Univariate Data Description: One Variable at a Time

- Topics:
  - Terminology for different types of variables
  - Summarizing different types of variables
    - Categorical: Frequency, proportion, and percentage
    - Quantitative: Central tendency, dispersion, asymmetry
    - Computation: Mean, variance, SD, skewness, and kurtosis
    - **Using SAS, STATA, and R to do all of this**
  - Bonus: Best practices for working with datasets

***Note: there is no separate example document for this unit; videos will demonstrate how to work with each program's files*

Types of Variables

- Goal: identify potential types of variables in quantitative data
  - Big picture: **categorical** or **quantitative**?
- This “types” taxonomy will guide **two things** about each variable:
  - What measures can be used to **summarize** its salient features
  - How it can be used in subsequent **analysis** (statistical models)
  - Note: this is related to traditional levels of measurement, but I am approaching it from more of a “how to model them” perspective
- Apparent purpose: Review **univariate descriptive statistics**
  - “univariate” = one variable at a time (as opposed to “multivariate”)
  - “descriptive” = not testing anything, just describing sample data
  - “statistics” = characteristics of a sample (from a population)
- Actual purpose: Give you some familiar ideas with which to begin to use **unfamiliar software** (using GSS practice data)
  - Watch my videos to see how I got results from SAS, STATA, and R! ☺

Categorical Variables: *Numbers are just **labels***

- **Binary** (dichotomous) = 2 choices (best coded as 0 or 1)
  - e.g., dead or alive; pregnant or not
- **Nominal** = 3+ unordered choices
  - e.g., favorite type of pet, degree program
- **Ordinal** = 3+ choices with some natural (undeniable) order, but the distances between the values don't mean anything
  - 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree
  - Equally ordinal values: 1, 20, 300, 4000
- Synonyms for a "**categorical**" variable: discrete variable, qualitative variable, grouping variable, factor variable in R, CLASS variable in SAS (stands for "classification variable")

Quantitative Variables: *Numbers are numbers*

- **Interval** measurement → equal distances between values
- Many quantitative variables have **1 or 2 boundaries**
  - **Binomial** = number of occurrences out of known possible
    - e.g., # correct on a test, has **2 boundaries**: 0 and total possible
    - a variable corrected for different possible totals (by computing proportion correct or rate of occurrence) would still be treated as binomial when predicted (just bounded by 0 and 1 instead)
  - **Count** = number of occurrences out of unknown possible
    - e.g., # of cigarettes smoked each day, has **1 boundary** of 0 (or 1) (only whole numbers used, maximum could be any positive number)
  - Btw, count variables have special cases involving **zero values**:
    - No zeros possible? → "zero-*truncated* count"
    - More zeros than expected (due to mixture)? → "zero-*inflated* count"

Quantitative Variables: *Numbers are numbers*

- Other **quantitative** variables are "**continuous**" (still with interval measurement in which the numbers are numbers)
  - But **continuous means unbounded →** can theoretically go on forever in either direction AND take non-integer values
  - Although in this semester's general linear models (GLMs) our predictors can be any type of variable, **all our outcomes must be *plausibly* continuous with interval measurement**
    - This is because GLMs use a *conditional* normal distribution (stay tuned)
    - Otherwise you need "general*ized* linear models" (from another class) by which you can choose different distributions for different variable types

- Don't worry, the key word is "**plausible**": Truly continuous and interval variables are rare, but there are many variations we often pretend are "continuous and interval enough"
  - These I like to call "**continu-ish**" variables…

Examples of Continu-ish Variables

- **Ordinal-treated-as-interval**: Values are really ordinal, but there are enough distinct values that people justify treating them as interval
  - e.g., one item on 1–4 ordinal scale? Most likely treated as ordinal
  - e.g., sum of 10 items? Likely treated as interval and continu-ish (even though there are no non-integer values, and range is 10–40)
  - e.g., mean of 10 items (better if items may be missing)? Likely treated as interval and continu-ish (non-integer values, but range is 1–4)
  - Binomial and count variables are often predicted as continu-ish ☹

- **Interval, but still likely continu-ish** (may be bounded in practice)
  - e.g., response time, heart rate → really is continuous with non-integer values (limited only by measurement precision) but is bounded at 0
  - e.g., latent trait estimates from measurement models (IRT, CFA,SEM) → non-integer values, but may have observed ceiling or floor effects

One Last Type of Variable: Ratio

- A **ratio scale** has a true zero point
  - Examples: length, height, volume, money
- Ratio scales allow references like "twice as long" or "half as much volume" to actually be meaningful
- Ratio scales do not apply to most quantitative variables in the social sciences (which tend to be interval at best)
  - e.g., a score of 50 vs 100 on an IQ test doesn't mean "half as intelligent" in the same way as a ratio scale
- For all intents and purposes, variables with ratio scaling can be treated as just another quantitative variable

Univariate Description by Variable Type

- For now, we focus on the possible values of each variable, and thus by what salient features we should describe it
  - **Two main types of variables: categorical or quantitative**
    - Distinctions among categorical predictors will **always** matter!
    - Distinctions among quantitative variables matter more when the variable is treated as a model outcome than when treated as a model predictor
      - How would you know which it is? It depends on your question (stay tuned)

- **Categorical** (numbers are labels): Binary, Ordinal, or Nominal
  - Just need to know **frequency** of each category
  - Often more understandable as **proportion**: frequency divided by total possible (proportion*100 becomes a "**percentage**")
  - Can be displayed graphically using a **bar graph**
  - **Value labels** make this information easier to digest or present

Nominal Variable for Marital Status: Description using **SAS** or **STATA**

In **SAS**, using **PROC FREQ**:

```
PROC FREQ DATA=work.Example1;
TABLE marital / MISSING;
RUN;
```

MISSING includes any missing values in table

marital: 5-Category Marital Status				
marital	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.Married	337	**45.91**	337	45.91
2.Widowed	17	2.32	354	48.23
3.Divorced	118	16.08	472	64.31
4.Separated	23	3.13	495	67.44
5.Never	239	32.56	734	100.00

In **STATA**, using **tabulate**:

```
tabulate marital, missing

 marital: 5-Category |
      Marital Status |      Freq.     Percent        Cum.
---------------------+-----------------------------------
           1.Married |        337       45.91       45.91
           2.Widowed |         17        2.32       48.23
          3.Divorced |        118       16.08       64.31
         4.Separated |         23        3.13       67.44
             5.Never |        239       32.56      100.00
---------------------+-----------------------------------
               Total |        734      100.00
```

Note: for your **HW 1**, these **percentages will need to be entered as proportions out of 1**. For instance, 45.59% should be entered as 0.4559 instead of 45.59.

Nominal Variable for Marital Status: Description using **R**

Frequencies in R, using table:

```
table(x=Example1$maritalLabeled,useNA="ifany")

  1.Married    2.Widowed   3.Divorced 4.Separated      5.Never
        337           17          118          23          239
```

