Estimation and Hypothesis Testing for Continuous Data

MAST90044

Thinking and Reasoning with Data

Dr Julia Polak

Chapter 5

School of Mathematics & Statistics
The University of Melbourne

Outline

Inference on the mean

- Introduction
- examples
- Hypothesis tests — comments

Inference on difference of means

- Independent samples
- Independent samples: assuming equal σ
- Independent samples: different σ
- Paired samples

Inference on the mean of a Normal population

The sample mean has a dual role:

(1) it indicates the centre of the sample distribution (& used for CI) and

(2) it gives an estimate of the population mean.

Investigate (2) here:

The population mean μ is unknown, and we wish to use the data (a random sample on X) to estimate it.

Inference on the mean of a Normal population

▷ **Example.** A random sample of $n = 400$ observations is to be obtained from a population with standard deviation $\sigma = 10$. If we observed the sample mean, $\bar{x} = 50.8$, what are plausible values for the unknown population mean μ?
In practice, μ is unknown, and we want to use $\bar{x}$ to estimate it.

* **point estimate**, $\hat{\mu} = \bar{x}$ (one-number guess, e.g. 50.8)

We know that μ will be “around” 50.8. $[\mu \neq \bar{x}$, but $\mu \approx \bar{x}]$

A point estimate is not enough! We want to know how close.

Inference on the mean of a Normal population

* **interval estimate** = confidence interval

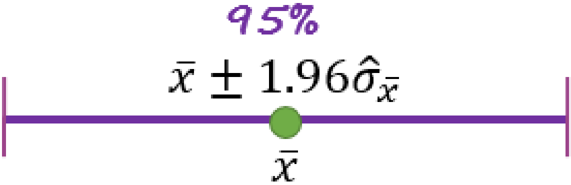
This specifies an interval of "plausible values" for μ,

We have $\bar{X} \overset{\mathrm{d}}{\approx} \mathrm{N}(\mu, \frac{100}{400})$,
$\mathrm{var}(\bar{X}) = \frac{\sigma^2}{n} = \frac{10^2}{400} = 0.25$, so that $\mathrm{sd}(\bar{X}) = 0.5$.

The (95%) confidence interval for μ is
$50.8 \pm 1.96 \times 0.5 = 50.8 \pm 0.98$ i.e. $(49.82, 51.78)$

So ("95%") plausible values for μ are $(49.82 < \mu < 51.78)$.

i.e. which values of μ could plausibly have led to the observed value $\bar{x} = 50.8$.


95%
$\bar{x} \pm 1.96\hat{\sigma}_{\bar{x}}$
$\bar{x}$

Inference on the mean of a Normal population

random sample:

$$X_i \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \sigma), \quad i = 1, \ldots, n$$

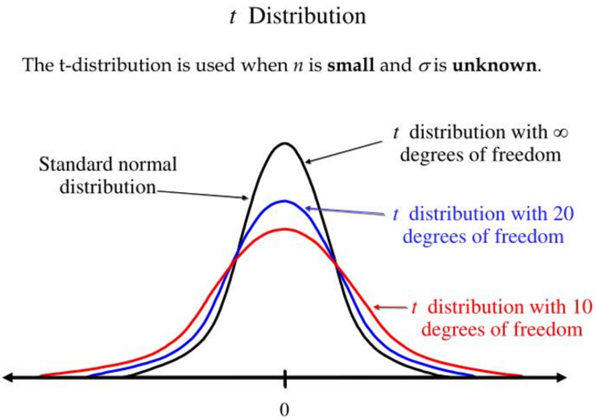
$$\bar{X} \overset{\mathrm{d}}{=} \mathrm{N}(\mu, \tfrac{\sigma}{\sqrt{n}}) \qquad (\bar{X} \text{ is an estimator for } \mu, \; sd(\bar{X}) \text{ measures imprecision})$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \overset{\mathrm{d}}{=} \mathrm{N}(0, 1).$$

But we need to estimate σ from the data. We use s. This means we lose some precision. And a modification of the distribution.

