Effective visualizations and the nature of data

STA 198: Introduction to Health Data Science

Yue Jiang

May 20, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

Announcements

- HW 01 released today (Wednesday)
- Lab 01 released tomorrow (Thursday - note: no corresponding live session)
- HW/Lab 01 due May 26 via Gradescope

Exploratory data analysis (EDA)

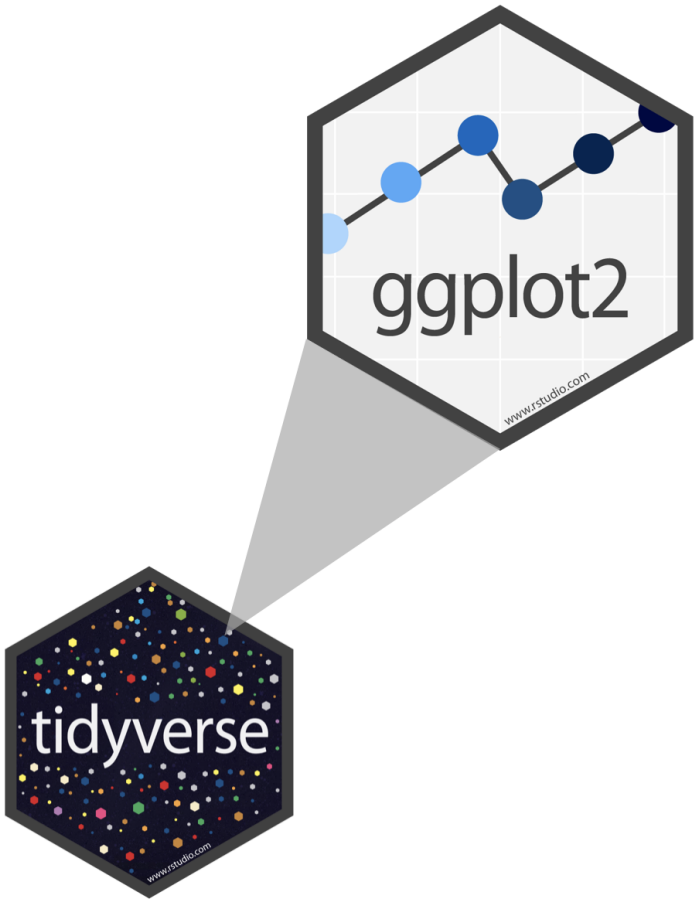
- Initial data analysis approach that summarizes main characteristics of dataset
- Often visual or in the form of basic summary statistics

Data visualization

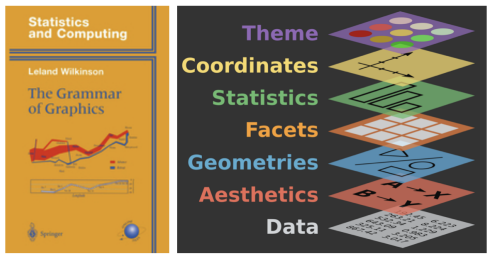
- The creation and study of the visual representation of data
- Many tools available (R is popular; many systems within R for data visualization)
- Creating visualizations helps us see patterns and identify potential data quality issues
- We will focus on ggplot2, a component of the tidyverse

"The simple graph has brought more information to the data analyst's mind than any other device." – John Tukey

