PROBLEMS

1. Cats weight 5kg on average with standard deviation 2 kg. The distribution is positively skewed. Describe the distribution of the average of a sample of 25 cats.

Sampling distribution of average 25 cat weights given population standard deviation known can be approximately be taken as a normal distribution having as expected value the true population mean and standard error of 2/sqrt(25). A 95% CI is [4.22, 5.78]

```
# As sigma is known to be 2. A 95% CI using standard normal statistic
test
lci <- 5 - qnorm(0.975)*2/sqrt(25)
uci <- 5 + qnorm(0.975)*2/sqrt(25)
lci;uci

## [1] 4.216014

## [1] 5.783986
```

2. Female height in Benin have average height 160 cm with variance 400. The heights of 25 women are measured: (a) Describe the distribution of the mean of such a sample and (b) Find the probability that the mean of such a sample is above 165 cm.

a) Sampling distribution of average 25 women height given population standard deviation known can be approximately be taken (25 units) as a normal distribution having as expected value the true population mean and standard error of 20/sqrt(25).
b) z0 = 165-160/(20/sqrt(25)) = 1.25. P(Z>z0) = 1-P(z<1.25) = 0.11.

```
# As sigma is known to be 2. A 95% CI using Student t derived statistic
test
lci <- 160 - qt(0.975,24)*20/sqrt(25)
uci <- 160 + qt(0.975,24)*20/sqrt(25)
lci;uci

## [1] 151.7444

## [1] 168.2556

uci

## [1] 168.2556

uci <- 160 + qnorm(0.95)*20/sqrt(25);uci

## [1] 166.5794
```

```
1-pnorm(1.25)

## [1] 0.1056498
```

3. A sample of 50 centenarians found that the age at which they died has $\sum_{i=1}^{50} x_i = 5200$ and $\sum_{i=1}^{50} x_i^2 = 540000$. Find an estimate for the mean life expectancy and standard deviation of all centenarians.

Mean sample life expentancy is 5200/50=104 years. Sample variance is ((540000/49)-(104^2)) = 204.41 years2. Mean life expentancy transformation t = (104- mu)/(14.3/sqrt(50)) is Student-t distributed with 49 d.f.

```
mm <- 5200/50
mm2 <- 540000/49
vv <- (mm2 - (mm^2))
ss <- sqrt( vv )
mm; mm2; vv; ss

## [1] 104

## [1] 11020.41

## [1] 204.4082

## [1] 14.29714

set.seed(12345)

lci <- 104 - qt(0.975, 49)*ss/sqrt(50)
uci <- 104 + qt(0.975, 49)*ss/sqrt(50)
lci;uci

## [1] 99.9368

## [1] 108.0632
```

4. Donkey's ears are known to have standard deviation 5 cm. A farmer samples 30 donkeys and finds they have ears with average length 20 cm. (a) Construct a 95% confidence interval for the mean length of this farmer's donkeys (b) A veterinarian claims the donkeys have mean length below 22 cm. Is there evidence to support this claim?

a) Sampling distribution of average 30 donkey ears given population standard deviation known can be approximately be taken (30 units) as a normal distribution having as expected value the true population mean and standard error of 5/sqrt(30). A 95% CI [18.21, 21.80] (see calculus below under normal distribution for the mean sampling distribution).

b) H0: mu = 22 H1: mu<22 One sided test with test statistic z0 = (20-22)/(5/sqrt(30)) = -2.19. P(Z<z0) = 0.014, thus H0 is rejected and veterinarian claim is supported by data.

```
lci <- 20 - qt(0.975, 29)*5/sqrt(30)
uci <- 20 + qt(0.975, 29)*5/sqrt(30)
lci;uci

## [1] 18.13297

## [1] 21.86703

# As sigma is known to be 5. A 95% CI using standard normal statistic
test
lci <- 20 - qnorm(0.975)*5/sqrt(30)
uci <- 20 + qnorm(0.975)*5/sqrt(30)
lci;uci

## [1] 18.21081

## [1] 21.78919

# b

t0 <- (20 - 22)/(5/sqrt(30));t0

## [1] -2.19089

pt(t0, 29)

## [1] 0.01832234

qt(0.05,29)

## [1] -1.699127

# It is z0 since population variance is known
pnorm(t0)

## [1] 0.01422987

qnorm(0.05)

## [1] -1.644854
```

5. Every week a dairy farm samples 50 eggs to check if any have imperfections. The data for the first 10 weeks is given below. Week 1 2 3 4 5 6 7 8 9 10. Number of eggs with imperfections 3 2 5 4 3 5 3 6 11 8 (a) Calculate an estimate for the proportion of eggs with imperfections (b) Find the upper and lower control limits at 5% confidence level. **

```
dada <- c(3,    2,  5,  4,  3,  5,  3,  6,  11, 8)
mm <- sum( dada )/500; mm

## [1] 0.1

sum( dada )
```

```
## [1] 50

prop.test( sum(dada), n = 500, p =0.1, correct = FALSE)

##
##  1-sample proportions test without continuity correction
##
## data:  sum(dada) out of 500, null probability 0.1
## X-squared = 0, df = 1, p-value = 1
## alternative hypothesis: true p is not equal to 0.1
## 95 percent confidence interval:
##  0.07667756 0.12942191
## sample estimates:
##   p
## 0.1

lci <- mm - qnorm(0.975)*sqrt(mm*(1-mm)/50)
uci <- mm + qnorm(0.975)*sqrt(mm*(1-mm)/50)
lci;uci

## [1] 0.01684577

## [1] 0.1831542
```

6. The weight of 30 chestnuts in grams is found to have $\sum_{i=1}^{30} x_i = 75$ and $\sum_{i=1}^{30} x_i^2 = 300$. Calculate an estimate for the mean and standard deviation of the weight of chestnuts.

