Lectures 2 and 3 - Goodness-of-Fit (GoF) Tests

Rui Castro

March 7, 2013

Often times we have some data and want to test if a particular statistical model (or model class) is a good fit. For instance, it is common to make normality assumptions about certain kinds of data for simplicity. Nevertheless one must check if these assumptions are reasonable. In Goodness-of-Fit (GoF) testing we strive to check the compatibility of the data with a fixed single model (simple GoF) or with a model in a given class (composite GoF).

Let $X_1, \ldots, X_n$ be i.i.d. samples from an unknown distribution F. If we wish to infer whether this sample comes from a certain hypothesized distribution F_0 the problem can be cast as the following hypothesis test:

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0 \ .$$

This is known as the *simple* Goodness-of-Fit (GoF)-problem, as the null hypothesis is simple (not-composite). The composite GoF problem arises when we want to test whether the distribution of the sample belongs to a certain class $\mathcal{F}$ of distribution functions. In this case we consider the testing problem

$$H_0 : F \in \mathcal{F} \quad \text{vs.} \quad H_1 : F \notin \mathcal{F} \ .$$

This null hypothesis is composite, which typically makes a formal analysis of testing procedures much more challenging. However, this is in general the type of tests we want to perform in practice. A typical application of such tests arises when we fit a linear model and want to check whether the normality assumption on the residuals is reasonable.

There are several approaches to GoF testing, and we will focus on two of them: (i) procedures based on the empirical CDF; (ii) Chi-squared type tests. Chi-square tests are an obvious choice when the hypothesized distributions (and data) are discrete or categorical, and the empirical CDF methods are typically more adequate for the continuous case.

1 Simple GoF Testing Using the Empirical CDF

The basic idea follows from the properties of the ECDF we seen in the previous lecture. Under H_0, for n sufficiently large, $\hat{F}_n$ is *uniformly* close to F_0. Recall that, under H_0, the Glivenko-Cantelli theorem tells us that

$$\sup_t |\hat{F}_n(t) - F_0(t)| \stackrel{\text{a.s.}}{\to} 0 \ ,$$

as $n \to \infty$. Hence, any discrepancy measure between $\hat{F}_n$ and F_0 can be used as a reasonable test statistic. A good statistic much be somewhat easy to compute and characterize, and give rise to a test that is powerful against most alternative distributions. Here are some important examples.

- The Kolmogorov-Smirnov (KS) test statistic

$$D_n := \sup_t |\hat{F}_n(t) - F_0(t)| \ .$$

- The Cramér-Von Mises (CvM) statistic

$$C_n := \int (\hat{F}_n(t) - F_0(t))^2 dF_0(t) \ .$$

- The Anderson-Darling (AD) statistic

$$A_n := \int \frac{(\hat{F}_n(t) - F_0(t))^2}{F_0(t)(1 - F_0(t))} dF_0(t) \ .$$

For any of these test statistics we should reject the null hypothesis if the observed value of the statistic is somewhat "large". To know what large means we need to characterize their distribution.

Although the above expressions might look somewhat complicated, they can be simplified significantly if F_0 is continuous. Note that $\hat{F}_n$ is piecewise constant (with the order statistics corresponding to the discontinuity points) and F_0 is a non-decreasing function. Therefore the maximum deviation between $\hat{F}_n(t)$ and $F_0(t)$ must occur in a neighborhood of the points $X_{(i)}$, and so

$$D_n = \max_{1 \le i \le n} \max\{|\hat{F}_n(X_{(i)}) - F_0(X_{(i)})|, |\hat{F}_n(X_{(i)}^-) - F_0(X_{(i)}^-)|\} \ ,$$

where $X_{(i)}^- := X_{(i)} - \epsilon$ for an arbitrarily small ϵ (there is a slight abuse of notation here). Now, taking into account that F_0 is continuous $F_0(X_{(i)}^-) = F_0(X_{(i)})$. Furthermore $\hat{F}_n(X_{(i)}) = i/n$ and $\hat{F}_n(X_{(i)}^-) = (i-1)/n$, therefore

$$D_n = \max_{1 \le i \le n} \max\{|i/n - F_0(X_{(i)})|, |(i-1)/n - F_0(X_{(i)})|\} \ .$$

For convenience, define $U_i = F_0(X_i)$, and let $U_{(i)}$ denote the corresponding order statistics (note that, since F_0 is monotone $U_{(i)} = F_0(X_{(i)})$ and the order is not perturbed). We can therefore write the above expression as

$$D_n = \max_{1 \le i \le n} \max\{|i/n - U_{(i)}|, |(i-1)/n - U_{(i)}|\} \ .$$

In a similar fashion we get simplified expressions for the CvM and AD statistics:

$$nC_n = \frac{1}{12n} + \sum_{i=1}^{n} \left(U_{(i)} - \frac{2i-1}{2n} \right)^2 \ , \tag{1}$$

and

$$nA_n = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \left[\log U_{(i)} + \log(1 - U_{(n-i+1)}) \right] \ .$$

In order be able to use these test statistics we must characterize their distribution under the null hypothesis (so that we can either devise tests of a certain level, or compute p-values). At this point it is useful to recall a very important basic result about distribution functions, the *Inverse Probability Transform*.

Proposition 1. *Let F be an arbitrary cumulative distribution function, and define $F^{-1} : [0,1] \to \mathbb{R}$ as*

$$F^{-1}(u) = \inf\{x : F(x) \geq u\} \ .$$

Then

- *If $U \sim$ Unif $([0,1])$ then $X = F^{-1}(U)$ is a random variable with distribution F.*
- *If F is continuous and X has distribution F then $F(X)$ is uniformly distributed over $[0,1]$.*

Now suppose that indeed we are working under H_0, which means the data $\{X_i\}$ are i.i.d. samples from the continuous distribution F_0. In that case U_i are i.i.d. uniform random variables in $[0,1]$. Therefore we conclude that, under H_0, the distribution of D_n, C_n and A_n does not depend on the underlying distribution F_0. In other words these statistics are *distribution free* under the null hypothesis, and so the computation of p-values can be easily done, as it suffices to study the case when the null hypothesis is the uniform distribution in $[0,1]$. This is what we will do next.

