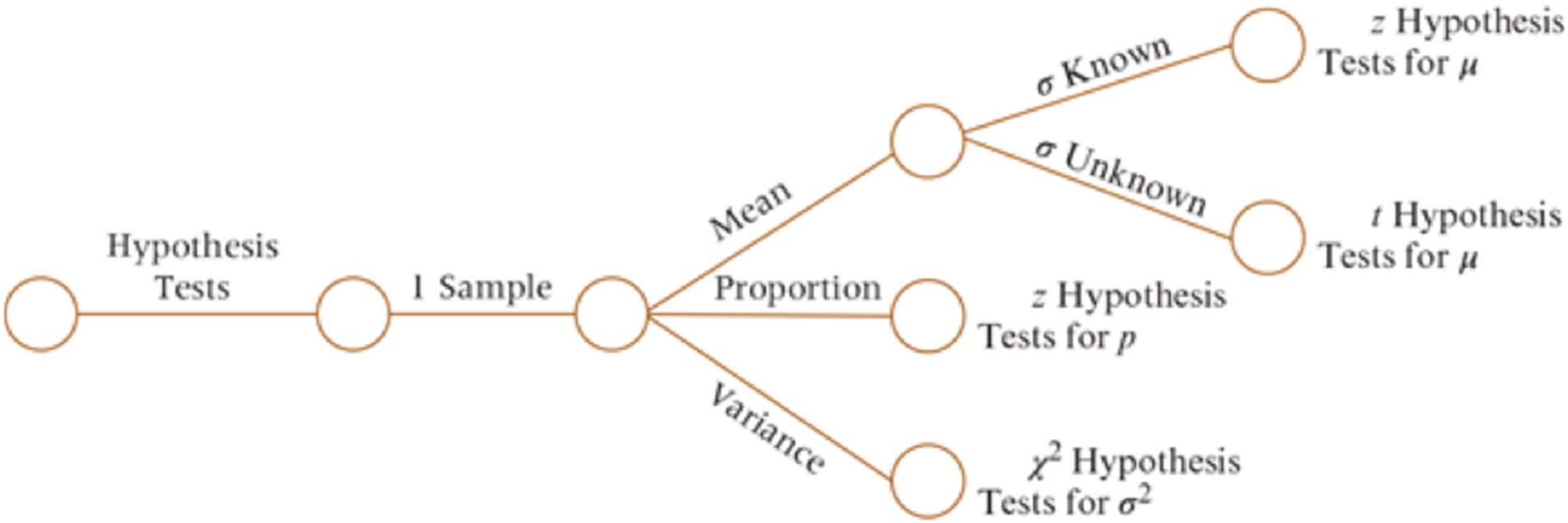
Comparing Two Populations

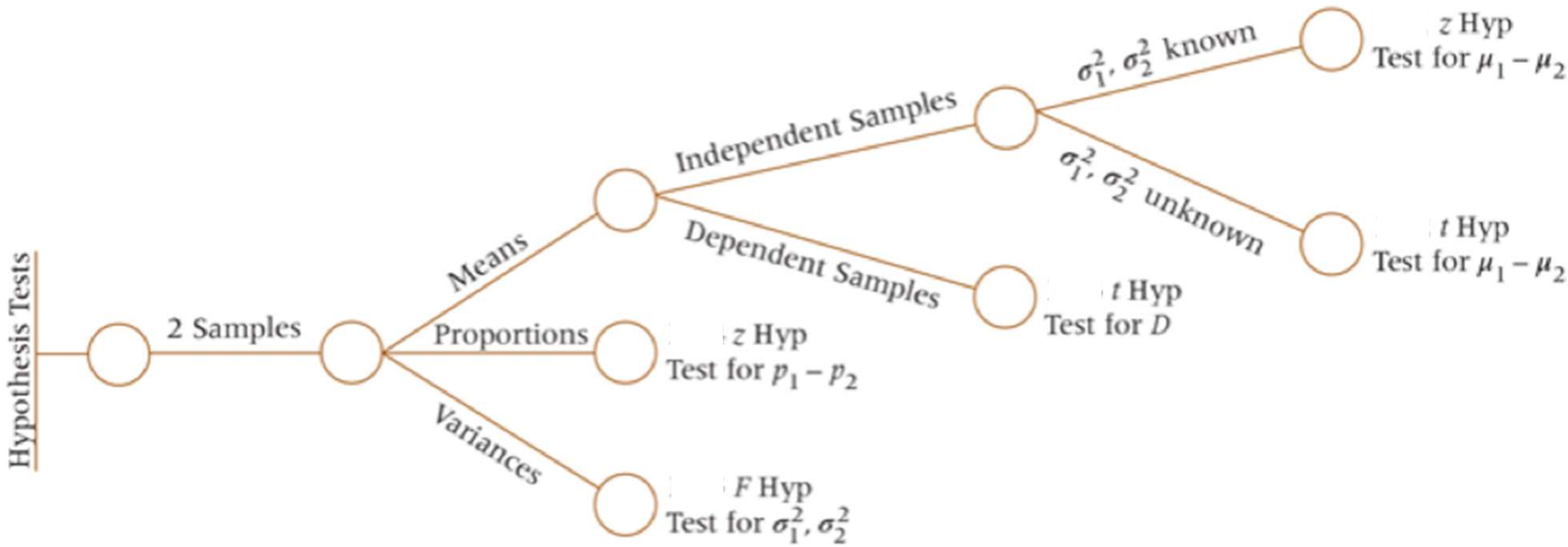
Sessions 9

Test Map



Hypothesis Tests
1 Sample
Mean
Proportion
Variance
σ Known
σ Unknown
z Hypothesis Tests for μ
t Hypothesis Tests for μ
z Hypothesis Tests for p
χ² Hypothesis Tests for σ²

Test Map



Hypothesis Tests
2 Samples
Means
Proportions
Variances
Independent Samples
Dependent Samples
σ_1^2, σ_2^2 known
σ_1^2, σ_2^2 unknown
z Hyp
Test for $\mu_1 - \mu_2$
t Hyp
Test for $\mu_1 - \mu_2$
t Hyp
Test for D
z Hyp
Test for $p_1 - p_2$
F Hyp
Test for σ_1^2, σ_2^2

Comparing Two Populations

A women's activist group wants to "prove" that women do not get paid as much per year as men. They hire a business analyst to test their theory. The business analyst takes a random sample of 87 professional working women and determines that the sample average annual salary per person is Rs.3,35,200. Assume that the standard deviation for this population is Rs.1,10,000. She also takes a random sample of 76 professional working men and determines that the sample average annual salary per person is Rs.5,72,700. Assume that the standard deviation for this population is Rs.1,70,000. Using these data, α = 0.001, test the hypothesis that women do get paid less per year than men.

