Ph.d. course in Basic Biostatistics – Day 1

Erik Parner, Department of Biostatistics, Aarhus University©

Structure of the course

Introduction: An example of possible time trend.

Statistical inference – the components
- Estimation
- Confidence intervals
- Statistics test and p-value

The normal (Gaussian) distribution
- Prediction intervals

Statistical models

Model validation

Analysis of continuous data – one sample

The structure of the course

The course will consider statistical analysis of different type of data in different settings:

Data to analyse	Type of analysis	Unpaired/Paired	Type	Day
Continuous	**One sample mean**	**Irrelevant**	**Parametric**	**Day 1**
			Nonparametric	Day 3
	Two sample mean	Non-paired	Parametric	Day 2
			Nonparametric	Day 2
		Paired	Parametric	Day 3
			Nonparametric	Day 3
	Regression	Non-paired	Parametric	Day 5
	Several means	Non-paired	Parametric	Day 6
			Nonparametric	Day 6
Binary	One sample mean	Irrelevant	Parametric	Day 4
	Two sample mean	Non-paired	Parametric	Day 4
		Paired	Parametric	Day 4
	Regression	Non-paired	Parametric	Day 7
Time to event	One sample: Cumulative risk	Irrelevant	Nonparametric	Day 8
	Regression: Rate/hazard ratio	Non-paired	Semi-parametric	Day 8

The document "Overview.pdf" gives more details.

Practical things

3 hours lecture and 3 hours exercise.

All analysis presented in the lecture at Day 1 are given in the Stata do-file "Day1.do", ect.

Standard analysis and formulation of the topic at Day 1 is given in the pdf-file: Standard1-1.pdf, ect.

Exercises in "Grupperum" in the Victor Albeck building.
- Form groups of size 4-5.
- Use the paper bin to indicate you need help.
- Check your solution with the teacher!
- Solutions to exercises day 1 is given in the Stata do-file "Exercise1.do" and possible supplementary pdf-files.
- Name tags are given here.
- Coffee will be served (approx. 9.30 am and 2 pm).

Practical things

An exam is given at the last Thursday, and a solution should be handed in 3 weeks after.

Statistical advice on you own data: http://bias.au.dk/

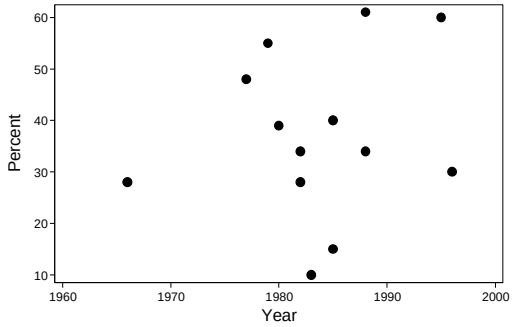
An example of possible time trend

A table from Altman (1998)

Year sampled		No obs.	No clustre	Percent
1966	Schor[4]	295	10	28
1977	Gore[5]	77	1	48
1979	White[6]	139	1	55
1980	Glantz[7]	79	2	39
1982	Felson[8]	74	1	34
1982	MacArthur[9]	114	1	28
1983	Tyson[10]	86	4	10
1985	Avram[11]	243	2	15
1985	Thorn[12]	120	4	< 40
1988	Murray[13]	28	1	61
1988	Morris[14]	103	1	34
1995	McGuigan[15]	164	1	60
1996	Welch[16]	145	1	30

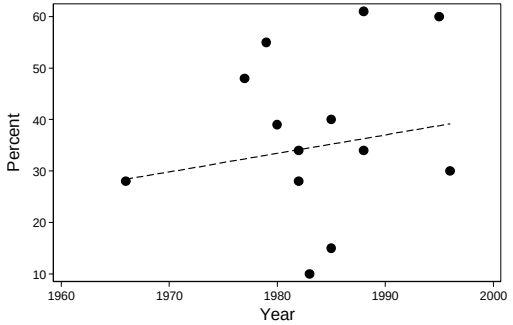
Any trend in the proportions?

Let's make a graph, ignoring for the moment the different sample size and clusters:



Any trend in the proportions?

A fitted line is added taking the the different sample size into account:



A slight increasing trend, however not statistical significant. So what is the statistical problem here?

More informations:

Table I. Summary of some reviews of the quality of statistics in medical journals, showing the percentage of 'acceptable' papers (of those using statistics)

Year published	First author	Number of papers	Number of Journals	% papers acceptable
1966	Schor[4]	295	10	28
1977	Gore[5]	77	1	48
1979	White[6]	139	1	55
1980	Glantz[7]	79	2	39
1982	Felson[8]	74	1	34
1982	MacArthur[9]	114	1	28
1983	Tyson[10]	86	4	10
1985	Avram[11]	243	2	15
1985	Thorn[12]	120	4	< 40
1988	Murray[13]	28	1	61
1988	Morris[14]	103	1	34
1995	McGuigan[15]	164	1	60
1996	Welch[16]	145	1	30

Altman (1998). Statistical reviewing in medical journals. Statist. Med. 17, 2661—2674.

