Chi Square Test and Analysis of Variance

S C Agarkar

V N BRIMS, Thane

Preamble

- Chi square test is to be used when we have more than two samples to compare. Chi square tests enable us to test whether more than two population proportions can be considered equal.
- If we classify a population into several categories with respect to two attributes (e. g. age and job performance) we can then use a chi square test to determine if two attributes are independent of each other.

Contingency tables

- The data collected is often tabulated putting the values in columns and rows. The table is described by a number of rows and columns.
- National Health Care Company samples its hospital employees' attitude toward job performance reviews (new quarterly reviews over present six monthly reviews).
- The data is collected from four different regions (northeast, southeast, central and west coast). This leads to a 2*4 contingency table as shown in next slide.

Sample response concerning review schedules

	Northeast	Southeast	Central	West coast	Total
Number who prefer present method	68	75	57	79	279
Number who prefer newer method	32	45	33	31	141
Total employees sampled ineach region	100	120	90	110	420

Observed and expected frequencies

- Suppose we symbolise the true proportions of samples as pN, pS, pC and pW.
- Null hypothesis would be H_0:pN = pS = pC = pW.
- Alternative hypothesis would be H1: They are not equal.
- If the null hypothesis is true we can estimate the proportion of the total workforce as (68+75+57+79)/ (100+120+90+110) = 0.664.
- If 0.664 is the estimate of population proportion who prefer the present method then (1-0.664) = 0.336 is the estimate of population proportion who would prefer the new method.

Proportion of sampled employees in each region preferring two methods

	Northeast	Southeast	Central	West coast
Total Number sampled	100	120	90	110
Estimated proportion preferring present method	X 0.664	X 0.664	X 0.664	X 0.664
Number expected to prefer present method	66	80	60	73
Total Number sampled	100	120	90	110
Estimated proportionpreferringnewmethod	X 0.336	X 0.336	X 0.336	X 0.336
Number expected to prefer new method	34	40	30	37

Comparing Observed and Expected Frequencies

	Northeast	Southeast	Central	West coast
Observed Frequency (Present method)	68	75	57	79
Expected Frequency (Present method)	66	80	60	73
Observed Frequency (New Method)	32	45	33	31
Expected Frequency (New Method)	34	40	30	37

