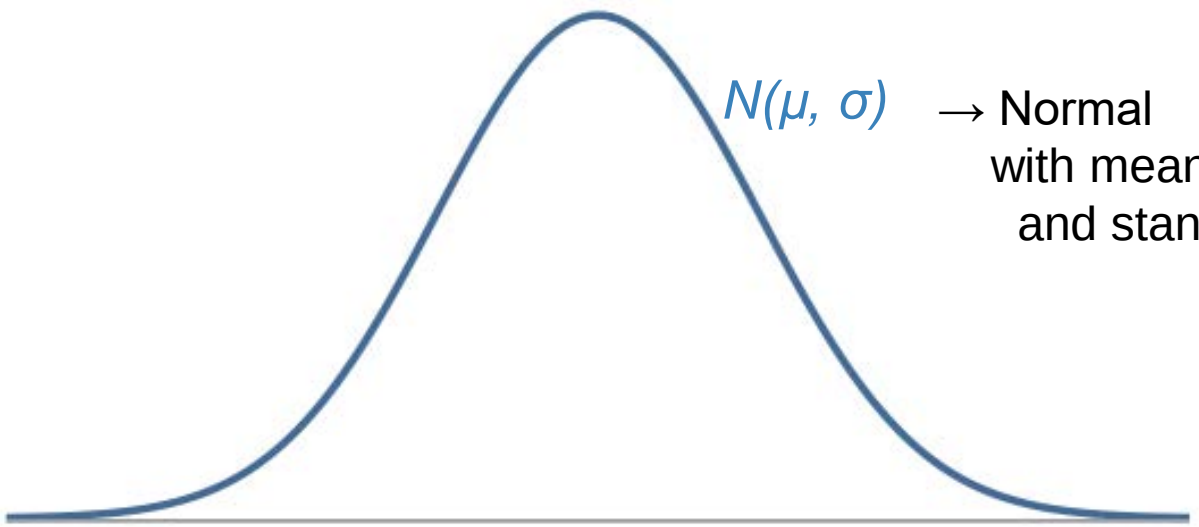
normal distribution

- unimodal and symmetric
  - bell shaped curve
- follows very strict guidelines about how variably around the mean



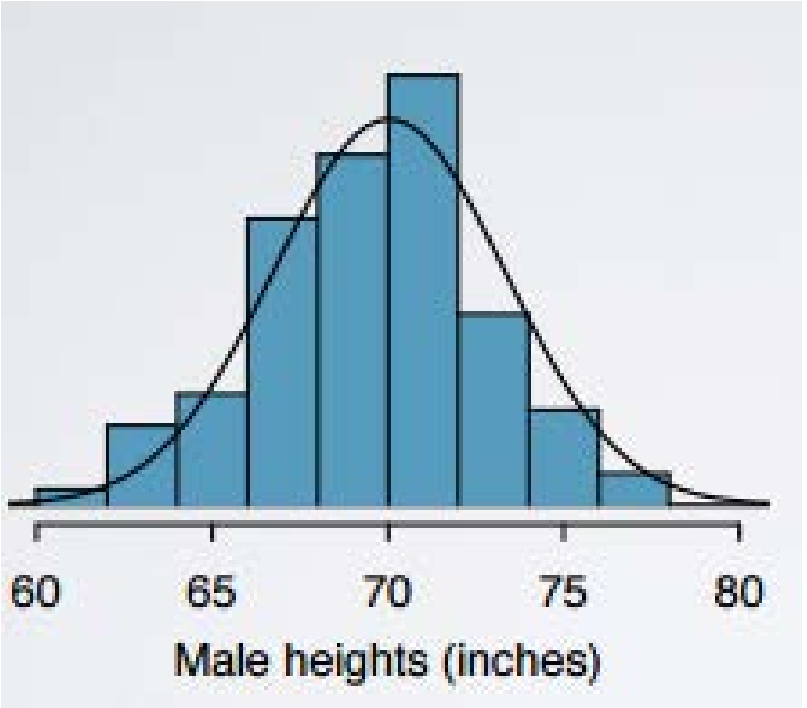
- many variables are nearly normal

normal distribution

many things closely follow a normal distribution.

