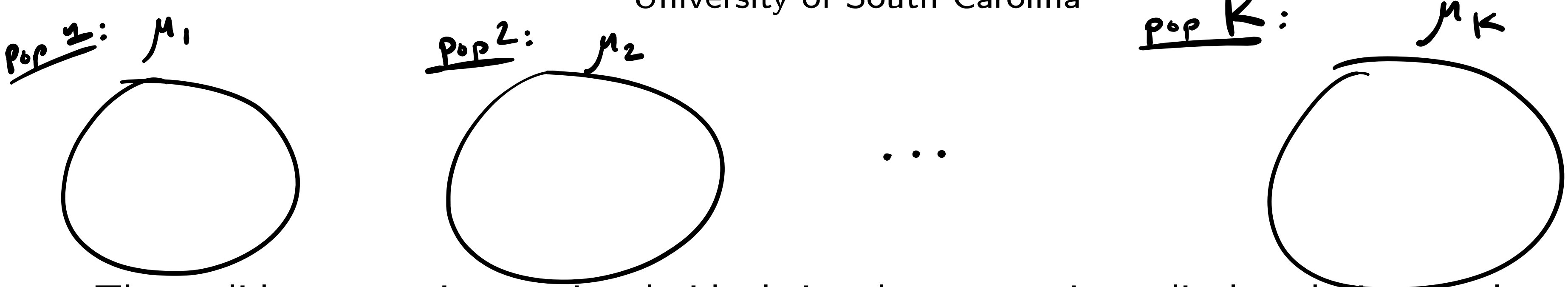
STAT 515 Lec 17 slides

Comparative experiments and analysis of variance

K = # of populations

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.

Test $H_0: \mu_1 = \mu_2 = \cdots = \mu_K$ (all means same) vs H_1: all means not the same.

Randomized

Comparative 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 comparative experiments breaks associations between measured and unmeasured variables, eliminating the problem of confounding variables.

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

Vocabulary for comparative experiments

- *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.

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].

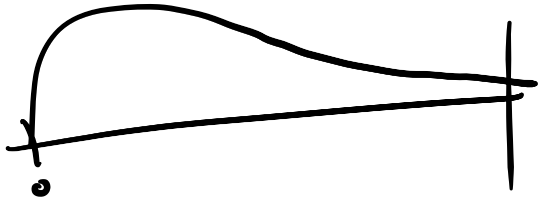
Steak	Packaging 1	log(# bact/cm^2)	Steak	Packaging	log(# bact/cm^2)
1	Commercial	Y_{11} = 7.66	10	Mixed Gas	7.41
6	Commercial	Y_{12} = 6.98	9	Mixed Gas	7.33
7	Commercial	Y_{13} = 7.80	2	Mixed Gas	7.04
12	Vacuum	Y_{21} = 5.26	8	CO_2	3.51
5	Vacuum	Y_{22} = 5.44	4	CO_2	2.91
3	Vacuum	Y_{23} = 5.80	11	CO_2	3.66

avg is 7.48

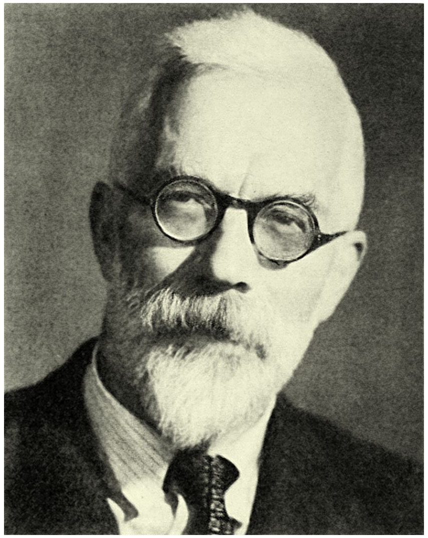
2

$$\text{Comm.} \; (\text{Avg} = 7.48) \quad \begin{pmatrix} 7.66 \\ 6.98 \\ 7.80 \end{pmatrix} \qquad \begin{pmatrix} 7.66 - 7.48 \\ 6.98 - 7.48 \\ 7.80 - 7.48 \end{pmatrix} \; \text{"residuals"}$$

$$\text{Vac} \; (\text{Avg} = 5.50) \quad \begin{pmatrix} 5.26 \\ 5.44 \\ 5.80 \end{pmatrix} \qquad \begin{pmatrix} 5.26 - 5.50 \\ 5.44 - 5.50 \\ 5.80 - 5.50 \end{pmatrix} \; \text{residuals}$$



