ML Week

0x01 Introduction

Jeff Abrahamson

2–5 novembre 2015

Definition

Supervised

Unsupervised

Reinforcement

Curse of Dimensionality

Addition rule: independent events

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

Addition rule: dependent events

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Multiplication rule: independent events

$$\Pr(A \cap B) = \Pr(A)\Pr(B)$$

Multiplication rule: dependent events

$$\Pr(A \cap B) = \Pr(A \mid B) \Pr(B)$$

Conditional probability

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

Conditional probability

$$\cup_i A_i = A \quad \wedge \quad A_i \cap A_j = \varnothing \implies$$

$$P(A_1 \mid B) = \frac{\Pr(B \mid A_1)\Pr(A_1)}{\sum_i \Pr(B \mid A_1)\Pr(A_1) + \cdots + \Pr(B \mid A_k)\Pr(A_k)}$$

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.

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

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

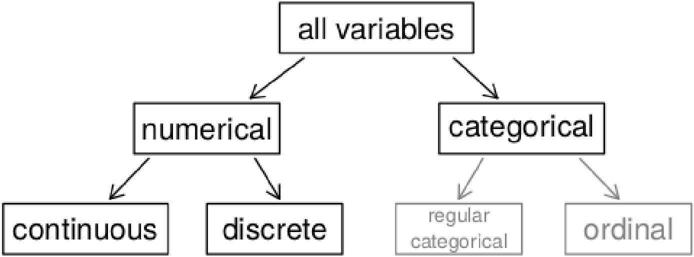
Study Types

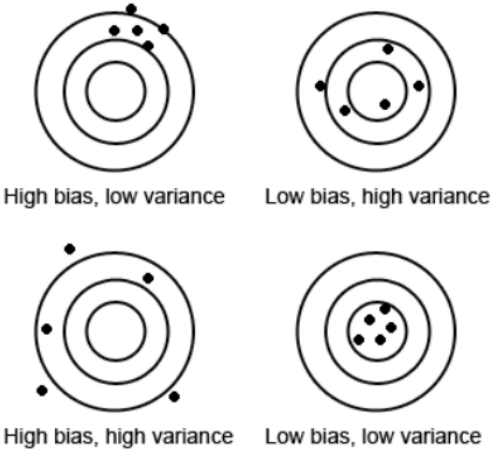
- Observational
- Experimental

What can go wrong?

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

Variable types



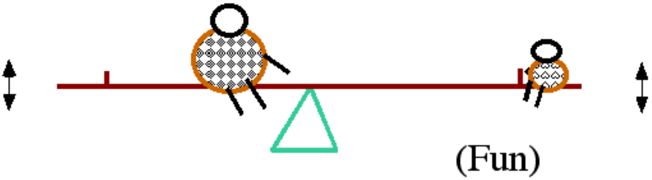

High bias, low variance
Low bias, high variance
High bias, high variance
Low bias, low variance

Mean

- Weighted and unweighted
- Centroid to physicists

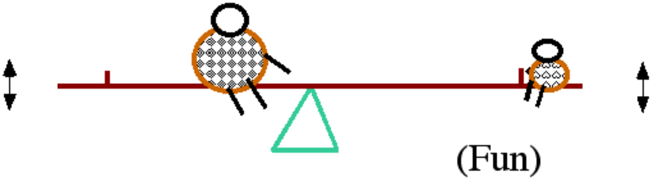
Mean

- Weighted and unweighted
- Centroid to physicists



Mean

- Weighted and unweighted
- Centroid to physicists

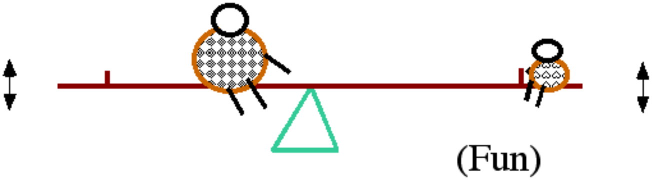


$$\mu = E(X) = \sum w_i x_i = \mathbf{w} \cdot \mathbf{x}$$

`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

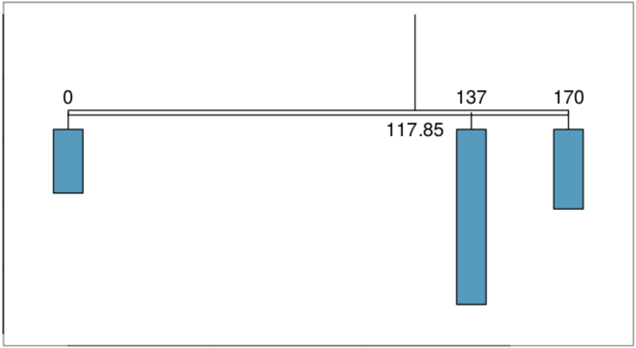
- Weighted and unweighted
- Centroid to physicists

