Hypothesis Testing, Estimation, Sample Size and Power

Tatsuki Koyama, Ph.D.

Biostatistics Subcore
Vanderbilt Digestive Disease Research Center

January 21, 2020

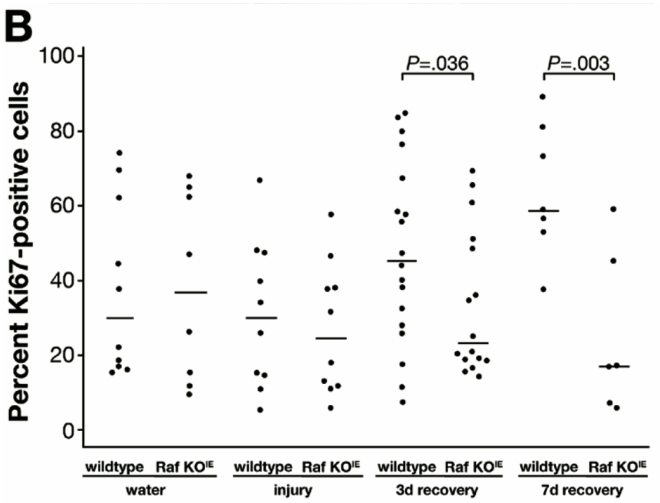
Statistics's role

Population A whole class of individuals on which we want to make a general statement.

Parameters Some numerical facts about the population

Sample A part of population that can be examined.

Statistics Numbers which can be computed from a sample.



The parameter of interest is the TRUE average proportion of Ki67-positive cells in Raf KO group.

The observed statistic is the sample average.

Edelblum et al (2008) *Gastroenterology*

Samples vary from one another. Statistics allows us to make inference about the unknown population ***parameters*** using the data at hand (a sample).

Hypothesis testing

Hypothesis is usually a statement about the population parameters (such as the population mean, difference of the population means, and the population proportions).

Examples:

- $\mu = 100$ (The population mean is 100.)
- $\pi = 0.2$ (The population proportion is 20%.)
- $\mu_1 - \mu_0 = 0$ (The mean of population 1 is the same as the mean of population 0.)

It can be about population distributions, but that is rare/unnecessary. (e.g., The true distribution is Normal.)

What is wrong with

> "There is a statistically significant difference in the means."

as a hypothesis?

Hypotheses

Null hypothesis is the statement you hope to reject/dismiss. (H_0)

- "Probability of success is less than 20%." when you want to say that the probability of success is ***greater than*** 20%.
- "There is no difference in the group means." when you want to say that the true means are ***different***.

Alternative hypothesis is the statement you want to use as a conclusion. (H_1 or H_a)

- $H_0 : \pi = 0.2$
  $H_1 : \pi > 0.2$
  This is an example of one-sided alternative.
- $H_0 : \mu_1 = \mu_2$
  $H_a : \mu_1 \neq \mu_2$
  This is an example of two-sided alternative.

Hypothesis Test

How we think when we conduct a hypothesis testing.

1. Compute the probability of acquiring the data we actually acquired assuming that H_0 is true.

   Strictly speaking, "acquiring the data we actually acquired ***or something more extreme***.
2. If that probability (p-value) is small, we say that something is wrong...
3. The data we have cannot be wrong, so what's wrong must be our assumption (i.e., H_0).

P-value

P-value is the probability of observing the data actually observed or something more extreme under H_0.

- Note that a p-value can be computed without ever referring to the alternative hypothesis.
- We can just compute the p-value, but a lot of times, we are required to make a go/no-go decision. So using the data, we decide to either "reject H_0" or "not reject H_0".
- The null hypothesis is a statement about the (true but unknown) population parameter, and it is either true or false. H_1 is a complement of H_0, and it is either true or false.
- Sometimes rejecting H_0 is correct, and sometimes it is not. The following table summarizes what happens when we reject or fail to reject H_0.

Errors

	Truth	
Conclusion	H_0 is true.	H_0 is not true.
Reject H_0	Type I error	Correct
Fail to reject H_0	Correct	Type II error

Type I error is an error of rejecting a true H_0.
We use α to denote the probability of this error.

Type II error is an error of failing to reject a false H_0.
We use β to denote the probability of this error.

Power is $1-\beta$: probability of correctly rejecting a false H_0.

