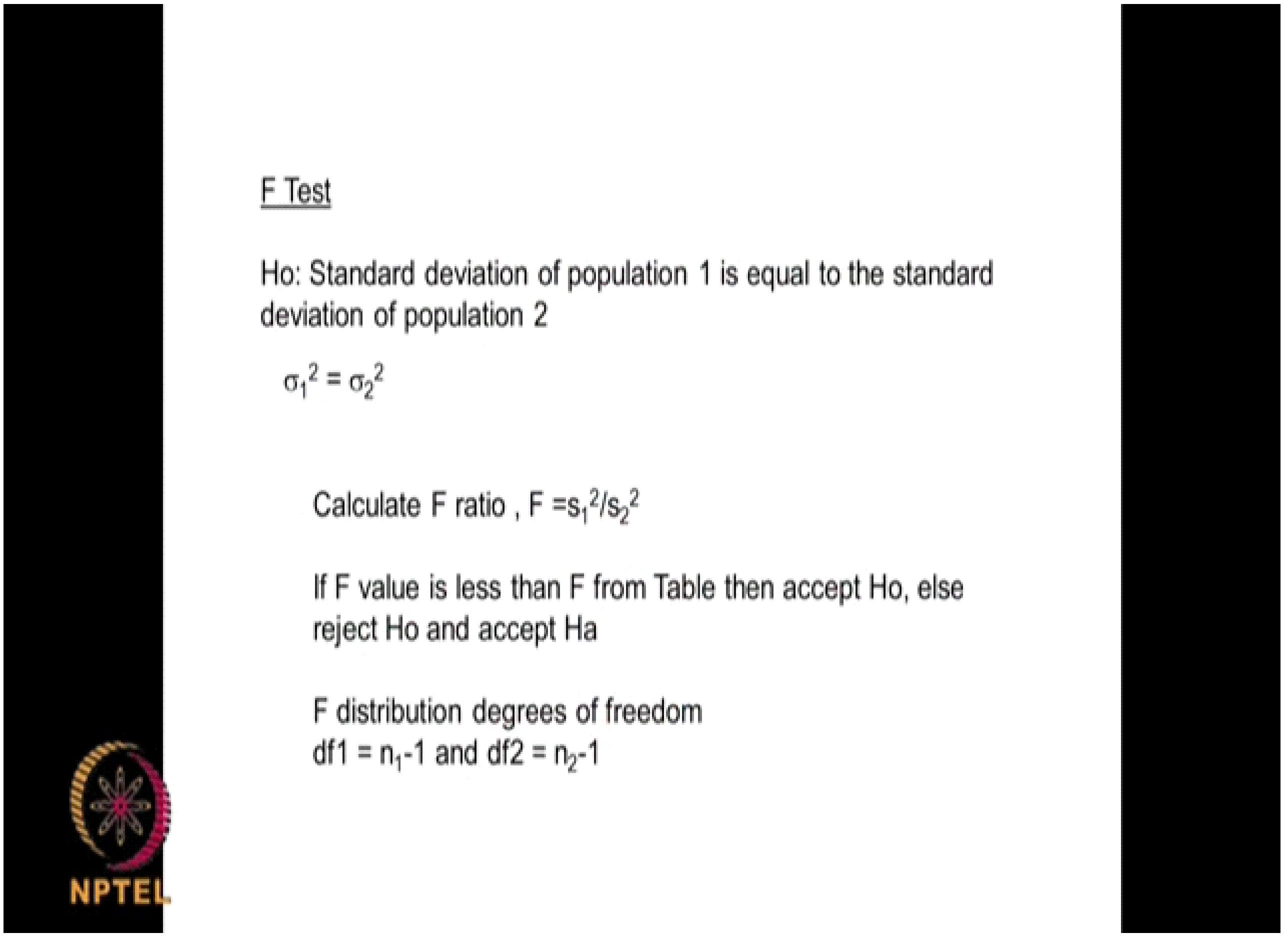
Biostatistics and Design of Experiments
Prof. MukeshDoble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 14
ANOVA

Welcome to the course on Biostatistics and Design of Experiments. Last class I introduced the terminology called ANOVA. We will continue on that same topic, ANOVA is nothing but analysis of variance. So ANOVA makes use of the F Test.

(Refer Slide Time: 00:27)



F Test

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

$\sigma_1^2 = \sigma_2^2$

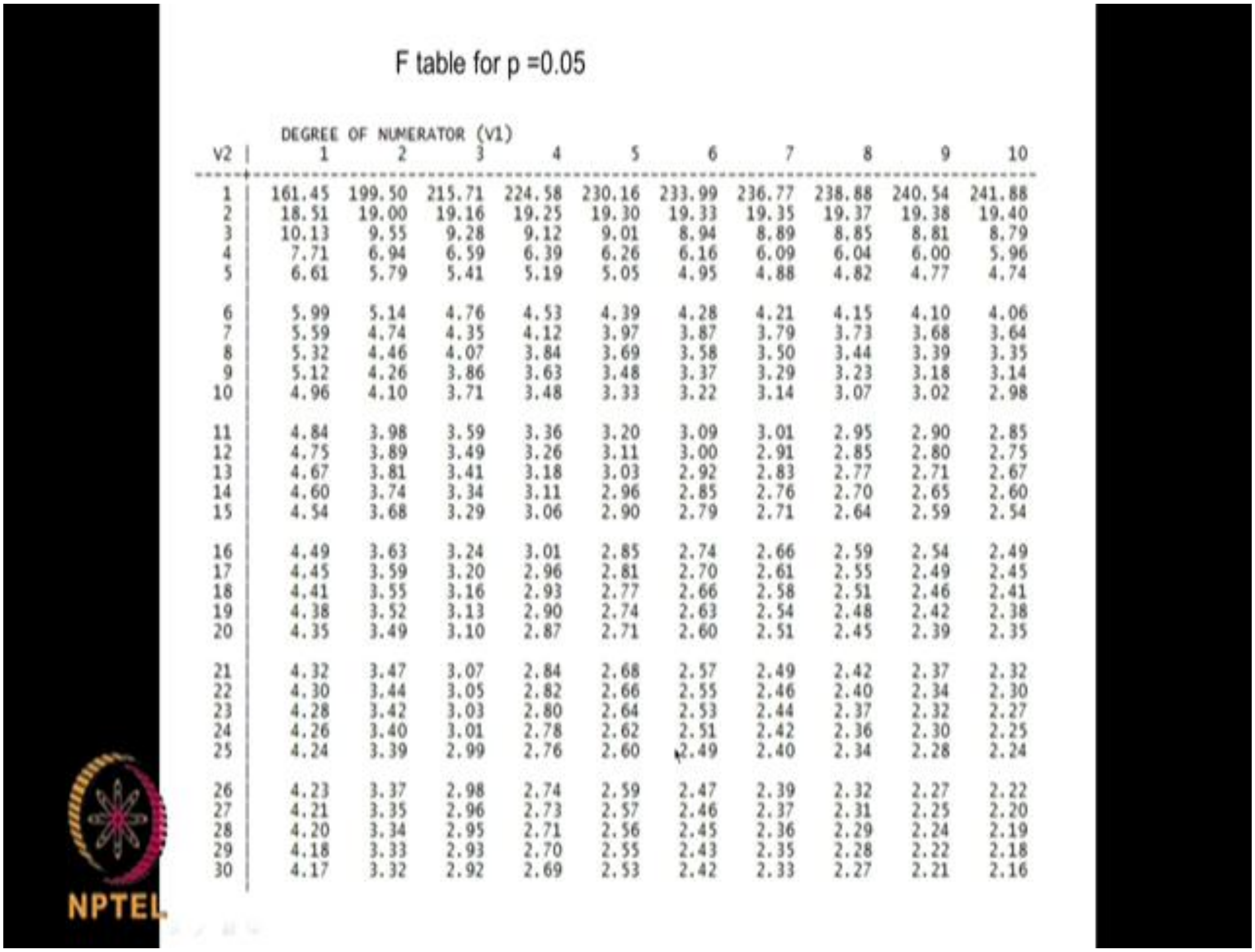
Calculate F ratio , $F = s_1^2/s_2^2$

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 $df2 = n_2 - 1$

Basically, it does a F ratio calculation, that is s_1^2 that is variance from sample 1 / variance from sample 2, so the null hypothesis will be the variances of the populations or equal, the alternate could be different or $\sigma_1^2 > \sigma_2^2$ and so on. We also have a table, F table. If the F value calculated is less than the F table, we do not reject the null hypothesis, we accept the null hypothesis. If the F value calculated is greater than the table, we reject the null hypothesis, hence accept the alternate hypothesis.

(Refer Slide Time: 00:34)



