Carson Slater
STA 5380 Homework #2

1. Computer screening of tax returns is used to flag tax forms that need to be investigated further. The method correctly flags 85% of all erroneous returns and incorrectly flags 5% of error-free returns. The pool of tax returns submitted for screening contains 15% with errors.

   (a) A tax return is randomly screened. What is the probability that it is flagged by the computer and has an error?
   Let F be the event a computer flags a tax form, and E be the event that a tax form contains an error. We know that $P(F|E) = 0.85$, and $P(E) = 0.15$. We also know that $P(F|E^c) = 0.05$. We want to find $P(F \cap E)$. By the definition of conditional probability, we know that

   $$\begin{aligned} P(F \cap E) &= P(F|E)P(E) \\ &= 0.85(0.15) \\ &= 0.1275. \end{aligned}$$

   (b) What is the probability that a random tax form is flagged by the computer?
   We know by the Law of Total Probability that

   $$\begin{aligned} P(F) &= P(F|E)P(E) + P(F|E^c)P(E^c) \\ &= 0.85(0.15) + 0.05(0.85) \\ &= 0.17. \end{aligned}$$

   (c) If a tax return is flagged by the computer, what is the probability that there is an error on it?
   By Bayes' Theorem, we have that

   $$\begin{aligned} P(E|F) &= \frac{P(F \cap E)}{P(F)} \\ &= \frac{0.1275}{0.17} \\ &= 0.75. \end{aligned}$$

   (d) If a tax return is not flagged by the computer, what is the probability that it is correct?
   We are looking for $P(E^c|F^c)$. Note that because $P(F|E^c) = 0.05$, we know that $P(F^c|E^c) = 0.95$. So then by the definition of conditional probability, we have that

   $$\begin{aligned} P(E^c|F^c) &= \frac{P(F^c|E^c)P(E^c)}{P(F^c)} \\ &= \frac{0.95(0.85)}{1 - 0.17} \\ &\approx 0.9729. \end{aligned}$$

2. For the following scenarios, identify the most likely probability model for the random variable that is described:

   (a) Suppose that for single launches of a space shuttle, there is a constant probability of O-ring failure (say 0.15). Out of ten launches, let X be the number of those involving an O-ring failure.

   The best probability model for this would be a Binomial probability model.

   $$X \sim \text{Binom}(10, 0.15).$$

   (b) In a particularly temperate location, the daily high temperature is equally likely to be any value over the range from 70°F to 100°F. Let X be the daily high temperature.

   The best probability model for this would be a Uniform probability model.

   $$X \sim \text{Unif}(70, 100).$$

   (c) Transmission line interruptions in a telecommunications network occur at an average rate of one per day. Let X be the number of interruptions in the next five-day work week.

   The best probability model for this would be a Poisson probability model.

   $$X \sim \text{Poisson}(5).$$

   (d) In a clinical study, volunteers are tested for a gene that has been found to increase the risk for a disease. The probability that a person carries the gene is 0.1. Let X be the number of people examined until a person who carries the gene is found.

   The best probability model for this would be a Geometric probability model.

   $$X \sim \text{Geo}(0.1).$$

   (e) Small aircraft arrive at a certain airport with a rate of 1 aircraft every 10 minutes. Let X be the length of time until the 15th aircraft arrives each day.

   The best probability model for this would be an Gamma probability model.

   $$X \sim \sum_{i=1}^{15} \text{Exp}\,(10) = \text{Gamma}(15, 10).$$

   (f) The distribution of resistance for resistors having a nominal value of 10 ohms is under investigation. An electrical engineer randomly selects 73 resistors and measures their resistance. Based on these 73 values, she determines that the resistance of the resistors has the following behavior: approximately 70% of the resistors have resistance within one standard deviation of 10 ohms, 95% are within two standard deviations, and none of the resistors have resistance greater than three standard deviations from 10 ohms.

   The best probability model for this would be a Normal probability model, given that the random variable is the resistance value of ohms for a particular resistor.

   $$X \sim \mathcal{N}(10, \sigma^2).$$

3. A typist makes a typographical error at the rate of 1 every 20 pages. Let X be the number of errors in a manuscript of 200 pages.

   (a) Write the binomial distribution model for X, and show how you will calculate the exact binomial probability that the manuscript has at least 5 errors.

   We know that the probability mass function for a binomial distribution is

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{for } x = 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

   So then knowing that $n = 200$ and $p = 0.05$, we can substitute and get

$$f(x) = \begin{cases} \binom{200}{x} (0.05)^x (0.95)^{200-x}, & \text{for } x = 1, 2, ..., n \\ 0, & \text{otherwise.} \end{cases}$$

   Since we want to find $P(X \geq 5)$, we will need to use the cumulative density function. Additionally, we know that $P(X \geq 5) = 1 - P(X < 5)$. We observe there are 10 trials of twenty pages. So using the CDF in R, we have:

```
pbinom(5, p = 0.05, size = 200, lower.tail = FALSE)

## [1] 0.9376575
```

   (b) Why is it reasonable to assume that X has a Poisson distribution? Calculate the Poisson approximation to the probability in (a).