- Customarily, we set α to 0.05 (or one-sided α to 0.025).
- ***Warning***: H_0 can never be shown to be true, i.e., even a p-value of 95% does not allow us to say "H_0 is shown to be true". Not even "H_0 seems to be true/believable/credible."

Ki67-positive cells example

Hypotheses:

$$H_0 : \mu_{wt} = \mu_{ko}$$
$$H_1 : \mu_{wt} \neq \mu_{ko}$$

Observed sample statistics

Wild type $N = 18, \bar{X}_{wt} = 48, sd_{wt} = 23$

Knockout $N = 16, \bar{X}_{ko} = 32, sd_{ko} = 18$

...

P-value is 0.019.

With these observed sample data, we reject H_0. We have enough evidence to claim that the true wild type mean is gerater. (with two-sided type I error rate of 5%).

- Strictly speaking, all we can say at this point is that the true means are different.

What are we saying?

We say, “The observed (sample) wild type mean is statistically significantly greater.”

Q: Is the true (population) wild type mean greater?
A: ***If*** the true means are equal, then the probability of observing what we observed is very small (0.019).

Q: So are you saying that the true wild type mean is greater?
A: No.

So what is the true difference of the means?

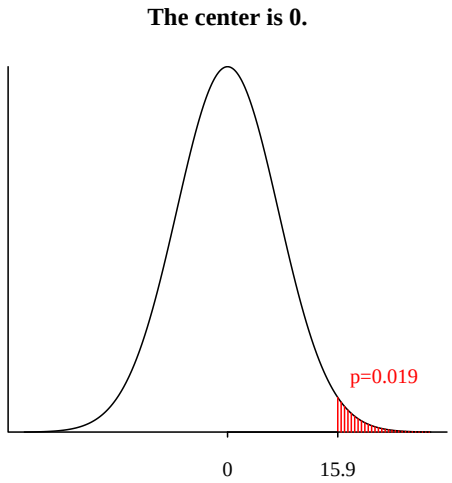
That is a more interesting research question than “Are the true means different?”.
$\Rightarrow$ Estimation

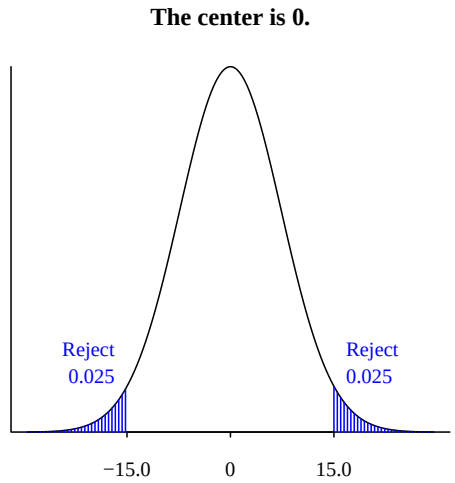
Estimation

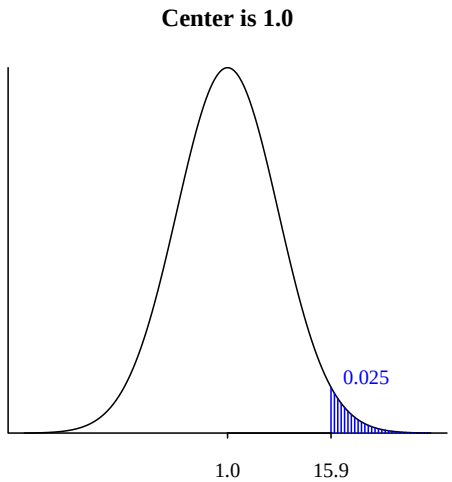
When we want to make inference about the ***population parameters***, we take a (representative) ***sample*** from the population and compute a ***statistic*** using the data from the sample. Using the sample statistic, we ***estimate*** the population parameter.

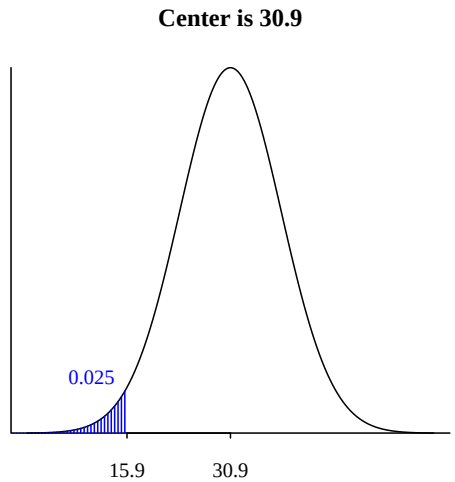
- Sample mean to estimate the population mean.
- Sample proportion to estimate the population proportion.
- Sample correlation to estimate the population correlation.
- Sample difference of means/proportions to estimate the population difference.

