Statistical Decision-Making

Hong Li

May 3, 2016
Objectives

- Explain steps in hypothesis testing
- Describe null and alternative hypothesis
- Define p-value (both one-sided and two-sided)
- Define Type-I and Type-II errors
- Define power
- Understand how to interpret confidence interval (CI) in the context of hypothesis testing
Example: Avastin in Gastric Cancer (AVAGAST)

- **Paper**
  - Bevacizumab in Combination with Chemotherapy As First-Line Therapy in Advanced Gastric Cancer: A Randomized Double-Blind Placebo-Controlled Phase III Study (Journal of Clinical Oncology Volume 29, Number 30, Oct. 20, 2011, pp3968-3976)

- **Purpose**
  - Evaluate the efficacy of adding bevacizumab to capecitabine-cisplatin in the first-line treatment of advanced gastric cancer

- **Study design**
  - Double-blind, placebo-controlled Phase III clinical trial
Example: Avastin in Gastric Cancer (AVAGAST), cont.

- Treatment: in addition to cisplatin and capecitabine
  - Treatment arm: Bevacizumab
  - Control arm: placebo
- Sample size
  - 774 in total with 387 in each arm
- Primary endpoint
  - Overall survival
  - Assessed every 3 months
Example: Avastin in Gastric Cancer (AVAGAST), cont.

- Overall response rate (ORR)
  - Overall response rate was improved significantly with the addition of bevacizumab (46% vs. 37.4% in the placebo group; p-value=0.0315)
Hypothesis Testing

- Statistical hypothesis is an assumption about a population parameter
- Assumption may or may not be true
- Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses
Hypothesis Testing, cont.

- There are two types of statistical hypotheses:
  - Null hypothesis ($H_0$): usually the hypothesis that sample observations result purely from chance
  - Alternative hypothesis ($H_A$): the hypothesis that sample observations are influenced by some non-random cause
Hypothesis Testing, cont.

- State the hypotheses. This involves stating the null and alternative hypotheses. The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.

- Formulate an analysis plan. The analysis plan describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.

- Analyze sample data. Find the value of the test statistic (mean score, proportion, t statistic, z-score, etc.) described in the analysis plan.

- Reject the null or fail to reject the null, depending on the p-value’s extremeness.
What Is Population Parameter?

- Statistician differentiate between that which is being estimated and the estimate itself.
- That which is being estimated is called a population parameter and we (typically) assume it has a true but unknown value.
- In this example, there are two population parameters:
  - True (but unknown) ORR for the population of patients represented by those in the study receiving bevacizumab- \( p_{bev} \).
  - True (but unknown) ORR for the population of patients represented by those in the study receiving placebo- \( p_{plac} \).
Hypothesis Testing: Assume a Null Condition

- Statistical testing starts with a statement of the null hypothesis

\[ H_0 : p_{\text{bev}} = p_{\text{plac}} \]

This can also be written

\[ H_0 : p_{\text{bev}} - p_{\text{plac}} = 0 \]
Hypothesis Testing: State Alternative Hypothesis

- Alternative hypothesis is the state of nature we will accept if the evidence in the data suggests the null is implausible

\[ H_A : p_{bev} > p_{plac} \]

This can also be written

\[ H_A : p_{bev} - p_{plac} > 0 \]

- This is called a one-tailed or one-sided test
- Note that we do not actually specify a value for \( p_{bev} - p_{plac} \) under the alternative. We simply state the direction of the effect (\( > 0 \), \( < 0 \), or \( \neq 0 \))
Hypothesis Testing: Formulate Analysis Plan

- Intuitively, it makes sense that any test for a comparison of ORRs should be constructed from their estimates, namely $\hat{p}_{bev}$ and $\hat{p}_{plac}$.

- From the results stated in the paper

  \[
  \hat{p}_{bev} = 46\% \quad \text{and} \quad \hat{p}_{plac} = 37.4\%
  \]

  or equivalently

  \[
  \hat{p}_{bev} - \hat{p}_{plac} = 8.6\%
  \]

- Question: are these estimates meaningfully different, or is the magnitude of their difference a chance occurrence—the luck of the draw?
Hypothesis Testing: Formulate Analysis Plan (cont.)

