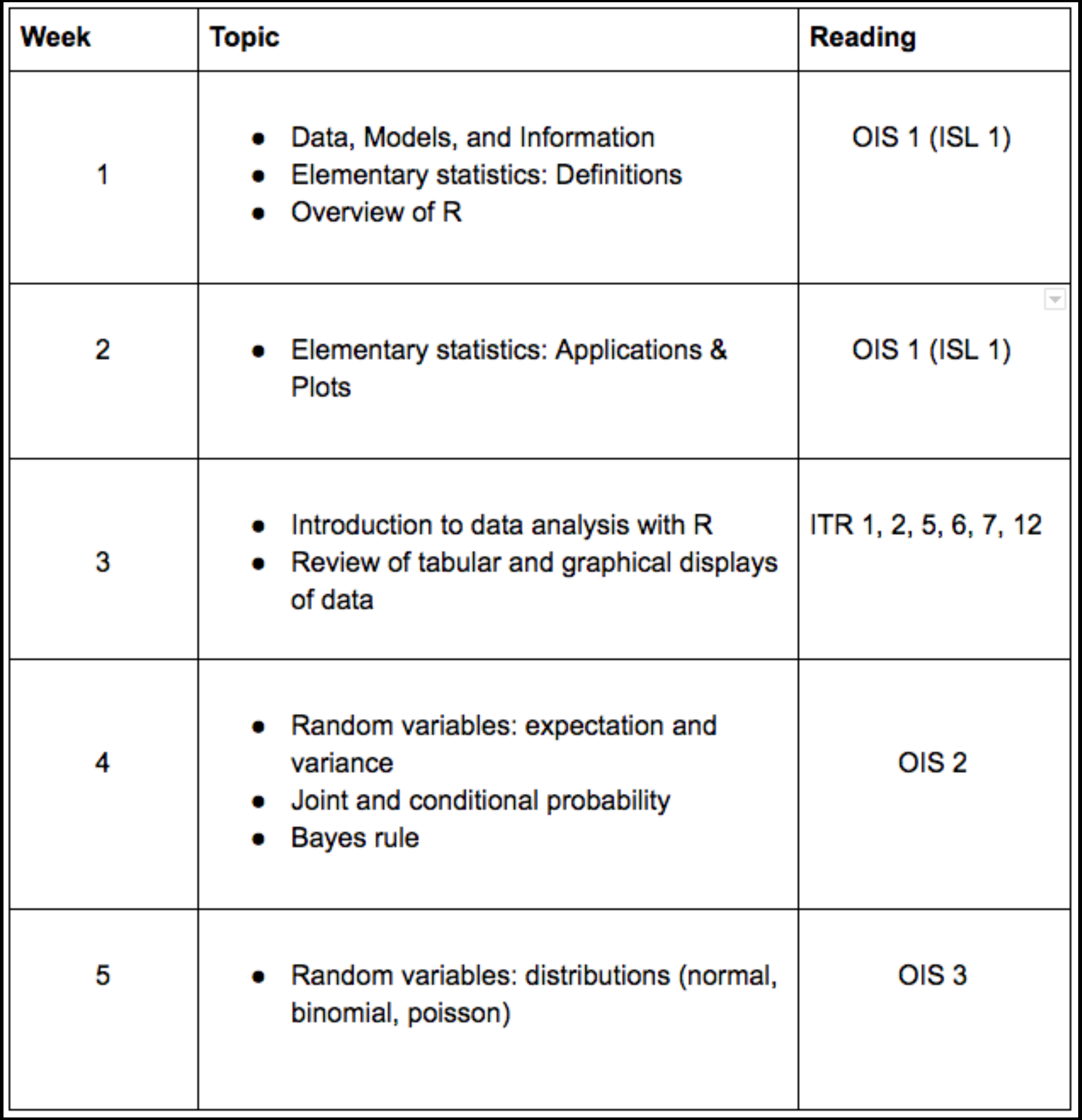
Welcome to Week #5!



Week	Topic	Reading
1	• Data, Models, and Information • Elementary statistics: Definitions • Overview of R	OIS 1 (ISL 1)
2	• Elementary statistics: Applications & Plots	OIS 1 (ISL 1)
3	• Introduction to data analysis with R • Review of tabular and graphical displays of data	ITR 1, 2, 5, 6, 7, 12
4	• Random variables: expectation and variance • Joint and conditional probability • Bayes rule	OIS 2
5	• Random variables: distributions (normal, binomial, poisson)	OIS 3

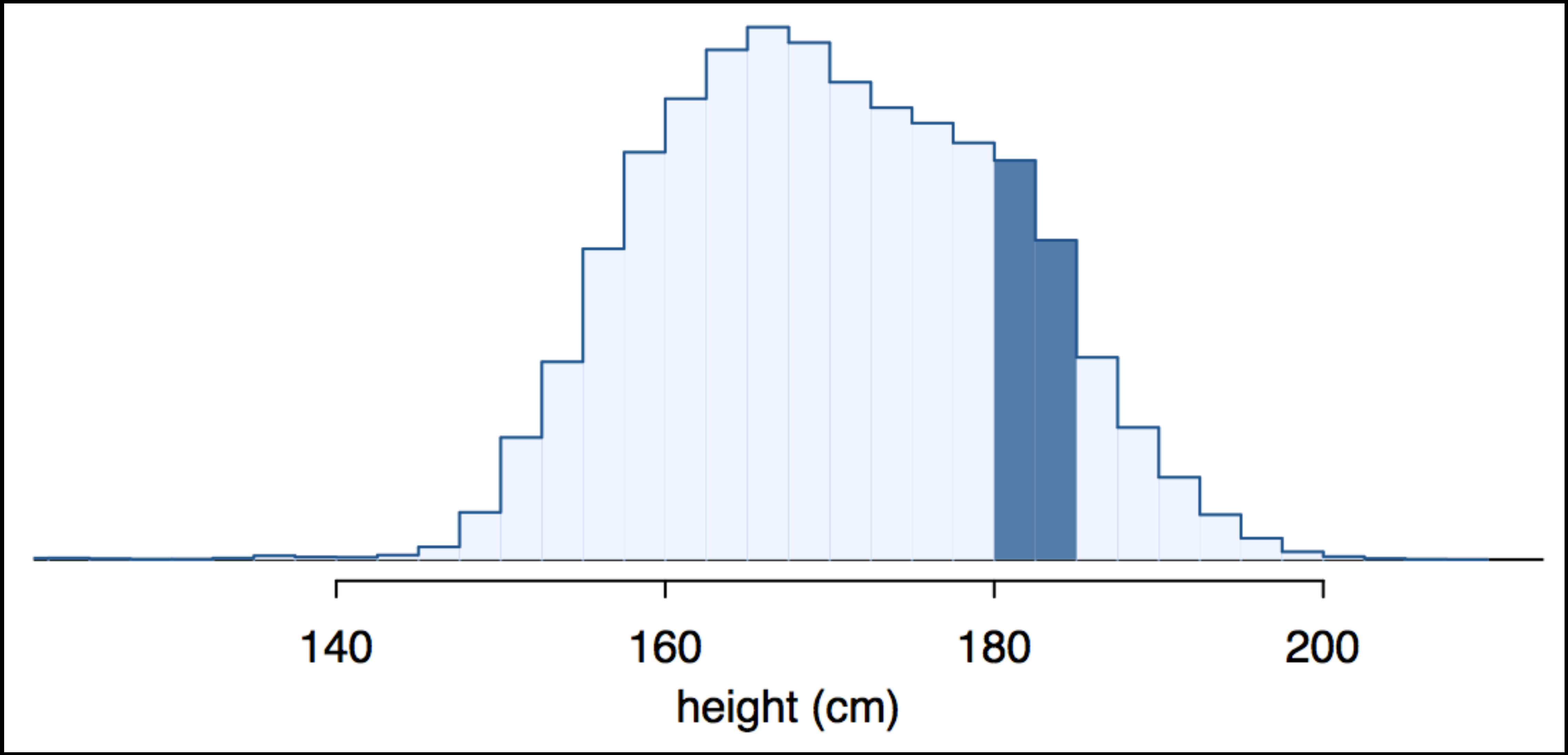
Getting into the fun stuff!

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.

