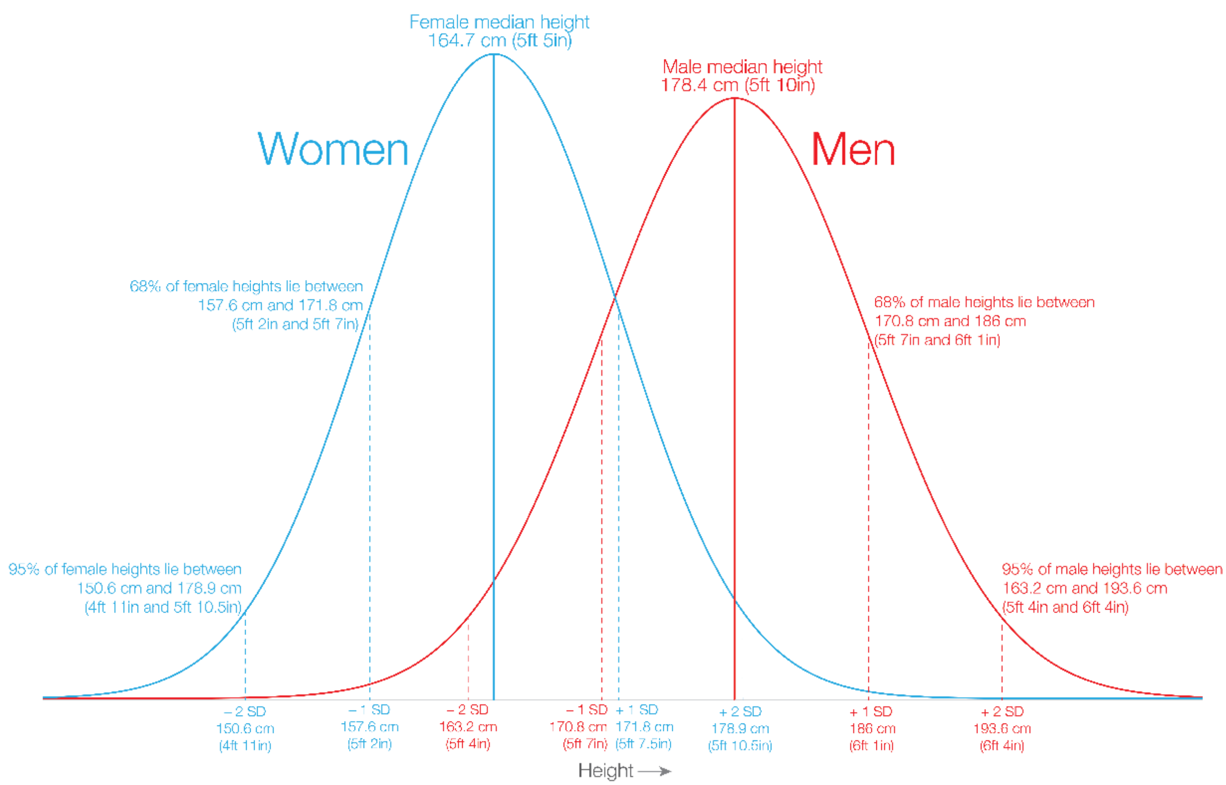
Unit 02 – Normal Distributions and Z-scores

Distributions of Normal Variables

In Unit 01 you learned to visualize the *shape* of a distribution of a random variable, X, using a *histogram.* In statistics, the most important distribution is the *Normal Distribution* with its familiar "bell" shape.

Example (*ourworldindata.org*) The worldwide distributions of $X =$ Height for Women and $Y =$ Height for Men are given by the following graphs.



Each of these curves (blue and red) is called a *probability density functions* or *pdf*. For a probability density function, the total area under the graph must equal 1 (representing 100%).

If you have a pdf for X, then the proportion of individuals in the population for which $a \leq X \leq b$ is the *area between* a and b. In other words,

$$P(a \leq X \leq b) =$$

Example Using the above pdf for Women's height, X, shade in the probability $P(164.7 \leq X \leq 171.8)$ and estimate its value "by eye".

All normal distributions have the same basic shape, but they differ depending on the mean μ and standard deviation σ of the variable, X, that they describe.

