Statistical tests, power and type I error

(Week 04 lecture notes)

Wei Q. Deng

Department of Statistical Sciences

July 25th 2018

Topics covered in the previous lecture

- Concepts of a statistical test
  - Structure of a hypothesis test
  - Test statistics
  - Rejection region and P-values
  - One-sided vs. two-sided test
- Examples of statistical tests
  - Statistical tests of μ (normal, t, and binomial)
  - Statistical tests to compare two samples
    - test for equality of means (student's t-tests)
    - test for equality of variances (F-test)

A statistical test

- Parameter of interest
- Null Hypothesis about a particular parameter value
- Alternative Hypothesis
- Test statistics with a known distribution under the null hypothesis
- Significance level
- P-value

The overall goal of a statistical test is to determine whether there is sufficient evidence (from data) to "reject" the null hypothesis (or what we believed to be the truth).

Review: The Null and Alternative Hypothesis

The null hypothesis, H_o

- States the assumption (numerical) to be tested
- Begin with the assumption that the null hypothesis is TRUE.
- Always contains the "=" sign

The alternative hypothesis, H_a

- Is the opposite of the null hypothesis
- Challenges the status H_o
- Never contains just the "=" sign
- Is generally the hypothesis that is believed to be true by the researcher

Practicing writing the null and alternative hypothesis

- Identify the parameter of interest (e.g. μ, p)
- Identify an established value for the null hypothesis
  - what is the convention
  - what is the prior belief
  - what is the expert opinion
- Decide whether the alternative hypothesis should be
  - two-sided (disagree with the prior belief without any subjective opinion on directionality)
  - one-sided (a clear preference for the parameter to be greater or less than some value)

Try the following for yourself

Define the parameter of interest and state the null and alternative hypotheses.

- Example 1: A new design for the braking system on a certain type of car has been proposed. For the current system, the true average braking distance at 40 mph under specified conditions is known to be 120 ft. It is proposed that the new design be implemented only if sample data strongly indicates a reduction in true average braking distance for the new design.

Try the following for yourself

- Example 2: A researcher thinks that if knee surgery patients go to physical therapy twice a week (instead of 3 times), their recovery period will be longer. Average recovery times for knee surgery patients is 8.2 weeks.

Try the following for yourself

- Example 3: A sample of 200 people has a mean age of 21 with a population standard deviation (σ) of 5. Test the hypothesis that the population mean age is 18.9 at $\alpha = 0.05$.

Test statistic

- The test can only be performed using **an estimator** of the parameter since we do not have knowledge of the parameter.
- We often use an estimator that has a **known sampling distribution**:
  - $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ either exactly or by CLT
  - $\hat{p} \sim \mathcal{N}(p, (1-p)p/n)$ by CLT
- This way we can compare whether the **observed test statistic** (t_{obs}) is more extreme than under the null hypothesis $\mu = \mu_o$ (one sample) or $\mu_1 - \mu_2 = \delta_o$
  - for one-sample z-test: $t_{\text{obs}} = \frac{\bar{x}-\mu_o}{\sigma/\sqrt{n}}$
  - for one-sample t-test: $t_{\text{obs}} = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$
  - for two-sample independent t-test: $t_{\text{obs}} = \frac{\bar{x}_1-\bar{x}_2-\delta_o}{s_p/\sqrt{1/n_1+1/n_2}}$
  - for two-sample paired t-test: $t_{\text{obs}} = \frac{\bar{d}-\delta_o}{s_D/\sqrt{n}}$ where $d_i = x_i - x_i'$ and $s_D^2 = \frac{\sum_{i=1}^{n}(d_i-\bar{d})^2}{n-1}$.

P-value

The probability of observing something more extreme than at the expected significance level α under the null hypothesis is the **P-value**.

- a conditional probability that the observed test statistic is more extreme than would be expected under the null hypothesis.
- Two-sided test:
  p-value $= \mathrm{P}(T > t_{\text{obs}} | \mu = \mu_o) + \mathrm{P}(T < -t_{\text{obs}} | \mu = \mu_o)$
- One-sided test: p-value $= \mathrm{P}(T > t_{\text{obs}} | \mu = \mu_o)$
- One-sided test: p-value $= \mathrm{P}(T < t_{\text{obs}} | \mu = \mu_o)$

Significance level

A significance level (α) is an established level of risk for making a mistake under the null

- Two-sided test: $\alpha = \mathsf{P}(T > \Phi^{1-\alpha/2} | \mu = \mu_o) + \mathsf{P}(T < \Phi^{\alpha/2} | \mu = \mu_o)$
- One-sided test: $\alpha = \mathsf{P}(T > \Phi^{1-\alpha} | \mu = \mu_o)$
- One-sided test: $\alpha = \mathsf{P}(T < \Phi^{\alpha} | \mu = \mu_o)$

A summary of p-value comparisons

- A few conventions regarding rejecting:
  - If P-value$<\alpha$, reject H_o.
  - If P-value$>\alpha$, do not or fail to reject H_o.
- Smaller P-value $\rightarrow$ less likely H_o is to be true
  - p-value$<0.001 \rightarrow$ very strong evidence against H_o
  - $0.001<$ p-value $<0.01 \rightarrow$ strong evidence against H_o.
  - $0.01<$ p-value $<0.05 \rightarrow$ moderate evidence against H_o
  - $0.05<$ p-value $<0.1 \rightarrow$ weak evidence against H_o
  - p-value$>0.1 \rightarrow$ no evidence against H_o

P-value and CI should agree about statistical significance

- You can use either P values or confidence intervals to determine whether your results are statistically significant.
- The confidence level is equivalent to $1 - \alpha$ level. So, if your significance level is 0.05, the corresponding confidence level is 95%.
  - If the P value is less than your significance (α) level, the hypothesis test is statistically significant.
  - If the CI does not contain the null hypothesis value (H_o value), the results are statistically significant.
  - If P-value$< \alpha$, the CI will not contain the null hypothesis value.

P-value and CI should agree about statistical significance

To understand why the results always agree, let's recall how both the significance level and confidence level work.

- The significance level defines the distance the sample mean must be from H_o to be considered statistically significant.
- The confidence level defines the distance for how close the confidence limits are to sample mean.
- Both the significance level and the confidence level define a distance from a limit to a mean. Guess what? The distances in both cases are exactly the same!

Three ways to make decisions

Let t_{obs} denote the test statistic T with the observed value plugged in under the null hypothesis $\theta = \theta_o$.

