3 Random Samples from Normal Distributions

Statistical theory for random samples drawn from normal distributions is very important, partly because a great deal is known about its various associated distributions and partly because the central limit theorem suggests that for large samples a normal approximation may be appropriate.

3.1 Useful theoretical results

Theorem 3.1 If $X_1,\ X_2, \ldots, X_n$ are independent random variables with $X_j \sim N(\mu_j, \sigma_j^2)$, then $Y = \sum_{j=1}^n \alpha_j X_j$ is also normally distributed with mean $\sum_{j=1}^n \alpha_j \mu_j$ and variance $\sum_{j=1}^n \alpha_j^2 \sigma_j^2$.
□

Proof Using moment generating functions - which you have just heard about:

$$\begin{aligned} M_{X_j}(t) &= \exp\left(\mu_j t + \tfrac{1}{2}\sigma_j^2 t^2\right) \\ M_Y(t) &= E\left(e^{t\sum_j \alpha_j X_j}\right) \\ &= \prod_{j=1}^n M_{X_j}(\alpha_j t), \text{ by independence,} \\ &= \exp\left(t\sum_{j=1}^n \alpha_j \mu_j + \tfrac{1}{2}t^2 \sum_{j=1}^n \alpha_j^2 \sigma_j^2\right). \end{aligned}$$

By uniqueness of the moment generating function the result follows.
■

Definition 3.1 A chi-square distribution with r degrees of freedom is a gamma distribution: $\chi^2(r)$ is the same as $\Gamma\left(\frac{r}{2}, \frac{1}{2}\right)$.

Note that the moment generating function is

$$M(t) = \left(\frac{1}{1-2t}\right)^{\frac{r}{2}}.$$

□

The density function $\chi^2(1)$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi}}\, x^{-\frac{1}{2}}\, e^{-\frac{1}{2}x}, \quad x > 0.$$

Using the above moment generating function, we see that $E(X) = M_X'(0) = 1$, $V(X) = 2$. You can see from the graphs below that χ^2-distributions have a distinctively skewed shape.

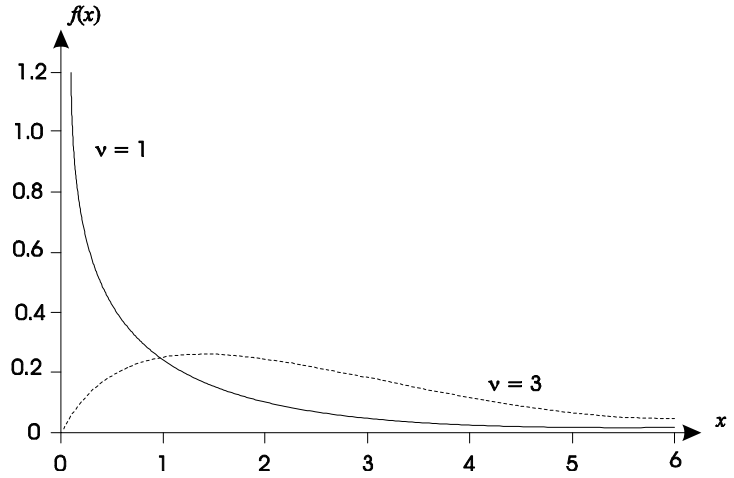

Figure 3.1 P.d.f.s of $\chi^2(1)$ and $\chi^2(3)$ distributions

The solid line is $\chi^2(1)$ and the dotted line is $\chi^2(3)$.

Lemma 3.1 If $X \sim N(0,1)$, then $Y = X^2 \sim \chi^2(1)$.
□

Proof Let $Y = X^2$. Then

$$F_Y(y) = P(Y \leq y) = P\left(X^2 \leq y\right) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) = F_X\left(\sqrt{y}\right) - F_X\left(-\sqrt{y}\right)$$

so that

$$f_Y(y) = F_Y'(y) = \frac{1}{2\sqrt{y}}\left(f_X\left(\sqrt{y}\right) + f_X\left(-\sqrt{y}\right)\right) = \frac{1}{\sqrt{2\pi y}}\, e^{-\frac{1}{2}y}, \quad y > 0.$$

