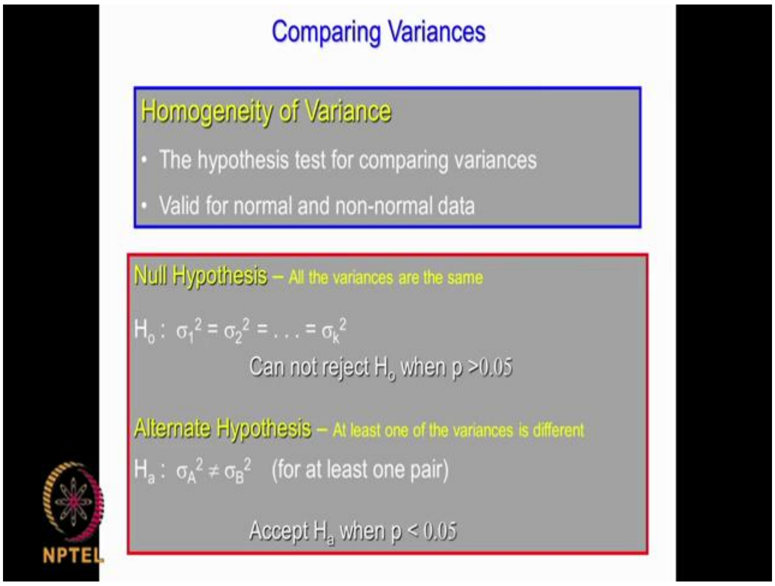
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 12
F- tests

Welcome to the course on Biostatistics and Design of Experiments. In this class, we will talk about f distribution, F- test. So far we talked about t distribution and comparing means, in the F-test we compare something called variances.

(Refer Slide Time: 00:27)



Comparing Variances

Homogeneity of Variance

- The hypothesis test for comparing variances
- Valid for normal and non-normal data

Null Hypothesis – All the variances are the same

$H_0: \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$

Can not reject H_0 when $p > 0.05$

Alternate Hypothesis – At least one of the variances is different

$H_a: \sigma_A^2 \neq \sigma_B^2$ (for at least one pair)

Accept H_a when $p < 0.05$

Variances are nothing but the spreads. Here in F-test or there is another test called analysis of variance, we are comparing variances this is way valid for normal and non-normal distribution data set also. So, when you are comparing variances the null hypothesis will be

$\sigma_1^2 = \sigma_2^2 =$ and so on, where σ s your standard deviation of the population so ---will be the variances. We cannot reject the null hypothesis if p is greater than 0.05 and the alternate hypothesis will be

$$\sigma_A^2 \neq \sigma_B^2$$

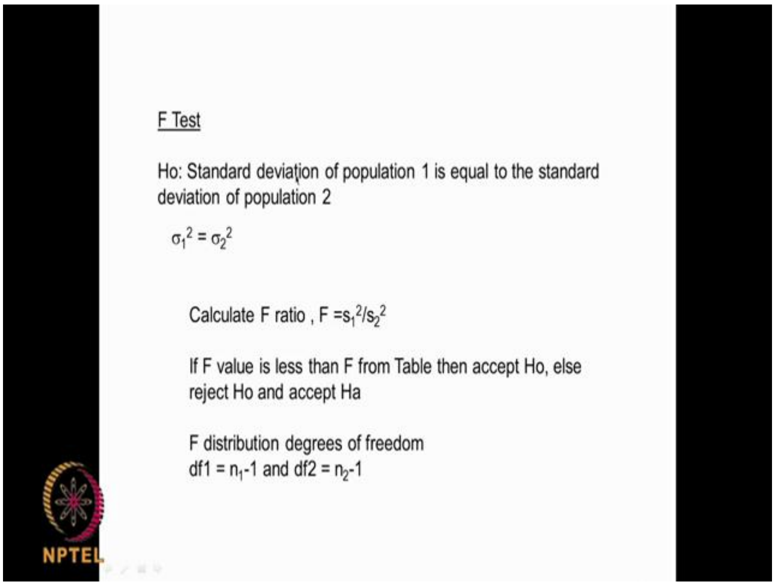
and so on. So here, when p is < 0.5 we will reject null hypothesis and we will accept the alternate hypothesis, so that is what a variance comparison is. The advantage of these variance comparison is it tells you about the spread of the data, that means what are the repeatability of the data. Suppose I take an instrument and I take a sample and then I do it 10 times I will get some variations or variance. Now is this variance too much or too little. If I use another instrument and I do the same sample 10 times, I will get again another set of variance, is these variances are very large or very small acceptable. So here the F-test is very, very useful. I am comparing 2 operators for example, I have a very well trained operator and I have a newly joined operator, I want to see what is the consistency of the results they produced. So I can test the quality of the newly joined operator with the respect to the operator who has been there for a very long time and I compare variances and I say yes, the variations that are the newly joined operator produces is comparable to the variations of the expert. If a newly joined operator comes in you give him a training and then we will give him a few samples and ask him to repeat and then we will compare it with the variations of the results from the expert and then we will conclude yes, this newly joined operator is well trained and the variations you produces are not very, very large it is comparable to the expert. In those situations for comparing instruments repeatability, for comparing the variations operators produced, comparing variations produced in one lab verses another lab, especially this is very, very useful when you are performing clinical trials so all these situations we use F-test also.

Suppose I for a two sample t test I should have only 2 samples, if I want to compare sample A with sample B, sample 1 with sample 2. Suppose I have 3 samples for example, if I had comparing 3 drugs or 2 drugs with the control or 1 drug with control under placebo. So I have more than 2 sets of samples. I might not be able to do 2 sample t test of course I can take 2 sets of sample at a time and do it t test. But if I want to go at 1 single shot comparing the data of these 3 sets what do I do, I look at the variances or the variations and there comes the something called F-test. So I can use the concept of analysis of variance and then I will say that

$$\sigma_1^2 = \sigma_2^2$$

that means σ_1^2 of the control is equal to square of the placebo is equal to of the drug that is the null hypothesis. Alternate hypothesis one of them could be different. In such situations again comparing variances are very very good, very very useful instead of using 2 sample t-test for a different combinations. What is F--test? So by H_0 here in F--test is standard deviation of population 1 = standard deviation of population 2 or actually, we should not use standard deviation, we should use variances because variances are additive, whereas standard deviations are not additive.

