August 31, 2017

CS 361: Probability & Statistics

Chapter 2

Administrative

- The web page is up at https://courses.engr.illinois.edu/cs361/
- Homework 1 has been posted and will be due September 11 at midnight
- Submissions should be made on compass, let me know if you haven't been added to compass

Recall - Standard deviation

- How close are our data, in some average sense, to the mean of the data?

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:

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

- Note that our formula requires us to know the mean of *{x}*

Root of the mean of the squared distance

- Squared distance from point *i* to the mean

$$d_i = (x_i - \text{mean}(\{x\}))^2$$

- These *N* distances form a dataset *{d}*, with mean

$$\text{mean}(\{d\}) = \frac{1}{N}\sum_{i=1}^{N} d_i$$

- So $\text{std}(\{x\}) = \sqrt{\text{mean}(\{d\})}$

Reminder

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

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
- Recall that translating the data doesn't change the standard deviation. So for our proof, we assume the mean is zero since we can create a mean zero dataset from any dataset by subtracting the mean from each item
- Then we will construct a worst case dataset, with the largest fraction of data lying k or more standard deviations from the mean

Proof

- Let's call the standard deviation of this dataset 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 from the mean
- 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)* which are 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

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

- 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

Variance

- Later, we will expand many of our definitions to include higher dimensional data
- The square of the standard deviation, or the **variance**, will allow us to do this more easily

Definition: 2.3 *Variance*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

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

Variance

- If we approximated each data item with the mean, the squared error of that approximation would be the variance

- Properties of variance

Definition: 2.3 *Variance*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

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

Useful Facts: 2.3 *Variance*

- $\text{var}(\{x + c\}) = \text{var}(\{x\})$.
- $\text{var}(\{kx\}) = k^2\text{var}(\{x\})$.

Outliers

- In walks a billionaire

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

- Suddenly our mean is around $90M
- It's more useful to think of this dataset as a bunch of people plus a billionaire rather than a group whose average net worth is $90M

Median

- The median

Definition: 2.4 *Median*

The median of a set of data points is obtained by sorting the data points, and finding the point halfway along the list. If the list is of even length, it's usual to average the two numbers on either side of the middle. We write

$$\text{median}(\{x_i\})$$

for the operator that returns the median.

- Ex: median({100,250, 1000}) = 250
- Ex: median of net worths, before the billionaire = $107,835
- After = $108,930
- Adding the billionaire had a big effect on mean, very little on median

Median

- A location parameter more robust in the presence of outliers
- With similar characteristics to the mean

Useful Facts: 2.4 *Median*

- median $(\{x + c\}) =$ median $(\{x\}) + c$.
- median $(\{kx\}) = k$median $(\{x\})$.

Outliers and scale

- Outliers also have an effect on standard deviation and variance
- Standard deviation without the billionaire is $9265, with is $3.014*10^8
- Standard deviation did what it was supposed to, inform us that there's large variation but it's not a great description of the data really in this case

Percentiles

- Percentiles

Definition: 2.5 *Percentile*

The k'th percentile is the value such that $k\%$ of the data is less than or equal to that value. We write percentile$(\{x\}, k)$ for the k'th percentile of dataset $\{x\}$.

- Quartiles

Definition: 2.6 *Quartiles*

The first quartile of the data is the value such that 25% of the data is less than or equal to that value (i.e. percentile$(\{x\}, 25)$). The second quartile of the data is the value such that 50% of the data is less than or equal to that value, which is usually the median (i.e. percentile$(\{x\}, 50)$). The third quartile of the data is the value such that 75% of the data is less than or equal to that value (i.e. percentile$(\{x\}, 75)$).

Interquartile range

Definition: 2.7 *Interquartile Range*

The interquartile range of a dataset $\{x\}$ is $\text{iqr}\{x\} = \text{percentile}(\{x\}, 75) - \text{percentile}(\{x\}, 25)$.

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

- 25th percentile: $100,360
- 75th percentile: $112,710
- IQR =$12,350
- With billionaire, IQR = $19,810

Interquartile range

- Scale that can handle some outliers
- Similar properties to variance

Useful Facts: 2.5 *Interquartile range*

- $\text{iqr}\{x + c\} = \text{iqr}\{x\}$.
- $\text{iqr}\{kx\} = k^2\text{iqr}\{x\}$.