Mean is 2.5 g and sample standard deviation is 2.024 g. According to the sample a 95% CI for chestnut weight is 2.5 +/- qt(0.975, 29)*2.024/sqrt(30) -> 1,74 to 3.26 g at 95% confidence level.

```
mm <- 75/30
mm2 <- 300/29
vv <- (mm2 - (mm^2))
ss <- sqrt( vv )
mm; mm2; vv; ss

## [1] 2.5

## [1] 10.34483

## [1] 4.094828

## [1] 2.023568

# A sample is not available, then theoretical formulas have to be applied
lci <- mm - qt(0.975, 29)*ss/sqrt(30)
uci <- mm + qt(0.975, 29)*ss/sqrt(30)
lci;uci

## [1] 1.744387
```

```
## [1] 3.255613
```

7. The number of sesame seeds on five big mac buns was: 150, 120, 130, 127, 133. Determine a 95% confidence interval for the mean number of sesame seeds on a big mac bun, and interpret it in context. From a long term study the variance is known to be 124.5 (so does not need to be calculated).

```
dada <- c(150, 120, 130, 127, 133)
mm <- mean( dada )
ss <- sd( dada )
mm; ss

## [1] 132

## [1] 11.15796

t.test( dada, mu=mm )

##
##  One Sample t-test
##
## data:  dada
## t = 0, df = 4, p-value = 1
## alternative hypothesis: true mean is not equal to 132
## 95 percent confidence interval:
##  118.1456 145.8544
## sample estimates:
## mean of x
##       132

lci <- mm - qt(0.975, 4)*ss/sqrt(5)
uci <- mm + qt(0.975, 4)*ss/sqrt(5)
lci;uci

## [1] 118.1456

## [1] 145.8544

# Now knowing true variance
lci <- mm - qnorm(0.975)*sqrt(124.5)/sqrt(5)
uci <- mm + qnorm(0.975)*sqrt(124.5)/sqrt(5)
lci;uci

## [1] 122.2198

## [1] 141.7802
```

8. Suppose that you manage a bank where the amounts of daily deposits and daily withdrawals are given by independent random variables with normal distributions. For deposits, the mean is £12,000 and the Standard deviation is £4,000; for withdrawals, the mean is £10,000 and the standard deviation is £5,000. (a) For a week, calculate the probability that the five withdrawals will

add up to more than £55,000. (b) For a particular day, calculate the probability that withdrawals will exceed deposits by more than £5,000.

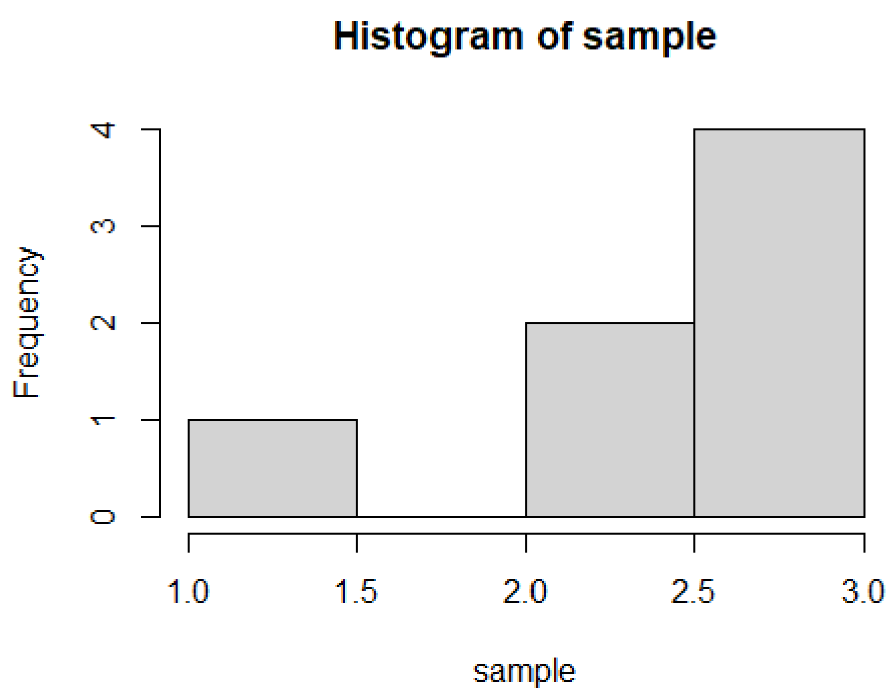
```
mm1 <- 12000; ss1 <- 4000
mm2 <- 10000; ss2 <- 5000

# a
smm1 <- 5*mm1; sss1 <- sqrt(5*((ss1)^2))
smm2 <- 5*mm2; sss2 <- sqrt(5*((ss2)^2)) #withdrawals
smm2;sss2

## [1] 50000

## [1] 11180.34

1-pnorm( 55000, mean = smm2, sd=sss2 )

## [1] 0.3273604

(55000-50000)/sss2

## [1] 0.4472136

#b
1-pnorm( 5000, mean = mm2-mm1, sd = sqrt( (ss1^2) + (ss2^2)))

## [1] 0.1371494

(5000+2000)/sqrt( (ss1^2) + (ss2^2))

## [1] 1.093216
```

**11. To forecast the yearly inflation (in percent, %), a simple random sample has been gathered: 1.5, 2.1, 1.9, 2.3, 2.5, 3.2, 3.0. It is assumed that the variable inflation follows a normal distribution. (a) By using these data, construct a 99% confidence interval for the mean of the inflation. (b) Experts have the opinion that the previous interval is too wide, and they want a total length of a unit. Find the level of confidence for this new interval. (c) Construct a confidence interval of 90% for the standard deviation.