(Refer Slide Time: 04:43)



F Test

Ho: Standard deviation of population 1 is equal to the standard deviation of population 2

$\sigma_1^2 = \sigma_2^2$

Calculate F ratio , F $=s_1^2/s_2^2$

If F value is less than F from Table then accept Ho, else reject Ho and accept Ha

F distribution degrees of freedom
df1 = n_1-1 and df2 = n_2-1

Why? Because you take a square root of the variance to arrive at the standard deviation, so obviously, it is non-linear. So, non-linear cannot be additive whereas variances are additive. The null hypothesis in a F--test is

$\sigma_1^2 = \sigma_2^2$, obviously, the alternate hypothesis could be $\sigma_1^2 \neq \sigma_1^2$

. We calculate something called F ratio, where F could be

$F =s_1^2/s_2^2$

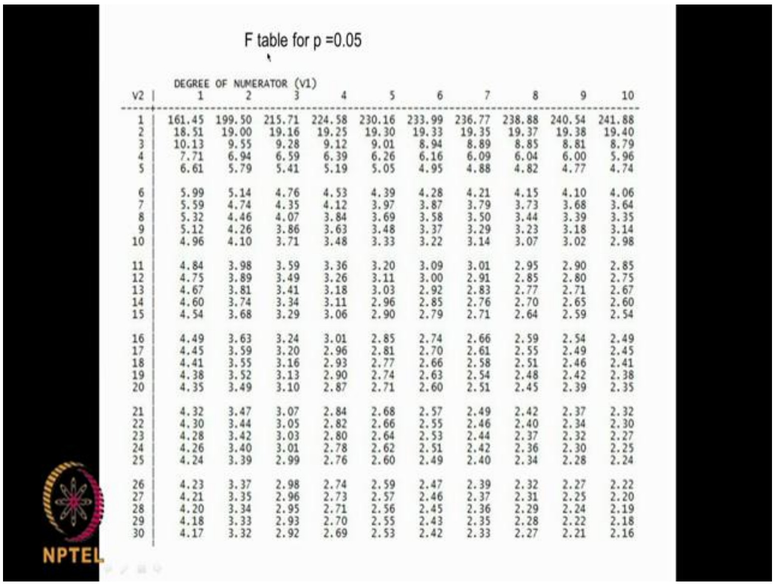
s_1^2 that is variance of sample 1 divided by the variance of sample 2 and then of course again we go to the table there is F table. There is F table which says, whether the calculated F value is greater than the table value or it is less than the table value. If the calculated F value is < the table value we will accept the H_0, if the calculated F value is > the table F value we will reject the H_0 and accept the alternate hypothesis. Now there will be 2 degrees of freedom because s_1 will have 1 set of degrees of freedom, s_2 also will have it is own set of degrees of freedom.

This table has interestingly row wise and column wise data representing both the degrees of freedom. So, degrees of freedom for

n_1-1 and df2 = n_2-1

s_1 will be n_1-1, degrees of freedom for s_2 will be n_2-1. Let me show you the table here.

(Refer Slide Time: 06:15)

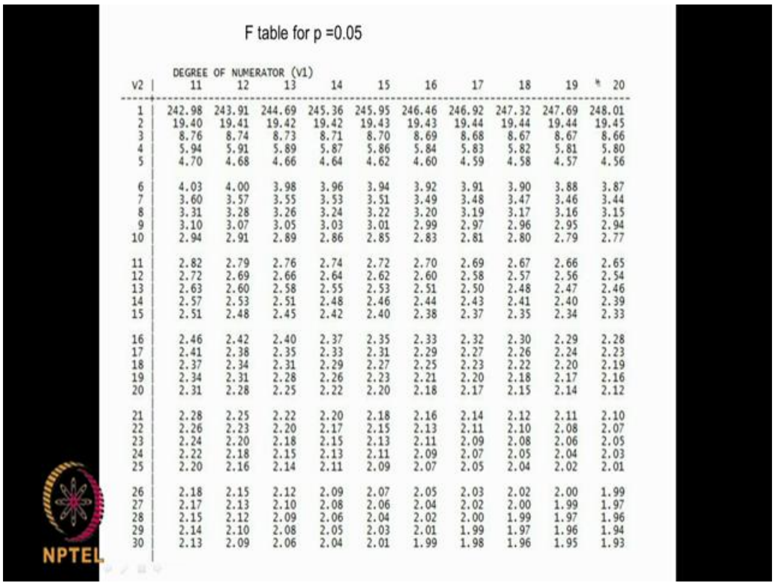


F table for p =0.05

	DEGREE OF NUMERATOR (V1)									
V2	1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16

You have the F table for p = 0.05 that means 95 % confidence interval, so degrees of freedom for the numerator, degrees of freedom for the denominator. So, if you want numerator more than 10 you go the next page here I am showing this right.

(Refer Slide Time: 06:32)



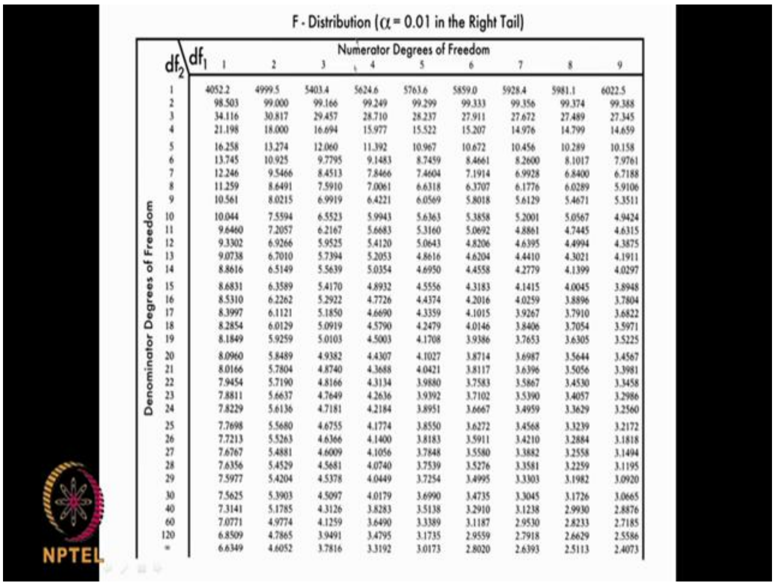
F table for p =0.05

DEGREE OF NUMERATOR (V1)