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{\mathrm{d}}{=} \mathrm{t}_{n-1}. \qquad (se = \hat{sd}(\bar{X}) = \tfrac{\hat{\sigma}}{\sqrt{n}}, \; \hat{\sigma} = S)$$

The t distribution



t Distribution

The t-distribution is used when n is **small** and σ is **unknown.**

Source: slideserve.com

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

A 95% confidence interval for μ is

$$\bar{x} \pm t_{n-1}^{0.975} \times \frac{s}{\sqrt{n}}$$

where $t_{n-1}^{0.975}$ is the 0.975 quantile of the t-distribution on $n-1$ df.

A $1-\alpha$ confidence interval for μ is

$$\bar{x} \pm t_{n-1}^{1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

$(se(\bar{X}) = \hat{sd}(\bar{X}) = \frac{\hat{\sigma}}{\sqrt{n}}, \ \hat{\sigma} = S)$

Inference on the mean of any population

random sample:

$$X_i \overset{\mathrm{d}}{=} \mathbb{D}(\mu, \sigma), \quad i = 1, \ldots, n$$

$$\bar{X} \overset{\mathrm{d}}{\approx} \mathrm{N}(\mu, \tfrac{\sigma}{\sqrt{n}})$$

for large* *n*; CLT

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \overset{\mathrm{d}}{\approx} \mathrm{N}(0, 1).$$

for just about any population!

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \overset{\mathrm{d}}{\approx} \mathrm{t}_{n-1}.$$

for roughly normal population.

but if *n* large: $s \approx \sigma$, $\mathrm{t}_{n-1} \approx \mathrm{N}$

problem only if *n* small and $\mathbb{D}$ very different from N.

Inference on the mean

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{\mathrm{d}}{\approx} \mathrm{t}_{n-1}.$$

Hypothesis test:

if $\mu = \mu_0$,

then $t - stat = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ is an observation on t_{n-1}.

P-value $= 2\,\mathrm{Pr}(\mathrm{t}_{n-1} > t)$ (for positive t)

Hypothesis testing

Soil pH: 17 samples from a field.

```
> pH <- c(6.0, 5.7, 6.2, 6.3, 6.5, 6.4, 6.9, 6.6, 6.8,
+          6.7, 6.8, 7.1, 6.8, 7.1, 7.1, 7.5, 7.0)
```

- Suppose we want the soil to be neutral, i.e. for the pH to be 7.
- Should we add some chemicals to change the pH of the soil?
- Is there sufficient evidence that the mean pH of the soil is different from 7?

We have 3 ways to answer the last question.

Hypothesis testing

- 95% confidence interval for μ is $(6.44, 6.91)$.

  So if $\mu = 7$ we have observed something unusual.

- Observed value of the test statistic is
  $$t = \frac{6.676 - 7}{0.455/\sqrt{17}} = -2.9326 \;\; (\bar{x} \text{ differs from } \mu_0 \text{ by about } 3\,\text{se}).$$
  The critical value is $t_{16}^{0.025} = -2.11$, it is smaller so we reject H_0.

- P-value could be found in R by: `> 2*pt(-2.9326, df=16)`, which gives `0.009758`; i.e. $P = 0.010$; and we reject H_0.

- There is significant evidence here that $\mu < 7$: $t_{16}^{0.025} = -2.11$, $P = 0.010$; 95% CI: $(6.44, 6.91)$.

