Vietnam National University, Ho Chi Minh City

University of Science

Faculty of Information Techonology

Introduction to Machine Learning

Model Evaluation

Duc Nguyen

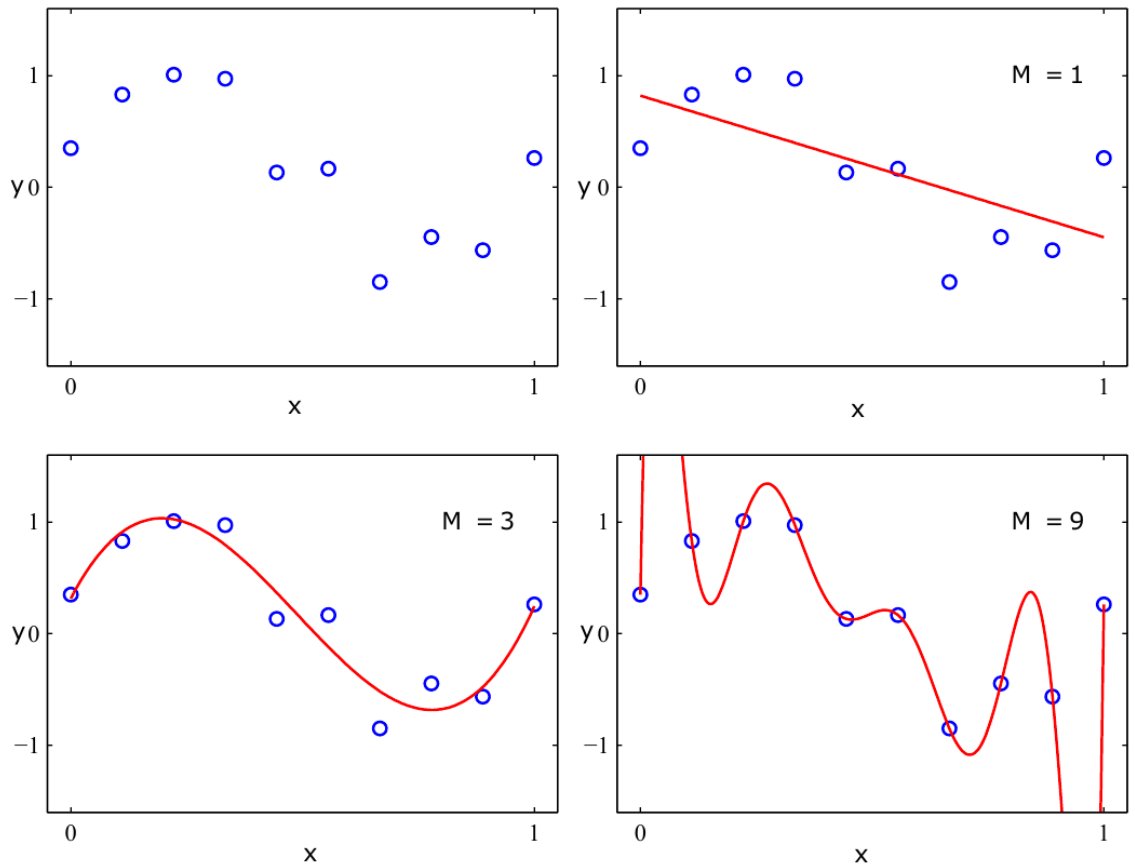
June 1, 2023

Contents

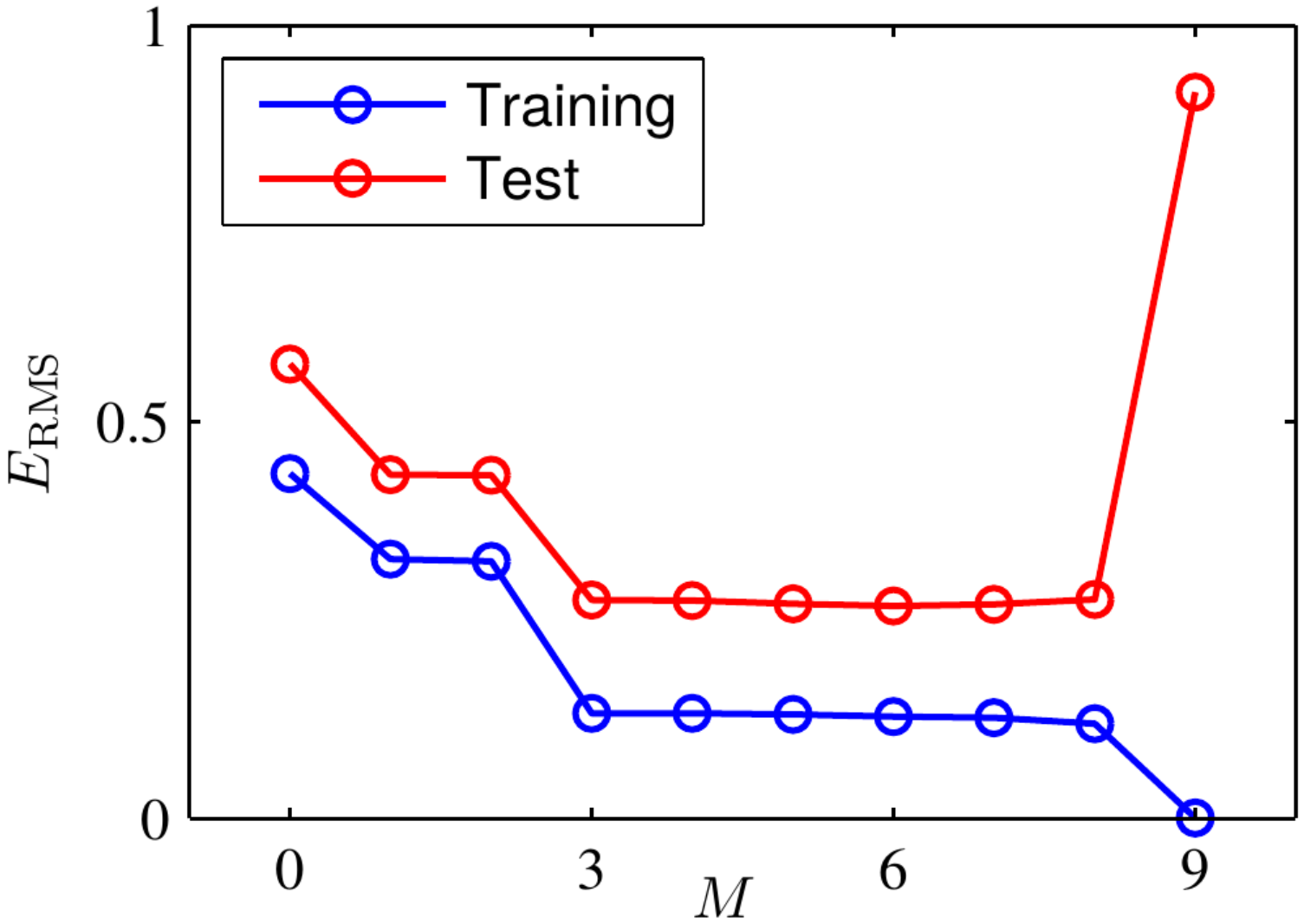
Capacity, Overfitting and Underfitting

Model Training

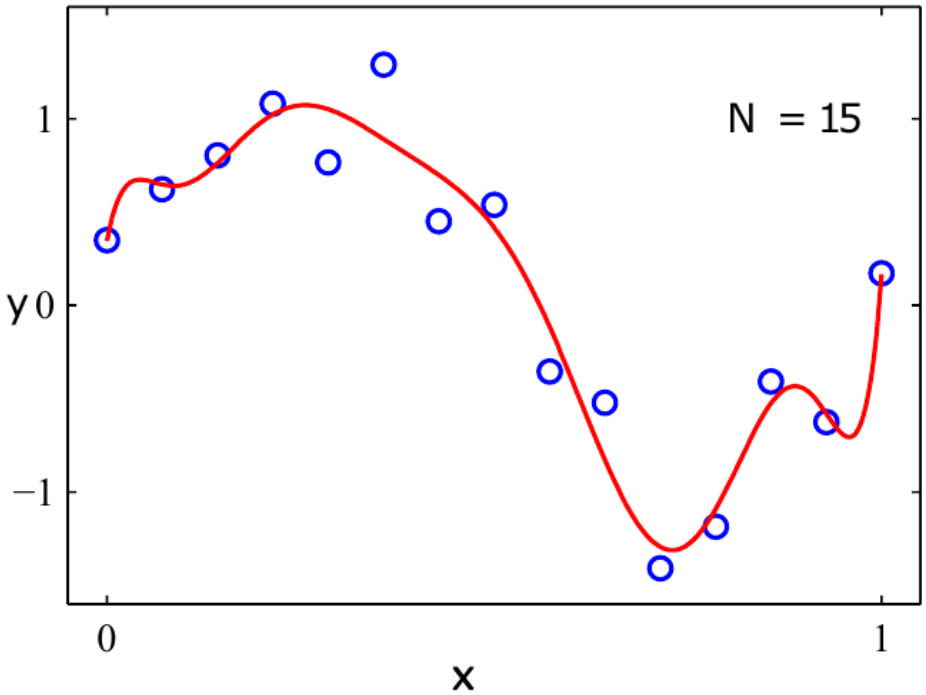
- Considering three hypothesis sets (polynomial functions) $\mathcal{H}_1, \mathcal{H}_3$ and $\mathcal{H}_9$ and the results of fitting the models to the data set $\mathcal{D}$

