Univariate Data Description: One Variable at a Time

- Topics:
  - Best practices for working with datasets and syntax
  - 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
  - **Note: there is no separate example document for this unit; a video will demonstrate how to work with the R file instead

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 in **.xlsx format** because it is easily viewable outside stats software and imports readily (even more so as .csv instead)

- **3 steps** to import external data (of any kind) into any stats program:
  - **Save** dataset to a folder and get the address to that folder
  - Copy the **folder address** into the right spot in the program syntax
  - Run **syntax to import** data into program's native format for analysis

How to Store Variables in Databases

- When entering data, do the following to save yourself later hassle:
  - Btw, it's fine—preferable, even!—to **use spreadsheets** (e.g., excel, google sheets) to enter 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 stats software)

- Put **variable names in the first row only** of the spreadsheet
  - **Do not use spaces or special characters** other than _ underscore (or . dot in R)
  - Use only as many characters as necessary to keep it unique
    - Use variable labels (as comments only in R code ☹) to add extra detail for clarification
  - **Start with a letter**, not a number (is a rule in most stats programs)
  - Use a **common stem** for a series of related variables
    - e.g., stress1, stress2, stress3…. wellbeing1, wellbeing2, wellbeing3…
    - Helpful when you need to refer to them as a series (e.g., find a mean across them)

How to Store Variables in Databases

- Enter numbers for grouping or ID variables, not text
  - **Text** variable = **string** variable = case- and space-sensitive
    - e.g., because "control group" is not the same as "Control Group "
  - In non-R software, add **value labels** to indicate what the numbers represent
    - Use the number in the value label for same label order alphabetically and numerically
      - e.g., group: 0 = "0. Control Group" 1 = "1. Alternative Group"
  - In R software, convert from number to text ("string") for value labels instead
  - **Do not mix numeric and text entries in the same variable**
    - Numbers will be read as text → becomes a "string" variable instead
  - **Do not bother with missing data codes** (e.g., −99 = missing)
    - You must define them as such to NOT be read as data in many programs
    - Just leave them missing values blank—make a new variable for differential missingness

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 of 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 (which is needed to use them)

- For example: People collected from different countries?
  - **Person-level data**: one row per person; include person ID, country ID, and person-level variables
  - **Separate country-level data**: one row per country; include country ID and country-level variables (when merged, country-level variables will replicate across people as needed to use them)

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

Types of Variables

- Goal: identify **potential types** of variables in quantitative data
  - Highest-order distinction: **categorical** or **quantitative**?

- This "**types**" taxonomy will guide **two processes** about each variable:
  - What indices can be used to **summarize** its salient features
  - How it can be used in subsequent **statistical analysis**
  - Note: this is related to traditional levels of measurement, but I am approaching it from more of a "how to use them" perspective

- **Apparent purpose of HW1: 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 of HW1: Use familiar ideas to begin to use unfamiliar R software**

Categorical Variables: Numbers are just labels

- **Binary** = choices of 0 or 1 (any 2 choices = "dichotomous")
  - e.g., dead or alive; pregnant or not

- **Nominal** = 3+ unordered choices
  - e.g., favorite type of pet, present 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

Quantitative Variables: Numbers are amounts

- **Interval** measurement → equal distances between values (that mean something)
- Many quantitative variables have **1 or 2 natural 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 (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 population 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 GLMs our predictors can be any type of variable, **GLM 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 a further class) to choose different distributions for different variable types (like categorical)
- Key word is "plausible": Truly continuous and interval variables are rare, but many variations we can 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 usually justify treating them as interval (it's easier)
  - 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 outcomes ☹

- **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

Welcome to R Syntax!

- Text after # in green are comments (= notes only to your future self)
- R is composed (almost) entirely of **functions**
  - **R = base functions + user-created** "**packages**" (of many specific functions)
  - Search "**CRAN**" with the R package name to find its documentation

- For instance, in the examples that follow, I used **two R packages**:

```
#####  Check to see if packages are downloaded, install if not; then load  #####

# To import excel .xls or .xlsx data as table
if (!require("readxl")) install.packages("readxl"); library(readxl)

# To summarize quantitative data
if (!require("psych")) install.packages("psych"); library(psych)
```

R Syntax: Set Working Directory and Import Data

- First need **`setwd`** function to tell R where to look for your data files

```
# Set working directory (to import and export files to)
# Paste in the folder address where your data file is saved in quotes
# Note the slashes are backwards relative to Windows file paths
setwd("C:/Dropbox/26_EDF9770/Lecture1/")
```

- R can import almost any kind of data, but you may need to find a specific package for other types of data (e.g., SAS, Stata, Access)
  - Use **`read_excel`** function from **`readxl`** package to import an .xlsx (or .xls) file

```
# Import "GSS_Example.xlsx" from sheet "Data" with first row as variable names
Example1 = read_excel(path="GSS_Example.xlsx", sheet="Data", col_names=TRUE)
# Convert to data frame to use for analysis
Example1 = as.data.frame(Example1)
```

R Syntax: Conducting Analyses

- For example: Imagine you were asked how your dinner was…
  and you'd like to answer "It's fine, not too spicy"

```
# Answer question about dinner dataset using answeR package in R
myanswer = answeR(data=dinner, spicy=FALSE, formula~response=fine)
summary(myanswer) # Print of saved result requested separately
```

  - Saved object (that you can then use in further commands) = **`myanswer`**
  - Saved object results do not print (why you need a **`summary`**-type command)
  - Function used = **`answeR;`** name of dataset given by **`data=dinner`**
  - Inside ( ) are **arguments**: option=choice, formula~outcome=predictors

Tips for Effective Syntax (in R and in general)

- Build a ***well-commented* syntax file** for Future You of **WHAT** you did and **WHY**
  - Add spaces and **comment section headers** to create **logical groupings** of commands

Example R Syntax with Comment Section Headers

```
# For EDF 99770 Spring 2026 Lecture 1 Slides

################################################################################
#####                 LECTURE 1 OPTIONS AND PACKAGES                       #####
################################################################################

# Set width of output and number of significant digits printed,
# number of digits before switching to scientific notation
options(width=120, digits=8, scipen=9)

#####  Check to see if packages are downloaded, install if not; then load  #####

# To import excel .xls or .xlsx data as table
if (!require("readxl")) install.packages("readxl"); library(readxl)

# To summarize quantitative data
if (!require("psych")) install.packages("psych"); library(psych)

################################################################################
#####            LECTURE 1 DATA IMPORT AND MANIPULATION                    #####
################################################################################

# Set working directory (to import and export files to)
# Paste in the folder address where your data file is saved in quotes
# Note the slashes are backwards relative to Windows file paths
setwd("C:/Dropbox/26_EDF9770/Lecture1/")

# Import "GSS_Example.xlsx" from sheet "Data" with first row as variable names
Example1 = read_excel(path="GSS_Example.xlsx", sheet="Data", col_names=TRUE)
# Convert to data frame to use for analysis
Example1 = as.data.frame(Example1)
```

