ANOVA and Linear Regression

Ling250: Data Science for Linguistics
C.M. Downey
Spring 2025

Going forward

Going forward

- For the last few statistical tests, we'll go into **very little math**
  - It gets a bit tricky, and this isn't a full statistics course

Going forward

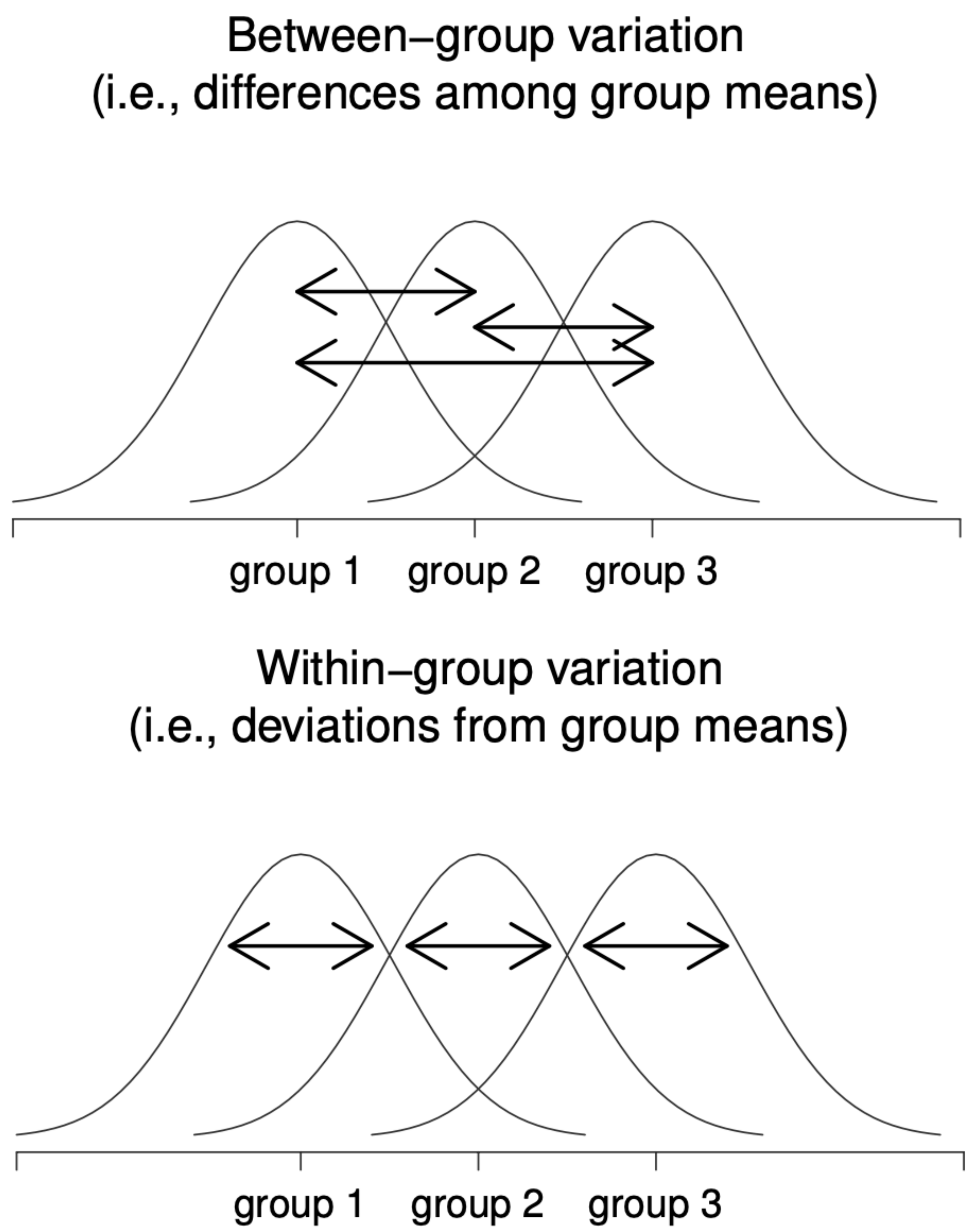
- For the last few statistical tests, we'll go into **very little math**
  - It gets a bit tricky, and this isn't a full statistics course
- We'll also make more **bad assumptions**, for simplicity
  - These tests make **stronger assumptions** about the data
  - There are ways to **check** those assumptions, but we won't
  - If you use these for real, there are **nitty-gritty details** to be considered

Going forward

- For the last few statistical tests, we'll go into **very little math**
  - It gets a bit tricky, and this isn't a full statistics course
- We'll also make more **bad assumptions**, for simplicity
  - These tests make **stronger assumptions** about the data
  - There are ways to **check** those assumptions, but we won't
  - If you use these for real, there are **nitty-gritty details** to be considered
- We'll focus on **when to use** these tests, and how to **use them in R**
  - Our goal is to practice the basics

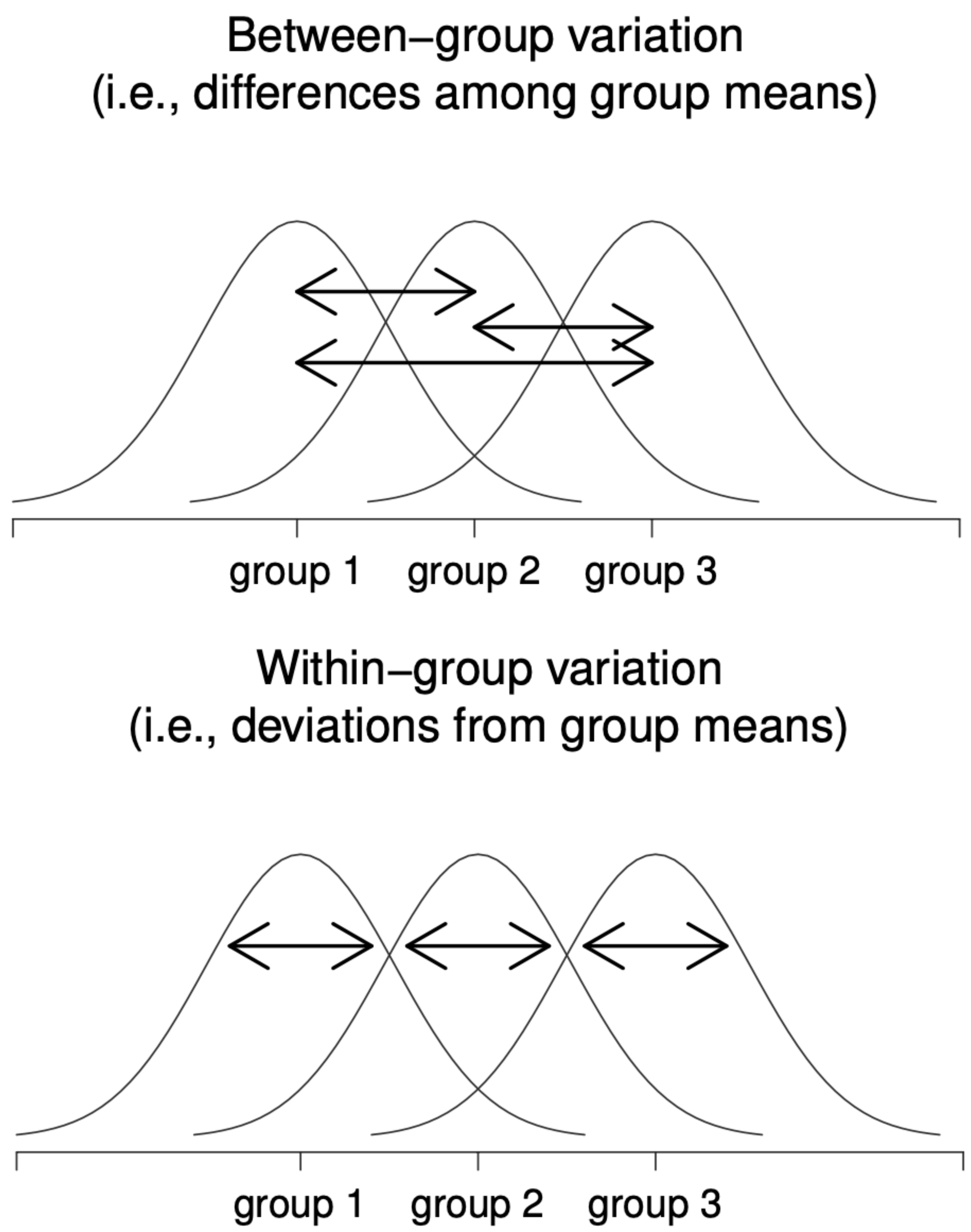
ANOVA

ANOVA


Between–group variation
(i.e., differences among group means)
group 1
group 2
group 3
Within–group variation
(i.e., deviations from group means)
group 1
group 2
group 3