To use the above test statistics one needs to characterize their distribution under the null. For small n these have been tabulated, and for large n we can rely on asymptotic properties. The asymptotic analysis is not the easiest, and requires the machinery of empirical processes theory, which is out of the scope of these notes. As mentioned before it suffices to study the case $F_0 = \text{Unif}[0,1]$. Let $U_i \sim \text{Unif}(0,1)$ and define $\hat{U}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{U_i \leq t\}$. Note that in this case

$$\begin{aligned} D_n &\stackrel{\text{D}}{=} \sup_t \left|\hat{U}_n(t) - \mathbb{P}(U \leq t)\right| \\ &= \sup_{t\in[0,1]} \left|\hat{U}_n(t) - t\right| \ . \end{aligned}$$

A well-known result from empirical processes theory states that the process $t \mapsto \sqrt{n}(\hat{U}_n(t) - t)$ converges in distribution to a process B_0, which is known as a *standard Brownian Bridge* on $[0,1]$. This is a Gaussian process defined for $t \in [0,1]$ with $\mathbb{E}[B_0(t)]$ and $\text{Cov}(B_0(s), B_0(t)) = \min\{s,t\} - st$. Now, with a bit of handwaving (a formal treatment requires the use of invariance principles), we have that

$$\sqrt{n}D_n \stackrel{\text{D}}{\to} \sup_t |B_0(t)| \ ,$$

as $n \to \infty$. Similarly

$$nC_n \stackrel{\text{D}}{\to} \int_0^1 B_0^2(t)dt \qquad \text{and} \qquad nA_n \stackrel{\text{D}}{\to} \int_0^1 \frac{B_0^2(t)}{t(1-t)}dt \ .$$

Fortunately, these asymptotic distributions can be studied analytically and we have

$$\lim_{n\to\infty} \mathbb{P}_{F_0}(\sqrt{n}D_n \leq \lambda) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2} \ ,$$

$$\lim_{n\to\infty} \mathbb{P}_{F_0}(nC_n > x) = \frac{1}{\pi}\sum_{j=1}^{\infty}(-1)^{j+1}\int_{(2j-1)^2\pi^2}^{4j^2\pi^2} \sqrt{\frac{-\sqrt{y}}{\sin(\sqrt{y})}}\frac{e^{-xy/2}}{y}dy \ .$$

Finally $nA_n \xrightarrow{\mathrm{D}} A$, with

$$A \stackrel{\mathrm{D}}{=} \sum_{j=1}^{\infty} \frac{Y_j}{j(j+1)} \ ,$$

where $Y_i \overset{\text{i.i.d}}{\sim} \chi^2_1$.

It is interesting to note that DKW inequality also gives us a crude characterization of the distribution of D_n under the null, namely for $\epsilon > 0$

$$\mathbb{P}_{F_0}\left(D_n \geq \epsilon\right) \leq 2\exp(-2n\epsilon^2) \ .$$

Therefore, if the observed value of the KS statistic is d_n the corresponding p-value of the KS test is upper-bounded by $2\exp(-2nd_n^2)$.

1.1 Consistency under the alternative

A very important property of the tests we have derived is that these are consistent under *any* alternative. This means that if the true distribution is not F_0 then eventually, as $n \to \infty$, we will reject the null hypothesis no matter what the true distribution is. We'll see this for the KS test.

Lemma 1. *Consider the KS test. This test rejects H_0 at level $0 < \alpha < 1$ if $D_n \geq \tau_\alpha$, where τ_α is the smallest real for which*

$$\mathbb{P}_{F_0}\left(D_n \geq \tau_\alpha\right) \leq \alpha \ .$$

Now, if the data $\{X_i\}_{i=1}^n$ comes from a distribution $F \neq F_0$ then

$$\mathbb{P}_{F}\left(D_n \geq \tau_\alpha\right) \to 1 \ ,$$

as $n \to \infty$.

Proof. The DKW inequality tells us almost immediately that

$$\tau_\alpha \leq \sqrt{\frac{1}{2n}\log(2/\alpha)} \ .$$

Also, since $F \neq F_0$ there is at least one point $a \in \mathbb{R}$ such that $F_0(a) \neq F(a)$. Putting these two facts together we get

$$\begin{aligned}
\mathbb{P}_F\left(D_n \geq \tau_\alpha\right) &= \mathbb{P}_F\left(\sup_t |\hat{F}_n(t) - F_0(t)| \geq \tau_\alpha\right) \\
&= \mathbb{P}_F\left(\sup_t |\hat{F}_n(t) - F(t) + F(t) - F_0(t)| \geq \tau_\alpha\right) \\
&\geq \mathbb{P}_F\left(|\hat{F}_n(a) - F(a) + F(a) - F_0(a)| \geq \tau_\alpha\right) \\
&\geq \mathbb{P}_F\left(|F(a) - F_0(a)| - |\hat{F}_n(a) - F(a)| \geq \tau_\alpha\right) \ ,
\end{aligned}$$

where the last step follows from the basic inequality $|x+y| \geq |x| - |y|$. Now

$$\begin{aligned}
&\mathbb{P}_F\left(|F(a) - F_0(a)| - |\hat{F}_n(a) - F(a)| \geq \tau_\alpha\right) \\
&= \mathbb{P}_F\left(|\hat{F}_n(a) - F(a)| \leq |F(a) - F_0(a)| - \tau_\alpha\right) \\
&= 1 - \mathbb{P}_F\left(|\hat{F}_n(a) - F(a)| > |F(a) - F_0(a)| - \tau_\alpha\right) \\
&\geq 1 - \mathbb{P}_F\left(|\hat{F}_n(a) - F(a)| \geq |F(a) - F_0(a)| - \tau_\alpha\right) \\
&\geq 1 - 2\exp\left(-2n\left(|F(a) - F_0(a)| - \tau_\alpha\right)^2\right) \ .
\end{aligned}$$

Now since τ_α clearly converges to zero it is immediate to see that

$$\lim_{n\to\infty} 1 - 2\exp\left(-2n\left(|F(a) - F_0(a)| - \tau_\alpha\right)^2\right) = 1 \ ,$$

concluding the proof. □

A similar argument also applies to C_n and A_n, but the DKW inequality can no longer be used. Nevertheless noting for any point $a \in \mathbb{R}$ the CLT implies that $\sqrt{n}|\hat{F}_n(a) - F(a)| = O_P(1)$ (meaning $\forall \delta > 0 \quad \exists c < \infty : P_F(\sqrt{n}|\hat{F}_n(a) - F(a)| \leq c) \geq 1 - \delta$) we can mimic the above proof without using the DKW inequality.

