Statistics / Sample Size / Data Methods

Concept to Commercialization Boot Camp
CNY Biotech Accelerator

Thomas T. John, PhD.

Lecturer - Department of Mathematics
Independent Statistical Consultant
thjohn@syr.edu

October 13, 2017

Statistics: Big Picture View

- Large "group" to be understood but the entirety cannot be reached
- Examples:
  all customers of a specific product/service,
  all patients of a clinical treatment/device/technology,
  all products/devices made by a manufacturer, etc.
- Entirety can't be reached: Why?
  Too much monetary cost,
  Too much time involvement,
  Really impossible or illogical
- Large "group" to be understood
  $\uparrow$ called "population"

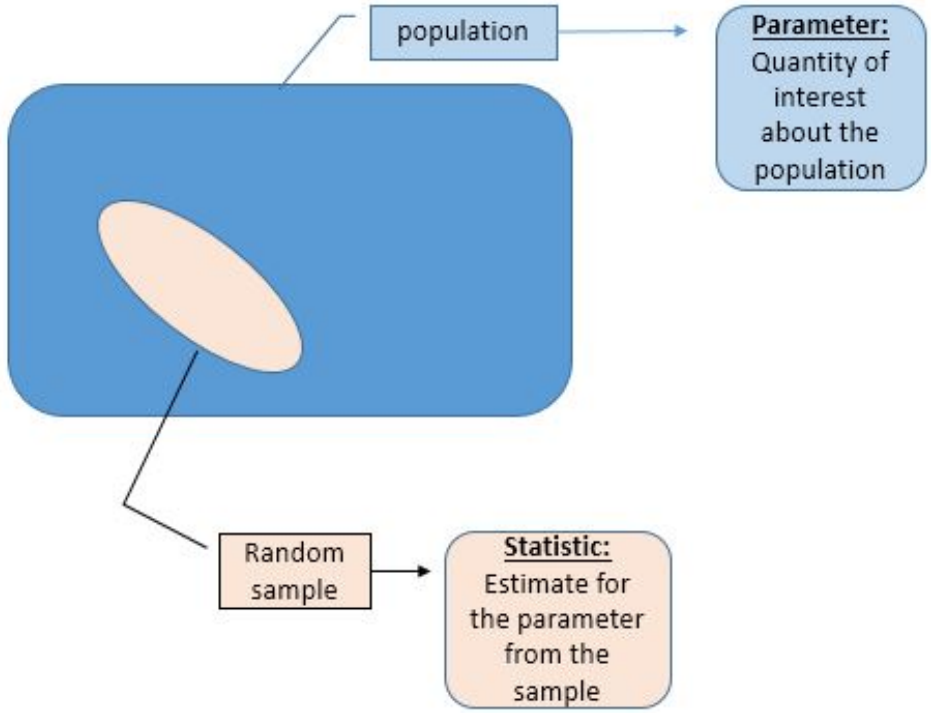
We sample instead...

- Core of statistical methodology is analogous to "tasting"

- A relatively "small" sample (relative to the "population") used to estimate and learn about the much bigger population.
- Hence a lot riding on proper sampling and we need to be careful with: study design, data collection, and sample size

Some terminologies

- Parameter: Quantity associated with the population we want to measure
- Statistic: Corresponding estimate we get from a sample
- Margin of Error: Assessment of error, due to sampling variability, of how the estimate might be off



Note: Keep in mind that "population" does not necessarily mean "living" things and could possibly be inanimate objects. Also, "population" refers to *ALL* members (the entire group) that the study wants to target and speak about.

Statistical Considerations

One should design the study properly

- Needs to be big enough, representative enough
  - so that we get a good read on the entire population
  - small samples have fluctuations (high variability)
    - in turn have high margin of error and low power
- Needs to be unbiased
  - sample not all concentrated in a few areas
  - sample not all leaning in certain directions

  So select subjects randomly
  - (analogous to stirring before sample tasting)
- "Controlled" and blinded
  - All aspects other than the study features need to be balanced across the board
  - Subjects and study handlers shouldn't know who/what is in treatment groups vs. control groups.

Statistical Control

- All other variables that might affect the study features should be controlled as best as possible.
- During the study design, carefully list factors that might influence the observations.
  – e.g., time of day, order of measurements, clothing worn, subjects/observers' familiarity, etc.
- If needed, use a randomization scheme to help balance out effects from extraneous factors
- A very simple randomization scheme:
  – Cyclical (say, 4 factors)
  $\Rightarrow$ ABCD, BCDA, CDAB, DABC, ABCD,...
- Complex factors $\Rightarrow$ requires complex randomization

Sample Size Determination

Margin of Error & Sample Sizes

Confidence intervals usually[1] have the form

$$\text{Estimate (Statistic)} \pm \underbrace{\text{Critical Value}^{\,2} \times \frac{\text{Standard Deviation}}{\sqrt{\text{Sample Size}}}}_{\text{this is called the margin of error}}$$

- Note that increase in sample size decreases margin of error (all else being same)
- All else fixed: sample size can be determined
- Quick rough example: 99% confidence, $\sigma = 2$, $n = 100$

$$\text{Margin of Error} = 2.576 \times \frac{2}{\sqrt{100}} \approx \frac{5}{10} = 0.5$$

[1] Technicalities: Assuming location parameter, for a single population, and a symmetric sampling distribution. Many confidence intervals in practice have this form.