- p-value $= \mathsf{P}(|T| > t_{\text{obs}})$
  - if p-value $< \alpha$, reject
  - o.w. fail to reject
- Compare $|t_{\text{obs}}|$ to $\Phi^{-1}(1 - \alpha/2)$ under the distribution of test statistics under the null
  - if $|t_{\text{obs}}| > \Phi^{-1}(1 - \alpha/2)$ reject
  - o.w. fail to reject
- Compare θ_o to the 95% or 99% confidence interval for θ
  - if θ_o is outside of the CI then reject
  - o.w. fail to reject

Topics covered in this lecture

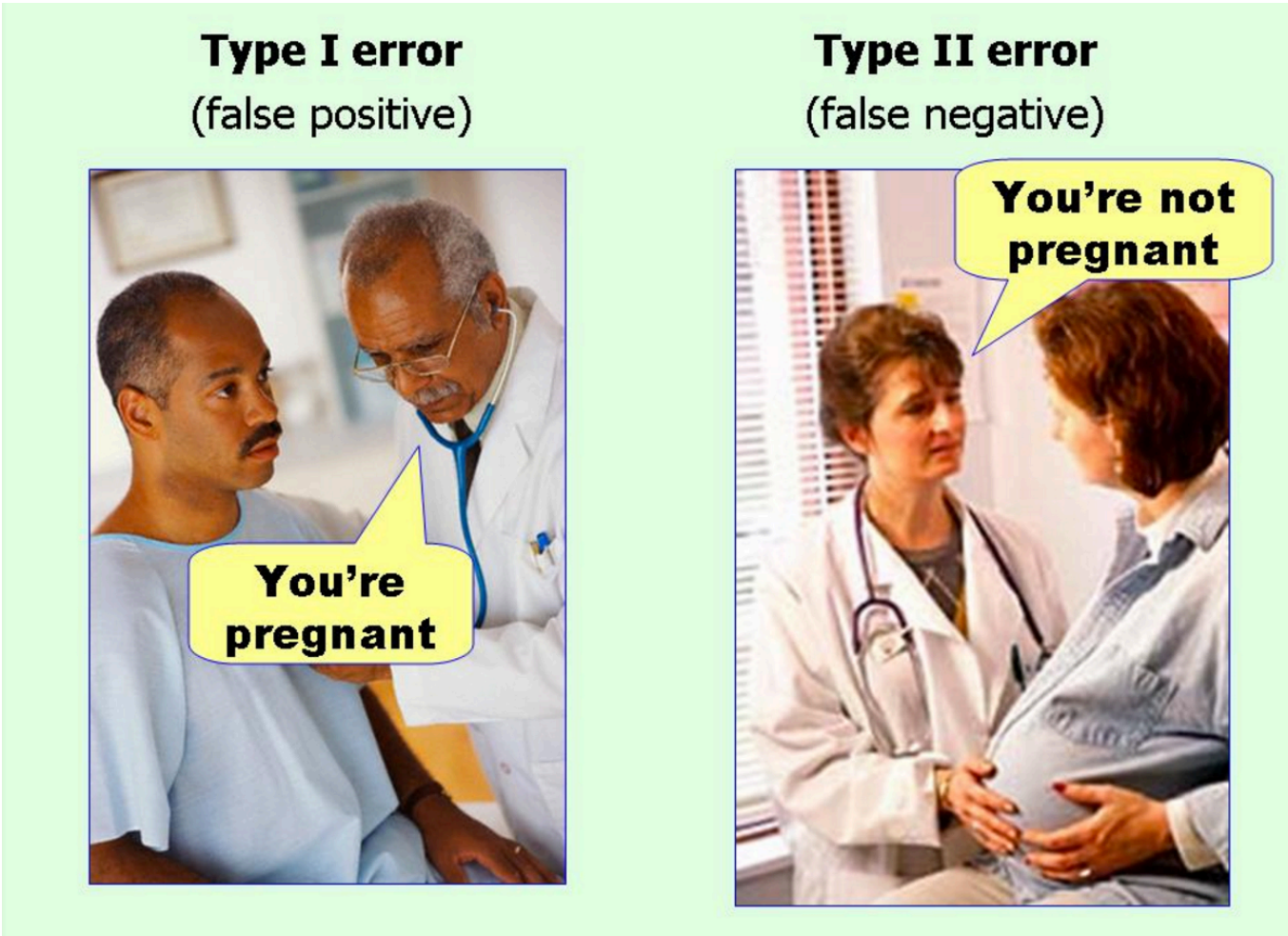
- Good and bad statistical tests: assessing errors of a test
  - Type I error
  - Type II error
- Statistical power
- Test for equality of means in more than two samples

Statistical evidence to inform decisions

		Reality/ Truth	
		Null is True H0: T	**Null is False H0: F**
Decision	**Reject Null**	Type I error "False Positive"	Correct Decision
	Fail to Reject Null	Correct Decision	Type II error "False Negative"

- Type I error = P(reject $H_o | H_o$ is true).
- Type II error = P(fail to rejct $H_o | H_o$ is false)
- Significance level α: maximum allowable probability of making type I error (usually 5% or 1%)

Error types



https://effectsizefaq.com/index/faqs/page/3/

Type I error

- Clearly, as long as $\alpha > 0$, there is α probability that we might reject the null hypothesis when it is true.
- Meaning even if the null hypothesis were true but we might reject it anyways because of the randomness in T
- To make sure we do not make this mistake, we require that p-value $< \alpha$.
- Let's see some examples of how to empirically assess type I error rates

Type I error simulation setup

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 5)$, where $n = 100$. We want to test the null hypothesis that the true mean $\mu = 0$ with a two-sided test (i.e. the alternative is $\mu \neq 0$).

- Step 1: Simulate $x_1, \ldots, x_{500}$ according to $\mathcal{N}(0, 5)$.
- Step 2: Perform one sample z-test ($\bar{X} \sim \mathcal{N}(\mu, 5/500)$) and obtain a p-value.
- Repeat Step 1 and 2, $B = 1000$ times, giving a vector of p-values $p_1, \ldots, p_{1000}$.
- What is your conclusion for each of the p-value?
- How many of those p-values do you expect to be less than $\alpha = 0.05$ or $\alpha = 0.01$?