Example1 = dataset
maritalLabeled = variable

`useNA="ifany"`
includes any missing values in table

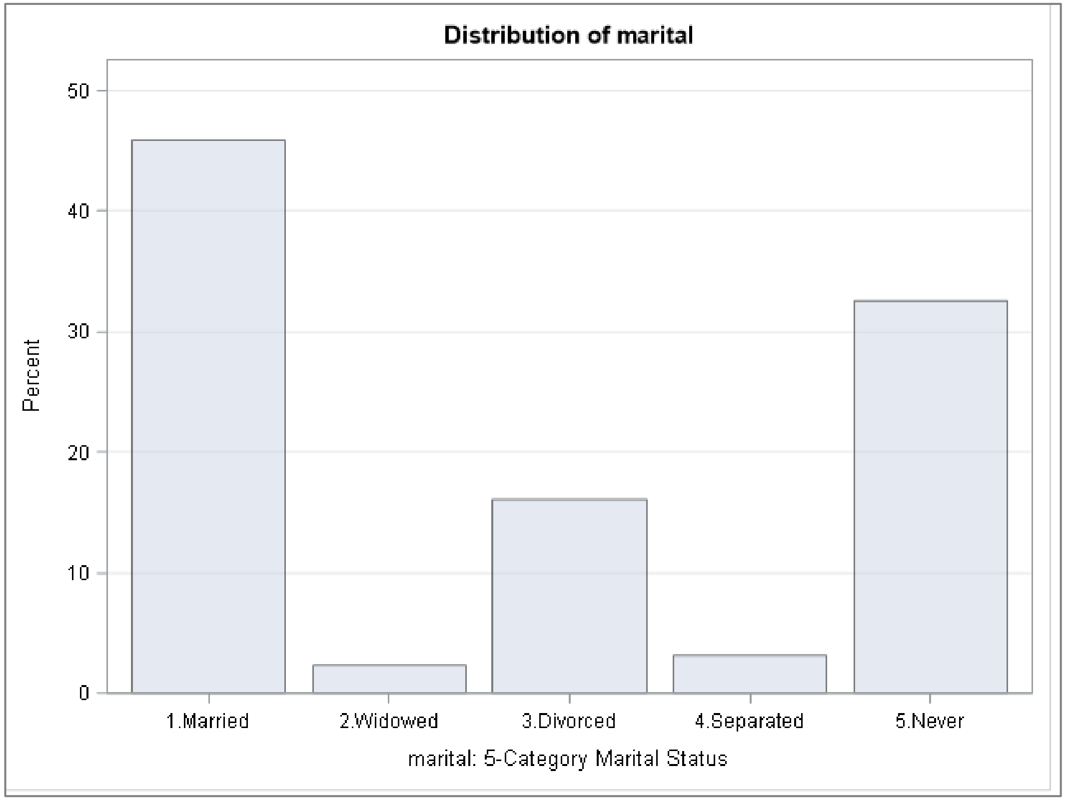
Proportions in R, using prop.table+table:

```
Prop.table(table(x=Example1$maritalLabeled,useNA="ifany"))

 1.Married    2.Widowed   3.Divorced 4.Separated      5.Never
     0.4591       0.0232       0.1608      0.0313       0.3256
```

Nominal Variable for Marital Status: Request a Bar Graph using **SAS**

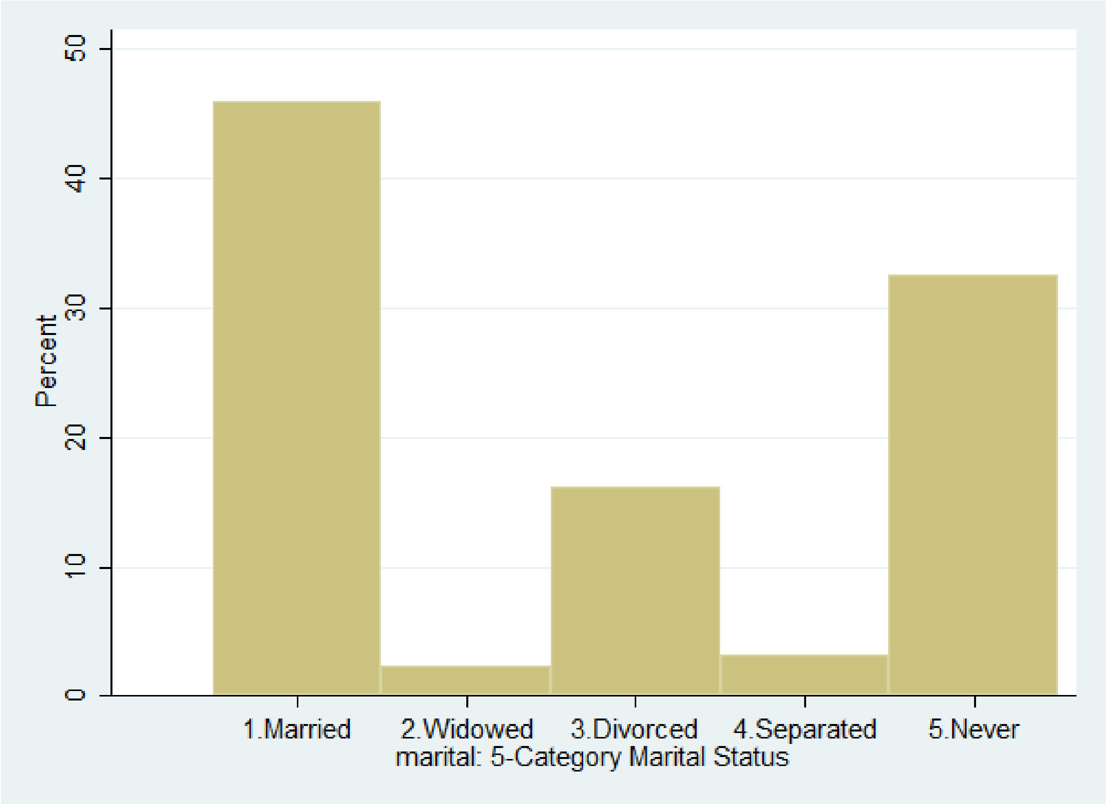
```
PROC FREQ DATA=work.Example1;
TABLE marital / PLOTS=FREQPLOT(TYPE=BAR SCALE=PERCENT);
RUN;
```

- x-axis (horizontal) shows each (labeled) observed category
- y-axis (vertical) shows percentage for each category
- Btw, further customization is available using PROC GPLOT instead
  - VBAR or HBAR



Nominal Variable for Marital Status: Request a Bar Graph using **STATA**

```
// histogram makes plots for categorical or quantitative variables
// marital: is discrete, in percent, x-axis goes 1 to 5 using value labels
   histogram marital, discrete percent xla(1/5, valuelabel)
```

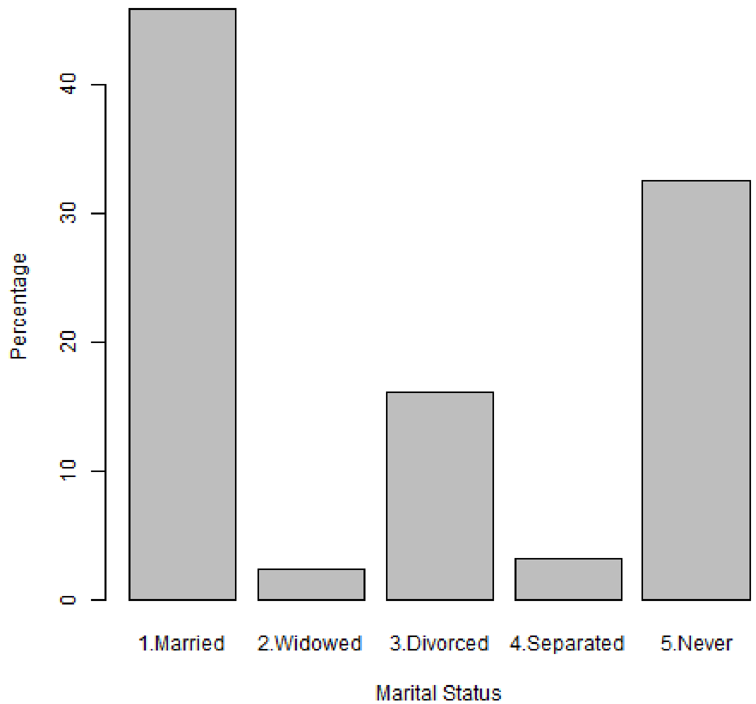


Further customization is available in the window this graph pops up in

Nominal Variable for Marital Status: Request a Bar Graph using **R**

```
# barplot can generate frequency plots for numeric variables
# here is a work-around to make it use our string maritalLabeled variable
barplot(height=table(x=Example1$maritalLabeled,useNA="ifany"),
        ylab="Frequency",xlab="Marital Status") # axis labels
```

```
# here is how to trick barplot
# into plotting percentages instead
barplot(height=prop.table
          (table(x=Example1$maritalLabeled,
           useNA="ifany"))*100,
        ylab="Percentage",
        xlab="Marital Status") # axis labels
```



Further customization is available in many places that I haven't tried to figure out yet, like ggplot…

What about Quantitative Variables?

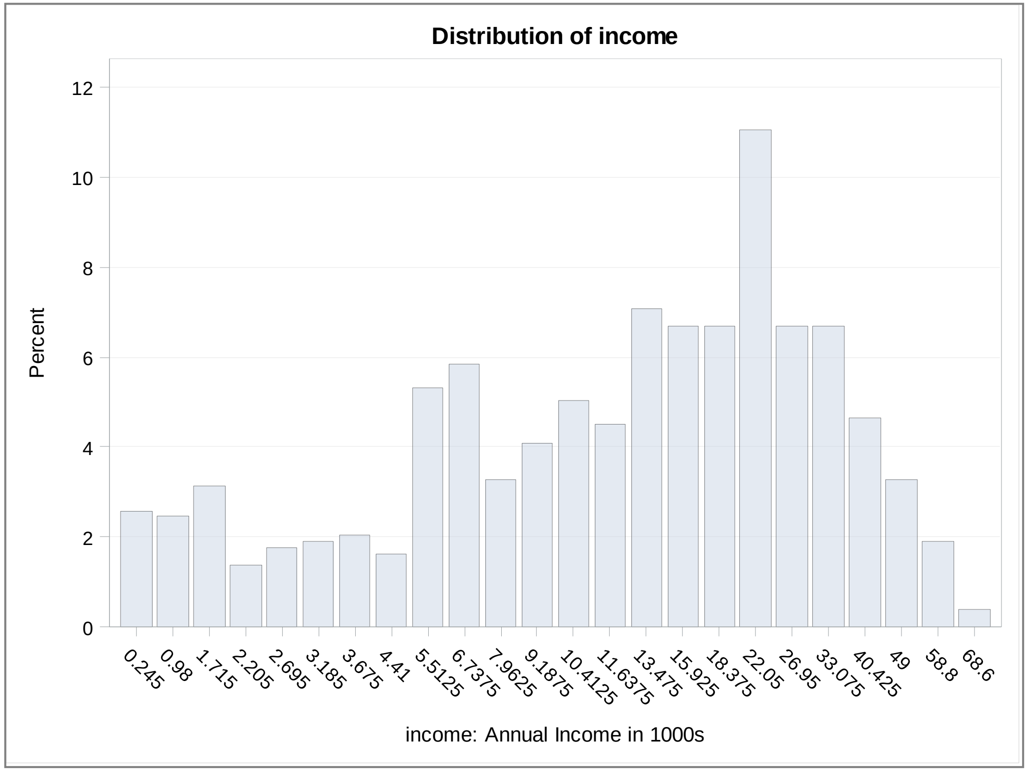
- **Quantitative variable**: numbers are numbers! (interval measurement)
  - May be bounded (binomial, count) or "continu-ish"

- For quantitative variables with **many observed values**, a frequency list of each distinct value is less useful (because interval is ignored)
  - For instance, consider annual income in $1000s (clearly from multiple choices, so it's "continu-ish" here):

income: Annual Income in 1000s				
income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0.245	19	2.59	19	2.59
0.98	18	2.45	37	5.04
1.715	23	3.13	60	8.17
2.205	10	1.36	70	9.54
2.695	13	1.77	83	11.31
3.185	14	1.91	97	13.22
3.675	15	2.04	112	15.26
4.41	12	1.63	124	16.89
5.5125	39	5.31	163	22.21
6.7375	43	5.86	206	28.07
7.9625	24	3.27	230	31.34
9.1875	30	4.09	260	35.42
10.4125	37	5.04	297	40.46
11.6375	33	4.50	330	44.96
13.475	52	7.08	382	52.04
15.925	49	6.68	431	58.72
18.375	49	6.68	480	65.40
22.05	81	11.04	561	76.43
26.95	49	6.68	610	83.11
33.075	49	6.68	659	89.78
40.425	34	4.63	693	94.41
49	24	3.27	717	97.68
58.8	14	1.91	731	99.59
68.6	3	0.41	734	100.00

What about Quantitative Variables?

- Bar graph: also not helpful...



Values are being treated as distinct categories without regard to the intervals between them...

income: Annual Income in 1000s				
income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0.245	19	2.59	19	2.59
0.98	18	2.45	37	5.04
1.715	23	3.13	60	8.17
2.205	10	1.36	70	9.54
2.695	13	1.77	83	11.31
3.185	14	1.91	97	13.22
3.675	15	2.04	112	15.26
4.41	12	1.63	124	16.89
5.5125	39	5.31	163	22.21
6.7375	43	5.86	206	28.07
7.9625	24	3.27	230	31.34
9.1875	30	4.09	260	35.42
10.4125	37	5.04	297	40.46
11.6375	33	4.50	330	44.96
13.475	52	7.08	382	52.04
15.925	49	6.68	431	58.72
18.375	49	6.68	480	65.40
22.05	81	11.04	561	76.43
26.95	49	6.68	610	83.11
33.075	49	6.68	659	89.78
40.425	34	4.63	693	94.41
49	24	3.27	717	97.68
58.8	14	1.91	731	99.59
68.6	3	0.41	734	100.00

What about Quantitative Variables?

- **Instead, we need a histogram**, which combines observations on the x-axis into “bins” (that you can and should choose)
  - Because different programs will bin differently, changing what it looks like…

- In **SAS**:

```
* VAR means variable, midpoints= start to end by increment;
PROC UNIVARIATE DATA=work.Example1;
VAR income;
HISTOGRAM income / MIDPOINTS=0 TO 70 BY 5;
RUN;
```

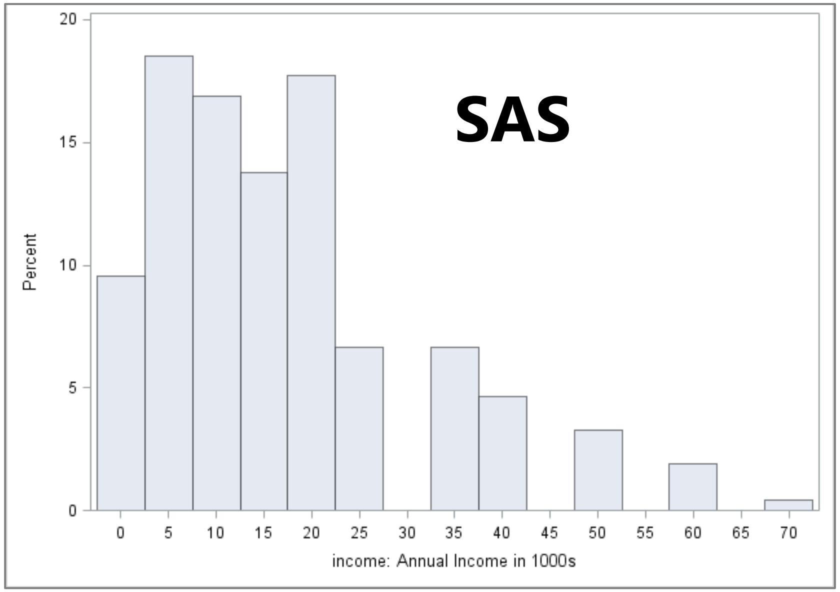
- In **STATA**:

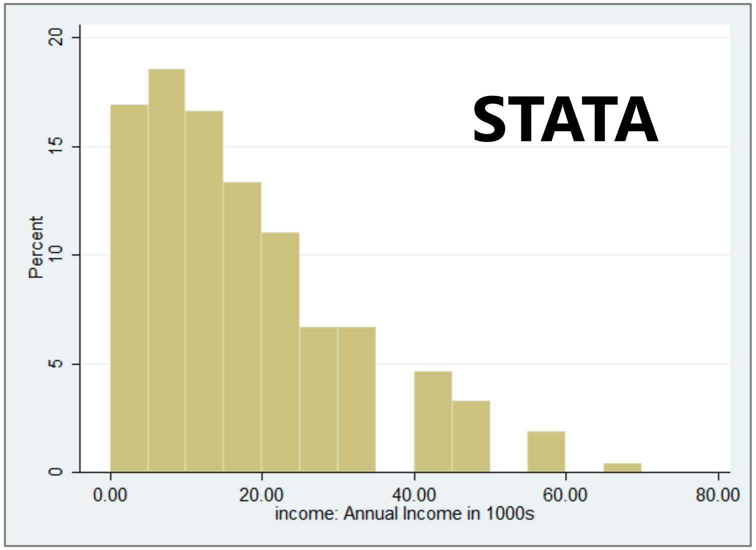
```
// histogram for income in percent in bins of 5 starting at 0
histogram income, percent width(5) start(0)
```

