Dr Isabella Gollini

Probability and Statistics

9 Statistical inference for normally distributed data

9.1 A confidence interval for the mean

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ distribution, where the mean μ and the variance σ^2 are unknown. (From now on we shall use lower case letters to denote sample values.)

We shall still use $\bar{x}$ as our estimate of μ, but we shall not be able to use the method of Section 8.2 for constructing a confidence interval for μ, where it was assumed that the value of σ was known. There the standard error of $\bar{x}$, considered as an estimator of μ, was given as $\sigma/\sqrt{n}$. Now we replace σ by its estimate, the sample standard deviation s, so that the *standard error* of $\bar{x}$ is given by

$$\frac{s}{\sqrt{n}} .$$

Recalling that $\bar{x} \sim N(\mu, \sigma^2/n)$, in Section 8.2 the construction of a confidence interval for μ was based upon the fact that $(\bar{x}-\mu)/(\sigma/\sqrt{n}) \sim N(0,1)$, using percentage points of the standard normal distribution. Now we shall use instead the t-distribution with $\nu = n-1$ degrees of freedom and, specifically, the result presented in Section 8.3 that

$$\frac{(\bar{x}-\mu)}{\frac{s}{\sqrt{n}}} \sim t_{n-1} .$$

As illustrated in Figure 1, it follows that

$$\Pr\left(-t_{n-1}(50\alpha) < \frac{(\bar{x}-\mu)}{\frac{s}{\sqrt{n}}} < t_{n-1}(50\alpha)\right) = 1-\alpha ,$$

where $t_{n-1}(50\alpha)$ is a percentage point of the t_{n-1} distribution, whose value may be looked up in Table 10 of *Lindley and Scott*.

Rearranging the terms within the brackets on the left hand side, we obtain the result that, whatever the values of μ and σ^2, as we repeatedly take random samples of size n,

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

Now given particular observed values of the sample mean $\bar{x}$ and the sample variance s^2, *a* $100(1-\alpha)\%$ *confidence interval for* μ is given by

$$\left(\bar{x} - t_{n-1}(50\alpha)\frac{s}{\sqrt{n}} \; , \; \bar{x} + t_{n-1}(50\alpha)\frac{s}{\sqrt{n}}\right) .$$

The most commonly used values of α are 0.05, for a 95% confidence interval, and 0.01, for a 99% confidence interval.

