Statistical methods

Lab 4

V. Gioia (and N.Torelli and G. Di Credico)

vincenzo.gioia@units.it

Office hour: Friday, 17.00 - 18.30

25/10/2024

Contents

Interval Estimation **1**
Comparing two population means or proportions 1
Example: Confidence Interval for comparing two means 1
Example: Confidence Interval for comparing two proportions 3

Basic concepts of hypothesis testing **5**
Test for the mean difference 5
Test of the equality of the variances 8
Pearson's chi-squared test 9
Pearson's chi-squared test: Test for independence 9
Goodness of fit test 13

Interval Estimation

Comparing two population means or proportions

We have two populations and we collect data about samples from these populations. Let's denote with $y_{i1}, i = 1, \ldots, n_1$, a certain observed characteristic in the first sample and let y_{i2}, $i = 1, \ldots, n_2$, the measurements on the same characteristic of the second sample. **Here, we are following Section 4.5 of Agresti and Kateri's book.**

Example: Confidence Interval for comparing two means

Let's consider a quantitative variable. Goal: **compare the population means μ_1 and μ_2**

- Parameter: $\mu_1 - \mu_2$
- Parameter estimate: $\bar{y}_1 - \bar{y}_2$
- Estimator: $\bar{Y}_1 - \bar{Y}_2$

Suppose to assume $Y_{i1} \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $i = 1, \ldots, n_1$ and $Y_{i2} \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $i = 1, \ldots, n_2$, with Y_{i1} and Y_{i2} independent. We will also assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Note,

- You cannot/don't assume the normality: you could consider the CLT
- It does not hold $\sigma_1^2 = \sigma_2^2 = \sigma^2$: there exists the Welch test (we can also verify by means of the hypothesis testing if this relation holds)
- In paired sample, under $n_1 = n_2$, Y_{i1} is not independent from Y_{i2}: you can still work on the difference

Two-sided confidence interval for the difference of two means

Let $\overline{Y}_1$ and $\overline{Y}_2$ be the sample mean for the two groups, respectively. Let us identify the pivotal-quantity. Note that you just know that $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but it is a parameter. You need to estimate it. Let's define the **pooled variance estimator**

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Then, the pivotal quantity is defined as

$$T = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Now we know that $T \sim t_{n_1+n_2-2}$ and the $(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$IC^{1-\alpha}_{\mu_1-\mu_2} = (\bar{y}_1 - \bar{y}_2 - t_{n_1+n_2-2;1-\alpha/2} \times s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \bar{y}_1 - \bar{y}_2 + t_{n_1+n_2-2;1-\alpha/2} \times s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}})$$

Example

Let us consider this example. A study makes use of a cognitive behavioral therapy to treat a sample of teenage girls who suffered from anorexia. The study, like most such studies, also had a control group that received no treatment. Then researchers analyzed how the mean weight change compared for the treatment and control groups. The girls in the study were randomly assigned to the cognitive behavioral therapy (Group1) or to the control group (Group2).

Let μ_1 and μ_2 denote the mean weight gains (in pounds) for these groups.

```
Anor <- read.table("http://stat4ds.rwth-aachen.de/data/Anorexia.dat",
                   header=TRUE)
# Get difference post-pre treatment for the group cb and c
cogbehav <- Anor$after[Anor$therapy == "cb"] - Anor$before[Anor$therapy == "cb"]
control <- Anor$after[Anor$therapy == "c"] - Anor$before[Anor$therapy == "c"]
# Get the 95% CI via t.test function
res <- t.test(cogbehav, control, var.equal = TRUE, conf.level = 0.95)
res$conf.int
```

```
## [1] -0.680137  7.593930
## attr(,"conf.level")
## [1] 0.95
```

- The mean weight change for the cognitive behavioral therapy could be as much as 0.68 pounds lower or as much as 7.59 pounds higher than the mean weight change for the control group.
- The interval includes 0: it is plausible that the population means are identical (we can also see the results underlying the hypothesis test)
- The confidence interval is relatively wide: sample sizes are not large.

```
n1 <- XXX
n2 <- XXX

s2 <- XXX

CI <- XXX
```