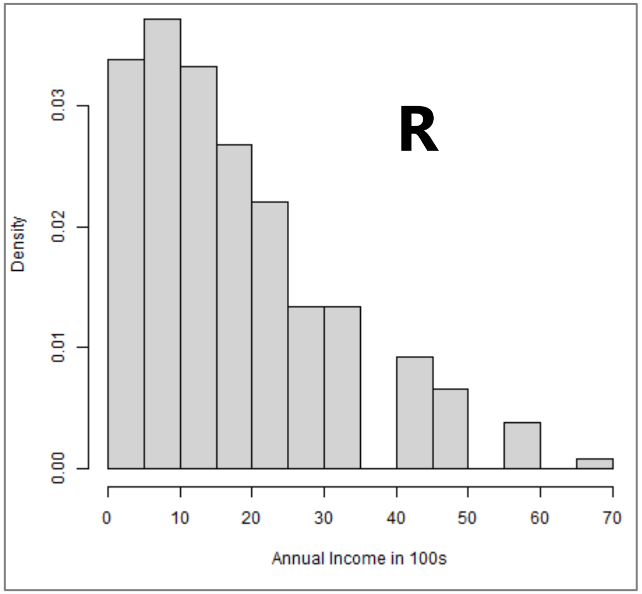
- In **R**:

```
# histogram for income in percent with 15 bins
hist(x=Example1$income, freq=FALSE, breaks=15,
     ylab="Density",xlab="Annual Income in 100s") # axis labels
```

Histogram for Quantitative Variable Income







Despite my trying to make them as equivalent as possible, the SAS version looks different than the STATA and R versions (that have a more similar shape, despite the difference in y-axis)

Big picture: There are more people who make less money than who make more money…

Summarizing Quantitative Variables: 3 Salient Features

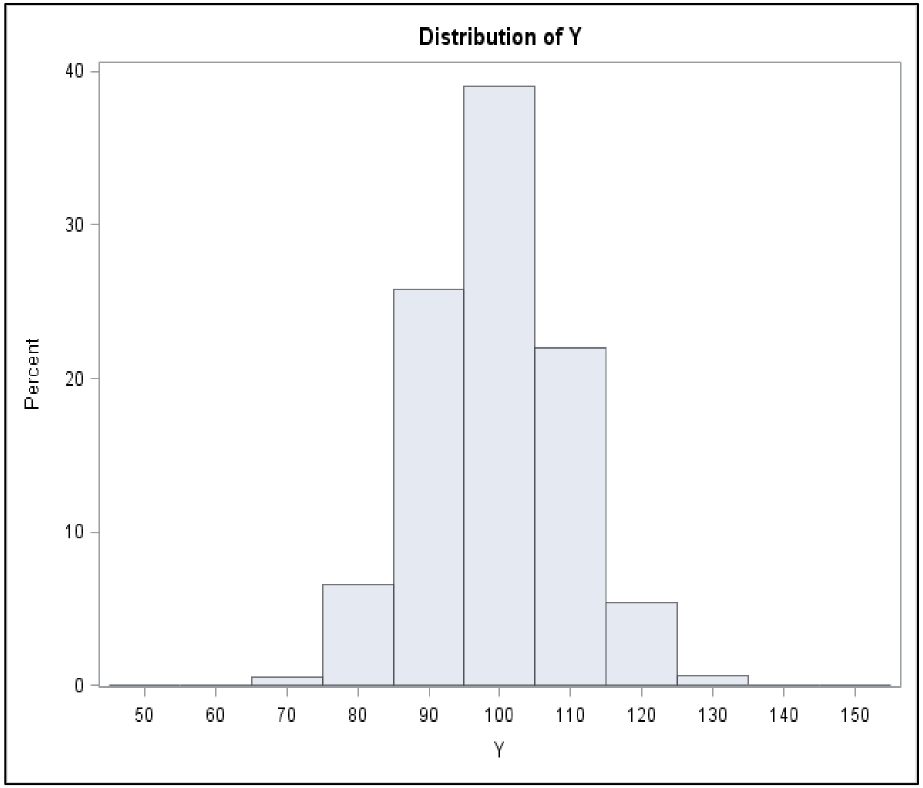
1. **Central tendency:** think "middle of distribution"; can be given by:
   - <u>Mean</u> = arithmetic average (abbreviated "M" in results)
   - Also by <u>Median</u> = middle value if ordered from most to least
   - Also by <u>Mode</u> = most frequent value; rarely mentioned in practice

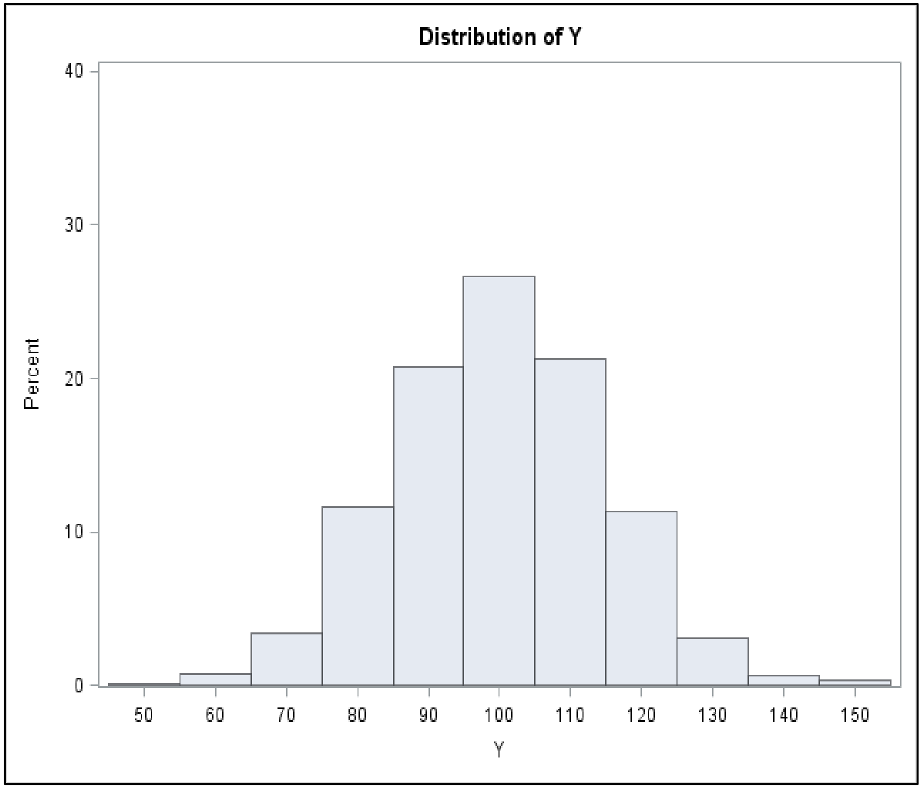
2. **Dispersion:** think "width of distribution", can be given by:
   - <u>Standard Deviation</u> (abbreviated "SD" in results) = average deviation of any given observation (e.g., person) from the mean
   - <u>Variance</u> (abbreviated "VAR" in results) = *squared* average deviation of any given observation (e.g., person) from the mean (so $VAR = SD^2$)
   - Also by <u>Inter-Quartile Range</u> = distance from 25th to 75th percentile

3. **Skewness:** think "asymmetry" (more values on one side than the other)
   - Is often caused by **natural boundaries** in practice (e.g., counts at 0)
   - Is something to factor into your analysis, but is not usually reported

Illustrating Differences in Dispersion (Mean = 100 in both histograms)

