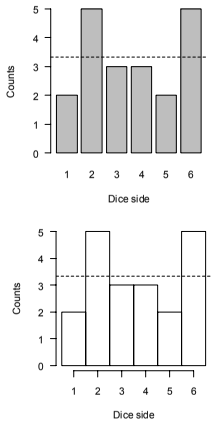
Empirical distributions

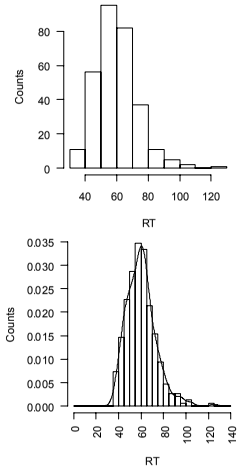


- Discrete probablity distribution

```
# simulated data set for 20 throws of a dice
set.seed(2)
x<-sample(1:6,size=20,replace=TRUE)

# counts outcomes
(counts<-table(x))
1 2 3 4 5 6
2 5 3 3 2 5

# barplot
barplot(counts,las=1,xlab="Dice
side",ylab="Counts")
abline(h=20/6,lty=2)

# histogram
hist(x,breaks=seq(.5,6.5,1),las=1,
        main="",xlab="Dice side",ylab="Counts")
abline(h=20/6,lty=2)
```

Note: for discrete distribution, the histogram represents the distribution of the observation

Empirical distributions

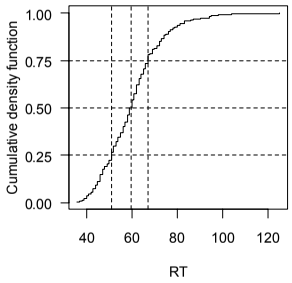


- Continuous probablity distribution

```
stern<-read.data("sternberg.dat")

# Frequency or ounts plot
hist(stern$rt,xlab="RT",ylab="Counts",
   main="",las=1)


# density plot
hist(stern$rt,breaks=seq(0,140,5),freq=FALSE,
        xlab="RT",ylab="",main="",las=2)
mtext("Counts",side=2,line=4)

lines(density(stern$rt))
```

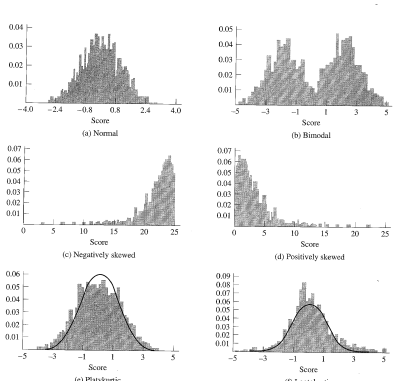
Note: for continuous distributions, the histogram is based on arbitrary bins.

[R] quantile and CDF plot



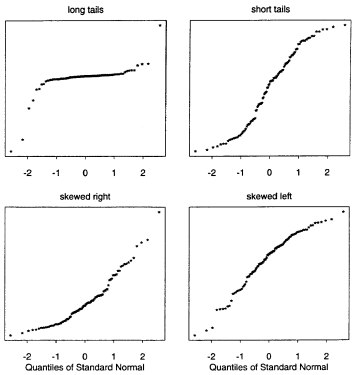

```
# quantile
p<-c(0.25,0.5,0.75)
(q<-quantile(stern$rt,c(0.25,0.5,0.75)))

# cumulative probability plot
plot(sort(stern$rt),

cumsum(rep(1/length(stern$rt),length(stern$rt))),
    type="s",xlab="RT",ylab="CDF",yaxt="n")
axis(2,at=seq(0,1,.25),las=1)
abline(h=p,v=q,lty=2)

# alternative method
plot(ecdf(stern$rt),xlab="RT",ylab="CDF",main="")
```


Skewness, kurtosis



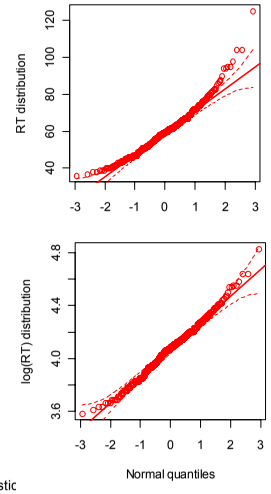
- Normal distribution (top left)
- Example of “deviations” from the assumption of normality:
  - Bimodal distribution (combination of two normal distributions)
  - Skewness is a measure of asymmetry (0=symmetric)
  - Kurtosis is a measure of tail length.

Probability plots



- Probability plot (also called rankit plot or theoretical QQ plot) represents the ordered data on the vertical axis against the corresponding normal scores on the horizontal axis.
- For normally distributed data, points should form a straight line. See graphs to see how to interpret deviations from the straight line.

[R] qqnorm, qq.plot



- Make a probability plot with the reaction time of the Sternberg dataset.
- The shape of the curve suggests data reaction time are skewed versus the left (this is indeed the case, see histogram in a previous slide)

```
qqnorm(stern$rt,main="",
    xlab="Normal quantiles",
    ylab="RT distribution")
```

or

```
library(car)
qq.plot(stern$rt,
    xlab="Normal quantiles",
    ylab="RT distribution")
```

- Make the probability plot of the log of the reaction time.

Hypothesis testing

Inferential statistics

- Inferential statistics: The branch of statistics concerned with methods that use a small set of data (sample) to make a decision (inference) about a larger set of data (population).
- Hypothesis testing:
  1. Formulation of the (null) hypothesis
  2. Realization of the experiment
  3. Computation of the statistics of interest from the sample
  4. Test of the hypothesis by comparing the statistic to the theoretical distribution assuming that the null hypothesis is true.
  5. Rejection of the null hypothesis if the test is statistically significant
- In hypothesis testing, the goal is to see if there is sufficient statistical evidence in the data to reject a presumed null hypothesis

The Null Hypothesis (H_0)

- The first step in hypothesis is to formulate an hypothesis about the variable of interest.
- In general, the hypothesis being tested – called the null hypothesis H_0- is ***counter*** to what we hope to demonstrate.

  For example, the null hypothesis might be that there are no difference between the means μ_1 and μ_2 of two groups, i.e. H_0: $\mu_1=\mu_2$, if you are interested to prove that there is a difference between the two groups.
- The philosophical argument introduced by Fisher for this approach, is that we can never prove something to be true but we can prove something to be false.

Hypothesis testing

- The second steps is
  - to formulate a statistic – a mathematical indicator - that will allow us to test the null hypothesis. For example, the t-statistic or the F-statistics allow us to realize test about means.
  - to compute the theoretical distribution associated with this statistic *under the assumption that the null hypothesis is true* (e.g., the Student or t-distribution for the t-statistic or the Fisher distribution for the F-statistic).
- This theoretical distribution will allow us to quantify the probability of observing the computed value for the statistic of interest under the assumption that the null hypothesis is true. The smaller this probability, the less likely the null hypothesis.
- The null hypothesis is said to be rejected if this probability is smaller that some arbitrary small threshold value (typically, 0.05 or 0.01). This value, referred to as the *level of significance* and noted by the Greek letter α , corresponds to the probability of rejecting erroneously the null hypothesis (Type 1 error).

