Statistical Thinking in Biomedical Research

Department of Biostatistics
School of Medicine
Vanderbilt University

`biostat.mc.vanderbilt.edu/biostat/twiki/bin/view/Main/ClinStat`

OCTOBER 14, 2004

Statistical Thinking in Biomedical Research

Thursday 5:30 - 7:00, Light Hall Room 208

---

1. Introduction to Statistical Concepts

   October 07

2. Tools for Formal Statistical Inference

   October 14

3. Statistical Modeling

   October 21

4. Experimental Design and Analyzing Serial Measurements

   October 28

5. Graphical Methods

   November 4

6. Miscellaneous issues and Demonstration of Statistical Software and Data Analysis

   November 11

7. Measuring Effects and Quantifying Evidence

   November 18

Section2
Tools for Formal Statistical Inference

Tatsuki Koyama

Contributor - **Robert D. Abbott**

Division of Biostatistics and Epidemiology

Department of Health Evaluation Sciences

University of Virginia School of Medicine

- The basic logic behind hypothesis testing
- The notion of a null and alternative hypothesis
- Sample size and statistical power
- The notion of a p-value
- The use of a confidence interval

Hypothesis Testing - Example 1-

24 hour total energy expenditure (mj/day) in groups of lean and obese women.[2]

Question : Is obese women's energy expenditure higher than lean women's energy expenditure?

Lean ($n = 13$)

6.13, 7.05, 7.48, 7.48, 7.53, 7.58, 8.08, 8.09, 8.11, 8.40, 10.15, 10.88

Obese ($n = 9$)

8.79, 9.19, 9.21, 9.68, 9.69, 9.97, 11.51, 11.85, 12.79

Answer : The energy expenditure varies from woman to woman. Some obese women have high energy expenditure and some don't. Some lean women do and some don't.

The lean women's average energy expenditure is 2.2 less than the obese women's average energy expenditure.

But we are not particularly interested in these $\mathbf{22\ (= 13 + 9)}$ women.

Population and Sample

- We would like to know the truth about the **population** of interest.
- But to test everybody in the population is usually impossible!
- So we take a small manageable **sample** from the population of interest.
- And we try to make inference about the population from the sample.

Can statistics lead you to the correct conclusion about the population of interest?
Even when we do everything correctly, the conclusion may be incorrect.

- The sample you have may not resemble the population of interest.
- And statistics cannot always lead to the correct conclusion ...
- ... but statistics can quantify how believable the conclusion is.

Back to the Example 1

Question

Is obese women's energy expenditure higher than lean women's energy expenditure?

Answer

If obese women's energy expenditure is equal to lean women's energy expenditure on average, the probability of observing a difference of $\mathbf{2.2}$ or more is $\mathbf{0.0009}$.

"Statistics means never having to say you're certain."

Summary

- To infer about the population of interest, we take a manageable sample.
- The sample you have may not represent the population well because of variability among individuals.
- Statistics cannot always lead you to the correct conclusion about the population, but it can quantify how believable your conclusion is.

Example 2

The Problem

The probability of survival from a rare form of cancer after six months from diagnosis is extremely low. Suppose that we all agree that no more than 3% of patients with this kind of cancer survive.

Suppose that we are interested in studying a new treatment for this type of cancer.

The Experiment

We would like to know if the new treatment is “promising” or “worthless.” Unfortunately, because this type of cancer is rare, it will be difficult to enroll a large sample, but we will be able to recruit a sample of 5 within a year.

Decision Rule

Suppose, for now, that we declare that the treatment has promise if one or more of the 5 patients survive after six months with the new treatment.

Null and Alternative Hypotheses

Definitions

Let θ be the probability that a patient survives.

1. The null hypothesis

(the hypothesis of ignorance)

$$H_0 : \text{The treatment is worthless.}$$

or

$$H_0 : \theta = 3\%$$

2. The alternative hypothesis

(the hypothesis we wish to prove)

$$H_a : \text{The treatment has promise.}$$

or

$$H_a : \theta > 3\%$$

One Sided and Two Sided Alternative

In this example, the alternative hypothesis is *one sided.* A *two sided* alternative hypothesis would be

$$H_0 : \text{The treatment has an effect.}$$

or

$$H_a : \theta \neq 3\%$$

- A direction is not specified in a two-sided alternative hypothesis.
- In this example, using a two-sided alternative hypothesis does not make sense.

Aside :

Do we ever use a two-sided alternative hypothesis against a placebo?

Type I and Type II Errors -page 1-

Possible experimental results:

		Actual truth about the treatment	
		Worthless	Promising
Our decision about the treatment	Worthless	a	$\boldsymbol{b}$
	Promising	$\boldsymbol{c}$	d

Note : a and d are correct decisions.

- $\boldsymbol{c} =$ decide the treatment has promise when it's actually worthless

  type I error (false positive decision)

- $\boldsymbol{b} =$ decide the treatment is worthless when it actually has promise

  type II error (false negative decision)

Type I and Type II Errors -page 2-

Goal: We wish to conduct our study so that the probability of making these wrong decisions is small. Type I and II errors are (usually) not of equal importance.

