Experimental Design and Data Analysis
Lectures 0 and 1

Eduard Belitser

VU Amsterdam

Lecture Overview

1. course organization
2. experimental design
3. recap probability theory and basic statistics
4. recap: examples in R

Course organisation

- Prerequisites: basic statistics course (e.g., Statistical Methods), basic probability, R knowledge.
- The first 1.5 lectures is a recap of what you are supposed to know. Test your prerequisite knowledge: exam (+ its solution) is available on canvas.
- All relevant information is on canvas: schedule, lecture slides, assignments (in due time), R manual(s) and suggestions additional literature.
- R is an open software, widely adopted in the academic community, it is a programming language (object oriented), a statistical package.
- RStudio is a powerful user interface for R.

Experimental design

What is experimental design?

- Experiments are performed with varied preconditions represented by ind. variables, also referred to as input variables or predictor variables.
- The change in predictors is hypothesized to result in a change in one or more dep. variables, also referred to as output or response variables.
- The experimental design may also identify control variables that must be held constant to prevent external factors from affecting the results.
- Experimental design involves also planning the experiment under statistically optimal conditions given the constraints of available resources.
- Main concerns in experimental design: validity, reliability, replicability, achieving appropriate levels of statistical power and sensitivity.
- Ronald Fisher: *The Arrangement of Field Experiments* (1926) and *The Design of Experiments* (1935).

Experimental design, randomization

- Statistics allows to generalize from data to a true state of nature, but statistical inference requires assumptions and mathematical modeling.
- The data should be obtained by a carefully designed (chance) experiment (or at least it must be possible to think about the data in this way).
- Any good design involves a chance element: “experimental units” are assigned to “treatments” by chance, or by randomization. The purpose is to exclude other possible explanations of an observed difference.
- We need probability to quantify the randomization. In practice, randomization is implemented with a random number generator. In R:

```
> x=rep(c("A","B"),each=5); x
 [1] "A" "A" "A" "A" "A" "B" "B" "B" "B" "B"
> sample(x)   # create a sequence of 5 A’s and 5 B’s in random order
 [1] "A" "B" "A" "B" "B" "A" "B" "A" "A" "B"
> rbinom(10,1,0.5)    # toss a fair coin 10 times
[1] 1 0 1 1 1 0 1 0 0 0
> rbinom(10,1,0.5)    # again toss a fair coin 10 times
[1] 1 0 0 0 0 1 0 1 1 0
> rbinom(5,1,0.8)  # toss a biased coin (success probability=0.8) 5 times
[1] 1 1 0 1 1
```

Examples, observational studies

EXAMPLE To compare two fertilisers we prepare 20 plots of land, apply the first fertilisers to 10 randomly chosen plots and the second one to the remaining plots. We plant a crop and measure the total yield from each plot.

EXAMPLE To compare two web designs we randomly select 50 subjects and measure the time needed to find some information. All 50 subjects perform this task with both designs, but for each subject the order of the two designs is based on tossing a coin.

EXAMPLE If an experiment involves subjects, then it could be wrong to assign "task A" to the first 10 subjects who arrive and "task B" to the last 10. (There may be a reason for arriving early.) Instead assign the tasks at random. Then an observed difference is due to the task (or chance).

Data obtained by registering an ongoing phenomenon, without randomization or applying other controls, is called observational.

EXAMPLE The incidence of lung cancer among 500 smokers is observed to be higher than among 500 non-smokers. Does this finding generalize to the full population? Does this show that smoking causes lung cancer?

Recap probability theory and basic statistics

(prerequisite for this course, if needed consult *Elementary Statistics*, by Mario Triola)

Probability distributions: continuous, discrete

- A probability distribution P determines the probability of different outcomes of a random variable.
- Probability distributions for:
  - discrete random variables which have finite or countable sets of possible outcome values (e.g., dice, coins, birthdays);
  - continuous random variables which have infinite sets of possible outcome values (e.g., temperature, length).
- The corresponding probability distributions: continuous, discrete.

Remark. Actually, there are distributions which neither continuous nor discrete.

Probability density functions

Examples of the probability density p of some continuous distributions (realised also in R with some default parameter values):

- normal distribution `norm` with parameters μ `mean`=0 and σ `sd`=1

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \mathbb{R}.$$

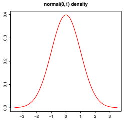
- exponential distribution `exp` with parameter λ (`lambda`=1)

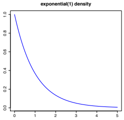
$$p(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

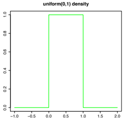
- uniform distribution `unif` with parameters minimum (`min`=a) and maximum (`max`=b) of the support interval

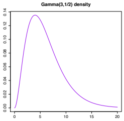
$$p(x) = \frac{1}{b-a}, \qquad a \leq x \leq b.$$

- Gamma distribution `gamma` with parameters shape `shape` and rate `rate`=1.









Probabilities of events – continuous distribution

If a random variable X has a distribution with the density $p(x)$, then

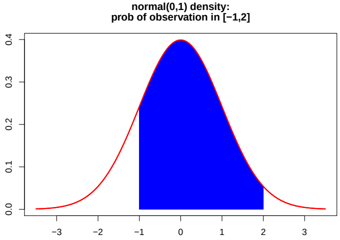
$$P\big(X \in I\big) = \int_I p(x)dx \qquad \text{for any interval} \quad I \subseteq \mathbb{R}.$$

In other words, the probability to have an outcome in some interval I is the area under the density function $p(x)$ over that interval.

Example. For $X \sim N(0,1)$,

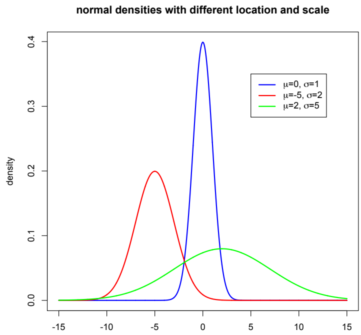
$$P(-1 \le X \le 2) = P\big(X \in [-1,2]\big)$$
$$= \int_{-1}^{2} p(x)dx = \int_{-1}^{2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 0.82.$$

In events for continuous distributions:
$<$ or $\le$ ($>$ or $\ge$) does not matter.



Location and scale, normal density

Two important characteristics of a population are location (or mean) μ and scale (or standard deviation) σ.



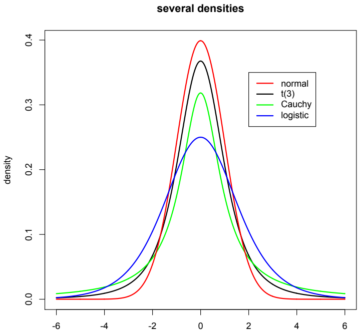
The normal density curve is given by

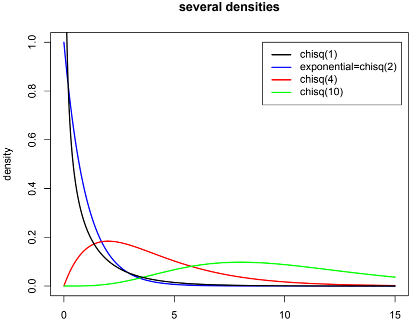
$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

The parameters μ and σ are the location and scale. Normal distributions with different μ and σ are still similar in a way.