Type I and II Errors

- An hypothesis test is a statistical decision; the conclusion will either be to reject the null hypothesis in favor of the alternative, or to fail to reject the null hypothesis. The ultimate decision may be correct or may be in error.

Hypothesis testing		Decision	
		H_0 true	H_0 false
True state	H_0 true	Correct	Type 1 error (α)
	H_0 false	Type 2 error (β)	Correct

- There are two types of errors, depending on which of the hypotheses is actually true (see Table):
  - A type 1 error is rejecting the null hypothesis when it is true. The probability of a Type I error is designated by the Greek letter alpha (α) and is called the Type I error rate.
  - A type 2 error is failing to reject the null hypothesis when it is false. A Type 2 error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. It is not an error in the sense that an incorrect conclusion was drawn since *no conclusion is drawn when the null hypothesis is not rejected*.

Rejecting the null hypothesis

- If the probability of observing the computed value of the statistic is smaller than α, we can reject the null hypothesis with less than α % of chances of making an error.
  - In this case, we say that the test is ***statistically significant*** at the α level.
  - In the context of a study where different treatments have been assigned to the various groups, we also say the ***effect*** of the factor of interest is statistically significant (the factor of interest is the independent variable that is manipulated by the experimenter, the effect is the difference between the mean values of the different groups). We will say more on this when we will study linear models. We can also say that the (mean value of dependent) variable ***depends*** on (the value of) the factor of interest.
- There is always a risk of making an error when rejecting an hypothesis. However, if the test is well done, we can guaranty that this risk is below some arbitrary level.

Effect size

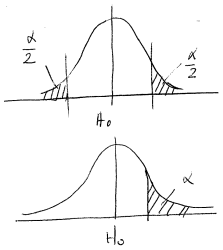
- A statistically significant test gives the indication that there is a good probability of observing this effect again if the study is repeated. In other words, statistical significance tests assess only the *reliability* of the observed effect.
- However, ***observing a statistically effect does not mean necessarily that this effect is relevant or meaningful***. Given large enough samples, any effect (e.g., a difference between the mean of two groups), even very small effects, can be made to be statistically significant.
- It is therefore often usefull to give an indication of the effect. In some cases, one might know from the context the meaningfulness of a significant effect. For example, let's assume that a diet followed by the mother has the effect of decreasing the weight of the babies by 500 g, this is certainly a meaningful difference but it would not be the case if the statistically significant difference was of 5 g.
- A typical way of measuring the size of an effect is to give the proportion of the total variance that is accounted by the effect. Various measures of effect sizes exist (e.g., R^2, η^2, ω^2).

Failing to reject the null hypothesis

- If the probability of observing the computed value of the statistic is larger than α, we cannot reject the null hypothesis.
- A null hypothesis is not proved because it is not rejected. The only valid interpretation in this case is that there is not enough evidence in the data to reject H_0 with less than α % of chances of making an error.
- For example, the observation that some data set is not sufficient to show convincingly that a difference between means is not zero does not prove that the difference is zero: No experiment can distinguish between the case of no difference between means and an extremely small difference between means.
- As stated by Fisher, we can never prove that an hypothesis is true: the fact that data are consistent with an hypothesis does not exclude that it might also be consistent with other other hypotheses making similar predictions. We can only prove that an hypothesis is false.
- The final conclusion once the test has been carried out is always given in terms of the null hypothesis. Even if we reject the null hypothesis (H_0), this does not mean that the alternative hypothesis (H_1) is true.

Alternative Hypothesis

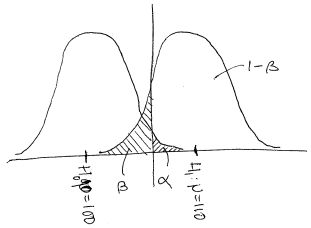
- The alternative hypothesis, H_1, is a statement of what a statistical hypothesis test is set up to establish
  - For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect compared to that of the current drug. In that case, the null hypothesis is H_0: $\mu_1 = \mu_2$ and the alternative hypothesis H_0:$\mu_1 \neq \mu_2$. To test this hypothesis, need to use a ***two-sided*** test.
  - The alternative hypothesis might also be that the new drug is better than the current drug. In this case, the null hypothesis is H_0:$\mu_1 - \mu_2 \leq 0$ and the alternative hypothesis H_0:$\mu_1 > \mu_2$. To test this hypothesis, we need to use a **one-sided** test.



- Note that these alternative hypothesis do not specify what is the difference between μ_1 and μ_2. Rejecting the null hypothesis rejected.

The minimal meaningful difference

- To compute the probability of accepting the null hypothesis when it is false (Type 2 error or β), we need to state a more specific alternative hypothesis. For example, the alternative hypothesis might be that the new drug bring a 15% improvement of the condition. In practice, however, we often lack a theory to predict the size of an effect, making it difficult to state specific alternative hypothesis.
- One possible way to determine an alternative hypothesis would be to decided beforehand what would be the *minimal meaningful difference* δ (or effect) that is relevant or interesting for the experimenter:
  - H_0: $\mu = \mu_0$
  - H_1: $\mu = \mu_0 + \delta$.

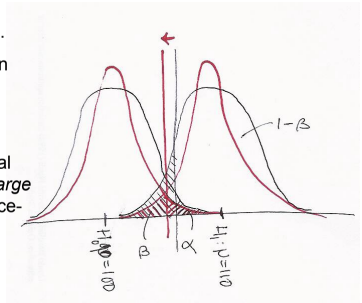


- For example, letîs assume that the null hypothesis is that the mean is 100 and the minimal meaning difference is 10, then it is possible to calculate the probabolity of a Type II error β and the power of the test 1- β.

Power and sample size

- Once a precise alternative hypothesis has been formulated, then it is also possible to compute:
  - The probability (β) of a type 2 error, i.e. the probability of *not* rejecting H_0 if H1 is true (when H_0 is false).
  - the power of the test (1-β), i.e. the probability of rejecting H_o if H_1 is true.