2 Composite GoF tests

As alluded before the composite GoF scenario is significantly more complicated. However, it is also more relevant from a practical standpoint, since we are hardly ever in the situation where we want to test for instance that a sample is from an exponential distribution with parameter 2, or from a normal distribution with parameters 1.05 and 0.56. Often, the composite GoF-problem comes down to testing

$$H_0 \ : \ F \in \underbrace{\{F_\theta \ : \ \theta \in \Theta\}}_{\mathcal{F}} \quad \text{vs.} \quad F \notin \{F_\theta \ : \ \theta \in \Theta\} \ ,$$

where $\Theta \subseteq \mathbb{R}^d$. In other words, the null hypothesis corresponds to distributions in a known parametric class. As an example F_θ may be the exponential distribution with mean θ.

Perhaps the simplest idea to come to mind in this case is to compare the "best" distribution in the class with the empirical CDF. This can be done by estimating the parameter θ from the data (denote this estimator by $\hat{\theta}_n$) and comparing $\hat{F}_n$ with $F_{\hat{\theta}_n}$, where $F \in \mathcal{F}$. Therefore, we get the following analogues of the KS and CvM test statistics,

$$\bar{D}_n = \sup_t |\hat{F}_n(t) - F_{\hat{\theta}_n}(t)| \quad \text{or} \quad \bar{C}_n = \int (\hat{F}_n(t) - F_{\hat{\theta}_n}(t))^2 dF_{\hat{\theta}_n}(t) \ ,$$

and similarly for the AD statistic.

Remark 1. *Plugging in a parameter estimate affects the distribution of these statistics, and the distribution under the null will be heavily influenced by the type of estimator you use. Therefore one can no longer use the asymptotic distributions given in the previous session. Practitioners often mistakenly plug in $\hat{\theta}_n$ and subsequently use an ordinary KS-test or CvM-test. This will result in inadequate testing procedures.* ***This is a common mistake, and it is a very serious one.***

2.1 Location-Scale Families

A family of distribution functions $F_{\mu,\sigma}$ is called a location-scale family if

$$F_{\mu,\sigma}(x) = F_{0,1}\left(\frac{x-\mu}{\sigma}\right) ,$$

where μ is the location parameter and σ is the scale parameter. The nice thing about such families of distributions is that we can devise estimators $\hat{\mu}$ and $\hat{\sigma}$ that are respectively location-scale and scale invariant (in particular the maximum-likelihood estimators do the trick). An estimator T is location-scale invariant in distribution if for $a > 0$ and $b \in \mathbb{R}$

$$T(aX_1 + b, \ldots, aX_n + b) \stackrel{D}{=} aT(X_1, \ldots, X_n) + b .$$

An estimator T is scale invariant in distribution if

$$T(aX_1 + b, \ldots, aX_n + b) \stackrel{D}{=} aT(X_1, \ldots, X_n) .$$

If we have a location-scale invariant estimator $\hat{\mu}$ and a scale invariant estimator $\hat{\sigma}$ then it is easy to show that the distributions of $(\hat{\mu}-\mu)/\sigma$ and $\hat{\sigma}/\sigma$ are not functions of μ and σ (such quantities are known as pivotal quantities). Therefore, if we manage to write our test statistic uniquely as a function of these quantities we will conclude that the test statistic distribution under the null is completely determined by $F_{0,1}$.

Suppose we wish to test whether a random sample $X_1, \ldots, X_n$ is drawn from the location-scale family of distributions $\mathcal{F} = \{F_{\mu,\sigma},\ \mu \in \mathbb{R}, \sigma > 0\}$. As an example we will focus on the Kolmogorov-Smirnov test statistic with estimated parameters

$$\bar{D}_n = \sup_t |\hat{F}_n(t) - F_{\hat{\mu},\hat{\sigma}}(t)| .$$

As before, we can re-write this as

$$\begin{aligned} \bar{D}_n &= \sup_t |\hat{F}_n(t) - F_{\hat{\mu},\hat{\sigma}}(t)| \\ &= \sup_t \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \le t\} - F_{\hat{\mu},\hat{\sigma}}(t) \right| \\ &= \sup_t \left| \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{(Z_i \le u\} - u) \right| , \end{aligned}$$

where

$$Z_i = F_{\hat{\mu},\hat{\sigma}}(X_i) = F_{0,1}\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right) = F_{0,1}\left(\frac{\frac{X_i-\mu}{\sigma} - \frac{\hat{\mu}-\mu}{\sigma}}{\hat{\sigma}/\sigma}\right) .$$

Therefore the distribution of Z_i does not depend on the unknown values of μ and σ (but it is not the uniform distribution). The distribution of the test statistic under H_0 can rarely be computed analytically, but can be approximated (to an arbitrary accuracy) by Monte Carlo simulation as follows:

1. Generate a random sample $X_1, \ldots, X_n$ from $F_{0,1}$.
2. Compute the test statistic for this random sample. This means that we compute $\hat{\mu}(X_1, \ldots, X_n)$ and $\hat{\sigma} = \hat{\sigma}(X_1, \ldots, X_n)$, compute $Z_i = F_{0,1}((X_i - \hat{\mu})/\hat{\sigma})$ and then compute $\bar{D}_n$.
3. Repeat steps (a) and (b) a large number of times.

3 Some composite GoF tests

Some tests are specifically designed for GoF with estimated parameters.

