Hypothesis testing and p-value pitfalls

Sandra Taylor, Ph.D.
August 14 & 28, 2019
We are video recording this seminar so please hold questions until the end.

Thanks
Seminar Objectives

- Understand framework of traditional null hypothesis significance testing
- Be able to correctly interpret p-values
- Understand confidence intervals
- Appreciate multiple testing issues and know corrections
Cardiovascular Disease Dataset

- 600 Subjects
- Presence/absence of coronary artery disease
- Demographics – age, sex, race, BMI
- Inflammatory biomarkers – CRP, LLPLA2, SAA, PTX3, FIBRIN, and HOMA

I will use this dataset to illustrate various points.
Primary and Secondary Aims

- **Primary Aim:** Do HOMA levels differ between CAD(+) and CAD(-) subjects?
  - Does the mean of HOMA levels differ between CAD(+) and CAD(-) subjects?

- **Secondary Aims:** Do CRP, LLPLA2, SAA, PTX3, and FIBRIN levels differ between CAD(+) and CAD(-) subjects?
The truth is out there.

If we had data from every person in our population we would know with certainty the difference in the group means.
Since we can’t observe every individual in a population, we collect a sample from the population.

We seek to make inferences (i.e., make decision regarding our hypothesis) about the entire population based on the sample.
Sampling yields variability

Between Subject Variability

- Values differ between subjects
- Standard deviation

Between Sample Variability

- Estimates differ between studies
- Standard error
Illustration of between study variability
How do we go from a sample to a decision? – Statistics!

1. Assume $H_0$ is true.
2. Sample the population.
3. Determine probability of observing sample data (i.e., conduct statistical test).
4. Reject or Fail to Reject $H_0$.
5. Infer about Population.
Null Hypothesis Significance Testing Framework

- In null hypothesis significance testing, we posit a null hypothesis
  - $H_0$: Mean CAD(+) = Mean CAD(-)

- We seek to reject the null hypothesis in favor of an alternative hypothesis.
  - $H_a$: Mean CAD(+) $\neq$ Mean CAD(-)

- Notice the simplicity of $H_a$
  - It's just that they aren't equal. No info on magnitude
## Hypothesis Testing: Ideas on Trial

<table>
<thead>
<tr>
<th>Courtroom</th>
<th>Hypothesis Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Presume innocent</td>
<td>- Assume null hypothesis is true</td>
</tr>
<tr>
<td>- Present and evaluate evidence</td>
<td>- Gather and evaluate evidence</td>
</tr>
<tr>
<td>- Jury verdict</td>
<td>- Statistical test result</td>
</tr>
<tr>
<td>- Guilty – ‘beyond a reasonable doubt’ standard avoids incorrect conviction</td>
<td>- Reject $H_0$ – significance level ($\alpha$) controls incorrect rejection</td>
</tr>
<tr>
<td>- Acquittal – not proof of innocent</td>
<td>- Fail to Reject $H_0$ – not unlikely to observe data</td>
</tr>
<tr>
<td>- Incorrect guilty verdict worse than incorrect acquittal</td>
<td>- Does not prove $H_0$ is true</td>
</tr>
<tr>
<td></td>
<td>- False positive worse than false negative</td>
</tr>
</tbody>
</table>
Absence of evidence is NOT evidence of absence!

**Courtroom**

**Conviction**: Beyond a reasonable doubt

**Acquittal**: Reasonable doubt – evidence insufficient

**Hypothesis Testing**

Reject $H_0$: Probability of observing data if null hypothesis is true is unlikely

Fail to Reject $H_0$: Probability of observing data if null hypothesis is true is not unlikely
## Hypothesis Testing: Ideas on Trial

<table>
<thead>
<tr>
<th></th>
<th>( H_0 ) False (Defendant is Guilty)</th>
<th>( H_0 ) True (Defendant is Innocent)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reject ( H_0 )</strong></td>
<td>Correct decision</td>
<td>Type I error (( \alpha ))</td>
</tr>
<tr>
<td><strong>Fail to Reject ( H_0 )</strong></td>
<td>Type II error (( \beta ))</td>
<td>Correct decision</td>
</tr>
</tbody>
</table>
Return to CAD Example