Type I error simulation R

```
n = 500; mu_o = 0; sigma = sqrt(5); B = 1000
pval <- NA
i = 1;
while(i <= B){
sample_cases <- rnorm(n, mu_o, sigma)
pval[i] <- 2*min(pnorm((mean(sample_cases) - mu_o)/(sqrt(sigma^2/n)), lower=F),
                 pnorm((mean(sample_cases) - mu_o) / (sqrt(sigma^2/n)), lower=T))
i = i + 1
}
sum(pval < 0.05)/B
```

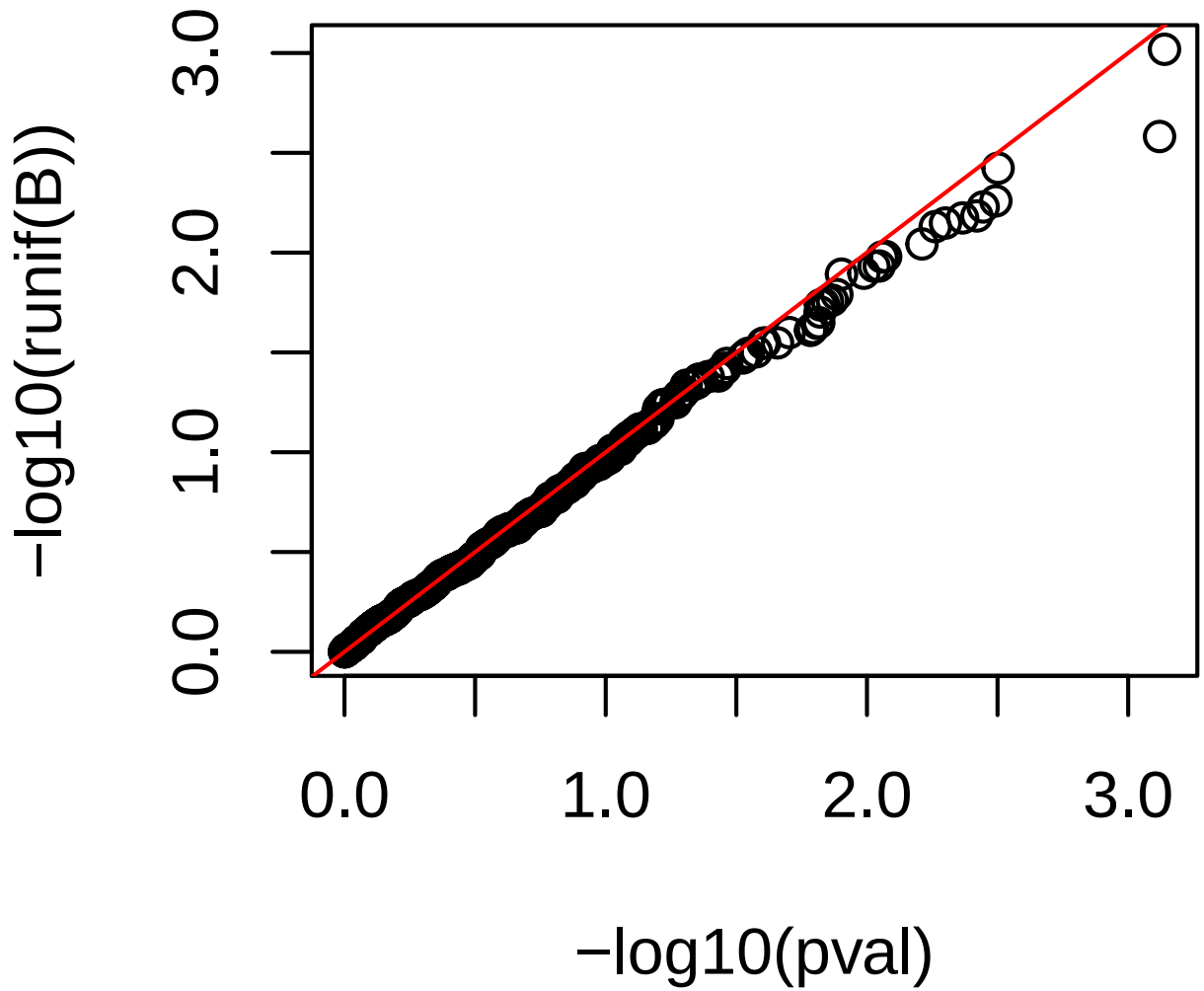
```
## [1] 0.051
```

```
sum(pval < 0.01)/B
```

```
## [1] 0.014
```

Type I error simulation R

```
qqplot(-log10(pval), -log10(runif(B))); abline(0,1,col=2)
```



Type I error simulation Python

```
import numpy as np
from scipy import stats
import random
n=500
mu_o=0
B=1000
pval=[]
for j in range(0, B):
  sample_cases = np.random.normal(mu_o, 5**0.5, n)
  z_scores = (sample_cases.mean())/(5.0/n)**(0.5)
  pval.append((1-stats.norm.cdf(abs(z_scores)))*2)
print(sum(i < 0.05 for i in pval))
```

```
## 53
```

```
print(sum(i < 0.01 for i in pval))
```

```
## 8
```

Type I error summary

- The above indicates the number of false positives is roughly as expected at α.
- When the type I error of a test is systematically higher than α, we call the test to have an **inflated type I error rate**
- When the type I error of a test is systematically lower than α, we call the test to be **conservative**
- Clearly we would like to use tests that have well-controlled type I error (close to α)

Type II error

Suppose the null hypothesis is $\theta = \theta_o$:

- Type II error captures the inability to reject a false null hypothsis, usually denoted by β.

$$\beta = \mathrm{P}(H_o \text{ not rejected} | H_o \text{ is false})$$

- Note that β is specific to the choice of θ under the alternative ($\theta' \neq \theta_o$):

$$\beta = \mathrm{P}(H_o \text{ not rejected} | \theta = \theta')$$

- So ideally, the smaller the β is for a particular μ', the better the test (always compare two tests for the same μ' and the same α).

An example of Type II error under the z-test

The rejection region of a one sample test

- under normality ($X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$)
- with σ^2 known
- at significance level α
- for a two-sided alternative

is $\bar{X} > \mu_o + \Phi^{-1}(1-\alpha/2)\sigma/n$ and $\bar{X} < \mu_o - \Phi^{-1}(1-\alpha/2)\sigma/n$\$.

An example of Type II error under the z-test

The type II error for a particular $\mu' \neq \mu_o$ is:

$$\beta(\theta') = \mathrm{P}(\mu_o - \Phi^{-1}(1-\alpha/2)\sigma/n < \bar{X} < \mu_o + \Phi^{-1}(1-\alpha/2)\sigma/n | \theta = \theta') \tag{1}$$

$$= \mathrm{P}\left(\frac{\mu_o - \mu'}{\sigma/n} - \Phi^{-1}(1-\alpha/2) < \frac{\bar{X} - \mu'}{\sigma/n} < \frac{\mu_o - \mu'}{\sigma/n} + \Phi^{-1}(1-\alpha/2)\right) \tag{2}$$

$$= \mathrm{P}\left(\frac{\mu_o - \mu'}{\sigma/n} - \Phi^{-1}(1-\alpha/2) < Z < \frac{\mu_o - \mu'}{\sigma/n} + \Phi^{-1}(1-\alpha/2)\right) \tag{3}$$

$$= \Phi\left(\frac{\mu_o - \mu'}{\sigma/n} + \Phi^{-1}(1-\alpha/2)\right) - \Phi\left(\frac{\mu_o - \mu'}{\sigma/n} - \Phi^{-1}(1-\alpha/2)\right) \tag{4}$$

From this expression, can you decrease the type II error for a particular μ'?
(Notice that I had corrected all α to $\alpha/2$ since this is clearly a two-sided tests as we do not know whether $\mu' > \mu$ or $\mu' < \mu$)

Type II error simulation setup

Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 5)$, where $n = 10$. We want to find the type II error for $\mu' = 0.5$ against the null hypothesis that the true mean $\mu = 0$ with a two-sided test (i.e. the alternative is $\mu \neq 0$).

