Basic Statistical Analysis

Rodrigo Labouriau

Aarhus University
Department of Mathematics
Applied Statistics Laboratory

Module 1, Day 2 - Statistical Models - 2024

[1]

General Remark

This material is only for internal use in the course.

Please, do not circulate and **do not record**.

Outline

An initial challenger: The Master Quiz problem

- Three boxes:
  One contains a BIG check, the other two are empty
- You choose one box,
  before you open the box the Master-Quiz says
  '*I give you a hint, the check is* **not** *here*'
  and he opens one of the remaining boxes, which is empty
- The Master-Quiz continues:
  Would you like to change and choose the other closed box?
- Question: Is it advantageous to change?

Review

Main topics from the last lectures

- The notion of random quantity, probability and independency
- What is a random variable and the distribution of a random variable
- The notion of expectation and its basic properties
- The notion of variance and its basic properties

The notion of expectation

- The idea of expectation can be easily understood for discrete variables taking a finite number of values
- X is a random variable taking n values, $x_1, x_2, \ldots, x_n$, with probabilities $p_1, p_2, \ldots, p_n$, respectively.
- The *expectation* or *expected value* of X is the sum of the possible values of X multiplied by their probabilities
- We use the symbol $\mathrm{E}(X)$ to denote the expectation of X and write

$$\mathrm{E}(X) = p_1 x_1 + p_2 x_2 + \cdots + p_n x_n \ .$$

The notion of expectation

Simple examples

- Example: binary trial
  X takes the values 0 and 1 with probabilities $(1-p)$ and p, respectively.
  The expectation of X is then,

$$\mathrm{E}(X) = (1-p)\,0 + p\,1 = p\ .$$

- Example: the binomial trial
  X takes the values $0, 1$, and 2
  with probabilities $(1-p)^2$, $2p(1-p)$ and p^2.
  The expectation is

$$\mathrm{E}(X) = (1-p)^2 0 + 2p(1-p)1 + p^2 2 = 2p\ .$$

- Remark: The random variable X represents the number of successes.

Basic Probability Theory

Expectation of continuous random variables

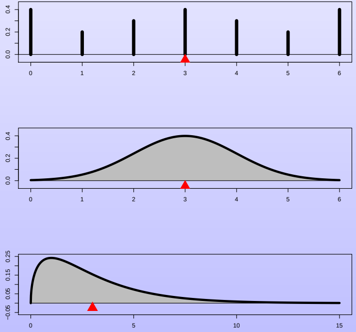
- If X continuous with density f then

$$E(X) = \int xf(x)dx$$

- The *expected value* of a continuous random variable X is the centre of mass of the graph of the density function

Basic Probability Theory

Physical interpretation of the expectation: Centre of gravity



Basic properties of the expectation

The expectation has the following basic properties:

- If the random variable X is equal to a constant c with probability 1, then $\mathrm{E}(X) = c$;
- If X and Y are random variables (with expectation well defined) and a, b are constants, then $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$;
- If X and Y are random variables (with expectation well defined) such that $X \leq Y$ with probability 1, then $\mathrm{E}(X) \leq \mathrm{E}(Y)$.
- (Jensens inequality) If ϕ is a convex real function and X is a random variable with finite expectation, then

$$\mathrm{E}\left\{\phi(X)\right\} \geq \phi\left\{\mathrm{E}(X)\right\} .$$

The notion of variance

The variance of a random variable X is defined by

$$\mathrm{Var}(X) = \mathrm{E}\left\{X - \mathrm{E}(X)\right\}^2 = \mathrm{E}(X^2) - \left\{\mathrm{E}(X)\right\}^2 \ .$$

Clearly, $\{X - \mathrm{E}(X)\}^2$ is a measure of the distance between the random variable X and its expectation.
Therefore, the expected value of this distance, i.e. the variance, is a measure of the dispersion of the data around its expected value.
The larger is the variance the more disperse is the data.

The variance of the binary variable X taking values 0 and 1 with probabilities $(1.p)$ and p is

$$\mathrm{Var}(X) = \mathrm{E}\left\{X - \mathrm{E}(X)\right\}^2 = \mathrm{E}\left\{X - p\right\}^2 = \cdots = \mathrm{E}(X^2) - p^2 \ .$$

To complete the calculation above we must compute the expectation of the random variable X^2. Note that $X^2 = X$, since X takes only the values 0 and 1. Therefore $\mathrm{E}(X^2) = \mathrm{E}(X)$. Replacing that in the last equation yields

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - p^2 = p - p^2 = p(1-p) \ .$$

The notion of covariance

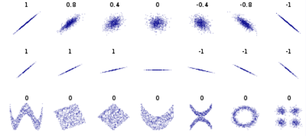
- The covariance of two random variables X and Y is defined by

$$\mathrm{Cov}(X, Y) = \mathrm{E}\left[\, \{X - \mathrm{E}(X)\}\, \{Y - \mathrm{E}(Y)\}\right] \ .$$

- The correlation of two random variables X and Y is defined by

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \ .$$

- If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$
  But, $\mathrm{Cov}(X, Y) = 0$ does not imply that X and Y are independent!



The notion of variance

The variance has the following basic properties:

1. If the random variable X is equal to a constant with probability 1, then $\text{Var}(X) = 0$;
2. If the random variable X has finite variance and b is a constant, then $\text{Var}(bX) = b^2\text{Var}(X)$;
3. If the random variables X and Y are independent, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$.
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

Three key distributions

- We studied three key distributions that will be the basic building blocks of (most of) the statistical models we will study
- Binomial distribution: study the occurrence of events, frequencies etc
- Poisson distribution: counting data
- Normal distribution: continuous measurements
- There are **many** other important distributions ...

Three key distributions:

The binomial Distribution

- Binomial distribution: Perform independently n times a basic binary trial with probability p of success and count the number of successes.
- Notation: $X \sim Bi(n, p)$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \ ,$$

  for $x = 0, 1, \ldots, n$.
- $E(X) = np$, $Var(X) = np(1-p)$
  The variance can be expressed as a function of the mean.

Binomial Distribution

The binomial coefficient

- For any set containing n elements, the number of distinct subsets each containing x elements of it that can be formed is given by the binomial coefficient ("n choose x")

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

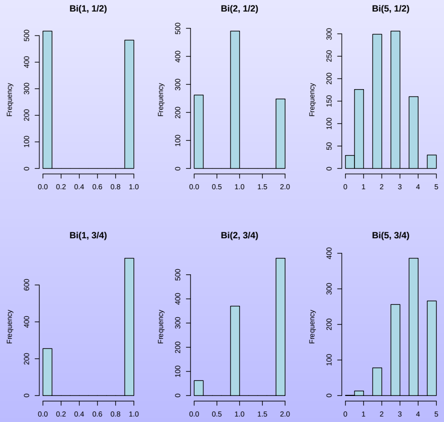
for $x = 0, 1, \ldots, n$.

- Curiosity: The binomial coefficient can be arranged to form the Pascal triangle

```
            1
        1       1
    1       2       1
  1     3       3     1
1   4       6       4   1
```

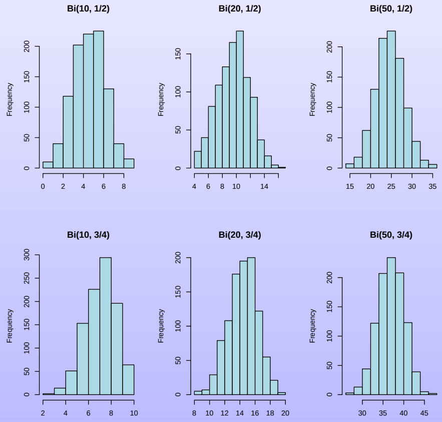
Three key distributions:

1000 simulations of the binomial distribution



Three key distributions:

1000 simulations of the binomial distribution



Three key distributions:

The Poisson distribution

- The Poisson distribution: describes the number of events
  (number of accidents, number of mutations in a fragment of DNA, number of worms in a portion of soil, etc.)
- This distribution was first used by Siméon-Denis Poisson
  Poisson, S.D., 1838. *Recherches sur la probabilité des jugements en matières criminelles et matière civile*
  (Study on the Probability of Judgments in Criminal and Civil Matters)
  to study the number of occurrences of an event during a time-interval of a given length, specifically the number of criminal and civil judgments
- The Poisson distribution takes positive integer values (*i.e.* 0,1,2, . . .) and depends on a single parameter, called the *intensity parameter* and usually denoted by λ

Three key distributions:

The Poisson distribution

- A random variable Y is said to follow a *Poisson distribution* with parameter λ ($\lambda > 0$) if

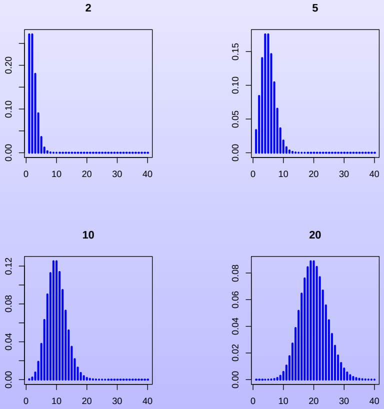
$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!},$$

for $y = 0, 1, 2, \ldots$.
Here $y! = y \cdot (y-1) \cdot \cdots \cdot 1$ and $0! = 1$.

