```
#* KLAUSUR WS 21
#*(a) Import the file melanoma.csv as a tibble called melanoma.
library(readr)
melanoma <- read_csv("melanoma.csv")
View(melanoma)

# (b) Determine type and scale of all variables. You find a description
#       of the dataset in the file melanoma description.pdf
#+  time: quantiative, discrete, absolute
#+  status: qualitative, discrete, nominal
#+  sex : qualitative ,discrete nominal
#+  age : quantitative , discrete absolute
#+  year: quantitative , discrete, interval
#+  thickness: quantitative, continious, ratio
#+  ulcer : qualtitative, discrete, nominal
melanoma

#(c) Change the values of the variables sex, status, ulcer to strings de-
#  scribing their values and add a new variable live.status describing
#  whether the patient is alive or dead.
library(tidyverse)
melanoma <- melanoma %>%
  mutate(sex = if_else(sex==1,'male','female')) %>%
  mutate(status = case_when(
    status == 1 ~ 'melanoma' ,
    status == 2 ~ 'alive',
    status == 3 ~ 'other reason'
  )) %>%
  mutate(ulcer = if_else(ulcer==1,'present','absent')) %>%
  mutate(live.status=if_else(status=='alive','alive','dead'))

melanoma

# (d) Create a contingency table for the variables sex and live.status.
chitest <- chisq.test(melanoma$sex,melanoma$live.status)
cont.tab <- chitest$observed
cont.tab

# (e) Evaluate the relative risks to survive at least 3 years for the va-
#  riable sex and interpret the values.
cont.tab2 <- melanoma %>% filter(time >= 365*3)
cont.tab2 <- chisq.test(cont.tab2$sex, cont.tab2$live.status)$observed
cont.tab2
#cont.tab2$sex alive dead
#female    91   19
#male      42   15
# relative risks for female
# alive -> outcome, female,male->exposure
#relative risk female
91/(91+19)  # women have a chance of 82% to survive atleast 3 years
#relative risk male
42/(42+15)  # men have a chance of 73% to survive atleast 3 years
# proportion between female and male
(91/(91+19))/(42/(42+15)) # 1.12
# interpretation: females have 1.12 higher risk to survive at least 3 years

# (f) Create a summary describing the distribution of the variable age
#     containing min, max, mean, the three quartiles depending on the
```

```
#     variable sex.
measures <- melanoma %>% group_by(sex) %>% summarise(
  min = min(age),
  max = max(age),
  mean = mean(age),
  q1 = quantile(age, 0.25, type=1),
  q2 = quantile(age, 0.5 , type=1),
  q3 = quantile(age, 0.75, type=1),
  iqr = IQR(age, type=1)
)
measures
# (g) Create side by side boxplots for the age of persons depending on
#     their sex and interpret the diagram.
boxplot(melanoma$age~melanoma$sex)
# both boxplots seem to be symmetric so the median is in the center of the
box
# the minimum of the female is less than of the males
# the maximum of the male is higher than of the females
# there is a extreme value in the data of the females
# iqr of female is 23 and iqr of male is 24 so the spread of the group of
males
# is higher
# (h) The csv file add.data.melanoma.csv contains data from another
#     study. Import the dataset as a tibble called add.data.melanoma.
library(readr)
add.data.melanoma <- read_csv("add.data.melanoma.csv")
View(melanoma)
add.data.melanoma
# (i) Is this dataset tidy?
# no its not tidy because the column sex_age_year contains multiple
# other attributes!
# each column must be a single attribute

# Probability
# Task 2
#2. In a computer science course the projects P1, P2, P3, P4 are offered.
