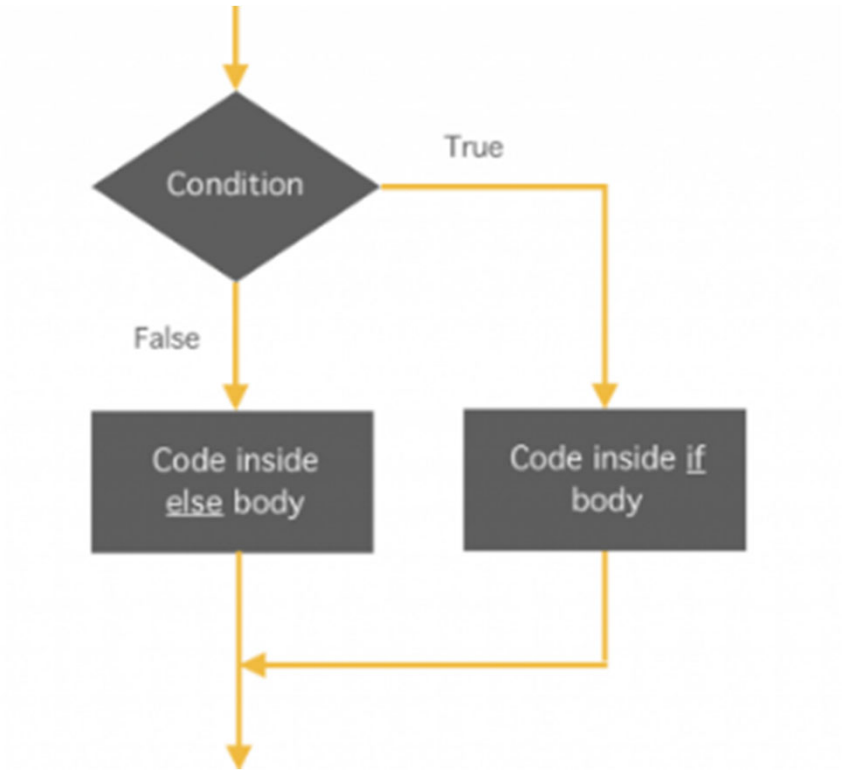
NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Final Exam Review

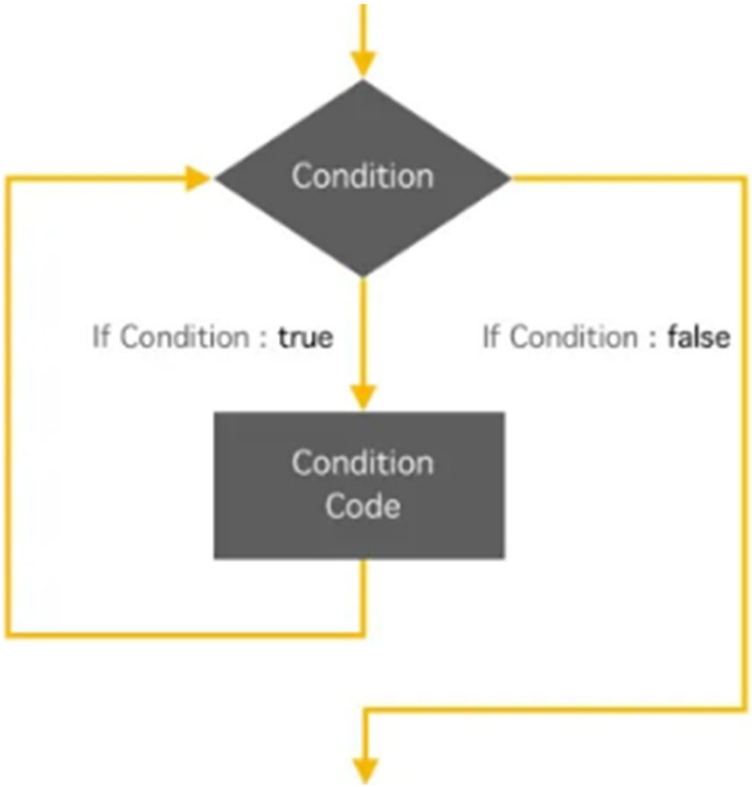
Conditional Statements

- We use a special computational data type called Boolean for True and False in Python
- Think of True/False as
  - Yes/No
  - 1/0
  - Not Empty/Empty…



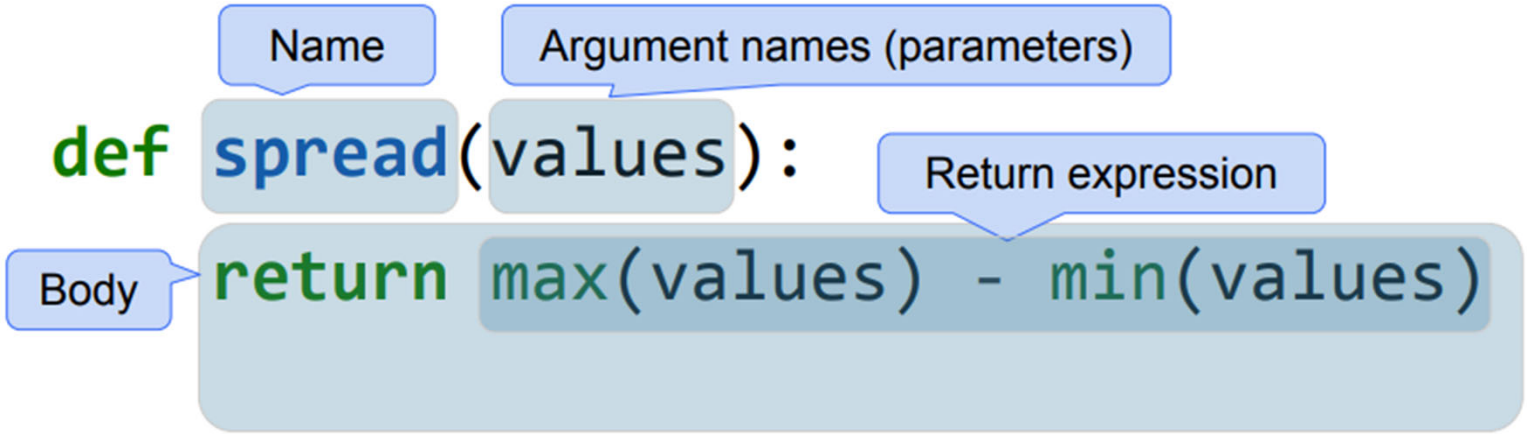
Loops

- We can repeatedly run a block of code in Python using a loop
- for loop
  - Runs the block of code for specified number of iterations
- while loop
  - Combines conditional statement and for loops
  - Runs block of code while the logical expression is True



Functions

- Input / Output
  - The input of the function appears within parentheses
  - Functions may not need an output
- If we want the function to output something, then we use the keyword **return**



Total Variation Distance

- Step 1
  - Take the difference between the proportions corresponding to each category
- Step 2
  - Apply absolute value transformation to obtain positive numbers
- Step 3
  - Add the transformed numbers. Divide the summation by 2.

