ELEMENTS OF A HYPOTHESIS TEST, SAMPLE SIZE AND POWER CALCULATION

ELEMENTS OF A HYPOTHESIS TEST (REVIEW)

Elements of a hypothesis test

In a hypothesis test, we need to decide whether or not to reject a null hypothesis based on the data we observe.

- The null hypothesis H_0: specifies one or a subset of statistical models (from a "universe" of possible models).
  - H_0 is often stated as constraints on parameter values, e.g., $H_0: \mu_1 = \mu_2 = \mu_3$.
- A decision rule: states for what values of the observed data we will reject H_0 (the collection of those values is called the **rejection region**).
  - The decision rule or, equivalently, the rejection region should be based on the data alone, and cannot depend on the model parameter values (unknown to us)
- Power: the probability of rejecting H_0, i.e., the probability of the rejection region
- Type I error: rejection of H_0 when H_0 is actually true (false positive)
- Type II error: failure to reject H_0 when H_0 is false (false negative)
- Level of a test, α: a pre-specified nominal upper bound for the probability of making a Type I error. (α is often set at 0.05, but other values can be used too)
- β: the probability of making a Type II error (when H_0 is false), equals (1 - power)

In a hypothesis test, we need to decide whether or not to reject a null hypothesis based on the data we observe. Here we list some basic elements of a hypothesis test.

The null hypothesis, H naught, specifies one or a subset of statistical models from a "universe" of possible models. For our examples, the null hypothesis is often stated as constraints on parameter values. For example, (mu 1 = mu 2 = mu 3). But remember each set of parameter values corresponds to a statistical model.

A decision rule states for what values of the observed data we will reject the null hypothesis. The collection of those values is called the rejection region.

The decision rule or, equivalently, the rejection region should be based on the data alone, and cannot depend on the values of the model parameters, since the parameter values are unknown when we performing a hypothesis test.

In other words, the decision rule or the rejection region have to remain the same no matter what the underlying model is.

The power of a test is the probability of rejecting a false null hypothesis, i.e., the probability of the rejection region.

For the same set of observed values, we either reject the null hypothesis or not. There are two types of mistakes we can make with our decision:

A Type I error or a false positive is the rejection of a null hypothesis when it is actually true.

A Type II error or a false negative is the failure to reject a null hypothesis when it is actually false.

The level of a test, often denoted by alpha, is a pre-specified nominal upper bound for the probability of making a Type I error. alpha=0.05 is often used, but other values of alpha such as 0.02, 0.01, and so on can be used too.

beta is often used to denote the probability of making a Type II error. We can only make a type II error when the null hypothesis is false, and in that case, beta equals 1 minus the power of the test.

Type I and Type II errors

H_0 true?	Reject H_0? Y	N	
Y	Type I error (α)	Correct $(1-\alpha)$	1
N	Correct $(1-\beta)$	Type II error (β)	1

This table summarizes Type I error and Type II error in hypothesis testing.

ASIDE: size versus level of a test

There is subtle difference between **the size** and **the level** of a test, even though some people use the two terms interchangeably.

The level of a test is the pre-specified nominal upper bound for the type I error rate; the size of a test is the actual maximum type I error rate under the null.

- "Under the null" means when the null hypothesis is satisfied. Note that many models can satisfy the null hypothesis. That is why we consider the maximum type I error rate (the probability of making a type-I error): the maximum is taken over all models satisfying the null hypothesis.

If the test is valid, the size of test should be under control (i.e., below the specified level), but in practice, a poorly designed test may not have the type I error under control (i.e., the size of the test can exceed the nominal level).

There is subtle difference between the size and the level of a test, even though some people use the two terms interchangeably.

The level of a test is the pre-specified nominal upper bound for the type I error rate; the size of a test is the actual maximum type I error rate under the null.

"Under the null" means when the null hypothesis is satisfied. Note that many models can satisfy the null hypothesis. That is why we consider the maximum type I error rate (the probability of making a type-I error): the maximum is taken over all models satisfying the null hypothesis.

If the test is valid, the size of test should be under control (i.e., below the specified level), but in practice, a poorly designed test may not have the type I error under control (i.e., the size of the test can exceed the nominal level).

Strategy for hypothesis testing

In a hypothesis test, we have to make the rejection decision based on data alone, not knowing whether the underlying statistical model satisfies H_0 or not. **The decision rule, or equivalently, the rejection region is fixed** no matter what the true underlying model is.