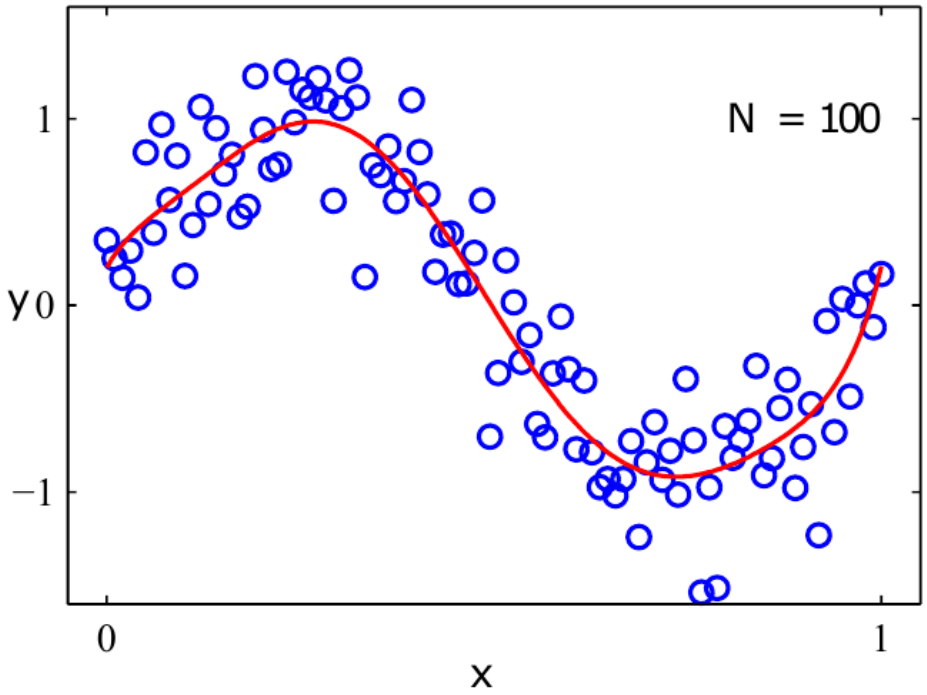


Model Performance



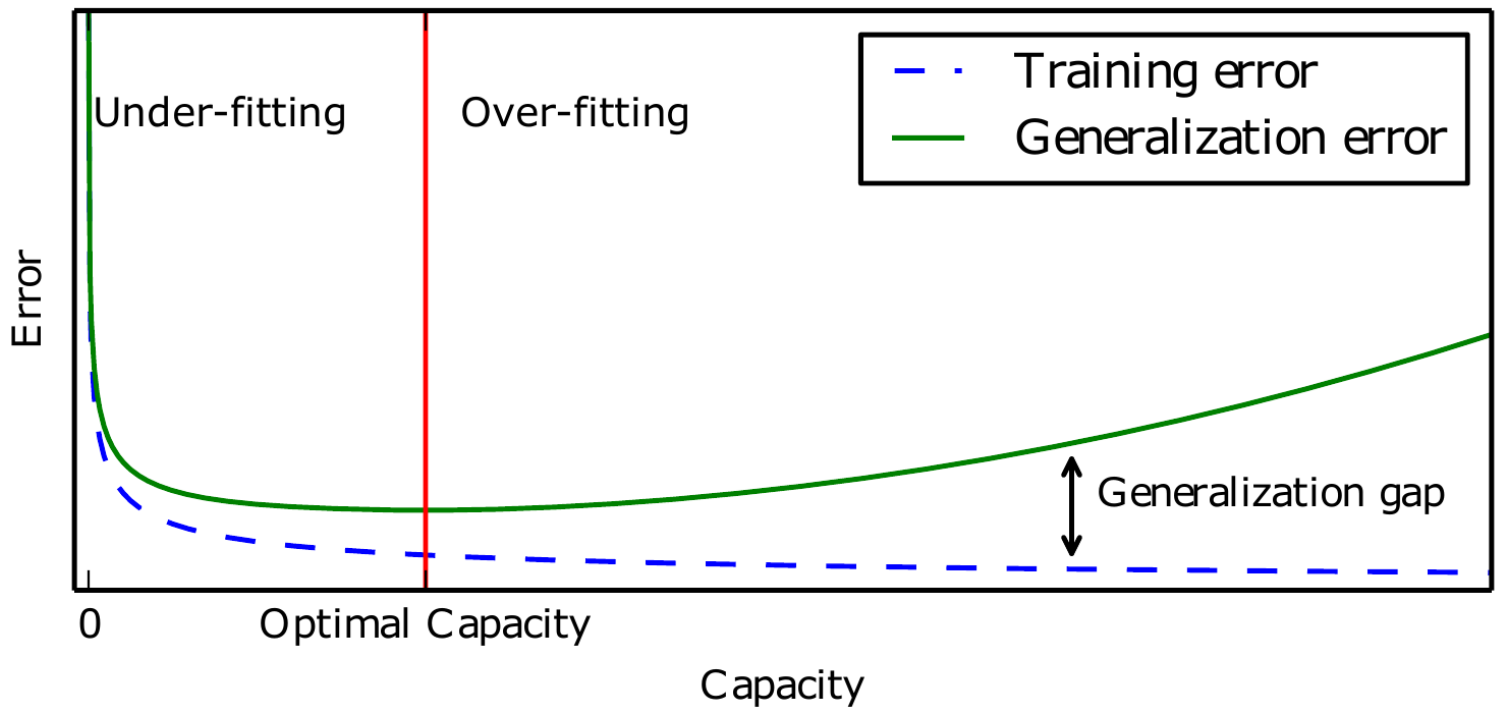
Increasing N





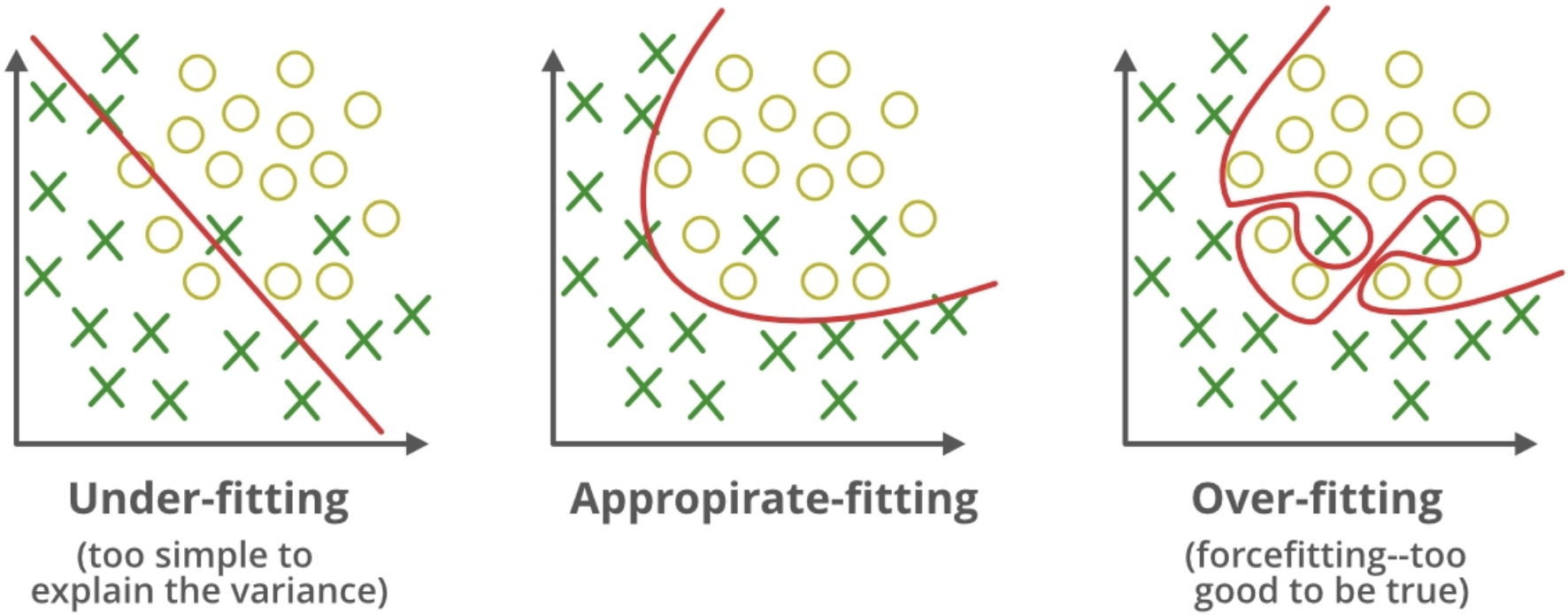
Generation and Capacity

- The criteria determining how well a machine learning model will perform:
  1. Make the training error small
  2. Make the gap between the training and test error small



Regularization I

- Regularization helps control model capacity
- However too much regularization can run the risk of underfitting



Regularization II

- There are 3 common types of regularization
  1. L2 (or weight decay)

$$R(W) = \sum_i \sum_j W_{i,j}^2$$

  2. L1

$$R(W) = \sum_i \sum_j |W_{i,j}|$$

  3. Elastic Net

$$R(W) = \sum_i \sum_j \beta W_{i,j}^2 + |W_{i,j}|$$

Classification Analysis

Confusion matrix

	Predicted: Positive	Predicted: Negative
Actual: Positive	TP	FN
Actual: Negative	FP	TN

- Accuracy:

$$\frac{TP + TN}{TP + TN + FN + FP}$$

- Precision

$$\frac{TP}{TP + FP}$$

- Recall:

$$\frac{TP}{TP + FN}$$

ROC Curve

Definition (ROC curve)

