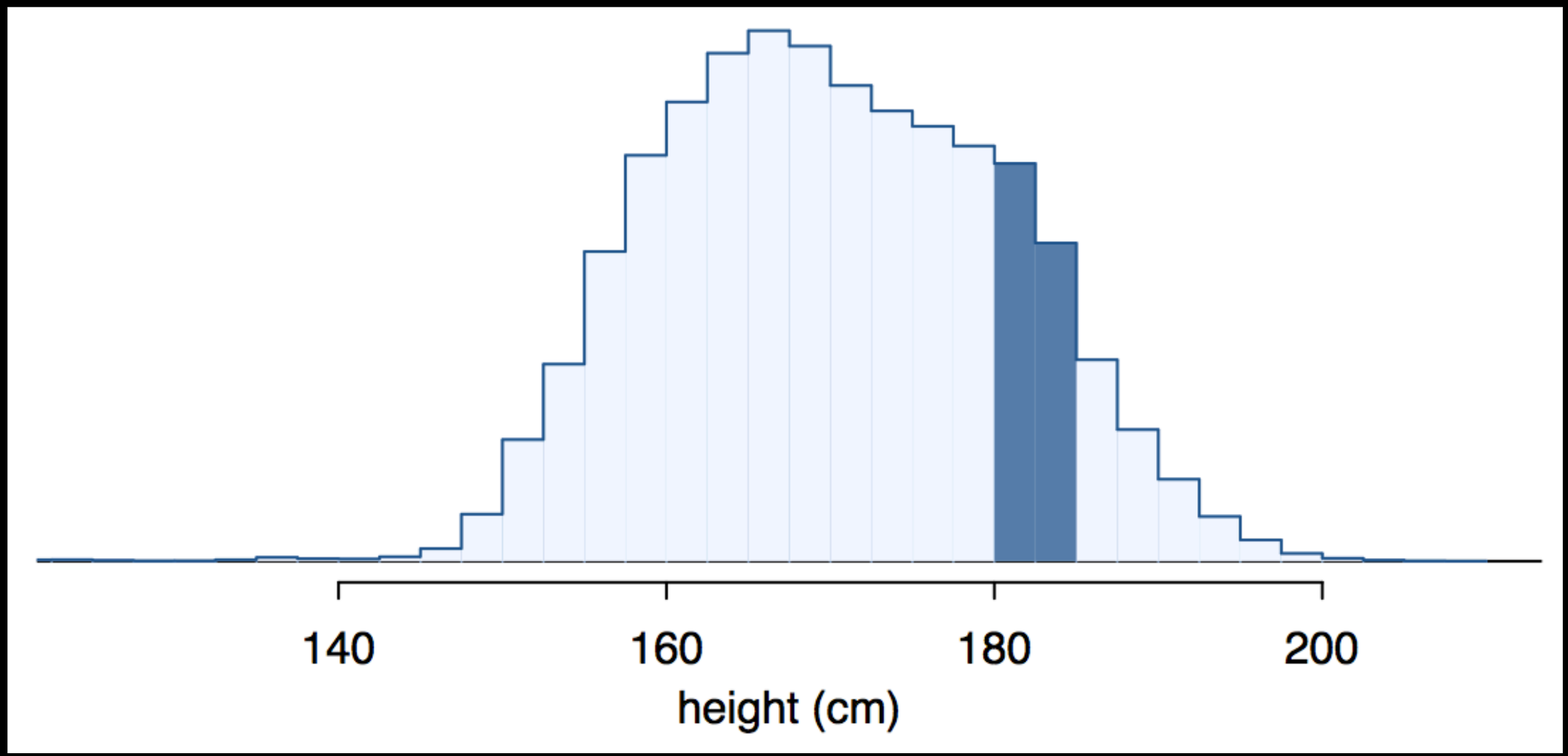
Moving on: Continuous distributions

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

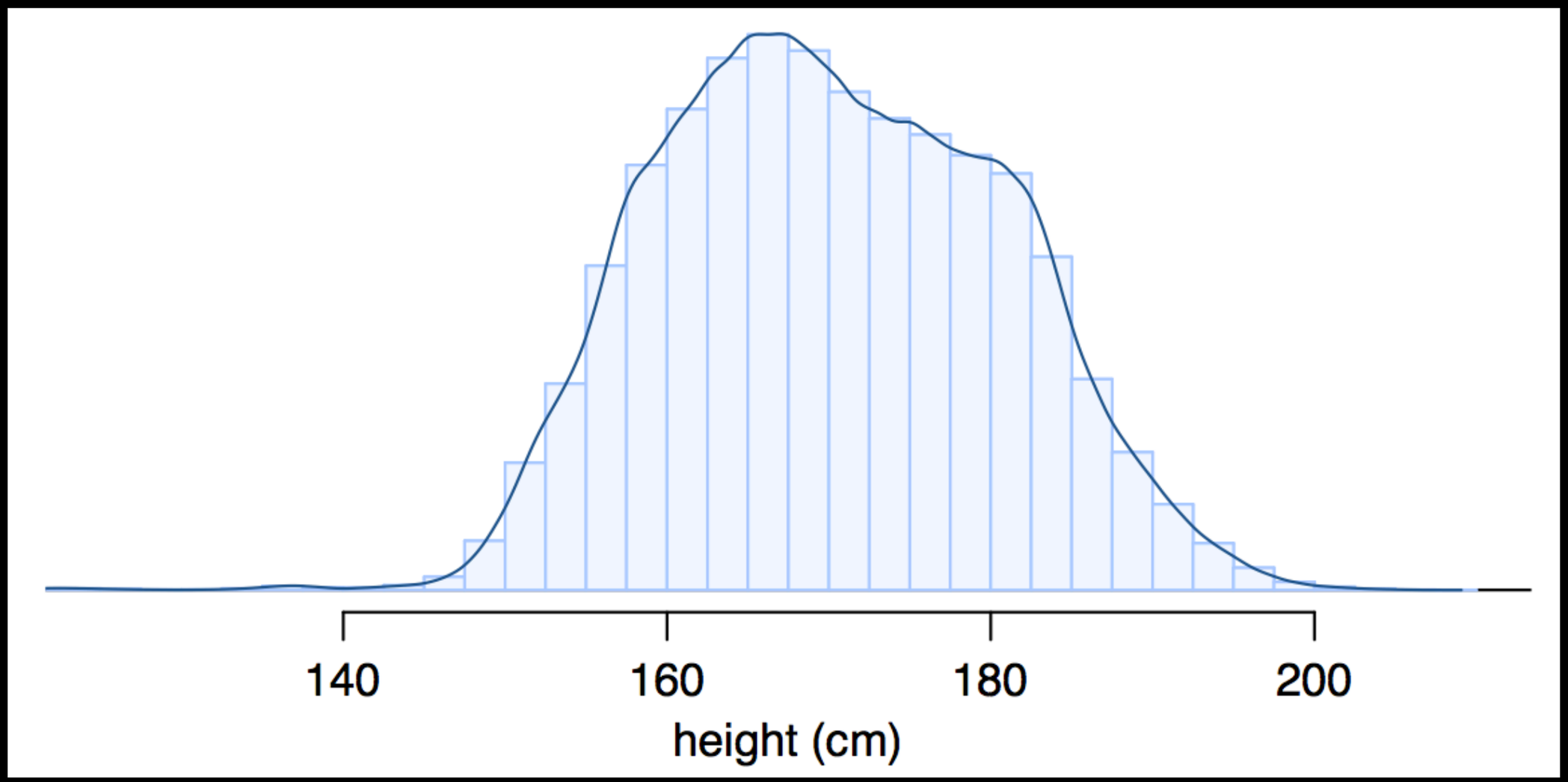
Moving on: Continuous distributions

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").



