Tutorial 2

Identities for Estimating Moments

1. Let $X_1, \ldots, X_n$ be a random sample, with sample average $\overline{X} = \frac{1}{n}\sum_{j=1}^n X_j$ and sample variance $S^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2$. Show that

$$S^2 = \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2$$

You may use:

$$\sum_{j=1}^n y_j = \frac{1}{2n} \sum_{j,k=1}^n (y_j + y_k)$$

and $x^2 + y^2 = (x-y)^2 + 2xy$.

2. Assume that $\mathbb{E}[X_i^4] < +\infty$ and set $\theta_1 = \mathbb{E}[X_i]$, $\theta_j = \mathbb{E}[(X_i - \theta_1)^j]$ for $j = 2, 3, 4$. Let $Y_j = X_j - \theta_1$, $\overline{Y} = \frac{1}{n}\sum_{j=1}^n Y_j$ and $\overline{Y^2} = \frac{1}{n}\sum_{j=1}^n Y_j^2$.

   (a) Compute $\mathbb{E}[\overline{Y}^4]$ and $\mathbb{E}[\overline{Y^2}^2]$ and $\mathbb{E}[\overline{Y^2}\,\overline{Y}^2]$ in terms of θ_1, θ_2, θ_3 and θ_4.

   (b) Show that

$$\mathrm{Var}(S^2) = \frac{1}{n}\left(\theta_4 - \frac{n-3}{n-1}\theta_2^2\right).$$

   (c) Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ population.

      i. Find expressions for $\theta_1, \theta_2, \theta_3, \theta_4$ in terms of μ and σ^2.

      ii. Hence compute $\mathrm{Var}(S^2)$ for a $N(\mu, \sigma^2)$ random sample.

3. Establish the following recursion relations for means and variances. Let $\overline{X}_n$ and S_n^2 be the mean and variance respectively of $X_1, \ldots, X_n$. Suppose another observation X_{n+1} becomes available. Show that

   (a)

$$\overline{X}_{n+1} = \frac{X_{n+1} + n\overline{X}_n}{n+1}$$

   (b)

$$nS_{n+1}^2 = (n-1)S_n^2 + \left(\frac{n}{n+1}\right)(X_{n+1} - \overline{X}_n)^2.$$

Parametric Families: Identifiability Let $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a family of probability distributions. The parametrisation θ is said to be *identifiable* if $\theta_1 \neq \theta_2 \Rightarrow \mathbb{P}_{\theta_1} \neq \mathbb{P}_{\theta_2}$. For example, let $\theta = (\mu, \sigma^2)$ and $\mathbb{P}_\theta$ denote the $N(\mu, \sigma^2)$ distribution. The parameterisation is *identifiable* since

$$(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2) \Rightarrow \exists A \in \mathcal{B}(\mathbb{R}) : \mathbb{P}_{\theta_1}(A) \neq \mathbb{P}_{\theta_2}(A)$$

where $\mathcal{B}(\mathbb{R})$ denotes the Borel subsets of $\mathbb{R}$.

On the other hand, the parametrisation $\theta = (\mu, \nu, \sigma^2)$ where $\mathbb{P}_\theta$ is $N(\mu - \nu, \sigma^2)$ is not identifiable, since $\theta_1 = (\mu, \nu, \theta)$ and $\theta_2 = (\mu + a, \nu + a, \theta)$ give the same distribution.

4. (a) Let $X_{ij} : i = 1, \dots, p; j = 1, \dots, b$ be independent with $X_{ij} \sim N(\mu_{ij}, \sigma^2)$. Let $\mu_{ij} = \nu + \alpha_i + \beta_j$. Let $\theta = (\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_b, \nu, \sigma^2)$ and $\mathbb{P}_\theta$ the distribution of $X_{11}, \dots, X_{pb}$. Is the parametrisation identifiable? Prove or disprove.

   (b) Now suppose that $(\alpha_1, \dots, \alpha_p)$ and $(\beta_1, \dots, \beta_b)$ are restricted to the sets $\sum_{i=1}^{p} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$. Is the parametrisation identifiable? Prove or disprove.

5. A measuring instrument is being used to obtain n independent determinations of a physical constant μ. Suppose that the measuring instrument is known to be biased by a positive constant θ units, where θ is unknown and that the errors are otherwise identically distributed normal random variables with known variance σ^2. Is the parametrisation identifiable? Prove or disprove.

6. The number of eggs laid by an insect follows a Poisson distribution with unknown mean μ. Once laid, each egg has an unknown chance p of hatching, independently of the others. An entomologist studies a set of n such insects, observing only the number of eggs hatching for each nest. Is the parametrisation identifiable?

