Course of Study Bachelor Computer Science	Exercises Statistics WS 2021/22

Sheet VII - Solutions

Discrete Random Variables and Distributions

1. A random experiment consists of tossing n (distinct) coins and recording the sequence of scores $X_1, X_2, ..., X_n$, where 1 denotes head and 0 denotes tail. Let Y denote the number of heads.

   (a) What is a suitable sample space Ω? How many elements contains Ω?

   (b) Express Y as a function on the sample space Ω.

   (c) Show that $|\{Y = k\}| = \binom{n}{k}$ for $k \in \{0, 1, ..., n\}$.

   (d) With n=5, explicitly list the elements in the event $\{Y = 3\}$.

   Answer: n different coins are tossed and $X_i =$ is the score for coin i. Furthermore, $Y = \sum_{i=1}^{n} X_i =$ Number of heads

   (a) $\Omega = \{(x_1, x_2, ..., x_n)|x_i \in \{0,1\} i = 1, 2, ..., n\}$, $|\Omega| = 2^n$

   (b) $Y = X_1 + X_2 + ... + X_n, Y : \Omega \mapsto \{0, 1\}$

   (c) Y = k : subset of $\{1, 2, ..., n\}$ with k elements

   (d) $\{Y = 3\} =$
   $\{(1,1,1,0,0), (1,1,0,1,0), (1,1,0,0,1), (1,0,1,0,1), (1,0,1,1,0),$
   $(1,0,0,1,1), (0,1,1,1,0), (0,1,1,0,1), (0,1,0,1,1), (0,0,1,1,1)\}$

2. Suppose that two fair, standard dice are tossed and the sequence of scores (X_1, X_2) recorded. Let $Y = X_1 + X_2$, denote the sum of the scores, $U = \min(X_1, X_2)$, the minimum score, and $V = \max(X_1, X_2)$ the maximum score.

   (a) Find the probability density function of (X_1, X_2).

   (b) Find the probability density function of Y.

   (c) Find the probability density function of U.

   (d) Find the probability density function of V.

   (e) Find the probability density function of (U, V).

Answer: Original data:
$$\begin{aligned} Y &= X_1 + X_2 \\ U &= \min(X_1, X_2) \\ V &= \max(X_1, X_2) \end{aligned}$$

(a) $P((X_1, X_2) = (i, j)) = \frac{1}{36}, i, j \in \{1, 2, 3, 4, 5, 6\}$

(b)

k	2	3	4	5	6	7	8	9	10	11	12
P(Y=k)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

(c)

k	1	2	3	4	5	6
P(U=k)	$\frac{11}{36}$	$\frac{9}{36}$	$\frac{7}{36}$	$\frac{5}{36}$	$\frac{3}{36}$	$\frac{1}{36}$

(d)

k	1	2	3	4	5	6
P(V=k)	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

(e) $P(i, i) = \frac{1}{36}, i = 1, 2, 3, 4, 5, 6$ and $P(i, j) = \frac{2}{36}, i < j = 2, 3, 4, 5, 6$

```
###################################################################
# Suppose that two fair, standard dice are tossed and the sequence
# of scores (X_1,X_2) are recorded. Let Y=X_1+X_2, denote the sum
# of the scores, U=min (X_1,X_2), the minimum score, and
# V=max (X_1,X_2) the maximum score.
#
# file: prob_rv_rolling_dice.R
###################################################################

library(gtools)
library(tidyverse)

# a) Find the probability density function of (X_1,X_2).
# expand.grid() create a data frame from all combinations of the
# supplied vectors or factors.
density.X1.X2 <-
  permutations(n = 6,r = 2, v = 1:6, repeats.allowed = TRUE) %>%
  as_tibble() %>%
  mutate(prob = 1/36)

# b) Find the probability density function of Y.
density.Y <- density.X1.X2 %>%
  mutate(Y = V1+V2) %>%
  group_by(Y) %>%
  mutate(prob.Y = sum(prob)) %>%
  select(Y, prob.Y) %>%
  unique()

# c) Find the probability density function of U.
density.U <- density.X1.X2 %>%
  rowwise() %>%
  mutate(U = min(V1,V2)) %>%
  group_by(U) %>%
  mutate(prob.U = sum(prob)) %>%
  select(U, prob.U) %>%
  unique()

# d) Find the probability density function of V.
density.V <- density.X1.X2 %>%
  rowwise() %>%
  mutate(V = max(V1,V2)) %>%
  group_by(V) %>%
  mutate(prob.V = sum(prob)) %>%
  select(V, prob.V) %>%
  unique()

# e) Find the probability density function of (U,V).
density.UV <- density.X1.X2 %>%
  rowwise() %>%
  mutate(UV = paste(min(V1,V2),max(V1,V2))) %>%
  group_by(UV) %>%
  mutate(prob.UV = sum(prob)) %>%
  select(UV, prob.UV) %>%
  unique()
```

```
##################################################################
# alternative solution
##################################################################
# a) Find the probability density function of (X_1,X_2).
# expand.grid() create a data frame from all combinations of the
# supplied vectors or factors.
x1 <- seq(1,6,1)
x2 <- seq(1,6,1)
x1_x2 <- expand.grid(x=x1,y=x2)
x1_x2_dens <- cbind(x1_x2,rep(1/36,36))
x1_x2_dens

# b) Find the probability density function of Y.
y <- x1_x2[,1]+x1_x2[,2]
y_dens <- table(y)/36
y_dens

# c) Find the probability density function of U.
# apply() returns a vector or array or list of values obtained by
# applying a function to margins of an array or matrix.
u <- apply(x1_x2,1,min)
u_dens <- table(u)/36
u_dens

# d) Find the probability density function of V.
v <- apply(x1_x2,1,max)
v_dens <- table(v)/36
v_dens

# e) Find the probability density function of (U,V).
uv <- cbind(x1_x2,u,v)
uv
#
uv_dens <- table(uv[,3],uv[,4])/36
uv_dens
```

3. R offers for a large number of probability distributions functions. The commands for each distribution are prepended with a letter to indicate the functionality:

   - "d" returns the height of the probability density function
   - "p" returns the cumulative density function
   - "q" returns the inverse cumulative density function (quantiles)
   - "r" returns randomly generated numbers

   Consider an urn with 100 balls, where are 30 balls of them are red. 20 balls are randomly drawn and let X be the number of red drawn balls.

