Data & Univariate Statistics

- Constants & Variables
- Operational Definitions
- Organizing and Presenting Data
- Tables & Figures
- Univariate Statistics
    - typicality, variability, & shape measures
- Combining Univariate Statistical Information

Measures of behaviors & characteristics 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

Qualitative (or Categorical) Variables

- Different values represent different categories / kinds

Quantitative (or Numerical) Variables

- Different values represent different amounts

Practice w/ Types of Variables

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

• species	qual
• age	quant
• major	qual
• # siblings	quant

Common Confusions

Quality is often a quantitative variable ➔ How much quality?

Watch out for #s that represent kinds ➔ perch = 1 & bass = 2

Color ➔ commonly qualitative but really quantitative (wavelength)

Conceptual & Operation Definitions

- With constants and 3 different types of variables, the “name” of the measure being examined might not be enough to properly identity “what is being measured how”
- Operational definition
  - describes a variable representiong aa behavior or characteristic by telling how it is actually measured or manipulated
  - often by specifying the “question” & the “response values”

Your turn -- tell the type of each below ...

- Gender -- “How strongly do you identify with the stereotypic characterization of your gender ?” quant

  1 = not at all 2 = somewhat 3 = very much 4 =totally

- Siblings -- “Are you an only child?” qual

  1 = no 2 = yes

- Major -- “How many different majors have you had at UNL or other colleges?” quant

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

Binary variables

- qualitative variables that have two categories/values
  - often 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!

Ways to Present Data

When we first get our data they are often somewhat messy

- usually just a “pile” of values for each participant
- hopefully with an “identifier” (name or number) for each
- there are three common ways of presenting the data

- Listing of complete ‘raw’ data
  - Pat got 80%, Kim got 75%, Dave got 90% ……
  - complete but very cumbersome
- Organized display of the data
  - frequency distribution table
  - frequency polygon
  - histogram
  - bar graph
- Statistical summary of the data
  - use a few values to represent the whole data set

A frequency distribution table starts with a count of how many participants got each value of the variable.

Example: A sample of student ages is taken:

Johnan, 21; Abdul, 18; Jesus, 18; Riley, 20; Pat, 19; Thorin, 24; Todd, 18; Zashesh, 22; Lilly, 19; Glenn, 20…

The ages (scores) are arranged into a **frequency distribution table**

- score values are listed in the "x" column from lowest (bottom) to highest (top)
- all score values in the range are listed (whether someone has that score or not)
- *f* (frequency) column tells how many there were of each score value
- sum of the *f* column should equal n (the number of participants/scores)

x	*f*
24	1
23	0
22	1
21	1
20	2
19	2
18	3
	n = 10

Notice that we have "organized" and summarized the data, but no longer know who has what scores.

This could also be done with a qualitative variable.

When working with continuous quantitative variables you'll have to pick how "precise" your score values will be -- in this case we chose "closest whole year".

We can augment the basic frequency table by adding columns of any of the following...

- cumulative frequency (how many with scores this large or smaller)
- the proportion of the sample with each score (f/n)
- cum proportion (what prop of sample has scores this large or smaller)
- % of the sample with each score (proportion x 100 or f/n x 100)
- cum % (what % of sample has scores this large or smaller)
- table might be completed using grouped scores (see below)

x	*f*	cum *f*	prop	cum prop	%	cum %
24	1	10	.1	1.0	10	100
23	0	9	0	.9	0	0
22	1	9	.1	.9	10	90
21	1	8	.1	.8	10	80
20	2	7	.2	.7	20	70
19	2	5	.2	.5	20	50
18 n = 10	3	3	.3	.3	30	30

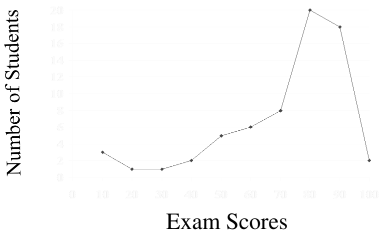
Grouped Frequency table

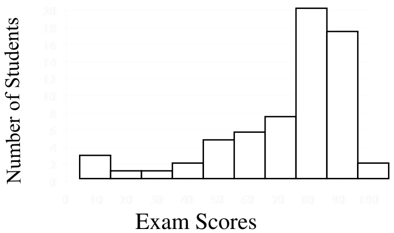
X	*f*	cum*f*	cum%
24-25	1	10	100
22-23	1	9	90
20-21	3	8	80
18-19	5	5	50

You can't do cumulative columns when working with a categorical variable -- there's no "right way" to line-up the score values !!

