

# Biostatistics II: Basic Analytic Statistics

## Hypothesis Testing

- What is a null hypothesis?
- How does a null hypothesis differ from your scientific hypothesis?
- What are the level and power of a test?
- What is the P value?
- What tests might you use in common settings?
- How should you report your findings?

Example from abstract of a recent paper on Women's Health Initiative, women 50-59:

“Estrogen therapy and coronary artery calcification”,  
Manson et al., *NEJM* 21 June 2007.

“The mean coronary-artery calcium score after trial completion was lower among women receiving estrogen (83.1) than among those receiving placebo (123.1) (P=0.02 by rank test).

After adjustment for coronary risk factors, the multivariate odds ratios for coronary-artery calcium scores of more than 0, 10 or more, and 100 or more in the group receiving estrogen as compared with placebo were 0.78 (95% confidence interval, 0.58 to 1.04), 0.74 (0.55 to 0.99), and 0.69 (0.48 to 0.98), respectively.”

Note that descriptive summary (group means) is supplemented by *inferences* that enable the reader to compare the two treatment groups.

The researchers present data on two questions:

- Do estrogen-treated women have different mean coronary artery calcification scores than the placebo group?
- Does the proportion of women with clinically elevated scores differ between the two groups?

Biostatistics lets us determine whether these differences could be accounted for just by chance variation.

Let's start with their main question: do means differ?

The *scientific hypothesis* might be that estrogen inhibits the formation of atherosclerosis during the early period after menopause.

The *null hypothesis* is that the population means do not differ, and any difference you see in the sample is just due to chance variation.

The *alternative hypothesis* is that the two population means really are different.

Hypothesis testing gives us a tool to decide between null and alternative as explanations for the data.

Let's return to our simple example, Bergen study.

We found that at oxygen uptake of 1 liter/min, the respiratory frequencies were:

	Mean	St Dev	Sample Size
Females	22	3.5	230
Males	17	4.5	143

Are population means different for men and women?

Clearly the sample means are not identical; but maybe this is explained just by individual variation!

Hypothesis testing offers us two choices:

1. Conclude that the difference between the two groups is so large that it is unlikely to be due to chance alone. *Reject* the null hypothesis and conclude that the groups really do differ.
2. Conclude that the difference between the two groups could be explained just by chance. *Accept* the null hypothesis, at least for now.

Note that you could be making an error either way!

# Hypothesis testing outcomes

Decision	Outcome if null hypothesis true	Outcome if null hypothesis false
Do not reject null hypothesis	Correct decision.	Type II error
Reject null hypothesis	Type I error	Correct decision

A *Type I error* occurs when there really is no difference, but you reject the null hypothesis.

Bad outcomes: you publish a paper, hard to undo the consequences!

We try to guard against these by setting a bound on how often we make such errors.

We determine in advance how big a probability of Type I error we can tolerate, called the *level* of the hypothesis test, usually written as  $\alpha$  (Greek letter alpha).

Often this is set at  $\alpha=0.05$ .

Sometimes people set level 0.01 but then it is hard to reject!

But you could also make a mistake if there really is a difference, and you fail to detect it. This is a *Type II error*.

A Type II error could happen if:

- The difference is so small it is just hard to detect.
- You didn't have a big enough sample size.  
(Uninformative null finding.)

You would need a big sample size if the difference is small, or there is so much unexplained variation that it swamps the difference in means.

Some people and some journals don't like to publish null findings. It is important, however, to publish them if the results are informative (based on enough data to rule out large differences).

For example, some people have claimed women are at higher risk of Alzheimer's disease. Careful analysis in large studies shows, however, that women have more AD only because they live longer. That is an important finding and needed to be published.

How do we carry out a hypothesis test?

We use statistical procedures to calculate the probability that we would see a difference or effect this big just by chance under the null hypothesis.

This probability is called the *P value*.

If the *P* value falls below the pre-determined level, we can reject the null hypothesis.

If the *P* value is above the pre-determined level, we cannot reject.

(If it is very close, some people might call it a “trend”.)

For our respiratory rate data, for example, the difference between the means is  $22 - 17 = 5$  breaths/min.

A statistical test finds the P value is  $< 0.001$ , so this is very unlikely (less than one chance in 1000) to have occurred just by chance.

We would *reject the null hypothesis* that men and women have the same respiratory rate at submaximal bicycle exercise and conclude that women in fact have a higher rate.

It's important to report not just that we found a difference, but how big it is, and the precision of our estimate.

So we could say:

Women had a faster respiratory rate under submaximal exercise than men ( $P < 0.001$ ), with women averaging 5 breaths more per minute (95% CI, 4.1 - 5.9 breaths per minute more than men).

