MA 354: Data Analysis I – Fall 2021
Homework 3:

Complete the following opportunities to use what we've talked about in class. These questions will be graded for correctness, communication and succinctness. Ensure you show your work and explain your logic in a legible and refined submission.

0. **Complete weekly diagnostics.**

1. Consider data originally from a study of the nesting horseshoe crabs (Brockmann, 1996). Each female crab in the study had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing nearby her. Explanatory variables thought possibly to affect this included the female crab's color, spine condition, weight, and carapace width. The response outcome for each female crab is her number of satellites. The sample is

Number of Satellites	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of Observations	62	16	9	19	19	15	13	4	6	3	3	1	1	1	1	1

It is believed that the distribution of the number of satellites for a female crab is distributed Poisson(λ) where the parameter λ is of interest.

(a) Calculate the method of moments estimator for λ.
Solution:

```
library(nleqslv)
#put data into the dataframe
crab.dat <- data.frame(x=c(rep(0,62), rep(1, 16), rep(2, 9), rep(3, 19),
                           rep(4, 19), rep(5, 15), rep(6, 13), rep(7, 4),
                           rep(8, 6), rep(9, 3), rep(10, 3), 11, 12,
                           13, 14, 15))



#MOM function
momfunc <- function(par, data){
  population.mean <- par
  sample.mean <- mean(data)
  return(sample.mean-population.mean)
}

answer <- nleqslv(x=1,
        fn = momfunc,
        data=crab.dat$x)

answer$x
```

```
## [1] 2.977011
```

(b) Find the maximum likelihood estimator for λ.
Solution:

```
LLfunc <- function(par, data){
  answer <- sum(dpois(x=data, lambda = par, log=T))
  -answer
}
```

```r
answer <- optim(par = 1,
      fn = LLfunc,
      data = crab.dat$x,
      method = "Brent",
      lower = 0,
      upper = 100)

answer
```

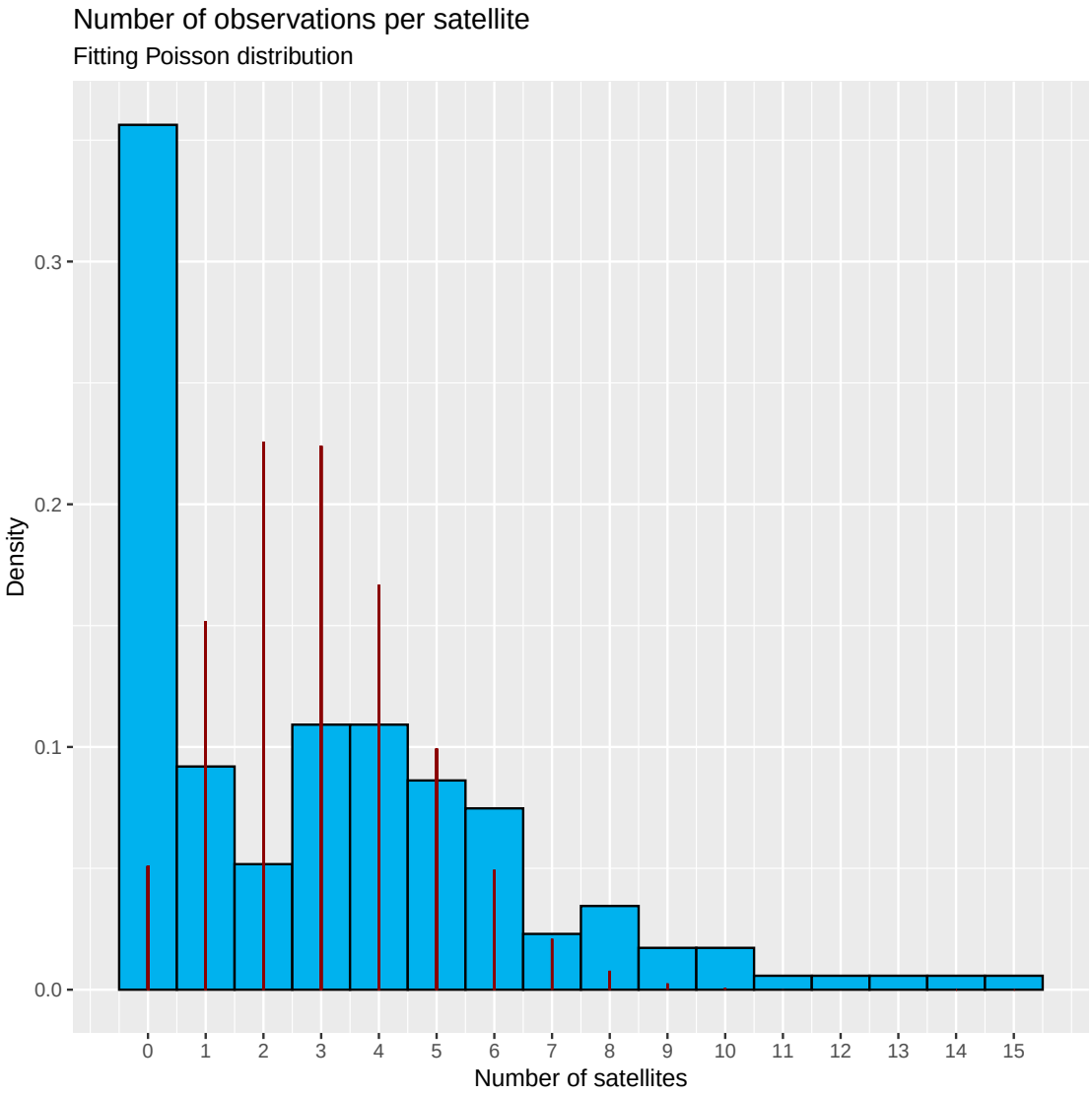
```
## $par
## [1] 2.977011
##
## $value
## [1] 505.4901
##
## $counts
## function gradient
##       NA       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

(c) Plot the data with the Poisson distribution fit with the MLE estimates. How well does the distribution fit the data?

