Sampling Distributions

Guy Lebanon

February 14, 2006

In this note we study the distributions of functions of iid samples $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The assumed scenario is that the samples are given to us, but we don’t know $\mu, \sigma$. We need to construct statistics that are functions of the provided samples and which will provide estimates for the unknown quantities. The statistic $T(X_1, \ldots, X_n)$ is a RV since it is a function of RVs. Uncovering its distribution is the first step in evaluation of the estimation the statistic provides. We will describe statistics that correspond to RVs with distributions $\chi^2, t$ and $F$. These distributions are important if the iid samples $X_1, \ldots, X_n$ are normally distributed. If they are not, we can still use the above distributions as approximations through the central limit theorem.

The most common estimator for the expected value $\mu$ is the empirical mean $\overline{X} = n^{-1} \sum_{i=1}^n X_i$ which is normally distributed $\overline{X} \sim N(\mu, \sigma^2/n)$. We can prove this by using the result that a linear combination of normal RVs is a normal RV (see the note on the moment generating function). The same result also proves that the standardized variables $(X_i - \mu)/\sigma \sim N(0, 1)$.

**Definition 1.** The Chi-squared distribution with $n$ degrees of freedom (dof) $\chi_n^2$ is a Gamma distribution with parameters $\alpha = n/2, \beta = 1/2$, with mgf $(1 - 2t)^{-n/2}$.

From the mgf (or from the formula for expectation and variance of a Gamma distribution), it is clear that $\chi_n^2$ has expectation $n$ and variance $2n$. The main use of the $\chi_n^2$ distribution is due to the following fact.

**Theorem 1.** If $Z_1, \ldots, Z_n$ are iid $N(0, 1)$ then $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$.

**Proof.** Consider first the case $n = 1$. The mgf of $Z_1^2$ is

$$E(e^{Z_1^2}) = \int_{-\infty}^{+\infty} e^{t^2} (2\pi)^{-1/2} e^{-t^2/2} dt = \int_{-\infty}^{+\infty} (2\pi)^{-1/2} e^{-(1-2t)^2/2} dt = \frac{1}{(1-2t)^{1/2}} \int_{-\infty}^{+\infty} \frac{e^{-z^2/2}}{(2\sqrt{\pi})^{-1}} dz = \frac{1}{(1-2t)^{1/2}} \cdot 1$$

which is the mgf of $\chi_1^2$. In the case $n > 1$, the mgf of the sum is the product of the mgfs (again, see mgf note) resulting in the $\chi_n^2$ mgf $\prod_{i=1}^n (1 - 2t)^{-1/2} = (1 - 2t)^{-n/2}$. \hfill $\square$

The most common estimator for the variance of a sample is $S^2(X_1, \ldots, X_n) = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2$. We will just write $S^2$, but remember that it is a function of the sample. The reason for the $n-1$ in the denominator and not $n$ will become clear later on.

**Theorem 2.** Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ and $S^2$ be defined as above. Then $\frac{S^2}{(n-1)\sigma^2} = \sum_{i=1}^n X_i^2/\sigma^2 \sim \chi_{n-1}^2$. Furthermore, $\overline{X}, S$ are independent RVs.

**Proof.** (from Degroot and Schervish) We first prove the result if $X_1, \ldots, X_n \sim N(0, 1)$. Consider the unit-norm vector $u = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$ and an orthonormal matrix $A$ built from $u$ using the Gram Schmidt procedure. Specifically, we look at $Y = AX$. $Y_1 = u^T X = \overline{X} \cdot X$. Since $A$ is an orthonormal matrix it preserves the norm and therefore $\sum_{i=1}^n X_i^2 = \sum_{i=1}^n Y_i^2$ and

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=1}^n X_i^2 - Y_1^2 = \sum_{i=1}^n X_i^2 - n\overline{X}^2 = \sum_{i=1}^n (X_i - \overline{X})^2.$$
The joint pdf of $Y_1, \ldots, Y_n$ is precisely the same as that of $X_1, \ldots, X_n$ since

$$ f_{Y_1, \ldots, Y_n}(y) = \frac{1}{|\text{det} A|} (2\pi)^{n/2} e^{-\frac{1}{2} \sum_i |A^{-1} y|^2} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2} \sum_i y_i^2}. $$

So, $Y_1, \ldots, Y_n$ are independent, and by the previous result, $\sum (X_i - \overline{X})^2$ and $\sqrt{n} \cdot \overline{X}$ are independent. Since $\overline{X}$ and $S^2$ are functions of independent RVs they are independent as well. Furthermore, $\sum (X_i - \overline{X})^2$ is shown to be a sum of squares of $n - 1$ iid standard normal RVs and so its distribution is $\chi^2_{n-1}$.

Now, assume $X_i$ are distributed normal, but not standard normal. From the above, it follows that the result holds for the standardized RVs $Z_i = (X_i - \mu)/\sigma$. However, $\sum_{i=1}^n (Z_i - \overline{Z})^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \overline{X})^2$ proving the result in the general case.

The problem with the above result is that we don’t know $\sigma^2$ and so we can’t use it in a statistic. If we replace $\sigma^2$ with $S^2$ the $\chi^2_{n-1}$ distribution turns into a $t$-distribution with $n - 1$ dof.

**Definition 2.** Let $Z \sim N(0,1), W \sim \chi^2_{\nu}$ be two independent RVs. Then the distribution of $\frac{Z}{\sqrt{W/\nu}}$ is known as a $t$-distribution with $\nu$ degrees of freedom, denoted $t_\nu$.

Using the above notations we have

$$ \sqrt{n} \left( \frac{\overline{Y} - \mu}{S} \right) = \frac{\sqrt{n}(\overline{Y} - \mu)/\sigma}{\sqrt{((n-1)S^2/\sigma^2)/(n-1)}} = \frac{Z}{\sqrt{W/(n-1)}} \sim t_{n-1}. $$

Finally, assume we have two populations $X_1, \ldots, X_n \sim N(\mu, \sigma^2), Y_1, \ldots, Y_m \sim N(\eta, \tau^2)$ and we are interested in comparing $\sigma^2$ to $\tau^2$ or looking at the magnitude of $\sigma^2/\tau^2$. This leads to the statistic that is the ratio of the two variance estimates $S_1^2/S_2^2$ which leads to the $F$-distribution.

**Definition 3.** Let $W_1 \sim \chi^2_p, W_2 \sim \chi^2_q$ be two independent RVs. Then $\frac{W_1}{W_2}$ has a distribution known as the $F$ distribution with $(p, q)$ dof denoted $F_{p,q}$.

We have

$$ \frac{S_1^2/\sigma^2}{S_2^2/\tau^2} = \frac{((n-1)S_1^2/\sigma^2)/(n-1)}{((m-1)S_2^2/\tau^2)/(m-1)} = \frac{W_1/(n-1)}{W_2/(m-1)} \sim F_{n-1,m-1}. $$