Statistical and Mathematical Methods

Statistical and Mathematical Methods for Data Science
DS5003

Dr. Nasir Touheed

Anova

Hypothesis testing

- We often use inferential statistics to make decisions or judgments about the value of a parameter, such as a population mean
- Based on a random sample, we can use Statistics to verify whether
  - a system has not been infected,
  - a hardware upgrade was efficient,
  - the mean age, of all cars in use has increased from the year 2011 mean of 9.0 years.
  - the proportion of defective products is at most 3%, as promised by the manufacturer.
- One of the most commonly used methods for making such decisions or judgments is to perform a hypothesis test.
- A hypothesis is a statement that something is true.

Hypothesis testing

- a hypothesis test involves two hypotheses: the null hypothesis and the alternative hypothesis.
- $H_0 =$ hypothesis (the null hypothesis) ,A hypothesis to be tested
- $H_A =$ alternative (the alternative hypothesis). A hypothesis to be considered as an alternative to the null hypothesis
- H_0 and H_A are simply two mutually exclusive statements.
- Hypothesis test: The problem in a hypothesis test is to decide whether the null hypothesis should be rejected in favor of the alternative hypothesis.
- Each test results either in acceptance of H_0 or its rejection in favor of H_A.

Hypothesis testing

- For Example, in the Ponam packaging of rice, the null hypothesis might be "the mean weight of all bags of packaged equals the advertised weight of 994 g,"
- the alternative hypothesis might be "the mean weight of all bags of Ponam rice packaged differs from the advertised weight of 994 g."
- The null hypothesis for a hypothesis test concerning a population mean, μ, always specifies a single value for that parameter.

$$H_0 : \mu = \mu_0,$$

where μ_0 is some number.

Alternative Hypothesis

- The choice of the alternative hypothesis depends on the hypothesis test.
- Three choices are possible for the alternative hypothesis.
  - If the primary concern is deciding whether a population mean, μ, is different from a specified value μ_0, we express the alternative hypothesis as

    $$H_a : \mu \neq \mu_0.$$

    A hypothesis test with this alternative hypothesis is a two-tailed test.
  - If the primary concern is deciding whether a population mean,μ , is less than a specified value μ_0, we express the alternative hypothesis as

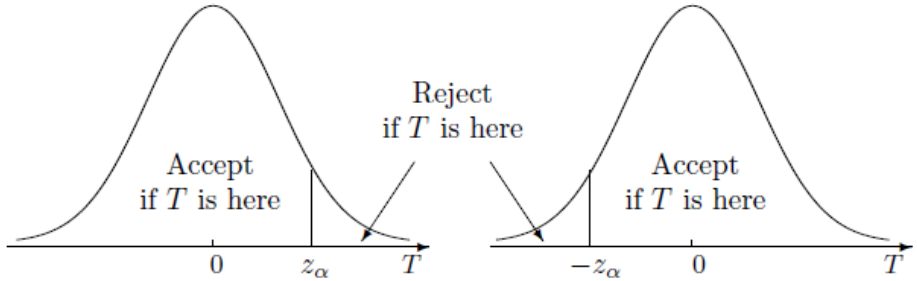
    $$H_a : \mu < \mu_0.$$

    A hypothesis test with this alternative hypothesis is left-tailed test.
  - If the primary concern is deciding whether a population mean, μ, is greater than a specified value μ_0, we express the alternative hypothesis as

