2-Sample Methods

Module 10

Statistics 251: Statistical Methods

Updated 2021

Comparing two groups

Comparisons:

(1) Two independent means
  (a) When $\sigma_1^2 \approx \sigma_2^2$: Pooled
  (b) When $\sigma_1^2 \neq \sigma_2^2$: Unpooled (also called Welch or Satterthwaite)
(2) Dependent means
(3) Two proportions (independent)

Independent means

This compares the means of two distinct (separate) groups of units or subjects. The wording used is **the difference of two (2) means**

While there are two cases for this (when variances are equal or unequal), we will only use the unequal variances (unpooled) method. If the two variances are unequal or equal, the unpooled is appropriate in either case. (In practice, a variance test is done to see if they are equal or not before deciding either pooled or unpooled; we will just learn unpooled)

Formula: df for use of t

Degrees of freedom for (unpooled) independent means is calculated rather than using $n-1$ or something similar:

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

Since we are using `R` software output, it will be calculated automatically and displayed in the output

Formula: CI

CI for the difference of two (independent) means:

$$\overline{X}_1 - \overline{X}_2 \pm t^\star(se) \text{ where } se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ and } t^\star = t_{\alpha/2,df}$$

Hypotheses

For the difference of two (independent) means[1]:

$$H_0 : \mu_1 = \mu_2 \quad H_a : \mu_1 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} \mu_2$$

$$Or$$

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} 0$$

Assumptions

(1) Independence (if random met, this is met)
(2) Randomization
(3) Each group of observations have approximate normal distribution

Formula: Test Statistic

$$t = \frac{\overline{X}_1 - \overline{X}_2}{se} \quad where\ se = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

With

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

Again, with software output, all of the values are calculated for you

Dependent means

This compares the mean of the difference between two measurements of the same unit or subject. The wording used is **the mean difference**. This analysis is for comparing measurements on the same subject/unit; once before a treatment and once again after the treatment, to detect if there is a difference due to the treatment.

Examples are weight loss programs, Coke vs. Pepsi, compare GDP of countries at 2 different dates (time is treatment)

Formula: CI

d_i: individual differences between measurements

$\overline{X}_d = \frac{\sum d_i}{n}$ sample mean difference (mean of the differences)

$s_d = \sqrt{\frac{\sum (d_i - \overline{X}_d)^2}{n-1}}$: sample standard deviation of the differences

CI for the mean difference:

$$\overline{X}_d \pm t^\star (se) \quad where\ se = \frac{s_d}{\sqrt{n}} \quad and \quad t^\star = t_{\alpha/2, df},\ df = n - 1$$

[1] In practice, the difference of means can be hypothesized to be equal to a value other than zero

Hypotheses

For the mean difference[2]:

$$H_0 : \mu_d = 0 \quad H_a : \mu_d \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} 0$$

Assumptions

(1) Randomization
(2) Independence (of units/subjects)
(3) *Differences* have approximate normal distribution
(4) Two measurements per unit/subject

Formula: Test Statistic

$$t = \frac{\overline{X}_d - 0}{se} \text{ where } se = \frac{s_d}{\sqrt{n}}$$

Two Proportions

This compares the proportions of two distinct (separate) groups of units or subjects. The wording used is **the difference of two (2) proportions**

The se for the test is different from the se for the CI

Formula: CI

CI for the difference of two (independent) proportions:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z^{\star}(se) \text{ where } se = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}} \text{ and } z^{\star} = z_{\alpha/2}$$

Hypotheses

For the difference of two (independent) proportions[3]:

$$H_0 : \pi_1 = \pi_2 \quad H_a : \pi_1 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} \pi_2$$

Or

$$H_0 : \pi_1 - \pi_2 = 0 \quad H_a : \pi_1 - \pi_2 \begin{pmatrix} \neq \\ > \\ < \end{pmatrix} 0$$

[2]In practice, the mean difference can be hypothesized to be equal to a value other than zero
[3]In practice, the difference of proportions can be hypothesized to be equal to a value other than zero

Assumptions

(1) Independent groups (if random met, this is met)
(2) Randomization
(3) success/failure condition to have normality
    (a) **either** $n_1 \geq 60$ **AND** $n_2 \geq 60$ or
    (b) $n_1\pi_1 \geq 5$, $n_1(1-\pi_1) \geq 5$, $n_2\pi_2 \geq 5$, **AND** $n_2(1-\pi_2) \geq 5$[4]

