January 17, 2018

Probability & Statistics

Tools for looking at data

Welcome to CS361

- Me: Rick Barber.
  Office 4209 Siebel, 1-3PM Tuesday
- TAs: Edward McEnrue, Farzaneh Khajouei
  Office hours TBA
- Webpage (eventually): https://courses.engr.illinois.edu/cs361/sp2018/
- Textbook: Link on webpage
- Piazza: Link on webpage

Grading

11 homeworks + take home final: 75% + 25%

Homework: approx 1 per week. Some are programming heavy, others aren't

Late policy: 10% off for each 12 hours late, e.g. 0 minutes to 12 hours late is 10% off, 12 hours to 24 hours is 20% off, etc.

Course content

Basic plots and statistics, probability theory, inference, and machine learning

Course gets progressively more mathematically difficult. By the end we are using a lot of linear algebra.

Exploratory data analysis

Datasets

- Literally a set of data
- In practice, data are multiple instances of an underlying phenomenon
- Examples
  -High temp in Champaign for every day in 2017
  -Every search query you've entered into Google + click results

Types of data

- Categorical: takes on a small set of prescribed values
- Examples
  -Eye color printed on your id
  -Survey 100 people walking on the sidewalk on rich vs famous

Types of data

- Ordinal: categorical data where we can say one item is larger than another
- Examples:
  - probably not eye color
  - your income tax bracket
  - doctor's office: pain on a scale from 1 to 10

Types of data

- Continuous: data can take on any numerical value within a range
- Examples:
  - height, weight, salary

A little more rigorously

- Dataset: a set of *d*-tuples (each item an ordered list of d dimensions)
- We will say our dataset {**x**} contains *N* items
- When we want to refer to *i*-th item, we will write $\mathbf{x}_i$
- We will assume all of our data items have *d* dimensions even if some are blank

What's going on here?

- Datasets can be huge and seemingly impenetrable
- Two basic approaches to finding out what's going on in a dataset
  - Visualizing
  - Summarizing

Visualizing data

- Tables

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

Index	Taste score	Index	Taste score
1	12.3	11	34.9
2	20.9	12	57.2
3	39	13	0.7
4	47.9	14	25.9
5	5.6	15	54.9
6	25.9	16	40.9
7	37.3	17	15.9
8	21.9	18	6.4
9	18.1	19	18
10	21	20	38.9

Visualizing data: bar charts

- Categorical data: bar charts
- A set of bars, one for each category
- Height is proportional to the number of items in the dataset whose value is that category

Example

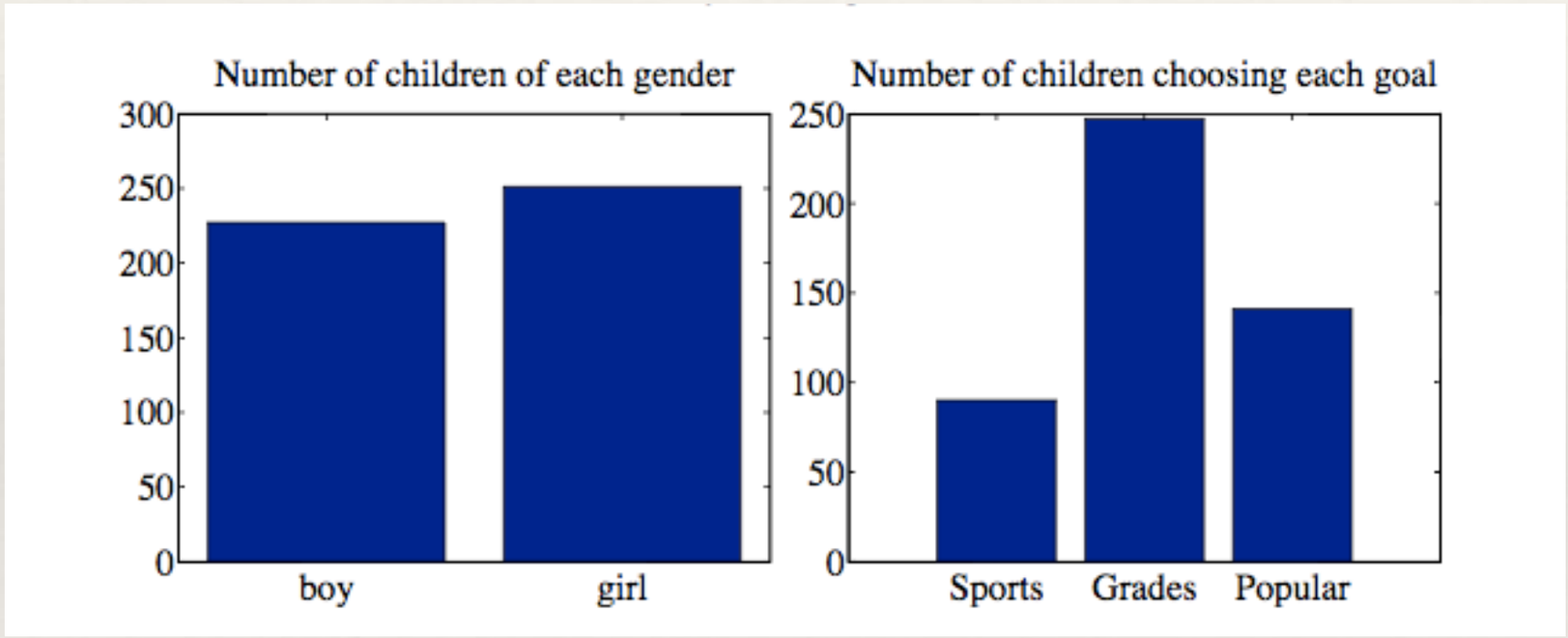
"The role of sports as a social determinant for children. Chase & Dunner.

Gender	Goal	Gender	Goal
Boy	Sports	Girl	Sports
Boy	Popular	Girl	Grades
Girl	Popular	Boy	Popular
Girl	Popular	Boy	Popular
Girl	Popular	Girl	Grades
Girl	Popular	Girl	Sports
Girl	Grades	Girl	Popular
Girl	Sports	Girl	Grades
Girl	Sports	Girl	Sports

What are the data types?

Visualizing data: bar charts

- Bar charts



Count of items in the dataset on the vertical axis

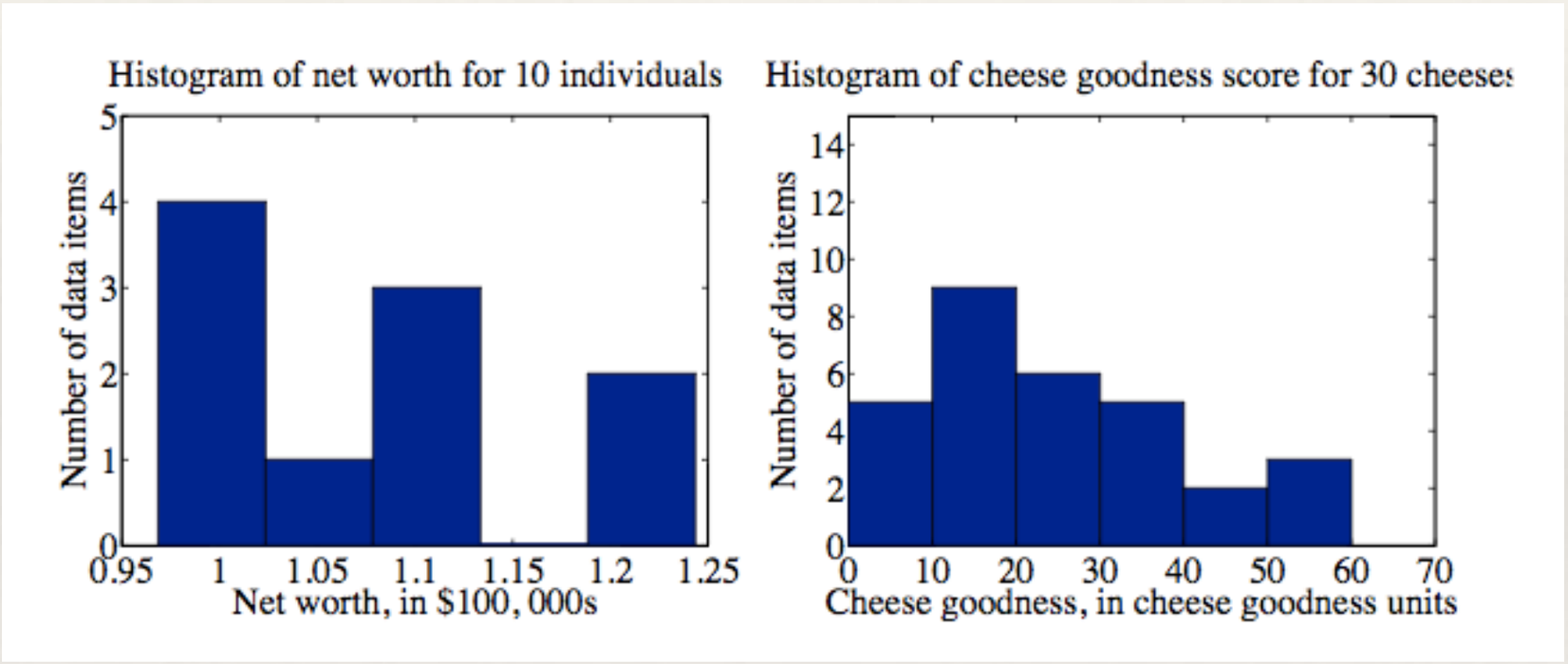
The values of one categorical variable on the horizontal axis

