UCLA STAT 110B
Applied Statistics for Engineering and the Sciences

•**Instructor**: Ivo Dinov,
Asst. Prof. In Statistics and Neurology

•**Teaching Assistants:** Brian Ng, UCLA Statistics

University of California, Los Angeles, Spring 2003

http://www.stat.ucla.edu/~dinov/courses_students.html

Categorical Data

Categorical Data is that which counts the number of outcomes falling into various categories.

•Binomial Experiment – consists of two categories

•Multinomial Experiment – consist of more than two categories

Binomial Experiment

- n independent trials
- Two possible outcomes (S) success and (F) failure
- p = Probability of success on each trial
- X = Number of successes in n trials

Binomial Distribution
Pdf, E[X], Var[X]

Multinomial Experiment

- n independent trials results in one of k possible categories labeled 1, …, k
- p_i = the probability of a trial resulting in the ith category, where $p_1+\ldots+p_k = 1$
- N_i = number of trials resulting in the ith category, where $N_1+\ldots+N_k = n$

Multinomial Cont'd

- The random variables $N_1,\ldots,N_k$ have a multinomial distribution

$$p(n_1,\ldots,n_k) = \frac{n!}{n_1!\cdots n_k!}\, p_1^{n_1} \cdots p_k^{n_k}$$

Multinomial Cont'd

- Expected Value: $E[N_i] = np_i = E_i$
- Variance: $Var [N_i] = np_i q_i$
- Covariance: $Cov [N_i, N_j] = -np_i p_j$

Testing Goodness of Fit with Specified Cell Probabilities

We wish to test whether the cell probabilities are specified by $p_1^o, \ldots, p_k^o$ where $p_1^o + \ldots + p_k^o = 1$.

We will use a test statistic to compare the observed cell count N_i to the expected cell count under H_o,

$E_i = np_i^o$

$H_o: p_1 = p_1^o$, (and) …., (and) $p_k = p_k^o$

H_a: Some $p_i \neq p_i^o$

Test Statistic

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

This is a Pearson's goodness-of-fit statistic

Rejection Region: $X^2 > \chi_\alpha^2$ where χ^2 is the chi-squared distribution with k-1 degrees of freedom.

General Rule: We want $np_i^o \geq 5$ for all cells

Example

A study is run to see whether the public favors the construction of a new dam. It is thought that 40% favor dam construction, 30% are neutral, 20% oppose the dam, and the rest have not thought about it. A random sample of 150 individuals are interviewed resulting in 42 in favor, 61 neutral, 33 opposed, and the rest have not though about it. Does the data indicate that the stated proportions are incorrect? Use α=0.01.

Example Cont'd

H_o: p_1=0.4, p_2=0.3, p_3=0.2, p_4=0.1

H_a: At least one probability is not as specified

Test Statistic: X^2

Rejection Region: $X^2 > \chi^2_{0.01,\,3} = 11.34$

	Favor	Neutral	Oppose	Unaware	Total
ni	42	61	33	14	150
pio	0.4	0.3	0.2	0.1	1
Ei	60	45	30	15	150

$$X^2 = \frac{(42-60)^2}{60} + \frac{(61-45)^2}{45} + \frac{(33-30)^2}{30} + \frac{(14-15)^2}{15} = 11.46$$

Since $X^2 = 11.46 > \chi_{0.01,3,}^2 = 11.34$, we reject H_o. Conclude that at least one of the true proportions differs from that hypothesized

Goodness of Fit for Distributions (Continuous and Discrete)

• Uses the concept of Maximum Likelihood Estimations (MLE)

• The range of a hypothesized distribution is divided into a set of k intervals (cells). After finding the MLE of unknown parameters, the cell probabilities are calculated and the χ^2 test performed

•Found in many computer packages - SOCR

Testing Normality

Many test procedures that we have developed rely on the assumption of Normality. There are many test for Normality of data. One uses the normal to provide cell probabilities for the chi-square goodness-of-fit test. A "better" test is based on the Normal Probability Plot

Testing Normality Cont'd

Recall: The NPP should be approx linear for normal data, and the correlation coefficient is a measure of linearity.

