Dr Isabella Gollini

Probability and Statistics

8 Statistical inference and sampling distributions

8.1 The sample mean and sample variance

Suppose that we are interested in the distribution of a random variable X in some population. Let μ and σ^2 denote the mean and variance of X, respectively.

The practical problem will be that we do not know the values of μ and σ^2 but will take a sample from the population in order to estimate or test hypotheses about μ and σ^2. This is a problem of *statistical inference* — we make inferences about population parameters from sample data.

Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from the population, so that $X_1, X_2, \ldots, X_n$ are independently and identically distributed r.v.s with mean μ and variance σ^2. Recall from Section 7.3 that the sample mean $\bar{X}$ has the properties that

$$E(\bar{X}) = \mu, \tag{1}$$

and

$$\text{var}(\bar{X}) = \frac{\sigma^2}{n} \ . \tag{2}$$

Equation (1) expresses the fact the $\bar{X}$ is an *unbiased estimator* of μ — if we take many random samples of size n then on average, in the long run, the value of $\bar{X}$ will give the true value of μ.

Now consider the sample variance s^2,

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \ .$$

First we prove an introductory lemma, which has a number of interesting and useful applications.

Lemma 1 *For any constant a,*

$$\sum_{i=1}^{n}(X_i - a)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(a - \bar{X})^2 \ . \tag{3}$$

Proof.

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - a)^2 &= \sum_{i=1}^{n}[(X_i - \bar{X}) - (a - \bar{X})]^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 - 2(a - \bar{X})\sum_{i=1}^{n}(X_i - \bar{X}) + \sum_{i=1}^{n}(a - \bar{X})^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(a - \bar{X})^2,
\end{aligned}$$

since $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$.

Corollary 2 *$\sum_{i=1}^{n}(X_i - a)^2$ as a function of a is minimized when $a = \bar{X}$.*

Proof. The two terms on the right hand side of Equation (3) are both non-negative. Only the second of them is a function of a. It attains its minimum value 0 when $a = \bar{X}$.

Corollary 3 *We may write*

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \ .$$

[This provides an alternative formula for calculating the sample variance.]

Proof. Put $a = 0$ in the result of Lemma 1 and rearrange.

Corollary 4

$$E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = (n-1)\sigma^2 \ .$$

Proof. Putting $a = \mu$ in the result of Lemma 1, we may write

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

Taking expectations,

$$\begin{aligned} E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) &= E\left(\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2\right) \\ &= \sum_{i=1}^{n} E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \ = \ (n-1)\sigma^2 \ , \end{aligned}$$

since $E[(X_i - \mu)^2] = \mathrm{var}(X_i) = \sigma^2$ and, using Equations (1) and (2), $E[(\bar{X} - \mu)^2] = \mathrm{var}(\bar{X}) = \sigma^2/n$.

It follows that from the result of Corollary 4 that

$$E(s^2) = \frac{1}{n-1}E\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2 \ .$$

Thus

$$E(s^2) = \sigma^2 \ . \tag{4}$$

Equation (4) expresses the fact the s^2 is an *unbiased estimator* of σ^2 — it provides an explanation of why the divisor $n - 1$ is used in the definition of sample variance.

	mean	variance
population/distribution	μ	σ^2
sample (unbiased estimator)	$\bar{X}$	s^2

Table 1: Population and sample

8.2 Estimation of the population mean for a normal distribution

Suppose further that the r.v. X is normally distributed, so that it has a $N(\mu, \sigma^2)$ distribution for some values of μ and σ^2. In other words, the r.v. X is normally distributed in the population from which we are sampling, with mean μ and variance σ^2.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from the population. We saw in Section 7.3 that the sampling distribution of the sample mean $\bar{X}$, which describes how the values of the sample mean vary from sample to sample if random samples of size n are taken repeatedly, is given by

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) , \tag{5}$$

from which it follows that

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) . \tag{6}$$

Suppose that we do not know the value of the population mean μ but that we wish to estimate it from a random sample of size n. A natural estimate of μ is the sample mean $\bar{x}$, which, as we noted in Section 8.1, provides an unbiased estimate of μ.

In statistical practice, we are concerned not only to estimate unknown parameter values but also to provide measures of the precision of our estimates. For the present we shall assume that we know the value of the population variance σ^2.

- Mostly, when we are sampling from a normal distribution, both the population mean μ and the population variance σ^2 are unknown. But this will not always be so.

