Overview of Hypothesis Testing

Paul Wakim, PhD

Chief, Biostatistics and Clinical Epidemiology Service

Clinical Center, National Institutes of Health

U.S. Department of Health and Human Services

3 November 2015

Outline

Fundamentals of Hypothesis Testing

Superiority vs. Non-Inferiority vs. Equivalence

Multiple Comparisons (Multiplicity Adjustment)

Bottom-Line Key Points

Outline

Fundamentals of Hypothesis Testing

Superiority vs. Non-Inferiority vs. Equivalence

Multiple Comparisons (Multiplicity Adjustment)

Bottom-Line Key Points

Question

Without ____?____ , there is no need for Statistics

Answer

Without variability, there is no need for Statistics

Variability

Variability

Variability

Variability

Statistical Inference

1) Draw a sample from the population of interest
2) Analyze the sample data
3) Make conclusion about the population based on results from the sample

Typical Setting of Statistical Inference

(non-Bayesian)

Null Hypothesis (H0)

Experimental = Control

or

Experimental – Control = 0

Alternative Hypothesis (H1 or HA)

Experimental ≠ Control

or

Experimental – Control ≠ 0

Question: Is there enough evidence to reject H0

– the hypothesis of no difference?

We expect (hope) to reject H0 in favor of HA

Correct and Incorrect Conclusion

Correct and Incorrect Conclusion

Correct and Incorrect Conclusion

Correct and Incorrect Conclusion

Correct and Incorrect Conclusion

Correct and Incorrect Conclusion

Alpha (α)

Probability of Making a Type I Error

Beta (β)

Probability of Making a Type II Error

Power to Detect an Effect

Non-technical definition (superiority trial):

Chance of concluding that the experimental treatment is more effective when in fact it is

Technical definition:

Probability of rejecting H0 when H0 is false

(i.e. when HA is true)

Different perspectives:

Regulatory agency, pharmaceutical company

Bottom line:

Most commonly used value for power:

Early-phase: 0.60 to 0.80 – Late-phase: 0.80 to 0.95

If you don’t change the sample size…

Experiment

Toss a coin

Null hypothesis (H0):

It’s the regular coin, with Head on one side and Tail on the other side

Alternative hypothesis (HA):

It’s the other coin, with Head on both sides

We assume “equipoise”, i.e. the coin is as likely to be a regular coin as to have 2 Heads

Experiment

Experiment

Experiment

Experiment

Experiment

Experiment

Experiment

Experiment

Definition of p-value

The p-value is the probability of obtaining a result as extreme or more extreme than the one obtained, if H0 were actually true

If p-value ≤ α (alpha), reject H0

If p-value > α (alpha), do not reject H0

Commonly used alpha levels: 0.05 or 0.01

Bayesian Approach

Bayesian Approach

If we start with the belief (prior) that H0 and HA are equally likely

What is the connection between

alpha (α) and p-value?

If the p-value is less than α (typically 0.05), the null hypothesis (e.g. of no difference) is rejected, and the result is declared statistically significant at the 5% alpha level

If the p-value is greater than α, the result is not statistically significant at the 5% alpha level

What is the connection between

p-value and sample size?

Randomized Controlled Trial

Objective: To compare 2 treatments

Data are collected. Analysis is done.

Result: p-value = 0.something

Distribution of the difference between

the 2 treatment groups, IF in fact there is no difference

Distribution of the difference between

the 2 treatment groups, IF in fact there is no difference

Distribution of the difference between

the 2 treatment groups, IF in fact there is no difference

Distribution of the difference between

the 2 treatment groups, IF in fact there is no difference

What's the point?

There are two ways to get statistically significant results… guaranteed!

Analyze a very large sample

What is the connection between

confidence intervals and hypothesis testing?