**

```
dada<-c(1.5, 2.1, 1.9, 2.3, 2.5, 3.2, 3.0)
mm <- mean(dada)
svar<-var(dada)
ss <- sd(dada)
mm; ss; svar

## [1] 2.357143

## [1] 0.599603

## [1] 0.3595238
```

```r
set.seed(12345)
sample <- rnorm(7,mean=mm, sd= ss)
hist( sample )
```




```r
t.test( dada, conf.level = 0.99 )

##
##  One Sample t-test
##
## data:  dada
## t = 10.401, df = 6, p-value = 4.627e-05
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  1.516933 3.197352
## sample estimates:
## mean of x
##  2.357143

lci <- mm - qt(0.995, 6)*ss/sqrt(7)
uci <- mm + qt(0.995, 6)*ss/sqrt(7)
lci;uci

## [1] 1.516933

## [1] 3.197352

#b + qt(0.995, 6)*ss/sqrt(n)=0.5  -> Absolute error 1/2, find sample size
nn <-((qt(0.995, 6)*ss)^2)/(0.5^2);nn
```

```
## [1] 19.76665

# P((-0.5)/(ss/sqrt(7)) < t(6)< (0.5)/(ss/sqrt(7))) -> Confidence
interval
pt((0.5)/(ss/sqrt(7)),6) - pt((-0.5)/(ss/sqrt(7)),6) # Student t between
-0.5/(ss/sqrt(7)) to 0.5/(ss/sqrt(7))

## [1] 0.9304989

#c
library(EnvStats)

## Warning: package 'EnvStats' was built under R version 4.0.5

##
## Attaching package: 'EnvStats'

## The following objects are masked from 'package:stats':
##
##     predict, predict.lm

## The following object is masked from 'package:base':
##
##     print.default

varTest( dada, conf.level=0.9)

## $statistic
## Chi-Squared
##    2.157143
##
## $parameters
## df
##  6
##
## $p.value
## [1] 0.1906129
##
## $estimate
##  variance
## 0.3595238
##
## $null.value
## variance
##        1
##
## $alternative
## [1] "two.sided"
##
## $method
## [1] "Chi-Squared Test on Variance"
##
```

```
## $data.name
## [1] "dada"
##
## $conf.int
##       LCL       UCL
## 0.1713162 1.3190445
## attr(,"conf.level")
## [1] 0.9
##
## attr(,"class")
## [1] "htestEnvStats"
```

```
lci <- (7-1)*svar/qchisq(0.95, 6)
uci <- (7-1)*svar/qchisq(0.05, 6)
lci;uci
```

```
## [1] 0.1713162
```

```
## [1] 1.319045
```

12. In the library of a university, the mean duration (in days, d) of the borrowing period seems to be 20d. A simple random sample of 100 books is analyzed, and the values 18d and 8d2 are obtained for the sample mean and the sample variance, respectively. Construct a 99% confidence interval for the mean duration of the borrowings to check if the initial population value is inside

```
mm <- 18
ss <- sqrt(8)

set.seed(12345)
sample <- rnorm(100,mean=mm, sd= ss)
t.test( sample, conf.level = 0.99 )
```

```
##
##  One Sample t-test
##
## data:  sample
## t = 59.289, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##  17.86543 19.52161
## sample estimates:
## mean of x
##  18.69352
```

```
lci <- mm - qt(0.995, 99)*ss/sqrt(100)
uci <- mm + qt(0.995, 99)*ss/sqrt(100)
lci;uci
```

```
## [1] 17.25714
```

```
## [1] 18.74286
```

```
# 20 days is not included in the 99%CI
```

13. The accounting firm Price Waterhouse periodically monitors the U.S. Postal Service's performance. One parameter of interest is the percentage of mail delivered on time. In a simple random sample of 332,000. mailed items, Price Waterhouse determined that 282,200 items were delivered on time (Tampa Tribune, March 26, 1995.) Use this information to estimate with 99% confidence the true percentage of items delivered on time by the U.S. Postal Service.

```
pp <- 282200/332000
lci <- pp - qnorm(0.995)*sqrt(pp*(1-pp)/332000)
uci <- pp + qnorm(0.995)*sqrt(pp*(1-pp)/332000)
lci;uci
```

```
## [1] 0.8484037
```

```
## [1] 0.8515963
```

```
prop.test(282200,332000, p = pp, correct=F )
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  282200 out of 332000, null probability pp
## X-squared = 1.063e-27, df = 1, p-value = 1
## alternative hypothesis: true p is not equal to 0.85
## 95 percent confidence interval:
##  0.8487813 0.8512106
## sample estimates:
##    p
## 0.85
```

