MEASURES OF CENTRAL TENDENCY

Dr. Moshood Omotayo

OUTLINE

- Importance of measures of central tendency
- Types of variables
- Measures of central tendency
- Tabular and graphical display of data

IMPORTANCE OF MEASURES OF CENTRAL TENDENCY

A group of goats is called a trip. A group of whales is called a school.

In life, we often need to group things. We also often need to use just one value to represent an entire group, e.g. *"the common man on the street"*

A measure of central tendency is the simplest way to summarize and describe data.

Question: How tall are men in Ibadan?

Answer: <u>On average</u>, men in Ibadan are 5 feet 8 inches, ± 2 inches.

TYPES OF VARIABLES (TYPES OF DATA)

Health-related variables are either

- Continuous
  - Variables that take on quantitative values that ‘continues’
  - Infinite range of values between two consecutive units e.g. 1.1, 1.11, 1.111 and so on
  - E.g., blood pressure, birthweight
  - Summarized with mean, median, percentiles
- Discrete
  - Variables that take on specific quantitative values, often integers. *Counts.*
  - E.g. Number of COVID-19 patients diagnosed today, number of patients seen in clinic today
- Categorical
  - May be binary (yes=1, no=0) or have more categories e.g. (1=underweight, 2=normal weight, 3=overweight, 4=obese)

Categorical data may also be nominal or ordinal

- Nominal data
  - Categorical variables whose values cannot be quantitatively ordered
  - Could be coded numerically or with alternate names, but the codes are just representations
  - Example:
    - Sex is often coded 1- female, 0-male;
    - Colors could be 1-red, 2-black, 3-green, 4-yellow
- Ordinal data
  - Variables whose values can be quantitatively ordered, and can be 'ranked'.
  - Example
    - HIV Disease stage could be coded 1- Stage 1, 2 – Stage II, 3 – Stage III, 4 – Stage IV
    - BMI could be coded 1=underweight, 2=normal weight, 3=overweight, 4= obese

POLL 1:

You measured systolic blood pressure (SBP) for 5,000 patients. You then group the SBP measured into 4 categories: 1) <90 mmHg, 2) between 90 and 120 mmHg, 3) between 120 and 130 mmHg, and 4) >130 mmHg. **What type of variable is your new SBP variable?**

a) Continuous

b) Binary

c) Ordinal categorical

d) Nominal categorical

e) I don’t know

POLL 1:

You measured systolic blood pressure (SBP) for 5,000 patients. You then group the SBP measured into 4 categories: 1) <90 mmHg, 2) between 90 and 120 mmHg, 3) between 120 and 130 mmHg, and 4) >130 mmHg. **What type of variable is your new SBP variable?**

a) Continuous

b) Binary

c) [✓ Ordinal categorical]

d) Nominal categorical

e) I don’t know

Ratio

- Relative magnitude of two quantities or a comparison of any two values. Simply one variable, the numerator, divided by another, the denominator
- The numerator and the denominator need not be related. You can use a ratio to compare the number of diabetics to non-diabetics in a study. You can also use it to compare the number of physicians to tech companies in a country
- Ratios are used as both descriptive measures and analytic tools in epidemiology
- As a descriptive measure, ratios can describe case-to-control ratio of study participants or ratio of adults to children
- As an analytic tool, ratios can be computed for occurrence of illness ,death or injury between two groups e.g risk ratio, odds ratio etc
- Proportions and rates are special types of ratios in epidemiology

Proportions

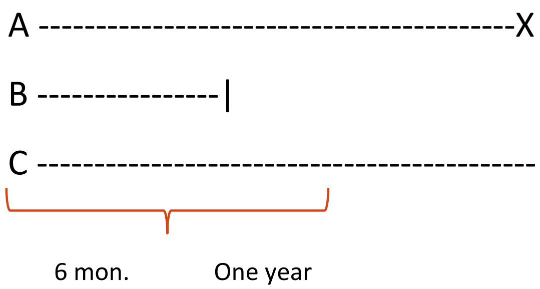
- Simply one variable, the numerator, divided by another, the denominator
- Note: in a proportion, the numerator is a portion of the denominator
- It is important to know what the numerator and denominator are when you encounter a proportion, but also the context in which they were measured.
- Ranges from 0 to 1.
- Prevalence
  - Proportion of cases of a disease in the population **at one time point**. Measure of disease burden, typically, in a cross-sectional study/survey. $Prevalence = \frac{\#\ existing\ cases}{\#\ individuals\ in\ study\ population}$
- Cumulative incidence
  - Proportion of **new** cases of a disease **during a period of time**. Measure of disease burden, typically in cohort studies.
  - $Cumulative\ incidence = \frac{\#\ new\ cases\ during\ \ period\ from\ t_0\ to\ t_1}{\#\ individuals\ at\ risk\ at t_0}$

Rate

- A measure of occurrence of a disease **per unit time**.
- The rate is particularly useful in cohort studies because individuals may begin the study at different times, may leave at different times, may spend different amounts of time in the study.
- Ranges from 0 to ∞.
- Incidence rate
  - Number of **new** cases of a disease **per unit time**.

$$Incidence\ rate(IR) = \frac{\#\ new\ cases\ during\ \ period\ from\ t_0\ to\ t_1}{Total\ time\ spent\ in\ study\ by\ participants}$$

Example



A --X

B ----------------|

C --

X = death

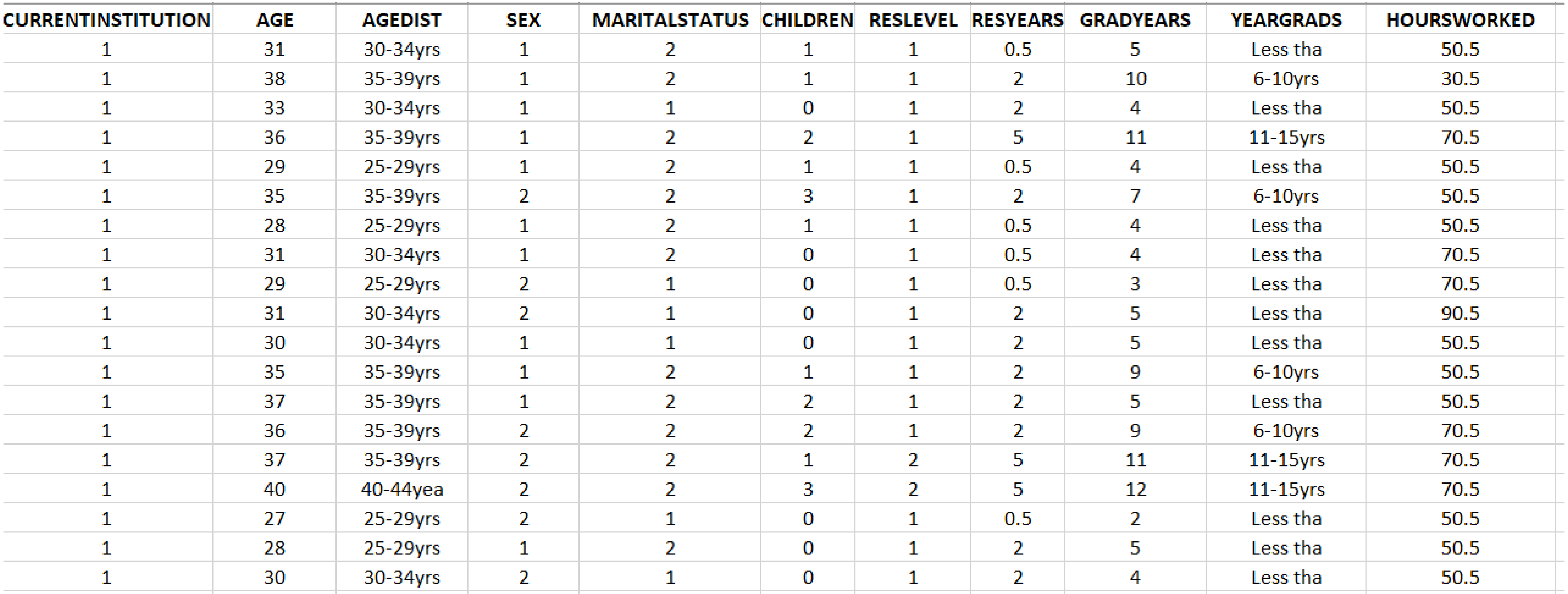
| = participant drop out

IR = 1 death per 2.5 person years
or 0.4 deaths per person year

Statisticians do three types of analysis

- Describe the population or sample (*description)*
- Predict an outcome (*prediction – much of machine learning and AI)*
- Clarify relationships of exposures and outcomes (*much of epidemiologic analysis and causal inference)*

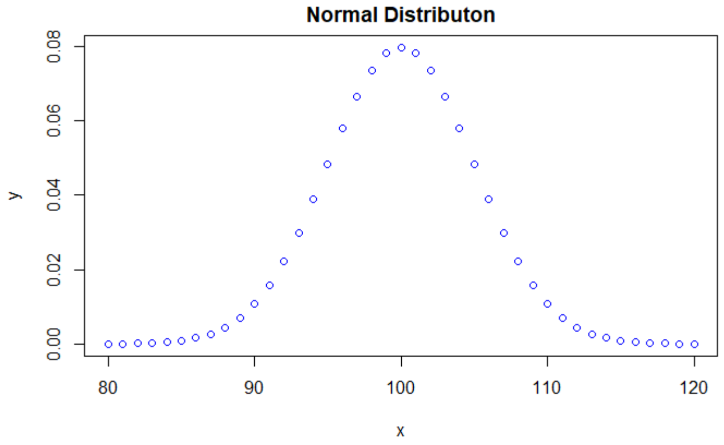
Here’s an example dataset showing different variable types



CURRENTINSTITUTION	AGE	AGEDIST	SEX	MARITALSTATUS	CHILDREN	RESLEVEL	RESYEARS	GRADYEARS	YEARGRADS	HOURSWORKED
1	31	30-34yrs	1	2	1	1	0.5	5	Less tha	50.5
1	38	35-39yrs	1	2	1	1	2	10	6-10yrs	30.5
1	33	30-34yrs	1	1	0	1	2	4	Less tha	50.5
1	36	35-39yrs	1	2	2	1	5	11	11-15yrs	70.5
1	29	25-29yrs	1	2	1	1	0.5	4	Less tha	50.5
1	35	35-39yrs	2	2	3	1	2	7	6-10yrs	50.5
1	28	25-29yrs	1	2	1	1	0.5	4	Less tha	50.5
1	31	30-34yrs	1	2	0	1	0.5	4	Less tha	70.5
1	29	25-29yrs	2	1	0	1	0.5	3	Less tha	70.5
1	31	30-34yrs	2	1	0	1	2	5	Less tha	90.5
1	30	30-34yrs	1	1	0	1	2	5	Less tha	50.5
1	35	35-39yrs	1	2	1	1	2	9	6-10yrs	50.5
1	37	35-39yrs	1	2	2	1	2	5	Less tha	50.5
1	36	35-39yrs	2	2	2	1	2	9	6-10yrs	70.5
1	37	35-39yrs	2	2	1	2	5	11	11-15yrs	70.5
1	40	40-44yea	2	2	3	2	5	12	11-15yrs	70.5
1	27	25-29yrs	2	1	0	1	0.5	2	Less tha	50.5
1	28	25-29yrs	1	2	0	1	2	5	Less tha	50.5
1	30	30-34yrs	2	1	0	1	2	4	Less tha	50.5

