MSSC 6010 / Comp. Probability

Instructor: Mehdi Maadooliat

Chapter 4

Department of Mathematics, Statistics and Computer Science

Special thanks to Prof. Ana Militino for providing the original slides of the book.

MARQUETTE UNIVERSITY
Be The Difference.

Chapter 4

Univariate Probability Distributions

4.2 Discrete Univariate Distributions

4.2.1 Discrete Uniform Distribution

The random variable X is said to follow a discrete uniform distribution with parameter n (where $n \in \mathbb{N}$) if the probability X takes on the value x is the same for all x, where $x = x_1, x_2, \ldots, x_n$.

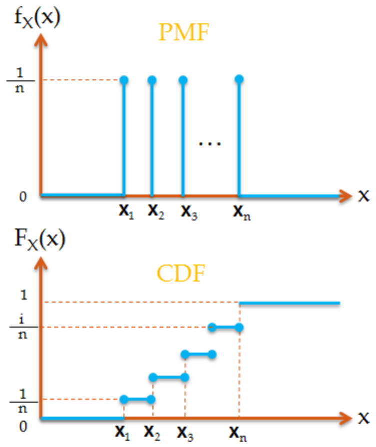
Discrete Uniform Distribution

$$\mathbb{P}(X = x_i|n) = \frac{1}{n}, \quad i = 1, 2, \ldots, n.$$

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\text{var}[X] = \frac{1}{n}\sum_{i=1}^{n}(x_i - E[X])^2$$

$$M_X(t) = \frac{1}{n}\sum_{i=1}^{n} e^{tx_i}$$



When $x_i = i$ for $i = 1, \ldots, n$, it can be shown that the $E[X] = \frac{n+1}{2}$ and that the $\text{var}[X] = \frac{n^2-1}{12}$ respectively.

Example 4.1 One light bulb is randomly selected from a box that contains a 40 watt light bulb, a 60 watt light bulb, a 75 watt light bulb, a 100 watt light bulb, and a 120 watt light bulb. Write the probability function for the random variable that represents the wattage of the randomly selected light bulb, and determine the mean and variance of that random variable.

Solution: The random variable X can assume the set of values $\Omega = \{40, 60, 75, 100, 120\}$. The probability density function for the random variable X is

$$\mathbb{P}(X = x|5) = \frac{1}{5} \quad \text{for} \quad x = 40, 60, 75, 100, 120.$$

The expected value of X, $E[X] = 79$, and the variance of X, $\text{var}[X] = 804$. S can be used to alleviate the arithmetic.

```
➢ Watts <- c(40,60,75,100,120)
➢ meanWatts <- (1/5)*sum(Watts)
➢ varWatts<- (1/5)*sum((Watts-meanWatts)^2)
➢ ans <- c(meanWatts, varWatts)
➢ ans
[1] 79 804
```

4.2.2 Bernoulli and Binomial Distributions

When the Tossing a coin a single time is an example of a **Bernoulli** trial. A Bernoulli trial is a random experiment with only two possible outcomes. The outcomes are mutually exclusive and exhaustive.

For example, success or failure, true or false, alive or dead, male or female, etc. A Bernoulli random variable X, can take on two values, where $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. The probability that X is a success is π, and the probability that X is a failure is $\varrho = 1 - \pi$. Box (4.2) gives the **pdf**, mean, and variance of a Bernoulli random variable.

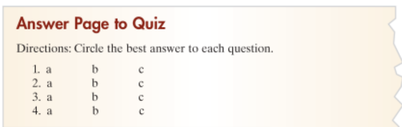
Bernoulli Distribution $\quad X \sim Bernoulli(\pi)$

$$\begin{aligned} \mathbb{P}(X = x|\pi) &= \pi^x (1-\pi)^{1-x}, \quad x = 0, 1 \\ E[X] &= \pi \\ \text{var}[X] &= \pi(1-\pi) \\ M_X(t) &= \pi e^t + \varrho \end{aligned} \tag{4.2}$$

4

THE BINOMIAL PROBABILITY DISTRIBUTION

- **Consider the following probability experiment. I give you a surprise four-question multiple-choice quiz.**
- **You have not studied the material, and therefore you decide to answer the four questions by randomly guessing.**
- **Here are some questions for you?**


Answer Page to Quiz

Directions: Circle the best answer to each question.

1. a b c
2. a b c
3. a b c
4. a b c

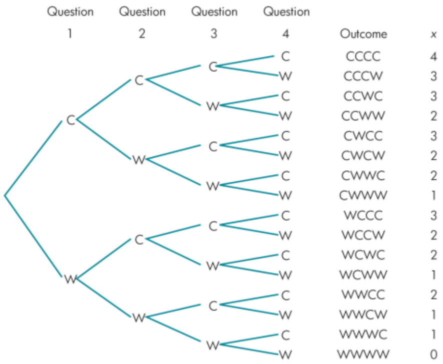
1. **How many of the four questions are you likely to have answered correctly?**
2. **How likely are you to have more than half of the answers correct?**
3. **What is the probability that you selected the correct answers to all four questions?**
4. **What is the probability that you selected wrong answers for all four questions?**
5. **If an entire class answers the quiz by guessing, what do you think the class “average” number of correct answers will be?**

5

THE BINOMIAL PROBABILITY DISTRIBUTION

MARQUETTE UNIVERSITY
Be The Difference.

- To find the answers to these questions, let's start with a tree diagram

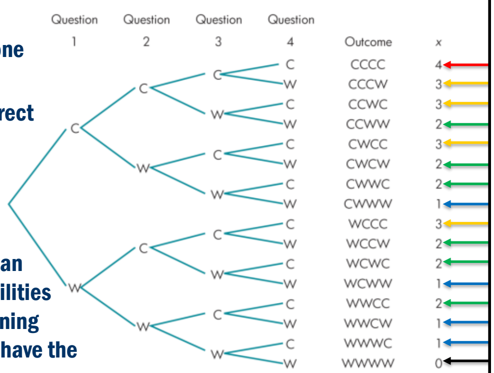


- Each of the four questions is answered with the correct answer (C) or with a wrong answer (W).
- x is the "number of correct answers" on one person's quiz when the quiz was taken by randomly guessing.

THE BINOMIAL PROBABILITY DISTRIBUTION

MARQUETTE UNIVERSITY
Be The Difference.

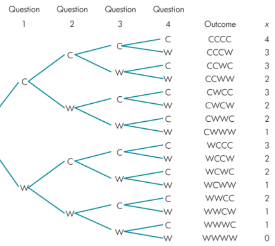
- Notice that:
  - The event $x = 4$, "four correct answers," is shown on the top branch.
  - The event $x = 0$, "zero correct answers," is shown on the bottom branch.
  - The event $x = 1$ occurs on four different branches.
  - The event $x = 2$ occurs on six branches.
  - The event $x = 3$ occurs on four branches.
  - Each individual question has only one correct answer.
  - The probability of selecting the correct answer to each question is $\frac{1}{3}$.
  - The probability that a wrong answer is selected is $\frac{2}{3}$.
  - The probability of each value of x can be found by calculating the probabilities of all the branches and then combining the probabilities for branches that have the same x values.



THE BINOMIAL PROBABILITY DISTRIBUTION

MARQUETTE UNIVERSITY Be The Difference.

- $P(x=0)$ **is the probability that the correct answers are given for zero questions.**
  - $P(x=0)=\frac{2}{3}\times\frac{2}{3}\times\frac{2}{3}\times\frac{2}{3}=\left(\frac{2}{3}\right)^4=\frac{16}{81}=\mathbf{0.198}$
  - **Note: Answering each individual question is a separate and independent event, thereby we can use:**
  - $P\ (A \text{ and } B)=P(A)\cdot P(B)$
- $P(x=4)$ **is the probability that correct answers are given for all four questions.**
  - $P(x=4)=\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}=\left(\frac{1}{3}\right)^4=\frac{1}{81}=\mathbf{0.012}$

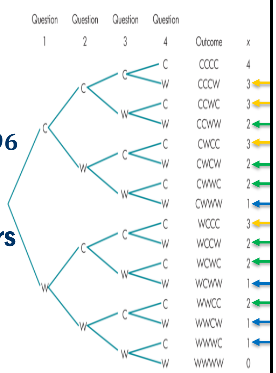


8

THE BINOMIAL PROBABILITY DISTRIBUTION

MARQUETTE UNIVERSITY Be The Difference.

