SAMPLE STATISTICS

A random sample of size n from a distribution $f(x)$ is a set of n random variables $x_1, x_2, \ldots, x_n$ which are independently and identically distributed with $x_i \sim f(x)$ for all i. Thus, the joint p.d.f of the random sample is

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_2) = \prod_{i=1}^{n} f(x_i).$$

A statistic is a function of the random variables of the sample, also know as the sample points. Examples are the sample mean $\bar{x} = \sum x_i/n$ and the sample variance $s^2 = \sum(x_i - \bar{x})^2/n$

A random sample may be regarded as a microcosm of the population from which it is drawn. Therefore, we might attempt to estimate the moments of the population's p.d.f $f(x)$ by the corresponding moments of the sample statistics.

To determine the worth of such estimates, we may determine their expected values and their variances. Beyond finding these simple measures, we might endeavour to find distributions of the statistics, which are described as their sampling distributions.

We can show, for example, that the mean $\bar{x}$ of a random sample is an unbiased estimate of the population moment $\mu = E(x)$, since

$$E(\bar{x}) = E\left(\sum \frac{x_i}{n}\right) = \frac{1}{n}\sum E(x_i) = \frac{n}{n}\mu = \mu.$$

Its variance is

$$V(\bar{x}) = V\left(\sum \frac{x_i}{n}\right) = \frac{1}{n^2}\sum V(x_i) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}.$$

Here, we have used the fact that the variance of a sum of independent random variables is the sum of their variances, since the covariances are all zero.

Observe that $V(\bar{x}) \to 0$ as $n \to \infty$. Since $E(\bar{x}) = \mu$, this implies that, as the sample size increases, the estimates become increasingly concentrated around the true population parameters. Such an estimate is said to be consistent

The sample variance, however, does not provide an unbiased estimate of $\sigma^2 = V(x)$, since

$$\begin{aligned}
E(s^2) = E\left\{\frac{1}{n}\sum(x_i - \bar{x})^2\right\} &= E\left[\frac{1}{n}\sum\left\{(x_i - \mu) + (\mu - \bar{x})\right\}^2\right] \\
&= E\left[\frac{1}{n}\sum\left\{(x_i - \mu)^2 + 2(x_i - \mu)(\mu - \bar{x}) + (\mu - \bar{x})^2\right\}\right] \\
&= V(x) - 2E\{(\bar{x} - \mu)^2\} + E\{(\bar{x} - \mu)^2\} = V(x) - V(\bar{x}).
\end{aligned}$$

Here, we have used the result that

$$E\left\{\frac{1}{n}\sum(x_i - \mu)(\mu - \bar{x})\right\} = -E\{(\mu - \bar{x})^2\} = -V(\bar{x}).$$

It follows that

$$E(s^2) = V(x) - V(\bar{x}) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \frac{(n-1)}{n}.$$

Therefore, s^2 is a biased estimator of the population variance and, for an unbiased estimate, we should use

$$\hat{\sigma}^2 = s^2 \frac{n}{n-1} = \frac{\sum (x_i - \bar{x})^2}{n-1}.$$

However, s^2 is still a consistent estimator, since $E(s^2) \to \sigma^2$ as $n \to \infty$ and also $V(s^2) \to 0$.

The value of $V(s^2)$ depends on the form of the underlying population distribution. It would help us to know exactly how the estimates are distributed. For this, we need some assumption about the functional form of the probability distribution of the population. The assumption that the population has a normal distribution is a conventional one, in which case, the following theorem is of assistance:

Theorem. Let $x_1, x_2, \ldots, x_n$ be a random sample from the normal population $N(\mu, \sigma^2)$. Then, $y = \sum a_i x_i$ is normally distributed with $E(y) = \sum a_i E(x_i) = \mu \sum a_i$ and $V(y) = \sum a_i^2 V(x_i) = \sigma^2 \sum a_i^2$.