Quantitative Summaries

MEAN, MEDIAN



Mean

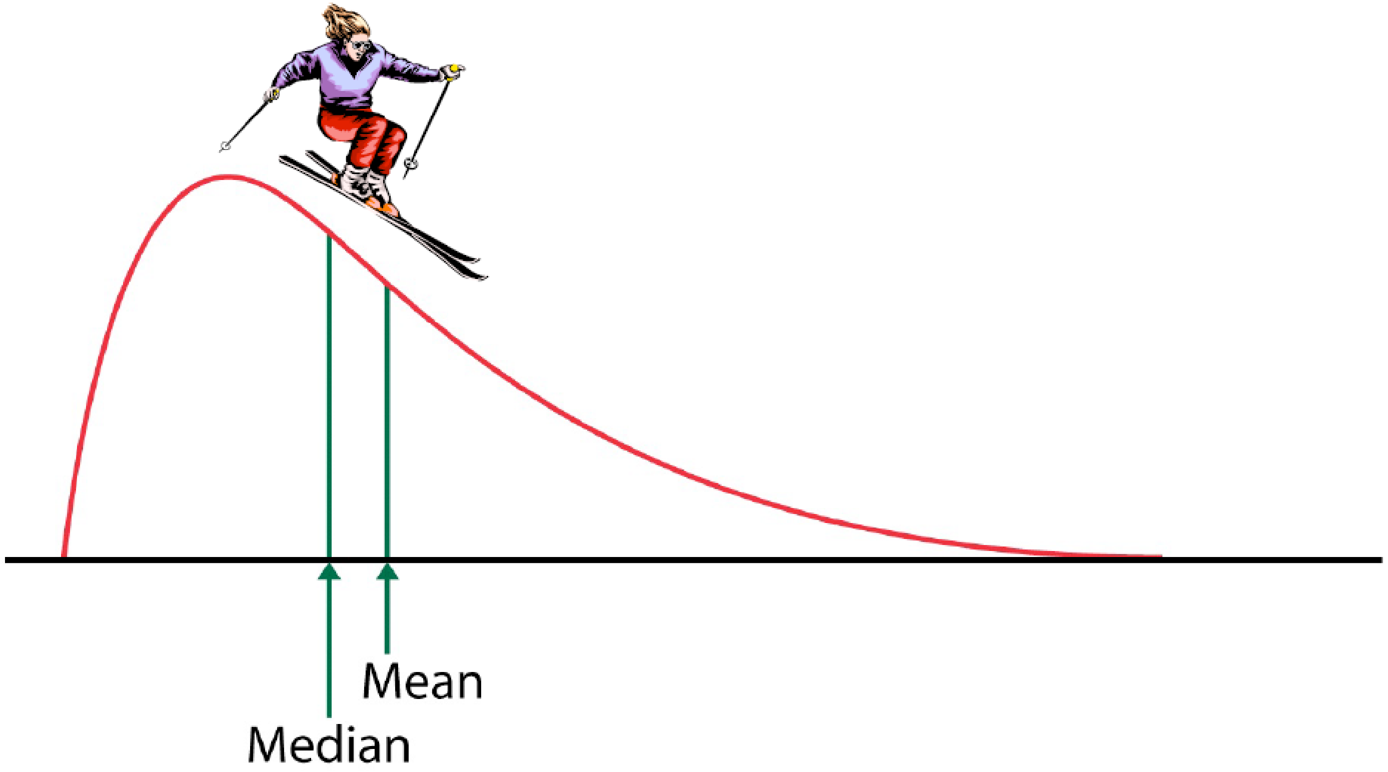
- Given a list of numbers, x, the mean value is simply the sum of the values divided by the number of values. $Mean, \bar{x} = \frac{\sum x}{n}$
- The mean is popularly called the **'average'**. This is the **arithmetic mean**.

Median

- Given a list of numbers, x, the median value is simply the value in the middle of the list if you organize the numbers in ascending or descending order.
- $median = \frac{(n+1)}{2}$th value of ordered observations

The median and the mean could be similar or close if the distribution of the variable is symmetric such as a normal distribution. If the distribution is not symmetric, the median is usually preferred over the mean. This is because the mean is sensitive to outliers.

Idealized right-skewed distribution: Mean larger than Median



The median is robust to outliers but the mean is not.

Take an example dataset for the ages of 10 resident doctors from our dataset.

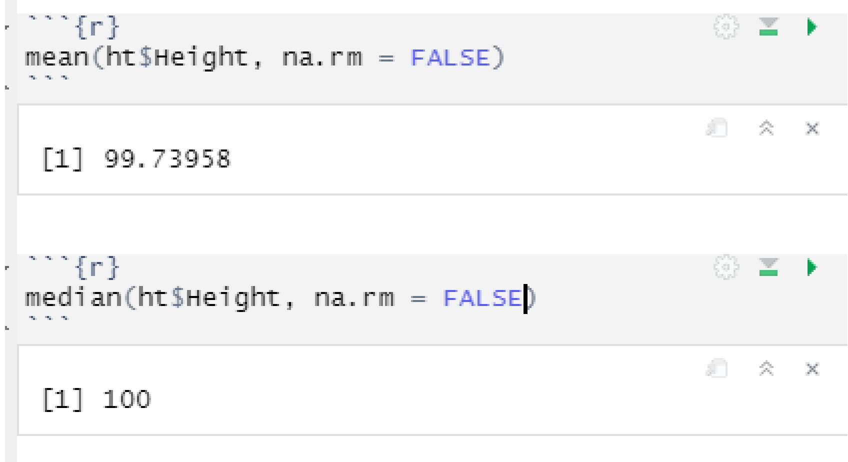
- 31, 38, 33, 36, 29, 35, 28, 31, 29, 31
- Mean = $\frac{31+\ 38+33+36+29+35+28+31+29+31}{10} = \frac{321}{10} = 32.1$
- Median = 31

If the last resident were switched with one whose age is 51, we get:

- Mean = $\frac{31+\ 38+33+36+29+35+28+31+29+51}{10} = \frac{341}{10} = 34.1,$ which is a meaningful change
- Median = 31, no change.

POLL 2

You recorded the heights of the kids attended to at the pediatric outpatient clinic each day this week. First, you plot the histogram for day 1. If your histogram indicates the data may be normally distributed, which average would you use?

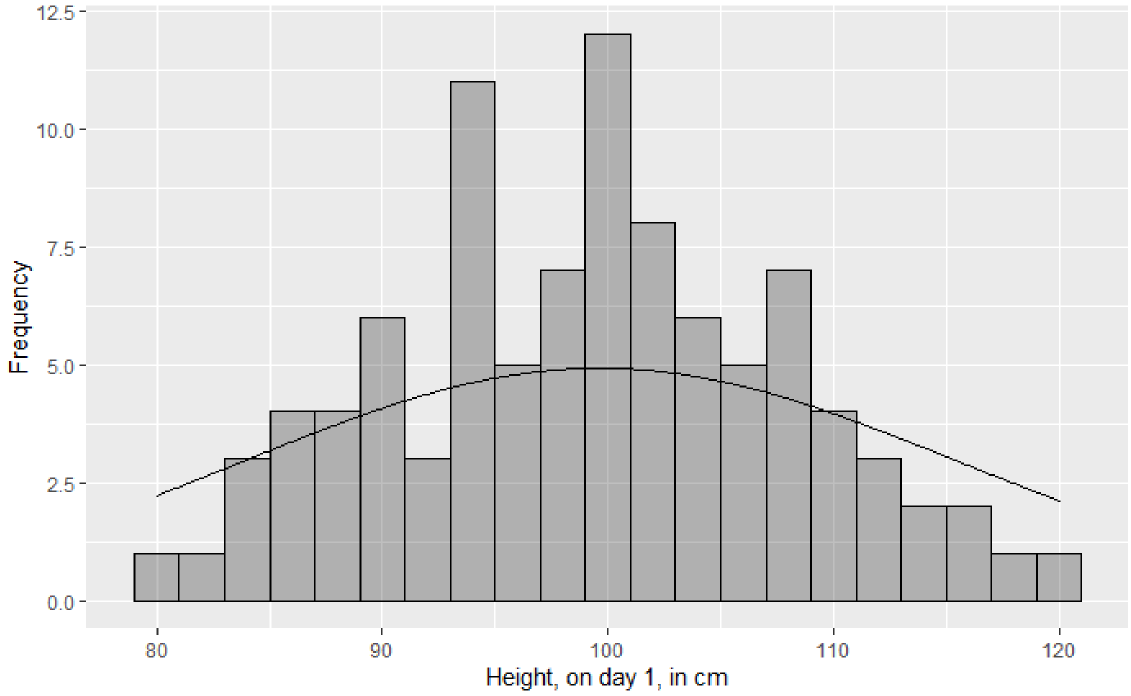


~~~~
```{r}
mean(ht$Height, na.rm = FALSE)
```
~~~~

```
[1] 99.73958
```

~~~~
```{r}
median(ht$Height, na.rm = FALSE)
```
~~~~

```
[1] 100
```

