Two sample problems

Chapters: 18 and 20.

Outline of presentation:

1) Confidence intervals and tests for difference of two means:

$$\mu_1 - \mu_2$$

2) Confidence intervals and tests for difference of two proportions:

$$p_1 - p_2$$

Summary: all intervals of the form

$$\text{estimate} \pm \text{multiplier} \times \text{SE}$$

where SE is standard error of the estimate.

All tests of the form

$$\frac{\text{estimate} - \text{target}}{\text{standard error}}$$

An example: 8 animals treated with a steroid, 10 controls.

Measured body weight gain.

Two samples: $n_1 = 8, n_2 = 10$.

Question: does steroid change body weight gain?

Data: two sample means: $\bar{x}_1 = 32.8$, $\bar{x}_2 = 40.5$.

Two sample sds: $s_1 = 2.6$, $s_2 = 2.5$.

Step 1: recognize that question has yes/no answer.

So: test hypothesis.

Introduce notation.

1) μ_1 is population mean weight gain of steroid treated animals.

2) μ_2 is population mean weight gain of untreated animals.

Frame null hypothesis: $\mu_1 = \mu_2$.

Equivalent hypothesis: $\mu_1 - \mu_2 = 0$.

Examine question to choose alternative:

Alternative does not predict direction so:

Alternative is $\mu_1 \neq \mu_2$.

Develop test statistic:

Numerator is measure of discrepancy:

$$(\bar{x}_1 - \bar{x}_2) - \text{target value of } \mu_1 - \mu_2$$

This is just

$$\bar{x}_1 - \bar{x}_2$$

Need standard error for difference.

Basic formula: SE of difference of two independent quantities is:

$$\sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$$

So: standard error of $\bar{x}_1 - \bar{x}_2$ is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Normally we must estimate this using

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This leads to the test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In our example get:

$$t = \frac{32.8 - 40.5}{\sqrt{\frac{2.6^2}{8} + \frac{2.5^2}{10}}} = -6.35$$

How do we compute a P-value?

Solution used in text: from t tables taking

$$df = \min\{n_1 - 1, n_2 - 1\} = 7$$

in our case.

In table C for 7 df largest value of t is 5.408 corresponding to $P = 0.001$ so

$$P < 0.001$$

From software $P \approx 0.00038$.

Conclusion: strong (*very highly significant*) evidence that steroid affects average weight gain.

Confidence interval for mean difference: $\mu_1 - \mu_2$?

Same formula as always:

$$\bar{x}_1 - \bar{x}_2 \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

In our example with 7 df get $t^* = 2.365$ for 95% CI. Also get standard error:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1.212$$

Degrees of Freedom; Controversy

There is not universal agreement on how to do this test.

Book gives two options and dismisses a third.

Option 1: Easy: use t statistic as above and df as above.

Option 2: **Satterthwaite's** approximation. Use t as above but

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Option 3: Assume $\sigma_1 = \sigma_2$ and use **pooled** estimate of standard error:

$$\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

and then take

$$df = n_1 + n_2 - 2$$

In our example pooling produces the standard error

$$\sqrt{\frac{7*2.6^2+9*2.5^2}{16}\left(\frac{1}{8}+\frac{1}{10}\right)} = 1.207$$

and

$$t = -6.38$$

with $df = 16$. The P value becomes much smaller, however. From software

$$P = 9.1 \times 10^{-6}$$

Option 2 gives $df = 14.86$ and $P = 1.36\times10^{-5}$.

Commentary:

1) Software usually does option 3 by default.

2) Better software also produces Option 2.

3) In this case not much difference in conclusions.

Comparing two proportions.

Example: two samples of praying mantis.

Brown: 65; Green: 45.

Of brown: 45 put on green leaves.

Of green: 25 put on brown leaves.

After 3 weeks: of the 45 brown on green leaves 26 still alive. Of the 25 green on brown leaves 16 still alive.

Question: difference in survival rates?

Common presentation of results:

Contingency table.

	Insect Type		
Status	Brown	Green	Total
Alive	26	16	42
Dead	19	9	28
Total	45	25	70

Model: Let X_1 be surviving number of brown.

Let X_2 be surviving number of green.

Each of X_1, X_2 is Binomial.

Numbers of trials $n_1 = 45$, $n_2 = 25$.

Population survival probabilities: p_1, p_2.

Null hypothesis $p_1 = p_2$.

Alternative $p_1 \neq p_2$.

Test statistic:

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Note: in denominator $\widehat{p} = (X_1 + X_2)/(n_1 + n_2)$ is overall success rate.

Called **pooled** estimate.

