STA303/1002 - Methods of Data Analysis II

(Week 02 lecture note)

Wei (Becky) Lin

Jan. 09, 2017

UNIVERSITY OF TORONTO

Learning objective this week

- Review of distribution theory and statistical inference
- t-test
  - one sample t-test
  - two samples t-test
  - paired t-test
  - testing of regression parameter in linear regression model: $H_0 : \beta_k = 0$
- one-way ANOVA

Review of Distribution Theory and Statistical Inference

Review of Distribution Theory

1. If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.
2. If $X_i \sim N(\mu_i, \sigma_i^2)$ independently, then

$$W = a + \sum_1^n a_i X_i \sim N(a + \sum_1^n a_i \mu_i, \sum_1^n a_i^2 \sigma_i^2)$$

Linear combination of N. RVs is still N.

3. If $Z_i \sim_{iid} N(0,1)$, then $\sum_1^n Z_i^2 \sim \chi_n^2$ (From $N(0,1) \to \chi_n^2$)
4. If $Z \sim N(0,1) \perp X \sim \chi_n^2$, then $T = \frac{Z}{\sqrt{X/n}} \sim t_n$.

$$\boxed{t_n = \frac{N(0,1)}{\sqrt{\chi_n^2/n}}}$$

5. If $X_1 \sim \chi_n^2 \perp X_2 \sim \chi_m^2$, then $\frac{X_1/n}{X_2/m} \sim F_{n,m}$

$$\boxed{F_{n,m} = \frac{\chi_n^2/n}{\chi_m^2/m}}$$

6. If $X_1, \ldots X_n \sim_{iid} N(\mu, \sigma^2)$ then
   - $\bar{X} \sim N(\mu, \sigma^2/n)$
   - $s^2 = \sum (X_i - \bar{X})^2/(n-1)$, and $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$
   - $\bar{X} \perp s^2$, and

lose 1 d.f. since s^2 use $\bar{x}$ instead of μ.

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)s^2}{\sigma^2}/(n-1)}} = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

Review: Hypothesis Testing

- The intent of hypothesis testing is formally examine two opposing conjectures (hypotheses), H_0 and H_a.
- These two hypotheses are mutually exclusive and exhaustive so that one is true to the exclusion of the other
- We accumulate evidence - collect and analyze sample information - for the purpose of determining whether there is enough evidence in a sample of data to infer the null hypothesis is true.

Review: The Null and Alternative Hypothesis

The null hypothesis, H_0

- 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_0
- Never contains just the "=" sign
- Is generally the hypothesis that is believed to be true by the researcher

One and Two sided Tests

Hypothesis tests can be one or two sided (tailed)

- One tailed testes are directional

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_a : \mu_1 - \mu_2 > 0$$ (or $\mu_1 - \mu_2 < 0$)

- Two tailed tests are not directional

$$H_0 : \mu_1 - \mu_2 = 0$$

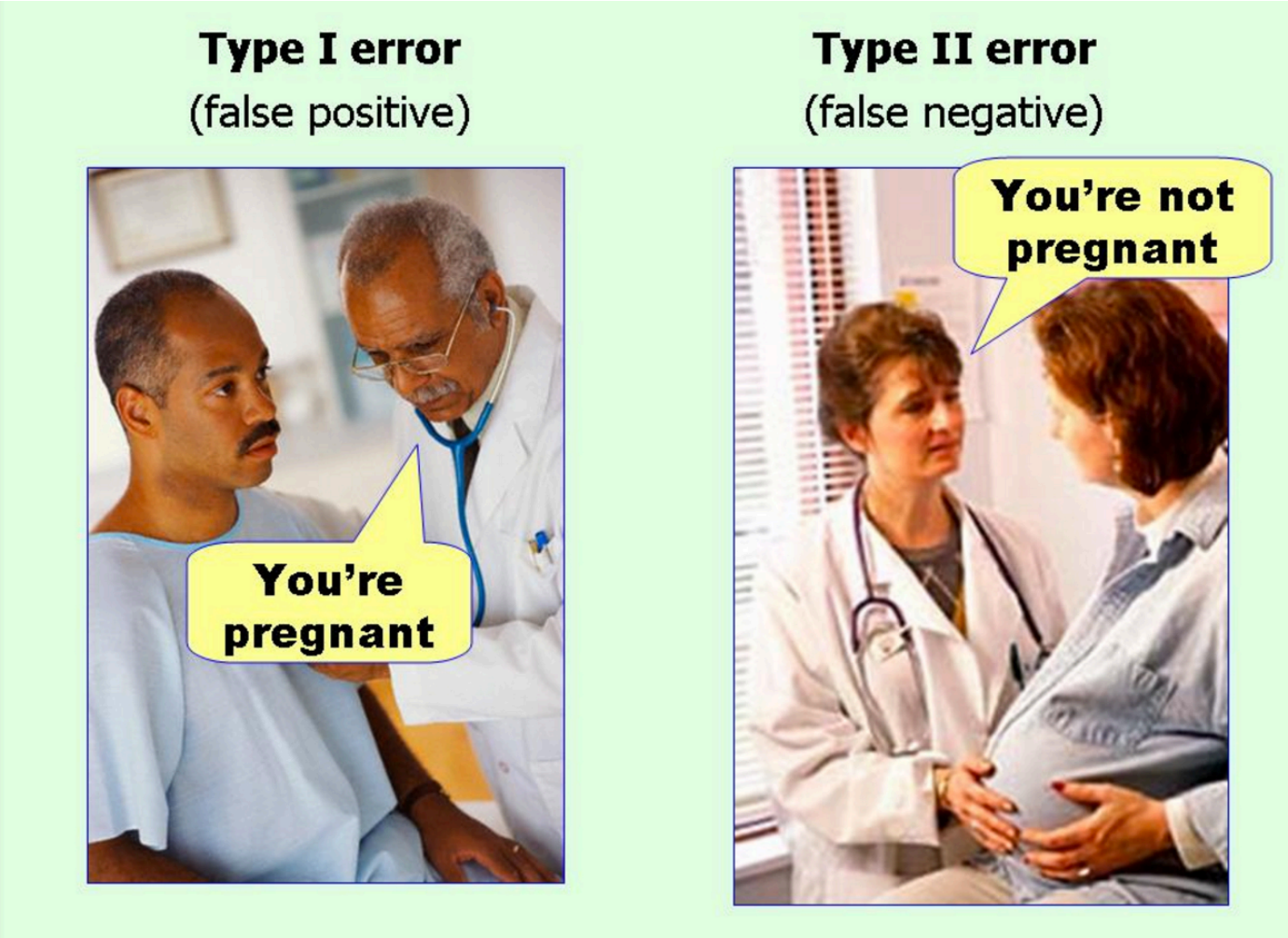
$$H_a : \mu_1 - \mu_2 \neq 0$$

Review: Hypothesis Testing

		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"

- α=P(Type I error) = P(reject $H_0|H_0$ is true).
- β=P(Type II error) = P(fail to rejct $H_0|H_0$ is false)
- Significance level α: maximum allowable probability of making type I error (usually 5%)
- Power = 1-β=P(correctly reject H_0 when it is false)
- Test statistics: a statistics (function of data variable) whose distribution under H_0 is known.

