Data Analysis and Statistical Inference
Introduction

Sta 101 - Spring 2015

Duke University, Department of Statistical Science

January 7, 2015

Dr. Çetinkaya-Rundel

Slides posted at bitly.com/sta101sp15

Teaching team

- Professor: Dr. Mine Çetinkaya-Rundel - *mine@stat.duke.edu*
- TAs:
  - Anthony Weishampel
  - Radhika Anand
  - Jialiang Mao
  - Christine Chai

Required materials

- OpenIntro Statistics, 2nd Edition
- i>clicker2 - See Google Doc for a list of students selling used clickers (link emailed)
- (optional) Calculator

Webpage

http://bit.ly/sta101sp15

Grading

Component	Weight
Attendance & participation + peer evaluation	7.5%
Problem sets	10%
Labs	10%
Readiness assessments	10%
Performance assessments	2.5%
Project 1	5%
Project 2	10%
Midterm 1	10%
Midterm 2	10%
Final	25%

- Grades may be curved at the end of the semester.
- Cumulative numerical averages of 90 - 100 are guaranteed at least an A-, 80 - 89 at least a B-, and 70 - 79 at least a C-, however the exact ranges for letter grades will be determined after the final exam.
- The more evidence there is that the class has mastered the material, the more generous the curve will be.

Course goals and objectives

- Recognize the importance of data collection, identify limitations in data collection methods, and determine how they affect the scope of inference.
- Use statistical software to summarize data numerically and visually, and to perform data analysis.
- Have a conceptual understanding of the unified nature of statistical inference.
- Apply estimation and testing methods to analyze single variables or the relationship between two variables in order to understand natural phenomena and make data-based decisions.
- Model numerical response variables using a single or multiple explanatory variables.
- Interpret results correctly, effectively, and in context without relying on statistical jargon.
- Critique data-based claims and evaluate data-based decisions.
- Complete two research projects: one that focuses on statistical inference and one that focuses on modeling.

Learning units and course outline

- *Unit 1 - Intro to data:* Observational studies and non-causal inference, principles of experimental design and causal inference, exploratory data analysis, and introduction to simulation-based statistical inference
- *Unit 2 - Probability & distributions:* Basics of probability and chance processes, Bayesian perspective in statistical inference, the normal and binomial distributions
- *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*
  - Project 1 & Midterm 2
- *Unit 6 - Simple linear regression:* Bivariate correlation and causality, introduction to modeling
- *Unit 7 - Multiple linear regression:* More advanced modeling with multiple predictors
  - Project 2 & Final

Course structure

- Set of learning objectives and required and suggested readings, videos, etc. for each unit
- Prior to beginning the unit, watch the videos and/or complete the readings and familiarize yourselves with the learning objectives
- Begin a new unit with a readiness assessment: individual, then team
- Class time: split between lecture, discussion/application, and lab
- Complement your learning with problem sets
- Wrap up a unit with a performance assessment

Teams

- Highly functional teams of learners based on survey and pre-test
- Team members first point of contact
- Application exercises, labs, team readiness assessments, projects
- Study together, but anything that is not explicitly a team assignment must be your own work
- Peer evaluations to ensure that all team members contribute to the success of the group and to address any potential issues early on
  - If you feel that there are issues within your team, you are encouraged to discuss it with your team members and to bring it to my or your TA's attention ASAP (don't wait till things get worse)

Clickers

Objective: Two-way communication and instant feedback

- Readiness assessments (graded for accuracy)
- Questions throughout lecture (graded for participation)
  - to get credit for the day you must respond to at least 75% of the questions
  - up to three unexcused late arrivals or absences will not affect your clicker grade
- Register your clicker
  - *https://www1.iclicker.com/register-clicker* (Student ID = Net ID)
  - grading starts Mon, Jan 26

Attendance & participation

Objective: Make you an active participant and help me pace the class

- Attendance and participation during class, as well as your activity on Piazza make up a non-insignificant portion of your grade in this class
- Might sometimes call on you during the class discussion, however it is your responsibility to be an active participant without being called on

Problem sets (PS)

Objective: Help you develop a more in-depth understanding of the material and help you prepare for exams and projects