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds

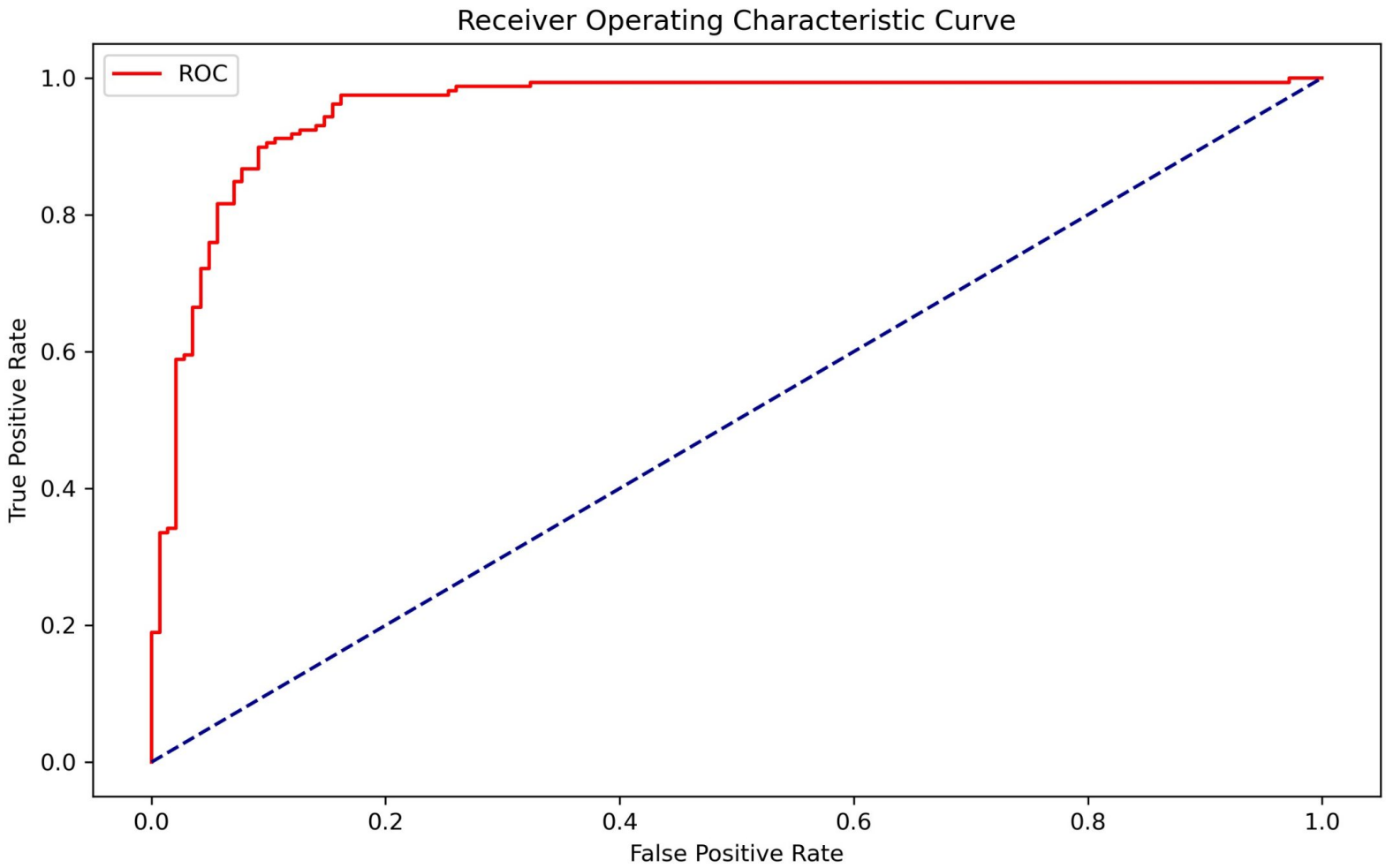
- This curve plots two parameters
  1. **True Positive Rate** also is called Recall:

$$TPR = \frac{TP}{TP + FN}$$

  2. **False Positive Rate** is defined as follow:

$$FPR = \frac{FP}{FP + TN}$$

Example


Receiver Operating Characteristic Curve
ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

Hypothesis Testing

"The tendency of modern scientific teaching is to neglect the great books, to lay far too much stress upon relatively unimportant modern work, and to present masses of detail of doubtful truth and questionable weight in such a way as to obscure principles."

– Ronald Fisher

On the shoulder of . . . Ronald A. Fisher

- R.A.Fisher (1890 -1962) was the statistician most responsible for the statistical methods used to analyze data today
- *How do you conduct scientific inquiry, whether it be developing methods of experimental design, finding the best way to estimate a parameter, or answering specific questions such as which fertilizer works best?*

Hypothesis

Hypothesis

In statistics, a **hypothesis** is a statement about a population, usually claiming that a population **parameter** takes a particular numerical value or falls in a certain range of values.

Example

Using a person's horoscope, the probability p that an astrologer can correctly predict which of three personality charts applies to that person equals $\frac{1}{3}$. In other words, astrologers' predictions correspond to random guessing.

Significance Test

Significance Test I

- **Significance test:** is a method for using data to summarize the evidence about a hypothesis.
- Before conducting a significance test, we identify
  1. The variable measured
  2. **Population** parameter of interest

Significance Test II

The Steps of a Significance Test

1. Assumptions
   - Each significance test makes certain assumptions or has certain conditions under which it applies
   - Foremost, a test assumes that the data production used randomization
   - Other assumptions may be about the sample size or about the shape of the population distribution

Significance Test III

2 Hypothesis:

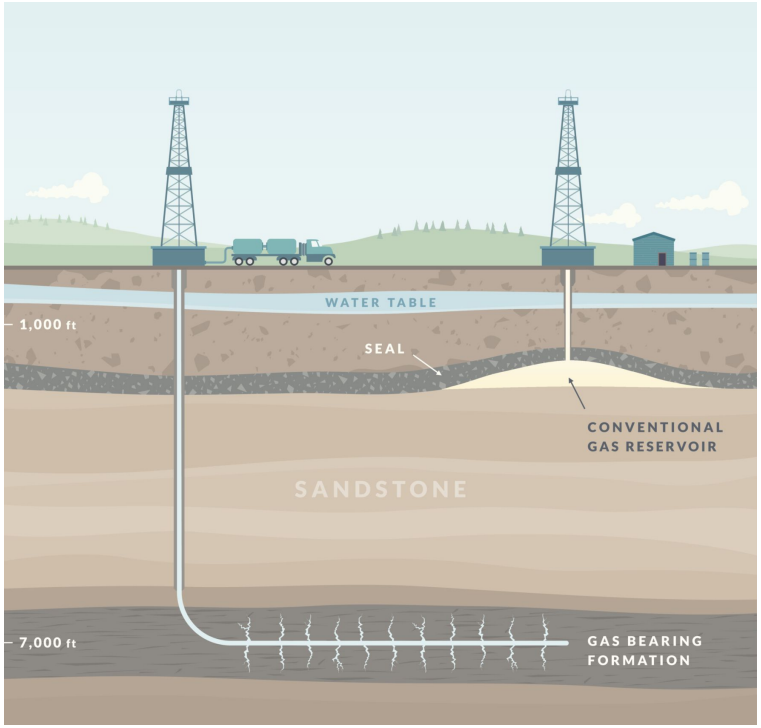
Null Hypothesis, Alternative Hypothesis

- The **null hypothesis** is a statement that the parameter takes a particular value.
- The **alternative hypothesis** states that the parameter falls in some alternative range of values.
- The symbol H_0 denotes the null hypothesis, and the symbol H_a denotes the alternative hypothesis.