- $P\ (x=1)$ **is the probability that the correct answer is given for exactly one question and wrong answers are given for the other three (there are four branches: CWWW, WCWW, WWCW, WWWC—and each has the same probability):**
  - $P(x=1)=4\times\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3}\times\frac{2}{3}=4\times\frac{1}{3}\times\left(\frac{2}{3}\right)^3=\mathbf{0.395}$
- $P\ (x=2)$ **is the probability that correct answers are given for exactly two questions and wrong answers are given for the other two (there are six branches) :**
  - $P(x=2)=6\times\frac{1}{3}\times\frac{1}{3}\times\frac{2}{3}\times\frac{2}{3}=6\times\left(\frac{1}{3}\right)^2\times\left(\frac{2}{3}\right)^2=\mathbf{0.296}$
- $P\ (x=3)$ **is the probability that correct answers are given for exactly three questions and wrong answers are given for the other one(there are four branches) :**
  - $P(x=3)=4\times\frac{1}{3}\times\frac{1}{3}\times\frac{1}{3}\times\frac{2}{3}=4\times\left(\frac{1}{3}\right)^3\times\frac{2}{3}=\mathbf{0.099}$



9

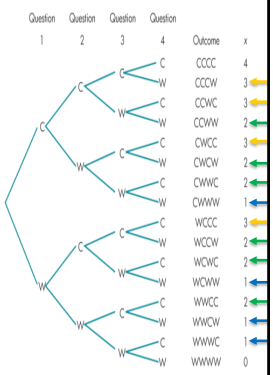
MARQUETTE UNIVERSITY
Be The Difference.

THE BINOMIAL PROBABILITY DISTRIBUTION

- $P(x=0) = \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} = \left(\frac{2}{3}\right)^4 = \frac{16}{81} = \mathbf{0.198}$
- $P(x=1) = 4 \times \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} = 4 \times \frac{1}{3} \times \left(\frac{2}{3}\right)^3 = \mathbf{0.395}$
- $P(x=2) = 6 \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} \times \frac{2}{3} = 6 \times \left(\frac{1}{3}\right)^2 \times \left(\frac{2}{3}\right)^2 = \mathbf{0.296}$
- $P(x=3) = 4 \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{2}{3} = 4 \times \left(\frac{1}{3}\right)^3 \times \frac{2}{3} = \mathbf{0.099}$
- $P(x=4) = \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = \left(\frac{1}{3}\right)^4 = \frac{1}{81} = \mathbf{0.012}$
- **In general:**
- $P(x=k) = \frac{4!}{k!(4-k)!} \times \left(\frac{1}{3}\right)^k \times \left(\frac{2}{3}\right)^{4-k}$, **for** $k = 0,1,2,3,4$
- **Probability distribution:**

x	$P(x)$
0	0.198
1	0.395
2	0.296
3	0.099
4	0.012
	1.000

Probability Distribution for the Four-Question Quiz



10

MARQUETTE UNIVERSITY
Be The Difference.

When a sequence of Bernoulli trials conforms to the following list of requirements it is called a **binomial experiment**.

1. The experiment consists of a fixed number (n) of Bernoulli trials.
2. The probability of success for each trial, denoted by π, is constant from trial to trial. The probability of failure is $\varrho = (1 - \pi)$.
3. The trials are independent.
4. The random variable of interest, X, is the number of observed successes during the n trials.

The probability that X is equal to x can be found in the following fashion. Any particular sequence of x successes occurs with probability $\pi^x(1-\pi)^{(n-x)}$ since there are x successes and $(n-x)$ failures. However, there are $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ possible sequences of x successes. Write $X \sim Bin(n, \pi)$ to indicate the random variable X follows a binomial distribution with parameters n and π. Box (4.3) gives the probability X is equal to x, the mean, the variance, and the moment generating function of a binomial random variable.

11

Binomial Distribution $X \sim Bin(n, \pi)$

$$\mathbb{P}(X = x|n,\pi) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \quad x = 0,1,2,\ldots,n.$$
$$E[X] = n\pi$$
$$\text{var}[X] = n\pi(1-\pi) \qquad (4.3)$$
$$M_X(t) = (\pi e^t + \varrho)^n$$

It is left as an exercise for the student to verify that $E[X] = n\pi$, $\text{var}[X] = n\pi(1-\pi)$, and that the moment generating function of a binomial random variable is $M_X(t) = (\pi e^t + \varrho)^n$.

Binomial Calculator

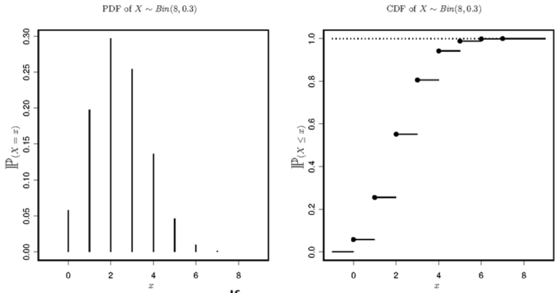


Figure 4.1: Left graph is the probability density function (pdf) of a binomial random variable with $n = 8$ and $\pi = 0.3$. Right graph is the cumulative distribution function (cdf) of a binomial random variable with $n = 8$ and $\pi = 0.3$.

Code to create graphs that represent the probability density function and the cumulative distribution function for a $Bin(8, 0.3)$ random variable follows. The graphs that are created are similar to those in Figure 4.1.

```
➢ par(mfrow=c(1,2), pty="s")
➢ plot(0:8, dbinom(0:8,8,0.3), type="h", xlab="x",
      ylab="P(X=x)", xlim=c(-1,9))
➢ title("PDF for X~Bin(8, 0.3)")
➢ plot(0:8, pbinom(0:8,8,0.3), type="n", xlab="x",
      ylab="P(X<=x)", xlim=c(-1,9), ylim=c(0,1))
➢ segments(-1,0,0,0)
➢ segments(0:8, pbinom(0:8,8,.3), 1:9,pbinom(0:8,8,.3))
➢ lines(0:7, pbinom(0:7,8,.3), type="p", pch=16)
➢ segments(-1,1,9,1, lty=2)
➢ title("CDF for X~Bin(8, 0.3)")
```

MARQUETTE UNIVERSITY
Be The Difference.

4.2.3 Poisson Distribution

- The Poisson distribution is very popular for modelling the number of times particular events occur in given times or on defined spaces.

- For example, one might count the number of phone calls to 911 between 1 A.M. and 2 A.M., the number of accidents at a busy street corner during a 24 hour period, or the number of typographical errors on a single page of this book.

- When the number of outcomes in a given continuous interval are counted, an approximate **Poisson process** with parameter $\lambda > 0$ results if the following conditions are satisfied:

14

MARQUETTE UNIVERSITY
Be The Difference.

POISSON PROCESS

(1) The number of outcomes in nonoverlapping intervals are independent. In other words, the number of outcomes in the interval of time $(0, t]$ are independent from the number of outcomes in the interval of time $(t, t+h]$ for any $h > 0$.

(2) The probability of two or more outcomes in a sufficiently short interval is virtually zero. In other words, provided h is sufficiently small, the probability of obtaining two or more outcomes in the interval $(t, t+h]$ is negligible compared to the probability of obtaining one or zero outcomes in the same interval of time.

(3) The probability of exactly one outcome in a sufficiently short interval or small region is proportional to the length of the interval or region. In other words, the probability of one outcome in an interval of length h is λh.

15

MARQUETTE UNIVERSITY Be The Difference.

The probability distribution of the Poisson random variable X, representing the number of outcomes in a given time interval or space region denoted by t is

$$\mathbb{P}(X = x|\lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \quad x = 0, 1, \ldots, \quad \lambda > 0. \qquad (4.4)$$

Poisson Distribution $X \sim Pois(\lambda)$

$$\mathbb{P}(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$
$$E[X] = \lambda$$
$$\text{var}[X] = \lambda$$
$$M_X(t) = e^{\lambda(e^t - 1)}$$

Poisson Calculator

The Poisson distribution can be used to approximate binomial prob. with $\lambda = n\pi$ provided $\pi \leq 0.1$ and $n\pi \leq 5$. See Example 4.8 on page 44 for an example of how the Poisson distribution is used to approximate the probabilities of a binomial distribution.

16

MARQUETTE UNIVERSITY Be The Difference.

FURTHER COMMENTS:

- Note that the parameter λ, referred to as the intensity parameter, represents the mean number of outcomes in either a fixed time interval or a fixed spatial region.

- The Poisson distribution is particularly appropriate for modelling "rare" phenomena or outcomes where the probability of success is small.

- However, whether or not data can be viewed as Poisson data depends on whether the proportions of 0's, 1's, 2's, and so on are similar to those predicted by the Poisson **pdf** given in Box (4.5).

- Given n independent Poisson random variables $X_1, X_2, \ldots, X_n$ with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectively, $Y = \sum_{i=1}^{n} X_i \sim Pois\left(\sum_{i=1}^{n} \lambda_i = \lambda\right)$.

17

Example 4.4 ▷ ***Poisson: World Cup Soccer*** ◁ The World Cup is played once every four years. National teams from all over the world compete. In 2002 and in 1998, thirty-six teams were invited; whereas, in 1994 and in 1990, only 24 teams participated. The data frame `Soccer` contains three columns: `CGT`, `Game`, and `Goals`. All of the information contained in `Soccer` is indirectly available from the FIFA World Cup website, located at http://fifaworldcup.yahoo.com/. The numbers of goals scored in the regulation 90 minute periods of World Cup soccer matches from 1990 to 2002 are listed in column `Goals`. There were a total of 575 goals scored during regulation time. The game in which the goals were scored is in column `Game`. There were 232 World Cup soccer games played from 1990 to 2002. There were 64 games played in each of 2002 and 1998 and 54 games played in each of 1994 and 1990.