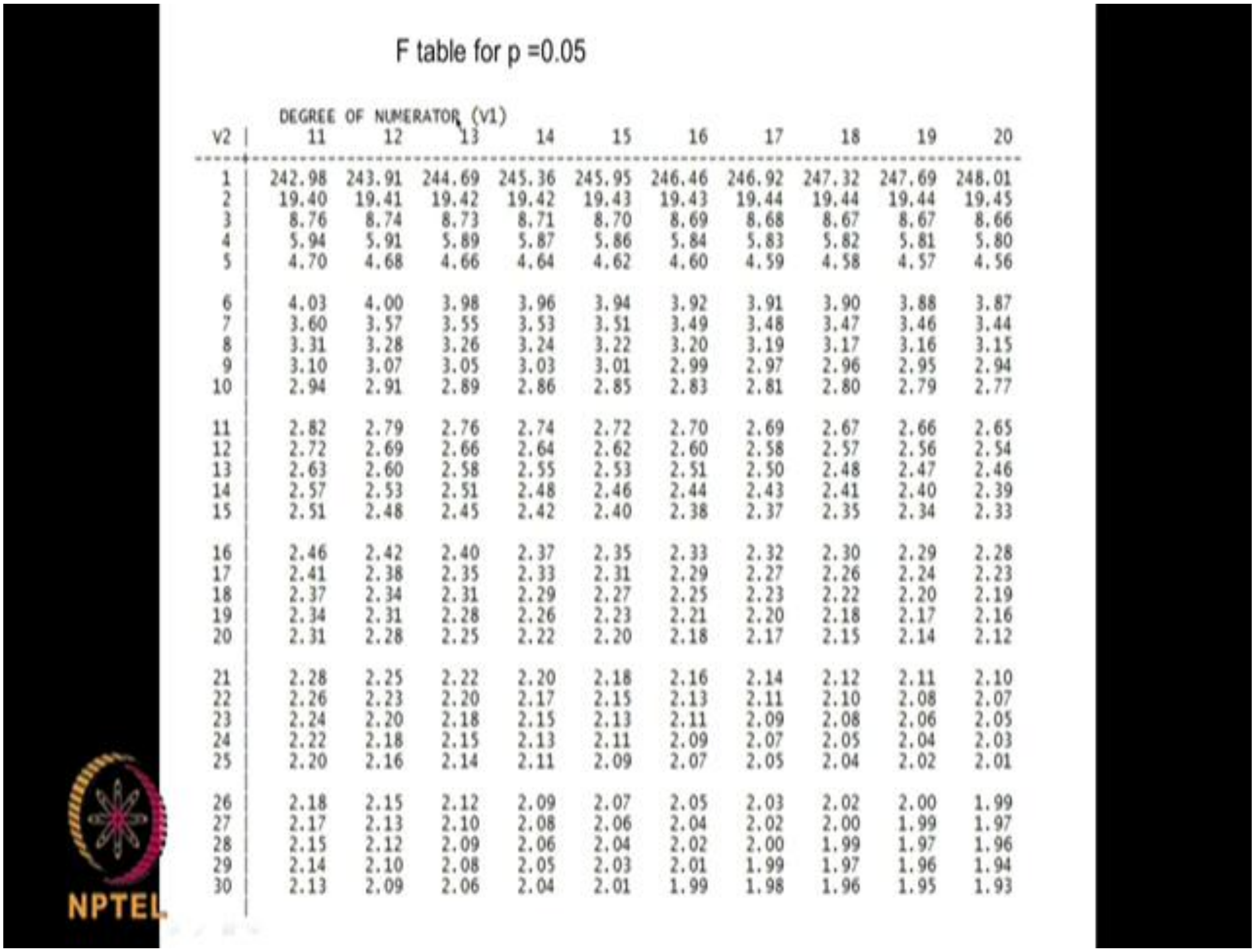
F table for p =0.05

	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

Now, I also mention the degrees of freedom for the numerator will be n- n_1-1 where n 1 is the number of data points for the numerator, sample and degrees of freedom for the denominator will be n_2 - 1, where n 2, is the number of data points for the sample 2 and there you have tables for p

= 0.5 that means in 95 % confidence we have F table, this is for the numerator and this is for the denominator, so this if we looked at the next page it continues like this so you select the F value.

(Refer Slide Time: 01:46)

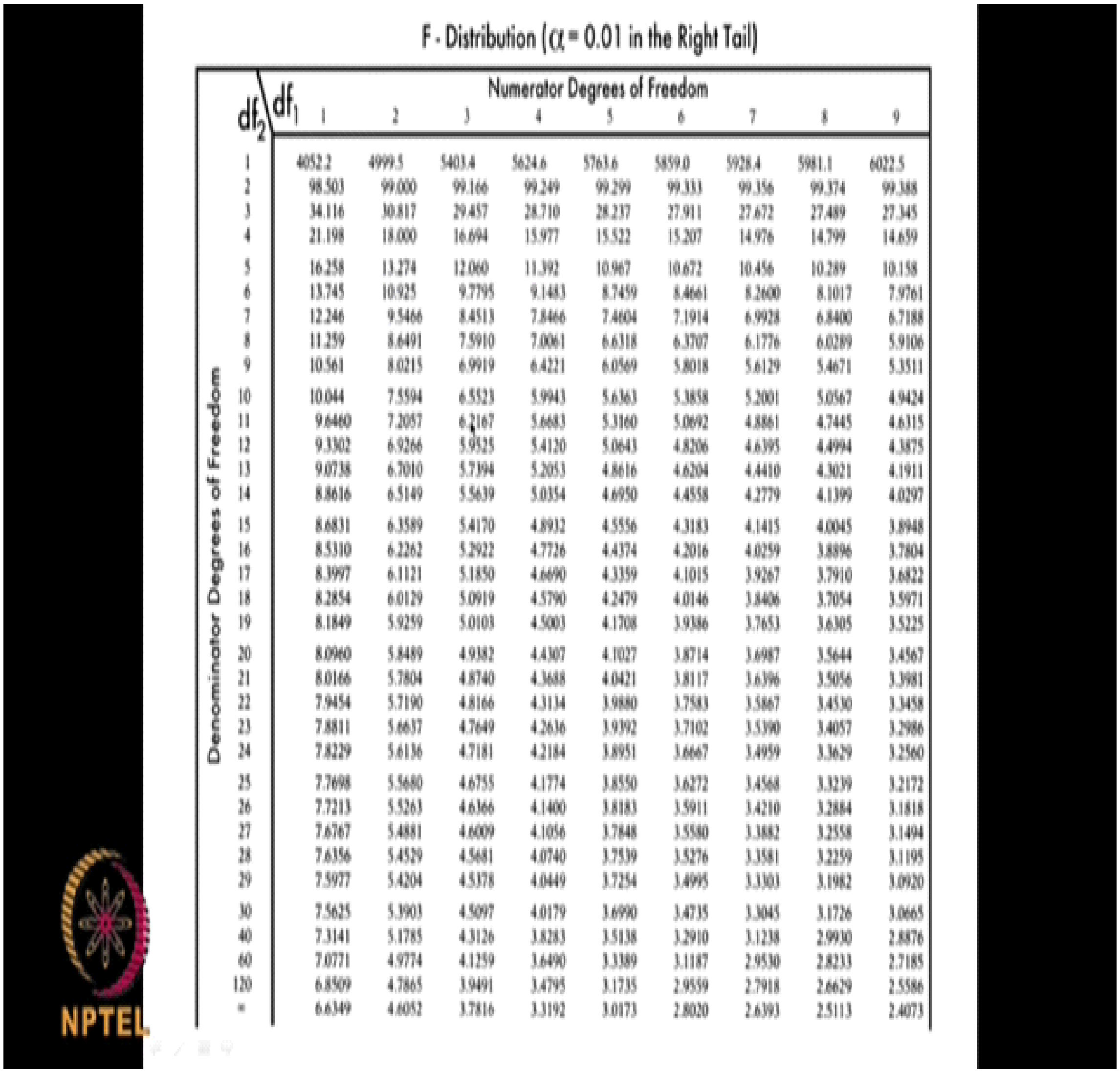


F table for p =0.05

	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

So this is for p = 0.05.

(Refer Slide Time: 01:53)



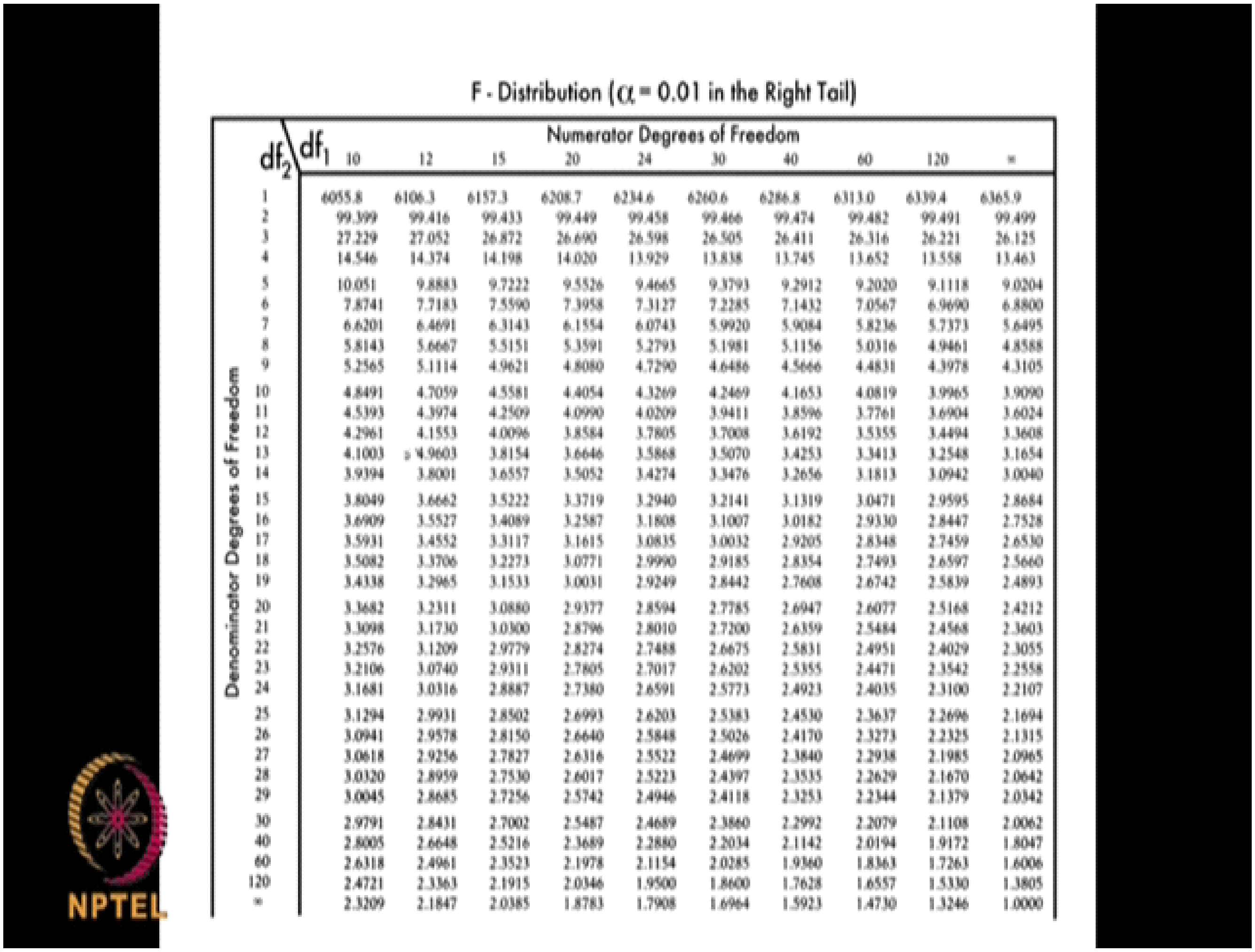
F - Distribution (α = 0.01 in the Right Tail)

df_2 \ df_1	Numerator Degrees of Freedom								
	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

Denominator Degrees of Freedom

Similarly we have p =0.01 and so on.

