Course Summary: *Statistics for Applications*

Yan Zeng

Version 1.0.1, last revised on 2015-11-07.

Abstract

Summary of Panchenko [1].

Contents

1 Method of Moments for Parametric Estimation

Suppose the parameter set $\Theta \subseteq \mathbb{R}$ and suppose that we can find a function g such that a function

$$m(\theta) = \mathbb{E}_\theta[g(X)]$$

has a continuous inverse m^{-1}. Here $\mathbb{E}_\theta$ denotes the expectation with respect to the distribution $\mathbb{P}_\theta$. Take

$$\hat{\theta} = m^{-1}(\overline{g}) = m^{-1}\left(\frac{g(X_1) + \cdots + g(X_n)}{n}\right).$$

as the estimate of θ_0. By Law of Large Numbers and the continuity of m^{-1}, we have

$$\overline{g} \to \mathbb{E}_{\theta_0}[g(X_1)] = m(\theta_0),\ \hat{\theta} = m^{-1}(\overline{g}) \to m^{-1}(m(\theta_0)) = \theta_0.$$

Typical choices of the function g are $g(x) = x$ or x^2. The quantity $\mathbb{E}[X^k]$ is called the kth moment of X and, hence, the name - *method of moments.* Sometimes, we can use either first moment or second moment to estimate parameter. The question is, which estimate is better? This motivates the concept of *consistency* and *asymptotic normality.*

Definition 1 (Consistency)**.** *We say that an estimate $\hat{\theta}$ is* **consistent** *if $\hat{\theta} \to \theta_0$ in probability as $n \to \infty$. We have shown above that by construction the estimate by method of moments is always consistent.*

Definition 2 (Asymptotic normality)**.** *We say that $\hat{\theta}$ is* **asymptotically normal** *if*

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2_{\theta_0})$$

where $\sigma^2_{\theta_0}$ is called the asymptotic variance of the estimate $\hat{\theta}$.

Theorem 1 (Asymptotic normality of method of moments). *The estimate $\hat{\theta} = m^{-1}(\overline{g})$ by the method of moments is asymptotically normal with asymptotic variance*

$$\sigma^2_{\theta_0} = \frac{Var_{\theta_0}[g(X_1)]}{(m'(\theta_0))^2}.$$

Proof. Use Taylor expansion and the Central Limit Theorem applied to $\overline{g}$:

$$\hat{\theta} - \theta_0 = m^{-1}(\overline{g}) - m^{-1}(m(\theta_0)) \approx (m^{-1})'(m(\theta_0))[\overline{g} - m(\theta_0)] + \frac{1}{2}(m^{-1})''(m(\theta_0))[\overline{g} - m(\theta_0)]^2.$$

□

Remark 1. *What this result tells us is that the smaller $\frac{Var_{\theta_0}[g(X_1)]}{(m'(\theta_0))^2}$ is the better is the estimate $\hat{\theta}$ in the sense that it has smaller deviations from the unknown parameter θ_0 asymptotically.*

2 Maximum Likelihood Estimator for Parametric Estimation

Definition 3 (Likelihood function)**.** *Intuitively, this is the probability to observe the sample $X_1, \cdots, X_n$:*

$$\varphi(\theta) = f(X_1|\theta) \times \cdots \times f(X_n|\theta) = \mathbb{P}_\theta(X_1) \times \cdots \times \mathbb{P}_\theta(X_n) = \mathbb{P}_\theta(X_1, \cdots, X_n).$$

For convenience, we often consider the **log-likelihood function**

$$\log \varphi(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$

Definition 4 (Maximum likelihood estimator)**.** *Let $\hat{\theta}$ be the parameter that maximizes $\varphi(\theta)$, i.e.*

$$\varphi(\hat{\theta}) = \max_{\theta} \varphi(\theta).$$

Then $\hat{\theta}$ is called the **maximum likelihood estimator** *(MLE).*

Why the MLE $\hat{\theta}$ converges to the unknown parameter θ_0? This is not immediately obvious. Defne $L(\theta) = \mathbb{E}_{\theta_0}[l(X|\theta)] = \mathbb{E}_{\theta_0}[\log f(X|\theta)]$.

Lemma 1. *We have, for any θ,*

$$L(\theta) \leq L(\theta_0).$$

Moreover, the inequality is strict $L(\theta) < L(\theta_0)$ unless

$$\mathbb{P}_{\theta_0}(f(X|\theta) = f(X|\theta_0)) = 1,$$

which means that $\mathbb{P}_\theta = \mathbb{P}_{\theta_0}$.

Proof. The intuition is that θ_0 is the maximizer of $L(\theta) = \mathbb{E}_{\theta_0}[\log f(X|\theta)]$. □

Theorem 2 (**Consistency of MLE**)**.** *Under some regularity conditions on the family of distributions, MLE $\hat{\theta}$ is consistent, i.e. $\hat{\theta} \to \theta_0$ as $n \to \infty$.*

Proof. 1) $\hat{\theta}$ is the maximizer of $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$ (by definition).

2) θ_0 is the maximizer of $L(\theta)$ (by Lemma).

3) $\forall\theta$ we have $L_n(\theta) \to L(\theta)$ by LLN.

