STAT 135
10. ANOVA

Spring 2022

Lecturer: Dr Rebecca Barter (*she/her*)
Office hours: Tu 9:30-10:30 (in person), Th 4-5pm (virtual)
Office: Evans 339

Email: rebeccabarter@berkeley.edu
Twitter: @rlbarter
GitHub: rlbarter

The χ^2-distribution

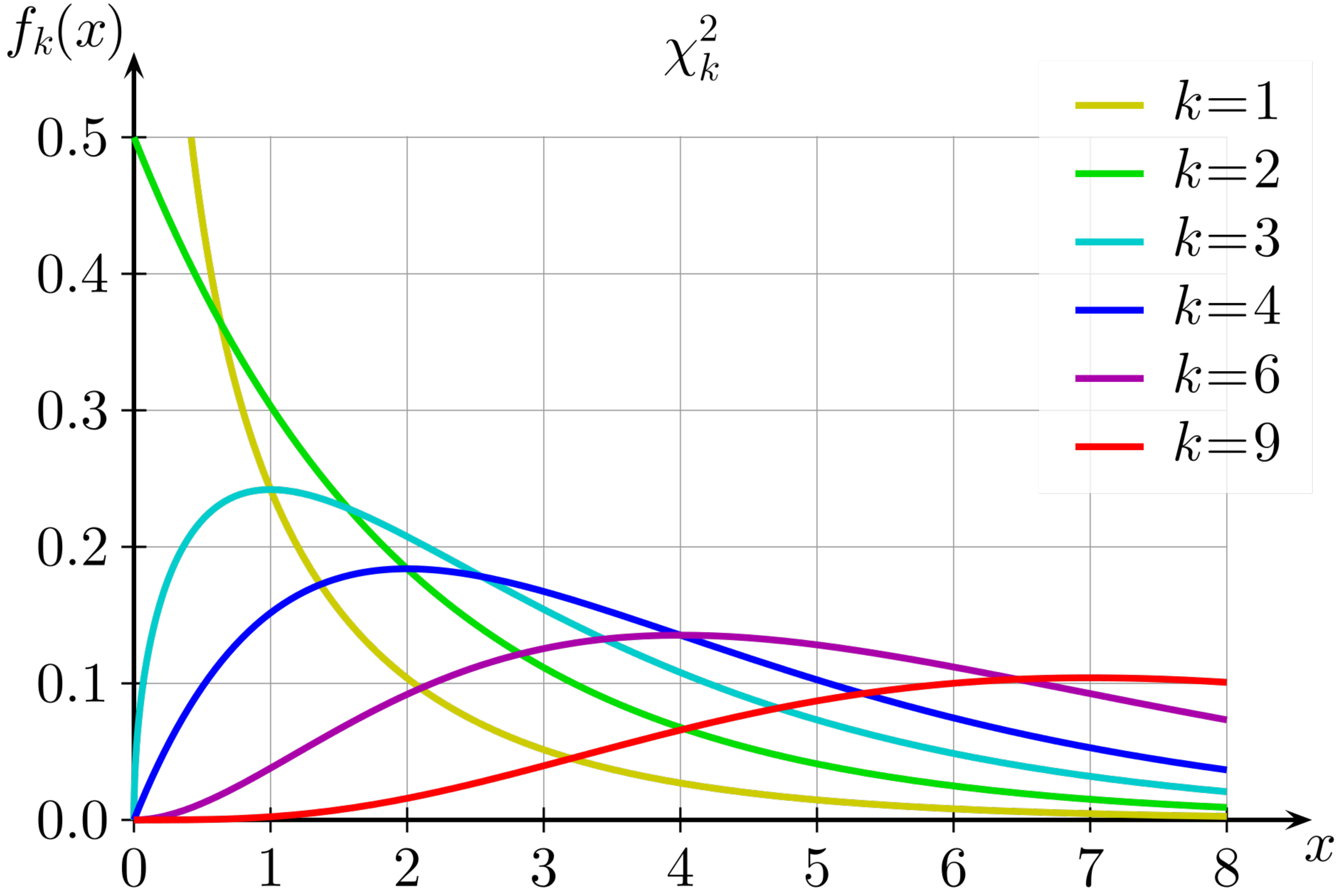
The χ^2-distribution

So far we have primarily dealt with test statistics that have **Normal**, t, or occasionally **Binomial** distributions

There are a whole class of tests for which the test statistic has a χ^2 distribution

Before we dive into these tests, let's take a look at the χ_k^2 distribution

There is one parameter k (degrees of freedom)



The χ^2-distribution and the Normal distribution

The χ^2 distribution is related to the sum of squared normal RVs

If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$

If $Z_i \sim N(0,1)$, then $\sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$

If $Z_i \sim N(0,1)$, then $\sum_{i=1}^{n} (Z_i - \bar{Z})^2 \sim \chi^2_{n-1}$

If $Y_i \sim N(0,\sigma^2)$, then $\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{n-1}$

If $X \sim \chi^2_{k_1}$ and $Y \sim \chi^2_{k_2}$, then $\frac{X/k_1}{Y/k_2} \sim F_{k_1,k_2}$

The χ^2-distribution and the Gamma distribution

The χ^2 distribution is related to the Gamma distribution

$$\text{If } U \sim \chi_n^2, \quad \text{then } U \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$$

Which means that the density of the χ_n^2 distribution is

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, \qquad x \geq 0$$

Analysis **O**f **Va**riance (ANOVA)

ANOVA

So far we have been looking at tests for comparing **two** populations

What if we have **multiple populations** whose means we want to compare?

"Do all of our groups come from populations with the same mean?"

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$$

H_1 : at least one population has different mean

ANOVA

We need to identify whether observed differences in the means are real and not just reflecting the underlying variation in the data

ANOVA answers this question by comparing:

1. The variability **between** the groups
2. The variability **within** the groups

If the variability between the groups is not significantly greater than the variability within the groups, then it is likely that the observed differences between the group means are simply due to chance.

