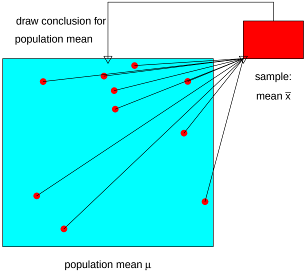
Biostatistics

Probability theory

Burkhardt Seifert & Alois Tschopp

Biostatistics Unit
University of Zurich

Probability theory



Link between sample and population:

- generalise results from sample to the population
- the population is a theoretical - usually infinite - quantity
- imagine one could observe the whole population (e.g. all human beings in the past, present and future) and handle it like a sample
- postulate that we would get **"true"** (population-) **characteristics**:

  probability ($\approx$ relative frequency; %): P
  expectation ($\approx$ mean $\bar{x}$): μ
  standard deviation ($\approx$ s): σ
  percentiles

- needed for statistical tests and confidence intervals

Probability theory

Intuitive:

Probability = relative frequency in the population

Formal:

$$\begin{array}{c}\text{Random experiment}\\ \downarrow \\ \text{Events} \\ \downarrow \\ \text{Probabilities}\end{array}$$

Random experiment

An experiment or observation that can be repeated numerous times under the same condition.

Examples:

- roll a dice
- flip a coin
- diagnose H_1N_1 in a person
- measure the body height of a student
- roll a dice twice
- measure the body height of 245 students

Events

Sample space Ω = set of all possible results of a random experiment

Examples:

Diagnosis $\longrightarrow \Omega = \{$ "sick", "healthy" $\}$
Roll the dice $\longrightarrow \Omega = \{1, 2, 3, 4, 5, 6\}$
Body height $\longrightarrow \Omega = \{x | x > 0\}$

Event A = subset of Ω

Examples:

$A = \{2, 4, 6\}$ even number on the dice
$A = \{1\}$
$A = \{\text{Body height} > 180 \text{ cm}\}$
$A = \{170 \text{ cm} \leq \text{Body height} \leq 180 \text{ cm}\}$
$A = \Omega$ = sure event
$A = \emptyset$ = impossible event

Events

Elementary event ω = element of Ω

Set-theoretic operations:

$A \cap B$	intersection	("and")
$A \cup B$	union	("or")
$A^c, \bar{A}, \neg A$	complement	("not A")

Relation:

$B \subset A$ ("included")

Probability

- $\mathrm{P}(A)$ = relative frequency of a measurable event A in Ω
- Probability can be defined formally based on:

Probability axioms

I. The probability of an event is a non-negative real number:

$$0 \leq \mathrm{P}(A) \qquad \text{for all } A \subseteq \Omega$$

II. Unit measure: the probability that some elementary event in the entire sample space will occur is 1: $\mathrm{P}(\Omega) = 1$

III. Additivity: Any countable sequence of pairwise disjoint events $A_1, A_2, \ldots$ (i.e. $A_i \cap A_j = \emptyset \quad \text{for all } i \neq j$) satisfies:

$$\mathrm{P}(A_1 \cup A_2 \cup \cdots) = \sum_i \mathrm{P}(A_i).$$

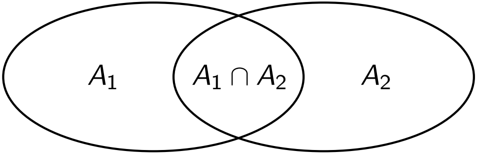
Consequence: $\mathrm{P}(A) \leq 1 \quad \text{for all } A \in \Omega$

Probability

Bonferroni inequality

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} P(A_i)$$

Since:



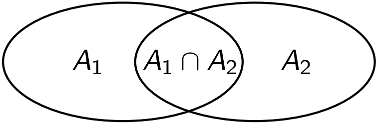
$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

Conditional probability

$\mathrm{P}(A_1|A_2)$ = Probability of some event A_1, given the occurrence of some other event A_2:

$$\mathrm{P}(A_1|A_2) = \frac{\mathrm{P}(A_1 \cap A_2)}{\mathrm{P}(A_2)}$$

$$\begin{aligned} \rightarrow \mathrm{P}(A_1 \cap A_2) &= \mathrm{P}(A_2)\,\mathrm{P}(A_1|A_2) \\ &= \mathrm{P}(A_1)\,\mathrm{P}(A_2|A_1) \end{aligned}$$