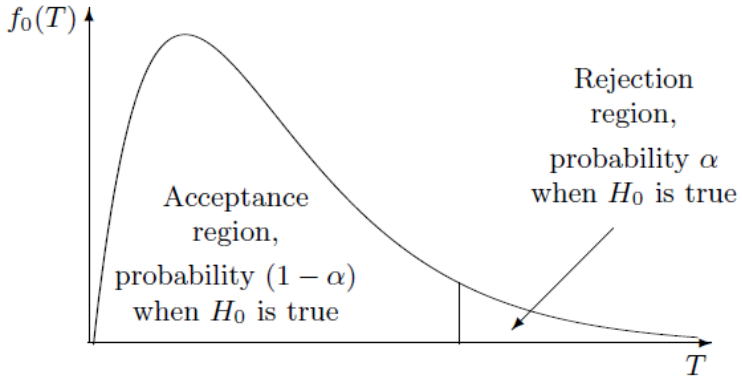
    $$H_a : \mu > \mu_0.$$

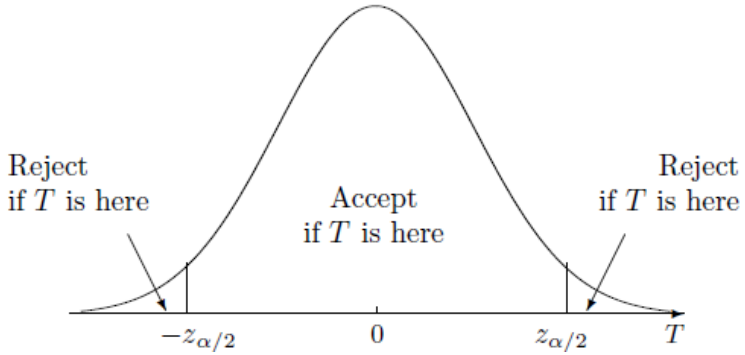
    A hypothesis test with this alternative hypothesis is right-tailed test.
- A hypothesis test is called a one-tailed test if it is either left tailed or right tailed.

Acceptance region and rejection region



(a) *Right-tail Z-test* (b) *Left-tail Z-test*





Hypothesis testing

- Poverty and Dietary Calcium Calcium is the most abundant mineral in the human body and has several important functions. Most body calcium is stored in the bones and teeth, where it functions to support their structure. Recommendations for calcium are provided in Dietary Reference Intakes, developed by the Institute of Medicine of the National Academy of Sciences. The recommended adequate intake (RAI) of calcium for adults (ages 19–50 years) is 1000 milligrams (mg) per day. Suppose that we want to perform a hypothesis test to decide whether the average adult with an income below the poverty level gets less than the RAI of 1000 mg.
  - Determine the null hypothesis for the hypothesis test.
  - Determine the alternative hypothesis for the hypothesis test.
  - Classify the hypothesis test as two tailed, left tailed, or right tailed.

Hypothesis testing

- A half-century ago, the average (U.S.) woman in her 20s was 62.6 inches tall. Suppose that we want to perform a hypothesis test to decide whether today's women in their 20s are, on average, taller than such women were a half-century ago.
  - Determine the null hypothesis for the hypothesis test.
  - Determine the alternative hypothesis for the hypothesis test.
  - Classify the hypothesis test as two tailed, left tailed, or right tailed.

Basic Logic of Hypothesis Testing

- Take a random sample from the population.
- If the sample data are consistent with the null hypothesis, do not reject the null hypothesis;
- if the sample data are inconsistent with the null hypothesis and supportive of the alternative hypothesis, reject the null hypothesis in favor of the alternative hypothesis.
- Type I and Type II Errors
  - Type I error: Rejecting the null hypothesis when it is in fact true.
  - Type II error: Not rejecting the null hypothesis when it is in fact false.

Decision:	H_0 is: True	H_0 is: False
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

Example

Consider again the Ponam Rice -packaging hypothesis test. The null and alternative hypotheses are, respectively,
$H_0 : \mu = 994g$ (the packaging machine is working properly)
$H_a : \mu = 994g$ (the packaging machine is not working properly),
where μ is the mean net weight of all bags of Ponam Rice s packaged.
Explain what each of the following would mean.
a. Type I error b. Type II error c. Correct decision
Now suppose that the results of carrying out the hypothesis test lead to rejection of the null hypothesis $\mu = 994g$, that is, to the conclusion that $\mu \neq 994g$. Classify that conclusion by error type or as a correct decision if
d. the mean net weight, μ, is in fact 994 g.
e. the mean net weight, μ, is in fact not 994 g.

Possible Conclusions for a Hypothesis Test

- Suppose that a hypothesis test is conducted at a small significance level.
  - If the null hypothesis is rejected, we conclude that the data provide sufficient evidence to support the alternative hypothesis.
  - If the null hypothesis is not rejected, we conclude that the data do not provide sufficient evidence to support the alternative hypothesis.
- When the null hypothesis is rejected in a hypothesis test performed at the significance level α, "the test results are statistically significant at the α level."
- Similarly, when the null hypothesis is not rejected in a hypothesis test performed at the significance level α, "the test results are not statistically significant at the α level."

