Non-Parametric Inference

Non – Parametric inference involves estimation and testing procedures when shape of the population distribution is unknown. The theory of Non – Parametric is mainly based on Order Statistics and the Probability Integral Transform (PIT). Commonly the tests are based on counts, ranks and runs. Most of the time nonparametric testing procedures are developed for following two purposes:

1. To test a hypothesis related to some location parameter.
2. To test a hypothesis related to equality of two or more populations.

Non-parametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of non-parametric methods may be necessary when data have a ranking but no clear numerical interpretation, such as when assessing preferences. In terms of levels of measurement, non-parametric methods result in "ordinal" data.

As non-parametric methods make fewer assumptions, their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Also, due to the reliance on fewer assumptions, non-parametric methods are more robust.

Another justification for the use of non-parametric methods is simplicity. In certain cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due both to this simplicity and to their greater robustness, non-parametric methods are seen by some statisticians as leaving less room for improper use and misunderstanding.

The wider applicability and increased robustness of non-parametric tests comes at a cost: in cases where a parametric test would be appropriate, non-parametric tests have less power. In other words, a larger sample size can be required to draw conclusions with the same degree of confidence.

Frank Wilcoxon Sign Test

One Sample Sign Test (Binomial Test)

Sign test is used to test for some hypothetical value of a population quantile, i.e. to test if a particular population quantile is equal to some hypothetical value. Population Median is also a quantile ($Q_{0.5}$). The most application of sign test is to test "if the population median is equal to a hypothetical value".

The procedure involves calculating median and then calculating the difference between median and the sample observations. Thereafter we count either positive or negative signs in the difference and then make use of Binomial Distribution to carry out the test.

Case 1: The following table represents observations on heights and weights of 15 females:

Observation	Height(inch's)	Weight(lbs)	Observation	Height(inch's)	Weight(lbs)
1	58	115	9	66	139
2	59	117	10	67	142
3	60	120	11	68	146
4	61	123	12	69	150
5	62	126	13	70	154
6	63	129	14	71	159
7	64	132	15	72	164
8	65	135			

We have data on heights and weights of 15 females. We will apply Binomial Test to test the following hypothesis:

H_0: The Height of the females can be taken to be equal to 64 inches. Vs

H_1: The Height of the females cannot be taken to be equal to 64 inches.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Height	Group 1	<= 64	7	.47	.50	1.000
	Group 2	> 64	8	.53		
	Total		15	1.00		

Conclusion: We see p-value is 1.00 > 0.05, hence we fail to reject H_0 and conclude that the heights of females can be taken to be 64 inches.

We will apply Binomial Test to test the following hypothesis:

H_0: The weight of the females can be taken to be equal to 135 lbs. Vs

H_1: The weight of the females cannot be taken to be equal to 135lbs.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Weight	Group 1	<= 135	8	.53	.50	1.000
	Group 2	> 135	7	.47		
	Total		15	1.00		

Conclusion: We see p-value is 1.00 > 0.05, hence we fail to reject H_0 and conclude that the weights of females can be taken to be 135 lbs.

Case 2: Win/Loss records of a certain basketball team during their 50 consecutive games are given in the following table:

Game	Outcome	Game	Outcome	Game	Outcome	Game	Outcome	Game	Outcome
1	1	11	1	21	0	31	0	41	1
2	1	12	1	22	1	32	1	42	0
3	1	13	1	23	1	33	1	43	0
4	1	14	0	24	1	34	1	44	0
5	1	15	1	25	1	35	1	45	1
6	1	16	0	26	0	36	1	46	1
7	0	17	1	27	1	37	1	47	0
8	1	18	1	28	1	38	0	48	1
9	1	19	1	29	1	39	0	49	1
10	1	20	0	30	0	40	1	50	1

We will use Sign Test to test the hypothesis that win and loss are equally likely:

H_0: Chance of winning is 50%.

H_1: Chance of winning is not 50%.