Analyze the number of goals scored during regulation play (90 minutes) of World Cup soccer matches to verify that the scores follow an approximate Poisson distribution. (?)

Solution:

- First, examine the data to see how well it conforms to the Poisson distribution. To calculate the observed number of goals scored during regulation time for the 232 World Cup soccer matches use `table()`.

- Next, let's verify that the mean and variance of Goals are approximately equal

```
➢ library("PASWR")
➢ attach(Soccer)
➢ mean(Goals, na.rm=TRUE)
[1] 2.478448
➢ var(Goals, na.rm=TRUE)
[1] 2.458408
```

Create a table to facilitate comparing the observed values (`OBS`) to the expected values (`EXP`) as well as the empirical proportions (`Empir`) to the theoretical proportions (`TheoP`) for a Poisson Distribution with $\lambda = 2.478448$, the mean number of goals per game. The empirical proportions are merely the number of goals in each category divided by the total number of goals.

R CODE:

```
➢ OBS <- table(Goals)
➢ Empir <- round(OBS/sum(OBS), 3)
➢ TheoP <- round(dpois(0:(length(OBS)-1),
      mean(Goals, na.rm="TRUE")),3)
➢ EXP <- round(TheoP*232, 0)
➢ ANS <- cbind(OBS, EXP, Empir, TheoP)
➢ ANS
  OBS EXP Empir TheoP
0  19  19 0.082 0.084
1  49  48 0.211 0.208
2  60  60 0.259 0.258
3  47  49 0.203 0.213
4  32  31 0.138 0.132
5  18  15 0.078 0.065
6   3   6 0.013 0.027
7   3   2 0.013 0.010
8   1   1 0.004 0.003
```


Code to represent a probability density function and cumulative distribution function for a $Pois(\lambda = 1)$ random variable similar to the one shown in Figure 4.3 on the facing page follows.

```
➢ par(mfrow=c(1,2), pty="s")
➢ plot(0:8, dpois(0:8,1), type="h", xlab="x",
      ylab="P",xlim=c(0,9), main="PDF")
➢ plot(0:8, ppois(0:8,1), type="n", xlab="x",
      ylab="F",xlim=c(0,9), ylim=c(0,1), main="CDF")
➢ segments(-1,0,0,0)
➢ segments(0:8, ppois(0:8,1), 1:9, ppois(0:8,1))
➢ lines(0:7, ppois(0:7,1), type="p", pch=16)
➢ segments(-1,1,9,1, lty=2)
```

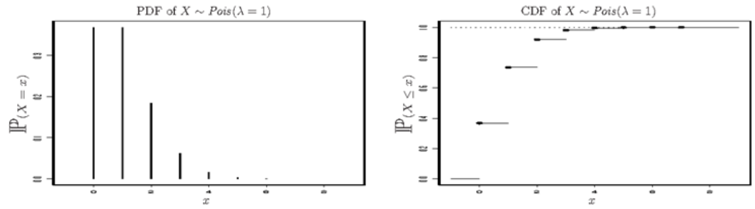


Figure 4.3: Left graph is the probability density function (pdf) of a Poisson random variable with $\lambda = 1$. Right graph is the cumulative distribution function (cdf) of a Poisson random variable with $\lambda = 1$.

Example 4.5 Given a random variable X that follows a Poisson distribution with parameter λ, find the mean and variance of X. Use the fact that:

$$e^{\lambda} = \sum_{r=0}^{\infty} \frac{\lambda^r}{r!} = 1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots$$

Solution:

$$E[X] = \sum_{r=0}^{\infty} r \frac{\lambda^r}{r!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} = \lambda,$$

$$\text{var}[X] = \sum_{r=0}^{\infty} (r-\lambda)^2 \frac{\lambda^r}{r!} e^{-\lambda},$$

rearranging terms

MARQUETTE UNIVERSITY Be The Difference.

SOLUTION CONT.

$$\text{var}[X] = \sum_{r=0}^{\infty} (r-\lambda)^2 \frac{\lambda^r}{r!} e^{-\lambda},$$

$$\begin{aligned}
\text{var}[X] &= e^{-\lambda} \left\{ \sum_{r=0}^{\infty} r^2 \frac{\lambda^r}{r!} + \sum_{r=0}^{\infty} \lambda^2 \frac{\lambda^r}{r!} - 2\lambda \sum_{r=0}^{\infty} r \frac{\lambda^r}{r!} \right\} \\
&= e^{-\lambda} \left\{ \sum_{r=1}^{\infty} r \frac{\lambda^r}{(r-1)!} + \lambda^2 e^{\lambda} - 2\lambda \cdot \lambda \sum_{r=1}^{\infty} \frac{\lambda^{r-1}}{(r-1)!} \right\} \\
&= e^{-\lambda} \left\{ \sum_{r=1}^{\infty} (r-1+1) \frac{\lambda^r}{(r-1)!} + \lambda^2 e^{\lambda} - 2\lambda^2 e^{\lambda} \right\} \\
&= e^{-\lambda} \left\{ \sum_{r=1}^{\infty} (r-1) \frac{\lambda^r}{(r-1)!} + \sum_{r=1}^{\infty} \frac{\lambda^r}{(r-1)!} + \lambda^2 e^{\lambda} - 2\lambda^2 e^{\lambda} \right\} \\
&= e^{-\lambda} \left\{ \lambda^2 + \lambda + \lambda^2 - 2\lambda^2 \right\} e^{\lambda} = \lambda.
\end{aligned}$$

Example 4.7 Telephone calls to a local 911 number are known to follow a Poisson distribution with an average of two calls per minute. Compute the probability that

(a) There will be zero calls during a one minute period.

(b) There will be less than five calls in a one minute period.

(c) There will be less than six calls in one hour.

Solution: The answers are:

(a) $\mathbb{P}(X = 0; \lambda = 2) = \frac{\lambda^0 e^{-\lambda}}{0!} = \frac{2^0}{0!} e^{-2} = 0.135.$

```
> dpois(0,2)
[1] 0.1353353
```

(b) $\mathbb{P}(X \leq 4; \lambda = 2) = \sum_{r=0}^{4} \frac{\lambda^r e^{-\lambda}}{r!} = e^{-2}\left(1 + 2 + \frac{2^2}{2!} + \frac{2^3}{3!} + \frac{2^4}{4!}\right) =$ 0.947.

```
> ppois(4,2)
[1] 0.947347
```

(c) Note that the time period changes from one minute to one hour (60 minutes). Consequently, the average number of calls in one hour is $\lambda' = 2 \times (60) = 120$.

$$\mathbb{P}(X \leq 5; \lambda' = 120) = \sum_{r=0}^{5} \frac{\lambda'^r e^{-\lambda'}}{r!}$$

$$= e^{-120}\left(1 + 120 + \frac{120^2}{2!} + \frac{120^3}{3!} + \frac{120^4}{4!} + \frac{120^5}{5!}\right) = 0.$$

```
> ppois(5,120)
[1] 0
```

24

4.2.4 Geometric Distribution

The geometric distribution, like the binomial distribution, is based on Bernoulli trials. However, the geometric distribution does not fix the number of trials prior to the experiment. The geometric distribution computes the probability the first success occurs after r failures instead of computing the probability of observing x successes in n trials. A random variable X that counts the number of Bernoulli trials that result in failure before the first success is called a **geometric** random variable. Clearly, the probability of a success after r failures is $\pi \times (1-\pi)^r$, which leads to the geometric probability distribution function where $\varrho = 1-\pi$ is the probability of failure as it was for the Bernoulli and binomial distributions. The pdf, mean, variance, and mgf for a geometric random variable are in (4.6).

Geometric Distribution

$$X \sim Geo(\pi)$$

$$\mathbb{P}(X = x; \pi) = \pi\varrho^x,\ x = 0, 1, \ldots$$

$$E[X] = \frac{\varrho}{\pi}$$

$$Var[X] = \frac{\varrho}{\pi^2}$$

$$M_X(t) = \frac{\pi}{1 - \varrho e^t}$$

Geometric Distribution in Wikipedia

25

4.2.5 Negative Binomial Distribution

The geometric random variable counted the number of failures prior to the first success. Quite often, the number of Bernoulli trials required to achieve some fixed number (r) of successes is the problem of interest. When the random variable X is defined as the number of failures prior to the r^{th} success, X has a **negative binomial** distribution written $X \sim NB(r, \pi)$. To find the $\mathbb{P}(X = x)$, first find the probability of $r-1$ successes in the first $x+r-1$ trials, and then multiply by the probability of success on the $(x+r)^{\text{th}}$ trial, $\binom{x+r-1}{r-1}\pi^{r-1}(1-\pi)^x \times \pi$. Combining like terms leads to the probability distribution for the negative binomial given in (4.7). The mean, variance, and **mgf** are also in (4.7):

Negative binomial Distribution in Wikipedia

Negative Binomial Distribution

$$X \sim NB(r, \pi)$$

$$\mathbb{P}(X = x|r, \pi) = \binom{x+r-1}{r-1}\pi^r \varrho^x,\ x = 0, 1, 2, \ldots$$

$$E[X] = r\frac{\varrho}{\pi}$$

$$Var[X] = r\frac{\varrho}{\pi^2}$$

$$M_X(t) = \pi^r(1-\varrho e^t)^{-r}$$

Useful Relationships

1. If n independent random variables $X_1, \ldots, X_n$ have a geometric distribution with parameter π, then the sum of the n independent random variables follows a negative binomial distribution with parameters (n, π).
2. If n independent random variables $X_1, \ldots, X_n$ have a negative binomial distribution with parameters r_i and π, then the sum of the n random variables is $NB\left(\sum_{i=1}^{n} r_i, \pi\right)$.
3. When $X \sim NB(r, \pi)$ and $r = 1$, a negative binomial random variable is the same as a geometric random variable with parameter π.