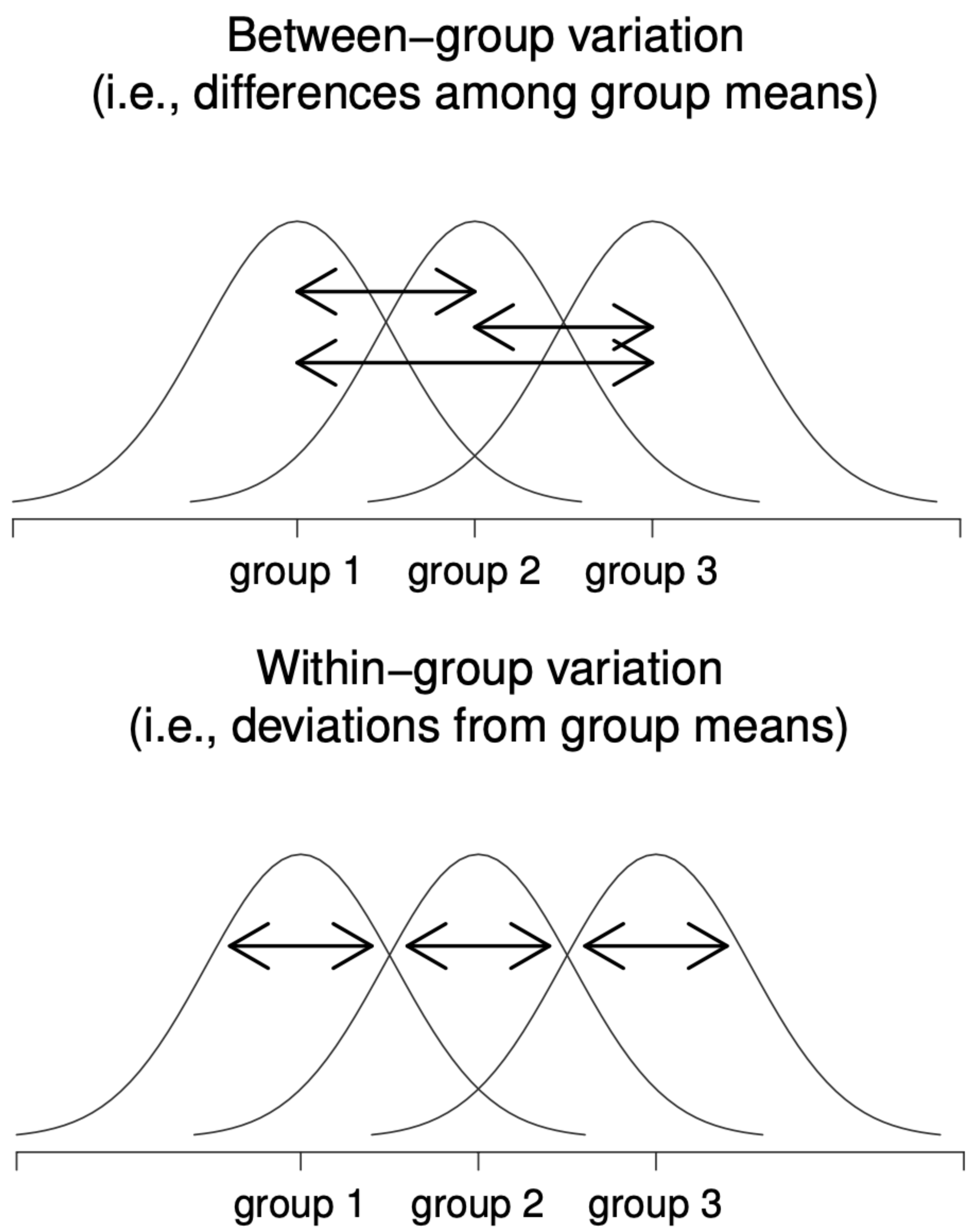
ANOVA

- ANOVA stands for **ANalysis Of VAriance**



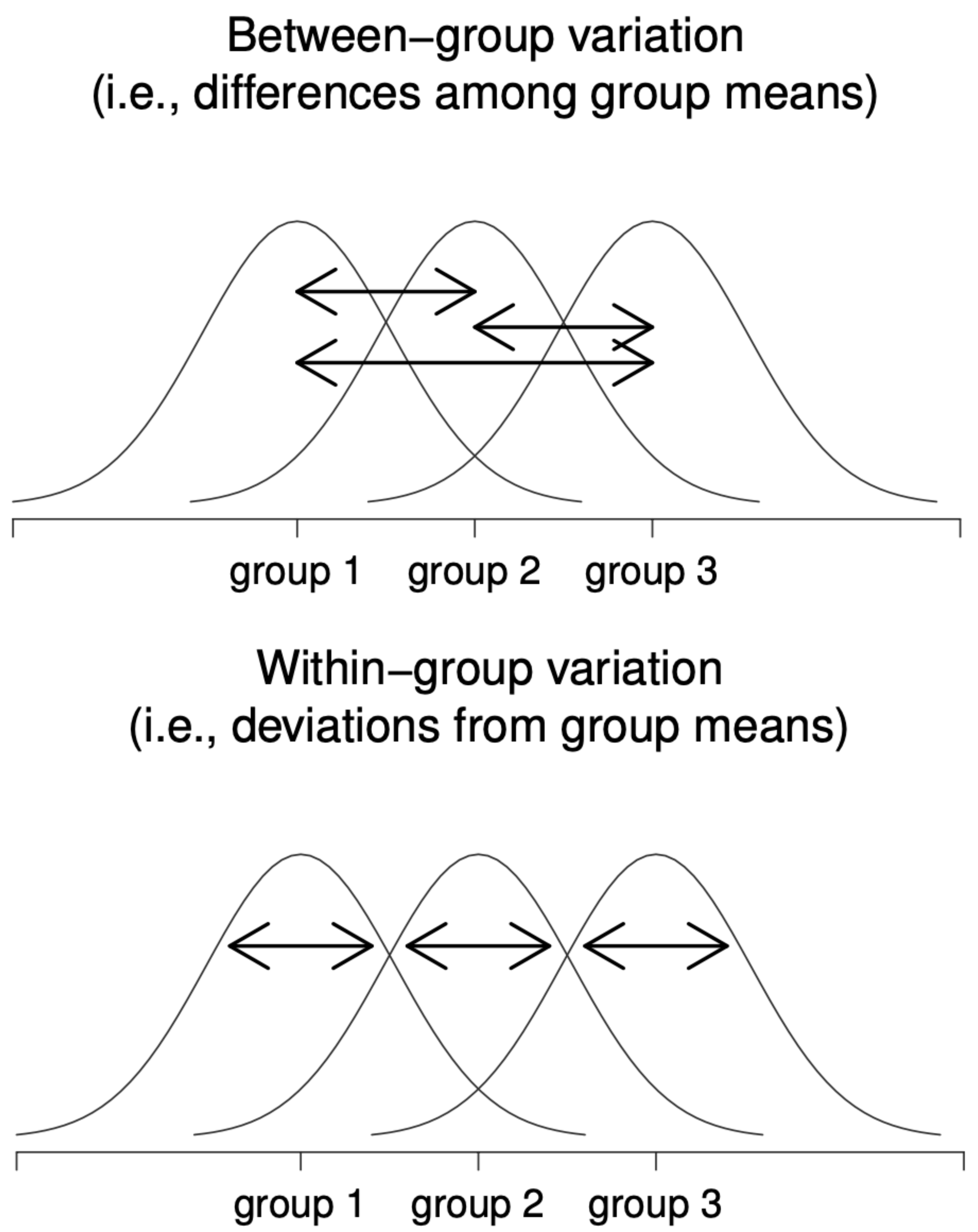
ANOVA

- ANOVA stands for **ANalysis Of VAriance**
- Like t-tests, checks whether samples have the **same population mean**



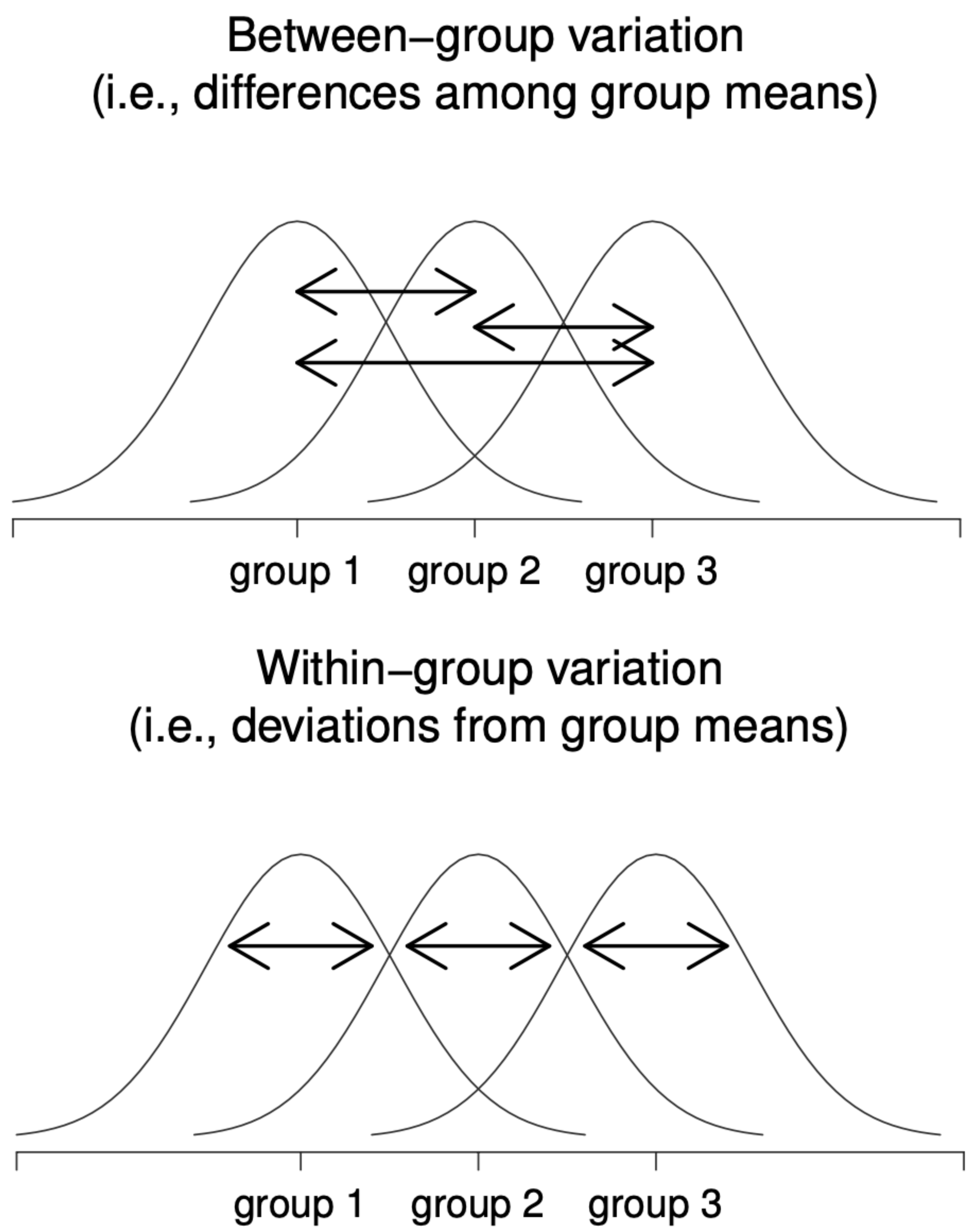
ANOVA

- ANOVA stands for **ANalysis Of VAriance**
- Like t-tests, checks whether samples have the **same population mean**
  - **Unlike** t-tests, tests this for **two or more** groups (t-tests assume two groups)



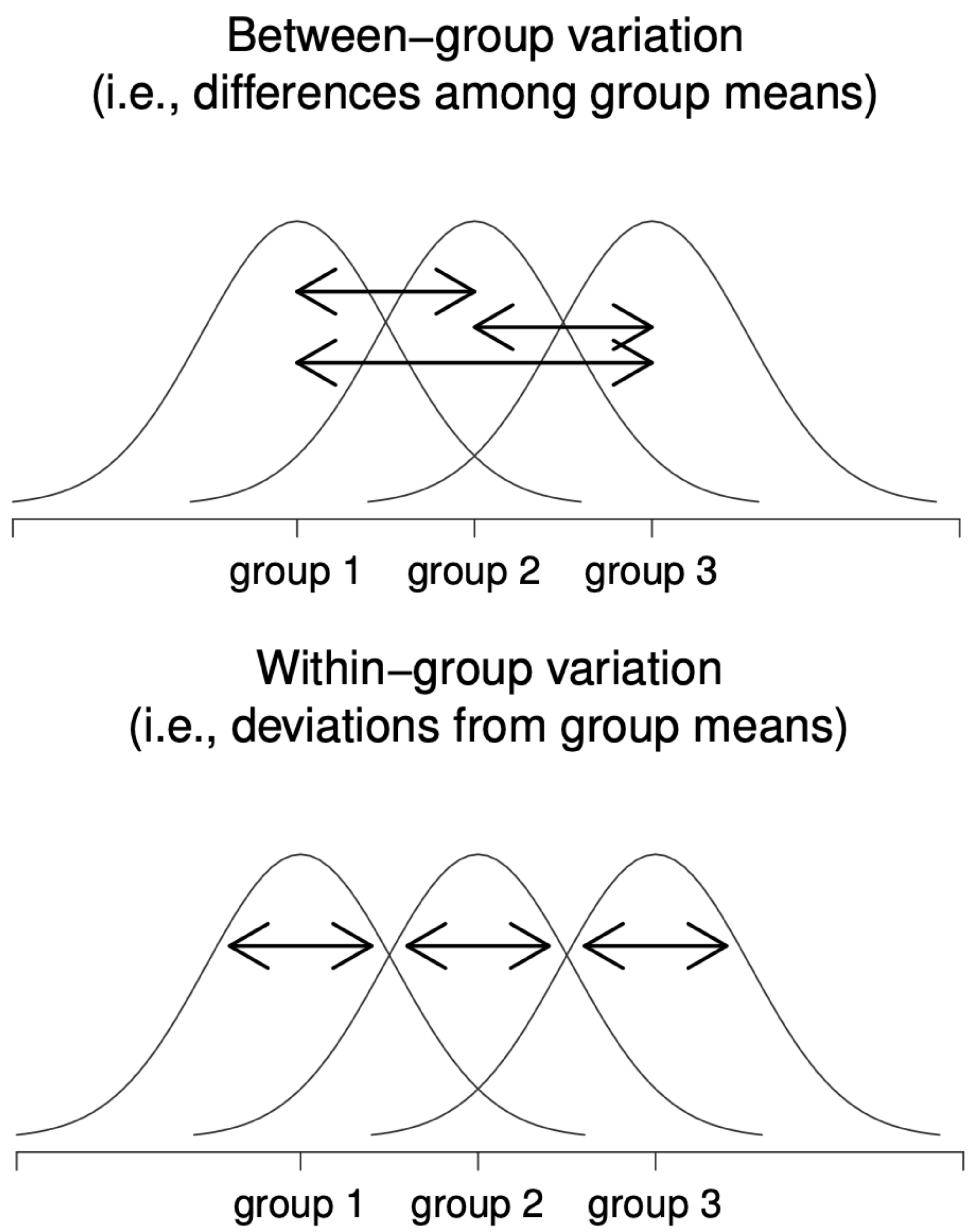
ANOVA

- ANOVA stands for **ANalysis Of VAriance**
- Like t-tests, checks whether samples have the **same population mean**
  - **Unlike** t-tests, tests this for **two or more** groups (t-tests assume two groups)
- Hypotheses



