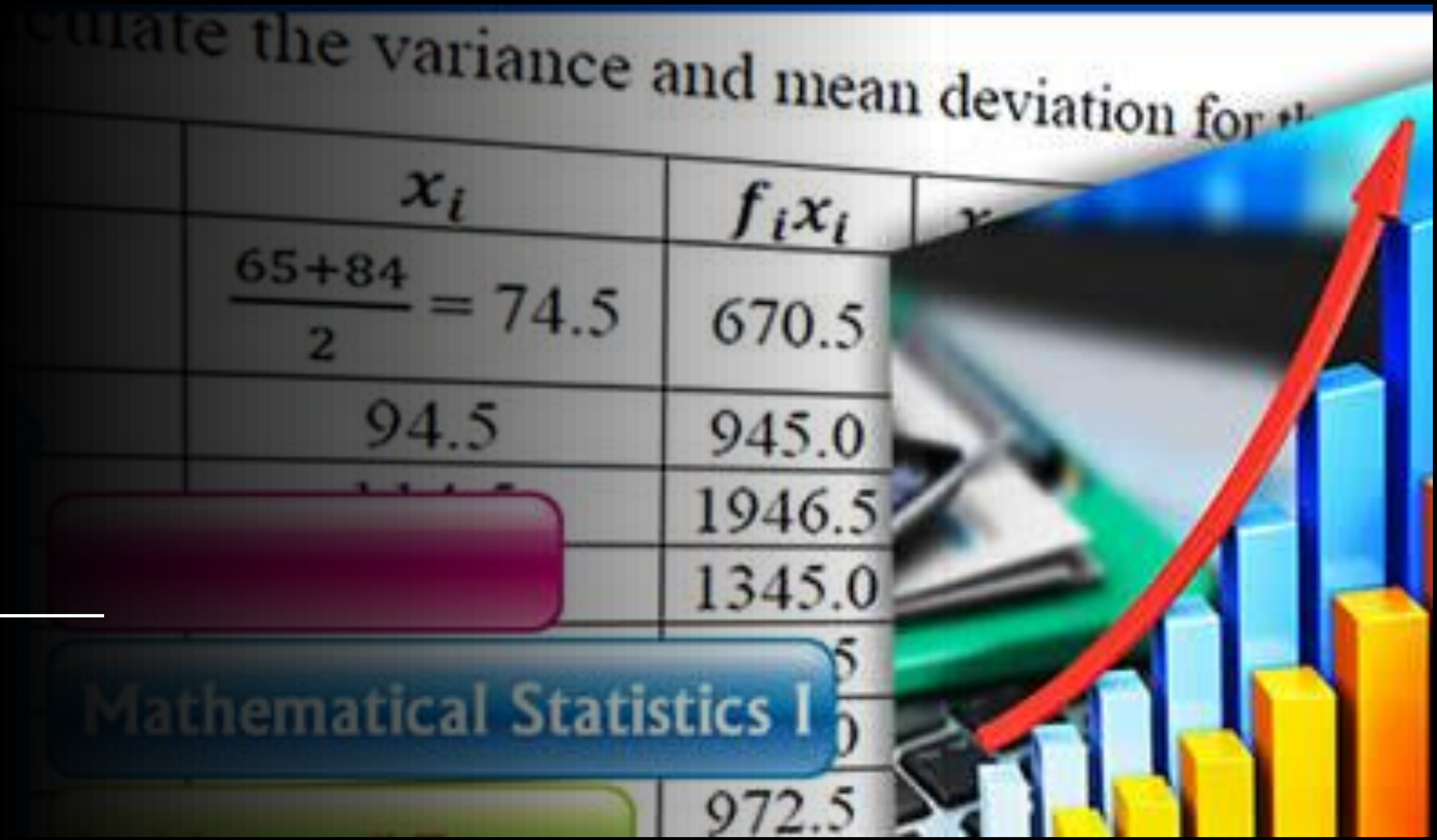
Hypothesis testing – Correlation analysis

Bindu K R


culate the variance and mean deviation for th
x_i
$f_i x_i$
$\frac{65+84}{2} = 74.5$
670.5
94.5
945.0
1946.5
1345.0
972.5
Mathematical Statistics I

Today's lecture

- Chi-square test
- Terminology
- Goodness of fit
- Example of the Chi-square test
- Performing the Chi-square test in spreadsheets
- Pros and Cons
- Usecases
- Conclusion

Chi-square Test(χ2 test)