Hazard and Survival

7. Let $T_1, \dots, T_m$ and $T'_1, \dots, T'_n$ be random samples with parent variables T and T' respectively, which are the survival times of two groups of patients receiving treatments A and B respectively. The *group survival* for the two groups is defined as $X = \min_{j=1,\dots,m} T_j$ and $Y = \min_{j=1,\dots,n} T'_j$ respectively. Let $S_X(t) = \mathbb{P}(X > t)$ and $S_Y(t) = \mathbb{P}(Y > t)$ denote the *group survival functions*. Assume that the groups are independent of each other and that T and T' have the same distribution.

   (a) Show that $S_Y(t) = S_X^{n/m}(t)$.

   (b) Extending from rationals to $\delta \in (0, +\infty)$ gives the *Lehmann model*: $S_Y(t) = S_X^{\delta}(t)$. Equivalently, $S_Y(t) = S_0^{n\delta}(t)$ and $S_X(t) = S_0^{m\delta}(t)$ for some survival function S_0. Suppose that X is a non negative continuous random variable with survival function $S_X(t) = S_0^{m\delta}(t)$. Compute the distribution function of $X' := -\log S_0(X)$.

   (c) Suppose that T and Y are two non-negative continuous random variables with survival functions $S_T(t)$ and $S_Y(t)$ respectively and densities $f_T(t)$ and $f_Y(t)$ respectively. Their *hazard functions* are defined as $\alpha_T(t) = \frac{f_T(t)}{S_T(t)}$ and $\alpha_Y(t) = \frac{f_Y(t)}{S_Y(t)}$ respectively. Show that $\alpha_Y = c\alpha_T$ if and only if $S_Y = S_T^c$. Such a model is known as the *Cox proportional hazard model*.

Order Statistics and Glivenko-Cantelli Lemma

8. Let $X_1, \dots, X_n$ be i.i.d. random variables, with c.d.f. F and density f. The ordered vector $X_{1:n} \leq X_{2:n} \leq \dots \leq X_{n:n}$ which is an ordering of $X_1, \dots, X_n$ from lowest to highest is the vector of *order statistics*.

   (a) Find the c.d.f. and density of $X_{k:n}$.

(b) Hence, if $X_1, \ldots, X_n$ be a random sample from a $U(0,1)$ distribution (uniform on the interval $(0,1)$), show that the density function for the jth order statistic $X_{j:n}$ is

$$f_{X_{j:n}}(x) = j\binom{n}{j}x^{j-1}(1-x)^{n-j} \qquad x \in [0,1]$$

(c) Hence prove (again for a $U(0,1)$ random sample) that for positive integer p,

$$\mathbb{E}\left[X_{j:n}^p\right] = j\binom{n}{j}\frac{\Gamma(j+p)\Gamma(n-j+1)}{\Gamma(n+p+1)}.$$

You may assume the Beta integral:

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

9. Let F be a continuous cumulative distribution function, $X_1, \ldots, X_n$ a random sample generated from F and $\widehat{F}_n$ the empirical distribution function. Let $D_n = \sup_{-\infty<x<+\infty} |F(x) - \widehat{F}_n(x)|$. Prove that for any $\epsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}\left(\sup_{-\infty<x<+\infty} |F(x) - \widehat{F}_n(x)| > \epsilon\right) = 0.$$

You may use the result from the previous tutorial that the distribution of D_n does not depend on the underlying F (and hence assume that the random sample is $U(0,1)$).

Short Answers

1.

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(X_j-\overline{X})^2 \\
&= \frac{1}{2n(n-1)}\sum_{j,k=1}^{n}\{((X_j-\overline{X})^2+(X_k-\overline{X})^2)\} \\
&= \frac{1}{2n(n-1)}\sum_{j,k=1}^{n}\{(X_j-X_k)^2+2(X_j-\overline{X})(X_k-\overline{X})\} \\
&= \frac{1}{2n(n-1)}\sum_{j,k=1}^{n}(X_j-X_k)^2
\end{aligned}$$

because $\sum_j (X_j-\overline{X})=0$.