MARQUETTE UNIVERSITY
Be The Difference.

4.3 Continuous Univariate Distributions

4.3.1 Uniform Distribution (Continuous) X is a **uniform** random variable defined on the interval $[a, b]$ if its **pdf** is given by

$$f(x|a, b) = \frac{1}{b-a}, \quad a \le x \le b.$$

Some common uses of the uniform distribution include random number generation. Box (4.6) gives the **pdf**

Uniform Distribution $\quad X \sim Unif(a, b)$

$$f(x|a,b) = \frac{1}{b-a}, \quad a \le x \le b$$

$$E[X] = \frac{b+a}{2}$$

$$\text{var}[X] = \frac{(b-a)^2}{12} \tag{4.6}$$

$$M_X(t) = \begin{cases} \dfrac{e^{tb} - e^{ta}}{t(b-a)} & \text{if } t \neq 0 \\ 1 & \text{if } t = 0 \end{cases}$$

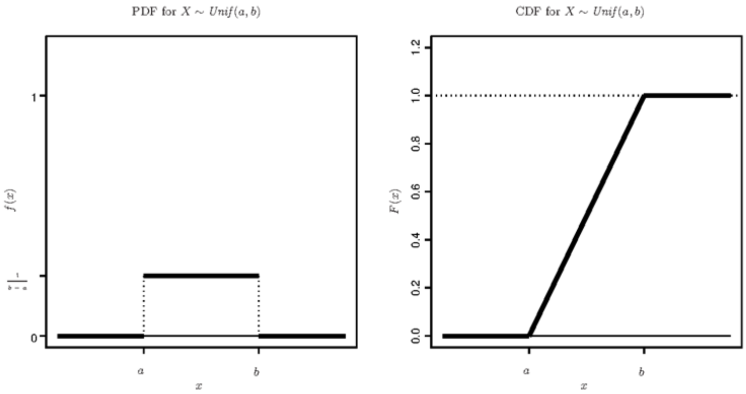


Figure 4.4: The **pdf** and **cdf** for the random variable $X \sim Unif(a, b)$

Example 4.12 Given a continuous random variable X defined over $[a, b]$ with **pdf** $f(x|a,b) = \frac{1}{b-a}$, $a \le x \le b$, find the expected value and the variance of X.

Solution: Using the definition for a continuous random variable

$$E[X] = \int_a^b x \cdot f(x)\,dx = \int_a^b \frac{x}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^2}{2}\Big|_a^b$$
$$= \frac{b^2-a^2}{2(b-a)} = \frac{(b+a)(b-a)}{2(b-a)} = \frac{b+a}{2}.$$

$$E\left[X^2\right] = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a} \cdot \frac{x^3}{3}\Big|_a^b = \frac{b^3-a^3}{3(b-a)}$$

$$\text{var}[X] = E\left[X^2\right] - (E[X])^2 = \frac{b^3-a^3}{3(b-a)} - \frac{(b+a)^2}{4}$$
$$= \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} - \frac{(b+a)^2}{4} = \frac{4(b^2+ab+a^2)}{12} - \frac{3(b+a)^2}{12}$$
$$= \frac{4b^2+4ab+4a^2-(3b^2+6ab+3a^2)}{12} = \frac{b^2-2ab+a^2}{12}$$
$$= \frac{(b-a)^2}{12}.$$

Generating Pseudo Random Numbers

The generation of pseudo random numbers is fundamental to any simulation study. The term "pseudo random" is used because once one value in such a simulation is known, the next values can be determined without fail, since they are generated by an algorithm. Most major statistical software systems have reputable pseudo random number generators. When using R, the user can specify one of several different random number generators including a user supplied random number generator. For more details, type `?RNG` at the R prompt. Generation of random values from named distributions is accomplished with the S command ***rdist***, where ***dist*** is the distribution name; however, it is helpful to understand some of the basic ideas of random number generation in the event a simulation does not involve a named distribution. When the user wants to generate a sample from a continuous random variable X with **cdf** F, one approach is to use the *Inverse Transformation Method*. This method simply sets $F_X(X) = U \sim \textit{Unif}(0,1)$ and solves for X assuming $F_X^{-1}(U)$ actually exists.

30

Example 4.14 Generate a sample of 1000 random values from a continuous distribution with **pdf** $f(x) = \frac{4}{3}x(2-x^2), \quad 0 \le x \le 1$. Verify that the mean and variance of the 1000 random values are approximately equal to the mean and variance of the given **pdf**.

Solution: First, the **cdf** is found, then $F_X(x)$ is set equal to u and solved.

$$F_X(x) = \int_0^x \frac{4}{3}t\left(2-t^2\right)dt = \frac{4}{3}\left(x^2 - \frac{x^4}{4}\right) = \frac{1}{3}x^2\left(4-x^2\right),$$

Solving for x in terms of u by setting $u = F_X(x)$:

$u = \frac{1}{3}x^2\left(4-x^2\right)$	
$3u = 4x^2 - x^4$	multiply by 3 and distribute x^2
$-3u+4 = x^4 - 4x^2 + 4$	multiply by -1 and add 4 to co
$-3u+4 = (x^2-2)^2$	factor
$\pm\sqrt{-3u+4} = x^2 - 2$	take the square root of both sides
$2 \pm \sqrt{-3u+4} = x^2$	add 2
$\pm\sqrt{2 \pm \sqrt{-3u+4}} = x$	take the square root of both sides

31

The mean and variance of the 1000 simulated random values using `set.seed(33)`

```
➢ set.seed(33)
➢ U <- runif(1000)
➢ X <- sqrt((2-sqrt(4-3*U)))
➢ mean(X)
[1] 0.6152578
➢ > var(X)
[1] 0.05809062

➢ f <- function(x){(4/3)*x*(2-x^2)}
➢ ex <- function(x){x*f(x)}
➢ ex2 <- function(x){x^2*f(x)}
➢ EX <- integrate(ex,0,1)
➢ EX2 <- integrate(ex2,0,1)
➢ VX <- EX2$value - EX$value^2
➢ c(EX$value,EX2$value,VX)
[1] 0.62222222 0.44444444 0.05728395
```

32

4.3.2 Exponential Distribution When observing a Poisson process such as that in Example 4.4 on page 25 where the number of outcomes in a fixed interval such as the number of goals scored during 90 minutes of World Cup soccer is counted, the random variable X, which measures the number of outcomes (number of goals), is modeled with the Poisson distribution. However, not only is X, the number of outcomes in a fixed interval, a random variable but also is the waiting time between successive outcomes. If W is the waiting time until the first outcome of a Poisson process with mean $\lambda > 0$, then the **pdf** for W is

$$f(w) = \begin{cases} \lambda e^{-\lambda w} & \text{if } w \geq 0 \\ 0 & \text{if } w < 0 \end{cases}$$

Proof: Since waiting time is nonnegative, $F(w) = 0$ for $w < 0$. When $w \geq 0$,

$$\begin{aligned} F(w) = \mathbb{P}(W \leq w) &= 1 - \mathbb{P}(W > w) \\ &= 1 - \mathbb{P}(\text{no outcomes in } [0, \text{w}]) \end{aligned}$$

33

Proof: Since waiting time is nonnegative, $F(w) = 0$ for $w < 0$. When $w \geq 0$,

$$\begin{aligned} F(w) = \mathbb{P}(W \leq w) &= 1 - \mathbb{P}(W > w) \\ &= 1 - \mathbb{P}(\text{no outcomes in } [0, \text{w}]) \\ &= 1 - \frac{(\lambda w)^0 e^{-\lambda w}}{0!} \\ &= 1 - e^{-\lambda w} \end{aligned}$$

Consequently, when $w > 0$, the **pdf** of W is $F'(w) = f(w) = \lambda e^{-\lambda w}$.

