GroupE_HM1

De Stasio, Ortese, Tavano, Carli

2024-11-06

Contents

CS - Chapter 1

Ex 1.1

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 $P(X < \lambda)$ and the median of X.
3. Find the mean and variance of X.

X is defined as:

$$X \sim Exp(\lambda)$$

Here, the formula for the cumulative density function:

$$c.d.f = \int_0^x \lambda exp(-\lambda x)dx$$
$$= \int_0^x \lambda exp(-\lambda t)dt$$
$$[-\exp(-\lambda t)]_0^x = -exp(-\lambda x) + exp(-0) = 1 - exp(-\lambda x)$$

Cumulative density function:

$$F(x) = 1 - exp(-\lambda x)$$

Calculate the probability of:

$$P(X < \lambda) = 1 - exp(-\lambda^2)$$

Quantile function:

$$1 - exp(-\lambda x) = q_\alpha$$
$$x_\alpha = -\frac{ln(1 - q_\alpha)}{\lambda}$$

Calculate the median, quantile = 0.5.

$$1 - exp(-\lambda x) = 0.5$$
$$x_{0.5} = \frac{-ln(0.5)}{\lambda}$$

Calculate Mean and Variance:

$$E(X) = \int_0^{+\infty} x\lambda exp(-\lambda x)dx = \frac{1}{\lambda}$$
$$V(X) = E(X^2) - (E(X))^2 = \frac{1}{\lambda^2}$$

Ex 1.6

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).

Variance of X:

$$V(X) = \sigma_x^2$$

Variance of Y:

$$V(Y) = \sigma_y^2$$

Covariance of X,Y

$$Cov(X,Y) = \sigma_{xy}^2$$

Compute the formula for the Variance of the sum:

$$V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$$

$$= \sigma_x^2 + \sigma_y^2 + 2\sigma_{xy}^2$$

Compute the formula for the Variance of the subtraction:

$$\begin{aligned} V(X-Y) &= V(X) + V(Y) - 2Cov(X,Y) \\ &= \sigma_x^2 + \sigma_y^2 - 2\sigma_{xy}^2 \end{aligned}$$

CS - Chapter 3

Ex 3.5

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. Read the help file for solve before trying it.
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 $x = A^{-1}y$. 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?

```
#b)
time_0 = Sys.time()
A_inverse = solve(A)         #inverse of A
x1 = A_inverse %*% y
time_end = Sys.time()
time_process = time_end - time_0
print(sprintf("Time of the explicit inverse method: %.8f", time_process))
```

```
## [1] "Time of the explicit inverse method: 0.46250296"
```

```
mean_abs_difference = mean(abs(x1- x.true))
print(sprintf("Approximate mean absolute 'error' - explicit inverse method: %.8e", mean_abs_difference)
```

```
## [1] "Approximate mean absolute 'error' - explicit inverse method: 2.95683270e-11"
```

```
#c)
time_0 = Sys.time()
x2 = solve(A,y)              #directly solve x
time_end = Sys.time()
time_process2 = time_end - time_0
print(sprintf("Time of the direct solve method: %.8f", time_process2))
```

```
## [1] "Time of the direct solve method: 0.11369419"
```

```
mean_abs_difference2 = mean(abs(x2- x.true))
print(sprintf("Approximate mean absolute 'error' - direct solve method: %.8e", mean_abs_difference2))
```

```
## [1] "Approximate mean absolute 'error' - direct solve method: 1.35614203e-12"
```

Directly solving Ax=y using solve (A, y) is faster than the explicit inverse method: solve(A) to get A^{-1}, and than performing matrix multiplication $A^{-1}y$. So, the first method is more efficient.

Ex 3.6

The empirical cumulative distribution function for a set of measurements $x_i : i = 1, ...n$ is

$$\hat{F}(x) = \frac{\#(xi \leq x)}{n}$$

where $\#(xi \leq x)$ denotes number of xi values 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.

a. Function to calculate empirical c.d.f. from a vector of observations x:

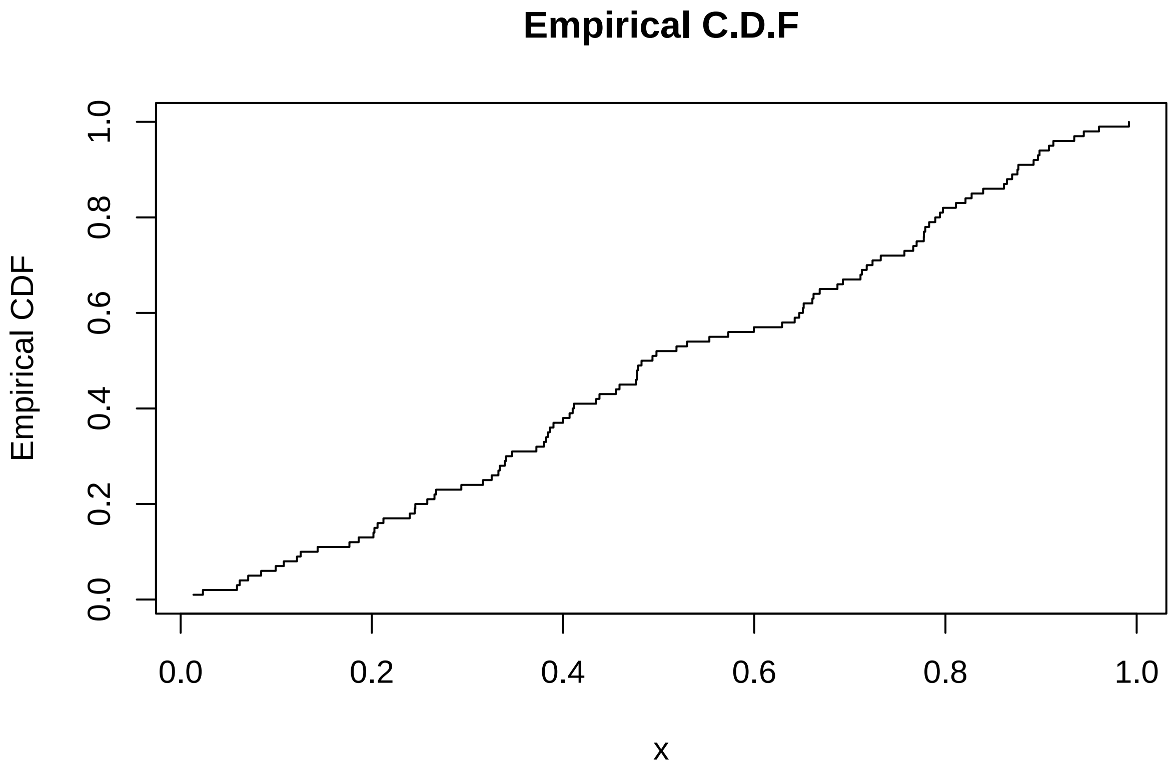
```r
empirical_cdf <- function(x) {

  sorted_x <- sort(x)

  ecdf_values <- sapply(x, function(val) {
    sum(sorted_x <= val) / length(sorted_x)
  })

  return(ecdf_values)
}
```

b. Modify the function to take an extra argument:

```r
empirical_cdf <- function(x, plot.cdf = FALSE) {

  sorted_x <- sort(x)

  ecdf_values <- sapply(x, function(val) {
    sum(sorted_x <= val) / length(sorted_x)
  })

  if (plot.cdf) {

    unique_vals <- sort(unique(x))
    ecdf_vals <- sapply(unique_vals, function(val) {
      sum(x <= val) / length(x)
    })

    plot(unique_vals, ecdf_vals, type = "s", xlab = "x", ylab = "Empirical CDF",
         main = "Empirical C.D.F")
  }

  return(ecdf_values)
}
```

