Statistics: Continuous Methods

STAT452/652, Spring 2013

Computer Lab 3

Thursday, 14 February, 2013
DMS 106
1:00 – 2:15PM

Goodness of Fit tests:

Chi-square, Kolmogorov-Smirnov,
Anderson-Darling, Shapiro-Wilk
with

Instructor: Ilya Zaliapin

Topic: Goodness of Fit (GoF) tests

Goals: Learn how to use and interpret the following tests

- Chi-square
- Kolmogorov-Smirnov, Anderson-Darling, Shapiro-Wilk

Assignments:

Use the data file Lab3_data_sets.MTW from the lab webpage.

1. Chi-square test for a uniform [0,1] sample in column C1. Perform the chi-square test with 3, 5, and 10 categories. Find the P-values, compare, discuss.

2. Anderson-Darling test for two exponential samples in columns C2 and C3. (Use the probability plot option, which shows the AD test results.) The two samples are from the exponential distribution with mean 5; C2 has length 100, while C3 has length 1000. For each sample, perform the AD test several times for different means of the theoretical distribution and find the limits of the mean corresponding to the P-value of above 5%. Compare results, discuss.

3. Chi-square test for a Normal(10,4) sample in column C4. Perform the chi-square test for the normal sample, using the cdf transformation that produces a uniform sample. Compare the chi-square P-value with that of KS, AD, and SW tests. Discuss.

4. Chi-square test for three samples in columns C7-C9. One of these samples in from the U[0,1] distribution, second significantly deviates from the U[0,1], and the third is overfit to the U[0,1]. Use the chi-square test to decide which sample is which.

Report:

A printed report for this Lab is due on **Thursday, February 21** in class. BW printouts are OK. Reports will not be accepted by mail.

1. Introduction

The methods considered in this Lab are focused on the following problem: Given a sample X_i, i=1,…,n and a distribution (cdf) $F(x)$ decide whether the sample is coming from this distribution. In other words we test the hypothesis

H_0: {X_i}, i =1,…,n are from the distribution $F(x)$

versus an alternate hypothesis

H_a: {X_i}, i =1,…,n are *not* from the distribution $F(x)$.

The considered methods will complement visual analysis that uses histograms, dot plots, ecdfs, and probability plots. We will focus here on continuous distributions $F(x)$.

2. Chi-square test

Recall that the essence of the chi-square test is to consider the observed numbers O_i of the sample values within the predefined k bins [x_{i-1}, x_i], i=1,…,k, and compare them with the expected numbers

$$E_i = n[F(x_i) - F(x_{i-1})].$$

The test statistic used in the chi-square test is defined as

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

It can be shown that if all E_i > 5 and we have sufficiently large sample, the statistic χ^2 is distributed approximately as a chi-square random variable with (k-1) degrees of freedom: χ^2_{k-1}.

Preliminary data transformations

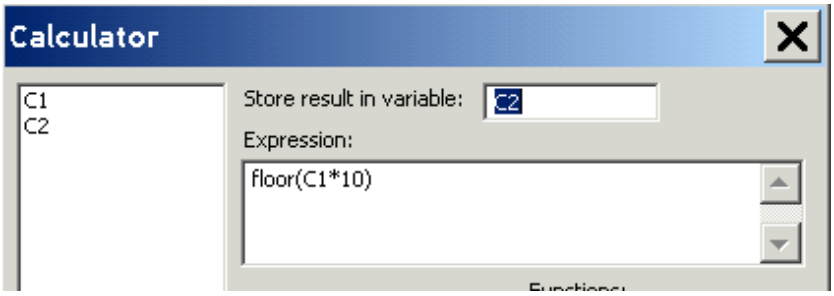
The chi-square test in Minitab works with *multinomial distributions*, thus if we want to apply it to continuous data, some preliminary work should be done. Specifically, we need to transform our continuous sample into a multinomial sample. For that we need to code our data replacing each sample value with the category (bin) index this value belongs to.

Case 1: $F(x)$ is the uniform distribution on [0,1]

For the uniform distribution it is convenient to choose an equidistant binning:

[0, 1/k), [1/k, 2/k), …, [1-1/k, 1).

The coding can be done using the **Calc/Calculator:**



The function ***floor(X*k)*** will map the sample values to their class indices (using *k* equidistant bins.)