Are the two samples coming from the same population or different populations?

Each sample $\bar{X}$ is distributed around its population mean μ (Central Limit Theorem)

If the two samples are from the same population, there will not be statistically significant difference between their means ie. $\mu_1 = \mu_2$

Thus,

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

Comparing Two Populations

σ_1 = 110000; σ_2 = 170000

$\bar{X}_1$ = 335200; $\bar{X}_2$ = 572700

n_1 = 87; n_2 = 76

α = 0.001

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 < 0$$

Since σ is known → Use Normal Distribution

$$z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$z = \frac{(335200 - 572700) - (0)}{\sqrt{\frac{110000^2}{87} + \frac{170000^2}{76}}} = -10.42$$

Comparing Two Populations

$\alpha = 0.001$

$H_0: \mu_1 - \mu_2 = 0$

$H_1: \mu_1 - \mu_2 < 0$

$$z = \frac{(335200 - 572700) - (0)}{\sqrt{\frac{110000^2}{87} + \frac{170000^2}{76}}} = -10.42$$

Critical Z value = - 3.09 → Z Value calculated from the data is to the left of the critical Z value → Reject H_0

$P(Z \leq -10.42) < \alpha$ → Reject H_0

Conclusion: Annual Salary paid to women is less than that paid to men. The difference is statistically significant.