In general, any linear function of a set of normally distributed variables is itself normally distributed. Thus, for example, if $x_1, x_2, \ldots, x_n$ is a random sample from the normal population $N(\mu, \sigma^2)$, then $\bar{x} \sim N(\mu, \sigma^2/n)$.

The general result is best expressed in terms of matrices. Let $\mu = [\mu_1, \mu_2, \ldots, \mu_n]' = E(x)$ denote the vector of the expected values of the elements of $x = [x_1, x_2, \ldots, x_n]'$ and let $\Sigma = [\sigma_{ij}; i, j = 1, 2, \ldots, n]$ denote the matrix of their variances and covariances. If $a = [a_1, a_2, \ldots, a_n]'$ is a constant vector of order n, then $a'x \sim N(a'\mu, a'\Sigma a)$ is a normally distributed random variable with a mean of

$$E(a'x) = a'\mu = \sum a_i x_i$$

and a variance of

$$V(a'x) = a'\Sigma a = \sum_i \sum_j a_i a_j \sigma_{ij} = \sum_i a_i^2 \sigma_{ii} + \sum_i \sum_{j \neq i} a_i a_j \sigma_{ij}.$$

An important case is when the vector $a = [a_1, a_2, \ldots, a_n]'$ becomes a vector of n units, denoted by $\iota = [1, 1, \ldots, 1]'$ and described as the summation vector. Then, if $x = [x_1, x_2, \ldots, x_n]'$ is the vector of a random sample with $x_i \sim N(\mu, \sigma^2)$ for all i, there is $x \sim N(\mu\iota, \sigma^2 I_n)$, where $\mu\iota = [\mu, \mu, \ldots, \mu]'$ is a vector with μ repeated n times and I_n is an identity matrix of order n. Writing this explicitly, we have

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \right).$$

Then, there is

$$\bar{x} = (\iota'\iota)^{-1}\iota'x = \frac{1}{n}\iota'x \sim N(\mu, \sigma^2/n)$$

and

$$\frac{1}{n}\sigma^2 = \frac{\iota'\{\sigma^2 I\}\iota}{n^2} = \frac{\sigma^2\iota'\iota}{n^2} = \frac{\sigma^2}{n},$$

where we have used repeatedly the result that $\iota'\iota = n$.

Even if we do not know the form of the distribution from which the sample has been taken, we can still say that, under very general conditions, the distribution of $\bar{x}$ tends to normality as $n \to \infty$. Thus we have

Theorem. *The Central Limit Theorem* states that, if $x_1, x_2, \ldots, x_n$ is a random sample from a distribution with mean μ and variance σ^2, then the distribution of $\bar{x}$ tends to the normal distribution $N(\mu, \sigma^2/n)$ as $n \to \infty$. Equivalently, $(\bar{x} - \mu)/(\sigma/\sqrt{n})$ tends in distribution to the standard normal $N(0, 1)$ distribution.

To describe the distribution of the sample variance, we need to define the chi-square distribution. If $x \sim N(0, 1)$ is distributed as a standard normal variable, then $x^2 \sim \chi^2(1)$ is distributed as a chi-square variate with one degree of freedom. Moreover

Theorem. The sum of two independent chi-square variates is a chi-square variate with degrees of freedom equal to the sum of the degrees of freedom of its additive components. In particular, if $x \sim \chi^2(n)$ and $y \sim \chi^2(m)$, then $(x + y) \sim \chi^2(n + M)$.

It follows that, if $x_1, x_2, \ldots, x_n$ is a random sample from a standard normal $N(0, 1)$ distribution, then $\sum x_i^2 \sim \chi^2(n)$. Moreover, if $x_1, x_2, \ldots, x_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution, then $\sum(x_i - \mu)^2/\sigma^2 \sim \chi^2(n)$.