Example: Confidence Interval for comparing two proportions

For two groups, let π_1 and π_2 denote the population proportions for a binary variable of interest. Suppose to have independent samples Y_{i1}, $i = 1, \ldots, n_1$ and Y_{i2}, $i = 1, \ldots, n_2$ from $Be(\pi_1)$ and $Be(\pi_2)$. To obtain a confidence interval of level $1 - \alpha$ for $\pi_1 - \pi_2$, we leverage the pivotal quantity

$$Z = \frac{\Pi_1 - \Pi_2 - (\pi_1 - \pi_2)}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2}}$$

which is approximately distributed as $\mathcal{N}(0,1)$. Above, we denoted with Π_1 and Π_2 the sample proportion estimators. So the realisations of a confidence interval is given by

$$IC_{1-\alpha}^{\pi_1-\pi_2} = (\hat{\pi}_1 - \hat{\pi}_2 \pm z_{1-\alpha/2}\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_1 + \hat{\pi}_2(1-\hat{\pi}_2)/n_2})$$

Example

A study used patients at six U.S. hospitals who were to receive coronary artery bypass graft surgery. The patients were randomly assigned to two groups. For one group, Christian volunteers were instructed to pray for a successful surgery with a quick, healthy recovery and no complications. The praying started the night before surgery and continued for two weeks. The other group did not have volunteers praying for them. The response was whether medical complications occurred within 30 days of the surgery

Prayer/Complications	Yes	No	Total
Yes	315	289	604
No	304	293	597

Let π_1 and π_2 denote, respectively, the probability of complications for those patients who had a prayer group and for those patients not having a prayer group.

```
success <- c(315, 304)
total <-  c(604, 597)
res <- prop.test(success, total, conf.level = 0.95, correct = FALSE)
res$conf.int
```

```
## [1] -0.04421536  0.06883625
## attr(,"conf.level")
## [1] 0.95
```

```
p1 <- success[1]/total[1]
p2 <- success[2]/total[2]

p1 - p2 + c(-1, 1) * qnorm(0.975) * sqrt(p1 * (1 - p1)/total[1] +
                                           p2 * (1 - p2)/total[2])
```

```
## [1] -0.04421536  0.06883625
```

Basic concepts of hypothesis testing

The null hypothesis for the parameter θ is usually expressed as

$$\mathrm{H}_0 : \theta = \theta_0$$

Complementary to the choice of H_0, we have to specify the alternative hypothesis H_1, specifying the values of the parameter which becomes reasonable when H_0 does not hold. Usually H_1 may be:

- $\mathrm{H}_1 : \theta \neq \theta_0$ (**two-sided alternative**)
- $\mathrm{H}_1 : \theta > \theta_0$ or $\mathrm{H}_1 : \theta < \theta_0$(**one-sided alternative**)

Test for the mean difference

Let's consider again the example on Anorexia, illustrated above. We have two groups and we want ask if the mean weight change between the two groups can be considered as equal.

We could set up a test with the following aim: do the cognitive behavioral group therapy have mean weight change equal to the mean weight change of the control group? Or, do the cognitive behavioral group therapy have mean weight greater than the control group?

Under the same assumptions detailed above we aim to compare their means, μ_1 and μ_2 through the following hypothesis test (consider $\alpha = 0.05$)

two-sided two-sample test

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \quad (\text{equivalently} \quad \mu_1 - \mu_2 \leq 0) \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

one-sided two-sample test

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \quad (\text{equivalently} \quad \mu_1 - \mu_2 \leq 0) \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$$

Assuming that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, and then the test statistic, under H_0 has the form

$$T = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \underset{H_0}{\sim} t_{n_1+n_2-2}$$

with $n_1 + n_2 - 2 = 13$.