Formula: Test Statistic

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{se} \text{ where } se = \sqrt{\hat{\pi}(1-\hat{\pi})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Where $\hat{\pi}$ without subscripts is the pooled proportion used when assuming the difference of proportions is equal to 0.

$$\hat{\pi} = \frac{X_1 + X_2}{n_1 + n_2}$$

X_1, X_2 are the successes from each group. If you are given percents, then you will have to calculate the successes by:

$$X_1 = n_1\hat{\pi}_1 \quad X_2 = n_2\hat{\pi}_2$$

Analyses

Thankfully, the basic process is still the same as for 1-sample methods. Make sure to follow the 5 steps to hypothesis testing:

(1) State hypotheses, check assumptions if requested
(2) State test statistic t, df, and *pvalue* from output
(3) State test results
(4) Make conclusion in context from results
(5) State possible error that could have been made and discuss it within the context

2 independent means

Some archaeologists theorize that ancient Egyptians interbred with several different immigrant populations over thousands of years. To see if there is any indication of changes in body structure that might have resulted, in a random sample they measured 30 skulls of male Egyptians dated from 4000 BCE and 30 others dated from 200 BCE. - Is there sufficient evidence that the mean breadth of males' skulls increased (as theorized by archaeologists) over this period? Conduct hypothesis test - Estimate the true difference of means with 95% confidence and interpret

Egypt data

```
head(egypt); tail(egypt)
```

```
  breadth    era
1     141 200BCE
2     141 200BCE
3     135 200BCE
4     133 200BCE
```

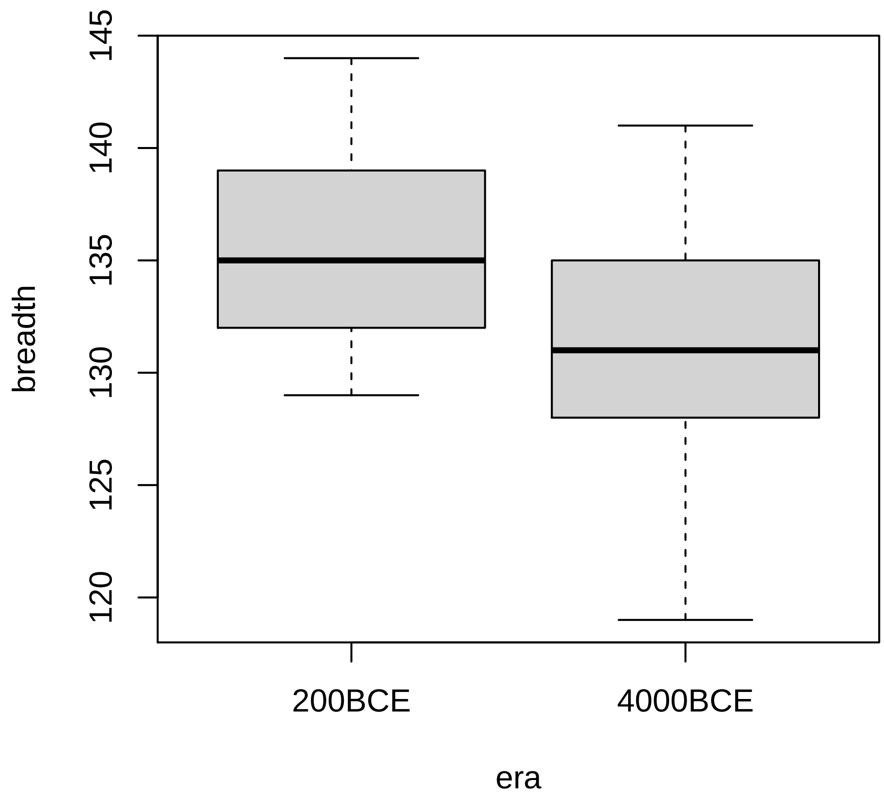
[4] Annoyingly, depending on the textbook, the values could be 5, 8, 10, 12, or 15. . . but 5 is good :-)

```
5       131  200BCE
6       140  200BCE
```

```
   breadth     era
55     130 4000BCE
56     138 4000BCE
57     128 4000BCE
58     127 4000BCE
59     131 4000BCE
60     124 4000BCE
```

