Introduction to Statistics
18.05 Spring 2022

T	T	T	H	H	T	H	H	H	T
H	T	H	T	H	T	H	T	H	T
H	T	T	T	H	T	T	T	T	H
H	T	T	H	H	T	H	H	T	H
T	T	H	H	H	H	T	H	T	H
T	T	T	H	T	H	H	H	H	T
T	T	T	H	H	H	T	T	T	H
H	H	H	H	H	H	H	T	T	T
H	T	H	H	T	T	T	H	H	T
H	T	H	H	H	T	T	T	H	H

Announcements/Agenda

Announcements

- Changed due date on Pset 6. Now due on Wed. 3/16. (Only one problem.)

Agenda

- Start statistics
- Define statistic
- Define likelihood
- Maximum likelihood estimates (MLE)

Statistics

Statistics is about the analysis of data. In practice, it is an art. It uses the mathematics or probability, but we will see that the conclusions we are able to draw are not rigorous in a mathematical sense.

- Data Collection:
  Informal Investigation / Observational Study / Formal Experiment
- Descriptive statistics : e.g. mean, median, variance of the data, histograms, box plots etc.
- Inferential statistics (the focus in 18.05): Draw conclusions or make decisions based on data.

To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

R.A. Fisher

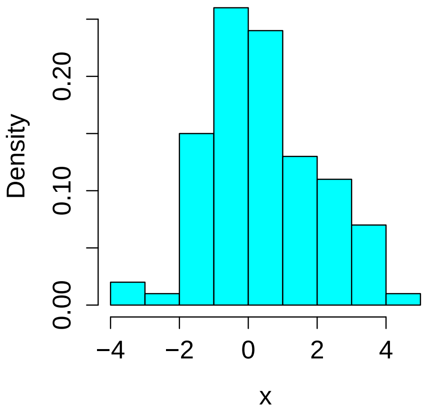
Question about data: Is it fair?

100 tosses of a coin. Is the coin fair?

T	T	T	H	H	T	H	H	H	T
H	T	H	T	H	T	H	T	H	T
H	T	T	T	H	T	T	T	T	H
H	T	T	H	H	T	H	H	T	H
T	T	H	H	H	H	T	H	T	H
T	T	T	H	T	H	H	H	H	T
T	T	T	H	H	H	T	T	T	H
H	H	H	H	H	H	H	T	T	T
H	T	H	H	T	T	T	H	H	T
H	T	H	H	H	T	T	T	H	H

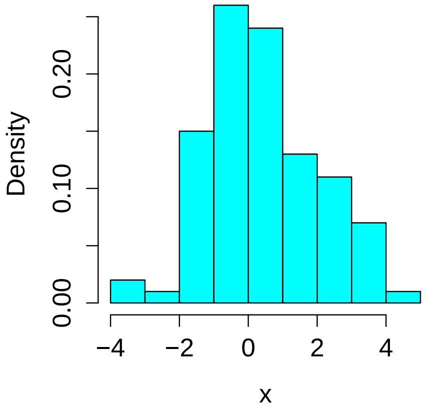
Question about data: Is it normal?

Does it have $\mu = 0$?   Is it normal?   Is it standard normal?



Question about data: Is it normal?

Does it have $\mu = 0$? Is it normal? Is it standard normal?



Sample mean $= 0.38$; sample standard deviation $= 1.59$

What is a statistic?

Definition. A statistic is anything that can be computed from the collected data. That is, a statistic must be observable.

- Point statistic: a single value computed from data, e.g sample average $\overline{x}_n$ or sample standard deviation s_n.

- Interval or range statistics: an interval $[a, b]$ computed from the data. (Just a pair of point statistics.) Often written as $\overline{x} \pm s$.

- **Important:** A statistic is itself a random variable since a new experiment will produce new data to compute it.

Concept question: Is it a statistic?

You believe that the lifetimes of a certain type of lightbulb follow an exponential distribution with parameter λ. To test this hypothesis you measure the lifetime of 5 bulbs and get data $x_1, \ldots x_5$.

Which of the following are statistics?

(a) The sample average $\overline{x} = \frac{x_1+x_2+x_3+x_4+x_5}{5}$.

(b) The expected value of a sample, namely $1/\lambda$.

(c) The difference between $\overline{x}$ and $1/\lambda$.

1. (a)
2. (b)
3. (c)
4. (a) and (b)
5. (a) and (c)
6. (b) and (c)
7. all three
8. none of them

Notation

Big letters X, Y, X_i are random variables.

Little letters x, y, x_i are data (values) generated by the random variables.