ggplot2 in tidyverse

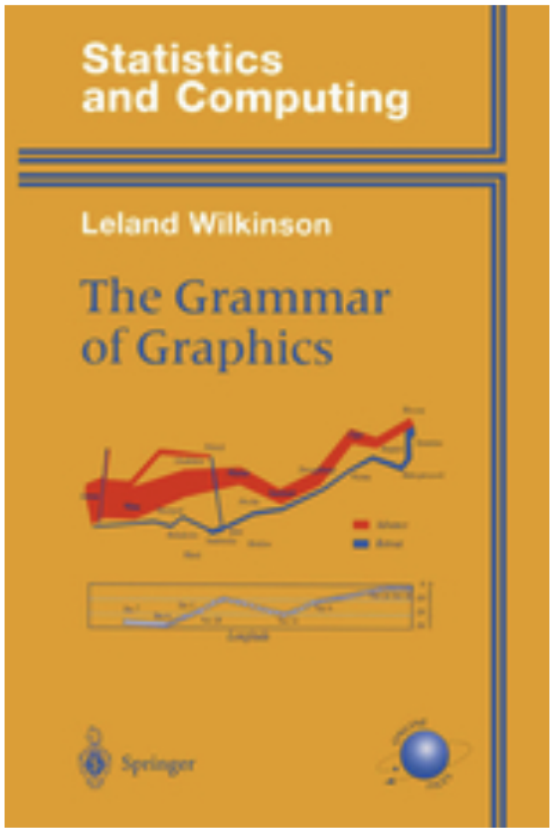


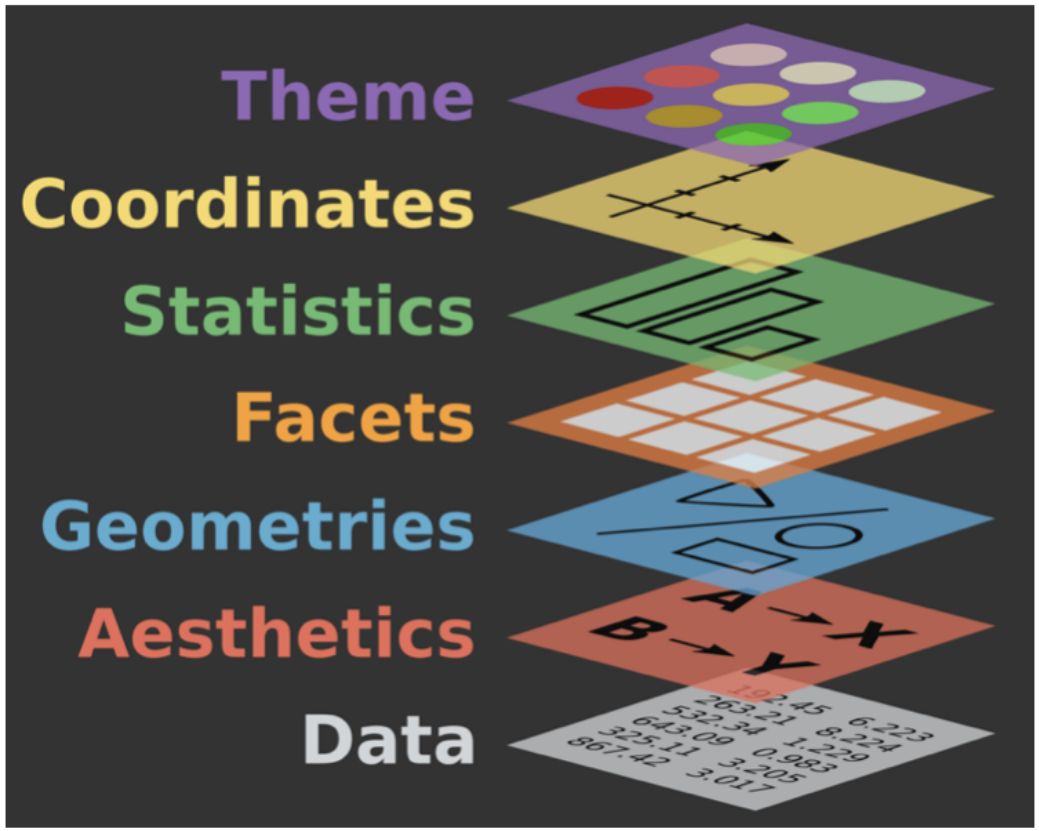
- The tidyverse is a group of R packages designed for data science
- All tidyverse packages share an underlying design philosophy
- Data visualization package in the tidyverse
- Inspired by *The Grammar of Graphics* (Wilkinson)



What is a *Grammar of Graphics*?

A system allowing for concise description of graphical components





What is a *Grammar of Graphics*?

A statistical graphic is

- data (which may be statistically summarized or transformed)
- mapped to aesthetic attributes (color, size, xy-position, etc.)
- using geometries (points, lines, bars, etc.)
- mapped onto a specific coordinate and/or facet system

The pre-recorded guided activity for Lab 01 on Thursday will walk you through creating some basic visualizations using this grammar (note: there is no corresponding live session for this lab). You will be creating visualizations to turn in as part of both HW 01 and Lab 01.

Categorical data

Nominal data

- Named categories without numeric meaning
- Only two categories: binary or dichotomous
- Breast cancer status, blood type, health insurance provider type, etc.

Ordinal data

- Ordered categories, but differences between values not easily measured
- Relative comparisons made about differences between levels
- Stage of colon cancer, Likert scale, frequency of smoking (often, sometimes, rarely, never), etc.

Numerical data

Count or rank data

- Discrete...counts. Or ranks (fairly self-explanatory)
- Number of alcoholic drinks consumed in past week, numerical rank of cancers by mortality, etc.

Continuous data

- Measureable quantities where difference between possible values can be arbitrarily small
- Data may lie within a range or be unbounded
- Birth weight, BMI, ppm ozone, etc.

Identifying data types

Table 1. Baseline Demographic and Clinical Characteristics, Stratified by Category of Chronic Kidney Disease (CKD), in the ACCORD Trial

Characteristics	No CKD $n = 6410$	Low GFR $n = 953$; Mean (SD)	Microalbuminuria $n = 2206$; Mean (SD)	Macroalbuminuria $n = 669$; Mean (SD)	*P*-Value
Age, y	62.0 (6.3)	67.5 (6.5)	62.8 (6.9)	62.9 (6.8)	<0.001
Male Sex, *n* (%)	3851 (60)	500 (53)	1515 (69)	426 (64)	<0.001
Race/ethnicity, *n* (%)					<0.001
White	4034 (63)	652 (68)	1330 (60)	372 (56)	
Black	1185 (19)	136 (14)	473 (21)	153 (23)	
Hispanic	451 (7)	66 (7)	169 (8)	50 (8)	
Other	740 (12)	99 (10)	234 (11)	94 (14)	
Log (UACR), μg/mg	2.2 (0.6)	3.0 (1.2)	4.3 (0.6)	6.6 (0.7)	<0.001
Creatinine, μg/mg	0.86 (0.18)	1.30 (0.20)	0.88 (0.19)	1.01 (0.30)	<0.001
eGFR*, mL/min/1.73 m^2	88 (14)	52 (7)	88 (15)	78 (20)	<0.001
BMI	32.1 (5.3)	32.2 (5.4)	32.5 (5.6)	32.3 (5.6)	0.02
DM duration, y	9.9 (7.2)	13.1 (8.5)	11.6 (7.7)	13.9 (8.0)	<0.001
HbA1c	8.22 (1.01)	8.25 (1.06)	8.47 (1.10)	8.60 (1.20)	<0.001
SBP, mm Hg	133.7 (16.0)	137.0 (17.6)	140.2 (17.2)	148.5 (19.0)	<0.001
DBP, mm Hg	74.9 (10.3)	71.5 (11.1)	75.8 (10.7)	76.8 (11.9)	<0.001
History of CVD, *n* (%)	2008 (31)	411 (43)	864 (39)	320 (48)	<0.001
ACEI use	3334 (52)	561 (59)	1266 (57)	401 (60)	<0.001
ARB use	1002 (16)	188 (20)	383 (17)	138 (21)	0.007