Case 2: Arbitrary continuous distribution $F(x)$

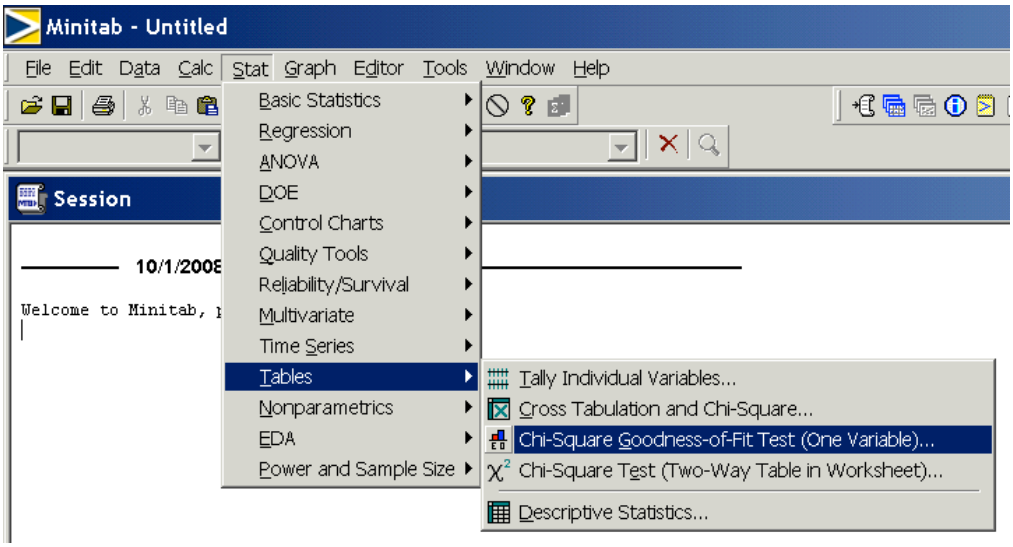
If $F(x)$ is not the uniform [0,1], we define

$$U_i = F(X_i).$$

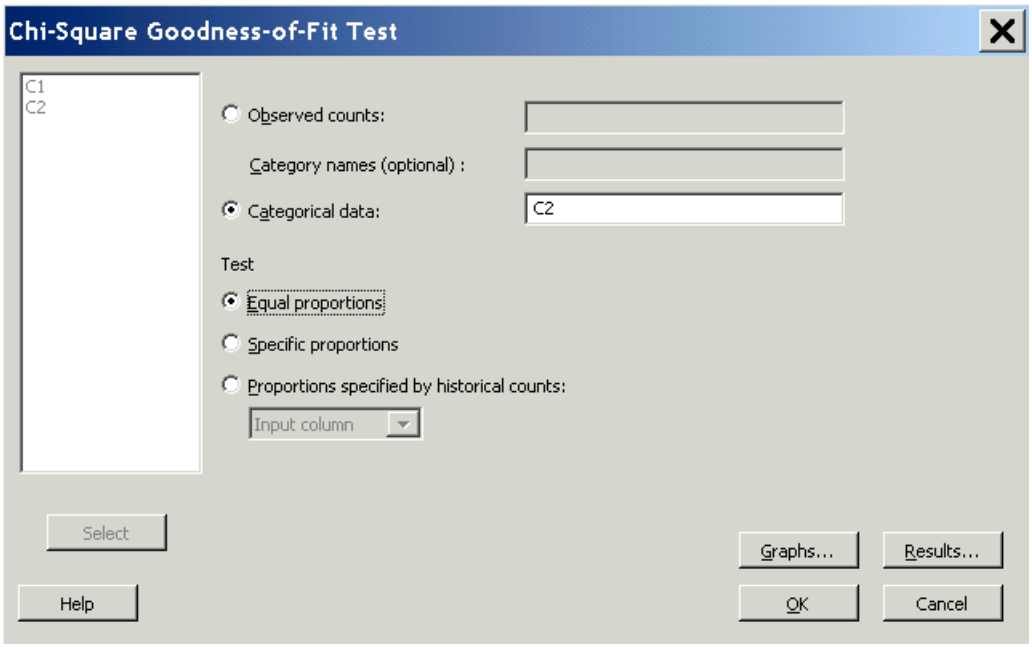
If the sample is actually from the distribution $F(x)$ (if the null hypothesis is true), then the sample U_i has the uniform distribution on [0,1] and we can apply the technique of Case 1.

The chi-square test is implemented in menu

Stat/Tables/Chi-Square Goodness-of-Fit test (One Variable):



We work with the index variable (not the original one!), and choose the **Categorical data** radio button; the test is **Equal proportions:**



The test results in some graphs (will be discussed in class) and the following (or similar) session outcome, with details of the analysis, and the resulting P-value:

Chi-Square Goodness-of-Fit Test for Categorical Variable: C2

Category	Observed	Test Proportion	Expected	Contribution to Chi-Sq
0	11	0.1	10	0.1
1	20	0.1	10	10.0
2	7	0.1	10	0.9
3	8	0.1	10	0.4
4	7	0.1	10	0.9
5	10	0.1	10	0.0
6	11	0.1	10	0.1
7	6	0.1	10	1.6
8	11	0.1	10	0.1
9	9	0.1	10	0.1

N	N*	DF	Chi-Sq	P-Value
100	0	9	14.2	0.115

Advantage of chi-square test:

- Can work with any distribution (discrete or continuous).