   It is reasonable to assume that X has a Poisson distribution because the probability of success is on the smaller side, and the number of trials is on the larger side. We have $\lambda = -0.05$, as the rate of success per each twenty page interval is 1. The probability of the typist making five or more errors in a 200 page document can be calculated as:

```
ppois(5, lambda = (200)*0.05, lower.tail = FALSE)

## [1] 0.932914
```

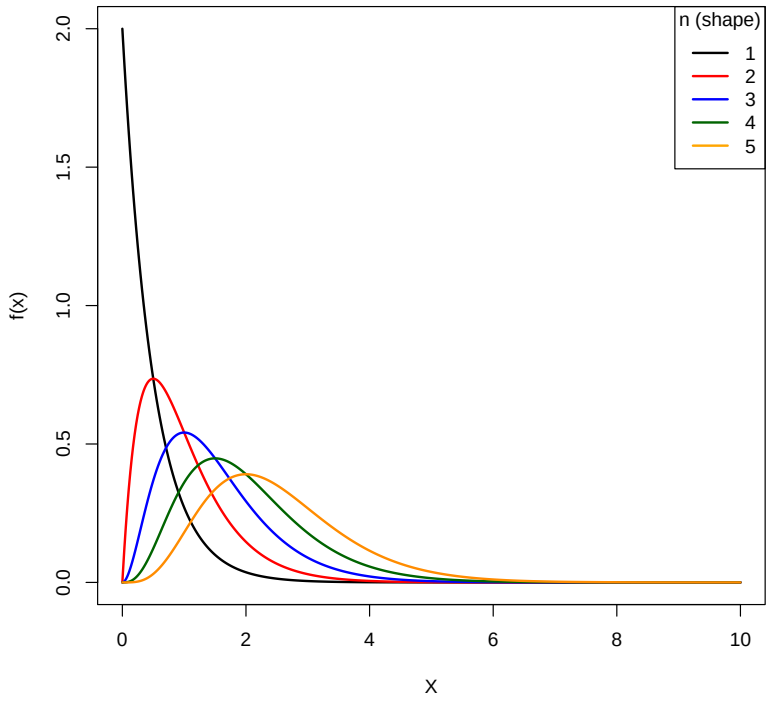
4. Plot the gamma distribution's pdf using the values below for n and α. You may plot all of the curves in each part on the same set of axes as long as they are each clearly marked.

   (a) $n = 1, 2, 3, 4, 5$ with $\alpha = 2$ for each.

```r
x <- seq(0, 10, len = 2000)
pdf1 <- dgamma(x, shape = 1, rate = 2)
pdf2 <- dgamma(x, shape = 2, rate = 2)
pdf3 <- dgamma(x, shape = 3, rate = 2)
pdf4 <- dgamma(x, shape = 4, rate = 2)
pdf5 <- dgamma(x, shape = 5, rate = 2)

df <- data.frame(cbind(x, pdf1, pdf2, pdf3, pdf4, pdf5))

plot(df$x, df$pdf1, type = "l", lwd = 2,
col = "black", xlab = "X", ylab = "f(x)")
lines(df$x, df$pdf2, col = "red", type = "l", lwd = 2)
lines(df$x, df$pdf3, col = "blue", type = "l", lwd = 2)
lines(df$x, df$pdf4, col = "darkgreen", type = "l", lwd = 2)
lines(df$x, df$pdf5, col = "darkorange", type = "l", lwd = 2)
legend("topright",
       legend = c("1", "2", "3", "4", "5"),
       lwd = 2,
       col = c("black", "red", "blue", "darkgreen", "orange"),
       title = expression(paste("n (shape)")))
```

