Type I and II Errors, Power

Kris Gunsalus

11/2/2020

Contents

Background Reading

Aho, Chapter 6.3-6.4: Type I and Type II Errors; Power

Which statistical test should I use?

In this course, we introduce you to several basic methods for **hypothesis testing** that allow us to determine whether observed differences between two or more groups are statistically different. The choice of an appropriate test will depend on what kind of question you are interested in asking.

As the course progresses, we are making our way through a variety of **parametric** tests for comparing differences between samples: t-tests, ANOVA, simple linear models, and linear models with mixed effects (i.e. interaction terms). We also go over **nonparametric tests** and tests for **correlations** and **categorical** data.

The diagram below illustrates how to choose the right test depending on your question. We will revisit this framework as we progress through the semester.

Type I and II Errors

When we make statistical inferences, we are making statements about the likelihood that some hypothesis is true based on a chosen significance threshold, based on data from a finite set of experimental samples. By convention, a significance cutoff of $\alpha = 0.05$ is often chosen, meaning that *5% of the time we are likely to get a value as extreme or more extreme than the one we have observed in our test sample* ***just by random chance***. That means that 5/100 times, a sample drawn from the null distribution will result in a false conclusion that the sample is NOT drawn from the null distribution. There are two types of errors we can make:

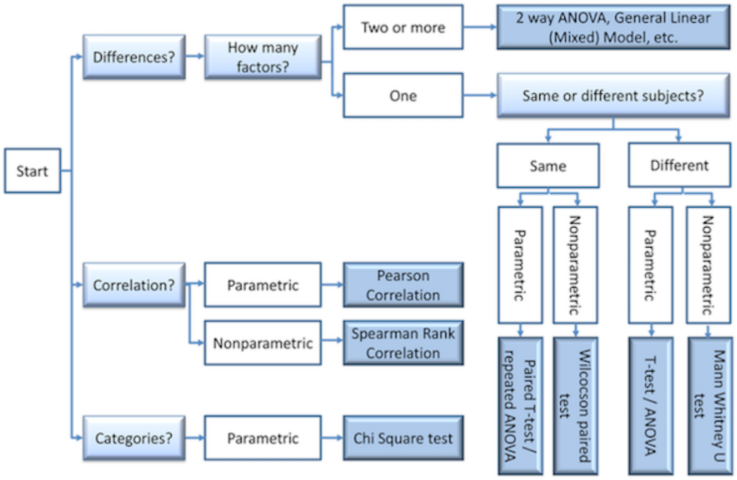


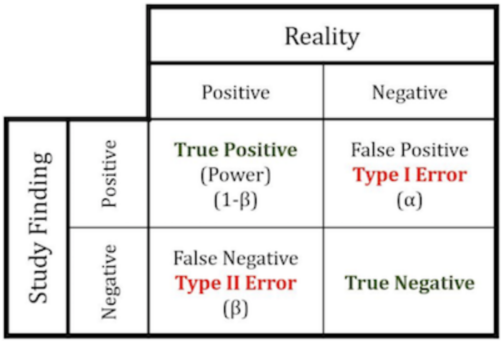
Figure 1: Decision tree for choosing a statistical test

Type I error (α): A "false positive"

- concluding that there is an effect when none exists
- incorrectly rejecting the null hypothesis (accept the alternative H_A) when H_o is true.
- probability of observing a value as extreme or more extreme by random chance, if the sample comes from the null distribution.

Type II error (β): A "false negative"

- concluding that there is no effect when there is one
- incorrectly accepting H_o (rejecting H_A) when the alternative hypothesis is true.
- probability that an extreme observation from the non-null distribution could overlap the null distribution just by chance.



		Reality	
		Positive	Negative
Study Finding	Positive	True Positive (Power) (1-β)	False Positive Type I Error (α)
	Negative	False Negative Type II Error (β)	True Negative

Figure 2: Possible outcomes in hypothesis testing

The **tradeoff** between Type I and Type II errors is illustrated by a diagram showing two different populations whose distributions partially overlap.