```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
crab.dat.plot <- crab.dat %>%
  mutate(f = dpois(x=x, lambda=2.977011))


sats <- 0:15
ggplot(crab.dat.plot, aes(x=x))+
  geom_histogram(aes(y=..density..), fill="deepskyblue2", color="black",
                 binwidth=1)+
  geom_linerange(aes(ymin=0, ymax=f), color="dark red")+
  scale_x_continuous("Number of satellites",
                     labels = as.character(sats), breaks = sats)+
  labs(y="Density",
       title="Number of observations per satellite",
       subtitle="Fitting Poisson distribution")
```



(d) Let's try another distribution - the zero-inflated Poisson distribution. Now, it is believed that the distribution of the number of satellites for a female crab is distributed $\text{Poisson}_0(\lambda, \sigma)$ where the parameters λ and σ are of interest. Find the method of moments estimators for both σ and λ.

$$f_X(x|\lambda,\sigma) = (1-\sigma)\frac{\lambda^x e^{-\lambda}}{x!}I(x \geq 1) + (\sigma + (1-\sigma)e^{-\lambda})I(x=0)$$

Hint: Noticing that this is a function of the Poisson PMF and rewriting it will help remarkably.
Solution:

```
zip.pois <- function(par, data){
  lambda <- par[1]
  pi <- par[2]

  EX1 <- (1-pi)*lambda
  EX2 <- (1-pi)*(lambda^2+lambda)

  xbar1 <- mean(data)
```

```
  xbar2 <- mean(data^2)

  c(EX1-xbar1, EX2-xbar2)
}


nleqslv(x = c(4.5,0.34), #best guess at parameter
        fn = zip.pois,
        data = crab.dat$x)
```

```
## $x
## [1] 5.4633205 0.4550912
##
## $fvec
## [1] 3.674039e-11 3.937686e-10
##
## $termcd
## [1] 1
##
## $message
## [1] "Function criterion near zero"
##
## $scalex
## [1] 1 1
##
## $nfcnt
## [1] 7
##
## $njcnt
## [1] 1
##
## $iter
## [1] 7
```

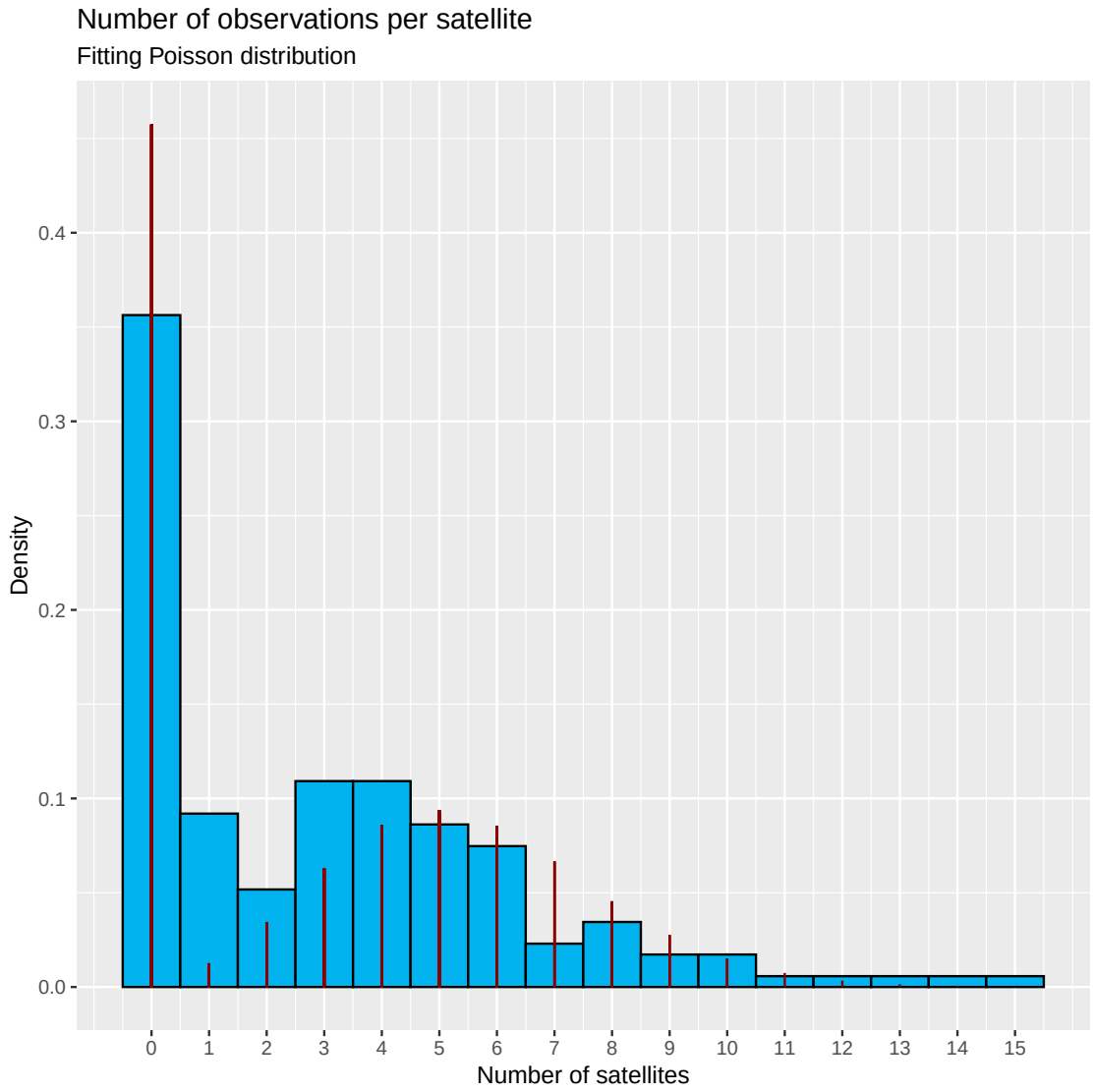
Let's graph the results to see how accurate this approach was!

```
library("VGAM")
```

```
## Loading required package:  stats4
## Loading required package:  splines
##
## Attaching package:  'VGAM'
## The following object is masked from 'package:tidyr':
##
##     fill
```

```
crab.dat.zim <- crab.dat %>%
          mutate(f = dzipois(x=x, lambda=5.4633205, pstr0=0.4550912))

ggplot(crab.dat.zim, aes(x=x))+
          geom_histogram(aes(y=..density..), fill="deepskyblue2", color="black",
                         binwidth=1)+
          geom_linerange(aes(ymin=0, ymax=f), color="dark red")+
          scale_x_continuous("Number of satellites",
                             labels = as.character(sats), breaks = sats)+
```

```
        labs(y="Density",
             title="Number of observations per satellite",
             subtitle="Fitting Poisson distribution")
```



It would appear that MOM inflates zero too much, and underestimates the other values.

(e) Find the maximum likelihood estimator for λ and σ.

