August 29, 2017

Probability & Statistics

Intro + Chapter 2

Welcome to CS361

- Me: Rick Barber.
  Office 4209 Siebel, 3-5PM Monday
- TAs: Edward McEnrue
  Office hours TBA
- Webpage (eventually): https://courses.engr.illinois.edu/cs361/fa2017/
- Textbook: https://tinyurl.com/cs361book
- Piazza

Grading

- 11 homeworks + take home final: 75% + 25%
- Homework: approx 1 per week
  Some programming in R
  Some allow work in groups

Course content

- Tools for dealing with data
- Probability and Random Variables
- Markov Chains and Simulation
- High dimensional data
- Machine Learning

Descriptive Statistics

- “What’s going on here?”
- Basic tools for visualizing and summarizing data

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 (an ordered list of d items)
- We will say our dataset {**x**} contains *N* items
- When we want to refer to *i*-th item, we will write x_i
- We will assume all of our data items have *d* dimensions even if some are blank

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

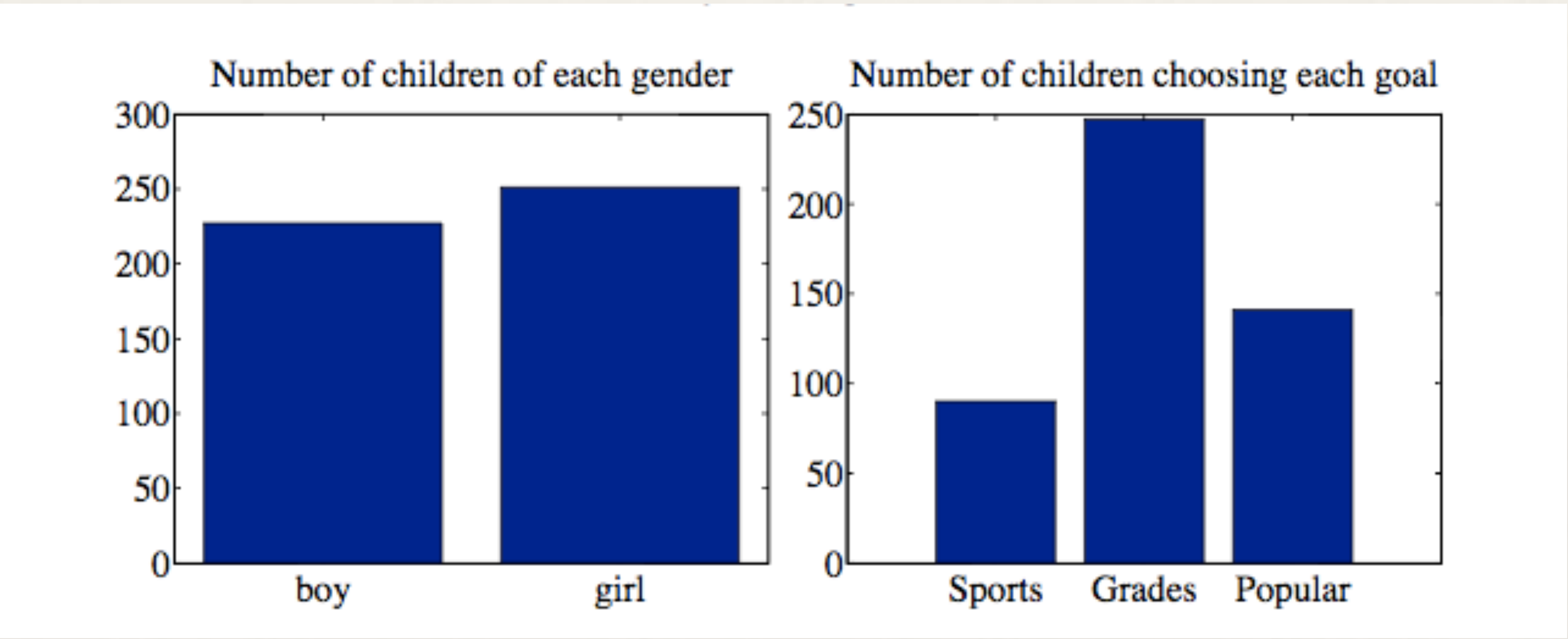
Index	Taste score	Index	Taste score
1	12.3	11	34.9
2	20.9	12	57.2
3	39	13	0.7
4	47.9	14	25.9
5	5.6	15	54.9
6	25.9	16	40.9
7	37.3	17	15.9
8	21.9	18	6.4
9	18.1	19	18
10	21	20	38.9

