STAT 515 Lec 17 slides

One-way analysis of variance

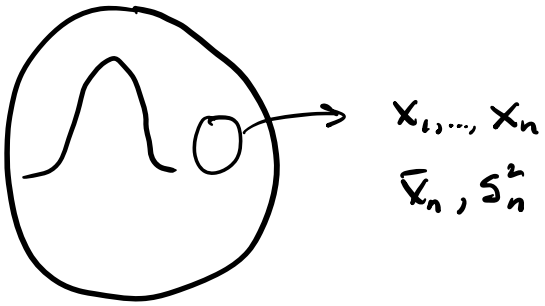
Karl Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

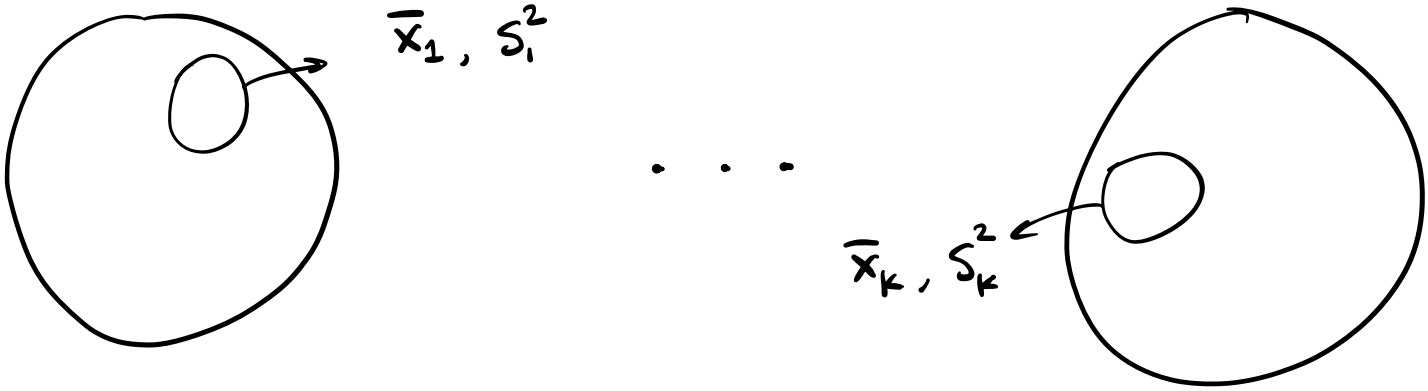
Normal

pop μ, σ^2, unknown



equal variances makes life easier

pop 1 μ_1, σ^2 ... pop K μ_K, σ^2



H_0: $\mu_1 = \cdots = \mu_K$

H_1: Not all the means are equal.

or $\mu_k \neq \mu_{k'}$ for some $k \neq k'$

$K = 3$

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: ~~$\mu_1 \neq \mu_2 \neq \mu_3$~~

$\mu_1 = 2$

$\mu_2 = -1$

$\mu_3 = -1$

Randomized experiments randomly assign subjects to different treatments.

Observational studies compare subjects existing in different circumstances.

Exercise: Experimental or observational?

1. Randomly assign plant clones to different drought conditions and measure CO_2 uptake.
2. Compare performance in school of children from different backgrounds.
3. Randomly assign tracts of a field to different fertilizers and compare yields.
4. Compare recycling habits of college students in Greenville and Columbia.

Observational studies are beset with the problem of *confounding variables*.

Confounding variable: An unrecorded property/circumstance associated with the outcome of interest as well as with a property/circumstance measured in the study.

Example: Family income and grades in school of children.

Is hours watching TV a confounding variable?

- Is hours watching TV associated with grades in school?
- Is hours watching TV associated with family income?

If yes to both, hours watching TV would be a confounder if ignored in the study.

The random assignment in randomized experiments breaks associations between measured and unmeasured variables, eliminating the problem of confounding variables.

Observational studies cannot establish causation—only association.
Randomized experiments *can* establish causation.

Vocabulary

- *Treatment*: A condition imposed by the investigator.
- *Experimental unit (EU)*: Each subject in the study—person, animal, etc.
- *Response*: Outcome measured on each EU after treatment applied.

Randomized experiment

Example: How to package a steak? Twelve steaks assigned to four different packagings (three to each) and bacteria per cm^2 recorded after nine days [1].

E.U.

$\bar{Y}_{MG\cdot} = 7.26$

Y_{ij}

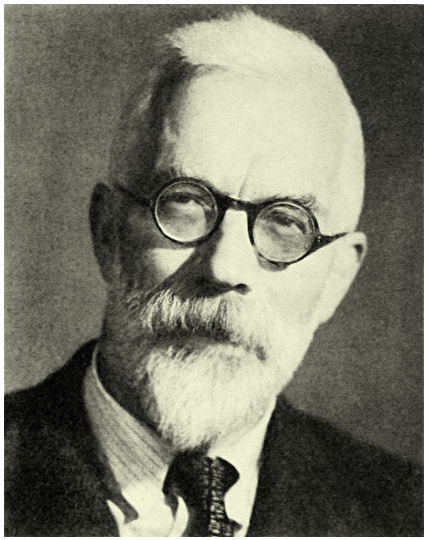
Steak	Packaging	log(# bact/cm^2)	Steak	Packaging	log(# bact/cm^2)
1	Commercial	7.66	10	Mixed Gas	7.41
6	Commercial	6.98	9	Mixed Gas	7.33
7	Commercial	7.80	2	Mixed Gas	7.04
12	Vacuum	5.26	8	CO_2	3.51
5	Vacuum	5.44	4	CO_2	2.91
3	Vacuum	5.80	11	CO_2	3.66

Treatments

Response values

$H_0: \mu_{comm} = \mu_{vacuum} = \mu_{MG} = \mu_{CO_2}$ vs

H_1: Not all means are equal.



Fischer

Example (cont): Here are the treatment means. How can we compare them?

Packaging	mean of log(# bact/cm^2)	
Commercial	7.48	$\bar{x}_{comm}$
Vaccuum	5.50	$\bar{x}_{vacuum}$
Mixed Gas	7.26	
CO_2	3.36	

Consider a randomized experiment with

- K treatment groups
- n EUs per treatment group

Assume:

"error term" or "noise"

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

ε : epsilon

Y_{ij}: Response for E.U. j in treatment group i

μ_i: Mean for treatment group i

Let

- K be the number of treatments.
- $n_1, \ldots, n_K$ be the numbers of EUs assigned to the treatments.
- $N = n_1 + \cdots + n_K$ be the total number of EUs.
- Y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$ be response for EU j in treatment group i.

