Statistics in Computer Science

Seminar Exercises

Stanislav Katina, Mojmir Vinkler

October 15, 2016

1 Location characteristics

Exercise 1 (Code Vectorization). *Implement two versions of* `mean` *function in R - thee first one is* `mean_slow = function(x)` *and use* `for` *loop to sum numbers. The second version is* `mean_fast = function(x)` *and use built-in function* `sum` *to sum numbers. Generate very long random vector of uniformly distributed random numbers and compare performance of both functions and also of built-in* `mean` *function.*

Hints. Use `runif(...)` for generating random numbers and `system.time({...})` for profiling.

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 [[1] *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].

[1] *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).

Exercise 2 (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.*

2 Spread (variability) characteristics

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

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

$$s^2 = s^2_{n-1} = s^2_x = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2;$$

under linear transformation sample variance changes like this[2]

$$s^2_y = s^2_{a+bx} = b^2 s^2_x,$$

i.e.

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

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

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

under linear transformation standard deviation changes like this

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

[2] Equation tells us that variance of shifted and rescaled variable y is equal to square of scale multiplied by variance of original variable 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** $S^2_{\overline{X}}$, with realization

$$s^2_{\overline{x}} = \frac{s^2_x}{n};$$

- **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*[3],
  - if $x_i \in \langle b_D^*, b_D) \vee (b_H, b_H^* \rangle$, these are *outer values*,
  - 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 3 (height of 10-year old girls)**.** *Calculate all spread characteristics for the sample with heights of* 10*-year old girls.*

3 Basics of Probability

Exercise 4 (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***[4]*.(Dalgaard 2008). If we put a member back to population after choosing it, we talk about* ***simple random sample with replacement***[5]*. 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)[6], sample(x=..., size=..., replace=...)

[3] Symbol $\vee$ means "or" and symbol $\wedge$ means "and".
[4] n-combination without replacement from N members of set $\mathcal{M}$.
[5] n-combination with replacement from N members of set $\mathcal{M}$.
[6] requires library utils

Exercise 5 (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 6 (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* $\Pr\left(a < X < b\right) = \Pr\left(X < b\right) - \Pr\left(X < a\right) = F_X\left(b\right) - F_X\left(a\right)$*, where* $a = \mu - k\sigma$*,* $b = \mu + k\sigma$*,* $k = 1, 2, 3.$[7] *Write a function that takes parameters* μ*,* σ*,* a *and* b *and calculates probability*

$$\Pr\left(a < X < b\right).$$

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

```
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.

Exercise 7 (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* $\Pr\left(a \leq X \leq b\right) = 1 - \alpha$ *is equal to* $0.90, 0.95$ *a* 0.99*. Value* $x_{1-\alpha}$ *is a quantile of standardized normal distribution, i.e.* $\Pr(Z = \frac{X-\mu}{\sigma} < x_{1-\alpha}) = 1 - \alpha, Z \sim N(0, 1)$*. Similarly to previous exercise, write a function that would take parameters* μ*,* σ *and* α *and return values* a *and* b *as a vector.*

Hints. qnorm(alpha)

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.

[7] Note that $\Pr\left(a < X < b\right) = \Pr\left(a \leq X \leq b\right)$ 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.

Exercise 8 (Interactive Normal Distribution)**.** *Create an interactive Shiny application that will use the function defined in exercise 6 and show probability $Pr(X > 0)$, where μ and σ can be interactively set by user.*

Hints. Download folder **normalplot-nographics** from study materials. Run it from R console with command `runApp('normalplot-nographics', display.mode="showcase")` from parent directory of **normalplot-nographics** (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
  numericInput("sig",
            "sigma:",
            min = 0.1,
            max = 3,
            step = 0.1,
            value = 1),
  numericInput("mu",
            "mean:",
            min = -4,
            max = 4,
            value = 0,
            step = 0.5),

  textOutput("myTextOutput")
))
```

Listing 2: **server.R**

```
library(shiny)

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

  output$myTextOutput <- renderText({
    # TODO: use your code from previous exercise to show probabilities
    paste0('My sigma is ', input$sig)
  })
})
```

Exercise 9 (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?*

Solution without 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 10 (Binomial Distribution)**.** *Let's assume that* $\Pr(\mathit{swirl}) = 0.533 = p_1$ *is the probability of having dermatological pattern swirl on right thumb of male population and* $\Pr(\mathit{other}) = 0.467 = p_2$ *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 11 (Normal approximation of binomial distribution)**.** [8]

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

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 $N(Np, Npq)$*.*

[8] 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)