Hypothesis Tests with the *t*-distribution

Some slides developed by Mine Çetinkaya-Rundel of OpenIntro
Translated from LaTeX to Google Slides by Curry W. Hilton of OpenIntro.

Recap: Hypothesis Tests

- Identify the research question, formalize in terms of parameters.
- We want to make a decision on whether we think H_0 or H_1 is correct.
- Find a statistics, compute the distribution of the statistic under the null.
- p-value: The probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true.
- If the p-value is low (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence reject H_0.
- If the p-value is high (higher than the significance level, α) then it is pretty likely to observe the data even if the null hypothesis were true, so we do not reject H_0.
- We never accept H_0 since we're not in the business of trying to prove it!

Example: Student Grades

- Last year, the average grade of students the first midterm of applied statistics was 68.7.
  This year, the first 20 students that were graded have an average grade of 63.7 with a standard deviation of 12.
  Is student performance different this year?

- H_o: Average student performance is the same as last year

- H_1: Average student performance is different than last year

Example

- Last year, the average grade of students the first midterm of applied statistics was 68.7.

  This year, the first 20 students that were graded have an average grade of 63.7 with a sample standard deviation of 12.

  Is student performance different this year?

- $H_o: \mu = 68.7$

- $H_1: \mu \neq 68.7$

Conditions

- *Independence*: We are told to assume that cases (rows) are independent.
- *Normality* : The distribution of the sample mean is nearly normal (CLT)
    - Sample size / skew: Sample size is not very small, distribution of grades does not appear extremely skewed.

Review: why do we need a large sample?

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal
- While this is a helpful special case, it’s inherently difficult to verify normality in small data sets

However, if we believe our original disstribution is normal, then

$$Z = \frac{\overline{X_n} - \mu}{SE = \sigma/\sqrt{n}} \sim N(0,1)$$

We do not know σ, and n is too small to get a reliable estimate using s.

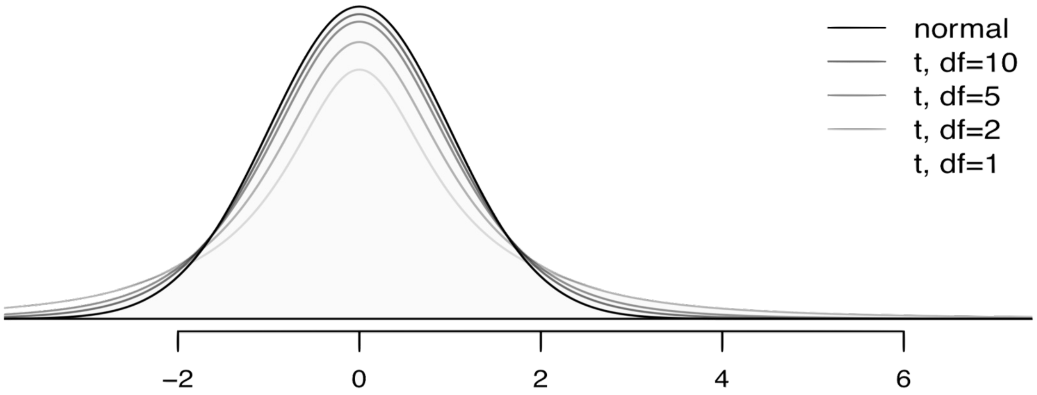
Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 30$) mean is the T statistic with $df = n - 1$