- A Poisson variable takes only non-negative integer values. The Poisson distribution describes typically counts (but there exist many other distributions for counts!!!).
- Notation: $Y \sim Po(\lambda)$
- $E(Y) = Var(Y) = \lambda$

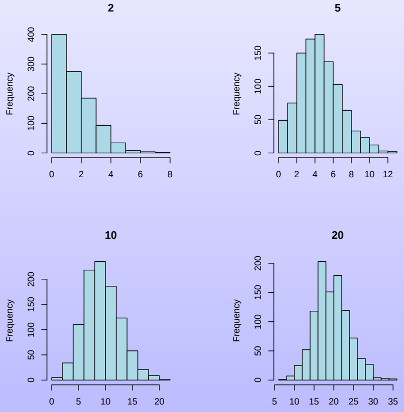
Three key distributions:

The probability function of the Poisson distribution



Three key distributions:

Simulated 1000 Poisson random variables



Three key distributions:

A classical example of Poisson distribution - Counts of alpha-particles

- Frequency of counts of alpha-particles emitted by the radioactive decay of a source of polonium, registered in time-intervals of 72 seconds

Counts:	0	1	2	3	4	5	6	7
Frequency:	57	203	383	525	532	408	273	139
Counts:	8	9	10	11	12	13	14	+ 15
Frequency:	45	27	10	4	0	1	1	0

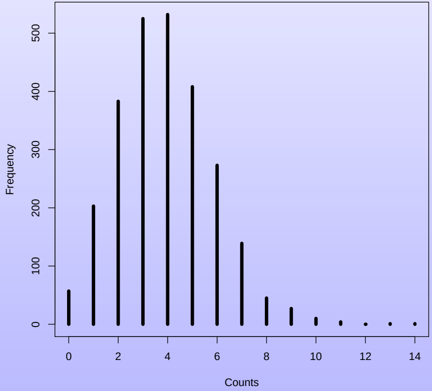
Rutherford, E. and Geiger, M. (1910).

- Mean of counts: 3.87
  Variance of counts: 3.74
- A reasonable estimate of λ is 3.87

(is the maximum likelihood estimate that we will study latter in this lecture)

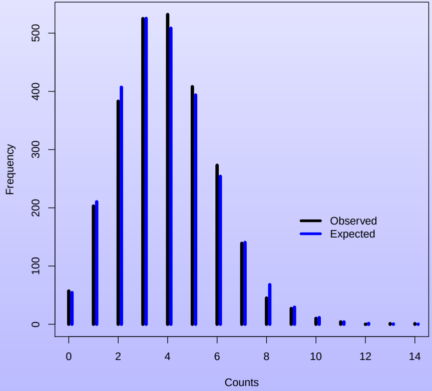
Three key distributions:

A classical example of Poisson distributed data - Counts of alpha-particles



Three key distributions:

A classical example of Poisson distributed data - Counts of alpha-particles



Three key distributions:

The normal distribution

- The normal distribution is one of the most used (and misused) distributions
- The normal distribution was used by Gauss to describe errors in astronomical measurements
  and is sometimes called the gaussian distribution
- The normal distribution was in fact used before Gauss by De Moivre and Laplace

Three key distributions:

The normal distribution

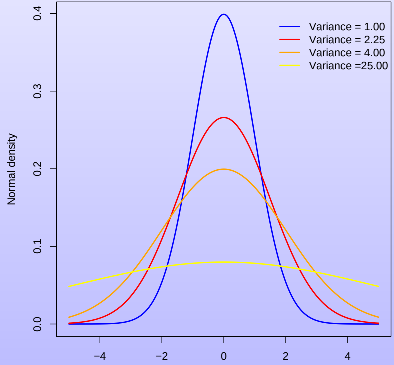
- Normal distribution: continuous distribution depending on two parameters, μ and σ^2 and probability density given by, for each real number x,

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

  Here μ is a real number and σ is a positive number ($\sigma > 0$).
- $E(X) = \mu$, $Var(X) = \sigma^2$
  The variance is not a function of the mean.

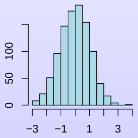
Three key distributions:

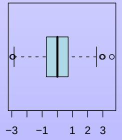
The density of the normal distribution



Three key distributions:

1000 simulated normal distributed variables



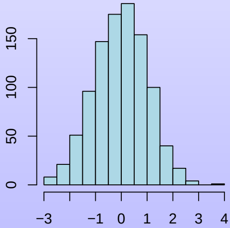


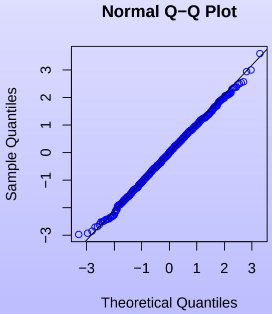
The Normal QQ-plot

- QQ-plot is a standard technique for informally checking the adjustment to a distribution
- Suppose you observe a sample (i.e. some values)

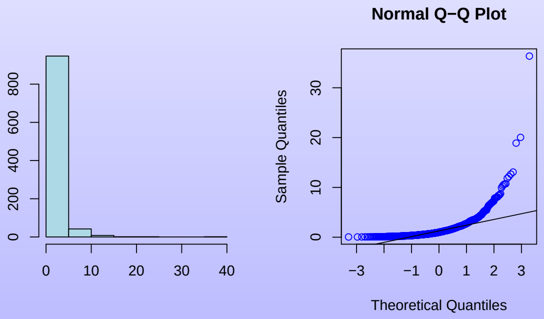
  The median is the value, say M, such that half of the observed values are smaller than M
- The 0.25-quantile is the value, say Q, such that 1/4=0.25 of the observed values are smaller than Q
- The α-quantile is the value, say Q, such that α of the observed values are smaller than Q
- The idea is to plot the observed (sample) quantiles against the quantiles one would expect for the putative distribution
- The assumed distribution adjusts well the data if, and only if, the QQ-plot is (approximately) linear (with a strait line crossing the origin and with steepness 1)

Normal QQ-plot of 1000 simulated normal observations

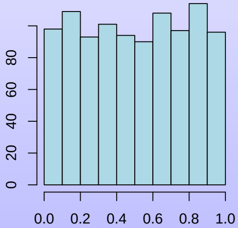


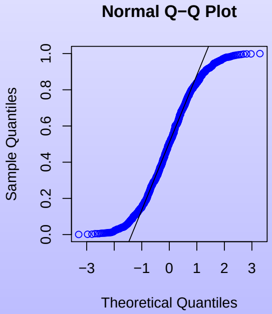


Normal QQ-plot of 1000 simulated **log**-normal observations



Normal QQ-plot of 1000 simulated uniform observations





Qq-plot: summary

- Qq-plot: plot the observed quantiles (sample quantiles or empirical quantiles) against the theoretical quantiles (theoretically calculated under the assumption that the data is normally distributed)
- In R use the functions: qqnorm (plots the qq-plot) and qqline (draws a reference line)
- x < − rnorm(1000)
  qqnorm (x)
  qqline (x)

The law of large numbers

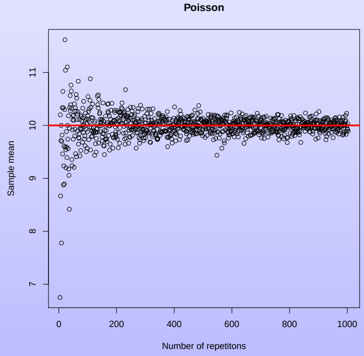
The general idea

Law of large numbers:
If we repeat independently many times an experiment generating the same random variable, then the mean of the observed values approximates the expectation of the random variable.
(under regularity conditions)

The law of large numbers

Means of Poisson simulated random variables, $\lambda = 10$,

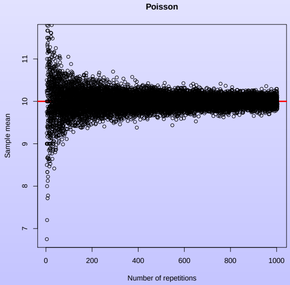
with different number of repetitions



The law of large numbers

Means of Poisson simulated random variables, $\lambda = 10$,

with different number of repetitions and 10 replicates for each number of repetitions



Question: Is this valid for other values of λ or for other distributions?

The law of large numbers

Precise formulation

- Suppose that $X_1, X_2, \ldots$ is a sequence of random variables independent and following the same distribution.
- We say that these random variables are *independent and identically distributed* (and some times denote that by *iid*).

- **Kolmogorov (strong) law of large numbers**:

  Let $X_1, X_2, \ldots$ be a sequence of iid random variables.
  If X_1 has finite expectation μ, then with probability 1

$$\frac{X_1 + \cdots + X_n}{n} \longrightarrow \mu$$

as $n \to \infty$ (*i.e.* as n increases arbitrarily).

The central limit theorem

- Consider $X_1, X_2, \ldots$ are independent and identically distributed random variables for which $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$, where $0 < \sigma^2 < \infty$
- Then the central limit theorem says that for n sufficiently large $X_1 + \cdots + X_n$ is approximately normally distributed!
- Equivalently, the central limit theorem says that, for n sufficiently large

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

follows approximately a standard normal distribution,

i.e. a normal distribution with mean 0 and variance 1, $N(0, 1)$

The central limit theorem

- Note that,

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma},$$

where $\bar{X} = 1/n \sum_{i=1}^{n} X_i$.

- The (sample) mean of the variables when properly standardized (i.e. subtracted the expectation and divided by the standard error and multiplying by $\sqrt{n}$) follows approximately a standard normal distribution.

