NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Week 13: Lecture 1
Correlation

NYU Center for Data Science

How can we measure the association between variables?

DS-UA 111
Data Science for Everyone

Week 13: Lecture 1
Correlation

Adapted from Adhikari, DeNero, Wagner, Milner

Announcements

- Please check Week 13 agenda on NYU Classes
  - Homework 3/4
  - Lab 8
  - Project Milestone
- Refer to the Calendar linked to NYU Classes

Announcements

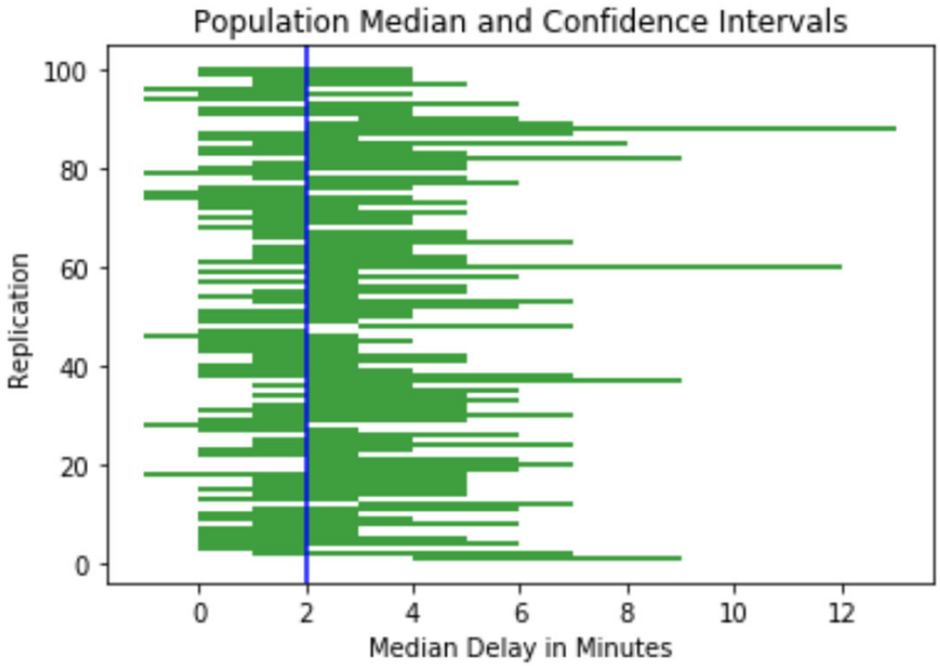
- Please check Week 13 agenda on NYU Classes
  - Homework 3/4
  - Lab 8
  - Project Milestone
- Refer to the Calendar linked to NYU Classes

Optional

- Lab 9
  - Correlation
  - April 22 – April 29
- Lab 10
  - Linear Regression
  - April 29 – May 6

Review

- For constructing a confidence interval for an unknown parameter
  1. Sample without replacement from the population to determine a sample. Larger samples are preferable to smaller sample.
  2. Sample with replacement from the sample to get a resample. Calculate the test statistic on the resample.
  3. Repeat Step 2 many times. Each replication generates another number.
  4. For an approximate 80% confidence interval, take the 10th and 90th percentiles of all the resample estimates.



Review

- If we take a p% confidence interval for hypothesis testing then p% of the time we expect the interval to contain the population parameter
  - So we have a false reject about

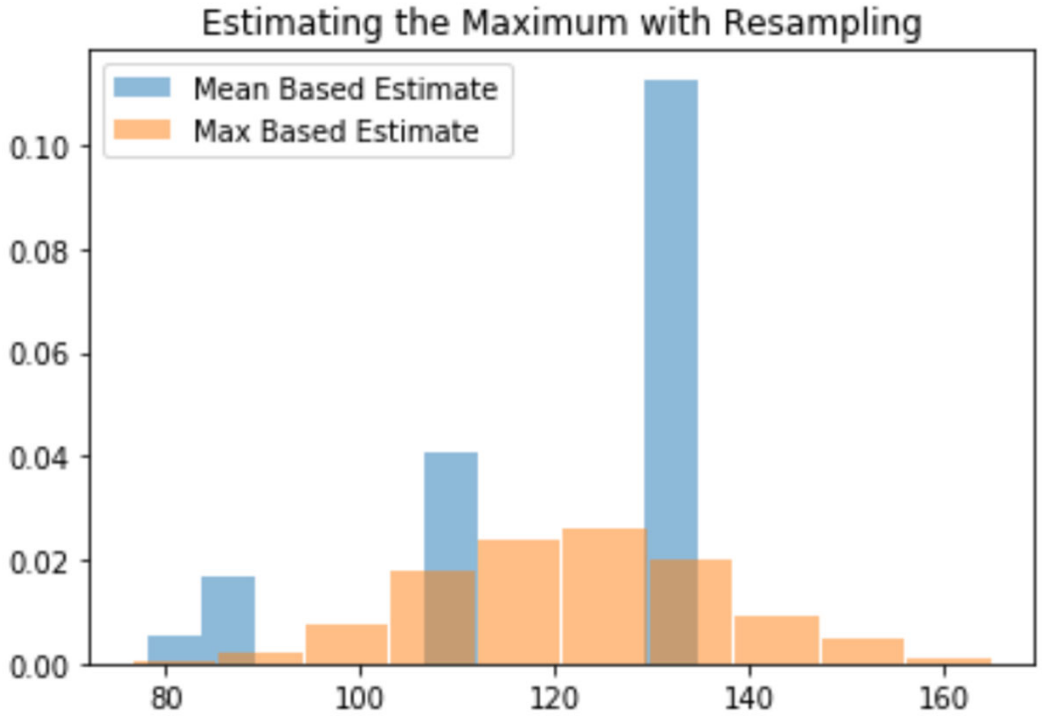
    (100-p)% of the time
- If we want to estimate any parameters related to rare elements of the population, then the confidence intervals from resamples might be inaccurate. Parameters like
  - Maximum Value
  - Minimum Value

