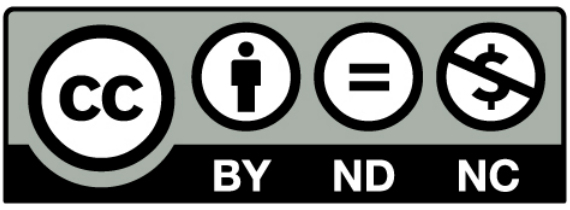
Research techniques

Applications of Probability models & Descriptive statistics

Rumen Manolov & Antonio Solanas



En gård i blom
När mina flickor kom
Jag höll andan och log

Contents

1 Aim of the document

The current document is intended to be a complement to the readings and in-class sessions of the Research techniques course (*Técnicas de investigación*). The focus is put specifically on the second part of the course, dealing with probability models and descriptive statistics. We offer examples that can be used exercises, accompanied by solutions in graphical and numerical form. This document contains all major topics of the course, but it does not substitute the textbooks recommended, attending sessions, or discussing with the teacher. Specifically, as a reading in English, we recommend consulting Gravetter and Wallnau (2009); in Spanish, Solanas, Salafranca, Fauquet, and Núñez (2005).

The examples used for working with the probability are either fictitious (but plausible) or stem from empirical data, if so specified. The univariate and bivariate analysis of categorical data refers to the answers gathered by students of Research techniques during the academic courses 2012-2013 and 2013-2014 via an interview to smokers. This data collection was part of the activities used to learn the methodological content of the course. The univariate and bivariate analysis of quantitative data refers to the partial grades obtained by the students during the academic course 2015-2016, maintaining the anonymity. The control in which the grades were obtained consisted of two parts: a multiple-choice test and open-ended questions, with the maximal score in both of them being 1.5 points.

For obtaining the output, R (R Core Team, 2013) code is used. More information about R in English is given by John Verzani (`http://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf`) and about R-Commander by John Fox and Milan Bouchet-Valat (`http://socserv.socsci.mcmaster.ca/jfox/Misc/Rcmdr/Getting-Started-with-the-Rcmdr.pdf`). In Spanish, we recommend the textbook by Peró, Leiva, Guàrdia, & Solanas (2012), in which it is explained how statistical content is implemented in R and R-Commander.

If any part of the content remains unclear, we encourage students to discuss it with their teachers in the course in order to achieve full understanding of the content.

2 Random variables

2.1 Moments of a distribution of a discrete random variable

$$E(X) \equiv \mu = \sum_{j=1}^{k} \left(x_j \times Prob(x_j)\right) = \frac{\sum_{i=1}^{n} y_i}{n} \quad units$$

$$Var(X) \equiv \sigma^2 = \sum_{j=1}^{k} \left((x_j - \mu)^2 \times Prob(x_j)\right) = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{n} \quad units^2$$

$$CV = \frac{\sigma}{\mu}$$

$$Skewness(X) \equiv \gamma_1 = \frac{\sum_{j=1}^{k} \left((x_j - \mu)^3 \times Prob(x_j)\right)}{\left(\sqrt{\sum_{j=1}^{k} \left((x_j - \mu)^2 \times Prob(x_j)\right)}\right)^3} = \frac{\frac{\sum_{i=1}^{n} (y_i - \mu)^3}{n}}{\sigma^3}$$

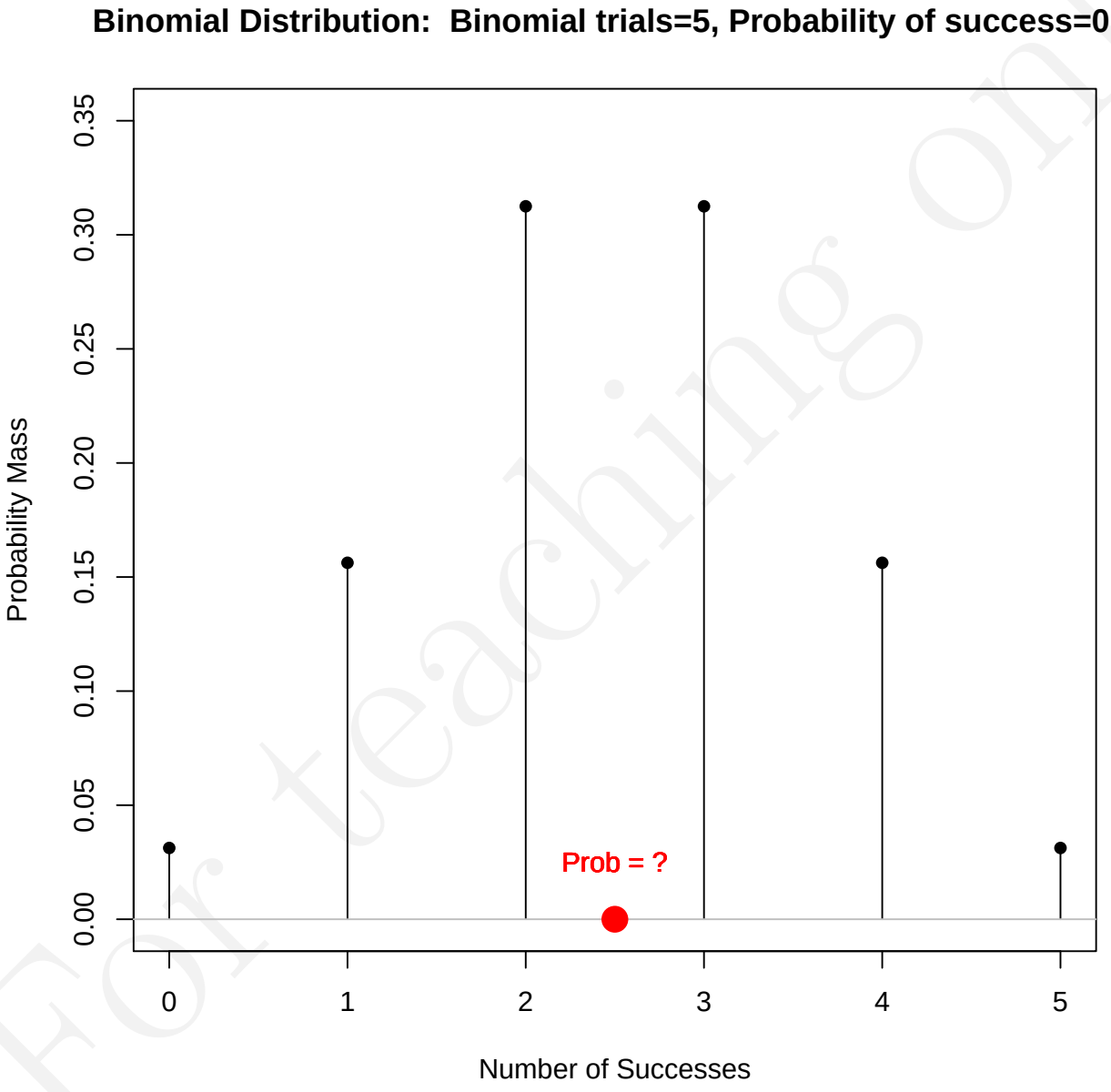
$$Kurtosis(X) \equiv \gamma_2 = \frac{\sum_{j=1}^{k} \left((x_j - \mu)^4 \times Prob(x_j)\right)}{\left(\sqrt{\sum_{j=1}^{k} \left((x_j - \mu)^2 \times Prob(x_j)\right)}\right)^4} - 3 = \frac{\frac{\sum_{i=1}^{n} (y_i - \mu)^4}{n}}{\sigma^4} - 3$$

where k denotes the number of different possible values of the random variable (e.g., $k = 6$ possible results for "number of heads" when tossing a coin 5 times: $x_1 = 0, x_2 = 1, x_3 = 2, x_4 = 3, x_5 = 4, x_6 = 5$). Note that we here define the random experience as tossing the coin five times, that is, there are five Bernoulli trials. Such a random experience can be repeated several times. Specifically, we here used n to denote the number of repetitions of the random experience. That is, the number of repetitions of the five Bernoulli trials. Each of these n repetitions of the random experience can lead to a different value of the random variable *number of heads in 5 tosses of coin*. The admissible values of the random variable are between 0 and 5. Therefore, in several repetitions (e.g., $n = 10$), some values of the random variable can appear more than once. For instance, for repetitions y_3 and y_{10} only one head is obtained, which would correspond to one of the six possible results of the random variable, $x_2 = 1$).

2.2 Mathematical expectancy: example

Example: 5 tosses of a fair coin. Binomial distribution: mass probability function

2.2.1 Plot of the mass probability function



2.2.2 Obtaining the probabilities

```
dbinom(c(0,1,2,3,4,5), size=5, prob=0.5)
```

```
## 0.03125 0.15625 0.3125 0.3125 0.15625 0.03125
```

Formula referring to discrete probability models for which there is a finite number k of possible values

$$
\begin{aligned}
E(X) \equiv \mu &= \sum_{j=1}^{k} \left(x_j \times Prob(x_j)\right) \\
&= (0 \times 0.03125) + (1 \times 0.15625) + (2 \times 0.3125) + (3 \times 0.3125) + (4 \times 0.15625) + (5 \times 0.03125) \\
&= 0 + 0.15625 + 0.625 + 0.9375 + 0.625 + 0.15625 \\
&= 2.5
\end{aligned}
$$

The previous calculations can be performed in R with a few lines of code:

```
values <- 0:5
massprobs <- dbinom(values, size=5, prob=0.5)
sum(values*massprobs)
```

```
## 2.5
```

2.2.3 Simulation of empirical distributions

Computing the mean for several repetitions of the random experiment, that is, for several sample size ($n = 10$, $n = 100$, and $n = 1000$), where the samples represent realizations of a binomial process.

It can be seen that this second expression for mathematical expectancy, referring to samples, leads to the same result as the first expression as sample size increases, technically when $n \to \infty$.

In the following we show the R code that can be used to draw samples of different sizes from a binomial distribution (i.e., the population) and to apply the formula $\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n}$ via the R function **mean()**.

```
ten <- rbinom(size=5, prob=0.5, n=10)
mean(ten)
```

```
## 3.1
```

```
onehundred <- rbinom(size=5, prob=0.5, n=100)
mean(onehundred)
```

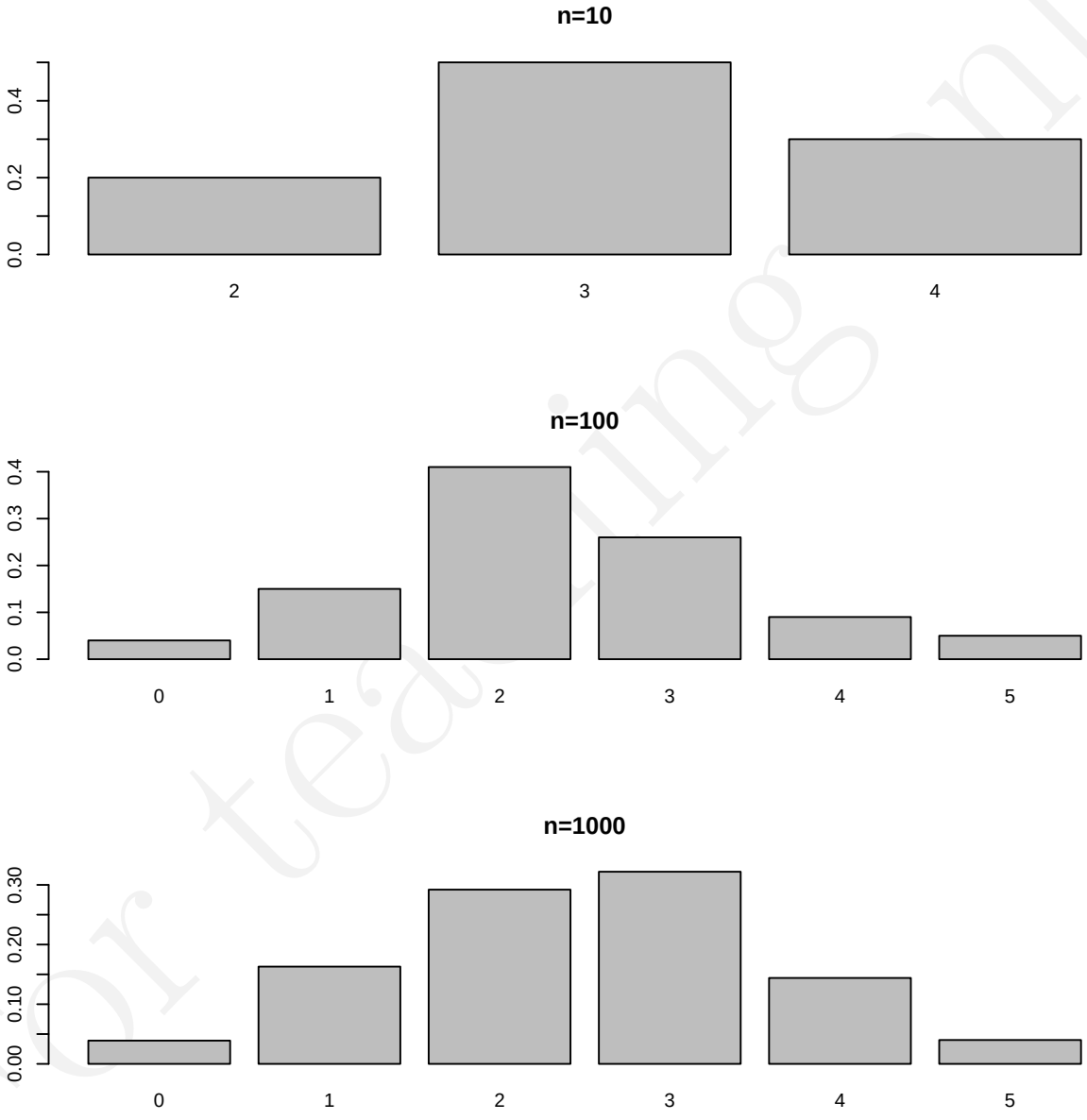
```
## 2.36
```

```
onethousand <- rbinom(size=5, prob=0.5, n=1000)
mean(onethousand)
```

```
## 2.489
```

To construct graphical representations of the relative frequency of appearance of each of the 6 possible results.

```
par(mfrow=c(3,1))
barplot(table(ten)/length(ten),main="n=10")
barplot(table(onehundred)/length(onehundred),main="n=100")
barplot(table(onethousand)/length(onethousand),main="n=1000")
```



2.2.4 Specific formula for the Binomial distribution

The mathematical expectancy for a random variable following the binomial distribution can be computed using a simpler formula than the one presented previously:

$$\begin{aligned} E(X) &= \pi \times n \\ &= 0.5 \times 5 \\ &= 2.5 \end{aligned}$$

3 Discrete probability models

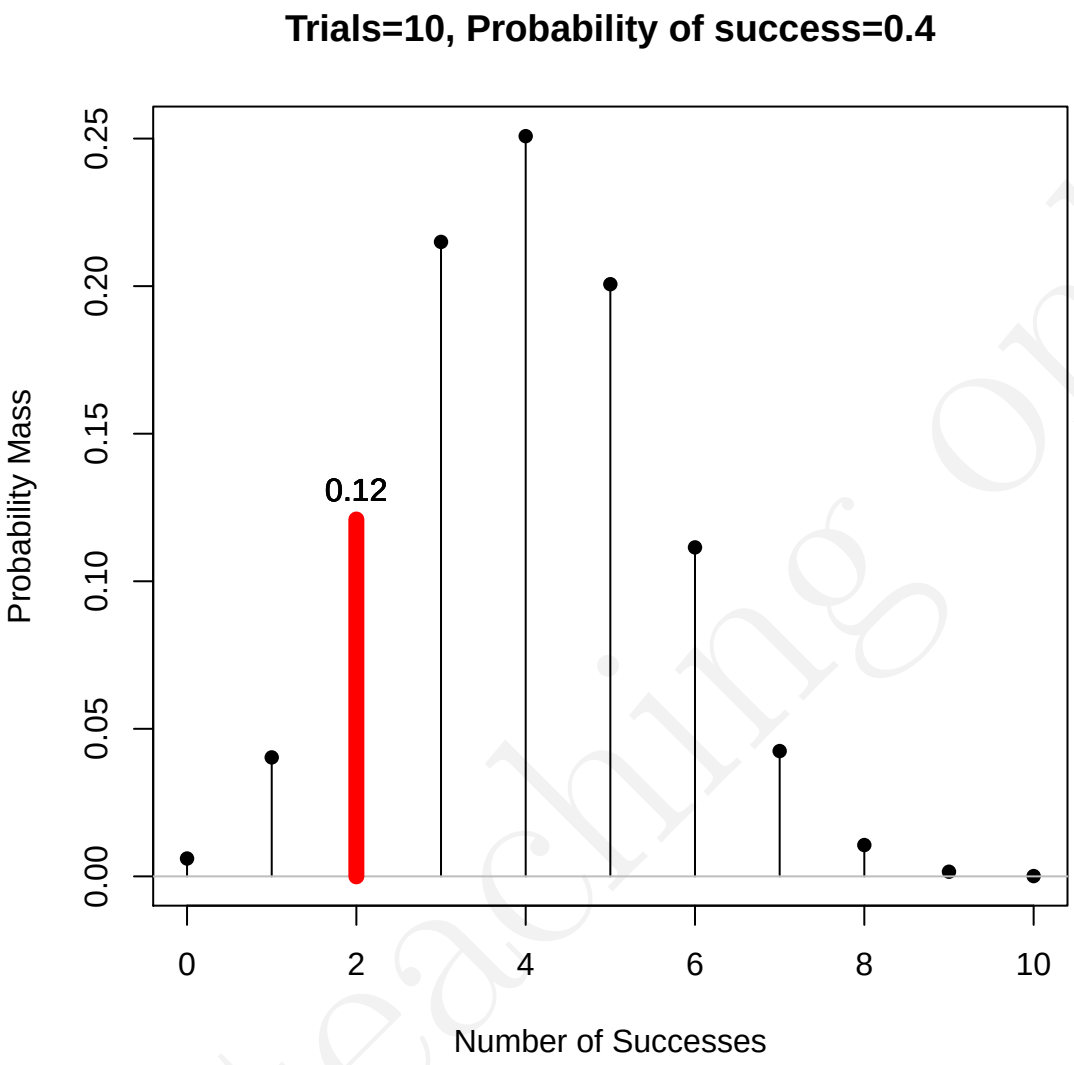
3.1 Binomial model

```
# Parameters of the model
n <- 10
pi <- 0.4
```

1. Mass probability function of 2:

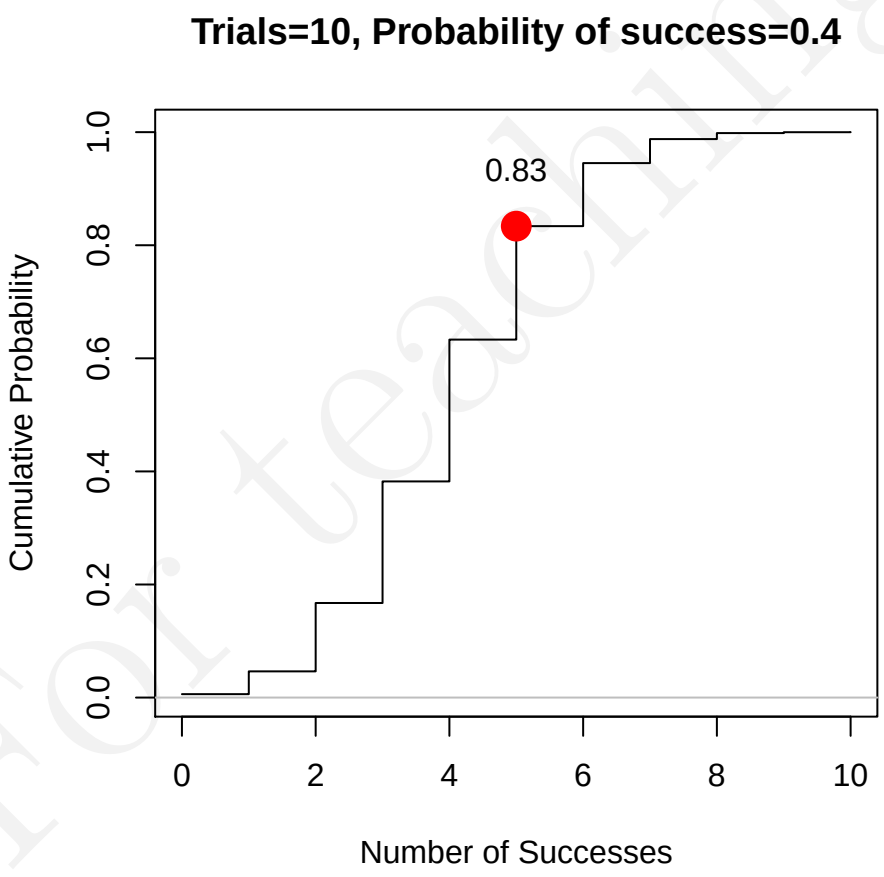
$$Prob(X = k) = \binom{n}{k} \times \pi^k \times (1-\pi)^{n-k}$$
$$= \frac{n!}{k! \times (n-k)!} \times \pi^k \times (1-\pi)^{n-k}$$

```
k <- 2
# 1st element
factorial(n)/(factorial(k)*factorial(n-k))

## [1] 45

# 2nd element
pi^k

## [1] 0.16

# 3rd element
(1-pi)^(n-k)

## [1] 0.01679616

# Find the product
(factorial(10)/(factorial(2)*factorial(8)))*0.4^2*0.6^8

## [1] 0.1209324

# Mass probability function
dbinom(c(2), size=10, prob=0.4)

## [1] 0.1209324
```



