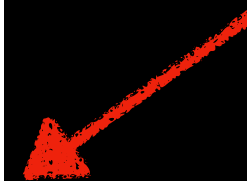


**Welcome to Week #5!**

Week	Topic	Reading
1	<ul style="list-style-type: none"> <li>• Data, Models, and Information</li> <li>• Elementary statistics: Definitions</li> <li>• Overview of R</li> </ul>	OIS 1 (ISL 1)
2	<ul style="list-style-type: none"> <li>• Elementary statistics: Applications &amp; Plots</li> </ul>	OIS 1 (ISL 1)
3	<ul style="list-style-type: none"> <li>• Introduction to data analysis with R</li> <li>• Review of tabular and graphical displays of data</li> </ul>	ITR 1, 2, 5, 6, 7, 12
4	<ul style="list-style-type: none"> <li>• Random variables: expectation and variance</li> <li>• Joint and conditional probability</li> <li>• Bayes rule</li> </ul>	OIS 2
5	<ul style="list-style-type: none"> <li>• Random variables: distributions (normal, binomial, poisson)</li> </ul>	OIS 3

Getting into the fun stuff!

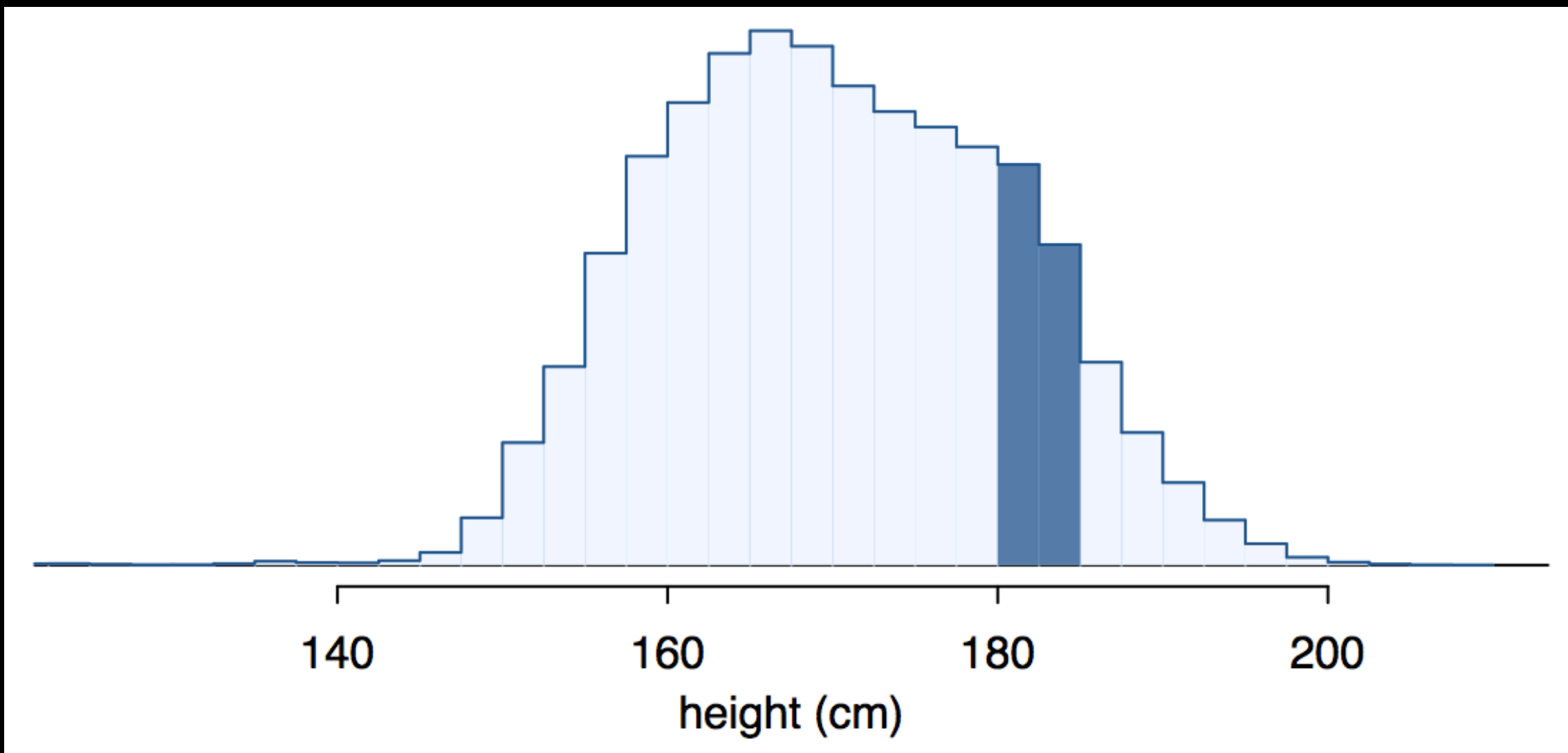


# Moving on: Continuous distributions

Many of the ideas we've messed with here can be applied to large distributions - which we can approximate as continuous.

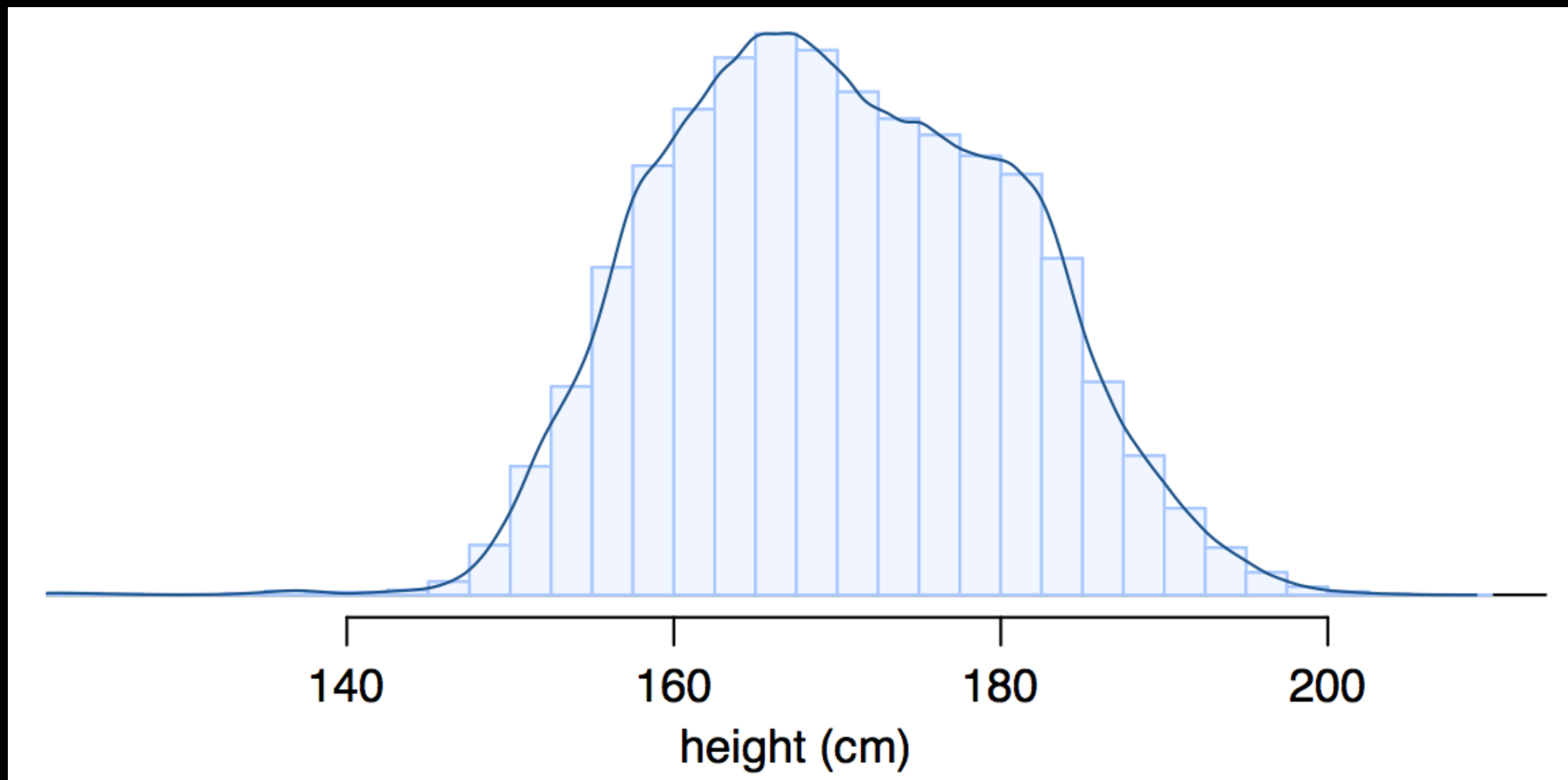
# Moving on: Continuous distributions

Below is a histogram of the distribution of heights of US adults. The proportion of data that falls in the shaded bins gives the probability that a randomly sampled US adult is between 180 cm and 185 cm (about 5'11" to 6'1").