## Interpretation of P value:

Sometimes people think a P value of 0.05 means: "Given the data we have, there is a 5% chance that there really is no difference."

The more accurate statement, however, is the converse: "Given that there really is no difference, the chance that we would get data as extreme as ours is only 5%."

The P value is the chance we calculate that we would see an outcome like this (or more extreme) if nothing at all were going on.

Most computers will print out the P value for you, so you don't have to look it up in tables.

We protect against Type I error by setting the level.

To keep Type II error small, we design study to have good *power*, or  $1 - [\text{Type II error}]$ .

If  $P[\text{Type II error}]$  is 0.10, then power is 0.90 or 90%.

A study with low power may fail to reject even if the null hypothesis is false (the dreaded “uninformative null”).

We want to design studies with good power. The next session will give you some explicit guidance on how to design studies with adequate power.

Four key factors affecting power are:

1. **Sample size:** The larger your sample, the greater your power. But, like the precision of a confidence interval, power only goes up as fast as  $\sqrt{n}$ .
2. **The difference in means:** It's easier to find big differences than small ones. Know what kind of difference you are looking for to plan a study.
3. **The variation of your measurements** also affects the study. If individuals vary a great deal within group, it will take a larger sample size to see the differences between groups.
4. **The level you require for the P value** affects power; if you make the level 0.01 instead of 0.05 it will be harder to reject and power will go down.

# Overview of the hypothesis testing process

State the null hypothesis

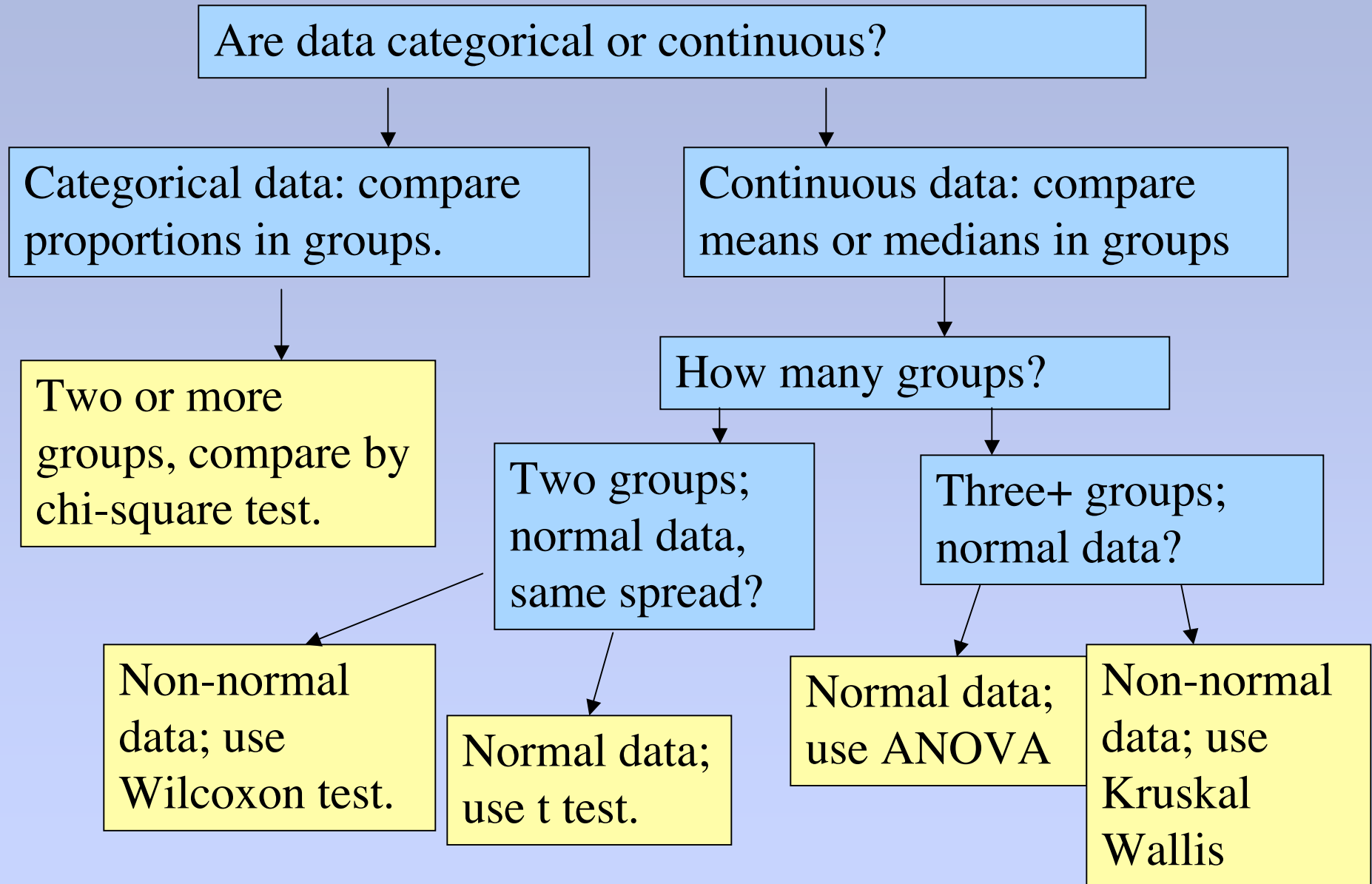
Look at the data and decide on an appropriate statistical test.

