Sums of Random Variables from a Random Sample

<u>Definition 5.2.1</u>

Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a statistic. The probability distribution of a statistic Y is called the sampling distribution of Y.

The definition of a statistic is very broad, with the only restriction being that a statistic cannot be a function of a parameter. Three statistics that are often used and provide good summaries of the sample are now defined.

<u>Definition</u>

The sample mean is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

<u>Definition</u>

The sample variance is the statistic defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The sample standard deviation is the statistic defined by $S = \sqrt{S^2}$.

The sample variance and standard deviation are measures of variability in the sample that are related to the population variance and standard deviation.

<u>Theorem 5.2.4</u>

Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. then

(a) $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

(b) $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Lemma 5.2.5

Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $Eg(X_1)$ and $\mathrm{Var}g(X_1)$ exists. Then

$$E\big(\sum_{i=1}^{n} g(X_i)\big) = n(Eg(X_1)).$$

and

$$\mathrm{Var}\big(\sum_{i=1}^{n} g(X_i)\big) = n(\mathrm{Var}g(X_1)).$$

THeorem 5.2.6

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

(a) $E\bar{X} = \mu$.

(b) $\mathrm{Var}\bar{X} = \frac{\sigma^2}{n}$.

(c) $ES^2 = \sigma^2$.

PROOF: We just prove part (c) here.

$$\begin{aligned} ES^2 &= E\big(\frac{1}{n-1}[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2]\big) \\ &= \frac{1}{n-1}(nEX_1^2 - nE\bar{X}^2) \\ &= \frac{1}{n-1}(n(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)) = \sigma^2. \end{aligned}$$

□

About the distribution of a statistic, we have the following theorems. Theorem 5.2.7

Let $X_1, \ldots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is

$$M_{\bar{X}}(t) = [M_X(t/n)]^n.$$

Example (Distribution of the mean)

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

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

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

The mgf of the sample mean a gamma(α, β) random sample is

$$M_{\bar{X}}(t) = [(\frac{1}{1-\beta(t/n)})^{\alpha}]^n = \Big(\frac{1}{1-(\beta/n)t}\Big)^{n\alpha},$$

which we recognize as the mgf of a gamma$(n\alpha, \beta/n)$, the distribution of $\bar{X}$.

If Theorem 5.2.7 is not applicable, because either the resulting mgf of $\bar{X}$ is unrecognizable or the population mgf does not exists. In such cases, the following convolution formula is useful.

<ins>Theorem 5.2.9</ins>

If X and Y are independent continuous random variables with pdfs $f_X(x)$ and $f_Y(y)$, then the pdf of $Z = X + Y$ is

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

Proof: Let $W = X$. The Jacobian of the transformation from (X, Y) to (Z, W) is 1. So the joint pdf of (Z, W) is

$$f_{Z,W}(z, w) = f_{X,Y}(w, z-w) = f_X(w) f_Y(z-w).$$

Integrating out w, we obtain the marginal pdf of Z and finish the proof. $\square$

<ins>Example (Sum of Cauchy random variables)</ins>

As an example of a situation where the mgf technique fails, consider sampling from a Cauchy distribution. Let U and V be independent Cauchy random variables, $U \sim Cauchy(0, \sigma)$ and $V \sim Cauchy(0, \tau)$; that is,

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

where $-\infty < U, V < \infty$. Based on the convolution formula, the pdf of $U + V$ is given by

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{\pi\sigma} \frac{1}{1 + (w/\sigma)^2} \frac{1}{\pi\sigma} \frac{1}{1 + ((z - w)/\tau)^2} dw, \\ &= \frac{1}{\pi(\sigma + \tau)} \frac{1}{1 + (z/(\sigma + \tau))^2}, \end{aligned}$$

where $-\infty < z < \infty$. Thus, the sum of two independent Cauchy random variables is again a Cauchy, with the scale parameters adding. It therefore follows that if $Z_1, \dots, Z_n$ are iid Cauchy(0,1) random variables, then $\sum Z_i$ is Cauchy$(0, n)$ and also $\bar{Z}$ is Cauchy(0,1). The sample mean has the same distribution as the individual observations.

Theorem 5.2.11

Suppose $X_1, \dots, X_n$ is a random sample from a pdf or pmf $f(x|\theta)$, where

$$f(x|\theta) = h(x)c(\theta) \exp(\sum_{i=1}^{k} w_i(\theta)t_i(x))$$

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

$$T_i(X_1, \dots, X_n) = \sum_{j=1}^{n} t_i(X_j), \quad i = 1, \dots, k.$$

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

$$f_T(u_1, \dots, u_k|\theta) = H(u_1, \dots, u_k)[c(\theta)]^n \exp(\sum_{i=1}^{k} w_i(\theta)u_i).$$

The open set condition eliminates a density such as the $N(\theta, \theta^2)$ and, in general, eliminates curved exponential families from Theorem 5.2.11.