Tips for Effective Syntax (in R and in general)

- Build a ***well-commented* syntax file** for Future You of **WHAT** you did and **WHY**
  - Add spaces and comment section headers to create logical groupings of commands

- **My example R syntax files will have the following structure:**
  - General options for **output** display
  - Install and load **packages**; load separate file of my functions (stay tuned)
  - Set "**working directory**" (file location) and **import data**
  - Create any **new variables** needed for data analysis
  - Optional: **Open external text file** to save (or "log") results to
  - Commands to do **analyses**—make a new instance for each!
  - If used: **Close external text file** of saved results

- I will give you "**starter**" **syntax** with this format for each homework
- **HW1 will require univariate descriptive statistics using R… let's review!**

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** variables 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 just labels) = **Binary**, **Ordinal**, or **Nominal**
  - Just need to know frequency of each category
  - Often more understandable as "**proportion**": frequency divided by total possible (proportions range from 0–1; proportion*100 becomes a "percentage")
  - Can be displayed graphically using a bar graph
  - Value labels (as strings in R) make this information easier to digest

Nominal Variable for Marital Status

Frequencies and **proportions** in R, using **`table`**:

data$variable
`Example1` = data
`maritalLabeled` = variable

```
# table prints frequency-only tables for categorical variables
# useNA="ifany" includes missing values too
print("R Frequency Table for Categorical Variable maritalLabeled")
table(x=Example1$maritalLabeled, useNA="ifany")
```

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

```
# prop.table converts tabled frequencies into proportions
print("R Proportion Table for Categorical Variable maritalLabeled")
prop.table(table(x=Example1$maritalLabeled, useNA="ifany"))
```

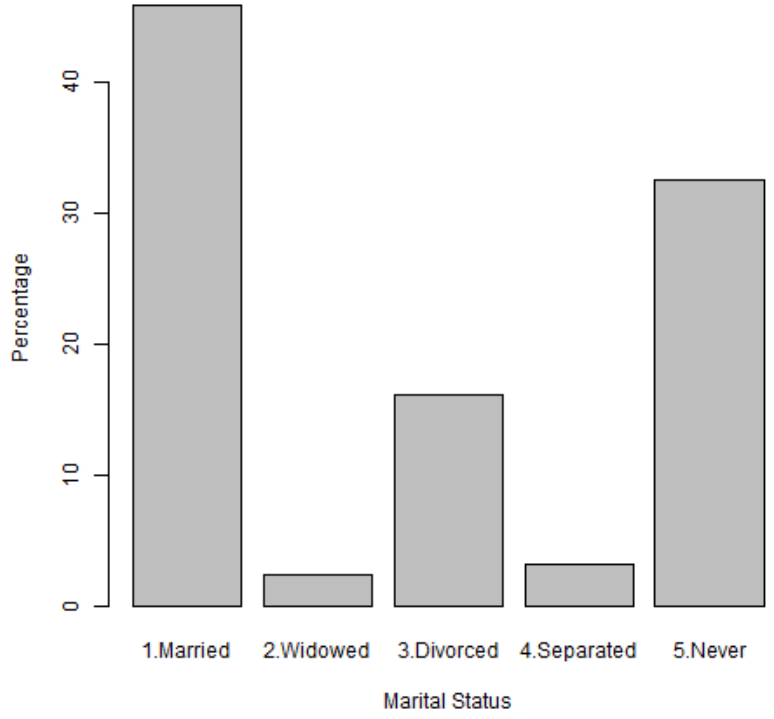
```
   1.Married   2.Widowed  3.Divorced 4.Separated     5.Never
0.459128065 0.023160763 0.160762943 0.031335150 0.325613079
```

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") # y and x axis labels


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

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



What about Quantitative Variables?

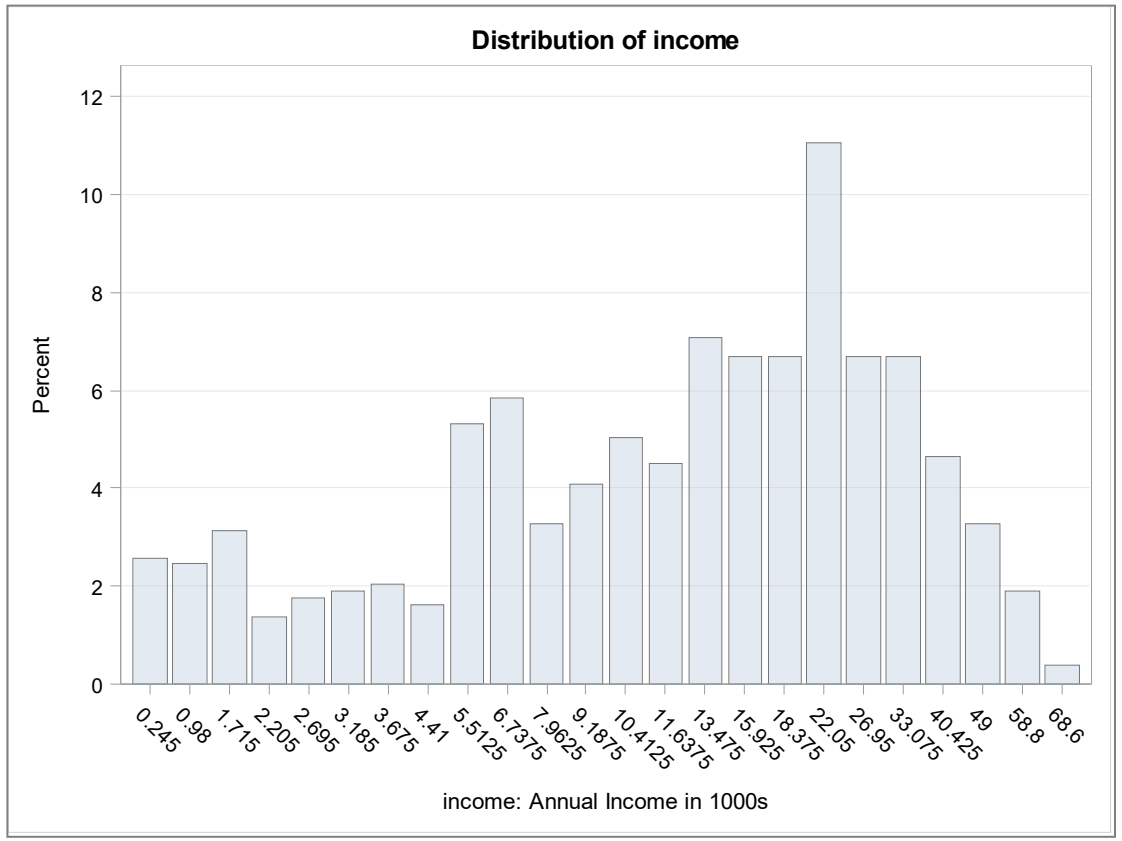
- Quantitative variable: **numbers are numbers**! (interval measurement)
  - May be bounded or "continu-ish"