(Refer Slide Time: 01:57)



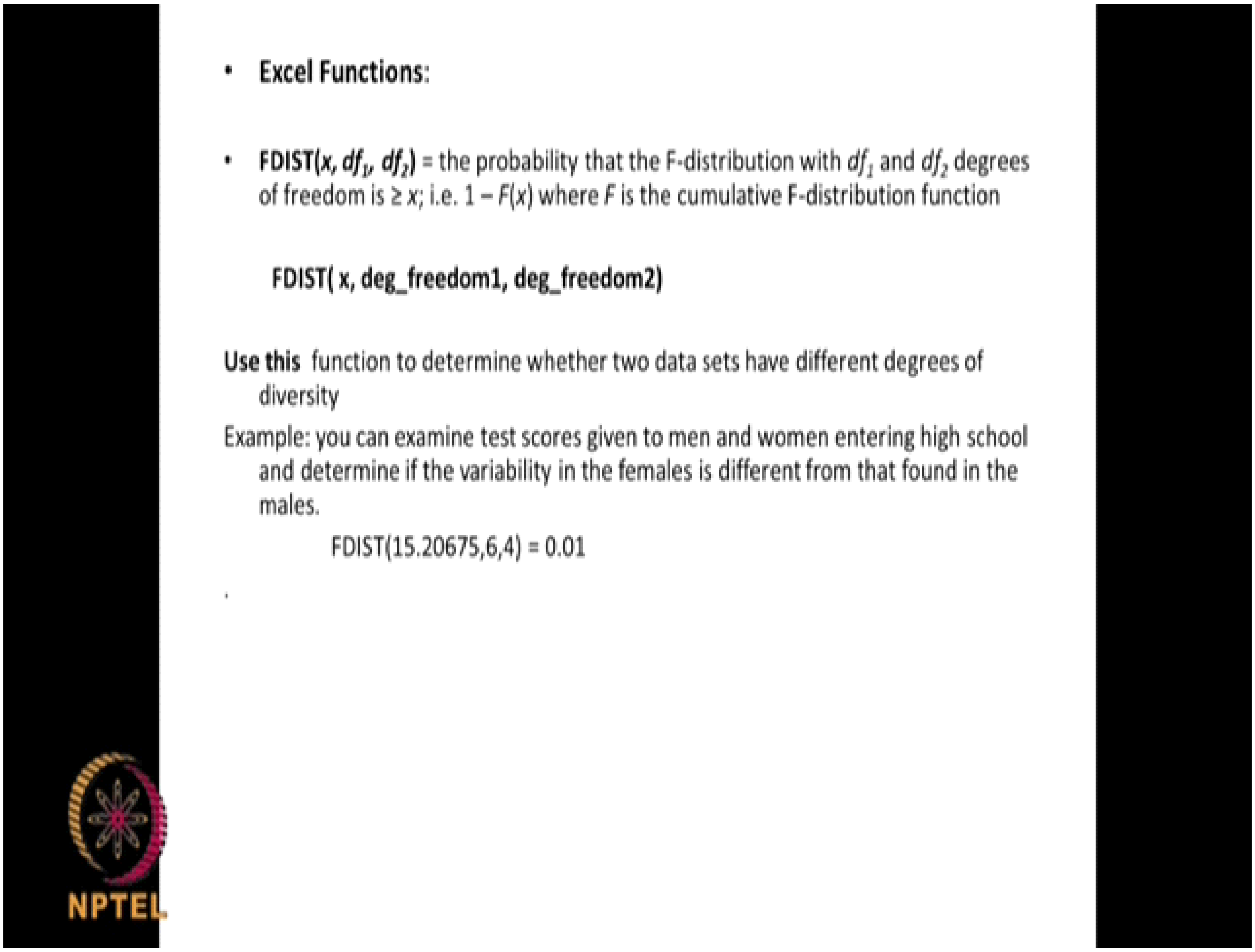
F - Distribution (α = 0.01 in the Right Tail)

Denominator Degrees of Freedom	Numerator Degrees of Freedom									
$df_2 \backslash df_1$	10	12	15	20	24	30	40	60	120	∞
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	4.9603	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

NPTEL

Actually, for any value of p you will be able to get it.

(Refer Slide Time: 02:03)



- **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)

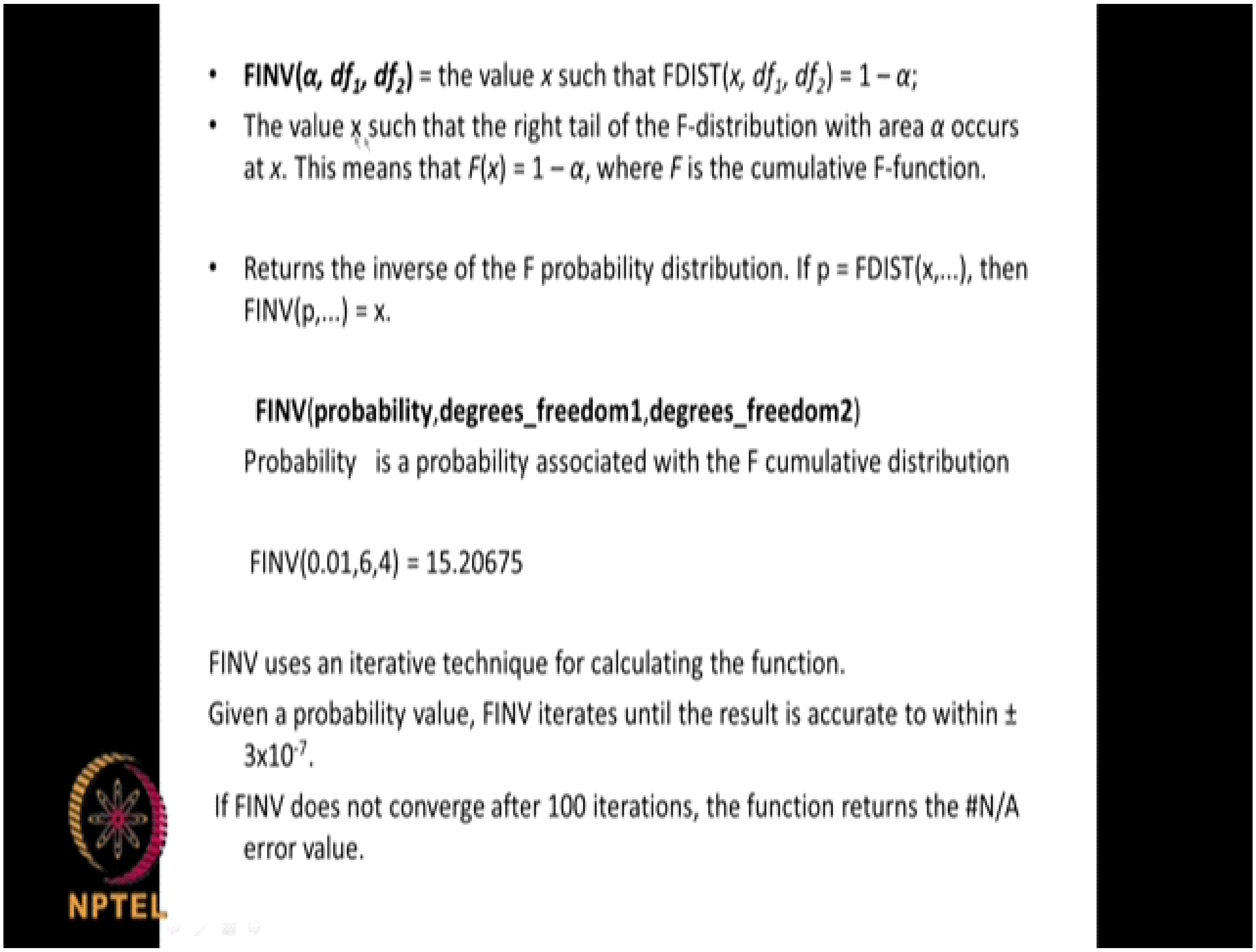
Use this function to determine whether two data sets have different degrees of diversity

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

Now we also have the Excel function, there are 2 Excel functions, one is called the FDIST other is called the FINV.

(Refer Slide Time: 02:08)



- **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

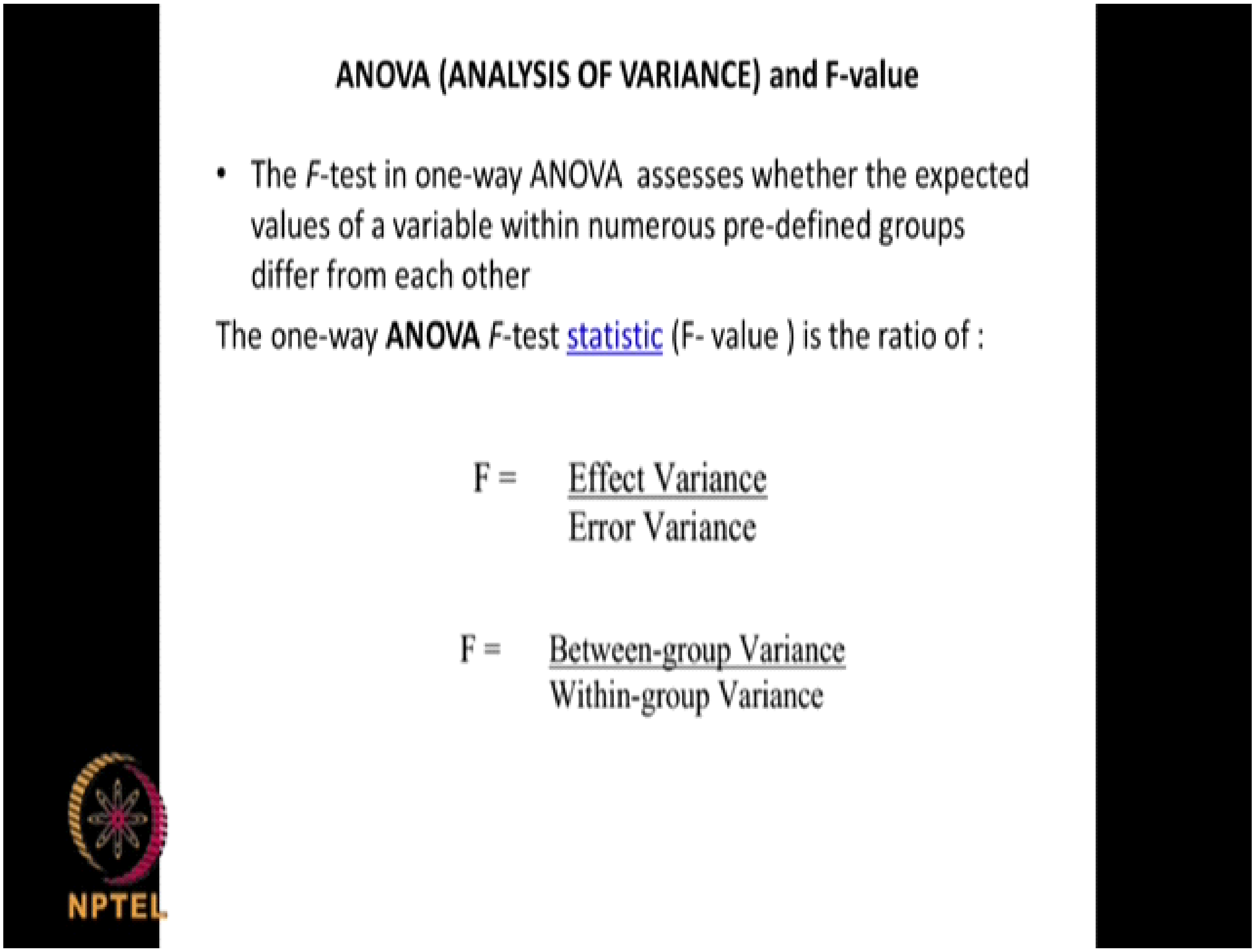
FINV uses an iterative technique for calculating the function.

