# SOC 3811/5811:
# BASIC SOCIAL STATISTICS

## Continuous Random Variables

# Random Variables

**Discrete Random Variable**

Can only equal a finite number of distinct values

*Example*: When you flip a coin 3 times, you can only get four possible values … the whole numbers 0 through 3

**Continuous Random Variable**

Can take any numeric value within a range of values

*Example*: The number of miles from campus students live can take on just about any value (0.12, 1.17, 2.00, etc.)

*Note*: Discrete random variables with lots of values (e.g., number of Facebook friends) are often treated as continuous; continuous variables that are rounded (e.g., age) may seem discrete

# Random Variables

**Today**

We <u>know</u> the actual probabilities associated with the random event that generates the theoretical distribution (e.g., coin flips)

We will learn to *describe* and *make practical use* of these theoretical distributions

**Later (and in Life)**

We <u>do not know</u> the actual probabilities associated with the random event (e.g., number of children per person, which candidate will win)

Our ability to *describe* and *make practical use* of theoretical distributions will allow us to <u>infer</u> those probabilities

# Continuous Random Variables

For discrete random variables we began by computing the probability of observing each possible outcome

We can’t do this for continuous random variables because there are (by definition) an infinite number of possible outcomes

Instead of determining the probability that *Y* equals particular values and specifying a probability distribution function, we determine the probability that *Y* falls within a certain range of the probability density function

# Continuous Random Variables

For the discrete random variable *Y*, the probability distribution function reports the probability of observing each possible value of *Y*

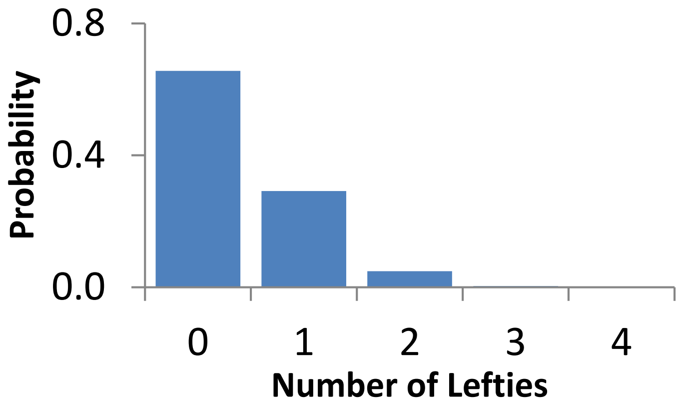
For a continuous random variable, the probability density function is a curve that provides information about the probability that *Y* falls between two values *a* and *b*

$P(a \leq Y \leq b)$ is the area under the curve over the interval between the values a and b