```
library(Rfast)
```

```
## Loading required package:  Rcpp
## Loading required package:  RcppZiggurat
##
## Attaching package:  'Rfast'
## The following object is masked from 'package:VGAM':
##
##     Rank
## The following object is masked from 'package:dplyr':
##
##     nth
```

```
## The following objects are masked from 'package:purrr':
##
##     is_integer, transpose
```

```
zip.mle(crab.dat$x)
```

```
## $iters
## [1] 5
##
## $loglik
## [1] -389.4734
##
## $param
##    lambda        pi
## 4.5774500 0.3496354
```

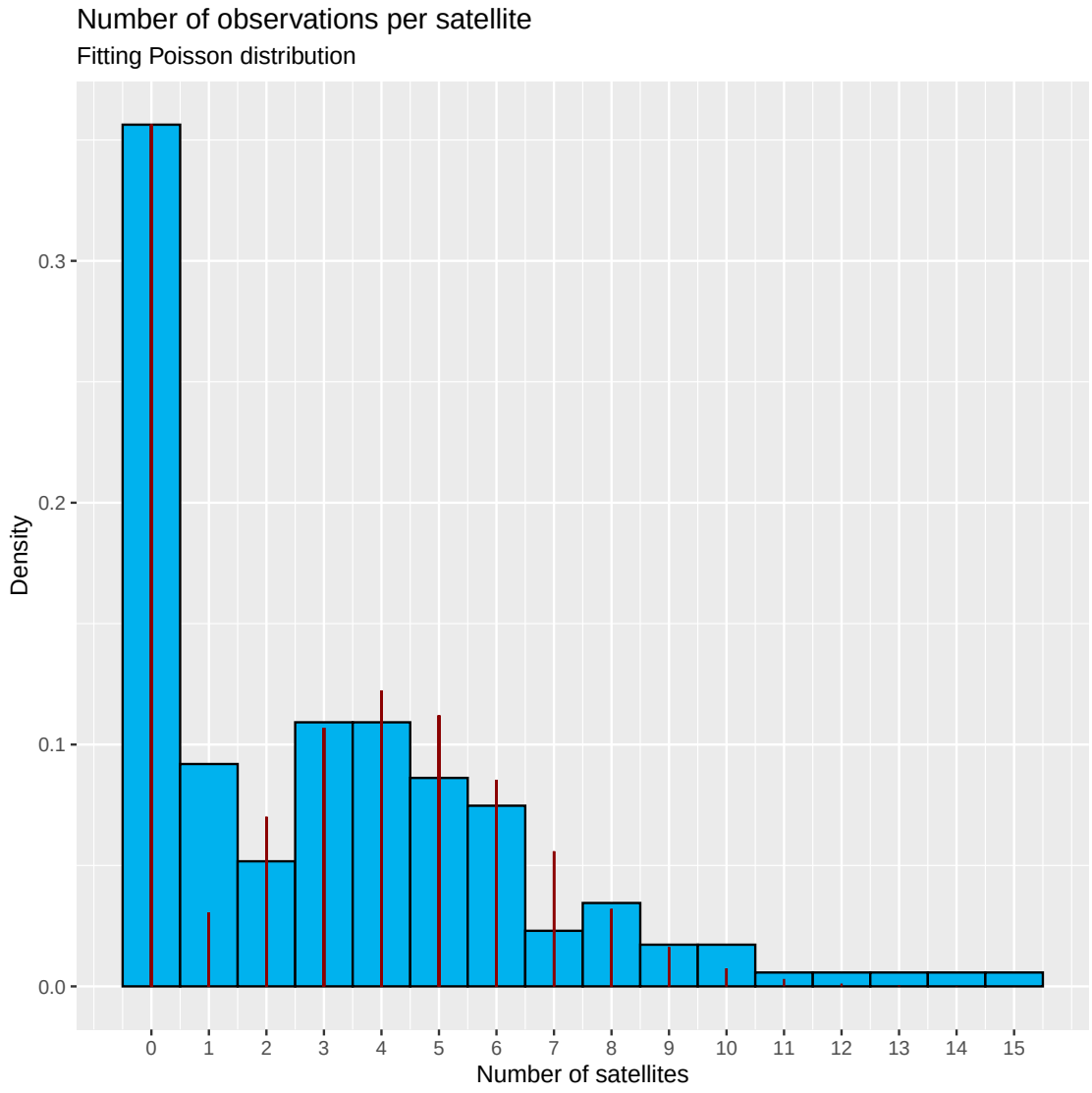
```
#MLE
dpois.ll<-function(par, data, neg=T){
  lambda <- par[1]
  pi <- par[2]
  ll <- sum(dzipois(x=data, lambda=lambda, pstr0=pi, log=T))
  ifelse(neg, -ll, ll)
}

optim(par = c(1,0),
      fn = dpois.ll,
      data=crab.dat$x)
```

```
## $par
## [1] 4.5770922 0.3496381
##
## $value
## [1] 389.4734
##
## $counts
## function gradient
##       71       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

(f) Plot a histogram of the data with the zero-inflated Poisson distribution fit with the MLE estimates. How well does the distribution fit the data?

```
crab.dat.zim <- crab.dat %>%
  mutate(f = dzipois(x=x, lambda=4.5770923, pstr0=0.3496381))

ggplot(crab.dat.zim, aes(x=x))+
  geom_histogram(aes(y=..density..), fill="deepskyblue2", color="black",
                 binwidth=1)+
  geom_linerange(aes(ymin=0, ymax=f), color="dark red")+
  scale_x_continuous("Number of satellites",
                     labels = as.character(sats), breaks = sats)+
```

```
labs(y="Density",
     title="Number of observations per satellite",
     subtitle="Fitting Poisson distribution")
```



2. The time to death for rats injected with a toxic substance, denoted by Y (measured in days), follows an exponential distribution with $\lambda = 1/5$. That is,

