HW-Section2

Rob McCulloch

July 11, 2021

1 Homework for Section 2

1.1 The Canadian Returns

The csv file conret.csv has monthly returns from 22 portfolios where each portfolio invests in firms from a specifice country.

```
cd = read.csv("http://www.rob-mcculloch.org/data/conret.csv")
print(dim(cd))
```

```
## [1] 107  23
```

```
head(cd)
```

```
##     date australia austria belgium canada denmark finland france germany
## 1 198802      0.01    0.00    0.25   0.07    0.04    0.02   0.21    0.16
## 2 198803      0.17    0.05    0.04   0.05    0.02    0.09  -0.06    0.00
## 3 198804      0.06    0.01    0.01   0.02    0.00    0.01   0.09   -0.02
## 4 198805      0.15   -0.05   -0.05  -0.04    0.09    0.06   0.07   -0.01
## 5 198806     -0.03   -0.03    0.01   0.08    0.00   -0.01   0.01    0.00
## 6 198807      0.06    0.03   -0.05  -0.02   -0.01   -0.01  -0.03    0.01
##   honkong irleland italy japan malaysia netherlands newzealand norway singapore
## 1    0.02    -0.01  0.06  0.08    -0.05        0.04      -0.16   0.08     -0.02
## 2    0.06     0.09  0.04  0.08     0.06        0.04       0.22   0.13      0.06
## 3    0.02     0.01 -0.02  0.01     0.09        0.03       0.02   0.04      0.06
## 4   -0.03     0.11 -0.11 -0.05     0.03       -0.05       0.03  -0.05      0.02
## 5    0.08     0.00  0.04 -0.04     0.12        0.01      -0.03   0.01      0.09
## 6    0.02    -0.01  0.03  0.05     0.00        0.03       0.02  -0.01      0.02
##   spain sweden switzerland    uk   usa
## 1  0.01   0.04        0.06  0.00  0.04
## 2  0.09   0.07       -0.02  0.05 -0.03
## 3  0.00   0.03       -0.01  0.03  0.01
## 4  0.01   0.03       -0.03 -0.03  0.01
## 5  0.00  -0.05        0.01 -0.03  0.05
## 6 -0.03   0.03       -0.02  0.01  0.00
```

(a)

Do the monthly returns on the portfolio of Canadian assets *look normal* ?

(b)

Assuming the Canadian returns are approximately IID normal get the sample mean and standard deviation.

Using the sample mean and standard deviation as estimates of the normal mean and normal standard deviation (μ and σ) use our normal model to come up with a value for the probability that the next Canadian

return is positive.

Note, to read the data into R you can use:

```
cdat = read.csv("http://www.rob-mcculloch.org/data/conret.csv")
# cdat is dataframe.
names(cdat)
```

```
##  [1] "date"        "australia"   "austria"     "belgium"     "canada"
##  [6] "denmark"     "finland"     "france"      "germany"     "honkong"
## [11] "irleland"    "italy"       "japan"       "malaysia"    "netherlands"
## [16] "newzealand"  "norway"      "singapore"   "spain"       "sweden"
## [21] "switzerland" "uk"          "usa"
```

```
canret = cdat$canada
summary(canret)
```

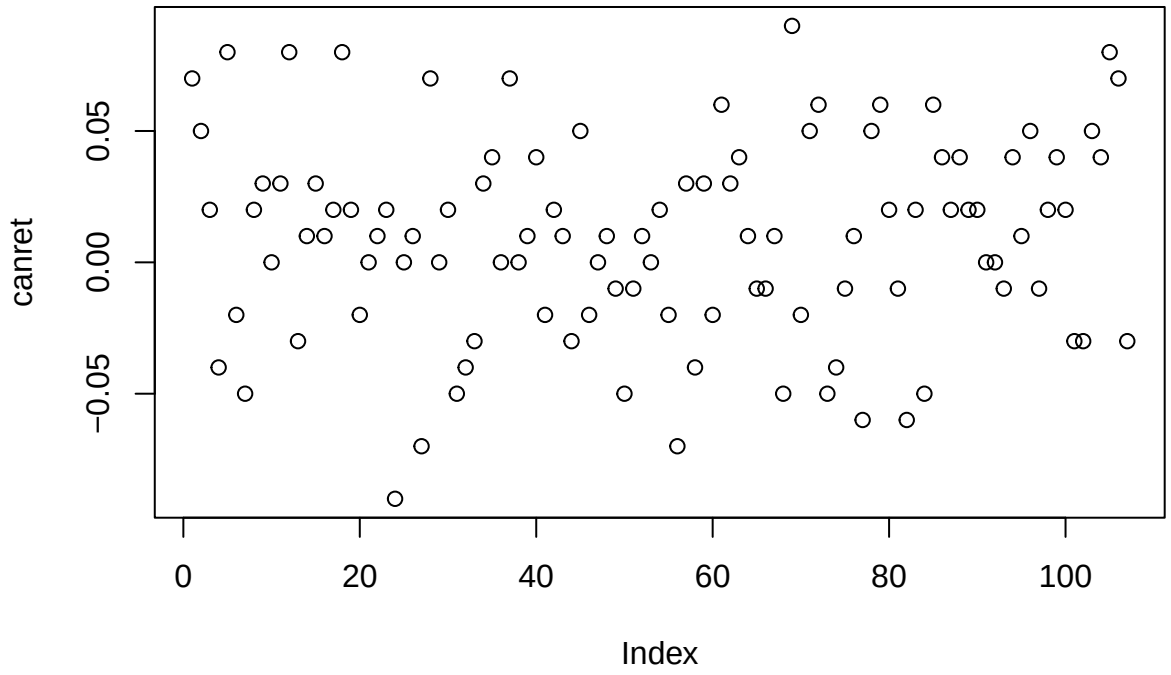
```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -0.090000 -0.020000  0.010000  0.009065  0.035000  0.090000
```

