One sample data analysis

Josep L. Carrasco
Bioestadística. Departament de Fonaments Clínics
Universitat de Barcelona

Setting

- One sample is collected in one population.
- The variable to analyze can be qualitative or quantitative.
- Here the parameters to evaluate will be:
    - **Proportion** if the variable is qualitative.
    - **Mean** if the variable is quantitative.
- The analysis will involve:
    - Confidence interval estimation.
    - Hypothesis testing.

Intervals

Proportion

Common approach

Example. What is the proportion of women in UB students? The variable to analyze is the gender (qualitative).

A sample of 500 students from the University of Barcelona is collected. A count of 288 women was observed.

- **Estimator**. The sample proportion.
- **Sample distribution**. Let X be the count of successes and n the total of subjects in the sample.

The sample proportion is $p = \frac{X}{n}$

Let π be the proportion in the population.

X follows a Binomial distribution with parameters n and π.

If the approximation Binomial to Normal is applied:

$$X \sim Bin\left(n, \pi\right) \longrightarrow N\left(n \cdot \pi, \sqrt{n \cdot \pi\left(1-\pi\right)}\right)$$

Remember that this approximation is acceptable if $n \cdot \pi > 5$ and $n \cdot (1-\pi) > 5$

With regards to relative frequency

$$p = \frac{X}{n} \sim N\left(\pi, \sqrt{\frac{\pi\left(1-\pi\right)}{n}}\right)$$

- The approach based on Normal approximation can be inaccurate if n is not large (continuity issue) or p is extreme.
- R uses **Wilson score interval** (WS) that improves the Normal approach.

We will use the *prop.test* function to obtain the confidence interval. The arguments are:

- x: counts of successes.
- n: counts of trials (sample size).
- conf.level: level of confidence, 95% by default.

```
prop.test(288,500)
```

```

    1-sample proportions test with continuity correction

data:  288 out of 500, null probability 0.5
X-squared = 11.25, df = 1, p-value = 0.0007962
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5312596 0.6195548
sample estimates:
    p
0.576
```

The sample estimate of the proportion of women is 0.576. The 95% confidence interval is $(0.531, 0.620)$

Alternative approach

Example. An assay aims to estimate the proportion of people having a genetic abnormality. A sample of 10 subjects was collected and the abnormality was found in 1 subject.

When $n \cdot \pi < 5$ or $n \cdot (1 - \pi) < 5$ the approximation to Normal distribution is too inaccurate.

In the example $n \cdot \pi = 1$ and $n \cdot (1 - \pi) = 9$.

- **Clopper-Pearson** (CP) approach is the exact alternative based on the Binomial distribution.

In R, CP method is implement in the *binom.test* function.

```r
binom.test(1,10)
```

```
    Exact binomial test

data:  1 and 10
number of successes = 1, number of trials = 10, p-value = 0.02148
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.002528579 0.445016117
sample estimates:
probability of success
                   0.1
```

Sample size

- Resources to carry out a survey/experiment are usually limited.
- The cost of the study is directly linked to the sample size.
- However, the success of the study is also linked to the sample size.
- When estimating parameters, "success" means the confidence interval would be informative → the interval must be narrow enough.
- Sample size computation: to optimize the sample size in such a way that the study aims are achieved.
- Confidence interval estimation: what is the minimum sample size needed to guarantee the confidence interval width will not exceed some concrete value?

$$n \geq \frac{Z^2 \cdot p \cdot q}{e^2}$$

- Z: quantile from the standard normal distribution. Its value depends on the confidence level. For example, $Z = 1.96$ for a confidence level of 95%.

- p: proportion/probability of success. Since its value is unknown (remember that we aimed to estimate it) we have to use a guess about it. Three options:

1) Use the estimate from a a pilot sample (with low sample size, say 20 o 30 subjects).
2) Use information about p from the literature to guess its value.
3) Use $p = 0.5$ which brings the largest value of n.

- q: the opposite of p, i.e. $(1 - p)$.

- e: half of the interval width.

Determining the sample size is a non trivial process. Let's use an example to show it.

Example. Following with the example on the genetic abnormality, let's suppose that the research aims to estimate the proportion of such abnormality. It is required that the length of the 95% confidence interval to be 0.1 as much. How many subjects must be sampled?

The level of confidence is 95%, so that $Z = 1.96$.

Concerning the value of p, researchers know that the probability of abnormality is low so it does not make any sense to set $p = 0.5$. Potential values should be lower than 0.2. To be more precise they decided to collect a small sample of 25 subjects and 4 of them showed the abnormality.

```r
binom.test(4,25)
```

```

	Exact binomial test

data:  4 and 25
number of successes = 4, number of trials = 25, p-value = 0.0009105
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.04537945 0.36082845
sample estimates:
probability of success 
                  0.16 
```

That gives an estimate of 0.16 with a 95% CI of 0.045 - 0.361.