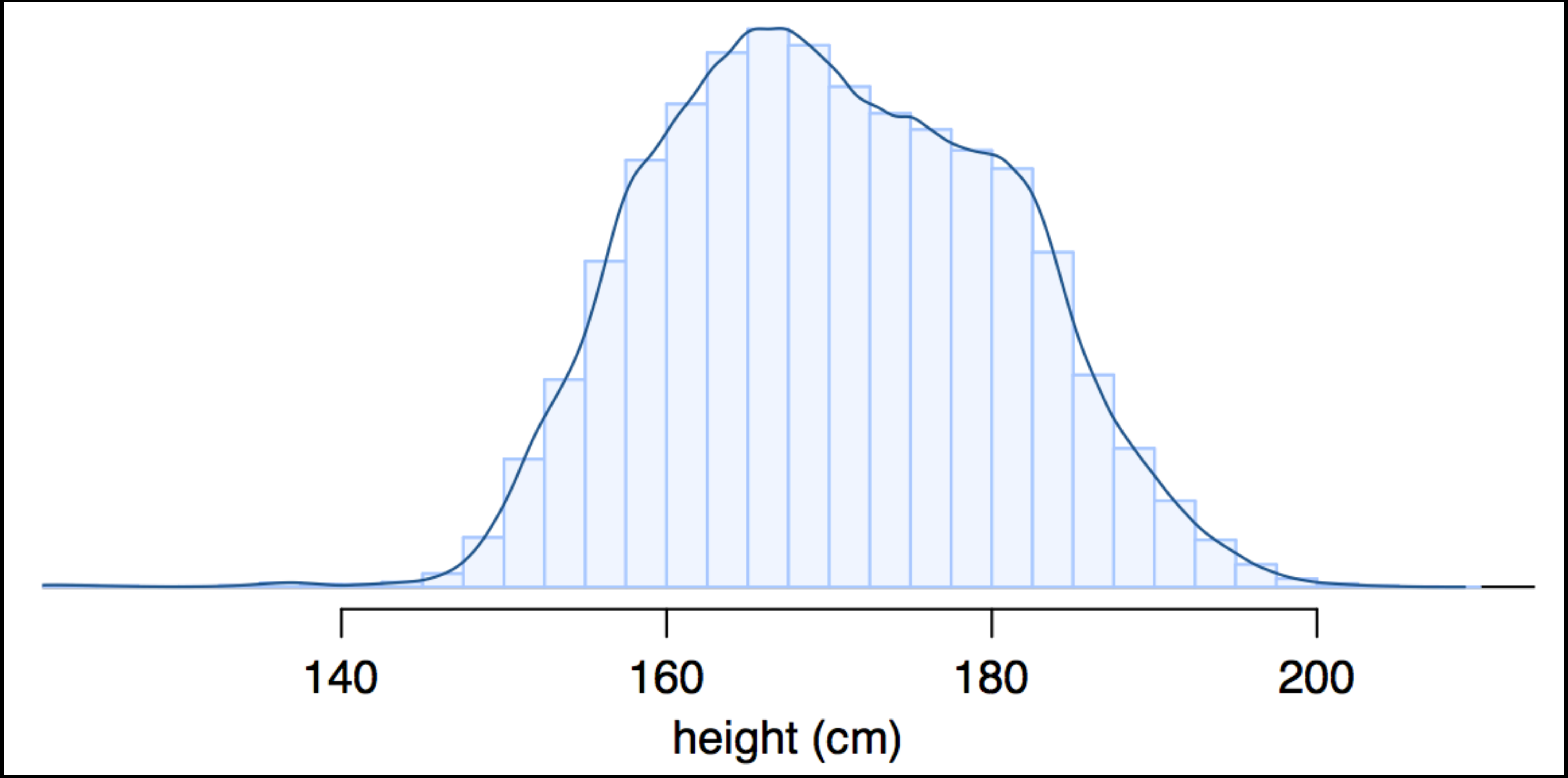
Moving on: Continuous distributions

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").



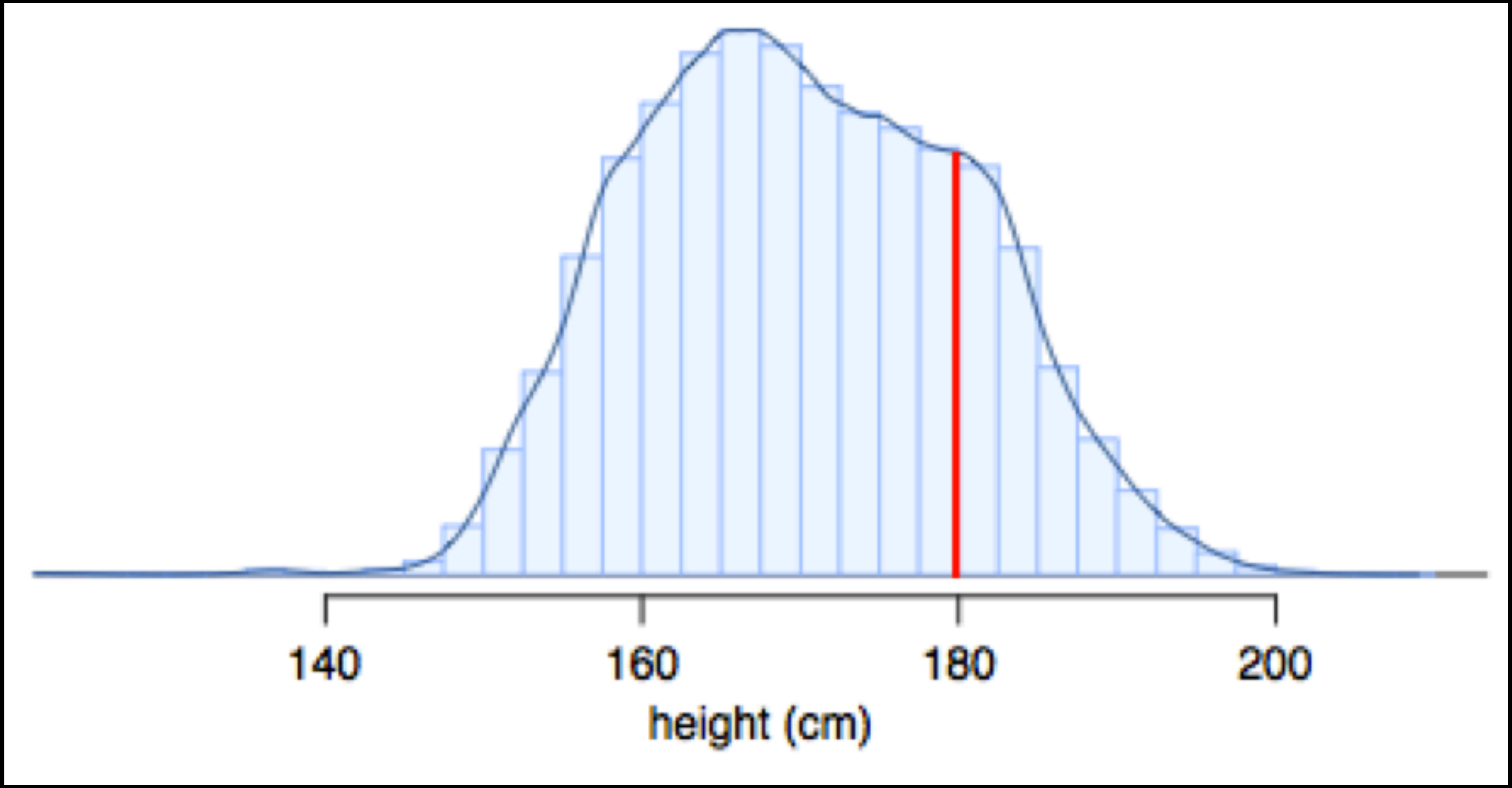
From histograms to continuous distributions

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



By definition...

Since continuous probabilities are estimated as “the area under the curve”, the probability of a person being exactly 180 cm (or any exact value) is defined as 0.



From discrete to continuous...