Significance Test IV

Example

- Through the use of fracking, a drilling method that uses high-pressure water and chemicals to extract oil and natural gas from underground rock formations, the United States has become the largest oil producer in the world



Significance Test V

- Despite its economic benefits, fracking is becoming more and more controversial due to its potential effects on the environment
- Some U.S. states and other countries have already banned it
- Let's investigate whether those who oppose the increased use of fracking in the United States are still in the minority
- Let p denote the proportion of people in the United States who oppose the increased use of fracking

Significance Test VI

- In the United States, the proportion of people who oppose the increased use of fracking is less than 0.50
  a. Is this a null or an alternative hypothesis?
  b. How can we express the hypothesis that the population proportion opposed to fracking may actually be 0.50?

Significance Test VII

3. Test statistic: describes **how far** that point estimate falls from the parameter value given in the null hypothesis
   - Particularly, let's consider the Z-distribution

$$z = \frac{\overline{x} - \mu_0}{se_0}$$

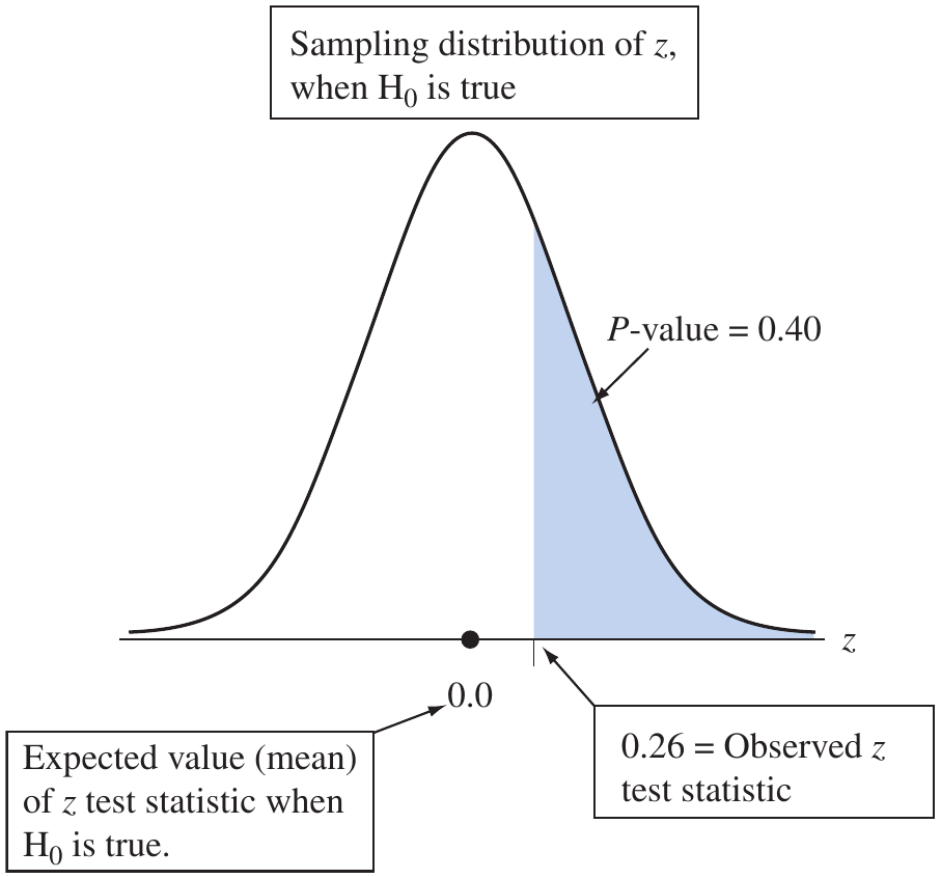
   - The estimate of the probability p is the sample proportion $\hat{p} = \frac{40}{116} = 0.345$. Which z value is: $z = 0.26$

Significance Test VIII

4 Compute the p-value:

Definition (P-Value)

The P-value is the probability that the test statistic equals the observed value or a value even more extreme. It is calculated by presuming that the null hypothesis H_0 is true.



Analysis of Variance (ANOVA)

Analysis of Variance

- The analysis of variance is a significance test of the null hypothesis of equal population means
- Consider g groups and their corresponding expected values $\mu_1, \ldots, \mu_g$
- ANOVA is a significance test for null hypothesis:

$$H_0 : \mu_1 = \mu_2 =, \ldots, = \mu_n$$

- Alternative hypothesis?

Analysis of Variance

- The analysis of variance is a significance test of the null hypothesis of equal population means
- Consider g groups and their corresponding expected values $\mu_1, \ldots, \mu_g$
- ANOVA is a significance test for null hypothesis:

$$H_0 : \mu_1 = \mu_2 =, \ldots, = \mu_n$$

- Alternative hypothesis?

H_a: at least 2 of the population means are unequal

Assumption

The assumptions for ANOVA are as follows:

1. The population distributions of the response variable for the g groups are normal, with the same standard deviation for each group.
2. **Randomization:** in the survey sample, independent random samples are selected from each of g population.

If H_0 is not rejected?

Assumption

The assumptions for ANOVA are as follows:

1. The population distributions of the response variable for the g groups are normal, with the same standard deviation for each group.
2. **Randomization:** in the survey sample, independent random samples are selected from each of g population.

If H_0 is not rejected?