F-distribution



$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$

Fisher

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

Packaging	mean of log(# bact/cm^2)	
Commercial	7.48	$= \bar{Y}_{1.}$
Vaccuum	5.50	$= \bar{Y}_{2.}$
Mixed Gas	7.26	$= \bar{Y}_{3.}$
CO_2	3.36	$= \bar{Y}_{4.}$

$K = 4$

$n_1 = 3, \; n_2 = 3, \; n_3 = 3, \; n_4 = 3$

$N = 12$

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.

Cell-means or one-way ANOVA model

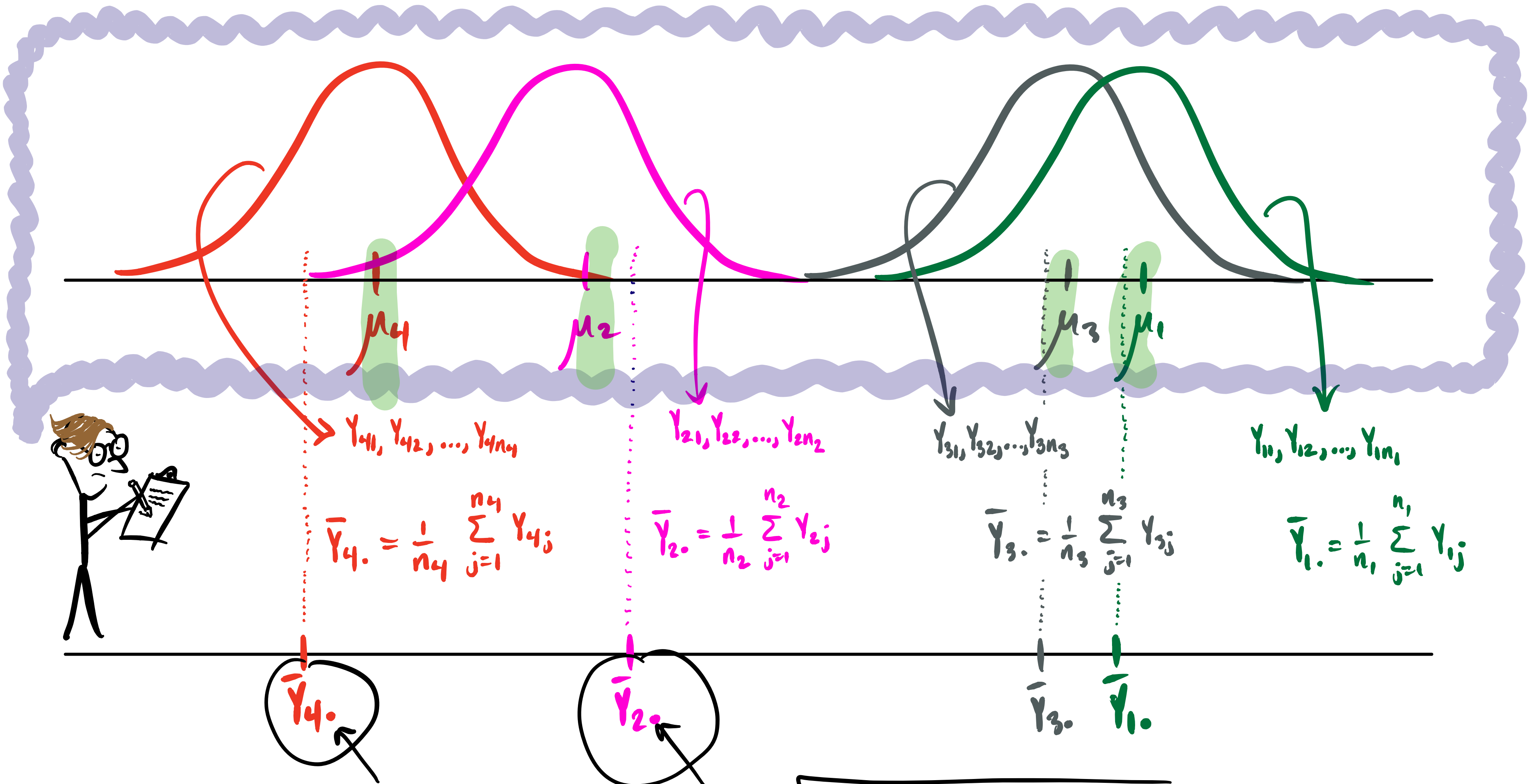
"epsilon", "error terms", "noise"