- The test is constructed based on the data \((\hat{p}_{bev} - \hat{p}_{plac})\)
- Goal of test: determine how unlikely it is to observe a difference in the estimated ORRs at least as large as the one we’ve observed, if in fact there really is no difference in ORRs
- Allows us to make a probabilistic statement about the likelihood of the observed difference in ORRs being attributable to chance and not to the intervention
- Requires understanding the sampling distribution of \(\hat{p}_{bev} - \hat{p}_{plac}\)
Sampling Distribution

- Sampling distribution of a statistic is the distribution of that statistic when derived from a random sample of size $n$.
- Example: the test is constructed under the null hypothesis.
  - Sampling distribution as a histogram of all possible values of $\hat{p}_{bev} - \hat{p}_{plac}$ you could observe if in fact the true ORRs, $p_{bev}$ and $p_{plac}$ are equal.
Example based on ORRs

- $p_{bev} = 46\%$ and $p_{plac} = 37.4\%$
- $H_0$: $p_{bev} = p_{plac}$
- Under $H_0$, the best estimate of the true underlying ORR (for either arm) is a pooled estimate
- Results from the paper

$$\hat{p}_{bev} = \frac{143}{311} = 0.46$$
$$\hat{p}_{plac} = \frac{111}{297} = 0.374$$

- The pooled estimate

$$\hat{p}_{pooled} = \frac{(143 + 111)}{(311 + 297)} = 0.418$$
Repeated Sampling under $H_0$

Simulate 10000 data sets with sample sizes of 311 and 297, and assume $p_{bev} = p_{plac} = p_{pooled} = 0.418$. For each sample, compute $p_{bev}$ and $p_{plac}$ and construct their difference.

<table>
<thead>
<tr>
<th>Sample no.</th>
<th>$\hat{p}_{bev}$ (n = 311)</th>
<th>$\hat{p}_{plac}$ (n = 297)</th>
<th>$\hat{p}<em>{bev} - \hat{p}</em>{plac}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.447</td>
<td>0.364</td>
<td>0.083</td>
</tr>
<tr>
<td>2</td>
<td>0.463</td>
<td>0.421</td>
<td>0.042</td>
</tr>
<tr>
<td>3</td>
<td>0.414</td>
<td>0.428</td>
<td>-0.014</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>9,999</td>
<td>0.415</td>
<td>0.431</td>
<td>-0.016</td>
</tr>
<tr>
<td>10,000</td>
<td>0.412</td>
<td>0.414</td>
<td>-0.002</td>
</tr>
</tbody>
</table>
Histogram of Resulting Differences
What Is P-value?

Question: How can we determine “likely” or “unlikely” by determining the probability-assuming the null hypothesis were true- of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed.

Answer: we estimate the probability of observing a difference in ORRs as or more extreme than the one we observed in our study. This probability is called a p-value.

The p-value is defined as the probability, under the assumption of null hypothesis, of obtaining a result equal to or more extreme than what was actually observed.
P-value for Our Example
P-value for Our Example (cont.)

- Approximate the theoretical distribution of the differences (blue line overlaying the histogram) using a normal distribution
- P-value is the red-shaded area in the plot
- Example: the p-value is approximately 0.016
  - This probability represents the likelihood of obtaining a sample proportion that is at least as extreme as our sample proportion in right tail of the distribution by chance if the null hypothesis is true
  - The likelihood of observing a difference of 8.6% (or one even larger) by chance alone is less than 2%
Other Alternatives

- Another one-tailed test would be

\[ H_A : p_{bev} < p_{plac} \]

- P-value would be approximated by the relative area to the left of the observed difference.
Other Alternatives (cont.)
P-value in the Paper

- A two-tailed or two-sided test is given by the alternative hypothesis

\[ H_A : p_{bev} \neq p_{plac} \]

- P-value is approximated by the sum of the relative areas to the left of -0.086 and to the right of 0.086. The authors performed a two-sided test and report a p-value of 0.0315.
P-value in the Paper (cont.)
Definition of P-value

- The probability of observing a result at least as extreme as the one you observed if the null is true
- P value is not the probability of making a mistake by rejecting a true null hypothesis
Make Decision

- Make a judgement about the null hypothesis based on p-value
- Small p-value: observing the outcome or one more extreme due to chance alone is highly improbable
  - Reject the null hypothesis in favor of the alternative (call finding statistically significant)
  - Common practice: p-value to be less than 0.05
- P-value is not small: observing the outcome or one more extreme due to chance alone is probable
  - Fail to reject the null
  - Unable to rule out random chance as a plausible explanation for any observed difference
Two types of errors one can commit in conducting a statistical test
Actions and Errors (cont.)

- **Type I error**
  - The probability of rejecting the null given that the null is true, denoted by $\alpha$

- **Type II error**
  - The probability of failing to reject the null given that the alternative is true, denoted by $\beta$
Power

- It is the probability of correctly rejecting the null in favor of the alternative. A well-designed study will strike a balance between acceptable levels of Type I error (usually 0.05) and power (often set at a minimum of 0.8).