Discrete

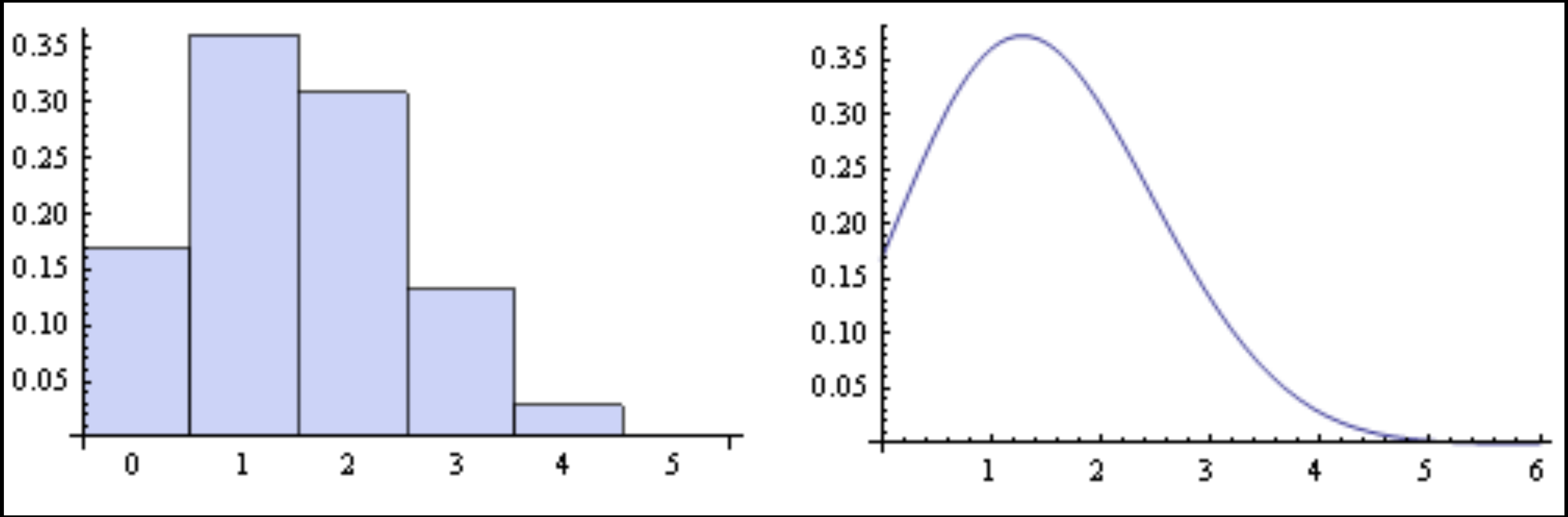
sum of probabilities must = 1

Continuous

total *area* must = 1

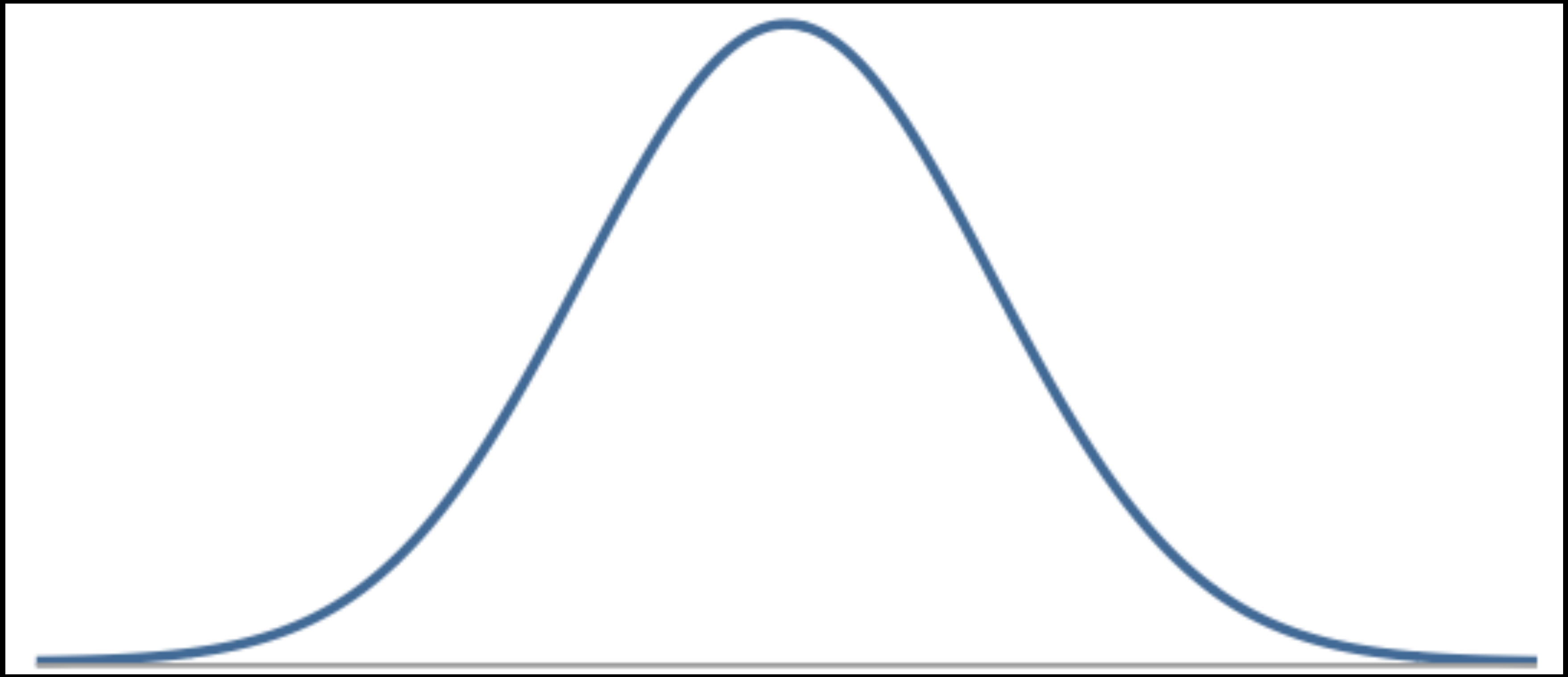
probability of a specific value = 0, e.g. $P(X = 2) = 0$

only intervals have probability, e.g. $P(1 < X < 2) = ?$



The Normal distribution

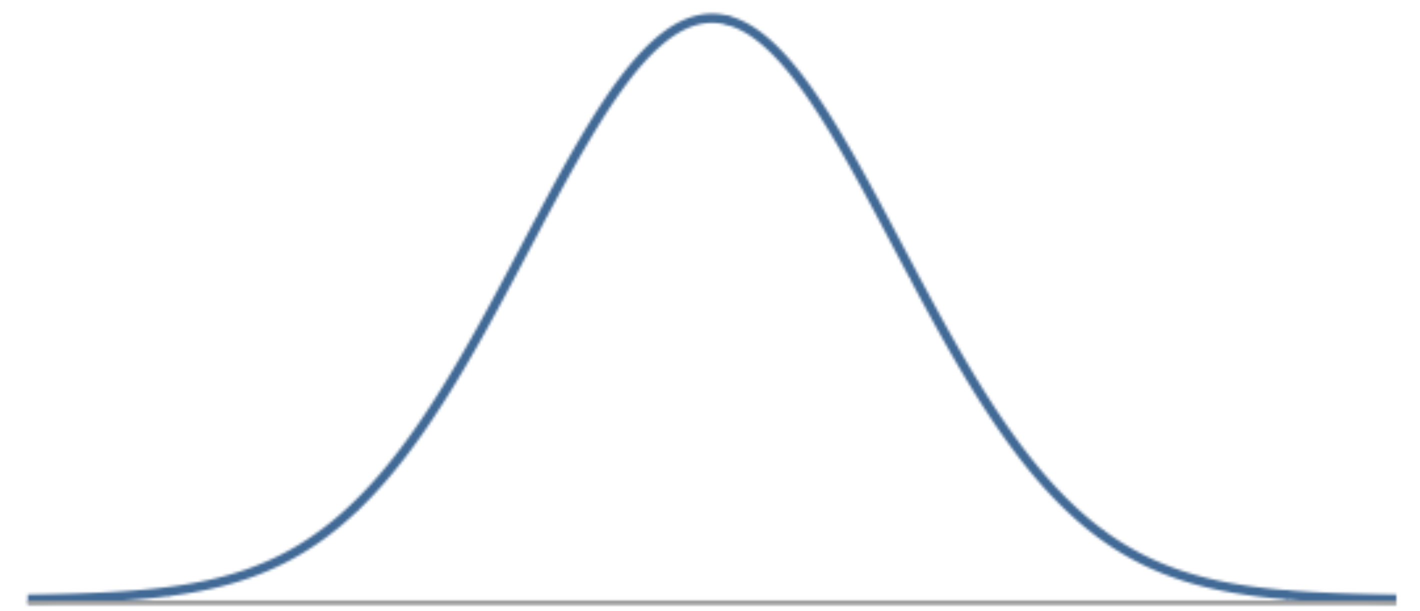
In Chapter 3, we look at the Normal distribution. The Normal distribution is the most famous continuous distribution.