Review: Hypothesis Testing



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

Parametric and Non-parametric Tests

- **Parametric Tests**: Relies on theoretical distributions of the test statistic under the null hypothesis and assumptions about the distribution of the sample data (i.e., normality)

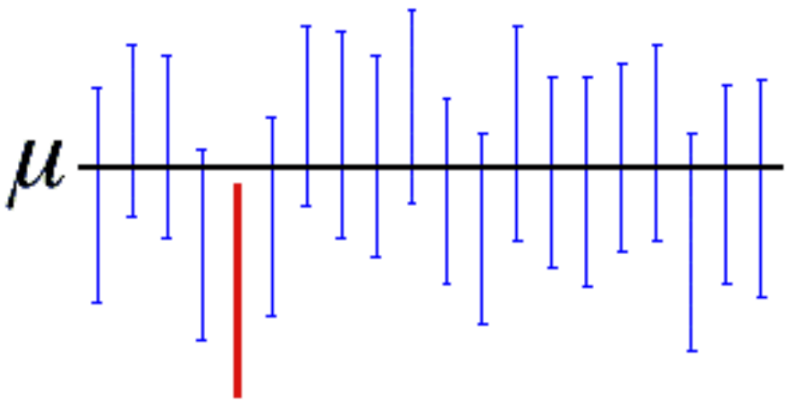
- **Non-Parametric Tests**: Referred to as "Distribution Free" as they do not assume that data are drawn from any particular distribution

Review: Confidence Interval for μ given σ^2

Two-side 95% CI for μ (assume σ^2 is known)

$$\bar{X} \pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = (\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

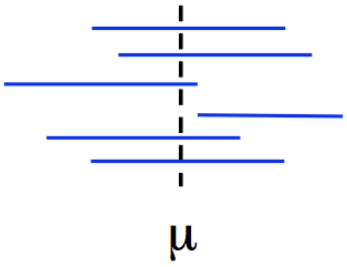
- A confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval which is calculated from the observations, in principle different from sample to sample.
- How frequently the observed interval contains the parameter is determined by the confidence level or confidence coefficient.
- More specifically, confidence level suggests that if CI are constructed across many separate data analyses of replicated experiments, the proportion of such intervals that contain the true value of the parameter will match the given confidence level.



Interpreting a 95% CI

- We calculate a 95% CI for a hypothetical sample mean to be between 20.6 and 35.4. Does this mean there is a 95% probability the true population mean is between 20.6 and 35.4?

- NO! Correct interpretation relies on the long-term frequency interpretation of probability.



More discussion on this topic:
`http://stats.stackexchange.com/questions/26450/why-does-a-95-confidence-interval-ci-not-imply-a-95-chance-of-containing-the`

Review: P-value and CI agree about statistical significance

(I) $H_0: \theta = \theta_0$ $H_a: \theta \neq \theta_0$

- two sided test

Decision
- P-value calculated from two tails
- $(1-\alpha)$ CI. (Low, Upp)

- 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. (i.e. Reject H_0)
  - If the CI does not contain the null hypothesis value (H_0 value), the results are statistically significant.
  - If P-value$< \alpha$, the CI will not contain the null hypothesis value.

(II) $H_0: \theta = \theta_0$ $H_a: \theta > \theta_0$

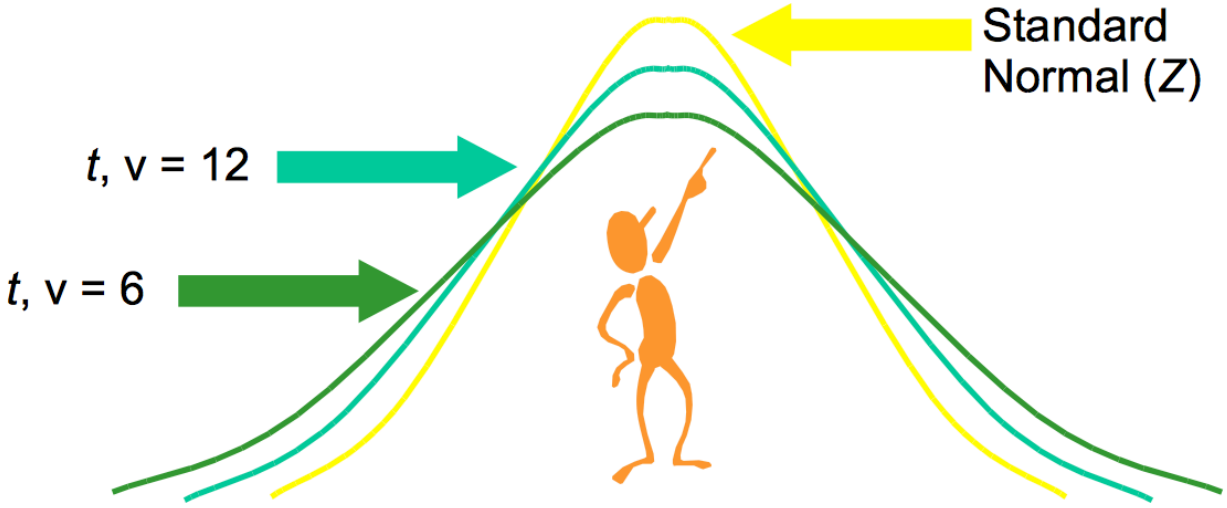
- one sided test

Decision
- P-value : right tail prob.
- $(1-\alpha)$ lower conf. bounds (Low, ∞)

(III) $H_0: \theta = \theta_0$ $H_a: \theta < \theta_0$

- one-sided test

Decision
- P. value : left tail prob
- $(1-\alpha)$ upper conf bound $(-\infty, \text{upp})$



The humble t-test

Types of t-test

- One sample t-test
  - Purpose: Tests whether the mean of a single population is equal to a target value

$$H_0 : \mu = \mu_0, \qquad H_a : \mu \neq (>, <) \ \ 0$$

  - Example: Is the mean height of female college students greater than 5.5 feet?
- Paired t-test
  - Purpose: Tests whether the mean of the differences between dependent or paired observations is equal to a target value

$$H_0 : \mu_d = \delta_0, \qquad H_a : \mu_d =\neq (>, <) \ \ \delta_0$$

  - If you measure the weight of male college students before and after each subject takes a weight-loss pill, is the mean weight loss significant enough to conclude that the pill works?

Types of t-test (be contd)

- Two sample t-test
  - Purpose: Tests whether the difference between the means of two independent populations is equal to a target value

$$H_0 : \mu_1 - \mu_2 = \delta_0, \qquad H_a : \mu_1 - \mu_2 \neq (>, <) \ \ \delta_0$$

  - Example: Does the mean height of female college students significantly differ from the mean height of male college students?
- t-test in regression output
  - Purpose: Tests whether the values of coefficients in the regression equation differ significantly from zero

$$H_0 : \beta_k = \beta_k^0, = 0 \qquad H_a : \beta_k \neq (>, <) \ \ \beta_k^0 = 0$$

  - Example: Are high school SAT test scores significant predictors of college GPA?

Comments on t-test

- An important property of the t-test is its robustness against assumptions of population normality. In other words, t-tests are often valid even when the assumption of normality is violated, but only if the distribution is not highly skewed. This property makes them one of the most useful procedures for making inferences about population means.
- However, with nonnormal and highly skewed distributions, it might be more appropriate to use nonparametric tests.