V2	11	12	13	14	15	16	17	18	19	20
1	242.98	243.91	244.69	245.36	245.95	246.46	246.92	247.32	247.69	248.01
2	19.40	19.41	19.42	19.42	19.43	19.43	19.44	19.44	19.44	19.45
3	8.76	8.74	8.73	8.71	8.70	8.69	8.68	8.67	8.67	8.66
4	5.94	5.91	5.89	5.87	5.86	5.84	5.83	5.82	5.81	5.80
5	4.70	4.68	4.66	4.64	4.62	4.60	4.59	4.58	4.57	4.56
6	4.03	4.00	3.98	3.96	3.94	3.92	3.91	3.90	3.88	3.87
7	3.60	3.57	3.55	3.53	3.51	3.49	3.48	3.47	3.46	3.44
8	3.31	3.28	3.26	3.24	3.22	3.20	3.19	3.17	3.16	3.15
9	3.10	3.07	3.05	3.03	3.01	2.99	2.97	2.96	2.95	2.94
10	2.94	2.91	2.89	2.86	2.85	2.83	2.81	2.80	2.79	2.77
11	2.82	2.79	2.76	2.74	2.72	2.70	2.69	2.67	2.66	2.65
12	2.72	2.69	2.66	2.64	2.62	2.60	2.58	2.57	2.56	2.54
13	2.63	2.60	2.58	2.55	2.53	2.51	2.50	2.48	2.47	2.46
14	2.57	2.53	2.51	2.48	2.46	2.44	2.43	2.41	2.40	2.39
15	2.51	2.48	2.45	2.42	2.40	2.38	2.37	2.35	2.34	2.33
16	2.46	2.42	2.40	2.37	2.35	2.33	2.32	2.30	2.29	2.28
17	2.41	2.38	2.35	2.33	2.31	2.29	2.27	2.26	2.24	2.23
18	2.37	2.34	2.31	2.29	2.27	2.25	2.23	2.22	2.20	2.19
19	2.34	2.31	2.28	2.26	2.23	2.21	2.20	2.18	2.17	2.16
20	2.31	2.28	2.25	2.22	2.20	2.18	2.17	2.15	2.14	2.12
21	2.28	2.25	2.22	2.20	2.18	2.16	2.14	2.12	2.11	2.10
22	2.26	2.23	2.20	2.17	2.15	2.13	2.11	2.10	2.08	2.07
23	2.24	2.20	2.18	2.15	2.13	2.11	2.09	2.08	2.06	2.05
24	2.22	2.18	2.15	2.13	2.11	2.09	2.07	2.05	2.04	2.03
25	2.20	2.16	2.14	2.11	2.09	2.07	2.05	2.04	2.02	2.01
26	2.18	2.15	2.12	2.09	2.07	2.05	2.03	2.02	2.00	1.99
27	2.17	2.13	2.10	2.08	2.06	2.04	2.02	2.00	1.99	1.97
28	2.15	2.12	2.09	2.06	2.04	2.02	2.00	1.99	1.97	1.96
29	2.14	2.10	2.08	2.05	2.03	2.01	1.99	1.97	1.96	1.94
30	2.13	2.09	2.06	2.04	2.01	1.99	1.98	1.96	1.95	1.93

It goes up to 20, so degrees of freedom for the denominator is here goes up to 30. So this is for p = 0.05.

(Refer Slide Time: 06:43)



F - Distribution (α = 0.01 in the Right Tail)

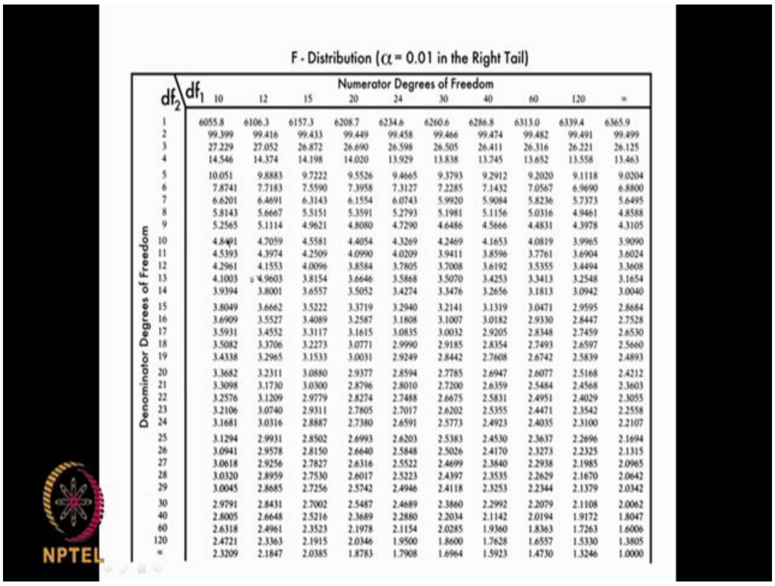
Numerator Degrees of Freedom

Denominator Degrees of Freedom

df_2 \ df_1	1	2	3	4	5	6	7	8	9
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920
30	7.5625	5.3903	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7185
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586
∞	6.6349	4.6052	3.7816	3.3192	3.0173	2.8020	2.6393	2.5113	2.4073

Similarly, we can have for 0.01 also the degrees of freedom for the numerator, degrees of freedom for the denominator going like this.

(Refer Slide Time: 06:51)



F - Distribution ($\alpha = 0.01$ in the Right Tail)

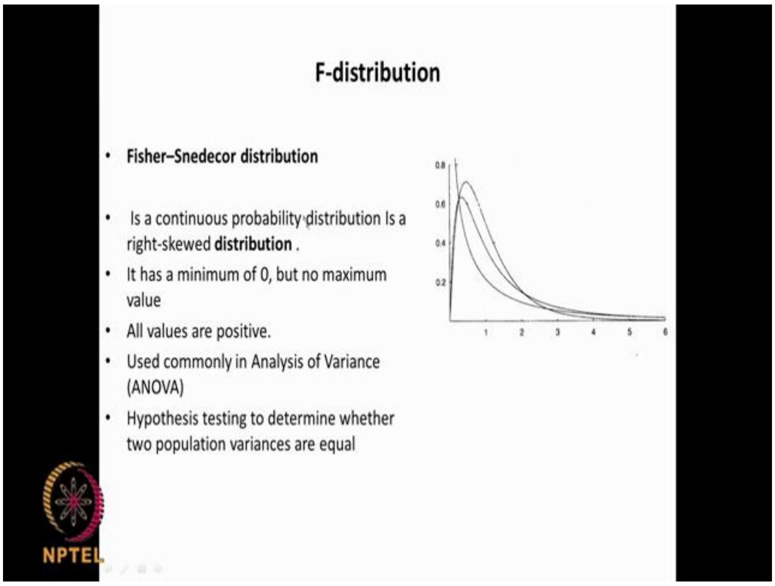
df_2 \ df_1	10	12	15	20	24	30	40	60	120	∞
	Numerator Degrees of Freedom									
1	6055.8	6106.3	6157.3	6208.7	6234.6	6260.6	6286.8	6313.0	6339.4	6365.9
2	99.399	99.416	99.433	99.449	99.458	99.466	99.474	99.482	99.491	99.499
3	27.229	27.052	26.872	26.690	26.598	26.505	26.411	26.316	26.221	26.125
4	14.546	14.374	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	10.051	9.8883	9.7222	9.5526	9.4665	9.3793	9.2912	9.2020	9.1118	9.0204
6	7.8741	7.7183	7.5590	7.3958	7.3127	7.2285	7.1432	7.0567	6.9690	6.8800
7	6.6201	6.4691	6.3143	6.1554	6.0743	5.9920	5.9084	5.8236	5.7373	5.6495
8	5.8143	5.6667	5.5151	5.3591	5.2793	5.1981	5.1156	5.0316	4.9461	4.8588
9	5.2565	5.1114	4.9621	4.8080	4.7290	4.6486	4.5666	4.4831	4.3978	4.3105
10	4.8491	4.7059	4.5581	4.4054	4.3269	4.2469	4.1653	4.0819	3.9965	3.9090
11	4.5393	4.3974	4.2509	4.0990	4.0209	3.9411	3.8596	3.7761	3.6904	3.6024
12	4.2961	4.1553	4.0096	3.8584	3.7805	3.7008	3.6192	3.5355	3.4494	3.3608
13	4.1003	3.9603	3.8154	3.6646	3.5868	3.5070	3.4253	3.3413	3.2548	3.1654
14	3.9394	3.8001	3.6557	3.5052	3.4274	3.3476	3.2656	3.1813	3.0942	3.0040
15	3.8049	3.6662	3.5222	3.3719	3.2940	3.2141	3.1319	3.0471	2.9595	2.8684
16	3.6909	3.5527	3.4089	3.2587	3.1808	3.1007	3.0182	2.9330	2.8447	2.7528
17	3.5931	3.4552	3.3117	3.1615	3.0835	3.0032	2.9205	2.8348	2.7459	2.6530
18	3.5082	3.3706	3.2273	3.0771	2.9990	2.9185	2.8354	2.7493	2.6597	2.5660
19	3.4338	3.2965	3.1533	3.0031	2.9249	2.8442	2.7608	2.6742	2.5839	2.4893
20	3.3682	3.2311	3.0880	2.9377	2.8594	2.7785	2.6947	2.6077	2.5168	2.4212
21	3.3098	3.1730	3.0300	2.8796	2.8010	2.7200	2.6359	2.5484	2.4568	2.3603
22	3.2576	3.1209	2.9779	2.8274	2.7488	2.6675	2.5831	2.4951	2.4029	2.3055
23	3.2106	3.0740	2.9311	2.7805	2.7017	2.6202	2.5355	2.4471	2.3542	2.2558
24	3.1681	3.0316	2.8887	2.7380	2.6591	2.5773	2.4923	2.4035	2.3100	2.2107
25	3.1294	2.9931	2.8502	2.6993	2.6203	2.5383	2.4530	2.3637	2.2696	2.1694
26	3.0941	2.9578	2.8150	2.6640	2.5848	2.5026	2.4170	2.3273	2.2325	2.1315
27	3.0618	2.9256	2.7827	2.6316	2.5522	2.4699	2.3840	2.2938	2.1985	2.0965
28	3.0320	2.8959	2.7530	2.6017	2.5223	2.4397	2.3535	2.2629	2.1670	2.0642
29	3.0045	2.8685	2.7256	2.5742	2.4946	2.4118	2.3253	2.2344	2.1379	2.0342
30	2.9791	2.8431	2.7002	2.5487	2.4689	2.3860	2.2992	2.2079	2.1108	2.0062
40	2.8005	2.6648	2.5216	2.3689	2.2880	2.2034	2.1142	2.0194	1.9172	1.8047
60	2.6318	2.4961	2.3523	2.1978	2.1154	2.0285	1.9360	1.8363	1.7263	1.6006
120	2.4721	2.3363	2.1915	2.0346	1.9500	1.8600	1.7628	1.6557	1.5330	1.3805
∞	2.3209	2.1847	2.0385	1.8783	1.7908	1.6964	1.5923	1.4730	1.3246	1.0000