2. (a)

$$\mathbb{E}[\overline{Y}^4]=\frac{1}{n^4}\sum_{j_1,j_2,j_3,j_4=1}^{n}\mathbb{E}[Y_{j_1}Y_{j_2}Y_{j_3}Y_{j_4}]=\frac{1}{n^3}\theta_4+3\left(\frac{n-1}{n^3}\right)\theta_2^2$$

We're using $\mathbb{E}[Y_j^4]=\theta_4$ and noting there are n such terms, for $j\neq k$ $\mathbb{E}[Y_j^2Y_k^2]=\mathbb{E}[Y_j^2]^2=\theta_2^2$ and noting there are n^2-n such terms - and that terms not of this form vanish since $\mathbb{E}[Y_j]=0$.

$$\mathbb{E}[\overline{Y^2}^2]=\frac{1}{n^2}\sum_{j_1,j_2=1}^{n}\mathbb{E}[Y_{j_1}^2Y_{j_2}^2]=\frac{1}{n}\theta_4+\frac{n-1}{n}\theta_2^2.$$

$$\mathbb{E}[\overline{Y^2}\,\overline{Y}^2]=\frac{1}{n^3}\sum_{j_1,j_2,j_3}\mathbb{E}[Y_{j_1}^2Y_{j_2}Y_{j_3}]=\frac{1}{n^2}\theta_4+\frac{n-1}{n^2}\theta_2^2$$

(b) $\mathbb{E}\left[\overline{Y}^2\right]=\frac{\theta_2}{n}$ and $\mathbb{E}\left[\overline{Y^2}\right]=\theta_2$. For $j\neq k$, $\mathbb{E}[(Y_j-Y_k)^2]=2\theta_2$. Since

$$S^2=\frac{1}{2n(n-1)}\sum_{j,k}(Y_j-Y_k)^2=\frac{1}{n-1}\sum_j(Y_j-\overline{Y})^2=\frac{n}{n-1}\left(\overline{Y^2}-\overline{Y}^2\right)$$

$$\begin{aligned}
\mathrm{Var}(S^2) &= \frac{n^2}{(n-1)^2}\mathrm{Var}\left(\overline{Y^2}-\overline{Y}^2\right) \\
&= \frac{n^2}{(n-1)^2}(\mathbb{E}\left[\overline{Y^2}^2+\overline{Y}^4-2\overline{Y^2}\,\overline{Y}^2\right]-\mathbb{E}[\overline{Y^2}]^2-\mathbb{E}[\overline{Y}^2]^2+2\mathbb{E}[\overline{Y^2}]\mathbb{E}[\overline{Y}^2]) \\
&= \frac{n^2}{(n-1)^2}\left(\left(\frac{1}{n}\theta_4+\frac{n-1}{n}\theta_2^2\right)+\left(\frac{1}{n^3}\theta_4+3\left(\frac{n-1}{n^3}\right)\theta_2^2\right)\right. \\
&\quad \left.-2\left(\frac{1}{n^2}\theta_4+\frac{n-1}{n^2}\theta_2^2\right)-\theta_2^2-\frac{\theta_2^2}{n^2}+\frac{2\theta_2^2}{n}\right) \\
&= \frac{n^2}{(n-1)^2}\left(\frac{(n-1)^2}{n^3}\theta_4-\theta_2^2\frac{(n-1)(n-3)}{n^3}\right) \\
&= \frac{1}{n}\left(\theta_4-\frac{n-3}{n-1}\,\theta_2^2\right)
\end{aligned}$$

(c) i. $\theta_1 = \mu$, $\theta_2 = \sigma^2$, $\theta_3 = 0$, $\theta_4 = 3\sigma^4$. The only one that may cause problems is the last one:

$$\theta_4 = \int y^4 \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} dy = 2\int_0^\infty y^4 \frac{1}{\sqrt{2\pi}\sigma} e^{-y^2/2\sigma^2} dy$$

substitute (for example) $x = \frac{y^2}{2\sigma^2}$ $dx = \frac{ydy}{\sigma^2}$

$$\theta_4 = \frac{4\sigma^4}{\sqrt{\pi}} \int_0^\infty z^{3/2} e^{-z} dz = \frac{4\sigma^4 \Gamma(5/2)}{\sqrt{\pi}} = 3\sigma^4$$

ii.

$$\text{Var}(S^2) = \frac{1}{n}\left(3 - \frac{n-3}{n-1}\right)\sigma^4 = \frac{2}{n-1}\sigma^4.$$

3. (a)

$$\overline{X}_{n+1} = \frac{1}{n+1}\sum_{j=1}^{n+1} X_j = \frac{1}{n+1}\sum_{j=1}^{n} X_j + \frac{1}{n+1}X_{n+1} = \frac{n}{n+1}\overline{X}_n + \frac{1}{n+1}X_{n+1}$$