When designing experiments, it is important to decide in advance what you will consider as acceptable Type I and Type II error rates. Depending on one's goals, either of these can be undesirable. For example:

- In many applications, Type I errors are considered especially bad, for instance when reporting results for RNAi screens or differential gene expression, since they can result in a lot of extra work that does not pan out.

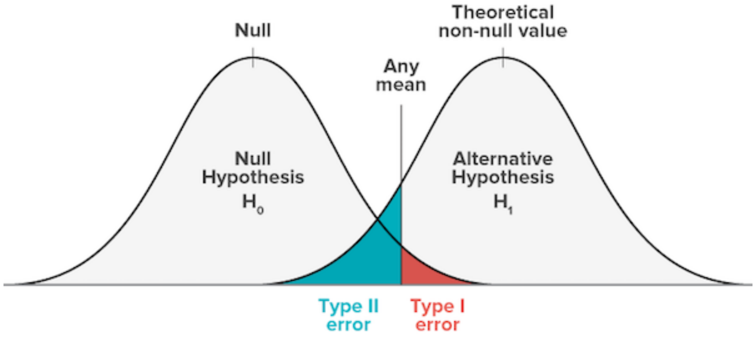


Figure 3: ***Errors in acceptance/rejection regions***

- On the other hand, casting a wide net in a primary screen at the expense of a few false positives may be ok if you want to minimize your chances of missing anything important, and you plan to double-check all of your primary data in a second round of testing (e.g. secondary RNAi screens with more replicates, or qPCR to confirm differential expression for genes of interest.)

Designing experiments

The choices you make in your experimental design should aim to **eliminate bias** and **reduce the influence of sampling error**.

Factors that reduce **bias**:

- Simultaneous controls
- Randomization
- Blinding

Factors that reduce **sampling error** increase both ***precision*** and ***power***:

- Replication
- Balance
- Blocking

Power Analysis

Power analysis tells us how confident we can be that the results we have observed are realistic. That is, how confident can we be, when we reject the null hypothesis, that our conclusion is reasonable? The power of an experiment depends on several variables:

- **Effect size (*E*)**: This is the expected (true) difference between the population means, $\mu_1 - \mu_2$. A greater effect size gives a better chance of distinguishing between two populations (i.e. getting "true positives").
- **Population variance (σ^2)**: Populations with less variation make it easier to distinguish between them for a given effect size.
- **Sample size (*n*)**: Since we know from the Central Limit Theorem that increasing sample size gives us a better estimate of the population mean (standard error), it also gives greater power.
- **Significance threshold (α)**: Relaxing our criteria for rejecting the null hypothesis will result in a larger acceptance region for a given effect size.

Given **any four of the five parameters** that define the power relationship, it is possible to **estimate the remaining one**. This allows us to answer a variety of related questions, for example:

- Given a chosen significance threshold and expected effect size, how big of a sample do I need to achieve a particular power?

- Given my sample size, how much power will I have to detect an effect of a particular size?
- Given a chosen critical value (significance threshold) and power, how big of an effect will I be able to reliably detect?

Goals for experimental power should always be specified in advance for a fixed level α and β.

Example: Alzheimer's and smoking

To illustrate the idea of power, and the factors that affect it, below will walk through **Example 6.10** from Ken Aho's book in class. The problem is:

- Alzheimer's seems to be negatively associated with moderate smoking!!! Possibly because nicotine may reduce apoptosis (programmed cell death) of neurons
- Researchers want to know if a sample size of 200 at $\alpha = 0.05$ is sufficent to detect a decrease of 7% on Alzheimer's in subjects that smoke 10-20 cigarettes per day, given that $\sigma = 45\%$.

What is the power of the experiment? In other words, what is the probability of rejecting the null hypothesis if the effect of smoking is a 7% decrease in Alzheimer's?