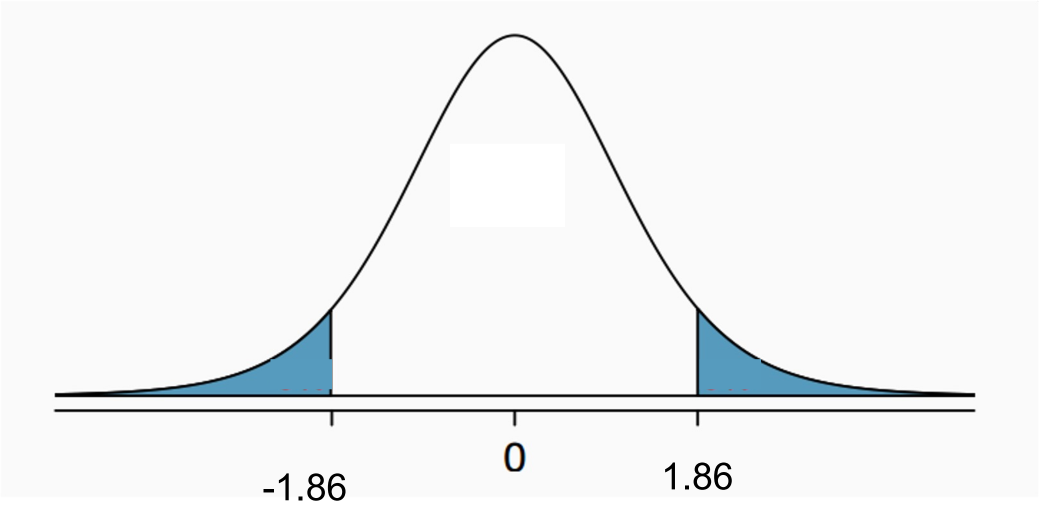
$$T_{df} = \frac{\overline{X_n} - \mu}{s/\sqrt{n}}$$

Under the null, $T_{df} \sim t_{n-1}$



Recap: Inference using the *t*-distribution

- In context:
- $n = 20, s = 12, \overline{x_n} = 63.7, \mu = 68.7$
- $T_{df} = \frac{\overline{x_n} - \mu}{s/\sqrt{n}} \sim t_{19}$



Paired observations

- Now assume that I have the grades for the midterm X and the
- final Y for applied statistics.

- I want to see if students did better or worse at the finals. My samples are "paired": The same student i has both a X_i and a Y_i

- I create the variable $X - Y$ and test if the mean is 0.

- This is called paired t-test.

Parameter and point estimate

- *Parameter of interest*: Average difference between the midterm and final scores of all students

$$\mu_{diff}$$

- *Point estimate*: Average difference between the midterm and final scores of sampled students

$$\bar{x}_{diff}$$

H_0 : Midterm and final scores for each student are on average the same: $\mu_{diff} = 0$

H_A : Midterm and final scores for each student are on average the same: $\mu_{diff} \neq 0$

$$T_{df} = \frac{point\ estimate\ \ - null\ value}{SE}, where\ df = n - 1$$

Nothing new here

- The analysis is no different than what we have done before
- We have data from one sample: differences.
- We are testing to see if the average difference is different than 0.

Two sample t-test

- Now assume that I want to compare the midterm grades of male and female students.
- I have graded 30 midterms from female students and 23 midterms from male students.
- Let X denote the grades of female students, Y denote the grades of male students.
- We assume X and Y have the **same** unknown variance.

Two sample t-test

- Now assume that I want to compare the midterm grades of male and female students.

- I have graded 30 midterms from female students and 20 midterms from male students.

- *Parameters of interest*: Average score of all female students minus average score of all male students.

- *Point estimates:* Average score of sampled female students minus average score of sampled male students

- $H_0: \mu_f - \mu_m = 0$ vs $H_1: \mu_f - \mu_m \neq 0$

Two sample t-test

- Now assume that I want to compare the midterm grades of male and female students.

- I have graded 30 midterms from female students and 20 midterms from male students.

- *Parameters of interest*: Average score of all female students minus average score of all male students.

- *Point estimates:* Average score of sampled female students minus average score of sampled male students

- $H_0: \mu_f - \mu_m = 0$ vs $H_1: \mu_f - \mu_m \neq 0$

Test statistics

Test statistic for inference on the difference of two small sample means

The test statistic for inference on the difference of two means where $\sigma_X = \sigma_y = \sigma$ is unknown is the *T* statistic.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

where

$$SE = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}\left(x_j - \bar{x}_2\right)^2}{n_1 + n_2 - 2}, \quad df = n_1 + n_2 - 2$$

Test statistics (cont.)

	M	F
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

In context...