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

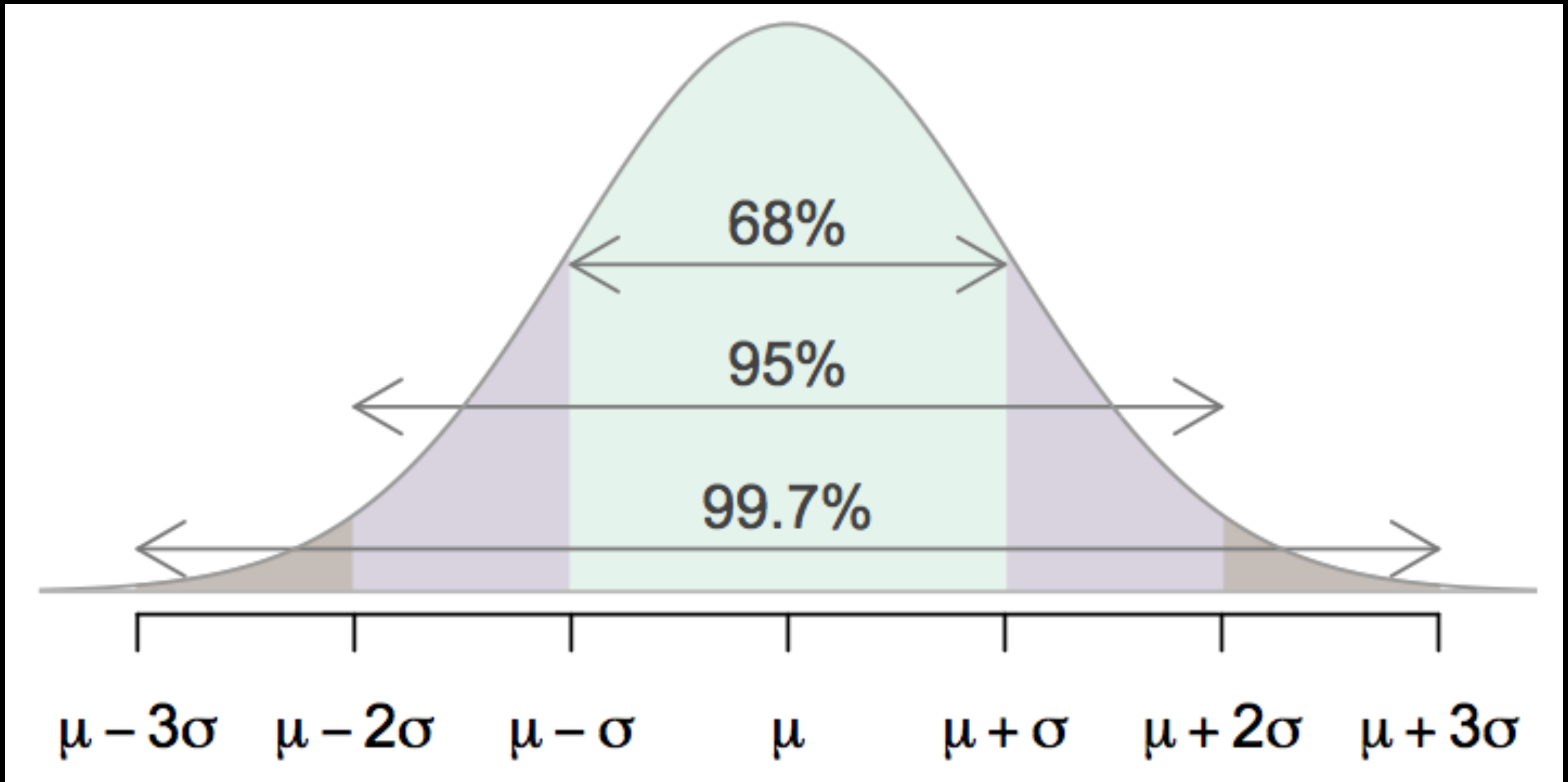
The Normal distribution...

is the most well known continuous distribution
Is unimodal and symmetric, bell shaped curve
has mean μ and standard deviation σ
has tails that extend infinitely in both directions
Many variables are nearly normal, but none are *exactly* normal