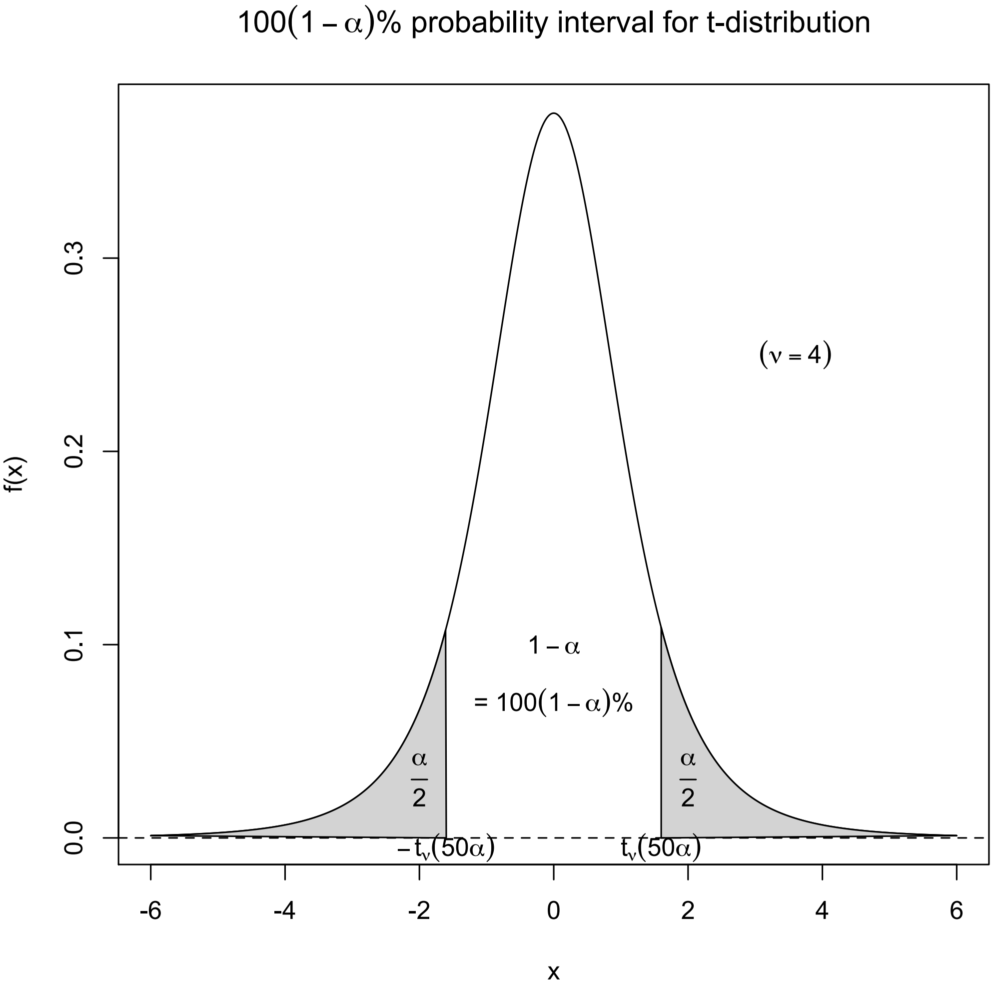

Figure 1: Symmetric probability interval for the t_ν distribution

- Under exam conditions you may need to calculate sample mean, sample variance and sample standard deviation using a hand calculator. You may do this in a number of ways. Using simple summation buttons, for the sample variance you may use either of the formulae

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

or

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right] ,$$

where the second formula follows from Corollary 3 of Section 8.1.

Example

Consider the following 10 measurements in centimetres of the lengths of skulls of fossil skeletons of an extinct species of bird.

$$5.82, 6.10, 5.63, 5.95, 6.24, 5.51, 6.06, 6.63, 6.19, 5.67$$

We assume that we have a random sample of observations and that the lengths of skulls for the given species are normally distributed with unknown mean μ and unknown variance σ^2.

We first calculate the sample mean, $\bar{x} = 5.98$, the sample variance, $s^2 = 0.11344$, and the sample standard deviation, $s = 0.3368$. The standard error of the sample mean is

$$\frac{s}{\sqrt{n}} = \frac{0.3368}{\sqrt{10}} = 0.1065.$$

In R:

```
skulls <- c(5.82, 6.10, 5.63, 5.95, 6.24,
            5.51, 6.06, 6.63, 6.19, 5.67)
n <- length(skulls)
n

## [1] 10

xbar <- mean(skulls)
xbar

## [1] 5.98

s <- sd(skulls)
s

## [1] 0.3368151

se <- s / sqrt(n)
se

## [1] 0.1065103
```

The degrees of freedom are $n - 1 = 9$. A 95% confidence interval for the mean skull length in the population, μ, is given by

$$\left(5.98 - t_9(2.5)\frac{0.3368}{\sqrt{10}}\ ,\ 5.98 + t_9(2.5)\frac{0.3368}{\sqrt{10}}\right)$$

i.e., using Table 10,

$$(5.98 - (2.262)(0.1065)\ ,\ 5.98 + (2.262)(0.1065)) \ = \ (5.74, 6.22)\ .$$

In R:

```
alpha <- 0.05
tval95 <- qt(1 - alpha / 2, df = n - 1)
tval95

## [1] 2.262157

xbar - tval95 * se

## [1] 5.739057

xbar + tval95 * se

## [1] 6.220943
```

The 99% confidence interval is:

```
alpha <- 0.01
tval99 <- qt(1 - alpha / 2, df = n - 1)
tval99

## [1] 3.249836

xbar - tval99 * se

## [1] 5.633859

xbar + tval99 * se

## [1] 6.326141
```

9.2 A confidence interval for the variance

Again let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a $N(\mu, \sigma^2)$ distribution, where the mean μ and the variance σ^2 are unknown.