Consider the identity

$$\begin{aligned}\sum(x_i - \mu)^2 &= \sum\left(\{x_i - \bar{x}\} + \{\bar{x} - \mu\}\right)^2 \\ &= \sum\{x_i - \bar{x}\}^2 + n\{\bar{x} - \mu\}^2,\end{aligned}$$

which follows from the fact that the cross product term is $\{\bar{x} - \mu\}\sum\{x_i - \bar{x}\} = 0$. This decomposition of a sum of squares features in the following result:

The Decomposition of a Chi-square statistic. If $x_1, x_2, \ldots, x_n$ is a random sample from a standard normal $N(\mu, \sigma^2)$ distribution, then

$$\sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\sigma^2} = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{\sigma^2} + n\frac{(\bar{x} - \mu)^2}{\sigma^2},$$

with

$$(1) \qquad \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{\sigma^2} \sim \chi^2(n),$$

$$(2) \qquad \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n - 1),$$

$$(3) \qquad n\frac{(\bar{x} - \mu)^2}{\sigma^2} \sim \chi^2(1),$$

where the statistics under (2) and (3) are independently distributed.

Definitions.

(1) If $u \sim \chi^2(m)$ and $v \sim \chi^2(n)$ are independent chi-square variates with m and n degrees of freedom respectively, then

$$F = \left\{ \frac{u}{m} \bigg/ \frac{v}{n} \right\} \sim F(m, n),$$

which is the ratio of of the chi-squares divided by their respective degrees of freedom, has an F distribution of m and n degrees of freedom, denoted by $F(m, n)$.

(2) If $x \sim N(0, 1)$ is a standard normal variate and if $v \sim \chi^2(n)$ is a chi-square variate of n degrees of freedom, and if the two variates are distributed independently, then the ratio

$$t = x \bigg/ \sqrt{\frac{v}{n}} \sim t(n)$$

has a t distributed of n degrees of freedom, denoted $t(n)$.

Notice that

$$t^2 = \frac{x^2}{v/n} \sim \left\{ \frac{\chi^2(1)}{1} \bigg/ \frac{\chi^2(n)}{n} \right\} = F(1, n).$$

CONFIDENCE INTERVALS

Consider a standard normal variate $z \sim N(0, 1)$. From the tables in the back of the book, we can find numbers a, b such that, for any $Q \in (0, 1)$, there is $P(a \leq z \leq b) = Q$. The interval $[a, b]$ is called a $Q \times 100\%$ confidence interval for z. We can minimise the length of the interval by disposing it symmetrically about the expected value $E(z) = 0$, since $z \sim N(0, 1)$ is symmetrically distributed about its mean of zero.

We can easily construct confidence intervals for the parameters underlying our sample statistics. Since they are concerned with fixed parameters, such confidence statements differ in a subtle way from those regarding random variables.

A confidence interval for the mean of the $N(\mu, \sigma^2)$ distribution. Let $x_1, x_2, \ldots, x_n$ be a random sample from a normal $N(\mu, \sigma^2)$ distribution. Then

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{and} \qquad \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Therefore, we can find numbers $\pm\beta$ such that

$$P\left(-\beta \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq \beta\right) = Q.$$

But, the following events are equivalent:

$$\begin{aligned}\left(-\beta \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq \beta\right) &\Longleftrightarrow \left(-\beta\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq \beta\frac{\sigma}{\sqrt{n}}\right)\\ &\Longleftrightarrow \left(-\beta\frac{\sigma}{\sqrt{n}} \leq \mu-\bar{x} \leq \beta\frac{\sigma}{\sqrt{n}}\right)\\ &\Longleftrightarrow \left(\bar{x}-\beta\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+\beta\frac{\sigma}{\sqrt{n}}\right).\end{aligned}$$

Hence

$$P\left(\bar{x}-\beta\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+\beta\frac{\sigma}{\sqrt{n}}\right) = Q.$$

This says that the probability that the random interval $[\bar{x}-\beta\sigma/\sqrt{n}, \bar{x}+\beta\sigma/\sqrt{n}]$ falls over the true value μ is Q. Equivalently, given a particular sample that has a mean value of $\bar{x}$, we are $Q \times 100\%$ confident that μ lies in the resulting interval.