- Questions from the textbook
- Show all your work to receive credit
- *Required format:* Use one of the following, no other submission types will be accepted
  - Type your answers in the text box on Sakai and attach any plots/images as separate files, properly named
  - Attach a PDF (not Word, Google Doc, etc.) of your answers
- Welcomed and encouraged to work with others, but turn in your own work
- No make-ups, excused absences (e.g. STINF) do not excuse homework
- Lowest PS score will be dropped

Labs

Objective: Give you hands on experience with data analysis using statistical software and provide you with tools for the projects

- Work in teams: author / discussants
- Must be present in lab session to get credit
- Lowest lab score will be dropped

Activity: Get started with R/RStudio

- Go to the course website, *http://bit.ly/sta101sp15*, click on the RStudio link (top right)
  - Make sure you're on the Duke network, not visitor
- Log in using your Net ID and password
- In the Console, generate a random number between 1 and 5, and introduce yourself to that many people sitting around you:

```
sample(1:5, size = 1)
```


Readiness assessments (RA)

Objective: Encourage you to watch the videos and/or complete the reading assignment and review the learning objectives prior to coming to class as well as evaluate your conceptual understanding of the unit's material

- 10 multiple choice questions, at the beginning of a unit
- Conceptual questions addressing the learning objectives of the new unit, assessing familiarity and reasoning, not mastery
- Take the individual RA using clickers, then re-take in teams
- Individual RA score 3/4 of grade, team RA score 1/4 & your input during the team portion will factor into your participation grade
- Lowest RA score will be dropped

Performance assessments (PA)

Objective: Evaluate your mastery of the material by the end of a unit and give you instant feedback on your performance.

- 10 multiple choice questions, at the end of a unit
- Taken individually on Sakai
- Lowest PA score will be dropped

Projects

Objective: Give you independent applied research experience using real data and statistical methods

- Project 1: For a parameter of interest to you, you will describe the relevant data, compute a confidence interval and conduct a hypothesis test, and summarize your findings in a written, fully reproducible, data analysis report
- Project 2: Use all (relevant) techniques learned in this class to analyze a dataset provided by me, and share your results in a poster session
- Must complete both projects and score at least 30% of the points on each project in order to pass this class

Exams

Midterm 1	Wed, Feb 18
Midterm 2	Wed, Mar 25
Final	Sat, May 2 (2-5pm)

- Exam dates cannot be changed, no make-up exams will be given
- If you cannot take the exams on these dates you should drop this class
- Calculator + cheat sheet allowed

Email & Piazza

- I will regularly send announcements by email, so make sure to check your email daily
- Any non-personal questions related to the material covered in class, problem sets, labs, projects, etc. should be posted on Piazza forum
- Before posting a new question please make sure to check if your question has already been answered, and answer others' questions
- Use informative titles for your posts
- It is more efficient to answer most statistical questions "in person" so make use of OH

Students with disabilities

Students with disabilities who believe they may need accommodations in this class are encouraged to contact the Student Disability Access Office at (919) 668-1267 as soon as possible to better ensure that such accommodations can be made

http://www.access.duke.edu/students/requesting/index.php

Late work policy

- Late work policy for problem sets and labs reports:
  - next day: lose 30% of points (within 24 hours of due date)
  - later than next day: lose all points
- Late work policy for projects: 10% off for each day late

Regrade policy

Regrade requests must be made *within 3 days* of when the assignment is returned, and must be submitted to me in writing

- These will be honored if points were tallied incorrectly, or if you feel your answer is correct but it was marked wrong
- No regrade will be made to alter the number of points deducted for a mistake
- There will be no grade changes after the final exam

Make up policy

- No make-up for attendance, individual and team readiness assessments, labs, problem sets, projects, or exams
- If the midterm exam must be missed due to a documented medical excuse, absence must be officially excused *in advance*, in which case the missing exam score will be imputed using the final exam score
- The final exam must be taken at the stated time
- You must take the final exam and turn in the projects in order to pass this course

Other policies

- Clickers may not be shared, and the clicker registered to a person may only be used by that person, failure to abide by this will result in a 0 clicker grade for everyone involved
- Use of disallowed materials (textbook, class notes, web references, any form of communication with classmates or other persons, etc.) during exams will not be tolerated

