Chapter 6: Probability Distributions

Yu Yang

School of Statistics
University of Minnesota

September 23, 2022

Example 6.1

Toss a coin 3 times. We know sample space (S) is

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

Let X = the number of heads (H).

What are the values X can take?

Example 6.1

Toss a coin 3 times. We know sample space (S) is

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

Let X = the number of heads (H).

What are the values X can take? 0, 1, 2, 3

The event that X takes on any one these values is **random** and can be assigned a probability. So, X is **variable** and X is random, so

X *is a* ***random variable****!*

Definitions

A **random variable** is a variable whose value is a numerical outcome of a random phenomenon.

Notation: We often use X, Y and Z to denote random variables and use x, y and z to denote their realized/observed values, respectively.

Example:
X = number of heads in 3 coin tosses
$x = 2$ (2 heads were observed in 3 coin tosses).

Types of Random Variable

1. Discrete: possible values can be listed

2. Continuous: possible values form an interval

Discrete Random Variable

A **discrete** random variable takes on values that can be listed.
Examples:

- $X = \#$ of M&Ms in a bag
- $X = \#$ of broken mirrors in a shipment
- $X = \#$ of accidents/day in a factory

We describe a random variable using its probability distribution.

Probability Distribution for a Discrete Random Variable

Probability distributions for discrete random variables have two properties:

1. They assign probabilities to each possible outcome.

   Example: $X = \#$ heads in 3 coin tosses. Then the probability distribution will assign a probability to all values of $X = 0, 1, 2, 3$.

   $P(X = 1) =$ Probability that the number of heads in 3 coin tosses is 1

2. The probabilities add up to 1.

$$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = 1$$

Example 6.2

Let X = the number of bases for a randomly selected at-bat. In 2015, the probability distribution of X (excluding walks) for a Minnesota Twins player was as follows (all probabilities rounded to three decimals):

x	0	1	2	3	4
probability	.753	.159	.051	.008	.029

(a) Is this a valid probability distribution?

Example 6.2

Let X = the number of bases for a randomly selected at-bat. In 2015, the probability distribution of X (excluding walks) for a Minnesota Twins player was as follows (all probabilities rounded to three decimals):

x	0	1	2	3	4
probability	.753	.159	.051	.008	.029

(a) Is this a valid probability distribution?
Yes, the probabilities add up to 1!

(b) For a randomly selected at-bat, what is the probability the player got at least one base?

(b) For a randomly selected at-bat, what is the probability the player got at least one base?

At least one base: $X \geq 1$

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\ &= .159 + .051 + .008 + .029 \\ &= .247 \end{aligned}$$

We can also calculate this a faster way!
If A is the event that $X \geq 1$, then $A^c = X < 1$. Since $P(A) = 1 - P(A^c)$,

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - .753 = .247$$

Center and Spread of a Probability Distribution

We can describe distributions of random variables just as we described distributions of data in a sample in Chapter 2. In particular, we can describe the center and spread of a probability distribution using the mean and standard deviation.

- Mean: $\mu = \mathrm{E}(X) =$ "Expected value of X"
  Measures the center tendency of the distribution of X
- Standard Deviation: $\sigma =$ "Standard Deviation of X"
  Measures the spread of the distribution of X.

Interpretation: μ is the "long-run" average outcome. That is, μ is what we expect the average to be for a very long series of repetitions. Similarly, σ is a "long-run" standard deviation.

Calculating μ and σ for Discrete Distributions

Let x represent the possible outcomes of discrete random variable X.
Then the formulas for the mean, variance, and standard deviation are:

$$\mu = \sum_x xP(X = x) = x_1 P(X = x_1) + x_2 P(X = x_2) + \dots$$

$$\sigma^2 = \sum_x (x - \mu)^2 P(X = x) = (x_1 - \mu)^2 P(X = x_1) + (x_2 - \mu)^2 P(X = x_2) + \dots$$

$$\sigma = \sqrt{\sigma^2}$$

Example 6.2 Continued

Recall: We let X be a random variable defined by the number of bases earned by a randomly selected Minnesota Twins at-bat from the 2015 season. The probability distribution of X is:

x	0	1	2	3	4
probability	.753	.159	.051	.008	.029

Find the mean and standard deviation of this probability distribution.

Example 6.2 Continued

$$
\begin{aligned}
\mu &= \sum_{x=0}^{4} xP(X=x) \\
&= (0)(.753)+(1)(.159)+(2)(.051)+(3)(.008)+(4)(.029) \\
&= .401, \\
\sigma^2 &= \sum_{x=0}^{4}(x-\mu)^2 P(X=x) \\
&= (0-.401)^2(.753)+(1-.401)^2(.159)+(2-.401)^2(.051) \\
&\quad +(3-.401)^2(.008)+(4-.401)^2(.029) \\
&= .738, \\
\sigma &= \sqrt{\sigma^2} \\
&= .859.
\end{aligned}
$$

Example 6.2 Continued

If we define a similar random variable Y for the Rangers, who hit many more home runs than the Twins, would we expect the standard deviation of Y to be smaller or larger than the standard deviation of X?

Example 6.2 Continued

If we define a similar random variable Y for the Rangers, who hit many more home runs than the Twins, would we expect the standard deviation of Y to be smaller or larger than the standard deviation of X?

Larger! Since Y has a much higher probability to take the value of 4 than X, the distribution of Y is more spread out than that of X, and therefore Y has a larger standard deviation.

Example 6.3

Let X denote the response of a randomly selected person to the question, "What is the ideal number of children for family to have?"
According to a recent General Social Survey, the probability distribution of X for men in the U.S. is as follows:

x	0	1	2	3	4
probability	0.04	0.03	0.57	0.23	0.13

Find and interpret the mean of this probability distribution.

Example 6.3

Let X denote the response of a randomly selected person to the question, "What is the ideal number of children for family to have?"
According to a recent General Social Survey, the probability distribution of X for men in the U.S. is as follows:

x	0	1	2	3	4
probability	0.04	0.03	0.57	0.23	0.13

Find and interpret the mean of this probability distribution.

$$\begin{aligned} \mu &= \sum xP(X = x) \\ &= (0)(.04) + (1)(.03) + (2)(.57) + (3)(.23) + (4)(.13) \\ &= 2.38, \end{aligned}$$

If we sample a large number of American men and compute the average of their ideal numbers of children, then in the long run that average will be close to 2.38.

Binary Response

Many experiments result in a binary response (two possible outcomes).

For instance, a person may

- *accept or decline* an offer from a bank for a credit card,
- *have or not have* health insurance,
- *pass or fail* an Introductory Statistics course.

Modeling the Number of Successes

We often summarize such variables by counting the *number* of the cases with an outcome of interest in a sample. For example,

- X = total number of heads out of 5 coin tosses
- X = total number of subjects (out of 50) that die during a clinical trial

Under certain conditions, a random variable X that counts the number of observations of a particular type has a probability distribution called the *binomial distribution*.

Binomial Distribution

Let the *discrete* random variable X be the number of "successes" in n independent trials where each trial has a success probability of p. Then X has a **binomial distribution**, which we denote by

$$X \sim Bin(n, p).$$

Note: Binomial distributions are uniquely specified by n and p.

Key Conditions

We need to highlight a few key conditions that $X \sim Bin(n, p)$ must satisfy:

1. The number of trials, n, is fixed.
2. The n trials are independent.
3. Each trial has only 2 possible outcomes, "success" and "failure".
4. The probability of success, p, is the same for each trial.

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let $X =$ number of heads.

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let $X =$ number of heads.
   Yes, $X \sim Bin(3, 0.5)$.

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let $X =$ number of heads.
   Yes, $X \sim Bin(3, 0.5)$.
2. Poll 100 people and let $X =$ number that supported the Iraq war.