   (a) Determine the distribution of X if the balls are drawn with resp. without replacement.

   (b) Plot the density of X.

   (c) Generate a sample of size 20 of values of X.

   (d) Compute $P(5 < X < 15)$.

   (e) Determine the 25% quantile, the median and the 75% quantile of X.

   Answer:

```
###########################################################################
# R offers for a large number of probability distributions functions.
# The commands for each distribution are prepended with a letter to
# indicate the functionality:
# "d"   returns the height of the probability density function
# "p"   returns the cumulative density function
# "q"   returns the inverse cumulative density function (quantiles)
# "r"   returns randomly generated numbers
# Consider an urn with 100 balls, where are 30 balls of them are red.
# 20 balls are randomly drawn and let X be the number of red drawn balls.
# a) Determine the distribution of X if the balls are drawn with resp.
# without replacement.
# b) Plot the density of X.
# c) Generate a sample of size 20 of values of X
# d) Compute P(5 < X < 15).
# e) Determine the 25% quantile, the median and the 75% quantile of X.
#
# file: prob_rv_rfunctions.R
###############################################################################

# with replacement: X ~ B(n=20,p=0.3)

# plot of the density
k <- 0:20
plot(k,dbinom(k,20,0.3), type = "h", main ="B(n=20,p=0.3", xlab="x",
     ylab="density")
# sample of size 20
rbinom(n=20,size = 20,prob = 0.3)
# P(5 < X < 15)
sum(dbinom(6:14, size = 20, prob = 0.3))
pbinom(14, size = 20, prob = 0.3) - pbinom(5, size = 20, prob = 0.3)
# quantile
qbinom(c(0.25,0.5,0.75), size = 20, prob = 0.3)

# without replacement: X ~ H(n=20,M=30,N=100)

# plot of the density
k <- 0:20
plot(k,dhyper(k,m=30,n=70,k=20), type = "h", main ="H(n=20,M=30,N=100",
     xlab="x", ylab="density")
# sample of size 20
rhyper(20,m=30,n=70,k=20)
# P(5 < X < 20)
sum(dhyper(6:14,m=30,n=70,k=20))
phyper(14,m=30,n=70,k=20) - phyper(5,m=30,n=70,k=20)
# quantile
qhyper(c(0.25,0.5,0.75),m=30,n=70,k=20)
```

4. In a game a player can bet 1$ on any of the numbers 1, 2, 3, 4, 5 and 6. Three dice are rolled. If the players number appears k times, where $k \geq 1$, the player gets k$ back plus the original stack of 1$. Over the long run, how many cents per game a player expects to win or lose playing this game?

   Answer: X = money per game, wanted: $E(X)$, data: $\begin{array}{rl} n = & 3 \\ p = & \frac{1}{6} \end{array}$ i.e.

   $p_1 = \mathcal{P}(\text{Player number occurs once}) = \binom{3}{1} \cdot \left(\frac{1}{6}\right)^1 \cdot \left(\frac{5}{6}\right)^2$,
   $p_2 = \mathcal{P}(\text{Player number occurs twice}) = \binom{3}{2} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{5}{6}\right)^1$,
   $p_3 = \mathcal{P}(\text{Player number occurs three times}) = \binom{3}{3} \cdot \left(\frac{1}{6}\right)^3 \cdot \left(\frac{5}{6}\right)^0$,
   $p_0 = \mathcal{P}(\text{Player number occurs 0 times}) = \binom{3}{0} \cdot \left(\frac{1}{6}\right)^0 \cdot \left(\frac{5}{6}\right)^3$

   Thus, we get
   $p_0 \cdot (0-1) + p_1 \cdot (2-1) + p_2 \cdot (3-1) + p_3 \cdot (4-1) = n \cdot p - p_0 = 3 \cdot \frac{1}{6} - p_0 = \frac{1}{2} - \left(\frac{5}{6}\right)^3 = -0.0787$

```
#########################################################
```

```
# In a game a player can bet 1$ on any of the numbers
# 1, 2, 3, 4, 5 and 6. Three dice are rolled. If the
# players number appears k times, where k >= 1, the player
# gets k$ back plus the original stack of 1$. Over the
# long run, how many cents per game a player expects to
# win or lose playing this game?
#
# file: prob_rv_exp_value.R
##########################################################

library(tidyverse)

# create a tibble with the different outcomes and their
# probabilities
R <-
  tibble(
    k=0:3
  )
R <-
 R %>%
  mutate(prob=choose(3,k)*(1/6)**k*(5/6)**(3-k)) %>%
  mutate(r = if_else(k==0,-1,-1+1+k))
R
# expected value
exp_r <- sum(R[,2]*R[,3])
exp_r
```

5. Consider a lottery of 20 tickets. Among the tickets there is are 1 first prize, 4 second prizes and 15 rivets. 5 tickets are drawn from the lottery drum. Determine the probability that

   (a) 2 rivets have been drawn.

   (b) 2 rivets, 2 second prizes and the first prize were drawn.

   (c) the 5th ticket drawn is the first ticket which is not a rivet.

   Calculate the probabilities if the tickets are drawn with replacement. What results will be obtained if the number of tickets in the drum is increased (with equal proportions of first prizes, second prizes and rivets)?

   Answer: Let N be the number of tickets, n the number of drawn tickets, FP the number of first prizes, SP the number of second prizes and R the number of rivets.