The main problem in the non-acceptable papers is that the statistical model does not correspond to the type and sampling of the data.

One sample, continuous outcome

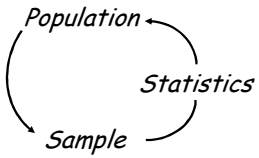
Suppose our scientific interest is

1. to **estimate** the expected (mean) BMI for 40-year old women.
2. to see if this mean could be 24.5 kg/m^2 .

We collect data on Body Mass Index for a random sample of eighty-five 40-year old women:

25.7	23.0	22.9	22.1	23.4	28.3	21.2	19.3	18.2
21.5	30.6	24.5	28.1	24.4	22.3	27.9	16.9	28.6
27.2	21.1	18.8	31.0	31.5	17.3	26.5	20.3	27.8
31.7	23.3	29.9	24.6	22.1	25.7	26.1	28.0	27.2
26.9	27.1	27.1	23.5	21.5	24.2	22.8	28.8	27.4
26.1	28.9	25.7	27.1	27.1	28.0	30.4	22.1	
25.7	25.5	31.0	21.7	27.2	18.6	23.4	23.7	
25.5	18.3	22.4	28.3	26.6	28.4	24.9	26.1	
23.6	21.0	30.8	27.7	26.0	19.5	27.5	21.8	
20.8	29.3	23.6	28.3	23.1	20.6	24.1	25.6	

What can we say about the population of 40- year old women?



Note that : "the mean BMI for 40-year old women" in the population is an **unknown quantity** - let us call it μ !

The purpose of the study is to **estimate** μ.

That is, to come up with a **relevant guess, based on the data** we have collected.

Under certain **assumptions** (we will **return to these later**) the best estimate is the sample average:

$$\hat{\mu} = \bar{x} = \frac{1}{85}(x_1 + x_2 + ... + x_{85}) = 24.99$$

^ : indicates that this is a (data-based) estimate

To avoid formulas that become very wide we will use the notation:

$$\bar{x} = \frac{1}{85}\sum_{i=1}^{85} x_i = \frac{1}{85}(x_1 + x_2 + ... + x_{85})$$

A **statistical model** is a set of assumptions about the sampling of the data and its variation.

Often the statistical analysis starts by a description of the data to find an appropriate statistical model.

Properties of the mean estimate

$$\hat{\mu} = \bar{x} = \frac{1}{85}\sum_{i=1}^{85} x_i = 24.99$$

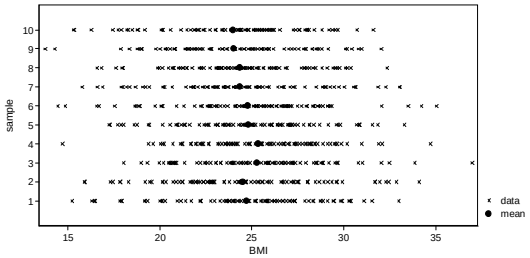
The statistical analysis depends on properties of the estimate.

Note, the estimate **depends on the sampled data**.

We would get another estimate, if we had sampled some other women!

Properties of the mean estimate

If we repeated sampling 10 studies of 85 observations we obtain:



Thus

- The estimate is a **random** quantity.
- The variation of the estimate is much smaller than the variation of the individual observations.

Properties of the mean estimate

$$\hat{\mu} = \bar{x} = \frac{1}{85}\sum_{i=1}^{85} x_i = 24.99$$

One usually quantifies the variation/uncertainty of the estimates by the **standard error**.

The **precise formula** for the standard error depends on the **assumptions/model**.

The estimate it is most often found by using a **computer program**.

In the special case of a random sample from the **normal distribution** we have:

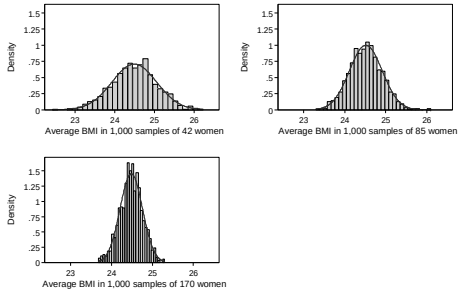
$$\text{sem} = \text{se}(\hat{\mu}) = \text{sd}/\sqrt{n}$$

$$= 3.58/\sqrt{85} = 0.388$$

sem is short for **s**tandard **e**rror of the **m**ean

Properties of the mean estimate

If we repeated sampling 1000 studies of 42, 85 and 170 observations we obtain:



Note

- The variation of the estimate becomes smaller as the sample size (42, 85 and 170) increase.
- The distribution of the estimates is symmetric.

The Confidence Interval

$$\text{sem} = \text{se}(\hat{\mu}) = \text{sd}/\sqrt{n}$$

From the formula we see, that the standard error decreases as the sample size increases – as $\sqrt{n}$.