- Without a proper control of probability of type I error, the conclusion is not valid.
- The experimenter can set the probability of type II error.
  - A small type II error probability costs more.
  - A large type II error probability may be unethical.

$\alpha =$ P[Type I Error]

$\beta =$ P[Type II Error]

To make $\alpha \downarrow$, you have to increase the sample size.

To make $\beta \downarrow$, you have to increase the sample size.

Important message:

Funding agencies are more interested in investing money in experiments that have low probabilities associated with these errors.

Ethical message:

Patient (and your) resources should not be wasted on experiments which have a high probability of leading to these errors. For example, would you want to be a participant in an experiment that utilized aggressive or invasive procedures that had little chance of showing that a treatment had promise when in fact the treatment was very effective? Should informed consent include such information?

Back to the Example 2

We will conclude that the new treatment is “promising” if one or more patient survive after six months.

When the new treatment is “worthless” ($p = 3\%$), the probability of concluding “promising” is about 0.14.

		Actual truth about the treatment	
		Worthless	Promising
Our decision about the treatment	Worthless	a	b
	Promising	0.14	d

Question: How can we make α smaller?

Answer: We can make it more difficult to declare that the treatment has promise when it is worthless by requiring that at least **2** patients must survive.

Now, P[Type I eror] $= 0.0085$.

In order to keep our type I error at an acceptable level, let’s propose that we proceed with our experiment by enrolling 5 patients. If 2 or more of the 5 patients survive, we will declare that the treatment has promise.

How About Type II Error?

Recall:

$$\begin{aligned}\text{P[Type II error]} &= \text{P[False negative decision]} \\ &= \text{P[We decide treatment is worthless} \\ &\qquad \text{when it is actually promissing]}\end{aligned}$$

But what do we mean by "promising?"
Is $\mathbf{5\%}$ promising? $\mathbf{10\%?}$ $\mathbf{20\%?}$ or $\mathbf{30\%?}$

Let's agree that $\mathbf{20\%}$ is promising. Then ...

		Actual truth about the treatment	
		Worthless	Promising
Our decision about the treatment	Worthless	a	$\mathbf{0.74}$
	Promising	$\mathbf{0.0085}$	d

Type II error probability of $\mathbf{0.74??}$
If we agree that $\mathbf{30\%}$ is promising, then $\beta = \mathbf{0.53}$.
But can we change what we mean by "promising?"

$n \uparrow$

We noted before that we can make $\alpha \downarrow$ and $\beta \downarrow$ by increasing the sample size, n.

For this example, now suppose that we have $n = 25$.
Now we declare that the treatment is promising if at least 3 out of these 25 survive after six months.
P[At least 3 survive
when the treatment is actually worthless] $= 0.038$

P[Fewer than 3 survive
when the treatment is promissing (20%)] $= 0.098$

		Actual truth about the treatment	
		Worthless	Promising
Our decision about the treatment	Worthless	a	0.098
	Promising	0.038	d

Summary:

If we declare that the treatment is promising when at least 3 of 25 patients survive, then the chance of committing a type I error will be less than 0.05. On the other hand, if the treatment is actually capable of saving the lives of 20% of patients with this form of cancer, then we will have more than 90% chance of concluding that the treatment has promise.

We say that the **power** of this study is about 0.90.

Power is probability of concluding that the treatment is effective when it is effective.

It is $1 - \beta$.

With $n = 5$, the power was $1 - .74 = .26$.

How can we increase the power?

$n \uparrow$

$\alpha \uparrow$

It also depends on the true effectiveness of the treatment. If the treatment can save 40% of the patients, then the power is 0.9996.

Power and Sample Size

			Power at		
N	R	α	$\theta = 10\%$	$\theta = 20\%$	$\theta = 30\%$
5	2	0.0085	0.082	0.26	0.47
10	2	0.035	0.26	0.62	0.85
20	3	0.021	0.32	0.79	0.96
25	3	0.038	0.46	0.90	0.99
40	4	0.032	0.58	0.97	≈ 1

If at least R patients out of N survive, we conclude that the treatment is promising.

Power is a function of N, α and what we mean by "promising."

So far we have only talked about the **design** of experiment.

Once the data are gathered, α, β, N no longer matter.

Inference

Suppose that we decide to proceed with our study based on $n = 25$. Suppose further that 4 patients survive.

Definition

p-value = probability that we observe what we have observed (or something more extreme) if the null hypothesis is true.

p-value tells us how unlikely our actual observation would be if the treatment is actually worthless.

We observed that 4 patients survived. So the p-value is the probability of observing 4 or more patients surviving if the treatment is worthless.

p-value = 0.0062.

The Conclusion

p-value = 0.0062 is very small. So our assumption which we used in computing the p-value, “the treatment is worthless,” is probably wrong.

Question:

Is the treatment promising?

Answer:

If the treatment is not promising then the probability that 4 (or more) out of 25 survive is very low. [0.0062 to be exact]

Question:

How promising is it?