   (a) X=number of rivets drawn: the random variable is hypergeometricly distributed (H(n=5,R=15,N=20)):

   $$P(X=2)=\frac{\binom{15}{2}\cdot\binom{5}{3}}{\binom{20}{5}}\approx 0.0726$$

   (b) If Y= number of second prizes and Z = number of first prizes we have

   $$P(X=2,Y=2,Z=1)=\frac{\binom{15}{2}\cdot\binom{4}{2}\cdot\binom{1}{1}}{\binom{20}{5}}\approx 0.041$$

(c) U = number of tickets drawn to get the first non rivet ticket. If U = 5 the first 4 tickets are rivets and the 5th ticket is a non rivet.

$$P(U=5) = \frac{\binom{15}{4}}{\binom{20}{4}} \cdot \frac{5}{16} \approx 0.088$$

If the tickets are drawn with replacement X is binomially distributed (B(n=5,p=15/20)), (X,Y,Z) is multinomially distributed and U is geomatricly distributed. We get

$$\begin{array}{lll} P(X=2) & = \binom{5}{2}\left(\frac{15}{20}\right)^2 \cdot \left(\frac{5}{20}\right)^2 & \approx 0.088 \\ P(X=2, Y=2, Z=1) & = \frac{5!}{2!2!1!} \cdot \left(\frac{15}{20}\right)^2 \cdot \left(\frac{4}{20}\right)^2 \cdot \left(\frac{1}{20}\right)^1 & \approx 0.03375 \\ P(U=5) & = \left(1 - \frac{15}{20}\right)^4 \cdot \frac{5}{20} & \approx 0.0791 \end{array}$$

Increasing the number of tickets with the same ratio of rivets, second and firts prizes we get

fp	sp	riv	pa-worepl	pa-wrepl	pb-worepl	pb-wrepl	pc-worepl	pc-wrepl
1	4	15	0.0677	0.0879	0.0406	0.0338	0.0880	0.0791
6	24	90	0.0853	0.0879	0.0348	0.0338	0.0804	0.0791
11	44	165	0.0865	0.0879	0.0343	0.0338	0.0798	0.0791
16	64	240	0.0870	0.0879	0.0341	0.0338	0.0796	0.0791
21	84	315	0.0872	0.0879	0.0340	0.0338	0.0795	0.0791
26	104	390	0.0873	0.0879	0.0340	0.0338	0.0794	0.0791
31	124	465	0.0874	0.0879	0.0340	0.0338	0.0794	0.0791
36	144	540	0.0875	0.0879	0.0339	0.0338	0.0793	0.0791
41	164	615	0.0875	0.0879	0.0339	0.0338	0.0793	0.0791
46	184	690	0.0876	0.0879	0.0339	0.0338	0.0793	0.0791
51	204	765	0.0876	0.0879	0.0339	0.0338	0.0793	0.0791

Mention that the results in case of drawing ticktes with replacement do not not depend on the total number of tickets, since the ratios rivets, second and first prizes are identical. Furthermore mention that the results without replacement tends to the resuls with replacement.

```
##########################################################################
# Consider a lottery of 20 tickets. Among the tickets there
# are a first prize, 4 second prizes and 15 rivets. 5 tickets
# are drawn from the lottery drum. Determine the probability that
# a) 2 rivets have been drawn.
# b) 2 rivets, 2 second prizes and the first prize were drawn.
# c) The 5th lot drawn is the first lot which is not a rivet.
# Calculate the probabilities if the lots are drawn with replacement.
# What results will be obtained if the number of lots in the drum is
# increased (with equal proportions of first prizes, second prizes
# and rivets)?
#
# file: prob_rv_rivet.R
###########################################################################

library(tidyverse)

fp <- 1
sp <- 4
riv <- 15
n <- 5

# without replacement
# a) hypergeometric distribution
# H(k, M, N+M) density:
#dhyper(x = number of white balls drawn,
#       m = number of white balls in the urn,
#       n        = number of black balls in the urn,
#       k        = the number of balls drawn from the urn)
pa <- dhyper(2,riv,fp+sp,n)
pa # = choose(15,2)*choose(5,3) / choose(20,5)

# b) generalised hypergeometric distribution
pb <- choose(riv,2)*choose(sp,2)*choose(fp,1) /
  choose(riv+fp+sp,n)
pb
# c) geometric distribution
pc <- (choose(riv,n-1)/choose(fp+sp+riv,n-1)) *
  (fp+sp)/(fp+sp+riv -(n-1))
pc

# results with and without replacement for increasing
# number of tickets but constant proportion of the
# prizes
results <-
  tibble(
    m = seq(from=1,to=55,by=5),
    fp = m,
    sp = 4*m,
    riv = 15*m,
    pa_worepl = dhyper(2,riv,fp+sp,5),
    pa_wrepl = dbinom(2,5,riv/(riv+fp+sp)),
    pb_worepl = choose(riv,2)*choose(sp,2)*choose(fp,1) /
      choose(riv+fp+sp,1+2+2),
    pb_wrepl = factorial(1+2+2)/
      (factorial(2)*factorial(2)*factorial(1))*
      (riv/(riv+fp+sp))^2 * (sp/(riv+fp+sp))^2 *
      (fp/(riv+fp+sp))^1,
    pc_worepl = (choose(riv,5-1)/choose(fp+sp+riv,5-1)) *
      (fp+sp)/(fp+sp+riv -(5-1)),
    pc_wrepl = dgeom(5-1,(fp+sp)/(fp+sp+riv))
  )
results
```

6. compare Heumann, Schomaker p 176, Exercise 8.6
   A company organizes a raffle at an end-of-year function. There are 4000 raffle tickets to be sold, of which 500 win a prize. The price of each ticket is 1.5 Euro. The value of the prizes varies between 80 Euro and 250 Euro with an average of 142 Euro.

   (a) An employee wants to have a 99% guarantee of receiving three prizes. How much money does he need to spend? Use R to solve the question.

   (b) Use R to plot the function which describes the relationship between the number of tickets bought and the probability of winning at least three prizes.

   (c) Given the value of the prizes and the costs of the tickets, it is worth taking part in raffle?

   Answer: Let X_n be the number of prize tickets if you have drawn n tickets randomly. X_n is hypergemtrically distributed $(\sim H(n, 500, 4000))$. The probability to get at least 3 prize tickets is $p_n = P(X_n >= 3) = 1 - P(X_n <= 2)$. Thus we are looking for the smallest n with$p_n \geq 0.99$.