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$SE = \sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \quad s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}\left(x_j - \bar{x}_2\right)^2}{n_1 + n_2 - 2}, \quad df = n_1 + n_2 - 2$$

$SE =$ 18.46, $T = 0.49$, $df = 51$

Find the p-value. What is the conclusion?

Meaning of significance

Suppose :

- $\bar{X}_n = 50, \; s = 2$
- $H_0 : \mu \leq 49.5, \; H_1 : \mu > 49.5.$
- Compute the p-value for $n = 100$ and $n = 10000?$

Meaning of significance

Suppose :

- $\bar{X}_n = 50, s = 2$
- $H_0 : \mu \leq 49.9,\ H_1 : \mu > 49.9$.
- Compute the p-value for $n = 100$ and for $n = 10000$?

Practical vs statistical significance

- Real differences between the point estimate and null value are easier to detect with larger samples.

- However, very large samples will result in statistical significance even for tiny differences between the sample mean and the null value (effect size), even when the difference is not practically significant.

- This is especially important to research: if we conduct a study, we want to focus on finding meaningful results (we want observed differences to be real but also large enough to matter)

Review of Hypothesis testing

- Hypothesis tests allow us to answer simple “yes-or-no” questions, such as:

- Is smoking independent from cardiovascular disease?

- Does the average blood pressure among mice in the treatment group equal the average blood pressure among mice in the control group?

Review of Hypothesis testing

- Hypothesis tests allow us to answer simple “yes-or-no” questions, such as:

  - Is smoking independent from cardiovascular disease?
  - Does the average blood pressure among mice in the treatment group equal the average blood pressure among mice in the control group?

1. Define the null and alternative hypotheses.
2. Construct the test statistic.
3. Compute the p-value.
4. Decide whether to reject the null hypothesis.

1. Define the null and alternative hypotheses

We divide the world into null and alternative hypotheses.
The null hypothesis, H_0, is the default state of belief about the world.

For instance:

1. Smoking is independent of cardiovascular disease.
2. There is no difference in the average blood pressures.

The alternative hypothesis, H_1, represents the complement of the null. For instance:

1. Smoking and cardiovascular disease are not independent.
2. There is a difference in the average blood pressures.

2. Construct the test statistic

The test statistic summarizes the extent to which our data are (in) consistent with H_0.

Let μ_t/μ_c. respectively denote the average blood pressure for the nt/nc mice in the treatment and control groups.

To test $H_0 : \mu_t = \mu_c$, we use a two-sample statistic:

$$T_{df} = \frac{\overline{x_t} - \overline{x_c}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_{\mathrm{t}})^2 + \sum_{j=1}^{n_2}\left(x_j - \bar{x}_{\mathrm{c}}\right)^2}{n_1 + n_2 - 2}$$

$$df = n_1 + n_2 - 2$$

3. Compute the p-value

The test statistic summarizes the extent to which our data are (in) consistent with H_0.

Let μ_t/μ_c. respectively denote the average blood pressure for the n_t $/n_c$ mice in the treatment and control groups.

To test H_0 : $\mu_t = \mu_c$, we use a two-sample statistic

3. Compute the p-value

- The p-value is the probability of observing a test statistic at least as extreme as the observed statistic, under the assumption that H_0 is true.

- A small p-value provides evidence against H_0.

- Suppose we compute $T = 2.55$ for our test of $H_0: \mu_{\mathrm{t}} = \mu_{\mathrm{c}}$ with 10 mice in the control group and 10 mice in the treatment group.