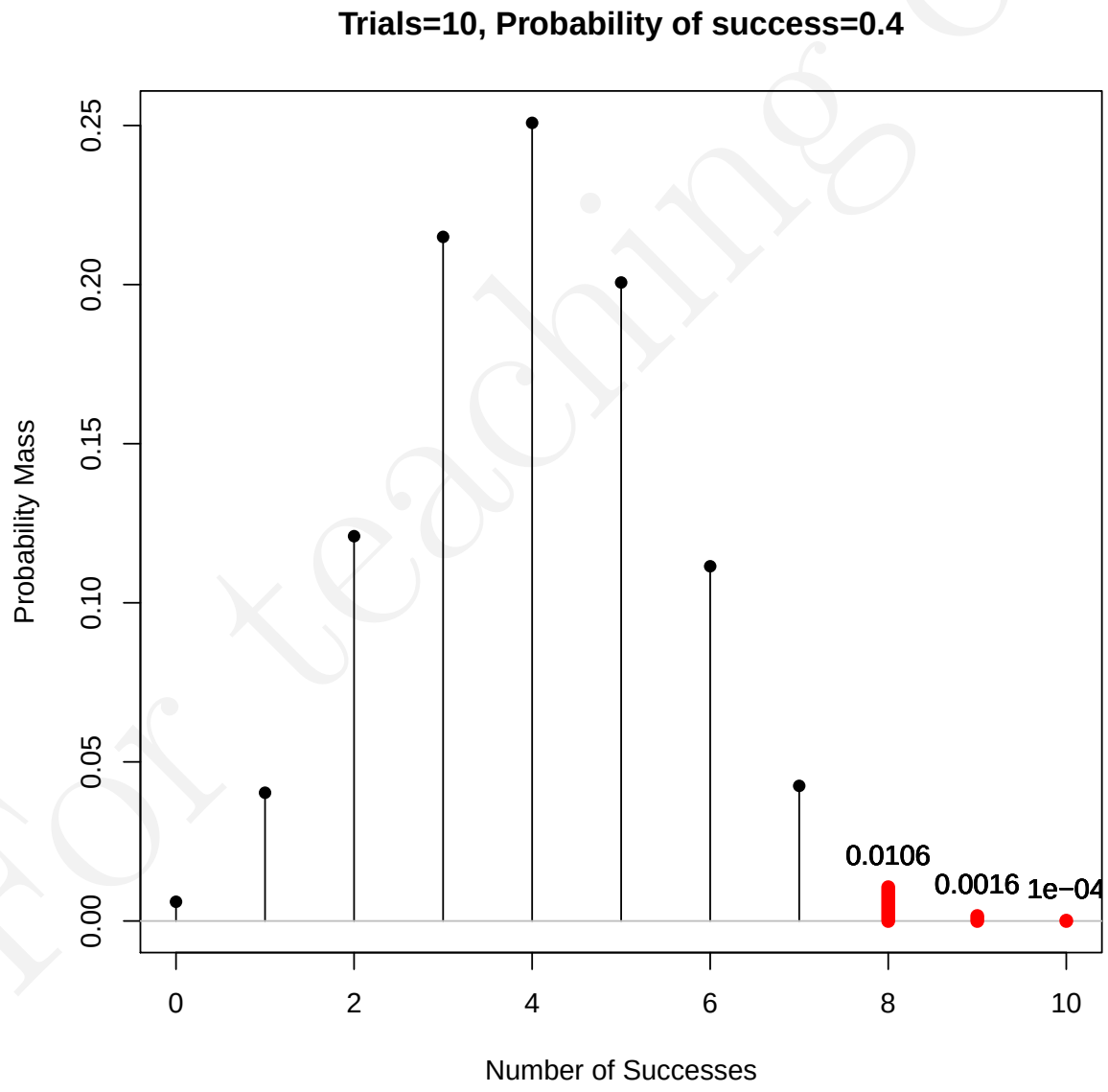
2. Distribution function of 5
$F(5) = \sum_{i=0}^{5} \left(Prob(X = i)\right)$

```
k <- 5
# Initiate sum to 0
distr.f.k <- 0
# Go on adding mass probability backwards
for (i in 0:k)
  distr.f.k <- distr.f.k +
               (factorial(n)/(factorial(i)*factorial(n-i)))*(pi^i)*((1-pi)^(n-i))
distr.f.k

## [1] 0.8337614

# Distribution function
pbinom(c(5), size=10, prob=0.4, lower.tail=TRUE)

## [1] 0.8337614
```

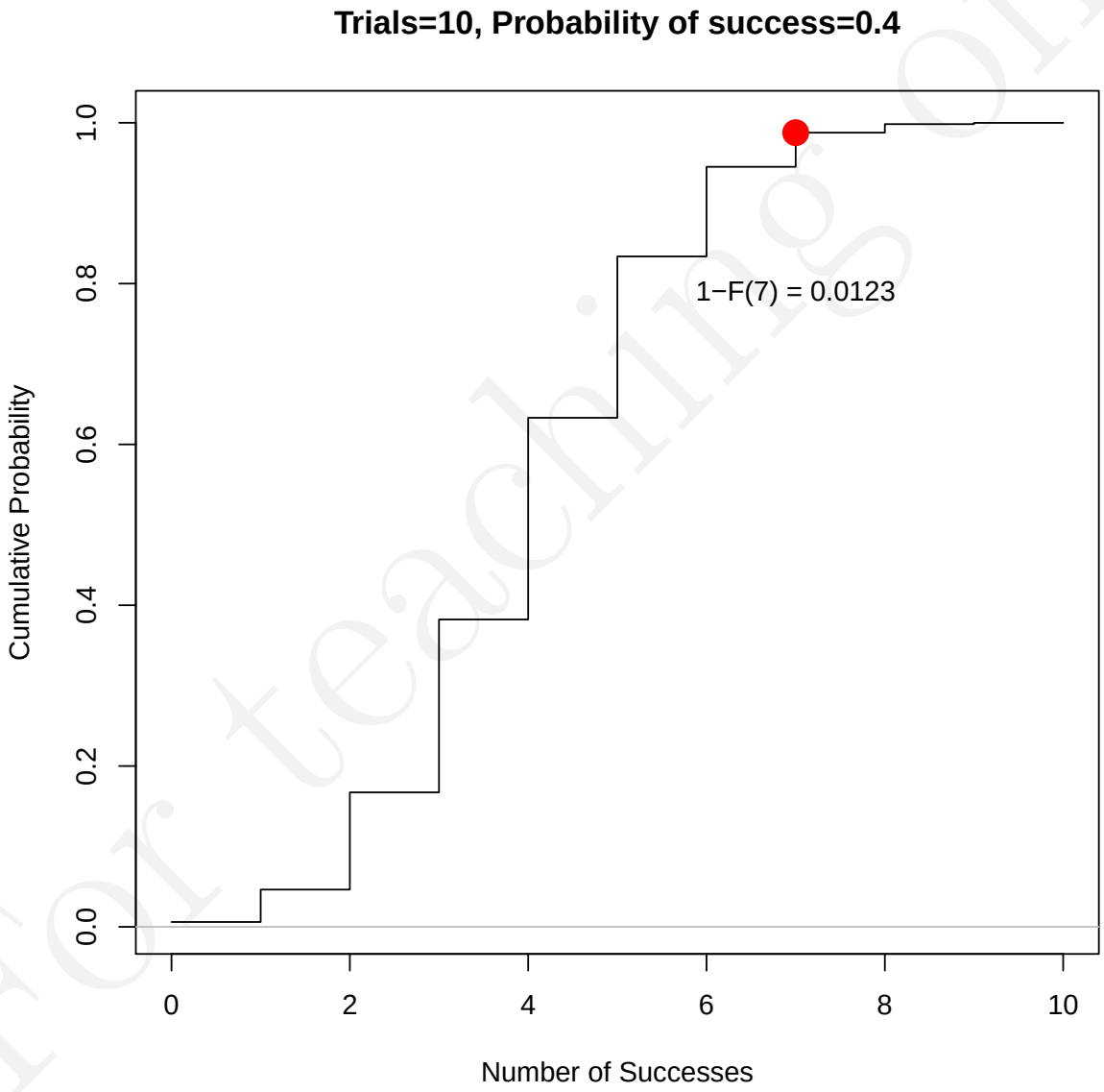


3. Survival function of 7
$S(7) = \sum_{i=7+1}^{10} \left(Prob(X = i)\right)$

```
k <- 7
# Initiate sum to 0
surv.f.k <- 0
# Go on adding mass probabilities forwards
for (i in (k+1):10)
  surv.f.k <- surv.f.k +
              (factorial(n)/(factorial(i)*factorial(n-i)))*(pi^i)*((1-pi)^(n-i))
surv.f.k

## [1] 0.01229455
```



```r
# Survival function
pbinom(c(7), size=10, prob=0.4, lower.tail=FALSE)

## [1] 0.01229455

# Complementary to distribution function
1 - pbinom(c(7), size=10, prob=0.4, lower.tail=TRUE)

## [1] 0.01229455
```

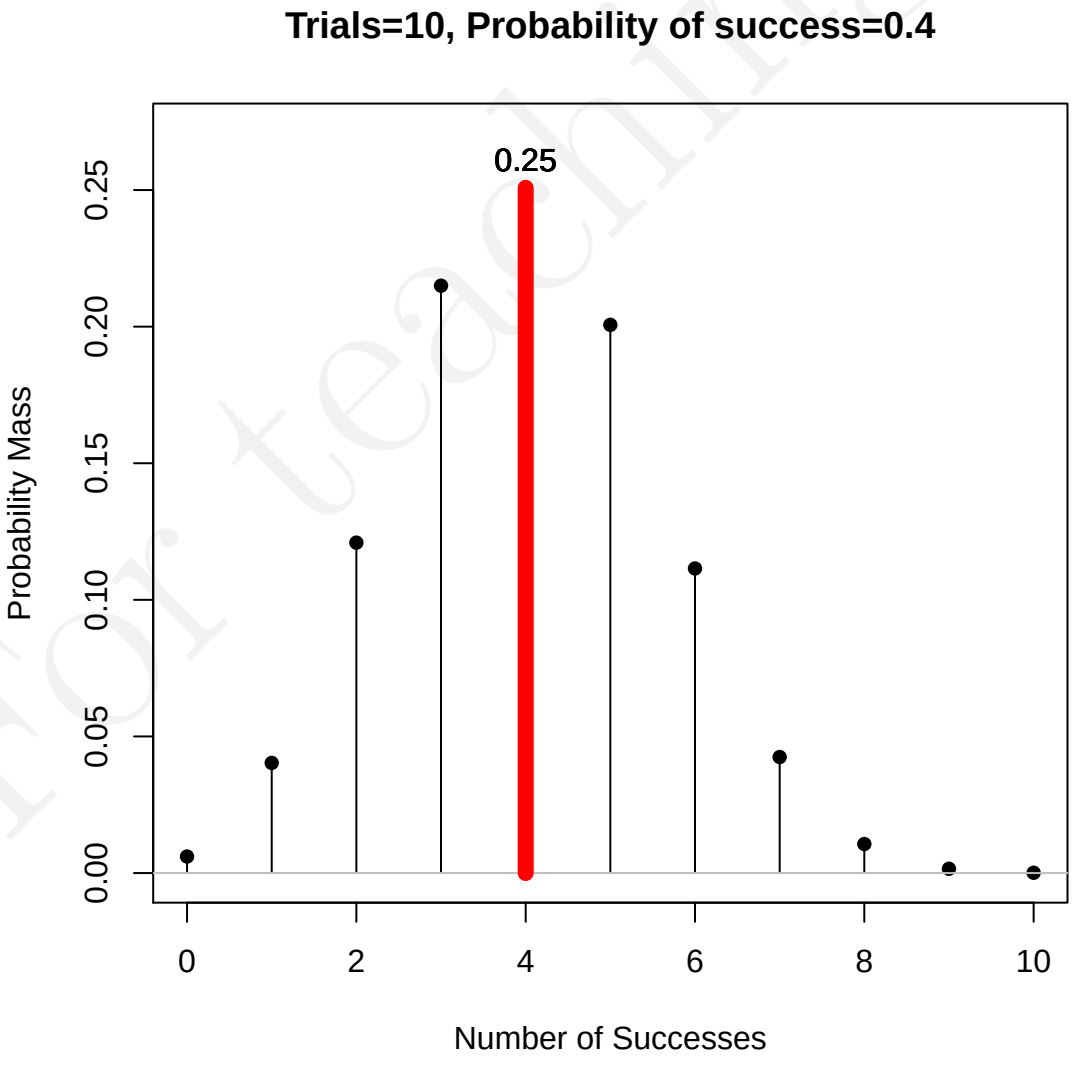


4. Mode of this binomial distribution: M
   It is an integer that satisfies $(n+1) \times \pi - 1 \leq M < (n+1) \times \pi$

$$\begin{aligned}(n+1) \times \pi - 1 &= (10+1) \times 0.4 - 1 \\ &= 11 \times 0.4 - 1 \\ &= 4.4 - 1 \\ &= 3.4\end{aligned}$$

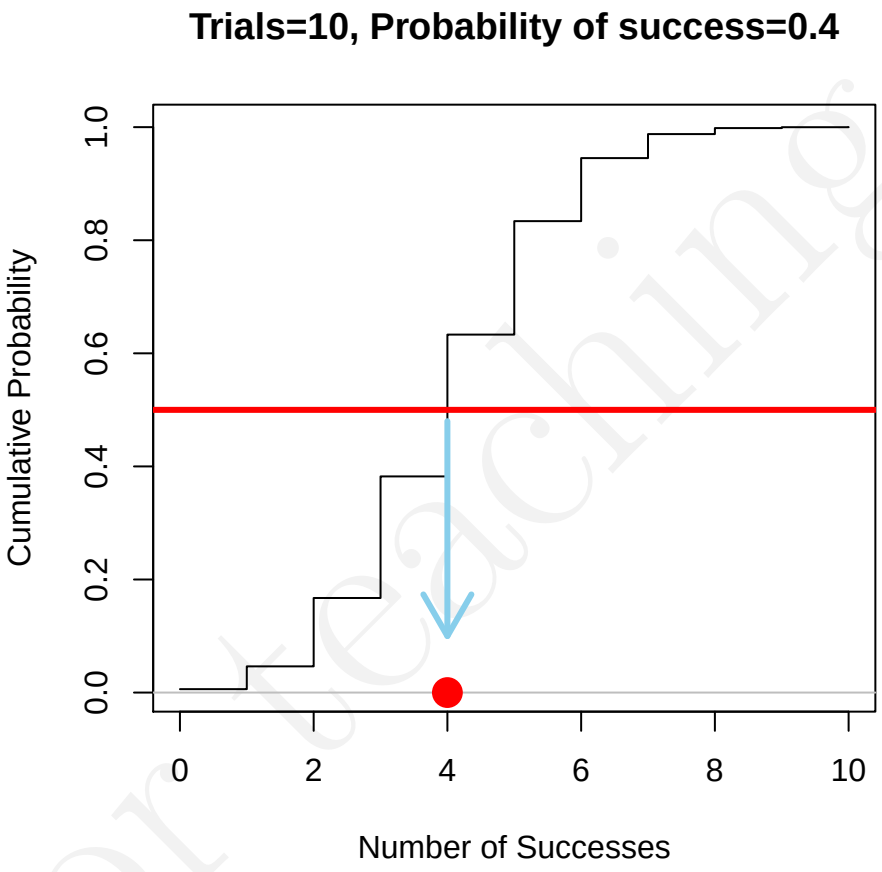
$$\begin{aligned}(n+1) \times \pi &= (10+1) \times 0.4 \\ &= 11 \times 0.4 \\ &= 4.4\end{aligned}$$

The integer between 3.4 and 4.4 is 4. It is the most probable value. In this example it also coincides with the mathematical expectancy $E(X) = n \times \pi = 10 \times 0.4 = 4$, given that the mathematical expectancy is a value that can actually be obtained. However, this is not necessarily the case for all n and π.



5. Median of this binomial distribution: Md
   It is the smallest possible value (i.e., integer) for the probability that it or a smaller value occurs is at least 0.5. That is, $Md(X) = min(k \in X | F(k) \geq 0.5)$.

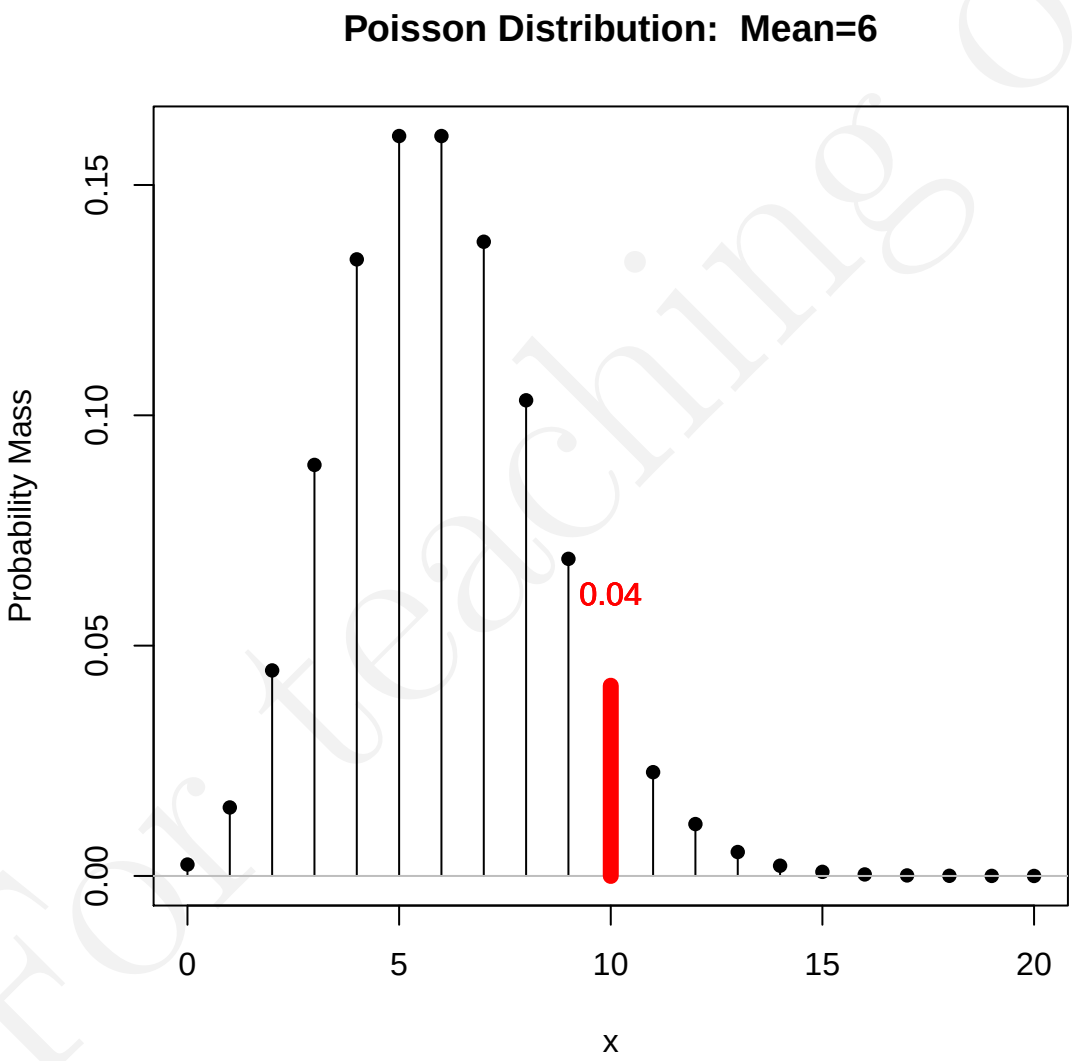
```
n <- 10
for (i in 0:n)
  if ((pbinom(i, size=10, prob=0.4)) >= 0.5) {md <- i; break}
paste("Median is ", md)

## [1] "Median is  4"
```



3.2 Poisson model: Example 1

This model is useful for modelling counts (averages) per time interval (e.g., hour, day, year) or per space interval (e.g., meter, kilometer), assuming that the probability that the event of interest takes place is constant throughout the whole interval and that whatever happens in a given interval does not affect (and is not affected) by whatever happens in the other intervals.

Imagine that in a village of 6000 individuals, the statistical records for the last 20 years suggest that there are, on average 6 cases of burnout detected each year (some years more, some years less). On January 1st, it was determined that there were 10 cases of burnout during the last year. What is the probability of that many cases if on average 6 are expected? (Note that there is no reason to suspect burnout is more likely in any specific period of the year, as could be the case for depression and its relation to Christmas holidays, for which the one of the assumptions of the Poisson model is questionable).



What is the probability of 10 or more cases if on average 6 are expected? (mind that we want the value of 10 to be included in the interval)

```
ppois(9, lambda = 6, lower.tail=FALSE)
```

```
## [1] 0.08392402
```

How many such years, with 10 or more cases, are expected to happen in a period of 25 years only due to random fluctuations, if the average of 6 per year is still correct?

```
25*ppois(9, lambda = 6, lower.tail=FALSE)
```

```
## [1] 2.0981
```

In this case, we used the Poisson model to approximate the Binomial distribution when there is a large number of trials ($n = 6000$) and the probability of success (here, burnout) is small: $\pi = 6/6000 = 0.001$. Using the Binomial distribution, we should obtain approximately the same results as before.