In Excel you would probably want to download the file conret.csv.

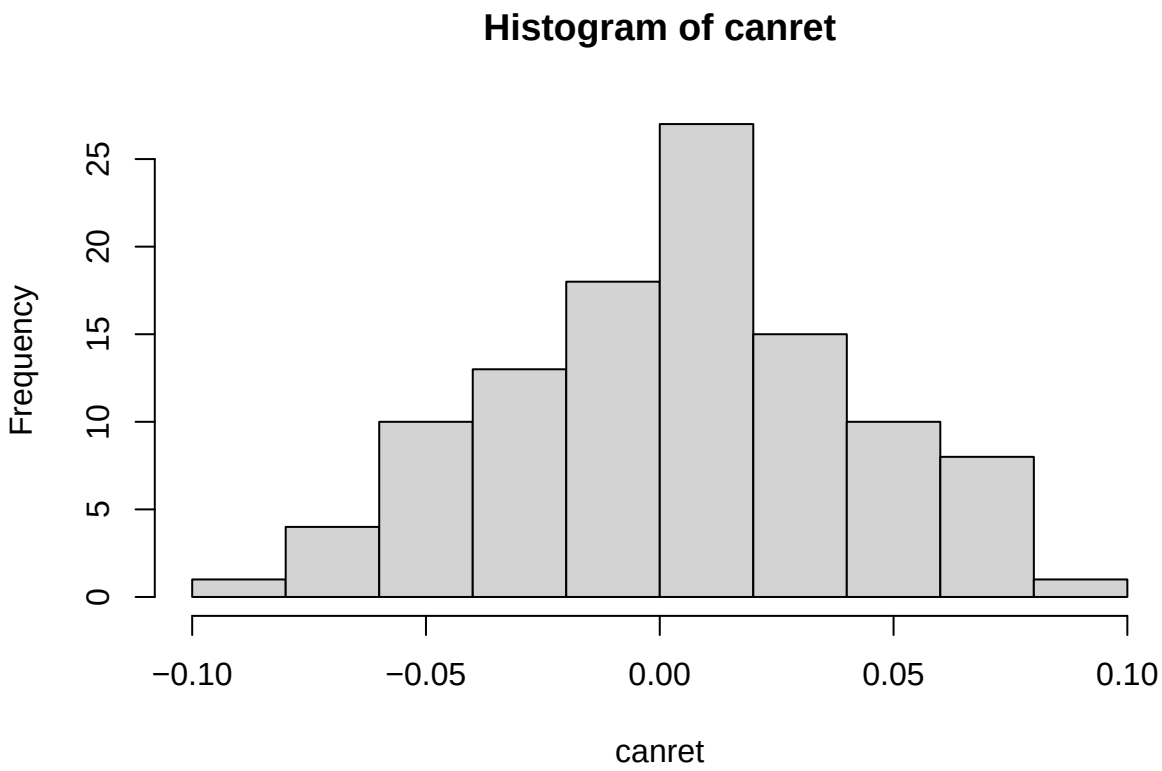
Solution

(a)

```
plot(canret)
```




```
hist(canret)
```



The is no obvious pattern in the sequence plot to the returns are plausibly IID draws.

The histogram has the bell shape, so the values look like draws from a normal.

(b)

Note that for parameter θ, we often use $\hat{\theta}$ to denote an estimate of θ.
So in the iid normal model, "$\bar{y}$ is $\hat{\mu}$" !!

```
muhat = mean(canret)
sighat = sd(canret)
pg0 = 1-pnorm(0,muhat,sighat)
cat("prob next return in positive: ", pg0)
```

```
## prob next return in positive:  0.5934895
```

That's actually pretty good!!

1.2 Return on the Fidelity Funds

Let's look at monthly returns on three different Fidelity funds.
These are in the file fidrest.csv.

```
ff = read.csv("http://www.rob-mcculloch.org/data/fidrets.csv")
print(names(ff))
```

```
## [1] "sp500"   "FidInc"  "FidVal"  "FidTech"
```

```
print(dim(ff))
```

```
## [1] 35  4
```

We have monthly returns on the sp500 and three different Fidelity funds.