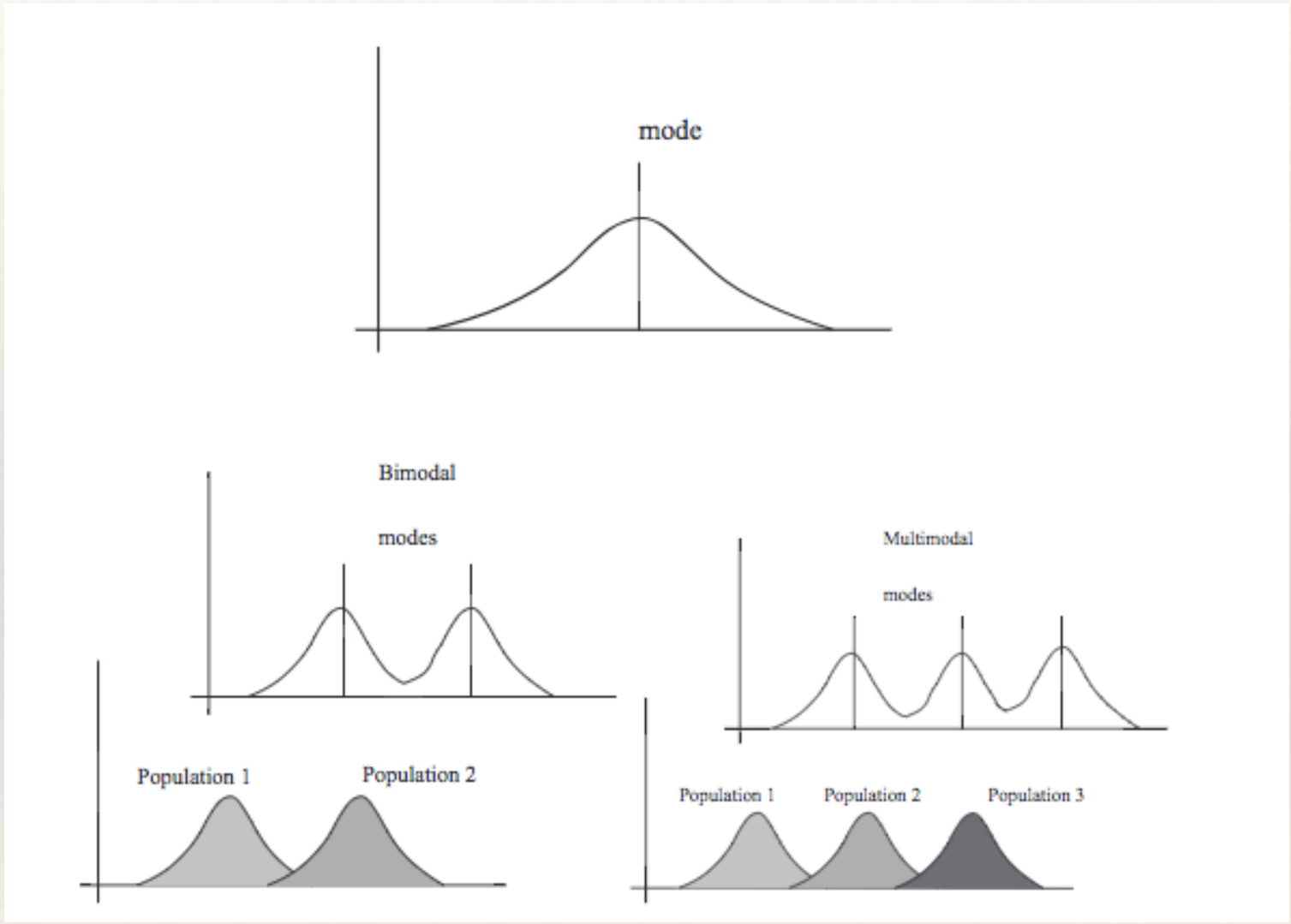
- **Correction**: iqr{kx} = k iqr{x}

A couple of notes on summary statistics

- Mean of effectively categorical numerical values: 2.6 children
- Precision: Average pregnancy of 32.833 weeks (to within 10 minutes)
- Often useful to calculate mean/median together. If they are very different, this is interesting

