Today’s Agenda – Wrapping up Ch.3 and 4

- Mean and frequency (Loose end Ch. 3)
- Variation (Not in text)
- Standard Deviation (Ch.4)
- Standard Deviation and Outliers (Ch. 4)
Consider the data set \{0,0,0, 4,4,7,10,10,10,10\}, n=10. Recall that we could find the mean by:

\[
\frac{1}{10} \times (0 + 0 + 0 + 4 + 4 + 7 + 10 + 10 + 10 + 10) = \underline{7}\\
\]

...and that we could describe the data in a frequency table

<table>
<thead>
<tr>
<th>Value</th>
<th>Frequency</th>
<th>Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>0.3, or 30%</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>0.4</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>
But if all we had was the frequency table, we could still get the mean.

There are four 10s, so instead of putting $10 + 10 + 10 + 10$ in the formula we could put in $\underline{\phantom{100}}$

\[
\frac{(0 + 0 + 0 + 4 + 4 + 7 + 10 + 10 + 10 + 10)}{10} = \frac{(3 \cdot 0 + 2 \cdot 4 + 1 \cdot 7 + 4 \cdot 10)}{10} = \frac{(0 + 8 + 7 + 40)}{10} = 5.5
\]
This becomes really important if we have _________ of the same number.

<table>
<thead>
<tr>
<th>Number of Children</th>
<th>Frequency</th>
<th>Relative Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>173,082</td>
<td>0.056</td>
</tr>
<tr>
<td>1</td>
<td>983,586</td>
<td>0.317</td>
</tr>
<tr>
<td>2</td>
<td>1,251,753</td>
<td>0.404</td>
</tr>
<tr>
<td>3</td>
<td>462,952</td>
<td>0.149</td>
</tr>
<tr>
<td>4</td>
<td>228,741</td>
<td>0.074</td>
</tr>
<tr>
<td>Total</td>
<td>3,100,114</td>
<td>1.000</td>
</tr>
</tbody>
</table>

We could even use the _________ frequency and skip division.

Mean = 0(0.056) + 1(0.317) + 2(0.404) + 3(0.149) + 4(0.074) = 1.868
Variation: Not everything can be controlled. Results may vary, even in a factory setting. Some bags will get more chips than others, we say there is _________ in the weights in each bag.

Image source: Failblog.org, “Quality Control Fail”
There are laws about the proportion of bags sold that can be under-weight.

A company needs to know the proportion that will be under but can’t afford to check every single bag.

Instead they check a sample of bags and hope it represents the population. (Like my survey of 39 students)
..but those samples are not going to be the same every time. Most of you have done this before during R-R-R-Roll Up The Rim season.

They say there’s a one in six chance of winning, but did you win on EXACTLY one in six cups. Did you win as much as your friends?
Mine: 0 / 3 = 0% Wins
Jason: 3 / 13 = 23% Wins
Emelie: 6 / 39 = 15% Wins
Each person’s roll up the rim season is different, why? Variability!
Why should you care?

When you’re doing a social study or experiment, your results aren’t going to be hard set.
If you did the same study tomorrow with similar subjects, you’d get different results. It would help if we had an idea how different we would expect these differences to be.

Image: xkcd.com
That’s what measures of spread like the __________ __________ (IQR) are for.

They help us measure how uncertain we are about our central values.

IQR is intuitive, works for a wide range of distributions, and has the __________ __________ for finding outliers.

But it’s tied to the median and related measures like the quartiles.
A spread measure based on the mean is the __________

__________

To deviate means the stray from the norm.

A standard deviation is the typical amount strayed from the mean.
When the distribution looks kind of like this...

about $\frac{2}{3}$ of the distribution is within 1 sd of the mean
about 95% is within 2 sd of the mean
about 99% is within 3 sd of the mean
Example: Grade 5 Reading Scores have a mean of 120 and a standard deviation (sd) of 25.

120 + 1sd = 145
120 – 1sd = 95
So about 2/3 of the grade 5s have a reading score between ___________
Example: Grade 5 Reading Scores have a mean of 120 and a standard deviation (sd) of 25.

$\text{120} + 2\text{sd} = 120 + 2(25) = \underline{160}$

$\text{120} - 2\text{sd} = 120 - 2(25) = \underline{70}$

So about 95% of the grade 5s have a reading score between 70 and 160.
Another way to determine outliers when using the mean and standard deviation is the 3 standard deviation rule.

Anything __________ _________ _________ below or above the mean is an outlier.
With the reading scores, anything below _______________ or above __________ __________ is an outlier.

Like the mean and standard deviation, this outlier measure is only appropriate for __________ data.
<table>
<thead>
<tr>
<th>Recommended for...</th>
<th>Symmetric Data, use by default.</th>
<th>Skewed Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spread / Variance Measure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Outlier Rule</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The ________ is the average squared difference between a value and the mean.

The standard deviation is the ________ _________ of the variance.

For XYZ data
21.15 = \(\sqrt{447.1}\)

<table>
<thead>
<tr>
<th></th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>32.9500</td>
<td>-.1780</td>
<td>4.7477</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>21.14518</td>
<td>.98133</td>
<td>3.48949</td>
</tr>
<tr>
<td>Variance</td>
<td>447.119</td>
<td>.963</td>
<td>12.177</td>
</tr>
</tbody>
</table>
Why squared? Because the average raw difference between values and the mean is zero.

Basically average – average = 0

So we square to make all of the difference values positive (a negative times a negative is positive)

After getting the average of the squares, we square root to get the standard deviation.

Because we used both a square and a square root, the standard deviation is the same scale as the mean.
The standard deviation is only used for ________ (or close) distributions.

When data is skewed the standard deviation breaks down because of ________ of the deviations becomes important.
Example: Right skewed distribution.

The first standard deviation below the mean (blue) covers more of the distribution than first one above (red).

So a standard deviation below implies something different than a standard deviation above.
Example: Right skewed distribution.

Since the mean is more than the median, there are more values below the mean. Does that imply that a deviation below the mean is ‘standard’?

For skew, avoid the whole mess and use the IQR.
Pop quiz:

If the distribution is **symmetric** and the data is **interval**, then the best measure of variability is:

- a) Interquartile range
- b) Standard Deviation

**Hint:** What is the default central measure? Which measure above is based on that?
Pop quiz:

If the data is ordinal, then which measure of variability/spread is not possible (without extra assumptions):

- a) Interquartile range
- b) Standard Deviation

Hint: The standard deviation is based on the mean. Do ordinals have means?
Pop quiz:
Which of the following standard deviations is/are impossible?

a) 40
b) 7 potatoes
c) -4

Hint: The standard deviation is the square root of the variance.
Getting standard deviation from SPSS is the same as getting the mean, median, and quartiles:

1. Analyze → Descriptive Statistics → Frequencies
2. Choose your variables, click on “Statistics”
3. Check off “Std. deviation” and “Variance”
Joy of stats: Distribution 19:40 – 23:00

Next Wednesday: Probability, start reading chapter 5.

NO CLASS OR OFFICE HOUR MONDAY,
HAVE A GOOD LONG WEEKEND!