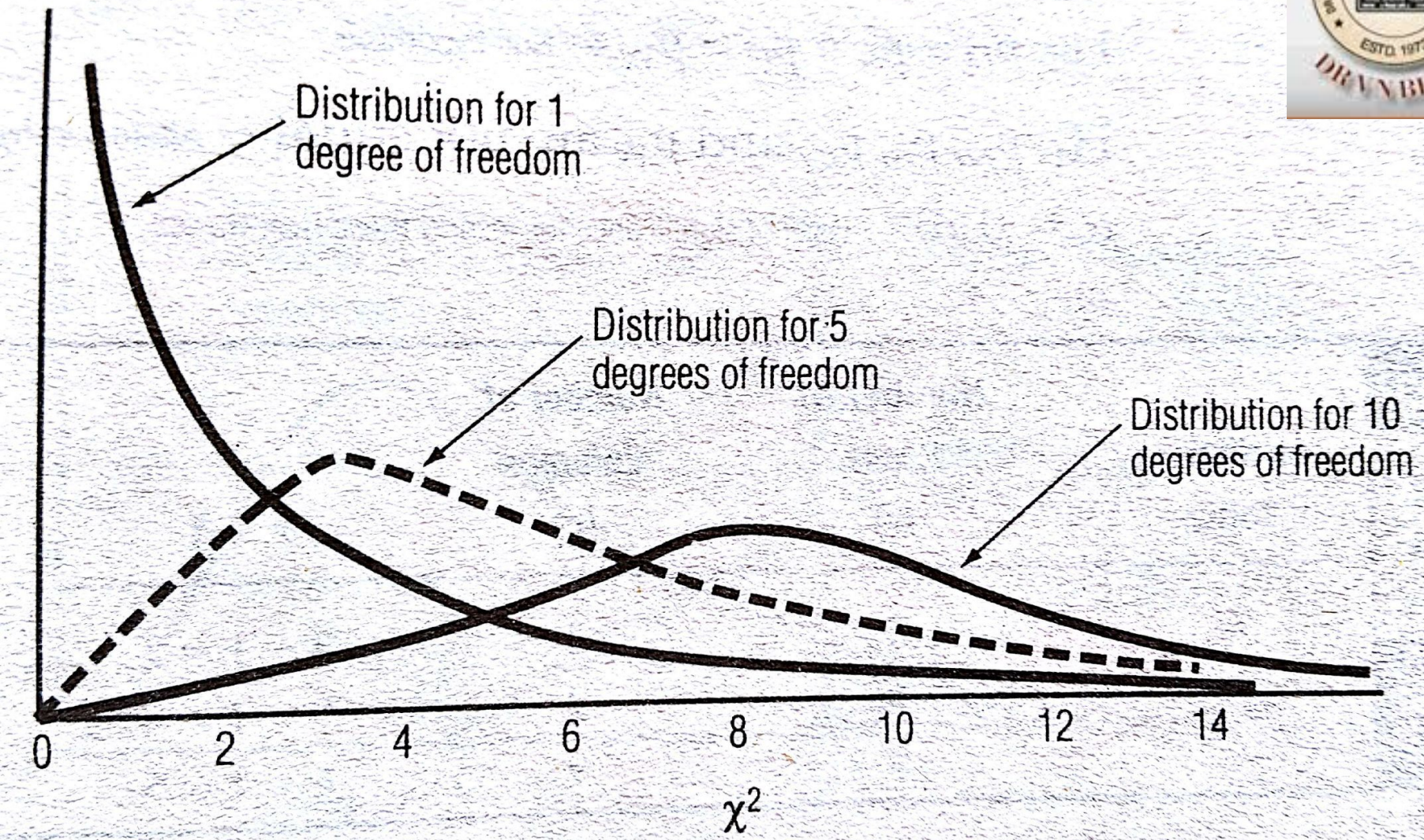
If the sets of observed and expected frequencies are nearly alike, we can reason intuitively that we will accept the null hypothesis.

Calculation of Chi square

f_0	f_e	f_0-f_e	$(f_0-f_e)^2$	$(f_0-fe)^2/f_e$
68	66	02	04	0.0606
75	80	-05	25	0.3125
57	60	-03	09	0.1500
79	73	06	36	0.4932
32	34	-02	04	0.1176
45	40	05	25	0.6250
33	30	03	09	0.3000
31	37	-06	36	0.9730

$x^2 = \Sigma (f_0 - f_e)^2 \div f_e = 3.032$

If this value was as large as 20 it would indicate a substantial difference between observed values and expected Values. A chi square of zero, on the other hand, indicates that observed frequencies exactly match the expected frequencies.



11.2 Chi-Square as a Test of In

Degrees of freedom

- In the contingency table of 3*4 the degrees of freedom are 6 (2x3) . In a contingency table of 6*9 the degrees of freedom are 40 (5x8). In general the degrees of freedom are (r-1)x(c-1).

	column1	column2	Column3	coulmn4	
Row1	yes	yes	yes	No	RT1
Row2	yes	yes	yes	No	RT2
Row3	No	No	No	No	RT3
	CT1	CT2	CT3	CT4	

Using Chi Square test

- Calculate the value of Chi square.
- Determine the degree freedom.
- Decide significance level.
- Find the value of the Chi square statistic.
- If the calculated value is less than tabulated value then the null hypothesis can be accepted.
- In the job review problem the calculated value is 3.032 and the tabulated value is 6.251. Hence we accept that there is no difference in four geographical regions.