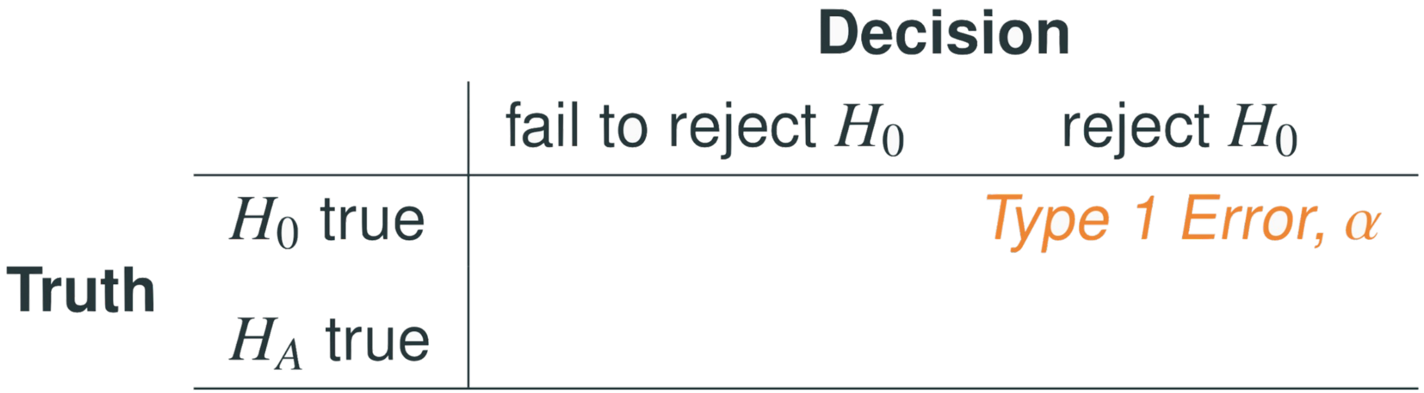
- Under H_0, $T \sim t_{m+n-2}$ for a two-sample t-statistic.

- The p-value is 0.02 because, if H_0 is true, we would only see $|T|$ this large 2% of the time.

4. Decide whether to reject the null hypothesis

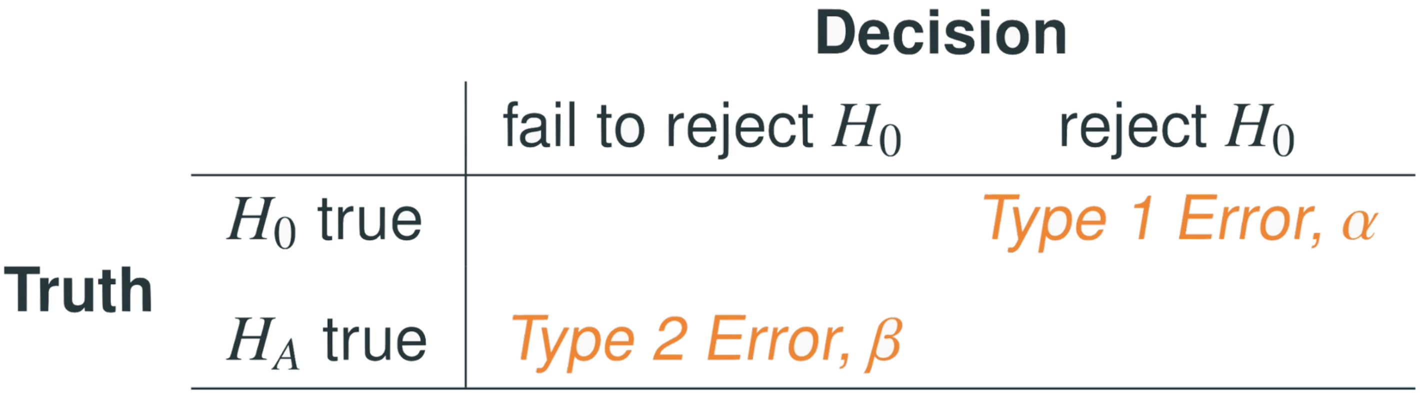
- A small p-value indicates that such a large value of the test statistic is unlikely to occur under H_0.
- So, a small p-value provides evidence against H_0.
- If the p-value is sufficiently small, then we will want to reject H_0 (and, therefore, make a potential “discovery”).
- How small is small enough?