One sample t-test

Assumption:

- The data variable Y is continuous.
- Data variable Y must follow the normal probability distribution.
- The sample is a simple random sample from its population.

t-test:

$$H_0 : \mu = 0, \qquad H_a : \mu \neq 0 \text{ or } \mu \geq (\leq) 0$$

$$t = \frac{\bar{y} - \mu}{S/\sqrt{n}} \text{ a.k.a signal-to-noise ratio} \sim_{H_0} t_{n-1}$$

where

$$S^2 = \frac{\sum_1^n (y_i - \bar{y})^2}{n-1}$$

1-α CI :

$$\left(\bar{y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}},\ \bar{y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right)$$

R example: one sample t-test

```r
mark= c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)
t.test(mark,mu=75)
```

```
##
##  One Sample t-test
##
## data:  mark
## t = -0.78303, df = 9, p-value = 0.4537
## alternative hypothesis: true mean is not equal to 75
## 95 percent confidence interval:
##  60.22187 82.17813
## sample estimates:
## mean of x
##      71.2
```

$H_0: \mu = 75$

$\Rightarrow$ $H_a: \mu \neq 75$

$\leftarrow$ (Low, Upp) = $(1-\alpha)$ CI $\leftarrow$ two sided test

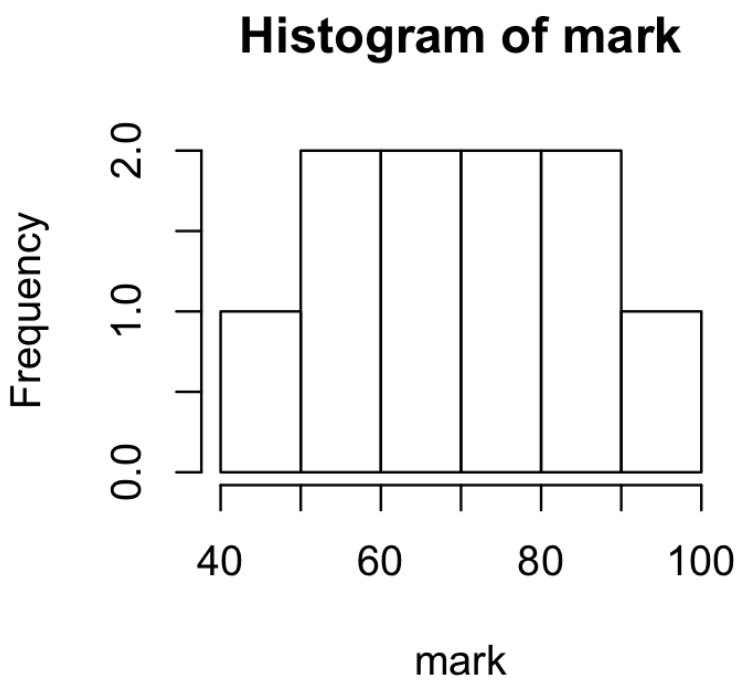
* $75 \in$ CI. $\longrightarrow$ fail to reject H_0.

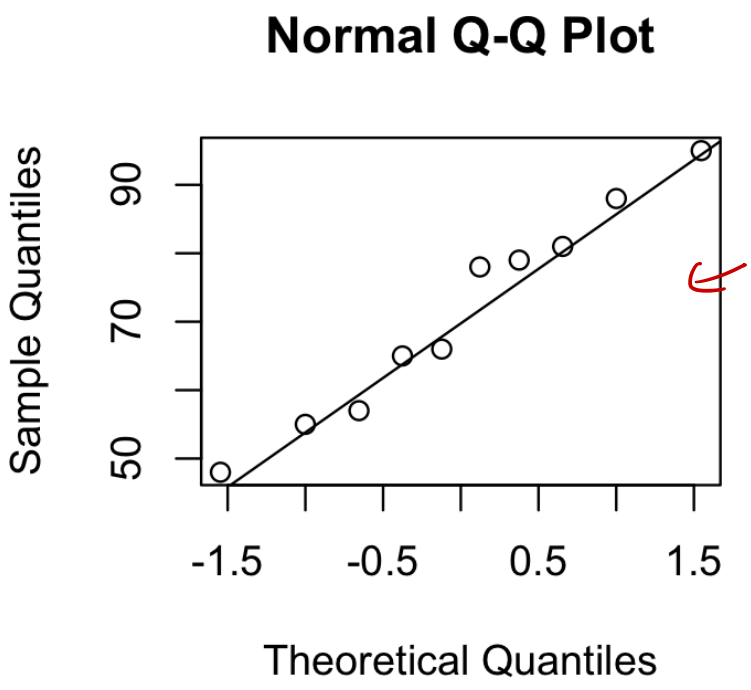
* P-val = 0.4537 $\nearrow$

R example: one sample t-test

```
mark= c(65, 78, 88, 55, 48, 95, 66, 57, 79, 81)
par(mfrow=c(1,2))
hist(mark)
qqnorm(mark); qqline(mark)
```

check Normality





looks fine!

Paired t-test

- Paired Design
  - Before-after data: blood pressure before and after taking one kind of medication
  - Twin data
  - Matched case-control
- Why pairing
  - Control extraneous noise
  - Each observation acts as a control
- Testing: $H_0 : \mu_d = \delta_0$ vs $H_a : \mu_d \neq (\geq, \leq)\delta_0$

sub	Y_1	Y_2	$Y_d = Y_1 - Y_2$
1	Y_{11}	Y_{12}	$Y_{11} - Y_{12}$
2	Y_{21}	Y_{22}	$Y_{21} - Y_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	Y_{n1}	Y_{n2}	$Y_{n1} - Y_{n2}$

become one sample t-test

$$T = \frac{\bar{y}_d - \delta_0}{S/\sqrt{n}} \sim_{H_0} t_{n-1}, S^2 = \frac{1}{n-1}\sum_{1}^{n}(y_{d(i)} - \bar{y}_d)^2$$

  - Assumptions:
    - Data are continous
    - The differences for the matched-pairs, follow a normal probability distribution.
    - The sample of pairs is a simple random sample from its population

R Example: Paired t-test

A school athletics has taken a new coach, and want to test the effectiveness of the new type of training proposed by comparing the average times of 10 runners in the 100 meters. Are below the time in seconds before and after training for each athlete.

- Want to test if the new coach has made substantial improvements to the team

before - After > 0: implies improvements

$$H_0 : \mu_b - \mu_a = 0 \qquad H_a : \mu_b - \mu_a > 0$$

No	1	2	3	4	5	6	7	8	9	10
Before	12.9	13.5	12.8	15.6	17.2	19.2	12.6	15.3	14.4	11.3
After	12.0	12.2	11.2	13.0	15.0	15.8	12.2	13.4	12.9	11.0

R Example: Paired t-test

```r
b = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
a = c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)
t.test(b,a,paired=T,alt="greater")
```

```
##
##  Paired t-test
##
## data:  b and a
## t = 5.2671, df = 9, p-value = 0.0002579
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.049675      Inf
## sample estimates:
## mean of the differences
##                    1.61
```

$H_0: \mu_b - \mu_a = 0$ $\quad H_a: \mu_b - \mu_a > 0$

$(\text{Low}, \infty) \leftarrow$ $H_a: \mu_b - \mu_a > 0$ one-sided test

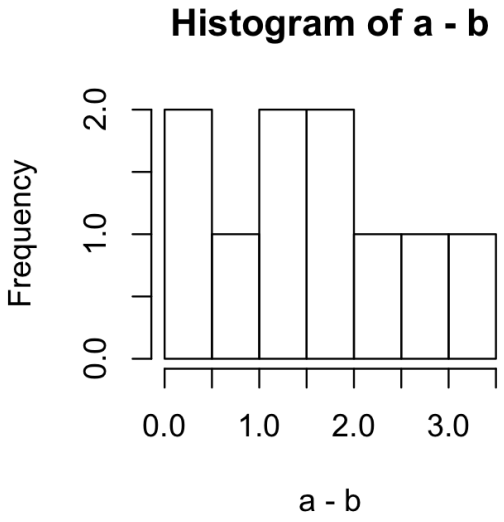
$\left\{ \begin{array}{l} 0 \notin (1.049675, \infty) \\ \text{P.val} = 0.0002579 \end{array} \right. \Rightarrow$ Reject H_0.