  For example, if the population consists of items coming off a production line, long experience of the process may have established what is the variance σ^2 of the measurement being taken, but the mean μ of the measurement may alter from time to time, and it is this that we are attempting to monitor and estimate.

From Equation (5), we note that the standard deviation of $\bar{X}$ is

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} ,$$

which is also known as the *standard error* when $\bar{X}$ is being considered as an estimator of μ.

- The standard error is the "root mean square error," a measure of the typical size of the error in the estimate $\bar{x}$ of μ.
- Note how the standard error decreases proportionally to $1/\sqrt{n}$. As the sample size increases, we may expect $\bar{x}$ to provide us with a more precise estimate of μ.

But we can go further and provide what is known as a *confidence interval* for μ.

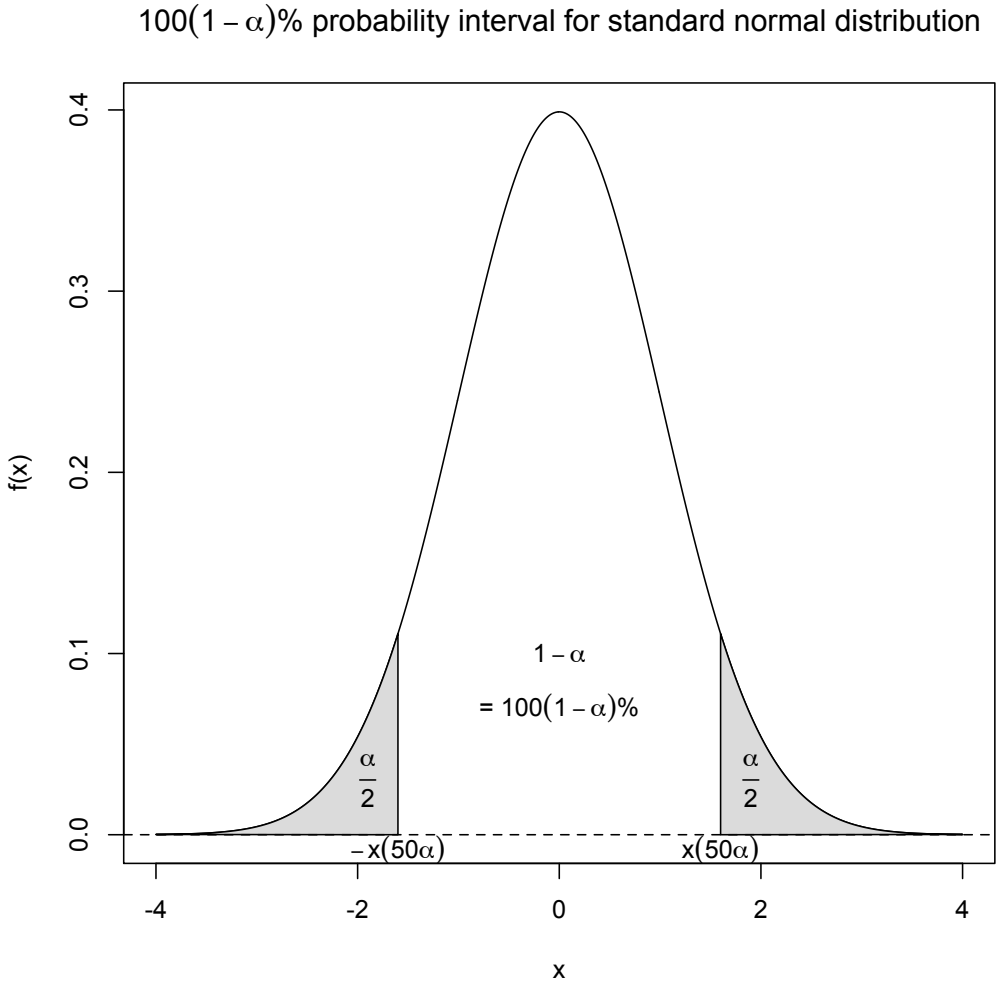

Figure 1: Symmetric probability interval for standard normal distribution

From the result of Equation (6), together with the definition of percentage points, as illustrated in Figure 1, we have that

$$\Pr\left(-x(50\alpha) < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < x(50\alpha)\right) = 1-\alpha\ .$$

Rearranging the terms within the brackets on the left hand side, we obtain the result that, whatever the value of μ, as we repeatedly take random samples of size n,

$$\Pr\left(\bar{X} - x(50\alpha)\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + x(50\alpha)\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \ .$$