If r is much less than one, we would conclude that the data doesn't come from a Normal distribution.

Ryan-Joiner Test

1. Order the data $x_{(1)},\ldots,x_{(n)}$
2. Compute the normal percentiles

$$y_i = \Phi^{-1}\left(\frac{i-.375}{n+.25}\right)$$

3. Compute the correlation coefficient, ***R***, for the $(y_i, x_{(i)})$ pairs and look up the distribution table for the Ryan-Joiner Statistics, A.12.

Ryan-Joiner Test

4. State the Null and Alternative Hypotheses

   H_o: The population is normal

   H_a: The population is not normal

5. Specify alpha and obtain critical values from Table A.12. Compare ***R*** to this value

Example

Consider the following data. Use the Ryan-Joiner test to test the assumption of normality at $\alpha = 0.10$

1.15	1.4	1.34	1.29	1.36	1.26	1.22	1.4
1.29	1.14	1.32	1.34	1.26	1.36	1.36	1.3
1.28	1.45	1.29	1.28	1.38	1.55	1.46	1.32

Raw Data

1.15; 1.4 1.34 1.29 1.36 1.26 1.22 1.4
1.29 1.14 1.32 1.34 1.26 1.36 1.36 1.3
1.28 1.45 1.29 1.28 1.38 1.55 1.46 1.32

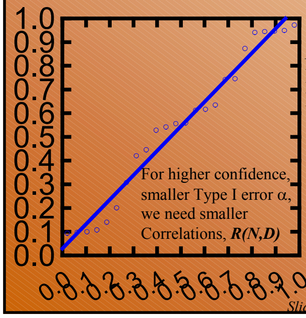
Normal(0,1) random sample:

0.036083496	1.15
-0.49035215	1.4
0.141456326	1.34
-1.18561	1.29
0.930130358	1.36
-1.01800921	1.26
-0.37025793	1.22
0.579551809	1.4
1.468404908	1.29
-1.13996599	1.14
1.078497074	1.32
0.993224744	1.34
1.026819753	1.26
-0.09336254	1.36
0.127025574	1.36
-1.10792949	1.3
-0.05052723	1.28
-0.16153136	1.45
-0.21241079	1.29
-0.10578143	1.28
-0.43351619	1.38
-0.67097827	1.55
-0.76809374	1.46
-0.03419266	1.32

Example

1.15	1.4	1.34	1.29	1.36	1.26	1.22	1.4
1.29	1.14	1.32	1.34	1.26	1.36	1.36	1.3
1.28	1.45	1.29	1.28	1.38	1.55	1.46	1.32

Corr(N(0,1), Data)=0.95541

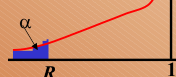


H_o: Data is Normal

R ~ Ryan-Joiner (α,n)

RJ(0.01,24)= 0.9408

RJ(0.10,24)= 0.9662



Since R_o=0.95541

→ R_o > Critical Value

→ Strong Correlation

→ Can't Reject H_o

Ascending Order Stats:

N(0,1)	Data
-1.18561	1.14
-1.13997	1.15
-1.10793	1.22
-1.01801	1.26
-0.76809	1.26
-0.67098	1.28
-0.49035	1.28
-0.43352	1.29
-0.37026	1.29
-0.21241	1.29
-0.16153	1.3
-0.10578	1.32
-0.09336	1.32
-0.05053	1.34
-0.03419	1.34
0.036083	1.36
0.127026	1.36
0.141456	1.36
0.579552	1.38
0.93013	1.4
0.993225	1.4
1.02682	1.45
1.078497	1.46
1.468405	1.55

