Measures of Dispersion, Skew, & Kurtosis (based on Kirk, Ch. 4)

{to be used in conjunction with “Measures of Dispersion Chart”}

Percentiles

percentile (P\(_{\%}\)): a score below which a specified percentage of scores falls (start with a percent, and convert to a score)

vs. percentile rank (P\(_R\)): percentage of scores that falls below a given score (start with a score, and convert to a percent)...we will return to this topic later

when computing the median, we are interested in computing the 50\(^{th}\) percentile (P\(_{50}\))

recall (from review): \(Mdn = P_{50} = X_{ll} + i(\frac{n}{2} - \sum f_i)\), where:

- \(X_{ll}\) is the real lower limit of the interval containing the median (the logic is that you are starting at the bottom of the interval and counting up into it)
- \(i\) is the size of the class interval (1 if you are working with ungrouped distributions)
- \(n\) is the number of scores in the entire distribution (thus, \(n/2\) is the location of the midpoint)
- \(\sum f_i\) is the number of scores below \(X_{ll}\) (this tells you how many scores up from the bottom that you’ve already come...so, taking this away from \(n/2\) tells you how many more scores you have to go)
- \(f_i\) is the number of scores there are in the interval containing the median (this tells you how many pieces to divide that interval into)

Example:

\(X = \{4, 4, 4, 5, 5, 5, 6, 7, 8, 8\}\)

\(Mdn (X) = \)

When computing the semi-interquartile range (Q), we need the 25\(^{th}\) and 75\(^{th}\) percentiles (P\(_{25}\) and P\(_{75}\)).

The formula for the median can be generalized to apply to any percentile, including the 25\(^{th}\) and 75\(^{th}\):
\[ P_{\%} = X_{ll} + i \left( \frac{n(P_R / 100) - \sum f_b}{f_i} \right) \]

where:

- \( X_{ll} \) is the real lower limit of the interval containing the percentile of interest (the logic is that you are starting at the bottom of the interval and counting up into it); consider the sample size to determine what interval (when working with raw, ungrouped data, this is just a score) to work with (e.g., e.g. if you have 10 cases, the 25\(^{\text{th}}\) percentile is associated with the .25(10) = 2.5\(^{\text{th}}\) case...use the real lower limit of the interval containing the 2.5\(^{\text{th}}\) case)
- \( i \) is the size of the class interval (1 if you are working with ungrouped distributions)
- \( n \) is the number of scores in the entire distribution
- \( P_R \) is the percentile rank you are working with (this is 50 for the median/Q\(_2\), 25 for Q\(_1\), 75 for Q\(_3\), etc.); thus, \( n(P_R / 100) \) tells you how many cases into the distribution (from the bottom) you must come to get to the case of interest (you actually use this to get \( X_{ll} \))
- \( \sum f_b \) is the number of scores below \( X_{ll} \) (this tells you how many scores up from the bottom that you’ve already come...so, taking this away from \( n(P_R / 100) \) tells you how many more scores you have to go)
- \( f_i \) is the number of scores there are in the interval containing the percentile of interest (this tells you how many pieces to divide that interval into)

Example:

\[ X = \{ 4, 4, 4, 5, 5, 5, 6, 7, 8, 8 \} \]

\[ Q_3 = \]

\[ Q_1 = \]

\[ Q = \]
**Problem:** Statistical packages compute percentiles differently from each other (and from the above method)

* see [http://www.maths.murdoch.edu.au/units/statsnotes/samplestats/quartilesmore.html](http://www.maths.murdoch.edu.au/units/statsnotes/samplestats/quartilesmore.html) if you have further interest in this topic

* Note that SPSS tends to give more extreme values (i.e., percentiles that are further away from the median), while Excel tends to give less extreme values (i.e., percentiles that are closer to the median); neither program uses the algorithm presented above (note, however, that the above algorithm is thought to be the best for normal distributions)

**Percentile Rank:** Again, here we are going from a score to a percent, so we just solve for $P_R$ in the above equation:

$$P_R = \frac{100}{n} \left( \sum f_i + \left( \frac{f_i (P_m - X_m)}{i} \right) \right)$$

Example: Using the above data set, find the percentile rank for the score of 8.

**More on Standard Deviation**

**Calculating Standard Deviation:**

Recall that the SD approximates the average distance of scores from the mean (deviations).

Why not just take the true average of these deviations?

The mean is the balancing point of all scores...there are equal positive deviations and negative deviations, so the sum of all deviations is equal to zero...always!...so this won’t give us any interesting information.

Why not take the absolute value of the deviations?

This is the right idea, but mathematicians have found that the resulting values do not work well with more advanced calculations (they are not mathematically tractable)

What’s the solution?

First square all the deviations and sum them (this gives us the sums of squares, SS). Then, take the average and convert back to original units by taking the square root.

**Variations on the Standard Deviation Formula:**

The formula given in the chart is for finding the standard deviation for a sample. It is purely a descriptive statistic:
We may also be interested in finding the standard deviation for a known population. In this case, the formula is the same, but now we’re comparing individual scores to a population mean instead of a sample mean:

\[ \sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}} \]

If our ultimate interest is in drawing an inference about a population based on a sample, we must use a slightly different formula. Statisticians (who make up populations and draw samples from these populations) have found that the formula for \( S \) will underestimate the true population standard deviation. They have found that dividing by \( n-1 \) (instead of \( n \)) compensates for this bias.

\[ \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \]

For any of the above formulas for standard deviation, it is noteworthy that a raw-score formula may also be applied. Below is the case for the sample SD:

\[ S = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \left( \sum_{i=1}^{n} X_i \right)^2}{n}} \]

This formula makes calculations easier (especially when there are many cases), but it is not as conceptually meaningful as the deviation formula.

**Effects of Linear Transformations of Scores on Standard Deviation**

What happens when we add a constant to all scores?
What happens when we multiply each score by a constant?

Standard Deviation and the Normal Curve

(see Kirk, p. 125)

Variance

Variance is another measure of dispersion that is often used in inferential statistics. It is simply the SD squared and is represented by $S^2$, $\sigma^2$, or $\hat{\sigma}^2$. For example:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

Skew

Skew refers to the asymmetry of a distribution and can be computed as:
Note that the deviations in the numerator of the formula are cubed. This means that each score can contribute a negative (scores to the left of the mean) or positive (scores to the right of the mean) value to Sk. Also note that the farther out a score is (i.e., in the tail), the greater “pull” it will have on the value of Sk.

For example, if the mean is 50, a score of 51 will add 1 to the sum of the cubed deviations, but a score of 25 will add \(-25^3\), or -15,625 to this sum.

If there are a lot of relatively extreme scores in one tail, the skew formula will reflect this.

Negative Skew (-) Symmetrical (0) Positive Skew (+)

**Kurtosis**

Kurtosis reflects how peaked or flat a distribution is and is calculated as:

\[
Kur = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{S^4} \frac{n}{n-3}
\]

platykurtic (-) mesokurtik (0) leptokurtic (+)

Examples (use data from chart):

Sk =

Kur =