Test the functions:

```r
set.seed(0)
x <- runif(100)

empirical_cdf(x, plot.cdf = TRUE)
```



```
##   [1] 0.93 0.22 0.32 0.56 0.95 0.14 0.94 0.98 0.63 0.58 0.04 0.16 0.12 0.66 0.35
##  [16] 0.75 0.52 0.70 1.00 0.33 0.77 0.97 0.17 0.62 0.10 0.23 0.36 0.01 0.34 0.89
##  [31] 0.30 0.50 0.57 0.51 0.13 0.85 0.65 0.81 0.08 0.71 0.41 0.84 0.60 0.79 0.55
##  [46] 0.54 0.80 0.02 0.47 0.72 0.67 0.48 0.87 0.43 0.19 0.05 0.07 0.25 0.53 0.64
##  [61] 0.39 0.96 0.24 0.45 0.27 0.61 0.21 0.49 0.74 0.06 0.90 0.29 0.86 0.31 0.28
##  [76] 0.46 0.92 0.88 0.37 0.76 0.99 0.42 0.69 0.38 0.26 0.73 0.15 0.68 0.09 0.20
##  [91] 0.11 0.18 0.03 0.59 0.91 0.78 0.82 0.44 0.40 0.83
```

FSDS - Chapter 2

Ex 2.8

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).

$$p = 0.01, \quad n = 100$$

a. Wrong calculation:

$$100(0.01) = 1.0$$

This calculation is wrong because it assumes that probabilities add linearly, ignoring the concept of cumulative probability; instead, the probability of catching a virus over multiple independent trips, should be calculated with a compound probability approach.

b. Right Calculation:

The probability of not catching a virus on a single trip is:

$$1 - p = 0.99$$

The probability of not catching a virus in all 100 trips (assuming independence) is:

$$(1 - p)^{100} = 0.99^{100}$$

Then, the probability of catching a virus at least once in 100 trips is:

$$P = 1 - (1 - p)^{100} = 1 - 0.99^{100} = 0.633$$

Ex 2.16

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?

b) Would you expect the Poisson distribution to better describe, or more poorly describe, the number of weekly admissions to the hospital for a rare disease? Why?

a.

$$Obs = (10, 8, 14, 7, 21, 44, 60)$$

$$E(X) = \frac{1}{n} \sum_i Obs_i = 23.24$$

$$V(X) = E(X^2) - E(X)^2 = \frac{1}{n}(\sum_i Obs_i^2) - 23.24^2 = 363.78$$

$$V(X) \neq E(X)$$

Variance and Mean are much different, thus Poisson doesn't describe the random variability of this phenomenon.

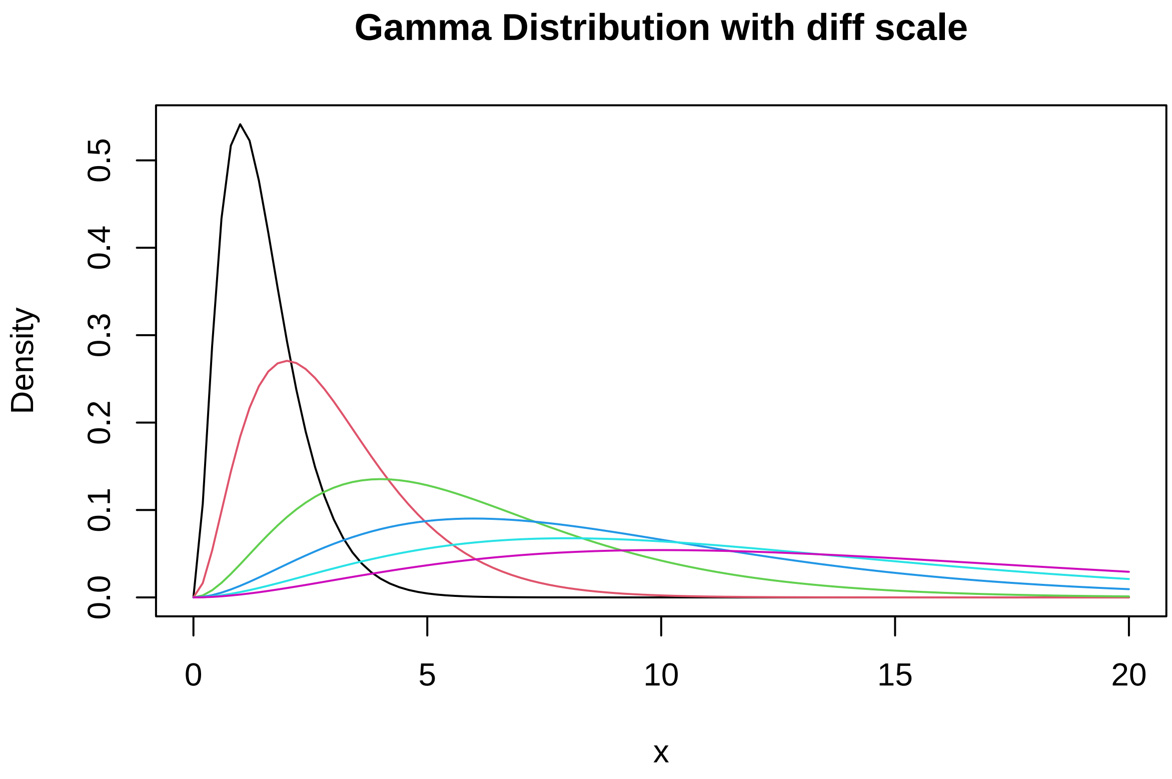
b. Poisson is suitable for counting events that occurs at random within a fixed interval, like in this case the number of weekly admissions (define interval) to the hospital for a rare events. So, mean and variance should be very similar and the model will be more efficient.

Ex 2.21

Plot the gamma distribution by fixing the shape parameter $k = 3$ and setting the scale parameter $= 0.5, 1, 2, 3, 4, 5$. What is the effect of increasing the scale parameter?