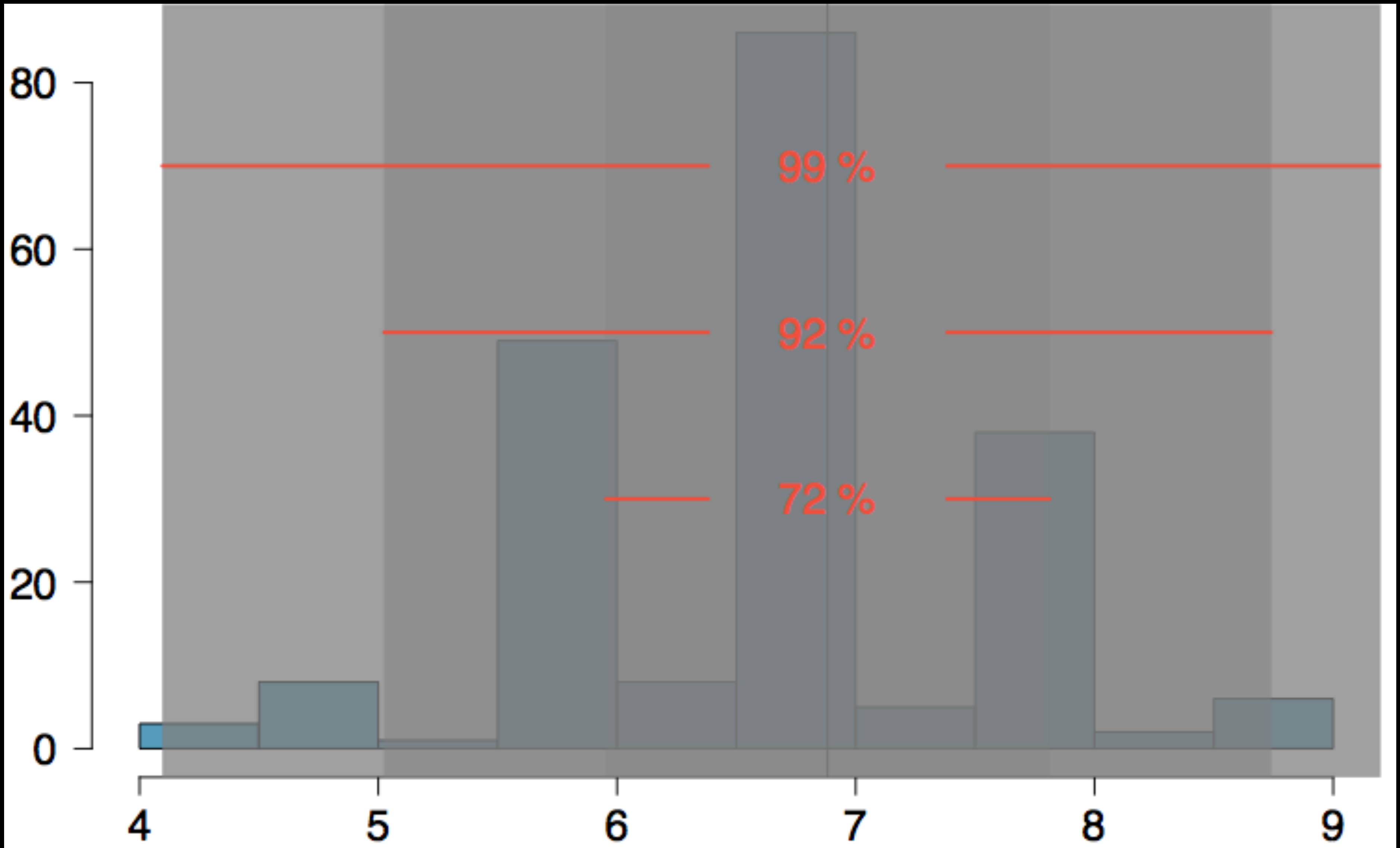
The source of the 68-95-99.7 Rule

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

Number of hours of sleep on school nights



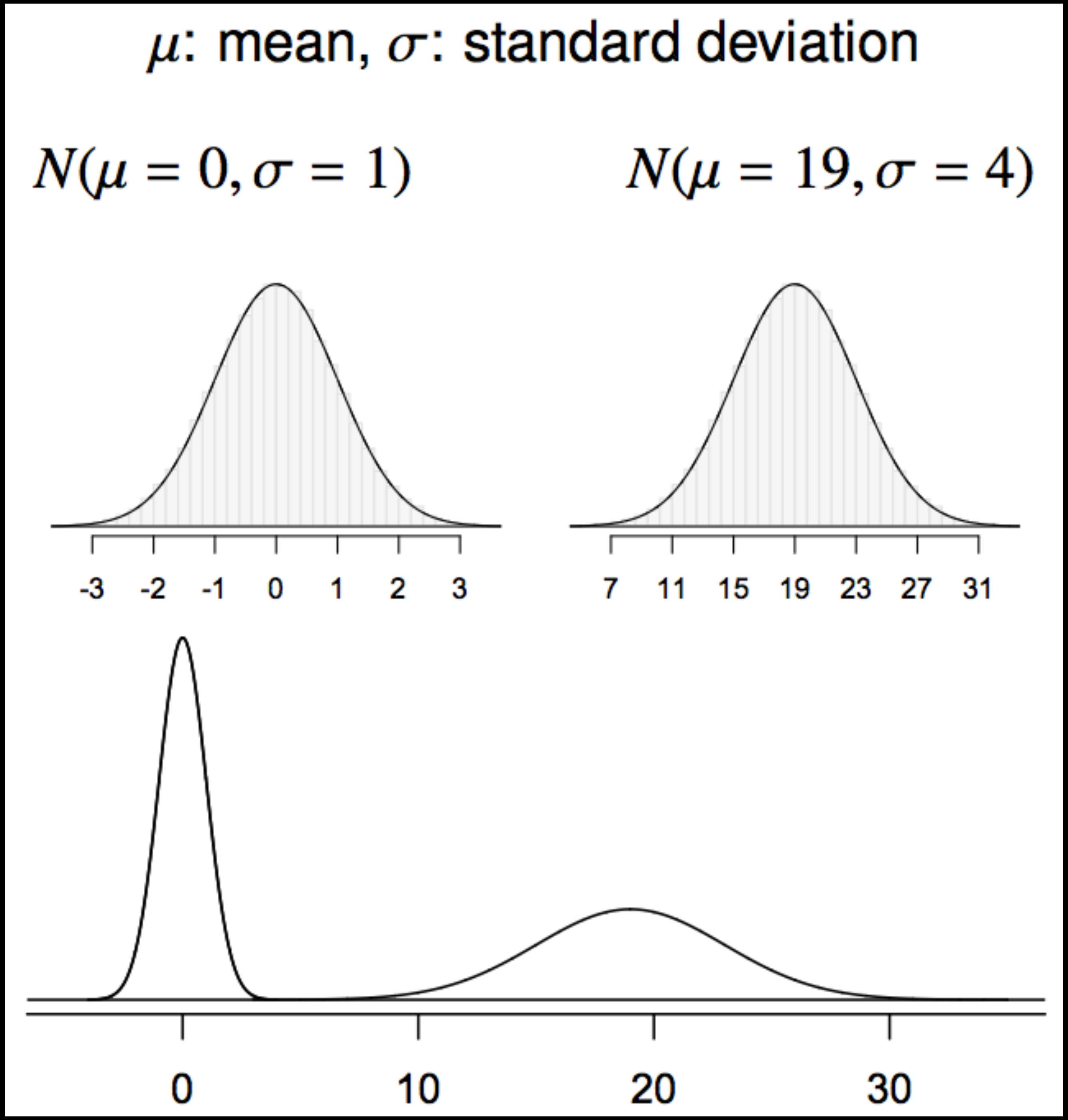
Mean = 6.88 hours, SD = 0.92 hrs

72% of the data are within 1 SD of the mean: 6.88 ± 0.93

92% of the data are within 2 SD of the mean: 6.88 ± 2 x 0.93

99% of the data are within 3 SD of the mean: 6.88 ± 3 x 0.93

Normal distributions with different parameters



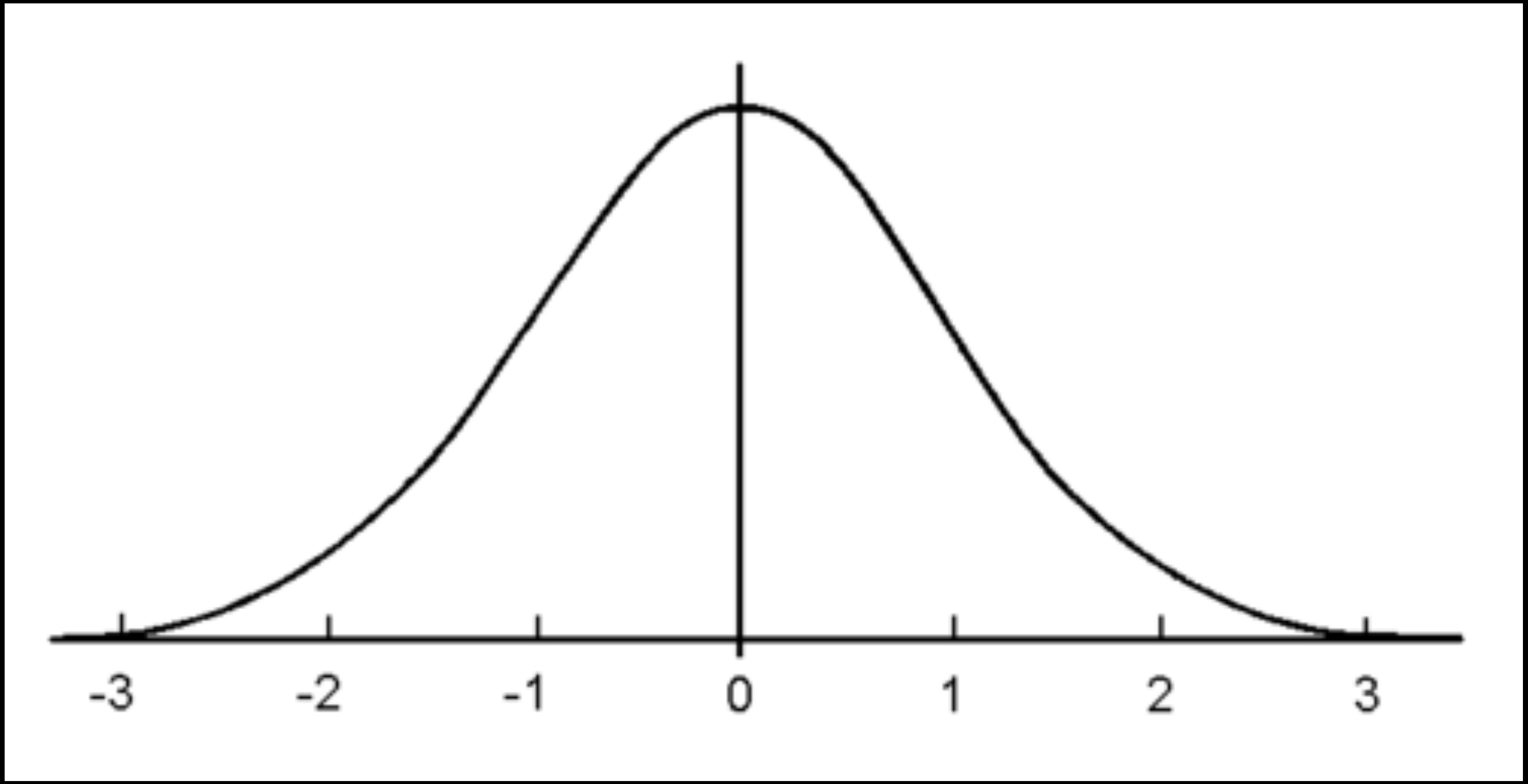
Why study the normal distribution?

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.

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

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

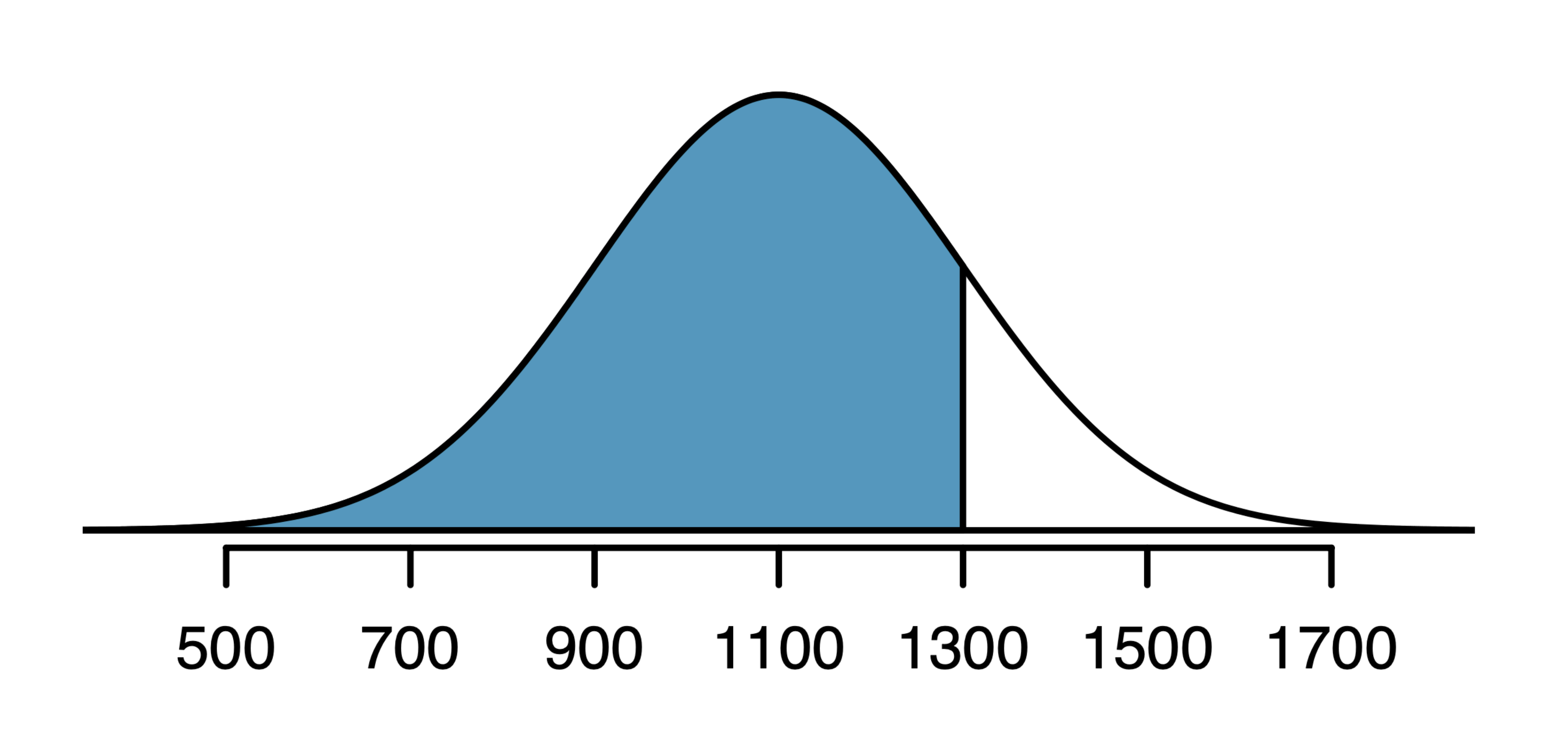
Observations that are more than 2 SD away from the mean ($|Z| > 2$) are generally considered unusual.