Denominator Degrees of Freedom

NPTEL

This table is very very useful and of course excel also can calculate these values so I will show you how to do that. But, manually if you are doing we can use these tables. Especially, the book quite I mentioned in the reference called Fisher, that gives you the data. So, we can use the F table to calculate the F value. Let us go slightly more into the fundamentals.

(Refer Slide Time: 07:14)



F-distribution

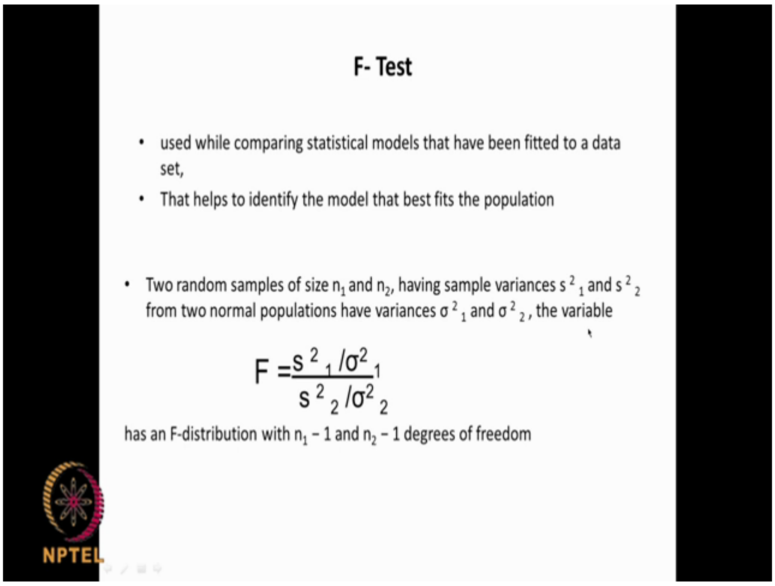
- **Fisher–Snedecor distribution**
- Is a continuous probability distribution Is a right-skewed **distribution** .
- It has a minimum of 0, but no maximum value
- All values are positive.
- Used commonly in Analysis of Variance (ANOVA)
- Hypothesis testing to determine whether two population variances are equal

We need to know what is F distribution? This is called a Fisher-Snedecor distribution, so it generally looks like this, at different values of degrees of freedom it will look like this.

It is a continuous probability distribution. It is a right skewed because on the right hand side it is skewed, it is not like a normal distribution going up and coming down like a bell shaped, but there is skew. It has a minimum of 0, but there is no maximum so it can be up and up, if very very low values of degrees of freedom it can go right up, very high and it is only positive there is no negative. This data is used very useful if you are doing a F-test, that means if you are comparing variances of 2 sample test and if you are doing an analysis of variance so we do that actually.

As I said the null hypothesis will be the variances of the 2 populations are equal, alternate hypothesis for a 2 tailed could be the variances are different, single tailed hypothesis could be variance of sample 1 is greater than variance of sample 2 or the variance of sample 1 is less than variance of sample 2. So, we can do in the same fashion. This is called an F distribution it is skewed like this as you can see here.

(Refer Slide Time: 08:41)



F- Test

- used while comparing statistical models that have been fitted to a data set,
- That helps to identify the model that best fits the population

- Two random samples of size n_1 and n_2, having sample variances $s^2{}_1$ and $s^2{}_2$ from two normal populations have variances $\sigma^2{}_1$ and $\sigma^2{}_2$, the variable

$$F = \frac{s^2{}_1/\sigma^2{}_1}{s^2{}_2/\sigma^2{}_2}$$

has an F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom

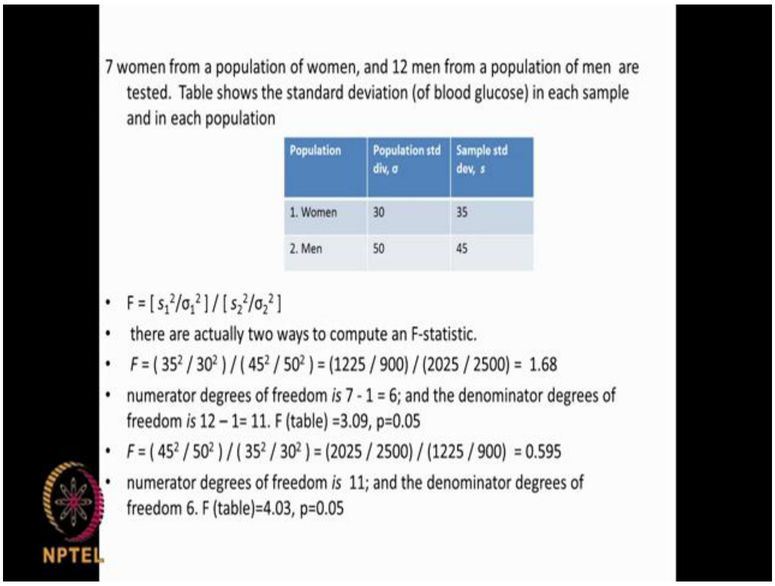
So, F-test as I said, it tries to identify the model it can be used for fitting the data, suppose I am doing a regression analysis it tells you I can find out what is the errors sum of squares. Errors sum of squares means, data that is not fitted to the model as against the total data. That way F-test is very powerful it is used in many, many situations actually and also we can use it if we are comparing as I said instruments, if you are comparing operators, if you are comparing more than 2 sets of samples for the population and so on actually.

