Statistics in Computer Science
Seminar Exercises

Stanislav Katina, Mojmir Vinkler

December 14, 2015

1 Introduction

Exercise 1 (Simple random sample). *In a simple random sample of size n from population of finite size N, each element has an equal probability of being chosen. If we avoid choosing any member of the population more than once, we call it* ***simple random sample without replacement***[1]*.(Dalgaard 2008). If we put a member back to population after choosing it, we talk about* ***simple random sample with replacement***[2]*. Let's have a set $\mathcal{M}$ with $N = 10$ elements and we want to choose $n = 3$ elements (a) without replacement and (b) with replacement. How many combinations there are? How do these combinations look like if $\mathcal{M} = \{1, 2, \ldots, 10\}$? Do the same for $N = 100$, $n = 30$ and set $\mathcal{M} = \{1, 2, \ldots, 100\}$.*

Solution without R code:
(a) Number of combinations is $\binom{N}{n}$. If $N = 10$ and $n = 3$, then $\binom{N}{n} = \frac{N!}{(N-n)!n!} = \binom{10}{3} = 120$.
(b) Number of combinations with replacement is $\binom{N+n-1}{n}$. If $N = 10$ and $n = 3$, then $\binom{N+n-1}{n} = \frac{(N+n-1)!}{(N-1)!n!} = \binom{10+3-1}{3} = 220$. If $N = 100$ a $n = 30$, then $\binom{N+n-1}{n} = \binom{100+30-1}{30} = 2.009491 \times 10^{29}$.

Hints. choose(n,k), combn(n,k)[3], sample(x=..., size=..., replace=...)

Exercise 2 (Simple random sample). *A group of people are labeled by their identification numbers (ID) from 1 to 30. Choose (a) randomly 5 people out of 30 without replacement, (b) randomly 5 people out of 30 with replacement and finally (c) randomly 5 people out of 30 without replacement, where people with ID between 28 and 30 have $4\times$ higher probability of being chosen than people with ID between 1 and 27.*

Exercise 3 (Normal distribution). *Let X be a random variable (it could represent for example adult height) and let's assume it is normally distributed with parameters μ (expectation or mean) and σ^2 (standard deviation) which is written as $X \sim N(\mu, \sigma^2)$, $\mu = 140.83, \sigma^2 = 33.79$. Normal distribution represents a probability distribution model for this random variable. Calculate probability $\Pr(a < X < b) = \Pr(X < b) - \Pr(X < a) = F_X(b) - F_X(a)$, where*

[1] n-combination without replacement from N members of set $\mathcal{M}$.
[2] n-combination with replacement from N members of set $\mathcal{M}$.
[3] requires library utils

$a = \mu - k\sigma$, $b = \mu + k\sigma$, $k = 1, 2, 3$.[4] *Plot probability density function and fill area between points a and b and label axes x and y as shown in figure 1.*

Partial solution:
$a = \mu - \sigma = 135.0171$, $b = \mu + \sigma = 146.6429$,
$\Pr(|X - \mu| > \sigma) = 0.3173, \Pr(|X - \mu| < \sigma) = 1 - 0.3173 = 0.6827$,
$a = \mu - 2\sigma = 129.2042$, $b = \mu + 2\sigma = 152.4558$,
$\Pr(|X - \mu| > 2\sigma) = 0.0455, \Pr(|X - \mu| < 2\sigma) = 1 - 0.0455 = 0.9545$,
$a = \mu - 3\sigma = 123.3913$, $b = \mu + 3\sigma = 158.2687$,
$\Pr(|X - \mu| > 3\sigma) = 0.0027, \Pr(|X - \mu| < 3\sigma) = 1 - 0.0027 = 0.9973$.

```
mu <- 0
sig <- 1
bin <- seq(mu-3*sig,mu+3*sig,by=sig)
pnorm(bin[7]) - pnorm(bin[1]) # 0.9973002
pnorm(bin[6]) - pnorm(bin[2]) # 0.9544997
pnorm(bin[5]) - pnorm(bin[3]) # 0.6826895
```

Probabilities $68.27 - 95.45 - 99.73$ are called (empirical) rule or 3-sigma rule.

Hints. Similar example for exponential distribution.

```
# probability density curve
x = seq(0, 4, by=0.01)
y = dexp(x)

# tell R we will plot 3 graphs in one row
par(mfrow=c(1,3))

# plot three graphs
for(p in c(0.2, 0.1, 0.05)){
    # plot probability density as a line
    plot(x, y, type='l', xlab=paste0('Area under the curve = ', p), ylab='density')

    # fill area under the curve
    p_quantile = qexp(1-p)
    pol_x = c(p_quantile, 4, seq(4, p_quantile, by=-0.01))
    pol_y = c(0, 0, dexp(seq(4, p_quantile, by=-0.01)))
    polygon(pol_x, pol_y, col = "grey")
}
```

Exercise 4 (Normal distribution). *Let $X \sim N(\mu, \sigma^2)$, where $\mu = 150, \sigma^2 = 6.25$. Calculate $a = \mu - x_{1-\alpha}\sigma$ and $b = \mu + x_{1-\alpha}\sigma$ so that $\Pr(a \le X \le b) = 1 - \alpha$ is equal to 0.90, 0.95 a 0.99. Value $x_{1-\alpha}$ is a quantile of standardized normal distribution, i.e. $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}) = 1 - \alpha, Z \sim N(0, 1)$. Plot density function, fill area between points a and b and label axes x a y as shown in figure 2.*

[4] Note that $\Pr(a < X < b) = \Pr(a \le X \le b)$ because probability of a point (here a and b) is zero for continuous random variables, i.e $\Pr(a) = \Pr(b) = 0$. This does not apply to discrete random variables.

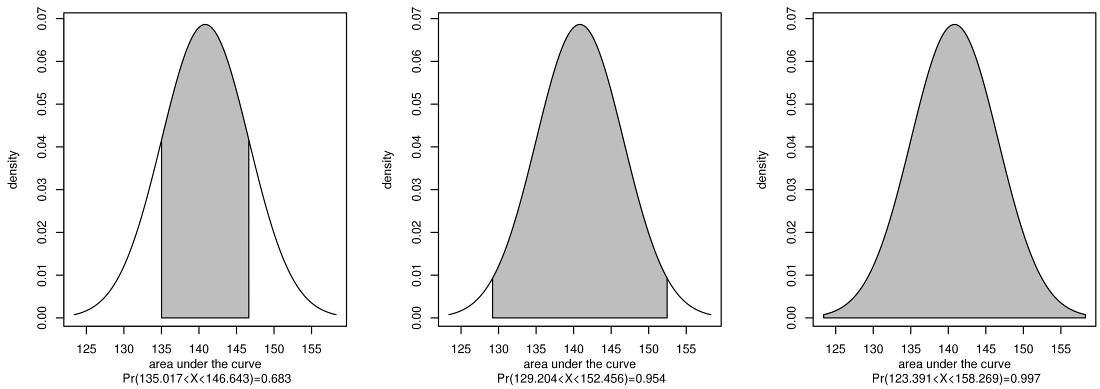

Figure 1: 3-sigma rule; density curve with colored area under this curve between corresponding quantiles on x-axis; volume of the area is equal to the probability of realization of random variable between these quantiles

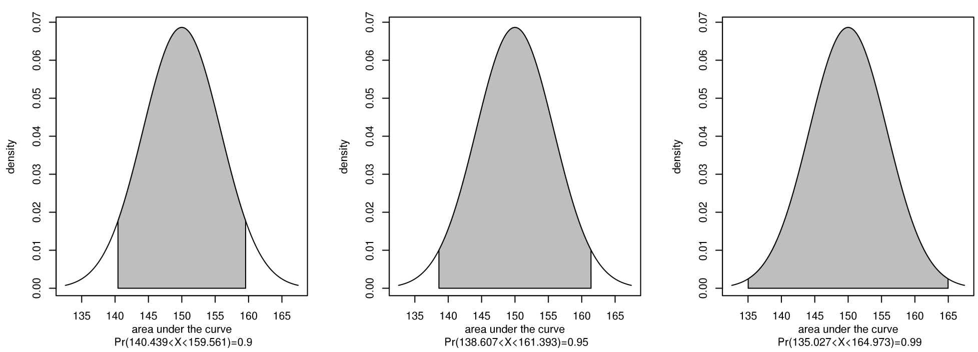

Figure 2: Adjusted 3-sigma rule; density curve with colored area under this curve between corresponding quantiles on x-axis; volume of the area is equal to the probability of realization of random variable between these quantiles

This gives us rule $90 - 95 - 99$ (so called **adjusted 3-sigma rule**). We used property $\Pr\left(u_{\alpha/2} < Z < u_{1-\alpha/2}\right) = \Phi\left(u_{1-\alpha/2}\right) - \Phi\left(u_{\alpha/2}\right) = 1 - \alpha$, where Φ is cumulative distribution function of normal distribution and in general $\alpha \in (0, 1/2)$; in the exercise we used $\alpha = 0.1$, 0.05 a 0.01. See figure 2.

Exercise 5 (Interactive Normal Distribution)**.** *Create an interactive Shiny application showing probability density function of normal distribution $N(\mu, \sigma^2)$, where $\mu = 0$ and σ can be interactively set by user. Fill area between points a and b, where $a = \mu - k\sigma$, $b = \mu + k\sigma$, $k = 1, 2, 3$ and make k interactive too. Finally, add graph title showing probability* $\Pr(a \leq X \leq b)$.

Hints. Create a folder **normalPlot** and copy following files **ui.R** and **server.R** into it. Run it from R console with command `runApp('normalPlot', display.mode="showcase")` from parent directory of **normalPlot** (use function `setwd` to set your working directory if needed). Use code from previous examples in **server.R** to finish the exercise.

