Science, Statistics, and Getting it Right: Interactive Discussion of Common Problems

PISA 2017 Workshop

Sponsored by:
ASIP Committee for Career Development & Diversity
ASIP Education Committee

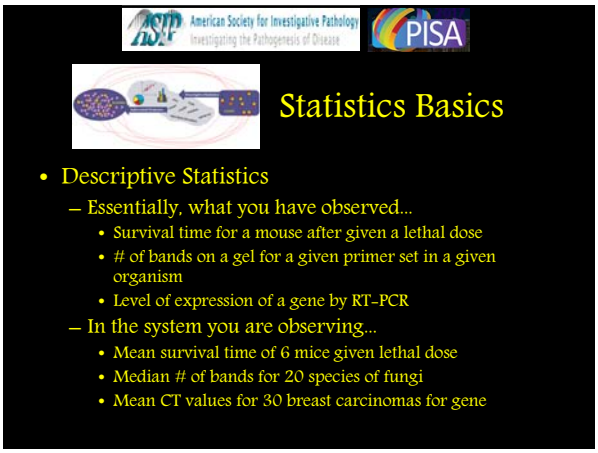
Dan Milner
Chief Medical Officer
American Society for Clinical Pathology
Special Thanks to Mark Sobel

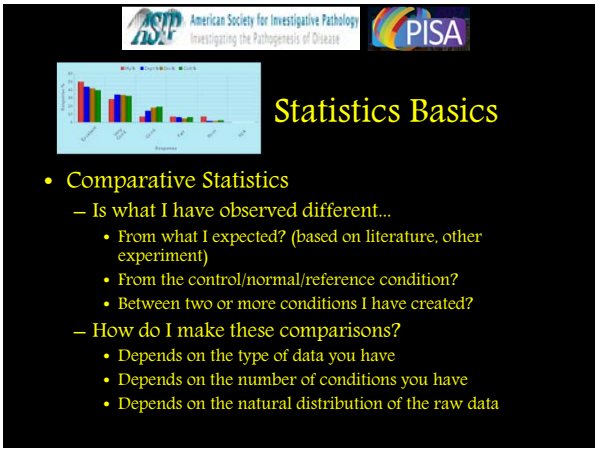
Overview

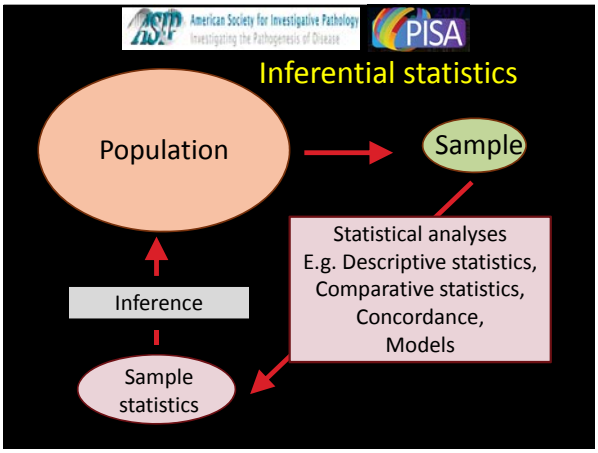
- Introduction
- Large group discussion of cases
- Questions and parting thoughts

Why Statistics for Science?

- Science is careful observation of the natural world
  - Responses to perturbations/alterations
  - Changes in the presence of variable conditions
- First observation… "Is this true?"
- Second observation… "This may be true?"
- Nth observation… "This is most likely true!"
- Statistics allows us to determine if what we have observed given the number of observations we have made is "significant" (i.e., likely to represent the natural world)


American Society for Investigative Pathology
Investigating the Pathogenesis of Disease
PISA
Statistics Basics
• Descriptive Statistics
– Essentially, what you have observed...
• Survival time for a mouse after given a lethal dose
• # of bands on a gel for a given primer set in a given organism
• Level of expression of a gene by RT-PCR
– In the system you are observing...
• Mean survival time of 6 mice given lethal dose
• Median # of bands for 20 species of fungi
• Mean CT values for 30 breast carcinomas for gene


