ISyE 6739 –Goodness of Fit and Contingency Table Tests (Chapter 9)

Instructor: Kamran Paynabar
H. Milton Stewart School of
Industrial and Systems Engineering
Georgia Tech

Kamran.paynabar@isye.gatech.edu
Office: Groseclose 436

Goodness of Fit

- The objective of Goodness of fit tests is to test hypotheses on the probability distribution of a population.
- The test is based on the difference of observed frequency and expected frequency.
- The test statistic follows a chi-square distribution when H0 is true

H0: The probability distribution of defects is Poisson

H1: The probability distribution of defects is NOT Poisson

EXAMPLE 9-12 Printed Circuit Board Defects
Poisson Distribution
The number of defects in printed circuit boards is hypothesized to follow a Poisson distribution. A random sample of $n = 60$ printed boards has been collected, and the following number of defects observed.

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

Goodness of Fit

H0: The probability distribution of defects is Poisson

H1: The probability distribution of defects is NOT Poisson

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

Interval 1 →

• Let O_i be the observed frequency in the ith class interval.

• Let E_i be the expected frequency in the ith class interval.

The test statistic is

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

Pearson, Karl (1900). "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling". Philosophical Magazine Series 5 50 (302): 157–175.

Expected Frequency

• Let O_i be the observed frequency in the ith class interval.

• Let E_i be the expected frequency in the ith class interval.

Interval 1 →

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

In order to calculate expected frequency

1. Use the sample data and estimate the parameter of the distribution (λ)
2. Calculate the probability of each interval (p_i)
3. Find the expected frequency $E_i = np_i$

Note: if the expected frequency for an interval is less than 3, the interval should be combined with the previous interval

Goodness of Fit

H0: The probability distribution of defects is Poisson

H1: The probability distribution of defects is NOT Poisson

Number of Defects	Observed Frequency
0	32
1	15
2	9
3	4

The test statistic is

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

If $X_0^2 > \chi^2_{\alpha,k-p-1}$ H0 is rejected

# of intervals

# of estimated parameters

Example

A sample of 60 observations was collected from a production line. Each observation is the number of produced parts until one defective is found. Use goodness of fit to investigate if the distribution of X is Geometric.

$$P(X = x) = (1-p)^{x-1}p$$

$$E_i = np_i$$

x	Obsreved (Oi)	pi	Ei
1	30		
2	15		
3	10		
4	5		

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Normality Test

Goodness of Fit – Normal Distribution

H0: The probability distribution of defects is Normal

H1: The probability distribution of defects is NOT Normal

1. Define interval boundaries so that the expected frequencies E_i are equal for all intervals
2. Calculate the observed frequency for each interval
3. Calculate the test statistic

Class Interval	Observed Frequency o_i	Expected Frequency E_i
$x < 4.948$	12	12.5
$4.948 \le x < 4.986$	14	12.5
$4.986 \le x < 5.014$	12	12.5
$5.014 \le x < 5.040$	13	12.5
$5.040 \le x < 5.066$	12	12.5
$5.066 \le x < 5.094$	11	12.5
$5.094 \le x < 5.132$	12	12.5
$5.132 \le x$	14	12.5
Totals	100	100

$n = 100$

$\overline{X} = 5.04$

$S = 0.08$

$$\chi_0^2 = \sum_{i=1}^{8} \frac{(o_i - E_i)^2}{E_i} = \frac{(12-12.5)^2}{12.5} + \frac{(14-12.5)^2}{12.5} + \cdots + \frac{(14-12.5)^2}{12.5} = 0.64$$

Normality Check

Histogram and Normal plot for residuals:

- Normality assumption

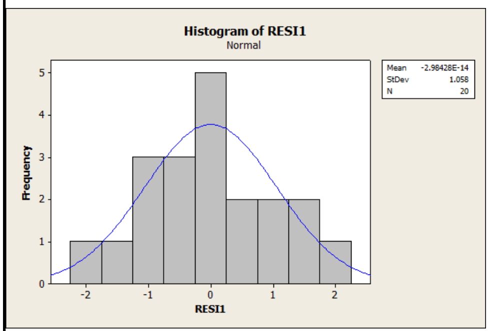
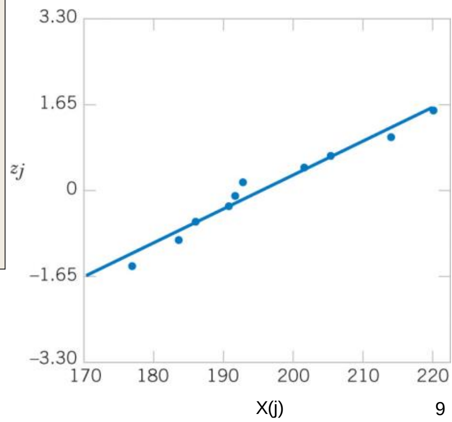