The exponential distribution is characterized by a lack of memory property and is often used to model lifetimes of electronic components as well as waiting times for Poisson processes. A random variable is said to be **memoryless** if

$$\mathbb{P}(X > t_2 + t_1 | X > t_1) = \mathbb{P}(X > t_2) \text{ for all } t_1, t_2 \geq 0. \qquad (4.8)$$

Exponential Distribution $\quad X \sim Exp(\lambda)$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$E[X] = \frac{1}{\lambda}$$

$$\text{var}[X] = \frac{1}{\lambda^2}$$

$$M_X(t) = (1 - \lambda^{-1} t)^{-1} \text{ for } t < \lambda$$

- Exponential Distribution Applet

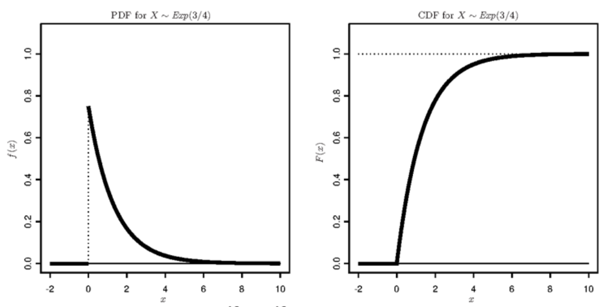


Figure 4.5: The **pdf** and **cdf** for the random variable $X \sim Exp(\lambda = 0.75)$

Example 4.16 Given $X \sim Exp(\lambda)$, find the mean and variance of X.

Solution:

$$E[X] = \int_0^\infty x\lambda e^{-\lambda x}\,dx.$$

Integrating by parts where $u = x$, and $dv = \lambda e^{-\lambda x}\,dx$ obtain

$$E[X] = -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty -e^{-\lambda x}\,dx$$
$$= 0 - \frac{1}{\lambda e^{\lambda x}}\Big|_0^\infty = \frac{1}{\lambda}.$$

Before finding the variance of X, find $E\left[X^2\right]$

$$E\left[X^2\right] = \int_0^\infty x^2\lambda e^{-\lambda x}\,dx \qquad (4.10)$$

MARQUETTE UNIVERSITY
Be The Difference.

$$E\left[X^2\right] = \int_0^\infty x^2\lambda e^{-\lambda x}\,dx$$

Note that $E[X] = \int_0^\infty x\lambda e^{-\lambda x}\,dx \Rightarrow \frac{E[X]}{\lambda} = \int_0^\infty xe^{-\lambda x}\,dx$ and integrate (4.10) by parts where $u = x^2$ and $dv = \lambda e^{-\lambda x}\,dx$:

$$E\left[X^2\right] = -x^2e^{-\lambda x}\Big|_0^\infty - \int_0^\infty -2xe^{-\lambda x}\,dx$$
$$= 0 + 2\frac{E[X]}{\lambda} = \frac{2}{\lambda^2}.$$

Using the fact that $\text{var}[X] = E\left[X^2\right] - (E[X])^2$, obtain $\text{var}[X] = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$. ■

Quite often, the **pdf** for the exponential is expressed as

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \quad x \geq 0, \quad \theta > 0,$$

where $\theta = \frac{1}{\lambda}$. Of course, the **mgf** is then written as $M_X(t) = (1-\theta t)^{-1}$ and the reparameterized mean and variance are θ and θ^2 respectively.

Note the relationship between the Poisson mean and the exponential mean. Given a Poisson process with mean λ, the waiting time until the first outcome has an exponential distribution with mean $\frac{1}{\lambda}$. That is if λ represents the number of outcomes in a unit interval, $\frac{1}{\lambda}$ is the mean waiting time for the first change.

REMEMBER

The exponential distribution is characterized by a lack of memory property and is often used to model lifetimes of electronic components as well as waiting times for Poisson processes. A random variable is said to be **memoryless** if

$$\mathbb{P}(X > t_2 + t_1 | X > t_1) = \mathbb{P}(X > t_2) \text{ for all } t_1, t_2 \geq 0. \tag{4.8}$$

If X denotes the lifetime of an electronic component following an exponential distribution with mean $\frac{1}{\lambda}$, (4.8) implies that the probability the component will work for $t_2 + t_1$ hours given that it has worked for t_1 hours is the same as the probability that the component will function for at least t_2 hours. In other words, the component has no memory of having functioned for t_1 hours. Equation (4.8) is equivalent to

$$\frac{\mathbb{P}(X > t_2 + t_1, X > t_1)}{\mathbb{P}(X > t_1)} = \mathbb{P}(X > t_2),$$

which is equivalent to

$$\mathbb{P}(X > t_2 + t_1) = \mathbb{P}(X > t_2)\mathbb{P}(X > t_1). \tag{4.11}$$

Since $\mathbb{P}(X > t_2 + t_1) = e^{-\lambda(t_2+t_1)} = e^{-\lambda t_2} e^{-\lambda t_1} = \mathbb{P}(X > t_2)\mathbb{P}(X > t_1)$

MARQUETTE UNIVERSITY Be The Difference.

Example 4.17 ▷ ***Exponential Distribution: Light Bulbs*** ◁ If the life of a certain type of light bulb has an exponential distribution with a mean of eight months, find

(a) The probability that a randomly selected light bulb lasts between three and twelve months.

(b) The 95$^{\text{th}}$ percentile of the distribution.

(c) The probability that a light bulb that has lasted for ten months will last more than twenty five months.

Solution: The answers are:
(a) Since $X \sim Exp\left(\lambda = \frac{1}{8}\right)$, the probability that a randomly selected light bulb lasts between three and twelve months is

$$\mathbb{P}(3 < X < 12) = \int_3^{12} \frac{1}{8} e^{-x/8}\, dx = -e^{-x/8}\Big|_3^{12} = -0.2231 + 0.6873 = 0.46$$

40

MARQUETTE UNIVERSITY Be The Difference.

R CODE:

```
➢ round(pexp(12,1/8) - pexp(3,1/8),4)
[1] 0.4642
➢ f1 <- function(x){(1/8)*exp(-x/8)}
➢ integrate(f1,3,12) # For R
0.4641591 with absolute error < 5.2e-15
```

(b) The 95$^{\text{th}}$ percentile is the value x_{95} such that

$$\int_{-\infty}^{x_{95}} f(x)\, dx = \int_0^{x_{95}} \frac{1}{8} e^{-x/8}\, dx = \frac{95}{100}$$

$$-e^{-x/8}\Big|_0^{x_{95}} = 1 - e^{-\frac{x_{95}}{8}} = \frac{95}{100}$$

$$e^{-\frac{x_{95}}{8}} = \frac{5}{100}$$

$$x_{95} = -8\ln(0.05) = 23.96586$$

```
➢ qexp(0.95,1/8)
[1] 23.96586
```

41

(c) The probability that a light bulb that has lasted for ten months will last more than twenty five months mathematically is written $\mathbb{P}(X > 25|X > 10)$. Because an exponential distribution is present, (4.8) can be used to say that this is equal to $\mathbb{P}(X > 15) = e^{-15/8} = 0.153355$. Solve the problem with S as follows.

```
➢ 1-pexp(15,1/8) ## applying memoryless property
[1] 0.153355

➢ (1-pexp(25,1/8))/(1-pexp(10,1/8)) ## without applying
  memoryless
[1] 0.153355
```


4.3.3 Gamma Distribution Some random variables are always nonnegative and yield distributions of data that tend to be skewed. The waiting time until a certain number of malfunctions in jet engines, and similar scenarios where the random variable of interest is the waiting time until a certain number of events take place yield skewed distributions. The **gamma** distribution is often used to model the waiting time until the α^{th} event in a Poisson process.

Before defining the gamma distribution, review the definition of the gamma function. The **gamma function**, $\Gamma(\alpha)$, is defined by:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx, \quad \alpha > 0 \tag{4.12}$$

Some of the more important properties of the gamma function include:

1. For $\alpha > 0$, $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$
2. For any positive integer, n, $\Gamma(n) = (n-1)!$
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

In Section 4 on page 64, it was proved that the waiting time until the first outcome in a Poisson process follows an exponential distribution. Now, let W denote the waiting time until the α^{th} outcome and derive the distribution of W in a similar fashion. Since waiting time is nonnegative, $F(w) = 0$ for $w < 0$. When $w \geq 0$,

$$\begin{aligned} F(w) = \mathbb{P}(W \leq w) &= 1 - \mathbb{P}(W > w) \\ &= 1 - \mathbb{P}(\text{fewer than } \alpha \text{ outcomes in } [0, \text{w}]) \\ &= 1 - \sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}}{k!} \end{aligned}$$

Consequently, when $w > 0$, the **pdf** of W is $F'(w) = f(w)$ whenever this derivative exists. It follows then that

$$\begin{aligned} f(w) = F'(w) &= -\sum_{k=0}^{\alpha-1} \frac{(\lambda w)^k e^{-\lambda w}(-\lambda) + e^{-\lambda w} k(\lambda w)^{k-1}\lambda}{k!} \\ &= \frac{\lambda(\lambda w)^{\alpha-1} e^{-\lambda w}}{(\alpha-1)!} = \frac{\lambda^\alpha w^{\alpha-1} e^{-\lambda w}}{\Gamma(\alpha)} \end{aligned}$$

Notice that different shapes are produced in Figure 4.7 for different values of α. For this reason, α is often called the shape parameter associated with the gamma distribution. The parameter λ is referred to as the scale parameter.