Given a probability value, FINV iterates until the result is accurate to within ± 3×10^{-7}.

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

FINV is exactly like your F table, given the probability I put in say 0.05 I give the numerator degrees of freedom denominators it gives you the F value, so if you don't have the table we can use the FINV function. Now FDIST, FDIST gives you the probability where here, I put in the ratio. If I put in the ratio that is F ratio give the degrees of freedom for numerator and denominator it calculates the probability. We have both functions available in an excel so we can make use of it. So as I said FINV is exactly like your F table, whereas FDIST calculates the probability and if the probability is high greater than 0.05 generally, there is no reason for you to reject the null hypothesis and similar to this, we also have the GraphPad also has got both the options I showed in the previous class when I come and start doing new problems, I will show you in this present class also.

(Refer Slide Time: 03:16)



ANOVA (ANALYSIS OF VARIANCE) and F-value

- The *F*-test in one-way ANOVA assesses whether the expected values of a variable within numerous pre-defined groups differ from each other

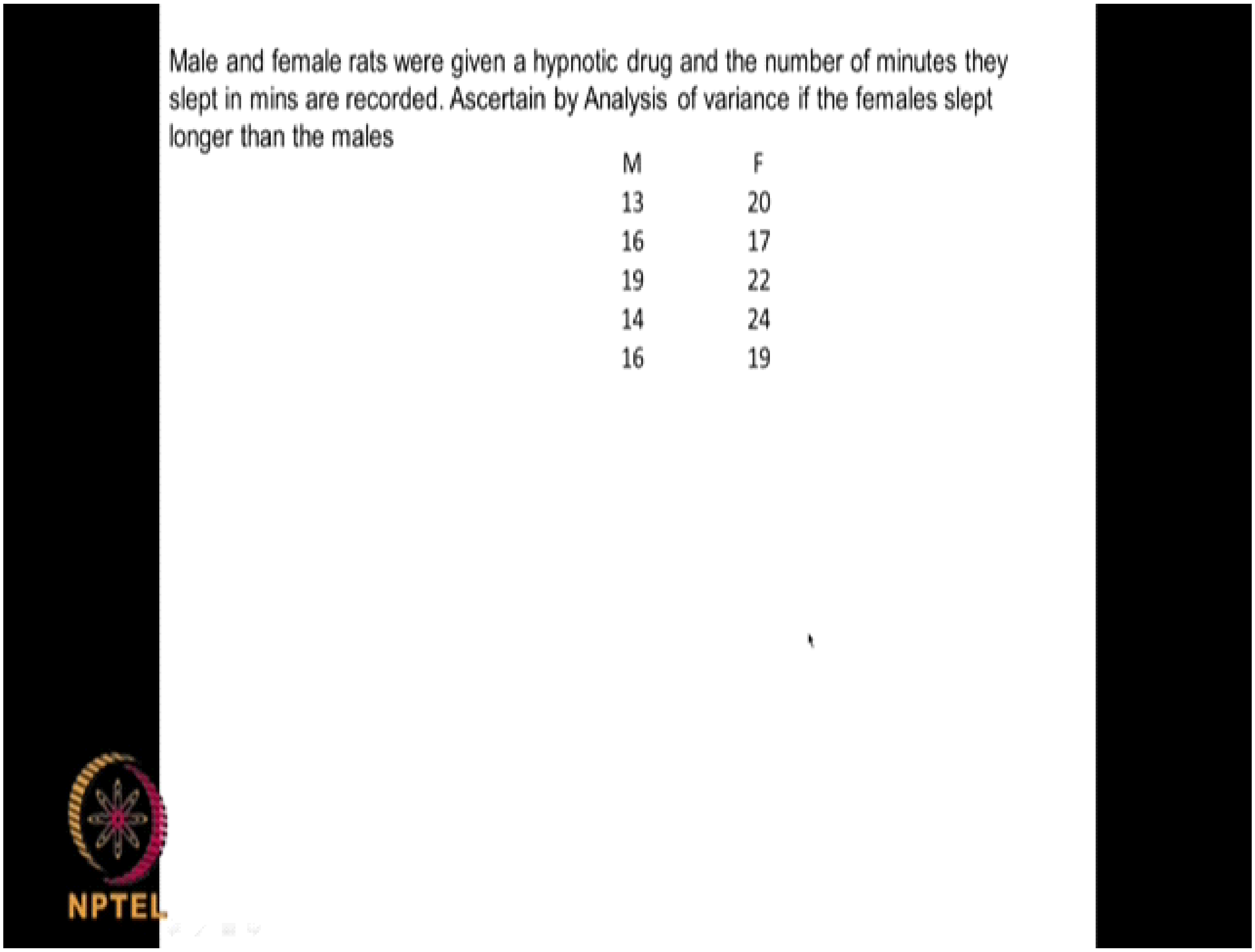
The one-way **ANOVA** *F*-test statistic (F- value) is the ratio of :

$$F = \frac{\text{Effect Variance}}{\text{Error Variance}}$$

$$F = \frac{\text{Between-group Variance}}{\text{Within-group Variance}}$$

We introduce the concept of ANOVA, ANOVA is nothing but Analysis Of Variance so we are taking a ratio of 2 variances, one is between group variance and within group variance. Suppose I am comparing 2 drugs, so between group variance will be comparison of variances between drugs, so within group variances is nothing but error. So, if there is a lot of variation when I within the group obviously, the with the within group variance or error variance will be very large, that means denominator will be very, very large, then F value will be small, so I will not be able to differentiate between 2 different drugs for example. So that is some important point one needs to keep in mind, so when I am repeating experiments. For example, different operators if they consistently there on is good then obviously, the error variance will be good, if the low in error variance will be low, that means the F value will be high. If there is lot of variation or standard deviation then obviously the error variance will be large, F value will be small. Obviously, we will not be able to reject the null hypothesis. Then I mentioned something called one-way ANOVA. Let us look at some problems here.

(Refer Slide Time: 04:36)

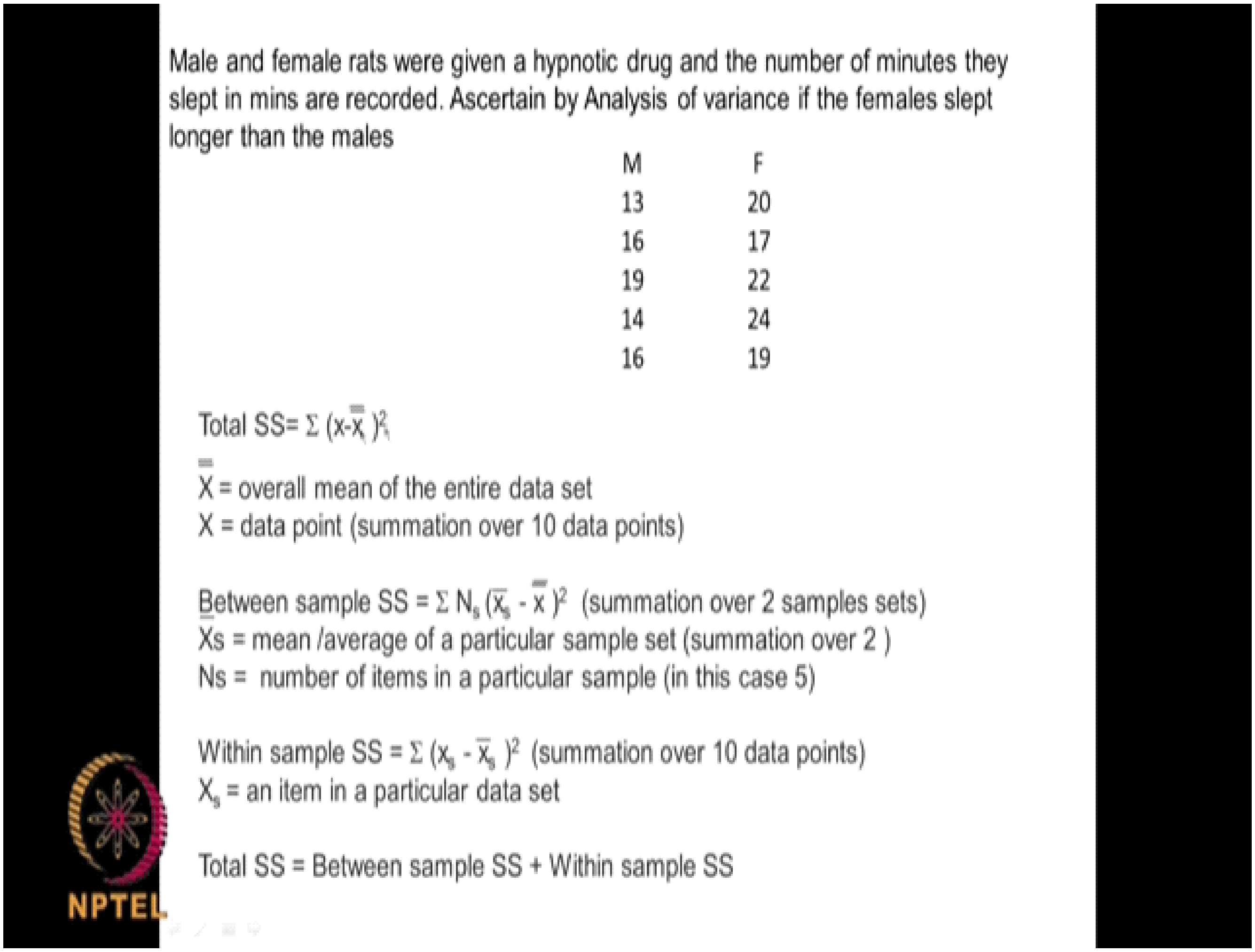


Male and female rats were given a hypnotic drug and the number of minutes they slept in mins are recorded. Ascertain by Analysis of variance if the females slept longer than the males

M	F
13	20
16	17
19	22
14	24
16	19

We have male and female rats, they are given hypnotic drug and then number of minutes they slept in minutes are recorded. Male rats they took 5 samples here, female rats they took 5 samples here and these are the time they took, so many minutes they took to sleep. Obviously, I want to know if the females slept longer than the males. The numerator in the F value will be between groups that means between male and female, within groups will be this. So the variations happening here that will come in the denominator, the variations between these 2 groups will come in the numerator. We need to do some calculation and then finally divide 1 set of variance with the another set of variance, so it is quite simple, let us do it very systematically.