14. Two independent groups, A and B, consist of 100 people each of whom have a disease. A serum is given to group A but not to group B, which are termed treatment and control groups, respectively; otherwise, the two groups are treated identically. Two simple random samples have yielded that in the two groups, 75 and 65 people, respectively, recover from the disease. To study the effect of the serum, build a 95% confidence interval for the difference ηA–ηB. Does the interval contain the case ηA = ηB?

```
prop.test(c(75,65),c(100,100),correct=F) # 0 is contained
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(75, 65) out of c(100, 100)
## X-squared = 2.381, df = 1, p-value = 0.1228
## alternative hypothesis: two.sided
## 95 percent confidence interval:
```

```
##  -0.02626185  0.22626185
## sample estimates:
## prop 1 prop 2
##   0.75   0.65

pp1 <- 75/100
pp2 <- 65/100

lci <- (pp1-pp2) - qnorm(0.975)*(sqrt((pp1*(1-pp1)/100)+(pp2*(1-
pp2)/100)))
uci <- (pp1-pp2) + qnorm(0.975)*(sqrt((pp1*(1-pp1)/100)+(pp2*(1-
pp2)/100)))
lci;uci

## [1] -0.02626185

## [1] 0.2262618
```

15. The lengths (in millimeters, mm) of metal rods produced by an industrial process are normally distributed with a standard deviation of 1.8mm. Based on a simple random sample of nine observations from this population, the 99% confidence interval was found for the population mean length to extend from 194.65mm to 197.75mm. Suppose that a production manager believes that the interval is too wide for practical use and, instead, requires a 99% confidence interval extending no further than 0.50mm on each side of the sample mean. How large a sample is needed to achieve such an interval?

```
ss <- 1.8

lci <- 194.65
uci <- 197.75
# lci <- mm - qnorm(0.995)*sscheck/sqrt(9)
# uci <- mm + qnorm(0.995)*sscheck/sqrt(9)

# qnorm(0.995)*1.8/sqrt(nn) = 0.5
nn <- ((qnorm(0.995)*1.8)^2)/(0.5^2);ceiling(nn)  # 86

## [1] 86
```

16. The mark of an aptitude exam follows a normal distribution with standard deviation equal to 28.2. A simple random sample with nine students yields the following results: $/sum\left(x_j\right) = 1{,}098$ **and** $/sum\left(x_j^2\right) = 138{,}148$**. a) Find a 90% confidence interval for the population mean μ. b) Discuss without calculations whether the length of a 95% confidence interval will be smaller, greater or equal to the length of the interval of the previous section. c) How large must the minimum sample size be to obtain a 90% confidence interval with length (distance between the endpoints) equal to 10?**

```
sigma<- 28.2
mm <- 1098/9
```

```
ss2 <-((138148/8)-(mm^2))
ss <- sqrt( ss2 )
mm; ss2; ss

## [1] 122

## [1] 2384.5

## [1] 48.83134

# a
lci <- mm - qnorm(0.95)*sigma/sqrt(9)
uci <- mm + qnorm(0.95)*sigma/sqrt(9)
lci;uci

## [1] 106.5384

## [1] 137.4616

#b
lci <- mm - qnorm(0.975)*sigma/sqrt(9)
uci <- mm + qnorm(0.975)*sigma/sqrt(9)
lci;uci

## [1] 103.5763

## [1] 140.4237

#c
nn <- ((qnorm(0.95)*28.2)^2)/(5^2);ceiling(nn)  # 87

## [1] 87
```

17. A 64-element simple random sample of petrol consumption (litres per 100 kilometers, in private cars has been taken, yielding a mean consumption of 9.36u and a standard deviation of 1.4u. Then: a) Obtain a 96% confidence interval for the mean consumption. b) Assume both normality (for the consumption) and variance σ2 = 2u2. How large must the sample be if, with the same confidence, we want the maximum error to be a quarter of litre?

```
mm <- 9.36
ss <- 1.4
# a
lci <- mm - qt(0.98,63)*ss/sqrt(64)
uci <- mm + qt(0.98,63)*ss/sqrt(64)
lci;uci

## [1] 8.992999

## [1] 9.727001

# b
sigma <- sqrt(2)
```

```
lci <- mm - qnorm(0.98)*sigma/sqrt(nn)
uci <- mm + qnorm(0.98)*sigma/sqrt(nn)
# qnorm(0.98)*sigma/sqrt(nn)=1/4
nn <- ((qnorm(0.98)*sigma)^2)/(0.25^2);ceiling(nn)  # 135

## [1] 135
```

18. A company makes two products, A and B, that can be considered independent and whose demands follow the distributions N(μA, σA=70u) and N(μB, σB=60u), respectively. After analysing 500 shops, the two simple random samples yield means of a = 156 and b = 128. (a) Build 95 and 98 percent confidence intervals for the difference between the population means. (b) What are the absolute errors? If sales are measured in the unit u = number of boxes, what is the unit of measure of the margin of error? (c) A margin of error equal to 10 is desired, how many shops are necessary? (d) If only product A is considered, as if product B had not been analysed, how many shops are necessary to guarantee a margin of error equal to 10?

```
mm1 <- 156
mm2 <- 128
sigma1 <- 70
sigma2 <- 60

# a
# Pooled mean does not apply
sigma <- sqrt(((70^2)/500)+((60^2)/500))
lci98 <- mm1- mm2 - qnorm(0.99)*sigma
uci98 <- mm1 -mm2 + qnorm(0.99)*sigma
lci98;uci98

## [1] 18.40822

## [1] 37.59178

lci95 <- mm1- mm2 - qnorm(0.975)*sigma
uci95 <- mm1 -mm2 + qnorm(0.975)*sigma
lci95;uci95

## [1] 19.91886

## [1] 36.08114

# b
qnorm(0.99)*sigma

## [1] 9.591778

qnorm(0.975)*sigma

## [1] 8.081139
```