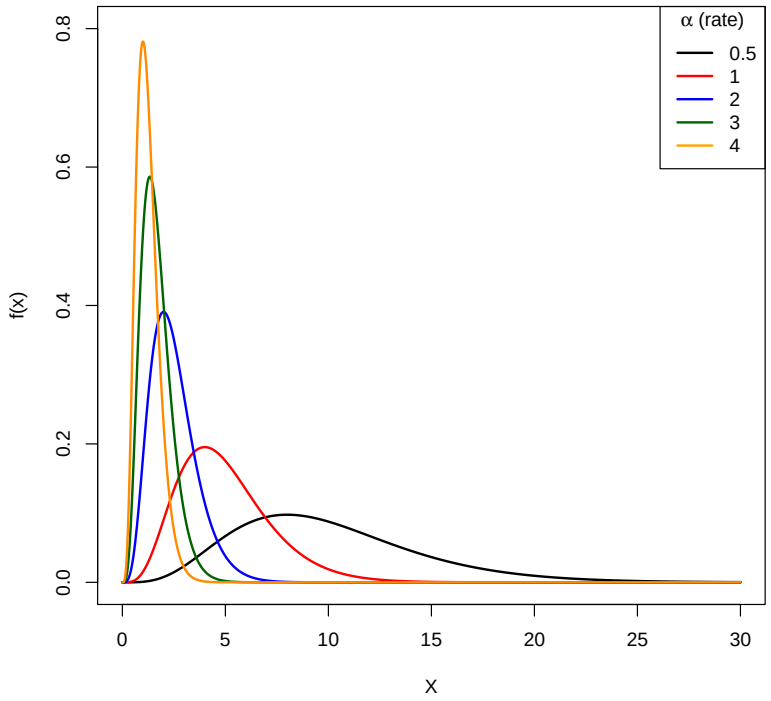


(b) $\alpha = 0.5, 1, 2, 3, 4$ with $n = 5$ for each.

```r
x <- seq(0, 30, len = 2000)
pdf1 <- dgamma(x, shape = 5, rate = 1/2)
pdf2 <- dgamma(x, shape = 5, rate = 1)
pdf3 <- dgamma(x, shape = 5, rate = 2)
pdf4 <- dgamma(x, shape = 5, rate = 3)
pdf5 <- dgamma(x, shape = 5, rate = 4)

df <- data.frame(cbind(x, pdf1, pdf2, pdf3, pdf4, pdf5))

plot(df$x, df$pdf1, type = "l", lwd = 2,
col = "black", xlab = "X", ylab = "f(x)",
xlim = c(0,30), ylim = c(0,0.8))
lines(df$x, df$pdf2, col = "red", type = "l", lwd = 2)
lines(df$x, df$pdf3, col = "blue", type = "l", lwd = 2)
lines(df$x, df$pdf4, col = "darkgreen", type = "l", lwd = 2)
lines(df$x, df$pdf5, col = "darkorange", type = "l", lwd = 2)
legend("topright",
       legend = c("0.5", "1", "2", "3", "4"),
       lwd = 2,
       col = c("black", "red", "blue", "darkgreen", "orange"),
       title = expression(paste(alpha, " (rate)")))
```



5. The pdf of the length of a hinge for fastening a door is $f(x) = 1.25$ for $74.6 < x < 75.4$. Determine the following:

   (a) $P(X < 74.8 \text{ or } X > 75.2)$

   For the following, we know that for any number that is not in the range $[74.6, 75.4]$, $f(x) = 0$. So then,

$$\begin{aligned} P(X) &= \int_{-\infty}^{74.8} 1.25\,dx + \int_{75.2}^{\infty} 1.25\,dx \\ &= \int_{74.6}^{74.8} 1.25\,dx + \int_{75.2}^{75.4} 1.25\,dx \\ &= 1.25x\Big|_{74.6}^{74.8} + 1.25x\Big|_{75.2}^{75.4} \\ &= 2\,(1.25)\,(0.2) \\ &= 0.5. \end{aligned}$$

   (b) $E(X)$

   We know that $X \sim U(74.6, 75.4)$, which implies the expected value of X is

$$\mu = \frac{(74.6 + 75.4)}{2} = 75.$$

   (c) $\text{Var}(X)$

   Likewise, we have that $Var(X)$ can be expressed as,

$$\sigma^2 = \frac{(74.6 - 75.4)^2}{12} = 0.05\bar{3}.$$

6. The pdf of the time to failure of an electronic component in a copier (in hours) is $f(x) = (1/1000)e^{-(1/1000)x}$ for $x > 0$. Determine the following:

   a) the probability that the component lasts more than 3000 hours before failure.

   We are interested in finding $P(X > 3000)$. We can find this in `R` (see code output for answer).

```
pexp(q = 3000, rate = 1/1000, lower.tail = FALSE)

## [1] 0.04978707
```

   b) the probability that the component fails between 1000 and 2000 hours.

   (See code output for answer).

```
pexp(q = 2000, rate = 1/1000, lower.tail = TRUE) -
  pexp(q = 1000, rate = 1/1000, lower.tail = TRUE)

## [1] 0.2325442
```

   c) the number of hours at which 10% of all components have failed.

   (See code output for answer).

```
qexp(0.1, rate = 1/1000, lower.tail = TRUE)

## [1] 105.3605
```

   d) the mean number of hours the component can be expected to last until failure.

   For any $X \sim \text{Exp}(\beta)$, we know that the mean value is $\frac{1}{\beta}$. In this case, $\beta = \frac{1}{1000}$, which means that $\mu = 1000$.

7. The speed of a file transfer from a server on campus to a personal computer at a student's home on a weekday evening is normally distributed with a mean of 60 kilobits/second and a standard deviation of 4 kilobits/second.