Academic Dishonesty

Any form of academic dishonesty will result in an immediate 0 on the given assignment and will be reported to the Office of Student Conduct. Additional penalties may also be assessed if deemed appropriate. If you have any questions about whether something is or is not allowed, ask me beforehand.
Some examples:

- Use of disallowed materials (including any form of communication with classmates or accessing the web) during exams and readiness assessments
- Plagiarism of any kind
- Use of outside answer keys or solution manuals for the homework

Tips for success

- Complete the reading before a new unit begins, and then review again after the unit is over.
- Be an active participant during lectures and labs.
- Ask questions - during class or office hours, or by email. Ask me, your TAs, and your classmates.
- Do the problem sets - start early and make sure you attempt and understand all questions.
- Start your projects early and and allow adequate time to complete them.
- Give yourself plenty of time time to prepare a good cheat sheet for exams. This requires going through the material and taking the time to review the concepts that you're not comfortable with.
- Do not procrastinate - don't let a unit go by with unanswered questions as it will just make the following unit's material even more difficult to follow.

To do

- Download or purchase the textbook
- Obtain and register your clicker
  - *https://www1.iclicker.com/register-clicker* (Student ID = Net ID)
- Complete the following by Friday, Jan 9, 11:59pm
  - Pretest
  - Getting to know you survey
  - Performance assessment 0 (on course policies etc., not graded, for practice with the quiz module on Sakai)
- Read the syllabus and let me know if you have any questions
- Watch/Read/Review the resources for Unit 1

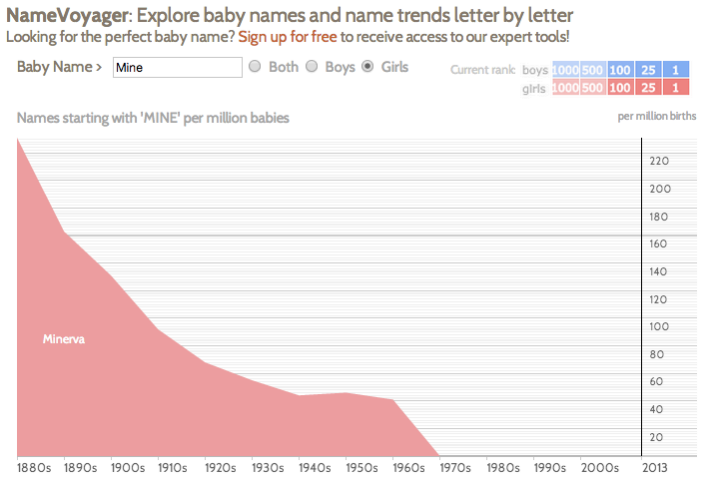
Baby names in the US

- Each year the Social Security Administration collects and releases data on the how many babies are given a certain name
- They released these data for years 1880 onwards for each gender
- For privacy reasons they restrict the list of names to those with at least 5 occurrences

Top 10 baby names for 2013

Rank	Male name	Female name
1	Noah	Sophia
2	Liam	Emma
3	Jacob	Olivia
4	Mason	Isabella
5	William	Ava
6	Ethan	Mia
7	Michael	Emily
8	Alexander	Abigail
9	Jayden	Madison
10	Daniel	Elizabeth