Total Variation Distance

- Step 1: Differences
- Step 2: Absolute Value
- Step 3: Summation

Hypothesis Testing

- Hypothesis Testing
  - We use tests to connect questions and answers about the data generating processes in a population
  - With hypothesis testing, we have two possibilities summarized by null hypothesis and alternative hypothesis

False Reject

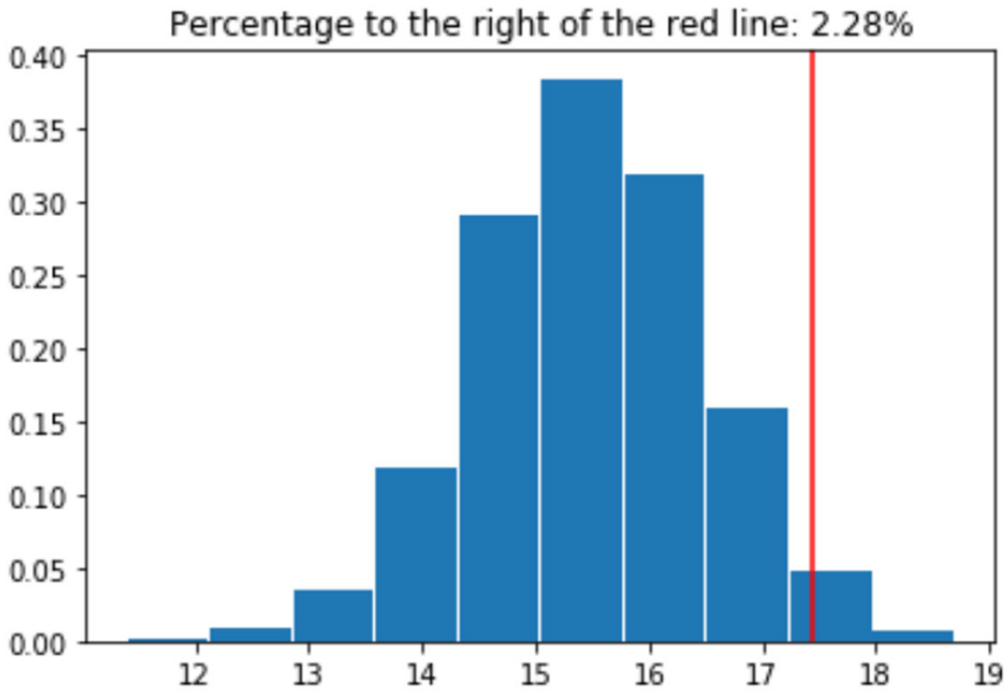
False Accept

	Null is True	**Alternative is True**
Test Favors the Null	Correct result	Error
Test Favors the Alternative	Error	Correct result

- The null hypothesis asserts that the processes follow a pattern. However, the alternative hypothesis asserts the processes do not follow the pattern.
- While we accept or reject the null hypothesis, we should think in terms of rejecting or failing to reject the null hypothesis. We must remember that hypothesis tests provide evidence not proof for the null hypothesis or alternative hypothesis

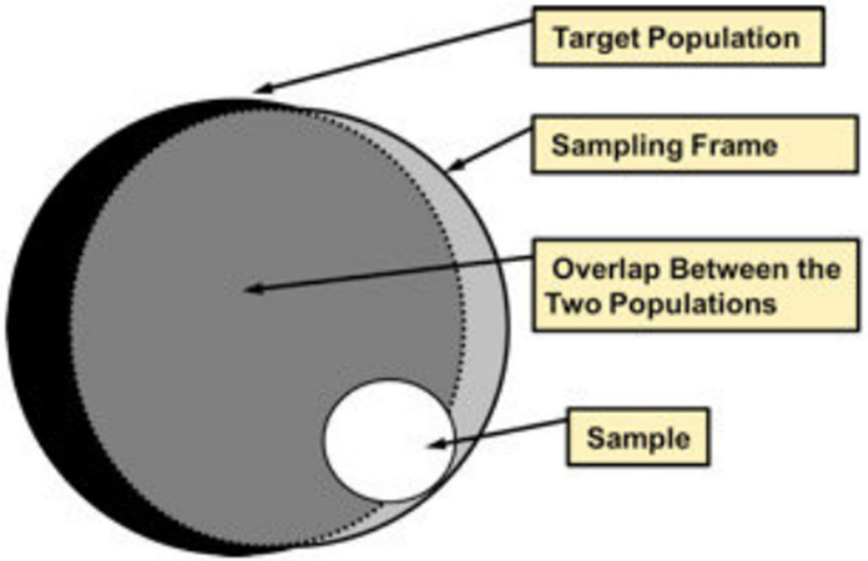
Observed Significance

- For hypothesis testing we need to compare the observed test statistic to the simulated test statistic under the null hypothesis. We check the left tail and right tail for outliers.
- We want to estimate the probability of a test statistic obtaining a value farther to the left in the left tail or farther to the right in the right tail.
- The number of simulated test statistics provides an estimate of the probability nicknamed p-value.



Population and Sample

- Researchers perform studies on a target population
- The sampling frame is the subset of the population eligible for inclusion in the sample
- The sample contains randomly or deterministically selected participants or observations from the sampling frame



Permutation Testing

Steps for Permutation Testing

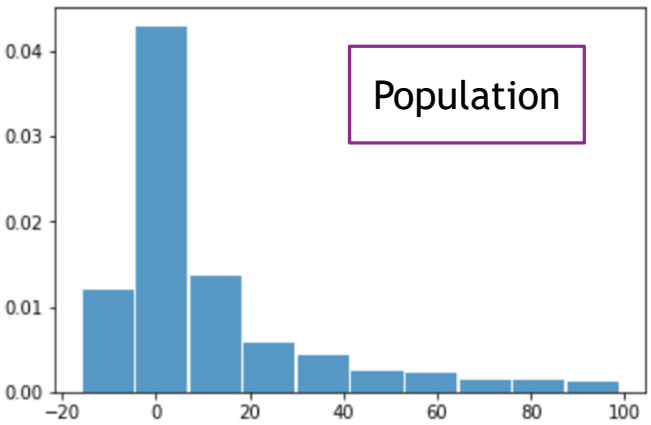
1. Fix a null hypothesis and alternative hypothesis
2. Determine a test statistic
3. Calculate the observed test statistic for the sample
4. Simulate test statistics under the null hypothesis with many trials
5. Calculate p-value for the observed test statistic with the empirical distribution