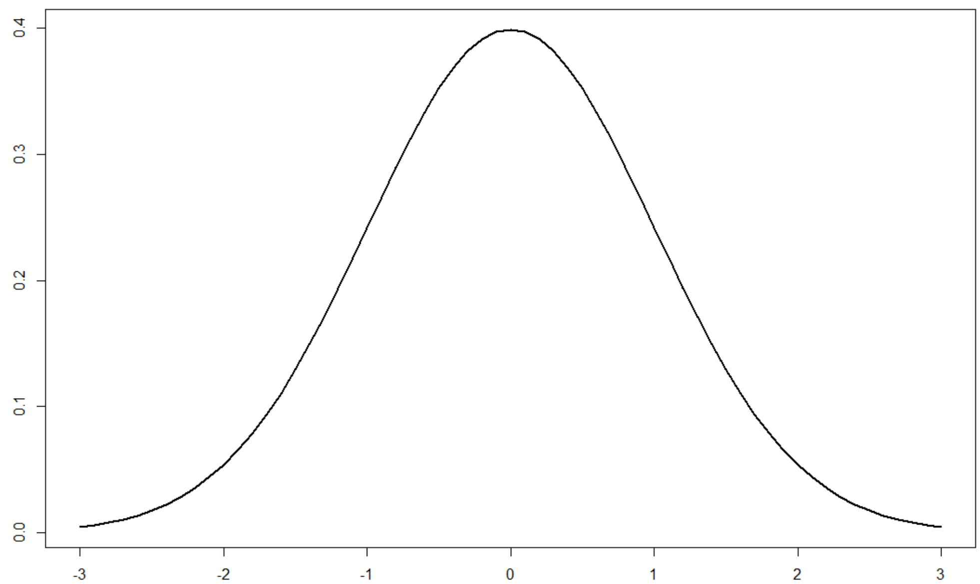
If X is a normally distributed variable with mean μ and standard deviation σ, then we write:

$$X \sim N(\mu, \sigma^2)$$

Standard Normal Distribution

For a *standard* normal distribution, we have: $\mu =$
$\sigma =$

The pdf of the *standard normal* distribution is given by the formula $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$.



Example Suppose $Z \sim N(0, 1)$. Find the probability $P(0 \leq Z \leq 1)$ in three ways:

a) By approximating the area with a single rectangle (based on the pdf).

b) Using the "Standard Normal Cumulative Probability Table".

c) Using the RStudio function `pnorm()`.

Example Suppose Z is a standard normal variable. Calculate the probabilities:

$P(Z \leq 1.96)$

$P(-1.96 < Z \leq 1.96)$

$P(Z = 1.96)$

$P(Z < 1.96)$

$P(Z > 1.96)$

Example If $Z \sim N(0, 1)$, find the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentile values of Z.

z-scores

When X is a *non-standard* normal variable (or something even weirder), we will often use the z-*score* or *standard score* to relate X back to the *standard* normal distribution.

Suppose X is a numerical variable. If $X = x$ for some individual, then the *z-score* for that individual is:

$z =$ or $z =$

Example Using the variable $X = \text{Height}$ for all students in MATH 1350 (in Fall 2021)

a. What is the z-score for a student whose height is $x = 165$ cm?

b. What is the height of a student whose z-score is $z = -1.25$?

In general, we can rearrange the above formulae to give

$x =$ or $x =$

If $z = 1$, then

If $z = -2$, then

The z-score indicates how many standard deviations a specific x is away from the mean value.

Parameters of Z (extra video content)

Suppose X is any numerical variable with mean μ and standard deviation σ.
Then the formula for the z-score defines a *new numerical variable.*

$$Z = \frac{X - \mu}{\sigma}$$

What are the mean and standard deviation of Z? We can answer this with a formal algebraic calculation.

$\mu_Z =$

$\sigma_Z =$

The pdf for Z therefore looks like:

To summarize: if X is a non-standard normal variable with mean μ and standard deviation σ, then the z-score defined by $Z = \frac{X-\mu}{\sigma}$ is a *standard* normal variable. Think of Z as a "scale model" of the original X.

Example Using the variable $X =$ Height for all students in MATH 1350:

a. Generate a histogram of X and a histogram of Z (the z-scores of X).

b. What proportion of students have z-score $z > 0$?

c. What proportion of students have a z-score $-1.0 < z < 1.0$?

d. What proportion of students have a z-score $-2.0 < z < 2.0$?