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (2-tailed)
Outcome	Group 1	Win	36	.72	.50	.003
	Group 2	Loss	14	.28		
	Total		50	1.00		

Conclusion: We see p-value is 0.003 < 0.05, hence we reject H_0 and conclude that win and loss are not equally likely.

Two Sample Sign Test (Sign Test)

The purpose of a two sample sign test which is generally referred as the sign test is to test whether two "related" samples are coming from the same population or not? In two sample sign test we need paired or related observations on two sample. This test is sometimes called as the non-parametric counterpart of the paired t – test. The intuition is as follows:

"If two paired samples are coming from the same population then the probability that sample observations of the first sample exceed or fall below the sample observations of the sample observations in the second sample. So if we calculate the pairwise difference between the sample observations from different samples and count the positive signs then we would expect approximately half of the signs would be positive if the samples come from the same population.

Case 3: Following data represents the marks given to the same set 22 students by two different professors in the same examination:

Student	**Professor A**	**Professor B**	**Student**	**Professor A**	**Professor B**
1	79	83	12	64	45
2	87	91	13	50	56
3	24	18	14	92	89
4	41	39	15	64	67
5	59	67	16	39	35
6	12	34	17	49	40
7	91	78	18	86	82
8	78	89	19	23	32
9	63	38	20	45	38
10	30	45	21	12	23
11	9	10	22	88	92

We will Use Sign Test to test if the grading of both the professors can be taken to be same.

H_0: Grading done by both professors are same.

H_1: Grading done by both professors are not same.

Frequencies

		N
Professor B - Professor A	Negative Differences[a]	10
	Positive Differences[b]	12
	Ties[c]	0
	Total	22

a. Professor B < Professor A

b. Professor B > Professor A,

c. Professor B = Professor A

Test Statistics[a]

	Professor B - Professor A
Exact Sig. (2-tailed)	.832[b]

a. Sign Test

b. Binomial distribution used.

Conclusion: We see p-value 0.832 > 0.05, hence we fail to reject H_0 and conclude that grading done by both professors are same.

Wald – Wolfowitz Run Test

A **run** is defined as a sequence of letters of one kind surrounded by a sequence of letters of other kind, and the number of element in a run is known as the length of the run.

Test Statistic: Let $Z_1,\ldots, Z_{n1+n2}$ be the combined ordered sample then define

U = # of runs in the combined ordered sample

Intuition: If two samples are coming from the same population then there would be a thorough mingling of X's and Y's and consequently # of runs in the combined ordered sample would be large. On the other hand if the samples are coming from two different populations then # of runs in the combined ordered samples would expected to be small. But "HOW SMALL??" so that we can accept the alternative. Under H_0:

$$P(U = u) = \frac{2C_{k-1}^{n_1-1}C_{k-1}^{n_1-1}}{C_{n_1}^{n_1+n_2}} if\ U = 2k$$

$$P(U = u) = \frac{C_{k-1}^{n_1-1}C_{k}^{n_2-1} + C_{k-1}^{n_1-1}C_{k}^{n_2-1}}{C_{n_1}^{n_1+n_2}} if\ U = 2k + 1$$

Decision Rule: Reject H_0 if $U \leq U_\propto$ where $U_\propto$ is obtained by solving the following.

$$P(U \leq U_\propto | H_0) \leq \propto$$

Case 4: Following is the data for prices in rupees of a certain commodity in a sample of 15 randomly selected shops from City A and those of 13 randomly selected shops from City B.

City A (prices in rupees)	City B (prices in rupees)
7.41	7.08
7.77	7.49
7.44	7.42
7.4	7.04
7.38	6.92
7.93	7.22
7.58	7.68
8.28	7.24
7.23	7.74
7.52	7.81
7.82	7.28
7.71	7.43
7.84	7.47
7.63	
7.68	

We will use Run Test to determine if the prices in City A and City B can be taken to be following same probability distribution.