Testing Homogeneity of Populations

*We wish to compare I multinomial populations, each with J categories. *

Take n_i samples from the ith population

Let N_{ij} be the number of observations from the ith population in the jth category. Hence, $\Sigma_j N_{ij} = n_i$

Place the data in a I x J table

Table

		Category				
		1	2	….	J	Total
Pop.	1	n11	n12	….	n1J	n1.
	2	n21	n22		n2J	n2.
	.	.	.	….	.	.
	.	.	.	….	.	.
	.	.	.	….	.	.
	I	nI1	nI2	….	nIJ	nI.
	Total	n.1	n.2	….	n.J	n

Corresponding to each cell, there is a cell probability p_{ij}=probability and outcome for the ith population falls into the jth category, where $\Sigma_j p_{ij} = 1$

		Category			
		1	2	….	J
Pop.	1	p11	p12	….	p1J
	2	p21	p22		p2J
	.	.	.	….	.
	.	.	.	….	.
	.	.	.	….	.
	I	pI1	pI2	….	pIJ

Test

H_o: $p_{1j} = p_{2j} = \ldots = p_{Ij}$, j = 1,…,J

H_a: Some $p_{ij} \neq p_{i'j}$

Under H_o, the common cell probability p_j is estimated by

$$\hat{p}_j = \frac{n_{.j}}{n}$$

Test Cont'd

The estimated expected cell frequency is

$$\hat{E}_{ij} = n_i \hat{p}_j = \frac{n_i n_{.j}}{n}$$

The test statistic is

$$X^2 = \sum_{rows} \sum_{columns} \frac{\left(n_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$$

Rejection Region: $X^2 > \chi^2_\alpha$ with d.f. = (I-1)(J-1)

Testing for Association

* Individuals are categorized by two categorical variables. We wish to determine whether these variables are associated. *

Row Categories – $A_1,\ldots,A_I$

Column Categories – $B_1,\ldots,B_J$

n = Total number of observations

n_{ij} = the number of individuals classified as A_i and B_j

Hence, $\Sigma\Sigma\, n_{ij} = n$

H_o: $P(A_i \cap B_j) = P(A_i)P(B_j)$ for all i,j

H_a: Some $P(A_i \cap B_j) \neq P(A_i)P(B_j)$

Expected Frequency:

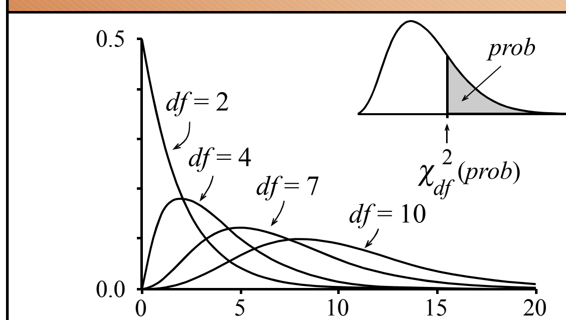
$$\hat{E}_{ij} = \frac{n_{i\cdot} \text{ x } n_{\cdot j}}{n}$$

Test Statistic:

$$X^2 = \sum_{rows} \sum_{columns} \frac{\left(n_{ij} - \hat{E}_{ij}\right)^2}{\hat{E}_{ij}}$$

Rejection Region: $X^2 > \chi^2_\alpha$ with d.f = (I-1)(J-1)

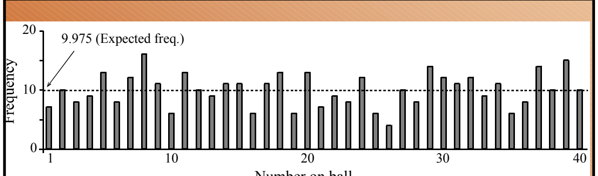
The Chi-square distribution



Lotto after 399 numbers have been drawn – Do some numbers appear more frequently in LOTTO?