An annotated version of the accompanying figure (Fig. 6.7) illustrates how the above variables affect the power to detect true differences in sample means. Panels (b-d) show that the following differences in the scenario would increase power:

(b) Larger effect size (E) – a bigger separation between population means
(c) Larger sample size (n) – this will decrease the SEM
(d) Lower stringency of the test (increase α) – increase the rejection region

Calculating power

Recall that the probability of a **Type I** error is the chosen significance threshold for an experiment, α. The probability of a **Type II** error is denoted by β. Now we can define the basic form of an equation to compute the **Power** for an experiment in terms of the significance threshold, effect size, variation, and sample size:

$$Power = 1 - \beta \propto \frac{E\alpha\sqrt{n}}{\sigma} = \frac{E\alpha}{\sigma/\sqrt{n}} = \frac{E\alpha}{SEM}$$

When designing experiments, **it is good practice to choose a desired significance and power for your experiment in advance**. Typical values are $\alpha = 0.05$ and $1 - \beta = 0.8$. In this case, we would like to detect outliers at a significance threshold (false positive rate) of 5%, with a power of 80%.

Estimating adequate sample size

Population mean estimates (review) Recall that when we learned about **confidence intervals**, we derived an **estimate for the population mean** as a function of the sample mean, sample size, and desired range $\gamma = 1 - \alpha$. We expect that $\gamma * 100\%$ of such intervals will contain the true population mean (e.g. the 95% CI):

$$\mu = \bar{X} \pm z_{1-(\alpha/2)} * \frac{\sigma}{\sqrt{n}}$$

Here, z is the **quantile function** for a standard normal distribution at probability $1 - \alpha/2$ and represents the z-score, or number of standard deviations away from the mean of a normal distribution.

The **margin of error**, $m = z_{1-(\alpha/2)} * \frac{\sigma}{\sqrt{n}}$, represents how far away the confidence bounds will be for a given CI.

Rearranging this equation allows us to estimate the required **sample size** to achieve a particular significance level and desired margin of error m:

