Spring 2021
SKKU Biostats and Big data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Lecture 23
Analysis of Variance (ANOVA)

Review: Key Points

Chapter 27: Inferences for Regression

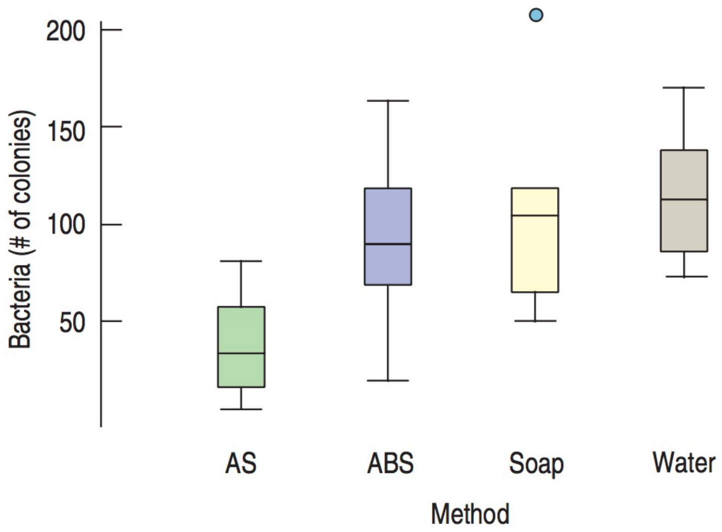
- $\mu_y = \beta_0 + \beta_1 x, y = \beta_0 + \beta_1 x + \varepsilon$
- Assumption and conditions:
  - Linearity Assumption, Equal Variance Assumption, Normal Population Assumption, Independence Assumption
- $SE(b_1) = \frac{s_e}{\sqrt{n-1}s_x}$, where $s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$, $s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$
- Hypothesis test: H_0: $\beta_1 = 0$, $t_{n-2} = \frac{b_1 - 0}{SE(b_1)}$
- 95% confidence interval for β: $b_1 \pm t^*_{n-2} \times SE(b_1)$
- Standard errors for predicted values: $\hat{y}_\nu \pm t^*_{n-2} \times SE$
  - Mean: $SE(\hat{\mu}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n}}$, Individual: $SE(\hat{y}_\nu) = \sqrt{SE^2(b_1) \times (x_\nu - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$

Example: Washing hands, bacteria colonies

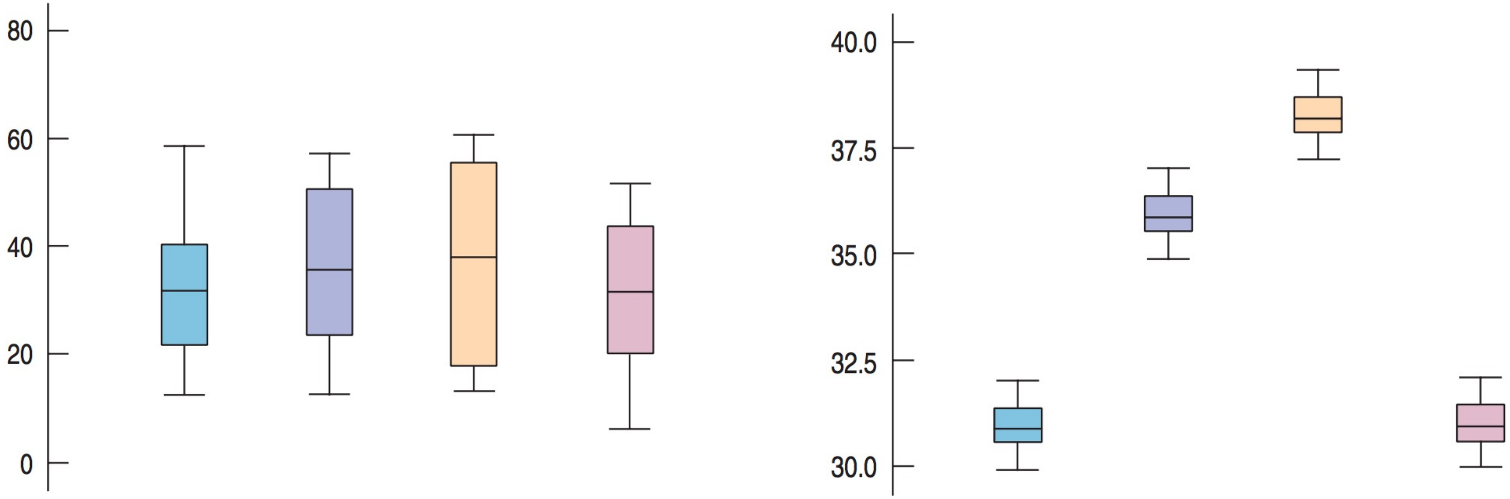
- How effective is washing hands with soap in eliminating bacteria?
- Comparing four different methods: with water only (Water), regular soap (Soap), antibacterial soap (ABS), spraying hands with antibacterial spray (AS)
- Question: *are there differences?*