Plot Gamma distribution:

```r
k <- 3
scale_vector <- c(0.5, 1, 2, 3, 4, 5)

curve(dgamma(x, shape = k, scale = scale_vector[1]), from = 0, to = 20,
      col = 1, ylab = "Density", xlab = "x",
      main = "Gamma Distribution with diff scale")

for (i in 2:length(scale_vector)) {
  curve(dgamma(x, shape = k, scale = scale_vector[i]), from = 0, to = 20,
        col = i, add = TRUE)
}
```



Increasing the scale, parameter, the curve generated becomes more and more flat.

Ex 2.26

Refer to Table 2.4 cross classifying happiness with family income.
a) Find and interpret the correlation using scores $(i)(1, 2, 3)$ for each variable, $(ii)(1, 2, 3)$ for family income and $(1, 4, 5)$ for happiness.

b) Construct the joint distribution that has these marginal distributions and exhibits independence of X and Y .

```
#joint distribution table
joint_distribution <- matrix(c(0.080, 0.198, 0.079,
                               0.043, 0.254, 0.143,
                               0.017, 0.105, 0.081),
                             nrow = 3, byrow = TRUE)

#marginal probabilities
row_marginals <- c(0.357, 0.440, 0.203) # for Family Income (X)
col_marginals <- c(0.140, 0.557, 0.303) # for Happiness (Y)

# a)
# (i): Using scores (1, 2, 3) for X and Y
family_income_scores <- c(1, 2, 3)
happiness_scores <- c(1, 2, 3)

#expected values and variances
E_X <- sum(row_marginals * family_income_scores)
E_Y <- sum(col_marginals * happiness_scores)
var_X <- sum(row_marginals * (family_income_scores - E_X)^2)
var_Y <- sum(col_marginals * (happiness_scores - E_Y)^2)

#covariance
cov_XY <- 0
for (i in 1:3) {
  for (j in 1:3) {
    cov_XY <- cov_XY + (family_income_scores[i] - E_X) *
                        (happiness_scores[j] - E_Y) * joint_distribution[i, j]
  }
}

# Correlation coefficient
correlation_i <- cov_XY / sqrt(var_X * var_Y)
print(paste("Correlation with scores (1,2,3) for both:", correlation_i))
```

```
## [1] "Correlation with scores (1,2,3) for both: 0.1906625886631"
```

```
#(ii): Using scores (1, 2, 3) for X and (1, 4, 5) for Y
happiness_scores_2 <- c(1, 4, 5)

#expected value and variance for the new Y
E_Y_2 <- sum(col_marginals * happiness_scores_2)
var_Y_2 <- sum(col_marginals * (happiness_scores_2 - E_Y_2)^2)

#recalculate covariance with new Y
cov_XY_2 <- 0
for (i in 1:3) {
  for (j in 1:3) {
    cov_XY_2 <- cov_XY_2 + (family_income_scores[i] - E_X) *
                              (happiness_scores_2[j] - E_Y_2) * joint_distribution[i, j]
  }
}

#correlation coefficient for (ii)
correlation_ii <- cov_XY_2 / sqrt(var_X * var_Y_2)
print("Correlation with scores (1,2,3) for Family Income and (1,4,5) for Happiness:")
```

```
## [1] "Correlation with scores (1,2,3) for Family Income and (1,4,5) for Happiness:"
```

```
print(correlation_ii)
```

```
## [1] 0.1897729
```

While higher income levels are associated with a slight increase in reported happiness, the relationship is weak. Adjusting the scale for Happiness (from (1,2,3) to (1,4,5), emphasizing the higher levels "Pretty happy" and "Very happy") has little impact on the correlation, suggesting that the observed association remains consistent across different scoring systems. So, family income is not a strongly determining factor for happiness: earning more is correlated with a slightly higher likelihood of being happier, but it's not a robust link.

```
#b)
print("Observed joint distribution:")
```

```
## [1] "Observed joint distribution:"
```

```
print(joint_distribution)
```

```
##       [,1]  [,2]  [,3]
## [1,] 0.080 0.198 0.079
## [2,] 0.043 0.254 0.143
## [3,] 0.017 0.105 0.081
```

```
#independent joint distribution
independent_distribution <- outer(row_marginals, col_marginals)

print("Independent joint distribution (assuming X and Y are independent):")
```

```
## [1] "Independent joint distribution (assuming X and Y are independent):"
```

```
print(independent_distribution)
```

```
##         [,1]     [,2]     [,3]
## [1,] 0.04998 0.198849 0.108171
## [2,] 0.06160 0.245080 0.133320
## [3,] 0.02842 0.113071 0.061509
```

The differences, between observed joint distribution and independent joint distribution, indicate potential dependencies or interactions between the two variables X and Y.

Ex. 2.52

The p.d.f. of a $N(\mu, \sigma^2)$ distribution can be derived from the standard normal p.d.f. ϕ shown in equation (2.9).

a) Show that the normal c.d.f. F relates to the standard normal c.d.f. 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).

(2.9):

$$\phi(z) = \frac{1}{\sqrt{(2\pi)}} e^{\frac{-z^2}{2}} \qquad -\infty < z < +\infty$$

$$f(y;\mu,\sigma) = (\frac{1}{\sqrt{2\pi\sigma}})e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

a) The c.d.f. F relates to the standard normal c.d.f. ϕ by

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

Since $F(y) = P(Y \le y)$, we can write:

$$Y = \sigma Z + \mu \Rightarrow Z = \frac{Y-\mu}{\sigma},$$

where $Y \sim N(\mu, \sigma^2)$. Therefore:

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

Since $Z = \frac{Y-\mu}{\sigma}$ is the standard normal variable, we find:

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

b) From (a), we have $F(y) = \phi\left(\frac{y-\mu}{\sigma}\right)$.

The p.d.f. $f(y)$ is the derivative of $F(y)$ with respect to y:

$$f(y) = \frac{d}{dy}F(y) = \frac{d}{dy}\phi\left(\frac{y-\mu}{\sigma}\right).$$

Using the chain rule, we get:

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

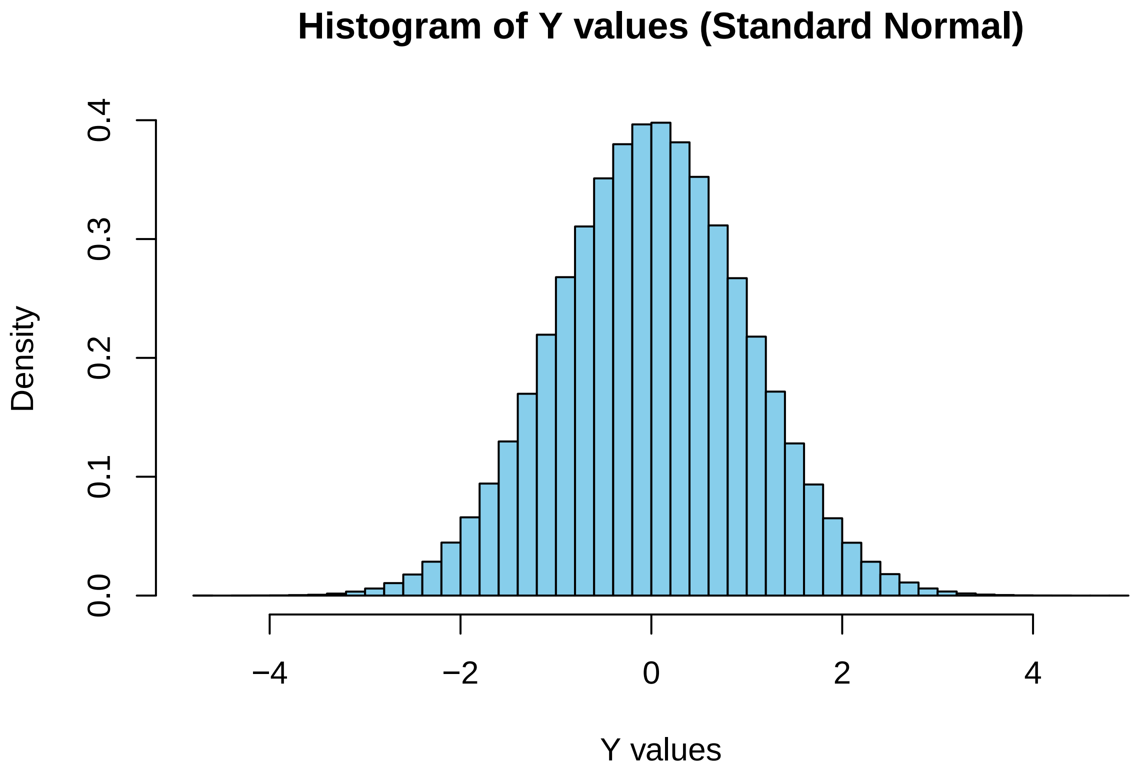
Substituting $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ with $z = \frac{y-\mu}{\sigma}$, we obtain:

$$f(y;\mu,\sigma) = \frac{1}{\sigma}\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}\right).$$

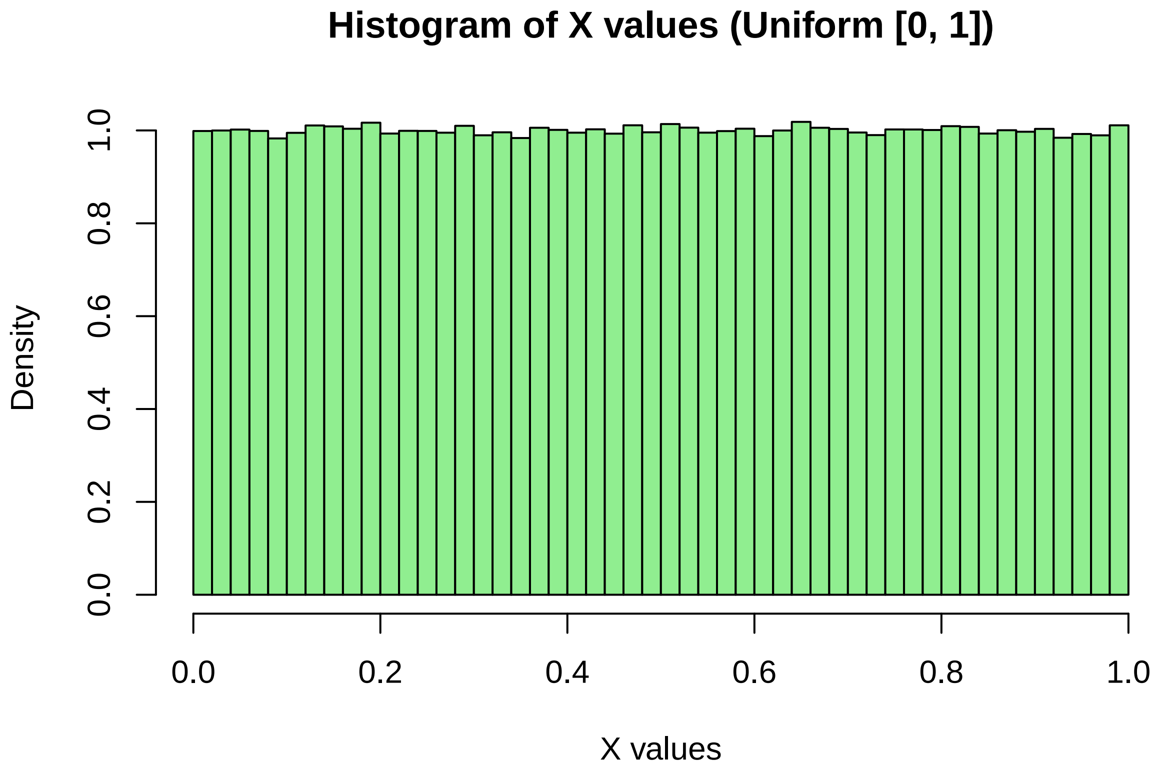
Ex 2.53

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.

```
set.seed(0)
n <- 1000000  #one million samples


y_values <- rnorm(n)   #standard normal random variables
x_values <- pnorm(y_values) #applying the cdf function


hist(y_values, breaks = 50, main = "Histogram of Y values (Standard Normal)",
     xlab = "Y values", col = "skyblue", probability = TRUE) #plot histogram of the Y values
```



```
hist(x_values, breaks = 50, main = "Histogram of X values (Uniform [0, 1])",
     xlab = "X values", col = "lightgreen", probability = TRUE)  #plot histogram of the X values
```



X follows a $Uniform(0, 1)$ distribution, the histogram confirms it.

Ex 2.70

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 p.d.f. is

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

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

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

Case where $\alpha = \beta = 1$:

$$\begin{aligned} f(y; 1, 1) &= \frac{\Gamma(1 + 1)}{\Gamma(1)\Gamma(1)} y^{1-1}(1 - y)^{1-1}, \quad 0 \le y \le 1 \\ &= \frac{\Gamma(2)}{\Gamma(1)^2} = \frac{1!}{1^2} = 1 \end{aligned}$$

By definition, this is defined over $[0, 1]$, so it is the p.d.f. of a Uniform distribution.

b) Show that $\mu = E(Y) = \frac{\alpha}{(\alpha+\beta)}$

$$\mu = E(Y) = \int_0^1 yf(y;\alpha,\beta)dy = \int_0^1 y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}dy$$

$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 y^{\alpha}(1-y)^{\beta-1}dy =$$

This integral correspond to the Beta distribution with parameters $\alpha+1$ e β.

