Statistics for Machine Learning and Big Data

An Introduction

Part 1: a probability and statistics primer

Jeff Abrahamson

7 April 2015

Probability

Review

- Counting
- Urns (sampling)
- Area

Words

Disjoint (mutually exclusive)

Words

Complement of an event (set)

Words

Independence

Words

Conditional probability

Words

Marginal probability

Words

Joint probability

Words

Conditional probability

Example

In any year, about .35% of the population is diagnosed with dread disease *D*.

The test for *D* isn't perfect:

- For people with *D*, 11% receive a (false) negative test result.
- For people without *D*, 7% receive a (false) positive test result.

We test person λ at random. The test says positive. What is the probability that λ has *D*?

Probability Density Function

```
import scipy.stats as ss
x = np.linspace(ss.norm.ppf(.01),
    ss.norm.ppf(.99), 100)
plt.plot(x, ss.norm.pdf(x))
plt.show()
```

Cumulative Density Function

```
import scipy.stats as ss
x = np.linspace(ss.norm.ppf(.01),
    ss.norm.ppf(.99), 100)
plt.plot(x, ss.norm.cdf(x))
plt.show()
```

Counting

Example: two children, what is $P(\exists \mathrm{boy}) = P(\exists b)$?

Pólya urn model

r red balls, b black balls in an urn.

Pólya urn model

r red balls, *b* black balls in an urn.

- Sampling with replacement and duplication (rich get richer)
- Variations: balls into boxes
  - Sampling with replacement
  - Sampling without replacement

Pólya urn model

r red balls, *b* black balls in an urn.

- Sampling with replacement and duplication (rich get richer)
- Variations: balls into boxes
  - Sampling with replacement
  - Sampling without replacement

Want to know sequence of colours selected, evolution of colour distribution in urn.

Sampling

- We don't know r, b, or even r/b. Estimate.

Area

- Continuous case.
- Events are regions.
- Same constraints: sum of areas is unity. Cover and disjoint.
- Zero probability events!

Distributions

Soon. . .

Statistics

What is statistics?

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

What is statistics?

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Sadly, sometimes people forget 1.

What is statistics?

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Statistics is about making 2–4 efficient, rigorous, and meaningful.

What is statistics?

1. Identify a question or problem.
2. Collect relevant data on the topic.
3. Analyze the data.
4. Form a conclusion.

Statistics is about making 2–4 efficient, rigorous, and meaningful.

OpenIntro Statistics, 2nd edition, D. Diez, C. Barr, M. Çetinkaya-Rundel, 2013.

What is data science?

(Exercise: Is this the same question as the last slide?)

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

What is data science?

(Exercise: Is this the same question as the last slide?)

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

What the public thinks.

What is data science?

(Exercise: Is this the same question as the last slide?)

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

Where we spend most of our time.

What is data science?

(Exercise: Is this the same question as the last slide?)

1. Define the question of interest
2. Get the data
3. Clean the data
4. Explore the data
5. Fit statistical models
6. Communicate the results
7. Make your analysis reproducible

The easiest part to forget.

What is data science?

(Exercise: Is this the same question as the last slide?)

http://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-that-is-a-big-problem-f

Significance

- Flip a coin N times. Get 49% heads. Is the coin fair?
- Congress/parliament has N members of whom m are male. Do we discriminate against women?
- Do we discriminate against women more or less than we discriminate against immigrants?
- Email classification: spam or not?

Significance

- Flip a coin N times. Get 49% heads. Is the coin fair?
- Congress/parliament has N members of whom m are male. Do we discriminate against women?
- Do we discriminate against women more or less than we discriminate against immigrants?
- Email classification: spam or not?
- Careful about what we conclude. Here we can't conclude anything about cause or source.

