Lecture 13: Sampling Distribution of the Sample Mean

Mathematical Statistics I, MATH 60061/70061

Thursday October 21, 2021

Reference: Casella & Berger, 5.2

Mean and variance of the sample mean $\bar{X}$

Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then

1. $E(\bar{X}) = \mu$,
2. $\mathrm{Var}(\bar{X}) = \sigma^2/n$,

We know the mean and variance of the sampling distribution of $\bar{X}$.

Can we say more about the sampling distribution?

MGF of $\bar{X}$

The MGF of $\bar{X}$ is

$$
\begin{aligned}
M_{\bar{X}}(t) &= E(e^{t\bar{X}}) \\
&= E(e^{t(X_1+\cdots+X_n)/n}) \\
&= E(e^{(t/n)X_1})\ldots E(e^{(t/n)X_n}) && [X_i\text{'s are independent}] \\
&= M_{X_1}(t/n)\ldots M_{X_n}(t/n) \\
&= [M_{X_1}(t/n)]^n && [X_i\text{'s are identically distributed}]
\end{aligned}
$$

Mean of Normal random variables

Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$. Then the MGF of the sample mean is

$$\begin{aligned} M_{\bar{X}}(t) &= \left[\exp\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right]^n \\ &= \exp\left(n\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right) \\ &= \exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right). \end{aligned}$$

Thus, $\bar{X}$ has a $\mathcal{N}(\mu, \sigma^2/n)$ distribution.

Mean of Poisson random variables

Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{Pois}(\lambda)$. Then the MGF of the sample mean is

$$\begin{aligned} M_{\bar{X}}(t) &= \left[e^{\lambda(e^{t/n}-1)}\right]^n \\ &= e^{n\lambda(e^{t/n}-1)}, \end{aligned}$$

which is the MGF of $\mathrm{Pois}(n\lambda)$ evaluated at (t/n). Thus, $n\bar{X}$ has a $\mathrm{Pois}(n\lambda)$ distribution.

Recall: If X has MGF $M_X(t)$, then for any constants a and b, the MGF of $a + bX$ is given by $M_{a+bX}(t) = e^{at} M_X(bt)$.

Convolution integrals

Let X and Y be independent random variables with PDFs $f_X(x)$ and $f_Y(y)$. The PDF of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z - x) f_X(x) dx = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy.$$

Convolution integrals

Let X and Y be independent random variables with PDFs $f_X(x)$ and $f_Y(y)$. The PDF of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x) dx = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy.$$

Using the LOTP and conditioning on X:

$$\begin{aligned}
F_Z(z) = P(X+Y \le z) &= \int_{-\infty}^{\infty} P(X+Y \le z \mid X = x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} P(Y \le z - x \mid X = x) f_X(x) dx \\
&= \int_{-\infty}^{\infty} F_Y(z-x) f_X(x) dx.
\end{aligned}$$

Differentiating the CDF t gives

$$f_Z(z) = \int_{-\infty}^{\infty} f_Y(z-x) f_X(x) dx.$$

Sum of Cauchy random variables

Let U and V be independent Cauchy random variables, $U \sim \text{Cauchy}(0, \sigma)$ and $V \sim \text{Cauchy}(0, \tau)$; that is

$$f_U(u) = \frac{1}{\pi\sigma} \frac{1}{1 + (u/\sigma)^2}, \quad f_V(v) = \frac{1}{\pi\tau} \frac{1}{1 + (v/\tau)^2},$$

for $-\infty < u < \infty$, $-\infty < v < \infty$.

Using the convolution formula, the PDF of $Z = U + V$ is given by

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{\pi\sigma} \frac{1}{1 + (u/\sigma)^2} \frac{1}{\pi\tau} \frac{1}{1 + ((z-u)/\tau)^2} du \\ &= \frac{1}{\pi(\sigma + \tau)} \frac{1}{1 + (z/(\sigma + \tau))^2}, \quad -\infty < z < \infty. \end{aligned}$$

Thus, the sum of two independent Cauchy random variables is again a Cauchy, with the scale parameters adding.

