Special topics in statistical testing

Krzysztof Podgórski
Department of Mathematics and Statistics
University of Limerick

October 5, 2009

One-sided vs. two-sided test

We consider the null hypothesis:

$$H_0 : \quad \mu = \mu_0,$$

where μ is unknown but true value of the mean while μ_0 is a known hypothesized value of the true mean.

One-sided vs. two-sided test

We consider the null hypothesis:

$$H_0 : \quad \mu = \mu_0,$$

where μ is unknown but true value of the mean while μ_0 is a known hypothesized value of the true mean.
We want to detect if $\mu \neq \mu_0$.

One-sided vs. two-sided test

We consider the null hypothesis:

$$H_0 : \quad \mu = \mu_0,$$

where μ is unknown but true value of the mean while μ_0 is a known hypothesized value of the true mean.
We want to detect if $\mu \neq \mu_0$. Alternatively, we can consider the null hypothesis:

$$H_0 : \quad \mu \leq \mu_0,$$

One-sided vs. two-sided test

We consider the null hypothesis:

$$H_0 : \quad \mu = \mu_0,$$

where μ is unknown but true value of the mean while μ_0 is a known hypothesized value of the true mean.
We want to detect if $\mu \neq \mu_0$. Alternatively, we can consider the null hypothesis:

$$H_0 : \quad \mu \leq \mu_0,$$

and we want to detect if $\mu > \mu_0$.

One-sided vs. two-sided test

We consider the null hypothesis:

$$H_0 : \quad \mu = \mu_0,$$

where μ is unknown but true value of the mean while μ_0 is a known hypothesized value of the true mean.
We want to detect if $\mu \neq \mu_0$. Alternatively, we can consider the null hypothesis:

$$H_0 : \quad \mu \leq \mu_0,$$

and we want to detect if $\mu > \mu_0$. In the first case, we deal with a two-sided test while the second case is described as a one-sided one.

Example 3.5.1

It is suspected that an acid–base titrimetric method has a significant indicator error and thus tends to give results with a positive systematic error (i.e. positive bias). To test this an exactly 0.1 M solution of acid is used to titrate 25.00 ml of an exactly 0.1 M solution of alkali, with the following results (ml):

25.06 25.18 24.87 25.51 25.34 25.41

Test for positive bias in these results.

For these data we have:

$$\text{mean} = 25.228 \text{ ml, standard deviation} = 0.238 \text{ ml}$$

Adopting the null hypothesis that there is no bias, H_0: $\mu = 25.00$, and using equation (3.1) gives:

$$t = (25.228 - 25.00) \times \sqrt{6}/0.238 = 2.35$$

From Table A.2 the critical value is $t_5 = 2.02$ ($P = 0.05$, one-sided test). Since the observed value of t is greater than this, the null hypothesis is rejected and there is evidence for positive bias.

Using a computer gives $P(t \geq 2.35) = 0.033$. Since this is less than 0.05, the result is significant at $P = 0.05$, as before.

The same example in **R**

```
x=c(25.06, 25.18, 24.87, 25.51, 25.34, 25.41)
t.test(x,mu=25, alternative="greater")
```

F-test for comparison of standard deviations

We consider the null hypothesis:

$$H_0 : \quad \sigma_1^2 = \sigma_2^2.$$

and we want to detect if $\sigma_1^2 \neq \sigma_2^2$.

F-test for comparison of standard deviations

We consider the null hypothesis:

$$H_0 : \quad \sigma_1^2 = \sigma_2^2.$$

and we want to detect if $\sigma_1^2 \neq \sigma_2^2$. Alternatively, in one-sided version the alternative is $\sigma_1^2 > \sigma_2^2$.

F-test for comparison of standard deviations

We consider the null hypothesis:

$$H_0 : \quad \sigma_1^2 = \sigma_2^2.$$

and we want to detect if $\sigma_1^2 \neq \sigma_2^2$. Alternatively, in one-sided version the alternative is $\sigma_1^2 > \sigma_2^2$.
We take the ratio

$$F = s_1^2 / s_2^2$$

and consider this having the so-called F-distribution (Fisher's distribution) with $(n_1 - 1)$ and $(n_2 - 1)$ degrees of freedom.

