Nonparametric Goodness-of-Fit Tests

We've seen in Chapter 10 how to test whether or not proportions $(p_1, \cdots, p_r) = p$ were equal to prescribed ones $(p_{1,0}, \cdots, p_{r,0}) = p_0$. Testing for $H_0 : p = p_0$ amounts to test if the unknown distribution P of data equals Multinomial (m, p_0), or equivalently $H_0 : P = \text{Multinomial}(m, p_0)$.

In other words, this is a goodness-of-fit test with

1) Discrete distributions (since we deal with count/bins here)
2) A parametric model (namely, the multinomial distributions)

In these notes, we explore the idea above for

1) Continuous distribution
2) Nonparametric models.

Notation: let $X_1, \dots, X_n \overset{iid}{\sim} P$ be real random variables.

Since we work in a nonparametric setting, we don't have parameter/formulas to describe distributions. Hence, we must work in full generality with its cumulative distribution function

$$F_X(t) = F(t) = \mathbb{P}(X \leq t) \quad , \quad t \in \mathbb{R}.$$

Recall that $F_X = F_Y \iff X$ and Y have the same distribution, so cdf's are good candidates for describing / testing equality of distributions.

We have $F(t) = \mathbb{P}(X \leq t) = \mathbb{E}\left[\mathbb{1}_{\{X \leq t\}}\right]$, An estimator of F is the empirical distribution function

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq t\}}.$$

Plug in / Method of moments

Indeed, we see that for any fixed $t \in \mathbb{R}$, the strong law of large numbers yields

$$F_n(t) \xrightarrow[n \to \infty]{a.s.} F(t).$$

Actually, the convergence is much stronger: it's uniform.

(3)

Thm (Glivenko - Cantelli)

$$\sup_{t \in \mathbb{R}} |F_n(t) - F(t)| \xrightarrow[n \to \infty]{a.s.} 0$$

Proof: let $\varepsilon > 0$. Then, fix $k > \frac{1}{\varepsilon}$ and consider "knot" points $x_0, \dots, x_k$ such that

$$-\infty = x_0 < x_1 \leq x_2 \leq \dots \leq x_{k-1} < x_k = +\infty$$

that define a partition of $\mathbb{R}$ into k disjoint intervals such that

$$F(x_j^-) \leq \frac{j}{k} \leq F(x_j) \quad , \quad 1 \leq j \leq k-1$$

where $F(t^-) = \mathbb{P}(X < t) = F(t) - \mathbb{P}(X = t)$

$$= \lim_{\substack{s \to t^- \\ s < t}} F(s) \, ,$$

Then, by construction, if $x_{j-1} < x_j$,

$$F(x_j^-) - F(x_{j-1}) \leq \frac{j}{k} - \frac{(j-1)}{k} = \frac{1}{k} < \varepsilon \, .$$

(4)

By by the strong LLN, we have the pointwise convergences

$$F_n(x_j) \xrightarrow[n\to\infty]{a.s.} F(x_j) \quad \text{and} \quad F_n(x_j^-) \xrightarrow[n\to\infty]{a.s.} F(x_j^-),$$

which rewrites as

$$|F_n(x_j) - F(x_j)| \xrightarrow[n\to\infty]{a.s.} 0 \quad \text{and} \quad |F_n(x_j^-) - F(x_j)| \xrightarrow[n\to\infty]{a.s.} 0.$$

Hence, looking at the maximum over all j in the finite set $\{1,\dots,k-1\}$,

$$\Delta_n = \max_{1\le j\le k-1} \left\{ |F_n(x_j) - F(x_j)|, |F_n(x_j^-) - F(x_j^-)| \right\} \xrightarrow[n\to\infty]{} 0.$$

Now, notice that any $t\in\mathbb{R}$ falls into some interval $[x_{j-1}, x_j)$.

$$x_{j-1} \le t < x_j.$$

We then have:

$$F_n(t) - F(t) \le F_n(x_j^-) - F(x_{j-1}) \le F_n(x_j^-) - F(x_j^-) + \varepsilon$$

and

$$F_n(t) - F(t) \ge F_n(x_{j-1}) - F(x_j^-) \ge F_n(x_{j-1}) - F(x_{j-1}) - \varepsilon,$$

and thus for any $t \in \mathbb{R}$,

$$|F_n(t) - F(t)| \leq \Delta_n + \varepsilon \xrightarrow[n\to\infty]{a.s} \varepsilon$$

As this holds for arbitrary $t \in \mathbb{R}$, it follows that

$$\forall \varepsilon > 0, \quad \limsup_{n\to\infty} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \varepsilon \quad a.s$$

All that remains to be done is to send $\varepsilon \to 0$. For this, write the event $A_\varepsilon = \left\{ \omega \mid \limsup_{n\to\infty} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| \leq \varepsilon \right\}$.