Remark. The normal curve is very specific! There are many "bell shaped" curves that are not normal.

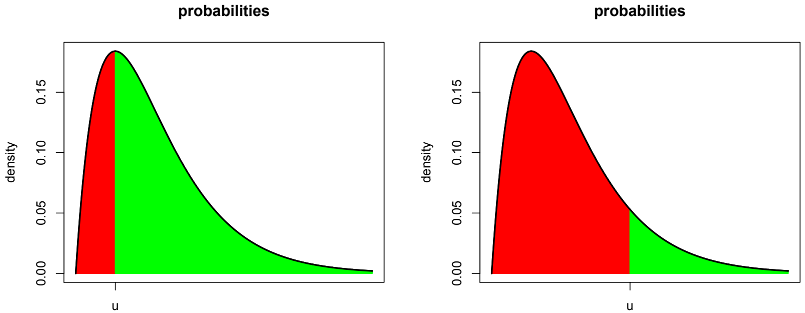
Other symmetric and asymmetric densities





Probabilities and quantiles

If a random variable X is distributed according to a density curve, the probability $P(X \leq u)$ is the (red) area under the density curve left of u. Likewise, $P(X \geq u)$ is the (green) area under the density curve right of u.



For distribution P, the quantile of level $\alpha \in (0, 1)$ is the number q_α such that $P(X \leq q_\alpha) = \alpha$, the upper quantile u_α such that $P(X \geq u_\alpha) = \alpha$.
For the standard normal distribution, the quantile and upper quantile are usually denoted by ξ_α and z_α.

Probability of events – discrete distribution

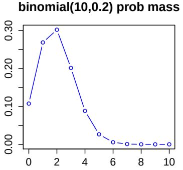
For discrete distributions we have a probability mass function p

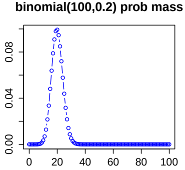
$$p(x) = P(X = x).$$

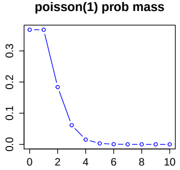
The probability to have an outcome in some set A is the sum

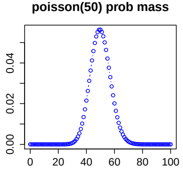
$$P\big(X \in A\big) = \sum_{x \in A} p(x).$$

Examples of discrete distributions are binomial and Poisson.









Probability mass functions for some discrete distributions

Discrete distributions (realised also in R):

- Binomial distribution `binom` with parameters n `size` and p `prob`

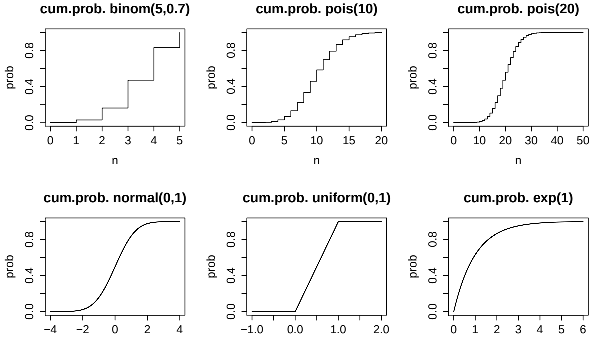
$$p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

- Poisson distribution `pois` with parameter λ `lambda`

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Cumulative distribution/probability function

- The cumulative distribution function (CDF) (sometimes also called cumulative probability function) of a random variable X is $F(u) = P(X \leq u)$= `pdist(u,par)` (continuos and discrete)
- Continuous distr.: $F(u) = \int_{-\infty}^{u} p(x)dx$; discrete: $F(u) = \sum_{x \leq u} p(x)$.
- Any other probability can be computed via $F(u)$, e.g., for any $a \leq b$, $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$.



R-commands for distributions

- in `R`, there is a number of continuous and discrete distributions `dist` with parameters `par`.
- Let $p(x)$ denote the density for continuous distribution and the probability mass function for discrete distribution.
- `ddist(x,par)` computes $p(x)$ (i.e., either density or mass function),
- `pdist(u,par)` computes the CDF $F(u) = P\big(X \leq u\big)$,
- `qdist(a,par)` (`a` $\in [0,1]$) computes the value `q` such that `pdist(q,par)=a`, this is the a-quantile. The α-quantile q_α is such number that $P(X \leq q_\alpha) = \alpha$.
- `rdist(size,par)` yields a random sample from `dist` with parameter `par` of size `size`.

Examples in R

```
> pnorm(2,mean=0,sd=1)-pnorm(-1,mean=0,sd=1) #P(-1<X<2)=P(X<2)-P(X<-1)
[1] 0.8185946
> pnorm(2)-pnorm(-1) # no need to set the default mean=0, sd=1
[1] 0.8185946
> rnorm(4) # generate 4 standard normals
[1]  0.5592590 -0.3570060 -0.7276720  0.8368255
> dbinom(1,size=5,prob=0.2) # this is P(X=1)
[1] 0.4096
> pbinom(1,size=5,prob=0.2) # this is P(X<=1)
[1] 0.73728
> dbinom(0,5,0.2)+dbinom(1,5,0.2) # indeed, P(X<=1)=P(X=0)+P(X=1)
[1] 0.73728
> rpois(3,lambda=5)
[1] 6 7 2
```

Expectation

- The expectation or mean $E(X)$ of a random variable X with probability distribution P is a location parameter of distribution P.
- For discrete random variable: $E(X) = \sum_x xp(x)$.
- For continuous random variable: $E(X) = \int xp(x)dx$.

Examples

Throwing a dice: $E(X) = \sum_x xp(x) = 1 \times \frac{1}{6} + \ldots + 6 \times \frac{1}{6} = 3\frac{1}{2}$.

Normal distribution: $E(X) = \int xp(x)dx = \int x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \ldots = \mu$.

Variance and standard deviation

- The variance of a probability distribution is a scale (or spread) parameter.
- For discrete random variable: $Var(X) = \sum_x (x - E(X))^2 p(x)$.
- For continuous random variable: $Var(X) = \int (x - E(X))^2 p(x) dx$.
- Definition: the standard deviation σ is the square root of the variance $\sigma = \sqrt{Var(X)}$.

Examples

Throwing dice:
$Var(X) = \sum_x (x - 3\frac{1}{2})^2 p(x) = (1 - 3\frac{1}{2})^2 \times \frac{1}{6} + \ldots + (6 - 3\frac{1}{2})^2 \times \frac{1}{6} = 2.92$.

Normal distribution:
$$Var(X) = \int (x - \mu)^2 p(x) dx = \int (x - \mu)^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} dx = \ldots = \sigma^2.$$

Expectation and variance for some distributions

	Expectation	Variance
Uniform(a,b)	$(a+b)/2$	$(b-a)^2/12$
Normal(μ,σ^2)	μ	σ^2
Exponential(λ)	$1/\lambda$	$1/\lambda^2$
Binomial(n,p)	np	$np(1-p)$
Poisson(λ)	λ	λ

Summarizing data and exploring distributions

Population and sample

- A population can be an actual population, e.g., the heights of all men in the Netherlands.
- It can also be the (imaginary) infinite number of outcomes obtained by repeating an experiment over and over, e.g., throwing a dice many times.
- A sample is a set of values (randomly) selected from a population.
- The population has a certain distribution, called the population distribution.
- From the sample we want to gain/extract information about this unknown population distribution.
- This is the main problem of statistics/data analysis.