Bayes' theorem

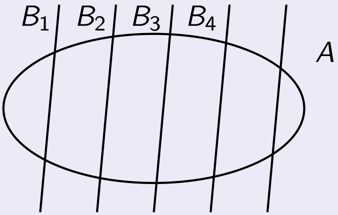
$$\mathrm{P}(A|B) = \frac{\mathrm{P}(B|A)\,\mathrm{P}(A)}{\mathrm{P}(B)}$$

Conditional probability

Law of total probability

Let $\{B_i : i = 1, 2, 3, \dots\}$ be a partition of Ω
(i.e. $B_i \cap B_j = \emptyset$ for all i, j and $B_1 \cup B_2 \cup \cdots \cup B_n = \Omega$), then:

$$P(A) = \sum_i P(A \cap B_i)$$



or, alternatively,

$$P(A) = \sum_i P(A|B_i)\,P(B_i)$$

Conditional probability

Definition: Independence

Two events A and B are (statistically) **independent** if and only if

$$P(A \cap B) = P(A)\,P(B) \quad \text{or} \quad P(B|A) = P(B)$$

Independence:

- formal simplification
- application of many mathematical laws

Examples:

- If a dice is rolled three times, the events of getting each time a 6 are independent:

$$P(\text{three times 6}) = \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216} = 0.0046$$

- If a dice is rolled three times getting at least once a 6:

$$P(\text{at least one 6}) = 1 - P(\text{no 6}) = 1 - \left(\frac{5}{6}\right)^3 = \frac{91}{216} = 0.42$$

Random variable X

Function that maps an elementary event in the sample space to a real number (Result of a random experiment).

Examples:

1. Roll the dice: Every elementary event is mapped to one of the numbers 1, 2, 3, 4, 5, 6.
   ("discrete random variable")
2. Body height: The result is a real number.
   ("continuous random variable")

The observed value $(X = x)$ is called **realisation**.

Random variable X

Definition: Sample

n realisations of a random variable X of interest: $x_1, \ldots, x_n$.

Events of interest and their probabilities:

$$\mathrm{P}(5 < X < 6), \quad \mathrm{P}(X \le c), \quad \mathrm{P}(a \le X \le b),$$
$$\mathrm{P}(X = x_i), \quad \text{if } X \text{ discrete}$$

Example: Flip a coin

- possible realizations $X = 0$ (heads), $X = 1$ (tails)
- sample $n = 2$
- distribution of number of "tails"
- possible samples x_1, x_2: 00 01 10 11

$$\begin{aligned}
\mathrm{P}(X_1 + X_2 = 1) &= \mathrm{P}(X_1 + X_2 = 1 | X_1 = 0)\,\mathrm{P}(X_1 = 0) \\
&\quad + \mathrm{P}(X_1 + X_2 = 1 | X_1 = 1)\,\mathrm{P}(X_1 = 1) \\
&= \mathrm{P}(X_2 = 1)\,\mathrm{P}(X_1 = 0) + \mathrm{P}(X_2 = 0)\,\mathrm{P}(X_1 = 1) \\
&= \frac{1}{4} + \frac{1}{4} = \frac{1}{2}
\end{aligned}$$

Binomial distribution

- sequence of n independent yes/no (1/0) experiments

$$
\begin{aligned}
&\mathrm{P}(X_i = 1) = p \\
&K = \sum_{i=1}^{n} X_i
\end{aligned}
$$

- all permutations of $x_1, \ldots, x_n$ with $K = k$ have the same probability

$$p^k (1-p)^{n-k}$$

- number of possible permutations with exactly k successes out of n known from combinatorics:

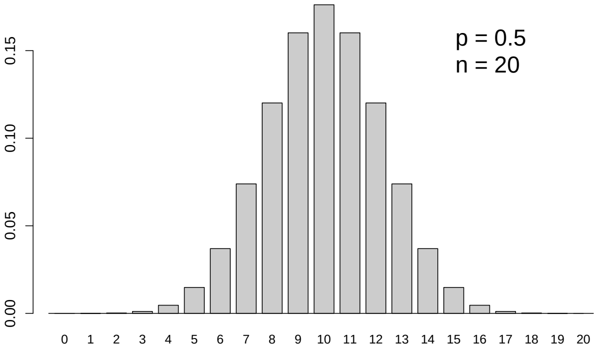
  binomial coefficient "n choose k"

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial distribution

- probability mass function

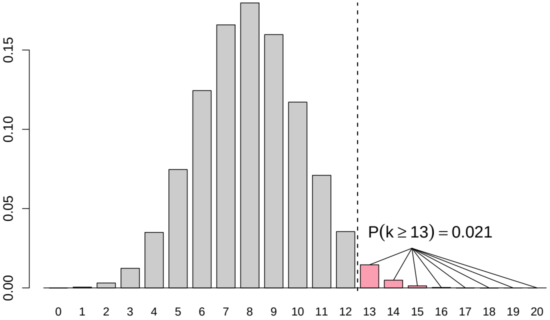
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \leq k \leq n$$



Example: mean number of recovered patients $\hat{p} = \frac{k}{n}$

A total of $n = 20$ patients are examined to test whether or not a new drug yields a probability of recovery higher than $p = 0.4$ (i.e. 40%).

The number k of recovered patients ($k = 0$ to 20 is possible) follows a binomial distribution. If one assumes a probability of $p = 0.4$, the following probability mass distribution for the number of recoveries arises:



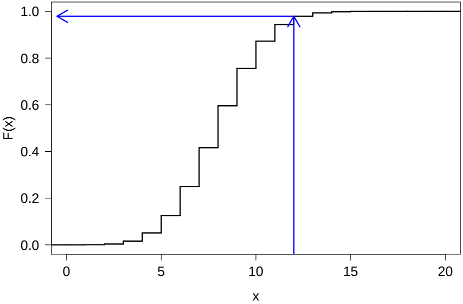
This means that 13 or more recoveries are expected with a probability of only 2.1%.

Cumulative distribution function

Events of the form $X \leq x$ are important as everything can be composed of them with elementary operations

Definition: Cumulative distribution function F of a random variable X

$$F(x) = \mathrm{P}(X \leq x)$$



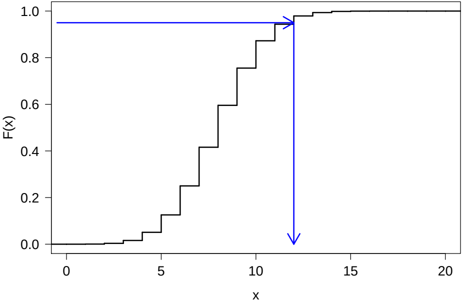
see also: **empirical** (cumulative) distribution function, for data

Cumulative distribution function

Properties of F:

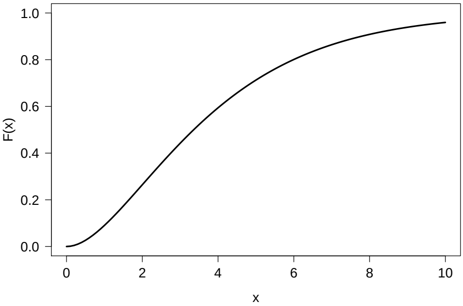
1. $F(-\infty) = 0, \qquad F(+\infty) = 1$
2. F monotone increasing
3. $\mathrm{P}(a < X \leq b) = F(b) - F(a)$

Percentiles of distributions are important for statistical tests.



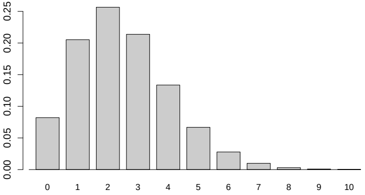
Cumulative distribution function

Continuous random variable (χ^2_4): F continuous

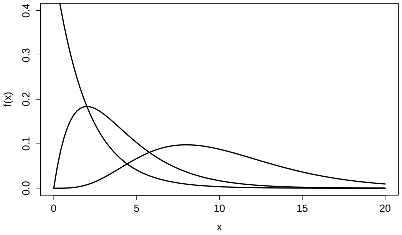


Definition: **Probability density** f

a) discrete variable: $f(x_i) = \mathrm{P}(X = x_i)$



b) continuous variable: $f(x) = F'(x)$



Analogy: histogram

Probability density

Properties:

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(t)dt = 1$
3. $\mathrm{P}(a < X \leq b) = F(b) - F(a) = \int_a^b f(t)dt$
4. $f(t)dt \approx \mathrm{P}(t < X \leq t + \Delta t)$

(stochastic) **independence** of X and Y

$\iff f_{XY}(x, y) = f_X(x)\, f_Y(y)$

(Population-) **Characteristics** of a cumulative distribution function F or random variable X, respectively

expectation $\mu = \mathsf{E}[X] = \int_{-\infty}^{\infty} x f(x) dx$

variance $\sigma^2 = \mathsf{E}\left[(X - \mathsf{E}[X])^2\right] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$

standard deviation $\sigma = \sqrt{\mathsf{E}\left[(X - \mathsf{E}[X])^2\right]}$

alpha-percentile $\mathbf{x}_{\boldsymbol{\alpha}}$ $F(x_\alpha) = \alpha$

If discrete: $\int \longrightarrow$ sums

$$\mu = \sum_{i=1}^{n} x_i \, \mathsf{P}(X = x_i)$$

sample characteristics = statistical estimates for population characteristics

Properties

1. Additivity of expectation:
$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y]$$
2. Non-additivity of variance:
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$
If X, Y are uncorrelated (i.e. $\rho = 0$) $\longrightarrow$ variance is additive
3. X, Y independent $\longrightarrow$ X, Y uncorrelated
   "$\longleftarrow$" not true (but valid for normal distributions)
4. $\mathrm{Var}(c\,X) = c^2\,\mathrm{Var}(X)$

Important consequence of (2) and (4):

$X_1, \ldots, X_n$ **independent**, identically distributed random variables, variance σ^2. Then:

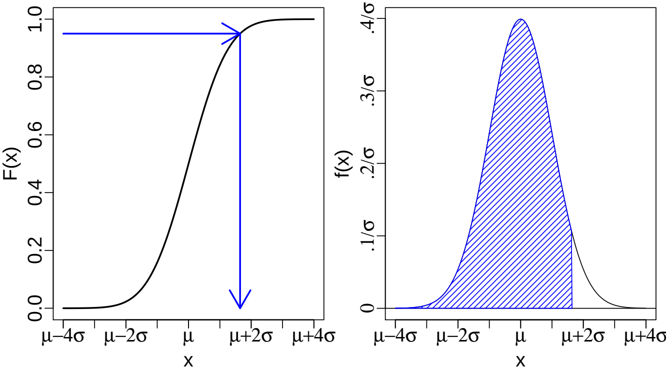
$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \frac{1}{n^2}\,\mathrm{Var}(X_i) = \frac{\sigma^2}{n}$$

$\boxed{\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}}$ "Square Root of n Law"

Important Distributions: Normal distribution $\mathcal{N}(\mu, \sigma^2)$

If $\mu = 0,\ \sigma^2 = 1$: Standard normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Normal distribution $\mathcal{N}(\mu, \sigma^2)$

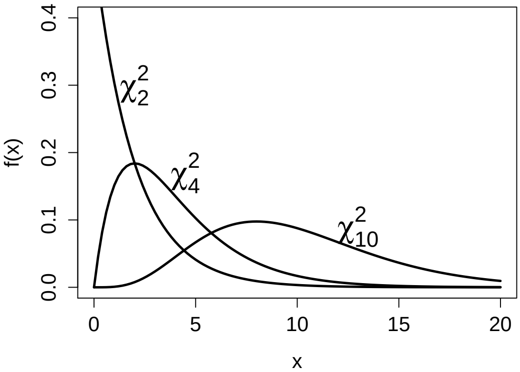
Properties:

- Central Limit Theorem $\longrightarrow$ omnipresent
- symmetric
- simple parameters μ, σ^2
- "light tails"
- assumption for many statistical methods

χ^2–distribution

$Z_1, \ldots, Z_\nu$ independent $\mathcal{N}(0,1)$

$$\chi^2_\nu = \sum_{i=1}^{\nu} Z_i^2 \quad \chi^2\text{–distributed with } \nu \text{ degrees of freedom}$$