H_0: Prices in City A and City B follow same probability distribution. Vs

H_1: Prices in City A and City B do not follow same probability distribution.

Frequencies

City		N
Price	City A	15
	City B	13
	Total	28

Test Statistics[a,b]

		Number of Runs	Z	Exact Sig. (1-tailed)
Price	Minimum Possible	14[c]	-.166	.436
	Maximum Possible	14[c]	-.166	.436

a. Wald-Wolfowitz Test

b. Grouping Variable: City

c. There are 1 inter-group ties involving 2 cases.

Conclusion: We see, p-value corresponding to both minimum and maximum possible runs are $0.436 > 0.05$, hence we fail to reject H_0 and conclude that, prices in both cities follow same probability distribution.

Mann – Whitney – Wilcoxon U – Test

U-test is a non-parametric test to test a hypothesis regarding equality of two population distributions.

Let $X_1 \dots X_{n1}$ be a random sample of size n from $X \sim F_X$ and $Y_1 \dots Y_{n2}$ be a random sample of size n from $Y \sim F_Y$. We are interested in testing a hypothesis regarding equality of two population distributions on the basis of the available samples. Set up the hypotheses as follows:

$$H_0: F_X = F_Y$$

$$H_1: F_X \neq F_Y \text{ (Two Tailed)}$$

Test Statistic: Let $Z_1 \dots Z_{n1+n2}$ be the combined ordered sample. Let T = sum of the ranks of Y's in the combined ordered sample then define:

$$U = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - 1$$

Intuition: If two samples are coming from different populations then one of the samples is expected to fall below or above the other sample, i.e. the sum of the ranks of sample units in one sample will be either very small or very large.

Decision Rule: For a pre-specified level of significance we reject H_0 if $U \geq U_{\propto/2}$ or $U \leq U'_{\propto/2}$ where$U_{\propto/2}$ and $U'_{\propto/2}$ are obtained by solving the following.

$$(U \geq U_{\propto/2} | \ H_0) \leq \propto/2 \qquad \text{and} \qquad (U \leq U'_{\propto/2} \ | \ H_0) \leq \propto/2$$

Case 5: An experiment on reading ability of students was conducted, where at the beginning of the year a class was randomly divided into two groups. One group was taught to read using a uniform method, where all the students progressed from one stage to the next at the same time, following the instructor's direction. The second group was taught to read using an individual method, where each student progressed at his own rate according to a programmed work book under the supervision of the instructor. At the end of the year each student was given a reading ability test and following were their scores:

First Group	227	176	252	149	16	55	234	194	247	92	184	147	88	161	171
Second Group	202	14	165	171	292	271	151	235	147	99	63	284	53	228	271

We will use U-test to test if two different teaching methods for reading ability can be taken as equally effective.

H_0: Two different teaching methods for reading ability are equally effective.

H_1: Two different teaching methods for reading ability are not equally effective.

Ranks

Flag		N	Mean Rank	Sum of Ranks
Scores	Group 1	15	14.47	217.00
	Group 2	15	16.53	248.00
	Total	30		

Test Statistics[a]

	Scores
Mann-Whitney U	97.000
Wilcoxon W	217.000
Z	-.643
Asymp. Sig. (2-tailed)	.520
Exact Sig. [2*(1-tailed Sig.)]	.539[b]

a. Grouping Variable: Flag

b. Not corrected for ties.

Conclusion: We see p-value is $0.539 > 0.05$, hence we fail to reject H_0 and conclude that 2 different teaching methods for reading ability are equally effective.

Run Test for Randomness

Assumptions:

1. The sample data are arranged according to some scheme (such as time series).
2. The data falls into two separate categories (such as above and below a specific value).
3. The runs test is based on the order in which the data occur; not on the frequency of the data.

Let $X_1 \dots X_n$ be a Sample of size n. We are interested in testing a hypothesis whether the sample is random or not. Set up the hypotheses as follows:

H_0: Data is random.