Frequency of Winning Numbers in LOTTO

1.	(7)	**2.**	(10)	**3.**	(8)	**4.**	(9)	**5.**	(13)	**6.**	(8)	**7.**	(12)	**8.**	(16)	**9.**	(11)	**10.**	(6)
11.	(13)	**12.**	(10)	**13.**	(9)	**14.**	(11)	**15.**	(11)	**16.**	(6)	**17.**	(11)	**18.**	(13)	**19.**	(6)	**20.**	(13)
21.	(7)	**22.**	(9)	**23.**	(8)	**24.**	(12)	**25.**	(6)	**26.**	(4)	**27.**	(10)	**28.**	(8)	**29.**	(14)	**30.**	(12)
31.	(11)	**32.**	(12)	**33.**	(9)	**34.**	(11)	**35.**	(6)	**36.**	(8)	**37.**	(14)	**38.**	(10)	**39.**	(15)	**40.**	(10)



Frequency of LOTTO winning numbers

Lotto after 399 numbers have been drawn – Do some numbers appear more frequently in LOTTO?

Number-range: [1:40]

Number of balls selected at each draw: 7

Number of samples: 57

Total number of balls selected: 57*7=399,

Expected value of each number: 399/40 = 9.975

Observed χ^2 statistics is x_0=30.97

df=40-1=39

P-value = 0.817

Conclusion: No evidence for departure from the null hypothesis.

Chi-Square Tests of Independence

An Example, Researchers in a California community have asked a sample of 175 automobile owners to select their favorite from three popular automotive magazines. Of the 111 import owners in the sample, 54 selected *Car and Driver*, 25 selected *Motor Trend*, and 32 selected *Road & Track*.

Of the 64 domestic-make owners in the sample, 19 selected *Car and Driver*, 22 selected *Motor Trend*, and 23 selected *Road & Track*. At the 0.05 level, is import/domestic ownership independent of magazine preference? What is the most accurate statement that can be made about the *p*-value for the test?

Chi-Square Tests of Independence

- First, arrange the data in a table.

	Car and Driver (1)	*Motor Trend (2)*	*Road & Track (3)*	**Totals**
Import (Imp)	54	25	32	**111**
Domestic (Dom)	19	22	23	**64**
Totals	**73**	**47**	**55**	**175**

- Second, compute the expected values and contributions to χ^2 for each of the six cells.
- Then to the hypothesis test ....

Chi-Square Tests of Independence

		Car and Driver (1)	*Motor Trend (2)*	*Road & Track (3)*
Import (Imp):	O -	54	25	32
	E -	46.3029	29.8114	34.8857
χ^2 contribution -		1.2795	0.7765	0.2387
Domestic (Dom) :	O -	19	22	23
	E -	26.6971	17.1886	20.1143
χ^2 contribution -		2.2192	1.3468	0.4140

$\Sigma\, \chi^2$ contributions = 6.2747

Chi-Square Tests of Independence

- I. Hypotheses:

H_0: Type of magazine and auto ownership are independent.

H_1: Type of magazine and auto ownership are not independent.

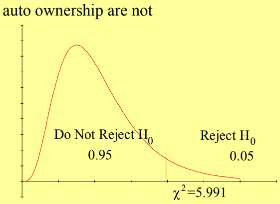
- II. Rejection Region:

$\alpha = 0.05$

df = $(r-1)\,(k-1)$

$= (2-1)\bullet(3-1)$

$= 1 \bullet 2 = 2$



If $\chi^2 > 5.991$, reject H_0.

Chi-Square Tests of Independence

- III. Test Statistic:

$\chi^2 = 6.2747$

- IV. Conclusion:

Since the test statistic of 6.2747 falls beyond the critical value of 5.991, we reject the null hypothesis with at least 95% confidence.

- V. Implications:

There is enough evidence to show that magazine preference is not independent from import/domestic auto ownership.

- *p*-value: In a cell on a Microsoft Excel spreadsheet, type:

=CHIDIST(6.2747,2). The answer is: ***p*-value = 0.043398**