GroupA_HW1

S. Cattonar, L.Ricatti, M. Rizwan, D. Rosa, A. Valle

2024-11-05

Contents

CS - Chapter 1

Ex 1.1

(Luca) *Exponential random variable, $X \geq 0$, has p.d.f. $f(x) = \lambda \exp(-\lambda x)$.*

1. *Find the c.d.f. and the quantile function for X.*
2. *Find $Pr(X < \lambda)$ and the median of X.*
3. *Find the mean and variance of X.*

Solution

1. C.D.F. and Quantile Function:

   The cumulative distribution function (c.d.f.) F(x) is:

   $$F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}, \quad x \geq 0$$

   The quantile function Q(p) is the inverse of F(x):

   $$Q(p) = -\frac{1}{\lambda} \ln(1 - p), \quad 0 \leq p < 1$$

2. Pr(X $<$ λ) and Median:

   Pr(X $<$ λ) = F(λ) = $1 - e^{-\lambda\lambda} = 1 - e^{-1} \approx 0.6321$

   For the median, we solve F(x) = 0.5:

   $$1 - e^{-\lambda x} = 0.5$$
   $$x = -\frac{1}{\lambda} \ln(0.5) = \frac{\ln(2)}{\lambda}$$

3. Mean and Variance:

   Mean: E[X] = $\int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$

   Variance: Var(X) = E[X^2] - (E[X])^2 = $\frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$

Comments on the solution:

1. The exponential distribution is characterized by its rate parameter λ, which determines both its shape and scale.

2. Interestingly, Pr(X $<$ λ) is always approximately 0.6321, regardless of the value of λ. This is a unique property of the exponential distribution.

3. The median of the distribution is $\frac{\ln(2)}{\lambda}$, which is always less than the mean ($\frac{1}{\lambda}$) due to the distribution's right-skewness.

4. The mean and variance are both functions of λ. As λ increases, both the mean and variance decrease, indicating that larger values of λ result in the distribution being more concentrated near zero.

5. The standard deviation of the distribution is equal to its mean, which is a distinctive feature of the exponential distribution.

Ex 1.6

(Luca) *Let X and Y be non-independent random variables, such that var(X)* $= \sigma_x^2$*, var(Y)* $= \sigma_y^2$ *and cov(X,Y)* $= \sigma_{xy}^2$*. Using the result from Section 1.6.2, find var(X + Y) and var(X - Y).*

Solution

Using the formula for linear transformations of random vectors from Section 1.6.2:

$\mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{A\ A^T})$

1. Define a random vector and its covariance matrix:
   $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$
   $\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix}$
2. For var(X + Y): Let $\mathbf{A} = \begin{pmatrix} 1 & 1 \end{pmatrix}$
   $\mathrm{var}(X+Y) = \mathbf{A\ A^T} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$
   $= \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} \sigma_x^2 + \sigma_{xy}^2 \\ \sigma_{xy}^2 + \sigma_y^2 \end{pmatrix} = \sigma_x^2 + 2\sigma_{xy}^2 + \sigma_y^2$
3. For var(X - Y): Let $\mathbf{A} = \begin{pmatrix} 1 & -1 \end{pmatrix}$
   $\mathrm{var}(X-Y) = \mathbf{A\ A^T} = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$
   $= \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \sigma_x^2 - \sigma_{xy}^2 \\ -(\sigma_{xy}^2 - \sigma_y^2) \end{pmatrix} = \sigma_x^2 - 2\sigma_{xy}^2 + \sigma_y^2$

Therefore:

$\mathrm{var}(X+Y) = \sigma_x^2 + 2\sigma_{xy}^2 + \sigma_y^2$

$\mathrm{var}(X-Y) = \sigma_x^2 - 2\sigma_{xy}^2 + \sigma_y^2$

Comments on the solution:

1. For the sum (X + Y), the covariance term is added twice, potentially increasing the overall variance if X and Y are positively correlated.
2. For the difference (X - Y), the covariance term is subtracted twice, potentially decreasing the overall variance if X and Y are positively correlated.
3. If X and Y are independent (i.e., $\sigma_{xy}^2 = 0$), the variances of their sum and difference would both simplify to $\sigma_x^2 + \sigma_y^2$.

Ex 3.5

(Devid) *Consider solving the matrix equation* $Ax = y$ *for* x*, where* y *is a known* n*-vector and* A *is a known* $n \times n$ *matrix. The formal solution to the problem is* $x = A^{-1}y$*, but it is possible to solve the equation directly, without actually forming* A^{-1}*. This question explores this direct solution.*

a. *First create an A, x and y satisfying $Ax = y$.*

```
set.seed(0)
n <- 1000
A <- matrix(runif(n*n), n, n)
x.true <- runif(n)
y <- A %*% x.true
```

The idea is to experiment with solving $Ax = y$ for x, but with a known truth to compare the answer to.

b. *Using solve, form the matrix $A-1$ explicitly and then form x1 = A^(-1)y. Note how long this takes. Also assess the mean absolute difference between x1 and x.true (the approximate mean absolute 'error' in the solution).*

c. *Now use solve to directly solve for x without forming A^(-1). Note how long this takes and assess the mean absolute error of the result.*

d. *What do you conclude?*

Solution

```r
set.seed(0)
n <- 1000

# Part (a): Generate A, x.true, and y
A <- matrix(runif(n * n), n, n)
x.true <- runif(n)
y <- A %*% x.true

# Part (b): Solve by forming A^-1 explicitly
start_time_1 <- Sys.time()
A_inv <- solve(A)
x1 <- A_inv %*% y
end_time_1 <- Sys.time()

time_taken_1 <- end_time_1 - start_time_1
error_1 <- mean(abs(x1 - x.true))

# Part (c): Solve directly without forming A^-1
start_time_2 <- Sys.time()
x2 <- solve(A, y)
end_time_2 <- Sys.time()

time_taken_2 <- end_time_2 - start_time_2
error_2 <- mean(abs(x2 - x.true))

# Part (d): Print results and conclusions
cat("Results:\n")
```

```
## Results:
```

```r
cat("1. Using explicit inverse (A^-1 * y):\n")
```

```
## 1. Using explicit inverse (A^-1 * y):
```

```r
cat("   Time taken:", time_taken_1, "\n")
```

```
##    Time taken: 0.588455
```

```r
cat("   Mean absolute error:", error_1, "\n\n")
```

```
##    Mean absolute error: 2.956833e-11
```

```r
cat("2. Using direct solve (solve(A, y)):\n")
```

```
## 2. Using direct solve (solve(A, y)):
```

```r
cat("   Time taken:", time_taken_2, "\n")
```

```
##    Time taken: 0.1141829
```

Directly solving (solve(A, y)) is faster than forming the inverse and also has comparable accuracy.

Ex 3.6

(Devid) *The empirical cumulative distribution function (ECDF) for a set of measurements $x_i : i = 1, \ldots, n$ is*

$$\hat{F}(x) = \frac{\#\{x_i < x\}}{n}$$

where $\#\{x_i < x\}$ denotes the number of x_i values that are less than x. When answering the following, try to ensure that your code is commented, clearly structured, and tested. To test your code, generate random samples using rnorm, runif, etc.

a. *Write an R function that takes an unordered vector of observations x and returns the values of the empirical c.d.f. for each value, in the order corresponding to the original x vector. See ?sort.int.*
b. *Modify your function to take an extra argument plot.cdf, that when TRUE will cause the empirical c.d.f. to be plotted as a step function over a suitable x range.*

Solution

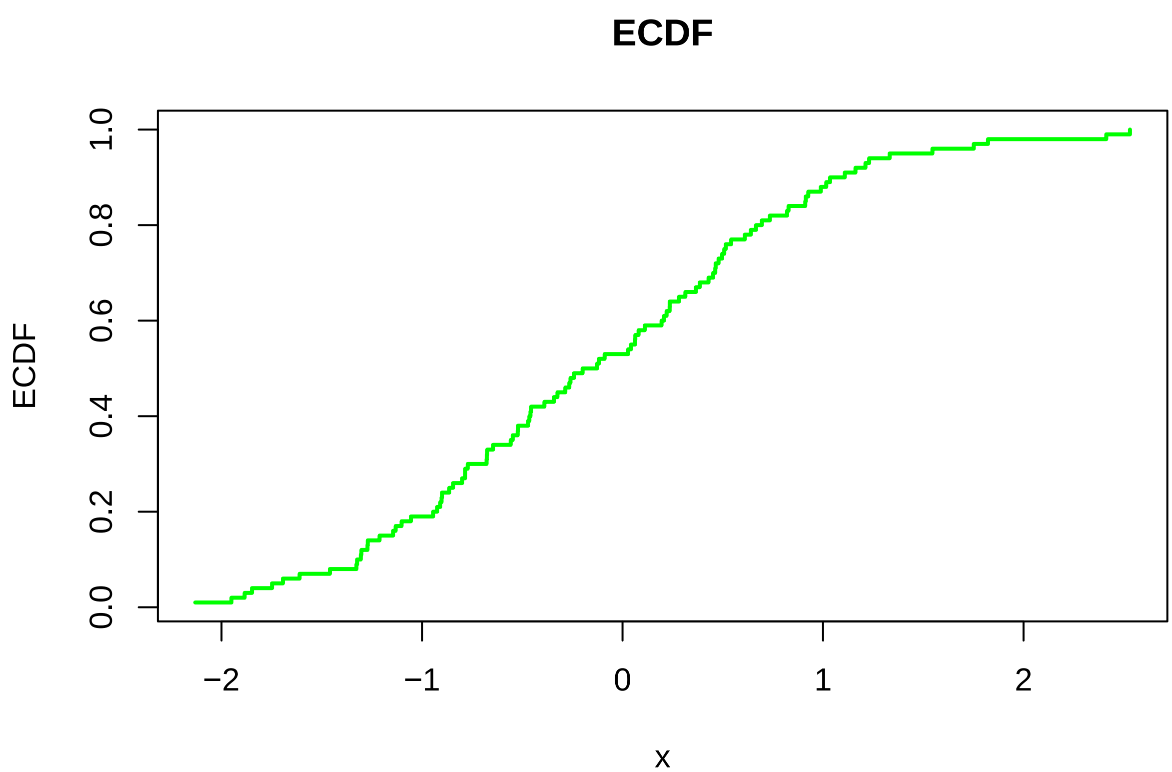
```r
compute_ecdf <- function(x, plot.cdf = FALSE) {
  n <- length(x)
  sorted_x <- sort(x)
  ecdf_values <- cumsum(table(cut(x, breaks = c(-Inf, sorted_x)))) / n

  # Match ECDF values back to the original order of x
```

```r
  ecdf_original_order <- ecdf_values[order(order(x))]

  # Part (b)
  if (plot.cdf) {
    plot(sort(x), ecdf_values, type = "s", col = "green", lwd = 2, xlab = "x", ylab = "ECDF", main =
        }
  return(ecdf_original_order)
}

# Testing the function
set.seed(18)
x <- rnorm(100)

ecdf_values <- compute_ecdf(x)
print(ecdf_values)
```

```
##   (0.9138,0.9265]     (1.752,1.823]   (-1.694,-1.611] (-0.3246,-0.2851]
##              0.87              0.98              0.07              0.46
## (-0.3895,-0.3421]   (0.3135,0.3662]    (-1.46,-1.327]     (1.823,2.413]
##              0.44              0.67              0.09              0.99
##  (0.0624,0.06382]     (1.332,1.546]    (-1.95,-1.885]   (0.8279,0.9114]
##              0.57              0.96              0.03              0.85
##   (-1.324,-1.305] (0.02814,0.04207] (-0.7999,-0.7853]     (1.162,1.212]
##              0.11              0.55              0.28              0.93
## (-0.9446,-0.9245]  (-0.7721,-0.678]      (1.23,1.332]   (0.4522,0.4626]
##              0.21              0.31              0.95              0.71
##   (-1.305,-1.302]     (1.035,1.108] (-0.7848,-0.7721]  (-0.678,-0.6772]
##              0.12              0.91              0.30              0.32
##   (0.4626,0.4643]    (-2.131,-1.95]   (-1.102,-1.056] (-0.1265,-0.1178]
##              0.72              0.02              0.19              0.52
## (-0.2652,-0.2595]   (-1.848,-1.748] (-0.9245,-0.9089]   (0.1946,0.2069]
##              0.48              0.05              0.22              0.61
##   (0.2819,0.3135]     (1.108,1.162]   (-1.748,-1.694]     (1.016,1.035]
##              0.66              0.92              0.06              0.90
##     (0.64,0.6658]   (-1.302,-1.272]    (0.542,0.6096]    (0.7348,0.821]
##              0.80              0.13              0.78              0.83
##   (0.6658,0.6949]     (2.413,2.531]   (0.4289,0.4522]   (0.9265,0.9887]
##              0.81              1.00              0.70              0.88
## (-0.1992,-0.1265]   (0.5074,0.5149]    (0.821,0.8279]   (-1.327,-1.324]
##              0.51              0.76              0.84              0.10
## (-0.3421,-0.3246] (-0.6458,-0.5579] (-0.9089,-0.9026]   (-1.144,-1.132]
##              0.45              0.35              0.23              0.17
##   (0.4789,0.4971] (-0.5579,-0.5477] (-0.1178,-0.0894]     (-Inf,-2.131]
##              0.74              0.36              0.53              0.01
## (-0.9026,-0.9007]      (1.212,1.23] (0.06382,0.08006]   (0.3852,0.4289]
##              0.24              0.94              0.58              0.69
## (-0.8641,-0.8453]    (0.2197,0.235]    (-1.611,-1.46] (-0.0894,0.02814]
##              0.26              0.63              0.08              0.54
## (-0.6772,-0.6747]    (0.5149,0.542]   (-1.271,-1.212]     (0.6096,0.64]
##              0.33              0.77              0.15              0.79
##  (-0.5218,-0.471]   (0.1107,0.1946] (-0.5229,-0.5218] (-0.2421,-0.1992]
##              0.39              0.60              0.38              0.50
## (-0.5477,-0.5229]   (-1.272,-1.271] (-0.4558,-0.3895]   (0.2351,0.2819]
```

```
##                0.37              0.14               0.43              0.65
##    (0.3662,0.3852] (-0.8453,-0.7999]  (0.04207,0.0624]   (0.4971,0.5074]
##                0.68              0.27               0.56              0.75
##   (-0.459,-0.4558]   (-1.132,-1.102] (-0.7853,-0.7848]     (1.546,1.752]
##                0.42              0.18               0.29              0.97
##    (0.4643,0.4789]  (-0.471,-0.4645] (-0.2851,-0.2652]   (-1.885,-1.848]
##                0.73              0.40               0.47              0.04
##     (0.235,0.2351]   (0.6949,0.7348]  (0.08006,0.1107] (-0.9007,-0.8641]
##                0.64              0.82               0.59              0.25
##    (-1.212,-1.144]  (-0.4645,-0.459]   (0.2069,0.2197] (-0.6747,-0.6458]
##                0.16              0.41               0.62              0.34
##    (0.9114,0.9138]  (-1.056,-0.9446] (-0.2595,-0.2421]    (0.9887,1.016]
##                0.86              0.20               0.49              0.89
```

```r
# Compute the ECDF with plotting
compute_ecdf(x, plot.cdf = TRUE)
```




```
##    (0.9138,0.9265]     (1.752,1.823]   (-1.694,-1.611] (-0.3246,-0.2851]
##               0.87              0.98               0.07              0.46
## (-0.3895,-0.3421]   (0.3135,0.3662]    (-1.46,-1.327]     (1.823,2.413]
##               0.44              0.67               0.09              0.99
##   (0.0624,0.06382]     (1.332,1.546]    (-1.95,-1.885]   (0.8279,0.9114]
##               0.57              0.96               0.03              0.85
##    (-1.324,-1.305] (0.02814,0.04207] (-0.7999,-0.7853]     (1.162,1.212]
##               0.11              0.55               0.28              0.93
```

```
## (-0.9446,-0.9245]  (-0.7721,-0.678]       (1.23,1.332]   (0.4522,0.4626]
##               0.21               0.31               0.95              0.71
##   (-1.305,-1.302]     (1.035,1.108] (-0.7848,-0.7721]  (-0.678,-0.6772]
##               0.12               0.91               0.30              0.32
##   (0.4626,0.4643]     (-2.131,-1.95]   (-1.102,-1.056] (-0.1265,-0.1178]
##               0.72               0.02               0.19              0.52
## (-0.2652,-0.2595]    (-1.848,-1.748] (-0.9245,-0.9089]   (0.1946,0.2069]
##               0.48               0.05               0.22              0.61
##   (0.2819,0.3135]     (1.108,1.162]   (-1.748,-1.694]     (1.016,1.035]
##               0.66               0.92               0.06              0.90
##     (0.64,0.6658]    (-1.302,-1.272]     (0.542,0.6096]    (0.7348,0.821]
##               0.80               0.13               0.78              0.83
##   (0.6658,0.6949]     (2.413,2.531]   (0.4289,0.4522]   (0.9265,0.9887]
##               0.81               1.00               0.70              0.88
## (-0.1992,-0.1265]   (0.5074,0.5149]     (0.821,0.8279]   (-1.327,-1.324]
##               0.51               0.76               0.84              0.10
## (-0.3421,-0.3246] (-0.6458,-0.5579] (-0.9089,-0.9026]   (-1.144,-1.132]
##               0.45               0.35               0.23              0.17
##   (0.4789,0.4971] (-0.5579,-0.5477] (-0.1178,-0.0894]     (-Inf,-2.131]
##               0.74               0.36               0.53              0.01
## (-0.9026,-0.9007]      (1.212,1.23] (0.06382,0.08006]   (0.3852,0.4289]
##               0.24               0.94               0.58              0.69
## (-0.8641,-0.8453]    (0.2197,0.235]     (-1.611,-1.46] (-0.0894,0.02814]
##               0.26               0.63               0.08              0.54
## (-0.6772,-0.6747]    (0.5149,0.542]   (-1.271,-1.212]     (0.6096,0.64]
##               0.33               0.77               0.15              0.79
##  (-0.5218,-0.471]   (0.1107,0.1946] (-0.5229,-0.5218] (-0.2421,-0.1992]
##               0.39               0.60               0.38              0.50
## (-0.5477,-0.5229]   (-1.272,-1.271] (-0.4558,-0.3895]   (0.2351,0.2819]
##               0.37               0.14               0.43              0.65
##   (0.3662,0.3852] (-0.8453,-0.7999]   (0.04207,0.0624]   (0.4971,0.5074]
##               0.68               0.27               0.56              0.75
##  (-0.459,-0.4558]   (-1.132,-1.102] (-0.7853,-0.7848]     (1.546,1.752]
##               0.42               0.18               0.29              0.97
##   (0.4643,0.4789]   (-0.471,-0.4645] (-0.2851,-0.2652]   (-1.885,-1.848]
##               0.73               0.40               0.47              0.04
##    (0.235,0.2351]   (0.6949,0.7348]  (0.08006,0.1107] (-0.9007,-0.8641]
##               0.64               0.82               0.59              0.25
##   (-1.212,-1.144]   (-0.4645,-0.459]   (0.2069,0.2197] (-0.6747,-0.6458]
##               0.16               0.41               0.62              0.34
##   (0.9114,0.9138]   (-1.056,-0.9446] (-0.2595,-0.2421]    (0.9887,1.016]
##               0.86               0.20               0.49              0.89
```