- Step 1: Simulate $x_1, \ldots, x_{10}$ according to $\mathcal{N}(\mu', 5)$.
- Step 2: Perform one sample z-test ($\bar{X} \sim \mathcal{N}(\mu_o, 5/10)$) and obtain a p-value.
- Repeat Step 1 and 2, $B = 1000$ times, giving a vector of p-values $p_1, \ldots, p_{1000}$.
- How many of those p-values do you expect to be less than $\alpha = 0.05$ or $\alpha = 0.01$?

Type II error simulation R

```
n = 10; mu_o = 0; sigma = sqrt(5); B = 1000; mu_p = 0.5
pval <- NA
i = 1;
while(i <= B){
sample_cases <- rnorm(n, mu_p, sigma)
pval[i] <- 2*min(pnorm((mean(sample_cases) - mu_o)/(sqrt(sigma^2/n)), lower=F),
                 pnorm((mean(sample_cases) - mu_o) / (sqrt(sigma^2/n)), lower=T))
i = i + 1
}
sum(pval > 0.05)/B
```

```
## [1] 0.889
```

```
sum(pval > 0.01)/B
```

```
## [1] 0.965
```

Type II error simulation R

Suppose we increase the sample size to $n = 100$:

```
n = 100; mu_o = 0; sigma = sqrt(5); B = 1000; mu_p = 0.5
pval <- NA
i = 1;
while(i <= B){
sample_cases <- rnorm(n, mu_p, sigma)
pval[i] <- 2*min(pnorm((mean(sample_cases) - mu_o)/(sqrt(sigma^2/n)), lower=F),
                 pnorm((mean(sample_cases) - mu_o) / (sqrt(sigma^2/n)), lower=T))
i = i + 1
}
sum(pval > 0.05)/B
```

```
## [1] 0.381
```

```
sum(pval > 0.01)/B
```

```
## [1] 0.623
```

Type II error simulation Python

```
import numpy as np
from scipy import stats
import random
n=10
mu_o=0
mu_p=0.5
B=1000
pval=[]
for j in range(0, B):
  sample_cases = np.random.normal(mu_p, 5**0.5, n)
  z_scores = (sample_cases.mean())/(5.0/n)**(0.5)
  pval.append((1-stats.norm.cdf(abs(z_scores)))*2)
print(sum(i > 0.05 for i in pval))
```

```
## 901
```

```
print(sum(i > 0.01 for i in pval))
```

```
## 978
```

Type II error simulation Python

Suppose we increase the sample size to $n = 100$:

```
import numpy as np
from scipy import stats
import random
n=100
mu_o=0
mu_p=0.5
B=1000
pval=[]
for j in range(0, B):
  sample_cases = np.random.normal(mu_p, 5**0.5, n)
  z_scores = (sample_cases.mean())/(5.0/n)**(0.5)
  pval.append((1-stats.norm.cdf(abs(z_scores)))*2)
print(sum(i > 0.05 for i in pval))
```

```
## 380
```

```
print(sum(i > 0.01 for i in pval))
```

```
## 611
```