Example. Let (1.2, 3.4, 0.6, 5.6) be a random sample from a normal $N(\mu, \sigma^2 = 9)$ distribution. Then $\bar{x} = 2.7$ and

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{2.7-\mu}{3/2} \sim N(0,1).$$

Hence

$$P\left(-1.96 \leq \frac{2.7-\mu}{3/2} \leq 1.96\right) = Q,$$

and it follows that $(0.24 \leq \mu \leq 5.64)$ is our 95% confidence interval.

A confidence interval for μ when σ^2 is unknown. Usually, we have to estimate σ^2. The unbiased estimate of σ^2 is $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/(n-1)$. With this estimate replacing σ^2, we have to replace the standard normal distribution, which is appropriate to $\sqrt{n}(\bar{x}-\mu)/\sigma$, by the $t(n-1)$ distribution, which is appropriate to $\sqrt{n}(\bar{x}-\mu)/\hat{\sigma}$.

To demonstrate this result, consider writing

$$\frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{n}} = \left\{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\Bigg/\sqrt{\frac{\sum(x_i-\bar{x})^2}{\sigma^2(n-1)}}\right\},$$

and observe that we can cancel the unknown value of σ from the numerator and the denominator.

Now, $\sum(x_i - \bar{x})^2/\sigma^2 \sim \chi^2(n-1)$, so the denominator contains the root of a chi-square variate divided by its $n-1$ degrees of freedom. The numerator contains a standard normal variate. That is to say, the statistic has the form of

$$\left\{N(0,1)\Bigg/\sqrt{\frac{\chi^2(n-1)}{n-1}}\right\} \sim t(n-1).$$

To construct a confidence interval, we proceed as before, except that we replace the numbers $\pm\beta$, obtained from the table of the $N(0,1)$ distribution, by the corresponding numbers $\pm b$, obtained from the $t(n-1)$ table. Our statement becomes

$$P\left(\bar{x} - b\frac{\hat{\sigma}}{\sqrt{n}} \leq \mu \leq \bar{x} + b\frac{\hat{\sigma}}{\sqrt{n}}\right) = Q.$$

A confidence interval for the difference between two means. Imagine a treatment that affects the mean of a normal population without affecting its variance. An instance of this might be the application of a fertiliser that increases the yield of a crop without adversely affecting its hardiness. We might wish to estimate the effect of the fertiliser; and, in that case, we would probably want to construct a confidence interval for the estimate.

To establish a confidence interval for the change in the mean, we would take samples from the population before and after treatment. Before treatment, there is

$$x_i \sim N(\mu_x, \sigma^2); \quad i = 1, \ldots, n \qquad \text{and} \qquad \bar{x} \sim N\left(\mu_x, \frac{\sigma^2}{n}\right),$$

and, after treatment, there is

$$y_j \sim N(\mu_y, \sigma^2); \quad j = 1, \ldots, m \qquad \text{and} \qquad \bar{y} \sim N\left(\mu_y, \frac{\sigma^2}{m}\right).$$

Then, on the assumption that the two samples are mutually independent, the difference between the sample means is

$$(\bar{x} - \bar{y}) \sim N\left(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right).$$

Hence

$$\frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m}}} \sim N(0,1).$$

If σ^2 were known, then, for any given value of $Q \in (0,1)$, we could find a number β from the $N(0,1)$ table such that

$$P\left\{(\bar{x} - \bar{y}) - \beta\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}} \leq \mu_x - \mu_y \leq (\bar{x} - \bar{y}) + \beta\sqrt{\frac{\sigma^2}{n} + +\frac{\sigma^2}{m}}\right\} = Q.$$

This would give a confidence interval for $\mu_x - \mu_y$. Usually, we have to estimate σ^2 from the sample information. We have

$$\sum \frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n-1) \qquad \text{and} \qquad \sum \frac{(y_j - \bar{y})^2}{\sigma^2} \sim \chi^2(m-1),$$

which are independent variates with expectations equal to the numbers of their degrees of freedom. The sum of independent chi-squares is itself a chi-square with degrees of freedom equal to the sum of those of its constituent parts. Therefore,

$$\frac{\sum(x_i-\bar{x})^2+\sum(y_j-\bar{y})^2}{\sigma^2}\sim\chi^2(n+m-2)$$

has an expected value of $n+m-2$, whence

$$\hat{\sigma}^2=\frac{\sum(x_i-\bar{x})^2+\sum(y_j-\bar{y})^2}{n+m-2}$$

is an unbiased estimate of the variance.