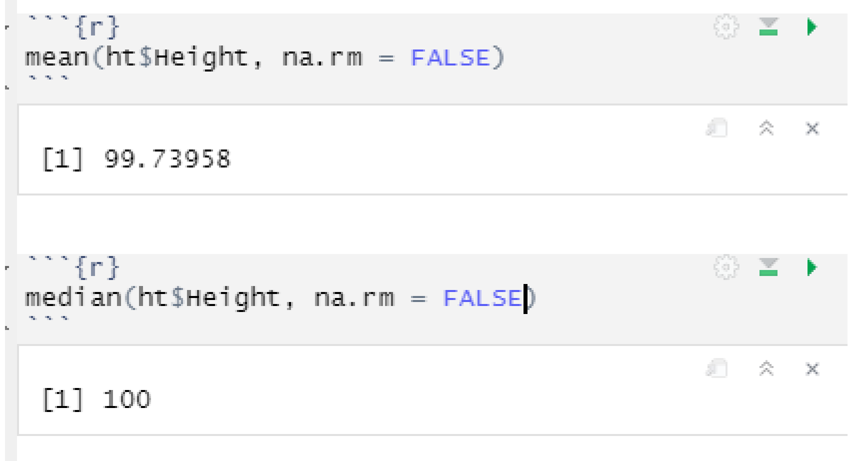


a) Mean

b) Median

POLL 2

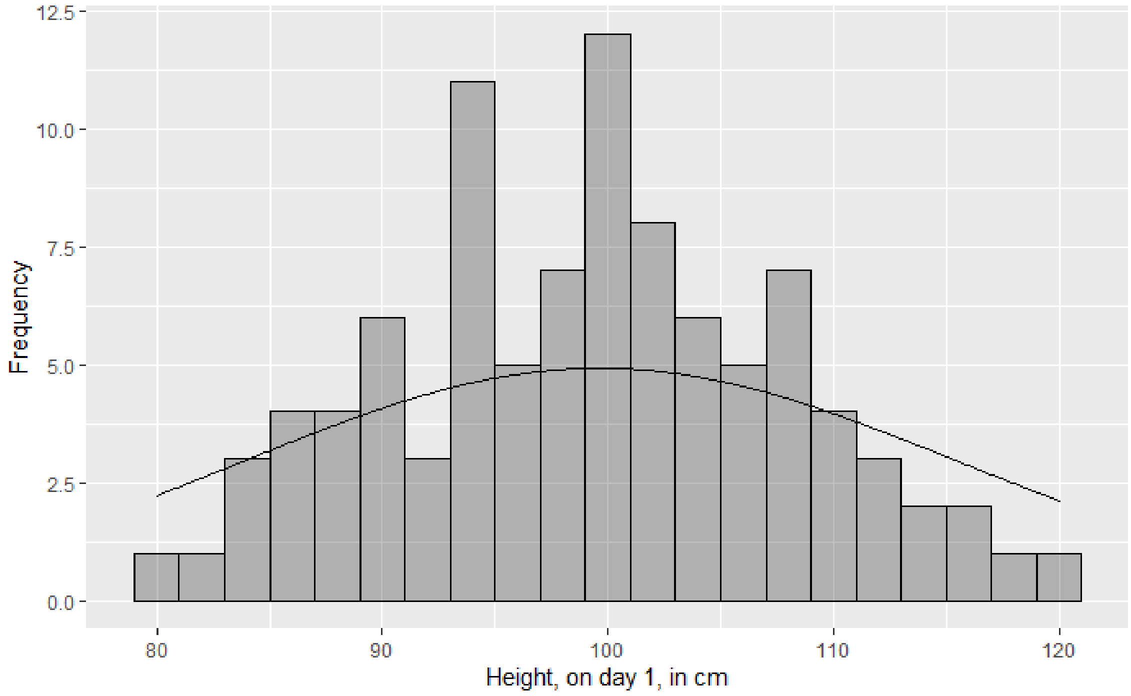
You recorded the heights of the kids attended to at the pediatric outpatient clinic each day this week. First, you plot the histogram for day 1. If your histogram indicates the data may be normally distributed, which average would you use?



~~~
```{r}
mean(ht$Height, na.rm = FALSE)
~~~

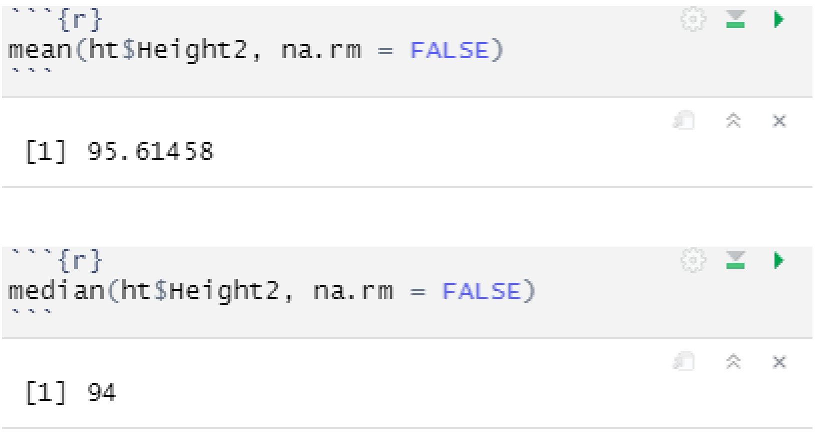
[1] 99.73958

```{r}
median(ht$Height, na.rm = FALSE)
```

[1] 100
```



[✓ a) Mean]

b) Median
```

# POLL 3

In a different study, you also have the data on the heights of kids at the outpatient clinic and plot the histogram. If your histogram indicates the data is not normally distributed, which average would you use?

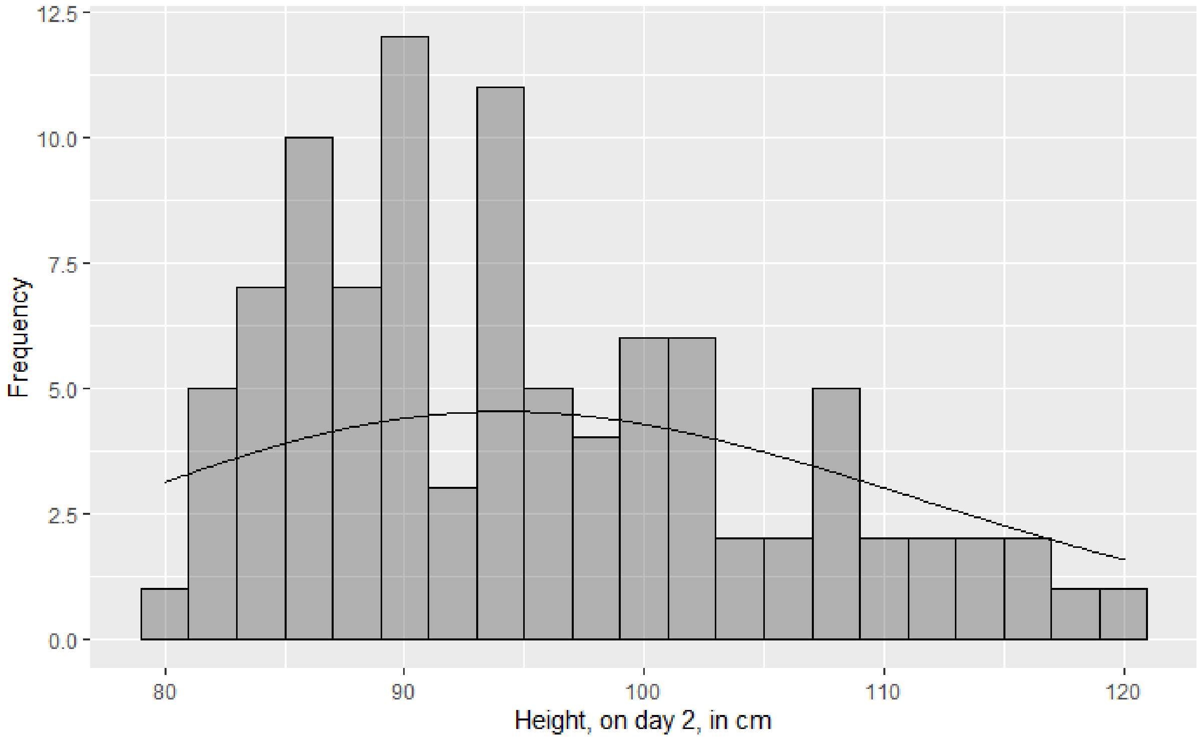


~~~
```{r}
mean(ht$Height2, na.rm = FALSE)
```
~~~

```
[1] 95.61458
```

~~~
```{r}
median(ht$Height2, na.rm = FALSE)
```
~~~

```
[1] 94
```

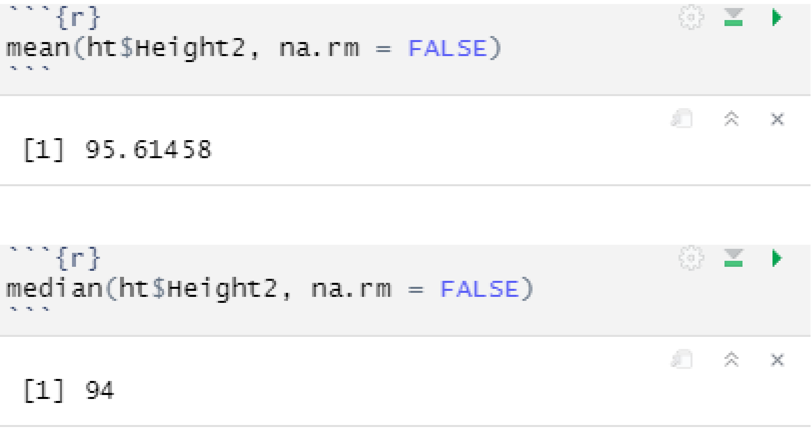


a) Mean

b) Median

POLL 3

In a different study, you also have the data on the heights of kids at the outpatient clinic and plot the histogram. If your histogram indicates the data is not normally distributed, which average would you use?

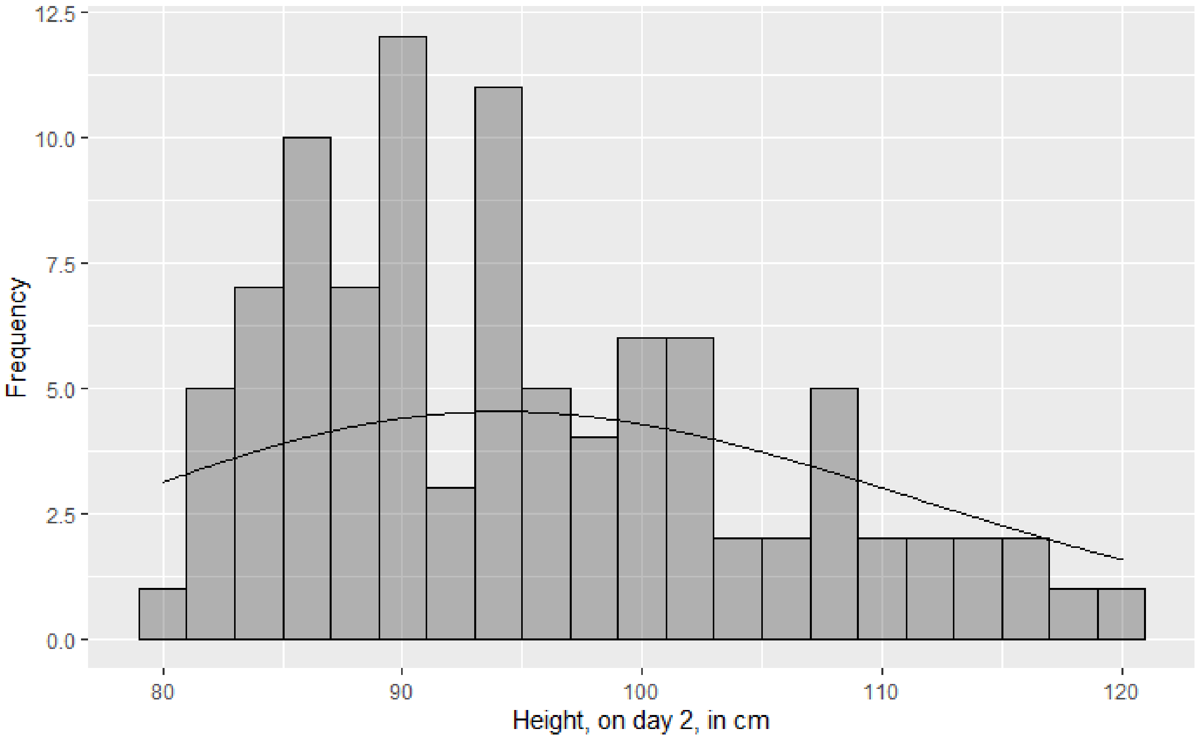


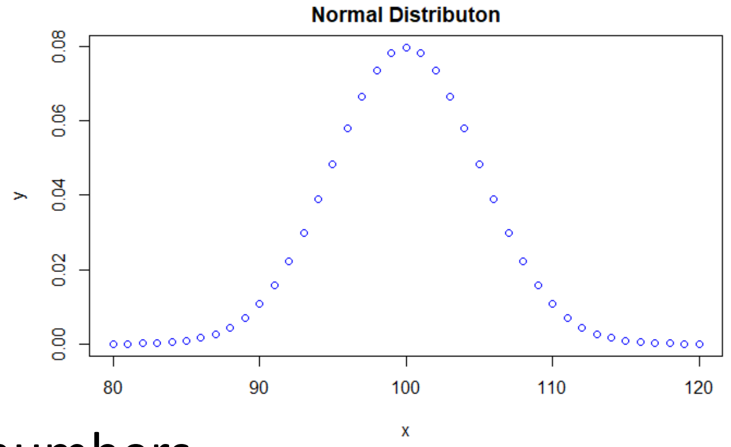
~~~
```{r}
mean(ht$Height2, na.rm = FALSE)
```

