Jan. 25, 2017.

Housekeeping

- *Hidden Figures* extra credit deadline extended by 1 week (see Canvas prompt for details)
- HW & quiz solutions online for those interested (thank you to students who contributed!)
- Homework due Friday (as usual, check Canvas)
- Writing assignment due Monday

Last time: Qualitative data representation

QUESTIONS?

This time:
- Finish qualitative visualization
- Sampling

"Final" points abt. qualitative data visualization:

- See pie charts on p. 19 — 19(a) on the left is "unsorted" (wedges of any size are interspersed in an unpredictable way), but 19(b) on the right is "sorted" — wedges are in decreasing order, going around clockwise.

WHICH IS MORE VISUALLY APPEALING?
— " — EASIER FOR THE VIEWER TO UNDERSTAND?

Key Idea: The distinction btwn. sorted & unsorted pie charts is like the distinction btwn. Pareto and bar charts.

The only time it is really necessary/preferable to use a bar chart over a Pareto chart is if the categories have some "other" natural order besides frequency... e.g., the categories are time periods, or follow another kind of spectrum or chunks of a spectrum (maybe physical location East → West, or temperature, etc.)

L4, ct'd.

Sampling (p. 19).

Recall: Diff. btwn. a sample & a population. (Why do we need sampling?)

A SAMPLE SHOULD HAVE THE SAME CHARACTERISTICS AS THE POPULATION IT REPRESENTS.

e.g., · age ✓ · religion · education level · income
· race · ethnicity · party affiliat'n · sex/gender
· geographic location

Sometimes, random sampling accomplishes this well (but statisticians always do reality checks!).

> DEF. A random sampling method gives each member of the populat'n an equal chance of being selected for the sample. (There are several different random sampling methods).

- Simple random sample: Assign a unique number to each member of the population, and pull numbers randomly from a hat (or from a random number generator) until the sample has the desired size.

· Simple random sampling, ct'd.

Example: Say I want a committee of 4 students chosen at random from among our classmates. I'll assign each of you a unique # from 1 – 26 (size of our class), and use a random # generator to choose 4 random #s from 1 – 26.

See p. 20 for the TI-83 & 84 random # generator.

Most r.n.g. computer programs generate random numbers in the interval [0, 1]. Example: 0.40581, etc.

Q: How to make such a program give random #s in the interval [1, 26] ?
(i.e., How to modify the output?)

A: Multiply the random # by 26, & take the next highest integer:

$$\underbrace{\text{ceil}}\left(26 \cdot \underline{\text{rand}()}\right)$$

"ceiling" of any # is the next highest integer.

(alt.) A: Depends on computer/calculator — maybe rand( ) has input.

- Stratified sample : divide the entire population into groups, then use simple rand. sampling within each group to choose a proportionate number of members of the sample. (The groups are called strata.)

Example.

STRATUM 1 | STRATUM 2

1 2 3 4 5 6 7 8 9 10 11 12 13 | 14 15 16 17 18 19 ... 26

Choose 2 from each stratum.

Q. Why might this appear to be a more effective way of getting a representative sample than simple random sampling? What's the difference/ what kind of samples does stratified sampling prevent?

A. With simple random sampling, it is possible that the sample consists of people whose IDs are all close together — stratified sampling (when there are > 2 strata) makes this impossible.

The probability (likelihood) that an individual in the above example will end up in the sample is :

(1st stratum: $\frac{2}{13}$) ; (2nd stratum: $\frac{2}{13}$).

Stratified ~~Cluster~~ sampling, ctd.

What if the strata are different sizes? — Remember, you can choose the strata (they don't just have to be based on ID #!).

For example: Suppose tht. in a company, there are the following staff:

- Male, full-time: 90
- Male, part-time: 18
- Female, full-time: 9
- Female, part-time: 63

TOTAL: 180

Suppose, further, that we are asked to take a sample of 40 staff, stratified into the above categories.

(1) Calculate ~~the percentage of each group~~ what proportion (percent) of the total each group/stratum comprises:

- % male, FT : $\frac{90}{180} = 0.5 = 50\%$
- % male, PT : $\frac{18}{180} = \frac{1}{10} = 0.1 = 10\%$
- % female, FT : $\frac{9}{180} = \frac{1}{20} = 0.05 = 5\%$
- % female, PT : $\frac{63}{180} = 0.35 = 35\%$

→

(2) Compute the # of people from each stratum that should go into the sample by multiplying the total # of people in ~~stratum~~ the sample by that stratum's proportion of the total # of people in the entire population.

In our sample, we wanted 40 people. So:

- Male, FT: $0.5(40) = \frac{1}{2}(40) = 20$. Choose 20 FT males at random.
- Male, PT: $0.1(40) = \frac{1}{10}(40) = 4$. Choose 4 PT males.
- Female, FT: $0.05(40) = \frac{1}{20}(40) = 2$. Choose 2 FT females.
- Female, PT: $0.35(40) = 14$. Choose 14 PT females.

... could "simplify" process by combining steps, e.g.:

- Take $\left(\frac{90}{180}\right) 40 = \frac{1}{2}(40) = 20$ FT males.
- Take $\left(\frac{18}{180}\right) 40 = \frac{1}{10}(40) = 4$ PT males.
- Take $\left(\frac{9}{180}\right) 40 = \frac{1}{20}(40) = 2$ FT females
- Take $\left(\frac{63}{180}\right) 40 = \frac{63 \cdot 2}{9} = 7 \cdot 2 = 14$ PT females.