Clearly, $\varepsilon' < \varepsilon \Rightarrow A_{\varepsilon'} \subset A_\varepsilon$, and $A_0 = \bigcap_{\varepsilon>0} A_\varepsilon$, so that

$$P(A_0) = P\left(\bigcap_{\varepsilon>0} A_\varepsilon\right) = \inf_{\varepsilon>0} P(A_\varepsilon)$$

But from what's above, $P(A_\varepsilon) = 1$ for all $\varepsilon > 0$. Hence, $P(A_0) = 1$, or equivalently

$$\boxed{P\left(\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |F_n(t) - F(t)| = 0\right) = 1}$$

Rk: Although Glivenko - Cantelli is a very strong result, it does not help us derive a testing procedure

Kolmogorov - Smirnov Test for Goodness-of-Fit

We introduce a (pseudo)-distance between the empirical distribution $P_n = \sum_{i=1}^{n} \delta_{X_i}$ and P, that simply is the L^∞ distance between their respective cdf's:

$$D_{KS}(P_n, P) = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

If F is given, this is a statistic that can actually be computed straightforwardly.

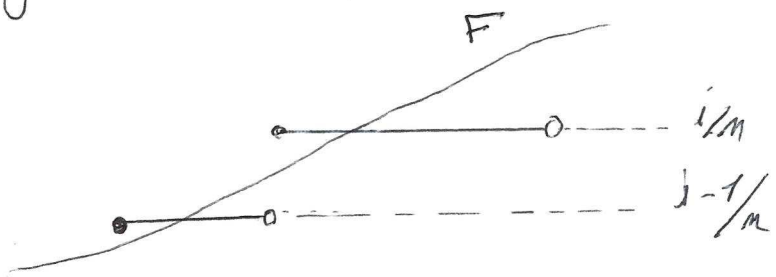
To see this, write $(X_{(1)}, \dots, X_{(n)})$ for the order statistic of $(X_1, \dots X_n)$. Namely it's just the sample points sorted by increasing order

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

(7)

Prop: $$D_{KS}(P_n, P) = \max_{1 \le i \le n} \max \left\{ \left| F(X_{(i)}) - \frac{i}{n} \right|, \left| F(X_{(i)}) - \frac{i-1}{n} \right| \right\}$$

Proof: F is increasing, and F_n is piecewise constant and taking values $\left\{ \frac{i}{n} \right\}_{0 \le i \le 1}$



We now describe the distribution of $D_{KS}(P_n, P)$.

Thm: Write $D_n = \sqrt{n}\, D_{KS}(P_n, P) = \sup_{t \in \mathbb{R}} \sqrt{n} \left| F_n(t) - F(t) \right|$.

Assume that P has a density

(i) If $U_n = \sup_{t \in [0,1]} \sqrt{n} \left| F_{U,n}(t) - t \right|$, where $F_{U,n}$ is the empirical distribution function associated to a i.i.d n-sample of the uniform distribution on $[0,1]$, then D_n has the same distribution as U_n. (Such tables are known)

(ii) (Kolmogorov, 1933) D_n converges in distribution towards the distribution with cdf
$$H(t) = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 t^2}$$

(iii) If P_0 is a distribution with density, and that $P \neq P_0$, then
$$D_n^0 = \sqrt{n}\, D_{KS}(P_n, P_0) \xrightarrow[n \to \infty]{\mathbb{P}} \infty$$

(8)

To prove the theorem, we'll need some results on the <u>quantile function</u>

$$F^{-1}(u) = \inf\{t \mid F(t) \geq u\}$$

Also called generalized inverse.

Lemma F and F^{-1} satisfy

1) $\forall t \in \mathbb{R},\ F^{-1}(F(t)) \leq t$
2) $\forall u \in (0,1),\ F(F^{-1}(u)) \geq u$
3) $F^{-1}(u) \leq t \iff u \leq F(t)$
4) If $U \sim \text{Uniform}([0,1])$, $F^{-1}(U)$ has the same distribution as X.

<u>Proof</u>: Let us write $\mathcal{T}_u = \{t \mid F(t) \geq u\}$, so that $F^{-1}(u) = \inf \mathcal{T}_u$.

1) $\forall t \in \mathbb{R},\ t \in \mathcal{T}_{F(t)}$, so $F^{-1}(F(t)) \leq t$.