■

Theorem 3.2 If $Y_1, Y_2, \ldots, Y_r$ are an independent random sample, each with a $\chi^2(1)$ distribution then

$$\sum_{j=1}^{r} Y_j \sim \chi^2(r).$$

□

Proof Just consider the moment generating functions. The m.g.f of a $\chi^2(r)$ random variable is $(1-2t)^{-r/2}$, so the m.g.f. of a $\chi^2(1)$ random variable is $(1-2t)^{-1/2}$. But the m.g.f. of a sum of r independent identically distributed random variables, each with m.g.f. $M(t)$ is $[M(t)]^r$, so the m.g.f. of $\sum_{j=1}^{r} Y_j$ is $(1-2t)^{-r/2}$, which is the m.g.f. of a $\chi^2(r)$ random variable. Therefore, by the uniqueness theorem for m.g.f.s,

$$\sum_{j=1}^{r} Y_j \sim \chi^2(r).$$

■

Corollary If $Y_1 \sim \chi^2(r)$ and $Y_2 \sim \chi^2(s)$, and are independent, then

$$Y_1 + Y_2 \sim \chi^2(r+s).$$

■

3.2 Independence of $\overline{X}$ and S^2 for normal samples.

One of the key results for normal random samples is the independence of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, and their relationship to the mean and variance parameters of a normal distribution.

Theorem 3.3 If $X_1, X_2, \ldots, X_n$ are independent, identically distributed random variables with normal distribution $N(\mu, \sigma^2)$, then $\overline{X}$ and S^2 are independent with distributions

(i) $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$;

(ii) $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

□

Proof There are various methods of proof. We will use one which delivers both independence and distribution within the same argument.

$$X_i \sim N\left(\mu, \sigma^2\right) \Longrightarrow Z_i = \frac{X_i - \mu}{\sigma} \sim N(0,1)$$

Now, from Theorem 3.1, we know that, if $\mathbf{Z}$ is a vector of normal random variables and $\mathbf{L}$ is a linear transformation, then $\mathbf{Y} = \mathbf{LZ}$ is also a vector of normal random variables. Suppose that $\mathbf{L}$ is orthogonal so that $\mathbf{L}^T\mathbf{L} = \mathbf{I}$. Then

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{Z}^T\mathbf{L}^T\mathbf{L}\mathbf{Z} = \mathbf{Z}^T\mathbf{Z} \quad \text{or} \quad \sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} Z_i^2.$$

Thus, if the joint p.d.f. of independent $N(0,1)$ variables Z_i, $i = 1, \ldots, n$ is

$$f_Z(\mathbf{z}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right), \quad \mathbf{Z} \in \mathbb{R}^n,$$

then joint p.d.f. of the Y_i, $i = 1, \ldots, n$ is

$$f_Y(\mathbf{y}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} y_i^2\right), \quad \mathbf{Y} \in \mathbb{R}^n,$$

and the Y_i variables are also independent and distributed as $N(0,1)$.

Now suppose we choose $\mathbf{L}$ such that its first row is

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right).$$

Then $Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i = \sqrt{n}\overline{Z}$, and

$$\sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2 = \sum_{i=1}^{n} Z_i^2 - n\overline{Z}^2 = \sum_{i=1}^{n} Y_i^2 - Y_1^2 = \sum_{i=2}^{n} Y_i^2,$$

which is independent of Y_1. Thus

$$\sum_{i=1}^{n} \left(Z_i - \overline{Z}\right)^2 \text{ is independent of } \overline{Z} \quad \Rightarrow \quad \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \text{ is independent of } \overline{X}$$

since $Z_i = \frac{X_i - \mu}{\sigma}$. The independence of $\overline{X}$ and S^2 is therefore proved.

(i) $Y_1 \sim N(0,1) \quad \Rightarrow \quad \sqrt{n}\overline{Z} \sim N(0,1) \quad \Rightarrow \quad \sqrt{n}\frac{\left(\overline{X} - \mu\right)}{\sigma} \sim N(0,1) \quad \Rightarrow$
$\overline{X} \sim N(\mu, \sigma^2/n)$;

(ii) From Theorem 3.2,

$$\sum_{i=2}^{n} Y_i^2 \sim \chi^2(n-1) \quad \Rightarrow \quad \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\Rightarrow \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

■

As we have just seen,$X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$ then

$$\sqrt{n}\frac{\left(\overline{X} - \mu\right)}{\sigma} \sim N(0,1).$$

When trying to make inferences from normal data we are often interested in μ, the location parameter, but the variance σ^2 is unknown. We need to find an estimator for the mean μ which does not contain the unknown variance parameter. To do this we are going to need to define something new and do a little more theory.