Based on the estimate and the standard error we can construct a **95% Confidence Interval** – here the **"large sample version"**:

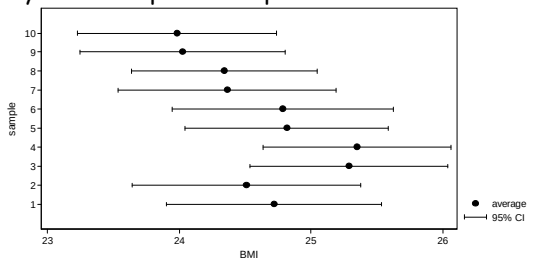
$$95\% - CI(\mu): \quad \hat{\mu} \pm 1.96 \cdot \text{se}(\hat{\mu})$$

$$95\% - CI(\mu): \quad 24.99 \pm 1.96 \cdot 0.388$$
$$= 24.99 \pm 0.760$$
$$= (24.23; 25.75)$$

The Confidence Interval

$$95\% - CI(\mu): \quad (24.23; 25.75)$$

Because the confidence interval depends on the data - it will vary from sample to sample:



The central property of a 95% confidence interval: 95% of the 95% confidence intervals will contain the true, but unknown, value of μ.

Purpose 1: the conclusion

Mean BMI: $24.99\,(24.23; 25.75)\,kg/m^2$

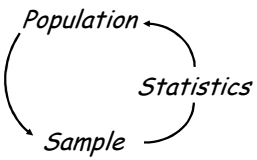
Based on the data our best guess is that the **mean BMI** for 40-year old women is: **24.99** kg/m^2.

But it might be as low as **24.23** kg/m^2 or as high as **25.75** kg/m^2.

Note, as **24.5** kg/m^2 is in the confidence interval, we cannot reject, that the mean BMI is **24.5** kg/m^2!

That is, the confidence interval also answers **purpose 2**!

The confidence interval quantifies information about the population:



Another way to answer purpose 2 is to **test** the hypothesis that the mean BMI is **24.5** kg/m^2, i.e. to test:

Hypothesis: μ = 24.5

This is a statistical test. The argument is somewhat more involved.

A statistical test – the basic version

We have observed: $\hat{\mu} = 24.99$ $\quad se(\hat{\mu}) = 0.388$

We want to test the hypothesis: $\mu = \mu_0 = 24.5$

The basic idea of a statistical test is to **compare** the **observed** with, what we **expect**, if the hypothesis is true – taking the **precision** of the observed into account:

$$z = \frac{\text{observed - expected}}{\text{standard error}} = \frac{\hat{\mu} - \mu_0}{se(\hat{\mu})}$$

Here we get: $z = \frac{24.99 - 24.50}{0.388} = 1.27$

Again this quantity, z, depends on the sample, i.e. **it is random**.

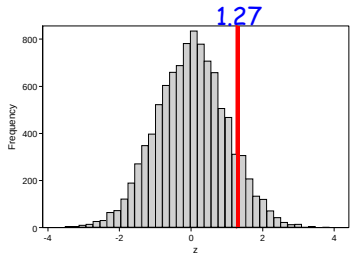
A statistical test – the basic version

$\hat{\mu} = 24.99$ $\quad se(\hat{\mu}) = 0.388$ $\quad$ Hypothesis: $\mu = \mu_0 = 24.5$

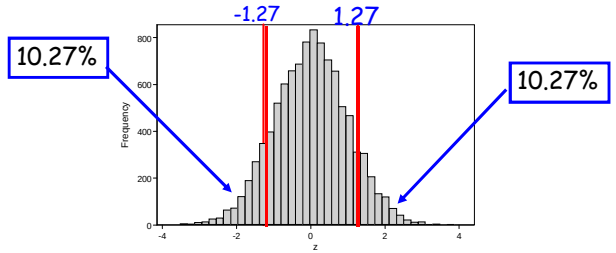
$$z_{obs} = (24.99 - 24.50)/0.388 = 1.27$$

If the **hypothesis is true,** then z follows a standard normal distribution (A result based on mathematical theory).

Numerically large values of z are critical for the hypothesis!



A statistical test – the basic version



The **p-value** = the probability of observing data at least as critical as what we have seen - given the hypothesis is true!

$$p \approx 2 \cdot 10.27\% = 20.54\%$$

Conclusion: Given the hypothesis is true, then there is a 21% chance of observing something as or more critical.

Interpretation of the p-value

The **p-value**= the probability of observing data at least as critical as what we have seen - given the hypothesis is true!

A **small p-value**, e.g. p=0.1% , implies that it is virtually impossible to observe the data we actually observed, if the hypothesis is true.

So we conclude that **the hypothesis cannot be true** – we **reject the hypothesis**.

If we do not reject the hypothesis, then we **might accept** it!

Note, a 'large' p-value is **not evidence,** that the hypothesis is true.

It only indicates that there is little evidence against it!

Maybe because we have little information
- **check the confidence interval**!

A statistical test – p-values and significance levels

The **p-value** = the probability of observing data at least as critical as what we have seen - given the hypothesis is true!

Usually, we will **reject the hypothesis** if the p-value is below a pre-specified value, the **significance level** (α).