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

Since $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, we can rewrite is as:

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

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

$$E(Y^2) = \int_0^1 y^2 f(y;\alpha,\beta)dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 y^{\alpha+1}(1-y)^{\beta-1}dy =$$

As before, we use the Beta distribution.

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

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

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

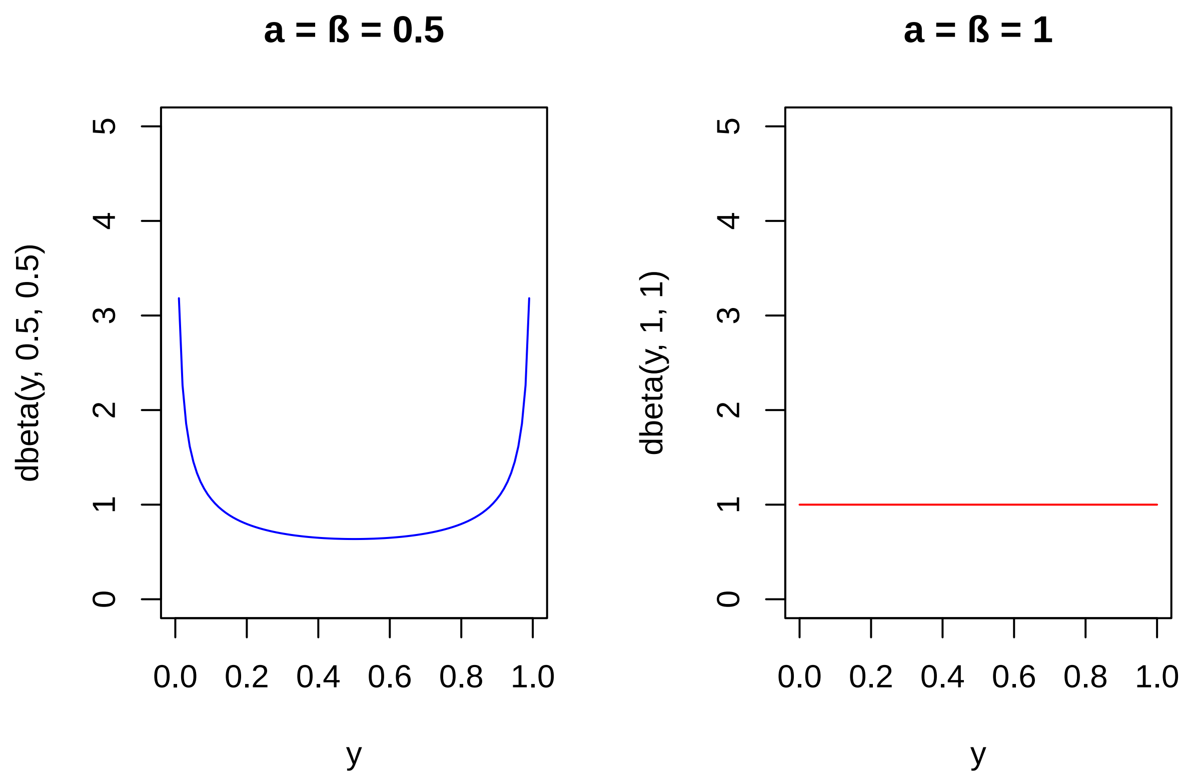
$$\frac{\alpha^3+\alpha^2+\alpha^2\beta-\alpha^2-\alpha^3-\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Since $\mu = \frac{\alpha}{\alpha+\beta}$ and $1-\mu = \frac{\beta}{\alpha+\beta}$, we can rewrite the variance as:

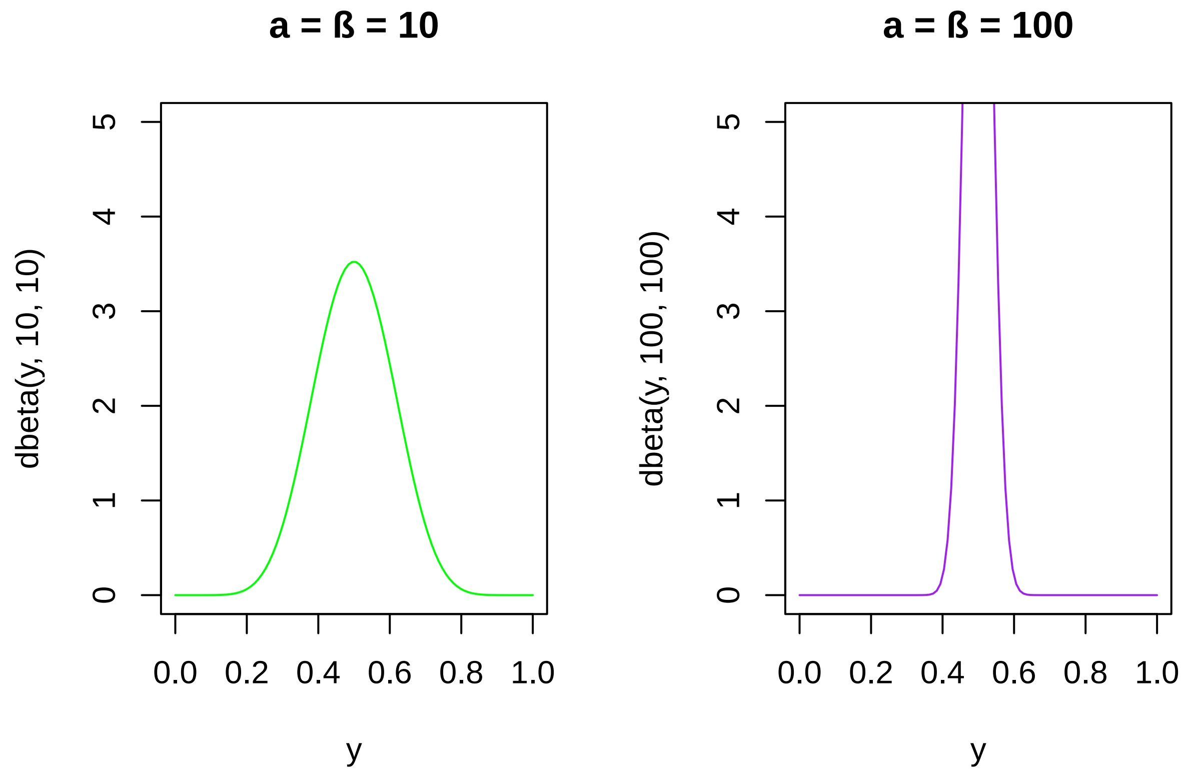
$$Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{\alpha+\beta+1} = \mu(1-\mu)\frac{1}{\alpha+\beta+1}$$

d) Using a function such as dbeta in R, plot the beta function 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 α and β on the shape and spread.