Therefore, since two functions L_n and L are getting closer, the points of maximum should also get closer which exactly means that $\hat{\theta} \to \theta_0$. □

Recall we defined the function $l(X|\theta) = \log f(X|\theta)$. Denote by $l'(X|\theta)$ and $l''(X|\theta)$ the derivatives of $l(X|\theta)$ with respect to θ. **Fisher information** of a random variable X with distribution $\mathbb{P}_{\theta_0}$ from the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is defined by

$$I(\theta_0) = \mathbb{E}_{\theta_0}[(l'(X|\theta_0))^2] = \mathbb{E}_{\theta_0}\left[\left(\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right)^2\right].$$

Lemma 2. *We have*

$$\mathbb{E}_{\theta_0}[l''(X|\theta_0)] = \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log f(X|\theta_0)\right] = -I(\theta_0).$$

Theorem 3 (**Asymptotic normality of MLE**)**.** *We have*

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow N\left(0, \frac{1}{I(\theta_0)}\right).$$

Proof. By the mean value theorem and the fact that $\hat{\theta}$ maximizes $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n l(X_i|\theta)$, we have

$$0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\theta_1)(\hat{\theta} - \theta_0)$$

for some $\theta_1 \in [\hat{\theta}, \theta_0]$. From here we have

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\theta_1)}.$$

Since $L_n'(\theta_0) \to \mathbb{E}_{\theta_0}[l'(X_1|\theta_0)] = L'(\theta_0) = 0$ (the last equality is by Lemma 1), we can apply the CLT to the numerator: $\sqrt{n}L_n'(\theta_0) \xrightarrow{d} N(0, \mathrm{Var}_{\theta_0}(l'(X_1|\theta_0)))$. The denominator converges to $\mathbb{E}_{\theta_0}[l''(X|\theta_0)] = -I(\theta_0)$. Finally,

$$\mathrm{Var}_{\theta_0}(l'(X_1|\theta_0)) = \mathbb{E}_{\theta_0}[(l'(X_1|\theta_0))^2] - (\mathbb{E}_{\theta_0}[l'(X_1|\theta_0)])^2 = I(\theta_0) - 0.$$

So

$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow N\left(0, \frac{1}{I(\theta_0)}\right).$$

□

3 Rao-Crámer Inequality and Efficient Estimators

Let us consider statistic

$$S = S(X_1, \cdots, Xn)$$

which is a function of the sample X_1, $\cdots$, X_n. Let us define a function

$$m(\theta) = \mathbb{E}_\theta[S]$$

where $\mathbb{E}_\theta$ is the expectation with respect to distribution $\mathbb{P}_\theta$.

Theorem 4 (The Rao-Crámer inequality). *We have*

$$Var_\theta(S) = \mathbb{E}_\theta[(S - m(\theta))^2] \geq \frac{(m'(\theta))^2}{nI(\theta)}.$$

This inequality becomes equality if and only if

$$S = \left[t(\theta) \sum_{i=1}^{n} l'(X_i|\theta) \right] + m(\theta)$$

for some function $t(\theta)$ and where $l(X_i|\theta) = \log f(X_i|\theta)$.

Definition 5 (Efficient estimators)**.** *Consider statistic $S = S(X_1, \cdots, X_n)$ and let*

$$m(\theta) = \mathbb{E}_\theta[S].$$

We say that S is an **efficient estimator** *of $m(\theta)$ if*

$$\mathbb{E}_\theta[(S - m(\theta))^2] = \frac{(m'(\theta))^2}{nI(\theta)},$$

i.e. equality holds in the Rao-Crámer's inequality.

In other words, efficient estimate S is the best possible unbiased estimate of $m(\theta)$ in the sense that it achieves the smallest possible value for the average squared deviation $\mathbb{E}_\theta[(S - m(\theta))^2]$ for all θ.

The condition for the equality to hold in the Rao-Crámer inequality means that efficient estimates do not always exist and they exist only if we can represent the derivative of log-likelihood l'_n as

$$l'_n = \sum_{i=1}^{n} l'(X_i|\theta) = \frac{S - m(\theta)}{t(\theta)},$$

where S does not depend on θ. In this case, S is an efficient estimate of $m(\theta)$.

Asymptotic efficiency of MLE. Suppose that the MLE $\hat{\theta}$ is unbiased:

$$\mathbb{E}[\hat{\theta}] = \theta.$$

If we take $S = \hat{\theta}$ and $m(\theta) = \theta$ then the Rao-Crámer's inequality implies that

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}.$$

Meanwhile, the asymptotic normality of MLE states that

$$\sqrt{n}(\hat{\theta} - \theta) \to N\left(0, \frac{1}{I(\theta)}\right).$$

In other words, $\mathrm{Var}(\hat{\theta}) \approx \frac{1}{nI(\theta)}$ for n sufficiently large. So, for large sample size n the MLE is almost best possible.

4 Bayesian Statistics

Definition 6 (Prior and posterior distributions)**.** *Think of parameter θ as a random variable and its distribution $\xi(\theta)$ is called* **prior distribution***. Let us emphasize that $\xi(\theta)$ is chosen a priori, i.e. before we even see the data.* **Posterior distribution** *of θ incorporates sample data and can be computed using Bayes formula*

$$\xi(\theta|X_1, \cdots, X_n) = \frac{f(X_1, \cdots, X_n|\theta)\xi(\theta)}{\int_\Theta f(X_1, \cdots, X_n|\theta)\xi(\theta)d\theta}.$$

Bayes estimators. The *Bayes estimator* $\hat{\theta}$ is defined as

$$\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n) = \mathbb{E}[\theta|X_1, \cdots, X_n] = \int \theta\xi(\theta|X_1, \cdots, X_n)d\theta.$$

The obvious motivation for this choice of $\hat{\theta}$ is that it is simply the average of the parameter with respect to posterior function that in some sense captures the information contained in the data and our prior intuition about the parameter. Let us summarize the construction of Bayes estimator.