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let $X =$ number of heads.
   Yes, $X \sim Bin(3, 0.5)$.
2. Poll 100 people and let $X =$ number that supported the Iraq war.
   Yes, $X \sim Bin(100, p)$ where p is unknown.

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let $X =$ number of heads.
   Yes, $X \sim Bin(3, 0.5)$.
2. Poll 100 people and let $X =$ number that supported the Iraq war.
   Yes, $X \sim Bin(100, p)$ where p is unknown.
3. Toss a coin until you get 10 heads. Let $X = \#$ of flips it takes.

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let $X =$ number of heads.
   Yes, $X \sim Bin(3, 0.5)$.
2. Poll 100 people and let $X =$ number that supported the Iraq war.
   Yes, $X \sim Bin(100, p)$ where p is unknown.
3. Toss a coin until you get 10 heads. Let $X = \#$ of flips it takes.
   No, n is not fixed. (Aside: X follows a negative binomial distribution in this case.)

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let X = number of heads.
   Yes, $X \sim Bin(3, 0.5)$.
2. Poll 100 people and let X = number that supported the Iraq war.
   Yes, $X \sim Bin(100, p)$ where p is unknown.
3. Toss a coin until you get 10 heads. Let $X = \#$ of flips it takes.
   No, n is not fixed. (Aside: X follows a negative binomial distribution in this case.)
4. Suppose an urn contains 3 red marbles and 2 blue marbles. Choose 2 marbles without replacement and let X = number that are blue.

Example 6.11

In the following experiments, does random variable X follow a binomial distribution? If yes, identify n and p. If not, explain why.

1. Toss a coin 3 times and let $X =$ number of heads.
   Yes, $X \sim Bin(3, 0.5)$.
2. Poll 100 people and let $X =$ number that supported the Iraq war.
   Yes, $X \sim Bin(100, p)$ where p is unknown.
3. Toss a coin until you get 10 heads. Let $X = \#$ of flips it takes.
   No, n is not fixed. (Aside: X follows a negative binomial distribution in this case.)
4. Suppose an urn contains 3 red marbles and 2 blue marbles. Choose 2 marbles without replacement and let $X =$ number that are blue.
   No, trials are not independent. (Aside: X follows a hypergeometric distribution in this case.)

Calculating Probabilities of the Binomial Distribution

Suppose $X \sim Bin(n, p)$ and let k be any value in $\{0, 1, \cdots, n\}$. Then

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

that is, (number of ways to arrange k successes among n trials) $\times$ (probability of any one of these arrangements) where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \qquad \text{and} \qquad k! = k \times (k-1) \times \cdots \times 2 \times 1.$$

Note 1: Another notation for $\binom{n}{x}$ is ${}_nC_x$.
Note 2: By definition, $0! = 1$.

Example 6.12

Suppose that the multiple choice portion of an exam contains 3 questions, each with 4 possible choices. An unprepared student decides to guess the answers at random. Let $X =$ number of questions the student guesses correctly.

(a) What is the distribution of X?

Example 6.12

Suppose that the multiple choice portion of an exam contains 3 questions, each with 4 possible choices. An unprepared student decides to guess the answers at random. Let $X =$ number of questions the student guesses correctly.

(a) What is the distribution of X?
$X \sim Bin(3, \frac{1}{4})$.

Example 6.12

Suppose that the multiple choice portion of an exam contains 3 questions, each with 4 possible choices. An unprepared student decides to guess the answers at random. Let X = number of questions the student guesses correctly.

(a) What is the distribution of X?
$X \sim Bin(3, \frac{1}{4})$.

(b) Let C be the event of a correct answer and W be the event of a wrong answer. Write down the sample space, S, for this experiment.

Example 6.12

Suppose that the multiple choice portion of an exam contains 3 questions, each with 4 possible choices. An unprepared student decides to guess the answers at random. Let X = number of questions the student guesses correctly.

(a) What is the distribution of X?
$X \sim Bin(3, \frac{1}{4})$.

(b) Let C be the event of a correct answer and W be the event of a wrong answer. Write down the sample space, S, for this experiment.
$S = \{CCC, CCW, CWC, WCC, WWC, WCW, CWW, WWW\}$

Example 6.12

Suppose that the multiple choice portion of an exam contains 3 questions, each with 4 possible choices. An unprepared student decides to guess the answers at random. Let X = number of questions the student guesses correctly.

(a) What is the distribution of X?
$X \sim Bin(3, \frac{1}{4})$.

(b) Let C be the event of a correct answer and W be the event of a wrong answer. Write down the sample space, S, for this experiment.
$S = \{CCC, CCW, CWC, WCC, WWC, WCW, CWW, WWW\}$

(c) Is each outcome in S equally likely?

Example 6.12

Suppose that the multiple choice portion of an exam contains 3 questions, each with 4 possible choices. An unprepared student decides to guess the answers at random. Let X = number of questions the student guesses correctly.

(a) What is the distribution of X?
$X \sim Bin(3, \frac{1}{4})$.

(b) Let C be the event of a correct answer and W be the event of a wrong answer. Write down the sample space, S, for this experiment.
$S = \{CCC, CCW, CWC, WCC, WWC, WCW, CWW, WWW\}$

(c) Is each outcome in S equally likely?
No. Correct guesses are less likely than wrong guesses, so, for example, CCC is far less likely than WWW.

Example 6.12 (Continued)

(d) Find $P(X = 1)$ *without* using binomial formula.

Example 6.12 (Continued)

(d) Find $P(X = 1)$ *without* using binomial formula.
$P(X = 1) = P(CWW, WCW, \text{ or } WWC) = 3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = 0.421875.$

Example 6.12 (Continued)

(d) Find $P(X = 1)$ *without* using binomial formula.
$P(X = 1) = P(CWW, WCW, \text{ or } WWC) = 3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = 0.421875.$

(e) Write down the probability distribution of X using the formula for binomial probabilities.

Example 6.12 (Continued)

(d) Find $P(X = 1)$ *without* using binomial formula.
$P(X = 1) = P(CWW, WCW, \text{ or } WWC) = 3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = 0.421875.$

(e) Write down the probability distribution of X using the formula for binomial probabilities.

$$P(X = x) = \binom{3}{x}\left(\frac{1}{4}\right)^{x}\left(\frac{3}{4}\right)^{3-x} \text{ for } x = 0, 1, 2, 3$$

Example 6.12 (Continued)

(d) Find $P(X = 1)$ *without* using binomial formula.
$P(X = 1) = P(CWW, WCW, \text{ or } WWC) = 3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = 0.421875.$

(e) Write down the probability distribution of X using the formula for binomial probabilities.

$$P(X = x) = \binom{3}{x}\left(\frac{1}{4}\right)^{x}\left(\frac{3}{4}\right)^{3-x} \text{ for } x = 0, 1, 2, 3$$

(f) Find the probability of getting one answer correct.

Example 6.12 (Continued)

(d) Find $P(X = 1)$ *without* using binomial formula.
$P(X = 1) = P(CWW, WCW, \text{ or } WWC) = 3\left(\frac{1}{4}\right)\left(\frac{3}{4}\right)\left(\frac{3}{4}\right) = 0.421875.$

(e) Write down the probability distribution of X using the formula for binomial probabilities.

$$P(X = x) = \binom{3}{x}\left(\frac{1}{4}\right)^{x}\left(\frac{3}{4}\right)^{3-x} \text{ for } x = 0, 1, 2, 3$$

(f) Find the probability of getting one answer correct.

$$P(X = 1) = \binom{3}{1}\left(\frac{1}{4}\right)^{1}\left(\frac{3}{4}\right)^{2} = 0.421875.$$

Example 6.12 (Continued)

(g) Find the probability of getting *at least* one answer correct.