Bacteria colonies

	Alcohol	AB Soap	Soap	Water
	51	70	84	74
	5	164	51	135
	19	88	110	102
	18	111	67	124
	58	73	119	105
	50	119	108	139
	82	20	207	170
	17	95	102	87
Treatment Means	37.5	92.5	106	117



Intuitions in testing whether the means of several groups are equal



- Which one does seem to have same vs. different means?
- The mean values in both figures are actually the same. Why do they look different?
- In the second figure, the variation *within* each group is so small that the differences *between* the means **stand out.**
    - Comparing the differences *between* groups with the variation *within* the groups: *F*-test!

Differences *between* and *within* groups

- Our goal is to compare two variances, one for *between* groups, and one for *within* group.
- We use the fact we learned in a previous lecture:

> ❖ For quantitative data, sample mean, $\bar{y}$
>
> - $Mean(\bar{y}) = \mu, SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$

- We will calculate σ in two different methods and compare: one using $\bar{y}$ (between), and the other using all the data (within).

$\bar{y}$

Level	n	Mean
Alcohol spray	8	37.5
Antibacterial soap	8	92.5
Soap	8	106.0
Water	8	117.0

$Var(\bar{y})$

$\frac{\sigma^2}{n}$

	Alcohol	AB Soap	Soap	Water
all data	51	70	84	74
	5	164	51	135
	19	88	110	102
	18	111	67	124
	58	73	119	105
	50	119	108	139
	82	20	207	170
	17	95	102	87
Treatment Means	37.5	92.5	106	117

s^2_{pooled}

σ^2

Differences *between* and *within* groups

- $n \times Var(\bar{y})$: Between Mean Square (MS_T), Treatment Mean Square
- s_{pooled}^2: Within Mean Square (MS_E), Error Mean Square
- We will use the ratio of these two (MS_T/MS_E) as a test!

$\bar{y}$

Level	n	Mean
Alcohol spray	8	37.5
Antibacterial soap	8	92.5
Soap	8	106.0
Water	8	117.0

$Var(\bar{y})$

$\frac{\sigma^2}{n}$

	Alcohol	AB Soap	Soap	Water
all data	51	70	84	74
	5	164	51	135
	19	88	110	102
	18	111	67	124
	58	73	119	105
	50	119	108	139
	82	20	207	170
	17	95	102	87
Treatment Means	37.5	92.5	106	117