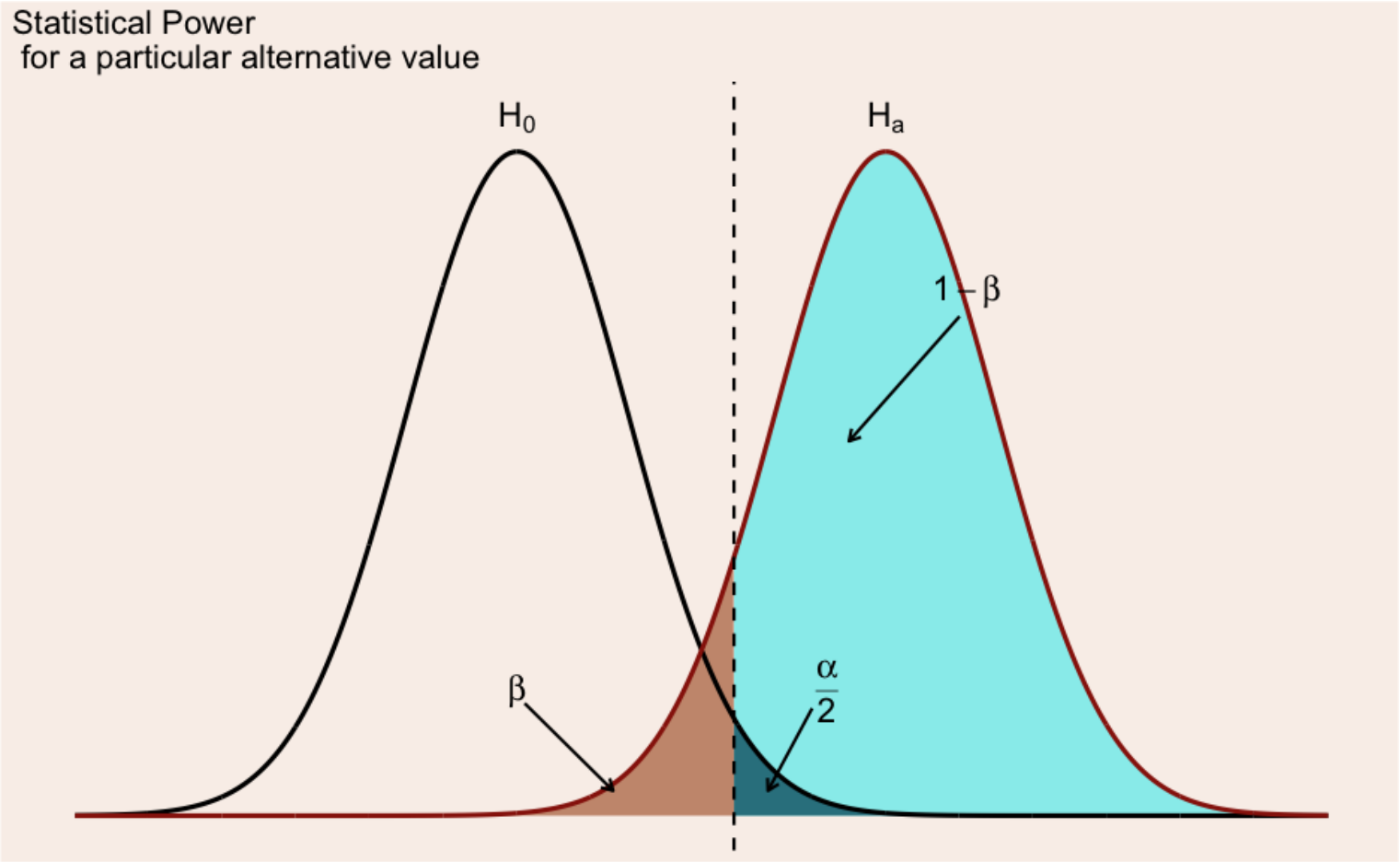
Statistical power

Contrary to type II error, statistical power is defined as:

$$\text{Power} = 1 - \beta = \text{P}(H_o \text{ is rejected at significance level } \alpha | H_o \text{ is false})$$

- Statistical power reflects the strength of a **test** to detect **a specific alternative** at some α value.
- The higher the statistical power (given an possible true parameter value), the more likely we are to reject the null hypothesis in a particular sample.
- Statistical power is considered **only under the alternative hypothesis**
- Type I error α is considered **only under the null hypothesis**

Statistical power



Statistical power, type I error, type II error and effect size

Suppose the null hypothesis is $\theta = \theta_o$ while a particular value of the alternative is θ'. The statistical power of rejecting $\theta = \theta_o$ while assuming $\theta = \theta'$ is denoted by $\beta(\theta')$.

- Denote $\delta = \theta_o - \theta'$
- The significance level is at α
- The sample size is n

How do we increase the statistical power by varying these values?

An example of Statistical power for the t-test (two-sided)

The statistical power of rejecting the null with a one sample **t-test** for a particular $\mu' > \mu_o$ is:

$$1 - \beta(\theta') = 1 - \mathrm{P}(\bar{X} < \mu_o - t_{(1-\alpha/2,n-1)} S/n | \mu = \mu') \tag{5}$$

$$= 1 - \mathrm{P}(\frac{\bar{X} - \mu'}{S/n} < \frac{\mu_o - \mu'}{S/n} - t_{(1-\alpha/2,n-1)}) \tag{6}$$

$$= 1 - \mathrm{P}(T < \frac{\mu_o - \mu'}{S/n} - t_{(1-\alpha/2,n-1)}) \tag{7}$$

$$\tag{8}$$

where $t_{(1-\alpha/2,n-1)}$ is the quantile value for a t-distributed random variable with degrees of freedom $n-1$.
From this expression, can you come up with ways to increase the statistical power?

An example of Statistical power for the t-test

- Notice that when $\mu_o = \mu'$, power is equal to α (i.e. power is always greater or equal to α)
- Everything else fixed, if α **decreases**, then the quantile increases, the power **decreases**.
- Everything else fixed, if $\mu' - \mu_o$ **increases**, then the quantile decreases, the power **increases**.
- Everything else fixed, if n **increases**, then the quantile decreases, the power **increases**.

An example of Statistical power for the t-test (one-sided)

Suppose the null hypothesis is $\theta = \theta_o$ while a particular value of the alternative is θ'. The statistical power of rejecting $\theta = \theta_o$ while assuming $\theta = \theta'$ is denoted by $\beta(\theta')$.

- Now we have $\mu' < \mu_o$
- Can you replicate the calculation above and conclude under what circumstances, the power would increase?

Statistical power simulation

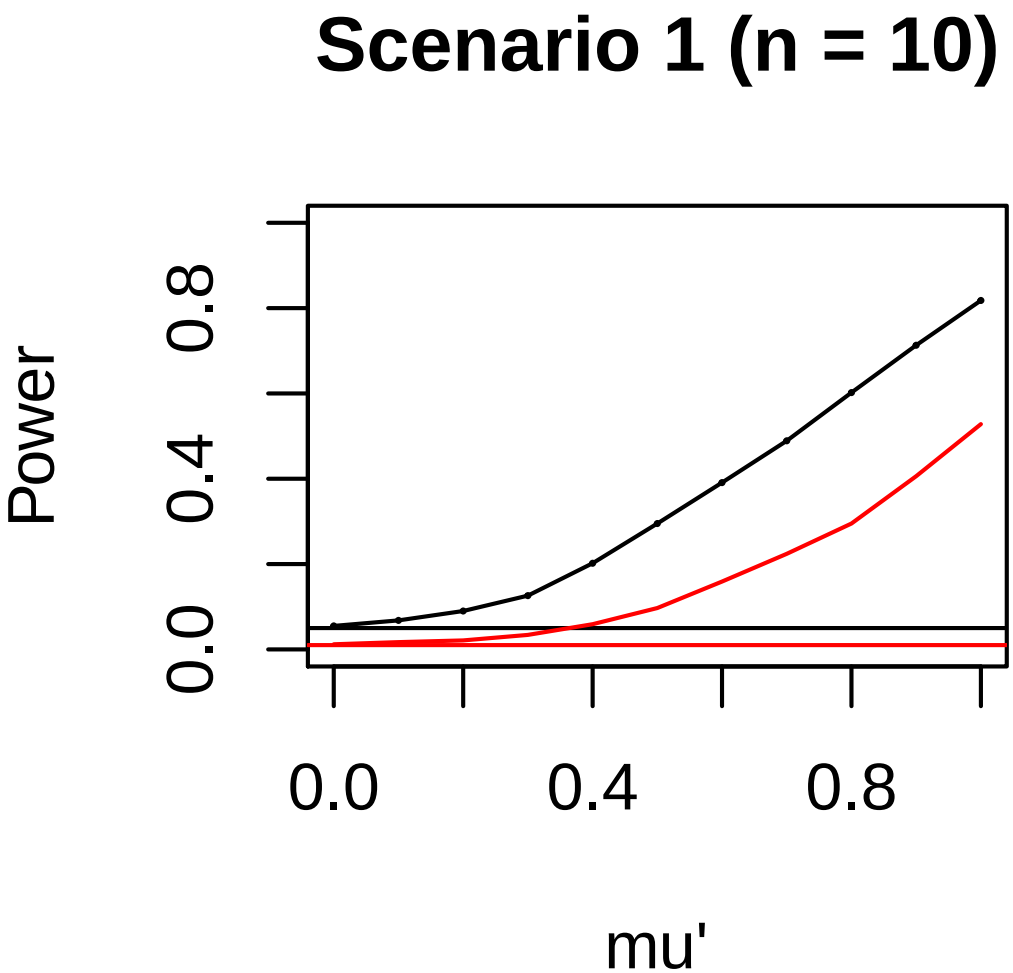
one sample t-test examples (assume $\sigma^2 = 1$ is unknown):

- Scenario 1: Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where $n = 10$. We want to find the power for each $\mu' > 0$ against the null hypothesis that the true mean $\mu = 0$ with a two-sided test (i.e. the alternative is $\mu \neq 0$).
- Scenario 2: Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where $n = 100$. We want to find the power for each $\mu' > 0$ against the null hypothesis that the true mean $\mu = 0$ with a two-sided test (i.e. the alternative is $\mu \neq 0$).
- Scenario 3: Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, where $n = 100$. We want to find the power for each $\mu' < 0$ against the null hypothesis that the true mean $\mu = 0$ with a one-sided test (i.e. the alternative is $\mu < 0$).