- Chi-square test is a non-parametric test (a non-parametric statistical test is a test whose model does not specify conditions about the parameter of the population from which the sample is drawn.).
- It is used for identifying the relationship between a categorical variable and denoted by χ2.
- 1900 **Karl Pearson** developed published a paper on the χ2 test

Chi-square Test(χ2 test)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

- It is a statistical test that can be used to determine what observed frequencies are significantly different from expected frequencies or not in one or more categories.
- In the mathematical expression, it is the ratio of experimentally observed result/frequencies (O) and the theoretically expected results (E) based on certain hypotheses, or it is calculated by dividing the overall deviation from the observed and expected frequencies by the expected frequencies.

Chi-square Test(χ2 test)

If there is no difference in observed and expected frequencies, then the chi-square value would be zero.

If there is a difference, then the value of chi-square would be more than zero.

Chi-square Test(χ2 test)

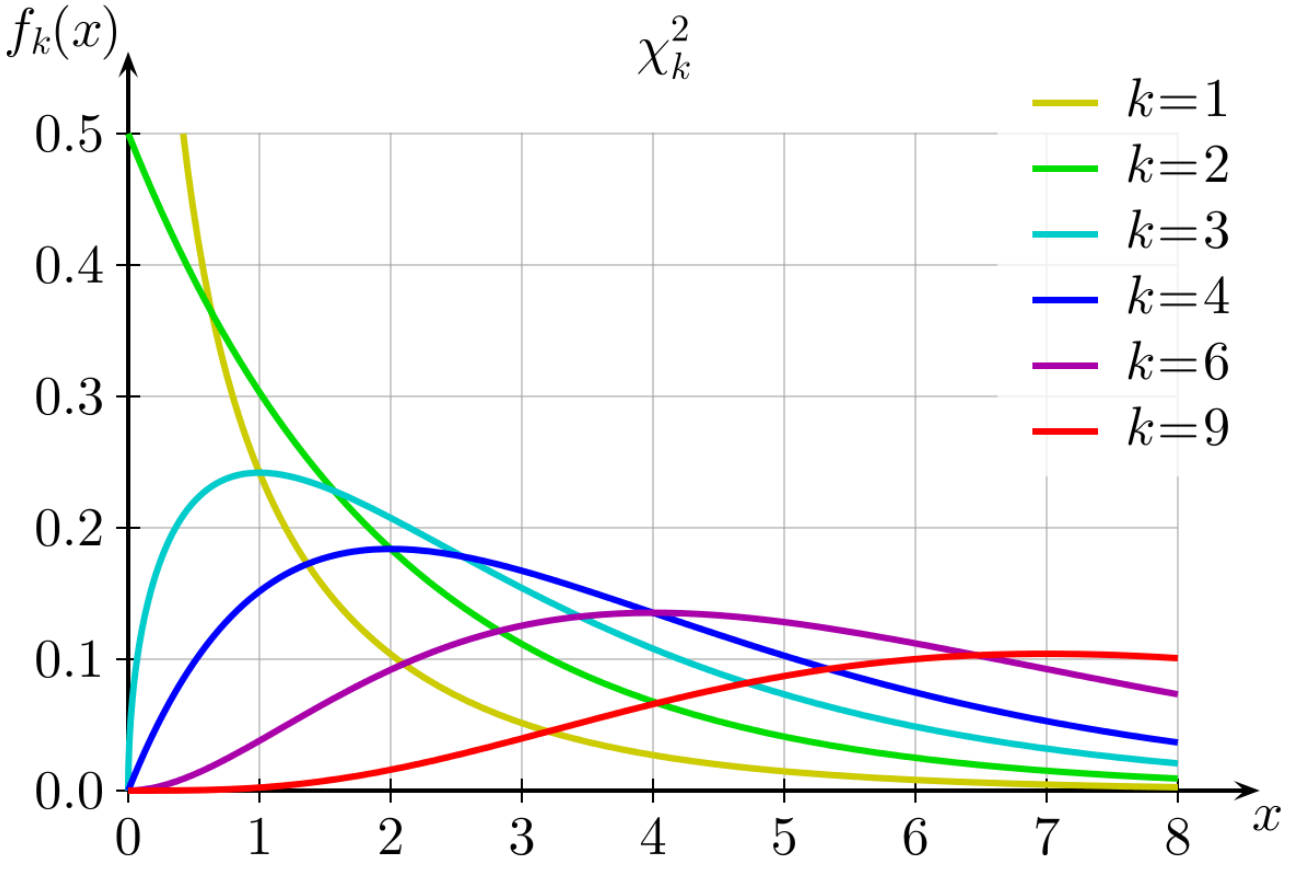
Three types of chi-square tests:

Goodness of fit

Test of independence

Test of homogeneity

|



Chi-square probability distribution graph

Terminology

- **Contingency table:** This is a cross table or two-way table.
- Its used to show the one variable in a row and another in a column with their frequency count.
- It is a type of frequency distribution table of the **categorical variables**.