[2] Critical value from an appropriate distribution for the desired confidence level

Sample Sizes (continued)

$$\text{Recall margin of error (MoE)} = \text{Critical Value} \times \frac{\text{Std. Dev.}}{\sqrt{\text{Sample Size}}}$$

- Standard deviation plays out differently for

Mean (Parameter for a Quantitative Variable)	vs.	Proportion (Parameter for a Categorical Variable)

- Sample size requirements for proportions are quite high
- For dichotomic (binary) proportions, the margin of error (MoE) is worst when proportions are 50%/50%.
- Quick example: 99% confidence, $n = 100$

$$\text{MoE} = 2.576 \times \frac{\sqrt{0.50 \times 0.50}}{\sqrt{100}} \approx \frac{2.5 \times 0.5}{10} = 0.125$$

Notice though: MoE of $\pm 12.5\%$ is a bit high for %'s.

Sample Sizes for Proportions

Below are sample sizes to estimate a proportion at 99% confidence level (critical value = 2.576)

MoE	Sample Size
± 5.0%	664
± 4.5%	820
± 4.0%	1,037
± 3.5%	1,355
± 3.0%	1,844
± 2.5%	2,654
± 2.0%	4,147
± 1.5%	7,373
± 1.0%	16,588
± 0.5%	66,349

Additional Notes on Sample Sizes for Proportions

- One intuitive rationalization of different behavior of proportions: For small samples, 1 observation being in one group vs. not, can make %'s (proportions) swing dramatically
  e.g, 8/15=53.3% to 9/15=60%
- Recall MoE for a proportion is

$$\text{margin of error (MoE)} = \text{Critical Value} \times \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

- Instead of the worst case scenario of $p = 0.50$, in 2-Stage or Multi-Stage or Sequential studies (e.g., usability testing), one uses the updated information on p as we proceed
- Note on usability sample sizes: It is known that a small sample of 5-8 is sufficient to detect major usability issues. **BUT** note that this is not the same as successful usage rate being significantly higher than the acceptance criteria. To claim significant difference, a bigger sample size will be needed.

Sample Sizes for Significance Testing

- Sample sizes for significance tests on population mean (quantitative variable), using a parallel logic, are based on: significance level (α), power $(1-\beta)$, and effect size.
- α controls falsely finding differences when they don't truly exist (finding defendant guilty, when truly innocent)
- β controls falsely not detecting differences when they do indeed exist (finding defendant not guilty, when truly guilty)
- Effect size is the true difference Δ that exists in the population desired to be detected by the study, scaled by the measure of spread. That is:

$$\text{Effect Size} = \frac{|\Delta|}{\sigma} = \frac{|\mu_1 - \mu_0|}{\sigma}$$

- Sample size computation formula (for simple scenarios):

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\text{Effect Size}} \right)^2$$

Sample Sizes for Significance Testing (cont'd)

- Sample size (for simple scenario - illustration):

$$n = \left(\frac{Z_{1-\alpha/2} + Z_{1-\beta}}{\text{Effect Size}} \right)^2 = \left(\frac{1.96 + 0.84}{0.5} \right)^2 = \left(\frac{2.8}{0.5} \right)^2 \approx \mathbf{32}$$

- Here 1.96 is for significance level $\alpha = 0.05$, 0.84 is for power of 0.80, and effect size of 0.5 is considered "medium"[3]
- To detect a "small" effect size of 0.2
  - sample size needed $(2.8/0.2)^2 = \mathbf{196}$.
- Rationale for the bigger sample size for smaller effect size: Detecting smaller differences require better resolution (i.e., lower noise and less variability) provided by larger sample sizes.
  - i.e., Bigger effects are detectable without looking "hard"

[3] Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences.*

Why $\alpha = 5\%$?