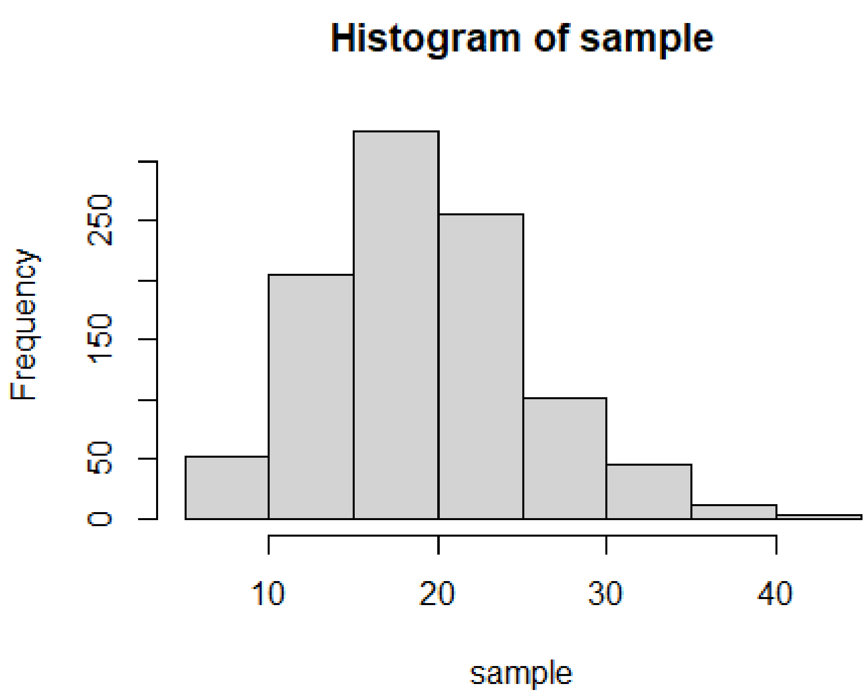
```r
# c
sigmapo <- sqrt(((70^2))+((60^2)))
nn <- ((qnorm(0.99)*sigmapo)^2)/(10^2);ceiling(nn)  # 461

## [1] 461

nn <- ((qnorm(0.975)*sigmapo)^2)/(10^2);ceiling(nn)  # 327

## [1] 327

# d - For A product
nn <- ((qnorm(0.99)*sigma1)^2)/(10^2);ceiling(nn)  # 266

## [1] 266

nn <- ((qnorm(0.975)*sigma1)^2)/(10^2);ceiling(nn)  # 189

## [1] 189
```

19. The lifetime of a machine (measured in years, y) follows a normal distribution with variance equal to 4y2. A simple random sample of size 100 yields a sample mean equal to 1.3y. Test the null hypothesis that the population mean is equal to 1.5y, by applying a two-tailed test with 5 percent significance level. What is the type I error? Calculate the type II error when the population mean is 2y.

```r
sigma <- sqrt(4)
mm <- 1.3

z0 <- (1.3 - 1.5)/(sigma/sqrt(100));z0

## [1] -1

2*(pnorm(z0))  # z0 negative  H0 fails to be rejected

## [1] 0.3173105

#b Type I error 0.05
# P(H0 acceptance| H0 false) = P( Z <-1.96-((2 - 1.5)/(2/sqrt(100)))) +
P(Z >1.96-((2 - 1.5)/(2/sqrt(100))))
z1l <-0-1.96-((2 - 1.5)/(2/sqrt(100))); z1l

## [1] -4.46

z1u<-1.96-((2 - 1.5)/(2/sqrt(100))); z1u

## [1] -0.54

pnorm(z1u) - pnorm(z1l)

## [1] 0.2945944
```

20. A company produces electric devices operated by a thermostatic control. The standard deviation of the temperature at which these controls actually operate should not exceed 2.0ºF. For a simple random sample of 20 of these controls, the sample standard deviation of operating temperatures was 2.39ºF. a) Stating any assumptions you need (write them), test at the 5% level the null hypothesis that the population standard deviation is not larger than 2.0ºF against the alternative that it is. b) Calculate the type II error at σ2=4.5ºF2.

```
ss <- 2.39
# H0: sigma= 4  H1: sigma > 4 Normal population  (n-1)ss^2/sigma^2 ~ X2(n-1)
# (n-1)ss^2/sigma^2
set.seed(12345)
sample <- rchisq(1000,19)
hist(sample)
```




```
chi<-(20-1)*(ss^2)/4;chi

## [1] 27.13248

qchisq(0.975,19)

## [1] 32.85233

qchisq(0.025,19)

## [1] 8.906516
```