If I take two random samples of size n 1 and n 2 and each of these has a variance of $s^2{}_1$ and $s^2{}_2$ and they are from the normal population and the population variances are $\sigma^2{}_1$ and $\sigma^2{}_2$ Then the F is actually given by

$$F = \frac{s^2{}_1/\sigma^2{}_1}{s^2{}_2/\sigma^2{}_2}$$

and the F distribution will have 2 degrees of freedom $n_1 - 1$ and $n_2 - 1$. So, $n_1 - 1$ relates to the numerator sample, $n_2 - 1$ relates to the denominator sample so you have to remember that. You can sometimes have different F values also if I take 1 the numerator as the denominator, denominator as the numerator, but when I read out from the table I will read out in a different way. I will show you an example later.

(Refer Slide Time: 10:14).



7 women from a population of women, and 12 men from a population of men are tested. Table shows the standard deviation (of blood glucose) in each sample and in each population

Population	Population std div, σ	Sample std dev, *s*
1. Women	30	35
2. Men	50	45

- $F = [s_1{}^2/\sigma_1{}^2] / [s_2{}^2/\sigma_2{}^2]$
- there are actually two ways to compute an F-statistic.
- $F = (35^2 / 30^2) / (45^2 / 50^2) = (1225 / 900) / (2025 / 2500) = 1.68$
- numerator degrees of freedom *is* 7 - 1 = 6; and the denominator degrees of freedom *is* 12 – 1= 11. F (table) =3.09, p=0.05
- $F = (45^2 / 50^2) / (35^2 / 30^2) = (2025 / 2500) / (1225 / 900) = 0.595$
- numerator degrees of freedom *is* 11; and the denominator degrees of freedom 6. F (table)=4.03, p=0.05

For example, I am looking at 7 women from a population and 12 men from a population, they are tested for something blood glucose for example, so it shows you the standard deviation of the blood glucose in each sample and in each population. This gives you the standard deviation of the population, this gives you sample standard deviation of the sample.

Here, I have taken 7 women and here I have taken 12 women these are the samples, these are the populations standard deviation from the some sort of a mean and so what do you do,

$F = [s_1^2 / \sigma_1^2]$. That (Refer Time: 10:54)

- $[s_1^2 / \sigma_1^2] / [s_2^2 / \sigma_2^2]$

. So, we can have

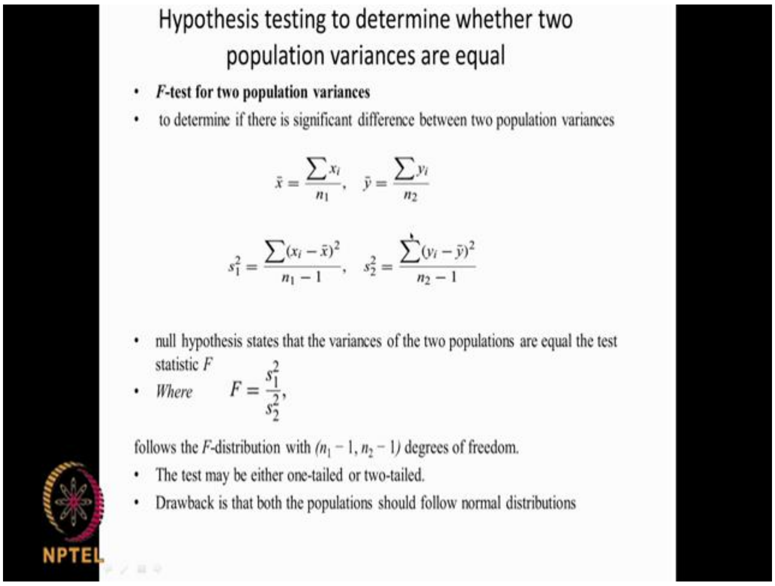
$(35^2 / 30^2) / (45^2 / 50^2)$

so I get 1.68 that is the F. Now the degrees of freedom is numerator is 7 - 1 6, denominator 12 - 1 11. Under 6 and 11 for 95 % confidence I get 3.09 so I put it as 3.09.

Now notice, instead of putting women on top here, I can also put men on top then women at the bottom, so it can be $(45^2 / 50^2) / (35^2 / 30^2)$ so I will get a different F value note that. Now the F value is 0.595. But then I should look under numerator degrees of freedom is 11 and 6 did you notice that, I can put numerator as denominator or denominator as numerator but when I look into the table I look at the corresponding df in that proper way. Here, I will look 11 and 6 instead of 6 and 11. So 11 and 6 if I go to the table, 11 and 6 is next page so it is 4.03 you understand. We get 4.03, so if I had taken men variance on the top and women variance on the bottom the degrees of freedom is 11 and 6 so I will get an F of 4.03. If I put women variance on the top and men variance at the bottom, I will get 6 and 11 for the F table is 3.09. I compare 3.09 with 1.68 so I cannot reject the null hypothesis, because the F calculated is lower than the table here and if look put the men variance on the top then I will get 0.595 this is the calculated F value and the table F value is 4.03, so I cannot reject the null hypothesis. So, either way I cannot reject the null hypothesis.

You need to see this you are doing ratio calculation, so instead of numerator you can have denominator, denominator can be numerator. But when you select the degrees of freedom, you should correspondingly change it and this table as you can see this gives you the numerator degrees of freedom, this gives you the denominator degrees of freedom as you can see that. So, you have to use properly the degrees of freedom. In either way here, in this particular problem we see that you cannot reject the null hypothesis because you are table F value is much larger than your calculated F value for 95 % confidence level, either way whether we put the variance of the men in the numerator and variance of the women in the denominator or if we put variance of the women in the numerator, variance of the denominator in the bottom. This table is very, very useful be it to 95 % confidence interval or be it 99 % confidence interval.