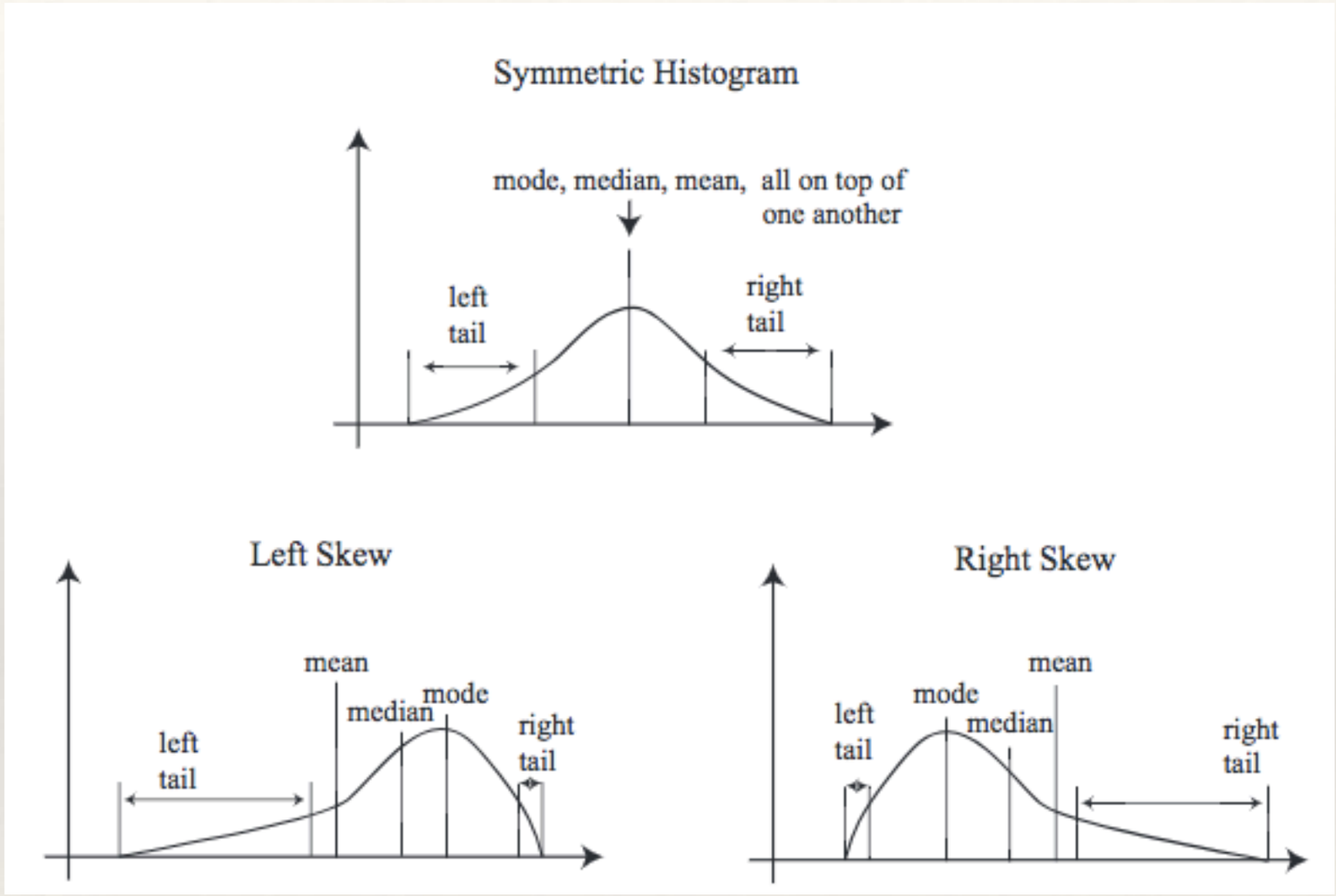
Modes

- Modes: peaks on histograms
- More than one mode could mean more than one population in the data