Compute the statistical test, look at P value.

P value  $< \alpha$ , reject null hypothesis. Report P value, estimated effect, and precision of estimate.

P value  $> \alpha$ , fail to reject null hypothesis. Report P value, power of study, and how big an effect you might have missed.

# Basic statistical procedures for comparing groups



Example 1. Comparing continuous measurements in two groups.

Many older people have complaints of cognitive problems, but do not meet the diagnostic criteria of Alzheimer's disease or other dementia. Some will go on to develop AD, some won't. Can we tell these two groups apart by findings on MRI?

Data on 12 people who went on to AD diagnosis and 11 who did not: normalized total entorhinal cortex volume (right + left), same for hippocampal volume.

Steps for hypothesis testing:

Step 1: Null hypothesis: the mean normalized volume is the same in both outcome groups.

Step 2: Look at data (see box plots, also stem-and-leaf, not shown.) Data look pretty normal for such small groups. Spread is very similar. OK to do t test.

Step 3: Carry out t-test.  $T=2.94$ ,  $p=0.008$ .

Step 4: Draw and report conclusions.

P value less than 0.05 so reject the null hypothesis. Those who went on to develop AD had lower entorhinal volume than those who did not.

Mean normalized entorhinal volume in those who developed AD was about 0.87, compared with 1.03 in those who did not.

Hippocampal volumes did not differ significantly. The sample sizes were small but the mean volumes were really very close, so it's at least a somewhat informative null. With these sample sizes, we'd have had about 80% power to detect a 1-sd difference in means between the two groups. (About half a volume-normalized unit of size.) The actual difference in means was almost zero.

Ref: Dickerson et al, *Neurobiol of Aging*, 2001.

Example 2. Population-based study of Alzheimer's disease.

A disease-free cohort of 642 people was followed about 4 years, then re-examined by neuropsychologist and neurologist. 57 of 362 women developed AD and 38 of 280 men.

Question: Was incidence of AD higher in women?

Hebert et al AJE 2000.

Step 1. State the null hypothesis. The proportion of men who develop AD is the same as the proportion of women.

Step 2. Look at data and determine appropriate statistical test. Proportions - we can do a chi-square test (but read on for details!)

Step 3. Carry out statistical procedure and get P value. The exact P value is 0.50. This means it's just like a coin-toss.

Step 4. P value is well over 0.05 so we can't reject the null hypothesis. Conclude there is no evident difference in proportions. Note: the actual analysis was "adjusted" for a complex sampling design, length of follow-up, and age of participant, got same result. Power would have been adequate to detect 16% vs 9% incidence.

## Some comments and cautions:

1. If the data aren't normal, the t-test may give wrong answers (either reject when it shouldn't or fail to reject when it should.) Two possible ways to fix: change the data (transform, e.g log or reciprocal) or change the test (Wilcoxon, other robust procedures.)
2. If you are comparing more than two groups, you may have lots of pair-wise comparisons. You need to take this into consideration both in planning the study and in doing the analysis. See "multiple comparisons procedures", e.g. Tukey studentized range.
3. It is easier to find differences in continuous data than in categorical data, given the same sample sizes. Don't turn continuous data into categorical unnecessarily.

More notes: Statistical procedures only quantify differences due to chance. Ask yourself what else could cause you to find or fail to find an effect.

1. Choice of participants? Clinic vs. community, consent to be in study vs. refuse to participate, urban vs. rural setting, etc. How generalizable is your finding? The people with MCI in example 1 came to a neurology clinic with complaints but were found not to have AD.

2. Other differences between the groups that you haven't taken into consideration? The women in Example 2 were older than the men, on average. Perhaps they also had lower education, or came back for follow-up later so had more time to develop AD.

3. Assumptions for procedure not met? Maybe your data are not normal or bell-shaped, or the variances of the two groups differ.