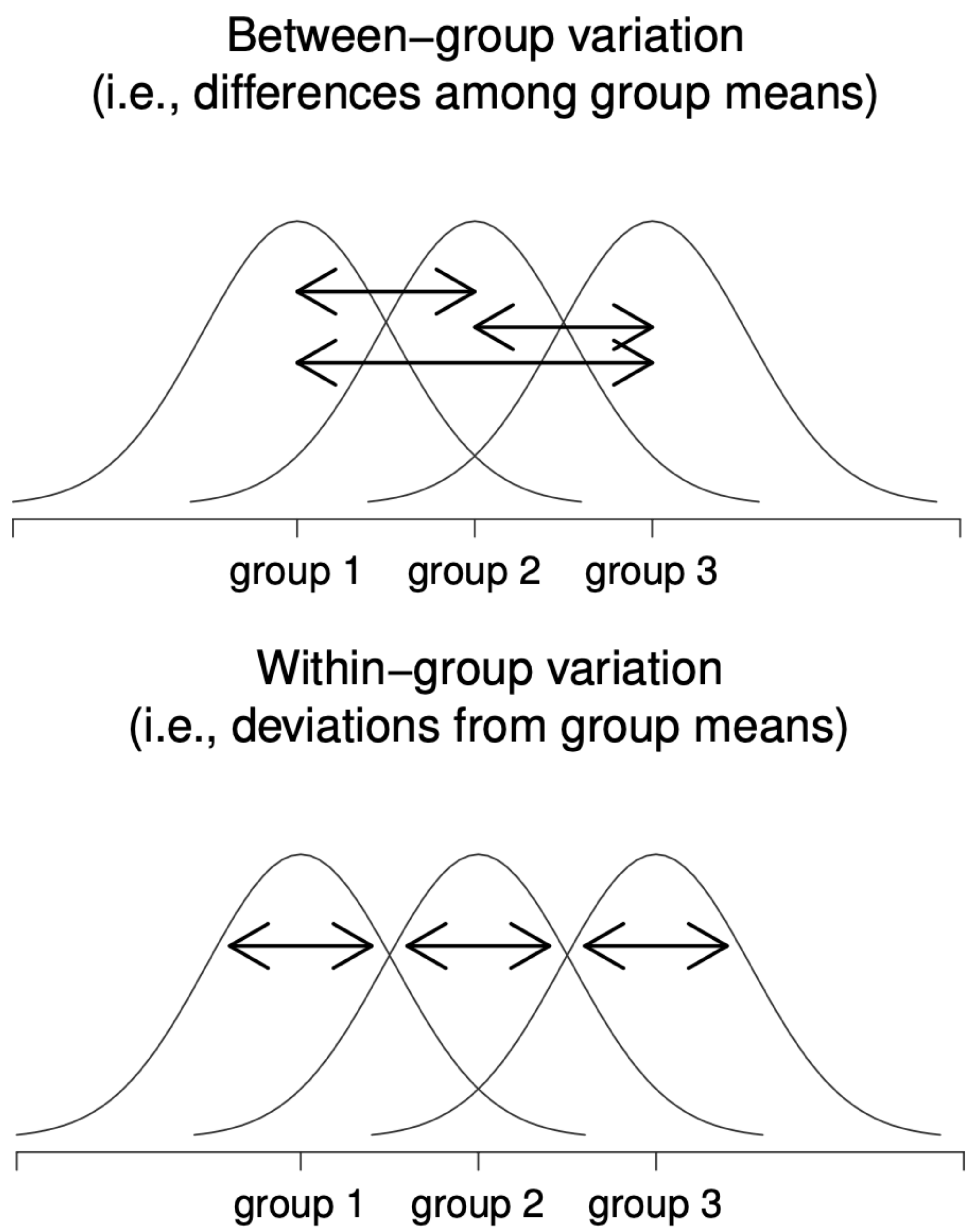
ANOVA

- ANOVA stands for **ANalysis Of VAriance**
- Like t-tests, checks whether samples have the **same population mean**
  - **Unlike** t-tests, tests this for **two or more** groups (t-tests assume two groups)
- Hypotheses
  - $H_0 : \mu_1 = \mu_2 \ldots = \mu_G$



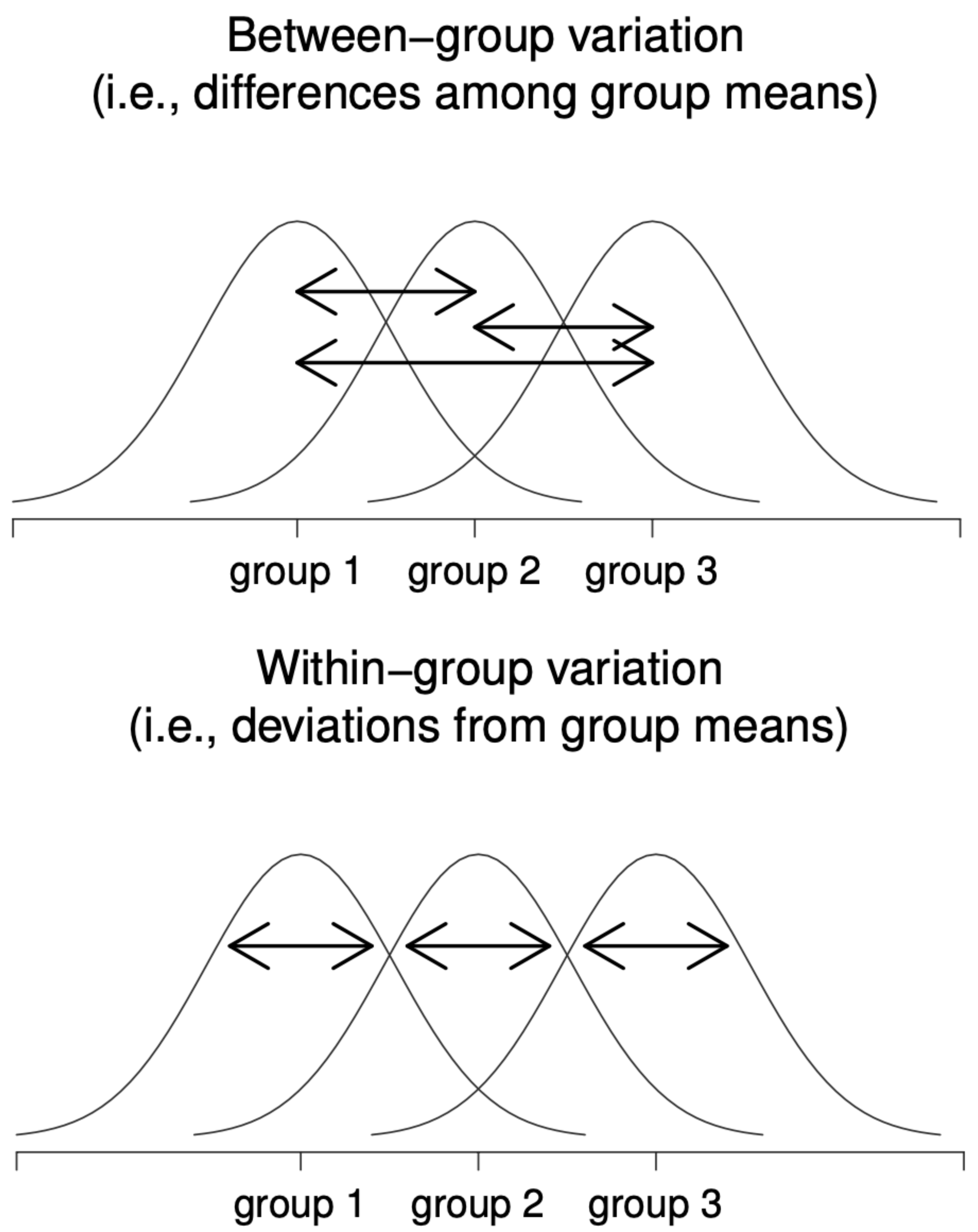
ANOVA

- ANOVA stands for **ANalysis Of VAriance**
- Like t-tests, checks whether samples have the **same population mean**
  - **Unlike** t-tests, tests this for **two or more** groups (t-tests assume two groups)
- Hypotheses
  - $H_0 : \mu_1 = \mu_2 \ldots = \mu_G$
  - H_1 : Not $(\mu_1 = \mu_2 \ldots = \mu_G)$

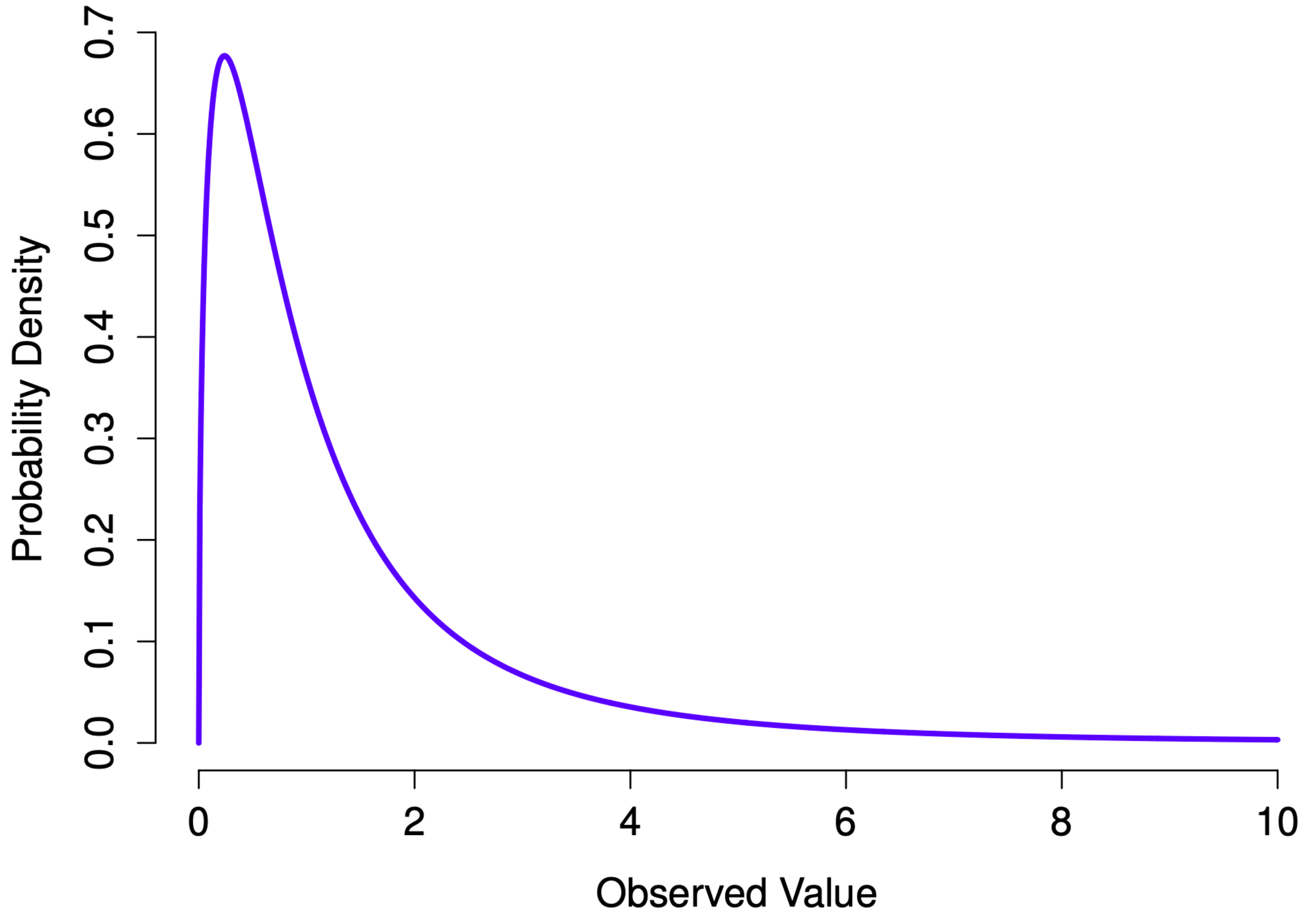


ANOVA

- ANOVA stands for **ANalysis Of VAriance**
- Like t-tests, checks whether samples have the **same population mean**
  - **Unlike** t-tests, tests this for **two or more** groups (t-tests assume two groups)
- Hypotheses
  - $H_0 : \mu_1 = \mu_2 \ldots = \mu_G$
  - H_1 : Not $(\mu_1 = \mu_2 \ldots = \mu_G)$
  - (H_1 is a **different hypothesis** than $\mu_1 \neq \mu_2 \ldots \neq \mu_G$)

