IMPROPER USE OF STATISTICS: MISUSES, MISINTERPRETATIONS AND BIAS

Douglas B. Brown, Ph.D.
IVT’s Statistics in Validation
June 20 – 22, 2017 (Philadelphia, PA)
The contents of this presentation represent the opinions and views of the speaker and do not necessarily represent any opinions or views from Charles River Laboratories, its subsidiaries, or any employee associated with Charles River Laboratories.
WHAT IS STATISTICS?

Definitions

• Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data (Wikipedia)

• The practice or science of collecting and analyzing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample (Google)

• Statistics is the study of the collection, analysis, interpretation and organization of data (The Oxford Dictionary of Statistical Terms, 2006)

• Statistics is the science of learning from data, and of measuring, controlling, and communicating uncertainty; and it thereby provides the navigation essential for controlling the course of scientific and societal advances (Marie Davidian and Thomas Louis, 2012)

• Statistics is the science of learning from data through not only the collection, organization, analysis, and interpretation of data, but also the presentation and communication of the results including its uncertainty (Douglas Brown)
WHAT IS STATISTICS?

Definitions

- **Data Collection** – one of the MOST CRUCIAL components of statistics
  - Great consideration must be dedicated to determine exactly what data is relevant to the issue under investigation and precisely how should that data be collected
  - Quality and reliability of any statistical analysis is directly related to the quality and reliability of the data
- **Data Organization** – enables the researcher to use data in an efficient and effective manner
  - Organized data allows for it to be easily assessed, identified, utilized, and understood
- **Data Analysis** – the *objective* evaluation of the results that can be directly derived from the data
  - Data analysis includes an understanding of the measurement limitations (i.e., uncertainty)
- **Data Interpretation** – involves the *subjective* communication of the results in the context of the issue under investigation
  - Interpretations offer explanations to the results and provide meaning for results that can be deduced from the data
- **Data Presentation** – presenting data accurately and effectively is absolutely essential in any analysis
  - Clearly communicating the results of a statistical study is crucial to the proper comprehension of the issue under investigation
WHY USE STATISTICS?

• We use data to study and describe (or summarize) past occurrences (descriptive statistics)
• We use data to evaluate patterns and create predictive models of future events (inferential statistics)
• The role statistics play is important when it comes to the conclusion of a research project.
  o Assists us in understanding and describing occurrences in our world
  o Enables us to draw **reliable conclusions** about observed (or evaluated) occurrences
  o Helps researchers to communicate research findings
    ➢ Lends support to hypotheses
    ➢ Gives credibility to the research methodology and conclusions
• For consumers of research, statistics...
  o Promotes their understanding of the research or project
  o Aids in their ability to evaluate the credibility and usefulness of information
  o Helps them to make appropriate decisions
WHAT IS THE IMPORTANCE OF STATISTICS IN SCIENTIFIC RESEARCH?

• Statistics guide researchers in the direction for proper characterization, summarization, presentation and interpretation of the results from their research.

• In other words, statistics provide the means for...
  o Determining the direction of current (and future) research projects
  o Considering the use of a representative sample or (if possible) the whole population
  o Identifying the techniques to be used in the data collection and observations
  o Describing the data collected and observed

• Statistics is very important when it comes to the conclusion of a research project.
  o Statistics assists us in understanding and describing occurrences in our world and enables us to draw reliable conclusions about those occurrences
  o Using statistics enable researchers to then communicate their finding lending support to their hypotheses and credibility to their research methodology and conclusions.
  o For consumers of research, statistics helps to promote their understanding of the research, aids in their ability to evaluate the credibility and usefulness of information, and make appropriate decisions.
STATISTICAL TESTS HELP TO ANSWER RESEARCH QUESTIONS...
...BUT YOU SHOULD NOT MISUSE STATISTICS

• Statistical tests can be misused, either by accident or design.
• This may lead to potential misinterpretation and ultimately, misrepresentation.
• Examples of this may include…
  o Altering the scales on a graph or chart to change the distribution of displayed data
  o Ignoring or removing high (or low) scores which may be considered to be inconvenient and thereby permitting the data may be presented in a more coherent manner
  o Focusing on certain variables while excluded others
  o Presenting correlation (the relationship between two variables) as causation (the reason for the result)
THERE ARE THREE KINDS OF LIES: LIES, DAMNED LIES, AND STATISTICS