s_{pooled}^2

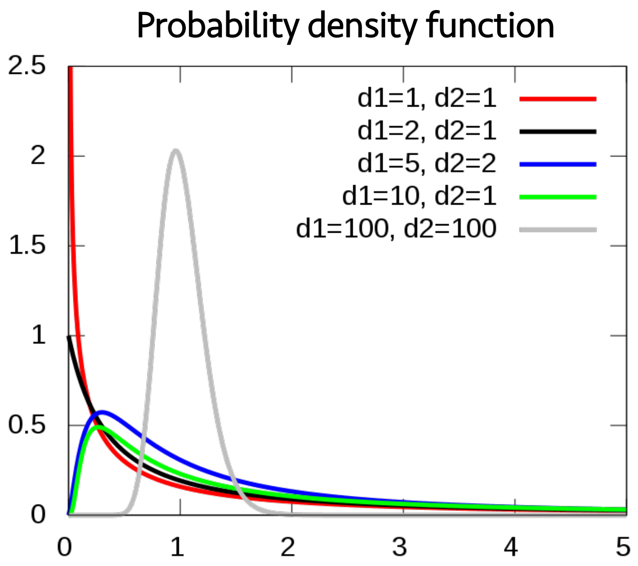
σ^2

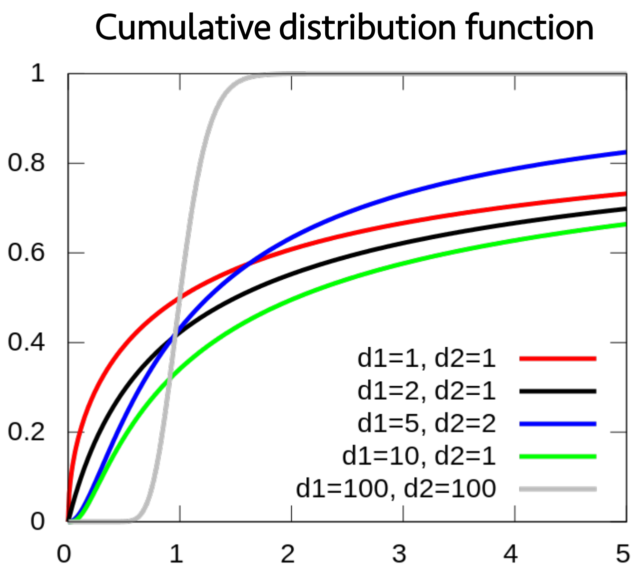
Hypothesis testing

- Null hypothesis:
  - H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$

- $MS_T/MS_E = 1$, if the null hypothesis is true
- MS_T/MS_E follows ***F*-distribution**
  - depends on *two* degrees of freedom
  - numerator *df* from MS_T: *k*-1
  - denominator *df* from MS_E: *k*(*n*-1) = *N*-*k*
    - Example data:
    - $df_1 = k-1 = 3$, $df_2 = N-k = 32-4 = 28$

- ***F*-test**: one-tailed test for the MS_T/MS_E ratio
  - also Analysis of Variance (**ANOVA**)





https://en.wikipedia.org/wiki/F-distribution

Quiz 23-1

https://forms.gle/bQoNGhr9UppVCtJx7

ANOVA table

- has a long history (nearly a century)

Analysis of Variance Table

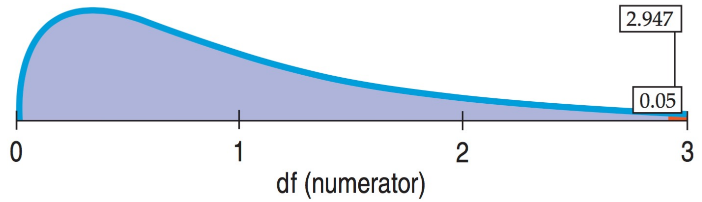
	Source	Sum of Squares	DF	Mean Square	F-ratio	P-value
Mean Square Between (Treatment): MS_T →	Method	29882 $k-1$	3	9960.64	7.0636	0.0011
Mean Square Within (Error, Residual): MS_E →	Error	39484 $N-k$	28	1410.14		
	Total	69366	31			

$n \times Var(\bar{y})$

s^2_{pooled}

?

- *F*-table



df (denominator)	1	2	3	4	5	6	7
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334
31	4.160	3.305	2.911	2.679	2.523	2.409	2.323
32	4.149	3.295	2.901	2.668	2.512	2.399	2.313

One-way ANOVA model

Analysis of Variance Table

Source	Sum of Squares	DF	Mean Square	F-ratio	P-value
Method	29882	3	9960.64	7.0636	0.0011
Error	39484	28	1410.14		
Total	69366	31			