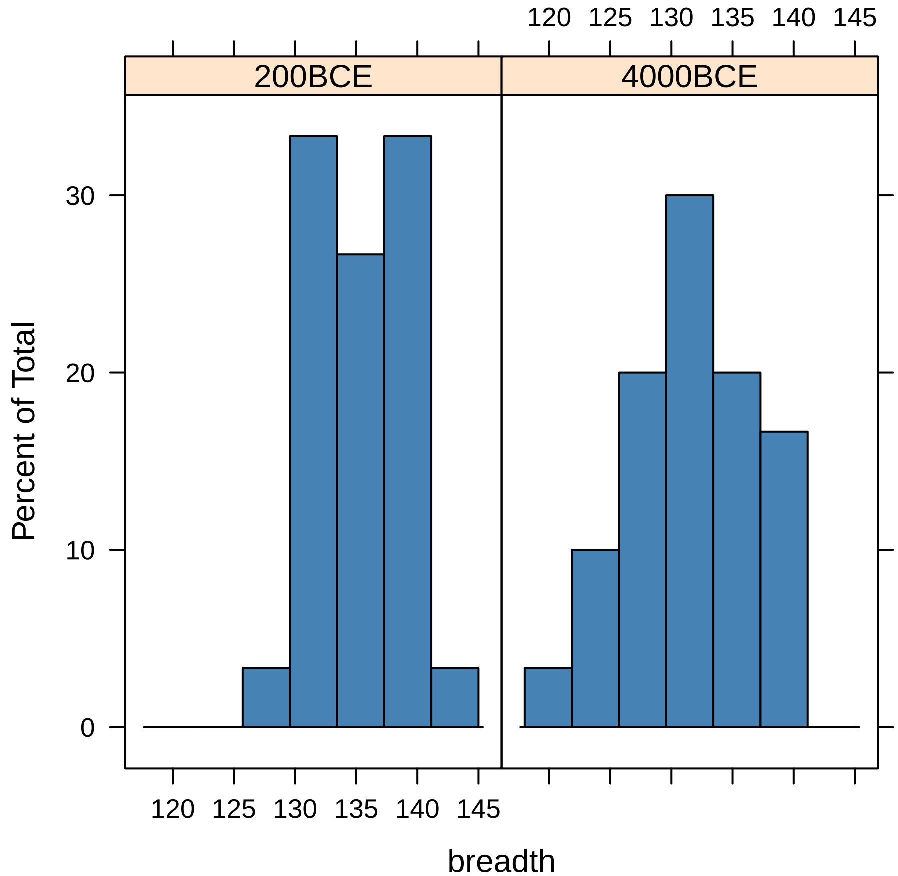
Egypt boxplots

```
boxplot(breadth~era,data=egypt)
```



Egypt histograms

```
histogram(~breadth|era,data=egypt,col='steelblue')
```



Egypt setup

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_a : \mu_1 - \mu_2 > 0$$

Assumptions:
(1) Independence: random so yes
(2) Randomization: yes
(3) Normality: $n_1 = n_2 = 30 \geq 30$ so yes

Organization of information:
$n_1 = 30$
$n_2 = 30$
$H_a :\ >$ (upper tail test)
$\alpha = 0.05$ (assumed because not specifically stated otherwise)

Egypt analysis test output

When doing one-tailed tests with software, the CIs are not the ones we want so a separate analysis is to be done to acquire proper CIs when doing one-tail tests (upper or lower)

```
with(egypt,t.test(breadth~era,alternative='g'))
```

```
    Welch Two Sample t-test

data:  breadth by era
t = 3.5797, df = 54.973, p-value = 0.000364
alternative hypothesis: true difference in means is greater than 0
```

```
95 percent confidence interval:
 2.27257     Inf
sample estimates:
 mean in group 200BCE mean in group 4000BCE
             135.6333              131.3667
```

Egypt analysis CI output

```
with(egypt,t.test(breadth~era))
```

```
    Welch Two Sample t-test

data:  breadth by era
t = 3.5797, df = 54.973, p-value = 0.000728
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.878030 6.655303
sample estimates:
 mean in group 200BCE mean in group 4000BCE
             135.6333              131.3667
```