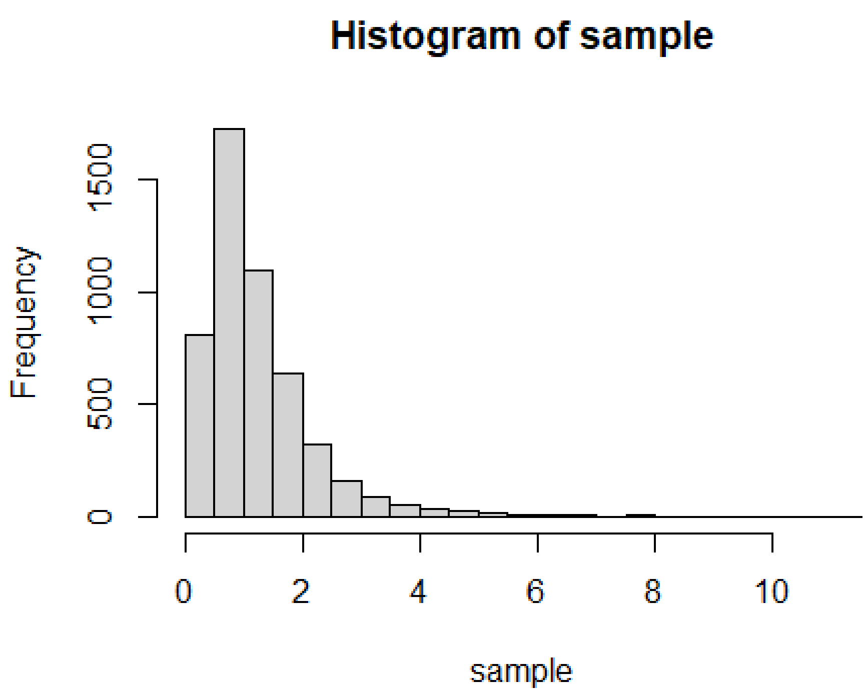
```r
pchisq(chi,19) # pvalue > 0.05 H0 can not be rejected

## [1] 0.898395

#b P( (n*(ss^2)/4.5)*(4.5/4) < 30.14 | H1) = P(X2(19)<30.14353*4/4.5)
# upper threshold
qchisq(0.95,19)

## [1] 30.14353

# P(X2(19) < 30.14353*4/4.5)
pchisq(30.14353*4/4.5,19)

## [1] 0.8904624
```

21. Let X = (X1,...,Xn) be a simple random sample with 25 data taken from a normal population variable X. The sample information is summarized in $\sum_{i=1}^{n} x_i = 105$ and $\sum_{i=1}^{n} x_i^2 = 579.24$.(a) Should the hypothesis H0: σ2 = 4 be rejected when H1: σ2 > 4 and α = 0.05?(b) And when H1: σ2 ≠ 4 and α = 0.05?

```r
sigma <- 4
mm <- 105/25
ss2 <-((579.24/24)-(mm^2))
ss <- sqrt( ss2 )
mm; ss2; ss

## [1] 4.2

## [1] 6.495

## [1] 2.548529

chi<-(25-1)*(ss2)/4;chi

## [1] 38.97

1-pchisq(chi,24) # pvalue < 0.05 H0 can  be rejected

## [1] 0.02750781

qchisq(0.95,24) #

## [1] 36.41503

qchisq(0.975,24) # two.sided alpha=0.05 upper threshold

## [1] 39.36408

qchisq(0.025,24) # two.sided alpha=0.05 lower threshold H0 can not be
rejected

## [1] 12.40115
```

22. The distribution of a variable is supposed to be normally distributed in two independent biological populations. The two population variances must be compared. After gathering information through simple random samples of sizes $n_X = 11, n_Y = 10$, respectively, we are given the value of the estimators $s_x^2 = 6.8$ and $s_y^2 = 7.1$. For α = 0.1, test: (a) H0: σX = σY against H1: σX < σY (b) H0: σX = σY against H1: σX > σY (c) H0: σX = σY against H1: σX ≠ σY.

```
set.seed(12345)
sample <- rf(5000,9,10)
hist( sample,30 )
```




```
f0 <-7.1/6.8;f0

## [1] 1.044118

#c)
1/qf(0.95,10,9)

## [1] 0.3187474

qf(0.05,9,10);qf(0.95,9,10)  # H0 fails to be rejected

## [1] 0.3187474

## [1] 3.020383

#b)
qf(0.9,9,10) # H0 fails to be rejected
```

```
## [1] 2.347306

#a)
qf(0.1,9,10) # H0 fails to be rejected

## [1] 0.4138532
```

23. Two simple random samples of 700 citizens of Italy and Russia yielded, respectively, that 53% of Italian people and 47% of Russian people wish to visit Catalonia within the next ten years. a) Should we conclude, with confidence 0.99, the Italians' desire is higher than the Russians'? b) Determine the critical region and make a decision. What is the type I error? c) Calculate the p-value and apply the methodology based on the p-value to make a decision.

```
p1 <- 0.53
p2 <- 0.47
(p1-p2)

## [1] 0.06

700*p1; 700*p2

## [1] 371

## [1] 329

prop.test(c(371,329),c(700,700),conf.level = 0.99,correct = F)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(371, 329) out of c(700, 700)
## X-squared = 5.04, df = 1, p-value = 0.02477
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.008717907  0.128717907
## sample estimates:
## prop 1 prop 2
##   0.53   0.47

lci <- (p1-p2)-qnorm(0.995)*sqrt((p1*(1-p1)+p2*(1-p2))/700);lci

## [1] -0.008717907

uci <- (p1-p2)+qnorm(0.995)*sqrt((p1*(1-p1)+p2*(1-p2))/700);uci

## [1] 0.1287179

prop.test(c(371,329),c(700,700),conf.level = 0.99,
alternative="greater",correct = F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(371, 329) out of c(700, 700)
## X-squared = 5.04, df = 1, p-value = 0.01238
## alternative hypothesis: greater
## 99 percent confidence interval:
##  -0.002062248  1.000000000
## sample estimates:
## prop 1 prop 2
##   0.53   0.47
```

```
uci <- (p1-p2)+qnorm(0.99)*sqrt((p1*(1-p1)+p2*(1-p2))/700);uci
```

```
## [1] 0.1220622
```

```
z0 <- (p1-p2)/sqrt((p1*(1-p1)+p2*(1-p2))/700);z0
```

```
## [1] 2.249046
```

```
1-pnorm(z0) # H0 can not be rejected pvalue: P(Z>z0) > 0.01 - H1 p1>p2
```

```
## [1] 0.01225477
```

```
qnorm(0.99) # H0 can not be rejected
```

```
## [1] 2.326348
```

```
pnorm(z0) # H0 can not be rejected pvalue: P(Z<z0) > 0.01 - H1 p1<p2
```

```
## [1] 0.9877452
```

```
prop.test(c(371,329),c(700,700),conf.level = 0.99,
alternative="less",correct = F)
```