Typically, there is not way to minimize both type I and type II error rates at the same time using the same rejection region:

- In most cases, the only way to minimize the type I error (false positive) rate is to make it 0 by never rejecting H_0 no matter what data we observe; the only way to minimize the type II error (false negative) rate is to make it 0 by always rejecting H_0 no matter what data we observe. Neither approach is sensible.

A sensible strategy is to choose a rejection region such that

- the type I error rate is bounded by a pre-specified α-level under models satisfying H_0, and
- the type II error is minimized (power maximized) under models not satisfying H_0.

In hypothesis testing, we have to make the rejection decision based on data, not knowing whether the underlying statistical model satisfy (H naught) or not.

The decision rule, or equivalently, the rejection region is fixed no matter what the true underlying model is.

In general, the underlying model for the data can be any model in the universe of all possible models. We want the decision rule to be sensible no matter what the true underlying model is.

Typically, there is no way to minimize both type I and type II error rates at the same time:

In most cases, the only way to minimize the type I error (false positive) rate is to make it 0 by never rejecting (H naught) no matter what data we observe;

The only way to minimize the type II error (false negative) rate is to make it 0 by always rejecting (H naught) no matter what data we observe.

Neither approach is sensible.

A sensible strategy is to choose a rejection region such that

the type I error rate is bounded by a pre-specified alpha-level under models satisfying the null hypothesis,

and the type II error rate is minimized under models not satisfying the null hypothesis.

Note that minimizing the type II error is equivalent to maximizing the power, since the power equals 1 minus the type II error for models not satisfying the null.

Strategy for hypothesis testing (continued)

The intuition for a good test is to the reject the null when the data observed is much more likely to happen under an alternative (non-null) model than under the a null model.

Often, our decision rule can be be formulated as comparing a summary statistic to a critical value (such as in a t-test or a F-test).

The intuition for a good test is to the reject the null when the data observed is much more likely to happen under an alternative (non-null) model than under the a null model.

Often, our decision rule can be be formulated as comparing a summary statistic to a critical value such as in a t-test or a F-test.

To reject or not to reject, connection to p-value

As an alternative to a simple reject/not-to-reject conclusion. We could also state our test result using a p-value.

Roughly speaking, the p-value the is probability that we could observe a data set as or more extreme than the one we actually have if the null hypothesis is true.

In practice, the p-value contains more information than a simple yes/no decision, since we can compare p-value to any α-level we want: if the p-value is 0.025, we know that we can reject H_0 at level $\alpha = 0.05$, but not at level $\alpha = 0.01$.

(But for power and sample-size calculation, we need to fix an α ahead of time.)

As an alternative to a simple reject/not-to-reject conclusion. We could also state our test result using a p-value.

Roughly speaking, the p-value the is probability that we could observe a data set as or more extreme than the one we actually have if the null hypothesis is true.

In practice, the p-value contains more information than a simple yes/no decision, since we can compare p-value to any alpha-level we want: if the p-value is 0.025, for example, we know that we can reject the null hypothesis at the level (alpha equals 0.05) but not at the level (alpha equals 0.01).

But for power and sample-size calculation, we need to fix an alpha ahead of time.

POWER AND SAMPLE SIZE CALCULATION

Power calculation: the two-sample comparison

Suppose we want to test $H_0: \mu_1 = \mu_2$, and the "truth" is that $|\mu_2 - \mu_1| = \delta > 0$.

Let $\Phi(z)$ be the probability function (c.d.f.), z_p the $(1-p)$th quantile of a standard normal random variable $Z \sim N(0,1)$:

- $\Phi(z) = Pr(Z \leq z)$, [$\texttt{pnorm}(\texttt{z}, \texttt{mean} = 0, \texttt{sd} = 1)$ in R], and
- z_p is the number such that $Pr(Z > z_p) = p$, [$\texttt{qnorm}(1-\texttt{p}, 0, 1)$ in R].

It can be shown that when $|\mu_2 - \mu_1| = \delta$, the power (i.e., the probability to reject H_0) of a two-sided two-sample t-test is approximately

$$\text{power}(\delta) \approx \Phi\left[z_{1-\alpha/2} + \sqrt{\frac{n}{2\sigma^2}} \cdot \delta\right]$$

The power depends on δ, α, σ^2, and n.