Traditionally, one uses a **5% significance level**.

That is, we reject the hypothesis, if we – **given the hypothesis is true** – only had a 5% chance of observing data as extreme as the data we actually observed.

The significance level is equal to the risk of making a **Type 1 error** – reject a hypothesis given it is true.

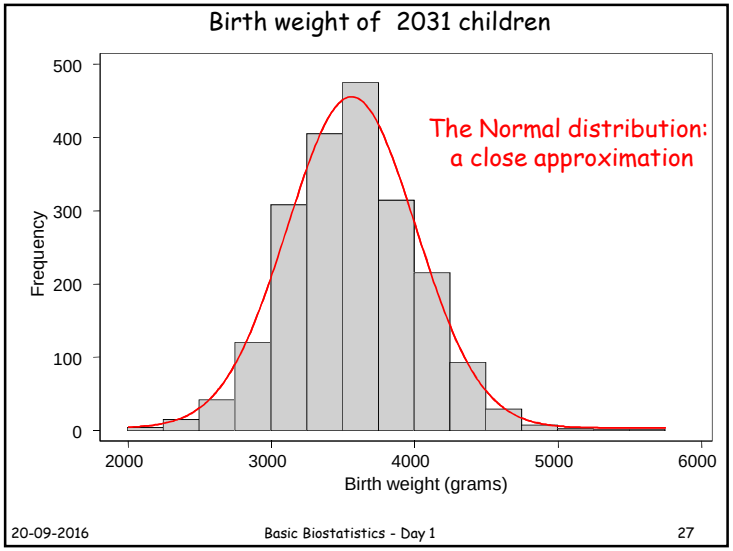
The Normal (Gaussian) distribution

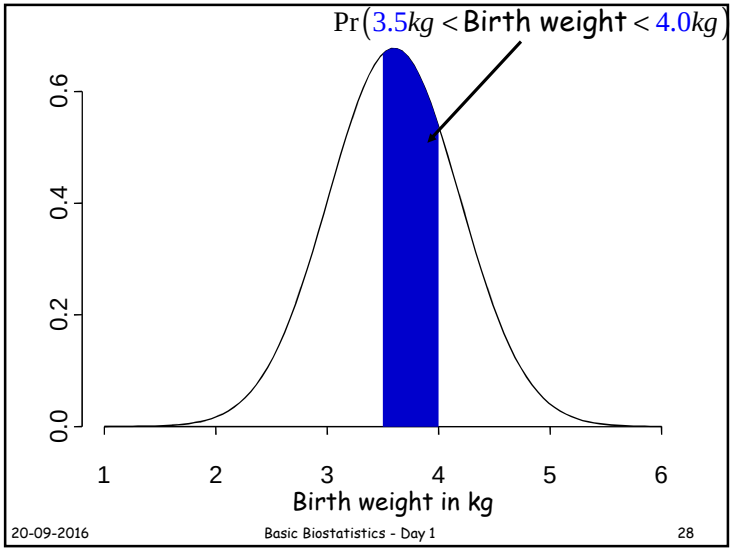
A central distribution due to two reasons:

- Many types of **data** are almost normally distributed (Maybe after a transformation).
- Many (more!) **estimates** are almost normally distributed, if they are based on a large number of observations (Maybe after a transformation).

Nothing is exactly normally distributed, but it is very often a real good approximation

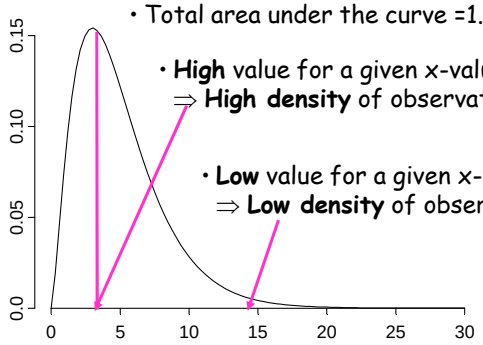
Relative quantities like **Odds Ratio**, **Relative Risk** or **Rate Ratio** should be analysed on **log-scale** (ln).





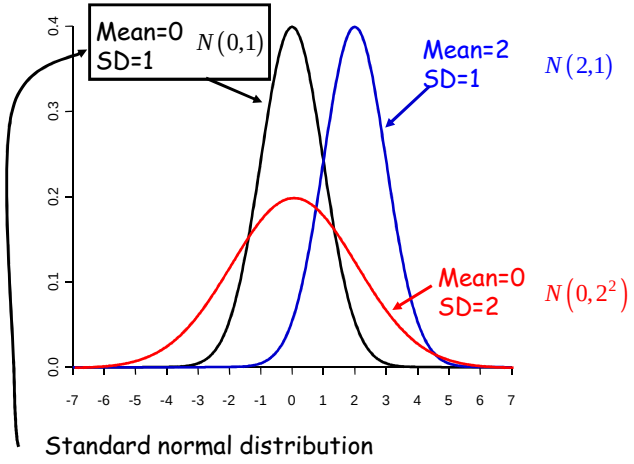
The density function:

- The probability of an observation in an interval = area under the curve.
- Total area under the curve =1.
- **High** value for a given x-value ⇒ **High density** of observations close to x.
- **Low** value for a given x-value ⇒ **Low density** of observations close to x.