ANOVA

We need to make the following **strong assumptions:**

1. The observations in each sample are independent
2. Each sample is **Normally distributed**
3. The populations have common variance σ^2

ANOVA

Suppose that we have a sample of J observations from each of I populations (groups)

G_1	G_2	$\dots$	G_I
Y_{11}	Y_{21}	$\dots$	Y_{I1}
Y_{12}	Y_{22}	$\dots$	Y_{I2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Y_{1J}	Y_{2J}	$\dots$	Y_{IJ}

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I$$

H_1 : at least one population has different mean

Where μ_k is the population mean for population k

ANOVA

To conduct ANOVA, we formulate a *model* for the data

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

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

- Y_{ij} is the observed measurement ("response") for the jth observation from group i
- μ is the overall global mean aggregated across all groups
- α_i is an offset from the global mean for each group such that $\mu_i = \mu + \alpha_i$
- ϵ_{ij} is a random error term for the jth observation in group i that allows the observation to be different from its group mean

ANOVA

To conduct ANOVA, we formulate a *model* for the data

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$$

What is this H_0 equivalent to in the context of this model?

Since $\mu_i = \mu + \alpha_i$, our H_0 is equivalent to:

$H_0 : \alpha_i = 0$ for all i

Notation

Let's define some notation

- Y_{ij} represents the jth observation in group i
- $\bar{Y}_i$ represents the mean in group i

$$\bar{Y}_i = \frac{1}{J}\sum_{j=1}^{J} Y_{ij}$$

- $\bar{Y}$ represents the mean across **all** observations

$$\bar{Y} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij} = \frac{1}{IJ}\sum_{i,j} Y_{ij}$$

Sum of squares

We can capture the different types of variance using sum of squares

Total sum of squares *(overall variance)*: compares each observation to the overall mean

$$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y})^2$$

Between sum of squares *(variance between groups)*: compares each group mean to the overall mean

$$SS_B = J \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2$$

Within sum of squares *(variance within groups)*: compares observation to its corresponding group mean

$$SS_W = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_i)^2$$

Sum of squares

Some algebra will reveal that the **total sum of squares** can be decomposed into the sum of the **between sum of squares** and the **within sum of squares**

$$SS_T = SS_B + SS_W$$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y})^2 = J \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2 + \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_i)^2$$

Sum of squares degrees of freedom

Each source of variability has its own associated degrees of freedom

$\boldsymbol{SS_T}$ compares the IJ observations to the overall mean, so it has $\boldsymbol{IJ-1}$ degrees of freedom

$$SS_T = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y})^2$$

$\boldsymbol{SS_B}$ compares the I group means to the overall mean, so it has $\boldsymbol{I-1}$ degrees of freedom

$$SS_B = J\sum_{i=1}^{I}(\bar{Y}_i-\bar{Y})^2$$

$\boldsymbol{SS_W}$ compares the IJ observations to the I group means, so it has $\boldsymbol{IJ-I} = I(J-1)$ degrees of freedom

$$SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_i)^2$$

Notice that $IJ-1=(IJ-I)+(I-1)$ So $df_T = df_B + df_W$

Sum of squares degrees of freedom

Recall that to test H_0, we want to compare the between sum of squares to the within sum of squares

Thus we define our test statistic to be

$$\boxed{F = \frac{SS_B/(I-1)}{SS_W/(I(J-1))} \overset{H_0}{\sim} F_{I-1,I(J-1)}}$$

ANOVA

In summary, if we want to test

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_I$$

H_1 : at least one population has different mean

We calculate the test statistic

$$F = \frac{SS_B/(I-1)}{SS_W/(I(J-1))} \quad \overset{H_0}{\sim} F_{I-1,I(J-1)}$$

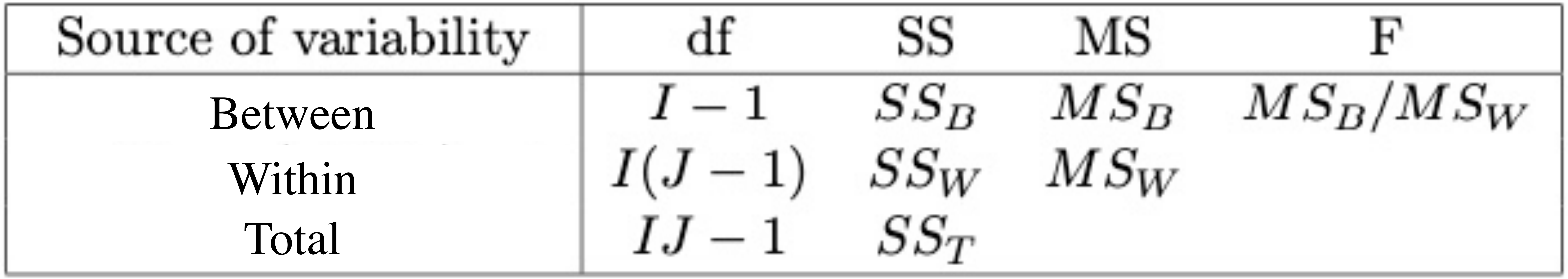
And the corresponding p-value

p-value = $= P(F \geq f)$

Note a more extreme test statistic is always **larger** for this test

ANOVA

It is common to summarize the data in the following table, from which the test statistic can just be read off



Source of variability	df	SS	MS	F
Between	$I-1$	SS_B	MS_B	MS_B/MS_W
Within	$I(J-1)$	SS_W	MS_W	
Total	$IJ-1$	SS_T		