Types of data summaries

A good summary of a data set shows the relevant information in a data set.

- numerical summaries (of what it estimates/investigates)
  - sample mean (population mean)
  - sample median (population median)
  - sample standard deviation (population standard deviation)
  - sample variance (population variance)
  - sample correlation(s) (population correlation(s))
  - ...
- graphical summaries
  - histogram (estimates probability density or probability mass)
  - boxplot (assess symmetry, range, outliers)
  - scatter plot(s) (assess relations between variables)
  - normal QQ-plot (checks normality)
  - empirical distribution function (cumulative prob. function)
  - ...

Data summaries and some useful R -commands

- Densities, probabilities and quantiles of many distributions can be computed in R. Commands in R: `dnorm(u,par)`, `pnorm(q,par)`, `qnorm(a,par)`, `rnorm(size,par)`, etc.
- Numerical summaries: sample mean, sample variance, sample median, sample standard deviation, sample α-quantile, etc. Commands in R: `mean(x)`, `var(x)`, `med(x)`, `sd(x)`, `quantile(x,a)`, `summary(x)`, `range(x)`, etc.
- Graphical summaries: histogram, boxplot, (normal) QQ-plot, scatter plot(s), empirical distribution function (cumulative histogram), etc. Commands in R: `hist(x)`, `boxplot(x)`, `qqnorm(x)`, `plot(x,y)`, `plot(ecdf(x))`, etc.

Study Assignment 0.

The boxplot of a sample is a box with `whiskers` and (possibly) `extremes`, from which you can see the scale of the data, its symmetry, whether there are extreems (outliers).

Complement graphical summaries with numerical summaries and vice versa.

Some numerical summaries: reminder

sample size		n
location	*mean*	$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$
	median	$\mathrm{med}(x) = \begin{cases} x_{((n+1)/2)}, & \text{if } n \text{ odd} \\ (x_{(n/2)} + x_{(n/2+1)})/2, & \text{if } n \text{ even} \end{cases}$
scale	*variance*	$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
	standard deviation	$s = \sqrt{s^2}$

Here $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ is the ordered sample.

Interpretation of location measures:

- mean – average value
- median – middle value in sorted values

Interpretation of scale measures:

- variance – average squared deviation from mean
- standard deviation – square root of variance

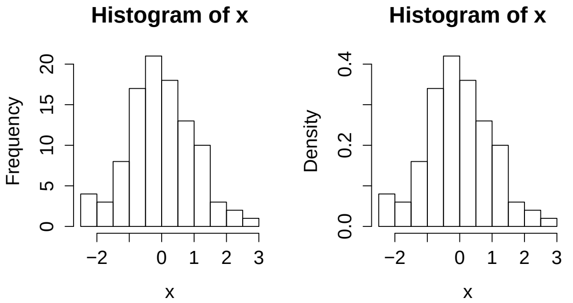
Histogram

The histogram of a sample of observed values $x_1, x_2, \ldots, x_N$ is a barplot, where the area of the bar over a cell (also called bin) C corresponds to the fraction

$$\frac{\text{number of observations in cell } C}{\text{sample size}} = \frac{\#\{1 \le i \le N : x_i \in C\}}{N}.$$

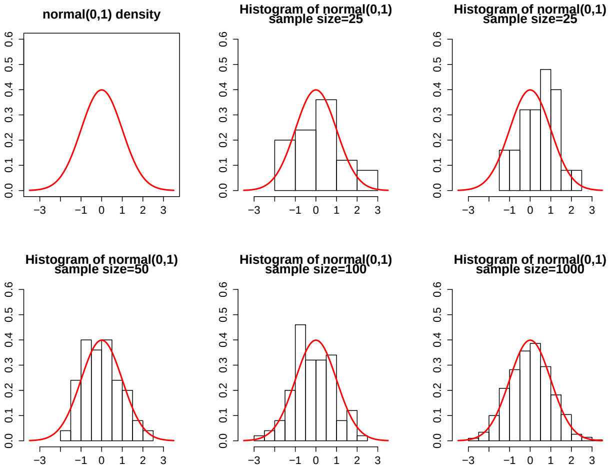
```
> x=rnorm(100); par(mfrow=c(1,2)) # two plots next to each other
> hist(x)} # frequencies on y-axis
> hist(x,prob=T) # probabilities on y-axis
```

Why are the Y-axes different?



Histogram versus density (1)

The histogram of a sample (from the true density p) varies around p. The smaller the sample, the bigger this variation.

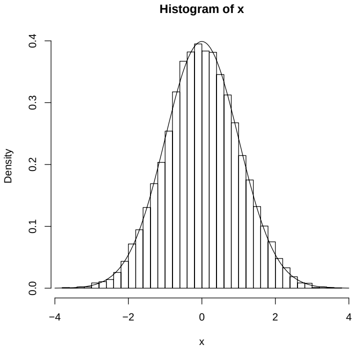


Histogram versus density (2)

For continuous distributions, the true population density can be seen as the smoothed (or limiting as sample size $\to \infty$) histogram of the population values.

The resemblance between the true normal(0,1) density and the histogram of a sample of size 10000.

You can think of the population here as consisting of infinitely many values.



Covariance, correlation, sample correlation

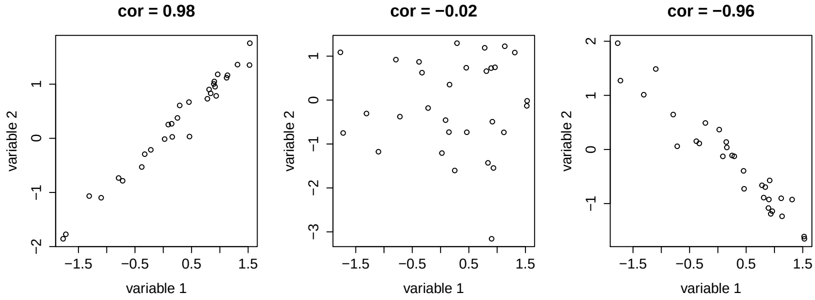
- The covariance between two random variables X and Y is $\mathrm{Cov}(X,Y) = E\Big[(X-EX)(Y-EY)\Big]$.
- The correlation between two variables X and Y quantifies the linear relation between them:

$$\rho = \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{E\Big[(X-EX)(Y-EY)\Big]}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}.$$

- In practice, the distribution of (X,Y) is almost never known. Instead, one has a sample $(x_1,y_1),\dots,(x_N,y_N)$ from the distributions of (X,Y).
- Then we can compute the sample covariance and sample correlation

$$\hat{c}_{x,y} = \frac{1}{n}\sum_{i=1}^{N}(X_i-\bar{X}_N)(Y_i-\bar{Y}_N), \quad \hat{\rho} = \frac{\sum_{i=1}^{N}(X_i-\bar{X}_N)(Y_i-\bar{Y}_N)}{\sqrt{\sum_{i=1}^{N}(X_i-\bar{X}_N)^2)\sum_{i=1}^{N}(Y_i-\bar{Y}_N)^2}}.$$

Correlation and scatter plot (1)

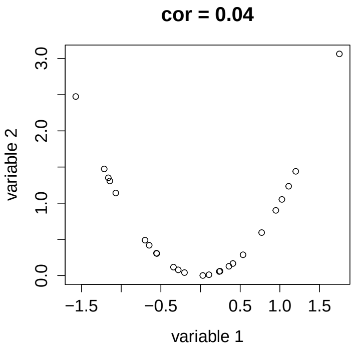


Correlation values:

- ≈ 1: linear relation (straight line) with positive slope (if =1, then perfect linear relation)
- ≈ −1: linear relation (straight line) with negative slope
- ≈ 0: no linear relation (but maybe some other relation?)