2) By definition, there exists a decreasing sequence $(t_n)_n$ of elements of $\mathcal{T}_u$ such that $t_n \xrightarrow[n\to\infty]{} F^{-1}(u)$. But since $F(t_n) \geq u$ and that F is right continuous, $F(F^{-1}(u)) \geq u$.

3) . If $u \leq F(t)$, then $t \in \mathcal{T}_u$ so that $F^{-1}(u) \leq t$

Conversely

. If $F^{-1}(u) \leq t$, then since F is non-decreasing, $F(F^{-1}(u)) \leq F(t)$. Using 2), we get $F(t) \geq u$.

4) For $t \in \mathbb{R}$, from 3) above,

$$\mathbb{P}(F^{-1}(U) \leq t) = \mathbb{P}(U \leq F(t))$$
$$= F(t),$$

and since the cdf characterises fully the distribution of X, we get the result.

Proof of the theorem

(i) If $U_i \sim \text{Uniform}([0,1])$, $F^{-1}(U_i) \overset{\mathcal{D}}{=} X_i$. As a consequence,

$$F_{U,n}(F(t)) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{U_i \leq F(t)}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{F^{-1}(U_i) \leq t}$$

$$\overset{\mathcal{D}}{=} F_n(t).$$

Taking suprema over $t \in \mathbb{R}$ we get that $D_n \overset{\mathcal{D}}{=} \sqrt{n} \sup_{t \in \mathbb{R}} \left| F_{U,n}(F(t)) - F(t) \right|$.

Since F is continuous (because P has a density), $\{F(t)\}_{t \in \mathbb{R}} = (0,1)$, so that we get (by "change of variable $u = F(t)$"), that $U_n \overset{\mathcal{D}}{=} D_n$.

(ii) Not proved here. See Billingsley for instance.

(iii) If $P_0 \neq P$, there exists $\varepsilon > 0$ and $t \in \mathbb{R}$ such that $|F(t) - F_0(t)| \geq \varepsilon$.

Without loss of generality, assume that $F(t) - F_0(t) \geq \varepsilon$.

Then,

$$D_n^0 \geq \sqrt{n}\,|F_n(t) - F_0(t)|$$

Triangle inequality $\Rightarrow$

$$\geq \sqrt{n}\,|F(t) - F_0(t)| - \sqrt{n}\,|F(t) - F_n(t)|$$

$$\geq \sqrt{n}\,\varepsilon - \sqrt{n}\,|F(t) - F_n(t)|.$$

But from the LLN, $|F_n(t) - F(t)| \xrightarrow[n\to\infty]{P} 0$, so

$$P\left(|F_n(t) - F(t)| \geq \frac{\varepsilon}{2}\right) \xrightarrow[n\to\infty]{} 0,$$

which yields $P\left(D_n^0 \geq \sqrt{n}\frac{\varepsilon}{2}\right) \xrightarrow[n\to\infty]{} 1$,

and hence $D_n^0 \xrightarrow[n\to\infty]{P} \infty$ $\left(\forall M > 0,\ P(D_n^0 \geq M) \xrightarrow[n\to\infty]{} 1\right)$

Application of these results to building the Goodness-of-Fit K.S. Test

Let P_0 be a fixed (known) distribution, with cdf F_0 that is continuous.

Consider the testing problem

$$H_0 : P = P_0 \quad \text{vs} \quad H_1 : P \neq P_0$$

Test statistic: Considering the previous results, we take

$$D_n^0 = \sqrt{n}\, D_{KS}(P_n, P_0)$$

$$= \sqrt{n} \max_{1 \le i \le n} \max \left\{ \left| F_0(X_{(i)}) - \frac{i}{n} \right|, \left| F_0(X_{(i)}) - \frac{i-1}{n} \right| \right\}$$

Rejection Region: Since $D_n^0 \xrightarrow[n \to \infty]{P} \infty$ under H_1, we take

$$R = \{ D_n^0 \ge s \} \quad \text{for some } s \in (0, \infty)$$

Calibration of s:

- If n is small ($n \le 100$), we use the tables of the distribution of U_n
- If n is large ($n > 100$), we use those of the asymptotic distribution with cdf H.

Rk : . The K.S. test has both nonasymptotic and asymptotic formulations

. The K.S. test cannot be used if P_0 does not have a density ! (i.e. F_0 continuous)

. Astonishingly, we often note that the K.S. test is more powerful than the Goodness-of-fit χ^2 test, especially if n is small

. As for χ^2 test, one can adapt K.S. to test if P belongs to some parametric family $\{P_\theta, \theta \in \Theta\}$: uniform distributions, Gaussians, Exponentials ...