20-09-2016 Basic Biostatistics - Day 1 29

Different normal distributions:



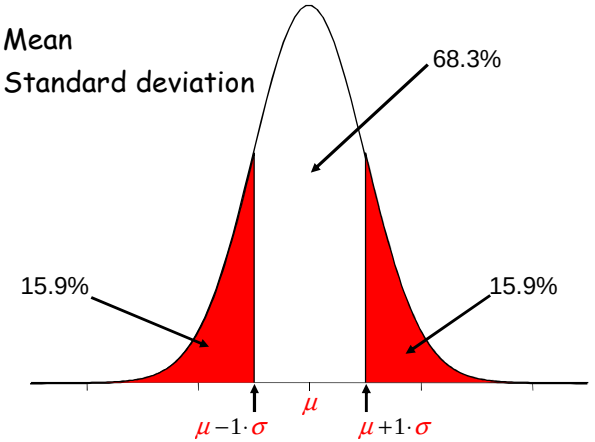
Standard normal distribution

20-09-2016 Basic Biostatistics - Day 1 30

How to understand the standard deviation (SD)?

μ = Mean

σ = Standard deviation

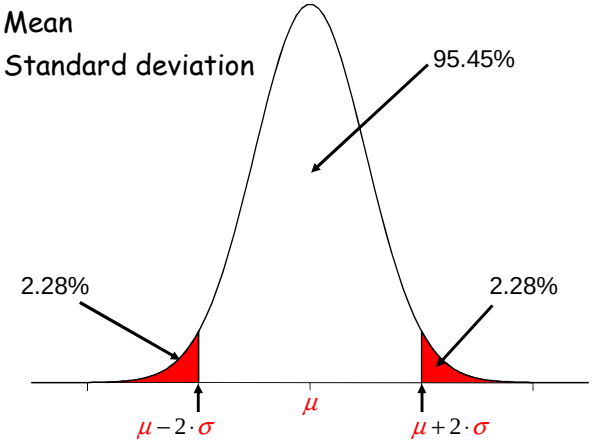


20-09-2016 Basic Biostatistics - Day 1 31

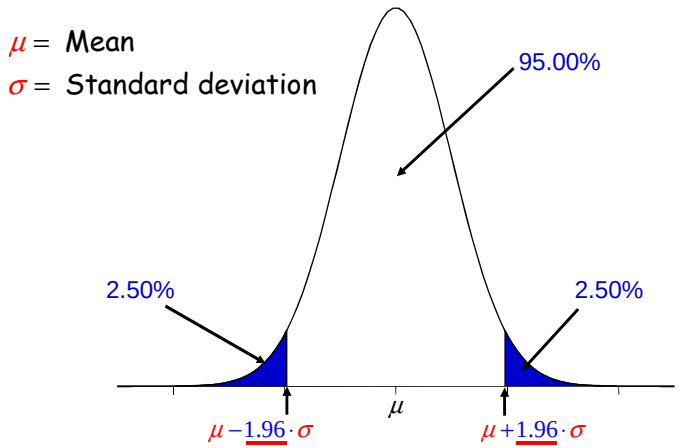
How to understand the standard deviation (SD)?

μ = Mean

σ = Standard deviation



20-09-2016 Basic Biostatistics - Day 1 32



Prediction interval (reference range)

A **95% Prediction Interval** (not a confidence interval!) is a interval that will contain 95% of the observations from a given population.

We saw that, **if** the data **follows a normal distribution**, then:

$$\mu \pm 1.96 \bullet \sigma$$

will contain 95% of the observations, i.e. it is a 95%-PI

If we want to base it on data, we have : $\bar{x} \pm 1.96 \bullet sd$

Here we get 95%PI: $24.99 kg/m^2 \pm 1.96 \bullet 3.58 kg/m^2$

$$= (17.97; 32.01)\, kg/m^2$$

That is, 95% of the women in a comparable population will have a BMI between 18 and 32 kg/m^2.
2.5% below 18 kg/m^2 and 2.5% above 32 kg/m^2 ,
if the data follow a normal distribution.

Statistical models

All results on the previous pages are based on a **statistical model**.

A statistical model is, like all other models, an **approximation** - that is a simplified description of the **real world**.

The **validity of the conclusions** depends on whether or not the model describes the important features of the process that generate the data.

As a consequence, **checking** the appropriateness of the underlying **assumptions** is an important part of the statistical analysis!

In the analysis of the BMI data, we have assumed that the data was a **random sample from a normal distribution**.