Mass probability function: $Prob(X = 10)$

```
dbinom(10, size=6000, prob=0.001)
```

```
## [1] 0.04128241
```

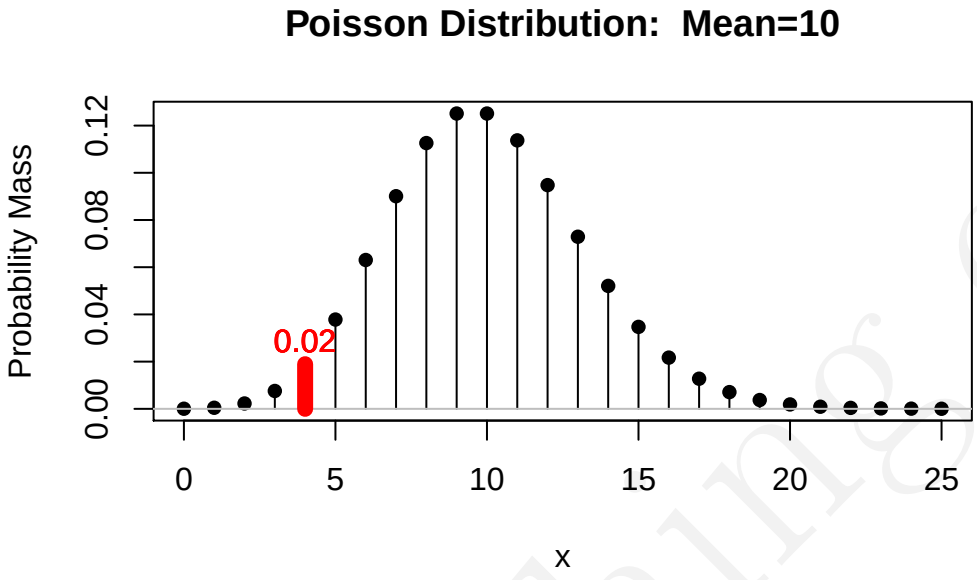
Survival function: $Prob((X - 1) > 10)$

```
pbinom(9, size=6000, prob=0.001, lower.tail=FALSE)
```

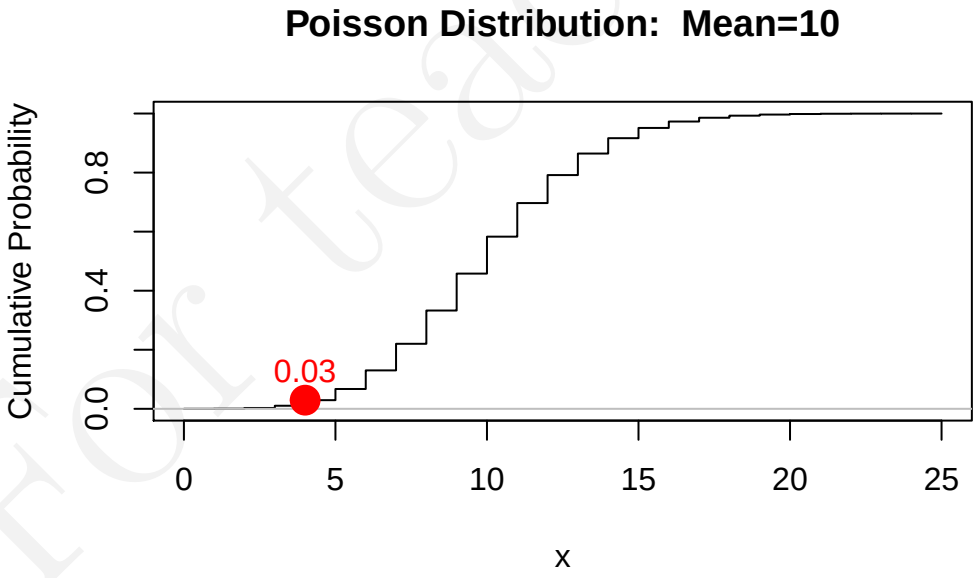
```
## [1] 0.08382071
```

3.3 Poisson model: Example 2

Imagine that in a given work place a worker has to serve 10 clients per hour, on average, and that it is not necessary to serve exactly 10 clients each specific hour, as long as the average criterion is met. The boss of this worker decides to observe the worker on a randomly chosen hour and in this specific hour only 4 clients are served. What is the probability of this happening in case the worker actually serves 10 clients on average?



What is the probability of serving 4 or less clients?

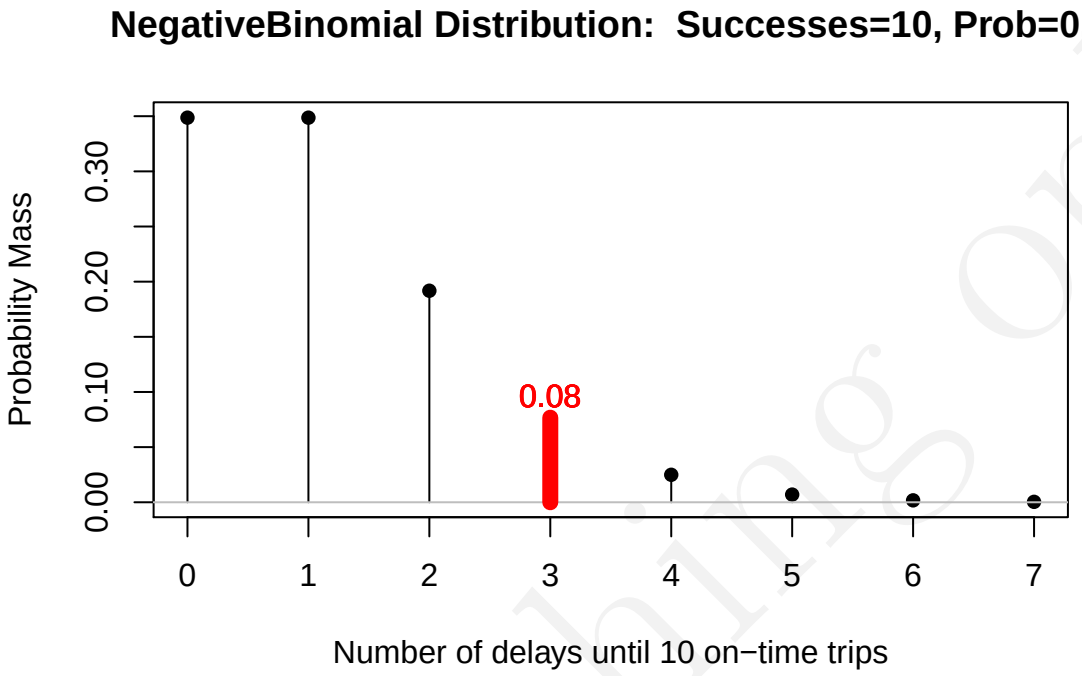


How many such hours, with 4 or less clients served, are expected to happen in a week with $5 \times 8 = 40$ hours, if the boss has to continue believing that the average of 10 clients per hour is actually met?

```
40*ppois(4, lambda=10, lower.tail=TRUE)

## [1] 1.170108
```

3.4 Negative binomial model

This model is useful when the probability of success is constant in each trial and we are looking for the probability of a number of repetitions before a specific amount of successes takes place. If the train is on time 90% of the trips (i.e., $\pi = 0.9$), what is the probability of 3 delays ($k = 3$), before reaching 10 journeys on time ($r = 10$)?



What is the probability of 3 or 4 or 5 delays, before reaching 10 journeys on time ($r = 10$)?
Adding mass probability function values:

```
dnbinom(3:5,  size = 10 ,  prob = 0.9)

## [1] 0.076709257 0.024930508 0.006980542

sum(dnbinom(3:5,  size = 10 ,  prob = 0.9))

## [1] 0.1086203
```

Subtracting distribution function values (mind the fact that we are interested in the value of 3 and the interval should include it):

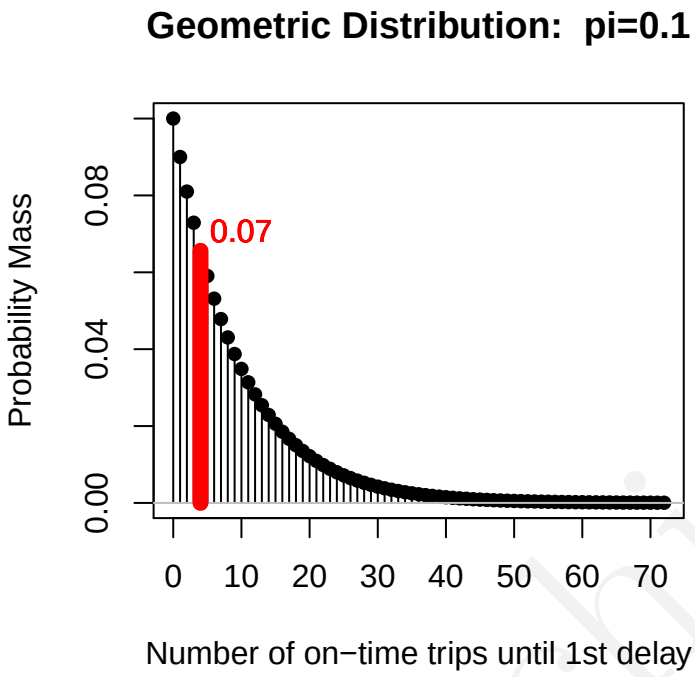
```
pnbinom(5,  size = 10 ,  prob = 0.9, lower.tail=TRUE) -
  pnbinom(2,  size = 10 ,  prob = 0.9, lower.tail=TRUE)

## [1] 0.1086203
```

Subtracting survival function values (mind the fact that we are interested in the value of 3 and the interval should include it):

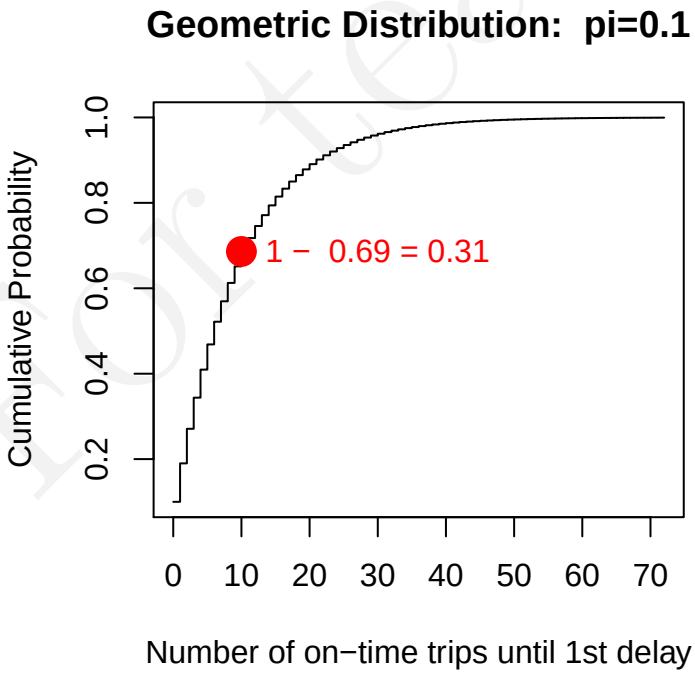
```
pnbinom(2,  size = 10 ,  prob = 0.9, lower.tail=FALSE) -
  pnbinom(5,  size = 10 ,  prob = 0.9, lower.tail=FALSE)

## [1] 0.1086203
```

3.5 Geometric model

This model is useful when the probability of success is constant in each trial and we are looking for the probability of a number of repetitions before the first success takes place. If the train has delays in 10% of the journeys (delay is a success here, $\pi = 0.1$), what is the probability of the first delay being already at journey 5 (i.e., after 4 repetitions)?



What is the probability of 10 or more trains without delay before the first delay takes place?

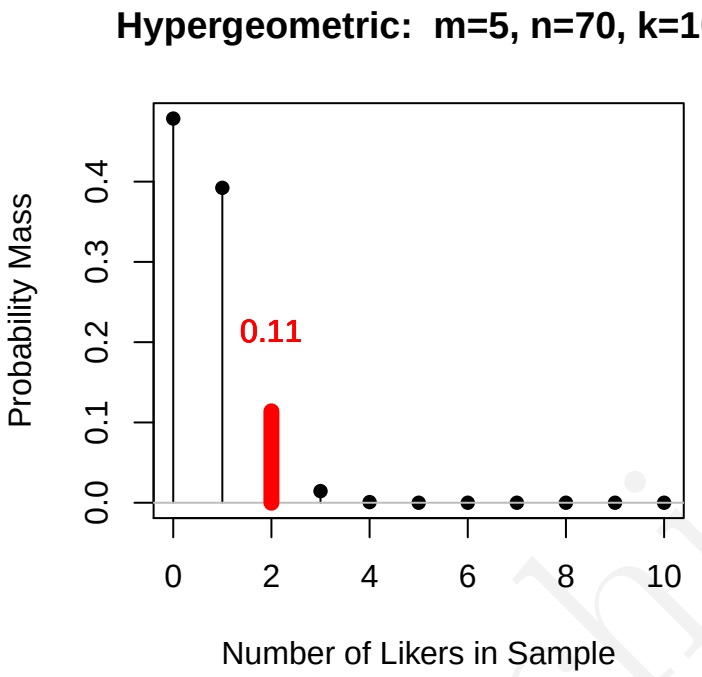



```
pgeom(c(10), prob=0.1, lower.tail=FALSE)
```

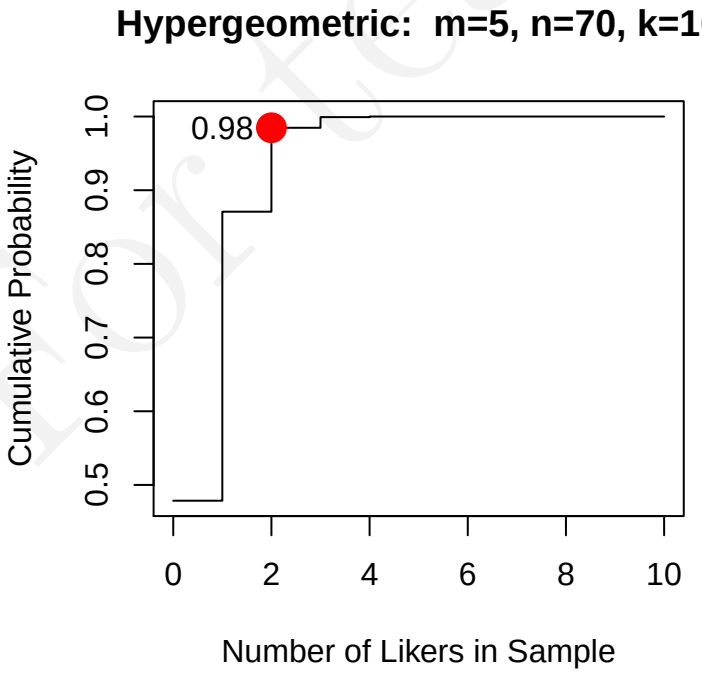
```
## [1] 0.3138106
```

3.6 Hypergeometric model

This model is useful when the probability of success is not constant in each extraction, given that sampling without replacement is performed. Imagine that there are 75 ($m + n = 75$) students in a class and 5 of them are fond of the Research techniques course (i.e., the "likers": $m = 5$). If 10 students are selected at random from this population ($k = 10$), what is the probability of two of them being "likers"?



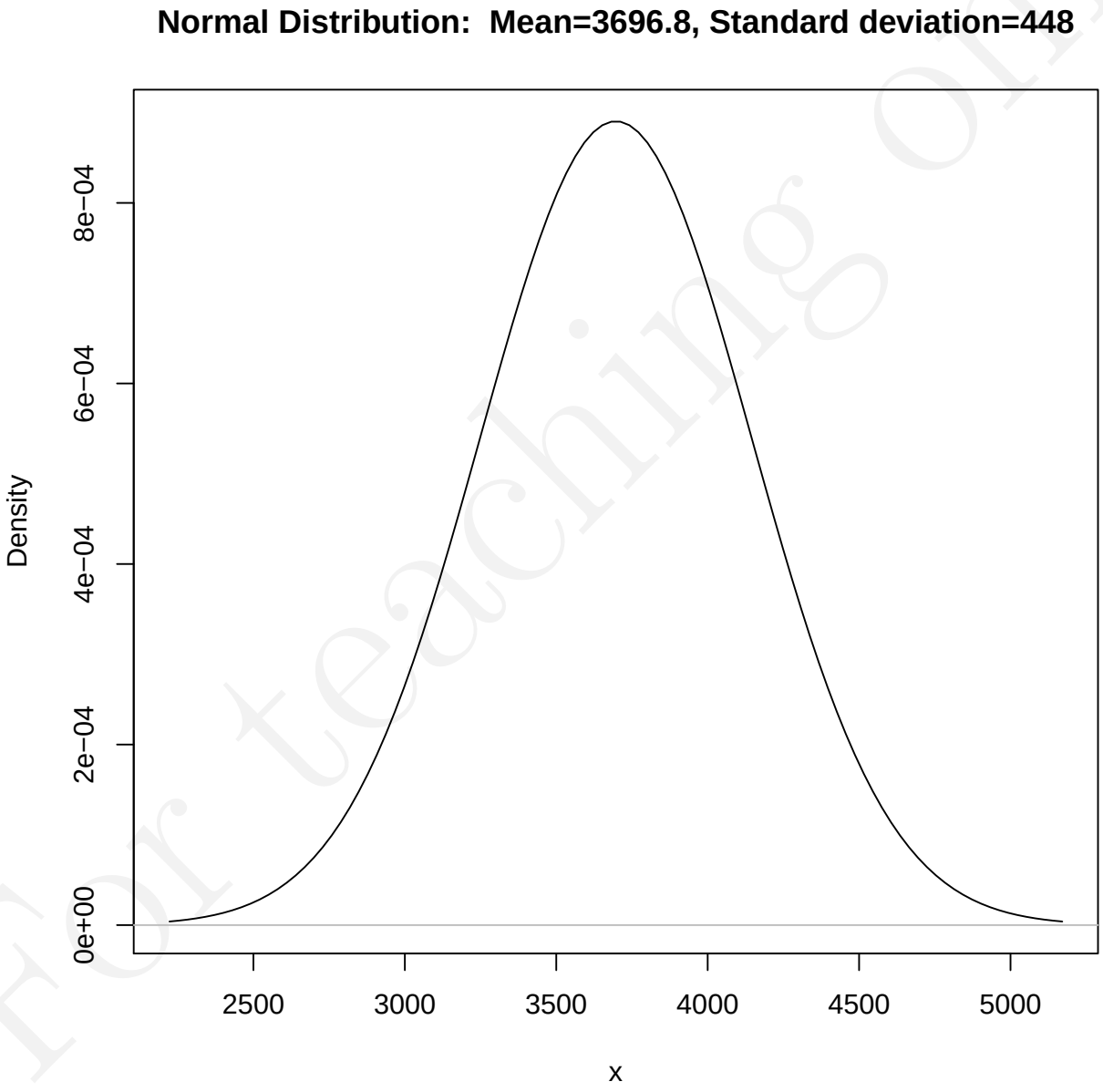
If 10 students are selected at random from this population ($k = 10$), what is the probability of two or fewer than two of them being "likers"?



4 Continuous probability models: Normal distribution

4.1 Example 1: Birthweight

According to the evidence available (Janssen et al., 2007; Wilcox, 2001) the weight of girls born after 41 weeks of gestation can be modelled via a normal distribution with a location parameter $\mu = 3696.8$ and a scale parameter $\sigma = 448$, both expressed in grams, that is $N(\mu = 3696.8, \sigma = 448)$ or $N(3696.8, 448)$. The density function of the model looks as shown below:



4.1.1 Density

The density for the value of interest (4500g) can be found in R and R-Commander using the following code:

```
dnorm(c(4500), mean=3696.8, sd=448)

## [1] 0.0001785039
```

Densities inform concentration of values; they do not quantify probabilities. However, a greater density indicates a greater probability. This can be illustrated looking at the density for a value close to the mean, e.g., 3700:

```
dnorm(c(3700), mean=3696.8, sd=448)

## [1] 0.0008904734
```

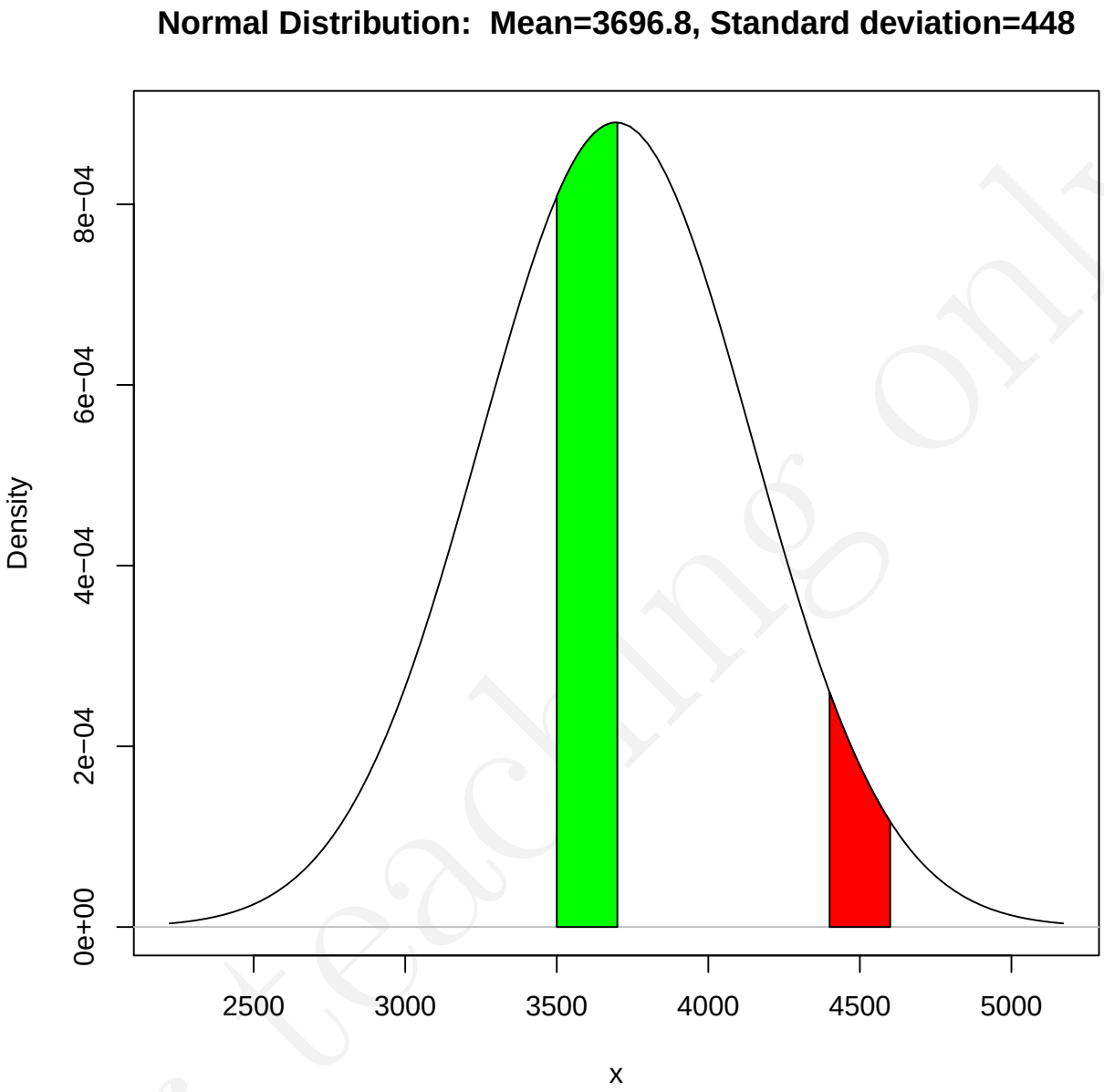
4.1.2 Probability

In order to illustrate that density is related to probability, we will show how a weight of approximately 4500g (between 4400g and 4600g) is less likely than a weight of approximately 3700g (between 3600g and 3800g). Note that in both cases, the interval of values has the same length: 200g. We will use the subtraction between distribution functions (i.e., the lower tail) to obtain the probabilities.

```
print("Probability of a weight between 4400 and 4600 grams = ")

## [1] "Probability of a weight between 4400 and 4600 grams = "

pnorm(c(4600), mean=3696.8, sd=448, lower.tail=TRUE) -
  pnorm(c(4400), mean=3696.8, sd=448, lower.tail=TRUE)

## [1] 0.03635286
```

```
print("Probability of a weight between 3500 and 3700 grams = ")

## [1] "Probability of a weight between 3500 and 3700 grams = "

pnorm(c(3700), mean=3696.8, sd=448, lower.tail=TRUE) -
  pnorm(c(3500), mean=3696.8, sd=448, lower.tail=TRUE)

## [1] 0.1726223
```

The values closer to the mean have greater density and the intervals around this values have greater probability. This is straightforward, given that the area under the curve (which represents probability) is larger near μ.



