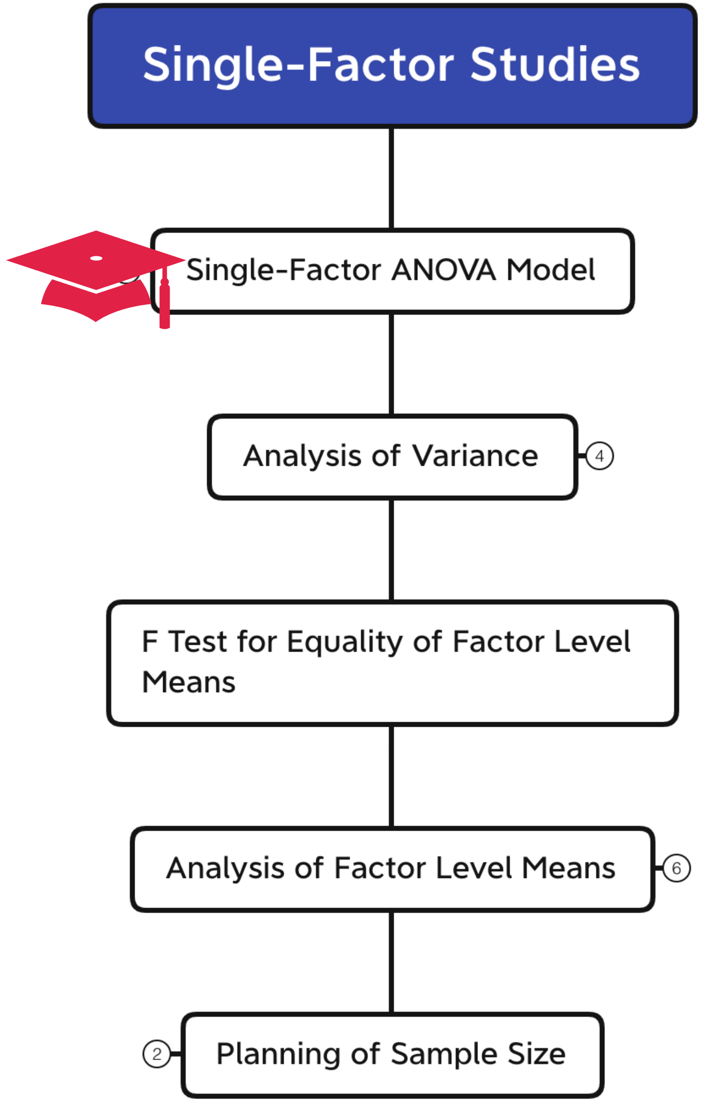
Lecture 2: Single-Factor Studies

STA 106: Analysis of Variance

Suggested reading: ALSM Chapter 16 & 17

Xiner Zhou
Department of Statistics, University of California, Davis


Single-Factor Studies
Single-Factor ANOVA Model
Analysis of Variance
4
F Test for Equality of Factor Level Means
Analysis of Factor Level Means
6
2
Planning of Sample Size

Example

(The Rehab Study)

The objective of the study:
A rehabilitation center researcher was interested in the relationship between
physical fitness prior to surgery of corrective knee surgery
Time required in physical therapy until successful rehabilitation

The study setup:
Patient records during the past year were examined: 24 male patients with age 18-30

		j									
i		1	2	3	4	5	6	7	8	9	10
1	Below Average	29	42	38	40	43	40	30	42 •		
2	Average	30	35	39	28	31	31	29	35	29	33
3	Above Average	26	32	21	20	23	22				

Substantive Research Questions of Interest:

- Whether the factor levels or treatments differ in terms of response?
- If the factor levels differ in terms of response, in what way do they differ or how do they differ?

These research questions lead to statistical questions usually performed in two steps, correspondingly.