Soil pH: 17 samples from a field.

```
> pH <- c(6.0, 5.7, 6.2, 6.3, 6.5, 6.4, 6.9, 6.6, 6.8,
+         6.7, 6.8, 7.1, 6.8, 7.1, 7.1, 7.5, 7.0)
> mean(pH); sd(pH)
[1] 6.67647
[1] 0.4548755
> mean(pH)+ qt(c(0.025,0.975),length(pH)-1)
+     *sd(pH)/sqrt(length(pH))
[1] 6.442595 6.910346
> t.test(pH,mu=7)
  One Sample t-test
data:  pH
t = -2.9326, df = 16, p-value = 0.009758
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
 6.442595 6.910346
sample estimates:
mean of x
  6.67647
```

Confidence Interval:

$\bar{x} = 6.676,$
$\hat{\sigma} = 0.455/\sqrt{17}$

$H_0: \mu = \mathbf{7}$
$H_1: \mu \neq 7$

Hypothesis testing :

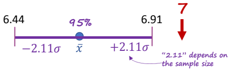
$$t = \frac{6.676 - 7}{0.455/\sqrt{17}} = -2.9326$$

P-value:

$p - value = 0.010$

Confidence Interval:

$\bar{x} = 6.676,$
$\hat{\sigma} = 0.455/\sqrt{17}$



$H_0\colon \mu = 7$
$H_1\colon \mu \neq 7$

Hypothesis testing :

$$t = \frac{6.676 - 7}{0.455/\sqrt{17}} = -2.9326$$

P-value:

$$p - value = 0.010$$

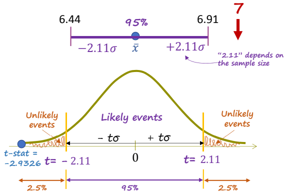
Confidence Interval:

$\bar{x} = 6.676,$
$\hat{\sigma} = 0.455/\sqrt{17}$

$H_0\colon \mu = 7$
$H_1\colon \mu \neq 7$

Hypothesis testing :

$$t = \frac{6.676 - 7}{0.455/\sqrt{17}} = -2.9326$$



P-value:

$p - value = 0.010$

Confidence Interval:

$\bar{x} = 6.676,$
$\hat{\sigma} = 0.455/\sqrt{17}$

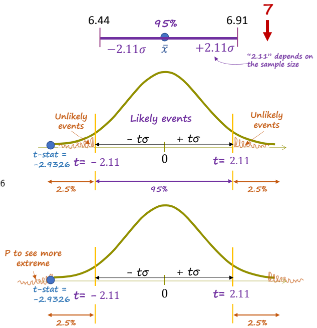
$H_0 : \mu = 7$
$H_1 : \mu \neq 7$

Hypothesis testing :

$$t = \frac{6.676 - 7}{0.455/\sqrt{17}} = -2.9326$$

P-value:

$p - value = 0.010$



Type of errors

- As we can specify different levels for a confidence interval, we can specify different levels for the test.
- This is just specifying what we mean by ''implausible''.
- To correspond to a 99% CI, we would reject H_0 if $P < 0.01$.
- We reject H_0 if $P < \alpha$,
  where α denotes the **significance level** of the test.
- Typically we use α=0.05, just as we typically use a 95% confidence interval. But we may choose α=0.01, 0.001, . . .
- If $P < \alpha$, we reject H_0; the result is **statistically significant.**

(One advantage of the P-value is that it gives a standard indication of the strength of the evidence against H_0; the smaller the P, the stronger the evidence against H_0.)

Terminology and conventions

	Accept H_0	Reject H_0
H_0 true	✓ Correct $(1-\alpha)$	Type I error (α)
H_1 true	Type II error (β)	✓ Correct $(1-\beta)$

Analogy with criminal trials:
the worse mistake is to convict an innocent person (Type I).

In deciding whether to accept or reject H_0, there is a risk of making two types of errors.

We set a small significance level (prob type I error) — usually 0.05; and then attempt to make the power (1 – prob type II error) as large as possible.

Terminology and conventions

$P < 0.05$	some evidence, significant (*)
$P < 0.01$	strong evidence, highly significant (**)
$P < 0.001$	very strong evidence, extremely significant (***)

- In deciding whether to accept or reject H_0, there is a risk of making two types of errors . . .
- We want α (Type I: reject H_0 when H_0 true) and β (Type II: accept H_0 when H_0 not true) to be small.

