Stat 145, Wed 10-Feb-2021 -- Wed 10-Feb-2021
Biostatistics
Spring 2021

Wednesday, February 10th 2021

Wk 2, We
Topic:: Shape/form of distribution
Topic:: Quantiles and mean
Read:: Lock5 2.1-2.2

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 of histogram is subject to bin size (more bins means thinner bins)
  density plot attempts to smooth things out (no bins anymore)

```
gf_density(~cds, data=ssurv)
```

  verbal description
    unimodal (describes number of major peaks) - form
    right-skewed (most-frequent words: symmetric, right-/left-skewed) - shape

- histogram of eruptions in faithful

Think about these

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)

Variable in these 4 cases is categorical

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)

More next time

- 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

— This is used in the WebWork H.W.

- 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