$$Y \sim \text{exponential}(\lambda = 1/5).$$

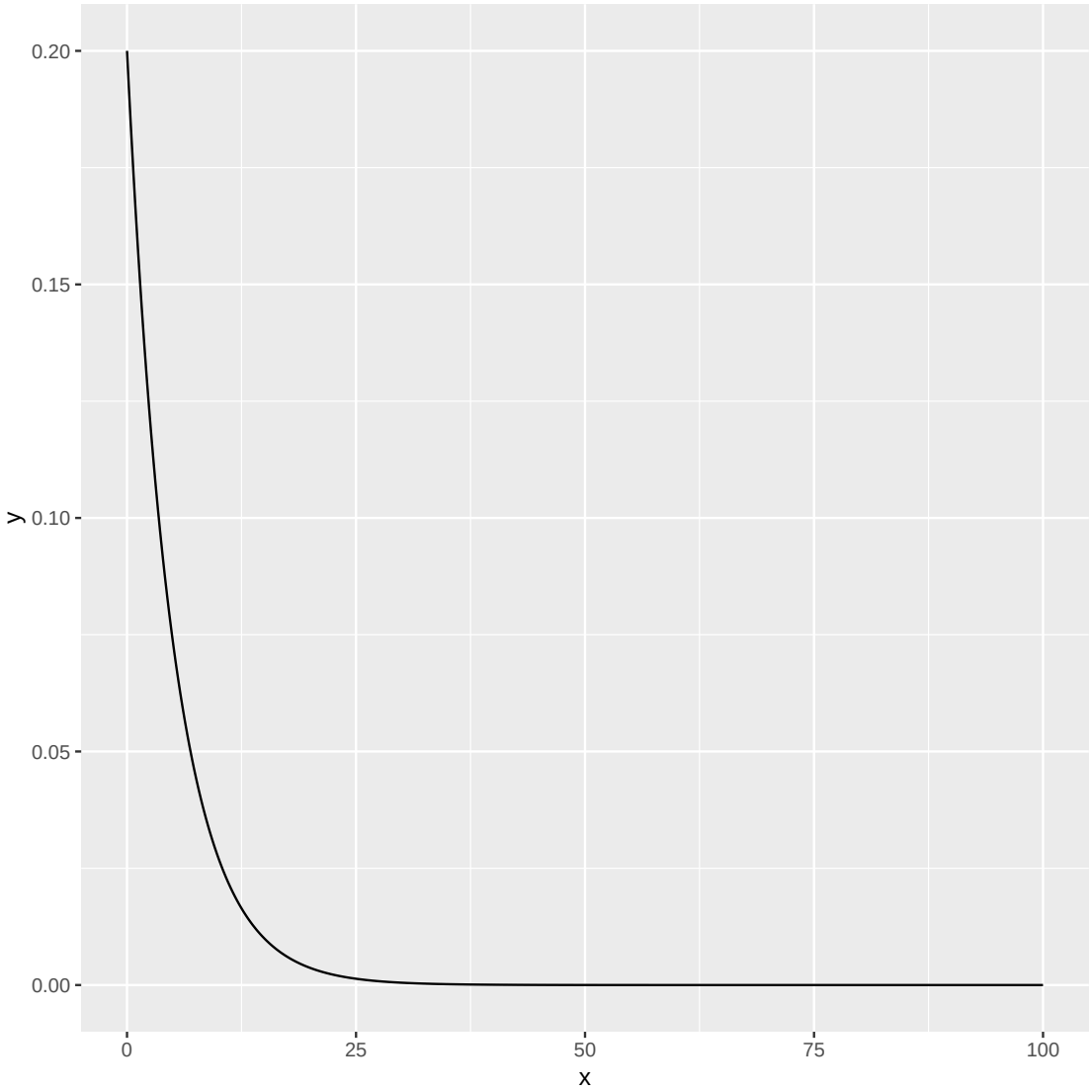
This is the population distribution. It describes the time to death for all individual rats in the population.

The **exponential distribution** serves as a very good model for measurements like waiting times or lifetimes.

$$\begin{aligned}
\lambda &\in \mathbb{R}^+ && \textbf{[Parameters]} \\
\mathcal{X} &= \{\omega : \omega \in \mathbb{R}^+\} && \textbf{[Support]} \\
f_X(x|\mu,\sigma) &= \lambda e^{-\lambda x}\ I(x \in \mathbb{R}^+) && \textbf{[PDF]} \\
F_X(x|\mu,\sigma) &= \left(1 - e^{-\lambda x}\right) I(x \in \mathbb{R}^+ && \textbf{[CDF]} \\
F_X^{-1}(p|\mu,\sigma) &= \frac{-\ln(1-p)}{\lambda} && \textbf{[Inverse CDF]} \\
E(X) &= \frac{1}{\lambda} && \textbf{[Expected Value]} \\
var(X) &= \frac{1}{\lambda^2} && \textbf{[Population Variance]}
\end{aligned}$$

(a) Plot the exponential($\lambda = 1/5$) population distribution.

```
x<-seq(0, 100, 0.1)
y<-dexp(x, 1/5)
ggdat<-data.frame(x,y)

ggplot(ggdat, aes(x=x, y=y))+
  geom_line()
```



(b) Mathematical statisticians can show the exact sampling distributions of $\bar{Y}$ are gamma; i.e.,

$$\bar{Y} \sim \text{gamma}(\alpha = n, \beta = \frac{1}{n\lambda}).$$

The gamma distribution is described below.

$$\alpha > 0, \beta > 0 \qquad \textbf{[Parameters]}$$

$$\mathcal{X} = (0, \infty) \qquad \textbf{[Support]}$$

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} I(x > 0) \qquad \textbf{[PDF]}$$