Where

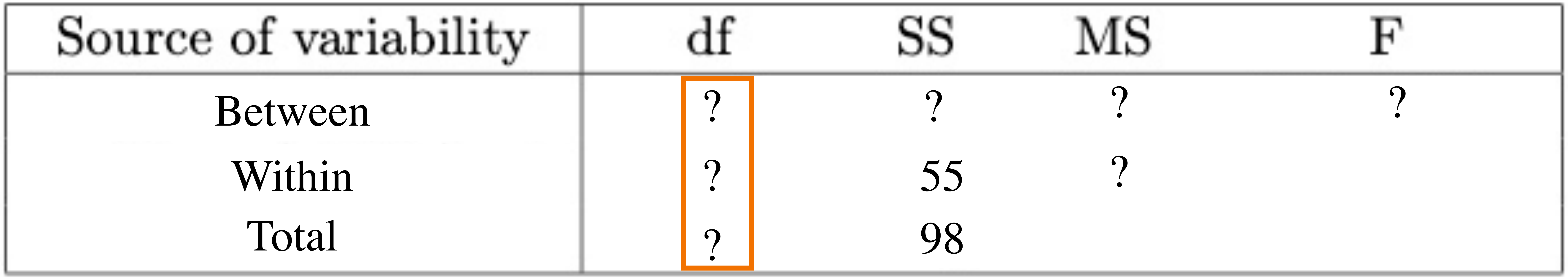
$$MS_B = SS_B/(I-1)$$

$$MS_W = SS_W/(I(J-1))$$

are the **mean sum of squares**

ANOVA table example

Suppose we have 6 groups and 4 samples from each group, but we only told you



Source of variability	df	SS	MS	F
Between	?	?	?	?
Within	?	55	?	
Total	?	98		

i.e. $SS_W = 55$ $SS_T = 98$

We know $I = 6$ and $J = 4$, so we can fill in the degrees of freedom

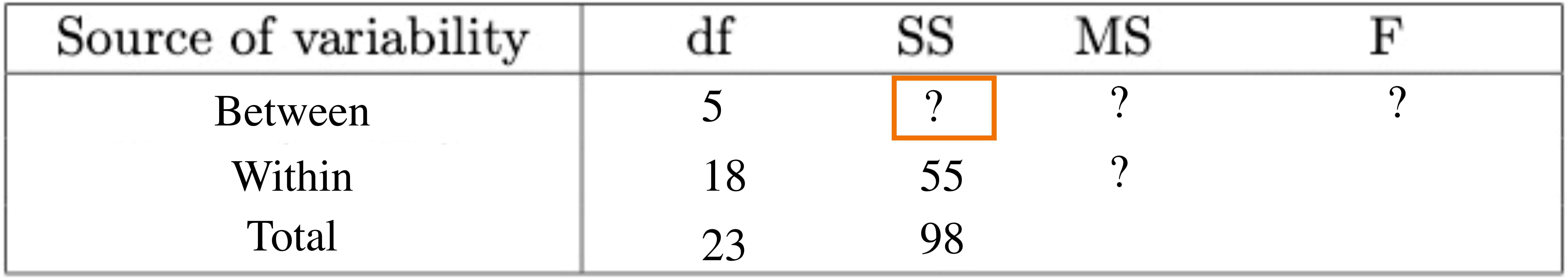
$$df_B = I - 1 = 5 \qquad df_W = I(J-1) = 6 \times 3 = 18$$

$$df_T = IJ - 1 = 6 \times 4 - 1 = 23$$

(Check that they all add up!)

ANOVA table example

Suppose we have 6 groups and 4 samples from each group, but we only told you



Source of variability	df	SS	MS	F
Between	5	?	?	?
Within	18	55	?	
Total	23	98		

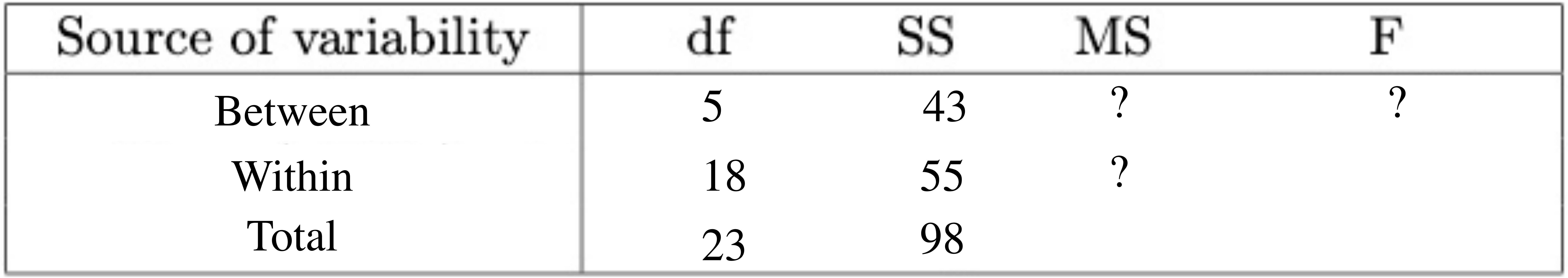
i.e. $SS_W = 55$ $SS_T = 98$

We know $SS_T = SS_B + SS_W$, so

$$SS_B = SS_T - SS_W = 98 - 55 = 43$$

ANOVA table example

Suppose we have 6 groups and 4 samples from each group, but we only told you

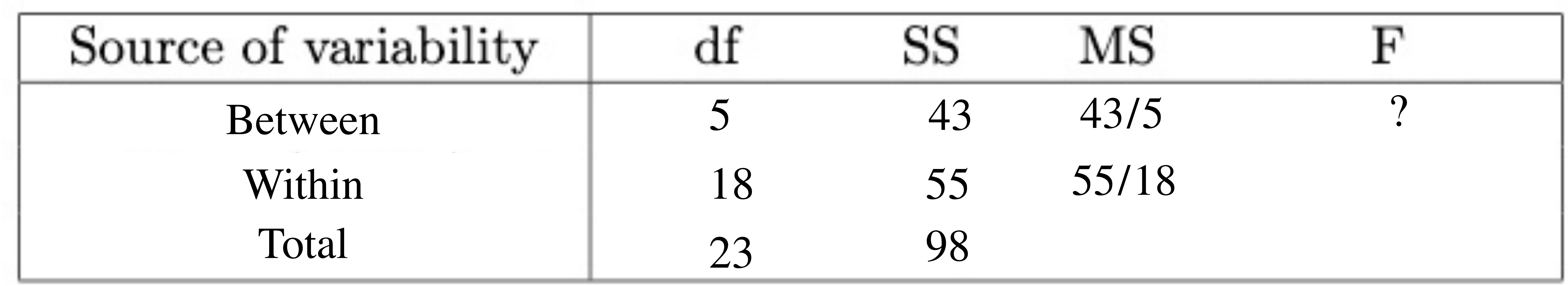


Source of variability	df	SS	MS	F
Between	5	43	?	?
Within	18	55	?	
Total	23	98		

The MS column just involves dividing the SS column by the df column

ANOVA table example

Suppose we have 6 groups and 4 samples from each group, but we only told you

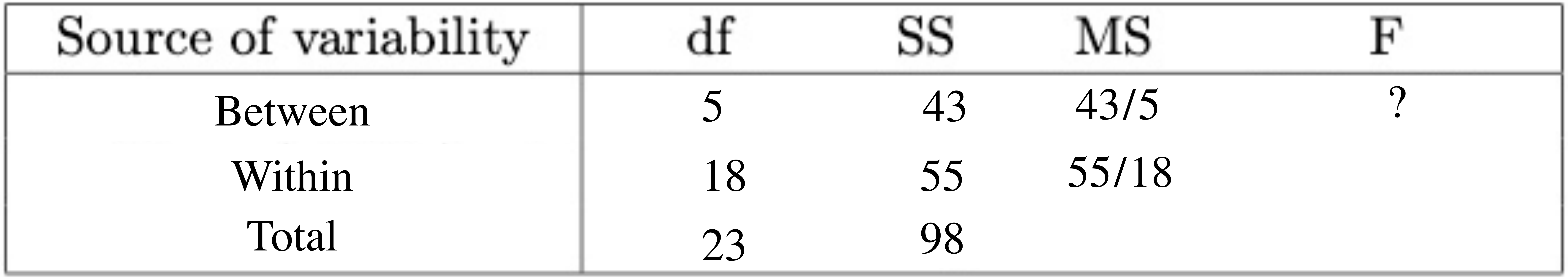


Source of variability	df	SS	MS	F
Between	5	43	43/5	?
Within	18	55	55/18	
Total	23	98		