Figure 11-10 **Normal probability plot of residuals, Example 11-7.**



Normal Probability Plot

- To construct a probability plot:
  - Sort the data observations in ascending order: $x_{(1)}, x_{(2)},\ldots, x_{(n)}$.
  - Calculate empirical cumulative distribution $p_{(j)} = (j - 0.5)/n$.
  - Find $z_j = z_{(1-p(j))}$
  - The paired numbers are plotted (z_j vs. $x_{(j)}$)
  - If the paired numbers form a straight line, it is reasonable to assume that the data follows the proposed distribution.

Example

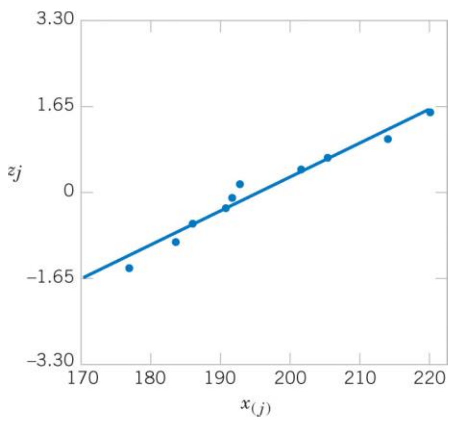


Table 6-6 Calculations for Constructing a Normal Probability Plot

j	$x_{(j)}$	$(j-0.5)/10$	z_j
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205		
9	214		
10	220		

Contingency Table Tests

- The objective of Contingency Table tests is to test if two categorical variable are independent.

H0: Two categorical variables are independent

H1: Two categorical variables are NOT independent

Table 9-2 An $r \times c$ Contingency Table

		Columns			
		1	2	…	c
Rows	1	O_{11}	O_{12}	…	O_{1c}
	2	O_{21}	O_{22}	…	O_{2c}
	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	…	O_{rc}

Table 9-3 Observed Data for Example 9-14

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

Contingency Table Tests

p_{ij} Probability that a randomly selected element falls in the ijth cell

u_i Probability that a randomly selected element falls in row i

v_j Probability that a randomly selected element falls in column j

$$\hat{u}_i = \frac{1}{n} \sum_{j=1}^{c} O_{ij}$$

$$\hat{v}_j = \frac{1}{n} \sum_{i=1}^{r} O_{ij}$$

$$\hat{p}_{ij} = \hat{u}_i \hat{v}_j$$

$$E_{ij} = n\hat{p}_{ij} = n\hat{u}_i \hat{v}_j$$

		Columns			
		1	2	...	c
Rows	1	O_{11}	O_{12}	...	O_{1c}
	2	O_{21}	O_{22}	...	O_{2c}
	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	...	O_{rc}

Table 9-3 Observed Data for Example 9-14

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

Table 9-4 Expected Frequencies for Example 9-14

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers				340
Hourly workers				160
Totals	200	200	100	500

$$\hat{u}_i = \frac{1}{n} \sum_{j=1}^{c} O_{ij}$$

$$\hat{v}_j = \frac{1}{n} \sum_{i=1}^{r} O_{ij}$$

Contingency Table Tests

$$\chi_0^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

		Columns			
		1	2	...	c
Rows	1	O_{11}	O_{12}	...	O_{1c}
	2	O_{21}	O_{22}	...	O_{2c}
	⋮	⋮	⋮	⋮	⋮
	r	O_{r1}	O_{r2}	...	O_{rc}

Table 9-3 Observed Data for Example 9-14

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers	160	140	40	340
Hourly workers	40	60	60	160
Totals	200	200	100	500

If $X_0^2 > \chi^2_{\alpha,(r-1)(c-1)}$ H0 is rejected

# of rows # of columns

Table 9-4 Expected Frequencies for Example 9-14

	Health Insurance Plan			
Job Classification	1	2	3	Totals
Salaried workers				340
Hourly workers				160
Totals	200	200	100	500