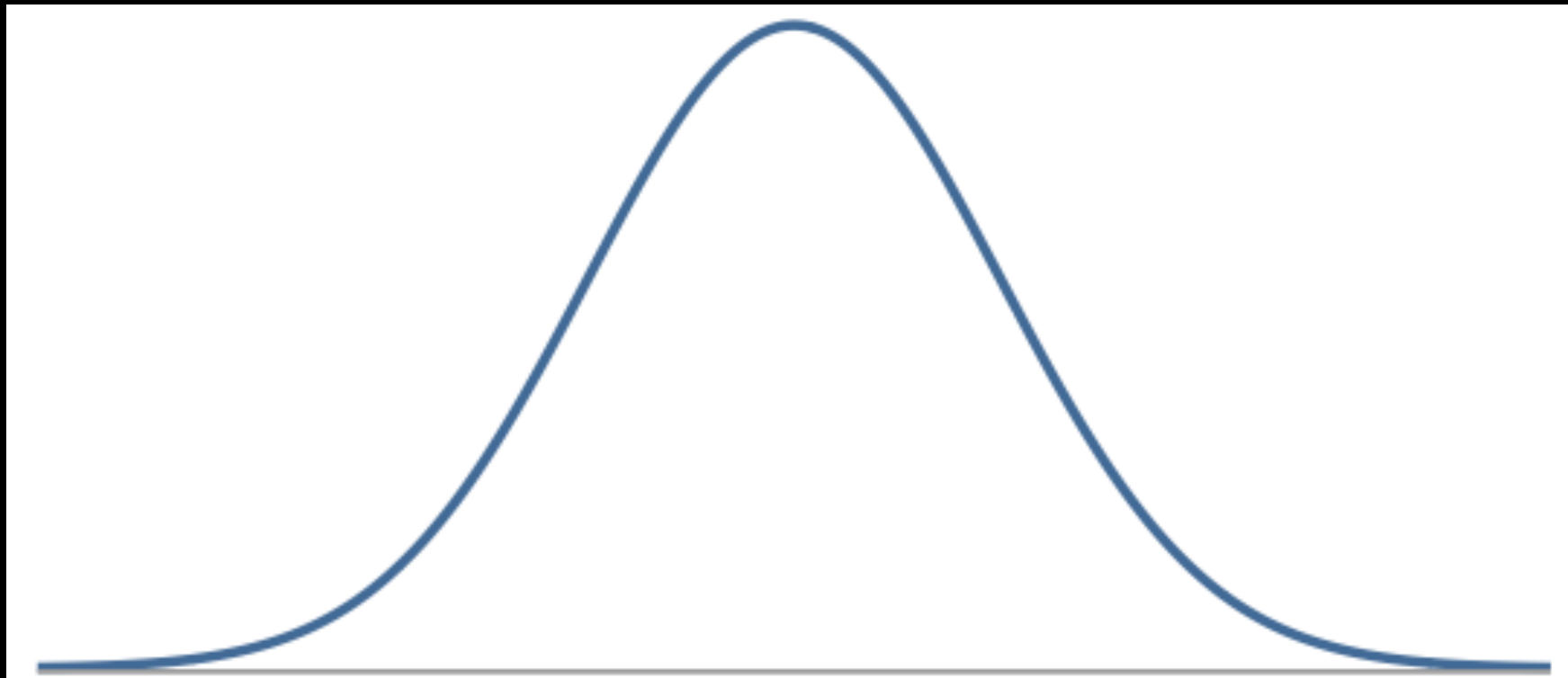
# From histograms to continuous distributions

Since height is a continuous numerical variable, its **probability density function** is a smooth curve.



# The Normal distribution

In Chapter 3, we look at the Normal distribution. The Normal distribution is the most famous continuous distribution.



To find areas under curves, we generally use a table or technology (i.e. calculator, stat program, etc.).

# The Normal distribution...

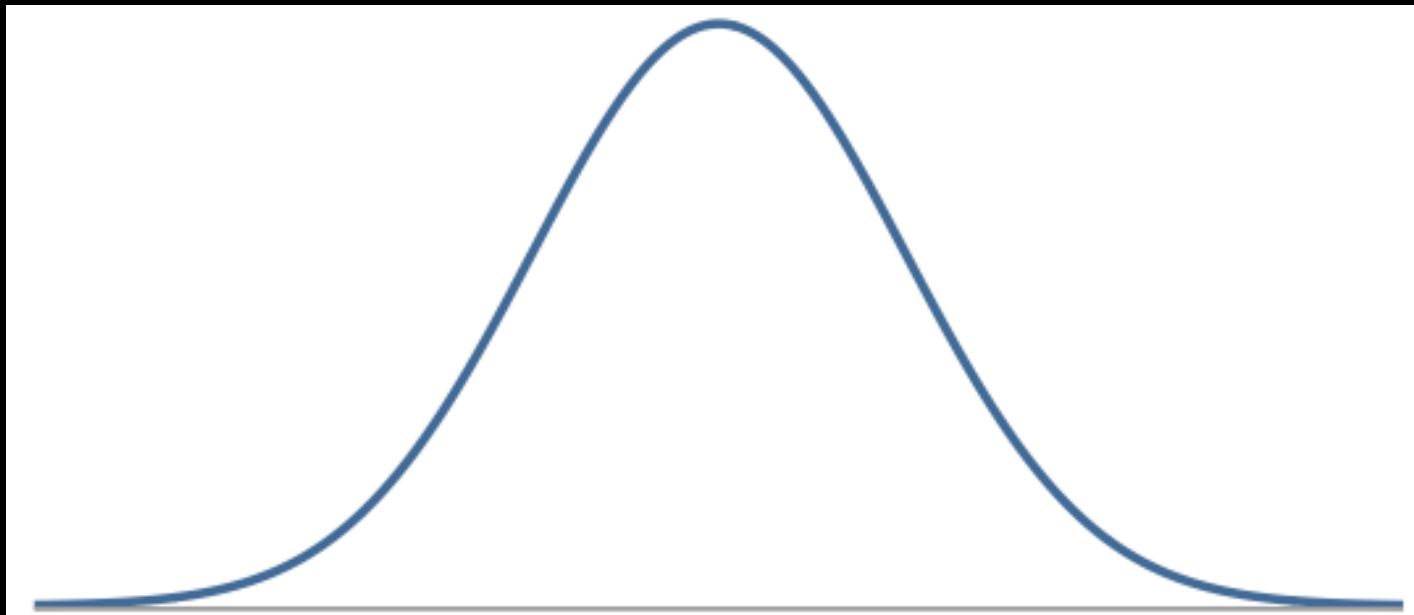
is the most well known continuous distribution