Area in Right Tail					Degrees of Freedom
0.20	0.10	0.05	0.025	0.01	
1.642	2.706	3.841	5.024	6.635	1
3.219	4.605	5.991	7.378	9.210	2
4.642	6.251	7.815	9.348	11.345	3
5.989	7.779	9.488	11.143	13.277	4
7.289	9.236	11.070	12.833	15.086	5
8.558	10.645	12.592	14.449	16.812	6
9.803	12.017	14.067	16.018	18.475	7
11.030	13.362	15.507	17.535	20.090	8
12.242	14.684	16.919	19.023	21.666	9
13.442	15.987	18.307	20.483	23.209	10
14.631	17.275	19.675	21.920	24.725	11
15.812	18.549	21.026	23.337	26.217	12
16.985	19.812	22.362	24.736	27.688	13
18.151	21.064	23.685	26.119	29.141	14
19.311	22.307	24.996	27.488	30.578	15
20.465	23.542	26.296	28.845	32.000	16
21.615	24.769	27.587	30.191	33.409	17
22.760	25.989	28.869	31.526	34.805	18
23.900	27.204	30.144	32.852	36.191	19
25.038	28.412	31.410	34.170	37.566	20
26.171	29.615	32.671	35.479	38.932	21
27.301	30.813	33.924	36.781	40.289	22
28.429	32.007	35.172	38.076	41.638	23
29.553	33.196	36.415	39.364	42.980	24
30.675	34.382	37.652	40.647	44.314	25
31.795	35.563	38.885	41.923	45.642	26
32.912	36.741	40.113	43.194	46.963	27
34.027	37.916	41.337	44.461	48.278	28
35.139	39.087	42.557	45.722	49.588	29
36.250	40.256	43.773	46.979	50.892	30

Problem to solve

- A brand manager is concerned that her brand share may be unevenly distributed throughout the country. In a survey in which the country was divided into 4 geographic regions. A random sampling of 100 consumer in each region was surveyed with the following results
- A) calculate chi square (1.723)
- B) state the null and alternative hypothesis.
- C) using the 0.05 level of significance, should the null hypothesis be rejected?

	A	B	C	D	Total
Purchase the brand	47	52	43	49	191
Do notpurchase the brand	53	48	57	51	209
Total	100	100	100	100	400

Contingency table with more than two rows

		<5	5-10	>10	Total
Fraction of cost	<25%	40	75	65	180
	25-50%	30	45	75	150
	>50%	40	100	190	330
	Total	110	220	330	660

The table shows data related to hospital stay classified by the type of insurance coverage and length of stay. The expected frequency of any cell is given by the equation fe= RT X CT/n, where RT is row total and CT is column total and n is the total number of observations.

Row	Column	fo	fe	fo-fe	(fo-fe)2	(fo-fe)2/fe
1	1	40	30	10	100	3.333
1	2	75	60	15	225	3.750
1	3	65	90	-25	625	6.944
2	1	30	25	05	25	1.00
2	2	45	50	-05	25	0.500
2	3	75	75	00	00	0.000
3	1	40	55	-15	225	4.091
3	2	100	110	-10	100	0.909
3	3	190	165	25	625	3.788

$$x^2 = \Sigma (fo-fe)^{2}/fe = 24.315$$

The acceptance region of the chi square value is 9.488. Since this value is more than that the null hypothesis is rejected.

Problem to solve

- To determine whether different income groups have different purchasing habits concerning a certain brand, a marketing researcher asked 4 income groups, Do you always, never or sometimes purchase the brand? If the results of the survey are as given in the table below should null hypothesis be accepted or rejected at 0.10 significance level?

	<$7,000	$ 7000-12999	$13000-19999	$20000+	Total
Always	25	40	47	46	158
Never	69	51	74	57	251
Sometimes	36	29	19	37	121
Total	130	120	140	140	530

Goodness of fit

- The chi square test can also be used to decide if a particular probability distribution such as Binomial, Poisson or Normal is the appropriate distribution for the data under consideration.
- We can obtain theoretical distribution of data and a distribution of observed data. Then using chi square test we can see if there is a goodness of fit among them.

The data table

- Assume that 100 candidates have been interviewed by three different executives. The result of the interview in terms of their score is shown in the following table.

Possible positive ratings from three interviewers	Number of candidates receiving each of these ratings
0	18
1	47
2	24
3	11

Calculating expected frequencies

- If it is assumed that the interview process can be approximated by a binomial distribution with p=.4 (40 % chance of any candidate getting positive rating). To test the hypothesis at .20 level of significance we must calculate chi square.
- For that we must calculate the expected frequencies using the cumulative binomial distribution table (these values are .7840, .3520 and .0640).