American Society for Investigative Pathology
Investigating the Pathogenesis of Disease
PISA
Statistics Basics
• Comparative Statistics
– Is what I have observed different...
• From what I expected? (based on literature, other experiment)
• From the control/normal/reference condition?
• Between two or more conditions I have created?
– How do I make these comparisons?
• Depends on the type of data you have
• Depends on the number of conditions you have
• Depends on the natural distribution of the raw data


American Society for Investigative Pathology
Investigating the Pathogenesis of Disease
PISA
Inferential statistics
Population
Sample
Statistical analyses
E.g. Descriptive statistics,
Comparative statistics,
Concordance,
Models
Inference
Sample
statistics

What test can we use?

What type of data do we have?

- Continuous dependent variable
  - Dependent variable is normally distributed → Parametric tests
  - Dependent variable is not normally distributed → Non-parametric tests
- Categorical dependent and independent variables
  - Unmatched/unpaired data for independent variables
  - Matched data for independent variables

Comparing 1 sample mean to a hypothesized mean	Comparing means in 2 independent samples	Comparing means in 2 paired observations	Comparing means in 3 or more independent samples
One sample t-test	Two independent samples t-test	Paired t-test	Analysis of variance (ANOVA)

What test should we use?

What type of data do we have?

- Interval or continuous dependent variable
  - Dependent variable is normally distributed → Parametric tests
  - Dependent variable is not normally distributed → Non-parametric tests
- Categorical dependent and independent variables
  - Unmatched/unpaired data for independent variables
  - Matched data for independent variables

Comparing 1 sample median to a hypothesized median	2 independent samples	Paired or correlated observations	3 or more independent samples
One sample median test	Wilcoxon-Mann Whitney test	Wilcoxon signed rank test	Kruskal Wallis test

What test should we use?

What type of data do we have?

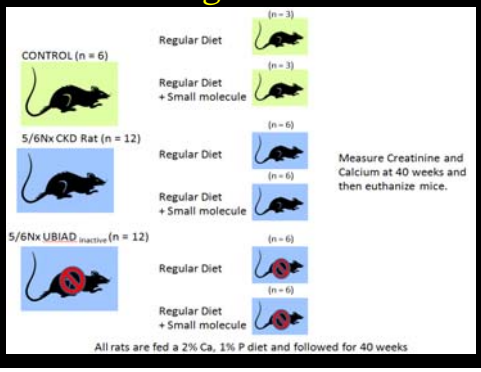
- Interval or continuous dependent variable
  - Dependent variable is normally distributed
  - Dependent variable is not normally distributed
- Categorical dependent and independent variables
  - Unmatched/unpaired data for independent variables
  - Matched data for independent variables

Comparing 1 sample proportion to a hypothesized proportion	Independent samples	Paired or correlated observations
Binomial test	Chi-square test, Fisher's exact test	McNemar's test

What Statistics Are (and are not)

- Description of the dataset we have and its parameters √
- Mathematical assessment of the significance of an observed result √
- Relevance of our data and findings to other reports in the literature √
- Tool for showing the importance of our work

Vignette 1



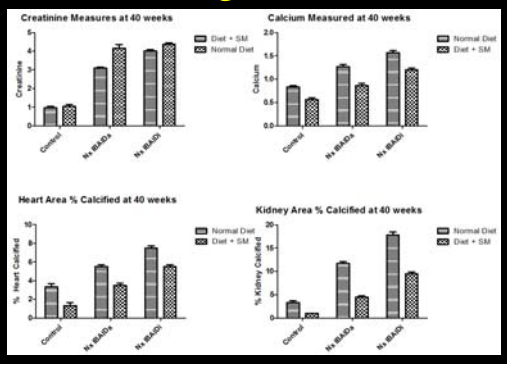
Vignette 1

- Common experimental model with:
  - Control: natural history of mouse with the wild-type genetic background
  - Experimental Model: Disease-specific mouse
  - Knockout: Loss of one specific gene (non-lethal)
  - Conditions: Regular diet or + small molecule
- Common challenge:
  - Small number of each group
  - Statistics of small numbers

American Society for Investigative Pathology
Investigating the Pathogenesis of Disease

PISA

Vignette 1



American Society for Investigative Pathology
Investigating the Pathogenesis of Disease

PISA

Vignette 1

- What kind of variables do you have?
- What kind of test can you perform?
- What kind of result are you looking for?
- How do you perform the test?
- What is the result of a two way ANOVA?