- **Gamma Distribution Applet**

Gamma Distribution $\quad X \sim \Gamma(\alpha, \lambda)$

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

$$E[X] = \frac{\alpha}{\lambda}$$

$$\text{var}[X] = \frac{\alpha}{\lambda^2}$$

$$M_X(t) = (1 - \lambda^{-1} t)^{-\alpha} \text{ for } t < \lambda$$

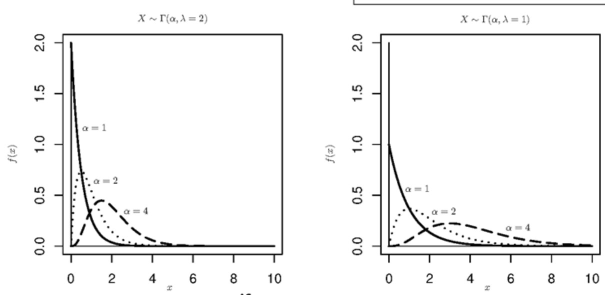


Figure 4.7: Graphical illustration of the **pdf**s of a $\Gamma(\alpha, 2)$ and a $\Gamma(\alpha, 1)$ random variable for $\alpha = 1, 2$, and 4 respectively.

MARQUETTE UNIVERSITY
Be The Difference.

Useful Relationships

1. Given $X \sim \Gamma(\alpha, \lambda)$. When $\alpha = 1$, the resulting random variable is $X \sim Exp(\lambda)$. That is, the exponential distribution is a special case of the gamma distribution.

2. Given $X \sim \Gamma(\alpha, \lambda)$. When $\alpha = \frac{n}{2}$ and $\lambda = \frac{1}{2}$, the resulting random variable has a chi-square distribution with n degrees of freedom.

3. Given $X \sim \Gamma(\alpha, \lambda)$. Provided α is a positive integer, the resulting distribution is known as the Erlang. In this case, the Erlang distribution gives the waiting time until the α^{th} occurrence when the number of outcomes in an interval of length t follows a Poisson distribution with parameter λt.

MARQUETTE UNIVERSITY
Be The Difference.

Example 4.20 Suppose that the average arrival rate at a local fast food drive through window is three cars per minute ($\lambda = 3$). Find

(a) The probability that at least five cars arrive in 120 seconds.

(b) The probability that more than one minute elapses before the second car arrives.

(c) If one car has already gone through the drive through, what is the average waiting time before the third car arrives?

Solution: The answers are:

(a) If the average number of car arrivals follows a Poisson distribution with a rate of three cars per minute, then the average rate of arrival for two minutes is six cars. Given that $X \sim Pois(\lambda = 6)$, the

$$\mathbb{P}(X \geq 5) = 1 - \mathbb{P}(X \leq 4) = 1 - \sum_{x=0}^{4} \frac{e^{-6}6^x}{x!} = 1 - 0.2850565 = 0.7149435.$$

```
➢ 1 - ppois(4,6)
[1] 0.7149435
```

MARQUETTE UNIVERSITY
Be The Difference.

(b) Let W represent the waiting time until the α^{th} outcome. It follows that $W \sim \Gamma(\alpha = 2, \lambda = 3)$. Consequently, the

$$\mathbb{P}(W > 1) = 1 - \mathbb{P}(W \leq 1) = 1 - \mathbb{P}(\Gamma(2,3) \leq 1)$$

$$= 1 - \int_0^1 \frac{3^2}{\Gamma(2)} x^{2-1} e^{-3x} dx \quad = 1 - \int_0^1 3x\, e^{-3x} 3\, dx$$

Using integration by parts where $u = 3x$ and $dv = 3e^{-3x} dx$,

$$\int_0^1 3x\, e^{-3x} 3\, dx = -3xe^{-3x} \Big|_0^1 + \int_0^1 3e^{-3x}\, dx$$

$$= -3e^{-3} + \left[-e^{-3x} \Big|_0^1 \right] = -3e^{-3} + \left[-e^{-3} + 1 \right]$$

$$= 1 - 4e^{-3} = 0.8008517.$$

In other words, $\mathbb{P}(W > 1) = 1 - 0.8008517 = 0.1991483$.

```
> 1 - pgamma(1,2,3)
[1] 0.1991483
gam23<-function(x){9*x*exp(-3*x)}
integrate(gam23,1,Inf)                          # R
0.1991483 with absolute error < 2.5e-05
```

48

MARQUETTE UNIVERSITY
Be The Difference.

(c) This problem is really asking for the mean of a $\Gamma(\alpha = 2, \lambda = 3)$ random variable. Note: $\alpha = 2$ since one car has already arrived and the problem requests the average waiting time until the third car arrives. Therefore, $E[X] = \frac{\alpha}{\lambda} = \frac{2}{3}$. In other words, there is an average wait of $\frac{2}{3}$ of a minute before the arrival of the third vehicle given one vehicle has already arrived. ■

49

COMMON CONTINUOUS DISTRIBUTIONS:

- Uniform
  - Wiki
- Exponential
  - Wiki
- Gamma
  - Wiki
- Weibull
  - Wiki
- Beta
  - Wiki
- Cauchy
  - Wiki
- Normal (μ=mean, σ^2=variance)
  - Wiki
- t (ν=df)
  - Wiki
- Chi-Square (ν=df)
  - Wiki
- F (ν_1=df$_1$, ν_2=df$_2$)
  - Wiki

4.3.7 Normal (Gaussian) Distribution

- The **normal** or **Gaussian** distribution is more than likely the most important distribution in statistical applications.
- This is due to the fact that many numerical populations have distributions that can be approximated with the normal distribution.
- Examples of distributions following an approximate normal distribution include physical characteristics such as the height and weight of a particular species. Further, certain statistics, such as the mean, follow an approximate normal distribution when certain conditions are satisfied.
- The **pdf**, mean, variance, and **mgf** for a normal random variable X with mean μ and variance σ^2 are provided in Box (4.14).

MARQUETTE UNIVERSITY
Be The Difference.

GAUSSIAN (NORMAL) DISTRIBUTION

$$\text{Normal Distribution} \quad X \sim N(\mu, \sigma)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty,$$

$$\text{where } -\infty < \mu < \infty, \text{ and } 0 < \sigma < \infty. \qquad (4.14)$$

$$E[X] = \mu$$

$$\text{var}[X] = \sigma^2$$

$$M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

- **Z-table ("D2L > Useful Links > Z, T and Chi^2 Tables")**
  - $P(Z \leq z)$, **where** Z **is a standard Normal,** $Z \sim N(\mu = 0, \sigma^2 = 1)$.
- **Normal calculator**

- The **pdf** for a normal random variable has an infinite number of centers and spreads, depending on both μ and σ, respectively.
- Although there are an infinite number of possible normal distributions, all normal distributions have a bell shape that is symmetric around the distribution's mean.

52

MARQUETTE UNIVERSITY
Be The Difference.

- Small values of σ produce distributions that are relatively close to the distribution's mean.
- On the other hand, values of σ that are large produce distributions that are quite spread out around the distribution's mean.
- Figure 4.8 on page 100 illustrates three normal distributions with identical means, μ, and increasing variances as the distributions are viewed from left to right.

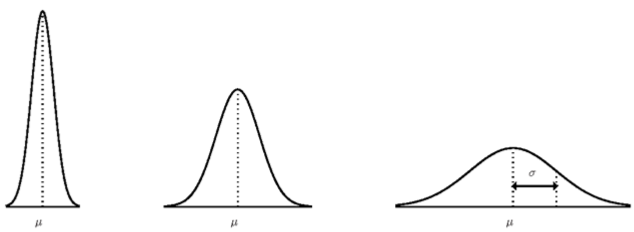


Figure 4.8: Three normal distributions each with an increasing σ value as read from left to right

Mean and Standard deviation of Normal distribution

53

The cdf for a normal random variable, X, with mean, μ, and standard deviation, σ, is

$$F(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt. \qquad (4.15)$$

A normal random variable with $\mu = 0$ and $\sigma = 1$, often denoted Z, is called a **standard normal** random variable. The cdf for the standard normal distribution, given in (4.17), is computed by first standardizing the random variable X, where $X \sim N(\mu, \sigma)$, using the change of variable formula in (4.16).

