# Confidence Intervals

In many applications one is given data

$$X_1, X_2, \ldots X_n$$

with the a priori knowledge that the data are drawn from a normally distributed population. However, the theoretical mean $\mu = E(X_j)$ and the theoretical variance $\sigma^2 = \mathrm{Var}(X_j)$ are unknown and must be estimated from the data. (In many real life applications the knowledge that the population is normally distributed is also unknown. But we take as a working hypothesis that the underlying data are normally distributed.)

The obvious estimate for the mean $\mu$ is the average of the data

$$\mu_n = \frac{1}{n} \sum_{j=1}^{n} X_j \tag{1}$$

This is an unbiased estimator, the *sample mean*, in the sense that $E(\mu_n) = \mu$. Furthermore, we know from the strong law of large numbers that

$$\mu_n \to \mu, \; n \to \infty \text{ a.s.}$$

That is, as the sample size $n$ tends to infinity our estimate $\mu_n$ converges to the theoretical mean $\mu$. Similarly, we construct the *sample variance* $S_n^2$ by

$$S_n^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \mu_n)^2 . \tag{2}$$

Note that in the sample variance we use the sample mean (since we don't know the theoretical mean $\mu$). Also we divide by $n-1$ instead of $n$. This is done so that the expected value of the sample variance is $\sigma^2$ (this forces $n-1$ in the denominator rather than $n$).[1]

For small sample sizes it is important to give an interval which we can say with a given probability that the theoretical mean $\mu$ lies inside. That is, we want to be able to say based solely upon the data that the theoretical mean $\mu$ satisfies the inequalities $a \leq \mu \leq b$ with probability $1 - \alpha$. (We assign the value of $\alpha$. To say it lies in the interval with 95% confidence we would take

[1] See if you can convince yourself of this fact.

$\alpha = 0.05$.) Thus we are looking for a statistic that constructs these intervals. Let

$$t_n = \frac{\mu_n - \mu}{\frac{S_n}{\sqrt{n}}} = \frac{\sqrt{n}(\mu_n - \mu)}{S_n} \tag{3}$$

The claim is that the distribution of $t_n$ is equal to the distribution the Student $T$-statistic with $n-1$ degrees of freedom. (Note the change from $n \to n-1$.) The Student $T$-statistic of $n$ degrees of freedom has density

$$f_T(t) = c_n \left(1 + t^2/n\right)^{-(n+1)/2} \tag{4}$$

where the normalization constant $c_n$ is given in terms of the $\Gamma$ function

$$c_n = \frac{1}{\sqrt{\pi n}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)}\,.$$

Let $X$ be a normally distributed random variable with mean 0 and variance 1. We compute the moment generating function of $X^2$. Recall that for any random variable $Y$ the moment generating function is

$$M_Y(t) = E(\mathrm{e}^{tY}).$$

Thus

$$\begin{aligned} M_{X^2}(t) &= E(\mathrm{e}^{tX^2}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{tx^2}\, \mathrm{e}^{-x^2/2}\, dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \mathrm{e}^{-x^2(1-2t)/2}\, dx \\ &= \frac{1}{\sqrt{1-2t}}\,. \end{aligned} \tag{5}$$

Now let $X_1, X_2, \ldots, X_n$ be independent $N(0,1)$ random variables and define

$$\chi^2 = X_1^2 + X_2^2 + \cdots + X_n^2.$$

This random variable is called *chi-squared* with $n$ degrees of freedom. The moment generating function for $\chi^2$ is

$$M_{\chi^2}(t) = E(\mathrm{e}^{t\chi^2})$$

$$
\begin{aligned}
&= E\left(\mathrm{e}^{t(X_1^2+\cdots+X_n^2)}\right) \\
&= E\left(\mathrm{e}^{tX_1^2}\right) E\left(\mathrm{e}^{tX_2^2}\right) \cdots E\left(\mathrm{e}^{tX_n^2}\right), \text{ by independence} \\
&= E\left(\mathrm{e}^{tX_1^2}\right)^n, \text{since } X_j \text{ are identically distributed} \\
&= \frac{1}{(1-2t)^{n/2}} \text{ by (5).} \qquad (6)
\end{aligned}
$$

We now wish to find a density whose moment generating function is given by (6).[2] Recall that the *gamma density* with parameters $\alpha$ and $\beta$ is zero for $x < 0$ and for $x > 0$ is by

$$
f_{\alpha,\beta}(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\, x^{\alpha-1}\mathrm{e}^{-x/\beta},\ x > 0, \alpha > 0, \beta > 0. \qquad (7)
$$

We showed earlier that it has mean $\alpha\beta$ and variance $\alpha\beta^2$. The moment generating function is

$$
\begin{aligned}
M_{\text{gamma}}(t) &= \int_0^\infty \mathrm{e}^{tx} f_{\alpha,\beta}(x)\, dx \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1}\mathrm{e}^{-(1/\beta-t)x}\, dx \\
&\qquad \text{make the substitution } u = (1/\beta - t)x, \\
&= \frac{1}{(1-\beta t)^\alpha} \frac{1}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1}\mathrm{e}^{-u}\, du \\
&= \frac{1}{(1-\beta t)^\alpha} \qquad (8)
\end{aligned}
$$

Comparing the moment generating function (6) with the moment generating function (8), we conclude that the distribution of $\chi^2$ with $n$ degrees of freedom has gamma density (7) with parameters $\alpha = n/2$ and $\beta = 2$.

We need one further preliminary result. If $X$ and $Y$ are independent random variables, $Y > 0$, with densities $f_X$ and $f_Y$, respectively, then the density of the random variable

$$
Z = \frac{X}{Y}
$$

[2] In an advanced course you will derive a formula that gives the *inverse* of a moment generating function. That is we have a formula that goes from $f_X \to M_X$; namely, the definition of $M_X$. What we also need is a formula for $M_X \to f_X$. Here we will have to just verify that the gamma density gives the desired result.

is

$$f_Z(z) = \int_0^\infty f_X(zy) y f_Y(y)\, dy \tag{9}$$

(You can most easily prove this by first writing down the distribution function $F_Z(z) = P(Z \leq z)$ in terms of the densities $f_X$ and $f_Y$.)

We are now ready to prove the following theorem

**Theorem:** If $X$, $Y_1$, $Y_2$, $\ldots Y_n$ are independent random variables with common normal density (of mean zero and variance 1), the variable

$$T_n = \frac{X\sqrt{n}}{\sqrt{Y_1^2 + \cdots + Y_n^2}}$$

has density $f_{T_n}$ given by (4).

To prove this we observe that $X\sqrt{n}$ is a normal random variable with mean 0 and variance $n$. The denominator is the square root of the chi-squared random variable. If $f_{\chi^2}$ is the density for $\chi^2$, a simple calculation shows that the density for $\sqrt{\chi^2}$ is

$$f_{\sqrt{\chi^2}}(x) = 2x f_{\chi^2}(x^2).$$

Thus we know the density of the random variable in the numerator and the density of the random variable in the denominator. We can now apply (9) to compute the density of $T_n$. This is straightforward manipulation of integrals. The result if (4).

To apply the student $T$-statistic to (3) we must know that the sample mean (1) is independent of the sample variance (2). As it turns out for *normal populations* the sample mean is independent of the sample variance. To prove this requires the use of the multivariate gaussian distribution. In the process of proving this one finds that that if the sample population is of size $n$, then the $T$-statistic used has $n-1$ degrees of freedom.[3] We will not prove this result about independence.

[3]This is due to the fact that there is the constraint $X_1 + \cdots + X_n = n\mu_n$.