

ML Week

0x01a 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

$$\cup_j A_j = A \quad \wedge \quad A_i \cap A_j = \emptyset \implies$$

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

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?

*[http://simplystatistics.org/2015/03/17/
data-science-done-well-looks-easy-and-that-is-a-big-problem-f](http://simplystatistics.org/2015/03/17/data-science-done-well-looks-easy-and-that-is-a-big-problem-for/)*

What is data science?



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

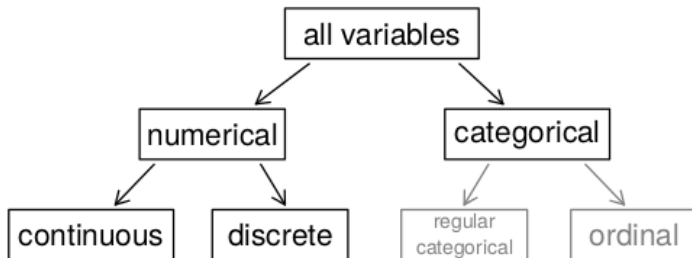
Study Types

- 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



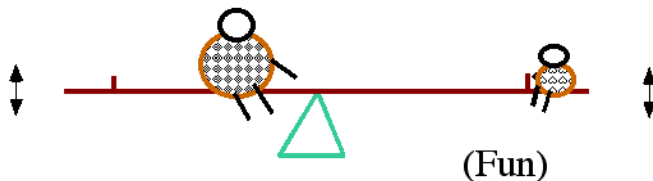
Low bias, low variance

Mean

- Weighted and unweighted
- Centroid to physicists

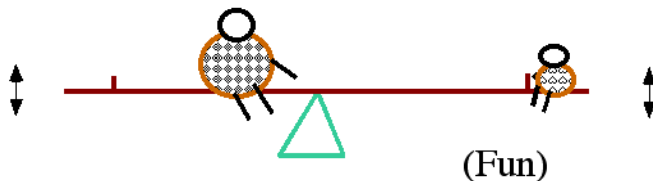
Mean

- Weighted and unweighted
- Centroid to physicists



Mean

- Weighted and unweighted
- Centroid to physicists

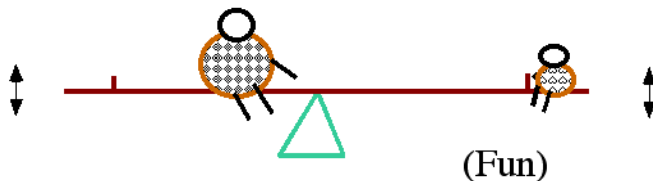


$$\mu = E(X) = \sum w_i x_i = \mathbf{w} \cdot \mathbf{x}$$

<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

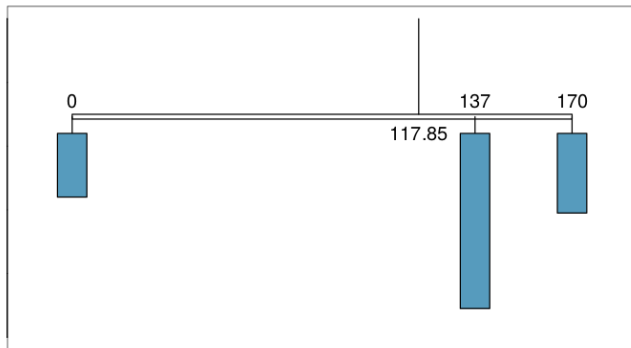
- Weighted and unweighted
- Centroid to physicists

$$\mu = E(X) = \int xf(x) dx$$

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

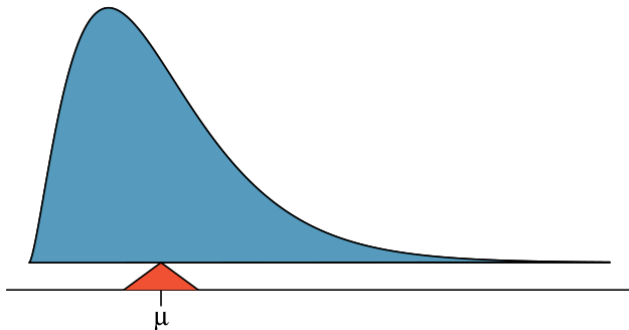
Mean

- Weighted and unweighted
- Centroid to physicists



Mean

- Weighted and unweighted
- Centroid to physicists



Deviation is distance from mean.

Variance is mean square of deviations

Standard deviation is square root of variance

Population statistics

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

Population statistics

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

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

Given a distribution X , the

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

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

or

$$\Pr(X = a)$$

Given a distribution X , the

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

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

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**.

Given a distribution X , the

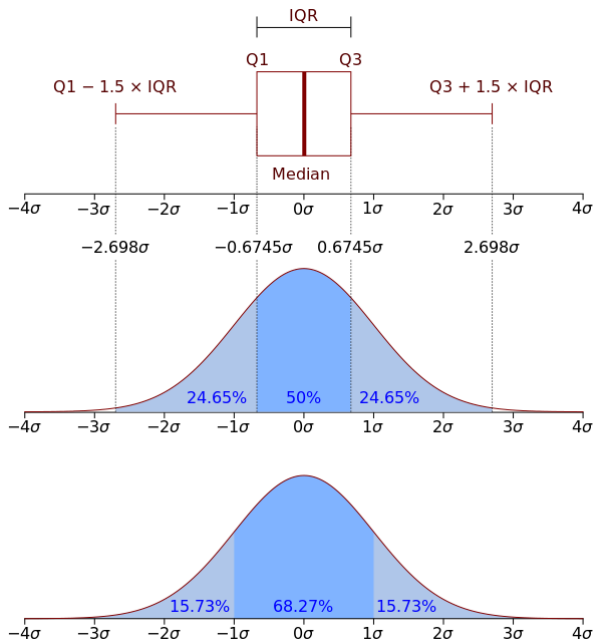
survival function (sf) is the probability that the variate takes a value greater than x :

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

Given a distribution X , the

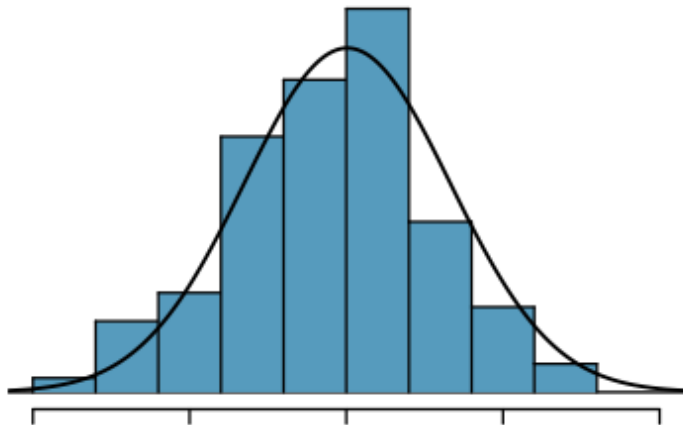
inverse survival function (isf) is the inverse of the survival function:

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



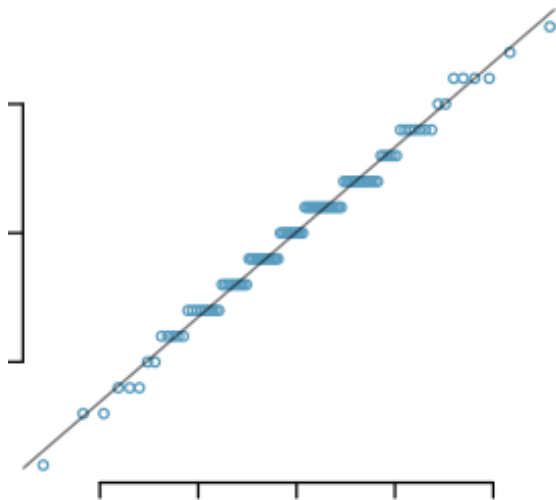
Evaluating Normal Approximations

Easy technique 1: visually compare to normal plot.