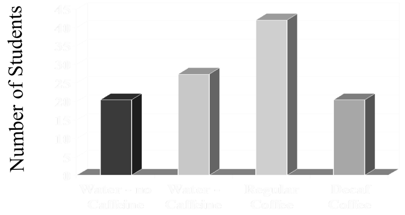
The frequency table can easily be changed into various types of graphs...

Frequency polygon & Histogram - use with quantitative data





Bar Graph -- use with qualitative data



Yep, the difference between a histogram and a bar graph is whether or not the bars are "snuggled"

Statistical Summaries --

The idea is to use a few summary values to describe the distribution of scores -- usually telling three things ...

- Typicality -- what's a typical or common score for these data
- Variability -- how much do the scores vary from "typical"
- Shape -- the shape of the distribution

The statistics we are about to explore are called…

Univariate Statistics

… because they are summarizing the information from a single variable.

Somewhat different univariate statistics are used for qualitative and quantitative variables. But, before we get into the specific statistics, let's consider how it is that they summarize the data ...

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 = (1 + 3 + 3 + 4 + 5 + 6) / 6 = 22/6 = 3.67

Means of quantitative & binary variables …

The mean is the most commonly used statistic to describe the "center" or "average" of a sample of scores.

The mean can be used with either quantitative or binary variables

- Quantitative variables – the mean tells the average
- Binary variables – "the decimal portion of the mean tells the proportion of the sample with the higher code value"

Huh ????

Say we entered a code of "1" for each participant who was a novice and a code of "2" for every expert.

We could compute the mean of the numbers as

codes for 9 participants 1 2 1 2 2 2 2 1 2

with a sum of 15 and a mean of 15/9 = 1.67

the ".67" tells us that 67% or 2/3 of the sample is experts

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

# categories -- used with categorical variables

Range -- largest score - smallest score

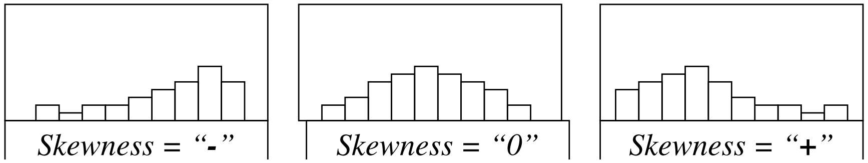
Standard Deviation (SD, S or std)

- – 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

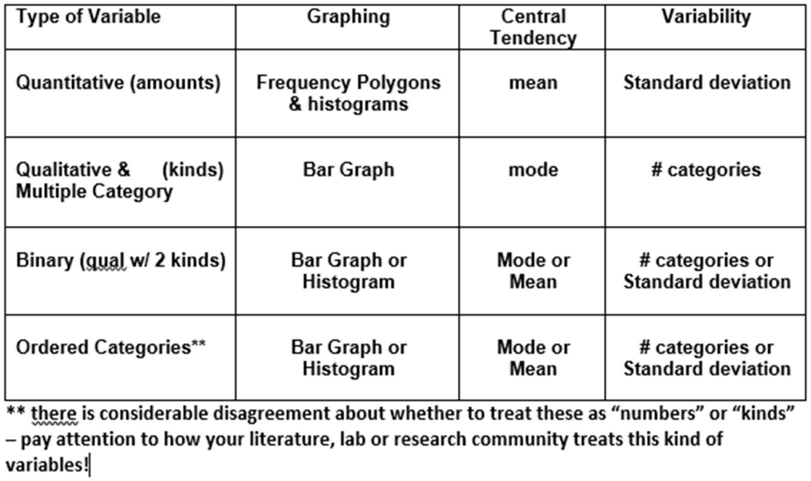
Skewness -- summarizes the symmetry of the distribution