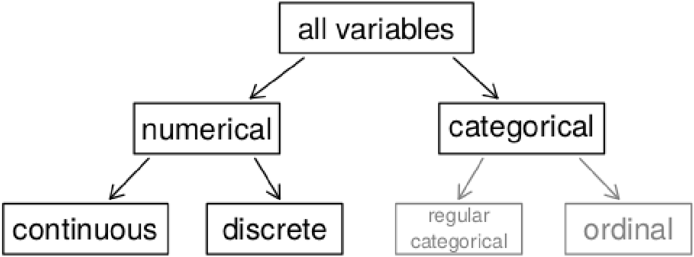
Words

A **case** or observational unit is an observation or a single sample. E.g., an email.

Words

A **variable** is a characteristic of an observational unit. E.g., number of lines, mean line length, number of words.

Words



Categorical is sometimes called nominal. Categorical variables with logical ordering (empty, half, full) or called ordinal.

Anecdote

Some properties of anecdote:

- is data
- haphazardly collected
- is generally not representative
- sometimes result of selective retention
- does not accumulate to be representative
- might be true (by chance)
- is ok to use as hypothesis, but be clear that hypothesis is anecdote

Study Types

- Observational
- Experimental

Study Types

- Observational
- Experimental

What can go wrong?

- Forgetting that association $\neq$ causation
- Not random
- Confounding variables

Observational studies

Can't generally conclude causation.

Experiment

We do stuff. Maybe randomised experiment.

Experiment

We do stuff. Maybe randomised experiment.

- Controlling
- Randomisation
- Replication
- Blocking (optional, advanced)

Experiment

We do stuff. Maybe randomised experiment.

- H_0: Independent (or null) hypothesis
- H_A: Alternative hypothesis

Simple random sampling

Uniform and independent.

Simple random sampling

Uniform and independent.

What can go wrong:

- Not actually random
- Convenience sample
- Non-response bias

Stratified Sampling

Group like with like, then usually simple random sampling of each stratum.

Stratified Sampling

Group like with like, then usually simple random sampling of each stratum.

What will go wrong:

- Harder to analyse than simple random sampling.

Cluster Sampling

Simple random sampling to form clusters, then simple random sampling.

Cluster Sampling

Simple random sampling to form clusters, then simple random sampling.

What will go wrong:

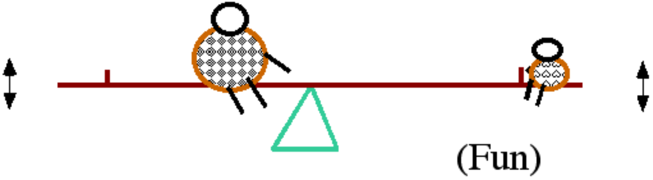
- Harder to analyse than simple random sampling.

Mean

- Weighted and unweighted
- Centroid to physicists

Mean

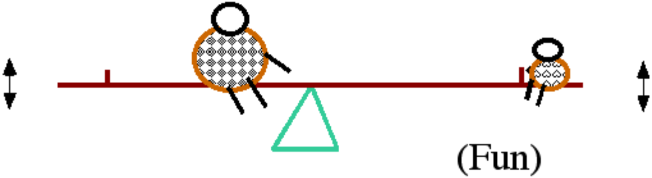
- Weighted and unweighted
- Centroid to physicists