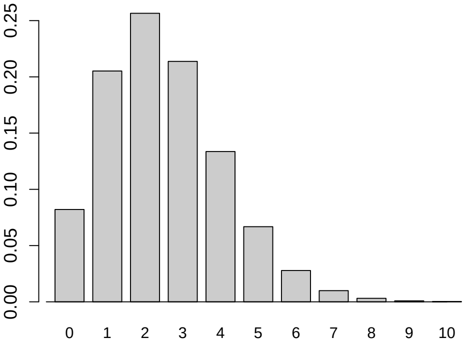
χ^2–distribution

Properties:

- $\mu = \nu, \quad \sigma^2 = 2\nu$
- $\nu = 2$: exponential distribution
- physics: modelling energy or the like
- statistics: important distribution for tests (contingency tables, goodness–of–fit)
- model for the variance of normally distributed data

Poisson–distribution (discrete)

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



Poisson–distribution

Properties:

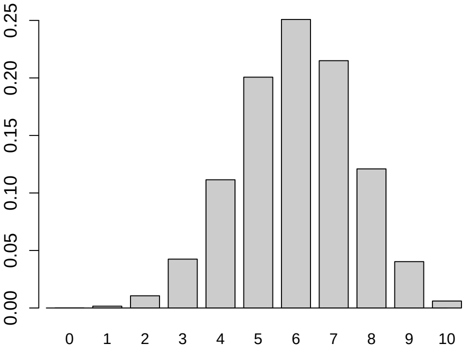
- $\mu = \lambda, \quad \sigma^2 = \lambda$
- modelling of rare events (radioactive decay, crime rate)

Number of crimes per day	full moon days Obs	full moon days Exp	new moon days Obs	new moon days Exp
0	40	45.2	114	112.8
1	64	63.1	56	56.4
2	56	44.3	11	14.1
3	19	20.7	4	2.4
4	1	7.1	1	0.3
5	2	2.0	0	0.0
6	0	0.5	0	0.0
7	0	0.1	0	0.0
8	0	0.0	0	0.0
9	1	0.0	0	0.0
Total number of days	183	183.0	186	186.0

Binomial–distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad 0 \le k \le n$$

- $\mu = n\,p, \quad \sigma^2 = n\,p\,(1-p)$



Law of Large Numbers ($n \longrightarrow \infty$)

(Always: independent random variables)

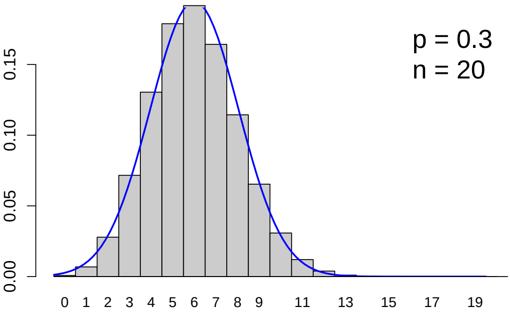
- Law of large numbers (**LLN**)

$$\bar{X} \longrightarrow \mu \qquad \text{in "probability"}$$

Central limit theorem

- Central limit theorem (**CLT**)

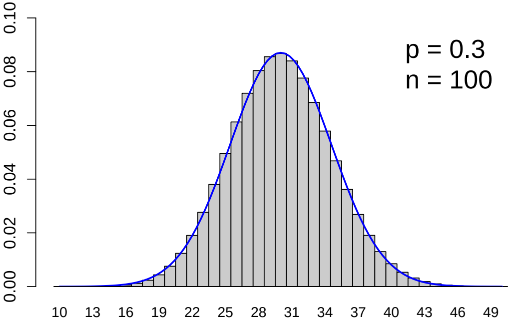
$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \longrightarrow \mathcal{N}(0, 1) \quad \text{"in distribution"}$$



Central limit theorem

- Central limit theorem (**CLT**)

