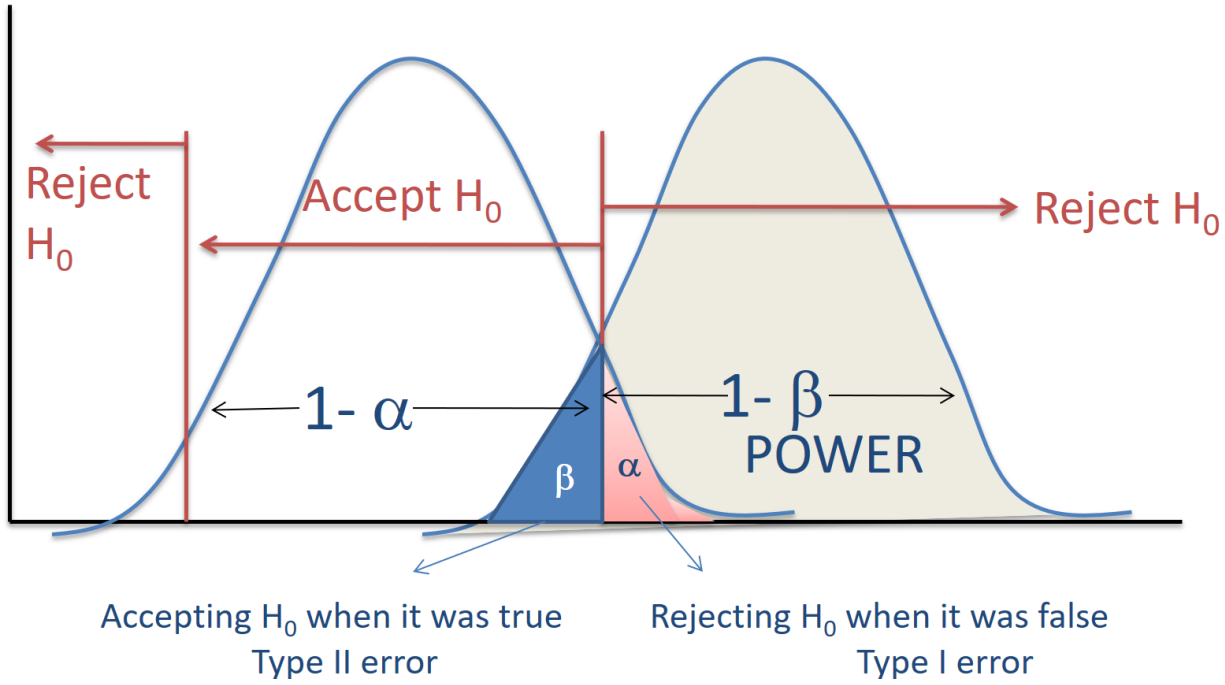
Hypothesis Testing

Conclusion from Statistical Test	Truth: H_0 **True**	Truth: H_0 **False**
Accept H_0	correct $1-\alpha$	β Type II Error false negative
Reject H_0	α Type I Error false positive	correct $1-\beta$ power

α = probability of a Type I error, usually fixed at 0.05
β = probability of a Type II error
α and β are inversely proportional
power = true positive, desired to be at least 0.8
fix α and make n large to increase power and keep β small



Multiplicity

- 3 or more treatment groups
- multiple outcomes from treatment
  - co-primary endpoints = study needs to demonstrate an effect on both clinical outcomes in order to conclude the drug is effective, e.g. improvement on cognitive function and daily function
- repeated measurements on the same outcome

Multiple Testing Errors

conducting m statistical tests at α=0.05 significance level

$(1-\alpha)^m$ probability of not making a type I error

$1-(1-\alpha)^m$ probability of making at least one type I error

Comparison-Wise Error Rate (CER) type I error rate for each comparison

Familywise Error Rate (FWER) type I error rate for entire group of comparisons

Multiple Comparison Strategies

$H_{01}: \mu_N = \mu_P$ $H_{02}: \mu_A = \mu_P$

$H_{A1}: \mu_N \neq \mu_P$ $H_{A2}: \mu_A = \mu_P$

All-or-Nothing

both of the null hypothesis need to be rejected

testing each individual co-primary endpoint doesn't increase FWER

e.g. if the first null hypothesis is true and the second is false,

P(false positive) = P(reject H_{01} & reject H_{02}) ≤ P(reject H_{01}) = 0.05

Either-Or

at least one of the null hypothesis need to be rejected

if one null hypothesis is rejected, the corresponding dose is the ideal dose

if both null hypothesis are rejected, the dose with the largest difference from the placebo is the ideal dose

if both null hypothesis are true, the FWER is increased

P(false positive) = P(reject H_{01} | reject H_{02}) ≤ P(reject H_{01}) + P(reject H_{02}) = 0.05 + 0.05 = 0.10

Composite Endpoints

multiple clinical outcomes combined into a single variable

e.g. treatment for cardiovascular disease uses composite endpoint of heart attack, stroke, or death

only one statistical test so FWER isn't increased

Adjusting for Multiplicity

Single-Step Procedures
test each null hypothesis independently of the other hypothesis
the order of testing isn't importance

Stepwise Procedures
pre-specified ordering: hypothesis are tested in a pre-determined order and doesn't depend on the observed data
data-driven ordering: testing sequence isn't determined before the study, order determined by p-values

Fisher's Least Significant Difference

H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$. All the group means are the same.
H_A: $\mu_1 \neq \mu_2 \neq \cdots \neq \mu_k$. At least one group's mean is different from the others.
if global ANOVA null hypothesis is rejected, perform pairwise comparisons without correcting any p-values

t-statistics use pooled standard deviation from all the groups and has $N - k$ degrees of freedom
N = total observations
k = number of groups

Bonferroni Correction

correct comparison-wise α-level to allow familywise comparison level to be controlled at 0.05

Method 1
correct comparison-wise α-level by the number of comparisons
compare p-value to adjusted α-level, $\frac{0.05}{m}$

Method 2
multiply observed p-values by number of comparisons
compare k$(p - value)$ to 0.05

$$p_{adjusted} = m \times p_{unadjusted}$$

Tukey-Kramer Correction

smaller p-values than Bonferroni
preferred method when looking at all pairwise comparisons

Dunnett's Test

focus is on pairwise comparisons with a control group
less conservative than Bonferroni corrections
smaller p-values than Tukey-Kramer for A vs P comparison

Holm Step-Down Algorithm

rank the null hypothesis by their corresponding p-values from smallest to largest

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

$$H_{(1)} \leq H_{(2)} \leq \cdots \leq H_{(m)}$$

Step 1	Reject $H_{(1)}$ if $p_{(1)} \leq \frac{0.05}{m}$
Step i (2 to $m-1$)	Reject $H_{(i)}$ if $p_{(i)} \leq \frac{0.05}{m-i+1}$
Step m	Reject $H_{(m)}$ if $p_{(m)} \leq 0.05$

If any null hypothesis is not rejected, stop and all remaining null hypothesis are also accepted.

Fixed-Sequence Procedure

natural order of the null hypothesis and testing order is fixed in advance
no adjusting for multiplicity as long as the proceeding tests had significant results
reject $H_{(i)}$ if $p_{(i)} \leq 0.05$
if any null hypothesis is not rejected, stop and all remaining null hypothesis are also accepted
FWER is controlled because a hypothesis is tested only if previous hypothesis are all rejected