- The power of a test increases with sample size.
- Once a *minimal meaningful difference* has been fixed, it is possible to found out the minimum sample size that will allow a test to reach the desired levels of significance and power.
- The sample size must be *large enough* to reveal the minimal meaningful difference *but not too large* to yield a significant result with a trivial difference-



Power (normal distribution)

- **What is the probability of rejecting the hull hypothesis if the difference between the null and alternative hypothesis is δ?**

$$Power = 1 - \beta = \Pr\left(Z < \frac{\delta}{\sigma/\sqrt{N}} - z_{1-\alpha/2}\right) = \Psi\left(\frac{\delta}{\sigma/\sqrt{N}} - z_{1-\alpha/2}\right)$$

where Ψ is the cumulative normal distribution and $z_{1-\alpha/2} = \Psi^{-1}(1-\alpha/2)$. For unilateral tests, replace $z_{1-\alpha/2}$ with $z_{1-\alpha}$.

- In the previous example, the difference between the observed mean and the null hypothesis is 10.26 = 60.26 – 50. The probability of rejecting the null hypothesis assuming that this difference is true is almost 100%:

$$\Pr\left(Z < \frac{10.26}{13.01/\sqrt{300}} - 1.96\right) = \Pr(Z < 11.7) \cong 1$$

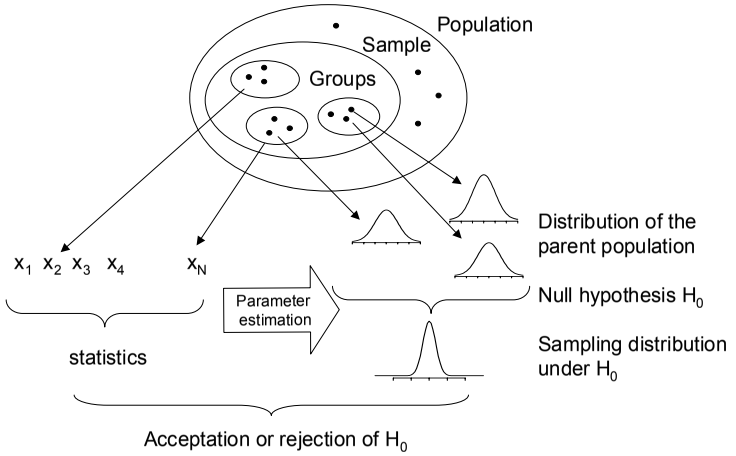
- Exercice. What is the power of the test if there were only 10 observations? Answer: 0.8 (80% of chance to reject H_0 is H_1 is true).

Power analysis in R

- Sample size planning for the power analytic approach is available for certain tests in R by default
  - `power.t.test` Power calculations for one and two sample t-tests
  - `power.prop.test` Power calculations two sample test for proportions
  - `power.anova.test` Power calculations for balanced one-way analysis of variance tests

- Other specialized packages
  - `pwr` package for some general linear model tests
  - `asypow` package for the asymptotic power of likelihood tests
  - `MBESS` package for behavioral and social sciences

Hypothesis testing summary



Hypothesis testing summary

Experimental (Descriptive statistics)	Theoretical (Inferential statistics)
Sample selection Treatment assignment Realization of the experiment (measures of the variable of interest)	Assumptions about the distribution of the variable of interest. Null hypothesis: e.g. no difference or no correlation between the groups.
Computation of the statistics of interest (sample statistic) from the measures (e.g., mean, difference between the two mean, or coefficient of correlation)	Computation the theoretical distribution (or sampling distribution) of the statistics of interest under the assumption that the null hypothesis (H0) is true.
Compare the sample statistics to the sampling distribution and reject H_0 if the probability of observing the sample statistics is lower that some value alpha (usually fixed at 0.05 or 0.01)	

Additonal issues: Effect size, power

Additional Reference

- Cohen (1994) The earth is round (p<.05). American Psychologist, 49(12):997-1003
- Nickerson (2000) Null Hypothesis Significance Testing: A Review of an Old and Continuing Controversy. Psychological Methods, 5(2):241-301.
- Bruce Thompson (1999) Why "Encouraging" Effect Size Reporting Is Not Working: the Etiology of Researcher Resistance to Changing Practices. Journal of Psychology, Vol. 133, 1999

Testing Means

- Parametric tests
  - Student t test,
  - one-way ANOVA
- Multiple comparisons
- Checking Assumptions
- Non-parametric test

Testing the Means

- Standard deviation of the population is known (=> normal distribution)
- Standard deviation of the population is unknown (=> Student or *t*-distribution)
  - One group
    - One sample T test
  - Two groups
    - Independent samples T test
    - Paired sample T test
  - Three or more groups
    - One-way ANOVA

One sample test

- **Objective:** Testing whether the mean of a sample $\{x_1,\ldots,x_N\}$ is different from some theoretical value. The variance of the parent population is unknown.
- **Assumptions:** The sample is normally distributed and measures are independent.
- **Null hypothesis:** H_0: $\mu=\mu_0$ => the theoretical mean μ is equal to the value μ_0.
- **Statistic:** The one sample t statistic is

$$t = \frac{m - \mu_0}{s/\sqrt{N}}$$

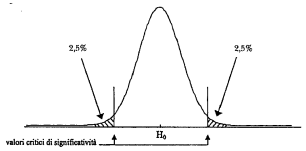
where *m* and s are the sample mean and sample standard deviation, and μ_0 is the theoretical value of the mean

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - m)^2}$$

- **Statistic theoretical distribution:** If the null hypothesis and our assumptions are true, then the *t*-statistic follow a *t*-distribution with N-1 degrees of freedom.

One versus two-tailed t-test

- **Two-tailed test:**
  - H_1: $\mu\neq\mu_0$ =>The alternative hypothesis is that the population mean μ is different from the value μ_0.
  - The lower and upper critical value are respectively $t_{\alpha/2}(N-1)$ and $t_{1-\alpha/2}(N-1)$
- **One-tailed test:**
  - H_1: $\mu>\mu_0$ => the alternative hypothesis is that the population mean μ is larger than the value μ_0 .
  - The critical value is $t_{1-\alpha}(N-1)$.
- **One-tailed test:**
  - H_1: $\mu<\mu_0$ => the alternative hypothesis is that the population mean μ is smaller than the value μ_0 .
  - The critical value is $t_{\alpha}(N-1)$.



t.test

```
t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95, ...)
```

Description

Performs one and two sample t-tests on vectors of data.

Main arguments

`x`	a (non-empty) numeric vector of data values..
`y`	an optional (non-empty) numeric vector of data values..
`alternative`	specify the alternative hypothesis: two-tailed = `"two.sided"` (default), one-tailed = `"greater"` or `"less"`.
`mu`	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
`paired`	a logical indicating whether you want a paired t-test
`var.equal`	a logical variable indicating whether to treat the two variances as being equal. If `TRUE` then the pooled variance is used to estimate the variance otherwise the Welch (or Satterthwaite) approximation to the degrees of freedom is used.
`conf.level`	confidence level of the interval.