Now given a particular observed value of the sample mean $\bar{x}$, *a* $100(1-\alpha)\%$ *confidence interval for* μ is given by

$$\left(\bar{x} - x(50\alpha)\frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + x(50\alpha)\frac{\sigma}{\sqrt{n}}\right) \ .$$

The most commonly used value of α is 0.05. As $x(2.5) = 1.96$, a 95% confidence interval for μ is given by

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

Another commonly used value of α is 0.01. As $x(0.5) = 2.5758$, a 99% confidence interval for μ is given by

$$\left(\bar{x} - 2.5758\frac{\sigma}{\sqrt{n}} \ , \ \bar{x} + 2.5758\frac{\sigma}{\sqrt{n}}\right) \ .$$

8.3 Further sampling distributions associated with normal data

We turn next to sampling distributions that will be used when sampling from a normal distribution where the population variance σ^2 is unknown. Firstly we deal with the sampling distribution of the sample variance s^2. This involves consideration of the distribution of sums of squares of normally distributed r.v.s.

As a preliminary, consider n independently and identically distributed r.v.s, $Z_1, Z_2, \ldots, Z_n$, each with a standard normal distribution, and let

$$W = \sum_{i=1}^{n} Z_i^2 \ .$$

The distribution of the r.v. W is given by the *chi-square distribution* with n degrees of freedom. The family of chi-square distributions with ν degrees of freedom ($\nu = 1, 2, \ldots$) is important in statistical analysis. We write the chi-square distribution with ν degrees of freedom as χ^2_ν. So, for example, $W \sim \chi^2_n$.

letter	pronunciation
μ	mu
ν	nu
σ	sigma
χ	chi

Table 2: Greek letters

Values of the chi-square distribution functions and their percentage points $\chi^2_\nu(P)$ are to be found in Tables 7 and 8, respectively, of *Lindley and Scott*. The p.d.f. is illustrated in Figure 2 for the case $\nu = 10$.