Listing 1: **ui.R**

```
# if missing, install with command ‘install.packages(’shiny’)‘
library(shiny)

# Define UI
shinyUI(fluidPage(

  # Application title
  titlePanel("Normal Distribution"),

  # Sidebar with a slider input for the number of bins
  sliderInput("sig",
              "sigma:",
              min = 0.1,
              max = 3,
              value = 1),
  sliderInput("k",
              "k:",
              min = 1,
              max = 3,
              value = 1,
              step = 1),

  plotOutput(’normalPlot’)
))
```

Listing 2: **server.R**

```
library(shiny)

# Define server logic
shinyServer(function(input, output) {

  output$normalPlot <- renderPlot({
    plot(c(1,2,3), c(1,2,3), main=’TODO’)
  })
})
```

Exercise 6 (Interactive Normal Distribution 2). *Extend previous example by allowing user to select if he wants use a and b equal to multiple of σ as in previous example or if he wants to set α, where $\Pr(a \le X \le b) = 1 - \alpha$. Let the user also choose if he wants to display normal distribution or exponential distribution*

Hints. Use control widgets `conditionalPanel` and `selectInput` in file **ui.R**.

Exercise 7 (Binomial Distribution). *Let's assume that number of people preferring treatment A over treatment B follows a binomial distribution with parameters p (probability of success) and N (number of independent trials) denoted $Bin(N, p)$, where $N = 20, p = 0.5$, i.e. people prefer both treatments equally. (a) What is the probability that 16 and more patients will prefer treatment A over treatment B? (b) What is the probability that 16 and more or 4 or less patients will prefer treatment A over treatment B?*

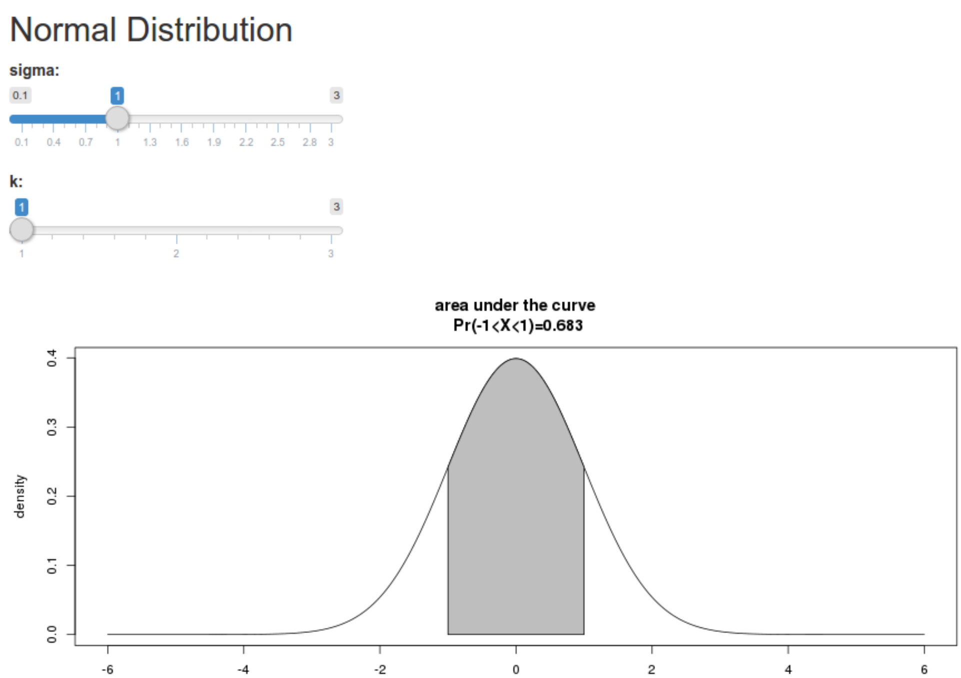

Figure 3: Screenshot from interactive Shiny application with normal distribution.

Solution without R code:
(a) $\Pr(X \geq 16) = 1 - \Pr(X < 16) = 1 - \Pr(X \leq 15) = 1 - \sum_{i:x_i \leq 15} \Pr(X = x_i) = 1 - \sum_{i:x_i \leq 15} \binom{N}{x_i} p^{x_i}(1-p)^{N-x_i} = 1 - \sum_{i:x_i \leq 15} \binom{20}{x_i} 0.5^{x_i}(1-0.5)^{20-x_i} = 0.0059$.
(b) $\Pr(X \leq 4, X \geq 16) = 1 - \sum_{i:x_i \leq 15} \Pr(X = x_i) + \sum_{i:x_i \leq 4} \Pr(X = x_i) = 0.012$. This probability is twice the previous one since $Bin(N, 0.5)$ is symmetric around 0.5.

Hints. pbinom(x, size=..., prob=...) gives you probability $Pr(X \leq x)$. How do you get probability $Pr(X \geq x)$ using this function?

Exercise 8 (Binomial Distribution)**.** *Let's assume that* $\Pr(swirl) = 0.533 = p_1$ *is the probability of having dermatological pattern swirl on right thumb of male population and* $\Pr(other) = 0.467 = p_2$ *is the probability of other patterns on right thumb of the same population. Random variable* X *represents number of swirls and* Y *number of other patterns, where* $X \sim Bin(N, p_1)$ *a* $Y \sim Bin(N, p_2)$*. Calculate*

1. $\Pr(X \leq 120)$ *if* $N = 300$

2. $\Pr(Y \leq 120)$ *if* $N = 300$

Exercise 9 (Normal approximation of binomial distribution)**.** [5]

Let $\Pr(man) = 0.515$ *be a proportion of men in population and* $\Pr(women) = 0.485$ *proportion of women. Let* X *represent number of men and* Y *number of women. Under the assumption of model* $Bin(N, p)$ *calculate*

[5] Approximation means "similar but not exactly equal", i.e. we approximate some distribution with a different one (that has certain advantages over the approximated one) or we approximate data with some distribution (that describes data with help of easily interpretable parameters)

1. $\Pr(X \leq 3)$ *if* $N = 5$

2. $\Pr(X \leq 5)$ *if* $N = 10$

3. $\Pr(X \leq 25)$ *if* $N = 50$*.*

Compare these probabilities with those approximated by normal distribution $N(Np, Npq)$*. Plot density function of normal distribution and superimpose it with probability distribution of binomial distribution. Plot cumulative distribution function of normal distribution and superimpose it with cumulative distribution function for binomial distribution. See figure 4.*

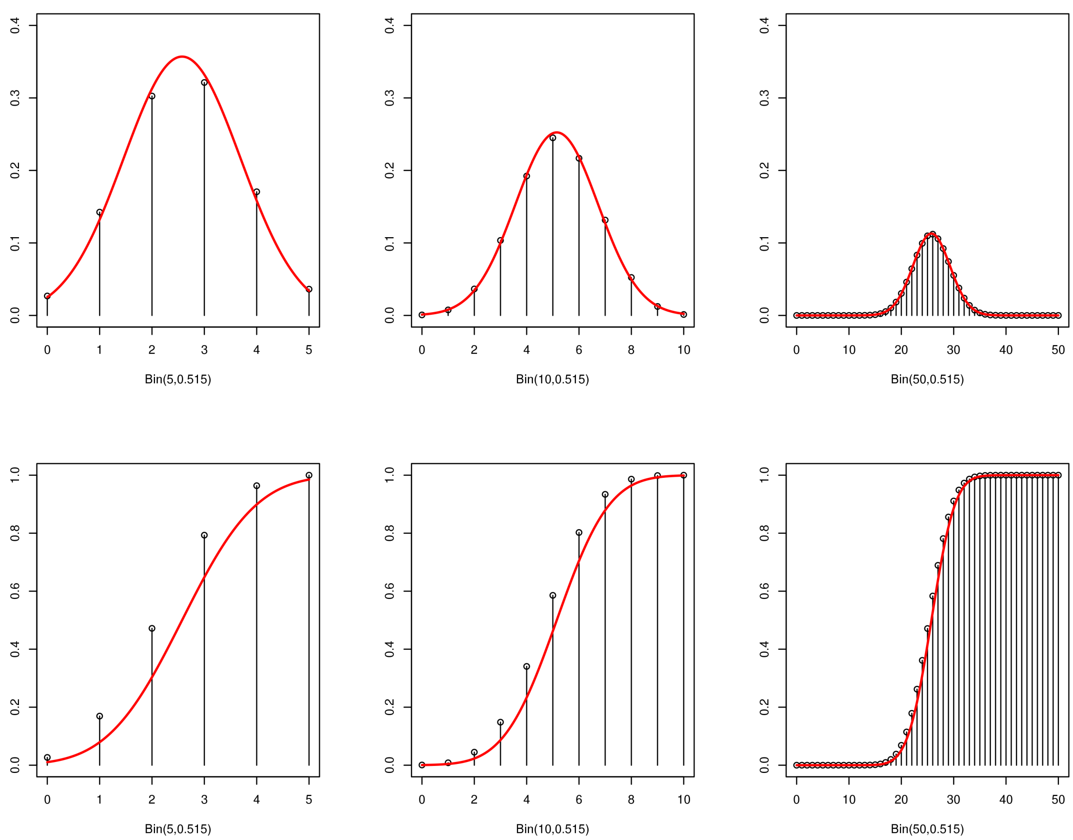

Figure 4: Approximating binomial distribution with normal distribution for $p = 0.515$, $N = 5, 10$ and 50; scatter plot superimposed with density (first row) and cumulative distribution function (second row).