Ki67 example

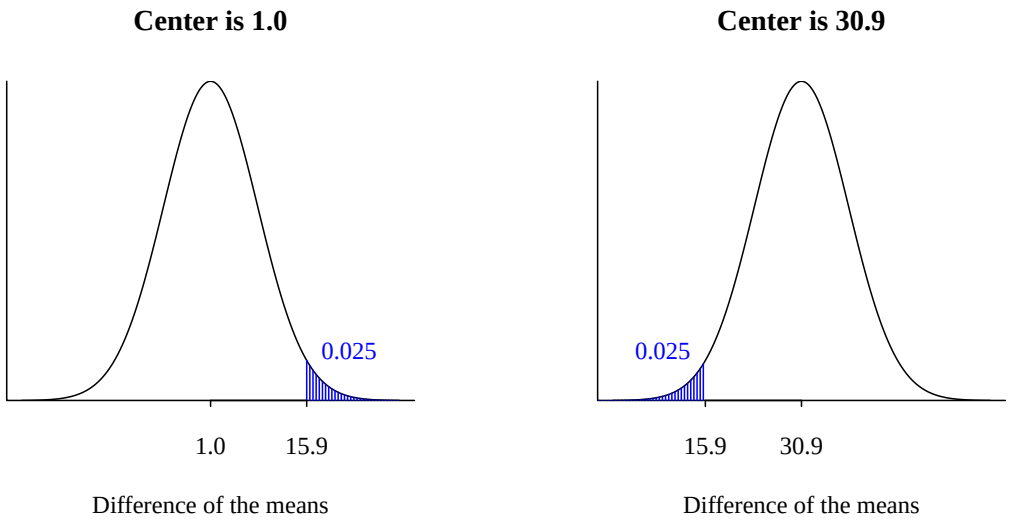








Confidence interval



If the true (unknown) mean difference is 1.0, observing a sample mean of 15.9 is barely plausible. And if the true mean difference is 30.9, observing a sample mean of 15.9 is barely plausible.

Any values between these two numbers would make observing $\bar{X}_{wt} - \bar{X}_{ko} = 15.9$ not very unusual. Let's cal this interval $(1.0, 30.9)$ a 95% confidence interval.

Interpretation of a confidence interval

A 95% confidence interval of the true difference of the means is $(1.0, 30.9)$.

- "Probability that the true difference of the means is between 1.0 and 30.9." is ***wrong***.
- "We are 95% confident that the true difference of the means is between 1.0 and 30.9."
- Remember the numbers 1.0 and 30.9 are specific to this particular sample. With a different sample, we will have a different confidence interval.
- Imagine repeating this experiment many times. Each sample will give us a different confidence interval. Most of (95%) these confidence intervals will contain the true unknown difference, but some of them (5%) won't.
- "We don't know if the true difference of the means is in $(1.0, 30.9)$. But we are using a process that produces intervals, 95% of which include the true difference." (We don't know if the one we have is one of them.)

This is the ***Frequentist*** inference.

Bayesian inference is more straightforward.

H_0 can never be shown to be true.

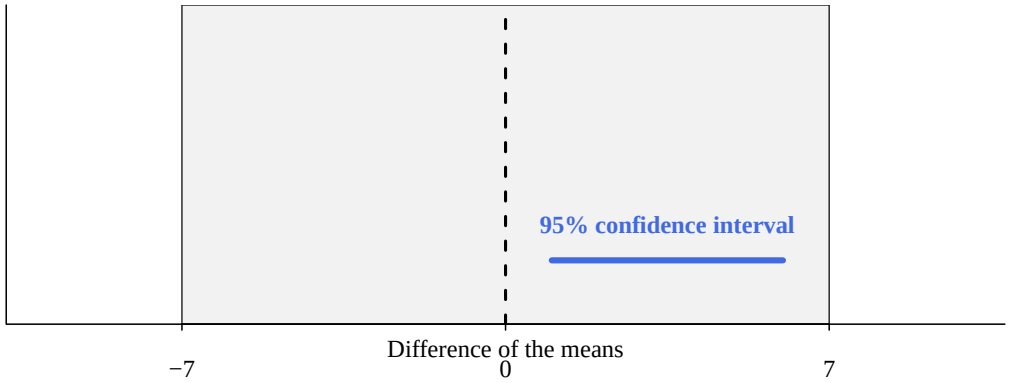
(H_0 is never true.)

