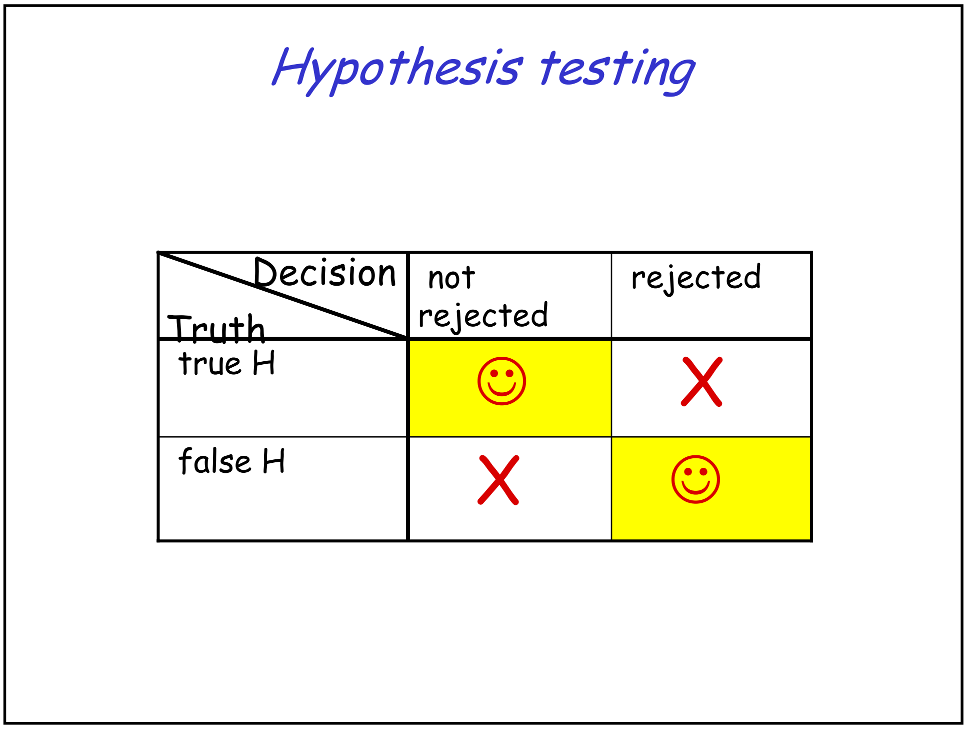
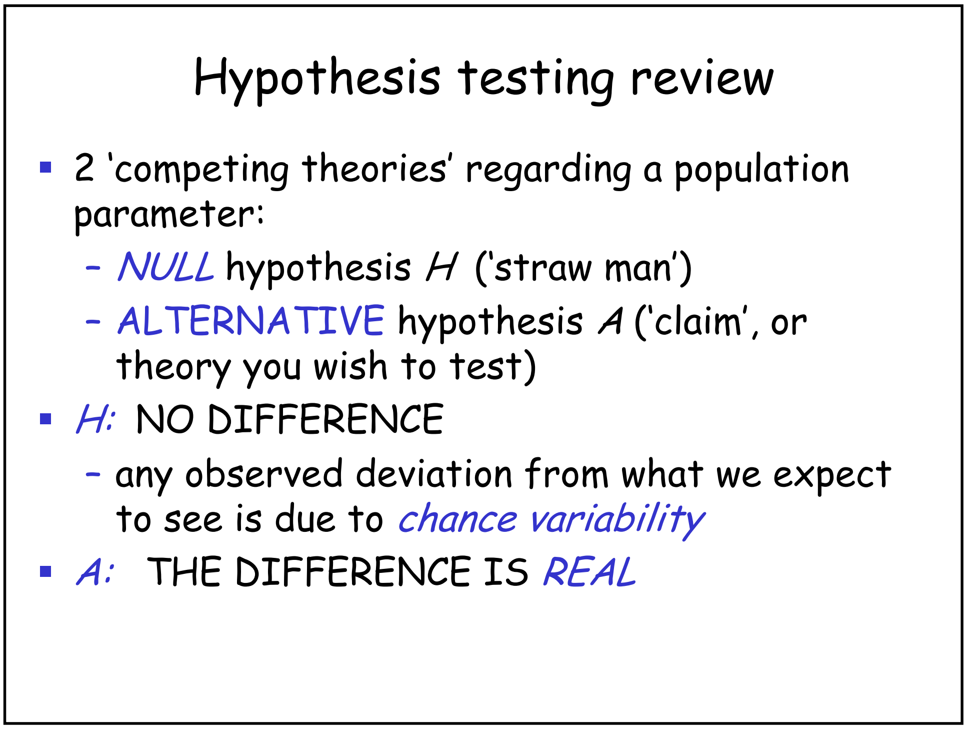


Hypothesis testing

Decision Truth	not rejected	rejected
true H	☺	✗
false H	✗	☺



Hypothesis testing review

- 2 'competing theories' regarding a population parameter:
  - *NULL* hypothesis *H* ('straw man')
  - ALTERNATIVE hypothesis *A* ('claim', or theory you wish to test)
- *H:* NO DIFFERENCE
  - any observed deviation from what we expect to see is due to *chance variability*
- *A:* THE DIFFERENCE IS *REAL*

Test statistic

- Measure how far the observed data are from what is expected *assuming the NULL H* by computing the value of a *test statistic* (TS) from the data
- The particular TS computed depends on the parameter
- For example, to test the population mean μ, the TS is the *sample mean* (or standardized sample mean)

Example

- An experiment is conducted to study the effect of exercise on the reduction of the cholesterol level in slightly obese patients considered to be at risk for heart attack. 80 patients are put on a specified exercise plan while maintaining a normal diet. At the end of 4 weeks the change in cholesterol level will be noted. It is thought that the program will reduce the average cholesterol reading by more than 25 points.
- Data:
  - sample mean = 27
  - sample SD = 18