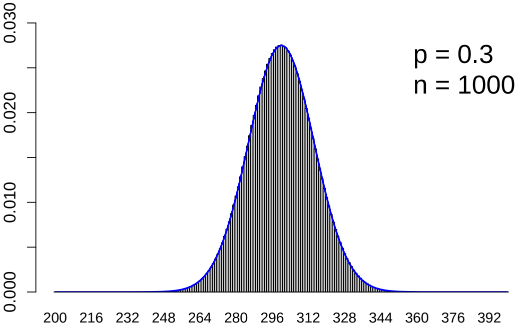
$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \longrightarrow \mathcal{N}(0,1) \quad \text{``in distribution''}$$



Central limit theorem

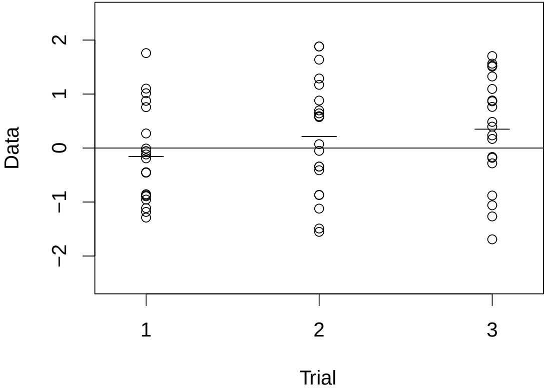
- Central limit theorem (**CLT**)

$$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \longrightarrow \mathcal{N}(0,1) \quad \text{“in distribution”}$$



Estimation procedures

- Sample characteristics such as e.g. the mean are random, they vary.



An estimator is a sample characteristic (statistic) which aims at approximating a population characteristic (parameter).

Estimation procedures

Studies cost money, time; data are often not available at will

- aim is a statistically efficient use of data
- use of "good" estimators for quantities of interest

Let $\hat{\theta}$ be an estimator for a parameter θ, based on a sample $x_1, \ldots, x_n$

Minimal requirement: Validity of LLN and CLT:

- $\hat{\theta} \longrightarrow \theta$ for $n \longrightarrow \infty$ in probability
  "$\hat{\theta}$ consistent"
- $\hat{\theta}$ for large n approximately normally distributed

Usually fulfilled!

Quantitatively: error $(\hat{\theta} - \theta)$ should be small!

Criterion 1: **Unbiasedness of $\hat{\theta}$**

$$\mathrm{E}\left[\hat{\theta} - \theta\right] = 0 \quad \text{or} \quad \mathrm{E}\left[\hat{\theta}\right] = \theta$$

i.e. on average you are right

If not: $\mathrm{E}\left[\hat{\theta} - \theta\right] = \text{bias of } \hat{\theta}$

Examples:

- n machines that can independently fail
- failure statistic, per day: $X_i = 0$, no failure
  $X_i = 1$, failure

Estimator $\hat{p}$ for failure probability p: $\hat{p} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\mathrm{E}[\hat{p}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] = p$$

Thus: no bias

Criterion 1: **Unbiasedness of $\hat{\theta}$**

- With two machines: probability that both fail:

  naïve: $\hat{p}^2 = \bar{x}^2$,

  but:

$$\begin{aligned} E[\bar{x}^2] &= \mathrm{Var}(\bar{x}) + (E[\bar{x}])^2 \\ &= \frac{1}{n}\mathrm{Var}(X_1) + (E[X_1])^2 \\ &= \frac{p(1-p)}{n} + p^2 \end{aligned}$$

$\text{Bias} = \dfrac{p(1-p)}{n} \neq 0$ for finite n

Criterion 2: **Minimum Variance Estimation**

Create an unbiased estimator $\hat{\theta}$ such that

$$\text{Var}(\hat{\theta}) = \text{ minimal}$$

Unbiased estimators with minimal variance are good.

Accuracy of the mean

n independent observations with variance σ^2

$$\longrightarrow \mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}$$

Standard error of the mean

$$\sigma_{\bar{x}} = SEM = SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Accuracy of fractions p

Example:

$n = 80$ individuals surveyed about asthma
$k = 7$ thereof are asthmatics

$$\hat{p} = \frac{k}{n} = 0.088 \quad \text{estimated prevalence}$$

How accurate is p determined? Binomial distribution!

$$\sigma^2_{\hat{p}} = \mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

$\longrightarrow s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

In the example with $\hat{p} = 0.088$: $s_{\hat{p}} = 0.032$
If $\hat{p} = 0.5$: $s_{\hat{p}} = 0.056$

This means, middle frequencies are more dispersed than extreme ones.