# Continuous Random Variables

**Probability Distribution Function**

Number Left Handed
Out of Four Babies

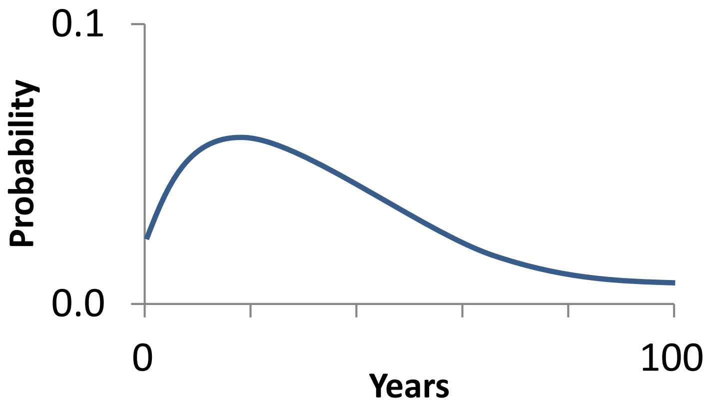



*Example*: P(Y=1) = 0.2916

*"The probability that Y equals 1 is 0.2916"*

**Probability Density Function**

Years from Birth
until Death



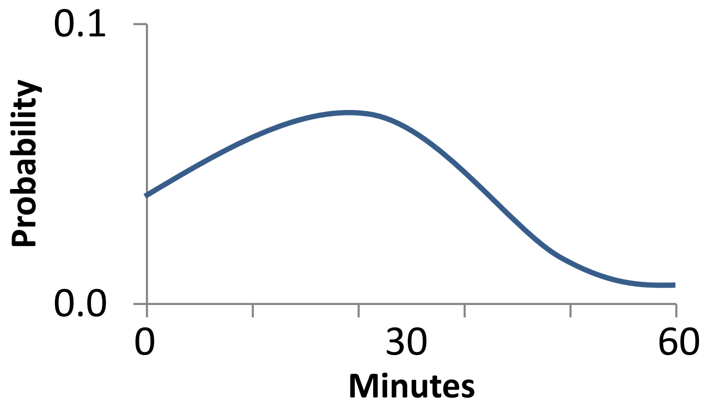

*Example*: P(20≤Y≤40) = 0.40

*"The probability that you live between 20 and 40 years is 0.40"*

# Continuous Random Variables

**Probability Density Function**

Time Waiting for your Pizza

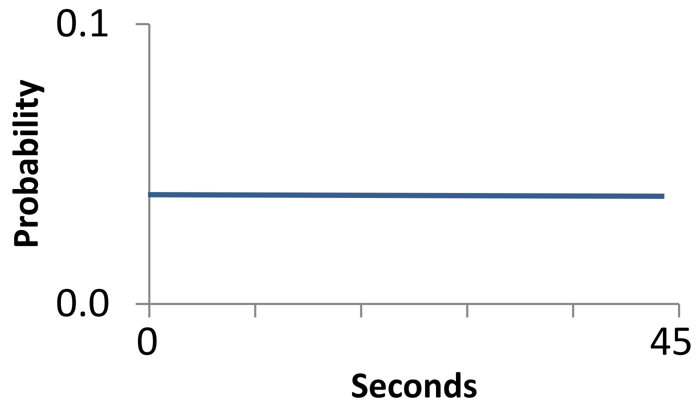



*Example*: P(30≤Y) = 0.70

*"The probability that you wait 30 minutes or longer is 0.70"*

**Probability Density Function**

Time Waiting at a Stop Light




*Example*: P(0≤Y≤45) = 1.00

*"The probability that you wait between 0 and 45 seconds is 1.00"*

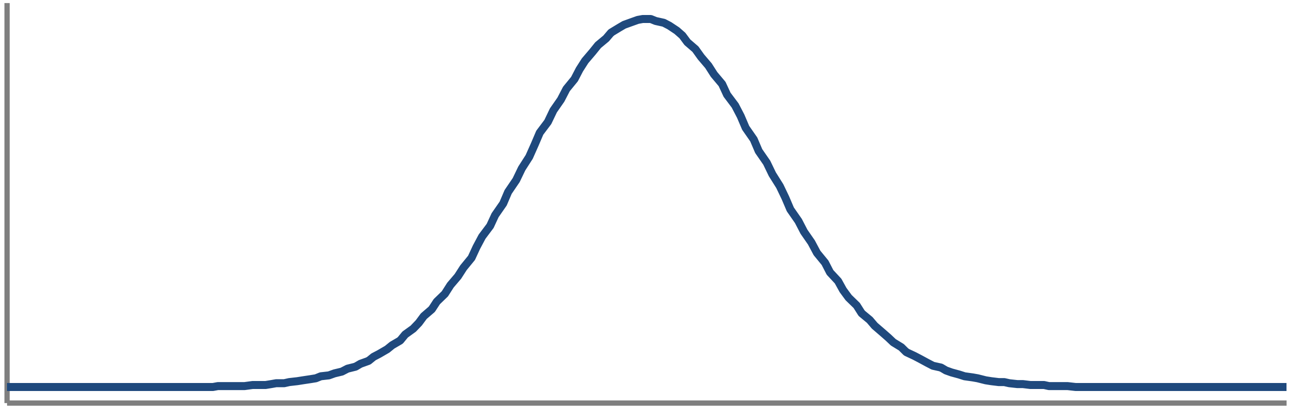
# Continuous Random Variables

As before, these are theoretical distributions … distributions of Y if we were to sample from the population an infinite number of times

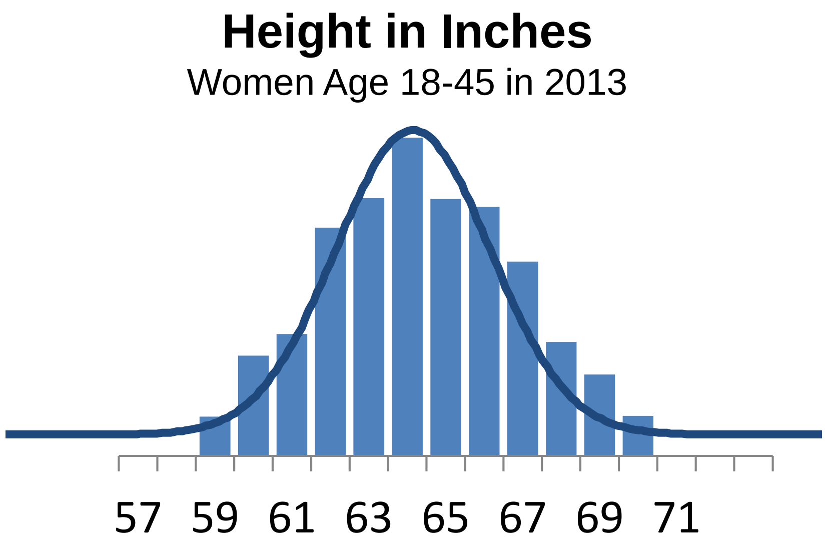
As with all distributions, we can describe distributions with respect to their central tendency and amount of variability

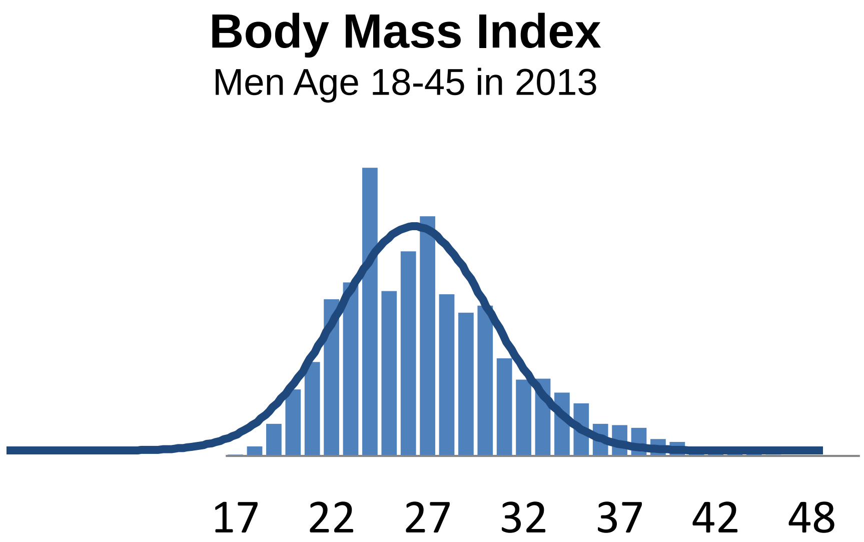
# Normal Random Variables

Just as a binomial random variables is a special (and very common) kind of discrete random variable, a **normal random variables** is a special (and very common) kind of continuous random variable

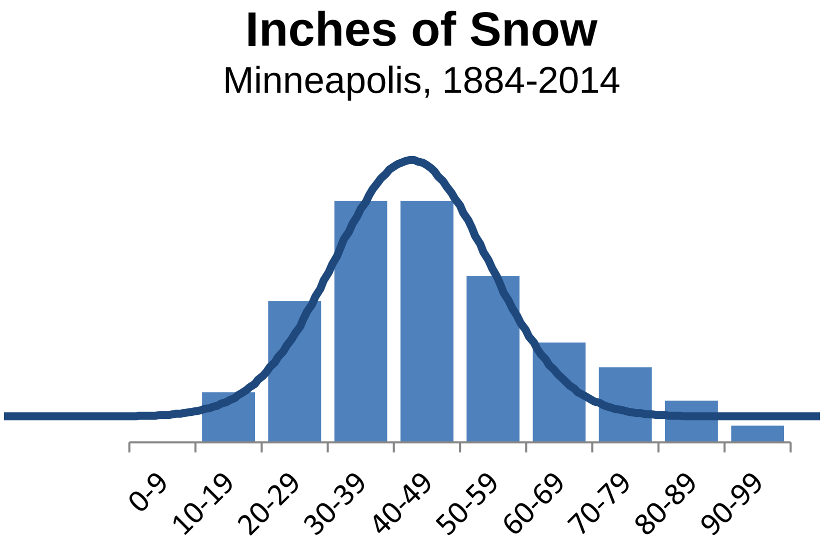


# Normal Random Variables







# Normal Random Variables

Any normal random variable Y is symmetric and can be characterized by its mean $\mu_Y$ and standard deviation $\sigma_Y$

Remember Z scores? $Z = \frac{(Y - \overline{Y})}{s_Y}$

For *any* normally distributed random variable, $\mu$=0 and $\sigma$=1:

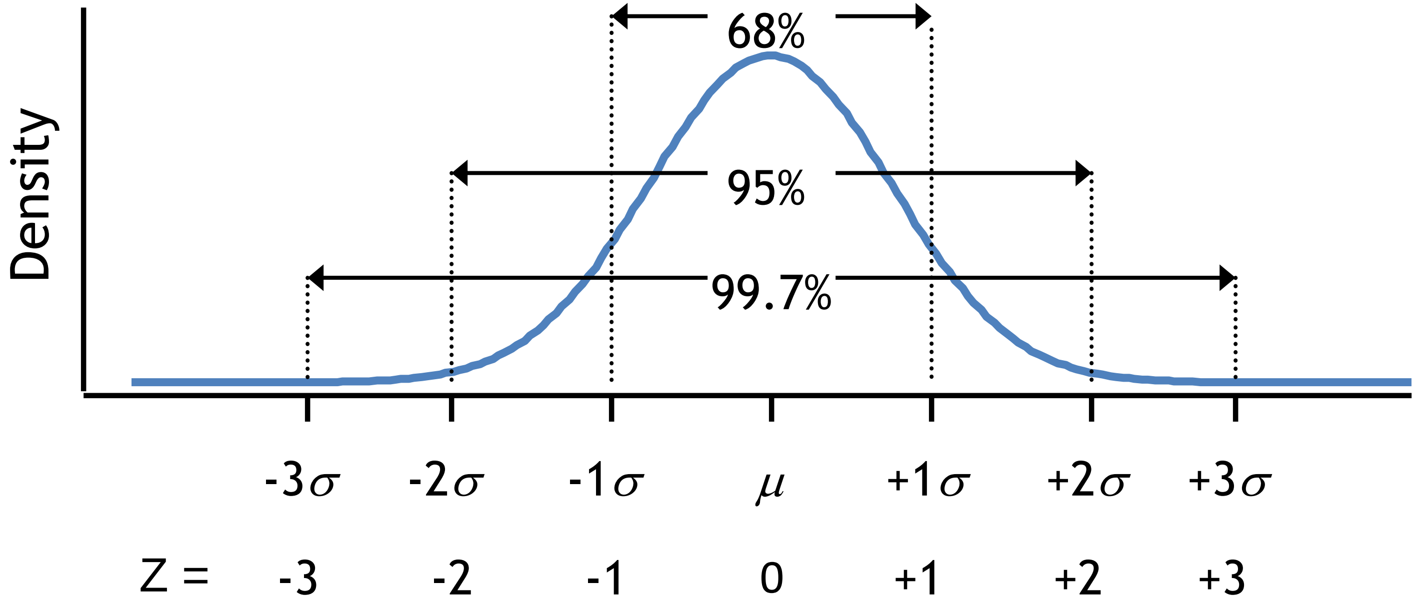
~68% of cases fall within the range -1Z and +1Z

~95% of cases fall within the range -2Z and +2Z

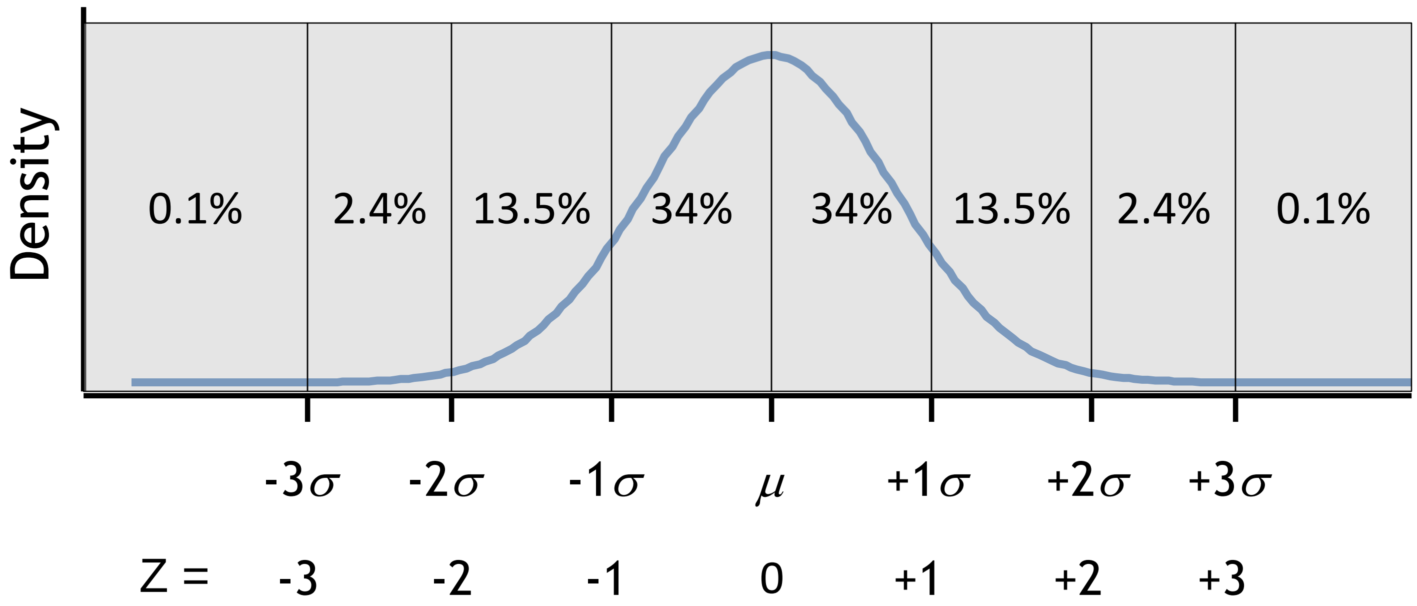
~99.7% of cases fall within the range -3Z and +3Z