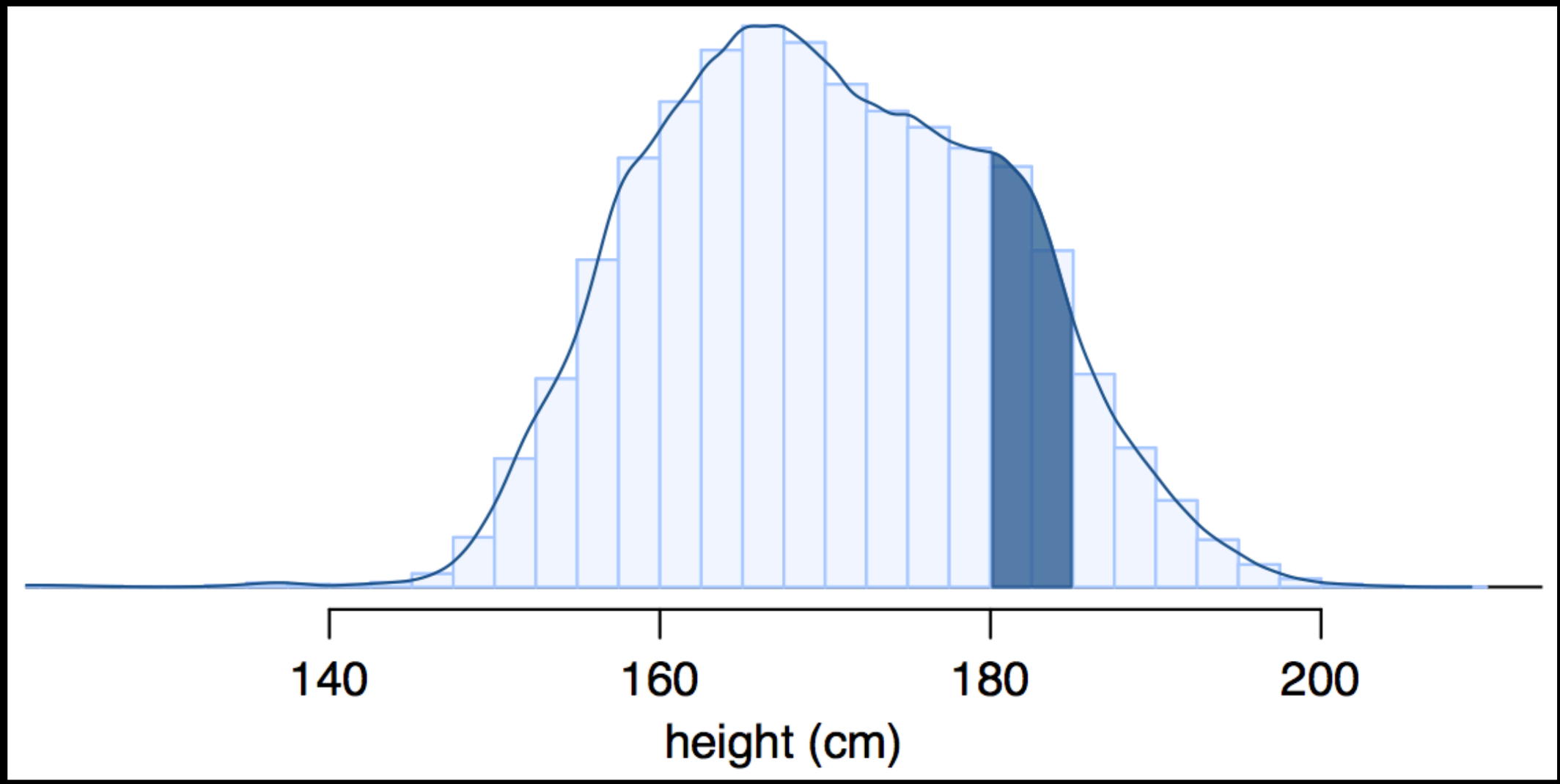
From histograms to continuous distributions

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



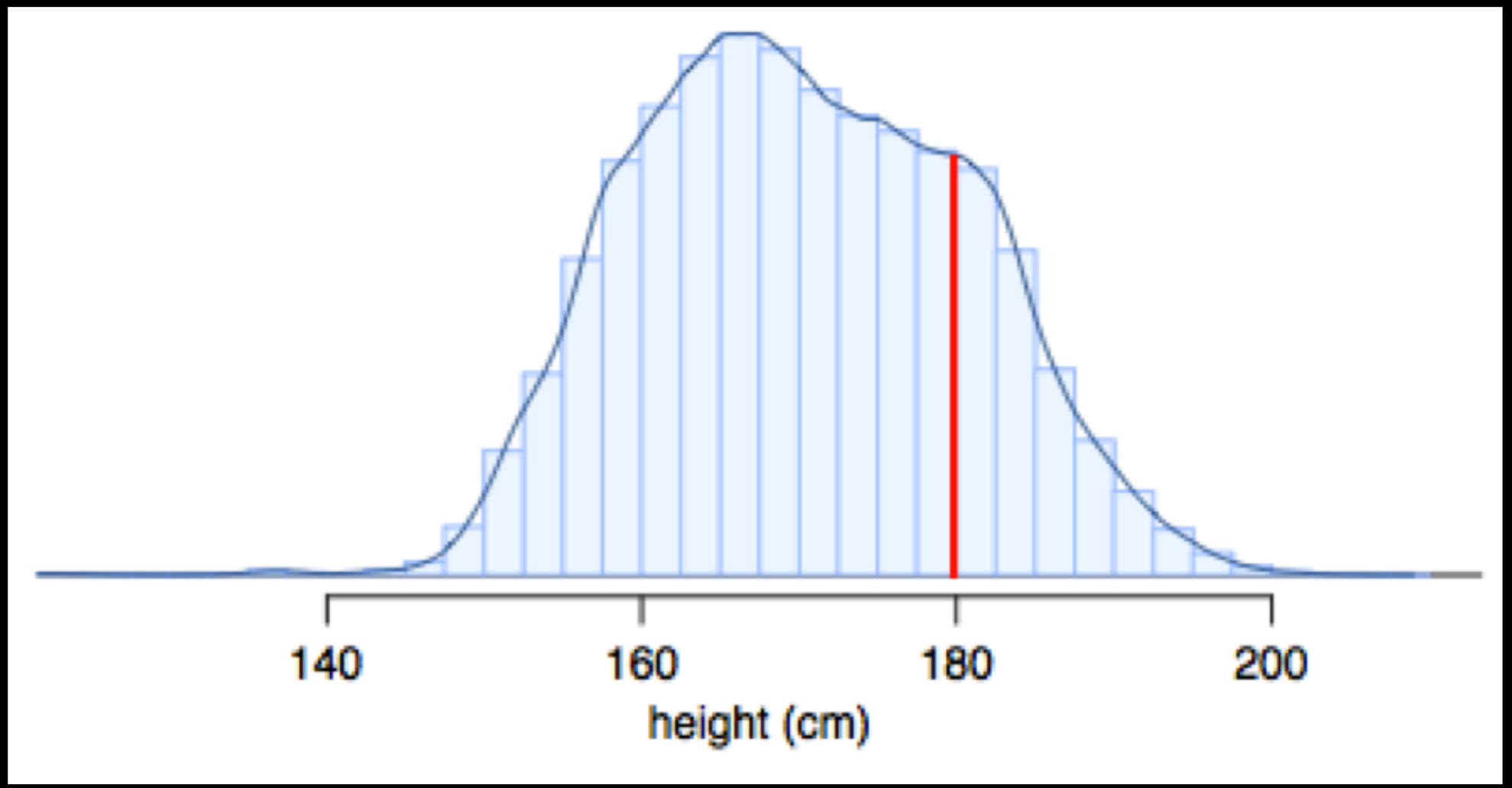
Probabilities from continuous distributions

Therefore, the probability that a randomly sampled US adult is between 180 cm and 185 cm can also be estimated as the shaded area under the curve.



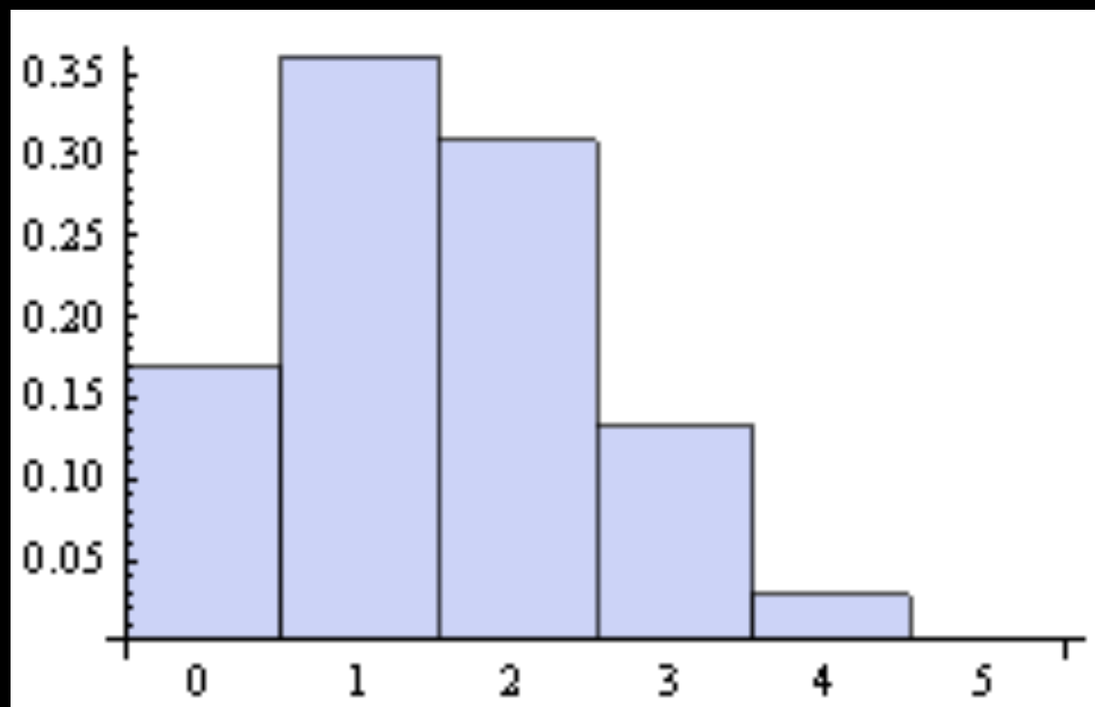
By definition...