In these cases one needs to show that by taking an estimator $\hat{\theta}$ of θ, the distribution of $\sqrt{n}\, D_{KS}(P_n, P_{\hat{\theta}})$ does not depend on P_θ.

We 'll see later on an example of these for Gaussians (Lilliefors)

Kolmogorov - Smirnov Test for Homogeneity

We consider the two-sample problem that consists in testing if two samples have the same distribution.
Namely, we have two independent sequences

$$\begin{cases} X_1, \dots, X_n \sim_{iid} P_X \\ Y_1, \dots, Y_m \sim_{iid} P_Y \end{cases},$$

where P_X and P_Y are unknown.

We let $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i \le t}$ and $G_m(t) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}_{Y_j \le t}$ denote the empirical distribution functions of the X_i's and Y_j's respectively.
Write

$$D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$$

Thm Assume that P_X and P_Y have a density.
(i) The distribution of $D_{n,m}$ does not depend on P_X and P_Y when $P_X = P_Y$. It only depends on n and m
(ii) If $P_X \neq P_Y$, then $D_{n,m} \xrightarrow[n,m \to \infty]{P} \infty$

Application of this result to building the K.S. test for homogeneity

We want to test

$$H_0 : P_X = P_Y \quad \text{vs} \quad H_1 : P_X \neq P_Y$$

Test statistic: $D_{n,m} = \sqrt{\frac{nm}{n+m}} \sup_{t \in \mathbb{R}} |F_n(t) - G_m(t)|$

Rejection Region: $\mathcal{R} = \{ D_{n,m} \geq \Delta \}$ for some $\Delta > 0$

Calibration of Δ: Using the tables of the K-S distribution with parameters (n, m)

Rk: The idea of comparing cdf and empirical cdf is pretty general. Instead of taking the sup-norm between these two functions, one can also consider:

- L^2 distance: $n \int_{\mathbb{R}} (F_n(t) - F(t))^2 f(t)\, dt$ (Cramér-von Mises test, see Homework 6)

- Weighted L^2 distance: $n \int_{\mathbb{R}} \frac{(F_n(t) - F(t))^2}{F(t)(1-F(t))} f(t)\, dt$ (Anderson-Darling test)

⊕ two-sample versions of these.

Lilliefors Test of Normality

Back to goodness-of-fit tests, we sometimes want to test whether data comes from a family of distributions (as opposed to a single P_0 of interest). For instance, one may wonder if P is a Gaussian / Exponential / Beta / Gamma / ... distribution.

Namely, given $X_1, \dots, X_n \overset{iid}{\sim} P$ and a parametric model $\{P_\theta \mid \theta \in \Theta\}$, you'd like to test

$$H_0 : P \in \{P_\theta \mid \theta \in \Theta\} \quad \text{vs} \quad H_1 : P \notin \{P_\theta \mid \theta \in \Theta\}.$$

Among these, probably the most important family is the Gaussian model

$$\{N(\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma > 0\}.$$

The Lilliefors test of Normality roughly boils down (⚠ a distribution with random parameters do not exist!) to apply a goodness-of-fit K.S. test to a distribution $\hat{P}_0$ "is Gaussian with mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and variance $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$".

Hypotheses: H_0 : P is Gaussian $\quad H_1$: P is not Gaussian

Test Statistic: let F_0 denote the cdf of the distribution $N(0,1)$.

We let $\hat{F}_0(t) = F_0\left(\frac{t-\bar{X}_n}{S_n}\right)$, and define the test statistic by

$$L_n = \sqrt{n} \sup_{t\in\mathbb{R}} \left| F_n(t) - \hat{F}_0(t) \right|$$

$$= \sqrt{n} \max_{1\le i\le n} \max\left\{ \left| F_0\left(\frac{X_{(i)}-\bar{X}_n}{S_n}\right) - \frac{i}{n} \right| ; \left| F_0\left(\frac{X_{(i)}-\bar{X}_n}{S_n}\right) - \frac{i-1}{n} \right| \right\}$$

Rejection Region: $R = \{ L_n \ge s \}$ for some $s > 0$

Calibration of s: If H_0 holds (meaning $P = N(\mu,\sigma^2)$ for some (μ,σ)), the distribution of $\frac{X_i - \bar{X}_n}{S_n}$ does not depend on (μ,σ), so the distribution of L_n does not depend on P under H_0.

⚠ The distribution of L_n under H_0 is not the K.S. distribution! So s is to be read in another table: that of the Lilliefors distribution.