[1] 95.61458
~~~

~~~
```{r}
median(ht$Height2, na.rm = FALSE)
```

[1] 94
~~~

a) Mean

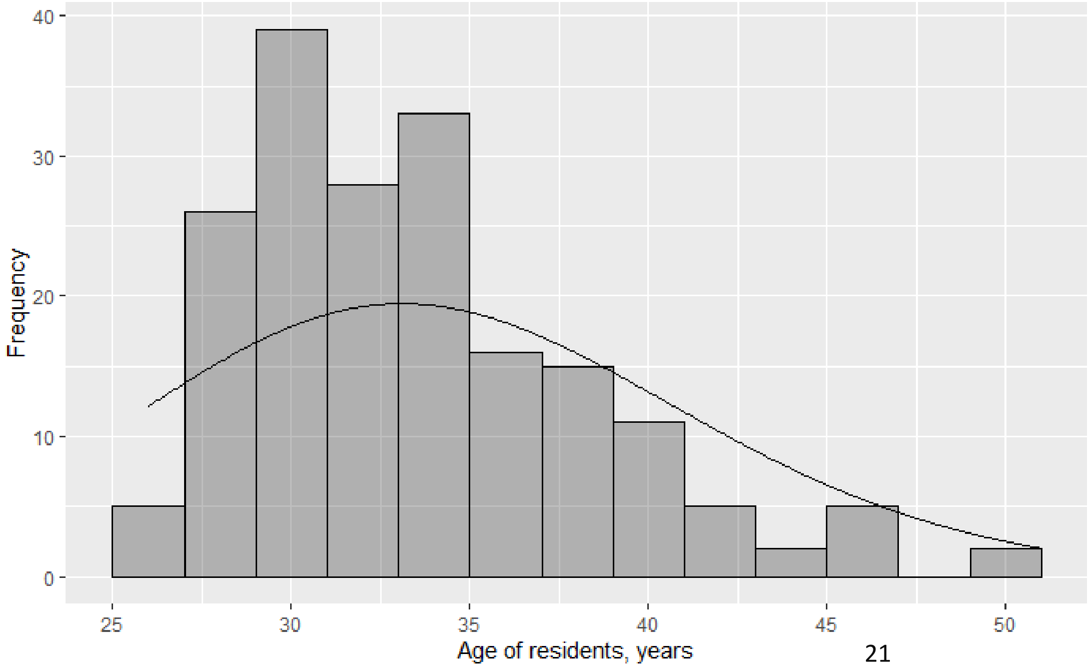
[✓ b) Median]





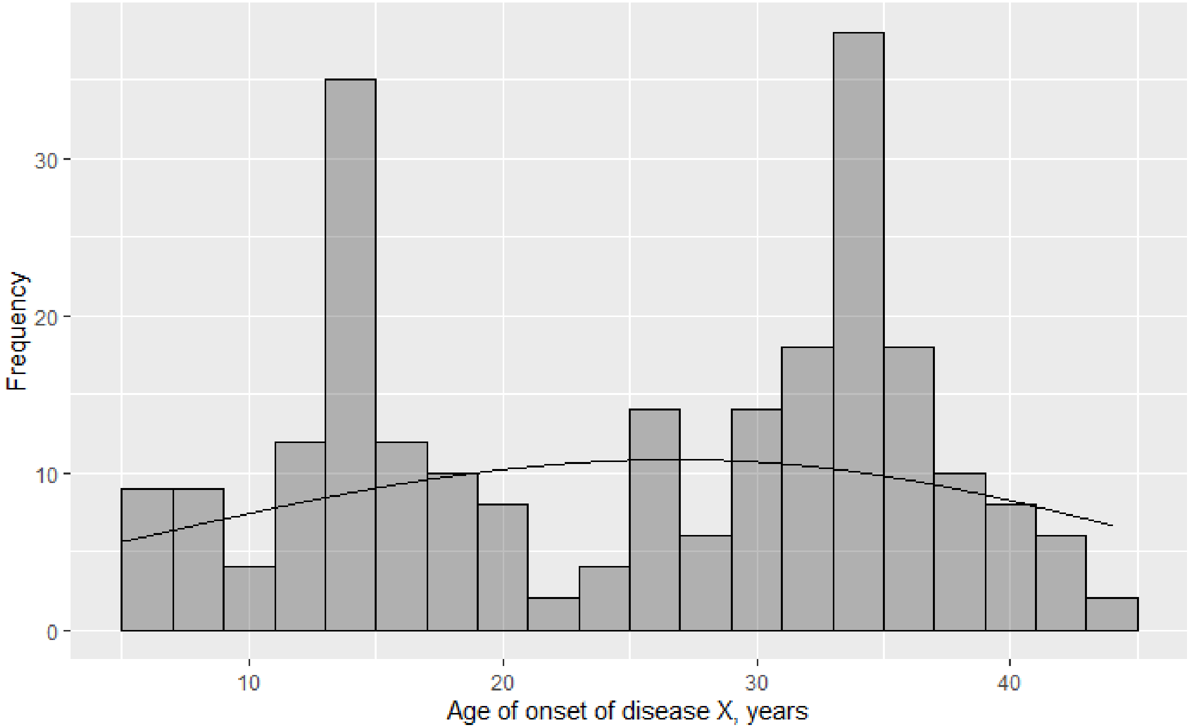
MODE

- The mode refers to the most commonly occurring of a set of numbers.
- In a normally distributed data where each point or bar represents a single unit of the variable, the mean, median and mode would be the same.



MODE

- Sometimes, there may be more than one mode in a dataset.
- For example, the age of onset of narcolepsy is bimodal
- If you don't plot this histogram, you would think the data was normally distributed *(as this density line erroneously shows)*, and you would miss important detail.



FREQUENCY TABLES

We use frequency tables to summarize categorical variables, whether nominal or ordered.

Age categories	N (%)
25 – 29 yrs	31 (16.6)
30 – 34 yrs	75 (40.1)
35 – 39 yrs	56 (29.9)
40 – 44 yrs	18 (9.6)
45 – 49 yrs	5 (2.7)
50 – 54 yrs	2 (1.1)

Sex	N (%)
Female	62 (33.2%)
Male	125 (66.8%)

Q: What can you infer from these frequency tables?

RELATIVE FREQUENCIES

Sometimes we want to see how one variable occurs in relation to another. Or we want to compare a variable in two populations of different sizes. For instance, look at these two cross-tabulations.

Within sex categories

Age categories	Female, N (%)	Male, N (%)	Total
25 – 29 yrs	18 (58%)	13 (42%)	31 (100%)
30 – 34 yrs	25 (33%)	50 (67%)	75 (100%)
35 – 39 yrs	12 (21%)	44 (79%)	56 (100%)
40 – 44 yrs	5 (28%)	13 (72%)	18 (100%)
45 – 49 yrs	2 (40%)	3 (60%)	5 (100%)
50 – 54 yrs	0 (0%)	2 (100%)	2 (100%)

Within age categories

Age categories	Female, N (%)	Male, N (%)
25 – 29 yrs	18 (29%)	13 (10%)
30 – 34 yrs	25 (40%)	50 (40%)
35 – 39 yrs	12 (19%)	44 (35%)
40 – 44 yrs	5 (8%)	13 (10%)