"factor"

lm(bacteria ~ packaging)

Cell-means or one-way ANOVA model

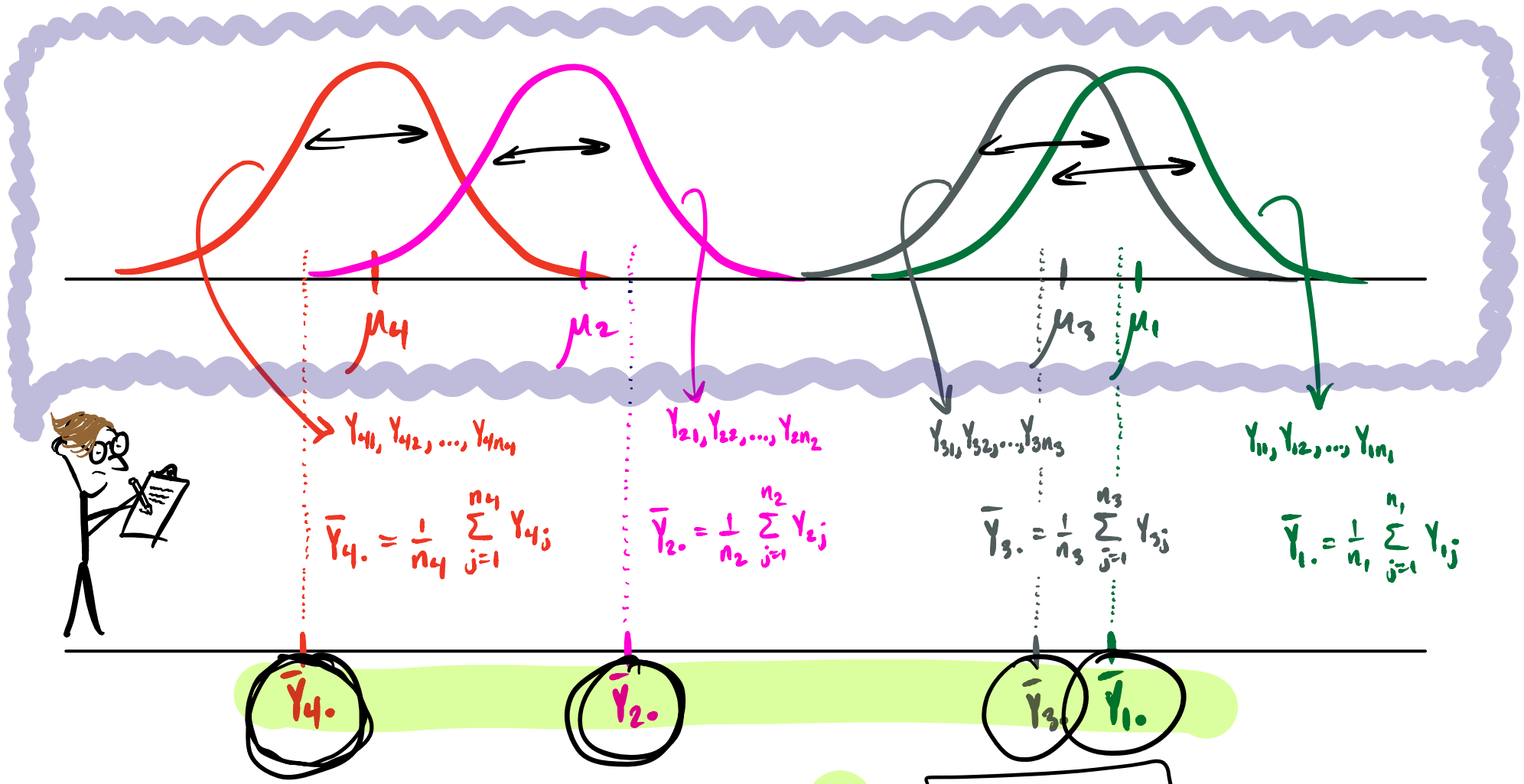
Assume

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, K,$$

where

- $\mu_1, \ldots, \mu_K$ are the population means for treatments $1, \ldots, K$.
- the ε_{ij} are independent Normal$(0, \sigma^2_\varepsilon)$.

$K = 4$



Estimate $\mu_1, \ldots, \mu_K$ with treatment means $\bar{Y}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}, \; i = 1, \ldots, K.$

Y_{ij}

"$\cdot$" indicates that we have summed over j

Research question: Do/does any of the treatments affect the response?