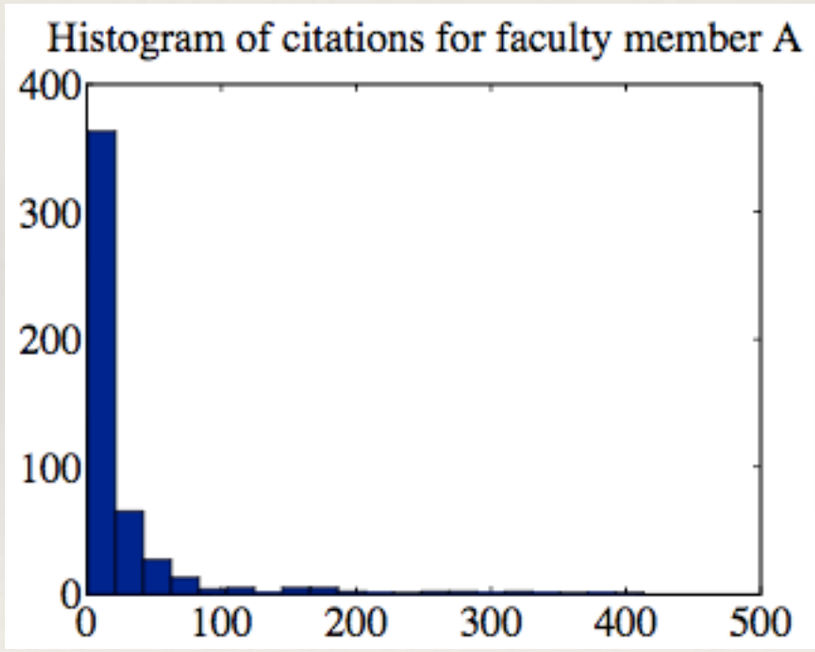
Tails and skew

- Tails: uncommon values (relatively) that are much larger or smaller than the mode



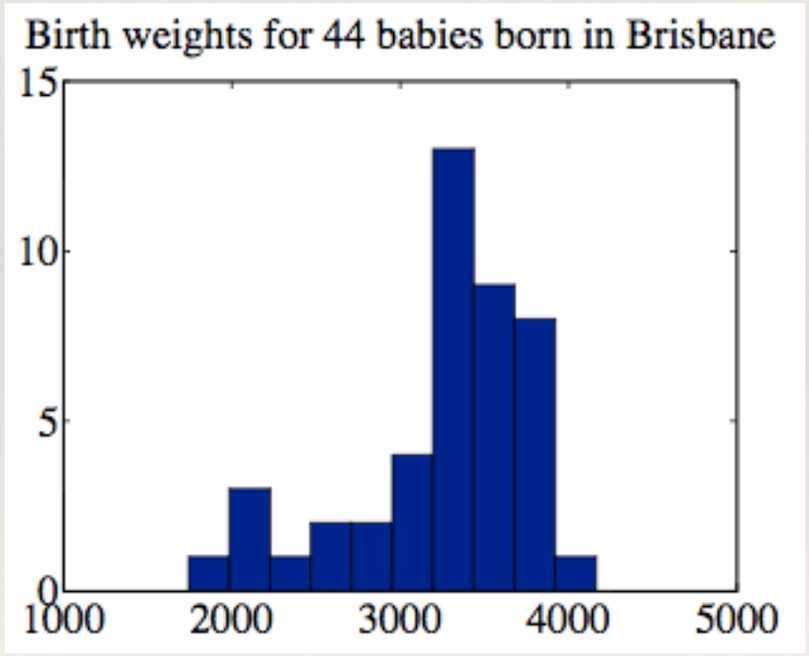
Skew

- Citations of scholars are right-skewed:
  x-axis is the number of citations, y-axis is the count of scholars with that many citations
- Mean: 24.7
- Median: 7.5



Skew

- Modern birth weights tend to be left-skewed
- Why might this be? Where are the missing heavy babies?



Skew

- Data is often but not always skewed due to real-world constraints
- Baby birth is one example
- Incomes are another since most people are paid a positive salary
- Grades on exams are as well since there is a maximum and minimum score

Standard coordinates

- We might like to compare two histograms that have different units or may measure entirely different quantities
- If we wanted to compare a histogram of internship earnings and GPA to see if there's something similar about the histograms, how to proceed?

Standard coordinates

- A transformation of the dataset $\{x\}$
- Create a new dataset $\{\hat{x}\}$ with standardized location — subtract the mean from each data item
- And standardized scale — divide each data item by the standard deviation
- $\{\hat{x}\}$ is dimensionless, has 0 mean, and unit standard deviation

