Data analysis with R software

Data-analyysi R-ohjelmistolla

Tommi Härkänen

National Institute for Health and Welfare (THL), Helsinki
E-mail: `tommi.harkanen@helsinki.fi`

University of Helsinki, January 29, 2013

Contents

What is randomness?

In applied sciences one can consider lack of predictability:

Deterministic result For example, a medical treatment cures all patients, but without treatment no one gets cured.

Random result For example, when treated, 60 % gets cured but without treatment 20 % gets cured.

Other examples of randomness

- Coin tossing or dice throwing
- Quantum mechanics
- Weather
- Stock market

Some examples on definitions of probability

A **unique event** cannot be predicted unless it is certain. One can form **subjective probabilities** prior to the observation.

If the process, which generates the data, can be **repeated**, the frequencies of different events can be calculated ⇒ **frequency probabilities**.

Continuous vs. discrete random variables

- A **discrete** random variable can have a countable number of values[1]. A single value can have a positive probability. For example,
  - Dice throwing (6 possible values)
  - Number of heads before the first tail in coin tossing (infinite number of possible values 0, 1, 2, ...)
- A **continuous** random variable has a **zero probability** for any single value. For example,
  - Height of a person.
  - Blood pressure.

[1] The values can be enumerated 1, 2, 3, ...

Characterization of probabilities

Cumulative distribution function The probability that a random variable X gets a value less or equal to x: $\mathbb{P}\{X \leq x\}$. For example, the probability that the height of a randomly chosen subject in the classroom is at most $x = 170$ cm.

Density function The point probability that a **discrete** random variable X equals some constant value x can be zero or positive. For example, the probability that a randomly chosen subject in the classroom is $x =$ male.
For a continuous random variable, value of a density function at x multiplied by a small positive constant $a > 0$ is approximately the probability that the value of the r.v. is within $[x, x + a]$.

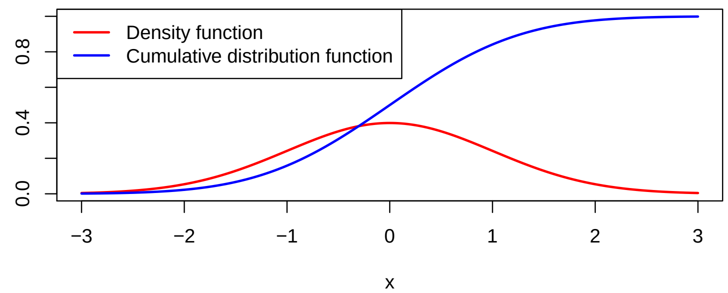
Distributions

- See help page `Distributions` for complete list.
- Generally four functions available for each distribution starting with letters
  - d Density function
  - p Qumulative distribution function
  - q Quantile function
  - r Random number generation
- For example, normal distribution has functions

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

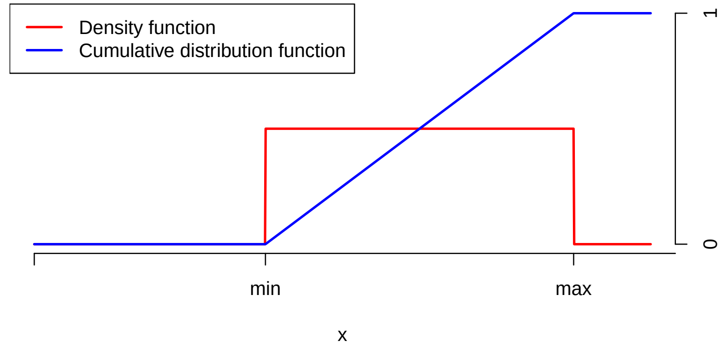
Normal distribution

```
pdf("normal_dist.pdf", width = 7, height = 3.5)
x <- seq(-3, 3, 0.01)
plot(x, dnorm(x, mean = 0, sd = 1), type = "l", col = "red",
    lwd = 2, ylim = c(0, 1), ylab = "")
lines(x, pnorm(x, mean = 0, sd = 1), type = "l", col = "blue",
    lwd = 2)
legend("topleft", lty = c(1, 1), lwd = 2, col = c("red", "blue"),
    cex = 1, legend = c("Density function", "Cumulative distribution func-
tion"))
dev.off()
```