Egypt test conclusion

Test statistic $t = 3.579$, $df = 54.973$, $pvalue = 0.000364$

Results: $pvalue = 0.000364 \leq \alpha(0.05) \therefore$ (therefore) H_0 is rejected

Conclusion: since the null is rejected, that means that there is evidence that the skull breadths have significantly increased over the period from 4000 BCE to 200 BCE

Error: since H_0 was rejected, a Type I error (reject null when null is true) could have been made; we think the the skull breadths have increased but they did not

Egypt CI interpretation

The CI: $(1.878030, 6.655303) \approx (1.88, 6.66)$

With 95% confidence, the true difference in mean skull breadths of Egyptian males from 4000 BCE to 200 BCE is 1.88 to 6.66 mm.

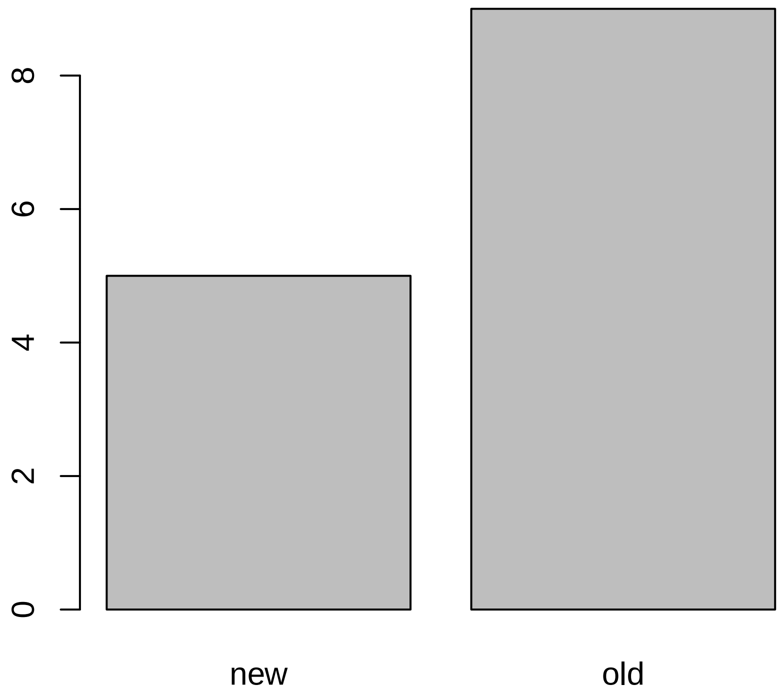
Alternative way to interpret: With 95% confidence, mean skull breadths of Egyptian males have increased from 4000 BCE to 200 BCE, 200 BCE skulls are 1.88 to 6.66 mm larger than the 4000 BCE skulls, indicating that immigrating populations did interbreed with the native Egyptians.

2 independent proportions

Sludge is a dried product remaining from processed sewage and is often used as a fertilizer on crops; there could be dangerous concentrations of nickel in the crops. A new method of processing sewage has been developed and a randomized experiment conducted to evaluate its effectiveness in removing heavy metals. Sewage of a known concentration of nickel is treated using both old and new methods and applied to 100 tomato plants that were randomly assigned to pots containing sewage sludge processed by one of the two methods and the nickel was measured in the tomatoes. Is there sufficient evidence that the concentration of nickel from the new treatment is less than the old treatment? Estimate the true difference of proportions with 95% confidence and interpret

Sludge graph

```
barplot(x)
```



Sludge data

```
nickel
```

```
      Toxic Non-toxic Total
New       5        45    50
Old       9        41    50
Total    14        86   100
```

Sludge setup

$$H_0 : \pi_1 - \pi_2 = 0 \quad H_a : \pi_1 - \pi_2 < 0$$

Assumptions:
(1) Independence: random so yes
(2) Randomization: yes
(3) Normality: $n_1 = n_2 = 50 \ngeq 60$ but there are at least 5 successes for new and old method so yes

Organization of information:
$n_1 = 50$
$n_2 = 50$
$H_a :\ <$ (lower tail test)
$\alpha = 0.05$ (assumed because not specifically stated otherwise)