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.



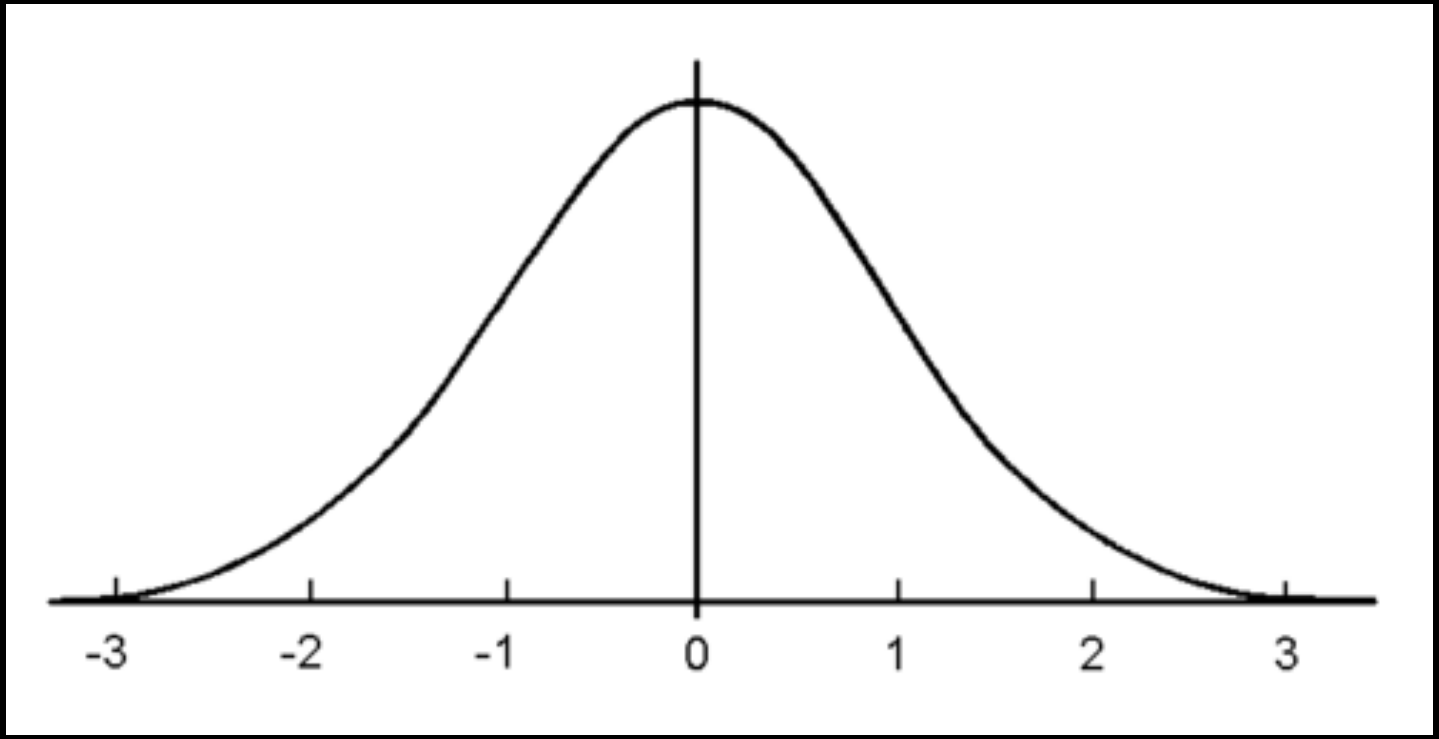
In R!

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 =

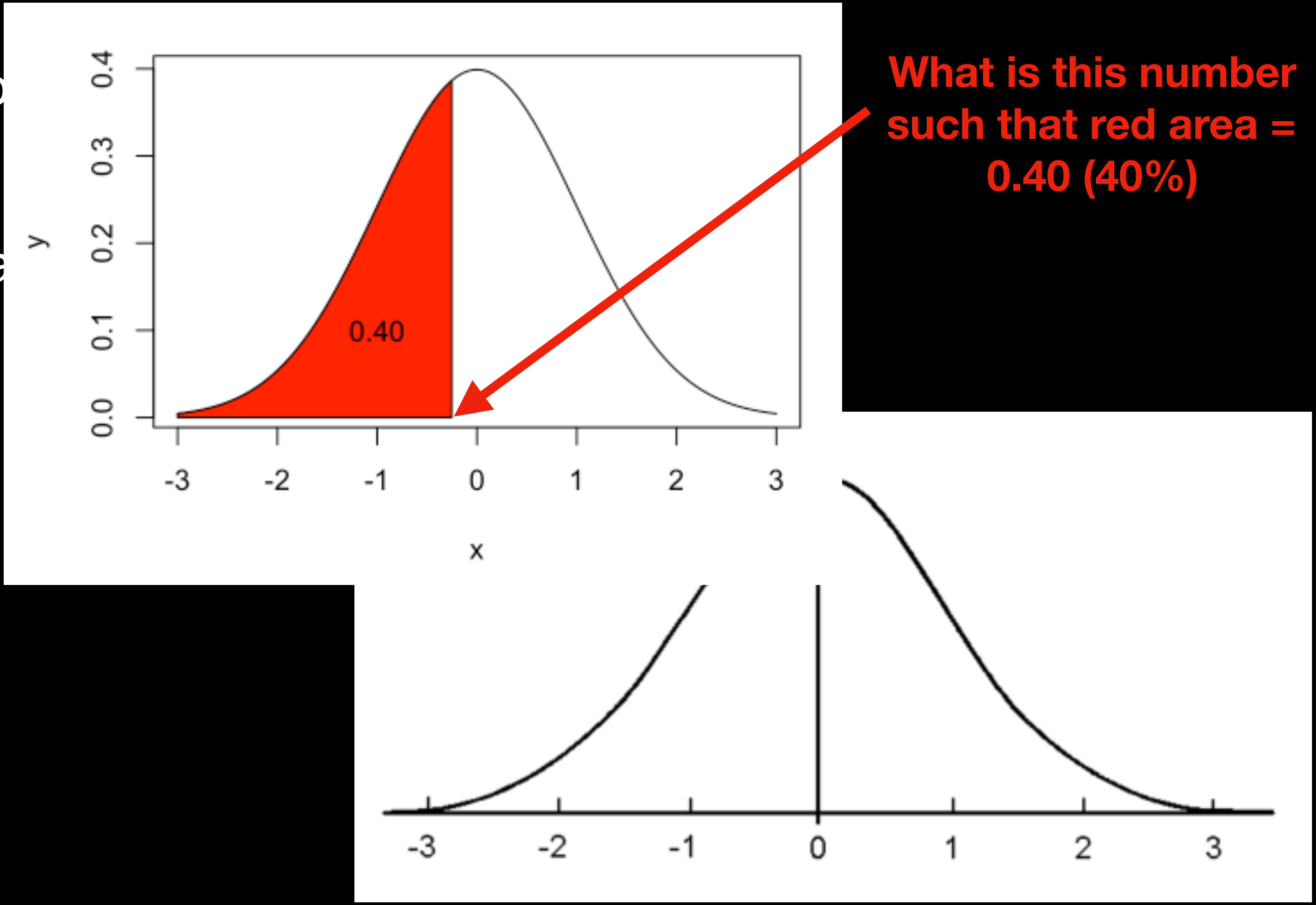


Finding percentiles from the standard normal curve

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

What Z-score co
i.e. P(Z < ?) = 0.