   (a) What is the probability that the file will transfer at a speed of 70 kilobits per second or more?
   (See code output for answer).

```
pnorm(70, mean = 60, sd = 4, lower.tail = FALSE)

## [1] 0.006209665
```

   (b) Given that the speed of the transfer is greater than 60 kilobits/second, what is the probability that the transfer is greater than 70 kilobits/second?
   Let S be the speed of a file transfer from a server on campus to a personal computer at a student's home on a weekday evening. We are interested in finding $P(S > 70|S > 60)$. So then we have that

$$\begin{aligned} P(S > 70|S > 60) &= \frac{P(S > 70 \cap S > 60)}{P(S > 60)} \\ &= \frac{P(S > 70)}{P(S > 60)}. \end{aligned}$$

   We can calculate these probabilities using R (see code output for answer).

```
pnorm(70, mean = 60, sd = 4, lower.tail = FALSE)/
    pnorm(60, mean = 60, sd = 4, lower.tail = FALSE)

## [1] 0.01241933
```

   (c) What file transfer time is faster than 80% of all other transfer times?
   (See code output for answer).

```
qnorm(.8, mean = 60, sd = 4)

## [1] 63.36648
```

8. A restaurant serves three fixed-price dinners costing \$12 \$15, and \$20. For a randomly selected couple (of one male and one female) dining at this restaurant, let X = the cost of the man's dinner, and Y = the cost of the woman's dinner. The joint pmf of X and Y is given in the following table:

		Y		
$f(x,y)$		12	15	20
	12	0.05	0.05	0.10
X	15	0.05	0.10	0.35
	20	0.00	0.20	0.10

(a) Compute the marginal pmf's of X and Y.

For any joint probability mass function, the marginal probabilities can be found by computing

$$f(X) = \sum_{\text{all } Y} f(X, Y).$$

So we have,

X	12	15	20	Y	12	15	20
$f(X)$	0.20	0.50	0.30	$f(Y)$	0.10	0.35	0.55

(b) What is the probability that the man's and the woman's dinner cost at most \$15 each?

Finding the answer to this question is the same as finding

$$\begin{aligned} P(X \leq 15, Y \leq 15) &= P(X = 12, Y = 12) + P(X = 15, Y = 12) + \\ &\quad P(X = 12, Y = 15) + P(X = 15, Y = 15) \\ &= 0.25. \end{aligned}$$

(c) Are X and Y independent? Justify your answer.

In this discrete case, X and Y are independent if and only if $P(X \cap Y) = P(X, Y) = P(X)P(Y)$. So then we have that

$$P(X = 12, Y = 12) \neq P(X = 12)P(Y = 12).$$

$\therefore$ X and Y are not independent.

(d) What is the expected total cost of the dinner for two people?

$$\begin{aligned} E[X + Y] &= E[X] + E[Y] \\ &= \sum_{i=1}^{3} x_i P(X = x_i) + \sum_{i=1}^{3} y_i P(Y = y_i) \\ &= \$15.9 + \$17.45 \\ &= \$33.35 \end{aligned}$$

9. A health-food store stocks two different brands of a certain type of grain. Let X = the amount (in lbs) of brand A on hand and Y = the amount of brand B on hand. Suppose the joint pdf of X and Y is

$$f(x,y) = \begin{cases} kxy, & x \geq 0, y \geq 0, 20 \leq x+y \leq 30 \\ 0, & \text{otherwise} \end{cases}$$

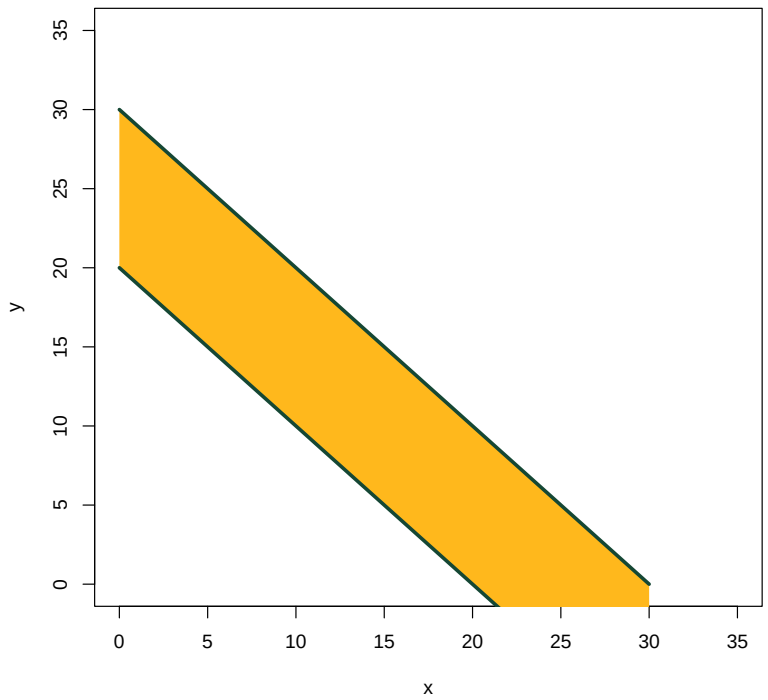
(a) Draw the support of the distribution, and set up the integral to find the value of k that makes this a valid pdf. (But do not work it out.)

```
x <- seq(0, 30, len = 2000)
y1 <- round(-x + 30, digits = 3)
y2 <- round(-x + 20, digits = 3)

# Lines
plot(x, y1, type = "l",
     xlim = c(0, 35), ylim = c(0, 35),
     ylab = "y")
lines(x, y2, type = "l")

# Fill area between lines
# Used Baylor Gold and Green for Hex Code Colors
polygon(c(x, rev(x)), c(y2, rev(y1)),
        col = "#FFB81C", lty = 0)

# Redraw the lines
lines(x, y1, col = "#154734", lwd = 3)
lines(x, y2, col = "#154734", lwd = 3)
```