H_1: Data is not random.

Test Statistic: Let M be the sample median then for each observation Define an indicator variable $\delta i = I\ (Xi > M)$, a realization of δi's can be 1001110010100011, define

U = # of run in the realization of δi's

Decision Rule: For a pre-specified level of significance we reject the hypothesis of randomness of data if $U \geq U_{\propto/2}$ or $U \leq U'_{\propto/2}$ where $U_{\propto/2}$ and $U'_{\propto/2}$ are obtained by solving the following.

$$P(U \geq U_{\propto/2} \mid H_0) \leq \propto/2 \qquad \text{and} \qquad P(U \leq U_{\propto/2}\ , \mid H_0) \leq \propto/2$$

If sample size is sufficiently large (>20) then we can use a normal approximation by using,

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0,1)$$

Under H_0, $E(U) = \frac{n+2}{2}$ and $Var(U) = \frac{n}{4}\left[\frac{n-2}{n-1}\right]$

Decision Rule: For a pre-specified level of significance we reject the hypothesis of randomness of data if $Z \geq Z_{\propto/2}$ or $Z \leq Z'_{\propto/2}$ where $Z_{\propto/2}$ and $Z'_{\propto/2}$ are obtained by solving the following:

$$1 - \Phi\ (Z_{\propto/2} \mid H_0) \leq \propto/2 \text{ and}$$

$$(Z'_{\propto/2} \mid H_0) \leq \propto/2$$

where Φ(.) is the CDF of a SNV.

Case 6: We will test the randomness of following sample of size 30 using Run Test:

15, 77, 01, 65, 69, 69, 58, 40, 81, 16, 16, 20, 00, 84, 22, 28, 26, 46, 66, 36, 86, 66, 17, 43, 49, 85, 40, 51, 40, 10

H_0: Sample is random.

H_1: Sample is not random.

Runs Test

	Sample
Test Value[a]	42
Cases < Test Value	15
Cases >= Test Value	15
Total Cases	30
Number of Runs	17
Z	.186
Asymp. Sig. (2-tailed)	.853

a. Median

Conclusion: We see p-value is 0.853 > 0.05, hence we fail to reject H_0 and conclude that sample is random.

One Sample Kolmogorov – Smirnov Test

One Sample KS Test is used to test if a sample taken from a specified population. The test statistic is calculated as a measure of Distance between the theoretical (to be tested) and empirical (observed) distribution functions.

Case 7: For the following four samples test if they are drawn from Normal, Exponential, Poisson and Uniform distributions respectively.

Observation	Sample 1	Sample 2	Sample 3	Sample 4
1	1.089781309	0.046136443	4	10.25068427
2	1.962787672	0.296905535	2	10.12379195
3	1.724451834	0.013852846	1	17.81733259
4	1.63955842	0.149763684	1	18.87337658
5	0.144050286	0.216846562	2	15.32347378
6	0.232942589	0.549152735	3	11.97729205
7	1.68271611	0.075868307	4	16.79090476
8	3.633887711	0.147932045	3	14.40535435
9	1.81341443	0.29035859	3	14.10547096
10	1.683039558	0.027180583	4	14.99055234
11	1.659612162	0.163903305	3	12.68408943
12	0.8396626	0.8104371	1	10.62609998
13	3.427254188	0.078686029	8	15.59840961
14	1.127955432	0.153359897	4	17.59452935
15	1.552543896	0.141724322	5	10.60249139
16	0.214796062	0.066255849	2	17.17324608
17	0.475882672	0.085298693	4	11.59441059
18	3.013061127	0.507875983	5	14.11860911
19	2.73502768	0.104899753	0	19.68738385
20	2.583921184	0.020127363	5	17.20303417

H_0: Sample is from normal distribution.

H_1: Sample is not from normal distribution.

