Sections 2.8, 2.9, and 3.2

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

2.8 Statistical inference

- Data is a sample from a larger population.
- Point of collecting and describing data is to *infer* about the population.
- Random sampling of data ensures that a representative collection of measurements has been taken, and that the data provide a reasonable “snapshot” of the population.
- Data can be used to formally assess population characteristics.

Example 2.8.1 Blood types in England

- $n = 3696$ blood types collected in England (published in 1939).
- 1634 were type A.
- In the *sample* $\frac{1634}{3696} = 0.44 = 44\%$ are type A.
- This is a good estimate of the percentage in the population *if the sample is representative*.
- If the sample is "bad" 44% still estimates the percentage in the population, but it may be biased.
- Estimating the population percentage as 44% is *inferring* a population characteristic from an imperfect sample.
- Questions: What is the population here? Is the sample representative?

Example 2.8.3 Alcohol and MOPEG

- $n = 7$ healthy men had MOPEG (3-methoxy-4-hydroxyphenylethyleneglycol, the major noradrenaline metabolite in the central nervous system) measured (pmol/ml) before and after drinking 80 gm of alcohol (about 6 drinks) at 8am.
- Population: All people? Men? Healthy men? Healthy men who have 6 drinks? Healthy men who have 6 drinks at 8am? Healthy men who have 6 drinks at 8am in a laboratory while being watched by scientists in white lab coats?
- The population is often narrower than we would like, but we are able to infer *something* anyway.
- If the results are conclusive, we can embark on a more ambitious study involving a more heterogeneous sample.

MOPEG

Table 2.8.2 Effect of alcohol on MOPEG

	MOPEG concentration		
Volunteer	Before	After	Change
1	46	56	10
2	47	52	5
3	41	47	6
4	45	48	3
5	37	37	0
6	48	51	3
7	58	62	4

Data collected from a population. We would like to infer whether MOPEG generally increases after consuming alcohol. Does it? Can we say this with certainty? For which population?

Proportions

- Variables with only two possible outcomes are said to be **dichotomous**.
- A **population proportion** is the fraction of *all population units* that exhibit the trait of interest, denoted p.
- We can take a random sample of n observational units and find the **sample proportion** of the n units with the trait of interest, denoted $\hat{p}$.
- Example 2.8.1. $\hat{p} = 0.44$ estimates the population proportion of blood type A in England.

Example 2.8.5 Lung cancer treatment

- Example 2.8.5: $n = 11$ patients with adenocarcinoma (type of lung cancer) treated with Mitomycin. $y = 3$ of the patients had a positive response (tumor shrunk more than 50%).
- $\hat{p} = \frac{3}{11} = 0.27$ estimates p, which is unknown.
- What is the population here?
- How good is this estimate?

Parameters versus statistics

- Sample characteristics estimate population characteristics.
- The sample mean $\bar{y}$ estimates the **population mean** μ, the average over *everyone in the population*.
- The sample standard deviation s estimates the population standard deviation σ.
- $\hat{p}$ estimates p.
- Sample medians estimate population medians, etc.
- The sample estimates are called **statistics**, their population counterparts are called **parameters.**

Example 2.8.6 Leaves on tobacco plants

An agronomist counted the number of leaves on $n = 150$ Havana tobacco plants

Table 2.8.4 Number of leaves on tobacco plants

Number of leaves	Frequency (number of plants)
17	3
18	22
19	44
20	42
21	22
22	10
23	6
24	1
Total	150

Example 2.8.6 Leaves on tobacco plants

- The sample mean is $\bar{y} = 19.8$ leaves.
- This estimates μ, where μ is average number of leaves grown on *all* Havana tobacco plants grown under the same conditions.
- The sample standard deviation is $s = 1.4$ leaves.
- This estimates σ, where σ is the standard deviation of leaves grown on *all* Havana tobacco plants grown under the same conditions.

Section 2.9 What's coming up...

- The mean is a number; the *density* is a function.
- However, *both* can be estimated from a sample of size *n*.
- Confidence interval gives plausible range of values for μ.
- Hypothesis tests allow us to assess evidence that μ is some fixed number, like $\mu = 15$ leaves.
- Chapters 3 (probability & random variables), 4 (normal distribution), and 5 (distribution of sample statistics) lay groundwork for these statistical tools.
- The next slide catalogues three population parameters and their sample estimates...

Statistics & population parameters they estimate

Table 2.9.1 Notation for some important statistics and parameters

Measure	Sample value (statistic)	Population value (parameter)
Proportion	$\hat{p}$	p
Mean	$\bar{y}$	μ
Standard deviation	s	σ

Chapter 2, review of important terms & ideas

- 2.1 numeric (continuous & discrete) vs. categorical (ordinal & nominal) variables; observational unit.
- 2.2 frequency distributions for categorical and continuous data: tables, bar charts, and histograms; shape: skewed vs. symmetric, modality.
- 2.3 Measures of center: sample mean $\bar{y}$ and sample median $\tilde{y}$; when to use which; what happens with skewed data.
- 2.4 Five number summary and boxplots; IQR; outliers.
- 2.5 Looking for association: categorical-categorical, categorical-numeric, numeric-numeric.
- 2.6 Measures of spread/dispersion: range, IQR, and s; empirical rule for $\bar{y}$ & s.
- 2.8 Inference: parameter vs. statistic.