Example 6.12 (Continued)

(g) Find the probability of getting *at least* one answer correct.

$$P(X \geq 1) = P(X = 1) + P(X = 2) + P(X = 3).$$

or

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - 0.421875 = 0.578125$$

Mean and Standard Deviation of the Binomial Distribution

The binomial probability distribution for n trials with probability p of success on each trail has mean μ and standard deviation σ given by :

$$\mu = np, \quad \sigma = \sqrt{np(1-p)}.$$

Example 6.12 (Continued)

(h) Find the mean and standard deviation of X.

Example 6.12 (Continued)

(h) Find the mean and standard deviation of X.

$\mu = np = 3(\frac{1}{4}) = 0.75, \quad \sigma = \sqrt{np(1-p)} = \sqrt{3(\frac{1}{4})(\frac{3}{4})} = 0.75$

If an unprepared student guesses the answers of 3 multiple choice questions (each question with 4 possible choices) randomly, the "long-run" average number of correct answers is $\mu = 0.75$.

Example 6.13

As of February 2018, 77% of Americans own a smartphone according to Pew Research Center's survey of smartphone ownership. Suppose 20 Americans are chosen at random and let X = number of those that own smartphones.

(a) What is the distribution of X?

Example 6.13

As of February 2018, 77% of Americans own a smartphone according to Pew Research Center's survey of smartphone ownership. Suppose 20 Americans are chosen at random and let $X =$ number of those that own smartphones.

(a) What is the distribution of X?

$X \sim Bin(20, 0.77)$

Example 6.13

As of February 2018, 77% of Americans own a smartphone according to Pew Research Center's survey of smartphone ownership. Suppose 20 Americans are chosen at random and let X = number of those that own smartphones.

(a) What is the distribution of X?
$X \sim Bin(20, 0.77)$

(b) Calculate the mean and standard deviation of X. Interpret the results.

Example 6.13

As of February 2018, 77% of Americans own a smartphone according to Pew Research Center's survey of smartphone ownership. Suppose 20 Americans are chosen at random and let $X =$ number of those that own smartphones.

(a) What is the distribution of X?
$X \sim Bin(20, 0.77)$

(b) Calculate the mean and standard deviation of X. Interpret the results.
$\mu = np = 20(0.77) = 15.4$.
Interpretation: If we pick many samples of size 20, on average we observe 15.4 who own smartphones.

$\sigma = \sqrt{np(1-p)} = \sqrt{20(0.77)(1-0.77)} = 1.882$
Interpretation: If we pick many samples of size 20, on average we observe the number of Americans that own smartphones is $\sigma = 1.882$ away from the mean $\mu = 15.4$.

Example 6.13 (Continued)

(c) Find the probability that exactly 12 of 20 randomly sampled Americans own smartphones.

Example 6.13 (Continued)

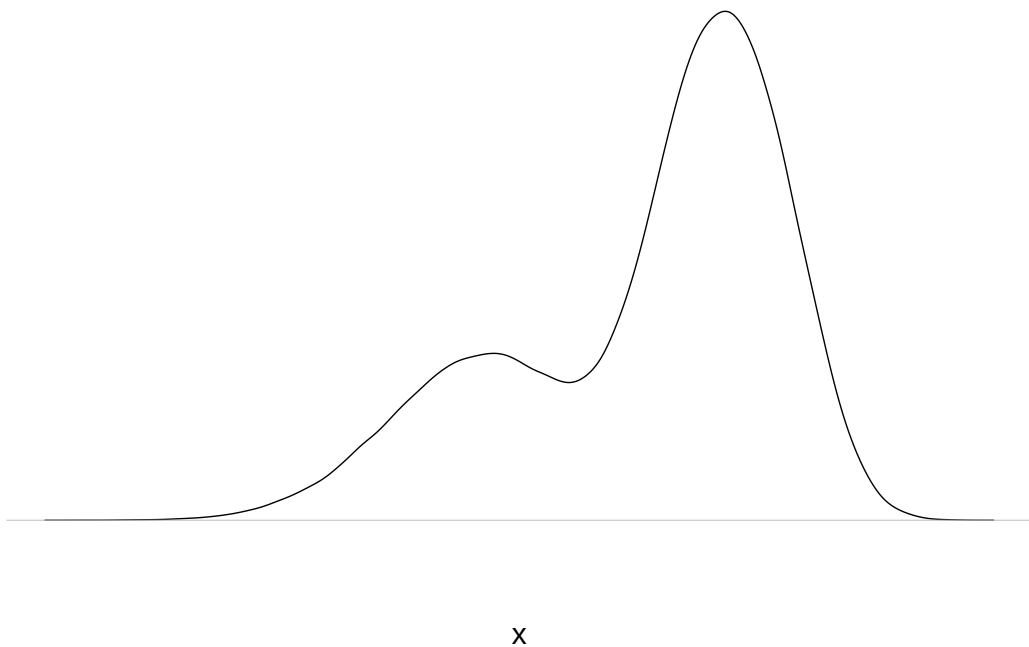
(c) Find the probability that exactly 12 of 20 randomly sampled Americans own smartphones.

$$P(X = 12) = \binom{20}{12}(0.77)^{12}(1 - 0.77)^{8} = 0.04285273.$$

Continuous Random Variables: Density Curves

A continuous random variable takes on values that form an interval. It is *impossible* to list its values. Therefore, instead of writing out the probability distribution as we did for discrete random variables, we describe the distribution using a density curve.

A **density curve** specifies the probability distribution of a continuous random variable.



Properties of a Density Curve:

1. always non-negative (≥ 0)

2. $P(a < X < b)$ = area under the curve above (a, b). Calculating this sometimes requires calculus.
   Therefore $P(X = a) = 0$ for all a.

3. Total area under density curve $= P(-\infty < X < \infty) = 1$.

Example 6.4

A certain bus is equally likely to be anywhere from zero to twenty minutes late.

Let X = the number of minutes that the bus is late.

Therefore X is *uniformly distributed* on the interval $[0, 20]$.

Question: Draw the density curve for X.

Total Area = 1

$$f(x) = \begin{cases} 1/20, & \text{if } 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

Example 6.4

What is the probability the bus is...

(a) At least 13 minutes late?

(b) Exactly 10 minutes late?

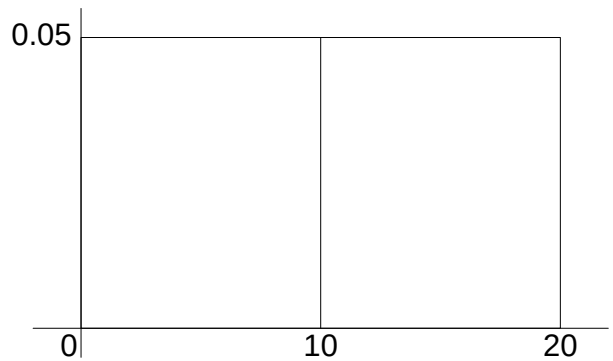
(c) Between 10 and 13 minutes late?

(d) Find x such that 75% of the waiting times are below x.

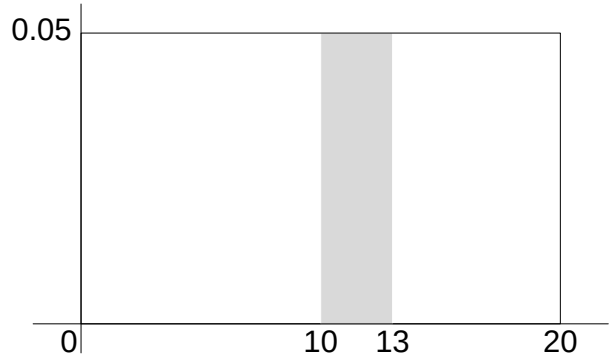
What is the probability the bus is...

