Statistics II: Introduction to Inference

Week 5: Sampling Distributions

Definition 1 (Sampling Distribution)**.** *Let $X_1, \cdots, X_n$ be a random sample from F_θ and $T(\mathbf{X}) = g(X_1, \cdots, X_n)$ be a statistic. Clearly $T(\mathbf{X})$ is random and the distribution of $T(\mathbf{X})$ depends on (but possibly different from) F_θ. The distribution of $T(\mathbf{X})$ is called sampling distribution.*

Example 1. *Let $X_1, \cdots, X_n$ be a random sample from $U(0, \theta)$ distribution. What is the distribution of $X_{(n)}$?*

Example 2. *If $X_1, \cdots, X_n$ be a random sample from location exponential distribution,*

$$f_\theta(x) = \lambda e^{-\lambda(x-\alpha)} \quad x > \alpha.$$

Derive the distribution of $X_{(1)}$.

1 Some useful ways of finding the sampling distributions

(I) MGF Technique.

Definition 2 (MGF)**.** *Let $X \sim F_\theta$ then the MGF of X is defined as*

$$M_X(t) = \mathbb{E}_\theta(e^{tX}) \quad \textit{for } t \in (-h, h).$$

Remarks.

1. MGF may or may not exist (i.e., the expectation may not exist for t in a neighborhood of zero).
2. For a random variable X, if MGF exists then all moments of X exists.
3. If exists, MGF provides a characterization of the distribution, i.e., MGF, if exists, uniquely identifies the distribution.

Example 3. *Let $X \sim N(\mu, \sigma^2)$ then show that the MGF of X is $M_X(t) = \exp\left\{t\mu + \dfrac{t^2\sigma^2}{2}\right\}$.*

Using the MGF, we can now derive the sampling distribution of $T(\mathbf{X}) = X_1 + \ldots + X_n$. Observe that

$$M_T(t) = \mathbb{E}\left(\exp\{tT(\mathbf{X}\}\right) = \{M_{X_1}(t)\}^n = \exp\left\{tn\mu + \frac{t^2 n\sigma^2}{2}\right\}.$$

By uniqueness of MGF, we can identify that $T \sim N(n\mu, n\sigma^2)$ distribution.

Example 4. *Let $X_1, \cdots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$. Find the distribution of $Z = \sum_{i=1}^{n} X_i$.*

(II) Change of Variable Technique. Let $\boldsymbol{X}$ be an absolutely continuous random variable. Let $\boldsymbol{X} \to \boldsymbol{Y}$ be a one-one transformation; i.e., $\boldsymbol{Y} = h(\boldsymbol{X})$. Define Jacobian of the transformation as $J = J_{\boldsymbol{X}\to\boldsymbol{Y}} = \left|\det\left(\dfrac{\partial \boldsymbol{X}}{\partial \boldsymbol{Y}}\right)\right|$. Suppose the pdf of $\boldsymbol{X}$ is known as $f_{\boldsymbol{X}}(\boldsymbol{x})$. Then the joint pdf of $\boldsymbol{Y}$ is:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = f_{\boldsymbol{X}}(h^{-1}(\boldsymbol{y}))\ J_{\boldsymbol{X}\to\boldsymbol{Y}}.$$

Example 5. *Let $X, Y \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$. Show that $\dfrac{X}{X+Y} \sim \text{Beta}(\alpha, \alpha)$.*

Hint: You can use the following transformation $(X, Y) \to \left(W = X + Y, Z = \dfrac{X}{X+Y}\right)$.

Example 6. *Let X be an absolutely continuous random variable with CDF F. Then show that $F(X) = U$ has an $U(0,1)$ distribution.*

Example 6 is useful for simulating different random variables.

Example 7. *Generate a random sample from* Exponential(λ) *with* $\lambda = 1.5$.