- Remember that the resamples cannot contain data outside of the sample.
- So if the sample is small, then the confidence intervals from resamples might be inaccurate

Review

- Suppose we want to estimate the average age of mothers in a population.
- We use bootstrap resampling to generate approximate 95% confidence interval for the average age of the mothers in the population

  26.9 years to 27.6 years
- True or False
  - About 95% of the mothers in the population were between 26.9 years and 27.6 years old.
  - False: We're estimating that their average age is in this interval.



Agenda

- Understanding Associations with Visualizations
  - Scatter-plots
- Understanding Associations with Number
  - Correlation

References

- Prediction
  - Chapter 15.1

Mean

- Among different statistics for estimate of parameters, the mean lends itself to resampling.
- We can make some observations about the mean that hold for any population.
  - The mean of numbers might not be contained among the numbers
    - For example the mean of integers could be a fraction
  - The mean is greater than the minimum and less than the maximum
  - The mean aggregates many numbers into one representative number

- Suppose we have data

  {2,3,9,9}

- The mean is

  $4.25 = (2 + 3 + 9 + 9) / 4$

- We can think of the mean as a weighted sum of the numbers. Here the weights reflect the frequency of the number

$$4.25 = 2*(¼) + 3*(¼) + 3*(¼) + 9*(¼)$$
$$= 2*(¼) + 3*(2/4) + 9*(¼)$$
$$= 2*0.25 + 3*0.5 + 9*0.25$$

Standard Deviation

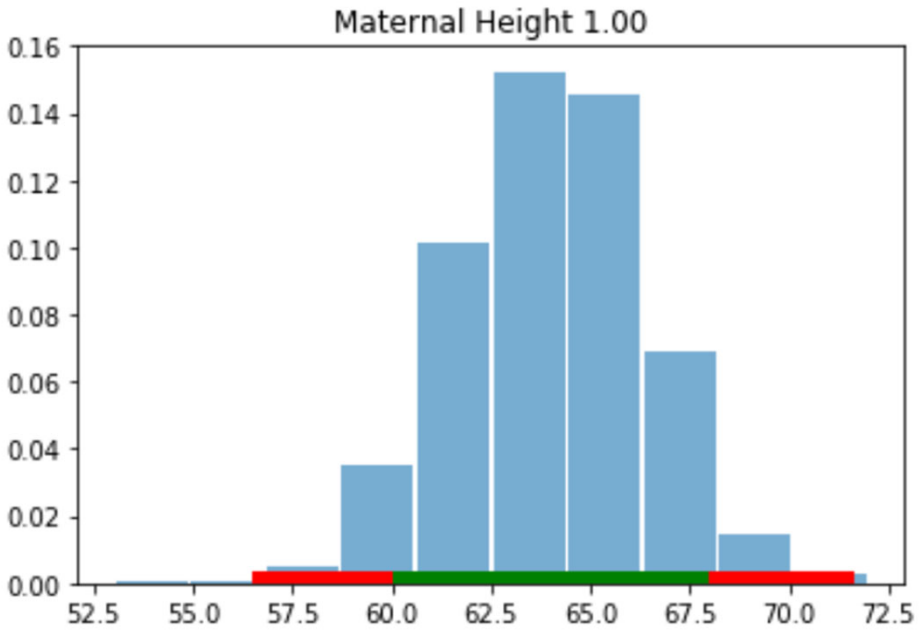
- The standard deviation measures the difference between numbers and the mean of the numbers.
- Regardless of the population we know that a certain amount of the data has to lie near the mean.
- The standard deviation bound tells use the fraction of data greater than