A random sample from a normal distribution. The three assumptions

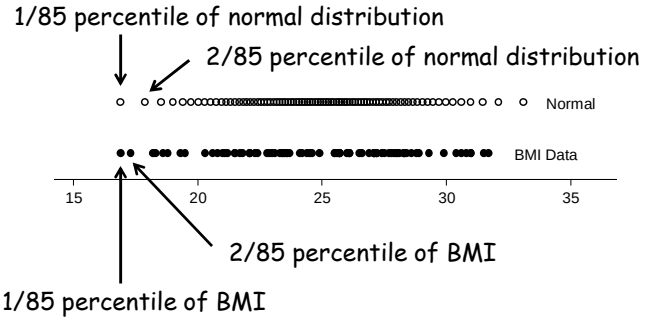
1. The observations are **independent** (knowing one observation will not alter the distribution of the others)
2. The observations come from the **same distribution**, e.g. they all have the same mean and variance.
3. This distribution is a **normal** distribution with unknown **mean**, μ, and **standard deviation**, σ. $N(\mu, \sigma^2)$

A random sample from a normal distribution
Checking the assumptions

1. Most often **independence** can only be checked by going through the **design - how was the data collected?**
   The assumption is violated if data/observational units
   cluster: some of the women are relatives, some are measured with one weight and some with another.
2. The **same distribution**: Do we have independent repetitions of the same "experiment"?
   If the data is collected over time, then one could plot the data as function of time.
3. Is the distribution **approximately normal**.
   Here one should make a **histogram** and a **Q-Q plot** of the data.

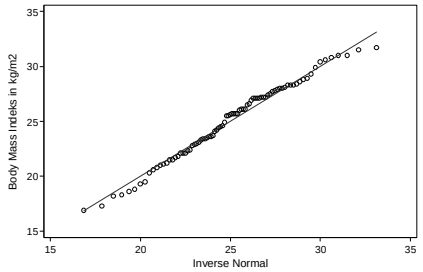
QQ plots

A QQ plots evaluate if the data follows approximately a normal distribution by comparing the percentiles of the data to the percentiles of a normal distribution with the same mean and standard deviation.



QQ plots

It is easier to evaluate if we plot the percentiles of the data to the percentiles of the normal distribution.



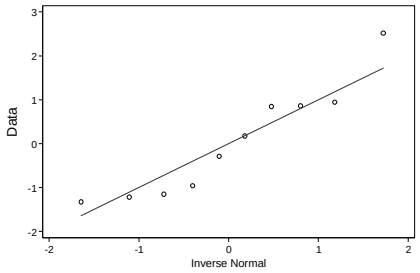
This is the QQ plot.

The points should be approximately on the identity line.

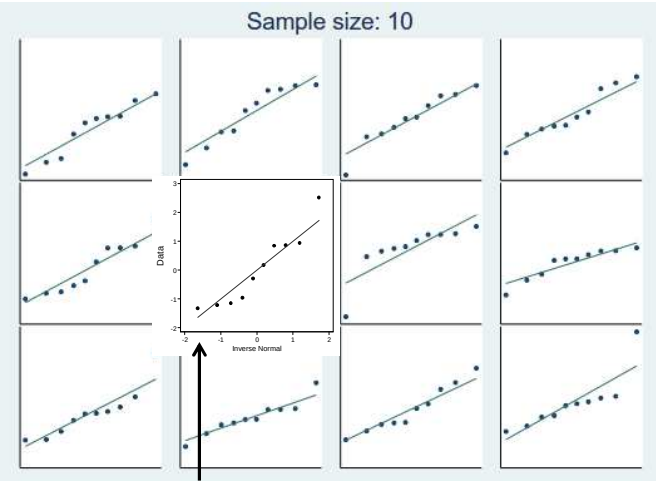
We conclude that BMI follows approximately a normal distribution.

QQ plot: The QQcatalog

For small sample sizes it may be less easy to evaluate. An example with 10 observations:



Experience may be obtained by comparing your QQ-plots to simulated QQ plots of similar sample size in the QQcatalog.pdf.



The QQ plot of the data similar to the simulated normals.

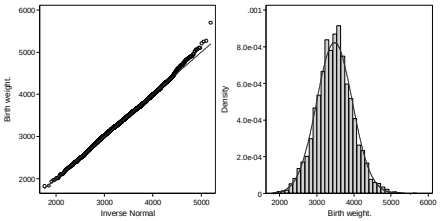
Test for normality

It is tempting to use a statistical test for normality.

However, these test for normality will

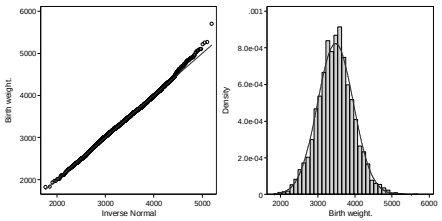
- usually reject for large samples
- often accept normality for small samples

Example: A random sample of birth weight born at term (37-40 weeks of gestation) in 2008.



The normal distributions seems a good approximation.

Test for normality



There is small excess number of babies with a large birth weight as compared to the normal distribution.

