PROJECT 1

Abstract

This is the first of 4 projects for the module WST221 of the University of Pretoria. The scope of this project is dealing with Chapters 7 and 8 of Bain, L.J. and Engelhardt M. Introduction to Probability and Mathematical Statistics 1992.

Question 1

Problem Statement

Explain the construct and purpose of a simulation study.

1. What is the purpose of a simulation study?

The purpose of a simulation study is to obtain empirical results about the real world performance of statistical methods and models given a variety of changing parameters. Morris, White and Crowther (2019). These results will be obtained by creating data sets by means of random sampling within a distribution of choice. Morris, White and Crowther (2019). We may also wish to gain perspective on how the behavior of a model or method will change in time or for example by varying the sample size. Algebraic and analytic results are a good starting point, but rarely will the phenomena with which the statistical model are dealing be discrete, instead most natural and social phenomena are evolving through time and thus with a simulation study we wish to understand and then adjust our models to better represent the fluidity of the phenomena that we are try to model.

2. What makes a good simulation study?

In an ideal world we would setup a simulation study in such a way to accommodate for every possible eventuality, which is not feasible given time and complexity constraints. A good simulation study should instead be one which accommodates for the most widely defined parameters in such a way as to be applicable to the phenomena under simulation. Over optimizing parameters to fit discrete data sets will inevitable lead to terrible real world results. Also when designing a good simulation study one should take note of the algebraic results obtain and in essence aim to verify and or expand them to be more applicable in the desired real world setting of choice.

3. What should we be careful of when setting up a simulation study?

When setting up a simulation study there are a plethora of decisions that need to be made that will directly influence the success of the simulation. For example: what sample size will be most effective? What type of data should be simulated? How random is our random sample really? Does our simulation represent in a computational way the natural flow in which the phenomena will occur in the real world? We also need to be careful with the interpretation of the algebraic results as to not use the results in a way its not intended to be used. Careful consideration should be given to the data being used or generated for the model since all processes in nature generate their own data sets and our generated data set should closely match that of which is found in nature.

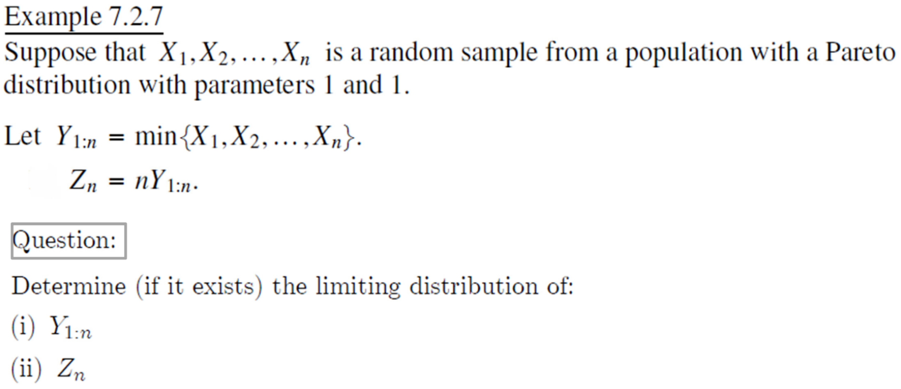
4. How large is large enough for a random sample?

This is a particularly hard question to answer in the absence of field specifics. Simulating a natural process which can be validated experimentally will be very different than simulating aspects of, for example, decision theory within a larger context of the financial markets. In essence the question becomes one of statistical significance with regards to the results obtained. There is also a case to be made for the relation between the size of the random sample used and the size of the population it is taken from and how we can accurately represent population characteristics in the choices of our sample size. From personal experience in the real world, one would always want to go as large as possible to account for as many variability as is possible. However this includes its own pitfalls as for example when using a large 30 year data set of intra-day GBP/USD prices, conventions like quote sizing becomes a major hurdle to ensure continuity in your analysis. Therefore careful consideration should be given when setting up and maintaining any sort of data set from which samples will be drawn.

Question 2

1. Problem Statement:

Design a simulation study to empirically show the result of Example 7.2.7 on Page 201 of Notes (2021*b*).



Example 7.2.7

Suppose that $X_1, X_2, \dots, X_n$ is a random sample from a population with a Pareto distribution with parameters 1 and 1.

Let $Y_{1:n} = \min\{X_1, X_2, \dots, X_n\}$.

$Z_n = nY_{1:n}$.

Question:

Determine (if it exists) the limiting distribution of:

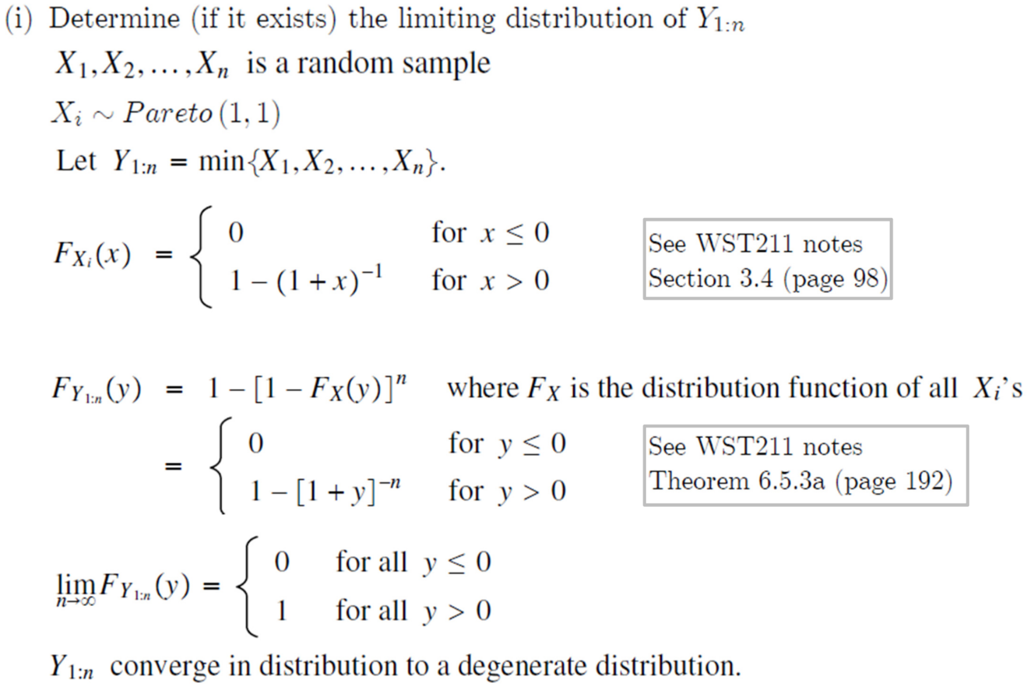
(i) $Y_{1:n}$

(ii) Z_n

Figure 1: Example 7.2.7

2. Analytic and Algebraic Results:

The following algebraic results are obtained in the Slides of video 6:(Slides, 2021)