95% Confidence Intervals

95% Confidence Intervals

(natural perspective)

There is a 95% chance that the true *unknow*n value is inside the confidence interval

We are 95% confident that the true *unknown* value is somewhere within the confidence interval

95% Confidence Intervals

(natural perspective)

95% Confidence Intervals

(frequentist's pure perspective)

There is a 95% chance that the confidence interval covers the true unknown value

95% Confidence Intervals

(frequentist’s pure perspective)

95% Confidence Intervals

What is the connection between

confidence intervals and hypothesis testing?

If the 95% confidence interval does not include the value of the null hypothesis (e.g. of zero difference), the result is statistically significant at the 5% alpha level

If it does, the result is not statistically significant at the 5% alpha level

Outline

Fundamentals of Hypothesis Testing

Superiority vs. Non-Inferiority vs. Equivalence

Multiple Comparisons (Multiplicity Adjustment)

Bottom-Line Key Points

Superiority

Clinical hypothesis:

Experimental treatment is more effective than

the control treatment

Statistical hypotheses:

Null hypothesis H0: Experimental = Control

Alternative hypothesis HA: Experimental ≠ Control

We expect (hope) to reject H0 in favor of HA

Superiority

Non-Inferiority

Clinical hypothesis:

Experimental treatment is not less effective than the control treatment

Statistical hypotheses:

Null hypothesis H0: Experimental $<$ Control $- \delta$

Alternative hypothesis HA: Experimental $\geq$ Control $- \delta$

We expect (hope) to reject H0 in favor of HA

Non-Inferiority

Equivalence

Clinical hypothesis:

Experimental treatment is as effective as the control treatment

Statistical hypotheses:

Null hypothesis H0:

Experimental $<$ Control $- \delta$ or Experimental $>$ Control $+ \delta$

Alternative hypothesis HA:

Control $- \delta \leq$ Experimental $\leq$ Control $+ \delta$

We expect (hope) to reject H0 in favor of HA

Equivalence

Outline

Fundamentals of Hypothesis Testing

Superiority vs. Non-Inferiority vs. Equivalence

Multiple Comparisons (Multiplicity Adjustment)

Bottom-Line Key Points

When to do multiplicity adjustment?

Formally, whenever there are more than one primary endpoint (or primary hypothesis), more than two treatment conditions, more than one dose vs. placebo, or more than one time point

Informally, whenever there are more than one secondary analysis, including subgroup analyses

The second way….

There are two ways to get statistically significant results… guaranteed!

Analyze a very large sample

Keep trying different statistical tests on different assessments (outcomes) or on different subgroups of the data

Outline

Fundamentals of Hypothesis Testing

Superiority vs. Non-Inferiority vs. Equivalence

Multiple Comparisons (Multiplicity Adjustment)

Bottom-Line Key Points

Bottom-Line Key Points

Statistical inference uses results from a sample from the population of interest to draw conclusions about the population

The null hypothesis is set up with the hope that it will be rejected

Alpha (α) is the chance of making a Type I error, i.e. of concluding that there is a difference when in fact there isn't

Beta (β) is the chance of making a Type II error, i.e. of concluding that there isn't a difference when in fact there is

Power = 1 – β = the chance of concluding that there is a difference when in fact there is

Bottom-Line Key Points (cont'd)

The investigator controls the chance of making a Type I error (alpha) and the chance of making a Type II error (beta) via the sample size

P-value is the probability of obtaining a result as extreme or more extreme than the one obtained, if there were no difference

Statistical significance does not mean clinical importance

Confidence intervals are very useful to better understand results

Multiplicity adjustment is needed with more than one primary hypothesis

Bayesian approach is gaining popularity as being more intuitive, and is worth considering

The End

Thank you for your attention

I hope this was worth your time

References

Piaggio G et al., Reporting of Noninferiority and Equivalence Randomized Trials: An Extension of the CONSORT Statement, JAMA, 2006, 295:1152-1160

Underwood D, The Profitable Pause, International Clinical Trials, August 2011, Issue 21, 56-60