Example 1. *The Lilliefors test (1967) is an adaptation of the KS-test for which the null hypothesis equals $H_0 : X_1, \ldots, X_n$ is a sample from a normal distribution with unknown parameters. The unknown (population) mean and variance are estimated by the sample mean and sample variance. Because the normal distributions form a location-scale family this test statistic is still distribution free - the distribution of this statistic under the null has been tabulated (by Monte-Carlo simulation). This is exactly the test and procedure described in Section 2.1 for the family of normal distributions.*

Example 2. *The Jarque-Bera test (1980) is a test for normality that is especially popular in the econometrics literature. This test is based on the sample kurtosis and skewness.*

$$\begin{aligned} \textit{skewness} \qquad & b_1 = \frac{\frac{1}{n}\sum_1^n (X_i - \bar{X}_n)^3}{s^3} \ , \\ \textit{kurtosis} \qquad & b_2 = \frac{\frac{1}{n}\sum_1^n (X_i - \bar{X}_n)^4}{s^4} \ . \end{aligned}$$

Under normality $\sqrt{n}b_1 \xrightarrow{D} N(0,6)$ and $\sqrt{n}(b_2 - 3) \xrightarrow{D} N(0,24)$. The Jarque-Bera statistic is defined by

$$JB = n(b_1^2/6 + (b_2 - 3)^2/24) \ .$$

Its limiting distribution (as n grows) is Chi-squared with 2 degrees of freedom.

Example 3. *The Shapiro-Wilk test is another powerful test for normality. The test statistic is*

$$W = \frac{\left(\sum_{i=1}^n a_i X_{(i)}\right)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2} \quad (\in (0,1]) \ ,$$

where the weights $a_1, \ldots, a_n$ are specified by an adequate formula. Under H_0, the numerator is an estimator for $(n-1)\sigma^2$, whereas the denominator is also an estimator for $(n-1)\sigma^2$. Hence, under H_0, $W \approx 1$. Under H_1, the numerator is tends to be smaller. Therefore, we reject the null hypothesis for small values of W.

Example 4. A simulation study to assess the performance of tests for normality. *We compute the fraction of times that the null hypothesis of normality is rejected for a number of distributions (in total we simulated 1000 times).*

```
Results for n=  20
         norm cauchy   exp    t5   t10   t15
Shapiro 0.054  0.852 0.852 0.182 0.095 0.080
KS      0.038  0.206 1.000 0.067 0.050 0.046
AD      0.043  0.863 0.799 0.166 0.092 0.074
CvM     0.050  0.864 0.751 0.157 0.081 0.070
JB      0.025  0.807 0.516 0.162 0.067 0.060
```

```
Results for n=  50
         norm cauchy   exp    t5   t10   t15
Shapiro 0.065  0.994 1.000 0.360 0.152 0.100
KS      0.062  0.472 1.000 0.066 0.045 0.054
AD      0.055  0.994 0.998 0.289 0.123 0.073
CvM     0.055  0.738 0.989 0.249 0.113 0.070
JB      0.043  0.993 0.953 0.396 0.172 0.106

Results for n=  200
         norm cauchy   exp    t5   t10   t15
Shapiro 0.054  1.000 1.000 0.825 0.362 0.223
KS      0.044  0.999 1.000 0.084 0.058 0.047
AD      0.052     NA    NA    NA 0.258 0.136
CvM     0.044  0.003 0.981 0.689 0.213 0.107
JB      0.049  1.000 1.000 0.869 0.436 0.291

Results for n=  5000
         norm cauchy exp    t5   t10   t15
Shapiro 0.056      1   1 1.000 1.000 0.997
KS      0.047      1   1 1.000 0.693 0.205
AD      0.058     NA  NA    NA    NA 0.989
CvM     0.061      0   0 0.067 1.000 0.962
JB      0.057      1   1 1.000 1.000 1.000
```

The R-code for obtaining this results is in the file `compare_goftests_normality.r`. *The AD test implementation appears to have some problems for large sample sizes. Although many textbooks only treat the KS-test/Lilliefors test, from this simulation study it appears that this is a rather poor testing procedure in practice. Indeed these tests have more theoretical interest than practical relevance. The JB and Shapiro-Wilk seem to work significantly better when testing normality. In [D'Agostino and Stephens (1986)] the authors warn that*

> *... for testing for normality, the Kolmogorov-Smirnov test is only a historical curiosity. It should never be used. It has poor power in comparison to specialized tests such as Shapiro-Wilk, D'Agostino-Pearson, Bowman-Shenton, and Anderson-Darling tests.*

As can be seen from this quote, there are many more specialized GoF-tests.

3.1 A Bootstrap approach for composite GoF tests

When tables and asymptotic distributional results are not available one has to resort to simulation-based techniques to approach the testing problem. The following method, studied by [Stute et al. (1993)], is an application of the (parametric) bootstrap and under some weak conditions it results in accurate asymptotic approximations (in n) of the p-values. It is presented here without any proof or performance guarantees.

(Input) $X_1, \ldots, X_n$ i.i.d. samples from some unknown distribution.

(i) Estimate $\hat{\theta}_n$ from $X_1, \ldots, X_n$ and construct the CDF $F_{\hat{\theta}_n}$.

(ii) Evaluate $\bar{D}_n$, $\bar{C}_n$ or $\bar{A}_n$. For the purpose of illustration let's use the KS statistic in what follows,

$$\bar{D}_n = \sup_t |\hat{F}_n(t) - F_{\hat{\theta}_n}(t)| \ .$$

(iii) Generate $B \in \mathbb{N}$ bootstrap samples of size n from $F_{\hat{\theta}_n}$. Denote these B samples by $\{X^*_{1,j}, \ldots, X^*_{n,j}\}$, $j \in \{1, \ldots, B\}$. The number of bootstrap samples B should be large to ensure a good approximation.

(iv) Compute $\bar{D}^*_j$, $\bar{C}^*_j$ or $\bar{A}^*_j$ as follows (example for the KS statistic)

$$\bar{D}^*_j = \sup_t |\hat{F}^*_j(t) - F_{\hat{\theta}^*_n}(t)| \ ,$$

where $\hat{\theta}^*_n$ is the estimate of θ obtained using $\{X^*_{1,j}, \ldots, X^*_{n,j}\}$ and $\hat{F}^*_j$ is the empirical CDF of $\{X^*_{1,j}, \ldots, X^*_{n,j}\}$.