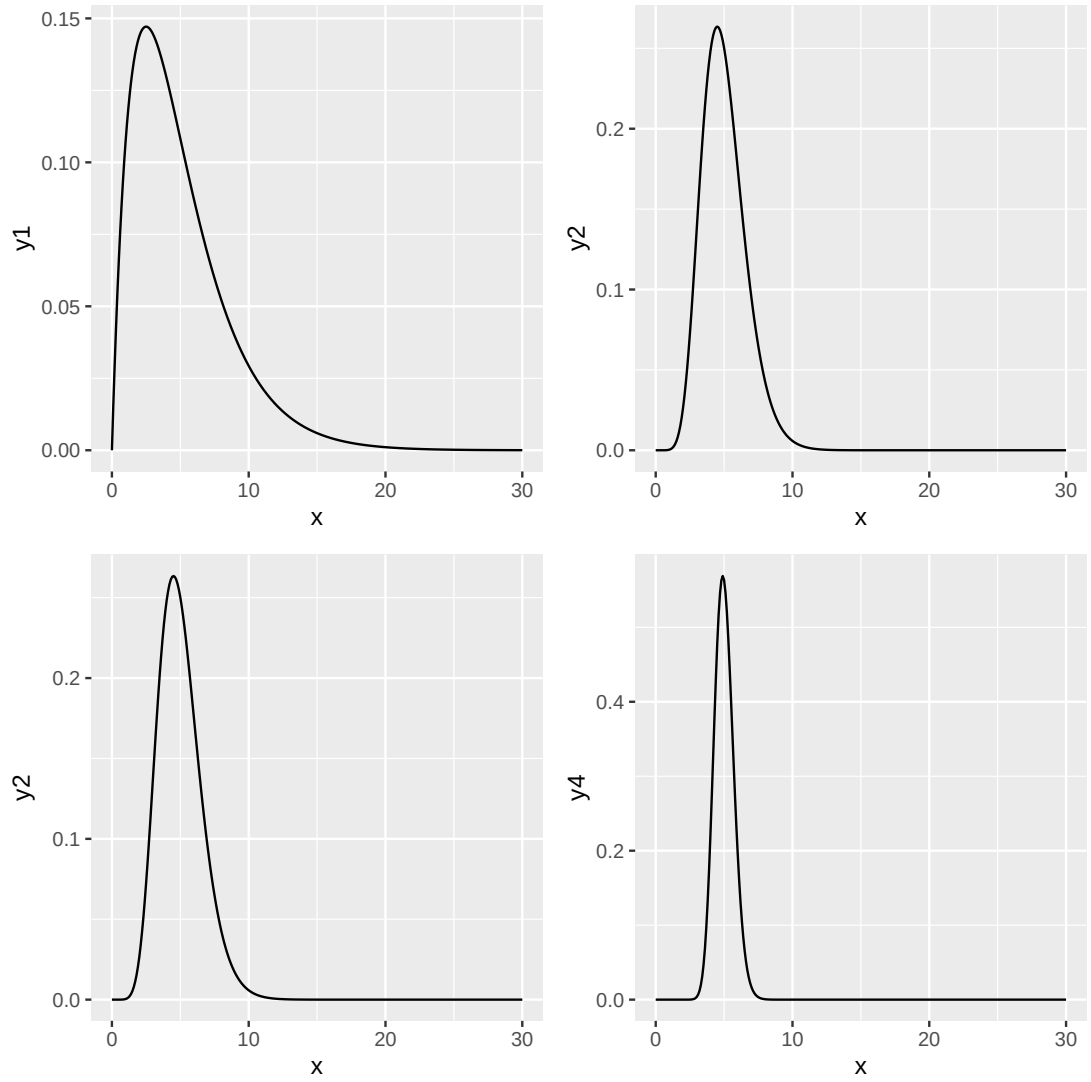
$$E(X) = \alpha\beta \qquad \textbf{[Population Mean]}$$

$$var(X) = \alpha\beta^2 \qquad \textbf{[Population Variance]}$$

You can ask `R` for the gamma distribution PDF using `dgamma(x=x,shape=alpha,scale=beta)`. Plot the exact sampling distribution for $n = 2$, $n = 10$, $n = 35$, and $n = 50$.

```r
library(patchwork)
ggdat <- data.frame(x1=seq(0, 100, 0.1))%>%
  mutate(y1 = dgamma(x=x1, shape=2, scale=1/(2*0.2)),
         y2 = dgamma(x=x1, shape=10, scale=1/(10*0.2)),
         y3 = dgamma(x=x, shape=35, scale=1/(35*0.2)),
         y4 = dgamma(x=x, shape=50, scale=1/(50*0.2)))

plot1<-ggplot(ggdat, aes(x=x, y=y1))+
  geom_line()+
  xlim(0,30)

plot2<-ggplot(ggdat, aes(x=x, y=y2))+
  geom_line()+
  xlim(0,30)

plot3<-ggplot(ggdat, aes(x=x, y=y2))+
  geom_line()+
  xlim(0,30)

plot4<-ggplot(ggdat, aes(x=x, y=y4))+
  geom_line()+
  xlim(0,30)

(plot1|plot2)/(plot3|plot4)
```

```
## Warning:  Removed 700 row(s) containing missing values (geom_path).
## Warning:  Removed 700 row(s) containing missing values (geom_path).
## Warning:  Removed 700 row(s) containing missing values (geom_path).
## Warning:  Removed 700 row(s) containing missing values (geom_path).
```