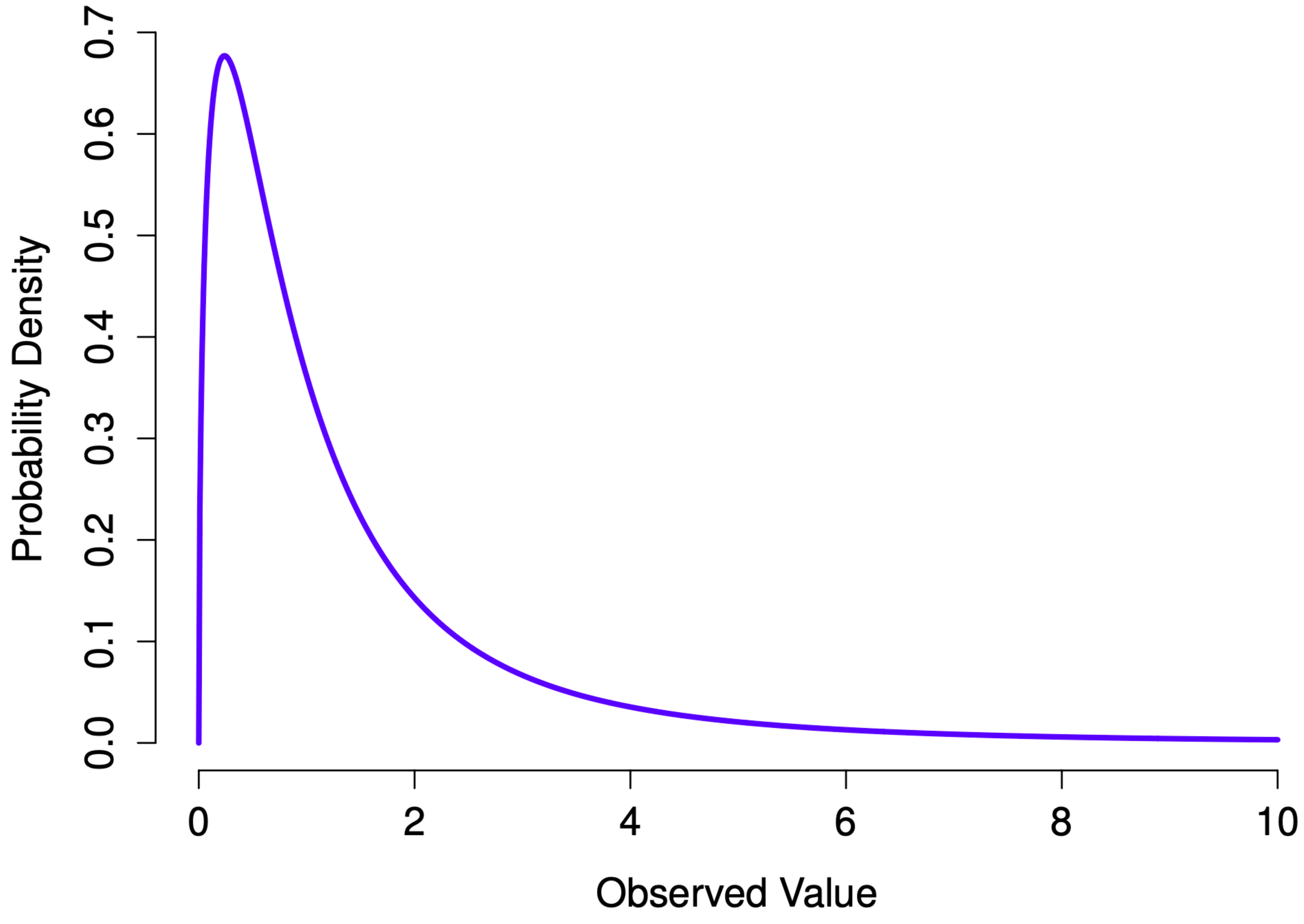


F-statistic


Probability Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
2
4
6
8
10
Observed Value

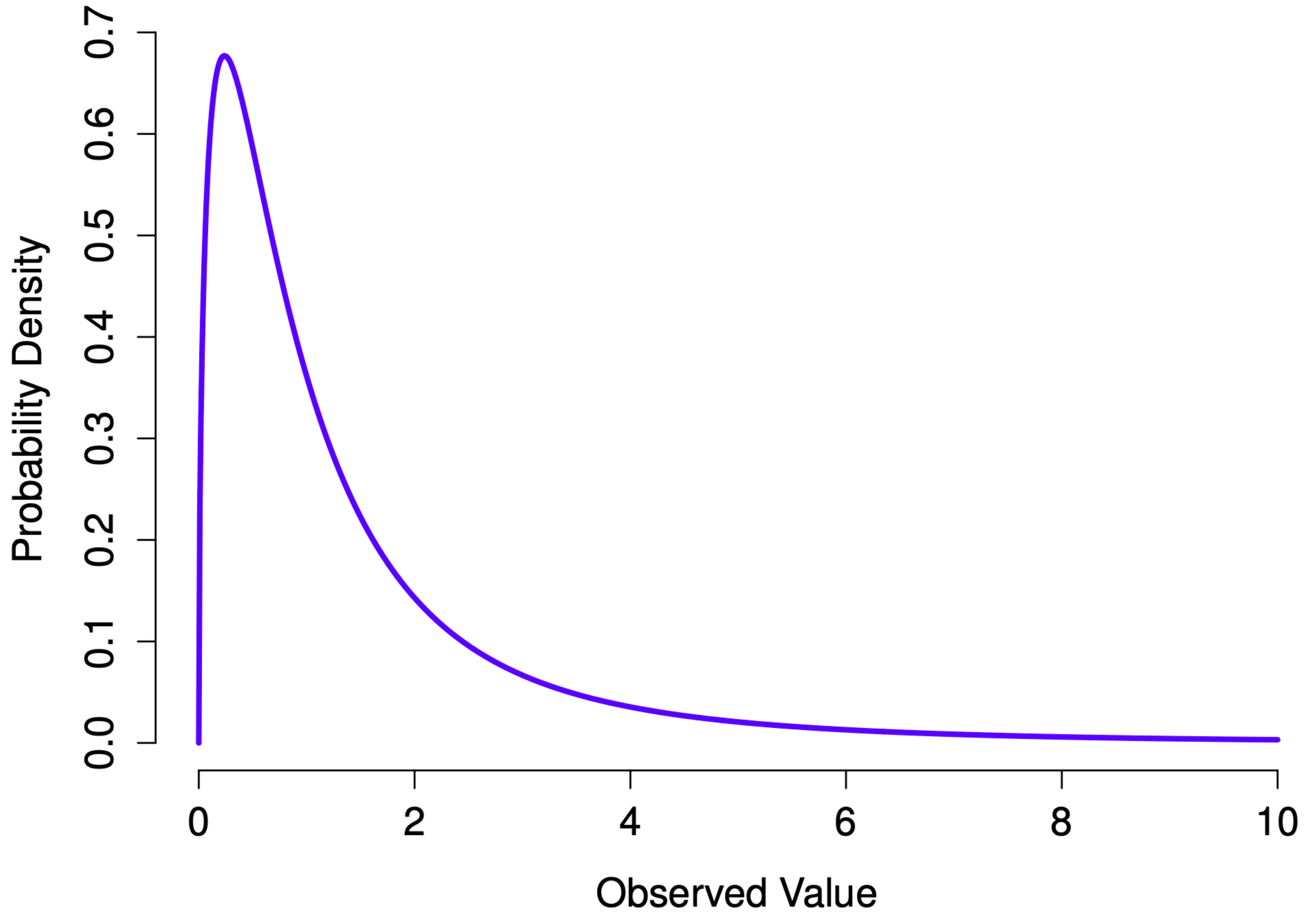
F-statistic

- We test the hypothesis with a **ratio** called the **F-statistic**
  - Compares the **between-group variation** (MS_b) with **within-group variation** (MS_w)
  - Computed as MS_b/MS_w
  - **F-distribution** gives us significance values
  - Higher F → lower p-value



F-statistic

- We test the hypothesis with a **ratio** called the **F-statistic**
  - Compares the **between-group variation** (MS_b) with **within-group variation** (MS_w)
  - Computed as MS_b / MS_w
  - **F-distribution** gives us significance values
  - Higher F → lower p-value
- Purposefully **glossing over** where MS_b and MS_w come from



Example in R

	drug	therapy	mood.gain
1	placebo	no.therapy	0.5
2	placebo	no.therapy	0.3
3	placebo	no.therapy	0.1
4	anxifree	no.therapy	0.6
5	anxifree	no.therapy	0.4
6	anxifree	no.therapy	0.2
7	joyzepam	no.therapy	1.4
8	joyzepam	no.therapy	1.7
9	joyzepam	no.therapy	1.3
10	placebo	CBT	0.6
11	placebo	CBT	0.9
12	placebo	CBT	0.3
13	anxifree	CBT	1.1
14	anxifree	CBT	0.8
15	anxifree	CBT	1.2
16	joyzepam	CBT	1.8
17	joyzepam	CBT	1.3
18	joyzepam	CBT	1.4

```
my.anova <- aov( mood.gain ~ drug, clin.trial )
```

Example in R

- Groups: **two drugs and a placebo** in a clinical trial
  - Encoded as categorical variable

	drug	therapy	mood.gain
1	placebo	no.therapy	0.5
2	placebo	no.therapy	0.3
3	placebo	no.therapy	0.1
4	anxifree	no.therapy	0.6
5	anxifree	no.therapy	0.4
6	anxifree	no.therapy	0.2
7	joyzepam	no.therapy	1.4
8	joyzepam	no.therapy	1.7
9	joyzepam	no.therapy	1.3
10	placebo	CBT	0.6
11	placebo	CBT	0.9
12	placebo	CBT	0.3
13	anxifree	CBT	1.1
14	anxifree	CBT	0.8
15	anxifree	CBT	1.2
16	joyzepam	CBT	1.8
17	joyzepam	CBT	1.3
18	joyzepam	CBT	1.4

```
my.anova <- aov( mood.gain ~ drug, clin.trial )
```

Example in R

- Groups: **two drugs and a placebo** in a clinical trial
  - Encoded as categorical variable
- Outcome measure: **mood improvement**

	drug	therapy	mood.gain
1	placebo	no.therapy	0.5
2	placebo	no.therapy	0.3
3	placebo	no.therapy	0.1
4	anxifree	no.therapy	0.6
5	anxifree	no.therapy	0.4
6	anxifree	no.therapy	0.2
7	joyzepam	no.therapy	1.4
8	joyzepam	no.therapy	1.7
9	joyzepam	no.therapy	1.3
10	placebo	CBT	0.6
11	placebo	CBT	0.9
12	placebo	CBT	0.3
13	anxifree	CBT	1.1
14	anxifree	CBT	0.8
15	anxifree	CBT	1.2
16	joyzepam	CBT	1.8
17	joyzepam	CBT	1.3
18	joyzepam	CBT	1.4

```
my.anova <- aov( mood.gain ~ drug, clin.trial )
```