The results obtained by using the t.test() function for the two-sided two-sample test are

```
res.two <- t.test(cogbehav, control, var.equal = TRUE)
res.two
```

```
##
##  Two Sample t-test
##
## data:  cogbehav and control
## t = 1.676, df = 53, p-value = 0.09963
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.680137  7.593930
## sample estimates:
## mean of x mean of y
##  3.006897 -0.450000
```

and for the one-sided two-sample test are

```
res.one <- t.test(cogbehav, control, var.equal = TRUE, alternative = "greater")
res.one
```

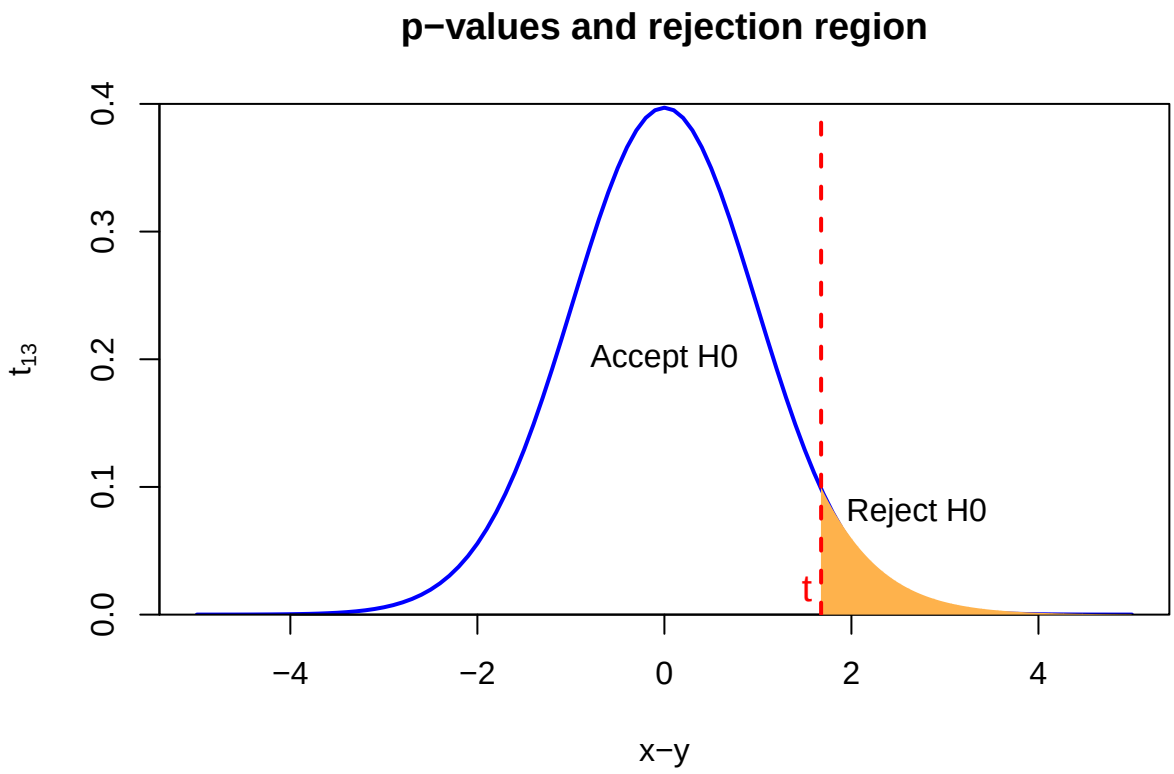
```
##
##  Two Sample t-test
##
## data:  cogbehav and control
## t = 1.676, df = 53, p-value = 0.04981
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.003879504         Inf
## sample estimates:
## mean of x mean of y
##  3.006897 -0.450000
```

Time to working with R: it's your turn Perform th test by hand (as ususally complete the XXX part)

```
testStat <- XXX

## two-sided
pvalue.two <- XXX

## one-sided
pvalue.one <- XXX
```