- blood pressure
- size of things produced by machines
- errors in measurements
- heights of people

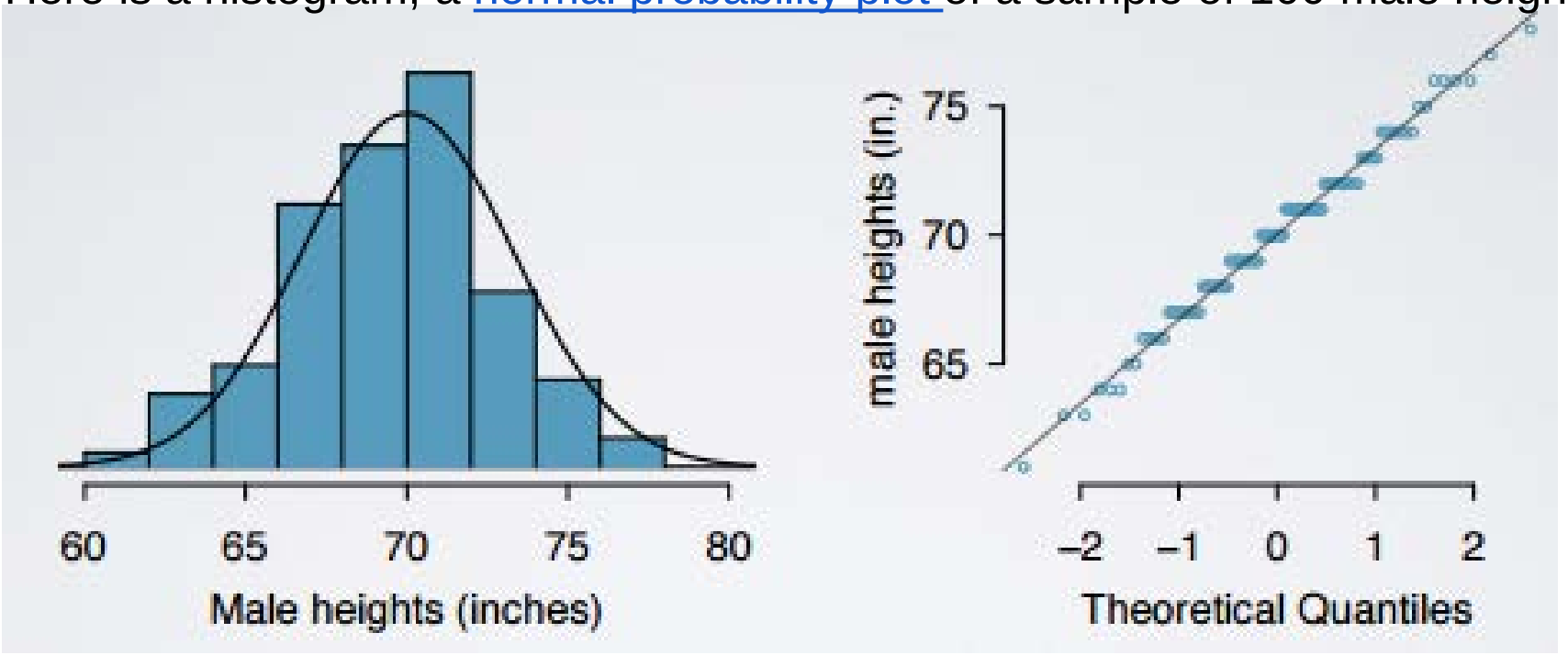
but none are exactly normal



evaluating normal distribution

normal probability plot

Here is a histogram, a normal probability plot of a sample of 100 male heights.