Example. Experiment: 10 flips of a coin:

X_i is the random variable for the i^{th} flip: either 0 or 1.

x_i is the actual result (data) from the i^{th} flip.
e.g. $x_1, \ldots, x_{10} = 1, 1, 1, 0, 0, 0, 0, 0, 1, 0$.

Reminder of Bayes' theorem

Bayes's theorem is the key to our view of statistics.

(Much more next week!)

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}.$$

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})}$$

Standard statistics question: Estimating a parameter

Example. Suppose we want to know the percentage p of people for whom cilantro tastes like soap.

Parameter of interest: p

Experiment: Ask n random people to taste cilantro.

Model: X_i is the response of the i^{th} person (1 if yes, 0 if no).
$X_i \sim \text{Bernoulli}(p)$.

Data: $x_1, \dots, x_n$ are the results of the experiment

Inference: Estimate p from the data.

Hypotheses: We can hypothesize that p has a specific value or a range of values.

Parameters of interest

Example. You ask 100 people to taste cilantro and 55 say it tastes like soap. Use this data to estimate p the fraction of all people for whom it tastes like soap.

So, p is the parameter of interest.

Don't know the value of p – can only hypothesize values.

Likelihood

You ask 100 people and 55 say cilantro tastes like soap. For a given p, the probability of 55 'successes' is the binomial probability

$$P(55 \text{ soap}|p) = \binom{100}{55} p^{55}(1-p)^{45}.$$

This the likelihood of p given the data. Often denoted $L(p)$.

In general,

$$\text{The likelihood of } p \ = L(p) = P(\text{data}|p)$$

NOTICE: The likelihood takes the data as fixed and computes the probability of the data for a given hypothetical value of$p$.

Maximum likelihood estimate (MLE)

The maximum likelihood estimate (MLE) is a way to estimate the value of a parameter of interest from the data.

The MLE is the value of p that maximizes the likelihood.

Different problems call for different methods of finding the maximum.

Here are a few –there are others:

1. Calculus: To find the MLE, solve $\frac{d}{dp}P(\text{data} \mid p) = 0$ for p. (We should also check that the critical point is a maximum.)

2. Sometimes the derivative is never 0 and the MLE is at an endpoint of the allowable range.

3. If the parameter takes only a finite number of values, you can compute the likelihood for each one and see which is the biggest.

Cilantro tasting MLE: Method 1

The MLE for the cilantro tasting experiment can be found by setting the derivative to 0:

$$\frac{dL(p)}{dp} = 0.$$

The calculation is not hard.

In this case, log likelihoods are easier, so we do that calculation instead.

Cilantro tasting MLE: Method 2: log likelihood

Because the log function turns multiplication into addition it is often convenient to use the log of the likelihood function

$$\text{log likelihood} = l(p) = \ln(\text{likelihood}) = \ln(P(\text{data} \mid p)).$$

Example.

$$\text{Likelihood } \; L(p) = P(\text{data}|p) = \binom{100}{55} p^{55}(1-p)^{45}$$

$$\text{Log likelihood } \; l(p) = \ln\left(\binom{100}{55}\right) + 55\ln(p) + 45\ln(1-p).$$

(Note first term is just a constant.)

The computation is posted with today's problems. The answer is

$$\hat{p} = 0.55$$

Adding a hat, is a standard way of indicating it is an estimate of the true value of p.

Board Question: Coins

(a) A box contains 3 coins. They land heads with, respectively, probability $p = 1/3,\ 1/2,\ 2/3$.

A coin is taken from the box. The mystery coin is tossed 80 times, resulting in 49 heads and 31 tails.

What is the likelihood of this data for each type of coin? Which coin gives the maximum likelihood?

(b) Now suppose you found a bent coin. It has an unknown probability p of landing heads. To estimate p you toss it 80 times getting 49 heads. Find the likelihood and log likelihood functions given this data. What is the maximum likelihood estimate for p?

Board question: continuous likelihood

For continuous likelihood: use the pdf instead of the pmf

Box of light bulbs.

Lifetime of each bulb $\sim \exp(\lambda)$, with unknown parameter λ.

For multiple independent data points, the likelihood is the product of the individual likelihoods.

(a) We test 5 light bulbs and find they have lifetimes of 2, 3, 1, 3, 4 years respectively. We assume the tests are independent.

(i) Find the likelihood and log likelihood functions (as functions of λ.)

(ii) What is the maximum likelihood estimate (MLE) for λ?

Reminder: An exponential distribution has pdf $f(x|\lambda) = \lambda \mathrm{e}^{-\lambda x}$

(b) Suppose we test 5 bulbs and find they have lifetimes x_1, x_2, x_3, x_4, x_5 years respectively. Redo Part (a) using these lifetimes.