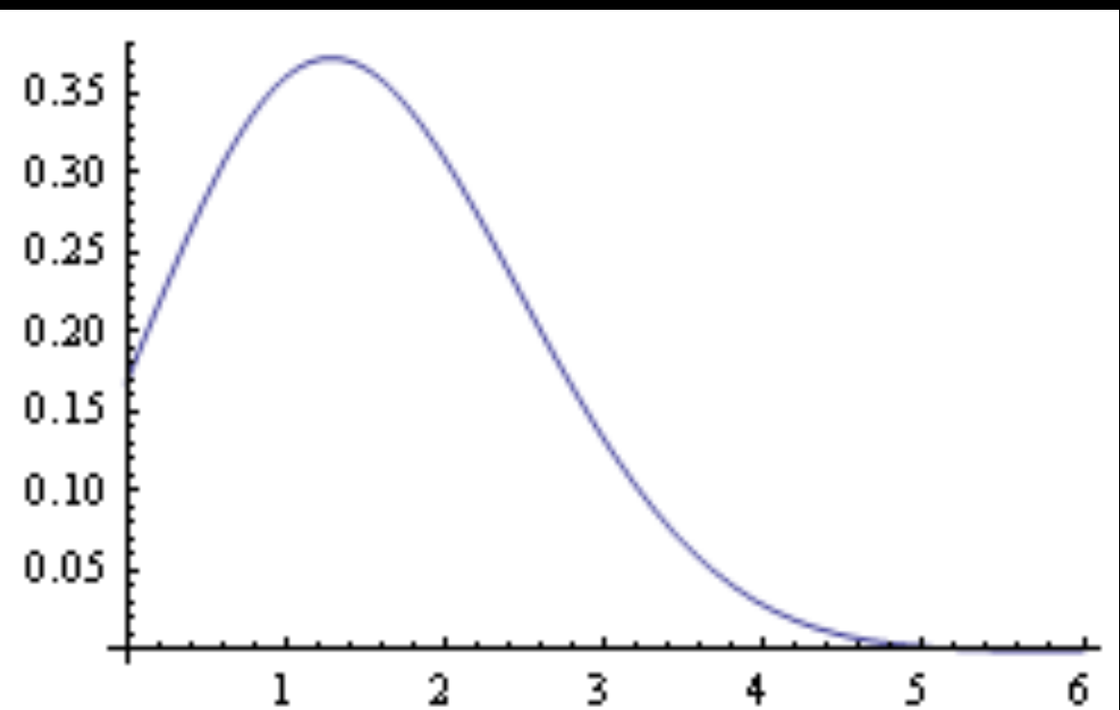
Since continuous probabilities are estimated as “the area under the curve”, the probability of a person being exactly 180 cm (or any exact value) is defined as 0.



From discrete to continuous...

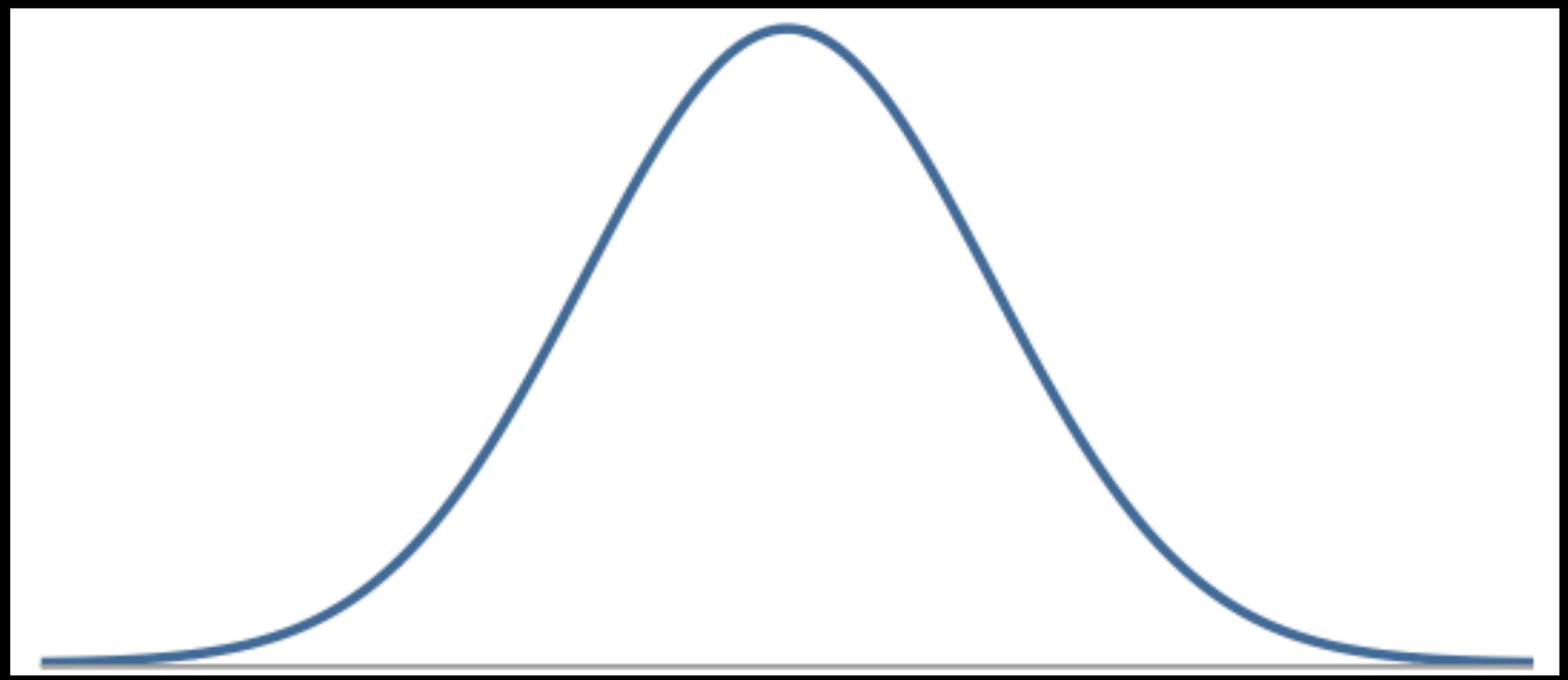
Discrete
sum of probabilities must = 1
Continuous
total *area* must = 1
probability of a specific value = 0, e.g. P(X = 2) = 0
only intervals have probability, e.g. P(1 < X < 2) = ?





The Normal distribution

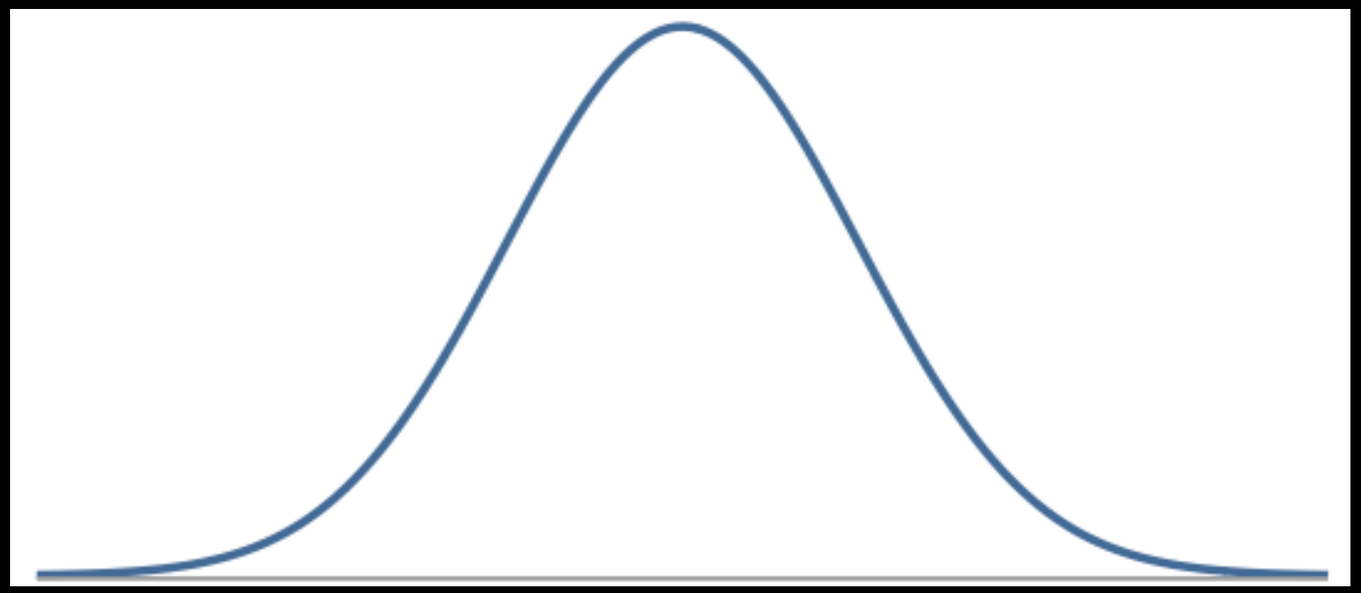
In Chapter 3, we look at the Normal distribution. The Normal distribution is the most famous continuous distribution.