FSDS - Chapter 2

Ex 2.8

(Devid) *Each time a person shops at a grocery store, the event of catching a cold or some other virus from another shopper is independent from visit to visit and has a constant probability over the year, equal to 0.01.*

(a) *In 100 trips to this store over the course of a year, the probability of catching a virus while shopping there is 100(0.01) = 1.0. What is wrong with this reasoning?*
(b) *Find the correct probability in (a).*

Solution

a)

The algorithm followed to compute the probability isn't correct. In fact, if we suppose to compute the same probability for 200 days, we will obtain a value of 2.0, but probability functions are defined in $\Omega \rightarrow [0, 1]$.

b)

The event of getting a cold at the supermarket in a single day can be described by a Bernoulli random variable:

$$X \sim \text{Be}(0.01)$$

The event of getting a cold at the supermarket over 100 days can be described by a Binomial random variable:

$$X \sim \text{Bin}(100, 0.01)$$

The probability of getting a virus, denoted P_v*, is given by:*

$$P_v = 1 - P(X = 0)$$

This can be calculated as:

$$P_v = 1 - P(X = 0) = 1 - \binom{100}{0} \cdot (0.01)^0 \cdot (1 - 0.01)^{100-0} \approx 0.6339677$$

Ex 2.16

(Luca) *Each day a hospital records the number of people who come to the emergency room for treatment.(a) In the first week, the observations from Sunday to Saturday are 10, 8, 14, 7, 21, 44, 60. Do you think that the Poisson distribution might describe the random variability of this phenomenon adequately. Why or why not?*

Solution To assess whether the Poisson distribution might adequately describe the random variability of emergency room visits, we'll examine the data and compare it to properties of the Poisson distribution.

```r
# Data
er_visits <- c(10, 8, 14, 7, 21, 44, 60)
days <- c("Sun", "Mon", "Tue", "Wed", "Thu", "Fri", "Sat")

# Basic statistics
mean_visits <- mean(er_visits)
var_visits <- var(er_visits)

# Print results
cat("Mean of visits:", round(mean_visits, 2), "\n")
```