In our case:

$$\widehat{p}_1 = \frac{26}{45} = 0.578 \quad \widehat{p}_2 = \frac{16}{25} = 0.64$$

and

$$\widehat{p} = \frac{26+16}{45+25} = 0.6$$

This gives

$$z = \frac{-0.0622}{\sqrt{0.6 \times 0.4 \left(\frac{1}{45} + \frac{1}{25}\right)}} = -0.51$$

Get P-value from normal tables: two sided.

$$P = 0.61$$

Interpretation: not much evidence of a difference in survival rates.

Ref: di Cesnola, A.P. (1904) *Biometrika*, **4**, 58–59.

Confidence interval for $p_1 - p_2$:

$$\hat{p_1} - \hat{p_2} \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Notice: No pooling. (In testing, pooling justified by null hypothesis.)

Commentary: text recommends: add 1 to each X_i and 2 to each n_i then do all arithmetic as above.

Not standard.

Improves coverage probability.

Large sample methods not recommended unless all of

$$n_1 p_1, n_1(1-p_1), n_2 p_2, n_2(1-p_2)$$

large enough. Book recommends all be at least 10.

Judged by all cell counts at least 10 in contingency table.

Matched pairs designs: instead of 2 independent samples, have 1 sample of pairs.

Example: look back at cross-fertilization of peas example.

Originally would have had 2 measurements for each parent.

Data reduced by subtraction to 1 sample problem!

Example: Pearson Lee data on father / son height.

Consists of $N = 1078$ pairs.

Denote: F_i father's height and S_i son's height in ith pair.

Problem: are sons taller than fathers?

Idea: μ_1 is population average height of sons.

μ_2 pop average height of fathers.

(At point in time when data collected!)

Point of next piece: illustrate merit of matched pairs design.

Treat the $N = 1078$ pairs as the population.

Then $\mu_1 = 68.68$, $\mu_2 = 67.69$, $\sigma_1 = 2.81$ and $\sigma_2 = 2.74$.

In the population the variables F and S are correlated:

$$\rho = 0.502$$

Consider two methods of comparing μ_1 and μ_2 based on sampling.

Method 1: take two samples of size $n_1 = n_2 = 9$, one of Fathers, other of Sons.

Method 2: take one sample of $n = 9$ pairs of Fathers and sons.

An explicit example:

For Method I: drew following 2 independent samples

Family # i	F_i	Family # i	S_i
128	70.01	635	78.25
251	68.32	574	70.70
756	65.24	564	69.20
150	69.52	778	69.12
257	64.07	160	70.82
⋮	⋮	⋮	⋮

Drew total of 1000 samples of $n = 9$ fathers and 1000 samples of $n = 9$ sons.

For Method II I drew the following sample of pairs:

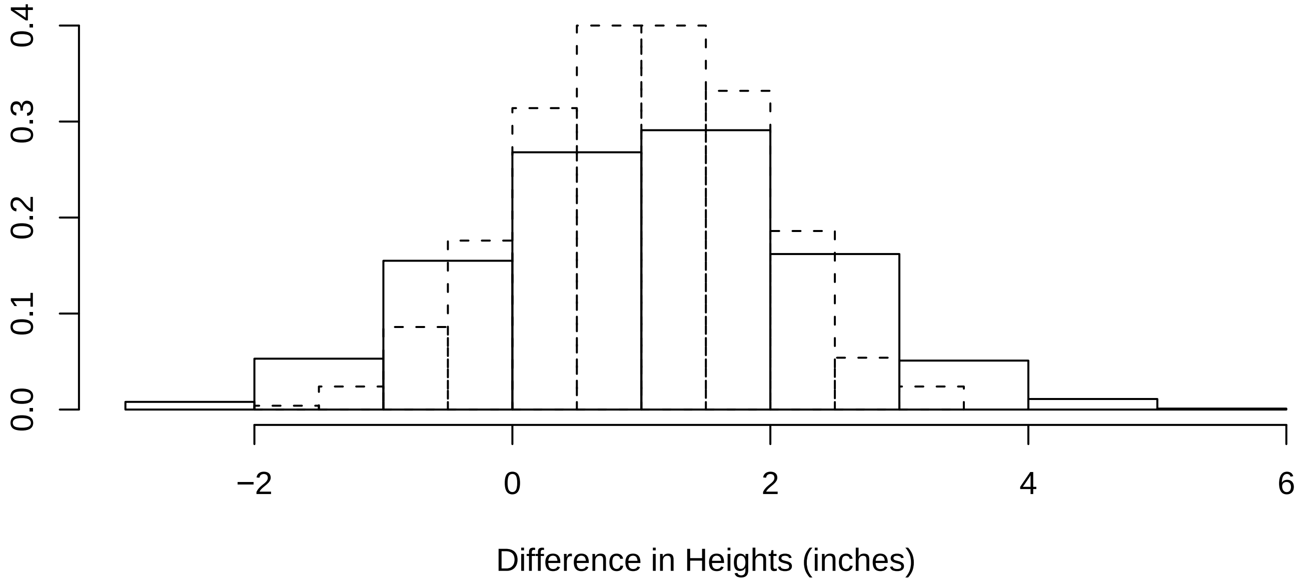
Family # i	F_i	S_i	$S_i - F_i$
851	69.07	78.36	9.29
53	65.83	67.07	1.24
919	65.68	67.68	2.00
475	64.68	66.79	2.11
754	64.34	69.23	4.88
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Repeated this 1000 times.