100% of cases fall within the range $-\infty$Z and + $\infty$Z

# Normal Random Variables



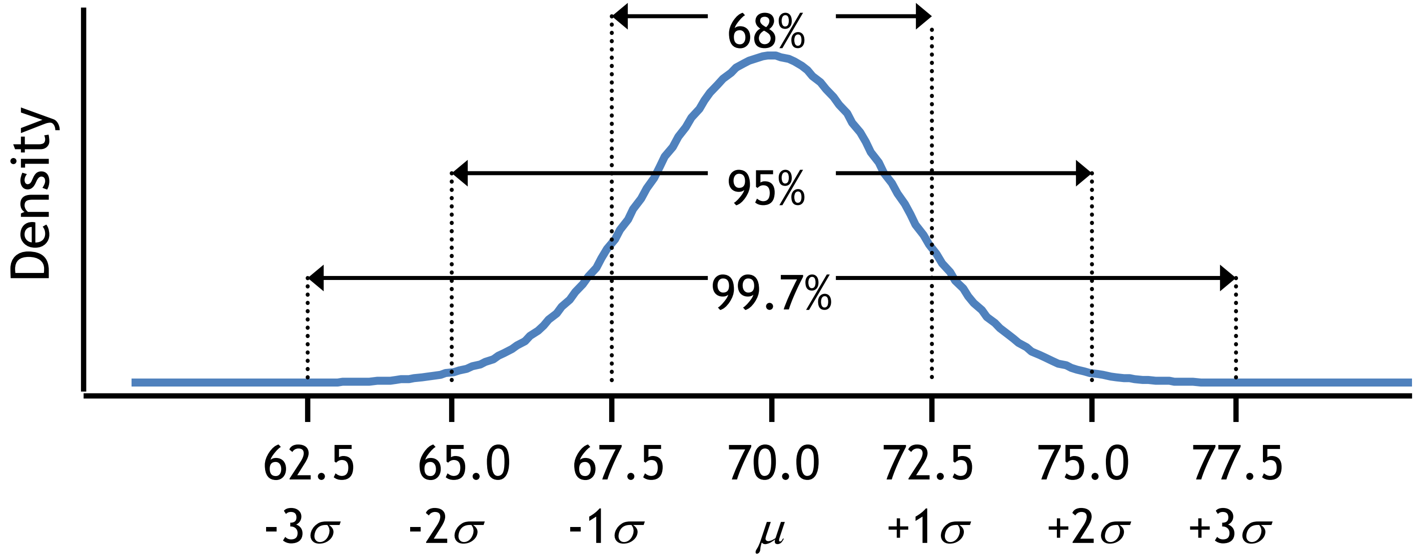
# Normal Random Variables




-3
-2
-1
0
1
2
3

# Normal Random Variables

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5



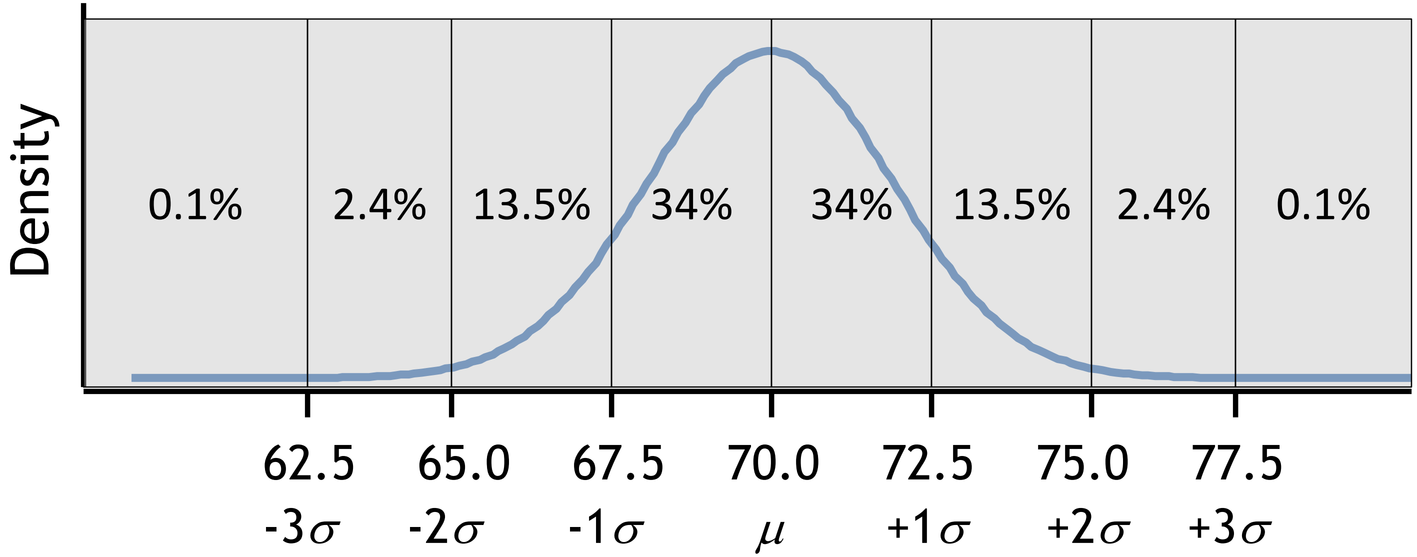

**What proportion of men is more than 75" tall?**

$Z = (Y - \mu_Y)/\sigma_Y = (75-70)/2.5 = 2$

$P(Z \leq 2) = 0.975$ and so $P(Z > 2) = 1 - 0.975 = 0.025$

# Normal Random Variables

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5



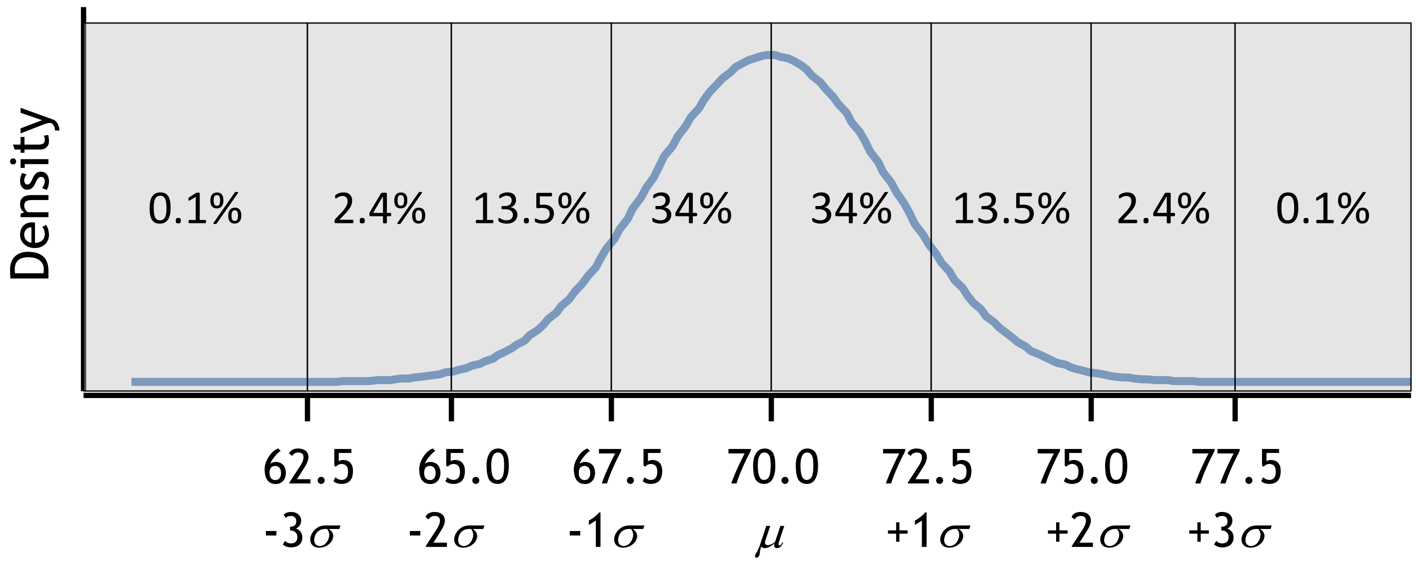

