Rob Abbott
Division of Biostatistics & Epidemiology
Dept. of Health Evaluation Sciences
University of Virginia School of Medicine

The Basic Logic in Statistical Inference

Objectives:

- To illustrate the basic logic behind hypothesis testing using a simple example.
- To illustrate the notion of a null and alternative hypothesis.
- To introduce the notion of a p-value and statistical power.
- To demonstrate how sample size can influence statistical power.
- To illustrate the use of a confidence interval.

The Problem:

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

the probability a patient survives is 0.03.

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

The Experiment:

We decide to recruit a sample of these types of cancer patients and apply the new treatment. Unfortunately, since this type of cancer is rare, it will be hard to enroll a large sample, but we are confident that we will be able to recruit a sample of five (based on past experience) within a year.

Decision Rule: Suppose (for now), that we decide to declare that the treatment has promise if one or more of the 5 patients survive after six months with our new treatment (20% survival).

Definitions: First, let

p = probability that a patient survives.

1. The null hypothesis (the hypothesis of ignorance) –

H_o: The treatment is worthless, or

H_o: $p=0.03$

2. The alternative hypothesis (the hypothesis we wish to prove) –

H_a: The treatment has promise, or

H_a: $p>0.03$

Note: The above alternative hypothesis is actually a one-sided alternative since it would be difficult to argue that the treatment promotes disease when there is already little promise of survival. Usually, we deal with two sided alternatives (e.g., where a treatment could either lower or raise blood pressure). Here we define the alternative as

H_a: The treatment has an effect

Without specifying a direction.

Possible experimental results:

		Actual truth about the treatment (beyond the border)	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***b***
	Promise	***c***	d

Note: a and d are correct decisions.

c = decide treatment has promise when it's actually worthless =

type I error (or false positive decision).

b = decide treatment is worthless when it actually has promise =

type II error (or false negative decision).

Goal: We wish to conduct (or design) our study so that the probability of making these wrong decisions is small.

Note: For this kind of experiment, when making a decision about a treatment, there are only two errors that can be made.

Important message: Funding agencies are more interested in investing money in experiments that are associated with 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?

Conclusion: It is important to design studies that have low type I (false positive) and type II (false negative) error rates.

First, some notation:

Let “D” denote the event that “a single patient dies”, then define

P(D) = P(a single patient dies) = probability a single patient dies.

Let “S” denote the event that “a single patient survives”, then define

$$p = P(S) = P(\text{a single patient survives})$$

Additive law of probability:

If two events (A and B) are mutually exclusive (they can't happen at the same time), then

$$P(A \text{ or } B) = P(A) + P(B)$$

For example: D = death and S = survival are mutually exclusive.

As a result, $P(D \text{ or } S) = P(D) + P(S)$

But, we also know that the $P(D \text{ or } S) = 1$.

Thus, $P(D) = 1 - P(S) = 1 - p$.

So, $p = P(S)$ and $1-p = P(D)$.

Multiplicative law of probability:

If individual outcomes are independent (i.e., they don't influence each other), then P(of a set of outcomes) equals the product of the probabilities of the individual outcomes.

Example #1: P(all 5 patients die) = P(DDDDD)

$= P(D)P(D)P(D)P(D)P(D) = (1-p)x(1-p)x(1-p)x(1-p)x(1-p) = (1-p)^5$

Example #2: P(DSDSD) = $(1-p)x(p)x(1-p)x(p)x(1-p) = (1-p)^3p^2$

Example #3: P(DDDDD or SDSDS) = P(DDDDD) + P(DSDSD)

$= (1-p)^5 + (1-p)^3p^2$

So, for any combination of deaths and survivors among our 5 patients, we can always express the probability of a specific set or sets of combinations as a product or sum of p's and (1-p)'s.

Back to the cancer example.

		Actual truth about the treatment: Worthless	Promise
Our decision about the treatment	Worthless	a	***b***
	Promise	***c***	d

c = decide treatment has promise when it's actually worthless =

type I error (or false positive decision).