Abbreviations: ACEI, Angiotensin Converting Enzyme Inhibitor; ARB, angiotensin-receptor blocker; BMI, body mass index; CVD, cardiovascular disease; DBP, diastolic blood pressure; DM, diabetes mellitus; eGFR, estimated glomerular filtration rate; GFR, glomerular filtration rate; HbA1c, hemoglobin A1c; SBP, systolic blood pressure; SD, standard deviation; UACR, urinary albumin:creatinine ratio.
*eGFR calculated using the CKD-Epi Formula.

Complications

In designing a study, what variable should we use for smoking exposure?

- Binary variable yes/no?
- Ordinal current/former/never smoker?
- Discrete number of cigarettes smoked in past week?
- Continuous measurement of lifetime pack-years?

In the real world, decisions are made based on sample size, statistical power, likelihood of measurement error, or simply convenience (this...happens a lot)

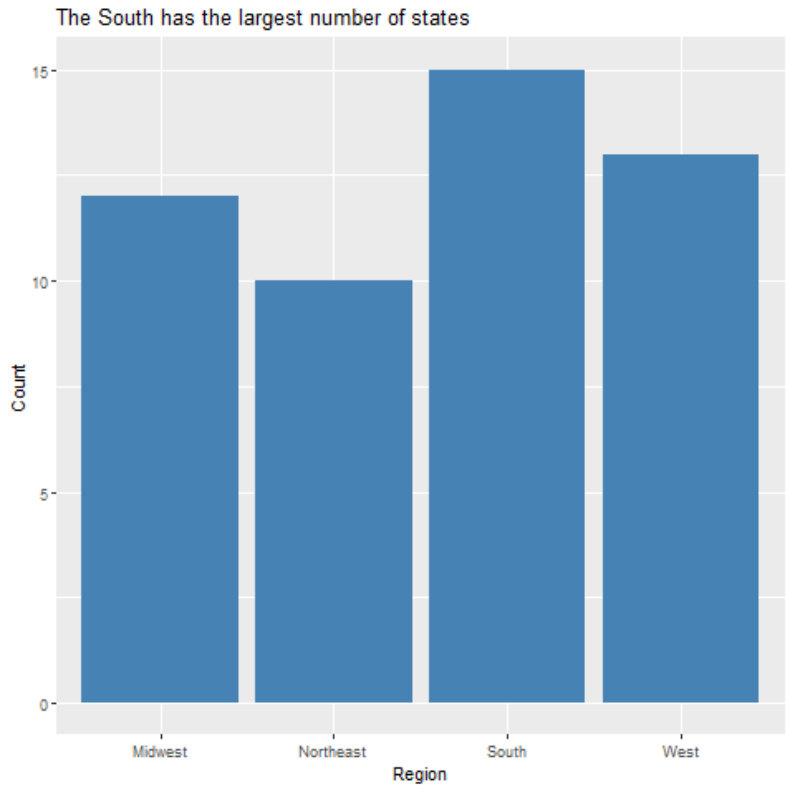
Visualizing CDC data

Let's take a look at some basic visualizations using state-level data collected by the CDC. We'll examine the following variables:

- State (categorical; nominal)
- HDI (categorical; ordinal)
- Region (categorical; nominal)
- Adult obesity % (numerical; continuous)
- Adequate aerobic activity % (numerical; continuous)

Bar charts

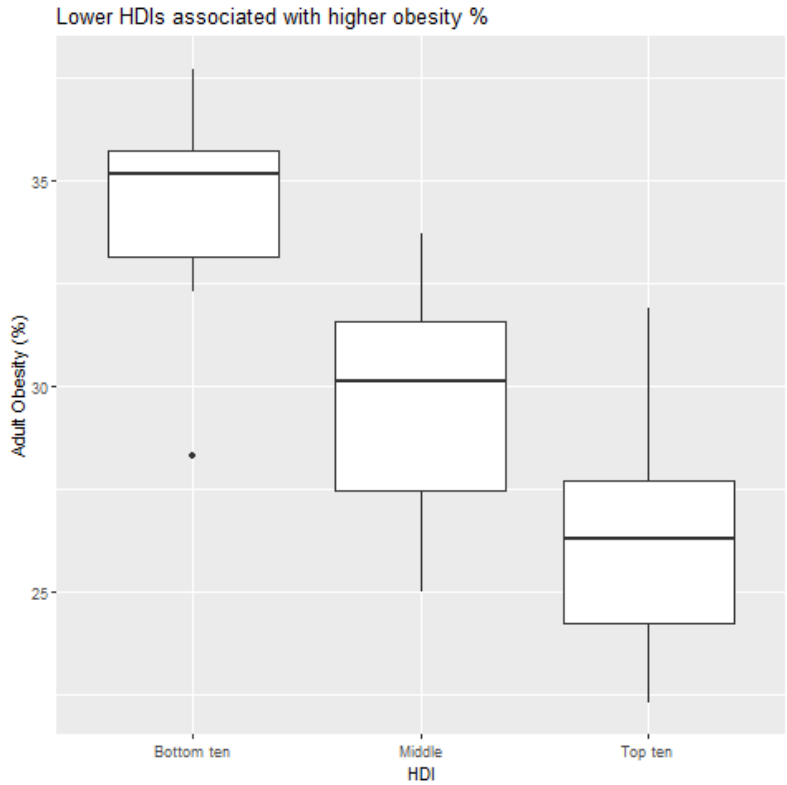
- Summarizes numerical variable by categories
- Visually depict frequency distributions for nominal or ordinal data
- Bars represent either frequency or relative frequency by category
- Separation between bars (non-continuous data)
- May contain error bars to indicate estimate variability



