MIE237 January 26-27 Labs

Neil Montgomery

January 25, 2016

Summary of you will do in this lab

You'll do some textbook questions using "paired" t procedures. Read the questions for background, but otherwise don't bother with what the book asks. Do what I ask here.

1. 9.94 - produce a 95% confidence interval for the difference between the two means.
2. 10.43 - perform the hypothesis test with the null hypothesis that there is no difference between mean absolute time differences under the two experimental conditions.
3. 10.45 - perform the hypothesis test with the null hypothesis that there is no difference between mean fuel consumption under the two experimental conditions.

Then you'll produce some 95% confidence intervals for proportions using the simulated data from the January 22 lecture. Details below.

The usual advice

I've told you where to get the textbook data. The PDF of this lab doesn't show all the code, but the `.Rmd` source file of the lab does. Data analysis consists of some graphical and/or numerical exploration, the analysis itself, a verification of assumptions, and a conclusion/interpretation.

"Paired" t procedure fully worked example

We'll look at 9.92 from the book. We looked at this one in class. While not part of the question or analysis, consider carefully why this experiment must be analyzed as a paired procedure, and try to think of an experimental design that would have answered the same question but using two independent samples.

Here is a numerical summary of the pairs.

	n	mean	sd
	12	40.58333	15.79101

To be honest I can't think of a worthwhile plot for only 12 numbers.

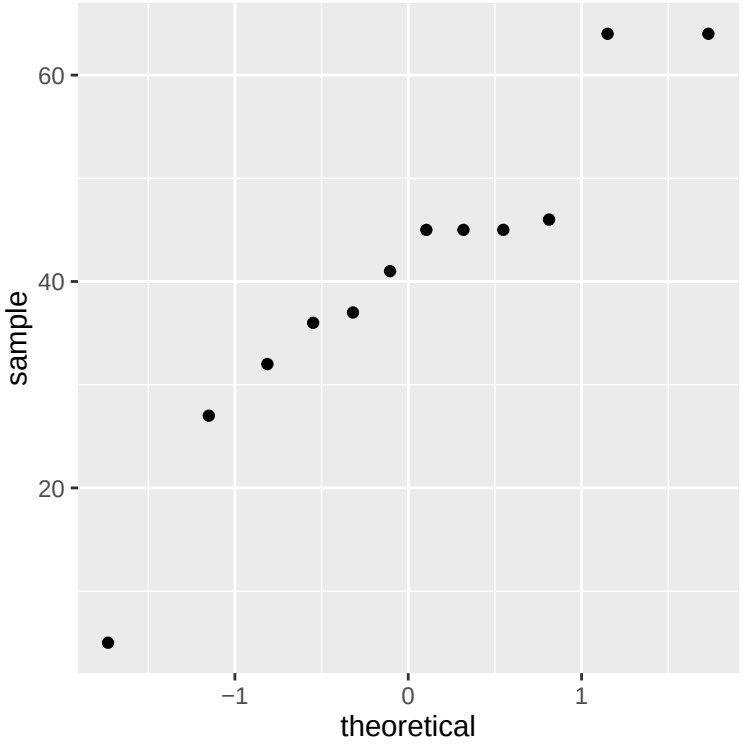
Here is the analysis for the mean difference in calcium.

```
##
##  One Sample t-test
##
## data:  calc_diff$Difference
## t = 8.9028, df = 11, p-value = 2.331e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  30.55020 50.61646
## sample estimates:
## mean of x
##  40.58333
```

The 95% confidence interval is [30.55, 50.62].

We need to verify the normality assumption. Here is a normal quantile plot of the data.



Again, with only 12 points it is difficult to say, but it seems no wild deviation from normality is present, so the calculations are probably valid.

Confidence interval for a proportion examples

Aside—lots of ways to produce this C.I.

The interval given in class is not actually implemented in base R. There's an R package `binom` dedicated to the many different ways to calculate a confidence interval for a proportion, which is an area of active statistical research. They are all approximations, and they all have their pros and cons. For example, suppose $n = 100$ and the number of "successes" is $k = 30$. Here is a table of 11 (slightly) different intervals:

```
library(binom)
kable(binom.confint(30, 100), row.names = TRUE)
```

	method	x	n	mean	lower	upper
1	agresti-coull	30	100	0.3000000	0.2186514	0.3961460

	method	x	n	mean	lower	upper
2	asymptotic	30	100	0.3000000	0.2101832	0.3898168
3	bayes	30	100	0.3019802	0.2143713	0.3917767
4	cloglog	30	100	0.3000000	0.2135522	0.3910559
5	exact	30	100	0.3000000	0.2124064	0.3998147
6	logit	30	100	0.3000000	0.2184030	0.3966128
7	probit	30	100	0.3000000	0.2168949	0.3950896
8	profile	30	100	0.3000000	0.2160309	0.3940967
9	lrt	30	100	0.3000000	0.2159984	0.3941141
10	prop.test	30	100	0.3000000	0.2145426	0.4010604
11	wilson	30	100	0.3000000	0.2189489	0.3958485

The one given in class is the "asymptotic" one. R's own built-in function is called `prop.test`.

Here are your options: 1. Just use the built-in `prop.test` as-is (accepting that it will be slightly different from your hand calculations with the class formula.) 2. Implement the class formula yourself, as it is very simple. 3. Install the `binom` package and ask it for the interval it calls `asymptotic`.

Trivial textbook example - 9.51

This is a rote textbook exercise with $n = 1000$ and $k = 228$ the "number heated by oil". The question is already asked in the form of a "numerical summary" and there is no dataset to explore.

I'll answer the question all three ways as outlined above.

1. The function `prop.test` gives the interval [0.203, 0.256].
2. The class formula can be implemented as follows:

```
n <- 1000
k <- 228
p_hat <- k/n
z <- qnorm(0.975)

(lower <- p_hat - z * sqrt(p_hat*(1 - p_hat)/1000))
```

```
## [1] 0.201997
```

```
(upper <- p_hat + z * sqrt(p_hat*(1 - p_hat)/1000))
```

```
## [1] 0.254003
```

3. Or I could have used `binom.confint` from the `binom` package to get the class formula as follows:

```
binom.confint(228, 1000, methods = "asymptotic")
```

```
##       method   x    n  mean    lower    upper
## 1 asymptotic 228 1000 0.228 0.201997 0.254003
```

To check the normality assumptions we note that $n\hat{p}$ and $n(1-\hat{p})$ both exceed 5, so the approximation is good.

Simulated data from class

The simulated data from class is closer to the kind of data one would encounter for estimating proportions. I've included the data in a spreadsheet as part of this lab. Look at the January 22 lecture slide source for ideas on numerical and graphical summaries.

Your specific tasks (in addition to the summaries, assumption verification, and conclusion) are toproduce 95% confidence intervals for the following proportions: 1. The proportion of gas mains that are made of Steel. 2. The proportion of gas mains that are under High pressure.