Exercise 10 (Bivariate normal distribution)**.** *Random vector* $(X, Y)^T$ *has bivariate normal distribution*

$$N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \ \textit{where } \boldsymbol{\mu} = (\mu_1, \mu_2)^T \ \textit{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

with density

$$f(x,y) = \frac{1}{2\pi\sqrt{\sigma_1^2\sigma_2^2(1-\rho^2)}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left\{ \frac{(x-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right\} \right\},$$

where $(x,y)^T \in \mathbb{R}^2,\ \mu_j \in \mathbb{R}^1,\ \sigma_j^2 > 0,\ j = 1,2,\ \rho \in \langle -1, 1\rangle$ *are parameters, then* $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$*. Expression in the exponent can be also written as*

$$-\frac{1}{2}\begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix}^T \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} x-\mu_1 \\ y-\mu_2 \end{pmatrix},$$

marginal distributions[6] *are* $X \sim N(\mu_1, \sigma_1^2)$ *and* $Y \sim N(\mu_2, \sigma_2^2)$*,* ρ *is a correlation coefficient*[7]*.*

Exercise 11 (Bivariate normal distribution)**.** *(1) Plot density of bivariate normal distribution* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *using function* ***`image()`*** *and superimpose it with contour graph of the same distribution using function* ***`contour()`****.*
(2) Plot density of bivariate normal distribution $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *using function* ***`persp()`****. Cut density into 12 intervals, where values in these intervals correspond to colors* ***`terrain.colors(12)`****.*
Use following parameters
(a) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0$*;*
(b) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1,\ \rho = 0.5$*;*
(c) $\mu_1 = 0, \mu_2 = 0,\ \sigma_1 = 1, \sigma_2 = 1.2,\ \rho = 0.5$*.*
See figure 5 for correct solution.

Hints.

1. create a function `bnorm(x,y,mu1,mu2,sigma1,sigma2,corr)` returning density of bivariate normal distribution

2. create vectors x and y with values from -3 to 3 and length n and their cartesian product $(x_i, y_j),\ i = 1, \ldots, n,\ j = 1, \ldots, n$ (represented as either $n^2 \times 2$ matrix or two vectors of length n^2)

3. apply your function `bnorm` on the cartesian and reshape it to a matrix Z with dimensions $n \times n$

4. use x, y and Z in `image` function to plot density

You don't have to use Greek letters in axis labels (use *mu_1* instead of μ_1), but if you want to, look up `expression` function. For coloring of `persp` look at the last example in `persp` documentation and modify it to your needs.

[6] Marginal distribution is a distribution of marginal random variable. Marginal distribution of multivariate normal distribution is again normal, which is a very useful property.

[7] From this example it is clear that to sufficiently describe bivariate normal distribution we need 5 parameters, i.e. mean and variance for marginal distributions of random variables X and Y and correlation coefficient $\rho = \rho(X,Y)$ describing the strength of linear relationship of X and Y.

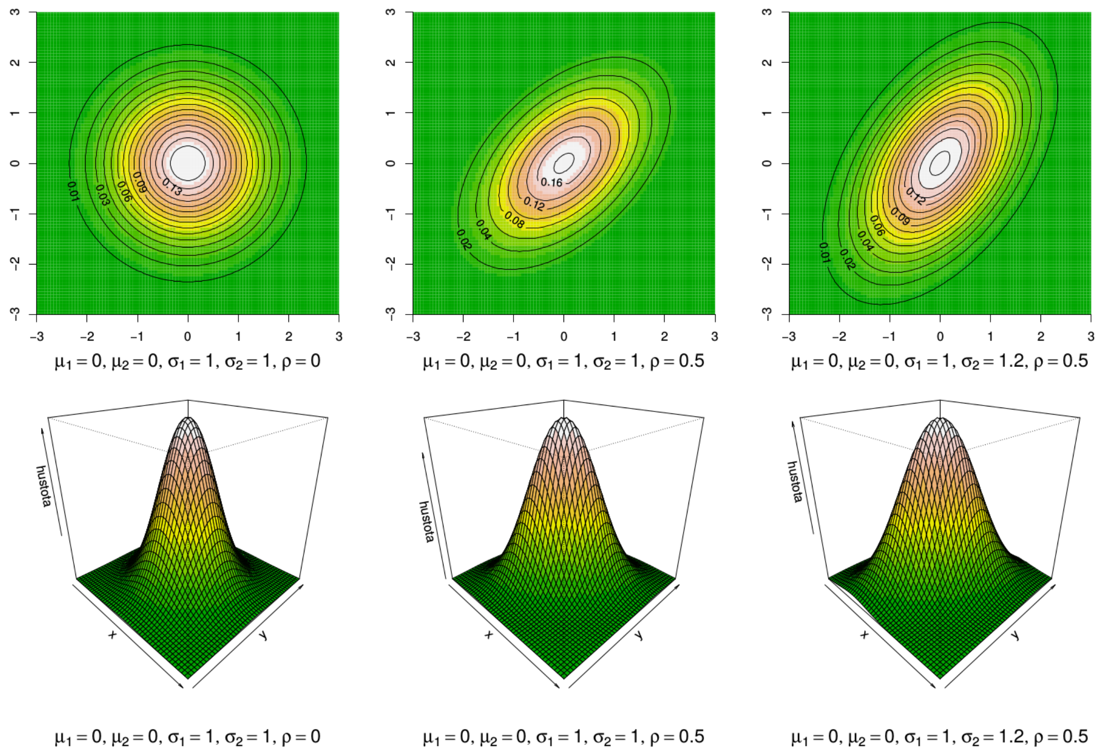

Figure 5: Density of bivariate normal distribution with different parameters (first row – contour graph, second row – perpective 3D graph displayed as surface); the larger absolute value of ρ, the more are contours different from circles (they are changing into ellipses); as the difference between σ_1 and σ_2 is getting larger, we say that difference in variability of X_1 and X_2 is increasing

Exercise 12 (Kernel density estimation)**.** *Simulate N samples from given distribution and plot them as + symbols on x-axis together with their kernel density estimation. Add theoretical normal distribution (resp. exponential distribution) with parameters estimated from data. Do it for (a) $N = 50$ (b) $N = 1000$ and following distributions*

1. $X \sim N(0, 1)$
2. $X \sim pN(-3, 1) + (1 - p)N(3, 1)$ *is a mixture of normals, where* $p = 0.3$
3. $X \sim Exp(\lambda)$, *where* $\lambda = 3$. *Plot theoretical exponential distribution with estimated parameter* $\hat{\lambda} = \frac{1}{\bar{X}} = \frac{N}{\sum_{i=1}^{N} X_i}$.

See figure 7 for correct solution.

Hints. To get the same results as in the figure, call `set.seed(5)` at the beginning of your script.
Useful functions: `density(..., from=, to=, n=)`, `plot(..., lty=2)`, `points(..., pch=3)`

Exercise 13 (Bivariate normal distribution)**.** *Simulating pseudorandom numbers from* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *can be done with R in several ways:*

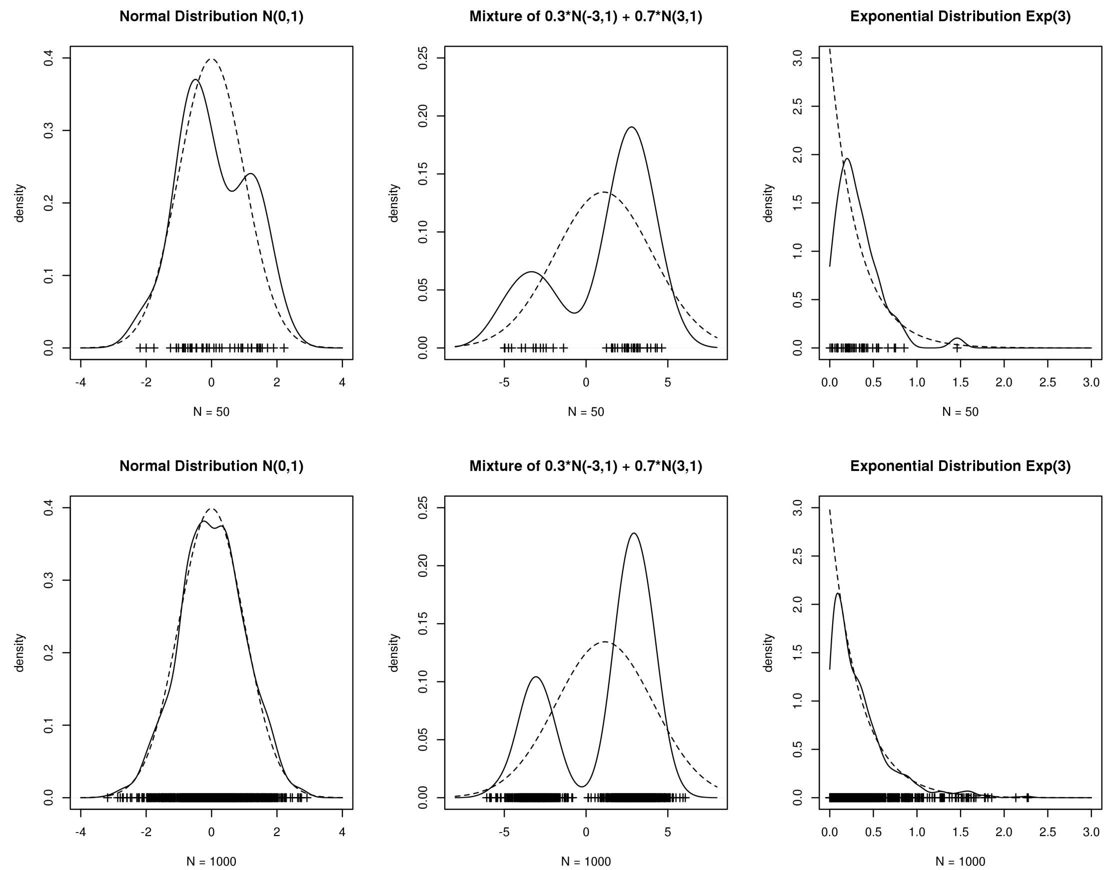

Figure 6: Kernel density estimation of several distributions (first row $n = 50$; second row $n = 1000$)