In addition to the support, to find the k that makes $f(x,y)$ a valid pdf, we must find the k that satisfies

$$\begin{aligned}
1 &= \int_0^{20}\int_{20-x}^{30-x} kxy\,dydx + \int_{20}^{30}\int_0^{30-x} kxy\,dydx \\
&= \frac{k}{2}\int_0^{20} x\left(y^2\Big|_{20-x}^{30-x}\right)\,dx + \frac{k}{2}\int_{20}^{30} x\left(y^2\Big|_0^{30-x}\right)\,dx \\
&= \frac{k}{2}\left(\int_0^{20} -20x^2+500x\,dx + \int_{20}^{30} x^3-60x^2+900x\,dx\right) \\
&= k\left(\left(-\frac{10}{3}x^3+125x^2\Big|_0^{20}\right)+\left(\frac{1}{8}x^4-10x^3+\frac{450}{2}x^2\Big|_{20}^{30}\right)\right) \\
&= k(\frac{70000}{3}+3750) \\
&= k(27083.\bar{3}).
\end{aligned}$$

This means that $k = \frac{1}{27083.\bar{3}}$ makes $f(x,y)$ a valid pdf.

(b) Set up the integral to find the marginal pdf of X. (But do not work it out.)
The marginal pdf of X would be written as follows:

$$\begin{aligned}
f_x(x) &= \int_{-\infty}^{\infty} \frac{1}{27083.\bar{3}} xy\,dy \\
&= \begin{cases} \displaystyle\int_{20-x}^{30-x} \frac{1}{27083.\bar{3}} xy\,dy, & 0 \le x \le 20 \\ \displaystyle\int_{0}^{30-x} \frac{1}{27083.\bar{3}} xy\,dy, & 20 \le x \le 30. \end{cases}
\end{aligned}$$

(c) Set up the integral with appropriate limits of integration to compute $P(X+Y \le 25)$. (But do not work it out).

$$P(X+Y\le 25) = \int_0^{20}\int_{20-x}^{25-x} \frac{1}{27083.\bar{3}} xy\,dydx + \int_{20}^{25}\int_0^{25-x} \frac{1}{27083.\bar{3}} xy\,dydx.$$

10. In a shaft and bearing assembly, the diameters of the bearings, X, are normally distributed with $\mu_x = 0.526$ inches and $\sigma_x = 3.0 \times 10^{-4}$ inches. The diameters of the shafts, Y, are also normally distributed with $\mu_y = 0.525$ inches and $\sigma_y = 4.0 \times 10^{-4}$ inches.

    (a) What is the distribution of the clearance, $X - Y$? What is its mean and standard deviation?

    Let $U = X - Y$, a linear combination of random variables X and Y. Because a linear combination of Normal random variables is still Normal, we know that U is Normal. Also, that

    $$\begin{aligned} \mathrm{E}[U] &= \mathrm{E}[X] - \mathrm{E}[Y] \\ &= 0.526 - 0.525 \\ &= 0.001. \end{aligned}$$

    Additionally because X and Y are normally distributed and $\mathrm{Cov}(X, Y) = 0$ due to their independence,

    $$\begin{aligned} \mathrm{Var}[U] &= \mathrm{Var}[X] + \mathrm{Var}[Y] - 2\,\mathrm{Cov}(X, Y) \\ &= (3.0 \times 10^{-4})^2 + (4.0 \times 10^{-4})^2 \\ &= 2.5 \times 10^{-7} \\ \implies \sigma_U &= 5 \times 10^{-4}. \end{aligned}$$

    So $U \sim \mathcal{N}(0.01, 5 \times 10^{-4})$.

    (b) What is the probability that the shaft will fit inside a bearing?

    We are looking for $P(U \geq 0)$. We can calculate this using `R`.

```
exp(pnorm(q = 0,
          mean = 0.001,
          sd = 5e-04,
          lower.tail = FALSE,
          log.p = TRUE))
```

```
## [1] 0.9772499
```

    (c) What is the probability that out of ten shaft-bearing pairs, at least nine will fit properly?