- 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 actually "continu-ish" here):

income: Annual Income in 1000s	Freq.	Percent	Cum.
0.25	19	2.59	2.59
0.98	18	2.45	5.04
1.72	23	3.13	8.17
2.21	10	1.36	9.54
2.69	13	1.77	11.31
3.19	14	1.91	13.22
3.67	15	2.04	15.26
4.41	12	1.63	16.89
5.51	39	5.31	22.21
6.74	43	5.86	28.07
7.96	24	3.27	31.34
9.19	30	4.09	35.42
10.41	37	5.04	40.46
11.64	33	4.50	44.96
13.48	52	7.08	52.04
15.93	49	6.68	58.72
18.38	49	6.68	65.40
22.05	81	11.04	76.43
26.95	49	6.68	83.11
33.08	49	6.68	89.78
40.42	34	4.63	94.41
49.00	24	3.27	97.68
58.80	14	1.91	99.59
68.60	3	0.41	100.00
Total	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…

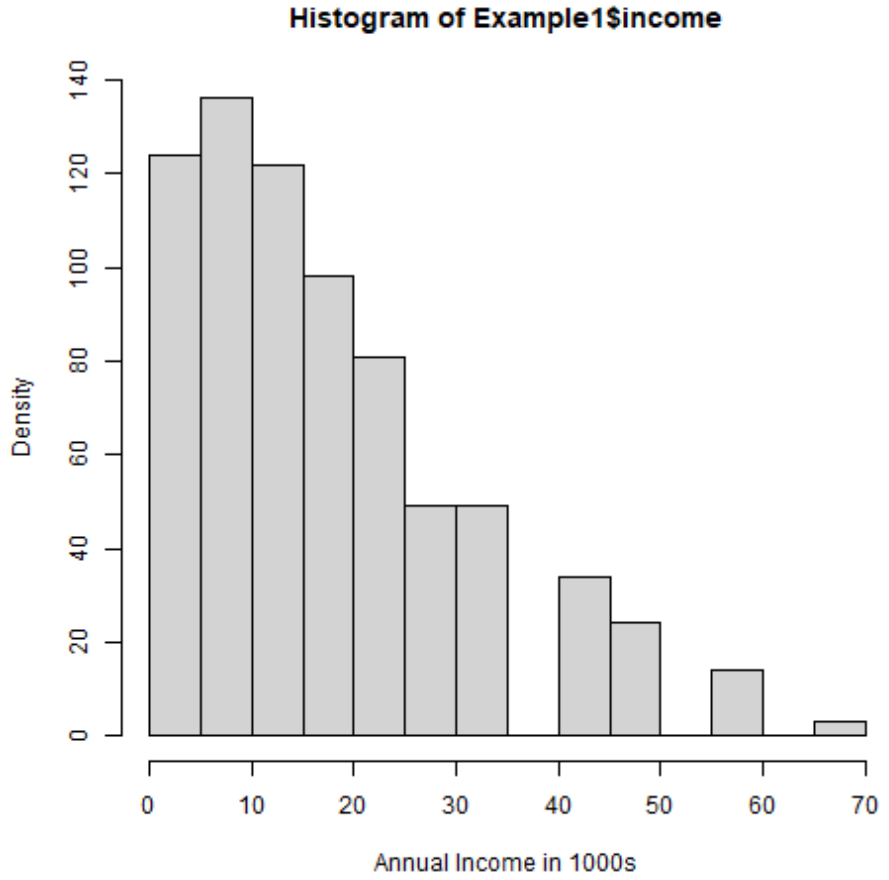


income: Annual Income in 1000s	Freq.	Percent	Cum.
0.25	19	2.59	2.59
0.98	18	2.45	5.04
1.72	23	3.13	8.17
2.21	10	1.36	9.54
2.69	13	1.77	11.31
3.19	14	1.91	13.22
3.67	15	2.04	15.26
4.41	12	1.63	16.89
5.51	39	5.31	22.21
6.74	43	5.86	28.07
7.96	24	3.27	31.34
9.19	30	4.09	35.42
10.41	37	5.04	40.46
11.64	33	4.50	44.96
13.48	52	7.08	52.04
15.93	49	6.68	58.72
18.38	49	6.68	65.40
22.05	81	11.04	76.43
26.95	49	6.68	83.11
33.08	49	6.68	89.78
40.42	34	4.63	94.41
49.00	24	3.27	97.68
58.80	14	1.91	99.59
68.60	3	0.41	100.00
Total	734	100.00	

What about Quantitative Variables?

- Instead: A histogram, which combines observations on the x-axis into "bins" (that you can and should choose)
  - Because different programs will create bins differently, changing what it looks like…
  - Note that the same "histogram" or "hist" command is still used

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



- Big picture: There are fewer people who make more money than who make less money!
- These pictures are designed to show the "**distribution**" of each quantitative variable, which we can summarize as follows…

Summarizing Quantitative Variables: 3 Salient Features

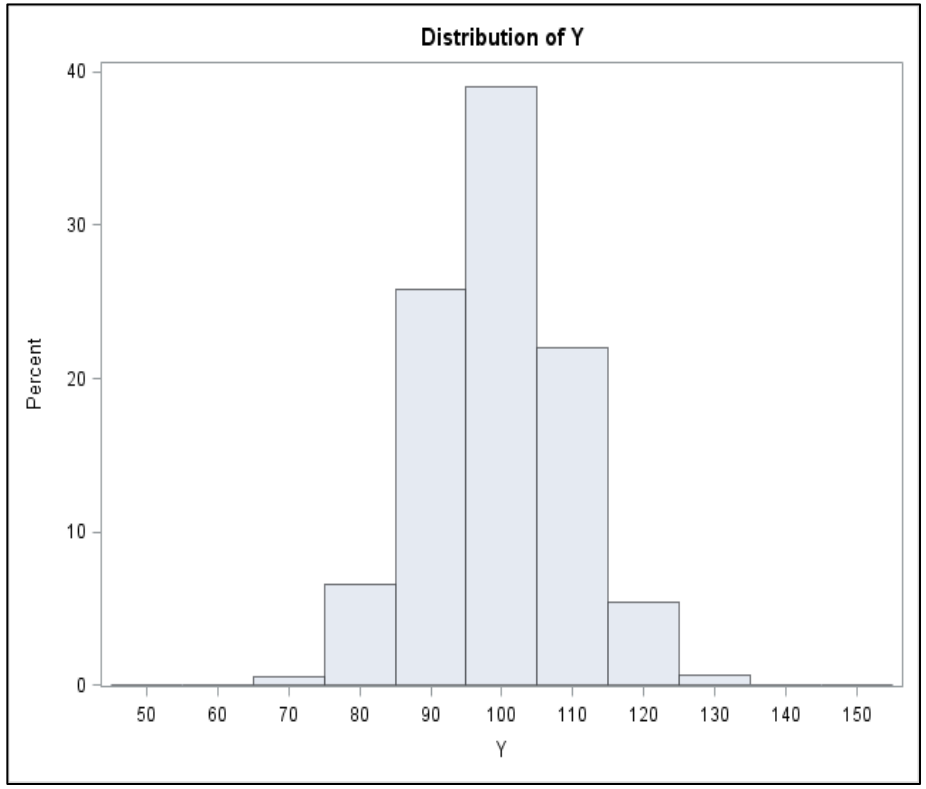
- **Central tendency**: think "**middle** of the distribution"; can be given by:
  - **Mean** = arithmetic average (abbreviated "$\boldsymbol{M}$" in results sections)
  - Also by **Median** = middle value ordered from most to least (50% percentile)
  - Also by **Mode** = most frequent value (but rarely mentioned in practice)

- **Dispersion**: think "**spread** of the distribution"; can be given by:
  - **Standard Deviation** (abbreviated "$\boldsymbol{SD}$" in results sections) = average deviation of any given observation (e.g., person) from the mean
  - **Variance** (abbreviated "$\boldsymbol{VAR}$" in results) = **squared** average deviation of any given observation (e.g., person) from the mean (so $VAR = SD^2$)
  - Also by **Inter-Quartile Range** (**IQR**) = distance from 25th to 75th percentile

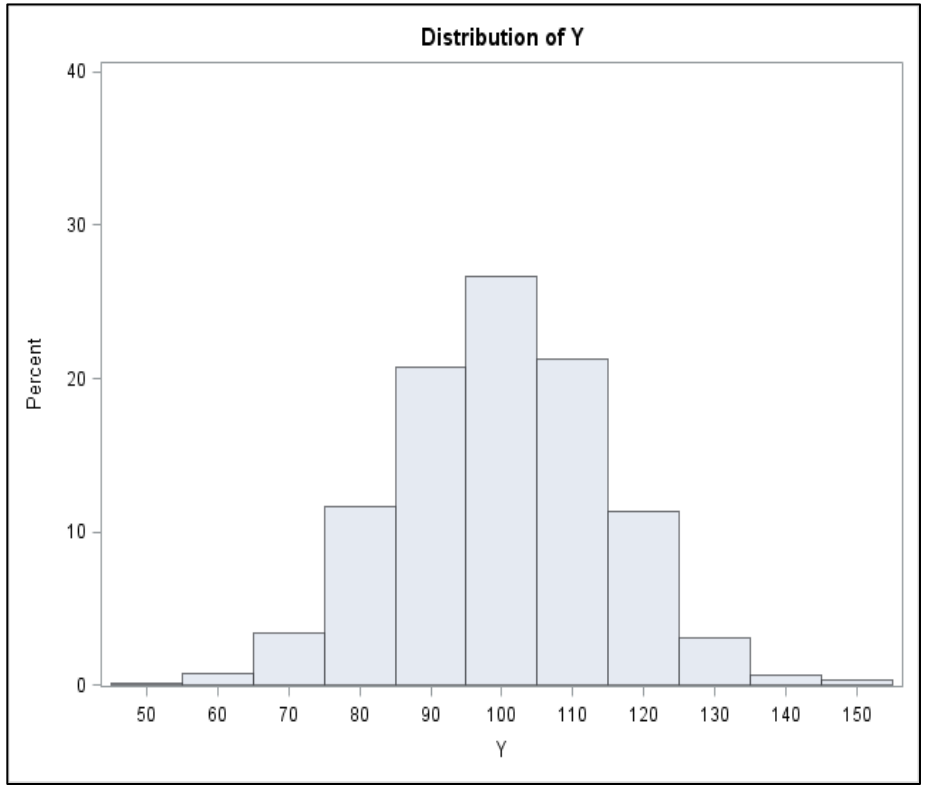
- **Skewness**: think "**asymmetry**" (more values on one side than the other)
  - Is often caused by natural boundaries in practice (e.g., counts at 0)
  - May be something to factor into your analysis, but is not usually reported

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

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

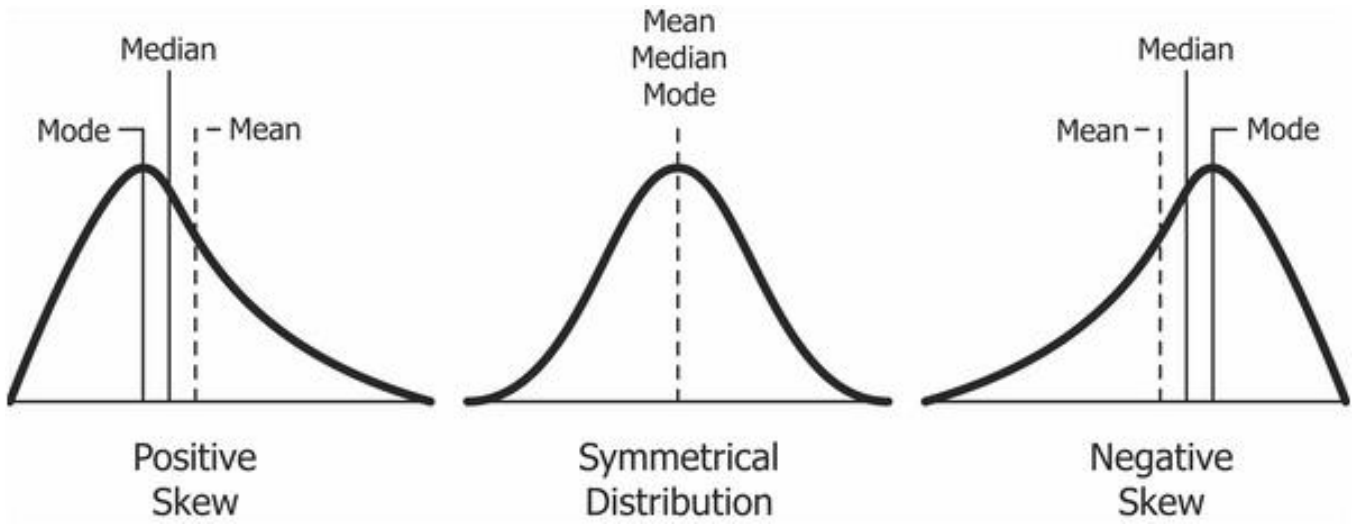


Standard Deviation (SD) = **15**, Variance (VAR) = SD*SD = 225



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 then it matters which is reported as an index of central tendency!

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://en.wikipedia.org/wiki/Skewness#/media/File:Relationship_between_mean_and_median_under_different_skewness.png

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 often made about population vs sample notation
  - **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 falls apart in describing the analysis model estimated and its results, in which mixing notation is more common (because people understand 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 used in computing descriptive statistics:
  - $\boldsymbol{y_i}$ = "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
  - $\boldsymbol{\overline{y}}$ = "y bar" = sample arithmetic* mean
    - Note the lack of an i subscript—this is because $\bar{y}$ is a constant, not a variable (as is N)

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

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

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

- Sample mean $\boldsymbol{\overline{y}}$ ("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):
- $Variance = s^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$

→ "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 $\boldsymbol{\sigma^2}$ ("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 person from the mean (e.g., SD describes a variable's expected dispersion)
  - Btw, in the denominator, $\boldsymbol{N - 1}$ is used instead of N to adjust for needing the sample mean first in order to calculate the sample variance; later this type of $\boldsymbol{N - 1}$ 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

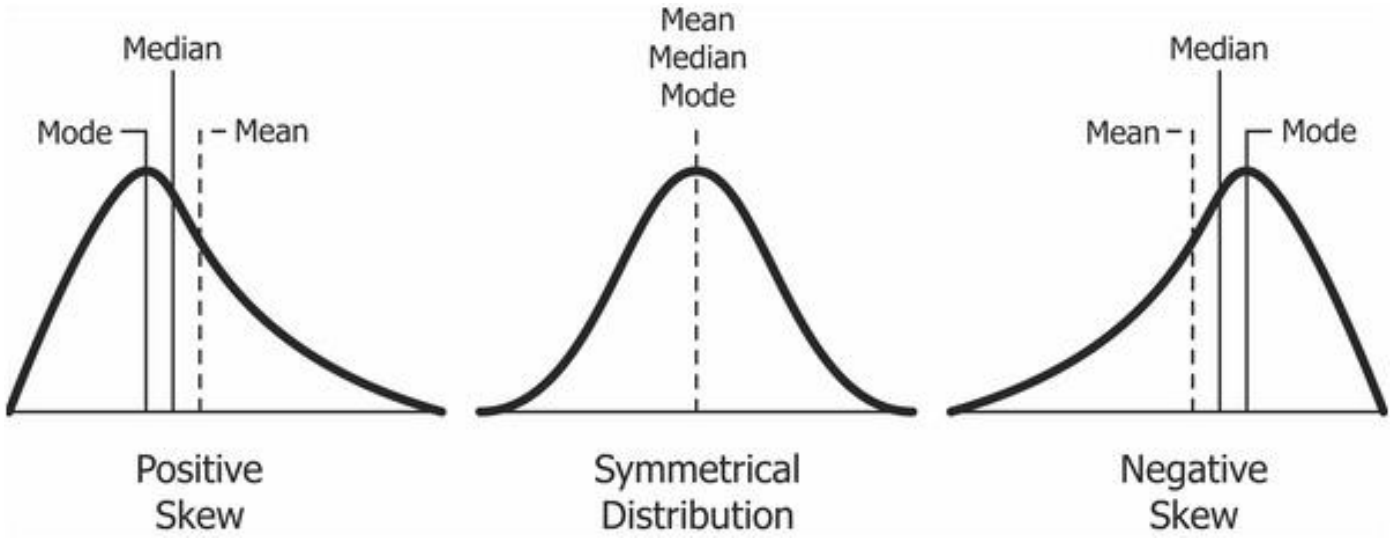
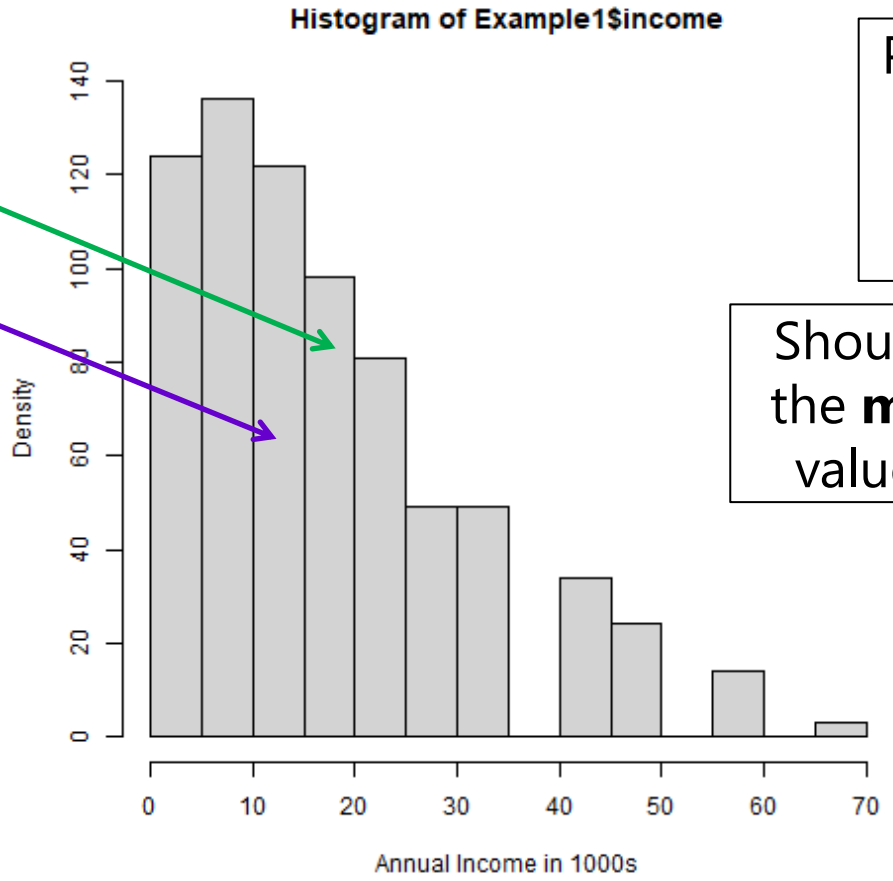