If the value of 0.16 is used:

```
(qnorm(0.975)^2)*0.16*0.84/(0.05^2)
```

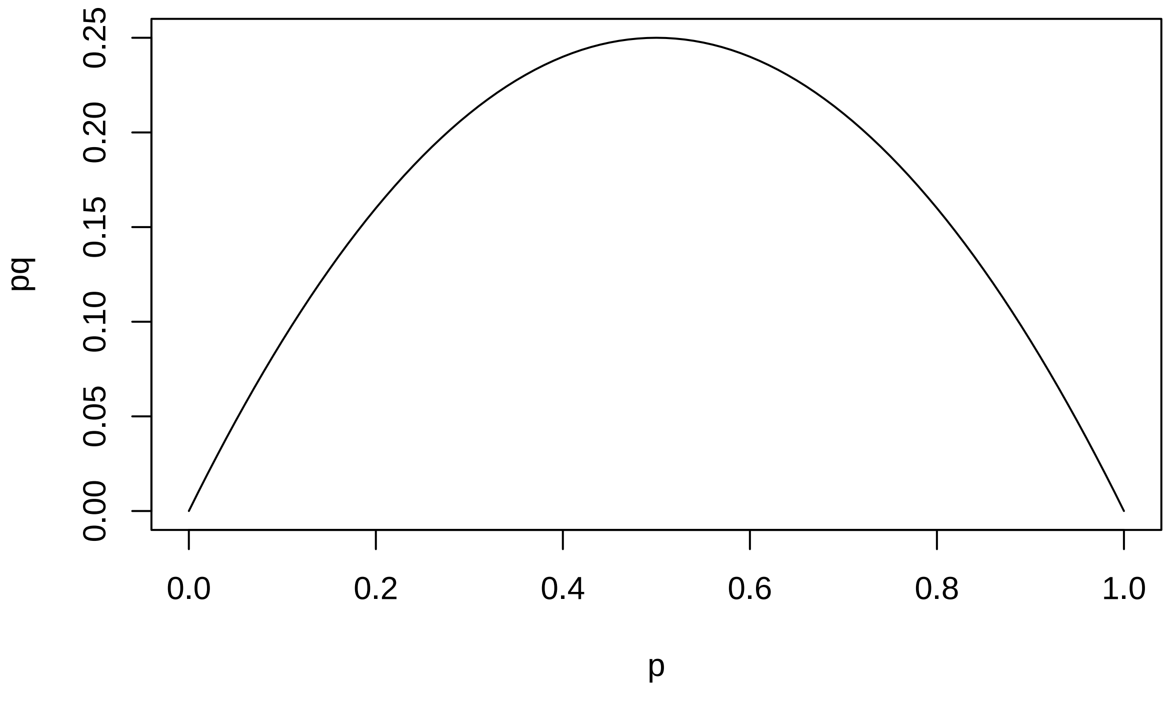
```
[1] 206.5168
```

So that 207 subjects are needed to guarantee that the confidence interval length will be 0.1 as much.

However this assertion will be true if the proportion in the sample with 207 subjects is lower than 0.16. Why? Let's draw the function $p \cdot q$ in relation to p.

```
p<-seq(0,1,0.01)
pq<-p*(1-p)

plot(p,pq,type="l")
```



The maximum of the function is reached at $p = 0.5$, that means that the maximum variability and eventually the largest sample size is given at $p = 0.5$. Conversely, lower values are found as p becomes more extreme.

Back to the example, if the sample proportion is higher than 0.16 the sample size will not be enough to guarantee an interval length of 0.1. Let's suppose that the sample of 207 subjects is collected and 42 of them have the abnormality.

```
x<-prop.test(42,207)
x$conf.int
```

```
[1] 0.1516244 0.2654695
attr(,"conf.level")
[1] 0.95
```

```
x$conf.int[2]-x$conf.int[1]
```

```
[1] 0.1138451
```

The length of the interval is higher than 0.1.

How to proceed? Two options:

- Use a safety margin. That means add a 10% or 20% more subjects than computed. In the case of the example, add between 20 and 40 subjects more.
- Use the worst situation (in terms of variability) that we could find. Remember the proportion's confidence interval from the pilot sample.

```
binom.test(4,25)
```

```
	Exact binomial test

data:  4 and 25
number of successes = 4, number of trials = 25, p-value = 0.0009105
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.04537945 0.36082845
sample estimates:
probability of success 
                  0.16 
```

The closest value to 0.5 in the confidence interval is 0.361. Use this value to compute the sample size.

```
(qnorm(0.975)^2)*0.361*0.639/(0.05^2)
```

```
[1] 354.4576
```

That gives a value of $n = 355$.

If the sample size is too large and unrealistic given that the resources to carry out the research, the restrictions about the length of the interval must be relaxed.

Mean

- **Estimator**. The sample mean.
- **Sample distribution**.

Let $x_1, \ldots, x_n$ the sample values. The expression of the sample mean is:

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

Give that x_i can be considered as independent and identically distributed, and following the Central Limit Theorem the sample distribution of the mean is a **Normal distribution**

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

- However this result is true if σ is known.
- How to proceed if σ is unknown?

Let s be the sample estimator of σ.

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

T follows a **t-Student** distribution with $n-1$ degrees of freedom.

t-Student distribution accounts for both mean and standard deviation sampling error.

Conditions:

- If the analyzed variable X follows a Normal distribution, the result is exact.
- If the analyzed variable X does not follow a Normal distribution, the result is asymptotic → large n needed (commonly $n \geq 30$).

Common approach

One of the two following conditions must be held:

- X follows a Normal distribution.
- n is large. In practice $n \geq 30$ is enough.

Example. What is the mean of BMI in a population?

A sample of 100 subjects was collected and their BMI was measured. You may download the data from the following link: bmi.txt

The sample mean and standard deviation were $m = 21.78$ and $s = 2.72$.

Since n is large enough we could apply the method based on t-student distribution. We have to use the *t.test(x,conf.level)* function.

- x: vector with data values
- *conf.level*: confidence level. By default 0.95.

```
t.test(bmi)
```

```

    One Sample t-test

data:  bmi
t = 80.178, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 21.24158 22.31962
sample estimates:
mean of x
  21.7806
```

The 95% confidence interval for the mean is 21.24 : 22.32

Alternative approach: bootstrap

- Bootstrap gets the sampling distribution by simulation.
- It is applicable to any estimator.
- It is a **free distribution approach**: no parametric model is assumed.
- Bootstrap is based on generating a huge number of samples (resamples) from the original sample using **sampling with replacement**.

Example. A sample of 15 values is obtained. We aim to estimate the mean. The distribution of the variable is unknown and unlikely to be a Normal distribution.