The MS column just involves dividing the SS column by the df column

ANOVA table example

Suppose we have 6 groups and 4 samples from each group, but we only told you

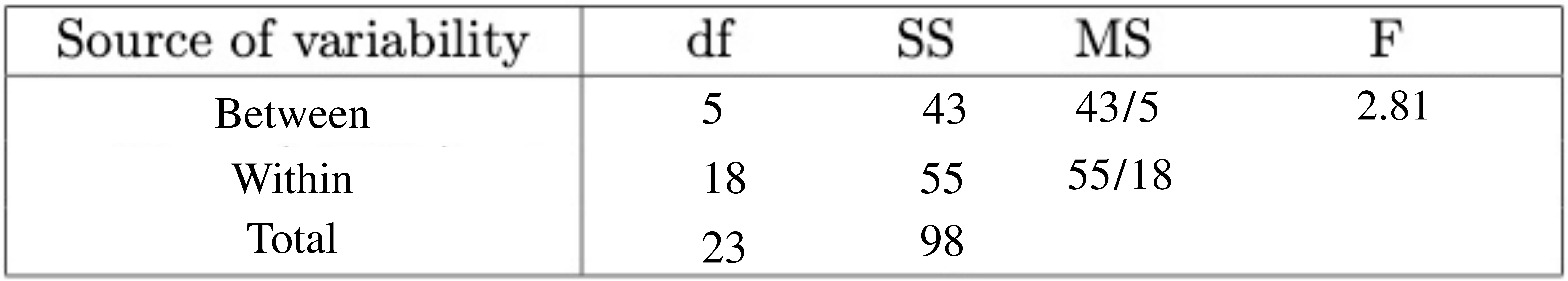


Source of variability	df	SS	MS	F
Between	5	43	43/5	?
Within	18	55	55/18	
Total	23	98		

To get the F value, divide MS_B by MS_W

ANOVA table example

Suppose we have 6 groups and 4 samples from each group, but we only told you

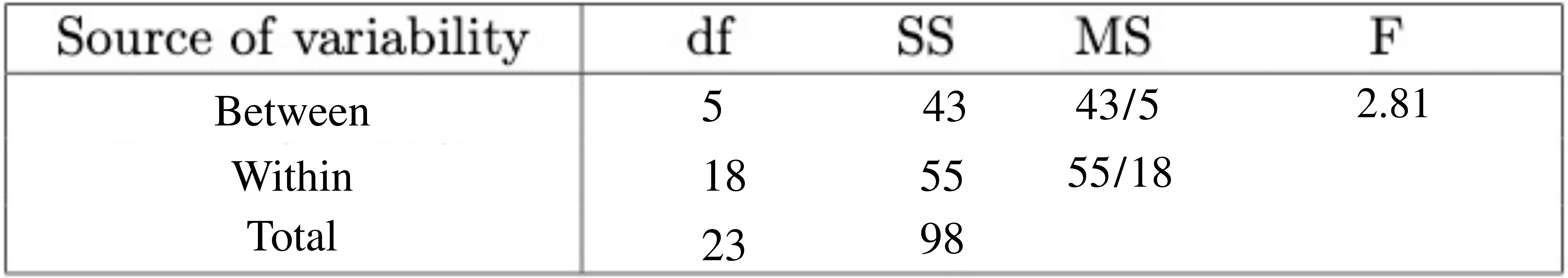


Source of variability	df	SS	MS	F
Between	5	43	43/5	2.81
Within	18	55	55/18	
Total	23	98		

To get the F value, divide MS_B by MS_W

ANOVA table example

Suppose we have 6 groups and 4 samples from each group, but we only told you



Source of variability	df	SS	MS	F
Between	5	43	43/5	2.81
Within	18	55	55/18	
Total	23	98		

So our p-value is

$$P(F \geq 2.81) = 0.048 \qquad \text{Where } F \sim F_{5,18}$$

Our evidence is borderline whether we should reject the null the the means of the 6 groups are the same, but technically less than 0.05

ANOVA example

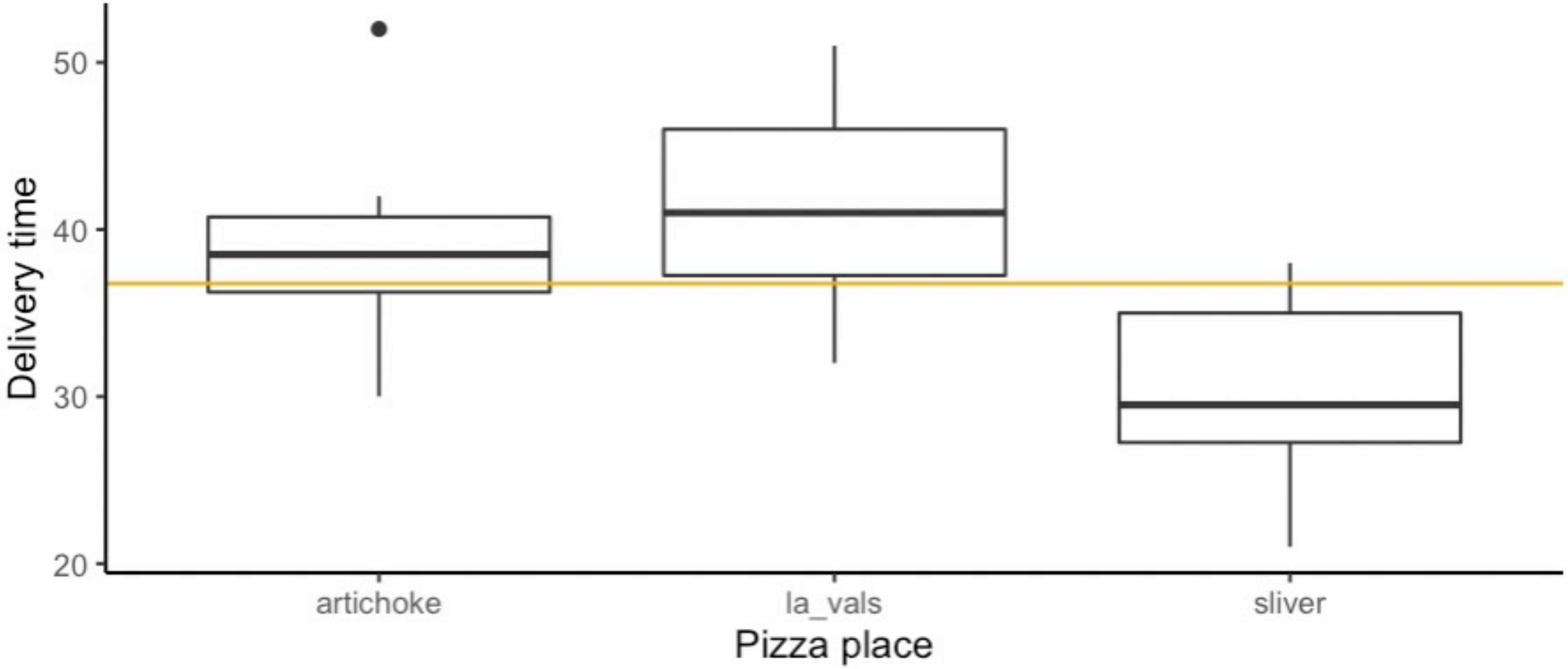
This week I went pizza crazy and ordered 10 pizzas from each of La Val's, Sliver and Artichoke Basille, and recorded how long it took them to deliver to my office in Evans.