Binomial Probabilities and expected frequencies

Positive ratings	Binomial probabilities of these outcomes
0	1.0 - 0.7840 = 0.2160
1	0.7840 - 0.3520 = 0.4320
2	0.3520 - 0.0640= 0.2880
3	0.0640

Possible Rating	Observed frequency	Possible outcomes	Numberof candidates	Expected Frequency
0	18	0.2160	100	22
1	47	0.4320	100	43
2	24	0.2880	100	29
3	11	0.0640	100	06

Calculating Chi Square value

Observed Frequency	Expected Frequency	fo-fe	$(fo-fe)^2$	$(fo-fe)^2/fe$
18	22	-4	16	0.7273
47	43	4	16	0.3721
24	29	-5	25	0.8621
11	06	6	25	4.1262

$$X^2 = \Sigma (fo-fe)^{2}/fe = 6.1282$$

The chi square value under 0.20 column corresponding to 3 degrees of freedom is 4.642. Since the above value is larger than that we reject the null hypothesis. It means the binomial distribution with p = 0.4 is not a good description of our observed frequencies.

Problem to solve

- At the 0.10 level of significance, can we conclude that the following 400 observations follows a Poisson distribution withλ =3?

- no. of arrivals per hour 012345 or more
- no. of hours205798857862
- Hint:
- Calculate Poission probability values for different cplumns.
- Calculate estimated frequencies
- Calculate chi square value
- At 0.10 level of significance and at 5 degrees of freedom the chi square value is 9.236.
- **Ans: the data is well described by Poisson distribution.**

Analysis of Variance

- Analysis of variance (abbreviated as ANOVA) enables us to test the significance of the difference between more than two sample means.
- ANOVA would be useful in situations like comparing the mileage achieved by five different brands of gasolene, testing which of four different testing methods produces the faster learning record or comparing the earnings of graduates of different business schools.

The situation

- The training director of a company is trying to evaluate three different methods of training new employees. The first method assigns each to an experienced employee for individual help in the factory, The second method puts all new employees in a training room separate from the factory and the third method uses training films and programmed learning materials. The training director chooses 16 new employees assigned at random to 3 training methods and records their daily production after they complete the programme. Based on the data (given below) help the director to find out if there are differences in effectiveness among these methods.

Method 1	**15**	**18**	**19**	**22**	**11**	**-**
Method 2	22	27	18	21	17	-
Method 3	18	24	19	16	22	15

Basic concepts in ANOVA

- Analysis of variance is based on a comparison of two different estimates of the variance (σ^2) of our overall population. We can calculate one of these estimates by examining the variance among the three sample means (Between-column variance).
- The other estimate of the population variance is determined by the variation within the three samples themselves (Within-column variance).
- Compare these two estimates of population variance. Since both are estimates of σ^2 they should be approximately equal in value when the null hypothesis is true. If the null hypothesis is not true these two estimates will differ considerably.

Steps in ANOVA

- Determine one estimate of the population variance from the variance among the sample means.
- Determine a second estimate of the population variance from the variance within the samples.
- Compare these two estimates. If they are approximately equal in value, accept the null hypothesis.
- Going back to the problem given earler we see that $x_1 = 17$ $x_2 = 21$ $x_3 = 19$; $n_1 = 5$, $n_2 = 5$ and $n_3 = 6$
- The grand mean$x$

$= (5/16)*17+(5/16)*21+ (6/16)*19=304/16= 19$

Between column variance

n	X	x	X-x	$(X-x)^2$	$n(X-x)^2$
5	17	19	-2	4	20
5	21	19	2	4	20
6	19	19	0	0	00

$\sigma^2 = \sum n_j (X-x)^2 / k-1$

σ^2= estimate of the population variance based on the variance among

sample means called between-column variance

n_j= the size of the jth sample, X = the sample mean of jth sample

x= the grand mean and K= the number of samples

Substituting the values in the above formula we get

$\sigma^2 = 40/ 3-1$

= 40/2

=20

It means the value of between-column variance = 20

Within-column variance

- The sample variation s^2 is given by the formula
- $s^2 = \sum (X-x)^2/n-1$
- Hence $s_1^2 = 70/ 5-1 = 17.5$;
- Similarly, $s_2^2 = 15.5$ and $s_3^2 = 12.0$
- Within column variance σ^2 is given by the formula
- $\sigma^2 = \sum (n_j-1/n_T-k)s_j^2$
- $= 4/13\ (17.5) + (4/13)\ (15.5) + (5/13)\ (12.0)$
- $= 192/13$
- $= 14.769$

The F Statistic

- Once two estimates of population variance are calculated then their ratio (called F) is taken.