Visualizing data: histograms

- Histograms: a generalization that can handle numerical/continuous data
- Example: Bar chart of CS summer internship earnings
- Binning

Visualizing data: histograms

- Histograms



Making histograms

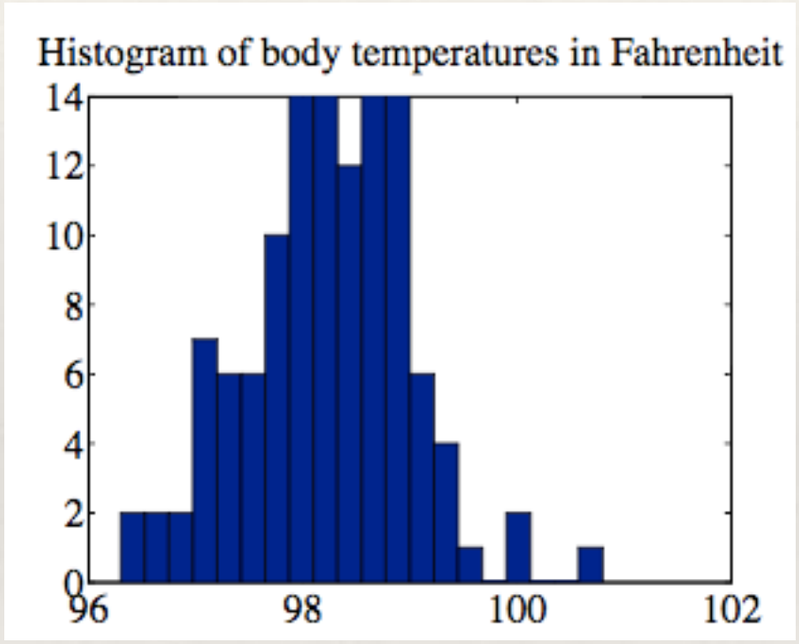
- If we want n intervals of the same size
- Interval size should be:

$$\frac{x_{max} - x_{min}}{n}$$

- Every data point should belong to exactly 1 interval
- So choose [0,1), [1,2), … or (0,1], (1, 2], …