We use the sample variance s^2 as an estimate of σ^2. As we saw in Section 8.1, it is an unbiased estimate. To construct a confidence interval for σ^2 we use the result of Section 8.3 that

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

Because the chi-square distributions are non-symmetrical, we will have to look up two different percentage points, as illustrated in Figure 2, to construct a $100(1-\alpha)\%$ probability interval. Using $\nu = n-1$, we look up in Table 8 of *Lindley and Scott* the percentage points $\chi^2_{n-1}(50\alpha)$ and $\chi^2_{n-1}(100-50\alpha)$.

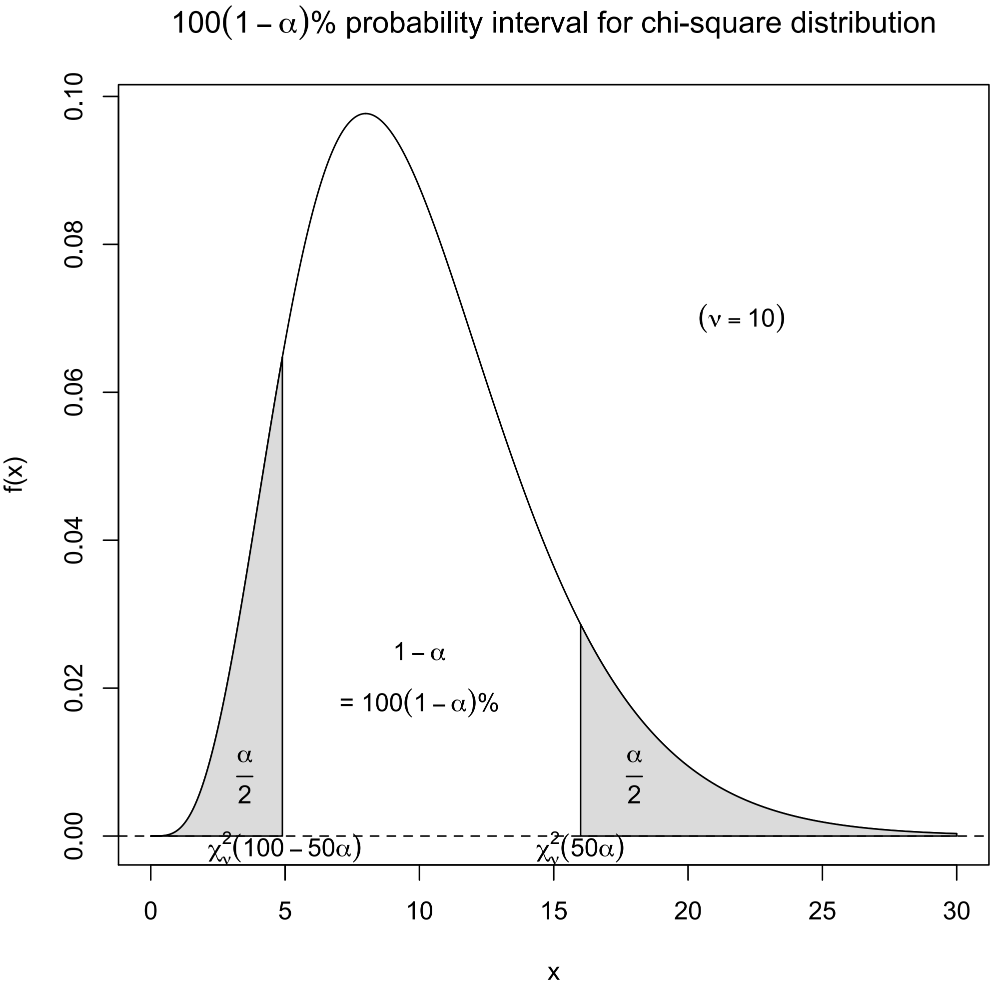

Figure 2: Probability interval for the χ^2_ν distribution, with equal tail probabilities