**CAD(+)**
- Mean = 3.5
- SD = 4.5

**CAD(-)**
- Mean = 2.7
- SD = 3.4
Does HOMA differ between CAD(+) and CAD(-) Groups?

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>SD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAD(+)</td>
<td>0.84</td>
<td>0.83</td>
<td>310</td>
</tr>
<tr>
<td>CAD(-)</td>
<td>0.67</td>
<td>0.73</td>
<td>290</td>
</tr>
</tbody>
</table>

- Define the Null (Ho) and Alternative (Ha) Hypotheses

Ho: Mean HOMA levels do not differ between CAD(+) and CAD(-)
Ha: Mean HOMA levels differ between CAD(+) and CAD(-)

- Calculate test statistic
  \[ t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \]
  - \( t = 2.77 \)

- Calculate the probability of observing a \( t \geq \pm 2.77 \) if the null hypothesis was true!
  \[ p\text{-value} = 0.006 \]

\( t \) is calculated using the formula:
\[ t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \]
What exactly are p-values?

- Probability that you would observe a test statistic at least extreme as you did if the null hypothesis is true
  - We know the distributions test statistics under $H_0$ which allows us to calculate p-values

- $P = 0.006$ – small probability so reject null hypothesis

- Did not prove alternative hypothesis
What’s so special about 0.05?

- Origin attributed to Ronald Fisher (1890-1962)
- English statistical evolutionary biologist
- Authored *Statistical Methods for Research Workers*
  - Very influential text
  - Provided probabilities between coarse bounds rather than very detailed tables – these were widely copied

“The value for which \( P=0.05 \) or 1 in 20; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant.”
What if we had a different sample?
Statistical vs. Clinical Significance

- Statistically significant is not necessarily clinically significant
- Not statistically significant is not necessarily not clinically significant
Point estimates and confidence intervals more informative

- P-values help in decision-making about the null but provide no additional useful information
- Point estimates – size and direction of differences/relationships
- Confidence intervals – precision of estimates
What are confidence intervals and what do they tell us?

- Define a range that includes the true value with a high degree of confidence, typically 95%.

- The confidence interval is NOT the probability that the true value is within the confidence limits.
  - The true value is either in the limits or not with probability 1 or 0.

- Repeated sampling and construction of confidence limits will encompass the true value 95% of the time.
Illustration of confidence intervals
Type II Errors and Power

- Significance level ($\alpha$) limits type I error
  - Set fairly low to minimize false positives (e.g., wrongly convicting an innocent person)

- Type II errors ($\beta$) are false negatives
  - failing to reject the null hypothesis when it is false

- Power is probability of rejecting Ho when it is false

- Power = $1 - \beta$
What determines the power of a test?

- **Size of the effect, e.g., difference between groups**
  - Larger effect \(\rightarrow\) more power

- **Variability of the data**
  - Greater variability \(\rightarrow\) less power

- **Sample size**
  - Larger sample \(\rightarrow\) more power

- **Significance level \((\alpha)\)**
  - Smaller significance level \(\rightarrow\) less power
How does sample size affect power?

- Assumes difference in means of 0.6 with SD = 1. So the two groups truly differ.

<table>
<thead>
<tr>
<th>Sample Size (Per group)</th>
<th>Number of Rejections (Power)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>18.0%</td>
</tr>
<tr>
<td>30</td>
<td>60.0%</td>
</tr>
<tr>
<td>50</td>
<td>86.0%</td>
</tr>
<tr>
<td>100</td>
<td>99.0%</td>
</tr>
</tbody>
</table>

- If you only have 10 samples per group, you will reject the null hypothesis about 18% of the time if the true difference in 0.6.
Hypothesis Testing: Summary

- Significance level controls type I error (false positives)
- Power controls type II error (false negatives)
- P-values aid in decision making about $H_0$
- Point estimates and confidence intervals are more informative than p-values
- Keep in mind between sample/study variation
- Keep in mind the sample size
Multiple Hypothesis Testing

- What is it?
- What does it mean to me?
- What do I do about it?
What is Multiple Testing?

- Conducting many hypothesis tests simultaneously
- **Examples:**
  - Comparing heart rate, respiratory rate, blood pressure, SOFA scores, mean arterial pressure, and additional laboratory values
  - Comparing multiple patient outcomes, e.g., 28-day mortality, in-hospital mortality, LOS, ICU LOS, ventilator days, readmissions
  - Evaluating scores from a battery of behavioral assessments
What does it mean to me?