permutation	value of T	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

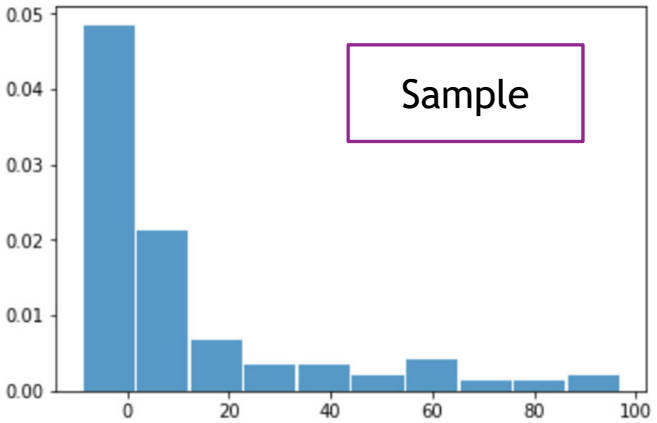
Resampling

Bootstrap Method

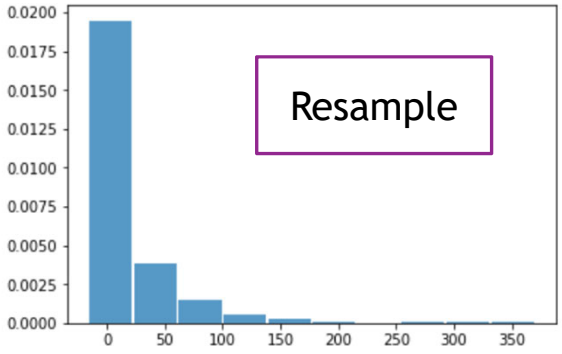
- Sample with replacement from a sample

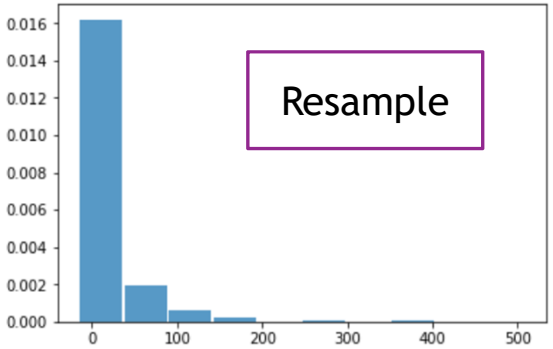


without replacement



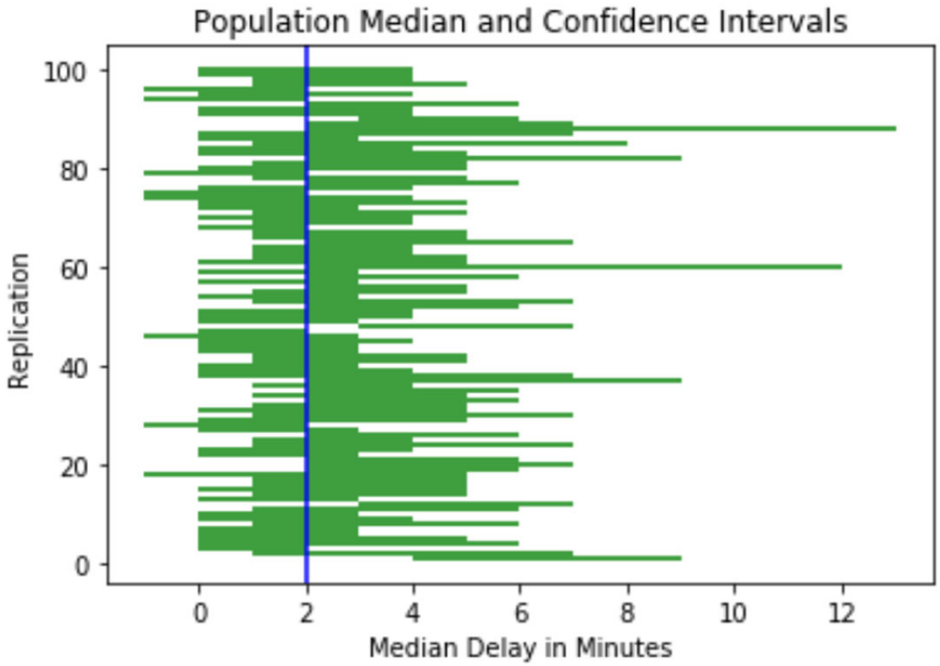
with replacement





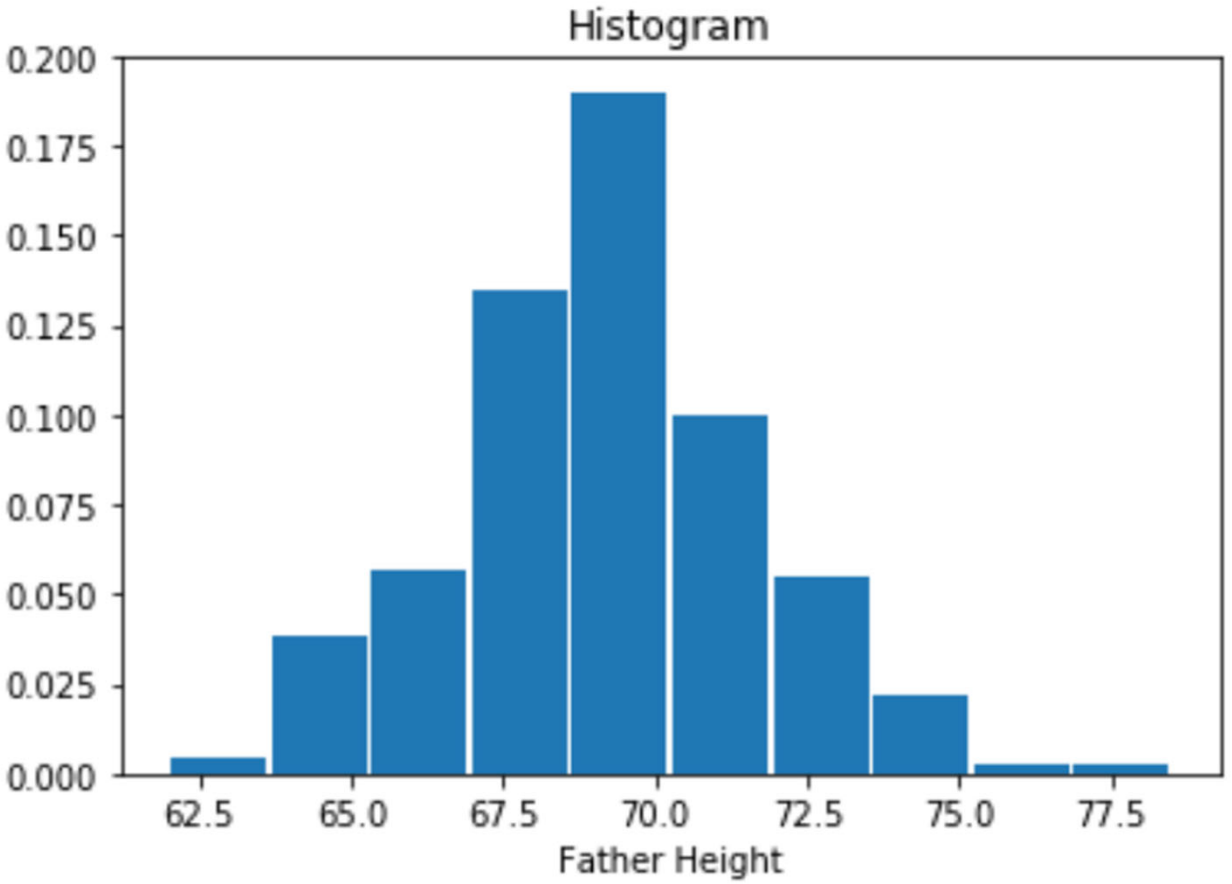
Confidence Intervals

- 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.