Hypothesis testing

Faculty Salaries The American Association of University Professors (AAUP) conducts salary studies of college professors and publishes its findings in AAUP Annual Report on the Economic Status of the Profession. Suppose that we want to decide whether the mean salaries of college faculty in private and public institutions are different.

a. Pose the problem as a hypothesis test.

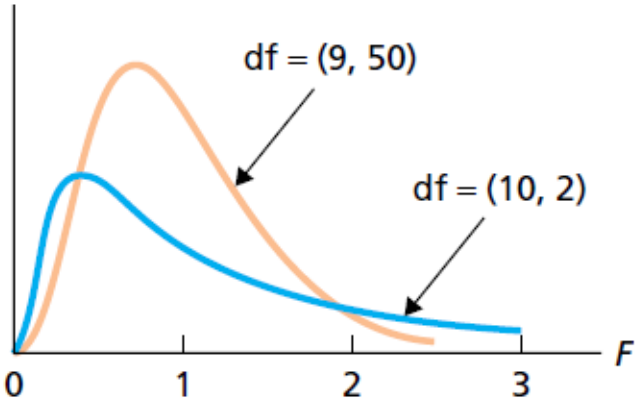
b. Explain the basic idea for carrying out the hypothesis test.

c. Suppose that 35 faculty members from private institutions and 30 faculty members from public institutions are randomly and independently selected and that their salaries are as shown in Table below, in thousands of dollars rounded to the nearest hundred. Discuss the use of these data to make a decision concerning the hypothesis test.

Sample 1 (private institutions)							Sample 2 (public institutions)					
97.3	85.9	118.8	93.9	66.6	109.2	64.9	59.9	115.7	126.1	50.3	133.1	89.3
83.1	100.6	99.3	94.9	94.4	139.3	108.8	82.5	67.1	60.7	79.9	50.1	81.7
158.1	142.4	85.0	108.2	116.3	141.5	51.4	83.9	102.5	109.9	105.1	67.9	107.5
125.6	70.6	74.6	69.9	115.4	84.6	92.0	54.9	41.5	59.5	65.9	76.9	66.9
97.2	55.1	126.6	116.7	76.0	109.6	63.0	85.9	113.9	70.3	90.1	99.7	96.7

F-distribution

- Analysis-of-variance procedures rely on a distribution called the F-distribution, named in honor of Sir Ronald Fisher.
- A variable is said to have an F-distribution if its distribution has the shape of a special type of right-skewed curve, called an F-curve.
- There are infinitely many F-distributions, and we identify an F-distribution (and F-curve) by its number of degrees of freedom, just as we do for t-distributions and chi-square distributions.
- An F-distribution, however, has two numbers of degrees of freedom instead of one.
- The first number of degrees of freedom for an F-curve is called the degrees of freedom for the numerator, and the second is called the degrees of freedom for the denominator.



F-distribution

Basic Properties of F-Curves
Property 1: The total area under an F-curve equals 1.
Property 2: An F-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
Property 3: An F-curve is right skewed.
Percentages (and probabilities) for a variable having an F-distribution are equal to areas under its associated F-curve.
To perform an ANOVA test, we need to know how to find the F-value having a specified area to its right. The symbol F_α denotes the F-value having area α to its right.