Now, if a test with significance α is desired reject H_0 if

$$\bar{D}_n > \bar{D}^*_{(B(1-\alpha)+1)} \ ,$$

where $\bar{D}^*_{(B(1-\alpha)+1)}$ denoted the $(B(1-\alpha)+1)$ order statistic of $\{\bar{D}^*_1, \ldots, \bar{D}^*_B\}$. This is just an estimate of the $t : P_{H_0}(\bar{D}_n > t) = \alpha$.

Alternatively you can compute an approximate p-value using

$$p \approx \frac{\#\{i \ : \ \bar{D}^*_j \geq \bar{D}_n\}}{B} \ ,$$

and reject H_0 if $p < \alpha$.

4 Chi-Square-type GoF tests

This is a simple approach to GoF for both discrete and continuous random variables. It has several advantages, namely

- Suitable for both continuous and discrete settings
- Easy to use, even in high dimensions (so far we have been discussing only the one-dimensional setting).

However, there is a drawback: for continuous random variables the procedure requires some arbitrary choices (that must be done before seeing any data). As a consequence some information is lost, and these tests no longer have the property of being consistent against any alternative.

First consider the simple GoF-problem. Suppose $S = \text{supp}(F_0)$. Fix k and let

$$S = \bigcup_{i=1}^{k} A_i \ ,$$

be a partition of S (meaning the sets A_i are disjoint). For discrete random variables there is a natural choice for such a partition. For continuous random variables a partition can be obtained

by forming appropriate cells using some knowledge about F_0. Usually one chooses the number of cells to satisfy $5 \leq k \leq 15$. It is often hard to fully justify a certainly chosen partition or value for k.

Define F_i to be the observed frequency in cell A_i

$$F_i = \#\{j : X_j \in A_i\} \ .$$

Under H_0, $e_i := \mathbb{E}_0 F_i = n\mathbb{P}_0(X \in A_i)$ (the subscript emphasizes that the expectation and probability have to be computed under the null hypothesis). Under H_0 we expect e_i and F_i to be close. Any discrepancy measure between these two quantities can serve as a basis for a test statistic. In particular we defined the chi-square statistic as

$$Q = \sum_{i=1}^{k} \frac{(F_i - e_i)^2}{e_i} \ .$$

It is not hard to show that Q converges to a χ^2-distribution with $k-1$ degrees of freedom. As a rule of thumb, this approximation is reasonable if all the expected cell frequencies e_i are at least 5.

4.1 A Generalized Likelihood-Ratio Test

Instead of a Chi-square test we can perform a generalized likelihood-ratio test to address this same testing problem: given the partition of the data in classes, the vector $(F_1, \ldots, F_k)$ has a multinomial distribution with parameters $\theta = (\theta_1, \ldots, \theta_k)$. Under H_0, $\theta_i = e_i/n$. The likelihood $L(\theta_1, \ldots, \theta_k)$ is therefore proportional to

$$\prod_{i=1}^{k} \theta_i^{F_i}, \qquad F_i \in \{0, 1, \ldots, n\}, \qquad \sum_{i=1}^{k} F_i = n, \qquad \sum_{i=1}^{k} \theta_i = 1 \ .$$

The Generalized Likelihood-Ratio (GLR) statistic is thus given by

$$\mathrm{GLR} = \frac{L(e_1, \ldots, e_k)}{\sup_\theta L(\theta)} = \frac{\prod_{i=1}^{k} (e_i/n)^{F_i}}{\prod_{i=1}^{k} (F_i/n)^{F_i}} = \prod_{i=1}^{k} \left(\frac{e_i}{F_i}\right)^{F_i} \ ,$$

since the maximum likelihood estimator for θ_i is given by F_i/n. We reject the null hypothesis if GLR is small. It is not hard to show that, asymptotically, the Chi-square and GLR-test are equivalent. This can be done by comparing $-2\log \mathrm{GLR}$ with Q as follows

$$\begin{aligned} -2\log \mathrm{GLR} &= -2\sum_{i=1}^{k} F_i \log\left(\frac{e_i}{F_i}\right) \\ &= -2\sum_{i=1}^{k} F_i \log\left(1 + \frac{e_i - F_i}{F_i}\right) \ . \end{aligned}$$

Now clearly $\frac{e_i - F_i}{F_i} = \frac{e_i/n - F_i/n}{F_i/n}$ converges in probability to zero as $n \to \infty$, so it makes sense to use a Taylor expansion of $\log(1+x)$ around 0. Namely we have $\log(1+x) = x - x^2/2 + o(x^2)$.

Therefore

$$\begin{aligned} -2\log \text{GLR} &= -2\sum_{i=1}^{k} F_i \log\left(1+\frac{e_i-F_i}{F_i}\right) \\ &= -2\sum_{i=1}^{k} F_i \left(\frac{e_i-F_i}{F_i} - \frac{1}{2}\left(\frac{e_i-F_i}{F_i}\right)^2 + o_P\left(\left(\frac{e_i-F_i}{F_i}\right)^2\right)\right) \\ &= \sum_{i=1}^{k}\left(\frac{(e_i-F_i)^2}{F_i} + o_P\left(\frac{(e_i-F_i)^2}{F_i}\right)\right) , \end{aligned}$$

where the final step follows from the simple facts $\sum_{i=1}^{k} F_i = \sum_{i=1}^{k} e_i = n$. We are almost done. Since

$$\frac{(e_i-F_i)^2}{F_i} = \frac{(e_i-F_i)^2}{e_i}\frac{e_i/n}{F_i/n} ,$$

and $\frac{e_i/n}{F_i/n}$ converges to 1 in probability Slutsky's theorem tells us that $\frac{(e_i-F_i)^2}{F_i}$ and $\frac{(e_i-F_i)^2}{e_i}$ converge in distribution to the same limit. Therefore we conclude that $-2\log \text{GLR} \xrightarrow{D} \chi^2_{k-1}$ as $n \to \infty$, as desired.

4.2 Composite χ^2-GoF tests

In the composite GoF case things get more complicated as you might expect. In particular the expected cell frequencies have to be estimated from the data. By plugging-in these estimators the distribution of Q under the null is going to change. The properties of that distribution depend on the properties of the estimator. If the estimators are chosen using a maximum likelihood principle then, under some mild assumptions (the same needed for Wilk's theorem regarding generalized likelihood ratio tests) the resulting limiting distribution will be chi-square with $k-1-s$ degrees of freedom, where s is the number of independent parameters that must be estimated for calculating the estimated expected cell frequencies. So in case we test for normality, $s=2$ (mean and variance).