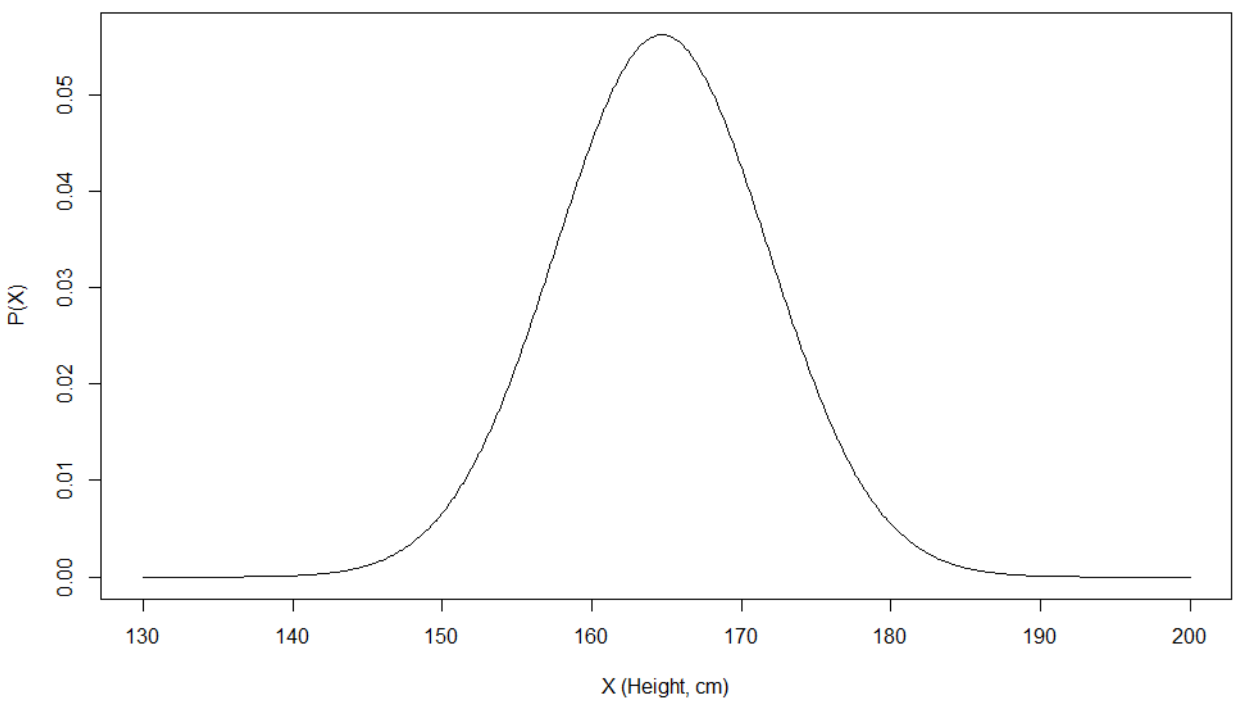
e. What is the sum of all the z-scores?

Non-Standard Normal Distributions

In the real world, most variables are *not* standard normal variables. In this section, we assume that:

- X is normally distributed
- X has mean $\mu \neq 0$
- X has standard deviation $\sigma \neq 1$

Example (Human Heights) According to *Our World in Data*, the height, X, of females around the world is approximately normally distributed with mean $\mu = 164.7 \text{ cm}$ and standard deviation $\sigma = 7.1 \text{ cm}$.



a) Calculate $P(170.0 \leq X \leq 180.0 \text{ cm})$.

b) Calculate the probability that a woman's height X is between 164.5 cm and 165.5 cm (which means it equals 165 cm when rounded to the nearest centimeter).

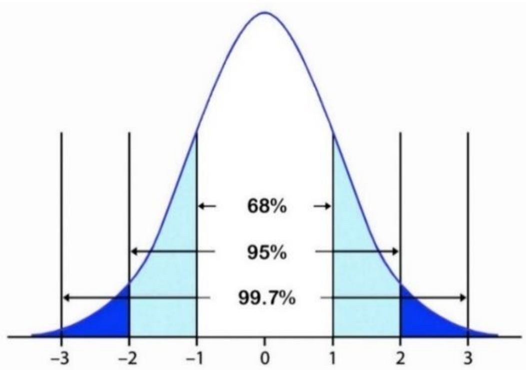
c) Calculate $P(X = 165.000\text{ cm})$

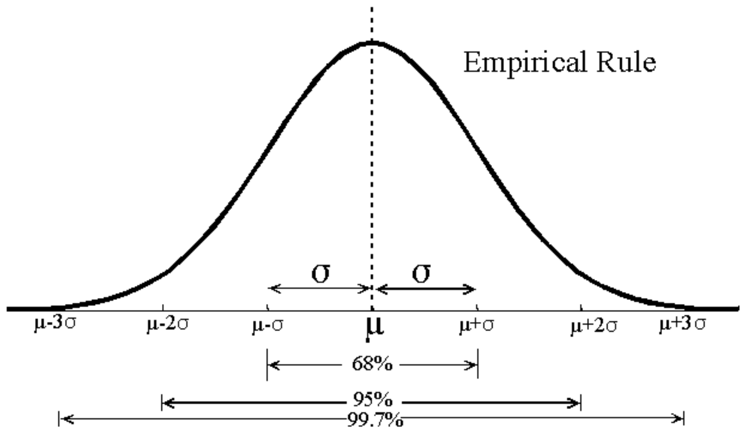
d) Calculate the height (i.e., y coordinate) of the probability density function of X at $x = 165.0$.

e) Find the 25th and 75th percentile of female heights.