I want to test the hypothesis that all three pizza places have the same average delivery time

La Val's	Sliver	Artichoke
32	27	52
48	32	40
51	37	41
47	21	36
41	25	32
35	28	38
37	36	37
41	38	42
43	29	39
38	30	30

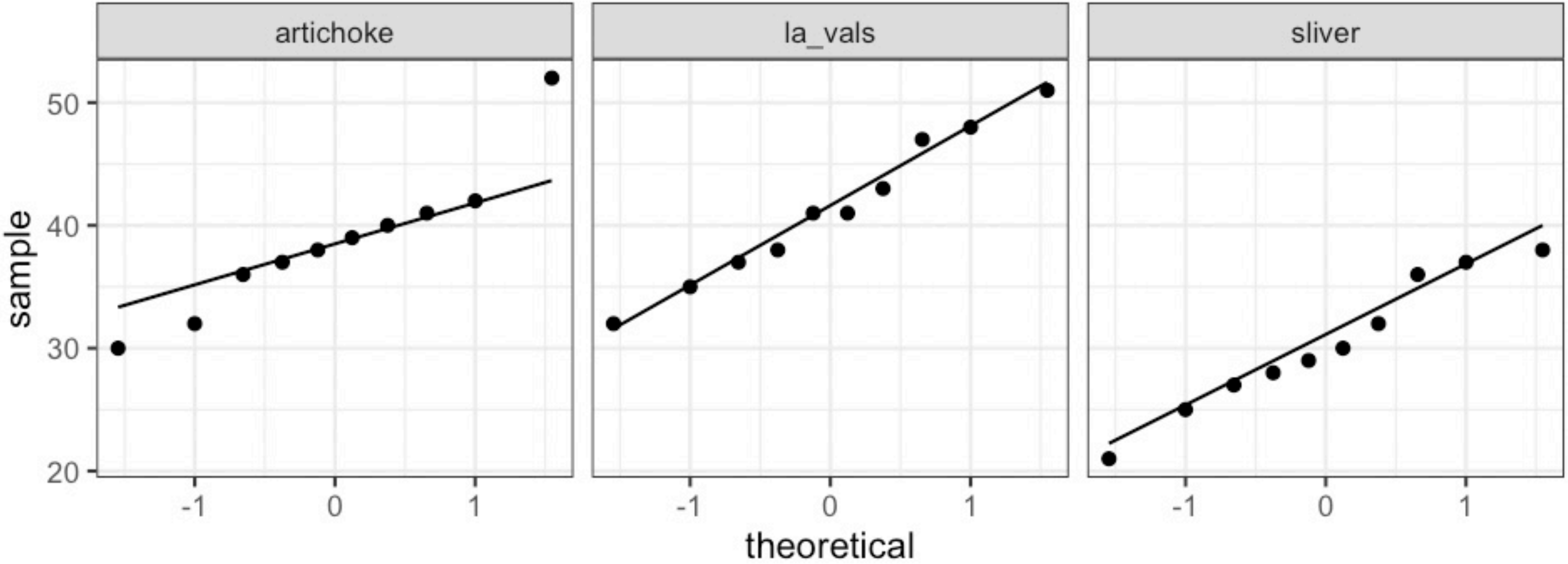
ANOVA example

To start with, let's *look* at our data! Do the means look different?