(b)

$$\begin{aligned} nS_{n+1}^2 &= \sum_{j=1}^{n+1}(X_j - \overline{X}_{n+1})^2 = \sum_{j=1}^{n}(X_j - \overline{X}_n)^2 + n(\overline{X}_n - \overline{X}_{n+1})^2 + (X_{n+1} - \overline{X}_{n+1})^2 \\ &= (n-1)S_n^2 + n\left(\frac{1}{n+1}\overline{X}_n - \frac{1}{n+1}X_{n+1}\right)^2 + (\frac{n}{n+1}X_{n+1} - \frac{n}{n+1}\overline{X}_n)^2 \\ &= (n-1)S_n^2 + \frac{n(1+n)}{(n+1)^2}(\overline{X}_n - X_{n+1})^2 = (n-1)S_n^2 + \frac{n}{n+1}\left(\overline{X}_n - X_{n+1}\right)^2. \end{aligned}$$

4. (a) Not identifiable: for example,

$$\mathbb{P}_{\nu,\sigma^2,\alpha_1,\ldots,\alpha_p,\beta_1,\ldots,\beta_b} = \mathbb{P}_{0,\sigma^2,\alpha_1+a\nu,\ldots,\alpha_p+a\nu,\beta_1+(1-a)\nu,\ldots,\beta_b+(1-a)\nu}$$

for any $a \in \mathbb{R}$.

(b) Yes - it is identifiable. Joint density is

$$\begin{aligned} &\frac{1}{(2\pi)^{pb/2}\sigma^{pb}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{ij}(x_{ij} - \nu - \alpha_i - \beta_j)^2\right\} \\ &= \frac{1}{(2\pi)^{pb/2}\sigma^{pb}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{ij} x_{ij}^2 - \sum_{ij} x_{ij}(\nu + \alpha_i + \beta_j) + \sum_{ij}(\nu + \alpha_i + \beta_j)^2\right)\right\} \end{aligned}$$

If it is not identifiable, then different $(\nu, \underline{\alpha}, \underline{\beta})$ yield the same $\nu + \alpha_i + \beta_j$ for each (i, j). If

$$\nu_1 + \alpha_{1i} + \beta_{1j} = \nu_2 + \alpha_{2i} + \beta_{2j} \qquad \forall (i, j)$$

plus zero sum conditions, then $\nu_1 = \nu_2$. Again, sum over j gives $\alpha_{1i} = \alpha_{2i}$ for each i and summing over i gives $\beta_{1j} = \beta_{2j}$. Hence it is identifiable.

5. Not identifiable; $\mathbb{P}_{\nu_1,\theta_1,\sigma^2} = \mathbb{P}_{\nu_2,\theta_2,\sigma^2}$ for all $(\mu_1, \theta_1), (\mu_2, \theta_2)$ such that $\mu_1 + \theta_1 = \mu_2 + \theta_2$.

6. The parametrisation is (μ, p). Let X denote number of eggs laid, Y the number that hatch. Then

$$\mathbb{P}(Y = y|X = x) = \binom{x}{y} p^y (1-p)^{x-y} \qquad \mathbb{P}(X = x) = \frac{\mu^x}{x!} e^{-\mu}$$
$$\mathbb{P}(Y = y, X = x) = \frac{x!}{y!(x-y)!} p^y (1-p)^{x-y} \frac{\mu^x}{x!} e^{-\mu} \qquad x \geq y$$

so that

$$\mathbb{P}(Y = y) = e^{-\mu} \frac{\mu^y p^y}{y!} \sum_{x=y}^{\infty} \frac{(1-p)^{x-y} \mu^{x-y}}{(x-y)!} = \frac{(\mu p)^y}{y!} e^{-\mu p}.$$

No not identifiable.

7. (a)

$$S_Y(t) = \mathbb{P}(\min(T_1', \ldots, T_n') > t) = \mathbb{P}(T > t)^n \qquad S_X(t) = \mathbb{P}(T > t)^m$$

from which the result follows directly.

(b)

$$\begin{aligned} F_{X'}(x) &= \mathbb{P}(X' \leq x) = \mathbb{P}(-\log S_0(X) \leq t) \\ &= \mathbb{P}(S_0(X) \geq e^{-t}) = \mathbb{P}(S_X(X) \geq e^{-m\delta t}) \\ &= \mathbb{P}(F_X(X) \leq 1 - e^{-m\delta t}) = 1 - e^{-m\delta t}. \end{aligned}$$

(c)

$$\alpha_T(t) = -\frac{d}{dt} \log S_T(t) \qquad \alpha_Y(t) = -\frac{d}{dt} \log S_Y(t).$$
$$\alpha_Y = c\alpha_T \Leftrightarrow -\frac{d}{dt} \log S_T(t) = -c\frac{d}{dt} \log S_Y(t) \Leftrightarrow -\frac{d}{dt} \log S_T(t) = -\frac{d}{dt} \log S_Y^c(t)$$