(Refer Slide Time: 14:02)



Hypothesis testing to determine whether two population variances are equal

- *F*-test for two population variances
- to determine if there is significant difference between two population variances

$$\bar{x} = \frac{\sum x_i}{n_1}, \quad \bar{y} = \frac{\sum y_i}{n_2}$$

$$s_1^2 = \frac{\sum (x_i - \bar{x})^2}{n_1 - 1}, \quad s_2^2 = \frac{\sum (y_i - \bar{y})^2}{n_2 - 1}$$

- null hypothesis states that the variances of the two populations are equal the test statistic *F*
- *Where* $F = \frac{s_1^2}{s_2^2}$,

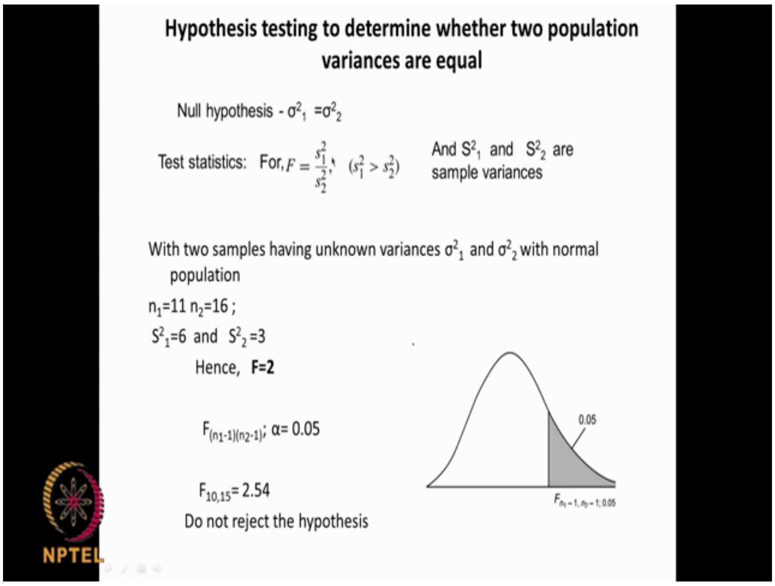
follows the *F*-distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom.

- The test may be either one-tailed or two-tailed.
- Drawback is that both the populations should follow normal distributions

So, F-test looks at 2 population variances and determines if there is significant difference between the 2 populations. If I am taking only samples I do not have any idea about my population. $\bar{x}$ So, will be as you know summation of all the x / n and for other data set it will be summation of y / n_2 and you know how to calculate the standard deviation right, x 1 - $\bar{x}$ and square summation / n_1 - 1, standard deviation of the other 1 y 1 -y bar square summation divided by n_2 - 1. F will be $[\, s_1^2/\sigma_1^2$ and then you look at F distribution n_2 -1 as the numerator that means that will be on the column, n 2 minus 1 as the denominator that will be on the row so do not forget that.

Whereas, if I do F = s_2^2 / s_1^2, then I will look at n_2 - 1 as the column and n_1 - 1 as the row do not forget that. So, we can have either 1 tailed test or a 2-tailed test. Of course, it is good that the data follows a normal distribution, it is not following then it is not very good and as I said I will teach you how to check whether the data is normal or non-normal later on there are some test for that. There are many test later on for non-normal distribution also. So, in that data the null hypothesis could be $\sigma^2_1 = \sigma^2_2$.

(Refer Slide Time: 15:34)



Hypothesis testing to determine whether two population variances are equal

Null hypothesis - $\sigma^2_1 = \sigma^2_2$

Test statistics: For, $F = \frac{s_1^2}{s_2^2}$, $(s_1^2 > s_2^2)$ And S^2_1 and S^2_2 are sample variances

With two samples having unknown variances σ^2_1 and σ^2_2 with normal population

n_1=11 n_2=16 ;

S^2_1=6 and S^2_2=3

Hence, **F=2**

$F_{(n1-1)(n2-1)}$; α= 0.05

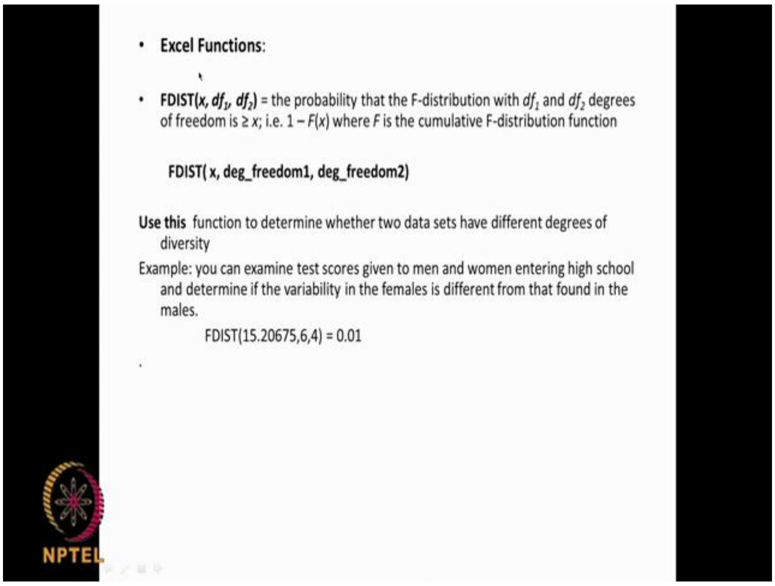
$F_{10,15}$= 2.54

Do not reject the hypothesis

The test statistics could be s 1 square by s 2 square and then numerator degrees of freedom is n_1 - 1, denominator is n_2 - 1. If I have 2 sets of samples n_1 is 11, n_2 is 16, S1 square is 6 and S 2 square is 3. So, the ratio test statistics F will be 6 by 3 that is 2 and then F value I am looking at 11, 11 and 16 that means 10 and 15 for α is equal to 95 % so 10 and 15, what is the of equation? Sorry what is a table? 10 and 15 will be here, 10 here 2.54, the numerator is 10, denominator is 15 degrees of freedom. I put it as 2.54 so F = 2 obviously, the calculated F is < table S so do not reject the null hypothesis.

Please remember these F values are for 1 tail only the right tail, it is only for the right tail. It is very simple we need to use this table and from the table just like t test we calculate F value and F value calculated is less than the table F value, we do not reject the null hypothesis. Now this excel function is also there that is called the f dist function.

(Refer Slide Time: 17:12)



- **Excel Functions:**
- **FDIST(*x*, *df*$_1$, *df*$_2$)** = the probability that the F-distribution with *df*$_1$ and *df*$_2$ degrees of freedom is ≥ *x*; i.e. 1 – *F*(*x*) where *F* is the cumulative F-distribution function

FDIST(x, deg_freedom1, deg_freedom2)

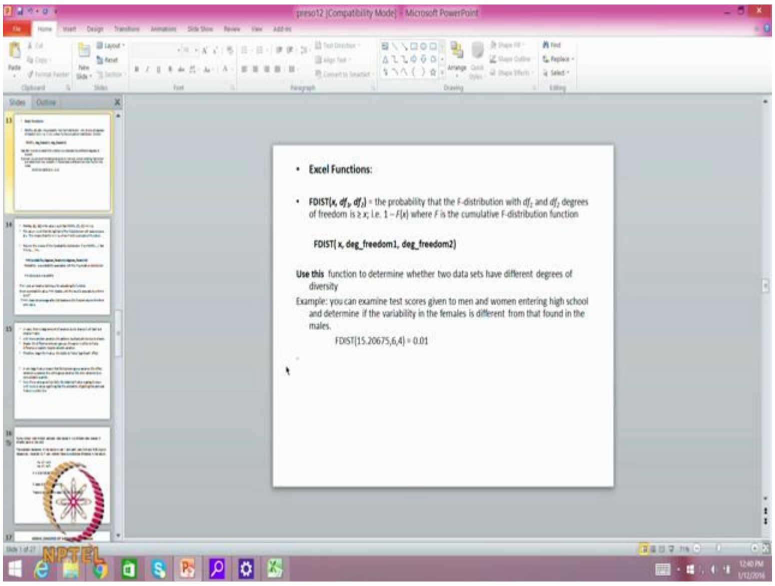
Use this function to determine whether two data sets have different degrees of diversity

Example: you can examine test scores given to men and women entering high school and determine if the variability in the females is different from that found in the males.