Another important fact about continuous random variables is that the probability of any specific value is practically zero. This is true even for the values with greater density. For instance the probability of an interval that includes the mean (between 3695g and 3700g) is as small as:

```
pnorm(c(3700), mean=3696.8, sd=448, lower.tail=TRUE) -
  pnorm(c(3696), mean=3696.8, sd=448, lower.tail=TRUE)

## [1] 0.00356196
```

In this example, we are still talking about intervals. In case we were interested in a specific value, we could define it as some value between 3696 and 3697. The probability is practically zero and is usually treated as if it were actually zero.

```
pnorm(c(3697), mean=3696.8, sd=448, lower.tail=TRUE) -
  pnorm(c(3696), mean=3696.8, sd=448, lower.tail=TRUE)

## [1] 0.0008904958
```

For all other values, farther away from μ the probabilities of such small intervals would be even smaller.

4.1.3 Standardizing

In the past, statistical tables were the source for probabilities and it is not feasible to have as many different tables, as there are possible values for the location and scale parameters of a normal distribution. This is why standardizing is used: so that all normal variables whose distribution has its own μ and σ values can be converted to the same normal distribution: the unitary one, denoted by Z or $N(0, 1)$. This notation makes clear that the new normal variable has a mean of zero (a change in location) and a standard deviation of one (a change in scale). Standardizing for the running example, with the mean and standard deviation as speficied above and for $x = 4500$ is performed as shown below:

$$\begin{aligned} Z &= \frac{x_i - \mu}{\sigma} \\ &= \frac{4500 - 3696.8}{448} \\ &= \frac{4500 - 3696.8}{448} \\ &= \frac{803.2}{448} \\ &= 1.792857 \end{aligned}$$

It can be shown that the probability of having a baby girl, born after 41 weeks of gestation, as large as or larger than 4500g is the same, within rounding error, regardless of whether we use the original variable or the standardized one:

```
pnorm(c(4500), mean=3696.8, sd=448, lower.tail=FALSE)

## [1] 0.03649788
```

```
pnorm(c(1.792857), lower.tail=FALSE)

## [1] 0.0364979
```

In the previous piece of R code, it should be noted that we did not specify the values for the location and the scale parameters. This is so, given that by default, R uses $\mu = 0$ and $\sigma = 1$.

Finally, in the next table, we illustrate how statistical tables were used to obtain probabilities. We first look at the row containing 1.7 as the first two digits of the Z value. Second, we look for the column containing the 0.09 value, as the nonzero digit is the third digit of our Z value. Third, the probability that the statistical table provides is the distribution function $F(1.79) = Prob(Z \leq 1.79)$. However, we are looking for the survival function $S(1.79) = Prob(Z > 1.79)$.

$$\begin{aligned} S(1.79) \equiv Prob(Z > 1.79) &= 1 - F(1.79) \\ &= 1 - 0.9633 \\ &= 0.0367 \end{aligned}$$

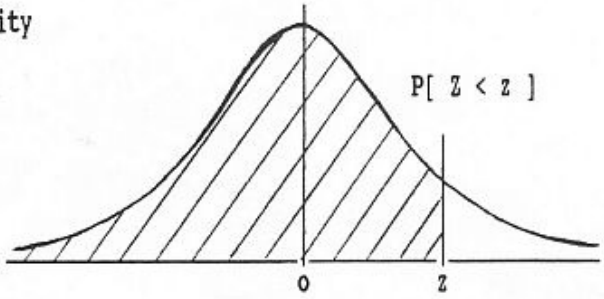
The difference with respect to the previously obtained probabilities is due to the fact that with R we were more precise, looking for $S(1.792857)$ rather than for $S(1.79)$.

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z
i.e.

$$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

Is it probable? Is *what* probable? What is *probable*?

cases, it is not only interesting to know whether such a large weight (or larger) is likely, but whether such an extreme weight is likely. "Extreme" means "so far away from the mean". Therefore, we would be interested in the probability of a weight as large as or larger than 4500g and the probability of a weight as small as or smaller than a weight, which is at the same distance from the mean (3696.8) as 4500 is, but below it. In order to find this equally distant value from the mean, we perform the following calculation: first, the distance between 4500 and the mean is $4500 - 3696.8 = 803.2$; second, the value that is 803.2g below the mean is $33696.8 - 803.2 = 2893.6$. In summary, when we talk about "extreme" results, we do not distinguish "small" from "large"; we deal with *bilateral* (*two-tailed*) probabilities. Otherwise, in case we wanted to know the probability of such a large or larger value, we would be dealing with *unilateral* (*one-tailed*) probabilities. We focus on the two-tailed case in this page.

```
pnorm(c(2893.6), mean=3696.8, sd=448, lower.tail=TRUE)

## [1] 0.03649788
```

This calculus was not necessary, given that the Normal distribution is symmetric and $Prob\big(X < (3696.8 - 803.2)\big) = Prob\big(X > (3696.8 + 803.2)\big)$, which is true for any equidistance from the mean, not only for 803.2. Therefore, the probability of "as extreme" weight is double the probability of an "as large" weight and double the probability of an "as small" weight.

But is that *probable*? It is possible! Beyond this obvious answer, it is difficult to give a more precise and sound response. In case-control designs, used in Neuropsychology, it is common to use probability models to assess how unfrequent or rare a result of an individual would be in the normative population (i.e., the one that, in the Neuropsychological context, presents no cognitive deficit). We could use the normal probability model here, because we have evidence that the population birthweight could be normally distributed.

```
pnorm(c(2893.6), mean=3696.8, sd=448, lower.tail=TRUE) +
  pnorm(c(4500), mean=3696.8, sd=448, lower.tail=FALSE)

## [1] 0.07299577
```

The result presented above suggests that a value as extreme as 4500g (i.e., a value as far away from the mean as 4500g) is expected to happen in approximately 7 out of 100 normative cases. It is still the researcher's decision to state whether this is very unfrequent (very rare) or not. In case we do not want to assume that the population is normal, the t distribution can be used, on the basis of the modification of the t statistic described by Crawford and colleagues (Crawford & Garthwaite, 2012; Crawford, Garthwaite, & Porter, 2010; Crawford & Howell, 1998; see also the following website `http://homepages.abdn.ac.uk/j.crawford/pages/dept/SingleCaseMethodsComputerPrograms.HTM`).

Another option is to emulate what is being done when making statistical decisions (i.e., to reject the null hypothesis or not), which are explained in the *Estadística* course. Specifically, when making statistical decisions it is common to treat probability values (p values) equal to or lower than 0.05 as an indicator that the null hypothesis can be rejected, as the probability of committing a mistake when rejecting it is low (0.05 or less). Thus, if our null hypothesis was that the girl with a weight of that is 803.2g away from the mean is part of the normative population with mean equal to 3696.8g and standard deviation equal to 448g, we would not reject the null hypothesis, as the probability of obtaining such a large difference from the mean or larger difference is approximately 0.07, that is, more than the risk we are willing to assume. In substantive terms, it could be stated that a baby girl born 4500g is not *that* different from the normative (typical) population of baby girls.

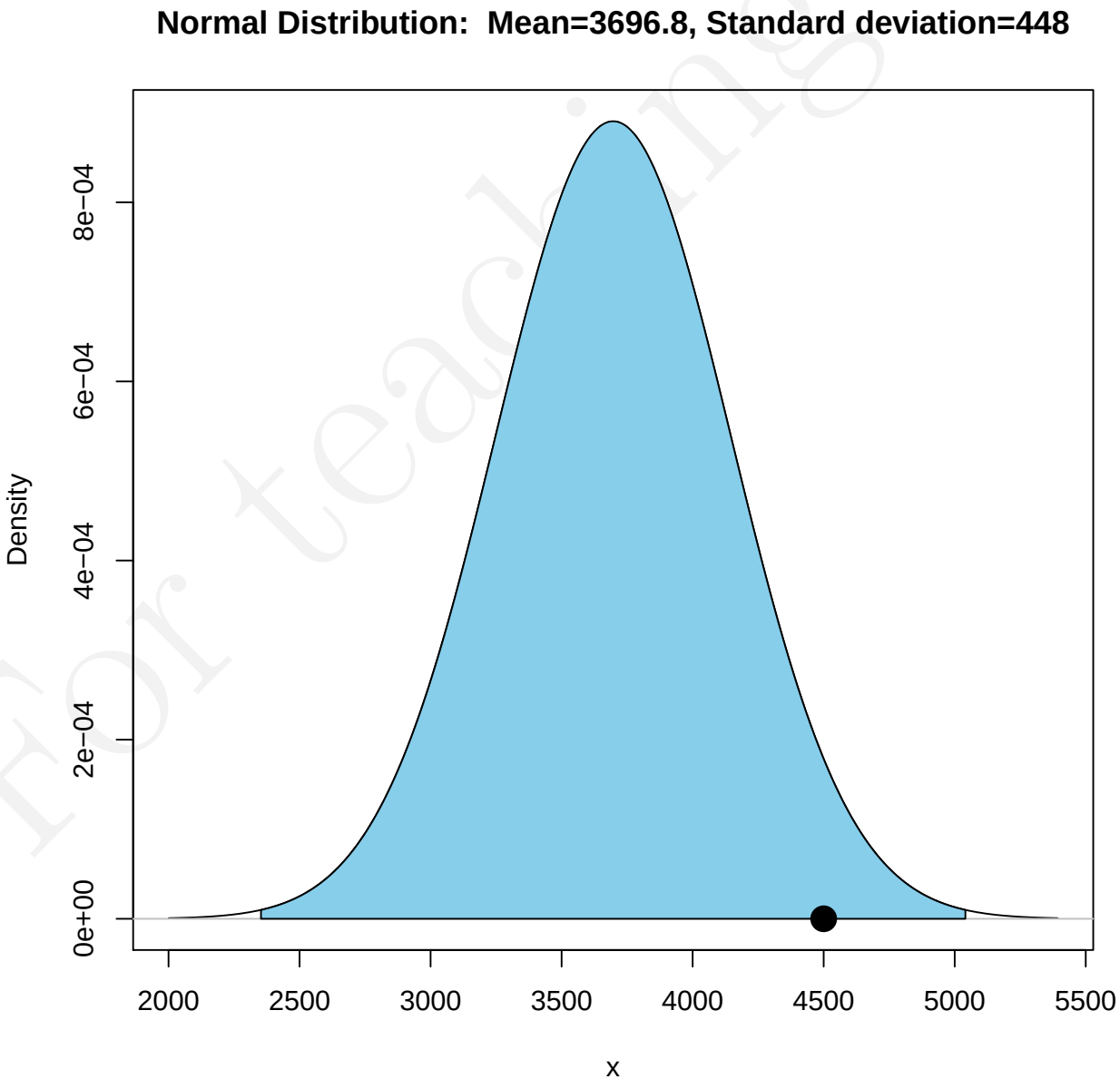
However, it should be noted that, in this last example, we are not using a one-sample test (i.e., we are not comparing a sample mean with a populational parameter). Therefore, we only "borrowed" the 0.05 convention for illustrative purposes. Actually, it depends on the topic, which probability should be considered small and which large risk for committing a Type I error.

Another option for assessing the degree to which a specific values is within the expected limits is the *statistical process control* rule according to which any value outside the interval $\mu \pm 3 \times \sigma$ is considered unlikely (see Pfadt & Wheeler, 1995, for *statistical process control* rules regading the number of values outside of the $\mu \pm \sigma$ or the $\mu \pm 2 \times \sigma$ intervals). Actually, according to the property of the normal distribution, within the interval $\mu \pm 3 \times \sigma$ there is the 99.7% of the area or 99.7% of the values are expected to be included; the values that are not included are the 0.15% most extreme small values and the 0.15% most extreme high values.

For the current example:

$$\begin{aligned} \mu \pm 3 \times \sigma &= 3696.8 \pm 3 \times 448 \\ &= 3696.8 \pm 1344 \\ &= [2352.8; 5040.8] \end{aligned}$$

Therefore, it can be seen that 4500 is within these limits and, according to this rule, is not considered as a value that is out of control.



4.2 Example 2: Intelligence quotient

Intelligence quotient: $\mu = 100$ and $\sigma = 15$

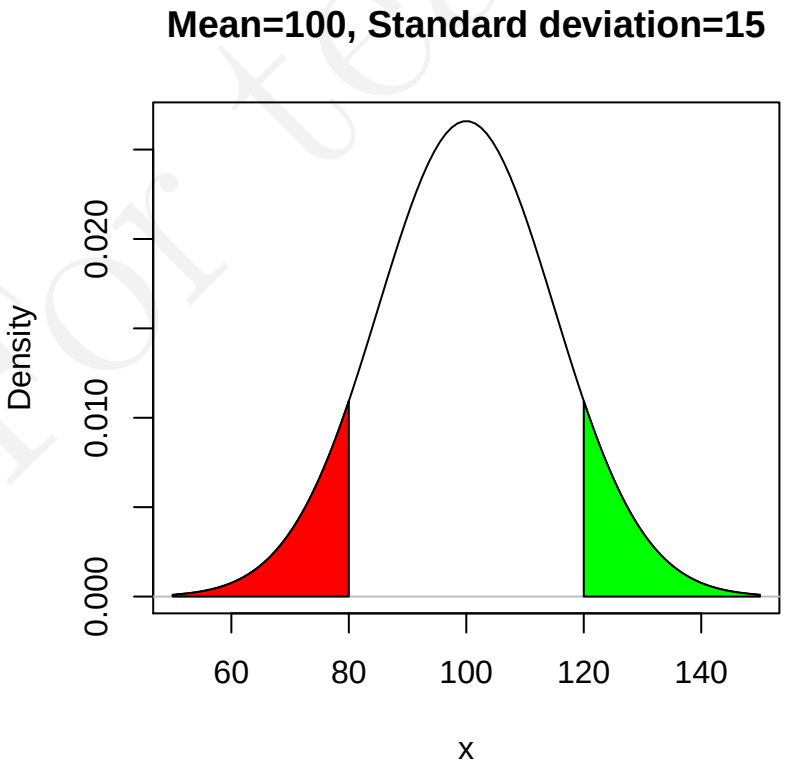
```
# Parameters of the model
mu <- 100
sigma <- 15
```

1. Find the probability that a randomly selected individual from the population has an IQ score lower than 80 or greater than 120.
   $Prob(X \leq 80) = F(80)$ or $Prob(X > 120) = S(120)$

```
# Mass probability function
pnorm(c(80), mean=100, sd=15, lower.tail=TRUE)

## [1] 0.09121122

pnorm(c(120), mean=100, sd=15, lower.tail=FALSE)

## [1] 0.09121122
```

   Both probabilities are equal given that they represent equidistant values from the mean (20 points below or above). Given that either result meets the condition, the answer is

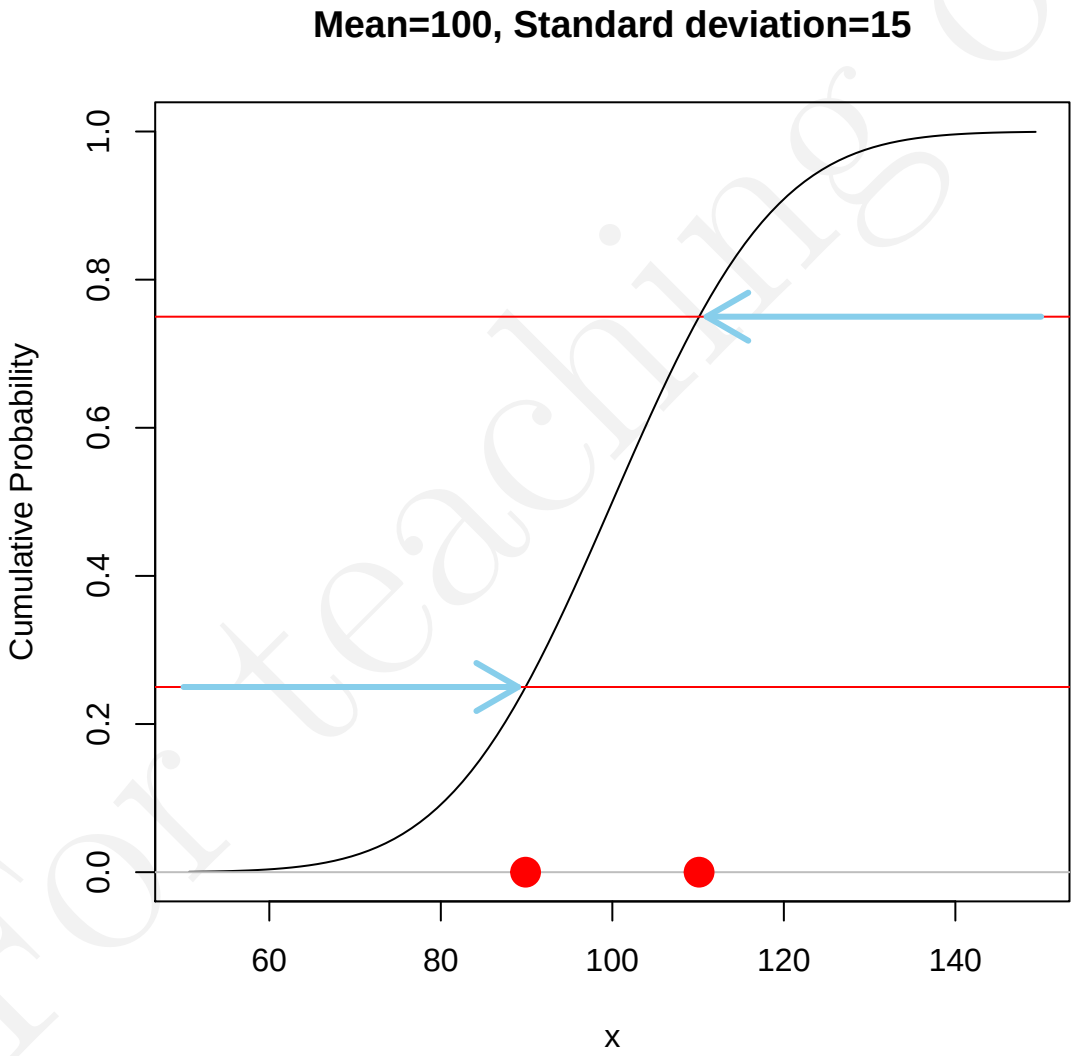
```
# Probability of two events that do not intersect
pnorm(c(80), mean=100, sd=15, lower.tail=TRUE) +
  pnorm(c(120), mean=100, sd=15, lower.tail=FALSE)

## [1] 0.1824224
```



2. Between which IQ scores are the central 50% of the individuals located (i.e., what is the interval of most representative values that includes 50% of the population).

In this case we have the cumulative probabilities (0.25 and 0.75, between which the central 50% of the individualsare located) and we are looking for the values. Thus, we are working with the inverse of the cumulative distribution function.