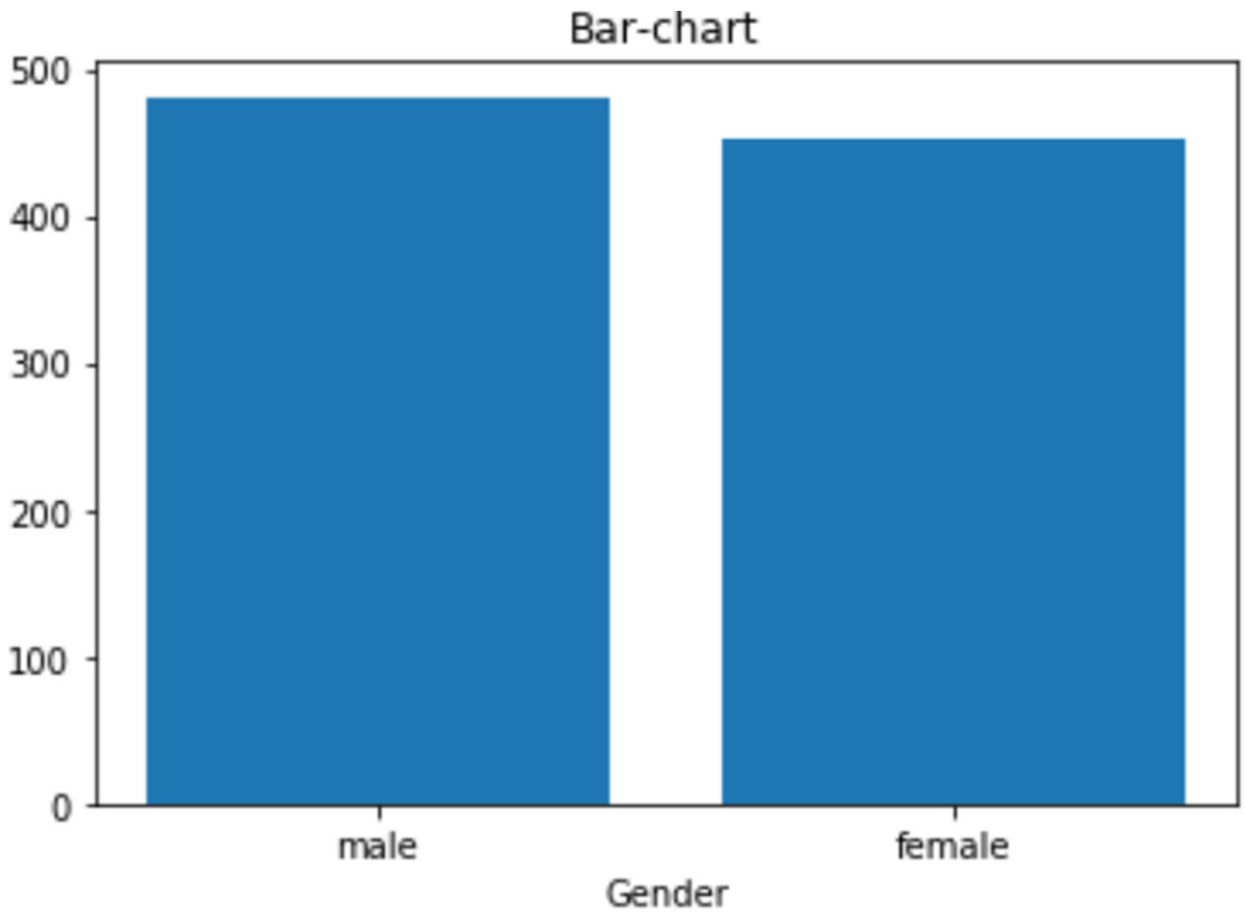
Histograms

plt.hist



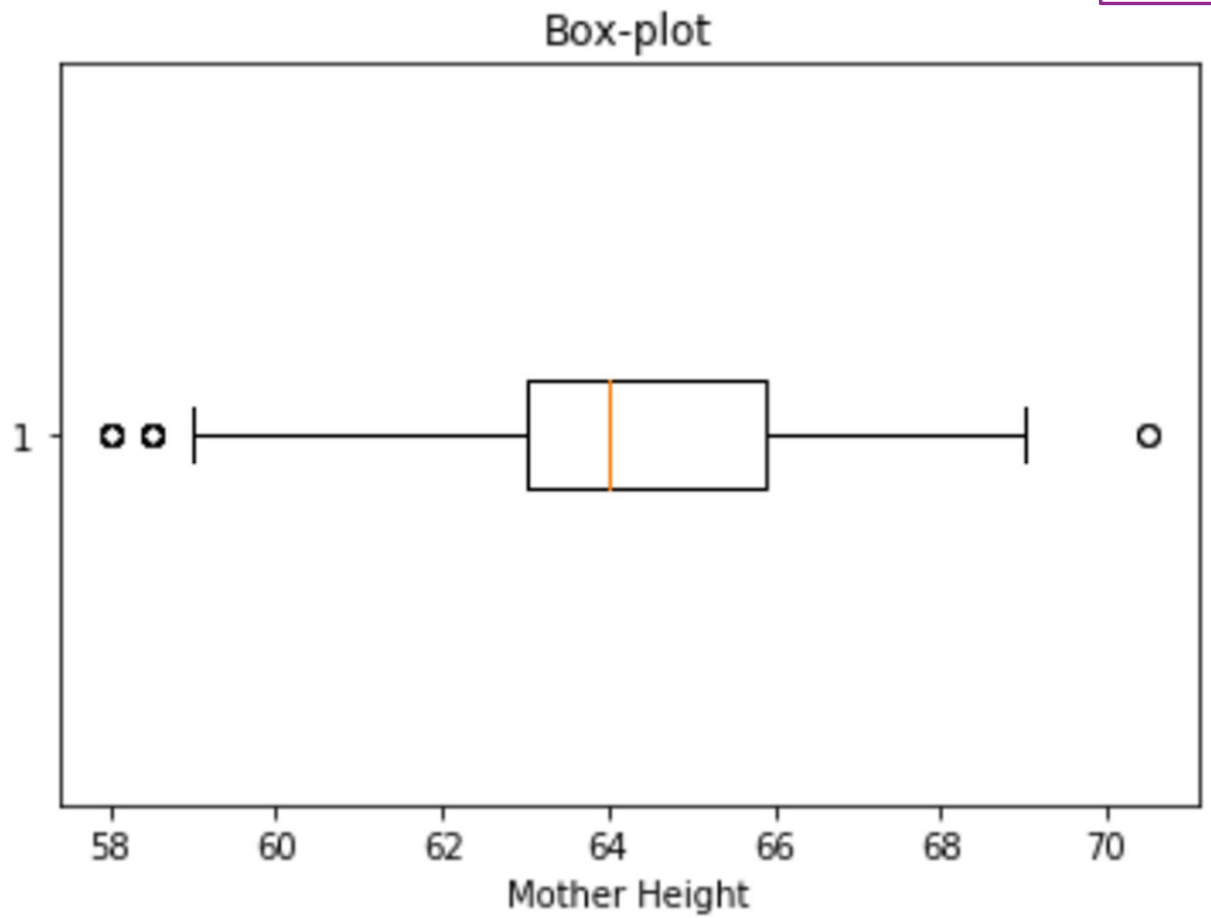
Bar-Charts

plt.bar



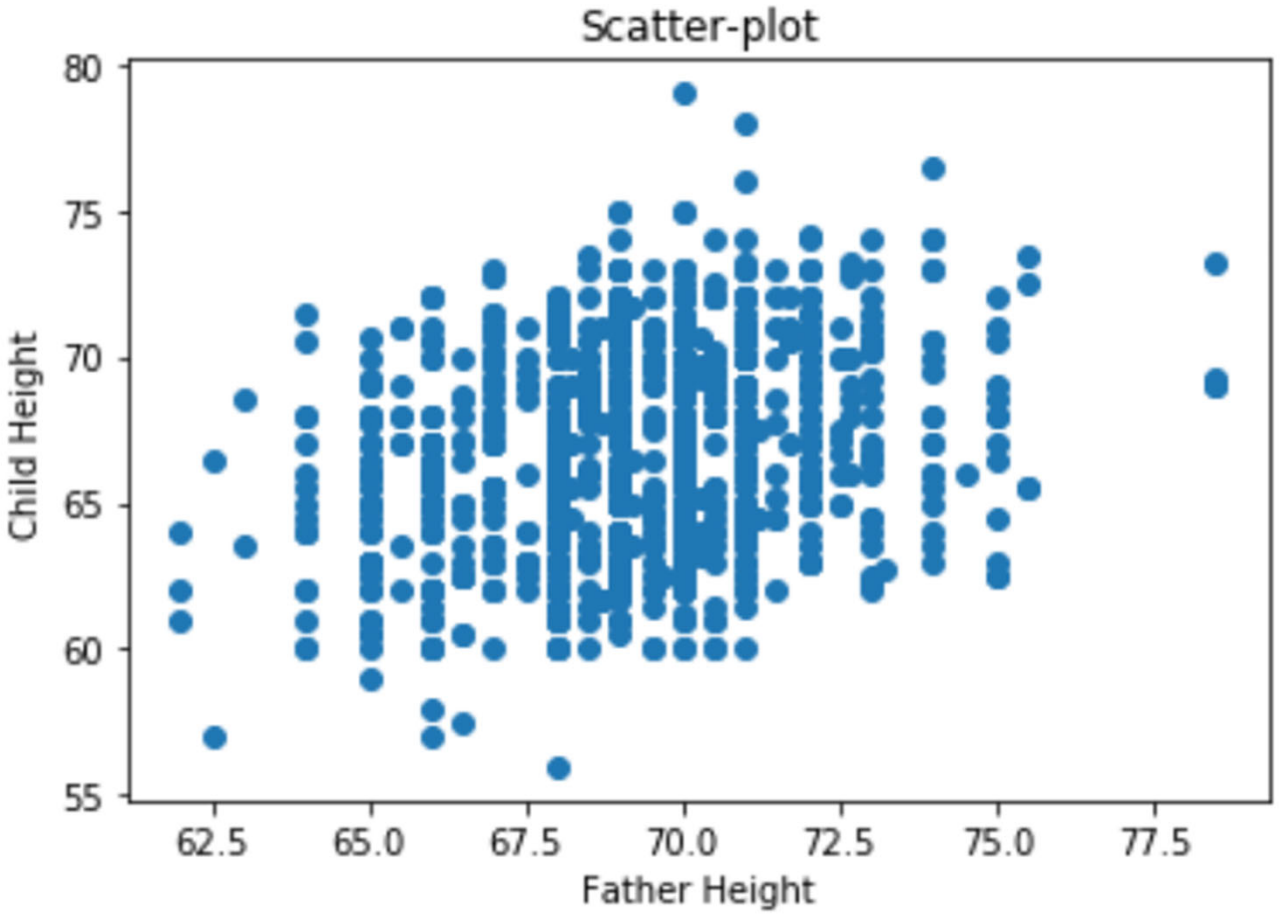
Box-Plots

plt.boxplot



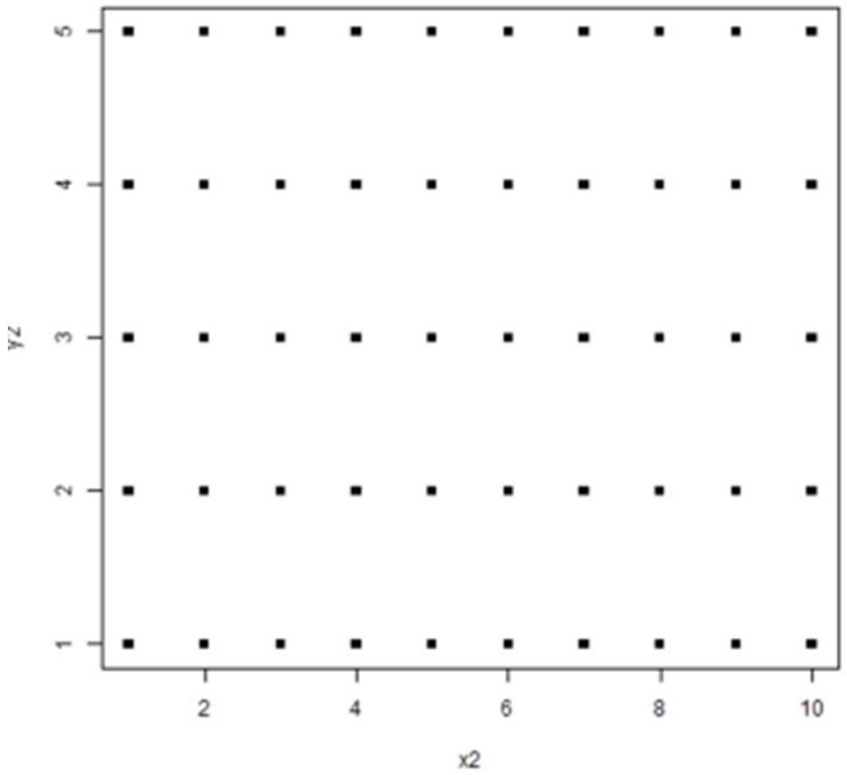
Scatter-Plots

plt.scatter



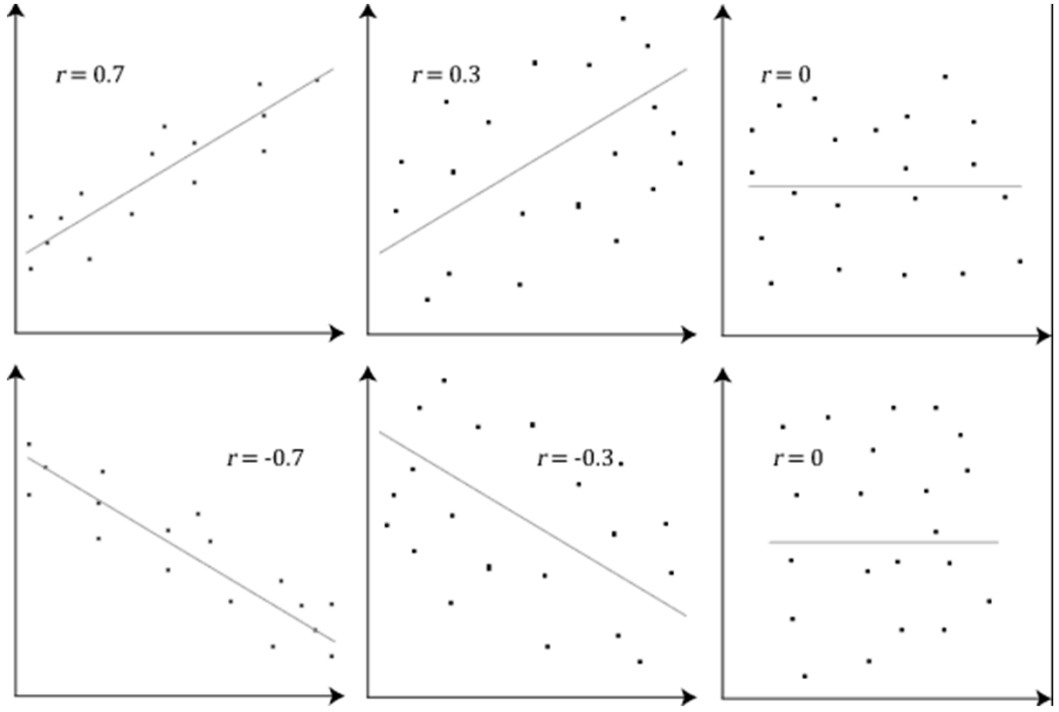
Over-Plotting

- Scatter-plots allow use to visualize two quantitative variables
- Be careful of over-plotting
  - With duplicate values we will miss data in the chart
  - With nearly duplicate values we will have a cluttered chart



Correlation

- Correlation measures the linear association between variables
- Linear means shaped like a line



Slope and Intercept

- Suppose we want to draw a line-chart through the scatter-plot to fit the pattern.
- The equation for a line is

Output = Intercept + Slope * Input

- The correlation helps us determine the slope of the line in standard units

$$\frac{\text{estimate of } y \ - \ \text{average of } y}{\text{SD of } y} = r \times \frac{\text{the given } x \ - \ \text{average of } x}{\text{SD of } x}$$