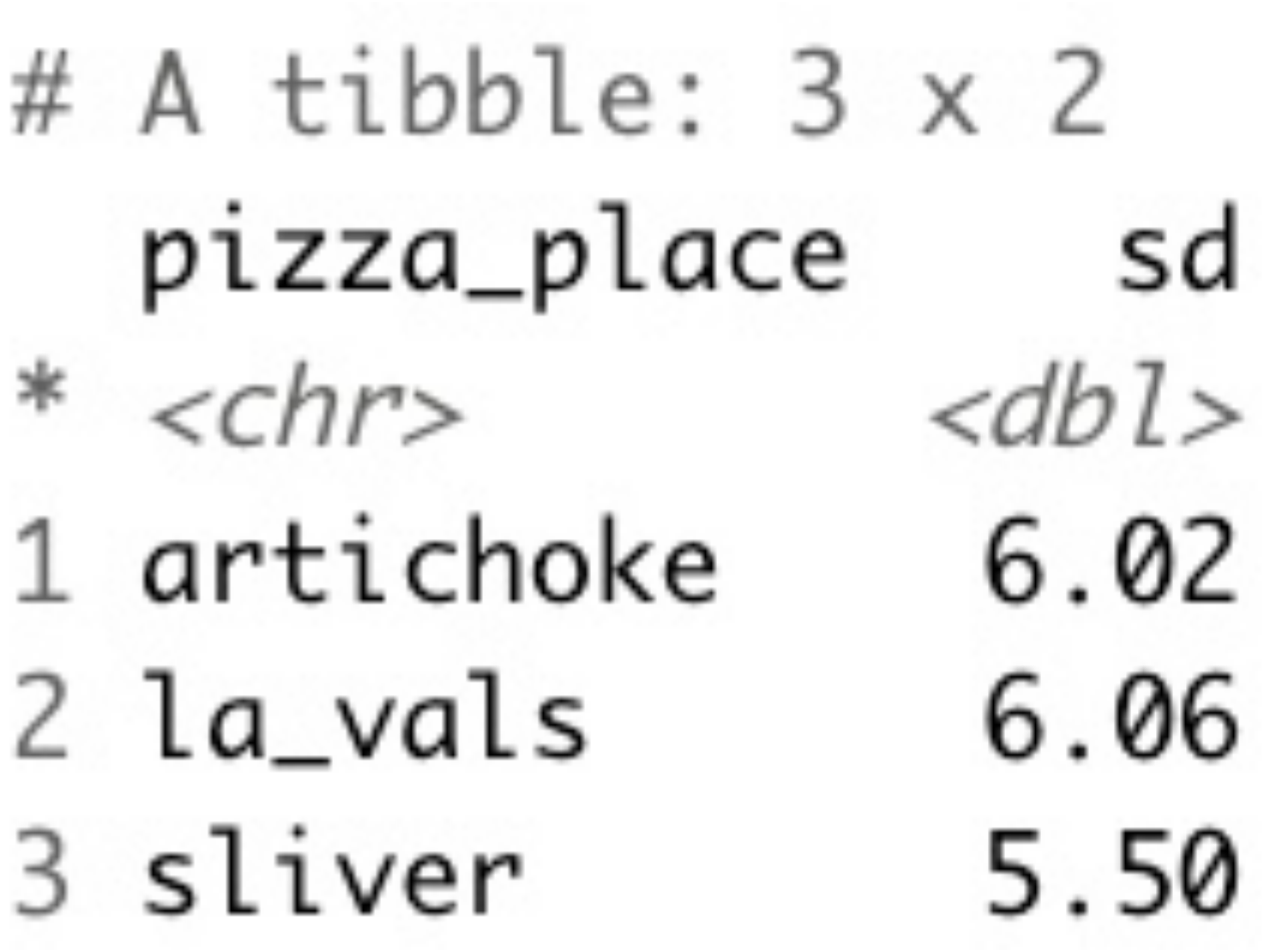
ANOVA example: assess normality

ANOVA assumed normality so let's check that each sample looks normal



ANOVA example: assess common variance

ANOVA assumed common variance, so let’s look at the sample SD



```
# A tibble: 3 x 2
  pizza_place    sd
* <chr>       <dbl>
1 artichoke    6.02
2 la_vals      6.06
3 sliver       5.50
```

ANOVA example: compute SS

We have 3 pizza places ($I = 3$) and 10 observations from each pizza place ($J = 10$)

	La Val’s	Sliver	Artichoke
	32	27	52
	48	32	40
	51	37	41
	47	21	36
	41	25	32
	35	28	38
	37	36	37
	41	38	42
	43	29	39
	38	30	30
Group mean	41.3	30.3	38.7
Global mean	36.8		

$$SS_B = J \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2$$

$$= 10\left[(41.3 - 36.8)^2 + (30.3 - 36.8)^2 + (38.7 - 36.8)^2\right]$$

$$= 661.1$$

ANOVA example: compute SS

We have 3 pizza places ($I = 3$) and 10 observations from each pizza place ($J = 10$)

	La Val's	Sliver	Artichoke
	32	27	52
	48	32	40
	51	37	41
	47	21	36
	41	25	32
	35	28	38
	37	36	37
	41	38	42
	43	29	39
	38	30	30
Group mean	41.3	30.3	38.7
Global mean	36.8		

$$SS_B = J\sum_{i=1}^{I}(\bar{Y}_i - \bar{Y})^2 \quad = 661.1$$

$$SS_W = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_i)^2$$

$$= (32 - 41.3)^2 + (48 - 41.3)^2 + \overset{\text{27other terms}}{\overset{\downarrow}{\ldots}} + (30 - 38.7)^2$$

$$= 928.3$$

ANOVA example: compute SS

We have 3 pizza places ($I = 3$) and 10 observations from each pizza place ($J = 10$)

	La Val’s	Sliver	Artichoke
	32	27	52
	48	32	40
	51	37	41
	47	21	36
	41	25	32
	35	28	38
	37	36	37
	41	38	42
	43	29	39
	38	30	30
Group mean	41.3	30.3	38.7
Global mean	36.8		

$$SS_B = J \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2 \quad = 661.1$$

$$SS_W = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_i)^2 \quad = 928.3$$

$$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y})^2$$

$$= (32 - 36.8)^2 + (48 - 36.8)^2 + \ldots + (30 - 36.8)^2$$

$$= 1589.4 \quad (= 661.1 + 928.3)$$

ANOVA example: compute SS

We have 3 pizza places ($I = 3$) and 10 observations from each pizza place ($J = 10$)

	La Val's	Sliver	Artichoke
	32	27	52
	48	32	40
	51	37	41
	47	21	36
	41	25	32
	35	28	38
	37	36	37
	41	38	42
	43	29	39
	38	30	30
Group mean	41.3	30.3	38.7
Global mean	36.8		

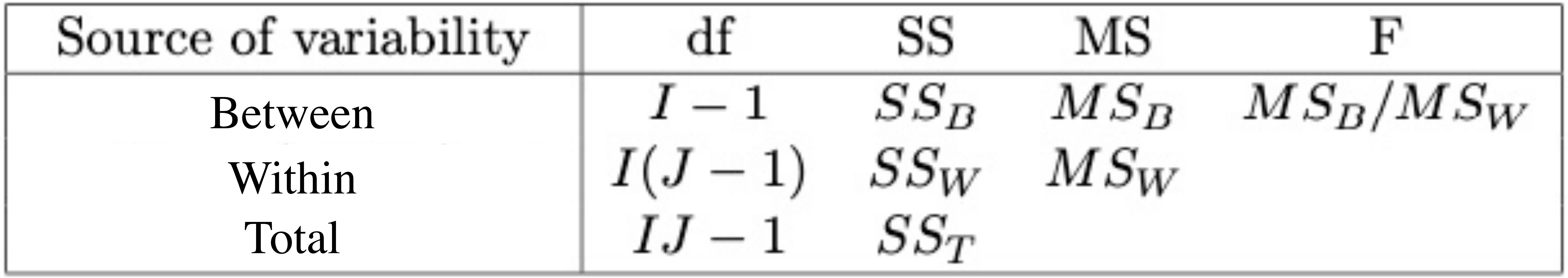
$$SS_B = J \sum_{i=1}^{I} (\bar{Y}_i - \bar{Y})^2 \quad = 661.1$$

$$SS_W = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y}_i)^2 \quad = 928.3$$