- Original sample:

```r
x=c(1,2,5:15,20,30)
x
```

```
 [1]  1  2  5  6  7  8  9 10 11 12 13 14 15 20 30
```

Let's generate three samples with replacement (resample) and compute the sample mean

```r
set.seed(2019) # Just to set the random seed and get the same results
m1=sample(x,replace=T)
m1
```

```
 [1] 11 15 12  7 15  1 10  9 15 14  5 12 11 15 12
```

```r
mean(m1)
```

```
[1] 10.93333
```

```r
m2=sample(x,replace=T)
m2
```

```
 [1] 20  1 10  5  9 20 20 30 10  5  6 30  7  5  6
```

```r
mean(m2)
```

```
[1] 12.26667
```

```r
m3=sample(x,replace=T)
m3
```

```
 [1]  7 30  9 11 14  1 20 30  9 14  8 14  8  1 30
```

```r
mean(m3)
```

```
[1] 13.73333
```

- Generate a large number of resamples and apply the estimator at each resample.

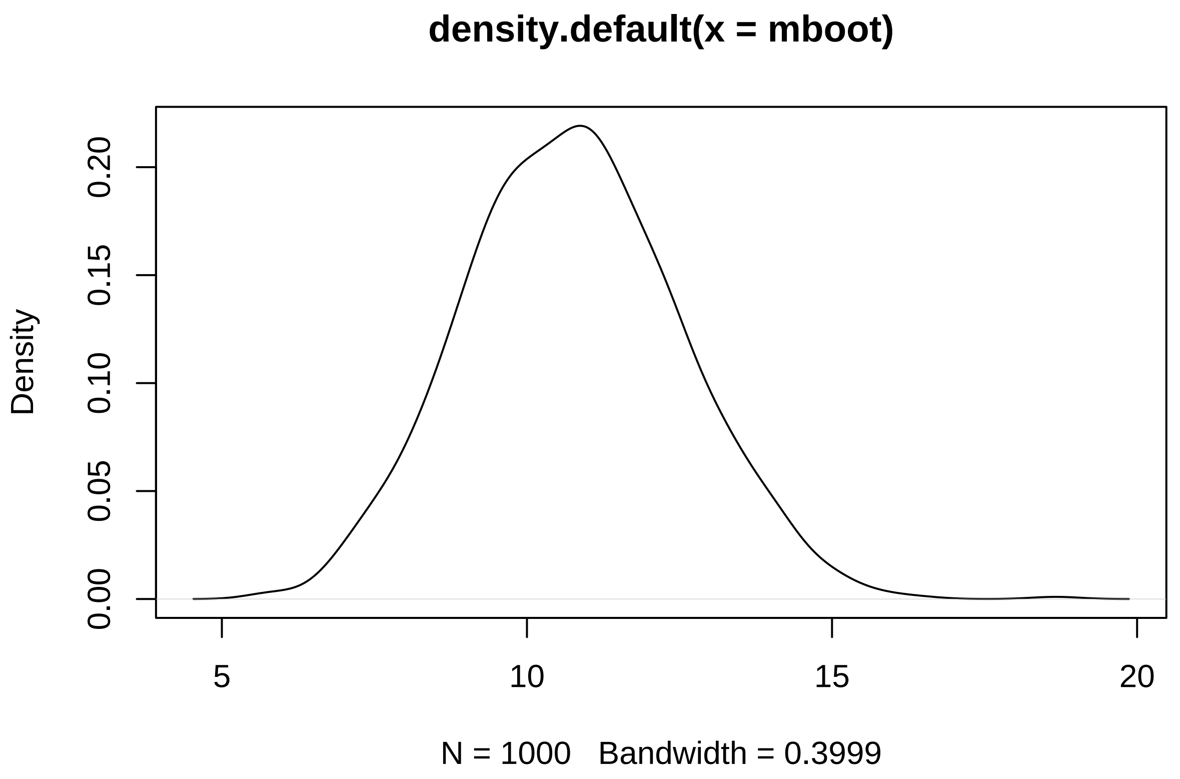
We can use the function *replicate(n,expr)* for repeated evaluation of an expression.

- **n**: number of replications.
- **expr**: expression to evaluate.

```
set.seed(2019) # Just to obtain the same results
mboot=replicate(1000,mean(sample(x,replace=T)))
```

We have generated the sample distribution of the mean.

```
plot(density(mboot))
```



To build a $(1-\alpha)$ % bootstrap confidence interval we'll use the *percentile* approach: compute the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap estimates.

```
quantile(mboot,probs=c(0.025,0.975))
```

```
    2.5%    97.5% 
 7.46500 14.26667 
```

Comments:

- Simulated sample distribution may be inaccurate with small samples.
- Percentile approach to estimate confidence intervals may give unrealistic intervals with biased estimators → there exists alternatives.
- Different executions lead to different intervals → set the random seed.

Sample size

Rule. Use the expression

$$n = \left(\frac{Z \cdot s}{e}\right)^2$$

Z: $1 - \alpha/2$ quantile of the standard Normal distribution.

s: sample standard deviation.

e: half of the interval length.

It is necessary to know the value of s.

1) Take a pilot sample.
2) Use previous knowledge on s.

Example. Following the case of BMI mean estimation, let's suppose that the sample of 100 subjects was a pilot study to obtain some knowledge about σ. Remember that the estimate was $s = 2.72$. The aim of the main study is to estimate the mean with a 95% confidence interval whose width must be lower than 0.5.

```r
Z<-qnorm(0.975)
e<-0.5/2

(Z*s/e)^2
```

```
[1] 453.572
```

The sample size must be 454 subjects at least.

Further Statistical Intervals

Prediction interval

- Aim: to predict a **future** observation of a variable using its probability model.
- The model parameters are unknown and must be estimated.
- Prediction needs to account for two kinds of variability:
  - Sampling error.
  - Randomness of the variable.

- Confidence intervals are not appropriate because they only account for sampling error.
- Confidence intervals are devoted to estimate parameters instead of make predictions of the variable.
- Only the Normal model case will be treated here.