Box plots

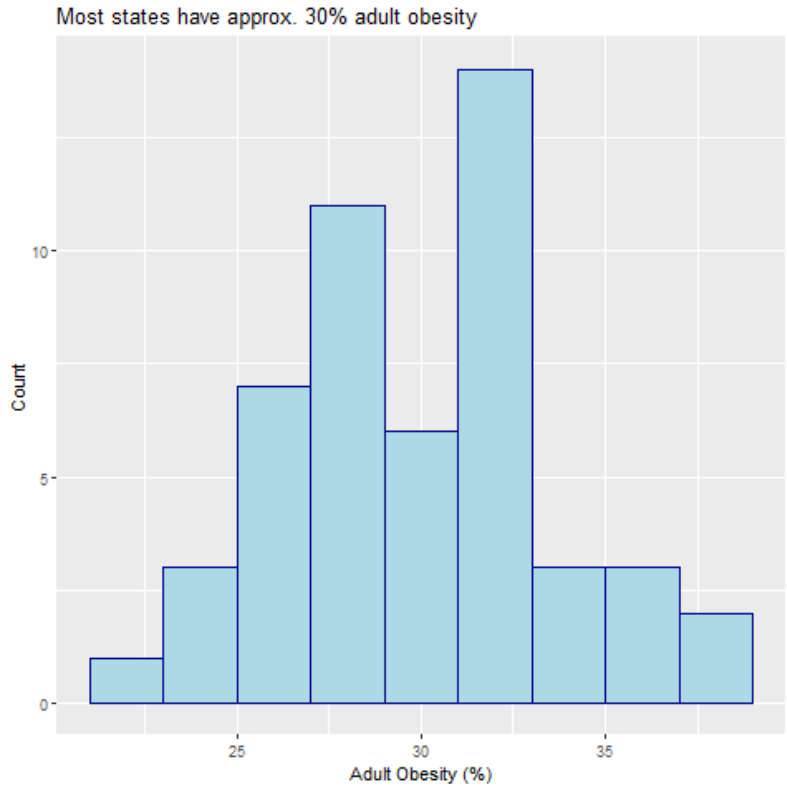
- Summarizes numerical variable
- Five-number summary: sample minimum, 25^{th} percentile, median, 75^{th} percentile, sample maximum
- Outliers
- Spread and skew

(more on all of these later)



Histograms

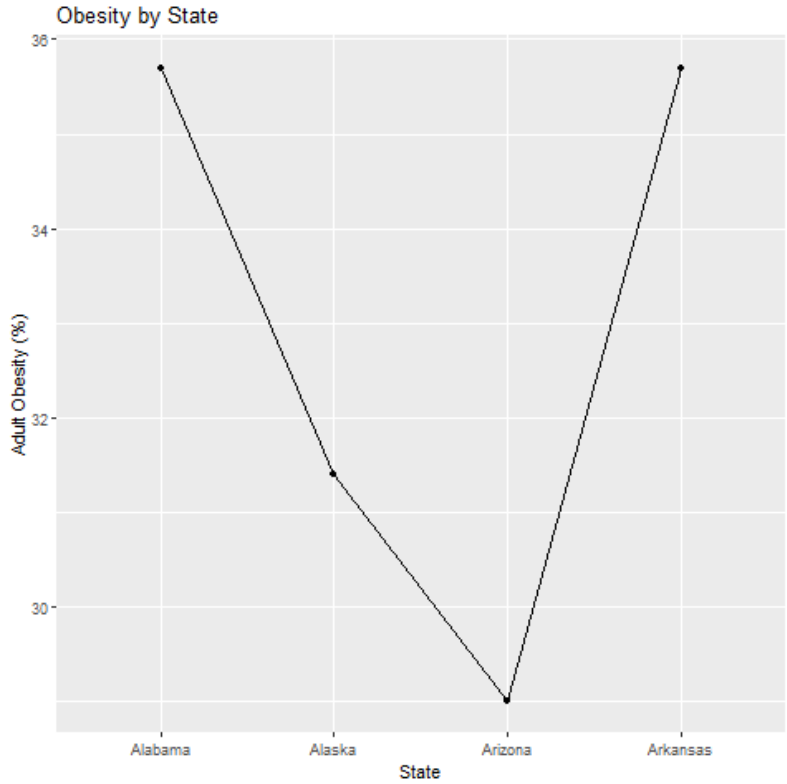
- Summarizes numerical variable
- Frequency distribution for discrete or continuous numerical data
- Each bar is proportional to frequency of the categories