1. At least 13 minutes late? $P(X \geq 13) = 7(1/20) = .35$

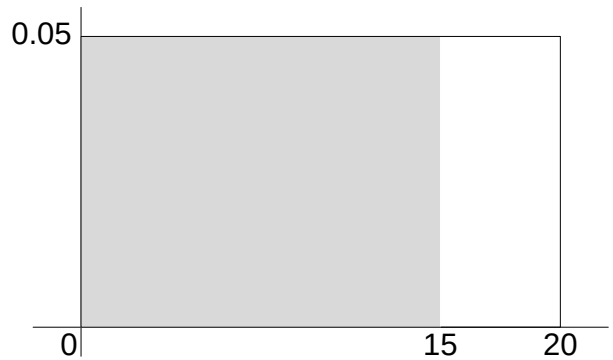
2. Exactly 10 minutes late?
   $P(X = 10) = 0$ (the area of a line is 0)



3. Between 10 and 13 minutes late?
   $P(10 \leq X \leq 13) = 3(1/20) = .15$

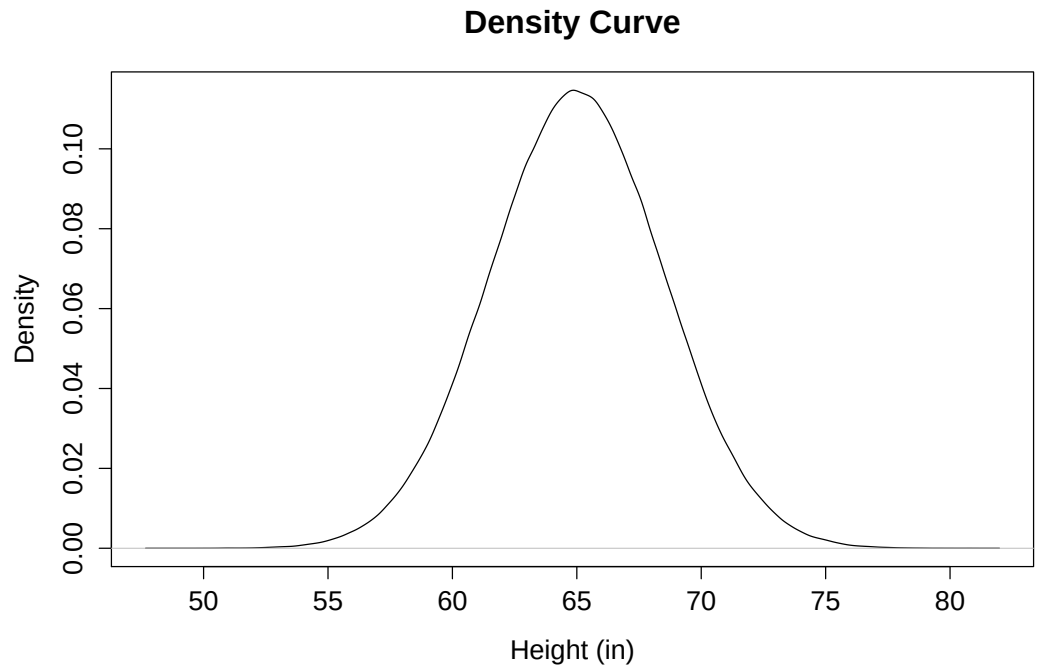


4. Find x such that 75% of the waiting times are below x.
   We want to find x such that $P(X \leq x) = 0.75$. We see that $P(X \leq x) = x(1/20)$, so $x = 15$ is the correct value.



Histograms as Discrete Approximations of Density Curves

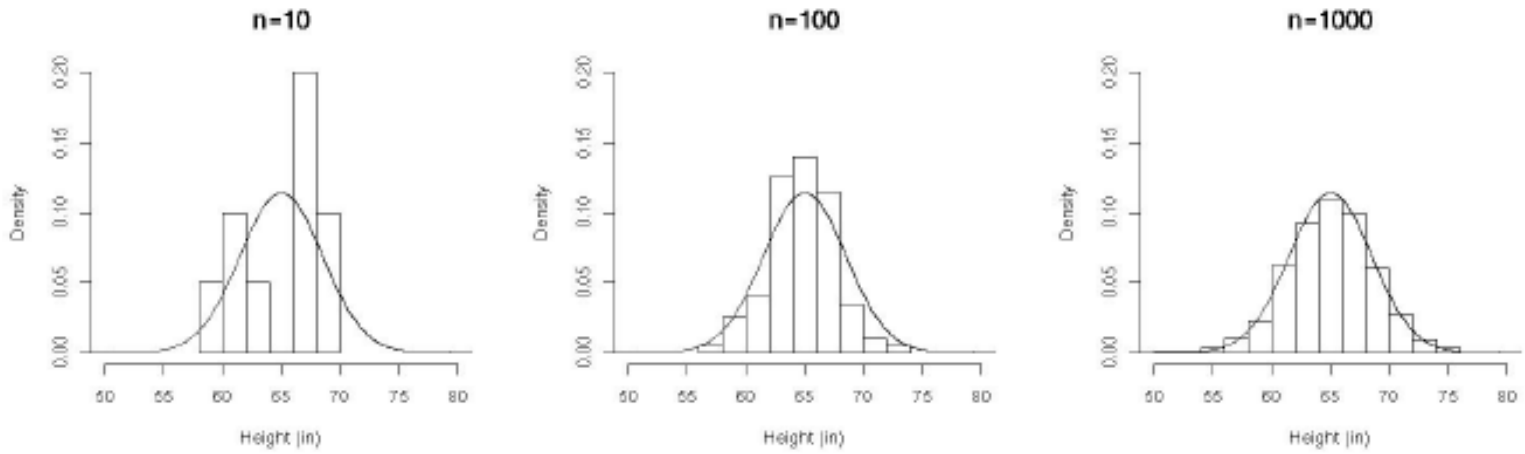
Let X = height (in inches) of a North American female. The following density curve specifies the probability distribution of X:



Suppose we don't have this information. All is not lost! To get an idea of what the distribution of heights is, we can **randomly sample** n North American women and measure their heights.

Example 6.5 Continued

Histograms for differently-sized samples:



As sample size increases, the shape of **the histogram for the sample** approaches **the true density curve**.

In general: the more data and finer the scale on the x-axis of a histogram, the better it approximates the true density curve.

Population: density
Sample: histogram

Center of a Density Curve

1. Median: the point that divides the area under the curve in half

2. Mean: the balance point, the point at which the curve would balance if made of solid material.

Spread of a Density Curve

We can measure the spread of a distribution using its standard deviation. This value if tough to eyeball, but can be calculated mathematically.

We will return to it later.

Notation Reminder:

$\mu =$ mean of a probability distribution

$\sigma =$ standard deviation of a probability distribution

$\bar{x} =$ sample mean

$s =$ sample standard deviation

Again, we need to remember the distinction between a **population** and a **sample** from that population.

The Normal Distribution

We now focus on a common continuous distribution, the **normal distribution**.

The distributions of many quantitative random variables are well-approximated by the normal distribution. It will therefore become an indispensable tool in approaching statistical inference.

Normal distributions are specified by μ and σ for

$$-\infty < \mu < \infty \quad \text{and} \quad 0 < \sigma < \infty.$$

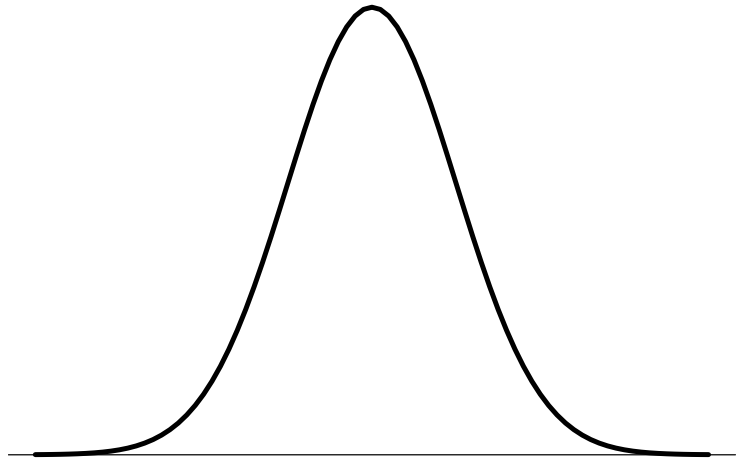
If we know that the random variable has a normal distribution and we also know μ and σ, then we know exactly what the probability distribution looks like. However, we rarely know μ and σ.