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		



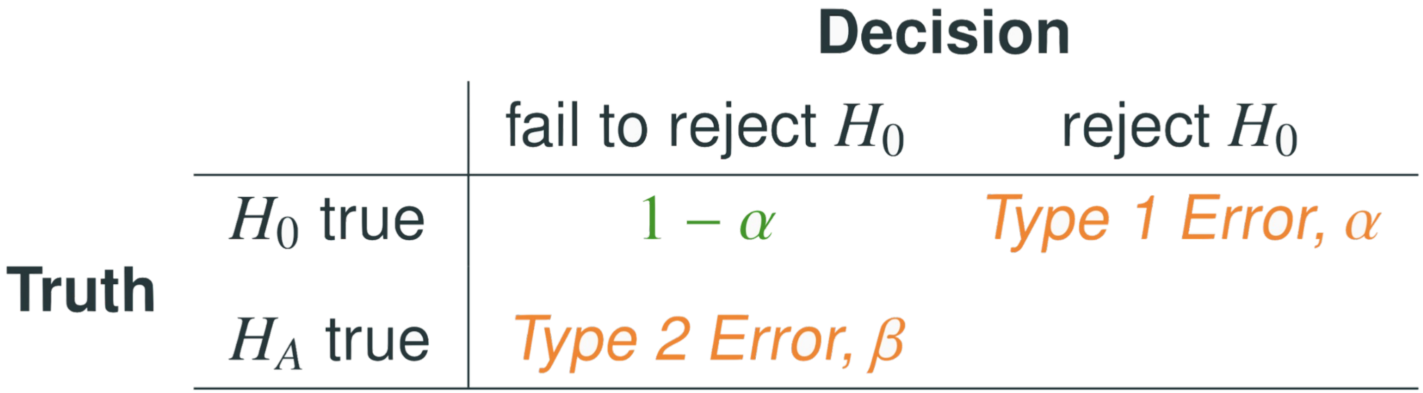
		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error*, α
	H_A true		

- Type 1 error is rejecting *H₀* when you shouldn't have, and the probability of doing so is *α* (significance level)



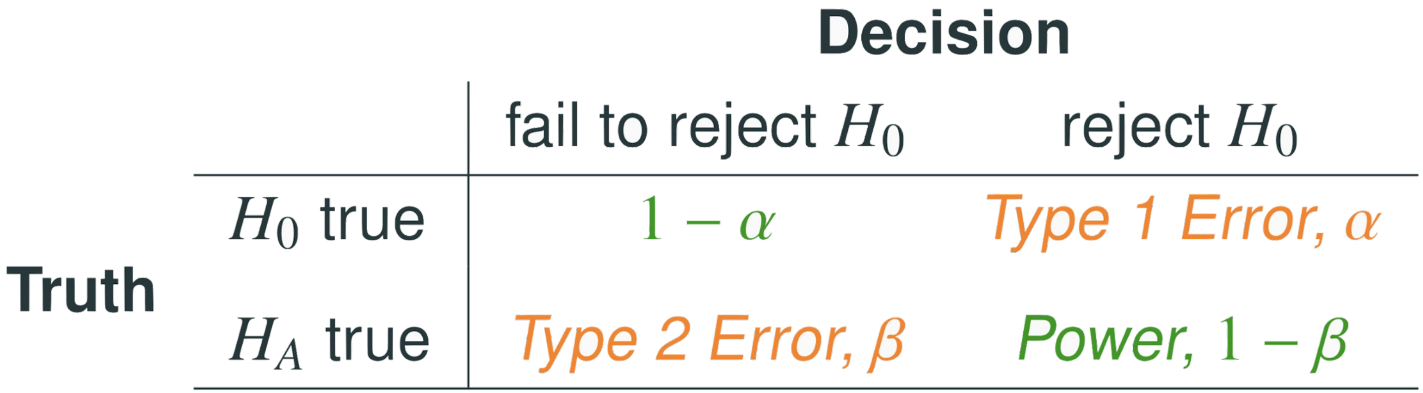
		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	

- Type 1 error is rejecting H_0 when you shouldn’t have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$