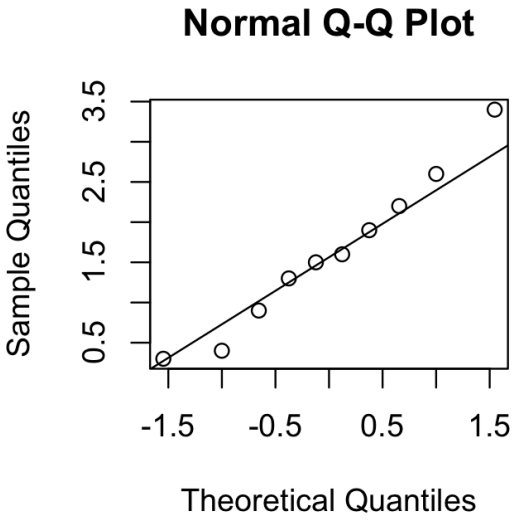
R Example: Paired t-test

```
b = c(12.9, 13.5, 12.8, 15.6, 17.2, 19.2, 12.6, 15.3, 14.4, 11.3)
a = c(12.0, 12.2, 11.2, 13.0, 15.0, 15.8, 12.2, 13.4, 12.9, 11.0)
#par(mfrow=c(1,2))
#hist(b-a); qqnorm(b-a); qqline(b-a)
shapiro.test(b-a)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  b - a
## W = 0.97176, p-value = 0.9067
```

After differencing, check normality.





For such small sample, it looks ok.

Two sample t-test

Assumptions:

- The data are continuous (not discrete).
- The data follow the normal probability distribution.
- The variances of the two populations are equal. (If not, the Aspin-Welch Unequal-Variance test is used.)
- The two samples are independent. There is no relationship between the individuals in one sample as compared to the other.
- Both samples are simple random samples from their respective populations.

$$X_i \sim N(\mu_1, \sigma_1^2), i = 1, 2.., m; \quad \perp \quad Y_j \sim N(\mu_2, \sigma_2^2), j = 1, 2, .., n$$

Which two sample t-test to use

$$H_0 : \mu_x - \mu_y = \delta_0 \qquad H_a : \mu_x - \mu_y \neq \delta_0$$

Determination of which two-sample t-test to use is dependent upon first testing the variance assumption:

$\sigma_1^2 = \sigma_2^2$

- Two sample t-test for Homogeneous Variances.(using pooled variance)

$$S_p^2 = \frac{\sum_1^m (x_i - \bar{x})^2 + \sum_1^n (y_j - \bar{y})^2}{(m-1) + (n-1)} = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2}$$

$$\boxed{t = \frac{(\bar{x} - \bar{y}) - \delta}{S_p\sqrt{1/m + 1/n}} \sim_{H_0:\delta=\delta_0 \rightarrow} t_{m+n-2}}$$

- Two-Sample t-test for Heterogeneous Variances. $(\sigma_1^2 \neq \sigma_2^2)$

$$\boxed{t = \frac{(\bar{x} - \bar{y}) - \delta_0}{\sqrt{S_x^2/m + S_y^2/n}} \sim_{H_0} t_{df}, \quad df = \frac{(S_x^2/m + S_y^2/n)^2}{\frac{(S_x^2/m)^2}{m-1} + \frac{(S_y^2/n)^2}{n-1}}}$$

  - Performs the t-test in exactly the same fashion as for homogeneous variances; but use **adjusted df**. Note that this can have a big effect on decision.

Variance Ratio F-test

- $(n_{max}-1)S_{max}^2/\sigma^2 \sim \chi^2_{n_{max}-1}$
- $(n_{min}-1)S_{min}^2/\sigma^2 \sim \chi^2_{n_{min}-1}$

$\Rightarrow S_{max}^2/S_{min}^2 \sim F_{n_{max}-1, n_{min}-1}$

For homogeneity of variance

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad H_a : \sigma_1^2 \neq \sigma_2^2$$

Test:

$$F^* = S_{max}^2/S_{min}^2 \sim_{H_0} F_{n_{max}-1, n_{min}-1}$$

- Reject H_0 if $F^* \overset{?}{\cancel{<}} F_{1-\alpha, n_{max}-1, n_{min}-1}$
- Example: $S_a^2 = 16, S_b^2 = 12.9, n_a = 12, n_b = 8$, $F^* = 16/12.9 = 1.24 < F_{0.95,11,7} = 3.603$, we fail to reject H_0, there is a chance that variances are homogeneous.
- **Warning**: requires that the two samples are drawn from normal populations. If the two samples are not normally distributed, do not use Variance Ratio F-test! Could consider use the Modified Levene Equal-Variance Test.

R: two sample t-test

```r
C<-c(10,11,12,11,10,11,11) # size for one type fish
W<-c(7,8,11,12,10,13,11,10,12) # size for another type fish
var.test(C,W)
```

$H_0: S^2_{max}/S^2_{min} = 1$

$H_a: S^2_{max}/S^2_{min} \neq 1$

```
##
##  F test to compare two variances
##
## data:  C and W
## F = 0.12605, num df = 6, denom df = 8, p-value = 0.0211
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.02709774 0.70583483
## sample estimates:
## ratio of variances
##          0.1260504
```

exclude 1

R: two sample t-test

```
C<-c(10,11,12,11,10,11,11) # size for captive fish
W<-c(7,8,11,12,10,13,11,10,12) # size for wild fish
(var(C)/7+var(W)/9)^2/( (var(C)/7)^2/6+(var(W)/9)^2/8)
```

$\sigma_1^2 \neq \sigma_2^2$

← calculate df for t-test

```
## [1] 10.43763
```

```
t.test(C,W,var.equal=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  C and W
## t = 0.59091, df = 10.438, p-value = 0.5672
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.134663  1.960060
## sample estimates:
## mean of x mean of y
##  10.85714  10.44444
```

Both fail to reject H_0.
No evidence to indicate unequal means.

Regression parameter testing in LR

- Simple linear regression with a dummary variable

$$Y = \beta_0 + \beta_1 I_B + \epsilon$$

$I_B = \begin{cases} 1 & \text{if at level } B \\ 0 & \text{if at level } A \end{cases}$

reference level.

  - Tesing whether $\beta_1 = \mu_B - \mu_A = 0$

$$t = \frac{(\bar{Y}_b - \bar{Y}_a) - 0}{\sqrt{MSE/S_{XX}}} \sim_{H_0} t_{n-2}, k = 0, 1, 2, .., p.$$

lose 2 d.f. since need to estimate β_0, β_1 in $E(Y) = \beta_0 + \beta_1 I_B$

- Multiple linear regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \ldots + \beta_p X_p + \epsilon, \epsilon \sim N(0, \sigma^2)$$

  - Tesing whether $\beta_k = 0$ or not.

$\# X = p$, take into account of β_0, $\Rightarrow p+1$

$$t = \frac{b_k - 0}{S(b_k)} \sim_{H_0} t_{n-p-1}, k = 0, 1, 2, .., p.$$

where S_k^2 is the $(k+1)$-th element on diagonal of estimated covariance matrix $\text{MSE}(X'X)^{-1}$.

Regression with Dummy Variable

- We will focus on the case where the predictor variable X is binary.
- Example: Height and gender

Sex	Height (cm)
male	183
male	178
female	172
female	154
female	159
female	165

X Y

$\longrightarrow$

iFemale	Height (cm)
0	183
0	178
1	172
1	154
1	159
1	165

Y has 2 subgroups by X.

Regression with Dummy Variable

- For a categorical variable with two levels, we need only one indicator variable
  - Continue the example: X takes value 1 for female and "0"for male.
- For a categorical variable with K levels, we need $K-1$ indicator variables.
  - For a variable takes value 1, 2 or 3.

$$X_1 = \begin{cases} 1, & X = 2 \\ 0, & X = 1 \ or \ 3 \end{cases} \quad , \quad X_2 = \begin{cases} 1, & X = 3 \\ 0, & X = 1 \ or \ 2 \end{cases}$$

  - Reference level is 1,this occurs when $X_1 = 0$ and $X_2 = 0$.

