Chi-Square Goodness-of-Fit Test

The test is applied when you have one categorical variable from a single population.
It is used to determine whether sample data are consistent with a hypothesized distribution.

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60%, veterans; and 10%, All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company.

Test statistic. Karl Pearson in 1900
The test statistic is a chi-square random variable (X^2) defined by the following equation.

$$X^2 = \Sigma \left[(O_i - E_i)^2 / E_i \right] \text{ ----- k categories gives } k - 1 \text{ degrees of freedom } \ldots$$

where O_i is the observed frequency count for the *i*th level of the categorical variable, and E_i is the expected frequency count for the *i*th level of the categorical variable. Sample Size Concerns – expected counts >= 5

Suppose a randomly-selected package of cards has 50 rookies, 45 veterans, and 5 All-Stars.
Is this consistent with Company's claim? Use a 0.05 level of significance.

Row	type	freq	nullvalues
1	rookie	50	0.3
2	veterans	45	0.6
3	all-star	5	0.1

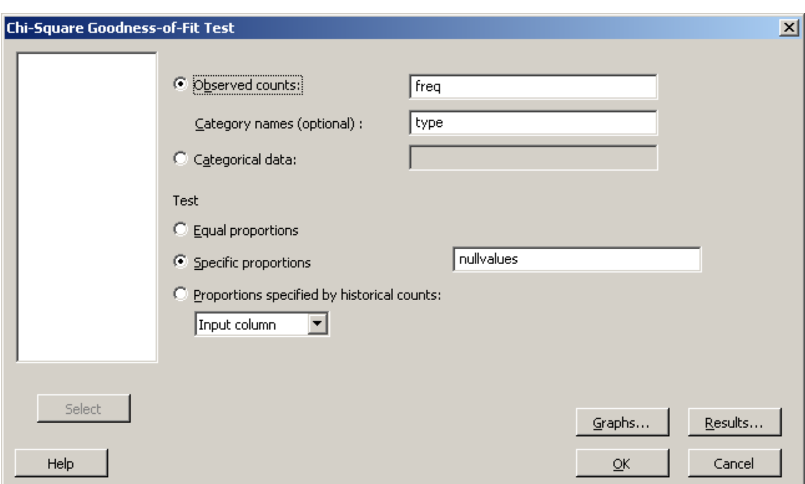
STAT > TABLES > GOF TEST

Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: freq

Using category names in type

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
rookie	50	0.3	30	13.3333
veterans	45	0.6	60	3.7500
all-star	5	0.1	10	2.5000

N	DF	Chi-Sq	P-Value
100	2	19.5833	0.000



Note – If just had the raw data would use the

O Categorical data

Minitab would determine the counts …

The FREQ procedure produces one-way to *n*-way frequency and contingency (cross tabulation) tables.

```
options nocenter ps=60 ls=78 nodate nonumber;
title1 'Chi square goodness of fit test';

data GoodFit;
     input type $ number; cards;
rookie      50
veterans    45
all-star     5
;

proc freq data=GoodFit order=data;
weight number;                        Note: if just had the raw data do not use weight!
tables type / chisq nocum testp=(0.3 0.6 0.1);
run; quit;
```

```
Chi square goodness of fit test

The FREQ Procedure

                                         Test
type        Frequency     Percent     Percent
---------------------------------------------
rookie            50       50.00       30.00
veterans          45       45.00       60.00
all-star           5        5.00       10.00


     Chi-Square Test
for Specified Proportions
-------------------------
Chi-Square        19.5833
DF                      2
Pr > ChiSq         <.0001

Sample Size = 100
```

Chi-square Test for Association

Situation:
A researcher wants to see if there is a "link" or "association" between two categorical (qualitative) variables.
A random sample of subjects is drawn from the population of interest.
The two categorical variables are recorded for each subject in the sample.

Examples: Is there an "association" between:

- which hand is dominant & which eye is dominant?
- which hand is dominant & involvement in an auto accident
- which hand is dominant & whether or not a baseball player makes it to the major leagues
- which hand is dominant & whether or not a golfer makes it to the professional circuit

Example -- dominant hand vs. premature birth status:
In a sample of 1000 people, 100 said they were left-handed and 900 said they were right-handed. Of the lefties, 27 were born prematurely and the other 73 were not. Of the righties, 23 were born prematurely and the other 877 were not. Do these results provide convincing evidence that there is an association between a person's dominant hand and whether or not he/she was born prematurely?

For descriptive purposes, we would display the data in a two-way table
(also called a 2-way classification table, contingency table, or cross-tabulation).

	Left	Right	Total
Premature	27	23	50
Not prem.	73	877	950
Total	100	900	1000

We would also report some relevant proportions or percentages:

Overall, 10% are left-handed.
Of those born prematurely, 54% are left-handed.
Of those not born prematurely, only 7.7% are left-handed.