Is unimodal and symmetric, bell shaped curve

has mean  $\mu$  and standard deviation  $\sigma$

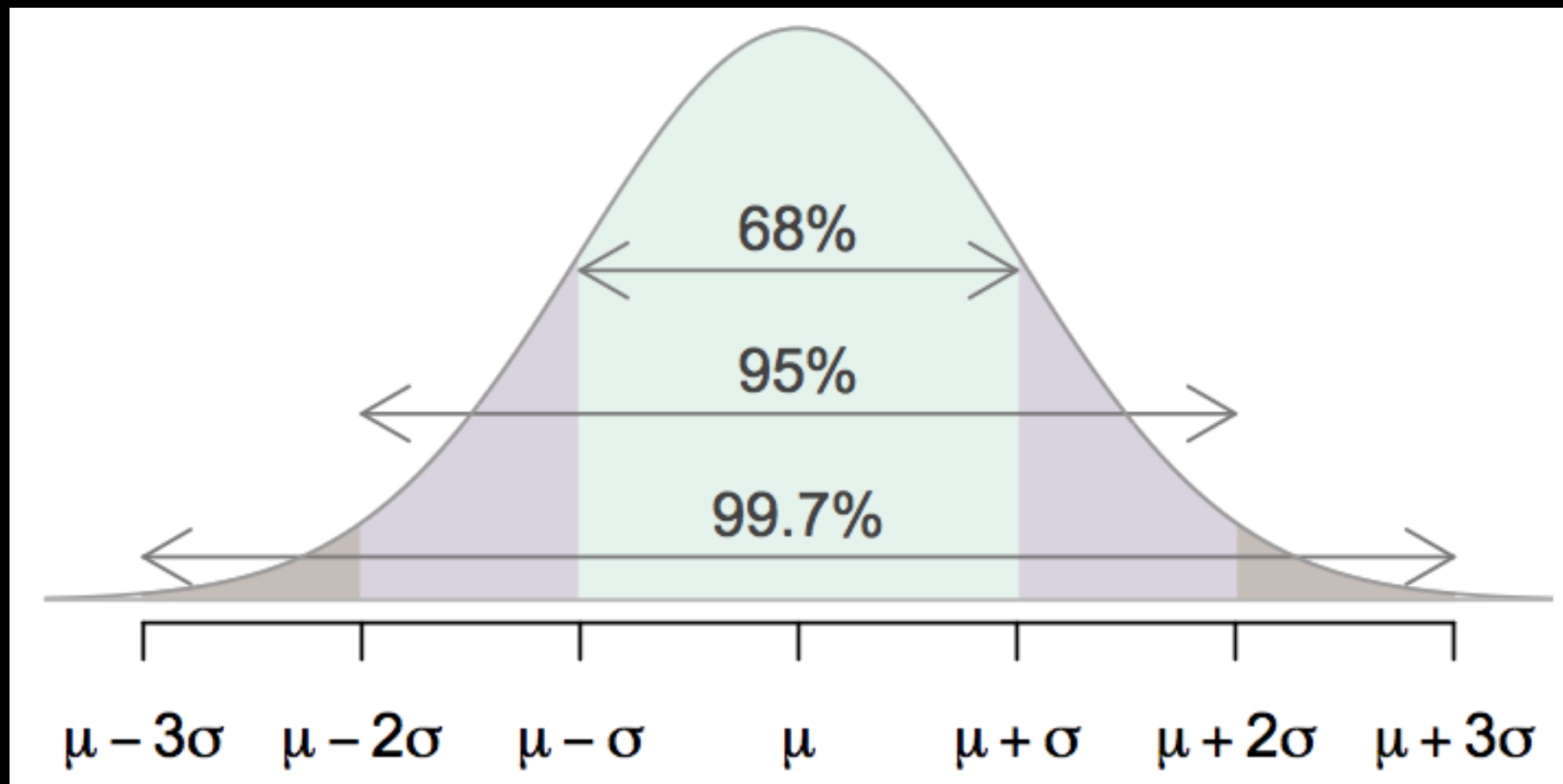
has tails that extend infinitely in both directions

Many variables are nearly normal, but none are *exactly* normal



# The source of the 68-95-99.7 Rule

For nearly normally distributed data,  
about 68% falls within 1 SD of the mean,  
about 95% falls within 2 SDs of the mean,  
about 99.7% falls within 3 SDs of the mean.



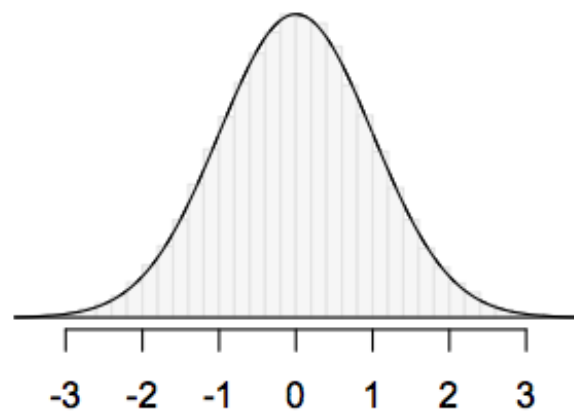
**And the total area  
under the curve = 1**



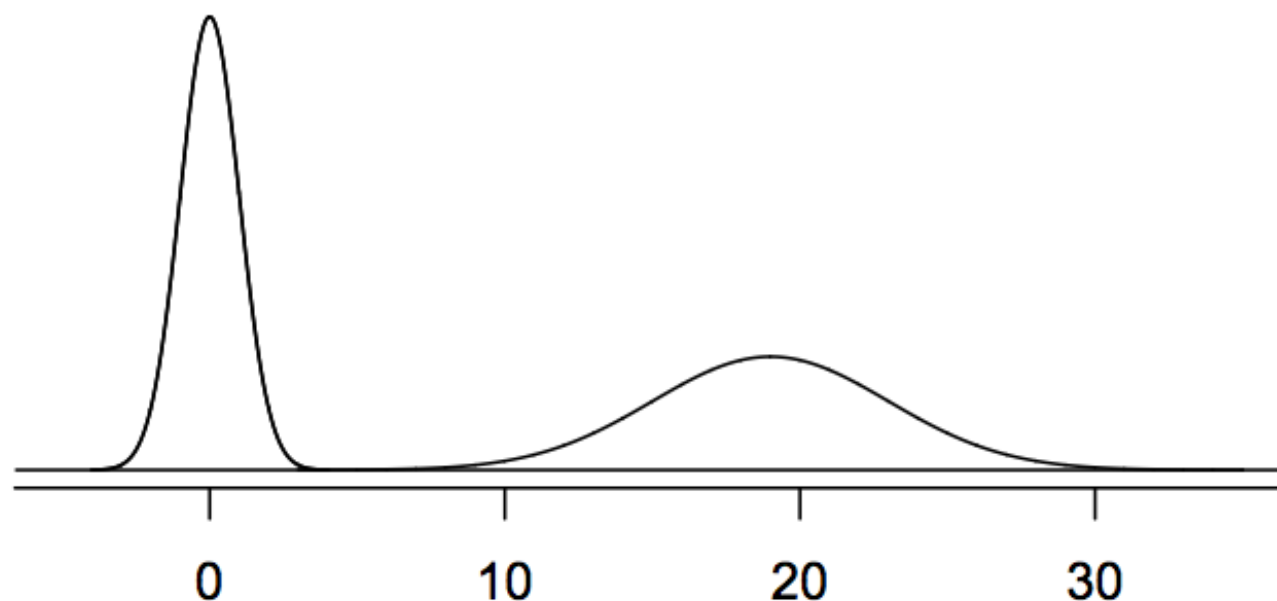
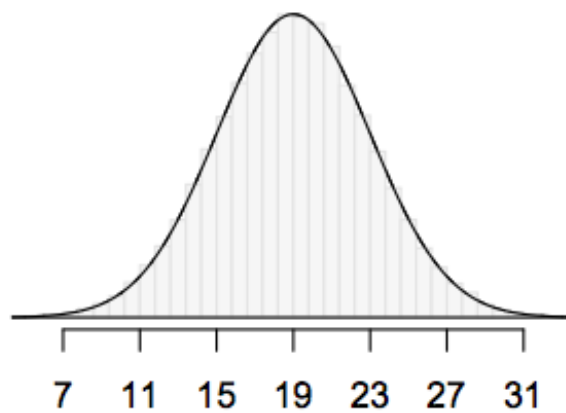
# Normal distributions with different parameters