F-distribution in **R**

```
n1=6
n2=12
r=seq(0.01,10,by=0.01)
plot(r,df(r,n1-1,n2-1))
```

Example 3.6.1

Example 3.6.1

A proposed method for the determination of the chemical oxygen demand of wastewater was compared with the standard (mercury salt) method. The following results were obtained for a sewage effluent sample:

	Mean (mg l^{-1})	Standard deviation (mg l^{-1})
Standard method	72	3.31
Proposed method	72	1.51

Example 3.6.1, cont.

For each method eight determinations were made.
(Ballinger, D., Lloyd, A. and Morrish, A. 1982. *Analyst* 107: 1047)

Is the precision of the proposed method significantly greater than that of the standard method?

We have to decide whether the variance of the standard method is significantly greater than that of the proposed method. F is given by the ratio of the variances:

$$F = \frac{3.31^2}{1.51^2} = 4.8$$

This is a case where a one-sided test must be used, the only point of interest being whether the proposed method is more precise than the standard method. In Table A.3 the number of degrees of freedom of the denominator is given in the left-hand column and the number of degrees of freedom of the numerator at the top. Both samples contain eight values so the number of degrees of freedom in each case is 7. The critical value is $F_{7,7} = 3.787$ $(P = 0.05)$, where the subscripts indicate the degrees of freedom of the numerator and denominator respectively. Since the calculated value of F (4.8) exceeds this, the variance of the standard method is significantly greater than that of the proposed method at the 5% probability level, i.e. the proposed method is more precise.

Example 3.6.1 in **R**

```
alpha=0.05
n1=8
n2=8

s1=3.31
s2=1.51
f=s1^2/s2^2

r=seq(0.01,10,by=0.01)
plot(r,df(r,n1-1,n2-1))
points(qf(1-alpha,n1-1,n2-1),0,pch="o",col="blue")
points(f,0,pch="*",col="red")
pvalue=1-pf(f,n1-1,n2-1)

pvalue
```

Example 3.3.1

In a comparison of two methods for the determination of chromium in rye grass, the following results (mg kg^{-1} Cr) were obtained:

Method 1: mean = 1.48; standard deviation 0.28
Method 2: mean = 2.33; standard deviation 0.31

For each method five determinations were made.
(Sahuquillo, A., Rubio, R. and Rauret, G. 1999. *Analyst* 124: 1)

Do these two methods give results having means which differ significantly?

The null hypothesis adopted is that the means of the results given by the two methods are equal. From equation (3.3), the pooled value of the standard deviation is given by:

$$s^2 = ([4 \times 0.28^2]) + [4 \times 0.31]^2)/(5 + 5 - 2) = 0.0873$$

$$s = 0.295$$

From equation (3.2):

$$t = \frac{2.33 - 1.48}{0.295\sqrt{\frac{1}{5} + \frac{1}{5}}} = 4.56$$

There are 8 degrees of freedom, so (Table A.2) the critical value $t_8 = 2.31$

Example 3.6.2

Example 3.6.2

In Example 3.3.1 it was assumed that the variances of the two methods for determining chromium in rye grass did not differ significantly. This assumption can now be tested. The standard deviations were 0.28 and 0.31 (each obtained from five measurements on a specimen of a particular plant). Calculating F so that it is greater than 1, we have:

$$F = \frac{0.31^2}{0.28^2} = 1.23$$

In this case, however, we have no reason to expect in advance that the variance of one method should be greater than that of the other, so a two-sided test is appropriate. The critical values given in Table A.3 are the values that F exceeds with a probability of 0.05, assuming that it must be greater than 1. In a two-sided test the ratio of the first to the second variance could be less or greater than 1, but if F is calculated so that it is greater than 1, the probability that it exceeds the critical values given in Table A.3 will be doubled. Thus these critical values are not appropriate for a two-sided test and Table A.4 is used instead. From this table, taking the number of degrees of freedom of both