1) library ***`library(MASS)`*** *and function* ***`mvrnorm()`****;*
2) library ***`library(mvtnorm)`*** *and function* ***`rmvnorm()`****;*
3) function ***`rnorm()`*** *and this algorithm – let $X_1 \sim N(0,1)$ and $X_2 \sim N(0,1)$; then $(Y_1, Y_2) \sim N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ is vector of means and σ_1^2, σ_2^2 and ρ are parameters of covariance matrix $\boldsymbol{\Sigma}$, where the strength of linear relationship Y_1 and Y_2 is given by magnitude and sign of ρ; $Y_1 = \sigma_1 X_1 + \mu_1$ a $Y_2 = \sigma_2(\rho X_1 + \sqrt{1-\rho^2} X_2) + \mu_2$.*

Simulate pseudorandom numbers Y_1 and Y_2 from $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the first method. Calculate kernel density estimation of $(Y_1, Y_2)^T$ using function ***`kde2d()`****. Plot it using function* ***`image()`*** *and superimpose it with contour graph of bivariate normal density of $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using function* ***`contour()`****. Use following parameter in simulations*
(a) $\mu_1 = 0, \mu_2 = 0$, $\sigma_1 = 1, \sigma_2 = 1$, $\rho = 0$; (1) $n = 50$ a (2) $n = 1000$;
(b) $\mu_1 = 0, \mu_2 = 0$, $\sigma_1 = 1, \sigma_2 = 1$, $\rho = 0.5$; (1) $n = 50$ a (2) $n = 1000$;
(c) $\mu_1 = 0, \mu_2 = 0$, $\sigma_1 = 1, \sigma_2 = 1.2$, $\rho = 0.5$; (1) $n = 50$ a (2) $n = 1000$.
See figure 7 for correct solution.

Hints. `kde2d(..., n=100, lims=c(-3, 3, -3, 3))`

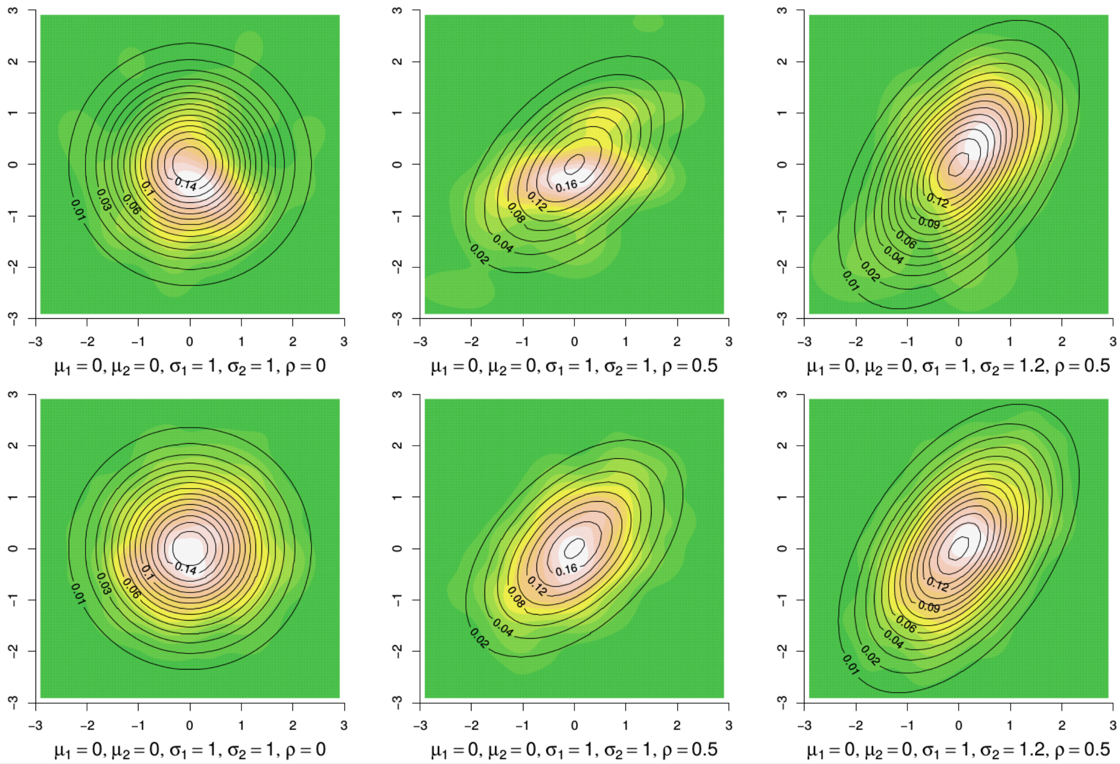

Figure 7: Density of bivariate normal distribution (first row $n = 50$; second row $n = 1000$)

Exercise 14 (Mixture of two normal distributions)**.** *Simulate pseudorandom numbers (1) from mixture of normal distributions* $pN_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + (1-p)N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ *where* $\boldsymbol{\theta}_1 = (\mu_{11}, \mu_{12}, \sigma_{11}^2, \sigma_{12}^2, \rho_1, \mu_{21}, \mu_{22}, \sigma_{21}^2, \sigma_{22}^2, \rho_2)^T$ *and (2) from bivaraite normal distribution* $N_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *where parameters represent combined vector of means a combined covariance matrix. i.e.* $\boldsymbol{\theta}_2 = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)^T$. *For (1) calculate kernel density estimation of* $(X, Y)^T$ *using function* ***kde2d()****.*

(a) Plot theoretical density (2) using function ***image()*** *and superimpose it with contour graph of theoretical density (2) using function* ***contour()****.*

(b) Plot theoretical density (1) using function ***image()*** *and superimpose it with contour graph of theoretical density (1) using function* ***contour()****.*

(c) Plot kernel density estimation of (1) using function ***image()*** *and superimpose it with contour graph of theoretical density (1) using function* ***contour()****.*

Cut density into 12 intervals, where values in these intervals correspond to colors ***terrain.colors(12)****. For simulation use these parameters* $\boldsymbol{\theta}_1 = (-1.2, -1.2, 1, 1, 0, 1, 1, 1, 1, 0)^T$ *and* $n = 100$.

Use parameters from simulation for case (1), i.e. $\widehat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1 = (\mu_{11}, \mu_{12}, \sigma_{11}^2, \sigma_{12}^2, \rho_1, \mu_{21}, \mu_{22}, \sigma_{21}^2, \sigma_{22}^2, \rho_2)^T = (-1.2, -1.2, 1, 1, 0, 1, 1, 1, 1, 0)^T$. *For case (2) estimate parameters from simulated data* $\widehat{\boldsymbol{\theta}}_2 = (\widehat{\mu}_1, \widehat{\mu}_2, \widehat{\sigma}_1^2, \widehat{\sigma}_2^2, \widehat{\rho})^T$.
See figure 8 for correct solution.

Hints. For parameter estimations in case (2) use functions `mean`, `std` and `cor`.

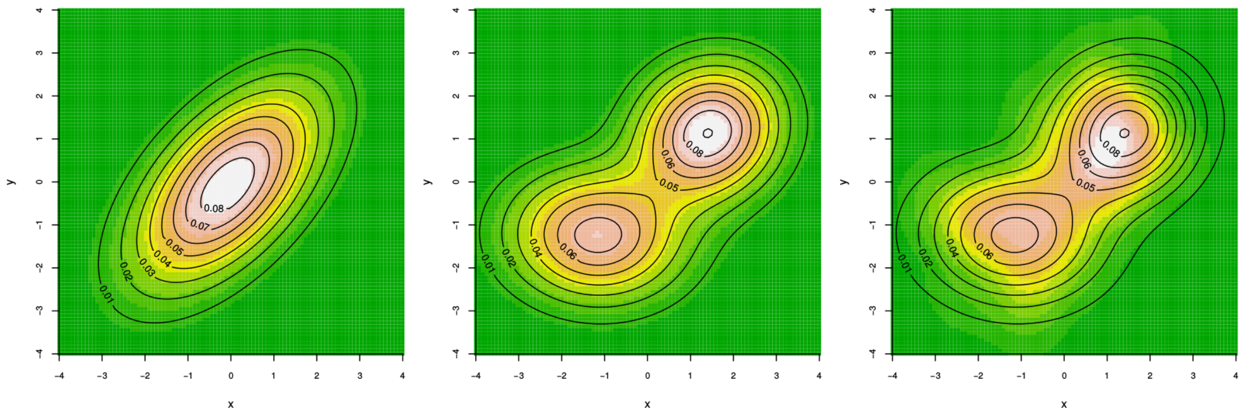

Figure 8: Combined density of bivariate normal distribution (left), mixture density of two bivariate normal distributions (middle) and kernel density estimation superimposed by mixture density of two bivariate normal distributions (right)

Exercise 15 (Poisson distribution)**.** *We have a data (Greenwood and Yule 1920) with number of injuries of factory workers in the following table*

n	0	1	2	3	4	≥ 5
m_n	447	132	42	21	3	2

where n is number of injuries and m_n number of workers with n injuries.

Calculate expected number of worker injuries under the assumption that random variable X representing injuries has Poisson distribution with parameter $\lambda = \frac{\sum_n nm_n}{\sum_n m_n}$, i.e. $X \sim Poiss(\lambda)$. Create a table with m_n and expected m_n and display it (figure 9).

Hints. Create a dataframe with two columns and proper row names, then print it. `data.frame(mn=..., expected_mn=..., row.names=...)`.

```
               0   1  2  3 4 >=5
mn           447 132 42 21 3   2
expected mn  406 189 44  7 1   0
```

Figure 9: Observed and expected frequencies of Poisson distribution.

Exercise 16 (Binomial distribution)**.** *In a study from 1889 based on medical records in Saxony professor Geissler (1889) recorded distribution of boys in families. The study included $M = 6115$ families with $N = 12$ children with following number of boys (n stands for number of boys and m_n number of families with n boys)*

Calculate expected m_n under the assumption that number of boys X in families follow binomial distribution $\pi = \frac{\sum_{n=0}^{N} nm_n}{NM}$ and $N = 12$ (i.e. $X \sim Bin(N, \pi)$).