estimate of *y* in standard units

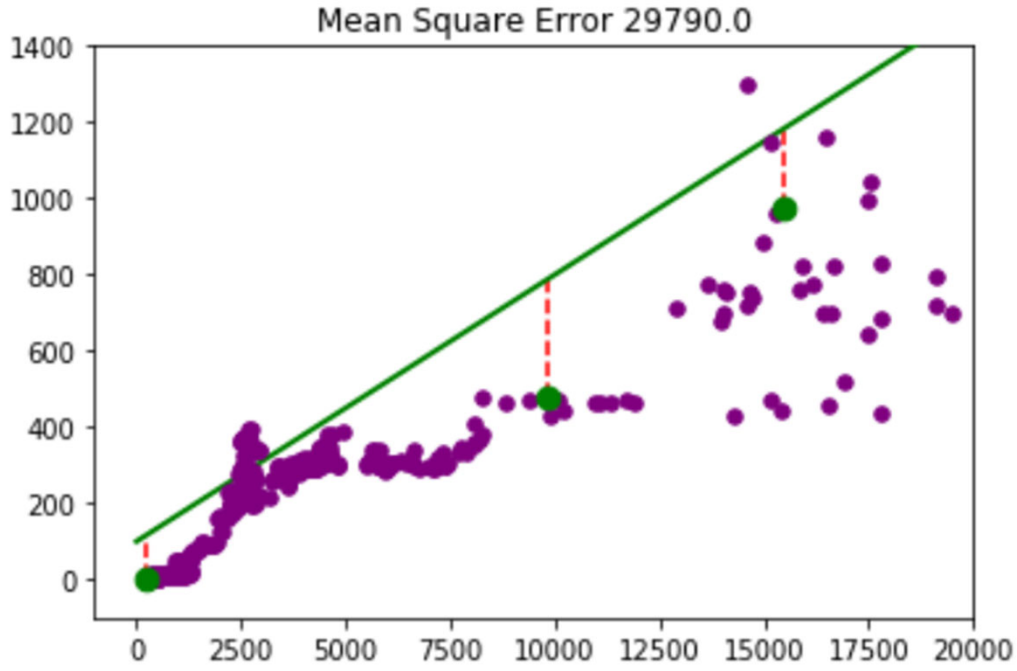
x in standard units

Least Squares Regression

- We can describe a line through a function of the form

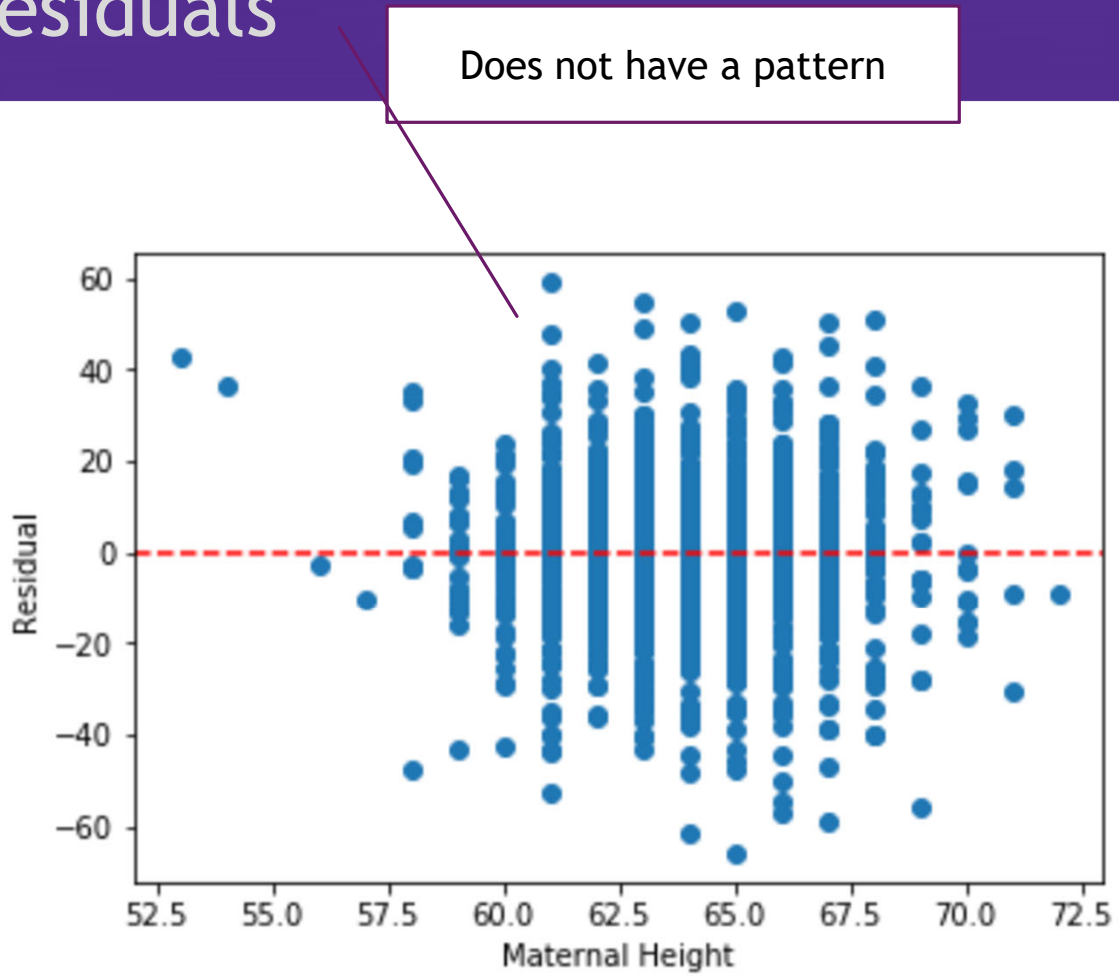
  Output = Intercept + Slope * Input

- The slope and intercept are the missing pieces in the model.
- We choose the slope and intercept that minimize the mean square error.



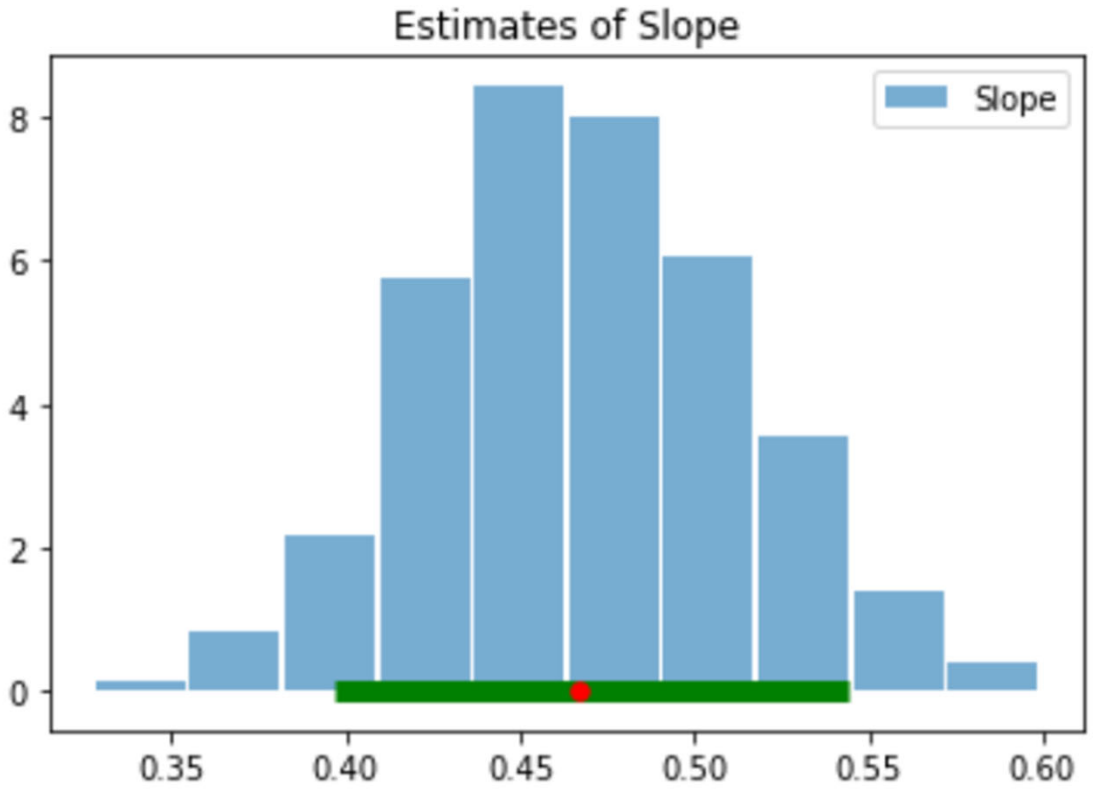
Residuals

- We can generate a scatter-plot to visualize the residuals. We want
  - About half the points above 0 and about half the points below 0