$\mu$ : mean,  $\sigma$ : standard deviation

$$N(\mu = 0, \sigma = 1)$$



$$N(\mu = 19, \sigma = 4)$$



# Why study the normal distribution?

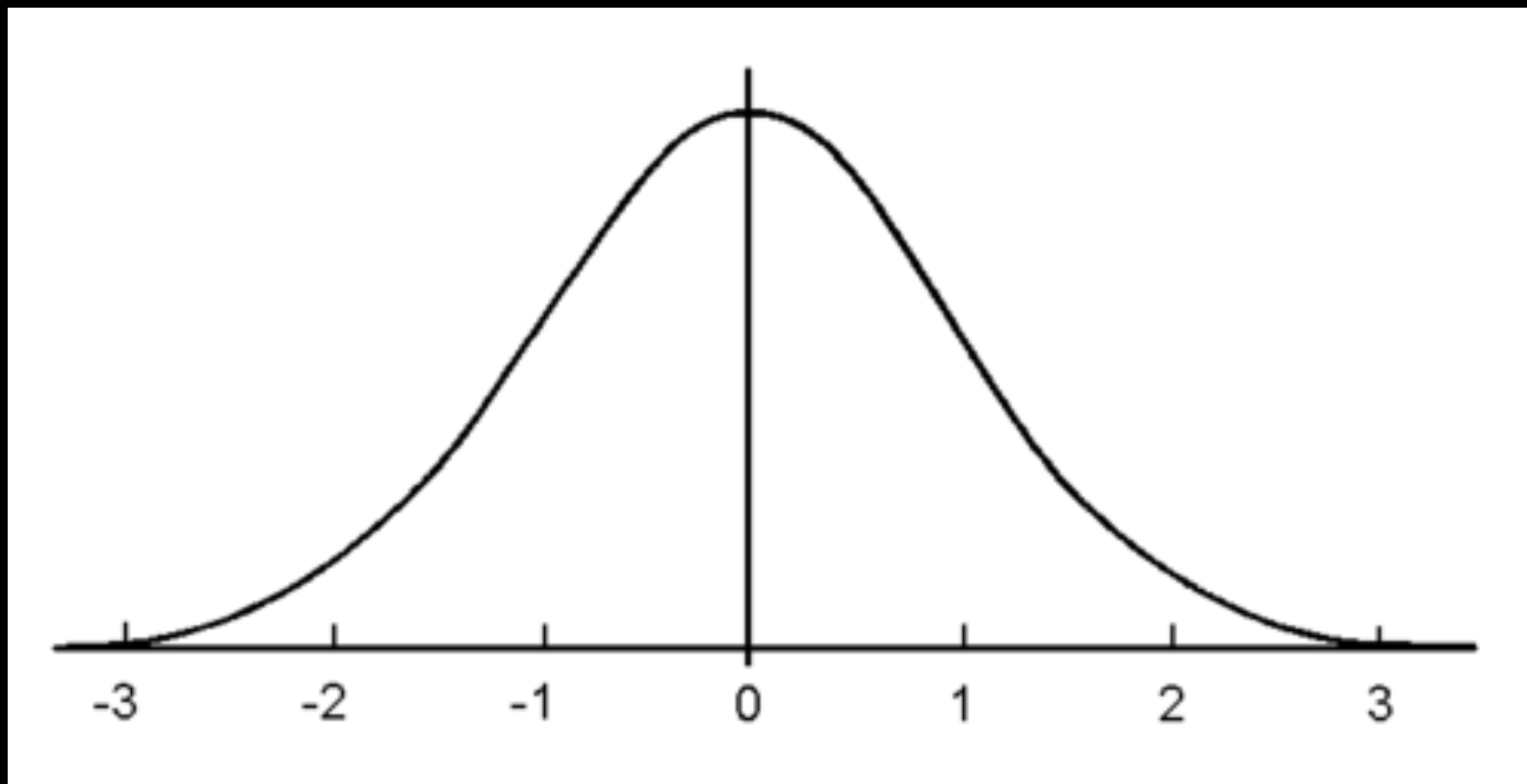
<https://galtonboard.com/probabilityexamplesinlife>

# The Standard Normal Curve

What units are on the horizontal axis?

Z-scores!

**A way to compare normal distributions**



# Standardizing with Z scores (cont.)

Z score of an observation is the *number of standard deviations* it falls above or below the mean.

$$Z = \frac{(\text{observation} - \text{mean})}{\text{SD}}$$

Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.

Observations that are more than 2 SD away from the mean ( $|Z| > 2$ ) are generally considered unusual.

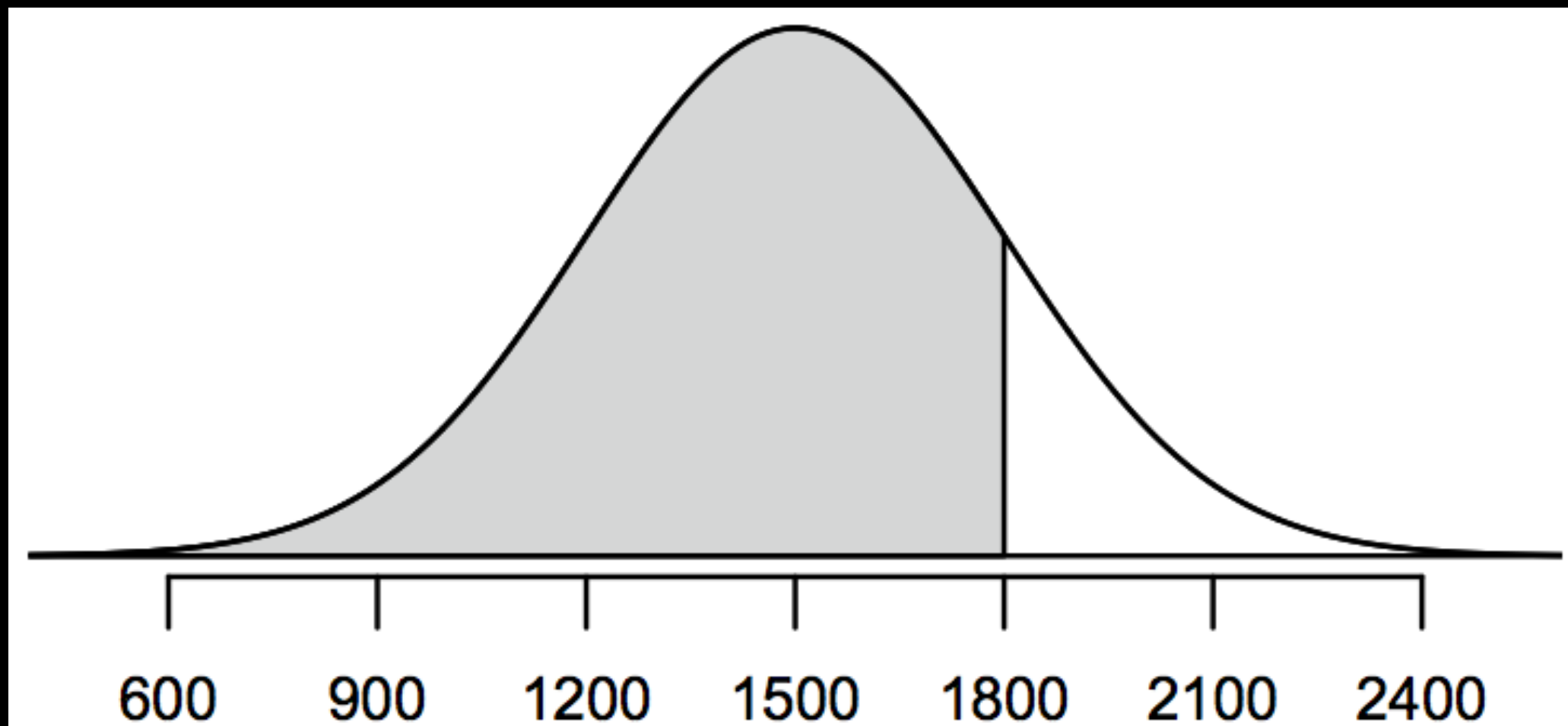
We need to (1) understand how to easily calculate these and (2) depict them on graphs

## In R!

# Percentiles

**Percentile** is the percentage of observations that fall below a given data point.

Graphically, percentile is the area below the probability distribution curve to the left of that observation.



In R!

# Finding percentiles from the standard normal curve

What Z-score corresponds to the 50th percentile?