  The significance level, α is usually pre-set at 0.05;
- We then do what we can to make the power $(1 - \beta)$ large (and hence β small).
- This will generally mean taking a bigger sample i.e. "sample size calculation" and "power calculation".

Sample size

▷ **Example.** A random sample of 340 observation is obtained from a Normal population with standard deviation 46. The observed sample mean is 220. Test the null hypothesis that $\mu{=}211$.

$$[n{=}340,\ \mu_0{=}211,\ \sigma{=}46,\ \bar{x}{=}220] \quad\Rightarrow\quad z = \frac{220-211}{46/\sqrt{340}} = 3.61.$$

Hence we reject H_0 (using significance level 0.05) since $|z| > 1.96$. There is significant evidence in this sample that $\mu \neq 211$.

▷ **Example.** $[n{=}25,\ \mu_0{=}211,\ \sigma{=}46;\ \bar{x}{=}220]$

$$P = 2\,\Pr(\bar{X} > 220), \text{ where } \bar{X} \overset{\mathrm{d}}{=} \mathrm{N}(211, \tfrac{46^2}{25}) \quad (\mathrm{H}_0 \text{ distribution}).$$

$$\therefore\ \ P = 2\,\Pr(\bar{X}_s > \frac{220-211}{46/\sqrt{25}}) = 2\,\Pr(\bar{X} > 0.978) = 0.328.$$

Since $P > 0.05$, we do not reject the null hypothesis $\mu = 211$. There is no significant evidence in this sample that $\mu \neq 211$.

$(|z| = 0.978) < 1.96$

Hypothesis tests and confidence intervals

A hypothesis test will reject H_0 at the level α whenever the null value being tested is outside a $100(1-\alpha)\%$ confidence interval.

A value outside the interval is "not consistent with the data".

When performing a hypothesis test, you should always calculate a confidence interval as well. And present both in your results.

Comparative inference

Here we consider the important standard case where inference is required to compare two populations (sub-populations, groups . . .)

It is common to consider the comparison of the effects of two *treatments* or *interventions* or *exposures* or *attributes*.

There are two main ways in which treatments can be compared:

- **Paired comparisons** — the two treatments are applied to pairs of experimental units which have been matched so as to be as alike as possible (even the same experimental unit at different times);
- **Independent samples** — the two treatments are applied to separate sets of experimental units randomly selected from the sample population.

Inference for two populations

Corn and fertiliser experiment:

10 of 28 plots randomly selected to receive fertiliser (45 kg/ha).

Corn yields (kilolitres per hectare):

0	2.13	0	2.82
0	0.54	0	2.39
0	2.32	0	0.46
0	2.58	0	1.56
0	1.92	45	2.08
0	2.66	45	1.03
0	3.85	45	5.24
0	1.91	45	7.18
0	1.04	45	8.38
0	2.96	45	7.03
0	3.28	45	7.06
0	2.98	45	4.44
0	3.31	45	6.92
0	3.05	45	3.46

"standard" test statistic: $t = \dfrac{\mathrm{est} - \theta_0}{\mathrm{se}}$

where se = estimate of $\mathrm{sd}(\mathrm{est})$.

Here $\mathrm{est} = \bar{Y}_1 - \bar{Y}_2$ and $\theta_0 = 0$.

and $$\mathrm{sd}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

but we don't know σ_1 or σ_2, so . . .

(1) use $\mathrm{sd}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}} \Rightarrow t \overset{\mathrm{d}}{\approx} t_k$.

(2) use $\mathrm{sd}(\bar{Y}_1 - \bar{Y}_2) = s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} \Rightarrow t \overset{\mathrm{d}}{=} t_{n_1+n_2-2}$.

(assuming $\sigma_1 = \sigma_2$ — as for every other model we consider!)