- **Type I error not controlled at 0.05**
  - Recall Type I error = probability of rejecting the null hypothesis when it is actually true

- **Prob(at least 1 significant result) =**
  
  
  \[ 1 - \text{Prob(no significant result)}^n = 1 - (1-0.05)^n \]

For 10 tests, \( \text{Prob} = 1-(1-0.05)^{10} = 0.40 \)

40% probability of at least 1 false positive across 10 tests
Probability of at least 1 false positive
What do I do about it?

<table>
<thead>
<tr>
<th>Host soluble mediators of inflammation</th>
<th>Deaths</th>
<th>Survivors</th>
<th>( p )</th>
<th>Holms-Bonferroni ( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-8</td>
<td>211.5 (110.4–410.8)</td>
<td>110.0 (78.5–165.5)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MIP-1β/CCL4</td>
<td>1,076.0 (570.5–2,501.0)</td>
<td>624.5 (397.5–1,087.5)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IL-1Ra</td>
<td>449.8 (145.1–1,425.3)</td>
<td>169.5 (93.6–397.5)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IL-6</td>
<td>361.3 (194.4–656.8)</td>
<td>208.0 (119.3–359.8)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IP-10/CXCL10</td>
<td>10,818.0 (6,326.9–16,913.8)</td>
<td>6,495.0 (3,301.5–11,846.3)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MIP-1α/CCL3</td>
<td>129.0 (73.0–295.0)</td>
<td>93.0 (65.8–156.3)</td>
<td>0.001</td>
<td>0.027</td>
</tr>
</tbody>
</table>

**Higher in participants who died**

<table>
<thead>
<tr>
<th>Host soluble mediators of inflammation</th>
<th>Deaths</th>
<th>Survivors</th>
<th>( p )</th>
<th>Holms-Bonferroni ( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>IL-5</td>
<td>22.0 (15.0–30.2)</td>
<td>31.0 (22.0–43.5)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>RANTES/CCL5</td>
<td>12,688.0 (7,340.8–15,191.9)</td>
<td>15,369.5 (12,732.5–16,552.3)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IL-13</td>
<td>27.0 (18.0–39.9)</td>
<td>39.0 (29.0–59.5)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>PDGF</td>
<td>93.5 (56.4–199.1)</td>
<td>201.0 (84.0–418.5)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>FGF</td>
<td>45.3 (37.0–54.0)</td>
<td>54.0 (43.8–69.0)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IL-7</td>
<td>28.5 (22.0–37.0)</td>
<td>35.0 (28.0–53.0)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IL-12p70</td>
<td>44.5 (35.4–58.1)</td>
<td>56.0 (42.0–76.8)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>IL-4</td>
<td>38.8 (26.8–55.1)</td>
<td>48.0 (36.8–63.3)</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>*TGF-β1</td>
<td>16.5 (12.0–36.3)</td>
<td>26.4 (15.7–55.4)</td>
<td>&lt;0.001</td>
<td>0.006</td>
</tr>
<tr>
<td>IL-17</td>
<td>56.0 (41.8–78.3)</td>
<td>64.5 (48.8–90.3)</td>
<td>&lt;0.001</td>
<td>0.019</td>
</tr>
<tr>
<td>IFNγ</td>
<td>45.0 (29.8–66.0)</td>
<td>54.0 (39.0–74.5)</td>
<td>0.001</td>
<td>0.031</td>
</tr>
</tbody>
</table>

**Lower in participants who died**

**No statistically significant difference between participants who died and those who survived**


Adjust p-values to control the overall error rate at desired level rather than controlling the error rate for just one hypothesis.
Multiple Testing Adjustment

- **Control Family-wise Type I Error**
  - Bonferroni adjustment
    - Use $\alpha' = \frac{\alpha}{n}$ where $n$ = number of tests
    - Simple, applicable anywhere, most conservative
  - Sequential procedures
    - Less conservative than Bonferroni
    - Holm’s step-down procedure

- **Control False Discovery Rate (FDR)**
  - Controls proportion of false positives out of all rejected hypotheses
  - Benjaminin & Hochburg procedure
Secondary Objectives:
CRP, LPPLA2, SAA, PTX3, FIBRIN