- $y_{ij} = \mu_j + \varepsilon_{ij}$, where *i*: each data point, *j*: group
- H$_0$: $\mu_1 = \mu_2 = \cdots = \mu_k$
- $\mu_j = \mu + \tau_j$, where τ_j: the deviation from the grand mean (μ = mean of mean)
- We can rewrite our null hypothesis, H$_0$: $\tau_1 = \tau_2 = \cdots = \tau_k$
- $\hat{\tau}_j = \bar{y}_j - \bar{\bar{y}}$, where, $\bar{y}_j$ is the observed group mean, $\bar{\bar{y}}$ is the observed grand mean.
- $\varepsilon_{ij} = y_{ij} - \bar{y}_j$
- $y_{ij} = \underbrace{\bar{\bar{y}}}_{\mu} + \underbrace{(\bar{y}_j - \bar{\bar{y}})}_{\tau_j} + \underbrace{(y_{ij} - \bar{y}_j)}_{\varepsilon_{ij}}$
- *Observations = Grand mean + Treatment effect + Residual*

Let's look at the data

Source	Sum of Squares	DF	Mean Square	F-ratio	P-value
Method	29882	3	9960.64	7.0636	0.0011
Error	39484	28	1410.14		
Total	69366	31			

Analysis of Variance Table

- $y_{ij} = \underbrace{\bar{\bar{y}}}_{\mu} + \underbrace{(\bar{y}_j - \bar{\bar{y}})}_{\tau_j} + \underbrace{(y_{ij} - \bar{y}_j)}_{\varepsilon_{ij}}$

y_{ij}: *Observations*

	Alcohol	AB Soap	Soap	Water
	51	70	84	74
	5	164	51	135
	19	88	110	102
	18	111	67	124
	58	73	119	105
	50	119	108	139
	82	20	207	170
	17	95	102	87
Treatment Means	37.5	92.5	106	117

$\bar{y}_j$

=

$\bar{\bar{y}}$: *Grand mean*

Alcohol	AB Soap	Soap	Water
88.25	88.25	88.25	88.25
88.25	88.25	88.25	88.25
88.25	88.25	88.25	88.25
88.25	88.25	88.25	88.25
88.25	88.25	88.25	88.25
88.25	88.25	88.25	88.25
88.25	88.25	88.25	88.25
88.25	88.25	88.25	88.25

+

$\bar{y}_j - \bar{\bar{y}}$: *Treatment effect*

Alcohol	AB Soap	Soap	Water
−50.75	4.25	17.75	28.75
−50.75	4.25	17.75	28.75
−50.75	4.25	17.75	28.75
−50.75	4.25	17.75	28.75
−50.75	4.25	17.75	28.75
−50.75	4.25	17.75	28.75
−50.75	4.25	17.75	28.75
−50.75	4.25	17.75	28.75

+

$y_{ij} - \bar{y}_j$: *Residual*

Alcohol	AB Soap	Soap	Water
13.5	−22.5	−22	−43
−32.5	71.5	−55	18
−18.5	−4.5	4	−15
−19.5	18.5	−39	7
20.5	−19.5	13	−12
12.5	26.5	2	22
44.5	−72.5	101	53
−20.5	2.5	−4	−30

Sum of Squares of all these values

$$SS_T = \sum\sum(\bar{y}_j - \bar{\bar{y}})^2$$

Treatment Sum of Squares

Sum of Squares of all these values

$$SS_E = \sum\sum(y_{ij} - \bar{y}_j)^2$$

Error Sum of Squares

- $MS_T = \frac{SS_T}{k-1}$, $MS_E = \frac{SS_E}{N-k}$, $F_{k-1,N-k} = \frac{MS_T}{MS_E}$