Image borrowed from: https://en.wikipedia.org/wiki/Skewness#/media/File:Relationship_between_mean_and_median_under_different_skewness.png

Example: Skewness = 1.16 in Income

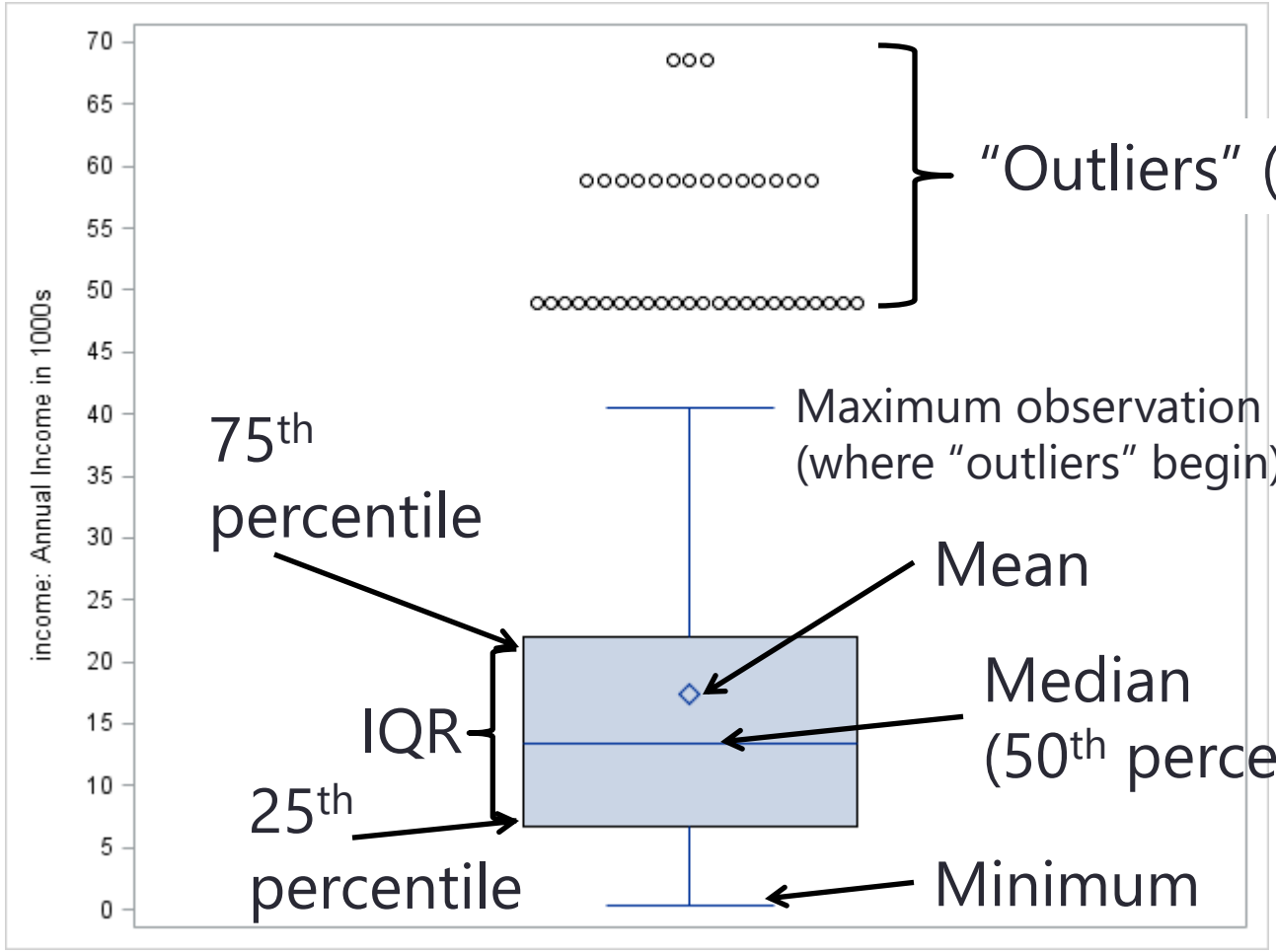
- **Central tendency:**
  - Mean (M) = 17.31
  - Median = 13.48
    - Btw, = 50th percentile
- **Dispersion:**
  - $VAR = SD^2$ = 190.21
  - SD = 13.79
  - Inter-quartile range:
    - IQR = 75th – 25th percentiles
    - IQR = 22.05 – 6.74 = 15.31



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

Should also report the **range**: the **minimum** and **maximum** values (0.25 and 68.60 here)

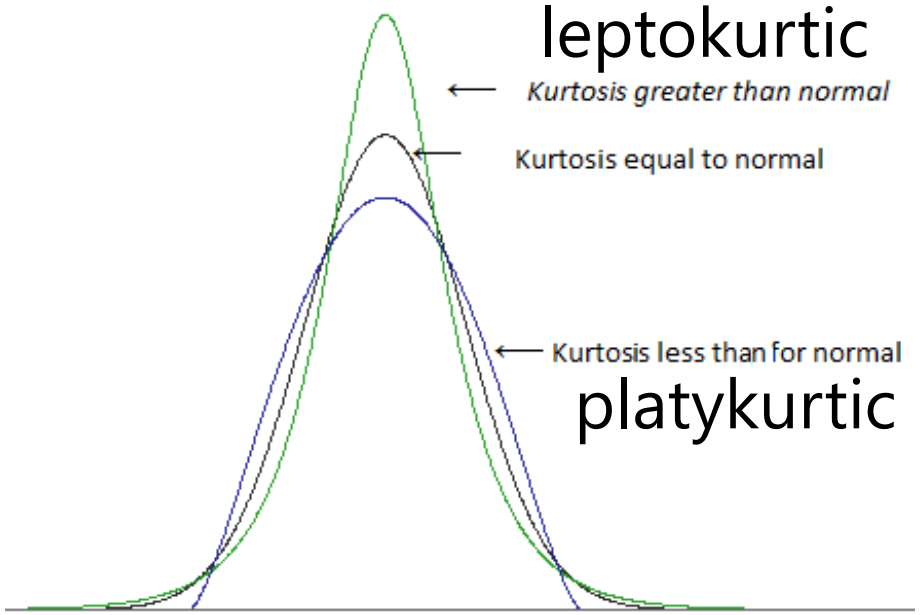
Summarizing Skewed Quantitative Variables using a "Box Plot"