(Refer Slide Time: 05:33)



Male and female rats were given a hypnotic drug and the number of minutes they slept in mins are recorded. Ascertain by Analysis of variance if the females slept longer than the males

M	F
13	20
16	17
19	22
14	24
16	19

Total SS= $\Sigma\ (x-\bar{\bar{x}}_i\ )^2_i$

$\bar{\bar{X}}$ = overall mean of the entire data set

X = data point (summation over 10 data points)

Between sample SS = $\Sigma\ N_s\ (\bar{x}_s - \bar{\bar{x}}\ )^2$ (summation over 2 samples sets)

Xs = mean /average of a particular sample set (summation over 2)

Ns = number of items in a particular sample (in this case 5)

Within sample SS = $\Sigma\ (x_s - \bar{x}_s\ )^2$ (summation over 10 data points)

X_s = an item in a particular data set

Total SS = Between sample SS + Within sample SS

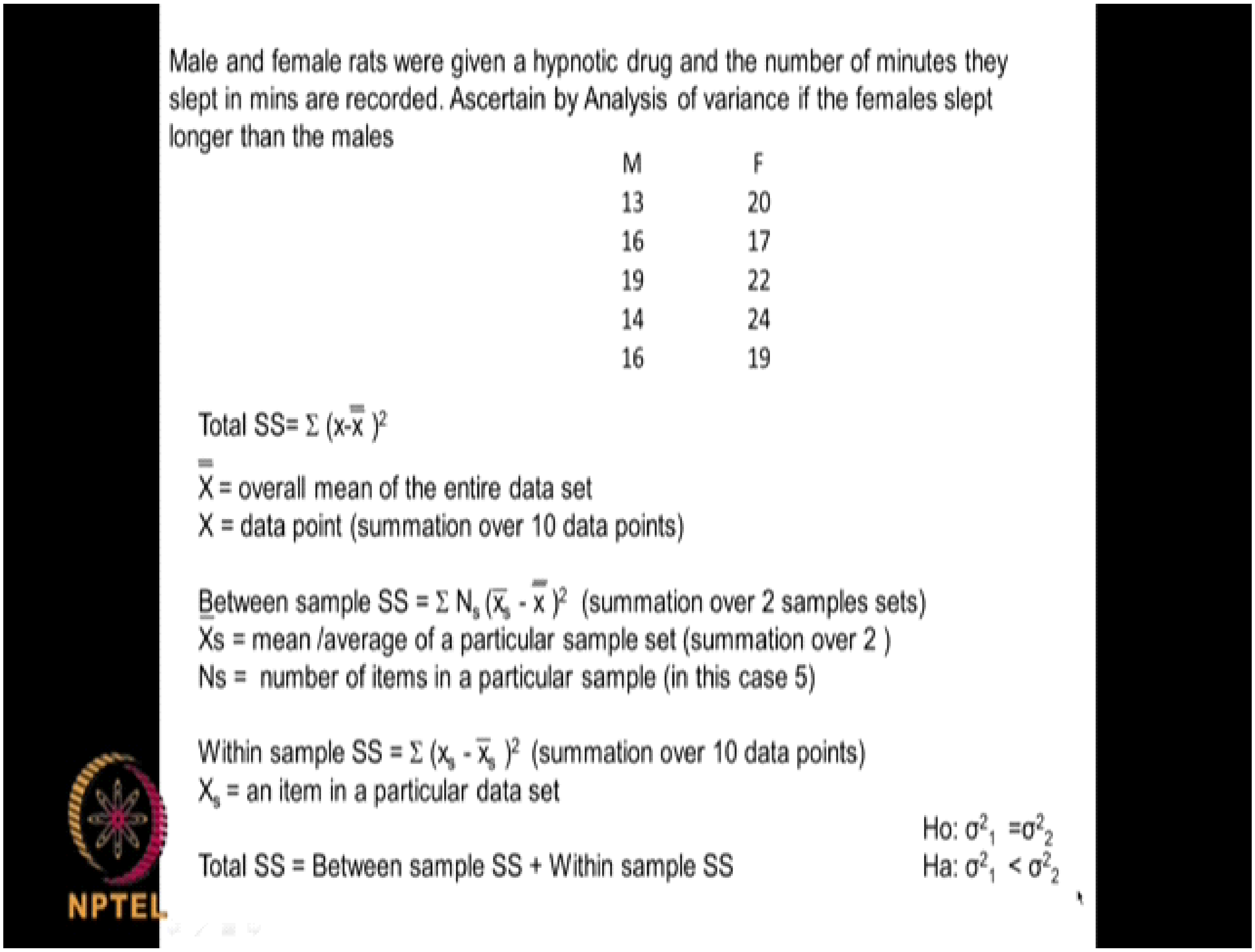
There is 3 different variances we need to calculate, the total sum of squares will be between sample sum of squares, between samples means between here groups the male and the female, plus within sample sum of squares that is within this, because as I said when we calculate F we do a between sample / within samples, we will do between sample variance / within sample variance. So if I calculate sum of squares, if I divide by degrees of freedom that will give you the variance, if take a calculate sum of squares within sample divided by its degrees of freedom I will get the variance. Total sum of squares will be some of these 2, how do you calculate total sum of squares? Summation of x is any of the value / x double bar, x double bar is over all mean, you will have a mean for male, that means if I add these 5 and / 5 I will get some mean right ?, I will call it Xs , so I will have a mean for x bar, I will have a mean for female, I will have a mean for male if I take mean of these 2 means I will get total mean over all mean or I add up all these and I divide by 10 also, I will get over all mean so the number of data points is 10. The total sum of squares here can be easily calculated so I will take a overall mean, that means I take a mean for male I take a mean for female then I take a mean of these 2 means that will give me the overall mean, x minus x double bar square summation that's called the total sum of squares, now u understood how to calculate that?

Now, between sample sum of squares between samples. So there is going to be a mean here, that is X s bar, this going to be another mean here, now we have the overall mean here. So what is the sum of squares between these 2, that is given by X s bar that is the mean of this data, x double bar is the over all means, square it, N s, N s how many data points are there? There are 5 samples right, so 5 here. in this case both the you have 5, you can have different sets also so 5. So when I do for male, I will take the average which are calculated for male, subtract from the overall mean then multiply by five plus I will subtract the average of the females here, from the overall mean square it, then again multiply by the 5, add those, that will call the between samples sum of squares, understand? Between sample sum of squares.

Now between samples sum of squares, so this is there will be 2 samples only here, suppose I had see 3 drugs or if I had male, female and infant you know then I may have 3 sets of samples, then I will have a 3 terms which needs to be added here I am adding. Within sample sum of squares, so it that is called more like an error you know. Within sample sum of squares how do I do? So I have a mean here, that is X s bar, subtract each one of these, square it up, then add it up, then I then I will have an average here mean, subtract from each one of these then square it up, add it up, so there will be 10 data points coming here right, you understood?. There is a mean here, X s bar, subtract from 16, square it, then mean here subtract from 14, square it, there is a mean here subtract from 19, square it, there is a mean here subtract from 16, square it, there is a mean here subtract from 13, square it, add all of them.

Now there will be another mean here subtract 19 from that, square it, another mean here subtract from 24, square it, another mean subtract from 22, square it, another mean subtract from 17, square it, another mean subtract from 20, square it, add up all of these. That is called the within sample. So the X s you select, if it is for male you will take the average which you calculated for male, if it is the female you will take the average from the female, summation over 10 data points, because there are 10 data points. Whereas when you are doing between sample there are only 2 sets because you have a mean here, you have a mean here, you are subtracting from the overall mean, so there are only 2 terms which you are adding, you understood how to do that? So, the degrees of freedom for samples or groups will be 2 samples, 2 - 1 is 1, understood ?. The degrees of freedom for total we have a say 5 + 5 10, 10 - 1 that will be 9. So within samples will be 9 - 1, 8.

(Refer Slide Time: 10:46)



Male and female rats were given a hypnotic drug and the number of minutes they slept in mins are recorded. Ascertain by Analysis of variance if the females slept longer than the males

M	F
13	20
16	17
19	22
14	24
16	19

Total SS= $\Sigma\,(x-\bar{\bar{x}})^2$

$\bar{\bar{X}}$ = overall mean of the entire data set
X = data point (summation over 10 data points)

Between sample SS = $\Sigma\, N_s\,(\bar{x}_s - \bar{\bar{x}})^2$ (summation over 2 samples sets)
Xs = mean /average of a particular sample set (summation over 2)
Ns = number of items in a particular sample (in this case 5)

Within sample SS = $\Sigma\,(x_s - \bar{x}_s)^2$ (summation over 10 data points)
X_s = an item in a particular data set