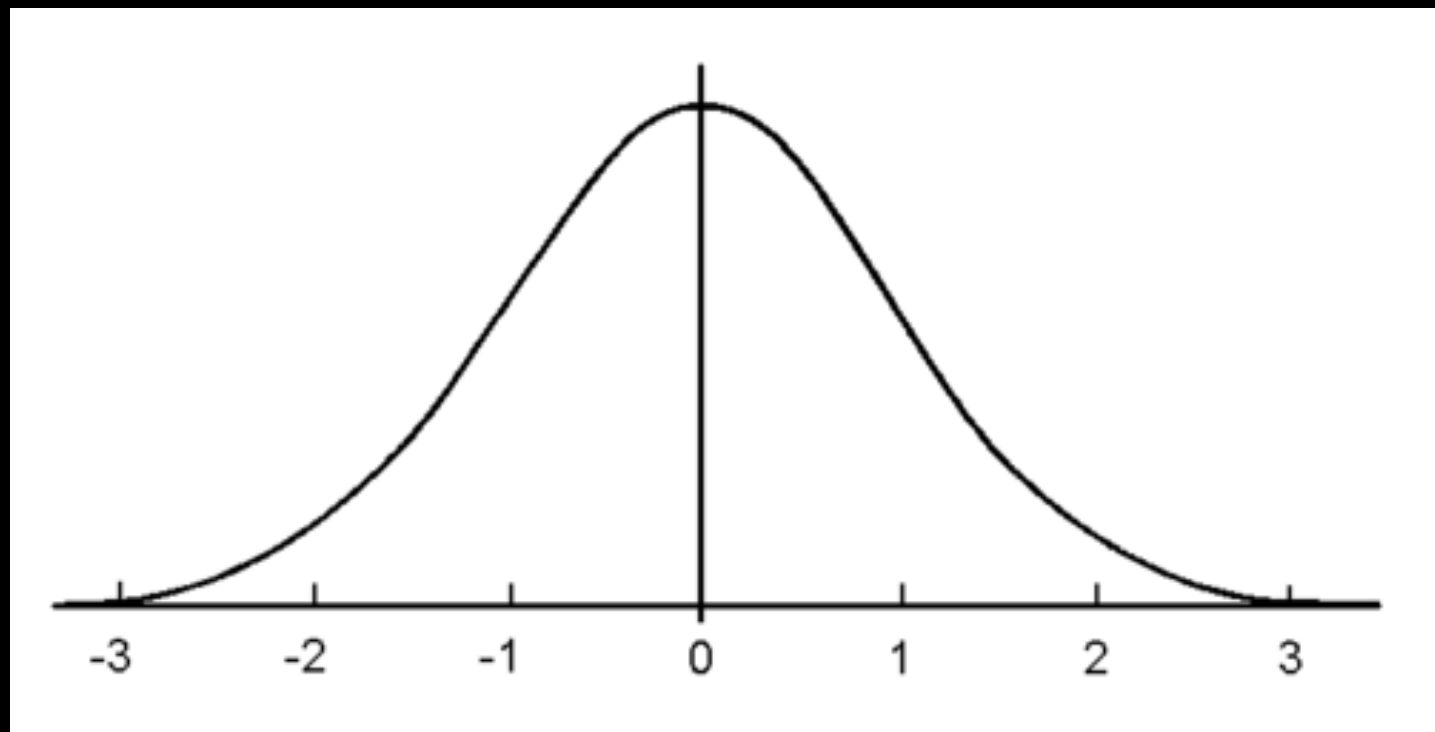
i.e.  $P(Z < ?) = 0.5$   $Z =$

What Z-score corresponds to the 20th percentile?

i.e.  $P(Z < ?) = 0.2$   $Z =$

What Z-score has 70% of the area to the *right* of it?

i.e.  $P(Z < ?) = 0.3$   $Z =$

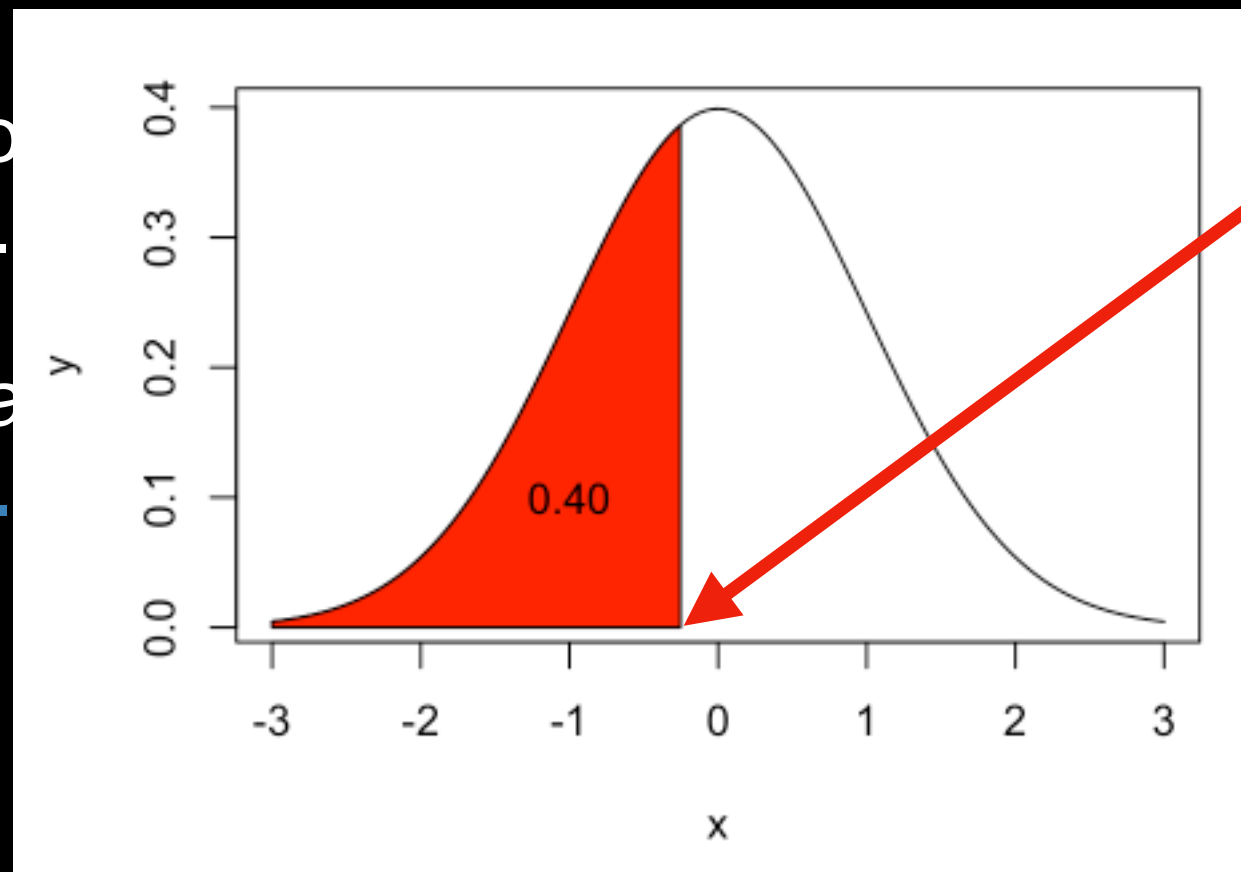


# Finding percentiles from the standard normal curve

What Z-score corresponds to the 50th percentile?  
i.e.  $P(Z < ?) = 0.5$   $Z =$