Terminology

- **Observed frequencies:** Are counts made from experimental data. In other words, you observe the data happening and take measurements.
- **Expected frequencies:** Are counts calculated using probability theory. Expected frequencies are calculated for each cell in the contingency table.

Terminology

- **Observed frequencies:** Are counts made from experimental data. In other words, you observe the data happening and take measurements.
- **Expected frequencies:** Are counts calculated using probability theory. Expected frequencies are calculated for each cell in the contingency table.

$$E_{ij} = \frac{T_i * T_j}{N}$$

Where,
•Eij: Expected frequency for ith row and jth column
•Ti: Total in the ith row
•Tj: Total in the jth row
•N: Grand Total

(row total * column total) / grand total

Terminology

Null Hypothesis (H0):

- It states that no association exists between the two cross-tabulated variables in the population.
- Hence, the variables are statistically independent.
- For example, if you compare two methods A and B for its goodness or which one works better, and if the assumption is that both methods are equally good, then this assumption is known as the **Null Hypothesis.**

Terminology

Alternate Hypothesis (HA):

- It proposes that the two variables are related to the population.
- If you assume that from two methods, method A is superior to method B or method B is superior to method A, then this assumption is known as **Alternative Hypothesis.**

Terminology

Degree of Freedom: The number of independent variates that make up the statistic is known as the degree of freedom of that statistic.

$$DOF = (r - 1) * (c - 1)$$

Where,

- r=numbers of rows
- c=number of columns

This will be used in the test of independence and test of homogeneity, not in the goodness of fit.

Terminology

- **Chi-square test Statistics:** A chi-squared statistic is a single number that tells you how much difference exists on your observed counts and the counts you would expect if there were no relationship at all in the population.
- **Chi-Square p-value:** Chi-square P-value will tell you if your test results are significant or not.

Types of Chi-square test

Goodness of fit:

➢Chi-Square goodness of fit test is a non-parametric test that is used to find out how the observed value of a given phenomenon is significantly different from the expected value.

➢In this test, you only have one variable from a single population

- **Null hypothesis (H0):** In the Chi-Square goodness of fit test, the null hypothesis assumes that there is no significant difference between the observed and the expected value (Source).
- **Alternative hypothesis (Ha):** In the Chi-Square goodness of fit test, the alternative hypothesis assumes that there is a significant difference between the observed and the expected value (Source).

Types of Chi-square test

Goodness of fit:

- **Null hypothesis (H0):** In the Chi-Square goodness of fit test, the null hypothesis assumes that there is no significant difference between the observed and the expected value (Source).
- **Alternative hypothesis (Ha):** In the Chi-Square goodness of fit test, the alternative hypothesis assumes that there is a significant difference between the observed and the expected value (Source).

Types of Chi-square test

Goodness of fit:

- A student rolled a fair 6-sided die 60 times and got the observed frequencies.

Die Value	Assumed Distribution	Observed Frequency
1	1/6	9
2	1/6	15
3	1/6	9
4	1/6	8
5	1/6	6
6	1/6	13

Types of Chi-square test

Goodness of fit:

- A student rolled a fair 6-sided die 60 times and got the observed frequencies.