Let $x_1, \dots, x_n$ be a sample from a Normal population. The mean and variance are estimated as $\bar{x}$, s^2.

The best predictor (less error) to estimate a new value x_{n+1} is the mean $\bar{x}$.

$$E\left(x_{n+1} - \bar{x}\right) = E\left(x_{n+1}\right) - E\left(\bar{x}\right) = \mu - \mu = 0$$

The variance of prediction error is:

$$V\left(x_{n+1} - \bar{x}\right) = V\left(x_{n+1}\right) + V\left(\bar{x}\right) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{1}{n}\right)$$

Note that $V\left(x_{n+1} - \bar{x}\right) \longrightarrow \sigma^2$ as $n \to \infty$

The $\left(1-\alpha\right)\%$ prediction interval is therefore defined as:

$$\bar{x} \pm t \cdot s\sqrt{\left(1 + \frac{1}{n}\right)}$$

where t stands for the $1-\alpha/2$ percentile of a t-Student distribution with $n-1$ degrees of freedom.

Notice that whether $n \to \infty$:

- Confidence intervals tends to 0.
- Prediction intervals tends to $\left(1-\alpha\right)\%$ probability interval.

Example. Let's follow with the BMI example. Remember that $n = 100$, $\bar{x} = 21.78$ and $s = 2.72$

The 95% prediction interval is:

```
n<-100
m<-mean(bmi)
s<-sd(bmi)

(l.pred<-m-qt(0.975,n-1)*s*sqrt((1+1/n)))
```

```
[1] 16.36352
```

```
(u.pred<-m+qt(0.975,n-1)*s*sqrt((1+1/n)))
```

```
[1] 27.19768
```

If a new subject is collected from the population, we have a confidence of 95% that its BMI will be within 16.36 and 27.2.

Tolerance interval

A tolerance interval is a probability interval estimated with some confidence.

$(1-p)\,\%$ probability interval: Interval that contains a probability of $(1-p)\,\%$.

In the Normal distribution case:

$$\mu \pm Z\sigma$$

where Z stands for the $1-p/2$ quantile of a standard Normal distribution.

For example, the interval $[-1.96, 1.96]$ in a standard Normal distribution (mean 0 and standard deviation 1) contains the 95% of probability, therefore it is a 95% probability interval.

The problem arises when μ and σ are unknown and need to be estimated → sampling error.

In the Normal distribution case the tolerance interval takes the following form:

$$\bar{x} \pm ks$$

where k is a value depending on the **tolerance level** $(1-p)$ and **confidence level** $(1-\alpha)$

Example Let's use again the BMI example. We wish to estimate the 80% tolerance interval with a 95% of confidence. That means the values of BMI that encloses the 80% of the population.

To estimate the tolerance interval we have to install and load the **tolerance** package and apply the **normtol.int(x,alpha,P,side)** with arguments:

- **x**. Vector data.
- **alpha**. The opposite of the confidence level. By default 0.05.
- **P**. Tolerance level. The proportion of the population to be covered by the interval.
- **side**. One-sided (1) or two-sided (2) interval. Default one sided.

```
library(tolerance)
tolint<-normtol.int(bmi,alpha=0.05,P=0.8,side=2)
l.tol<-tolint$`2-sided.lower`
u.tol<-tolint$`2-sided.upper`
c(l.tol,u.tol)
```

```
[1] 17.81265 25.74855
```

The interval 17.81 - 25.75 includes the 80% of BMIs of this population with a 95% of confidence.

Example. In an experiment it is aimed to determine the diameter in mm of a concrete part of a machinery. A sample of 25 parts are randomly collected from the factory and their diameter is accurately measured.

You may download the data from the following link: diameter.txt

The sample mean and standard deviation are

```
mean(x)
```

```
[1] 14.7236
```

```
sd(x)
```

```
[1] 1.619138
```

The 95% confidence interval of the mean

```
t.test(x)$conf.int
```

```
[1] 14.05525 15.39195
attr(,"conf.level")
[1] 0.95
```

We have a confidence of 95% that the *true* mean process is within 14.06mm and 15.39mm.

The 95% prediction interval

```
LL=mean(x)-qt(0.975,24)*sd(x)*sqrt(1+(1/25))
UL=mean(x)+qt(0.975,24)*sd(x)*sqrt(1+(1/25))
c(LL,UL)
```

```
[1] 11.31568 18.13152
```

If we take a new part from the factory, with a confidence of 95%, its diameter will be within 11.32mm and 18.13mm

The 90% tolerance interval with 95% confidence

```
library(tolerance)
normtol.int(x,alpha=0.05,P=0.9,side=2)
```

```
  alpha   P   x.bar 2-sided.lower 2-sided.upper
1  0.05 0.9 14.7236      11.13737      18.30983
```

We have a confidence of 95% that 90% of diameters of parts produced at the factory are within 11.14mm and 18.31mm.

Hypothesis testing

One proportion

Example. The efficacy (proportion of success) in a year of a new treatment is assessed. Researchers think that the efficacy must be greater than 25% in order to move the production of the treatment to the next stage.

The hypotheses are:

$$H_0 : \pi_0 = 0.25$$
$$H_1 : \pi_0 > 0.25$$

Common approach

- Based on the standard normal distribution
- Applicability conditions. All of them must be met:

1) $n \cdot \pi_0 > 5$
2) $n \cdot (1 - \pi_0) > 5$

Additionally it could be claimed the condition of $n \geq 30$.

Example. A sample of 40 subjects underwent the new treatment. After a year of follow up 12 of them experienced a successful result.

```
n<-40
p0<-0.25
n*p0
```

```
[1] 10
```

```
n*(1-p0)
```

```
[1] 30
```

The applicability conditions are met.