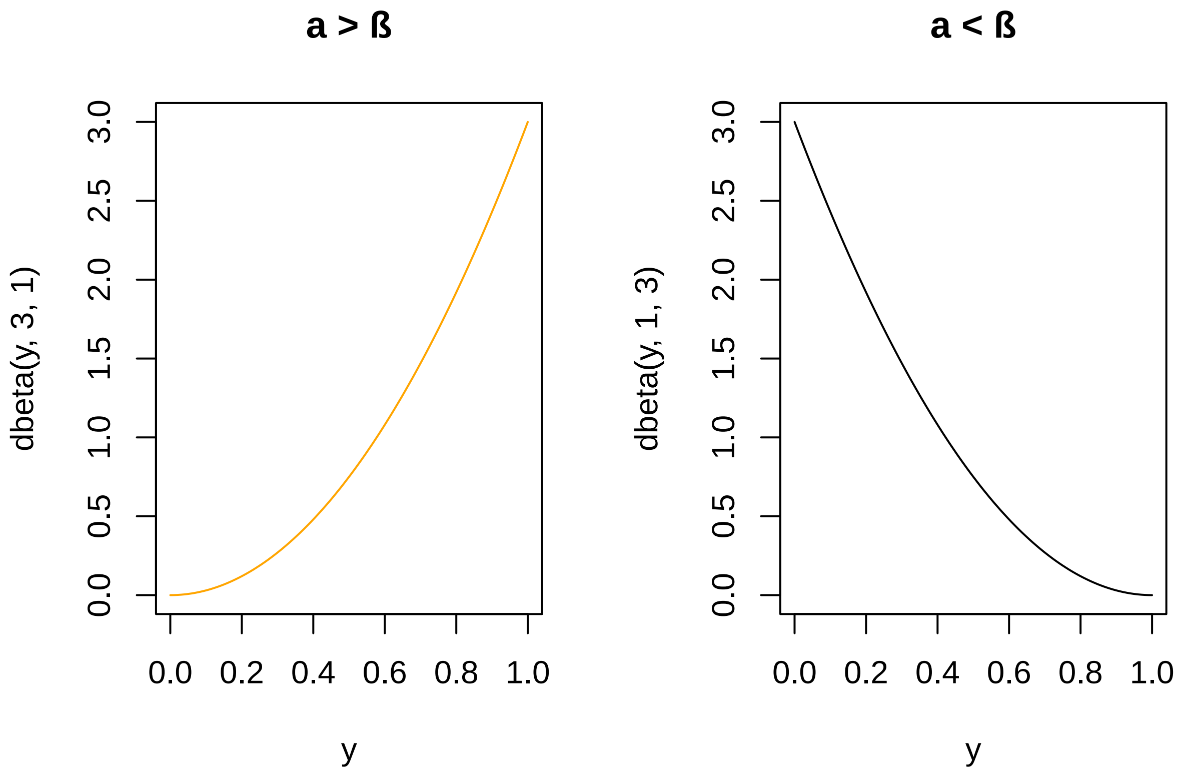
```
y <- seq(0, 1, length = 100)

# (i) Plot for different values of   =
par(mfrow = c(1, 2)) # arrange plots in 2x2 grid
plot(y, dbeta(y, 0.5, 0.5), type = "l", col = "blue", ylim = c(0, 5), main = "  =   = 0.5")
plot(y, dbeta(y, 1, 1), type = "l", col = "red", ylim = c(0, 5), main = "  =   = 1")
```



```
plot(y, dbeta(y, 10, 10), type = "l", col = "green", ylim = c(0, 5), main = "  =   = 10")
plot(y, dbeta(y, 100, 100), type = "l", col = "purple", ylim = c(0, 5), main = "  =   = 100")
```



```
# (ii) Plot for some values of   >   and   <
par(mfrow = c(1, 2))
plot(y, dbeta(y, 3, 1), type = "l", col = "orange", main = "  >  ")
plot(y, dbeta(y, 1, 3), type = "l", col = "black", main = "  <  ")
```



Discussion:

$(\alpha = \beta)$:

- Symmetry : the Beta distribution is symmetric around y = 0.5
- Spread:
  - = 1, the distribution is uniform over (0, 1).
  - < 1: the distribution becomes U-shaped (with peaks near 0 and 1), implying values close to 0 and 1 are more likely.
  - > 1 : the distribution becomes more peaked around 0.5, with lower spread as α and β increase.

$(\alpha \neq \beta)$:

- Asymmetry: the Beta distribution becomes skewed:
  - $(\alpha \leq \beta)$: skews to the right, favoring values closer to 1
  - $(\alpha \leq \beta)$: skews to the left, favoring values closer to 0
- Spread:
  - $(\alpha + \beta)$ smaller, the spread is broader
  - $(\alpha + \beta)$ increases, the distribution concentrates around a mode defined by:

$$\text{Mode} = \frac{\alpha - 1}{\alpha + \beta - 2}$$

- $(\alpha + \beta)$ larger, the distribution becomes more sharply, reducing the variance and increasing the probability density around the mode.

In summary, the parameters α and β determinate the skewness and the sum of them influences the spread.

FDFS - 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 sample mean = 70 and a sample standard deviation = 11.

a) Describe the center and spread of the (i) population distribution, (ii) sample data distribution. What shape does the sample data distribution probably have? Why?

The population distribution center is given by the mean = 72 and the spread is given by the standard deviation = 12. The sample data distribution center is given by the sample mean = 70 and the spread is given by the sample standard deviation = 11. The sample data distribution will probably have the same shape of the population distribution because the sample tends to reflect the population's shape, especially with a large sample size.

b) Find the center and spread of the sampling distribution of the sample mean for $n = 100$. What shape does it have and what does it describe?

The sampling distribution of the sample mean is approximately a normal distribution with

$$\mu_{\overline{y}} = \mu = 72$$

and

$$\sigma_{\overline{y}} = \frac{\sigma}{\sqrt{n}} = \frac{12}{\sqrt{100}} = 1.2$$

This holds true for the central limit theorem, when n is sufficiently large, like in this case. The sampling distribution of the sample mean describes how we expect the sample mean to vary when repeatedly taking a random sample of size

$$n$$

from the population.

c) Explain why it would not be unusual to sample a person of age 60 in Sunshine City, but it would be highly unusual for the sample mean to be 60, for a random sample of 100 residents.

Since the the population mean is 72 and the standard deviation is 12, we expect most individual ages to fall within one standard deviation of the mean, which means between 60 and 84 years. So it's not unusual to sample one person in this range. However the sampling distribution for the sample mean of a sample of size 100 has mean 72 and standard deviation of 1.2, so 60 is 10 standard deviations away from the expected value, which is highly unusual.

d) Describe the sampling distribution of the sample mean: (i) for a random sample of size $n = 1$; (ii) if you sample all $90,000$ residents.