<table>
<thead>
<tr>
<th></th>
<th>$H_0$ true</th>
<th>$H_A$ true</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>$\alpha$</td>
<td>$1 - \beta$ = Power</td>
</tr>
<tr>
<td>Fail to reject $H_0$</td>
<td>$1 - \alpha$</td>
<td>$\beta$</td>
</tr>
</tbody>
</table>
The power of a hypothesis test is affected by three factors

- **Sample size**
  - Other things being equal, the greater the sample size, the greater the power of the test

- **Significance level** ($\alpha$)
  - The higher the significance level, the higher the power of the test
  - Increase the significance level, reduce the region of acceptance. Less likely to accept the null hypothesis when it is false; i.e., less likely to make a Type II error. Hence, the power of the test is increased
The power of a hypothesis test is affected by three factors:

- The "true" value of the parameter being tested
  - The greater the difference between the "true" value of a parameter and the value specified in the null hypothesis, the greater the power of the test.
Common Endpoints to Compare Groups

- **Difference in means**
  - Compares a continuous variable between two groups
  - Null value is 0
  - two-sample t-test or Wilcoxon rank sum test (independent groups)
  - paired t-test or Wilcoxon signed rank test (paired groups)

- **Difference in proportions**
  - Compares a categorical variable between two groups
  - Null value is 0
  - Chi-square test or Fisher’s exact test (independent groups)
  - McNemar’s test (paired groups)
Odds Ratio (OR)

- OR = odds of event \( E \) for subjects in group A divided by odds of event \( E \) for subjects in group B
- Compares a binary variable between groups A and B
- OR > 0
- Null value is 1
- In our example, \[ \text{OR} = \frac{0.46/(1-0.46)}{0.374/(1-0.374)} = 1.43 \]
- There is a 43% increase in the odds of response comparing patients randomized to bevacizumab to patients randomized to placebo
Hazard Ratio (HR)
- \( HR = \text{hazard of death in group A divided by hazard of death in group B (‘hazard’ = ‘risk’)} \)
- Compares a time-to-event variable between groups A and B
- \( HR > 0 \)
- Null value is 1
- Log rank test or Cox proportional hazards regression
"On the basis of a systematic literature review, it was assumed that median OS in the placebo group would be 10 months. The study was powered to test the hypothesis that the addition of bevacizumab would improve median OS to 12.8 months, equivalent to a hazard ratio (HR) of 0.78, 509 deaths were necessary to ensure 80% power for a two-sided log-rank test at a significance level of 0.05"
Confidence Interval (CI)

- CI is a type of interval estimate of a population parameter
- Estimated range of values that provides a measure of uncertainty associated with an estimated parameter
- The wider the interval the greater the uncertainty in the estimate
- 95% CI means
  - If you were able to replicate the exact same study an infinite number of times, 95% of the resulting CIs would contain the true parameter of interest
"PFS was prolonged significantly in the bevacizumab group compared with the placebo group (HR, 0.8; 95% CI, 0.67 to 0.93; p-value=0.037)"

- The best estimate of the true HR is 0.80, but our data are consistent with a HR ranging from 0.67 to 0.93
- CI does not contain 1 which means we have even greater faith that the benefit observed is real and not just a chance occurrence
S be the statistic used to perform a test that compares two groups. For example, \( S \) could be a difference in sample means, a difference in sample proportions, an estimated OR, or an estimated HR. If the test is significant at \( \alpha \)-level 0.05, then the 95% CI constructed with \( S \) will not contain the relevant null value.
Subgroup Analysis of Overall Survival in the Intent-to-Treat Population

<table>
<thead>
<tr>
<th>Category</th>
<th>Subgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>Region</td>
<td>Asia</td>
</tr>
<tr>
<td></td>
<td>Europe</td>
</tr>
<tr>
<td></td>
<td>Pan-America</td>
</tr>
<tr>
<td>ECOG performance status</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>≥ 1</td>
</tr>
<tr>
<td>Site of primary disease</td>
<td>Stomach</td>
</tr>
<tr>
<td></td>
<td>GEJ</td>
</tr>
<tr>
<td>Histologic type</td>
<td>Intestinal</td>
</tr>
<tr>
<td></td>
<td>Diffuse</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
</tr>
<tr>
<td>Disease status</td>
<td>Locally advanced*</td>
</tr>
<tr>
<td></td>
<td>Metastatic</td>
</tr>
<tr>
<td>Disease measurability</td>
<td>Measurable</td>
</tr>
<tr>
<td></td>
<td>Nonmeasurable</td>
</tr>
<tr>
<td>Prior gastrectomy</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>No. of metastatic sites at baseline</td>
<td>≥ 1</td>
</tr>
<tr>
<td></td>
<td>≥ 2</td>
</tr>
</tbody>
</table>

Ohtsu A et al. JCO 2011;29:3968-3976