Line plots

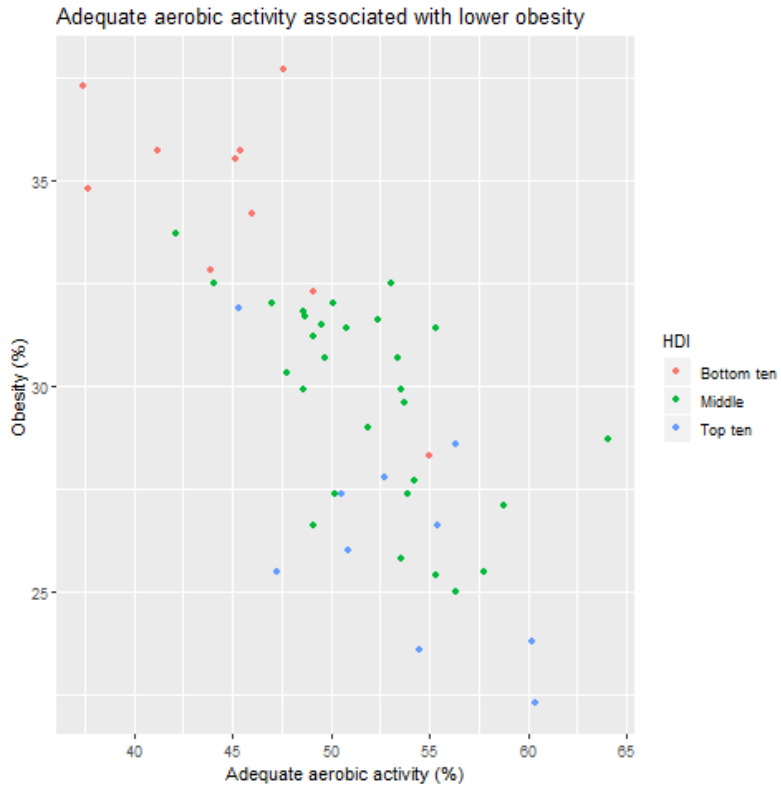
- Summarizes numerical variable (most often used through time)
- Each value on one axis corresponds to only one measurement on the other
- Often used to depict change over time and connected with line

Is the line graph displayed on the right useful?



Scatterplots

- Summarizes two numerical variables (can be extended)
- Depicts relationship between multiple continuous measurements
- Can add color, shape, transparency, etc. to further differentiate by category