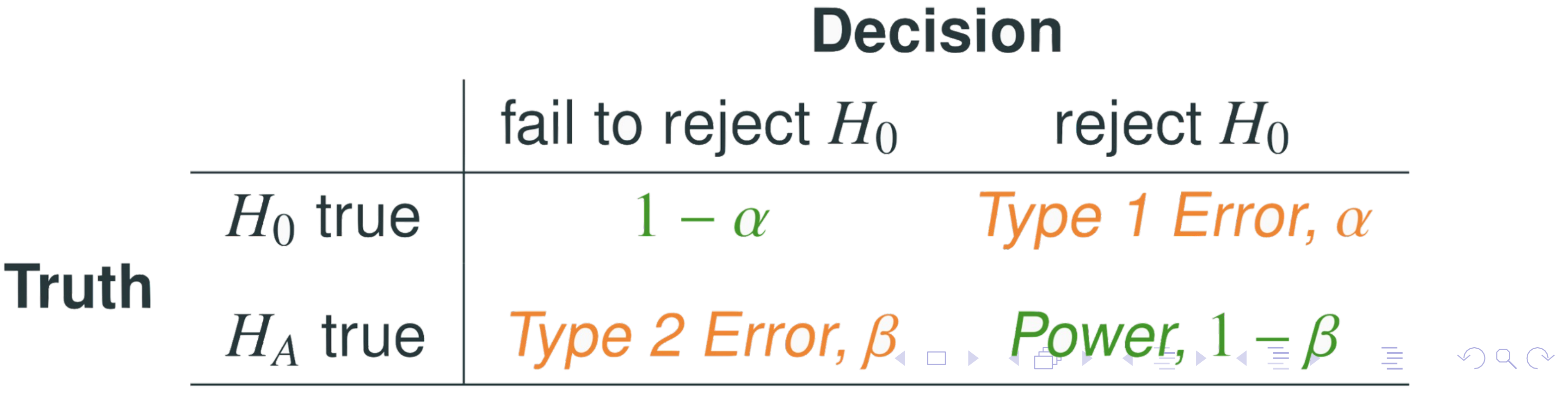


		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1 - \alpha$	*Type 1 Error*, α
	H_A true	*Type 2 Error*, β	*Power*, $1 - \beta$

- Type 1 error is rejecting H_0 when you shouldn't have, and the probability of doing so is α (significance level)
- Type 2 error is failing to reject H_0 when you should have, and the probability of doing so is β (a little more complicated to calculate)
- Power of a test is the probability of correctly rejecting H_0, and the probability of doing so is $1 - \beta$
- In hypothesis testing, we want to keep α and β low, but there are inherent trade-offs

Type I error

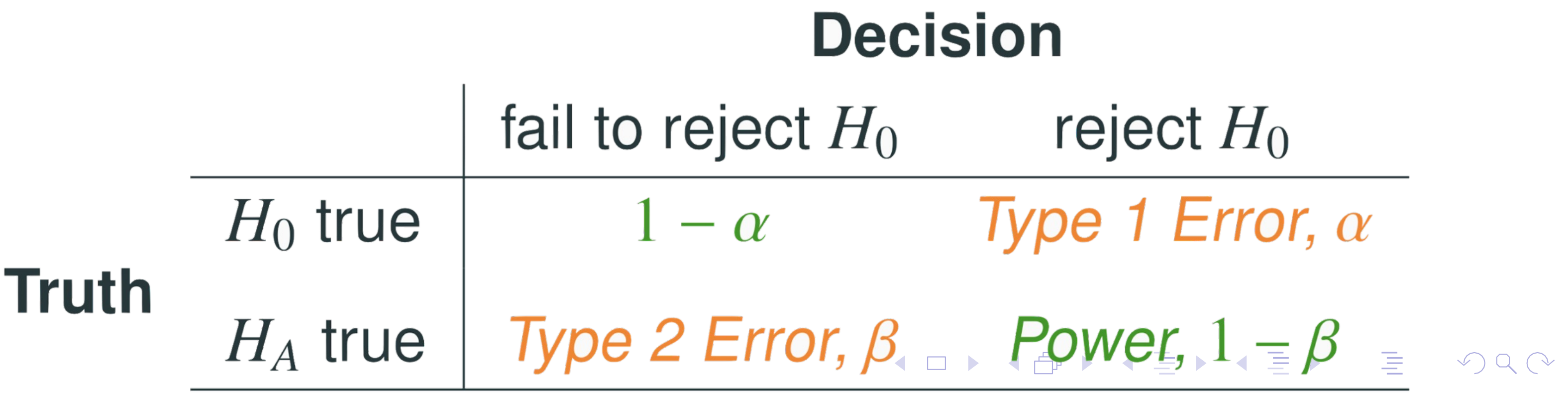
- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a significance level of 0.05, $\alpha = 0.05$.
- Type I error rate:
  - P (Type I error) = P (Reject $H_0 | H_0$ is true) $\leq \alpha$
- Increasing α increases the Type I error rate.
- When we select α we control for the tolerance we have for type I errors.