Quiz 23-2

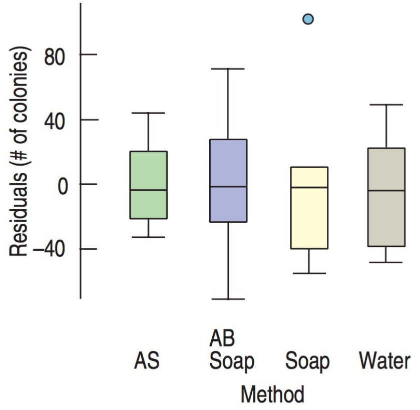
https://forms.gle/cdzfzo2eGbq1DvYm7

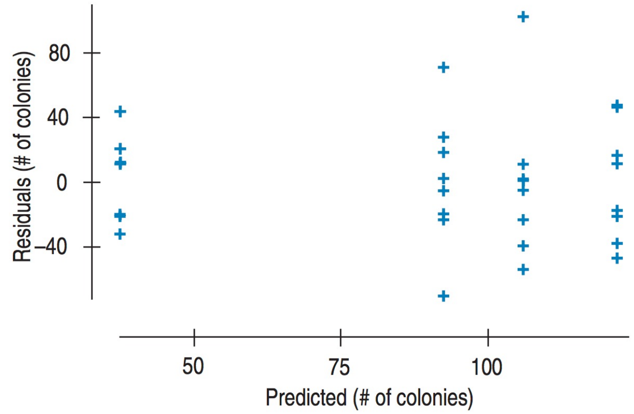
Quiz 23-3

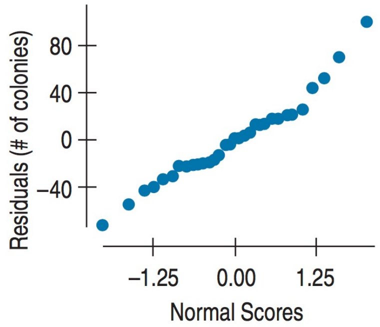
https://forms.gle/gwrmqwzfrU9cLiqd6

Assumptions and Conditions

- Independence Assumption
  - 1. Groups must be independent of each other.
  - 2. Data within each treatment group must be independent.
  - Randomization Condition
- Equal Variance Assumption
  - Similar Spread Condition
- Normal Population Assumption
  - Nearly Normal Condition
  - Normal probability plot (using all the residuals)







Comparing means

- Pooled *t*-test
- But a little tweaks:
    - s_{p} : calculated based on the whole group.
    - *df: N-k*
- Otherwise, the same.

- H_0: $\mu_W - \mu_{ABS} = 0$
- $SE(\mu_W - \mu_{ABS}) = s_{\mathrm{p}}\sqrt{\frac{1}{n_W} + \frac{1}{n_{ABS}}}$
- $df = N - k$

Pooled *t*-test

- This is simpler than two-sample *t*-test, but has a big assumption
    - "The variances of the two groups are the same."
    - Advantages:
        - This has a large degrees of freedom than two-sample t-test.
        - Simpler formula for degrees of freedom
    - Disadvantages:
        - The assumption of equal variances is a strong one, and is often not true, and difficult to check.

- $s^2_{\text{pooled}} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$
- $SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}} = s_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- df = n_1+n_2-2

Correction for multiple comparisons

- If you do multiple tests, the error rate adds up!
- When, α = 0.05, if you have 10 tests, the error rate will be $1-(1-0.05)^{10}=0.4013$
- That's not ideal. We still want to control the total error rate under α!

- Family-wise error rate (FWER)
  - There are many methods to control the FWER $\leq \alpha$.
  - The most popular one: **Bonferroni**
    - $\alpha_{FWER} \; = \; 1-(1-\alpha_{each\ test})^m \leq m \cdot \alpha_{each\ test}$ (Boole's inequality), *m*: # tests
    - Therefore, $\alpha_{each\ test} = \frac{\alpha_{FWER}}{m}$

Key Points

Chapter 28: Analysis of Variance

- Based on the comparison between variances *between* vs. *within* groups: $n \cdot Var(\bar{y})$ vs. s^2_{pooled}
- MS_T/MS_E follows the *F*-distribution with *k*-1 and *N*-*k* as two degrees of freedom
- Another way of presenting ANOVA
  - *Observations = Grand mean + Treatment effect + Residual*
  - $y_{ij} = \underbrace{\bar{\bar{y}}}_{\mu} + \underbrace{(\bar{y}_j - \bar{\bar{y}})}_{\tau_j} + \underbrace{(y_{ij} - \bar{y}_j)}_{\varepsilon_{ij}}$
  - Treatment Sum of Squares: $SS_T = \sum\sum(\bar{y}_j - \bar{\bar{y}})^2$
  - Error Sum of Squares: $SS_E = \sum\sum(y_{ij} - \bar{y}_j)^2$
  - $MS_T = \frac{SS_T}{k-1}, \quad MS_E = \frac{SS_E}{N-k}, \quad F_{k-1,N-k} = \frac{MS_T}{MS_E}$
- Comparing means between two groups using Pooled *t*-test with s_p (based on the whole group) and d*f* = *N*-*k*