Sludge analysis test output

When doing one-tailed tests with software, the CIs are not the ones we want so a separate analysis is to be done to acquire proper CIs when doing one-tail tests (upper or lower)

```r
t.test(sludge.new,sludge.old,alternative='l')
```

```
	Welch Two Sample t-test

data:  sludge.new and sludge.old
t = -1.1489, df = 92.56, p-value = 0.1268
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf 0.03569674
sample estimates:
mean of x mean of y
     0.10      0.18
```

Sludge analysis CI output

When doing one-tailed tests with software, the CIs are not the ones we want so a separate analysis is to be done to acquire proper CIs when doing one-tail tests (upper or lower)

```r
t.test(sludge.new,sludge.old)
```

```
	Welch Two Sample t-test

data:  sludge.new and sludge.old
t = -1.1489, df = 92.56, p-value = 0.2536
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2182892  0.0582892
sample estimates:
mean of x mean of y
     0.10      0.18
```

Sludge test conclusion

Test statistic $t = -1.1489$, $df = 92.56$, $pvalue = 0.1268$

Results: $pvalue = 0.1268 \nleq \alpha(0.05)$ $\therefore$ (therefore) H_0 is not rejected

Conclusion: since the null is not rejected, that means that the concentration of nickel from the new treatment is not significantly less than the old treatment

Error: since H_0 was not rejected, a Type II error (not rejecting null when null is false) could have been made; we think the new method is no better than the old method at nickel removal

Sludge CI interpretation

The CI: $(-0.2182892, 0.0582892) = (-21.83\%, 5.83\%)$

With 95% confidence, the true difference in proportions of plants with nickel from new vs. old treatment is between -21.83% and 5.83%. Since the CI includes 0, we say that there is no significant difference between the two treatments

Dependent means

Trace metals in drinking water affect the flavor; high concentrations can be a health hazard. A randomized study looked at six river locations along the South Indian River (6 units) and the zinc concentration in mg/L was measured for both surface and bottom water at each location. Is there sufficient evidence the true mean difference in concentration in bottom water differs from that of surface water? Let $\alpha = 0.10$. Estimate the true mean difference with 90% confidence and interpret

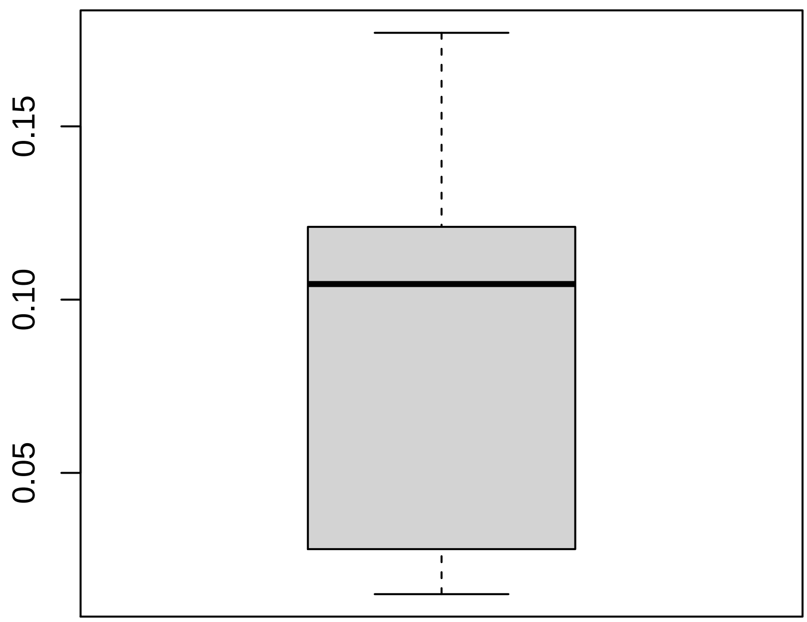
River data

```
zinc
```

```
                1     2     3     4     5     6
Bottom     0.430 0.266 0.567 0.531 0.707 0.716
Surface    0.415 0.238 0.390 0.410 0.605 0.609
Difference 0.015 0.028 0.177 0.121 0.102 0.107
```