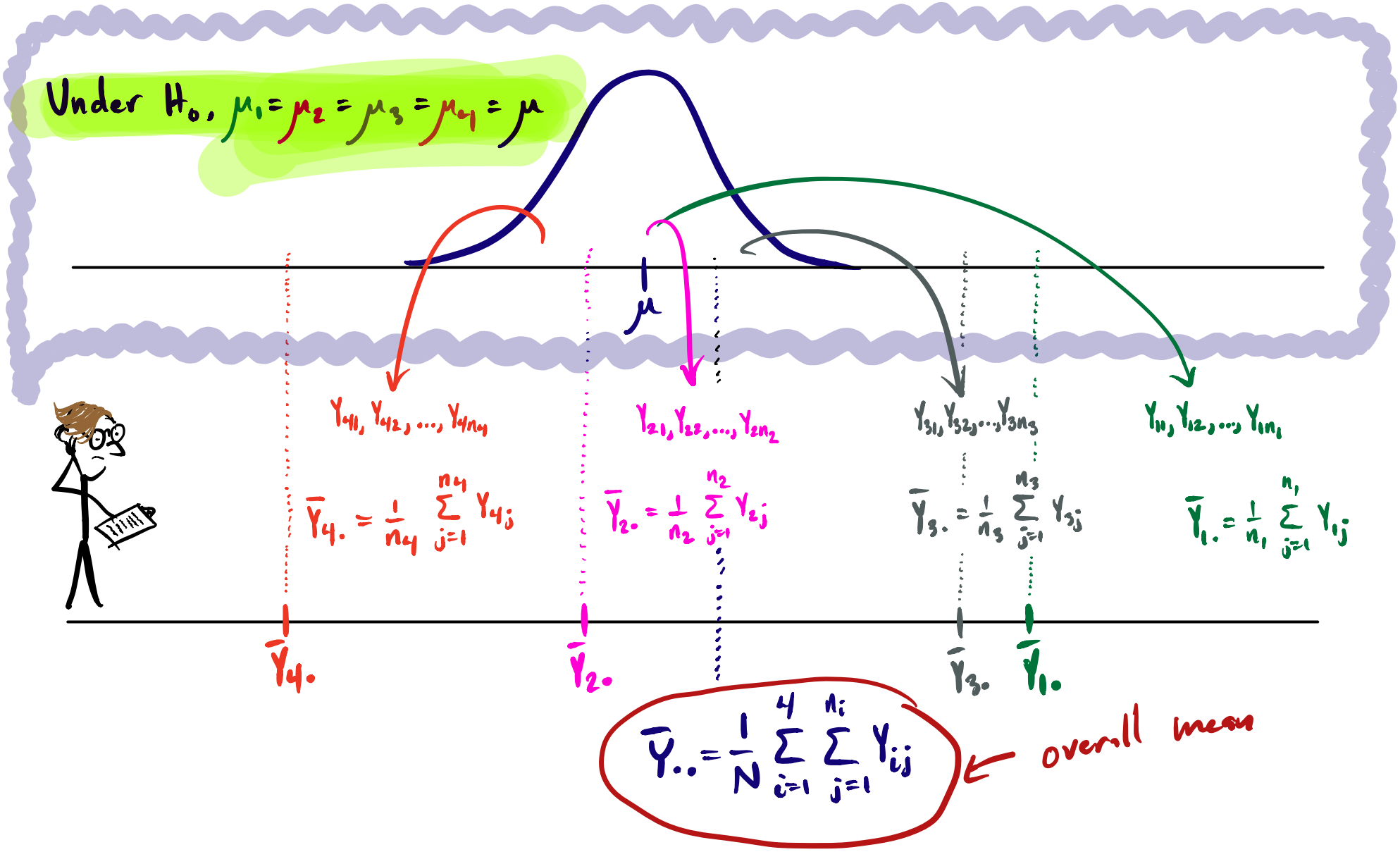
Central hypotheses in cell means model

H_0: $\mu_1 = \cdots = \mu_K$

H_1: $\mu_i \neq \mu_{i'}$ for some $i \neq i'$, i.e. not all treatment means are equal

To build a test statistic, we look at the spread of $\bar{Y}_{1\cdot}, \ldots, \bar{Y}_{K\cdot}$.

treatment group means



Estimate overall mean with $\bar{Y}_{\cdot\cdot} = N^{-1} \sum_{i=1}^{K} \sum_{j=1}^{n_i} Y_{ij}$.

Analysis of variance (ANOVA): Decomposition of the variability in Y_{ij} into

1. *Between-treatment variation*: Variability due to treatment effects.
2. *Within-treatment variation*: Variability due to differences among EUs.

Represents "total" variability in Y_{ij}.

within

$$SS_{Total} = SS_{trt} + SS_{error}$$

Between-treatment

Sums of squares (SS)

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

$$SS_{trt} = \sum_{i=1}^{K} n_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2$$

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

mean for group i.

Sums of squares for decomposing variation in the Y_{ij}

$$SS_{\text{Total}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 \qquad \text{(Total variation)}$$

$$SS_{\text{Treatment}} = \sum_{i=1}^{K} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \qquad \text{(Between-treatment)}$$

$$SS_{\text{Error}} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i.})^2 \qquad \text{(Within-treatment)}$$

We have

$$\underbrace{SS_{\text{Total}}}_{\text{Total}} = \underbrace{SS_{\text{Treatment}}}_{\text{Between}} + \underbrace{SS_{\text{Error}}}_{\text{Within}}.$$

Sampling distributions of scaled sums of squares

Under the cell means model under H_0: $\mu_1 = \cdots = \mu_K$, we have

$$SS_{\text{Total}} / \sigma_\varepsilon^2 \sim \chi^2_{N-1}$$
$$SS_{\text{Treatment}} / \sigma_\varepsilon^2 \sim \chi^2_{K-1}$$
$$SS_{\text{Error}} / \sigma_\varepsilon^2 \sim \chi^2_{N-K}.$$

Define the treatment and error *mean squares* as

Treatment Mean Square

Error Mean Square

$$MS_{\text{Treatment}} = SS_{\text{Treatment}} / (K-1)$$
$$MS_{\text{Error}} = SS_{\text{Error}} / (N-K).$$

Lastly, define the *F-statistic* as

$H_0: \mu_1 = \cdots = \mu_K$

larger values reflect more evidence against $H_0: \mu_1 = \cdots = \mu_K$

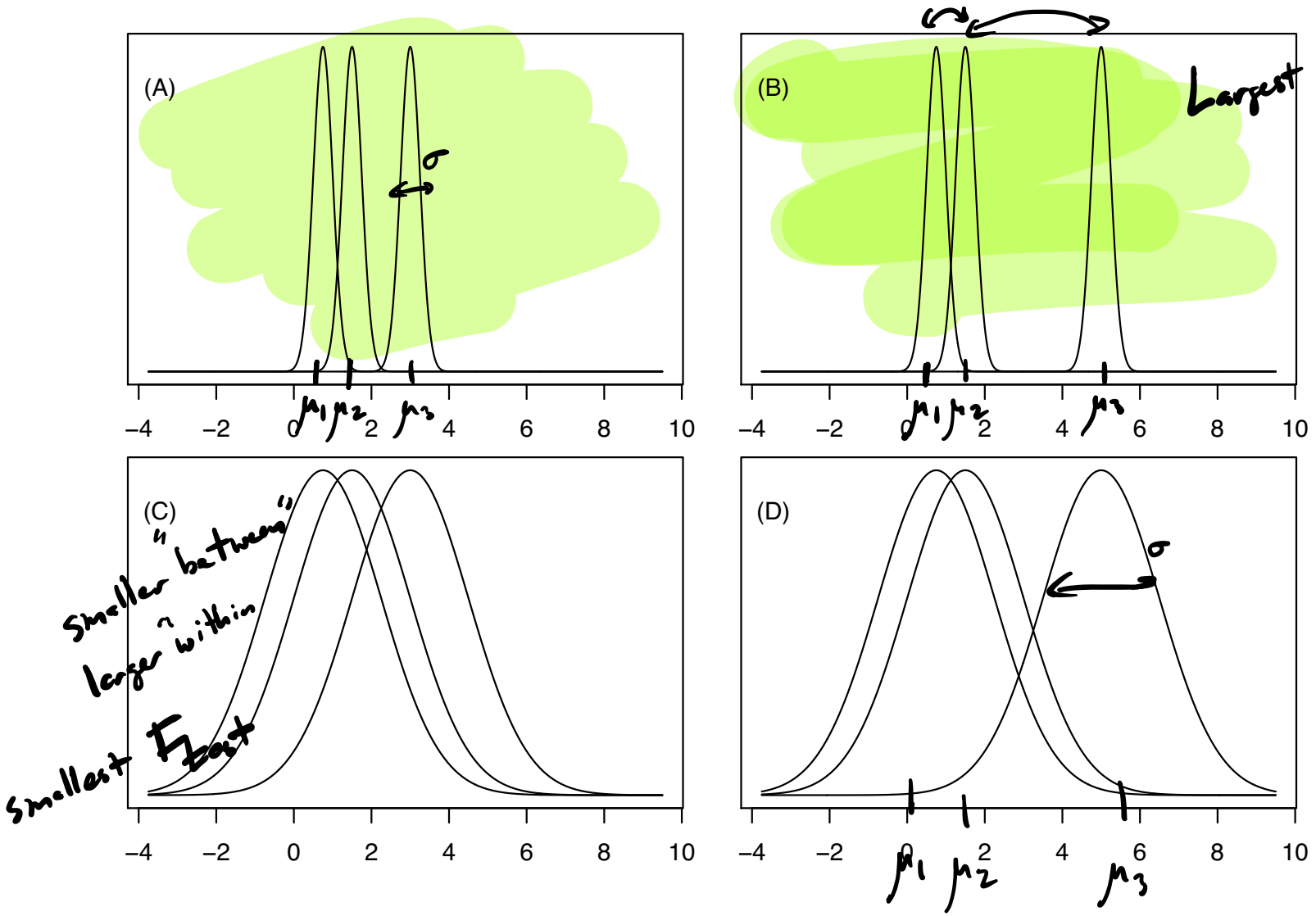
$$F_{\text{test}} = \frac{MS_{\text{Treatment}}}{MS_{\text{Error}}}. = \frac{\text{Between-trt-variation}}{\text{Within-trt-variation}}$$

