Descriptive statistics, TTest, and ANOVA

Stephen Opiyo

PATIRA DATA SCIENCE
DISCOVERING KNOWLEDGE FROM DATA

THE OHIO STATE UNIVERSITY

Overview

- Measure of central tendency (summary statistics)
- Measure of variability
- Hypothesis testing
- TTest and ANOVA

Why study statistics

Need to understand and use data to make decisions.
To do this, we must:

- Decide weather existing data is adequate.
- If necessary collect more data.
- Summarize the available data in a useful information.
- Analyze the available data.
- Draw conclusions, make decisions.

Two types of statistics

- Descriptive statistics:
  - data used for descriptive purposes no predictions.
  - use tables and graphs.
- Inferential statistics:
  - employ data to draw inferences (conclusions).
  - sample data employed to infer populations.
  - sample must be random.

Measures of central tendency

- A single value that attempts to describe a set of data by identifying the central position.
- The most common measures of central tendency are:
  - Mean
  - Median
  - Mode

What is mean?

The "mean" is the average score or value. Example the average age of participants in this workshop.

- Inferential mean of a sample: $X=(\sum X)/n$
- Mean of a population: $\mu=(\sum X)/N$

Problem with mean

- The main problem associated with the mean value of some data is that it is sensitive to outliers.
- Example, the average age of participants might be affected if there was one with the age of 80.
- Mean is affected when data is skewed (i.e., the frequency distribution of the data is skewed).

Mean of students

Student group 1	Age group 1	Student group 2	Age group 2
Schmuggles	28	Schmuggles	28
Bopsey	31	Bopsey	31
Pallitto	49	Pallitto	49
Homer	42	Homer	42
Schnickerson	30	Schnickerson	30
Levin	39	Levin	39
Honkey-Doorey	32	Honkey-Doorey	32
Zingers	80	Amy	30
Boehmer	48	Boehmer	48
Queenie	40	Queenie	40
Mean	**41.9**		**36.9**

The median

- Because the mean (average) can be sensitive to extreme values, the median is sometimes useful and more accurate.
- The median is simply the middle value among some scores of a variable. (no standard formula for its computation)

What is the median?

Student	Weight
Schmuggles	165
Bopsey	213
Pallitto	189
Homer	187
Schnickerson	165
Levin	148
Honkey-Doorey	251
Zingers	308
Boehmer	151
Queenie	132
Googles-Boop	199
Calzone	227
Mean	**194.6**

Rank order and choose middle value.

If even then average between two in the middle

(187 + 189) = 188

Median = 188

Weight
132
148
151
165
165
187
189
199
213
227
251
308

The Mode

- The most frequent response or value for a variable.
- Multiple modes are possible: bimodal or multimodal.

Figuring the Mode

Student	Weight
Schmuggles	165
Bopsey	213
Pallitto	189
Homer	187
Schnickerson	165
Levin	148
Honkey-Doorey	251
Zingers	308
Boehmer	151
Queenie	132
Googles-Boop	199
Calzone	227

What is the mode?

Answer: 165

Variability

- Measure of central tendency only gives partial information about a data set.
- It is important to describe how much the observations differ from one another.
- Measures of dispersion give us information about how much our variables vary from the mean, because if they don't it makes it difficult to infer anything from the data. Dispersion is also known as the spread or range of variability.

Measure of dispersion

- Measures of dispersion tell us about variability in the data.
- Basic question: how much do values differ for a variable from the min to max, and distance among scores in between. We use:
  - Range
  - Standard Deviation
  - Variance

Measure of Variability

- Most commonly employed measure of variability are the range, standard deviation, and variance.
- The range:
  The difference between the highest and the lowest scores in a distribution
  - Example: 79, 56, 74, 22, 35, 87, 60, 60, 61, 62, 62, 62, 62, 21, 74,23

    The range is (**87 - 21 = 66)**)

The standard deviation

- A standardized measure of distance from the mean.
- Very useful and something you do read about when making predictions or other statements about the data.