Standard coordinates

- Recipe for computing standard coordinates from your dataset

Definition: 2.8 *Standard coordinates*

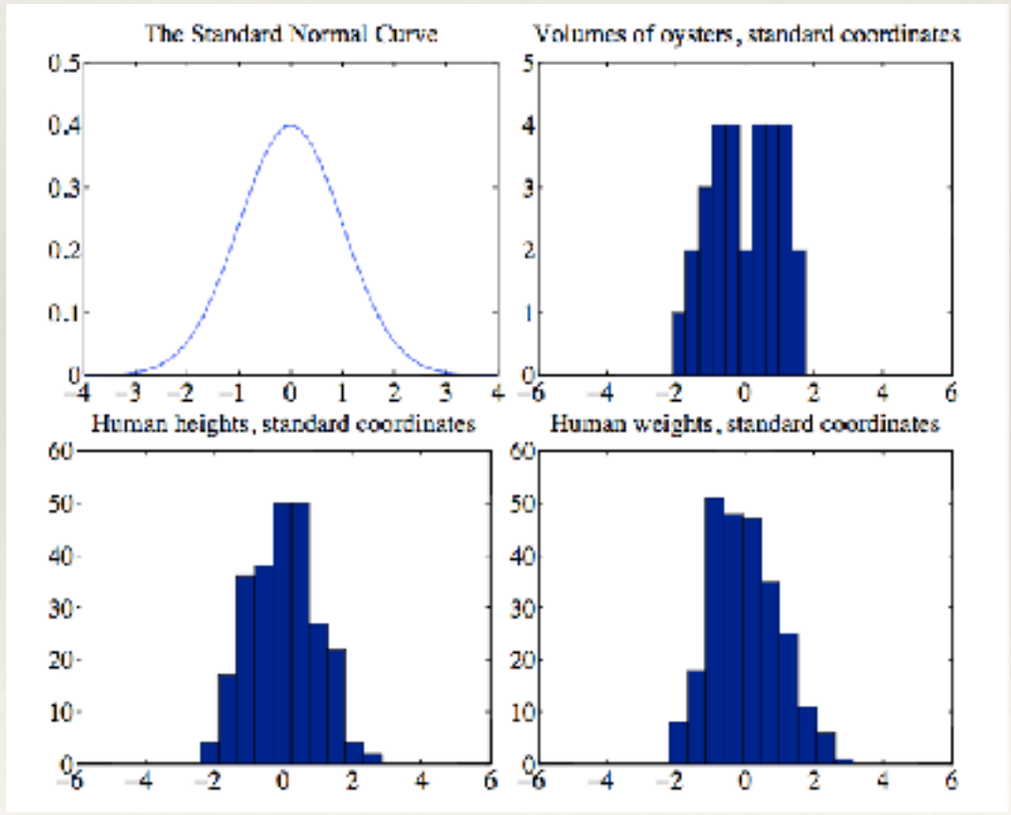
Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \text{mean}(\{x\}))}{\text{std}(\{x\})}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Standard normal data

- A wide variety of data, when standardized, will have a particular look and even fit a particular mathematical curve. What features do these histograms have?



Definition: 2.9 *Standard normal data*

Data is standard **normal data** if, when we have a great deal of data, the histogram is a close approximation to the **standard normal curve**. This curve is given by