Now using $S_T(0) = S_Y(0) = 1$ gives:

$$S_T(t) = S_Y^c(t) \qquad \forall t \geq 1.$$

8. (a)

$$\mathbb{P}(X_{k:n} \leq x < X_{k+1:n}) = F_{X_{k:n}}(x) - F_{X_{k+1:n}}(x)$$

and

$$\begin{aligned} F_{X_{k:n}}(x) - F_{X_{k+1:n}}(x) &= \binom{n}{k} \mathbb{P}(X_1 \leq x, \ldots, X_k \leq x, X_{k+1} > x, \ldots, X_n > x) \\ &= \binom{n}{k} F(x)^k (1 - F(x))^{n-k}. \end{aligned}$$

To compute $F_{X_{k:n}}(x)$, we need $F_{X_{n:n}}(x)$, but this is easy:

$$F_{X_{n:n}}(x) = F(x)^n.$$

Therefore:

$$F_{X_{k:n}}(x) = \sum_{j=k}^{n} \binom{n}{j} F(x)^k (1-F(x))^{n-k}.$$

To compute the density, take a derivative:

$$\begin{aligned}
f_{X_{k:n}}(x) &= \sum_{j=k}^{n} \binom{n}{j} \left(jF(x)^{j-1}(1-F(x))^{n-j} - (n-j)F(x)^j(1-F(x))^{n-j-1} \right) f(x) \\
&= nf(x) \sum_{j=k}^{n} \left\{ \binom{n-1}{j-1} F(x)^{j-1}(1-F(x))^{n-j} - \binom{n-1}{j} F(x)^j (1-F(x))^{n-j-1} \right\} \\
&= n\binom{n-1}{k-1} F(x)^{k-1}(1-F(x))^{n-k} f(x)
\end{aligned}$$

so that:

$$f_{X_{k:n}}(x) = \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x).$$

(b) For $U(0,1)$, $F(x) = x$ for $0 \le x \le 1$ and $f(x) = \mathbf{1}_{[0,1]}(x)$ so that:

$$f_{X_{k:n}}(x) = n\binom{n-1}{k-1} x^{k-1}(1-x)^{n-k} \mathbf{1}_{[0,1]}(x)$$

as required.

(c)

$$\mathbb{E}[X_{j:n}^p] = j\binom{n}{j} \int_0^1 x^p x^{j-1}(1-x)^{n-j} dx = j\binom{n}{j} \frac{\Gamma(j+p)\Gamma(n-j+1)}{\Gamma(n+p+1)}.$$

Using $\Gamma(n+1) = n!$, it follows that

$$\mathbb{E}[X_{j:n}^p] = j\frac{n!}{j!(n-j)!} \frac{(j+p-1)!(n-j)!}{(n+p)!} = \frac{\prod_{k=0}^{p-1}(j+k)}{\prod_{k=1}^{p}(n+k)}.$$

9. First, for fixed ϵ, we consider the following grid: $x_1 = \inf\{z : F(z) \ge \epsilon\}$, $x_j = \inf\{z > x_{j-1} : F(z) - F(x_{j-1}) \ge \epsilon$, define M as the smallest integer such that $1 \ge F(x_M) > 1-\epsilon$. Since F is continuous, $F(x_j) - F(x_{j-1}) = \epsilon$ for $j = 2, \ldots, M$.

   Now, if $|\widehat{F}_n(x_j) - F(x_j)| \le \epsilon$ and $|\widehat{F}_n(x_{j+1}) - F(x_{j+1})| \le \epsilon$, then it is straightforward that $\sup_{x \in [x_j, x_{j+1}]} |\widehat{F}_n(x) - F(x)| \le 2\epsilon$. Therefore

$$\begin{aligned}
\mathbb{P}\left(\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| > \epsilon \right) &\le \mathbb{P}\left(\max_{j\in\{1,\ldots,M\}} |\widehat{F}_n(x_j) - F(x_j)| > \frac{\epsilon}{2} \right) \\
&\le \sum_{j=1}^{M} \mathbb{P}\left(|\widehat{F}_n(x_j) - F(x_j)| > \epsilon \right) \\
&\le M \times \frac{4}{\epsilon^2} \times \sup_x \frac{F(x)(1-F(x))}{n} \le \frac{1}{n\epsilon^3} \overset{n\to+\infty}{\longrightarrow} 0
\end{aligned}$$

   using the fact that $\mathbb{E}[\widehat{F}_n(x)] = F(x)$ and $\mathrm{Var}(\widehat{F}_n(x)) = \frac{F(x)(1-F(x))}{n} \le \frac{1}{4n}$.