P(type I error) = P(false positive decision) =

P(we decide treatment has promise when it's actually worthless)

= P(one or more patients survive when treatment is worthless)

= P(SDDDD or DSDDD or…SSSSS)

= P(SDDDD) + P(DSDDD) +… P(SSSSS)

$= 1 - (1-p)^5$ (recall, p = 0.03 when the treatment is worthless)

$1 - (1-0.03)^5 = 0.14$

i.e., We have a 14% chance of committing a type I error.

		Actual truth about the treatment: Worthless	Promise
Our decision about the treatment	Worthless	a	***b***
	Promise	***0.14***	d

Question: How can we make it harder to declare that the treatment has promise when the treatment is worthless?

Answer: Let's make it harder to declare the treatment promising when it's actually worthless by requiring that at least 2 patients must survive.

Then, P(type I error) = P(false positive decision) =

= P(2 or more patients survive when treatment is worthless)

= P(SSDDD or SDSDD or… SSSSS)

= P(SSDDD) + P(SDSDD) +… P(SSSSS)

$= 1 - [5 \times (1-p)^4 p + (1-p)^5] = 0.01$ (when p = 0.03)

		Actual truth about the treatment	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***b***
	Promise	***0.01***	d

So, if we decide to declare that the treatment has promise if 2 or more patients survive, then there will be about a 1% chance of declaring that it has promise when it is actually worthless.

Conclusion: In order to keep our type I error at an acceptable level (14% vs 1%), let's propose that we proceed with our experiment by enrolling 5 patients. If 2 or more of the 5 patients survive, then we will declare that the treatment has promise. Here, type I error looks great!

But what about the other error; i.e.,

		Actual truth about the treatment	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***b***
	Promise	***0.01***	d

b = decide treatment is worthless when it actually has promise =

type II error (or false negative decision).

P(type II error) = P(false negative decision) =

P(we decide treatment is worthless when it actually has promise) =

P(one or no patients survive)

P(SDDDD or DSDDD or DDSDD or DDDSD or DDDDS or DDDDD)

$= 5 \times (1-p)^4 p + (1-p)^5$, when $p>0.03$.

Note: $p>0.03$ corresponds to the treatment having promise.

Note: We can use this last formula for calculating P(false negative decision) for any value of p where

p = P(a single patient survives)

For example, if we proceed with our experiment as planned; i.e., enroll n=5 subjects and decide to declare the treatment effective if at least 2 patients survive,

what will be the probability of our saying the treatment is worthless when, <u>in fact</u>, the treatment is actually capable of saving the lives of 20% of patients (not 3%) with this kind of cancer; i.e., when p=0.2?

That is, what is

P(we decide the treatment is worthless when p=0.2)?

Solution: P(one or no patients survive when p=0.2) =

$$5 \times (1-p)^4 p + (1-p)^5 = 0.74 \quad \textbf{\textit{BAD!}}$$

		Actual truth about the treatment	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***0.74***
	Promise	***0.01***	d

Similarly, when p = 0.3,

P(one or no patients survive when p=0.3) = $5 \times (1-p)^4 p + (1-p)^5 = 0.53$

		Actual truth about the treatment	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***0.53***
	Promise	***0.01***	d

STILL BAD!

Question: Is there a way to lower the P(false negative decision) from 0.74 to say less than 0.1?

What is in control in our experiment (or we wish is in control)?

SAMPLE SIZE (n)

That is, choose n (just by replacing “5” with “n” in the preceding arguments) so that

P(type II error) = P(false negative decision)

= P(we decide the treatment is worthless when $p=0.2$) < 0.1

The usual procedure for selecting n:

Step 1: We select a value of n and a decision rule for declaring the treatment promising so that the probability of declaring it promising when it is not is low (usually < 0.05); i.e., we select n and a decision rule so that

$$P(\text{type I error}) = P(\text{false positive decision}) < 0.05$$

		Actual truth about the treatment	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***b***
	Promise	***<0.05***	d

Example #1: If we choose n=5, then in order to keep the type I error rate below 5%, our decision rule requires that we declare that the treatment is promising only when we observe at least 2 survivors among the 5 patients.