The discrepancy between 54% and 7.7% with the overall percentage of 10% suggests that there is some connection between handedness and premature birth. However, we need to do a statistical test to determine whether this discrepancy could be an ordinary coincidence.

The test for a two-way table is called a chi-square test of homogeneity or a chi-square test of independence. The hypotheses are stated a bit differently for these two tests, but the calculations are the same.

Hypotheses – Test of homogeneity:

H_0: The proportion of lefties is the same among those who were born prematurely as it is among those who were not born prematurely.
NO ASSOCIATION B/W PREMATURE STATUS & HANDEDNESS

H_1: The proportion of lefties is different for those who were born prematurely than it is for those who were not born prematurely.
YES ASSOCIATION B/W PREMATURE STATUS & HANDEDNESS

Q: What frequencies would have been expected if the null hypothesis is true?

A: We would have expected 10% in each group to be left-handed, since the overall percentage of lefties is 10%. That means that we'd expect 10% of the 50 who were born prematurely to be lefty: 10% of 50 = 5 lefties. The other 45 born prematurely would be expected to be righties. Similarly, we'd expect 10% of the 950 not born prematurely to be left-handed: 10% of 950 = 95, and 90% of 950 = 855 to be right-handed. Below are these expected cell frequencies based on the null hypothesis.

Expected cell frequencies

	Left	Right	Total
Premature	5	45	50
Not prem.	95	855	950
Total	100	900	1000

Q: Are the differences between the observed frequencies and expected frequencies just due to coincidence? Or, are these differences large enough for us to conclude that the percentage of lefties is different among those born prematurely than for those who aren't born prematurely?

A: Continue with the chi-square test of homogeneity to find out.

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

This statistic follows a chi-square distribution with (R-1)(C-1) degrees of freedom, where R is the number of rows and C is the number of columns.

O denotes the observed frequency or count for each cell
E denotes the expected frequency or count (if H_0 is true).

A range for the p-value can be found using tables or exactly using technology. The test is always right-tailed.

Let's test at $\alpha = 0.05$ in this example.
Since there are 2 rows and 2 columns, that means we have df = (2-1)(2-1) = 1.

Calculate the value of the test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(27-5)^2}{5} + \frac{(23-45)^2}{45} + \frac{(73-95)^2}{95} + \frac{(877-855)^2}{855}$$

$$= \frac{(22)^2}{5} + \frac{(-22)^2}{45} + \frac{(-22)^2}{95} + \frac{(22)^2}{855} = 96.80 + 10.76 + 5.09 + 0.57 = 113.22$$

Determine the p-value. From the table, we can say that the p-value is smaller than 0.005.
Make a statistical decision and state a conclusion:

Reject H_0.
Conclude that the percentage of lefties among those born prematurely is different than for those who were not born prematurely.

(In a 2x2 table, we can be more specific.
(In this case we can conclude that the percentage of lefties is greater among those who are born prematurely.)

Notes:

- In order for the chi-square distribution to apply accurately, the expected frequencies should all be at least 5.
- A formula for calculating the expected frequencies is:
  $$E = \frac{(\text{Row Total})(\text{Column Total})}{n}$$
  (This is especially useful for larger tables where there are more than 2 rows and 2 columns.)
- For a test of homogeneity, the goal is to compare proportions for two or more populations. For a test of independence, the goal is to find out if there is any association between two qualitative variables. The hypotheses for a test of independence are stated in the following manner:

H_0: The variables … are independent. No Association b/w …
H_A: The variables … are dependent. Yes Association b/w …

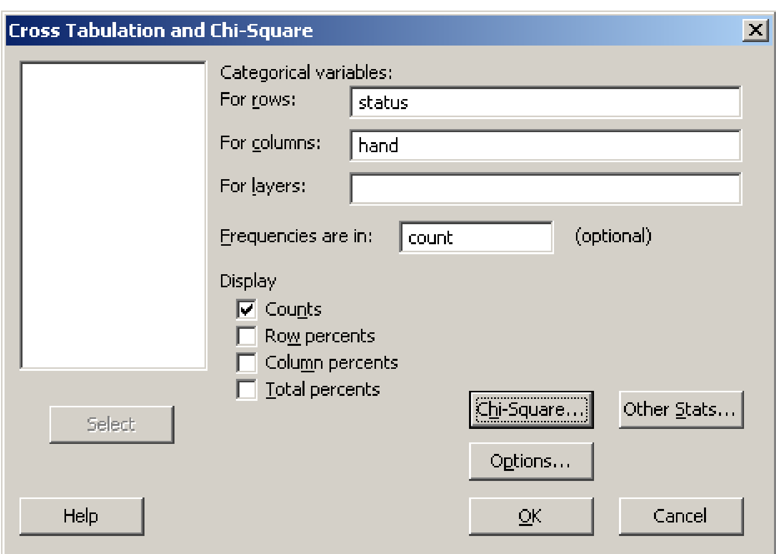
In the previous example, for a test of independence we would have used –
H_0: Dominant hand and premature birth status are independent.
H_A: Dominant hand and premature birth status are dependent.