Copyright 2015 Jeff Abrahamson, for private use by course students only

Mean

- Weighted and unweighted
- Centroid to physicists

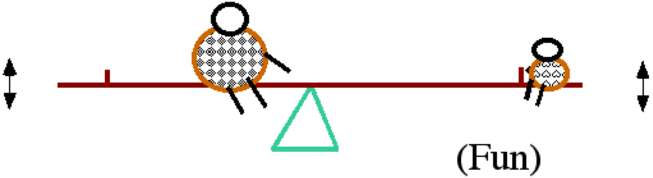


$$\mu = E(X) = \sum w_i x_i = \mathbf{w} \cdot \mathbf{x}$$

`http://telescopes.stardate.org/images/research/teeter-totter/TT4.gif`

Mean

- Weighted and unweighted
- Centroid to physicists



$$\mu = E(X) = \sum \Pr(X = x_i) x_i$$

http://telescopes.stardate.org/images/research/teeter-totter/TT4.gif

Mean

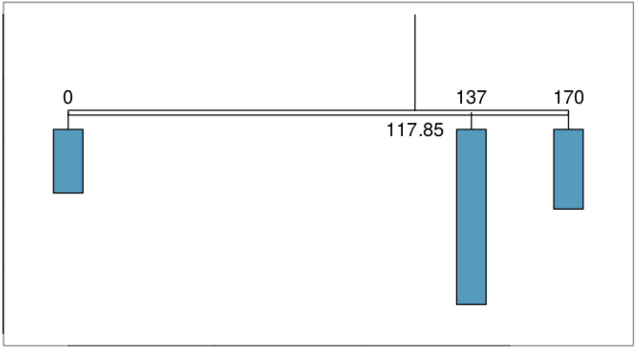
- Weighted and unweighted
- Centroid to physicists

$$\mu = E(X) = \int x f(x)\,\mathrm{d}x$$

http://telescopes.stardate.org/images/research/teeter-totter/TT4.gif

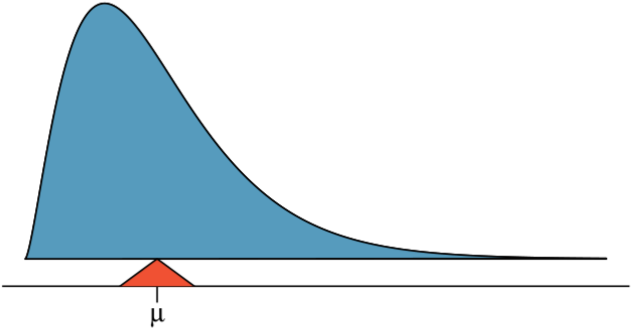
Mean

- Weighted and unweighted
- Centroid to physicists



Mean

- Weighted and unweighted
- Centroid to physicists



Variation

Deviation is distance from mean.

Variation

Variance is mean square of deviations

Variation

Standard deviation is square root of variance

Variation

$$s^2 = \frac{(\overline{x} - x_1)^2 + \cdots (\overline{x} - x_n)^2}{n - 1}$$

Variation

$$\sigma^2 = \frac{(\overline{x} - x_1)^2 + \cdots (\overline{x} - x_n)^2}{n}$$

Variation

$$\mathrm{Var}(X) = \sigma^2 = (\overline{x} - x_1)^2 \Pr(X = x_1) + \cdots (\overline{x} - x_n)^2 \Pr(X = x_n)$$

Linear combinations of random variables

Assuming *X* and *Y* are independent:

$$\mathbf{E}(aX + bY) = a\mathbf{E}(X) + b\mathbf{E}(Y)$$

$$\mathbf{Var}(aX + bY) = a^2\mathbf{Var}(X) + b^2\mathbf{Var}(Y)$$

Games

Multiple choice exam, 10 questions, all have four choices of which precisely one is correct. Guess at random.

- Probability of getting 100%?
- Probability of getting 0%?
- Expected score?
- Standard deviation

We'll come back to: if student gets 9 of 10 correct, do we really believe that they guessed at random?

Games

Multiple choice exam, 10 questions, all have four choices of which precisely one is correct. Guess at random.

- Probability of getting 100%? $\left(\frac{1}{4}\right)^{10}$
- Probability of getting 0%?
- Expected score?
- Standard deviation

We'll come back to: if student gets 9 of 10 correct, do we really believe that they guessed at random?

Games

Multiple choice exam, 10 questions, all have four choices of which precisely one is correct. Guess at random.

- Probability of getting 100%? $\left(\frac{1}{4}\right)^{10}$
- Probability of getting 0%? $\left(\frac{3}{4}\right)^{10}$
- Expected score?
- Standard deviation

We'll come back to: if student gets 9 of 10 correct, do we really believe that they guessed at random?

Games

Multiple choice exam, 10 questions, all have four choices of which precisely one is correct. Guess at random.