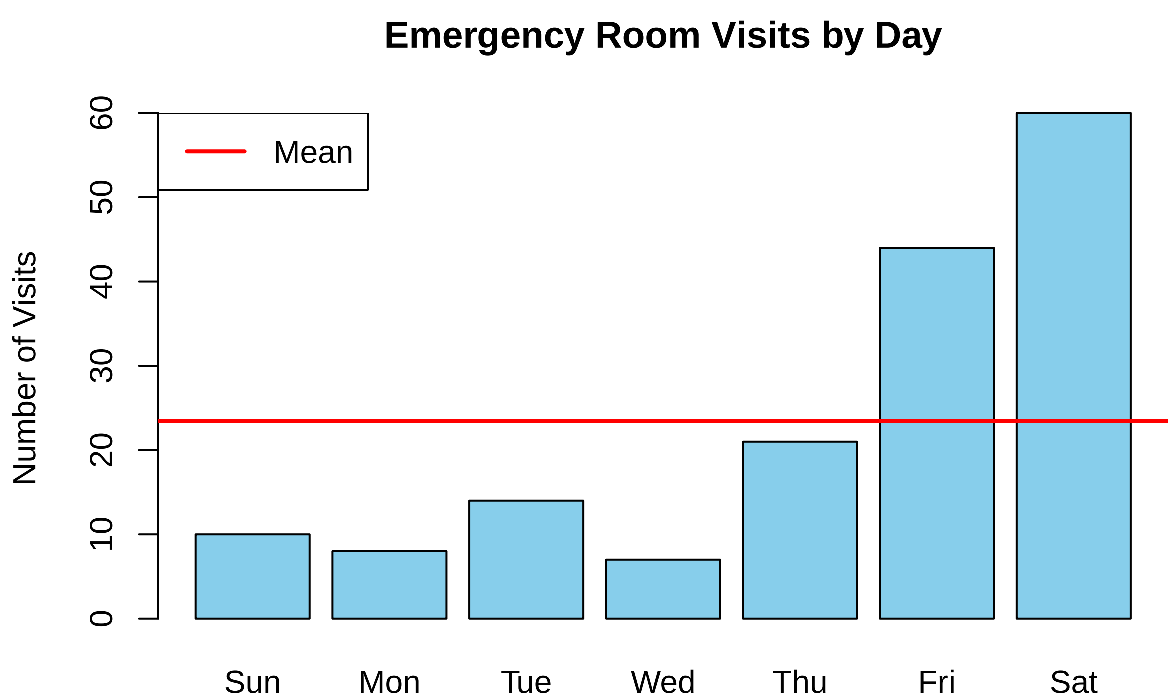
a)

```
## Mean of visits: 23.43
```

```r
cat("Variance of visits:", round(var_visits, 2), "\n")
```

```
## Variance of visits: 423.95
```

```r
# Plot the data
barplot(er_visits, names.arg = days, main = "Emergency Room Visits by Day",
        ylab = "Number of Visits", col = "skyblue")

# Add a line for the mean
abline(h = mean_visits, col = "red", lwd = 2)
legend("topleft", legend = "Mean", col = "red", lwd = 2)
```



1. The Poisson distribution has the property that its mean and variance are equal. In our data:

   - Mean: 23.4285714
   - Variance: 423.952381 The large difference between these values suggests that the Poisson distribution may not be appropriate.

2. The bar plot shows a clear increasing trend throughout the week, with a sharp increase on Friday and Saturday. This pattern is not consistent with the Poisson distribution, which assumes a constant rate of events.

3. The Poisson distribution assumes:

   - Events occur independently

- The average rate of occurrences is constant

In this case, the number of ER visits doesn't satisfy these assumptions:

- There may be dependencies (e.g., a local event affecting multiple people)
- The rate clearly varies by day of the week

4. The data shows overdispersion (variance much larger than the mean), which is not characteristic of the Poisson distribution.

Given these observations, we can conclude that the Poisson distribution does not adequately describe the random variability of emergency room visits in this hospital. A more complex model that accounts for day-of-week effects and overdispersion (such as a negative binomial distribution or a time series model) would likely be more appropriate.

b) Solution

Yes, we would expect the Poisson distribution to better describe the number of weekly admissions to the hospital for a rare disease.

1. **Rare events**: The Poisson distribution is particularly well-suited for modeling rare events. A rare disease, by definition, occurs infrequently.
2. **Independence**: Admissions for a rare disease are more likely to be independent of each other, especially if the disease is not contagious.
3. **Constant rate**: The occurrence of a rare disease is less likely to be affected by day-of-week patterns or other cyclical factors that we observed in general ER admissions.
4. **No simultaneous occurrences**: With rare diseases, the probability of two or more admissions occurring simultaneously is extremely low, which aligns with another assumption of the Poisson distribution.
5. **Lower variance**: Rare events typically have a lower variance, which is more likely to be closer to the mean.
6. **Small numbers**: The Poisson distribution is often used to model count data when the counts are small, which is likely the case for weekly admissions of a rare disease.

To illustrate this point, we can simulate weekly admissions for a hypothetical rare disease:

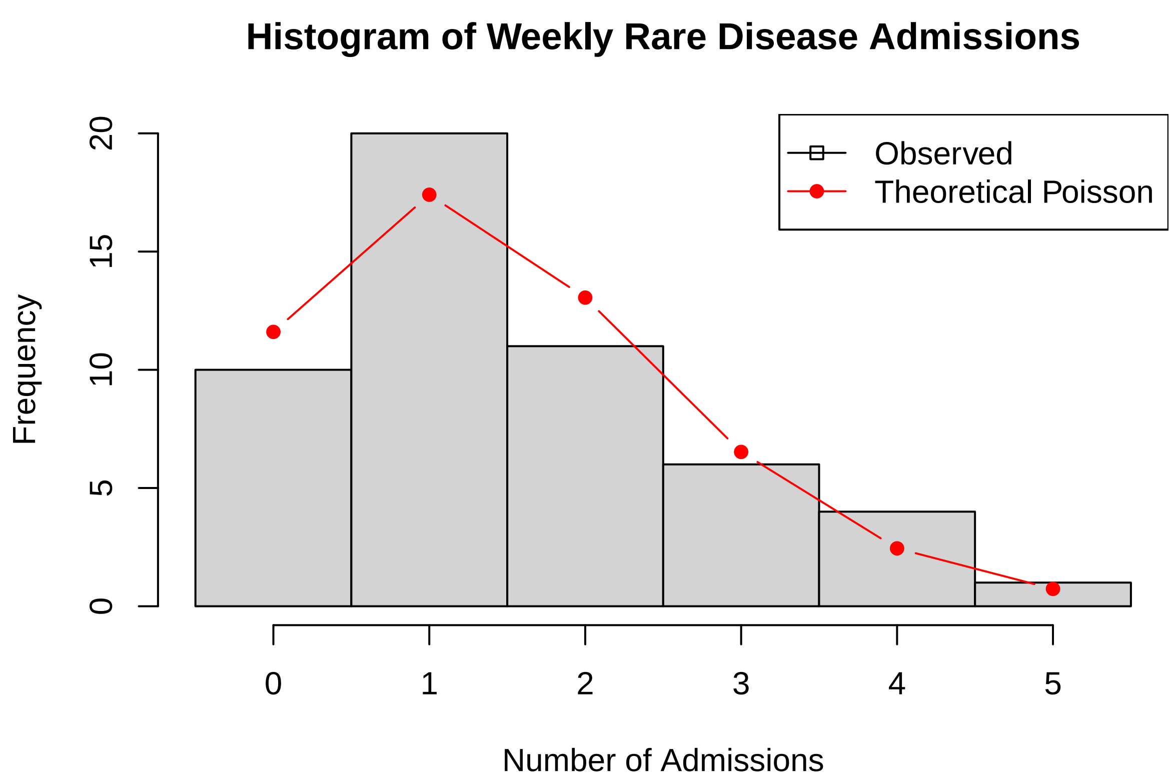
```r
set.seed(123)
weeks <- 52
lambda <- 1.5  # Average 1.5 admissions per week for the rare disease
rare_disease_admissions <- rpois(weeks, lambda)

# Basic statistics
mean_admissions <- mean(rare_disease_admissions)
var_admissions <- var(rare_disease_admissions)

# Print results
cat("Mean of admissions:", round(mean_admissions, 2), "\n")
```

```
## Mean of admissions: 1.56
```

```r
cat("Variance of admissions:", round(var_admissions, 2), "\n")
```

```
## Variance of admissions: 1.58
```

```r
# Plot the data
hist(rare_disease_admissions, breaks = seq(-0.5, max(rare_disease_admissions) + 0.5, by = 1),
     main = "Histogram of Weekly Rare Disease Admissions",
     xlab = "Number of Admissions", ylab = "Frequency")

# Overlay Poisson distribution
x <- 0:max(rare_disease_admissions)
lines(x, dpois(x, lambda) * weeks, col = "red", type = "b", pch = 16)
legend("topright", legend = c("Observed", "Theoretical Poisson"),
       col = c("black", "red"), lty = 1, pch = c(22, 16))
```



Comments on Solution:

1. The mean (1.56) and variance (1.58) of the simulated data are much closer to each other, which is characteristic of the Poisson distribution.
2. The histogram of simulated admissions closely follows the theoretical Poisson distribution (red line), indicating a good fit.
3. The number of admissions per week is small and varies within a narrow range, which is typical for rare events and well-described by the Poisson distribution.