Compare expected and observed frequencies - do you see any difference? Display m_n and expected m_n in a table and visualize both observed and expected frequencies in one graph (see

n	0	1	2	3	4	5	6	7	8	9	10	11	12
m_n	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

figure 10 or use your own imagination). Calculate probability that family will have ≥ 4 *and* ≤ 6 *boys from theoretical distribution (i.e.* $Pr(4 \geq X \leq 6)$*).*

Hints. Add legend with `legend('topleft', c('observed', 'expected'), col=c('black', 'red'), lty=1, bty='n')`. For stairs graph use argument `type='S'` or `type='s'`

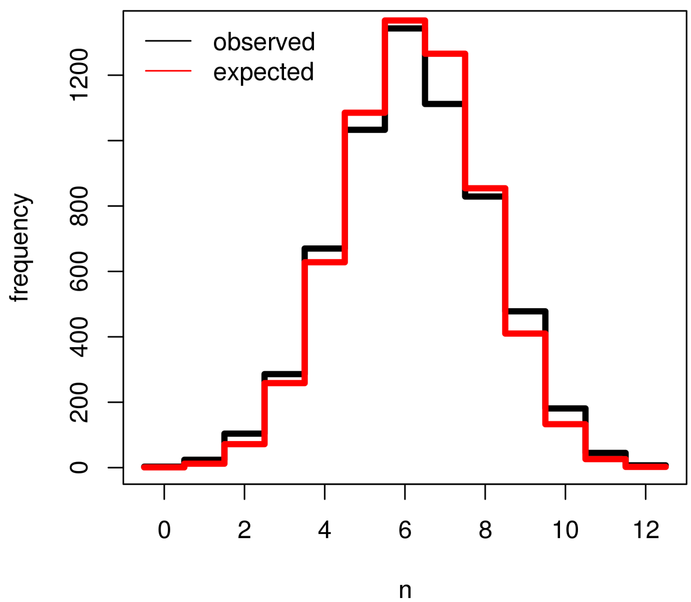

Figure 10: Observed and expected frequencies of Binomial distribution

Exercise 17 (Approximating Binomial distribution by Poisson)**.** *If each of 50 million people in Italy drives a car next week (independently), then probability of dying in an accident will be 0.000002, where number of fatalities has binomial distribution* $Bin(50mil, 0.000002)$*. Approximate this Binomial distribution with Poisson that has the same mean. Plot both probabilities for values* $\{50, 51, \ldots, 150\}$*.*

Exercise 18 (Moments of Bernoulli distribution)**.** *Random variable* X *has Bernoulli distribution (*$X \sim Ber(p)$*) if* $Pr(X = 1) = 1 - Pr(X = 0) = 1 - q = p$*. Derive its expected value* $\mathrm{E}X = \sum_k kPr(X = k)$ *and variance* $\mathrm{Var}X = \mathrm{E}[(X - \mathrm{E}X)^2] = \mathrm{E}[X^2] - (\mathrm{E}X)^2$

Exercise 19 (Moments of Binomial distribution)**.** *Random variable* X *has Binomial distribution (*$X \sim Bin(N, p)$*) if*

$$Pr(X = k) = \binom{N}{k} p^k (1-p)^{N-k} = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}$$

for $k \in \{0, \dots, N\}$. Derive its expected value $\mathrm{E}X = \sum_k kPr(X = k)$ *and variance* $\mathrm{Var}X = \mathrm{E}[(X - \mathrm{E}X)^2] = \mathrm{E}[X^2] - (\mathrm{E}X)^2$

Hints. The hard way involves Binomial theorem, for the easy way you need to realize relationship between Binomial and Bernoulli random variable and use results from previous exercise.

Exercise 20 (Contingency table)**.** *356 people have been polled on their smoking status (Smoke) and their socioeconomic status (SES). For each person it was determined whether or not they are current smokers, former smokers, or have never smoked. Also, for each person their socioeconomic status was determined (low, middle, or high). The data file* `smoker.csv` *(available from study materials) contains only two columns - Smoke and SES. Load this data into R and create contingency table with* `table()` *function.*

These observations (after converting to probability) form joint distribution $Pr(X = x, Y = y)$, where X is a discrete random variable for smoker and Y for SES. Calculate marginal distributions $Pr(X = x) = \sum_y Pr(X = x, Y = y)$, $x \in \{current, former, never\}$ and $Pr(Y = y)$, $y \in \{high, low, middle\}$. Also calculate and compare $Pr(X = x|Y = high)$ with $Pr(X = x|Y = low)$ and $Pr(Y = y|X = smoker)$.

Create a table with expected frequencies using assumption that X and Y are independent (figure 11). Visualize both tables (observed and expected) with `mosaicplot` *function (figure 12).*

Hints. To calculate marginal probabilities use either special `table` functions `margin.table` and `prop.table` or simply `rowSums` and `colSums`.

```
        High Low Middle
current   51  43     22
former    92  28     21
never     68  22      9
```

```
            High      Low   Middle
current 68.75281 30.30337 16.94382
former  83.57022 36.83427 20.59551
never   58.67697 25.86236 14.46067
```

Figure 11: Observed and expected frequencies for contingency table

Exercise 21 (Number of contingency tables)**.** *How many contingency tables of size 4×2 there are for $N = 20$?*

Exercise 22 (Multinomial distribution)**.** *We have following random variables (1) socioeconomic status (high - H, low - Lo), (2) political affiliation (democrat - D, republican - R) and (3) political philosophy (libral - Li, convservatism - C). Let's denote their interactions like this: X_1 (H-D-Li), X_2 (H-D-C), X_3 (H-R-Li), X_4 (H-R-C), X_5 (Lo-D-Li), X_6 (Lo-D-C), X_7 (Lo-R-Li) a X_8 (Lo-R-C). We have random sample of size $N = 50$. Probabilities p_j are in the following table*

	D-Li	*D-C*	*R-Li*	*R-C*	*all*
H	0.12	0.12	0.04	0.12	0.4
Lo	0.18	0.18	0.06	0.18	0.6
all	0.30	0.30	0.10	0.30	1.0

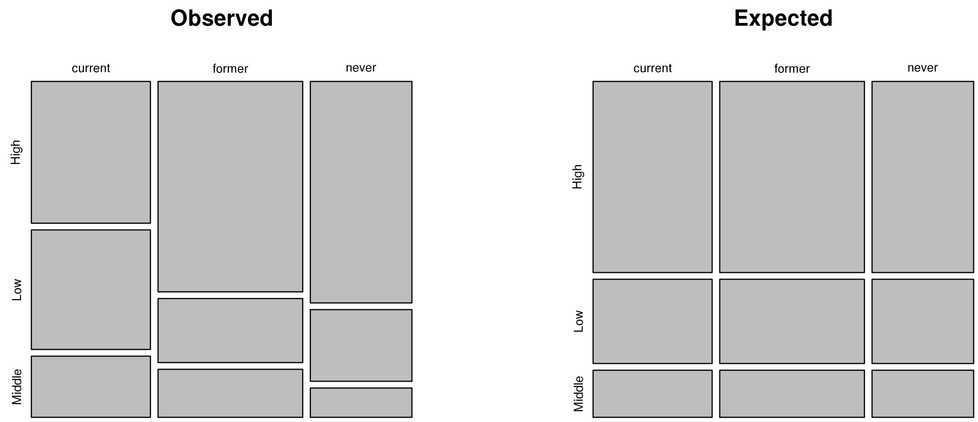

Figure 12: Mosaic plot of contingency table

Calculate $\mathrm{Var}[X_1]$, $\mathrm{Var}[X_3]$, $\mathrm{Cov}\,[X_1, X_3]$, *Cor* $[X_1, X_3]$ *and expected frequencies* $Np_j, j = 1, 2, \ldots, 8$.

Hints. $X = (X_1, X_2, \ldots, X_8) \sim Mult(N, \mathbf{p})$, where $N = 50$, $\mathbf{p} = (p_1, p_2, \ldots, p_8)^T$, we know that $X_j \sim Bin(N, p_j)$, p_j are given by table above $j = 1, 2, \ldots, 8$.

Exercise 23 (Product-multinomial distribution)**.** *Let's have the same probabilities as in previous exercise, but now with two separate random samples - first with size* $N_1 = 30$ *from group H and the second with size* $N_2 = 20$ *from group Lo. Denote interactions of random variables* $X_{11} = X_{1|1}$ *(H-D-Li),* $X_{12} = X_{2|1}$ *(H-D-C),* $X_{13} = X_{3|1}$ *(H-R-Li),* $X_{14} = X_{4|1}$ *(H-R-C),* $X_{21} = X_{1|2}$ *(Lo-D-Li),* $X_{22} = X_{2|2}$ *(Lo-D-C),* $X_{23} = X_{3|2}$ *(Lo-R-Li) and* $X_{24} = X_{4|2}$ *(Lo-R-C), where* $\mathbf{X}_1 = (X_{11}, X_{12}, X_{13}, X_{14})^T$ *and* $\mathbf{X}_2 = (X_{21}, X_{22}, X_{23}, X_{24})^T$*. Then* $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ *has product-multinomial distribution with* $K = 2$, $N_1 = 30, J_1 = 4$, $N_2 = 20, J_2 = 4$*. Notation* $X_{j|k}$*, where* $j = 1, 2, 3, 4$ *and* $k = 1, 2$ *highlights the fact, that distribution is conditioned by socioeconomic status (high – H, low – Lo), i.e. distribution in table columns is conditioned by its row. Realizations* $X_{j|k}$ *are denoted as* $n_{j|k} = n_{kj}$*, similarly probabilities* $X_{j|k} = X_{kj}$ *as* $p_{j|k} = p_{kj}$*. Calculate conditional probabilities* $p_{j|k}$*, expected frequencies* $N_k p_{kj}$*,* $\mathrm{Var}[X_{13}]$*,* $\mathrm{Cov}\,[X_{21}, X_{23}]$ *and Cor* $[X_{11}, X_{23}]$*.*

Exercise 24 (Binomial distribution, simulation study)**.** *Generate* $M = 1000$ *pseudo-random numbers from distribution* $Bin(5, 0.5)$*. Create a table with observed and theoretical probabilities (for* $n = 0, 1, \ldots, 5$*).*

r	*0*	*1*	*2*	*3*	*4*	*5*
observed probabilities	*0.031*	*0.158*	*0.302*	*0.324*	*0.161*	*0.024*
theoretical probabilities	*0.031*	*0.156*	*0.312*	*0.312*	*0.156*	*0.031*

Superimpose histogram of simulated numbers with theoretical density function. See figure 13.