Total SS = Between sample SS + Within sample SS

Ho: $\sigma^2_1 = \sigma^2_2$
Ha: $\sigma^2_1 < \sigma^2_2$

Here, your Ho will be

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

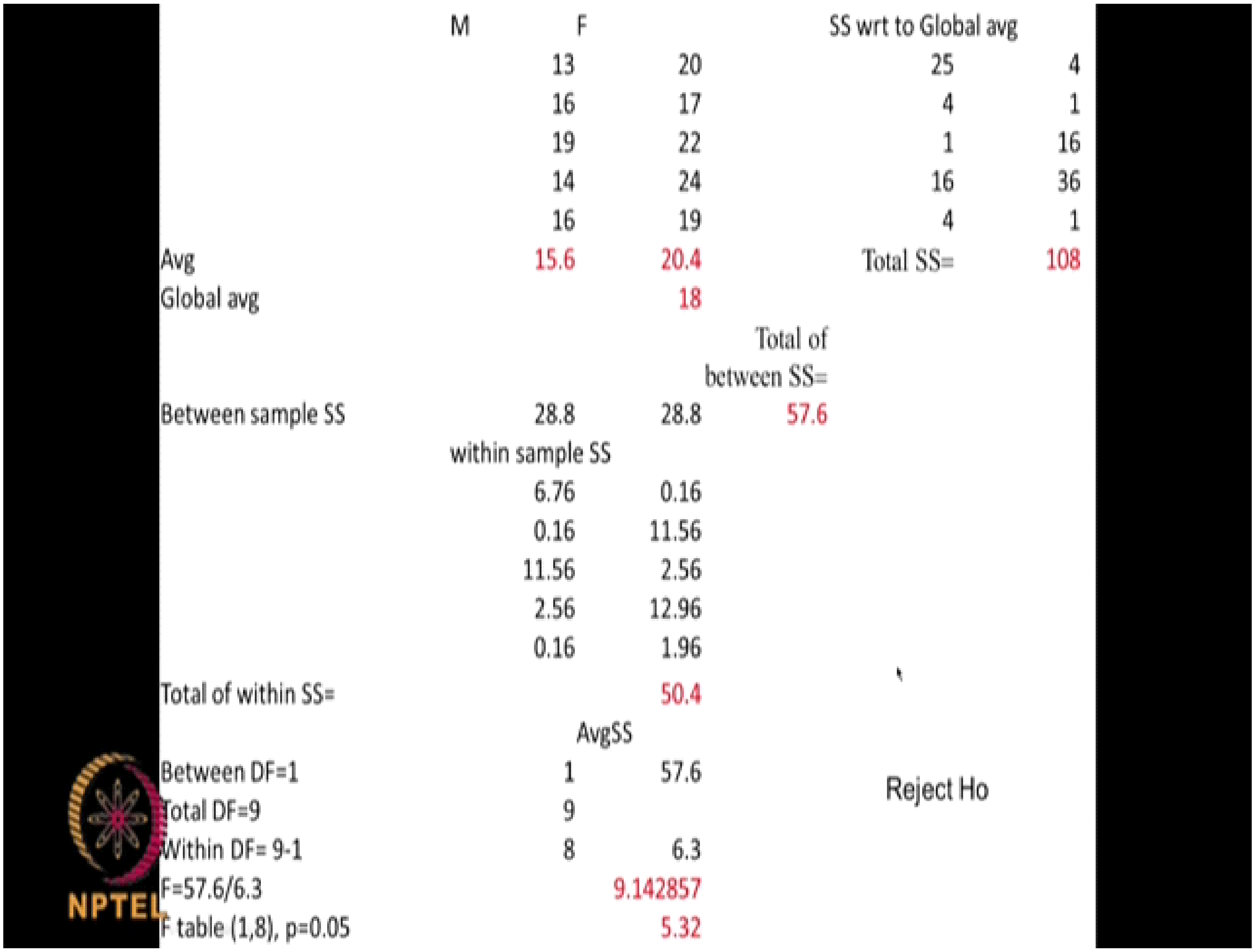
,

but then there is no difference, where as H a will be

$$\square_1^2 > \square_2^2$$

sigma 1 square less than sigma 2 square, that means male slept less then female, H a we can do at 95 or 99 depending upon what you want to take it up as.

(Refer Slide Time: 11:08)



	M	F			SS wrt to Global avg	
		13	20		25	4
		16	17		4	1
		19	22		1	16
		14	24		16	36
		16	19		4	1
Avg		15.6	20.4		Total SS=	108
Global avg			18			
				Total of between SS=		
Between sample SS		28.8	28.8	57.6		
	within sample SS					
		6.76	0.16			
		0.16	11.56			
		11.56	2.56			
		2.56	12.96			
		0.16	1.96			
Total of within SS=			50.4			
			AvgSS			
Between DF=1		1	57.6		Reject Ho	
Total DF=9		9				
Within DF= 9-1		8	6.3			
F=57.6/6.3			9.142857			
F table (1,8), p=0.05			5.32			

So that is what we are doing now. I took the same male data female data, average of male is 15.6, average of female is 20.4. The global average, that mean you add up all these and divide by 2 you get 18, so x double bar is 18, x double bar is 18. If I want to calculate between sum of squares, between sample sum of squares, so what do I do? I subtract from the global average in fact, that's what I am doing. So, 15.6 -18 ^{2} x 5 why do I need to do 5, because there 5 data sets here and this one how do I get I get 20.4 - 18 that is 2.4 x 5 that comes out to be 28. If I add that will give me total of between sum of squares here between the samples, understand ? 2 data sets will be there only.

So how do I get this? So I know the average of all the males, I know average of females, I know the global average that means total average so I will subtract 15.6 - 18, square it, multiply by 5, so this approximately so this approximately about 2.5 so 28.8, then we have 20.4 - 18, square it, multiply by 5, this will give you total of between sum of squares. How do I calculate total sum of squares? What is total sum of squares for each one of them, I subtract from the global mean, square it up. What do I do here this is the global mean so 13 - 18 ^{2}, 13 - 18 5, 5 x 5 25, you understand ? 20 - 18 is to 2, 2 ^{2}is 4, 16 -18 is to 2, 2 ^{2}4, 17 - 18 is 1, 1 ^{2} is 1, 19 - 18 is 1, 1 ^{2} is 1, 22 - 18 is 4, 4 ^{2} is 16, 14 -18 is 4, 4 ^{2} is 16, 24 - 18 is 6, 6 ^{2} is 36, 16 - 18 is 2, 2 ^{2} is 4, 19 - 18 is 1, 1 ^{2}, so I add up all these that will give you total sum of squares that is total sum of squares I got.

I got between sample sum of squares, I got total sum of squares.Now, within sample sum of squares how do I do? So within sample sum of squares as I said here so for each set of sample I will subtract from this and square it up, between sample, within sample I will subtract from the mean of each sample set and square it up right. So I will do 13 - 15.6 ^{2}, 16 ^{2} 15.6 s^{2}, 19 - 15.6 ^{2}, 14 - 15.6 ^{2}, 16 - 15.6 ^{2}. Whereas here I will do 20 - 20.4 ^{2}, 17 - 11.56 ^{2}, 22 - 2.56 ^{2} I will get if I add up all these, I will get total of within sum of squares. For between, you have a degrees of freedom 1, for total as I said there are 10 data points 5 male, 5 female so degrees of freedom is 9 so within will be 9 - 1 8. If I want to do average sum of squares for between what do I do this is between 57.6 / 1 is this if I have the total degrees of freedom is 9 within is 9 -1, 8 so 50.4 / 8 is 6.3. So we have between we have within and as I told you within is generally called error within is generally called error, between is actually looking at 2 different data sets. So what do I do, between divided by the error 57.6 / 6.3 I get 9.14.

Now what will be the F table degrees of freedom numerator is 1 degrees of freedom, denominator is 8 degrees of freedom so I look into the table of F 1, 8 for so I will look under 1 , 8, p = 0.05, 5.32, 5.32 F calculated 57.6 is / 6.3 is 9.1, so you reject the null hypothesis do you understand how to do this problem? It is not very difficult to do, it is quite simple to do actually, the most important point is we need to understand 2 important concepts, concept number 1 is we have one sample set, another sample set this case its male, female, it could be drug a, drug b, it could be anything operator 1, operator 2. So it will be called between and this variation is called within. So you are dividing between by within to get F value. You are dividing between and within to get F value, and then that is what you are comparing with the table.

Now your question is how do you calculate the between and within? So there is something called total sum of squares, total sum of squares is I take average of this data set, I take average of this data set and then I take an average of both these, that is average total of the entire or it can be will call it global average. So that global average subtracted from each one of these terms, square it up and add all of them, there will be 10 sets of data here that is called the total sum of squares. The degrees of freedom for total sum of squares will be, because I have 5, 5 10, 10 - 1 is 9, so the total degrees of freedom will be 9.

Now how do I calculate between samples what do I do? I know an average here so that is a average so he that average I subtract from each one of these terms, square it up, add up and then for the female, I have an average so I subtract from each one of these term, square it up, add up, then multiply by 5 here, because the 5 data points then add up the overall, that will give you me the between sample sum of squares now the degrees of freedom is 1 because I have 2 sets of samples or 2 groups, male group, female group so I will divide by 1.

Now how do I calculate? Sometimes, I need not calculate within sample also because in this particular case total sum of squares - between sample sum of squares will give me the within sample sum of squares and the degrees of freedom for this will be 8, because total will be 10 - 1, 9, between will be 2 - 1, 1. 9 -1 is 8. So within sample will have degrees of freedom at 8. How do I calculate within sample sum of squares? It is quite simple so within sample sum of squares I will subtract from each one of these terms here, and then square it up. I have an average, average minus global square and 5 term that will give me the between, and for within I will, from the average I keep subtracting for each one of them, square it up, it will be 10 terms because 5 for here, and 5 for here, that will give me the within samples sum of squares, understand?. These are the numerics of this so I have the time taken by the male rats in minutes, time taken by the female rats in minutes, I take the average, average is s sum it up, divide by 5, sum it up, divide by 5, now the global average will be sum it up, divided by 2.

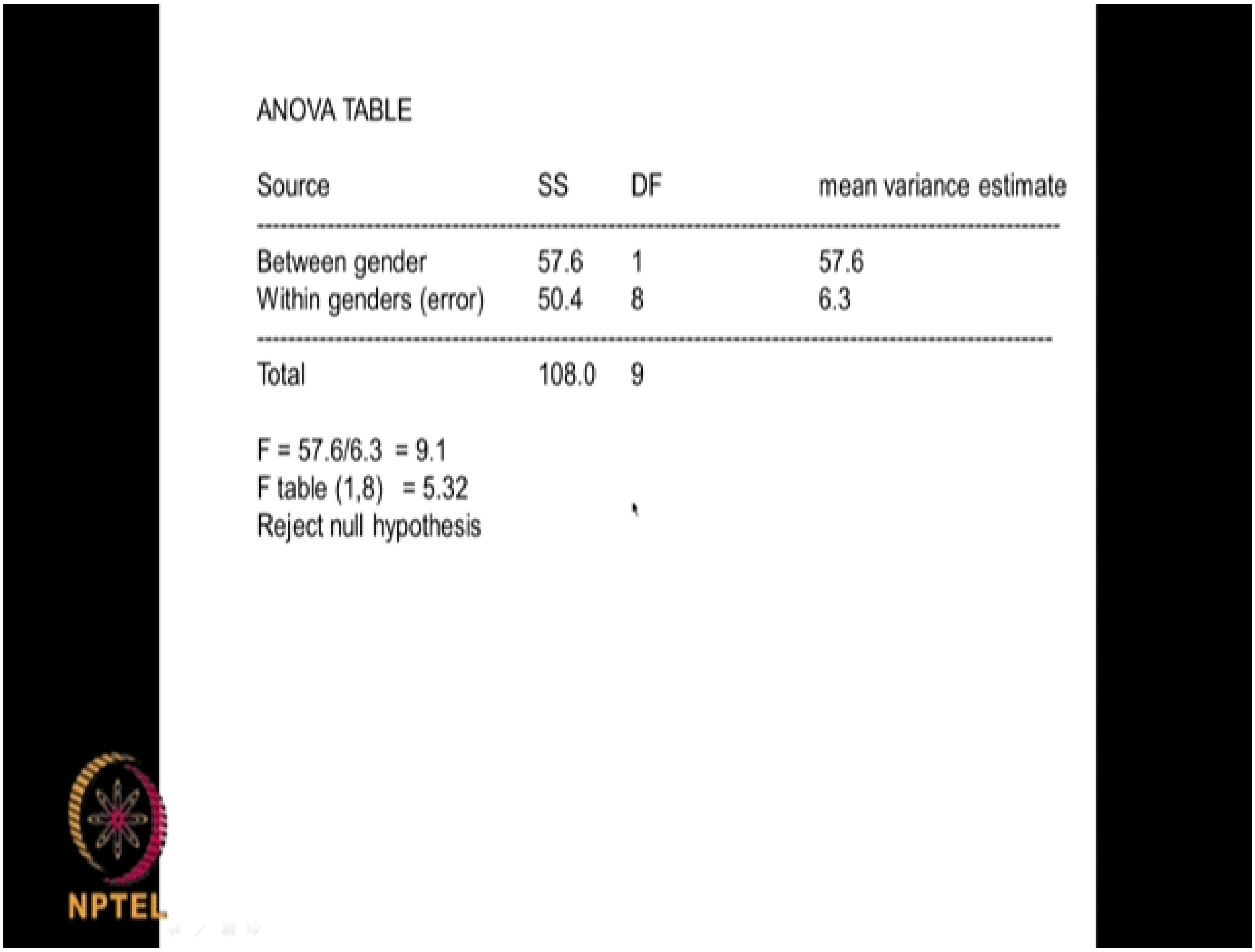
So between samples sum of squares how do I do? 28.8 - 15.6, sorry 18 - 15.6, square it up, and then so 18 - 15.6, square it up, multiply by 5. For the female 18 - 20.4, square it up, multiply by 5 this.So total of between this plus this, understand? We get 20.4 - 18, that comes to 1.6, 2.5 then multiply by 5 that should come here. Similarly, here 15.6 -18 that will come to 2.4, square it up and then multiply by 5, you will get this answer actually, then you add all these, this degrees of freedom will be 1 because we have male rats and female rats 2 data, 2 groups so 1 is the degree of freedom.