  Mean – z * (Standard Deviation)

and less than

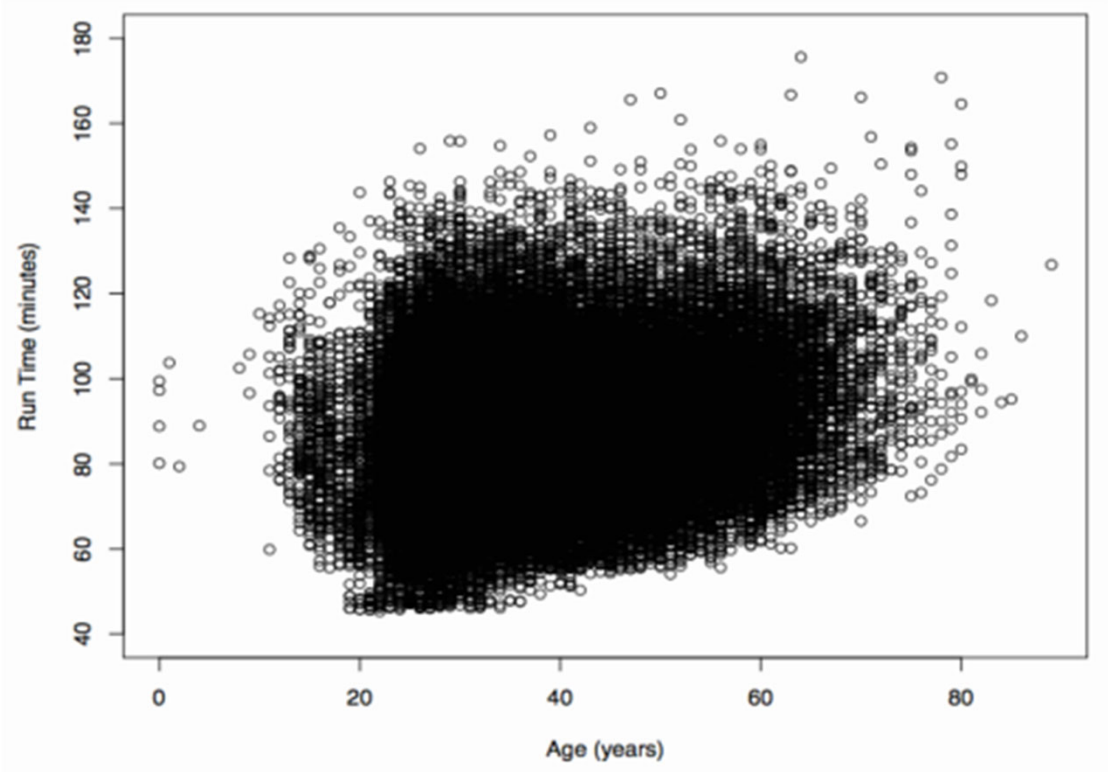
  Mean + z * (Standard Deviation)