Regression with Dummy Variable

The Example of height vs gender, model

$$Y_i = \beta_0 + \beta_1 X + \epsilon_i$$

- Assume assumptions hold for this model
- X here is the female indicator

$$X = \begin{cases} 1, & Female \\ 0, & Male \end{cases}$$

- The slope β_1 gives the true difference in means between males and females, which can be estimated by b_1 from data.
- Proof:
  - $E(Y|female) = E(Y|X == 1) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$
  - $E(Y|male) = E(Y|X == 0) = \beta_0 + \beta_1 \times 0 = \beta_0$
  - $\beta_1 = E(Y|female) - E(Y|male) = E(Y|X == 1) - E(Y|X == 0)$

$\hat{\beta}_1 = b_1 = \bar{Y}_{female} - \bar{Y}_{male}$

Example: height vs gender (Lin. Regression Approach)

```r
X = c( rep(0,2),rep(1,4))        # X==1 for female, X==0 for male
Y= c( 183,178,172,154,159,165)   # height (cm) of a subject
summary(lm(Y~X))                 # summary of a SLR model
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##    1    2    3    4    5    6
##  2.5 -2.5  9.5 -8.5 -3.5  2.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  180.500      4.918  36.701 3.29e-06 ***
## X            -18.000      6.023  -2.988   0.0404 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.955 on 4 degrees of freedom
## Multiple R-squared:  0.6906, Adjusted R-squared:  0.6133
## F-statistic:  8.93 on 1 and 4 DF,  p-value: 0.0404
```

=162.5−180.5

```r
c(mean(Y[X==0]), mean(Y[X==1]))
```

```
## [1] 180.5 162.5
```

Example: height vs gender (two sample t-test)

Two sample t-test

```r
Ym=c(183,178); Yf=c(172,154,159,165)
t.test(Yf,Ym,var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  Yf and Ym
## t = -2.9883, df = 4, p-value = 0.0404
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -34.723617  -1.276383
## sample estimates:
## mean of x mean of y
##     162.5     180.5
```

Example: height vs gender (ANOVA Approach)

