STATISTICS FOR DATA SCIENCE

HYPOTHESIS and INFERENCE

Dr. Deepa Nair
Department of Science and Humanities

STATISTICS FOR DATA SCIENCE

UNIT-4 HYPOTHESIS and INFERENCE

Session-10

Chi-squared Test

Dr. Deepa Nair
Department of Science and Humanities

The Chi-Square Test for Homogeneity:

- Sometimes several multinomial trials are conducted, each with the same set of possible outcomes.
- The null hypothesis is that the probabilities of the outcomes are the same for each experiment.

The Chi-Square Test for Homogeneity:

Example:

- Use the following data to test the null hypothesis that the proportions of pins that are too thin, OK, or too thick are the same for all the machines.

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Homogeneity:

Example:

	T00 thin	OK	Too thick	Total
Machine 1	10	102	8	120
Machine 2	34	161	5	200
Machine 3	12	79	9	100
Machine 4	10	60	10	80
Total	66	402	32	500

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Homogeneity:

Example:

Expected Values

	T00 thin	OK	Too thick	Total
Machine 1				120
Machine 2				200
Machine 3				100
Machine 4				80
Machine 5	66	402	32	500

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Homogeneity:

Example:

Expected Values

	T00 thin	OK	Too thick	Total
Machine 1	15.84	96.48	7.68	120
Machine 2	26.40	160.8	12.8	200
Machine 3	13.2	80.4	6.40	100
Machine 4	10.58	64.32	5.12	80
Machine 5	66	402	32	500

Solution:

$$\chi 2 = (10 - 15.84)2/15.84 + \cdots + (10 - 5.12)2/5.12$$
$$= 34.1056/15.84 + \cdots + 23.8144/5.12$$
$$= 15.5844$$

Solution:

- Since there are four rows and three columns, the number of degrees of freedom is $(4-1)(3-1) = 6$.
- To obtain the *P*-value, we consult the chi-square table Looking under six degrees of freedom, we find that the upper 2.5% point is 14.449, and the upper 1% point is 16.812.
- Therefore $0.01 < P < 0.025$. It is reasonable to conclude that the machines differ in the proportions of pins that are too thin, OK, or too thick.

The Chi-Square Test for Independence:

- In some cases, both row and column totals are random. In either case, we can test the null hypothesis that the probabilities of the column outcomes are the same for each row outcome, and the test is exactly the same in both cases.

STATISTICS FOR DATA SCIENCE

Chi-squared Test

The Chi-Square Test for Independence:

Example:

- The cylindrical steel pins in previous example are subject to a length specification as well as a diameter specification.

- With respect to the length, a pin may meet the specification, or it may be too short or too long. A total of 1021 pins are sampled and categorized

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Example:

Observed Values

Length	T00 thin	OK	Too thick	Total
Too Short	13	117	4	134
OK	62	664	80	806
Too Long	5	68	8	81
Total	80	849	92	1021

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Example:

Expected Values

Length	T00 thin	OK	Too thick	Total
Too Short				134
OK				806
Too Long				81
Total	80	849	92	1021

STATISTICS FOR DATA SCIENCE

Chi-squared Test

Example:

Expected Values

Length	T00 thin	OK	Too thick	Total
Too Short	10.50	111.43	12.07	134
OK	63.15	670.22	72.63	806
Too Long	6.35	67.36	7.30	81
Total	80	849	92	1021

STATISTICS FOR DATA SCIENCE

Chi-squared Test

$$\chi^2 = \frac{(13 - 10.50)^2}{10.50} + \cdots + \frac{(8 - 7.30)^2}{7.30}$$

$$= \frac{6.25}{10.50} + \cdots + \frac{0.49}{7.30} = 7.46$$

- Since there are three rows and three columns, the number of degrees of freedom is $(3 - 1)(3 - 1) = 4$.

- To obtain the *P*-value, we consult the chi-square table Looking under four degrees of freedom, we find that the upper 10% point is 7.779. We conclude that $P > 0.10$.

- There is no evidence that the length and thickness are related.

Dr. Deepa Nair

Department of Science and Humanities

deepanair@pes.edu