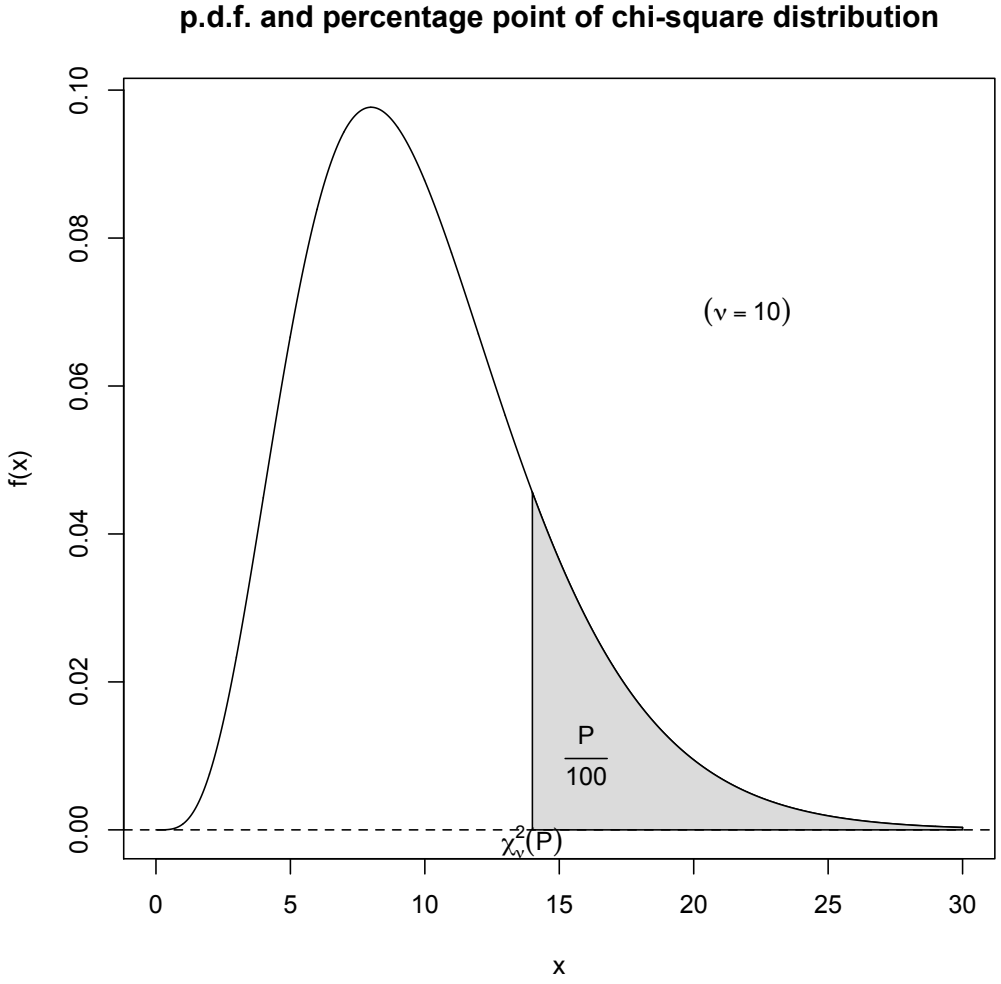

Figure 2: The chi-square distribution with ν degrees of freedom (χ^2_ν)

- Note that chi-square distributions are non-symmetric distributions, restricted to take positive values (Figure 3).
- If $W \sim \chi^2_\nu$ then $\Pr(W > \chi^2_\nu(P)) = P\% = P/100$.

For example, from Table 8, $\chi^2_{10}(20) = 13.44$, so that if $W \sim \chi^2_{10}$ then $\Pr(W > 13.44) = 0.2$.

`R` may also be used to calculate percentage points of the chi-square distributions. Use the `qchisq` function to calculate values of the inverse of the cumulative distribution function. If F denotes the c.d.f. of the χ^2_ν distribution then, by the same argument as used for the standard normal distribution in Section 7.4,

$$\chi^2_\nu(P) = F^{-1}(1 - P/100) \ .$$

In the following `R` session, the value 13.4420 is obtained for $\chi^2_{10}(20)$.

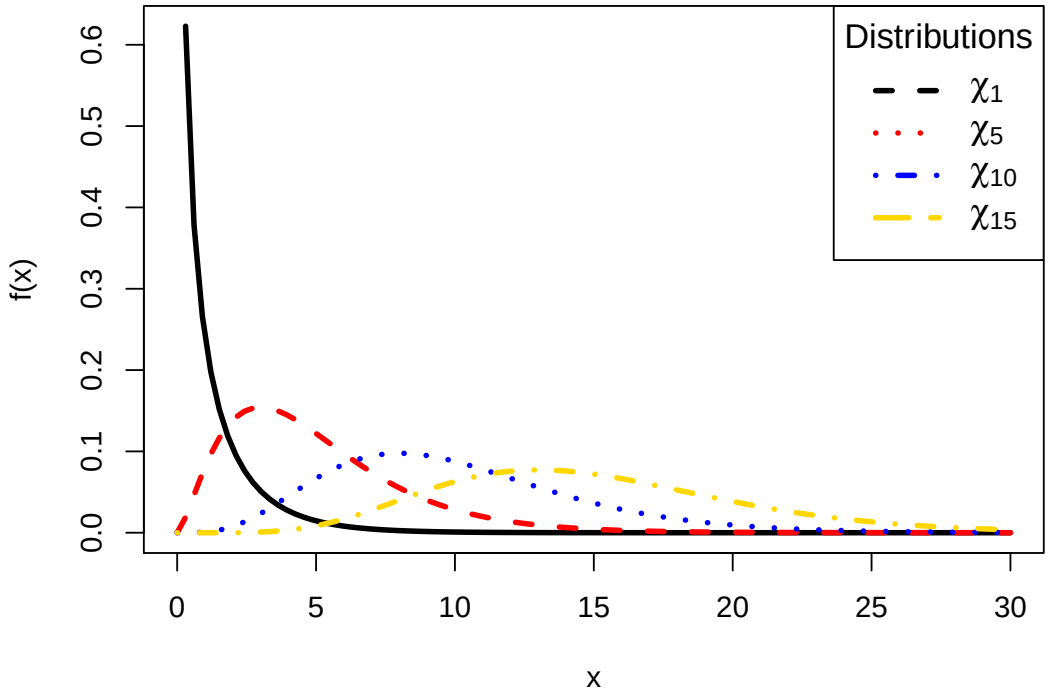

Figure 3: The chi-square distribution with $\nu = 1, 5, 10, 15$ degrees of freedom (χ^2_ν)

```
P <- 20
qchisq(1 - P / 100, 10)

## [1] 13.44196
```

Now let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ distribution. Define

$$Z_i = \frac{X_i - \mu}{\sigma} \qquad (1 \leq i \leq n).$$

Then $Z_1, Z_2, \ldots, Z_n$ are independently and identically distributed r.v.s, each with a standard normal distribution. Hence

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n} Z_i^2 \ \sim \ \chi^2_n \ . \tag{7}$$