(i) Determine (if it exists) the limiting distribution of $Y_{1:n}$

$X_1, X_2, \ldots, X_n$ is a random sample

$X_i \sim Pareto\,(1,1)$

Let $Y_{1:n} = \min\{X_1, X_2, \ldots, X_n\}$.

$$F_{X_i}(x) = \begin{cases} 0 & \text{for } x \le 0 \\ 1-(1+x)^{-1} & \text{for } x > 0 \end{cases}$$

See WST211 notes
Section 3.4 (page 98)

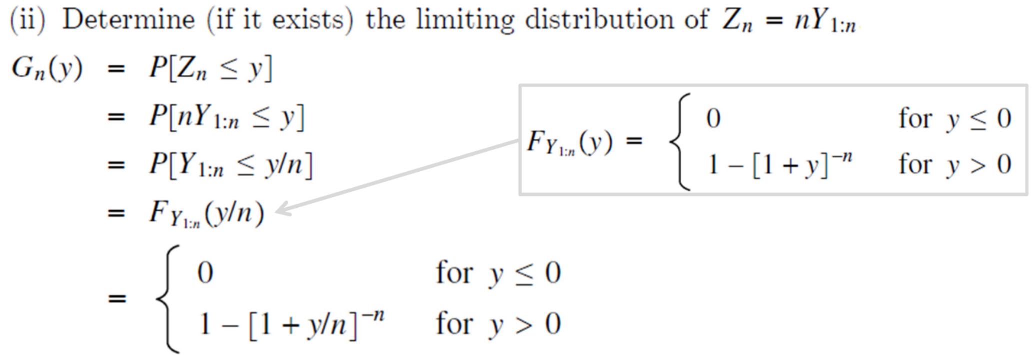
$$F_{Y_{1:n}}(y) = 1-[1-F_X(y)]^n \quad \text{where } F_X \text{ is the distribution function of all } X_i\text{'s}$$

$$= \begin{cases} 0 & \text{for } y \le 0 \\ 1-[1+y]^{-n} & \text{for } y > 0 \end{cases}$$

See WST211 notes
Theorem 6.5.3a (page 192)

$$\lim_{n\to\infty} F_{Y_{1:n}}(y) = \begin{cases} 0 & \text{for all } y \le 0 \\ 1 & \text{for all } y > 0 \end{cases}$$

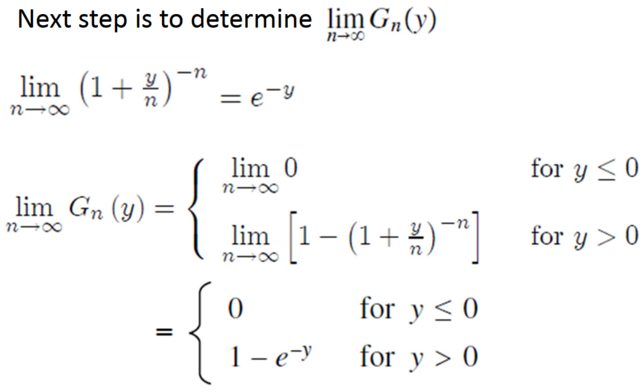
$Y_{1:n}$ converge in distribution to a degenerate distribution.



(ii) Determine (if it exists) the limiting distribution of $Z_n = nY_{1:n}$

$$\begin{aligned} G_n(y) &= P[Z_n \le y] \\ &= P[nY_{1:n} \le y] \\ &= P[Y_{1:n} \le y/n] \\ &= F_{Y_{1:n}}(y/n) \\ &= \begin{cases} 0 & \text{for } y \le 0 \\ 1-[1+y/n]^{-n} & \text{for } y > 0 \end{cases} \end{aligned}$$

$$F_{Y_{1:n}}(y) = \begin{cases} 0 & \text{for } y \le 0 \\ 1-[1+y]^{-n} & \text{for } y > 0 \end{cases}$$



Next step is to determine $\lim_{n\to\infty} G_n(y)$

$$\lim_{n\to\infty} \left(1+\tfrac{y}{n}\right)^{-n} = e^{-y}$$

$$\lim_{n\to\infty} G_n(y) = \begin{cases} \lim_{n\to\infty} 0 & \text{for } y \le 0 \\ \lim_{n\to\infty} \left[1-\left(1+\frac{y}{n}\right)^{-n}\right] & \text{for } y > 0 \end{cases}$$

$$= \begin{cases} 0 & \text{for } y \le 0 \\ 1-e^{-y} & \text{for } y > 0 \end{cases}$$

Do note that in these algebraic results obtained in the figures above, the CDF of the General Pareto Distribution are used. In this project the standard version of the Pareto Distribution will be used instead, which is: $F_X(x) = \begin{cases} 1-\left(\frac{\theta}{x}\right)^{\alpha} & x \ge \theta \\ 0 & x < \theta \end{cases}$ with $\theta :=$ Scale Parameter with $\theta > 0$ and $\alpha :=$ Shape

Parameter with $\alpha > 0$.