Tutorials on the LLN and the CLT

- Tutorial 5 - Demonstration of the law of large numbers
- Tutorial 6 - Demonstration of the central limit theorem
- Tutorial 7 - Demonstration of the failure of the central limit theorem (if wrongly applied)
- Please, run the tutorials (after the lecture), modify the parameters used there and re-run ...

Statistical Models

Three simple examples

- Toss a coin twice and count the number of heads
  Discussed before in this course
  Natural to choose a binomial distribution
  Additional example of binomial model: Play a game several times
- The number of particles counted
  Rutherford, E. and Geiger, M. (1910).
  Can "deduce" the distribution to be a Poisson
  Additional example of Poisson model: Telomerase activity
- Weights of seeds of *Vicia graminea*
  the weights of 100 seeds were recorded
  Which distribution should we use here?

Statistical Models

A simple example (Oh no! again!)

- Recall one of our first examples: The binomial trial
  Toss a coin twice and count the number of tails
- We performed the experiment four times
  Result: 1, 2, 0, 1
- We can think on this result as four random variables
  Y_1, Y_2, Y_3 and Y_4
- In this execution of the experiment we observed that
  $Y_1 = 1$, $Y_2 = 2$, $Y_3 = 0$ and $Y_4 = 1$

Statistical Models

A simple example (Oh no! again!)

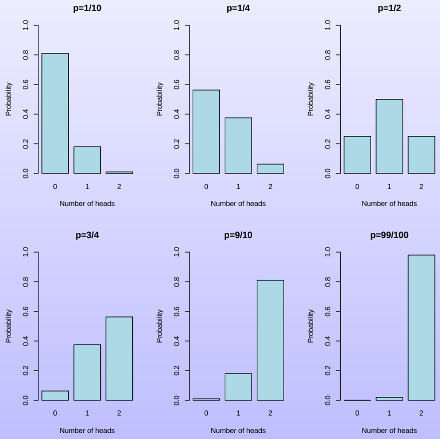
In general:

- Y_1, Y_2, Y_3 and Y_4 are independent
- Y_1, Y_2, Y_3 and Y_4 follow the same distribution
- The distribution of Y_1 (and also Y_2, Y_3 and Y_4) has probability function

$$f_{Y_1}(y) = P(Y_1 = y) = \begin{cases} (1-p)^2, \text{ if } y = 0, \\ 2p(1-p), \text{ if } y = 1, \\ p^2, \text{ if } y = 2\,. \end{cases}$$

  Here p can be any number between 0 and 1
- Each value of p determines a distribution (using the formula above),
- The class of all such distributions is a parametric statistical model and p is a parameter.

Statistical Models



Statistical Models

A model for the first example:

In summary:

- The results: 1, 2, 0, 1
  are viewed as realisations of four independent and identically distributed (iid) random variables
  Y_1, Y_2, Y_3 and Y_4
- $Y_1 \sim Bi(2, p)$
- In a short form:
  $Y_1, \ldots, Y_4$ iid
  $Y_i \sim Bi(2, p)$, for $i = 1, \ldots, 4$
- Here p is a parameter (to be estimated)

Statistical model

Another example of (binomial) statistical model

Data on 112 trials of a game (previous courses)

```
0 1 0 0 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 0
1 1 1 0 1 1 1 0 0 1 1 0 1 0 1 1 0 1 0 0
1 1 1 1 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 0
1 0 1 1 1 1 0 1 1 0 1 1 1 0 1 1 1 1 0 1
0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1 1
1 1 0 0 1 1 0 1 1 0 1 0
```

- Here 1="success" (get the check) and 0="failure" (don't get the check)
- 77 successes out of 112 trial
- Statistical model:
  The results are represented by 112 independent random variables,
  $X_1, X_2, \dots, X_{112}$,
  where $X_i \sim Bi(1, p)$, for $i = 1, \dots, 112$.
  Here p is a parameter (to be estimated).

Statistical Models

A classical example - Counts of alpha-particles

- Frequency of 10,097 counts of alpha-particles emitted by the radioactive decay of a source of polonium, registered in time-intervals of 72 seconds

Counts:	0	1	2	3	4	5	6	7
Frequency:	57	203	383	525	532	408	273	139
Counts:	8	9	10	11	12	13	14	+ 15
Frequency:	45	27	10	4	0	1	1	0

  Rutherford, E. and Geiger, M. (1910).

- Data: $2, 1, 3, 5, 3, 4, \ldots, 3$
- $Y_1, Y_2, \ldots Y_{10097}$
  are independent and identically distributed random variables representing each of the results (counts)
- Which distribution each of this random variables follows?

Statistical Models

Distribution of the alpha-particles counts

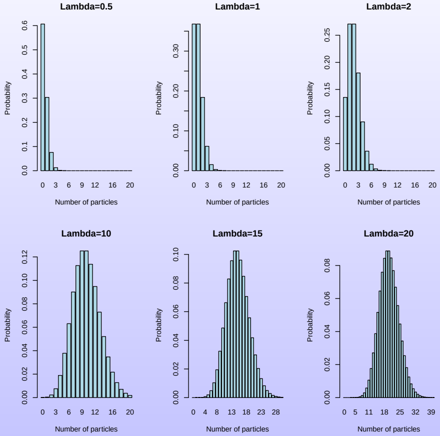
- Let us assume that
  - The time of arrival of a particle in the counter is homogeneously distributed in the observation interval
  - The number of particles that arrive in two disjoint intervals are independent
  - The particles do not arrive at the same time

    (the probability of two or more particles arrive in the counter in a short interval divided by the probability that only one particle arrives in this interval tends to zero as the length of the interval approaches zero)
- Under these assumptions it can be shown that the number of particles arriving in the counter is distributed according to a Poisson distribution! (formal prove involves a proper formulation of the problem as a stochastic process and the solution of a differential equation, not at the level of this course!)

Statistical Models

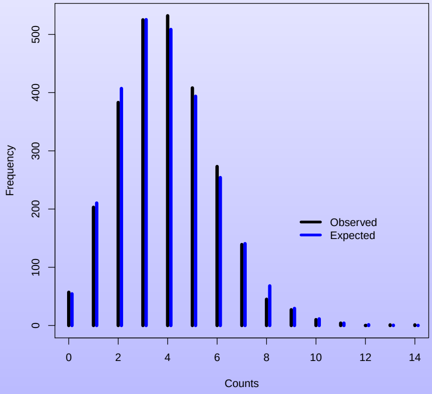
A classical example - Counts of alpha-particles

- Frequency of 10,097 counts of alpha-particles emitted by the radioactive decay of a source of polonium, registered in time-intervals of 72 seconds
- Data: $2, 1, 3, 5, 3, 4, \ldots, 3$
- $Y_1, Y_2, \ldots Y_{10097}$
  are independent and identically distributed random variables representing each of the results (counts)
- $Y_1 \sim Po(\lambda)$
- In short:
  $Y_1, Y_2, \ldots, Y_{10097}$ are independent
  $Y_i \sim Po(\lambda)$, for $i = 1, \ldots, 10097$
- Here λ is a parameter (to be estimated)


Lambda=0.5
Lambda=1
Lambda=2
Lambda=10
Lambda=15
Lambda=20
Probability
Number of particles

Counts of alpha-particles and

the expected number of counts under a Poisson distribution with $\lambda = 3.87$



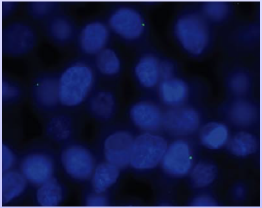
Statistical Models

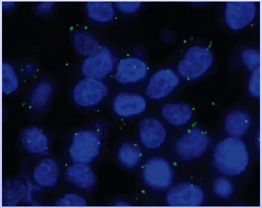
Additional example of Poisson model: Telomerase activity

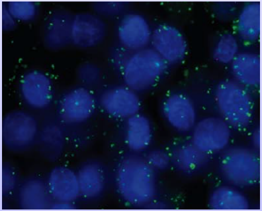
- Essay to classify a type of skin tumour:
  Benignant or malignant
- Samples of tumours are taken from a patient
- Telomerase (an enzyme) : marker of cell division activity
  Augmented telomerase activity is an evidence of cancer
- In the experimental setup used the
  telomerase activity produces luminescent spots
  that are observed in the microscope
- Large telomerase activity $\Longrightarrow$ Large number of signals (spots)
- Count the number of spots per microscope field

Statistical Models

Additional example of Poisson model: Telomerase activity





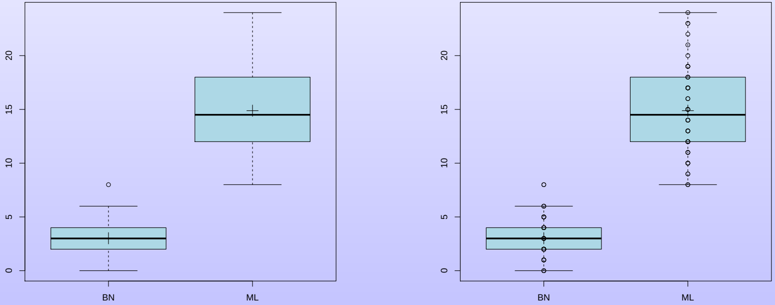


Statistical Models

Additional example of Poisson model: Telomerase activity

- In a pilot test:
  Counted the number of signals of 50 microscopic fields for a benignant tumour
  and
  Counted the number of signals of 50 microscopic fields for a malignant tumour
- Means: 3.02 and 14.88 for benignant and malignant, respectively
  Sample Variances: 3.20 and 16.88 for benignant and malignant, respectively
- We will assume the counts to be Poisson distributed
  (not necessarily with the same parameter for the two types of tumours)

Box-plot and box plot superposed with the dot-plot of the number of signals per type of tumour



Statistical Models

Additional example of Poisson model: Telomerase activity

- Statistical model for the data on the benignant tumour: The results are represented by 50 independent random variables,
  $X_{b1}, X_{b2}, \ldots, X_{b50}$, whith
  $X_{bi} \sim Po(\lambda_b)$, for $i = 1, \ldots, 50$
- Statistical model for the data on the malignant tumour: The results are represented by 50 independent random variables,
  $X_{m1}, X_{m2}, \ldots, X_{m50}$, whith
  $X_{mi} \sim Po(\lambda_m)$, for $i = 1, \ldots, 50$
- How would you describe a statistical model representing the entire data?