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Raw P-value</th>
<th>Bonferroni</th>
<th>Holm’s</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRP</td>
<td>0.0557</td>
<td>0.279</td>
<td>0.194</td>
<td>0.093</td>
</tr>
<tr>
<td>LLPLA2</td>
<td>0.0855</td>
<td>0.428</td>
<td>0.194</td>
<td>0.107</td>
</tr>
<tr>
<td>SAA</td>
<td>0.0486</td>
<td>0.243</td>
<td>0.194</td>
<td>0.093</td>
</tr>
<tr>
<td>PTX3</td>
<td>0.8117</td>
<td>1.000</td>
<td>0.812</td>
<td>0.812</td>
</tr>
<tr>
<td>Fibrin</td>
<td>0.0361</td>
<td>0.180</td>
<td>0.181</td>
<td>0.093</td>
</tr>
</tbody>
</table>
Interpretation & Reporting
P-value Points to Remember

- Probability of observing data more extreme than you did *if the null hypothesis is true*
- **NOT** the probability that the null hypothesis *is* true
- Absence of evidence is **NOT** evidence of absence
  - Particularly important for small studies
  - Non-significant P values do not distinguish between group differences that are truly negligible and group differences that are non-informative because of large standard errors.
- **P-values provide no information about the magnitude of differences.**
Reporting & Interpretation

Suppose \( p = 0.006 \)

- We could state, “Mean HOMA levels were significantly higher in subjects with CAD (\( p = 0.006 \)). Log transformed mean [95% CI] values were 0.84 [0.75, 0.93] and 0.67 [0.59, 0.72] for CAD(+) and CAD(-) groups respectively.”

- Also report sample sizes: \( n = 310 \) and \( 290 \), for CAD(+) and CAD(-)
Now suppose $p = 0.32$

- Would not want to say “CAD status had no effect on HOMA levels” or “HOMA levels did not differ by CAD status.”

- We could state, “Evidence was not sufficient to reject the null hypothesis of no difference in mean HOMA levels by CAD status ($p = 0.32$). Log transformed mean [95% CI] values were 0.84 [0.75, 0.92] and 0.79 [0.65, 0.85] for CAD(+) and CAD(-) groups respectively.”

- Again, report sample sizes.
What if we see...

Scenario 1
- CAD(+): 0.84 [0.54, 1.14], n = 20
- CAD(-): 0.42 [0.12, 0.72], n = 18

Scenario 2
- CAD(+): 0.85 [0.83, 0.88], n = 2000
- CAD(-): 0.80 [0.78, 0.82], n = 1800
New Guidelines for Statistical Reporting in the *Journal*

David Harrington, Ph.D., Ralph B. D'Agostino, Sr., Ph.D., Constantine Gatsonis, Ph.D., Joseph W. Hogan, Sc.D., David J. Hunter, M.B., B.S., M.P.H., Sc.D., Sharon-Lise T. Normand, Ph.D., Jeffrey M. Drazen, M.D., and Mary Beth Hamel, M.D., M.P.H.

The *Journal’s* revised policies on P values rest on three premises: it is important to adhere to a prespecified analysis plan if one exists; the use of statistical thresholds for claiming an effect or association should be limited to analyses for which the analysis plan outlined a method for controlling type I error; and the evidence about the benefits and harms of a treatment or exposure should include both point estimates and their margins of error.
Significance tests should be accompanied by confidence intervals for estimated effect sizes, measures of association, or other parameters of interest.

P values adjusted for multiplicity should be reported when appropriate and labeled as such in the manuscript.

When appropriate, observational studies should use pre-specified accepted methods for controlling family-wise error rate or false discovery rate when multiple tests are conducted.
Help is Available

- **CTSC Biostatistics Office Hours**
  - Every Tuesday from 12 – 1:30 in Sacramento
  - Sign-up through the CTSC Biostatistics Website

- **EHS Biostatistics Office Hours**
  - Every Monday from 2-4 in Davis

- **Request Biostatistics Consultations**
  - CTSC - www.ucdmc.ucdavis.edu/ctsc/
  - MIND IDDRC -
    www.ucdmc.ucdavis.edu/mindinstitute.centers/iddrc/cores/bbrd.html
  - Cancer Center and EHS Center
Selected References

- Ioannidis 2005. Why most published research findings are false *PLoS Medicine* 2(8) e124