  - Comparable differences from 0 throughout the points
  - No discernible trend or pattern
- Otherwise we should explore other explanatory variables



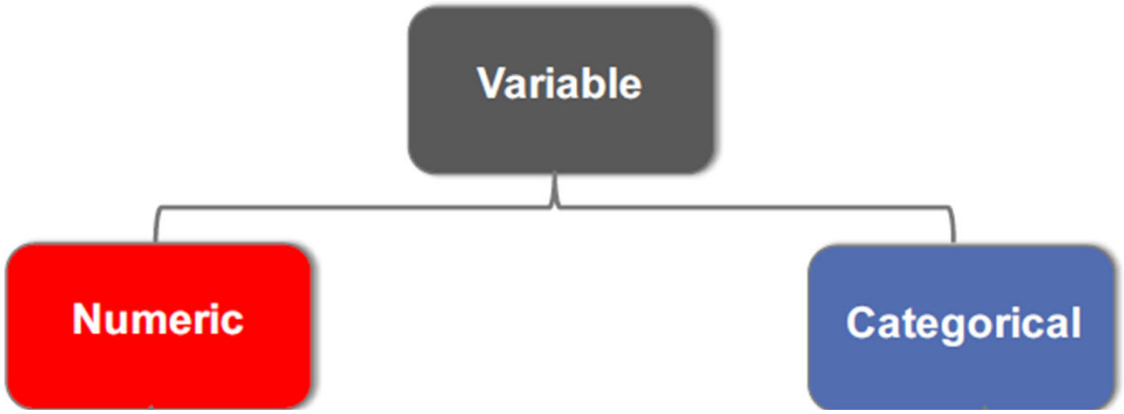
Inference for Regression

- We determine the slope and intercept through fitting the line to the data. The data is a sample from the population.
- We can quantify the variation across samples in the slope and interval through resampling.
- Bootstrap resampling allows us to generate many slopes and intercepts across replications