Comparing Two Populations

Critical Value Method:

Critical Z value = - 3.09

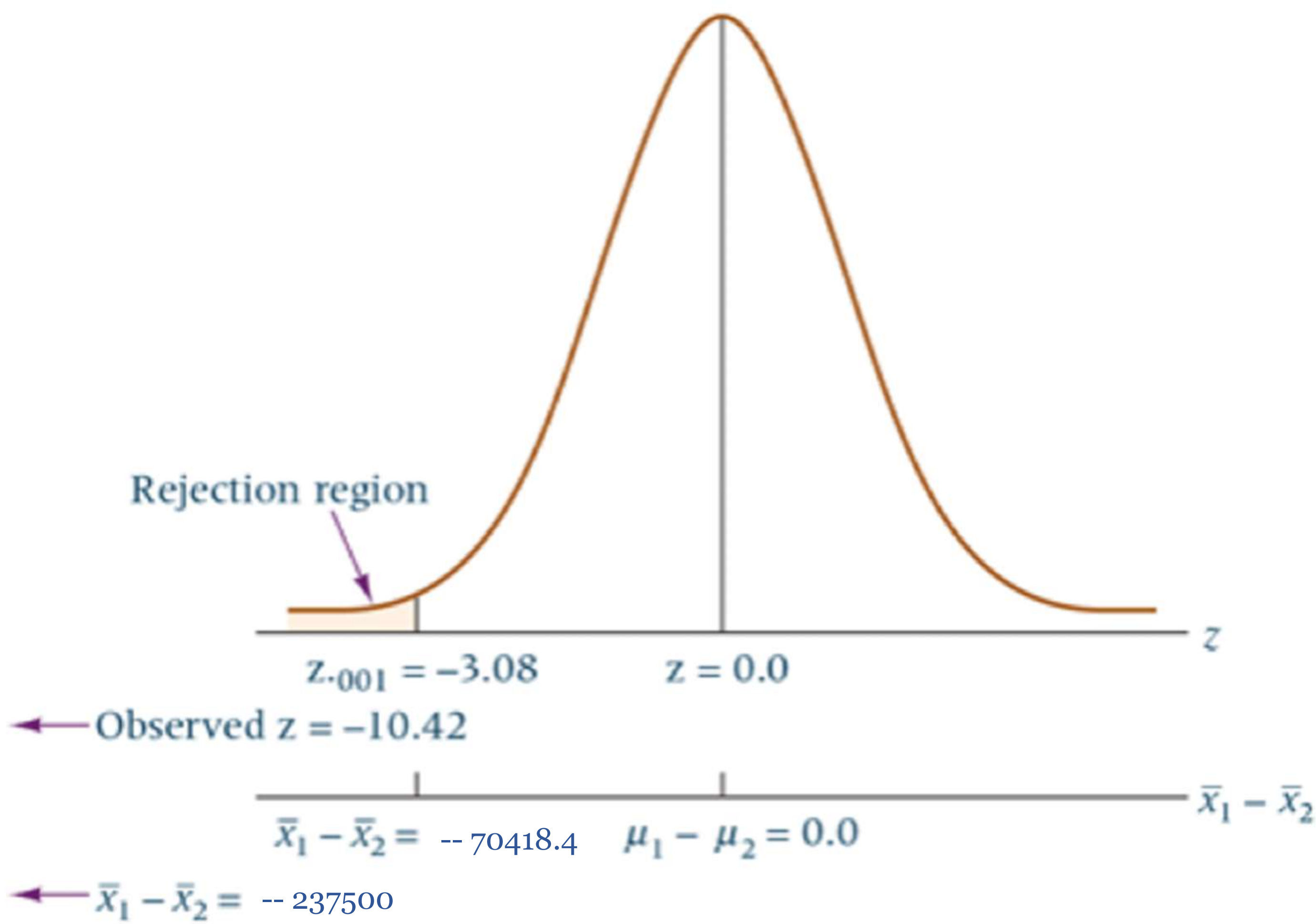
For this critical value the critical value for X1 – X2 can be calculated as:

$$(\overline{X_1} - \overline{X_2}) = 0 - 3.09\sqrt{\frac{110000^2}{87} + \frac{170000^2}{76}} = -70418.4$$