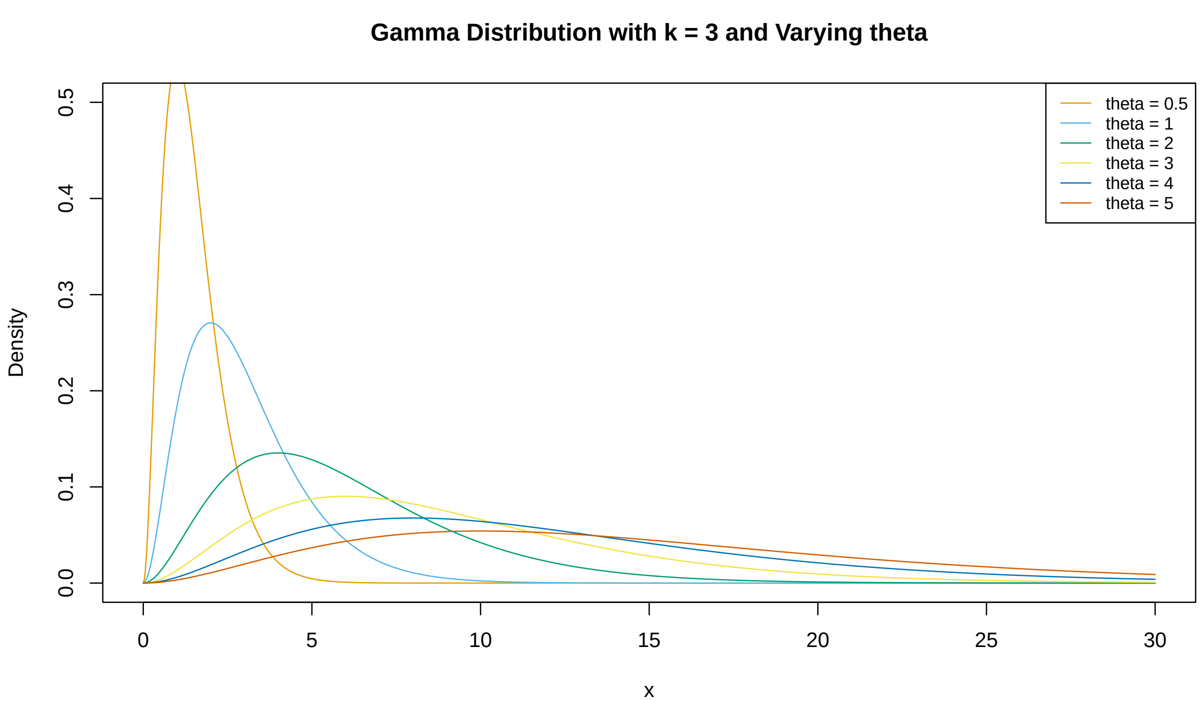
Ex 2.21

(Luca) *Plot the gamma distribution by fixing the shape parameter $k = 3$ and setting the scale parameter $\theta = 0.5, 1, 2, 3, 4, 5$. What is the effect of increasing the scale parameter? (See also Exercise 2.48.)*

Solution

To visualize the effect of increasing the scale parameter on the gamma distribution, we'll create a plot showing multiple gamma distributions with a fixed shape parameter and varying scale parameters.

```r
# Set parameters
k <- 3  # Shape parameter
theta <- c(0.5, 1, 2, 3, 4, 5)  # Scale parameters
colors <- c("#E69F00", "#56B4E9", "#009E73", "#F0E442", "#0072B2", "#D55E00")

# Create x values
x <- seq(0, 30, length.out = 1000)

# Plot
plot(x, dgamma(x, shape = k, scale = theta[1]), type = "l", col = colors[1],
     main = "Gamma Distribution with k = 3 and Varying theta",
     xlab = "x", ylab = "Density", ylim = c(0, 0.5))

# Add lines for other scale parameters
for (i in 2:length(theta)) {
  lines(x, dgamma(x, shape = k, scale = theta[i]), col = colors[i])
}

# Add legend
legend("topright", legend = paste("theta =", theta), col = colors, lty = 1, cex = 0.8)
```



Comments on the solution:

1. **Shape of the distribution**: As we increase the scale parameter θ, we observe:

   a. The peak of the distribution shifts to the right (towards larger x values).
   b. The height of the peak decreases.
   c. The distribution becomes wider.

2. **Interpretation**:

   - A larger scale parameter θ indicates greater variability and a shift in the distribution.
   - This can be explained by the relationship between the parameters of the gamma distribution:
     - The mean of the gamma distribution is $\mu = k\theta$
     - The variance is $\sigma^2 = k\theta^2$
   - As θ increases:
     - The mean increases linearly ($k\theta$)
     - The variance increases quadratically ($k\theta^2$)
   - This quadratic increase in variance relative to the mean explains the greater spread and variability we observe with larger θ values.

Ex 2.26

Refer to Table 2.4 cross classifying happiness with family income.

Solution

Add comments to the solution.

Ex 2.52

(Devid) *The pdf f of a $N(\mu, \sigma^2)$ distribution can be derived from the standard normal pdf ϕ shown in equation* (2.9). *(a) Show that the normal cdf F relates to the standard normal cdf Φ by $F(y) = \Phi[(y - \mu)/\sigma]$. (b) From (a), show that $f(y) = (1/\sigma)\phi[(y - \mu)/\sigma]$, and show this is equation* (2.8).

Solution

a) *The cdf F of a normal distribution $N(\mu, \sigma^2)$ is defined as:*

$$F(y) = P(Y \le y) = \int_{-\infty}^{y} f(t)\, dt$$

where f is the pdf of $N(\mu, \sigma^2)$

To express $F(y)$ in terms of the standard normal cdf Φ, we can standardize the variable Y to convert it to the std normal form

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

$$F(y) = P(Y \le y) = P\left(\frac{Y - \mu}{\sigma} \le \frac{y - \mu}{\sigma}\right)$$

Since $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$, the probability $P(Z \le \frac{y-\mu}{\sigma})$, is the definition of the std normal cdf Φ

$$F(y) = \Phi\left(\frac{y - \mu}{\sigma}\right)$$

b) *The pdf $f(y)$ is the derivative of the cdf $F(y)$*

$$f(y) = \frac{d}{dy}F(y)$$

From (a) we know

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right)$$

Differentiate $F(y)$ with respect to y

$$f(y) = \frac{d}{dy}\Phi\left(\frac{y-\mu}{\sigma}\right) = \Phi'\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma}$$

Sixnce $\Phi'(z) = \phi(z)$, where $\phi(z)$ is the std normal pdf, we have:

$$f(y) = \phi\left(\frac{y-\mu}{\sigma}\right)\frac{1}{\sigma}$$

Ex 2.53

(Devid) *If Y is a standard normal random variable, with cdf Φ, what is the probability distribution of $X = \Phi(Y)$? Illustrate by randomly generating a million standard normal random variables, applying the cdf function $\Phi()$ to each, and plotting histograms of the (a) y values, (b) x values.*