Corresponding to Figure 2 we have the probability statement

$$\Pr\left(\chi^2_{n-1}(100-50\alpha) < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1}(50\alpha)\right) = 1-\alpha \ .$$

Rearranging the terms within the brackets on the left hand side, we obtain the result that, whatever the values of μ and σ^2, as we repeatedly take random samples of size n,

$$\Pr\left(\frac{(n-1)s^2}{\chi^2_{n-1}(50\alpha)} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1}(100-50\alpha)}\right) = 1-\alpha \ .$$

Now given a particular observed value of the sample variance s^2, *a* $100(1-\alpha)\%$ *confidence interval for* σ^2 is given by

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1}(50\alpha)} \ , \ \frac{(n-1)s^2}{\chi^2_{n-1}(100-50\alpha)}\right) \ .$$

As in the case of the confidence interval for the mean, the most commonly used values of α are 0.05, for a 95% confidence interval, and 0.01, for a 99% confidence interval.

Example (continued)

Using the same data as in Section 9.1, an unbiased estimate of the population variance σ^2 of the skull lengths in the population is given by $s^2 = 0.11344$. A 95% confidence interval for σ^2 is given by

$$\left(\frac{(n-1)s^2}{\chi^2_9(2.5)} \, , \, \frac{(n-1)s^2}{\chi^2_9(97.5)} \right)$$

i.e., using Table 8,

$$\left(\frac{(9)(0.11344)}{19.02} \, , \, \frac{(9)(0.11344)}{2.700} \right) \; = \; (0.054, 0.378) \; .$$

```
alpha <- 0.05
chisq25 <- qchisq(alpha / 2, df = n - 1, lower.tail = FALSE)
chisq975 <- qchisq(1 - alpha / 2, df = n - 1, lower.tail = FALSE)
(n - 1) * s^2 / chisq25

## [1] 0.05367253

(n - 1) * s^2 / chisq975

## [1] 0.3780936
```

- The chi-square distribution theory that underlies the construction of the "standard" confidence interval for σ^2 is highly sensitive to the assumption that the data are normally distributed. In other words, the method is not *robust* against departures from normality — it will provide unreliable results if the data are not at least approximately normally distributed.

- The method for constructing a confidence interval for μ is more robust. As we saw in Section 7.3, using the Central Limit Theorem, if n is reasonably large then $\bar{x}$ is approximately normally distributed even if the underlying data are not, which provides at least some indication that the method of constructing confidence intervals based on the t-distribution might be reasonably robust.

9.3 Introduction to hypothesis testing

Example (continued)

Suppose that over a long period of time a large number of skulls of fossil skeletons from a given species have been measured and it is well established that the mean length of skull in the population was 5.65 cm.

Suppose further that the data used in Section 9.1 and 9.2 are the skull lengths of a newly found sample of fossil skeletons, where it is being investigated whether the new sample comes from the same species. Are the recorded measurements consistent with the hypothesis that the newly found sample is from the same species, or is the observed sample mean $\bar{x} = 5.98$ significantly different from the population mean $\mu = 5.65$?

To formalize the situation, we assume that the observed data are a random sample from a $N(\mu, \sigma^2)$ distribution, where μ and σ^2 are unknown — this is our *statistical model* for the data. We test the *null hypothesis* $H_0 : \mu = 5.65$ against the *alternative hypothesis* $H_1 : \mu \neq 5.65$.

To test the null hypothesis H_0 we use the *test statistic*

$$t = \frac{(\bar{x} - 5.65)}{\frac{s}{\sqrt{n}}},$$

which under H_0 has the t_{n-1} distribution, where n is the sample size. We shall reject H_0 if $|t|$ is too large.

In the present case, $n = 10$, $\bar{x} = 5.98$ and $s = 0.3368$. The calculated value of the test statistic, which under H_0 has the t_9 distribution, is

$$t = \frac{(5.98 - 5.65)}{\frac{0.3368}{\sqrt{10}}} = 3.0983.$$

The p-value of the test statistic is the probability p under the null hypothesis of observing the value actually obtained or a more extreme one.