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

Visualizing data: bar charts

- Bar charts

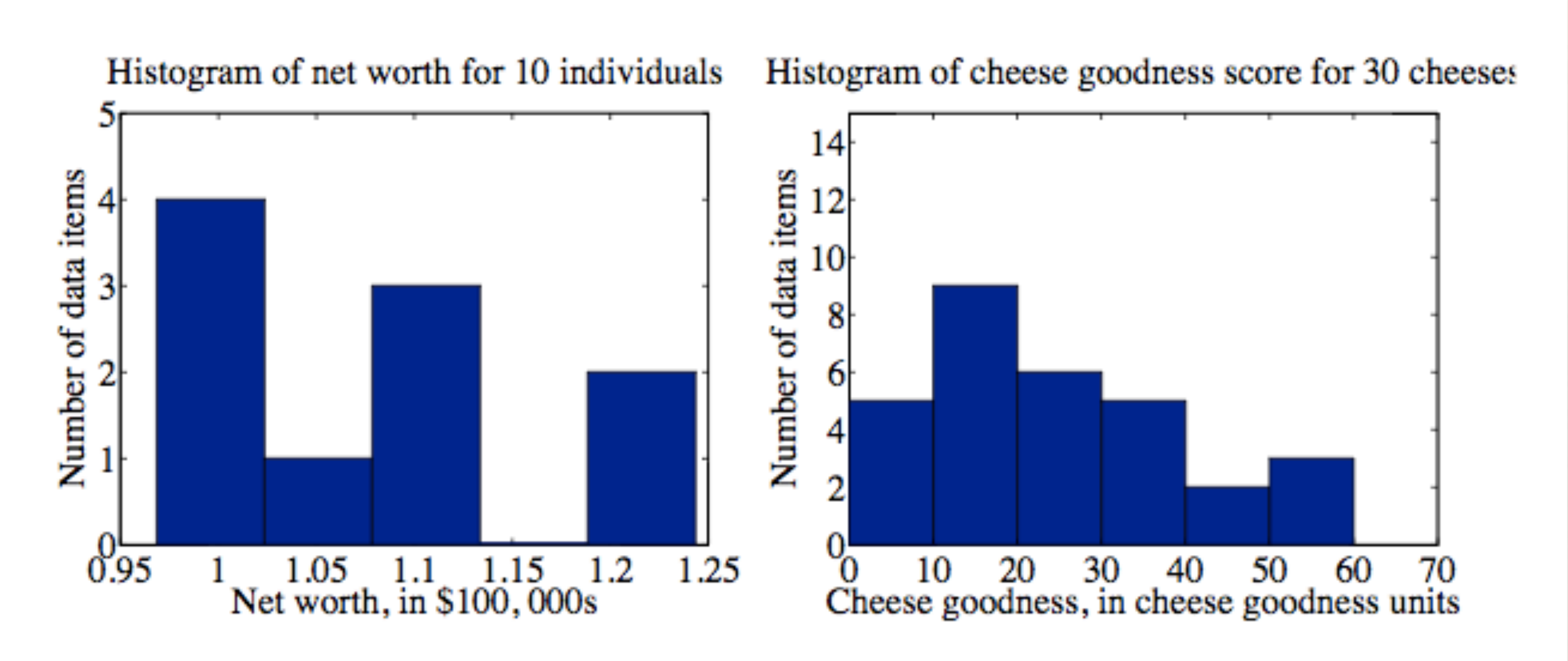


Visualizing data: histograms

- Histograms: a generalization
- Bar chart of CS summer internship earnings
- Prefer fewer and taller bars
- Solution: binning

Visualizing data: histograms

- Histograms



Making histograms

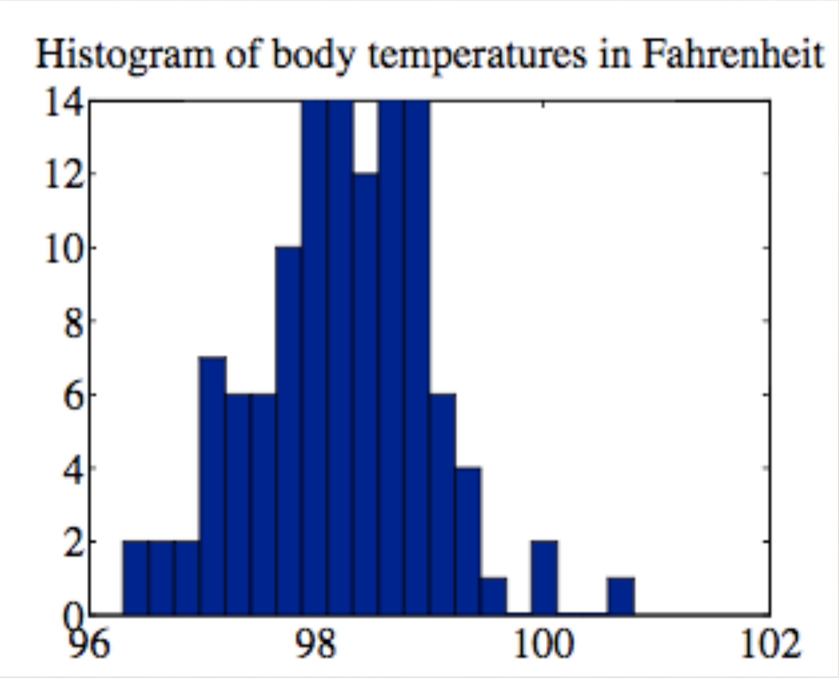
- If we want *n* intervals of the same size
- Interval size should be:

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

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

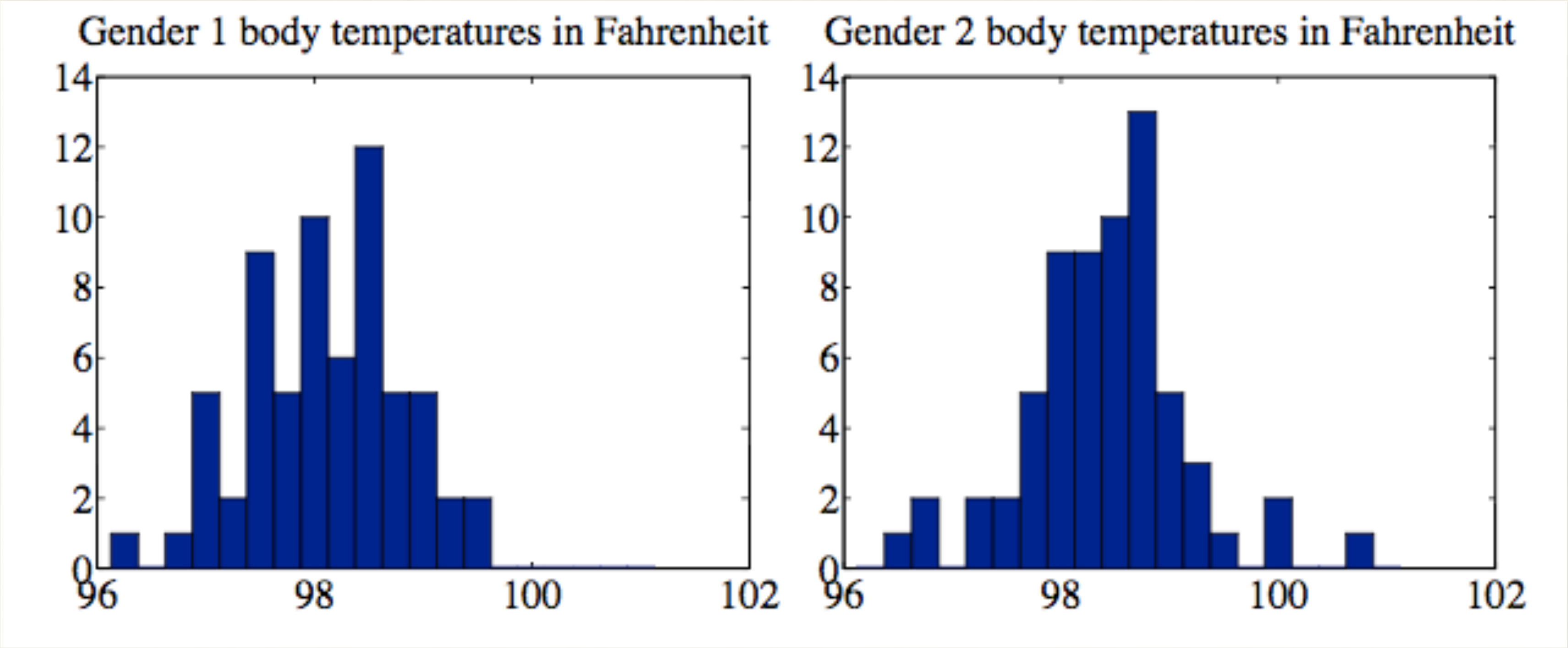
Conditional histograms

- 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

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

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

- A "location parameter" or "measure of central tendency"

The mean

- Scaling data scales the mean: $\text{mean}(\{kx_i\}) = k\text{mean}(\{x_i\})$.
- Translating data translates the mean:

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

- The sum of signed distances from the mean is 0:

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

The mean

- The number which minimizes the "sum of squared distances" from the data

$$\underset{\mu}{\arg\min} \sum_i (x_i - \mu)^2 = \text{mean}\left(\{x_i\}\right)$$

- Which is why it's a good "location parameter"
- It is close to the data in a meaningful way

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
- 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:

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

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 $k\sigma$ 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 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, we see that the fraction of points 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