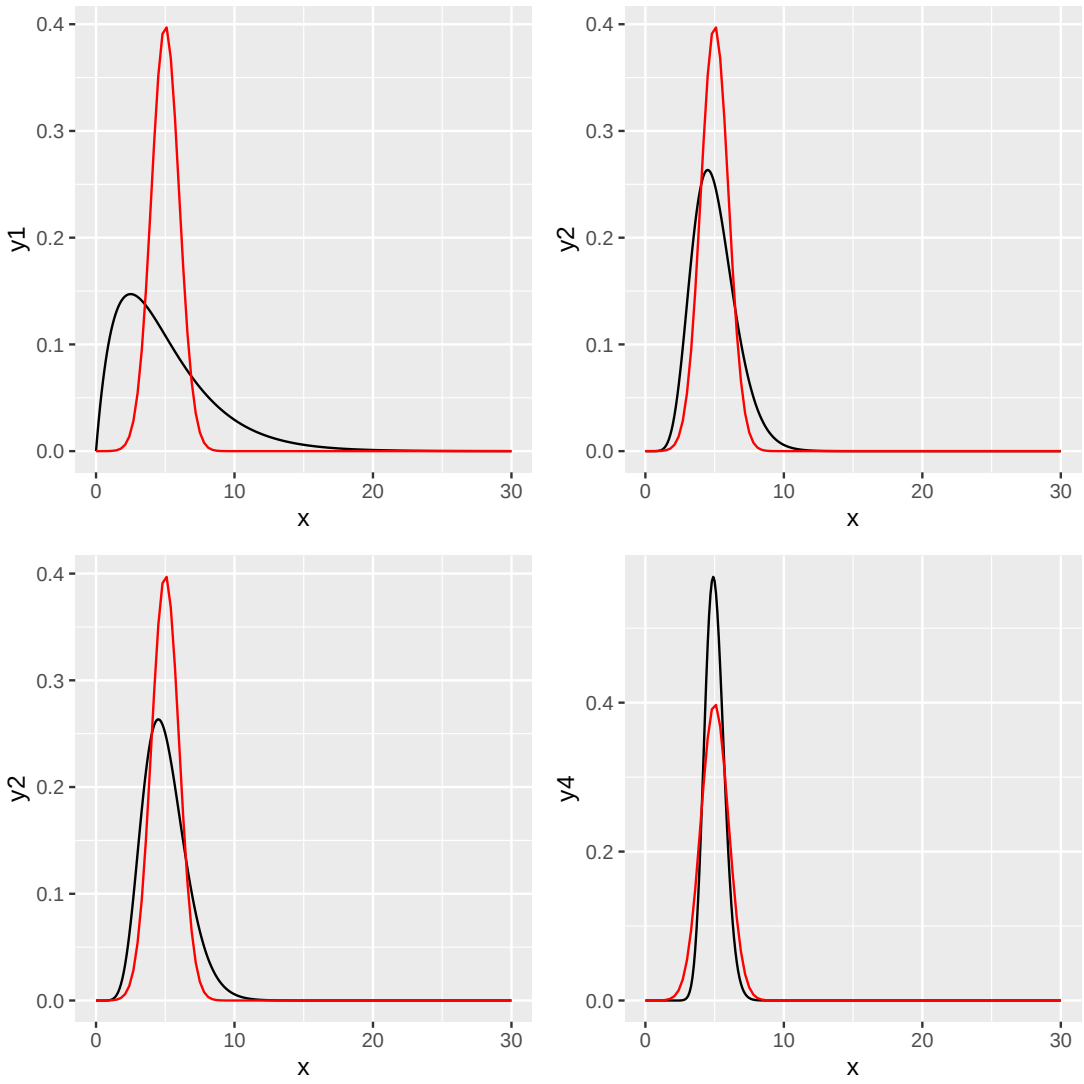


(c) The Central Limit Theorem says that as n increases, the sampling distribution of $\bar{Y}$ can be well approximated with a Gaussian distribution. Superimpose the approximate sampling distribution of $\bar{Y}$ for $n = 2$, $n = 10$, $n = 35$, and $n = 50$.

```
plot1 <- plot1+
  geom_function(fun=dnorm, args=list(mean=5, sd=1),
                color="red")

plot2 <- plot2+
  geom_function(fun=dnorm, args=list(mean=5, sd=1),
                color="red")

plot3 <- plot3+
  geom_function(fun=dnorm, args=list(mean=5, sd=1),
                color="red")

plot4 <- plot4+
```

```
  geom_function(fun=dnorm, args=list(mean=5, sd=1),
                color="red")

(plot1|plot2)/(plot3|plot4)
```

```
## Warning:  Removed 700 row(s) containing missing values (geom_path).
## Warning:  Removed 700 row(s) containing missing values (geom_path).
## Warning:  Removed 700 row(s) containing missing values (geom_path).
## Warning:  Removed 700 row(s) containing missing values (geom_path).
```



(d) Find the probability that a randomly selected rat injected with the toxic substance lives between 1 and 3 days.

```
# P(x < 3) & P(x > 1) = 1-P(x<1)-P(x>3) = 1-P(x<=1)-(1-P(x>3))
1-pgamma(q=1, shape=1, scale=1/0.2)-(1-pgamma(q=3, shape=1, scale=1/0.2))
```

```
## [1] 0.2699191
```

(e) Find the exact probability, using the exact sampling distribution, that two randomly selected rats injected with the toxic substance live between 1 and 3 days on average.

```
# P(x < 3) & P(x > 1) = 1-P(x<1)-P(x>3) = 1-P(x<=1)-(1-P(x>3))
1-pgamma(q=1, shape=2, scale=1/(2*0.2))-(1-pgamma(q=3, shape=2, scale=1/(2*0.2)))
```

```
## [1] 0.2758208
```

(f) Find the approximate probability, using the Central Limit Theorem, that two randomly selected rats injected with the toxic substance live between 1 and 3 days on average. Comment on connection between the results and the assumptions of Central Limit Theorem.

(g) Under what conditions would the approximate probability calculated in part (f) better match the exact probability in part (e)?

3. Below you will load and summarize a dataset containing 575 observations of drug treatments. The data includes the following

   - ID – Identification Code (1 - 575)
   - AGE – Age at Enrollment (Years)
   - BECK – Beck Depression Score (0.000 - 54.000)
   - HC – Heroin/Cocaine Use During 3 Months Prior to Admission (1 = Heroin & Cocaine; 2 = Heroin Only, 3 = Cocaine Only; 4 = Neither Heroin nor Cocaine)
   - IV – History of IV Drug Use (1 = Never; 2 = Previous; 3 = Recent)
   - IV3 – Recent IV use (1 = Yes; 0 = No)
   - NDT – Number of Prior Drug Treatments (0 - 40)
   - RACE – Subject's Race (0 = White; 1 = Non-White)
   - TREAT – Treatment Randomization (0 = Short Assignment; 1 = Long Assignment)
   - SITE – Treatment Site (0 = A; 1 = B)
   - LEN.T – Length of Stay in Treatment (Days Admission Date to Exit Date)
   - TIME – Time to Drug Relapse (Days Measured from Admission Date)
   - CENSOR – Event for Treating Lost to Follow-Up as Returned to Drugs (1 = Returned to Drugs or Lost to Follow-Up; 0 = Otherwise)
   - etc.

   (a) Load the data provided in the "quantreg" package for `R` (Koenker, 2021).

   ```
   library("quantreg")
   data("uis")
   ```

   (b) Is there evidence that patients receiving drug treatments are at least mildly depressed on average? That is, is there evidence that the average BECK depression score is greater than 13, $\mu > 13$?

      i. What is the null hypothesis for this test?
      Solution: $H_0 : \mu = 13$

      ii. What is the alternative hypothesis for this test?
      Solution: $H_a : \mu > 13$

      iii. What is the sample mean BECK score for these data?
      Solution:

      ```
      mu <- mean(uis$BECK)
      paste("The mean of BECK is", mean(mu))
      ## [1] "The mean of BECK is 17.367427826087"
      ```

      iv. What is the test statistics for this test?
      Solution: Mean would be our test statistics.

      ```
      mean(uis$BECK)
      ## [1] 17.36743
      median(uis$BECK)
      ## [1] 17
      #Mean is basically equal to median, so we don't really care. c:
      ```

      v. At what value of $\bar{X}$ does the rejection region start for $\alpha = 0.05$?

$$se = \frac{s}{\sqrt{n}} \qquad \text{(Formula for finding t-value)}$$

$$t = \frac{\bar{X} - \mu}{se} \qquad \text{(Formula for finding t-value)}$$

$$\bar{X} = t * se + \mu \qquad \text{(Solving for } \bar{X}\text{)}$$

```
n<-nrow(uis)
se = sd(uis$BECK)/sqrt(n)
(value<-(qt(.95, n-1)*se)+mu)
```

```
## [1] 18.00866
```