3. Methodology and Design of Simulation Study:

The goal here is to construct a simulation of the CDF of a Least Order Statistic as well as a linear combination of the Least Order Statistic with both from a random sample from a *PAR(1,1)* distribution. We will also study the limiting behavior of the model with emphasis on the sample size as it tends to infinity. We will conduct this simulation study in r:

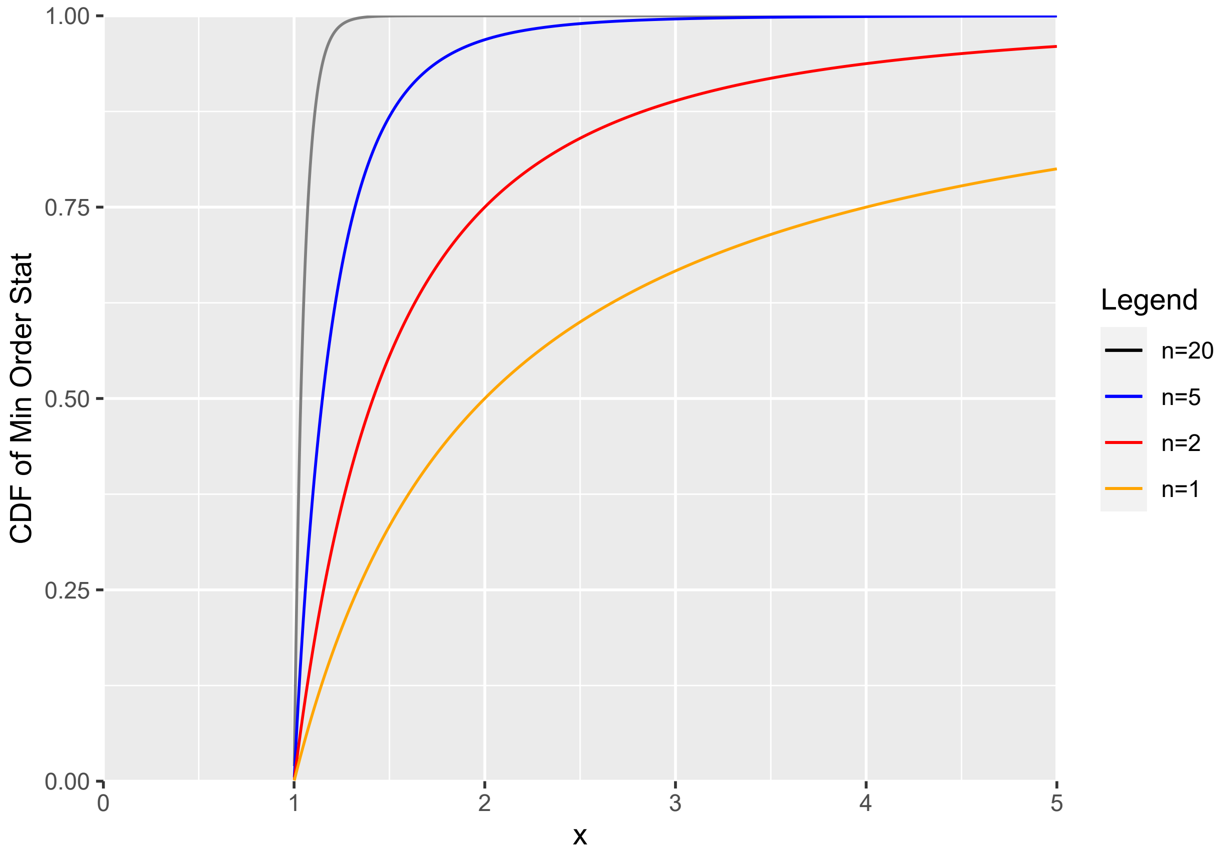
For question (i):

```r
q=5
qrel = q*1000
x=0
cdfmin=0
constrainedcdfmin = 0
for(k in 1:qrel){
  x[k] = c(k/1000)
}
y = c( 0, 5)
x = x[ x > y[1] & x <= y[2]]

cdfmin = function(r){
  cdmin=0
  for(i in 1:qrel){
    cdmin[i] = 1-(1/(i/1000))^r
    if (between(cdmin[i], 0.0001, q)){
      constrainedcdfmin[i] = cdmin[i]
    }
  }
  return (constrainedcdfmin)
}

colors <- c("n=20" = "black", "n=5" = "blue", "n=2" = "red", "n=1" = "orange")
df = data.frame(x, cdfmin(20), cdfmin(5), cdfmin(2), cdfmin(1))

ggplot(df, aes(x)) +
  geom_line(aes(y=cdfmin(20), color = "n=100")) +
  geom_line(aes(y=cdfmin(5), color = "n=5")) +
  geom_line(aes(y=cdfmin(2), color = "n=2")) +
  geom_line(aes(y=cdfmin(1), color = "n=1")) +
  labs(x = "x",
       y = "CDF of Min Order Stat",
       color = "Legend") +
  scale_color_manual(values = colors) +
  expand_limits(x = 0, cdfmin = 0) +
  scale_x_continuous(expand = c(0, 0), limits = c(0,5), oob = scales::censor) +
  scale_y_continuous(expand = c(0, 0), limits = c(0,1), oob = scales::censor)
```



4. Discussion fo Results Obtained:

Now we can see that as n increases we steadily approach a step function between y=0 and y=1 on the graph, but in order to validate this we shall investigate the limiting behavior of the CDF of the smallest order statistic as n goes to infinity and therefore: $\lim_{n\to\infty} F_{Y_{1:n}} = \lim_{n\to\infty} 1-(\frac{1}{x})^n$ This approach is laid out in Definition 7.2.1 on page 196 Notes (2021b).

```
#Implementing Ryacas Packages for symbolic math
n <- ysym("n")
x <- ysym("x")
lim(1-x^-n, n, Inf) #limit of the CDF of the smallest order statistic
```

```
## y: 1-x^(-Infinity)
```

$\therefore y = 1 - x^{-\infty} = 1 \ \forall \, n \in \mathbb{R}$

$\Rightarrow \lim_{n\to\infty} 1-(\frac{1}{x})^n = \begin{cases} 1 & x \geq 1 \\ 0 & x < 1 \end{cases}$

Now this distribution is not a valid CDF since it is not continuous from the right as is required by Theorem 2.2.1b on page 33 of Notes (2021a). However from Definition 7.2.1 on page 196 Notes (2021b) we see that this discontinuity is where the distribution function of the Least Order Statistic is not equal to a valid CDF or the CDF of the Limiting Distribution, but we only need the limit of the Least Order Statistic's distribution to be equal to the valid Limiting Distribution where the Limiting Distribution are continuous. Thus we have the limit of the CDF of the Least Order Statistic to be equal to the Limiting Distribution for all points where the Limiting Distribution is continuous.