(a)

Compute the sample means and standard deviation for each of the 4 returns columns.
What do these summaries tell us about the data?

(b)

Assumming we can model each of the four return types as IID normal, estimate each normal mean and standard deviation (μ, σ) using the sample quantities.

Plot the 4 estimated normal pdfs. How do the funds (and sp500) compare ?

Solution

(a)

```
cat("means:\n")
```

```
## means:
```

```
apply(ff,2,mean)
```

```
##       sp500      FidInc      FidVal     FidTech
## 0.014976037 0.004645065 0.012619036 0.016955382
```

```
cat("sds: \n")
```

```
## sds:
```

```
apply(ff,2,sd)
```

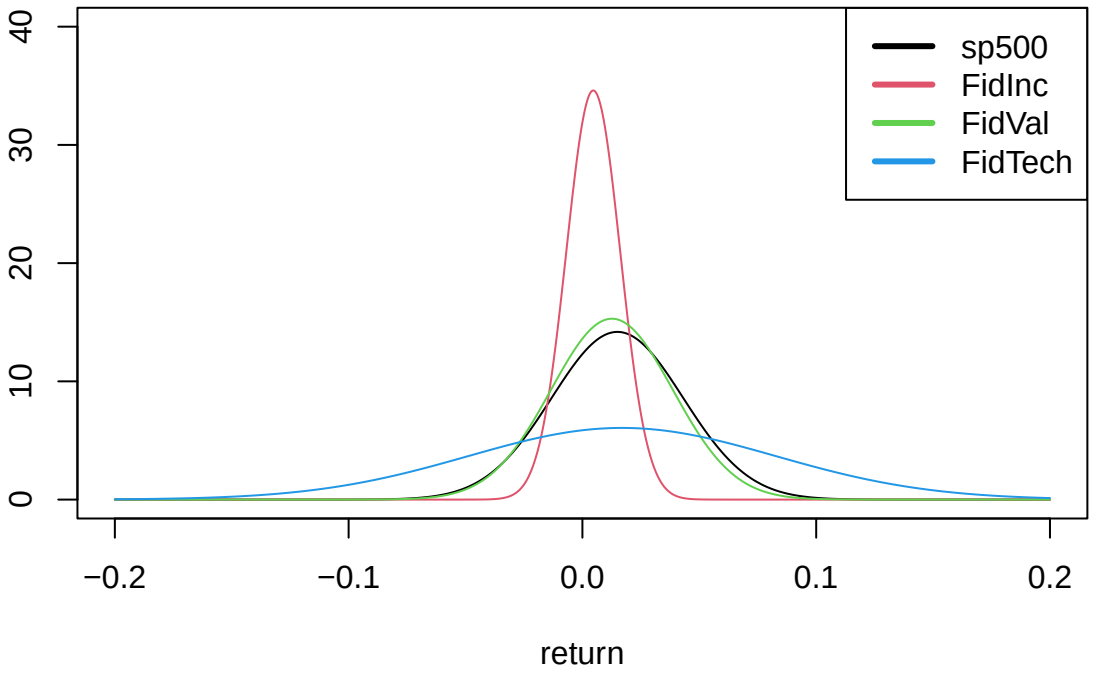
```
##      sp500     FidInc     FidVal    FidTech
## 0.02812769 0.01152790 0.02607556 0.06581641
```

Looking at the three Fidelity funds, the mean returns and standard deviations go up as we go from and *income fund* to a *value fund* to a *tech fund*.
Just going by the names, this makes sense as a tech fund is probably riskier than an income fund.

The sp500 mean and sd look pretty good! The mean is in between the value fund and the tech fund but the sd looks closer to the value fund.

(b)

```
mvec = apply(ff,2,mean)
svec = apply(ff,2,sd)
rvals = seq(from=-.2,to=.2,length.out=1000)
plot(c(-.2,.2),c(0,40),type="n",xlab="return",ylab="")
for(i in 1:4) {
   lines(rvals,dnorm(rvals,mvec[i],svec[i]),col=i)
}
legend("topright",legend=names(ff),co=1:4,lwd=rep(3,4))
```



A pretty dramatic story about how the returns differ.

1.3 The Audit

You manage the process that fills boxes of cereal.

You are about to be audited !!!

The audit means that they will take a sample of 10 boxes and see how much cereal goes in them.

If the *average* weight of the 10 weights is in the interval (330,370) you pass the audit.

As a good quality engineer, you have done statistical analysis of the process and feel that a good model is:

$$W \sim N(345, 15^2), \quad IID$$

That is, the amount of cereal going into the boxes looks like IID draws from an normal with $\mu = 345$ and $\sigma = 15$.

(a)

Let $\bar{W}$ represent the average weight from the audit.

$\bar{W}$ is a random variable !!

Given your model, what is the distribution of $\bar{W}$?