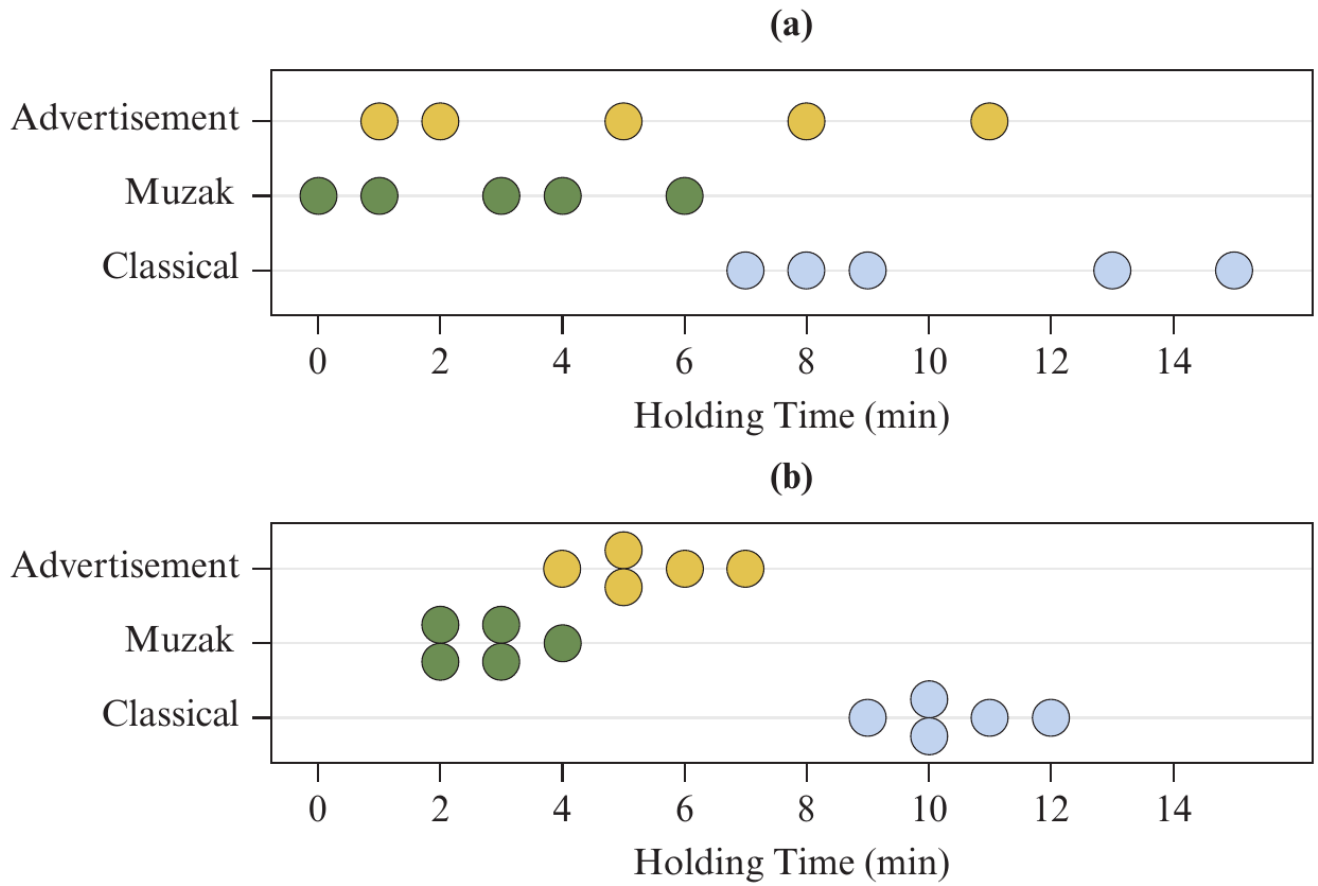
The population distribution does not depend on the group to which a subject belongs.

Why is it called “Analysis of Variance”?

- Analysis of Variance method is used to compare population mean?

Why is it called “Analysis of Variance”?

- The test statistic uses evidence about two types of variability.



ANOVA F Test Statistic I

A key idea of ANOVA is the decomposition of the total variability into two components:

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

ANOVA F Test Statistic II

Let $n_1, n_2, \ldots, n_g$ be the sample sizes of the groups, total observations:

$$m = \sum_{i=1}^{g} n_i$$

Let μ be the data mean:

$$\mu = \frac{1}{m} \sum_{i=1}^{g} \sum_{j=1}^{n_i} x_{ij} = \sum_{i=1}^{g} \frac{n_i}{m} \mu_i$$

ANOVA F Test Statistic III

Total variability of all observations

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\mu)^2 = \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left((x_{ij}-\mu_i)+(\mu_i-\mu)\right)^2 \tag{1}$$

$$= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left((x_{ij}-\mu_i)^2 - 2(x_{ij}-\mu_i)+(\mu_i-\mu)^2\right) \tag{2}$$

$$= \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}_{V_I} + \underbrace{\sum_{i=1}^{g} n_i(\mu_i-\mu)^2}_{V_B} \tag{3}$$

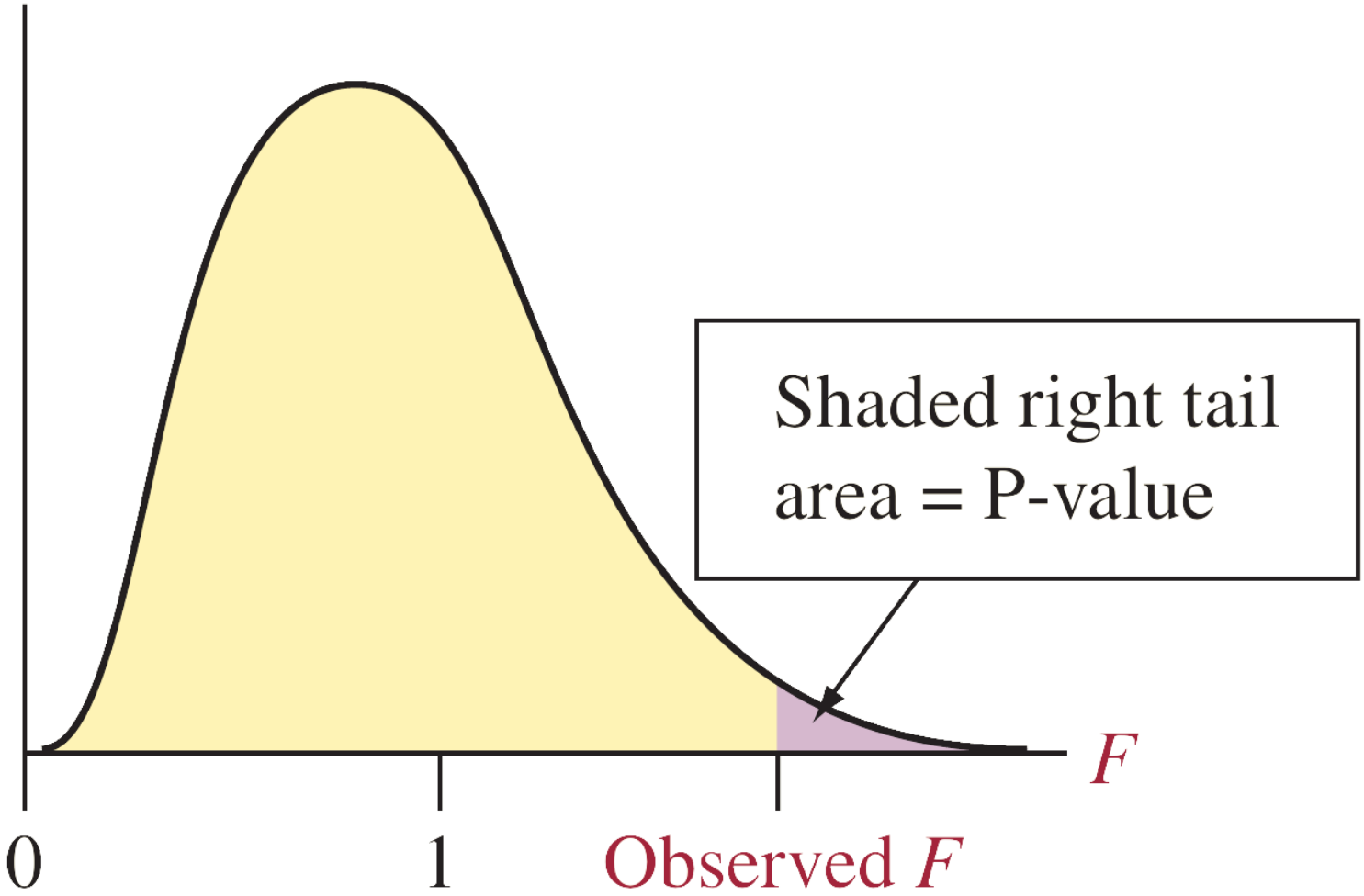
ANOVA F Test Statistic IV

The F statistic of ANOVA:

$$F = \frac{V_I / df_{V_I}}{V_B / df_{V_B}}$$

where $df_{V_I} = g - 1$ and $df_{V_B} = m - g$ are degrees of freedom of V_I and V_B.