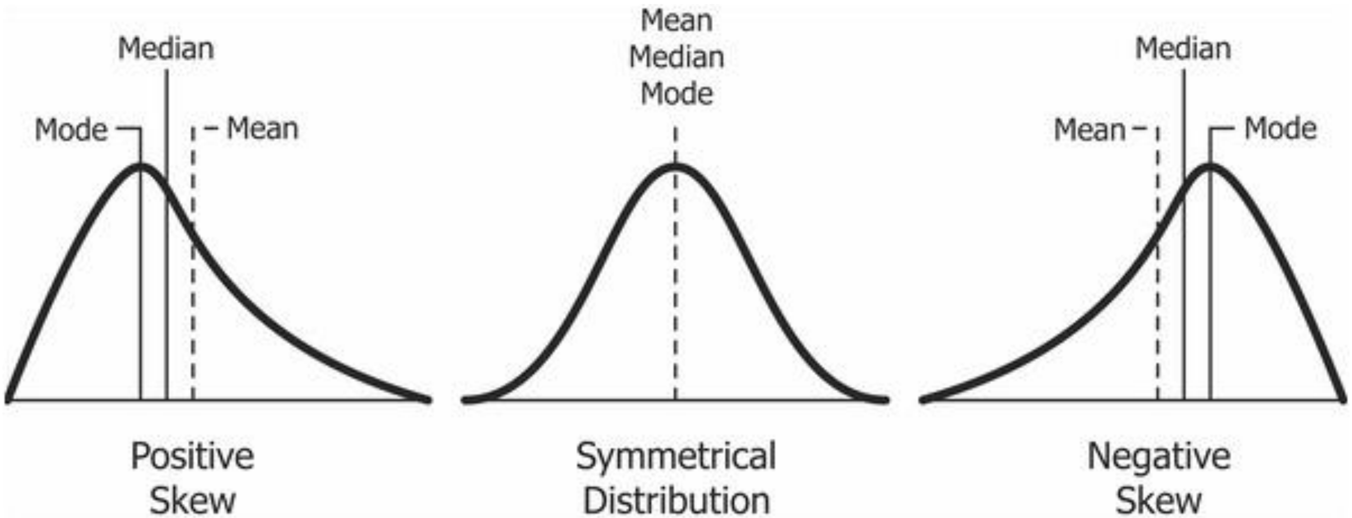
Standard Deviation (SD) = **10**, Variance (VAR) = SD*SD = 100	Standard Deviation (SD) = **15**, Variance (VAR) = SD*SD = 225





Salient Feature #3 of Quantitative Variables: Skewness (Asymmetry)

- **Skewness** can be positive, 0(=symmetric), or negative; **skewness is named for where the tail is headed!**



Note: Mean, median, and mode will diverge in asymmetric variables, so which is reported as an index of central tendency then matters!

Skewness is often caused by natural boundaries (e.g., count variables are often positively skewed). Positive skewness can also result from "floor effects" (and negative skewness can result from "ceiling effects) in binomial-type variables (or both, which is "bimodal").

Image borrowed from: https://www.kullabs.com/classes/subjects/units/lessons/notes/note-detail/9958

Caveats: Population vs. Sample Notation

- Numeric characteristics of the population are called "**parameters**"
  - You almost NEVER know these unless you make up ("simulate") the data ☺
- Numeric characteristics of a specific sample are called "**statistics**"
  - Thus, results sections typically report "descriptive statistics" by that name

- In intro classes, a big deal is usually made about the difference between population and sample notation for univariate (and bivariate) statistics
  - **Population** notation usually uses **Greek** letters (e.g., pandemic alphabet)
  - **Sample** notation usually uses **Roman** letters (e.g., English alphabet)
  - This distinction in notation is important to maintain in SOME contexts, such as when describing the results of simulation studies (i.e., research examining the uses of quantitative methods, where the goal is to see how accurately a given technique returns the known population values)
  - This distinction in notation starts to fall apart in describing the analysis model estimated and its results, in which a mixture of notation is more common (because people understand that you only have a sample)
  - I present both in what follows to link to what you've likely seen before…

Calculating the Arithmetic* Mean of Quantitative (or Binary) Variables

- Sample notation:
  - $\boldsymbol{y_i}$= "y sub i" = outcome y for person i
  - $\boldsymbol{N}$ = "big N" = number of persons in the sample
  - $\boldsymbol{y_N}$= "y sub N" = last person in the sample
  - $\boldsymbol{\bar{y}}$ = "y bar" = sample arithmetic* mean
    - Note the lack of an i subscript—this is because $\bar{y}$ is a constant, not a variable

- How to calculate a **sample mean** (abbreviated $\boldsymbol{M}$ in results):

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{\sum_{i=1}^{N} y_i}{N}$$

→ "Start at $i = 1$, sum over all the y values up to N, then divide that total by N"

- **Sample mean $\boldsymbol{\bar{y}}$** ("y bar") is an estimator of **population mean $\boldsymbol{\mu}$** ("mu")

** Yes, there are other kinds of means (geometric, harmonic, weighted)…*

Calculating the Variance (Dispersion) of Quantitative Variables

- Notation to calculate **variance** (abbreviated VAR in results):

$$Variance = s^2 = \frac{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}{N-1}$$

→ "Start at $i = 1$, subtract $\bar{y}$ from each y value, square that result, sum until N, then divide by $N - 1$"

  - **Sample variance $\boldsymbol{s^2}$** is an estimator of **population variance $\boldsymbol{\sigma_e^2}$** ("sigma squared")

- Squaring maintains absolute magnitudes, but because squared units are less interpretable than raw-data units, the **standard deviation** (SD, the square root of variance) can be more intuitive: **$\boldsymbol{SD}$ is the average distance for any given unit from the mean** (e.g., SD describes a variable's dispersion across persons)

- Btw, in the denominator for variance, $\boldsymbol{N-1}$ is used instead of N to adjust for needing the sample mean in order to calculate the sample variance; later on this term will be called "**denominator degrees of freedom (DF)**"

Calculating the Skewness of Quantitative Variables (Asymmetry)

- **Skewness** is calculated with the same pattern, but cubed (without common special notation, btw):

$$\text{Skewness} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \bar{y}}{s}\right)^3$$