In this case,

$$\begin{aligned} p &= \Pr(|t| \geq 3.0983) \\ &= \Pr(t \geq 3.0983) + \Pr(t \leq -3.0983) \\ &= 2\Pr(t \geq 3.0983), \end{aligned}$$

by the symmetry of the t distribution. If F denotes the distribution function of the t_9 distribution then, using Table 9 of *Lindley and Scott* with $\nu = 9$,

$$\begin{aligned} p &= 2[1 - F(3.0983)] \\ &\approx 2(1 - 0.9936) \\ &\approx 0.0128. \end{aligned}$$

The p-value is used as a measure of the *significance* of the value obtained of the test statistic. The smaller the p-value the more significant is the value of the test statistic and the stronger the evidence against the null hypothesis. If, as in this case, $p < 0.05$, we say that "we reject H_0 at the 5% significance level." If $p < 0.01$, which is not true in the present case, we would say that "we reject H_0 at the 1% significance level."

In the present case, because we have obtained a value of the test statistic that is significant at the 5% level (and almost significant at the 1% level), we have strong evidence to reject the null hypothesis that the sample comes from a population whose mean value is 5.65. There is strong evidence that the newly found sample does not come from the given species.

The test may readily be carried out using `R` by using the function `t.test` with the vector of data and the the hypothesized mean (in this case $\mu = 5.65$).

```
t.test(skulls, mu = 5.65)

##
##  One Sample t-test
##
## data:  skulls
## t = 3.0983, df = 9, p-value = 0.01276
## alternative hypothesis: true mean is not equal to 5.65
## 95 percent confidence interval:
##  5.739057 6.220943
## sample estimates:
## mean of x
##      5.98
```

The value of the t-statistic and its p-value are shown. There is no need to calculate the p-value separately, but as an exercise we may verify the p-value as follows:

```
pvalue <- 2 * (1 - pt(3.0983, 9))
pvalue

## [1] 0.01275741
```

The use of p-values has become commonplace since the development of statistical packages. Before, and nowadays too if you are not using a computer, the significance of a test statistic was evaluated by comparing its value with percentage points of the appropriate distribution.

In the present case $p \leq 0.05$ and we reject H_0 at the 5% significance level if and only if $|t| \geq t_9(2.5)$. For any α with $0 < \alpha < 1$, $p \leq \alpha$ and we reject H_0 at the $100\alpha\%$ significance level if and only if $|t| \geq t_9(50\alpha)$. Conventionally $\alpha = 0.05, 0.01$ and 0.001 are used, i.e., 5%, 1% and 0.1% significance levels. From Table 10, $t_9(2.5) = 2.262$ and $t_9(0.5) = 3.250$. Since $t = 3.0983 \geq t_9(2.5)$ but $t = 3.0983 < t_9(0.5)$, we reject H_0 at the 5% significance level but not at the 1% significance level.

`R` code to do the t-test with the 99% confidence interval:

```r
t.test(skulls, mu = 5.65, conf.level = 0.99)

##
##  One Sample t-test
##
## data:  skulls
## t = 3.0983, df = 9, p-value = 0.01276
## alternative hypothesis: true mean is not equal to 5.65
## 99 percent confidence interval:
##  5.633859 6.326141
## sample estimates:
## mean of x
##      5.98
```

Extra Examples

Questionnaire times

A certain telephone company conducts a survey. They stored the times to complete the questionnaire of a random sample of 7 people, the times in minutes are:

$$7, 10, 17, 11, 11, 10, 11.$$

We assume that the times to complete the questionnaire are normally distributed, with unknown mean and unknown variance.

- Calculate the 95% confidence interval for the mean time
- Calculate the 95% confidence interval for the variance
- The telephone company claims that the questionnaire should take 10 minutes to be completed the questionnaire. Is there any reason to assert the claim is not true?

Exam grades

Rumour has that the exam have been very successful, and that the grades are normally distributed with mean 85. Not feeling quite confident that she can believe this rumour Jane obtains the grades of a random sample of 9 students. The grades are:

$$79, 75, 84, 63, 98, 52, 87, 99, 83.$$

- On the basis of this set of data can Jane infer that the rumour of mean grade of 85 is false?