Notation

$X \sim N(\mu, \sigma)$: X is "normally distributed" with mean μ and standard deviation σ.

$Z \sim N(0, 1)$: Z is "standard normal" (normal with $\mu = 0$ and $\sigma = 1$).

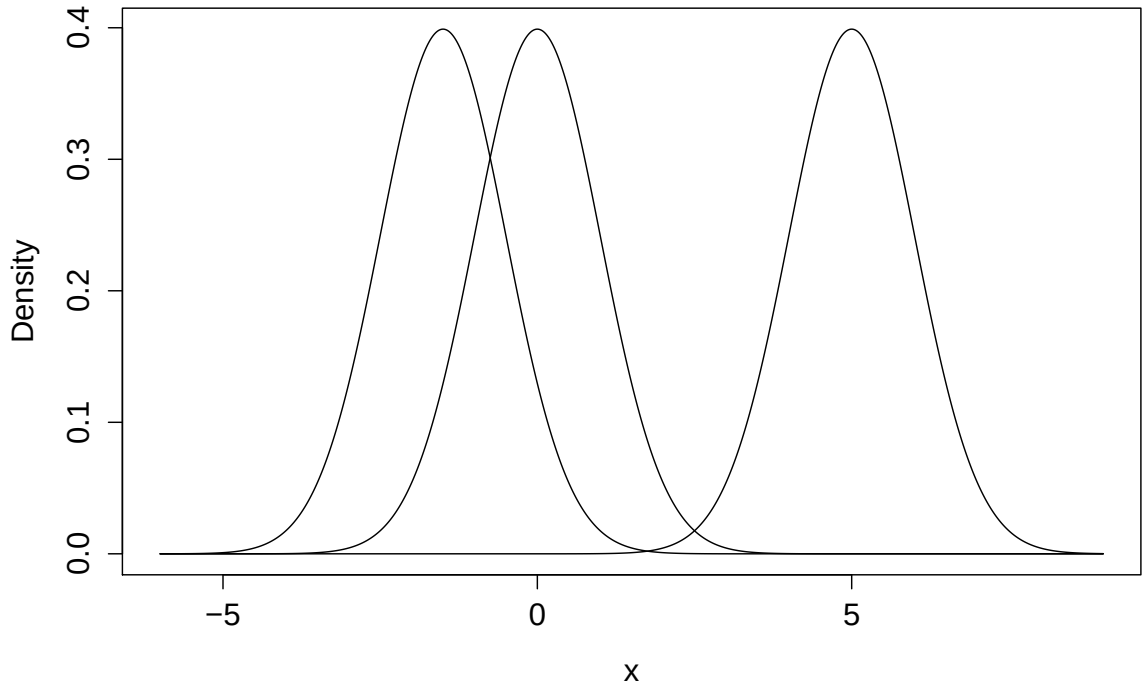
Normal Distribution: Density Curve



Features:

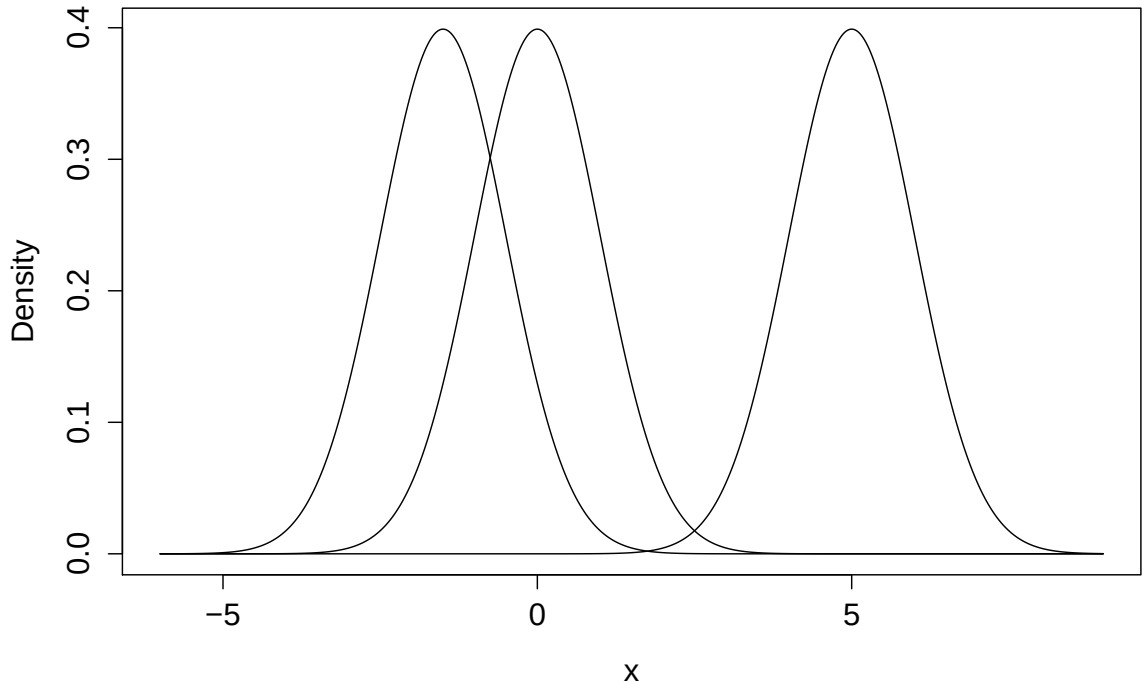
- bell shaped
- symmetric
- centered at μ (where is this on the plot?)
- mean = median = μ.

Changing μ



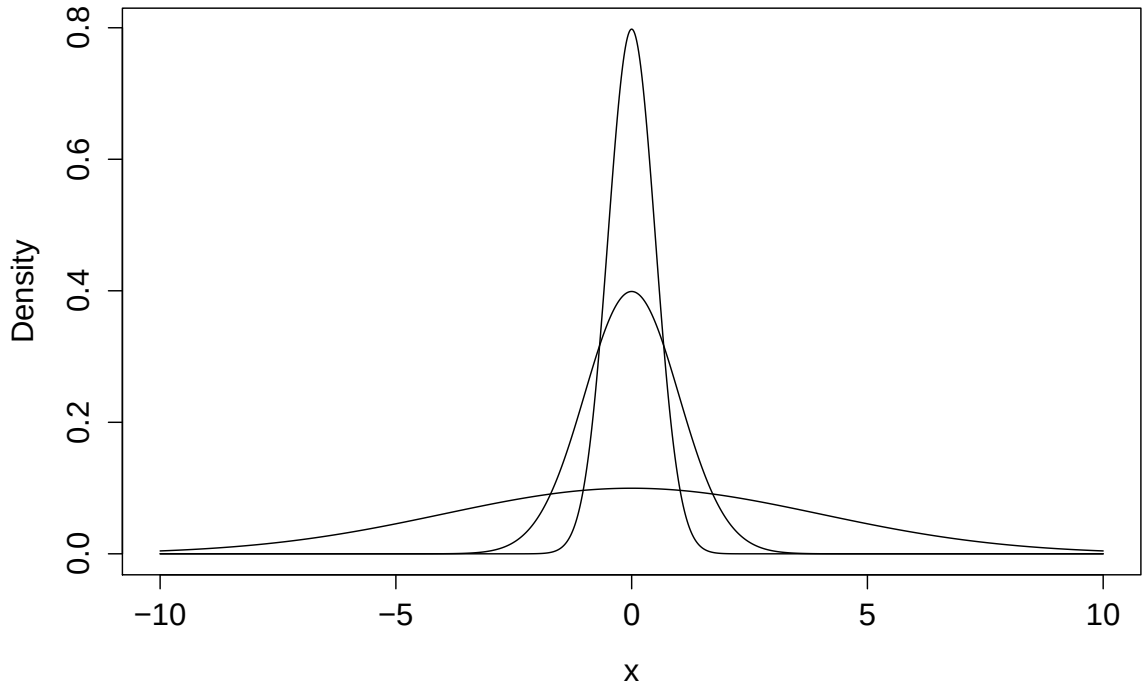
Observations:

Changing μ



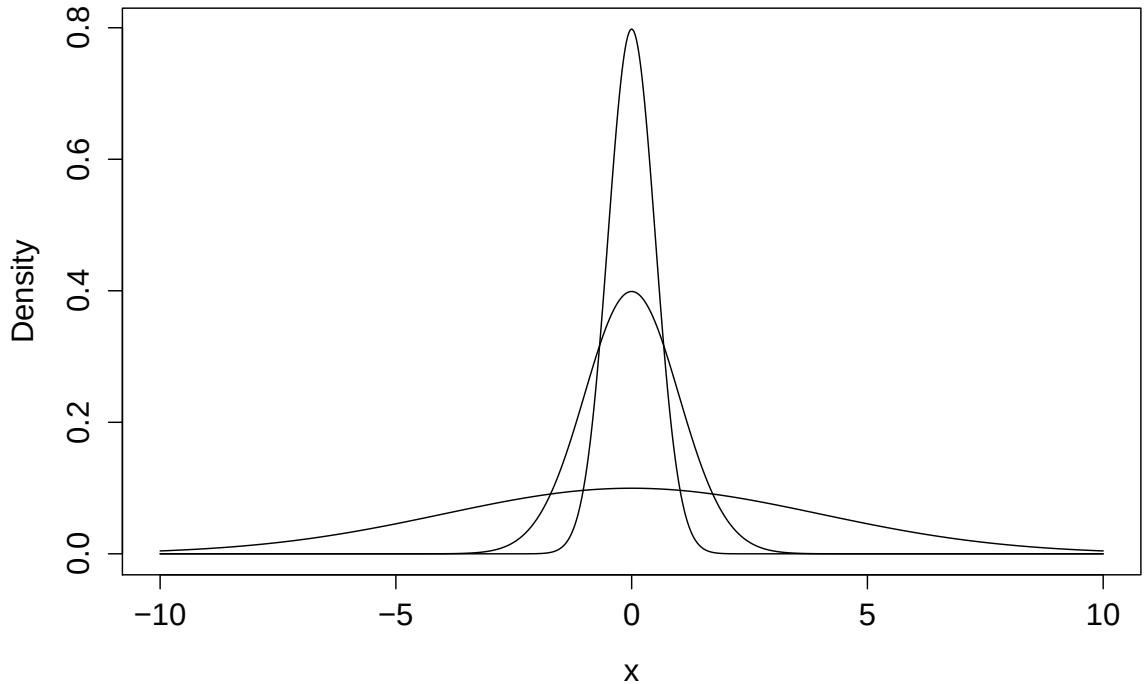
Observations: μ shifts the normal curve.

Changing σ



Observations:

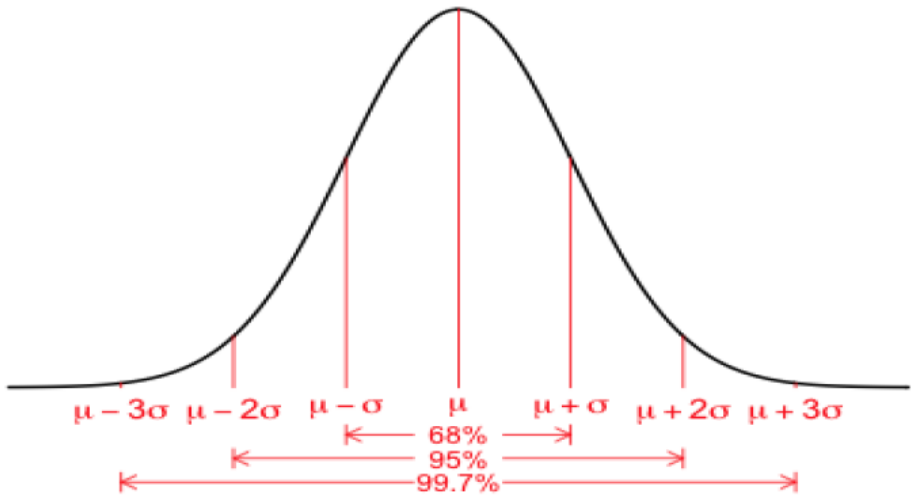
Changing σ



Observations: σ determines the spread of the normal curve. The larger σ is, the more flat (spread out) the normal curve.

The 68-95-99.7 Rule

All normal densities share common properties. One of these is the **68-95-99.7 Rule**



For any normal distribution

- Approximately 68% of the distribution falls within 1 σ of μ
- Approximately 95% of the distribution falls within 2 σ of μ
- Approximately 99.7% of the distribution falls within 3 σ of μ

Example 6.6: Applying the 68-95-99.7 Rule

The time a customer has to wait for their food to arrive at a local restaurant has a normal distribution with mean of 16 minutes and standard deviation of 4 minutes.

(a) What proportion of customers wait longer than 16 minutes?

(b) What proportion of customers wait between 12 and 20 minutes?

(c) What is the probability that a customer waits between 12 and 24 minutes?

(d) The shortest 2.5% of waiting times are smaller than what value?

(e) What is the probability that a customer waits exactly 20 minutes?

(f) Can we compute the probability that a customer waits less than 21 minutes using the 68-95-99.7 Rule?

Example 6.6: Applying the 68-95-99.7 Rule

(a) What proportion of customers wait longer than 16 minutes?
$P(X > 16) = 1/2$ since the median $= \mu = 16$

(b) What proportion of customers wait between 12 and 20 minutes?
$P(12 \leq X \leq 20) = P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$

(c) What is the probability that a customer waits between 12 and 24 minutes?

$$\begin{aligned} P(12 \leq X \leq 24) &= P(\mu - \sigma \leq X \leq \mu + 2\sigma) \\ &= P(\mu - \sigma \leq X \leq \mu + \sigma) + P(\mu + \sigma < X \leq \mu + 2\sigma) \\ &= 0.68 + \frac{0.95 - 0.68}{2} \\ &= 0.815 \end{aligned}$$

Example 6.6: Applying the 68-95-99.7 Rule

(d) The shortest 2.5% of waiting times are smaller than what value?
We know $P(X \leq \mu - 2\sigma) = 0.025$, so the shortest 2.5% of waiting times are smaller than $8 = \mu - 2\sigma$.

(e) What is the probability that a customer waits exactly 20 minutes?
$P(X = 20) = 0$ because X is a continuous random variable.

(f) Can we compute the probability that a customer waits less than 21 minutes using the 68-95-99.7 Rule?
No. Thankfully we have other tools for computing probabilities for the normal distribution.

Normal Distribution Calculations

Goal: Calculate quantities such as $P(X \leq a), P(a < X < b)$, etc using the normal curve.

To accomplish this, we will need to use **Table A** in Appendix A.

Big Note: There is only one table but infinitely many normal distributions (because there are infinitely many combinations of μ and σ). To use the table, we convert every normal distribution to the **Standard Normal** distribution.

Standardizing Normal Random Variables

Consider **any** μ and σ. Suppose $X \sim N(\mu, \sigma)$. Then we can transform X to the standard normal scale:

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

Standardizing Observations

z-score: Let x be an observation from a normal distribution with mean μ and standard deviation σ. Then

$$z = \frac{x - \mu}{\sigma}$$

is the **z-score** of x.