Exercise:

$F_{test} = \frac{\text{Between}}{\text{Within}}$

$K = 3$

F_{stat}.



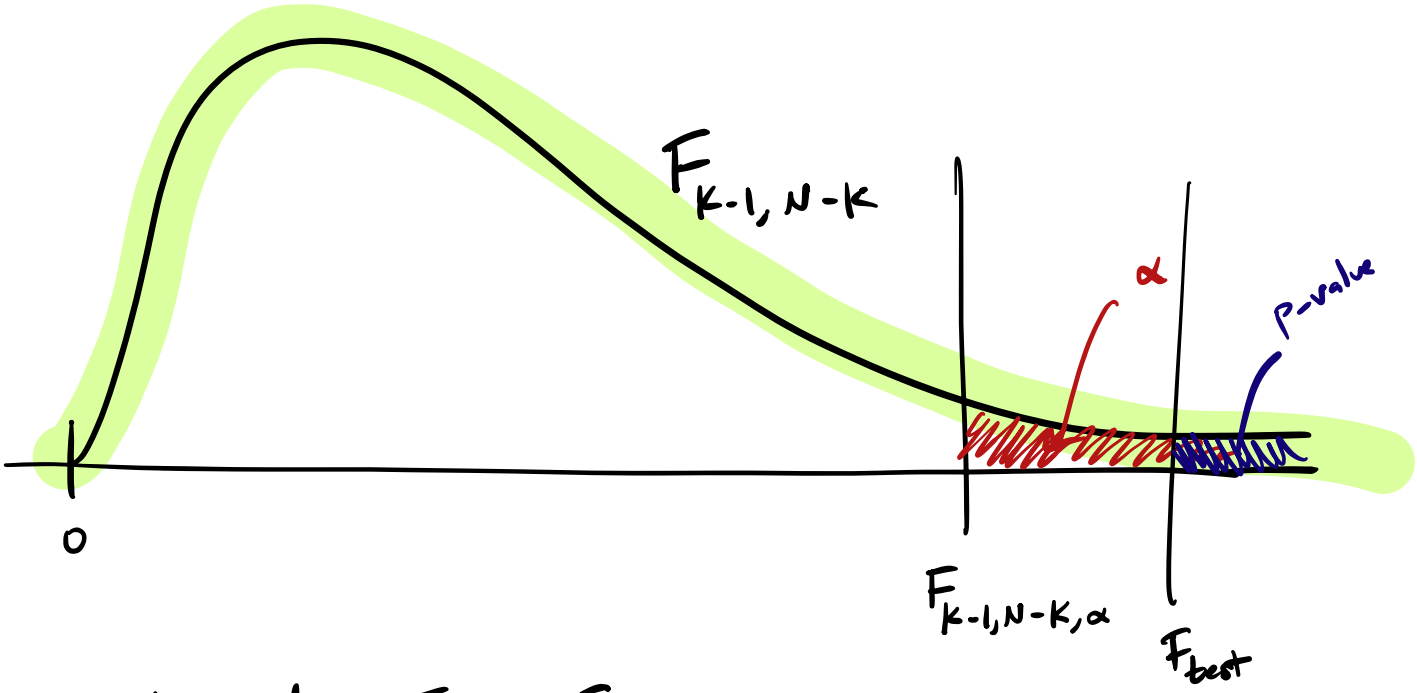
i) Largest F_{test}? ii) Smallest? iii) Two with larger $MS_{\text{Treatment}}$? iv) Larger MS_{Error}?

$H_0: \mu_1 = \cdots = \mu_K$

$$F_{test} = \frac{MS_{Trt}}{MS_{Error}} = \frac{\text{Between}}{\text{Within}} \sim F_{K-1,\,N-K}$$

$MS_{Trt} = \chi^2_{K-1}/(K-1)$

$MS_{Error} = \chi^2_{N-K}/(N-K)$



Reject H_0 if $F_{test} > F_{K-1,N-K,\alpha}$

Let $W_1 \sim \chi^2_{df_1}$ and $W_2 \sim \chi^2_{df_2}$, W_1 and W_2 independent, then

$$\frac{W_1/df_1}{W_2/df_2} \sim F_{df_1,\,df_2}$$

"numerator df" df_1

"denominator df" df_2

Sampling distribution of F-statistic

Under H_0: $\mu_1 = \cdots = \mu_K$, we have

$$F_{\text{test}} \sim F_{K-1,N-K},$$

where $F_{K-1,N-K}$ is the F-dist. with num. df $K-1$ and denom. df $N-K$.

F-test for significance of treatment effect

We reject H_0: $\mu_1 = \cdots = \mu_K$ at significance level α if $F_{\text{test}} > F_{K-1,N-K,\alpha}$.