Continuous uniform distribution

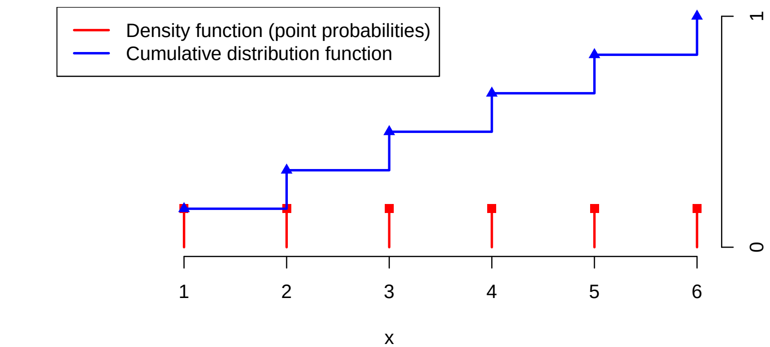
- Observations are between a minimum and maximum (a.s.)
- All intervals of the same length on the distribution's support [min, max] are equally probable.



Discrete uniform distribution

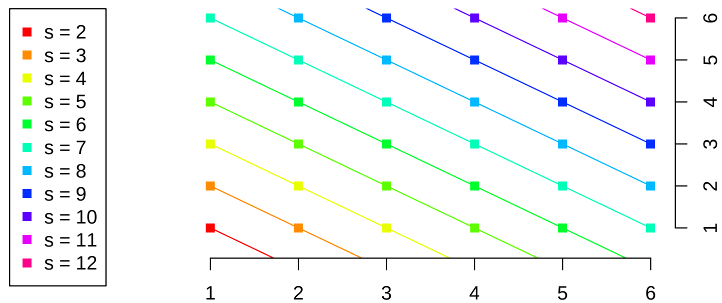
- Observations are between a minimum and maximum **integer values**
- All intervals of the same length on the distribution's support $\{\min, \min + 1, \ldots, \max\}$ are equally probable.

Example In dice rolling minimum is 1 and maximum 6:



Sum of two discrete uniformly distributed random variables

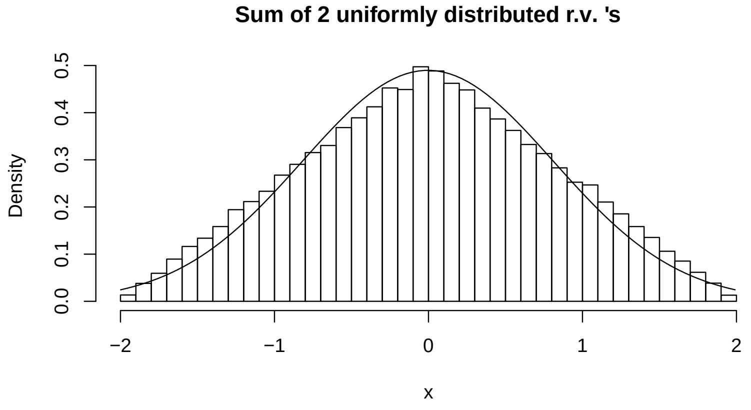
Example Assume rolling two independent dices X_1 and X_2 (min=1 and max=6).
What is the probability of the sum $S := X_1 + X_2$ being equal to s?



- 6×6 different outcomes $(X_1, X_2) = (x_1, x_2)$ with equal probabilities 1/36
- The sum s is between a $2 \times$ minimum and $2 \times$ maximum
- Maximum probability is at $s = 7$: $\mathbb{P}\{S = 7\} = 6/36 \approx 0.167$

Central limit theorem (CLT)

Consider a sum of independent random variables.
The more terms there are in the sum, the closer the distribution of the sum resembles normal distribution.
Example: random variables with uniform distribution on [-1, 1].



What does CLT mean in practice?

Many estimators are sums over observed values, which are independent.
E.g. sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Therefore sampling distributions of estimators are often normal, if the sample size is large.

