M4056 Lecture Notes. **September 8, 2010**

Sampling from a normal distribution II

Before getting to the meat of this lecture, we provide some orientation. Bear in mind the kinds of inferential tasks we might face:

i) We are sampling from a population known to be normal and with known variance, seeking to determine the mean. The previous lecture contained all the information we need make the optimal guess about the mean and to compute confidence intervals. We will examine the practical details later.

ii) We are sampling from a population known to be normal but with unknown variance, seeking to determine the mean. This is a much more common situation than *i*). In this case, we can still make an optimal guess about the mean, but we cannot find confidence intervals without information incorporating some estimate of the variance. The present lecture begins to address the mathematical theory that supports this.

iii) We are sampling from a population not known to be normal. The theory we are presenting now has no bearing on this.

Throughout today's lecture, we will be dealing with a sample $X_1, \ldots, X_n$ of size n drawn from an $n(\mu, \sigma^2)$ distribution. $\overline{X}$ denotes the sample mean and S^2 denotes the sample variance. Last Friday, we saw that $\overline{X}$ would be $n(\mu, \sigma^2/n)$. Today, we will show:

Fact 1. *$\overline{X}$ and S^2 are independent. (5.3.1.a)*

Fact 2. *$(n-1)S^2/\sigma^2$ has the same distribution as the sum of the squares of $n-1$ independent random variables, each with a $n(0,1)$ distribution. (5.3.1.c)*

Before proving Fact 1, we shall review some facts about independent random variables. Suppose X and Y are random variables—possibly vector-valued, e.g., $X = (X_1, \ldots, X_m)$, $Y = (Y_1, \ldots, Y_n)$. Then, X and Y are said to be *independent* if any one (hence all) of the following equivalent conditions is satisfied:

i) the joint *pdf* $f_{X,Y}(x,y)$ factors as $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$,

ii) the joint *cdf* $F_{X,Y}(x,y)$ factors as $F_{X,Y}(x,y) = F_X(x) \cdot F_Y(y)$,

iii) for any measurable sets $A \subseteq \mathbb{R}^m$ and $B \subseteq \mathbb{R}^n$,

$$P(X \in A \;\&\; Y \in B) = P(X \in A) \cdot P(Y \in B).$$

Lemma. *If X and Y are independent random variables and g and h be continuous scalar functions of X and Y, respectively, then $U = g(X)$ and $V = h(Y)$ are independent.*

Proof. Let $A_u = \{\, x \in \mathbb{R}^m \mid g(x) \leq u \,\}$. Then, $g(X) \leq u \Leftrightarrow X \in A_u$, so $P(g(X) \leq u) = P(X \in A_u)$. This trick broadens the applicability of condition *iii*). We use it for both X and Y:

$$\begin{aligned} F_{U,V}(u,v) &= P(U \leq u \;\&\; V \leq v) \\ &= P(g(X) \leq u \;\&\; h(Y) \leq v) \\ &= P(g(X) \leq u) \cdot P(h(Y) \leq v) \\ &= F_U(u) \cdot F_V(v). \end{aligned}$$

Proof of Fact 1. Note that

$$X_1 - \overline{X} = -\sum_{i=2}^{n} X_i - \overline{X}, \qquad (*)$$

since $\sum_{i=1}^{n} X_i - \overline{X} = 0$. Let $Y_1 = \overline{X}$ and let $Y_i := X_i - \overline{X}$ for $i = 2, 3, \ldots, n$. Then by $(*)$, S^2 is a function $Y_2, \ldots, Y_n$. Thus, it suffices to show that Y_1 is independent of $\vec{Y} = (Y_2, \ldots, Y_n)$. Now, $X_1 = Y_1 - \sum_{i=2}^{n} Y_i$, and $X_i = Y_i + Y_1$. We see that $\vec{X}$ is a linear function of $\vec{Y}$, and $\left|\frac{\partial X}{\partial Y}\right| = n$. Viewing $\vec{x}$ as a function of $\vec{y}$ by the formulae just stated in upper-case letters,

$$\begin{aligned} f_{\vec{Y}}(\vec{y}) &= n \cdot f_{\vec{X}}(\vec{x}) \\ &= n \cdot \frac{1}{(2\pi)^{n/2}} \cdot \exp\left(-(1/2)\sum_{i=1}^{n} x_i^2\right) \end{aligned}$$

The sum appearing in the argument of the exponential function can be rewritten as follows (see homework, below):

$$\sum_{i=1}^{n} x_i^2 = n y_1^2 + \left(\sum_{i=2}^{n} y_i\right)^2 + \sum_{i=2}^{n} y_i^2. \qquad (**)$$

Since this is a sum of a function of y_1 and a function of $y_2, \ldots, y_n$, after applying the exponential function, the result factors as a function of y_1 alone times a function of $y_2, \ldots, y_n$.

We now turn to Fact 2. The kind of distribution mentioned there is called "chi squared with $n-1$ degrees of freedom." We will investigate the properties of such a distribution later, and for now the name is not important. (As Feynman says, if we know the name of a bird in all the languages of the world, we know something about language but we don't know anything about the bird; see http://www.youtube.com/watch?v=0XgmrMZ0h54, minute 6:15.)

Proof of Fact 2. The proof is by induction on the sample size. In case $n = 2$, you will show in a homework problem that S_2^2/σ^2 is distributed as the square of a standard normal variable. To carry out the rest of the proof, set $\overline{X}_k$ and S_k^2 be the statistics for a sample of size k. Assume Fact 2 is known for samples of size $n = k$. We shall show that it is true of samples of size $n = k + 1$. We will treat the case where $\mu = 0$ and $\sigma^2 = 1$, and in your homework you will generalize. In your homework, you will verify that with $\mu = 0$ and $\sigma = 1$:

$$k S_{k+1}^2 = (k-1) S_k^2 + \frac{k}{k+1}(X_{k+1} - \overline{X}_k)^2 \qquad (***)$$

Now, S_k^2 is independent of $\overline{X}_k$ (by Fact 1) and independent X_{k+1} (since all the X_i are independent). Thus, S_k^2 is independent of $\frac{k}{k+1}(X_{k+1} - \overline{X}_k)^2$. Moreover, $X_{k+1} - \overline{X}_k$ is a sum of a $n(0,1)$ and a $n(0, 1/k)$ variable, so it is normal with variance $(k+1)/k$, and therefore $\sqrt{\frac{k}{k+1}}(X_{k+1} - \overline{X}_k)$ is $n(0,1)$. By the inductive hypothesis, $(k-1)S_k^2$ is distributed as a sum of the squares of $k-1$ independent $n(0,1)$ variables. Thus, $k S_{k+1}^2$ is distributed as a sum of the squares of k independent $n(0,1)$ variables, and Fact 2 is proved.

Homework

1. Prove the equality $(*)$.
2. Prove the equality $(**)$ by writing the x_is in terms of the y_is, expanding, canceling and collecting terms.
3. Prove the $n = 2$ case of Fact 2. Here is a sketch. Note that

$$S_2^2 = (X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 = (X_1 - X_2)^2/2 = ((X_1 - \mu) - (X_2 - \mu))^2/2.$$

   Now $X_2 - \mu$ is symmetric, so S^2 is half the square of the sum of two $n(0, \sigma^2)$ variables.
4. Prove the equality $(***)$. One way to do this is to use Theorem 5.2.4.b.
5. Deduce the general case of Fact 2 from the case we have proved (with $\mu = 0$, $\sigma^2 = 1$).