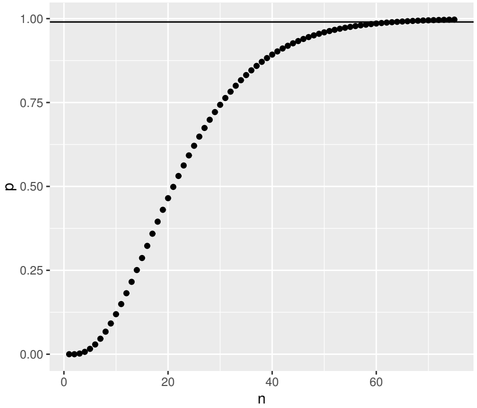
```
##########################################################
# A company organizes a raffle at an end-of-year function.
# There are 4000 raffle tickets to be sold, of which 500
# win a prize. The price of each ticket is 1.5 Euro. The
# value of the prizes varies between 80 Euro and 250 Euro
# with an average of 142 Euro.
#
# file: prob_rv_raffle.R
###########################################################

library(tidyverse)

# a) An employee wants to have a 99% guarantee of receiving
# three prizes. How much money does he need to spend?

# X number of drawn prize tickets, X ~ H(n,M=500,N=4000)
# with n number of drawn tickets
dis <-
  tibble(
    n=seq(1,4000),
    p=1-phyper(2,500,3500,n) # prob. of at least 3 wins
  )
dis
# 99% quantile
min(which(dis$p >=0.99))
# or
dis %>%
  filter(p >= 0.99) %>%
  summarise(money = min(n))

# b) Use R to plot the function which describes the relationship
# between the number of tickets bought and the probability of
# winning at least three prizes.
plot(x=dis$n, y=dis$p,
     type = "l", xlim = c(0,75),
     xlab = "n", ylab = "p")
abline(a=0.99, b=0)
# diagram applying ggplot()
dis %>%
  filter(n <= 75) %>%
  ggplot(mapping=aes(x=n,y=p))+
  geom_point()+
  geom_abline(slope = 0, intercept = 0.99)
```



You will get n=64 as the minimum n with $p_n \geq 0.99$. If you buy 64 tickets you must pay $64 \cdot 1.5 = 96$ Euro but you will get a return of $3 \cdot 142 = 426$ Euro with a probability of 0.99.
The total value of all prizes is $500 \cdot 142 = 71000$ Euro. If all tickets are sold the company will get $400 \cdot 1.5 = 6000$ Euro. Certainly the company have not paid the regular prize for all prizes.

Continous Random Variables and Distributions

1. The time T (in minutes) required to perform a certain job is uniformly distributed over the interval $[15, 60]$.

   (a) Find the probability that the job requires more than 30 minutes.

   (b) Given that the job is not finished after 30 minutes, find the probability that the job will require more than 15 additional minutes.

   Answer:

   (a) $f(x) = \frac{1}{b-a} = \frac{1}{60-15} = \frac{1}{45}$ i.e. $\int_{30}^{60} \frac{1}{45} dx = \frac{1}{45} \cdot [x]_{30}^{60} = \frac{30}{45} = \frac{2}{3}$

   (b) $X =$ total working time and we become
   $P(X > 45|X > 30) = \frac{P(X>45)}{P(X>30)} = \frac{\int_{45}^{60} \frac{1}{45} dx}{\int_{30}^{60} \frac{1}{45} dx} = \frac{\frac{1}{45} \cdot [x]_{45}^{60}}{\frac{30}{45}} = \frac{\frac{15}{45}}{\frac{30}{45}} = 0.5$

2. compare Heumann, Schomaker, p. 150, exercise 7.4
   A quality index summarizes different features of a product by means of a score. Different experts may assign different quality scores depending on their experience with the product. Let X be the quality index for a tablet. Suppose the respective probability density function is given as follows:

$$f(x) = \begin{cases} cx(2-x) & \text{if } 0 \le x \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

   (a) Determine c such that f is a proper probability density function.

   (b) Determine the cumulative distribution function.

   (c) What is probability that the score is less equal 1.5 and bigger 0.5?

   (d) Calculate the expectation and the variance.

   Answer:

   (a) From

$$1 = \int_0^2 cx(2-x)\text{dx} = c\int_0^2 (2x - x^2)\text{dx} = c(x^2 - \frac{x^3}{3})\bigg|_0^2 = \frac{4c}{3}$$

   we get c=0.75.

   (b) If $0 \le x \le 2$ we have

$$F(x) = 0.75\int_0^x u(2-u)\text{du} = 0.75x^2(1 - \frac{x}{3}) \Rightarrow$$

$$F(x) = \begin{cases} 0 & 0 \leq x \\ 0.75x^2(1-\frac{x}{3}) & 0 < x \leq 2 \\ 1 & 2 < x \end{cases}$$

(c)

$$P(0.5 < X \leq 1.5) = F(1.5) - f(0.5) = 0.75(\frac{9}{4}(1-\frac{1}{2}) - \frac{1}{4}(1-\frac{1}{6})) = \frac{11}{16}$$

(d)

$$\begin{aligned} E(X) &= 0.75 \int_0^2 x^2(2-x)\text{dx} = 0.75\,(\frac{2x^3}{3} - \frac{x^4}{4})\Big|_0^2 = 1 \\ E(X^2) &= 0.75 \int_0^2 x^3(2-x)\text{dx} = 0.75\,(\frac{x^4}{2} - \frac{x^5}{5})\Big|_0^2 = 1.2 \Rightarrow \\ Var(X) &= E(X^2) - (E(X))^2 = 1.2 - 1 = 0.2 \end{aligned}$$

Normal Distributions

1. R offers a number of functions for calculating with normal distributions. Call them up in the RStudio Help area with the keyword Normal and familiarize yourself with them.
   R has four in built functions to generate normal distribution.

   (a) The function qnorm() gives height of the probability distribution at each point for a given mean and standard deviation. Apply this function to create a plot the density of the normal distribution with mean 2.5 and standard deviation 1.5.

   (b) pnorm() gives the probability of a normally distributed random number to be less that the value of a given number (cumulative distribution function). Apply this function to create a plot of the normal distribution function with mean 2.5 and standard deviation 1.5.

   (c) qnorm() takes the probability value and gives a number whose cumulative value matches the probability value (quantile). Apply this function to plot the quantiles of the normal distribution with mean 2.5 and standard deviation 1.5.

   (d) rnorm() is used to generate random numbers whose distribution is normal. It takes the sample size as input and generates that many random numbers. Draw a histogram to show the distribution of the generated numbers which are normally distributed with mean 2.5 and standard deviation 1.5 .