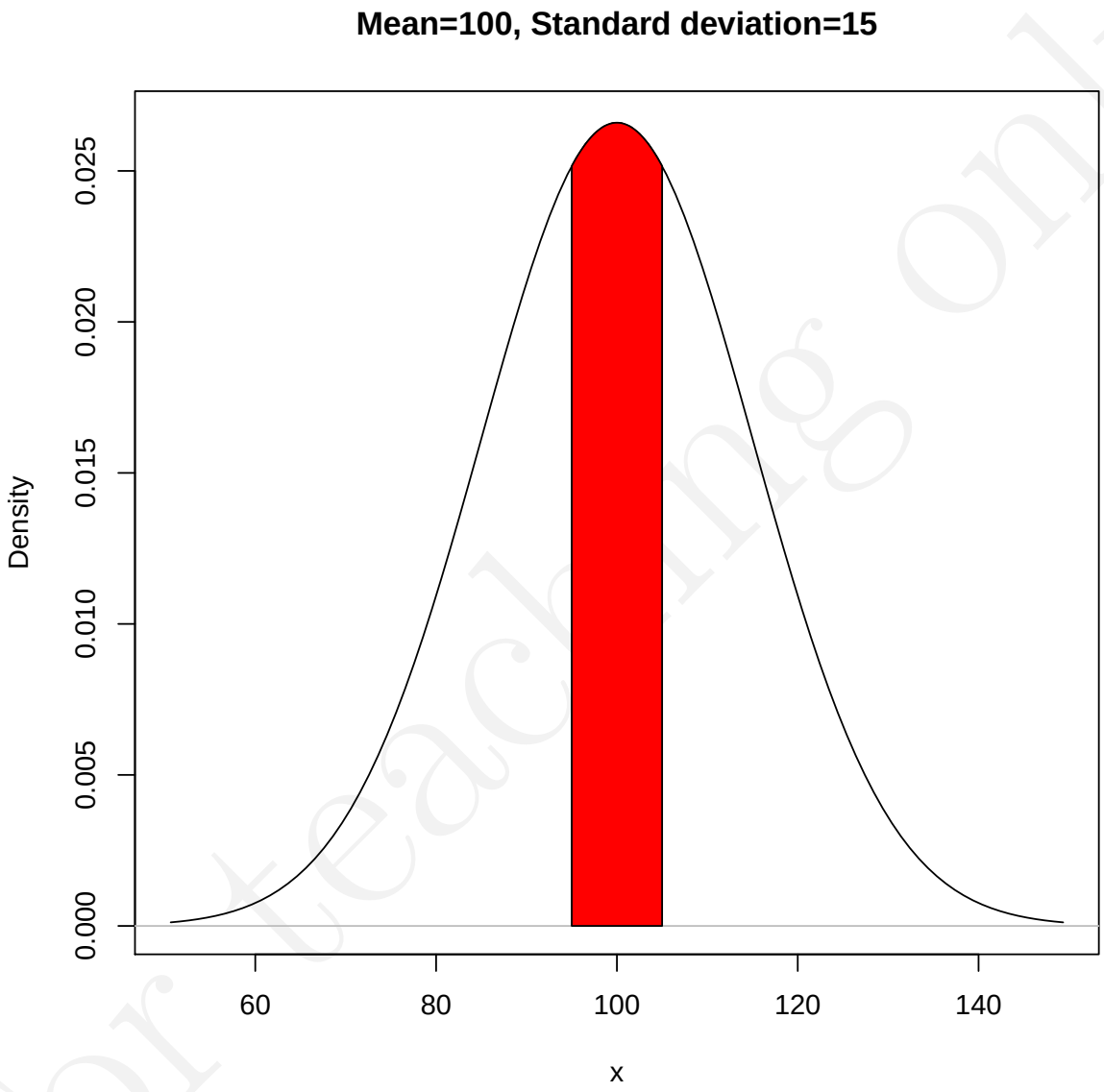
```
# Quantiles
qnorm(c(0.25,0.75), mean=100, sd=15, lower.tail=TRUE)

## [1]  89.88265 110.11735
```



Without computing the quantiles and due to the property of the Normal distribution, an approximate answer could have been given. Specifically, we know that the central 68.27% of the individuals has IQ values between $\mu \pm \sigma = 100 \pm 15 = [85, 115]$. That is, the lower limit of this interval is value one for which the distribution function is equal to $(1 - 0.6827)/2 = 0.1587$, so $Q_{0.1587} = 85$. Given that we are looking for $Q_{0.25}$ this value should be greater than 85. Analogously, the upper limit of this interval is value one for which the distribution function is equal to $0.6827 + (1 - 0.6827)/2 = 0.6827 + 0.8414$, so $Q_{0.8414} = 115$. Given that we are looking for $Q_{0.75}$, this value should be smaller than 115.

3. A sample of 250 individuals ($n = 250$) is extracted at random from the population $N(\mu = 100, \sigma = 15)$. How many of them are expected to have IQ values between 95 and 105?

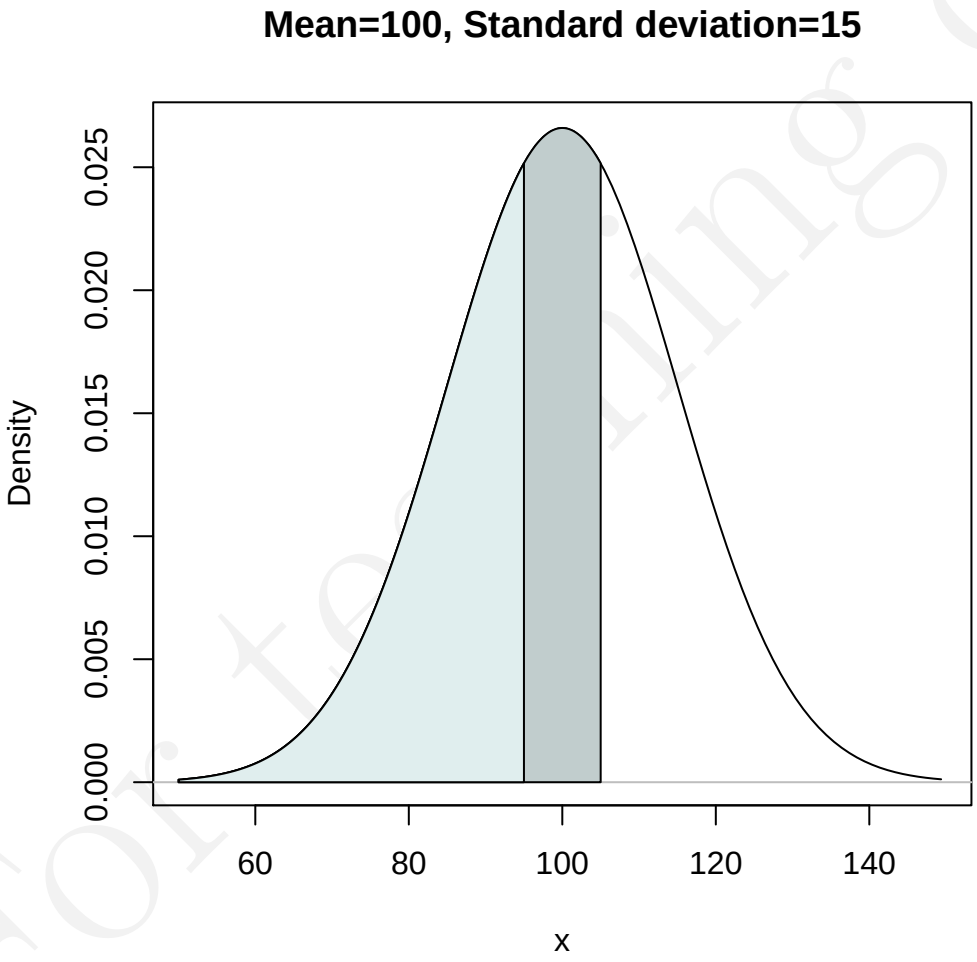
   First we need to find the area under the curve with limits 95 and 105.



This can be done in two ways. One option is to use the distribution function (i.e., lower tail of the density function).

```
# Subtracting the distribution function of the higher value
#  from the distribution function of the lower value
pnorm(c(105), mean=100, sd=15, lower.tail=TRUE) -
  pnorm(c(95), mean=100, sd=15, lower.tail=TRUE)

## [1] 0.2611173
```

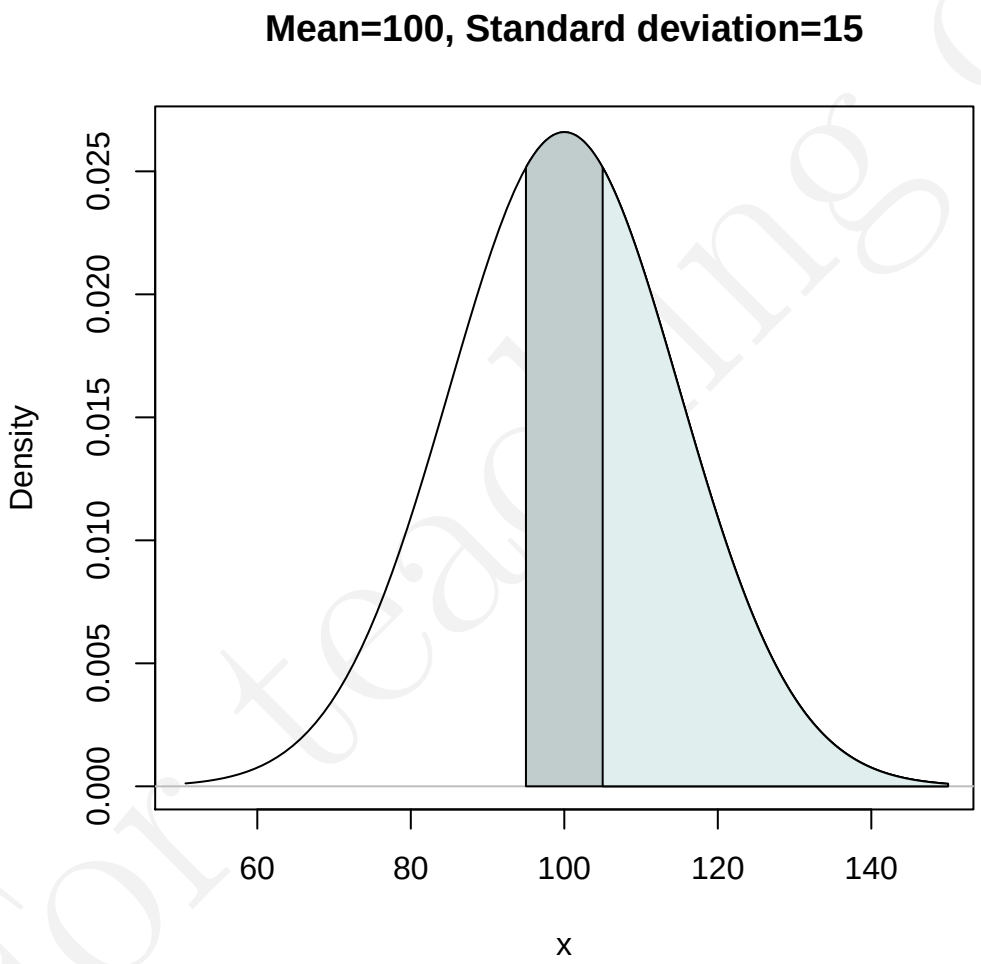
The white area is not part of the calculation. The lighter color is the area that is subtracted. The darker color is the area that remains after the subtraction (i.e., the area of interest) - it is the intersection between the two distribution functions.



Another option is to use the survival function (i.e. upper tail of the density function).

```
# Subtracting the survival function of the lower value
#  from the survival function of the lower value
pnorm(c(95), mean=100, sd=15, lower.tail=FALSE) -
  pnorm(c(105), mean=100, sd=15, lower.tail=FALSE)

## [1] 0.2611173
```

The white area is not part of the calculation. The lighter color is the area that is subtracted. The darker color is the area that remains after the subtraction (i.e., the area of itnerest) - it is the intersection between the two survival functions.



Second, we need to multiply this probability by the sample size in order to know how many individuals are expected to have IQ values between 95 and 105: $Expected[n(95 < X < 105)] = Prob(95 < X < 105) \times n = 0.2611173 \times 250 \simeq 65.28$. Therefore, approximately 65 people are expected to have an IQ in the range in a random sample of 250.

Actually, we could use R as a calculator. The signs for summation and subtraction are the classical ones (+ and −, repsectively), as is also the case for division (/). Multiplication is done using the asterisk sign (*), whereas elevating a number to a given power is achieved using the circumflex or hat sign (^). Out of this operations, only multiplication is needed here.

```
# Specific value
0.2611173*250

## [1] 65.27933

# Rounding to the nearest integer
round(0.2611173*250,0)

## [1] 65

# Rounding downwards
floor(0.2611173*250)

## [1] 65

# Rounding upwards
ceiling(0.2611173*250)

## [1] 66
```

4. We have previously mentioned that approximately 68% of the area is included in the interval $\mu \pm \sigma$. What proportion of the area (i.e., what proportion of the the individuals in the population) is included in the intervals $\mu \pm 2 \times \sigma$ and $\mu \pm 3 \times \sigma$. Note that this property is general for all values of μ and σ of a normal distribution.

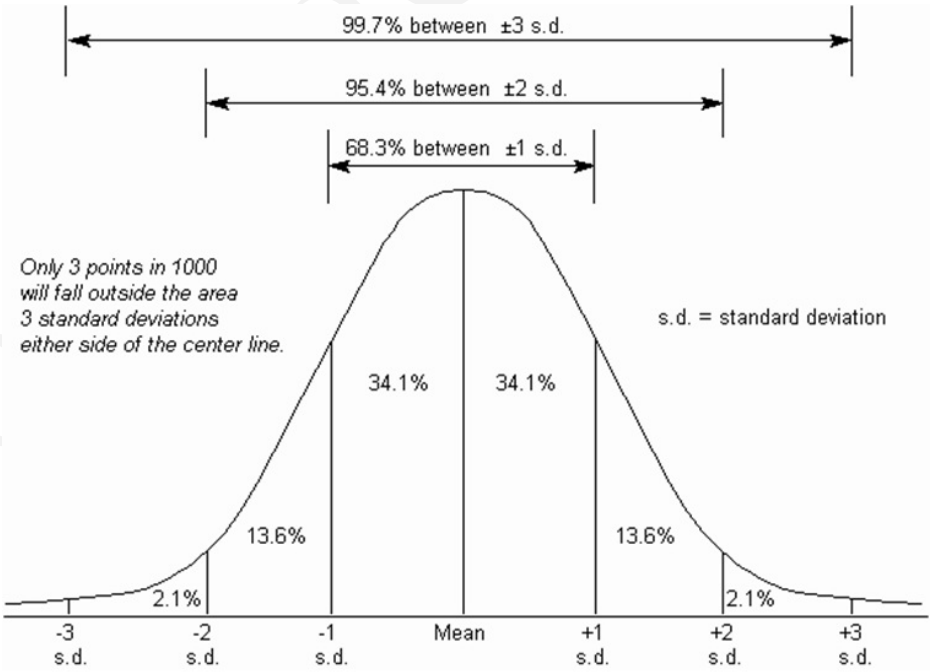
We first need to obtain the values that correspond to $\mu - 2 \times \sigma$ and $\mu + 2 \times \sigma$. The lower limit of this interval is marked by $\mu - 2 \times \sigma = 100 - 2 \times 15 = 70$, whereas the upper limit is $\mu + 2 \times \sigma = 100 + 2 \times 15 = 130$. The area within this limits can be found subtracting two distribution functions.

```
# Subtracting the distribution function of the higher value
#  from the distribution function of the lower value
pnorm(c(130), mean=100, sd=15, lower.tail=TRUE) -
  pnorm(c(70), mean=100, sd=15, lower.tail=TRUE)

## [1] 0.9544997
```

We then need to obtain the values that correspond to $\mu - 3 \times \sigma$ and $\mu + 3 \times \sigma$. The lower limit of this interval is marked by $\mu - 2 \times \sigma = 100 - 3 \times 15 = 55$, whereas the upper limit is $\mu + 2 \times \sigma = 100 + 3 \times 15 = 145$. The area within this limits can be found subtracting two distribution functions.

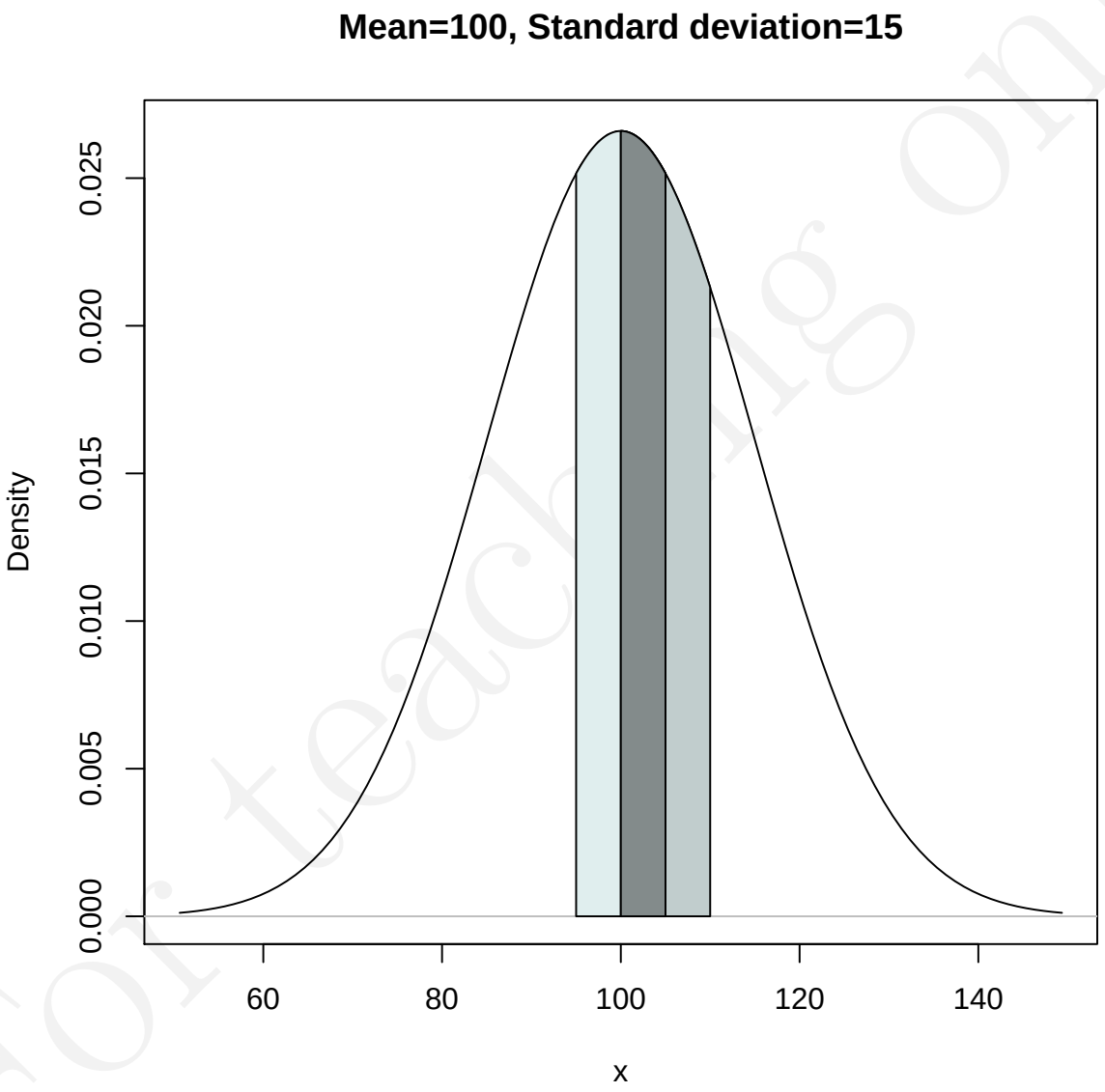
```
# Subtracting the distribution function of the higher value
#  from the distribution function of the lower value
pnorm(c(145), mean=100, sd=15, lower.tail=TRUE) -
  pnorm(c(55), mean=100, sd=15, lower.tail=TRUE)

## [1] 0.9973002
```

These proportions are all reflected in the following figure:



5. Find the probability that a randomly selected individual from the population has an IQ score in the interval [95; 105] or in the interval[100; 110].

   In this case, a value in either of the intervals would meet the condition. Therefore, we can add the probabilities of the two intervals to obtain the probability that the value is contained in either of them. However, the two intervals intersect. Specifically, they share the values in the interval [100; 105]: the darkest area in the density plot below.



There are three ways to obtain the probability. First, to obtain $Prob(95 < X < 105) + Prob(100 < X < 110) - Prob(100 < X < 105)$ via distribution functions.

```
# Subtracting the distribution function of the higher value
#  from the distribution function of the lower value
P.95.105 <- pnorm(c(105), mean=100, sd=15, lower.tail=TRUE) -
              pnorm(c(95), mean=100, sd=15, lower.tail=TRUE)
P.100.110 <- pnorm(c(110), mean=100, sd=15, lower.tail=TRUE) -
              pnorm(c(100), mean=100, sd=15, lower.tail=TRUE)
P.100.105 <- pnorm(c(105), mean=100, sd=15, lower.tail=TRUE) -
              pnorm(c(100), mean=100, sd=15, lower.tail=TRUE)
P.95.105 + P.100.110 - P.100.105

## [1] 0.3780661
```

Second, the same result can be obtained using the survival functions.

```
# Subtracting the distribution function of the higher value
#  from the distribution function of the lower value
P.95.105 <- pnorm(c(95), mean=100, sd=15, lower.tail=FALSE) -
              pnorm(c(105), mean=100, sd=15, lower.tail=FALSE)
P.100.110 <- pnorm(c(100), mean=100, sd=15, lower.tail=FALSE) -
              pnorm(c(110), mean=100, sd=15, lower.tail=FALSE)
SP.100.105 <- pnorm(c(100), mean=100, sd=15, lower.tail=FALSE) -
              pnorm(c(105), mean=100, sd=15, lower.tail=FALSE)
P.95.105 + P.100.110 - P.100.105

## [1] 0.3780661
```

Finally, the result can be obtained computing the probability of the area that is colored and not white, given it represents the union of the two events that intersect:

$$\begin{aligned} &Prob\big((95 < X < 105) \cup (100 < X < 110)\big) \\ &\quad = Prob(95 < X < 105) + Prob(100 < X < 110) - Prob\big((95 < X < 105) \cap (100 < X < 110)\big) \\ &\quad = Prob(95 < X < 105) + Prob(100 < X < 110) - Prob(100 < X < 105) \\ &\quad = Prob(95 < X < 110). \end{aligned}$$

The probability of obtaining an IQ score in that interval can be obtained via the distribution function.

```
# Subtracting the distribution function of the higher value
#  from the distribution function of the lower value
pnorm(c(110), mean=100, sd=15, lower.tail=TRUE) -
  pnorm(c(95), mean=100, sd=15, lower.tail=TRUE)

## [1] 0.3780661
```

6. Find the probability that an individual extracted at random from the population has an IQ above 120, if it is known that her IQ is above 110.