Example 3.6.1 in **R**

```
alpha=0.05
n1=5
n2=5

s1=0.31
s2=0.28
f=s1^2/s2^2

r=seq(0.01,10,by=0.01)
plot(r,df(r,n1-1,n2-1))
points(c(qf(alpha/2,n1-1,n2-1),qf(1-alpha/2,n1-1,n2
points(f,0,pch="*",col="red")
pvalue=min(pf(f,n1-1,n2-1),1-pf(f,n1-1,n2-1))

pvalue
```

Outliers

Every experimentalist is familiar with the situation in which one (or possibly more) of a set of results appears to differ unreasonably from the others in the set. Such a measurement is called an **outlier**. In some cases an outlier may be attributed to a human error. For example, if the following results were given for a titration:

12.12, 12.15, 12.13, 13.14, 12.12 ml

then the fourth value is almost certainly due to a slip in writing down the result and should read 12.14. However, even when such obviously erroneous values have been removed or corrected, values which appear to be outliers may still occur. Should they be kept, come what may, or should some means be found to test statistically whether or not they should be rejected? Obviously the final values presented for the mean and standard deviation will depend on whether or not the outliers are rejected. Since discussion of the precision and accuracy of a method depends on these final values, it should always be made clear whether outliers have been rejected, and if so, why.

The ISO recommended test for outliers is **Grubbs' test**. This test compares the deviation of the suspect value from the sample mean with the standard deviation of the sample. The suspect value is the value that is furthest away from the mean.

Grubb's and Dixon's tests

In order to use Grubbs' test for an outlier, that is to test H_0 : all measurements come from the same population, the statistic G is calculated:

$$G = |\text{suspect value} - \bar{x}|/s \tag{3.8}$$

where $\bar{x}$ and s are calculated with the suspect value *included*.

The test assumes that the population is normal.

In order to use Dixon's test for an outlier, that is to test H_0 : all measurements come from the same population, the statistic Q is calculated:

$$Q = |\text{suspect value} - \text{nearest value}|/(\text{largest value} - \text{smallest value}) \tag{3.9}$$

This test is valid for samples size 3 to 7 and assumes that the population is normal.

Example 3.7.1

Example 3.7.1

The following values were obtained for the nitrite concentration (mg l^{-1}) in a sample of river water:

$$0.403,\ 0.410,\ 0.401,\ 0.380$$

The last measurement is suspect: should it be rejected?

The four values have $\bar{x} = 0.3985$ and $s = 0.01292$, giving

$$G = |0.380 - 0.3985|/0.01292 = 1.432$$

From Table A.5, for sample size 4, the critical value of G is 1.481 ($P = 0.05$). Since the calculated value of G does not exceed 1.481, the suspect measurement should be retained.

n fact, the suspect value in this data set would have to be considerably lower before : was rejected. It can be shown, using trial and error, that for the data set

$$0.403,\ 0.410,\ 0.401,\ b$$

where $b < 0.401$), the value of b would have to be as low as 0.356 before it was

Examples 3.7.2-3

Example 3.7.2

If three further measurements were added to those given in the example above so that the complete results became:

$$0.403, 0.410, 0.401, 0.380, 0.400, 0.413, 0.408$$

should 0.380 still be retained?

The seven values have $\bar{x} = 0.4021$ and $s = 0.01088$. The calculated value of G is now

$$G = |0.380 - 0.402|/0.01088 = 2.031$$

The critical value of G ($P = 0.05$) for a sample size 7 is 2.020, so the suspect measurement is now rejected at the 5% significance level.

Example 3.7.3

Apply Dixon's test to the data from the previous example.

$$Q = |0.380 - 0.400|/(0.413 - 0.380) = 0.606$$

The critical value of Q ($P = 0.05$) for a sample size 7 is 0.570. The suspect value 0.380 is rejected (as it was using Grubbs' test).

Outliers in **R**

The tests for outliers come in a contributed package called “outliers”. In order to use it one has to download the package to the computer. It can be done for the command line by using `install.package("outliers")` or can by using a convenient interface of the software.

```
x=c(0.403,0.410,0.401,0.380)

grubbs.test(x)

# Grubbs test for one outlier

#data:  x
#G = 1.4316, U = 0.0892, p-value = 0.09124
#alternative hypothesis: lowest value 0.38 is an ou
```

Outliers in **R**

```
dixon.test(x)

# Dixon test for outliers

#data:  x
#Q = 0.7, p-value = 0.1721
#alternative hypothesis: lowest value 0.38 is an ou

x=c(0.403,0.410,0.401,0.380,0.400,0.413,0.408)

grubbs.test(x)
dixon.test(x)
```

The chi-squared test for proportions

The chi-squared test for proportions

- Supposed that we have several categories of events that are observed.

The chi-squared test for proportions

- Supposed that we have several categories of events that are observed.
- Assume that n observations are made in total and there is k disjoint categories to which they can be classified.

The chi-squared test for proportions

- Supposed that we have several categories of events that are observed.
- Assume that n observations are made in total and there is k disjoint categories to which they can be classified.
- Let O_i represent the observed count in the ith category.

The chi-squared test for proportions

- Supposed that we have several categories of events that are observed.
- Assume that n observations are made in total and there is k disjoint categories to which they can be classified.
- Let O_i represent the observed count in the ith category.
- Let p_i be the proportion of the count that is expected in the ith category, so that the expected count for this category is $E_i = np_i$.

The chi-squared test for proportions

- Supposed that we have several categories of events that are observed.
- Assume that n observations are made in total and there is k disjoint categories to which they can be classified.
- Let O_i represent the observed count in the ith category.
- Let p_i be the proportion of the count that is expected in the ith category, so that the expected count for this category is $E_i = np_i$.
- Consider the quantity

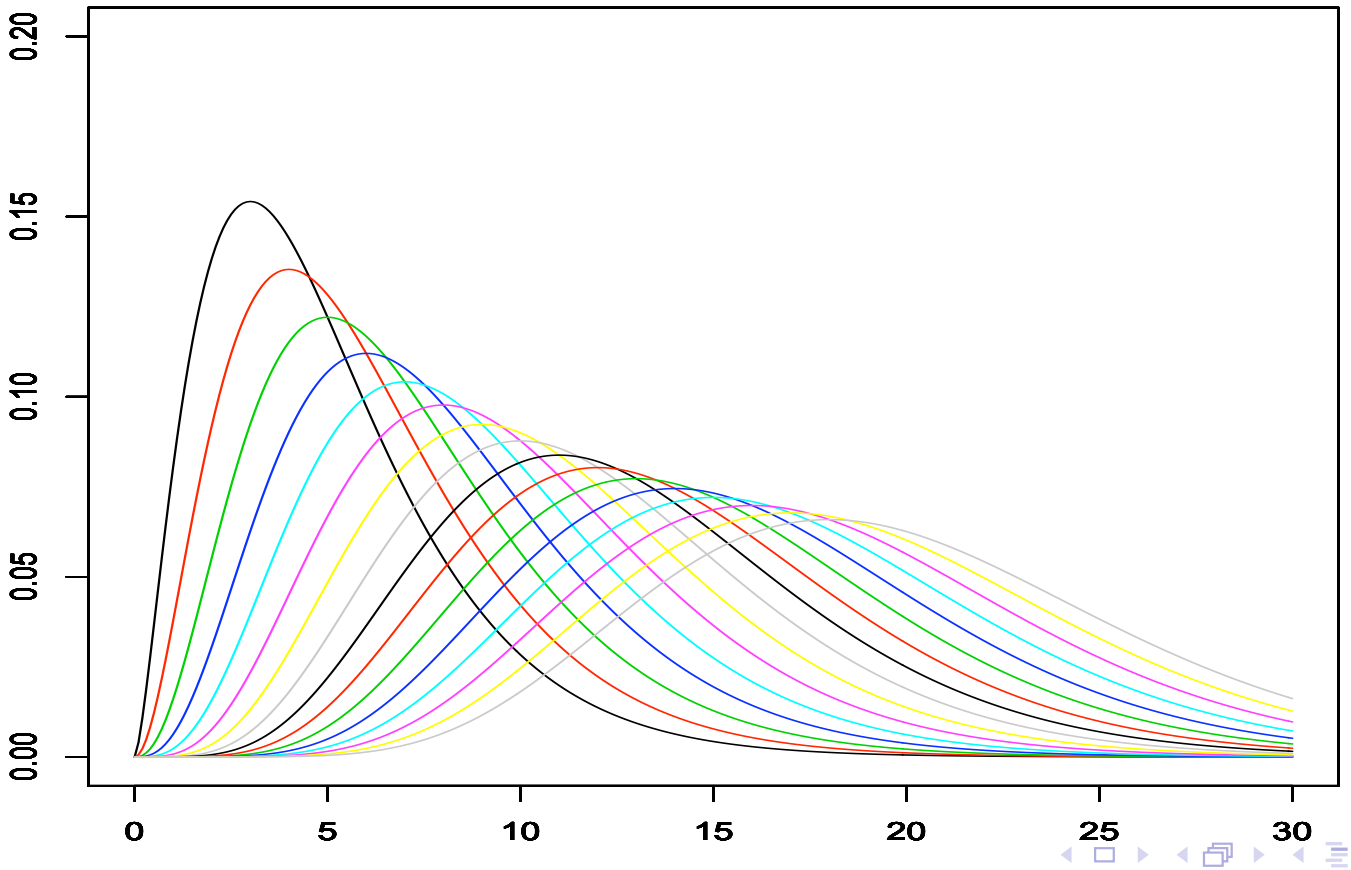
$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

and reject assumed proportions p_i if χ^2 is too large.

The chi-square distribution

The chi-square distribution

- It can be shown that for large sample sizes ($n > 50$) the variable χ^2 is distributed according to χ^2-distribution with $k - 1$ degrees of freedom.



Example 3.11.1

The numbers of glassware breakages reported by four laboratory workers over a given period are shown below. Is there any evidence that the workers differ in their reliability?

Numbers of breakages: 24, 17, 11, 9

The null hypothesis is that there is no difference in reliability. Assuming that the workers use the laboratory for an equal length of time, we would expect, from the null hypothesis, the same number of breakages by each worker. Since the total number of breakages is 61, the expected number of breakages per worker is $61/4 = 15.25$. Obviously it is not possible in practice to have a non-integral number of breakages: this number is a mathematical concept. The nearest practicable 'equal' distribution is 15, 15, 15, 16, in some order. The question to be answered is whether the difference between the observed and expected frequencies is so large that the null hypothesis should be rejected. That there should be *some* difference between the two sets of frequencies can be most readily appreciated by considering a sequence of throws of a die: we should, for example, be most surprised if 30 throws of a die yielded exactly equal frequencies for 1, 2, 3, etc. The calculation of χ^2 is shown below.

Observed frequency, O	Expected frequency, E	$O - E$	$(O - E)^2/E$
24	15.25	8.75	5.020
17	15.25	1.75	0.201
11	15.25	−4.25	1.184
9	15.25	−6.25	2.561
Totals	61	0	$\chi^2 = 8.966$

Report form **R**

```
qchisq(0.95,3)
[1] 7.814728

1-pchisq(8.966,3)
[1] 0.02974637
```

Testing for normality of distribution

The chi-square test could be used for testing normality by dividing the range of data into bins and compare the count in each bin with the corresponding probabilities based on the normal distribution.

Testing for normality of distribution

The chi-square test could be used for testing normality by dividing the range of data into bins and compare the count in each bin with the corresponding probabilities based on the normal distribution.
Unfortunately, one needs relatively large data sample size in order to use chi-squared test (> 50), thus there is a need for a small sample size procedure.

Normal probability plots

- One simple graphical way of comparing data to normal distributions is by plotting empirical quantile vs. corresponding normal quantile.

Normal probability plots

- One simple graphical way of comparing data to normal distributions is by plotting empirical quantile vs. corresponding normal quantile.
- Recall that the p-quantile for a given (empirical) distribution is the number below of which there is $100p\%$ of probability (of the data).

Normal probability plots

- One simple graphical way of comparing data to normal distributions is by plotting empirical quantile vs. corresponding normal quantile.
- Recall that the p-quantile for a given (empirical) distribution is the number below of which there is $100p\%$ of probability (of the data).

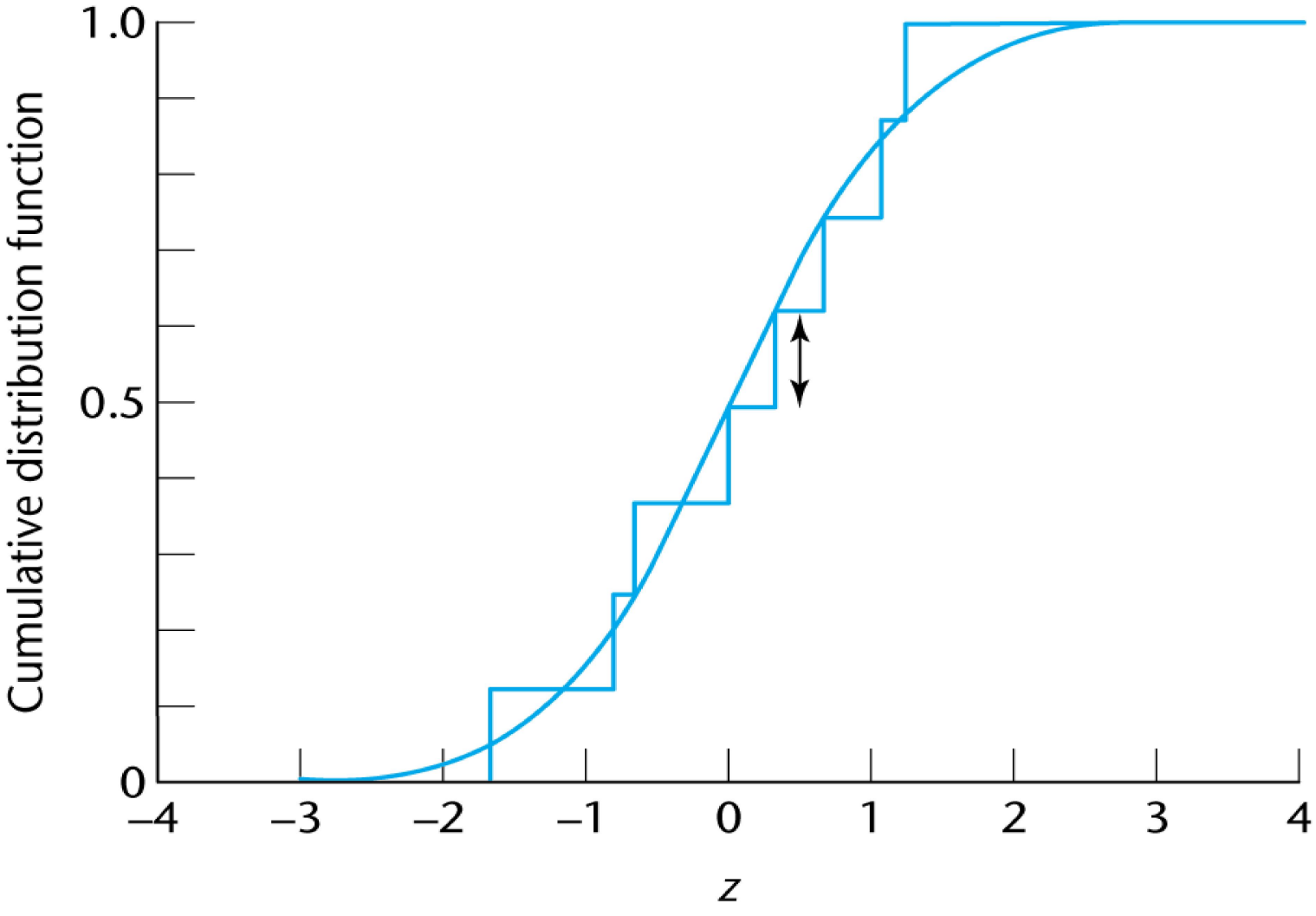
Consider data from Example 3.12.1

```
x=c(109,89,99,99,107,111,86,74,115,107,134,113,110,88,104)
```

The normal probability plot is obtained in **R** using

```
qqnorm(x)
```

Kolmogorov-Smirnov test



Example 3.12.1

Eight titrations were performed, with the results 25.13, 25.02, 25.11, 25.07, 25.03, 24.97, 25.14 and 25.09 ml. Could such results have come from a normal population?

First we estimate the mean and the standard deviation [with the aid of equations (2.1) and (2.2)] as 25.07 and 0.0593 ml respectively. The next step is to transform the x-values into z-values by using the relationship $z = (x - 25.07)/0.059$, obtained from equation (3.13). The eight results are thus transformed into 1.01, –0.84, 0.67, 0, –0.67, –1.69, 1.18 and 0.34. These z-values are arranged in order of increasing size and plotted as a stepped cumulative distribution function with a step height of $1/n$, where n is the number of measurements. Thus, in this case the step height is 0.125 (i.e. 1/8). (Note that this is not quite the same approach as that used in Example 3.12.1.) Comparison with the hypothetical function for z (Table A.2) indicates (Figure 3.6) that the maximum difference is 0.132 when $z = 0.34$. The critical values for this test are given in Table A.14. The table shows that, for $n = 8$ and $P = 0.05$, the critical value is 0.288. Since $0.132 < 0.288$ we can accept the null hypothesis that the data come from a normal population with mean 25.07 and standard deviation 0.059.

The value of this Kolmogorov–Smirnov test statistic, together with its P-value, can be obtained directly from Minitab in conjunction with a normal probability plot.

Kolmogorov-Smirnov test in **R**

```
x=c(25.13,25.02,25.11,25.07,25.03,24.97,25.14,25.09)

ks.test(x,"pnorm",mean(x),sd(x))

One-sample Kolmogorov-Smirnov test

data:  x
D = 0.1321, p-value = 0.995
alternative hypothesis: two-sided
```

Types of errors

- Consider the testing statistical hypothesis

$$H_0 : \mu = 3.0\%$$

against

$$H_1 : \mu = 3.5\%$$

Types of errors

- Consider the testing statistical hypothesis

$$H_0 : \mu = 3.0\%$$

against

$$H_1 : \mu = 3.5\%$$

- Significance level α (typically 5%) represents chances of rejecting H_0 given that it is true.

Types of errors

- Consider the testing statistical hypothesis

$$H_0 : \mu = 3.0\%$$

against

$$H_1 : \mu = 3.5\%$$

- Significance level α (typically 5%) represents chances of rejecting H_0 given that it is true.
- We refer to this as the probability of Type I Error.

Types of errors

- Consider the testing statistical hypothesis

$$H_0 : \mu = 3.0\%$$

against

$$H_1 : \mu = 3.5\%$$

- Significance level α (typically 5%) represents chances of rejecting H_0 given that it is true.
- We refer to this as the probability of Type I Error.
- The other error we can make is saying that we do not reject H_0 while in fact H_1 is true. This is called Type II Error.

Types of errors

- Consider the testing statistical hypothesis

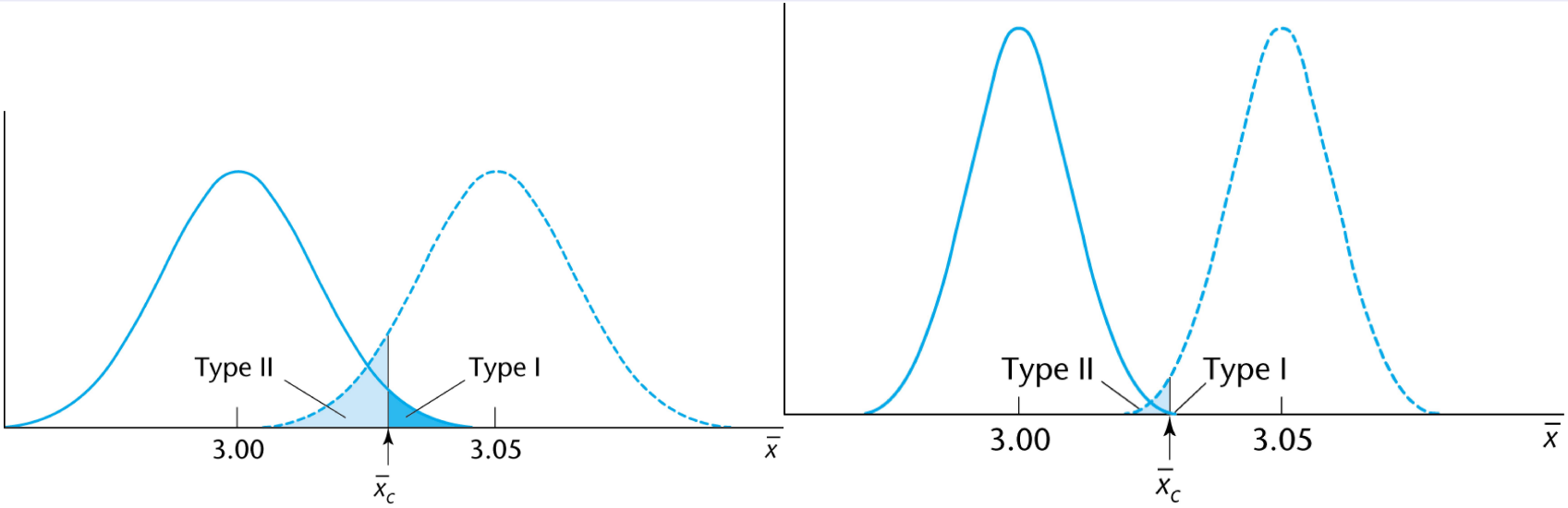
$$H_0 : \mu = 3.0\%$$

against

$$H_1 : \mu = 3.5\%$$

- Significance level α (typically 5%) represents chances of rejecting H_0 given that it is true.
- We refer to this as the probability of Type I Error.
- The other error we can make is saying that we do not reject H_0 while in fact H_1 is true. This is called Type II Error. One minus probability of making Type II Error is called the power of the test.

Calculation of the power



We reject H_0 if $t = (\bar{x} - 3.0)/(0.036/\sqrt{n}) > 1.96$, where $n = 4$, or equivalently if $\bar{x} > 0.036 * 1.96/\sqrt{n} + 3.0 = 3.035$.
If H_1 is true, then $\bar{x} > 3.035$ is given by chances that the normal variable that has mean 3.05 and variance 0.036/2 exceeds 3.035 which is

```
1-pnorm(3.035,3.05,0.036/2)
0.7976716
```

Thus the power of this test is about 80%.

The power and sample size

In general, the power depends on the sample size. We reject H_0 if $t = (\bar{x} - 3.0)/(0.036/\sqrt{n}) > 1.96$, where $n = 4$, or equivalently if $\bar{x} > 0.036 * 1.96/\sqrt{n} + 3.0$.
If H_1 is true, then $\bar{x} > 0.036 * 1.96/\sqrt{n} + 3.0$ is a function of n that can be plotted as follows

```
n=1:100
plot(n,1-pnorm(0.036*1.96/sqrt(n)+3.0,3.05,0.036/2),col="red")
```