One sample t-test

```
> stern<-read.table("sternberg.dat",header=TRUE)
> t.test(stern$rt,mu=50)

        One Sample t-test

data:  stern$rt
t = 13.6587, df = 299, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to
50
95 percent confidence interval:
 58.78175 61.73825
sample estimates:
mean of x
    60.26
```

Manual computations

```
> (m<-mean(stern$rt))
[1] 60.26
> (sm<-sd(stern$rt)/sqrt(length(stern$rt)))
[1] 0.7511694
> (mean(stern$rt)-50)/sm
[1] 13.65870
> m+c(-1,1)*qt(.975,299)*sm
58.78175 61.73825
```

- Test if the mean reaction time sid different from 50 (i.e., H_0: μ=50, two-tailed test).
- Set μ_0=50. By default, the t.test function does two-tailed test.
- value of the statistic t, number of degrees of freedom of the theoretical t-distribution, and p value
- The P value is extremely small
  p-value = Pr(|t| > 13.65) ≈0
  therefore the null hypothesis can be rejected with confidence.
- alternative hypothesis H_1 is μ≠50, this means two-tailed test.
- 95% confidence interval around the mean

$$t = \frac{m - \mu_0}{s/\sqrt{N}} = \frac{60.26 - 50}{0.7512} = 13.659$$

Independent-samples T Test

- **Usage.** The Independent-Samples T Test procedure compares means for two groups of cases. Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors. This is not the case if you compare average income for males and females. A person is not randomly assigned to be a male or female. In such situations, you should ensure that differences in other factors are not masking or enhancing a significant difference in means. Differences in average income may be influenced by factors such as education and not by sex alone.

- **Example.** Patients with high blood pressure are randomly assigned to a placebo group and a treatment group. The placebo subjects receive an inactive pill and the treatment subjects receive a new drug that is expected to lower blood pressure. After treating the subjects for two months, the two-sample t test is used to compare the average blood pressures for the placebo group and the treatment group. Each patient is measured once and belongs to one group.

Independent samples T test

- **Objective:** Testing the difference between the means of two *independent* samples $\{x_{11},\ldots,x_{1N1}\}$ and $\{x_{21},\ldots,x_{2N2}\}$ where N_1 and N_2 are the size of the first and second sample.
- **Assumptions:** Both samples are normally distributed and have approximately the same variance.
- **Null hypothesis:** H_0: $\mu_1 = \mu_2$, the two means are equal.
- **Statistic:** The two-sample t statistic is

$$t = \frac{m_1 - m_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$

where m_1 and m_2 are the means of the first and second ample, and s_p is the pooled estimate of the standard deviation.

$$s_p = \sqrt{\frac{\sum_{i=1}^{N_1}(x_{1i} - m_1)^2 + \sum_{i=1}^{N_2}(x_{2i} - m_2)^2}{N_1 + N_2 - 2}} = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}}$$

- **Theoretical statistic distribution:** If the null hypothesis and our assumptions are true, then the *t*-statistic follow a *t*-distribution with N_1+N_2-2 degrees of freedom.

Welch's test

Welch's test is a variant of Student's test for independent samples when the variances are not equal. In this case, the t statistic follows a student distribution with v degrees of freedom:

$$v = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)}{\left(\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}\right)}$$

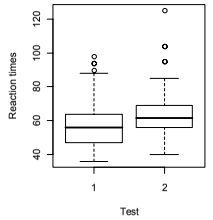
where v is an approximation to the effective degrees of freedom given by the Welch–Satterthwaite equation.

Welch, B L (1947) "The generalization of "Student's" problem when several different population variances are involved", Biometrika 34 (1–2): 28–35,

Independent sample T test (SPSS)

```
> boxplot(stern$rt[stern$test==1], stern$rt[stern$test==2],
+     ylab="Reaction times",xlab="Test")

> # same but using a formula
> boxplot(rt~test,stern,ylab="Reaction times",xlab="Test")
```

The boxplot represents the distribution of the reaction times when the test number was present (TEST=1) or missing (TEST=2) among the previously presented set of numbers. The question is whether the difference between the two means is statistically significant.




```
> t.test(stern$rt[stern$test==1],stern$rt[stern$test==2],
+ var.equal=TRUE)

  Two Sample t-test

data:  stern$rt[stern$test == 1] and stern$rt[stern$test == 2]
t = -3.6496, df = 298, p-value = 0.0003101
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -8.270783 -2.475884
sample estimates:
mean of x mean of y
 57.57333  62.94667
```

$$s_p = \sqrt{\frac{149\,(13.0792^2 + 12.4133^2)}{298}} = 12.7506$$

$$s_d = s_p\sqrt{1/150 + 1/150} = 12.7506 \times 0.1155 = 1.4727$$

$$t = \frac{m_1 - m_2}{s_d} = \frac{-5.373}{1.472} = -3.65$$

By default, t. test assumes that variances are not equal and use Welch (or Satterthwaite) approximation. Setting `var.equal=TRUE` will force N_1+N_2-1 degrees of freedom.

The null hypothesis can be rejected with confidence because the P value (Type 1 error) is extremely small (P≈0, two-tailed T test).

Paired T test

- **Usage.** The Paired-Samples T Test procedure compares the means of two variables for a single group. It computes the differences between values of the two variables for each case and tests whether the average differs from 0.
- **Example.** In a study on high blood pressure, all patients are measured at the beginning of the study, given a treatment, and measured again. Thus, each subject has two measures, often called before and after measures. An alternative design for which this test is used is a matched-pairs or case-control study. Here, each record in the data file contains the response for the patient and also for his or her matched control subject. In a blood pressure study, patients and controls might be matched by age (a 75-year-old patient with a 75-year-old control group member).

Paired T test

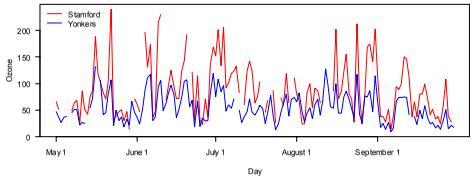
- **Objective:** Testing the difference between the means of two variables measured on the same experimental unit $\{(x_{11}, x_{21}),\ldots,(x_{1N}, x_{2N})\}$ where N is the number of experimental units.
- **Assumptions:** The differences $d_i = x_{1i} - x_{2i}$ are independent and normally distributed.
- **Null hypothesis:** H_0: $\mu_1 = \mu_2$, the two means are equal. Note that this is equivalent to say that the mean m_d of the differences d_i is zero(H_0: $m_d = 0$)
- **Statistic:** Let's compute the difference $d_i = x_{1i} - x_{2i}$ for each pair of sample (i =1,…,N). The statistic is

$$t = \frac{m_d}{s_d / \sqrt{N}}$$

  where m_d and s_d are the mean and standard deviation of d_i respectively.
- **Theoretical statistic distribution:** If the null hypothesis and our assumptions are true, then the *t*-statistic follow a *t*-distribution with *N*-1 degrees of freedom.
- Note that the procedure for paired t-test corresponds to a one-sample t test for the differences d_i.