Empirical Rule for Normal Distributions

If X is a *normal* variable, we can say exactly how many individuals lie within certain z-score ranges.





The following three facts are called the *Empirical Rule*:

- 68% of all individuals will have a standard z-score between -1 and +1

  In other words, $P(\mu - \sigma \leq X \leq \mu + \sigma) \approx$

- 95% of all individuals will have a standard z-score between -2 and +2

  In other words, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx$

- 99.7% of all individuals will have a standard z-score between -3 and +3.

  In other words, $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx$

Example What interval of heights contains 95% of all women worldwide?

Usual and Unusual Values

For normal (i.e., bell-shaped) distributions, we say that *ordinary* values fall within two standard deviations of the mean and *unusual* values are more than 2 standard deviations from the mean.

Note: "unusual" is not the same thing as "outlier"!

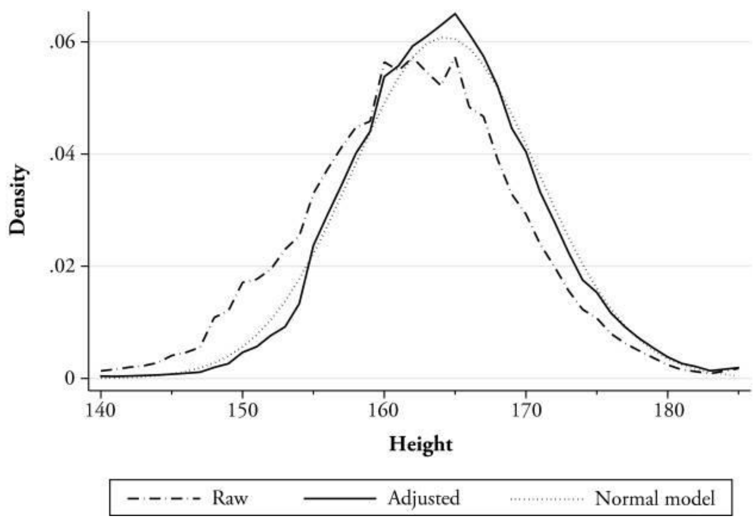
Usual values: $-2 \leq Z \leq +2$

Unusual values: $Z < -2 \text{ or } Z > 2$

For a normal distribution, the percentage of unusual values is ________%.

Example Suppose $X = \text{Height}$ where the population is all soldiers in the Italian military who were born in 1900 (this data set was collected during the First World War). The parameters for X are:

$\mu = 164.3 \text{ cm}$
$\sigma = 6.56 \text{ cm}$



a) Based on the pdf curve, what fraction of soldiers were between 170.0 cm and 172.0 cm?

b) What percentage of soldiers were greater than 170.86 cm in height?

c) Approximately what range of heights contains 99.7% of all soldiers' heights?

d) What range of heights is considered *unusual* for this population of soldiers?

e) Challenging: what height x would be the *upper limit* for *outliers* in this population? What percentage of soldiers are above this limit?

Example Sketch a normal distribution with mean 100 and standard deviation 15. (This is the distribution of IQ scores across a large population of individuals.) What range of IQ scores is *unusually* high?

Chebyshev's Theorem

The *Empirical Rule* describes the spread of any *normally distributed* variable, X, in terms of σ. (The Empirical Rule only applies to normal variables.)

Chebyshev's Theorem is a more general fact about numerical variables (not just *normal* variables)

According to Chebyshev's Theorem, the fraction of a population for which X is within k standard deviations of the mean is always at least $1 - \frac{1}{k^2}$.

Example Chebyshev's Theorm implies that the fraction of a population whose height is within $k = 2$ standard deviations of the mean is at least $1 - \frac{1}{2^2} = 0.25 = 25\%$.

Example Using Chebyshev's theorem, what is the smallest possible fraction of men's heights worldwide that are in the range 163.2 cm to 193.6 cm?

Example For the variable $X = \text{Age}$ for students in MATH 1350, what proportion of students' lie within $k = 3$ standard deviations of the mean age?

What would Chebyshev's Theorem tell us for $k = 3$?

What would Chebyshev's Theorem tell us for $k = 1$?