What are the conclusions? Take a look to https://doi.org/10.1080/00031305.2016.1154108

```r
library(RColorBrewer)
plotclr <- brewer.pal(6, "YlOrRd")

curve(dt(x, n1 + n2 - 2), xlim = c(-5, 5), ylim = c(0, 0.4),
      main = "p-values and rejection region", col = "blue",
      lwd = 2, xlab = "x-y",  ylab = expression(t[13]),  yaxs="i")
cord.x <- c(qt(0.95, n1 + n2 - 2),seq(qt(0.95, n1 + n2 - 2), 5, 0.01), 5)
cord.y <- c(0, dt(seq(qt(0.95, n1 + n2 - 2), 5, 0.01), 13), 0)
polygon(cord.x, cord.y, col = plotclr[3], border = NA )


abline(v = res.one$statistic, lty = 2, lwd = 2, col = "red")
text(0, 0.2, paste("Accept", expression(H0)))
text(2.7, 0.08, paste("Reject", expression(H0)))
text(as.double(res.one$statistic) - 0.15, 0.02, "t", col = "red", cex = 1.2)
```



Test of the equality of the variances

Let us suppose to have two normally distributed populations. The test of the equality of the variances is useful for example when we want verify the assumption of equal variances before carrying out the test for the mean difference when the variances are unknown.

For two independent samples with sample sizes n_1 and n_2, respectively, from $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ we want test:

$$\begin{cases} \mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 \\ \mathrm{H}_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Under H_0 (equal variances) the test statistic

$$T = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}$$

has an F-distribution with numerator degrees of freedom $n_1 - 1$ and denominator degrees of freedom $n_2 - 1$ (denoted with F_{n_1-1,n_2-1}). S_1^2 and S_2^2 are the unbiased sample variance estimators.

Then, by computing s_1^2 and s_2^2 and so $t_0 = s_1^2/s_2^2$, we reject H_0 if $t_0 < f_{n_1-1,n_2-1,\frac{\alpha}{2}}$ or if $t_0 > f_{n_1-1,n_2-1,1-\frac{\alpha}{2}}$.

$$p - value = 2\min(P(T < t_0), P(T > t_0))$$

```
# Test for equality of variance using the var.test function
var.test(cogbehav, control, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  cogbehav and control
## F = 0.83696, num df = 28, denom df = 25, p-value = 0.6449
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3805749 1.8090469
## sample estimates:
## ratio of variances
##          0.8369592
```

```
# Test for equality of variance by hand
rationvar <- var(cogbehav)/var(control) # Test statistic
pv_bi <- 2 * min(pf(rationvar, n1 - 1, n2 - 1, lower = FALSE),
                 1 - pf(rationvar, n1 - 1, n2 - 1, lower = FALSE))
pv_bi
```

```
## [1] 0.6448602
```

Pearson's chi-squared test

This is a class of test applied to sets of categorical data to evaluate whether any observed difference between the sets arose by chance. It is suitable for unpaired data from large samples. Pearson's chi-squared test is used to assess three types of comparison:

- **Goodness of fit**: establishes whether an observed frequency differs from a theoretical distribution;
- **Homogeneity**: test if two or more sub-groups of a population share the same distribution of a single categorical variable;
- **Independence**: determines whether two categorical variables are associated

In all the cases the test statistic is, under H_0, distributed according to the Chi-square distribution with said degrees of freedom.

Pearson's chi-squared test: Test for independence

Question: is there a relationship (association) between two categorical variables?

We want carry out the hypothesis test

$$\begin{cases} H_0 : \text{there is no relationship between the categorical variables} \\ H_1 : \text{there is relationship between the categorical variables} \end{cases}$$

H_1 is not one-sided or two-sided: sometimes it is referred to as 'many-sided' since it allows any kind of difference. However, H_1 says that there is a relationship but does not specify any particular kind of relationship.

To test H_0, we compare the observed counts with the expected counts, that is the counts that we would expect if H_0 were true. If the observed counts are far from the observed counts, there is evidence against H_0.

Then, the data are organised in a two-way table of (observed) counts or contingency table. Thus, let X and Y be the two categorical variables, with s and t categories, respectively.