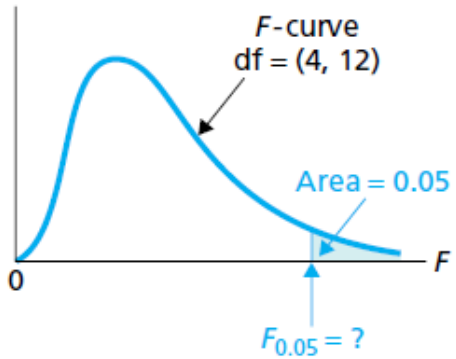
F-distribution

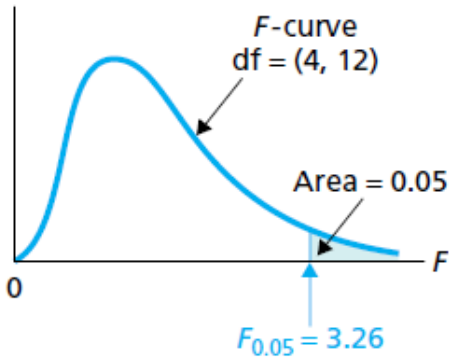
dfd	α	dfn 1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	α	dfd
	0.10	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	0.10	
	0.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	0.05	
1	0.025	647.79	799.50	864.16	899.58	921.85	937.11	948.22	956.66	963.28	968.63	976.71	984.87	993.10	997.25	1001.41	1005.60	1009.80	1014.02	0.025	1
	0.01	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5	6055.8	6106.3	6157.3	6208.7	6234.6	6260.6	6286.7	631.9	6339.4	0.01	
	0.005	16211	20000	21615	22500	23056	23437	23715	23925	24091	24224	24426	24630	24836	24940	25044	25148	25253	25359	0.005	
	0.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	0.10	
	0.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	0.05	
2	0.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	0.025	2
	0.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	0.01	
	0.005	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40	199.42	199.43	199.45	199.46	199.47	199.47	199.48	199.49	0.005	
	0.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	0.10	
	0.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	0.05	
3	0.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	0.025	3
	0.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	0.01	
	0.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	0.005	
	0.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	0.10	
	0.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	0.05	
4	0.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	0.025	4
	0.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	0.01	
	0.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	0.005	
	0.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	0.10	
	0.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	0.05	
5	0.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	0.025	5
	0.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	0.01	
	0.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	0.005	
	0.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	0.10	
	0.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	0.05	
6	0.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	0.025	6
	0.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	0.01	
	0.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	0.005	
	0.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	0.10	
	0.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	0.05	
7	0.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.41	4.36	4.31	4.25	4.20	0.025	7
	0.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	0.01	
	0.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	8.18	7.97	7.75	7.64	7.53	7.42	7.31	7.19	0.005	
	0.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	0.10	
	0.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	0.05	
8	0.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	0.025	8
	0.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	0.01	
	0.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	0.005	

F-distribution

Finding the *F*-Value Having a Specified Area to Its Right

For an F-curve with df = (4, 12), find $F_{0.05}$; that is, find the F-value having area 0.05 to its right, as shown in Fig. 16.2(a).





		dfn								
dfd	α	1	2	3	4	5	6	7	8	9
	0.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	0.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
12	0.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	0.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	0.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20

Analysis of variance (ANOVA)

- Analysis of variance (ANOVA) provides methods for comparing several population means, that is, the means of a single variable for several populations.
- The simplest kind of ANOVA, one-way analysis of variance.
- This type of ANOVA is called one-way analysis of variance because it compares the means of a variable for populations that result from a classification by one other variable, called the factor.
- The possible values of the factor are referred to as the levels of the factor.

For example, suppose that you want to compare the mean energy consumption by households among the four regions of the Karachi. The variable under consideration is "energy consumption," and there are four populations: households in the four regions. The four populations result from classifying households in the Karachi by the factor "region," whose levels are East, Central, South, and West.