If I repeatedly take a sample of size n=1 I get a sampling distribution of the sample mean that strictly corresponds to the population distribution, since the mean of a sample of size 1 is the value of the sample itself. If I repeatedly take a sample of the same size as the population I always get the same value, which is the population mean, so the distribution would just have mean equal to the population mean and standard deviation of 0.

Ex 3.24

Construct a population distribution that is plausible for Y = number of alcoholic drinks in the past day.
a) Simulate a single random sample of size $n = 1000$ from this population to reflect results of a typical sample survey. Summarize how the sample mean and standard deviation resemble those for the population. (Alternatively, you can do this and part (b) using an app, such as the Sampling Distribution for the Sample Mean (Discrete Population) app at www.artofstat.com/web-apps using the Build Custom Distribution option.)

Since the number of drinks is an integer we should use a discrete distribution. Also the minimum value is 0, so the distribution should only admit non-negative values. We could use a Poisson distribution with parameter $\lambda = 0.5$.

```
n = 1000
lambda = 0.5
set.seed(123)
sample = rpois(n,lambda)

sample_mu = mean(sample)
sample_sigma = sd(sample)
sample_mu
```

```
## [1] 0.489
```

```
sample_sigma
```

```
## [1] 0.6930796
```

The sample mean and sample standard deviation tend to be similar to the true parameter, given a large sample size.

b) Now draw $10,000$ random samples of size 1000 each,to approximate the sampling distribution of Y . Report the mean and standard deviation of this simulated sampling distribution, and compare to the theoretical values. Explain what this sampling distribution represents.

```
n = 1000
R = 10000
lambda = 0.5
set.seed(123)
mu = array(0, dim = (R))
for(i in 1:R){
  sample = rpois(n,lambda)
  mu[i] = mean(sample)
}
mean = mean(mu)
sd = sd(mu)
mean
```

```
## [1] 0.4998635
```

```
sd
```

```
## [1] 0.02225101
```

```
theoretical_mean = lambda
theoretical_sd = sqrt(lambda/n)
theoretical_mean
```

```
## [1] 0.5
```

```
theoretical_sd
```

```
## [1] 0.02236068
```

This sampling distributions represents how the sample mean could vary when taking a random sample of size 1000 from the population. It's close to the values that we were expecting thanks to the Central Limit Theorem.

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?

To determine the required sample size n for estimating a population proportion π with a desired standard error SE, we can use the formula for standard error of the sample proportion.

$$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$$

So we insert our hypothesis into the previous formula.

We want:

$$0.4 = \sqrt{\frac{0.50(1-0.50)}{n}}$$

$$0.4^2 = \frac{0.50(1-0.50)}{n}$$

$$0.0016 = \frac{0.25}{n}$$

And finally,

$$n = 156.25$$

Since n must be a whole number, round up to $n = 157$ to ensure the standard error is at most 0.04.

The required sample sizen should be at least 157 to achieve a standard error of 0.04 when the population proportion π is around 0.50.

FDFS - 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.

```
students = read.csv("students.csv")
head(students)
```

```
##   subject gender age hsgpa cogpa dhome dres tv sport news aids veg affil ideol
## 1       1      0  32   2.2   3.5     0  5.0  3     5    0    0   0     2     6
## 2       2      1  23   2.1   3.5  1200  0.3 15     7    5    6   1     1     2
## 3       3      1  27   3.3   3.0  1300  1.5  0     4    3    0   1     1     2
## 4       4      1  35   3.5   3.2  1500  8.0  5     5    6    3   0     3     4
## 5       5      0  23   3.1   3.5  1600 10.0  6     6    3    0   0     3     1
## 6       6      0  39   3.5   3.5   350  3.0  4     5    7    0   1     1     2
##   relig abor affirm life
## 1     2    0      0    1
## 2     1    1      1    3
## 3     2    1      1    3
## 4     1    1      1    2
## 5     0    1      0    2
## 6     1    1      1    3
```

```
#a) C.I. for the mean weakly number of hours spent watching tv
tv = students$tv
n = length(tv)
alpha = 0.05
mean_tv = mean(tv)
sd_tv = sqrt(var(tv))
lower_bound = mean_tv - qt(1- alpha/2, df = n-1) * (sd_tv/ sqrt(n))
upper_bound = mean_tv + qt(1- alpha/2, df = n-1) * (sd_tv/ sqrt(n))
cat("a) Confidence interval at 95%: [", round(lower_bound, 2), ",", round(upper_bound, 2), "]\n")
```

```
## a) Confidence interval at 95%: [ 5.53 , 9 ]
```

```
#b) compare females and males on the mean weekly number of hours spent watching TV
females <- students$tv[students$gender == 0]
males <- students$tv[students$gender == 1]

# (test di Welch)
t_test_result <- t.test(females, males, var.equal = FALSE)

print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  females and males
## t = -0.84995, df = 56.249, p-value = 0.399
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.980854  2.013112
## sample estimates:
## mean of x mean of y
##  6.500000  7.983871
```

From the results of Welch's t-test, we can draw the following conclusions:

- Difference in averages: The average hours per week spent watching TV for females is 6.5, while for males it is 7.98.
- Hypothesis: The null hypothesis is that there is no difference in the weekly averages of hours spent in front of TV between the two groups (females and males).
- Value of p: The value of $p = 0.399$ indicates that we do not have sufficient statistical evidence to reject the null hypothesis at a conventional level of significance (e.g, α=0.05). This means that there is no significant difference between the mean TV hours for males and females in the considered sample.
- Confidence Interval: The 95% confidence interval for the difference of the averages ranges from -4.98 to 2.01. Since the interval includes zero, this further strengthens the conclusion that the difference is not significant.

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.

(4.13)

$$(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

a)

We define:

Total number of students who have used alcohol: $n_1 = 955 + 994 = 1949$
Total number of students who have not used alcohol: $n_2 = 5 + 322 = 327$
Number of students who have used both alcohol and marijuana: $x_1 = 955$
Number of students who have used marijuana but not alchool: $x_2 = 5$

Calculate proportions:

Proportion of students who have used marijuana among those who have used alcohol: $\hat{p}_1 = \frac{x_1}{n_1} = 0.490$
Proportion of students who have used marijuana among those who have not used alcohol: $\hat{p}_2 = \frac{x_2}{n_2} = 0.015$

Using formula (4.13) to compute Confidence Interval, with $z_{\alpha/2} = 1.96$.

$$C.I. = (0.490 - 0.015) \pm 1.96\sqrt{\frac{0.490(1-0.490)}{1949} + \frac{0.015(1-0.015)}{327}}$$
$$C.I. = [0.4488, 0.5006]$$

The resulting confidence interval does not contain 0, it suggests a statistically significant difference in marijuana use between students who have used alcohol and those who have not, with a 95% level of confidence.

b)

```
substance = read.csv("substance.csv")
head(substance)
```

```
##   alcohol cigarettes marijuana count
## 1     yes        yes       yes   911
## 2     yes        yes        no   538
## 3     yes         no       yes    44
## 4     yes         no        no   456
## 5      no        yes       yes     3
## 6      no        yes        no    43
```

```
pi_1 <- 0.490
pi_2 <- 0.015
z <- qnorm(0.975)
n1 <- 1949
n2 <- 327

lower_b = (pi_1 - pi_2) - z * sqrt(((pi_1*(1-pi_1))/n1) + ((pi_2*(1-pi_2))/n2))
upper_b = (pi_1 - pi_2) + z * sqrt(((pi_1*(1-pi_1))/n1) + ((pi_2*(1-pi_2))/n2))
lower_b
```

```
## [1] 0.4491907
```

```
upper_b
```

```
## [1] 0.5008093
```

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 $\frac{1}{\sqrt{n}}$ in estimating the population proportion. (Hint: To show this "$\frac{1}{\sqrt{n}}$ 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 $n = \frac{1}{M^2}$ is a safe sample size for estimating a proportion to within M with 95% confidence.