Assume

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

where

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



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

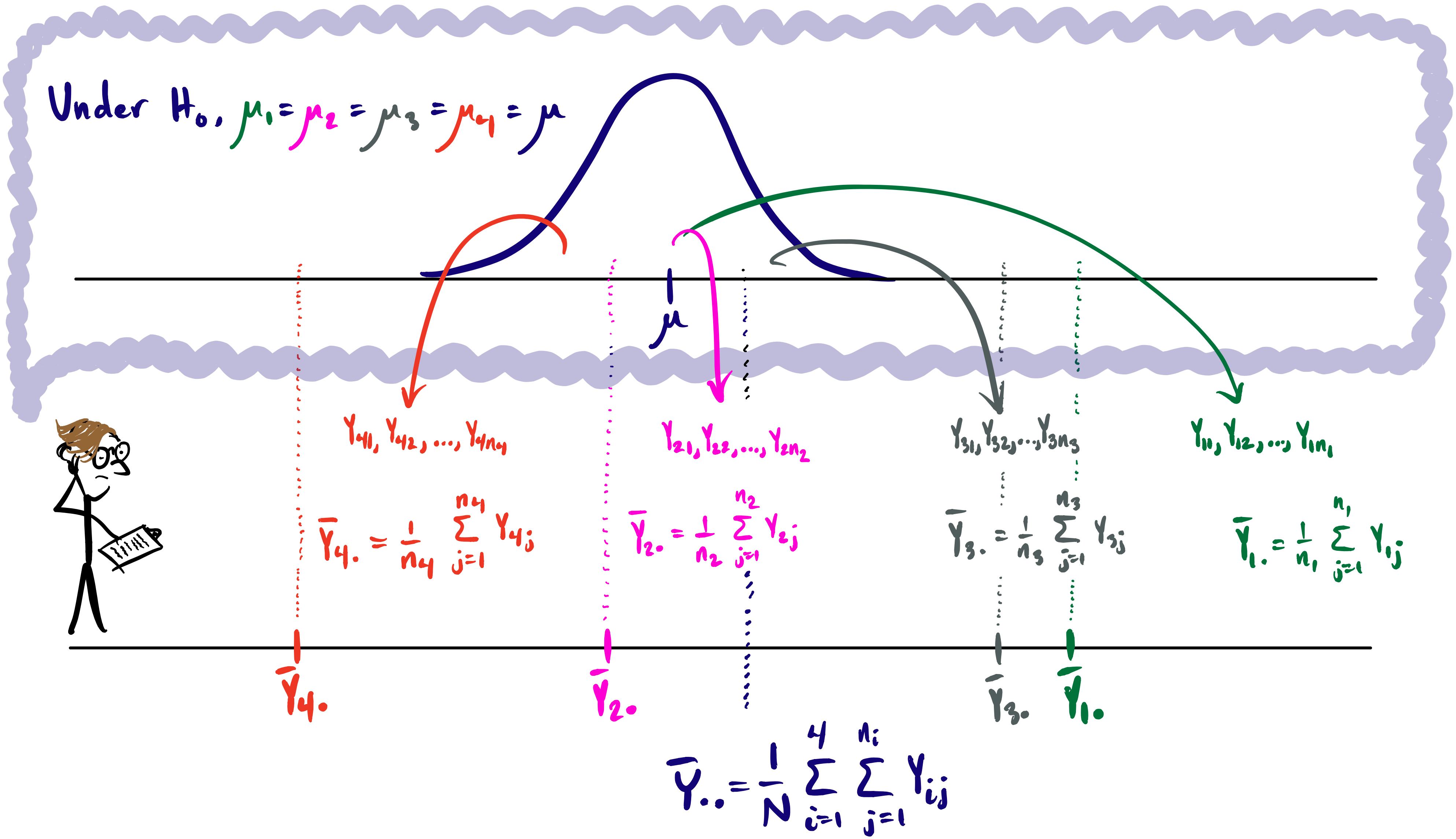
Research question: Do/does any of the treatments affect the response?