American Society for Investigative Pathology
Investigating the Pathogenesis of Disease

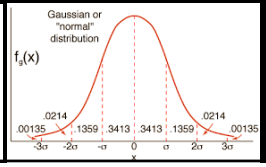
PISA

Vignette 1

- What kind of variables do you have?

Mouse (categorical) : controls, NX UBAID+, NX UBAID-
Diet (categorical): normal, normal + small molecule
Creatinine (continuous), Calcium (continuous),
heart % area calcified (continuous), kidney % area calcified (continuous)

Categorical vs. Continuous variables



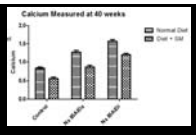
American Society for Investigative Pathology
Investigating the Pathogenesis of Disease

PISA

Vignette 1

- What kind of test can you perform?

Differences in UBIAD mouse calcium levels (excluding control):
Continuous variable, two groups = T-test (GD) or Rank Sum (NND)
Difference in mouse % heart calcified (all mice):
Continuous variable , three groups = ANOVA (GD) or Kruskal-Wallis (NND)
Difference in diet creatinine levels (excluding knock out):
Continuous variable, two groups = T-test (GD) or Rank Sum (NND)
Difference by mouse and diet in calcium level (all mice):
Continuous variable, three by two groups: two way ANOVA (GD) or others*

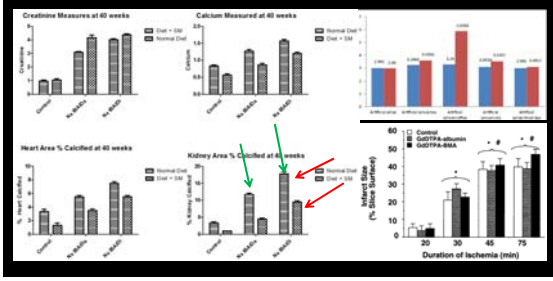


American Society for Investigative Pathology
Investigating the Pathogenesis of Disease

PISA

Vignette 1

- What kind of result are you looking for?

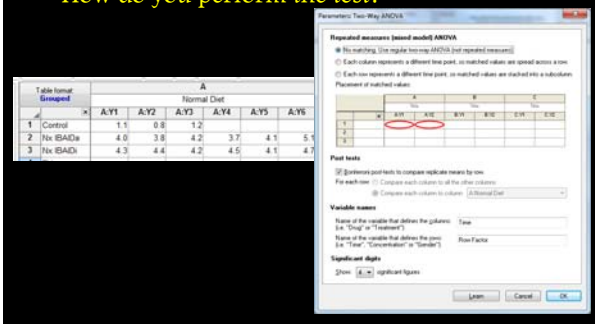


American Society for Investigative Pathology
Investigating the Pathogenesis of Disease

PISA

Vignette 1

- How do you perform the test?



Vignette 1

- What is the result of a two way ANOVA?

Table Analyzed	Calcium			
Two-way ANOVA				
Source of Variation	% of total variation	P value		
Interaction	0.57	0.4388		
Column Factor	25.28	< 0.0001		
Row Factor	60.75	< 0.0001		
Source of Variation	P value summary	Significant?		
Interaction	ns	No		
Column Factor	***	Yes		
Row Factor	***	Yes		
Source of Variation	Df	Sum-of-squares	Mean square	F
Interaction	2	0.01800	0.009000	0.8526
Column Factor	1	0.8008	0.8008	75.87
Row Factor	2	1.925	0.9623	91.17
Residual	24	0.2533	0.01056	
Number of missing values	6			
Bonferroni posttests				
Normal Diet vs Diet + SM				
Row Factor	Normal Diet	Diet + SM	Difference	95% CI of diff.
Control	0.8333	0.5667	-0.2667	-0.4826 to -0.05077
Nx IBAIDa	1.267	0.8667	-0.4000	-0.5527 to -0.2473
Nx IBAIDi	1.567	1.200	-0.3667	-0.5193 to -0.2140
Row Factor	Difference	t	P value	Summary
Control	-0.2667	3.179	P < 0.05	*
Nx IBAIDa	-0.4000	6.743	P<0.001	***
Nx IBAIDi	-0.3667	6.181	P<0.001	***