- Because research is based on trust, research should be conducted in an ethical manner; findings ought to be presented truthfully.
- Deliberately misusing statistics should be neither acceptable nor tolerated by your organization and the scientific community.
- Misleading statistics are often employed to defend poor sampling techniques (or to fit a preconceived model or theory) that may be the result of biased or skewed data.
- With inexperienced researchers and analysts, a few of the main errors which may be encountered are
  - Inappropriate or unintentionally misused statistical analysis (or test)
  - Making improper inferences
  - Bias data collection, analysis and conclusions
MISUSE OF STATISTICS

• Misused statistics occurs when a statistical argument asserts a falsehood.
• Misused statistics may result from mistakes during an analysis
• This may result in poor decisions or conclusions and, ultimately, failed strategies
• While in some cases, the misuse may be accidental, in others, the intentional misuse of statistics is to achieve some gain or advantage.
• Misuses in statistics may occur because:
  o The source is a subject matter expert, not a statistics expert. The source may incorrectly use a method or interpret a result.
  o The source is a statistician, not a subject matter expert. An expert should know when the numbers being compared describe different things.
  o The subject being studied is not well defined.
  o Data quality is poor.
MISUSE OF STATISTICS

FAULTY ANALOGY – APPLES & ORANGES

• Comparing things that are not comparable or using unfair or impractical criteria of comparison.

• Here, the comparison or analogy is technically valid, but it has little or no practical meaning.

• Hence, the implied comparison has been designed to be misleading.

Example

Broccoli has significantly less fat than the leading candy bar!

While both broccoli and candy bars can be considered snacks, comparing the two in terms of fat content and ignoring the significant difference in taste, leads to the false comparison.
MISUSE OF STATISTICS

OVERCOMPLEXITY

• Graphs and data visualizations that are too complex to be interpreted by your audience.
• This may prevent data from being challenged and validated.

Example
• The pie chart looks cool, but is almost impossible to read.
  • Far too many similar colors & there are too many variables
  • Almost impossible to tell how one sliver compares to the others
• The correct usage of this pie chart would be to show how much of the total beer market one brewery owns.
MISUSE OF STATISTICS
DATA DREDGING or DATA FISHING

• This is a data mining technique where extremely large volumes of data are analyzed for the purposes of discovering relationships and/or patterns between data points.
• Data dredging has valid applications for exploratory data analysis.
• Seeking a relationship between data is not a data misuse per se, however, doing so without a hypothesis is.
• It is generally considered poor practice to draw conclusions using data dredging as it tends to find random patterns that are meaningless.

• Data dredging may be thought as a self-serving technique often employed for the unethical purpose of circumventing traditional data mining techniques in order to seek additional data conclusions that may not exist.
• As such, the ethical issues associated with data dredging is because it is an easy way to create a research paper that looks valid but is essentially auto-generated.
MISUSE OF STATISTICS
DATA DREDGING or DATA FISHING (example)

• The study included 18 different measured items—weight, cholesterol, sodium, blood protein levels, sleep quality, well-being, etc.—from 14 people.

• Each item had a small chance of paying off in the form of a “significant” result.

• The conventional cutoff for being “significant” is 0.05, which means that there is just a 5% chance that a difference between data sets exists.

• Increase the number of items searched, the better your chances of getting a false positive.

• So how many items do you need? \[ P(\text{success}) = 1 - (1 - p)^n \]

• With \( n=18 \) measurements, \( p = 0.05 \), there was about a 60% chance of getting something “significant”.

• This is also known as p-hacking—fiddling with your experimental design and data to push \( p \) under 0.05—and it’s a big problem.

MISINTERPRETATION OF STATISTICS

- Misinterpretation may occur unintentionally or intentionally.
  - When incorrect interpretation does not deliberately occur, is usually is the result of a lack of understanding or unwittingly arriving at the wrong conclusion through poor use of data.
  - If there is an intent to deceive or control a situation through partial truths and/or outright falsehoods, then that is simply “scientific fraud”.
  - Deliberate misinterpretation resulting in fraud is unethical and often illegal; when discovered and proven, fraud may ultimately end the scientific careers of those who engage in this practice.