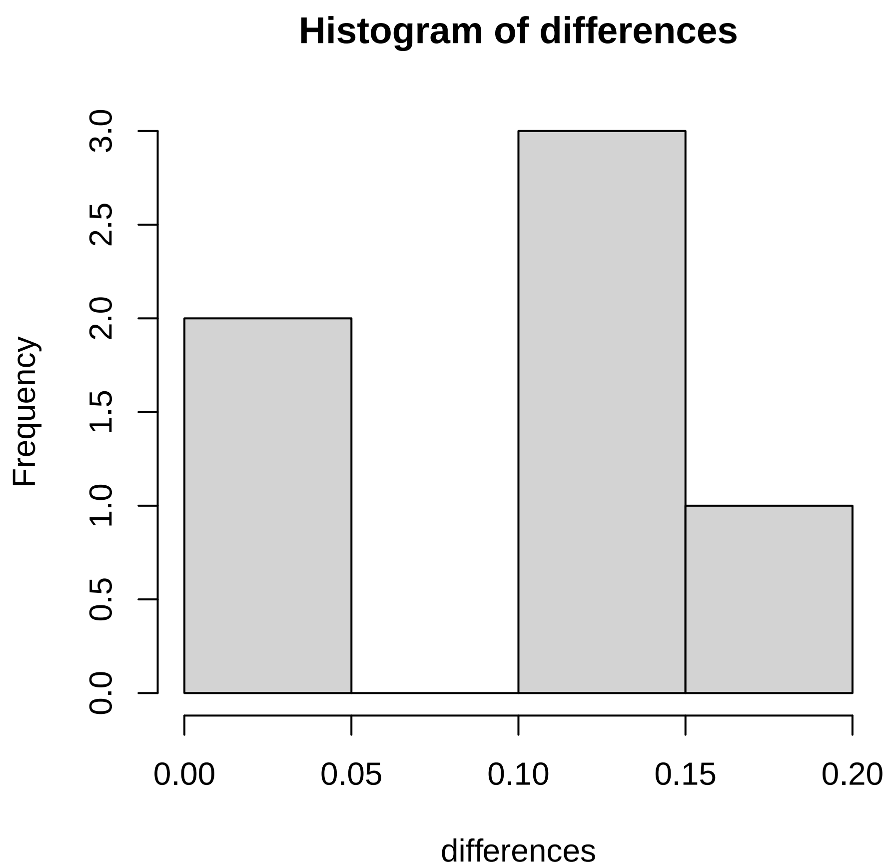
River boxplot

```
boxplot(differences)
```



River histogram

```
hist(differences)
```



River setup

$$H_0 : \mu_D = 0 \quad H_a : \mu_D \neq 0$$

Assumptions:
(1) Randomization: yes
(2) Independence (of units/subjects): random met so yes
(3) *Differences* have approximate normal distribution (boxplots are ok)
(4) Two measurements per unit/subject: yes

Organization of information:
$n = 6$ (6 sites)
$H_a : \neq$ (two-tail test)
$\alpha = 0.10$ (specifically stated)

River analysis output

When doing two-tailed tests with software, the CIs are the proper ones we want and no additional analysis adjusting for one- vs. two-tailed test is necessary

```
t.test(bottom,surface,paired=T,conf.level=.9)
```

```
    Paired t-test

data:  bottom and surface
t = 3.6998, df = 5, p-value = 0.014
alternative hypothesis: true difference in means is not equal to 0
```

```
90 percent confidence interval:
 0.04174206 0.14159127
sample estimates:
mean of the differences
             0.09166667
```

River test conclusion

Test statistic $t = 3.6998$, $df = 5$, $pvalue = 0.014$

Results: $pvalue = 0.014 \leq \alpha(0.10) \therefore$ (therefore) H_0 is rejected

Conclusion: since the null is rejected, that means the true mean difference in concentration in bottom water significantly differs from that of surface water

Error: since H_0 was rejected, a Type I error (reject null when null is true) could have been made; we think the mean difference in concentration in bottom water significantly differs from that of surface water when it does not

River CI interpretation

The CI: $(0.04174206, 0.14159127) \approx (0.0417, 0.1416)$

With 90% confidence, the true mean difference in zinc concentration between bottom and surface water is between 0.0417 and 0.1416 mg/L.

Alternative way to interpret: With 90% confidence, zinc concentration is between 0.0417 and 0.1416 mg/L higher at the bottom than at the surface