**Packet 7: Summarizing Quantitative Data** Textbook pages: 48 – 50; 53 – 72

After completing this material, you should be able to:

- describe the distribution of a quantitative variable by discussing its shape, center, spread, and unusual characteristics.
- calculate (using StatCrunch) measures of center and measures of spread.
- apply the Empirical Rule or Chebyshev's Rule to a distribution when discussing the standard deviation.
- compare distributions using boxplots.

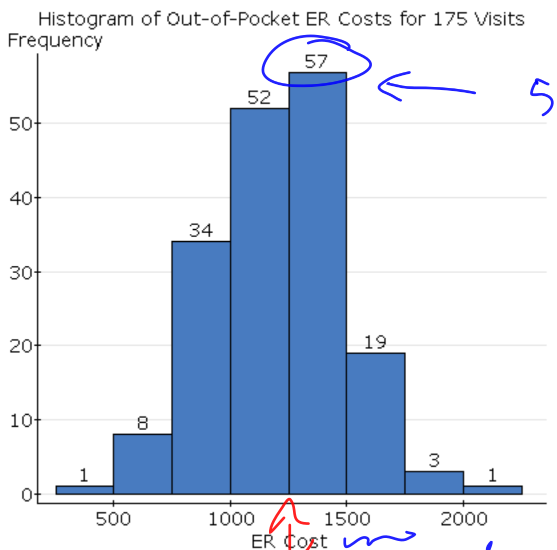
**Recall:** What is a quantitative variable?

Measures a variable using real numbers.
for which averages make sense.

To summarize a quantitative variable, we need a new graphical display – a bar graph cannot be used. We will first look at **histograms** for graphically summarizing quantitative data. What exactly is a histogram?

The numeric values are divided into nonoverlapping classes (or bins) that span all the data, & every observation is in one & only one class.

**Example:** *Money* magazine undertook a study in 2009 to estimate the average cost for a visit to a hospital emergency room. A random sample of 175 emergency room visits in a certain urban area was taken, and the out-of-pocket costs associated with that visit were recorded. A histogram for the collected data is given below:




57 observations from $1250 to $1500.

Classes are ranges of data values

[250, 500[, [500, 750[

500 is here – every obs. is in a unique class.

Bin widths are at the user's discretion.

$250 wide

**When summarizing or describing a distribution, the following four characteristics must be discussed:**

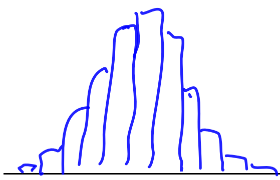
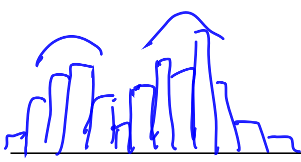
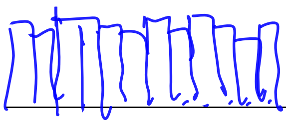
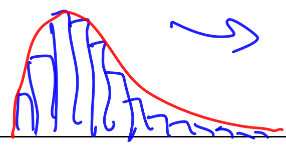
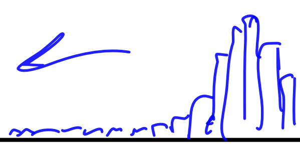
1. shape – hoping for normal – like this one!
2. center – typical values
3. spread – summarizes the variability in the data
4. outliers – odd or extreme values; we worry about mistakes or data errors.

**When asked to describe a distribution, make sure you address these four characteristics *in context* and *in complete sentences*.**

Let's consider each of these four characteristics individually.