Correlation and scatter plot (2)

Example of two variables that have correlation close to 0, but a clear relation:

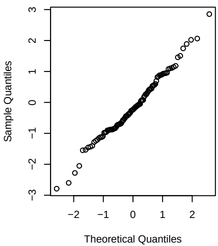


Such a figure is called a scatter plot of `variable 1` (horizontal) versus `variable 2` (vertical).

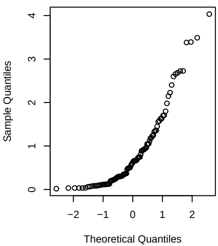
QQ-plots

- A QQ-plot can reveal whether data (approximately) follows a certain distribution P (often this is the normal distribution: `qqnorm(x)`).
- It plots the ordered data $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(N)}$ versus the quantiles $q_{1/N}, q_{2/N}, \ldots, q_{N/N}$ of the distr. P, i.e., $P(X \leq q_\alpha) = \alpha$ for $X \sim P$. (Actually, R uses the quantiles at $\frac{i}{n+1}$ (or another slight adaptation) rather than at $\frac{i}{n}$)
- If $X_i \sim P$, then approx. a fraction $\frac{i}{N}$ of the population should be smaller than the $\frac{i}{N}$-quantile $q_{i/N}$, i.e., the plot points should follow a straight line.
- If the points are approximately on a straight line, then the data can be assumed to be sampled from P, possibly with different location and scale.

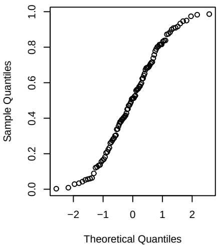
Normal Q–Q Plot



Normal Q–Q Plot



Normal Q–Q Plot



Shapiro Wilk test for normality

Setting: A sample $X_1, \ldots, X_n$ from an unkown distribution P.

Hypotheses: H_0 : P *is a normal distribution* versus H_1 : P *is not a normal*

Test statistic: with certain constants $a_1, \ldots, a_n$,

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

Distribution of W under H_0: known, but complicated to write down. H_0 is rejected for "small" values of W. It is always the left-sided test.

In R: `shapiro.test(x)`

Note: this test complements the graphical check by a normal QQ-plot.

Example — expensescrime (1)

The data `expensescrime` *were obtained to determine factors related to state expenditures on fighting criminality (courts, police, etc.). The variables are:* `state` *(indicating the state in the USA),* `expend` *(state expenditures on fighting criminality in $1000),* `bad` *(number of persons under judicial supervision),* `crime` *(crime rate per 100000),* `lawyers` *(number of lawyers in the state),* `employ` *(number of persons employed in the state) and* `pop` *(population of the state in 1000).*

```
> expensescrime=read.table("expensescrime.txt",header=TRUE)
> head(expensescrime)
  state expend   bad crime lawyers employ   pop
1    AK    360   5.1  5877    1749   2796   525
2    AL    498  34.4  3942    6679  13999  4083
3    AR    219  19.2  3585    3741   7227  2388
4    AZ    728  31.3  7116    7535  14755  3386
5    CA   6539 336.2  6518   82001 118149 27663
6    CO    602  25.7  6919   11174  12556  3296
```

Apart from numerical and graphical summaries of the columns separately, we can consider bivariate summaries to see the relation between pairs of columns.

Example — expensescrime (2)

The correlation between all pairs of variables, excluding the first column:

```
> round(cor(expensescrime[,-1]),3)
        expend   bad crime lawyers employ   pop
expend   1.000 0.834 0.334   0.968  0.977 0.953
bad      0.834 1.000 0.373   0.832  0.871 0.920
crime    0.334 0.373 1.000   0.375  0.311 0.275
lawyers  0.968 0.832 0.375   1.000  0.966 0.934
employ   0.977 0.871 0.311   0.966  1.000 0.971
pop      0.953 0.920 0.275   0.934  0.971 1.000
```

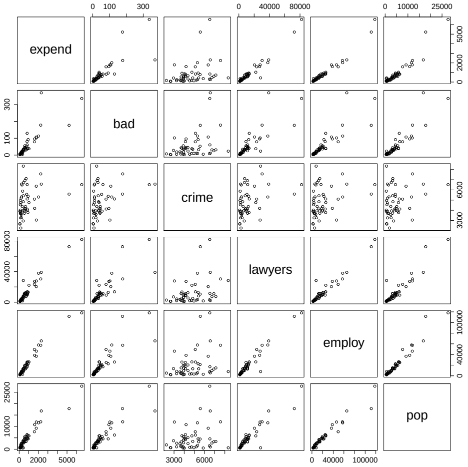
Ingredients of R-code:

- `expensescrime[,-1]` removes column 1 from `expensescrime`,
- `cor(expensescrime[,-1])` produces pairwise correlations between remaining columns,
- `round(cor(expensescrime[,-1]),3)` rounds the numbers to 3 decimals.

Example — expensescrime (3)

The scatter plots of all pairs of variables, excluding the first column:

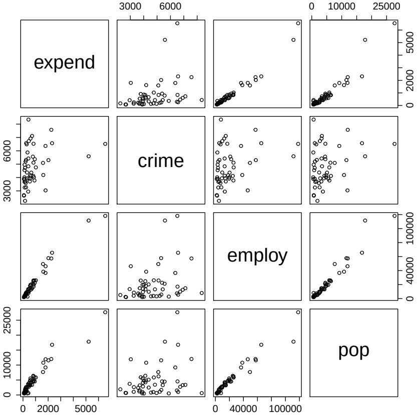
```
> pairs(expensescrime[,-1])
```



Example — expensescrime (4)

The scatter plots of the variables `expend`, `crime`, `employ`, `pop`:

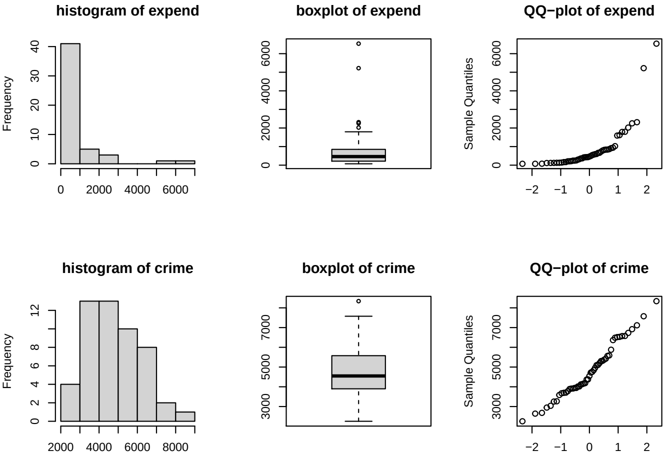
```
> pairs(expensescrime[,c(2,4,6,7)])
```



`expensescrime[,c(2,4,6,7)]` selects columns 2, 4, 6 and 7.