Btw, One More Feature of Quantitative Variables: Kurtosis

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

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



→ 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: `describe` in R

```
# describe (from psych package) prints descriptive statistics for quantitative variables
# quant= requests list of quantiles, IQR requests inter-quartile range
# [ , c()] part says use all rows, but just columns named in c()
print("R Descriptive Statistics for Quantitative Variables income and age")
describe(x=Example1[ , c("income","age")], quant=c(.25,.50,.75), IQR=TRUE)
```

```
       vars   n  mean    sd median trimmed   mad   min  max range skew kurtosis   se   IQR Q0.25  Q0.5 Q0.75
income    1 734 17.30 13.79  13.47   15.49 12.71  0.24 68.6 68.35 1.16     1.08 0.51 15.31  6.74 13.47 22.05
age       2 734 42.06 13.38  41.00   41.57 14.83 18.00 75.0 57.00 0.29    -0.77 0.49 20.00 32.00 41.00 52.00
```

```
# embedding describe inside a print function allows better control of number of digits printed
print(describe(x=Example1[ , c("income","age")], quant=c(.25,.50,.75), IQR=TRUE), digits=3)

# to make sure it is using the describe function from the psych package, write it this way
print(psych::describe(x=Example1[ , c("income","age")], quant=c(.25,.50,.75), IQR=TRUE), digits=3)
```

```
       vars   n   mean     sd median trimmed    mad    min  max  range  skew kurtosis    se    IQR  Q0.25   Q0.5 Q0.75
income    1 734 17.303 13.792 13.475  15.493 12.713  0.245 68.6 68.355 1.156    1.075 0.509 15.312  6.737 13.475 22.05
age       2 734 42.063 13.378 41.000  41.573 14.826 18.000 75.0 57.000 0.293   -0.769 0.494 20.000 32.000 41.000 52.00
```

Note what's missing... Any guesses why?

Describing Quantitative Variables: R

```
        vars   n  mean    sd median trimmed   mad   min  max range skew kurtosis   se   IQR Q0.25  Q0.5 Q0.75
income     1 734 17.30 13.79  13.47   15.49 12.71  0.24 68.6 68.35 1.16     1.08 0.51 15.31  6.74 13.47 22.05
age        2 734 42.06 13.38  41.00   41.57 14.83 18.00 75.0 57.00 0.29    -0.77 0.49 20.00 32.00 41.00 52.00
```

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

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

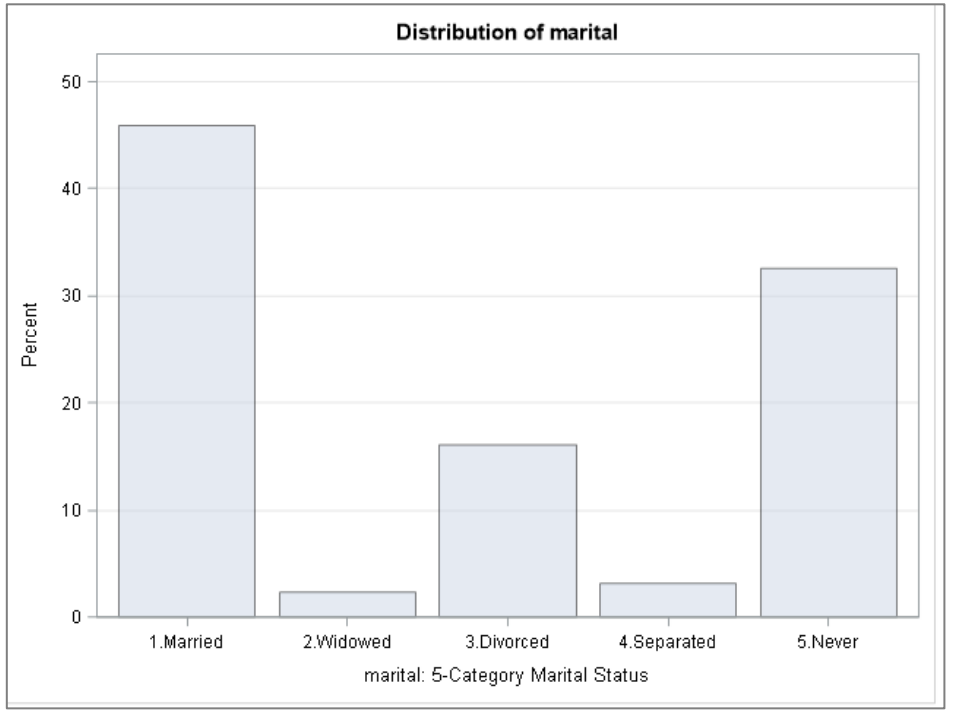
However, the results are then context-free (i.e., you only know what the result value means based on the code run immediately before it)—below is what the console would show

```
> # describe does not include variance, so here is a base R command to do so
> var(x=Example1$income)
[1] 190.20905
>
> # likewise, here are base R commands to get the mean and SD separately with more precision
> # can have more than one command on a line if separated by a semi-colon
> mean(x=Example1$income); sd(x=Example1$income)
[1] 17.302875
[1] 13.79163
```