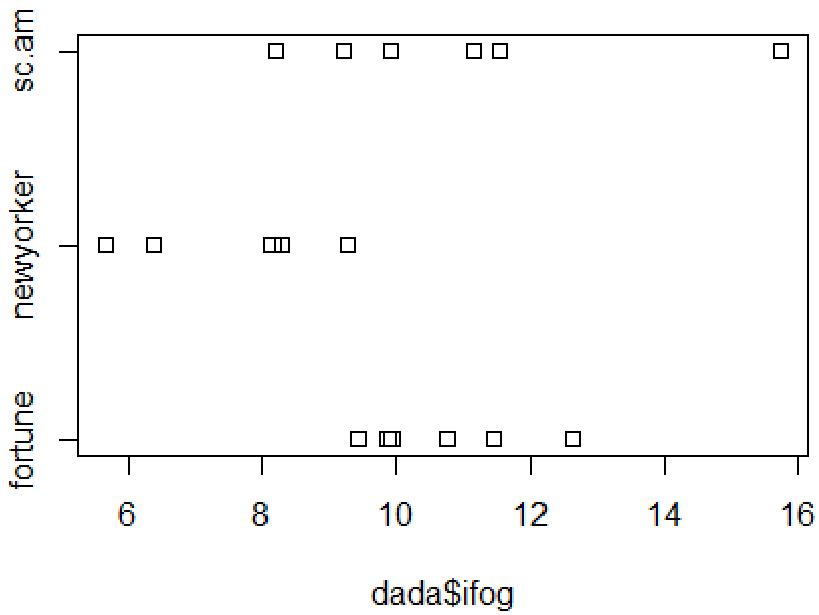
```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  c(371, 329) out of c(700, 700)
## X-squared = 5.04, df = 1, p-value = 0.9876
## alternative hypothesis: less
## 99 percent confidence interval:
##  -1.0000000  0.1220622
## sample estimates:
## prop 1 prop 2
##   0.53   0.47
```

```
lci <- (p1-p2)-qnorm(0.99)*sqrt((p1*(1-p1)+p2*(1-p2))/700);lci
```

```
## [1] -0.002062248
```

24. The fog index is used to measure the reading difficulty of a written text: The higher the value of the index, the more difficult the reading level. We want to know if the reading difficulty index is different for three magazines: Scientific American, Fortune, and the New Yorker. Three independent random samples of 6 advertisements were taken, and the fog indices for the 18 advertisements were measured, as recorded in the following table. Apply an analysis of variance to test whether the average level of difficulty is the same in the three magazines.

```
dada <- data.frame( ifog = c(15.75,11.55,
11.16,9.92,9.23,8.20,12.63,11.46,10.77,
9.93,9.87,9.42,9.27,8.28,8.15,6.37,6.37,5.66), magazin =
c(rep("sc.am",6),rep("fortune",6),rep("newyorker",6)))

dada$magazin <- factor(dada$magazin)
kruskal.test( dada$ifog ~ dada$magazin )

##
##  Kruskal-Wallis rank sum test
##
## data:  dada$ifog by dada$magazin
## Kruskal-Wallis chi-squared = 9.6357, df = 2, p-value = 0.008084

oneway.test( dada$ifog ~ dada$magazin )

##
##  One-way analysis of means (not assuming equal variances)
##
## data:  dada$ifog and dada$magazin
## F = 10.05, num df = 2.0000, denom df = 9.3867, p-value = 0.004644

stripchart(dada$ifog ~ dada$magazin)
```




```
pairwise.wilcox.test( dada$ifog, dada$magazin, alternative="greater")

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
compute
## exact p-value with ties

##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity
correction
##
## data:  dada$ifog and dada$magazin
##
##           fortune newyorker
## newyorker 1.00    -
## sc.am     1.00    0.03
##
## P value adjustment method: holm

pairwise.wilcox.test( dada$ifog, dada$magazin, alternative="less")

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
compute
## exact p-value with ties
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
compute
## exact p-value with ties

##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity
correction
##
## data:  dada$ifog and dada$magazin
##
##           fortune newyorker
## newyorker 0.0075  -
## sc.am     0.9372  0.9935
##
## P value adjustment method: holm
```

```
pairwise.wilcox.test( dada$ifog, dada$magazin, alternative="two.sided")
```

```
## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
compute
## exact p-value with ties

## Warning in wilcox.test.default(xi, xj, paired = paired, ...): cannot
compute
## exact p-value with ties

##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity
correction
##
## data:  dada$ifog and dada$magazin
##
##           fortune newyorker
## newyorker 0.015   -
## sc.am     0.937   0.040
##
## P value adjustment method: holm
```

