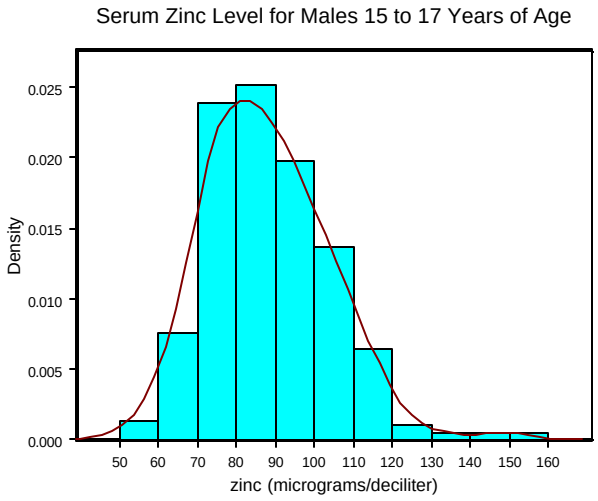
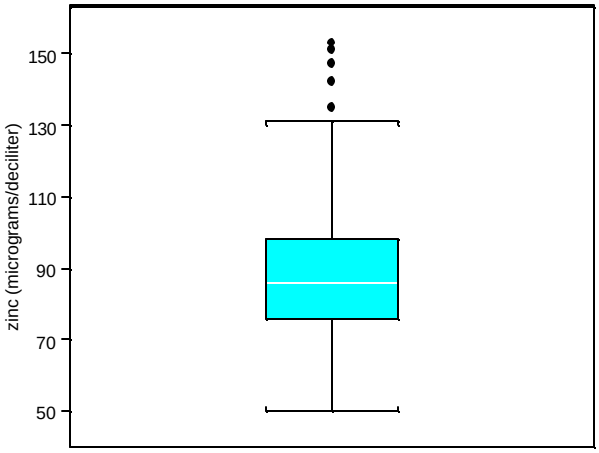
3. Numerical Summary Measures

The last chapter described ways of using tables and graphs to summarize our data. Here, we describe numerical summaries:

- Measures of Central Tendency (i.e., where's the "center" of the data?)
- Measures of Dispersion (i.e., how "spread out" is the data?)
- Proportions of data falling within certain intervals (Chebychev's rule and the empirical rule)

Note: Check out the letters of the Greek alphabet: http://www.mathacademy.com/pr/prime/articles/greek/ . We'll use a few throughout the course.





3.1. Measures of Central Tendency

- Arithmetic mean
  - uses all the data ("efficient")
  - better suited to symmetric distributions
  - sensitive to outliers

- Median—the "middle" value of the data (when ordered). a.k.a. 50^{th} p-tile or 2^{nd} quartile (Q2)
  - better suited to skewed distributions
  - "resistant" to outliers (POB "robust")

- Mode(s)—most frequently occurring value(s)
  - o can be used for all types of data
  - o may not exist
  - o not used much

E.g. Using S-Plus to get numerical summaries:

- Menu: Statistics>Data Summaries>Summary Statistics...
- Report window:

```
   ***  Summary Statistics for data in:
serzinc ***

                  zinc
     Min:  50.0000000
 1st Qu.:  76.0000000
    Mean:  87.9372294
  Median:  86.0000000
 3rd Qu.:  98.0000000
     Max: 153.0000000
 Total N: 462.0000000
   NA's :   0.0000000
Std Dev.:  16.0046888
```

3.2. Measures of Dispersion

- Range: (max. – min.)
  - o sensitive to outliers, of course!
- Interquartile range
  - o Q3 – Q2 (roughly—the range of the “middle half of the data”)
  - o not sensitive to outliers

- Determining kth percentile—data value for which at least k percent of the data values fall at or below and at least (100-k) percent fall at or above
    1) order the data
    2) calculate n*k/100
    3) if n*k/100 is an integer, average the (n*k/100)th and the ((n*k/100) + 1)th ordered data value
    4) if n*k/100 is not an integer, round up to next highest integer and use this ordered data value

- Standard deviation (and variance)
    - same units as mean
    - use when using the mean (symmetric distributions)
    - sensitive to outliers

3.3. Grouped Data

Left to you to read.

3.4. Chebychev's Inequality

Can the mean and standard deviation be used to tell us more than the "center" and "spread" of our data? Yes. We discuss 2 ways to use these measures to answer questions about the proportion of data falling within certain intervals. Before we get to Chebychev's inequality, we discuss the "empirical rule:"

- roughly 67% of the observations fall within $\bar{x} \pm s$
- roughly 95% fall within $\bar{x} \pm 2s$
- "almost all" fall within $\bar{x} \pm 3s$

E.g. Serum Zinc levels revisited.

The empirical rule is a rule-of-thumb for symmetric, unimodal ("bell-shaped") distributions. Chebychev's inequality is similar in spirit to the empirical rule, but Chebychev applies to all distributions!

Chebychev's Inequality tells us that, for $k \geq 1$, at least $(1-(1/k)^2)*100\%$ of the data falls within k standard deviations of their mean.

- at least $(1-(1/2)^2)*100\% = 75\%$ of the data falls within 2 s.d. of their mean

- at least $(1-(1/3)^2)*100\% \cong 89\%$ of the data falls within 3 s.d. of their mean.
- Note that Chebychev says at least 0% of the data falls within 1 s.d. of the mean! (thanks!) But, the empirical rule would say about 67%. Because, we do not assume anything about our distribution, it is not surprising that Chebychev is more conservative.

Summary

- Mean, median, or mode?
- IQR, range, standard deviation?
- Chebchev's or empirical rule?