http://www.ssa.gov/oact/babynames



http://www.babynamewizard.com/voyager

FiveThirtyEight Life

ICYMI | 1:40 PM | MAY 29, 2014

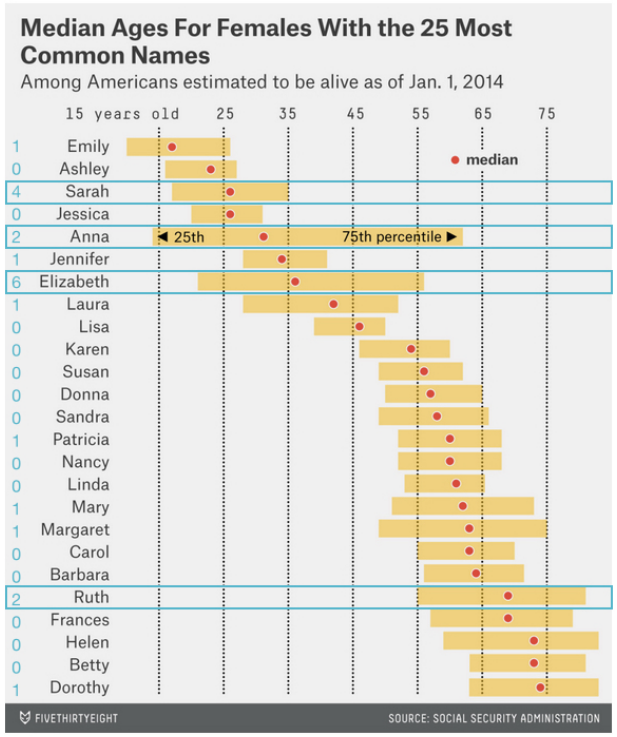
How to Tell Someone's Age When All You Know Is Her Name

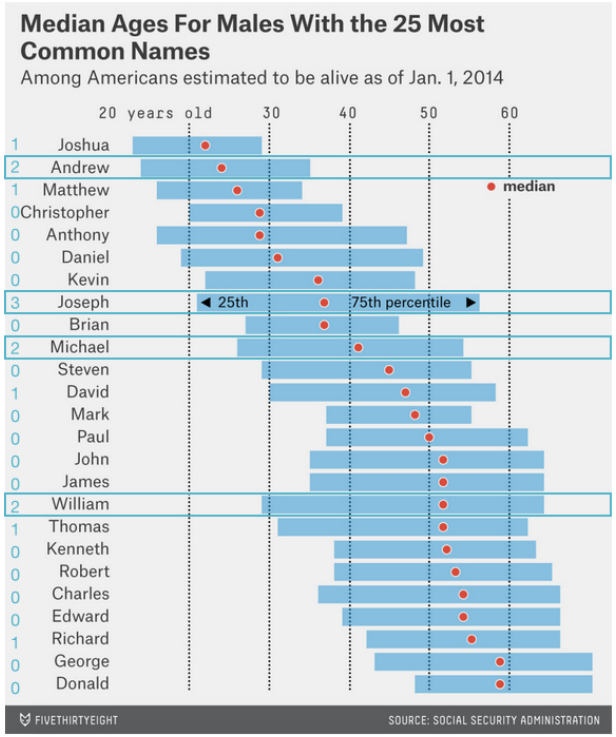
By NATE SILVER and ALLISON MCCANN

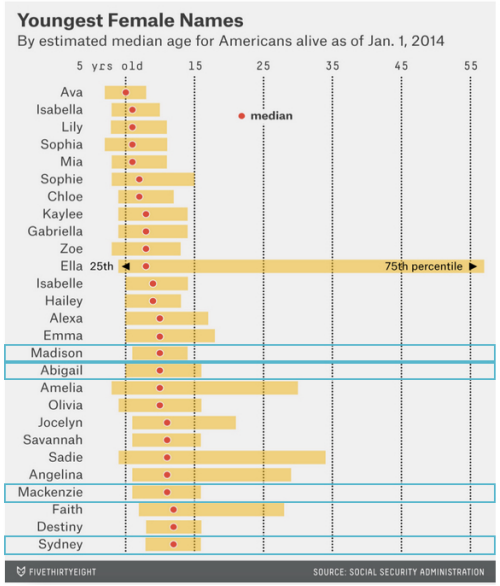
Picture Mildred, Agnes, Ethel and Blanche. Perhaps you imagine the Golden Girls or your grandmother's poker game. These are names for women of age, wisdom and distinction. The median living Mildred in the United States is now 78 years old.