FDIST(15.20675,6,4) = 0.01

We give the probability the value. So, F dist gives you the probability of that the Fdistribution is d f 2 and d f 2 degrees of freedom, this gives you cumulative actually. So, we will give the F value, we will give the degrees of freedom and it will give you the probability. For example, let us take this particular problem, so it will be 2, 10 and 15. Let us go to F distribution I will show you in excel also.

(Refer Slide Time: 17:54)



- **Excel Functions:**
- **FDIST(*x*, *df*$_1$, *df*$_2$)** = the probability that the F-distribution with *df*$_1$ and *df*$_2$ degrees of freedom is ≥ *x*; i.e. 1 – *F*(*x*) where *F* is the cumulative F-distribution function

FDIST(x, deg_freedom1, deg_freedom2)

Use this function to determine whether two data sets have different degrees of diversity

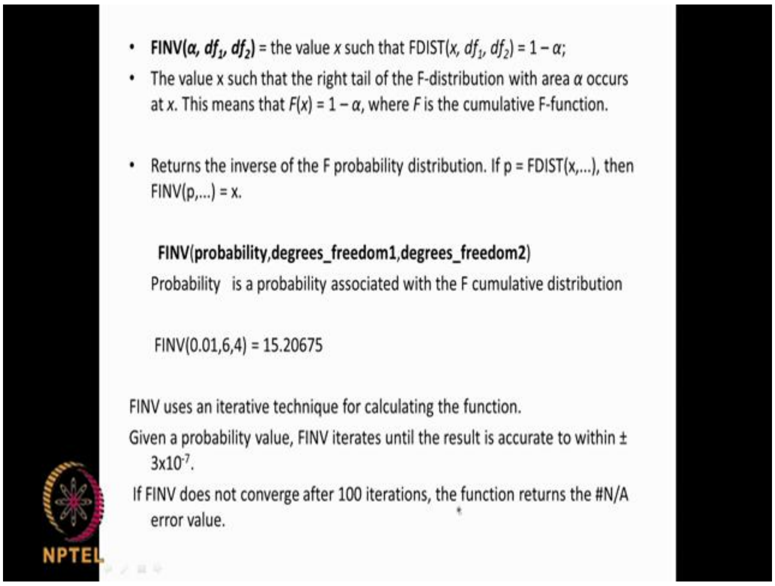
Example: you can examine test scores given to men and women entering high school and determine if the variability in the females is different from that found in the males.

FDIST(15.20675,6,4) = 0.01

So, let us look at the previous case, as such sorry will see **FDIST** 2 is my test statistics, degrees of freedom is 10 , 15. 2 is my F statistics 10 and 15 or the numerator and the denominator degrees of freedom. It gives you my p value that is 0.10914, that is what this F distribution gives, sorry, yeah so it gives you the probability value. So, it gives you the pre-value obviously, p a is 0.109 which is much very large. So obviously, we reject the null hypothesis. This **FDIST** function in excel gives you the p value. Now F inverse is also there, F inverse. So, it gives you the value of x such the right tail of their distribution will have an area of α.

Basically, it gives you the inverse of the F probability distribution. So, you have the degrees of freedom here, degrees of freedom here and it will give you the given the alpha, it will calculate what will be the so in this particular problem if I say F inverse 0.05, 10 and 15 it gives you 2.54 see it gives you this. So, for a prob p = 95 % or p is equal to 0.05. Whatever that table is giving so it will give you 2.54. As I said 10 and 15 it gives you 2.54. I can use the **FINV** function to calculate the value exactly the area exactly just like the table.

(Refer Slide Time: 20:32)



- **FINV(α, df_1, df_2)** = the value x such that FDIST(x, df_1, df_2) = 1 − α;
- The value x such that the right tail of the F-distribution with area α occurs at x. This means that $F(x) = 1 - \alpha$, where F is the cumulative F-function.
- Returns the inverse of the F probability distribution. If p = FDIST(x,...), then FINV(p,...) = x.

FINV(probability,degrees_freedom1,degrees_freedom2)

Probability is a probability associated with the F cumulative distribution

FINV(0.01,6,4) = 15.20675

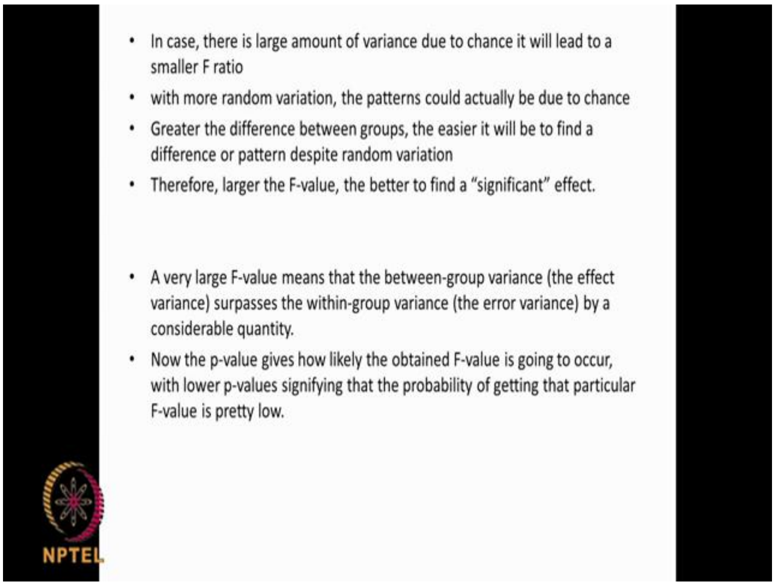
FINV uses an iterative technique for calculating the function.

Given a probability value, FINV iterates until the result is accurate to within ± $3x10^{-7}$.

If FINV does not converge after 100 iterations, the function returns the #N/A error value.

Here I will mention the probability 0.05. Whereas, the **FDIST** will give you the probability, where I will put the ratio here 2.54 and then or 2 here and then put the degrees of freedoms here 10 and 15. So I can use **FDIST** which will give me the area or the probability or if you want to given the probability it will calculate the F value using **FINV** function. We can use both either the **FDIST** or **FINV** both are quite good so this can give. F inv it gives you the probability that means it will calculate the probability and if the probability calculate is less than 0.05 we can we can reject the null hypothesis. So, you use this particular thing or we use the **FDIST** it will calculate the area under the curve.