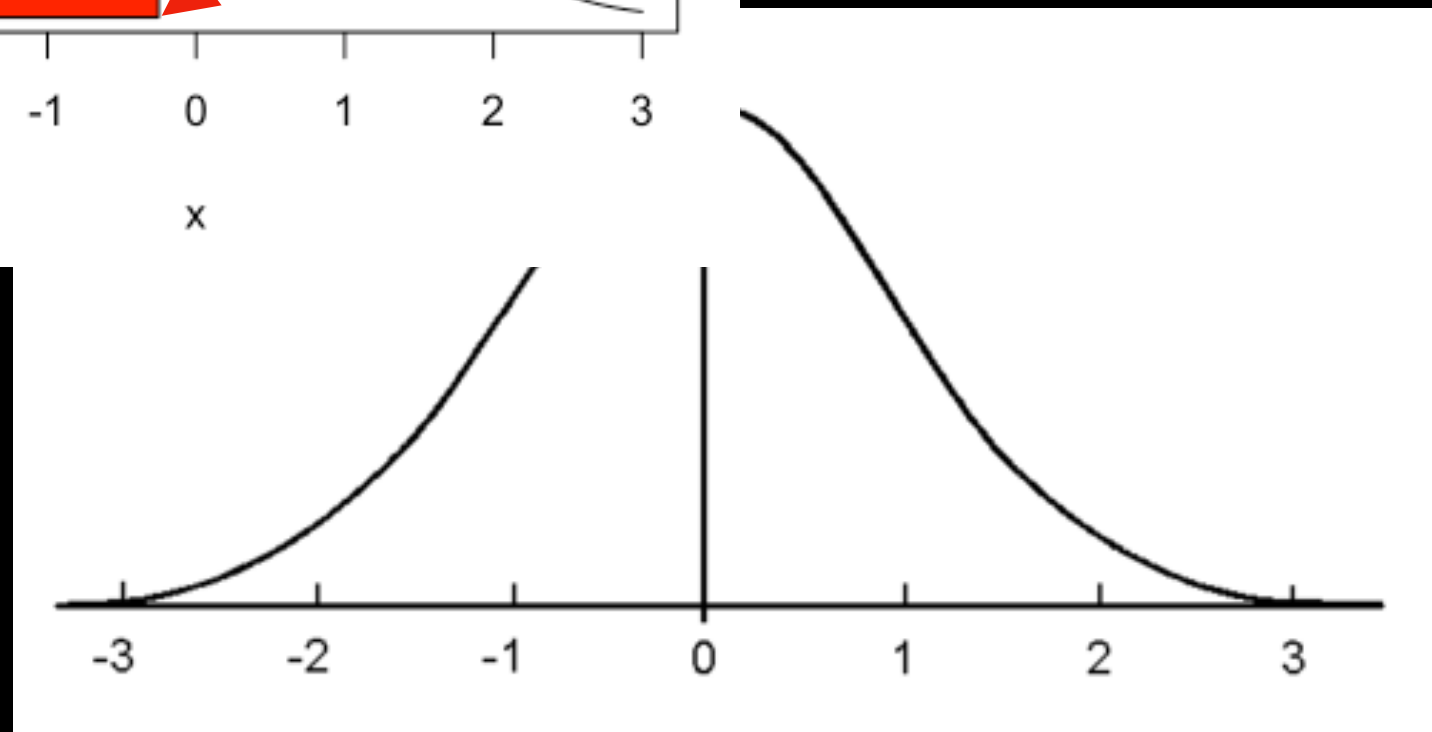
What Z-score corresponds to the 40th percentile?  
i.e.  $P(Z < ?) = 0.4$

What Z-score has a right-tail area of 0.40?  
i.e.  $P(Z > ?) = 0.4$



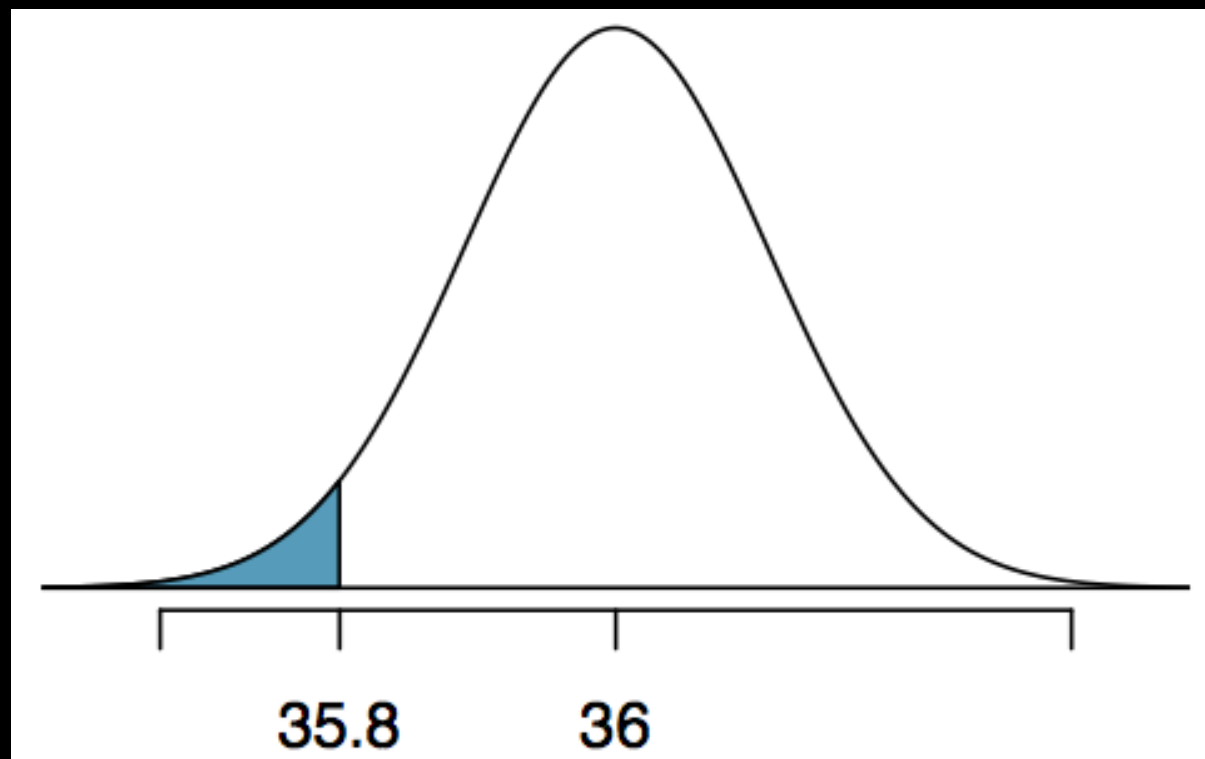
What is this number  
such that red area =  
0.40 (40%)

In R!



# Quality control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. If the amount of ketchup in the bottle is below 35.8 oz. then the bottle fails the quality control inspection. What is the probability that a randomly selected bottle fails quality control (that is, what percent of bottles fail quality control)?



**Try in R or by hand  
(you can just write out what you'd calculate)**



# Is it Normal? The Normal probability plot

• The Normal probability plot is a graphical tool used to check if a dataset follows a normal distribution. It plots the sample quantiles against the theoretical quantiles of a normal distribution. If the data points fall approximately on a straight line, the data is likely normally distributed.

• The plot is constructed by ordering the sample data and plotting them against the theoretical quantiles of a standard normal distribution. The theoretical quantiles are generated from a standard normal distribution table or software.

• The plot is useful for identifying outliers and non-normality in the data. Points that deviate significantly from the straight line indicate potential outliers or non-normality.