Answer:

```
2###########################################################
# R offers a number of functions for calculating with normal
# distributions. Call them up in the RStudio Help area with
# the keyword Normal and familiarize yourself with them.
# R has four in built functions to generate normal distribution.
#
# file: prob_nd_rfunctions.R
###########################################################

# The function dnorm() gives height of the probability density
# at each point for a given mean and standard deviation. Apply
# this function to create a plot the density of the normal
# distribution with mean 2.5 and standard deviation 1.5.
x <- seq(-10, 10, by = .1)
y <- dnorm(x, mean = 2.5, sd = 1.5)
plot(x,y, type = ”l”)

# pnorm() gives the probability of a normally distributed random
# variable to be less that the value of a given number (cumulative
# distribution function). Apply this function to create a plot of
# the normal distribution function with mean 2.5 and standard
# deviation 1.5.
x <- seq(-10,10,by = .1)
y <- pnorm(x, mean = 2.5, sd = 1.5)
plot(x,y, type = ”l”)

# qnorm() takes the probability value and gives a number whose
# cumulative value matches the probability value (quantile).
# Apply this function to plot the quantiles of the normal distribution
# with mean 2.5 and standard deviation 1.5.
x <- seq(0, 1, by = 0.02)
y <- qnorm(x, mean = 2, sd = 1.5)

plot(x,y,type = ”l”)

# rnorm() is used to generate random numbers whose distribution is
# normal. It takes the sample size as input and generates that many
# random numbers. Draw a histogram to show the distribution of the
# generated numbers which are normally distributed with mean 2.5 and
# standard deviation 1.5.
y <- rnorm(100,mean = 2.5, sd = 1.5)
hist(y, main = ”Normal Distribution (mean=2.5, sd=1.5)”)
```

2. If scores are normally distributed with a mean of 35 and a standard deviation of 10, what percent of the scores is:

   (a) greater than 34?

   (b) smaller than 42?

   (c) between 28 and 34?

Answer: Score $S \sim N(\mu, \sigma^2)$ with $\mu = 35$ and $\sigma = 10$

(a) $P(X > 34) = 1 - P(X \leq 34) = 1 - P\left(\frac{X-35}{10} \leq \frac{34-35}{10}\right) =$
$1 - \Phi\left(\frac{34-35}{10}\right) = 1 - \Phi(-0.1) = 1 - (1 - \Phi(0.1)) = \Phi(0.1) = 0.53983$
R-Code: 1-pnorm(34,35,10)

(b) $P(X < 42) = P(X \leq 42) = \Phi\left(\frac{42-35}{10}\right) = \Phi(0.7) = 0.758036$
R-Code: pnorm(42,35,10)

(c) $P(28 \leq X \leq 34) = P(X \leq 34) - P(X \leq 28) =$
$= \Phi\left(\frac{34-35}{10}\right) - \Phi\left(\frac{28-35}{10}\right) = \Phi(-0.1) - \Phi(-0.7) =$
$1-\Phi(0.1)-(1-\Phi(0.7)) = \Phi(0.7)-\Phi(0.1) = 0.758036-0.539828 =$
0.2182

R-Code: pnorm(34,35,10) - pnorm(28,35,10)

3. Assume a normal distribution with a mean of 70 and a standard deviation of 12. What limits would include the middle 65% of the cases?

Answer: $X \sim N(70, 12^2)$ and $P(u \leq X \leq o) = 0.65$ i.e. $P(X \leq u) = 0.175$ and $P(X \leq o) = 0.175 + 0.65 = 0.825$ i.e. $\Phi\left(\frac{o-70}{12}\right) = 0.825$ i.e. $\frac{o-70}{12} = u_{0.825}$ (the 82.5%-Quantil of the $N(0,1)-$distribution)

With $u_{0.80} = 0.8416$ and $u_{0.85} = 1.0364$ we become
$u_{0.825} \approx u_{0.80} + \frac{1}{1}(u_{0.85} - u_{0.80}) = 0.939$

Totally $o = 70 + 12 \cdot u_{0.825} = 81.268$ and $u = 70 - 12 \cdot u_{0.825} = 58.732$

R-Code: qnorm(0.825,70,12) qnorm(0.825,70,12)

4. Suppose that weights of bags of potato chips coming from a factory follow a normal distribution with mean 12.8 ounces and standard deviation 0.6 ounces. If the manufacturer wants to keep the mean at 12.8 ounces but adjust the standard deviation so that only 1% of the bags weigh less than 12 ounces, how small does he need to make that standard deviation?

Answer: $X \sim N(12.8,\ 0.6^2)$
We keep the expectation 12.8 but adjust the standard deviation, such that $P(X < 12) = 0.01$. We become $P\left(\frac{X-12.8}{s} < \frac{12-12.8}{s}\right) = 0.01$ i.e. with the standard normal distribution we become $\Phi\left(\frac{-0.8}{s}\right) = 0.01$. With $\Phi(-x) = 1 - \Phi(x) : 1 - \Phi\left(\frac{0.8}{s}\right) = 0.01$ i.e.
$\Phi\left(\frac{0.8}{s}\right) = 0.99 \Rightarrow \frac{0.8}{s} = 2.3263$ i.e. $s = \frac{0.8}{2.3263} \approx 0.3439$

R-Code: (12-12.8)/qnorm(0.01,0,1)

5. Peter and Paul agree to meet at a restaurant at noon. Peter arrives at a time normally distributed with mean 12:00 and standard deviation 5 minutes. Paul arrives at a time normally distributed with mean 12:02 and standard deviation 2 minutes. Assuming the two arrivals are independent, find the probabilty that

(a) Peter arrives before Paul

(b) both men arrive within 3 minutes of noon

(c) the two men arrive within 3 minutes of each other

Answer: Arrival time Peter: $X \sim N\left(12, \left(\frac{1}{12}\right)^2\right)$

Arrival time Paul: $Y \sim N\left(12\frac{1}{30}, \left(\frac{1}{30}\right)^2\right)$

(a) $P(X < Y) = P(X - Y < 0)$ and
$X - Y \sim N\left(12 - 12\frac{1}{30}, \left(\frac{1}{12}\right)^2 + \left(\frac{1}{30}\right)^2\right) = N\left(-\frac{1}{30}, \left(\frac{1}{12}\right)^2 + \left(\frac{1}{30}\right)^2\right)$
$P(X - Y < 0) = \Phi\left(\frac{0+\frac{1}{30}}{\sqrt{\left(\frac{1}{12}\right)^2+\left(\frac{1}{30}\right)^2}}\right) = 0.644309$