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error, α*
	H_A true	*Type 2 Error, β*	*Power, $1-\beta$*

Type II error

- If the alternative hypothesis is actually true, what is the chance that we make a Type II Error, i.e. we fail to reject the null hypothesis even when we should reject it?
- The answer is not obvious, but
    - If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
    - If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- The probability of correctly rejecting the null is the **power** of the test.



		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	$1-\alpha$	*Type 1 Error,* α
	H_A true	*Type 2 Error,* β	*Power,* $1-\beta$

Multiple testing

- Now assume we want to test multiple hypotheses
$$H_{01}, \ldots, H_{0m}$$

- If we reject all null hypotheses for which the p-value falls below 0.05, then how many Type I errors will we make?

A thought experiment

- Suppose that we flip a fair coin ten times, and we wish to test H_0: the coin is fair.

- We'll probably get approximately the same number of heads and tails.

- The p-value probably won't be small. We do not reject H_0.

- But what if we flip 1,024 fair coins ten times each?

Multiple hypotheses testing

- Suppose we test $H_{01}, \ldots, H_{0m}$, all of which are true, and reject any null hypothesis with a p-value below 0.05.
- Then we expect to falsely reject approximately $0.05 \times m$ null hypotheses.
- If $m = 10{,}000$, then we expect to falsely reject 500 null hypotheses by chance!
- That's a lot of Type I errors, i.e. false discoveries/false positives!
- Example: Genome-wide association studies.

Family-wise error rate

The probability of making at least one type 1 error

	Fail to reject H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

FWER =1 − P (do not falsely reject any null hypothesis)=

$$= 1 - P(\bigcap_{j=1}^{m} \text{do not falsely reject } H_{0j})$$

Family-wise error rate

	Fail to reject H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

FWER =1 $-$ P (do not falsely reject any null hypothesis)=

$= 1 \ - P(\bigcap_{j=1}^{m}$ do not falsely reject $H_{0j})$

If the tests are independent and all H_{0j} are true then

FWER = 1 - $\sum_{j=1}^{m} P(do\ not\ falsely\ reject\ H_{0j}) = 1 - (1 - a)^m$

Multiple hypotheses testing

FWER = P (falsely reject at least one null hypothesis) =

$$P\left(\cup_{j=1}^{m} A_j\right) \leq \sum_{j=1}^{m} P\left(A_j\right)$$

where A_j is the event that we falsely reject the $j-th$ null hypothesis. If we only reject hypotheses when the p-value is

$$\text{FWER} \leq \sum_{j=1}^{m} P(A_j) \leq \sum_{j=1}^{m} \frac{\alpha}{m} = \alpha$$

because $P\ (A_j\ ) \leq\ \alpha/m$

This is the Bonferroni Correction: to control FWER at level α, reject any null hypothesis with p-value below α/m

Example: Video Games and ADHD

- If we reject H_{0j} if the p-value is less than $\alpha = 0.05$, we will conclude that TV, VG-C, VG-I significantly affect YAS, VG-C and VG-I significantly affect Inattention and ADHD.
- However, we have tested multiple hypotheses, so the FWER is greater than 0.05 .