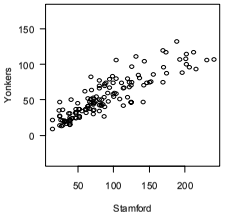
Ozone data set

The ozone data set gives the maximum ozone concentration in Stamford and Yonkers for each day of a five months period. Missing data are coded with the value –999.




```
ozone<-read.table("ozone.dat",header=TRUE)        # read data
ozone[ozone==-999]<-NA                            # replace missing data with NA
ozone$month<-ordered(ozone$month,                 # order factor month
          levels=c("May","June","July","August","September"))

> table(stmf=is.na(ozone$stmf),                   # count missing data
+     ykrs=is.na(ozone$ykrs))
       ykrs
stmf    FALSE TRUE
  FALSE   132    4
  TRUE     16    1

x<-ozone[,c("stmf","ykrs")]                       # plot time series
matplot(row(x),x,type="l",xaxt="n",xlab="Day",ylab="Ozone",
          las=1,lty=1,col=c("red","blue"))
axis(1,at=which(ozone$day==1),lab=paste(ozone$month,ozone$day)[ozone$day==1])
legend("topleft",c("Stamford","Yonkers"),lty=1,lwd=2,col=c("red","blue"),bty="n")
```


Paired t-test



- Plot Stamford again Yonkers concentrations

```
> plot(ozone$stmf,ozone$ykrs,las=1,asp=1
     xlab="Stamford",ylab="Yonkers")
```

- Observations in Stamford and Yonkers are correlated (which should be expected since they were taken on the same day).

```
> t.test(ozone$stmf,ozone$ykrs, paired=TRUE)

        Paired t-test

data:  ozone$stmf and ozone$ykrs
t = 13.0441, df = 131, p-value < 2.2e-16
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 29.66627 40.27312
sample estimates:
mean of the differences
                  34.9697
```

- Cases (days) where there is a missing value for one of the two cities are excluded from this analysis. The number of degrees of freedom (131) corresponds to number of days where the ozone concentration has been measured for both cities minus 1.
- The test is highly significant (P≈0, two-tailed T test), indicating that the mean m_d = 34.97 is not equal to zero and, therefore, that the two means are different.

Aanalysis of variance (ANOVA)

- One-way ANOVA
- Effect size
- Multiple Comparisons
- Checking Assumptions
- Transformations
- Non-parametric tests

Experimental designs and the ANOVA

- The Analysis of Variance (ANOVA) is one of the most common statistical techniques in psychology and medical sciences
- The ANOVA can be used to analyze the effect of manipulations (or experimental factors) on the mean value of the observations in a large variety of situations (or experimental design).
- Terminology:
  - *Completely randomized*, *block randomized* or *repeated-measure designs* refer to the way treatments have been assigned to the experimental units
  - *One-way*, *two-way, n-way* ANOVA refer to the number of experimental factors.
  - *Factorial* ANOVA refers to a particular way of analyzing experimental designs with two or more experiment factors which allow one to assess the possible interaction effect of interaction.

One-way ANOVA

- The aim of the one-way ANOVA is to test whether the means of two or more group are equal or not.
- In the *one-way* ANOVA, there is a *unique* independent variable or *experiment factor* (A) that is manipulated by the experimenter. The dependent variable (Y) is measured in different experimental conditions which correspond to different values or *level* of the experimental factor. In other words, the value of the dependent variable (or the level of the experimental factor) determines the group to which the observation belong.
- The experimental units must have been assigned randomly to each treatment (fully randomized design).

One-way ANOVA

- **Objective:** Compare the means of *g* samples $\{y_{11},\ldots,y_{1n1}\},\ldots,\ \{y_{g1},\ \ldots,y_{gng}\}$ where N_i is the sample size of the i^{th} sample ($N=n_1+\ldots+n_g$)
- **Assumptions:** Data are independent and normally distributed. The variance of each group is approximately equal.
- **Null hypothesis:** H_0: $\mu_1 = \ldots = \mu_g$, the means of all groups are equal.
- **Statistic:** The F-statistic is the ratio

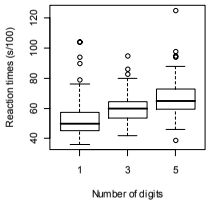
$$F = \frac{SS_B / g-1}{SS_E / N-g} = \frac{MS_B}{MS_E}$$

  where SS_B is the treatment (or between-group) sum of square, SS_E is the error (or within-group) sum of squares.
- **Theoretical statistic distribution:** If the null hypothesis and our assumptions are true, then the F-statistic follows a Fisher distribution with g-1 and N-g degrees of freedom.

Sums of squares

$\sum y_i^2$	Uncorrected sum of square
$SST = \sum (y_i - \bar{y})^2 = SSY$	Total sum of square (also called corrected sum of square for the y_i)
$SSR = \sum (\hat{y}_i - \bar{y})^2$	Explained sum of square
$RSS = \sum (y_i - \hat{y}_i)^2 = \sum \varepsilon_i^2 = SSE$	Residual sum of squares (or sum of squares of the errors, or unexplained sum of squares)
$SSX = \sum (x_i - \bar{x})^2$	Corrected sum of squares for the x_i
$SXY = \sum (x_i - \bar{x})(y_i - \bar{y})$	Corrected sum of squares of the product

Why the ANOVA works?



It can be proved that the total sum of square (SST) is equal to the treatment sum of square plus the error sum of square

$$SS_T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - m)^2 = SS_B + SS_E$$

- MS_B is an estimate of the variance of the means:

$$MS_B = \frac{SS_B}{g-1} = \frac{1}{g-1} \sum_{i=1}^{g} n_i (m_i - m)^2$$

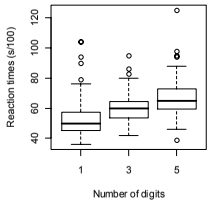
where m_i is the mean of the i^{th} sample and m is the general (or grand) mean. Note that MST=0 if H_0 is true.