One-Sample Kolmogorov-Smirnov Test

		Sample1
N		20
Normal Parameters[a,b]	Mean	1.6618
	Std. Deviation	1.02850
Most Extreme Differences.	Absolute	.141
	Positive	.141
	Negative	-.108
Test Statistic		.141
Asymp. Sig. (2-tailed)		.200[c,d]

a. Test distribution is Normal.
b. Calculated from data.
c. Lilliefors Significance Correction.
d. This is a lower bound of the true significance.

Conclusion: We see p-value is 0.200 > 0.05, hence we fail to reject H_0 and conclude that sample is from Normal distribution.

H_0: Sample is from Exponential distribution.

H_1: Sample is not from Exponential distribution.

One-Sample Kolmogorov-Smirnov Test

		Sample2
N		20
Exponential parameter.[a,b]	Mean	.1973
Most Extreme Differences	Absolute	.136
	Positive	.136
	Negative	-.085
Kolmogorov-Smirnov Z		.607
Asymp. Sig. (2-tailed)		.855

a. Test Distribution is Exponential.

b. Calculated from data.

Conclusion: We see p-value is 0.855 > 0.05, hence we fail to reject H_0 and conclude that sample is from Exponential distribution.

H_0: Sample is from Poisson distribution.

H_1: Sample is not from Poisson distribution.

One-Sample Kolmogorov-Smirnov Test

		Sample3
N		20
Poisson Parameter[a,b]	Mean	3.2000
Most Extreme Differences	Absolute	.055
	Positive	.055
	Negative	-.053
Kolmogorov-Smirnov Z		.248
Asymp. Sig. (2-tailed)		1.000

a. Test distribution is Poisson.

b. Calculated from data.

Conclusion: We see p-value is 1.00 > 0.05, hence we fail to reject H_0 and conclude that sample is from Poisson distribution.

H_0: Sample is from Uniform distribution.

H_1: Sample is not from Uniform distribution.

One-Sample Kolmogorov-Smirnov Test

		Sample4
N		20
Uniform Parameters[a,b]	Minimum	10.12
	Maximum	19.69
Most Extreme Differences	Absolute	.147
	Positive	.147
	Negative	-.066
Kolmogorov-Smirnov Z		.660
Asymp. Sig. (2-tailed)		.777

a. Test distribution is Uniform.

b. Calculated from data.

Conclusion: We see p-value is 0.77 > 0.05, hence we fail to reject H_0 and conclude that sample is from Uniform distribution.

Two Sample Kolmogorov – Smirnov Test

Two Sample KS Test is used to test if two samples are taken from same population. The test statistic is calculated as a measure of Distance between the empirical (observed) distribution functions of the samples.

Case 8: For the following two samples test if they can be taken to be coming from same population:

Observation	Sample 1	Sample 2
1	0.075204597	1.319177696
2	0.282203071	0.255423126
3	0.473605304	0.250284353
4	0.171775727	0.941835437
5	0.084642496	3.078396099
6	0.601160542	0.270368067
7	0.212552515	0.413272132
8	0.294969478	0.05425652
9	0.026919861	1.340734424
10	0.054462148	0.127618122
11	0.076084169	0.060699583
12	0.021943532	0.208278913
13	0.486042232	0.104869289
14	0.083376869	1.126610877
15	0.62800881	1.179774988
16	1.317637268	2.015836491
17	0.431532897	0.43267859
18	0.151809043	0.686019322
19	0.645182388	1.210587738
20	0.018898663	0.230682213

H_0: Two samples are coming from the same population.

H_1: Two samples are not coming from the same population.

Frequencies

	Flag	N
Sample	Sample1	20
	Sample2	20
	Total	40

Test Statistics[a]

		Sample
Most Extreme Differences	Absolute	.400
	Positive	.400
	Negative	.000
Kolmogorov-Smirnov Z		1.265
Asymp. Sig. (2-tailed)		.082

a. Grouping Variable: Flag

Conclusion: We see that p-value is 0.082 > 0.05, hence we fail to reject H_0 and conclude that the 2 samples are coming from the same population.