Also from Definition 7.2.2 on page 197 Notes (2021b) we see that that we have constructed a Limiting Distribution for the CDF of the Least Order Statistic from a $PAR(1,1)$ distribution that is degenerate at the point $c = 1$ which also implies that $P(X = c) = 1$

Thus by Definition 7.2.3 on page 199 Notes (2021b) we have stochastic convergence to the constant c=1.

For question (ii):

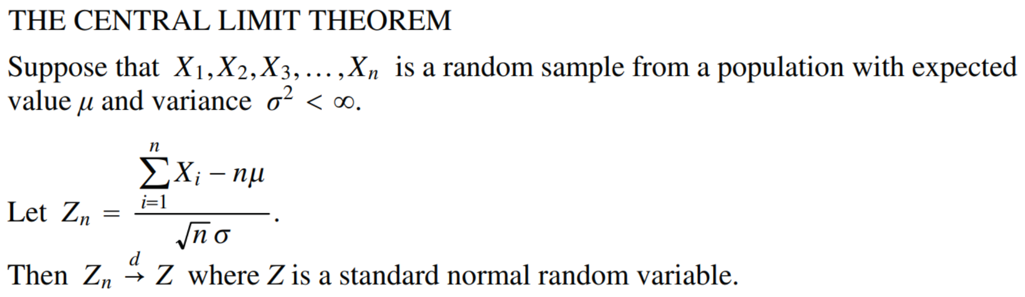
We were instructed on Discussion Board that it is not required to complete this sub section of the example.

Question 3

1. Problem Statement:

Design a simulation study to empirically show the result of Theorem 7.3.2 on page 204 of Notes (2021b), based on data from a Gamma distribution.

Theorem 7.3.2:



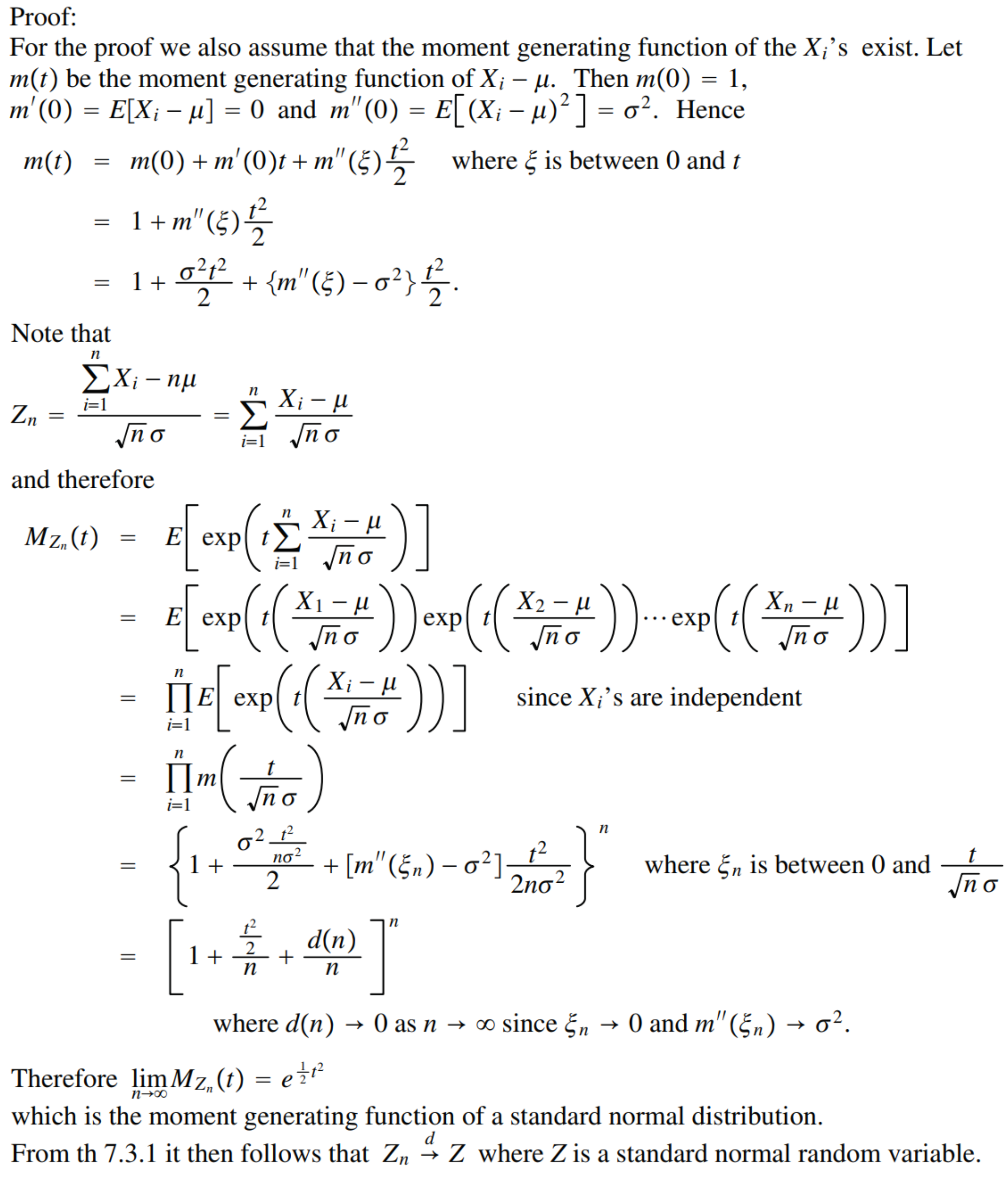
THE CENTRAL LIMIT THEOREM

Suppose that $X_1, X_2, X_3, \ldots, X_n$ is a random sample from a population with expected value μ and variance $\sigma^2 < \infty$.