Putting $a = \mu$ in the result of Lemma 1, we find

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\mu - \bar{X})^2$$

so that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{n(\bar{X} - \mu)^2}{\sigma^2} \ . \tag{8}$$

From Equation (6) it follows that

$$\frac{n(\bar{X}-\mu)^2}{\sigma^2} = \left(\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right)^2 \sim \chi^2_1 \ . \tag{9}$$

It turns out that the two terms on the right hand side of Equation (8) are independently distributed and, furthermore, that

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2_{n-1} \ . \tag{10}$$

Equation (10) may be written equivalently as

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \ . \tag{11}$$

- Equation (11) specifies the sampling distribution of the sample variance s^2 and is fundamental to making inferences about σ^2.
- The degrees of freedom in Equation (10)/(11) are $n-1$, one less than the sample size n.
- According to the results of Equations (7), (9) and (10), in Equation (8) the term on the left hand side, which has the χ^2_n distribution, has been partitioned into two independently distributed terms on the right hand side, which have the χ^2_{n-1} and χ^2_1 distributions, respectively.

Another family of distributions that is important in statistical analysis is the family of *t-distributions* with ν degrees of freedom ($\nu = 1, 2, \ldots$). We write the t-distribution with ν degrees of freedom as t_ν.

Equation (6) has a fundamental role in making inferences about μ when σ is known. But when σ is unknown, it is replaced on the left-hand side of Equation (6) by the estimate s. It turns out that the resulting r.v. has the t_{n-1} distribution:

$$\frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1} \ . \tag{12}$$

Equation (12) is fundamental in making inferences about μ when σ is unknown.

Values of the distribution functions of the t-distributions and their percentage points $t_\nu(P)$ are to be found in Tables 9 and 10, respectively, of *Lindley and Scott*. The p.d.f. is illustrated in Figure 4 for the case $\nu = 4$.

- The t-distributions are symmetric distributions, rather similar in shape to the $N(0,1)$ distribution, but they differ from the $N(0,1)$ distribution in that they have longer tails, although for ν large there is little difference between the t_ν distribution and the $N(0,1)$ distribution (Figure 5).