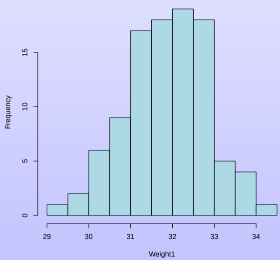
Statistical Models

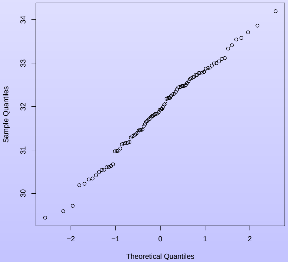
Weight seeds of *Vicia graminea*

- We recorded the weights of 100 seeds of *Vicia graminea*
  Automatic weight measurements
- Data:
  31.788, 32.475, 31.155, 29.444, ... , 30.543, 32.496, 31.130
- Sample mean = 31.90
  Sample variance = 0.959
- We assume the weights independent and identically distributed
- The results can be represented by 100 random variables $X_1, X_2, \ldots, X_{100}$
- Which type of distribution each of these random variables has?
  We look at the data!

Statistical Models

Weight seeds of *Vicia graminea*





Statistical Models

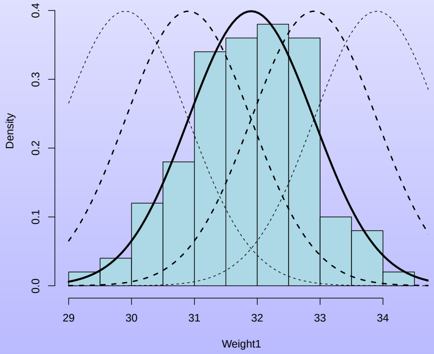
Weight seeds of *Vicia graminea*

- The weights of 100 seeds represented by 100 independent random variables
  $X_1, X_2, \ldots, X_{100}$
- $X_i \sim N(\mu, 1)$ for $i = 1, \ldots, 100$
- Each value of μ determines one distribution
  μ is a parameter indexing the distributions in the model
- Alternative model:
  $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, 100$

  Here the distributions of the model are indexed by **two** parameters μ and σ^2

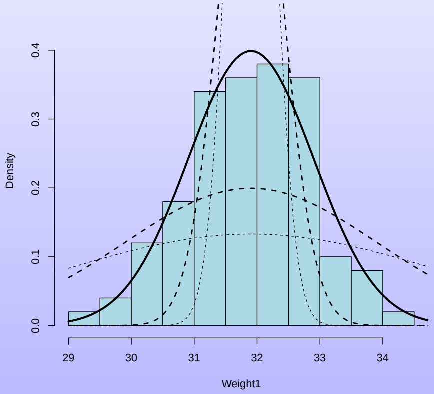
Statistical Models

Weight seeds of *Vicia graminea*



Statistical Models

Weight seeds of *Vicia graminea*



Statistical Models

Weight seeds of *Vicia graminea*

Model 1: $X_1, \ldots, X_{100}$ iid; $X_1 \sim N(\mu, 1)$
Model 2: $X_1, \ldots, X_{100}$ iid; $X_1 \sim N(\mu, \sigma^2)$

Questions for discussion:

1) Which of the two models would represent (fit) the data best?

2) Which of the two models gives a more compact representation of the data?

Statistical Models

General framework

- The probability law of the random quantity in study, Y, is unknown and we determine it in two steps:
- Choose a class of distributions that are good candidates for being the distribution of Y (or for well approximate the probability law of Y).
- This class of distributions is called the *statistical model*.
- The next step: determine the best candidate **in the parametric model** for representing the probability law of Y, on the basis of observations of the experiment.
- This step is called *parametric point estimation* or simply *(point) estimation*.

Statistical Models

- There many general techniques for deriving good point estimates.
- We concentrate only on maximum likelihood estimation.
- Maximum likelihood estimation produces good estimators in many cases (but, not always!) and is by far the most popular estimation method.

Parameter Estimation for a simple binomial model

A simpler particular case

- We considered the Master Quiz "experiment"
- Statistical model:
  The results are represented by 112 independent random variables,
  $X_1, X_2, \ldots, X_{112}$,
  where $X_i \sim Bi(1, p)$, for $i = 1, \ldots, 112$.
  Here p is a parameter (to be estimated).
- In order to make things simpler,
  we will work with the first four observations:
  0, 1, 0, 0
  (the other observations will be ignored for the moment,
  we will make calculate things for this case and then generalise)
- What is the probability of observing **this** result?

Parameter Estimation for a simple binomial model

The probability of observing a particular result

- What is the probability of observing $X_1 = 0$, $X_2 = 1$, $X_3 = 0$ and $X_4 = 0$?
- We assumed $X_1 \sim Bi(1, p)$, then $P(X_1 = 0) = 1 - p$ and
  $X_2 \sim Bi(1, p)$, then $P(X_2 = 1) = p$,
  $X_3 \sim Bi(1, p)$, then $P(X_3 = 0) = 1 - p$,
  $X_4 \sim Bi(1, p)$, then $P(X_4 = 0) = 1 - p$
- Since $X_1, \ldots, X_4$ are independent,
  $P\left(X_1 = 0, X_2 = 1, X_3 = 0 \text{ and } X_4 = 0\right)$ is

  $$P\left(X_1 = 0\right).P\left(X_2 = 1\right).P\left(X_3 = 0\right).P\left(X_4 = 0\right),$$

  which is

  $$(1 - p).p.(1 - p).(1 - p),$$

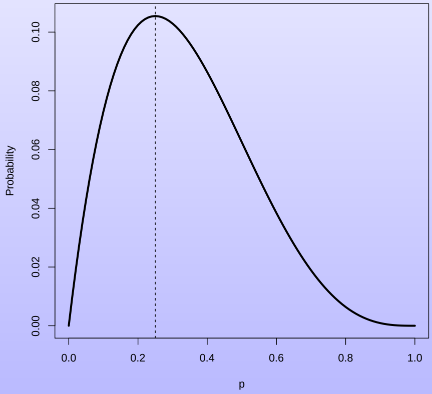
  or equivalently

  $$p(1 - p)^3$$

- This probability depends on the parameter p

Parameter Estimation for a simple binomial model

The probability of observing 0, 1, 0, 0 as a function of p



Parameter Estimation for a simple binomial model

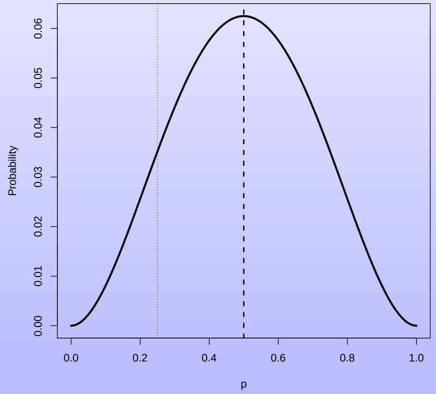
The probability of observing a particular result

- The probability of observing 0, 1, 0 , 0 is $p(1-p)^3$
- The probability of observing 1, 0, 0 , 0 is $p(1-p)^3$

- The probability of observing 1, 1, 0 , 0 is $p^2(1-p)^2$
- The probability of observing 1, 0, 1 , 0 is $p^2(1-p)^2$

Parameter Estimation for a simple binomial model

The probability of observing 1, 1, 0, 0 as a function of p



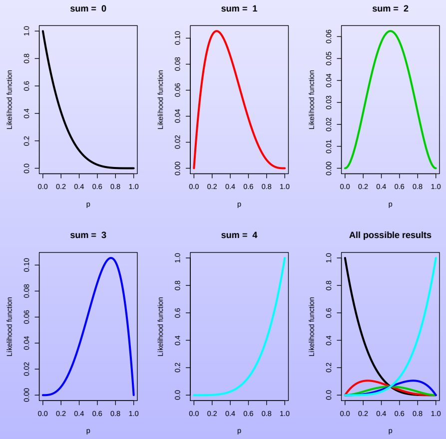
Parameter Estimation for a simple binomial model

The likelihood function

- Note that the form of the function that describes the probability as a function of p depends on the data
- The function that expresses the probability of observing a result to p depends on the sum of the results (number of successes observed)
- Notation x_+ sum of the results (eg. for the result 0,1, 0, 0 $x_+ = 0 + 1 + 0 + 0 = 1$)
  x_+ is the number of observed successes
- $P($ observing a particular result $) = p^{x_+}(1-p)^{4-x_+}$
- Viewing the observed results as fixed (we know the observed results) the function that expresses the probability of observing a result in terms of p is called the *likelihood function*
- In our example, $L(p) = p^{x_+}(1-p)^{4-x_+}$

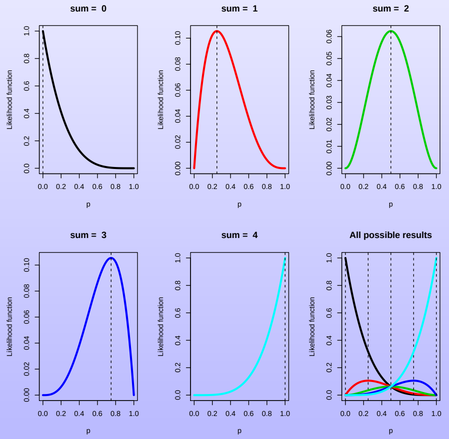
Parameter Estimation for a simple binomial model

The likelihood function



Parameter Estimation for a simple binomial model

The likelihood function and their maxima



Parameter Estimation for a simple binomial model

The idea of the maximum likelihood estimate

- For the results were we observed 1 success ($x_+ = 1$) out of 4 had a maximum at $p = 1/4$
- $x_+ = 0 \implies$ maximum at $0/4 = 0$
  $x_+ = 1 \implies$ maximum at $1/4$
  $x_+ = 2 \implies$ maximum at $2/4 = 1/2$
  $x_+ = 3 \implies$ maximum at $3/4$
  $x_+ = 4 \implies$ maximum at $4/4 = 1$
- Fisher's idea:
  Estimate the parameter p by the value that maximises the likelihood function

Parameter Estimation for a simple binomial model

The log-likelihood function

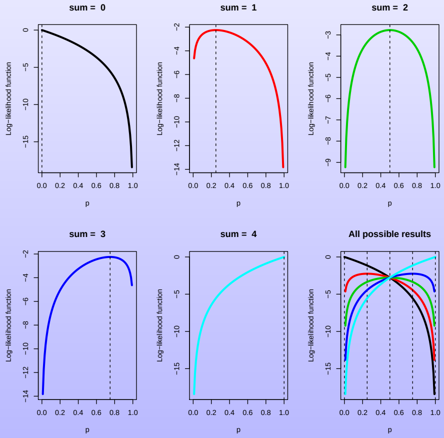
- There is a standard way to calculate the maximum likelihood estimate
- The task is to find $\hat{p}$ such $L(\hat{p})$ takes a maximum value
- Find the maximum of L is equivalent to find the maximum of

$$l(p) = \log\left(L(p)\right)$$

- The function l is called the *log-likelihood function*
- In our example,
$l(p) = \log\left[p^{x_+}(1-p)^{4-x_+}\right] = \cdots = [\log(p) - \log(1-p)]\, x_+ + 4\log(1-p)$

Parameter Estimation for a simple binomial model

The log-likelihood function and their maxima



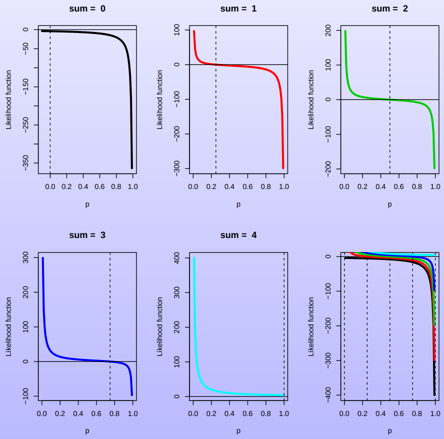
Parameter Estimation for a simple binomial model

The score function and the score equation

- We use an old trick to calculate a maximum: Differentiate and equate to zero
- In our example, $l(p) = \log\left[p^{x_+}(1-p)^{4-x_+}\right] = \cdots = [\log(p) - \log(1-p)]\, x_+ + 4\log(1-p)$
- The derivative (i.e. the inclination of a tangent of the graph of the function) of the function l is $S(p) = \frac{\partial}{\partial p} l(p) = \left[\frac{1}{p} + \frac{1}{1-p}\right] x_+ - \frac{4}{1-p}$
- The function S is called the *score function*
- Equating the score function yields, $\left[\frac{1}{\hat{p}} + \frac{1}{1-\hat{p}}\right] x_+ = \frac{4}{1-\hat{p}}$ which has solution $\hat{p} = \frac{x_+}{4}$
- Conclusion: in general $\hat{p} = \frac{x_+}{n}$ for this simple binomial model

Parameter Estimation for a simple binomial model

The score function and the m.l.e.



Parameter Estimation for a simple Poisson model

The statistical model

- $Y_1, \ldots Y_n$ iid $Y_1 \sim Po(\lambda)$
- The likelihood function for observations $y_1, \ldots y_n$ is
  $L(\lambda) = \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} \cdot \ldots \cdot \frac{e^{-\lambda}\lambda^{y_n}}{y_n!}$
- The log-likelihood is
  $l(\lambda) = -n\lambda + \log(\lambda)\sum_i y_i - \sum_{i=1}^n \log(y_i)$
- The score function is
  $S(\lambda) = \frac{\partial}{\partial\lambda} l(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^n y_i$
- Equating the score function to zero yields
  $\frac{1}{\hat{\lambda}}\sum_{i=1}^n y_i = n$
  which has solution $\hat{\lambda} = \frac{\sum_{i=1}^n y_i}{n}$
- The sample mean is the maximum likelihood estimate for λ

Parameter Estimation for a gaussian model

The statistical model

- $Y_1, \dots, Y_{100}$ iid, $Y_1 \sim N(\mu, 1)$
- The likelihood function is
  $L(\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-(y_1 - \mu)^2\right) \cdot \dots \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-(y_{100} - \mu)^2\right)$
- The log-likelihood function is
  $l(\mu) = n\frac{1}{\sqrt{2\pi}} - \sum_{i=1}^{n}(y_i - \mu)^2$
- The score function is
  $S(\mu) = \dots = -2\sum_{i=1}^{n}(y_i - \mu)$
  equating to zero yields,
  $0 = S(\hat{\mu}) = -2\sum_{i=1}^{n}(y_i - \hat{\mu})$
  which has solution

$$\hat{\mu} = 1/n \sum_{i=1}^{n} y_i$$

Parameter Estimation

Summing up

- We used the same procedure to estimate a parameter in a statistical model
- First we calculate the probability of the observed data as a function of the parameter
- This function is called the likelihood function
- We then found the value of the parameter that maximizes the likelihood function
- Note that there exist other general techniques for obtaining estimates of parameters

The Master Quiz problem

- Three boxes:
  One contains a BIG check, the other two are empty
- You choose one box,
  before you open the box the Master-Quiz says
  '*I give you a hint, the check is* **not** *here*'
  and he opens one of the remaining boxes, which is empty
- The Master-Quiz continues:
  Would you like to change and choose the other closed box?

- Question: Is it advantageous to change?

Three types of arguments

- Theoretical evidences: change boxes!

  (but, who cares to academic reflections?)

- Concensual evidences: It doesn't matter to change, don't change.

  (the majority uses to say that, but I didn't!)

- Experimental evidences:
  5 essays changing $\longrightarrow$ 3 successes
  5 essays not changing $\longrightarrow$ 1 success.

  (can we conclude with this data? Need more data to be convinced?)

Data on 112 trials changing boxes (previous courses)

0	1	0	0	1	1	1	1	1	1	0	0	1	1	1	1	0	1	0	0
1	1	1	0	1	1	1	0	0	1	1	0	1	0	1	1	0	1	0	0
1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	1	1	0
1	0	1	1	1	1	0	1	1	0	1	1	1	0	1	1	1	1	0	1
0	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1
1	1	0	0	1	1	0		1	1	0	1	0							

- 77 successes out of 112 trial
  $\hat{p} = 77/112 = 0.6875$
- Convinced that the probability of getting the check is larger than $1/2$?
- Can this result be explained by mere random fluctuation?

- Changing box strategy $\rightarrow 77/112 = 0.6875$
  $1/2 = 0.5000$
  $2/3 = 0.6666\ldots$
- Hypothesis: "$p = 1/2$" (I will call that the **null hypothesis**)
  Alternative hypothesis: "$p \neq 1/2$" (the negation of the null hypothesis)
- **Whether** the null hypothesis
  or the alternative hypothesis is correct.
- Want to decide, on the basis of the data available, which hypothesis is correct.

A technical parenthesis

- What is the probability of throwing 1 time a fair coin and get a tails? Answer: $1/2$
- What is the probability of throwing 2 times a fair coin and get two tails? Answer: $1/2x1/2 = 1/4 = 1/2^2$
- What is the probability of throwing 112 times a fair coin and get 112 tails? Answer: $1/2^{112} = 1.92593x10^{-34}$
- What is the probability of throwing 112 times a fair coin and get exactly one head (*i.e.* 111 heads)? Answer: $1/2x1/2^{111}x112 = 2.157042x10^{-34}$

$X \sim Bi(112, 1/2)$, the probabilities $P(X = x)$ are given below

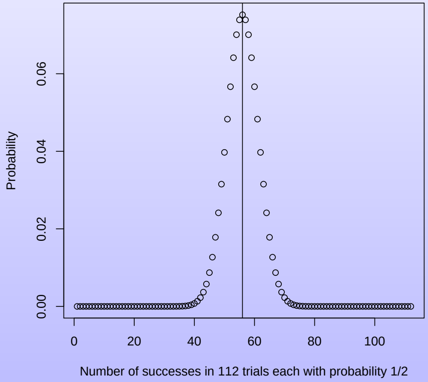
```
  [1] 1.925930e-34 2.157042e-32 1.197158e-30 4.389580e-29 1.196160e-27 2.583707e-26
  [7] 4.607610e-25 6.977238e-24 9.157625e-23 1.058214e-21 1.089961e-20 1.010691e-19
 [13] 8.506649e-19 6.543576e-18 4.627243e-17 3.023132e-16 1.832774e-15 1.034978e-14
 [19] 5.462385e-14 2.702443e-13 1.256636e-12 5.505263e-12 2.277177e-11 8.910692e-11
 [25] 3.304382e-10 1.163142e-09 3.892053e-09 1.239691e-08 3.763348e-08 1.090073e-07
 [31] 3.015869e-07 7.977460e-07 2.019295e-06 4.895259e-06 1.137428e-05 2.534839e-05
 [37] 5.421740e-05 1.113655e-04 2.198003e-04 4.170569e-04 7.611288e-04 1.336617e-03
 [43] 2.259518e-03 3.678286e-03 5.768221e-03 8.716423e-03 1.269566e-02 1.782795e-02
 [49] 2.414201e-02 3.153242e-02 3.973085e-02 4.830025e-02 5.665991e-02 6.414330e-02
 [55] 7.008249e-02 7.390517e-02 7.522491e-02 7.390517e-02 7.008249e-02 6.414330e-02
 [61] 5.665991e-02 4.830025e-02 3.973085e-02 3.153242e-02 2.414201e-02 1.782795e-02
 [67] 1.269566e-02 8.716423e-03 5.768221e-03 3.678286e-03 2.259518e-03 1.336617e-03
 [73] 7.611288e-04 4.170569e-04 2.198003e-04 1.113655e-04 5.421740e-05 2.534839e-05
 [79] 1.137428e-05 4.895259e-06 2.019295e-06 7.977460e-07 3.015869e-07 1.090073e-07
 [85] 3.763348e-08 1.239691e-08 3.892053e-09 1.163142e-09 3.304382e-10 8.910692e-11
 [91] 2.277177e-11 5.505263e-12 1.256636e-12 2.702443e-13 5.462385e-14 1.034978e-14
 [97] 1.832774e-15 3.023132e-16 4.627243e-17 6.543576e-18 8.506649e-19 1.010691e-19
[103] 1.089961e-20 1.058214e-21 9.157625e-23 6.977238e-24 4.607610e-25 2.583707e-26
[109] 1.196160e-27 4.389580e-29 1.197158e-30 2.157042e-32 1.925930e-34
```

- We can use the formula below with $n = 112$ and $p = 1/2$

$$P(X = x) = \frac{x!}{n!(n-x)!} p^x (1-p)^{n-x} \text{ , for } x = 0, 1, \ldots, n.$$

- R-command: dbinom(x=0:112, size=112, prob=1/2)

Probability function



- General idea:
  Reject the null hypothesis ($p = 1/2$) when
  the relative frequency of successes is far from $1/2$
- First proposal:
  Reject the null hypothesis when
  the number of successes is more than $56 + 1 = 56$ or
  the number of successes is less than $56 - 1 = 55$
- What is the probability of wrongly rejecting the null hypothesis when the null hypothesis is actually true?

  (this is the probability of making the so called type 1 error)
- We can use the formula below with $n = 112$ and $p = 1/2$

$$P(X = x) = \frac{n!}{x!(n-x)!}\, p^x(1-p)^{n-x}\ ,\ \text{for } x = 0, 1, \ldots, n.$$

$X \sim Bi(112, 1/2)$, the probabilities $P(X \leq x)$ are given below

```
  [1] 1.925930e-34 2.176301e-32 1.218921e-30 4.511472e-29 1.241275e-27 2.707834e-26
  [7] 4.878393e-25 7.465077e-24 9.904132e-23 1.157256e-21 1.205686e-20 1.131260e-19
 [13] 9.637909e-19 7.507367e-18 5.377980e-17 3.560930e-16 2.188867e-15 1.253865e-14
 [19] 6.716250e-14 3.374068e-13 1.594043e-12 7.099305e-12 2.987107e-11 1.189780e-10
 [25] 4.494161e-10 1.612558e-09 5.504612e-09 1.790152e-08 5.553500e-08 1.645423e-07
 [31] 4.661292e-07 1.263875e-06 3.283170e-06 8.178429e-06 1.955271e-05 4.490110e-05
 [37] 9.911850e-05 2.104840e-04 4.302842e-04 8.473411e-04 1.608470e-03 2.945086e-03
 [43] 5.204605e-03 8.882891e-03 1.465111e-02 2.336753e-02 3.606319e-02 5.389114e-02
 [49] 7.803315e-02 1.095656e-01 1.492964e-01 1.975967e-01 2.542566e-01 3.183999e-01
 [55] 3.884824e-01 4.623875e-01 5.376125e-01 6.115176e-01 6.816001e-01 7.457434e-01
 [61] 8.024033e-01 8.507036e-01 8.904344e-01 9.219668e-01 9.461089e-01 9.639368e-01
 [67] 9.766325e-01 9.853489e-01 9.911171e-01 9.947954e-01 9.970549e-01 9.983915e-01
 [73] 9.991527e-01 9.995697e-01 9.997895e-01 9.999009e-01 9.999551e-01 9.999804e-01
 [79] 9.999918e-01 9.999967e-01 9.999987e-01 9.999995e-01 9.999998e-01 9.999999e-01
 [85] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
 [91] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
 [97] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
[103] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
[109] 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00 1.000000e+00
```

- R-command: pbinom(q=0:112, size=112, prob=1/2)

- First proposal for a rejection rule:
  Reject the null hypothesis when
  the number of successes is more than $56 + 1 = 57$ or
  the number of successes is less than $56 - 1 = 55$
- Probability of wrongly rejecting the null hypothesis when the null hypothesis is actually true:
  $P(X < 55) + P(X > 57) = 2P(X \leq 54) = 2 * 0.3884824 = 0.7769648$
- That is, even if the null hypothesis is true,
  we would wrongly reject it in around 77% of the cases!
  We have been too strict!
- New rule:
  Reject the null hypothesis when
  the number of successes is more than $56 + 2 = 58$ or
  the number of successes is less than $56 - 2 = 54$

- New rule:
  Reject the null hypothesis when
  the number of successes is more than $56 + 2 = 58$ or
  the number of successes is less than $56 - 2 = 54$

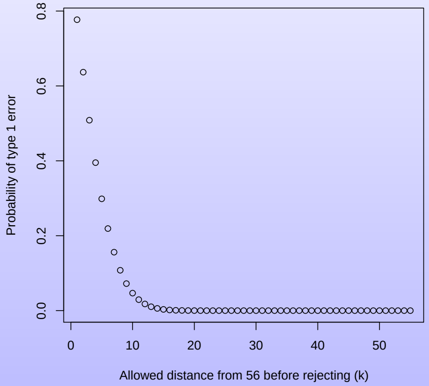
- Probability of wrongly rejecting the null hypothesis when the null hypothesis is true:
  $P(X < 54) + P(X > 58) = 2P(X \leq 53) = 2 * 0.3183999 = 0.6367998$

- Better, but still more than half of the cases with wrong rejection (under the null hypothesis) !

- General rule:
  Reject the null hypothesis when
  the number of successes is more than $56 + k$ or
  the number of successes is less than $56 - k$,
  for $k = 1, 2, 3, \ldots$.

- Probability of wrongly rejecting the null hypothesis
  when the null hypothesis is true:
  $P(X < 56 - k) + P(X > 56 + k) = 2P(X \leq 56 - k - 1)$
  $P(\text{type 1 error})$.

Probability of error type 1



- General rule:
  Reject the null hypothesis when
  the number of successes is more than $56 + k$ or
  the number of successes is less than $56 - k$,
  for $k = 1, 2, 3, \ldots$.
- Probability of wrongly rejecting the null hypothesis
  when the null hypothesis is true:
  $P(X < 56 - k) + P(X > 56 + k) = 2P(X \leq 56 - k - 1)$

  P(type 1 error).
- The probability of type 1 error decreases (quickly) with k.
- Idea: choose k large enough in order to make the probability of type 1 error small.
- Convention: Fix the probability of type 1 error
  $\alpha = P(\text{ type 1 error }) = 0.1$ or 0.05 or $0.01 \ldots$
- Using $k = 10$ yields a probability of type 1 error of
  $0.04673507 \approx 0.05$

- Null hypothesis:
  The probability of getting the check when changing the box is 1/2
  Alternative hypothesis:
  The probability of getting the check when changing the box is not 1/2
- Rejection rule:
  Reject the null hypothesis when the number of successes
  is smaller than 46 or larger than 66.
- This rejection rule implies that the probability of rejecting the null hypothesis when the null hypothesis is true is 0.047 i.e. approx. 5% .
- We observed 77 successes, and therefore reject the null hypothesis!
  Conclusion: the probability of success when changing box is **not** 1/2.