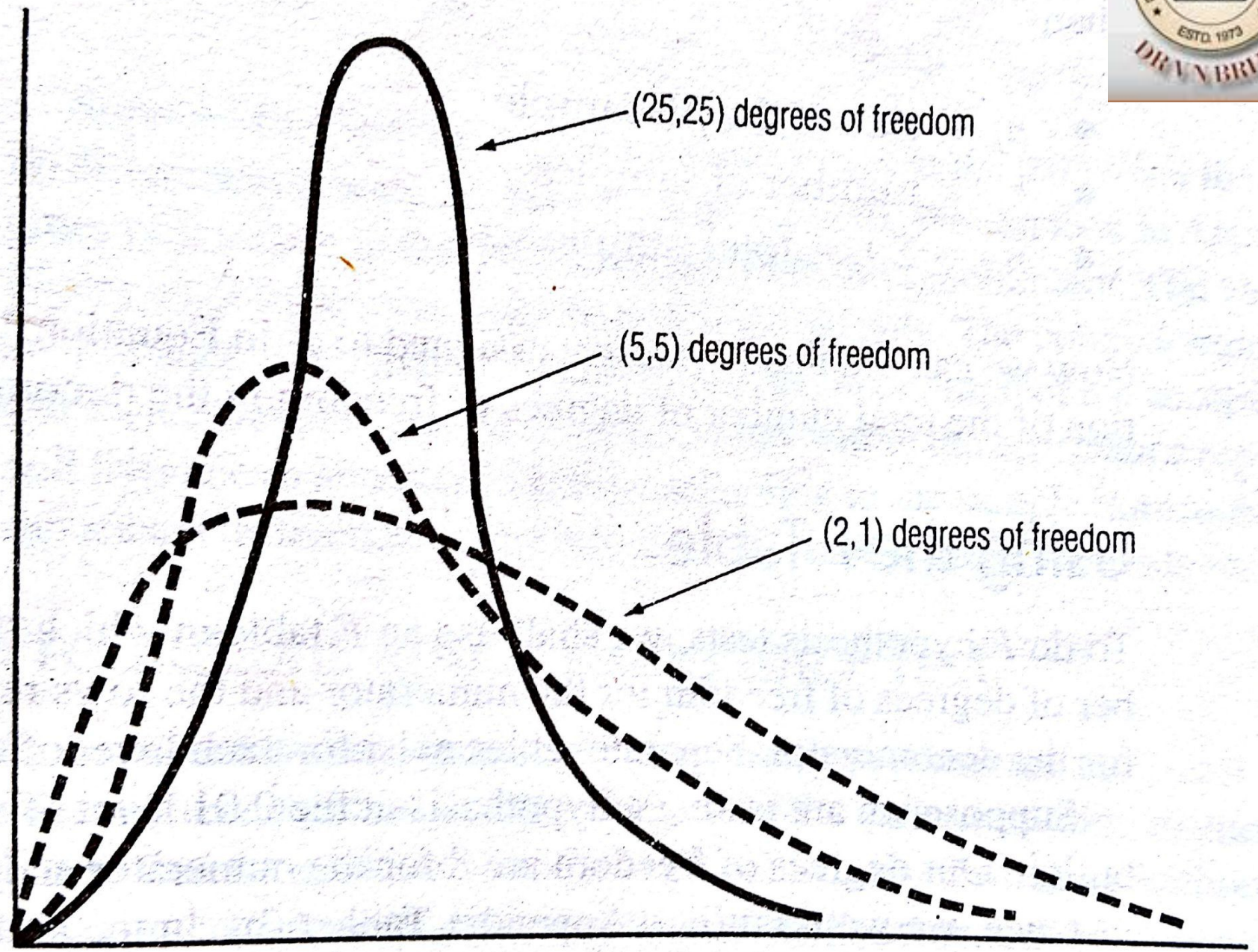
F = Between-column variance/Within-column variance

F = 20/14.769 = 1.354

- In the above formula the numerator is the variation among sample means of the three methods. It is a good estimator of population variance. The denominator is based on the variance within the samples. It is also a good estimator of population variance.
- The numerator and denominator should be about equal if the null hypothesis is true.