- Misinterpretation in statistics may occur because:
  - Data is not correctly understood
  - Incomparable definitions are used
  - Information is deliberately mistaken
MISINTERPRETATION OF STATISTICS
CORRELATION VERSUS CAUSATION

• This is the case of an invalid assumption that because two things are correlated that one causes the other.
  o **Correlation** is a relationship between two things
  o **Causation** is a specific relationship between two things where one causes the other

• Assuming that correlation implies causation, is a logical fallacy also known as questionable cause.

**Example**

• Consider data that shows that gray cars are statistically involved in less accidents than other colors of car in a particular city.
  o It might be tempting to conclude that gray is a safer color for cars
  o However, there are a large number of alternative factors that could explain the data
    ➢ People who buy gray cars may be older or more conservative in their driving habits
    ➢ Gray cars may tend be more prevalent in certain conservative model type compared to non-gray, flashy cars
MISINTERPRETATION OF STATISTICS

FLAWED CORRELATION

• If you measure enough variables, eventually it will appear that some of them correlate.

Example

• A study has found in a particular area there is a correlation between an increase in car accidents in a month (A), AND an increase in bear attacks during that same month of (B). Possible explanations include:
  o Car accidents (A) cause bear attacks (B)
  o Bear attacks (B) cause car accidents (A)
  o Car accidents (A) and bear attacks (B) partly cause each other
  o Car accidents (A) and bear attacks (B) are caused by a third factor (C)
  o Bear attacks (B) are caused by a third factor (C) which correlates to car accidents (A)
  o The correlation is only chance

• Rationally, car accidents do not cause bear attacks.

• (A) and (B) are likely a result of a third factor, that being an increased population, due to possibly high tourism season during the month.

• It would be absurd to say that they cause each other, but this type of misinterpretation occurs
MISINTERPRETATION OF STATISTICS

GARBAGE IN, GARBAGE OUT

- Processes, procedures, methods, technologies, etc. require meaningful input to produce meaningful results.
- It is not uncommon for people to believe that technology can *magically* optimize results based on data without thinking about the nature of the data or the problem at hand.
- It is not possible to carry out an accurate statistical analysis using poor data.
- Also, it is not possible to generate meaningful results when good data is being evaluated by incorrect/inappropriate statistical methods.
“Bias” simply means prejudice.

This typically occurs when the data being collected is acquired from elicited attitudes or opinions.

For example, in a questionnaire, if the non-respondents (those who haven’t answered the questionnaire) are composed of a large percentage of a different socio-economic group, it could introduce bias (or systematic error) because there is an under-representation of that group in the study.

This type of bias is often observed where there are people who have

- Very strong opinions
- Great interest in the results of the research
- May derive some benefit from the results
- Loyalty or allegiance to express

over people who are may not have the desire to complete the questionnaire because they do not share the same passion, views or interests.
• Bias may also appear in a laboratory setting.
  o Only the “best analysts” perform the qualification and/or validation of a test method
  o Data collection was only acquired during a specific time of day in which temperature, humidity, or other environmental influences could impact the results

  This is applicable when these environment conditions are not going to be controlled during routine testing or production

• Bias in research leads to unrepresentative outcomes.

"You are completely free to carry out whatever research you want, so long as you come to these conclusions."
Displaying data is essential. However, be careful when evaluating comparison graphs.

- Data comparison may be displayed using different scales, different starting values (other than zero), methods of calculation (e.g., methods of calculations), and even dimension visualization.

**Example 1**

Both graphs display exactly the same data. However, the graph on the left makes the change appear to be larger than it really is because the vertical axis DOES NOT start at 0. The tick marks on left graph represent units of 1 while the tick marks on the right graph represent units of 20.

**Example 2**

The pie chart on the left is misleading due to its 3D presentation. In this pie chart, Item C appears to be as large as Item A. In fact, on a flat pie chart, Item C is actually less than half of Item A.
Biased scaling may be displayed in the form of a truncated graph (i.e., the vertical axis does not start at 0).

