Pitfalls and Biases

in design, analysis and ...

Jiayang Sun

Department of Statistics
Case Western Reserve University
Cleveland, OH 44106

jsun@case.edu

FESC Research Seminar, May 10, 2006
Abstract

Via some fun, interactive examples, I’ll illustrate some practical issues with designing experiments or protocols, analyzing REAL data (vs. ideal data) and interpreting "statistical outcomes". The lessons to be learned are that

1. It is not just the T-test or P-values;
2. Be aware of biases created by incomplete data or an arbitrary experiment;
3. ...
**Ideal World**

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

Classical Statistical Methods

Sample

Population
Real World
Correlations, Outliers, Missing Values, Censored or Truncated Obs.

If whether $X$ is missing depends on $X$, then there is sampling bias and hence

\[
\text{target population} \neq \text{sampled population}
\]
Biases

1. Respondent Bias
2. Experimenter Bias
3. Procedure Bias
4. Sampling Bias
5. Interpretation Bias
1. Respondent Bias - *Lies; Hawthorne, Placebo, and Anchoring Effects*
2. Experimenter Bias
3. Procedure Bias - *mistakes*
4. Sampling Bias - *Self selection, None ignorable none-responses; and Length or size bias ...*
5. Interpretation Bias

Need better designs
Need better analyses
Need better understanding of statistics
Respondent Bias: Example 1.1

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?
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 1.2

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?
Example 1.2

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

Solution: set a control group & a treatment group
Example 1.3

Do you cheat in your exams? Surprising, the answers are all no – surprised?
Example 1.3

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”.

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

\[
P(\text{yes}|Q1) = \frac{P(\text{yes}) - P(\text{yes}|Q2)P(Q2)}{P(Q1)}
\]

\[
= \frac{0.52 - 0.32 \cdot 0.5}{0.5} = 0.72
\]
Example 1.4

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?”
Example 1.4

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 1.5

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?
Example 1.5

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: Example 2.1

Role of $H_1$ in study

$H_1$ is often the model specified by the scientist or experimenter. She/he may have a conscious or unconscious tendency to bias the result in favor of $H_1$. 
Experimenter Bias: Example 2.1

Role of $H_1$ in study

$H_1$ is often the model specified by the scientist or experimenter. She/he may have a conscious or unconscious tendency to bias the result in favor of $H_1$.

Such biases may arise in

- assigning subjects
- obtaining measurements
- interpreting the results

A possible solution/remedy: use *randomization* in assignment (in a *double blind* version)
Example 2.2

Publication Pressure

Editors of journals are (generally) unlikely to publish articles that “accept $H_0$”. So, perform many tests for comparisons, but only publish those that turned out to be significant.
Example 2.2

Publication Pressure

Editors of journals are (generally) unlikely to publish articles that “accept $H_0$”. So, perform many tests for comparisons, but only publish those that turned out to be significant.

Data Ransacking

How to catch the problem? Idea:

- ask how many experiments that he/she had done
- repeat the experiment independently
Basic Principles of Good Experimental Design

- Simplicity
- Avoidance of biases
  - use of random allocation for forming treatment groups
  - responsiveness to the scientific question(s)
  - replication
- Adequate sample size ...
Procedure Bias: Example 3.1

Data obtained over time were treated as a simple random sample and a simple t-test is performed. What’s wrong with this procedure?
Data obtained over time were treated as a simple random sample and a simple t-test is performed. What’s wrong with this procedure?

Wrong Significance

Solution: Use the time series model (or spatial model for spatial data) that incorporates the correlation structure
Interpretation Bias: Example 4.1

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?
**Interpretation Bias: Example 4.1**

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?

**Simpson Paradox - Aggregation Bias**

<table>
<thead>
<tr>
<th></th>
<th>Job 1</th>
<th>Job 2</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>salary</td>
<td>$8/hr</td>
<td>$18/hr</td>
<td>$10/hr</td>
</tr>
<tr>
<td>#</td>
<td>20</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>Men</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>$10/hr</td>
<td>$20/hr</td>
<td></td>
</tr>
<tr>
<td>#</td>
<td>80</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>
Interpretation Bias: Example 4.1

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?

**Simpson Paradox - Aggregation Bias**

<table>
<thead>
<tr>
<th>Job 1</th>
<th>Job 2</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>salary</td>
<td>#</td>
<td>salary</td>
</tr>
<tr>
<td>Men</td>
<td>$8/hr</td>
<td>20</td>
</tr>
<tr>
<td>Women</td>
<td>$10/hr</td>
<td>80</td>
</tr>
</tbody>
</table>

Due to pooled results, what’s true for the whole population may not be true for some (or all) of the subpopulations.
Example 4.2

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?
Example 4.2

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!

$$\alpha = P\{\text{type I error}\} = P\{\text{Rej } H_0|H_0\} = P\{H_1|H_0\} = 0.05$$

He is right if $H_1$ is indeed true:

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

If this last probability is 0.95 as he claimed, then

$$P(H_0|H_1) = P(H_1|H_0)$$
Example 4.2

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!