Accuracy of estimators

Variation of an estimator = standard error SE

- can be obtained from computer printouts
- conveys an impression about the accuracy of the statistic

Maximum likelihood estimation

Unbiased estimators with minimal variance do not always exist.

General alternative:

Maximum likelihood estimation

- general, successful estimation principle
- algorithmic procedures exist
- theory thereto is complex
- **assumption of a distribution model** $f(x, \theta)$ for data is necessary to estimate parameter θ

Maximum likelihood estimation: Idea

1. Given data $x_1, \dots, x_n$ (independent)
2. Probability to observe $x_1 \cong f(x_1, \theta)$
3. Probability to observe $x_1, \dots, x_n$:
   product, since random variables x_i are independent

   $$L(\theta) = f(x_1, \theta) \times f(x_2, \theta) \times \cdots \times f(x_n, \theta)$$

   $L(\theta)$ is called **likelihood function**, θ is the argument,
   x_i are given data
4. For which value θ is the agreement with the data $x_1, \dots, x_n$ maximal?
5. Determine θ such that $L(\theta)$ is maximal
   ("maximum likelihood estimator" for θ)
6. Mathematically often easier to maximise $\log L(\theta)$.

ML estimation: Example

Flip a coin 10 times, observe "heads" as result 4 times.
How large is the probability to throw "heads"?

Heuristically: probability $\hat{p} = 0.4$
Probability distribution for 4 times "heads" is according to binomial distribution proportional to

$$L(p) = p^4(1-p)^6$$

Maximisation:

$$L'(p) = 4p^3(1-p)^6 - 6p^4(1-p)^5 = 0$$

$$\longrightarrow 4(1-p) = 6p$$

$$\longrightarrow \text{maximum likelihood estimator: } \hat{p}_{ML} = 0.4$$

Thus?
In this example plausible.

ML estimation: Example

Sample $x_1, \ldots, x_n$ originating from normal distribution $\mathcal{N}(\mu, \sigma^2)$

Maximum likelihood estimators for μ, σ^2 ?

$$L(\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$\begin{aligned}
\log L(\mu, \sigma^2) &= -n \log\sqrt{2\pi} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(x_i-\mu)^2 \\
\frac{\partial \log L}{\partial \mu} &= \frac{1}{\sigma^2}\sum(x_i-\mu) = 0 \\
\longrightarrow \hat{\mu}_{ML} &= \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i
\end{aligned}$$

as known and to be expected.

ML estimation: Example

Maximum likelihood estimators for σ^2:

$$
\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 = 0
$$

$$
\longrightarrow \hat{\sigma}_{ML} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
$$

Attention: $\hat{\sigma}^2_{ML} = \frac{n-1}{n} s^2$!

s^2 is unbiased, but $\hat{\sigma}^2_{ML}$ is not.

Properties of ML estimators

Good properties:

1. ML method is never worse than any other method (for $n \longrightarrow \infty$).
2. ML method is applicable for most (even complex) problems.
3. ML estimators are consistent.
4. If $\hat{\theta}$ is a ML estimator for θ, then $h(\hat{\theta})$ is a ML estimator for $h(\theta)$.
5. ML estimators are approximately normally distributed.

And the bad news?

1. Many properties only asymptotically valid ($n \longrightarrow \infty$)
2. One needs a parametric probability model for data,
   e.g. normal distribution

Where does chance come from?

- **Random sample**: "Drawing" of individuals from the population
  - chance is not a measurement error but inter-individual variation
- Representativeness (generalisability)
  - volunteers are not representative for the population of all patients
  - patients from university hospitals are not representative
- **Randomisation**: random splitting in two or more groups
  - chance arises from the computer (pseudo random) or physical random process

Where does chance come from?

- Independence
- succession of patients in the hospital is random
- violated with patients from a single family,
- or, when doctors have an effect on the result (cluster)

- With repeated measurements for the same patient (pre-post comparisons, several locations, e.g. arteries or longitudinal studies) the patient is the observational unit and not the single measurement.