Chi – Square Tests

Chi-Square test is perhaps the most widely used statistical test. Most of the common Chi – Square tests viz. goodness of fit, independence of attributes etc. do not assume any assumption on the shape of the distribution and hence are nonparametric in nature. For this case study we will be dealing with the following four Chi-Square tests.

Karl Pearson's Goodness of Fit

It is very powerful non-parametric test for testing the significance of discrepancy between theory and experiment. It enables to find out if the deviation of the experiment form theory is just by chance or is it really to inadequacy of the theory to fit the observed data. Let $X_1 \ldots X_n$ be a random sample of size n from ~ F_X. We are interested in testing a hypothesis if $F_X = F_X^0$ where F_X^0 is some hypothetical distribution. Let us set up the null and alternative hypotheses as follows:

$$H_0: F_X = F_X^0$$

$$H_1: F_X \neq F_X^0$$

Test Statistic: Classify the sample data into k different groups/classes and observe the frequencies (; i = 1… k) in the different classes. Now obtain the expected frequencies (; i = 1… k) for these groups using the distribution as F_X^0. Define

$$\chi^2 = \sum_{i=1}^{K} \left[\frac{(f_i - e_i)^2}{e_i}\right] \sim \chi^2_{n-1}$$

where $\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} e_i$

Decision Rule: Reject H$_0$ if$\chi^2 \leq \chi_{\propto/2}(n-1)$ where $\chi^2{}_{\propto}(n-1)$ is obtained by solving the following.

$$P(\chi^2 \geq \chi_{\propto}{}^2(n-1) \mid H_0) \leq \propto$$

Case 9: A sample survey of 800 families each with 4 children was conducted and following distribution was observed.

# of Male Children	# of Female Children	Type	# of Families
0	4	1	32
1	3	2	178
2	2	3	290
3	1	4	236
4	0	5	64

Is the observed distribution consistent with the hypothesis that male and female births are equally probable?

H$_0$: Male and female birth are equally probable.

H$_1$: Male and female births are not equally probable.

Type

	Observed N	Expected N	Residual
1	32	50.0	-18.0
2	178	200.0	-22.0
3	290	300.0	-10.0
4	236	200.0	36.0
5	64	50.0	14.0
Total	800		

Test Statistics

	Type
Chi-Square	19.633[a]
df	4
Asymp. Sig.	.001

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 50.0.

Conclusion: We see p-value is 0.001 < 0.05, hence we reject H_0 and conclude that male and females births are not equally probable.

Independence of Attributes

The Chi-Square test for independence of attributes is based on exactly the same ideas as the goodness of fit test. Indeed they are the same tests.

Case 10: Out of 8000 graduates in a town 800 are females, out of 1600 graduate employees 120 are females.

We will use Chi – Square test to test if any sex discrimination is made in the employment.

Sex	Employment	Frequency
Male	Employed	1480
Male	Not Employed	5720
Female	Employed	120
Female	Not Employed	680

H_0: No sex discrimination appears in employment.

H_1: Sex discrimination appears in employment.

Sex * Employment Crosstabulation

			Employment		
			Employed	Not Employed	Total
Sex	Female	Count	120	680	800
		Expected Count	160.0	640.0	800.0
	Male	Count	1480	5720	7200
		Expected Count	1440.0	5760.0	7200.0
Total		Count	1600	6400	8000
		Expected Count	1600.0	6400.0	8000.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	13.889[a]	1	.000		
Continuity Correction[b]	13.544	1	.000		
Likelihood Ratio	14.785	1	.000		
Fisher's Exact Test				.000	.000
N of Valid Cases	8000				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 160.00.

b. Computed only for a 2x2 table

Conclusion: We see p-value is 0.00 < 0.05, hence we reject H_0 and conclude that sex discrimination is appears in employment.