5 Probability Plotting and Quantile-Quantile Plots

Probability plots provide a visual way to do empirically do GoF tests. These are not formal tests, but provide a quick tool to check if a certain distributional assumption is somewhat reasonable.

Let $\mathcal{F}$ be a location-scale distribution family, that is

$$\mathcal{F} = \{F_{a,b} : a \in \mathbb{R}, b > 0\} ,$$

for some distribution F. In the above $F_{a,b}$ is the CDF of $a+bX$ when $X \sim F_{0,1}$, that is

$$F_{a,b}(x) = F_{0,1}\left(\frac{x-a}{b}\right) .$$

As an example, a $\mathcal{N}(\mu,\sigma^2)$ random variable is obtained from a standard normal random variable Z by the linear transformation $Z \mapsto \mu + \sigma Z$.

Let $X_1, \ldots, X_n$ be data from some distribution. Recall that $\hat{F}_n(X_{(i)}) = i/n$. Therefore

$$F_{a,b}^{-1}(\hat{F}_n(X_{(i)})) = F_{a,b}^{-1}\left(\frac{i}{n}\right) \ .$$

Now, if the data comes from a distribution in $\mathcal{F}$ then $\hat{F}_n(X_{(i)}) \approx F_{a,b}(X_{(i)})$, and so

$$X_{(i)} \approx F_{a,b}^{-1}\left(\frac{i}{n}\right) = a + bF_{0,1}^{-1}\left(\frac{i}{n}\right) \ .$$

Said differently, if the data comes from a distribution in $\mathcal{F}$ we expect the points $\left(X_{(i)}, F_{0,1}^{-1}\left(\frac{i}{n+1}\right)\right)$ to lie approximately in a straight line. Note that we replaced i/n by $i/(n+1)$: this is to ensure that we stay away from evaluating $F_{0,1}^{-1}(1)$, which can be infinite. The plot of the above points is commonly called the quantile-quantile plot.

The use of probability plots requires some training, but these are very commonly used and helpful. If we want to test for a distribution F_θ that is not in a location-scale family, then the preceding reasoning implies that the points $\left(F_\theta^{-1}\left(\frac{i}{n+1}\right), x_{(i)}\right)$ should be on a straight line if θ is known. If θ is unknown, an estimator for it can be plugged in. Most software packages can generate such plots automatically. As pointed out in [Venables and Ripley (1997)] (page 165), a QQ-plot for e.g. a t_9-distribution can be generated by executing the following code in r:

```
plot(qt(ppoints(x),9), sort(x))
```

(We assume the data are stored in a vector `x`). Adding the command `qqline(x)`, produces a straight line through the lower and upper quartiles. This helps assess whether the points are (approximately) on a straight line. You may also want to consider the function `qq.plot` in the library `car`, which gives a direct implementation of the QQ-plot.

Example 5. *Comparing QQ-normality plots and AD-test for normality. We simulate samples of size* $10, 50, 100, 1000, 5000, 10000$ *from a standard normal and* t_{15} *distribution. Figures 5 and 2 show QQ-normality plots for both cases respectively. Reading these figure from upper-left to lower-right, the corresponding AD p-values are:*

normal	t_{15}
0.671	*0.021*
0.278	*0.815*
0.493	*0.381*
0.925	*0.001*
0.186	*0.001*
0.339	*9.98e-08*

For almost every purpose in practice, the difference between a t_{15} *and a Normal distribution is of no importance. However, as we obtain sufficiently many data, hypothesis tests will always detect any fixed deviation from the null hypothesis, as can be seen very clearly from the computed p-values of the AD test. The* `R`*-code for producing these figures is in the file* `compare_qq_testing.r`*.*

For large datasets there are difficulties with the interpretation of QQ-plots, as indicated by the following theorem.

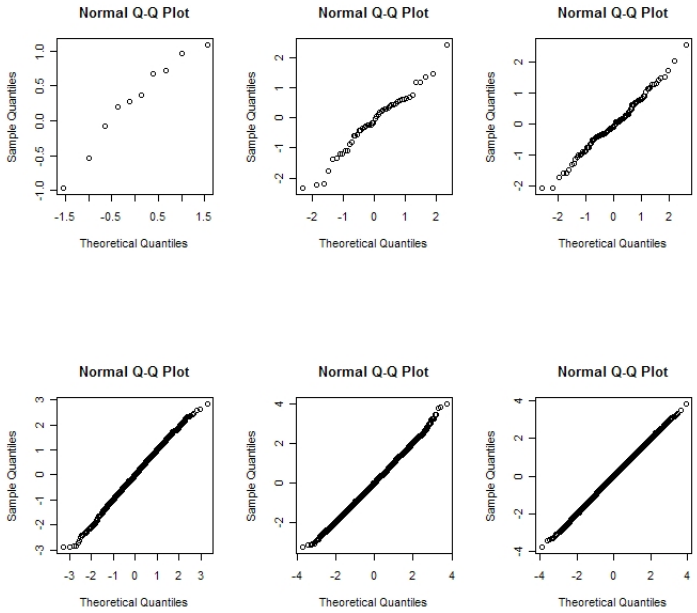

Figure 1: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a standard normal distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

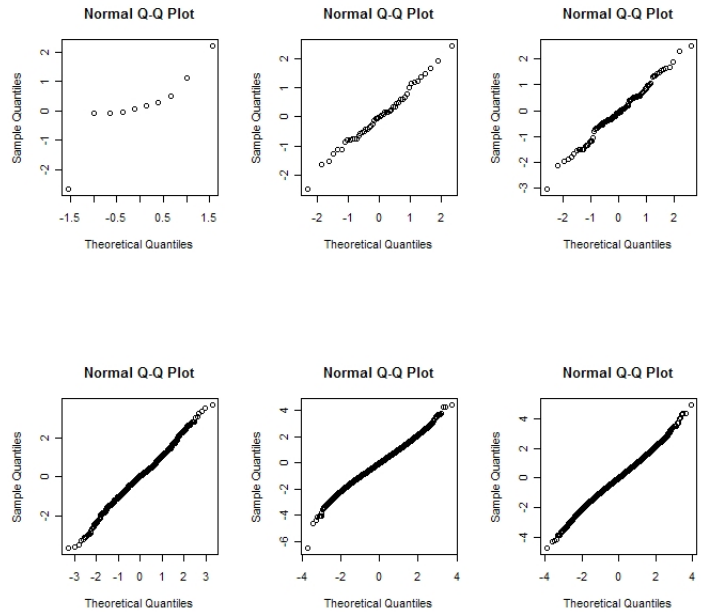

Figure 2: QQ-plots for samples of sizes $10, 50, 100, 1000, 5000, 10000$ from a t_{15} distribution. The upper-left figure is for sample size 10, the lower-right is for sample 10000.