Example in R

- Groups: **two drugs and a placebo** in a clinical trial
  - Encoded as categorical variable
- Outcome measure: **mood improvement**
- R command: **`aov()`**
  - `formula`: outcome variable distributed by group
  - Returns **complex R object**

	drug	therapy	mood.gain
1	placebo	no.therapy	0.5
2	placebo	no.therapy	0.3
3	placebo	no.therapy	0.1
4	anxifree	no.therapy	0.6
5	anxifree	no.therapy	0.4
6	anxifree	no.therapy	0.2
7	joyzepam	no.therapy	1.4
8	joyzepam	no.therapy	1.7
9	joyzepam	no.therapy	1.3
10	placebo	CBT	0.6
11	placebo	CBT	0.9
12	placebo	CBT	0.3
13	anxifree	CBT	1.1
14	anxifree	CBT	0.8
15	anxifree	CBT	1.2
16	joyzepam	CBT	1.8
17	joyzepam	CBT	1.3
18	joyzepam	CBT	1.4

```
my.anova <- aov( mood.gain ~ drug, clin.trial )
```

Example in R

```
> print(my_anova)
Call:
   aov(formula = mood_gain ~ drug, data = clin_trial)

Terms:
                    drug Residuals
Sum of Squares  3.453333  1.391667
Deg. of Freedom        2        15

Residual standard error: 0.3045944
Estimated effects may be unbalanced

> summary(my_anova)
            Df Sum Sq Mean Sq F value   Pr(>F)    
drug         2  3.453  1.7267   18.61 8.65e-05 ***
Residuals   15  1.392  0.0928                     
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example in R

- The result of `print` is **not very informative**
  - **Residuals** is the variability that is **not explained by drug**
  - (i.e. **within-group variability**)

```
> print(my_anova)
Call:
   aov(formula = mood_gain ~ drug, data = clin_trial)

Terms:
                    drug Residuals
Sum of Squares  3.453333  1.391667
Deg. of Freedom        2        15

Residual standard error: 0.3045944
Estimated effects may be unbalanced

> summary(my_anova)
            Df Sum Sq Mean Sq F value   Pr(>F)    
drug         2  3.453  1.7267   18.61 8.65e-05 ***
Residuals   15  1.392  0.0928                     
---
Signif. codes:  
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example in R

- The result of `print` is **not very informative**
  - **Residuals** is the variability that is **not explained by drug**
  - (i.e. **within-group variability**)
- Can get **p-values** with `summary(my_anova)`
  - F-statistic is here too

```
> print(my_anova)
Call:
   aov(formula = mood_gain ~ drug, data = clin_trial)

Terms:
                    drug Residuals
Sum of Squares  3.453333  1.391667
Deg. of Freedom        2        15

Residual standard error: 0.3045944
Estimated effects may be unbalanced

> summary(my_anova)
            Df Sum Sq Mean Sq F value   Pr(>F)
drug         2  3.453  1.7267   18.61 8.65e-05 ***
Residuals   15  1.392  0.0928
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the result

Interpreting the result

- If we get a significant result, what does that mean?
  - Null hypothesis: $\mu_1 = \mu_2 \ldots = \mu_G$
  - Alternative hypothesis: **anything else!**

Interpreting the result

- If we get a significant result, what does that mean?
  - Null hypothesis: $\mu_1 = \mu_2 \ldots = \mu_G$
  - Alternative hypothesis: **anything else!**
- Rejecting the Null **doesn't tell us which groups are distinct** from which other groups
  - (Only that they **aren't all the same!**)

Interpreting the result

- If we get a significant result, what does that mean?
  - Null hypothesis: $\mu_1 = \mu_2 ... = \mu_G$
  - Alternative hypothesis: **anything else!**
- Rejecting the Null **doesn't tell us which groups are distinct** from which other groups
  - (Only that they **aren't all the same!**)
- Null can be rephrased as "drugs don't have an affect on mood"

Interpreting the result

- If we get a significant result, what does that mean?
  - Null hypothesis: $\mu_1 = \mu_2 \ldots = \mu_G$
  - Alternative hypothesis: **anything else!**
- Rejecting the Null **doesn't tell us which groups are distinct** from which other groups
  - (Only that they **aren't all the same!**)
- Null can be rephrased as "drugs don't have an affect on mood"
- What if we want to be **more specific?**

Multiple comparisons

possibility:	is $\mu_P = \mu_A$?	is $\mu_P = \mu_J$?	is $\mu_A = \mu_J$?	which hypothesis?
1	✓	✓	✓	null
2	✓	✓		alternative
3	✓		✓	alternative
4	✓			alternative
5		✓	✓	alternative
6		✓		alternative
7			✓	alternative
8				alternative

Multiple comparisons

- 3 groups gives us **8 possible hypotheses!**
  - ANOVA **combines** 7 of them

possibility:	is $\mu_P = \mu_A$?	is $\mu_P = \mu_J$?	is $\mu_A = \mu_J$?	which hypothesis?
1	✓	✓	✓	null
2	✓	✓		alternative
3	✓		✓	alternative
4	✓			alternative
5		✓	✓	alternative
6		✓		alternative
7			✓	alternative
8				alternative

Multiple comparisons

- 3 groups gives us **8 possible hypotheses!**
  - ANOVA **combines** 7 of them
- Can test all with **pair-wise t-tests**

```
> pairwise.t.test( x = clin.trial$mood.gain,    # outcome variable
+                  g = clin.trial$drug,         # grouping variable
+                  p.adjust.method = "none"     # which correction to use?
+ )

        Pairwise comparisons using t tests with pooled SD

data:  clin.trial$mood.gain and clin.trial$drug

         placebo anxifree
anxifree 0.15021 -
joyzepam 3e-05   0.00056
```

Problems with pair-wise tests

Problems with pair-wise tests

- The issue with pair-wise tests: **p-value "fishing"**
  - We do a **large number** of pair-wise comparisons, so a few are likely to be **significant by random chance!**
  - This is a type of **post-hockery peril** that Rusico talks about in hw4

Problems with pair-wise tests

- The issue with pair-wise tests: **p-value "fishing"**
  - We do a **large number** of pair-wise comparisons, so a few are likely to be **significant by random chance!**
  - This is a type of **post-hockery peril** that Rusico talks about in hw4
- There are methods to **correct for this problem**
  - Idea: p-values adjusted as if it were **one big test** rather than multiple
  - The book recommends **Holm Correction** (can be applied in R)

Problems with pair-wise tests

- The issue with pair-wise tests: **p-value "fishing"**
  - We do a **large number** of pair-wise comparisons, so a few are likely to be **significant by random chance!**
  - This is a type of **post-hockery peril** that Rusico talks about in hw4
- There are methods to **correct for this problem**
  - Idea: p-values adjusted as if it were **one big test** rather than multiple
  - The book recommends **Holm Correction** (can be applied in R)
- Why not just run corrected pair-wise tests rather than ANOVA?
  - ANOVA encourages **well-formulated hypotheses**

Corrected pair-wise t-tests

```
> pairwise.t.test(
+     x = clin_trial$mood_gain,
+     g = clin_trial$drug,
+     p.adjust.method = "holm"
+ )