1. Choose prior distribution of θ, $\xi(\theta)$.
2. Compute posterior distribution $\xi(\theta|X_1, \cdots, X_n)$.
3. Find the expectation of the posterior $\hat{\theta} = \mathbb{E}[\theta|X_1, \cdots, X_n]$.

Conjugate prior distributions. Often for many popular families of distributions the prior distribution $\xi(\theta)$ is chosen so that it is easy to compute the posterior distribution. This is done by choosing $\xi(\theta)$ that resembles the likelihood function $f(X_1, \cdots, X_n|\theta)$.

5 Sufficient Statistics

5.1 Neyman-Fisher factorization criterion

Definition 7 (Sufficient statistics)**.** *$T = T(X_1, \cdots, X_n)$ is called* **sufficient statistics** *if*

$$\mathbb{P}_\theta(X_1, \cdots, X_n|T) = \mathbb{P}'(X_1, \cdots, X_n|T),$$

i.e. the conditional distribution of the vector $(X_1, \cdots, X_n)$ given T does not depend on the parameter θ and is equal to $\mathbb{P}'$.

Sufficient statistics:

- Gives a way of compressing information about underlying parameter θ.
- Gives a way of improving estimator using sufficient statistics (see below).

Theorem 5 (Neyman-Fisher factorization criterion). *$T = T(X_1, \cdots, X_n)$ is sufficient statistics if and only if the joint p.d.f. of $(X_1, \cdots, X_n)$ can be represented as*

$$f(x_1, \cdots, x_n|\theta) = f(x_1|\theta)\cdots f(x_n|\theta) = u(x_1, \cdots, x_n)v(T(x_1, \cdots, x_n), \theta)$$

for some function u and v.

5.2 Rao-Blackwell theorem

Consider $\delta = \delta(X_1, \cdots, X_n)$, some estimator of unknown parameter θ_0, which corresponds to a true distribution $\mathbb{P}_{\theta_0}$ of the data. Suppose that we have a sufficient statistics $T = T(X_1, \cdots, X_n)$. Consider a new estimator of θ_0 given by

$$\delta'(X_1, \cdots, X_n) = \mathbb{E}_{\theta_0}[\delta(X_1, \cdots, X_n)|T(X_1, \cdots, X_n)].$$

Then δ' does not depend on θ_0 and depends on the data only through T.

Theorem 6 (The Rao-Blackwell theorem). *We have*

$$\mathbb{E}_{\theta_0}[(\delta' - \theta_0)^2] \leq \mathbb{E}_{\theta_0}[(\delta - \theta_0)^2].$$

5.3 Minimal jointly sufficient statistics

Definition 8. *Statistics* $(T_1, \cdots, T_k)$ *are* **minimal jointly sufficient** *if given any other jointly sufficient statistics* $(r_1, \cdots, r_m)$ *we have*

$$T_1 = g_1(r_1, \cdots, r_m), \cdots, T_k = g_k(r_1, \cdots, r_m),$$

i.e. T's can be expressed as a functions of r's.

Proposition 1. *Given any jointly sufficient statistics* $(r_1, \cdots, r_m)$ *the MLE* $\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)$ *is always a function of* $(r_1, \cdots, r_m)$*. As a consequence, if MLE* $\hat{\theta}$ *is jointly sufficient then it is minimal.*

Proof. $\hat{\theta}$ is the maximizer of the likelihood which by factorization criterion can be represented as

$$f(x_1, \cdots, x_n|\theta) = u(x_1, \cdots, x_n)v(r_1, \cdots, r_m, \theta).$$

But maximizing this over θ is equivalent to maximizing $v(r_1, \cdots, r_m, \theta)$ over θ, and the solution of this maximization problem depends only on $(r_1, \cdots, r_m)$, i.e. $\hat{\theta} = \hat{\theta}(r_1, \cdots, r_m)$. □

N.B. If we measure the quality of an estimate via the average squared error loss function, then Rao-Blackwell theorem tells us that we can improve any estimator by conditioning it on the sufficient statistics. This means that any "good" estimate must depend on the data only through this minimal sufficient statistics, otherwise, we can always improve it.