Single-Factor ANOVA Model

r : number of factor levels

Y_{ij} : observed value of the outcome or response variable for the jth unit in ith factor level

Subscript $i = 1 \ldots r$: ith factor level

n_i : number of units in ith factor level

Subscript $j = 1 \ldots n_i$: jth unit in a given factor level

$n_T = \sum_{i=1}^{r} n_i$: total number of units in the study

Single-Factor ANOVA Model

Assume: observed value of response variable is the sum of two components

Treatment mean = factor level mean μ_i — fixed term

Error term ϵ_{ij} — random term

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- ϵ_{ij} error term

But not technically errors (mistakes), it reflects all other extraneous factors that influence the response but are not measured or not considered in current study

We assume $\epsilon_{ij}'s$ are independently and identically distributed as $N(0,\sigma^2)$, for all i, j

$$\epsilon_{ij} \sim N(0,\sigma^2)$$

Single-Factor ANOVA Model

Assume: observed value of response variable is the sum of two components

Treatment mean = factor level mean μ_i — fixed term

Error term ϵ_{ij} — random term

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- μ_i

$$E\left(Y_{ij}\right) = E\left(\mu_i + \varepsilon_{ij}\right) = \mu_i + E\left(\varepsilon_{ij}\right) = \mu_i$$

Mean response of ith factor level or treatment

Interpretation:

Experimental study

μ_i is the mean response that would be obtained if the ith treatment were applied to all units in the population under study

Observational study

μ_i is the mean response for ith factor level subpopulation

Single-Factor ANOVA Model

Assume: observed value of response variable is the sum of two components

Treatment mean = factor level mean μ_i — fixed term

Error term ϵ_{ij} — random term

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

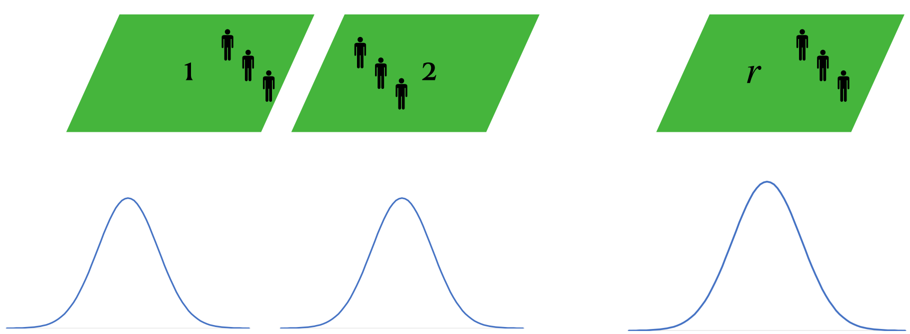
- What is the distribution of Y_{ij}?

$$Y_{ij} = \mu_i + \varepsilon_{ij} \sim N(\mu_i, \sigma^2)$$

Different treatment means depend on which treatment this unit is from

same variance of error term: homogeneity of error variance

Y_{ij} is a random draw from a normal population with mean μ_i and variance σ^2



r sub-populations corresponding to *r* treatments, each one follows a normal distribution with different means, but same variance

Single-Factor ANOVA Model

Assume: observed value of response variable is the sum of two components

Treatment mean = factor level mean μ_i — fixed term

Error term ϵ_{ij} — random term

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

- Unknown parameters

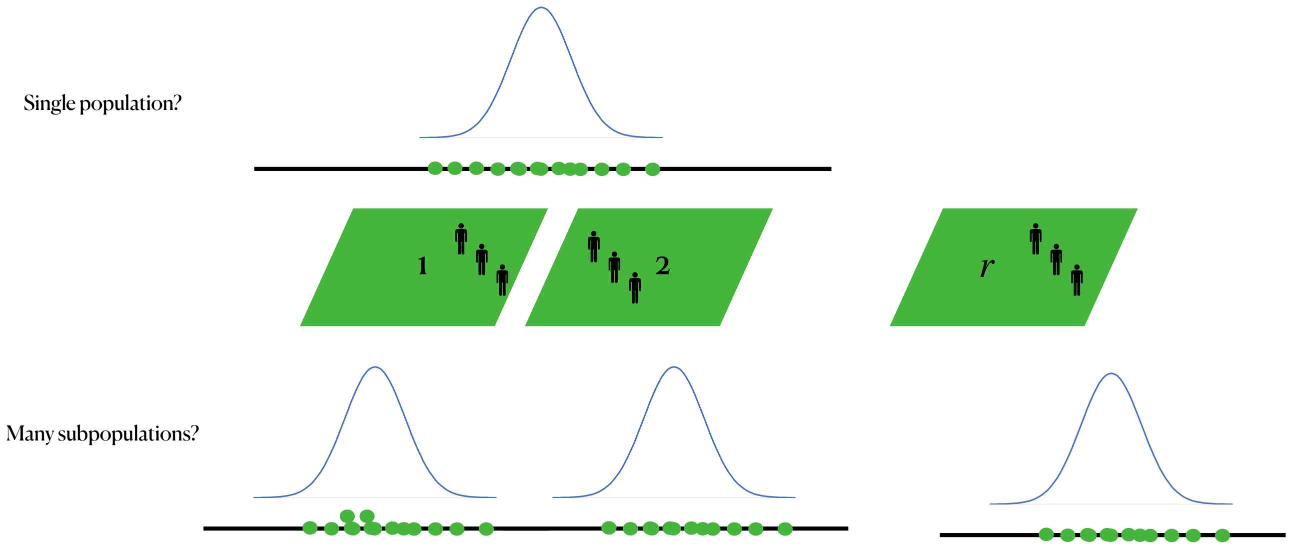
Treatment mean = factor level mean $\mu_1 \ldots \mu_r$