- The mean square error MSE is an estimate within-group variance

$$MS_E = \frac{SS_E}{N-g} = \frac{1}{N-g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} n_i (y_{ij} - m_i)^2$$

- Note that we have assumed that the variance was the same for all groups.
- F becomes larger if the differences between the means are important and, therefore, the variance between groups is larger. F becomes smaller if there is much uncertainty about the means (that is, a larger within group variance)

Example



- Test if the mean reaction time (rt) depends on the number of digits memorized.

```
> boxplot(rt~ndigits,stern,outlier=FALSE,
+   ylab="Reaction times (s/100)",
+   xlab="Number of digits")
```

```
> fit<-aov(rt~ndigits,stern)
> anova(fit)
Analysis of Variance Table

Response: rt
           Df Sum Sq Mean Sq F value    Pr(>F)
ndigits     2   9257    4629   33.24 9.395e-14 ***
Residuals 297  41357     139
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The ANOVA table indicates that the null hypothesis can be rejected with high confidence (F(2,297)=33.240,P<0.001).

$$F = \frac{MS_B}{MS_E} = \frac{4628.61}{139.25} = 33.240$$

Multiple comparisons

- The F-test in the ANOVA tells us that there is at least one group that has a mean that is statistically different from the means of the other groups. Problem: How can we identify which groups have means that are statistically different one from another?

  One approach would be to use the t-test to make several two-by-two (or pairwise) comparisons. The problem is that the probability of making a false discovery (i.e., the risk of finding a statistically significant difference between two groups when the null hypothesis is true) increases with the number of tests.

- Note that the problem is extremely general and occurs not only when comparing means but when one makes multiple tests. When making multiple tests to find out some effect (data snooping), we always increase the risk of making a false discovery.

Bonferroni's inequality

Probability of making a false discovery in a series of n tests

n	Bonferroni limit	Independent tests
1	0.05	0.05
2	0.10	0.10
3	0.15	0.14
5	0.25	0.23
10	0.50	0.40

- **Bonferroni's inequality**: There is a theorem that says that the probability α of making a false discovery in a series of n test is smaller than the sum of the probabilities α_i of making a Type I error in the individual tests.

$$\alpha \leq \alpha_1 + \cdots + \alpha_n$$

- If all tests are independent, then we can compute the probability of making a false discovery in a series of n tests:

$$\alpha = 1 - (1 - \alpha')^n$$

where α' is the risk of a Type I error in the individual tests

Comparing means

- There are several techniques to adjust the level of confidence of the individual tests to control for the risk of a false discovery when comparing means between different groups. We need however to distinguish between the following situations:
  - **A priori or planned-comparisons:** One has identified in advances which pairs of means (or combinations of means) should be tested.
  - **Post-hoc comparisons:** In absence of specific hypothesis, one wants to make all pairwise comparison between the means of several treatment. Note that the number of comparisons (tests) in post-hoc comparisons is usually much higher than in planned-comparison.

A priori or planned comparisons

- In general, this situation involves a limited number of tests. In the case of comparison between means, the usual approach, based on Bonferroni inequality, is to use t-test with an alpha level α_i for the individual test equals to the desired alpha level α of protection divided by the number of test

$$\alpha_i = \frac{\alpha}{n}$$

where α_i is the adjusted alpha level for i^{th} t tests. Note this this method is extremely general and could be used to adjust the alpha level each time we make multiple tests.

- A particular type of hypothesis that one might want to test is, for example, whether the difference between the mean of a subset of groups is different from the mean from another subset of groups. In that case, one might use *contrasts* (we will not see contrasts in this class). Another possibility is to use a t test between the two subsets (remember to adjust the alpha according to the number of test performed).

Pairwise comparisons

Error rate	Method
Simultaneous confidence intervals	BSD or HSD
Strong familywise	REGWR
False discovery rate	SNK
Experimentwise	Protected LSD
Comparisonwise	LSD

Display 5.2: Summary of pairwise comparison methods

- A pairwise comparison test examines the differences of every pair of means. Note that there are g(g-1)/2 pairs of means for g groups.
- Pairwise comparisons test:
  - Sheffe's method
  - Bonferroni Significant Difference (BSD)
  - *Tukey Honest Significant Difference* (HSD)
  - Ryan-Eino-Gabriel-Welsh-Range (REGWR)
  - Student-Newman-Keuls (SNK)
  - Least Significant Difference (LSD)
- The most used test is the HSD test. It is more powerful than Sheffe's method and BSD when all pairs are considered. The other tests might be more powerful but provides less protection against Type I error (see definition of error rates in Statistical books).

Tukey HSD

- The function TukeyHSD can be used to compute *Tukey Honest Significant Difference*

```
> fit<-aov(rt~ndigits,stern)
> TukeyHSD(fit, which="ndigits")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = rt ~ ndigits, data = stern)

$ndigits
     diff      lwr      upr     p adj
3-1  7.38 3.449061 11.31094 0.0000407
5-1 13.59 9.659061 17.52094 0.0000000
5-3  6.21 2.279061 10.14094 0.0006921
```

- The **multicomp** package offers a convenient interface to perform multiple comparisons in a general context.

  Reference: Bretz F, Hothorn T, Westfall P (2010) Multiple Comparisons Using R. Routledge.

Effect size

- A statistically significant test does mean that the differences across goups are large, relevant or meaningful, in particular if the sample size is large (see Hypothesis testing).
- Several measures of the strenght of association of effect size exist:
  - Cohen's d and f
  - R^2, η^2, ω^2
- The computation of the effect size depend on the experimental design. We consider only the simplest cases.

Cohen's d and f

- One-sample t test

$$d = \frac{\mu - \mu_0}{\sigma} \approx \frac{m - \mu_0}{s}$$

- Two-sample t-test

$$d = \frac{\mu_2 - \mu_1}{\sigma} \approx \frac{m_2 - m_1}{s}$$

- one-way ANOVA

$$f = \frac{\sigma_B}{\sigma} = \sqrt{\frac{\frac{1}{p}\sum_{i=1}^{p}(\mu_i - \mu)^2}{\sigma}} \approx \left(\frac{p-1}{N-p}\right)\frac{MSB - MSE}{MSE}$$

values of f=.10 is a small effect size, f=.25 is a medium effect size, and f=.40 or larger is a large effect size.

Effect size (η^2)

- The most common measures of effect sizes express treatment magnitude as a proportion of the total variability that is associated with the effect
- The simplest measures of effect size is η^2 (also called R^2):

$$\eta^2 = R^2 = \frac{SS_B}{SS_T}$$

- For example, in the previous example, the stimulus length explains 18% of the variance of the observed reaction times (η^2 = 9257.2/50613.72 = 0.183).
- Note that it is possible to compute confidence intervals around effect size (Smithson, 2003) but we won't cover it here (see Tabachnick & Fiddell, 2007, for an introductory statistical textbood that covers it)

Effect size (ω^2)

- The simplest measures of effect η^2 is a descriptive statistics based on the proportion of variance accoundted in a particular sample and tends to overestimate the size of the effect in the population.

- An alternative measures, called ω^2, is an unbiased parameter estimate that generalizes to the population

$$\hat{\omega}^2 = \frac{SS_B - df_B MS_B}{SS_T + MS_E} = \frac{df_B(F-1)}{df_B(F-1)+N}$$

where N is the total number of subjects (both formulae are equivalent).

- Because it adjusts the overstimation of effect size by η^2, it is always smaller than η^2. In the previous example,

$$\hat{\omega}^2 = \frac{9257.22 - 2\times 139.247}{50613.72 + 139.247} = \frac{2\times(33.240-1)}{2\times(33.240-1)+300} = 0.177$$

Effect size

- The object returned by anova is a data.frame. The elements of this table can be selected and manipulated.