F	ρ_F
normal	1
uniform	0.98
double exp	0.98
t_3	0.90
t_5	0.98
χ^2_3	0.96
exponential	0.90
logistic	0.97

Table 1: Limiting values for r_n

Theorem 1. *Let r_n be the correlation coefficient of the pairs $\left(X_{(i)}, \Phi^{-1}\left(\frac{i}{n+1}\right)\right)$, where Φ is the distribution function of the standard normal distribution. Let F be the true CDF with variance σ^2, then*

$$\lim_{n\to\infty} r_n = \frac{1}{\sigma}\int_0^1 F^{-1}(x)\Phi^{-1}(x)dx =: \rho_F \quad a.s\ .$$

See theorem 26.10 in [DasGupta (2008)]. Table 1 provides values for the limiting value for various distribution. We conclude that asymptotically we get a perfect straight line in case of normality (as it should be). However, for many other distributions we obtain a correlation coefficient that is very close to one. It is hard to distinguish a set of points with correlation coefficient 0.97 from a set with correlation coefficient equal to 1. The difference is mainly in the tails. For small sample sizes, probability plots help to assess whether normality holds *approximately*, which is often all we need (for example in assessing approximate normality of residuals of linear models).

6 Exercises

Exercise 1. *The Cramer-Von Mises test statistic is defined by*

$$C_n := \int (\hat{F}_n(t) - F_0(t))^2 dF_0(t)\ .$$

Verify the following computational formula for C_n:

$$nC_n = \frac{1}{12n} + \sum_{i=1}^{n}\left(U_{(i)} - \frac{2i-1}{2n}\right)^2\ ,$$

*with $U_{(i)} = F_0(X_{(i)})$. To make your task easier you may make the extra assumption that F_0 is strictly increasing. (**Note:** The following fact might come in handy: $\sum_{i=1}^{n}(2i-1)^2 = \frac{4n^3}{3} - \frac{n}{3}$).*

Exercise 2. *Suppose we wish to test whether data are from a Normal distribution. More precisely, assume $x_1, \ldots, x_n$ is are realizations of independent random variables $X_1, \ldots, X_n$ with a $N(\mu, \sigma^2)$-distribution. The parameters μ and σ^2 are unknown.*

The Cramer-Von Mises test statistic in this case is defined by

$$\tilde{C}_n = \int \left[F_n(x) - \Phi\left(\frac{x-\bar{X}_n}{S_n}\right)\right]^2 d\Phi\left(\frac{x-\bar{X}_n}{S_n}\right)\ .$$

Here $\bar{X}_n$ and S_n are respectively the sample mean and standard deviation of $X_1, \ldots, X_n$ respectively.

a) Show that

$$\tilde{C}_n = \int_0^1 \left[\frac{1}{n} \sum \mathbf{1}\{Y_i \leq y\} - y \right]^2 dy \ ,$$

where $Y_i = \Phi((X_i - \bar{X}_n)/S_n)$.

b) Is this test distribution free?

c) Explain why the following piece of code outputs $n\tilde{C}_n$

```
cvm.normality <- function(x)
{   n<-length(x)
    un<-pnorm(sort(x),mean=mean(x),sd=sd(x))
    dn<-(2*(1:n)-1)/(2*n)
    1/(12*n)+sum((dn-un)^2)
}
```

Exercise 3. *In the lecture slides a simulation study was performed to assess the power of various tests for normality, including the Crámer - Von Mises test. According to the results of this simulation study, for sample size $n = 5000$ the implementation of this test within the* `nortest` *library fails to reject data from a standard exponential distribution. This is rather strange, since with such a large sample size, discerning between a normal and exponential distribution should be very easy. It turns out there is an error in the implementation. Use the code below to assess the power of the Crámer - Von Mises test for normality against the alternative of an exponential distribution.*

```
cvm.test.asymp<-function(x,alpha)
# This function computes the Cramer - Von Mises statistic
# H_0 : data are from a normal distribution with unknown parameters.
# Test statistic equals
# Y = n C_n = n \int (F_n - Phi(x-mean(x)/sd(x)))^2 d Phi(x-mean(x)/sd(x)),
# where F_n is the ECDF and Phi the CDF of
# a standard normal random variable.
# Critical values are taken from Stephens (Ann. Stat. 4, 1976), table 4.
# These are based on asymptotics (n to infinity).
# Alpha should be below 0.15.
# If alpha is not equal to 0.010, 0.025, 0.050, 0.100 or 0.15,
# we use linear interpolation of critical values
# Function returns value of test statistic and
# a 1 if normality is rejected, 0 else
{
    T <- cvm.normality(x)
    a <- c(0.010, 0.025, 0.050, 0.100, 0.15)
    q <- c(0.178, 0.148, 0.126, 0.104, 0.091)
    CV <- approx(a, q, alpha, 'linear')
```

```
    rej <-(T>=CV$y)
    c(T,rej)
}
```

Exercise 4. *A quality control engineer has taken 50 samples of size 13 from a production process. The number of defectives for these samples are recorded below.*

Number of defects	*Number of samples*
0	*11*
1	*13*
2	*14*
3	*10*
4	*1*
5	*1*
6 or more	*0*

a) Suppose we want to check if a Poisson distribution is a good model for the number of defects in each sample. If X_i denotes the number of defects in the i-th sample, then we henceforth assume that $X_1, \ldots, X_n$ is a random sample from a Poisson distribution, say with parameter μ. Show that the maximum likelihood estimate for the μ is given by $\hat{\mu} = 1.6$.

