Statitics of repeated measurements

Krzysztof Podgórski
Department of Mathematics and Statistics
University of Limerick

September 14, 2009

Titration experiment

Table 1.1 Random and systematic errors

Student	Results (ml)					Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise, unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased

This is also given in the text file `Table1_1.txt` contents of which is given below

```
A    10.08    10.11    10.09    10.10
B    9.88    10.14    10.02    9.80
C    10.19    9.79    9.69    10.05
D    10.04    9.98    10.02    9.97
```

Reading data from a file to **R**

```
#Reading the data from
Titra=read.table("Table1_1.txt", row.names = 1)

Titra

#Listing the first row
Titra[1,]

#and the last column
Titra[,5]
```

Mean and standard deviation

Student	Results (ml)				
A	10.08	10.11	10.09	10.10	10.12
B	9.88	10.14	10.02	9.80	10.21
C	10.19	9.79	9.69	10.05	9.78
D	10.04	9.98	10.02	9.97	10.04

Two criteria were used to compare these results, the average value (technically know as a measure of location) and the degree of spread (or dispersion). The average valu used was the **arithmetic mean** (usually abbreviated to the **mean**), which is the su of all the measurements divided by the number of measurements.

The mean, $\bar{x}$, of n measurements is given by $\bar{x} = \dfrac{\sum x_i}{n}$ (2.1)

In Chapter 1 the spread was measured by the difference between the highest ar lowest values (the **range**). A more useful measure, which utilizes all the values, is tl **standard deviation**, s, which is defined as follows:

The standard deviation, s, of n measurements is given by

$$s = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)} \qquad (2.2)$$

Means and standard deviations – counting on fingers

Example 2.1.1

Find the mean and standard deviation of A's results.

	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
	10.08	−0.02	0.0004
	10.11	0.01	0.0001
	10.09	−0.01	0.0001
	10.10	0.00	0.0000
	10.12	0.02	0.0004
Totals	50.50	0	0.0010

$$\bar{x} = \frac{\sum x_i}{n} = \frac{50.50}{5} = 10.1 \text{ ml}$$

$$s = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)} = \sqrt{0.001/4} = 0.0158 \text{ ml}$$

Note that $\sum(x_i - \bar{x})$ is always equal to 0.

Means and standard deviations much faster and better

```
#Comuting means

rowMeans(Titra)

#       A       B       C       D
#10.0950  9.9600  9.9300 10.0025

#and standard deviation

apply(Titra,1,sd)
#         A          B          C          D
#0.01290994 0.15055453 0.23036203 0.03304038
```

Bias and precision using mean and standard deviation

Bias and precision using mean and standard deviation

Task

Classify bias and precision using means and standard deviation of measurements.

Variance, coefficient of variation - relative standard deviation

Variance, coefficient of variation - relative standard deviation

- variance = squared standard deviation

Variance, coefficient of variation - relative standard deviation

- variance = squared standard deviation
- coefficient of variation = relative standard deviation (in percentage)

The empirical distribution of repeated measurements

The empirical distribution of repeated measurements

- frequency table

The empirical distribution of repeated measurements

- frequency table
- histogram and dotchart - graphical representation of the empirical distribution

Nitrate ion concentration from Table 2.1

Table 2.1 Results of 50 determinations of nitrate ion concentration, in μg ml^{-1}

0.51	0.51	0.51	0.50	0.51	0.49	0.52	0.53	0.50	0.47
0.51	0.52	0.53	0.48	0.49	0.50	0.52	0.49	0.49	0.50
0.49	0.48	0.46	0.49	0.49	0.48	0.49	0.49	0.51	0.47
0.51	0.51	0.51	0.48	0.50	0.47	0.50	0.51	0.49	0.48
0.51	0.50	0.50	0.53	0.52	0.52	0.50	0.50	0.51	0.51

Also in the file `Table2_1.txt`

```
0.51   0.51   0.51   0.50   0.51   0.49   0.52   0.53   0.50   0.47
0.51   0.52   0.53   0.48   0.49   0.50   0.52   0.49   0.49   0.50
0.49   0.48   0.46   0.49   0.49   0.48   0.49   0.49   0.51   0.47
0.51   0.51   0.51   0.48   0.50   0.47   0.50   0.51   0.49   0.48
0.51   0.50   0.50   0.53   0.52   0.52   0.50   0.50   0.51   0.51
```

The mean concentration

Reading data

```
#Getting data in a vector
x=scan("Table2_1.txt")
mean(x)
#[1] 0.4998
sd(x)
#[1] 0.01647385
```

Dotchart and histogram in **R**

```
#Dotchart
dotchart(x)
#Histogram and frequency table
Histogr=hist(x)
Histogr
```

The normal distribution – theoretical distribution of repeated measurements

The normal distribution – theoretical distribution of repeated measurements

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

The normal distribution – theoretical distribution of repeated measurements

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- The density represents distribution of probability for a random variable associated with it.

The normal distribution – theoretical distribution of repeated measurements

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- The density represents distribution of probability for a random variable associated with it.
- The area under the density represents the probability so the that the total area under it is equal to one.

The normal distribution – theoretical distribution of repeated measurements

It is not only the table values that can be explored for the standard normal distribution using **R**. Recall that the normal distribution is defined by the density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

- The density represents distribution of probability for a random variable associated with it.
- The area under the density represents the probability so the that the total area under it is equal to one.
- The area accumulated up to certain value z represents probability that a corresponding random variable takes value smaller than z and this probability defines the cumulative distribution function $F(z)$ which is tabularized.

Normal distribution in **R**

The following code explores various aspects of the standard normal distribution

```
#Plotting the density function of the standard normal variable

z=seq(-3,3,by=0.01)
plot(z,dnorm(z),type="l",col="red",lwd=4)

#Plotting the cumulative distribution function (that one from the table)

plot(z,pnorm(z),type="l",col="red",lwd=4)

#And plotting them one at the top of the other


par(mfrow=c(2, 1))

plot(z,dnorm(z),type="l",col="red",lwd=4)

plot(z,pnorm(z),type="l",col="red",lwd=4)

#Side by side

par(mfrow=c(1, 2))

plot(z,dnorm(z),type="l",col="red",lwd=4)

plot(z,pnorm(z),type="l",col="red",lwd=4)
```

Empirical vs. theoretical

Empirical vs. theoretical

- The theoretical one can be compared with empirical by taking μ equal to the sample mean $\bar{x}$ and σ equal to sample standard deviation s.

Empirical vs. theoretical

- The theoretical one can be compared with empirical by taking μ equal to the sample mean $\bar{x}$ and σ equal to sample standard deviation s.
- The following code compares empirical percentages with theoretical

```
quantile(x,c(0.16,0.84))

qnorm(c(0.16,0.84),mean(x),sd(x))
```

Example 2.2.1

Example 2.2.1

If repeated values of a titration are normally distributed with mean 10.15 ml and standard deviation 0.02 ml, find the proportion of measurements which lie between 10.12 ml and 10.20 ml.

Standardizing the first value gives $z = (10.12 - 10.15)/0.02 = -1.5$.
From Table A.1, $F(-1.5) = 0.0668$.

Standardizing the second value gives $z = (10.20 - 10.15)/0.02 = 2.5$.
From Table A.1, $F(2.5) = 0.9938$.

Thus the proportion of values between $x = 10.12$ to 10.20 (which corresponds to $z = -1.5$ to 2.5) is $0.9938 - 0.0668 = 0.927$.

No standardization needed in **R**

```
pnorm(c(10.12,10.20),10.15,0.02)
```

Not everything is normal, unfortunately – lognormal distribution

```
Concentr=scan("Figure2_5.txt")

hist(Concentr)
hist(Concentr,nclass=30)
```

Distribution of the sample mean

```
MatrConc=matrix(Concentr,ncol=4)
ConcM=rowMeans(MatrConc)

hist(ConcM)

MatrConc=matrix(Concentr,ncol=25)
ConcM=rowMeans(MatrConc)

hist(ConcM)
```

Two distributional effects of taking sample mean

Two distributional effects of taking sample mean

- Reduction in standard deviation (increased precision)

Two distributional effects of taking sample mean

- Reduction in standard deviation (increased precision)
- Distribution is becoming normal even if original is not

For a sample of n measurements,

$$\text{standard error of the mean (s.e.m.)} = \sigma/\sqrt{n} \qquad (2.5)$$

As expected, the larger n is, the smaller the value of the s.e.m. and consequently the smaller the spread of the sample means about μ.

The term 'standard error of the mean' might give the impression that $\sigma/\sqrt{n}$ gives the difference between μ and $\bar{x}$. This is not so: $\sigma/\sqrt{n}$ gives a measure of the variability of $\bar{x}$, as we shall see in the next section.

Another property of the sampling distribution of the mean is that, *even if the original population is not normal*, the sampling distribution of the mean tends to the normal distribution as n increases. This result is known as the **central limit theorem**. This theorem is of great importance because many statistical tests are performed on the mean and assume that it is normally distributed. Since in practice we can assume that distributions of repeated measurements are at least approximately normally distributed, it is reasonable to assume that the means of quite small samples (say >5) are normally distributed.