Let's obtain the p-value with the function **prop.test**. The syntax is similar to that of the confidence interval, but now we have to account for the one-sided alternative hypothesis (argument *alt* in the function) and null hypothesis π value (argument p in the function).

```
prop.test(12,40,p=0.25,alt="greater")
```

```

	1-sample proportions test with continuity correction

data:  12 out of 40, null probability 0.25
X-squared = 0.3, df = 1, p-value = 0.2919
alternative hypothesis: true p is greater than 0.25
95 percent confidence interval:
 0.1862508 1.0000000
sample estimates:
  p 
0.3 
```

The significance level is set to the common value of 5%.

Remember the rule:

- $P - value < \alpha \rightarrow$ Reject the null hypothesis
- $P - value \geq \alpha \rightarrow$ Do not reject the null hypothesis

The p-value is 0.29, greater than 0.05 so the null hypothesis is not rejected.

We could not demonstrate that the efficacy is greater than 25%.

Alternative approach

Example. Let's suppose the sample was of 15 instead of 40, and after a year of follow up 5 of them experienced a successful result.

Applicability conditions are not met:

```
n<-15
p0<-0.25
n*p0
```

```
[1] 3.75
```

```
n*(1-p0)
```

```
[1] 11.25
```

- Alternative: Binomial test.

```
binom.test(5,15,p=0.25,alt="greater")
```

```

	Exact binomial test

data:  5 and 15
number of successes = 5, number of trials = 15, p-value = 0.3135
alternative hypothesis: true probability of success is greater than 0.25
95 percent confidence interval:
 0.141664 1.000000
sample estimates:
probability of success
             0.3333333
```

The p-value is 0.31, greater than 0.05 so the null hypothesis is not rejected.

Sample size

In the hypothesis contrast setting the minimum sample size is computed to reach a lower bound of power level.

- What difference is aimed to detect? We name this difference as δ. It means what difference must be detected as significant.

- What is the aimed power? Power $= 1 - \beta$. Common values are above 70%.

Example. Following with the example of the new treatment's efficacy let's suppose that researchers wish to detect a difference of 10% with a power of 80%. Given that they wanted to demonstrate the proportion was greater than 25%, that ultimately means they want to reject the null hypothesis with a probability of 80% (power) if the true proportion is 35% or greater.

Let's use the function *pwr.p.test* from *pwr* package to compute the sample size needed.

The main arguments of this function are:

- *h*: effect size. A value based on the difference to detect. To compute it we will use the function *ES.h*.
- *power*: the aimed power of the test.
- *alternative*: the direction of the alternative hypothesis.

```
library(pwr)
h<-ES.h(0.35,0.25)
pwr.p.test(h=h,power=0.8,alternative="greater")
```

```
     proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.2189061
              n = 129.0186
      sig.level = 0.05
          power = 0.8
    alternative = greater
```

It gives a sample size of $n = 130$ subjects.

Furthermore, it is also possible to compute the power of a test. In the example we began with a sample of 40 subjects and the null hypothesis was not rejected. What was the power to detect a difference of 0.1 greater than 0.25? To answer that question just replace the argument *power* by *n*.

```r
h<-ES.h(0.35,0.25)
pwr.p.test(h=h,n=40,alternative="greater")
```

```
     proportion power calculation for binomial distribution (arcsine transformation)

              h = 0.2189061
              n = 40
      sig.level = 0.05
          power = 0.3972893
    alternative = greater
```

The power was about 40%, that means that was a probability of 40% of rejecting the null hypothesis if the true proportion would have been of 0.35.

Multinomial: more than one proportion

Example. Do the births in a hospital follow a uniform time distribution? A sample of 1062 births in a hospital were collected.

Time	9h-13h	13h-17h	17h-21h	21h-1h	1h-9h
Number of births	150	207	180	210	315
Observed proportions	0.1412	0.1949	0.1695	0.1977	0.2966

The hypotheses are:

$$\begin{aligned} H_0 : &\quad \boldsymbol{\pi} = \boldsymbol{\pi_0} \\ H_A : &\quad \boldsymbol{\pi} \neq \boldsymbol{\pi_0} \end{aligned}$$

Notice that $\boldsymbol{\pi}$ is a vector of proportions, and $\boldsymbol{\pi_0}$ is the vector of proportions proposed in the null hypothesis.

In the example there are six time sections, five of them are 4 hours length that implies a $\frac{1}{6}$ of a day. The remaining time section is 8 hours length that suppose a $\frac{1}{3}$ of a day. Thus, the hypotheses according to test whether the births are uniformly distributed along a day are:

$$\begin{aligned} H_0 : &\quad \pi_1 = \pi_2 = \pi_3 = \pi_4 = \tfrac{1}{6} \cap \pi_5 = \tfrac{1}{3} \\ H_A : &\quad \pi_1 \neq \tfrac{1}{6} \cup \pi_2 \neq \tfrac{1}{6} \cup \pi_3 \neq \tfrac{1}{6} \cup \pi_4 \neq \tfrac{1}{6} \cup \pi_5 \neq \tfrac{1}{3} \end{aligned}$$

- It is always **two-sided**.

Common approach

- Statistical test (chi-square test):

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

- O_i: observed frequencies in the sample.
- E_i expected frequencies if H_0 is true.
- Distribution under H_0: χ^2 distribution with $K-1$ degrees of freedom where K is the number of categories.
- Conditions. All $E_i \geq 5$
- Pearson's residuals. Useful when the null hypothesis is rejected.

$$r_i = \frac{O_i - E_i}{\sqrt{E_i}}$$

To solve the test we use the function **chisq.test(x,p)**:

- x: vector with observed frequencies
- p: proportions under the null hypothesis

```
# Observed frequencies
fab=c(150,207,180,210,315)
# Create test object
test=chisq.test(fab,p=c(rep(1/6,4),1/3))
# Check the expected frequencies. All >5 --> OK
test$expected
```

```
[1] 177 177 177 177 354
```

```
# Test results
test
```

```

    Chi-squared test for given probabilities

data:  fab
X-squared = 19.703, df = 4, p-value = 0.0005714
```

$P-value < 0.05 \longrightarrow H_0$ is rejected.