What Z-score ha
i.e. P(Z < ?) = **0.**

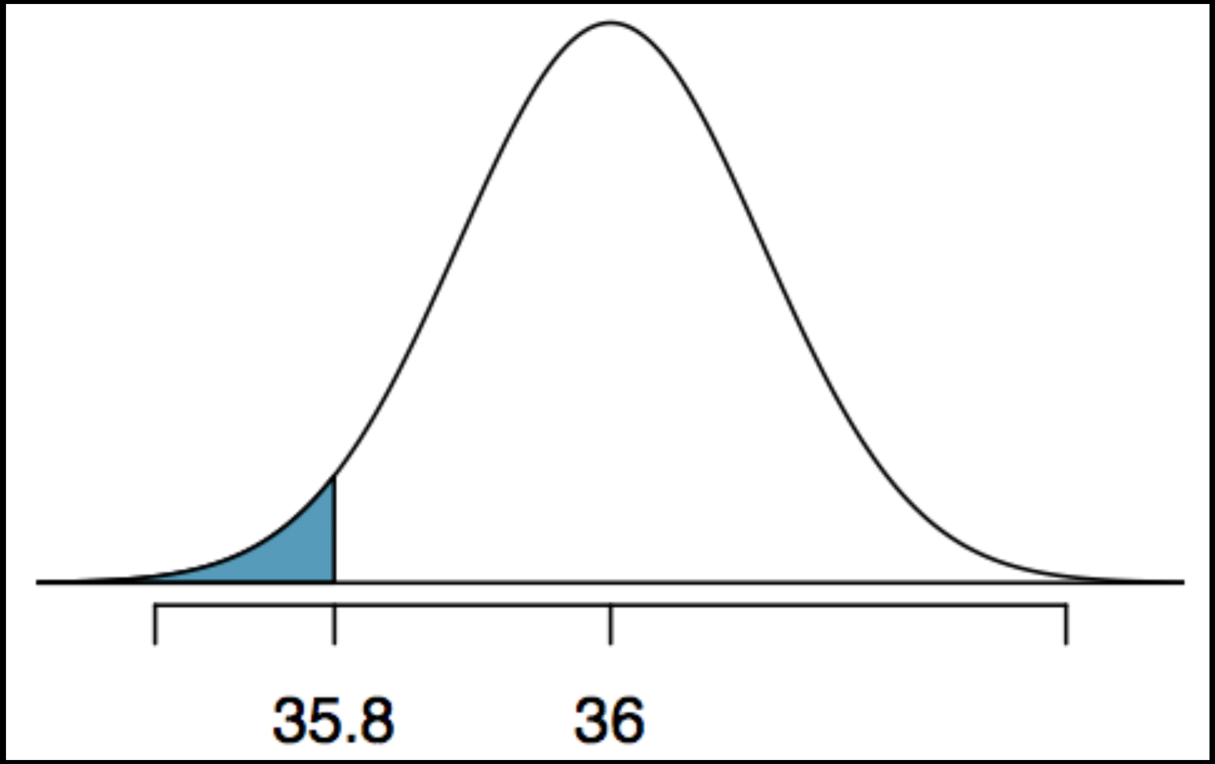


What is this number such that red area = 0.40 (40%)

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)?

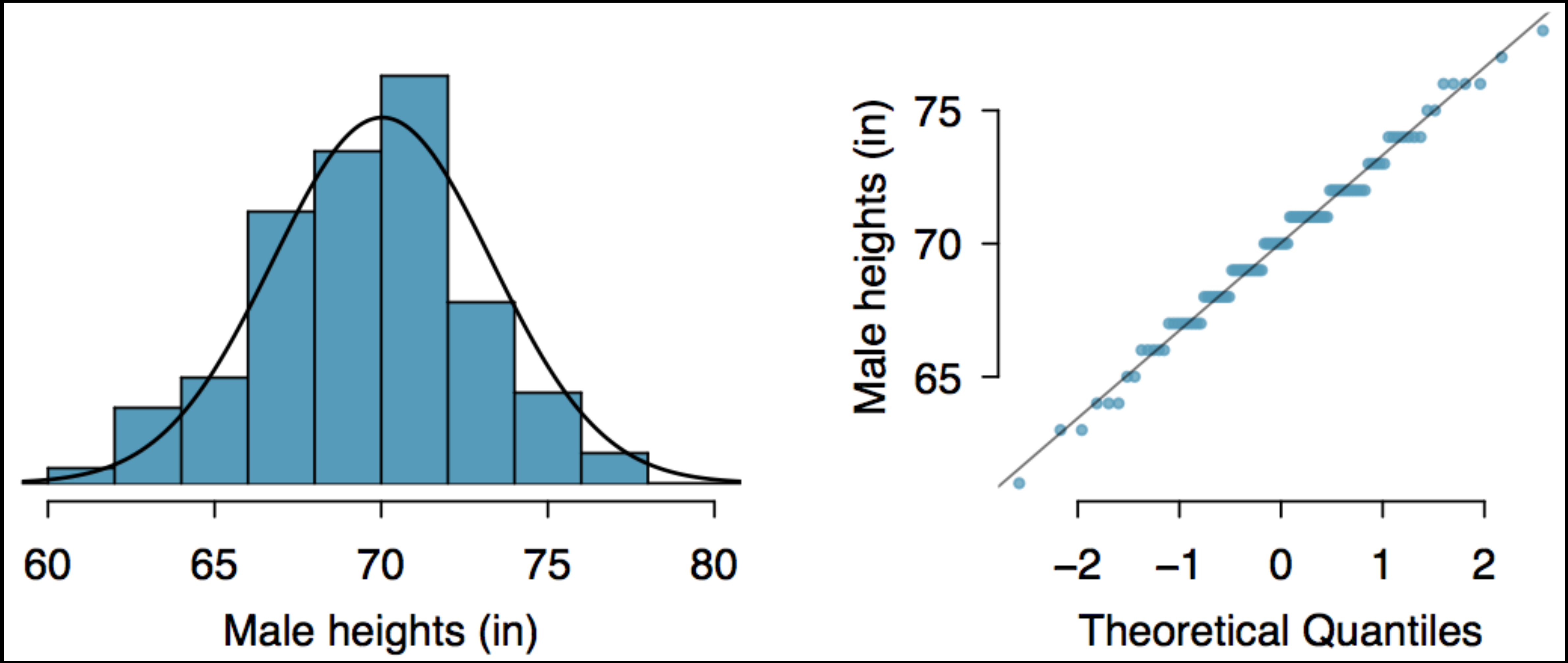


Try in R or by hand
(you can just write out what you'd calculate)

Is it Normal? The Normal probability plot

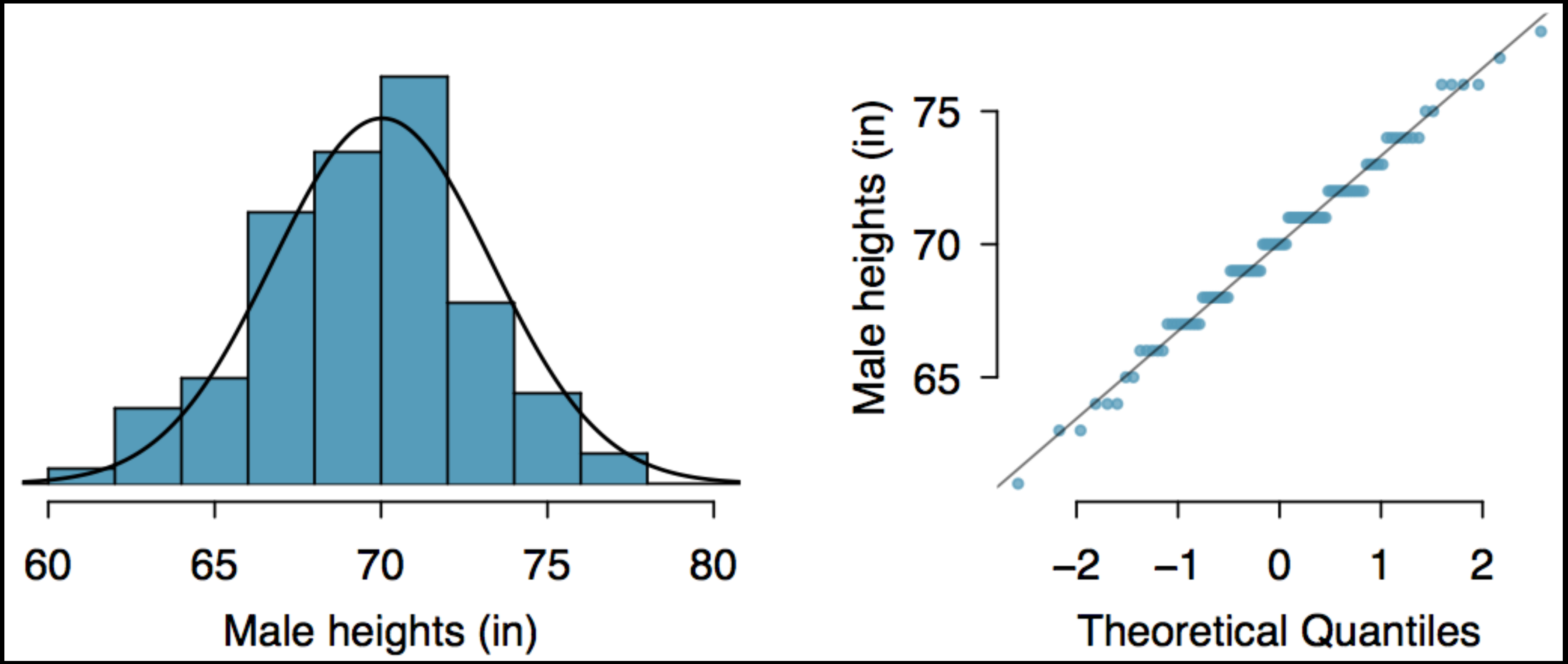
Is it Normal? The Normal probability plot