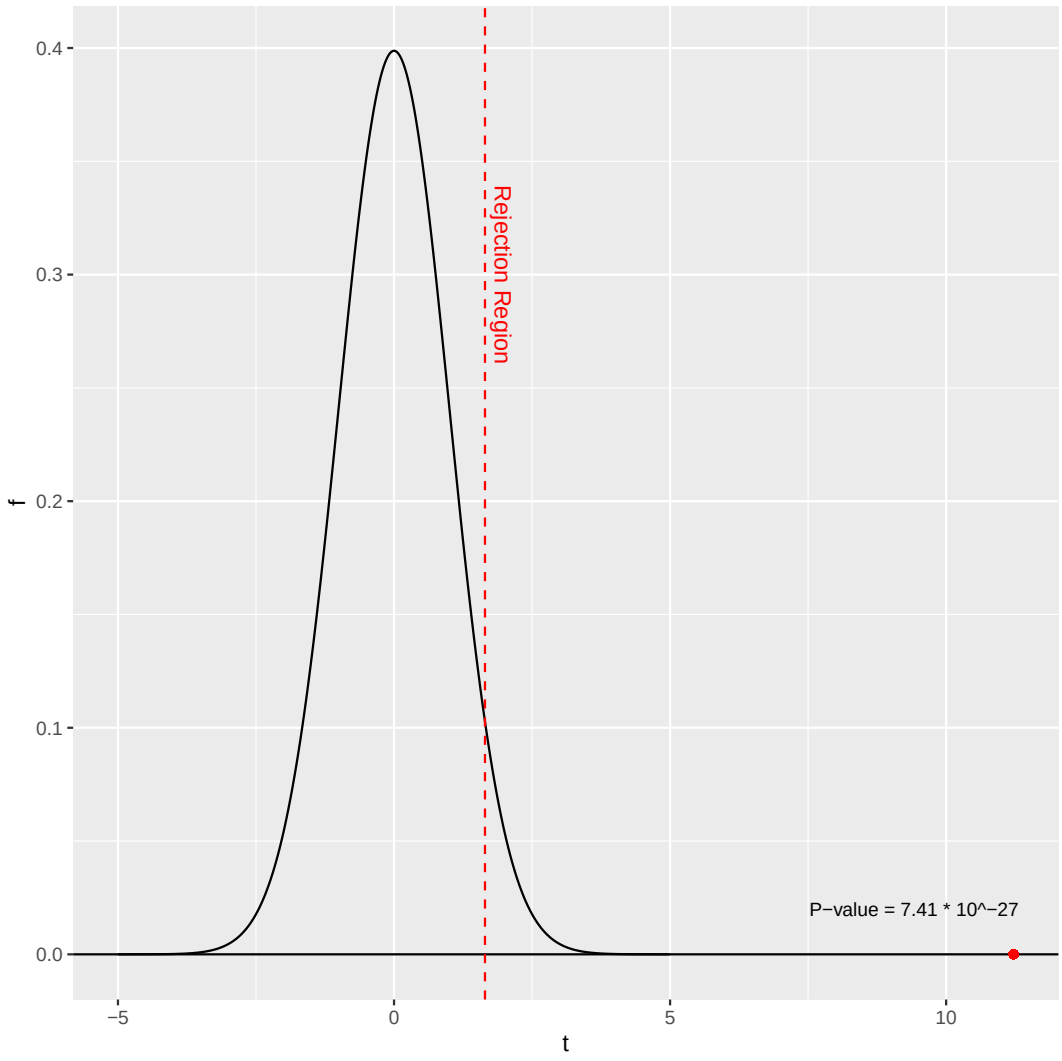
vi. What is the p value for this test?

```
n=nrow(uis)
tstat <- t.test(x=uis$BECK,
       mu = 13,
       alternative = "greater")

tstat$p.value
```

```
## [1] 7.412123e-27
```

vii. Graph the results of this test.

```
ggdat <- data.frame(t=seq(-5,5,length=500))%>%
  mutate(f=dt(x=t, df=n-1))

ggplot(data=ggdat, aes(x=t, y=f))+
  geom_line() +
  geom_hline(yintercept=0)+
  geom_point(aes(x=tstat$statistic, y=0), color="red")+
  geom_vline(xintercept=qt(p=0.95, df=n-1),
             linetype="dashed", color="red")+
  annotate("text", x=1.2*qt(p=0.95, df=n-1), y=0.30,
           label="Rejection Region", angle="270",
           color="red")+
  annotate("text", x=tstat$statistic-1.8, y=0.02,
           label="P-value = 7.41 * 10^-27",
           size=3)
```



viii. Report a 95% confidence interval for the average BECK depression score and interpret it in the context of this question.

```
ci<-t.test(x=uis$BECK,
       alternative = "two.sided")

paste("The confidence interval is [", ci$conf.int[1], ", ",
      ci$conf.int[2],"]", sep="")
## [1] "The confidence interval is [16.6029755170859, 18.131880135088]"
```

(c) Is there a significant difference in the length of stay in treatment by treatment site?
Solution:

$$H_0 : \mu_0 = \mu 1$$
$$H_a : \mu_0 \neq \mu 1$$

```
res <- t.test(uis$LEN.T ~ uis$SITE, data = uis, var.equal = TRUE)
res
```

```
##
##  Two Sample t-test
##
## data:  uis$LEN.T by uis$SITE
## t = -7.8446, df = 573, p-value = 2.137e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -65.03172 -38.98756
## sample estimates:
## mean in group 0 mean in group 1
##         84.9275        136.9371
```

4. Below you will load and summarize a dataset containing 53 observations of prostate cance patients. In this research, a number of possible predictor variables were measured before surgery. The patients then had surgery to determine nodal involvement.

   - r – Nodal Involvement (0=No, 1=Yes)
   - aged – Age Group (0=Less than 60, 1=At least 60)
   - stage – Palpitation Result Severity (0=Less severe, 1=More severe)
   - grade – Biopsy Result Severity (0=Less severe, 1=More severe)
   - xray – X-ray Result Severity (0=Less severe, 1=More severe)
   - acid – the level of acid phosphatase in the blood serum

   The treatment strategy for a patient diagnosed with cancer of the prostate depend highly on whether the cancer has spread to the surrounding lymph nodes (nodal involvement). It is common to operate on the patient to get samples from the nodes which can then be analysed under a microscope.

   (a) Load the data provided in the "boot" package for `R` (Canty and Ripley, 2021).

```
library("boot")
data("nodal")
```

   (b) Is there evidence that less than half of prostate cancer patients have nodal involvment?

      i. What is the null hypothesis for this test?
      Solution: $H_0 : \hat{P} = 0.5$

      ii. What is the alternative hypothesis for this test?
      Solution: $H_0 : \hat{P} \neq 0.5$

      iii. What is the sample proportion of patients with nodal involvement for these data?
      Solution: Something

```
table(nodal$r)[2]/nrow(nodal) #.37%
```

```
##         1
## 0.3773585
```

      iv. What is the test statistics for this test?
      Solution: Z-value

      v. At what value of $\hat{P}$ does the rejection region start for $\alpha = 0.05$?

