

STAT 518 --- Chapter 6 --- Goodness-of-Fit Tests

- Often in statistics, we assume a sample comes from a particular distribution.
- Goodness-of-fit tests help us determine whether the assumed distribution is reasonable for the data we have.

Section 6.1: Kolmogorov Goodness-of-Fit Test

- Recall that the empirical distribution function (e.d.f.) of a sample is an estimate of the cumulative distribution function (c.d.f.)

for the population that the sample came from.

- If $S(x)$ is close to the c.d.f. $F^*(x)$ of our assumed distribution, then our assumption is reasonable.
- If $S(x)$ is far from the c.d.f. $F^*(x)$ of our assumed distribution, then our assumption should be rejected.

Picture:

- **How to measure the distance between $S(x)$ and $F^*(x)$?**

- **Kolmogorov suggested using the maximum vertical discrepancy between $S(x)$ and $F^*(x)$ as a test statistic:**

- **In Chapter 6 we will see several tests that use a type of maximum vertical discrepancy.**

- **For the Kolmogorov Goodness-of-Fit test, we assume only that we have a random sample X_1, X_2, \dots, X_n .**

Test Statistic (depends on the alternative hypothesis):

- **The null distribution of T is tabulated in Table A13 for $n \leq 40$.**

- **This approach is _____ if $F(x)$ is continuous and _____ if $F(x)$ is discrete.**

- **Pages 435-436 describe an adjustment to improve the test if $F(x)$ is discrete, but we will not cover this.**

- **If $n > 40$, an asymptotic null distribution can be used (see equation (5) on page 431).**

Possible Hypotheses and Decision Rules

H₀:

H₀:

H₀:

H₁:

H₁:

H₁:

- The corresponding rejection rules in each case are:

- The P-values for each case are approximated by interpolation within Table A13 or found using R.

Example 1: Ten observations are obtained in a sample which supposedly comes from a Uniform(0, 1) distribution. The sorted sample is: 0.203, 0.329, 0.382, 0.477, 0.480, 0.503, 0.554, 0.581, 0.621, 0.710. Is there evidence that the hypothesized distribution is incorrect? (Use $\alpha = 0.05$.)

Picture:

Example 2: A medical team collecting counts of tumors on kidneys has used a Poisson(1.75) distribution to model the counts in the past. They gather such counts on 18 kidneys and obtain this sample: 2, 2, 4, 1, 3, 1, 4, 0, 2, 2, 1, 1, 0, 2, 2, 3, 3, 3. Is there evidence that the count distribution is actually stochastically larger than previously thought? (Use $\alpha = .05$.)

In R, see the function `ks.test` to perform this test.

A Confidence Band for the True Population c.d.f.

- From Table A13, it is easy to obtain an upper function and lower function that form a $(1 - \alpha)100\%$ confidence band.**
- We can be, say, 95% confident that the entire true c.d.f. will fall within the 95% confidence band.**

- To form the band, simply draw the e.d.f $S(x)$, and add to each point to form the upper boundary of the band.
- Subtract $\frac{1}{\sqrt{n}}$ from $S(x)$ at each point to form the lower boundary.
- If $\frac{1}{\sqrt{n}}$ goes below 0, simply draw the lower boundary at 0 in those places.
- If $\frac{1}{\sqrt{n}}$ exceeds 1, simply draw the upper boundary at 1 in those places.
- See example in R using Example 1 data.

Properties of the Kolmogorov Test

- The two-sided Kolmogorov test is more powerful against all differences between the hypothesized $F^*(x)$ and the true $F(x)$.
- However, the test is less powerful for finite sample sizes.
- The Kolmogorov test is more powerful than the chi-squared goodness-of-fit test (covered in Chapter 4) when the data are ordinal.
- The Kolmogorov test is exact even for small samples, while the chi-square test requires a large sample size.

Sec. 6.2 --- Lilliefors Goodness-of-Fit Tests

- With the Kolmogorov test, the hypothesized distribution must be completely specified, including parameter values.
- In some cases, we may want to test whether the data may come from some distribution (e.g., normal, exponential, etc.) without knowing what the specific parameter values may be.
- Lilliefors introduced Kolmogorov-type tests to allow testing goodness of fit when the parameters are not specified, but rather _____ from the data.

Lilliefors Test for Normality

- Assume that we have a random sample X_1, X_2, \dots, X_n .

Hypotheses: H_0 :

vs. H_1 :

- Compute the normalized sample values

where

• Let $S(x)$ be the e.d.f. of the Z_i 's, and let $F^*(x)$ be the c.d.f. of the standard normal distribution; then the Lilliefors test statistic is:

- If T_1 is excessively _____, then H_0 is rejected.
- The null distribution of T_1 is unknown and has been approximated by random number generation.