  - H0 The die is fair
  - Ha The die is not fair

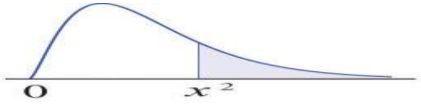
Types of Chi-square test

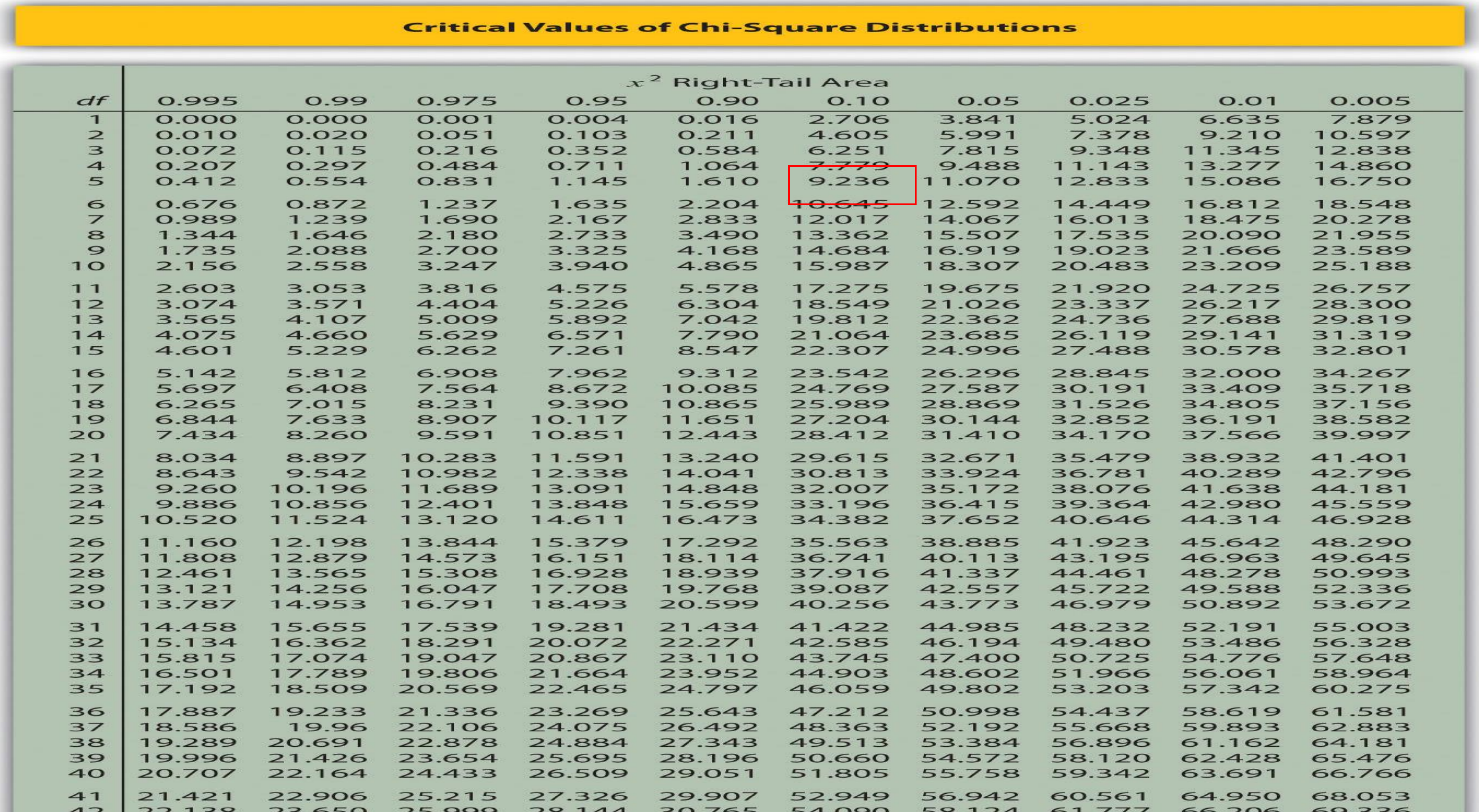
Goodness of fit:

- A student rolled a fair 6-sided die 60 times and got the observed frequencies.

Die Value	Assumed Distribution	Observed Frequency	expected	o-e	(o-e)^2/e
1	1/6	9	10	-1	0.1
2	1/6	15	10	5	2.5
3	1/6	9	10	-1	0.1
4	1/6	8	10	-2	0.4
5	1/6	6	10	-4	1.6
6	1/6	13	10	3	0.9
		60			5.6

