Probability and Statistical Inference

1. Introduction & exploratory data analysis

Sta 111 - Summer II 2017

Duke University, Department of Statistical Science

June 1, 2017

Course outline

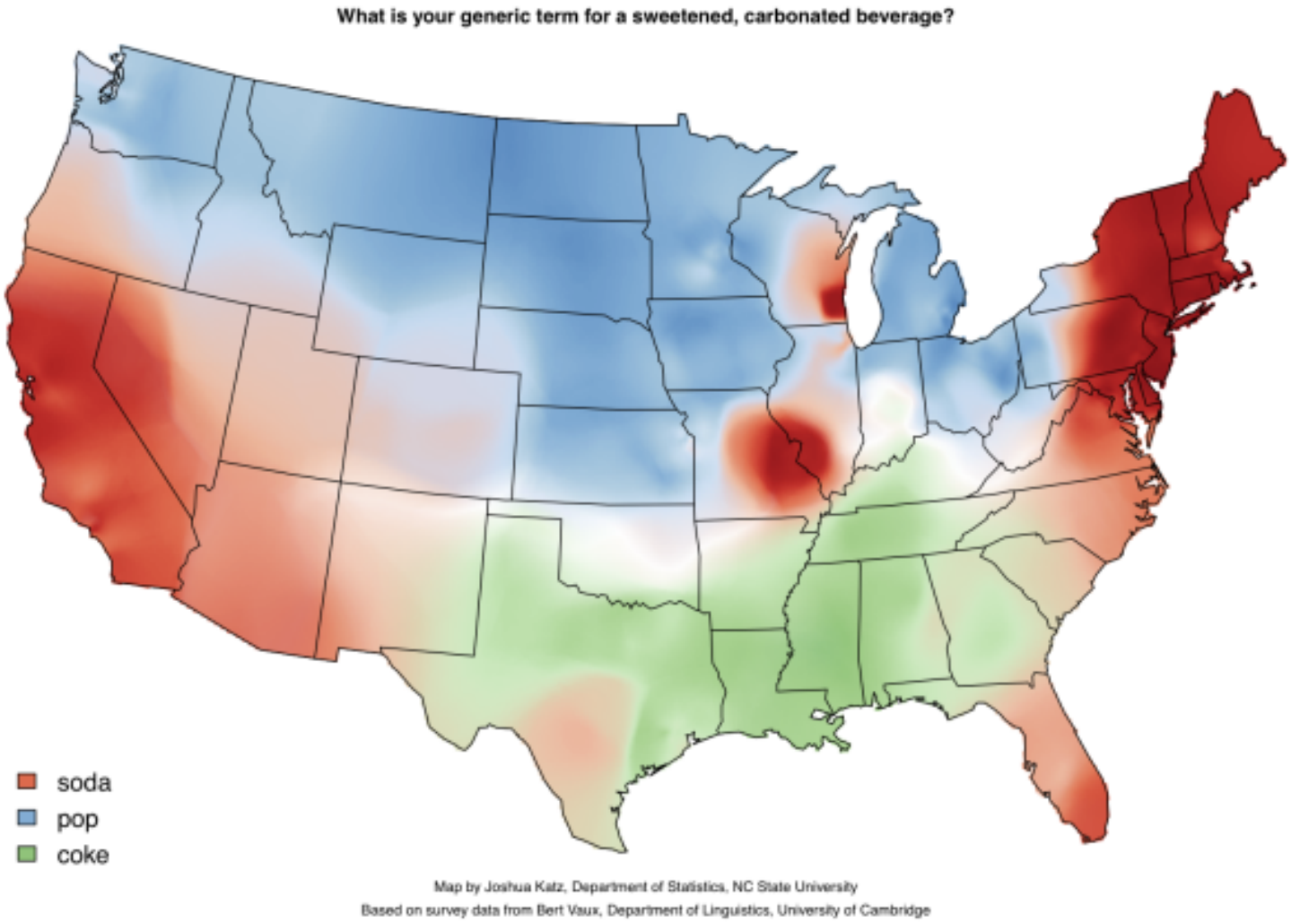
General information at *https://wangronglu.github.io/sta111*

- Mathematics behind statistics
  - *Ch 2-3 Probability & distributions:* Basics of probability, conditional probability, Bayes' theorem, binomial and normal distributions.
- Statistical inference
  - *Unit 3 - Framework for inference:* CLT, sampling distributions, and introduction to theoretical inference.
  - Midterm 1
  - *Unit 4 - Statistical inference for numerical variables*
  - *Unit 5 - Statistical inference for categorical variables*
  - Midterm 2
- Modeling
  - *Unit 6 - Simple linear regression:* Bivariate correlation and causality, introduction to modeling.
  - *Unit 7 - Multiple linear regression:* More advanced modeling with multiple predictors.
  - Final Exam

From a past Sta 101 survey...

Do you see anything out of the ordinary?

Describe the spatial distribution of preferred sweetened carbonated beverage drink.



http://spark.rstudio.com/jkatz/SurveyMaps

Describing distributions of numerical variables

- *Shape*: skewness, modality
- *Center*: an estimate of a *typical* observation in the distribution (mean, median, mode, etc.)
  - Notation: μ: population mean, $\bar{x}$: sample mean
- *Spread*: measure of variability in the distribution (standard deviation, IQR, range, etc.)
- *Unusual observations*: observations that stand out from the rest of the data that may be suspected outliers

Clicker question

Which of these is most likely to have a roughly symmetric distribution?

(a) salaries of a random sample of people from North Carolina
(b) weights of adult females
(c) scores on an well-designed exam
(d) last digits of phone numbers

Mean vs. median

Clicker question

How do the mean and median of the following two datasets compare?

Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

(a) $\bar{x}_1 = \bar{x}_2$, $median_1 = median_2$
(b) $\bar{x}_1 < \bar{x}_2$, $median_1 = median_2$
(c) $\bar{x}_1 < \bar{x}_2$, $median_1 < median_2$
(d) $\bar{x}_1 > \bar{x}_2$, $median_1 < median_2$
(e) $\bar{x}_1 > \bar{x}_2$, $median_1 = median_2$

Standard deviation and variance

- Most commonly used measure of variability is the *standard deviation*, which roughly measures the average deviation from the mean
  - Notation: σ: population standard deviation, s: sample standard deviation
- Calculating the standard deviation, for a population (rarely, if ever) and for a sample:

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

- Square of the standard deviation is called the *variance*.

Why divide by $n - 1$ instead of n when calculating the sample standard deviation?

Lose a "degree of freedom" for using an estimate (the sample mean, $\bar{x}$), in estimating the sample variance/standard deviation.

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

Range and IQR

Clicker question

True / False: The range is always at least as large as the IQR for a given dataset.

(a) Yes

(b) No

Is the range or the IQR more robust to outliers?

Robust statistics

- Mean and standard deviation are easily affected by extreme observations since the value of each data point contributes to their calculation.
- Median and IQR are more robust.
- Therefore we choose median&IQR (over mean&SD) when describing skewed distributions.

Box plot

A *box plot* visualizes the median, the quartiles, and suspected outliers. An *outlier* is defined as an observation more than $1.5 \times$ IQR away from the quartiles.