An experiment

- A researcher has a hypothesis.
- He/she plans and executes an experiment, and collects data in order to test the validity of the hypothesis.
- **Question**: Does the data support the hypothesis?
- For example, coin tossing:
  - Hypothesis is "Probability of heads in is 0.5."
  - Experiment is "Toss a coin 10 times."
  - Data are the proportion of heads.
- For example, medical experiment:
  - Treatment group and control group.
  - **Null hypothesis, H0**: "No difference between groups" i.e. the treatment has no effect.
  - **Alternate hypothesis, H1**: "Some differences between groups."
  - Experiment is "Administer the treatment to the treatment group and some placebo treatment to the control group."
  - Data are the recovery status of the patients (and the group indicator).

Experiments should be repeatable

- Other researchers should be able to repeat the experiment in (identical) conditions.

- In frequentist inference one assumes (infinitely) many hypothetical experiments, in which the null hypothesis is true.
- The outcome of the observed experiment is then compared with the distribution of the outcomes of the hypothetical experiments.
- If the observed outcome seems to be very rare, then the empirical evidence does not support the null hypothesis.

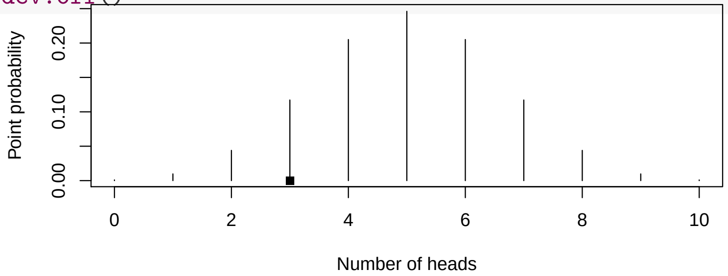
- Observational vs. experimental studies: observational studies can be unique $\Rightarrow$ the idea of repeated experiments (or new samples of study subjects) in identical conditions unrealistic.

Distribution of hypothetical experiments

Return to the experiment with 10 tosses of a coin. 3 heads were observed.

- **H0**: Probability of head is 0.5.
- The probability distribution of number of heads in 10 tosses is **Binomial**.

```
x <- 0:10
pdf("binom10.pdf", width = 7, height = 3.5)
plot(x, dbinom(x, size = 10, p = 0.5), type = "h", xlab = "Number of heads",
    ylab = "Point probability")
points(3, 0, pch = 15)
dev.off()
```



One- and two-sided tests

"How often hypothetical experiments would generate data at least as rare as the observed outcome?"

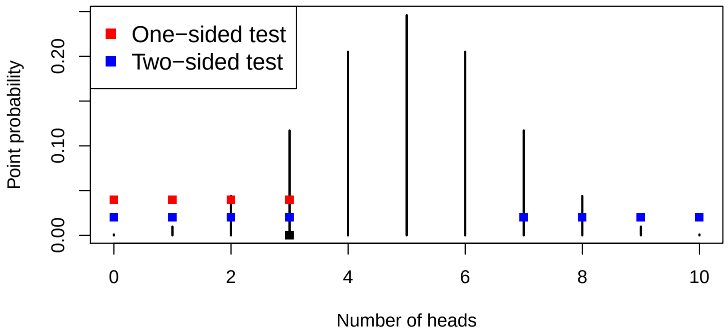
One-sided test: Calculate the probability of sampling distribution at the observed value and at the more extreme values.

Two-sided test: If the sampling distribution is symmetric, calculate the p-value from both tails of sampling distribution.

Example: binomial test

The experiment with 10 tosses of a coin. 3 heads were observed.

Note that the expected number of heads is $10 \times 0.5 = 5$ and the sampling distribution is symmetric.
The p-value of the one-sided test is 0.17 and of the two-sided test 0.34.



Sample size of the experiment

The researcher considers the coin to be "unfair" if the probability of head is outside [0.45, 0.55].

- Are 10 tosses enought to detect the bias?
- If not, then how many are needed?

Recall that the standard error (SE) of the estimated proportion is

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{p \times (1-p)}{n}}. \tag{1}$$

With $n = 10$ tosses and $p = 0.5$, $\mathrm{SE}(\hat{p}) = 0.16$.

Approximately 95 % of the sampling distribution lies between $p \pm 1.96 \times \mathrm{SE}(\hat{p})$. Desired sample size can be solved from (1): $n = (1.96^2/a^2) \times p \times (1-p)$, where $a = 0.05$ is the desired accuracy.

The researcher decides to toss the coin $n = 384$ times.

