STAT 139
Data and Samples

Populations

Random and representative samples

Sampling methods

Populations

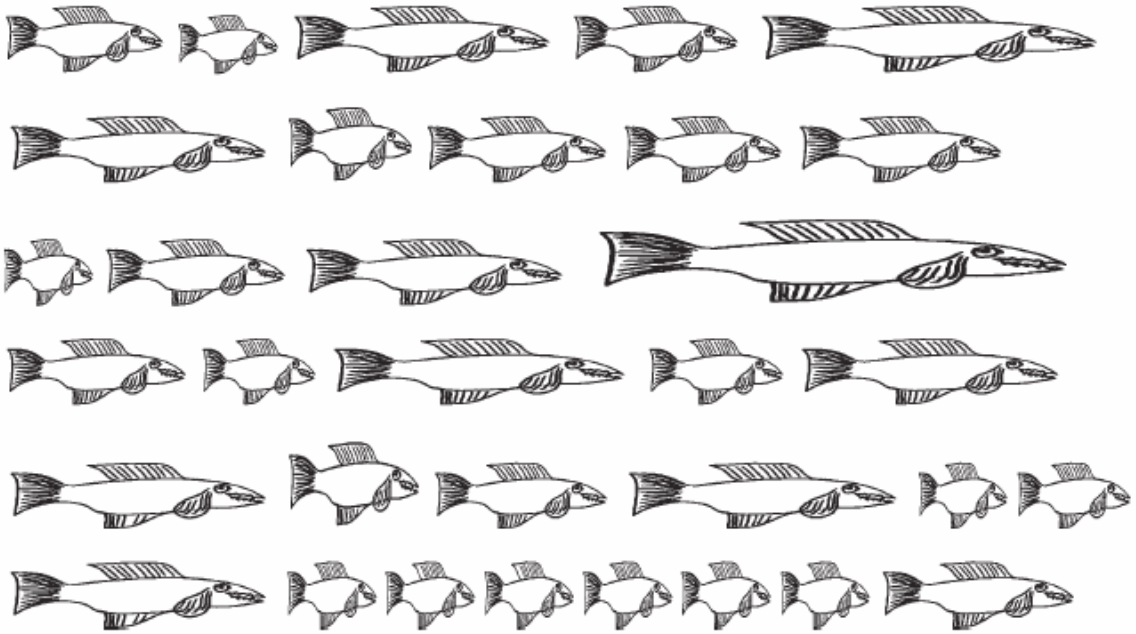
- are often theoretical, and unknowable
- are unmeasurable in total

Example: ALL sparrows of Pennsylvania

Some "populations" may be very small (subsets)

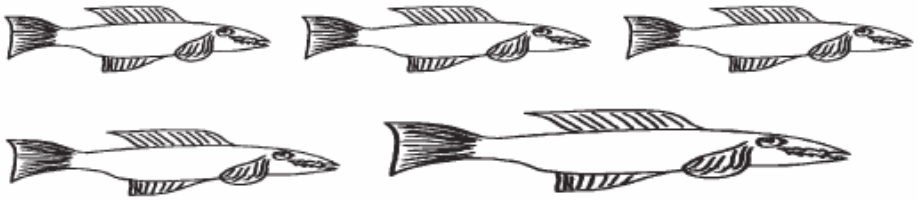
- Left-handed Pisces black republican male Texans who crochet
- Birds who did not survive a storm (Bumpus)

Population



Population

Sample 1



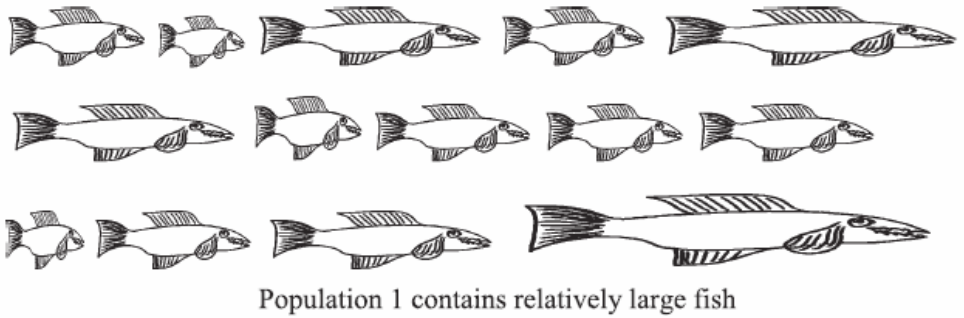
Sample 1

Sample 2

Sample 2

Different samples of the same population will be different (often a little). They may be VERY different if...