(b) $P(11.95 \leq X \leq 12.05, 11.95 \leq Y \leq 12.05) =$
$= P(11.95 \leq X \leq 12.05) \cdot P(11.95 \leq Y \leq 12.05) =$
$= \left(\Phi\left(\frac{12.05-12}{\frac{1}{12}}\right) - \Phi\left(\frac{11.95-12}{\frac{1}{12}}\right)\right) \cdot \left(\Phi\left(\frac{12.05-12\frac{1}{30}}{\frac{1}{30}}\right) - \Phi\left(\frac{11.95-12\frac{1}{30}}{\frac{1}{30}}\right)\right) =$
$(0.72575 - (1 - 0.72575)) \cdot (0.69146 - (1 - 0.99379)) = 0.3094$

(c) $P\left(|X - Y| \leq \frac{1}{20}\right) = P(-0.05 \leq X - Y \leq 0.05) =$
$= \Phi\left(\frac{0{,}05+\frac{1}{30}}{\sqrt{\left(\frac{1}{12}\right)^2+\left(\frac{1}{30}\right)^2}}\right) - \Phi\left(\frac{-0{,}05+\frac{1}{30}}{\sqrt{\left(\frac{1}{12}\right)^2+\left(\frac{1}{30}\right)^2}}\right) = \Phi(0.93) - \Phi(-0.19) =$
$0.82 - (1 - 0.58) = 0.40$

```
#####################################################################
# Peter and Paul agree to meet at a restaurant at noon. Peter
# arrives at a time normally distributed with mean 12:00 and
# standard deviation 5 minutes. Paul arrives at a time normally
# distributed with mean 12:02 and standard deviation 2 minutes.
# Assuming the two arrivals are independent, find the following
# probabilie.
#
# file: prob_nd_peter_paul.R
######################################################################
m1 <- 12; m2 <- 12+2/60
s1 <- 5/60; s2 <- 2/60

# a) Peter arrives before Paul
p <- pnorm(0, m1-m2, sqrt(s1^2+s2^2))
p

# b) both men arrive within 3 minutes of noon
p <- (pnorm(12+3/60, m1, s1) - pnorm(12-3/60, m1, s1)) *
  (pnorm(12+3/60, m2, s2) - pnorm(12-3/60, m2, s2))
p

# c) the two men arrive within 3 minutes of each other
p <- pnorm(3/60, m1-m2, sqrt(s1^2+s2^2)) - pnorm(-3/60, m1-m2, sqrt(s1^2+s2^2))
p
```

6. The weight of a melon, X, in kg is $N(\mu = 1.2, \sigma^2 = 0.3^2)$, i.e. normally distributed with expectation 1.2 kg and standard deviation 0.3 kg. The weight Y for a pineapple is in kg $N(\mu = 0.6, \sigma^2 = 0.2^2)$. We assume that a melon and a pineapple are chosen independently of each other.

   (a) Which distribution has the total weight of the two fruits?

   (b) Calculate the probability that the total weight of the two fruits does not exceed 2.0 kg.

(c) The melon costs 2 euro per kg and the pineapple 4 euro per kg. Give an expression for the total price Z using X and Y. What is the distribution of Z?

(d) Calculate the probability that the price Z is higher than 4 euro.

Answer:

(a) $X+Y$ is normally distributed with expectation $E[X]+E[Y]=1.8$ and variance $Var(X)+Var(Y)=0.3^2+0.2^2=0.13$

(b) $P(X+Y\leq 2.0)=\Phi\left(\frac{2.0-1.8}{\sqrt{0.13}}\right)\approx\Phi(0.55)\approx 0.71$

(c) $Z=2X+4Y$. The price is a linear combination of normally distributed random variables and thus also normhjally distributed. The expectation is $2E[X]+4E[Y]=2\cdot 1.2+4\cdot 0.6=4.8$ and the variance is $2^2Var(X)+4^2Var(Y)=4\cdot 0.3^2+16\cdot 0.2^2=1$

(d) $P(Z>4)P=1-P(Z\leq 4)=1-\Phi((4-4.8)/\sqrt{1})=1-\Phi(-0.8)=\Phi(0.8)=0.7881$

```
##################################################################
# The weight of a melon, X, in kg is N(mu=1.2,sigma^2=0.3^2),
# i.e. normally distributed with expectation 1.2 kg and standard
# deviation 0.3 kg. The weight Y for a pineapple is in kg
# N(mu=0.6, sigma^2=0.2^2). We assume that a melon and a pineapple
# are chosen independently of each other.
#
# file: prob_nd_fruits.R
###################################################################

mu_m <- 1.2
sigma_m <- 0.3
mu_p <- 0.6
sigma_p <- 0.2

# a) Which distribution has the total weight of the two fruits?
# S = X + Y
mu_s <- mu_m + mu_p
sigma_s <- (sigma_m^2+sigma_p^2)^0.5

# b) Calculate the probability that the total weight of the two fruits
# does not exceed 2.0 kg.
pnorm(2,mu_s,sigma_s)

# c) The melon costs 2 euro per kg and the pineapple 4 euro per kg.
# Give an expression for the total price Z using X and Y. What is the
# distribution of Z?
mu_z <- 2*mu_m + 4*mu_p
sigma_z <- (2^2*sigma_m^2 + 4^2*sigma_p^2)^0.5

# d) Calculate the probability that the price Z is higher than 4 euro.
1-pnorm(4,mu_z,sigma_z)
```

Central Limit Theorem

1. A machine consists of the three modules A, B and C. The machine works only if all three modules are working and if no error occured

during the construction phase. The probabilities that the modules A, B and C are defect are 1%, 1% and 5%. The probability for an error during the construction phase 2%. The four kinds of errors occur independently of each other.

(a) Calculate the expectation and the variation of the number of defect machines in a lot of 1000 randomly chosen machines.

(b) The producer is thinking about guaranteeing that not more than 110 machines are defect i such a lot. With which approximate probability can this guarantee promise be kept?

(c) Each defect machine provokes an extra cost of 100 euro. The producer considers to buy a better module C (at a higher price) but with an error rate of is 1%.
How high can the additional cost for each machine for module C be, in order to say that it is (according to the expectation) profitable to buy the more expensive module C?