Independent samples: assuming equal σ

If $X_1 \overset{\mathrm{d}}{=} \mathrm{N}(\mu_1, \sigma)$ and $X_2 \overset{\mathrm{d}}{=} \mathrm{N}(\mu_2, \sigma)$,

$$\bar{X}_1 - \bar{X}_2 \overset{\mathrm{d}}{=} \mathrm{N}\left(\mu_1 - \mu_2,\ \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

Inference is based on:

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{\mathrm{d}}{=} \mathrm{t}_{n_1+n_2-2}.$$

where $S^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$.

Independent samples: assuming equal σ

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \overset{\mathrm{d}}{=} \mathrm{t}_{n_1+n_2-2}.$$

A (95%) CI for $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm c_{0.975}(\mathrm{t}_{n_1+n_2-2})\ s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad [\text{est} \pm \text{“2”se}]$$

If H_0 is true, i.e. $\mu_1 = \mu_2$, then

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{is an observation on } \mathrm{t}_{n_1+n_2-2}.$$

```
> t.test(y1, y2, ...)
```

```
> t.test(y ~ g, data=d, ...)
```

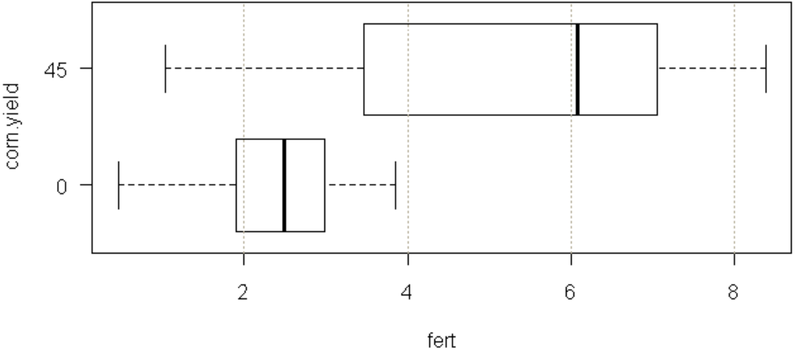
`alternative`	$\neq$	"two.sided", "less", "greater"
`mu`	0	(null hypothesis value, μ_0)
`conf.level`	0.95	
`paired`	F	paired or independent
`var.equal`	F	

```
>?t.test
```

```
> corn.fert <- data.frame(fert = c(rep("0",18), rep("45",10)),
+ corn.yield = c(2.13,0.54,2.32,2.58,1.92,2.66,3.85,
+ 1.91,1.04,2.96,3.28,2.98,3.31,3.05,
+ 2.82,2.39,0.46,1.56,2.08,1.03,5.24,
+ 7.18,8.38,7.03,7.06,4.44,6.92,3.46))

> plot(corn.fert,horizontal=T,las=1)
> grid(col="darkgray",nx=NULL,ny=NA)

> tapply(corn.fert$corn.yield, corn.fert$fert, length)
 0 45
18 10
> tapply(corn.fert$corn.yield, corn.fert$fert, mean)
    0    45
2.320 5.282
> tapply(corn.fert$corn.yield, corn.fert$fert, sd)
        0         45
0.9476597 2.4599584
```


corn.yield
45
0
2
4
6
8
fert

Independent samples: different σ

A (95%) CI for $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 \pm c_{0.975}(t_\nu)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

while a test statistic for testing H$_0$: $\mu_1 = \mu_2$ is given by

$$t = \frac{\bar{x}_2 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

which has (approximately) a t distribution with ν df, when H$_0$ is true.

Independent samples: different σ

$$\nu = \frac{(V_1 + V_2)^2}{\frac{V_1^2}{n_1-1} + \frac{V_2^2}{n_2-1}}, \quad \text{where} \quad V_1 = \frac{s_1^2}{n_1},\ V_2 = \frac{s_2^2}{n_2}.$$

Corn yield and fertiliser:

95% confidence interval for $\mu_1 - \mu_2$ is $(1.17, 4.75)$.