- Probability of getting 100%? $\left(\frac{1}{4}\right)^{10}$
- Probability of getting 0%? $\left(\frac{3}{4}\right)^{10}$
- Expected score? $\mathbf{E}(X) = 10\left(\frac{1}{4}\right) = \frac{5}{2}$
- Standard deviation

We'll come back to: if student gets 9 of 10 correct, do we really believe that they guessed at random?

Games

Multiple choice exam, 10 questions, all have four choices of which precisely one is correct. Guess at random.

- Probability of getting 100%? $\left(\frac{1}{4}\right)^{10}$
- Probability of getting 0%? $\left(\frac{3}{4}\right)^{10}$
- Expected score? $\mathbf{E}(X) = 10\left(\frac{1}{4}\right) = \frac{5}{2}$
- Standard deviation $\mathbf{Var}(X) = 10\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) = \frac{30}{16} = \frac{15}{8}$

We'll come back to: if student gets 9 of 10 correct, do we really believe that they guessed at random?

Experimental design

A telephone survey company dials digits at random (some of them aren’t valid telephone numbers) rather than using a phone directory.

Experimental design

A statistics student wants to determine if social network usage among his peers (students at his school) affects academic performance. Possibilities:

- He messages his class snapchat group.
- He posts paper fliers at his university.
- He gets a list of students from the registrar. (What about opt-in/opt-out policies? What if he doesn't know how many have opted out/failed to opt in?
- Other?

Experimental design

Due to a flu epidemic, 10% of students in a class miss the (final) exam. The professor decides to offer a make-up exam but to require 11% of the students who took the original exam to take the make-up exam as well in order to calibrate the new exam and make results comparable.

Reasoning about data

Students at an elementary school are given a questionnaire that they are required to return after their parents have completed it. One of the questions asked is, “Do you find that your work schedule makes it difficult for you to spend time with your kids after school?” Of the parents who replied, 85% said “no”. Based on these results, the school officials conclude thata great majority of the parents have no difficulty spending time with their kids after school.

Reasoning about data

A survey is conducted on a simple random sample of 1,000 women who recently gave birth, asking them about whether or not they smoked during pregnancy. A follow-up survey asking if the children have respiratory problems is conducted 3 years later, however, only 567 of these women are reached at the same address. The researcher reports that these 567 women are representative of all mothers.

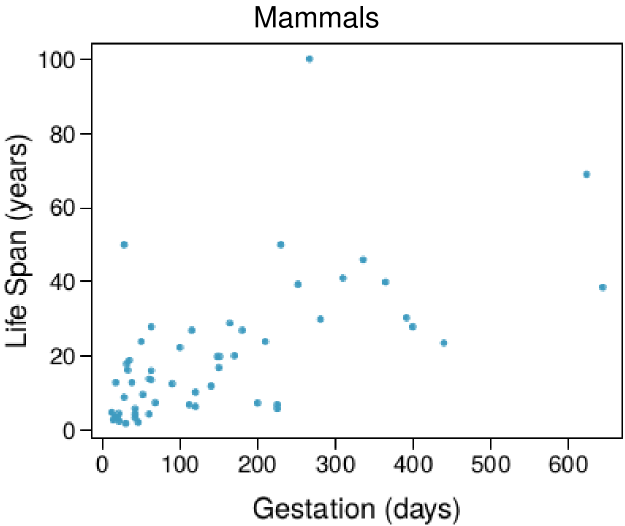
Reasoning about data

An orthopedist administers a questionnaire to 30 of his patients who do not have any joint problems and finds that 20 of them regularly go running. He concludes that running decreases the risk of joint problems.

Reasoning about data

The news reports that 27 year old Andreas Lubitz, co-pilot of crashed flight 4U 9525, had researched suicide before killing himself.

Example


Mammals
Life Span (years)
0
20
40
60
80
100
Gestation (days)
0
100
200
300
400
500
600

Questions?

`purple.com/talk-feedback`

Break