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



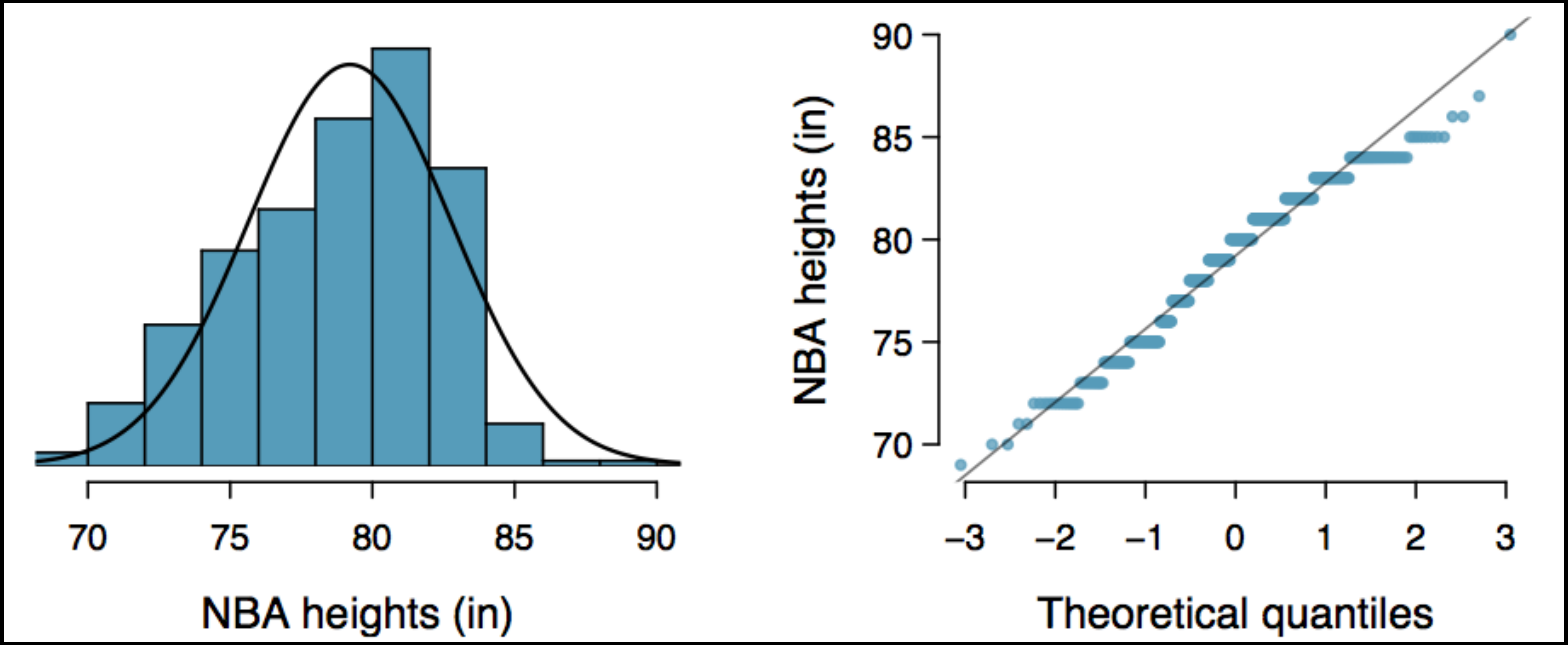
Anatomy of a normal probability plot

- 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.



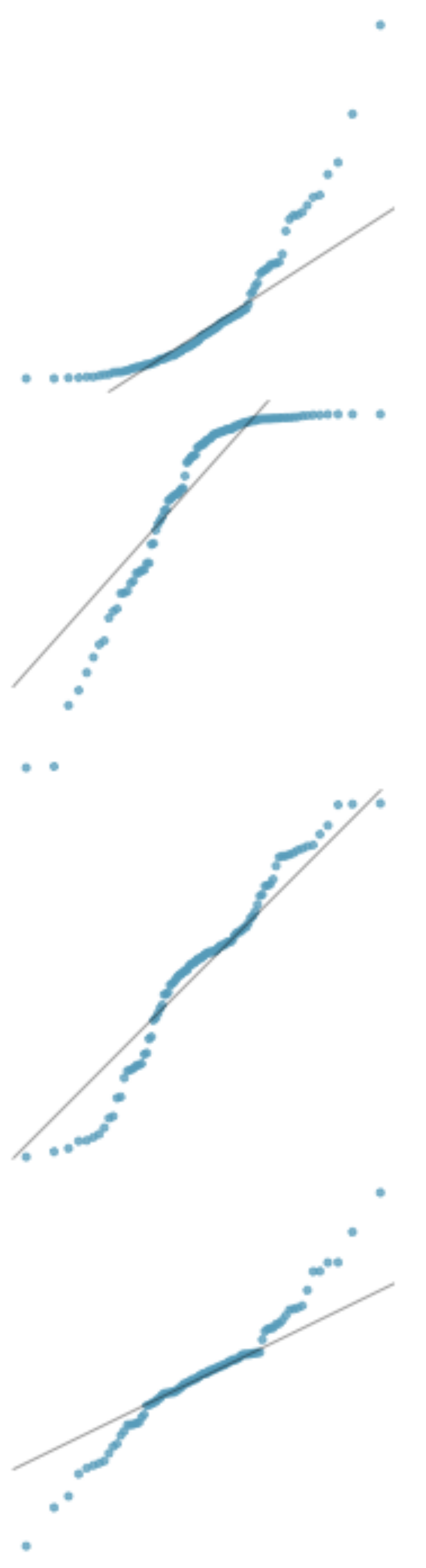
Practice

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

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

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

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

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

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

= n*(n-1)*(n-2)*…*1

=(n-k)*(n-k-1)*…*1

“n choose k”

=k*(k-1)*(k-2)*…*1

The Binomial formula: Factorials

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

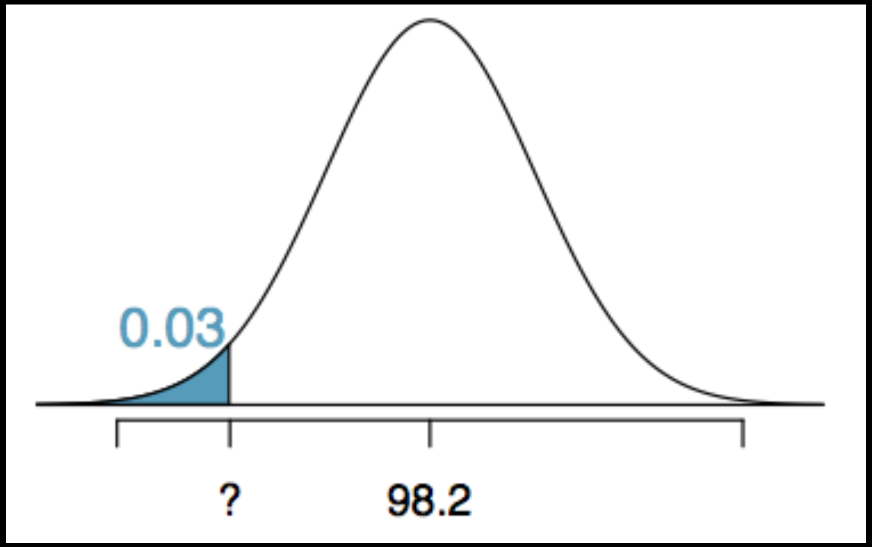
If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle *fails* the quality control inspection. Recall $\mu = 36$, $\sigma = 0.11$.

What percent of bottles pass the quality control inspection?

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

Finding cutoff points

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

(a) 97.3^{o}F

(b) 99.1^{o}F

(c) 99.4^{o}F

(d) 99.6^{o}F