Example: Video Games and ADHD

vs	Internet	TV	VG-C	VG-I
Young’s Addiction Scale	0.804	**0.040**	**< 0.001**	**<0.001**
Conner’s Scale: Oppositional	0.096	0.397	0.917	0.826
Conner’s Scale: Inattention	0.289	0.311	**0.001**	**<0.001**
Conner’s Scale: Hyperactivity	0.901	0.397	0.800	0.142
Conner’s Scale: ADHD	0.115	0.343	**0.018**	**0.020**

- If we reject H_{0j} if the p-value is less than $\alpha = 0.05$, we will conclude that TV, VG-C, VG-I significantly affect YAS, VG-C and VG-I significantly affect Inattention and ADHD.
- However, we have tested multiple hypotheses, so the FWER is greater than 0.05 .

Example: Video Games and ADHD

vs	Internet	TV	VG-C	VG-I
Young’s Addiction Scale	0.804	0.040	< **0.001**	**<0.001**
Conner’s Scale: Oppositional	0.096	0.397	0.917	0.826
Conner’s Scale: Inattention	0.289	0.311	**0.001**	**<0.001**
Conner’s Scale: Hyperactivity	0.901	0.397	0.800	0.142
Conner’s Scale: ADHD	0.115	0.343	**0.018**	**0.020**

- Using the Bonferroni correction we will reject p-values less than $\alpha/20 = 0.0025$.

- If we reject H_{0j} if the p-value is less than 0.0025, we will conclude that VG-C, VG-I significantly affect YAS, VG-C and VG-I significantly affect Inattention.

Holm's method for controling FWER

- Compute p-values, $p_1, \ldots, p_m$ for the m null hypotheses $H_{01}, \ldots, H_{0m}$.
- Order the m p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
- Define

$$P_L = min_j \, P_{(j)} < \frac{\alpha}{m+1-j}$$

- Reject all null hypotheses H_{0j} for which $p_j < p_{(L)}$
- Holm's method controls the FWER at level α.

Bonferroni vs Holm

Consider m = 5

p-values

$p_1 = 0.006$, $p_2 = 0.918$, $p_3 = 0.012$, $p_4 = 0.601$, $p_5 = 0.756$.

Then

$p_{(1)} = 0.006$, $p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$, $p_{(5)} = 0.918$.

- Bonferroni?
- Bonferroni - Holm?

Bonferroni vs Holm

- Bonferroni is simple ... reject any null hypothesis with a p-value below α/m.
- Holm is slightly more complicated, but it will lead to more rejections while controlling FWER!!
- Holm is a better choice

The False Discovery Rate

	Fail to reject H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

- The FWER rate focuses on controlling $P(V > 1)$, i.e., the probability of falsely rejecting any null hypothesis.
- This is a tough ask when m is large! It will cause us to be super conservative (i.e. to very rarely reject).
- Instead, we can control the false discovery rate:
  - FDR $= E(V/R)$

The False Discovery Rate

$$\text{FDR} = E\left(\frac{V}{R}\right) = E\left(\frac{\text{number of false rejections}}{\text{total number of rejections}}\right)$$

- A scientist conducts a hypothesis test on each of m = 20, 000
  - drug candidates.
- She wants to identify a smaller set of promising candidates to investigate further.
- She wants reassurance that this smaller set is really “promising”, i.e. not too many falsely rejected H0’s.
- FWER controls P(at least one false rejection).
- FDR controls the fraction of candidates in the smaller set that are really false rejections. This is what she needs!

Benjamini-Hochberg procedure for controlling FDR

1. Specify q, the level at which to control the FDR.
2. Compute p-values $p_1, \dots, p_m$ for the null hypotheses $H_{01}, \dots, H_{0m}$.
3. Order the p-values so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
4. Define $L = max_j : p_{(j)} < qj/m$.
5. Reject all null hypotheses H_{0j} for which $p_{(j)} \leq p_{(L)}$.

Then, FDR $\leq q$.

FWER vs FDR

Consider m = 5

p-values

$p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$

Then

$p_{(1)} = 0.006, p_{(2)} = 0.012, p_{(3)} = 0.601, p_{(4)} = 0.756, p_{(5)} = 0.918.$

- Bonferroni?
- Bonferroni-Holm?
- Benjamini-Hochberg?