Stat 145, Wed 8-Sep-2021 -- Wed 8-Sep-2021
Biostatistics
Spring 2021

Wednesday, September 8th 2021

Wk 2, We
Topic:: Summary statistics
Read:: Lock5 2.3-2.4

Warmup question Would you expect home-sale prices in Grand Rapids to be
- symmetric?
- right-skewed?
- left-skewed?

Administrative:
- difficulties to date with
  submitting to Gradescope?
  accessing WebWork?
- access to bolded dataset names in the textbook
  examples
    AllCountries, in Exercise 2.56
    April14Temps, in Example 2.15
- when you use an RStudio calculation in homework, cite
  the command you typed
  the result it gave

Q: Look at April14Temps data. It is arranged like this

	Year	DesMoines	SanFrancisco
1	1995	56.0	51.0
2	1996	37.5	55.3
3	1997	37.2	55.7
4	1998	56.0	48.7
5	1999	54.3	56.2

Would the data be understood the same way if it were arranged like this?

measurement

	City	Temp	Year
1	SanFrancisco	48.7	1998
2	SanFrancisco	55.7	1997

```
3     DesMoines      56.0     1995
4     DesMoines      54.3     1999
5     SanFrancisco   55.3     1996
6     SanFrancisco   51.0     1995
7     DesMoines      37.2     1997
8     SanFrancisco   56.2     1999
      DesMoines      56.0     1998
      DesMoines      37.5     1996
```

Quantiles/percentiles
 - concept arises for (single) quantitative var. (not for a categorical var.)

 - English monarchs data: years is quantitative
    em = read.csv("http://scofield.site/teaching/data/csv/monarchReigns.csv")
    gf_dotplot(~years, data=em)   # produces a dotplot; compare w/ histogram
      gf_dotplot(~years, data=em, dotsize=.3)
    qdata(~years, .5, data=em)   # produces .5-quantile = 50th percentile
    median(~years, data=em)   # also gives median
    qdata(~years, c(.1,.2,.3), data=em)   # produces .1-, .2-, .3-quantiles

 - terms
    median of a variable = 50th percentile of that variable
    1st quartile (Q1) = 25th percentile of that variable
    3rd quartile (Q3) = 75th percentile of that variable
    5-number summary
      gives: min, Q1, median, Q3, max
      qdata(~years, data=em)
    box-and-whisker plot
      gf_boxplot(~years, data=em)
      range = max - min   (the distance between smallest and largest values)
      IQR = Q3 - Q1   (IQR = interquartile range)
      automated outlier-flagging: the 1.5-IQR rule

Mean = average
 - formula
 - command: mean(~years, data=em)
 - sensitive to outliers
    different from median, which is "resistant to outliers"
    app at istats.shinyapps.io/MeanvsMedian/
    observations
      right-skewed corresponds to mean larger than median

left-skewed corresponds to mean smaller than median
when symmetric, mean and median are roughly equal
- where median and mean are located on histogram/dotplot

Commands introduced today: checked ones

✓ qdata - for finding quantiles of a quantitative variable
median - specifically finds the median of a quantitative variable
mean - finds the mean of a quantitative variable
✓ favstats - finds a number of values
gf_dhistogram - like histogram, but scales area to be 1
gf_density - smoothed-out histogram, area equals 1
gf_percents - like bar graph, but gives relative frequencies, not frequencies
✓ gf_dotplot - for quantitative variable without too many values
gf_boxplot - for quantitiative variable, visual depiction of 5-number summary

rep - produces a list copying a value a specified number of times
sample - produces a list drawn from a list of values

```
--------------
FURTHER THOUGHTS (not covered in class?)

Examples of bias
 - In surveys: scenarios
    "Local library is sponsoring talk by Planned Parenthood representative.
    Do you think our community should sanction baby-killers?"
      leading questions
    Ann Landers on whether parents would choose to have children in do-over
      voluntary response bias
    Literary digest survey leading into 1936 election
      poor sampling frame
    "Do you take elicit drugs?"
      embarrassing question
    "How old were you when you stopped taking baths?"
      imperfect recall
    "Do you prefer this first soft drink, or the second one?"
      order of presentation should be random to avoid bias
    "Which candidate did you vote for?", asked outside only during hours 7-9 am
      convenience sample

 - In experiments
    measuring instrument not calibrated
    order of treatment

experiments and observational studies
 - both types of studies may have explanatory/response vars
 - observational study does not attempt to assign explanatory values
    ==>  when difference appears significant, cannot rule out lurking vars
     in presence of significant difference only say vars have an association
 - blocking
    identifying specific (non-factor) variables to even out
      example: soil, sunlight in agricultural studies
      example: sex, smoking status, age in drug studies
    matched pairs: each "case" contributes two values
      case might be a person: contributes "control" and "treatment" values
      case might be identical twins: one twin is "control" for the other
      case might be "married couple": one spouse is "control" for the other

Measures of "center" (or "central tendancy")
 - what they are
```

```
    mode = location/value occurring most frequently
      meaningful for both categorical and quantitative variables
    median = 50th percentile
      meaningful for quantitative variables only
      resistant to outliers
    mean = average
      meaningful for quantitative variables only
      sensitive to outliers
 - visualizing on a distribution
    mean is balancing point
    median cuts values/area in half

Measures of "spread"
 - what they are
    range: sensitive to outliers
    IQR: resistant to outliers
    standard deviation: sensitive to outliers

Q3: 5-number summary has 4 other numbers besides the median.
    Are these other numbers resistant to outliers, or are they sensitive?
```