McNemar's Test

McNemar's test is a statistical test used on nominal data. It is applied to 2 × 2 contingency tables with a dichotomous trait, with "matched pairs" of subjects, to determine whether the row and column marginal frequencies are equal ("marginal homogeneity"). It is named after Quinn McNemar(1947).

The test is applied to a 2 × 2 contingency table, which tabulates the outcomes of two tests on a sample of n subjects, as follows:

	Test 2 positive	Test 2 negative	Row total
Test 1 positive	a	b	a + b
Test 1 negative	c	d	c + d
Column total	a + c	b + d	N

The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same, i.e. $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$. Thus the null and alternative hypotheses are given as follows:

$$H_0: p_b = p_c$$

$$H1: p_b \neq p_c$$

Here p_a, etc., denote the theoretical probability of occurrences in cells with the corresponding label. The McNemar test statistic is given as follows:

$$\chi^2 = \frac{(b - c)^2}{b + c}$$

The statistic with Yates's correction for continuity is given by:

$$\chi^2 = \frac{(|b - c| - 0.5)^2}{b + c}$$

Under the null hypothesis, with a sufficiently large number of discordances (cells b and c), χ^2 has a chi-squared distribution with 1 degree of freedom. If either b or c is small ($b + c < 25$) then χ^2 is not well-approximated by the chi-squared distribution and we then use the Binomial distribution.

If the $\chi 2$ result is significant, this provides sufficient evidence to reject the null hypothesis, in favour of the alternative hypothesis that $p_b \neq p_c$, which would mean that the marginal proportions are significantly different from each other.

Case 11: A researcher attempts to determine if a drug has an effect on a particular disease. Counts of individuals are given in the table, with the diagnosis (disease: present or absent) before treatment given in the rows, and the diagnosis after treatment in the columns. The test requires the same subjects to be included in the before-and-after measurements (matched pairs).

Effect of Treatment	**After:** present	**After:** absent	Row total
Before: present	101	121	222
Before: absent	59	33	92
Column total	160	154	314

Using McNemar's Test, test the hypothesis of "marginal homogeneity", i.e. there was no effect of the treatment.

H_0: There is no significant effect of treatment on disease.

H_1: There is significant effect of treatment on disease.

Before * After Crosstabulation

			After		
			Absent	Present	Total
Before	Absent	Count	33	59	92
		Expected Count	45.1	46.9	92.0
	Present	Count	121	101	222
		Expected Count	108.9	113.1	222.0
Total		Count	154	160	314
		Expected Count	154.0	160.0	314.0

Chi-Square Tests

	Value	Exact Sig. (2-sided)
McNemar Test		.000[a]
N of Valid Cases	314	

a. Binomial distribution used.

Conclusion: We see p-value is 0.00 < 0.05, hence we conclude that there is significant effect of treatment on disease.

Cochran – Mantel – Haenszel Test

The Cochran–Mantel–Haenszel test (which is sometimes called the Mantel–Haenszel test) is used for repeated tests of independence. There are three nominal variables; we want to know whether two of the variables are independent of each other, and the third variable identifies the repeats. We will cover only the case when repeated individual tests for independence have 2×2 only contingency tables.

Null and Alternative Hypotheses: The null hypothesis is that the two nominal variables that are tested within each repetition are independent of each other; having one value of one variable does not mean that it's more likely that we will have one value of the second variable.

The null hypothesis of the Cochran–Mantel–Haenszel test is that the odds ratios within each repetition are equal to 1. The odds ratio is equal to 1 when the proportions are the same, and the odds ratio is different from 1 when the proportions are different from each other.