Example — expensescrime (5)

Histograms, boxplot and QQ-plots of the two columns (`expend` and `crime`) of the `expensescrime` data. Column `crime` seems to follow a normal distribution.



Start Lecture 1. Recap basic stat. concepts: estimation, CI, CLT

The sample mean and its distribution, CLT

- The sample mean of a sample $X_1, \dots, X_n$ of sample size n is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i; \text{ for binomial data } X_1, \dots, X_n \sim \text{Bin}(1, p), \quad \bar{X} = \hat{p}.$$

- If $X_1, \dots, X_n \sim N(\mu, \sigma^2)$-distribution, then $\bar{X} \sim N(\mu, \sigma^2/n)$ exactly.
- When the sample is taken from any distribution with expectation μ and variance σ^2, $\bar{X}$ still has approximately $N(\mu, \sigma^2/n)$-distribution ($\bar{X}$ is asymptotically normal). This is the Central Limit Theorem (CLT):

$$\bar{X} \sim N(\mu, \sigma^2/n), \quad \text{or} \quad \sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1), \quad \text{appr. (for large } n).$$

- The CLT is a fundamental result of probability theory.
- Example: for binomial data $X_1, \dots, X_n \sim \text{Bin}(1, p)$, $E(X_i) = p$, $\bar{X} = \hat{p}$, $\sigma^2 = Var(X_i) = p(1 - p) \approx \hat{p}(1 - \hat{p})$, so that approximately (for large n)

$$\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1 - \hat{p})}} \sim N(0, 1).$$

Standardizing the mean

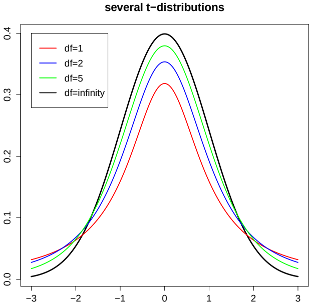
- Any normal random variable $X \sim N(\mu, \sigma^2)$ can be standardized into a standard $N(0,1)$-variable by $Z = (X - \mu)/\sigma \sim N(0,1)$.
- Converse is also true: if $Z \sim N(0,1)$, then $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$.
- General fact: if $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ are independent, then $V = aX + bY + c \sim N(a\mu_x + b\mu_y + c, a^2\sigma_x^2 + b^2\sigma_y^2)$.
- As $\bar{X} \sim N(\mu, \sigma^2/n)$ (exactly or approximately), standardizing $\bar{X}$ yields

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

The t-distribution

- In a real data set $X_1, \ldots, X_n$, the population standard deviation σ is unknown and needs to be estimated by the sample standard deviation s.
- This uncertainty influences the distribution of the resulting statistics $\frac{\bar{X}-\mu}{s/\sqrt{n}}$.
- If $X_1, \ldots, X_n$ is a sample from $N(\mu, \sigma^2)$, then the random variable $T = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$, a t-distribution with $n-1$ degrees of freedom.
- $t_{n-1} \neq N(0,1)$, but $t_{n-1} \approx N(0,1)$ for big n.

For any other generating distribution, $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ is still approximately $N(0,1)$, but often t_{n-1}-distribution is used instead. This does no harm to inference on μ as it only leads to more conservative quantiles in testing and confidence intervals.



Estimation – the concepts

- Suppose we assume that our population of interest has a certain distribution with an unknown parameter, e.g., its mean μ or a fraction p.
- A point estimate for the unknown parameter is a function of only the observed data $(X_1, \dots, X_n)$, seen as a random variable. frequentist POV
- We denote estimators by a hat: $\hat{\mu}$, $\hat{p}$, etc.
- Examples of point estimates: $\hat{\mu} = \bar{X}$, the sample proportion $\hat{p}$.
- A confidence interval (CI) of level $1 - \alpha$ for the unknown parameter is a random interval based only on the observed data $(X_1, \dots, X_n)$ that contains the true value of the parameter with probability at least $1 - \alpha$.

Estimating the mean, CI

- Recall that $\bar{X} \sim N(\mu, \sigma^2/n)$ for $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$ distribution.
- The upper quantile z_α of the $N(0,1)$-distribution is such z_α that $P(Z \geq z_\alpha) = \alpha$ for $Z \sim N(0,1)$, (in R: z_α=`qnorm(1-alpha)`). Then

$$1-\alpha = P\big(|Z| \leq z_{\alpha/2}\big) = P\Big(\tfrac{|\bar{X}-\mu|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\Big)$$
$$= P\Big(\bar{X} - z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\tfrac{\sigma}{\sqrt{n}}\Big).$$

- In other words, $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = [\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}]$ is the confidence interval of μ of level $1-\alpha$.
- If the standard deviation σ is unknown, we estimate it by s and the confidence interval is based on a t-distribution and the upper t-quantile t_α =`qt(1-alpha,df=n-1)` (i.e., $P(T \geq t_\alpha) = \alpha$ for $T \sim t_{n-1}$.).
- The t-confidence interval of level $1-\alpha$ for μ then becomes

$$\bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = \Big[\bar{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\Big].$$

The t-CI's are (nearly) always used, since σ is almost never known in practice. In view of CLT, this can be used also for non-normal data.

Margin of error for the mean

- The $(1-\alpha)$-confidence interval for μ

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \text{or} \qquad \bar{X} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

- The margin of error is thus $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or $E = t_{\alpha/2}\frac{s}{\sqrt{n}}$.
- Remark 1. If we take larger n, the confidence interval will be smaller (shorter), i.e., gaining more accuracy at the same confidence level.
- Remark 2. If σ (or s) is smaller, the confidence interval will be shorter, again yielding more accuracy at the same confidence level.
- Remark 3. If we take bigger α, the confidence interval will be shorter. Warning: more accuracy at the cost of a lower confidence level.

Minimal sample size

- Question: how big should the sample size be in order to obtain a margin of error at most E? (This is the same as having the CI length at most $2E$.)
- Answer: n must satisfy $z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq E$ or $t_{\alpha/2}\frac{s}{\sqrt{n}} \leq E$, or equivalently

$$\sqrt{n} \geq \frac{z_{\alpha/2}\sigma}{E} \quad \text{or} \quad \sqrt{n} \geq \frac{t_{\alpha/2}s}{E}, \qquad \text{so that}$$

$$n \geq \frac{(z_{\alpha/2})^2\sigma^2}{E^2} \quad \text{or} \quad n \geq \frac{(t_{\alpha/2})^2 s^2}{E^2} \approx \frac{(z_{\alpha/2})^2 s^2}{E^2}.$$

- Remark. For large n we have $t_{\alpha/2} \approx z_{\alpha/2}$ and $s \approx \sigma$. Actually, it makes sense to use $z_{\alpha/2}$ in the second formula instead of $t_{\alpha/2}$, because $t_{\alpha/2}$ depends on (unknown) n as well.