(b)

Plot the density (pdf) of $\bar{W}$. Add vertical lines indicating the region where you would pass the audit.

Does it look like you will pass the audit?

(c)

What is the probability that you pass the audit?

(d)

Give an interval such that there is a 95% chance that $\bar{W}$ will be in it.

(e)

Suppose you pass the audit if the average of 100 boxes is in (330,370).
Give an interval such that there is a 95% chance that $\bar{W}$ will be in it.
How does this interval compare to the one in part (d)?

Solution

(a)

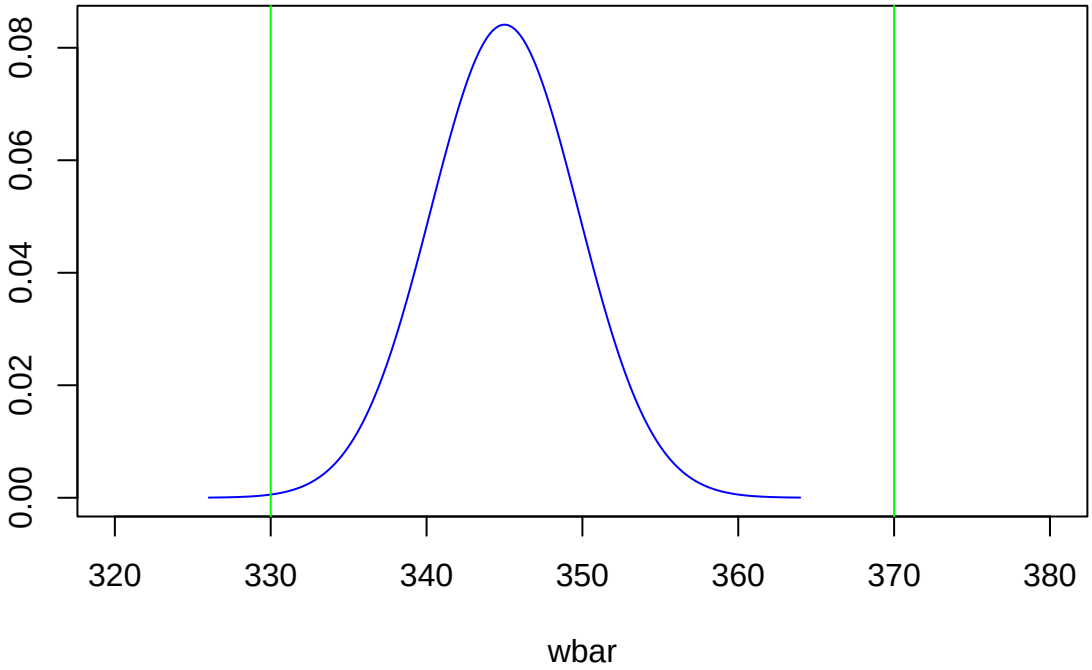
```
sigma = 15; n=10
wbarsigma = sigma/sqrt(n)
print(wbarsigma)
```

```
## [1] 4.743416
```

$$\bar{W} \sim N(345, 4.74^2)$$

(b)

```
mu = 345
wvec = seq(from = mu-4*wbarsigma,to=mu+4*wbarsigma,length.out=1000)
plot(wvec,dnorm(wvec,mu,wbarsigma),type="l",col="blue",xlim=c(320,380),ylab="",xlab="wbar")
abline(v=330,col="green")
abline(v=370,col="green")
```



Looks like there is a good chance you will pass the audit!!

(c)

```
pnorm(370,mu,wbarsigma) - pnorm(330,mu,wbarsigma)
```

```
## [1] 0.9992172
```

(d)

```
wbarsigma = 15/sqrt(10)
mu + 2*wbarsigma*c(-1,1)
```

```
## [1] 335.5132 354.4868
```

```
cat("width of the interval is: ",4*wbarsigma)
```

```
## width of the interval is:  18.97367
```

(e)

```
wbarsigma = 15/sqrt(100)
mu + 2*wbarsigma*c(-1,1)
```

```
## [1] 342 348
```

```
cat("width of the interval is: ",4*wbarsigma)
```

```
## width of the interval is:  6
```

The width of the interval is 6 which is much smaller than 18.97.

1.4 Confidence interval from the Audit Data

The audit takes place and is based on 100 observations.

We can read the data into R:

```
ad100 = read.csv("http://www.rob-mcculloch.org/data/audit100.csv")
print(names(ad100))
```

```
## [1] "x"
```

```
ad100 = ad100$x
print(length(ad100))
```

```
## [1] 100
```

```
summary(ad100)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   291.7   335.2   345.7   344.8   354.5   381.6
```

(a)

What are the sample mean and sample standard deviation of the 100 weights?

(b)

Get the sequence plot and histogram of the 100 weights.

(c)

Using the sample mean as the estimate of the true mean, what is the associated standard error?

(d)

What is the 95% confidence interval for the true mean based on the 100 observations and the sample mean estimator?