Now if I want to do the total sum of squares what do I do? Every point I subtract from the 18 because 18 is the global average, square it up, So 13 - 18 subtract, square, like that I do so I will get total for each then add up all, this will give you total sum of squares, degrees of freedom for total sum of squares is 9 because we have 10 data points. And then how do we do with within sample basically you do each one of them compare it with the group or sample average, so 15.6 - 13 ^{2}, 16 - 15.6 ^{2}, 19 - 15.6 ^{2}, like that, you get here it will be 20 - 20.4 ^{2}, 17 - 20.4 ^{2} and so on. You add up all of them that will be give you total of within. Now this is like an error sum of squares and this is like the group effect, where as this is like an error effect, actually if you add up these 2 you will get the total sum of squares.

Now between because we have 2 sets of groups we have degrees of freedom is 1, and for the total as I said degrees of freedom is 9, for within 9 - 1 is 8 and the F ratio will be 57.6 / 1, that is the average 57.6. Here it is, 50.4 / 8 that comes to 6 6.3. So, the F ratio is 57.6 / 6.3 that comes to be 9.14. If you go to the F table for p = 0.05, 1 degree of freedom for the numerator, 8 degrees freedom for denominator, it comes to 5.32. This table value is less than F value calculated so we reject the null hypothesis. So it is quite simple problem to do, the softwares Excel or the GraphPad will not be able to do, if you give the F value the GraphPad or Excel can calculate what will be the p value or if we give the p value it will give you the F value as 5.32, but this type of calculations it cannot do.

But there are commercial softwares which can do this type of calculation and tell you it is not very difficult, so you just put it on excel, we can do that this is called a one way ANOVA, because you are comparing male with the female rats, male rats with the female rats, that’s the one way ANOVA. The most important point which we need to keep in mind is the within sum of squares should be quite small, so that when you do the F ratio, you are doing the between divided by within it should be sufficiently large then only your value will be much larger than the table value. If the error sum of squares is very large or if the degrees of freedom for error is small also, your denominator term will become large so the F value will not be large. You will not be able to differentiate between 2 sets of group or 2 sets of sample, so you need to keep that point very much in mind. This is an interesting problem, which can be attacked with one way ANOVA. Now the same problem can be done as a two sample t test also because we have 2 sets of sample right so we can do it as a two sample t test um.

(Refer Slide Time: 24:47)



ANOVA TABLE

Source	SS	DF	mean variance estimate
Between gender	57.6	1	57.6
Within genders (error)	50.4	8	6.3
Total	108.0	9	

F = 57.6/6.3 = 9.1

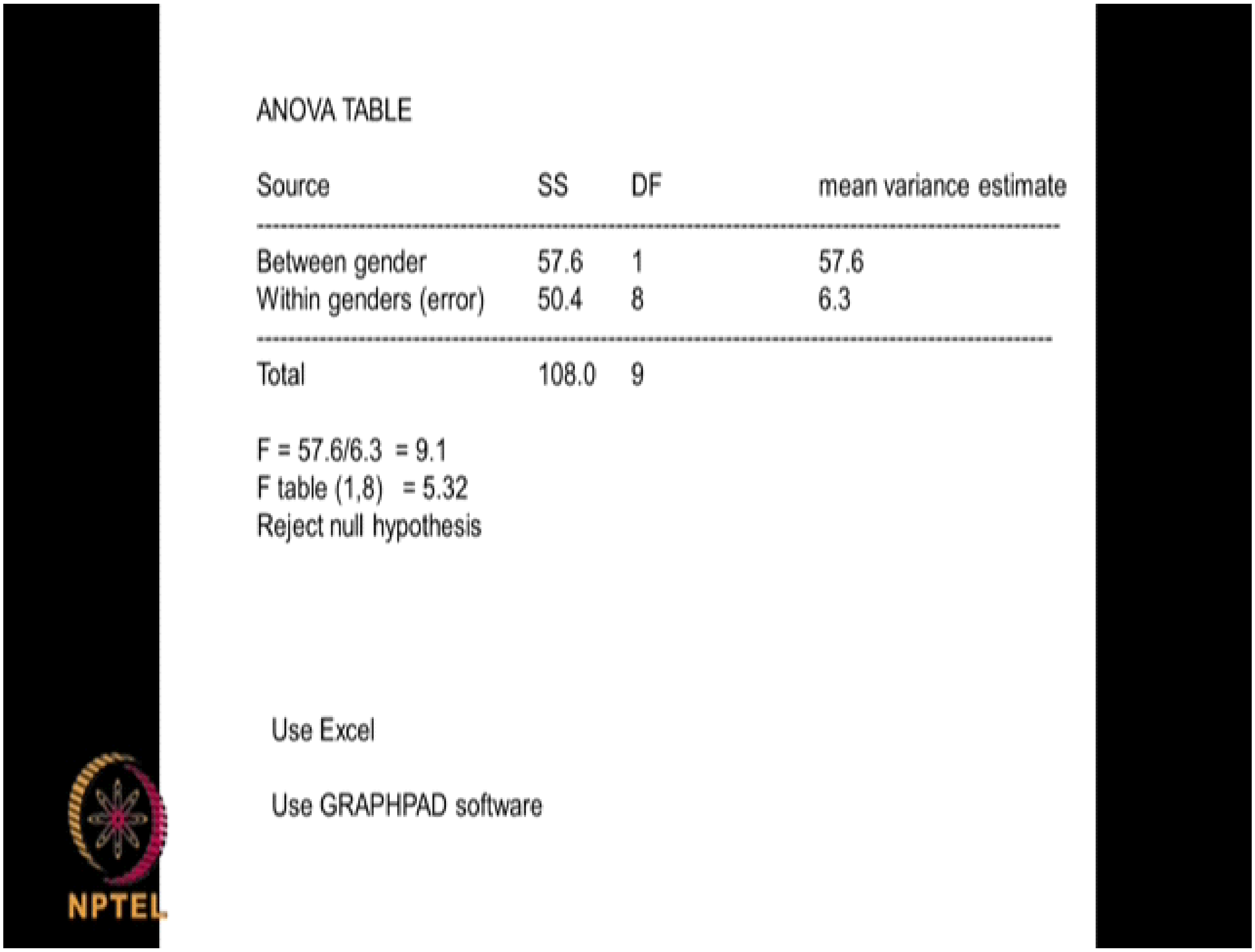
F table (1,8) = 5.32

Reject null hypothesis

Before that, I need to show this ANOVA table which is sort of a summary of whole thing, it is very nice. Between we have a degrees of freedom of 1 and the sum of squares is 57.6, sum of squares is 57.6, degrees of freedom 1, mean value is 57.6 / 1, 57.6, total is 108, degrees of freedom for total is 9, so total is 108, degrees of freedom is 9. Now within, that is within sum of squares is 50.4, degrees of freedom is 8, 50.4, degrees of freedom is 8. So the mean value is 50.4 / 8, 6.3, as you can see here they all add up, DF of gender and within gender it will come out to be 9, if you add up sum of squares these 2 will come up to 8. So this is called a summary table, this is very, very important, you will come across in these in many in the softwares hand, and when you do the F value your doing 57.6 / 6.3 that comes to be 9.1, the F table for 1 , 8 is 5.32, so we reject the null hypothesis.

After this preliminary calculation, we need to prepare this ANOVA table, this gives you summary and we can do this nice looking calculation, it tells you the whole story. And as you can see if I add up DF of a gender and error, within gender is called error, it will give you the total. Similarly, sum of squares if I add up between gender and within gender it gives the total. It's this is very important, even if you do this calculation which is in the background, we need to give the summary table, ANOVA table, which tells you in one short what is the status of your analysis. Now the same problem we can do it using 2 sample t test also, right? Two sample t test also.

(Refer Slide Time: 26:45)



ANOVA TABLE

Source	SS	DF	mean variance estimate
Between gender	57.6	1	57.6
Within genders (error)	50.4	8	6.3
Total	108.0	9	