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

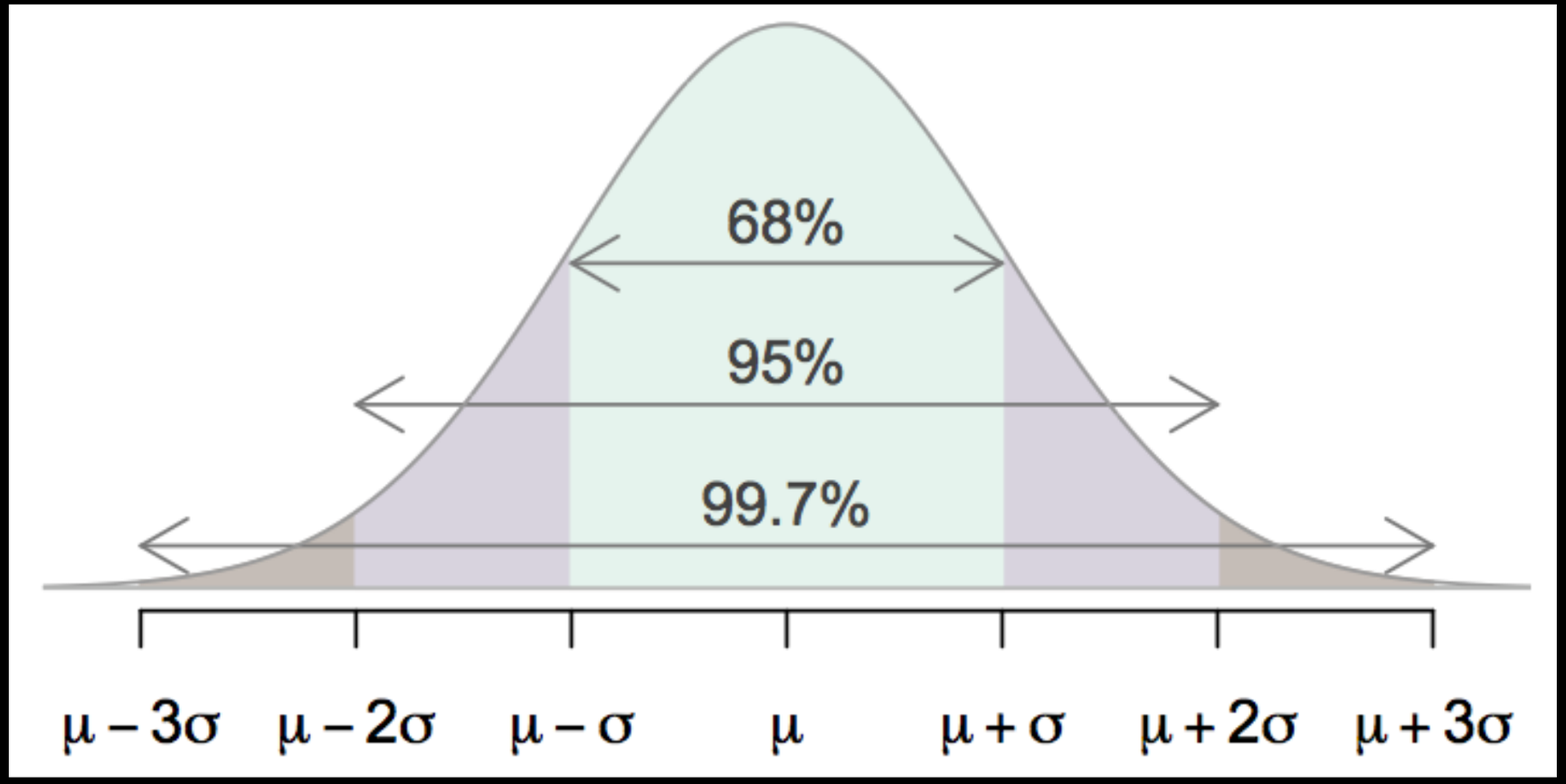
The Normal distribution...

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



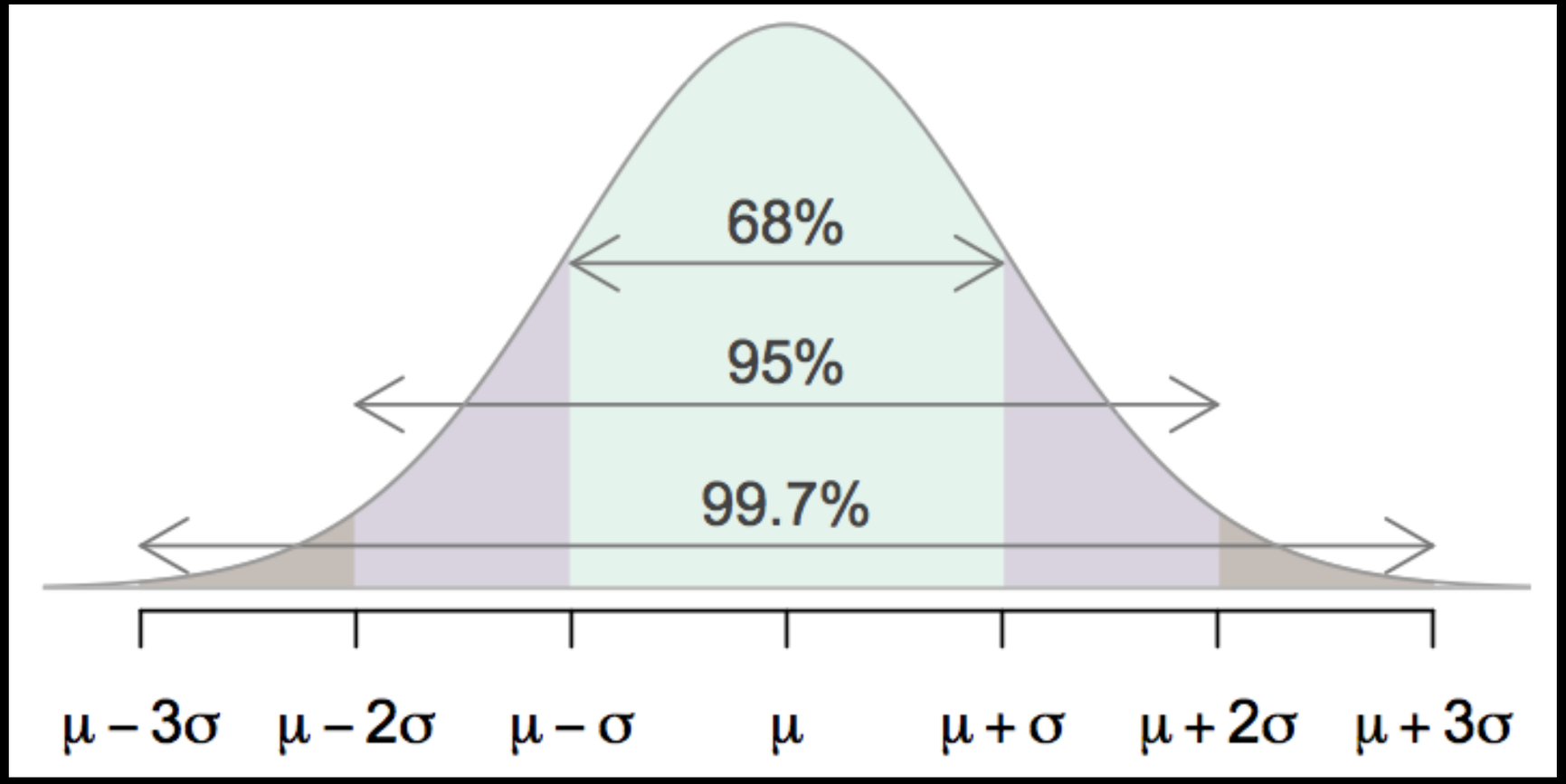
The source of the 68-95-99.7 Rule

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.