Central hypotheses in cell means model

$$\boxed{H_0: \quad \mu_1 = \cdots = \mu_K}$$
$$H_1: \quad \mu_i \neq \mu_{i'} \text{ for some } i \neq i', \text{ i.e. not all treatment means are equal}$$

Wrong: $H_1: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_K$

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



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

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.

observed responses.

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

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

overall mean

overall

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

mean for group i

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

mean for group i

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

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

$$\begin{aligned} \mathrm{SS}_{\text{Total}} / \sigma_\varepsilon^2 &\sim \chi^2_{N-1} \\ \mathrm{SS}_{\text{Treatment}} / \sigma_\varepsilon^2 &\sim \chi^2_{K-1} \\ \mathrm{SS}_{\text{Error}} / \sigma_\varepsilon^2 &\sim \chi^2_{N-K}. \end{aligned}$$

Define the treatment and error *mean squares* as

$$\begin{aligned} \mathrm{MS}_{\text{Treatment}} &= \mathrm{SS}_{\text{Treatment}} / (K-1) \\ \mathrm{MS}_{\text{Error}} &= \mathrm{SS}_{\text{Error}} / (N-K). \end{aligned}$$

Lastly, define the *F-statistic* as

Reject $H_0: \mu_1 = \cdots = \mu_K$ if this is "large enough".

$$F_{\text{test}} = \frac{\mathrm{MS}_{\text{Treatment}}}{\mathrm{MS}_{\text{Error}}}. = \frac{\text{Between treatments}}{\text{within treatments}}$$

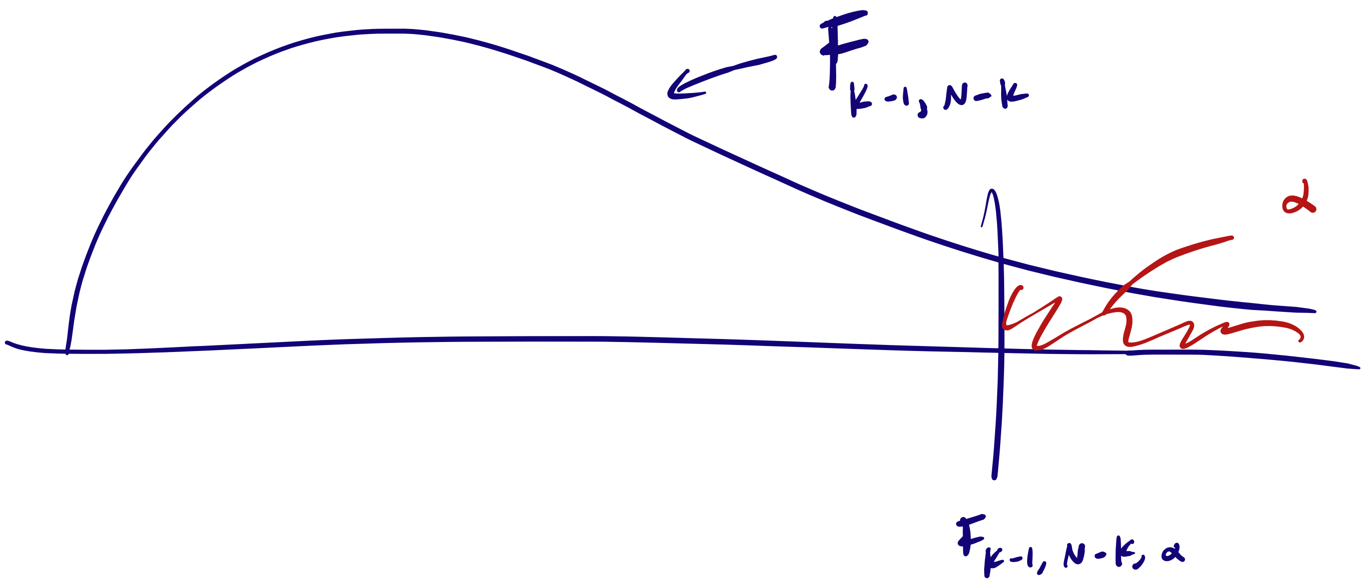
$H_0: \mu_1 = \cdots = \mu_K$ $\qquad$ $\bar{Y}_{i\cdot}$ is our estimate of μ_i.