Suppose we want to test null hypothesis that the two means, (mu 1) and (mu 2) are equal in a two-group or two-sample comparison, and the truth is that the difference of (mu 2) and (mu 1) equals delta.

Let (phi of z) be the c.d.f. of a standard normal random variable, (z p) be the (1-p)th quantile of a standard normal random variable.

Then it can be shown that the power (i.e., the probability to reject H naught) of the two-sided two-sample t-test is given in this formula.

The power depends on delta, alpha, sigma squared, and the sample size n.

Understanding the power formula

For comparing two group mean parameters, we test $H_0: \mu_1 = \mu_2$. We will look at the distribution of a summary statistic:

$$Z = \frac{\overline{y}_2 - \overline{y}_1}{SE(\overline{y}_2 - \overline{y}_1)}$$

The numerator of Z, $\overline{y}_2 - \overline{y}_1$, estimates $\mu_2 - \mu_1$, and denominator of Z is the standard error of the numerator.

In fact, Z summarizes all the useful information in the data set about $\mu_2 - \mu_1$. To test $H_0: \mu_1 = \mu_2$, we can thus focus on the distribution of Z under different models.

In large samples, Z is approximately normally distributed.

On the next slide, we will examine the distribution of Z under the null when $\mu_2 - \mu_1 = 0$ and under an alternative model where $\mu_2 - \mu_1 = \delta > 0$.

For comparing two group mean parameters, we test (H naught) (mu 1) equals (mu 2). We will look at the distribution of a summary statistic Z.

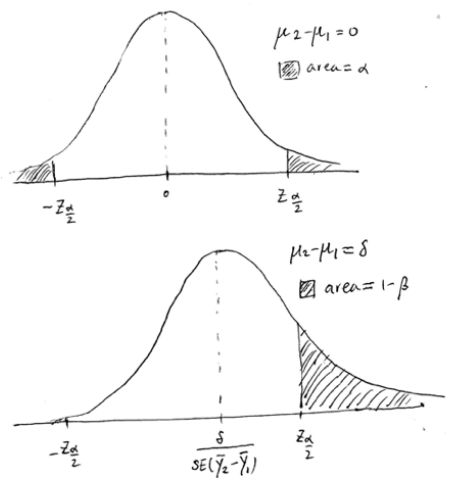
The numerator of Z, (y bar 2) minus (y bar 1), estimates (mu 2) minus (mu 1), and denominator of Z is the standard error of the numerator.

In fact, Z summarizes all the useful information in the data set about (mu 2) minus (mu 1). To test the null hypothesis that (mu 1) equals (mu 2), we can thus focus on the distribution of Z under different models.

In large samples, Z is approximately normally distributed.

On the next slide, we will examine the distribution of Z under the null when (mu 1) equals (mu 2) and under an alternative model where (mu 2) minus (mu 1) equals to a delta that is greater than 0.

Illustration of the power of a two-sample comparison



The distribution of Z under the null and under an alternative model

In the top plot, we show the distribution of Z when the null hypothesis is true.

If the null hypothesis is true, that is if (mu 2) minus (mu 1) equals 0, then Z follows a standard normal distribution.

The two shaded tail areas represent the rejection region for the test, having a total area alpha.

In the bottom plot, we show the distribution of Z under one possible alternative model.

In this case, the null hypothesis is false and (mu 2) minus (mu 1) is greater than 0. Z still follows an normal distribution, but mean of Z is shifted to delta, the true value of (mu 2 minus mu 1), divided by the standard error of (y 2 bar minus y 1 bar).

Under this different normal density curve under the alternative model, the probability of the rejection region is given by the power formula we showed earlier.

Note that the rejection region remains the same under both models, since our decision rule has to be the same for all models. The probability of the rejection region changes as the underlying model changes.

Intuition behind the power formula

Using a two-sided equal-variance two-sample t-test, the power/probability to reject $H_0: \mu_1 = \mu_2$, when actually $|\mu_2 - \mu_1| = \delta$, is approximately

$$\text{power}(\delta) \approx \Phi\left[z_{1-\alpha/2} + \sqrt{\frac{n}{2\sigma^2}} \cdot \delta\right]$$

The power depends on δ, α, σ^2, and n. Note that the power increases as

- n increases (the sample size per group, Kuehl's r);
- δ increases (the difference between μ_1 and μ_2); and
- σ^2 decreases (the variability within group).
- α increase (at the cost of increase type I error rate)

Note:

- In fact, the power depends on σ and σ^2 only through the ratio δ/σ, which can be thought as the signal to noise ratio.
- Also, this power formula assumes equal population variation in the two groups.