    Let $W \sim \mathrm{Binom}(10, P(U \geq 0))$, where W is the number of shaft-bearing pairs that fit properly. We can find the probability that at least nine fit using the following `R` code.

```
pbinom(q = 9,
       size = 10,
       prob = exp(pnorm(q = 0, mean = 0.001, sd = 5e-04,
                        lower.tail = FALSE, log.p = TRUE)),
       lower.tail = FALSE)
```

```
## [1] 0.794431
```

11. The coefficient of linear expansion of brass is to be determined as a laboratory exercise. For a brass bar that is L_1 meters long at T_1 °C and L_2 meters long at T_2 °C, this coefficient is

$$\alpha = \frac{L_2 - L_1}{L_1(T_2 - T_1)}.$$

Suppose that the equipment to be used in the laboratory is thought to have a standard deviation for repeated length measurements of about 0.00005 m and a standard deviation for repeated temperature measurements of about 0.1°C. Repeated measures are also normally distributed.

(a) The values $T_1 \approx 50$°C, $T_2 \approx 100$°C, $L_1 \approx 1.00000$ m, and $L_2 \approx 1.00095$ m are obtained. Approximate the standard deviation of α by simulating its distribution, and plot a histogram of the distribution. Show your code. **We found that the standard deviation for alpha $\sigma_\alpha \approx 1.414 \times 10^{-6}$.**

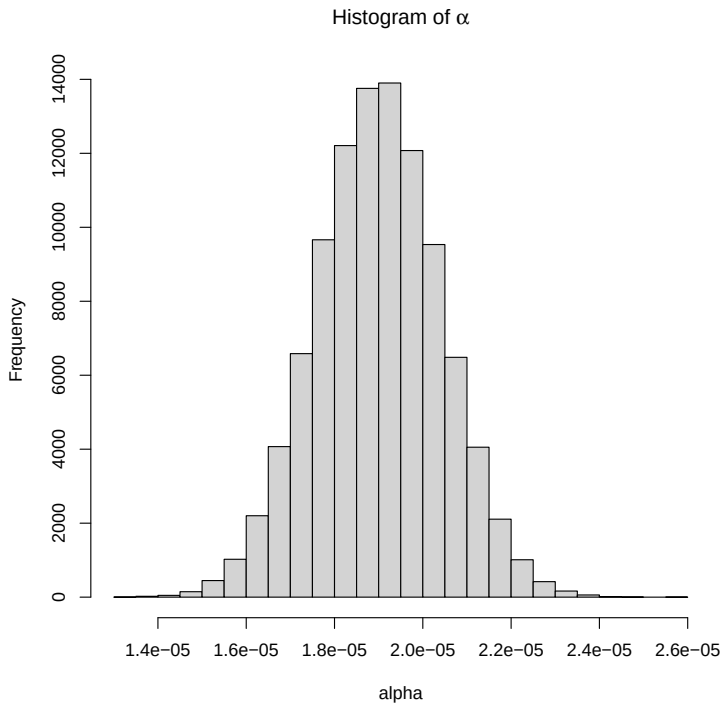
```
set.seed(613)
L1 <- rnorm(1e5, mean = 1.00000, sd = 5e-5)

alpha <- (rnorm(1e5, mean = 1.00095, sd = 5e-5) -
 L1) / ((L1)*(rnorm(1e5, mean = 100, sd = 0.1) -
      rnorm(1e5, mean = 50, sd = 0.1)))

sd(alpha)

## [1] 1.414597e-06

hist(alpha,
     main = expression(paste("Histogram of ", alpha)))
```



(b) Approximate the standard deviation of α by using the propagation of error formulas. Show your work.

We assume all random variables to be independent. In this case, the propagation error formulas for

$$\sigma_\alpha \approx \sqrt{\left(\frac{\partial \alpha}{\partial L_1}\right)^2 \sigma_{L_1}^2 + \left(\frac{\partial \alpha}{\partial L_2}\right)^2 \sigma_{L_2}^2 + \left(\frac{\partial \alpha}{\partial T_1}\right)^2 \sigma_{T_1}^2 + \left(\frac{\partial \alpha}{\partial T_2}\right)^2 \sigma_{T_2}^2}.$$

We can find the partial derivatives as such,

$$\begin{aligned}
\frac{\partial \alpha}{\partial L_1} &= -\frac{L_2}{L_1^2(T_2 - T_1)} = -\frac{1.00095}{1^2(100-50)} = -0.020019 \\
\frac{\partial \alpha}{\partial L_2} &= \frac{1}{L_1(T_2 - T_1)} = 0.02 \\
\frac{\partial \alpha}{\partial T_1} &= \frac{L_2 - L_1}{L_1(T_2 - T_1)^2} = \frac{1.00095 - 1}{1^2(100-50)^2} = 3.8 \times 10^{-7} \\
\frac{\partial \alpha}{\partial T_1} &= -\frac{L_2 - L_1}{L_1(T_2 - T_1)^2} = -\frac{1.00095 - 1}{1^2(100-50)^2} = -3.8 \times 10^{-7},
\end{aligned}$$

and use propagation of error formulas to estimate σ_α as,

$$\begin{aligned}
\sigma_\alpha \approx ((-0.020019)^2(5 \times 10^{-5})^2 + (0.02)^2(5 \times 10^{-5})^2 + \\
(3.8 \times 10^{-7})^2(0.1)^2 + (-3.8 \times 10^{-7})^2(0.1)^2)^{\frac{1}{2}}
\end{aligned}$$

giving $\sigma_\alpha \approx 1.416 \times 10^{-6}$.