Solution

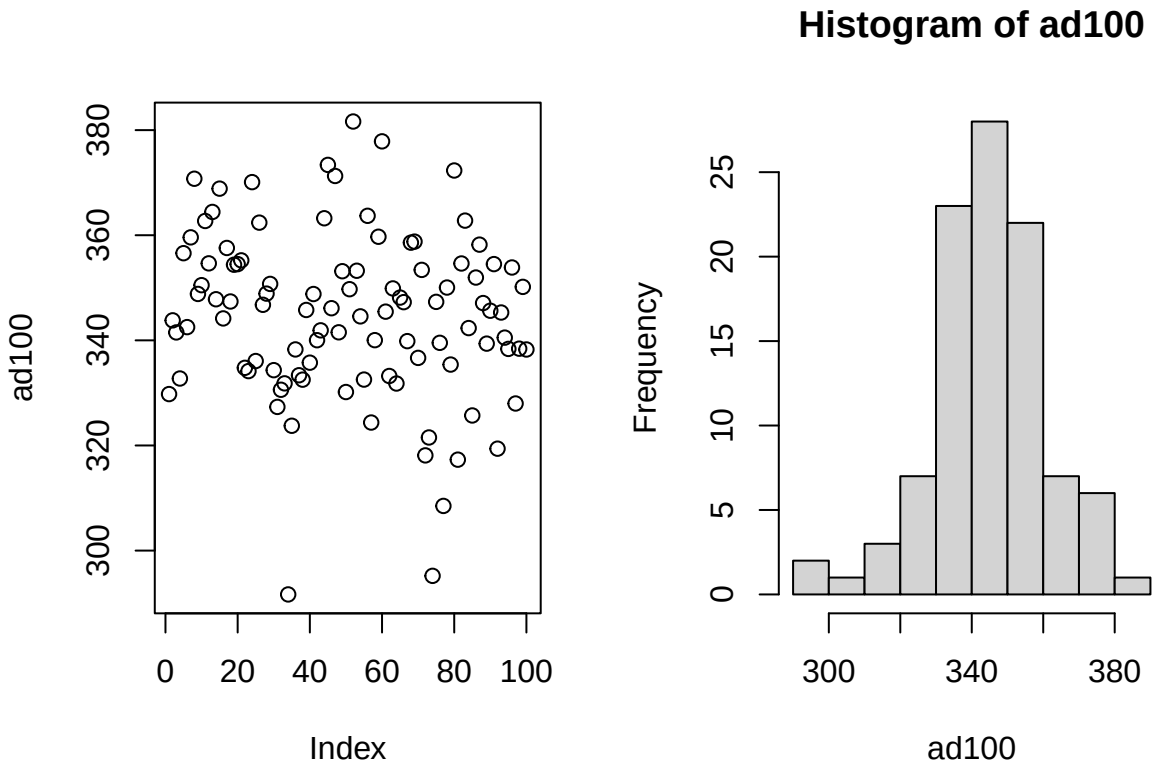
(a)

```
cat("the sample mean and sd are:",mean(ad100),sd(ad100),"\n")
```

```
## the sample mean and sd are: 344.7709 15.86704
```

(b)

```
par(mfrow=c(1,2))
plot(ad100)
hist(ad100)
```



(c)

```
sterr = sd(ad100)/sqrt(100)
cat("sterr is:",sterr,"\n")
```

```
## sterr is: 1.586704
```

which is, of course, the sample standard deviation divided by 10.

(d)

```
mean(ad100) + 2*sterr*c(-1,1)
```

```
## [1] 341.5975 347.9443
```

Note also:

```
t.test(ad100)
```

```
##
##  One Sample t-test
##
```

```
## data:  ad100
## t = 217.29, df = 99, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  341.6225 347.9193
## sample estimates:
## mean of x
##  344.7709
```

1.5 USA Mean Return

Assuming the returns on usa are iid normal, get the 95% confidence for the true mean return.

(use the data conret.csv)

Solution

The mean return is 0.0135 and the sample standard deviation is 0.0332.

So, the standard error of the mean $se(\bar{y})$ is
0.0332/sqrt(107) = 0.003217

The 95% CI is 0.0135 +/- 2(0.003217)=
= 0.0135 +/- 0.006435 $\approx$ .0135 +/- .0064 = (0.0071, 0.0199).

Big!!

Note that the function to do this in `R` is `t.test`.

```
temp = read.csv('http://www.rob-mcculloch.org/data/conret.csv',header=T)$usa
t.test(temp)
```

```
##
##  One Sample t-test
##
## data:  temp
## t = 4.1826, df = 106, p-value = 5.954e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.007078809 0.019837078
## sample estimates:
##  mean of x
## 0.01345794
```

1.6 Tokyo Level

Get the data tokyo_sub.csv from the webpage.

This data is a time-series of daily levels of a Japanese stock index.

(a)

Do the time-series plot of the levels.

