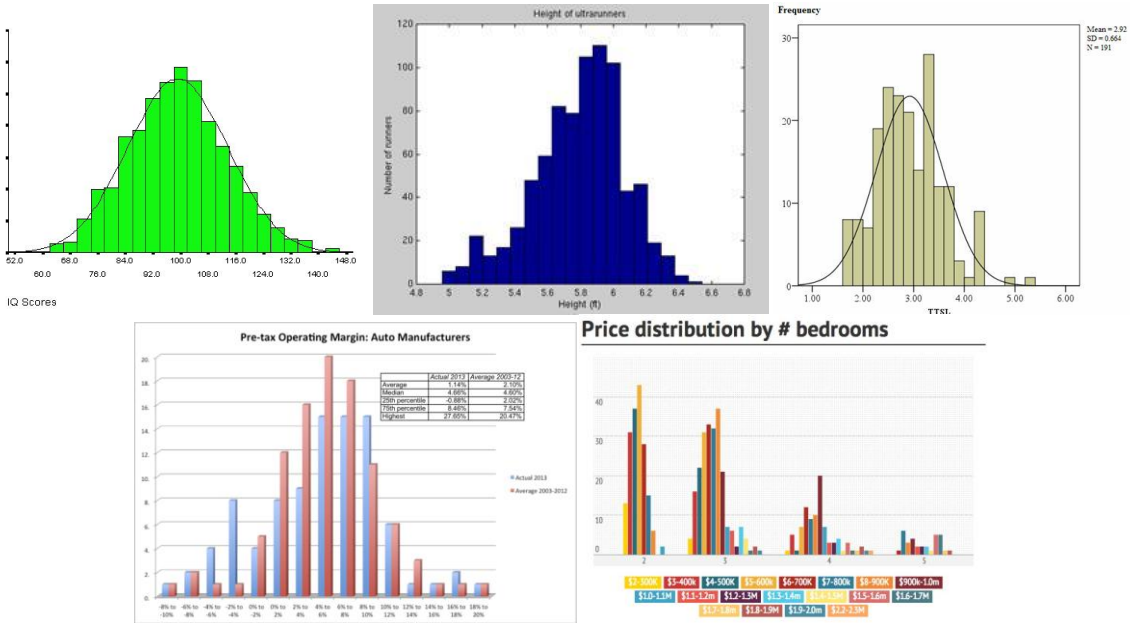
David Youngberg
BSAD 210—Montgomery College

LECTURE 09: THE NORMAL DISTRIBUTION AND THE CLT

I. Examples[1]



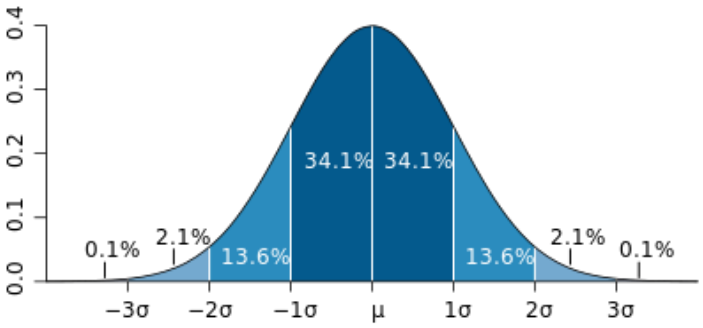
a. There are many other examples: crop yields; customer traffic; sales data; product quality; and so on.
b. All of these examples have a distribution we call a *normal distribution*—a bell-shaped distribution that is symmetric around the mean.
    i. By "symmetric" we mean each side of the mean has the same shape. This renders the mean equal to both the median and the mode.
    ii. While none of these empirical examples have *exactly* the perfect bell-curved shape, many of them approximate it. Thus we analyze data, we assume an ideal bell-curve.

---

[1] IQ: http://www.psychology.emory.edu/clinical/bliwise/Tutorials/SOM/smmod/scalemea/print2.htm
Height: http://ib.berkeley.edu/courses/ib162/Week1.htm
Time-to-stop line (time taken to determine to stop for a yellow light):
http://www.fhwa.dot.gov/publications/research/safety/09049/
Pre-tax Net Profit Margin (Operating Margin) of Auto Makers:
http://aswathdamodaran.blogspot.com/2013/09/valuation-of-week-1-tesla-test.html
Price distribution of homes sold in Berkeley by number of bedrooms:
http://www.berkeleyside.com/2013/02/15/berkeley-house-prices-tick-up-after-years-of-slump/

c. While many variables have a normal distribution, there are some that don't. Examples:
    i. Proportion of Americans who voted for each candidate in the last Presidential election.
    ii. The number of people along a beach.
    iii. Income.

II. Empirical Rule
a. As mentioned, the z-score rests on the assumption of a normal distribution.
b. This distribution follows the *empirical* rule:
    i. About 68% of all observations are within one standard deviation of the mean;
    ii. About 95% of all observations are within two standard deviations of the mean; and
    iii. About 99.7% of all observations are within three standard deviations of the mean.



    iv. This graphical representation shows how each segment of the normal distribution breaks down.
    v. Note anything beyond these three standard deviations is an outlier.

III. Qualities of a Normal Distribution
a. *Skew*—measurement of distribution symmetry
    i. Symmetric means the tails are equally disperse. Mean equals median: e.g. height. Normal distributions are symmetric.
    ii. A positive skew means there is a long tail (few extreme values) on the right. Mean is greater than median: e.g. salary
    iii. A negative skew means there is a long tail (few extreme values) on the left. Mean is less than median: e.g. test score
b. *Kurtosis*—measurement of the "peakness" of the distribution
    i. If zero, the peak resembles that of a normal distribution.
    ii. If negative, the peak is flatter than the normal. More of the variance is due to observations near the mean.

iii. If positive, the peak is sharper than the normal. More of the variance is due to extreme observations.

IV. Chebyshev's Theorem

a. A less precise, but more general, form of the empirical rule is *Chebyshev's Theorem*: the <u>minimum</u> percent of observations that fall within z standard deviations is equal to:

$$\left(1 - \frac{1}{z^2}\right) \times 100$$

…assuming z > 1

b. This Theorem is true of *any* distribution—bell-shaped or not—but does not give a precise percent.

V. Central Limit Theorem

a. The *Central Limit Theorem (CLT)* states that the sample means of large-sized will be normally distributed regardless of the shape of the distribution.

i. In other words, suppose you take 100 samples of a population, with each sample having many observations in it. If you take the average of each sample, you'll get 100 sample averages. Those 100 averages will form a normal distribution, with a few averages being very low or very high and many being right in the middle.

b. What's interesting is that the CLT works for any population distribution.