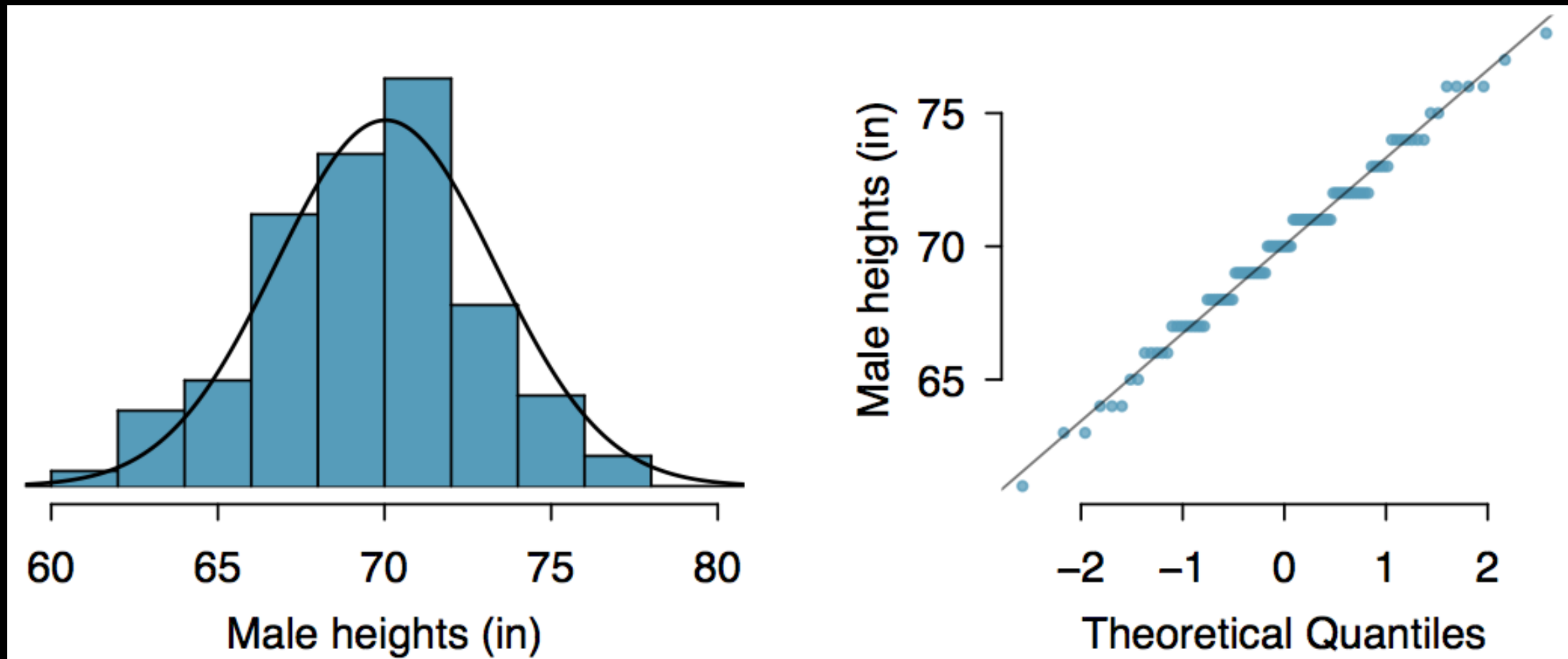
• The plot is also used to compare the distribution of two or more datasets. If the points for different datasets fall on the same straight line, they are likely from the same normal distribution.

• The plot is a simple and effective tool for checking normality, and it is widely used in statistical analysis.

• The plot is a graphical representation of the data, and it is easy to interpret. It provides a visual check for normality, which is a common assumption in many statistical tests.

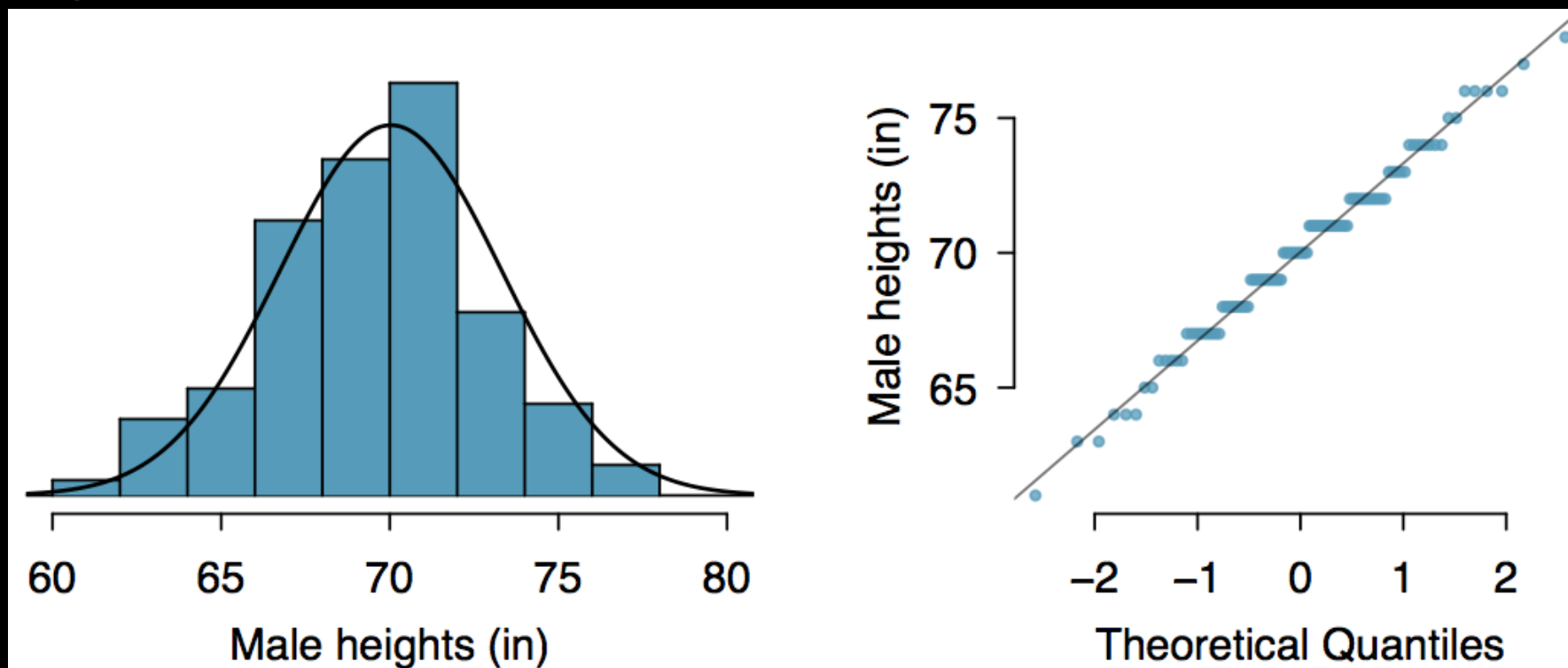
# Is it Normal? The Normal probability plot

A histogram and normal probability plot of a sample of 100 male heights.



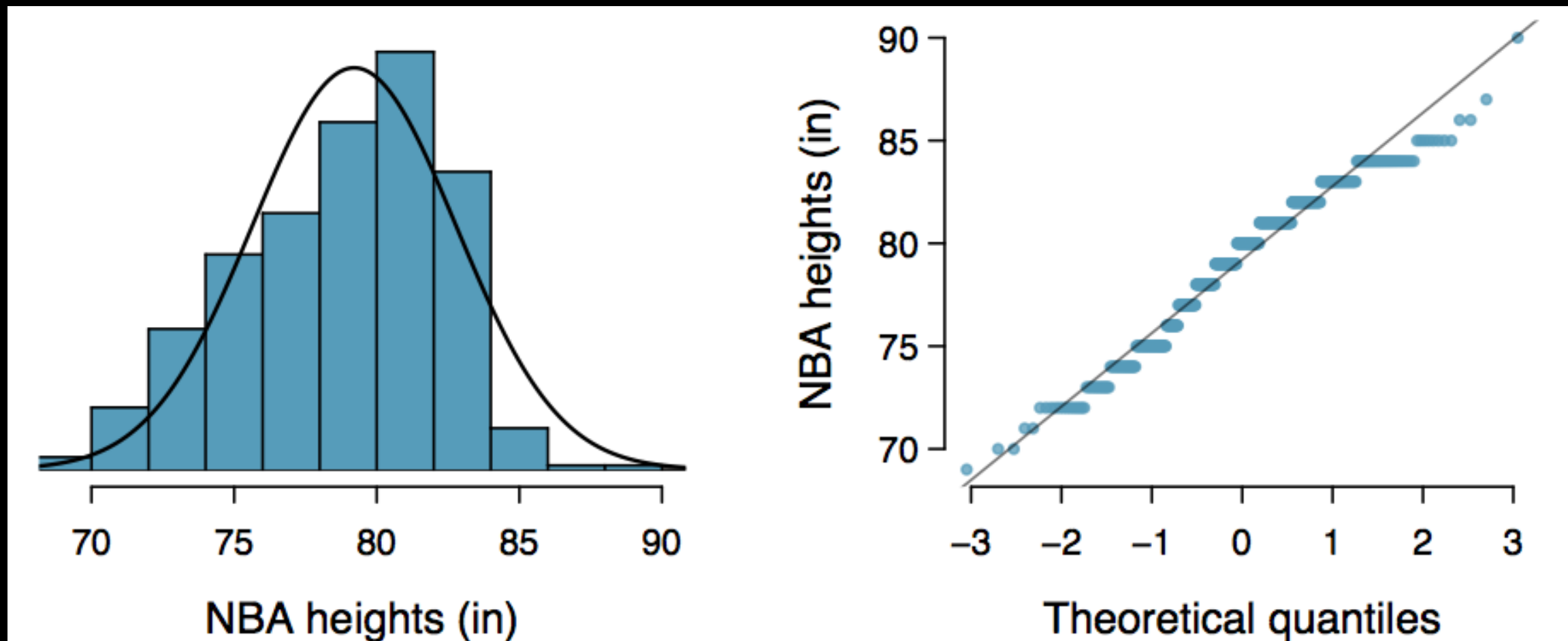
# Anatomy of a normal probability plot

- Data are plotted on the y-axis of a normal probability plot, and theoretical quantiles (following a normal distribution) on the x-axis.
- If there is a linear relationship in the plot, then the data follow a nearly normal distribution.
- Constructing a normal probability plot requires calculating percentiles and corresponding z-scores for each observation, which is tedious. Therefore we generally rely on software when making these plots.