Not rejecting H_0 does not allow you to say, "H_0 is true." -Why not?
You can always 'not reject H_0' by conducting a bad experiment.

If you want to 'not reject H_0'

- Choose $n = 1$
- Make data noisy (large variance)

If you want to show two things are about the same, show that the difference is smaller than the clinically (biologically) meaningful difference.



$$H_0 : \mu_{wt} - \mu_{ko} < -7 \text{ or } \mu_{wt} - \mu_{ko} > 7$$
$$H_1 : -7 \leq \mu_{wt} - \mu_{ko} \leq 7$$

Sample size and power

To test

$$H_0 : \mu_t - \mu_c = 0$$
$$H_1 : \mu_t - \mu_c > 0$$

take a random sample of size N from each group.

- Large N: Expensive. Unethical.
- Small N: Wasteful. Unethical.

N is a function of

- Type I error rate
- Power
- Alternative (where the power is set)
- Variability of data (Usually unknown)
- Analysis method

In some situation, N is also a function of

- Number of hypotheses (multiplicity)
- Correlation of the data (before-after)

How to reduce sample size

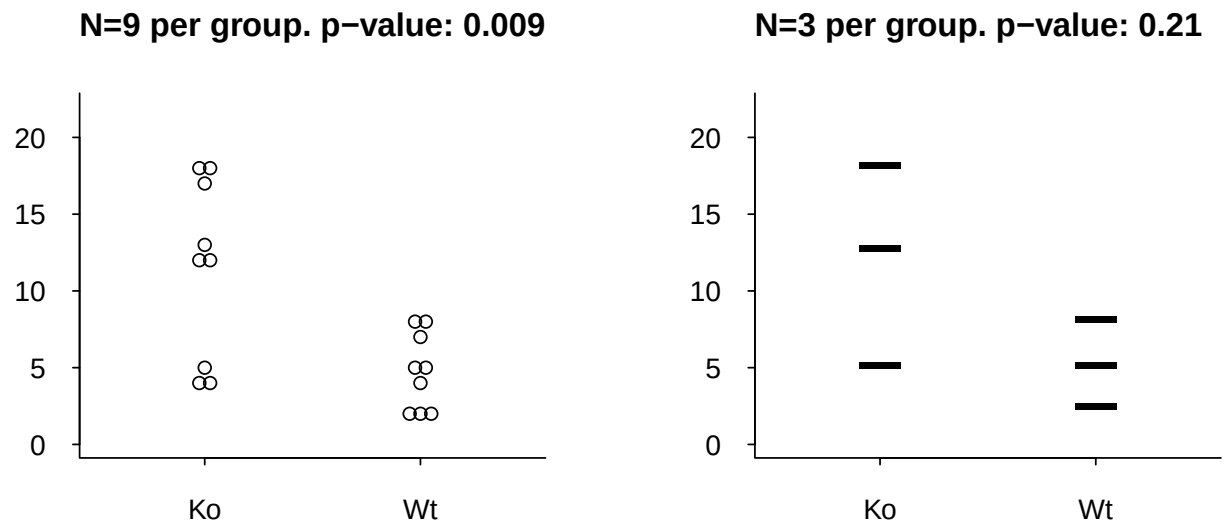
Factors that affect sample size and power.

Everything else being equal...

- Sample size $\uparrow$ ... Power $\uparrow$
- Type I error rate (α) $\downarrow$ ... Power $\downarrow$
- Difference to detect $\uparrow$ ... Power $\uparrow$
- Standard deviation $\uparrow$ ... Power $\downarrow$
- Correlation between repeated measures (Must be > 0.5) $\uparrow$ ... Power $\uparrow$

Cautions

1. (Cohen's d)
   - "To detect an effect size of 2"
   - "To detect a standardized difference of 2"
2. Duplicates and triplicates
   - $N = 9$ is different from 3 sets of $N = 3$ unless within varaiance is as large as beteen variance (no cluster).



Many statistical tests require a 'random sample,' and correlated (grouped) samples do not satisfy this.

Hypothesis Testing, Estimation, Sample Size and Power

Tatsuki Koyama, Ph.D.

Biostatistics Subcore
Vanderbilt Digestive Disease Research Center

January 21, 2020