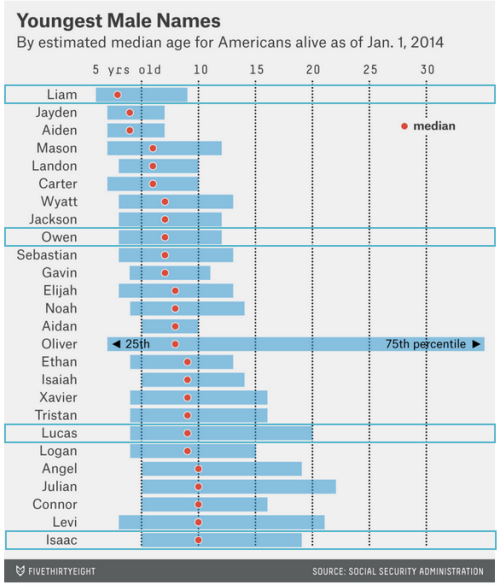
Now imagine Madison, Sydney, Alexa and Hailey. They sound like the starting midfield on a fourth-grade girls' soccer team. And they might as well be: the median American females with these names are between 9 and 12 years old.

http://fivethirtyeight.com/features/how-to-tell-someones-age-when-all-you-know-is-her-name







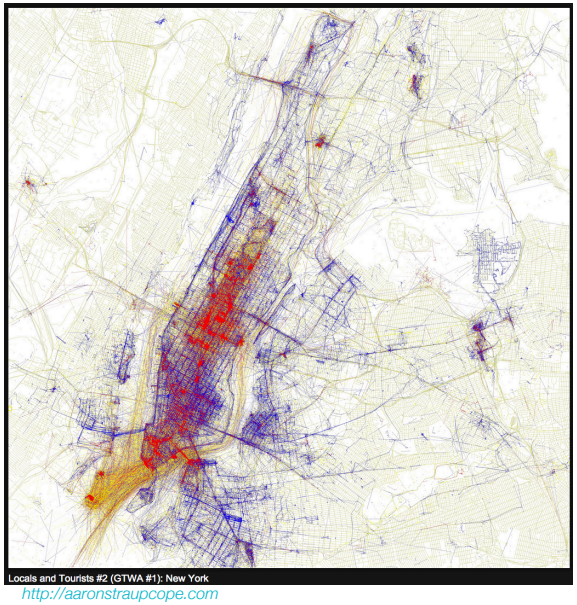


Clicker question

Do you geotag your posts on social networking sites, like Facebook, Twitter, Instagram, etc.?

(a) yes
(b) no

Maps based on clicker tags



http://aaronstraupcope.com

tourists
local
both

'Lucky' woman who won lottery four times outed as Stanford University statistics PhD

By RACHEL QUIGLEY

Last updated at 7:02 PM on 9th August 2011

She was called the luckiest woman in the world.

But now that luck is being called into question by some who think that winning the lottery four times is more than just a coincidental spell of good fortune.

Joan R. Ginther, 63, from Texas, won multiple million dollar payouts each time.

Lady luck? Joan Ginther has won the Texas lottery four times, but is it remarkable good fortune or beating the system?

What do statisticians do? Here are a few examples:

Help animals
Collaborate with other scientists to find ways to protect endangered species.

Combat disease
Help researchers understand how prevalent diseases are among various populations and why.

Build winning professional sports teams
Work for professional sports teams to help them pick the next season's new players.

Make drugs safe
Help develop new medicines that are safe and effective.

Jobs in statistics are not only fun and exciting, but also smart for your future:

Job growth is strong
Jobs for statisticians will grow **27 percent between 2012 and 2022**, much faster than the growth rate of 18 percent for computer occupations, and 11 percent for all occupations.

Wages are high
The median salary for data scientists with less than three years of experience is **$80,000 and $150,000** for those with nine or more years of experience, according to the Burtch Works 2014 report.

http://thisisstatistics.org

Activity: Class survey

- One of your first tasks in this class is to help design a survey. This survey will be completed anonymously. It will (ideally) have information on variables you are interested in. When writing your question consider whether you would feel comfortable answering it on an anonymous survey.
- Work with 3-4 classmates to come up with a survey question, and add it to Google Doc linked below. Make sure that the wording of the question is clear, and (if categorical) the answer choices make sense.

http://bit.ly/sta101sp15_ClassSurvey

- Before adding a question check to make sure that it hasn't already been added. If your question is already there, but you can suggest a clearer / better wording, add it as "alternative wording" underneath the original question.