{*Hint: Calculate the CDF F_X of X.*

1. *Generate a random sample between 0 and 1, say u (u acts as a realization of the RV $F_X(x) = u$).*
2. *Using inverse transformation we get a realization of X.*}

Example 8 (Box-Muller Transform). *Let $U_1, U_2 \overset{iid}{\sim} \text{Uniform}(0,1)$. Define $Z_1 = \sqrt{-2\log U_1}\cos(2\pi\, U_2)$ and $Z_2 = \sqrt{-2\log U_2}\sin(2\pi\, U_2)$. Show that $Z_1, Z_2 \overset{iid}{\sim} N(0,1)$.*

Hint: Use polar transformation $(Z_1, Z_2) \to (r, \theta)$ such that $Z_1 = r\cos(\theta)$ and $Z_2 = r\sin(\theta)$.

(III) Distribution Function Technique Suppose we are interested in finding the distribution of $T(\mathbf{X})$. Let us find the distribution function of $Y = T(\mathbf{X})$ directly.

$$\begin{aligned}
\mathbb{P}(T(X) \le y) &= \mathbb{P}\left(T(X) \in (-\infty, y]\right) \\
&= \mathbb{P}\left(X \in T^{-1}((-\infty, y]) = \mathcal{A}(y)\right) \\
&= \begin{cases} \sum_{x: x \in \mathcal{A}(y)} f_X(x), & \text{if } X \text{ is discrete} \\ \int_{\mathcal{A}(y)} f_X(x)dx, & \text{if } X \text{ is continuous,} \end{cases}
\end{aligned}$$

where $\mathcal{A}(y) = T^{-1}((-\infty, y]) = \{x : T(x) \le y\}$.

Example 9. *Let $X \sim N(0,1)$. Show that the distribution of $Y = X^2$ is* Gamma$\left(\frac{1}{2}, \frac{1}{2}\right)$.

2 Some Important Sampling Distributions

Example 10. *(Gamma distribution) If $X_1, \cdots, X_n \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$ then $\sum_{i=1}^{n} X_i \sim \text{Gamma}(n\alpha, \beta)$.*

2.1 The χ^2 distribution

Definition 3. *(χ^2 distribution) Let $X_1, \cdots, X_n \overset{iid}{\sim} N(0,1)$. Then $T = \sum_{i=1}^{n} X_i^2 \sim \chi^2_{(n)}$. By Examples 9 and 10, $\chi^2_{(n)} \equiv \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$.*

Properties:

1. From the properties of Gamma distribution, $\mathbb{E}\left(T \mid T \sim \chi^2_{(n)}\right) = n$ and $\text{Var}\left(T \mid T \sim \chi^2_{(n)}\right) = 2n$.
2. (Additive property) Let $T_i \overset{ind}{\sim} \chi^2_{(n_i)}$ for $i = 1, 2$. Then $T_1 + T_2 \sim \chi^2_{(n_1+n_2)}$.
3. If $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, and define $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ then $\frac{nS_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

The proof to show $\frac{nS_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Step 1. Let $Z_1, \dots, Z_n$ be a random sample from $N(0,1)$ distribution. Define

$$nS_z^2 = \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}_n^2.$$

Define $\mathbf{Z} = (Z_1, \dots, Z_n)'$ in vector form. Then $\sum_i Z_i^2 = \mathbf{Z}'\mathbf{Z}$ and $n\bar{Z}_n^2 = \mathbf{Z}'\mathbf{1}\mathbf{1}'\mathbf{Z}/n$, where $\mathbf{1}$ is a vector of ones. Then

$$nS_z^2 = \mathbf{Z}'\left(I - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{Z} = \mathbf{Z}'H\mathbf{Z},$$

where $H = I - \mathbf{1}\mathbf{1}'/n$.

Step 2. Consider the transformation $\boldsymbol{Y} = W\boldsymbol{Z}$, where

$$W = \begin{bmatrix} \mathbf{W}'_{(1)} \\ W_{(2)} \end{bmatrix}$$

and $W_{(1)} = [1/\sqrt{n},\ 1/\sqrt{n}, \dots,\ 1/\sqrt{n}]'$ is a $n \times 1$ vector and $W_{(2)}$ is a $(n-1) \times n$ matrix such that $W'W = I_n$.