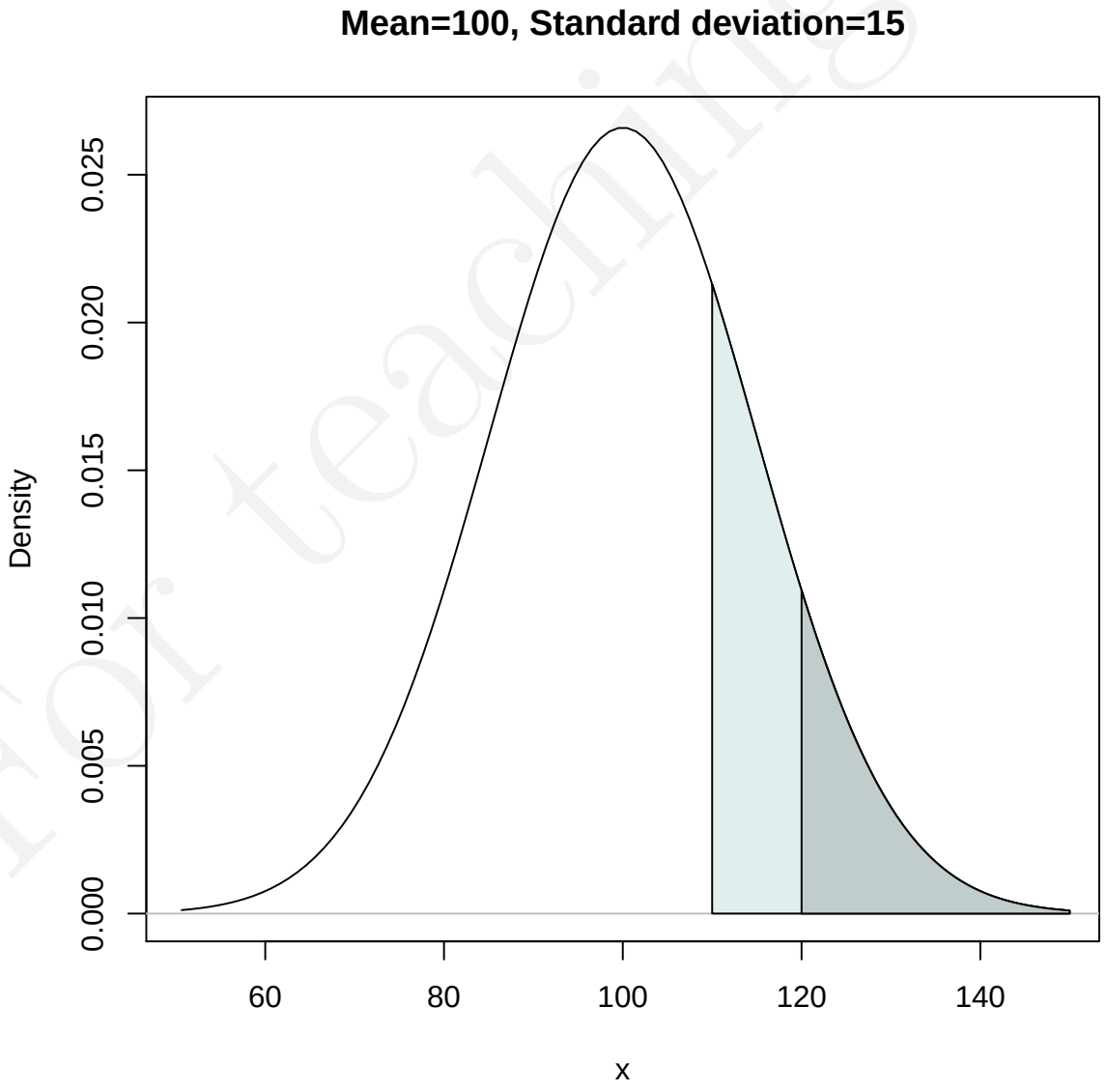
In this case we are dealing with a conditional probability: finding the probability of an event conditional on the fact that there is already some knowdlege about this event. The general formula for the conditional probability is

$$Prob(A|B) = \frac{Prob(A \cap B)}{Prob(B)}$$

In the current question, this formula can be translated into:

$$Prob(X > 120|X > 110) = \frac{Prob(X > 120 \cap X > 110)}{Prob(X > 110)}$$

Therefore, in the numerator, we need to find the probability of the intersection of the two events. It is the area they share, the one with the darker color in the density function below.



It can be seen that the intersection of $Prob(X > 120)$ and $Prob(X > 110)$ is $Prob(X > 120)$. This probability can be found using the survival function.

```
# Survival function
pnorm(c(120), mean=100, sd=15, lower.tail=FALSE)

## [1] 0.09121122
```

In the denominator, we need to find $Prob(X > 110)$, which can also be obtained using teh survival function.

```
# Survival function
pnorm(c(110), mean=100, sd=15, lower.tail=FALSE)

## [1] 0.2524925
```

With the values obtained, we apply the formula for the conditional probability:

$$\begin{aligned} Prob(X > 120|X > 110) &= \frac{Prob(X > 120 \cap X > 110)}{Prob(X > 110)} \\ &= \frac{Prob(X > 120)}{Prob(X > 110)} \\ &\simeq \frac{0.09}{0.25} \\ &= 0.36 \end{aligned}$$

7. Two individuals are extracted at random from the population. Find the probability that one of them has IQ below 70 and the other one has an IQ greater than 130.

In this question we are dealing with two independent events, given that the IQ scores of the both individuals sampled at random do not have any dependence among themselves. Thus, we will have to multiply the $Prob(X < 70)$ by $Prob(X > 130)$. Using the property of the normal distribution, we know that approximately 95.4% of the individuals have scores in the interval [70; 130], given that $\mu \pm 2 \times \sigma = 100 \pm 2 \times 15 = [70; 130]$. Therefore we know that $Prob(X < 70) = Prob(X > 130) \simeq \frac{1-0.954}{2} = \frac{0.046}{2} = 0.023$. In order to obtain the exact probabilities, we will use the distribution function of 70 and the survival function of 130.

```
# Distribution funtion of 70
pnorm(c(70), mean=100, sd=15, lower.tail=TRUE)

## [1] 0.02275013

# Survival function of 130
pnorm(c(130), mean=100, sd=15, lower.tail=FALSE)

## [1] 0.02275013
```

The probability that we are looking for is apparently $0.02275013 \times 0.02275013 = 0.0005175684$. However, either of the two individuals can have the below 70 score or the above 130 score. Therefore, there are two different ways in which the condition can be met. Thus, the probability we are looking for is $0.0005175684 \times 2 = 0.001035137$. That is, such a result is expected to happen in approximately 1 of every 1000 random extractions of two individuals.

5 Categorical data

The data presented below correspond to the answers of smokers (most of the participants) regarding the reason which could make them quit smoking:

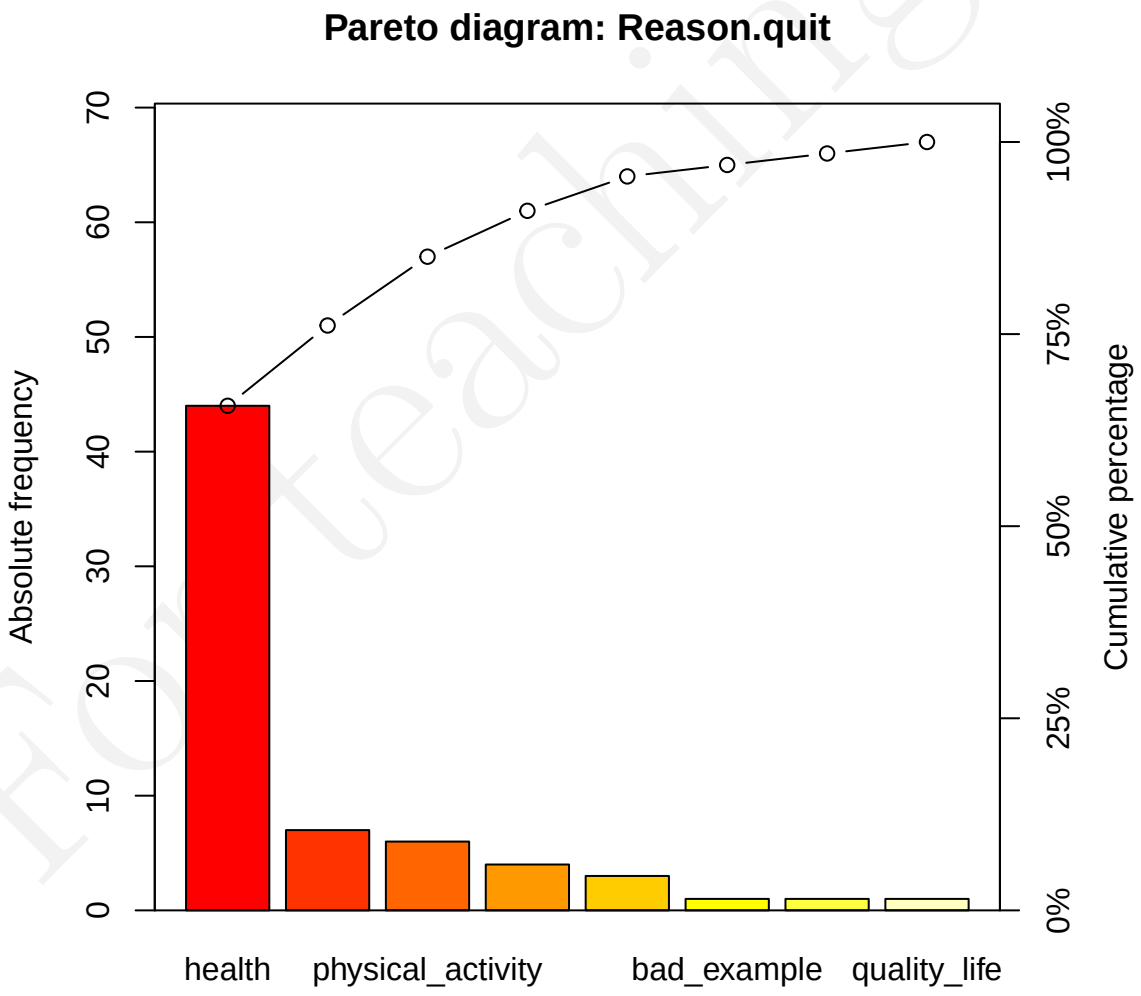
Id	Gender	Reason.quit		Id	Gender	Reason.quit
1	female	health		35	female	physical activity
2	female	not smoking		36	female	health
3	female	health		37	male	money
4	male	not smoking		38	female	health
5	female	not smoking		39	female	health
6	female	health		40	female	health
7	female	health		41	male	physical activity
8	male	health		42	male	physical activity
9	female	health		43	female	bad example
10	female	health		44	male	physical activity
11	male	quality of life		45	male	pregnancy.child
12	male	health		46	female	health
13	female	health		47	male	health
14	female	health		48	male	health
15	female	bad experience		49	female	health
16	female	health		50	female	health
17	female	health		51	female	health
18	female	health		52	male	health
19	female	health		53	male	physical activity
20	female	money		54	female	money
21	female	health		55	female	health
22	female	health		56	female	money
23	male	health		57	female	health
24	male	health		58	female	money
25	female	health		59	female	health
26	male	health		60	female	money
27	male	health		61	female	health
28	female	pregnancy.child		62	female	health
29	male	health		63	male	pregnancy.child
30	female	physical activity		64	female	health
31	female	health		65	female	health
32	female	pregnancy.child		66	female	health
33	female	health		67	female	health
34	male	money				

5.1 Univariate analysis

5.1.1 Graphical description

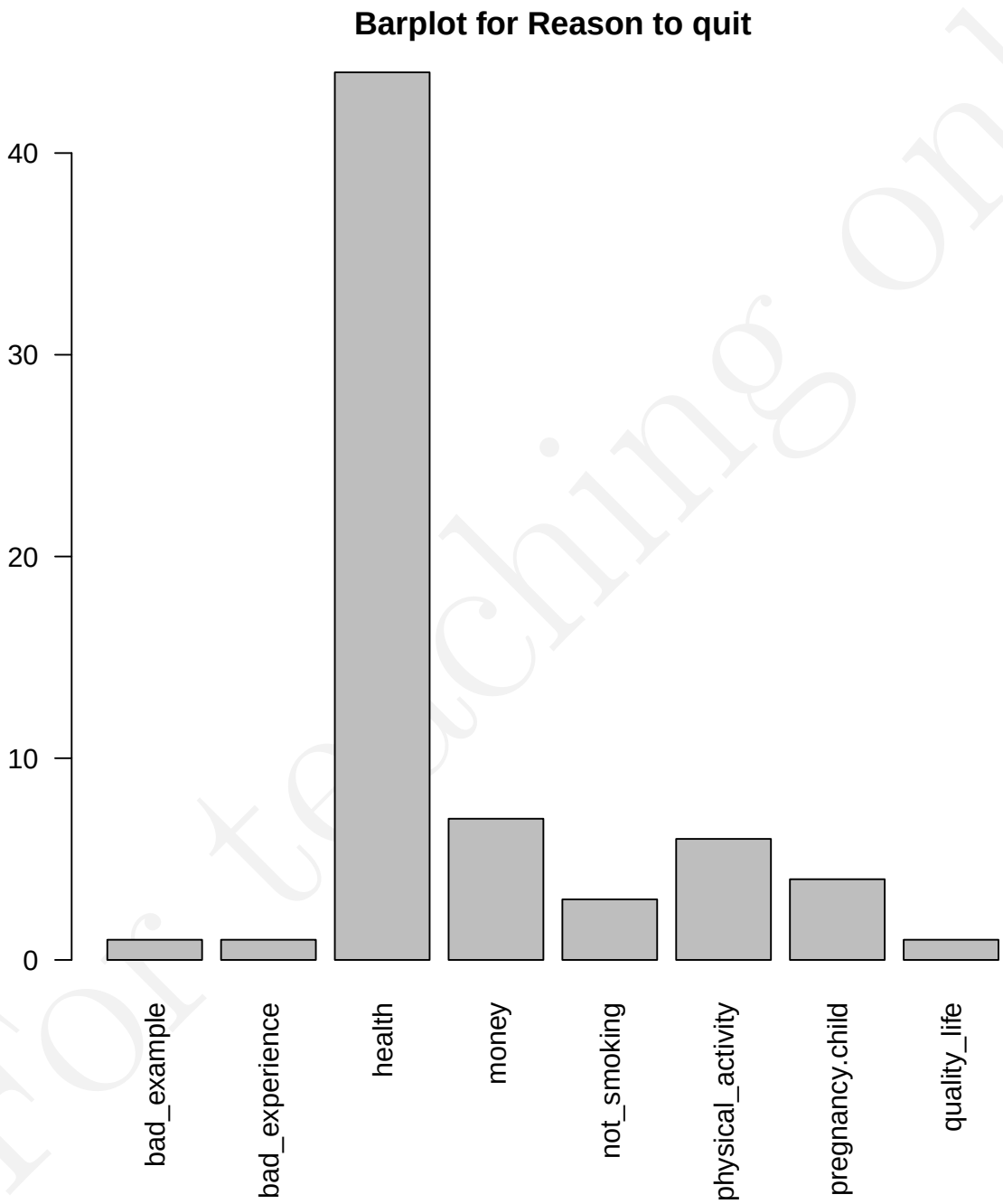
Sorting the categories in descending order according to their frequency:

```
Datos <- read.table(file.choose(),header=TRUE,fill=TRUE)
require(RcmdrPlugin.EACSPIR)
.tabla <- table(Datos$Reason.quit)
.tabla <- .tabla[order(-.tabla)]
par(mar=c(5,4,4,4))
.x <- barplot(.tabla, main='Pareto diagram: Reason.quit',
              ylab='Absolute frequency',ylim=c(0,sum(.tabla)*1.05),
              col=heat.colors(length(.tabla)))
lines(.x[1:length(.tabla)],cumsum(.tabla),type='b')
box()
axis(4,at=seq(0,max(cumsum(.tabla)),length=5),
     labels=paste(seq(0,1,length=5)*100,'%',sep=''))
mtext('Cumulative percentage', 4, line=2.5, las=3)
```

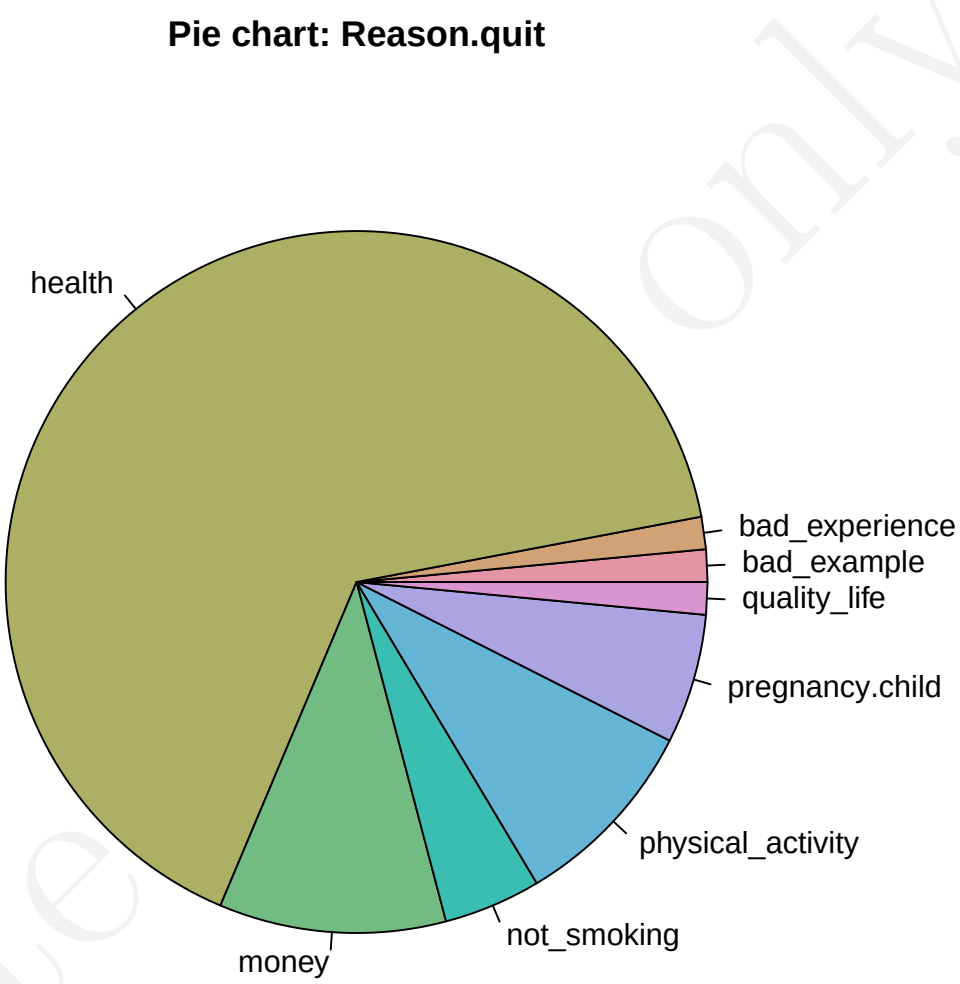


Alphabetical order of categories:

```
par(mar=c(10, 3, 3, 3))
barplot(table(Datos$Reason.quit), xlab=" ", ylab="Frequency",
        main="Barplot for Reason to quit",las=2)
```



```
library(colorspace)
pie(table(Datos$Reason.quit), labels=levels(Datos$Reason.quit),
    main="Pie chart: Reason.quit",
    col=rainbow_hcl(length(levels(Datos$Reason.quit))))
```



The most frequent category is clear in all of the graphical representations, as also is the fact that there is variability in the sense that several categories are present, but this variability is not that marked, given the relative frequency of the modal category.