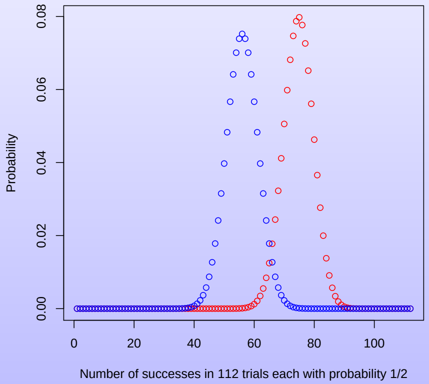
- Another question:
  Is the probability of getting the check equal to 2/3?

  (as I claimed)
- Null hypothesis: $p = 2/3$
  Alternative hypothesis: $p \neq 2/3$
- The probability law (under the null hypothesis) changes!
  We can use the formula for the binomial distribution with $n = 112$ and $p = 2/3$

$$P(X = x) = \frac{x!}{112!(112-x)!}(2/3)^x(1-2/3)^{112-x}$$

for $x = 0, 1, \ldots, 112$.

Probability function



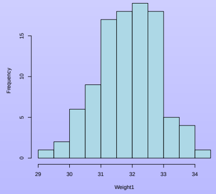
Blue $\rightarrow p = 1/2$ Red $\rightarrow p = 2/3$

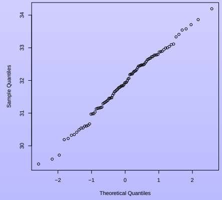
- Rejection rule:
  Reject the null hypothesis when the number of successes is smaller than 71 or larger than 91.
- Probability of rejection (under the null hypothesis):
  $P(X < 71) + P(X > 91) = F(70) + 1 - F(90) = 0.04346528$
- Conclusion:
  We do not have evidences to reject the null hypothesis, i.e. to reject that the probability of getting the check is $2/3$.

The example of *Vicia graminea*

Revisited (partial data)

- We recorded the weights of 100 seeds of *Vicia graminea*
- Statistical model:
  The results can be represented by 100 independent random variables $X_1, X_2, \ldots, X_{100}$, with $X_i \sim N(\mu, 1)$, for $i = 1, \ldots, 100$
- μ is a parameter indexing the distributions in the model

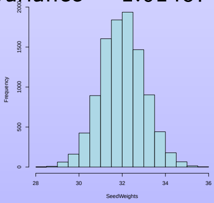


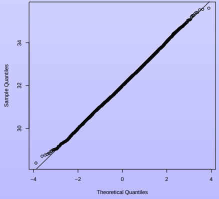


The example of *Vicia graminea*

Revisited and **enlarged** (complete data)

- We recorded in fact the weights of 10,000 seeds of *Vicia graminea*
- Statistical model:
  The results can be represented by 10,000 independent random variables $X_1, X_2, \ldots, X_{10000}$, with
  $X_i \sim N(\mu, 1)$, for $i = 1, \ldots, 10000$
  μ is a parameter indexing the distributions in the model
- Sample mean = m.l.e. for $\mu = 32.00303$
  Sample variance = 1.01487





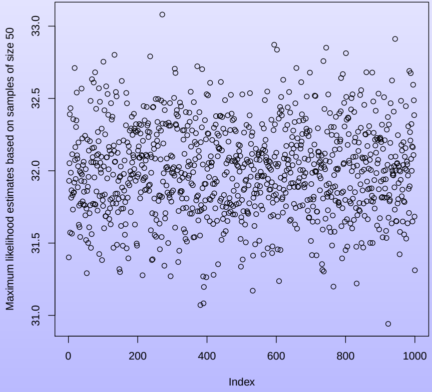
The example of *Vicia graminea*

An "experiment" on the behaviour of estimates

- Having so many observations (10,000 !!!)
  why not try to make an experiment on estimates
- Idea:
  Divide the observations in smaller non-overlaping subgroups,
  say of 50 observations
  and make estimation based on each of these groups
- Doing that we would obtain 200 estimates
  (one for each of the 200 non-overlaping subgroups of 50 observations)
- If we had really really good estimates,
  we should get the same result each time (???)
- We use the "best" estimate: the maximum likelihood estimate
  (*i.e.* the sample mean)

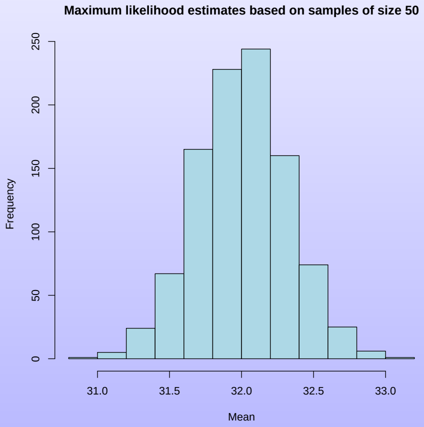
The example of *Vicia graminea*

500 estimates based on samples of size 50



The example of *Vicia graminea*

500 estimates based on samples of size 50



The example of *Vicia graminea*

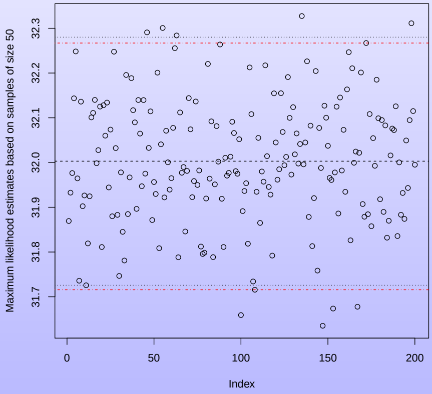
Results of the study on the behaviour of estimates

- The estimated values were not constant, but oscillated in a certain range
- 95% of the estimates where between 31.7157 and 32.26705 (an interval that contained the value 32.00303, i.e. the m.l.e. using the whole dataset)
- The interval [31.7157 , 32.26705] gives an idea of how much the estimate we used oscillates if we used 50 observations to estimate
- But, this way to evaluate the quality of the estimate is **not** feasible in practice (it is not aways that we have 10,000 observation to play with É)

  Therefore, we make some theoretical calculations that will yield an interval of this type

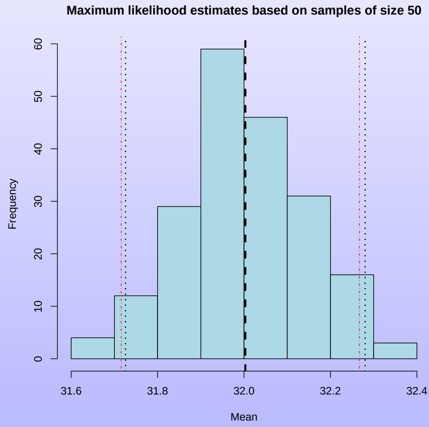
The example of *Vicia graminea*

500 estimates based on samples of size 50



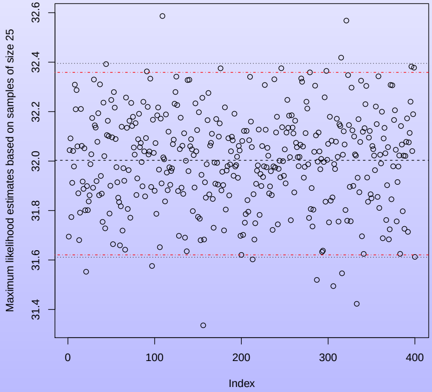
The example of *Vicia graminea*

500 estimates based on samples of size 50



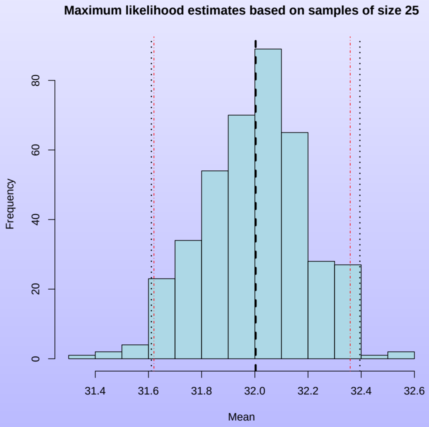
The example of *Vicia graminea*

500 estimates based on samples of size 25



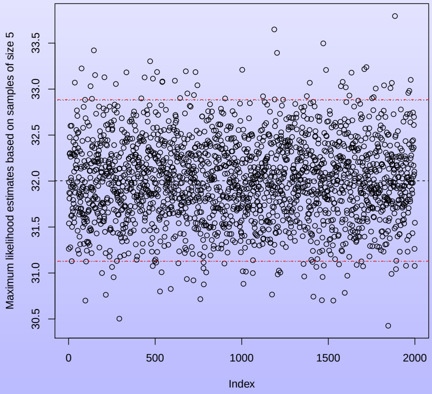
The example of *Vicia graminea*

500 estimates based on samples of size 25



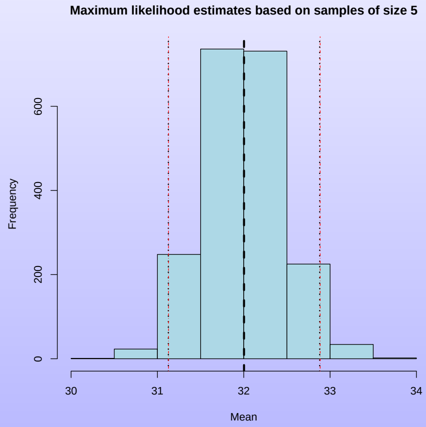
The example of *Vicia graminea*

2,000 estimates based on samples of size 5



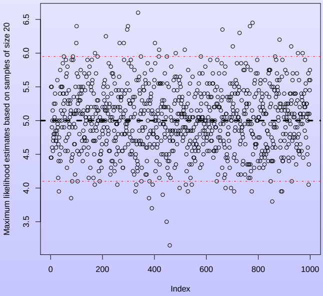
The example of *Vicia graminea*

2,000 estimates based on samples of size 5

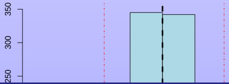


An experiment on the behaviour of estimates

Simulated Poisson data: 100 simulations, of sample size 20, $\lambda = 5$

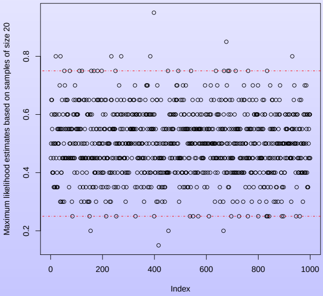


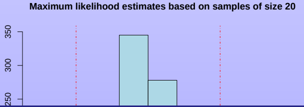
Maximum likelihood estimates based on samples of size 20