Example #2: If we select n=25, then in order to keep the type I error rate below 5%, our decision rule requires that we declare that the treatment is promising only when we observe at least 3 survivors among the 25 patients.

Step 2: Next, for the sample size and decision rule selected in Step 1, we derive the

P(type II error) = P(false negative decision for various values of p>0.03),

and select the n and the corresponding decision rule that yields an acceptable type II error rate < 0.1.

		Actual truth about the treatment	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***<0.1***
	Promise	***<0.05***	d

Example #1: For n=5 and the decision rule selected in Step 1, the chance of a false negative decision is 74% when p = P(S) = 0.2.

Example #2: For n=25 and the decision rule selected in Step 1, the chance of a false negative decision is 9.9% when p = P(S) = 0.2.

Note: For this experiment, n=25 is the smallest sample size that will yield a decision rule that is associated with a type I error below 5% and a type II error rate (when p = P(S) = 0.2) of less than 10%.

That is, for n=25, if we decide to declare that the treatment has promise only when at least 3 patients survive, then our chances of making the wrong decision can be summarized in the following table.

		Actual truth about the treatment	
		Worthless	Promise
Our decision about the treatment	Worthless	a	***<0.1***
	Promise	***<0.05***	d

Definition of Statistical Power:

Power is simply 1 – P(type II error)

= P(we decide the treatment has promise when it's actually effective)

Example #1: For n=5 and the selected decision rule, if p = P(S) = 0.2, power equals

$$1 - P(\text{false negative decision}) = 1 - 0.74 = 0.26.$$

Example #2: For n=25 and the selected decision rule, if p = P(S) = 0.2, power equals

$$1 - P(\text{false negative decision}) = 1 - 0.099 > 0.90.$$

Note: Having P(false negative decision) small means we have a powerful experiment because P(true positive decision) will be high.

Summary:

If we declare that the treatment has promise only when at least 3 of 25 patients survive, than the chance of committing a type I error will be less than 5%. If, on the other hand, the treatment is actually capable of saving the lives of 20% of patients with this form of cancer, then we will have more than a 90% chance of concluding that the treatment has promise.

That is, we'll have a good chance of saying that the treatment has promise if it is truly effective.

Now, suppose we decide to proceed with our study based on n=5. Suppose further, that 2 patients survive.

Definition: p-value =

P(result at least as extreme as that observed if the treatment is worthless)

Here, p-value =

P(2 or more patients survive if the treatment is worthless) = 0.01

Interpretation:

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

If the p-value is small, our observation would be too inconsistent with the hypothesis that the treatment is worthless; i.e., the data suggest that the treatment is not worthless.

On the other hand, if the p-value is large, the observed data is very likely to have occurred when the treatment is worthless; i.e., the data don't support us in saying the treatment is effective.

For example, in our study, if only one patient survived, then

p-value =

P(one or more patients survive if the treatment is worthless) = 0.14.

What about questions of how effective is the treatment?

Given that we actually observed 2 patients surviving among 5, one might be pleased by the 40% survival rate. But, how good is this as an estimate of the actual rate beyond the border?

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

For a 95% confidence interval, we say that we are 95% confident that the unknown survival probability falls within the interval (a,b).

For our example, and for various sample sizes, where the observed survival rate is 40%.

n	95% confidence interval
5	(5.2, 85.4)
10	(12.1,73.8)
15	(16.3, 67.8)
20	(19.1, 64.0)
25	(21.1, 61.4)
30	(22.6, 59.4)
35	(23.8, 57.9)
40	(24.8, 56.7)