Data, Univariate Statistics & Statistical Inference

- Constants & Variables
- Univariate Statistics
  - typicality, variability, & shape measures
- Combining Univariate Statistical Information
- Parameters & Statistics
- Samples & Variability

Measures are either Variables or Constants

Constants

– when all the participants in the sample have the same value on that measure/behavior

Variables

– when at least some of the participants in the sample have different values on that measure

– either qualitative or quantitative (more later)

How do we get constants ?

Homogeneous group

- everybody really has the same value

Imprecise measure

- age for a group of 1st graders if measured as "# years old " vs. "# days old"

Inadequate sample

- smaller samples are more likely to have constants
- non-representative (incomplete) samples more likely to have constants

Two Main Types of Variables

Qualitative (or Categorical)

– Values are mutually exclusive

– Different values represent different categories / kinds

– Discrete

Quantitative (or Numerical)

– Values are mutually exclusive

– Different values represent different amounts

– Discrete or Continuous

- discrete
  – no "partial counts" just "whole numbers"
  – e.g., how many siblings do you have
- continuous
  – fractions, decimals, parts possible
  – must decide on level of precision
  – e.g., how tall are you = 6' 5'11" 5'10.65"

Practice w/ Types of Variables

Is each is "qual" or "quant" and if quant whether discrete or continuous ?

• gender	qual
• age	quant - continuous
• major	qual
• # siblings	quant - discreet

Your Turn -- make up one of each kind…

qualitative

quantitative - discreet

quantitative - continuous

Oh yeah, there's really two more variables types …

Binary variables

- qualitative variables that have two categories/values
  - often because we are "combining categories" (e.g., if we grouped married, separated, divorced & widowed together as "ever married" and grouped used "single" as the other category
- two reasons for this…
  - categories are "equivalent" for the purpose of the analysis -- simplifies the analysis
  - too few participants in some of the samples to "trust" the data from that category
- often treated as quantitative because the statistics for quantitative variables produce "sensible results"

The other of two other variables type …

Ordered Category Variables

- multiple category variables that are formed by "sectioning" a quantitative variable
  - age categories of 0-10, 11-20, 21-30, 31-40
  - most grading systems are like this 90-100 A, etc.
- can have equal or unequal "spans"
  - could use age categories of 1-12 13-18 19-21 25-35
- can be binary -- "under 21" vs. "21 and older"
- often treated as quantitative because the statistics for quantitative variables produce "sensible results"
  - but somewhat more controversy about this among measurement experts and theorists

A quick word of caution about the category "other" !!!

Many categorical variables have lots of possible categories, with widely ranging frequency or likelihood!!!

Pets → dogs & cats head the list. Fish, birds & rodents are fairly common. And we all know a few folx with some "less common pets"!!

The same is true of all demographic and describing variables, race/ethnicity/heritage and sex/gender/orientation being important examples.

An unfortunate, but common, practice is to "bundle" less frequent categories into an "other" category:

Pets → 1=dogs 2=cats 3=fish 4=rodents 5=other

For any variable, the "other" category likely bundles several different categories that are not really equivalent!

Combing them can provide very misleading results, especially when you are examining how other variables are related to this categorical variable!!!!

But the fact remains that some categories are less common! What are we supposed to do???

First → We have to carefully define the population we are interested in studying – we can't study "everybody" or "everything" in every study – gotta make active, informed choices, based on the literature you are reading and the research community you are working in!

Second → use stratified sampling techniques to get large-enough samples of the categories you want to study – it can be "expensive" in time & mondy– but a bad sample is the easiest way to end up with poor statistical conclusion validity!

Measures of Typicality (or Center)

• the goal is to summarize the entire data set with a single value

stated differently …

• if you had to pick one value as your "best guess" of the next participant's score, what would it be ???

Measures of Variability (or Spread)

• the goal is to tell how much a set of scores varies or differs

stated differently …

• how accurate is "best guess" likely to be ???