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$\hat{p} = \frac{\sqrt{n} z \sqrt{p_0(1-p_0)}}{n} + p_0$$

```
p0 <- 0.5
phat <- table(nodal$r)[2]/nrow(nodal)
n <- nrow(nodal)
z1 <- qt(.975, n-1)
z2 <- -z1

(sqrt(n)*z1*sqrt(p0*(1-p0)))/(n) + p0 #for upper bound
```

```
## [1] 0.6378171
```

```
(sqrt(n)*z2*sqrt(p0*(1-p0)))/(n) + p0 #for lower bound
```

```
## [1] 0.3621829
```

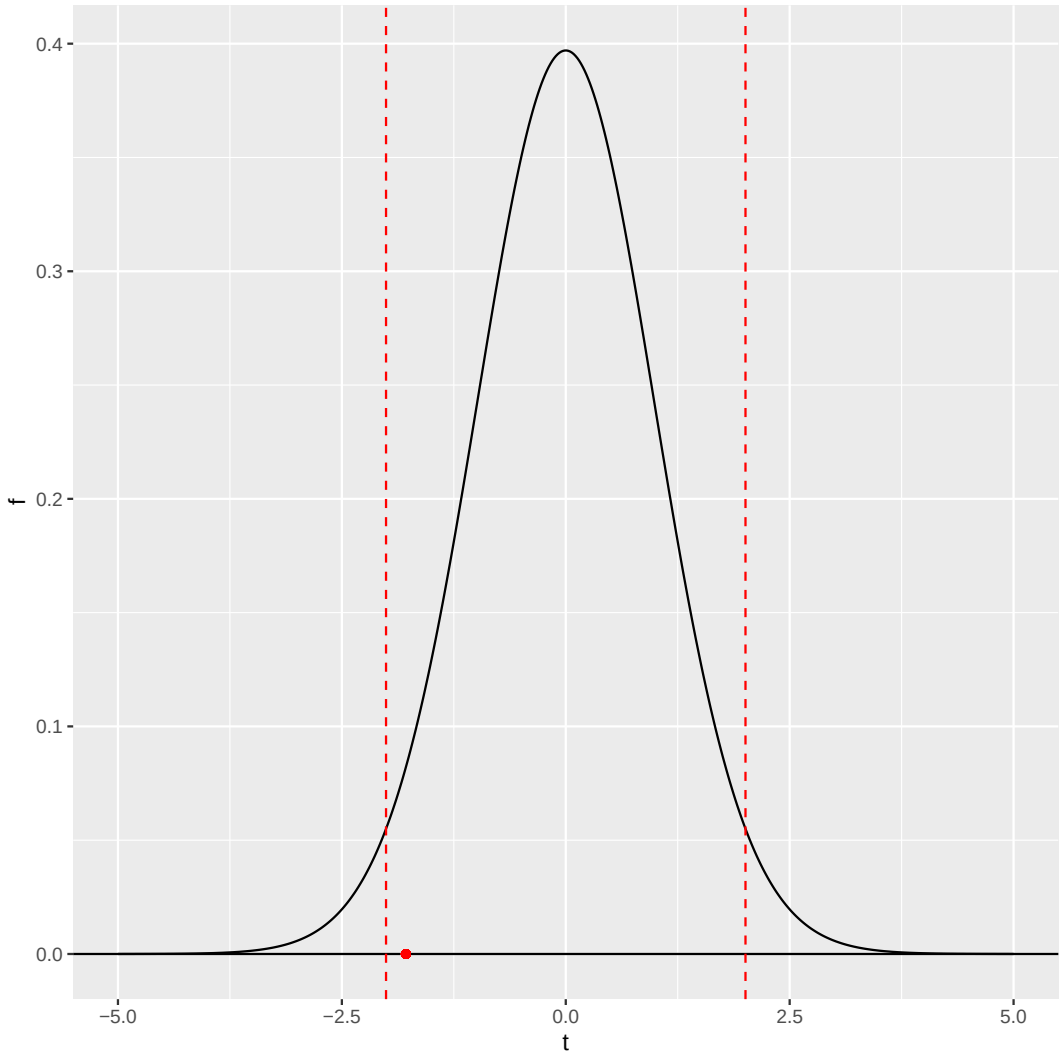
vi. What is the p value for this test?

```
zAns <- (phat-p0)/sqrt((p0*(1-p0))/(n))
zAns
```

```
##         1
## -1.785687
```

vii. Graph the results of this test.

```
ggdat <- data.frame(t=seq(-5,5,length=500))%>%
  mutate(f=dt(x=t, df=n-1))

ggplot(data=ggdat, aes(x=t, y=f))+
  geom_line() +
  geom_hline(yintercept=0)+
  geom_point(aes(x=zAns, y=0), color="red")+
  geom_vline(xintercept=qt(p=0.975, df=n-1),
             linetype="dashed", color="red")+
  geom_vline(xintercept=qt(p=0.025, df=n-1),
             linetype="dashed", color="red")
```



viii. Report a 95% confidence interval for the proportion of prostate cancer patients with nodal involvement and interpret it in the context of this question.

```
#no clue
```

(c) Clearly, it would be preferable if an accurate assessment of nodal involvement could be made without surgery. Is there a significant difference in the nodal involvement of patients with any of the severity indicators?

Bonus 1: Use the gganimate package (Pedersen and Robinson, 2020) for `R` to create a plot that demonstrates the Central Limit Theorem for the Poisson, Binomial, Exponential, and Gaussian distributions in a 2×2 grid as here.

Note that we can't add GIFs to the .pdf document, so you'll have to email me your code for this part. You'll find `transition_time()` helpful for creating your animation and `gganimate_save()` helpful for saving your animation.

Bonus 2: Compare the effectiveness of the t-interval with the bootstrap interval. In a loop, generate 1000 datasets, evaluate a t and bootstrapping confidence interval for each set of data, and track whether you've captured the true population mean. An effective answer here would evaluate this several times varying sample size and the data generating distribution.

References

Brockmann, H. J. (1996). Satellite male groups in horseshoe crabs, limulus polyphemus. *Ethology*, 102(1):1–21.

Canty, A. and Ripley, B. D. (2021). *boot: Bootstrap R (S-Plus) Functions*. R package version 1.3-28.

Koenker, R. (2021). *quantreg: Quantile Regression*. R package version 5.86.

Pedersen, T. L. and Robinson, D. (2020). *gganimate: A Grammar of Animated Graphics*. R package version 1.0.7.