        Pairwise comparisons using t tests with pooled SD

data:  clin_trial$mood_gain and clin_trial$drug

          anxifree joyzepram
joyzepram 0.0011   -
placebo   0.1502   9.1e-05

P value adjustment method: holm
```

ANOVA assumptions

ANOVA assumptions

- Data points are **independent** of one another
  - e.g. the same person isn't in more than one group

ANOVA assumptions

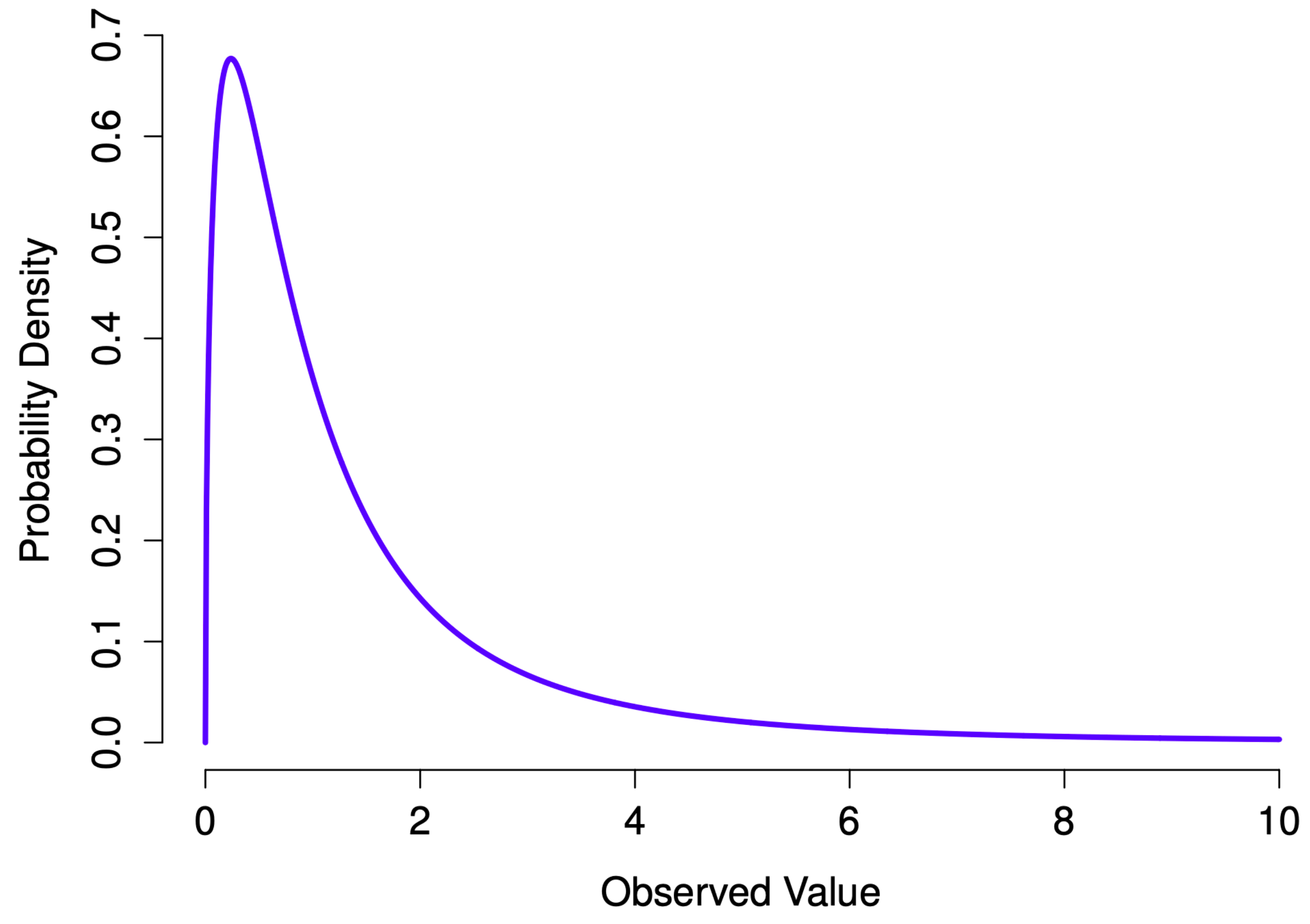
- Data points are **independent** of one another
  - e.g. the same person isn't in more than one group
- All groups have the **same variance** ("homogeneity of variance")
  - Like t-tests, there's a version of ANOVA that **removes this assumption**
  - Called the **Welch One-way Test**
  - R command is named unhelpfully: `oneway.test()`

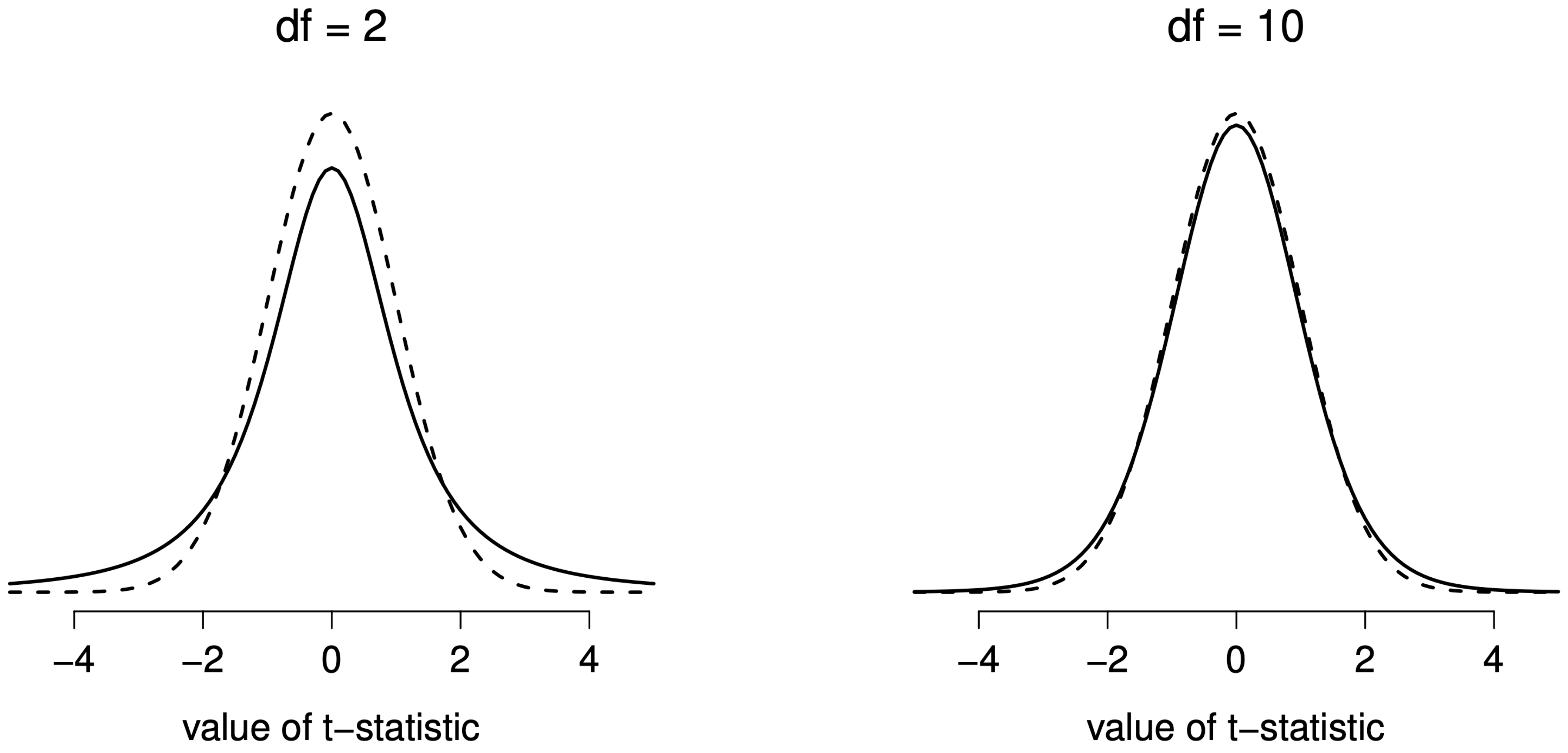
ANOVA assumptions

- Data points are **independent** of one another
  - e.g. the same person isn't in more than one group
- All groups have the **same variance** ("homogeneity of variance")
  - Like t-tests, there's a version of ANOVA that **removes this assumption**
  - Called the **Welch One-way Test**
  - R command is named unhelpfully: `oneway.test()`
- Groups are **Normally distributed**
  - This is a **strong assumption**, but there are **methods to check it**
  - (We won't bother checking Normality in this course. **This is a <u>simplification</u>**)

ANOVA and t-test

- With **two groups** the ANOVA is **mathematically equivalent** to the **Student's t-test**
  - Student's t-test can be considered a **simplified case** of ANOVA
  - Of mathematical interest: equivalence between test based on F-statistic and t-statistic





Linear regression

Linear regression

$$y = \beta x + \alpha$$

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)

$$y = \beta x + \alpha$$

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)
- Independent variables are also called **predictors**

$$y = \beta x + \alpha$$

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)
- Independent variables are also called **predictors**
- Dependents are also called **responses**

$$y = \beta x + \alpha$$

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)
- Independent variables are also called **predictors**
- Dependents are also called **responses**
- "Linear" refers to the fact that effects of predictors are **summed together**

$$y = \beta x + \alpha$$

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)
- Independent variables are also called **predictors**
- Dependents are also called **responses**
- “Linear” refers to the fact that effects of predictors are **summed together**

$$\underbrace{y}_{\text{response}} = \beta \underbrace{x}_{\text{predictor}} + \alpha$$

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)
- Independent variables are also called **predictors**
- Dependents are also called **responses**
- “Linear” refers to the fact that effects of predictors are **summed together**

learned
coefficients

$$y = \beta x + \alpha$$

response

predictor

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)
- Independent variables are also called **predictors**
- Dependents are also called **responses**
- “Linear” refers to the fact that effects of predictors are **summed together**

slope

intercept / bias

$$y = \beta x + \alpha$$

response

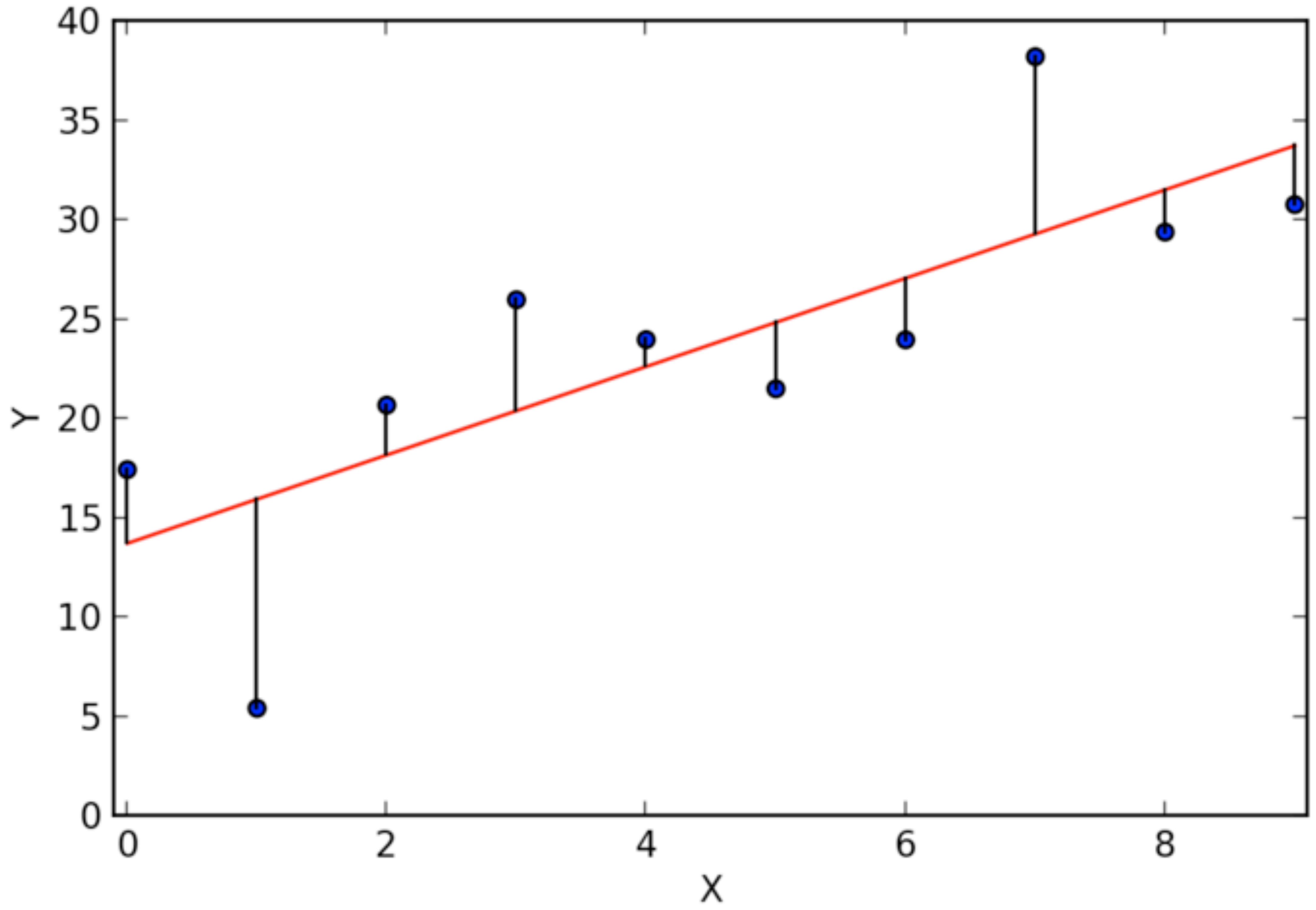
predictor

Linear regression

- Model the change in an observed **dependent variable** (y) as a function of one or more **independent variables** (x)
- Independent variables are also called **predictors**
- Dependents are also called **responses**
- “Linear” refers to the fact that effects of predictors are **summed together**

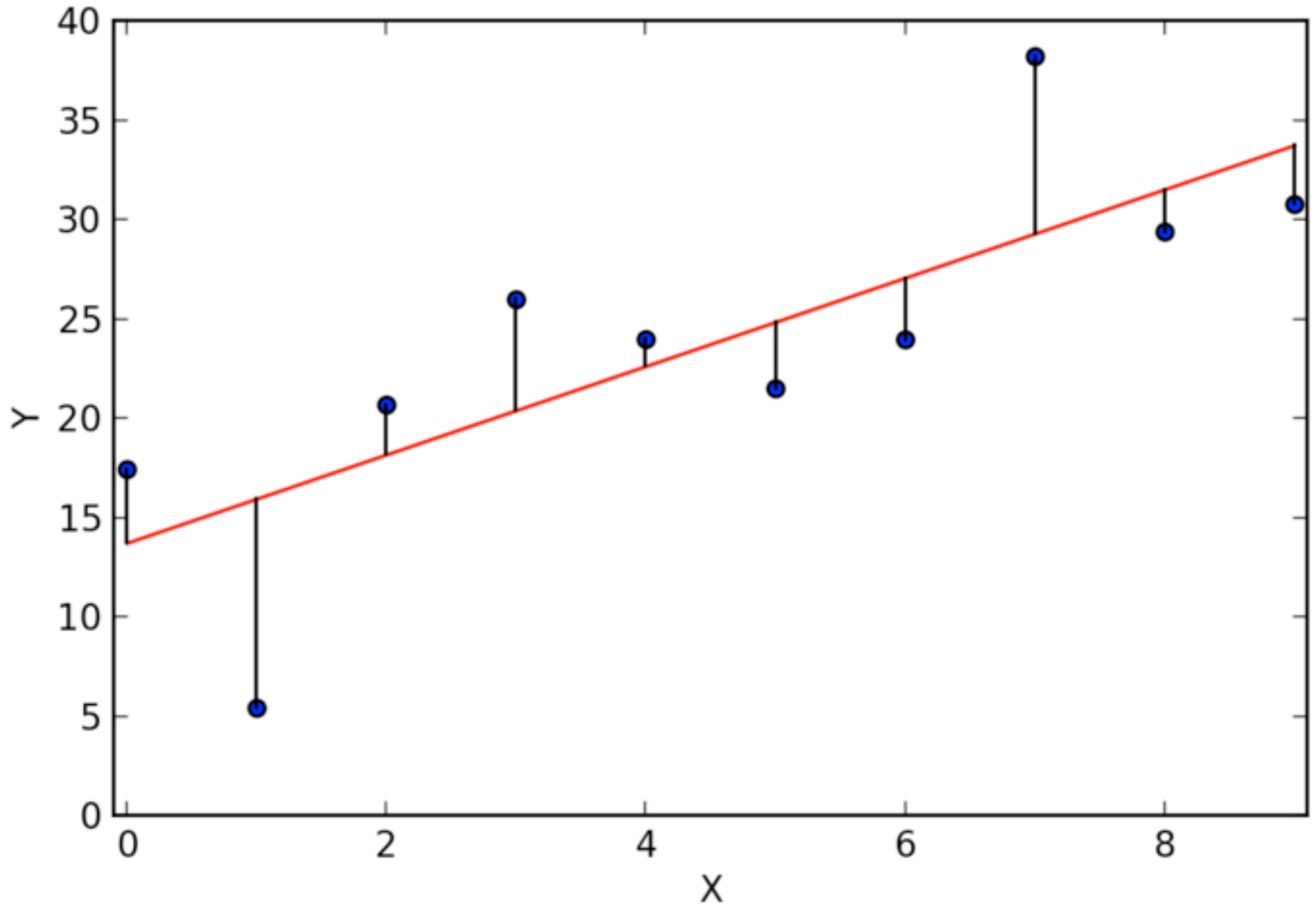
$$\underbrace{y}_{\text{response}} = \overbrace{\beta}^{\text{slope}}\underbrace{x}_{\text{predictor}} + \overbrace{\alpha}^{\text{intercept / bias}} \qquad (y = mx + b)$$

Error minimization



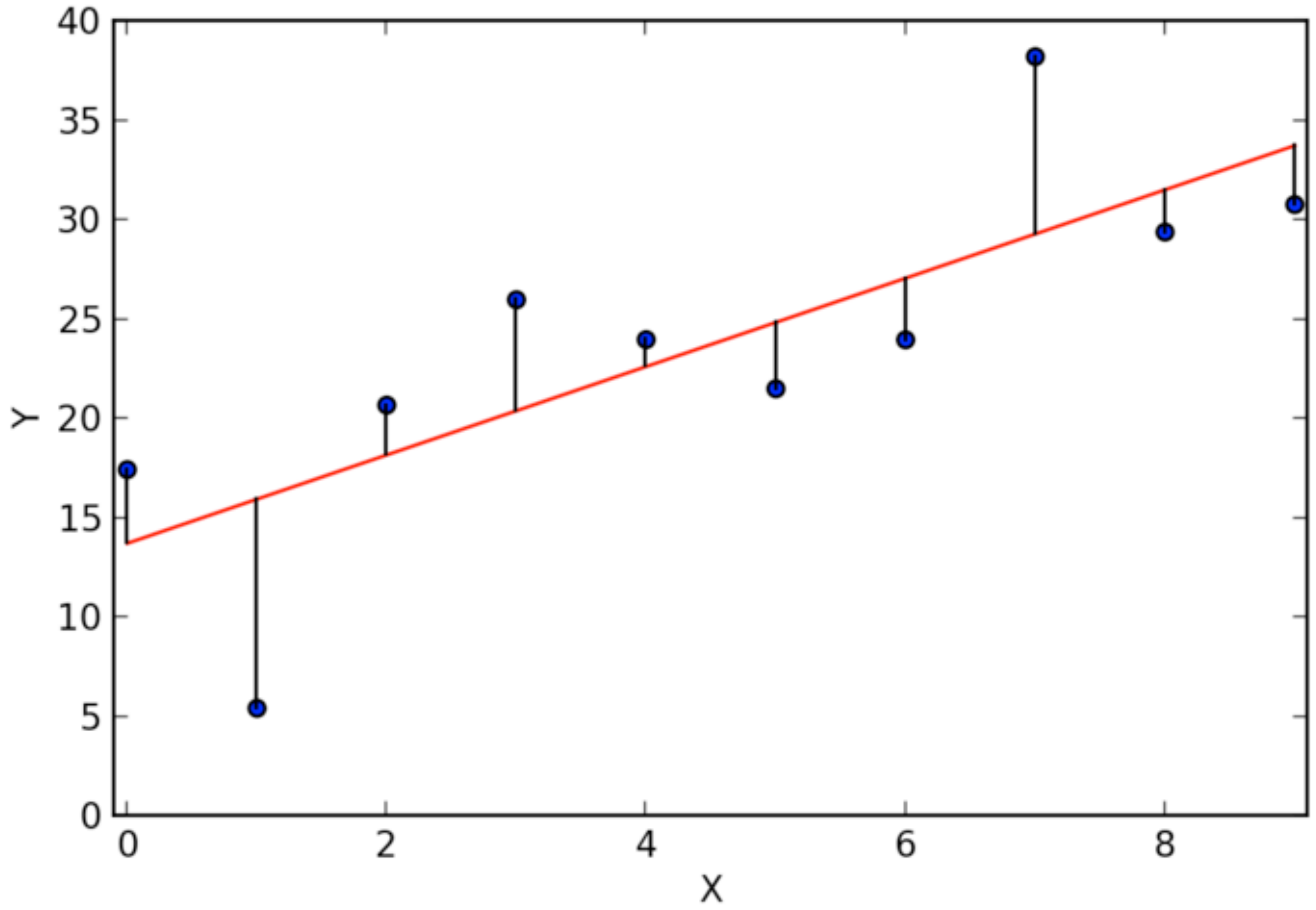