Let $Z_1, \ldots, Z_n$ be a random sample from a $\text{Cauchy}(0, 1)$.

- $\sum_{i=1}^{n} Z_i$ is $\text{Cauchy}(0, n)$.
- $\bar{Z}$ is $\text{Cauchy}(0, 1)$.

The dispersion in the distribution of $\bar{Z}$ is the same, regardless of the sample size n.

This is in sharp contrast to the more common situation (when the population has finite variance σ^2), where $\text{Var}(\bar{X}) = \sigma^2/n$ decreases as the sample size increases.

Sample from an exponential family

Suppose that $X_1, \ldots, X_n$ is a random sample from a PDF/PMF $f(x \mid \theta)$, where

$$f(x \mid \boldsymbol{\theta}) = h(x) c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right)$$

is a member of an **exponential family**. Define statistics $T_1, \ldots, T_k$ by

$$T_j(\boldsymbol{X}) = T_j(X_1, \ldots, X_n) = \sum_{i=1}^{n} t_j(X_i), \quad j = 1, \ldots, k.$$

If the set $\{(w_1(\boldsymbol{\theta}), w_2(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ contains an open subset of $\mathbb{R}^k$, then the distribution of $(T_1, \ldots, T_k)$ is an exponential family of the form

$$f_T(u_1, \ldots, u_k \mid \boldsymbol{\theta}) = H(u_1, \ldots, u_k)[c(\boldsymbol{\theta})]^n \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) u_j\right).$$

Proof for the discrete case

The joint PMF of $X_1, \ldots, X_n$ is

$$\begin{aligned}\prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[h(x_i) c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x_i) \right) \right] \\ &= \prod_{i=1}^{n} h(x_i) [c(\boldsymbol{\theta})]^n \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) \sum_{i=1}^{n} t_j(x_i) \right).\end{aligned}$$

Then, the PMF of $(T_1, \ldots, T_k)$ is

$$\begin{aligned}f_T(u_1, \ldots, u_k \mid \boldsymbol{\theta}) = P(T_1 = u_1, \ldots, T_k = u_k) &= \sum_{\boldsymbol{x}: T(\boldsymbol{x}) = \boldsymbol{u}} \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta}) \\ &= \sum_{\boldsymbol{x}: T(\boldsymbol{x}) = \boldsymbol{u}} \prod_{i=1}^{n} h(x_i) [c(\boldsymbol{\theta})]^n \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) \sum_{i=1}^{n} t_j(x_i) \right) \\ &= \left[\sum_{\boldsymbol{x}: T(\boldsymbol{x}) = \boldsymbol{u}} \prod_{i=1}^{n} h(x_i) \right] [c(\boldsymbol{\theta})]^n \exp\left(\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) u_j \right)\end{aligned}$$

Sum of Bernoulli random variables

Suppose $X_1, \ldots, X_n \sim \mathrm{Bern}(p)$. The joint PMF is

$$\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = \prod_{i=1}^{n} \left[(1-p)\exp\left(x_i \log \frac{p}{1-p}\right)\right]$$
$$= (1-p)^n \exp\left(\log \frac{p}{1-p} \sum_{i=1}^{n} x_i\right)$$

$\mathrm{Bern}(p)$ is a member of an exponential family with $h(x) = 1$, $c(p) = (1-p)$, $w_1(p) = \log(p/(1-p))$, and $t_1(x) = x$.

The statistic $T_1(X_1, \ldots, X_n) = X_1 + \cdots + X_n$ has PMF

$$P(T_1 = u_1) = (1-p)^n \exp\left(\log \frac{p}{1-p} \cdot u_1\right)$$

This is the PMF of $\mathrm{Bin}(n, p)$.