Now let's take a closer look at this power formula for a two-sided equal-variance two-sample t-test. We see that the power depends on delta, alpha, sigma squared and n.

Note that the power increases as n, the sample size per group, increases.

The power increases as delta, the magnitude of the difference between the two mean parameters, increases.

The power increases as sigma squared, the variability within each group. decreases.

And finally, the power increases as alpha, the type I error rate, increases.

All these make intuitive sense.

In fact, the power depends on delta and sigma squared only through the ratio (delta over sigma), which can be thought as the signal to noise ratio.

Also, this power formula assumes equal population variation in the two groups.

Sample size calculation

If any four of these four quantities (power, n, δ, σ^2, α) are known, the fifth can be solved for.

For example, the power formula can be re-arranged to give the sample size corresponding to a particular power, δ, σ^2, and α-level:

$$n = 2\sigma^2 \left(\frac{z_\beta + z_{\alpha/2}}{\delta} \right)^2 .$$

If any four of these four quantities (power, sample size n, delta, sigma squared, and alpha) are known, the fifth can be solved for.

For example, the power formula can be re-arranged to give the sample size corresponding to a particular power, delta, sigma squared, and alpha.

Example

Suppose we will be comparing the change in diastolic blood pressure of a group of patients who have been taking a drug for six weeks to that of a control group. Assume we know that $\sigma^2 = 42.5$ (this is the MSE from the analysis based on the bp0 data). How many patients per group would be needed to have 80% power to detect a 10-unit difference in blood-pressure change between two groups?

(Work out your own solution before moving on to the next slide.)

Now let's do some examples.

Suppose we will be comparing the change in diastolic blood pressure of a group of patients who have been taking a drug for six weeks to that of a control group. Assume we know that (sigma squared equals 42.5). This is the MSE from the analysis based on the bp0 data.

How many patients per group would be needed to have (80 percent) power to detect a 10-unit difference in blood-pressure change between two groups?

Please pause the lecture and work out your own solution before moving on to the next slide.

Example

Suppose we will be comparing the change in diastolic blood pressure of a group of patients who have been taking a drug for six weeks to that of a control group. Assume we know that $\sigma^2 = 42.5$ (this is the MSE from the analysis based on the bp0 data). How many patients per group would be needed to have 80% power to detect a 10-unit difference in blood-pressure change between two groups?

$$1 - \beta = 0.8 \Longrightarrow z_\beta = z_{0.2} = 0.4$$
$$\alpha = 0.05 \Longrightarrow z_{\alpha/2} = z_{0.025} = 1.96$$

So, $n = 2 \cdot 42.5 \cdot (\frac{0.84+1.96}{10})^2 = 6.7$, which we round up to 7.

We would need 7 patients in each group to have 80% power to detect a 10-unit difference in blood-pressure change between groups.

To compute the sample size, we need to know the power, alpha-level, delta, and sigma-squared. From the problem description, we see that power is 0.8, so that beta equals 1 minus power is 0.2. alpha-level is not mentioned, we can use alpha equals 0.05 here. delta equals 10 units. And finally, sigma squared is 42.5.

Plugging these information into the sample size formula, we get n equal 6.7, which we would round up to 7.

To conclude the problem: we would need 7 patients in each group to have 80 percent power to detect a 10-unit difference in blood-pressure change between two groups.

Power and sample size calculation using R

Once we understand the ideas behind the power and sample size formula.
In practice, we do not need to do the calculations by hand.