45 – 49 yrs	2 (3%)	3 (2%)
50 – 54 yrs	0 (0%)	2 (2%)
Total	62 (100%)	125 (100%)

What the results tables for scientific papers often look like

Age distribution , years	Pregnant	Male, N (%)	Total
25 – 29 yrs	18 (58%)	13 (42%)	31 (100%)
30 – 34 yrs	25 (33%)	50 (67%)	75 (100%)
35 – 39 yrs	12 (21%)	44 (79%)	56 (100%)
40 – 44 yrs	5 (28%)	13 (72%)	18 (100%)
45 – 49 yrs	2 (40%)	3 (60%)	5 (100%)
50 – 54 yrs	0 (0%)	2 (100%)	2 (100%)

POLL 4

As an epidemiologist, the health department tasked you with developing a metric for monitoring the risk of death during childbirth in your city. From review of death certificates, you found that in the last 12 months 10 women died due to causes related to childbirth but could not determine whether they had a livebirth or stillbirth. You also found from review of birth records that there were 1000 live births in the city in the last 12 months. You presented a metric of 10 maternal deaths/1000 live births. What type of metric did you present to the health department.

a) Descriptive Proportion

b) Analytic Ratio

c) Descriptive Ratio

d) Rate

e) Analytic Proportion

POLL 4

As an epidemiologist, the health department tasked you with developing a metric for monitoring the risk of death during childbirth in your city. From review of death certificates, you found that in the last 12 months 10 women died due to causes related to childbirth but could not determine whether they had a livebirth or stillbirth. You also found from review of birth records that there were 1000 live births in the city in the last 12 months. You presented a metric of 10 maternal deaths/1000 live births. What type of metric did you present to the health department.

a) Descriptive Proportion

b) Analytic Ratio

[✓ c) Descriptive Ratio]