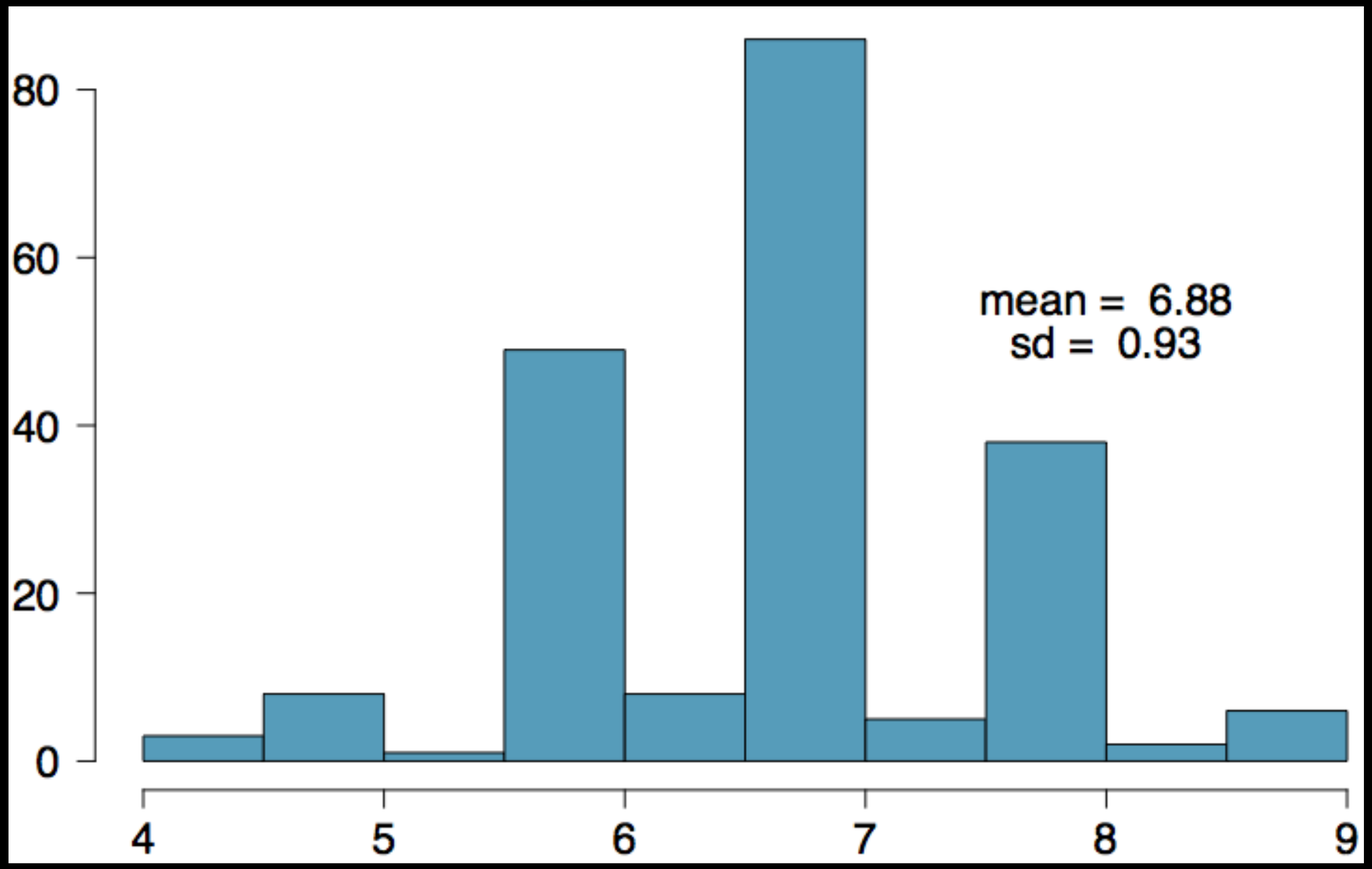
The source of the 68-95-99.7 Rule

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.