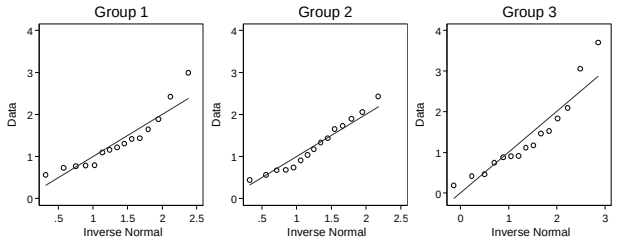
However, for most purposes the normal distributions is a very good approximation.

The Shapiro-Wilk test for normal data provides a p-values of 0.00001. For all 34,208 birth at that year the p-values is even 3e-16.

For large samples a test for normality will usually reject.

Test for normality

Example: 3 groups with concentration of triglyceride. Sample size in each group is 15.

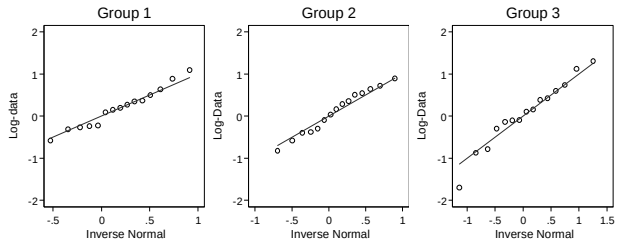


The Shapiro-Wilk test for normal data provides a p-values of 0.07, 0.55 and 0.06.

Test for normality

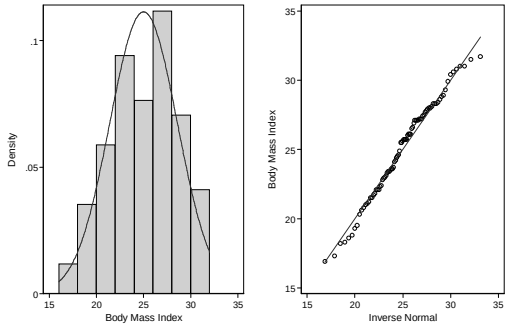
Concentration measurements are often analyzed on the log-scale.

The log-transformed QQ plots are a bit nicer.



For small samples a test for normality may have limited power.

A random sample from a normal distribution Histogram and Q-Q plot



If the data follows a normal distribution, then points in the Q-Q plot (right, which we most often use) would lie close to a straight line – **it looks okay here**!

Summary: A random sample from a normal distribution Estimation

The model contains **two unknown parameters**:

the **mean**, μ, and the **standard deviation**, σ!

Under the model they are estimated by:

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ and } \hat{\sigma} = \text{sd} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

That is the **average** (the observed mean) and the **observed standard deviation**.

An **approximate confidence interval** for μ is given by:

$$95\% - CI(\mu): \quad \hat{\mu} \pm 1.96 \cdot \text{se}(\hat{\mu})$$

$\text{se}(\hat{\mu}) = \text{sem} = \text{sd}/\sqrt{n}$ is the standard error of the mean

A random sample from a normal distribution Approximate inference

If one have more than **sixty** observations from a normal distribution one can rely on **approximate** methods

$$95\% - CI(\mu): \quad \hat{\mu} \pm 1.96 \cdot \text{se}(\hat{\mu})$$

$\text{se}(\hat{\mu}) = \text{sem} = \text{sd}/\sqrt{n}$ is the standard error of the mean

The test of the hypothesis: $\mu = \mu_0$ is based on

$$z_{obs} = \frac{\hat{\mu} - \mu_0}{\text{se}(\hat{\mu})}$$

and the p-value is given as $2 \cdot \Pr(\text{standard normal} \geq |z_{obs}|)$

This is, however, not the method that is implemented in standard statistical packages.

A random sample from a normal distribution
Exact inference

Under the normal model one can make **exact** inference using the t-distribution:

$$95\%-CI(\mu): \quad \hat{\mu} \pm t_{0.975} \cdot se(\hat{\mu})$$

Where $t_{0.975}$ is the upper 97.5-percentile in a t-distribution with **n-1 degrees** of freedom, d.f.=n-1.

The test of the hypothesis: $\mu = \mu_0$ is based on $z_{obs} = \frac{\hat{\mu} - \mu_0}{se(\hat{\mu})}$

and the p-value is given as

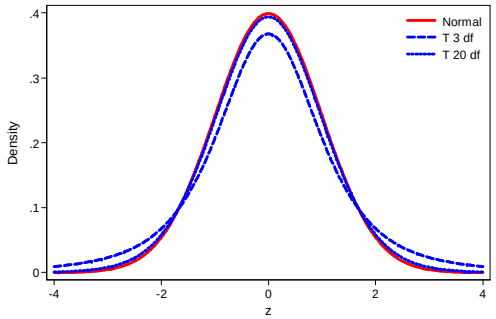
$$2 \cdot \Pr\left(\text{t-distribution with d.f.} = (n-1) \geq |z_{obs}|\right)$$

Nowadays this is done by computer!