Population 1

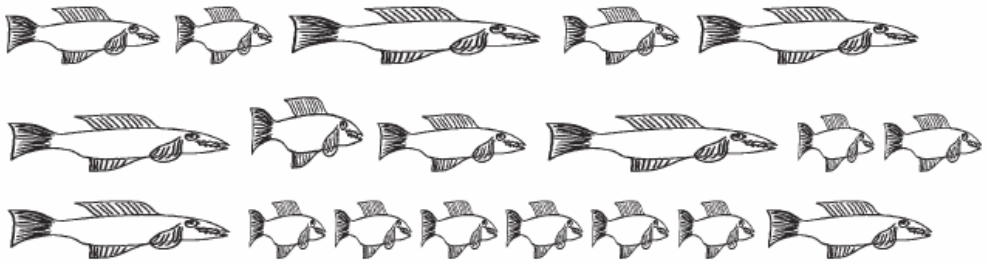


Population 1 contains relatively large fish

Random sample of Population 1

Random sample from population 1

Population 2



Population 2 contains smaller fish

Random sample of Population 2

Random sample from population 2

But samples of different populations can be very similar too!

Samples

If time, instruments, and money unlimited, the best sample size would be ...?

- as large as possible
- the whole population

BUT time, money, etc. are limited

- sampling is necessary
- sampling should be as random as possible
- truly random samples will be representative
- statistics makes most of samples

Samples and Treatments

To test a new drug

- some people given the drug (treatment)
- some not given the drug (control)

- people are different genetically
- different people may respond differently
- effects can be measured in different ways
- overall benefit is assessed

Data

Direct Measurements

- stature, weight, cholesterol level, blood pressure
- 100 yard dash time, grip strength
- amount of drug taken, snowfall amount

Counts

- number who die, survive shipwrecks
- number of heart attacks
- number of cigarettes per day
- count of certain DNA markers

Data

Dichotomous (Binary): Y/N, T/F

- treatment drug taken, drug abuser, smoker, pregnant
- DNA marker (Y/N) ?

Subjective Qualifiers

- mood/attitude, hair loss/gain
- any self-assessment

Sample data + Statistics = inferences

We generalize results from the sample to the population
(inferential statistics)
We can "predict" something about the population
- election polls

- characteristics (mean, standard deviation)
- treatment effects (drugs, toothpaste, diet, etc.)
- different responses in groups (M vs. F)
- relationships among different variables (correlation)

Causality

We can only get at causes by experimenting

(prospective study)

- we must control the treatment
- we should vary the treatment (dosage effects)

In humans, can be hard to do:

Can we randomly assign people to smokers?

Can we randomly assign people to eat broccoli?

Bad sampling (that gets headlines)

- Ask people on the street if they eat broccoli
- Weigh them / measure blood pressure
- Ask them if they are near their ideal weight
- Look for differences in eaters vs. non-eaters
- **Publish if broccoli effect**

"junk science"

BROCCOLI LINKED TO BETTER HEALTH

EIGHT GLASSES OF WATER DAILY LOWERS BLOOD PRESSURE

CHOCOLATE CAN HELP YOU LOSE WEIGHT

Bad sampling (that gets headlines)

Does (dark) chocolate improve your health?

A few scientists created the "Institute of Diet and Health"

Controlled experiment using small numbers (5 M 11 F)

- Diets: 1. Chocolate and low carb 2. low carb 3. control

- **Measure lots of characteristics (18)**
  - weight, BP, sleep, cholesterol, sodium
  - Look for differences 21 days later
- **Chocolate group lost 10% more weight**
- **Publish in peer-reviewed internet journal ($650)**

  "Chocolate with high Cocoa content as a weight-loss accelerator"
- **No sample size given**
- **Wrote press release**

Bad sampling (that gets headlines)

Dark chocolate improves your health!

- **Many newspapers, web sites, and a few television stations repeated the claims**
- Few asked about sample or methods

Excellent News: Chocolate Can Help You Lose Weight!