**What proportion of men is more than 75" tall?**

$Z = (Y - \mu_Y)/\sigma_Y = (75-70)/2.5 = 2$

$P(Z \leq 2) = 0.975$ and so $P(Z>2) = 1 - 0.975 = 0.025$

# Normal Random Variables

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5



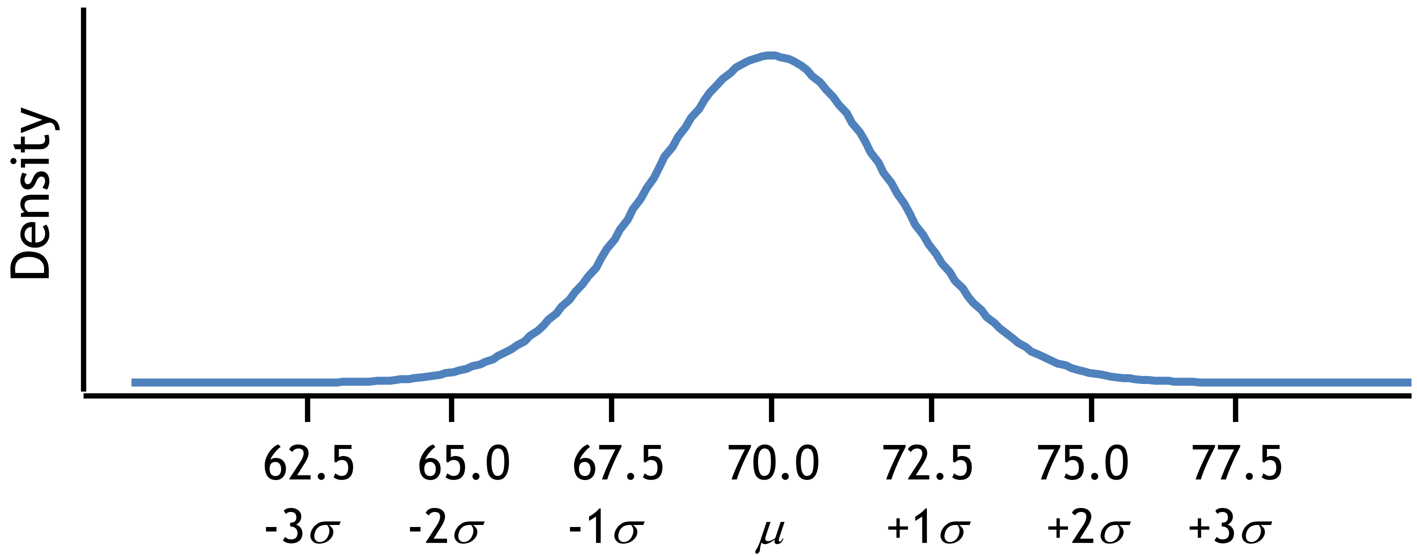

**What proportion of men is between 67.5" and 75" tall?**

$Z_{75}$ = (75–70)/2.5 = 2 ; $Z_{67.5}$ = (67.5–70)/2.5 = -1

P(67.5"≤ Z ≤ 75") = P(Z≤2) — P(Z≤-1) = 0.975 — 0.160 = 0.815

# Normal Random Variables

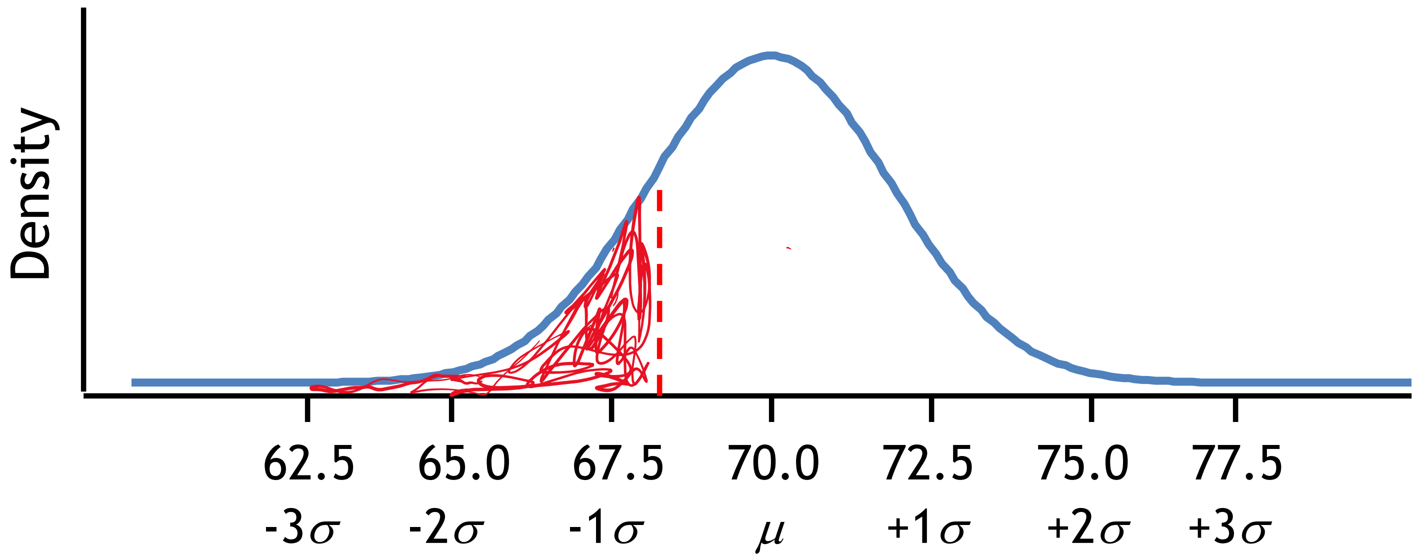
Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5