→ Skewness will be 0 if the variable is symmetric

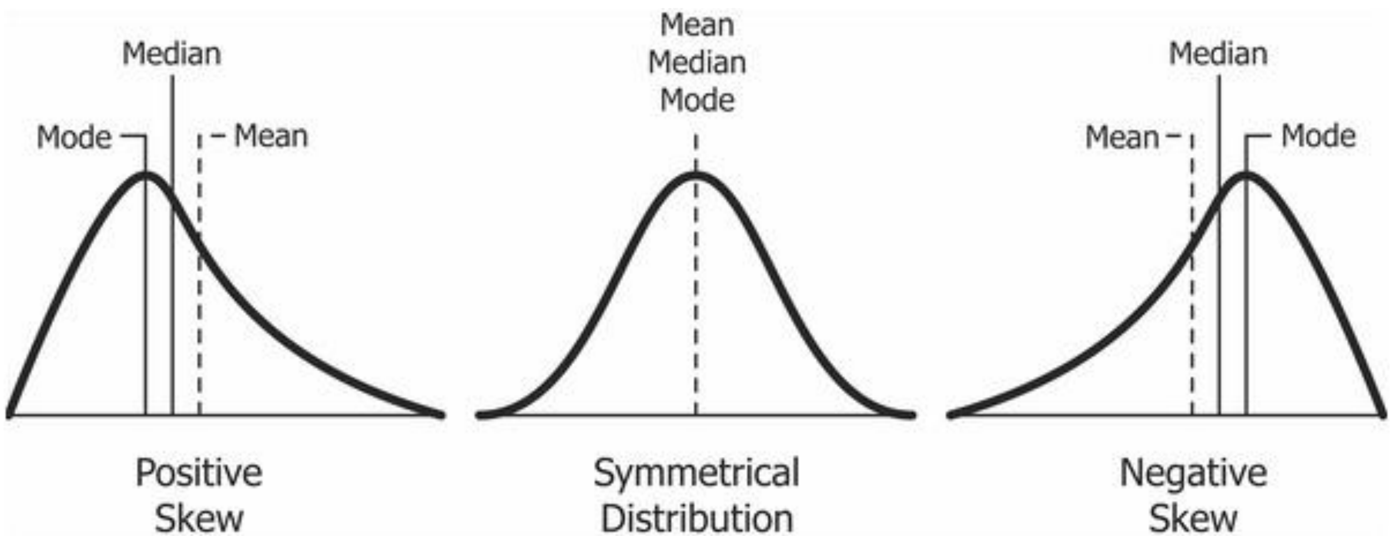
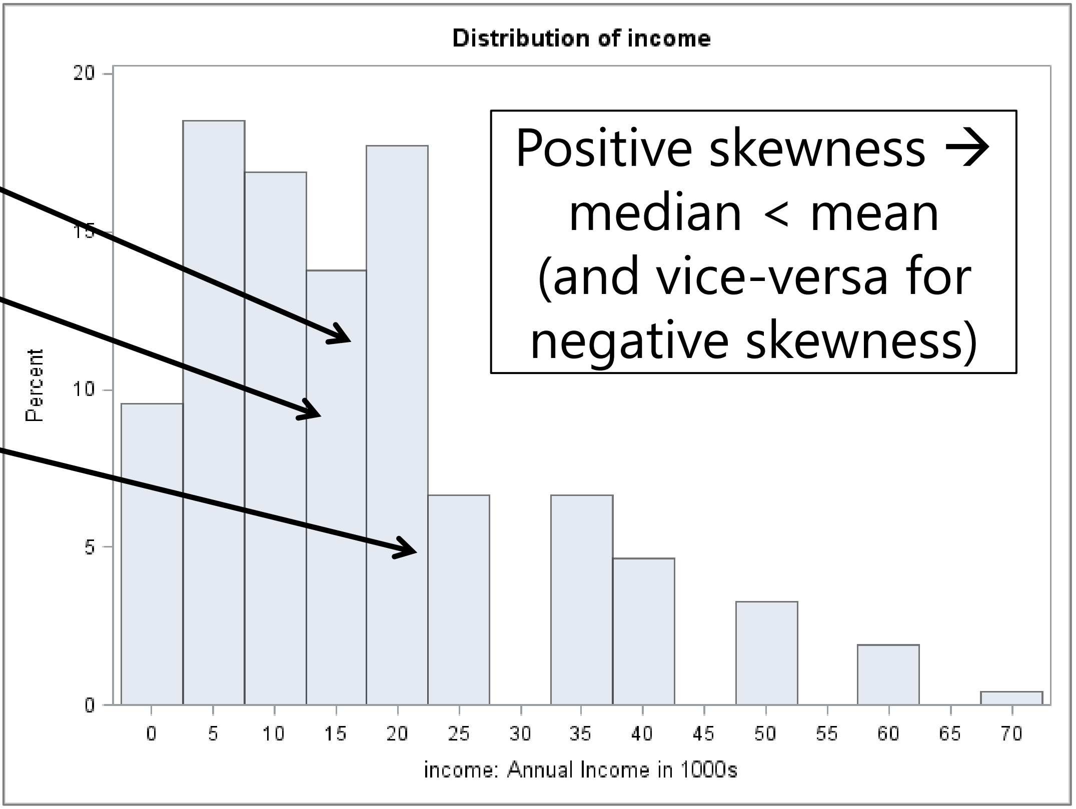


Image borrowed from: https://www.kullabs.com/classes/subjects/units/lessons/notes/note-detail/9958

Example: Skewness = 1.16 in Income

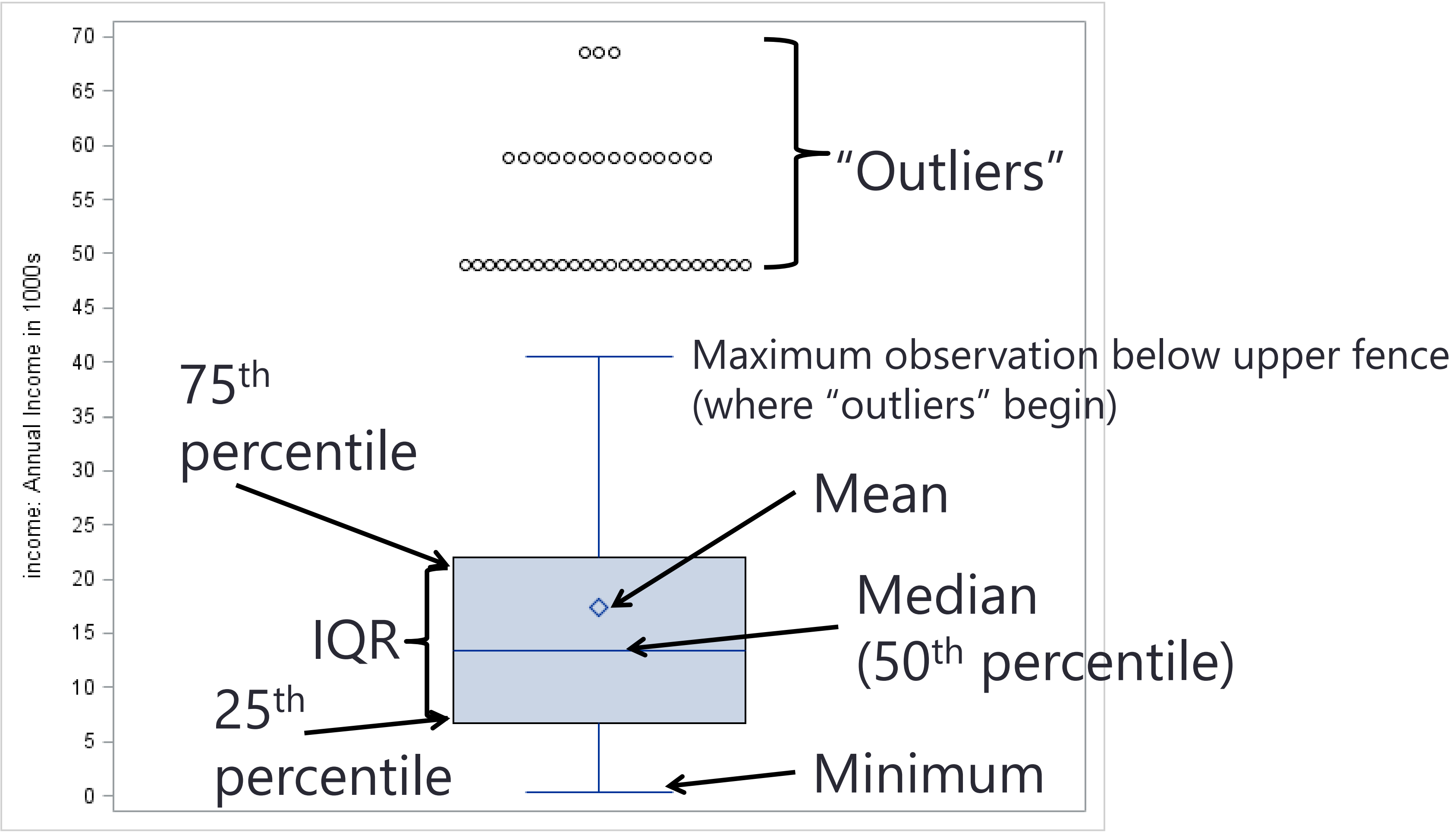
- **Central tendency:**
  - Mean (M) = 17.31
  - Median = 13.48
    - Btw, = 50th percentile
  - Mode = 22.05
- **Dispersion:**
  - $VAR = SD^2$ = 190.21
  - SD = 13.79
  - Inter-quartile range:
    - IQR = 75th − 25th percentiles
    - IQR = 22.05 − 6.74 = 15.31



Positive skewness → median < mean (and vice-versa for negative skewness)

Should also report the **range**: the **minimum** and **maximum** values (0.245 and 68.60 here)

Summarizing Skewed Quantitative Variables using a "**Box Plot**"