ANOVA F Test Statistic V



Example: I

The airline recently conducted a randomized experiment to analyze whether callers would remain on hold longer, on average

1. An advertisement about the airline and its current promotions
2. Muzak
3. Vivaldi' Four Seasons

For each call, they randomly selected one of the three recordings to play and then measured the number of minutes that the caller remained on hold before hanging up (these calls were purposely not answered)

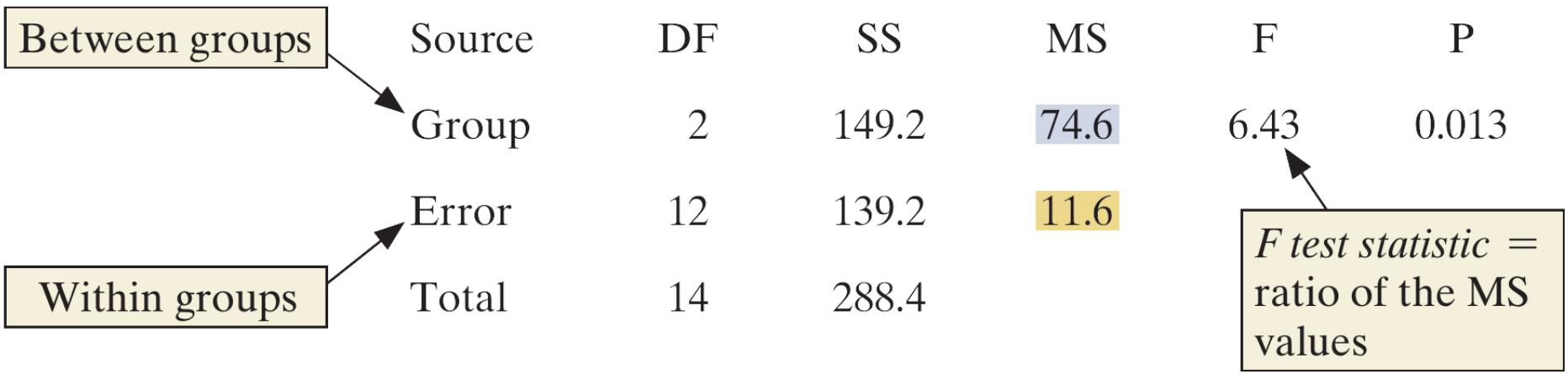
Example: II

Recording	Holding Time Observations	Sample Size	Mean	Standard Deviation
Advertisement	5, 1, 11, 2, 8	5	5.4	4.2
Muzak	0, 1, 4, 6, 3	5	2.8	2.4
Classical	13, 9, 8, 15, 7	5	10.4	3.4

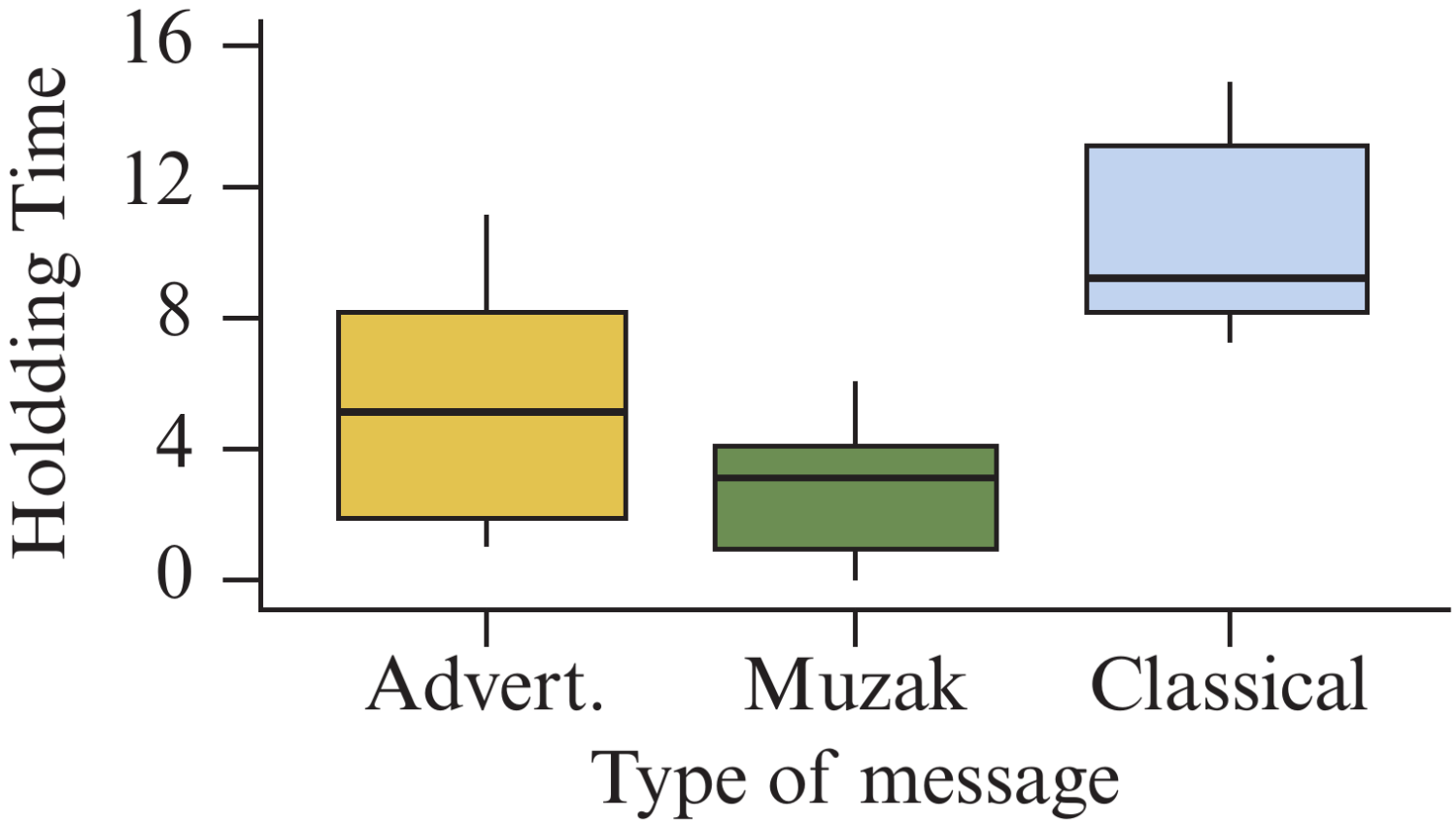
- Hypotheses?
- F-test?
- P-value?
- Conclusion?