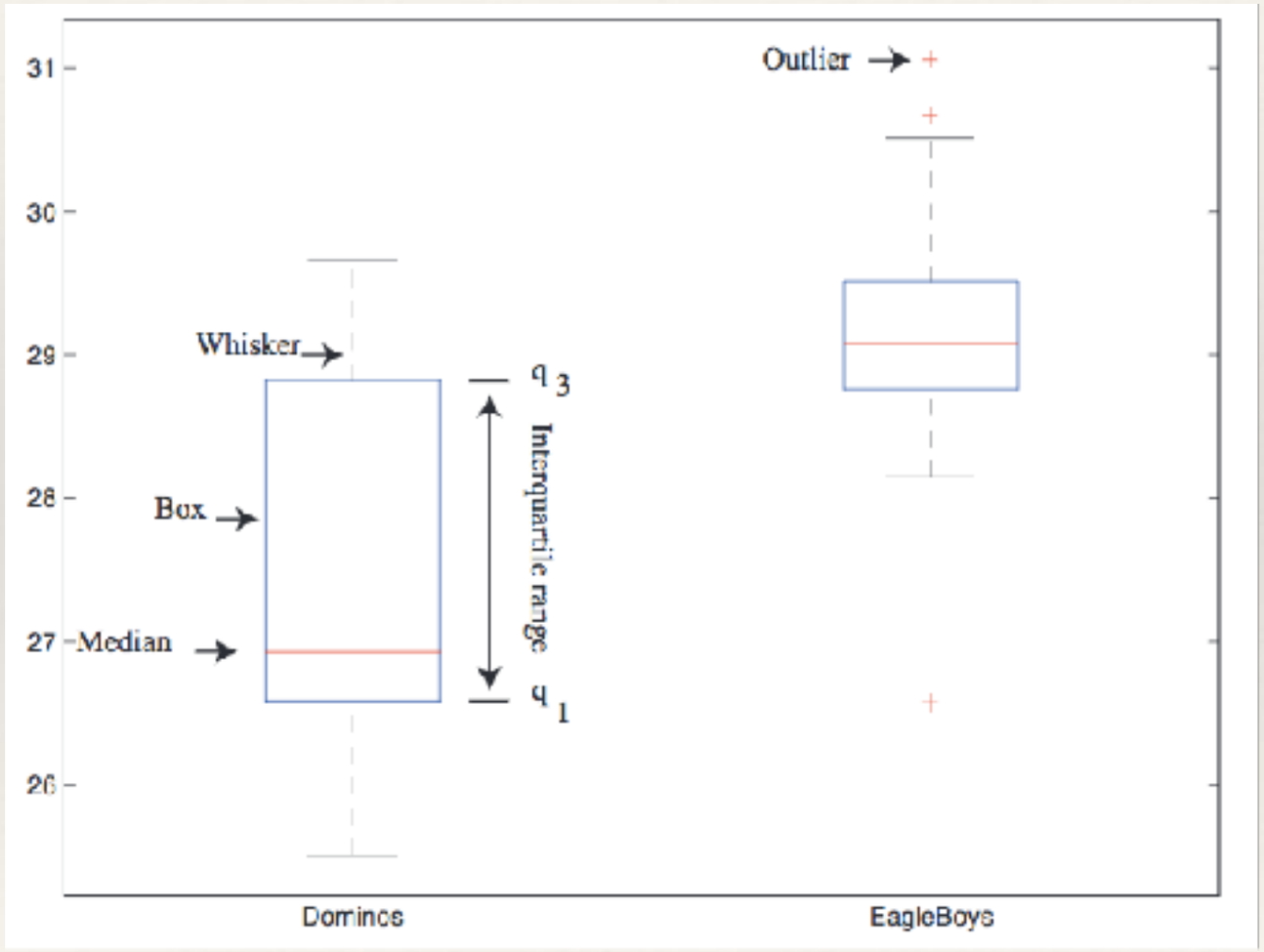
$$y(x) = \frac{1}{\sqrt{2\pi}} e^{(-x^2/2)}$$

Definition: 2.10 *Normal data*

Data is **normal data** if, when we subtract the mean and divide by the standard deviation (i.e. compute standard coordinates), it becomes standard normal data.

Boxplots

- Another type of visualization



How to make a box plot

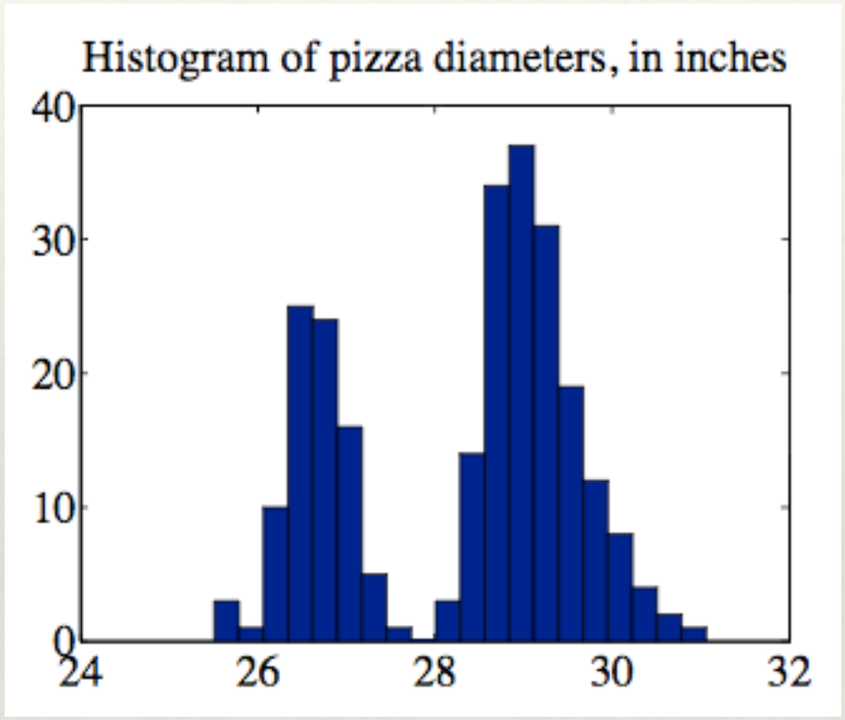
- Height of the box is from q1 to q3, width is whatever makes it look nice
- Identify the median
- Use a rule for outliers: bigger than q3 + 1.5(q3-q1) or smaller than q1 - 1.5(q3-q1) for example
- Whiskers extend to the largest data item which isn't an outlier and smallest data item which isn't an outlier
- Outlier data points indicated