ANOVA result

```r
X = c( rep(0,2),rep(1,4))        # X==1 for female, X==0 for male
Y= c( 183,178,172,154,159,165)   # height (cm) of a subject
anova(lm(Y~X))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value Pr(>F)
## X          1  432.0  432.00  8.9302 0.0404 *
## Residuals  4  193.5   48.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$= (-2.9883)^2$

$$t = \frac{N(0,1)}{\sqrt{\chi^2_n/n}}$$

$$\Downarrow$$

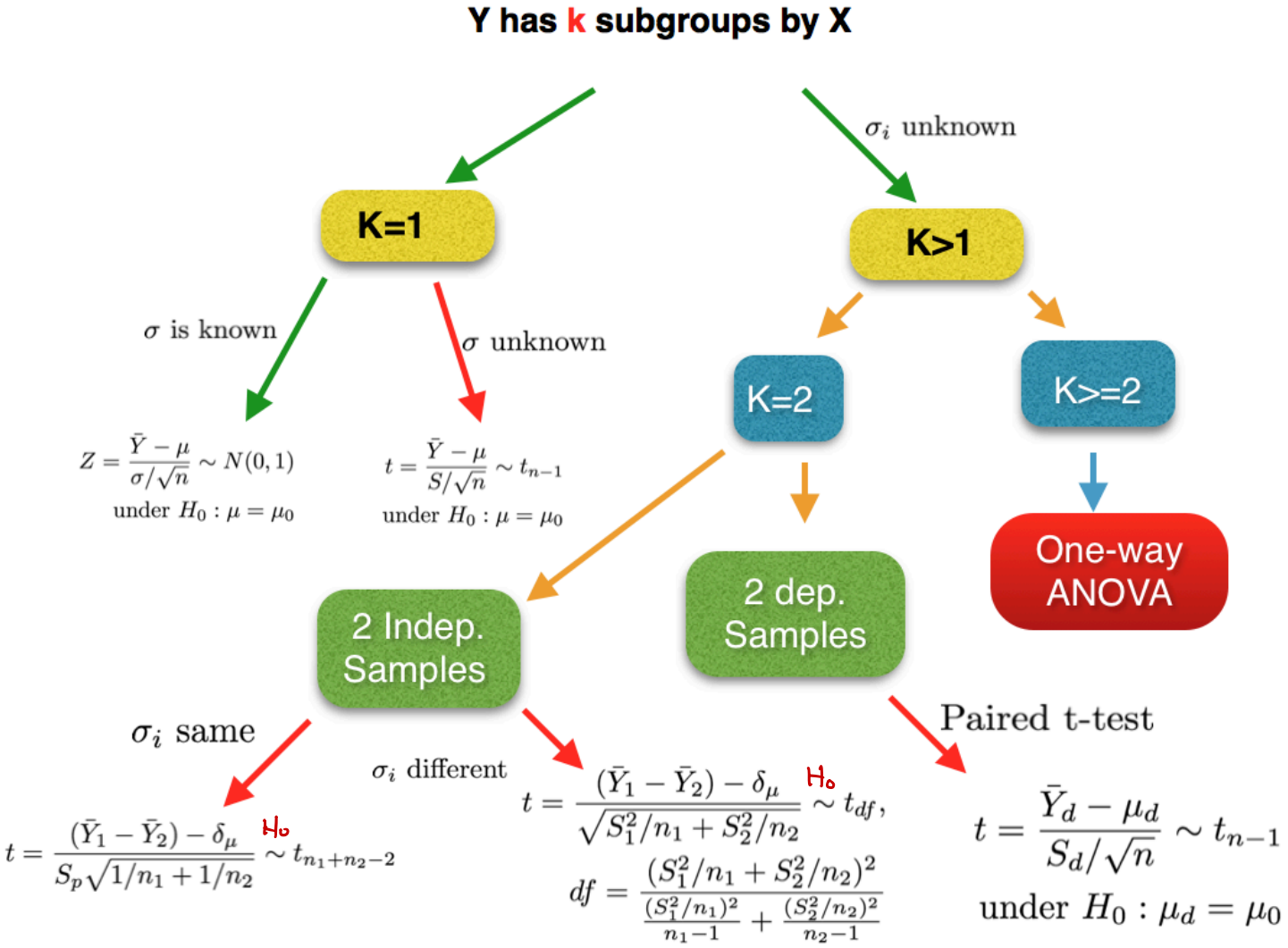
$$t^2 = \frac{N(0,1)^2/1}{\chi^2_n/n} = \frac{\chi^2_1/n}{\chi^2_n/n} = F_{1,n}$$

* For two sample t-test, equivalent approaches

① LR
② two-sample t-test
③ ANOVA

Summary of all t-test

- Y is the variable of interest. Y has K subgroups by X.
- X is a factor variable. e.g. X is gender or blood type information.



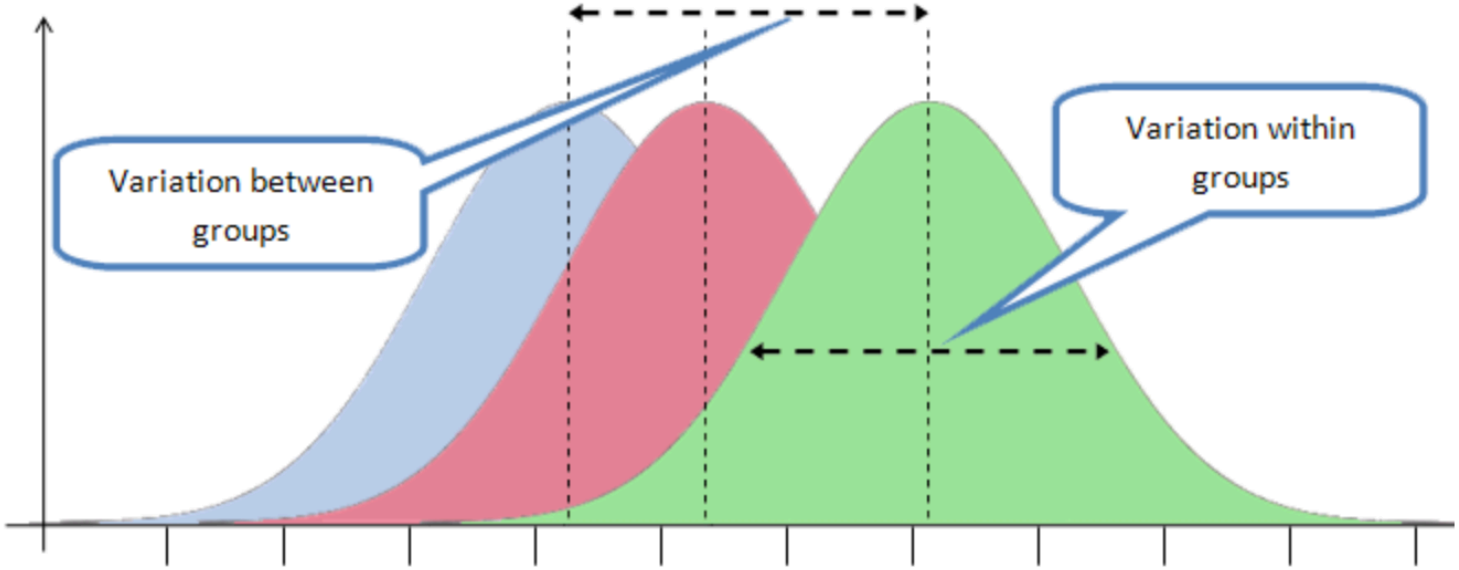
Summary of all t-test

	Type of Data		
Goal	Gaussian	Non-Gaussian	Binomial
Compare one group to a hypothetical value	One sample t-test	Wilcoxon Test	Binomial Test
Compare two paired groups	Paired t-test	Wilcoxon Test	McNemar's Test
Compare two unpaired groups	Two sample t-test	Wilcoxon-Mann-Whitney Test	Chi-Square or Fisher's Exact Test

Nonparametric fallbacks

- Independent samples $\rightarrow$ Mann-Whitney/Wilcoxon rank-sum test, R: **wilcox.test()** (with paired=FALSE)
- Paired t-test $\rightarrow$ Wilcoxon signed-rank test, R: **wilcox.test()** (with paired=TRUE)
- Kruskal Wallis Test for One Way Anova by Ranks. R: **kruskal.test(y~A)** (where y is numeric and A is a factor)
- Resampling method, such as randomization distribution (STA305)

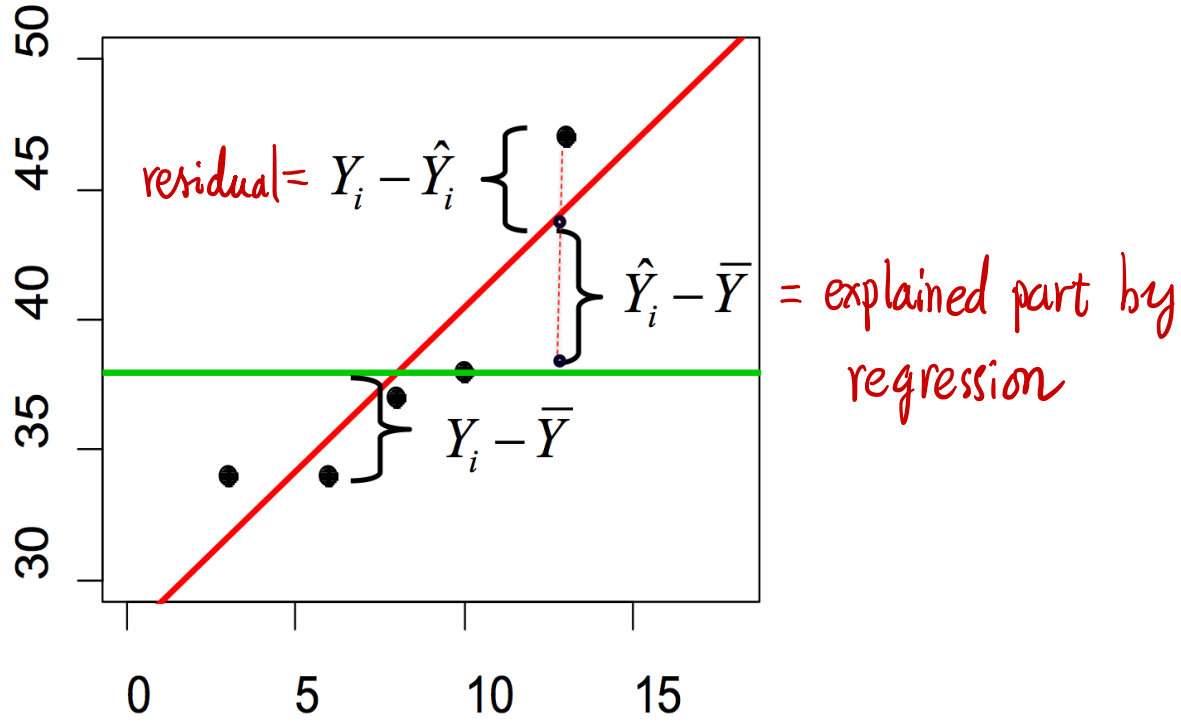
Take a break, and see you on Thursday



One-way ANOVA

Basic Framework of ANOVA

Review: ANOVA in Linear regression model



$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}), \quad \hat{Y}_i = X_i\beta$$

$$\underbrace{\sum_i (Y_i - \bar{Y})^2}_{SS_{Tot}} = \underbrace{\sum_i (Y_i - \hat{Y}_i)^2}_{SSE} + \underbrace{\sum_i (\hat{Y}_i - \bar{Y})^2}_{SSR}$$

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

$i=1$ $\quad Y_{11}, Y_{12}, Y_{13}, \cdots Y_{1n_1} \rightarrow$ Grp 1 : n_1

$i=2$ $\quad Y_{21}, Y_{22}, Y_{23}, \cdots Y_{2n_2} \rightarrow$ Grp 2 : n_2

$i=3$ $\quad Y_{31}, Y_{32}, Y_{33}, \cdots \cdots Y_{3n_3} \rightarrow \cdots$

$\vdots$

$i=r$ $\quad Y_{r1}, Y_{r2}, Y_{r3}, \cdots \cdots Y_{rn_r} \rightarrow \cdots$

$\Rightarrow$

Y	Treatment
Y_1	1
Y_2	2
Y_3	1
Y_4	3
$\vdots$	$\vdots$

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

$\hookrightarrow$ for treatment level, j = replication for i^{th} treatment.

- 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: $n = \sum_{i=1}^{r} n_i = n_1 + n_2 + \cdots + n_r$
- Denote $\bar{Y}_{i.}$ be i-the level mean and $\bar{Y}_{..}$ be the grand mean.

S_i^2 = sample variance for i^{th} group

$$= \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i.})^2$$

ANOVA: SS break down

- $\bar{Y}_{..}$ = grand mean = $\frac{\text{sum of all}}{n}$ $= \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} Y_{ij}}{n}$

- Total Sum of Squares:

$$SS = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2$$

- $\bar{Y}_{i.}$ = i^{th} group mean $= \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$

$\Downarrow$

$\boxed{\sum_{j=1}^{n_i} Y_{ij} = n_i \bar{Y}_{i.}}$

- Sum of Squares within groups (for error):

$$SSE = SS_W = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 = \sum_{i=1}^{r}(n_i - 1)S_i^2$$

- Sum of Squares between groups (for treatments):

$$SST = SS_B = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2 = \sum_{i=1}^{r} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

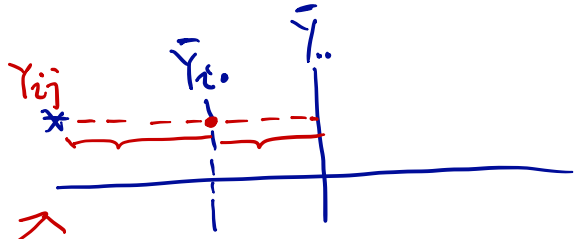
ANOVA: SS break down

Show $SS = SS_E + SS_T$ (in some books: $SS = SS_W + SS_B$)

Proof:

$$SS = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\underbrace{Y_{ij} - \bar{Y}_{i.} + \bar{Y}_{i.} - \bar{Y}_{..}})^2$$