(compared to $(1.63, 4.29)$ assuming equal σ).

For a test of H_0: $\mu_1 = \mu_2, \quad t = 3.66$,

resulting in a P-value of 0.004.

```
> t.test(corn.yield ~ fert, data=corn.fert)

        Welch Two Sample t-test

data:  corn.yield by fert
t = -3.6598, df = 10.507, p-value = 0.004047
alternative hypothesis: true diff in means is not equal to 0
95 percent confidence interval:
 -4.753588 -1.170412
sample estimates:
 mean in group 0 mean in group 45
           2.320            5.282

           group 0      group 45
count           18            10
mean          2.32          5.28
sd            0.95          2.46
```

Paired samples

For independent samples, look at the difference between samples; for paired samples, look at the sample of differences.

Effect of gaps in plantations on pine needle blight

Ten blocks each had 2 plots—one bordering a large gap and the other with no gap.

Disease scores were:

Block	1	2	3	4	5	6	7	8	9	10
Large gap	4.1	3.3	3.8	3.8	4.3	2.7	4.0	3.4	2.1	1.4
No gap	3.3	3.3	3.2	3.0	3.1	2.4	3.4	3.1	2.3	1.2

Paired samples

Ignoring pairs:

```
> large.gap <- c(4.1, 3.3, 3.8, 3.8, 4.3, 2.7, 4, 3.4, 2.1, 1.4)
> no.gap <- c(3.3, 3.3, 3.2, 3, 3.1, 2.4, 3.4, 3.1, 2.3, 1.2)
> t.test(large.gap, no.gap, var.equal=T)

         Two Sample t-test
 data:  large.gap and no.gap t = 1.2468, df = 18, p-value = 0.2284
 alternative hypothesis: true difference in means is not equal to 0
 95 percent confidence interval:
  -0.3150982  1.2350982
 sample estimates: mean of x mean of y
      3.29      2.83
```

conclude that there is no significant effect of gaps (wrongly!)

Paired samples

Calculate differences:

block	1	2	3	4	5	6	7	8	9	10
large gap	4.1	3.3	3.8	3.8	4.3	2.7	4.0	3.4	2.1	1.4
no gap	3.3	3.3	3.2	3.0	3.1	2.4	3.4	3.1	2.3	1.2
difference	0.8	0.0	0.6	0.8	1.2	0.3	0.6	0.3	-0.2	0.2

. . . and look at the sample of differences.

(We could do a sign test: $P = 2\Pr(X \geqslant 8)$ where $X \overset{\mathrm{d}}{=} \mathrm{Bi}(9, 0.5)$.
This gives $P = 0.039$, and significant evidence against H_0.
But it doesn't estimate the mean difference.
though you could estimate the median of the differences.)

Paired samples

Perform inference on the differences (using t):

```
> diff <- large.gap - no.gap
> t.test(diff)

    One Sample t-test
data:  diff
t = 3.4674, df = 9, p-value = 0.007078
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1598909 0.7601091
sample estimates:
mean of x
     0.46
```

Correctly conclude that there is a significant effect of gaps.

Paired samples

OR use paired=T:

```
>  t.test(large.gap, no.gap, paired=T)

        Paired t-test

data:  large.gap and no.gap
t = 3.4674, df = 9, p-value = 0.007078
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1598909 0.7601091
sample estimates:
mean of the differences
                   0.46
```

Correctly conclude that there is a significant effect of gaps.

Paired samples

Why is pairing so effective here?

- Greater variability between pairs than within pairs;
- Ignoring pairs results in larger unexplained variation;
- Differencing removes substantial variation;
- The gap-effect is hidden by the variation between pairs.

Message?

- pairing is more efficient . . . so pair if you can;
- determine whether the data are paired: it affects the analysis.