Let $Z_n = \dfrac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma}$.

Then $Z_n \overset{d}{\to} Z$ where Z is a standard normal random variable.

2. Analytic and Algebraic Results:



Proof:
For the proof we also assume that the moment generating function of the X_i's exist. Let $m(t)$ be the moment generating function of $X_i - \mu$. Then $m(0) = 1$, $m'(0) = E[X_i - \mu] = 0$ and $m''(0) = E\left[(X_i - \mu)^2\right] = \sigma^2$. Hence

$$\begin{aligned} m(t) &= m(0) + m'(0)t + m''(\xi)\frac{t^2}{2} \qquad \text{where } \xi \text{ is between 0 and } t \\ &= 1 + m''(\xi)\frac{t^2}{2} \\ &= 1 + \frac{\sigma^2 t^2}{2} + \{m''(\xi) - \sigma^2\}\frac{t^2}{2}. \end{aligned}$$

Note that

$$Z_n = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} = \sum_{i=1}^{n} \frac{X_i - \mu}{\sqrt{n}\,\sigma}$$

and therefore

$$\begin{aligned} M_{Z_n}(t) &= E\left[\exp\left(t\sum_{i=1}^{n}\frac{X_i - \mu}{\sqrt{n}\,\sigma}\right)\right] \\ &= E\left[\exp\left(t\left(\frac{X_1 - \mu}{\sqrt{n}\,\sigma}\right)\right)\exp\left(t\left(\frac{X_2 - \mu}{\sqrt{n}\,\sigma}\right)\right)\cdots\exp\left(t\left(\frac{X_n - \mu}{\sqrt{n}\,\sigma}\right)\right)\right] \\ &= \prod_{i=1}^{n} E\left[\exp\left(t\left(\frac{X_i - \mu}{\sqrt{n}\,\sigma}\right)\right)\right] \qquad \text{since } X_i\text{'s are independent} \\ &= \prod_{i=1}^{n} m\left(\frac{t}{\sqrt{n}\,\sigma}\right) \\ &= \left\{1 + \frac{\sigma^2 \frac{t^2}{n\sigma^2}}{2} + [m''(\xi_n) - \sigma^2]\frac{t^2}{2n\sigma^2}\right\}^n \qquad \text{where } \xi_n \text{ is between 0 and } \frac{t}{\sqrt{n}\,\sigma} \\ &= \left[1 + \frac{\frac{t^2}{2}}{n} + \frac{d(n)}{n}\right]^n \end{aligned}$$

where $d(n) \to 0$ as $n \to \infty$ since $\xi_n \to 0$ and $m''(\xi_n) \to \sigma^2$.

Therefore $\lim_{n\to\infty} M_{Z_n}(t) = e^{\frac{1}{2}t^2}$
which is the moment generating function of a standard normal distribution.
From th 7.3.1 it then follows that $Z_n \overset{d}{\to} Z$ where Z is a standard normal random variable.

The proof, from Notes (2021b), for Theorem 7.3.2 clearly shows the result of the theorem holds for large values of n.

3. Methodology and Design of Simulation Study:

The goal here is to construct a simulation of the algebraic results obtained from Theorem 7.3.2. I decided to create 3 different random samples each from a *~GAM(2,1)* distribution with increasing number of samples in each to visually be able to inspect if Z_n does have a *~N(0,1)* distribution for n getting ever larger.

For each of the 3 random samples I constructed a plot of the density function of Z_n vs that of a *~N(0,1)* distributed random sample in order to visually gain perspective on the effect that the changing sample size has on the density function of Z_n. The following code in was constructed to that effect:

```r
zn1=0
zn2=0
zn3=0

#------------------------------------------------------------
set.seed(1492)
x1 = rgamma(500, 2, 1)
x1_mean = mean(x1)
x1_sd = sd(x1)
for(i in 1:500){
  zn1[i] = (sum(rgamma(i, 2, 1)) - i*x1_mean)/(sqrt(i)*x1_sd)
}
zn1_mean = mean(zn1)
#------------------------------------------------------------
set.seed(1492)
x2 = rgamma(1000, 2, 1)
x2_sum = sum(x2)
x2_mean = mean(x2)
x2_sd = sd(x2)
for(i in 1:1000){
  zn2[i] = (sum(rgamma(i, 2, 1)) - i*x2_mean)/(sqrt(i)*x2_sd)
}
zn2_mean = mean(zn2)
#------------------------------------------------------------
set.seed(1492)
x3 = rgamma(1800, 2, 1)
x3_sum = sum(x3)
x3_mean = mean(x3)
x3_sd = sd(x3)
for(i in 1:1800){
  zn3[i] = (sum(rgamma(i, 2, 1)) - i*x3_mean)/(sqrt(i)*x3_sd)
}
zn3_mean = mean(zn3)
#------------------------------------------------------------
colors1 = c("Z500" = "black", "N(0,1)" = "red")
set.seed(1492)
ggplot(data.frame(x=zn1), aes(x, color="black")) +
  geom_density() +
  geom_function(fun = dnorm, colour = "red") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  geom_vline(xintercept = zn1_mean, linetype = "dashed", color = "black") +
  annotate("text", x=zn1_mean+0.2, y=0.1, angle=90,
           label=paste("Mean of Z500")) +
  annotate("text", x=0-0.2, y=0.1, angle=90, color = "red",
           label=paste("Mean of N(0,1)")) +
  labs(title = "Simulation of Zn for n = 500 vs. ~N(0,1)",
       caption = "Figure 1",
       x = "x",
       y = "PDF of Z500",
       color = "Legend") +
  scale_color_manual(values = colors1) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
     plot.caption = element_text(hjust = 0.5, colour = "blue"))
```