$$F_{test} = \frac{MS_{Trt}}{MS_{Error}} = \frac{SS_{Trt}/(K-1)}{SS_{Error}/(N-K)}$$

large when treatment means are spread out

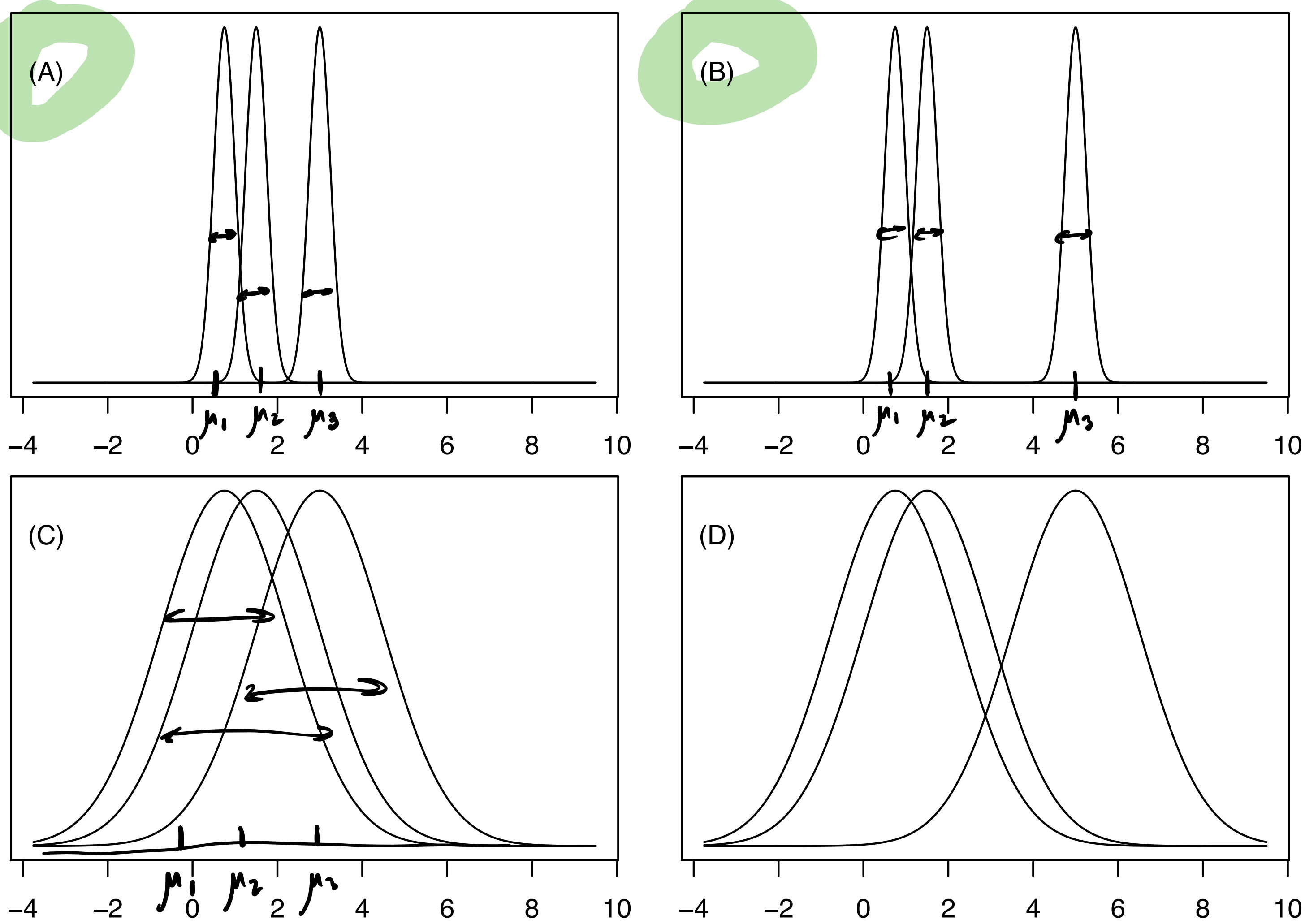
$$= \frac{\sum_{i=1}^{K} n_i \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 / (K-1)}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2 / (N-K)}$$