Pizzas

- We have a dataset containing data about large pizzas from two different pizza chains
- The data is a set of 3-tuples: (chain, diameter, crust type)

Pizzas

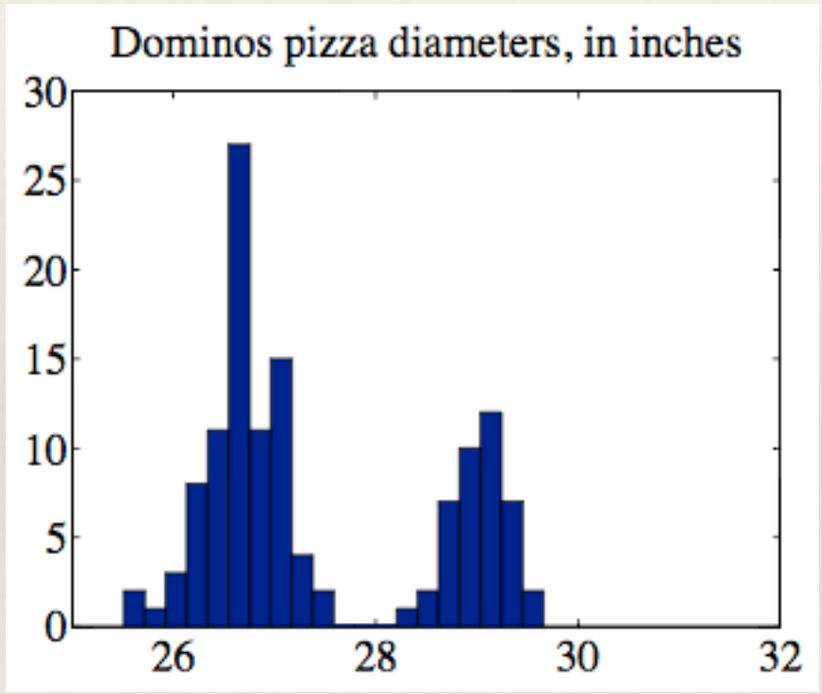
- Diameters of all pizzas

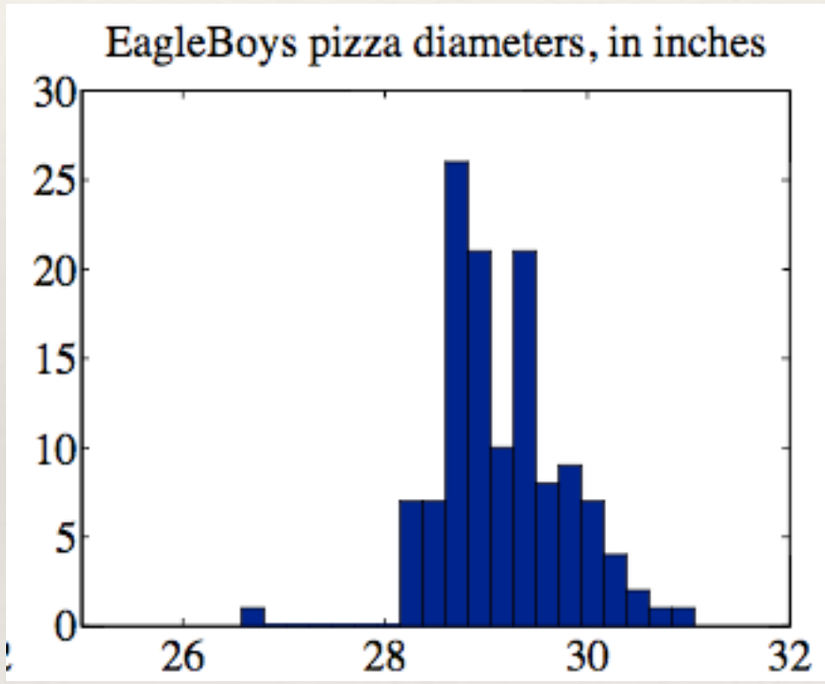


- What are the properties of this histogram?