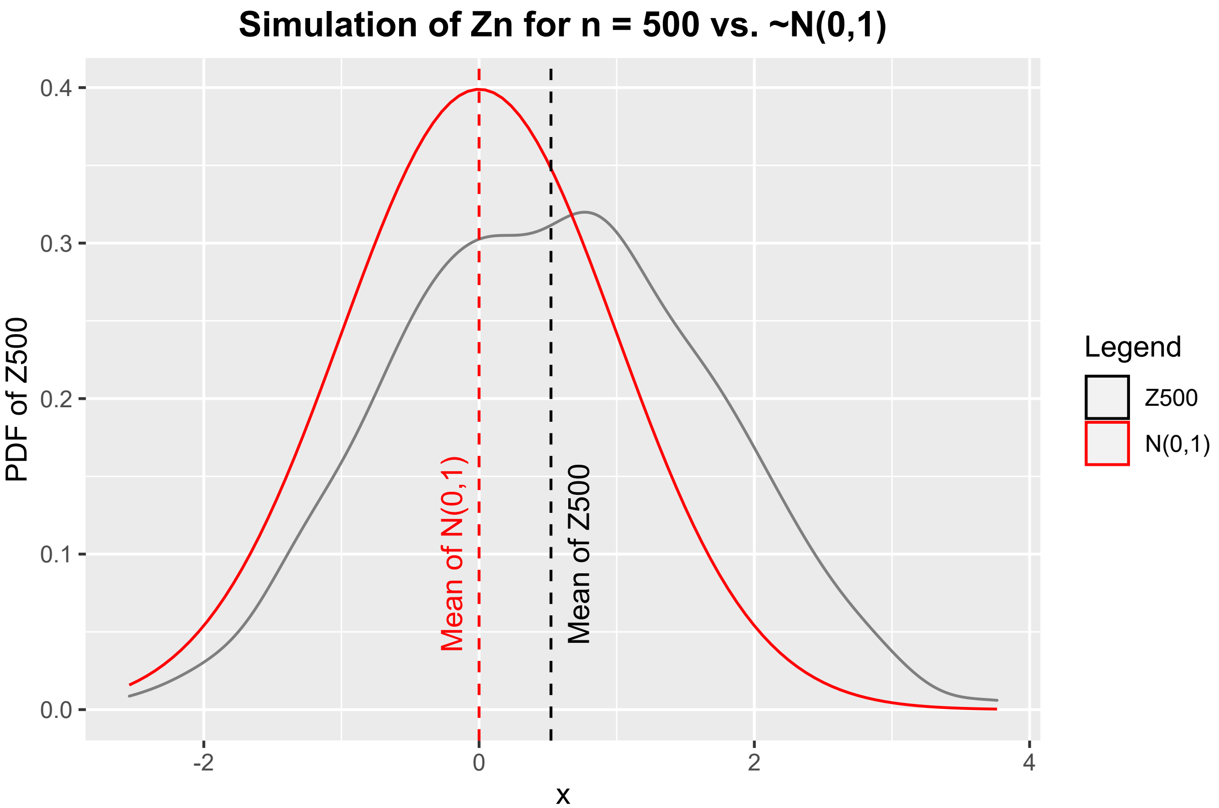


Figure 1

```
colors2 = c("Z1000" = "black", "N(0,1)" = "red")
set.seed(1492)
ggplot(data.frame(x=zn2), aes(x, color="black")) +
  geom_density() +
  geom_function(fun = dnorm, colour = "red") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  geom_vline(xintercept = zn2_mean, linetype = "dashed", color = "black") +
  annotate("text", x=zn2_mean+0.2, y=0.1, angle=90,
           label=paste("Mean of Z1000")) +
  annotate("text", x=0-0.2, y=0.1, angle=90, color = "red",
           label=paste("Mean of N(0,1)")) +
  labs(title = "Simulation of Zn for n = 1000 vs. ~N(0,1)",
       caption = "Figure 2",
       x = "x",
       y = "PDF of Z1000",
       color = "Legend") +
  scale_color_manual(values = colors2) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
     plot.caption = element_text(hjust = 0.5, colour = "blue"))
```