If the four numbers in a 2×2 test of independence are labelled like this:

a	b
c	d

and $(a + b + c + d) = n$, the equation for the Cochran–Mantel–Haenszel test statistic can be written like this:

$$\chi^2_{MH} = \frac{\left[\left|\sum\left\{a - \frac{(a+b)(a+c)}{n}\right\}\right| - 0.5\right]^2}{\sum(a + b)(a + c)(b + d)(c + d)/(n^3 - n^2)}$$

The numerator contains the absolute value of the difference between the observed value in one cell (a) and the expected value under the null hypothesis, $(a + b)(a + c)/n$, so the numerator is the squared sum of deviations between the observed and expected values. It doesn't matter how we arrange the 2×2 tables, any of the four values can be used as a. The 0.5 is subtracted as a continuity correction. The denominator contains an estimate of the variance of the squared differences.

The test statistic, $\chi_{MH}{}^2$, gets bigger as the differences between the observed and expected values get larger, or as the variance gets smaller (primarily due to the sample size getting bigger). It is chi-square distributed with one degree of freedom.

Some statisticians recommend that we should test the homogeneity of the odds ratios in the different repeats, and if different repeats show significantly different odds ratios, then we shouldn't do the Cochran–Mantel–Haenszel test. The most common way to test the homogeneity of odds ratios is with the Breslow–Day test.

Other statisticians tells us that it's perfectly okay to use the Cochran–Mantel–Haenszel test when the odds ratios are significantly heterogeneous. The different recommendations depend on what is our goal. If our main goal is hypothesis testing, then the Cochran–Mantel–Haenszel test is perfectly appropriate. If our main goal is estimation, then it would be inappropriate to combine the data using the Cochran–Mantel–Haenszel test.

Case 12: McDonald and Siebenaller (1989) surveyed allele frequencies at the Lap locus in the mussel Mytilus trossulus on the Oregon coast. At four estuaries, samples were taken from inside the estuary and from a marine habitat outside the estuary. There were three common alleles and a couple of rare alleles; based on previous results, the biologically interesting question was whether the Lap ("94") allele was less common inside estuaries, so all the other alleles were pooled into a "non-Lap"("non-94") class.

There are three nominal variables: allele (94 or non-94), habitat (marine or estuarine), and area (Tillamook, Yaquina, Alsea, or Umpqua). The following table shows the number of 94 and non-94 alleles at each location.

Location	Allele	Marine	Estuarine
Tillamook	94	56	69
	non-94	40	77
Yaquina	94	61	257
	non-94	57	301
Alsea	94	73	65
	non-94	71	79
Umpqua	94	71	48
	non-94	55	48

Using Cochran–Mantel–Haenszel test, test the null hypothesis that at each area, there is no difference in the proportion of Lap alleles between the marine and estuarine habitats, after controlling for area.

H_0: There is no difference in the proportion of Lap alleles between the marine and estuarine habitats, after controlling for area.

H_1: There is difference in the proportion of Lap alleles between the marine and estuarine habitats, after controlling for area.

Allele * Habitat * Location Crosstabulation

Count

Location			Habitat: Estuarin	Habitat: Marine	Total
Alsea	Allele	94	65	73	138
		Non-94	79	71	150
Tillamook	Allele	94	69	56	125
		Non-94	77	40	117
Umpqua	Allele	94	48	71	119
		Non-94	48	55	103
Yaquina	Allele	94	257	61	318
		Non-94	301	57	358
Total	Allele	94	439	261	700
		Non-94	505	223	728
	Total		944	484	1428

Tests of Conditional Independence

	Chi-Squared	df	Asymp. Sig. (2-sided)
Cochran's	5.338	1	.021
Mantel-Haenszel	5.050	1	.025

Under the conditional independence assumption, Cochran's statistic is asymptotically distributed as a 1 df chi-squared distribution, only if the number of strata is fixed, while the Mantel-Haenszel statistic is always asymptotically distributed as a 1 df chi-squared distribution. Note that the continuity correction is removed from the Mantel-Haenszel statistic when the sum of the differences between the observed and the expected is 0.

Conclusion: We see p-value is 0.021 < 0.05, hence we reject H_0 and conclude that there is difference in the proportion of Lap alleles between the marine and estuarine habitats, after controlling for area.