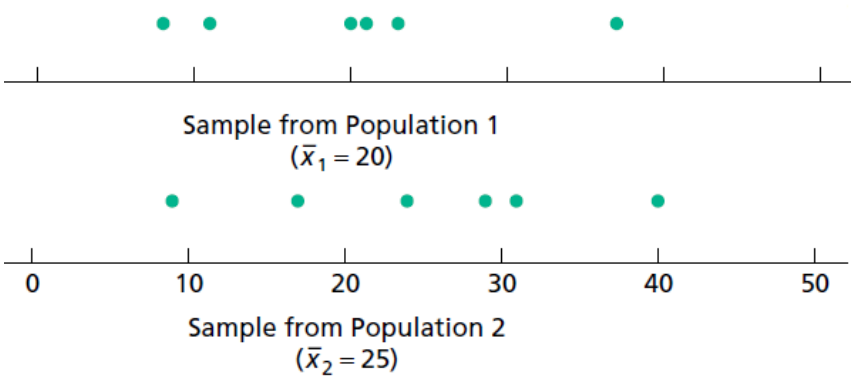
Conditions for One-Way ANOVA

Assumptions for One-Way ANOVA

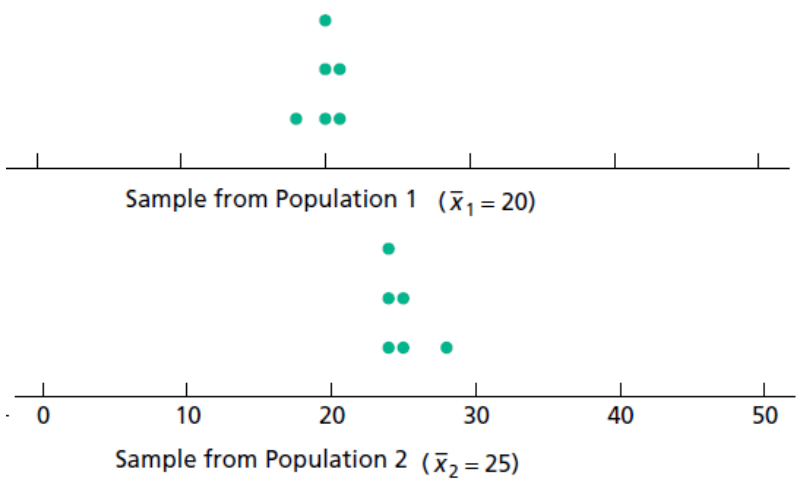
1. Simple random samples: The samples taken from the populations under consideration are simple random samples.
2. Independent samples: The samples taken from the populations under consideration are independent of one another.
3. Normal populations: For each population, the variable under consideration is normally distributed.
4. Equal standard deviations: The standard deviations of the variable under consideration are the same for all the populations.

Analysis of variance (ANOVA)

Sample from Population 1	21	37	11	20	8	23
Sample from Population 2	24	31	29	40	9	17



Sample from Population 1	21	21	20	18	20	20
Sample from Population 2	25	28	25	24	24	24



Analysis of variance (ANOVA)

- We will assume that there there are k populations under consideration and the ANOVA requirements are satisfied.
- Let $n_1, \ldots, n_k$ denote the number of elements in the samples from each of the k populations.
- Let $x_1, \ldots, x_k$denote the sample means for each of the k populations.
- Let $s_1, \ldots, s_k$ denote the sample standard deviations for each of the k populations

Measuring the variation amongst the means

- First calculate $\overline{x}$, the sample mean for the entire sample.
- Next we calculate the treatment sum of squares:

$$SSTR = n_1(\overline{x}_1 - \overline{x})^2 + \ldots + n_k(\overline{x}_k - \overline{x})^2$$

. SSTR is a measure of the variation between the sample means.

- From this we can now calculate the mean square treatment:

$$MSTR = \frac{SSTR}{k-1}$$

Measuring variation within the samples

- First we calculate the error sum of squares:

$$SSE = (n_1 - 1)s_1^2 + \ldots + (n_k - 1)s_k^2.$$

- Next we calculate the mean square error:

$$MSE = \frac{SSE}{n - k}$$

- The MSE measures the variation within the entire sample.

F-Statistics

- The F-statistic is the ratio of the variation among the sample means to the variation among the entire sample; thus,

$$F = \frac{MSTR}{MSE}$$

- If F is large, then we see that the variation amongst the means is greater than the variation within the sample itself. This is strong evidence against the claim that the the population means are equal.

F-Statistics

the summary of the singer height's data set:

$$\begin{pmatrix} part & mean & sd & n \\ Alto & 64.89 & 2.79 & 35 \\ Bass & 70.72 & 2.36 & 39 \\ Soprano & 64.25 & 1.87 & 36 \\ Tenor & 69.15 & 3.21 & 20 \end{pmatrix}$$

Compute the F -statistic for this set
The sample mean for the entire population is x =
SSTR =
MSTR =
SSE=
MSE=
F=

F-Statistics

The sample mean for the entire population is $\bar{x} = 67.12$.