Error variance σ^2

➡ How do we estimate them?

Fitting the ANOVA Model (Estimate Model Parameters $\mu_1 \ldots \mu_r$)

Treatment mean = factor level mean $\mu_1 \ldots \mu_r$?



Fitting the ANOVA Model (Estimate Model Parameters $\mu_1 \dots \mu_r$)

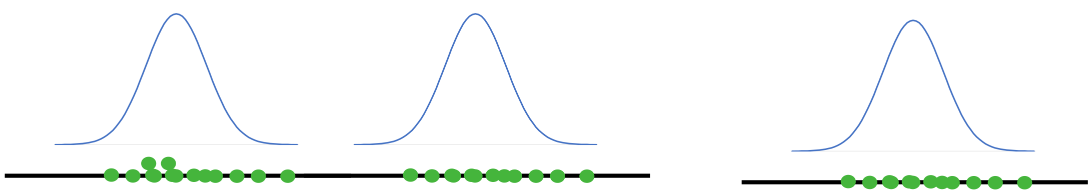
Treatment mean = factor level mean $\mu_1 \dots \mu_r$?

Least Squares Method:

The estimates for the centers of populations $\hat{\mu}_1 \dots \hat{\mu}_r$ should minimize the dispersion in the data so that each observation Y_{ij} is as close as possible to its corresponding mean μ_i

$$Q(\mu_1 \dots \mu_r) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{j=1}^{n_1} (Y_{ij} - \mu_1)^2 + \sum_{j=2}^{n_i} (Y_{ij} - \mu_2)^2 + \dots + \sum_{j=1}^{n_r} (Y_{ij} - \mu_r)^2$$

How to measure dispersion in the data: sum of squared deviations of observations away from its population means



$$\hat{\mu}_1 = \text{arg min}_{\mu_1} Q(\mu_1 \dots \mu_r) \quad \hat{\mu}_2 = \text{arg min}_{\mu_2} Q(\mu_1 \dots \mu_r) \quad \hat{\mu}_r = \text{arg min}_{\mu_r} Q(\mu_1 \dots \mu_r)$$

Fitting the ANOVA Model (Estimate Model Parameters $\mu_1 \ldots \mu_r$)

Treatment mean = factor level mean $\mu_1 \ldots \mu_r$?

Least Squares Method:

$\hat{\mu}_1 = \arg\min_{\mu_1} Q(\mu_1 \ldots \mu_r)$ $\quad\hat{\mu}_2 = \arg\min_{\mu_2} Q(\mu_1 \ldots \mu_r)$ $\quad\hat{\mu}_r = \arg\min_{\mu_r} Q(\mu_1 \ldots \mu_r)$

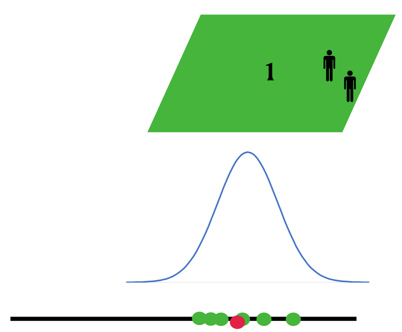
Why it's a simpler than it looks?