An experiment on the behaviour of estimates

Simulated binomial data: 100 simulations, of sample size 20, $p = 1/2$





An experiment on the behaviour of estimates

Concluding:

- The estimated values are typically not constant,
  but oscillated in a certain range
  (in fact estimates are random quantities, they depended on the data)

- It is possible to study the range of variation of estimates,
  provided we have many repetitions of the experiment
  (or we have many observations and we artificially split the data in non-overlapping subsets)

- I claim that it is possible (in many situations) to calculate a theoretical interval that contains the true value of the parameter with a high probability (a probability that we pre-specify).
  This interval is called a *confidence interval*.

Confidence Intervals

The general idea

- Idea:
  Find a region around the estimate such that the probability that the region contains the actual value of the parameter is high.

- The probability that the region contains the parameter is pre-fixed and is called the *coverage probability*
  Typical values used: 0.90, **0.95**, 0.99

- If this region is an interval, we call it a *confidence interval*.

Confidence Intervals for a Normal Sample with Known Variance

- We consider first the situation where the data arises from a single normal distribution with known variance σ_0^2, but with an unknown expected value μ (to be estimated)
- This is the situation that we encountered in the example of the weights of *Vicia graminea*
  (there we assumed $\sigma_0^2 = 1$)
- In symbols: $X_1, \ldots, X_n$ iid, $X_1 \sim N(\mu, \sigma_0^2)$
- I claim that

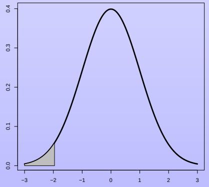
$$\left[\bar{X} - \frac{1.96\sigma_0}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma_0}{\sqrt{n}}\right]$$

  is a confidence interval for the mean μ with a coverage $\alpha = 0.95$
  (*i.e.* 95%)

Confidence Intervals for a Normal Sample with Known Variance

The z_α values

- The $\alpha -$ *quantil* of the standard normal distribution is the number z_α such that if $X \sim N(0,1)$, then $P(X \leq z_\alpha) = \alpha$
- $\int_{-\infty}^{z_\alpha} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) dx = \alpha$
- In R use the function "qnorm"
- $z_{0.025} = 1.96$



Confidence Intervals for a Normal Sample with Known Variance

- $X_1, \ldots, X_n$ iid, $X_1 \sim N(\mu, \sigma_0^2)$
- $\bar{X} = \frac{1}{n}(X_1 + \ldots X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$
- $$\left[\bar{X} - \frac{1.96\sigma_0}{\sqrt{n}}, \bar{X} + \frac{1.96\sigma_0}{\sqrt{n}}\right]$$
  is a confidence interval for the mean μ with a coverage of 0.95 (*i.e.* 95%)
- In general
  $$\left[\bar{X} - \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}\right]$$
  is a confidence interval for the mean μ with a coverage of α

Confidence Intervals for a Normal Sample with Known Variance

Preliminary calculations

- $X_1, \ldots, X_n$ iid, $X_1 \sim N(\mu, \sigma^2)$
- $\bar{X} = \frac{1}{n}(X_1 + \ldots X_n) = \frac{1}{n}\sum_{i=1}^n X_i$
- $\sum_{i=1}^n X_i \sim N(n\mu, n\sigma^2)$
- $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n)$
- Conclusion:

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

Confidence Intervals for a Normal Sample with Known Variance

Constructing the confidence interval

- $$\begin{aligned} P\left(\bar{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} \leq \mu\right) &= P\left(\bar{X} - \mu \leq \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right) \\ &= P\left(\frac{\sqrt{n}\,(\bar{X} - \mu)}{\sigma} \leq z_{1-\alpha/2}\right) \\ &= 1 - \alpha/2 \end{aligned}$$
- Therefore, $P\left(\mu < \bar{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right) = 1 - (1 - \alpha/2) = \alpha/2$
- Analogously, $P\left(\mu > \bar{X} + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}}\right) = \alpha/2$
- Therefore:

$$P\left(\mu < \bar{X} - \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} \text{ and } \bar{X} + \frac{\sigma z_{1-\alpha/2}}{\sqrt{n}} < \mu\right) = \alpha/2 + \alpha/2 = \alpha$$

Confidence Intervals for a Normal Sample with Known Variance

A general form of the CI

- In general
$$\left[\bar{X} - \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}\right]$$
is a confidence interval for the mean μ with a coverage of α

- For coverage 0.95 use $z_{1-\alpha/2} = z_{1-0.95/2} = z_{0.025} = 1.96$
For coverage 0.95 use $z_{1-\alpha/2} = z_{1-0.90/2} = z_{0.05} = 1.64$
For coverage 0.99 use $z_{1-\alpha/2} = z_{1-0.99/2} = z_{0.005} = 2.58$
- In R we use
qnorm(p=0.005, lower.tail=F)

Confidence Intervals for a Normal Sample with Known Variance

The example revisited:

- In the example of the weight of seeds we had:
  Sample mean = m.l.e. for μ = 32.00303
  Sample variance = 1.01487
  $n = 10,000$
- Assuming the variance $\sigma_0 = 1$ a confidence interval with coverage 0.95 for μ is

$$\left[32.00303 - \frac{1.96}{\sqrt{10,000}}, 32.00303 + \frac{1.96}{\sqrt{10,000}}\right] = [31.98343, 32.02263]$$

- **Interpretation:**
  We have evidence that the value of the average (μ) is contained in the interval $[31.98343, 32.02263]$
  with probability 0.95

The t- and the χ^2-distributions

- Suppose that $X_1, \ldots, X_k$ are iid with $X_1 \sim N(0,1)$, then $X_1^2 + \cdots + X_k^2$ follows a known distribution called the *Chi-square distribution* with k degrees of freedom
- Suppose that $X \sim N(0,1)$ and Z is chi-square distributed with k degrees of freedom, then $\frac{Z}{\sqrt{X/k}}$ has a known distribution called the *t-distribution* with k degrees of freedom
- There are tables for the t- and the χ^2-distributions

Confidence Intervals for a Normal Sample with Unknown Variance

A general form of the CI

- In the case where the variance was known the CI was of the form

$$\left[\bar{X} - \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}, \bar{X} + \frac{z_{1-\alpha/2}\sigma_0}{\sqrt{n}}\right]$$

- When we do not know the variance, we replace it by an estimate. For a sample $X_1, \ldots, X_n$ we use the sample variance s^2 given by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$
- In general

$$\left[\bar{X} - \frac{t_{n-1}(1-\alpha/2)s}{\sqrt{n}}, \bar{X} + \frac{t_{n-1}(1-\alpha/2s}{\sqrt{n}}\right]$$

where $t_{n-1}(1-\alpha/2)$ is the $1-\alpha/2$-quantile of the t-distribution with $n-1$ degrees of freedom

(we consult a table or use qt(p, df, lower.tail = F) in R)

Confidence Intervals for a Normal Sample with Unknown Variance

The example revisited:

- In the example of the weight of seeds, we had:
  Sample mean = m.l.e. for $\mu = 32.00303$
  Sample variance = 1.01487
  $n = 10,000$
- Using $t_{9999}(1 - 0.025) = 1.960201$
  and $s = \sqrt{1.01487}$
- A confidence interval with coverage 0.95 for μ is

$$\left[32.00303 - \frac{1.96020 * \sqrt{1.01487}}{\sqrt{10,000}}, 32.00303 + \frac{1.96020 * \sqrt{1.01487}}{\sqrt{10,000}}\right] = [31.98328, 32.02278]$$

  which is close to [31.98343, 32.02263]

- **Interpretation:**
  We have evidences that the value of the average (μ) is contained in the interval $[31.98328, 32.02278]$
  with probability 0.95

Summary and Practice

What you should know

- The idea of the central limit theorem and the law of large numbers
- The notion of parametric statistical models
- The idea of the maximum likelihood estimate
- The idea behind hypotheses tests and the interpretation of the results of a test
- The idea of confidence intervals

Summary and Practice

Tutorials on the LLN and the CLT

- Tutorial 4 - On the normal distribution
- Tutorial 5 - Demonstration of the law of large numbers
- Tutorial 6 - Demonstration of the central limit theorem
- Tutorial 7 - Demonstration of the failure of the central limit theorem (if wrongly applied)
- Tutorial 8 - Confidence intervals based on the normal distribution
- Tutorial 9 - Simple hypotheses tests based on the normal distribution
- Please, run the tutorials, modify the parameters used there and re-run ...