**What proportion of men is shorter than 68"?**

# Normal Random Variables

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5



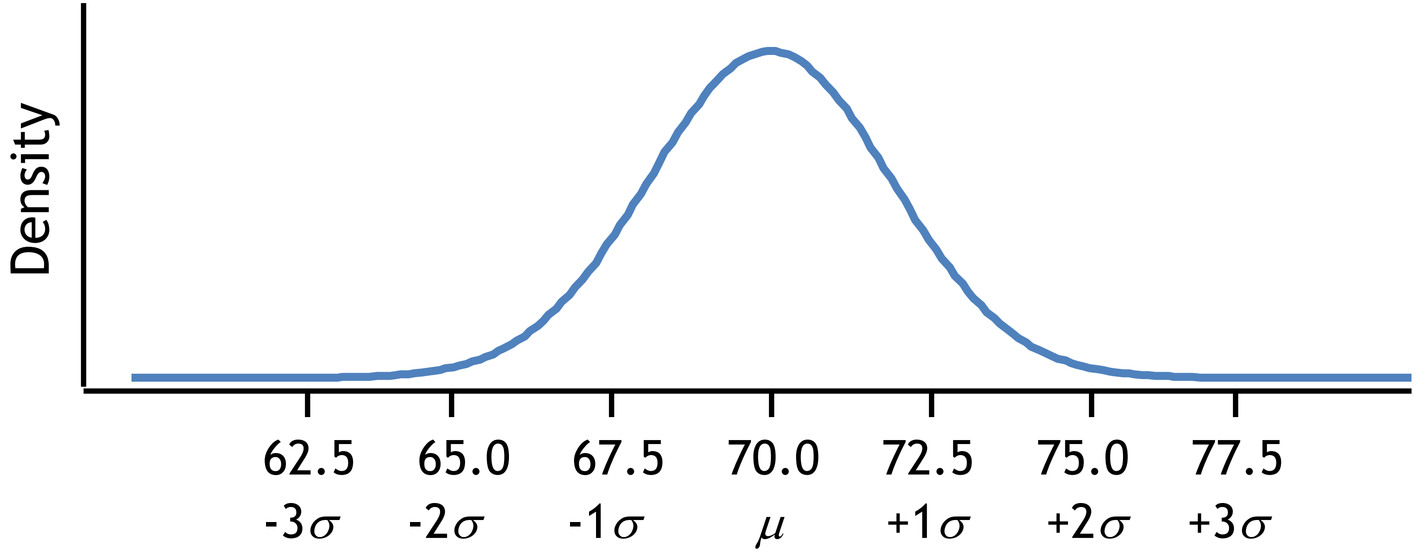

**What proportion of men is shorter than 68"?**

68" corresponds to a Z Score of (68-70)/2.5 = -0.8

Area to the left of -0.8 = P(Z<-0.8) = 0.212 … so 21.2%

# Normal Random Variables

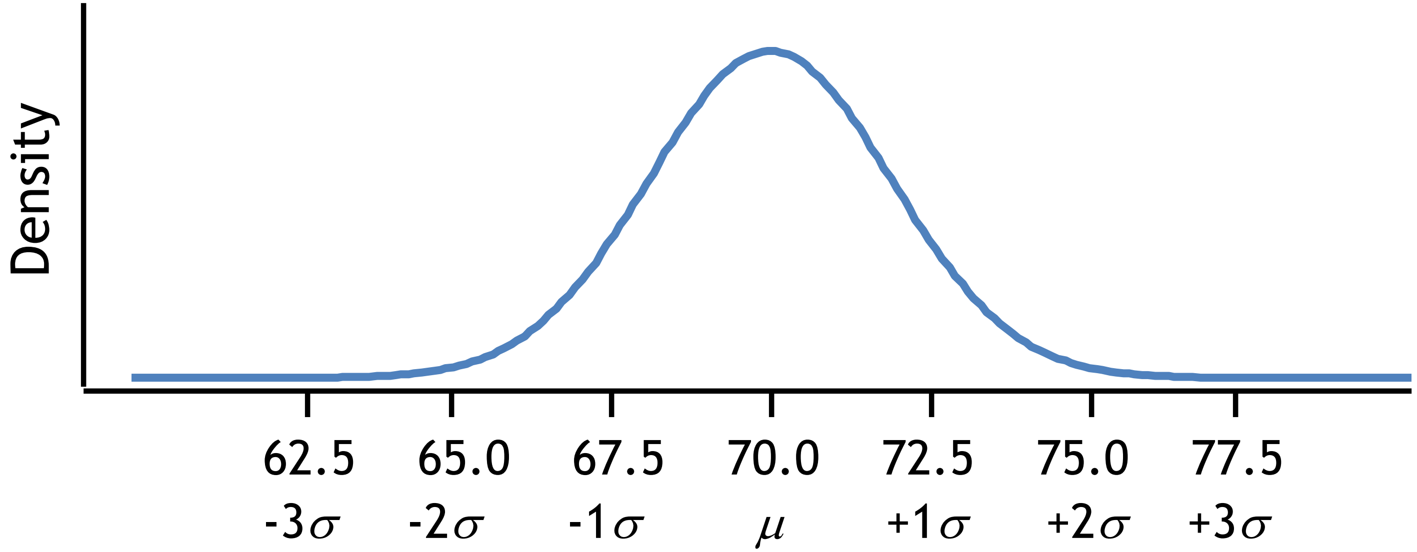
Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5




