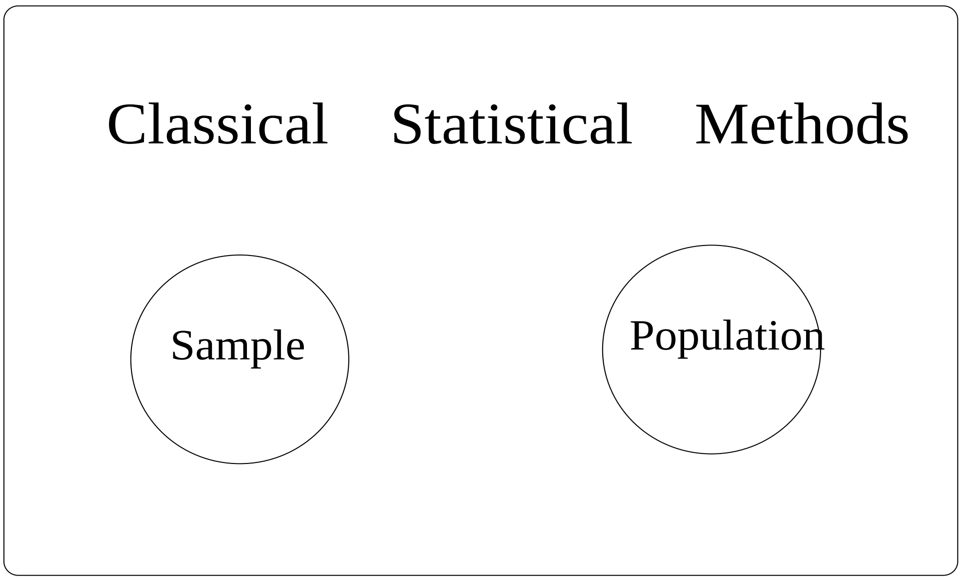
Good Apples?
Sampling Biases

Jiayang Sun
Department of Statistics
Case Western Reserve University
Cleveland, OH 44106

jiayang@sun.cwru.edu
http://sun.cwru.edu/~jiayang

Ideal World



- Normal
- Simple Random Sample
- Sampled Population = Target Population
- ...

Real World

Outliers, Censored or Truncated Observations
Missing Values, Biased Sample

$\{X$ is missing?$\}$ depends on $X \Longrightarrow$ Biased Sample !

Biases

- Respondent Bias
  Hawthorne effect, Placebo effect, Lies and Anchoring Effect ...

- Experimenter Bias

- Interpretation Bias

- Sampling Bias
  Self selection, none response; Length or size bias ...

Respondent Bias

Example 1: A group of workers were observed by the plant manager to determine how long it takes to perform a certain task. The time was much less than the manager expected. What might explain the difference?

Hawthorne Effect

By merely informing individuals that they are included in a study, they tend to show the effect (response) they think the researcher is looking for

Example 2: Researchers are interested in studying the effect of a new drug on cancer victims. Due to ethical considerations, they could only use the drug on the most desperately ill patients. The condition of the patients was found to improve – surprised?

Placebo Effect

Tendency to improve even though the treatment may be totally inert

A way to help control this effect is:

- set a control group
- set a treatment group

Example 3: Do you cheat in your exams? Surprising, the answers are all no – surprised?

Lies

Respondent may have a tendency to lie or refuse to respond, especially when asked personal or sensitive questions

Warner's Randomized Response Technique

Goal: Estimate proportion of cheaters

1. Set two questions:

Q1: Cheat?

Q2: Registered democrat?

2. Let the respondent toss a coin in deciding which Q to answer and tell interviewer only "Yes" or "No".

$$P(\text{yes}) = P(\text{yes}|Q1)P(\text{Q1}) + P(\text{yes}|Q2)P(\text{Q2})$$

$\Longrightarrow$

$$\begin{aligned} P(\text{yes}|\text{Q1}) &= \frac{P(\text{yes}) - P(\text{yes}|\text{Q2})P(\text{Q2})}{P(\text{Q1})} \\ &= \frac{0.52 - 0.32 \cdot 0.5}{0.5} = 0.72 \end{aligned}$$

Example 4: A Gallup poll sponsored by the disposable-diaper industry found that 84% of adults in the US felt that it would not be fair to ban disposable diapers.

"It is estimated that disposable diapers account for less than 2% of the trash in today's landfill. In contrast, beverage containers, third-class mail, and yard waste are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?"

Anchoring Effect

Happens when the Q suggests answer

Example 5: You are driving along a highway when you hear a voice on your right say, " Dear, you are almost out of gas." You looked and said: "No, I am not." "Yes, you are." What happened?