Steps in hypothesis testing (I)

1. Identify the population parameter being tested

   Here, the parameter being tested is the population mean cholesterol reading μ

2. Formulate the NULL and ALT hypotheses

   H: $\mu = 25$ *(or* $\mu \leq 25$*)*

   A: $\mu > 25$

3. Compute the TS

   $t = (27 - 25)/(18/\sqrt{80}) = .99$

Hypothesis Truth vs. Decision

Decision Truth	not rejected	rejected
true H	☺ specificity	✗ Type I error (False +) α
false H	✗ Type II error (False -) β	☺ Power 1 - β; sensitivity

Some terminology

- The chance of rejecting a NULL which is *true* is α; this type of mistake is called a *Type I error* or *false positive*
- The chance of not rejecting a NULL which is false is β; this type of mistake is called a *Type II error* or a *false negative*
- In medical contexts, these quantities are referred to with other terminology:
  - The *specificity* of a test is the chance that the test result is negative given that the subject is negative; this is just $1 - \alpha$
  - The *sensitivity* of a test is the chance that the test result is positive given that the subject is positive; this is just $1 - \beta$, also called *power*

p-value

- Decide on whether or not to *reject* the NULL hypothesis *H* based on the chance of obtaining a TS *as or more extreme* (as far away from what we expected or even farther, in the direction of the ALT) than the one we got, *ASSUMING THE NULL IS TRUE*
- This chance is called the *observed significance level*, or *p-value*
- A TS with a p-value less than some pre-specified false positive *level* (or *size*) α is said to be ‘statistically significant’ at that level

p-value interpretation

- The interpretation of a p-value is a little tricky
- In particular, it does ***NOT*** tell us the probability that the NULL hypothesis is true
- The p-value represents the chance that we would see a difference as big as we saw (or bigger) *if* there were really nothing happening other than chance variability
- 'a single convenient number giving a measure of the degree of surprise which the experiment should cause a believer of the null hypothesis' (Hodges and Lehmann)

Steps in hypothesis testing (II)

4. Compute the p-value

   Here, $p = P(T_{79} > .99) = .16$

5. (Optional) *Decision Rule:* REJECT H if the p-value $\leq \alpha$

   (This is a type of argument by contradiction)

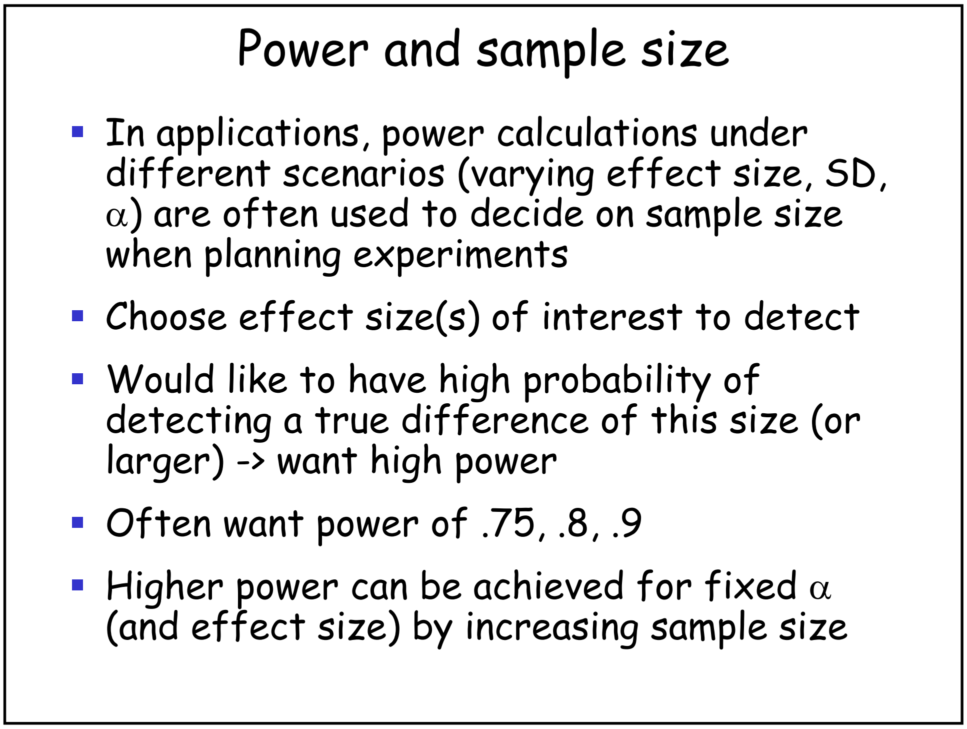
   A typical value of α is .05, but there's no law that it needs to be. If we use .05, the decision here will be DO NOT REJECT H

Power

- Not only do you want to have a low FALSE positive rate, but you would also like to have a high TRUE positive rate – that is, high *power*, the chance to find an effect (or difference) if it is really there
- Statistical tests will not be able to detect a true difference if the *sample size* is too small compared to the *effect size* of interest

Estimating power

- To compute or estimate power of a study, you need to be able to specify :
  - α level of the test,
  - the sample size n,
  - the effect size δ (or a specific alternative),
  - and the SD σ (or an estimate s)



Power and sample size

- In applications, power calculations under different scenarios (varying effect size, SD, α) are often used to decide on sample size when planning experiments
- Choose effect size(s) of interest to detect
- Would like to have high probability of detecting a true difference of this size (or larger) -> want high power
- Often want power of .75, .8, .9
- Higher power can be achieved for fixed α (and effect size) by increasing sample size