Evaluating Normal Approximations

Easy technique 2: normal probability plot.



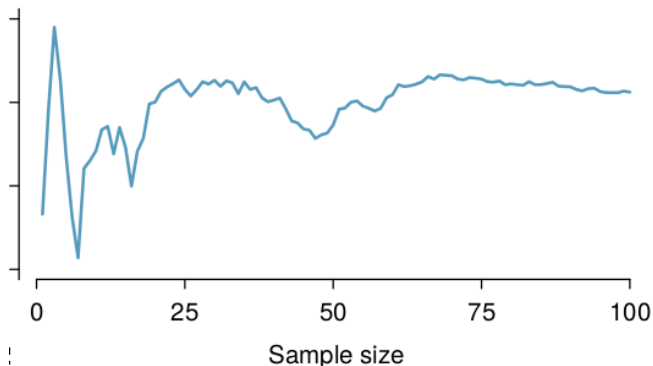
Also known as a quantile-quantile plot.

$$\bar{X} \neq \mu$$

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

Inference Concepts

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 \bar{x} from one sample to the next.

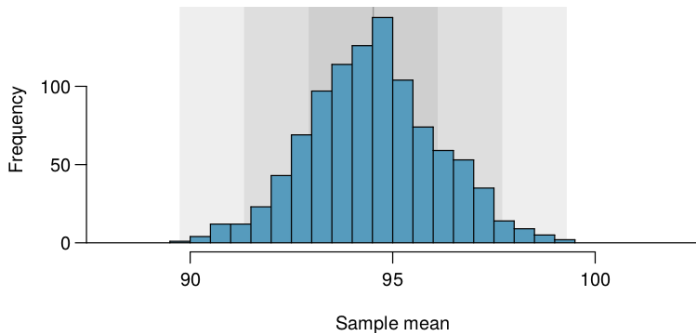
Inference Concepts

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

Sampling variation. Change of \bar{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

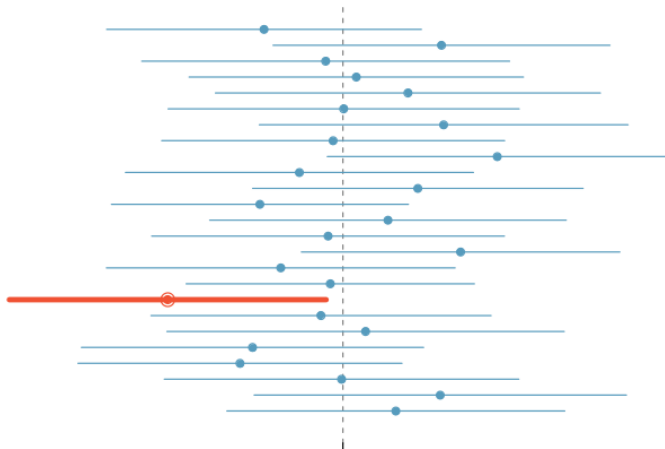
Sample n points, choose an interval around the sample mean.

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.

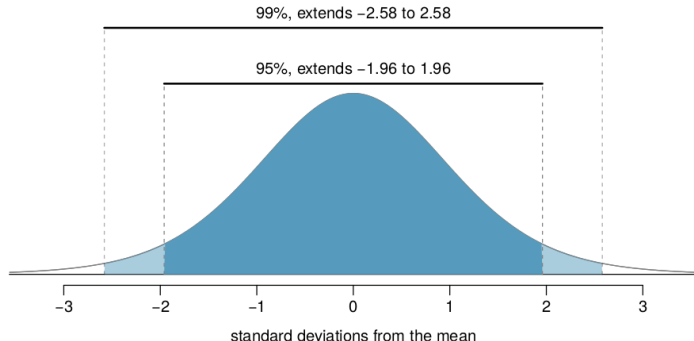
A 95% confidence interval means if we sample repeatedly, about 95% of the samples will contain the population mean.



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`