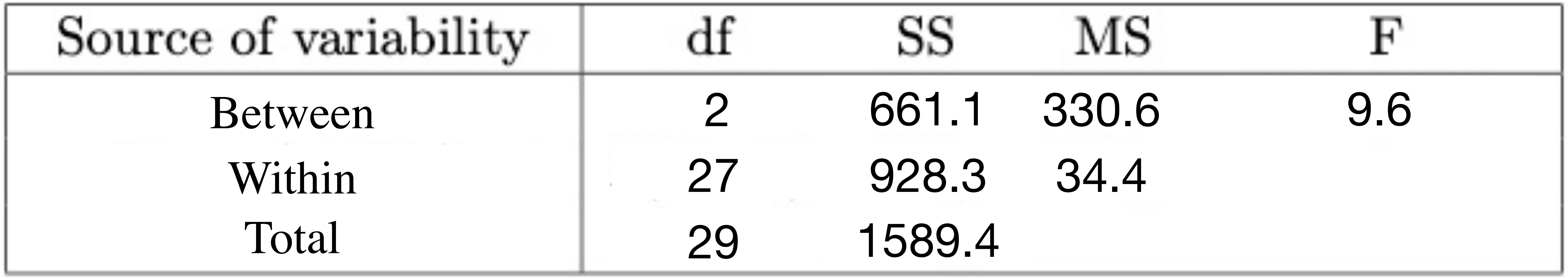
$$SS_T = \sum_{i=1}^{I} \sum_{j=1}^{J} (Y_{ij} - \bar{Y})^2 \quad = 1589.4$$

ANOVA example: fill in table

We have 3 pizza places ($I = 3$) and 10 observations from each pizza place ($J = 10$)



Source of variability	df	SS	MS	F
Between	$I-1$	SS_B	MS_B	MS_B/MS_W
Within	$I(J-1)$	SS_W	MS_W	
Total	$IJ-1$	SS_T		



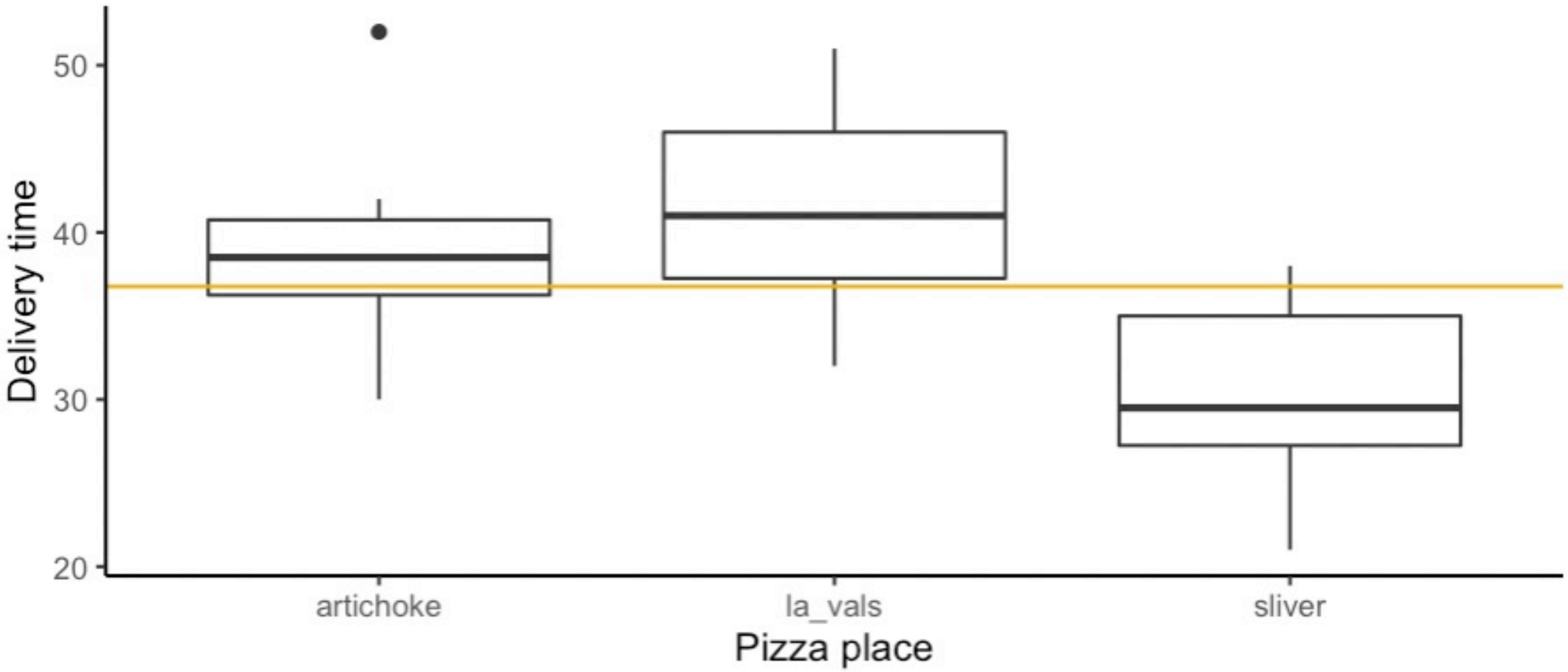
Source of variability	df	SS	MS	F
Between	2	661.1	330.6	9.6
Within	27	928.3	34.4	
Total	29	1589.4		

P-value = $P(F \geq 9.6) = 0.0007$ Where $F \sim F_{I-1,I(J-1)} = F_{2,27}$

So we reject the null that all of the means are the same!

ANOVA example

That we rejected the null isn't terribly surprising…



ANOVA example in R

See anova.R

Non-parametric ANOVA

The Kruskal-Wallis test

Kruskal-Wallis

The Kruskal-Wallis test is a generalization of the Mann-Whitney test to multiple groups

It tests whether each sample comes from populations with the same distribution:

$$H_0 : G_1 = G_2 = \ldots . = G_I$$

The observations are assumed to be independent, but no distributional form is assumed

Notation

Similarly to the Mann-Whitney test, the observations are pooled together and ranked.

