# The 68% Rule and the 95% Rule

Colm Mulcahy

Math 107-03, Spring 2020, Spelman College

3-6 Apr 2020

# Key fact about standard deviation: the 68% rule

**For large data sets, about two thirds (actually 68%) of the data is within one std dev of the mean.**

We're talking about the middle two-thirds here: we are assuming the data is symmetric and unimodal. In this situation the mean, median and mode are equal! This is an instance of the Bell Curve.

If 3000 women's heights have mean 5 feet 5 inches, with std dev 2 inches, then about 2000 of the women's heights would be between 5 feet 3 inches and 5 feet 7 inches.

Now assume that 6000 men's heights have mean 5 feet 9 inches, with std dev 3 inches, then fill in the blanks: "about XXX of the men's heights would be between YYY and ZZZ."

Get XXX is 4000, and YYY is 5 feet 6 inches, ZZZ is 6 feet.

# Key fact about standard deviation: the other 32%

**For large data sets, about one third (actually 32%) of the data is NOT within one std dev of the mean.**

If 3000 women's heights had mean 5 feet 5 inches, with std dev 2 inches, then about 1000 of the women's heights would NOT be between 5 feet 3 inches and 5 feet 7 inches.

These other women are either shorter (under 5 feet 3 inches) or taller (over 5 feet 7 inches). It's fair to assume those are roughly equally split, with about 500 women (about 16%) in each category).

Similarly, if 6000 men's heights had mean 5 feet 9 inches, with std dev 3 inches, then about 2000 of the men's heights would NOT be between 5 feet 6 inches and 6 feet. The other roughly 2000 are divided between 1000 men (about 16%) under 5 feet 6 inches and another 1000 men (about 16%) over 6 feet.

# Top 16% and the bottom 84% (and vice versa)

**For large data sets, about 16% of the data is more than one std dev above the mean.** Similarly, about 16% of the data is more than one std dev below the mean.

If new born babies weigh 7lbs on average, with std dev 1lb, then about 16% of them weigh over 8lbs. The others weigh under 8lbs. Hence 84% (100% − 16%, or 16% + 68%) weigh under 8lbs. Hence, an 8lb baby would be said to be "at the 84th percentile."

Similarly, a 7lb baby (right in the middle, weightwise) would be said to be "at the 50th percentile (or second quartile)."

A 6lb baby would be said to be "at the 16th percentile."

It would be nice to be able to say what baby weight corresponds to the 75th percentile (*aka* the third quartile), or what percentile correponds to a baby weight of 6.5lbs. We don't have the tools yet: we need to learn about "standard z-scores and bell curve tables".

# Key fact about standard deviation: the 95% rule

**For large data sets, about 95% of the data is within TWO std devs of the mean.**

If 3000 women's heights have mean 5 feet 5 inches, with std dev 2 inches, then since two std devs is 4 inches, we have about 2850 = 95% of 3000 of the women's heights would be between 5 feet 1 inches and 5 feet 9 inches.

If 6000 men's heights have mean 5 feet 9 inches, with std dev 3 inches, then use the 95% rule to fill in the blanks: "about XXX of the men's heights would be between YYY and ZZZ."

Get XXX is 5700, YYY is 5 feet 3 inches, ZZZ is 6 feet 3 inches.

# Key fact about standard deviation: the other 5%

**For large data sets, about 5% of the data is NOT within TWO std devs of the mean.**

If 3000 women's heights had mean 5 feet 5 inches, with std dev 2 inches, then about 150 (namely 5%) of the women's heights would NOT be between 5 feet 1 inches and 5 feet 9 inches. These other women are either shorter (under 5 feet 1 inches) or taller (over 5 feet 9 inches). It's fair to assume those are roughly equally split, with about 75 women (or 2.5% of the total) in each category).

If 6000 men's heights had mean 5 feet 9 inches, with std dev 3 inches, then about 300 (namely 5%) of the men's heights would NOT be between 5 feet 3 inches and 6 feet 3 inches. These roughly 300 are divided between 150 (that's 2.5%) men under 5 feet 3 inches and another 150 (2.5%) men over 6 feet 3 inches.

# Top 2.5% and the bottom 97.5% (and vice versa)

**For large data sets, about 2.5% of the data is more than two std devs above the mean.** Likewise, about 2.5% of the data is more than two std devs below the mean.

If new born babies weigh 7lbs on average, with std dev 1lb, then about 2.5% of them weigh over 9lbs. The others weigh under 9lbs. Hence 97.5% (100% − 2.5%, or 2.5% + 95%) weigh under 9lbs. A 9lb baby would be said to be "at the 97 and a half-th percentile."

A 5lb baby would be said to be "at the 2 and a half-th percentile."

We've seen how much we can deduce about a set of data knowing only its mean and standard deviation.

*We are assuming the data is symmetric and unimodal. The mean, median and mode are equal, and we have a Bell Curve.* See page 391 (Section 6C) in the text, and especially the pictures there.