Conclusion. The births are not distributed uniformly across the daytime. Question: where are the differences?

```
# Pearson residuals
test$residuals
```

```
[1] -2.0294443  2.2549381  0.2254938  2.4804319 -2.0728266
```

- A residual greater than 2 is indicative of a significant distance to H_0.
- Negative residual $\longrightarrow$ Observed $<$ Expected
- Positive residual $\longrightarrow$ Observed $>$ Expected
- 1st and 5th time intervals: there are less births than expected.
- 2nd and 4th time intervals: there are more births than expected.

Alternative approach

- **Monte carlo** simulation significance test procedure consist of the comparison of the observed data with random samples generated in accordance with the hypothesis being tested.
- The outcome of the test is determined by the rank of the test result with observed data relative to the test results of the random samples.

Example. Is the genotype of a concrete gene in a population according to the following proportions?

$$\pi_{AA} = 0.25 \; ; \; \pi_{Aa} = 0.5 \; ; \; \pi_{aa} = 0.25$$

A sample of 15 subjects is collected. The following frequencies are observed:

Genotype	AA	Aa	aa
Observed	4	11	0

```
#Observed frequencies
fab=c(4,11,0)
#test
test=chisq.test(fab,p=c(0.25,0.5,0.25))
# Expected frequencies < 5
test$expected
```

```
[1] 3.75 7.50 3.75
```

Expected counts are lower than 5. Apply the alternative approach.

```
test<-chisq.test(fab,p=c(0.25,0.5,0.25),simulate.p.value=T)
# Test results
test
```

```

    Chi-squared test for given probabilities with simulated p-value (based
    on 2000 replicates)

data:  fab
X-squared = 5.4, df = NA, p-value = 0.07746
```

Using a significance level of 5% the null hypothesis is not rejected.

Handycap: every execution gives a different p-value.

Solutions:

- Set the random seed.

```
set.seed(2022)
test<-chisq.test(fab,p=c(0.25,0.5,0.25),simulate.p.value=T)
# Test results
test
```

```

	Chi-squared test for given probabilities with simulated p-value (based
	on 2000 replicates)

data:  fab
X-squared = 5.4, df = NA, p-value = 0.08446
```

- Increase the number of simulations to make the result more steady. Default value is 2000 simulations. Let's use 10000.

Two executions that give similar results.

```
test<-chisq.test(fab,p=c(0.25,0.5,0.25),simulate.p.value=T, B=10000)
# Test results
test
```

```

	Chi-squared test for given probabilities with simulated p-value (based
	on 10000 replicates)

data:  fab
X-squared = 5.4, df = NA, p-value = 0.07509
```

```
test<-chisq.test(fab,p=c(0.25,0.5,0.25),simulate.p.value=T, B=10000)
# Test results
test
```

```

	Chi-squared test for given probabilities with simulated p-value (based
	on 10000 replicates)

data:  fab
X-squared = 5.4, df = NA, p-value = 0.07639
```

- Combine the two solutions.

```
set.seed(2022)
test<-chisq.test(fab,p=c(0.25,0.5,0.25),simulate.p.value=T, B=10000)
# Test results
test
```

```
	Chi-squared test for given probabilities with simulated p-value (based
	on 10000 replicates)

data:  fab
X-squared = 5.4, df = NA, p-value = 0.07729
```

Sample size

To compute the sample size we will use the function *pwr.chisq.test* from the *pwr* package.

The main arguments are:

- *w*. Effect size. Related to the difference to detect. Consider a value of 0.1 as **small**, 0.3 as **medium**, and 0.5 as **large**.
- *df*: degrees of freedom, number of categories minus 1.

In the example, what is the sample size necessary to detect a small effect with a power of 80%? Remember that were five categories so four degrees of freedom.

```
library(pwr)
pwr.chisq.test(w = 0.1, df = 4, power = 0.8)
```

```
     Chi squared power calculation 

              w = 0.1
              N = 1193.529
             df = 4
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations
```

What was the power in the example to detect a small effect? Remember the sample size was 1062.

```r
pwr.chisq.test(w = 0.1, df = 4, N = 1062)
```

```

     Chi squared power calculation 

              w = 0.1
              N = 1062
             df = 4
      sig.level = 0.05
          power = 0.7453805

NOTE: N is the number of observations
```

Mean

Example. The gum blood flow in 30 health patients has a velocity mean of 95 mm/s. Is the velocity greater in patients with gingivitis?

Download the data from this link: flow.txt

Common approach

- Based on the t-student distribution with $n-1$ degrees of freedom.
- Applicability conditions. At least one must be met:

1) Variable in analysis follows a Normal distribution.
2) $n \geq 30$.

The hypotheses are:

$$\begin{aligned} H_0: &\quad \mu = 95 \\ H_A: &\quad \mu > 95 \end{aligned}$$

Let's obtain the p-value with the function **t.test(x,alternative,mu)**. The syntax is similar to that of the confidence interval, but now we have to account for the one-sided alternative hypothesis and null hypothesis μ value.

```
t.test(y,alternative="greater",mu=95)
```

```

    One Sample t-test

data:  y
t = 3.0762, df = 29, p-value = 0.002271
alternative hypothesis: true mean is greater than 95
95 percent confidence interval:
 107.4129      Inf
sample estimates:
mean of x
 122.7283
```

Null hypothesis is rejected at significance level $\alpha = 5\%$. We may affirm that the gum blood flow mean is higher in patients with gingivitis.

Alternative approaches

1) **Bootstrap**.

- Use the bootstrap resamples to generate the sample distribution of parameter.
- Problem: we have to generate the distribution under the null hypothesis!