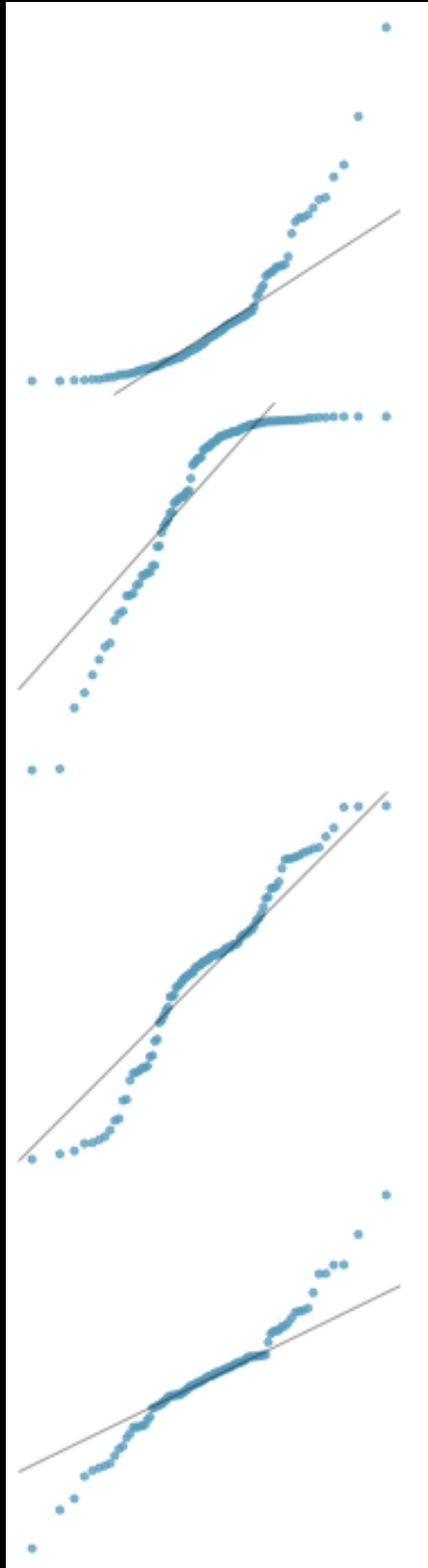
# Practice

Below is a histogram and normal probability plot for the NBA heights from the 2008-2009 season. Do these data appear to follow a normal distribution?



Why do the points on the normal probability have jumps?

# Normal probability plot and skewness



Right skew - Points bend up and to the left of the line.

Left skew - Points bend down and to the right of the line.

Short tails (narrower than the normal distribution) - Points follow an S shaped-curve.

Long tails (wider than the normal distribution) - Points start below the line, bend to follow it, and end above it.

**Is it normal?**  
**Simulations & Fish Data!**

**The Binomial formula: A quick aside for next week!**

# The Binomial formula

If  $p$  represents probability of success,  $(1-p)$  represents probability of failure,  $n$  represents number of **independent** trials, and  $k$  represents number of successes

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1 - p)^{(n-k)}$$



# The Binomial formula

If  $p$  represents probability of success,  $(1-p)$  represents probability of failure,  $n$  represents number of **independent** trials, and  $k$  represents number of successes

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1 - p)^{(n-k)}$$

WHAT



# The Binomial formula

If  $p$  represents probability of success,  $(1-p)$  represents probability of failure,  $n$  represents number of **independent** trials, and  $k$  represents number of successes

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{(n-k)}$$

WHAT

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Still WHAT

# The Binomial formula: Factorials

If  $p$  represents probability of success,  $(1-p)$  represents probability of failure,  $n$  represents number of **independent** trials, and  $k$  represents number of successes

“ $n$  choose  $k$ ”

$$= k * (k-1) * (k-2) * \dots * 1$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$= n * (n-1) * (n-2) * \dots * 1$$

$$= (n-k) * (n-k-1) * \dots * 1$$

# The Binomial formula: Factorials

$$n! = \text{factorial}(n) \quad \boxed{\binom{n}{k} = \frac{n!}{k!(n-k)!}} \quad = \text{choose}(n,k)$$

$$\binom{3}{2} = \frac{3*2*1}{(2*1)*(3-2)}$$

“3 choose 2”

$$\binom{7}{3} = \frac{7*6*5*4*3*2*1}{(3*2*1)*[(7-3)*(7-3-1)*(7-3-2)*(7-3-3)]}$$

“7 choose 3”



**If we have time: more practice problems!**

# Practice

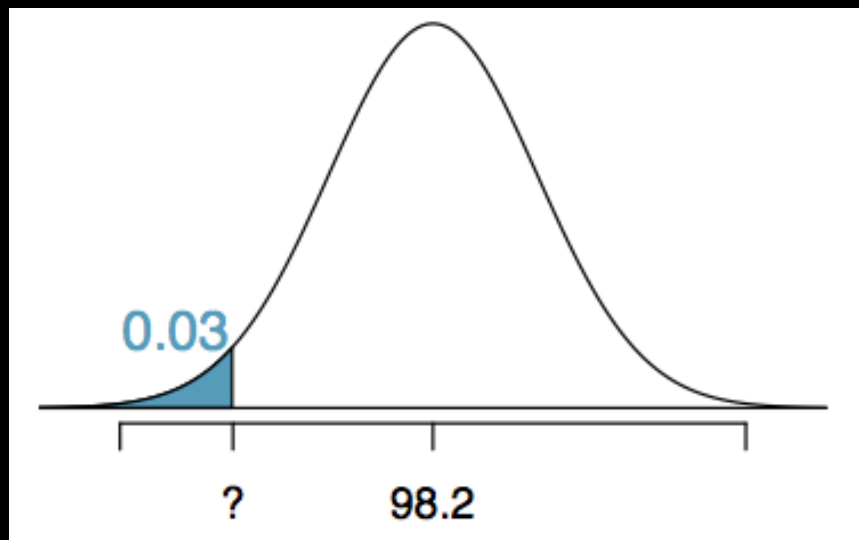
If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle *fails* the quality control inspection. Recall  $\mu = 36$ ,  $\sigma = 0.11$ .

What percent of bottles pass the quality control inspection?

- |           |            |
|-----------|------------|
| (a) 1.82% | (d) 93.09% |
| (b) 3.44% | (e) 96.56% |
| (c) 6.88% |            |

# Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean  $98.2^{\circ}\text{F}$  and standard deviation  $0.73^{\circ}\text{F}$ . What is the cutoff for the lowest 3% of human body temperatures?



Mackowiak, Wasserman, and Levine (1992), A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick.



# Practice

Body temperatures of healthy humans are distributed nearly normally with mean  $98.2^{\circ}\text{F}$  and standard deviation  $0.73^{\circ}\text{F}$ . What is the cutoff for the *highest* 10% of human body temperatures?

- |                            |                            |
|----------------------------|----------------------------|
| (a) $97.3^{\circ}\text{F}$ | (c) $99.4^{\circ}\text{F}$ |
| (b) $99.1^{\circ}\text{F}$ | (d) $99.6^{\circ}\text{F}$ |