Hints. For histogram plot use `breaks` and `probability` arguments.

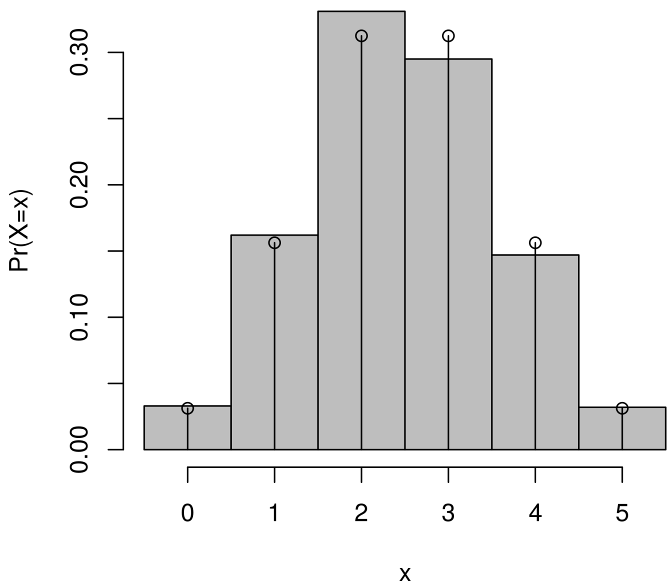

Figure 13: Histogram superimposed with theoretical distribution of $Bin(5, 0.5)$

Exercise 25 (Normal distribution, simulation study)**.** *If random variable* $X \sim N(150, 6.25)$*, what distribution has its arithmetic average* $\overline{X}_n$*? Verify your result using simulations for* $n = 30$*. Make 5000 simulations (each simulation involves generating 30 realizations of* X*) and for each simulation calculate arithmetic average* $\overline{x}_m$*,* $m = 1, 2, \dots, M$*, where* $M = 5000$*. Superimpose their histogram (normalized) with theoretical density function for* $\overline{X}_n$ *you derived. Calculate* $\Pr(\overline{X}_n > 151)$ *from simulated data and compare this result with theoretical (expected) probability. See figure 14.*

Exercise 26 (Normal distribution, simulation study)**.** *Let* $X \sim N(\mu_1, \sigma_1^2)$ *a* $Y \sim N(\mu_2, \sigma_2^2)$*. Then* $\overline{X}_{n_1} - \overline{Y}_{n_2} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$*. Use* $\mu_1 = 100, \sigma_1 = 10, \mu_2 = 50, \sigma_2 = 9$ *and (a)* $n_1 = 4, n_2 = 5$*, (b)* $n_1 = 100, n_2 = 81$*. Make* $M = 1000$ *simulations and for each calculate difference* $\overline{x}_m - \overline{y}_m, m = 1, 2, \dots, M$*. Superimpose histogram (normalized) of the differences with theoretical density function of difference* $\overline{X}_{n_1} - \overline{Y}_{n_2}$*. For both cases (a) and (b) calculate* $\Pr(\overline{X}_{n_1} - \overline{Y}_{n_2}) < 52$ *from simulated data and compare this result with theoretical probability.*

Exercise 27 (Maximum-likelihood, Poisson distribution)**.** *Let* $X \sim Poiss(\lambda)$*. Simulate realizations* $x_1, \dots, x_n$*. Use* $n = 100$ *and* $\lambda = 4$*. Plot log-likelihood function of Poisson distribution*

$$l(\lambda|\mathbf{x}) = \sum_{i=1}^{n} x_i \ln \lambda - n\lambda.$$

for $\lambda \in [2, 6]$*. Find maximum likelihood estimate of* λ *and mark it in graph with dashed line. See figure 15.*

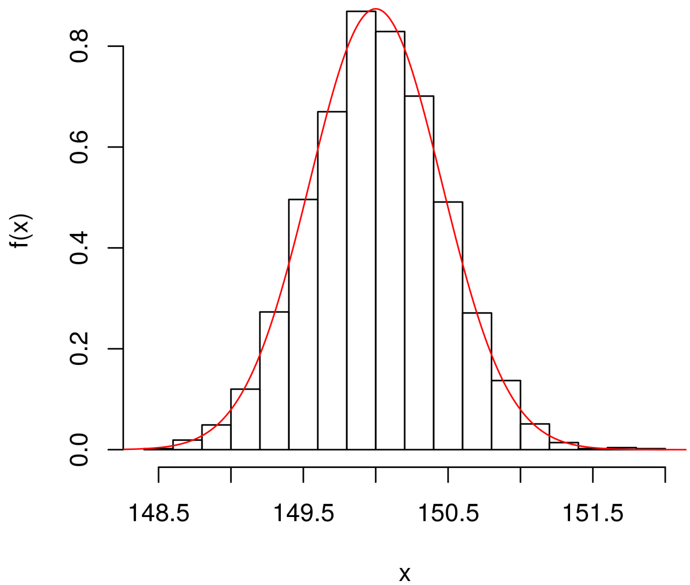

Figure 14: Histogram superimposed with theoretical distribution of $N(150, 6.25/n)$

Hints. Start with writing a function `loglikehood = function(x, lambda)` that calculates log-likelihood for **single** λ. Then in a for loop or using `sapply` function calculate log-likelihood for $\lambda \in [2, 6]$.

To find maximum of a function, either use formula for MLE from the lecture or find it empirically from vector of log-likelihoods using function `which.max(lambdas)`. To plot a vertical line use `abline(v=..., lty='dashed')`.

Exercise 28 (Maximum-likelihood, Poisson distribution). *Use data from exercise 15 for log-likelihood function from the previous exercise and plot it. Note that the log-likelihood function will have to be slightly modified to work with* ***counts*** *of realizations from exercise 15. See figure 15.*

Exercise 29 (Maximum-likelihood, Binomial distribution). *Let $X \sim Bin(N, p)$ and x its realization*[8]*. Formulate its likelihood function $L(p|x)$ and log-likelihood function $l(p|x)$. Use log-likelihood function to derive maximum-likelihood estimate $\hat{p} = \underset{p}{\operatorname{argmax}}\, l(p|x)$ of parameter p and Fisher information $\mathcal{I}(\hat{p}) = -\frac{\partial^2}{\partial p^2} l(p|x)|_{p=\hat{p}}$.*

Exercise 30 (Quadratic approximation of log-likelihood function). *Let $X \sim Bin(N, p)$ and x its realization. Plot scaled log-likelihood function of binomial distribution with p on x-axis*

[8] We work with just **one** realization, so $L(p|x) = \prod_{i=1}^{1} f(x|p)$. You would get the same results if you worked with N realizations with Bernoulli distribution $Ber(p)$.

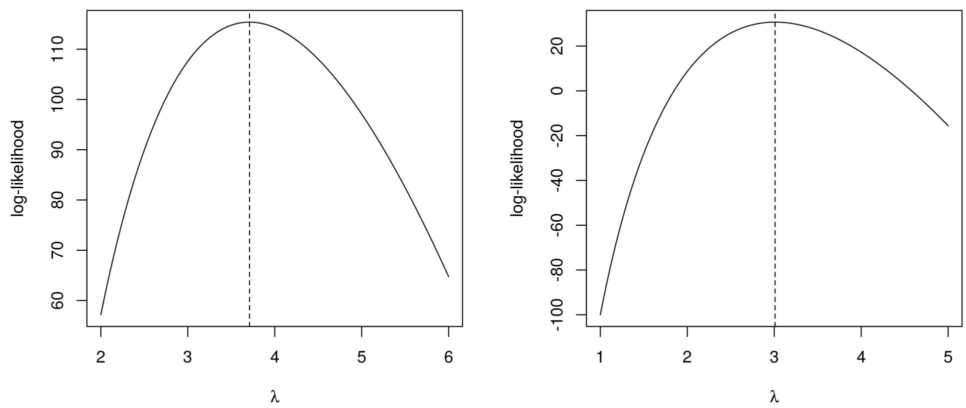

Figure 15: Left: Log-likelihood function for simulated data from $Poiss(\lambda = 4)$. Right: Log-likelihood function for number of injuries of factory workers.

and scaled log-likelihood $l^*(p|\mathbf{x}) = l(p|\mathbf{x}) - \max(l(p|\mathbf{x}))$ *on y-axis. Compare* $l^*(p|\mathbf{x})$ *with its quadratic approximation calculated using Taylor approximation* $\ln(\frac{L(p|\mathbf{x})}{L(\widehat{p}|\mathbf{x})}) \approx -\frac{1}{2}\mathcal{I}(\widehat{p})(p-\widehat{p})^2$*. Use (a)* $x = 8, N = 10$*, (b)* $x = 80, N = 100$ *and (c)* $n = 800, N = 1000$ *with reasonable ranges for both axes. See figure 16.*

Hints. Use `lines(..., lty="dashed")` for the dashed line.