Interpretation:

- By definition: z = number of standard deviations (σ) that x is away from the mean (μ).
- Almost all observations will be within 3σ of μ. Therefore, any observation with a z-score close to or more than 3 is an unusual observation (it falls far from μ).

Computation after Standardization

Suppose $X \sim N(\mu, \sigma)$ and we want to calculate $P(X < a)$ for some value a. Then, standardization guarantees that

$$P(X < a) = P\left(\frac{X - \mu}{\sigma} < \frac{a - \mu}{\sigma}\right) = P(Z < z^*)$$

where $z^* = \frac{a-\mu}{\sigma}$ is the z-score for a. Therefore, after standardizing any normal random variable, we can use the standard normal distribution for probability calculations.

Table A

Table A can be used to calculate areas under the curve for the standard normal distribution, $N(0, 1)$. Specifically, Table A allows us to approximate $P(Z < z^*)$ for any value z^* where $Z \sim N(0, 1)$.

Table A (page A-1):

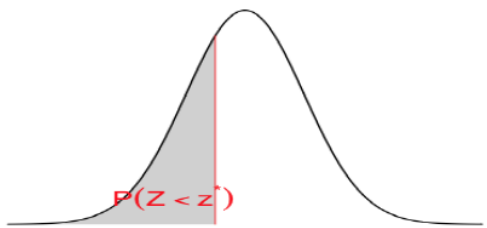
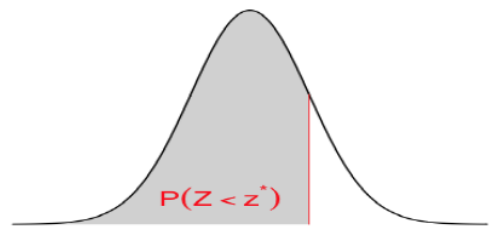


Table A (page A-2):



We can now approximate **any** area beneath the normal curve corresponding to **any** normal distribution.

Example 6.7: Computation with Table A

Let Z be a standard normal random variable.

(a) What is the probability Z falls below -2.63?

(b) What is the probability Z is at least 2.63?

(c) What is the probability Z is greater than -1.31?

(d) Find the proportion of the distribution that falls between -.97 and 1.31.

(e) What value marks the 57th percentile?

Drawing a picture of the density can help us understand the probability we are trying to calculate.

Example 6.7: Computation with Table A

Let Z be a standard normal random variable.

(a) What is the probability Z falls below -2.63?
$P(Z < -2.63) = .0043$

(b) What is the probability Z is at least 2.63?
$P(Z \geq 2.63) = 1 - P(Z < 2.63) = 1 - .9957 = .0043.$
Notice that $P(Z \geq 2.63) = (Z < -2.63)$ (by symmetry).

(c) What is the probability Z is greater than -1.31?
$P(Z > -1.31) = 1 - P(Z \leq -1.31) = 1 - .0951 = .9049.$
or $P(Z > -1.31) = (Z \leq 1.31) = .9049.$

Example 6.7: Computation with Table A

Let Z be a standard normal random variable.

(d) Find the proportion of the distribution that falls between -.97 and 1.31.
$P(-.97 < Z < 1.31) = P(Z < 1.31) - P(Z < -.97) =$
$.9049 - .1660 = .7389.$

(e) What value marks the 57th percentile?
Find a such that $P(Z < a) = .57$. From Table A we get
$P(Z < .18) = .57$, so $a = .18$.

Steps for Calculating Normal Probabilities:

When given a value x and asked to find a probability or proportion, there are 3 steps to follow:

1. Draw a picture.
2. Convert x to a z-score using $z = \dfrac{x - \mu}{\sigma}$.
3. Use Table A to find $P(Z \leq z)$ where $Z \sim N(0, 1)$.

Steps for Calculating x Values for Normal Distributions

When given a probability or proportion and asked for the corresponding x value:

1. Draw a picture.

2. Use Table A to find the z-score for the specified probability

3. Unstandardize the z-score. Since $z = \dfrac{x - \mu}{\sigma}$, we see that $x = \mu + z\sigma$.

Example 6.8: More General Computations

Let X = SAT score and suppose that SAT scores are known to follow a normal distribution with mean $\mu = 1026$ and standard deviation $\sigma = 209$.

(a) Write down the distribution of X.

(b) Calculate and interpret the z-score for an SAT score of 1100.

(c) What proportion of students score lower than 1100?

(d) What is the probability that a randomly selected student received a score of at least 820?

(e) What score must you earn to be in the top 10%?

Example 6.8: More General Computations

(a) Write down the distribution of X.
$X \sim N(1026, 209)$

(b) Calculate and interpret the z-score for an SAT score of 1100.
$z = \frac{x-\mu}{\sigma} = \frac{1100-1026}{209} = 0.35$. This means that an SAT score of 1100 is 0.35 standard deviations above the mean.

(c) What proportion of students score lower than 1100?
$P(X \leq 1100) = P(\frac{X-\mu}{\sigma} \leq \frac{1100-\mu}{\sigma}) = P(Z \leq 0.35) = 0.6368$

(d) What is the probability that a randomly selected student received a score of at least 820?
$P(X > 820) = 1 - P(X \leq 820) = 1 - P(Z \leq \frac{820-1026}{209}) =$
$1 - P(Z \leq -0.99) = 1 - .1611 = 0.8389.$

(e) What score must you earn to be in the top 10%?
A z-score of $z = 1.28$ puts you in the top 10%. This corresponds to an SAT score of $x = z\sigma + \mu = 1.28(209) + 1026 = 1293.5 \approx 1294$.

Assessing Normality

Not every continuous random variable follows a Normal Distribution. Moreover, there is no way to *guarantee* that a random variable is normally distributed.

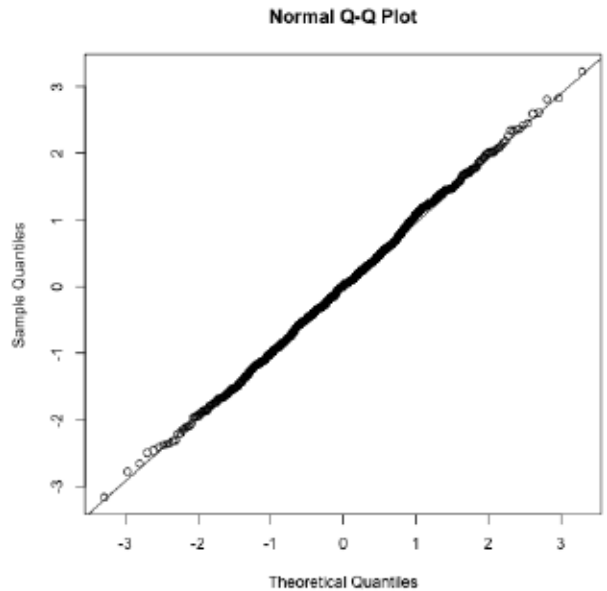
However, using graphs we can assess whether it is reasonable to *assume* that a random variable follows a normal distribution.

Assessment Tools:

- Histogram/Box plot: Check for skewness and outliers. We know normal distributions are symmetric.
- Q-Q plot: Check for skewness, outliers, and "heavy-tailedness".

Q-Q Norm Plot

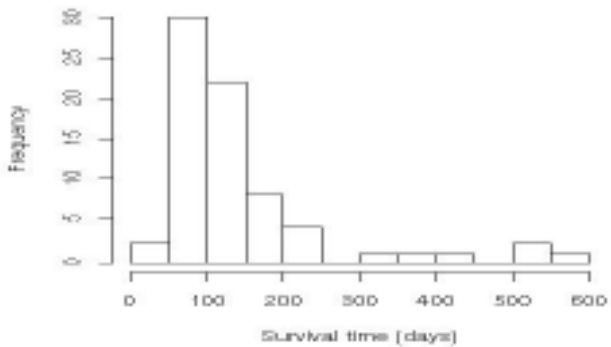
The QQ Plot plots the quantiles of the data ("sample quantiles") against the quantiles of a normal distribution ("theoretical quantiles").
If the points lie along the plotted line we can safely assume the data are normally distributed.