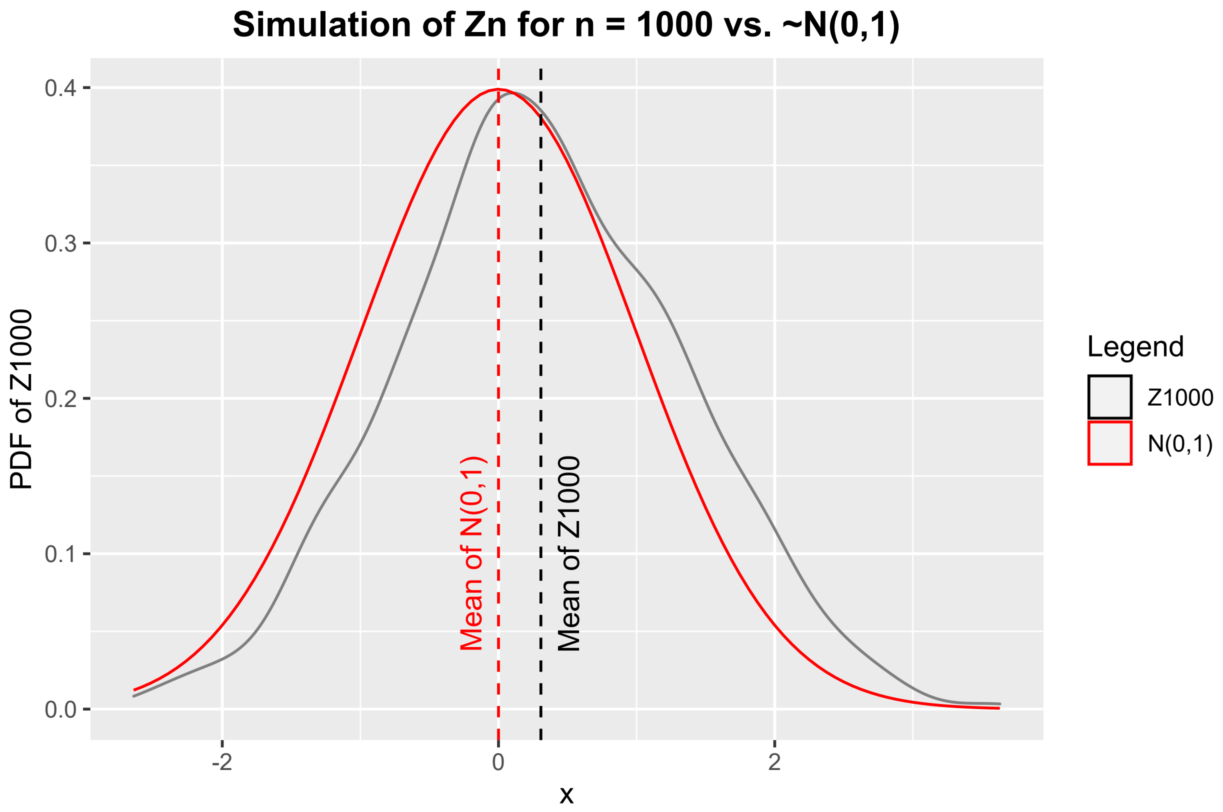


Figure 2

```
colors3 = c("Z1800" = "black", "N(0,1)" = "red")
set.seed(1492)
ggplot(data.frame(x=zn3), aes(x, color="black")) +
  geom_density() +
  geom_function(fun = dnorm, colour = "red") +
  geom_vline(xintercept = 0, linetype = "dashed", color = "red") +
  geom_vline(xintercept = zn3_mean, linetype = "dashed", color = "black") +
  annotate("text", x=zn3_mean-0.2, y=0.1, angle=90,
           label=paste("Mean of Z1800")) +
  annotate("text", x=0+0.2, y=0.1, angle=90, color = "red",
           label=paste("Mean of N(0,1)")) +
  labs(title = "Simulation of Zn for n = 1800 vs. ~N(0,1)",
       caption = "Figure 3",
       x = "x",
       y = "PDF of Z1800",
       color = "Legend") +
  scale_color_manual(values = colors3) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        plot.caption = element_text(hjust = 0.5, colour = "blue"))
```

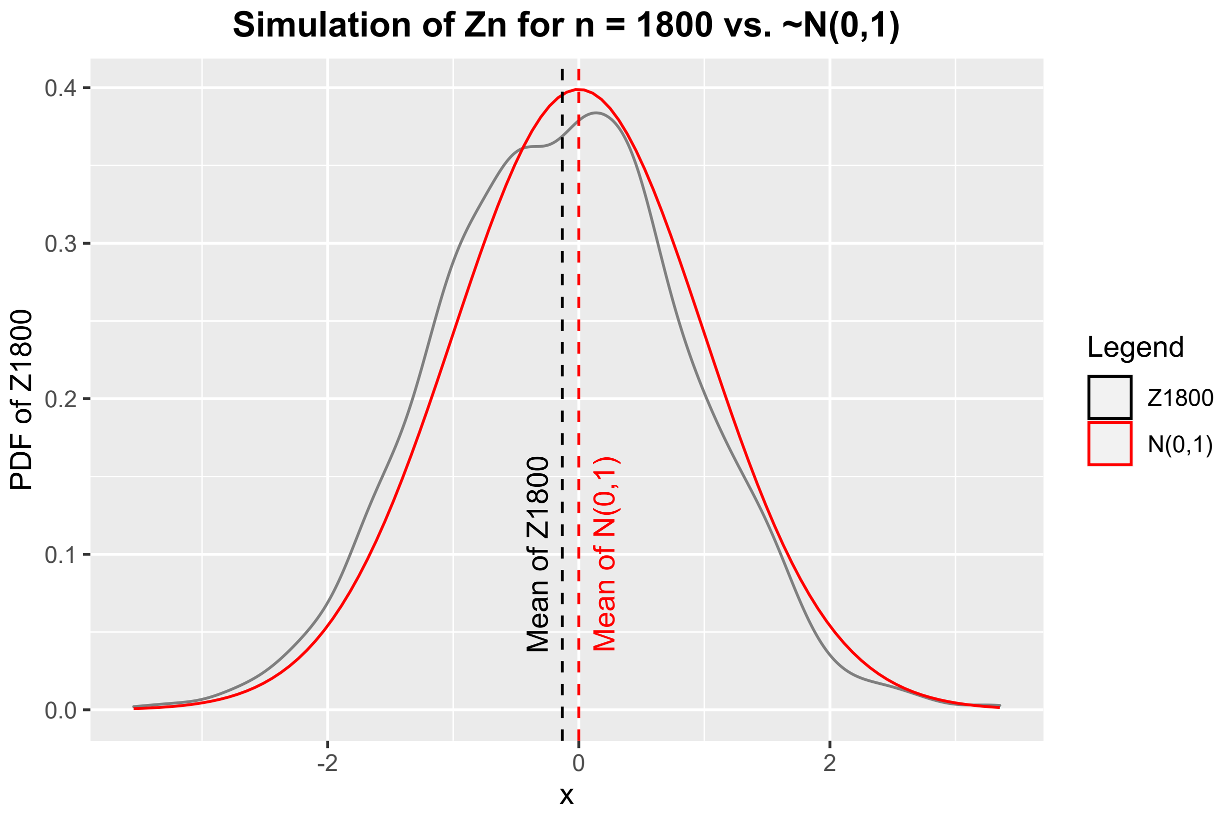


Figure 3

4. Discussion fo Results Obtained:

As we look through the progression of the simulation from Figure 1 to Figure 3 for ever increasing values of the sample size n, we can clearly see that Z_n is converging in distribution to a *N(0,1)* distributed random variable since we can clearly see how the density function of Z_n is approaching the density function of a *N(0,1)* distributed random variable as n grows ever larger.

Thus our simulation confirmed the result stated in Theorem 7.3.2 and Z_n will infact converge in distribution to the *N(0,1)* distribution.

Question 4

1. Problem Statement:

Suppose we have two machines in a factory that manufactures drink cans. It would be important for us to make sure that the variation in the can sizes are not too big. Also we have two independent samples from a Normal distribution with the same mean and variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 3$. Use a simulation study to get the distribution of $S_1^2 + S_2^2$, and use it to calculate the probability $P(S_1^2 + S_2^2 < 4.2)$. Include the theoretical methodology and design, the code and results (include graphics) as well as a discussion of the results.