Definition 3.2 If $U \sim N(0,1)$ and $V \sim \chi^2(r)$ are independent, then

$$T = \frac{U}{\sqrt{V/r}} \sim t(r)$$

has a t-distribution with r degrees of freedom. This defines the distribution $t(r)$.
□

Figure 3.2 shows a graph of $t(7)$ compared with a standard $N(0,1)$ distribution. You can see that it is very similar, but has fatter tails.

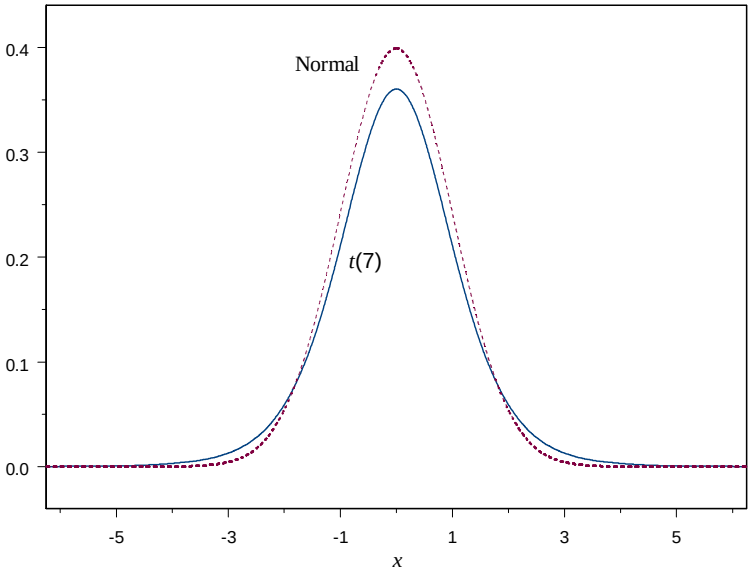

Figure 3.2 Normal and $t(7)$ distributions

Properties of the t-distribution

(i) The pdf of the $t(r)$ distribution is

$$f(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{(\pi r)}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}}, \ t \in \mathbb{R}.$$

(ii) If $r = 1$ then $t(1)$ is a Cauchy distribution without finite mean or variance.

(iii) As $r \to \infty$ then $t(r) \to \mathrm{N}(0, 1)$.

Referring again to Figure 3.2, the distributions are not so very different in shape, and the higher the number of degrees of freedom the closer the t-distribution approaches to the standard normal distribution.

These results are all leading to the fact that, if we replace σ by S, then we know the distribution of the resulting estimator and the mean, variance and all quantiles are tabulated in either tables or any statistical package, including R.

Theorem 3.4 If $X_1, X_2, \ldots, X_n$ are a normal random sample then

$$\frac{\sqrt{n}(\overline{X} - \mu)}{S} \sim t(n-1).$$

□

Proof $\overline{X}$ and S^2 are independent and

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1) \ , \ \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Using the definition of the t-distribution, the result follows with the unknown σ cancelling.
■

3.3 Estimators and confidence intervals

Z-statistics

Given a set of data which we regard as plausibly normal we might wish to find a point estimate of the mean μ. The previous section suggests that $\overline{X}$ is an obvious candidate. We also need to know what is the likely error range. If we had a different estimate, how close to the estimator $\overline{X}$ should it lie to be regarded as a plausible estimator of μ? For example consider the radiocarbon dating sample in Data Set 2.4. Without the outlier we have seen that it is plausible that the data is a normal random sample. With the outlier, $\overline{x} = 2622$, without $\overline{x} = 2505.9$. In either case, how reliable is our estimate, can we trust it? To within what error bounds? Should we include or exclude the outlier? If we exclude the outlier then the data is plausibly normal and we have a possible measure of spread in terms of the sample standard deviation s, although the variance σ^2 is unknown and is often referred to as a nuisance parameter. We need some theory, making use of the previous sections.

Definition 3.3 An *estimator* of a parameter θ is a statistic, say a function $A(X_1, X_2, \ldots, X_n)$ of the random sample, which does not depend on any unknown parameters in the model and which we use to give a point estimate of the parameter from the data.
□

An example of this is the way we use $\overline{X}$ to estimate the mean of a distribution.
If the estimator is to have any use at all, it should have some nice properties. For example, we know that $\overline{X} \xrightarrow{P} \mu$ by the weak law of large numbers, ensuring that $\overline{X}$ is a sensible estimator for μ.