**Shape of the distribution**

When describing a distribution, the first think we should consider is what shape the distribution has. We will consider **five** common shapes (shown below):

| Shape | Histogram | Description |
|---|---|---|
|  | Normal Dist. One mode or hump | A symmetric, uni-modal distribution. |
|  | Bi-modal Dist ("two modes") | Maybe there are two different types of obs. units |
|  | Uniform Dist (same heights) | A symmetric dist. w/o a mode - def. of fairness |
|  | Skewed Right (follow the tail) | Like a $\chi^2$ for example. Long tail to the right |
|  | Skewed Left | |

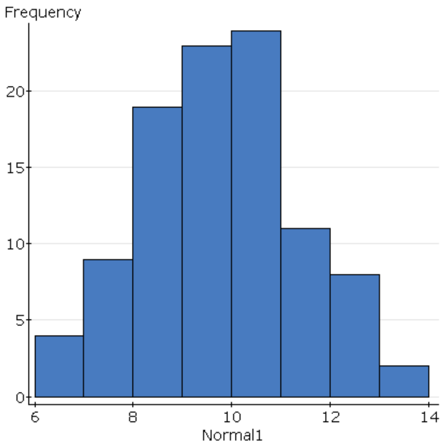
**Measures of Center**

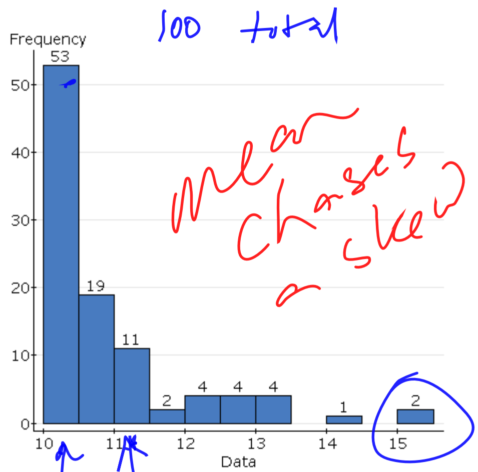
Once we know the shape of a distribution, it is common to summarize it by finding a "typical" value of the distribution – these values are generally referred to as measures of center. There are two common measures of center which are used:

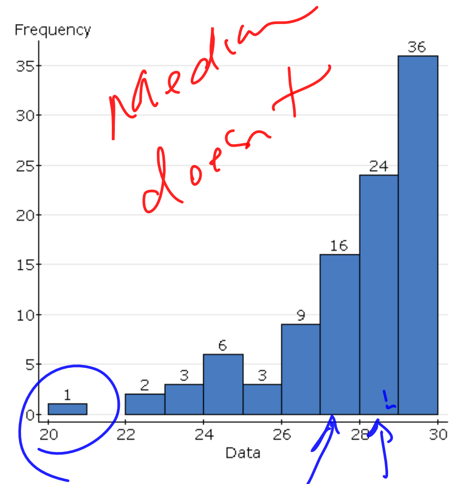
| Measure of Center | Notation | Description |
|---|---|---|
| mean - numerical average (chases the skew) | $\mu$ - pop. $\bar{y}$ - sample | The numerical average of the data - Sensitive to skew |
| median - middle value of (sorted) data | $Q_2$ (non standard) | The value such that half are below & half are above. insensitive to skew |

The calculation of these measures, while not difficult, can be tedious. Instead of calculating these summary statistics by hand, we will rely on the use of StatCrunch.

How does the shape of the distribution affect measures of center?




mean
median
True for symmetric

median here

mean here - chasing the skew

chasing skew is the mean

median here

**Measures of Spread**

Unfortunately, a measure of center doesn't adequately describe a distribution. We also must have some idea how the values in the distribution vary. This requires a measure of spread. There are three common measures of spread which are used:

| Measure of Spread | Notation | Description |
|---|---|---|
| Range sensitive to skew | non standard | The distance between max and min. 100% of data within that span. |
| Interquartile range insensitive to skew | IQR = $Q_3 - Q_1$ (3rd quartile 75% below; 1st 25% below) | The distance between the middle 50% of data - Range of middle 50% |
| Standard deviation sensitive to skew | $\sigma$ - pop std. dev.<br>$s$ - sample std. dev. | Typical deviation — unsurprising deviation - from the mean. |

The calculation of these measures can be quite difficult – the formula for standard deviation is quite tedious. Instead of calculating these summary statistics by hand, we will rely on the use of StatCrunch.

