Stat 145, Tue 7-Sep-2021 -- Tue 7-Sep-2021
Biostatistics
Spring 2021

Tuesday, September 07th 2021

Due:: PS01 due at 11 pm
Due:: PS02 due at 11 pm

Tuesday, September 07th 2021

Wk 2, Tu
Topic:: Shape/form of distribution
Topic:: Quantiles and mean
Read:: Lock5 2.1-2.2
HW:: WW Descriptive1 due Sat.

administrative:
- take attendance
- mention quiz

distributions - gives values taken, and how often
- univariate
- "displayed" using
  categorical vars: frequency tables (tally()), bar graphs (gf_bar())
    frequency vs. relative frequency
  quant vars: frequency tables (tally()), histograms (gf_histogram()), other?
- form/shape
  number of modes
  symmetric vs. skew (quantitative only)

bivariate displays
- used to begin investigations of associations
- two categorical vars.
  contingency table
  side-by-side bar graphs
- one categorical, one quantitative
  contingency table?

```
    side-by-side dotplots/histograms
 - two quantitative
    scatterplot
    association can be equated with "a nonhorizontal pattern exists"
    positive vs. negative association (not the only possibilities)

relationships/associations between variables:
 ssurv <- read.csv("http://scofield.site/teaching/data/csv/ssurv.csv")

 tally:
   tally(~selfhandedness | sex, data=ssurv)
   tally(~selfhandedness | sex, data=filter(ssurv, selfhandedness!=""))

   tally(~selfhandedness | dadhandedness, data=filter(ssurv, dadhandedness!=""))

 addmargins():
   addmargins(tally(~selfhandedness | dadhandedness, data=ssurv))     or, an alternate method
   tally(~selfhandedness | dadhandedness, data=ssurv) %>% addmargins()
                                                          ↑
                                                    uses "piping"
 gf_bar()
   gf_bar(~selfhandedness|dadhandedness, data=filter(ssurv, dadhandedness!=""))

 gf_histogram()          quantitative      categorical
   gf_histogram(~speedtickets | oncampus, data=ssurv)
   gf_histogram(~speedtickets | oncampus, data=filter(ssurv, oncampus!=""))

Associations between variables
 - context: bivariate data
    from ssurv.csv
      sex and handedness: two categorical vars
      We used tally(~ sex | handedness, data = ssurv) to obtain contingency table
      proportion of left-handed women to all women in sample: 15/139 } no apparent
      proportion of left-handed men to all women in sample: 16/140 } association

      speeding tickets vs. off-campus: one quantitative, one categorical var.
        side-by-side histograms (see above for command)
```

speeding tickets vs. number of cds : two quantitative vars.

Scatterplot using # of cds as explanatory variable comes from command

gf_point(speedtickets ~ cds, data = ssurv)

Further notes not covered(?) in class on Sept. 7

Example:

- histogram of cds in ssurv

```
gf_histogram(~cds, data=ssurv, color="black", bins=10)
gf_dhistogram(~cds, data=ssurv, color="black", bins=10)

features
  the color switch is unnecessary, but delineates the bins
  it isn't obvious there are 10 bins, since some are empty
  gf_dhistogram doesn't give count in bins; proportionally adjusts area = 1

shape is subject to bin size (more bins means thinner bins)
density plot attempts to smooth things out
  gf_density(~cds, data=ssurv)
verbal description
  unimodal (describes number of major peaks)
  right-skewed (most-frequent words: symmetric, right-/left-skewed)
```

- histogram of eruptions in faithful

```
Q: Would you expect home-sale prices in Grand Rapids to be
    symmetric?
    right-skewed?
    left-skewed?

Discuss: Is there a variable you can think of that would be left-skewed?
```

- histogram of randomnum in ssurv

```
    gf_histogram(~randomnum, bins=20, data=filter(ssurv, randomnum <= 20))
      might have expected a flat (uniform) distribution

Uniform distributions (all values occur equally) can arise in categorical data
```

- coin flips (H, T)

```
  coin = c("H","T")
```

```
    resultOfFlips = sample(coin, 500, replace=TRUE)
    tally(~resultOfFlips)
    gf_bar(~resultOfFlips)
    gf_percents(~resultOfFlips)

 - rock, paper, scissors?
    see StatKey: One Categorical Variable, under Descriptive Statistics

 - days of the week for births in 2015
    scofield only can do this example using data frame all2015Births

 - when distribution of categorical variable is not uniform
    shape isn't generally relevant (due to resequencing of bars)
    can still identify mode(s)

Quantiles/percentiles
 - concept for quantitative vars only
 - English monarchs: 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
    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
      fivenum(~years, data=em)
    box-and-whisker plot
      gf_boxplot(~years, data=em)

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:
 qdata - for finding quantiles of a quantitative variable
 median - specifically finds the median of a quantitative variable
 fivenum - delivers the 5-number summary of a quantitative variable
 mean - finds the mean of a quantitative variable
 sample - produces a list drawn from a list 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
```