A starting point for considering the likely error using the normal distribution is given by

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1).$$

Definition 3.4 A *Z-statistic* is a statistic with a standard normal distribution (as above).
□

The main use of Z-statistics stems from the facts that, for a general distribution, the Central Limit Theorem implies asymptotically that

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1),$$

and that the standard normal distribution involves no unknown parameters: it can be (and is) tabulated.

We can use the Z-statistic to calculate a range of plausible values for μ, under the assumption that σ^2 is known.

Definition 3.5 *Confidence interval*

Let $\mathbf{X}$ represent a vector of random variables with entries X_i. If $(a(\mathbf{X}), b(\mathbf{X}))$ is a random interval such that

$$P(a(\mathbf{X}) < \mu < b(\mathbf{X})) = 1 - \alpha,$$

then a realisation of that interval, $(a(\mathbf{x}), b(\mathbf{x}))$ is said to be a $100(1-\alpha)\,\%$ confidence interval for μ.
□

It is not easy to get to grips with what is meant by a confidence interval. Clearly one cannot say that the parameter μ has probability $(1-\alpha)$ of lying within the calculated interval $(a(\mathbf{x}), b(\mathbf{x}))$ because the ends of the interval are fixed numbers, as is μ, and without random variables being present, probability statements cannot be made: either μ lies between the two numbers or it doesn't, and we have no way of knowing which. The only viable interpretation is to say that we have used a procedure which, if repeated over and over again, would give an interval containing the parameter $100(1-\alpha)\,\%$ of the time: the rest of the time we will be unlucky.

Central $100(1-\alpha)\%$ confidence intervals using Z-statistics are found as follows. Remembering that $Z \sim N(0,1)$, choose $z_{\alpha/2}$ such that

$$\begin{aligned} P\left(Z \leq z_{\alpha/2}\right) &= 1 - \frac{\alpha}{2} \\ \Longrightarrow P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) &= 1 - \alpha. \end{aligned}$$

If $Z = \dfrac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}$ as above, then

$$\begin{aligned} P\left(-z_{\alpha/2} \leq \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \leq z_{\alpha/2}\right) &= 1 - \alpha \\ \Longrightarrow \quad P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) &= 1 - \alpha. \end{aligned}$$

Hence the appropriate random interval is

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\ ,\ \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\right)$$

and the $100\,(1-\alpha)\,\%$ confidence interval is

$$\left(\overline{x} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\ ,\ \overline{x} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\right).$$

The most common value of α in use is 0.05, in which case $z_{\alpha/2} = z_{0.025} = 1.960$.

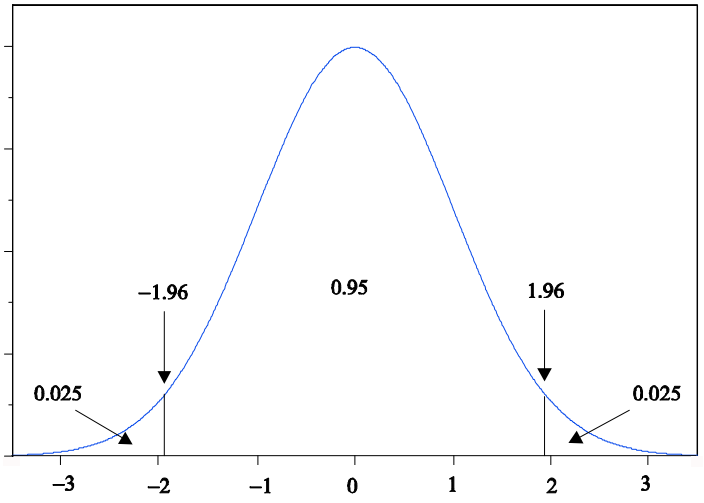

Figure 3.3 95% interval for $N(0,1)$

Example 3.1 *Radioactive-carbon dating*

In order to estimate the age of the site, we need to take the following steps.

(i) Check that the data are plausibly normal. We did this in Example 2.8 using a normal probability plot. We decided that we should leave out one point because it was a clear outlier.

(ii) Estimate the mean of the distribution by the sample mean and write $\widehat{\mu} = \overline{x} = 2505.86$.