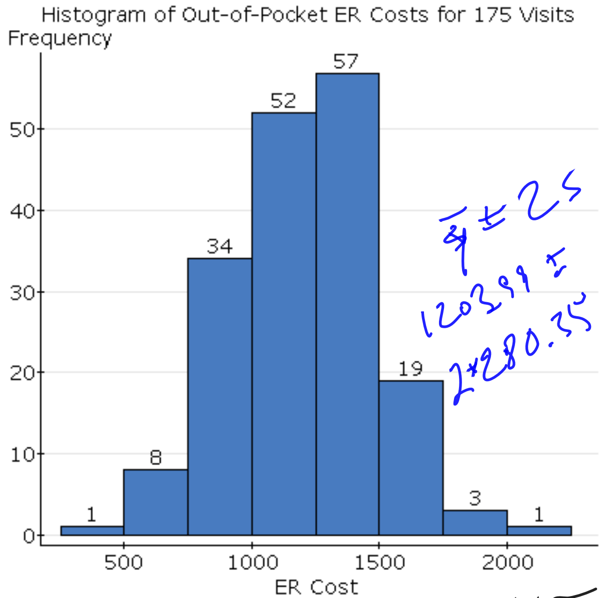
be aware of these

**Unusual Observations** Outliers. → risk skewing things.

Unusual observations are often referred to as *outliers*. When determining if unusual observations are present in the data, look for observations which do not follow the overall patter of the data. These will generally be observations which are in the tail of the distribution – either very large or very small. get rid of these

When are they real, or in error?

**Example:** *Money* magazine undertook a study in 2009 to estimate the average cost for a visit to a hospital emergency room. A random sample of 175 emergency room visits in a certain urban area was taken, and the out-of-pocket costs associated with that visit were recorded. A histogram for the sample is given below, as well as summary statistics:




$\bar{y} \pm 2s$

$1203.99 \pm 2 * 280.35$

Describe the distribution of out-of-pocket costs.

The dist. of out of pocket costs looks relatively normal. The mean cost for these 175 patients was \$1203.99, while 50% of those sampled paid \$1214 or less. The minimum cost was \$471 while the max was \$2157. The middle 50% of cost were between \$1000 & \$1414. According to the empirical rule 95% of patients paid between \$643.29 & \$1764.67. No apparent outliers.

center

spread

Range IQR

**Summary statistics:**

| Column | n | Mean | ~~Variance~~ | Std. dev. | Std. err. | Median | Range | Min | Max | Q1 | Q3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ER Cost | 175 | 1203.9886 | ~~78595.161~~ | 280.34828 | 21.192338 | 1214 | 1686 | 471 | 2157 | 1000 | 1414 |

---

How can the standard deviation be interpreted?

works for all!

**Empirical Rule**

$\mu \pm 1\sigma$

68% between $\bar{y} \pm s$

95% " $\mu \pm 2\sigma$

99.7% " $\mu \pm 3\sigma$

**Chebyshev's Rule**

This rule works for all distributions:

at least 75% between 2 std. dev. from the mean

$\mu \pm 2\sigma$

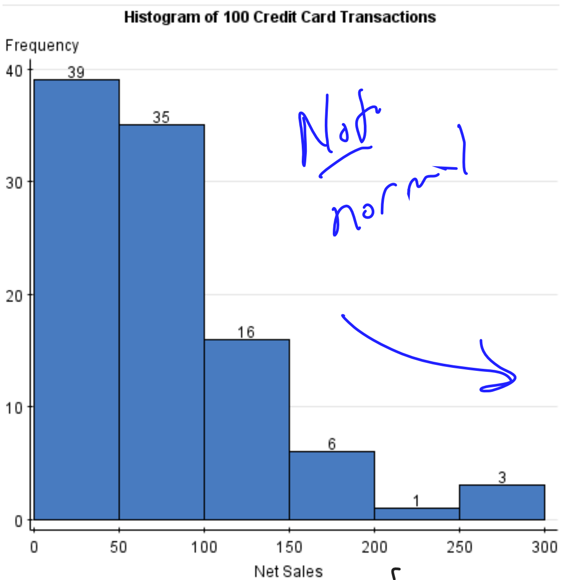
at least $\boxed{90\%}$ $\mu \pm 3\sigma$

less powerful

In the case of normal, use E.R.!