The next slides introduce the F-distributions...

The F-distributions

The F-distribution with num. df $\nu_1 > 0$ and den. df $\nu_2 > 0$ has pdf given by

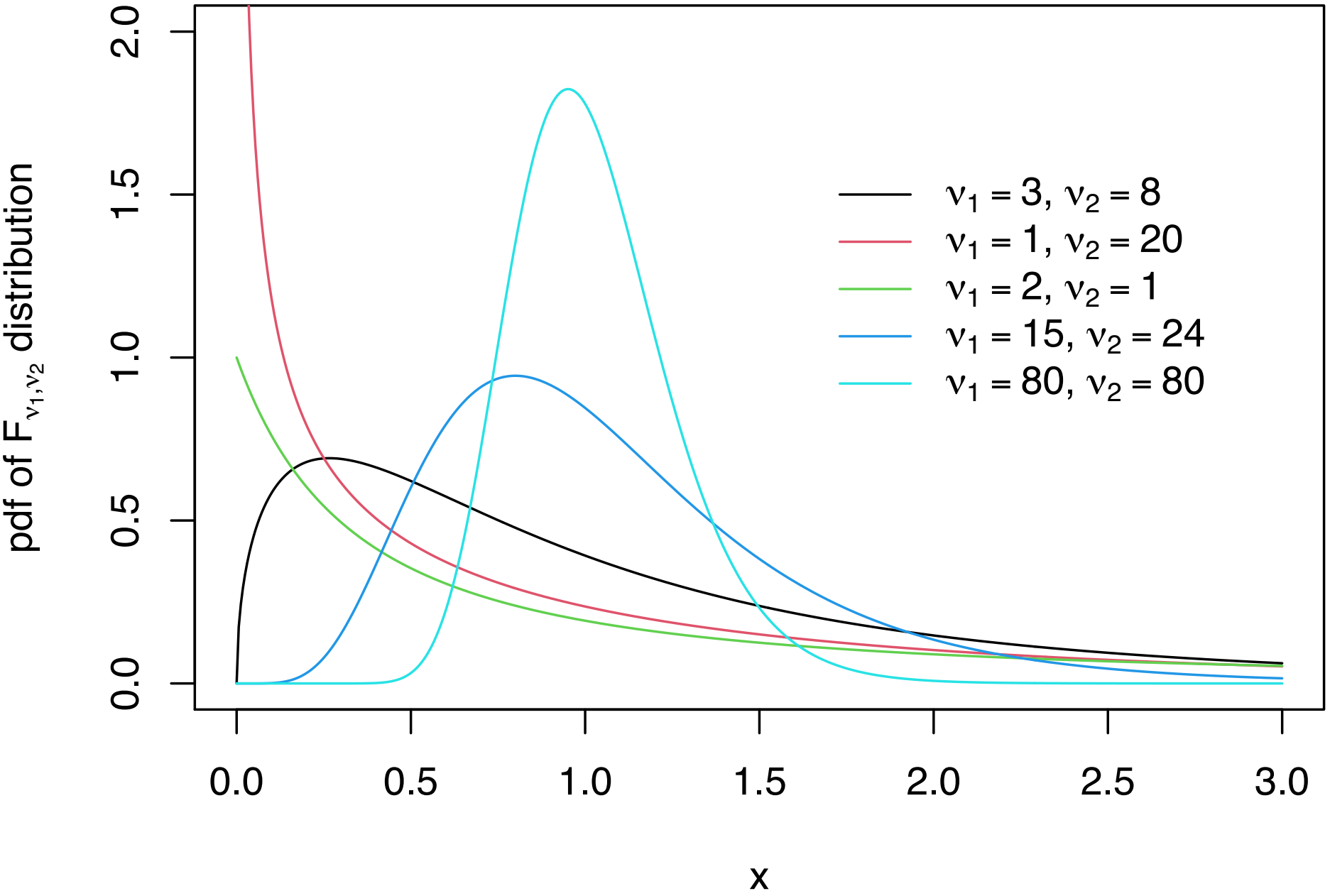
$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1} \left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}, \quad x > 0.$$

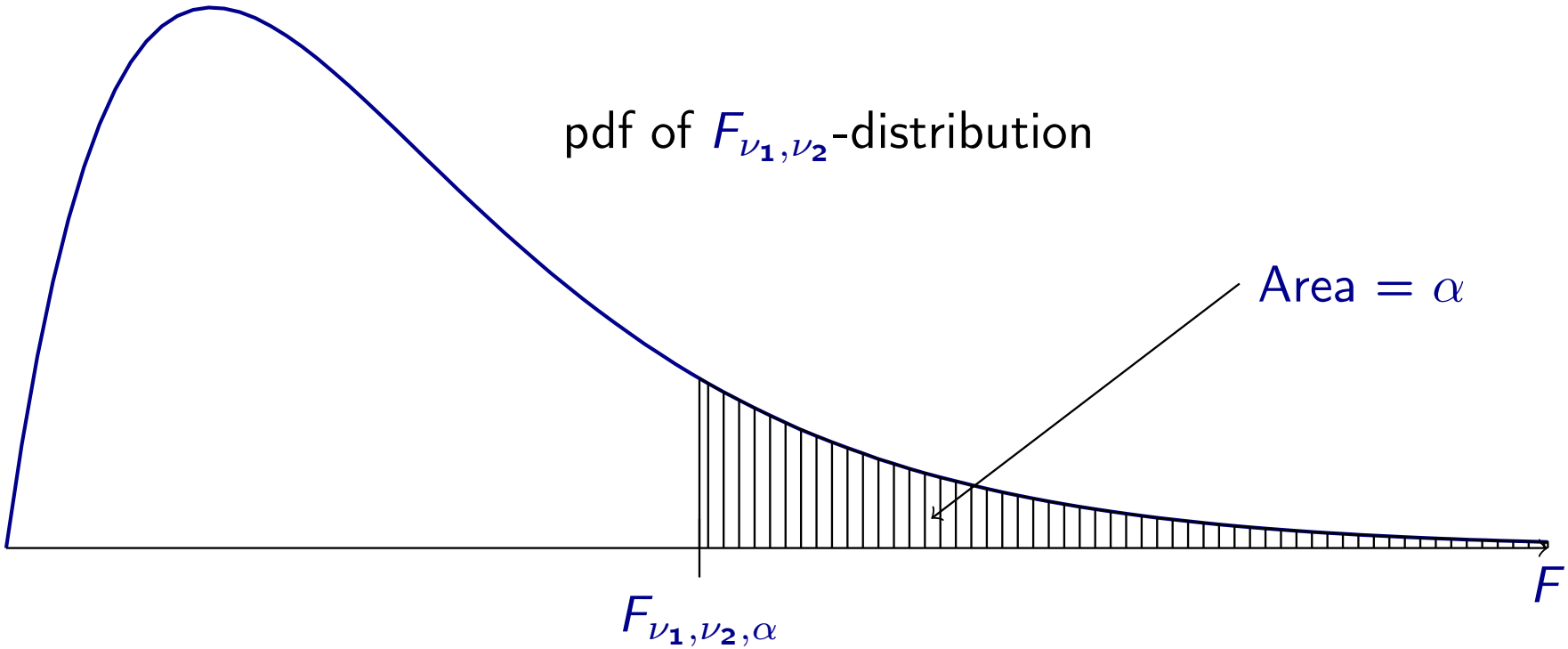
We write $X \sim F_{\nu_1,\nu_2}$.