Parallax Bias

The angle at which an instrument is read can bias the observations

Experimenter Bias

- Role of H_A/ Presence in study

H_A is often the model specified by the scientist or experimenter. She/he may have a concious or unconsious tendency to bias the result in favor of H_A

Such biases may arise in
– assigning subjects to control or treatment G
– obtaining measurements
– interpreting the results

They may be avoided (reduced) by
the use of randomizatin in assignment, in a double blind version

Joe

- Publication Pressure/ Data Ransacking

Editors of journals are (genreally) unlikely to publish articles that "accept H_0"

$\Longrightarrow$ data ransacking

i.e. perform many tests for comparisons, but only publish those that turned out to be significant

☻ Idea:

- ask how many experiments that he/she had done
- have an independent scientist to repeat the experiment

Interpretation Bias

Example 6: An investigator studying earnings of men and women at a company found that (1) average salary for men is higher than for women; and (2) In each job category, women earn more on average than men. How could this happen?

	Job 1		Job 2		Overall
	Hr Salary	#	Hr Salary	#	
Men	\$8	20	\$18	80	
Women	\$10	80	\$20	20	

Simpson Paradox - Aggregation Bias

Due to pooled results
What's true for the whole population may not be true for some (or all) of the subpopulations

Example 7: A scientist has published 1000 papers. With $\alpha = 0.05$ and in all he rejected H_0. He then claimed that he expected 950 papers to be correct. Is the claim justified?

No!

$$\alpha = P\{\text{type I error}\} = P\{\text{Rej } H_0 | H_0\} = 0.05$$

He is right if H_A is indeed true:

$$P\{\text{Rej } H_0 | H_1\} = 1 - P\{\text{Acp } H_0 | H_1\} = 0.95$$

$$\boxed{\text{Wrong as } P(A|B) \neq P(B|A)}$$

Small Prior Probability/ Bayes Theorem

Example 8: 1 % of employees of a company use drug X, all employees are tested for the presence of this drug. The probability of finding a positive test result among users is .95, the probability of a negative test result among non-users is .90. If an employee has a positive test, what is the probability they are a drug users?

Does knowing that the test is positive increase the chance of being a drug user?

Sampling Bias

- Self-selection, non-response

sampled population $\neq$ target population

- Length or Size Bias

If the larger or longer items have bigger chance of being selected, the mean will bias upwards.

Examples:

1. Tobacco Law Suit

2. Scleroderma

• Data: time from diagnosis of scleroderma (a rare disease) to death.

• Target Population: all cases of scleroderma diagnosed in Michigan from 1980 to 1991.

• When we estimate survival curves for patients diagnosed in '80-'85 versus '86-'91, we find that those diagnosed between '80 and '85 live significantly longer.

Our sources of information included hospital databases and responses from private physicians. Unfortunately (for us), because hospitals records don't always go back to 1980, and physicians don't always remember patients they saw many years ago, we are more likely to identify patients who are still alive (and thus have more current hospital records or doctor visits). We feel that the result is entirely due to our length-biased sample. (If anything, medical care has improved during this period, so an unbiased sample should give the reverse result.)

Solutions for Sampling Bias

Simple Solution:

- with length biased sampling and $Y > 0$:

$$\mu_x = \mu_y \left(1 + \frac{\sigma_y^2}{\mu_y^2}\right)$$

General Solution:

- Ideally,

$$Y_1, \ldots, Y_N \quad \sim \quad f_\theta$$

- In reality, Y is observed with a probability $w(y)$ if $Y = y$. The observed obs are

$$X_1, \ldots, X_n \quad \not\sim \quad f_\theta$$

Instead,

- $n \sim$ Binomial(N, κ), $\kappa = \int w f_\theta$;

- and given n:

$$X_1, \ldots, X_n \sim f^*_{w,\theta}(x) = \frac{w(x) f_\theta(x)}{\kappa(w, \theta)}.$$

Likelihood

- N is known, the likelihood is

$$\left[\prod_{i=1}^{n} f^*_{w,\theta}(x_i)\right](1-w)^{N-n}$$

- N is unknown, the conditional likelihood is

$$\left[\prod_{i=1}^{n} f^*_{w,\theta}(x_i)\right] = \left[\prod_{i=1}^{n} \frac{w(x_i)f_\theta(x_i)}{\kappa(w,\theta)}\right]$$

Results

- Parametric: easy

Example: $Y \sim N(\mu, \sigma^2), w(y) = ce^y$.
Then

$$\kappa = \exp\{\frac{\sigma^2 + 2\mu}{2}\}$$

and the Likelihood equation is

$$\begin{aligned}
\sigma^4 + \sigma^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \\
\sigma^2 + \mu &= \bar{X}
\end{aligned}$$

which gives the MLE:

$$\begin{aligned}
\hat{\mu} &= \bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2
\end{aligned}$$

- Semiparametric: $\boxed{\text{MM algorithm}}$

- Nonparametric: need covariates ?

• Connection to censoring and truncated observations

• Testing problems:

$$H_0: \text{ no bias} \quad vs \quad H_A: \text{ some bias}$$

Conclusion

- Know Context

Who? Individuals measured and observed

What? has been measured and observed

Why? Study Purpose

- Do something about the bias!