Use any time! Any case!

**Example:** Pelican Stores, a division of National Clothing, is a chain of women's apparel stores operating throughout the country. The chain recently sampled 100 in-store credit transactions in order to gain insight about the spending habits of their customers. A histogram of these transactions is given below, as well as various summary statistics:




Not normal

Outlier? Don't see any.

**Summary statistics:**

| Column | n | Mean | Variance | Std. Dev. |
|---|---|---|---|---|
| Net Sales | 100 | 77.6005 | 3098.5854 | 55.66494 |

| Median | Range | Min | Max | Q1 | Q3 |
|---|---|---|---|---|---|
| 59.705 | 274.36 | 13.23 | 287.59 | 39.6 | 101.4 |

The distribution of _credit card sales_ has a shape that is _skewed right_. The average purchase in this sample of _100_ transactions is _$77.60_, and 50% of the transactions have an amount of _$59.71_ or less. The minimum purchase was _$13.23_, while the maximum was _$287.59_. The middle _50_% of purchases in this sample were between $39.60 and _$101.40_. According to _Chebyshev's_ Rule, we expect _75_% of purchases at this store to fall between _$0_ and _$188.93_ (two standard deviations from the mean).

$77.60 \pm 2 * 55.66$

left side is negative! So put 0 to the left

## Comparing Distributions

Textbook pages: 88 – 95

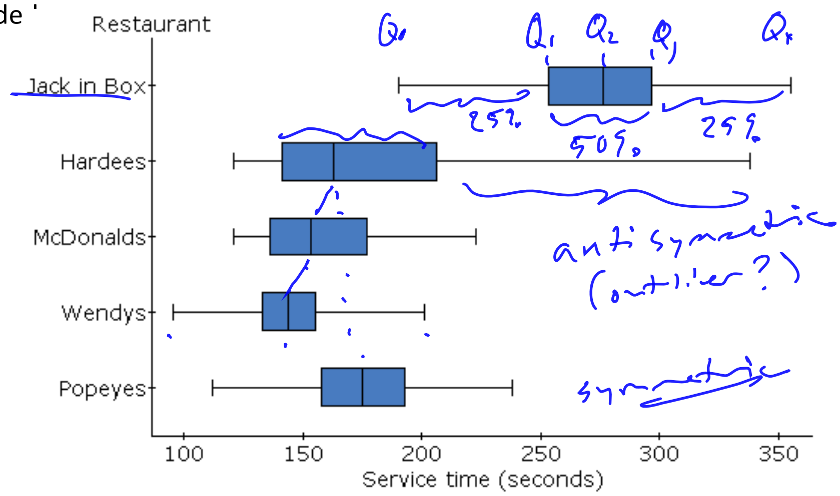
Generally, we can answer much more interesting questions when we compare two or more distributions. It can be cumbersome to compare several different histograms, so a different graphical display called a **boxplot** is often used. A boxplot is based on the 5-number summary which consists of the following five staitsitcs:

Min "$Q_0$" $Q_1$ Median "$Q_2$" $Q_3$ Max "$Q_4$"

These statistics are then used to construct a boxplot. A generic boxplot is shown below:

IQR

**Example:** A large number of fast-food restaurants with drive-through windows offer drivers and their passengers the advantages of quick service. To measure the quality of service, an organization called QSR planned a study in which the amount of time taken by a sample of drive-through customers at each of five restaurants (Popeye's, Wendy's, McDonald's, Hardee's, and Jack in the Box) was recorde '




- Which fast food chains had the most similar median drive through times?

M & H ; H & P ; M & W

- Which fast food chain has a distribution of drive thru times that is likely right skewed?

Which are asymmetric?
Hardees

- Which fast food chain had the smallest innerquartile range of drive through times?

Wendys

- Which fast food chain had the most variability according to their ranges? (Biggest range)

Hardees

- Which fast food chain was responsible for the overall fastest service in the sample? Which was responsible for the overall slowest service?

Wendys wins on almost every measure - smallest min
" max
" median
" IQR

(We receive no funding or food from Wendys)