**What proportion of men is between 66" and 74"?**

# Normal Random Variables

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5



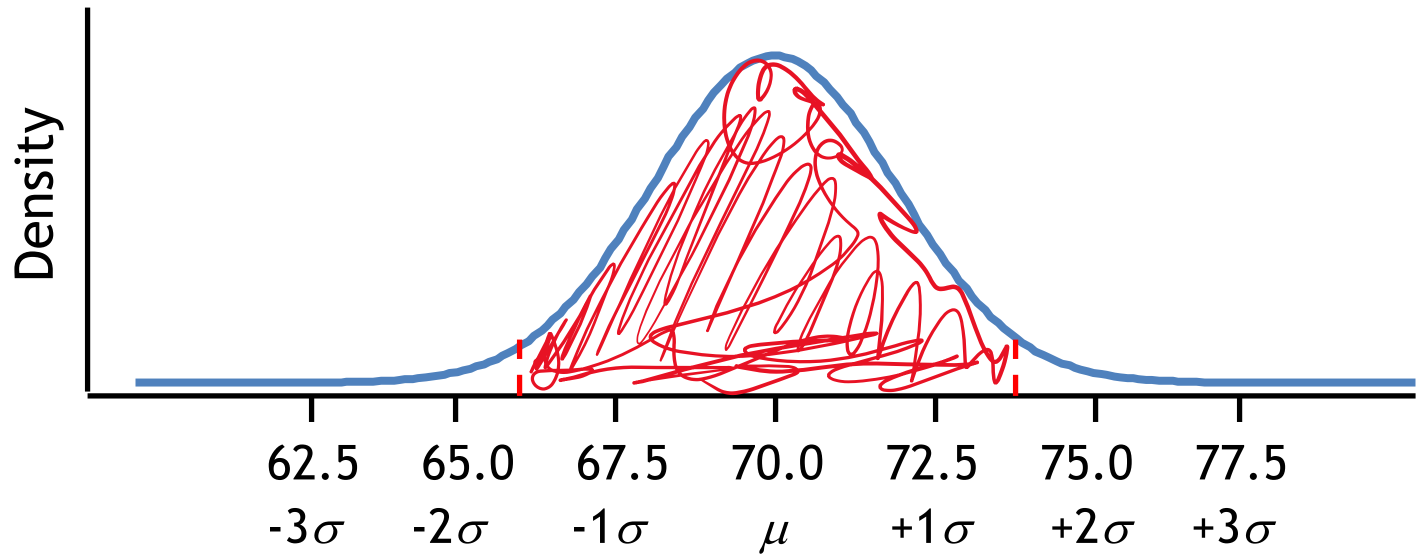

**What proportion of men is between 66" and 74"?**

66" corresponds to a Z Score of (66-70)/2.5 = -1.6

74" corresponds to a Z Score of (74-70)/2.5 = 1.6

# Normal Random Variables

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5



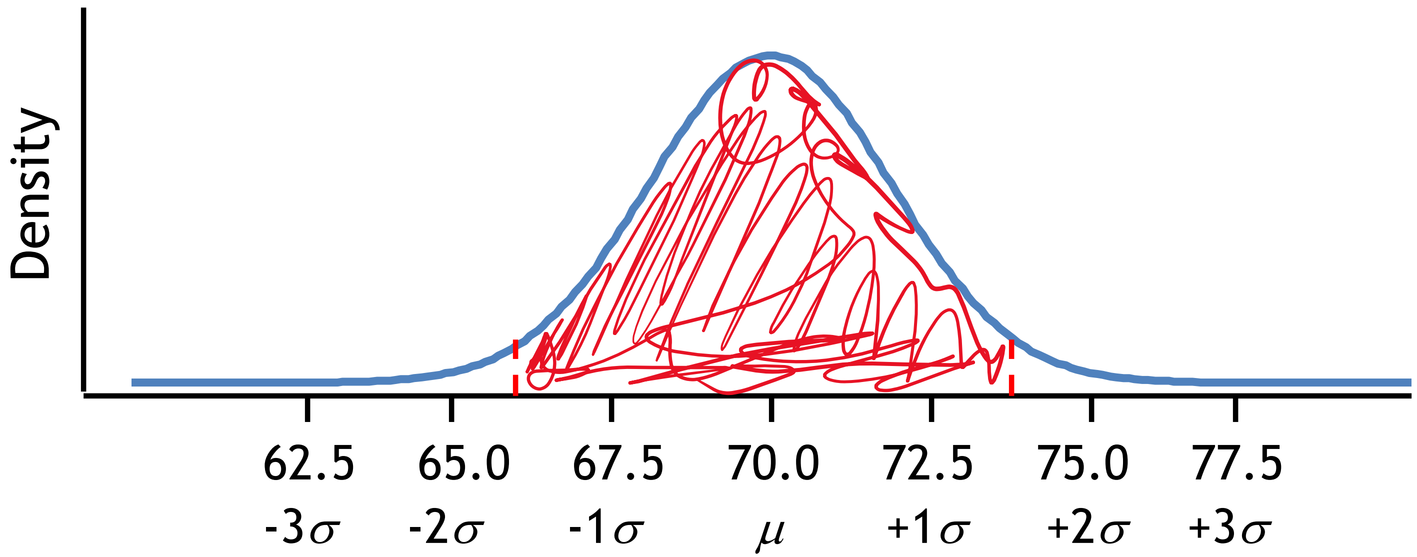

**What proportion of men is between 66" and 74"?**

Area to the left of -1.6 = P(Z<-1.6) = 0.055

Area to the left of 1.6 = P(Z<1.6) = 0.945

# Normal Random Variables

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5

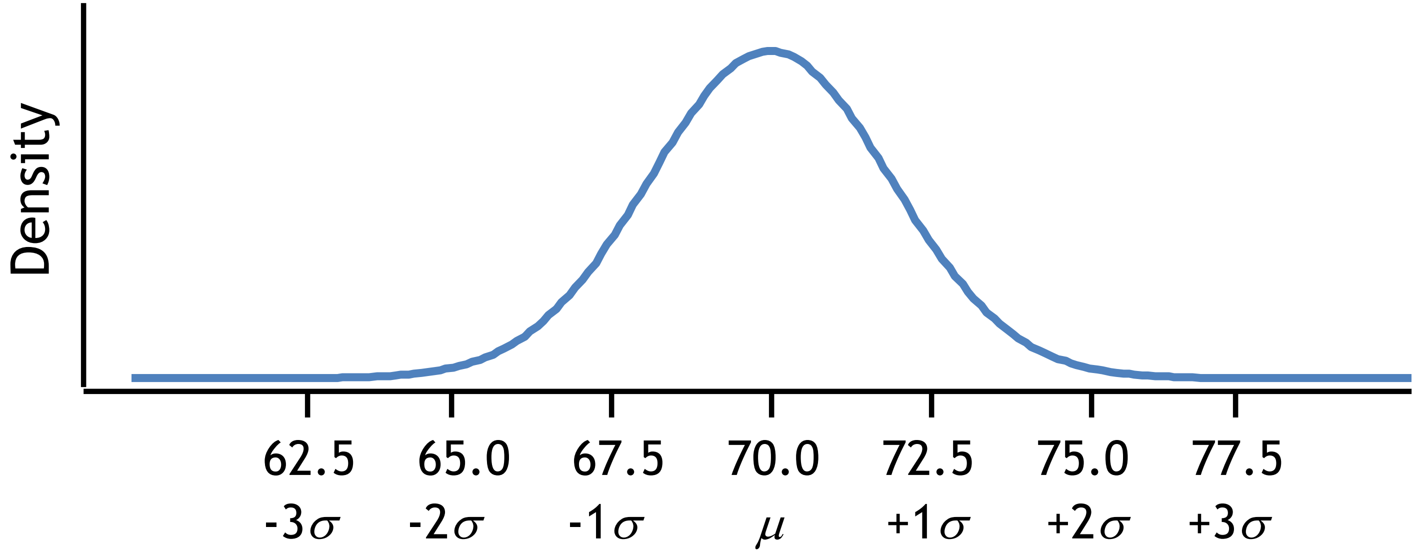



**What proportion of men is between 66" and 74"?**

Thus, P(-1.6<Z<1.6) = 0.945 – 0.055 = 0.89 … or 89%

# Worksheet

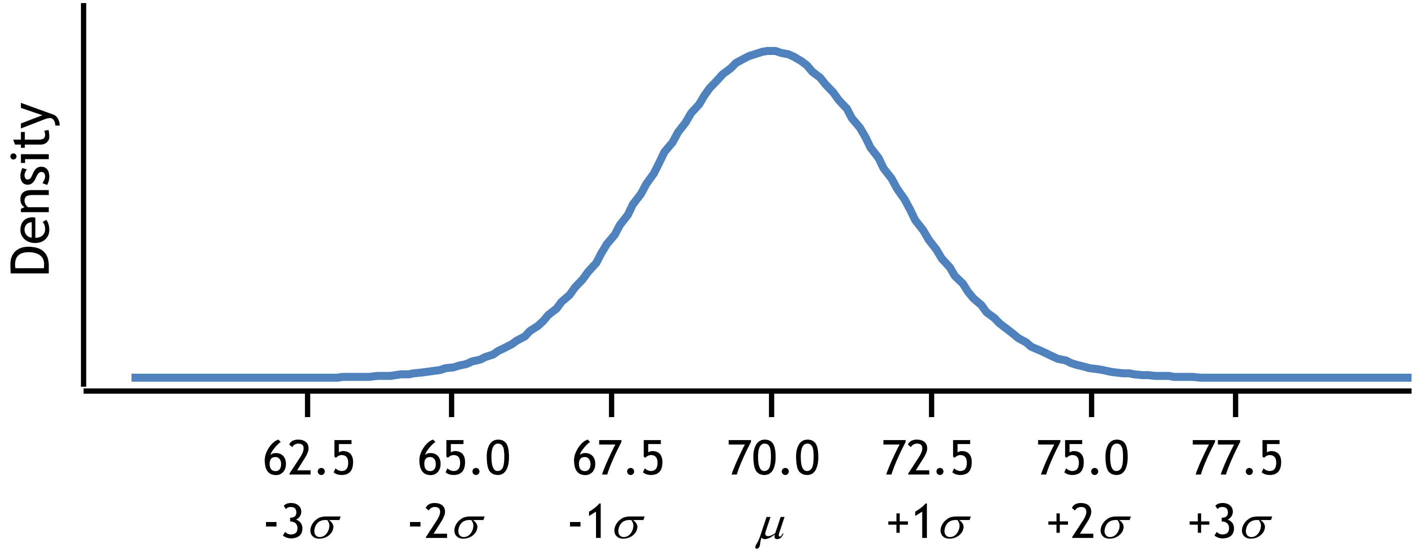
Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5




**What proportion of men is taller than me (69.5")?**

# Worksheet

Men's height in inches in the United States is normally distributed with $\mu$=70 and $\sigma$=2.5




**What proportion of men is between 66" and 71" tall?**

# Worksheet

The distributions of times for 19-34 year men and women to complete a half marathon are (more or less) normal

**Men**: $\mu_{Y_{MEN}}$ (in minutes) is 125 with $\sigma_{Y_{MEN}}$ of 25

**Women**: $\mu_{Y_{WOMEN}}$ is 140 with $\sigma_{Y_{WOMEN}}$ of 23

What proportion of women finish half marathons ahead of the average man?