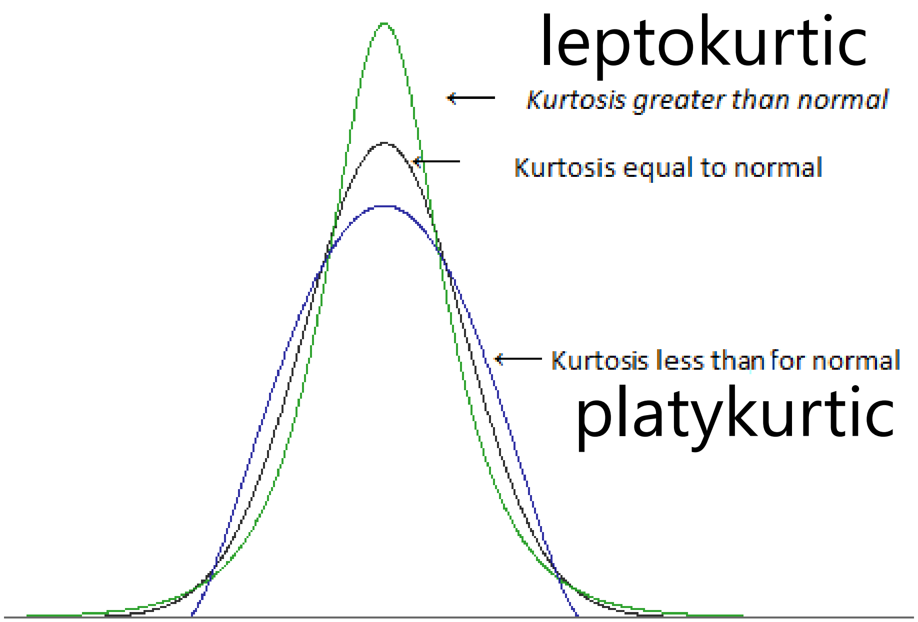


Btw, One More Feature of Quantitative Variables: Kurtosis

- **Kurtosis** is calculated with the same pattern, but fourth-d:

$$\text{Kurtosis} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \bar{y}}{s}\right)^4 - 3$$

→ Kurtosis will also be 0 if the variable is symmetric



Note: Extent of kurtosis is hard to differentiate from variance in real data, so don't worry about this one

Image borrowed from: https://stats.stackexchange.com/a/143522/124771

Describing Quantitative Variables: **SAS**

```
* PROC UNIVARIATE prints ALL descriptive statistics;
PROC UNIVARIATE DATA=work.Example1;
VAR income; * VAR = variables;
RUN;
```

Moments			
N	**734**	Sum Weights	734
Mean	**17.3028747**	Sum Observations	12700.31
Std Deviation	**13.7916296**	Variance	**190.209048**
Skewness	**1.16073362**	Kurtosis	1.10205445
Uncorrected SS	359175.104	Corrected SS	139423.232
Coeff Variation	79.7071579	Std Error Mean	0.50905834

Coefficient of variation $= SD/M$

Basic Statistical Measures			
Location		Variability	
Mean	17.30287	Std Deviation	13.79163
Median	13.47500	Variance	190.20905
Mode	22.05000	Range	68.35500
		Interquartile Range	15.31250

Quantiles (Definition 5)	
Level	Quantile
100% Max	68.6000
99%	58.8000
95%	49.0000
90%	40.4250
75% Q3	22.0500
50% Median	13.4750
25% Q1	6.7375
10%	2.6950
5%	0.9800
1%	0.2450
0% Min	0.2450

```
* PROC MEANS prints most common descriptive statistics;
PROC MEANS DATA=work.Example1 NDEC=3 N MEAN STDDEV VAR MIN MAX;
VAR income; * VAR = variables;
RUN;
```

Variable	Label	N	Mean	Std Dev	Variance	Minimum	Maximum
income	income: Annual Income in 1000s	734	17.303	13.792	190.209	0.245	68.600

Describing Quantitative Variables: **STATA**

```
summarize income, detail  // detail gives more info

 income: Annual Income in 1000s
-------------------------------------------------------------
      Percentiles      Smallest
 1%        .245           .245
 5%         .98           .245
10%       2.695           .245       Obs                 734
25%      6.7375           .245       Sum of Wgt.         734

50%      13.475                      Mean           17.30287
                        Largest      Std. Dev.      13.79163
75%       22.05           58.8
90%      40.425           68.6       Variance        190.209
95%          49           68.6       Skewness        1.15836
99%        58.8           68.6       Kurtosis       4.086398
```

```
summarize income

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      income |       734    17.30287    13.79163       .245       68.6
```

Describing Quantitative Variables: R

```
# describe prints sample descriptive statistics
# embedding describe in print allows control of number of digits printed
# quant= requests list of quantiles, IQR requests inter-quartile range
print(describe(x=Example1$income,quant=c(.25,.50,.75),IQR=TRUE),digits=4)

  vars   n    mean      sd median trimmed     mad   min  max  range
1    1 734 17.3029 13.7916 13.475 15.4929 12.7133 0.245 68.6 68.355

   skew kurtosis     se     IQR  Q0.25   Q0.5 Q0.75
1 1.156   1.0753 0.5091 15.3125 6.7375 13.475 22.05
```

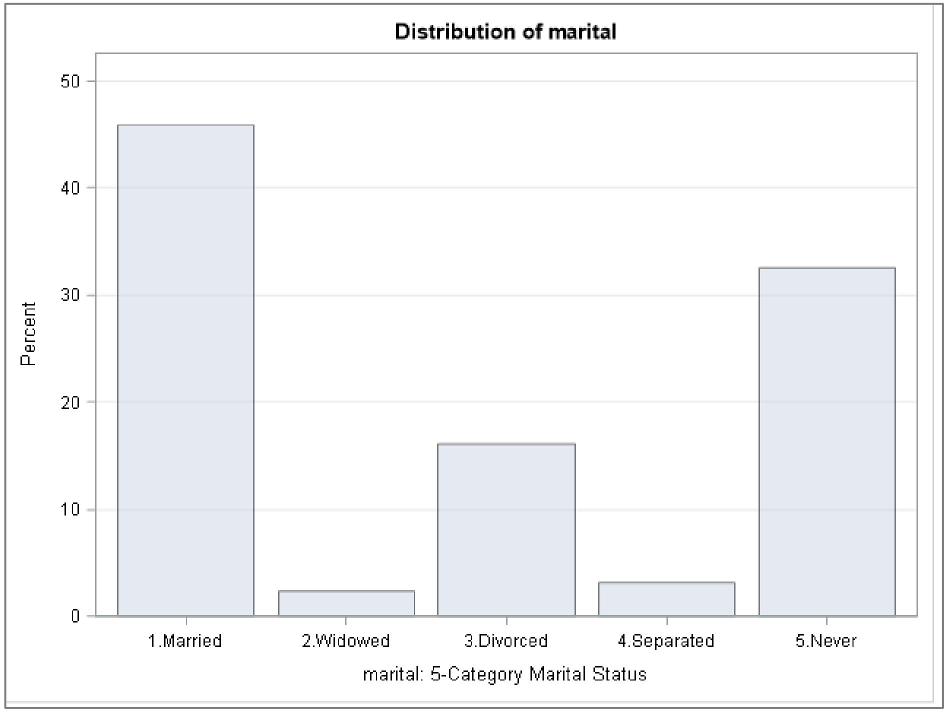
Note what's missing…
Any guesses why?

```
# describe does not include variance, so here is a command to do so
var(Example1$income)

[1] 190.20905

# likewise, here are commands to get the mean and SD by themselves
# can have more than one command on a line if separated by semi-colon
mean(Example1$income); sd(x=Example1$income)

[1] 17.302875
[1] 13.79163
```

Means for Categorical Variables?