Example: III

ANOVA table of F test



Example: IV


Holdding Time
16
12
8
4
0
Advert.
Muzak
Classical
Type of message

Amounts of Fertilizer and Manure

- A large field was portioned into 20 equal-size plots.
- Each plot was planted with the same amount of corn seed, using a fixed spacing pattern between the seeds.
- The goal was to study how the yield of corn later harvested from the plots (in metric tons) depends on the level of nitrogen-based fertilizer and manure
  - The fertilizer level was low ($45kg/ha$) or high ($135kg/ha$)
  - The manure level was low ($84kg/ha$) or high ($168kg/ha$)

Analysis

Fertilizer Level	Manure Level	Plot					Sample Size	Mean	Std. Dev.
		1	2	3	4	5			
High	High	13.7	15.8	13.9	16.6	15.5	5	15.1	1.3
High	Low	16.4	12.5	14.1	14.4	12.2	5	13.9	1.7
Low	High	15.0	15.1	12.0	15.7	12.2	5	14.0	1.8
Low	Low	12.4	10.6	13.7	8.7	10.9	5	11.3	1.9

Analysis

Fertilizer Level	Manure Level	Plot					Sample Size	Mean	Std. Dev.
		1	2	3	4	5			
High	High	13.7	15.8	13.9	16.6	15.5	5	15.1	1.3
High	Low	16.4	12.5	14.1	14.4	12.2	5	13.9	1.7
Low	High	15.0	15.1	12.0	15.7	12.2	5	14.0	1.8
Low	Low	12.4	10.6	13.7	8.7	10.9	5	11.3	1.9

Hypothesis?

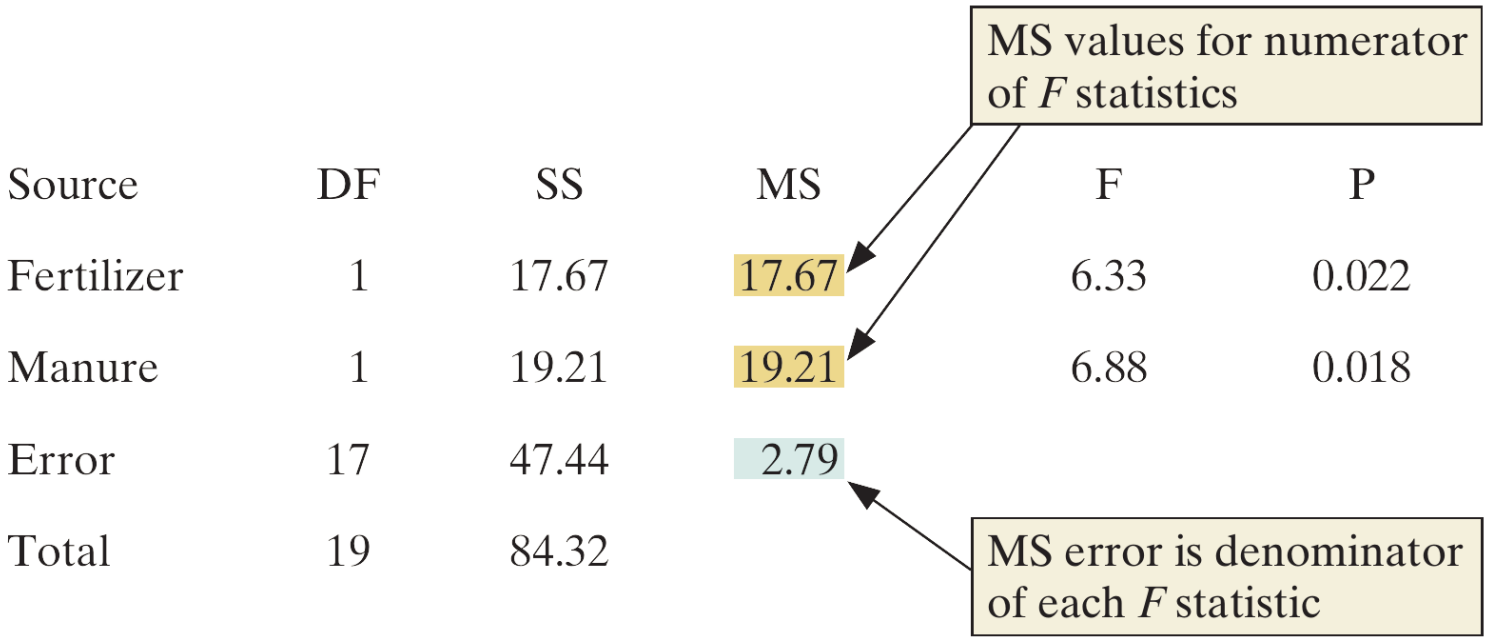
Analysis

Fertilizer Level	Manure Level	Plot					Sample Size	Mean	Std. Dev.
		1	2	3	4	5			
High	High	13.7	15.8	13.9	16.6	15.5	5	15.1	1.3
High	Low	16.4	12.5	14.1	14.4	12.2	5	13.9	1.7
Low	High	15.0	15.1	12.0	15.7	12.2	5	14.0	1.8
Low	Low	12.4	10.6	13.7	8.7	10.9	5	11.3	1.9

	Fertilizer	
Manure	Low	High
Low	11.3	13.9
High	14.0	15.1

Two-way ANOVA

Figure 1: Two-way ANOVA for corn yield



ANOVA and regression I

		Indicator Variables		
Fertilizer	**Manure**	f	m	**Mean of y**
High	High	1	1	$\alpha + \beta_1 + \beta_2$
High	Low	1	0	$\alpha + \beta_1$
Low	High	0	1	$\alpha + \beta_2$
Low	Low	0	0	α

Regression model for the mean corn yield

$$\mu_y = \alpha + \beta_1 f + \beta_2 m$$

ANOVA and regression II

Prediction model:

$$\hat{\mu}_y = 11.6 + 1.9f + 2.0m$$

	Fertilizer	
Manure	**Low**	**High**
Low	11.6	11.6 + 1.9 = 13.5
High	11.6 + 2.0 = 13.6	11.6 + 1.9 + 2.0 = 15.5

Summary

- Classification Model Evaluation: Accuracy, Precision, Recall, ROC
- Significance Level is a thresold of p-value such that we reject H_0
- When data provides enough evidence that rejects H_0, the test result is statistically significance
- The ANOVA F test is robust to moderate breakdowns in the population normality and equal standard deviation assumptions