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1) \qquad (4.16)$$

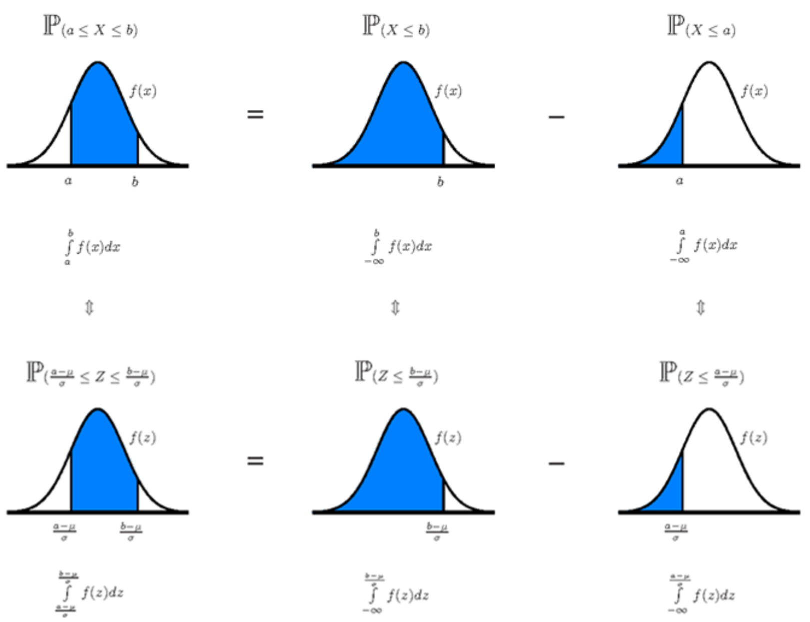
$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}\left(Z \leq \frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{(x-\mu)}{\sigma}} e^{-\frac{z^2}{2}} dz \qquad (4.17)$$

Neither the integral for (4.17) nor the integral for (4.15) can be computed with standard techniques of integration. However, (4.17) has been numerically evaluated and tabled. Further, any normal random variable can be converted to a standard normal random variable using (4.16). The process of computing $\mathbb{P}(a \leq X \leq b)$ where

$X \sim N(\mu, \sigma)$ is graphically illustrated in Figure 4.9 on page 104. Throughout the text, the convention z_α is used to represent the value of the standard normal random variable Z that has α of its area to the left of said value. In other words, $\mathbb{P}(Z < z_\alpha) = \alpha$. Another notation that is also used in the text is $\mathbf{\Phi}(z_\alpha) = \alpha$. Basically, the $\mathbf{\Phi}(\text{value})$ is the same as $\mathbb{P}(Z < \text{value})$. That is, Φ is the **cdf** of the standard normal distribution. Likewise, $\Phi^{-1}(\alpha) = z_\alpha$. The Φ notation for the **cdf** and inverse **cdf** is used more in Chapter 10.

To find the numerical value of X_α, where $X \sim N(\mu, \sigma)$ and α is the area (or probability) to the left of the value X_α, use the S command `qnorm(p, mean=MValue, sd=SValue)` where `p` is the area or probability (this is equivalent to α) to the left of X_α, `MValue` is the value of the mean, and `SValue` is the value of the standard deviation. Note that if one is dealing with the standard normal distribution, the `mean=MValue` or `sd=SValue` arguments are not needed.

56



57

MARQUETTE UNIVERSITY Be The Difference.

Example 4.21 Scores on a particular standardized test follow a normal distribution with a mean of 100 and standard deviation of 10.

(a) What is the probability that a randomly selected individual will score between 90 and 115?

(b) What score does one need to be in the top 10%?

(c) Find the constant c such that $\mathbb{P}(105 \le X \le c) = 0.10$.

- **Solution:**

(a) To find $\mathbb{P}(90 \le X \le 115)$, first draw a picture representing the desired area such as the one in Figure 4.10 on page 112. Note that finding the area between 90 and 115 is equivalent to finding the area to the left of 115 and from that area, subtracting the area to the left of 90. In other words,

$$\mathbb{P}(90 \le X \le 115) = \mathbb{P}(X \le 115) - \mathbb{P}(X \le 90).$$

58

MARQUETTE UNIVERSITY Be The Difference.

To find $\mathbb{P}(X \le 115)$ and $\mathbb{P}(X \le 90)$, one can standardize using (4.16). That is,

$$\mathbb{P}(X \le 115) = \mathbb{P}\left(Z \le \frac{115 - 100}{10}\right) = \mathbb{P}(Z \le 1.5),$$

and

$$\mathbb{P}(X \le 90) = \mathbb{P}\left(Z \le \frac{90 - 100}{10}\right) = \mathbb{P}(Z \le -1.0).$$

Using the S commands `pnorm(1.5)` and `pnorm(-1)`, find the areas to the left of 1.5 and -1.0 to be 0.9332 and 0.1586 respectively.

Consequently,

$$\begin{aligned}\mathbb{P}(90 \le X \le 115) &= \mathbb{P}(-1.0 \le Z \le 1.5)\\ &= \mathbb{P}(Z \le 1.5) - \mathbb{P}(Z \le -1.0)\\ &= 0.9332 - 0.1587 = 0.7745.\end{aligned}$$

59

(b) Finding the value c such that 90% of the area is to its left is equivalent to finding the value c such that 10% of its area is to the right. That is, finding the value c that satisfies $\mathbb{P}(X \leq c) = 0.90$ is equivalent to finding the value c such that $\mathbb{P}(X \geq c) = 0.10$. Since the `qnorm()` function refers to areas to the left of a given value by default, solve

$$\mathbb{P}(X \leq c) = \mathbb{P}\left(Z = \frac{X-100}{10} \leq \frac{c-100}{10}\right) = 0.90 \text{ for } c.$$

Using `qnorm(.9)`, find the Z value (1.2816) such that 90% of the area in the distribution is to the left of that value. Consequently, to be in the top 10%, one needs to be more than 1.2816 standard deviations above the mean.

$$\frac{c-100}{10} \overset{\text{set}}{=} 1.2816$$

and solve for $c \Rightarrow c = 112.816$.

To be in the top 10%, one needs to score 112.816 or higher.

(c) $\mathbb{P}(105 \leq X \leq c) = 0.10$ is the same as

$$\mathbb{P}(X \leq c) = 0.10 + \mathbb{P}(X \leq 105) = 0.10 + \mathbb{P}\left(Z \leq \frac{105-100}{10}\right).$$

Using `pnorm(.5)`,

$$\mathbb{P}\left(Z \leq \frac{105-100}{10}\right) = \mathbb{P}(Z \leq 0.5) = 0.6915.$$

It follows then that $\mathbb{P}(X \leq c) = 0.7915$. Using `qnorm(.7915)`, gives 0.8116.

$$\mathbb{P}(X \leq c) = \mathbb{P}\left(Z = \frac{X-100}{10} \leq \frac{c-100}{10}\right) = 0.7915$$

is found by solving $\frac{c-100}{10} = 0.8116 \Rightarrow c = 108.116$

Note that a Z value of 0.8116 has 79.15% of its area to the left of that value.

MARQUETTE UNIVERSITY
Be The Difference.

The following S commands can be used to solve (a),(b), and (c) respectively.

(a) $\mathbb{P}(90 \leq X \leq 115)$

```
> pnorm(115,100,10) - pnorm(90,100,10)
[1] 0.7745375
```

(b) $\mathbb{P}(X \leq c) = 0.90$

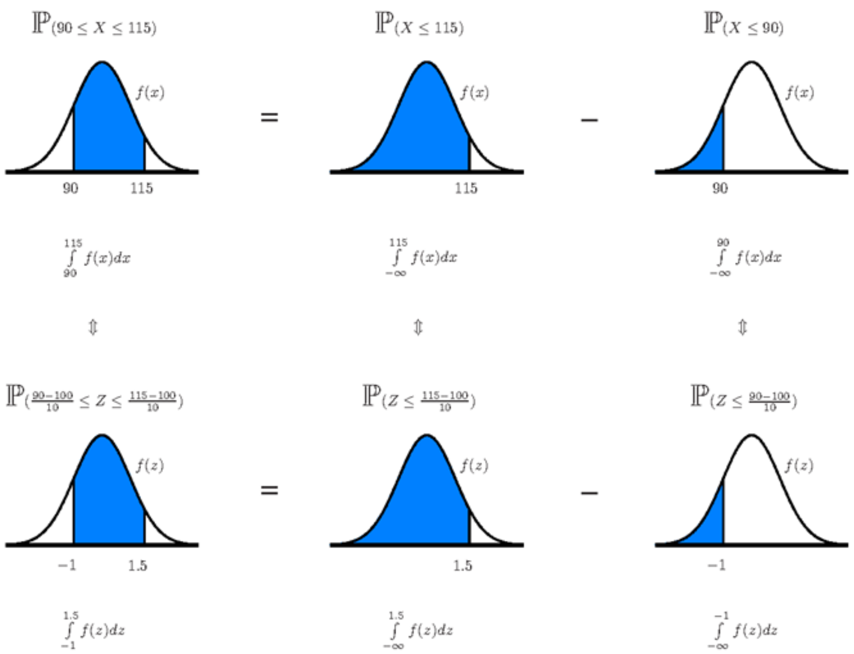
```
> qnorm(.90,100,10)
[1] 112.8155
```

(c) $\mathbb{P}(105 \leq X \leq c) = 0.10$

```
> qnorm(.10 + pnorm(105,100,10),100,10)
[1] 108.1151
```

62

MARQUETTE UNIVERSITY
Be The Difference.