F-distributed rv as ratio of chi-squared rvs

If $W_1 \sim \chi^2_{\nu_1}$ and $W_2 \sim \chi^2_{\nu_2}$ are independent, then

$$\frac{W_1/\nu_1}{W_2/\nu_2} \sim F_{\nu_1,\nu_2}.$$


pdf of F_{ν_1,ν_2} distribution
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
x
$\nu_1 = 3, \nu_2 = 8$
$\nu_1 = 1, \nu_2 = 20$
$\nu_1 = 2, \nu_2 = 1$
$\nu_1 = 15, \nu_2 = 24$
$\nu_1 = 80, \nu_2 = 80$



Can use function `qf()` to look up the values, e.g.

$$F_{3,8,0.05} = \texttt{qf(.95,3,8)} = 4.066181$$
$$F_{3,8,0.01} = \texttt{qf(.99,3,8)} = 7.590992$$

Can get area under the curve to the left with the `pf()` function.

Analysis of variable

The *ANOVA table* is a table presenting all of these values:

Source	df	Sum of Sq	Mean Sq	F	p-value
Treatment	$K-1$	$SS_{\text{Treatment}}$	$MS_{\text{Treatment}}$	F_{test}	$P(F > F_{\text{test}})$
Error	$N-K$	SS_{Error}	MS_{Error}		where $F \sim F_{K-1,N-K}$
Total	$N-1$	SS_{Total}			

$\frac{SS_{Trt}}{K-1}$

$\frac{MS_{Trt}}{MS_{Error}}$

$\frac{SS_{Error}}{N-K}$

ingredients for constructing F_{test}.

Exercise: Get the ANOVA table for the steaks data using `lm()` and `anova()`.

```
# read in the data and format it for ANOVA:

bacteria <- c(7.66,6.98,7.80,
              5.26,5.44,5.80,
              7.41,7.33,7.04,
              3.51,2.91,3.66)

packaging <- c(rep("Commercial",3),
               rep("Vacuum",3),
               rep("Mixed Gas",3),
               rep("C02",3))

packaging <- as.factor(packaging)

# estimate model with lm() function and retrieve ANOVA table:

model <- lm(bacteria ~ packaging)
anova(model)
```

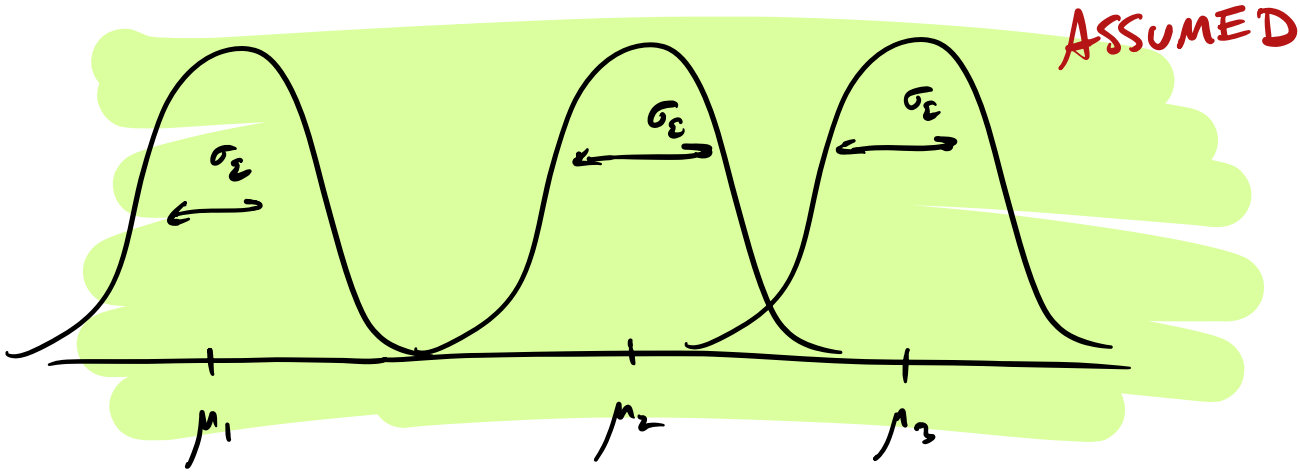
One-way ANOVA model.

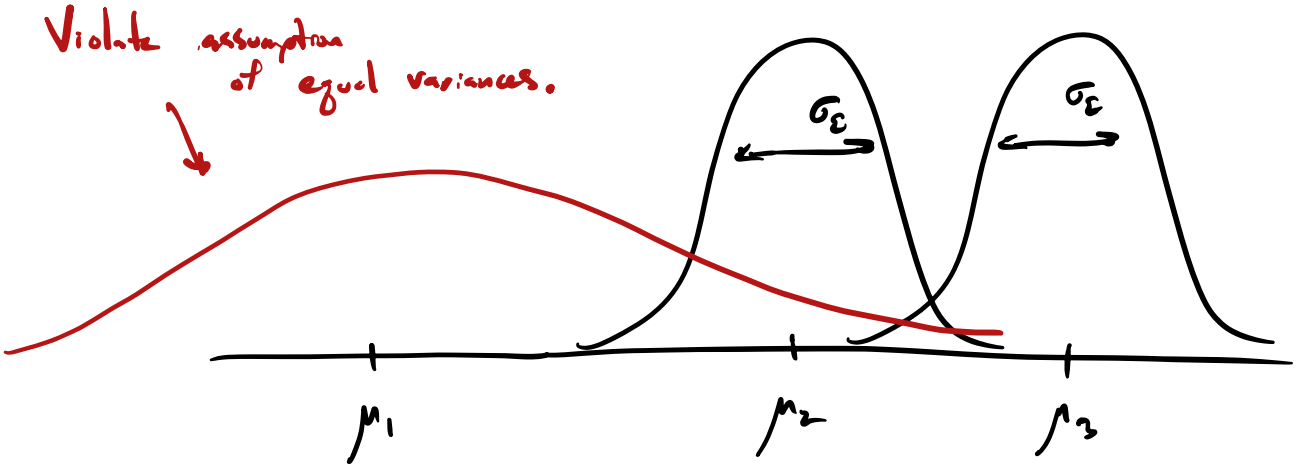
$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