d) Rate

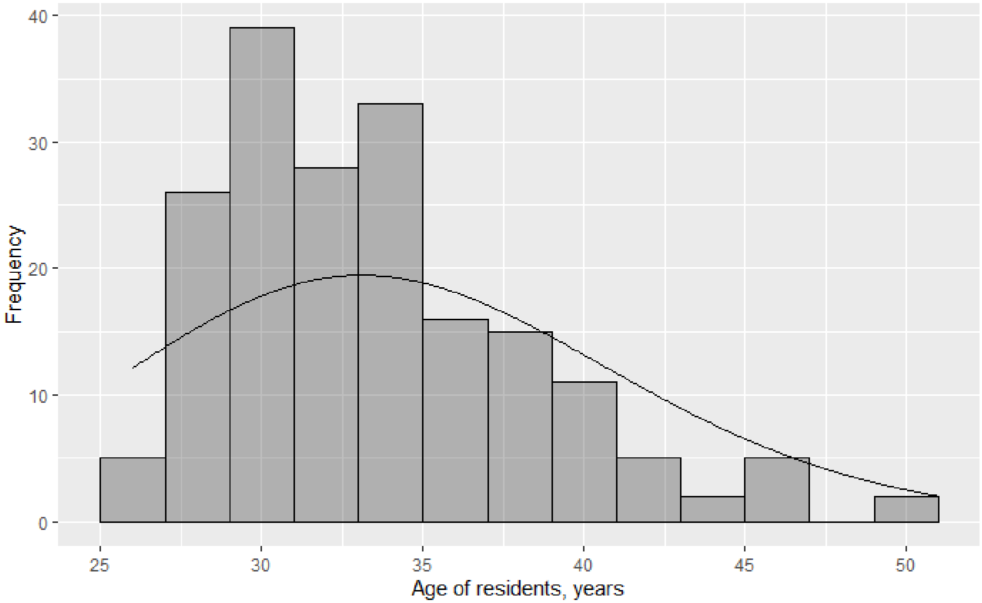
e) Analytic Proportion

Basic principles for graphical displays

- Make sure it shows the data you really intend to show. Ask someone else to take a quick glance.
- Help the reader to think about the substance.
- Avoid distorting the data – especially by using weird axes.
- Encourage the eye to make comparisons – color, axes
- Be closely integrated with the statistical and verbal descriptions of the data

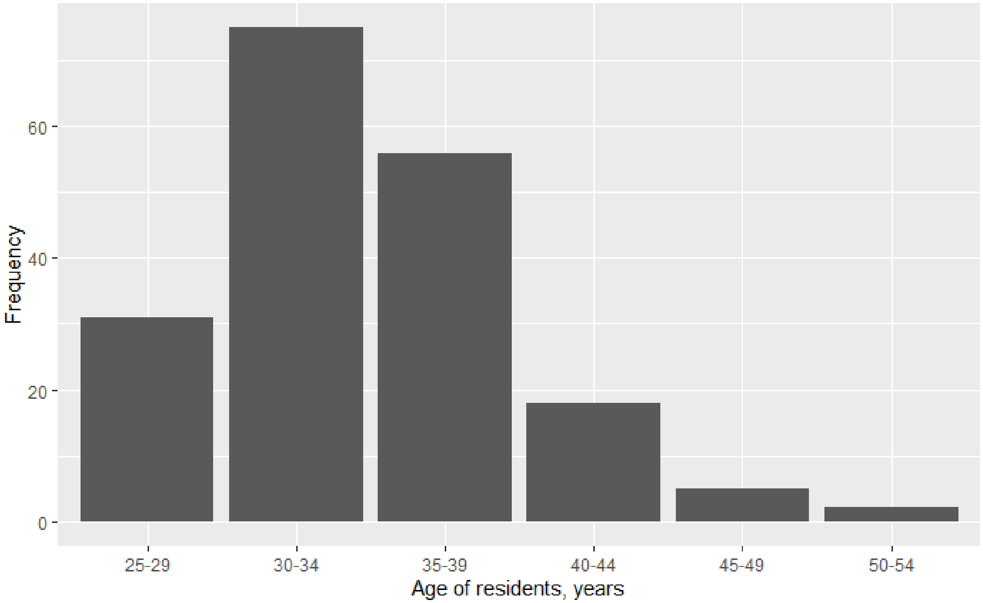
The histogram is great to present continuous data. The bar chart is great for categorical data.

Histogram, with density line



You want to carefully choose the number of bins so that patterns in the data can be obvious. 5 – 10 bins is common.

Bar chart



Histograms are sensitive to bin width.

Check out these two histograms showing the length of time of service calls at a bank.

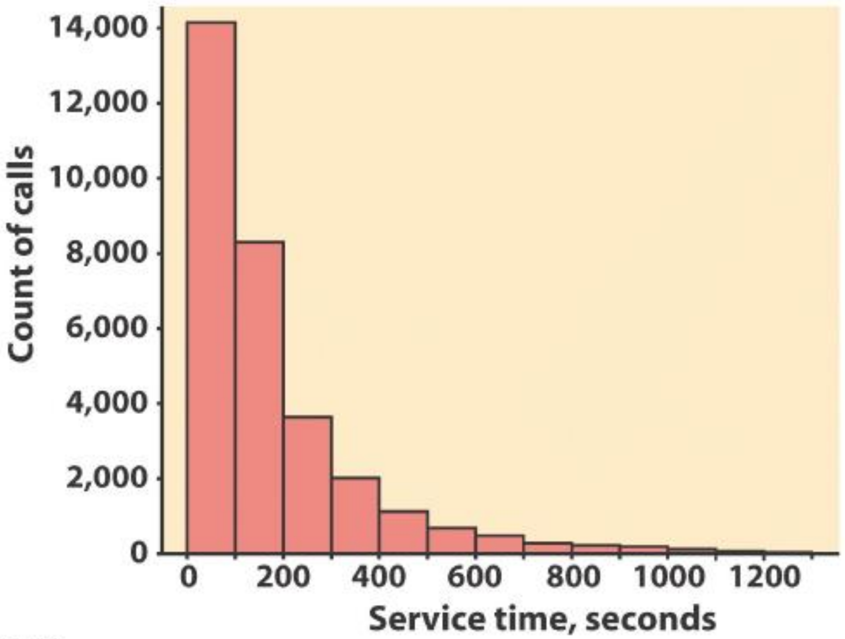


Figure 1-6
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

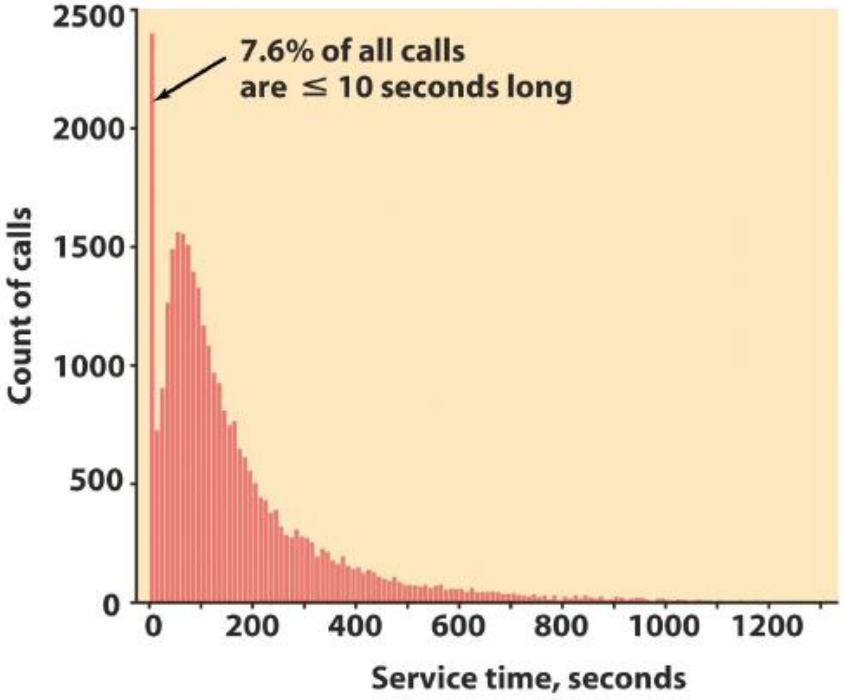
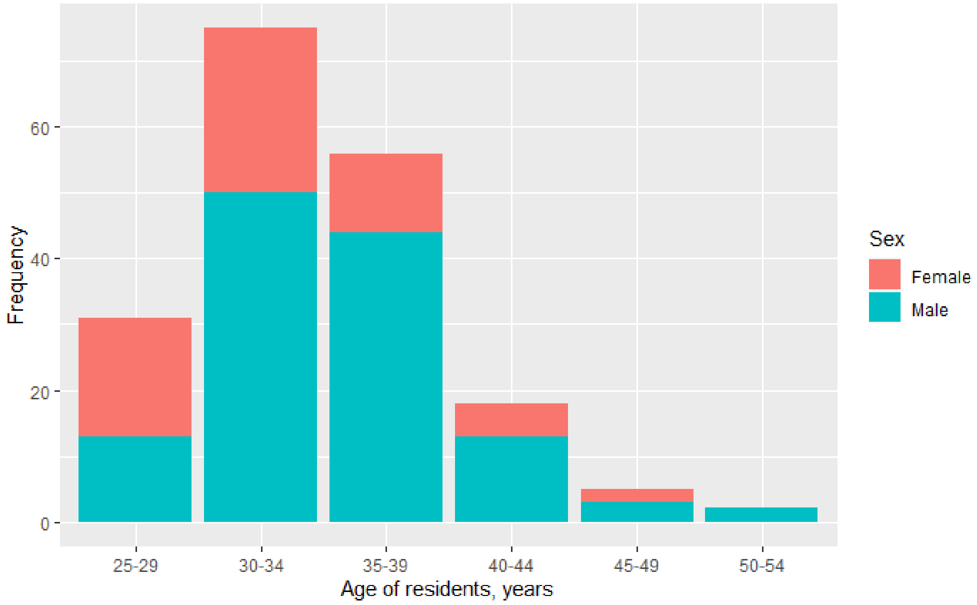


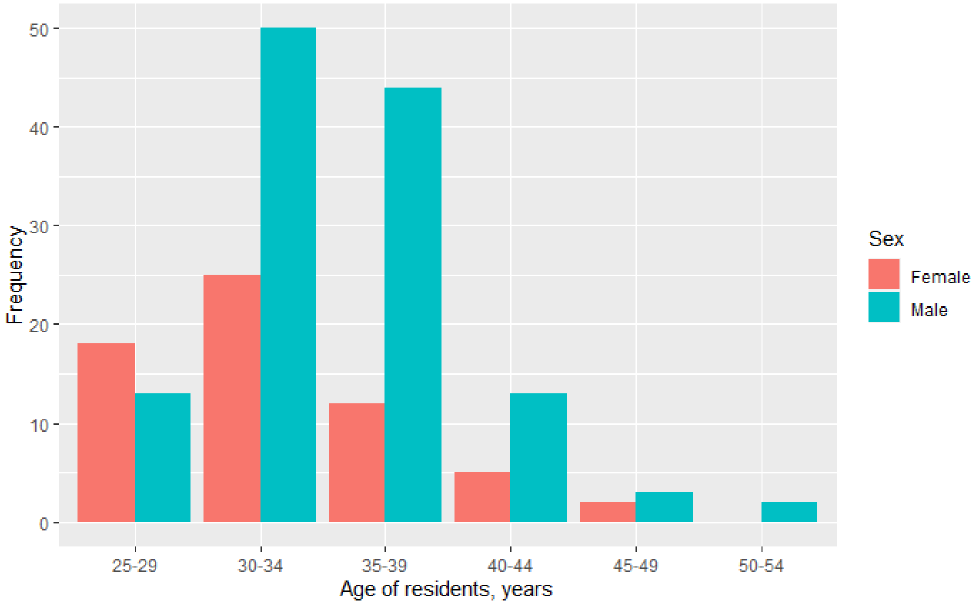
Figure 1-2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

The bar chart is also great for comparisons.

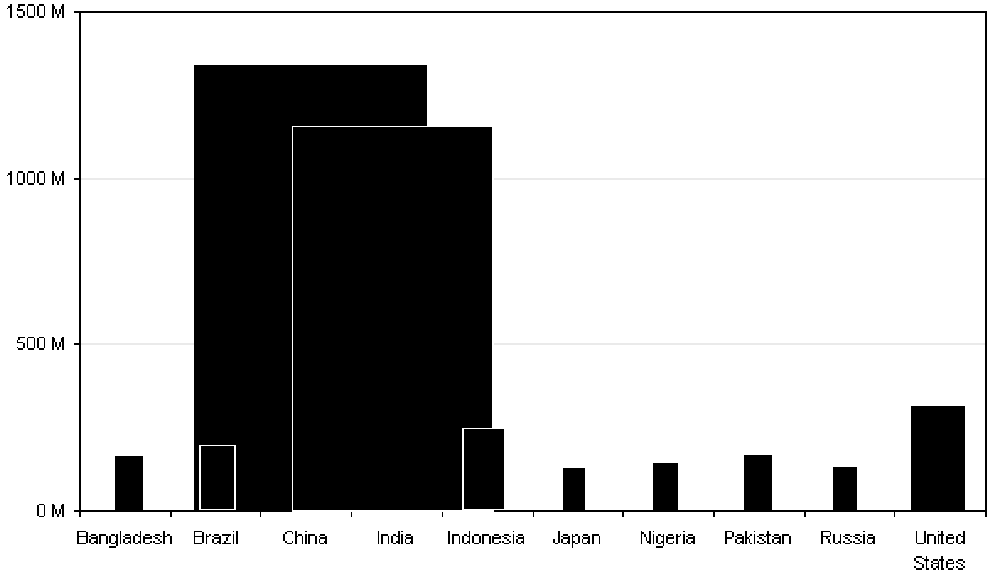
Stacked bar chart

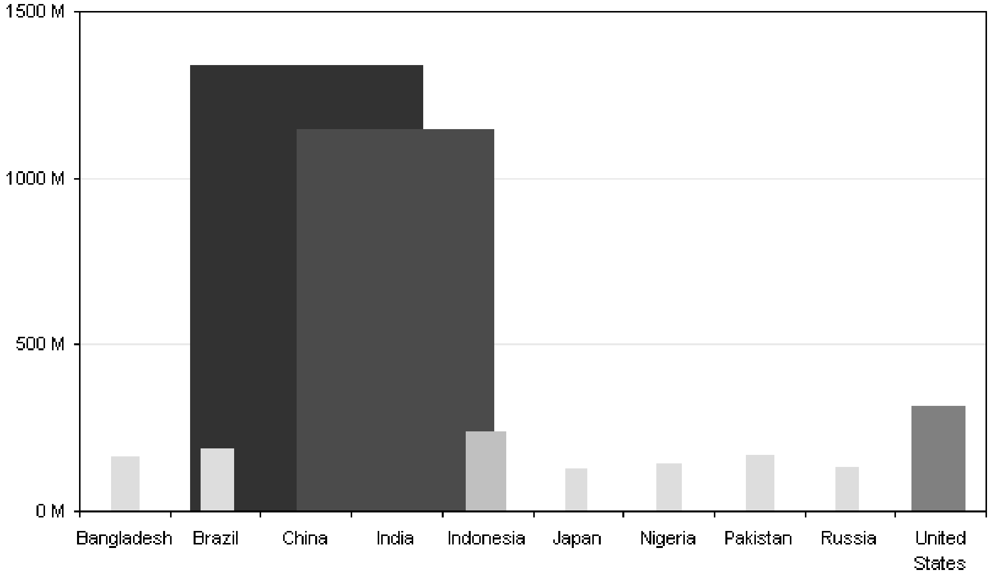


Side-by-side



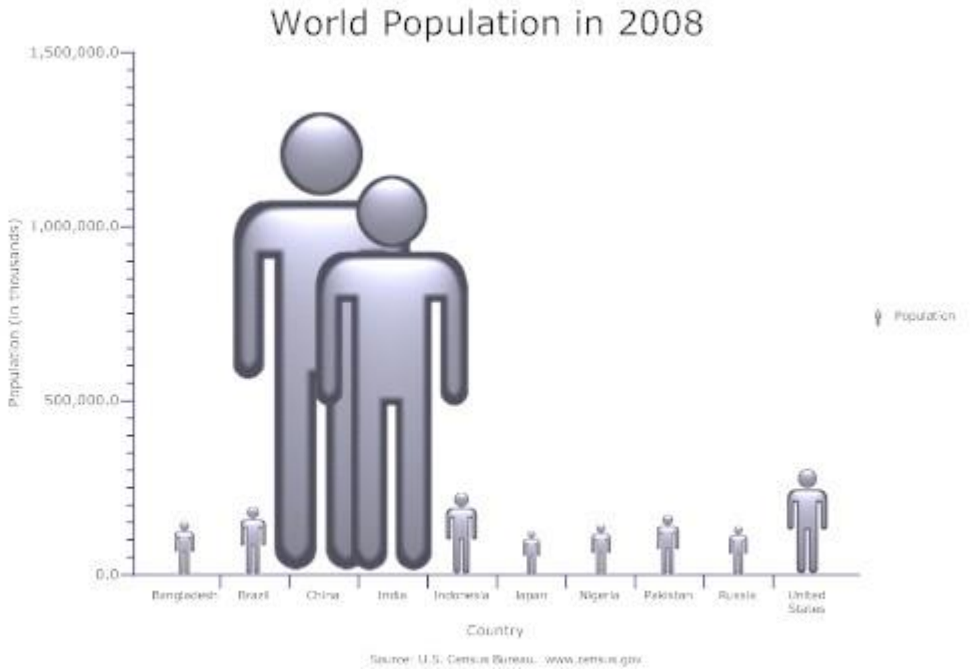
Let’s look at a few bad graphs

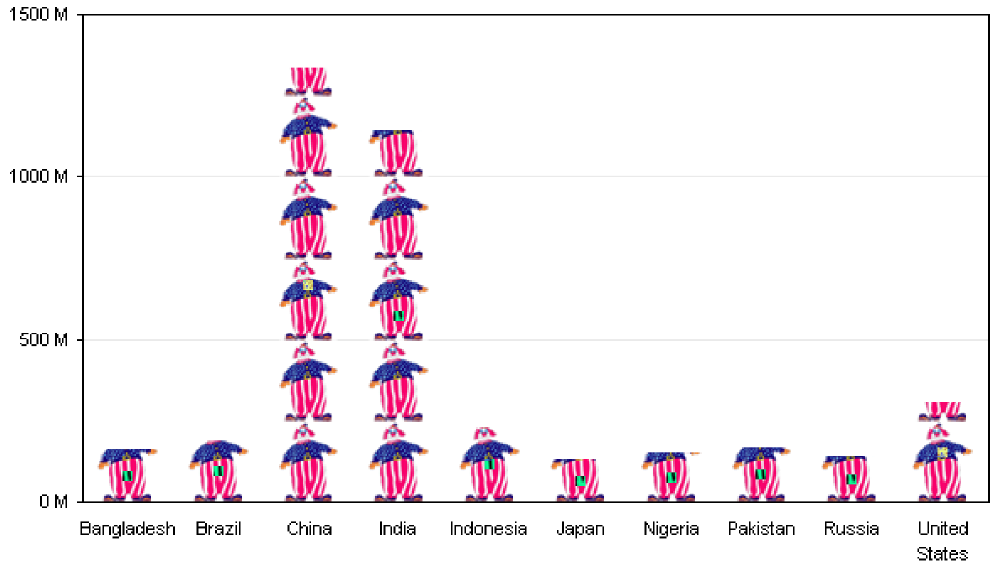




Source: https://peltiertech.Com/bad-bar-chart-practices-or-send-in-the-clowns/

Let’s look at a few bad graphs

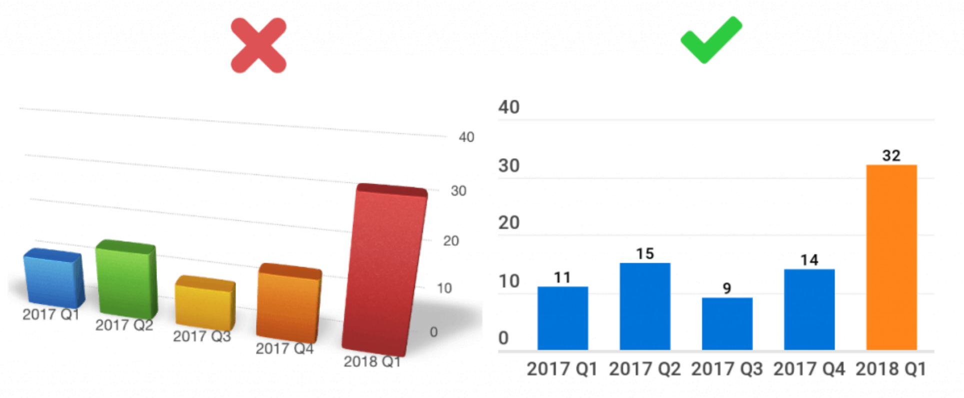


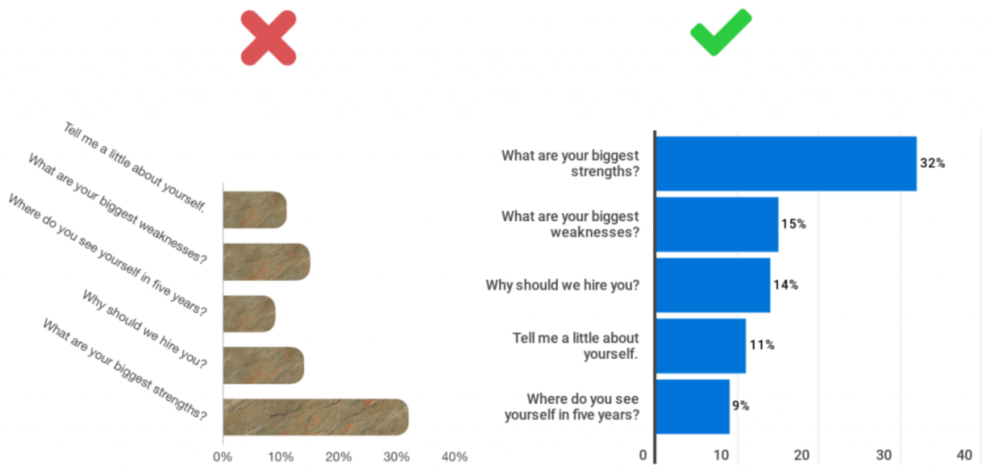


Look at this axis and the images used instead of bars!

Source: https://peltiertech.Com/bad-bar-chart-practices-or-send-in-the-clowns/