6 Estimates of Parameters of Normal Distribution

Definition 9 (χ^2_n distribution with n degrees of freedom)**.** *This is the distribution of the sum* $X_1^2 + \cdots + X_n^2$*, where* X_i*'s are i.i.d. standard normal, which is also a gamma distribution* $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$*.*

Definition 10 (Fisher distribution $F_{k,m}$ with degree of freedom k and m)**.** *Consider* X_1, $\cdots$, X_k *and* Y_1, $\cdots$, Y_m *all independent standard normal random variables. Distribution of the random variable*

$$Z = \frac{X_1^2 + \cdots + X_k^2}{Y_1^2 + \cdots + Y_m^2}$$

is called **Fisher distribution with degree of freedom** k **and** m*, and it is denoted as* $F_{k,m}$*. The p.d.f. of Fisher distribution with* k *and* m *degrees of freedom is given by*

$$f_{k,m}(t) = \frac{\Gamma\left(\frac{k+m}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\Gamma\left(\frac{m}{2}\right)} t^{\frac{k}{2}-1}(1+t)^{-\frac{k+m}{2}}.$$

Definition 11 (Student distribution or t-distribution with m degrees of freedom)**.** *Consider* X *and* Y_1, $\cdots$, Y_m *all independent standard normal random variables. The distribution of the random variable*

$$Z = \frac{X}{\frac{1}{m}(Y_1^2 + \cdots + Y_m^2)}$$

is called the **Student distribution** *or* t**-distribution with** m **degrees of freedom** *and it is denoted by* t_m*. Its p.d.f. is*

$$f_Z(t) = \frac{t}{m} f_{1,m}\left(\frac{t^2}{m}\right) = \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{1}{\sqrt{m}} \left(1 + \frac{t^2}{m}\right)^{-\frac{m+1}{2}}.$$

Let us consider a sample $X_1, \cdots, X_n \sim N(\alpha, \sigma^2)$. By LLN we know that

$$\overline{X} \to \alpha,\ \overline{X}^2 - (\overline{X})^2 \to \sigma^2,\ n \to \infty.$$

The question is: *how close are these estimates to actual values of unknown parameters?* As a first step, we consider the case of $X \sim N(0, 1)$.

Theorem 7. *If $X_1, \cdots, X_n$ are i.i.d. standard normal, then sample mean $\overline{X}$ and sample variance $\overline{X}^2 - (\overline{X})^2$ are independent:*

$$\sqrt{n}\cdot\overline{X} \sim N(0,1),\ n(\overline{X}^2 - (\overline{X})^2) \sim \chi^2_{n-1}.$$

Proof. The key idea is to consider $Y = AX$, where A is an orthogonal matrix, such that $Y_1 = \sqrt{n}\cdot\overline{X}$. Then

$$n(\overline{X}^2 - (\overline{X})^2) = X_1^2 + \cdots + X_n^2 - Y_1^2 = Y_1^2 + \cdots + Y_n^2 - Y_1^2 = Y_2^2 + \cdots + Y_n^2.$$

□

Corollary 1. *Suppose $X_1, \cdots, X_n$ are i.i.d. with distribution $N(\alpha_0, \sigma_0^2)$. Then*

$$A = \frac{\sqrt{n}(\overline{X} - \alpha_0)}{\sigma_0} \sim N(0,1),\ B = \frac{n(\overline{X^2} - (\overline{X})^2)}{\sigma_0^2} \sim \chi^2_{n-1}$$

and the random variables A and B are independent.

Moreover, if we look at the p.d.f. of χ^2_{n-1} distribution and choose the constants c_1 and c_2 so that the area in each tail is $(1-\alpha)/2$, then the interval

$$\left[\frac{n(\overline{X^2} - (\overline{X})^2)}{c_2}, \frac{n(\overline{X^2} - (\overline{X})^2)}{c_1}\right]$$

is the α confidence interval for the unknown variance σ_0^2. If we look at the p.d.f. of t_{n-1} distribution and choose the constants $-c$ and c so that the area in each tail is $(1-\alpha)/2$, then the interval

$$\left[\overline{X} - c\sqrt{\frac{1}{n-1}(\overline{X^2} - (\overline{X})^2)}, \overline{X} + c\sqrt{\frac{1}{n-1}(\overline{X^2} - (\overline{X})^2)}\right]$$

is the α confidence interval for the unknown mean α_0.

7 Testing Hypotheses

7.1 Bayes decision rule

Definition 12 (Bayes decision rule)**.** *Given hypothesis $H_1, \cdots, H_k$, let us consider k nonnegative weights $\xi(1), \cdots, \xi(k)$ that add up to one: $\sum_{i=1}^k \xi(i) = 1$. Then the* **Bayes error** *of a decision rule δ is defined as*

$$\alpha(\xi) = \sum_{i=1}^k \xi(i)\alpha_i = \sum_{i=1}^k \xi(i)\mathbb{P}_i(\delta \neq H_i),$$

which is simply a weighted error. Decision rule δ that minimizes $\alpha(\xi)$ is called **Bayes decision rule***.*

Theorem 8. *Assume that each distribution $\mathbb{P}_i$ has p.d.f. $f_i(x)$. Then*

$$\delta = H_j,\ \textit{if}\ f_j f_j(X_1)\cdots f_j(X_n) = \max_{1\le i\le k} \xi(i) f_i(X_1)\cdots f_i(X_n)$$

is the Bayes decision rule. In other words, we choose hypothesis H_j if it maximizes the weighted likelihood function

$$\xi(i) f_i(X_1)\cdots f_i(X_n)$$

among all hypotheses. If this maximum is achieved simultaneously on several hypotheses, we can pick any one of them, or at random.