(iii) Use a Z-statistic to find a 95% confidence interval which gives a range of plausible values for the mean age. This is

$$\left(\overline{x} - \frac{1.96\sigma}{\sqrt{n}}\ ,\ \overline{x} + \frac{1.96\sigma}{\sqrt{n}}\right),$$

and, putting in $n = 7$ and $\overline{x} = 2505.86$, we find a central 95% confidence interval

$$\left(2505.86 - \frac{1.96\sigma}{\sqrt{7}}\ ,\ 2505.86 + \frac{1.96\sigma}{\sqrt{n}}\right),$$

Unfortunately we are no better off. We cannot obtain the confidence interval because we do not know σ, so what should we do? We would like to replace σ by s, the sample standard deviation, but can we? $\left[\text{Recall that } S^2 = \frac{1}{n-1}\sum (x_i - \overline{x})^2 .\right]$

■

We know from Theorem 3.4 that, if $X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution $N(\mu, \sigma^2)$, then

$$T = \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} \sim t\,(n-1)$$

We can now look for a confidence interval by replacing the Z-statistic with the t-statistic. Writing $t_{\alpha/2}\,(n-1)$ for the $1 - \frac{\alpha}{2}$ quantile from the distribution $t(n-1)$,

$$P\left(-t_{\alpha/2}\,(n-1) < \frac{\sqrt{n}\left(\overline{X} - \mu\right)}{S} < t_{\alpha/2}\,(n-1)\right) = 1 - \alpha.$$

Re-arranging gives the random interval

$$\left(\overline{X} - \frac{t_{\alpha/2}}{\sqrt{n}}\,S,\ \overline{X} + \frac{t_{\alpha/2}}{\sqrt{n}}\,S\right),$$

and the $100\,(1-\alpha)\,\%$ confidence interval is the realisation of this interval.

Example 3.2 *Radioactive-carbon dating*

For the carbon-dating example, $n = 7$ and $t_{0.025}(6) = 2.447$, from a t-distribution with 6 degrees of freedom, $s = 56.44$. Plugging these values into the formula results in a 95% confidence interval of (2453.5, 2558.3), thereby giving a range of plausible values for μ.

■

3.4 Application of Central Limit Theorem

The Central Limit Theorem states that, for any random sample $X_1, X_2, \ldots, X_n$ such that the sample size n is sufficiently large, we have

$$\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma} \dot{\sim} N\,(0,1)\,.$$

Notation: $\dot{\sim}$ means 'approximately distributed as'. Provided we are dealing with moderate to large sample sizes we can therefore use the approximate normality to find confidence intervals, using approximate Z-statistics.

Example 3.3 *Binomial Proportion*

In an opinion poll prior to a Staffordshire South East by-election, of 688 constituents

chosen at random 368 said they would vote Labour (53.5%). The newspapers are perfectly happy to use these data to estimate p, the probability that a constituent selected at random would vote Labour, but they rarely, if ever, give any idea of the quality of the estimate. Let us see how to obtain a 95% confidence interval for p.

First identify the random sample. Constituents questioned are labelled 1,..., 688. Let

$$X_i = \begin{cases} 1, & \text{if } i^{\text{th}} \text{ constituent says "I will vote Labour",} \\ 0, & \text{otherwise.} \end{cases}$$

Then X_i has a Bernoulli distribution $B(1,p)$, the sample size n is 688, and $E(X_i) = p$, $V(X_i) = p(1-p)$. We know that p can be estimated by the sample mean $\overline{x} = \dfrac{368}{688} = 0.535$. We can also apply the Central Limit Theorem to find an approximate confidence interval using the asymptotic normality with $\mu = p,\ \sigma^2 = p(1-p)$. Thus

$$\frac{\sqrt{n}\left(\overline{X} - p\right)}{\sqrt{p\left(1-p\right)}} \dot{\sim} N(0,1).$$

The 95% random interval is of the form

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{0.025},\ \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{0.025}\right)$$

but unfortunately σ is a function of p. We could solve a quadratic inequality for p, but, since $n = 688$ is large, we will replace σ by its estimator $\sqrt{\overline{x}(1-\overline{x})}$. This gives (0.498, 0.572) as a 95% confidence interval for p, with point estimate 0.535.

If we required a 99% confidence interval we would use $z_{0.005} = 2.576$ to replace 1.960, and get a wider interval (0.486, 0.584) about which we are slightly more confident.

■