Statistical power simulation R (scenario 1)

```
n = 10; mu_o = 0; B = 1000;
mu_seq <- seq(0,1,0.1)
power1 <- data.frame("alpha5" = NA, "alpha1" = NA)
for (j in 1:length(mu_seq)){
p_value1 <- replicate(B, t.test(rnorm(n, mu_seq[j], 1), mu = 0)$p.value)
power1[j,] <- c(sum(p_value1 < 0.05)/B, sum(p_value1 < 0.01)/B)
}
```

Statistical power simulation R (scenario 1)

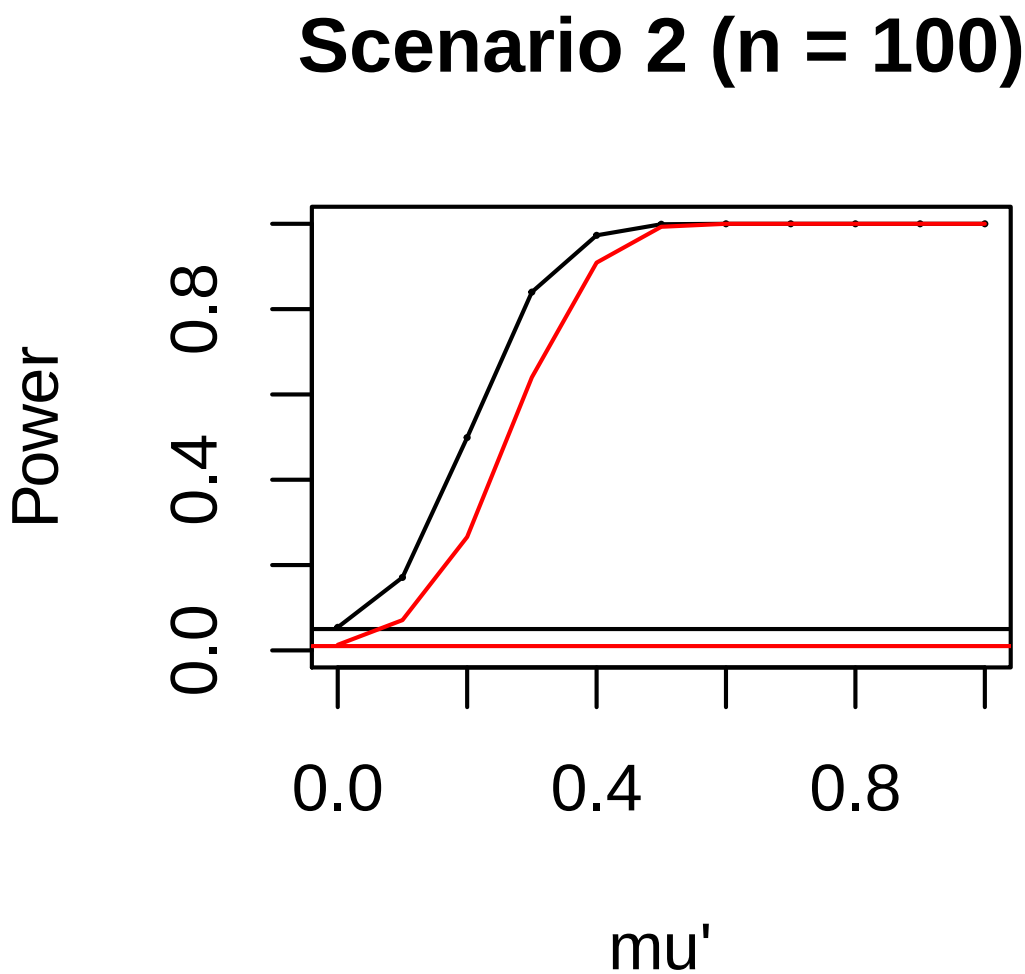


Statistical power simulation R (scenario 2)

Suppose we increase the sample size to $n = 100$:

```
n = 100; mu_o = 0; B = 1000;
mu_seq <- seq(0,1,0.1)
power2 <- data.frame("alpha5" = NA, "alpha1" = NA)
for (j in 1:length(mu_seq)){
p_value2 <- replicate(B, t.test(rnorm(n, mu_seq[j], 1), mu = 0)$p.value)
power2[j,] <- c(sum(p_value2 < 0.05)/B, sum(p_value2 < 0.01)/B)
}
```

Statistical power simulation R (scenario 2)