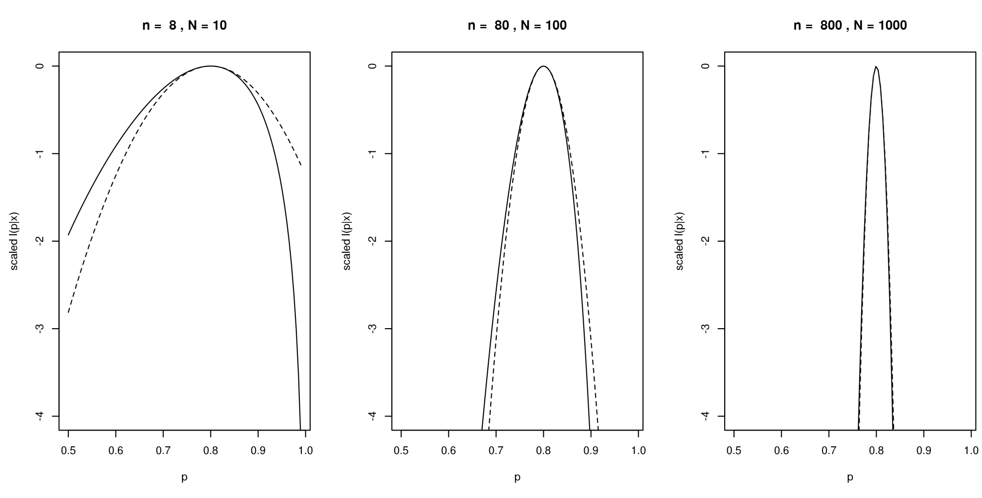

Figure 16: Compare scaled log-likelihood function (full line) with its quadratic approximation (dashed line).

2 Location characteristics

Realizations will be denoted as $x_1, x_2, \ldots, x_n$, **sorted realizations** as $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Then we define following estimations of location characteristics (sample location characteristics):

- **sample minimum** $X_{\min}$, with realization $x_{\min} = x_{(1)}$;
- **sample maximum** $X_{\max}$, with realization $x_{\max} = x_{(n)}$;
- **sample (arithmetic) mean** $\overline{X}$, with realization $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\sum_{j=1}^{n_j} x_j f_j, n_j \leqslant n$, where f_j are frequencies (counts) of x_j and $n = \sum_j f_j$;
- **sample mode** X_{mod}, with realization x_{mod} is the most common value (in case of discrete variable it is value x in which probability function has its maximum; in case of continuous variable it is value x in which density function has its maximum);
- **sample median** $\widetilde{X}$(robust estimation of location), with realization

$$\widetilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even;} \end{cases}$$

  distribution is *symmetric*, if $\overline{x} = \widetilde{x} = x_{\text{mod}}$, distribution is *positively skewed* (right), if $\overline{x} > \widetilde{x} > x_{\text{mod}}$ and distribution is *negatively skewed* (left), if $\overline{x} < \widetilde{x} < x_{\text{mod}}$;

- **sample quartiles** there are three
  - **sample first (lower) quartile** Q_1, with realization $\widetilde{x}_{0.25}$ is a value that splits off the lowest 25% of data from the highest 75%,

  $$\Pr\left[x_{\min}, \widetilde{x}_{0.25}\right] = \Pr\left[X \leq \widetilde{x}_{0.25}\right] = \frac{1}{4}, \Pr\left[\widetilde{x}_{0.25}, x_{\max}\right] = \Pr\left[X \geq \widetilde{x}_{0.25}\right] = \frac{3}{4};$$

  - **sample second quartile** (median) Q_2, with realization $\widetilde{x}_{0.5} = \widetilde{x}$ is a value that splits off the lowest 50% of data from the highest 50%,,

  $$\Pr\left[x_{\min}, \widetilde{x}_{0.5}\right] = \Pr\left[X \leq \widetilde{x}_{0.5}\right] = \frac{1}{2}, \Pr\left[\widetilde{x}_{0.5}, x_{\max}\right] = \Pr\left[X \geq \widetilde{x}_{0.5}\right] = \frac{1}{2};$$

  - **sample third (upper) quartile** Q_3, with realization $\widetilde{x}_{0.75}$ is a value that splits off the lowest 75% of data from the highest 25%,,

  $$\Pr\left[x_{\min}, \widetilde{x}_{0.75}\right] = \Pr\left[X \leq \widetilde{x}_{0.75}\right] = \frac{3}{4}, \Pr\left[\widetilde{x}_{0.75}, x_{\max}\right] = \Pr\left[X \geq \widetilde{x}_{0.75}\right] = \frac{1}{4};$$

- **sample deciles** $\widetilde{X}_k$, with realizations $\widetilde{x}_k$ splits data to ten buckets, i.e. $k/10$ of data are lower than a decile and $(10-k)/10$ are higher, where $k \in \{0, 1, \ldots, 10\}$;

- **sample percentile** $\widetilde{X}_p$ (read as $100p$-percentile), with realization $\widetilde{x}_p$ defined as

$$\widetilde{x}_p = \begin{cases} x_{(k+1)} & \text{for } k \neq np, \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right) & \text{for } k = np, \end{cases}$$

  where $k = \lfloor np \rfloor$, which is floor of np;

- **sample five-number summary** $(X_{\min}, Q_1, Q_2, Q_3, X_{\max})^T$, with realizations $(x_{\min}, \widetilde{x}_{0.25}, \widetilde{x}_{0.50}, \widetilde{x}_{0.75}, x_{\max})^T$.

Robust location characteristics (resistant to outliers) are

- **sample γ-truncated arithmetic average** $\overline{X}_g$, with realization $\overline{x}_g$ that is calculated as

$$\overline{x}_g = \frac{1}{n-2g}\left(x_{(g+1)} + x_{(g+2)} + \ldots + x_{(n-g)}\right),$$

  where $g = \{\gamma n\}, g = \lfloor \gamma n \rfloor, \gamma = 0.1, 0.2$. More than $\gamma 100$ % observations must be replaced for the γ-truncated average to become large or small relative to the original [[9] *breakdown point* $\overline{x}_g$ is therefore γ],

- **sample γ-winsorized arithmetic average** $\overline{X}_w$, with realization $\overline{x}_w$ is defined as

$$\overline{x}_w = \frac{1}{n}\left((g+1)x_{(g+1)} + x_{(g+2)} + \ldots + (g+1)x_{(n-g)}\right).$$

  More than $\gamma 100$ % must be replaced for the γ-winsorized average to become large or small relative to the original [*breakdown point* $\overline{x}_w$ is therefore γ].

Exercise 31 (height of 10-year old girls)**.** *Let's have $n = 12$ heights (in cm) of randomly sampled 10-year old girls sorted by height (**order** denoted as r_i for $x_{(i)}$; in case the values are equal, r_i is calculated as average of their order numbers).*

Table 1: Sorted realizations x_i and their order r_i for heights of 10-year old girls

i	1	2	3	4	5	6	7	8	9	10	11	12
$x_{(i)}$	131	132	135	141	141	141	141	142	143	146	146	151
r_i	1	2	3	5.5	5.5	5.5	5.5	8	9	10.5	10.5	12

Then $\overline{x} \doteq 140.83$, $\widetilde{x} = \frac{1}{2}\left(x_{(6)} + x_{(7)}\right) = 141$, $\widetilde{x}_{0.25} = \frac{1}{2}\left(x_{(3)} + x_{(4)}\right) = 138$, *where* $k = \lfloor 12 \times 0.25 \rfloor = 3$, $Q_3 = \widetilde{x}_{0.75} = \frac{1}{2}\left(x_{(9)} + x_{(10)}\right) = 144.5$, *where* $k = \lfloor 12 \times 0.75 \rfloor = 9$.
Write functions for calculation of all location characteristics. Verify your functions on characteristics above. Don't use built-in functions for location characteristics such as `mean`, `quantile`, etc. Use $\gamma = 0.1$ for truncated and winsorized arithmetic averages.

[9] *Breakdown point* represents number of observations we need to significantly change value of location characteristics. For γ-truncated and γ-winsorized arithmetic average it is γn observations, for median $n/2$ observations and for simple arithmetic average changing just one observation is enough (that's the reason we say that arithmetic average is very sensitive to outliers).

3 Spread (variability) characteristics

Then we define following estimations of spread (variability) characteristics (sample spread characteristics):

- **sample variance** S^2, with realization

  $$s^2 = s_{n-1}^2 = s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2;$$

  under linear transformation sample variance changes like this[10]

  $$s_y^2 = s_{a+bx}^2 = b^2 s_x^2,$$

  i.e.

  $$\begin{aligned} s_y^2 &= s_{a+bx}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(a + bx_i - \overline{a+bx}\right)^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(a + bx_i - (a + b\overline{x}))^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(b(x_i - \overline{x}))^2 = b^2 s_x^2; \end{aligned}$$

- **sample standard deviation** S, with realization

  $$s = s_{n-1} = s_x = \sqrt{s_x^2};$$

  under linear transformation standard deviation changes like this

  $$s_y = s_{a+bx} = |b|\, s_x,$$

- **coefficient of variation** V_k, with realization v_k represents normalized form of standard deviation (inversion of *signal-to-noise ratio*; fraction of variability to mean)

  $$v_k = \frac{s_x}{\overline{x}};$$

  it is usually denoted in percentage points, i.e. $100 \times (s_x/\overline{x})$ % and can be used only for realizations with positive values;

- **sample variance of arithmetic average** $S_{\overline{X}}^2$, with realization

  $$s_{\overline{x}}^2 = \frac{s_x^2}{n};$$

[10] Equation tells us that variance of shifted and rescaled variable y is equal to square of scale multiplied by variance of original variable x.