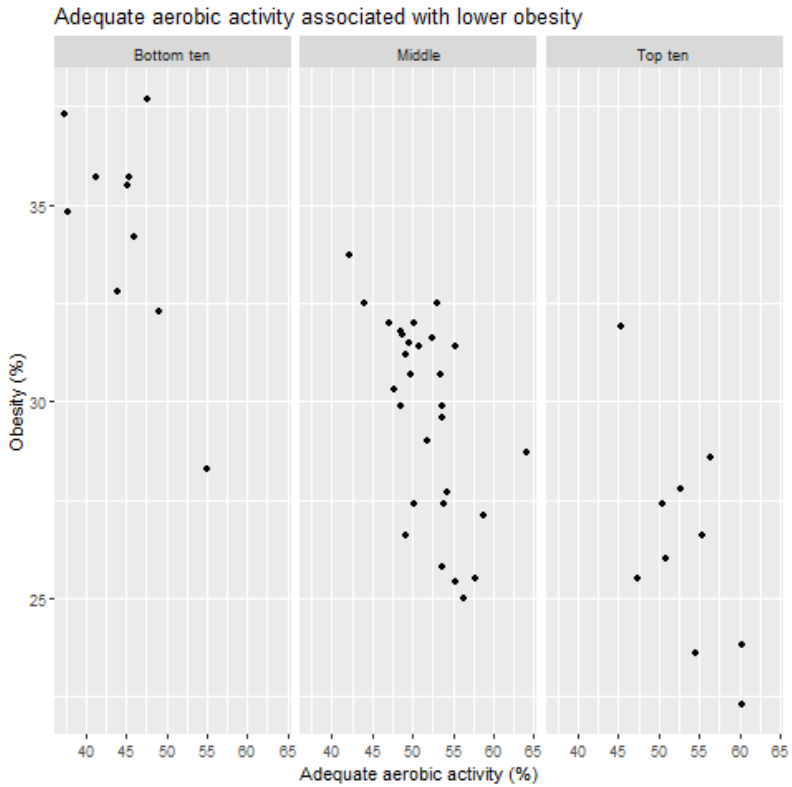


Some best practices

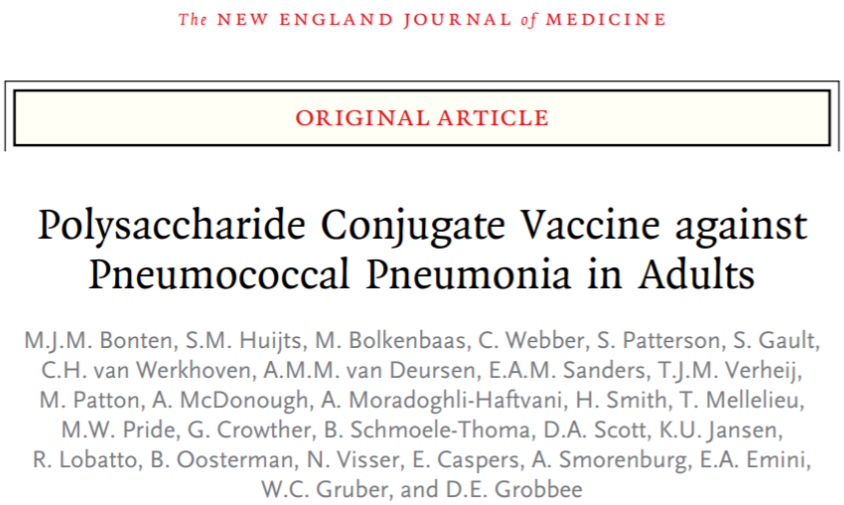
- Keep it simple
- Summarize and highlight
- Tell a story with the plot (use "active titles")
- If possible, replace text with visuals

Scatterplots

- Depicts relationship between multiple continuous measurements
- Can add color, shape, transparency, etc. to further differentiate by category



Reminder: the population vs. a sample



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Polysaccharide Conjugate Vaccine against Pneumococcal Pneumonia in Adults

M.J.M. Bonten, S.M. Huijts, M. Bolkenbaas, C. Webber, S. Patterson, S. Gault, C.H. van Werkhoven, A.M.M. van Deursen, E.A.M. Sanders, T.J.M. Verheij, M. Patton, A. McDonough, A. Moradoghli-Haftvani, H. Smith, T. Mellelieu, M.W. Pride, G. Crowther, B. Schmoele-Thoma, D.A. Scott, K.U. Jansen, R. Lobatto, B. Oosterman, N. Visser, E. Caspers, A. Smorenburg, E.A. Emini, W.C. Gruber, and D.E. Grobbee

- Population and research question: Is the PCV13 vaccine effective against community-acquired pneumonia in adults aged 65 or older?
- Sample: 84,496 adults 65 years of age or older recruited in a trial between September 2008, and January 2010 at 101 sites throughout the Netherlands

Parameters and statistics

Statistics

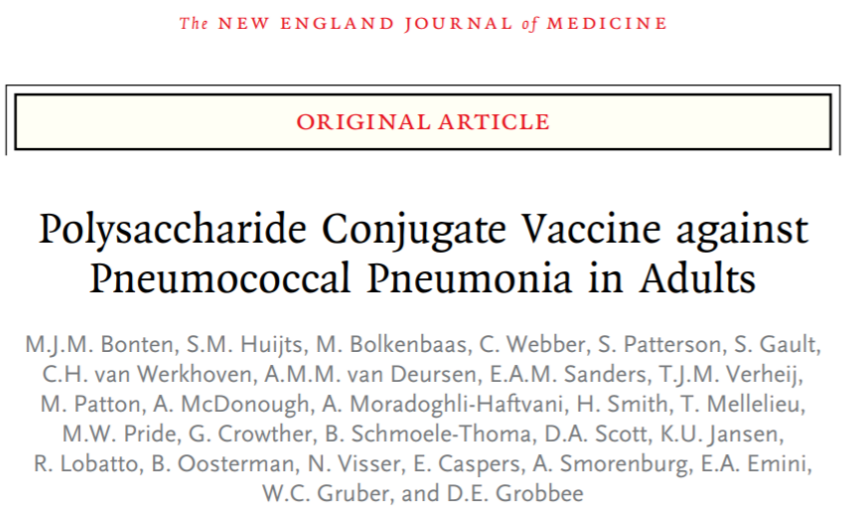
- Attribute of a sample
- Function of the observed values at hand
- Confusingly, both the function and the values (sorry)
- Written in Roman letters

Parameters

- Attribute of the population of interest
- Not computable directly (unless entire population is perfectly measured)
- Written in Greek letters

A statistic used to estimate a population parameter is an estimator

Numerical summary statistics



The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Polysaccharide Conjugate Vaccine against Pneumococcal Pneumonia in Adults

M.J.M. Bonten, S.M. Huijts, M. Bolkenbaas, C. Webber, S. Patterson, S. Gault, C.H. van Werkhoven, A.M.M. van Deursen, E.A.M. Sanders, T.J.M. Verheij, M. Patton, A. McDonough, A. Moradoghli-Haftvani, H. Smith, T. Mellelieu, M.W. Pride, G. Crowther, B. Schmoele-Thoma, D.A. Scott, K.U. Jansen, R. Lobatto, B. Oosterman, N. Visser, E. Caspers, A. Smorenburg, E.A. Emini, W.C. Gruber, and D.E. Grobbee

- Population parameter of interest: vaccine efficacy among all adults aged 65 or older
- Sample statistic collected: proportion of vaccinated adults in the trial who became ill with community-acquired pneumonia

Mean

- Sample mean: the arithmetic average of values in the sample:

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Population mean μ is calculated the same way, but would involve sum over every observation in the population (rarely feasible)
- The sample mean is a point estimate of the population mean
- Not the exact population mean (unless lucky), but for a representative sample, it's a pretty good guess
- As the sample size gets larger, on average $\bar{x}$ gets closer and closer to μ