Estimating a proportion, CI, minimal sample size

- We want to estimate a population proportion p, based on a sample $X_1, \dots, X_n \sim \mathrm{Bin}(1, p)$. The point estimate for p is $\hat{p} = \bar{X}$.
- The $(1-\alpha)$-confidence interval for p is $\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (based on CLT).
- To ensure a margin of error at most E, the minimal sample size must satisfy $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq E$ or $n \geq z^2_{\alpha/2}\hat{p}(1-\hat{p})/E^2$.
- Example: trains in time. We want take a sample trains to estimate the fraction p of trains that arrive in time. This fraction was estimated as 0.95. We want a 98% confidence interval for p with length at most 3% (0.03). Question: how many trains should we have in the sample? Answer. A CI length of 3% means $2E = 0.03$ so that $E = 0.015$. Next, $\hat{p} = 0.95$ and $1 - \hat{p} = 0.05$. For a 98% interval we have $z_{\alpha/2} =$ `qnorm(0.99)`=2.326. Hence, the minimal sample size must satisfy

$$n \geq \frac{z^2_{\alpha/2}\hat{p}(1-\hat{p})}{E^2} = \frac{(2.326)^2 \times 0.95 \times 0.05}{(0.015)^2} = 1142.5.$$

  which is found in R by `qnorm(0.99)^2*0.95*0.05/(0.015)^2`. In words: we need at least 1143 trains to ensure a 98%-CI of length at most 0.03.

Recap basic stat. concepts: hypothesis testing

Hypothesis testing: the concepts

- Null hypothesis H_0 and alternative hypothesis H_1 about the world.
- A statistical test based on the observed data $X = (X_1, \dots, X_n)$ chooses between H_0 and H_1. The claim of interest is usually represented by H_1.
- Precisely, for some test statistic $T = T(X)$ and critical region K, we reject H_0 (and accept H_1) if $\{T(X) \in K\}$ (the strong outcome), otherwise do not reject H_0 (the weak outcome).
- A test statistic $T = T(X)$ summarizes the data $X = (X_1, \dots, X_n)$ in a relevant way. Critical region K is chosen in such a way that $T(X)$ is hardly ever expected to take values in K if H_0 were true.
- In general, to construct a good K we need to know the distribution of $T(X)$ under H_0. This is usually the main difficulty in constructing tests.
- The test (and test statistic) is not unique. Different tests are possible for the same pair of hypothesis H_0, H_1, with different performances.

Hypothesis testing: *p*-values

- 3 ways to test, say $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$, test stat. $T(X)$, level α:
  - by checking whether $T(X) \in K_\alpha = \{T(X) < t_\alpha\}$ or not;
  - by comparing the *p*-value to α: $p = P(T(X) \leq t) \leq \alpha$ or not;
  - by checking whether μ_0 is in the (relevant) $(1-\alpha)$-CI for μ or not.

  By using *p*-values is the most common way: e.g., for the realized value $T(x) = t$ and $T \sim t_{n-1}$, check whetheer $p = P(T \leq t) \leq \alpha$ or not.
- Given observed value t of the test statistic, the *p*-value is the probability under H_0 of observing a value for T that is at least as extreme as t. A small *p*-value indicates that the observed data is unlikely if H_0 were true.
- When the *p*-value is below the chosen significance level α (e.g., 0.05), reject H_0 (strong outcome), otherwise do not reject H_0 (weak outcome).
- If H_0 is rejected, the data are said to be statistically significant at level α.
- By construction, under H_0, the *p*-value is like a uniform draw from $[0,1]$. Let us show this for our example. Let $p(t) = P(T \leq t) = F_T(t)$ for $T \sim t_{n-1}$, then the (random) *p*-value is $\tilde{p} = p(T(X)) = F_T(T(X))$, and for any $\alpha \in (0,1)$, $P(\tilde{p} \leq \alpha) = P(F(T(X)) \leq \alpha) = P(T(X) \leq F_T^{-1}(\alpha)) = F(F^{-1}(\alpha)) = \alpha$.

Example: the one sample t-test(s)

- Data $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The t-test is for testing about μ.
  1. $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ (`t.test(data,mu=`μ_0`,alt="g")`)
  2. $H_0 : \mu \geq \mu_0$ vs. $H_1 : \mu < \mu_0$ (`t.test(data,mu=`μ_0`,alt="l")`)
  3. $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ (`t.test(data,mu=`μ_0`)`)
- In all 3 cases, at the border of H_0 and H_1 (i.e. for $\mu = \mu_0$), the

  test statistic $T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ has t-distribution with $n-1$ degrees of freedom.
- The p-value for observed value $T(x) = t$ of the test statistic is
  1. $p = P(T \geq t)$ under H_0 (i.e., assuming that $T \sim t_{n-1}$);
  2. $p = P(T \leq t)$ under H_0;
  3. $p = P(|T| \geq |t|) = 2 \min\{P(T \geq t), P(T \leq t)\}$ under H_0.
- For testing, say, situation 3, $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, we reject H_0 if

  either $|T(x)| > |t_{\alpha/2}|$,
  or $p = P(|T| \geq |t|) < \alpha$ under H_0,
  or μ_0 does not belong to the CI $\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$.

Hypothesis testing: types of errors, power of the test

- Statistical tests $\psi = 1\{T(X) \in K\} \in \{0,1\}$ make two types of errors:
  - Error of the first kind (type I error): rejecting H_0 while it is true.
  - Error of the second kind (type II error): not rejecting H_0 while it is false.
- It is desirable to construct tests with small probability of type I error P_{H_0}(type I error) ($\leq 5\%$). P_{H_0}(type I error) is called the level of this test.
- P_{H_1}(type II error) depends (among others) on the amount of data.
- The probability of correct (i.e., when H_0 is not true) rejecting H_0 is called the power of the test. Under H_1, power $= 1 - P_{H_1}$(type II error).
- Different test statistics can yield different statistical power of the test.
- Higher sample sizes typically yield higher power.
- The Neyman-Pearson paradigm: tests with high statistical power are preferred, while controlling the level of the test by a fixed margin (5%).

The power of a test is specified for each possibility under H_1. E.g., if $H_0 : \mu \leq 0$ then the power can be calculated in each $\mu > 0$. A *good* test (that is, a test based on a *good* test statistic) has high power in all positive μ-values, relative to other tests.

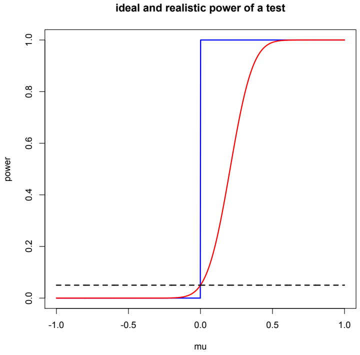
When H_0 is true, power $= P_{H_0}$(do not reject H_0) $= P_{H_0}$(type I error) $\leq \alpha$.

Ideal test and realistic test

The ideal test ψ_{ideal} makes no errors:

- never falsely reject (no error of type I): $\psi_{ideal} = 0$ on H_0;
- always reject when H_1 is true (no error of type II): $\psi_{ideal} = 1$ on H_1.

The power of the ideal test and a realistic test for $H_0 : \mu \leq 0$ vs. $H_1 : \mu > \mu_0$. The dashed line is the level of the test, here 0.05.



One can think of probability of type I error as the proportion of false positives (or the false positive rate), and probability of type II error as the proportion of false positives (or the false negative rate) in binary classification.

Practical significance

- Statistical significance is about generalization: an observed effect is not due to chance, it should be observed again if a new experiment were performed.
- In practice, this boils down to practical significance which is about the relation between the size of the effect and the available information.