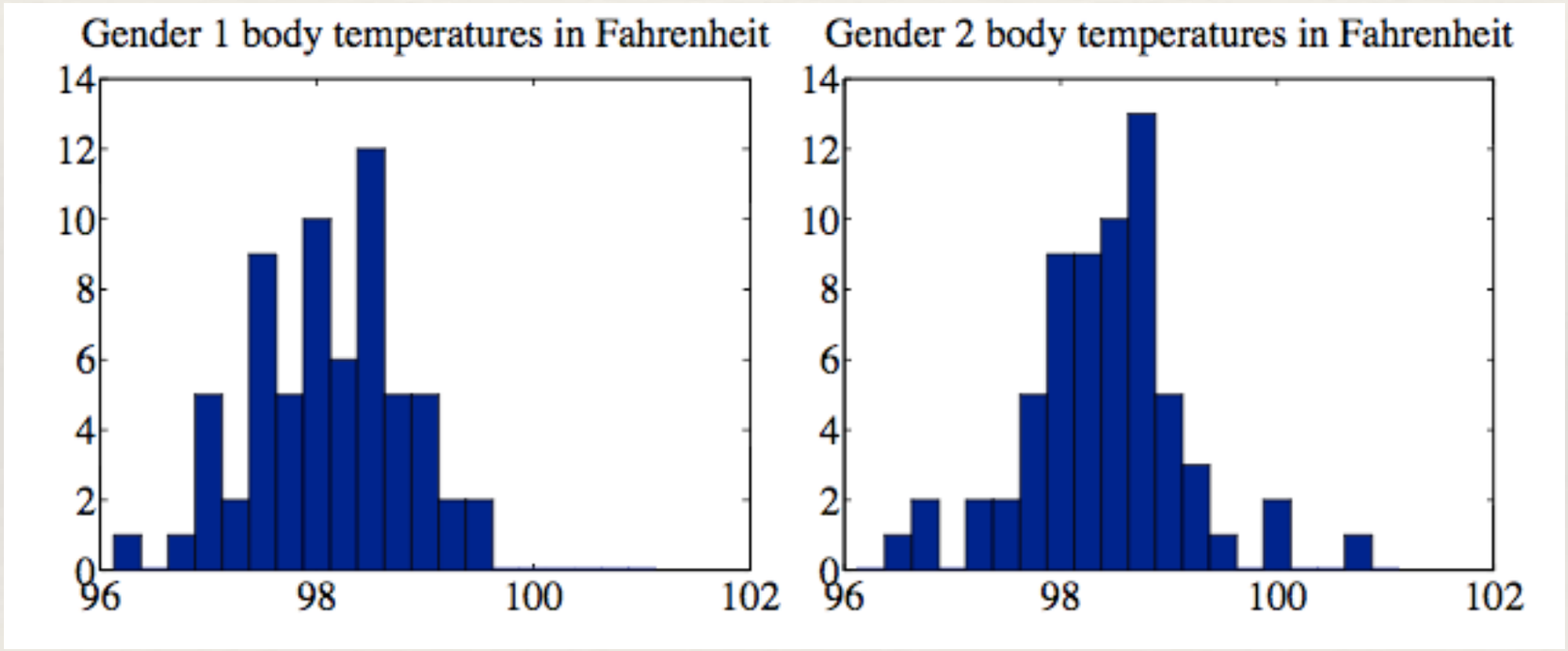
Conditional histograms

- What's happening



Conditional histograms

- Conditioning on another variable



Summarizing data

- Assume continuous data
- That can be added, subtracted, and multiplied by a constant and have a reasonable interpretation
- There are a number of "summary statistics" that might tell us something about the whole set of data
- Where is the data? How spread out is it?

The mean: definition

The mean, aka the average

Definition: 2.1 *Mean*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\text{mean}\left(\{x\}\right) = \frac{1}{N}\sum_{i=1}^{i=N} x_i.$$

A “location parameter” or “measure of central tendency”

The mean: properties

Scaling data scales the mean

$$\text{mean}\left(\{kx_i\}\right) = k\text{mean}\left(\{x_i\}\right).$$

Translating data translates the mean

$$\text{mean}\left(\{x_i + c\}\right) = \text{mean}\left(\{x_i\}\right) + c.$$

The sum of signed distances from the mean is 0

$$\sum_{i=1}^{N}(x_i - \text{mean}\left(\{x_i\}\right)) = 0.$$

The mean: properties

The mean is the number which minimizes the sum of squared distances from the data, i.e.

$$\arg\min_{\mu} \sum (x_i - \mu)^2 = \text{mean}\left(\{x_i\}\right)$$

which is to say it is “close” to the data in some optimal sense

Which is why it’s a good “location parameter”

Proof

- We want to find the number μ that minimizes

$$\sum_{i=1}^{N}(x_i - \mu)^2$$

Proof

$$\frac{d}{d\mu}\sum_{i=1}^{N}(x_i-\mu)^2=-2\sum_{i=1}^{N}(x_i-\mu)$$

$$-2\sum_{i=1}^{N}(x_i-\mu)=0$$

Proof

$$\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = 0$$

$$\sum_{i=1}^{N} x_i - N\mu = 0$$

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The mean

- A good single number summary for where data is located
- In a pinch, a good prediction for an unknown data item