Mean of normally distributed r.v.'s

SD is known

Assume that n independent r.v.'s are sampled from $\mathrm{N}(\mu, \sigma^2)$. Recall that parameter μ is expectation and σ^2 is variance (σ is standard deviation SD).

The sample mean $\bar{x}$ is also normally distributed $\mathrm{N}(\mu, \sigma^2/n)$. The SE of the sample mean is $\sigma/\sqrt{n}$.
Large $n \Rightarrow$ small SE.

Example: $n = 16$, $\sigma^2 = 4$ and $\bar{x} = 3$. Let the null hypothesis be $\mathrm{H}_0 : \mu = \mu_0 = 2$. Then test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{3-2}{2/4} = 2.$$

The tail probability of one-sided test is
`1 - pnorm(2, mean=0, sd=1)` $= 0.023$.

Student's t test

SD is unknown

Typically σ is unknown as well as μ. If the sample size n is large then the sample SD s is close to the true σ.

If n is small, then the uncertainty of s needs to be accounted for using t distribution with $n-1$ degrees of freedom (df) instead of normal distribution.

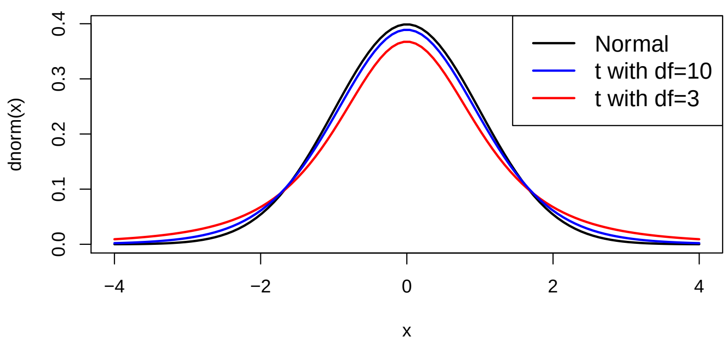
Example: $n = 16$, $s^2 = 4$ and $\bar{x} = 3$. Let the null hypothesis be $\mathrm{H}_0 : \mu = \mu_0 = 2$. Then test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{3-2}{2/4} = 2.$$

The tail probability of one-sided test is
`1 - pt(2, df=16-1)` $= 0.032$.

The t distribution vs. normal distribution

If n is small, then s is imprecise, thus the sampling distribution contains more extreme values than normal distribution. The smaller degrees of freedom (df), the more extreme values. If df (and n) are large, then t distribution is close to normal distribution (s is close to σ).



Comparing two independent samples: t test

SD unknown, possibly unequal

Test statistic is now

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Notations are analogous to the one sample case. Example:

```
with(iris, t.test(Sepal.Length[Species == "versicolor"], Sepal.Length[Species ==
    "virginica"]))

##
##  Welch Two Sample t-test
##
## data:  Sepal.Length[Species == "versicolor"] and Sepal.Length[Species == "virginica"
## t = -5.629, df = 94.03, p-value = 1.866e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.882 -0.422
## sample estimates:
## mean of x mean of y
##     5.936     6.588
```

Samples are not normally distributed

Some non-parametric tests

Median test: Calculate the **median** of the joined data sets, create a binary variable and do the χ^2 test for the 2×2 table.

Mann-Whitney test: More efficient than the median test. Does not assume normality. Based on ranks of observations. Install package `exactRankTests`, where the function `wilcox.exact` can be found.

Paired two-sample tests

The arguments are generally two vectors of the same length.

The t test: Use the **t.test** function and **paired=TRUE** argument.

Sign test: Calculate the number of pairs i for which $x_i < y_i$. This number should be close to 50 % of the number of pairs. Compare with Binomial distribution as with the binomial test above.

Wilcoxon test: More efficient than the sign test. Does not assume normality, but requires a **symmetric** distribution. (If distribution is not symmetric, then use the sign test.) Install package `exactRankTests`, where the function `wilcox.exact` can be found. Use option `paired=TRUE`.

Two measurements on the same subjects

Any changes between measurements?

For each subject i there are two measurements x_i and y_i. For example, a measurement before medical treatment and the other after the treatment.

- The measurements are often correlated.
- Positive correlation means e.g. that subjects i who had high level of symptoms at the time of the first measurement x_i, tend to have high level of symptoms also at the second measurement y_i.
- The correlation must be accounted for in analyses.