Vignette 1: Interim Summary

- When designing an experiment, determine what the data will look like and what statistical tests will be used.
- When possible, calculate a sample size based on 80% power and alpha of 0.05
  - Normally distributed data
  - If non-normal, can approximate with normal
- For any given question, there is usually only one "correct" statistical test

Vignette 2

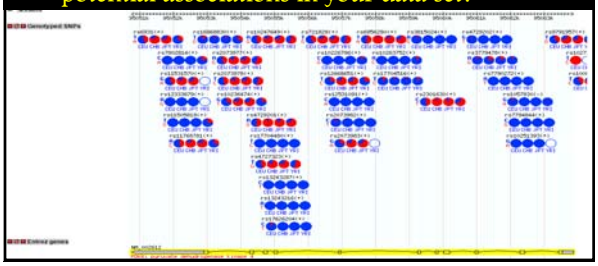
- 1175 healthy subjects (43% Caucasian, 33% African or African American, 24% Hispanic/Latino)
- College students provided DNA sequencing and detailed questionnaire
- Follow up survey every 5 years for the next 25 years
- Sequence each patient to 40X coverage, comparative genomic hybridization
  - Mutations included insertions/deletions, single nucleotide polymorphisms, and gene duplication.
- Questionnaire: diabetes, hypertension, malignancy, infections, diet, and exercise habits.
- Free pedometer connected to the internet
- Pyruvate dehydrogenase kinase 4 (PDK4) mutation in mice shown to decrease activity and lead to obesity

Vignette 2

- How would you go about investigating any potential associations in your data set?
- What considerations (statistical) are important in thinking about this question?
- How should the pedometer data be parsed at the individual variable level for analysis?

Vignette 2

- How would you go about investigating any potential associations in your data set?



Vignette 2

- What considerations (statistical) are important in thinking about this question?

Number of Samples: 1175 samples (n) but 100K+ variables! - > Hypothesis generation…

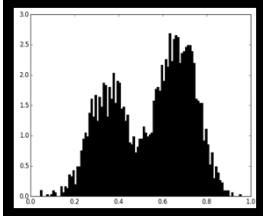
Distribution of SNPs: HapMap vs. our data set… New SNPs?

Other variables: Levels of exercise during college, lifestyle, etc.

Censored Data: Highly active individuals die of unrelated causes

Vignette 2

- How should the pedometer data be parsed at the individual variable level for analysis?

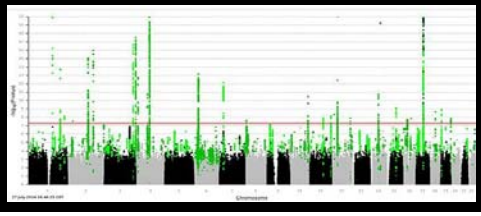


	Major Allele	Minor Allele
<50%	A	B
>50%	C	D

Vignette 2

	Major Allele	Minor Allele
<50%	A	B
>50%	C	D

- Analysis:
  - Fishers Exact test for EACH SNP (100K+)
  - Alternative: T-test for each SNP (pedometer)



Vignette 2: Interim Summary

- Genetic studies have a common problem:
  - n (number of patients) << k (number of variables)
- Same or similar statistical tests with special modifications (corrections)
- Should be considered "hypothesis generating" exercises
  - Always need experimental or epidemiological confirmation and, in best cases, biological evidence
- Gene expression data have similar challenges with equivalent solutions

Summary and Conclusions

- Your observations (science) are always important...
  - Positive OR Negative results
- If your experiments were designed correctly...
  - Including statistics BEFORE you begin
- And you complete them in a reproducible manner...
  - Repetition by others is required to get to the Nth observation!

Questions?