Answer:

(a) $P(A \text{ and } B \text{ and } C \text{ ok and no construction error}) =$
$= P(A \text{ ok}) \cdot P(B \text{ ok}) \cdot P(C \text{ ok}) \cdot P(\text{no construction error}) =$
$= 0.99 \cdot 0.99 \cdot 0.95 \cdot 0.98 = 0.91247$
$X = \#$ defect machines in a set of 1000 machines. Then $X \sim B(n = 1000, p = 0.08753)$. We become $E(X) = n \cdot p = 87.53$ and $Var(X) = n \cdot p \cdot (1 - p) = 87.53 \cdot (1 - 0.08753) \approx 79.86$

(b) Approximation with a normal distribution:
$P(X \leq 110) \approx \Phi\left(\frac{110+0.5-87.53}{\sqrt{79.86}}\right) \approx \Phi(2.58) = 0.995$

(c) Cost for each: $100 \cdot 0.0875 = 8.753$
The better module C: $p_c = 0.01$ and $\hat{p}$= probability that a machine is defect. We become:
$\hat{p} = 1 - 0.99 \cdot 0.99 \cdot 0.99 \cdot 0.98 \approx 0.0491$ Costs with this better module: $100\hat{p} = 4.91$ i.e. max. additional costs for module C: 8.753-4.91=3.843, thus 3.84 euro.

```
##################################################################
# A machine consists of the three modules A, B and C. The machine
# works only if all three modules are working and if no error
# occured during the construction phase. The probabilities that
# the modules A, B and C are defect are 1%, 1% and 5%. The
# probability for an error during the construction phase 2%. The
# four kinds of errors occur independently of each other.
#
# file: prob_cl_modules.R
##################################################################

# a) Calculate the expectation and the variation of the number of
# defect machines in a lot of 1000 randomly chosen machines.
```

```
n <- 1000
pa <- 0.01
pb <- 0.01
pc <- 0.05
pcp <- 0.02
# prob. of no error
p_no_error <- (1-pa)*(1-pb)*(1-pc)*(1-pcp)
# expected value
exp_n_def <- n*(1-p_no_error)
# variance
var_n_def <- n*(1-p_no_error)*p_no_error
p_no_error; exp_n_def; var_n_def

# b) The producer is thinking about guaranteeing that not more
# than 110 machines are defect in such a lot. With which approximate
# probability can this guarantee promise be kept?
pbinom(110,n,1-p_no_error) # 0.9936782
# approximation by a normal distribution
# without continuity correction
pnorm(110,mean=exp_n_def,sd=(var_n_def)^0.5) # 0.9940429
# with continuity correction
pnorm(110.5,mean=exp_n_def,sd=(var_n_def)^0.5) # 0.9949242

# c) Each defect machine provokes an extra cost of 100 euro.
# The producer considers to buy a better module C (at a higher
# price) but with an error rate of is 1%. How high can the
# additional cost for each machine for module C be, in order to
# say that it is (according to the expectation) profitable to buy
# the more expensive module C?
p_no_error_new <- (1-pa)*(1-pb)*(1-0.01)*(1-pcp)
100*(1-p_no_error) - 100*(1-p_no_error_new)
```

2. An airline knows that over the long run, 90% of passengers who reserve seats show up for their flight. On a particular flight with 300 seats, the airline accepts 324 reservations.

   (a) Assuming that passengers show up independently of each other, what is the chance that a passenger with a reservation do not get a seat?

   (b) How many reservations can be given, if the airline will accept an overbooking probability of 1%?

   Answer: Let n be the number of flight tickets, which are sold, and let $Y_i = 1$, if the person having flight ticket i shows up, and otherwise $Y_i = 0,\ 1 \leq i \leq n$.
   Y_i are Bernoulli-distributed stochastic variables with $P(Y_i = 0) = 0.05 = 1 - P(Y_i = 1)$ for all $1 \leq i \leq n$.

   Let $S_n = \sum_{i=1}^{n} Y_i$ be the number of passengers showing up. We assume that the passengers show up independent of each other. In this case we have $E[S_n] = n \cdot E[Y_1]$ und $Var(S_n) = n \cdot Var(Y_1)$ with $E[Y_1] = P(Y_1 = 1) = 0.95$ and $Var[Y_1] = P(Y_1 = 1) \cdot P(Y_1 = 0) = 0.95 \cdot 0.05$
   All passengers must become a seat with a probability higher than 0.99. The plane has totally 300 seats. Thus, the following equation must be fulfilled: $P(S_n > 300) < 0.01$
   Using the central limit theorem, we become $W = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} \sim N(0, 1)$ approximately i.e.

$P(S_n > 300) = P\left(\frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} > \frac{300 - E[S_n]}{\sqrt{Var(S_n)}}\right) < 0.01$

This can be expressed as:

$P\left(W > \frac{300.5 - 0.95 \cdot n}{\sqrt{n \cdot 0.95 \cdot 0.05}}\right) < 0.01$ and $P\left(W \leq \frac{300.5 - 0.95 \cdot n}{\sqrt{n \cdot 0.95 \cdot 0.05}}\right) \geq 0.99$

We become $\frac{300.5 - 0.95 \cdot n}{\sqrt{n \cdot 0.95 \cdot 0.05}} \geq \Phi^{-1}(0.99) \approx 2.33$ with $\Phi(z) = P(Z \leq z)$, which, together with $2.33 \cdot \sqrt{0.95 \cdot 0.05} \approx 0.508$, can be written as: $0.95 \cdot n + 0.508\sqrt{n} - 300.5 \leq 0$

We solve a second degree equation in order to determine n (first to determine $\sqrt{n}$), and become

$\sqrt{n} = \frac{-0.508 + \sqrt{0.508^2 + 4 \cdot 0.95 \cdot 300.5}}{2 \cdot 0.95} \approx \frac{33.29}{1.9} = 17.5$