Y/X	x_1	$\cdots$	x_j	$\cdots$	x_t	Total
y_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1t}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
y_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{it}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
y_s	n_{s1}	$\cdots$	n_{sj}	$\cdots$	n_{st}	$n_{s\cdot}$
Total	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot t}$	$n_{\cdot\cdot}$

- n_{ij}= number of couples (y_i, x_j), $i = 1, \ldots, s$, $j = 1, \ldots, t$ (absolute frequencies)
- $n_{i\cdot} = \sum_{j=1}^{t} n_{ij}, i = 1, \ldots, s$ (marginal frequencies of y_i)
- $n_{\cdot j} = \sum_{i=1}^{s} n_{ij}, k = 1, \ldots, t$ (marginal frequencies of x_j)

- $n_{\cdot\cdot} = n = \sum_{i=1}^{s} \sum_{j=1}^{t} n_{ij}$ (total sample size)

The test statistic that allows the comparison between observed and expected counts is the Pearson's chi-squared statistic, which is a measure of how far the observed counts in the two-way table are from the expected counts. It is always positive and it is zero only when the observed counts are exactly equal to the expected counts.

Then

- Large value of the statistic are evidence against H_0
- Small values of the statistic do not provide evidence against H_0

Note that even though H_1 is many-sided, the Pearson's chi-squared test is one-sided because any violation of H_0 tends to produce a large value in the statistics.

Then, under H_0 leveraging the independence

$$P(X = x_i, Y = y_j) = P(X = x_i) \times P(Y = y_j),$$

for any $i = 1, \ldots, s$ and $j = 1, \ldots, t$, we can obtain the table of expected frequencies, denoted as n^*_{ij}, under the H_0

Y/X	x_1	$\cdots$	x_j	$\cdots$	x_t
y_1	n^*_{11}	$\cdots$	n^*_{1j}	$\cdots$	n^*_{1t}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_i	n^*_{i1}	$\cdots$	n^*_{ij}	$\cdots$	n^*_{it}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_s	n^*_{s1}	$\cdots$	n^*_{sj}	$\cdots$	n^*_{st}

where

$$n^*_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

Then, the X^2 Pearson statistic is

$$X^2 = \sum \frac{(N_{ij} - n^*_{ij})^2}{n^*_{ij}} \underset{H_0}{\dot{\sim}} \chi^2_{(s-1)(t-1)}$$

Then, using the observed frequencies, n_{ij}, and the expected frequencies, n^*_{ij}, we can obtain the observed test statistics

$$t_0 = \sum \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

By fixing the significance level α, we reject H_0 if $t_0 > \chi^2_{(s-1)(t-1);1-\alpha}$ or equivalently if

$$t_0 \in \mathcal{R}_\alpha = (\chi^2_{(s-1)(t-1);1-\alpha}, +\infty)$$

Example

Let us consider the following example to assess whether two qualitative chracteristics are independent. The blood sample can be classified into four groups A, B, AB and O. To assess if the distribution of the blood sample is independent by the membership to the Italian macro-region ("South", "Center", "North"), we consider a sample of 760 subjects. The following table shows the joint distribution of the blood sample (X) and the membership to the macro-region. (Y).

	A	B	AB	O
South	50	70	30	100
Center	114	30	10	100
Nord	116	27	13	100

Basically you observed the table of absoulte frequencies

	x_1	x_2	x_3	x_4	Total
y_1	n_{11}	n_{12}	n_{13}	n_{14}	$n_{1\cdot}$
y_2	n_{21}	n_{22}	n_{23}	n_{24}	$n_{2\cdot}$
y_3	n_{31}	n_{32}	n_{33}	n_{34}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	$n_{\cdot 4}$	n

you can derive the relative frequencies $f_{jk} = n_{jk}/n$, $f_{j\cdot} = n_{j\cdot}/n$ and $f_{\cdot k} = n_{\cdot k}/n$ and perform the following hypothesis test

$$\begin{cases} H_0 : f_{jk} = f_{j\cdot} f_{\cdot k} \ \ \forall j = 1, \dots, J, \ \ k = 1, \dots, K \\ H_1 : \exists!(j,k) : f_{jk} \neq f_{j\cdot} f_{\cdot k} \end{cases}$$

```
n <- 760
obs_freq <- matrix(c(50, 70, 30, 100,
                     114, 30, 10, 100,
                     116, 27, 13, 100),3,4,T)
colnames(obs_freq) <- c("A", "B", "AB", "O")
rownames(obs_freq) <- c("South", "Central", "North")
chisq.test(obs_freq)
```

```
##
##  Pearson's Chi-squared test
##
## data:  obs_freq
## X-squared = 70.99, df = 6, p-value = 2.561e-13
```