#   Each of the 60 students must sign up for one of the projects offered.
#   All projects are equally popular among the students. Determine the
#   probability that
#   (a) exactly 15 students register for every of the four projects
#   (b) more than 15 students sign up for project P1.
#   Assume that the number of places in the projects is unlimited:
#  Hint: The R functions factorial(n) and choose(n,m) evaluate n! and n
over m

# SOLUTION:
# a) auf spicker
# b) P(X > 15)
1- pbinom(15,60,1/4) # 0.4312

# Task 3
# year, an introductory computer course is held at the beginning
#of the winter semester. From many years of experience, we know that
#about 11% of the registered course participants do not show up for
#the course. Since each participant needs his own computer during the
#course, no more participants can take part in the course than there are
#free computers. In total, there are ten rooms with 22 seats each and a
#total of 240 first-year students. Using an approximation by the central
```

```
#limit theorem, calculate

#(a) the probability that all students who are present for the course
#will find a seat if all first-year students have registered for the
#course.
# 11% of the registered do not show up
# total pcs = 10*22 == 220 seats
# total of 240 first year students
n <- 240
pnorm(220,n*0.89, sqrt(n*0.89*(1-0.89)))

#(b) the minimum number of computers needed so that there is at
#least a 99% probability that all students who show up will have a
#computer?
qnorm(0.99, 240*0.89, sqrt(n*0.89*(1-0.89)))
#  (c) how many registrations may be accepted at most, if with probabi-
#  lity 0.99 all students who show up for the course will find a place
#  in the course with 220 places.
library(tidyverse)
tibble(
  n = 220:240,
  p = pnorm(220, n*0.89, sqrt(n*0.89*(1-0.89)))
) %>% filter(p>=0.99) %>% filter(n==max(n))

# Inferential
# Task 5
# A company produces chocolade bars with a standard weight of 100 gr.
#As a measure of quality controls he weighs 15 bars and obtains the
#following results:
#  98.32,97.26,99.85,99.52,95.73,95.56,100.49,98.19,95.16,
#  98.26,96.46,100.23,99.76,98.58,97.43
sample <- c(98.32,97.26,99.85,99.52,95.73,95.56,100.49,98.19,95.16,
            98.26,96.46,100.23,99.76,98.58,97.43)
mu0 <- 100
#(a) What is an appropriate hypothesis regarding the expected weight
# µ for a two-sided-test?
# H0: m0 == 100, H1: m0 != 100

#  (b) If weights can be assumed to be nomally distributed, which test
#should used to test these hypothesis?
# an appropriate test would be t -test because the sd is unknown
# and we are testing mu
# (normal model)

#  (c) Conduct the test that was suggested to be used in b) at a 5%
#level. What is your test decision. Specify the p-value.
alpha <- 0.05;
t.test(x = sample, mu = mu0, alternative="two.sided",
       conf.level = 1 -alpha)
# pvalue is = 0.0007251 which is much lower than alpha,
# so we are rejecting the null hypothesis

#(d) Based on the sample, the producer changes the settings in produc-
#  tion. To check whether the correction has led to an improvement,
#he again takes 15 chocolate bars and weighs them.
#100.14,100.05,96.51,98.70,98.22,101.06,103.55,100.16,
#100.60,102.85,103.15,100.66,102.52,102.09,100.84
#What is an appropriate hypothesis for comparing the expected
```

```
#weights of the two samples?
sample2 <- c(100.14,100.05,96.51,98.70,98.22,101.06,103.55,100.16,
             100.60,102.85,103.15,100.66,102.52,102.09,100.84)

#  (e) Provide an appropriate statistical test to test the hypothesis and
#perform at the 5% level. Assume that the variances of the popu-
#  lations of the two samples are equal. What is your test decision?
#  Specify the p-value.
# H0: mu1 >= mu2 , H1: mu1 < mu 2
t.test(sample, sample2, alternative="less", paired = F, var.equal =T,
       conf.level = 1-alpha)
# pvalue = 0.0002228 much lesser than alpha -> reject H0

#(f) In question e) the population variances of the two samples are
#assumed to be equal. Verify that the variances are equal using an
#appropriate test at the 10% level.
var.test(sample, sample2, alternative = "two.sided",
         conf.level = 1 -0.1)
```