Standard deviation and variance

$$S = \sqrt{\frac{(X \quad \overline{X})^2}{(n-1)}} \qquad Var_x = \frac{\Sigma(x-\bar{x})^2}{(n-1)} = \frac{\Sigma(x-\bar{x})(x-\bar{x})}{(n-1)} = \frac{\Sigma x^2 - n\bar{x}^2}{(n-1)}$$

$\sqrt{\ }$ =square root
Σ=sum (sigma)
X=score for each point in data
$\overline{X}$=mean of scores for the variable
n=sample size (number of observations or cases

Measure of Variability

- Standard deviation: Is a measure of how spread out scores are
- Example: 79, 56, 74, 22, 35, 87, 60, 61, 62, 21, 23 calculate the SD
- Start by calculating mean $\overline{X}$

Sample No	Score X
1	21
2	22
3	23
4	35
5	56
6	60
7	61
8	62
9	74
10	79
11	87
Total	580
	Mean $\overline{X}$ =(580/11)=52.7

$$\sqrt{\frac{\Sigma(X - \overline{X})^2}{(n-1)}}$$

where:
X = each score
$\overline{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

Measure of Variability

- Standard deviation: Is a measure of how spread out scores are
- Subtract the mean from each score $X - \bar{X}$

Score X	$X - \bar{X}$
21	21 - 52.7 =-317
22	22 - 52.7 =-307
23	23 - 52.7 =-297
35	35 - 52.7 =177
56	56 - 52.7 =3.3
60	60 - 52.7 =7.3
61	61 - 52.7 =8.3
62	62 - 52.7 =9.3
74	74 - 52.7 =213
79	79 - 52.7 =263
87	87 - 52.7 =34.3
Mean = $\bar{X}$ =(580/11)=52.7	

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

Measure of Variability

- Standard deviation: Is a measure of how spread out scores are
- Square the difference $(X - \bar{X})^2$

Score X	$X - \bar{X}$	$(X - \bar{X})^2$
21	21 - 52.7 =-317	$(-31.7)^2$ =1004.8
22	22 - 52.7 =-307	$(-30.7)^2$ =942.5
23	23 - 52.7 =-297	$(-29.7)^2$ =882.1
35	35 - 52.7 =177	$(-17.7)^2$ =313.3
56	56 - 52.7 =3.3	$(3.3)^2$ =10.9
60	60 - 52.7 =7.3	$(7.3)^2$ =53.3
61	61 - 52.7 =8.3	$(8.3)^2$ =68.9
62	62 - 52.7 =9.3	$(9.3)^2$ =86.5
74	74 - 52.7 =213	$(21.3)^2$ =453.7
79	79 - 52.7 =263	$(26.3)^2$ =691.7
87	87 - 52.7 =34.3	$(34.3)^2$ =1176.5
Mean = $\bar{X}$ =(580/11)=52.7		

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

Measure of Variability

- Standard deviation: Is a measure of how spread out scores are
- Sum the square of the difference $\Sigma(X - \bar{X})^2$

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

Score X	$X - \bar{X}$	$(X - \bar{X})^2$	
21	21 - 52.7 =-317	$(-31.7)^2$ =1004.8	1004.8
22	22 - 52.7 =-307	$(-30.7)^2$ =942.5	942.5
23	23 - 52.7 =-297	$(-29.7)^2$ =882.1	882.1
35	35 - 52.7 =177	$(-17.7)^2$ =313.3	313.3
56	56 - 52.7 =3.3	$(3.3)^2$ =10.9	10.9
60	60 - 52.7 =7.3	$(7.3)^2$ =53.3	53.3
61	61 - 52.7 =8.3	$(8.3)^2$ =68.9	68.9
62	62 - 52.7 =9.3	$(9.3)^2$ =86.5	86.5
74	74 - 52.7 =213	$(21.3)^2$ =453.7	453.7
79	79 - 52.7 =263	$(26.3)^2$ =691.7	691.7
87	87 - 52.7 =34.3	$(34.3)^2$ =1176.5	1176.5
Mean = $\bar{X}$ =(580/11)=527			$\Sigma(X - \bar{X})^2$ =5684.2

Measure of Variability

- Standard deviation: Is a measure of how spread out scores are
- Calculate the standard deviatioı $\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$