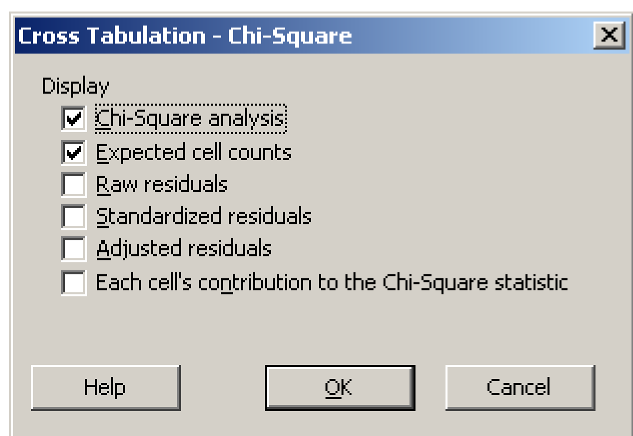
```
Row  status   hand   count
  1  pre      left      27
  2  pre      right     23
  3  pre-not  left      73
  4  pre-not  right    877
```

Tabulated statistics: status, hand

```
Using frequencies in count


Rows: status   Columns: hand

           left  right   All

pre          27     23    50
              5     45    50

pre-not      73    877   950
             95    855   950

All         100    900  1000
            100    900  1000

Cell Contents:      Count
                    Expected count


Pearson Chi-Square = 113.216, DF = 1, P-Value = 0.000
Likelihood Ratio Chi-Square = 66.294, DF = 1, P-Value = 0.000
```

STAT > TABLES > CROSS TABULATION …





Note: if raw data no need for the using Frequencies

```
options nocenter ps=60 ls=78 nodate nonumber;
title1 'Chi Square Two-way Table test';

data example;
input status $ hand $ number; cards;
pre      left      27
pre      right     23
pre-not  left      73
pre-not  right    877
;

proc freq; title3 'Chi square analysis using Proc Freq';
weight number;                            Note: if just raw data no need for weight
tables status*hand / chisq expected cellchi2 norow nocol nopercent;
run;
```

```
Chi Square Two-way Table test

Chi square analysis using Proc Freq

The FREQ Procedure

Table of status by hand

status          hand

Frequency      |
Expected       |
Cell Chi-Square|left    |right   |  Total
---------------+--------+--------+
pre            |     27 |     23 |     50
               |      5 |     45 |
               |   96.8 | 10.756 |
---------------+--------+--------+
pre-not        |     73 |    877 |    950
               |     95 |    855 |
               | 5.0947 | 0.5661 |
---------------+--------+--------+
Total               100      900    1000


Statistics for Table of status by hand

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1    113.2164    <.0001
```

EXAMPLE

In the paper "Color Association of Male and Female Fourth-Grade School Children" (J. Psych., 1988, 383-388), children were asked to indicate what emotion they associated with the color red. The response and the sex of the child are noted and summarized below.

	anger	happy	love	pain
female	26	19	38	17
male	30	11	34	25

[a] What emotion (anger, happy, love, pain) do you associate with the color red?

[b] What is the marginal distribution for the emotion associated with the color red?

[c] What are the conditional distributions for the emotion associated with the color red?

- conditional distribution given female
- conditional distribution given male

[d] Are the three distributions in [b] and [c] the same?

[e] At the 1% significance level, do the data provide sufficient evidence to conclude that
[e] an association exists between sex and the emotion associated with the color red in the population?

EXAMPLE

A controversial issue in sports is the use of the "instant replay" for making decisions on plays that are extremely close or hard to call by an official. A survey of players in each of four professional sports was conducted, asking them if they felt "instant replays" should be used to decide close or controversial calls.

The results are as follows:

	Favor -- Instant Replay	Oppose -- Instant Replay
Baseball (MLB)	12	12
Soccer (MLS)	4	8
Basketball (NBA)	16	24
Football (NFL)	18	6

[a] What is the marginal distribution for the "feelings" about "instant replays"?

[b] What are the conditional distributions for the "feelings" about "instant replays"?

- conditional distribution given NFL
- conditional distribution given MLB
- conditional distribution given NBA
- conditional distribution given MLS

[c] Are the five distributions in [a] and [b] the same?

[d] At the 5% significance level, do the data provide sufficient evidence to conclude that
[d] an association exists between sport and "feelings" about "instant replays"?