1) Proportional Hazard models (Cox) → Covariates multipl. Hazard; $\log(h) \propto \sum x_i$
2) Accelerated Failure time models (AFT) → Cov multipl Survival; $\log(S) \propto \sum x_i$

M = 1 univariate

Cox: t_i N patients, M covariates x_{ij} ⇒ $S(t)$ wrt baseline group ($x_{ij}=0$ all)

MODEL exp fact hazar.: $h(t; \vec{x}_i) = h_0(t) \cdot e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} \dots e^{\beta_M x_{iM}} = h_0(t) \cdot f(x_i)$ proportional Hazard assumption

- x_{ij} not $x_{ij}(t)$; $x_{ij}(x_k)$
- β_j fit parameter; not $\beta_j(t)$
- $h_0(t)$ Baseline Hazard-function
  - ↳ not explicitly needed (semiparametric model)
  - ↳ robust against estimated Δ model

Goal: find best β_j of model to fit data; get HR

$$\frac{h(t; \vec{x}_i)}{\sum_{\substack{k \in n_i \\ t_k \geq t_i}} h(t, \vec{x}_k)} = L_i = \frac{\cancel{h_0(t_i)} \prod_j^M \exp(\beta_j x_{ij})}{\sum_{\substack{k \in n_i \\ t_k \geq t_i}} \cancel{h_0(t)} \prod_j^M \exp(\beta_j x_{kj})} \rightarrow L(\vec{\beta}) = \prod_i^N L_i^{\delta_i}$$

h_0 disappears, not

$\delta_i = 0$ otherw $\delta_i = 1$

- censored ignored BUT count for sum other L_i
- current t_i irrelevant, just order n_i (idem to wampar rank sum)

- $LL(\vec{\beta})$ → easier to maximize; Newton-Raphson iterative $\frac{\partial LL}{\partial \beta}$
- $\mathrm{Cov}(\vec{\beta}) = -I^{-1}$ (2nd deriv. of LL)
- $\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{\mathrm{Cov}(\hat{\beta})_{jj}}$ → Test Wald if $\hat{\beta}_j \neq 0$ or Likel. Quotient Test = LogLik Diff Test
  - ↳ $W_j = (\beta_j / s_j)^2 \sim \chi^2_1$ → p-value
  - ↳ $LL - LL\{\beta_j = 0\} \sim \chi^2$ → p-value
  - $W_j > \chi^2_{1;0.95}$ ↳ survival signif. affected by covariate x_j

Hazard ratio (two groups w. diff. covariates)

$HR_{ij} = \exp(\beta_j) = h(t; x_{ij}=1) / h(t; x_{ij}=0)$

$HR_i = h(t; \vec{x}_i^*) / h(t; \vec{x}_i) = \exp(\beta_1 (x_{i1}^* - x_{i1})) \dots \exp(\beta_M (x_{iM}^* - x_{iM}))$

↳ how much risk increase if covar increased by 1

↳ factor to multiply treated group wrt to control, rest constant

HR_j: < 1 longer survival; > 1 shorter survival; = 1 no diff } wrt control group; ($W_j > W_{crit}$)

Logistic regression

binary criterion y (probab. (TCP)) ; covariates x_i (dose, age ...) predictors

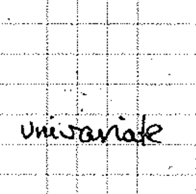
↳ $P(y=1) \in [0,1]$

univariate $P(y=1) = \dfrac{\exp(b_0 + b_1 x)}{1 + \exp(b_0 + b_1 x)}$ b_0: constant, b_1: slope, x: dose $x_{50} = -\frac{b_0}{b_1} \rightarrow y(x_{50}) = 0.5$

multivariate $P(y=1) = \dfrac{\exp(\vec{b} \cdot \vec{x})}{1 + \exp(\vec{b} \cdot \vec{x})}$ $\vec{b} = (b_0, b_1, \dots b_M)$ y_i N points $= \{0, 1\}$

$\vec{x} = (1, x_1, \dots x_M)$ x_j M predictors

" $L_i(\vec{b})$



Goal: find optimum $\vec{b}$ to fit data; get x_{50}

- $L_i(\vec{b}) = e^{\vec{b} \vec{x}_i} / (1 + e^{\vec{b} \vec{x}_i}) \rightarrow L(\vec{b}) := P(y_1 \dots y_N) = \prod_{i=1}^N L_i(b)^{y_i} (1 - L_i(b))^{1-y_i}$
- $LL(\vec{b})$ → easier max → NR, iterat
- $\mathrm{Cov}(\vec{\beta})$ ... 2nd deriv. LL → $\hat{b}_m \pm z_{\alpha/2} \sqrt{\mathrm{Cov}(\hat{b})_{mm}}$; $\hat{b} \cdot \vec{x} \pm z_{\alpha/2} \sqrt{\vec{x} \cdot \mathrm{Cov}(\hat{b}) \vec{x}^T}$
- H_0: $\hat{b}_m = 0$ → Wald stat $W_m = (\beta_m / s_m)^2 \sim \chi^2_1$ or Likquotient test

Two group comparison

- Two dummy predictors e_i (0/1 group nr), xe_i ($x_i \cdot e_i$); x_i: dose & combine all data
  1) log regr. $\{x_i; e_i; xe_i\}$; If $xe_i = 0 \rightarrow b_1|_1 = b_1|_2$; otherwise $b_1|_1 \neq b_1|_2$; b_0 no info
  2) If $b_1 = b_2$, log regr $\{x_i; e_i\}$ If $e_i \neq 0 \rightarrow b_{01} \neq b_{02}$ ⇒ implies $TCD_{50}|_1 \neq TCD_{50}|_2$

OR 3) Lik. Quotient Test $\{x_i, e_i, xe_i\}$ vs $\{x_i\}$ → signif ≡ regr. line different but don't know which parameter