$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$

where:
X = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

Score X	$X - \bar{X}$	$(X - \bar{X})^2$	
21	21 - 52.7 = -317	$(-31.7)^2$ = 1004.8	1004.8
22	22 - 52.7 = -307	$(-30.7)^2$ = 942.5	942.5
23	23 - 52.7 = -297	$(-29.7)^2$ = 882.1	882.1
35	35 - 52.7 = 177	$(-17.7)^2$ = 313.3	313.3
56	56 - 52.7 = 3.3	$(3.3)^2$ = 10.9	10.9
60	60 - 52.7 = 7.3	$(7.3)^2$ = 53.3	53.3
61	61 - 52.7 = 8.3	$(8.3)^2$ = 68.9	68.9
62	62 - 52.7 = 9.3	$(9.3)^2$ = 86.5	86.5
74	74 - 52.7 = 213	$(21.3)^2$ = 453.7	453.7
79	79 - 52.7 = 263	$(26.3)^2$ = 691.7	691.7
87	87 - 52.7 = 34.3	$(34.3)^2$ = 1176.5	1176.5
Mean = $\bar{X}$ = (580/11) = 52-7			$\Sigma(X - \bar{X})^2$ = 5684.2

Standard deviation = $\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$ = sqrt((5684.2)/(11-1)) = **23.84**

Variance = standard deviation X standard deviation = 23.84 X 23.84 = **568.3**

Hypothesis Testing, TTest, and ANOVA

Hypothesis Testing

- A **Hypothesis** is a claim or statement about the value of single population characteristics or values of several population characteristics. In addition, a **hypothesis** is a statement that something is true.
- **Null hypothesis:** A hypothesis to be tested. We use the symbol H_0 to represent the null hypothesis
- **Alternative hypothesis:** A hypothesis to be considered as an alternative to the null hypothesis. We use the symbol H_a *or* H_1 to represent the alternative hypothesis.

Hypothesis testing

- The form of null hypothesis is:
  - H_o: population characteristic = hypothesized value.
- The alternative hypothesis will have on of the following three forms:
  - H_a: population characteristic > hypothesized value.
  - H_a: population characteristic < hypothesized value.
  - H_a: population characteristic ≠ hypothesized value.

Hypothesis Testing

The critical concepts are these

1. The procedure begins with the assumption that the ***H_0*** is true.

2. The goal is to determine whether there is enough evidence to infer that ***H_a*** or **H_1** is true, or ***H_0*** is not likely to be true.

3. There are two possible decisions:

- Conclude that there is **enough** evidence to support the **alternative** hypothesis. Reject the null.
- Conclude that there is **not enough** evidence to support the **alternative** hypothesis. Fail to reject the null.

Interpretation

- *P*-value answer the question: What is the probability of the observed test statistic … **when *H_0* *is true*?**
- Thus, smaller and smaller *P*-values provide stronger and stronger evidence against H_0
- Small *P*-value ⇒ strong evidence

Interpretation

Conventions

$P > 0.10 \Rightarrow$ non-significant evidence against H_0

$0.05 < P \leq 0.10 \Rightarrow$ marginally significant evidence

$0.01 < P \leq 0.05 \Rightarrow$ significant evidence against H_0

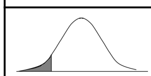
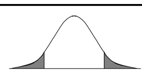
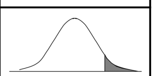
$P \leq 0.01 \Rightarrow$ highly significant evidence against H_0

Examples

$P = 0.27 \Rightarrow$ non-significant evidence against H_0

$P = 0.01 \Rightarrow$ highly significant evidence against H_0

Example: summary of One- and Two-Tail Tests

One-Tail Test (left tail)	Two-Tail Test	One-Tail Test (right tail)
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$
		

Summary: Hypothesis Testing

<u>The Steps:</u>

1. Define your hypotheses (null, alternative)
2. Specify your null distribution
3. Do an experiment
4. Calculate the p-value of what you observed
5. Reject or fail to reject (~accept) the null hypothesis

Hypothesis testing

- Examples of methods used in hypothesis testing:
  - Z-test
  - TTest
  - Chi-square test
  - ANOVA
  - Non-parametric statistics

T-Test

Overview of T-test

- What is the main use of the *t*-test?
- How is the distribution of t related to the unit normal?
- When would we use a t-test instead of a z-test
- Why might we prefer one to the other?
- What are the chief varieties or forms of the *t*-tests?

Overview of T-test continued

- Identify the appropriate version of *t* to use for a given design.
- Compute and interpret *t*-tests appropriately.

Background

- The *t*-test is used to test hypotheses about means when the population variance is unknown (the usual case). Closely related to Z, the unit normal.
- Comes in 3 varieties:
  - Single sample
  - Independent samples
  - Dependent samples

What kind of *t* is it?

- Single sample *t* – we have only 1 group; want to test against a hypothetical mean.
- Independent samples *t* – we have 2 means, 2 groups; no relation between groups, e.g., people randomly assigned to a single group.
- Dependent *t* – we have 2 means. Either same people in both groups, or people are related, e.g., husband-wife, left hand-right hand, hospital patient and visitor.

Assumptions of t-Test

- Dependent variables are interval or ratio.
- The population from which samples are drawn is normally distributed.
- Samples are randomly selected.
- The t-statistic is robust (it is reasonably reliable even if assumptions are not fully met.

Independent *t* Test

- Compares the difference between **two means** of two **independent** groups.
- The comparison distribution is a difference between means to a distribution of *differences between means.*
  - Population of measures for Group 1 and Group 2
  - Sample means from Group 1 and Group 2
  - Population of differences between sample means of Group 1 and Group 2

Independent *t* Test

The *t* Test compares the averages and standard deviations of two samples to see if there is a significant difference between them.

We start by calculating a number, *t*

t is calculated using the equation:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

Where:

$\bar{x}_1$ is the mean of sample 1
s_1 is the standard deviation of sample 1
n_1 is the number of individuals in sample 1
$\bar{x}_2$ is the mean of sample 2
s_2 is the standard deviation of sample 2
n_2 is the number of individuals in sample 2

Worked Example: Random samples were taken for heights of students in C1 and D8

Their recorded heights are shown below…

	Students in C1					Students in D8				
Student Height (cm)	145	149	152	153	154	148	153	157	161	162
	154	158	160	166	166	162	163	167	172	172
	166	167	175	177	182	175	177	183	185	187

Step 1: Work out the mean height for each sample

C1: $\overline{x_1}$ = 161.60 D8: $\overline{x_2}$ = 168.27

Step 2: Work out the difference in means

$\overline{x_2} - \overline{x_1}$ = 168.27 – 161.60 = 6.67

Step 3: Work out the standard deviation for each sample

C1: s_1 = 10.86 D8: s_2 = 11.74

Step 4: Calculate s^2/n for each sample

C1: $\frac{(s_1)^2}{n_1} = 10.86^2 \div 15 = 7.86$

D8: $\frac{(s_2)^2}{n_2} = 11.74^2 \div 15 = 9.19$

Step 5: Calculate $\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}$

$$\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}} = \sqrt{(7.86 + 9.19)} = 4.13$$

Step 6: Calculate *t* (Step 2 divided by Step 5)

$$t = \frac{\overline{x_2} - \overline{x_1}}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} = \frac{6.67}{4.13} = 1.62$$

Step 7: Work out the number of degrees of freedom

d.f. = $n_1 + n_2 - 2$ = 15 + 15 – 2 = 28

Step 8: Find the critical value of *t* for the relevant number of degrees of freedom

Use the 95% (p=0.05) confidence limit

Critical value = ?

t Distribution critical values Table

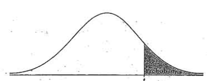


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	[illegible]	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	[illegible]	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level *C*											

Step 7: Work out the number of degrees of freedom

d.f. = $n_1 + n_2 - 2$ = 15 + 15 – 2 = 28

Step 8: Find the critical value of *t* for the relevant number of degrees of freedom

Use the 95% (p=0.05) confidence limit

Critical value = 2.048

Our calculated value of t = 1.62 is below the critical value of 2.048 for 28 df, therefore, there is no significant difference between the height of students in samples from C1 and D

ANOVA:
Analysis of Variance

46

Overview

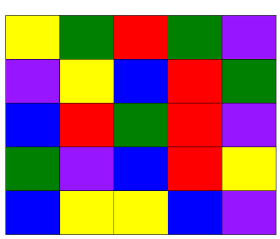
- Completely randomized and randomized complete block design
- What is ANOVA?
- When is it useful?
- How does it work?
- One-way ANOVA
- Post-hoc Tests
- Two-way ANOVA