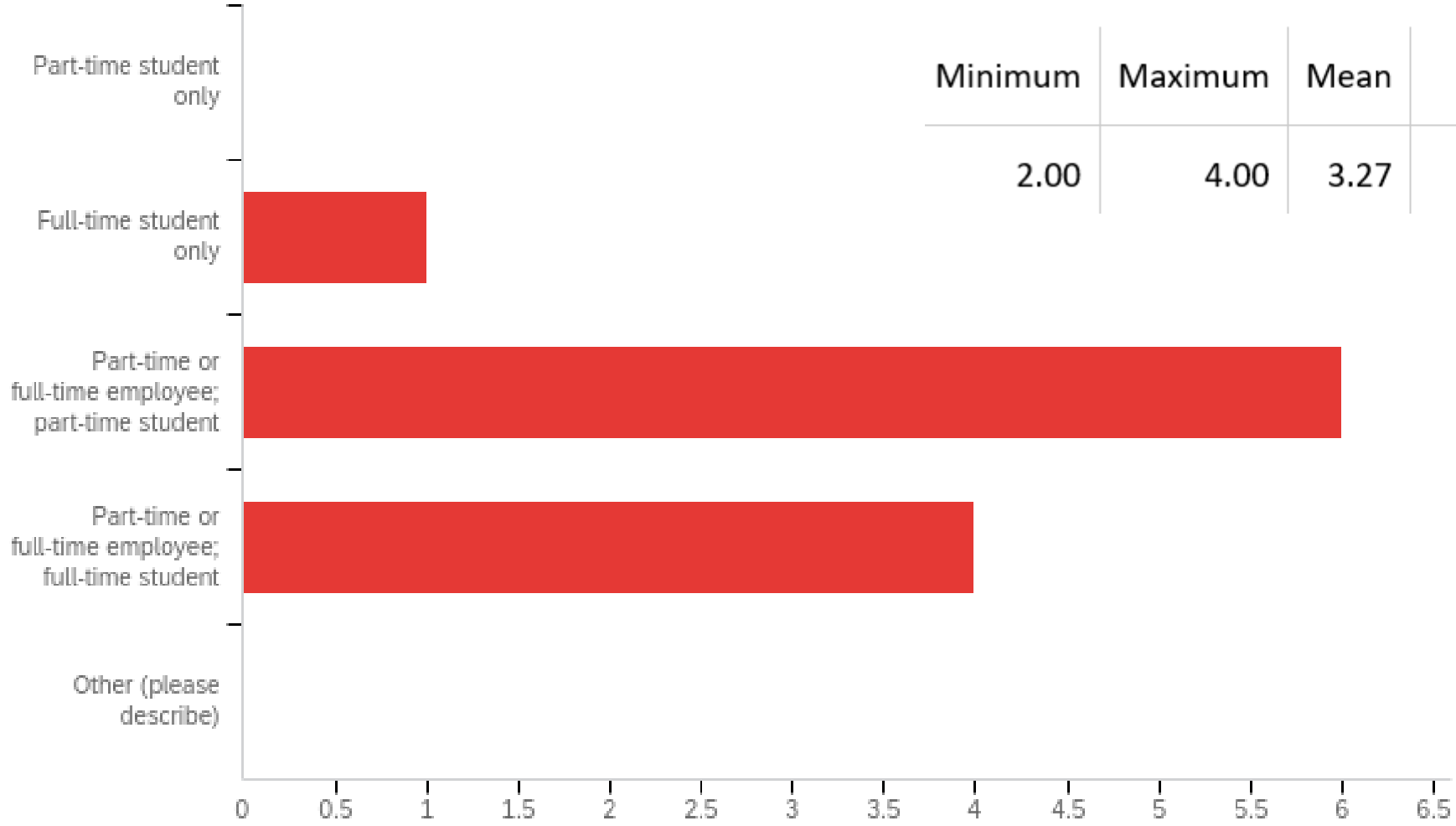
Means for Categorical Variables?

- For binary variables (coded 0 or 1), the mean is calculated the same way but it is called the "proportion $\boldsymbol{p}$" 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 any sense!
  - e.g., for nominal marital status, M = 2.74… ?!?
  - Software will give it to you anyway (user beware) ☺



Means for Nominal Variables? (a Qualtrics fail)

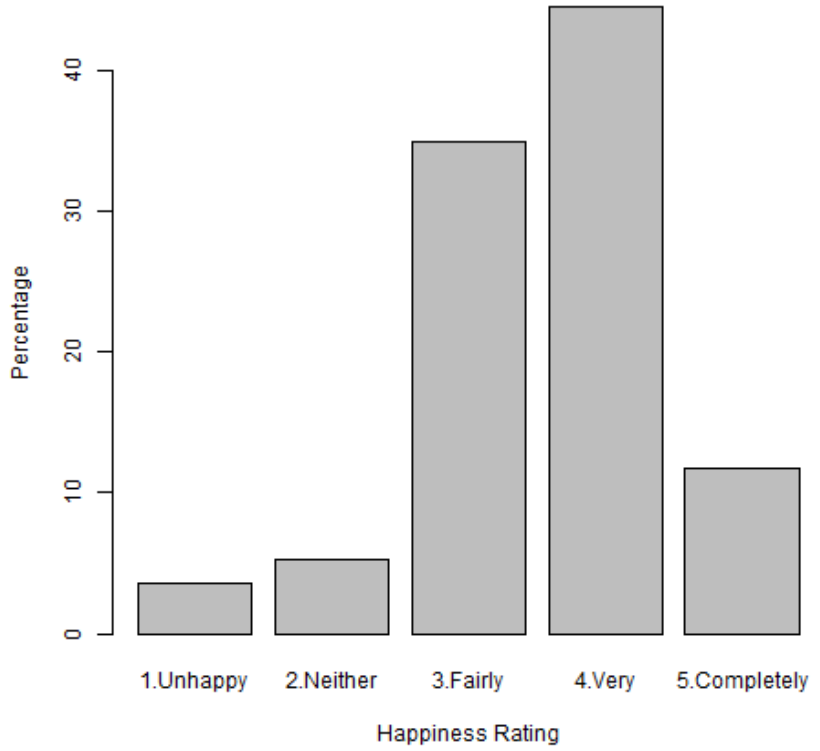


Minimum	Maximum	Mean	Std Deviation	Variance
2.00	4.00	3.27	0.62	0.38

Means for Ordinal Variables?

- What about **means 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"(*Lick-ert*, not *Like-ert*)



- Using a mean assumes equal distances between the options (interval measurement)
  - Stay tuned for ordinal predictors… whether to think of them as ordinal or "interval enough" is an empirical question!

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 = $\frac{\boldsymbol{1-2p}}{\sqrt{\boldsymbol{p*q}}}$

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 percentage of each category is enough to report (i.e., the pairwise variance and skewness values are not helpful)

Wrapping Up

- Which univariate descriptive statistics are relevant varies by type of variable:

  - **Quantitative variables (numbers are amounts):**
    - 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; mean → variance and skewness
    - Nominal with 3+ categories: % of each category; mean and SD make no sense
    - Ordinal with 3+ categories may be treated as quantitative, but this assumes interval measurement (i.e., equal distances between the numbers used as labels)
    - Bar graphs of % in each category make a good visual