Truncated graphs are not always misleading when used to demonstrate small changes that cannot normally be visualized.

However, using an altered scaling to demonstrate a significant change when there is actually very little change is deceptive.

*Identical data*; is display on both graphs. However, in the truncated bar graph on the right, the data is displaying significant differences, whereas in the regular bar graph on the right, differences are barely visible.
Biased scaling may also be displayed without any scaling.

The lack of a starting value on the vertical axis makes it unclear whether the graph is truncated.

The lack of tick marks prevents the observer from knowing whether the graph bars are properly scaled.

Without a scale, the visual difference between the bars can be easily manipulated to fit a conclusion.
Biased labelling occurs when the use of suggestive or loaded words appear in the graph's title, axis labels, and/or caption causing the reader to be inappropriately lead toward a conclusion.

- Graph A is the original data
- Graph B has a reduced vertical axis scaling leading the observer to view a steeper increase year over year
- Graph C has an increased vertical axis scaling leading the observer to view a shallower or flatter increase year over year
The manner in which questions are phrased has a HUGE impact on the way an audience answers them.

Specific wording patterns have a persuasive effect and induce respondents to answer in a predictable manner.

Do not ask/write leading questions
- These questions tend to lead the reader to one side of the argument.
- Use neutral wording
  - Instead of asking “How cold is the room?”, ask “What is the temperature of the room?”

Avoid “loaded” questions
- These questions force the respondent into an answer that does not accurately reflect their opinion.
- Here, the question directs or guides the respondent toward an answer.
  - Where do you enjoy drinking beer?
- By answering this question, the respondent is announcing that they drink beer. However, many people dislike beer or will not drink alcohol and therefore can’t answer the question truthfully.
Do not ask/write “double-barreled” questions

- These questions ask the respondent to answer 2 questions at once.
- Questions should be asked so that only one item is being measured each time.
  - How satisfied or dissatisfied are you with the pay and work benefits of your current job?
  - How satisfied or dissatisfied are you with the pay and work benefits of your current job?
STATISTICAL BIAS

BIASED SAMPLING

- A biased sampling method occurs if it *systematically* favors certain outcomes over others.
- This type of bias may be intentional, but often it is not.
- Random sampling may not be random if patterns tend to be repeated.
  - For example, pulling production material vials (n = 300) from the same general time point
    - Selecting a vial from #10-40 (beginning), #130-160 (middle), #250-280 (end)
- If a researcher asks people to voluntarily participate in a survey, the resulting sample is called a "voluntary response sample." (similar to biased polling)
- *Voluntary response samples are always biased*: they only include people who choose volunteer, whereas a random sample would need to include people whether or not they choose to volunteer.
- Often, voluntary response samples oversample people who have strong opinions and under sample people who don't care much about the topic of the survey.
- Thus inferences from a voluntary response sample are not as trustworthy as conclusions based on a random sample of the entire population under consideration.
**REVIEW**

- **MISUSE**
  - Faulty Analogy – Comparing things that are not comparable
  - Over-complexity – Creating graphs & data visualizations that are too complex to be interpreted by the audience
  - Data Dredging - Data mining where extremely large volumes of data are analyzed for the purposes of discovering some relationship

- **MISINTERPRETATION**
  - Correlation versus Causation – Assuming a correlation implies causation
  - Flawed Correlation – Linking 2 variables that have to practical association
  - Garbage In, Garbage Out – Meaningful data must be analyzed by appropriate statistical methods

- **BIAS**
  - Misleading Data Visualization – Displaying data in a manner that over-emphasizes/de-emphasizes results
  - Biased Scaling – Altered scaling may demonstrate a significant change when there is actually very little change
  - Biased Labelling – Axis labels and/or caption causing the reader to be inappropriately leading toward a conclusion
  - Faulty Polling – Specific wording patterns have a persuasive effect & induce predictable answers
  - Sampling Bias – Sampling systematically favors certain outcomes over others
Clarity and a point of reference are always important when it comes to using statistics.
CONTACT US

Douglas B. Brown, Ph.D.
doug.brown@crl.com
610-407-1071

358 Technology Drive
Malvern, PA
19343

askcharlesriver@crl.com
www.criver.com
877.CRIVER.1