Thus we have $n \leq 17.5^2 \approx 306.25$. Thus, the airline should sell at most 306 flight tickets.

```
################################################################
# An airline knows that over the long run, 90% of  passengers
# who reserve seats show up for their flight. On a particular
# flight with 300 seats, the airline accepts 324 reservations.
#
# file: prob_cl_airline.R
################################################################

# a) Assuming that passengers show up independently of each other,
# what is the chance that a passenger with a reservation do not
# get a seat?
n <- 324; p <- 0.9
# exact value
p_ex <- 1-pbinom(300.5,n,p)
p_ex
# approx. value
m <- n*p; s <- sqrt(n*p*(1-p))
p_app <- 1-pnorm(300.5,m,s)
p_app

# b) How many reservations can be given, if the airline will
# accept an overbooking probability of 1%?

# exact bound
n <- seq(301,350,1)
p_ex <- pbinom(300,n,p)
o_ex <- n[max(which(p_ex >= 0.99))]
o_ex

# approx. bound
m <- n*p; s <- sqrt(n*p*(1-p))
p_app <- pnorm(300.5,m,s)
o_app <- n[max(which(p_app >= 0.99))]
o_app

# alternative solution: solving the equation
# (300.5 - p*n)^2 = p*(1-p)*u^2_0.99*n
# quadratic equation: n^2 + a*n +b =0
u_099 <- qnorm(0.99,0,1)
#u_099
#p
a <- -(2*p*300.5 + p*(1-p)*u_099^2)/(p^2)
b <- (300.5/p)^2
# a;b
n1 <- -a/2 + sqrt((a/2)^2 -b)
n2 <- -a/2 - sqrt((a/2)^2 -b)
n1; n2
300.5-n1*p
300.5-n2*p

# solution without solving the above equation
library(tidyverse)
tibble(
  n = 300:324,
  p = 1-pbinom(300, size = n, prob = 0.9),
  p.app = 1-pnorm(300.5, mean = n*0.6, sd = sqrt(n*0.9*0.1))
) %>% filter(p >= 0.01) %>% filter(n == min(n)) # n < 321!!
```

3. As a new residential area with 1000 domestic homes is going to be built, the number of required parking lots is calculated in the following way: We assume that there is no relation between the number of cars in different homes. Furthermore, we assume that a domestic home has no car with probability 0.2, one car with probability 0.7 and two cars with probability 0.1. The number of parking lots should be planned in such way that the probability that each car gets a parking lot is 0.99.

   How many parking lots should be built?

   Answer: Let $X_i,\ 1 \leq i \leq 1000$ be the number of cars in household i. Then $X_i,\ 1 \leq i \leq 1000$ are independent stochastic variables with the same discrete distribution.
   $P(X_i = 0) = 0.2$ and $P(X_i = 1) = 0.7$ und $P(X_i = 2) = 0.1$, $i = 1, \ldots, 1000$
   Let Y be the total number of cars in the 1000 households:
   $Y = \sum_{i=1}^{1000} X_i$. The central limit theorem gives that:
   $Z = \frac{\left(\sum_{i=1}^{1000} X_i\right) - 1000 \cdot E[X_1]}{\sqrt{1000 \cdot Var(X_1)}}$ is an approximatively normally distributed variable is, with $Z \sim N(0,1)$.

   We have $E[X_1] = \sum_{x=0}^{2} xP(X_1 = x) = 1 \cdot 0.7 + 2 \cdot 0.1 = 0.9$ and $Var(X_1) = E(X_1^2) - (E(X_1))^2 = 1^2 \cdot 0.7 + 2^2 \cdot 0.1 - 0.9^2 = 0.29$
   The smalest number of parking lots to build, so that all cars get a parking lot with probability 0.99, is given by: $P(Y \leq y) = 0.99$. We become the equation
   $0.99 = P(Y \leq y) = P\left(\sum_{i=1}^{1000} X_i \leq y\right) = P\left(\frac{\left(\sum_{i=1}^{1000} X_i\right) - 1000 \cdot 0.9}{\sqrt{1000 \cdot 0.29}} \leq \frac{y - 1000 \cdot 0.9}{\sqrt{1000 \cdot 0.29}}\right) \approx$
   $P\left(Z < \frac{y-900}{17.03}\right)$
   This gives $\frac{y-900}{17.03} = \Phi^{-1}(0.99) = 2.33$
   Thus $y = 900 + 2.33 \cdot 17.03 \approx 939.7$. Thus, 940 parking lots must be built.

```
#################################################################
# As a new residential area with 1000 domestic homes is going
# to be built, the number of required parking lots is calculated
# in the following way: We assume that there is no relation between
# the number of cars in different homes. Furthermore, we assume
# that a domestic home has no car with probability 0.2, one car
# with probability 0.7 and two cars with probability 0.1. The number
# of parking lots should be planned in such way that the probability
# that each car gets a parking lot is 0.99. How many parking lots
# should be built?
#
# file: prob_cl_res_area.R
#################################################################

p0 <- 0.2
p1 <- 0.7
p2 <- 0.1
n <- 1000
# expected values
EX <- 0*p0 + 1*p1 + 2*p2
EX2 <- 0^2*p0 + 1^2*p1 + 2^2*p2
# variance
```

```
VarX <- EX2 - (EX)^2
EX
VarX
# 99% quantile
qnorm(0.99,n*EX,(n*VarX)^0.5) # 939.6163
```

4. Starting in the origin, a particle is moving along the integer axes in this way:
   At each time point $1, 2, 3, \ldots$, the particle is moving either one step to the left or one step to the right with the same probability. There is also a third alternative: The particle stays constant, without moving. This alternative has the probability p and $0 < p < 1$. The movement of the particle is independent of earlier movements.

   How should p be chosen, if we want that the probability is 1% that the particle at time point 100 is located to the right of point 15?

   Answer: Let S_n be the location of the particle at time n.
   Then we have $S_n = \sum_{i=1}^{n} X_i$, with X_i independent, identically distributed random variables and
   $P(X_i = 0) = p$ and $P(X_i = 1) = P(X_i = -1) = (1-p)/2$.

   We look for p such that $P(S_{100} > 15) = 0.01$. We have $E[X_i] = 0$ and $Var(X_i) = E[X_i^2] = 1 - p$.
   The central limit theorem gives that $S_n/\sqrt{n(1-p)}$ converges towards a $N(0,1)-$distribution. Thus:
   $P(S_{100} > 15) = 1 - P\left(\frac{S_{100}}{\sqrt{100(1-p)}} \leq \frac{15}{\sqrt{100(1-p)}}\right) \approx 1 - \Phi\left(\frac{3}{2\sqrt{1-p}}\right)$
   This must be equal to 0.01. Thus, $3/2\sqrt{1-p}$ must be equal to the 99%-quantile of the $N(0,1)-$distribution, i.e. $3/2\sqrt{1-p} = 2.3263$, and this gives $p \approx 0.58$.

   R-Code: 1-(1.5/qnorm(0.99,0,1))$\hat{2}$