The results obtained by mean of the **chisq.test** function are obtained as

```
mx <- colSums(obs_freq)/n; mx
```

```
##          A          B         AB          O
## 0.36842105 0.16710526 0.06973684 0.39473684
```

```r
my <- rowSums(obs_freq)/n; my
```

```
##     South   Central     North
## 0.3289474 0.3342105 0.3368421
```

```r
exp_freq <- outer(my, mx) * n; exp_freq
```

```
##                A        B       AB         O
## South   92.10526 41.77632 17.43421  98.68421
## Central 93.57895 42.44474 17.71316 100.26316
## North   94.31579 42.77895 17.85263 101.05263
```

```r
chi2 <- sum((obs_freq - exp_freq)^2/exp_freq); chi2
```

```
## [1] 70.99012
```

```r
pchisq(chi2, (ncol(obs_freq) - 1) * (nrow(obs_freq) - 1), lower.tail = FALSE)
```

```
## [1] 2.561274e-13
```

Exercise What if a remove the "South" macro-region. Are the conclusions of the test different?

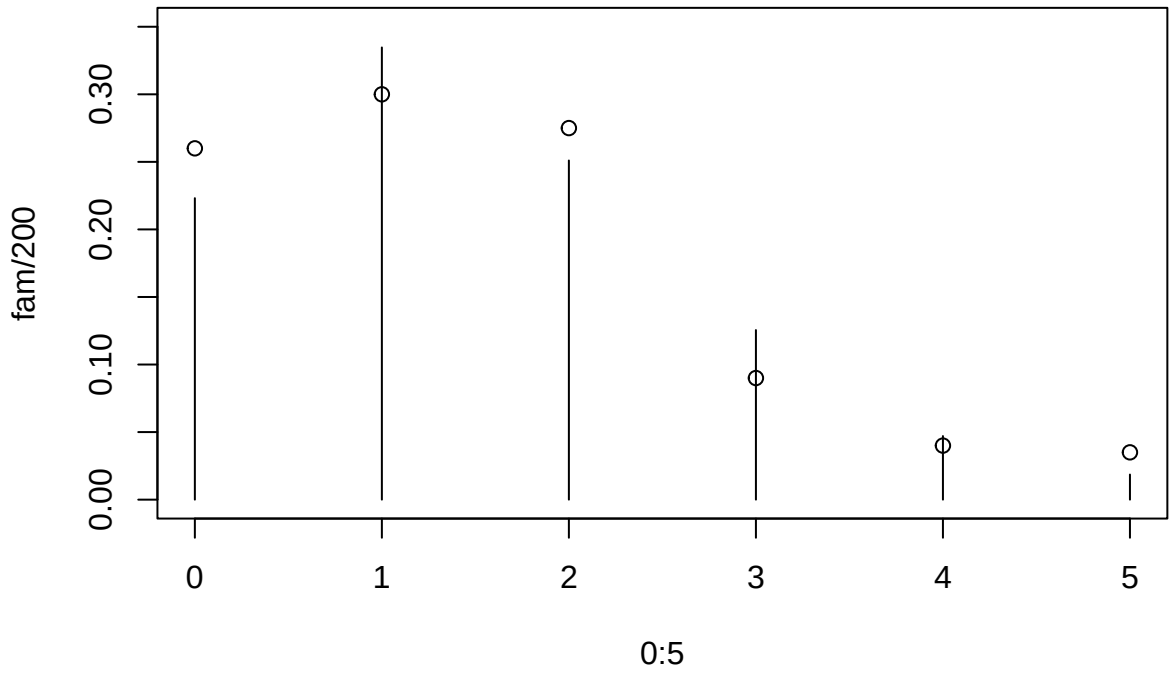
Goodness of fit test

The number of children in a household survey of 200 households is

Number of children	0	1	2	3	4	≥ 5
Number of families	52	60	55	18	8	7

Goal: perform an hypothesis testing to assess whether it is reasonable to assume that X (number of children in a family) is distributed according to a $Poisson(\lambda)$, with $\lambda = 1.5$.

```
child <- 0:5
fam <- c(52, 60, 55, 18, 8, 7)
lambda <-  1.5
plot(0:5, fam/200, ylim = c(0,0.35))
segments(0:5,rep(0,5), 0:5,
         c(dpois(0:4, lambda),ppois(4, lambda, lower = FALSE)))
```




```
c(dpois(0:4, lambda), ppois(4, lambda, lower = FALSE))
```

```
## [1] 0.22313016 0.33469524 0.25102143 0.12551072 0.04706652 0.01857594
```

The null hypothesis $H_0 : Y \sim \mathcal{P}(\lambda = 1.5)$ translates into

$$H_0 : \pi_0 = 0.2231, \quad \pi_1 = 0.3347, \quad \pi_2 = 0.2510, \quad \pi_3 = 0.1255, \quad \pi_4 = 0.0471 \quad \pi_5 = 0.0186$$

