**Inferences about μ1 - μ2: paired data**

Examples:

* Which grocery store has lower prices on average: Wal-Mart or HyVee?
  + - Sample the same items at each store (i.e. “match” or “pair” the items)
* Does an allergy relief medicine provide relief on average to patients?
  + - Sample the same person before and after they take the medicine.
* Have overall sales ($) at McDonald’s stores improved on average with their new advertising campaign?
  + - Sample the sales of the same restaurants before and after the promotion started.
* Which of two types of running shoes has the least wear after 1 month?
  + - Sample n individuals and have each individual wear one type of shoe on one foot and the other type of shoe on the other foot. Randomly assign the shoe type to each foot.

Notice that items are paired in each sample. Often, this is referred to as dependent samples. Compare this to the last section when we had “independent” samples.

Matched pairs are usually a better way to do an experiment. An explanation of why will be given shortly.

Setup

Suppose a random sample of size n is taken. Let Y11, Y21, …, Yn1 represent the sample from population #1. Let Y12, Y22, …, Yn2 represent the sample from population #2. Suppose Yi1 and Yi2 have something in common with respect to how the sample was obtained for i = 1, …, n; thus, because Yi1 and Yi2 have something in common, they are dependent random variables.

Transforming the observations

To take advantage of the paired observations, we are now going to use the observed values in the form of

Di = Yi1 – Yi2

for i = 1, …, n (notice the same sample size for each sample). We can then use the usual one sample methods developed earlier in this course where “D” replaces the usual “Y”! Then

 and 

are the sample mean and sample standard deviation, respectively. Also, the parameter we are interested in estimating is denoted symbolically as μd. Notice this helps to differentiate this section’s statistical methods from those in the previous section with independent samples.

One can show that working with dependent samples generally results in a smaller variance for the sample mean than working with independent samples; i.e., . Intuitively, this is because the inherent variability that you would have from one sampled item to another sampled item is partially removed due to the pairing.

Question: What type of effect will this result have on CIs? Which is more desirable then – working with dependent or independent samples?

In a more general setting, one will refer to this type of experimental design as a “randomized complete block” experiment.

Using what was learned earlier in the last chapter, we can develop a CI for μd!

CI for μd for paired observations: If the differences Di for i = 1, …, n have a normal distribution with a mean μd and variance , a (1-α)100% CI for μd is



where tα/2,n-1 is the 1 – α/2 quantile from a t distribution with n – 1 degrees of freedom,  is the observed value for , and  is the observed value for .

Notes:

* This is the same CI as discussed in for one population mean, but calculated on the difference between paired observations!
* Please see previous discussions about its use then. For example, what happens to the need for the normal probability distribution when there is a “large” sample size?

Example: CPT (cpt.R)

A pharmaceutical company is conducting clinical trials on a new drug used to treat schizophrenia patients. Ten healthy male volunteers were given 3mg of the drug.

Before the drug was administered (time = 0) and 4 hours after (time = 4), a psychometric test called the Continuous Performance Test (CPT) was administered. The CPT involves the following:

* A subject sits in front a computer screen.
* Randomly generated numbers appear on the computer screen.
* Each image is slightly blurred.
* One number appears every second for 480 seconds.
* Subjects are required to press a button whenever the number 0 appears.
* We want to examine the number of hits (i.e. the number of correct responses).

Does the number of hits change after the drug is administered? If the number of hits goes down, this could mean:

* drug causes drowsiness
* drug causes blurred vision
* some other effect

If the number of hits go up, possibly this is some type of stimulant.

Find a 99% CI for μd to help answer the above question, where μd represents the average difference in hits for all people in the population of interest taking the drug.

The data set:

| **Subjects** | **Time 0** | **Time 4** | **di** | **(di-)2** |
| --- | --- | --- | --- | --- |
| **1** | 47 | 45 | 2 | 0.64 |
| **2** | 34 | 32 | 2 | 0.64 |
| **3** | 60 | 58 | 2 | 0.64 |
| **4** | 59 | 57 | 2 | 0.64 |
| **5** | 63 | 60 | 3 | 0.04 |
| **6** | 44 | 38 | 6 | 10.24 |
| **7** | 49 | 47 | 2 | 0.64 |
| **8** | 53 | 51 | 2 | 0.64 |
| **9** | 46 | 42 | 4 | 1.44 |
| **10** | 41 | 38 | 3 | 0.04 |
|  |  |  | 28 | 15.60 |

Then

**** = = 28/10 = 2.8



Let’s plot the data!