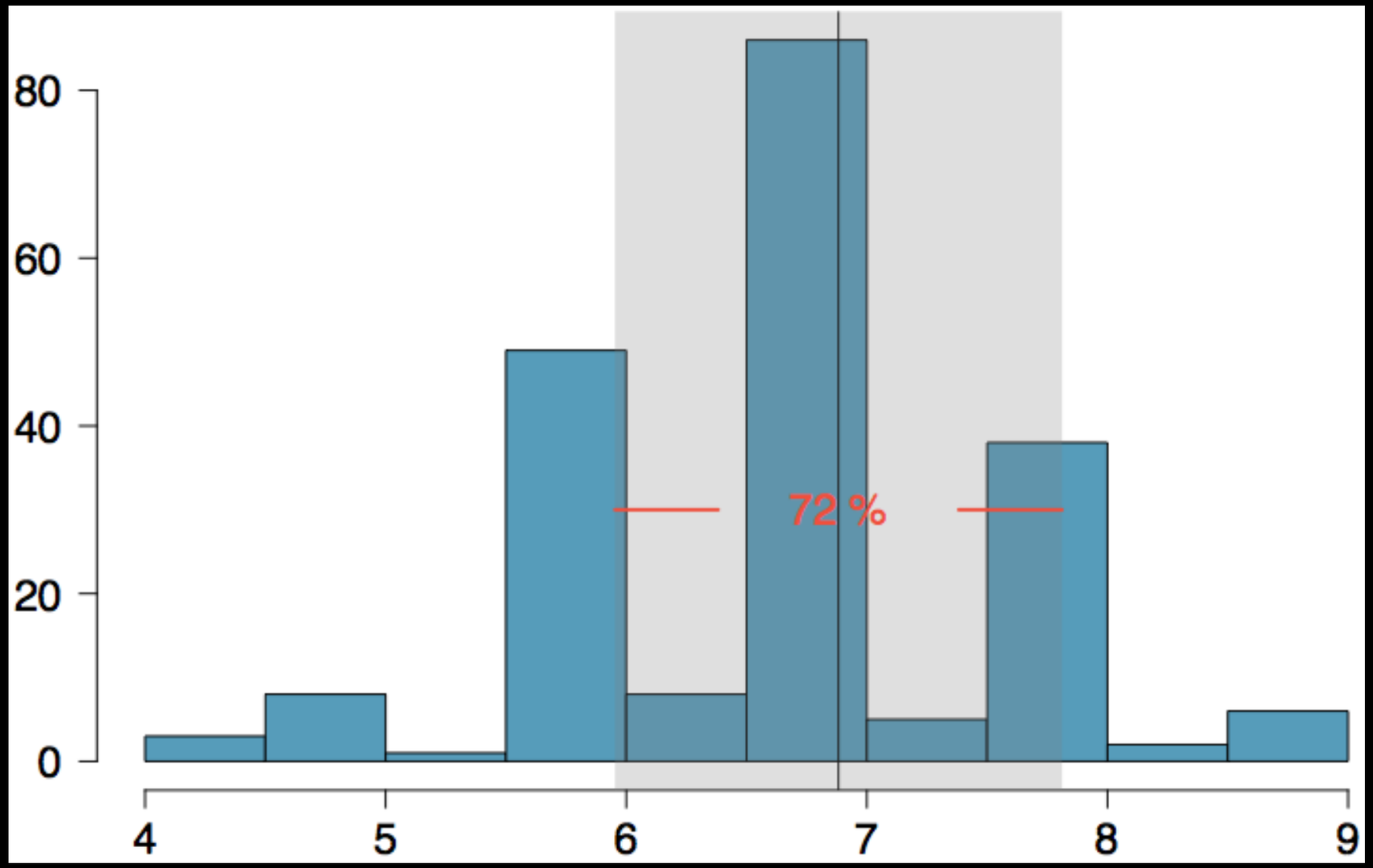
And the total area under the curve = 1

Number of hours of sleep on school nights



Mean = 6.88 hours, SD = 0.92 hrs

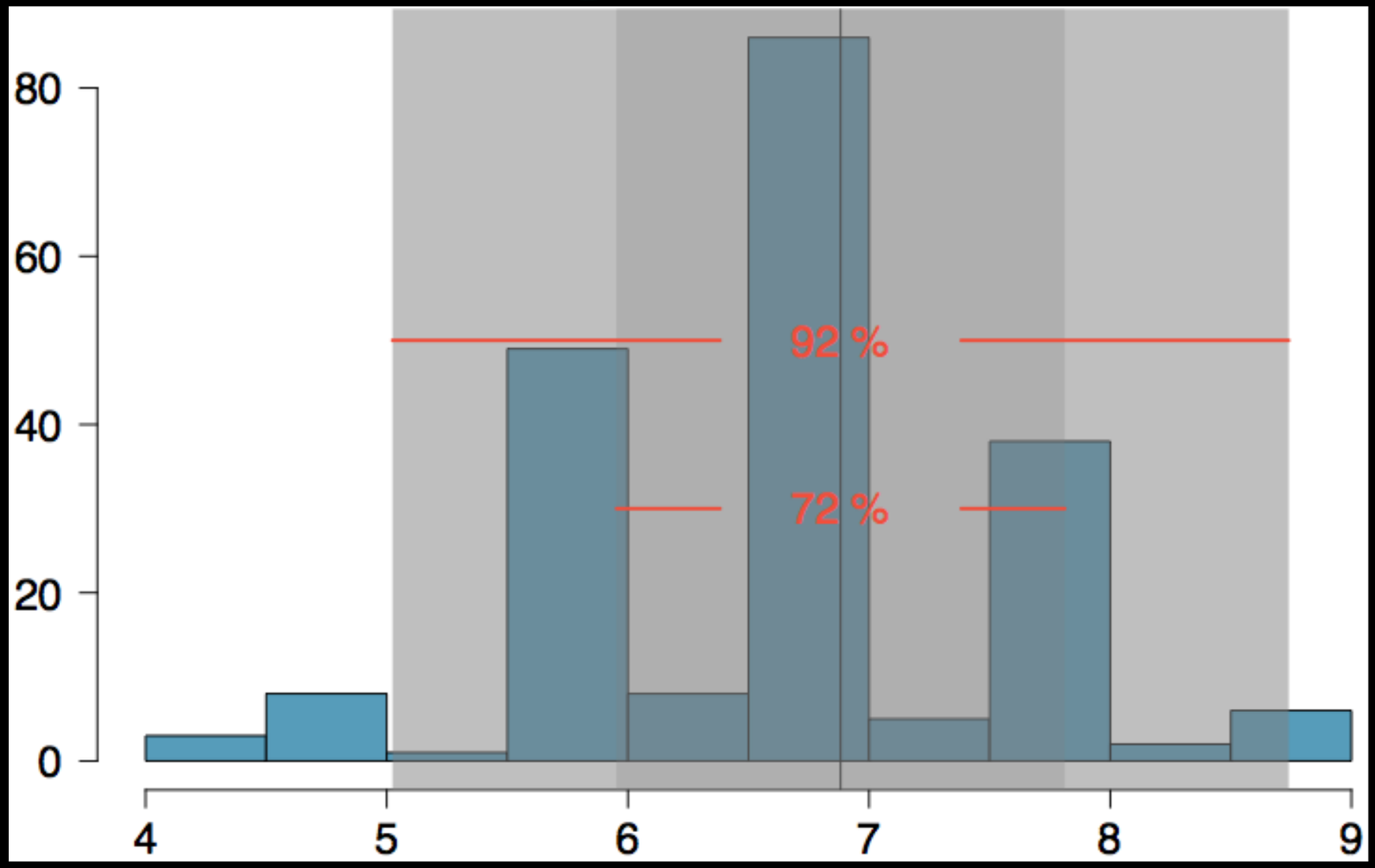
Number of hours of sleep on school nights



Mean = 6.88 hours, SD = 0.92 hrs

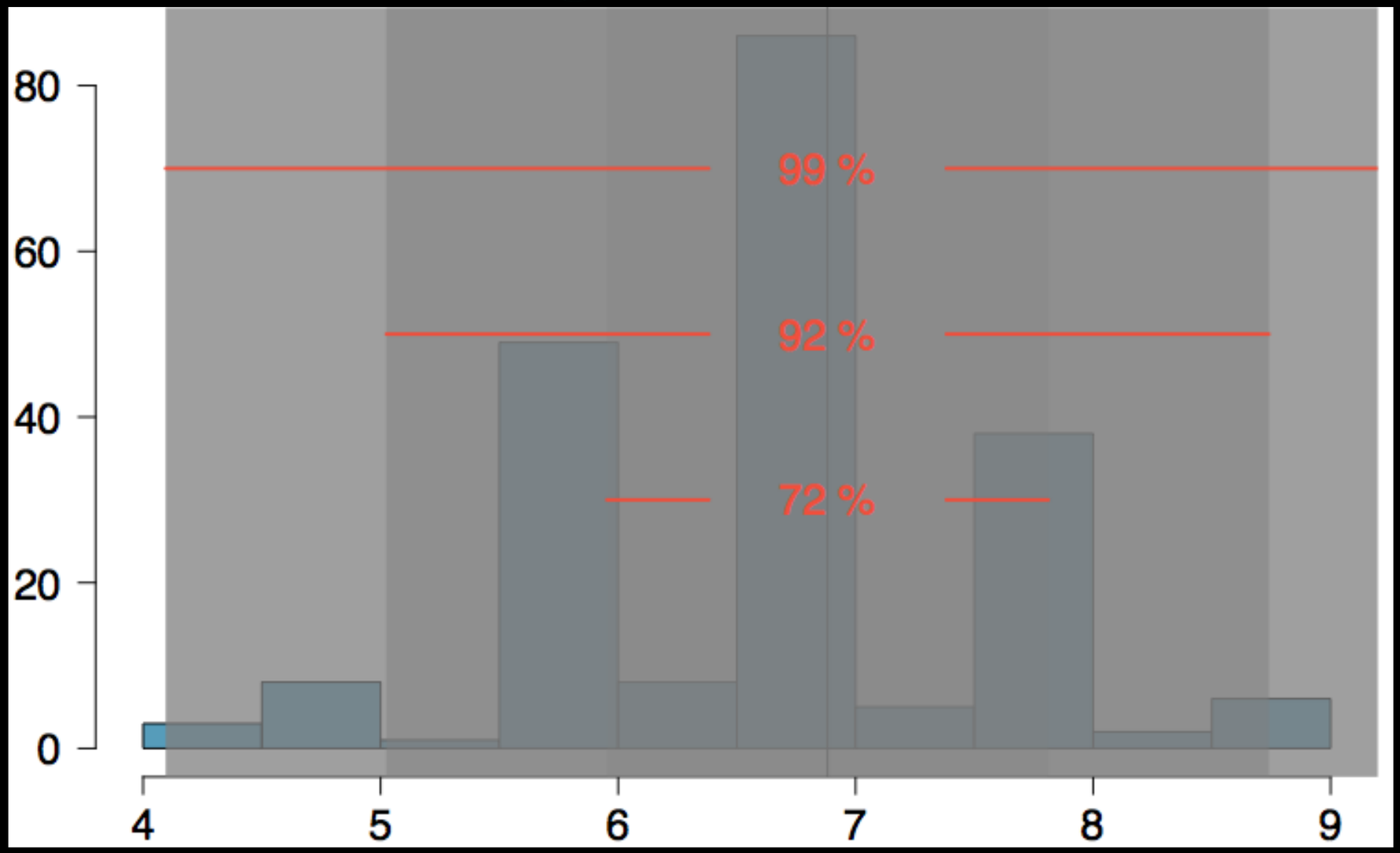
72% of the data are within 1 SD of the mean: 6.88 ± 0.93

