556: MATHEMATICAL STATISTICS I

GENERAL RESULTS FOR THE SAMPLE MEAN AND VARIANCE STATISTICS

THEOREM
Suppose that $X_1, ..., X_n$ is a random sample from a distribution, say with finite expectation μ and variance σ^2. Consider the sample mean and sample variance statistics $\overline{X}$ and s^2 and denote

$$T_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad\qquad T_2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Then

(a) $\quad E_{f_{T_1}}[T_1] = \mu$

(b) $\quad Var_{f_{T_1}}[T_1] = \dfrac{\sigma^2}{n}$

(c) $\quad E_{f_{T_2}}[T_2] = \sigma^2$

Proof. (a) and (b) follow from elementary properties of expectations and variances for independent random variables. For (c), note that

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2.$$

Hence

$$\begin{aligned}
E_{f_{T_2}}[T_2] &= \frac{1}{n-1} E_{f_{\underset{\sim}{X}}}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} E_{f_{\underset{\sim}{X}}}[X_i^2] - nE_{f_{\underset{\sim}{X}}}[\overline{X}]^2\right] \\
&= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right] \qquad (1) \\
&= \sigma^2
\end{aligned}$$

where line (1) follows from the fact that for any random variable X

$$\sigma^2 = E_{f_X}[X^2] - E_{f_X}[X]^2 = E_{f_X}[X^2] - \mu^2$$

and the result of parts (a) and (b).
∎

SAMPLING FROM A NORMAL FAMILY

Recall the fundamental transformation results for Normal random variables:

(i) If $X \sim N(0,1)$, then

$$X^2 \sim \chi_1^2 \equiv Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$$

(ii) If $X_1, \ldots, X_r \sim N(0,1)$ are independent random variables, then

$$Y = \sum_{i=1}^{r} X_i^2 \sim \chi_r^2 \equiv Gamma\left(\frac{r}{2}, \frac{1}{2}\right)$$

(iii) If $Y_1 \sim \chi_{r_1}^2$ and $Y_2 \sim \chi_{r_2}^2$ are independent random variables, then

$$Y = Y_1 + Y_2 \sim \chi_{r_1+r_2}^2$$

THEOREM

Suppose that $X_1, ..., X_n$ is a random sample from a normal distribution, say $X_i \sim N(\mu, \sigma^2)$. Define the sample mean and sample variance statistics $\overline{X}$ and s^2 as the random variables

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad\qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

Then

(a) $\overline{X} \sim N(\mu, \sigma^2/n)$

(b) $\overline{X}$ is independent of $\left\{X_i - \overline{X}, i = 1, ..., n\right\}$, and $\overline{X}$ and s^2 are independent random variables

(c) The random variable

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

has a **chi-squared distribution** with $n-1$ degrees of freedom.

Proof. (a) Proof straightforward using mgfs.

(b) The joint pdf $X_1, ..., X_n$ is the normal density

$$f_{X_1,...,X_n}(x_1, ..., x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}$$

Consider the multivariate transformation to $Y_1, ..., Y_n$ where

$$\left.\begin{aligned} Y_1 &= \overline{X} \\ Y_i &= X_i - \overline{X},\ i = 2, ..., n \end{aligned}\right\} \iff \left\{\begin{aligned} X_1 &= Y_1 - \sum_{i=2}^{n} Y_i \\ X_i &= Y_i + Y_1,\quad i = 2, ..., n \end{aligned}\right.$$

Thus $\underset{\sim}{Y} = A\underset{\sim}{X}$, or equivalently $\underset{\sim}{X} = A^{-1}\underset{\sim}{Y}$, where A is the $n \times n$ matrix with (i,j)th element

$$[A]_{ij} = \begin{cases} 1 - \frac{1}{n} & i = j \ \text{ and } i \neq 1, \\ \frac{1}{n} & i = 1 \\ -\frac{1}{n} & \text{otherwise} \end{cases}$$

that is, we have a linear transformation. Note that

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\mu)^2=\sum_{i=1}^{n}(x_i-\overline{x}+\overline{x}-\mu)^2 &= \sum_{i=1}^{n}\left[(x_i-\overline{x})^2+2(x_i-\overline{x})(\overline{x}-\mu)+(\overline{x}-\mu)^2\right]\\ &= \sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2\end{aligned}$$

where $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ is the observed sample mean. Thus the joint pdf of $X_1,...,X_n$ takes the form

$$f_{X_1,..,X_n}(x_1,..,x_n)=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-\mu)^2\right]\right\}.$$

Now

$$x_1-\overline{x}=-\sum_{i=2}^{n}(x_i-\overline{x})=-\sum_{i=2}^{n}y_i$$

and so

$$\sum_{i=1}^{n}(x_i-\overline{x})^2=(x_1-\overline{x})^2+\sum_{i=2}^{n}(x_i-\overline{x})^2=\left(-\sum_{i=2}^{n}y_i\right)^2+\sum_{i=2}^{n}y_i^2$$