b) Under the assumption of a Poisson distribution, the expected number of samples that have k defects ($k = 0, 1, \ldots$) can be estimated by $\hat{E}_k = 50e^{-\hat{\mu}}\hat{\mu}^k/(k!)$. Compute the value of the test statistic

$$Q = \sum_{i \in \mathcal{I}} \frac{(F_i - \hat{E}_i)^2}{\hat{E}_i} \ ,$$

Here $\mathcal{I}$ denotes the 7 classes of number of defects, i.e. $\{0, 1, 2, \ldots, 5, 6$ or more$\}$ defects. F_i denotes the observed count in class i.

c) Asymptotically, under the null hypothesis, Q has a χ^2-distribution with $7-1-1 = 5$ degrees of freedom (we lose one degree of freedom from the estimation of μ). Check whether the Chi-square test rejects the null hypothesis that the data are a random sample from a Poisson distribution. Use significance level $\alpha = 0.05$.

Exercise 5. *Consider 31 measurements of polished window strength data for a glass airplane window. In reliability tests, researchers often rely on parametric assumptions to characterize observed lifetimes. In this exercise we will see if a normal or Weibull distribution is appropriate. The data can be found on the website*
`http://atomic.phys.uni-sofia.bg/local/nist-e-handbook/e-handbook/eda/section4/eda4291.htm`

a) Check if the data can be assumed normally distributed. Use both the Anderson-Darling statistic and a normal probability plot.

b) The Weibull distribution function is given by

$$F_{\beta,\gamma}(x) = \begin{cases} 0 & x < 0 \\ 1 - \exp\left(-\left[\frac{x}{\gamma}\right]^{\beta}\right) & x \geq 0 \end{cases} .$$

Here $\gamma > 0$ is the scale parameter and $\beta > 0$ is the shape parameter. Show that for $x \geq 0$

$$\ln(x) = \frac{1}{\beta}\ln(-\ln(1 - F_{\beta,\gamma}(x))) + \ln\gamma.$$

c) *Note the above expression gives a linear relation between $\ln(x)$ and $\ln(-\ln(1 - F_{\beta,\gamma}(x)))$. Use this fact to devise a probability plot to check if it is sensible to assume data is a sample from a Weibull distribution. In particular take $x = X_{(i)}$ in that expression and use $F_{\beta,\gamma}(X_{(i)}) \approx \hat{F}_n(X_{(i)}) \approx \frac{i}{n+1}$. Make this plot for the glass airplane window data.*

d) *Write a function in* `R` *that computes*

$$C_n \equiv C_n(X_1, \ldots, X_n) = \int (\hat{F}_n(t) - F_{\hat{\beta},\hat{\gamma}}(t))^2 dF_{\hat{\beta},\hat{\gamma}}(t)$$

for data $X_1, \ldots, X_n$. Here $\hat{F}_n$ denotes the empirical distribution function of $X_1, \ldots, X_n$ and $(\hat{\beta}, \hat{\gamma})$ is the maximum likelihood estimator for (β, γ). Note: If the data are in a vector `x`*, then the following R-code computes the MLE for the Weibull distribution:*

```
library(MASS)
th.hat<-fitdistr(x, 'weibull')
beta.hat <- as.numeric(th.hat$estimate[1])
gamma.hat <-as.numeric(th.hat$estimate[2])
```

e) *Test whether the data follow a Weibull distribution, using the test statistic C_n. Implement the parametric bootstrap to compute a p-value.*

(Adaptation of exercise 6.2 in [Kvam and Vidakovic (2007)]).

Exercise 6. *Suppose we have data $x_1, \ldots, x_n$ that are assumed to have been sampled from a distribution F. We wish to test $H_0 : F = F_0$. The Kolmogorov-Smirnov test was shown to be consistent under any alternative. In this exercise we will show that this is not true for a χ^2-type test. Let F_1 be a distribution function such that $F_0 \neq F_1$. Fix $k \in \mathcal{N}$ and suppose $A_i, i = 1, \ldots, k$ is a partition of supp(F_0). Let G_i denote the number of observations with values in A_i. Define*

$$p_{0i} = P_{F_0}(X \in A_i) \qquad p_{1i} = P_{F_1}(X \in A_i)$$

and

$$e_{0i} = np_{0i} \qquad e_{1i} = np_{1i}.$$

The χ^2-test statistic is given by

$$Q = \sum_{i=1}^{k} \frac{(G_i - e_{0i})^2}{e_{0i}}.$$

a) *Show that under F_1,*

$$Q/n \xrightarrow{P} \sum_{i=1}^{k} \frac{(p_{1i} - p_{0i})^2}{p_{0i}} =: c.$$

b) *Argue that if $c > 0$ the test is consistent against F_1.*

c) *What if $c = 0$?*

7 Useful R commands

Goodness of fit

Test	R-function	within package
Chi-square GOF	chisq.test	stats
KS GOF	ks.test	stats
KS-distribution	ksdist	PASWR
QQ-plot	qq.plot	car

Specialized test for normality.

Test	R-function	within package
Shapiro-Wilk	shapiro.test	stats
Anderson Darling	ad.test	nortest
Cramer-Von Mises (for composite normality)	cvm.test	nortest
Jarque-Bera (for composite normality)	jarque.bera.test	tseries

References

[D'Agostino and Stephens (1986)] D'Agostino, R.B. and Stephens, M.A. (1986) *Goodness-of-Fit Techniques*, New York: Marcel Dekker.

[Stute et al. (1993)] Stute, W. González-Manteiga, W. and Quindimil, M.P. (1993) *Bootstrap Based Goodness-Of-Fit-Tests* Metrika **40**, 243–256.

[DasGupta (2008)] DasGupta, A. (2008) *Asymptotic Theory of Statistics and Probability*, Springer. chapters 26 and 27

[Kvam and Vidakovic (2007)] Kvam, P.H. and Vidakovic, B. (2007) *Nonparametric Statistics with Applications to Science and Engineering*, Wiley.

[Venables and Ripley (1997)] Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with S-PLUS*, second edition, Springer.