Data Types

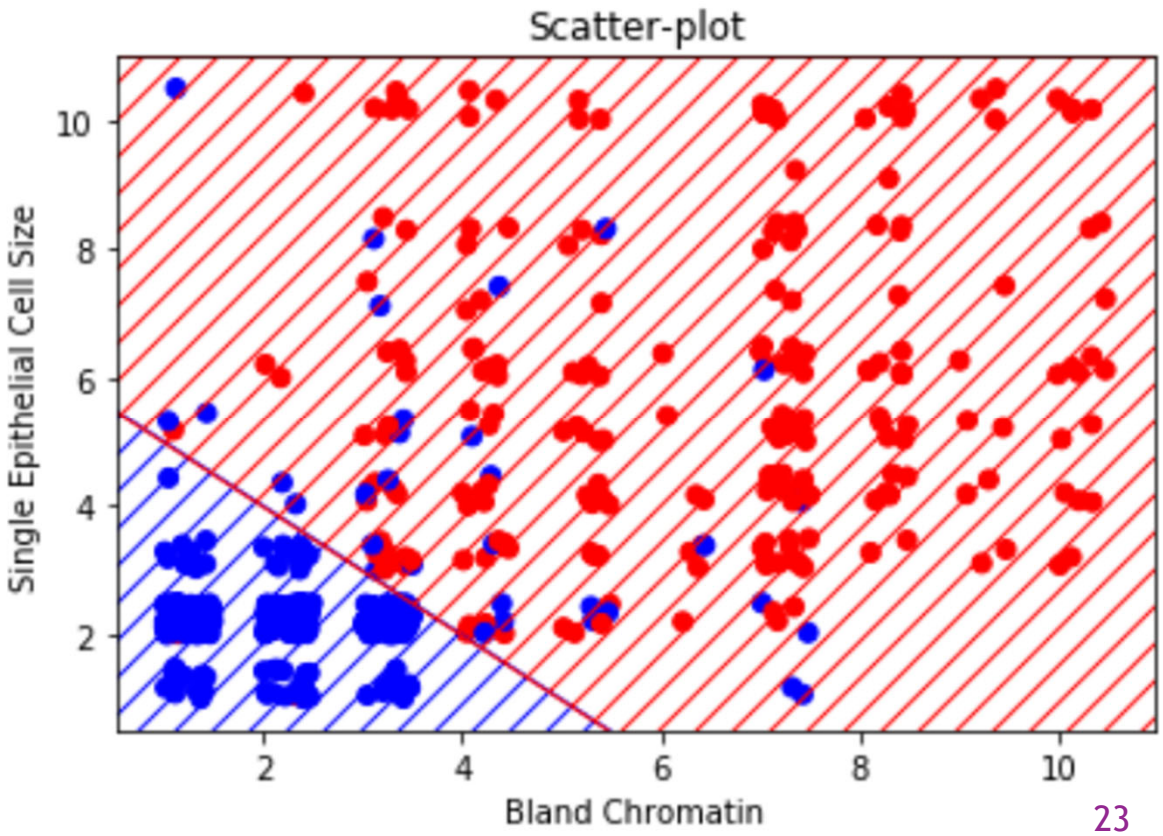
Statistical Data Types not Computational Data Types



- We study data with different properties. We divide these properties into two types
  - Numbers
    - We call it Quantitative Data
  - Categories
    - We call it Qualitative Data

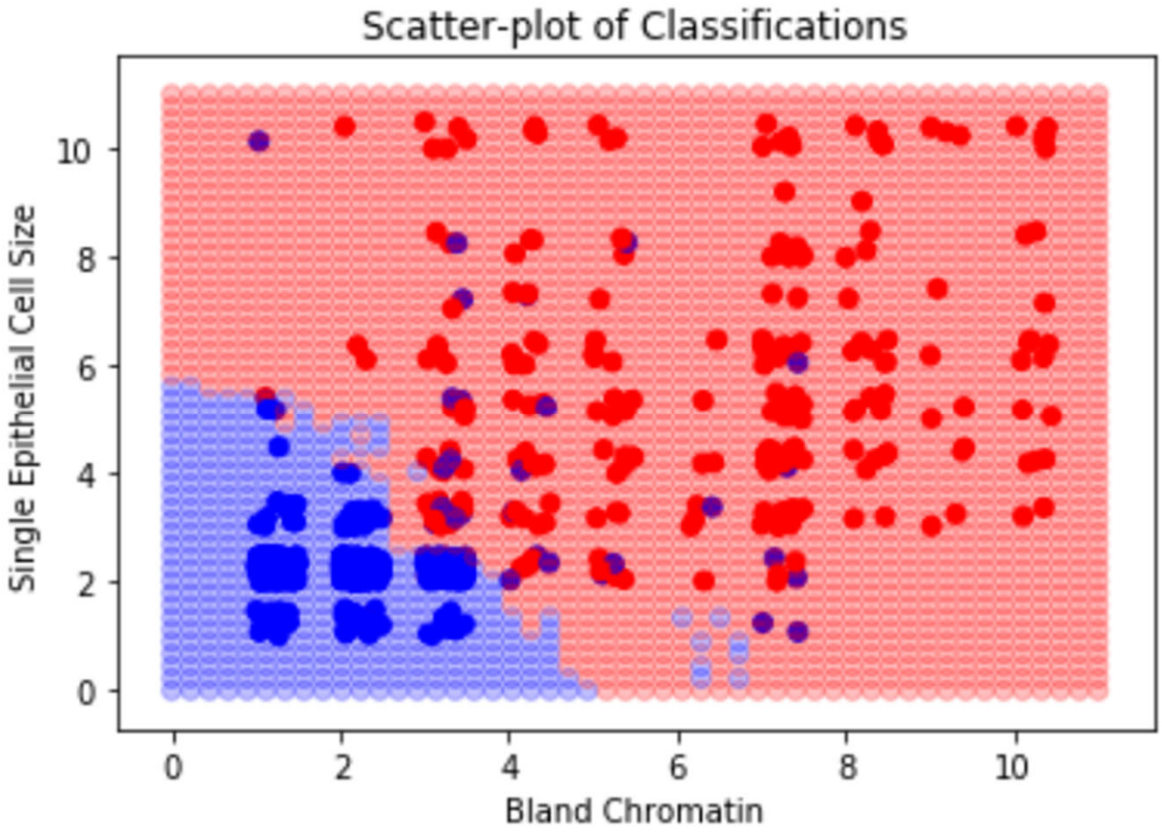
Classification

- We need to use the data to determine a boundary that separates the regions.
- We can compare determining the boundary in classification to fitting the line in regression



Nearest Neighbors

- We determine the category of the unlabeled record from the categories of the nearest labeled records.
- If we predict categories for many unlabeled records then we can determine the boundary



Training and Testing

- We determine the boundary on the training set
- We calculate the accuracy on the testing set.
- We should contrast in-sample accuracy and out-of-sample accuracy