large when the values within each treatment group are spread out.



Reject $H_0: \mu_1 = \cdots = \mu_K$ if $F_{test} > F_{K-1, N-K, \alpha}$.

Exercise: $K = 3$



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

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. . .

optional

The F-distributions

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

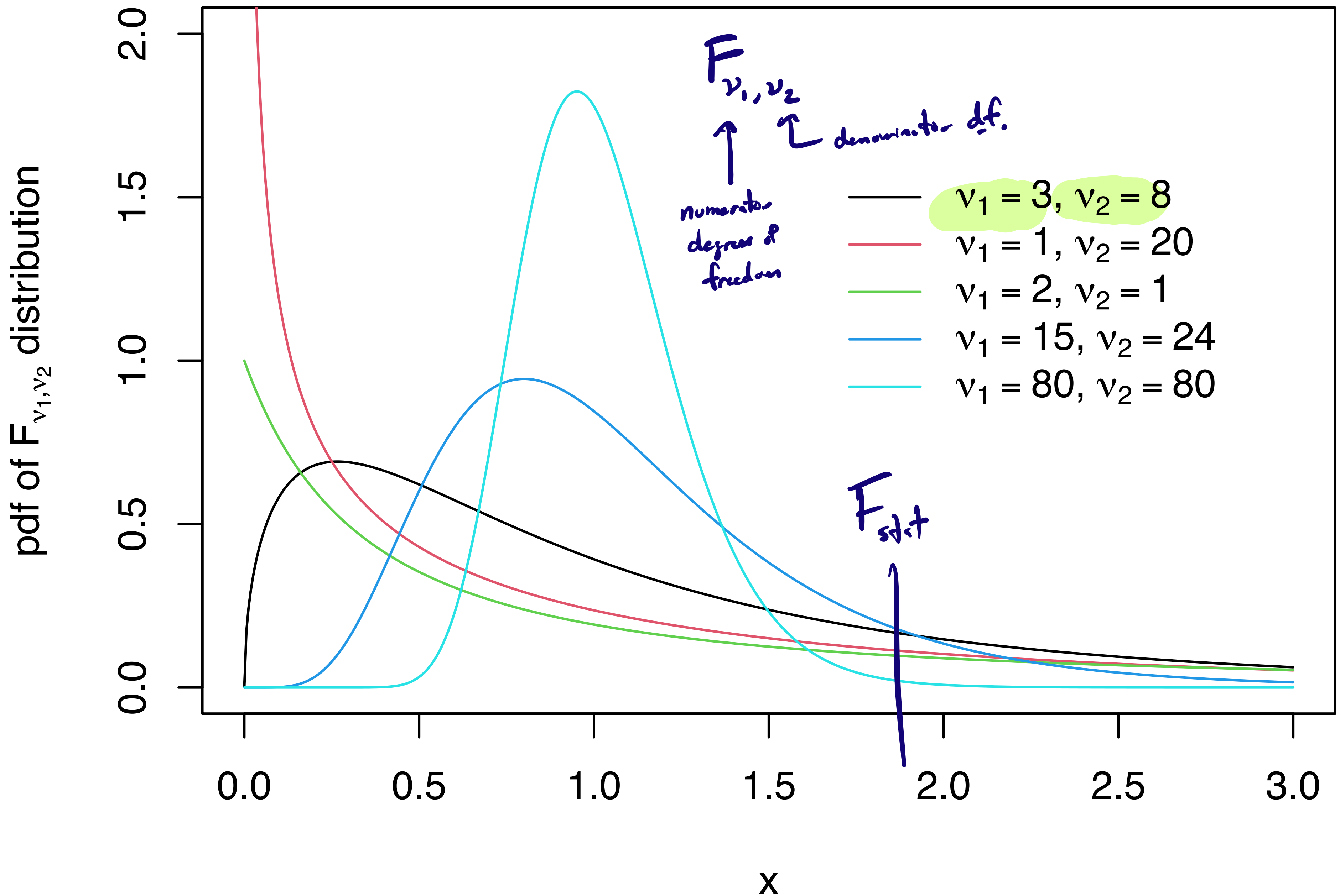
$$f(x) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}\left(1+\frac{\nu_1}{\nu_2}x\right)^{-\frac{\nu_1+\nu_2}{2}}, \quad x > 0.$$