The Jacobian of the transformation is n, so the joint density of $Y_1,...,Y_n$ is given by the multivariate transformation theorem as

$$\begin{aligned}f_{Y_1,..,Y_n}(y_1,..,y_n) &= n\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\left(-\sum_{i=2}^{n}y_i\right)^2+\sum_{i=2}^{n}y_i^2+n(y_1-\mu)^2\right]\right\}\\ &= n\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\left(-\sum_{i=2}^{n}y_i\right)^2+\sum_{i=2}^{n}y_i^2\right]\right\}\times\exp\left\{-\frac{n}{2\sigma^2}(y_1-\mu)^2\right\}\end{aligned}$$

Hence

$$f_{Y_1,..,Y_n}(y_1,..,y_n) = f_{Y_2,..,Y_n}(y_2,..,y_n)f_{Y_1}(y_1)$$

and therefore Y_1 is independent of $Y_2,...,Y_n$. Hence $\overline{X}$ is **independent** of the random variables terms $\left\{Y_i=X_i-\overline{X}, i=2,...,n\right\}$. Finally, $\overline{X}$ is also independent of $X_1-\overline{X}$ as

$$X_1-\overline{X}=-\sum_{i=2}^{n}\left(X_i-\overline{X}\right)$$

and s^2 is a function only of $\left\{X_i-\overline{X},\ i=1,...,n\right\}$. As $\overline{X}$ is independent of these variables, $\overline{X}$ and s^2 are also independent.

(c) The random variables that appear as sums of squares terms that joint pdf are

$$\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2}+\frac{n\left(\overline{X}-\mu\right)^2}{\sigma^2}$$

or $V_1 = V_2 + V_3$, say. Now, $X_i \sim N(\mu, \sigma^2)$, so therefore

$$\frac{(X_i-\mu)^2}{\sigma^2} \sim N(0,1) \quad \Longrightarrow \quad \frac{(X_i-\mu)^2}{\sigma^2} \sim \chi_1^2 \equiv Ga\left(\frac{1}{2},\frac{1}{2}\right) \quad \Longrightarrow \quad V_1 = \frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \sim \chi_n^2$$

as the X_is are independent, and the sum of n independent $Ga(1/2,1/2)$ variables has a $Ga(n/2,1/2)$ distribution. Similarly, as $\overline{X} \sim N(\mu, \sigma^2/n)$, $V_3 \sim \chi_1^2$ By part (b), V_2 and V_3 are independent, and so the mgfs of V_1, V_2 and V_3 are related by

$$M_{V_1}(t) = M_{V_2}(t)M_{V_3}(t) \Longrightarrow M_{V_2}(t) = \frac{M_{V_1}(t)}{M_{V_3}(t)}$$

As V_1 and V_3 are Gamma random variables, M_{V_1} and M_{V_3} are given by

$$M_{V_1}(t) = \left(\frac{1/2}{1/2-t}\right)^{n/2} \quad \text{and} \quad M_{V_3}(t) = \left(\frac{1/2}{1/2-t}\right)^{1/2}.$$

So therefore

$$M_{V_2}(t) = \left(\frac{1/2}{1/2-t}\right)^{(n-1)/2}$$

which is also the mgf of a Gamma random variable, and hence

$$V_2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

and the result follows. ∎

Alternative inductive proof of (c): Let $\overline{X}_k$ and s_k^2, $k = 1, 2, \ldots, n$ denote the sample mean and sample variance random variables derived from the first k variables. Now, for $k \geq 2$, it can be shown after some manipulation that

$$(k-1)s_k^2 = (k-2)s_{k-1}^2 + \left(\frac{k-1}{k}\right)(X_k - \overline{X}_{k-1})^2 \tag{2}$$

For $k = 2$

$$(2-1)s_2^2 = \frac{1}{2}(X_2 - X_1)^2 = \left(\frac{X_2 - X_1}{\sqrt{2}}\right)^2 = Z^2$$

say, where $Z \sim N(0,1)$. Thus $s_2^2 \sim \chi_1^2$. Now for the inductive hypothesis, presume that

$$(k-1)s_k^2 \sim \chi_{k-1}^2$$

so that, using the identity in (2),

$$ks_{k+1}^2 = (k-1)s_k^2 + \left(\frac{k}{k+1}\right)(X_{k+1} - \overline{X}_k)^2$$

The two terms on the right hand side are independent (using the result in (b)); the first term is χ_{k-1}^2 distributed, the second term is χ_1^2 distributed, so ks_{k+1}^2 is χ_k^2 distributed and the inductive argument is completed.