From Sir.Fisher, Father of Modern Statistics

> "If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."
>
> — Sir Ronald Fisher (1926)

Statistical Analysis Methods

Analysis Methods

- Exploratory analyses must be done first
  - Numerical summaries
  - Graphical summaries (e.g., boxplots)
  - Summarize in total and sliced by groups
  - Goal: Detect anomalies, data entry errors, overall sense
- Clean data as needed, structure, and organize
- Sometimes outliers are not "bad"
  – could provide important information
- Select appropriate inferential analyses methods
  - Inferential: findings extending to the "population"
  - Chosen based on research study goals
  - ...and associated data/variable types

Methods by Variable Types

Variable Type	**Analysis Method**
Quantitative variable compared between 2 groups	Paired t or 2-sample t
Dichotomous (binary categorical) compared between 2 groups	2-sample proportions
Quantitative compared between 2+ groups	One-Way ANOVA
Quantitative compared against 2 factors (each of 2+ groups)	Two-Way ANOVA
Quantitative response against 1 or more predictors	Linear Regression (OLS)
Categorical variable against another categorical	Chi-Square Test of Association

Methods Illustration

A few illustrations of analyses to follow. Analyses are driven by the context and the variable types. Data used: "Heart"[4] dataset from the UC-Irvine repository:

	Age	Gender	ChestPainType	RestingBP	SerumChol	BlSugarHigh	RestingECG	MaxHeartRate	ExercIndAng
1	70	1	4	130	322	0	2	109	
2	67	0	3	115	564	0	2	160	
3	57	1	2	124	261	0	0	141	
4	64	1	4	128	263	0	0	105	
5	74	0	2	120	269	0	2	121	
6	65	1	4	120	177	0	0	140	

[4] Source: archive.ics.uci.edu/ml/datasets/Statlog+ (Heart); Donor: David W. Aha (aha at ics.uci.edu); Principal Investigator: Robert Detrano, M.D., Ph.D., V.A. Medical Center, Long Beach and Cleveland Clinic Foundation.

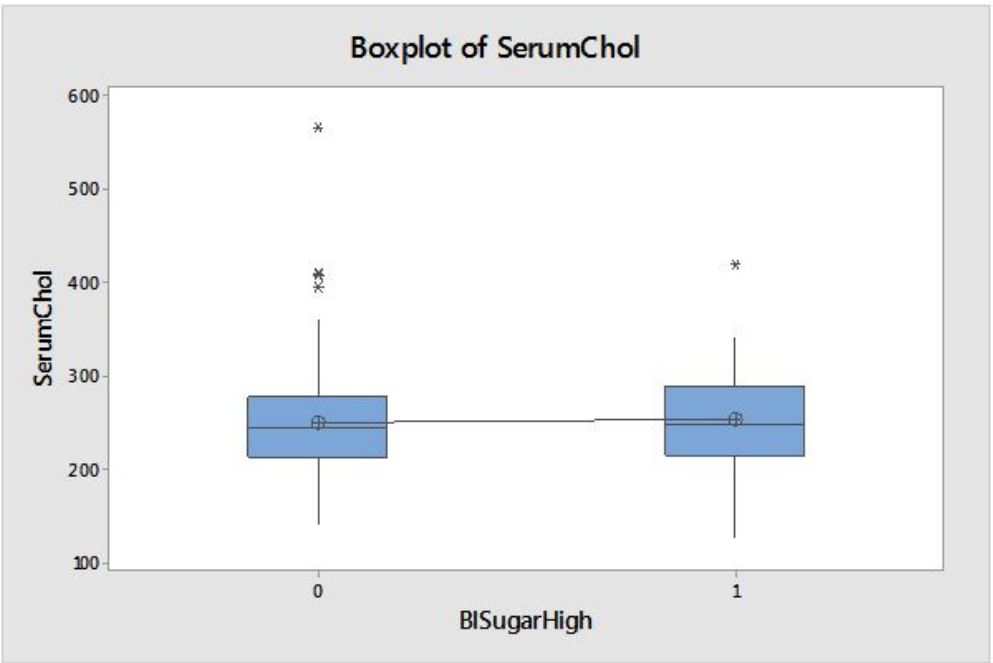
Statlog (Heart) Analyses: 2-sample t

Comparison of Serum Cholesterol (in mg/dl) between subjects with high blood sugar and low blood sugar:

```
Two-sample T for SerumChol

BlSugarHigh    N   Mean  StDev  SE Mean
0            230  249.1   51.8      3.4
1             40  252.8   51.5      8.1

Difference = μ (0) - μ (1)
Estimate for difference:  -3.66
95% CI for difference:  (-21.37, 14.05)
T-Test of difference = 0 (vs ≠):
      T-Value = -0.41  P-Value = 0.680  DF = 53
```



"1" = fasting blood sugar higher than 120 mg/dl; "0" = fasting blood sugar $<$ 120 mg/dl;

Statlog (Heart) Analyses: 1-Way ANOVA

Comparison of Maximum Heart Rate between 4 chest pain type groups:

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
ChestPainType	3	18797	6265.8	13.27	0.000
Error	266	125562	472.0		
Total	269	144359			