Method. Given that a random sample $x_1, \cdots, x_n$:

1) Center the values in relation to the sample mean

$$z_i = x_i - \overline{x}$$

2) Add to the new z_i values the mean on the null hypothesis

$$z_i^* = z_i + \mu_0$$

3) Generate $j = 1, \cdots, B$ bootstrap resamples with z_i^* and compute the mean at each sample $\overline{Z}_j^*$.
4) Compute the P-value as the relative position of the original sample mean in the set of $\overline{Z}_j^*$.

If the alternative hypothesis is:

- Greater $>$

$$P = \frac{\#\left(\overline{Z}^* > \overline{X}_{obs}\right)}{B}$$

where $\overline{X}_{obs}$ is the mean of x_i.

- Lower $<$

$$P = \frac{\#\left(\overline{Z}^* < \overline{X}_{obs}\right)}{B}$$

- Difference $\neq$

$$P = \frac{\#\left(abs\left(\overline{Z}^{*}\right) < abs\left(\overline{X}_{obs}\right)\right)}{B}$$

Example. A psychological questionnaire has been designed to be filled out in less than 30 minutes as average. To check that a sample of 12 subjects filled out the questionnaire and the time in minutes was recorded.

Download the data from this link: time.txt

The hypotheses are:

$$\begin{aligned} H_0: &\quad \mu = 30 \\ H_A: &\quad \mu < 30 \end{aligned}$$

Conditions. The time to fill out the questionnaire can not be considered to follow a Normal distribution. Furthermore, the sample size is low.

Let's test the hypothesis using bootstrap.

1) Center the data and add the mean on the null hypothesis

```
z<-y-mean(y)+30
```

2) Generate the bootstrap samples. Here the random seed is set to generate the same resamples so you will be able to obtain the same p-value as here when executing the code.

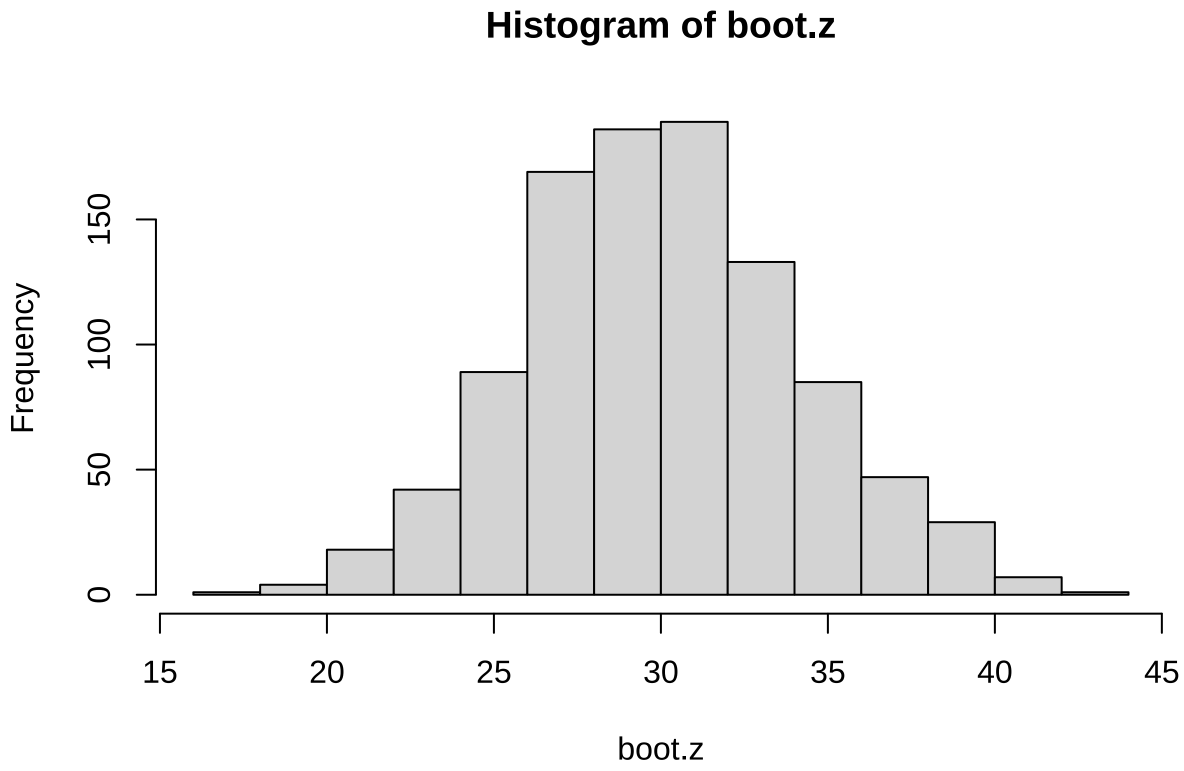
```
set.seed(2020)
boot.z<-replicate(1000,mean(sample(z,replace=T)))
```

3) Compute the p-value

```
sum(boot.z<mean(y))/1000
```

```
[1] 0.165
```

```
hist(boot.z)
```



Using a significance level of 5% the null hypothesis is not rejected.

2) **Non-parametric test**. One-sample Wilcoxon Signed Rank Test.

- Use the ranks of the variable.
- It is widely accepted that the test is about the median rather than the mean.

```
wilcox.test(y,mu=30,alternative="less")
```

```

    Wilcoxon signed rank test with continuity correction

data:  y
V = 26.5, p-value = 0.1728
alternative hypothesis: true location is less than 30
```

Using a significance level of 5% the null hypothesis is not rejected.

Sample size

To compute the sample size we need to answer the following questions before:

- What is the difference to detect? Let's name such difference as δ.
- How much power is desired? $1 - \beta$
- What is the variance of the variable? σ^2

To know the value of σ we may proceed as:

1) Take a pilot sample.
2) Use previous knowledge on σ.