The Jacobian of the transformation is $J = |\det(W)|^{-1} = 1$ as

$$\frac{\partial \mathbf{Y}}{\partial \mathbf{Z}} = W, \qquad \text{and} \qquad \det(W) = \pm 1.$$

Then

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= f_{\boldsymbol{Z}}(W^{-1}\boldsymbol{y}) \\ &= \frac{1}{(2\pi)^{n/2}}\, e^{-\boldsymbol{z}'\boldsymbol{z}/2} \\ &= \frac{1}{(2\pi)^{n/2}}\, e^{-\boldsymbol{y}'\boldsymbol{y}/2}, \end{aligned}$$

as $\mathbf{z}'\mathbf{z} = \mathbf{y}'(W'W)^{-1}\mathbf{y} = \mathbf{y}'\mathbf{y}$. Therefore, $\boldsymbol{Y} \sim N_n(\mathbf{0}, I_n)$. Now,

$$\boldsymbol{Y}'\boldsymbol{Y} = \boldsymbol{Z}'W'W\boldsymbol{Z} = \boldsymbol{Z}'\left(\frac{1}{n}\mathbf{1}\mathbf{1}' + \underbrace{I_n - \frac{1}{n}\mathbf{1}\mathbf{1}'}_{H}\right)\boldsymbol{Z} = Y_1^2 + nS_z^2.$$

(recall from Step 1 that $\boldsymbol{Z}'H\boldsymbol{Z} = nS_z^2$).

Hence, $Y_1^2 = \dfrac{1}{n}\left(\displaystyle\sum_{i=1}^n Z_i\right)^2 = n\bar{Z}_n^2$ and $\displaystyle\sum_{i=2}^n Y_i^2 = nS_z^2$. It implies $nS_z^2 \sim \chi^2_{(n-1)}$. Also, as $\bar{Z}_n = g(Y_1)$ and $S_z^2 = h(Y_2, \dots, Y_n)$, we have $\bar{Z}_n \perp S_n^2$.

Step 3. When $X_1, \cdots, X_n$ are iid from $N(\mu, \sigma^2)$. Then consider the transformation $Z_i = \sigma^{-1}(X_i - \mu)$. Then $Z_i \overset{iid}{\sim} N(0,1)$. Further,

$$\frac{S_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n (Z_i - \bar{Z}_n)^2 = S_z^2.$$

Thus, the result follows from step 2. ■

2.2 The F distribution

Definition 4 (F distribution)**.** *Let $X \sim \chi^2_{(m)}$ and $Y \sim \chi^2_{(n)}$ independently. Define $F = \dfrac{X/m}{Y/n}$. Then $F \sim F_{m,n}$, i.e. F distribution with degrees of freedom m and n.*

Properties:

1. If $F \sim F_{m,n}$ then $F^{-1} \sim F_{n,m}$.
2. $\mathbb{E}(F) = \dfrac{n}{n-2}$; $n > 2$ and $\text{Var}(F) = \dfrac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$; $n > 4$.
3. If $F \sim F_{m,n}$ then $\dfrac{mF/n}{1+mF/n} \sim \text{Beta}(\frac{m}{2}, \frac{n}{2})$.

2.3 The t distribution

Definition 5 (t distribution)**.** *Let $X \sim N(0,1)$ and $Y \sim \chi^2_{(m)}$ independently. Then $T = \dfrac{X}{\sqrt{Y/m}} \sim t_{(m)}$; i.e. t distribution with m degrees of freedom.*

Properties:

1. $t_{(1)} \equiv \text{Cauchy}(0,1)$.
   Implication: Let $X, Y \overset{iid}{\sim} N(0,1)$ then $\dfrac{X}{\mid Y \mid} \sim \text{Cauchy}(0,1)$.
2. $\mathbb{E}(T) = 0$; for $m > 1$ and $\text{Var}(T) = \dfrac{m}{m-2}$; for $m > 2$.
3. Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\dfrac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}} \sim t_{(n-1)}$.