2. Analytic and Algebraic Results:

Since $\frac{(n-1)S_1^2}{\sigma_1^2}$ ~χ^2(n-1) distributed and $\frac{(n-1)S_2^2}{\sigma_2^2}$ ~χ^2(n-1)

3. Methodology and Design of Simulation Study:

The goal with this simulation is to first construct two independent random samples with parameters as indicated in the problem statement. We will them simulate, for increasing values of n, the sample variance as well as the sum of the two independent sample variances. Then we will plot this result to be able to graphically assess if we can discern the distribution that this linear combination of sample variances will have. When we have established this distribution we will then proceed to calculate $P(S_1^2 + S_2^2 < 4.2)$ as is required in the problem statement.

```r
n=500
sampvar1=0
sampvar2=0
sampleSum=0
samplemean1=0
samplemean2=0
#
set.seed(1495)
for(i in 2:n){
  samplemean1[i] = (sum(rnorm(i, 1, 1)))/i
  samplemean2[i] = (sum(rnorm(i, 1, sqrt(3))))/i
  sampvar1[i] = (sum((rnorm(i, 1, 1) - samplemean1[i])^2))/(i-1)
  sampvar2[i] = (sum((rnorm(i, 1, sqrt(3)) - samplemean2[i])^2))/(i-1)

  sampleSum[i] = sampvar1[i] + sampvar2[i]
}

sampleSumMean = mean(sampleSum)
sampleSumVar = var(sampleSum)

colors1 = c("S1^2-S2^2" = "black", " ~N(4.10,1.18)" = "red")
set.seed(1495)
ggplot(data.frame(x=sampleSum), aes(x, color = "black")) +
  geom_density() +
  geom_function(fun = dnorm, args = list(mean = sampleSumMean,
          sd = sampleSumVar), colour = "red") +
  geom_vline(xintercept = sampleSumMean, linetype = "dashed", color = "black") +
  labs(title = "Simulation S1^2+S2^2 vs. ~N(4.10,1.18)",
       caption = "Figure 1",
       x = "x",
       y = "Density of S1^2+S2^2",
       color = "Legend") +
  scale_color_manual(values = colors1) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        plot.caption = element_text(hjust = 0.5, colour = "blue"))
```

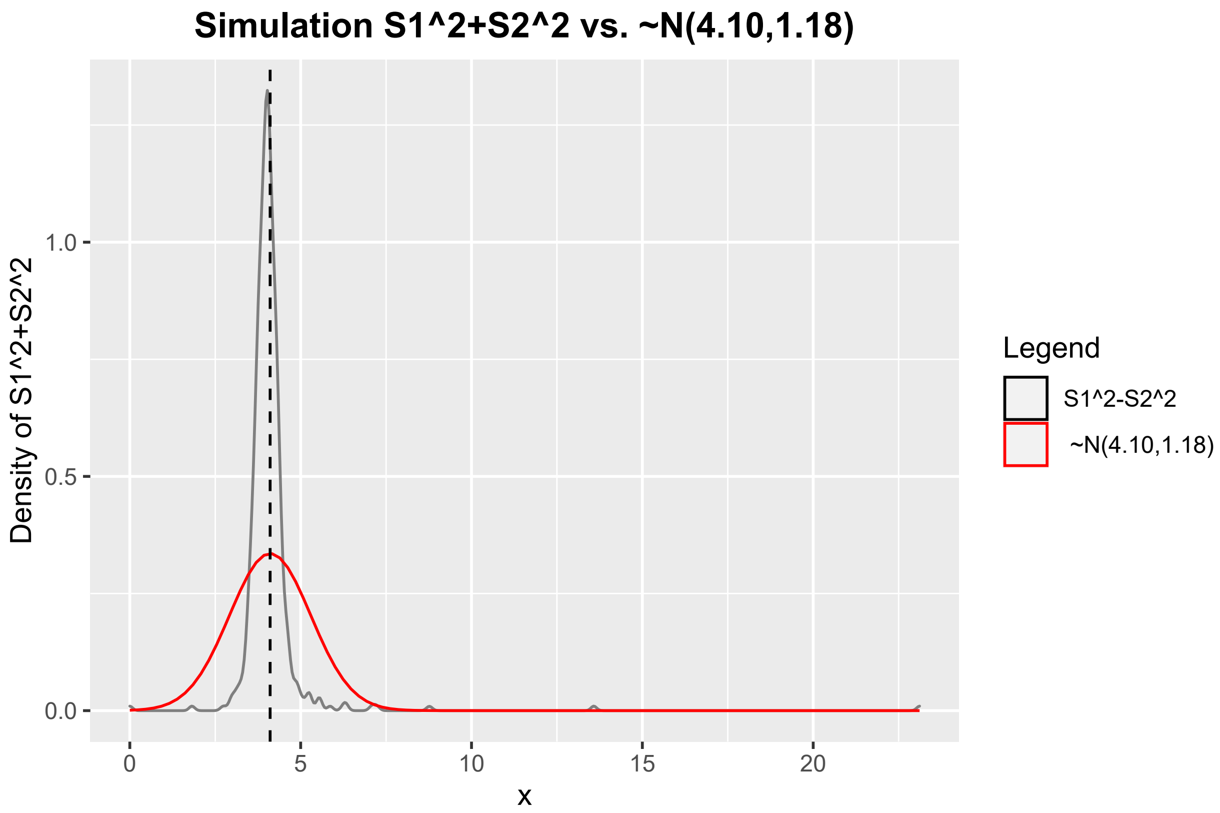


Figure 1

4. Discussion fo Results Obtained:

However, it does seem that the distribution of the sum of the two sample variances will be normally distributed since we can see that its center is very close to that of a normally distributed random variable and the shape is also conforming to that.

References

Morris, T.P., I.R. White and M.J. Crowther. 2019. "Using simulation studies to evaluate statistical methods." *Statistics in Medicine* 38(11):2074–2102.