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

Number of hours of sleep on school nights

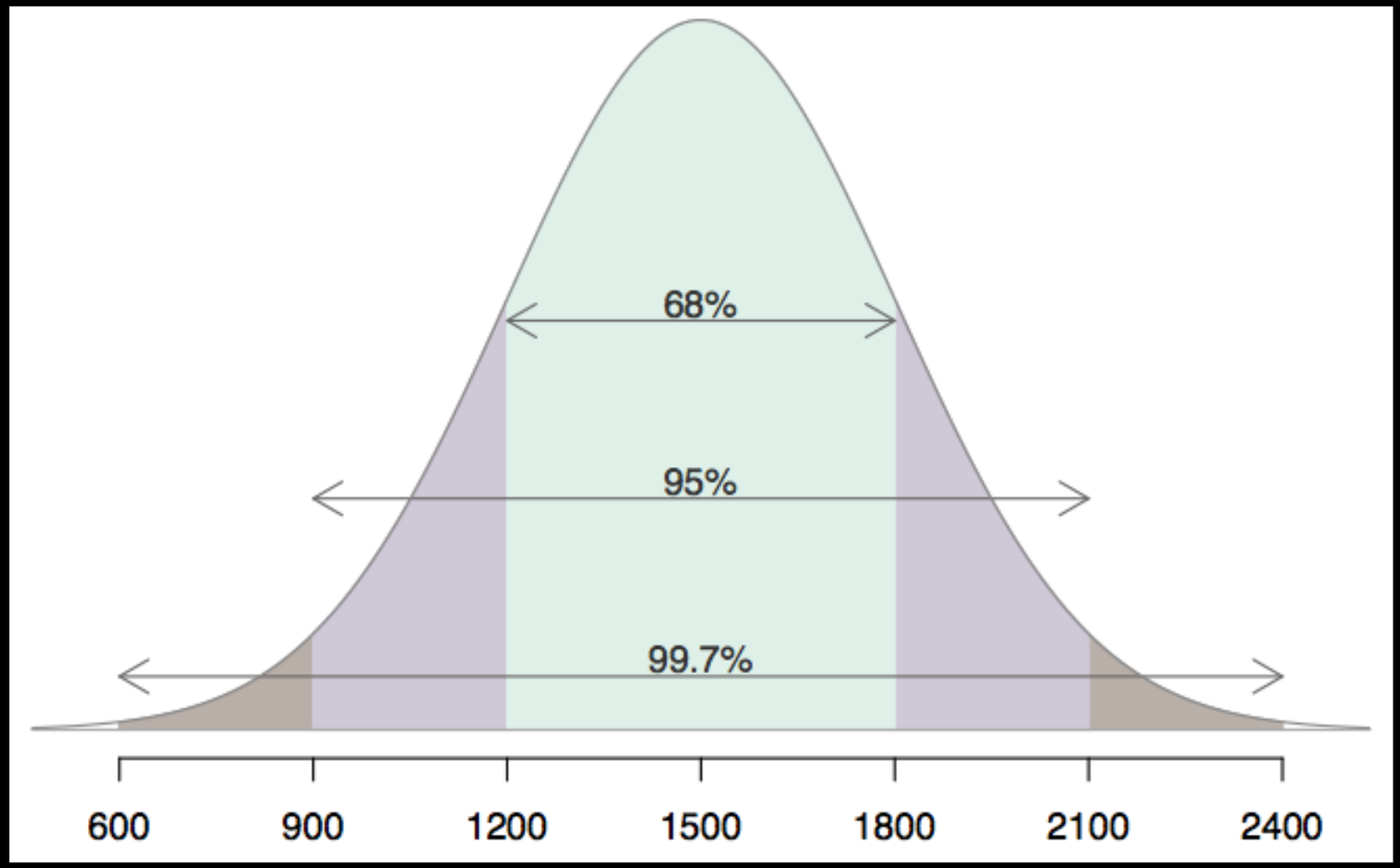


Mean = 6.88 hours, SD = 0.92 hrs
72% of the data are within 1 SD of the mean: 6.88 ± 0.93
92% of the data are within 2 SD of the mean: 6.88 ± 2 x 0.93
99% of the data are within 3 SD of the mean: 6.88 ± 3 x 0.93

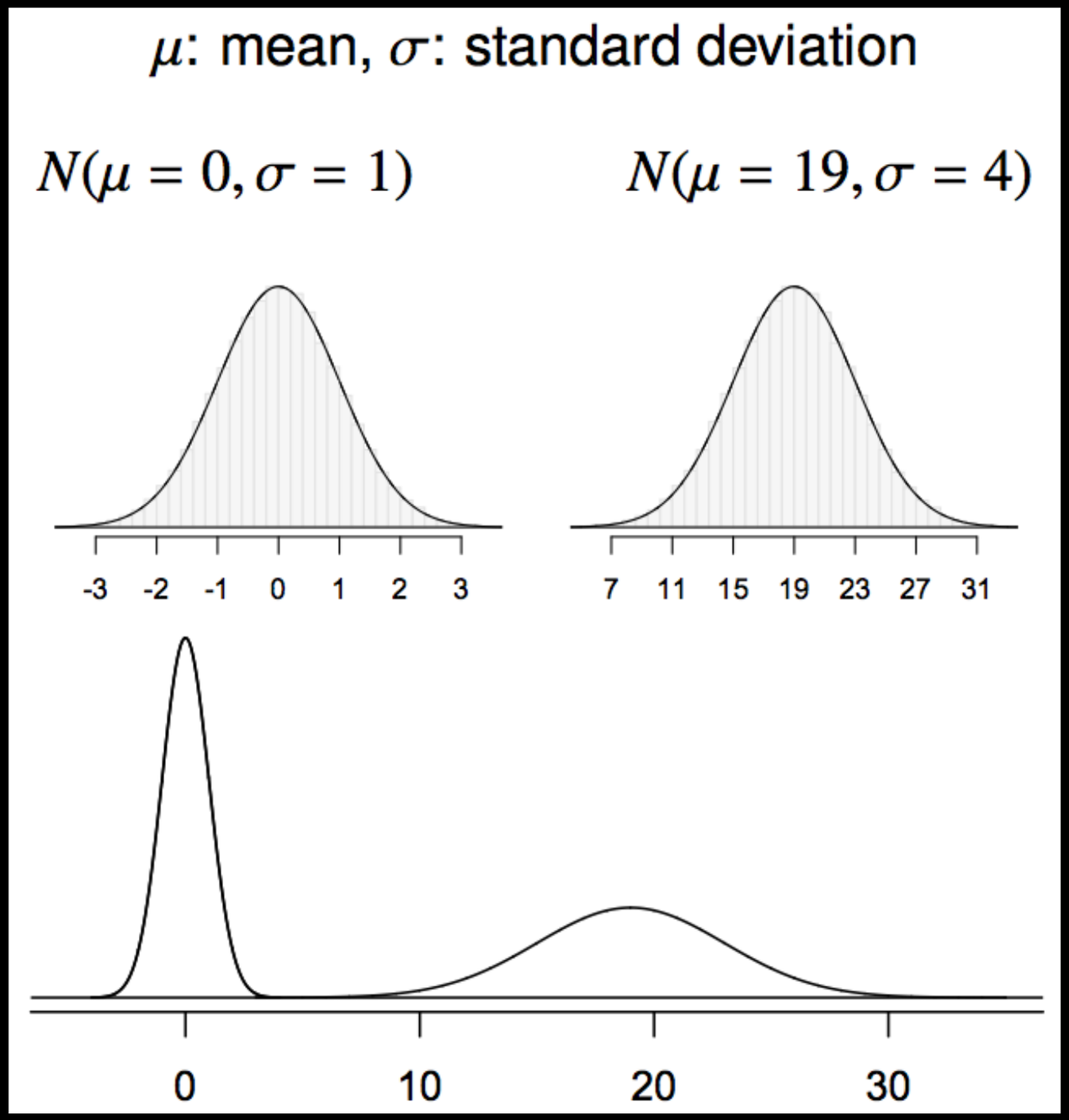
Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

~68% of students score between 1200 and 1800 on the SAT.
~95% of students score between 900 and 2100 on the SAT.
~99.7% of students score between 600 and 2400 on the SAT.



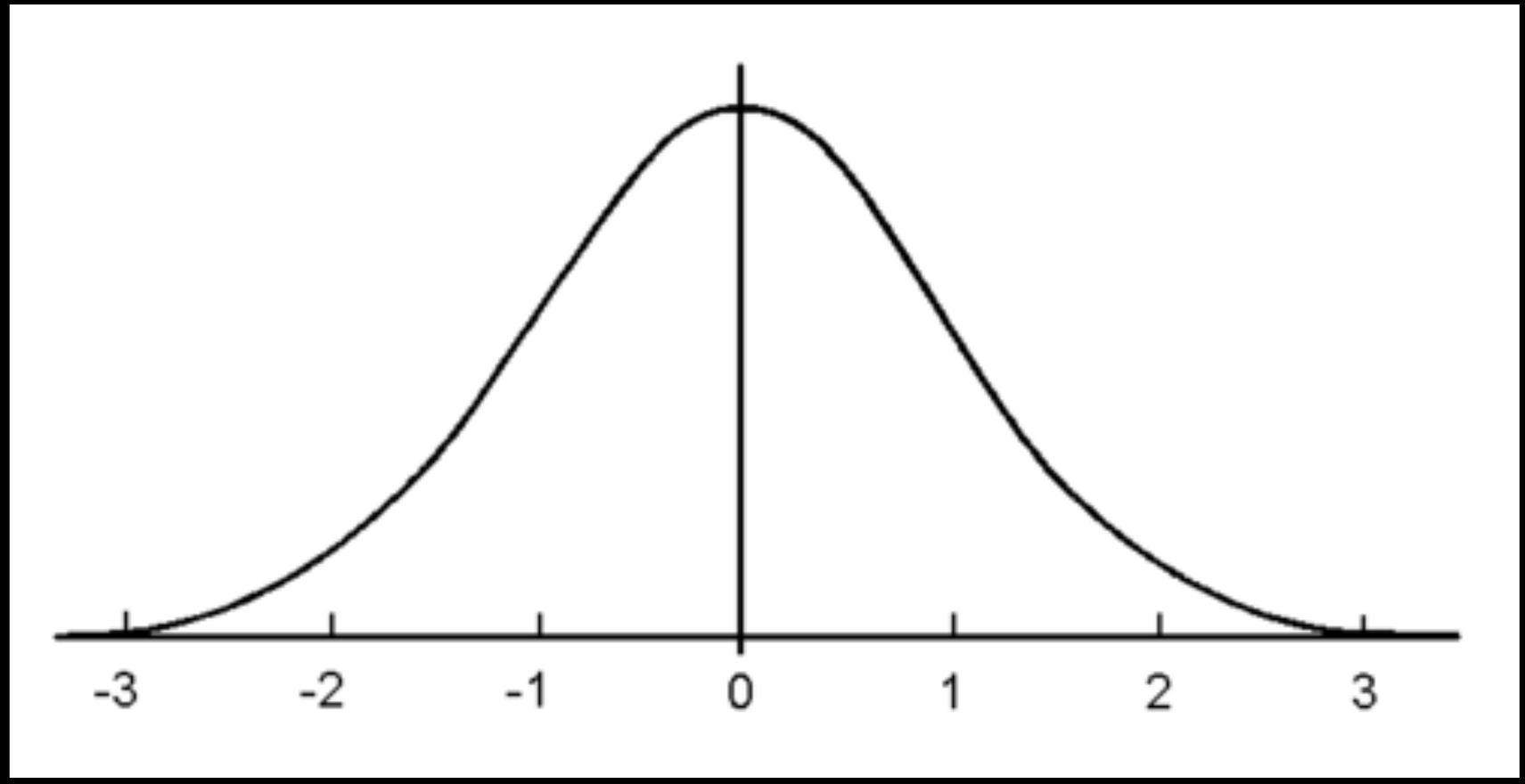
Normal distributions with different parameters