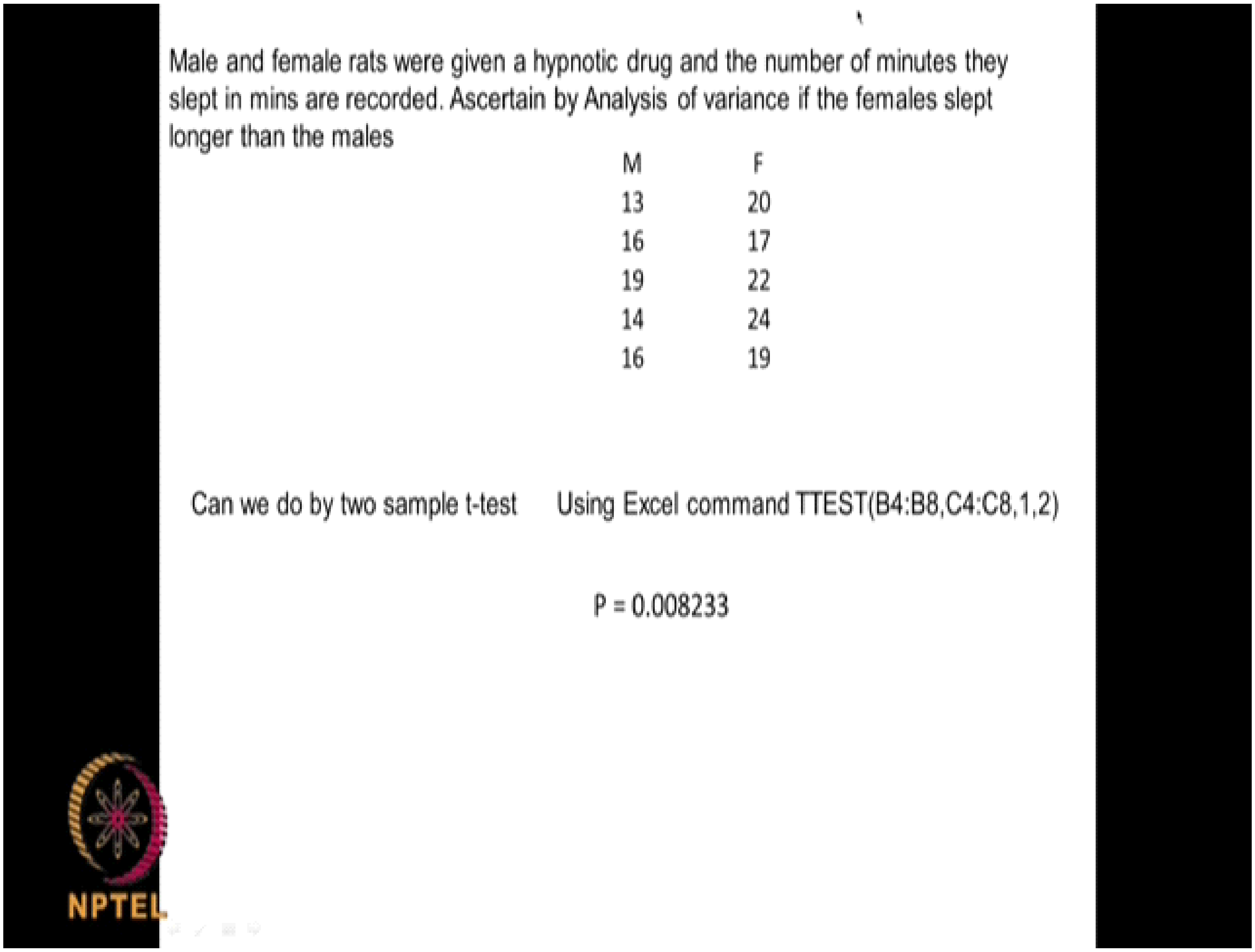
F = 57.6/6.3 = 9.1
F table (1,8) = 5.32
Reject null hypothesis

Use Excel

Use GRAPHPAD software

So, same problem it can do it in the two sample t test.

(Refer Slide Time: 26:56)



Male and female rats were given a hypnotic drug and the number of minutes they slept in mins are recorded. Ascertain by Analysis of variance if the females slept longer than the males

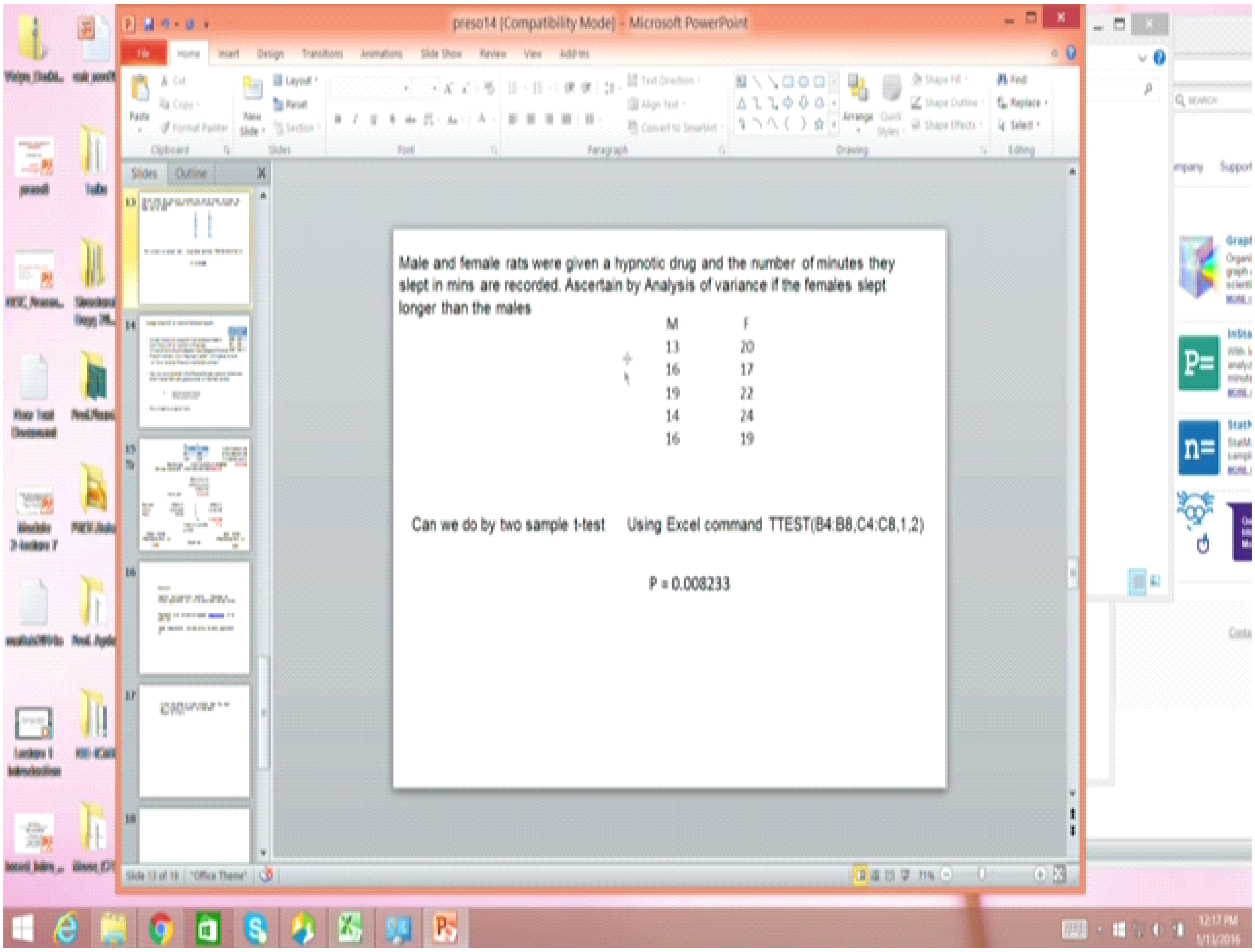
M	F
13	20
16	17
19	22
14	24
16	19

Can we do by two sample t-test Using Excel command TTEST(B4:B8,C4:C8,1,2)

P = 0.008233

Of course, we can use Excel or we can use both the commands and see what do you get.

(Refer Slide Time: 27:11)



Male and female rats were given a hypnotic drug and the number of minutes they slept in mins are recorded. Ascertain by Analysis of variance if the females slept longer than the males

M	F
13	20
16	17
19	22
14	24
16	19

Can we do by two sample t-test Using Excel command TTEST(B4:B8,C4:C8,1,2)

P = 0.008233

We use the Excel we can copy this control c, so we can do it here so we sayT t e s t array 1 , array 2 ,so one tailed test, because we want to know whether female rats sleep longer then type 1. 1 means paired, if you remember 2 means equal variance, 3 means unequal variance. Let us try 2 for example, that means equal variance, so it gives you 0.008, as our p value so obviously, p is very small reject the null hypothesis. Let us do the other one unequal variance also, this comma, this comma, one tailed comma, 3 that is unequal variance also we can do. So you can see you will get different p values but still in both the cases and p value is very small so we can reject the null hypothesis. We can use the GraphPad software also and check it out. The GraphPad software as you know that is the t test we do continue compare 2 means continue.

We will do this control C control Vthen again go control C control V. It is an unpaid t test it is not a paid t test calculate now so obviously as I said GraphPad I gives you in only a 2 tail test but it is giving you p value. We can divide this p by 2 for a 1 tail. So that comes out to be 0.0082 that is what we also got it right 0.0082. So the difference is considered to be statistically significant. We can use 2 sample t test also and solve this problem. But I am later on going to show you that sometimes ANOVA is much more accurate in some situations then 2 sample t test but this particular problem in whether use ANOVA whether we use 2 sample t test of equal variance or unequal variance we see that p value is very, very small 0.008 in that order 0.0082 like so we can reject the null hypothesis. But in some cases we find ANOVA may be more accurate. So I talked about one way ANOVA which is very useful quite powerful we can compare as different groups in this particular case the groups were male and female. Degrees of freedom was 1, we are talking about male male and female. So we have a sorry male and female the degrees of freedom is 1 then we calculated total sum of squares which is calculated from the global mean x double bar is global mean. So we will ma say 13 - global mean square, 16 - global mean square and so on that will give you a total sum of squares the degrees of freedom for total sum of squares is 9 because there are 10 data points. When we have 2 sets of samples or groups here the degrees of freedom is 2 -1 is 1. If I want to calculate between sum of squares sample sum of squares what I will do I will have a mean here. That mean minus global mean square but I will multiply by 5 here because there are 5 data point then this mean minus global mean square multiply by 5, add both that will give me the between sample sum of squares and the degree of freedom is 1 here.

Now we want to calculate within sample sum of squares I will have a mean for male alone every time I have subtract from each 1 of these data points, square it up, summit up, then for the female I will have a mean here and subtract from each 1 of these data point, square it up and summit up, over all summation. Here I will do that this sum of the degrees of freedom for within sample is the total is 9 between sample is 1 so within sample is 9 - 1 is 8. I have the ANOVA table so total I told you how to calculate the degrees of freedom is 9 for between sample or between gender sum of squares I know I have only 1 degree of freedom I divide this term by this term to get mean variance estimate within gender, the degrees of freedom is 8 I know how to calculate I told you divide by 8 I will get this F value is ratio of this by this and F table is given by 5.32 for 1 and 8 degrees of freedom then I compare the F table value with the F value calculated and I in this particular case I will say I will reject the null hypothesis. So this one way ANOVA is very powerful. Instead of male female I could also have an infant. So there I have 3 groups 2 sample t test is very difficult to do whether I take male and female alone, female and infant alone, male and infant alone and do two sample t test or I have to use a one way ANOVA. So I have 3 groups. So the degrees of freedom will be 3 - 1 2 in this particular case. I can use these ANOVA which is very powerful to analyze that sort of data. So the H_0 will be

$$Ho: \sigma^2_1 \ =\sigma^2_2$$

to sigma 3 square that means 1 means male, 2 means female, 3 could be infant and H_a could be if I am looking at some difference it could be

$$Ha: \sigma^2_1 \ < \sigma^2_2$$

one of them is not equal to. So let us continue with the more of this 1 way ANOVA and we look at 1 more problem in the next class.

Thank you very much.

Key words: Variance, Null hypothesis, between group variance, within group variance, total sample sum of squares, between sample sum of squares, within sample sum of squares, global mean, alternate hypothesis, degrees of freedom, error, ANOVA, ANOVA table.