Here is a histogram of $\bar{S} - \bar{F}$.

Solid lines: two independent samples.

Dotted lines: sample of pairs.



Numerical summary of this Monte Carlo experiment.

Method 1 outcomes:

$\bar{F}$	$\bar{S}$	$\bar{S} - \bar{F}$
68.15	69.78	1.63
66.63	9.98	3.35
68.11	7.98	-0.13
68.48	69.36	0.88
67.17	67.32	0.15
⋮	⋮	⋮

Method 2 outcomes:

$\bar{F}$	$\bar{S}$	$\bar{S} - \bar{F}$
67.76	67.68	-0.07
68.66	68.01	-0.65
68.96	69.15	0.20
66.33	67.72	1.39
68.33	67.90	-0.43
⋮	⋮	⋮

To compare: examine mean and sd of the last columns:

Get

Independent		Matched Pair	
Mean	SD	Mean	SD
1.046	1.302	0.958	0.932

Major point: both means close to $\mu_1 - \mu_2 = 0.997$.

But: SD for matched pairs is smaller.

Formula for SE of difference of two independent means:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1.308$$

Formula for SE of difference in paired sample:

$$\sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2\rho * \sigma_1\sigma_2}{n}} = 0.925$$

Notice great match of theory to Monte Carlo.

Example problem: Does too much sleep impair intellectual performance.

10 subjects tested twice each.

Once after two normal night's sleep ,

Once after two nights of 'extended sleep'.

Data on test for vigilance: low scores are alert:

Subject	1	2	3	4	5	6	7	8	9	10
Normal	8	9	14	4	12	11	3	26	3	11
Extended	8	9	15	2	21	16	9	38	10	11
Diff	0	0	-1	2	-9	-5	-6	-12	-7	0

WARNING: I might put in a column of differences even if data are not paired.

Null: pop mean difference $\mu_N - \mu_E$ in vigilance scores is 0.

Alternative: $\mu_N < \mu_E$.

Summary statistics: $\bar{N} - \bar{E} = -3.8$; $s = 4.66$.

Test statistic:

$$t = \frac{-3.8 - 0}{4.66/\sqrt{10}} = -2.58$$

One sided alternative. P-value in lower tail. 9df.

$$P = 0.015$$

In tables best approx is $0.01 < P < 0.02$.

What if: had used 20 subjects. 10 assigned to Normal, 10 to Extended at random?

Could have presented same data (but probably without row 'Subject').

Analysis: not paired, so 2 sample t test.

Hypotheses unchanged!

$$\bar{x}_N = 10.1, s_N = 6.81, \bar{x}_E = 13.9, s_E = 9.92$$

Two sample t statistic is

$$t = \frac{10.1 - 13.9}{\sqrt{\frac{6.81^2}{10} + \frac{9.92^2}{10}}} = -1.00$$

which gives

$$P = 0.172$$

In tables $0.15 < P < 0.2$. Not significant.

Summary points:

1) for original description of experiment paired analysis right, two sample analysis wrong. (Only 10 subjects.)

2) Since two variables positively correlated paired design is better.

3) conclusion is that extra sleep does seem to worsen vigilance.

4) but if we had collected same data in unpaired design would have concluded no real evidence that extra sleep worsens vigilance.

Another example:

Studying gopher tortoise burrows to see which are active.

Two methods of evaluation of ‘active’ compared.

Camera versus ‘experience’.

Data: 151 burrows judged by ‘experience’. 107 rated active.

114 judged by cameras. 48 rated active.

Problem: evaluation methods equivalent?

Assume: burrows assigned to evaluation method at random.

If X_1 is number judged active by experience then X_1 is Binomial with $n_1 = 151$, and some p_1. We estimate

$$\hat{p}_1 = X_1/n_1 = 107/151 = 0.7086.$$

Then X_2 number judged active by camera is Binomial $n_2 = 114$,

$$\hat{p}_2 = 48/114 = 0.44211$$

Null hypothesis: $p_1 = p_2$.

Alternative: $p_1 \neq p_2$.

Pooled estimate of p assuming $p_1 = p_2$ is

$$\hat{p} = \frac{107 + 48}{151 + 114} = 0.5849.$$

Test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{0.5849(1 - 0.5849)\left(\frac{1}{151} + \frac{1}{114}\right)}} = 4.70$$

Get two sided P-value; less than 0.006 in Table A.