In R, the function `power.t.test` uses a more sophisticated algorithm than the above approximation, but usually gives similar answers. Here,

```
power.t.test(delta=10, sd=sqrt(42.5), power=0.8, sig.level=0.05, type="two.sample");
##
##      Two-sample t test power calculation
##
##               n = 7.760289
##           delta = 10
##              sd = 6.519202
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

gives $n = 7.76$ which we will round up to 8.

Type [? power.t.test] in R to find out more details on how to use power.t.test.

Once we understand the ideas behind the power and sample size formula. In practice, fortunately for us, we do not need to do the calculations by hand.

In R, the function power.t.test uses a more sophisticated algorithm than the above approximation, but usually gives similar answers.

Here it gives n equals 7.75, which we will round up to 8.

In the power formula we derived earlier, we used normal approximation. The R function works on the t-distribution directly. So the two will give slightly different results in small-sample situation.

More than two groups

In the single-factor ANOVA setting, we need to consider comparisons among more than two groups. The calculations, based on the noncentral F-distribution, become more complicated (see Kuehl, p. 63-65).

In a single-factor ANOVA, factors affecting the power/sample size include

1. number of groups (treatments)
2. number of observations in each group
3. population mean level of each group
4. population variance within each group (common to all groups)
5. significance level
6. power of the test

Given 1–5, we can determine the power of the ANOVA F-test; Given 1, 3–6, we can determine 2 (group size).

In the single-factor ANOVA setting, we need to consider comparisons among more than two groups. The calculations, based on the noncentral F-distribution, become more complicated.

In a single-factor ANOVA, factors affecting the power/sample size include

1.number of groups (treatments)

1.number of observations in each group

1.population mean level of each group

1.population variance within each group (common to all groups)

1.significance level

1.power of the test

Given 1–5, we can determine the power of the ANOVA F-test; Given 1, 3–6, we can determine 2 (group size).

Example.

Example 2 (Ex. 2.3 on p. 64) Ear inflammation in rabbits as a function of three treatments. How many rabbits would we need for 90% power to detect a treatment effect, if truly $\mu_1 = 0.8$, $\mu_2 = 0.1$, and $\mu_3 = 0$? Assume $\sigma^2 = 0.22$.

- Using Kuehl's approach (p. 64), we find $n = 9$.
- In R, the following code gives $n = 8.4$ (round up to 9).

```
means <- c(0.8, 0.1, 0)
power.anova.test(groups=3, between.var=var(means), within.var=0.22, power=0.9)
##
##      Balanced one-way analysis of variance power calculation
##
##           groups = 3
##                n = 8.417699
##      between.var = 0.19
##       within.var = 0.22
##        sig.level = 0.05
##            power = 0.9
##
## NOTE: n is number in each group
```

Let's look at an example.

In this experiment, a researcher plan to study the ear inflammation in rabbits in three treatment groups. Based on numbers from previous studies, we assume treatment effects to be (mu 1 equals 0.8), (mu 2 equals 0.2), (mu 3 equals 3), and assume the with-in group variance to be (sigma squared = 0.22). Suppose we need to have 90 percent power in the ANOVA test using a completely randomized design.

We can compute the sample size needed using the R function power.anova.test. We find that the sample size needed is 8.4 per group, which we will round up to 9.

Example (continued)

Using the two-sample approach. Given: $\sigma^2 = 0.22$; $\alpha = 0.05$; and power $= 0.9$. Set $\delta = \mu_1 - \mu_3 = 0.8$. Then

$$n = 2\sigma^2 \left(\frac{z_\beta + z_{\alpha/2}}{\delta}\right)^2 = 2 \cdot 0.22 \cdot \left(\frac{1.28 + 1.96}{0.8}\right)^2 = 7.2$$

which rounds up to 8 rabbits.

These different approaches usually give similar answers.

If we use a two-sample approach and ask what the sample size need to detect the greatest mean difference among groups, that is, the mean difference between group 1 and group 3.

We will get n equal 7.2, which we round up to 8 rabbits.

These different approaches usually give similar answers.

Example (continued)

If we know we can only afford 6 rabbits per group, other settings being the same. We can use power.anova.test to compute the power of the study:

```
means = c(0.8, 0.1, 0);
power.anova.test(n = 6, groups=3, between.var=var(means), within.var=0.22)
##
##      Balanced one-way analysis of variance power calculation
##
##          groups = 3
##               n = 6
##     between.var = 0.19
##      within.var = 0.22
##       sig.level = 0.05
##           power = 0.7418642
##
## NOTE: n is number in each group
```

We find that the power of the ANOVA F-test will be 0.74.

If we know we can only afford 6 rabbits per group, other settings being the same. We can use power.anova.test to compute the power of the study. We find that the power of the ANOVA F-test will be 0.74.

Summary

1. Review of elements of hypothesis testing
2. Power and sample size formulas for two-sample *t*-test
3. Power and sample size calculation in R using power.t.test and power.anova.test

In this lecture, we reviewed the basic elements of hypothesis testing.

We discussed the ideas behind the power and sample size calculations using the two-sample t-test as example.

We learned how to do power and sample size calculation in R using the R functions power.t.test and power.anova.test.