In the special case of two simple hypotheses, the Bayes decision rule becomes **likelihood ratio test**: $f_i(X) = f_i(X_1)\cdots f_i(X_n)$,

$$\delta = \begin{cases} H_1, & \frac{f_1(X)}{f_2(X)} > \frac{\xi(2)}{\xi(1)}; \\ H_2, & \frac{f_1(X)}{f_2(X)} < \frac{\xi(2)}{\xi(1)}; \\ H_1 \text{ or } H_2, & \frac{f_1(X)}{f_2(X)} = \frac{\xi(2)}{\xi(1)}. \end{cases}$$

7.2 Most powerful test

Definition 13. *When we only have two hypotheses H_1 and H_2, the error of type 1*

$$\alpha_1 = \mathbb{P}_1(\delta \neq H_1)$$

is also called **size** *or* **level of significance** *of decision rule δ and one minus type 2 error*

$$\beta = 1 - \alpha_2 = 1 - \mathbb{P}_2(\delta \neq H_2) = \mathbb{P}_2(\delta = H_2)$$

is called the **power of** *δ.*

Ideally, we would like to make errors of all types as small as possible, but it is clear that there is a trade-off: if we want to decrease type 1 error, we will have to predict hypothesis 1 more often; but this will make a type 2 error more often if hypothesis 2 is actually the true one.

Given $\alpha \in [0, 1]$, we consider the class of decision rules

$$K_\alpha = \{\delta : \mathbb{P}_1(\delta \neq H_1) \leq \alpha\}$$

and will try to find $\delta \in K_\alpha$ that makes the type 2 error $\alpha_2 = \mathbb{P}_2(\delta \neq H_2)$ as small as possible.

Theorem 9 (Most powerful test for two simple hypotheses). *Assume that there exist a constant c such that*

$$\mathbb{P}_1\left(\frac{f_1(X)}{f_2(X)} < c\right) = \alpha,$$

then the decision rule

$$\delta = \begin{cases} H_1 : \frac{f_1(X)}{f_2(X)} \geq c, \\ H_2 : \frac{f_1(X)}{f_2(X)} < c \end{cases}$$

is the most powerful in class K_α.

Proof. Key idea of the proof: if the most powerful rule δ with controlled error of type 1 happens to be a Bayes rule, what does it look like? □

7.3 Randomized most powerful test

The condition in Theorem 9 is not always fulfilled, especially when we deal with discrete distributions. But if we look carefully at the proof of that theorem, this condition was only necessary to make sure that the likelihood ratio test has error type 1 exactly equal to α. The next theorem shows that the most powerful test in class K_α can always be found if one randomly breaks the tie between two hypotheses in a way that ensures that the error of type 1 is equal to α.

Theorem 10 (Randomized most powerful test). *Given any $\alpha \in [0, 1]$ we can always find $c \in [0, \infty)$ and $p \in [0, 1]$ such that*

$$\mathbb{P}_1\left(\frac{f_1(X)}{f_2(X)} < c\right) + (1 - p)\mathbb{P}_1\left(\frac{f_1(X)}{f_2(X)} = c\right) = \alpha.$$

In this case, the most powerful test $\delta \in K_\alpha$ is given by

$$\delta = \begin{cases} H_1, & \frac{f_1(X)}{f_2(X)} > \frac{\xi(2)}{\xi(1)}; \\ H_2, & \frac{f_1(X)}{f_2(X)} < \frac{\xi(2)}{\xi(1)}; \\ H_1 \text{ or } H_2, & \frac{f_1(X)}{f_2(X)} = \frac{\xi(2)}{\xi(1)}. \end{cases}$$

where in the last case of equality we break the tie at random by choosing H_1 with probability p and choosing H_2 with probability $1 - p$.

7.4 Uniformly most powerful test

We assume that the sample $X_1, \cdots, X_n$ has distribution $\mathbb{P}_{\theta_0}$ that comes from a set of probability distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Our hypotheses will be

$$H_1 : \theta \in \Theta_1 \subseteq \Theta,\ H_2 : \theta \in \Theta_2 \subseteq \Theta.$$

Given some decision rule δ, consider the **power function** of δ:

$$\Pi(\delta, \theta) = \mathbb{P}_\theta(\delta \neq H_1).$$

Ideally, we would like to minimize the power function for all $\theta \in \Theta_1$ and maximize it for all $\theta \in \Theta_2$.

Definition 14. *If we can find $\delta \in K_\alpha = \{\delta : \sup_{\theta\in\Theta_1} \Pi(\delta, \theta) \leq \alpha\}$ such that*

$$\Pi(\delta, \theta) \geq \Pi(\delta', \theta) \textit{ for all } \theta \in \Theta_2 \textit{ and all } \delta' \in K_\alpha,$$

then δ is called the **uniformly most powerful (UMP) test***.*

Suppose the parameter $\Theta \subseteq \mathbb{R}$ is a subset of the real line and that probability distributions $\mathbb{P}_\theta$ have p.d.f. $f(x|\theta)$. Given a sample $X = (X_1, \cdots, X_n)$, the likelihood function is given by

$$f(X|\theta) = \Pi_{i=1}^n f(X_i|\theta).$$

Definition 15. *The set of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ has* **monotone likelihood ration** *(MLR) if we can represent the likelihood ratio as*

$$\frac{f(X|\theta_1)}{f(X|\theta_2)} = V(T(X), \theta_1, \theta_2)$$

and for $\theta_1 > \theta_2$ the function $V(T, \theta_1, \theta_2)$ is strictly increasing in T.