Let’s look at a few bad graphs

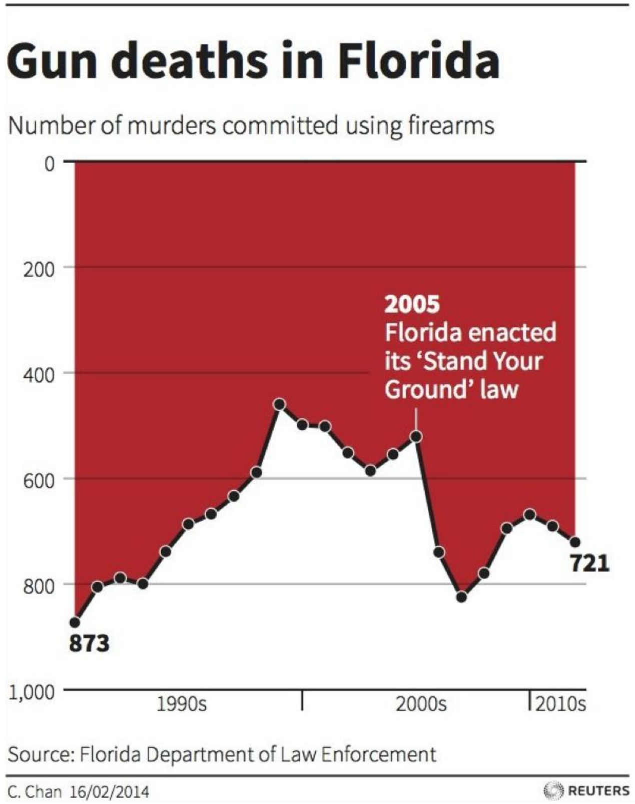


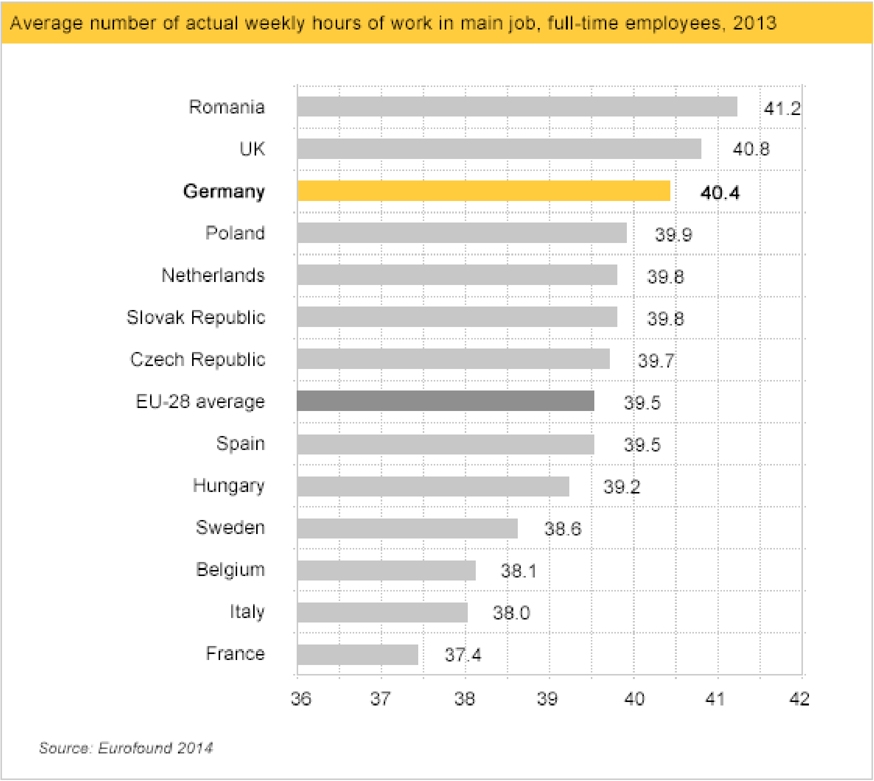


Careful when you use colors and labels. It should be clear what they mean!

Source: https://peltiertech.Com/bad-bar-chart-practices-or-send-in-the-clowns/

Let’s look at a few bad graphs



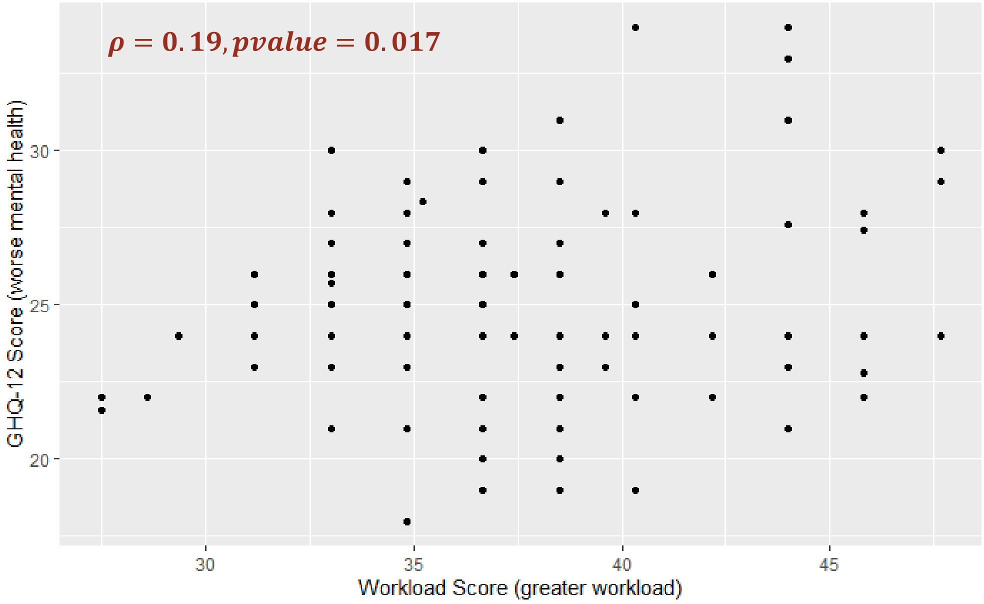


Watch out for the story the makers of the graphs are trying to tell. Be a skeptic! Also, please don’t invert the axes! Start the axes from 0 as often as possible!

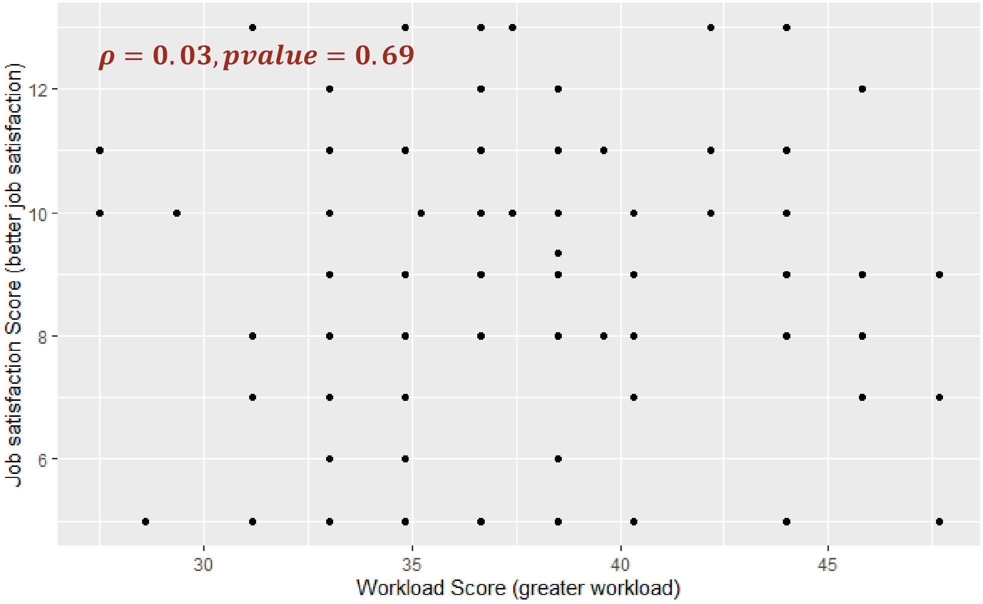
Source: https://www.callingbullshit.org/tools/tools_misleading_axes.html

The scatter plot is useful for visualizing the relationship of two continuous variables.

Positive correlation



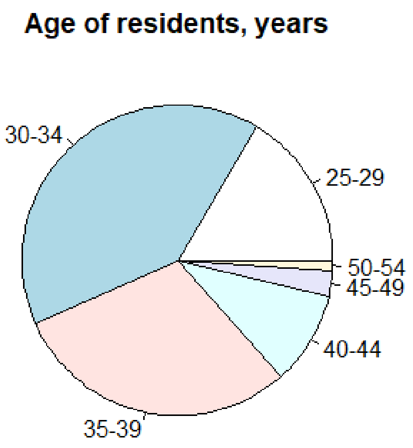
No correlation



The correlation summarizes the same information as the scatter plot but numerically.

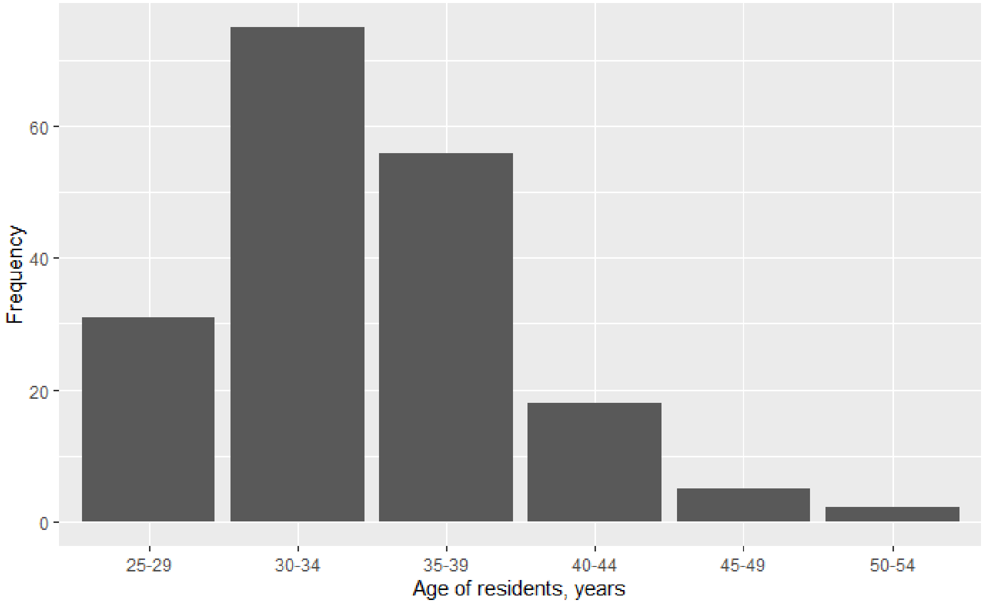
Pie charts are the worst! Avoid them!

Pie chart

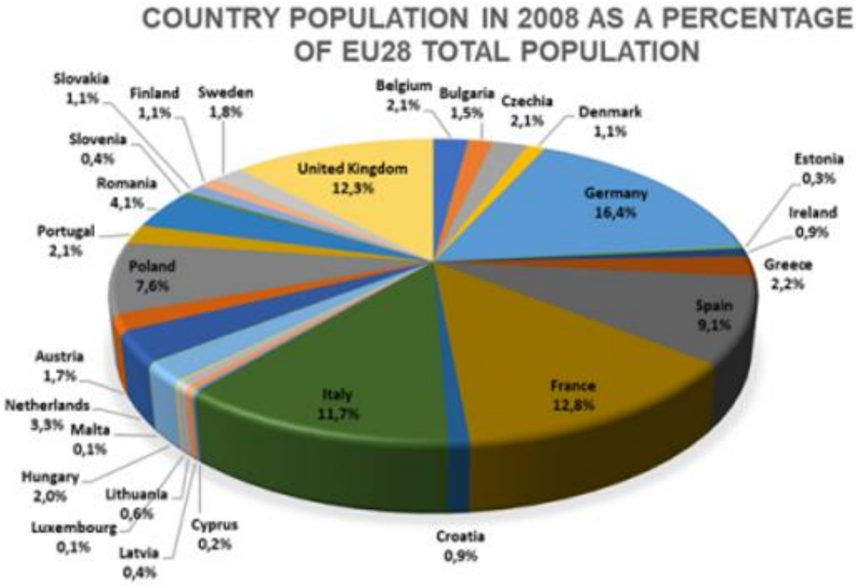


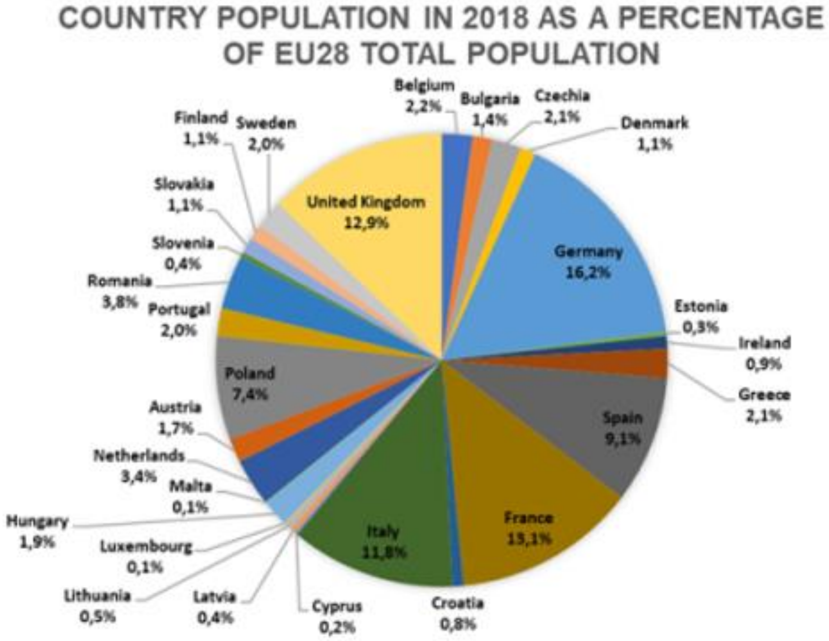
Its easier to see the different sizes of bars than to see the different sizes of pies. Especially if the pies are >3. People who are color-blind cannot read pie charts.

Bar chart

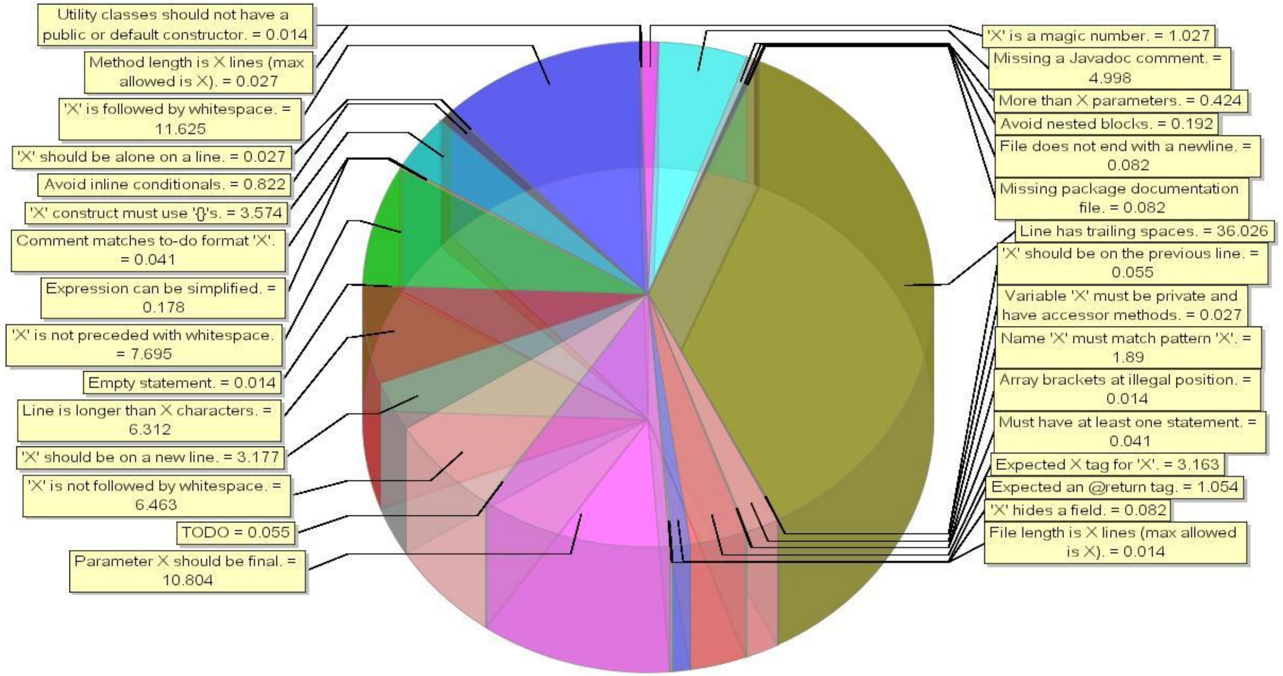


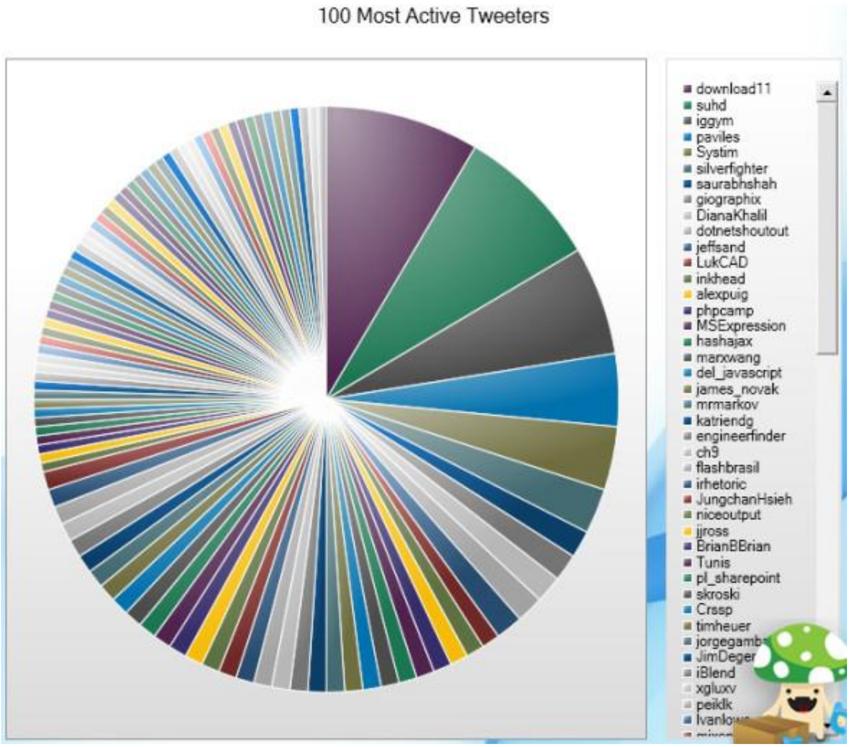
Let’s look at a few bad graphs





Let’s look at a few bad graphs





Source: https://visuanalyze.wordpress.com/2013/02/05/bad_piecharts/

WRAP-UP

What are your key takeaways from today’s class?

- Can you classify variables?
- Can you prepare data tables?
- Can you calculate basic measures of central tendencies and summarize frequency distributions?
- Can you use a graph to tell your story?