47

Completely randomized design

- Subjects are randomly assigned to treatments
- Relies on randomization to control for the effects of extraneous variables

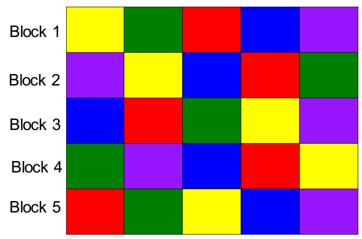
- **Five treatments**
  - T1 = Blue
  - T2 = Red
  - T3 = Army green
  - T4 = Yellow
  - T5 = Purple



48

Randomized complete block design

- Subjects are divided into subgroups called blocks, such that the variability within blocks is less than the variability between blocks



49

Definition

- <u>ANOVA:</u> ANALYSIS OF VARIANCE is the analysis of variation in an experimental outcome, especially of a statistical variance in order to determine the contributions of given factors or variables to the variance.
- <u>Remember:</u> Variance = the square of the standard deviation.

50

Data set

- Any data set has variability
- Variability exists within groups
- Variability exists between groups
- Question that ANOVA allows us to answer: Is this variability significant, or merely by chance?

51

ANOVA model Assumptions

- Independence (response variables y_i are independent).
- Normality (response variables are normally distributed).
- Homoscedasticity (the response variables have the same variance).

52

ANOVA model assumptions

- Test for normality
  - **Shapiro-Wilk Test**
- Testing variances
  - **Levene's Test for Equality of Variances**

53

Test for normality by Shapiro-Wilk

- A Shapiro-Wilk
  - You fail reject the null when your observed value is **greater** than your critical value (that's right, the critical region on this test is in the small tail)
  - For example, P-value is greater than 0.05, conclude normality.

54

ANOVA Notation

- k = number of groups
- n = number of observations in each group
- x_{ij} = one observation j in group i
- $\bar{x}$ = mean over all groups x
- $\bar{x}_i$ = mean for group i
- SS = Sum of Squares
- MS = Mean of Squares
- F = Between MS/Within MS

55

ANOVA usefulness

- Similar to t-test
- More versatile than t-test
- Compare one parameter (response variable) between two or more groups

56

What does ANOVA do?

At its simplest ANOVA tests the following:

H_0: The means of all the groups are equal.

H_a: Not all the means are equal

- doesn't say how or which ones differ.
- Can follow up with "multiple comparisons tests"

57

What ANOVA does not do

- Tell which groups are different
  - Post-hoc test of mean differences required
- Compare multiple parameters for multiple groups (so it cannot be used for multiple response variables)

58

ANOVA variations

- One-Way
- Factorial (Two-Way, Three-Way) ANOVA
- Repeated measures ANOVA
- MANOVA (Multiple analysis of variance)
  - multiple response variables

59

Summary

- ANOVA:
  - allows us to know if variability in a data set is between groups or merely within groups.
  - is more versatile than t-test.
  - can compare multiple groups at once.
  - cannot process multiple response variables.
  - does not indicate which groups are different.

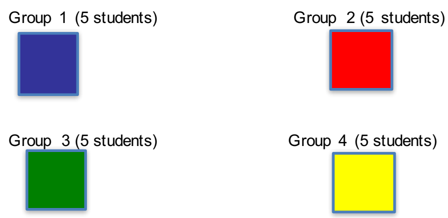
60

One-Way ANOVA

- One factor (manipulated variable)
- One response variable
- Two or more groups to compare

61

Between and within groups



62

How ANOVA works

ANOVA measures two sources of variation in the data and compares their relative sizes

- variation BETWEEN groups
  - for each data value look at the difference between its group mean and the overall mean

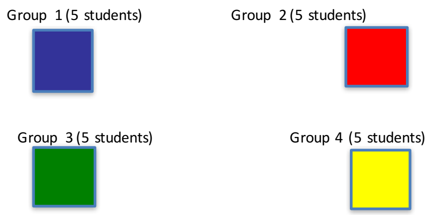
$$(\bar{x}_i - \bar{x})^2$$

- variation WITHIN groups
  - for each data value we look at the difference between that value and the mean of its group

$$(x_{ij} - \bar{x}_i)^2$$

63

Between and within groups



Between groups $(\bar{x}_i - \bar{x})^2$ Within groups $(x_{ij} - \bar{x}_i)^2$

64

F-statistic

The ANOVA F-statistic is a ratio of the Between Group Variation divided by the Within Group Variation:

$$F = \frac{Between}{Within} = \frac{MSG}{MSE}$$

A large F is evidence ***against*** **H_0**, since it indicates that there is more difference between groups than within groups.