Chi-Square Distribution sheet





Critical Values of Chi-Square Distributions

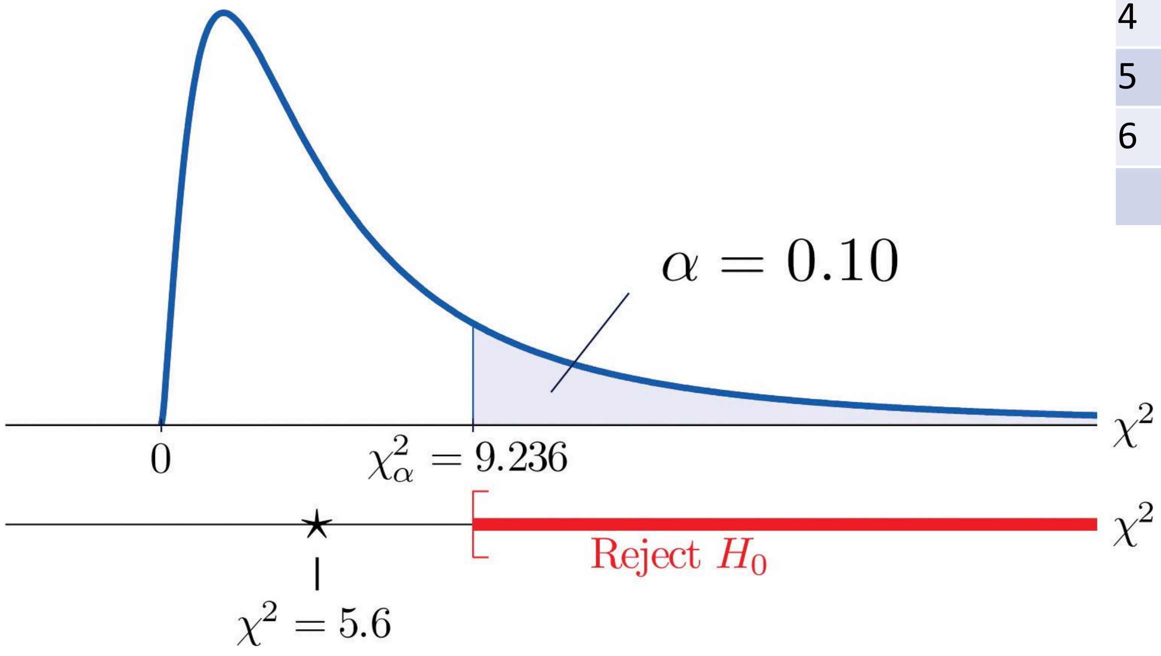
	x^2 Right-Tail Area									
df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
31	14.458	15.655	17.539	19.281	21.434	41.422	44.985	48.232	52.191	55.003
32	15.134	16.362	18.291	20.072	22.271	42.585	46.194	49.480	53.486	56.328
33	15.815	17.074	19.047	20.867	23.110	43.745	47.400	50.725	54.776	57.648
34	16.501	17.789	19.806	21.664	23.952	44.903	48.602	51.966	56.061	58.964
35	17.192	18.509	20.569	22.465	24.797	46.059	49.802	53.203	57.342	60.275
36	17.887	19.233	21.336	23.269	25.643	47.212	50.998	54.437	58.619	61.581
37	18.586	19.96	22.106	24.075	26.492	48.363	52.192	55.668	59.893	62.883
38	19.289	20.691	22.878	24.884	27.343	49.513	53.384	56.896	61.162	64.181
39	19.996	21.426	23.654	25.695	28.196	50.660	54.572	58.120	62.428	65.476
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
41	21.421	22.906	25.215	27.326	29.907	52.949	56.942	60.561	64.950	68.053
42	22.138	23.650	25.999	28.144	30.765	54.090	58.124	61.777	66.206	69.336

Types of Chi-square test

Goodness of fit:

- A student rolled a fair 6-sided die 60 times and got the observed frequencies.

Die Value	Assumed Distribution	Observed Frequency	expected	o-e	(o-e)^2/e
1	1/6	9	10	-1	0.1
2	1/6	15	10	5	2.5
3	1/6	9	10	-1	0.1
4	1/6	8	10	-2	0.4
5	1/6	6	10	-4	1.6
6	1/6	13	10	3	0.9
		60			5.6



Since 5.6 < 9.236 the decision is not to reject H_0.
The data do not provide sufficient evidence, at the 10% level of significance, to conclude that the die is loaded.

Example 1

Distribution of various ethnic groups in the population of a particular state based on a decennial U.S. census. Five years later a random sample of 2,500 residents of the state was taken, and census was taken.

Test, at the 1% level of significance, whether there is sufficient evidence in the sample to conclude that the distribution of ethnic groups in this state five years after the census had changed from that in the census year.

Example 1

Ethnicity	Assumed Distribution	Observed Frequency
White	0.743	1732
Black	0.216	538
American-Indian	0.012	32
Hispanic	0.012	42
Asian	0.008	133
Others	0.009	23

Thank you