- R_{ij} represents the rank of Y_{ij} in the combined sample
- $\bar{R}_i$ represents the mean rank in group i

$$\bar{R}_i = \frac{1}{J} \sum_{j=1}^{J} R_{ij}$$

- $\bar{R}$ represents the mean rank across **all** observations

$$\bar{R} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} R_{ij} = \frac{IJ+1}{2}$$

Kruskal-Wallis

As in ANOVA, define the between sum of squares (a measure of the variance of the group means) as

$$SS_B = J \sum_{i=1}^{I} (\bar{R}_i - \bar{R})^2$$

We can use a scaled version of SS_B as our test statistic because

$$\boxed{K = \frac{12}{IJ(IJ+1)} SS_B \sim \chi^2_{I-1}}$$

A larger SS_B provides stronger evidence against H_0

P-value = $P(K \geq k)$ where $K \sim \chi^2_{I-1}$

Kruskal-Wallis

Our test statistic K can also be computed using:

$$\boxed{K = \frac{12}{IJ(IJ+1)}\left(J\sum_{i=1}^{I}\bar{R}_i^2\right) - 3(IJ+1) \sim \chi^2_{I-1}}$$

This formula is equivalent to the formula from the previous slide

A larger SS_B provides stronger evidence against H_0

P-value = $P(K \geq k),$ where $K \sim \chi^2_{I-1}$

Kruskal-Wallis

As in ANOVA, define the between sum of squares (a measure of the variance of the group means) as

$$SS_B = J \sum_{i=1}^{I} (\bar{R}_i - \bar{R})^2$$

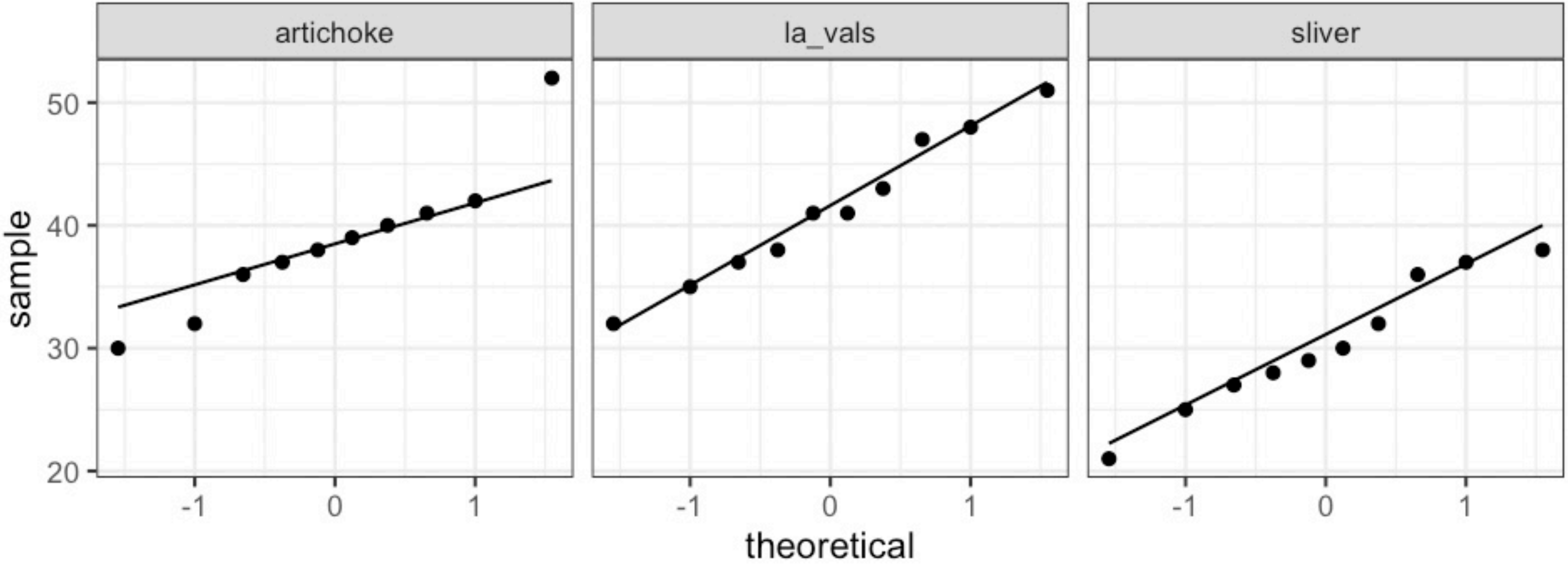
We can use SS_B to test the null hypothesis that the groups have the same distribution because

$$K = \frac{12}{IJ(IJ+1)} SS_B \sim \chi^2_{I-1}$$

A larger SS_B provides stronger evidence against H_0

Kruskal-Wallis example

If we didn't feel comfortable assuming our pizza delivery times were normal



ANOVA example: compute SS

Original data

	La Val's	Sliver	Artichoke
	32	27	52
	48	32	40
	51	37	41
	47	21	36
	41	25	32
	35	28	38
	37	36	37
	41	38	42
	43	29	39
	38	30	30
Group mean	41.3	30.3	38.7
Global mean	36.8		

Ranks

	La Val's	Sliver	Artichoke
	9	3	30
	28	9	21
	29	15	23
	27	1	12.5
	23	2	9
	11	4	18
	15	12.5	15
	23	18	25
	26	5	20
	18	6.5	6.5
Group mean	20.9	7.6	18

ANOVA example: compute SS

$I = 3, \quad J = 10$

$$K = \frac{12}{IJ(IJ+1)} \left(J \sum_{i=1}^{I} \bar{R}_i^2 \right) - 3(IJ+1) \sim \chi^2_{I-1}$$

$$= \frac{12}{3 \times 10(3 \times 10 + 1)} \left(10(20.9^2 + 7.6^2 + 18^2)\right) - 3(3 \times 10 + 1)$$

$$= 12.6$$

P-value = $P(K \geq 12.6),$ where $K \sim \chi^2_2$

$$= 0.0018$$

Ranks

	La Val's	Sliver	Artichoke
	9	3	30
	28	9	21
	29	15	23
	27	1	12.5
	23	2	9
	11	4	18
	15	12.5	15
	23	18	25
	26	5	20
	18	6.5	6.5
Group mean	20.9	7.6	18

ANOVA example in R

See anova.R