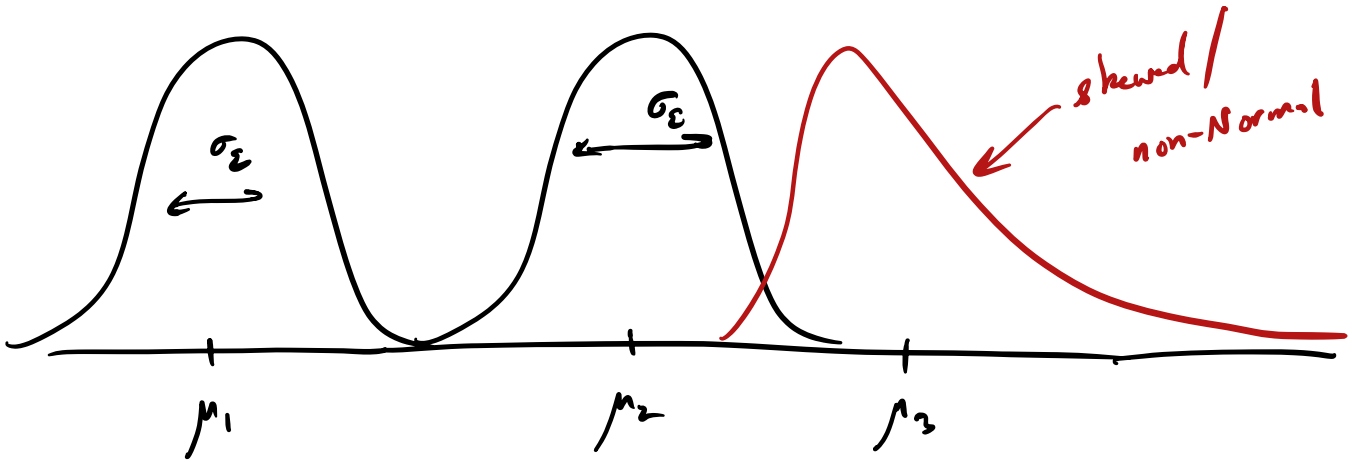
$$\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$$

μ_i: trt i mean

ε_{ij}: noise






σ_ε
σ_ε
skewed / non-Normal
μ_1
μ_2
μ_3

Consider the assumptions of the model

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad j = 1, \ldots, n_i, \quad i = 1, \ldots, K,$$

where the ε_{ij} are independent Normal$(0, \sigma^2_\varepsilon)$.

(A.1) The responses are Normally distributed around the treatment means.

To check: Look at a QQ plot of the residuals.

(A.2) The responses have the same variance in all treatment groups.

To check: Look at the residuals versus fitted values plot.

(A.3) The responses are independent from each other.

Cannot check: Trust the random assignment of EUs to treatments.

Residuals are defined on the next slide...

$\varepsilon_{ij} = Y_{ij} - \mu_i$ $\quad (\ Y_{ij} = \mu_i + \varepsilon_{ij}\ )$

Define the *residuals* as $\hat{\varepsilon}_{ij} = Y_{ij} - \bar{Y}_{i\cdot}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, K$.

Example (cont): This table includes the residuals from the steak experiment.

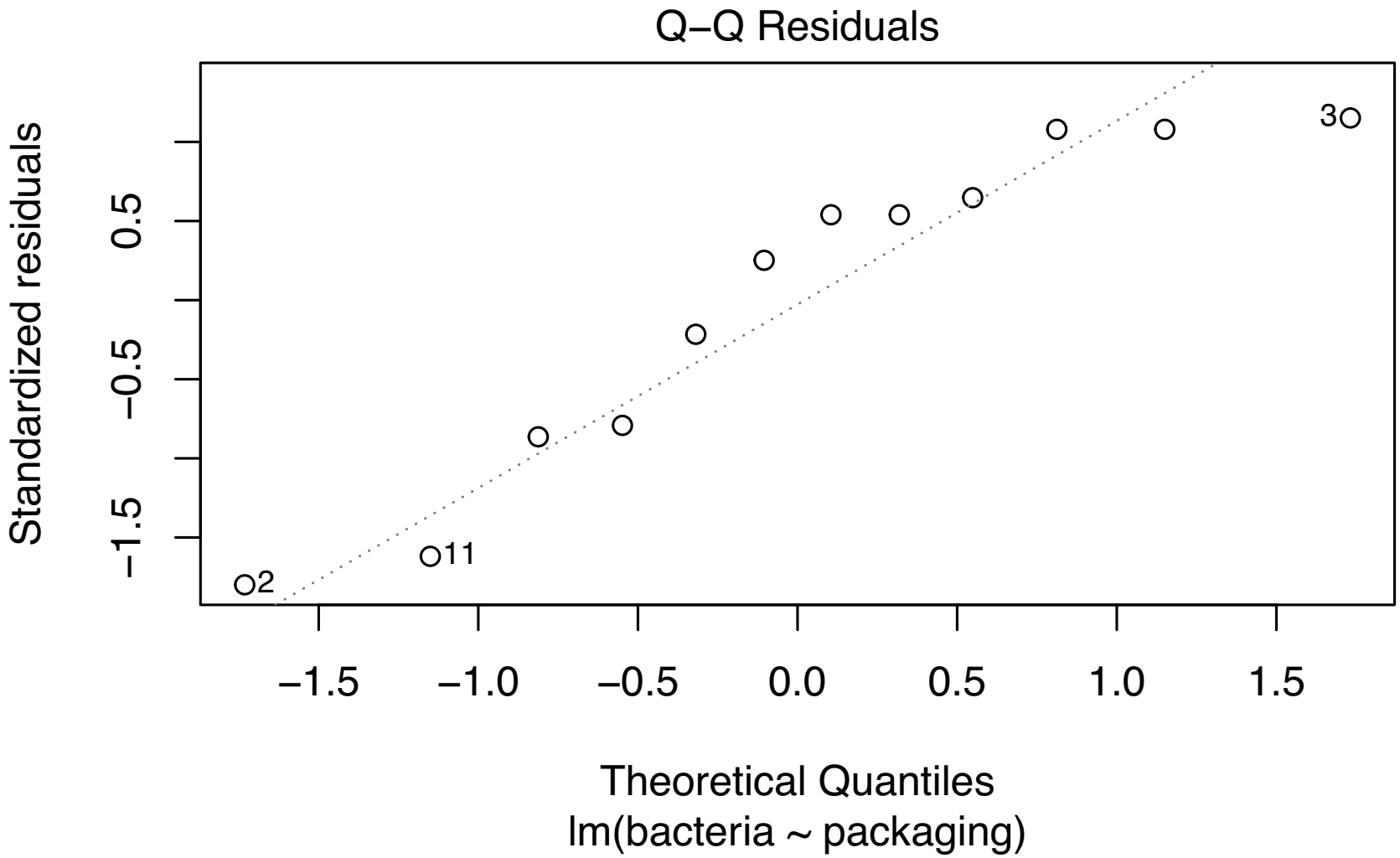
Steak	Packaging	log(# bact/cm^2)	$\bar{Y}_{i\cdot}$	$\hat{\varepsilon}_{ij}$
1	Commercial	7.66	7.48	0.18
6	Commercial	6.98	7.48	−0.50
7	Commercial	7.80	7.48	0.32
12	Vacuum	5.26	5.50	−0.24
5	Vacuum	5.44	5.50	−0.06
3	Vacuum	5.80	5.50	0.30
10	Mixed Gas	7.41	7.26	0.15
9	Mixed Gas	7.33	7.26	0.07
2	Mixed Gas	7.04	7.26	−0.22
8	CO_2	3.51	3.36	0.15
4	CO_2	2.91	3.36	−0.45
11	CO_2	3.66	3.36	0.30