where $\pi_j = P(X = j)$, $j = 0, \ldots, 4$, and $\pi_5 = P(X \geq 5)$.

The test is based on the comparison between the observed frequencies and the expected frequencies under H_0 and it makes use of the chi-square test statistics. The needed quantities are reported in the table below

Number of Children	0	1	2	3	4	≥ 5
Observed frequencies (N_i)	52	60	55	18	8	7
Expected frequencies ($N \times \pi_i$)	44.6260	66.9390	50.2043	25.1021	9.4133	3.7152
$\frac{(N_i - N\pi_i)^2}{N\pi_i}$	1.2185	0.7193	0.4581	2.0094	0.2122	2.9043

which are obtained in R as

```
exp <- c(dpois(0:4, lambda), ppois(4, lambda, lower = FALSE)) * 200
round(exp, 4)
```

```
## [1] 44.6260 66.9390 50.2043 25.1021  9.4133  3.7152
```

```
chisq_el <- (fam-exp)^2/exp
round((fam-exp)^2/exp, 4)
```

```
## [1] 1.2185 0.7193 0.4581 2.0094 0.2122 2.9043
```

The test is based on

$$\chi^2 = \sum_{j=1}^{5} \frac{(N_j - N\pi_j)^2}{N\pi_j}$$

that for large samples $\chi^2 \dot{\sim} \chi^2_{K-1}$. The reject region is

$$\mathcal{R} = \{\chi^2 > \chi^2_{K-1;1-\alpha}\} = \{\chi^2 > \chi^2_{K-1;1-\alpha}\} = \{\chi^2 > 11.0705\}$$

The observed test statistic is

$$\chi^2 = \sum_{j=1}^{5} \frac{(N_j - N\pi_j)^2}{N\pi_j} = 7.521784$$

This value does not belong to the reject region and we do not have evidence to discard H_0. P-value, having denoted with $Q = \chi^2_{K-1}$; is

$$p - value = P(Q > \chi^2) = 1 - P(Q \leq \chi^2) = 1 - F_Q(7.521784) = 0.18$$

```
chisq.obs <- sum(chisq_el)
chisq.obs
```

```
## [1] 7.521784
```

```
pchisq(chisq.obs, df = 5, lower=FALSE)
```

```
## [1] 0.1846351
```

```r
chisq.test(fam, p = exp/200)
```

```
## Warning in chisq.test(fam, p = exp/200): L'approssimazione al Chi-quadrato
## potrebbe essere inesatta
```

```
##
##  Chi-squared test for given probabilities
##
## data:  fam
## X-squared = 7.5218, df = 5, p-value = 0.1846
```