SUMMARY

- Health-related variables can be continuous, discrete or categorical. Categorical variables can be nominal or ordinal.
- A measure of central tendency is a central or typical value in a data distribution, and the most common measures are the arithmetic mean, median and mode
- The mean is the most sensitive to outliers. There can be more than one mode
- A ratio is the relative magnitude of two quantities which can be unrelated
- A proportion is a ratio in which the numerator is a part of the denominator
- A rate is a ratio in which the denominator is a time measure
- Axes and colors should be carefully thought through in graphical displays
- Scatter plots provide a good visual assessment of correlation between two variables
- Beware of pie charts, particularly when dealing with more than 3 variables
- A histogram is a graphical summary for continuous variables and is sensitive to bin sizes

- A histogram is a graphical summary for continuous variables and is sensitive to bin sizes

- A bar graph is a graphical summary for categorical variables

ACKNOWLEDGEMENT

These slides draw considerably from the following sources:

- Professors' classnotes
  - David Harrington, Harvard
  - Murray Mittleman, Harvard
- Textbook(s)
  - Medical Statistics by Betty Kirkwood and Jonathan Sterne
  - Biostatistics and epidemiology: a primer for health and biomedical professionals by Wassertheil-Smoller S and Smoller J
- Websites
  - Boston University School of Public Health website: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/Menu/index.html
  - Statistical tools for high-throughput data analysis (STHDA): http://www.sthda.com/english/
  - Penn State University's Statistics course websites https://online.stat.psu.edu/stat507/lesson/12/12.3