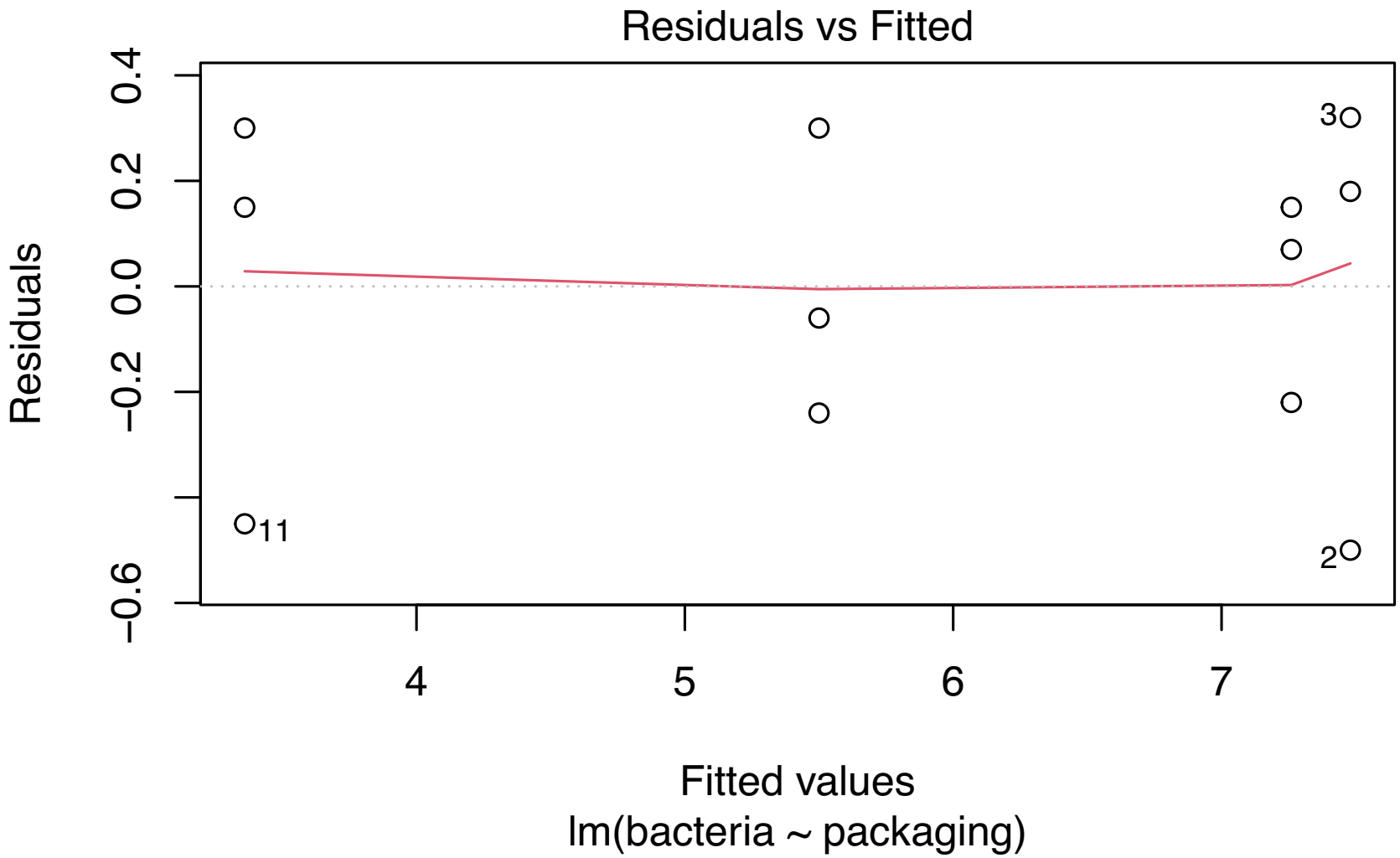
$\bar{Y}_{comm.}$

make normal QQ plot of these.

Y_{ij}

```
plot(model,which=2) # qq plot of residuals
plot(model,which=1) # residuals versus fitted values plot
```


Q–Q Residuals
Standardized residuals
−1.5
−0.5
0.5
3
11
2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Theoretical Quantiles
lm(bacteria ~ packaging)


Residuals vs Fitted
Residuals
0.4
0.2
0.0
−0.2
−0.6
4
5
6
7
3
2
11
Fitted values
lm(bacteria ~ packaging)


R. O. Kuehl.
Design of Experiments: Statistical Principles of Research Design and Analysis.

Duxbury/Thomson Learning, 2000.
Google-Books-ID: mIV2QgAACAAJ.