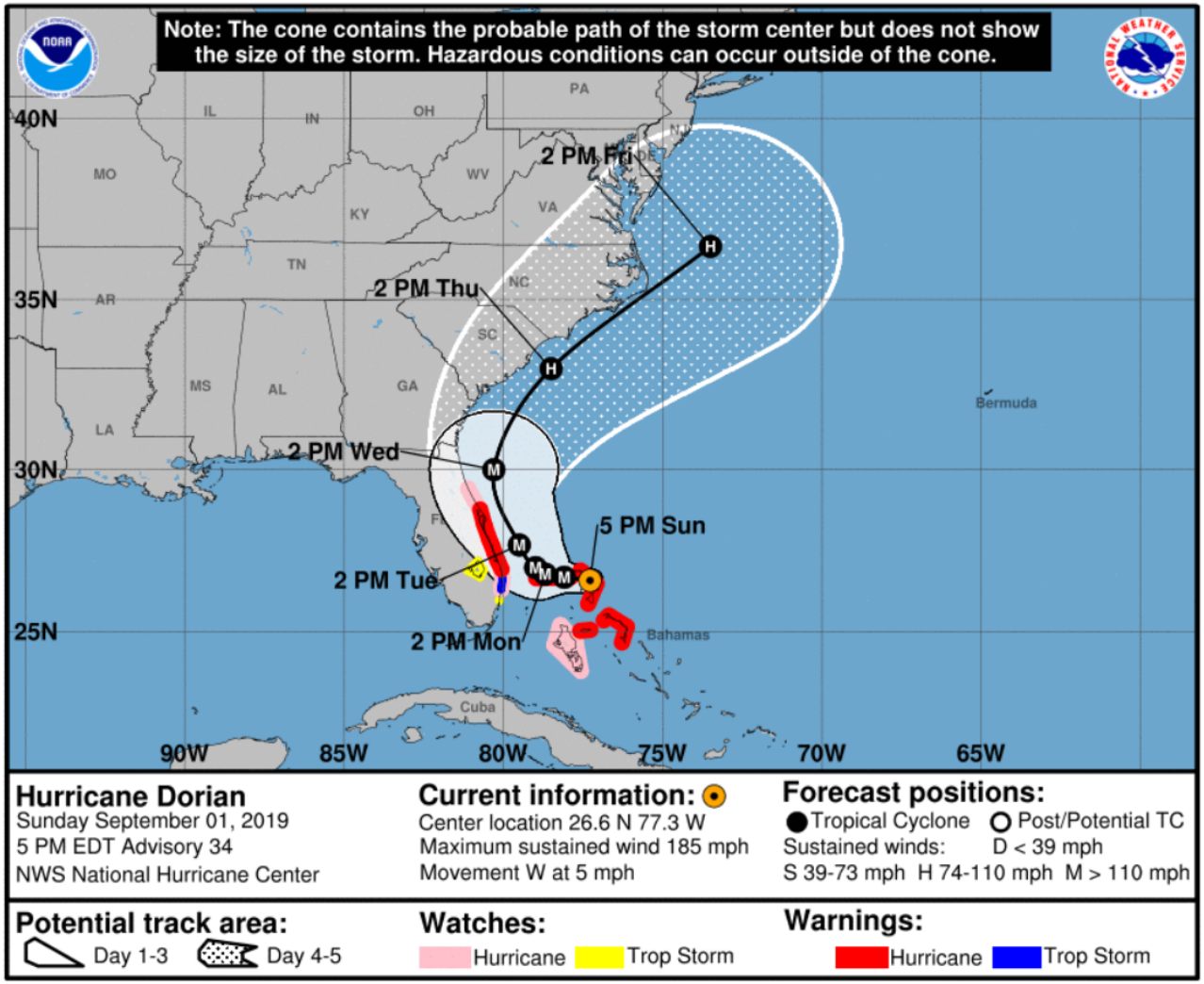
Median

- Sample median: the 50^{th} percentile
- Middle number of observations are ranked in numerical order
- For odd number of observations, it is the exact middle value; otherwise, it is the arithmetic average of the middle two
- More robust to extreme values or outliers when compared to the mean

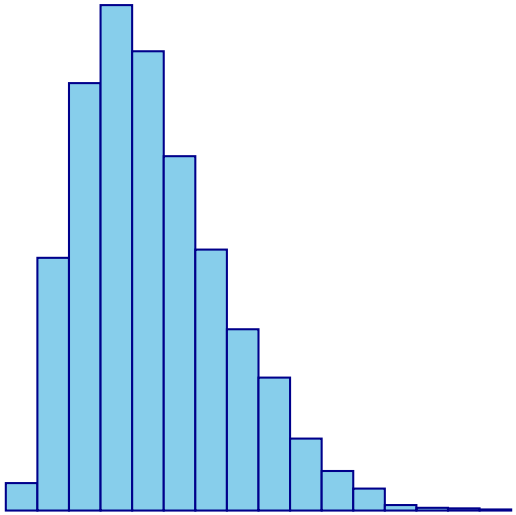
Mode

- Sample mode: the most frequent value in the dataset
- There does not have to be only one mode (we can have bimodal or trimodal or other multimodal distributions)

Are point estimates of location enough?



Skewness



Right-skewed distribution

Skewed distributions are not symmetric; they can be right or left skewed depending on which side the "tail" is on.