$$\alpha = P \{ \text{type I error} \} = P \{ \text{Rej } H_0 | H_0 \} = P \{ H_1 | H_0 \} = 0.05$$

He is right if $H_1$ is indeed true:

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

If this last probability is 0.95 as he claimed, then

$$P(H_0|H_1) = P(H_1|H_0) \quad \text{WRONG!}$$
Example 4.3: 1% of employees of a company use drug D, all employees are tested for the presence of this drug. The probability of finding a positive test result among users is 0.95, the probability of a negative test result among non-users is 0.90. If an employee has a positive test, what is the probability they are a drug user? Does knowing that the test is positive increase the chance of being a drug user?
Example 4.3: 1% of employees of a company use drug D(P(D)), all employees are tested for the presence of this drug. The probability of finding a positive test result among users is .95(P(+|D)) , the probability of a negative test result among non-users is .90(P(−|ND)) . If an employee has a positive test, what is the probability they are a drug users?

\[ P(D|+) = \]

Does knowing that the test is positive increase the chance of being a drug user?
Example 4.3: 1 % of employees of a company use drug D\((P(D))\), all employees are tested for the presence of this drug. The probability of finding a positive test result among users is \(.95(P(+|D|))\), the probability of a negative test result among non-users is \(.90(P(−|ND|))\). If an employee has a positive test, what is the probability they are a drug users?

\[
P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|ND)P(ND)} = \frac{(0.95)(0.01)}{(0.95)(0.01) + (1-0.9)(1-0.1)} = 0.0876
\]

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

Yes, 0.087 > 0.01, but not much. Q: what’s the probability if this person is tested positive again?

Solution: repeat the test!
Sampling Bias

- Self-selection, non-response
  sampled population ≠ target population

- Length or Size Bias
  The biasing function is proportional to the length or size of an object

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

- General bias
  not MAR
Example 5.1: Scleroderma

- Data: time from diagnosis of scleroderma (a rare disease) to death.
- 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.
Scleroderma: What happened

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.)
Example 5.2

- Tobacco Law Suit

  - Scott Zeager, Dean of Public Health at JH for Government, Blue-Cross & Blue-Shield
  - Don Rubin, Professor at Harvard for Tobacco Companies
    “Use Rubin’s multiple imputation method”

- Scientists differ on what triggered advent of humans
  “Early ancestors were prodded into existence in response to abrupt environmental changes during Pliocene epoch.” Sounds correct?
Solutions for Sampling Bias

Simple Solution: with l.b. sampling and $Y > 0$:

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

General Solution:

- Ideally, $Y_1, \ldots, Y_N \sim f_{\theta}$
- In reality, $Y$ is observed with a probability $w(y)$ if $Y = y$.

The observed obs are, given $n$:

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

2 cases: $N$ is known and unknown.
**Results**

- **Parametric:** easy

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

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

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

- **Semiparametric:** MM algorithm (Sun & Woodroofe, 97); survival data (Sun & Wang, 04)

- **Nonparametric:** need covariates, or two-stage sampling (Sun and Wang, 04)
Results - continued

- **Connection** to censoring and truncated observations

- **Testing** problems:

  \[ H_0 : \text{no bias} \quad vs \quad H_1 : \text{some bias} \]

  (Sun & Woodroofe, 99; Sun & Woodroofe, 04)
Procedure Bias: mistakes and solutions

- T-test requires that data are IID normal
  If any part of the assumption is wrong, the resulting p-value is misleading.

- If data are biased, use bias-corrected procedures

- Many problems are simultaneous in nature
  
  *Example:* 300,000 voxels, if a standard point-wise P-value is used and each pixel is claimed to be activated when its p-value < 0.05. How many false activation points are there under the null?

- A good data analysis involves using more than one statistics procedure

  EDA → model → analysis → diagnostics → final model → inferences → interpretation
Procedure Bias: mistakes and solutions

- T-test requires that data are IID normal
  If any part of the assumption is wrong, the resulting p-value is misleading.

- If data are biased, use bias-corrected procedures

- Many problems are simultaneous in nature
  Example: 300,000 voxels, if a standard point-wise P-value is used and each pixel is claimed to be activated when its p-value < 0.05. How many false activation points are there under the null? $5\% \times 300,000 = 15,000$. Data ransacking! Use procedures that correct for the multiplicity.

- A good data analysis involves using more than one statistics procedure
  EDA → model → analysis → diagnostics → final model → inferences → interpretation
Conclusion

For statisticians and scientists:

- Know context
  - Who? Individuals measured and observed
  - What? has been measured and observed
  - Why? Study Purpose

- Have good designs

- Avoid bias

- Do something about the bias if there is one.
Conclusion

For statisticians and scientists:

- Know context
  - Who? Individuals measured and observed
  - What? has been measured and observed
  - Why? Study Purpose

- Have good designs

- Avoid bias

- Do something about the bias if there is one.

For scientists:

- Involve a statistician or use statistical strategies early (from the design of an experiment to the analysis of the resulting data) not just later (for the analysis part only).