65

F-Ratio = MS_{Bet}/MS_{w}

- If:
  - The ratio of Between-Groups MS: Within-Groups MS is LARGE→ reject H_0→ there *is* a difference between groups
  - The ratio of Between-Groups MS: Within-Groups MS is SMALL→ do not reject H_0→ there is *no* difference between groups

66

P-values

- Use F-table to determine p
- Use df for numerator and denominator
- Choose level of significance
- If calculated F values > critical value, reject the null hypothesis (for one-tail test)

67

F-table

VassarStats: Table of Critical F Values (p.1)
[top entry for .05 level; bottom entry for .01 level]

df denominator	df numerator 1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	161	199	216	225	230	234	237	239	241	242	243	244	245	245
	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42
	99	99	99	99	99	99	99	99	99	99	99	99	99	99
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.98	26.92
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64
	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96
	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53
	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24
	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03
	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86
	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74
	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64
	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55
	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48
	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56

Computing ANOVA F statistic

			WITHIN		BETWEEN	
			difference:		difference	
		group	data - group mean		group mean - overall mean	
data	group	mean	plain	squared	plain	squared
5.3	1	6.00	-0.70	0.490	-0.4	0.194
6.0	1	6.00	0.00	0.000	-0.4	0.194
6.7	1	6.00	0.70	0.490	-0.4	0.194
5.5	2	5.95	-0.45	0.203	-0.5	0.240
6.2	2	5.95	0.25	0.063	-0.5	0.240
6.4	2	5.95	0.45	0.203	-0.5	0.240
5.7	2	5.95	-0.25	0.063	-0.5	0.240
7.5	3	7.53	-0.03	0.001	1.1	1.188
7.2	3	7.53	-0.33	0.109	1.1	1.188
7.9	3	7.53	0.37	0.137	1.1	1.188
TOTAL				1.757		5.106
TOTAL/df				**0.25095714**		**2.55275**

overall mean: 6.44; MSG = 5.106/(3-1)= 2.5528; MSW=(1.757/(10-3) = 0.251
F = 2.5528/0.251 = 10.21575

69

F-table

VassarStats: Table of Critical F Values (p. 1)
[top entry for .05 level; bottom entry for .01 level]

df denominator	df numerator: 1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	161	199	216	225	230	234	237	239	241	242	243	244	245	245
	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056	6083	6107	6126	6143
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.40	19.41	19.42	19.42
	99	99	99	99	99	99	99	99	99	99	99	99	99	99
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71
	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.34	27.23	27.13	27.05	26.98	26.92
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87
	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.45	14.37	14.31	14.25
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64
	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89	9.82	9.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96
	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72	7.66	7.60
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53
	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47	6.41	6.36
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24
	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67	5.61	5.56
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03
	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11	5.05	5.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86
	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71	4.65	4.60
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74
	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40	4.34	4.29
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64
	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16	4.10	4.05
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55
	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96	3.91	3.86
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48
	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80	3.75	3.70
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42
	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67	3.61	3.56

R ANOVA Output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	2	5.127	2.552	10.22	0.00839 **
Residuals	7	1.757	0.251		
Total	9	6.884			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

1 less than number of groups

number of data values - number of groups (equals df for each group added together)

71

R ANOVA Output

Analysis of Variance

Source	DF	SS	MS	F	P
treatment	2	5.127	2.564	10.22	0.00839
Error	7	1.757	0.251		
Total	9	6.884			

$$\sum_{obs}(x_{ij}-\bar{x}_i)^2$$

$$\sum_{obs}(x_{ij}-\bar{\bar{x}})^2$$

$$\sum_{obs}(\bar{x}_i-\bar{\bar{x}})^2$$

SS stands for sum of squares
MS stands for mean squares

72

R ANOVA Output

```
Analysis of Variance
Source      DF        SS         MS        F       P
treatment    2     5.127      2.564    10.22   0.00839
Error        7     1.757      0.251
Total        9     6.884
```

MSG = SSG / DFG
MSE = SSE / DFE

F = MSG / MSE

P-value comes from F(DFG,DFE)

(P-values for the F statistic are in Table)

73

R^2 Statistic

R^2 gives the percent of variance due to between group variation

$$R^2 = \frac{SS[Between]}{SS[Total]} = \frac{SSG}{SST}$$

We will discuss more on R^2 when we study regression.