As per data: $\bar{X}_1 - \bar{X}_2$ = -237500 which is to the left of -70418.4 → Reject Null

Can also be solved constructing a confidence interval around $\mu_1 - \mu_2$ = 0

Comparing Two Populations



Population σ unknown

A coffee manufacturer is interested in estimating the difference in the average daily coffee consumption of regular-coffee drinkers and decaffeinated-coffee drinkers. Its analyst randomly selects 13 regular-coffee drinkers and asks how many cups of coffee per day they drink. He randomly locates 15 decaffeinated-coffee drinkers and asks how many cups of coffee per day they drink. The average for the regular-coffee drinkers is 4.35 cups, with a standard deviation of 1.20 cups. The average for the decaffeinated-coffee drinkers is 6.84 cups, with a standard deviation of 1.42 cups. The analyst assumes, for each population, that the daily consumption is normally distributed, and he constructs a 95% confidence interval to estimate the difference in the averages of the two populations.

$\bar{X}_1$ = 4.35; $\bar{X}_2$ = 6.84; S_1 = 1.20; S_2 = 1.42

n_1 = 13; n_2 = 15

α = 0.05

σ unknown;

Population is normally distributed

Use t-distribution

Population σ unknown

$\bar{X}_1 = 4.35$; $\bar{X}_2 = 6.84$; $S_1 = 1.20$; $S_2 = 1.42$

$n_1 = 13$; $n_2 = 15$

$\alpha = 0.05$

If σ_1 and σ_2 are equal:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$\text{df} = n_1 + n_2 - 2$$

If σ_1 and σ_2 are not equal:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\text{df} = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Population σ unknown

$\bar{X}_1 = 4.35$; $\bar{X}_2 = 6.84$; $S_1 = 1.20$; $S_2 = 1.42$

$n_1 = 13$; $n_2 = 15$

$\alpha = 0.05$

Using Confidence Intervals:

$$(\bar{x}_1 - \bar{x}_2) - t\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le$$

$$(\bar{x}_1 - \bar{x}_2) - t\sqrt{\frac{s_1^2(n_1-1)+s_2^2(n_2-1)}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\mathrm{df} = n_1 + n_2 - 2$$

For s known: df = 26, t0.025, 26 = 2.056

Confidence Interval: $-3.52 < \mu_1 - \mu_2 < -1.46$

The researcher is 95% confident that the difference in population average daily consumption of cups of coffee between regular- and decaffeinated-coffee drinkers is between 1.46 cups and 3.52 cups. The point estimate for the difference in population means is 2.49 cups, with a margin of error of 1.03 cups.

Population σ unknown

At the Haribhai Manufacturing Company, new employees are expected to attend a three-day seminar to learn about the company. At the end of the seminar, they are tested to measure their knowledge about the company. Management decided to experiment with a different training procedure of training new employees in two days by using pre-recorded training videos and having no question-and-answer session. If this procedure works, it could save the company lakhs of rupees over a period of several years. However, there is some concern about the effectiveness of the two-day method, and company managers would like to know whether there is any difference in the effectiveness of the two training methods.

To test the difference in the two methods, the managers randomly select one group of 15 newly hired employees to take the three-day seminar (method A) and a second group of 12 new employees for the two-day pre-recorded videos method (method B). **Table below** shows the test scores of the two groups. Using $\alpha = 0.05$, the managers want to determine whether there is a significant difference in the mean scores of the two groups. They assume that the scores for this test are normally distributed and that the population variances are approximately equal.