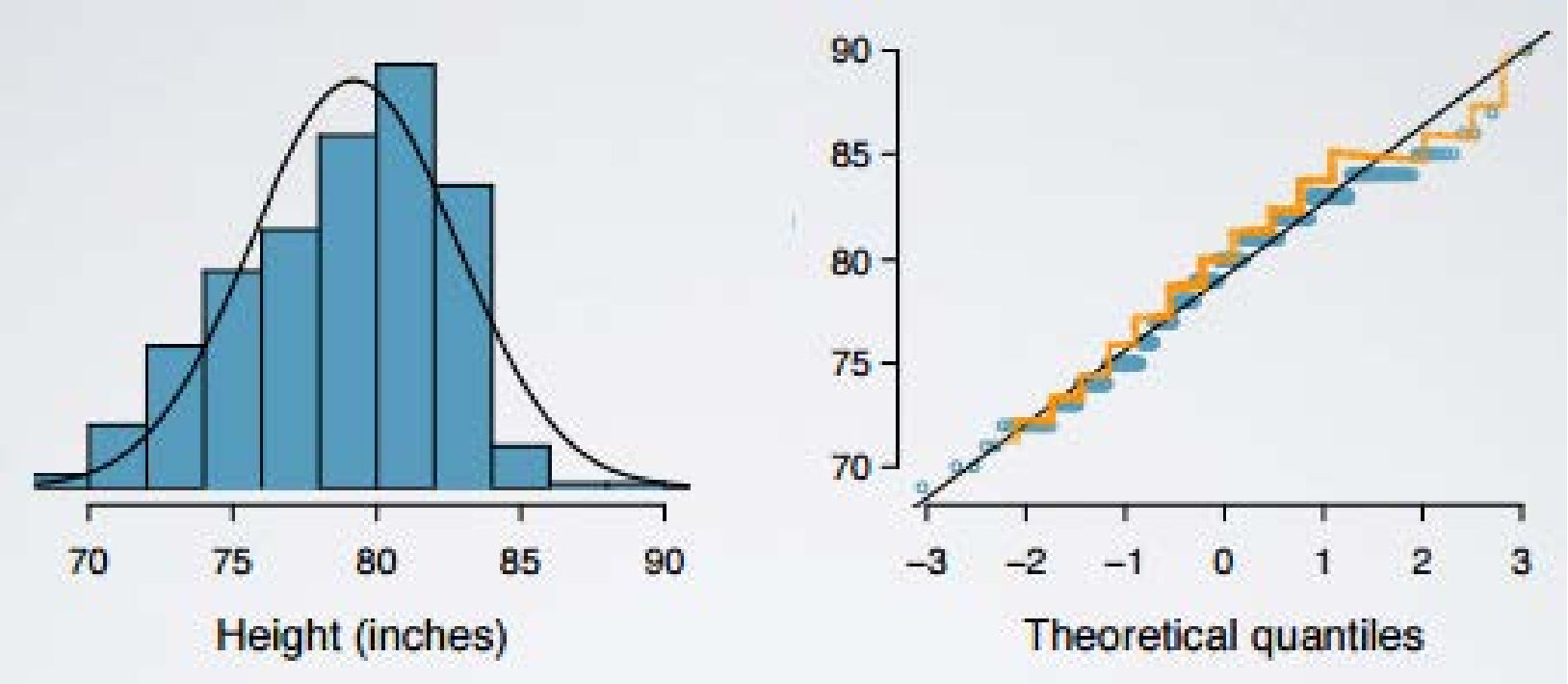


- The **normal probability plot** is a graphical technique for assessing whether or not a data set is approximately normally distributed.
- The **data** are plotted against a **theoretical normal distribution** in such a way that the points should form an **approximate straight line**. Departures from this straight line indicate departures from normality.

How to form a normal probability plot

- Data are plotted on the y-axis. The theoretical quantiles follow a normal distribution on the x-axis.
- If there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution.
- Since a one-to-one relationship would appear as **straight line** on a scatter plot, **the closer the points are to a perfect straight line**, the most confident we can be that the data follow the normal model.
- 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.