74

Where's the Difference?

Once ANOVA indicates that the groups do not all appear to have the same means, what do we do?

75

Multiple Comparisons

Once ANOVA indicates that the groups do not all have the same means, we can compare them two by two using the 2-sample t test

- We need to adjust our p-value threshold because we are doing multiple tests with the same data.
- There are several methods for doing this.
- If we really just want to test the difference between one pair of treatments, we should set the study up that way.

76

Post-hoc Comparisons of Treatments

- If differences in group means are determined from the *F*-test, researchers want to compare pairs of groups. Three popular methods include:
  - Fisher's LSD - Upon rejecting the null hypothesis of no differences in group means, LSD method is equivalent to doing pairwise comparisons among all pairs of groups.
  - Tukey's Method - Specifically compares all $t(t-1)/2$ pairs of groups.
  - Bonferroni's Method - Adjusts individual comparison error rates so that all conclusions will be correct at desired confidence/significance level. Any number of comparisons can be made. Very general approach can be applied to any inferential problem

77

Factorial ANOVA

- ANOVA in which there are more than one independent variable (factor)
- Seldom uses more than three or four factors
- Examples: Testing the effect of fertilizer rates, and types on yield; testing the effect of temperature and light on bacteria growth.

78

Components of Factorial ANOVA

- **Interaction** – present when the differences between the groups of one independent variable on the dependent variable vary according to the level of a second independent variable [e.g., crop yield, fertilizer, varieties]
- **Main effects –** test of each independent variable (IV) when all other independent variables are held constant

79

Questions of Interest

- Generally, the questions of interest here (i.e. hypotheses to be tested) concern three questions regarding the potential effects of the factors on the response variable.
- ***Question 1:*** Do the effects that factors *a* and *b* have on the response variable ***interact***, i.e. is there a significant ***interaction*** between factors *a* and *b* ?

80

Questions of Interest

- ***Question 1:*** If we conclude there is a significant interaction then we conclude **the effects of both factors a and b are significant!**
- When we have an interaction we cannot consider the effect of either factor independently of the other, therefore both factors matter.

81

Questions of Interest

- If there is **not a significant interaction** effect then we can consider the main effects separately, i.e. we ask the following:
- ***Question 2:*** Does factor *a* alone have a significant effect?
- ***Question 3:*** Does factor *b* alone have a significant effect?

82

Tests of Hypotheses

Just as we had Sums of Squares and Mean Squares in ***One-way ANOVA***, we have the same in ***Two-way ANOVA:***

Recall, Mean Squares are measures of variability across the levels of the relevant factor of interest.

In balanced ***Two-way ANOVA***, we measure the overall variability in the data by:

83

Two-way ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	*F*-ratio	*P-value*
Factor *A*	$a - 1$	SS_A	MS_A	$F_A = MS_A / MS_E$	Tail area
Factor *B*	$b - 1$	SS_B	MS_B	$F_B = MS_B / MS_E$	Tail area
Interaction	$(a - 1)(b - 1)$	SS_{AB}	MS_{AB}	$F_{AB} = MS_{AB} / MS_E$	Tail area
Error	$ab(n - 1)$	SS_E	MS_E		
Total	$abn - 1$	SS_T			

This is our initial focus which is the p-value for Question 1: Is there an interaction effect?

84

Tests of Hypotheses

- If the interaction ***is not*** statistically significant (i.e. *p-value* > 0.05) then we conclude the main effects (if present) are independent of one another.
- We can then test for significance of the main effects separately, again using an F-test.
- If a main effect is significant we can then use multiple comparison procedures as usual to compare the mean response for different levels of the factor while holding the other factor fixed.

85