```
> x<-anova(aov(rt~ndigits,stern))
> is.data.frame(x)
[1] TRUE
```

- Compute Cohen's f

```
> dfB<-x["ndigits","Df"]
> dfE<-x["Residuals","Df"]
> MSB<-x["ndigits","Sum Sq"]
> MSE<-x["Residuals","Mean Sq"]
> sqrt(dfB/dfE*(MSB-MSE)/MSE)
[1] 0.468838
```

- Compute eta^2

```
> x["ndigits","Sum Sq"]/sum(x[,"Sum Sq"])
[1] 0.1828994
```

- Compute omega^2

```
> dfB<-x["ndigits","Df"]
> F<-x["ndigits","F value"]    # F ratio
> N<-sum(x[,"Df"])+1
> dfB*(F-1)/(dfB*(F-1)+N)
[1] 0.1769103
```


The assumptions of the ANOVA

- Independence of the observations.
  - Technically, this assumption states the the residuals ε_{ij} are independent: $\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{kl})=0$. For any two observations within an experimental treatment, we assume that knowing how one of these observations stands relative to the treatment mean tell us nothing about the other observation. This is one of the reasons why subjects are randomly assigned to groups. Violation of this assumption can have serious consequence for an analysis.
- Homogeneity of the variances.
  - The variances inside each group are equal ($\mathrm{Var}(\varepsilon_{ij}) = \sigma^2$).
- Normality.
  - Observations are distributed normally around their means: $y_{ij} \sim N(\mu_i, \sigma)$ or, equivalently, $\varepsilon_{ij} \sim N(0, \sigma)$.

- The number of treatments (groups) is fixed.
  - It is in principle possible to consider situations where the number of groups considered in the analysis constitutes only a subset of all possible groups/treatment. In the following, we shall however assume that all possible groups/treatment have been considered.

Checking the assumptions

- The quality of our inference depends on the validity of our assumptions. The main assumptions are:
  1. data/errors are independent
  2. data/errors are normally distributed
  3. data/errors have constant variance (across groups)

- Independence is the most important of these assumptions, and also the most difficult to accommodate when it fails.
- Normality is the least important assumption (for t-test or ANOVAs), particularly for large sample size.
- Balanced designs (same number of data in each group) are less susceptible to the effects of non-normality and non-constant variance.

- There are two main ways of dealing with violations of the assumptions:
  - Transforming the data
  - Using non parametric tests

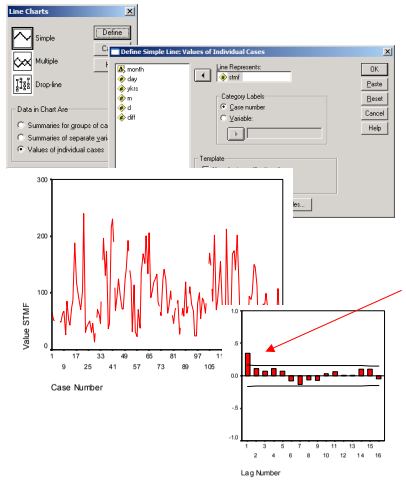
Independence

- Independence is the most important of these assumptions, and also the most difficult to accommodate when it fails. Example of dependences
  - serial dependence
    - time series,
    - learning or fatigue effect
    - drift in measuring instrument
  - spatial association

- A graphical method to identify serial dependence is to plot the data (or the residuals) versus their time or the order of acquisition

- To deal with possible order (learning or fatigue) effects, it is good practice to randomize the order of presentation of stimuli and vary this order across participants.

Example. Serial dependence



- Ozone data: plot maximum ozone concentration versus the day for Stamford. To plot all cases,
  - Select **Graphs/Line...**
  - Select **Values of individual cases** and **Simple**
  - Select Line Represents (stmf) and Case number in Category Labeks

- Auto correlation plot:
  - Select **Graphs/Times Series/ Autocorrelations**
  - Select **Variables** (stmf)

- There is some evidence that consecutive data are positively correlated at lag 1 (bar above the horizontal line).

Syntax

```
GRAPH
  /LINE(SIMPLE)=VALUE( stmf ) .
ACF  VARIABLES= stmf
  /NOLOG
  /MXAUTO 16
  /SERROR=IND .
```


Non-constant variance

- Several tests exist to test equality of variance across samples (homoscedasticity):
  - Bartlett's test
  - Levene's test
  - Brown-Forsythe's test

- Bartlett's test is sensitive to departures from normality. That is, if the samples come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. The Levene test and Brown–Forsythe test are alternatives to the Bartlett test that are less sensitive to departures from normality.

- The most common deviations from constant variance are those were the residual variation depends on the mean. You can plot the standard deviation against the mean for each group to check if there is a problem.

- To stabilize the variance when it increases with the mean, you can transform the variable with the square root or logarithm functions.

levene.test (library car)

```
levene.test(y, data, ...)
levene.test(y, group, ...)
```

Description

Test for homeogeneity of variance across groups.

Main arguments

`y`	response variable for the default method, or formula object. If `y` is a formula, the variables on the right-hand-side of the model must all be factors and must be completely crossed.
`group`	factor defining groups.
`data`	data frame containing dependent and factors referred to in the formula.

Notes

This function is part of the library car (J. Fox).

Assessing non-normality

- Normality is the least important assumption (for t-test or ANOVAs), particularly for large sample size.
- Graphical methods
  - Histograms
  - boxplot (useful for identifying quickly outliers).
  - probability blot (also named theoretical quantile-quantile plot or rankit plots)
- Formal methods
  - Chi square test
  - Kolmogorov-Smirnov test
- Outliers are an extreme form of non-normality and might have an important effect. An outlier is an observation "different" from the bulk of the data. Reviewing all the methods to deal with outliers is outside the scope of this course. If some data is unequivocally an outlier that needs to be deal with, it is possible
  - to replace the outlier with the mean value inside the group
  - to remove the outliers from the analysis

Kolmogorov-Smirnov test – ks.test

```
ks.test(x, y, ..., alternative = c("two.sided", "less", "greater"),
        exact = NULL)
```

Description

Performs one or two sample Kolmogorov-Smirnov tests.

Main arguments

`x` a (non-empty) numeric vector of data values..

`y` either a numeric vector of data values, or a character string naming a cumulative distribution function or an actual cumulative distribtution function such as `pnorm`

`alternative` specify the alternative hypothesis: two-tailed = `"two.sided"` (default), one-tailed = `"greater"` or `"less"`.

Details

If `y` is numeric, a two-sample test of the null hypothesis that `x` and `y` were drawn from the same *continuous* distribution is performed. Alternatively, `y` can be a character string naming a continuous (cumulative) distribution function, or such a function. In this case, a one-sample test is carried out of the null that the distribution function which generated `x` is distribution `y` with parameters specified by ...

[R] testing assumptions

- Test normality