Theorem 11 (**Uniformly most powerful test for one-sided hypothesis**)**.** *Suppose that we have monotone likelihood ration with $T = T(X)$ and we consider one-sided hypotheses:*

$$H_1 : \theta \leq \theta_0,\ H_2 : \theta > \theta_0.$$

For any level of significance $\alpha \in [0, 1]$, we can find $c \in \mathbb{R}$ and $p \in [0, 1]$ such that

$$\mathbb{P}_{\theta_0}(T(X) > c) + (1-p)\mathbb{P}_{\theta_0}(T(X) = c) = \alpha.$$

Then the following test δ^* will be the uniformly most powerful test with level of significance α:

$$\delta^* = \begin{cases} H_1 : T < c \\ H_2 : T > c \\ H_1 \textit{ or } H_2 : T = c \end{cases}$$

where in the last case of $T = c$ we randomly pick H_1 with probability p and H_2 with probability $1 - p$.

8 χ^2 Goodness-of-Fit Test

8.1 Pearson's theorem

Let us consider r boxes $B_1, \cdots, B_r$. Assume that we throw n balls $X_1, \cdots, X_n$ into these boxes randomly independently of each other with probabilities

$$\mathbb{P}(X_i \in B_1) = p_1, \cdots, \mathbb{P}(X_i \in B_r) = p_r.$$

where probabilities add up to one $p_1 + \cdots + p_r = 1$. Let ν_j be a number of balls in the jth box. On average, the number of balls in the jth box will be np_j, so random variable ν_j should be close to np_j. One can also use CLT to describe how close ν_j is to np_j (note $\frac{\nu_i - np_i}{\sqrt{np_i}} \to N(0, 1 - p_i)$). The next results tells us how we can describe in some sense the closeness of ν_j to np_j simultaneously for all $j \leq r$. The main difficulty in this theorem comes from the fact that random variables ν_j for $j \leq r$ are not independent.

Theorem 12. *We have that the random variable*

$$\sum_{j=1}^{n} \frac{(\nu_j - np_j)^2}{np_j} \to \chi^2_{r-1}$$

converges in distribution to χ^2_{r-1} distribution with $(r-1)$ degrees of freedom.

8.2 χ^2 goodness-of-fit test for discrete distribution

Suppose that we observe an i.i.d. sample $X_1, \cdots, X_n$ of random variables that can take a finite number of values $B_1, \cdots, B_r$ with some unknown probabilities

$$p_1 = \mathbb{P}(X = B_1), \cdots, p_r = \mathbb{P}(X = B_r).$$

Suppose we want to test the hypothesis

$$\begin{cases} H_1 : p_i = p_i^0, \ i = 1, \cdots, r; \\ H_2 : \text{otherwise, i.e. for some } i, \ p_i \neq p_i^0. \end{cases}$$

By Pearson' theorem and CLT, as sample size n increases the distribution of $T = \sum_{i=1}^{r} \frac{(\nu_i - np_i^0)^2}{np_i^0}$ under hypothesis H_1 will approach χ^2_{r-1} distribution and under hypothesis H_2 it will shift to ∞. Therefore, the following test looks very natural

$$\delta = \begin{cases} H_1 : T \leq c \\ H_2 : T > c, \end{cases}$$

where the threshold c is so chosen that the type 1 error is equal to the level of significance α:

$$\alpha = \mathbb{P}_1(\delta \neq H_1) = \mathbb{P}_1(T > c) \approx \chi^2_{r-1}(c, \infty).$$

8.3 χ^2 goodness-of-fit test for continuous distribution

Let $X_1, \cdots, X_n$ be the sample from unknown distribution $\mathbb{P}$ and consider the following hypothesis:

$$\begin{cases} H_1 : \mathbb{P} = \mathbb{P}_0 \\ H_2 : \mathbb{P} \neq \mathbb{P}_0 \end{cases}$$

for some particular $\mathbb{P}_0$. To use the result from previous section, we will discretize the set of possible values of X's by splitting it into a finite number of intervals $I_1, \cdots, I_r$ and define

$$p_j^0 = \mathbb{P}_0(X \in I_j), \ j = 1, \cdots, r.$$

Instead of testing H_1 vs. H_2, we will consider the following weaker hypotheses

$$\begin{cases} H_1' : \mathbb{P}(X \in I_j) = p_j^0, \text{ for all } j \leq r \\ H_2' : \text{otherwise.} \end{cases}$$

The rule of thumb about how to split into subintervals $I_1, \cdots, I_r$ is to have the expected count in each subinterval

$$np_i^0 = n\mathbb{P}_0(X \in I_i) \geq 5.$$

8.4 χ^2 goodness-of-fit test for composite hypotheses

Suppose that we observe an i.i.d. sample $X_1, \cdots, X_n$ of random variables that can take a finite number of values $B_1, \cdots, B_r$ with some unknown probabilities

$$p_1 = \mathbb{P}(X = B_1), \cdots, p_r = \mathbb{P}(X = B_r).$$

Suppose we want to test the hypothesis that this distribution comes from a parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$. In other words, if we denote $p_j(\theta) = \mathbb{P}_\theta(X = B_j)$, we want to test

$$\begin{cases} H_1 : p_j = p_j(\theta), \text{ for all } j \leq r \text{ for some } \theta \in \Theta; \\ H_2 : \text{otherwise.} \end{cases}$$

One way to approach this problem is as follows:
Step 1. Assuming that hypothesis H_1 holds, we can find an estimate θ^* of this unknown θ.
Step 2. Try to test whether indeed the distribution $\mathbb{P}$ is equal to $\mathbb{P}_{\theta^*}$ by using the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)}$$

in χ^2 goodness-of-fit test.