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



Skewness = "-" *Skewness = "0"* *Skewness = "+"*

A quick summary…

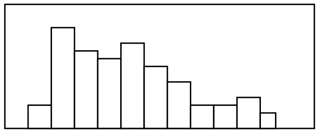


Type of Variable	Graphing	Central Tendency	Variability
Quantitative (amounts)	Frequency Polygons & histograms	mean	Standard deviation
Qualitative & (kinds) Multiple Category	Bar Graph	mode	# categories
Binary (qual w/ 2 kinds)	Bar Graph or Histogram	Mode or Mean	# categories or Standard deviation
Ordered Categories**	Bar Graph or Histogram	Mode or Mean	# categories or Standard deviation

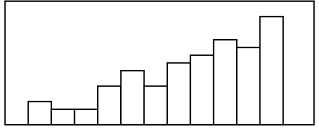
** there is considerable disagreement about whether to treat these as "numbers" or "kinds" – pay attention to how your literature, lab or research community treats this kind of variables!

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 < $\bar{X}$ = 55



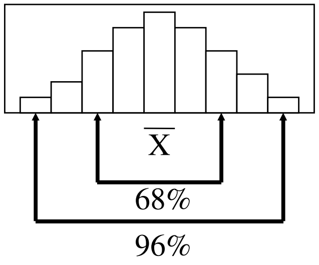
$\bar{X}$ = 56 < Med = 72

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 ...

$\bar{X}$=10 SD=5	$\bar{X}$=20 SD=3
68% 5-15 96% 0-20	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

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 …

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

Inferential std from sample

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 the Exam 1 scores vary? std

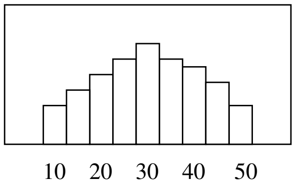
The mean Exam 1 score was 82% this semester. How much will this mean likely vary from the true mean of all Exam 1 scores? SEM

The average depression score of patients currently receiving treatment in the PCC is 73.2. How much does this vary from the true 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

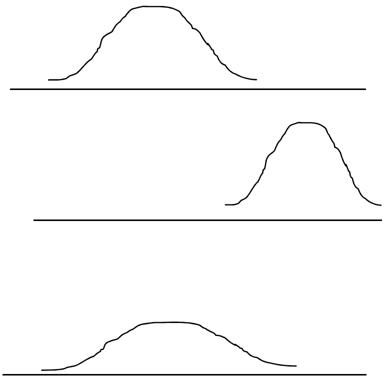
Normal Distributions and Why We Care !!

- As we mentioned earlier, we can organize the sample data into a histogram, like on the right.
- However, this does not provide a very efficient summary of the data.
- Univariate statistics provide formulas to calculate more efficient summaries of the data (e.g., mean and standard deviation)
- 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!!).

- Normal distributions generally look like well-drawn versions of those shown to the right.
- All normal distributions...
  - are symmetrical
  - have known proportions of the cases within certain regions of the distribution (68% & 96% stuff)
- Normal distributions differ in their …
  - centers (means)
  - spread or variability around the mean (standard deviation)



Nearly all the statistics we'll use in this class assume that the data are normally distributed. The less accurate this assumption, the greater the chance that our statistical analyses and their conclusions will be misleading.

A bit about computational notation…

The summation sign → Σ ← is the main symbol used.

- It means to sum, or add up, whatever is to the right of the sign
- The two versions you'll see when during hand calculations of the univariate stats are ΣX & ΣX^2

Also → N ← means the number of participants/numbers

Participant #	X	X^2
1	5	25
2	4	16
3	3	9
4	4	16
N= 4	$\Sigma X = 16$	$\Sigma X^2 = 66$

Calculating the mean …

$$\text{Mean} = \frac{\Sigma X}{N} = \frac{16}{4} = 4$$

Calculating sum of squares

$$SS = \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 66 - \frac{16^2}{4} = 2$$