(Change of Topic)

# Combinations of Random Variables

It is often useful to combine (e.g., add or subtract) random variables

*Example*: Time to commute to and from work

The time that it takes to drive to work is a normal random variable $Y_{TO}$, with $\mu_{TO}$=30 minutes and $\sigma_{TO}$=10

The time that it takes to drive home from work is also a normal random variable $Y_{FROM}$, with $\mu_{FROM}$=25 and $\sigma_{FROM}$=15

What is the distribution of total commute time?

What is the distribution of the difference between commutes to and from work?

# Combinations of Random Variables

In general, for two random variables Y and Z:

Mean of Y + Z = $\mu_{Y+Z} = \mu_Y + \mu_Z$

and

Mean of Y - Z = $\mu_{Y-Z} = \mu_Y - \mu_Z$

(Note: These rules are true for discrete or continuous random variables, and are true whether or not the variables are independent)

# Combinations of Random Variables

For two independent random variables *Y* and *Z*:

Variance of $Y + Z = \sigma^2_{Y+Z} = \sigma^2_Y + \sigma^2_Z$

and

Variance of $Y - Z = \sigma^2_{Y-Z} = \sigma^2_Y + \sigma^2_Z$

(*Note*: These rules for combinations of variance only hold for independent random variables, but they work for discrete or continuous variables.)

# Combinations of Random Variables

*Example*: Time to commute to and from work

The time that it takes to drive to work is a normal random variable $Y_{TO}$, with $\mu_{TO}$=30 minutes and $\sigma_{TO}$=10

The time that it takes to drive home from work is also a normal random variable $Y_{FROM}$, with $\mu_{FROM}$=25 and $\sigma_{FROM}$=15

What is the distribution of total commute time?

$$\mu_{TO+FROM} = \mu_{TO} + \mu_{FROM} = 30 + 25 = 55 \text{ minutes}$$

$$\sigma^2_{TO+FROM} = \sigma^2_{TO} + \sigma^2_{FROM} = 10^2 + 15^2 = 325$$

# Combinations of Random Variables

*Example*: Time to commute to and from work

The time that it takes to drive to work is a normal random variable $Y_{TO}$, with $\mu_{TO}$=30 minutes and $\sigma_{TO}$=10

The time that it takes to drive home from work is also a normal random variable $Y_{FROM}$, with $\mu_{FROM}$=25 and $\sigma_{FROM}$=15

What is the distribution of the difference between commutes to and from work?

$$\mu_{TO\text{-}FROM} = \mu_{TO} - \mu_{FROM} = 30 - 25 = 5 \text{ minutes}$$

$$\sigma^2_{TO\text{-}FROM} = \sigma^2_{TO} + \sigma^2_{FROM} = 10^2 + 15^2 = 325$$

# Worksheet

The distributions of times for 19-34 year men and women to complete a half marathon are (more or less) normal

**Men**: $\mu_{Y_{MEN}}$ (in minutes) is 125 with $\sigma_{Y_{MEN}}$ of 25

**Women**: $\mu_{Y_{WOMEN}}$ is 140 with $\sigma_{Y_{WOMEN}}$ of 23

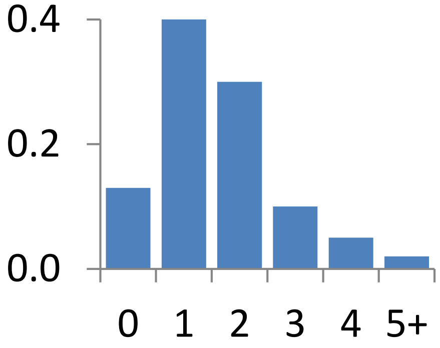
What is the distribution of the difference between men's and women's race times?

# Reality

# Data

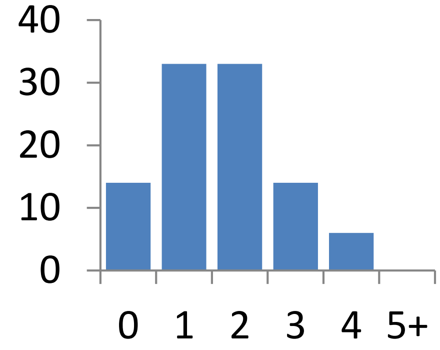
# Knowledge

**Theoretical Distribution**
for the **Population** of
**All** Possible Responses




2954462187388221348
8711102682140050851
9291758426493659357
1797402968538728539
4401671369193970834
0165785879715135279
6365488407871392124
2920353540363160907
9546090581563252173
8998220343490929571
8159244365888845385
0229635205640107762
4385245611536981630
9196626803867353805
4956868092019177602
1078342587201850530
2250228297490991053
5575276026995705920
9224813120745604486

**Observed Distribution**
for a **Sample** of
**100** Sets of Responses



# Random Variables

**Today**

We know the actual probabilities associated with the random event that generates the theoretical distribution (e.g., coin flips)

We will learn to *describe* and *make practical use* of these theoretical distributions

**Later (and in Life)**

We do not know the actual probabilities associated with the random event (e.g., number of children per person, which candidate will win)

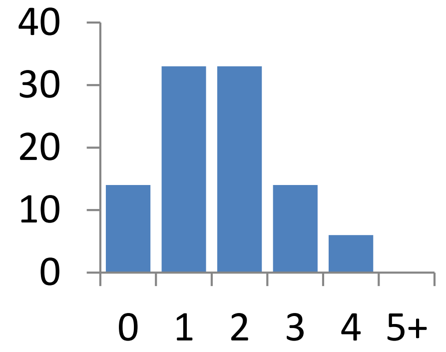
Our ability to *describe* and *make practical use* of theoretical distributions will allow us to infer those probabilities

Reality | Data | Knowledge

?

29544621873882213480
87111026821400508510
92917584264936593570
17974029685387285390
44016713691939708340
01657858797151352790
63654884078713921240
29203535403631609070
95460905815632521730
89982203434909295710
81592443658888453850
02296352056401077620
43852456115369816300
91966268038673538050
49568680920191776020
10783425872018505300
22502282974909910530
55752760269957059200
92248131207456044860

**Observed Distribution**
for a **Sample** of
**100** Sets of Responses



# Want More?

This is a good reading:

http://www.stat.cmu.edu/~cshalizi/36-220/lecture-7.pdf

David Lane's Book

http://onlinestatbook.com/2/normal_distribution/normal_distribution.html

Gerard Dallal's Book

http://www.jerrydallal.com/LHSP/normal.htm