If we use the estimate in place of the unknown value of σ^2, we get

$$\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\dfrac{\hat{\sigma}^2}{n}+\dfrac{\hat{\sigma}^2}{m}}}=\left\{\frac{(\bar{x}-\bar{y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}\Bigg/\sqrt{\frac{\sum(x_i-\bar{x})^2+\sum(y_j-\bar{y})^2}{\sigma^2(n+m-2)}}\right\}$$

$$\sim\frac{N(0,1)}{\sqrt{\frac{\chi^2(n+m-2)}{n+m-2}}}=t(n+m-2).$$

This is the basis for determining a confidence interval that uses an estimated variance in place of the unknown value.

A confidence interval for the variance. If $x_i\sim N(\mu,\sigma^2); i=1,\ldots,n$ is a random sample, then $\sum(x_i-\bar{x})^2/(n-1)$ is an unbiased estimate of the variance and $\sum(x_i-\bar{x})^2/\sigma^2\sim\chi^2(n-1)$. Therefore, by looking in the back of the book at the appropriate chi-square table, we can find numbers α and β such that

$$P\left(\alpha\leq\frac{\sum(x_i-\bar{x})^2}{\sigma^2}\leq\beta\right)=Q$$

for some chosen $Q\in(0,1)$. From this, it follows that

$$P\left(\frac{1}{\alpha}\geq\frac{\sigma^2}{\sum(x_i-\bar{x})^2}\geq\frac{1}{\beta}\right)=Q\Longleftrightarrow P\left(\frac{\sum(x_i-\bar{x})^2}{\beta}\leq\sigma^2\leq\frac{\sum(x_i-\bar{x})^2}{\alpha}\right)=Q$$

and the latter provides a confidence interval for σ^2.

We ought to choose α and β so as to minimise the length of the interval $[\alpha^{-1},\beta^{-1}]$. The chi-square is an asymmetric distribution, so it is tedious to do so. The distribution becomes increasingly symmetric as the sample size n increases, and so, for large values of n, we may choose α and β to demarcate equal areas within the two tails of the distribution.

The confidence interval for the ratio of two variances. Imagine a treatment that affects the variance of a normal population. We might also wish to allow for the possibility that the mean is also affected. Let $x_i\sim N(\mu_x,\sigma_x^2); i=1,\ldots,n$

be a random sample taken from the population before treatment and let $y_j \sim N(\mu_y, \sigma_y^2); j = 1, \ldots, m$ be a random sample taken after treatment. Then

$$\sum \frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi^2(n-1) \qquad \text{and} \qquad \sum \frac{(y_j - \bar{y})^2}{\sigma^2} \sim \chi^2(m-1),$$

are independent chi-squared variates, and hence

$$F = \left\{ \frac{\sum (x_i - \bar{x})^2}{\sigma_x^2 (n-1)} \bigg/ \frac{\sum (y_j - \bar{y})^2}{\sigma_y^2 (m-1)} \right\} \sim F(n-1, m-1).$$

It is possible to find numbers α and β such that $P(\alpha \leq F \leq \beta) = Q$, where $Q \in (0,1)$ is some chose probability value. Given such values, we may make the following probability statement:

$$P\left(\alpha \frac{\sum (y_j - \bar{y})^2 (n-1)}{\sum (x_i - \bar{x})^2 (m-1)} \leq \frac{\sigma_y^2}{\sigma_x^2} \leq \beta \frac{\sum (y_j - \bar{y})^2 (n-1)}{\sum (x_i - \bar{x})^2 (m-1)} \right) = Q.$$