> time0 <- c(47, 34, 60, 59, 63, 44, 49, 53, 46, 41)

> time4 <- c(45, 32, 58, 57, 60, 38, 47, 51, 42, 38)

> cpt <- data.frame(time0, time4)

> cpt

time0 time4

1 47 45

2 34 32

3 60 58

4 59 57

5 63 60

6 44 38

7 49 47

8 53 51

9 46 42

10 41 38

> boxplot(x = time0 - time4, data = cpt, main = "Box and

dot plot", ylab = "Difference in hits", xlab = "",

ylim = c(0,8), pars = list(outpch=NA))

> stripchart(x = cpt$time0 - cpt$time4, lwd = 2, col =

"red", method = "jitter", vertical = TRUE, pch = 1,

main = "Dot plot", add = TRUE)



> hist(x = cpt$time0 - cpt$time4, main = "CPT data", xlab

= "Difference in hits", freq = FALSE, xlim = c(0,6))

> curve(expr = dnorm(x = x, mean = mean(cpt$time0 –

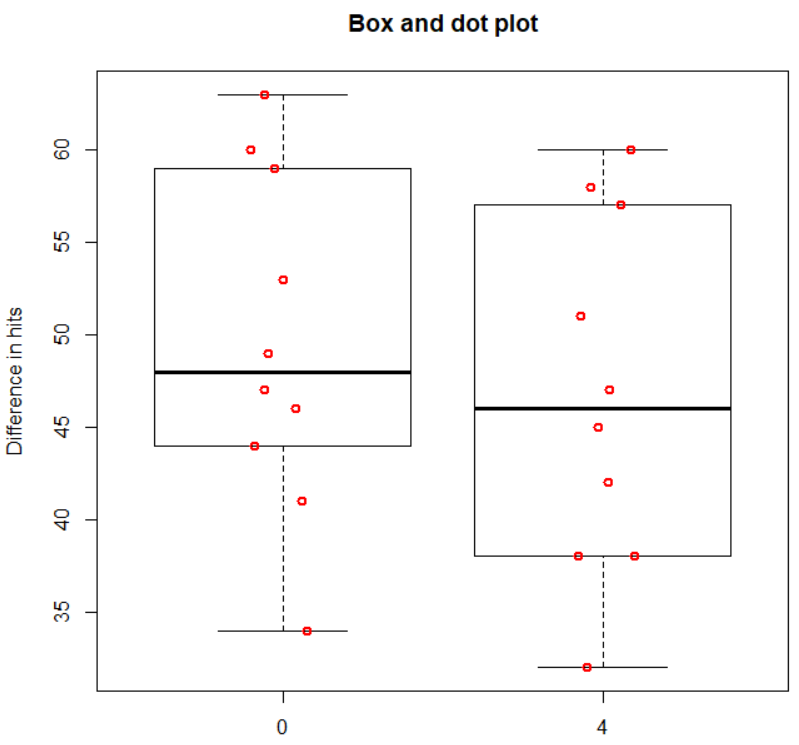
cpt$time4), sd = sd(cpt$time0 - cpt$time4)), col =

"red", add = TRUE)



Questions:

* + Why didn’t I look at separate plots for hits at each time?



* + Do you think there is a change in the mean number of hits?
  + What do you think about the normal probability distribution approximation?
  + Is the normal probability distribution for each population needed?

If the normal probability distribution is not appropriate, there are other statistical methods that one could use. For example, one could use a non-parametric test, like a Wilcoxon test (conclusion here does not change), that does not have distributional assumptions. There are whole courses on these types of methods!

We will proceed with the assumption that the normal probability distribution assumption is approximately satisfied.

Below is my R code and output:

> d <- cpt$time0 - cpt$time4

> dbar <- mean(d)

> s.d <- sd(d)

> alpha <- 0.01

> n <- length(d)

> qt(p = 1 - alpha/2, df = n-1)

[1] 3.2498

> lower <- dbar - qt(p = 1 - alpha/2, df = n-1) \* s.d /

sqrt(n)

> upper <- dbar + qt(p = 1 - alpha/2, df = n-1) \* s.d /

sqrt(n)

> data.frame(lower, upper)

lower upper

1 1.447 4.153

The 99% CI is





1.4470 < μd < 4.1530

I am 99% confident that the mean difference in the number of hits is between 1.4470 and 4.1530.