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.

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

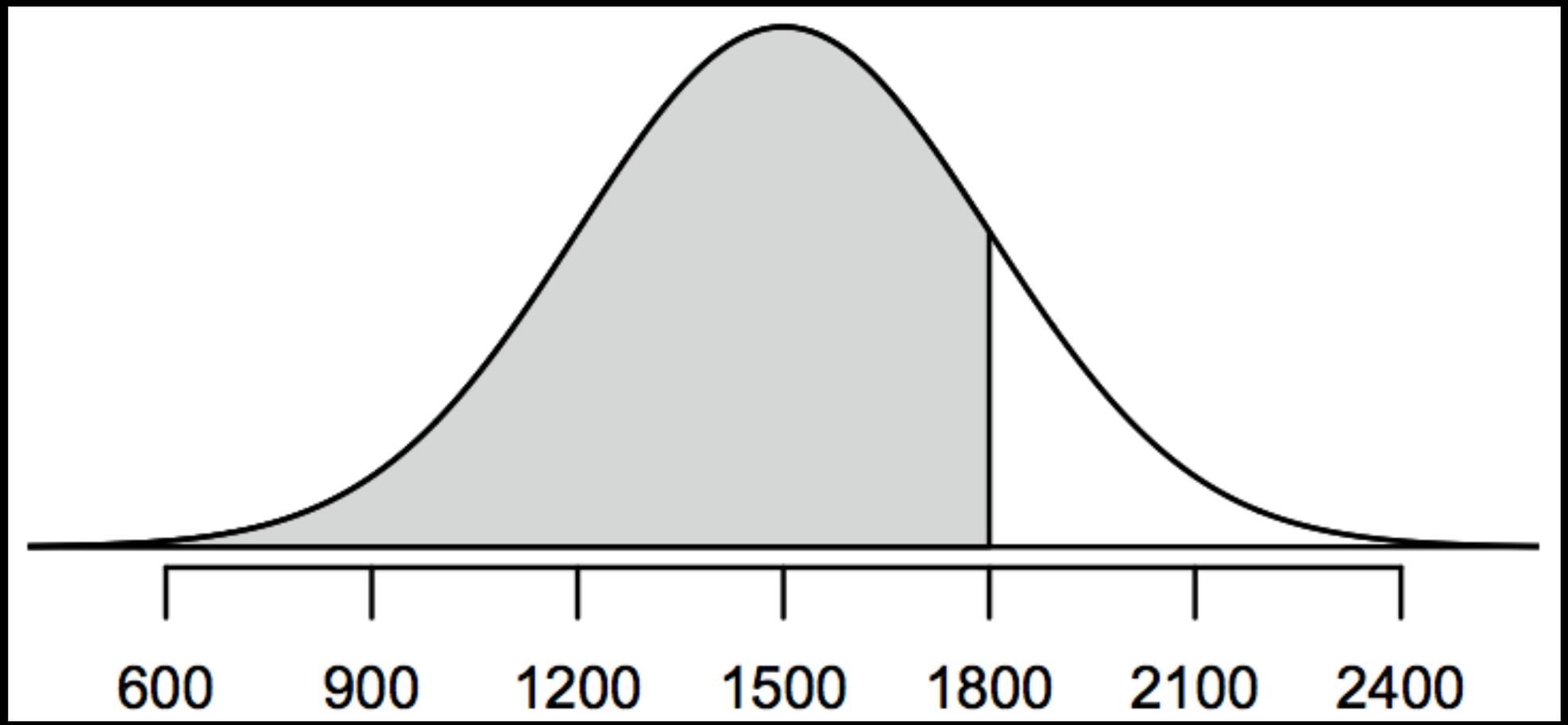
Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.

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

In R!

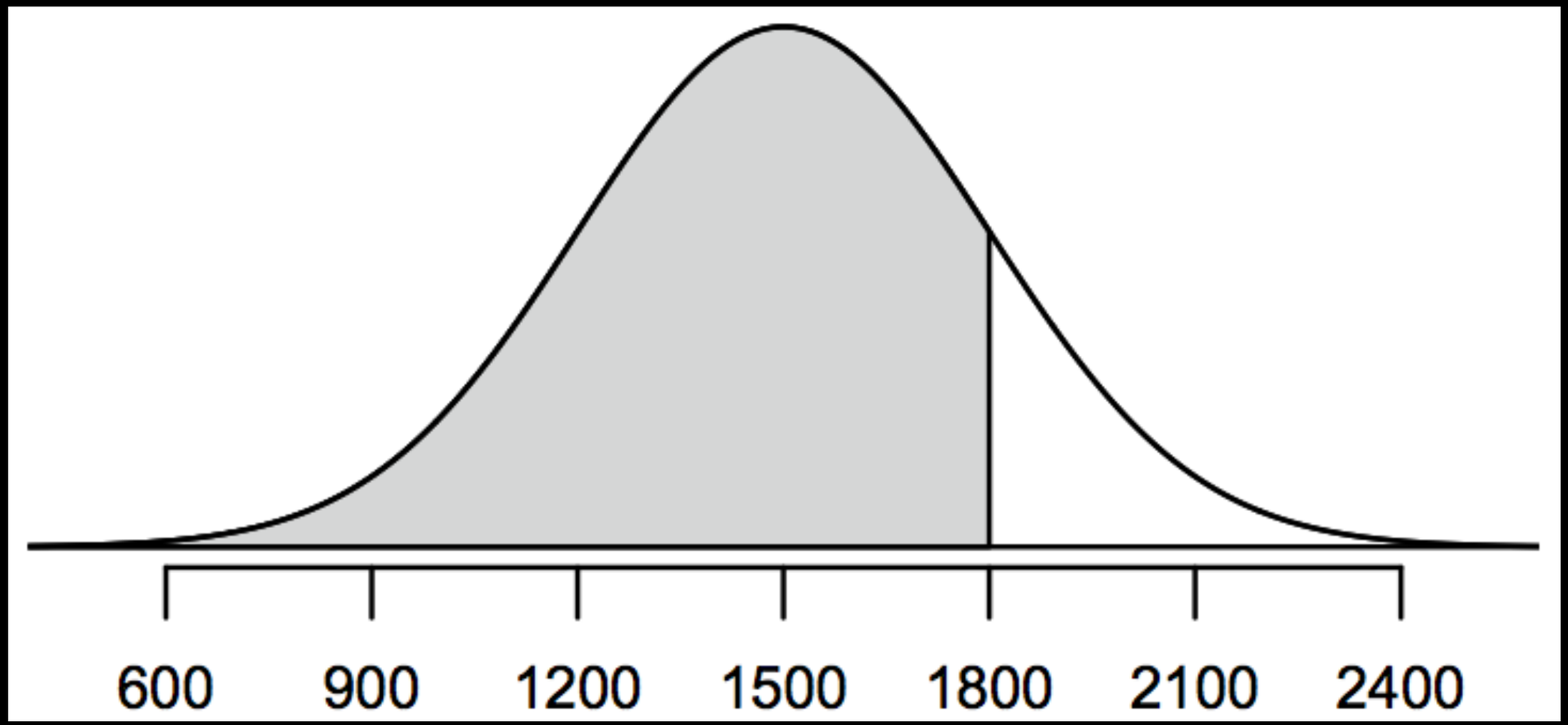
Percentiles

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.



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!