63

Example 4.26 ▷ *Normal Distribution: Cell Phone Compo*
◁ Most mobile appliances today allow the consumer to switch from the built in speaker and microphone to an external source. A manufacturer of cell phones wants to package an external speaker and microphone for hands free operation. A new company has patented a component that allows the on-resistance flatness for both the microphone and speaker to be lower than ever before. The cell phone company requires that the on-resistance flatness be less than 0.7 ohms ($\boldsymbol{\Omega}$). If it is known that 50% of the components from the new company have an ohm rating of 0.5 $\boldsymbol{\Omega}$ or less and 10% have an ohm rating of 0.628 $\boldsymbol{\Omega}$ or greater and the distribution of the ohm ratings is normal, then:

(a) Find the mean and standard deviation for the distribution of the ohm rating of the components.
(b) If a component is selected at random, what is the probability that its on-resistance flatness will be less than 0.7 $\boldsymbol{\Omega}$?
(c) If 20 components are selected at random, what is the probability that at least 19 components will have on-resistance flatness values less than 0.7 $\boldsymbol{\Omega}$?

Solution: Let X = the ohm rating of the patented components. (a) Because a normal distribution is symmetric, the mean equals the median. It is known that 50% of the components have an ohm rating of 0.5 $\boldsymbol{\Omega}$ or less, so $\mu_X = 0.5$. To calculate the standard deviation of the components' ohm ratings, use the fact that "10% have an ohm rating of 0.628 $\boldsymbol{\Omega}$ or greater."

$$\text{This means that} \quad \mathbb{P}(X \le 0.628) = 0.9,$$

$$\text{which implies} \quad \mathbb{P}\left(Z = \frac{X - 0.5}{\sigma} \le \frac{.628 - .5}{\sigma}\right) = 0.9.$$

$$\text{Because } \mathbb{P}(Z \le 1.28) = 0.9, \text{ set} \quad \frac{0.628 - 0.5}{\sigma} = 1.28$$

$$\text{and solve for } \sigma. \quad \frac{0.628 - 0.5}{1.28} = \sigma$$

$$\text{Therefore} \quad \sigma = 0.1.$$

(b) Calculate that the probability a component has an on-resistance flatness less than 0.7 $\boldsymbol{\Omega}$.

$$\begin{aligned}\mathbb{P}(X \leq 0.7) &= \mathbb{P}\left(Z = \frac{X - 0.5}{.1} \leq \frac{0.7 - 0.5}{0.1}\right) \\ &= \mathbb{P}(Z \leq 2) \\ &= 0.97725\end{aligned}$$

The answer computed with S is

```
> p <- pnorm(0.7,0.5,0.1)
> p
[1] 0.97725
```


(c) Calculate the probability that at least 19 of the 20 components will have an on-resistance flatness value less than 0.7 $\boldsymbol{\Omega}$. Let $Y \sim Bin(20, 0.97725)$.

$$\mathbb{P}(Y \geq 19) = \sum_{i=19}^{20} \binom{20}{i} (0.97725)^i (1 - 0.97725)^{20-i} = 0.9250$$

To compute the answer with S type

```
> sum(dbinom(19:20,20,p))
[1] 0.92497
```

Quantile-Quantile Plots for Normal Distributions

- Many of the techniques presented later in the book assume the underlying distribution is normal.
- One of the more useful graphical procedures for assessing distributions is the quantile-quantile (QQ) plot.
- To help determine whether the underlying distribution is normal, use the S function `qqnorm()`.
- To understand the `qqnorm()` function, one needs to have some understanding of S's `quantile()` function.
- Recall that the cumulative distribution function (**cdf**) is $F(x) = P(X \leq x)$.
- The `quantile()` function is the inverse of the **cdf**, where this exists; that is $Q(u) = F^{-1}(u)$.

- The `qqnorm()` function works by first computing the quantiles of the points $(i - 1/2)/n$ for the standard normal distribution.
- The ordered sample values are then plotted against the quantiles.
- When the resulting plot is linear, it indicates the sample values have a normal distribution.
- To help assess the linearity of the `qqnorm()` plot, it is often quite helpful to plot a straight line through the 25^{th} and 75^{th} percentiles also referred to as the first and third quartiles using the S function `qqline()` which connects the pair of points (First Quartile Standard Normal, First Quartile Data), (Third Quartile Standard Normal, Third Quartile Data).
- For example, consider the values stored in the variable `scores` of the data frame `Score` and reported in Table 4.2 on page 123 which are the scores a random sample of twenty college freshmen received on a standardized test.

MARQUETTE UNIVERSITY
Be The Difference.

Table 4.2: Standardized scores (data frame Score)

119	107	96	107	97	103	94	106	87	112
99	99	90	106	110	99	105	100	100	94

- The points $(i - 1/2)/n$ are calculated as

  $(1{-}1/2)/20 = 0.025, (2{-}1/2)/20 = 0.075, \ldots, (20{-}1/2)/20 = 0.975,$

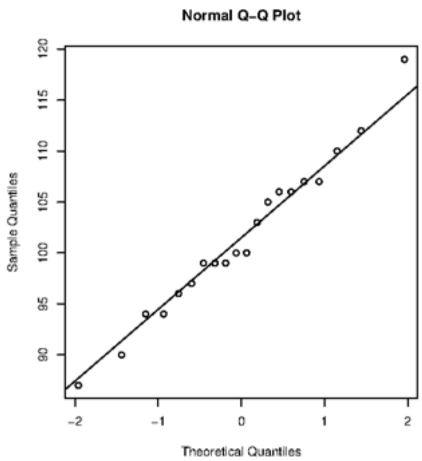
  corresponding standard normal quantiles of $\{0.025, 0.075, \ldots, 0.975\}$ are computed with `qnorm()` to be $\{-1.96, -1.44, \ldots, 1.96\}$ respectively.
- The S function `qqnorm()` plots the quantiles $\{-1.96, -1.44, \ldots, 1.96\}$ versus the ordered values in the sample, $\{87, 90, \ldots, 119\}$ as shown in Figure 4.11 on page 124.
- The pair of points (First Quartile Standard Normal, First Quartile Data), (Third Quartile Standard Normal, Third Quartile Data) are (-0.637, 96.75) and (0.637, 106.25) respectively. Note how the line in Figure 4.11 on page 124 created using the S function `qqline()` goes through the points (-0.637, 96.75) and (0.637, 106.25).

70

MARQUETTE UNIVERSITY
Be The Difference.

To compute the pairs of values plotted in an S quantile-quantile plot, issue the following commands.

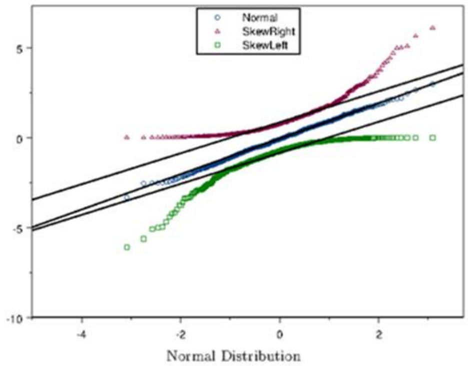
```
> attach(Score)
> par(pty="s")
> X <- (1:20-1/2)/20
> Xs <- qnorm(X)
> Ys <- sort(scores)
> plot(Xs,Ys)
> quantile(Xs,c(0.25, 0.75))
       25%        75%
-0.6371739  0.6371739
> quantile(Ys,c(0.25, 0.75))
   25%    75%
 96.75 106.25
```



71

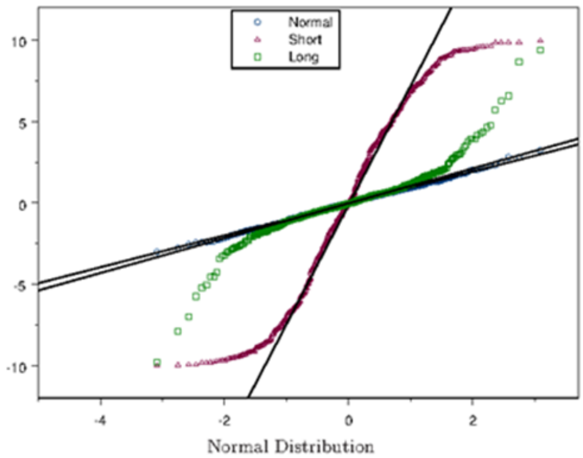
MARQUETTE UNIVERSITY
Be The Difference.

It is possible to tell from a quantile-quantile plot whether the distribution has shorter or longer tails than a normal distribution. In addition, the quantile-quantile plot will show whether a distribution is skewed and in which direction the distribution is skewed. The right quantile-quantile plots in Figure 4.12 on the facing page illustrate how distributions that have a positive skew will appear as upward opening U shapes in the quantile-quantile plot, while distributions with a negative skew have downward facing U shapes.



MARQUETTE UNIVERSITY
Be The Difference.

The left quantile-quantile plots in
Figure 4.12 on the next page illustrate how distributions that have short tails relative to the normal distribution will have an S shape while distributions with tails longer than the normal distribution will have an inverted S shape.



MARQUETTE UNIVERSITY
Be The Difference.

QUESTIONS?

- **ANY QUESTION?**

74