The t distributions

The 97.5- percentiles in some t-distributions:

d.f.	1	2	5	10	20	60	84	∞
$t_{0.975}$	12.71	4.30	2.57	2.23	2.09	2.00	1.99	1.96



The BMI example
Exact inference

Using a computer we get: $95\%-CI(\mu): \ (24.22; 25.76)$

Hypothesis: $\mu = \mu_0 = 24.5$ p-value=**20.89%**

The approximate results we found before are very close to these exact – the sample is "large".

Summary: A random sample from a normal distribution
Confidence interval of σ

$$\hat{\sigma} = \text{sd} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The precision of the estimate of σ is genreally given by the degrees of freedom, *df*, which in general is the number of observations minus the number of unknown parameters describing the mean. Here *df=n-1*.

Finding the 95% CI for s is a bit complicated, as it involves the upper and lower 2.5 percentile in a chi-squared distribution with *df* degrees of freedom:

$$\hat{\sigma} \cdot \sqrt{\frac{df}{\chi^2_{df}(0.975)}} \leq \sigma \leq \hat{\sigma} \cdot \sqrt{\frac{df}{\chi^2_{df}(0.025)}}$$

Stata: QQ-plot, centile, CI and t-test

The main commands use today (more details can be found in the file day1.do).

```
. use bmiwomen.dta, clear
( data on body Mass Index for 40 year old women)
. qnorm bmi
. * Confidence interval for the mean
. ci bmi
 Variable |  Obs      Mean  Std. Err.   [95% Conf. Interval]
----------+--------------------------------------------------
      bmi |   85  24.99176  .3883103   24.21957    25.76396
```



```
. * Student's t-test for the hypothesis:  H_0: Mean BMI = 24.5
. ttest bmi=24.5
One-sample t test
------------------------------------------------------------------------------
Variable | Obs      Mean  Std.Err. Std.Dev. [95% Conf.Interval]
---------+--------------------------------------------------------------------
     bmi |  85  24.99176 .3883103  3.580044  24.21957  25.76396
------------------------------------------------------------------------------
    mean = mean(bmi)                                t =   1.2664
Ho: mean = 24.5                         degrees of freedom =   84
  Ha: mean < 24.5        mean != 24.5           Ha: mean > 24.5
Pr(T < t) = 0.8956   Pr(|T| > |t|) = 0.2089  Pr(T > t) = 0.1044

. * Prediction interval
. centile bmi, centile(2.5 97.5) meansd
                          -- Normal, based on mean and std. dev.--
Variable |  Obs  Percentile   Centile     [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   85        2.5   17.97501     16.66924    19.28077
         |            97.5   32.00852     30.70275    33.31429
```



```
. * Confidence interval of sigma.
. display 3.580044*sqrt(84/invchi2(84,0.975))
3.1109505
. display 3.580044*sqrt(84/invchi2(84,0.025))
4.2170376
* In Stata 14 the syntax is: ci variance bmi, sd
```


Stata: Creating a log-file

Log files are documentation of the statistical analyses.
Here is a simple, but typical example.

```
. ************************************************************
. * File: BMI analysis.do
. * Task: Compare the mean BMI to a hypothesis of 24.5.
. * Author: Erik Parner. 14-1-2015.
. ************************************************************
. * Change the directory.
. cd "D:\Teaching\BasicBiostat\Examples\BMI"
D:\Teaching\BasicBiostat\Examples\BMI
.
. * Close the log file, if there is one open.
. capture log close
.
. * Start a new log file.
. log using "BMI analysis.log" ,text replace
  *** output omitted ***
. * Read the data.
. use bmiwomen.dta,clear
( data on body Mass Index for 40 year old women)
```

```
. * Checking if the data follows a normal distribution using
. * a QQ plot.
. qnorm bmi
.
. * Student's t-test for the hypothesis: Mean BMI = 24.5
. ttest bmi=24.5

One-sample t test
------------------------------------------------------------------------------
Variable| Obs      Mean  Std.Err.  Std.Dev.  [95%Conf.Interval]
--------+---------------------------------------------------------------------
    bmi |  85  24.99176  .3883103  3.580044  24.21957  25.76396
------------------------------------------------------------------------------
  *** output omitted ***
. * Close the log file.
. log close
 *** output omitted ***
```

The video "Structure of the do file" gives more details.

The BMI example - formulation in Methods and Results

Methods:
Data was analyzed as one sample from a normal distribution based on the **Students t**. The assumption of normality was checked by a Q-Q plot.

Results:
The mean BMI was 25.0 kg/m^2 with a 95%-CI (24.2; 25.8) kg/m^2. The mean BMI was not statistically different from 24.5 kg/m^2 (p=21%).

Alternatively, with more focus on the expected 24.5 kg/m^2

The average BMI was 25.0 kg/m^2.
This was 0.5 (95%-CI: -0.3; 1.3) kg/m^2 more than the expected 24.5 kg/m^2. This difference was not statistically significant (p=21%).

The BMI example - conclusion

The conclusions will **always depend on setting**,

- What was **the purpose** of the study?
  Are we interested in **clinically relevant** effects/differences or on smaller differences that could have implications for the **etiology** ?
- How large effects/differences are relevant?
- What is known a priory?

Here you, the researcher, come in……