Measures of Shape

• primarily telling if the distribution is "symmetrical" or "skewed"

Measures of Typicality or Center (our "best guess")

Mode -- the "most common" score value

- used with both quantitative and categorical variable

Median -- "middlemost score" (1/2 of scores larger & 1/2 smaller

- used with quantitative variables only
- if an even number of scores, median is the average of the middlemost two scores

Mean -- "balancing point of the distribution"

- used with quantitative variables only
- the arithmetic average of the scores (sum of scores / # of scores)

Find the mode, median & mean of these scores… 1 3 3 4 5 6

Mode = 3 Median = average of 3 & 4 = 3.5

Mean = $\Sigma X / n = (1 + 3 + 3 + 4 + 5 + 6) / 6 = 22/6 = 3.67$

Measures of Variability or Spread -- how good is "best guess"

categories -- used with categorical variables

Range -- largest score - smallest score

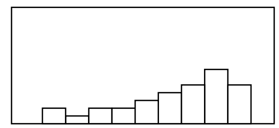
Standard Deviation (SD, S or std)

- average difference from mean of scores in the distribution
- most commonly used variability measure with quant vars
- pretty nasty formula -- we'll concentrate on using the value
- "larger the std the less representative the mean"

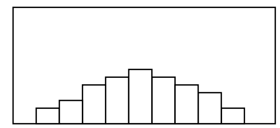
Measures of Shape

Skewness -- summarizes the symmetry of the distribution

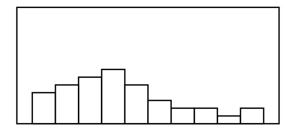
- skewness value tells the "direction of the distribution tail"
- mean & std assume distribution is symmetrical



Skewness = "-"



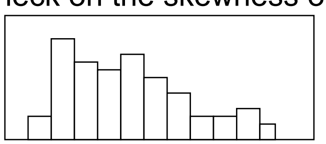
Skewness = "0"



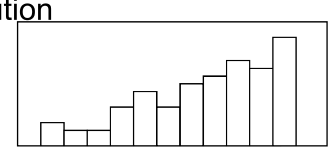
Skewness = "+"

Using Median & Mean to Anticipate Distribution Shape

- When the distribution is symmetrical mean = median (= mode)
- Mean is influenced (pulled) more than the median by the scores in the tail of a skewed distribution
- So, by looking at the mean and median, you can get a quick check on the skewness of the distribution



Med = 42 < $\overline{X}$ = 55



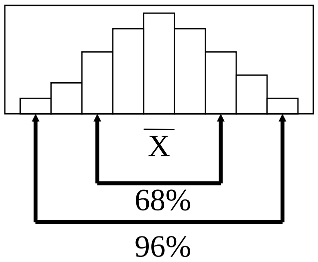
$\overline{X}$ = 56 < Med = 72

Your turn -- what's the skewness of each of the following distributions ?

- mean = 34 median = 35 — 0 skewness
- mean = 124 median = 85 — + skewness
- mean = 8.2 median = 16.4 — - skewness

Combining Information from the Mean and Std

How much does the distribution of scores vary around the mean ?



If the distribution is symmetrical

- 68% of the distribution falls w/n +/- 1 SD of the mean
- 96% of the distribution falls w/n +/- 2 SD of the mean

Tell me about score ranges in the following distributions ...

$\overline{X}$=10 SD=5
68% 5-15 96% 0-20

$\overline{X}$=20 SD=3
68% 17-23 96% 14-26

"Beware Skewness" when combining the mean & std !!!

Consider the following summary of a test

- mean %-correct = 85 std = 11
- so, about 68% of the scores fall within 74% to 96%
- so, about 96% of the scores fall within 63% to 107%

Anyone see a problem with this ?!? **107%** ???!!??

What "shape" do you think this distribution has ? - skewed

Which will be larger, the mean or the median? Why think you so ??

mean < mdn

Here's another common example…

How many times have you had stitches ?

- Mean = 2.3, std = 4 68% 0-7.3 96% 0-10.3

Be sure ALL of the values in the score range are possible !!!

When you're doing the +2/-2 Std check for skewness, you have to be sure to consider the "functional range" of the variable for the population you are working with.