Using Table A14, reject H_0 if

The P-value is approximated via interpolation in Table A14, or by using R.

Example 1: The data on page 445 consist of 50 observations. At $\alpha = 0.05$, is it reasonable to claim that the data follow a normal distribution?

Test Statistic: $T_1 =$

Decision Rule:

P-value

Conclusion:

See examples on simulated data on course web page.

Lilliefors Test for Exponential Distribution

- **Assume that we have a random sample X_1, X_2, \dots, X_n .**

Hypotheses: H_0 :

vs. H_1 :

- **Compute the standardized sample values**
- **Let $S(x)$ be the e.d.f. of the Z_i 's, and let $F^*(x)$ be the c.d.f. of the standard exponential distribution:**
- **Then the Lilliefors test statistic is:**
- **If T_2 is excessively _____, then H_0 is rejected.**
- **The null distribution of T_2 is complicated, but has been tabulated in Table A15.**

Using Table A15, reject H_0 if

The P-value is approximated via interpolation in Table A15.

Example 2: The exponential distribution is a common model for waiting times between occurrences of some random phenomenon. A built-in data set in R gives 272 waiting times between eruptions of the Old faithful geyser. At $\alpha = 0.05$, is it reasonable to claim that the waiting times follow an exponential distribution?

Test Statistic: $T_2 =$

Decision Rule:

Conclusion:

- Lilliefors also presented a similar goodness-of-fit test for a gamma distribution with unknown parameter values. More such tests could be developed similarly.
- The Shapiro-Wilk test (`shapiro.test` in R) is another test for normality, based not on the e.d.f. but on the correlation between observed order statistics and expected order statistics under normality.
- The Shapiro-Wilk test tends to have better power than the Lilliefors test for detecting departures from normality.

Sec. 6.3 --- Smirnov Test for Two Samples

- Suppose we have independent random samples (denoted X_1, \dots, X_n and Y_1, \dots, Y_m) from two populations.
- The Smirnov test (also called the Kolmogorov-Smirnov two-sample test) uses the e.d.f. of each sample to test whether the two samples come from the same distribution.
- Let $F(\cdot)$ represent the c.d.f. of the X_i 's and let $G(\cdot)$ represent the c.d.f. of the Y_i 's.
- Assume the samples are mutually independent and the measurement scale is at least ordinal.
- Let $S_1(x)$ represent the e.d.f. of the X_i 's and let $S_2(x)$ represent the e.d.f. of the Y_i 's.

Possible Hypotheses and Decision Rules

H_0 :

H_0 :

H_0 :

H_1 :

H_1 :

H_1 :

These alternative hypotheses can also be stated as:

H_1 :

H_1 :

H_1 :

Test Statistic (depends on the alternative hypothesis):

- The null distributions of T_1 , T_1^+ and T_1^- are based on the fact that all orderings of X 's and Y 's are equally likely if H_0 is true. The null distributions are tabulated in Table A19 for $n = m$ and in Table A20 if $n \neq m$.

- The corresponding rejection rules in each case are:

- The P-values for each case are approximated by interpolation within the tables or found using R.

- If $n = m$, equation (5) on page 458 can be used to obtain a more exact p-value.

- If n is too large for the tables, an approximation given at the end of the tables can be used.

Example 1: Consider the 2 samples given on page 460:

X_i 's: 7.6, 8.4, 8.6, 8.7, 9.3, 9.9, 10.1, 10.6, 11.2

Y_i 's: 5.2, 5.7, 5.9, 6.5, 6.8, 8.2, 9.1, 9.8, 10.8, 11.3, 11.5, 12.3, 12.5, 13.4, 14.6

Is it reasonable to believe that the two samples follow identical distributions?

Test statistic:

P-value

Conclusion:

Example 2: In a test to compare gasoline types, a driver measured gas mileages for four tanks of “Unleaded” (21.7, 21.4, 23.3, 22.8) and for four tanks of “Premium” (23.1, 23.5, 22.9, 23.4). Is there evidence that mileage tends to be less for “Unleaded” than for “Premium”?

H₀:

H₁:

Test statistic:

P-value

Conclusion:

- This Smirnov test is _____ if the data are continuous and _____ if the data are discrete.
- The Mann-Whitney test was another test used to compare two independent samples, but that was sensitive to differences in center.
- The Smirnov test is designed to detect any sort of difference in distribution, and is in fact _____ against any type of departure from the null hypothesis.