$$= \sum_{i=1}^{r}\sum_{j=1}^{n_i}\left\{(Y_{ij} - \bar{Y}_{i.})^2 + 2(Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{i.} - \bar{Y}_{..}) + (\bar{Y}_{i.} - \bar{Y}_{..})^2\right\}$$

$$= \sum_{i=1}^{r}\underbrace{\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2} + \sum_{i=1}^{r}(\bar{Y}_{i.} - \bar{Y}_{..})\underbrace{\left[\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})\right]}$$

$$+ \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$= \left(\sum_{j=1}^{n_i} Y_{ij}\right) - n_i\bar{Y}_{i.}$$

$$= \left(\sum_{j=1}^{n_i} Y_{ij}\right) - \left(\sum_{j=1}^{n_i} Y_{ij}\right)$$

$$= 0$$

$$= \sum_{i=1}^{r}(n_i - 1)S_i^2 + \sum_{i=1}^{r} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$= \underset{\substack{\parallel \\ SS_E}}{SS_W} + \underset{\substack{\parallel \\ SS_T}}{SS_B}$$

ANOVA: SS within groups

Lemma 1: $\frac{SS_E}{\sigma^2} \sim \chi^2_{n-r}$ and $E\left(\frac{SS_E}{n-r}\right) = \sigma^2$

Proof: • $SS_E = SS_W = \sum_{i=1}^{r} (n_i - 1) S_i^2$, $\quad Y_{ij} \sim N(\mu_i, \sigma^2)$

• $\frac{(n_i-1) S_i^2}{\sigma^2} \sim \chi^2_{n_i-1}$

$$\Rightarrow \frac{SS_E}{\sigma^2} = \sum_{i=1}^{r} \frac{(n_i-1) S_i^2}{\sigma^2} = X_1 + X_2 + \cdots + X_r, \quad X_i \overset{\text{indep.}}{\sim} \chi^2_{n_i-1}$$

$$\sim \chi^2_{\sum_{i=1}^{r}(n_i-1)} = \chi^2_{(n-r)}$$

$$\boxed{X \sim \chi^2_n \Rightarrow E(X) = n, \quad Var(X) = 2n}$$

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-r} \Rightarrow E\left(\frac{SSE}{\sigma^2}\right) = n-r \Leftrightarrow E\left(\frac{SSE}{n-r}\right) = \sigma^2$$

ANOVA: SS_T is Independent of SS_W

Lemma 2: $SS_E \perp SS_T$

- For group i: $\bar{Y}_{i\cdot} \perp\!\!\!\perp S_i^2$
- different groups: independent $\Rightarrow$ $S_i^2 \perp\!\!\!\perp \bar{Y}_{j\cdot}$, $i \neq j$

$\Rightarrow$ $S_i^2 \perp\!\!\!\perp$ all sample group means $(\bar{Y}_{1\cdot}, \bar{Y}_{2\cdot}, \cdots \bar{Y}_{r\cdot})$

$\Rightarrow$ $S_i^2 \perp\!\!\!\perp \bar{Y}_{\cdot\cdot}$ since $\bar{Y}_{\cdot\cdot} = \frac{1}{r}(\bar{Y}_{1\cdot} + \bar{Y}_{2\cdot} + \cdots + \bar{Y}_{r\cdot})$

$\Downarrow$ $\qquad\qquad$ $\Downarrow$

$SS_E = SS_W = \sum_{i=1}^{r} (n_i - 1) S_i^2$ $\quad \perp\!\!\!\perp \quad$ $SS_T = SS_B = \sum_{i=1}^{r} n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$

So $SS_E \perp\!\!\!\perp SS_T$

$\Rightarrow$ $SS = SS_E +_{ind} SS_T$

ANOVA: SS within groups

Show $E\left(\frac{SS_E}{n-r}\right) = \sigma^2$

proof:

$$SS_E = \sum_{i=1}^{r} (n_i - 1) S_i^2$$

$$\Rightarrow \frac{SSE}{\sigma^2} = \sum_{i=1}^{r} \boxed{\frac{(n_i-1)S_i^2}{\sigma^2}} \sim \chi^2_{n_i-1}$$

$$\Rightarrow \frac{SSE}{\sigma^2} \sim \chi^2_{n-r}$$

$$\Rightarrow E\left(\frac{SSE}{\sigma^2}\right) = n-r$$

$$\Rightarrow E\left(\frac{SSE}{n-r}\right) = \sigma^2$$

,

$$\sum_{i=1}^{r} (n_i - 1)$$
$$= (n_1 - 1 + n_2 - 1 + \cdots + n_r - 1)$$
$$= (n_1 + n_2 + \cdots + n_r) - r$$
$$= n - r$$

ANOVA: total sum of squares under H_0

$SS/\sigma^2 \sim \chi^2_{n-1}$ under $H_0 : \mu_1 = \ldots = \mu_r$

- $SS = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2$

- under H_0: $\mu_1 = \mu_2 = \cdots = \mu_r = \mu$

$\Rightarrow \quad Y_{ij} \sim N(\mu, \sigma^2)$

$$\Rightarrow \quad \frac{\sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})^2}{n-1} = S^2_{pop}$$

$\hookrightarrow$ need to use $\bar{Y}_{..}$ to estimate μ

so $$\boxed{\frac{(n-1)\, S^2_{pop}}{\sigma^2} = \frac{SS}{\sigma^2} \sim \chi^2_{n-1}}$$

$$\underset{\chi^2_{n-1}}{\frac{SS}{\sigma^2}} = \underset{\chi^2_{n-r}}{\frac{SS_E}{\sigma^2}} + \text{ind } \underset{\chi^2_{r-1}}{\frac{SS_T}{\sigma^2}}$$

ANOVA: E(SS_T)

$$E(SS_T) \geq (r-1)\sigma^2$$

$$\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$$

$$\Downarrow$$

$$\sum_{i=1}^{r} (n_i \bar{Y}_{i\cdot}) = \sum_{i=1}^{r} \left(\sum_{j=1}^{n_i} Y_{ij}\right) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij} = n\bar{Y}_{\cdot\cdot}$$

$$\bar{Y}_{\cdot\cdot} = \frac{\sum_i \sum_j Y_{ij}}{n}$$