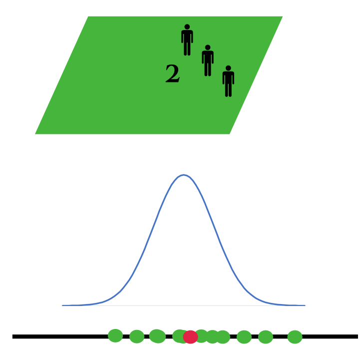
When minimizes with respect to μ_1, only the first term matters, other terms are constants

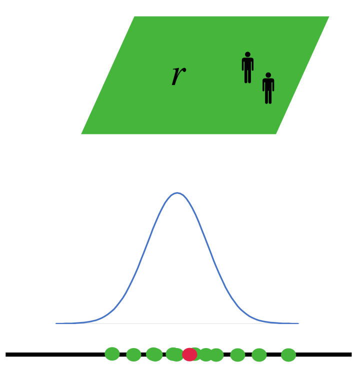
$$\frac{dQ}{d\mu_1} = \frac{d}{d\mu_1}\sum_{j=1}^{n_1}(Y_{ij} - \mu_1)^2 = -2\sum_{j=1}^{n_1}(Y_{ij} - \mu_1) = 0$$

Least squares estimates (LSE):

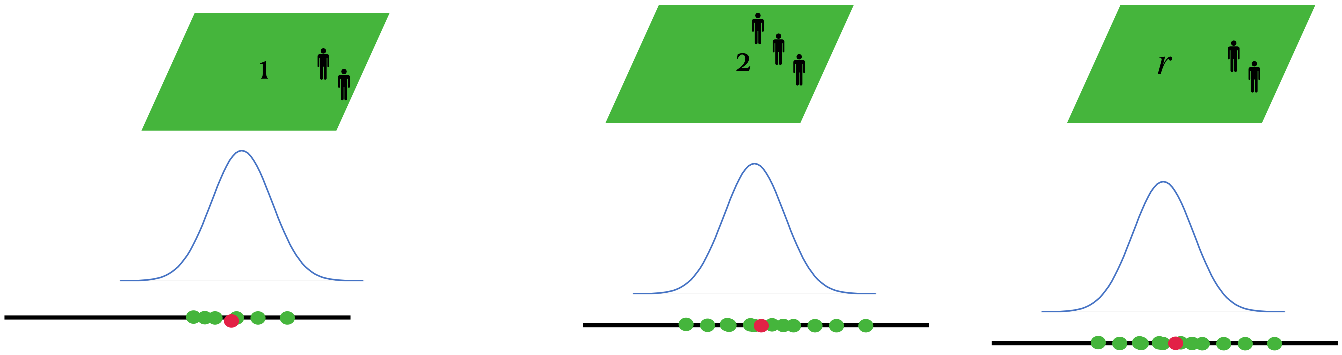
$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n_1} Y_{1j}}{n_1} = \overline{Y_{1.}} \qquad \hat{\mu}_2 = \frac{\sum_{i=1}^{n_2} Y_{2j}}{n_2} = \overline{Y_{2.}} \qquad \hat{\mu}_r = \frac{\sum_{i=1}^{n_r} Y_{rj}}{n_r} = \overline{Y_{r.}}$$







Fitting the ANOVA Model (Estimate Model Parameters $\mu_1 \ldots \mu_r$)



- fitted value for an observation Y_{ij}

  ANOVA model's "best guess" or "best prediction" for Y_{ij}

  $\hat{Y}_{ij} = \hat{\mu}_i = \overline{Y_{i.}}$

- residual e_{ij} corresponds to observation Y_{ij} is

$$e_{ij} = \underbrace{Y_{ij} - \hat{Y}_{ij}} = Y_{ij} - \bar{Y}_{i.}$$

Difference between observed value and fitted value which is estimated factor level mean