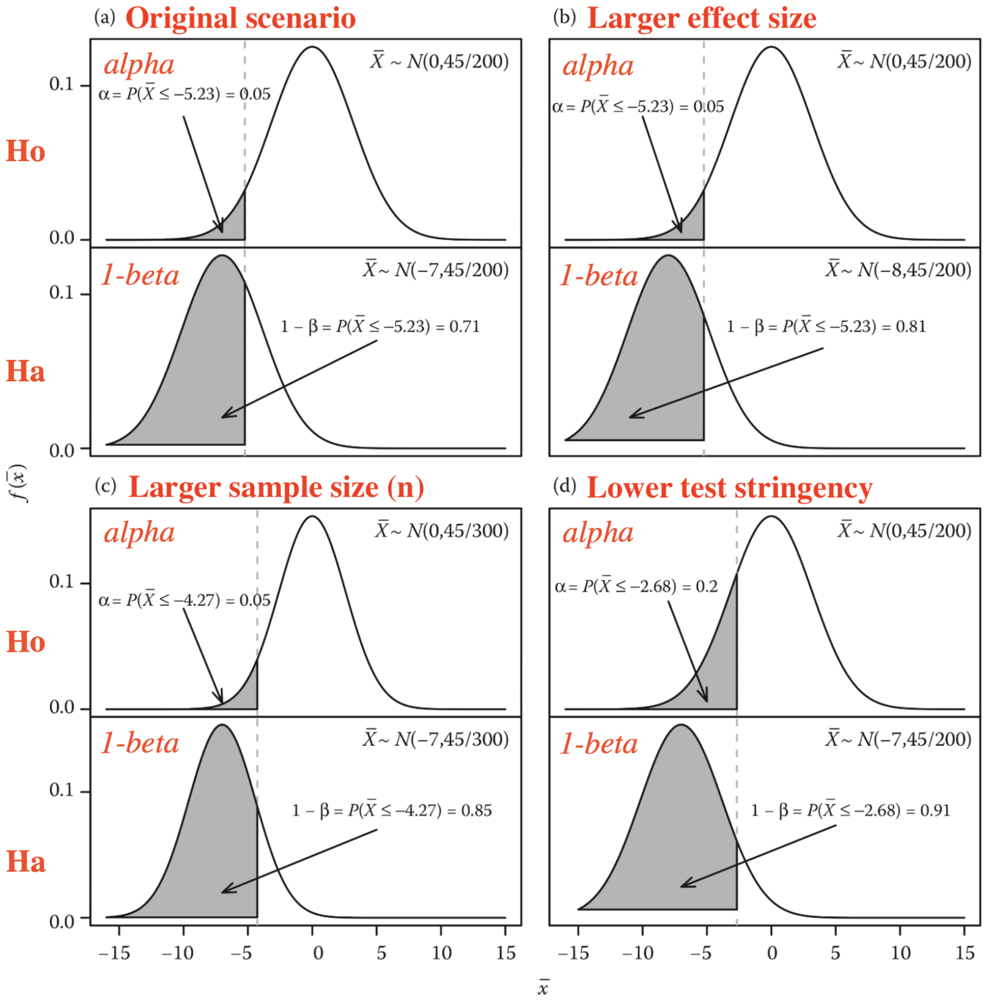


Figure 4: ***Aho, Figure 6.7 (Example 6.10)***

$$n = \left(\frac{z_{1-(\alpha/2)}\sigma}{m} \right)^2$$

Power Similarly, for **power calculations**, we can estimate the sample size required to detect a particular **effect size** $(\mu - \mu_o)$, given a predetermined **power** $(1-\beta)$ and **Type I error** (α).

For a **one-tailed** z-test, this gives:

$$n \approx \frac{(z_{1-\alpha} + z_{1-\beta})^2\sigma^2}{(\mu - \mu_o)^2}$$

For a **two-tailed** z-test, this gives:

$$n \approx \frac{(z_{1-(\alpha/2)} + z_{1-\beta})^2\sigma^2}{(\mu - \mu_o)^2}$$

These equations differ only in whether the Type I error, α, is found on one side or is split between both sides of the null distribution. Recall that for $\alpha = 0.05$, $z_{1-\alpha/2} \approx 1.96$ and $z_{1-\alpha} \approx 1.645$. Similarly, for 80% power, $1 - \beta = 0.8$ and $z_{1-\beta} \approx 0.842$.

Note that **we either need to know the population variance, or have a good estimate of it** (when the sample size is large, this can be estimated using the sample variance).

Power for t-tests

Power analyses can be performed in a similar way for t-tests. In this case the null distribution follows a "non-central" t-distribution (where the expectation value can be non-zero) that will depend on the degrees of freedom. Again, some estimate for the population variance is needed.

Cohen's d

Above we discussed the **effect size** in terms of the true difference between means, $\mu - \mu_o$ (or, more generally, $(\mu_X - \mu_Y) - D_o$). When the true effect size or population variance are not known, as is often the case, we can use an alternative measure of effect size, Cohen's d[1]:

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}$$

where σ is the common group variance (the variance of the two groups is expected to be the same). Cohen's d is a **normalized** effect size that measures the difference between two groups in standard deviations.

Cohen's d can be used to make power calculations assuming "small", "medium", or "large" effect sizes. (For his work, Cohen set the boundaries between these categories at $d = 0.2$ and $d = 0.8$). It is common to set $d = 0.5$, as anything larger than this is considered to be a moderate to large effect.

Power calculations in R

R functions for power calculations will compute, given any four arguments, a value for the missing variable. For example, they can be used to compute the expected power given a certain sample size and expected effect size, or the required sample size to achieve a given desired power, etc.

In practice, formulas for power estimates will depend on the type of experimental scenario and statistical test one is performing (e.g. t-test, proportions test, ANOVA, etc.). Base R and the `asbio` and `pwr` packages provide a variety of functions for performing different kinds of power tests. Not surprisingly, these use different

[1] Cohen (1988), *tatistical Power Analysis for the Behavioral Sciences (2nd ed)*

names for the command arguments, so make sure to check the documentation. Additional packages are available for different applications in statistical genomics.