EXAMPLE Suppose that a coin has probabilities $1/2 - 10^{-10}$ and $1/2 + 10^{-10}$ to land HEAD or TAIL.
If we use the coin to decide who will kick-off in a soccer game, then TAIL has a slight advantage, but the difference is negligible. A statistical test based on observing 100 tosses of the coin will not reject H_0, but a test based on observing 10^{21} coin tosses almost certainly will.

Recap: examples in R

Example of one sample right-sided t-test – crime rate

We want to test whether the mean crime rate (recal column `crime` from the dataset `expencescrime`) is bigger than 4500. Use `t.test` to do the t-test in R:

```
> x=expensescrime$crime; n=length(x); t.test(x,mu=4500,alt="g")
       One Sample t-test
data:  x
t = 1.5583, df = 50, p-value = 0.06273
alternative hypothesis: true mean is greater than 4500
95 percent confidence interval:
 4477.224      Inf
sample estimates:
mean of x
 4801.843
```

The R-output gives $\bar{X} = 4801.843$, the value of the test statistics $t = 1.5583$ (or `t=(mean(x)-4500)/(sd(x)/sqrt(n))`), the p-value $p \approx 0.63$ (or `1-pt(t,n-1)`). Conclude that the mean crime rate is not greater than 4500. Interestingly, also confidence interval $[4477.224, +\infty)$ is given in the R-output. But why is `Inf` in it?

Point and interval estimation, one sample two-sided t-test

Given a random sample $X_1, \ldots, X_N$ from a population with mean μ and unknown variance σ^2, we wish to estimate μ, construct a CI for it, and to test $H_0 : \mu = \mu_0$ against $H_1 ; \mu \neq \mu_0$ for some given number μ_0, e.g., $\mu_0 = 0$.

```
> mu=0.2; x=rnorm(50,mu,1)  # creating artificial data
> t.test(x,mu=0)
      One Sample t-test
data:  x
t = 2.4211, df = 49, p-value = 0.01922
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.05219746 0.56202370
sample estimates:
mean of x
0.3071106
```

For this (synthetic) data $X_1, \ldots, X_n \sim N(0.2, 1)$, we read off from the above R-output the estimate $\bar{X} \approx 0.31$, the 95% CI $[0.052, 0.562]$, p-value ≈ 0.019. $H_0 : \mu = 0$ is rejected because 1) $|t| = 2.42 > |t_{\alpha/2}| = \mathtt{qt}(0.975, 49) \approx 2.01$, or because 2) p-value$=0.01922 < 0.05$, or because 3) $0 \notin [0.052, 0.562]$.

Standard error and confidence interval

The standard error $\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$ of the estimator $\bar{X}$ is a measure of its precision. By CLT, this estimator is approximately normally distributed, hence

$$\texttt{Estimate} \pm 1.96 \times \texttt{Std.Error} \quad \text{gives an approx. 95\% CI.}$$

The bigger the sample size n, the smaller the standard error and the confidence intervals. The estimates get more precise, as more information is available.

Generate estimates $\bar{X}$ from standard normal sampes (i.e., the true $\mu = 0$):

sample size	Estimate	Std.Error
10	0.3564	0.3604
50	0.2198	0.1510
100	0.1098	0.1067
1000	-0.007433	0.031466

In all cases the true value 0 is in the 95% confidence interval

$$\texttt{Estimate} \pm 1.96 \times \texttt{Std.Error}.$$

The margin $m = 1.96 \times \texttt{Std.Error}$ is based on the asymptotic normality of $\bar{X}$ and the fact that s is a good estimator of σ. If in the CI we use the upper t-quantile $t_{0.025,n-1}$ instead of $z_{0.025} \approx 1.96$, the CI will be bigger (i.e., more "conservative") because always $t_{\alpha,n-1} > z_\alpha$, but $t_{\alpha,n-1} \to z_\alpha$ as $n \to \infty$.

Recap binomial and (appr.) normal tests for a proportion

Setting: $X \sim \text{Bin}(n, p)$, e.g., the number of successes in n trials, p is the success proportion (or the probability of success). We want to test about p.

Hypotheses: $H_0 : p \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} p_0$ versus $H_1 : p \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} p_0$.

Test statistic: X or $T = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$, where $\hat{p} = \frac{X}{n}$.

Distribution under H_0: $X \sim \text{Bin}(n, p_0)$ (exactly) or $T \sim N(0, 1)$ (approx.)

In R: `binom.test(x,n,p=`p_0`,alt=...)` `prop.test(x,n,p=`p_0`,alt=...)`

Testing for binomial data: example trains on time

- In a (fictive) sample of 100 trains arriving at Amsterdam Central station, we observe a sample proportion $\hat{p} = 0.89$ (89/100) trains arriving in time.
- We want to test whether this is significantly lower than the reported 95% for the Netherlands. Hence, we test $H_0 : p \geq 0.95$ versus $H_1 : p < 0.95$.
- This is a binomial sample with $n = 100$ and p unknown. One can use the exact binomial test `binom.test` or the proportion `prop.test`.

The exact binomial test:

```
> binom.test(89,100,p=0.95,alt="l")
[ some output is deleted ]
     p-value = 0.01147
```

The approximate proportion test:

```
> prop.test(89,100,p=0.95,alt="l")
    [ some output is deleted ]
    p-value = 0.005808
```

The p-values in both tests < 0.05 (although different). Conclusion: reject H_0.

Example continued: trains on time

Now perform the two-sided test $H_0 : p = 0.95$ versus $H_1 : p \neq 0.95$.

The exact binomial test:

```
> binom.test(89,100,p=0.95)
   [ some output is deleted ]
    p-value = 0.01739
```

The approximate proportion test:

```
> prop.test(89,100,p=0.95)
    [ some output is deleted ]
   p-value = 0.01162
```

The p-values in both tests < 0.05 (although different). Conclusion?

The influence of the sample size: suppose we had found 890 trains arriving in time amongst 1000 trains:

The exact binomial test:

```
> binom.test(890,1000,p=0.95)
   [ some output is deleted ]
    p-value = 3.786e-14
```

The approximate proportion test:

```
> prop.test(890,1000,p=0.95)
   [ some output is deleted ]
    p-value < 2.2e-16
```

e-14 $= 10^{-14} = 0.00000000000001$, 3.786e-14$= 0.00000000000003786$. The same deviation from H_0 in more data yields a lower p-value.

Tests for a difference in proportions

Setting: X_1 successes in a sample of size n_1 taken from population 1 and X_2 successes in a sample of size n_2 from population 2. We want to test about the difference in population success proportion p_1 and p_2.

Hypotheses: $H_0 : p_1 - p_2 \left\{ \begin{array}{c} = \\ \leq \\ \geq \end{array} \right\} 0$ versus $H_1 : p_1 - p_2 \left\{ \begin{array}{c} \neq \\ > \\ < \end{array} \right\} 0.$

Test statistic: $T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, where $\hat{p}_1 = \frac{X_1}{n_1}$, $\hat{p}_2 = \frac{X_2}{n_2}$, $\bar{q} = 1 - \bar{p}$,

$\bar{p} = \frac{x_1+x_2}{n_1+n_2}$ is the pooled sample fraction (the best estimate of p under $H_0 : p_1 = p_2 = p$).