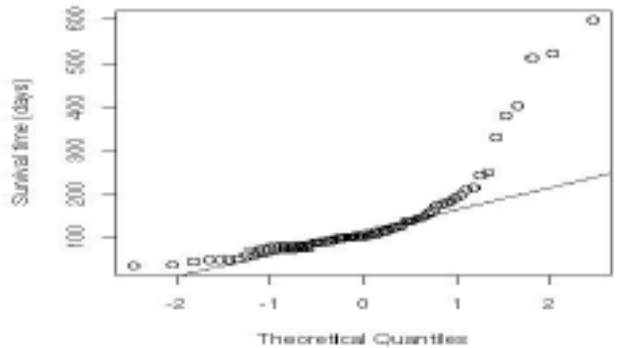



```
> samp <- rnorm(1000)
> qqnorm(samp)
> qqline(samp)
```

Example 6.10

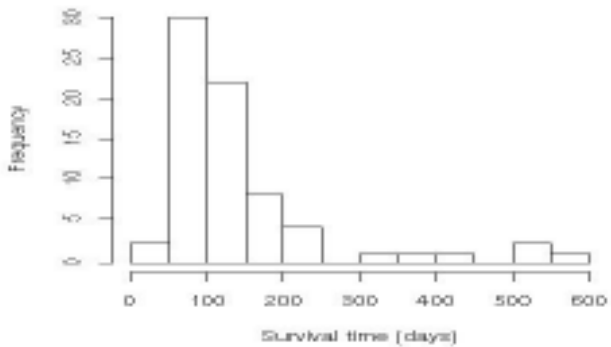
Two examples from unspecified distributions:
We will use histograms and Q-Q plots to assess whether that data are normally distributed and characterize the sample distributions.

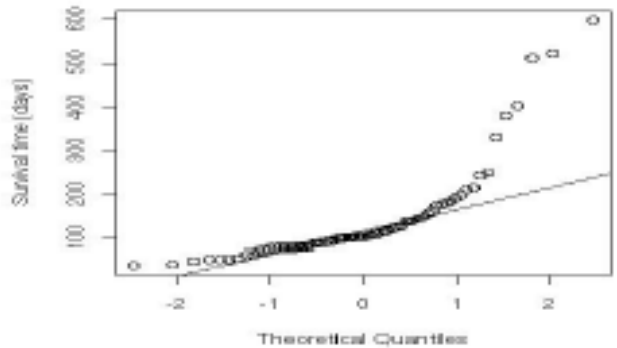




Example 6.10

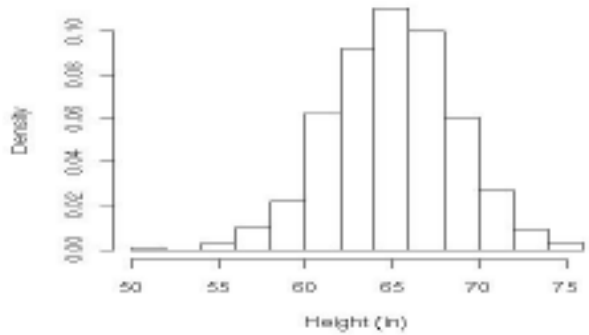
Two examples from unspecified distributions:
We will use histograms and Q-Q plots to assess whether that data are normally distributed and characterize the sample distributions.

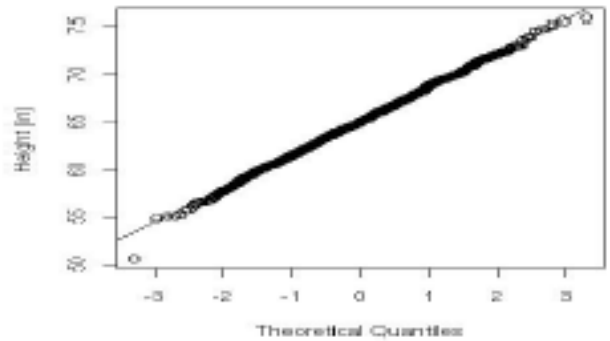




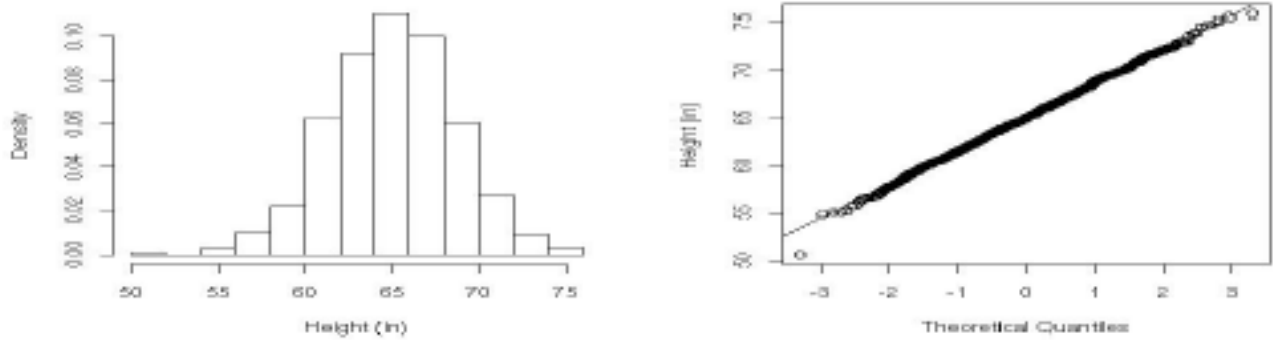
The distribution is right skewed and contains outliers. Therefore, it is likely that the distribution is not normal

Example 6.10


Density
0.00
0.02
0.04
0.06
0.08
0.10
50
55
60
65
70
75
Height (in)


Height (in)
50
55
60
65
70
75
-3
-2
-1
0
1
2
3
Theoretical Quantiles

Example 6.10



With the exception of a single outlier on the lower tail, most of the data points approximately form a straight line in the Q-Q plot. Thus it is reasonable that the data comes from a normal distribution.

The Normal Distribution in R

Assume $Z \sim N(0, 1)$.

(a) Calculate $P(Z \leq 1.645)$

```
> pnorm(1.645)
[1] 0.950015
```

(b) Calculate $P(Z \geq -0.971)$

```
> 1 - pnorm(-0.971)
[1] 0.8342259
> pnorm(-0.971, lower.tail=F)
[1] 0.8342259
```

Note: R calculates lower tail probabilities by default!

(c) Find the 97.5th percentile

```
> qnorm(0.975)
[1] 1.959964
```

The Normal Distribution in R

Assume $X \sim N(21, 2)$.

(a) Calculate $P(X < 19)$

```
> pnorm(19, mean=21, sd=2)
[1] 0.1586553
```

(b) Calculate $P(17 < X < 25)$

```
> pnorm(25, mean=21, sd=2) - pnorm(17, mean=21, sd=2)
[1] 0.9544997
```

(c) Find the 60th percentile

```
> qnorm(.60, mean=21, sd=2)
[1] 21.50669
```

Generating Normal Data

```
#random sample of 100 observations (n=100) from N(0,1)
> x <- rnorm(100)

#random sample of 10 observations (n=10) from N(90,5)
> y1 <- rnorm(10,mean=90,sd=5)

#random sample of 100 observations (n=100) from N(90,5)
> y2 <- rnorm(100,mean=90,sd=5)

#random sample of 1000 observations (n=1000) from N(90,5)
> y3 <- rnorm(1000,mean=90,sd=5)

> hist(y1)
> hist(y2)
> hist(y3)
```