(c) What percentage of the variability in α comes from each of the four variables, L_1, L_2, T_1, and T_2?

We know that the percentage of the variability in α is the proportion of each term's variance over σ_α^2. So we have

$$\begin{aligned}
\left(\frac{\left(\frac{\partial \alpha}{\partial L_1}\right)^2 \sigma_{L_1}^2}{\sigma_\alpha^2}\right) &\approx 0.49969, \\
\left(\frac{\left(\frac{\partial \alpha}{\partial L_2}\right)^2 \sigma_{L_2}^2}{\sigma_\alpha^2}\right) &\approx 0.49874, \\
\left(\frac{\left(\frac{\partial \alpha}{\partial T_1}\right)^2 \sigma_{T_1}^2}{\sigma_\alpha^2}\right) &\approx 7.202 \times 10^{-4}, \\
\left(\frac{\left(\frac{\partial \alpha}{\partial T_2}\right)^2 \sigma_{T_2}^2}{\sigma_\alpha^2}\right) &\approx 7.202 \times 10^{-4}.
\end{aligned}$$

12. Derive the mean and variance of the geometric random variable, $X = \#$ of trials until first success, using the definitions of mean and variance for a discrete random variable.

For any discrete random variable, we know $E[X] = \sum_{\text{all } x} x f_X(x)$. So we have for a geometric random variable,

$$\begin{aligned}
E[X] &= \sum_{k=1}^{\infty} k\,(1-p)^{k-1}\,p \\
&= p\sum_{k=1}^{\infty} k\,(1-p)^{k-1} \\
&= -p\sum_{k=1}^{\infty} \frac{d}{dp}(1-p)^{k} \\
&= -p\frac{d}{dp}\frac{(1-p)}{(1-(1-p))} \\
&= -p\frac{d}{dp}\frac{(1-p)}{p} \\
&= -p\left(-\frac{1}{p^2}\right) \\
&= \frac{1}{p}.
\end{aligned}$$

For any random variable, we know $\text{Var}[X] = E[X^2] - E[X]^2$. We already know that $E[X]^2 = \frac{1}{p^2}$. So for a geometric random variable, we can find $E[X^2]$ by,

$$\begin{aligned}
E[X^2] &= E[X(X-1)] + E[X] \\
&= p\left(\sum_{k=1}^{\infty}(k-1)k\,(1-p)^{k-1}\right) + \frac{1}{p} \\
&= -p\frac{d}{dp}\left(\sum_{k=1}^{\infty}(k-1)\,(1-p)^{k}\right) + \frac{1}{p} \\
&= -p\frac{d}{dp}\left[(1-p)^2\left(\sum_{k=2}^{\infty}(k-1)\,(1-p)^{k-2}\right)\right] + \frac{1}{p} \\
&= p\frac{d}{dp}\left[(1-p)^2\frac{d}{dp}\left(\sum_{k=2}^{\infty}(1-p)^{k-1}\right)\right] + \frac{1}{p} \\
&= p\frac{d}{dp}\left[(1-p)^2\frac{d}{dp}\left(\sum_{k=1}^{\infty}(1-p)^{k}\right)\right] + \frac{1}{p} \\
&= p\frac{d}{dp}\left[(1-p)^2\frac{d}{dp}\left(\frac{1}{p}-1\right)\right] + \frac{1}{p} \\
&= p\frac{d}{dp}\left(\frac{(1-p)^2}{p^2}\right) + \frac{1}{p} \\
&= p\left(\frac{-2(1-p)}{((1-p)-1)^3}\right) \\
&= \frac{2(p-1)}{-p^2} + \frac{1}{p} \\
&= \frac{2(1-p)}{p^2} + \frac{1}{p}.
\end{aligned}$$

So then we have

$$\begin{aligned}
\text{Var}[X] &= E[X^2] - E[X]^2 \\
&= \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} \\
&= \frac{2-2p+p-1}{p^2} \\
&= \frac{1-p}{p^2}.
\end{aligned}$$

13. A random variable X has the following pdf:

$$f(x) = \begin{cases} 2x^{-3}, & \text{if } x \geq 1 \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the cdf of X.

The cdf of X would be written as,

$$\begin{aligned} F_X(x) &= \int_1^x 2t^{-3}\, dt \\ &= -\frac{1}{t^2}\Big|_1^x \\ &= 1 - \frac{1}{x^2} \end{aligned}$$

(b) Give a formula for the pth quantile of X. Use it to find the median (i.e., the $50th$ quantile) of X.