25. The following table is based on data from the U.S. Department of Labor, Bureau of Labor Statistics. a) Use the data in the table, coming from a simple random sample, to test the claim that occupation is independent of whether the cause of death was homicide. Use a significance α = 0.05 and apply a nonparametric chi-square test. b) Does any particular occupation appear to be most prone to homicides? If so, which one?

```
dada <- data.frame(
deaths=c(82,92,107,9,70,29,59,42),occu=c(1,1,2,2,3,3,4,4), homicide =
rep(c(1,2),4));dada
```

```
##   deaths occu homicide
## 1     82    1        1
## 2     92    1        2
## 3    107    2        1
## 4      9    2        2
## 5     70    3        1
## 6     29    3        2
## 7     59    4        1
## 8     42    4        2
```

```
dada$homicide <- factor( dada$homicide, labels=c("Homicide", "Other"))
dada$occu <- factor( dada$occu, labels=c("Police",
"Cashiers","TaxiD","Guards"))

tdada <- xtabs( deaths ~ homicide +occu, data=dada)
tdada
```

```
##           occu
## homicide   Police Cashiers TaxiD Guards
##   Homicide     82      107    70     59
##   Other        92        9    29     42
```

```
chisq.test(tdada)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tdada
## X-squared = 65.524, df = 3, p-value = 3.875e-14
```

```
prop.table(tdada,2)
```

```
##           occu
## homicide       Police   Cashiers      TaxiD     Guards
##   Homicide 0.47126437 0.92241379 0.70707071 0.58415842
##   Other    0.52873563 0.07758621 0.29292929 0.41584158
```

26. World War II Bomb Hits in London. To carry out an analysis, South London was divided into 576 areas. For the variable N ≡ number of bombs in the k-th area (any), a simple random sample (x1,...,x576) was gathered and grouped in the following table. By applying the chi-square goodness-of-fit methodology, (1) Test at 95% confidence whether N can be supposed to follow a Poisson distribution. (2) Test at 95% confidence whether N can be supposed to follow a Poisson distribution with λ = 0.8.

```
dada <- data.frame( nb.bombs=c(0,1,2,3,4,5),nb.areas=c(229, 211, 93, 35,
7, 1));dada
```

```
##   nb.bombs nb.areas
## 1        0      229
## 2        1      211
```

```
## 3         2       93
## 4         3       35
## 5         4        7
## 6         5        1

elambda<-weighted.mean(dada$nb.bombs,dada$nb.areas)

dada$exp.areas <- 576*dpois(dada$nb.bombs,lambda=elambda)
dada$exp.areas[6] <- 576-sum(dada$exp.areas[1:5])

dada$cchiest<-((dada$nb.areas-dada$exp.areas)^2)/dada$exp.areas

chiest<-sum(dada$cchiest);chiest

## [1] 1.17238

# H0: Data Poisson distributed
1-pchisq( chiest, 4 )

## [1] 0.8826248

dada$exp.areas8 <- 576*dpois(dada$nb.bombs,lambda=0.8)
dada$exp.areas8[6] <- 576-sum(dada$exp.areas8[1:5])
dada

##   nb.bombs nb.areas  exp.areas      cchiest  exp.areas8
## 1        0      229 227.531392 0.0094791715 258.8134833
## 2        1      211 211.335581 0.0005328718 207.0507867
## 3        2       93  98.146299 0.2698460299  82.8203147
## 4        3       35  30.386730 0.7003800140  22.0854172
## 5        4        7   7.055946 0.0004435984   4.4170834
## 6        5        1   1.544051 0.1916982530   0.8129146

dada$cchiest8<-((dada$nb.areas-dada$exp.areas8)^2)/dada$exp.areas8
# H0: Data Poisson distributed Lambda = 0.8
chiest8<-sum(dada$cchiest8);chiest8

## [1] 13.86616

1-pchisq( chiest8, 5 )  # H0 Rejected

## [1] 0.01648234
```

27. To test if a coin is fair or not, it has independently been tossed 100,000 times (the outputs are a simple random sample), and 50,347 of them were heads. Should the fairness of the coin, as null hypothesis, be rejected when α = 0.1? (a) Apply a parametric test. By using a computer, plot the power function. (b) Apply the nonparametric chi-square goodness-of-fit test.

```
#a)
z0<-((50347/100000)-0.5)/(0.5*0.5/sqrt(100000));z0
```

```
## [1] 4.389241

2*(1-pnorm(z0)) # Two.sided H0 rejected

## [1] 1.137468e-05

# b)
prop.test(50347,100000, p=0.5 ,conf.level=0.9, correct=F)

##
##  1-sample proportions test without continuity correction
##
## data:  50347 out of 1e+05, null probability 0.5
## X-squared = 4.8164, df = 1, p-value = 0.02819
## alternative hypothesis: true p is not equal to 0.5
## 90 percent confidence interval:
##  0.5008693 0.5060706
## sample estimates:
##       p
## 0.50347

pp<-50347/100000;pp

## [1] 0.50347

chis <- (((50347-50000)^2)/50000)+(((100000-50347-50000)^2)/50000);chis

## [1] 4.81636

1-pchisq(chis,1) # Two.sided H0 rejected

## [1] 0.02819082
```