Solution

Y is a standard normal variable, $Y \sim N(0, 1)$ The cdf of a standard normal variable, $\Phi(y) = P(Y \leq y)$, gives the probability that Y takes on a value less than or equal to y By defininig $X = \Phi(Y)$, we're transforming Y by it's own cdf, so X take values in [0,1] Since Y is std normal, and so is a continuos random variable

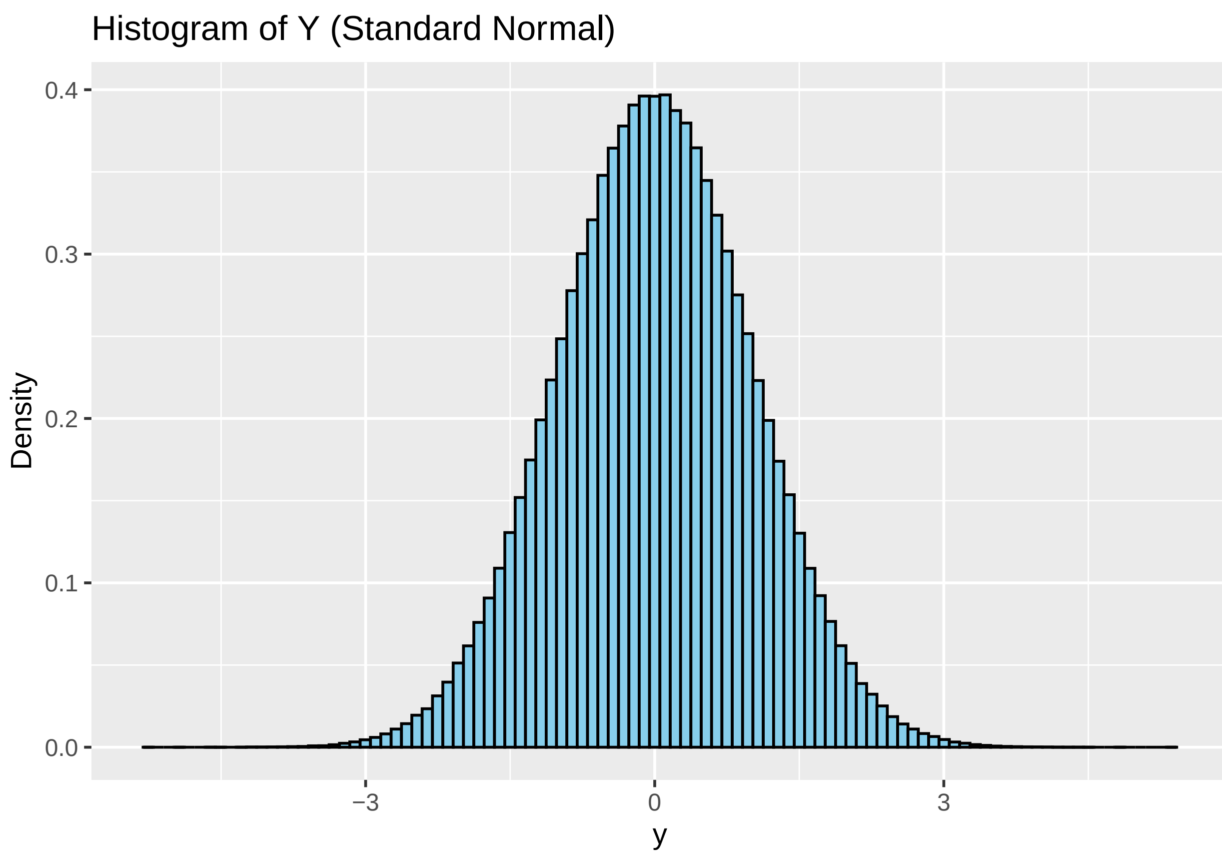
$$F_X = P(X \leq x) = P(F_Y(Y) \leq x)$$

We have to notice that $F_Y(Y) \leq x$ iff $Y \leq F_Y^{-1}(x)$, thus

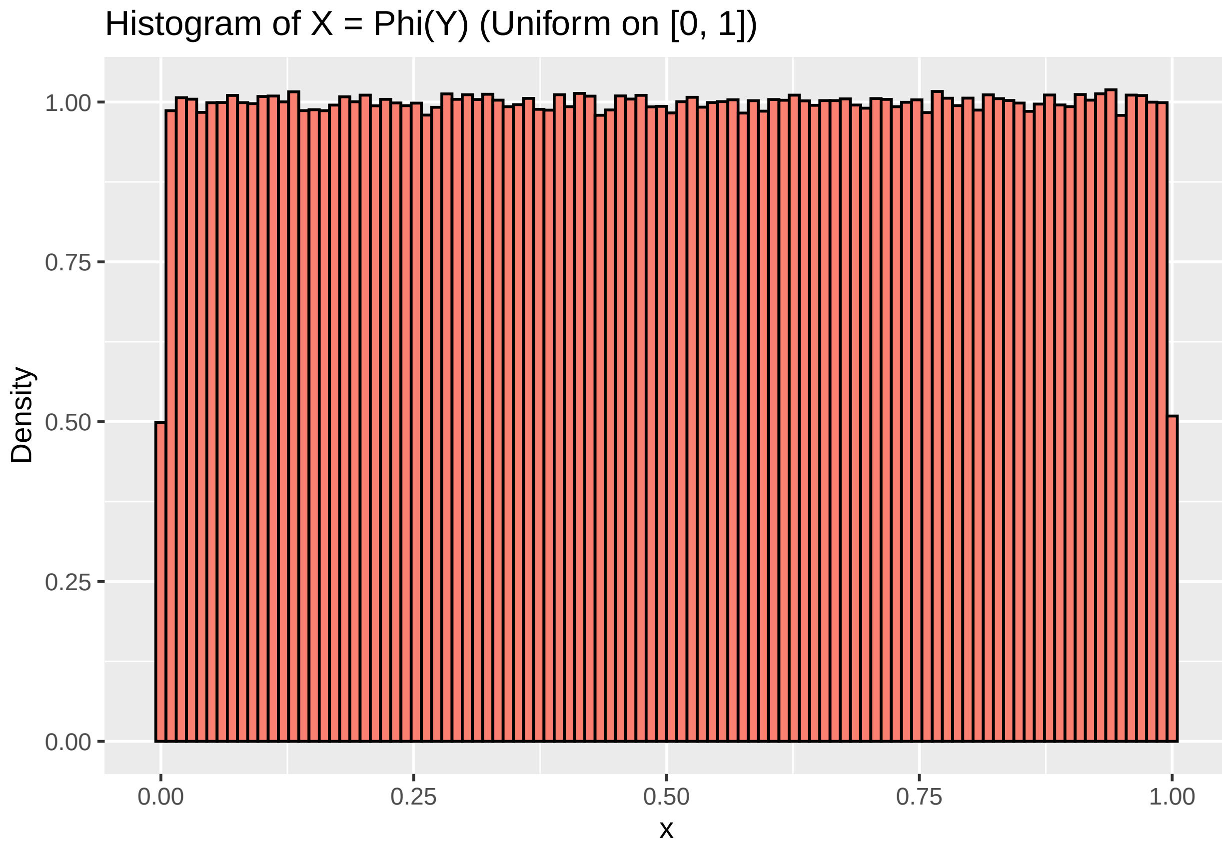
$$F_X = P(Y \leq F_Y^{-1}(x)) = F_Y(F_Y^{-1}(x)) = x$$

This shows that the cdf of X is $F_X(x) = x$, which is the cdf of a $U(0, 1)$

```r
set.seed(18)
n_samples <- 1e6
Y <- rnorm(n_samples)

X <- pnorm(Y)

if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
library(ggplot2)

ggplot(data.frame(Y), aes(x = Y)) +
  geom_histogram(aes(y = after_stat(density)), bins = 100, fill = "skyblue", color = "black") +
  ggtitle("Histogram of Y (Standard Normal)") +
  xlab("y") +
  ylab("Density")
```




```r
ggplot(data.frame(X), aes(x = X)) +
  geom_histogram(aes(y = after_stat(density)), bins = 100, fill = "salmon", color = "black") +
  ggtitle("Histogram of X = Phi(Y) (Uniform on [0, 1])") +
  xlab("x") +
  ylab("Density")
```



Ex 2.70

(Devid) *The beta distribution is a probability distribution over (0, 1) that is often used in applications for which the random variable is a proportion. The beta pdf is*

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, 0 \leq y \leq 1,$$

for parameters α and β, where $\Gamma()$ denotes the gamma function.