Let $0 \leq Q \leq 1$ be the quantile of X (e.g. if we were interested in the $50th$ quantile, $Q = 0.5$). The formula for the quantile would be,

$$Q = \left(1 - \frac{1}{x^2}\right).$$

Solving for the median, we would get

$$\begin{aligned} 0.5 &= \left(1 - \frac{1}{x^2}\right) \\ -0.5 &= -\frac{1}{x^2} \\ x &= \sqrt{\frac{1}{0.5}} = \sqrt{2}. \end{aligned}$$

So the median value of this random variable X would be $\sqrt{2}$, as $-\sqrt{2}$ is not in the support of the pdf.

14. Find the cdf of the exponential random variable with $\alpha = 5$. Show your work.

The pdf of a random variable $X \sim \text{Exp}(\alpha)$ is

$$f(x) = \begin{cases} \alpha e^{-\alpha x}, & \text{for } x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

To find the cdf, we need to evaluate $\int_{x \in \mathcal{X}} f(x)\, dx$ on the subset of its support, $\mathcal{X} = (0, x)$. When $\alpha = 5$, we have

$$\begin{aligned} \int_{-\infty}^{x} f(t)\, dt &= \int_{0}^{x} 5e^{5t}\, dt \\ &= 5 \int_{0}^{-5x} \frac{-1}{5} e^{u}\, du \\ &= -\int_{0}^{-5x} e^{u}\, du \\ &= -\left(e^{u} \Big|_{0}^{-5x} \right) \\ F_X(x) &= -e^{-5x} + 1. \end{aligned}$$

15. Consider a random variable X with has pdf $f_1(x)$ with probability p_1 and pdf $f_2(x)$ with probability p_2, where $p_1 + p_2 = 1$. We can think of first observing a Bernoulli r.v. Y, which equals 1 with probability p_1 and 0 with probability $1 - p_1$. If $Y = 1$, then $X = X_1 \sim f_1(x)$, and if $Y = 0$, then $X = X_2 \sim f_2(x)$. We say that X has a *mixture* distribution. This is used to model samples drawn from a heterogeneous population formed by a mixture of two different populations and to model data contaminated by outliers.

(a) Show that the pdf of X is

$$f(x) = p_1 f_1(x) + p_2 f_2(x).$$

We can think of $P(X|Y_i)$ as an exhaustive partition of the sample space $\mathcal{X}$, because $p_2 = 1 - p_1 \iff p_1 + p_2 = 1 \iff p_1 + (1 - p_1) = 1$. We let Y_1 be the event where $Y = 1$, and Y_2 be the event where $Y = 0$. By the Law of Total Probability, we have it that,

$$P(X) = \sum_{i=1}^{2} P(X_i|Y_i)P(Y_i).$$

We know that $P(X|Y_1) = F_1(x)$ where $Y = 1$, and $P(X|Y_0) = F_2(x)$ where $Y = 0$. We can So then we have it that

$$P(X) = F_1(x)p_1 + F_2(x)p_2.$$

This means the pdf could be written as:

$$f(X) = f_1(x)p_1 + f_2(x)p_2.$$

(b) Let μ_1 and μ_2 be the means of $f_1(x)$ and $f_2(x)$, respectively. Show that

$$E(X) = \mu = p_1\mu_1 + p_2\mu_2.$$

We have that X is a mixture of X_1 and X_2. So then

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x \sum_{i=1}^{2} p_i f_i(x)\, dx \\ &= \sum_{i=1}^{2} p_i \int_{-\infty}^{\infty} x f_i(x)\, dx \\ &= \sum_{i=1}^{2} p_i E[X_i] \\ &= p_1\mu_1 + p_2\mu_2 \\ &= \mu. \end{aligned}$$

(c) Let σ_1^2 and σ_2^2 be the variances of $f_1(x)$ and $f_2(x)$, respectively. Show that

$$Var(X) = \sigma^2 = p_1\sigma_1^2 + p_2\sigma_2^2 + p_1\mu_1^2 + p_2\mu_2^2 - (p_1\mu_1 + p_2\mu_2)^2.$$

We know that the variance can be written as,

$$\begin{aligned}
\sigma^2 &= E[X^2] - \mu^2 \\
\implies &= \left(\sum_{i=1}^{2} p_i(E[X_i^2])\right) - \mu^2 \\
&= \sum_{i=1}^{2} p_i(\sigma_i^2 + \mu_i^2) - \mu^2 \quad (\text{from } \sigma_i^2 = E[X_i^2] - \mu_i^2, \text{ therefore } E[X_i^2] = \sigma_i^2 + \mu_i^2) \\
&= p_1\sigma_1^2 + p_2\sigma_2^2 + p_1\mu_1^2 + p_2\mu_2^2 - \mu^2 \\
&= p_1\sigma_1^2 + p_2\sigma_2^2 + p_1\mu_1^2 + p_2\mu_2^2 - (p_1\mu_1 + p_2\mu_2)^2.
\end{aligned}$$