- **sample standard deviation of arithmetic average (sample standard error)** $S_{\overline{X}}$, with realization
$$s_{\overline{x}} = \frac{s_x}{\sqrt{n}};$$

- **sample skewness** B_1, with realization
$$b_1 = \frac{n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left[n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i-\overline{x})^3}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{3/2}},$$
distribution is *symmetric*, if $b_1 = 0$, *positive skewness* (density on the left side is steeper than the right side), if $b_1 > 0$ a *negative skewness* (density on the right side is steeper than the left side), if $b_1 < 0$;

- **sample kurtosis** B_2, with realization
$$b_2 = \frac{n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left[n^{-1}\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{2}} - 3 = \frac{n\sum_{i=1}^{n}(x_i-\overline{x})^4}{\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right]^{2}} - 3,$$
distribution is normal (*mesokurtic*), if $b_2 = 0$, pointy (*leptokurtic*), if $b_2 > 0$ and flat (*platykurtic*), if $b_2 < 0$;

- **sample sum of squares** $SS = \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, with realization
$$SS_{\text{obs}} = \sum_{i=1}^{n}(x_i-\overline{x})^2,$$

- **sample sum of absolute deviation** $SAD = \sum_{i=1}^{n}|X_i - \widetilde{X}_{0.5}|$, with realization
$$SAD_{\text{obs}} = \sum_{i=1}^{n}|x_i - \widetilde{x}_{0.5}|;$$

- **sample arithmetic average deviation** $MAD = \frac{1}{n}\sum_{i=1}^{n}|X_i - \widetilde{X}_{0.5}|$, with realization
$$MAD_{\text{obs}} = SAD_{\text{obs}}/n;$$

- **sample range** $D = X_{\max} - X_{\min}$, with realization
$$d_{\text{obs}} = x_{\max} - x_{\min};$$

- **sample interquartile range** $D_Q = Q_3 - Q_1$, with realization
$$d_Q = \widetilde{x}_{0.75} - \widetilde{x}_{0.25};$$
distribution is (between quartiles) *symmetric*, if $\widetilde{x}_{0.75} - \widetilde{x}_{0.50} = \widetilde{x}_{0.50} - \widetilde{x}_{0.25}$, *positively skewed*, if $\widetilde{x}_{0.75} - \widetilde{x}_{0.50} > \widetilde{x}_{0.50} - \widetilde{x}_{0.25}$ and *negatively skewed*, if $\widetilde{x}_{0.75} - \widetilde{x}_{0.50} < \widetilde{x}_{0.50} - \widetilde{x}_{0.25}$;

- **sample decile range** $D_D = \widetilde{X}_{0.9} - \widetilde{X}_{0.1}$, with realization

$$d_D = \widetilde{x}_{0.9} - \widetilde{x}_{0.1};$$

- **sample percentile range** $D_P = \widetilde{X}_{0.99} - \widetilde{X}_{0.01}$, with realization

$$d_P = \widetilde{x}_{0.99} - \widetilde{x}_{0.01}.$$

Robust spread characteristics (variability) are

- **sample γ-truncated variance** S_g^2, with realization s_g^2 calculated as

$$s_g^2 = \frac{1}{n-2g-1} \sum_{i=g+1}^{n-g} x_{(i)};$$

  more than $\gamma 100$ % must be replaced so that γ-truncated variance changes to large or small relative to the original s^2 [*breakdown point* s_g^2 is γ]; it applies that $s_g^2 < s^2$ because truncating removes outliers;

- **sample γ-winsorized variance** S_w^2, with realizations s_w^2; more than $\gamma 100$ % must be replaced so that *gamma*-winsorized variance changes to large or small relative to the original s^2 [*breakdown point* s_w^2 is γ]; it applies that $s_w^2 < s^2$ because winsorization pulls outliers closer to the mean;

- **sample quartile coefficient of variation** $V_{k,Q} = (Q_3 - Q_1)/Q_2$, with realization $v_{k,Q}$ calculated as

$$v_{k,Q} = \frac{\widetilde{x}_{0.75} - \widetilde{x}_{0.25}}{\widetilde{x}}.$$

Other robust spread characteristics characterized by modified boundaries are

- **sample robust minimum and maximum** ("inner boundaries") $X^*_{\min} = B_D = Q_1 - 1.5 D_Q$ and $X^*_{\max} = B_H = Q_1 + 1.5 D_Q$, with realizations defined as

$$x^*_{\min} = b_D = \widetilde{x}_{0.25} - 1.5\left(\widetilde{x}_{0.75} - \widetilde{x}_{0.25}\right),$$

$$x^*_{\max} = b_H = \widetilde{x}_{0.75} + 1.5\left(\widetilde{x}_{0.75} - \widetilde{x}_{0.25}\right),$$

  values outside of boundaries are considered to be *suspicious, potential outliers*;

- **sample robust minimum and maximum** ("outer boundaries") defined as $B^*_H = Q_1 - 3(Q_3 - Q_1)$, $B^*_H = Q_3 + 3(Q_3 - Q_1)$, with realizations $b^*_D = \widetilde{x}_{0.25} - 3d_Q$, $b^*_H = \widetilde{x}_{0.75} + 3d_Q$;

  - if there are any $x_i < b^*_D \vee x_i > b^*_H$, we call them *distant values*[11],
  - if $x_i \in \langle b^*_D, b_D) \vee (b_H, b^*_H\rangle$, these are *outer values*,

[11] Symbol $\vee$ means "or" and symbol $\wedge$ means "and".

– if $x_i \in \langle b_D, b_H \rangle$, these are *inner values* or *values close to median*;

– normal distribution has these properties $B_H - B_D = Q_3 + 1.5D_Q - Q_1 + 1.5D_Q = 4D_Q \doteq 4.2$; probability of $x_i \notin \langle B_D, B_H \rangle$ is then 0.04;

- **sample robust skewness** B_{1Q} and B_{1O} and their variance under asymptotic normality $B_{1\cdot}$, where $\cdot = Q$ or O, with realizations defined as

  – quartile skewness

$$b_{1Q} = \frac{(\widetilde{x}_{0.75} - \widetilde{x}_{0.50}) - (\widetilde{x}_{0.50} - \widetilde{x}_{0.25})}{\widetilde{x}_{0.75} - \widetilde{x}_{0.25}}, Var_{as}(b_{1Q}) = 1.84,$$

  – octile skewness

$$b_{1O} = \frac{(\widetilde{x}_{0.875} - \widetilde{x}_{0.50}) - (\widetilde{x}_{0.50} - \widetilde{x}_{0.125})}{\widetilde{x}_{0.875} - \widetilde{x}_{0.125}}, Var_{as}(b_{1O}) = 1.15.$$

Exercise 32 (height of 10-year old girls)**.** *Calculate all spread characteristics for the sample with heights of* 10*-year old girls.*

Exercise 33 (Hypothesis testing)**.** *Let* $X \sim Bin(N, p)$ *be a random variable with* $N = 1000$ *and* $p = 0.5$*. Make 100000 simulations* $x_1, \ldots, x_{100000}$ *and for each simulation calculate maximum likelihood estimate of p, i.e.* $\hat{p} = \frac{x}{N}$*. Plot histogram of* $\hat{p}$ *and add vertical line at value 0.534. Calculate percentage of* $\hat{p}$ *that are higher than 0.534.*

Exercise 34 (Hypothesis testing)**.** *Out of 1000 newborn children, 534 of them were boys and 466 girls. Test null hypothesis that probabilities of having a boy or a girl are equal. Calculate p-value, confidence interval, test statistic and critical region. Use Wald statistic and Likelihood statistic.*

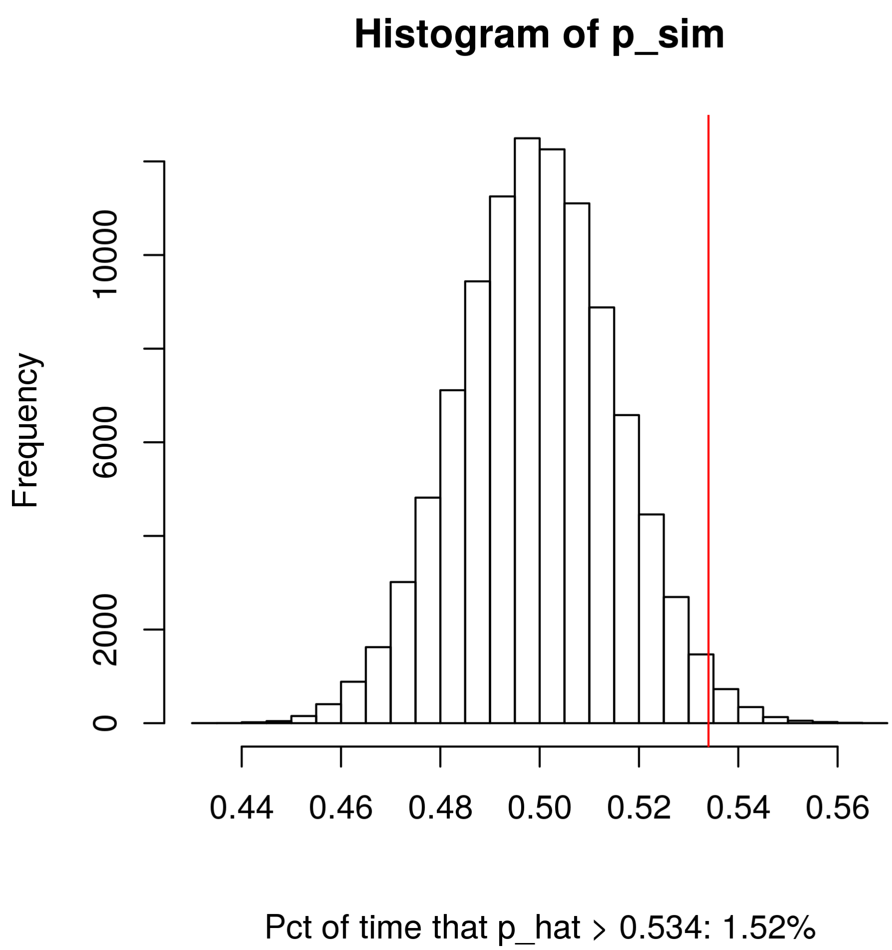

Figure 17: Histogram of simulations of $\hat{p}$ with line at 0.534.