This approach looks natural, the only question is what estimate θ^* to use and how the fact that θ^* also depends on the data will affect the convergence of T. It turns out that if we let θ^* be the MLE, i.e. θ that maximizes the likelihood function

$$\varphi(\theta) = p_1(\theta)^{\nu_1} \cdots p_r(\theta)^{\nu_r},$$

then the statistic

$$T = \sum_{j=1}^{r} \frac{(\nu_j - np_j(\theta^*))^2}{np_j(\theta^*)}$$

converges to χ^2_{r-s-1} distribution with $r - s - 1$ degrees of freedom, where s is the dimension of the parameter set Θ. Very informally, by dimension we mean the number of free parameters that describe the set Θ.

General families. We could use a similar test when the distributions $\mathbb{P}_\theta$, $\theta \in \Theta$ are not necessarily supported by a finite number of points. In this case, we use discretization as before and test the derived hypotheses.

8.5 χ^2 test of independence

Suppose we have an i.i.d. sample $X_1, \cdots, X_n$ with some distribution $\mathbb{P}$ on $\mathcal{X} = \{(i, j) : i = 1, \cdots, a; j = 1, \cdots, b\}$. Then each X_i is a pair (X_i^1, X_i^2) where X_i^1 can take a different values and X_i^2 can take b different values. Let N_{ij} be a count of all observations equal to (i, j). We would like to test the independence of these two features:

$$\mathbb{P}(X = (i, j)) = \mathbb{P}(X^1 = i)\mathbb{P}(X^2 = j).$$

If we introduce the notations

$$\mathbb{P}(X = (i, j)) = \theta_{ij}, \ \mathbb{P}(X^1 = i) = p_i, \ \mathbb{P}(X^2 = j) = q_j,$$

our hypotheses can be formulated as

$$\begin{cases} H_1 : \theta_{ij} = p_i q_j \text{ for some } (p_1, \cdots, p_a) \text{ and } (q_1, \cdots, q_b) \\ H_2 : \text{otherwise.} \end{cases}$$

These hypotheses fall into the case of composite χ^2 goodness-of-fit test from previous lecture because our random variables take $r = a \times b$ possible values and we want to test that their distribution comes from the family of distributions with independent features described by the hypothesis H_1. Since p_i's and q_j's should add up to 1, the dimension of the parameter set is $s = (a-1) + (b-1)$. Therefore, if we find the MLE for the parameters of this model then the chi-squared statistic:

$$T = \sum_{i,j} \frac{(N_{ij} - np_i^* q_j^*)^2}{np_i^* q_j^*} \to \chi^2_{r-s-1} = \chi^2_{ab-(a-1)-(b-1)-1} = \chi^2_{(a-1)(b-1)}.$$

Denote $N_{i+} = \sum_j N_{ij}$ and $N_{+j} = \sum_i N_{ij}$. Then explicit calculation gives $p_i^* = \frac{N_{i+}}{n}$ and $q_j^* = \frac{N_{+j}}{n}$. Therefore, the chi-square statistic T in this case can be written as

$$T = \sum_{i,j} \frac{(N_{ij} - N_{i+}N_{+j}/n)^2}{N_{i+}N_{+j}/n}$$

and the decision rule is given by

$$\delta = \begin{cases} H_1 : T \le c \\ H_2 : T > c \end{cases}$$

where the threshold is determined from the condition $\chi^2_{(a-1)(b-1)}(c, +\infty) = \alpha$.

8.6 χ^2 test of homogeneity

Suppose that the population is divided into R groups and each group (or the entire population) is divided into C categories. We would like to test whether the distribution of categories in each group is the same.

Formally, if we denote $\mathbb{P}(\text{Category}_j|\text{Group}_i) = p_{ij}$ so that for each group $i \le R$ we have $\sum_{j=1}^{C} p_{ij} = 1$, then we want to test the following hypotheses:

$$\begin{cases} H_1 : p_{ij} = p_j,\ i = 1, \cdots, R; \\ H_2 : \text{otherwise.} \end{cases}$$

If the observations $X_1, \cdots, X_n$ are sampled independently from the entire population then the homogeneity over groups is the same as the independence of groups and categories:

$$\mathbb{P}(\text{Category}_j|\text{Group}_i) = \mathbb{P}(\text{Category}_j) \Longleftrightarrow \mathbb{P}(\text{Group}_i, \text{Category}_j) = \mathbb{P}(\text{Category}_j)\mathbb{P}(\text{Group}_i).$$

This means that to test homogeneity we can use the independence test from previous lecture.

9 Kolmogorov-Smirnov Test

Let us denote by $F(x) = \mathbb{P}(X \le x)$ a cumulative distribution function and consider what is called an *empirical distribution function*:

$$F_n(x) = \mathbb{P}_n(X \le x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}}.$$

It is easy to show that $F_n(x) \to F(x)$ (by LLN) and this approximation holds uniformly over all $x \in \mathbb{R}$:

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \to 0.$$

The key observation in the Kolmogorov-Smirnov test is that the distribution of this supremum does not depend on the distribution $\mathbb{P}$ of the sample.