To compute the sample size we will use the function *pwr.t.test* from *pwr* package. The main arguments are:

- *d*. Effect size. The difference to detect δ divided by the standard deviation σ.
- *type*. Type of t test. In this case is **"one.sample"**.

Example. Let's follow with the gum blood flow example.

Let's suppose that we didn't made any previous analysis so we are at the beginning of the study. We do not have any information about σ. A pilot sample of 10 subjects is collected and a value of $s = 45$ is obtained.

We wish to detect as significant at level $\alpha = 5\%$ a difference of $\delta = 10$ with a power of 80%.

```
pwr.t.test(d=10/45,type="one.sample",power=0.8, alternative = "greater")
```

```

     One-sample t test power calculation 

              n = 126.5606
              d = 0.2222222
      sig.level = 0.05
          power = 0.8
    alternative = greater
```

The sample size necessary is 127.

Goodness of fit to probability distribution models

A sample of n data has been collected from the random variable X.

Let $F_0\left(X\right)$ be a concrete probability distribution function (model).

Hypotheses

$$\begin{array}{rl} H_0: & X \sim F_0\left(X\right) \\ H_A: & X \nsim F_0\left(X\right) \end{array}$$

- Most common case: $F_0\left(X\right)$ is the Normal distribution.

Alternatives of analysis:

- Exploratory analysis using Q-Q plots.
- Kolmogorov-Smirnov test.
- Shapiro-Wilks

Example. Let's test if the 100 BMI values from the former example come from a Normal model.

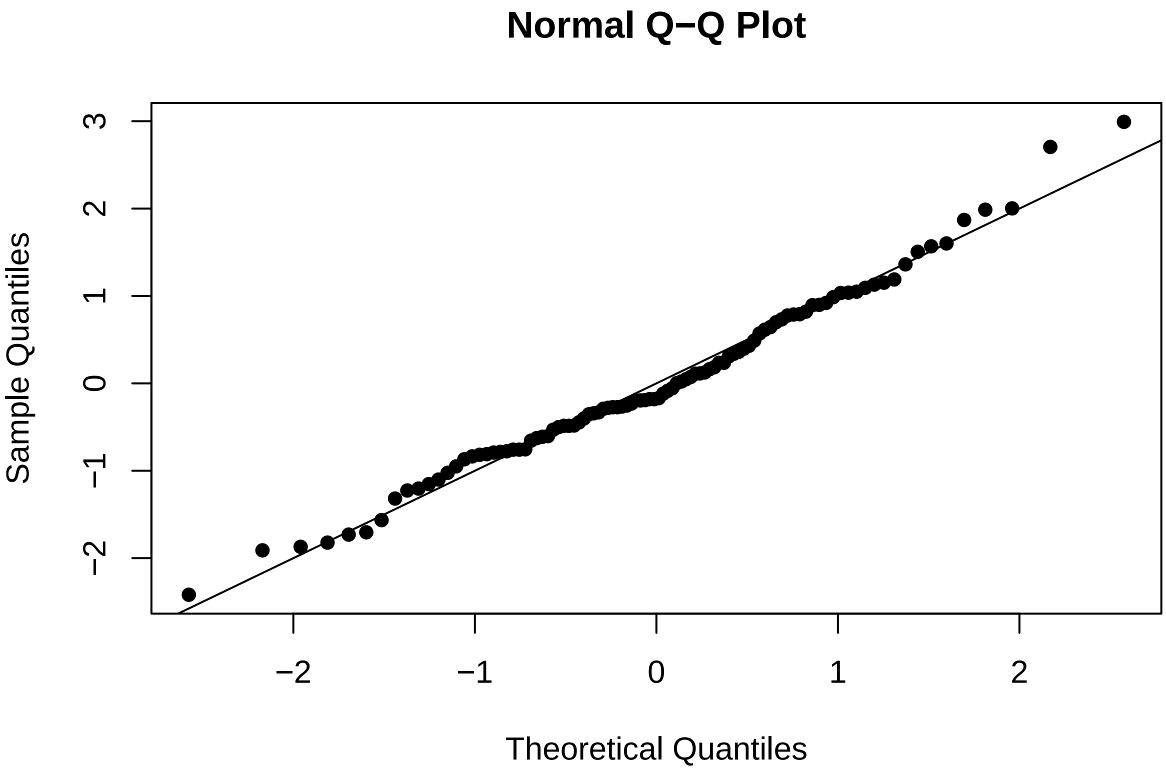
Remember you can download the data from the following link: bmi.txt

Exploratory analysis using Q-Q plots

Let's use the **qqnorm** function that is specific to draw QQ plots for Normal models.

The function *scale* is to standardize the values. This is quite convenient because the function **qqnorm** is going to compare the sample quantiles to those from a standardized Normal model (mean 0, standard deviation 1).

```
qqnorm(scale(bmi),pch=16)
abline(0,1)
```



Most of the points line up on the concordance line. So, it is reasonable to assume that data may come from a Normal model.

Kolmogorov-Smirnov test.

- KS-test is based on comparing the observed (empirical) cumulative probability function with the theoretical one.
- The test statistic D is the maximum distance between the two functions.
- D in a Q-Q plot would be the maximum distance to the concordance line.
- Limitations:
  - Low power with small sample size.
  - Too high power with large sample size.

The test is implemented in R with the function **ks.test**. The first argument is the data values, the next arguments are the model under the null hypothesis and the values of the related parameters.

```
ks.test(x,"pnorm",mean(x),sd(x))
```

```

    One-sample Kolmogorov-Smirnov test

data:  x
D = 0.12777, p-value = 0.809
alternative hypothesis: two-sided
```

The hypothesis of Normal model is not rejected.

Shapiro-Wilks

- It is considered one of the most powerful tests for checking normality.
- Only valid for Normal model

```
shapiro.test(bmi)
```

```

    Shapiro-Wilk normality test

data:  bmi
W = 0.9868, p-value = 0.4246
```

The hypothesis of Normal model is not rejected.