Error minimization

- Model coefficients are selected to **minimize the error** between the predicted line and observed datapoints



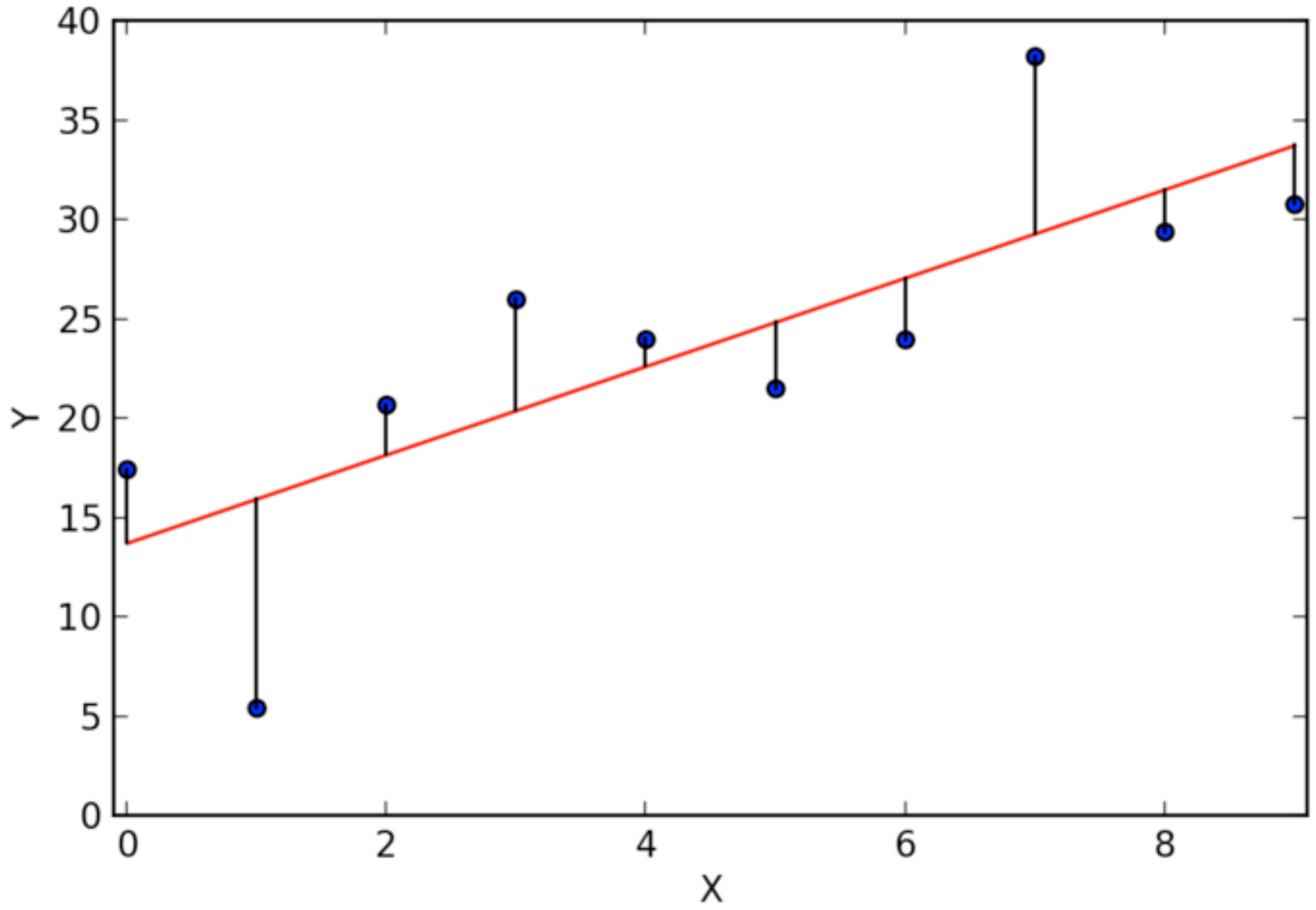
Error minimization

- Model coefficients are selected to **minimize the error** between the predicted line and observed datapoints
- This is called the **residual error**



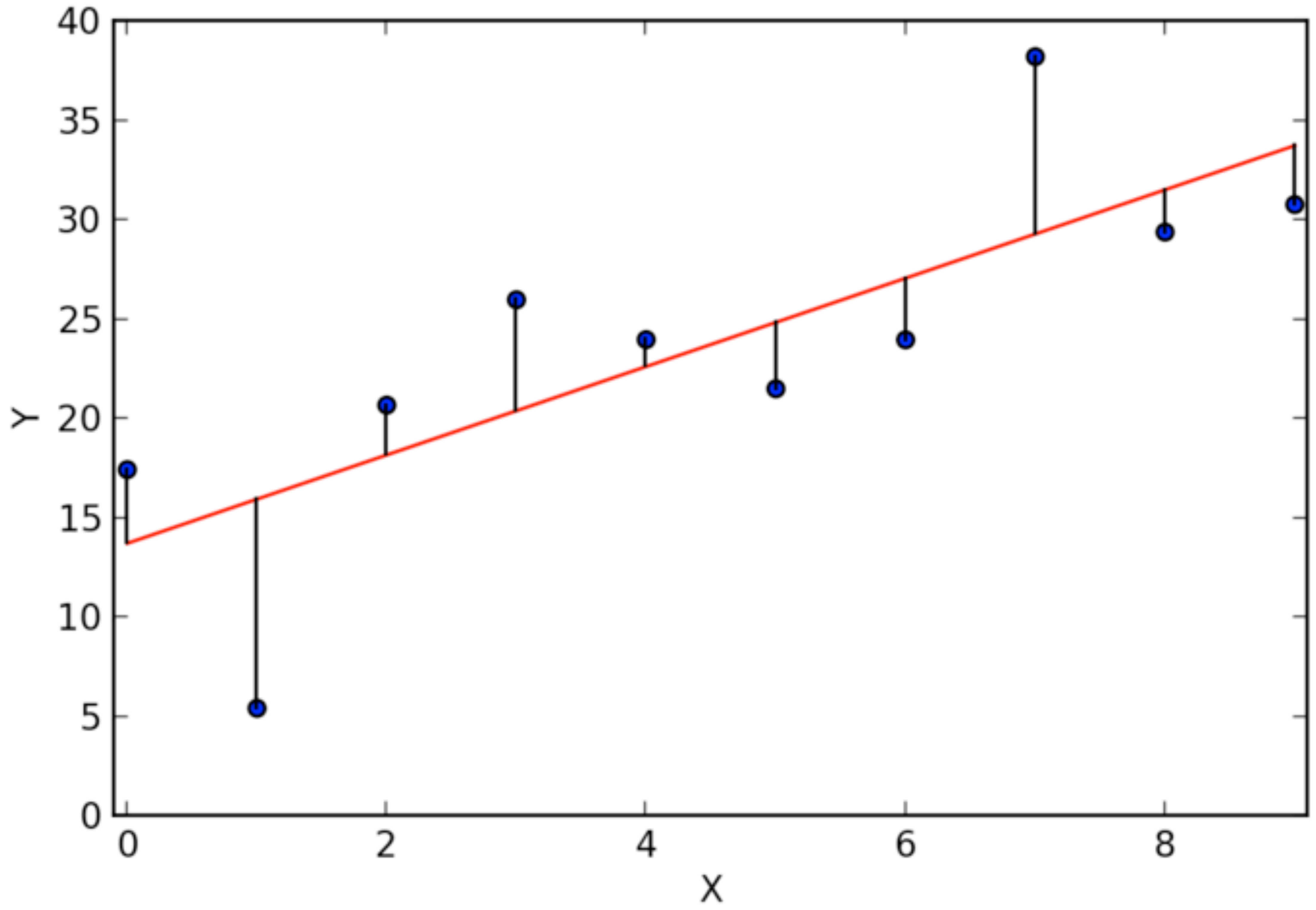
Error minimization

- Model coefficients are selected to **minimize the error** between the predicted line and observed datapoints
- This is called the **residual error**
- Yields the **“best fit line”**



Error minimization

- Model coefficients are selected to **minimize the error** between the predicted line and observed datapoints
- This is called the **residual error**
- Yields the "**best fit line**"
- Sometimes explicitly modeled:
  - $y = \beta x + \alpha + \epsilon$
  - where ϵ is residual error



Multivariable regression

Multivariable regression

- Formula straightforwardly generalizes to **multiple predictors**
  - $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon$
  - Matrix notation: $Y = X\beta + \epsilon$
  - Can be **solved in the same way**

Multivariable regression

- Formula straightforwardly generalizes to **multiple predictors**
  - $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon$
  - Matrix notation: $Y = X\beta + \epsilon$
  - Can be **solved in the same way**
- R code
  - `model = lm(y ~ variable)`
  - `model = lm(y ~ variable_1 + variable_2)`

Categorical variables

Categorical variables

- A predictor x can be **categorical**, also known as a **factor**
  - e.g. which pond is a fish sampled from out of {pond_1, pond_2}

Categorical variables

- A predictor x can be **categorical**, also known as a **factor**
  - e.g. which pond is a fish sampled from out of {pond_1, pond_2}
- Regression software usually converts this to a **binary “dummy” variable**
  - Pond 1 : $x = 0$
  - Pond 2: $x = 1$

Categorical variables

- A predictor x can be **categorical**, also known as a **factor**
  - e.g. which pond is a fish sampled from out of {pond_1, pond_2}
- Regression software usually converts this to a **binary "dummy" variable**
  - Pond 1 : $x = 0$
  - Pond 2: $x = 1$
- What is the result of this "dummy" encoding?

Categorical variables

Categorical variables

- **Pond 1** case
  - $y = \beta * x + \alpha = \beta * \mathbf{0} + \alpha = \alpha$

Categorical variables

- **Pond 1** case
  - $y = \beta * x + \alpha = \beta * \mathbf{0} + \alpha = \alpha$
- **Pond 2** case
  - $y = \beta * x + \alpha = \beta * \mathbf{1} + \alpha = \beta + \alpha$

Categorical variables

- **Pond 1** case
  - $y = \beta * x + \alpha = \beta * \mathbf{0} + \alpha = \alpha$
- **Pond 2** case
  - $y = \beta * x + \alpha = \beta * \mathbf{1} + \alpha = \beta + \alpha$
- With the dummy encoding, α is the **mean of pond 1**
  - Implicitly assumes **pond 1** is the **“baseline/control” group**
  - β is the **difference** between pond 1 and pond 2

Categorical variables

Categorical variables

- **Pond 1** case
  - $y = \beta_2 * 0 + \beta_3 * 0 + \alpha = \alpha$

Categorical variables

- **Pond 1** case
  - $y = \beta_2 * 0 + \beta_3 * 0 + \alpha = \alpha$
- **Pond 2** case
  - $y = \beta_2 * \mathbf{1} + \beta_3 * 0 + \alpha = \alpha + \beta_2$

Categorical variables

- **Pond 1** case
  - $y = \beta_2 * 0 + \beta_3 * 0 + \alpha = \alpha$
- **Pond 2** case
  - $y = \beta_2 * \mathbf{1} + \beta_3 * 0 + \alpha = \alpha + \beta_2$
- **Pond 3** case
  - $y = \beta_2 * 0 + \beta_3 * \mathbf{1} + \alpha = \alpha + \beta_3$

Categorical variables

- **Pond 1** case
  - $y = \beta_2 * 0 + \beta_3 * 0 + \alpha = \alpha$
- **Pond 2** case
  - $y = \beta_2 * \mathbf{1} + \beta_3 * 0 + \alpha = \alpha + \beta_2$
- **Pond 3** case
  - $y = \beta_2 * 0 + \beta_3 * \mathbf{1} + \alpha = \alpha + \beta_3$
- ($n - 1$ dummy variables used to represent n values)

R example