Theorem 13. *The distribution of* $\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ *does not depend on* F.

Next, we note that for a fixed x the CLT implies that

$$\sqrt{n}(F_n(x) - F(x)) \to N(0, F(x)(1 - F(x))$$

because $F(x)(1-F(x))$ is the variance of $I_{\{X \le x\}}$. It turns out that if we consider $\sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$, it will also converge to some distribution.

Theorem 14 (Kolmogorov-Smirnov distribution). *We have*

$$\mathbb{P}(\sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le t) \to H(t) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t}$$

where $H(t)$ is the c.d.f. of Kolmogorov-Smirnov distribution.

If we formulate our hypotheses in terms of cumulative distribution functions:

$$\begin{cases} H_1 : F = F_0 \text{ for a given } F_0 \\ H_2 : \text{otherwise,} \end{cases}$$

then based on Theorem 13 and 14, the **Kolmogorov-Smirnov test** is formulated as follows:

$$\delta = \begin{cases} H_1 : D_n \le c \\ H_2 : D_n > c \end{cases}$$

where $D_n = \sqrt{n}\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| = \sqrt{n}\sup_{0\le y\le 1} \left|\frac{1}{n}\sum_{i=1}^n I_{\{U_i \le y\}} - y\right|$ and the threshold c depends on the level of significance α and can be found from the condition

$$\alpha = \mathbb{P}(\delta \ne H_1 | H_1) = \mathbb{P}(D_n \ge c | H_1).$$

Theorem 13 shows D_n does not depend on the unknown distribution F and, therefore, it can be tabulated. Another way to find c, especially when the sample size is large, is to use Theorem 14 which tells that the distribution of D_n can be approximated by the Kolmogorov-Smirnov distribution and, therefore,

$$\alpha = \mathbb{P}(D_n \ge c | H_1) \approx 1 - H(c).$$

10 Simple Linear Regression

A *simple linear regression* model for the response variable Y is

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where ε is independent of X and is $N(0, \sigma^2)$. Using maximum likelihood estimation, we have the estimates

$$\hat{\beta}_1 = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - (\overline{X})^2},\ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X},\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2.$$

Explicit calculation shows

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n(\overline{X^2} - \overline{X}^2)}\right),\ \hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\overline{X}^2}{n(\overline{X^2} - \overline{X}^2)}\right)\sigma^2\right),$$

and

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\overline{X}\sigma^2}{n(\overline{X^2} - \overline{X}^2)}.$$

10.1 Joint distribution of the estimates

Proposition 2. *$\hat{\sigma}^2$ is independent of $\hat{\beta}_0$ and $\hat{\beta}_1$; $\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.*

10.2 Statistical inference in simple linear regression

Proposition 3. *If we find constants c_1 and c_2 such that $\chi^2_{n-2}(0, c_1) = \frac{\alpha}{2}$ and $\chi^2_{n-2}(c_2, \infty) = \frac{\alpha}{2}$, then the $1-\alpha$ confidence interval for σ^2 is*

$$\left[\frac{n\hat{\sigma}^2}{c_2}, \frac{n\hat{\sigma}^2}{c_1}\right].$$

If we find c such that $t_{n-2}(-c, c) = 1 - \alpha$, the $1-\alpha$ confidence interval for β_1 is

$$\left[\hat{\beta}_1 - c\sqrt{\frac{\hat{\sigma}^2}{(n-2)(\overline{X^2} - \overline{X}^2)}}, \hat{\beta}_1 + c\sqrt{\frac{\hat{\sigma}^2}{(n-2)(\overline{X^2} - \overline{X}^2)}}\right]$$

and the $1-\alpha$ confidence interval for β_0 is

$$\left[\hat{\beta}_0 - c\sqrt{\frac{\hat{\sigma}^2}{n-2}\left(1 + \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2}\right)}, \hat{\beta}_0 + c\sqrt{\frac{\hat{\sigma}^2}{n-2}\left(1 + \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2}\right)}\right]$$

Proposition 4. *For the prediction $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, we have*

$$\hat{Y} - Y \sim N\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(\overline{X} - X)^2}{n(\overline{X^2} - \overline{X}^2)}\right]\right),$$

$$\frac{\hat{Y} - Y}{\sqrt{\sigma^2\left[1 + \frac{1}{n} + \frac{(\overline{X} - X)^2}{n(\overline{X^2} - \overline{X}^2)}\right]}} \Big/ \sqrt{\frac{1}{n-2}\frac{n\hat{\sigma}^2}{\sigma^2}} \sim t_{n-2}$$

and the $1-\alpha$ prediction interval for Y is

$$\left[\hat{Y} - c\sqrt{\frac{\sigma^2}{n-2}\left(n + 1 + \frac{(\overline{X} - X)^2}{\overline{X^2} - \overline{X}^2}\right)}, \hat{Y} + c\sqrt{\frac{\sigma^2}{n-2}\left(n + 1 + \frac{(\overline{X} - X)^2}{\overline{X^2} - \overline{X}^2}\right)}\right]$$

11 AdaBoost Algorithm for Classification Problem

References

[1] Dmitry Panchenko. MIT OpenCourseWare: *18.443. Statistics for Applications, Fall 2003.* http://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2003/. 1