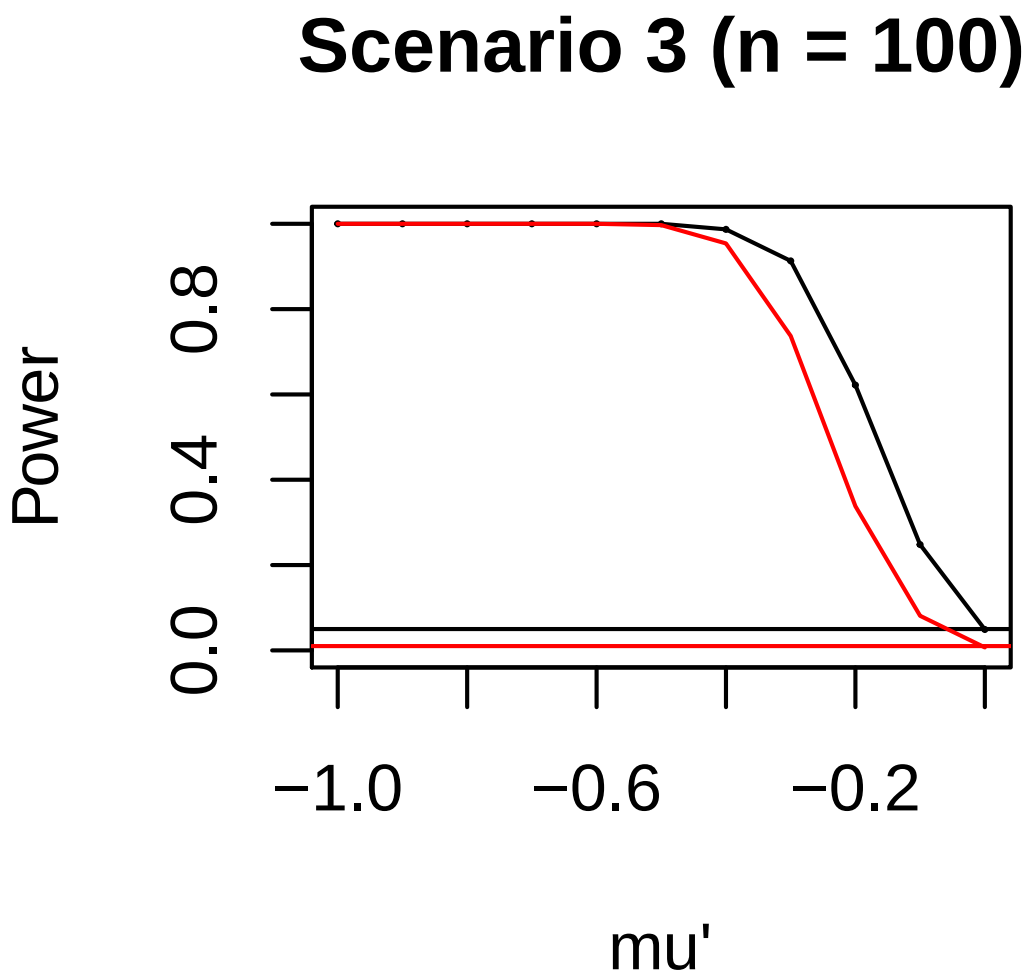


Statistical power simulation R (scenario 3)

Suppose the test is one-sided with $n = 100$:

```
n = 100; mu_o = 0; B = 1000;
mu_seq <- seq(-1,0,0.1)
power3 <- data.frame("alpha5" = NA, "alpha1" = NA)
for (j in 1:length(mu_seq)){
p_value3 <- replicate(B, t.test(rnorm(n, mu_seq[j], 1), mu = 0, alternative = "less")$p.val
power3[j,] <- c(sum(p_value3 < 0.05)/B, sum(p_value3 < 0.01)/B)
}
```

Statistical power simulation R (scenario 3)



Statistical power simulation Python (scenario 1)

```
import numpy as np
from scipy import stats
import random
from __future__ import division
n=10
mu_o=0
mu_seq=np.arange(0.0, 1.0, 0.1)
B=1000
power1=list()
for i in range(0, len(mu_seq)):
  pval=[]
  for j in range(0, B):
    sample_cases = np.random.normal(mu_seq[i], 5**0.5, n)
    z_scores = (sample_cases.mean())/(5.0/n)**(0.5)
    pval.append((1-stats.norm.cdf(abs(z_scores)))*2)
  power1.append([round(sum(i < 0.05 for i in pval)/B,3), round(sum(i < 0.01 for i in pval)/
```

Statistical power simulation Python (scenario 2)

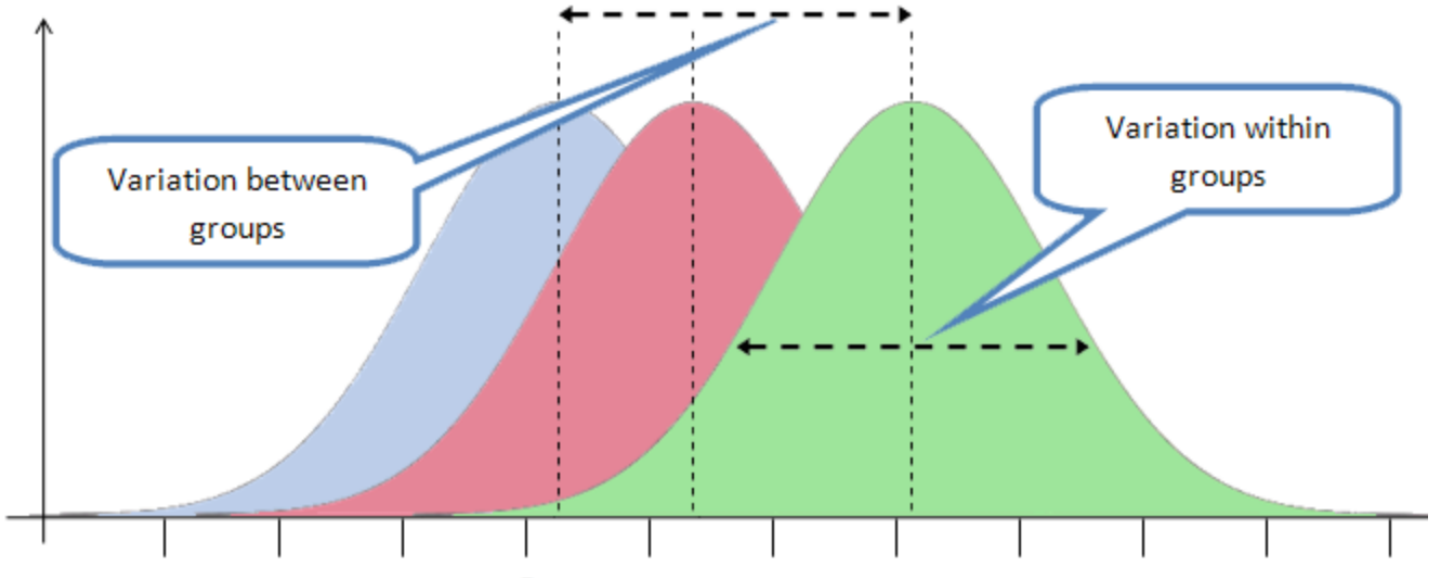
Suppose the sample size $n = 100$:

```
import numpy as np
from scipy import stats
import random
from __future__ import division
n=100
mu_o=0
mu_seq=np.arange(0.0, 1.0, 0.1)
B=1000
power2=list()
for i in range(0, len(mu_seq)):
  pval=[]
  for j in range(0, B):
    sample_cases = np.random.normal(mu_seq[i], 5**0.5, n)
    z_scores = (sample_cases.mean())/(5.0/n)**(0.5)
    pval.append((1-stats.norm.cdf(abs(z_scores)))*2)
  power2.append([round(sum(i < 0.05 for i in pval)/B,3), round(sum(i < 0.01 for i in pval)/
```

Statistical power simulation Python (scenario 3)

Suppose the test is one-sided with sample size $n = 100$:

```
import numpy as np
from scipy import stats
import random
from __future__ import division
n=100
mu_o=0
mu_seq=np.arange(-1.0, 0, 0.1)
B=1000
power3=list()
for i in range(0, len(mu_seq)):
  pval=[]
  for j in range(0, B):
    sample_cases = np.random.normal(mu_seq[i], 5**0.5, n)
    z_scores = (sample_cases.mean())/(5.0/n)**(0.5)
    pval.append(stats.norm.cdf(z_scores))
  power3.append([round(sum(i < 0.05 for i in pval)/B,3), round(sum(i < 0.01 for i in pval)/
```



How about equality of means in more than two samples?