Residuals are approximations or estimation for the error term

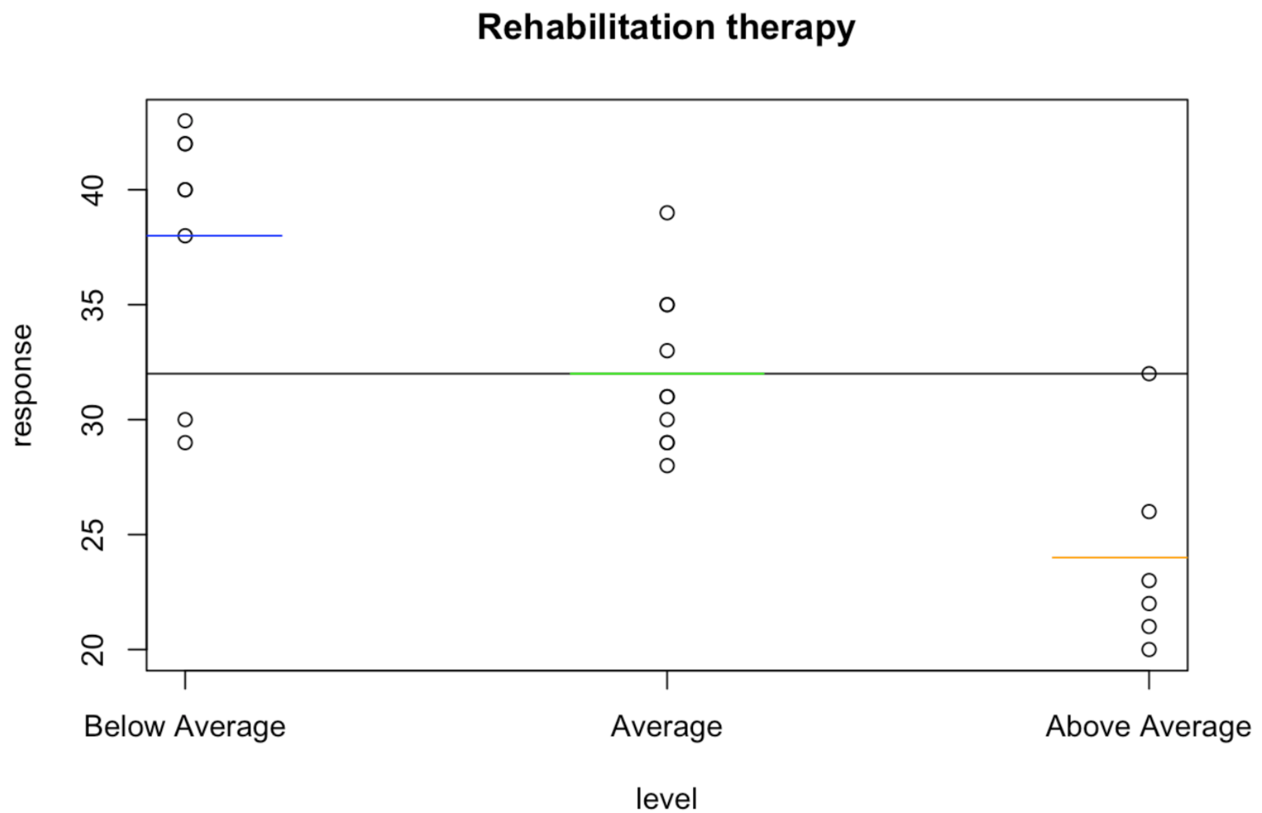
$$\hat{\epsilon}_{ij} = e_{ij}$$

Residuals are highly useful for checking whether assumptions of ANOVA Model is appropriate for the data at hand

Residuals sum to 0 for each treatment :

For ith treatment: $\sum_{j=1}^{n_i} e_i = \sum_{j=1}^{n_i} Y_{ij} - \overline{Y_{i.}} = 0$

Example



Do the factor level means appear to differ?

Does the variability of the observations within each factor level appear to be approximately the same for all factor levels?

Example

Fitted values and Residuals

Treatment levels	Response|fitted value|residual
1	29 | 38 | -9
1	42 | 38 | 4
1	38 | 38 | 0
1	40 | 38 | 2
1	43 | 38 | 5
1	40 | 38 | 2
1	30 | 38 | -8
1	42 | 38 | 4
2	30 | 32 | -2
2	35 | 32 | 3
2	39 | 32 | 7
2	28 | 32 | -4
2	31 | 32 | -1
2	31 | 32 | -1
2	29 | 32 | -3
2	35 | 32 | 3
2	29 | 32 | -3
2	33 | 32 | 1
3	26 | 24 | 2
3	32 | 24 | 8
3	21 | 24 | -3
3	20 | 24 | -4
3	23 | 24 | -1
3	22 | 24 | -2

Do residuals sum to zero within each treatment ?