(a) Show that the uniform distribution is the special case $\alpha = \beta = 1$

(b) show that $\mu = E(Y) = \alpha/\alpha + \beta$

(c) Find $E(Y^2)$. Show that var$(Y) = \alpha\beta/(\alpha+\beta)^2(\alpha+\beta+1) = \mu(1-\mu)/(\alpha+\beta+1)$. For fixed $\alpha + \beta$, note that var(Y) decreases as μ approaches 0 or 1.

(d) Using a function such as dbeta in R, plot the beta pdf for (i) $\alpha = \beta = 0.5, 1.0, 10, 100$,(ii) some values of $\alpha > \beta$ and some values of $\alpha < \beta$. Describe the impact of and on the shape and spread.

Solution

a) *The assumption will be proof by the moment generating function $E(e^{tx}) = \int_0^1 e^{tx} f(x; \alpha, \beta)dx$*

$$f(y; 1, 1) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)} y^0(1-y)^0 = 1$$

$$E(e^{tx}) = \int_0^1 e^{tx} f(x;1,1)dx = \frac{e^t - 1}{t} = \frac{e^{tb} - e^{ta}}{t(b-a)}, b = 1, a = 0$$

But $\frac{e^{tb}-e^{ta}}{t(b-a)}$ *is the mgf of a* $U(0,1)$

b)

$$\mu = E(Y) = \int_{-\infty}^{+\infty} y \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{-\infty}^{+\infty} y^{\alpha}(1-y)^{\beta-1} dy =$$

We can notice that the solution of the integral is as a combination of Gamma function

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} =$$

Remembering that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}$$

c) *We start computing* $E(Y^2)$

$$E(Y^2) = \int_{-\infty}^{+\infty} y^2 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_{-\infty}^{+\infty} y^{\alpha+1}(1-y)^{\beta-1} dy =$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+\beta+2)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\alpha(\alpha+1)\Gamma(\alpha)}{(\alpha+\beta)(\alpha+\beta+1)\Gamma(\alpha+\beta)} =$$

$$= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

Using the equality $var(Y) = E(Y^2) - E(Y)^2$

$$E(Y^2) - E(Y)^2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \left(\frac{\alpha}{\alpha+\beta}\right)^2 =$$

$$= \frac{(\alpha^2+\alpha)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta)^2(\alpha+\beta+1)} =$$

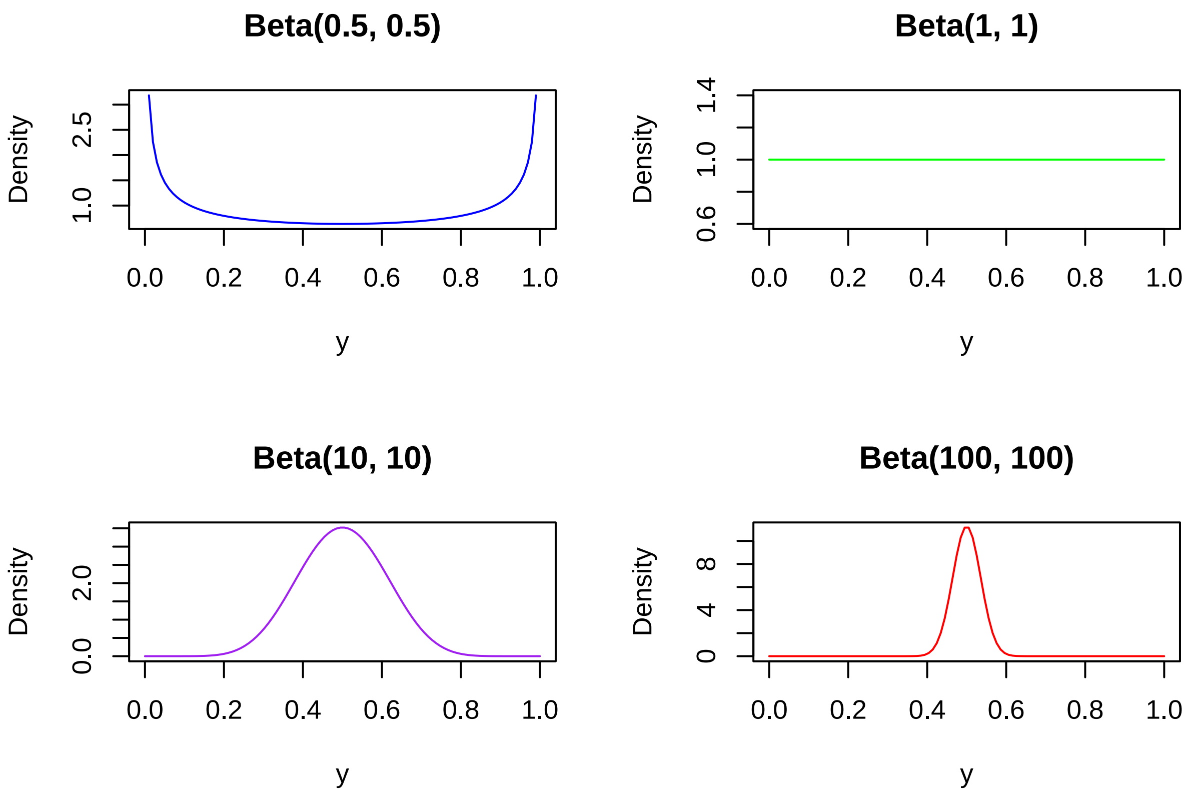
$$= \frac{\alpha^3 + \alpha^2\beta + \alpha^2 + \alpha\beta - \alpha^3 - \alpha^2\beta - \alpha^2}{(\alpha+\beta)^2(\alpha+\beta+1)} =$$

$$= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

We can write this result as $\frac{\alpha}{\alpha+\beta}\left(1 - \frac{\alpha}{\alpha+\beta}\right)\frac{1}{(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{(\alpha+\beta+1)}$