Because we reject the null hypothesis that $\theta = 3\%$, we conclude that it’s better than $\theta = 3\%$. And because 4 out of 25 patients survived, our estimate is $\theta = 4/25 = 16\%$.

But why is it just a “estimate?”

Confidence Interval

Definition:

A confidence interval is a range of survival probabilities (in our example -it can be used for many things we might like to estimate) that could conceivably have produced the observed results with reasonalbe probability.

For a $\mathbf{90\%}$ confidence interval, we say that we are " $\mathbf{90\%}$ confident" that the unknown survival probability falls within the interval. For our data, a $\mathbf{90\%}$ confidence interval of the true survival probability is $\mathbf{(.06,\ .28)}$.

Hypothesis testing and confidence interval are closely related.

Usually, "the null hypothesis is rejected"

$\leftrightarrow$ "a confidence interval does not include the value under null hypothesis"

two-sided $\boldsymbol{\alpha = 0.05 \leftrightarrow 95\%}$ confidence interval

two-sided $\boldsymbol{\alpha = 0.01 \leftrightarrow 99\%}$ confidence interval

Interpreting a Confidence Interval

What is " 95% confident?"

- In a frequentist paradigm, as opposed to Bayesian, the unknown parameters (e.g., mean) are fixed constants.
- So the true unknown parameter is either **in** or **out** of a confidence interval. (It can't be " 95% in.")

Is $\mathbf{5}$ in $\mathbf{(3.3,\ 5.1)}$?

Don't say, "the probability that the true mean is in $\mathbf{(3.3,\ 5.1)}$ is 95%."

A confidence (95%) does not refer to the one confidence interval that you just calculated, but it refers to the **method** that you used to calculate the confidence interval.

If we repaeat the procedure of " taking a sample and calculating a confidence interval" many, many times, about 95% of the resulting confidence intervals contain the true, unknown parameter.

What Information is Missing from a P-value?

Suppose that the null hypothesis of interest is

$$H_0 \ : \ \text{True Mean} = 10,$$

Suppose that the standard deviation is known to be $\mathbf{10}$.

And you want to show that the true mean is higher than $\mathbf{10}$.

Situation 1: Suppose that you get a p-value $= \mathbf{0.011}$.
Situation 2: Suppose that you get a p-value $= \mathbf{0.046}$.

Which is more preferable?

What Information is Missing from a P-value?

Situation 1: (p-value is 0.011.)

Situation 2: (p-value is 0.046.)

Situation 1:

Sample size is 10. The sample average is 18.0

Situation 2:

Sample size is 100. The sample average is 12.0

Situation 1:

A 95% confidence interval is $(11.8,\ 24.2)$

Situation 2:

A 95% confidence interval is $(10.1,\ 13.9)$

Miscellaneous Topic 1 : Multiplicity

Recall that

α = P[Type I Error]

= P[Incorrectly conclude that the treatment has promise.]

= 0.05 (usually)

Example :

To characterize the role of TGFβ signaling on prostate androgen responsiveness. ... ... we observe “apoptotic index” and “proliferation index” on each tissue sample from wild type and knockout groups.

H_0 : There is no difference between wild type and knockout groups.

H_a : There is some difference.

$\alpha = 0.05$ means that you would conclude that there is some difference when actually there is no difference with a probability, 0.05.

Multiplicity

For apoptotic index, α is 0.05.

For proliferation index, α is 0.05.

There are 2 chances to make an incorrect conclusion.

Consequently, the probability of incorrectly rejecting at least one H_0 is 0.0975.

In order to control "overall" α at 0.05, each test needs to be conducted at 0.025.

Number of tests	P[Type I Error]	adjusted α
1	0.05	0.05
2	0.0975	0.025
3	0.143	0.017
4	0.185	0.013
5	0.226	0.010
6	0.265	0.009
8	0.337	0.006
10	0.400	0.005

Miscellaneous Topic 2 : Test of Equivalence

"Absence of evidence is not evidence of absence"[1]

H_0: The two treatments are the same.

H_a: The two treatments are different.

If we cannot reject the above H_0, it either means that the two treatments are not different or that the sample size was too small. **Don't say, "we accept H_0."**

Hypotheses for "Test of equivalence":

H_0: The difference is greater than Δ or smaller than $-\Delta$.

H_a: the difference is between $-\Delta$ and Δ.

Then by rejecting H_0, we can conclude that the two treatments are "equivalent".

Choosing a good Δ is not an easy task.

Suppose we're interested in showing equivalence of two treatments' response rate.

"$\Delta = 80$ percentage point" is meaningless.

"$\Delta = 5$ percentage point" requires a large sample.

References

[1] G. Altman, D and M. Bland, J. Absence of evidence is not evidence of absence. *British Medical Journal*, pages 311–485, 1995.

[2] A. Prentice, A. Black, W. Coward, H. Davies, G. Goldberg, P. Murgatroyd, J. Ashford, M. Sawyer, and R. Whitehead. High-levels of energy-expenditure in obese women. *British Medical Journal*, 292:983–987, April 1986.