Basic Framework of one-way ANOVA

- This is an extension of two-sample ($\sigma_1^2 = \sigma_2^2$) t-test.
  - The pooled-variance t-test tests the null hypothesis that two population means are equal, i.e. $H_0 : \mu_1 = \mu_2$.
- The **one-way Analysis of Variance (ANOVA)** can test the equality of several population means. That is:

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k, \quad H_a : \exists i, j \ s.t. \ \mu_i \neq \mu_j$$

  - Assumptions
    - Normal populations, i.e. assume normality for each population.
    - Equality of population variances, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$.

Basic Framework of one-way ANOVA

- The term one- way, also called one-factor, indicates that there is a single explanatory variable ("treatment") with two or more levels, and only one level of treatment is applied at any time for a given subject.

- When the factor variable has exactly two levels, one-way ANOVA and two independent samples t-test always come to the the same conclusions regardless of which method we use.

- ANOVA: compare means by analysing variability. Below is a bit history of ANOVA
  - It goes back to early work by Fisher in 1918 on mathematical genetics.
  - Further developed by R.A. Fisher in 1920.
  - The convenient acronym ANOVA was coined much later, by John W. Tukey (1915–2000), the pioneer of exploratory data analysis (EDA)

ANOVA set up

$$Y_{ij} \sim N(\mu_i, \sigma^2), i = 1, 2, \ldots, r; j = 1, 2, \ldots, n_i$$

- μ_i is the mean for a factor variable at level i
- For each i, we draw ni independent samples Y_{jj}
- Total sample size: $n = \sum_{i=1}^{r} n_i$
- Denote $\bar{Y}_{i.}$ be i-the level mean and $\bar{Y}_{..}$ be the grand mean.

ANOVA: Hypothesis of equal means

- **Null and alternative:**

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_r, \quad H_a : \exists i \neq j, s.t.\ \mu_i \neq \mu_j$$

- **Test Statistic**

$$F^* = \frac{MST}{MSE} = \frac{SS_T/(r-1)}{SS_E/(n-r)} \sim F_{r-1,n-r} \text{ under } H_0$$

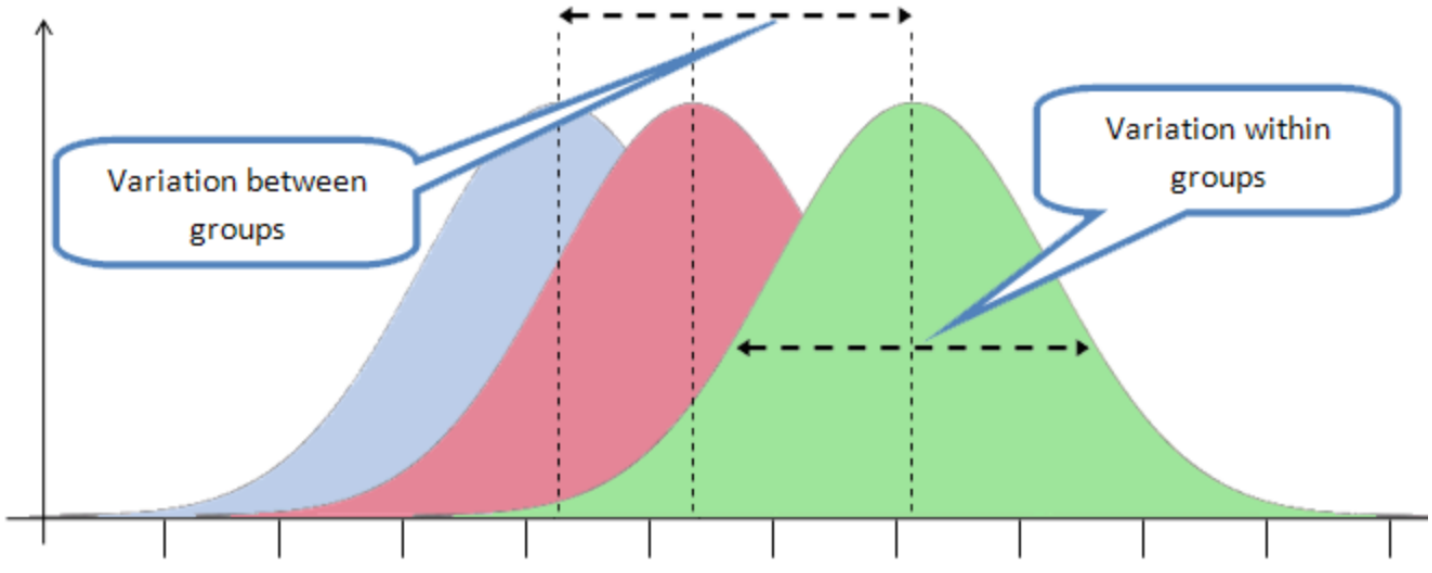
  - If H_0 is true, $F^* \approx 1$, we don't have evidence to reject H_0.
  - If H_0 is false, $MST > MSE$, F^* increases. So large values of observed F^* are evidence against H_0, and we test H_0 using a one-tailed test.

- **Decision:** rejecting at significance level α if F^* is too big, that is, we reject H_0 if

$$F^* > F_{1-\alpha,r-1,n-r}$$

One-way ANOVA

Why analysing variance help to compare group means



next week schedule

- *Monday*: Midterm solution and review
- *Wednesday*: Regression: moving from estimation to prediction
  - regression as a supervised learning technique
  - examples of regression applications
  - training and testing
  - cross-validation
  - optimization

week of August 6 schedule

- *Monday*: civic holiday no lecture
- *Wednesday*: simple and multiple linear regression
  - Through constructing a regression model with unknown parameters
  - Estimating unknown parameters to obtain an **estimated model**
  - Use the estimated model to **predict a new observation**

final week schedule

- *Monday*: review lecture + office hour 2-6pm
- *Thursday* (August 16th): office hour 10am-2pm
- *Friday*: Final Exam scheduled for 9am-12pm