Questions:

* Is there sufficient evidence to indicate a difference in the mean number of hits? If so, is there a positive or negative difference and what does it mean in terms of the problem?
* What would you expect the plots to look like if there was not enough evidence to indicate a difference between the mean hits?

Remember that the confidence level DOES NOT MEAN there is a 99% probability that this CI contains μd. Instead, it means that if many more similar samples were taken, we would expect on average about 99% of the CIs to contain μd.

Below is an easier way to have R perform the calculations:

> t.test(x = time0, y = time4, data = cpt, conf.level =

0.99, paired = TRUE)

Paired t-test

data: time0 and time4

t = 6.7254, df = 9, p-value = 8.602e-05

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

99 percent confidence interval:

1.447 4.153

sample estimates:

mean of the differences

2.8

If the same function was used with the object d, we would get the same answers:

> t.test(x = d, conf.level = 0.99)

One Sample t-test

data: d

t = 6.7254, df = 9, p-value = 8.602e-05

alternative hypothesis: true mean is not equal to 0

99 percent confidence interval:

1.447 4.153

sample estimates:

mean of x

2.8

which is done here only to illustrate the equivalence of a paired data confidence interval and an interval for one population mean.

What about hypothesis testing?

Tests can be performed using confidence intervals, test statistics, and p-values in a similar manner as we have seen before, but now using the differences!

Example: CPT (cpt.R)

Notice that we have used (time 0 hits) – (time 4 hits) (i.e., di = yi1 - yi2 where 1 = time 0 and 2 = time 4). Perform a hypothesis test to determine if there was any type of an effect on the average hits at the different times.

CI Method using α = 0.01:

1. Ho:μd = 0  
   Ha:μd ≠ 0
2. The 99% CI is 1.4470 < μd < 4.1530
3. Reject Ho because 0 is not in the interval.
4. There is sufficient evidence to indicate a difference between the average hits at the two time periods.

Of course, you could also perform the hypothesis test using the test statistic and p-value methods.

Test statistic: Note that



Thus, the test statistic to be used here replaces the random variables with their observed values for what is in the middle of the inequality above:



Notice that μd is replaced with 0. This is done because most often we will hypothesize Ho:μd = 0 or Ho:μd ≤ 0 or Ho:μd ≥ 0.

P-value: 2×P(T>|t|) where T has n – 1 degrees of freedom.

Example: CPT (cpt.R)

Test statistic method:

1. Ho:μd = 0  
   Ha:μd ≠ 0
2. 
3. ±t0.005,9 = ±3.2498



Reject Ho because 6.7254 > 3.2498.



1. There is sufficient evidence to indicate a difference between the average hits for the two time periods. Thus, the drug is having an effect on the patients.

P-value method:

1. Ho:μd = 0  
   Ha:μd ≠ 0
2. 2×P(T > |6.7254|) = 8.6 \* 10-5
3. α = 0.01
4. Reject Ho because 8.6 \* 10-5 < 0.01
5. There is sufficient evidence to indicate a difference between the average hits for the two time periods. Thus, the drug is having an effect on the patients.

What if the independent sample methods of earlier in this chapter were used? Below is the R code and output:

> t.test(x = time0, y = time4, data = cpt, conf.level =

0.99, paired = FALSE)

Welch Two Sample t-test

data: time0 and time4

t = 0.6687, df = 17.969, p-value = 0.5122

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

99 percent confidence interval:

-9.2552 14.8552

sample estimates:

mean of x mean of y

49.6 46.8

As you can see from the large p-value, the conclusions would have been quite different!

For this example, it may have more appropriate to perform a one-tail test. Specifically, suppose it was desired to determine if the drug may have a particular type of detrimental effect (hits go down after administered). The hypotheses would be

Ho:μd ≤ 0  
Ha:μd > 0

because a positive mean represents that time 4 is less than time 0. Notice that the type of error that I am guarding against here with α = 0.01. The alternative = "greater" argument needs to be used with t.test():

> t.test(x = time0, y = time4, data = cpt, conf.level =

0.99, paired = TRUE, alternative = "greater")

Paired t-test

data: time0 and time4

t = 6.7254, df = 9, p-value = 4.301e-05

alternative hypothesis: true difference in means is greater than 0

99 percent confidence interval:

1.6253 Inf

sample estimates:

mean of the differences

2.8

The confidence interval is 1.63 < μd < ∞ (different from before), the test statistic is 6.7254 (same as before), and the p-value is 4.3 × 10-5 (half as much as before). Why do these items change or stay the same? What are the conclusions from this test?