Do the time-series plot of the difference of the levels.
(for example the first two values are 10743 and 10760, so the first difference is 10760 - 10743 = 17)

Do the histogram of the difference of the levels.

Which one could be iid normal?

Note:

```
temp = c(1,1,2,4)
diff(temp)
```

```
## [1] 0 1 2
```

(b)

For the rest of this question, assume we are modeling the differences of the levels as iid normal.

Let's call this variable D.

$D_t \sim N(\mu_D, \sigma_D^2)$.

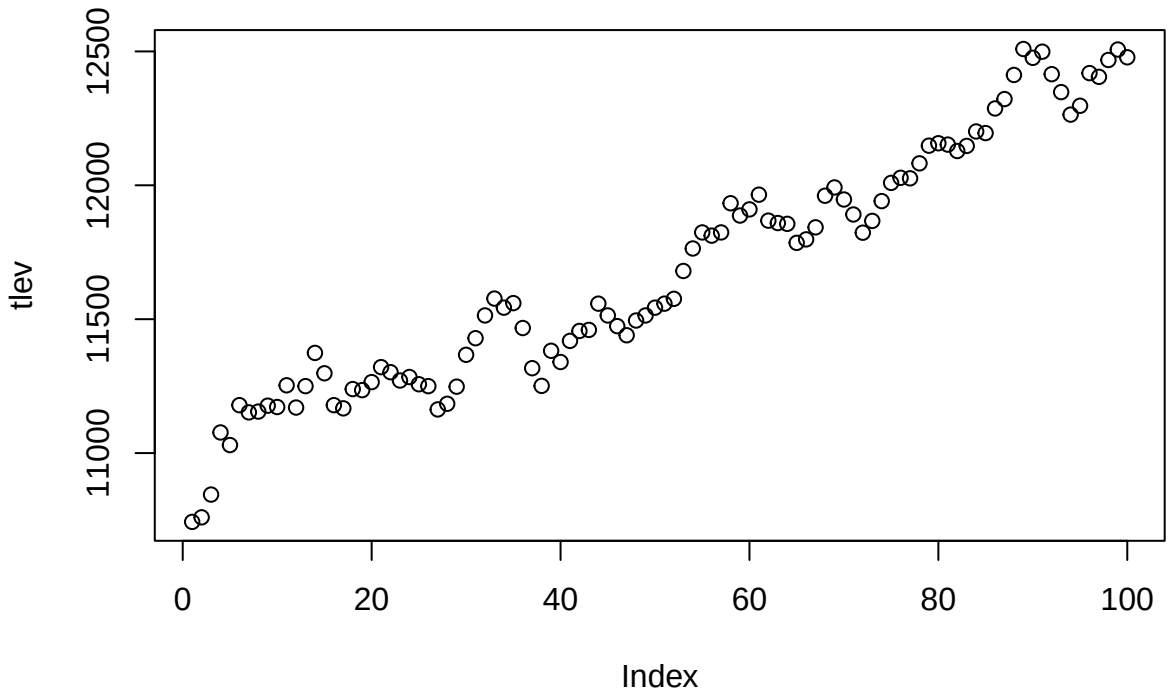
So, for example, D_2 turned out to be 17.\ (note that there is no D_1 !!).

Give the 95% confidence interval for μ_D.