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

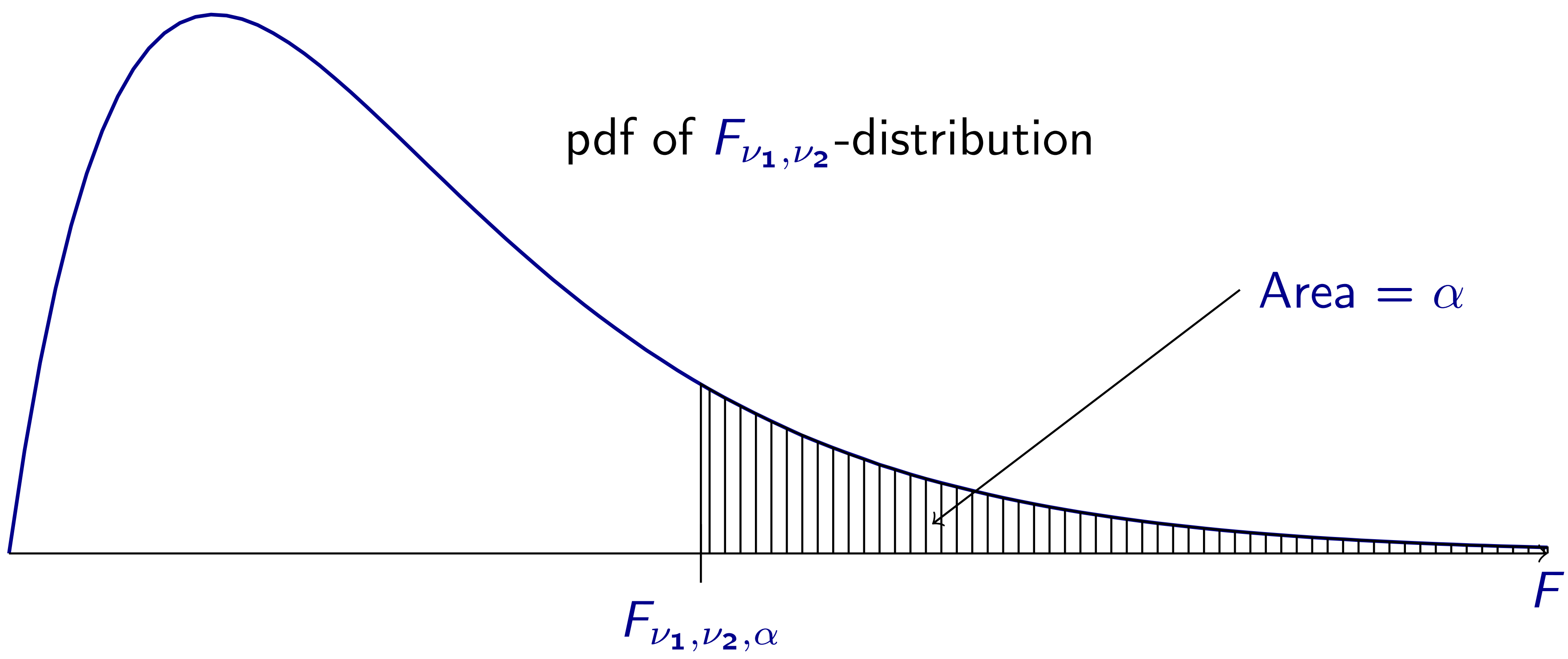
F-distributed rv as ratio of chi-squared rvs