Then, $var(Y) = \frac{\mu(1-\mu)}{(\alpha+\beta)^2(\alpha+\beta+1)}$

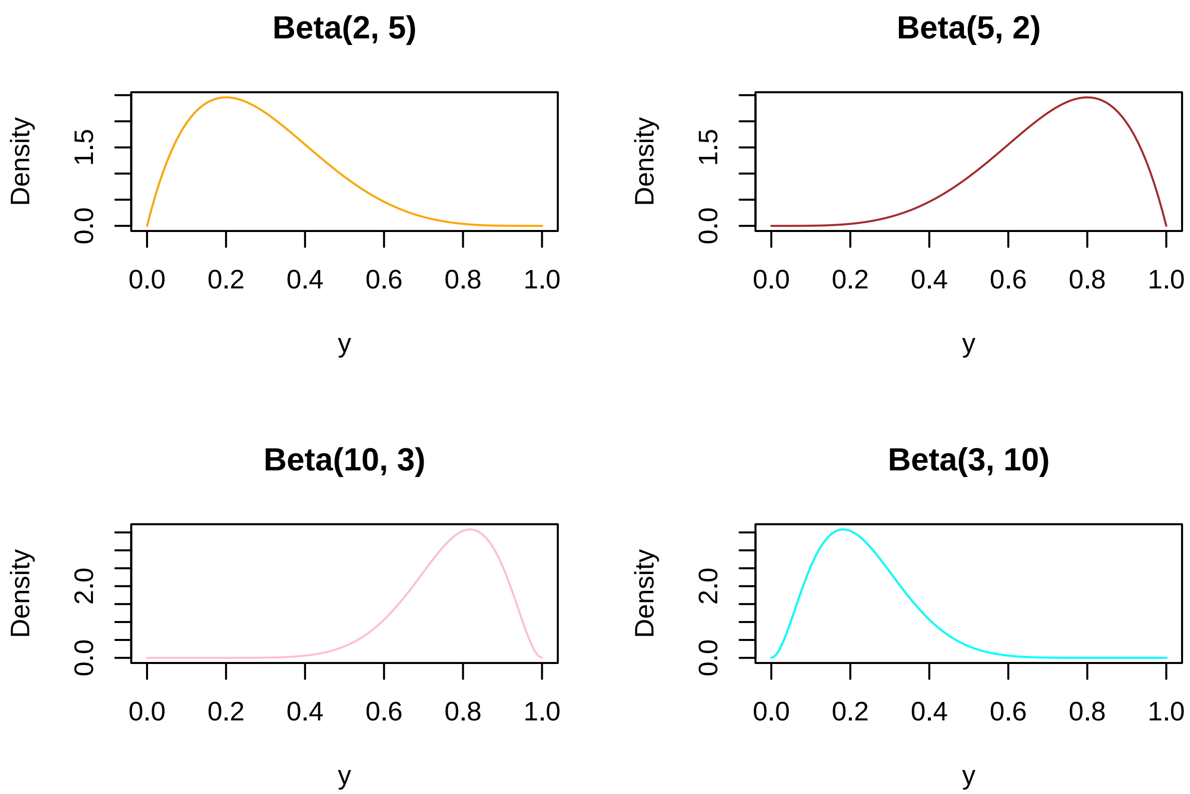
```r
par(mfrow = c(2, 2))
y <- seq(0, 1, length.out = 100)


plot(y, dbeta(y, 0.5, 0.5), type = "l", col = "blue", main = "Beta(0.5, 0.5)",
     ylab = "Density", xlab = "y")

plot(y, dbeta(y, 1, 1), type = "l", col = "green", main = "Beta(1, 1)",
     ylab = "Density", xlab = "y")

plot(y, dbeta(y, 10, 10), type = "l", col = "purple", main = "Beta(10, 10)",
     ylab = "Density", xlab = "y")

plot(y, dbeta(y, 100, 100), type = "l", col = "red", main = "Beta(100, 100)",
     ylab = "Density", xlab = "y")
```



d)

```r
par(mfrow = c(2, 2))

plot(y, dbeta(y, 2, 5), type = "l", col = "orange", main = "Beta(2, 5)",
     ylab = "Density", xlab = "y")

plot(y, dbeta(y, 5, 2), type = "l", col = "brown", main = "Beta(5, 2)",
     ylab = "Density", xlab = "y")

plot(y, dbeta(y, 10, 3), type = "l", col = "pink", main = "Beta(10, 3)",
```

```
     ylab = "Density", xlab = "y")

plot(y, dbeta(y, 3, 10), type = "l", col = "cyan", main = "Beta(3, 10)",
     ylab = "Density", xlab = "y")
```



Interpretation of the coefficients:

- $\alpha = \beta = 0.5$ *the distribution has a U-shape, with higher density near 0 and 1*
- $\alpha = \beta = 1$ *is the Uniform distribution*
- $\alpha = \beta > 1$ *the distribution becomes more peaked around the center*
- $\alpha > \beta$ *the distribution skews toward 1, with higher density on the right side of the interval*
- $\alpha < \beta$ *the distribution skews toward 0, with higher density on the left side of the interval*

FSDS - Chapter 3

Ex 3.18

Sunshine City, which attracts primarily retired people, has 90,000 residents with a mean age of 72 years and a standard deviation of 12 years. The age distribution is skewed to the left. A random sample of 100 residents of Sunshine City has $y = 70$ and $s = 11$... **Solution**

Add comments to the solution.

Ex 3.28

A survey is planned to estimate the population proportion π supporting more government action to address global warming. For a simple random sample, if π may be near 0.50, how large should n be so that the standard error of the sample proportion is 0.04? **Solution**

Add comments to the solution.

Ex 3.24 (use R)

Construct a population distribution that is plausible for Y = number of alcoholic drinks in the past day. Use the following steps:

Solution

Add comments to the solution.

FSDS - Chapter 4

Ex 4.14

Using the Students data file, for the corresponding population, construct a 95% confidence interval (a) for the mean weekly number of hours spent watching TV; (b) to compare females and males on the mean weekly number of hours spent watching TV. In each case, state assumptions, including the practical importance of each, and interpret results.

Solution

Add comments to the solution.

Ex 4.16

The Substance data file at the book's website shows a contingency table formed from a survey that asked a sample of high school students whether they have ever used alcohol, cigarettes, and marijuana. Construct a 95% Wald confidence interval to compare those who have used or not used alcohol on whether they have used marijuana, using (a) formula (4.13); (b) software. State assumptions for your analysis, and interpret results.

Solution

Add comments to the solution.

Ex 4.48

For a simple random sample of n subjects, explain why it is about 95% likely that the sample proportion has error no more than $1/\sqrt{n}$ in estimating the population proportion. (Hint: To show this "$1/\sqrt{n}$ rule," find two standard errors when $\pi = 0.50$, and explain how this compares to two standard errors at other values of π.) Using this result, show that $n = 1/M^2$ is a safe sample size for estimating a proportion to within M with 95% confidence.

Solution

Add comments to the solution.

FSDS - Chapter 5

Ex 5.2

When a government does not have enough money to pay for the services that it provides, it can raise taxes or it can reduce services. When the Florida Poll asked a random sample of 1200 Floridians which they preferred, 52% (624 of the 1200) chose raise taxes and 48% chose reduce services. Let π denote the population proportion of Floridians who would choose raising taxes. Analyze whether this is a minority of the population ($\pi < 0.50$) or a majority ($\pi > 0.50$) by testing $H0 : \pi = 0.50$ against $Ha : \pi \neq 0.50$. Interpret the P -value. Is it appropriate to "accept H0 ? Why or why not? **Solution**

Add comments to the solution.

Ex 5.12

The example in Section 3.1.4 described an experiment to estimate the mean sales with a proposed menu for a new restaurant. In a revised experiment to compare two menus, on Tuesday of the opening week the owner gives customers menu A and on Wednesday she gives them menu B. The bills average \$22.30 for the 43 customers on Tuesday ($s = 6.88$) and \$25.91 for the 50 customers on Wednesday ($s = 8.01$). Under the strong assumption that her customers each night are comparable to a random sample from the conceptual population of potential customers, show how to compare the mean sales for the two menus based on (a) the P -value of a significance test, (b) a 95% confidence interval. Which is more informative, and why? (When used in an experiment to compare two treatments to determine which works better, a two-sample test is often called an A/B test.).

Solution

Add comments to the solution.

Ex 5.50

A random sample of size 40 has $y = 120$. The P -value for testing $H0$: $\mu = 100$ against $H : \mu \neq 100$ is 0.057. Explain what is incorrect about each of the following interpretations of this P-value, and provide a proper interpretation. **Solution**

Add comments to the solution.