ANI

Posted: 31/03/2015 16:21 IST | Updated: 31/03/2015 16:21 IST

The study was weak

Then the scientists revealed all the facts about their study.

AFTERWARDS, a retraction notice from the Editorial Office of the *International Archives of Medicine*

Abstract

This article accidentally appeared online for some days in the same url where this notice can now be found. A public disclaimer has been published in the publisher's website explaining how it all happened. **Indeed that manuscript was finally rejected and never published as such.**

We are sorry for the inconvenience. We are taking measures to prevent this kind of mistakes from happening again.

An (almost) inescapable truth

Sample size is the single most important thing!

Larger samples are almost always more representative

Effects are easier to detect in larger samples
(statistical power)

All population estimates are better with larger samples
(mean, standard deviation, treatment effects, etc.)

Math and Probabilities

We use probability to judge statistical significance

If a result is unlikely based on random probability

- it is a statistically significant result

The probability of heads coming up in a coin toss?

- important because I have psychic powers!

Math and Probabilities

The probability (*p*) of 2 heads coming up in 2 coin tosses?

$= 0.5 \times 0.5 = 0.5^2 = ½ \times ½$

$0.5^2 =$?

0.25 or ¼

The probability of tossing two heads in a row = 0.25

$0.3^2 =$?

0.09

(Binomial probabilities, coin tosses)

When a number is between 0 and 1, its square is smaller

p of 3 heads in a row: $0.5 \times 0.5 \times 0.5 = 0.5^3 = 1/8$ or 0.125

Paul the Octopus and probability (*p*)

(2010)

Telegraph.co.uk

Home | News | Sport | Finance | Lifestyle | Comment | Travel | Culture | Technology

UK | World | Politics | Celebrities | Obituaries | Weird | Earth | Science | Health News | Educatio

USA | Barack Obama | Asia | China | Central Asia | Europe | Australasia | Middle East | Afric

HOME > NEWS > WORLD NEWS > EUROPE > GERMANY

Mahmoud Ahmadinejad attacks Octopus Paul

Mahmoud Ahmadinejad, the Iranian leader, says Paul the Octopus, the sea creature that correctly predicted the outcome of World Cup games, is a symbol of all that is wrong with the western world.

Published: 1:30PM BST 27 Jul 2010

Email | Print

The Iranian president accused Octopus Paul of spreading 'western propaganda and superstition' Photo: REUTERS/EPA

He claims that the octopus is a symbol of decadence and decay among "his enemies".

Germany
News
World News
Europe
Iran
World Cup 2010

How good was Paul?

2008 and 2010:

11 of 14 correct

$p < 0.022$

In 2010, all 7 correct

$p = 0.5^7 < 0.008$

Results involving Germany

Euro 2008

Opponent	Stage	Date	Prediction	Result	Outcome
Poland	group stage	8 June	Germany	2–0	Correct
Croatia	group stage	12 June	Germany	1–2	Wrong
Austria	group stage	16 June	Germany	1–0	Correct
Portugal	quarter-finals	19 June	Germany	3–2	Correct
Turkey	semi-finals	25 June	Germany	3–2	Correct
Spain	final	29 June	Germany[3]	0–1	Wrong

World Cup 2010

Opponent	Stage	Date	Prediction	Result	Outcome
Australia	group stage	13 June	Germany[37]	4-0	Correct
Serbia	group stage	18 June	Serbia[37]	0–1	Correct
Ghana	group stage	23 June	Germany[37]	1–0	Correct
England	round of 16	27 June	Germany[15]	4–1	Correct
Argentina	quarter-finals	3 July	Germany[30]	4–0	Correct
Spain	semi-finals	7 July	Spain[38]	0–1	Correct
Uruguay	3rd place play-off	10 July	Germany[39]	3–2	Correct

Sometimes, hard to understand what statistical significance means

Next time

"Read" (work through) the Introductions to R and R Studio: VCH1-2011; VCH1-2014;

Additional files are found under R Resources

Setting up *R* and *R Studio*

The PCs here in the Lab get wiped with every boot.

If you want to install on your own computer:

Windows and Mac:

1. Download and install *R* (version 3.3.2)
***R* home page: http://cran.r-project.org/**

2. Download and install R Studio (version 1.0.136)
RStudio home page: https://www.rstudio.com/products/rstudio/download/

Start *R* Studio
(Vista Users: Right-click the icon and choose Compatibility, run as administrator)