is a least $1 - (1 / z^2)$



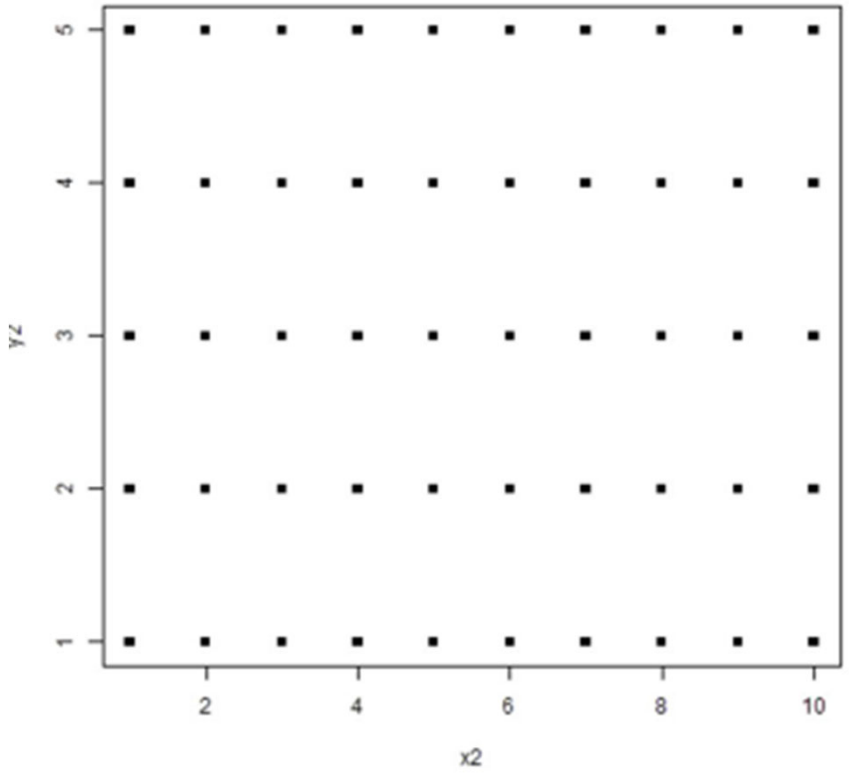
Scatter-Plots

- We use scatter-plots to visualize two quantitative variables.
- The horizontal coordinate corresponds to one variable. The vertical coordinate corresponds to the other variable



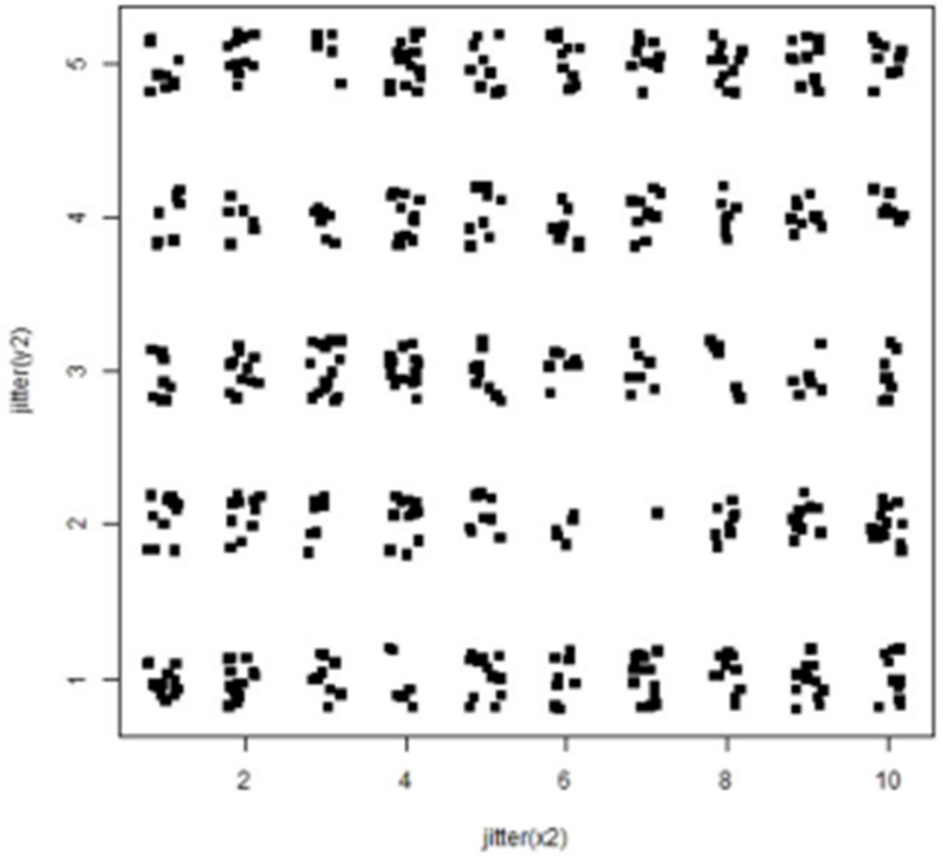
Over-Plotting

- With too many numbers plotted in the same chart, the data gets obscured through over-plotting
- For example, how could the chart to the right contain 1000 records of data?



Over-Plotting

- We can try to fix over-plotting in different ways
  - Jittering the points
  - Adjusting the saturation of the colors
  - Splitting into different charts
  - Grouping together the data



Correlation

- We want to understand the association between variables to make predictions.
- Scatter-plots help us to visualize the association with charts.
- Correlation helps us to quantify the association with a number

- Trend
  - Positive
  - Negative
- Pattern
  - Linear
  - Non-linear

Correlation

- Correlation measures the linear association between variables
- The number comes from the transformation of the data to standard units
- The values of *r* range from $-1 \leq r \leq 1$
  - r = 1: scatter is perfect straight line sloping up
  - r = -1: scatter is perfect straight line sloping down
  - r = 0: No linear association; uncorrelated

Correlation

Correlation Coefficient (r) =

average of	product of	x in standard units	and	y in standard units

Summary

- Understanding Associations with Visualizations
  - Scatter-plots
- Understanding Associations with Number
  - Correlation

Goals

- Generate a scatter-plot to assess the association between variables
- Use standard units to compute the correlation
- Understand some limitations of correlation