```
> pond_data
   fertilizer pond     algae
1   7.6990327    1 13.910031
2   6.2655237    1 11.260629
3   3.0759136    1  5.283725
4   4.2035496    1  7.677377
5   9.5885167    1 19.556104
6   3.2948473    1  2.923524
7   1.0155064    1  1.319341
8   1.5696621    1  3.223186
9   3.7088986    2  7.985676
10  2.5777049    2 11.427004
11  1.4913182    2  6.370520
12  5.8812233    2 16.597473
13  7.6915106    2 17.517383
14  9.6907026    2 25.386514
15  9.6002026    2 24.594571
16  0.4020754    2  4.575538
>
> model = lm(algae ~ fertilizer + pond, data=pond_data)
```

R example

- We have **two predictors:**
  - fertilizer (continuous)
  - pond (categorical)

```
> pond_data
   fertilizer pond     algae
1   7.6990327    1 13.910031
2   6.2655237    1 11.260629
3   3.0759136    1  5.283725
4   4.2035496    1  7.677377
5   9.5885167    1 19.556104
6   3.2948473    1  2.923524
7   1.0155064    1  1.319341
8   1.5696621    1  3.223186
9   3.7088986    2  7.985676
10  2.5777049    2 11.427004
11  1.4913182    2  6.370520
12  5.8812233    2 16.597473
13  7.6915106    2 17.517383
14  9.6907026    2 25.386514
15  9.6002026    2 24.594571
16  0.4020754    2  4.575538
>
> model = lm(algae ~ fertilizer + pond, data=pond_data)
```

R example

- We have **two predictors:**
  - fertilizer (continuous)
  - pond (categorical)
- **One outcome**: algae (continuous)

```
> pond_data
   fertilizer pond     algae
1   7.6990327    1 13.910031
2   6.2655237    1 11.260629
3   3.0759136    1  5.283725
4   4.2035496    1  7.677377
5   9.5885167    1 19.556104
6   3.2948473    1  2.923524
7   1.0155064    1  1.319341
8   1.5696621    1  3.223186
9   3.7088986    2  7.985676
10  2.5777049    2 11.427004
11  1.4913182    2  6.370520
12  5.8812233    2 16.597473
13  7.6915106    2 17.517383
14  9.6907026    2 25.386514
15  9.6002026    2 24.594571
16  0.4020754    2  4.575538
>
> model = lm(algae ~ fertilizer + pond, data=pond_data)
```

R example

- We have **two predictors:**
  - fertilizer (continuous)
  - pond (categorical)
- **One outcome**: algae (continuous)
- Categorical variable gets encoded with **dummy variables**

```
> pond_data
   fertilizer pond     algae
1   7.6990327    1 13.910031
2   6.2655237    1 11.260629
3   3.0759136    1  5.283725
4   4.2035496    1  7.677377
5   9.5885167    1 19.556104
6   3.2948473    1  2.923524
7   1.0155064    1  1.319341
8   1.5696621    1  3.223186
9   3.7088986    2  7.985676
10  2.5777049    2 11.427004
11  1.4913182    2  6.370520
12  5.8812233    2 16.597473
13  7.6915106    2 17.517383
14  9.6907026    2 25.386514
15  9.6002026    2 24.594571
16  0.4020754    2  4.575538
>
> model = lm(algae ~ fertilizer + pond, data=pond_data)
```

R example

- We have **two predictors:**
  - fertilizer (continuous)
  - pond (categorical)
- **One outcome**: algae (continuous)
- Categorical variable gets encoded with **dummy variables**
- algae $= \beta_f \cdot$ fertilizer $+ \beta_p \cdot$ pond2 $+ \alpha$
  - α is the "default" intercept

```
> pond_data
   fertilizer pond     algae
1   7.6990327    1 13.910031
2   6.2655237    1 11.260629
3   3.0759136    1  5.283725
4   4.2035496    1  7.677377
5   9.5885167    1 19.556104
6   3.2948473    1  2.923524
7   1.0155064    1  1.319341
8   1.5696621    1  3.223186
9   3.7088986    2  7.985676
10  2.5777049    2 11.427004
11  1.4913182    2  6.370520
12  5.8812233    2 16.597473
13  7.6915106    2 17.517383
14  9.6907026    2 25.386514
15  9.6002026    2 24.594571
16  0.4020754    2  4.575538
>
> model = lm(algae ~ fertilizer + pond, data=pond_data)
```

R example

- We have **two predictors:**
  - fertilizer (continuous)
  - pond (categorical)
- **One outcome**: algae (continuous)
- Categorical variable gets encoded with **dummy variables**
- algae $= \beta_f \cdot$ fertilizer $+ \beta_p \cdot$ pond2 $+ \alpha$
  - α is the "default" intercept
- **Null hypothesis** is that α and β are zero

```
> pond_data
   fertilizer pond     algae
1   7.6990327    1 13.910031
2   6.2655237    1 11.260629
3   3.0759136    1  5.283725
4   4.2035496    1  7.677377
5   9.5885167    1 19.556104
6   3.2948473    1  2.923524
7   1.0155064    1  1.319341
8   1.5696621    1  3.223186
9   3.7088986    2  7.985676
10  2.5777049    2 11.427004
11  1.4913182    2  6.370520
12  5.8812233    2 16.597473
13  7.6915106    2 17.517383
14  9.6907026    2 25.386514
15  9.6002026    2 24.594571
16  0.4020754    2  4.575538
>
> model = lm(algae ~ fertilizer + pond, data=pond_data)
```

R example

```
> summary(model)

Call:
lm(formula = algae ~ fertilizer + pond, data = pond_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.62339 -0.74624 -0.06466  0.78589  1.52331 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.12995    0.50861   0.256    0.802    
fertilizer   2.04681    0.07992  25.610 1.65e-12 ***
pond2        3.37159    0.50021   6.740 1.38e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9967 on 13 degrees of freedom
Multiple R-squared:  0.9827,	Adjusted R-squared:   0.98 
F-statistic: 368.3 on 2 and 13 DF,  p-value: 3.58e-12
```

R example

```
> summary(model)

Call:
lm(formula = algae ~ fertilizer + pond, data = pond_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.62339 -0.74624 -0.06466  0.78589  1.52331 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.12995    0.50861   0.256    0.802    
fertilizer   2.04681    0.07992  25.610 1.65e-12 ***
pond2        3.37159    0.50021   6.740 1.38e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9967 on 13 degrees of freedom
Multiple R-squared:  0.9827,	Adjusted R-squared:   0.98 
F-statistic: 368.3 on 2 and 13 DF,  p-value: 3.58e-12
```

Just means the intercept is **not significantly different from zero**

R example

```
> summary(model)

Call:
lm(formula = algae ~ fertilizer + pond, data = pond_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.62339 -0.74624 -0.06466  0.78589  1.52331 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.12995    0.50861   0.256    0.802    
fertilizer   2.04681    0.07992  25.610 1.65e-12 ***
pond2        3.37159    0.50021   6.740 1.38e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9967 on 13 degrees of freedom
Multiple R-squared:  0.9827,	Adjusted R-squared:   0.98 
F-statistic: 368.3 on 2 and 13 DF,  p-value: 3.58e-12
```

Just means the intercept is **not significantly different from zero**

$$\text{algae} = 2.05 \cdot \text{fertilizer} + 3.37 \cdot \text{pond2} + 0.13$$