Standard deviation

- How close are our data to the mean?
- Root of the mean of the squared distances

Definition: 2.2 *Standard deviation*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is is:

$$\text{std}\left(\{x_i\}\right) = \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \text{mean}\left(\{x\}\right))^2} = \sqrt{\text{mean}\left(\{(x_i - \text{mean}\left(\{x\}\right))^2\}\right)}.$$

Standard deviation

- A “scale parameter”
- How wide the spread of the data is
- Larger standard deviation means values much larger or smaller than the mean
- We can talk about a data item *j* being within *k* standard deviations of the mean

$$\text{abs}\left(x_j - \text{mean}\left(\{x\}\right)\right) \leq k\text{std}\left(\{x_i\}\right).$$

Properties of standard deviation

Translating the data does not change the standard deviation

$$\text{std}(\{x_i + c\}) = \text{std}(\{x_i\})$$

Scaling data scales the standard deviation

$$\text{std}(\{kx_i\}) = k\,\text{std}(\{x_i\})$$

For any dataset, there can only be a few items that are many standard deviations from the mean

For any dataset, there is at least one item that is at least one standard deviation from the mean

Standard deviation

For any dataset, there are at most $\frac{N}{k^2}$ items that are k standard deviations from the mean

To prove this, we will assume the mean is 0 which we can do since translating the data does nothing to the standard deviation as we said

Then we will construct a worst case dataset, with the largest fraction of data lying k or more standard deviations from the mean

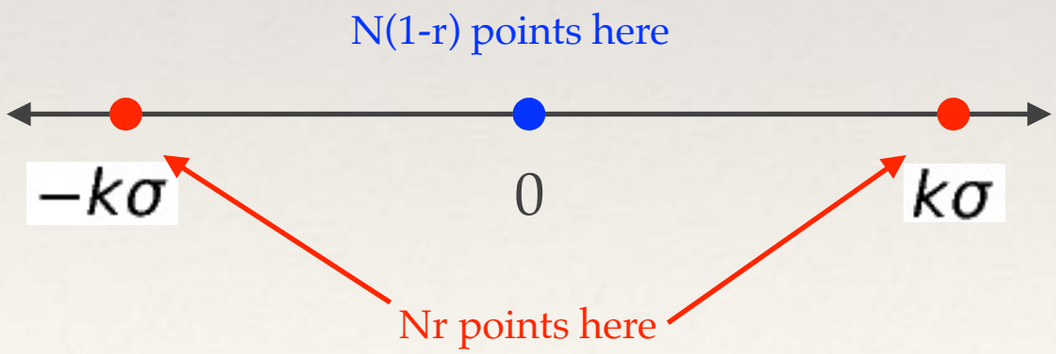
Proof

We will construct a dataset with mean 0 and standard deviation of σ

We want to construct a dataset with the largest possible fraction of points k or more standard deviations from the mean. Call this fraction r

The Nr points that are k or more standard deviations away should be exactly $k\sigma$ away or else we aren't getting as many points as we could

Likewise the other N(1-r) points should be at 0



Proof

So for our dataset with Nr points $k\sigma$ from the mean and N(1-r) 0 from the mean

We have Nr points where $(x_i - \mu)^2 = k^2\sigma^2$

And the other N(1-r) where the contribution the the standard deviation is 0

So our standard deviation for the whole dataset is

$$\sigma = \sqrt{\frac{Nrk^2\sigma^2}{N}}$$

Proof

Solving for r in

$$\sigma = \sqrt{\frac{Nrk^2\sigma^2}{N}}$$

We get

$$r = \frac{1}{k^2}$$

Proof

And since this was the maximum fraction of points we could have chosen to put *k* standard deviations away, we see that the fraction of points *r* that is at least k standard deviations from the mean is given by

$$r \leq \frac{1}{k^2}$$

Which is to say the largest number of points that could possibly be that far from the mean is

$$\frac{N}{k^2}$$

So what does that mean?

- This is true for any dataset
- So for any dataset we know that at most 100% of the data is 1 standard deviation away
- At most 25% is 2 standard deviations away
- At most 11% is 3 standard deviations away, etc.
- But the data must be very unusual to achieve even this, usually even less will be far from the mean