(Refer Slide Time: 21:47)



- In case, there is large amount of variance due to chance it will lead to a smaller F ratio
- with more random variation, the patterns could actually be due to chance
- Greater the difference between groups, the easier it will be to find a difference or pattern despite random variation
- Therefore, larger the F-value, the better to find a "significant" effect.

- A very large F-value means that the between-group variance (the effect variance) surpasses the within-group variance (the error variance) by a considerable quantity.
- Now the p-value gives how likely the obtained F-value is going to occur, with lower p-values signifying that the probability of getting that particular F-value is pretty low.

So, in case there is large amount of variance due to chance it will lead to a smaller F ratio. Of course, in as you know numerator denominator, if the denominator variances are very large obviously, the ratio will become small. So, the F ratio also will become small so there would not be any significant. What does this statement means? Suppose I am an operator I do repeat the same sample 4 times or 5 times and each time I get large variation, obviously, my variance will be very large.

Whereas, I can take an export operator who repeats the same sample 4,5 times his variance is not very large, it is very, very small. So, if I am dividing 1 by other I will get much difference. If the variances are very large I am going to get much smaller f ratio. So, I will not be able to differentiate between 2 sets of drugs for example or 2 sets of a instruments for example. If the repeatability is good, then if there is some variations between sample 1 and sample 2 then I will be able to see the difference. If the experimental errors are very large then I will not be able to see variations between sample 1 and sample 2, that is what this particular statement is all is about. That means, if I am doing experiments the variations in my experimental measurements should be very, very, very less. With more random variation the patterns could actually be due to chance, so if there is lot of variation it could be just greater the difference between the groups, then the easier it will be to find a difference or pattern despite random variation. If the variations are very, very large, even if the variations because of error or chance random is very large will be able to find differences.

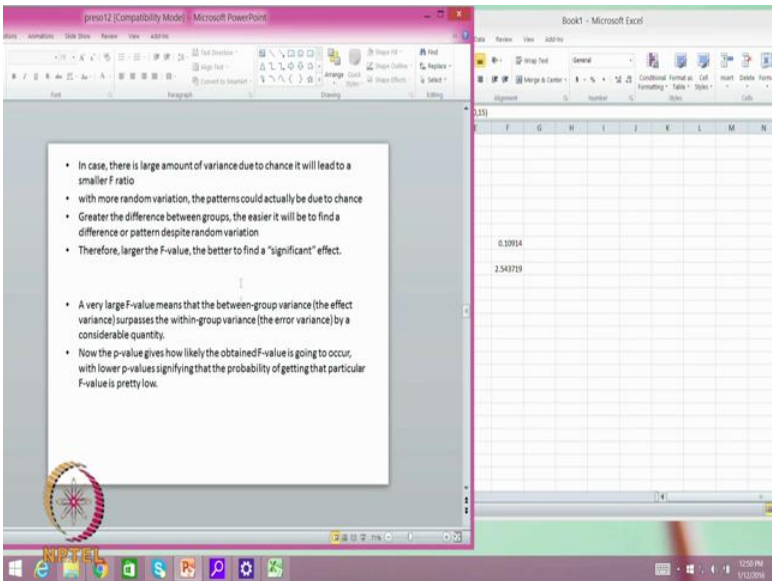
But if the variations are between 2 sets of samples is small and the variations repeatability is large, then I will not be able to do because my F ratio will be come out very small. So, larger the f ratio better to find a significant effect. If I am comparing 2 sets of drugs or 2 sets of instruments, I need to have larger F ratio. My variations repeatability should be very less, that means within sample error should be very less. I am introducing a new word that is called within sample error. So, within sample error should be less whereas, between sample that means sample a and sample b or instrument a and instrument b so that is called between sample variance. Whereas, within sample variance means variations in the repeatability. So, 1 is called between samples that is a and b instrument, 1 instrument 2, operator 1 and 2, that should be much larger than within sample error that is called the sampling error or chance error or random error.

So, between sample verses within sample that is what you are trying to do in F ratio and it is very powerful because we can use this for comparing like I said 2 drugs or 3 drugs or many drugs, 2 operators or 3 operators or many operators or many instruments, sample done in different location and so on actually. Always we should have less variation within sample to see a difference between samples, that is very important statement. A very large F value means that between group variance surpasses the within group you understand, between sample error variance within sample variance. Always within sample variance should be very, very small in order to find between sample variance or between group variance. This within sample variance is also called error variance. Whereas, between sample or between group is also called effect variance that means effect I am comparing between drug a and drug b, temperature of fermentation 30 ° or 40 °, fermentation at pH of 3 and 4. So if I want to see a real variation and statistically significant variation between pH at 3 and 4.

When I do a repeatability, those variation should be much smaller then only I will be able to see a difference between effect that means p h 3 and 4 or temperature 30 and 40 or drug 1 and drug 2 or operator a and operator b. So, within sample variance should be much smaller than between sample variance. That means I should have a good assay method, I should have good instruments so that the errors when I repeat are much, much less. This is very key in F-test. In fact, it is very key in many studies where we are comparing samples, comparing drugs, comparing operators, comparing assays, comparing instruments, comparing bio processors, comparing yield of plants and so on.

The p value gives you how likely the obtained f value is going to occur, p value we can use this f dist that gives you the p value. You can use **FDIST** function, where x is the ratio of the variances d f numerator and denominator. The p value gives how likely obtained the F value is going to occur. So, if the p value > 0.05 then we will say it is not greatly significant, if the p value is very small we will say it is greatly significant. You can use, as I said **FDIST** function in excel to do this particular job for you. We can also use the graph pad software.

(Refer Slide Time: 27:55)



- In case, there is large amount of variance due to chance it will lead to a smaller F ratio
- with more random variation, the patterns could actually be due to chance
- Greater the difference between groups, the easier it will be to find a difference or pattern despite random variation
- Therefore, larger the F-value, the better to find a "significant" effect.
- A very large F-value means that the between-group variance (the effect variance) surpasses the within-group variance (the error variance) by a considerable quantity.
- Now the p-value gives how likely the obtained F-value is going to occur, with lower p-values signifying that the probability of getting that particular F-value is pretty low.

0.10914

2.543719

As you can see here, it gives you the F value. So, we can say calculate, continue so here we can say calculate p value from F or calculate F value given a probability. As you can see **FDIST** and F inverse both can be done here. So, we can say here for example, so F ratio so we give a probability say 0.05 and then the degrees of freedom 10 and degrees of freedom 15 in that problem compute F value so you see 2.543, it gives you here 2.543. That is equivalent to your f inverse. So, if you want to do f dist then we obviously the other.

Here what do we do? We give the F value and then this will give you the probability value equivalent to f dist, so it gives you 0.1091 understand. We can use both these functions in this software, so this gives you the exactly whatever the **FDIST** function gives you, it gives you the probability value for given F ratio and degrees of freedom. Whereas this one gives you the F value, when I give the probability of 0.05 and the 2 numerator degrees of freedom and the denominator degrees of freedom.

So, we will continue on this F distribution further in the next class also.

Thank you very much.