Population σ unknown

At the Haribhai Manufacturing Company…

Method A					Method B			
56	50	52	44	52	59	54	55	65
47	42	47	51	53	52	57	64	53
42	45	43	48	44	53	56	53	57

t-Test: Two-Sample Assuming Equal Variances

	Variable 1	*Variable 2*
Mean	47.7333333	56.5
Variance	19.4952381	18.2727273
Observations	15	12
Pooled Variance	18.9573333	
Hypothesized Mean Difference	0	
df	25	
t Stat	-5.1987657	
P(T<=t) one-tail	1.1152E-05	
t Critical one-tail	1.70814076	
P(T<=t) two-tail	2.2303E-05	
t Critical two-tail	2.05953855	

Population σ unknown

When the two populations are dependent:

Examples:

- Before and after studies
- Survey on siblings, spouses, etc.

$$t = \frac{\bar{d} - D}{\frac{s_d}{\sqrt{n}}}$$

$$\text{df} = n - 1$$

n = number of pairs

d = sample difference

$\bar{d}$ = sample mean difference

D = population mean difference

s_d = standard deviation of sample difference

$$\bar{d} = \frac{\sum d}{n}$$

$$s_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}} = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

Paired Sample t-test

A stock market investor is interested in determining whether there is a significant difference in the P/E (price to earnings) ratio for companies from one year to the next. In an effort to study this question, the investor randomly samples nine companies from the *BSE* and records the P/E ratios for each of these companies at the end of year 1 and at the end of year 2.

Company	Year 1 P/E	Year 2 P/E
1	8.9	12.7
2	38.1	45.4
3	43.0	10.0
4	34.0	27.2
5	34.5	22.8
6	15.2	24.1
7	20.3	32.3
8	19.9	40.1
9	61.9	106.5

t-Test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	30.6444444	35.6777778
Variance	268.135278	837.544444
Observations	9	9
Pearson Corr	0.6743572	
Hypothesized	0	
df	8	
t Stat	-0.6990947	
P(T<=t) one-	0.25215353	
t Critical one	1.85954804	
P(T<=t) two-	0.50430706	
t Critical two	2.30600414	

Comparing Population Proportion

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{(\bar{p} \cdot \bar{q}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where,

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

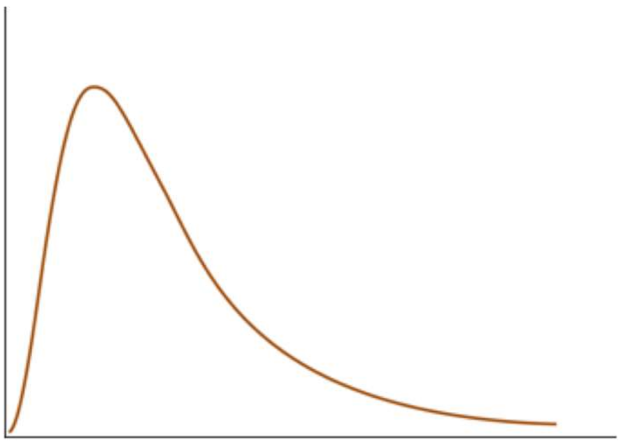
$$\bar{q} = 1 - \bar{p}$$

Comparing Population Variance

$$F = \frac{s_1^2}{s_2^2}$$

$$\mathrm{df}_{\mathrm{numerator}} = v_1 = n_1 - 1$$

$$\mathrm{df}_{\mathrm{denominator}} = v_2 = n_2 - 1$$



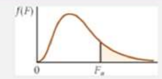
F distribution for $df_n = 6$ and $df_d = 30$

Excel Formula

=F.DIST.RT(F value, v_1, v_2)

=F.INV.RT(α, v_1, v_2)

Percentage Points of the *F* Distribution



$\alpha = .10$

Numerator Degrees of Freedom

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	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	62.5
2	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	9.4
3	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	5.1
4	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	3.8
5	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	3.1
6	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	2.7
7	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	2.5
8	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	2.3
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.2
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.1
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.0
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.9

$\alpha = .025$

Numerator Degrees of Freedom

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20
1	647.79	799.48	864.15	899.60	921.83	937.11	948.20	956.64	963.28	968.63	976.72	984.87	993.08
2	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
3	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
4	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
5	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	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
7	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
8	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
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67

$\alpha = .05$

Numerator Degrees of Freedom

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.90	245.90	248.00
2	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
3	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
4	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	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
6	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
7	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
8	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
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94