Base R

Power functions include:

- `power.t.test()` - *t*-tests
- `power.prop.test()` - proportions tests
- `power.anova.test()` - ANOVA

asbio

The `asbio` package implements a function for a *z*-test, `power.z.test()`, as an alternative to `power.t.test()`, which can be used if the population sigma is known and the sample size is large.

pwr

Functions in this package begin with `pwr` instead of `power`. The package does not contain a function for a *z*-test, but it is much more general and contains functions for *t*-tests (with even or uneven sample sizes), proportions, Chi-square, ANOVA, correlation, and general linear models. The package is pretty well-documented, and several sites provide **tutorials** and **vignettes** with examples for different use cases.

NOTE: The `pwr` package uses **Cohen's** ***d*** for the effect size, rather than **delta** (the true difference in means). Tests can be run using a pre-computed value for *d*, or using "small", "medium", or "large" effect sizes. The function `cohen.ES()` will provide a standard value for each of these categories for a given flavor of test.

Example

Aho, Ex. 6.10 (also see Fig. 6.7)

We will set up this example and go through it together in class.

First, compute power by hand:

```
# =================================================== #
# set up variables

effect.size = -7  # effect size = Exp(X) under H_A
n = 200
sigma = 45
alpha = 0.05
type = "one.sample" # one or two sample
alt = "one.sided"  # one- or two-sided

# set Exp(X) = 0 under null H_o

# critical value (z*) for lower-tail test at alpha=0.05
z.crit = qnorm(alpha)  # -1.644854

# check alpha using standard normal distribution
pnorm(0,abs(z.crit),lower.tail=T)  # 0.05
## [1] 0.05

# compute SEM for sample size
sem = sigma/sqrt(n)
```

```r
sem
## [1] 3.181981

# =================================================== #
# manual power calculation

# compute power using area under the curve for H_A

# get value of critical x at z.crit for H_A
# want P( X.bar  <= z.crit * sem)
# percent difference for lower-tail significance
x.crit = z.crit*sem  # x-value at critical z-score
x.crit
## [1] -5.233892

# Expected X.bar (pop. mean) under H_A is (mu_o - mu_A): Exp(X) = -7
pwr = pnorm(x.crit, mean = effect.size, sd = sem)  # power = 0.71
pwr
## [1] 0.7105643

# check z-score for H_A at expected power
qnorm(pwr, effect.size, sem)  # alpha = P(X.bar <= -5.24)
## [1] -5.233892
```

Compute power directly in R

Given any 4 of the 5 variables that go into the power equation, we can use `power.t.test()` to compute the missing value. Since n is large, we could also use the `power.z.test()` command from the `asbio` package. These give slightly different results, as the t-test is a bit more conservative. (They also use different names for their arguments, and the objects the produce are also different.)

NOTE: *Effect size used for these functions should be given as a positive number, otherwise these functions will not work as expected.*

```r
# ================================================== #
# provide expected effect size as a positive number
power.t.test(n, delta = abs(effect.size), sd = sigma, sig.level = alpha,
             type="one.sample", alternative="one.sided", strict=T)
```

```
##
##      One-sample t test power calculation
##
##               n = 200
##           delta = 7
##              sd = 45
##       sig.level = 0.05
##           power = 0.7079982
##     alternative = one.sided
```

```r
# note that arguments for this command differ
power.z.test(n, effect = abs(effect.size), sigma = sigma,
             alpha = alpha, test="one.tail", strict=T)
```

```
## $sigma
## [1] 45
##
## $n
## [1] 200
##
## $power
## [1] 0.7105643
##
## $alpha
## [1] 0.05
##
## $effect
## [1] 7
##
## $test
## [1] "one.tail"
```

What if you change different variables that influence power?