```
> ks.test(stern$rt,y="pnorm",mean=mean(stern$rt),sd=sd(stern$rt))
        One-sample Kolmogorov-Smirnov test
data:  stern$rt
D = 0.0822, p-value = 0.03465
alternative hypothesis: two-sided
Warning message:
In ks.test(stern$rt,y="pnorm", mean=mean(stern$rt),sd(stern$rt)):
cannot compute correct p-values with ties
```

The presence of ties generates a warning, since continuous distributions do not generate them.

- Test equality of variance

```
> library(car)
> levene.test(stern$rt, as.factor(stern$ndigits))
Levene's Test for Homogeneity of Variance
      Df F value Pr(>F)
group    2  1.9511 0.1439
       297
```


Transformations

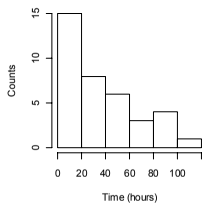
- Scores might be transformed to better for the following reaons:
  1. to achieve homogeneity of error variance
  2. to achieve normality of error effects
  3. to obtaine additivity effects
- Possible transformations include
  - The square root transformation: y'=sqrt(y)
  - the logarithmic transformation: y'=log(y) or y'=log(y+1)
  - the reciprocal transformation: y'=1/x or y'=1/(y+1)
  - etc.
- Given the relative robustness of the ANOVA to deviations from the normality or homoscadesticity assumptions, it is not clear whether transforming thendata is desirable. Still, it is conventional to transfrom some types of data.

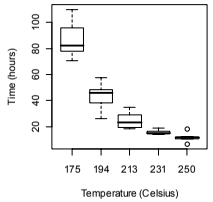
Data with non-normal distributions

Data	Transformation
Proportions (Binomial)	$\arcsin\left(\sqrt{p}\right)$
Coefficient of correlations	$\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$
Counts (Poisson)	$\sqrt{y}$

- Some data do not have a normal distributions and a variance that depends on the mean. For example,
  - Proportions
  - Coefficients of correlations
  - Poisson distributed data (e.g. counts, interspike intervals)
- To perform a t-test or an ANOVA on data of this type, it is necessary transform the variables with the functions indicated in the table.
- For proportions and counts, it is also possible to use generalized linear models (logistic regression, log-linear models, etc.).

Example - Resin Data Set





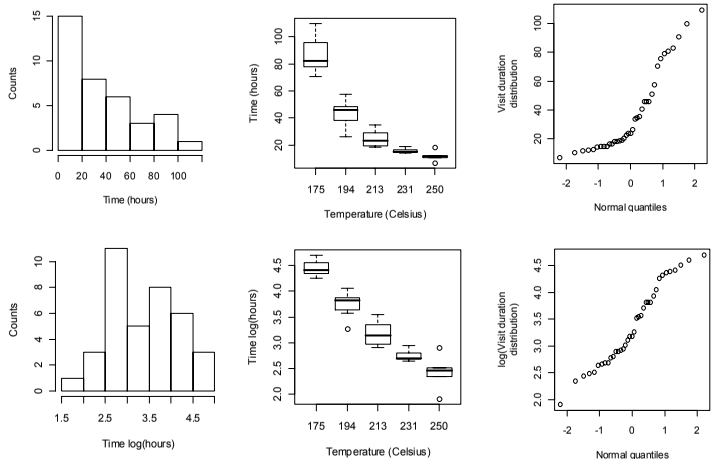
- The Resin Data Set: This data set represents the results of a test of the lifetime (in hours) of an encapsulating resin used in the construction of integrated chips. Thirty seven units were assign at random to one of five different temperature stress (temp variable, in Celsius). The time variable gives the time (in hours) and the time until the units failed.
- Histogram (and probability plot) show that data are not normally distributed and boxplot shows that variance vary with the mean. The same plots with the $\log_{10}$ of the lifetime shows that the data satisfy better the assumptions of the ANOVA (next slide).
- The ANOVA on $\log_{10}$ of the lifetime show that the null hypothesis can be rejected with confidence.

ANOVA

LOGTIME

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.538	4	.884	96.363	.000
Within Groups	.294	32	9.178E-03		
Total	3.831	36			

Resin Data Set



Non-parametric tests

- Non-parametric tests can be used when assumptions of parametric tests are not satisfied.

	Parametric	Non-parametric*
2 independent samples	T test	*Mann-Whitney U test*
2 dependent samples	Paired T test	*Wilconxon test* Sign test
N independent samples	One-way ANOVA	*Kruskal-Wallis test* Median test

- Non-parametric tests are less powerful than parametric tests.

wilcox.test (R)

```
wilcox.test(x, y = NULL, alternative = c("two.sided", "less", "greater"),
            mu = 0, paired = FALSE, exact = NULL, correct = TRUE,
            conf.int = FALSE, conf.level = 0.95, ...)
```

Description

Performs one and two sample Wilcoxon tests on vectors of data; the latter is also known as 'Mann-Whitney' test. .

Main arguments

`y`	an optional (non-empty) numeric vector of data values..
`alternative`	specify the alternative hypothesis: two-tailed = `"two.sided"` (default), one-tailed = `"greater"` or `"less"`.
`mu`	a number indicating the true value of the mean (or difference in means if you are performing a two sample test).
`paired`	a logical indicating whether you want a paired t-test data frame containing dependent and factors referred to in the formula.
`exact`	a logical indicating whether an exact p-value should be computed.
`correct`	a logical indicating whether to apply continuity correction in the normal approximation for the p-value.

kruskal.test (R)

```
kruskal.test(x, g, ...)
kruskal.test(formula, data, subset, na.action, ...)
```

Description

Performs a Kruskal-Wallis rank sum test. . .

Main arguments

`x`	a numeric vector of data values, or a list of numeric data vectors.
`g`	a vector or factor object giving the group for the corresponding elements of `x`. Ignored if `x` is a list.
`formula`	a formula of the form `lhs ~ rhs` where `lhs` gives the data values and `rhs` the corresponding groups.
`data`	an optional matrix or data frame containing the variables in the formula.

[R] Non parametric tests

```
> wilcox.test(stern$rt[stern$test==1],stern$rt[stern$test==2])

        Wilcoxon rank sum test with continuity correction

data:  stern$rt[stern$test == 1] and stern$rt[stern$test == 2]
W = 8198, p-value = 4.831e-05
alternative hypothesis: true location shift is not equal to 0

> kruskal.test(rt~ndigits,stern)

        Kruskal-Wallis rank sum test

data:  rt by ndigits
Kruskal-Wallis chi-squared = 73.7036, df = 2, p-value < 2.2e-16
```