Thus

$$SSTR = 35(64.89 - 67.12)^2 + 39(70.72 - 67.12)^2 + 36(64.25 - 67.12)^2 + 20(69.15 - 67.12)^2 = 1058.4379$$

Therefore

$$MSTR = \frac{1058.4379}{3} = 352.813$$

$$SSE = (34)(2.79)^2 + (38)(2.36)^2 + (35)(1.87)^2 + (19)(3.21)^2 = 794.47$$

Therefore

$$MSE = \frac{794.47}{126} = 6.31$$

$$F = \frac{352.813}{6.31} = 55.91$$

F-Statistics

Northeast	Midwest	South	West
13	15	5	8
8	10	11	10
11	16	9	6
12	11	5	5
11	13		7
	10		
11.0	12.5	7.5	7.2

← *Means*

SSTR and SSE calculation

Region	Size n_j	Mean $\bar{x}_j$	$\bar{x}_j - \bar{x}$	$(\bar{x}_j - \bar{x})^2$	$n_j(\bar{x}_j - \bar{x})^2$
Northeast	5	11.0	1.2	1.44	7.20
Midwest	6	12.5	2.7	7.29	43.74
South	4	7.5	−2.3	5.29	21.16
West	5	7.2	−2.6	6.76	33.80
					105.90

Region	Size n_j	Variance s_j^2	$n_j - 1$	$(n_j - 1)s_j^2$
Northeast	5	3.5	4	14.0
Midwest	6	6.7	5	33.5
South	4	9.0	3	27.0
West	5	3.7	4	14.8
				89.3

One-Way ANOVA Identity

The total sum of squares equals the treatment sum of squares plus the error
sum of squares: $SST = SSTR + SSE$.
This last identity, called the ANOVA Identity, is very important:

$$\underbrace{SST}_{\text{total variation}} = \underbrace{SSTR}_{\text{treatment variation}} + \underbrace{SSE}_{\text{sample's variation}}$$

In order to compute the F-statistic, we need $SSTR$ and SSE. This identity shows us that we can compute SST and $SSTR$ (for example) and then find SSE by

$$SSE = SST - SSTR.$$

Analysis of variance (ANOVA)

Source	df	SS	$MS = SS/\text{df}$	F-statistic
Treatment	$k-1$	$SSTR$	$MSTR = \frac{SSTR}{k-1}$	$F = \frac{MSTR}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	SST		

Source	df	SS	$MS = SS/\text{df}$	F-statistic
Treatment	3	105.9	35.3	6.32
Error	16	89.3	5.581	
Total	19	195.2		

Purpose To perform a hypothesis test to compare k population means, $\mu_1, \mu_2, \ldots, \mu_k$

Assumptions

1. Simple random samples
2. Independent samples
3. Normal populations
4. Equal population standard deviations

Step 1 The null and alternative hypotheses are, respectively,

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

H_a: Not all the means are equal.

Step 2 Decide on the significance level, α.

Step 3 Compute the value of the test statistic

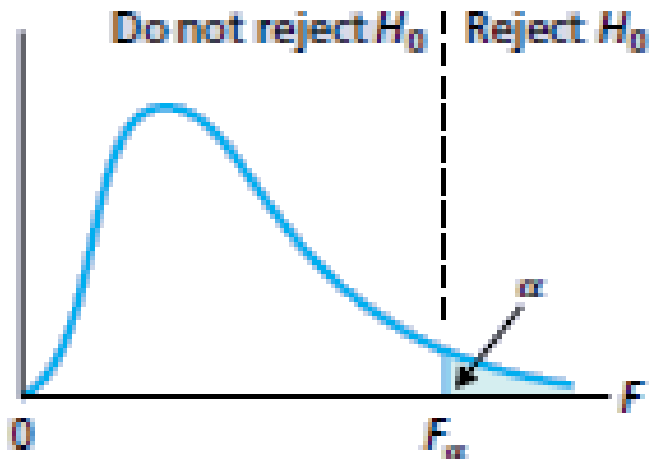
$$F = \frac{MSTR}{MSE}$$

and denote that value F_0. To do so, construct a one-way ANOVA table:

Source	df	SS	$MS = SS/\text{df}$	F-statistic
Treatment	$k-1$	$SSTR$	$MSTR = \frac{SSTR}{k-1}$	$F = \frac{MSTR}{MSE}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$	
Total	$n-1$	SST		

Analysis of variance (ANOVA)

Step 4 The critical value is F_α with df$=(k-1, n-k)$. Use Table VIII to find the critical value.



Step 5 If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.