- Increase effect size => increase power (reduce Type II error)
- Increase sample size => increase power (reduce Type II error)
- Raise α => lower stringency (increase Type I error)

We can compute the new power by hand, or use the `power.z.test()` command:

```
# ==================================================== #
# increase effect size
# ==================================================== #
# what happens if E = -8? => increase power
# (keep alpha the same)
pnorm(x.crit, effect.size - 1, sem)  # power = 0.81
```

```
## [1] 0.8076595
```

```
power.z.test(n, effect = -(effect.size-1), sigma = sigma,
             alpha = alpha, test="one.tail", strict=T)
```

```
## $sigma
## [1] 45
##
## $n
## [1] 200
##
## $power
## [1] 0.8076595
##
## $alpha
## [1] 0.05
##
## $effect
## [1] 8
##
## $test
## [1] "one.tail"
```

```
# ==================================================== #
# increase sample size
# ==================================================== #
# what if sample size = 300? => more power for same E
# (keep alpha the same)
```

```
n = 300

# get x-bar and SEM
sem = sigma / sqrt(n)
sem
```

```
## [1] 2.598076
```

```
x.crit = qnorm(0.05)*sem
x.crit
```

```
## [1] -4.273455
```

```
# power
pnorm(x.crit, effect.size, sem)  # power = 0.85
```

```
## [1] 0.8530139
```

```
power.z.test(n, effect = -effect.size, sigma = sigma,
             alpha = alpha, test="one.tail", strict=T)
```

```
## $sigma
## [1] 45
##
## $n
## [1] 300
##
## $power
## [1] 0.8530139
##
## $alpha
## [1] 0.05
##
## $effect
## [1] 7
##
## $test
## [1] "one.tail"
```

```
# ================================================== #
# relax stringency: raise alpha
# ================================================== #
# raising alpha increases Type I error
# what happens to power? => power goes down
alpha = 0.2
z.crit = qnorm(alpha)  # -0.842

# check alpha2 using standard normal distribution
pnorm(0,abs(z.crit),lower.tail=T)
```

```
## [1] 0.2
```

```
x.crit = z.crit*sem
x.crit
```

```
## [1] -2.186596
```

```
pwr = pnorm(x.crit, mean = effect.size, sd = sem)  # power = 0.913
pwr
```

```
## [1] 0.9680359
```

```
qnorm(pwr, effect.size, sem) # check power
```

```
## [1] -2.186596
```

```
# power is the same with the z-test power function
power.z.test(n, effect = -effect.size, sigma = sigma,
             alpha = alpha, test="one.tail", strict=T)
```

```
## $sigma
## [1] 45
##
## $n
## [1] 300
##
## $power
## [1] 0.9680359
##
## $alpha
## [1] 0.2
##
## $effect
## [1] 7
##
## $test
## [1] "one.tail"
```

Design for a targeted power

What if you want to design the experiment for power = 0.8, for the same sample size and effect size? What is the Type II error? What happens to the Type I error?

```
# ================================================== #
# increase desired power to 0.8
# ================================================== #
# if raise desired power without increasing effect size,
#   => alpha goes up (less stringent)
x.bar2 = qnorm(0.8, effect.size, sem) # effect size -4.32
x.bar2
```

```
## [1] -4.813404
```

```
# what significance level is this?
alpha2 = x.bar2 / sem
alpha2
```

```
## [1] -1.85268
```

```
pnorm(0,abs(alpha2),lower.tail=T)  # 0.087
```

```
## [1] 0.03196412
```

```
# what significance level is this?
alpha2 = x.bar2 / sem
alpha2
```

```
## [1] -1.85268
```

```r
pnorm(0,abs(alpha2),lower.tail=T)  # 0.087
```

```
## [1] 0.03196412
```

```r
# ================================================== #
# using power.t.test command
# now supply power and ask what new alpha is
power.t.test(n, delta = -effect.size, sd = sigma,
             sig.level = NULL, power = 0.8,
            type="one.sample", alternative="one.sided", strict=T)
```

```
## Warning in pt(qt(sig.level/tside, nu, lower.tail = FALSE), nu, ncp = sqrt(n/
## tsample) * : full precision may not have been achieved in 'pnt{final}'
```

```
##
##      One-sample t test power calculation
##
##               n = 300
##           delta = 7
##              sd = 45
##       sig.level = 0.03251671
##           power = 0.8
##     alternative = one.sided
```

```r
# gives alpha = 0.088
```