First of all, we evaluate the standard error of the proportion $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$, which tells us the typical deviation of the sample proportion from the true population proportion p for a sample size n.

Then, we aim to maximize that when p = 0.5, since $p(1-p)$ is maximized at p = 0.5. Hence,

$$SE(\hat{p}) = \sqrt{\frac{0.5(1-0.5)}{n}} = \frac{1}{2\sqrt{n}}$$

This result tells us that, at most, the standard error of $\hat{p}$ is $\frac{1}{2\sqrt{n}}$. This value represents the worst-case scenario for the variability of the sample proportion.

For a 95% confidence interval, we typically consider an interval of about two standard errors around the sample proportion $\hat{p}$ to capture the true proportion p approximately 95% of the time. So, the margin of error for a 95% confidence interval can be approximated by:

$$ME = 2SE(\hat{p})$$

Using $SE(\hat{p}) = \frac{1}{2\sqrt{n}}$, we get:

$$ME = 2\frac{1}{2\sqrt{n}} = \frac{1}{\sqrt{n}}$$

Therefore, regardless of the true value of p, it is approximately 95% likely that the sample proportion $\hat{p}$ will be within $\frac{1}{\sqrt{n}}$ of the true proportion p.

In conclusion, we determine the sample size that ensures we can estimate the proportion p within a margin of error M with 95% confidence.

$$\frac{1}{\sqrt{n}} \leq M$$

Solving for n we obtain:

$$\sqrt{n} \geq \frac{1}{M} => n \geq \frac{1}{M^2}$$

This means that if we choose a sample size of $n = \frac{1}{M^2}$, we can be confident that the sample proportion $\hat{p}$ will estimate the population proportion p within a margin of error M with approximately 95% confidence. This sample size ensures that the error will be no larger than M in the worst-case scenario (i.e., when p=0.5).

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 H_0 \$ \$ = 0.50 against H_a: $\pi \neq 0.50$. Interpret the P-value. Is it appropriate to "accept H_0"? Why or why not?

$\hat{\pi} = \frac{624}{1200} = 0.52$ $\pi_{H_0} = 0.50$ $n = 1200$

Calculate Test Statics:

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Thus,

$$z_{oss} = \frac{0.52 - 0.50}{\sqrt{\frac{0.50(1-0.50)}{1200}}} = 1.387$$

Find p-value:

$$P(|Z| > z_{oss}) = 0.165$$

```
#compute p-value using R
z<- (0.52-0.50)/sqrt((0.50^2)/1200)
q=2*(1- pnorm(z))
q
```

```
## [1] 0.1658567
```

The P-value of approximately 0.165 indicates that there is a 16.5% probability of obtaining a sample proportion as extreme as (or more extreme than) 0.52 if the true population proportion were actually 0.50.

If we choose to fix $\alpha = 0.05$, 0.025 or 0.001, in all cases the P-value (0.165) is greater than any typical significance level, thus we do not reject the null hypothesis. There is not enough evidence to conclude that the true proportion of Floridians who prefer raising taxes differs from 50%.

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).

a) Let's compute the statistical test, using the pooled estimator in order to compute the P-value.

```
mean1 <- 22.30
mean2 <- 25.91
n <- 43
m <- 50
s <- 6.88
t <- 8.01
#pooled estimator
st= sqrt(((n-1) *s^2 + (m-1)*t^2)/(n + m - 2))
print(st)
```

```
## [1] 7.50962
```

```
t_stat<- (mean1-mean2-0)/(st*sqrt(1/n+1/m))
print(t_stat)
```

```
## [1] -2.311357
```

```
q=2*(1- pt(abs(t_stat), df = n + m - 2))
print(q)
```

```
## [1] 0.02307139
```

The observed P-value is 0.023. This can be informative with regards to the significance level of the test we decide to use. For example, if we choose $\alpha = 0.05$, then H_0 will be refused. Since the P-value is not enough we compute the confidence interval with 1-$\alpha = 0.95$.

```
alpha = 0.05
confidence_interval <- (mean1 - mean2) + (c(-1,1) * qt(1 - alpha/2, df = n+m-2)  * st * sqrt(1/n + 1/m)
confidence_interval
```

```
## [1] -6.7124295 -0.5075705
```

Since 0 is not included in the confidence interval, we have strong evidence that the means of the two population are different.

Ex 5.50

A random sample of size 40 has $\bar{y} = 120$. The P-value for testing $H_0 : \mu = 100$ against $H_a : \mu \neq 100$ is 0.057. Explain what is incorrect about each of the following interpretations of this P-value, and provide a proper interpretation.
a) The probability that H_0 is correct equals 0.057.
b) The probability that $y = 120$ if H_0 is true equals 0.057.
c) The probability of Type I error equals 0.057.
d) We can accept H_0 at the $\alpha = 0.05$ level.

(a)

No, the p-value doesn't correspond to the probability that H_0 is true, since H_0 is not event an event and cannot be evaluated as True or False.

The p-value of 0.057 means that if H_0 is true, there is a α * chance of obtaining a sample mean greater or equal than 120.

(b)

The p-value is not the probability to observe the sample mean exactly equal to 120.

If H_0 is true, the probability of obtaining a sample mean of 120 or more is 0.057.

(c)

The probability of a Type I error corresponds to the significance level of the test α, not to the p-value.

The p-value, instead is the measure of the evidence against H_0.

(d)

The result of statistical test, and p-value in particular should never taken without considering context-specific knowledge.

Since the p-value is slightly above the significance level of 0.05, we do not have enough evidence to reject H_0 at the 5% significance level. Thus, we fail to reject the null hypothesis.

Correct interpretation: If the population mean is 100 (H_0 is true), there is a 0.057 chance of obtaining a sample mean greater or equal to 120. Since the p-value is slightly above the significance level of 0.05, it is not reasonable to reject the null hypothesis at 5% level.