5.1.2 Numerical description

```
# Absolute frequencies
table(Datos$Reason.quit)

##
##       bad_example    bad_experience            health             money
##                 1                 1                44                 7
##       not_smoking physical_activity   pregnancy.child      quality_life
##                 3                 6                 4                 1

#Relative frequencies
round(table(Datos$Reason.quit)/sum(table(Datos$Reason.quit)),2)

##
##       bad_example    bad_experience            health             money
##              0.01              0.01              0.66              0.10
##       not_smoking physical_activity   pregnancy.child      quality_life
##              0.04              0.09              0.06              0.01

# Mode: A central tendency measure
names(table(Datos$Reason.quit))[which(table(Datos$Reason.quit)==max(table(Datos$Reason.quit)))]

## [1] "health"
```

The odds of the modal category *health* can be found via the following formula:

$$\begin{aligned} Odds_{health} &= \frac{n_{health}}{n_{total} - n_{health}} \\ &= \frac{44}{67 - 44} \\ &= \frac{44}{23} \\ &\simeq 1.91 \end{aligned}$$

This value means that for every person stating some other reason there are approximately 2 individuals mentioning *health* as the reason that would make them quit smoking. In R this value is found as follows:

```
freq <- 0
for (i in 1:length(Datos$Reason.quit)) if (Datos$Reason.quit[i]=="health") freq <- freq + 1
freq/(sum(table(Datos$Reason.quit))-freq)

## [1] 1.913043
```

In the following we illustrate the calculation of two indices of dispersion (scatter, variability). For both of them 0 would mean minimal variability (all individuals are in the same category). The maximal value for the variation ratio approaches 1 as the modal category becomes less relatively frequent, whereas the maximal value for the index of qualitative variation is 1 when all categories are equally represented. First, the variation ratio.

$$\begin{aligned} VR &= 1 - \frac{n_{health}}{n_{total}} \\ &= 1 - \frac{44}{67} \\ &\simeq 1 - 0.66 \\ &= 0.34 \end{aligned}$$

In R this value is found as follows:

```
freq <- 0
for (i in 1:length(Datos$Reason.quit)) if (Datos$Reason.quit[i]=="health") freq <- freq + 1
1 - freq/sum(table(Datos$Reason.quit))

## [1] 0.3432836
```

Second, we obtain the value of the index of qualitative variation (the term k in the expression below denotes the number of categories, p_i is the proportion of each category, and B denotes Blau's index):

$$\begin{aligned}
IQV &= \frac{B \times k}{k-1} \\
&= \frac{(1-\sum_{i=1}^{k} p_i^2) \times k}{k-1} \\
&= \frac{(1-(0.01^2+0.01^2+0.66^2+0.10^2+0.04^2+0.09^2+0.06^2+0.01^2)) \times 8}{8-1} \\
&= \frac{(1-(0.0001+0.0001+0.4356+0.01+0.0016+0.0081+0.0036+0.001)) \times 8}{8-1} \\
&= \frac{(1-0.46) \times 8}{7} \\
&= \frac{0.54 \times 8}{7} \\
&= \frac{4.32}{7} \\
&\simeq 0.62
\end{aligned}$$

In R this value is found as follows:

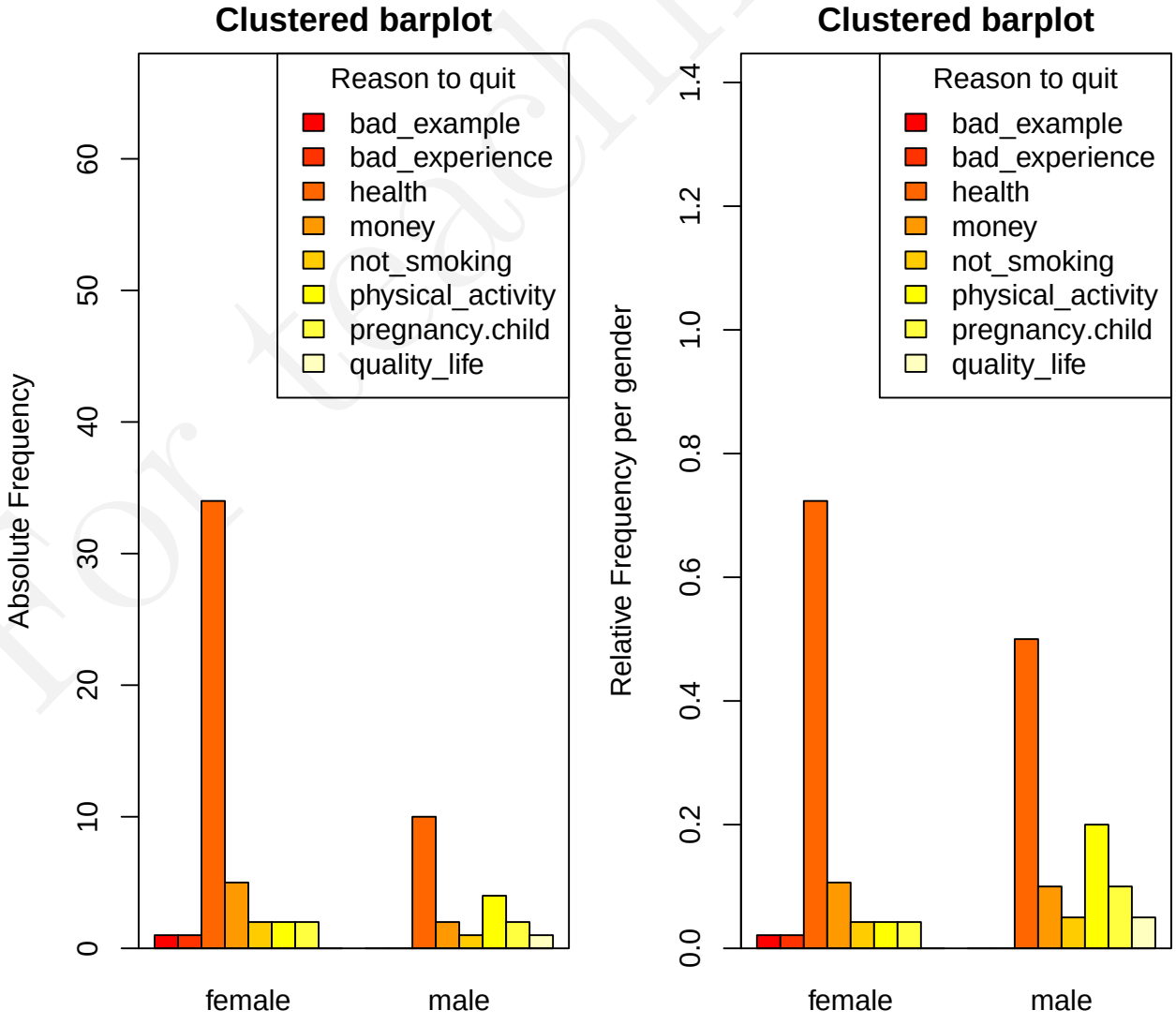
```
((1-sum(prop.table(table(Datos$Reason.quit))^2))*length(table(Datos$Reason.quit)))/
  (length(table(Datos$Reason.quit))-1)

## [1] 0.621201
```

5.2 Bivariate analysis

5.2.1 Graphical description

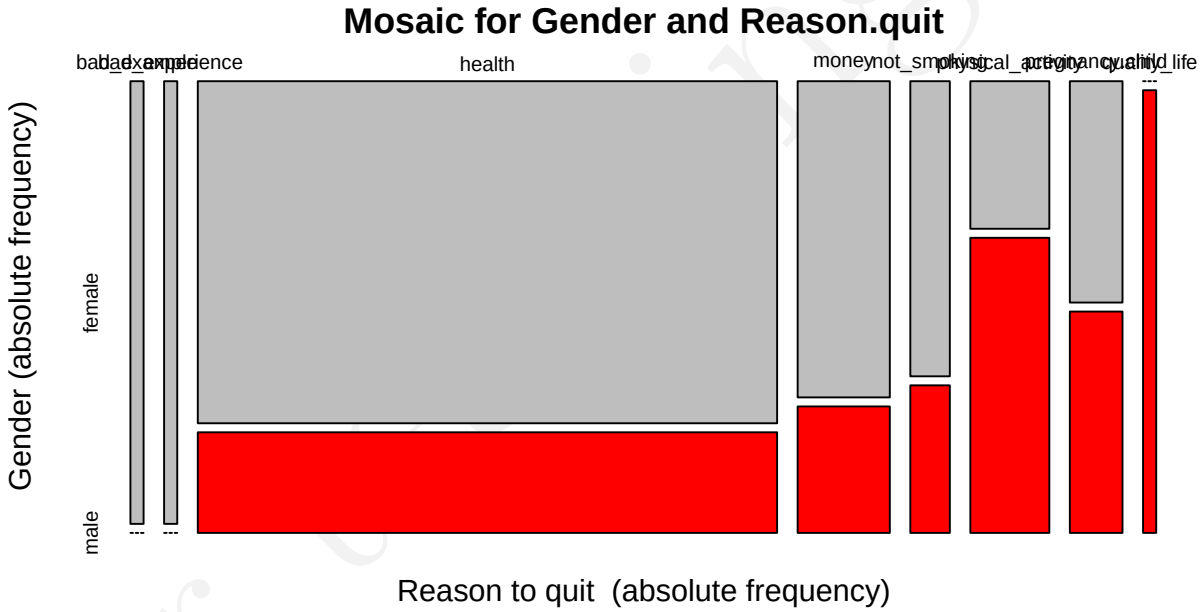
Clustered barplot (on the basis of absolute or relative frequencies per gender). Using relative frequencies is expected to make the interpretation easier, as the hights of bars for the different genders become comparable:

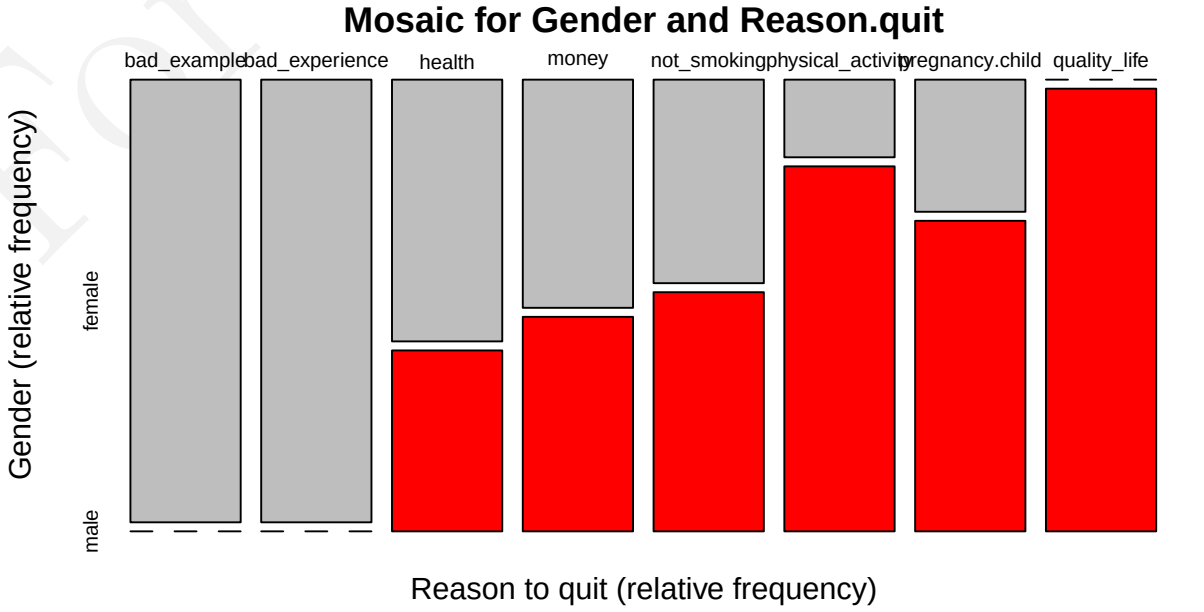
```
par(mfrow=c(1,2),mar=c(2, 4, 2, 1))
.Tabla <- xtabs(~Reason.quit+Gender, data=Datos)
barplot(.Tabla,beside=TRUE,main='Clustered barplot',
  ylab='Absolute Frequency',xlab='Gender',ylim=c(0,max(.Tabla)*2),
  col=heat.colors(length(levels(Datos$Reason.quit))),legend.text=TRUE,
  args.legend=list(x='topright',title='Reason to quit'))
box()

.Tabla2 <- xtabs(~Reason.quit+Gender, data=Datos)/sum(xtabs(~Reason.quit+Gender, data=Datos))
.Tabla2[1:8] <- .Tabla[1:8]/sum(.Tabla[1:8])
.Tabla2[9:16] <- .Tabla[9:16]/sum(.Tabla[9:16])
barplot(.Tabla2,beside=TRUE,main='Clustered barplot',
  ylab='Relative Frequency per gender',xlab='Gender',ylim=c(0,max(.Tabla2)*2),
  col=heat.colors(length(levels(Datos$Reason.quit))),legend.text=TRUE,
  args.legend=list(x='topright',title='Reason to quit'))
box()
```



Mosaic diagram using absolute frequencies or relative frequencies for the two categorical variables. Although the names of the *Reason.quit* categories are not easily read, the visual inspection of the data suggest that there is some association between the two variables:

```
par(mfrow=c(2,1),mar=c(2, 2, 2, 2))
.Tabla <- xtabs(~Reason.quit+Gender, data=Datos)
mosaicplot(.Tabla,main='Mosaic for Gender and Reason.quit',
           xlab="Reason to quit  (absolute frequency)",
           ylab= "Gender (absolute frequency)", col=dim(.Tabla))

.Tabla3 <- xtabs(~Reason.quit+Gender, data=Datos)/sum(xtabs(~Reason.quit+Gender, data=Datos))
for (i in 1:8)
.Tabla3[i] <- .Tabla2[i]/sum(.Tabla2[i],.Tabla2[i+8])
for (i in 9:16)
.Tabla3[i] <- .Tabla2[i]/sum(.Tabla2[i],.Tabla2[i-8])
mosaicplot(.Tabla3,main='Mosaic for Gender and Reason.quit',
           xlab="Reason to quit (relative frequency)",
           ylab= "Gender (relative frequency)",col=dim(.Tabla))
```





5.2.2 Numerical description

We begin the bivariate analysis of *Reason.quit* and *Gender* (i.e., the exploration of whether these two variables are related and to what degree) with the contingency table:

```
obs.freq <- xtabs(~Reason.quit+Gender, data=Datos)
obs.freq
```

```
##                    Gender
## Reason.quit         female male
##   bad_example            1    0
##   bad_experience         1    0
##   health                34   10
##   money                  5    2
##   not_smoking            2    1
##   physical_activity      2    4
##   pregnancy.child        2    2
##   quality_life           0    1
```

We compare this table to the one that would have been obtained (for the same number of male and female, and for the same number of individuals pointing health, money, etc. as the main reason for quitting):

```
exp.freq <- chisq.test(obs.freq, correct=FALSE)$expected
```

The values obtained are similar, but not exactly the same. In order not to compare the observed frequencies $f_{o.ij}$ and the expected frequencies $f_{e.ij}$ only visually, we use the chi-square statistic for all $I = 2$ columns and $J = 8$ rows:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{o.ij}-f_{e.ij})^2}{f_{e.ij}} \\ &= \frac{(1-0.07)^2}{0.07}+\frac{(1-0.07)^2}{0.07}+\frac{(34-30.9)^2}{30.9}+\frac{(5-4.9)^2}{4.9}+\ldots+\frac{(1-0.3)^2}{0.3} \\ &= 8.95\end{aligned}$$

In R this is achieved via:

```
# Using the information from the statistical test
chisq.test(obs.freq, correct=FALSE)$statistic
```

```
## X-squared
##  8.949378
```

```
# Applying the formula on the basis of the two tables
sum(((obs.freq-exp.freq)^2)/exp.freq)
```

```
## [1] 8.949378
```

Given that $\chi^2 \neq 0$ there is some association between the two variables (i.e., they are not indepedent). In order to assess the magnitude of association, we will compute Cramér's V index, taking into account the fact that q denotes the number of categories in the variable that has fewer categories: $q = \min I, J$

$$
\begin{aligned}
V &= \sqrt{\frac{\frac{\chi^2}{n}}{q-1}} \\
&= \sqrt{\frac{\frac{8.95}{67}}{2-1}} \\
&= \sqrt{\frac{0.13}{1}} \\
&= \sqrt{0.13} \\
&\simeq 0.37
\end{aligned}
$$

Finally, we can compute the odds ratio for the *health* reason, comparing men and women. First, we compute the odds for *health* in men:

$$
\begin{aligned}
Odds_{health:men} &= \frac{n_{health:men}}{n_{total:men} - n_{health:men}} \\
&= \frac{10}{20-10} \\
&= \frac{10}{10} \\
&= 1
\end{aligned}
$$

Second, we compute the odds for *health* in women:

$$
\begin{aligned}
Odds_{health:women} &= \frac{n_{health:women}}{n_{total:women} - n_{health:women}} \\
&= \frac{34}{47-34} \\
&= \frac{34}{13} \\
&\simeq 2.62
\end{aligned}
$$

Third, we compute the odds ratio as

$$
OR = \frac{Odds_{health:women}}{Odds_{health:men}} = 2.62
$$

That is, it is approximataley 2 and a half times more likely to state *health* as a reason for quitting smoking if the person as a woman compared to being a man. It is another way of stating that the two variables are related and another way of quantifying the strength of association.

In R this is achieved via:

```
# Using the information from the statistical test
freq_m <- 0
for (i in 1:length(Datos$Reason.quit))
  if ((Datos$Reason.quit[i]=="health") && (Datos$Gender[i]=="male"))
    freq_m <- freq_m + 1
freq_f <- 0
for (i in 1:length(Datos$Reason.quit))
  if ((Datos$Reason.quit[i]=="health") && (Datos$Gender[i]=="female"))
    freq_f <- freq_f + 1
(freq_f/(sum(obs.freq[,1])-freq_f))/(freq_m/(sum(obs.freq[,2])-freq_m))

## [1] 2.615385
```

6 Quantitative data

6.1 Univariate analysis

6.1.1 Example 1: Multiple-choice questions

1. Numerical analysis

```
quantile(results_mc) # quantiles
```

```
##      0%     25%     50%     75%    100%
## 0.1300 0.4925 0.6000 0.9000 1.1000
```

```
mean(results_mc) # Mean
```

```
## [1] 0.65
```

```
median(results_mc) # Median
```

```
## [1] 0.6
```

```
# Mode for a numeric variable
frecu <- as.data.frame(table(results_mc))
frecu2 <- as.matrix(frecu)
as.numeric(frecu2[which(frecu$Freq == max(frecu$Freq)),1])
```

```
## [1] 0.9
```

```
(quantile(results_mc,probs=0.75)[[1]] +
   quantile(results_mc,probs=0.25)[[1]]) /2 # Midhinge
```

```
## [1] 0.69625
```

```
require(e1071) # Package necessary for computing skewness
skewness(results_mc)
```

```
## [1] -0.283752
```

```
max(results_mc)-min(results_mc) # Range
```

```
## [1] 0.97
```

```
IQR(results_mc) # Interquartile range
```

```
## [1] 0.4075
```

```
sqrt(mean((results_mc-mean(results_mc))^2)) # Std dev not equal to sd(results_mc)
```

```
## [1] 0.2510229
```

```
median(abs(results_mc-median(results_mc)))  # MAD not equal to mad(results_mc)
```

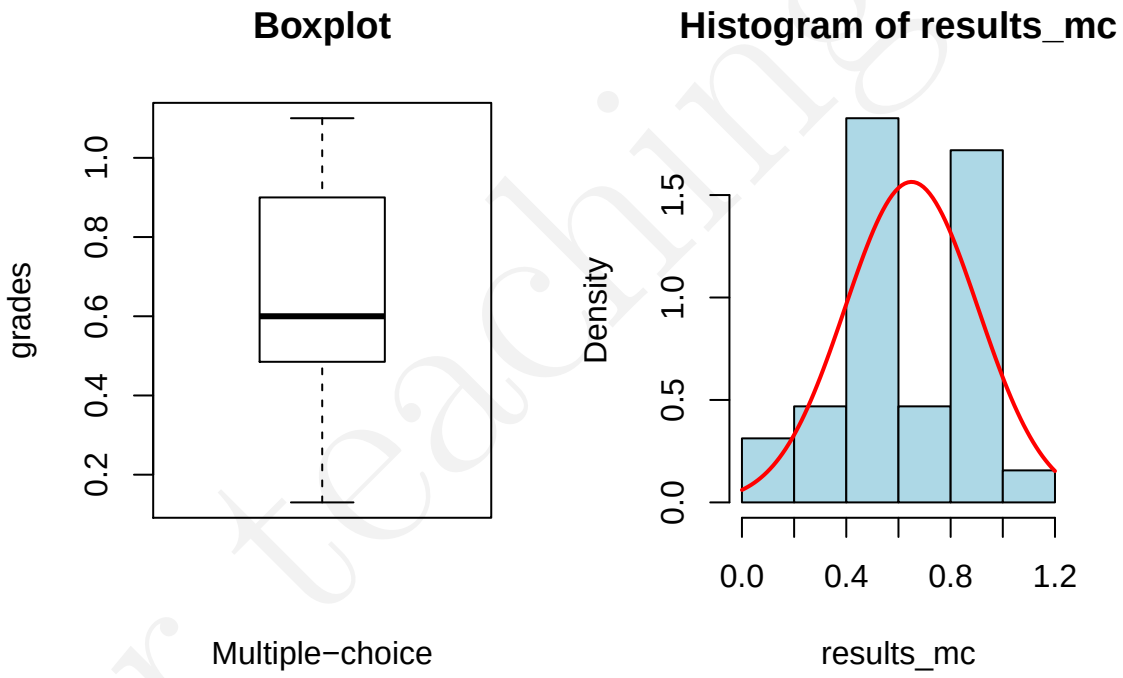
```
## [1] 0.215
```

```
# Coefficient of variation
sqrt(mean((results_mc-mean(results_mc))^2)) / abs(mean(results_mc))
```

```
## [1] 0.3861891

# Quartile coefficient of variation
(quantile(results_mc,probs=0.75)[[1]] - quantile(results_mc,probs=0.25)[[1]]) /
  (quantile(results_mc,probs=0.75)[[1]] + quantile(results_mc,probs=0.25)[[1]])

## [1] 0.2926391
```

2. Graphical analysis

```
# Graphical representation
par(mfrow=c(1,2))
boxplot(results_mc, main="Boxplot", ylab="grades", xlab="Multiple-choice")
hist(results_mc,col="lightblue",freq=FALSE) # histogram
curve(dnorm(x, mean=mean(results_mc), sd=sd(results_mc)),
          col="red", lwd=2, add=TRUE, yaxt="n")
```




```
stem(results_mc)

##
##   The decimal point is 1 digit(s) to the left of the |
##
##    1 | 33
##    2 | 7
##    3 | 0
##    4 | 0377
##    5 | 000777
##    6 | 0003
##    7 | 7
##    8 | 033
##    9 | 000000337
##   10 |
##   11 | 0
```