Pizas

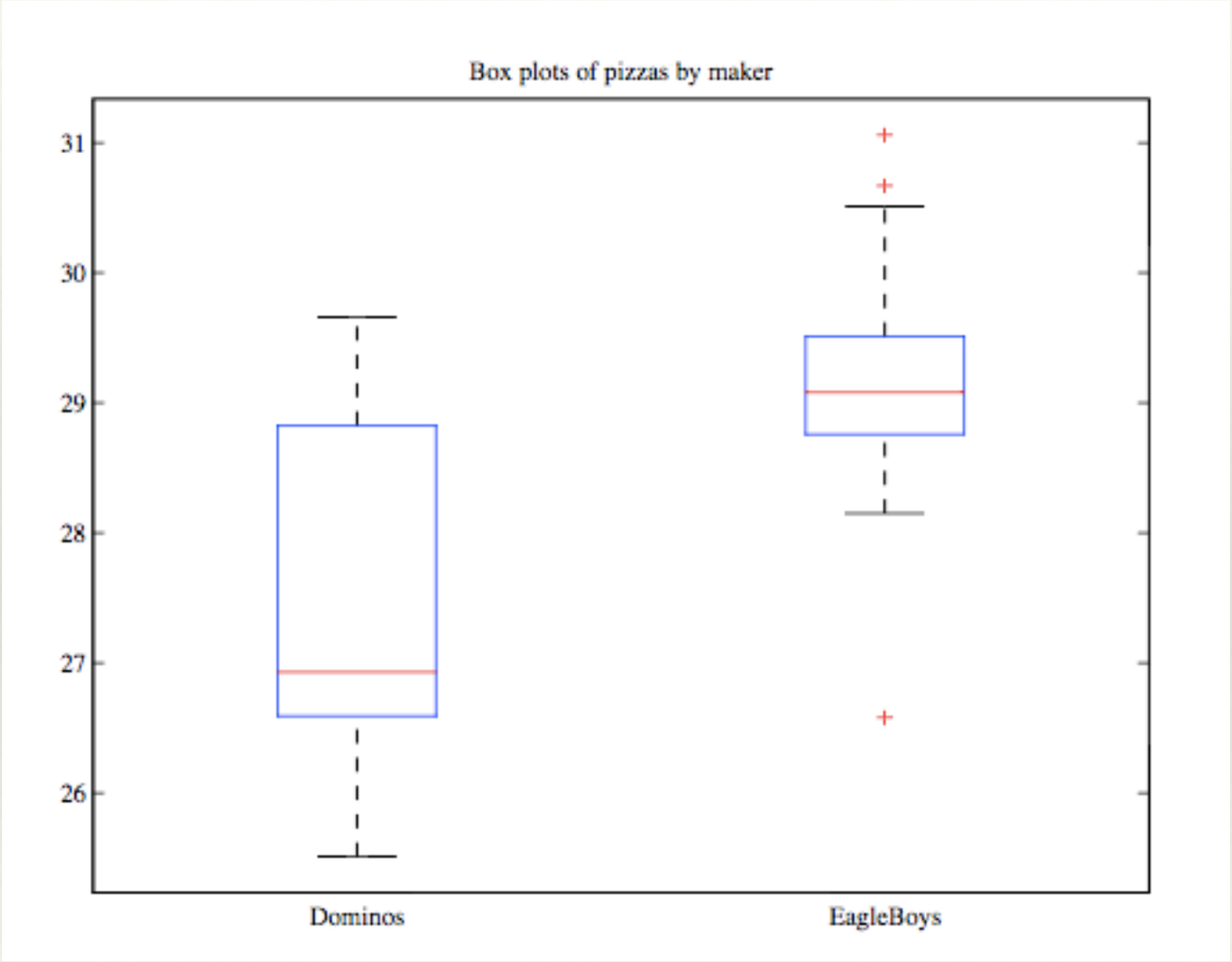
- Class conditional pizza data





Pizzas

- Boxplots



Pizzas

- More conditioning