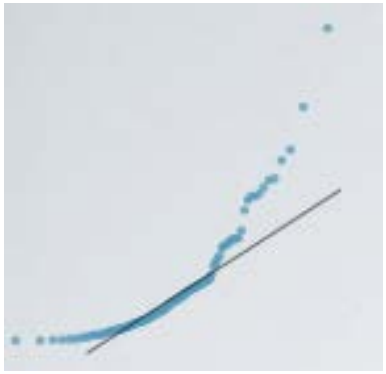
Heights of NBA players



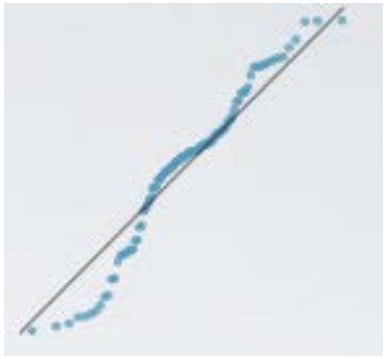
Example of data that do NOT really follow a normal distribution.

normal probability plot

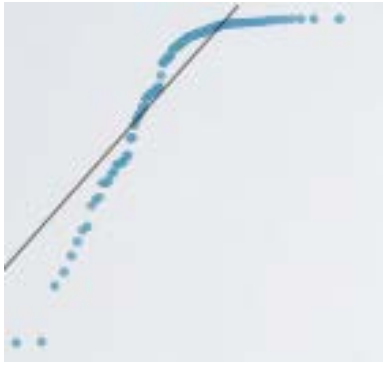
Just like histogram, normal probability plot also reveal shapes of distributions.



In the **right skew** distribution, points bend up and to the left of the line.



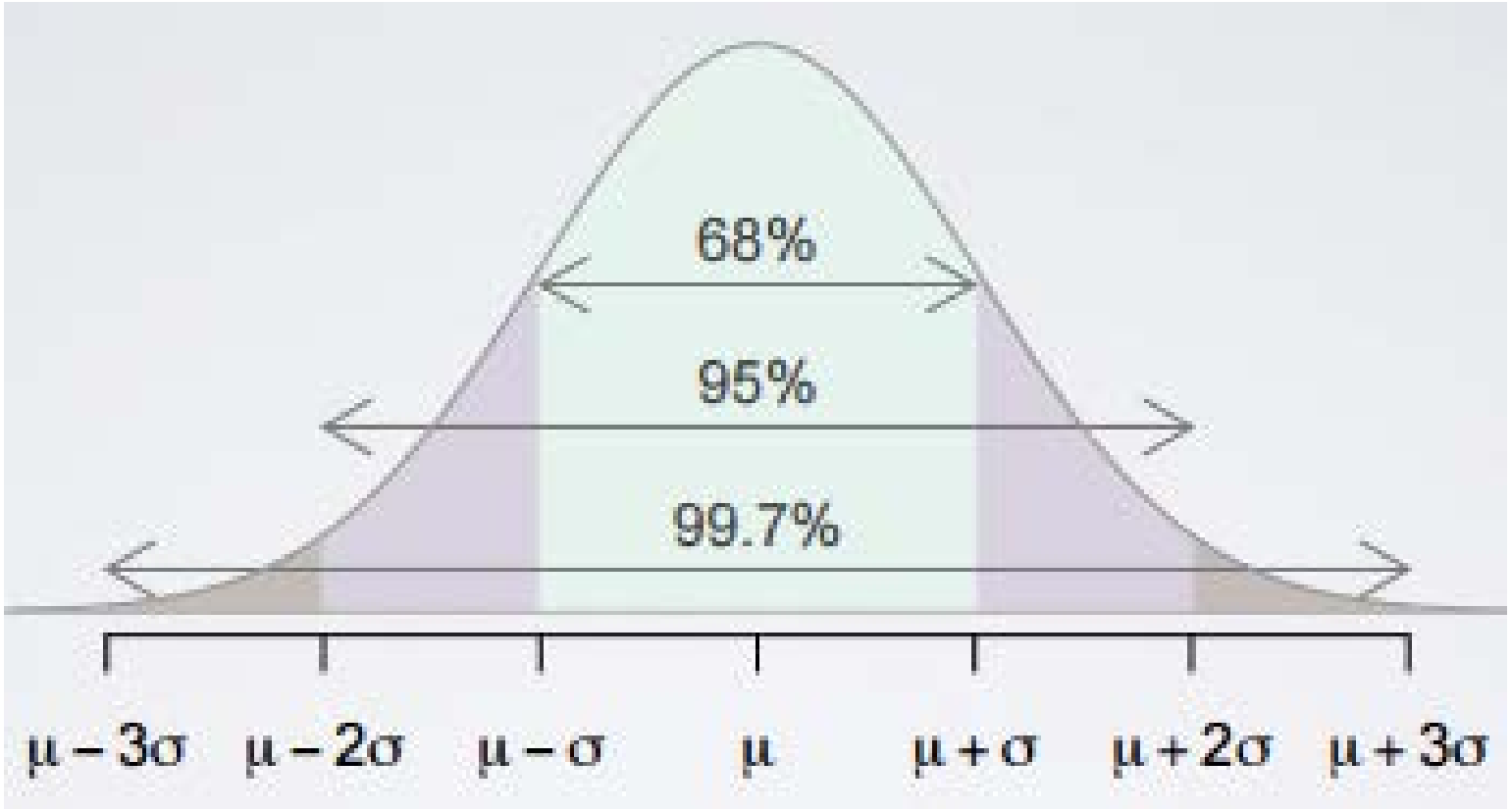
In the **short tails** (narrower than the normal distribution) distribution, points follow an S shaped-curve