proof:

$$SS_T = \sum_i^r n_i (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^r n_i (\bar{Y}_{i\cdot}^2 - 2\bar{Y}_{i\cdot}\bar{Y}_{\cdot\cdot} + \bar{Y}_{\cdot\cdot})^2$$

$$= \sum_{i=1}^r n_i \bar{Y}_{i\cdot}^2 - 2\bar{Y}_{\cdot\cdot} \underbrace{\sum_{i=1}^r (n_i \bar{Y}_{i\cdot})} + \sum_{i=1}^r n_i \bar{Y}_{\cdot\cdot}^2,$$

$$= \sum_{i=1}^r n_i \bar{Y}_{i\cdot}^2 - 2n\bar{Y}_{\cdot\cdot}^2 + n\bar{Y}_{\cdot\cdot}^2$$

$$= \sum_{i=1}^r n_i \bar{Y}_{i\cdot}^2 - n\bar{Y}_{\cdot\cdot}^2$$

$$\boxed{E(SS_T) = \sum_{i=1}^r n_i E(\bar{Y}_{i\cdot}^2) - n E(\bar{Y}_{\cdot\cdot}^2)}$$

- $E(\bar{Y}_{i\cdot}^2) = Var(\bar{Y}_{i\cdot}) + \{E(\bar{Y}_{i\cdot})\}^2 \leftarrow \bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$

$$= \frac{1}{n_i^2} \cdot n_i \sigma^2 + \mu_i^2$$

Defn $\boxed{\bar{\mu} = \frac{\sum_{i=1}^r n_i \mu_i}{n}}$

- $E(\bar{Y}_{\cdot\cdot}^2) = Var(\bar{Y}_{\cdot\cdot}) + \{E(\bar{Y}_{\cdot\cdot})\}^2$

$$= \frac{1}{n^2} n\sigma^2 + \left(\frac{\sum_{i=1}^r n_i \mu_i}{n}\right)^2 = \frac{\sigma^2}{n} + \bar{\mu}^2$$

ANOVA: E(SS_T) (be contd)

So $E(SS_T) = \sum_{i=1}^{r} n_i E(\bar{Y}_{i\cdot}^2) - n E(\bar{Y}_{\cdot\cdot}^2)$

$$= \sum_{i=1}^{r} n_i \left(\frac{\sigma^2}{n_i} + \mu_i^2\right) - n\left(\frac{\sigma^2}{n} + \bar{\mu}^2\right)$$

$$= r\sigma^2 + \sum_{i=1}^{r} n_i \mu_i^2 - \sigma^2 - n\bar{\mu}^2, \quad n = \sum_{i=1}^{r} n_i$$

$$= (r-1)\sigma^2 + \sum_{i=1}^{r} (n_i \mu_i^2 - n_i \bar{\mu}^2)$$

$$= \underbrace{(r-1)\sigma^2}_{>0} + \underbrace{\sum_{i=1}^{r} n_i (\mu_i - \bar{\mu})^2}_{>0}$$ check it.

$$\geq (r-1)\sigma^2$$

under H_0

$E\left(\frac{SS_T}{r-1}\right) = \sigma^2$

so $E(SS_T) \geq (r-1)\sigma^2$, "=" when $\underbrace{\mu_i = \mu \text{ for all } i}_{H_0}$

Mean Sum of Squares (MS*)

H_0 : all the group means are equal. H_a : at least two groups have different means.

- $SS_E/\sigma^2 \sim \chi^2_{n-r}$
- $SS_T/\sigma^2 \sim_{H_0} \chi^2_{r-1}$
- $SS/\sigma^2 \sim_{H_0} \chi^2_{n-1}$

Define *Mean Sum of Squares*: $MS = SS*/df(SS*)$

- Mean Sums of Squares for total: $MS = SS/df(SS) = SS/(n-1)$
- Mean Sums of Squares for Treatment:
  $MST = SS_T/df(SS_T) = SS_T/(r-1), SS_T = SS_B$
- Mean Sums of Squares for Error:
  $MSE = SS_E/df(SS_E) = SS_E/(n-r), SS_E = SS_W$

ANOVA: Hypothesis of equal means

H_0 : True. $E\left(\frac{SS_T}{r-1}\right) = \sigma^2$

H_0 : T/F : $E\left(\frac{SS_E}{n-r}\right) = \sigma^2$

- **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**

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

$E(SST) \geq (r-1)\sigma^2$

$\Rightarrow E(MST) \geq \sigma^2$

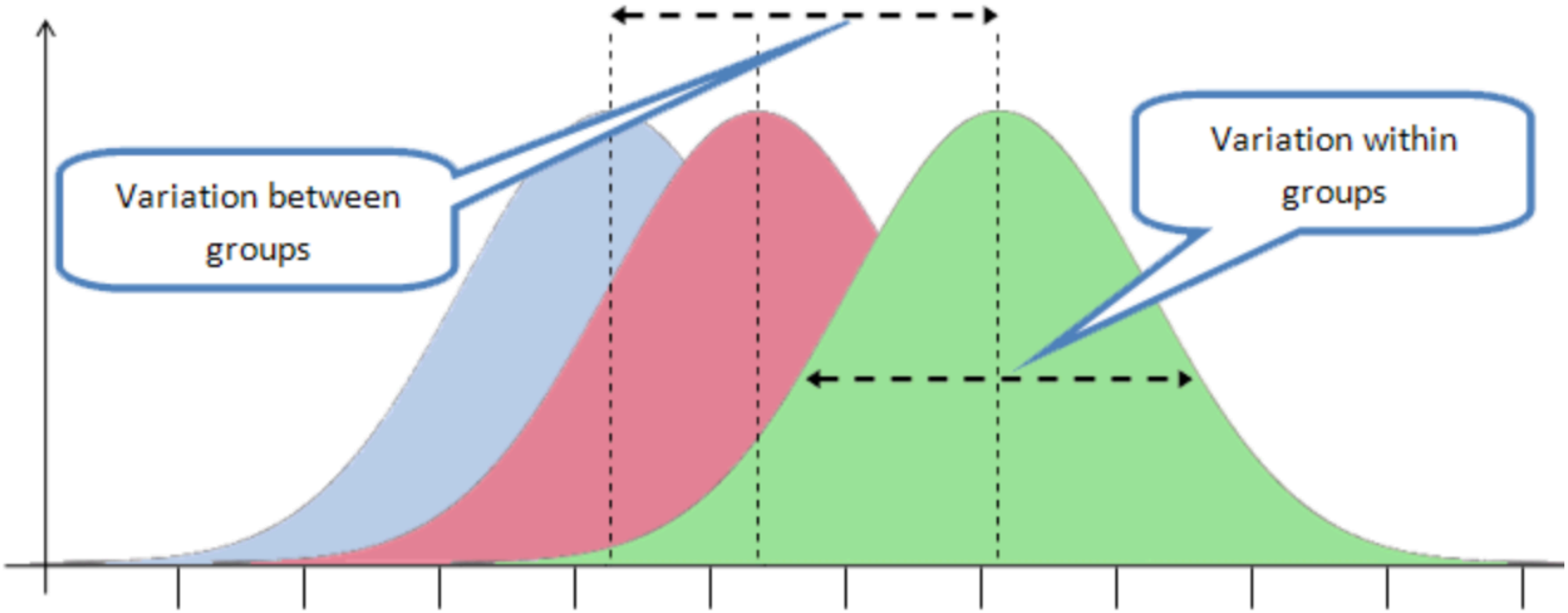
  - If H_0 is true, $F^* \approx 1$, we don't have evidence to reject H_0.
  - If H_0 is false, $MST > MSE$, F^* increases. So large values of observed F^* are evidence against H_0, and we test H_0 using a one-tailed test.

- **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



- Equal means $\Rightarrow$ $SS_B\ (=SS_T) \approx SS_E\ (=SS_W)$.
- unequal means $\Rightarrow$ $SS_B\ (=SS_T) > SS_E\ (=SS_W)$: evidence

One-way ANOVA Table

Source	df	SS	Mean Square	F
Treatments	r-1	SS_T	MST=SS_T/(r-1)	F=MST/MSE
Residual	n-r	SS_E	MSE	
Total	n-1	SS		

Model Equations for one-way ANOVA

- $Y_{ij} = \mu_i + \epsilon_{ij}$, (i=1,2,. . . , r, j =1,2,. . . , n_i)
- $\epsilon_{ij} \sim N(0, \sigma^2)$ $\Rightarrow$ $Y_{ij} \sim N(\mu_i, \sigma^2)$, check $E(Y_{ij})$ and $V(Y_{ij})$
- μ_i is the main effect for the i-th treatment.

After Lecture This Week

Practice problems

- Check if you receive A0 submission notice from Crowdmark or not, if not, write an email to instrutor.
- Try the posted posted practice problem (solution is available too)
- Try all the R example in slides.
- Review all types of t-test and one-way ANOVA analysis.

Topics for next week:

- R Example: one-way ANOVA
- Discussion about one-way ANOVA
- Two-way ANOVA