Minimum, maximum, and range

- Sample minimum and maximum: the smallest and largest observations in the dataset
- Sample range: the difference between the sample maximum and minimum

Quantiles

- Cutpoints dividing data into equal-sized groups (tertiles, quartiles, quintiles, percentiles, etc.)
- First quartile (Q1) and third quartile (Q3) cut off the bottom and top 25%, resepectively
- Interquartile range (IQR): Q3 - Q1; shows the width of the middle 50% of the data
- The sample minimum, Q1, Q2 (median), Q3, and maximum are sometimes called the five number summary

Outliers

- Observations numerically distant from others (definitions vary)
- Statistical methods robust to outliers (e.g., the median) can be used if outliers are problematic
- Should be noted and handled carefully! (e.g., maternal ages of 11 vs. 111 in a dataset)

Standard deviation

- Sample standard deviation: most common measure of spread, based on deviations around the mean

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Population SD σ is calculated the same way, but requires sum over everyone in the population (with $\bar{x}$ replaced by μ)
- Same units as original dataset for easier interpretation
- Often used to express confidence (e.g., a margin of error for a poll being around $\pm$ 2 SD of the mean)
- Squared deviations weight larger deviations more heavily, and also so positive and negative deviations do not cancel out

Variance

- Sample variance: approximately the average squared deviation from the mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Estimate of the population variance σ^2
- Division by $n-1$ instead of n to avoid bias in small samples (don't worry about this right now: more details in STA 240/432 for interested students)

How big are most values?

For a distribution of any shape, most of the data are within "average $\pm$ a few SDs."

Chebychev's inequality tells us the proportion of values in the range "average $\pm$ k SDs" is

$$\text{at least } 1 - \frac{1}{k^2}.$$

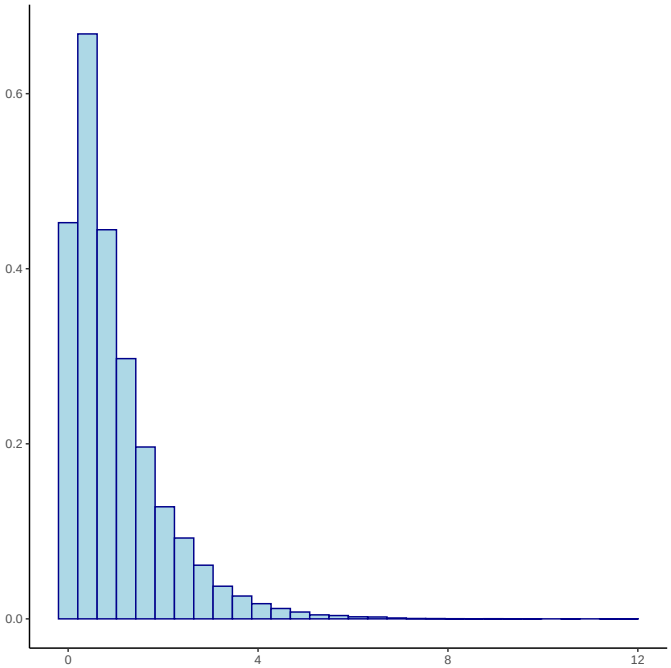
Chebychev's bounds

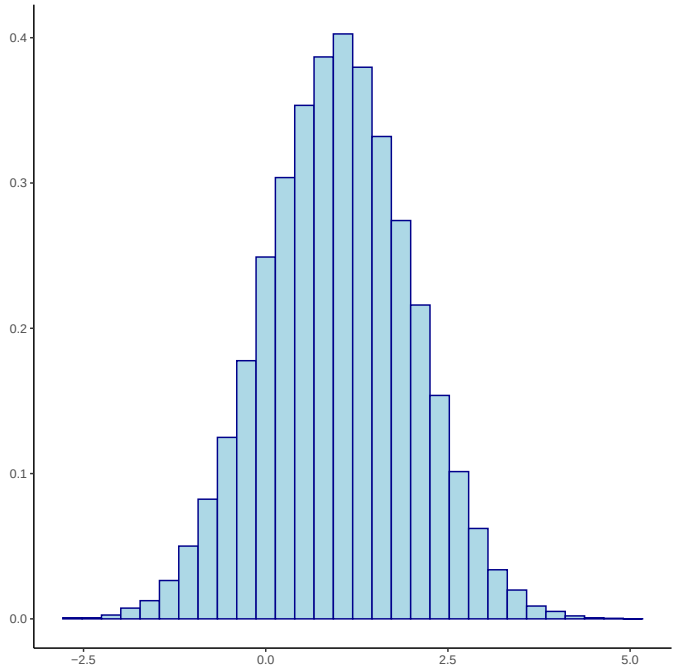
Range	*Proportion*
Average $\pm$ 2 SDs	at least $1 - \frac{1}{4} = 75\%$
Average $\pm$ 3 SDs	at least $1 - \frac{1}{9} = 89\%$
Average $\pm$ 4 SDs	at least $1 - \frac{1}{16} = 94\%$
Average $\pm$ 5 SDs	at least $1 - \frac{1}{25} = 96\%$

If we know the exact distribution (coming soon), we can often calculate better bounds

However, these bounds hold for *any* distribution (that has a well-defined mean and variance)

Why not always use means and SDs?





These two distributions have the same mean and standard deviation, but are clearly very different!