Introduction to Statistics
18.05 Spring 2014

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

Three 'phases'

- Data Collection:
  Informal Investigation / Observational Study / Formal Experiment
- Descriptive statistics
- Inferential statistics (the focus in 18.05)

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.

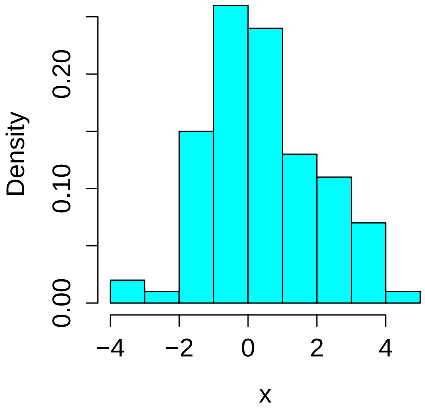
R.A. Fisher

Is it 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

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

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!)

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

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

Estimating a parameter

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

Experiment: Ask n random people to taste cilantro.

Model:
$X_i \sim \text{Bernoulli}(p)$ is whether the i^{th} person says it tastes like soap.

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

Inference: Estimate p from the data.

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.

Likelihood

For a given value of p the probability of getting 55 'successes' is the binomial probability

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

Definition:

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

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

Maximum likelihood estimate (MLE)

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

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

Different problems call for different methods of finding the maximum.

Here are two –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.

Log likelihood

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

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

Example.

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

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

(Note first term is just a constant.)

Board Question: Coins

A coin is taken from a box containing three coins, which give heads with probability $p = 1/3$, $1/2$, and $2/3$. The mystery coin is tossed 80 times, resulting in 49 heads and 31 tails.

(a) What is the likelihood of this data for each type on coin? Which coin gives the maximum likelihood?

(b) Now suppose that we have a single coin with unknown probability p of landing heads. Find the likelihood and log likelihood functions given the same data. What is the maximum likelihood estimate for p?

Continuous likelihood

Use the pdf instead of the pmf

Example. Light bulbs

Lifetime of each bulb $\sim \exp(\lambda)$.

Test 5 bulbs and find lifetimes of $x_1, \ldots, x_5$.

(i) Find the likelihood and log likelihood functions.

(ii) Then find the maximum likelihood estimate (MLE) for λ.

Board Question

Suppose the 5 bulbs are tested and have lifetimes of 2, 3, 1, 3, 4 years respectively. What is the maximum likelihood estimate (MLE) for λ?

Work from scratch. Do not simply use the formula just given.

Set the problem up carefully by defining random variables and densities.