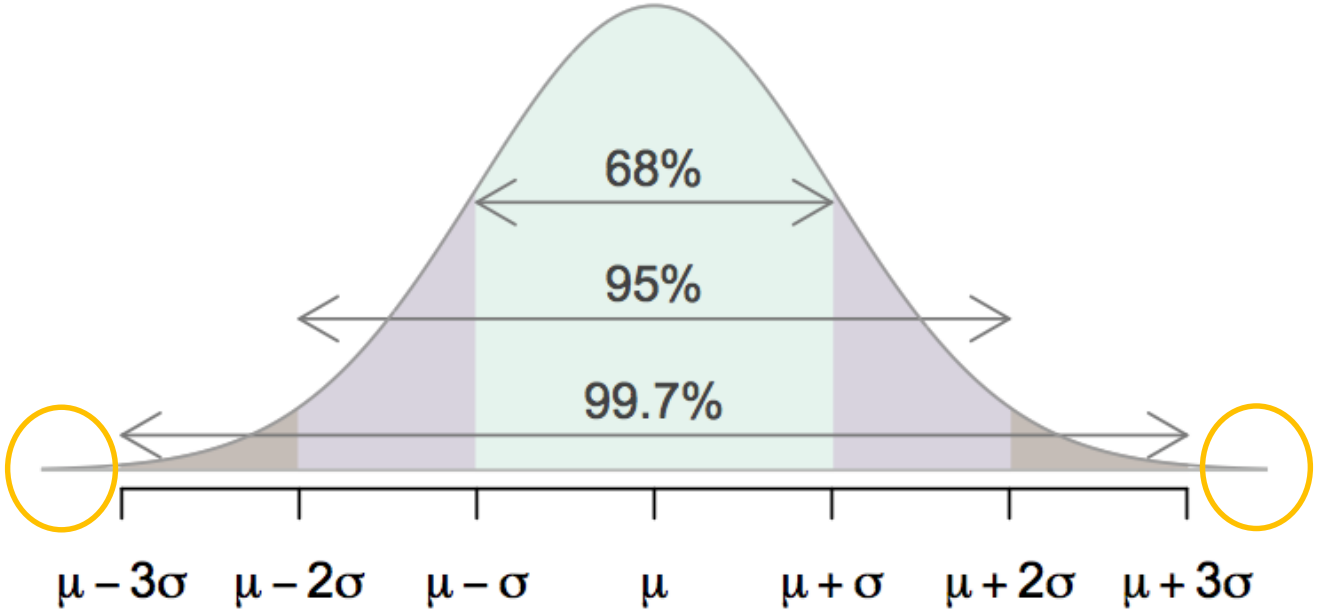
In the **left skew** distribution, points bend down and to the right of the line.

In the **long tials** (wider than the normal distribution) distribution, points start below the line, bend to follow it, and end above it.

In addition, you can also use the **68-95-99.7% rule** to evaluate normality by assessing whether the distribution follows what's required by this rule.



68-95-99.7% rule



For nearly normally distributed data,

- about 68% falls within 1 SD of the mean,
- about 95% falls within 2 SD of the mean,
- about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.