The F Distribution

- Like the t distribution, the F distribution is a whole family of distributions.
- Each distribution is identified by a pair of DF (degrees of freedom). Recall that for t distribution, there is only one DF.
- The first number refers to the number of DF in the numerator of F ratio while the second to DF in the denominator.
- F distribution has a single mode. They are usually skewed towards the right but become symmetrical as both the DFs increase.


(25,25) degrees of freedom
(5,5) degrees of freedom
(2,1) degrees of freedom

Calculating DFs

- The numerator of the F ratio is the value of between column variance. In our calculation we used three values to compute it. Once we know two of them the third is automatically determined. Thus the no. of DF for numerator in F ratio is one less than the number of samples (s-1).
- For calculating the denominator (within column variance) of a F ratio we used all three samples. In each sample one DF is lost. Hence no. of DF for denominator is ($n_T - k$), that is 16-3=13

Using F table

- To test the hypothesis using F test we have to use values given in F tables. Separate tables exists for each level of significance. In these tables columns show the DFs for numerator while rows show DFs for denominator.
- In the table prepared for 0.01 level of significance we find that the value corresponding to DFs 2 and 13 is 6.70.

Degrees of Freedom for Denominator	Degrees of Freedom for Numerator																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40			
1	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,982	6,023	6,056	6,106	6,157	6,209	6,235	6,261	6,287	6,313	6,339	6,36
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Testing the hypothesis

- Suppose the director of training wants to test at 0.01 level of significance the hypothesis that there is no difference among the three training methods.
- From the table we found that value of F ratio for DF 2 and 13 as 6.70. It sets the upper limit of the acceptance region.
- Through calculation we found the value of F ratio as 1.354. Since this value lies in the acceptance region, the director can accept the null hypothesis.
- The conclusion is according to the sample information there is no difference in the effects of three training methods on employee productivity.

Use of Computers for F test

- For convenience we had taken a smaller sample in our problem. In actual practice the size may be quite large and calculation would be tedious. In such cases use of computers is advocated.
- A software package for social sciences called SPSS (Statistical Packages for Social Sciences) is developed. Using SPSS one can undertake ANOVA easily.

Problem to solve

- The manager of an assembly line in a clock manufacturing plant decided to study how different speeds of the conveyor belt affect the rate of defective units produced in an 8 hour shift. To examine this, he ran the belt at 4 different speeds for 8 hour shifts each and measured the number of defective units found at the end of the shift. The result is given below.
  - Sp 1: 36,34,37,35,33; Sp 2: 29,34,34,36,32
  - Sp 3: 31,35,32,33,39; Sp 4: 36,28,34,32,30
- Calculate the mean number of defective units for each speed, determine the grand mean (33.5), estimate between column variance, (8.333), calculate within column variance, (7.375), calculate F ratio (1.130). At the 0.05 level of significance do four different conveyor belt speeds produce the same rate of defective clocks per shift? (Ans. Yes, the same).

One more problem to solve

- The study compared the effects of 4 one- month point-of-purchase promotions sales. Below are the unit sales for 5 stores using all 4 promotions in different months,

Free sample: 77 86 80 88 84

One-pack gift: 95 92 88 91 89

Cents off:72 77 68 82 75

Refund by mail: 80 84 79 70 82

- Calculate the mean unit sale for each promotion and then determine the grand mean (81.95), estimate between-column variance, (238.05), calculate within-column variance (21.05), calculate F ratio (11.31). At the 0.05 level of significance do the promotions produce different effects on sale? (Ans. Yes, different).

One more problem to solve

- A research company has designed three different systems to clean up oil spills. The following table contains the results, measured by how much surface are in sq meters is cleared in 1 hour.

System a 556063565955

System B 575364 49 62

System C 6652 61 57

The data were found by testing each method in several trials. Are the three systems equally effective? Use 0.05 level of significance.

Hints: Calculate between column variance (4.4667), calculate within-column variance (26), calculate F ratio (0.17). At the 0.05 level of significance the critical value of F ratio is 3.89.

Thank you. Remember, there is so much variation in nature.