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
$\nu_1 = 3, \nu_2 = 8$
$\nu_1 = 1, \nu_2 = 20$
$\nu_1 = 2, \nu_2 = 1$
$\nu_1 = 15, \nu_2 = 24$
$\nu_1 = 80, \nu_2 = 80$
F_{ν_1, ν_2}
numerator degrees of freedom
denominator d.f.
F_{stat}



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 variance

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

$\frac{MS_{Treatment}}{MS_{Error}}$

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

"Residual"

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

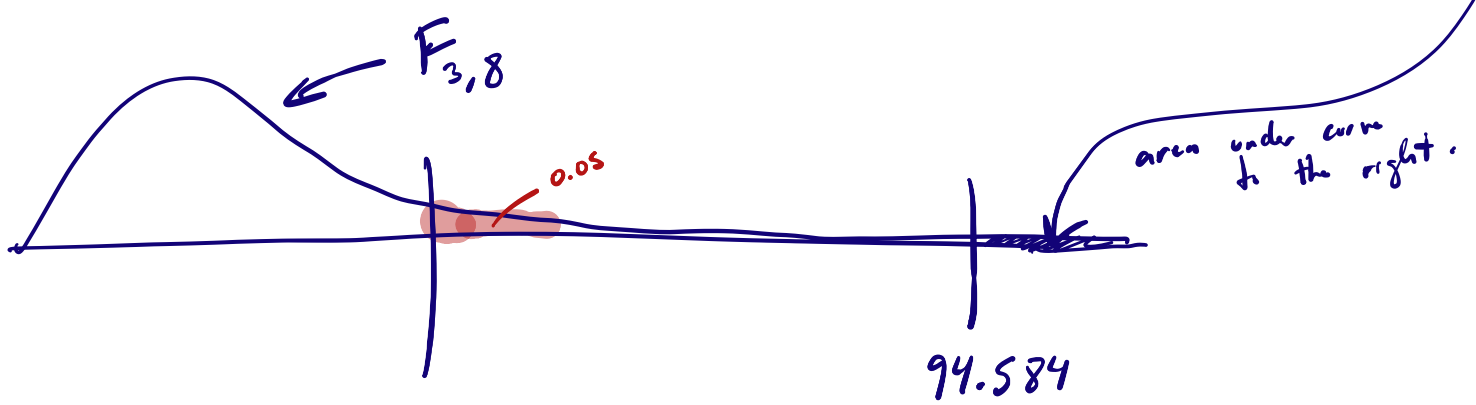
$\boxed{K = 4, \quad N = 12}$

	Df	SS	MS	F	p-val
packaging	3 (K-1)	32.873	10.9576	94.584	small
Residual	8 (N-K)	0.927	0.1154		

$\frac{32.873}{3}$

$\frac{0.927}{8}$

$\frac{10.9576}{0.1154}$



$F_{3,8,0.05} = 4.066$

$\text{qf}(0.95, 3, 8) = 4.066$

```
# 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("CO2",3))

packaging <- as.factor(packaging)

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

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

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$(0, \sigma_\varepsilon^2)$.

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

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

deviations of response values around treatment means.

(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.

Use `plot()` on the output of `lm()`.

Exercise: Check the diagnostic plots for the steaks example.

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

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

$k = 3$

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

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