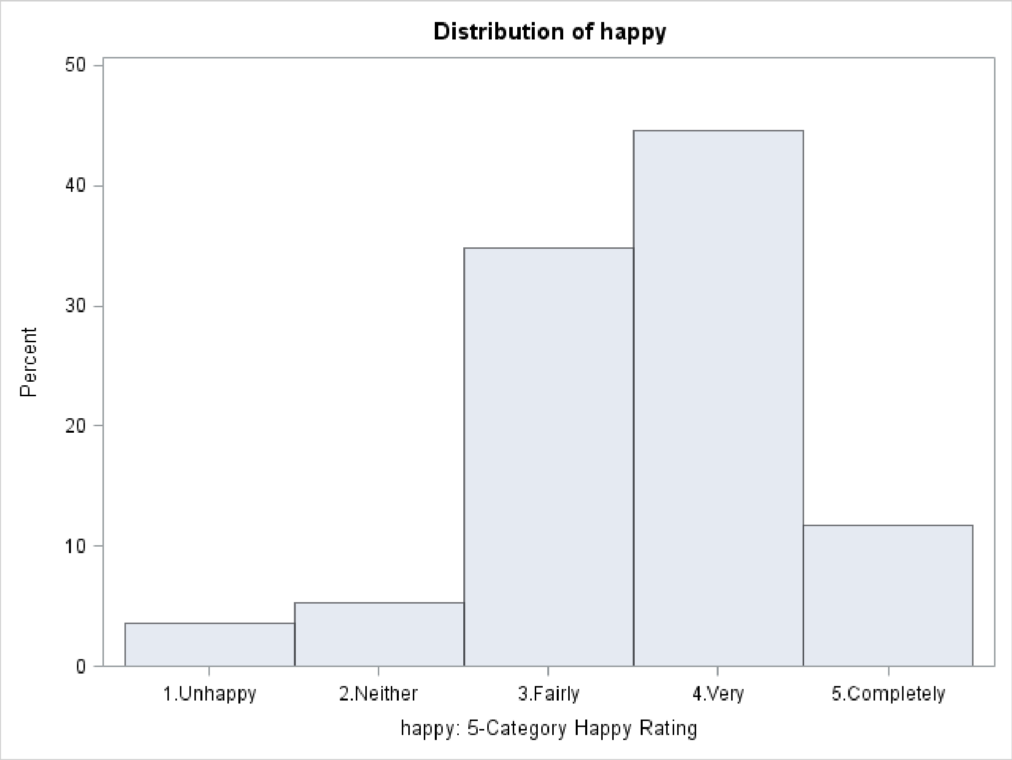
- For **binary variables** coded 0 or 1, **the mean** is calculated the same way but it **is called the** "**proportion**" instead
  - e.g., 0=alive, 1=dead? Mean = "death rate"
  - This is fine because there is only one interval to consider

- For **nominal variables** with 3+ options, **a single mean does not make sense!**
  - e.g., for nominal marital status, *M* = 2.74… ?!?
  - Software will give it to you anyway (user beware) ☺



Means for Ordinal Variables?

- What about **means (or SDs) for ordinal variables**?
  - Should give you pause….



- For example, for **self-rated happiness on a 5-point scale**:
  - Mean (M) = 3.56
    Median and Mode = 4
  - Known as "Likert scale" (Like-ert, not Lick-ert)
- Using a mean (or SD) **assumes equal distances between the options** (interval measurement)
  - *Stay tuned for how to think about ordinal predictors…*

Variances for Categorical Variables?

- For **binary variables coded 0 or 1**, variance and skewness are not separate properties (as they are in quantitative variables)
  - If $\boldsymbol{p}$ = proportion of 1 values, and $\boldsymbol{q}$ = proportion of 0 values:
  - Mean $\overline{\boldsymbol{y}} = \boldsymbol{p}$, variance $\boldsymbol{s^2} = \boldsymbol{p} * \boldsymbol{q}$, and skewness = $\frac{\mathbf{1-2}\boldsymbol{p}}{\sqrt{\boldsymbol{p*q}}}$

Mean and Variance of a Binary Variable

Mean ($\boldsymbol{p}$)	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

- For variables with >2 categories, **each pair of categories** would have its **own $\boldsymbol{p}$ and $\boldsymbol{q}$** (and thus variance/skewness)
  - So the **percent for each category is enough to report** (i.e., the pairwise variance and skewness values are not helpful)

Wrapping Up

- What kind of **univariate descriptive statistics** are relevant to report depends on the type of variable to be described:
  - **Quantitative variables (numbers are numbers):**
    - If “symmetric enough”: Min, Max, Mean, SD (or SD^2 = variance)
    - If not, add median (for central tendency) and IQR (for dispersion) that are “robust” to outliers (extreme values) or general skewness
    - Binned-value histograms, boxplots, or violin plots make good visuals
  - **Categorical variables (numbers are just labels):**
    - **Binary** (0 or 1): Mean (= proportion of 1 values); variance and skewness are then determined by the mean (i.e., they are redundant)
    - **Nominal** with **3+ categories**: **percentage** of each category; a single mean (or variance or skewness) makes no kind of sense
    - **Ordinal** with **3+ categories** may be treated as quantitative, but keep in mind this assumes equal distances between the numbers used as labels
    - Bar graphs of the percentage in each category make a good visual

Best Practices for Working with Datasets and Variables in Statistical Software

- Quantitative data can be stored in a variety of formats
- We will use data stored in excel (with **.xlsx** extension) because it is viewable outside of specific statistical software, but it can easily be imported into any package

- <u>3 steps to import .xlsx data into either stats program:</u>
  1. Save dataset to a folder and get the address to that folder
  2. Copy the folder address into the program syntax
  3. Run (execute) the syntax to import the .xlsx data into the program's native format (SAS, STATA, R) for use in analysis

- Historically this has been the hardest step, so I have made new videos using Example 1 to walk you through the process…

Best Practices for Working with Datasets and Variables in Statistical Software

At least 3 useful pieces of information will be stored for each variable (see demo in videos describing use of software for example 1):

1. **Variable name** = column name (required)
   - No spaces or special characters, must start with letter
   - To be referred to when requesting info or results about that column

2. **Variable label** = column description (optional; not used in R)
   - Longer text label that can document the variable in more detail
   - e.g., how it was created, # categories, which version or metric

3. **Value label** = verbal labels that go with the numbers (optional)
   - Used for categorical variables only (in which numbers are labels)
   - Makes results easier to read (i.e., don’t have to remember values)
   - Can make a text-only (string) variable as a workaround (I did in R)

How to Store Variables in Databases

- When entering data, there are things you can do up front to save yourself a lot of tedious hassle later:
  - Btw, it's fine—preferable—to use spreadsheets (e.g., excel) to enter the data, no matter how you plan to analyze it
    - But keep in mind that "meaningful" formatting will not transfer (e.g., coloring cells yellow will mean nothing in SAS, STATA, or R)
- Put **variable names** in the first row of the spreadsheet
  - Do not use spaces or special characters other than _
  - Use only as many characters as necessary to keep it unique
    - Use variable labels to add extra detail for clarification
  - Start with a letter, not a number (is rule in stats programs)
  - Use a common stem for a series of related variables
    - e.g., stress1, stress2, stress3…. wellbeing1, wellbeing2, wellbeing3…
    - This is helpful when you need to refer to them as a series

How to Store Variables in Databases

- Enter **numbers for categorical variables**, not text
  - Text variable = string variable = case- and space-sensitive
    - e.g., “control group” is not the same as “Control Group ”
  - Add **value labels** to indicate what the numbers represent
    - It can be helpful to use the number in the value label so that the order of the labels is the same alphabetically and numerically
      - e.g., group: 0 = “0. Control Group” 1 = “1. Alternative Group”
  - Do not mix numeric and text entries in the same variable
    - Numbers will be read as text → becomes a string variable instead
  - IMHO: Do not use missing data codes (e.g., −99 = missing)
    - You must define them as such for -99 to NOT be read as data
    - Just leave them missing values blank—if you need to keep track of reasons for missing values, use a new categorical variable to do so

How to Store Variables in Databases

- Tips for handling entry of dependent data more easily
  - Create a unique ID variable for each level of sampling
  - Create separate databases for each level sampling—you can easily merge them together so that the values of the higher-level variables are replicated automatically across the rows of the lower-level database (as is needed)

- For example: people collected from different countries?
  - Person-level database: one row per person; include person ID, country ID, and all person-level variables
  - *Separate* country-level database: one row per country; include country ID and all country-level variables (when merged, will replicate across people)

- For example: multiple occasions from same person?
  - Occasion-level database: one row per occasion; include occasion ID, person ID, and all variables measured per occasion
  - *Separate* person-level database: one row per person; include person ID and all person-level variables (when merged, will replicate across occasions)