6.1.2 Example 2: Open-ended questions

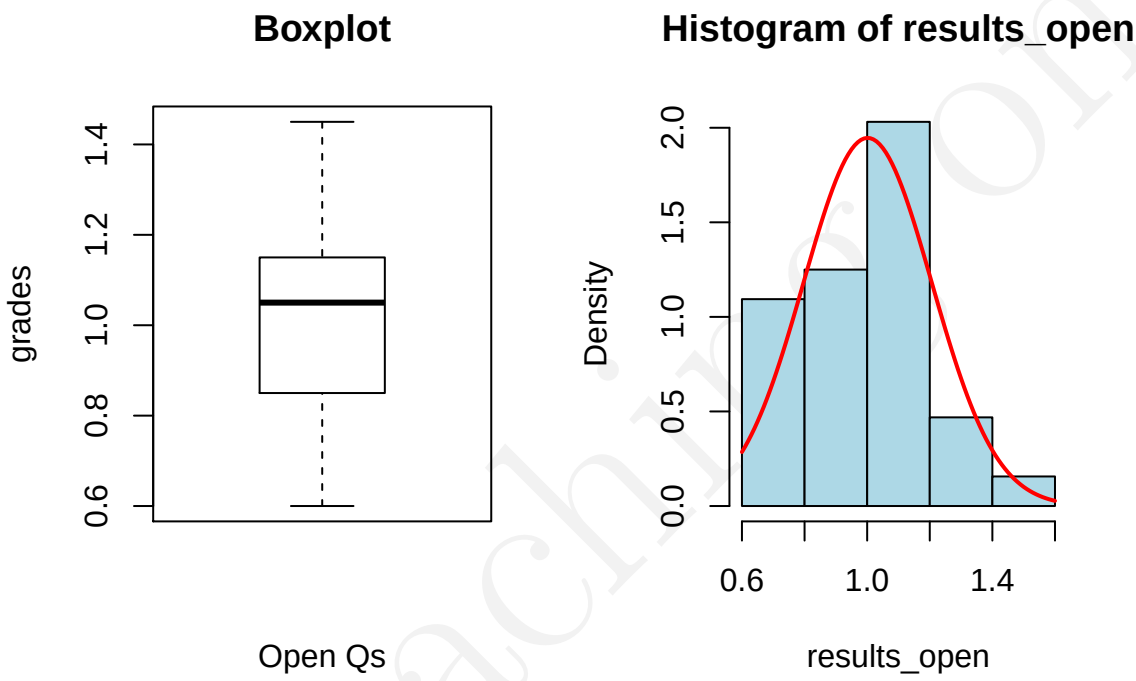
1. Numerical analysis

```
quantile(results_open) # quantiles

##   0%  25%  50%  75% 100%
## 0.60 0.85 1.05 1.15 1.45

mean(results_open) # Mean

## [1] 1.001562

median(results_open) # Median

## [1] 1.05

# Mode for a numeric variable
frecu <- as.data.frame(table(results_open))
frecu2 <- as.matrix(frecu)
as.numeric(frecu2[which(frecu$Freq == max(frecu$Freq)),1])

## [1] 1.05

(quantile(results_open,probs=0.75)[[1]] +
   quantile(results_open,probs=0.25)[[1]]) /2 # Midhinge

## [1] 1

require(e1071) # Package necessary for computing skewness
skewness(results_open)

## [1] 0.008236599
```

```
max(results_open)-min(results_open) # Range

## [1] 0.85

IQR(results_open) # Interquartile range

## [1] 0.3

sqrt(mean((results_open-mean(results_open))^2)) # St dev not equal to sd(results_mc)

## [1] 0.2017441

median(abs(results_open-median(results_open))) # MAD not equal to mad(results_mc)

## [1] 0.15

# Coefficient of variation
sqrt(mean((results_open-mean(results_open))^2)) / mean(results_open)

## [1] 0.2014294

# Quartile coefficient of variation
(quantile(results_open,probs=0.75)[[1]] - quantile(results_open,probs=0.25)[[1]]) /
  (quantile(results_open,probs=0.75)[[1]] + quantile(results_open,probs=0.25)[[1]])

## [1] 0.15
```

2. Graphical analysis

```
# Graphical representation
par(mfrow=c(1,2))
boxplot(results_open, main="Boxplot", ylab="grades", xlab="Open Qs")
hist(results_open,col="lightblue",freq=FALSE) # histogram
curve(dnorm(x, mean=mean(results_open), sd=sd(results_open)),
          col="red", lwd=2, add=TRUE, yaxt="n")
```




```
stem(results_open)

##
##   The decimal point is 1 digit(s) to the left of the |
##
##    6 | 05
##    7 | 5555
##    8 | 055
##    9 | 00055
##   10 | 05555555
##   11 | 555
##   12 | 0005
##   13 | 00
##   14 | 5
```

6.1.3 Summary: Indices

Central tendency				
Index	**Formula**	**Basis**	**Resistant?**	**Dimension**
Mean	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	Moments	No	Units (e.g., m, kg)
Median	$Md = x_{(\frac{n-1}{2})}$ for odd n $Md = \frac{x_{(\frac{n}{2})} + x_{(\frac{n+1}{2})}}{2}$ for even n where the x values are sorted from smallest, 1, to largest, n	Position	Yes	Units (e.g., m, kg)
Mode	The most frequent value x_i for which $max(f_j)$, with n measurements $(i = 1, 2..., n)$ and k different values $(j = 1, 2, ..., k; k \leq n)$	Frequency	Yes	Units (e.g., m, kg)
Midhinge	$\bar{Q} = \frac{Q_1 + Q_3}{2}$	Position	Yes	Units (e.g., m, kg)

Dispersion = Scatter = Variability = Heterogeneity				
Index	**Formula**	**Basis**	**Resistant?**	**Dimension**
Variance	$Var \equiv s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$	Moments	No	Units^2 (e.g., m^2, kg^2)
Standard deviation	$SD \equiv s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$	Moments	No	Units (e.g., m, kg)
MAD	$MAD = Md\|x_i - Md(x_i)\|$	Position	Yes	Units (e.g., m, kg)
Range	$Range = max(x_i) - min(x_i)$	Position	No	Units (e.g., m, kg)
Interquartile range	$IQR = Q_3 - Q_1$	Position	Yes	Units (e.g., m, kg)
Coefficient of variation	$CV = \frac{s}{abs(\bar{x})}$	Moments	No	None (pure number)
Quartile coefficient of variation	$CV_Q = \frac{Q_3 - Q_1}{Q_3 + Q_1}$	Position	Yes	None (pure number)

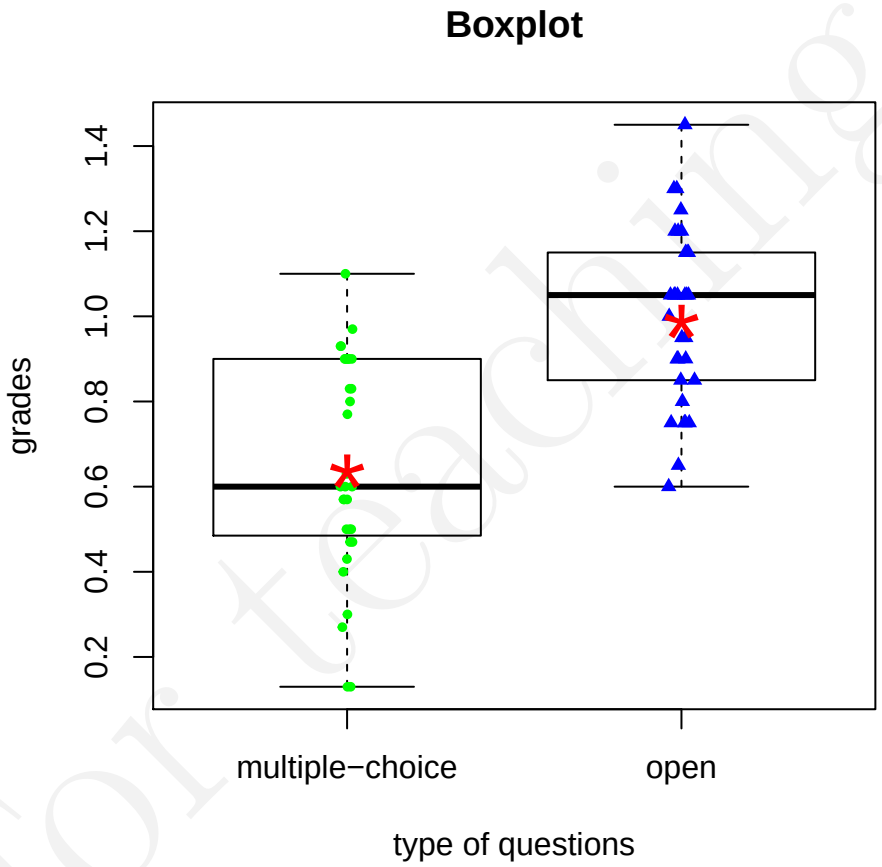
Shape = Skewness & Kurtosis				
Index	**Formula**	**Basis**	**Resistant?**	**Dimension**
Skewness	$\gamma_1 = \frac{\frac{\sum_{i=1}^{n}(y_i - \mu)^3}{n}}{\sigma^3}$	Moments	No	None (pure number)
Yule's index	$H_1 = \frac{Q_3 + Q_1 - 2 \times Md}{2 \times Md}$	Position	Yes	None (pure number)
Kelly's index	$H_1 = \frac{P_{10} + P_{90} - 2 \times Md}{2 \times Md}$	Position	Yes	None (pure number)
Kurtosis	$\gamma_2 = \frac{\frac{\sum_{i=1}^{n}(y_i - \mu)^4}{n}}{\sigma^4} - 3$	Moments	No	None (pure number)
K_2	$K_2 = \frac{P_{90} - P_{10}}{1.9 \times (P_{75} - P_{25})}$	Position	Yes	None (pure number)
K_3	$K_3 = \frac{P_{87.5} - P_{12.5}}{1.7 \times (P_{90} - P_{10})}$	Position	Yes	None (pure number)

6.2 Bivariate analysis

6.2.1 Bivariate analysis: Comparing grades according to type of question

1. Graphical representation: Based on position

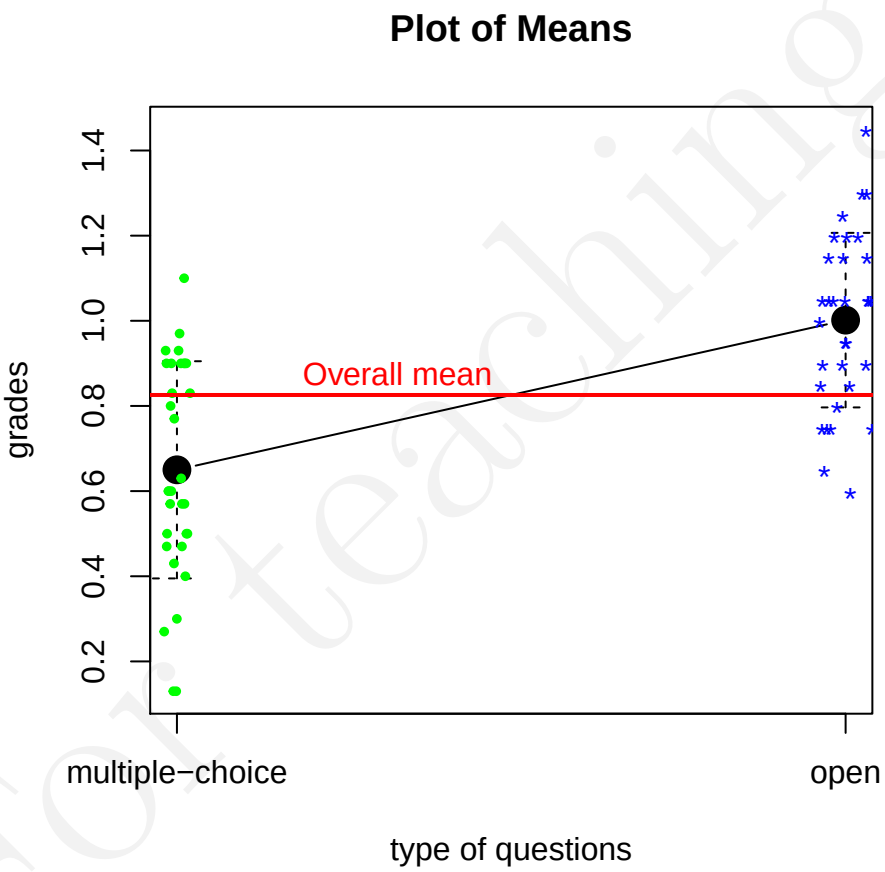
```
# Boxplot: Based on position indices
boxplot(as.numeric(alldata[,1])~alldata[,2],
        main="Boxplot", ylab="grades", xlab="type of questions")
points(jitter(as.numeric(rep(2,length(results_open)),factor=0.1)),
       as.numeric(alldata[(alldata[,2]=="open"),1]),col="blue",pch=17,cex=0.8)
points(2,mean(results_open),pch="*",col="red",cex=3)
points(jitter(as.numeric(rep(1,length(results_mc)),factor=0.1)),
       as.numeric(alldata[(alldata[,2]=="multiple-choice"),1]),col="green",pch=20,cex=0.8)
points(1,mean(results_mc),pch="*",col="red",cex=3)
```




```
# Compare numerically: Position indices
require(Rcmdr)
numSummary(as.numeric(alldata[,1]), groups=alldata[,2],
           statistics=c("quantiles", "IQR"), quantiles=c(0,.25,.5,.75,1))

##                    IQR   0%    25%  50%  75% 100% data:n
## multiple-choice 0.4075 0.13 0.4925 0.60 0.90 1.10     32
## open            0.3000 0.60 0.8500 1.05 1.15 1.45     32
```

2. Graphical representation: Based on moments

```
# Plot of within-group means (points) and standard deviations (bars)
require(Rcmdr)
plotMeans(as.numeric(alldata[,1]), as.factor(alldata[,2]),
          error.bars="sd", ylab="grades",
          xlab="type of questions",
          ylim=c(min(as.numeric(alldata[,1])),max(as.numeric(alldata[,1]))))
points(jitter(as.numeric(rep(2,length(results_open)),factor=0.1)),
       as.numeric(alldata[(alldata[,2]=="open"),1]),col="blue",pch="*")
points(jitter(as.numeric(rep(1,length(results_open)),factor=0.1)),
       as.numeric(alldata[(alldata[,2]=="multiple-choice"),1]),col="green",pch=20,cex=0.8)

abline(h=mean(as.numeric(alldata[,1])),col="red",lwd=2)
text(1.33,(mean(as.numeric(alldata[,1]))+0.05),"Overall mean",col="red")
```



3. Eta-squared An index based on moments, for comparing two or more groups via quantifying the amount of variability of the quantitative variable explained by the grouping variable.

$$\eta^2 = \frac{SC_{factor}}{SC_{total}} = \frac{\sum_{j=1}^{a} n_j \times (\bar{x}_j - \bar{x})^2}{\sum_{j=1}^{a}\sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2}$$
$$= \frac{SC_{factor}}{SC_{factor} + SC_{error}} = \frac{\sum_{j=1}^{a} n_j \times (\bar{x}_j - \bar{x})^2}{\sum_{j=1}^{a} n_j \times (\bar{x}_j - \bar{x})^2 + \sum_{j=1}^{a}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2}$$

where x_{ij} is each value $i (i = 1, 2, ..., n_j)$, n_j is the number of measurements in group j, and a is the number of groups $j (j = 1, 2, ..., a)$.

```
# Install package can be done via the following line of code
#install.packages("BaylorEdPsych")
#Load package
require(BaylorEdPsych)
# Obtain eta-squared
analysis <- aov(as.numeric(alldata[,1])~alldata[,2])
EtaSq(analysis)

##                  Eta^2 Partial Eta^2
## alldata[, 2] 0.3733769     0.3733769
```

```
# (Not squared) Between groups variability: explained by "type of question"
mean(as.numeric(alldata[(alldata[,2]=="open"),1])) -
  mean(as.numeric(alldata[,1]))

## [1] 0.1757812

mean(as.numeric(alldata[(alldata[,2]=="multiple-choice"),1])) -
  mean(as.numeric(alldata[,1]))

## [1] -0.1757812

# Not squared) Intra-group variability: unexplained
as.numeric(alldata[(alldata[,2]=="open"),1]) -
  mean(as.numeric(alldata[(alldata[,2]=="open"),1]))

##  [1]  0.1984375  0.2984375  0.0484375 -0.1015625  0.2984375 -0.1015625
##  [7] -0.4015625 -0.0515625  0.0484375 -0.2515625  0.0484375  0.4484375
## [13] -0.2515625  0.1484375  0.0484375 -0.1015625 -0.0015625  0.1984375
## [19] -0.2515625  0.0484375 -0.0515625 -0.2515625  0.0484375  0.1484375
## [25] -0.3515625  0.1984375 -0.1515625 -0.1515625  0.2484375  0.1484375
## [31] -0.2015625  0.0484375

as.numeric(alldata[(alldata[,2]=="multiple-choice"),1]) -
  mean(as.numeric(alldata[(alldata[,2]=="multiple-choice"),1]))

##  [1]  0.15  0.18 -0.02 -0.52  0.45 -0.08  0.25  0.18 -0.25 -0.15 -0.15
## [12] -0.15 -0.22  0.28  0.25 -0.38 -0.05  0.25  0.25  0.28 -0.18 -0.35
## [23] -0.52  0.25 -0.18  0.25 -0.05 -0.05  0.12 -0.08 -0.08  0.32
```

4. Cohen's d An index based on moments, for comparing two groups in terms of a standardized mean difference.

   Shorter version of the formula for groups of the same size

$$d = \frac{\bar{x}_{group1} - \bar{x}_{group2}}{s_{pooled}} = \frac{\bar{x}_{group1} - \bar{x}_{group2}}{\sqrt{\frac{s^2_{group1} + s^2_{group2}}{2}}}$$

```
# Numerator: difference in means
num <- mean(results_open) - mean(results_mc)
num

## [1] 0.3515625

# Denominator: pooled estimation of within-group variability
denom <- sqrt( (mean((results_mc-mean(results_mc))^2)+
        mean((results_open-mean(results_open))^2)) /2 )
denom

## [1] 0.2277204

# Cohen's d
num/denom

## [1] 1.543834
```

   Version of the formula for groups of different size

$$d = \frac{\bar{x}_{group1} - \bar{x}_{group2}}{S_{pooled}} = \frac{\bar{x}_{group1} - \bar{x}_{group2}}{\sqrt{\frac{(n_{group1}-1)\times S^2_{group1} + (n_{group2}-1)\times S^2_{group2}}{n_{group1}+n_{group2}-2}}}$$

```
# Numerator: difference in means
num <- mean(results_open) - mean(results_mc)
num

## [1] 0.3515625

# Denominator: pooled estimation of within-group variability
denom <- sqrt( ((length(results_open)-1)*mean((results_open-mean(results_open))^2)  +
               (length(results_mc)-1)*mean((results_mc-mean(results_mc))^2) ) /
                  (length(results_open) + length(results_mc) - 2) )
denom

## [1] 0.2277204

# Cohen's d
num/denom

## [1] 1.543834
```

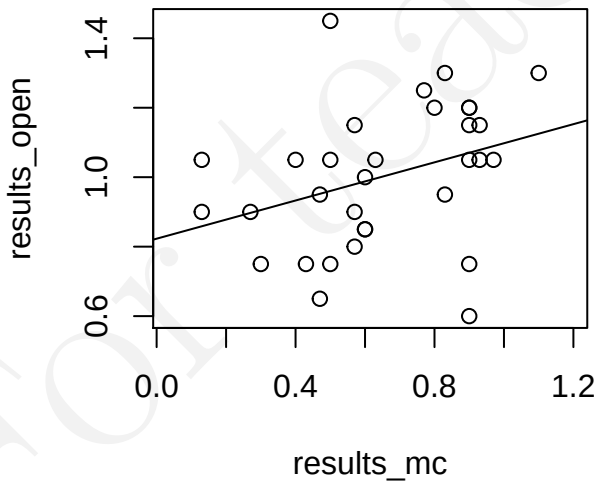
6.2.2 Bivariate analysis: Correlation between grades

1. Numerical results

```
# Pearson's product moment correlation coefficient
cor(results_mc,results_open)

## [1] 0.3418579

# Spearman's rank correlation coefficient
cor(results_mc,results_open,method="spearman")

## [1] 0.3938791

# Kendall's tau-a
cor(results_mc,results_open,method="kendall")

## [1] 0.2789386
```

2. Graphical results

```
# Graphical representation: Scatterplot
plot(results_mc,results_open,asp=1)
abline(lm(results_open~results_mc))
```



6.2.3 Summary: Correlation indices

Index	Formula	Data	Resistant?	Quantifies
Pearson's	$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\times\sum_{i=1}^{n}(y_i-\bar{y})^2}}$	Interval scale	No	Linear relation
Spearman's	$r_s = \frac{\sum_{i=1}^{n}(r_i-\bar{r})\times(s_i-\bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i-\bar{r})^2\times\sum_{i=1}^{n}(s_i-\bar{s})^2}}$	Few tied ranks	Yes	Monotonic relation
Kendall's	$\tau_a = \frac{Concordances-Discordances}{\frac{n\times(n-1)}{2}}$	Ordinal scale	Yes	Degree of concordance

7 References

Crawford, J. R., & Garthwaite, P. H. (2012). Single-case research in neuropsychology: A comparison of five forms of *t*-test for comparing a case to controls. *Cortex*, *48*, 1009-1016.

Crawford, J. R., Garthwaite, P. H., & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. *Cognitive Neuropsychology*, *27*, 245-260.

Crawford, J. R., & Howell, D. C. (1998). Comparing an individual's test score against norms derived from small samples. *The Clinical Neuropsychologist*, *12*, 482-486.

Gravetter, F. J., & Wallnau, L. B. (2009). *Statistics for the behavioral sciences* (8th ed.). Belmont, CA: Wadsworth.

Janssen, P. A., Thiessen, P., Klein, M. C., Whitfield, M. F., McNab, Y. C., & Cullis-Kuhl, S. C. (2007). Standards for the measurement of birth weight, length and head circumference at term in neonates of European, Chinese and South Asian ancestry. *Open Medicine*, *1*, e74-e88.

Peró, M., Leiva, D., Guàrdia, J., & Solanas, A. (Eds.) (2012). *Estadística aplicada a las ciencias sociales mediante R y R-Commander*. Madrid: Garceta.

Pfadt, A., & Wheeler, D. J. (1995). Using statistical process control to make data-based clinical decisions. *Journal of Applied Behavior Analysis*, *28*, 349-370.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

Solanas, A., Salafranca, Ll., Fauquet, J., & Núñez, M. I. (2005). *Estadística descriptiva en ciencias del comportamiento*. Madrid: Thomson.

Wilcox, A. J. (2001). On the importance–and the unimportance–of birthweight. *International Journal of Epidemiology*, *30*, 1233-1241.