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

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

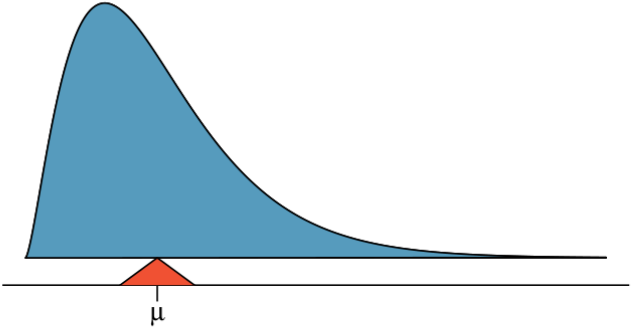
Mean

- Weighted and unweighted
- Centroid to physicists



Mean

- Weighted and unweighted
- Centroid to physicists



Population statistics

Deviation is distance from mean.

Population statistics

Variance is mean square of deviations

Population statistics

Standard deviation is square root of variance

Population statistics

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

Population statistics

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

Population statistics

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

Words

Given a distribution X, the

probability distribution function (pdf) (continuous) or **probability mass function (pmf)** is the probability that the variate has value x:

$$\Pr(a \leqslant X \leqslant b)$$

or

$$\Pr(X = a)$$

Words

Given a distribution X, the

cumulative probability function (cdf) is the probability that the variate is less than x:

$$\Pr(X \leqslant x) = \int_{-\infty}^{x} \mathrm{pdf}(x)\,\mathrm{d}x \text{ or } \sum_{i \leqslant x} \Pr(X = x)$$

Words

Given a distribution X, the

percent point function (ppf) is the inverse of the cdf. Given a probability, what's x? Also called the **inverse distribution**.

Words

Given a distribution X, the

survival function (sf) is the probability that the variate takes a value greater than x:

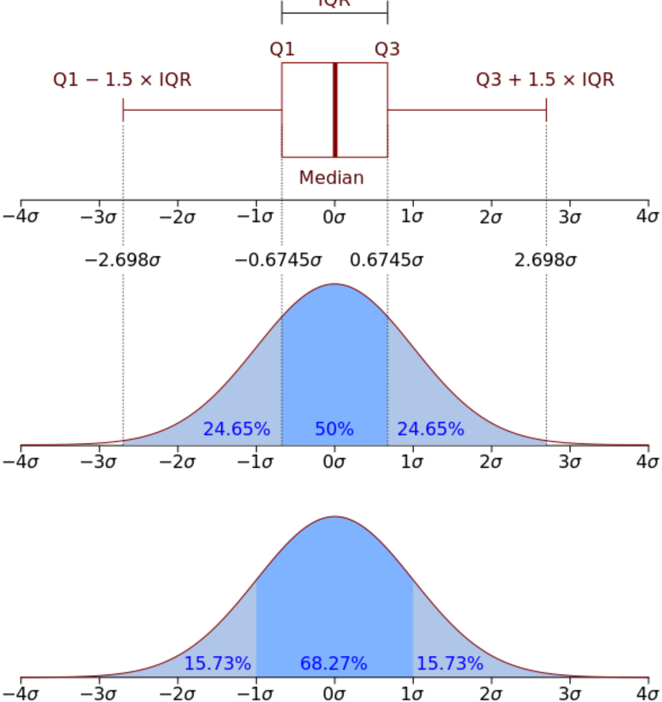
$$\text{ss}(x) = \Pr(X > x) = 1 - \text{cdf}(x)$$

Words

Given a distribution X, the

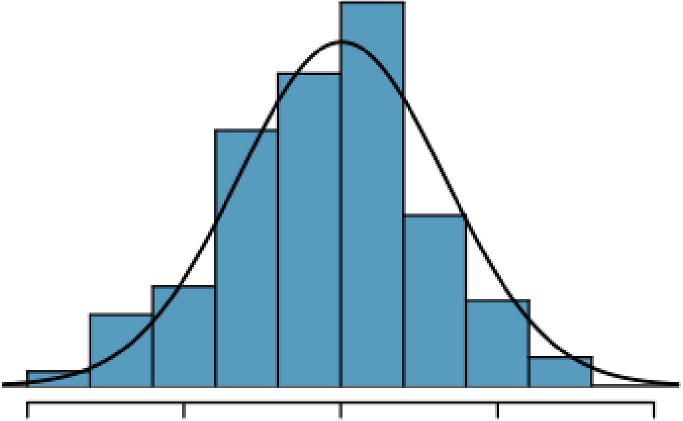
inverse survival function (isf) is the inverse of the survival function:

$$\text{isf}(\alpha) = \text{ppf}(1 - \alpha)$$


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
15.73%
68.27%
15.73%

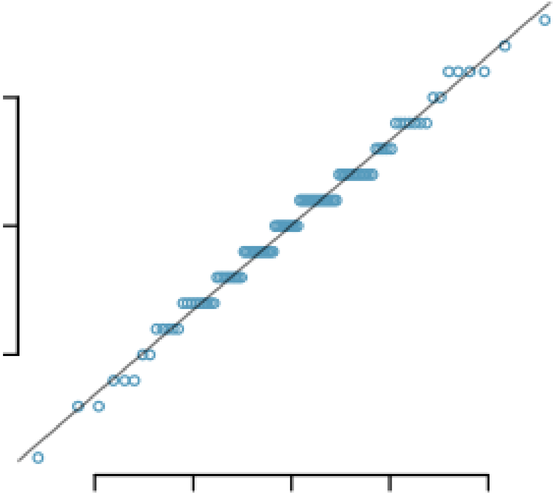
Evaluating Normal Approximations

Easy technique 1: visually compare to normal plot.



Evaluating Normal Approximations

Easy technique 2: normal probability plot.



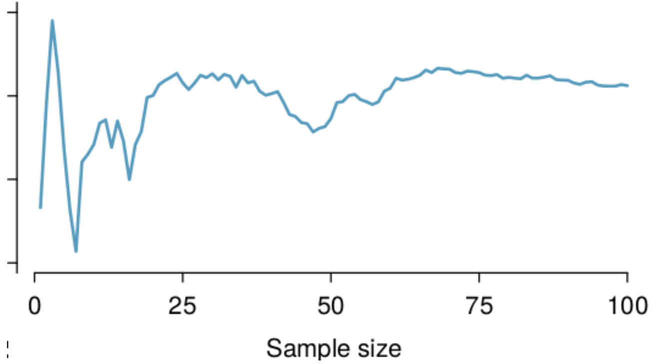
Also known as a quantile-quantile plot.

$$\overline{x} \neq \mu$$

Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Running mean. Sequence of partial sums (divided by number in sum).



Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Sampling variation. Change of $\overline{x}$ from one sample to the next.

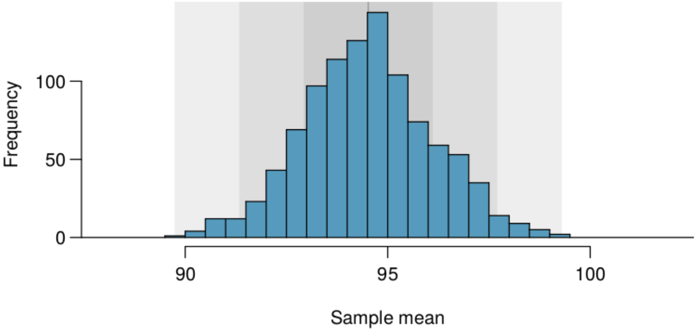
Inference Concepts

Running mean. Sequence of partial sums (divided by number in sum).

Sampling variation. Change of $\overline{x}$ from one sample to the next.

Sampling distribution. The distribution of possible point samples of a fixed size from a given population.

Sampling distribution



Confidence intervals

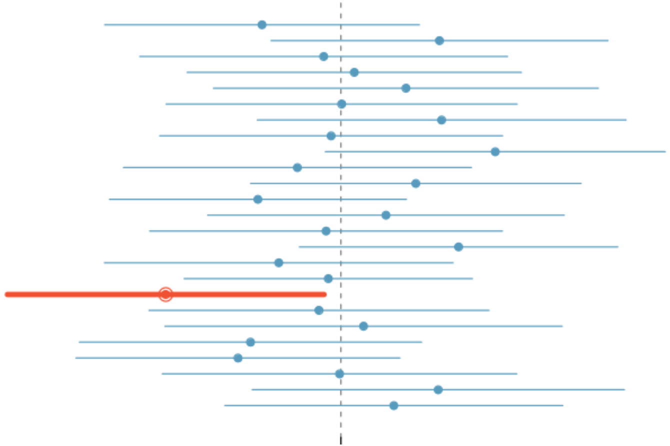
Sample *n* points, choose an interval around the sample mean.

A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.

Confidence intervals

Sample n points, choose an interval around the sample mean.

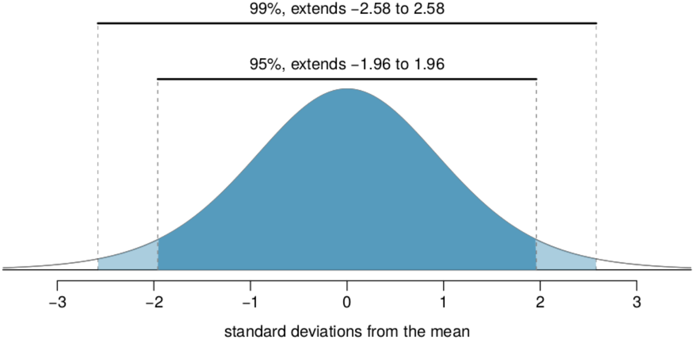
A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.



Confidence intervals

Sample *n* points, choose an interval around the sample mean.

A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.



Questions?

`purple.com/talk-feedback`