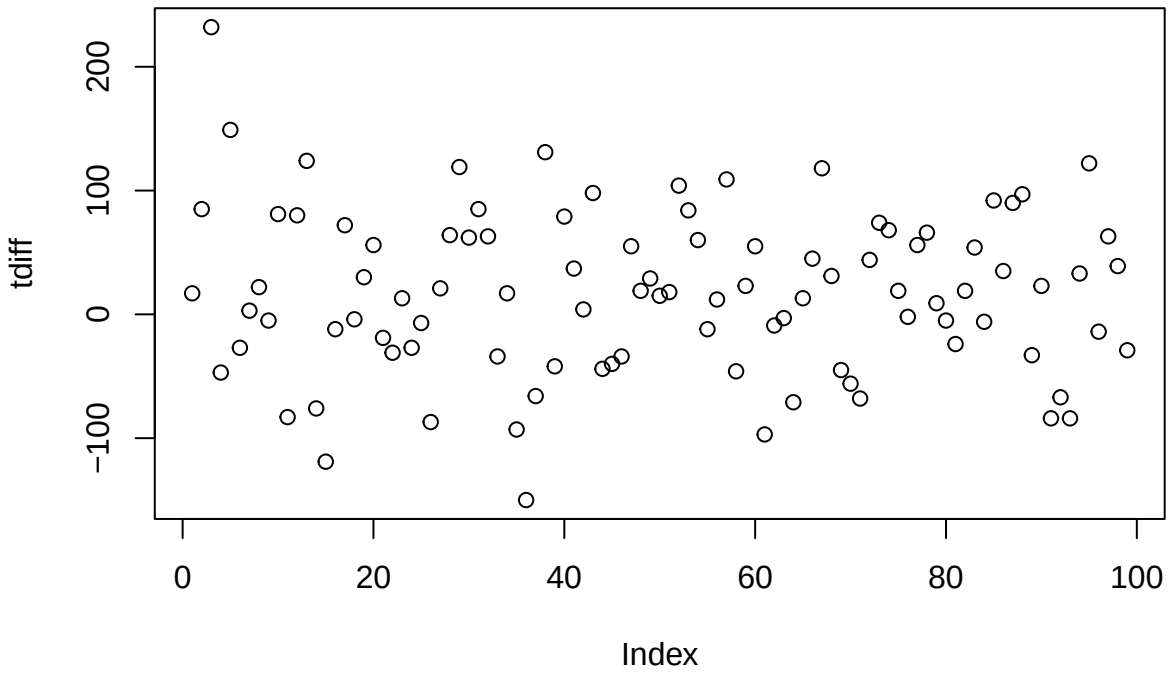
Solution

(a)

```
tlev = read.csv("http://www.rob-mcculloch.org/data/tokyo_sub.csv")$level
tdiff = diff(tlev)
plot(tlev)
```

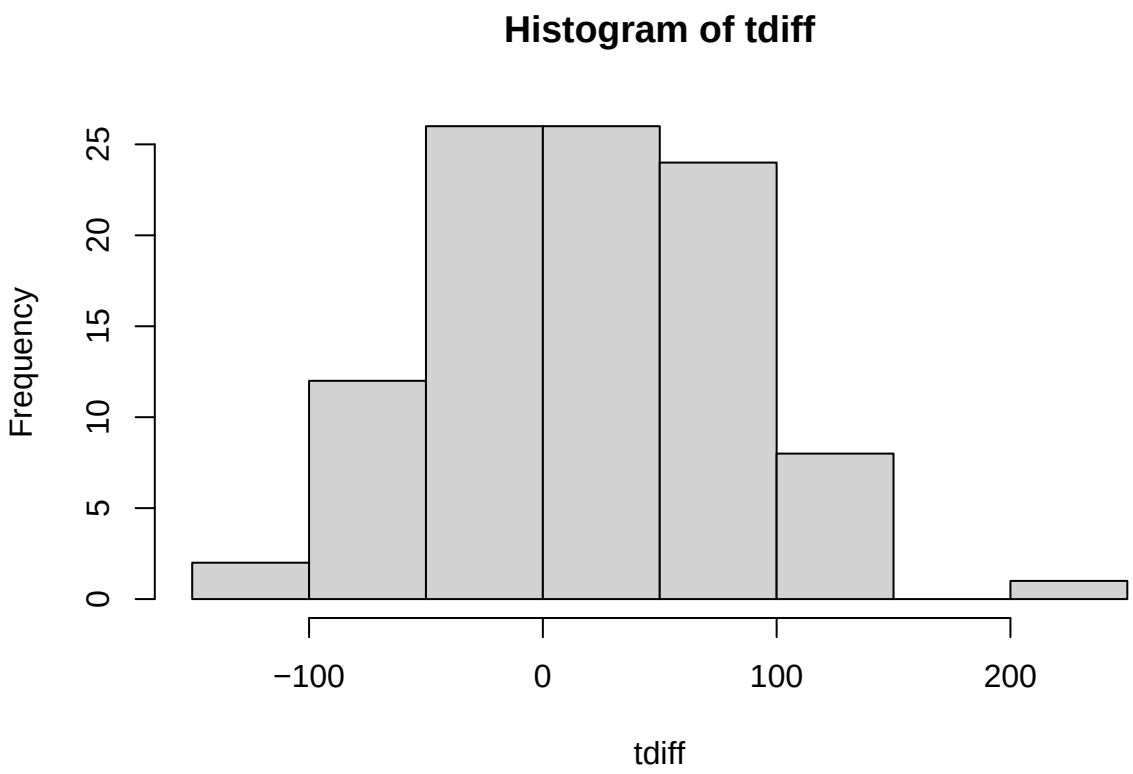



```
plot(tdiff)
```



The level is wandering up, definitely not IID.
Difference could be IID!!

```
hist(tdiff)
```



Could be IID normal!!!

(b)

```
t.test(tdiff)
```

```
##
##  One Sample t-test
##
## data:  tdiff
## t = 2.6579, df = 98, p-value = 0.009182
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   4.440487 30.610018
## sample estimates:
## mean of x
##  17.52525
```

1.7 CI-for-proportion-of-up-ticks

Suppose we are modeling whether or not a stock price goes up each month as iid Bernoulli(p).

We observed that out of the last 99 months, the price went up 62 time so that our estimate of p is $62/99 = 0.63$.

What is the 95% confidence interval for p = Prob(price goes up)?

Is it big?

Solution

```
se = sqrt(.63*(1-.63)/99)
print(2*se)
```

```
## [1] 0.09704732
```

```
print(se)
```

```
## [1] 0.04852366
```

```
.63 + 2*se*c(-1,1)
```

```
## [1] 0.5329527 0.7270473
```

1.8 Sample Size for Acceptable Error

Suppose you think an election is close.