Distribution of T under H_0: $N(0,1)$ (approximately).

In R: `prop.test(c(x1,x2),c(n1,n2),alt=...)`

Example: compare two proportions of defective items

We test whether the proportions of defective items in two manufacturing processes are (significantly) different. In a sample of 1000 items in process A we find 20 defective items, and in a sample of 1500 items in process B we find 19 defective ones. Question: is there a significant difference in (population) proportions p_A and p_B of defective items for precesses A and B?

Thus the sample proprtions are $\hat{p}_A = \frac{20}{1000} = 0.02$ and $\hat{p}_B = \frac{19}{1500} = 0.013$, but are they significantly different? We apply the approximate proportion test:

```
> prop.test(c(20,19),c(1000,1500))
   [ some output is deleted ]
  p-value = 0.1989
```

Conclusion? Do not reject $H_0 : p_A = p_B$.

Suppose we found the same sample proportions but in larger samples:

```
> prop.test(c(200,190),c(10000,15000))
  [ some output is deleted ]
  p-value = 5.85e-06
```

Now we do reject $H_0 : p_A = p_B$. Why?

More information (estimates, CI's) can be extracted from the complete R-output.

Two sample t-test

- Given two populations with means μ and ν, we wish to test $H_0 : \mu\left\{\begin{array}{c} = \\ \leq \\ \geq \end{array}\right\}\nu$ versus $H_1 : \mu\left\{\begin{array}{c} \neq \\ > \\ < \end{array}\right\}\nu$. Take a sample $X_1, \ldots, X_M$ from the first population and, independently, $Y_1, \ldots, Y_N$ from the second.
- The test is based on $\bar{X}_M - \bar{Y}_N$ which is a reasonable estimate for $\mu - \nu$. If it deviates from 0 too much (in the relevant direction), we reject H_0.
- How different? $\bar{X}_M - \bar{Y}_N$ will not exactly be $\mu - \nu$. The estimation error depends on M and N and the standard deviations of the populations.
- T-statistic: $\bar{X}_M - \bar{Y}_N$ is divided by an estimate $S_{M,N}$ of its standard error.

  $$\text{under } H_0, \quad T = \tfrac{\bar{X}_M - \bar{Y}_N}{S_{M,N}} \sim t_{M+N-2}, \quad S^2_{M,N} = S^2_{X,Y}\left(\tfrac{1}{M} + \tfrac{1}{N}\right).$$

  where $S^2_{X,Y} = \frac{1}{M+N-2}\left(\sum_{i=1}^M (X_i - \bar{X}_M)^2 + \sum_{j=1}^N (Y_j - \bar{Y}_N)^2\right)$.
- Then T is compared to the critical value (quantile from t_{M+N-2}-distrib.), or the p-value (computed by using t_{M+N-2}-distribution) is compared to α.

The standard t-test assumes that the two populations are (approx.) normal. If the sample sizes M and N are large, then the test performs well even without this assumption, but the test is unreliable for M,N less than 20.
The quantity $S^2_{M,N}$ is called pooled sample variance.

Two sample two-sided t-test: implementing in R

For example, we test $H_0 : \mu = \nu$ against $H_1 : \mu \neq \nu$ by the two sample t-test:

```
> mu=0;nu=0.5
> x=rnorm(50,mu,1);y=rnorm(50,nu,1) #creating artificial data
> t.test(x,y)
        Welch Two Sample t-test
data:  x and y
t = -2.4339, df = 96.574, p-value = 0.01677
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.85202520 -0.08659066
sample estimates:
 mean of x  mean of y
0.06552453 0.53483246
```

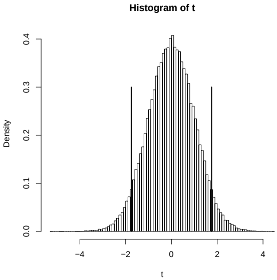
The observed $t = -2.4339$, so that the corresponding p-value is

$$P(|T| > |t|) = 2P(T > |t|) = 2(1 - P(T \leq |t|)) \approx 0.0167.$$

This can be found in the above output, and we could also compute this directly in R as `2*(1-pt(2.4339,98))=0.01674788`. We thus reject H_0 as p-value $\approx 0.017 < 0.05$.

p-value for two sample t-test

We can also evaluate this p-value by simulating from the null hypothesis.



We generate a population of T-values under H_0 by repeating the sampling. The p-value of the observed value t is approximately the fraction of this population that is bigger than $|t|$ or smaller than $-|t|$.

```
> mu=nu=0; t=numeric(100000)
> for(i in 1:100000){x=rnorm(50,mu,1);y=rnorm(50,nu,1);t[i]=t.test(x,y)[[1]]}
> sum(abs(t)>=abs(-2.4339))/length(t)  ##cf. 2*(1-pt(2.4339,98))=0.01674788
[1] 0.01744
```

Different test statistics

EXAMPLE The t-test is for testing the population mean μ of a normal population, $H_0 : \mu = \mu_0$. Given a sample $X_1, \ldots, X_n$, the test statistic is

$$T = \frac{\bar{X} - \mu_0}{S_X/\sqrt{n}}, \quad \text{where } \bar{X} = \frac{1}{n} \sum\nolimits_{i=1}^{n} X_i, \;\; S_X^2 = \frac{1}{n-1} \sum\nolimits_{i=1}^{n} (X_i - \bar{X})^2.$$

When T is very different from 0, reject H_0. The critical value for T that acts as border between rejecting and not rejecting H_0 is based on the distribution of T under H_0. For t-test, this distribution is the t_{n-1}-distribution.

EXAMPLE For testing $H_0 : \mu = 0$ we can as well use the sign test. Given a sample $X_1, \ldots, X_n$ from the population, the test statistic for the sign test is

$$T = \#(X_i < 0).$$

If T is very different from $\frac{n}{2}$, we reject H_0. The critical value comes from the $\mathrm{Bin}(n, \frac{1}{2})$-distribution, the distribution of number of heads in throwing n times a fair coin.

Comparing powers of different tests

Assume we have a normal sample and test $H_0 : \mu = 0$ using the t-test and the sign test. We can compare the power in $\mu = 0.5$ of the two tests by simulation. Recall that always power $= P_{H_1}(\text{reject } H_0)$.

```
> B=1000; n=50
> psign=numeric(B)   ## will contain p-values of sign test
> pttest=numeric(B)  ## will contain p-values of t-test
> for(i in 1:B) {
+  x=rnorm(n,mean=0.5,sd=1) ## generate data under H1 with mu=0.5
+  pttest[i]=t.test(x)[[3]]                     ## extract p-value
+  psign[i]=binom.test(sum(x>0),n,p=0.5)[[3]] }   ## extract p-value
> sum(psign<0.05)/B
[1] 0.746
> sum(pttest<0.05)/B
[1] 0.937
```

The power in $\mu = 0.5$ for the t-test (0.937) is higher than for the sign test (0.746) when we reject for p-values smaller than the level 0.05. Why? Because for normal data, the t-test has better performance than the sign test.

To finish

We discussed

1. course organization
2. experimental design
3. recap probability theory and basic statistics
4. recap: examples in R

Study the exam to test your prerequisite knowledge and Assignment 0 (not to be submitted) to learn how to make assignment reports to submit.

Next time bootstrap methods, one sample tests.