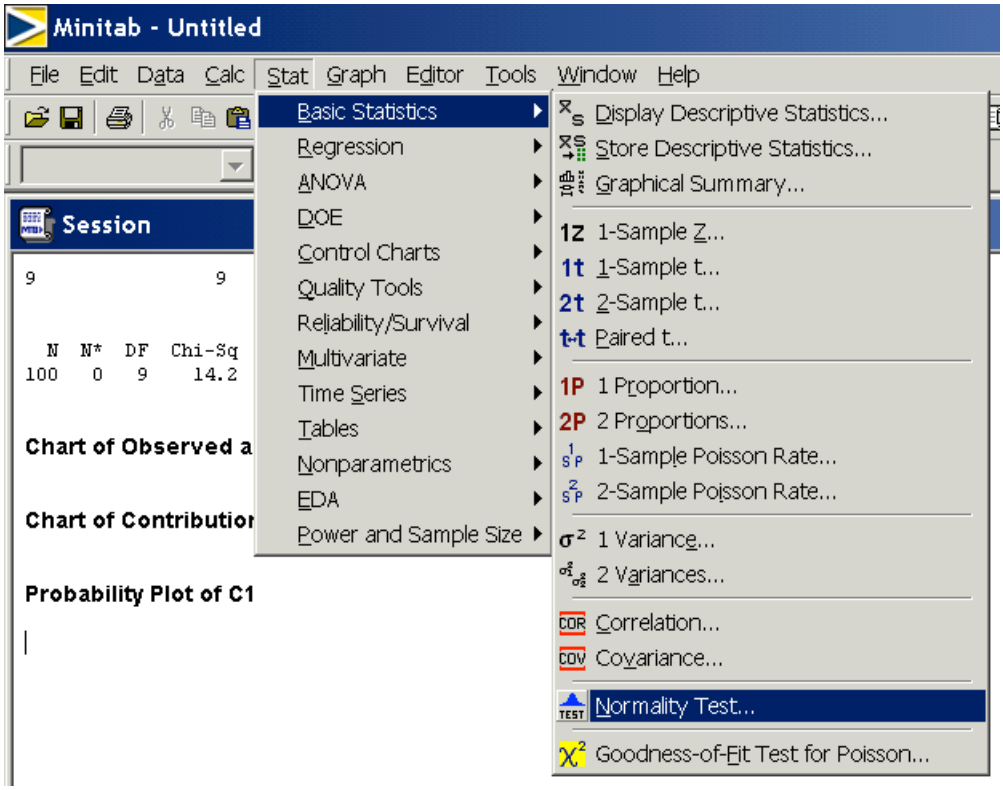
Disadvantages of chi-square test:

- Requires a large number of observations (to ensure convergence).
- Results depend on the chosen bins.

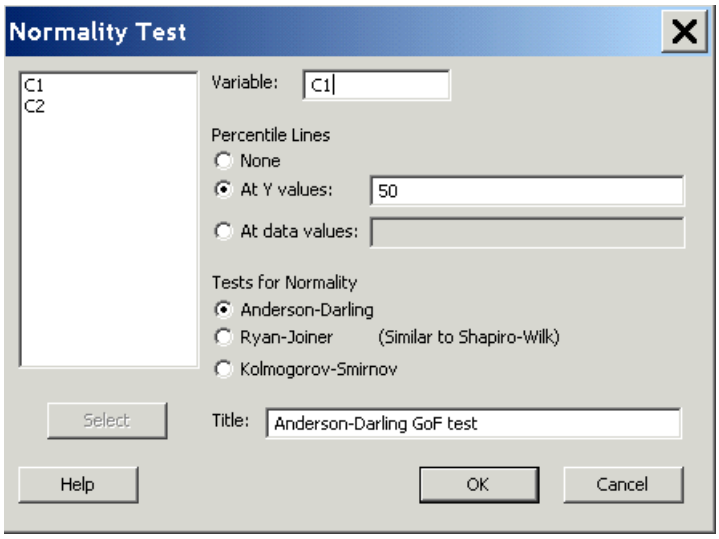
2. Kolmogorov-Smirnov, Anderson-Darling, and Shapiro-Wilk tests

These tests are implemented in Minitab only for testing the Normal distribution (although KS ad AD tests can be applied to other distributions as well).

The tests can be accessed via menu **Stat/Basic Statistics/Normality Test:**



In the next menu, you choose the variable to analyze, test to perform, and some other options that will be discussed in class:



The test results are summarized in the output figure: