

Two-sample t-tests

Colin Aitken

University of Edinburgh

c.g.g.aitken@ed.ac.uk

<http://www.maths.ed.ac.uk/~cgga>



References

- Lucy, D. (2005) *Introduction to statistics for forensic scientists*. John Wiley and Sons Ltd.
- Maindonald, J. and Braun, J. (2003) *Data analysis and graphics using R*. Cambridge University Press.
- ‘It is easy to lie with statistics. It is hard to tell the truth without statistics.’ (Andrejs Dunkels)
- ‘... technology tends to overwhelm common sense.’ (D.A. Freedman)



Data set A

Table 1: 25 observations from a Normal distribution from mean 1.51674 and standard deviation 9×10^{-5}

1.516755	1.516604	1.516660	1.516608	1.516762
1.516740	1.516811	1.516745	1.516648	1.516753
1.516693	1.516821	1.516815	1.516676	1.516886
1.516768	1.516964	1.516655	1.516681	1.516771
1.516690	1.516795	1.516805	1.516652	1.516818



Data set *B*

Table 2: 25 observations from a Normal distribution from mean 1.51674 and standard deviation 9×10^{-5}

1.516720	1.516601	1.516725	1.516680	1.516661
1.516700	1.516741	1.516711	1.516680	1.516719
1.516688	1.516799	1.516767	1.516612	1.516726
1.516733	1.516729	1.516597	1.516600	1.516813
1.516792	1.516859	1.516941	1.516778	1.516716



Data set C

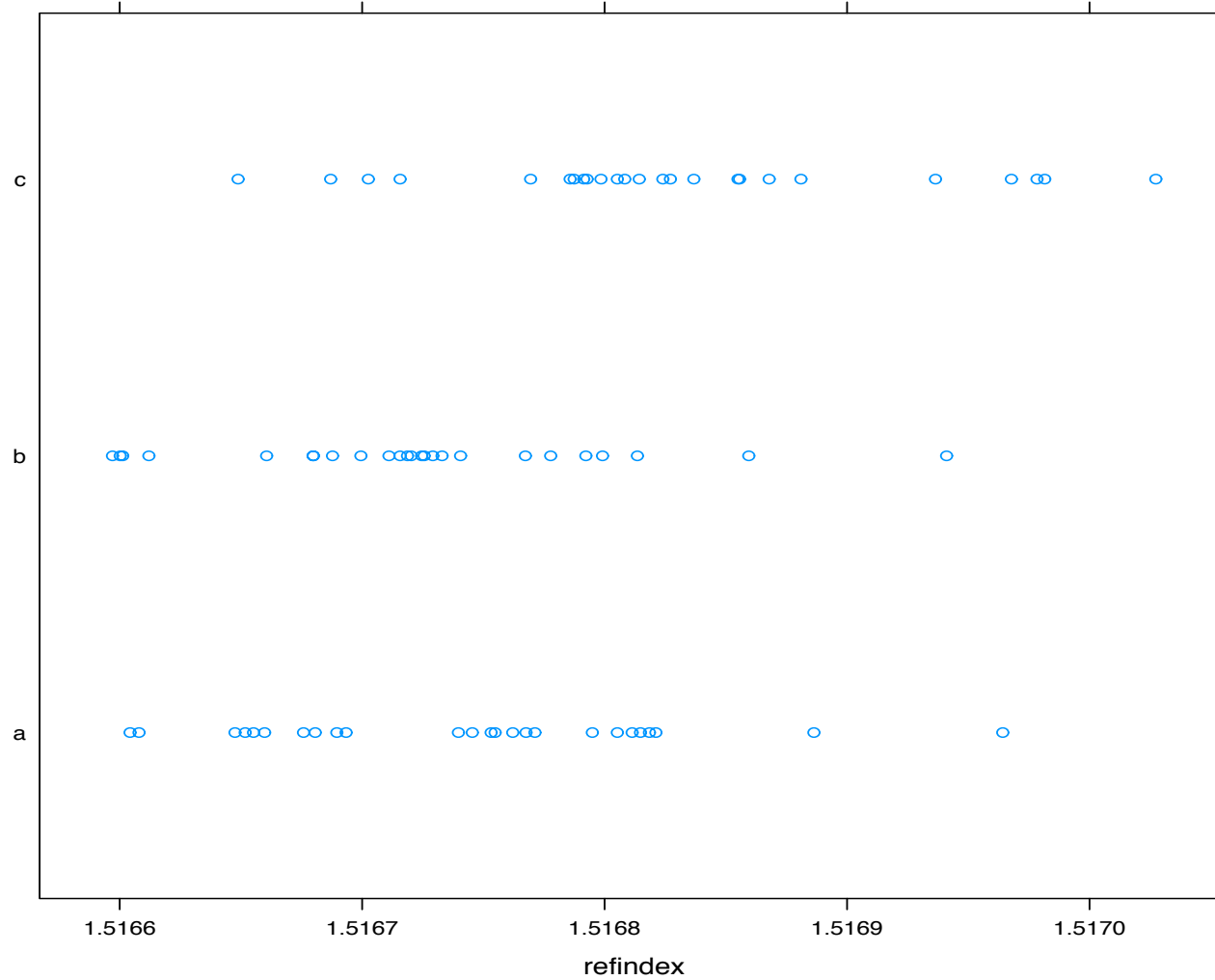
Table 3: 25 observations from a Normal distribution from mean 1.51682 and standard deviation 9×10^{-5}

1.516856	1.516798	1.516769	1.516793	1.516981
1.516808	1.516814	1.516978	1.516855	1.516805
1.516786	1.517027	1.516827	1.516703	1.516787
1.516649	1.516968	1.516716	1.516837	1.516824
1.516936	1.516687	1.516868	1.516881	1.516791



Dotplots

Dotplots of three data sets, .



Two-sample t-test

```
t.test(seta, setb, var.equal=T)
```

Two Sample t-test

```
data:  seta and setb t = 0.8209, df = 48, p-value = 0.4157
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.836589e-05  6.751187e-05
sample estimates: mean of x mean of y
 1.516743  1.516723
95% confidence interval is
```

$$(-2.836589 \times 10^{-5}, 6.751187 \times 10^{-5})$$

$$= (-0.00002836589, 0.00006751187).$$



Two-sample t-test

```
t.test(seta, setc, var.equal=T)
```

Two Sample t-test

```
data: seta and setc t = -3.3698, df = 48, p-value = 0.001492
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-1.385410e-04 -3.499646e-05
```

```
sample estimates: mean of x mean of y
```

```
1.516743 1.516830
```

95% confidence interval is

$$(-1.385410 \times 10^{-4}, -3.499646 \times 10^{-5})$$

$$= (-0.0001385410, -0.00003499646).$$



Two-sample t-test - background

- Two independent random samples of measurements.
- Measurements have a Normal distribution.
- The variation of the measurements in the underlying population or populations is the same.
- The question is whether the means of the two populations are the same.
- (Populations may also be known as sources.)



Two-sample t-test - forensic science background

- Control and recovered samples are independent and random samples from the same or from different populations.
- Fragments of glass of known source are known as control fragments. Fragments of glass of unknown source are known as recovered fragments.
- The question is whether the means of the two populations are the same.
- If it is decided the population means are equal then there is an inference that the populations are the same - the control and recovered samples come from the same source.



Comments on random samples

Comments on sampling from Evett and Weir *Interpreting DNA evidence* (1998)

“Of course, a real crime laboratory would not attempt ... to take a random, representative, stratified sample of individuals to address the question of issue. In the vast majority of cases the laboratory will have one or more *convenience* samples. Such a sample may be of laboratory staff members, or from blood donor samples with the cooperation of a local blood bank, or from samples from victims and suspects examined in the course of casework.

“...(I)n the forensic context, we will generally be dealing, not with random but, with convenience samples. Does this matter? The first response to that question is that every case must be treated according to the circumstances within which it has occurred, and the next response is that it is always a matter of judgement. ... In the last analysis, **the scientist must also convince a court of the reasonableness of his or her inference within the circumstances as they are presented as evidence.**”(pp. 44-45.)



Two-sample t-test - notation

- Control measurements are denoted x_1, \dots, x_m .
- Recovered measurements are denoted y_1, \dots, y_n .
- The mean of the population from which the control measurements were taken is μ_x .
- The mean of the population from which the recovered measurements were taken is μ_y .
- If the control and recovered measurements come from the same source then $\mu_x = \mu_y$.
- The variances of the measurements from the one or two sources are assumed equal and denoted σ^2 .



Two-sample t-test - notation continued

- The sample mean of the control measurements is

$$\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i.$$

- The sample mean of the recovered measurements is

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j.$$

- The sample variance of the control measurements is

$$s_x^2 = \frac{1}{(m-1)} \sum_{i=1}^m (x_i - \bar{x})^2.$$

- The sample variance of the recovered measurements is

$$s_y^2 = \frac{1}{(n-1)} \sum_{j=1}^n (y_j - \bar{y})^2.$$



Two-sample t-test - summary calculations: A v C

- $\bar{x} = 1.516743.$

- $\bar{y} = 1.516830.$

- $s_x^2 = 7.624648e - 09 = 7.624648 \times 10^{-9}.$

- $s_y^2 = 8.950903e - 09 = 8.950903 \times 10^{-9}.$

The two sample variances s_x^2 and s_y^2 are taken to be estimates of the same population variance σ^2 .



Standard error - single sample

The *standard error* is defined as the standard deviation of the sampling distribution of the statistic. A special case is the standard error of the mean *SEM*. The sampling distribution of the mean is the distribution of means from repeated independent random samples. Consider a sample of size n with standard deviation s . The estimate of the *SEM* is

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

The *SEM* indicates the extent to which the sample mean may be expected to vary from one sample to another and may be thought of as a measure of precision. The variation decreases as the square root of the sample size increases. Thus, to double the precision of an estimate of a mean it is necessary to double the sample size.



Standard error - two samples

Comparison of the means of two different samples, which may or may not be from the same population. With two independent samples of sizes m and n the comparison is in the form of a difference

$$\bar{x} - \bar{y}$$

where \bar{x} and \bar{y} denote the respective sample means. Denote the corresponding standard errors by SEM_x and SEM_y . The standard error of the difference SED is computed using the formula

$$SED = \sqrt{SEM_x^2 + SEM_y^2}$$

A reasonable assumption is that the standard deviations in the populations from which the samples are drawn are equal, with common s.d. s .

Then

$$SEM_x = \frac{s}{\sqrt{m}}, \quad SEM_y = \frac{s}{\sqrt{n}}$$

and

$$SED = s\sqrt{\frac{1}{m} + \frac{1}{n}}.$$



Two-sample t-test - estimate of variance

The best estimate s^2 of σ^2 using s_x^2 and s_y^2 is given by

$$s^2 = \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2}.$$



Test for equality of variance: A v C

- $s_x^2 = 7.624648e - 09 = 7.624648 \times 10^{-9}$.

- $s_y^2 = 8.950903e - 09 = 8.950903 \times 10^{-9}$.

Test the equality of the variances by taking their ratio, the larger divided by the smaller. The null hypothesis is that the two population variances are equal. If the data are Normally distributed and the null hypothesis is true then the ratio has an F -distribution with a pair of degrees of freedom $(n_1 - 1, n_2 - 1)$ where (n_1, n_2) are the sample sizes of the data sets for the numerator and denominator, respectively.



Test for equality of variance

Test of equality of variance for sets A and C :

Variance for Set $A = 7.624648e - 09$

Variance for Set $C = 8.950903e - 09$

The ratio of higher to lower = $8.95093/7.624648 = 1.173947$.

The significance probability is determined with reference to an F -distribution with $(24, 24)$ degrees of freedom. The significance probability is $0.3488036 = 0.35$.

This is the probability of obtaining a value for the ratio of two sample variances, both estimated with 24 degrees of freedom, from Normal populations with the same variance, as large as 1.174.



Two-sample t-test

Estimate of common variance s^2 is

$$\begin{aligned}s^2 &= \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} \\ &= \frac{24 \times 7.624648 \times 10^{-9} + 24 \times 8.950903 \times 10^{-9}}{24 + 24} \\ &= \frac{7.624648 \times 10^{-9} + 8.950903 \times 10^{-9}}{2} \\ &= 8.287775 \times 10^{-9} \\ s &= 0.00009103722.\end{aligned}$$



Two-sample t-test

Test statistic

$$\begin{aligned}t &= \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{m} + \frac{1}{n}}} \\ &= \frac{1.516743 - 1.516830}{0.00009103722 \sqrt{\left(\frac{1}{25} + \frac{1}{25}\right)}} \\ &= \frac{-0.000087}{0.00002574921} \\ &= -3.369762.\end{aligned}$$



Normal theory approximation

From Maindonald and Braun:

“For random samples from a distribution that is close to symmetric, the approximation is often adequate, even for samples as small as 3 or 4. In practice, we may know little about the population from which we are sampling. even if the main part of the population distribution is symmetric, occasional aberrant values are to be expected. **Such aberrant values ... make it more difficult to detect genuine differences.** The take-home message is that, especially in small samples, the probabilities and quantiles can be quite imprecise. They are rough guides, intended to assist researchers in making a judgement.”

In forensic science, when testing, a genuine difference is indicative of different sources which is something that we wish to be less difficult to detect, not ‘more difficult’.



Two-sample t-test - interpretation

```
t.test(seta, setc, var.equal=T)
```

Two Sample t-test

```
data: seta and setc t = -3.3698, df = 48, p-value = 0.001492
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.385410e-04 -3.499646e-05
sample estimates: mean of x mean of y
 1.516743  1.516830
95% confidence interval is
```

$$(-1.385410 \times 10^{-4}, -3.499646 \times 10^{-5})$$

$$= (-0.0001385410, -0.00003499646).$$



Two-sample t-test - derivation

- Confidence intervals for a mean difference, or a difference of means, have the form difference $\pm t$ -critical value \times standard error of the difference.
- The t -statistic has the form

$$t = \frac{\text{difference}}{\text{standard error of difference}}.$$

- Given t , the p -value, or significance probability for a two-sided test is defined as

$$Pr(T > t) + Pr(T < -t)$$

where T has a t -distribution with the appropriate number of degrees of freedom. **A small p -value corresponds to a large value of $|t|$, and is regarded as evidence that the true difference is non-zero and this leads to rejection of the *null hypothesis*.**



What is a small p -value?

- At what point is a p -value small enough to be convincing? Conventionally, $p = 0.05 (= 5\%)$ is used as the cut-off.
- However, 0.05 is too large if results from the experiment are to be made the basis for a recommendation for changes to farming practice or to medical treatment.
- 0.05 may be too small when deciding which effects merit further experimental or other investigation.
- In forensic science, hypothesis testing is suggestive of a principle that a person is guilty until proven innocent: the null hypothesis is that the control and recovered samples come from the same source.



Confidence intervals

Often, an interval is wanted that most often, when samples are taken in the way that our samples are taken, will include the population mean. There are two common choices for the long run proportion of similar samples that should contain the population mean: 95% and 99%. A later example uses 90%.

Confidence intervals can be used as the basis for tests of hypotheses. If the confidence interval for the population mean does not contain zero the hypothesis that the population mean is zero is rejected.



Confidence intervals

The general form for a confidence interval is

difference $\pm t$ -critical value \times standard error of the difference.

$$\bar{x} - \bar{y} \pm t_{m+n-2}(0.025) s \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)}$$

where $t_{m+n-2}(0.025)$ is the 2.5% point of the t -distribution with $(m + n - 2)$ degrees of freedom; *i.e.*, the point of the t -distribution such that

$$Pr(T > t(0.025)) + Pr(T < -t(0.025)) = 0.05; \quad t(0.025) > 0.$$



Degrees of freedom

The 'degrees of freedom' is the number associated with the estimation of the variance; the denominator in the formulae for the variance.

Consider a random sample of size n . The denominator of the estimate of the variance is $(n - 1)$, hence this is the degrees of freedom for any associated t -test. Alternatively, there are initially n independent observations. Then the sample mean is estimated. This leaves $(n - 1)$ independent observations.

For a two-sample test, there are $m + n$ independent observations. Two sample means are estimated. There are then $(m + n - 2)$ independent observations left.



Confidence interval example

$$\begin{aligned} & \bar{x} - \bar{y} \pm t_{m+n-2}(0.025)s\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)} \\ &= (1.516743 - 1.516830) \pm 2.010635 \times 0.00009103722\sqrt{\left(\frac{1}{25} + \frac{1}{25}\right)} \\ &= -0.000087 \pm 2.010635 \times 0.00002574921 \\ &= -0.000087 \pm 0.00005177226 \\ &= (-0.000139, -0.00003523) \end{aligned}$$



Unequal variances

A procedure known as the *Welch* procedure is used. Details are given in Curran, Hicks and Buckleton (2000) *Forensic interpretation of glass evidence* (CRC Press).

The Welch statistic is the difference in means divided by the standard error of the difference:

$$t = (\bar{x} - \bar{y}) / SED;$$

$$SED = \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}.$$



Unequal variances continued

If the two variances are unequal this does not have a t -distribution but critical values are quite well approximated by a t -distribution with degrees of freedom given by a suitable approximation. Note also that if $m = n$ the statistics is the same as for the t -test based on the pooled estimate of variance. The degrees of freedom are reduced.

$$\begin{aligned} SED &= \sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}} \\ &= \sqrt{\frac{s_x^2 + s_y^2}{n}} \\ &= \sqrt{\frac{s_x^2 + s_y^2}{2}} \sqrt{\frac{2}{n}}. \end{aligned}$$



Unequal variances continued

Let $\nu_x = s_x^2/m$ with $\eta_x = (m - 1)$ degrees of freedom and $\nu_y = s_y^2/n$ with $\eta_y = (n - 1)$ degrees of freedom. The approximate degrees of freedom for Welch's test is then

$$\eta = \frac{(\nu_x + \nu_y)^2}{\frac{\nu_x^2}{\eta_x} + \frac{\nu_y^2}{\eta_y}}.$$

The modified statistic is the default in R and is an option in Excel.



Paired t-test

Example from Finkelstein and Levin *Statistics for Lawyers*, Second edition, Springer, 2001.

USA: The federal Clean Air Act requires that before a new fuel or fuel additive is sold in the United States, the producer must demonstrate that the emission products generated will not cause a vehicle to fail to achieve compliance with certified emission standards. To estimate the difference in emission levels, the EPA requires, among other tests, a Paired-difference test in which a sample of cars is first driven with the standard fuel, and then with the new fuel and the emission levels compared. EPA then constructs a 90% confidence interval for the average difference. If the interval includes 0 the new fuel is eligible for a waiver.



Paired t-test - data

Nitrous Oxide emissions for sixteen cars driven first with a standard fuel (B) and then with Petrocoal P , a gasoline with a methanol additive; Finkelstein and Levin, p. 227.

Note the test is a paired test, the same 16 cars are used for the two types of fuel. It would be possible to use two sets of 16 cars but then differences in emissions may have been due to differences in cars as well as in differences in the effects of fuel.



Paired t-test - data

Nitrous Oxide emissions for sixteen cars driven first with a standard fuel (B) and then with Petrocoal P , a gasoline with a methanol additive; Finkelstein and Levin, p. 227

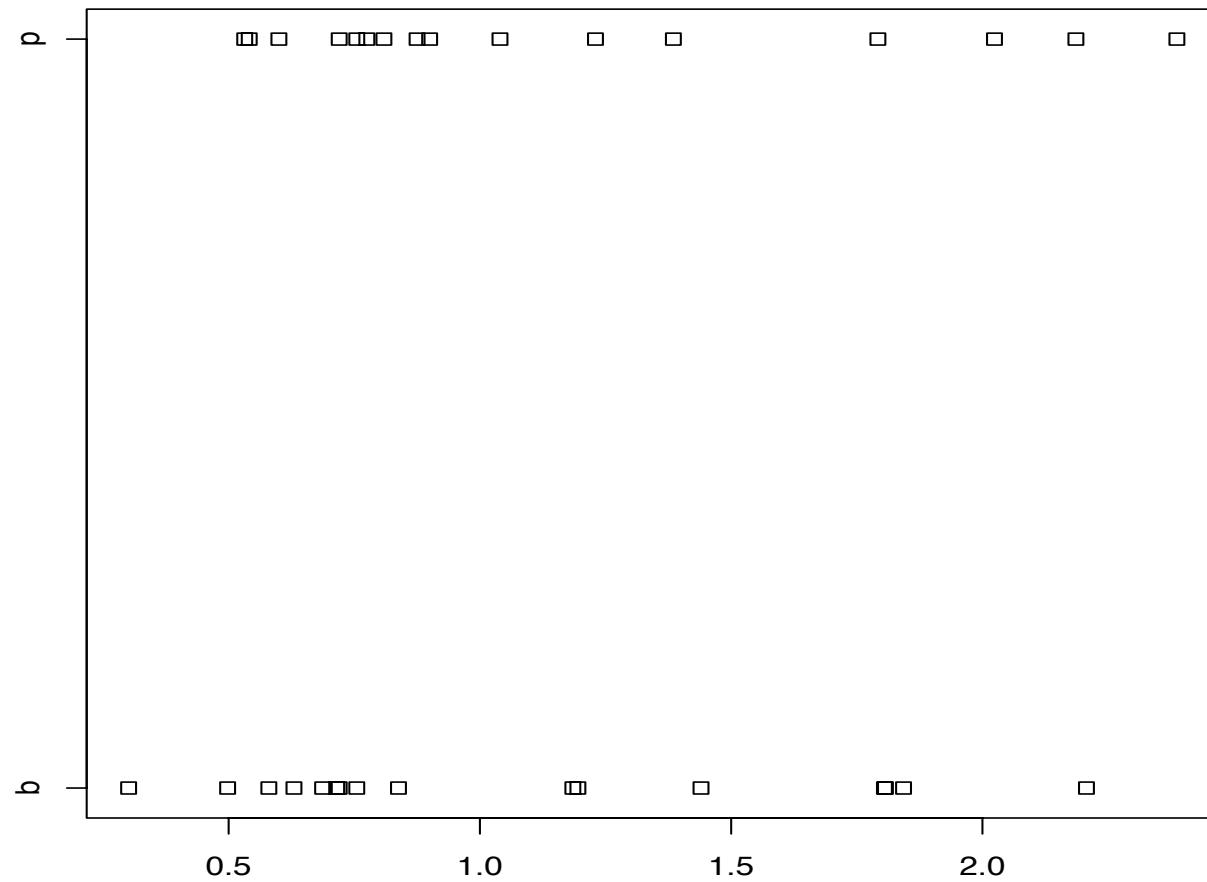
Table 4: Emission data for NO_x

B	1.195	1.185	0.755	0.715	1.805	1.807	2.207	0.301
P	1.385	1.230	0.755	0.775	2.024	1.792	2.387	0.532
$(P - B)$	+0.190	+0.045	0.000	+0.060	+0.219	-0.015	+0.180	+0.231
Sign	+	+	tie	+	+	-	+	+
B	0.687	0.498	1.843	0.838	0.720	0.580	0.630	1.440
P	0.875	0.541	2.186	0.809	0.900	0.600	0.720	1.040
$P - B$	+0.188	+0.043	+0.343	-0.029	+0.180	+0.020	+0.090	-0.400
Sign	+	+	+	-	+	+	+	-



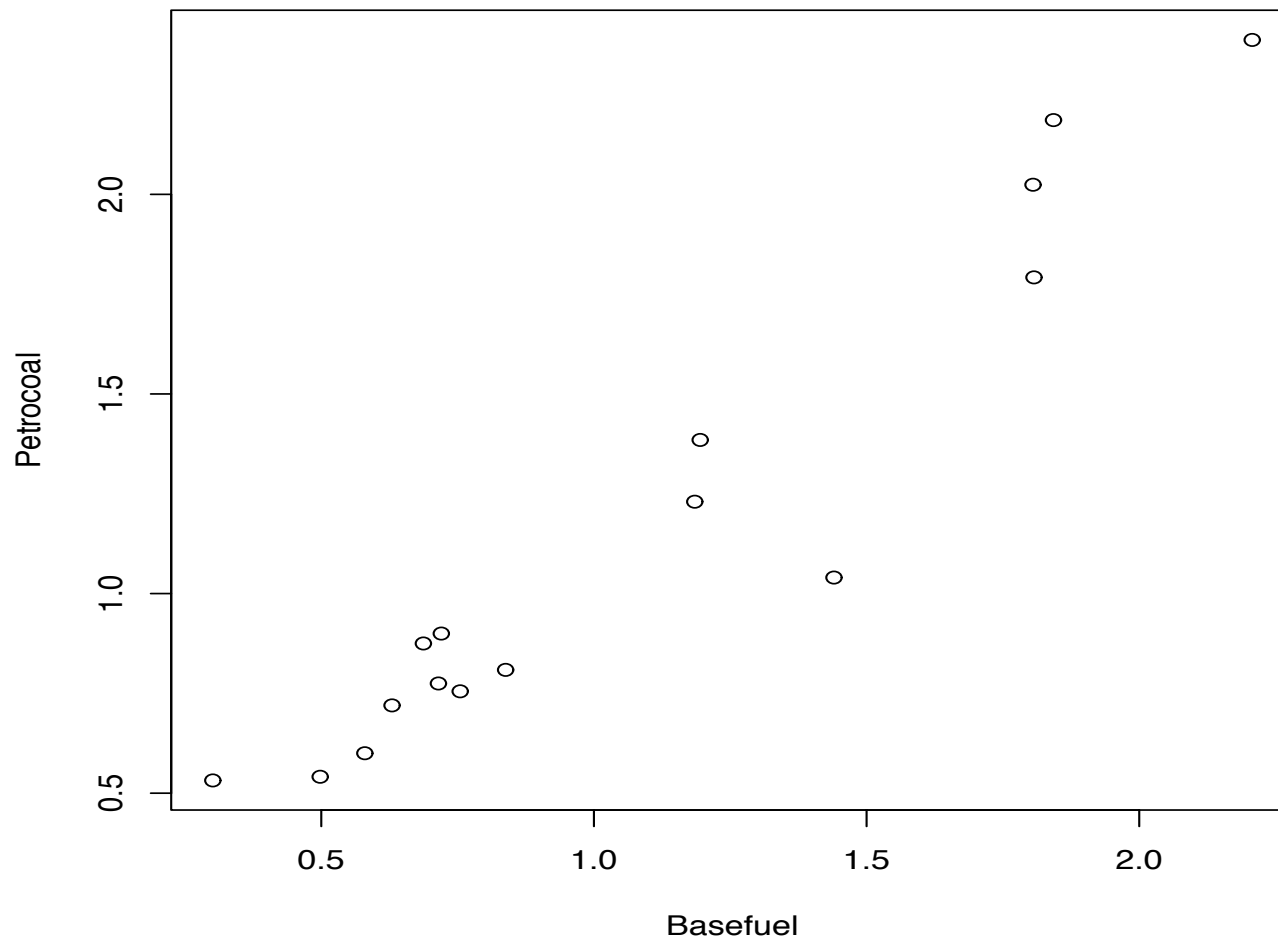
Paired t-test - dotplots

Dotplots of emission data



Paired t-test - scatter plot

Scatterplots of emission data



Two-sample t-test: WRONG!

```
t.test(nitrousoxide$Basefuel, nitrousoxide$Petrocoal,  
var.equal=T, conf.level=0.90)
```

Two Sample t-test

```
data: nitrousoxide$Basefuel$ and nitrousoxide$Petrocoal$  
t = -0.3984, df = 30, p-value = 0.6931
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
90 percent confidence interval:  
-0.4421429  0.2740179
```

```
sample estimates:  
mean of x mean of y  
1.075375  1.159438
```



Paired-sample t-test

```
> nitrousdiff
 [1]  0.190  0.045  0.000  0.060  0.219 -0.015  0.180  0.231  0.188  0.04
[11]  0.343 -0.029  0.180  0.020  0.090 -0.400
> nitrousdiff.n
 [1] 16
> nitrousdiff.se <- sd(nitrousdiff)/sqrt(nitrousdiff.n)
> nitrousdiff.se
 [1] 0.04179269
> mean(nitrousdiff)
 [1] 0.0840625

> mean(nitrousdiff) + qt(c(0.05, 0.95),15)*nitrousdiff.se

 [1] 0.01079781 0.15732719
```

The 90% confidence interval does not contain zero. There is evidence that emission data for NO_x for P is greater than for B .



Paired-sample t-test

```
mean(nitrousdiff)/nitrousdiff.se [1] 2.011416
```

```
> 2*(1 -pt(2.04116, 15))
```

```
[1] 0.05923676
```

The 90% confidence interval does not contain zero. Thus the corresponding two-sided test for zero difference is rejected at the 10% level. Significance probability is 0.06 or 6%. There is evidence that emission data for NO_x for P is greater than for B .



Summary - 1

From M&B: ‘The formal methodology of hypothesis testing may seem contorted. A small p -value makes the null hypothesis appear implausible. It is not a probability statement about the null hypothesis itself, or for that matter its alternative. All it offers is an assessment of implications that flow from accepting the null hypothesis. A straw man is set up that $\mu = 0$. The typical goal is to knock down this straw man. By its very nature, hypothesis testing lends itself to various abuses.’



Summary - 2

How to choose between a two-sample test and a paired test:

- If sample sizes m, n are unequal, test is two-sample one.
- If sample sizes m, n are equal test may be either two-sample or paired.
- If re-ordering one sample and not the other makes no difference to the statistical analysis the test is a two-sample one.
- If there is a natural pairing, the test is a paired one