For example…

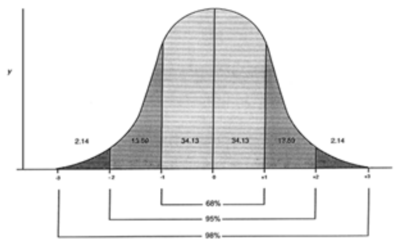
Age

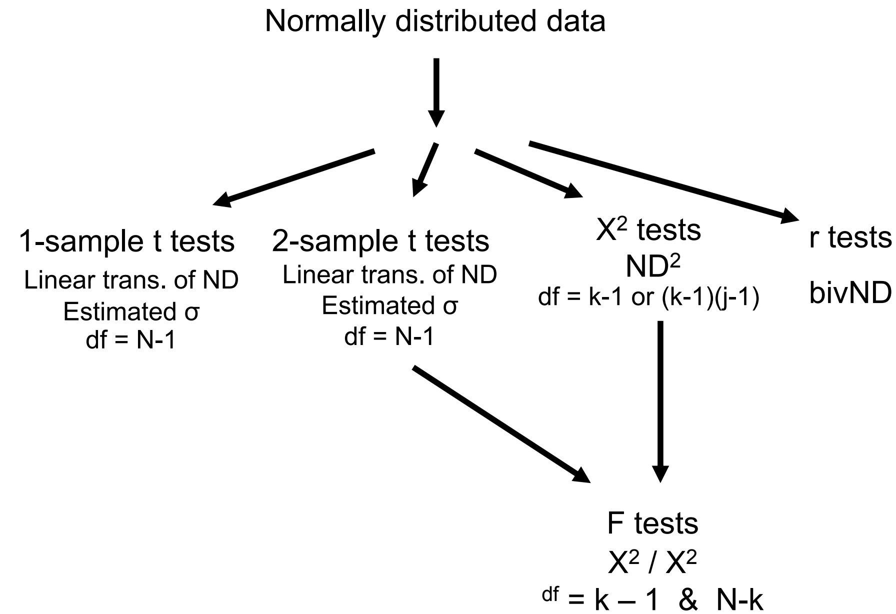
- lowest possible numerical value is 0
- but among college students the minimum is around 17
- so, what about a distribution from a "college sample" with…
  - mean = 20 and std = 1.5 17-23 -- seems ok
  - mean = 20 and std = 3 14-26 – 14 seems young → + skew
- so, what about a distribution from a "sample of retirees" with…
  - mean = 96 and std = 8 80-112 – seems a bit old & - skew
  - mean = 76 and std = 8 60-92 – seems ok

Normal Distributions and Why We Care !!

- Statistics provide formulas to calculate more efficient summaries of the data
- These stats are then the bases for other statistics that test research hypotheses (e.g., r, t, F, X^2)
- The "catch" is that the formulas for these statistics (and all the ones you will learn this semester) depend upon the assumption that the data come from a population with a normal distribution for that variable.
- Data have a normal distribution if they have a certain shape, which is represented by a really ugly formula (that we won't worry about!!).

$$f(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$





Some new language …

Parameter -- summary of a population characteristic

Statistic -- summary of a sample characteristic

Just two more …

Descriptive Statistic -- calculated from sample data to describe the sample

Inferential Statistic -- calculated from sample data to infer about a specific population parameter

Reviewing descriptive and inferential statistics …

The major difference between descriptive and inferential statistics is intent – what information you intend to get from the statistic

Descriptive statistics

- obtained from the sample
- used to describe characteristics of the sample
- used to determine if the sample represents the target population by comparing sample statistics and population parameters

Inferential statistics

- obtained from the sample
- used to describe, infer, estimate, approximate characteristics of the target population

Parameters – description of population characteristics

- usually aren't obtained from the population (we can't measure everybody)
- ideally they are from repeated large samplings that produce consistent results, giving us confidence to use them as parameters

Another look at variability and inference ...

From those sample data we compute …

inferential mean

- our best guess of the population mean
- also our best guess of the score for any member of the population

inferential std

- our best guess of the variability of individual scores around the population mean
- the smaller the standard deviation…
  - the less scores vary around the mean in the population
  - the better the mean is as a guess of each person's score

Is there any way to estimate the accuracy of our inferential mean???

Yep -- it is called the Standard Error of the Mean (SEM)

and it is calculated as … Inferential std from sample

$$\text{SEM} = \frac{\text{std}}{\sqrt{n}}$$

sample size

The SEM tells the average sampling mean sampling error -- by how much is our estimate of the population mean wrong, on the average

This formula makes sense ...

- the smaller the population std, the more accurate will tend to be our population mean estimate from the sample
- larger samples tend to give more accurate population estimates

So now you know about the two important types of variation…

- variation of population scores around the population mean
  - estimated by the inferential standard deviation (std)
- variation in sample estimates of the population mean around the true population mean
  - estimated by the standard error of the mean (SEM)

When would we use each (hint: they're in pairs) …

The mean Exam 1 score was 82% this semester. How much do this semesters Exam 1 scores vary?	std
The mean Exam 1 score was 82% this semester. How much will this mean likely vary from the pop mean of all Exam 1 scores?	SEM
The average depression score of patients currently receiving treatment in the PCC is 73.2. How much does this vary from the pop mean of all the patients ever seen there?	SEM
The average depression score of patients currently receiving treatment in the PCC is 73.2. How much do the patient's scores vary from each other?	std