Probability

- The **probability** of an event E occurring is the long-run proportion of times it will occur in repeated experiments.
- Denoted $\Pr\{E\}$.
- $0 \leq \Pr\{E\} \leq 1$.
- $\Pr\{E\} = 1$ means E always occurs; e.g. E = lab rat has two eyes.
- $\Pr\{E\} = 0$ means E never occurs; e.g. E = lab rat speaks fluent Finnish.
- Example 3.2.1. E = "tails" on fair coin toss, $\Pr\{E\} = 0.5$. Half the tosses will be tails.

Probability and sampling a population

- Consider a population with proportion p of a characteristic.
- *Randomly* choose one member from the population.
- Let E = randonly chosen member has characteristic.
- Then $\Pr\{E\} = p$.
- Example 3.2.3. Large population of *Drosophila melanogaster* (fruit fly) kept in lab. Proportion that are black is $p = 0.3$; proportion gray is $1 - p = 0.7$.
- Say E = randomly sampled fruit fly is black.
- $\Pr\{E\} = 0.3$.

Rolling a fair 6-sided die

- Say I roll one fair 6-sided die.
- E = "roll a 7.5." $\Pr\{E\}$?
- E = "roll a number between zero and ten." $\Pr\{E\}$?
- E = "roll a 6," i.e. $E = \{6\}$. $\Pr\{E\}$?
- E = "roll a 1 or a 6," i.e. $E = \{1, 6\}$. $\Pr\{E\}$?
- E = "roll an even number," i.e. $E = \{2, 4, 6\}$. $\Pr\{E\}$?
- If I roll the die 100,000 times and find the sample proportion of times I rolled an even number, what would this sample proportion be close to?

Frequency interpretation in more detail

- What do I (and the book) mean by $\Pr\{E\}$ is the “long-run proportion of times E occurs in repeated experiments” ?
- We will show shortly that the probability of sampling two flies of the same color is $0.7 \times 0.7 + 0.3 \times 0.3 = 0.58$.
- Let’s look at what happens why we try repeating the experiment “randomly sample two flies” over and over and over and over again...
- After each sample we will update the sample proportion of times two flies of the same color were are sampled.

Cumulatively estimating $\hat{p}$

Table 3.2.1 Partial results of simulated sampling from a *Drosophila* population

Sample number	Color: 1st Fly	Color: 2nd Fly	Did E occur?	Relative frequency of E (cumulative)
1	G	B	No	0.000
2	B	B	Yes	0.500
3	B	G	No	0.333
4	G	B	No	0.250
5	G	G	Yes	0.400
6	G	B	No	0.333
7	B	B	Yes	0.429
8	G	G	Yes	0.500
9	G	B	No	0.444
10	B	B	Yes	0.500
⋮	⋮	⋮	⋮	⋮
20	G	B	No	0.450
⋮	⋮	⋮	⋮	⋮
100	G	B	No	0.540
⋮	⋮	⋮	⋮	⋮
1,000	G	G	Yes	0.596
⋮	⋮	⋮	⋮	⋮
10,000	B	B	Yes	0.577

Plotting $\hat{p}$ versus number of experiments

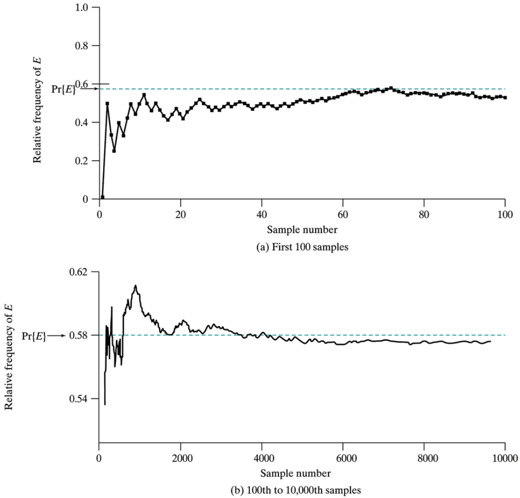

Figure 3.2.1 Results of sampling from fruitfly population. Note that the axes are scaled differently in (a) and (b).

What is happening?

- As more and more information (data!) are collected, we can estimate the probability $p = 0.58$ almost perfectly.
- In some textbooks, probability is defined as a limit

$$\Pr\{E\} = \lim_{n\to\infty} \frac{\text{\# times } E \text{ occurs out of } n \text{ experiments}}{n} = \lim_{n\to\infty} \hat{p}.$$

- This is "long-run proportion."
- The previous two slides allow n to get really large, but not infinite.
- We see that as n gets large, $\hat{p} \to \Pr\{E\}$ ("gets arbitrarily close to").
- We can replace '$\to$' by '$=$' only at $n = \infty$.
- Next lecture: probability trees and probability rules.