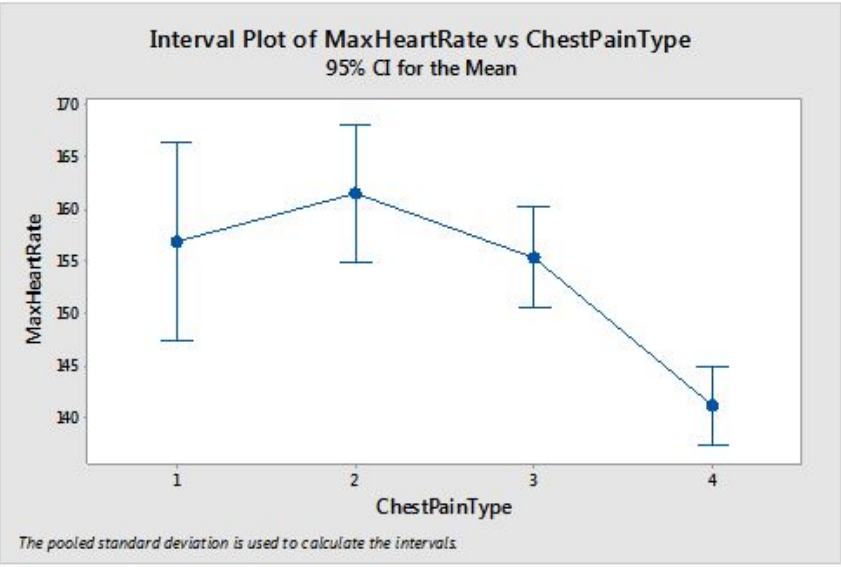
Model Summary

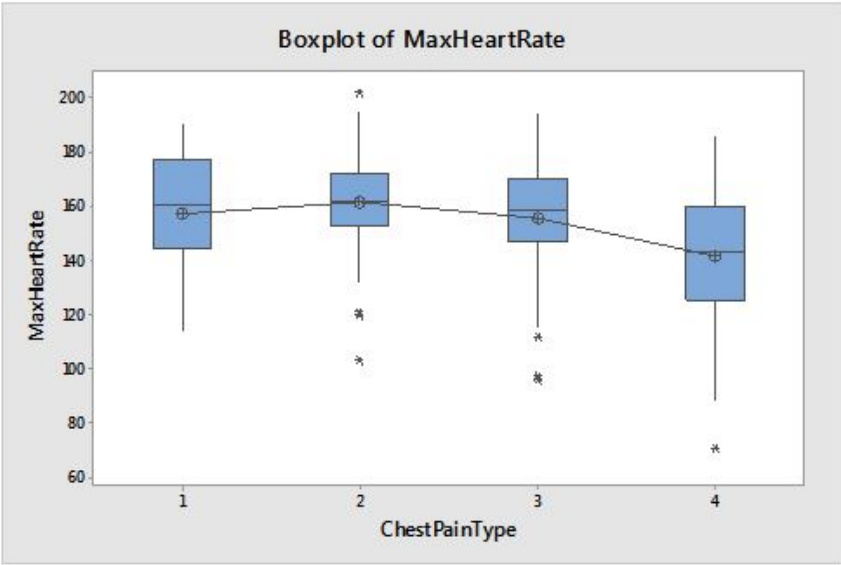
S	R-sq	R-sq(adj)	R-sq(pred)
21.7264	13.02%	12.04%	10.49%

Means

ChestPainType	N	Mean	StDev	95% CI
1	20	156.90	21.74	(147.33, 166.47)
2	42	161.55	20.11	(154.95, 168.15)
3	79	155.38	19.28	(150.57, 160.19)
4	129	141.20	23.55	(137.44, 144.97)

Pooled StDev = 21.7264





Chest Pain Types: 1=typical angina; 2=atypical angina; 3=non-anginal pain; 4=asymptomatic

Advanced Methods

- Categorical response against 1 or more predictors:
  - Typical scenario arising in classification issues
  - e.g., predict loan applicant could default, predict email could be spam, predict website visitor likely to buy, etc.
  - A few popular methods: Logistic Regression, Classification Tree, Support Vector Machines
  - Called supervised learning methods/algorithms
- Several quantitative variables (characteristics/features) of a large group, with the hypothesis that there might be some natural groupings internally:
  - Typical scenario arising in customer segmentation
  - Methods: k-Means (centroid), Hierarchical (linkage)
  - Called unsupervised learning methods/algorithms
- Time series: Analyses of a certain variable tracked overtime (e.g., ARIMA, Exponential Smoothing, Holt-Winters)
- Data reduction methods: Large number of variables to manageable number of variables (e.g., principal component, factor, SEM, PLS)