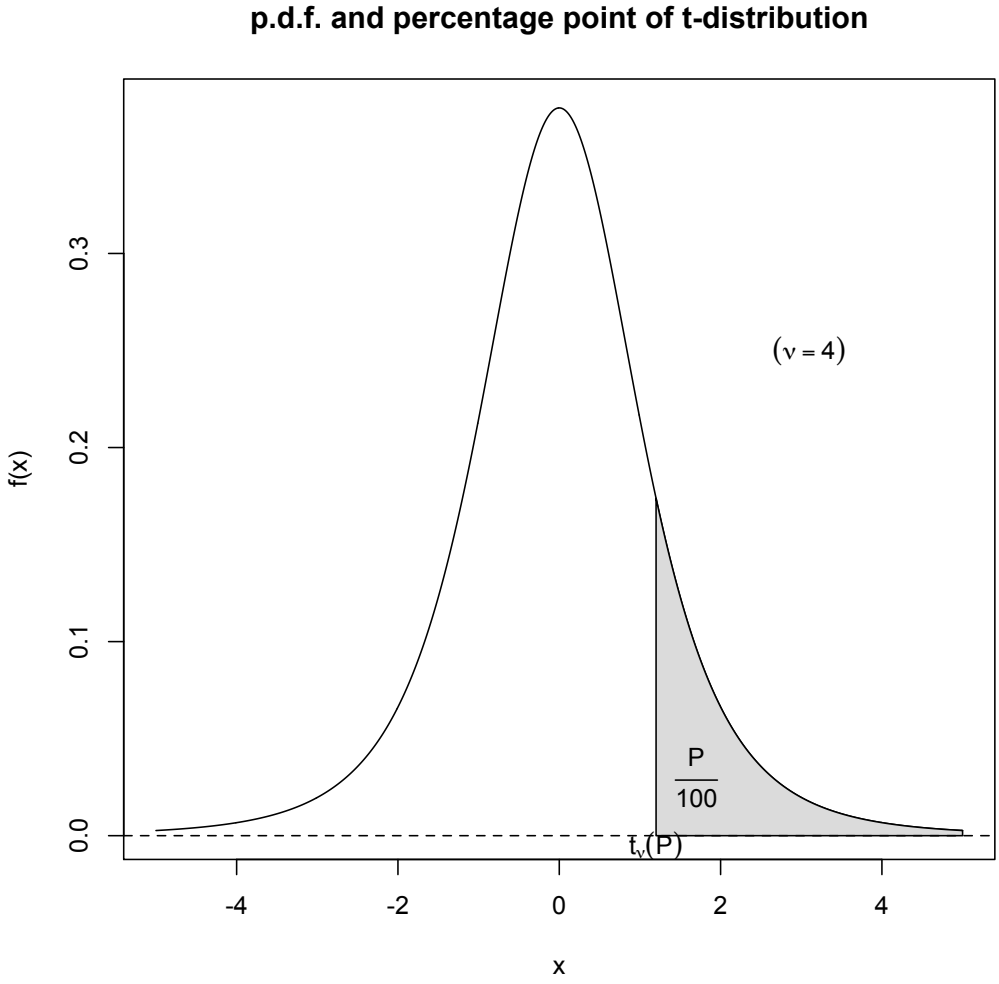

Figure 4: The t-distribution with ν degrees of freedom (t_ν)

- As $\nu \to \infty$, the t_ν distribution converges to the $N(0,1)$ distribution, and, for given P, $t_\nu(P) \downarrow x(P)$, the corresponding percentage point of the $N(0,1)$ distribution.
- In Table 10 of *Lindley and Scott*, the final row, labelled $\nu = \infty$, gives percentage points of the $N(0,1)$ distribution.

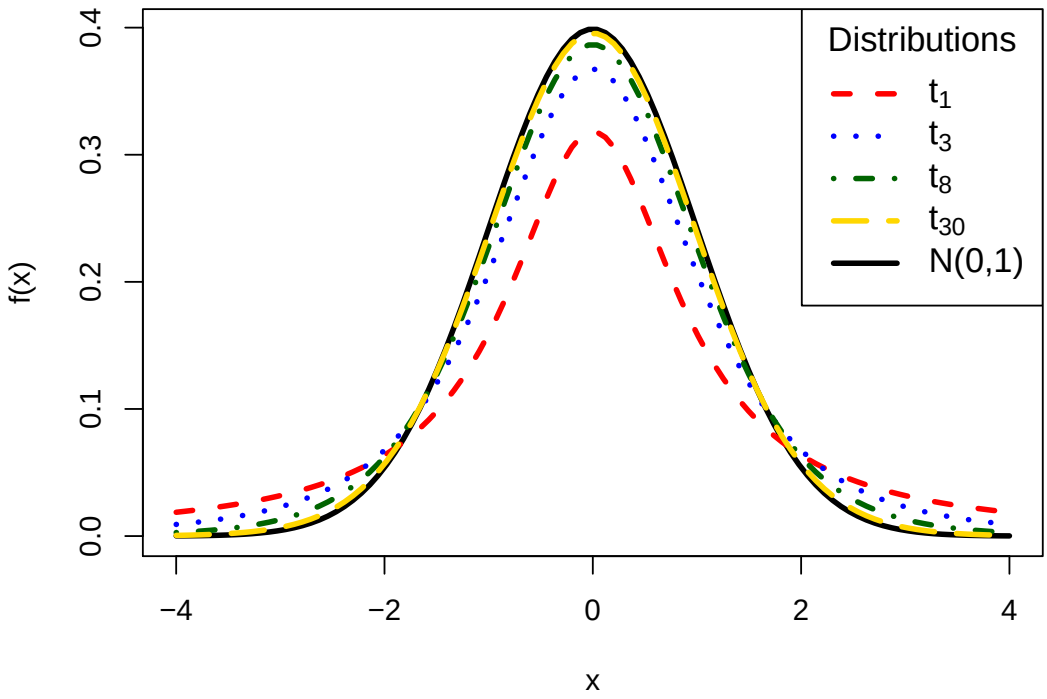

Figure 5: Comparison normal distribution and t-distributions with $\nu = 1, 3, 8, 30$ degrees of freedom (t_ν)

For example, from Table 10, $t_{10}(2.5) = 2.228$.

In `R`, use the function `qt` to calculate values of the inverse of the cumulative distribution function. If F denotes the c.d.f. of the t_ν distribution then

$$t_\nu(P) = F^{-1}(1 - P/100) \ .$$

In the following `R` session, the value 2.22814 is obtained for $t_{10}(2.5)$.

```
P <- 2.5
qt(1 - P / 100, 10)

## [1] 2.228139
```

It follows that if the r.v. $T \sim t_{10}$ then

$$\Pr(-2.22814 < T < 2.22814) = 0.95 \ .$$