You think that if you take a poll, you are likely to get a $\hat{p}$ (sample proportion) for Candidate A close to .5.

Since the election is close, you are thinking the usual $\pm.03$ for sample sizes of about 1,000 will be to big.

What sample size do you need to have a $\pm$ of .01?

Solution

We would like to have an n such that $2\sqrt{\frac{.5(.5)}{n}} = .01$
$\frac{1}{\sqrt{n}} = .01$
$\sqrt{n} = 100$
$n = 10000$.

check:

```
2*sqrt(.5*.5/10000)
```

```
## [1] 0.01
```

1.9 Equality of Proportions

(a)

We have a random sample of 1000 women voters.
550 of them say they will vote Democrat.

What is your 95% confidence interval for the true proportion of women voters who would vote Democrat?

(b)

We have a random sample of 1000 male voters.
450 of them say they will vote Democrat.

What is your 95% confidence interval for the true propotion of male voters who would vote Democrat?

(c)

Do you think the true proportion of women voters that would vote Democrat is equal to the true proportion for men?

Let p_{men} be the true population proportion for men and p_{women} be the true population proportion for women.

Here is the R output for inferring about the difference in two probabilities (or proportions).

The confidence interval (-.14, -.055) is for $p_{men} - p_{women}$.
What does the confidence interval tell us about the difference in probabilities?

```
prop.test(c(450,550),n=c(1000,1000))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(450, 550) out of c(1000, 1000)
## X-squared = 19.602, df = 1, p-value = 9.537e-06
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.14460645 -0.05539355
## sample estimates:
## prop 1 prop 2
##   0.45   0.55
```

Solution

(a)

```
prop.test(c(550),c(1000))
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  c(550) out of c(1000), null probability 0.5
## X-squared = 9.801, df = 1, p-value = 0.001744
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.5185315 0.5810796
## sample estimates:
##    p
## 0.55
```

Let's check this with the formula we know for the standard error of $\hat{p}$.

```
phat = 550/1000
cat('phat is ',phat,'\n')
```

```
## phat is  0.55
```

```
sephat = sqrt(phat*(1-phat)/1000)
cat('se of phat is ',sephat)
```

```
## se of phat is  0.01573213
```

```
cat('confidence interval is:\n')
```

```
## confidence interval is:
```

```
print(phat + 2*sephat*c(-1,1))
```

```
## [1] 0.5185357 0.5814643
```

Which agrees with what we got from the R function prop.test.

(b)

```
prop.test(c(450),c(1000))
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  c(450) out of c(1000), null probability 0.5
## X-squared = 9.801, df = 1, p-value = 0.001744
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.4189204 0.4814685
## sample estimates:
##    p
## 0.45
```

(c)

The confidence interval suggests the difference $p_{women} - p_{men}$ is at least -.05 and could be as large as -.15 which is pretty big.

1.10 Testing the USA Mean

Previously we computed the 95% confidence interval for the true mean return of the USA porfolio using the data from the file conret.csv.

(a)

Test the hypothesis $H_0 : \mu = 0$ at level .05.

(b)

Test the hypothesis $H_0 : \mu = .1/12$ at level .05.

Solution

(a)

```
temp = read.csv('http://www.rob-mcculloch.org/data/conret.csv',header=T)$usa
n = length(temp)
cat("sample size is: ",n,"\n")
```

```
## sample size is:  107
```

```
usam = mean(temp)
usasd = sd(temp)
cat("mean and sd of usa are: ",usam,usasd,"\n")
```

```
## mean and sd of usa are:  0.01345794 0.03328275
```

```
se = usasd/sqrt(n)
tval = (usam-0)/se
cat("t test stat is: ",tval,"\n")
```

```
## t test stat is:  4.182649
```

Tstat is 4.2, clear reject.

Level .05 tells us to use 2 as our cutoff and 4.2 is much bigger than 2.

Using the `R` function `t.test` we have:

```
temp = read.csv('http://www.rob-mcculloch.org/data/conret.csv',header=T)$usa
t.test(temp)
```

```
##
##  One Sample t-test
##
## data:  temp
## t = 4.1826, df = 106, p-value = 5.954e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.007078809 0.019837078
## sample estimates:
##  mean of x
## 0.01345794
```

which gives us the same numbers.

(b)

```
munot = .1/12
tval1 = (usam - munot)/se
cat("taval for testing mu = .1/12 :",tval1,"\n")
```

```
## taval for testing mu = .1/12 : 1.592699
```

Tval is 1.6, fail to reject.

```
t.test(temp,mu=.1/12)
```

```
##
##  One Sample t-test
##
## data:  temp
## t = 1.5927, df = 106, p-value = 0.1142
```

```
## alternative hypothesis: true mean is not equal to 0.008333333
## 95 percent confidence interval:
##  0.007078809 0.019837078
## sample estimates:
##  mean of x
## 0.01345794
```

1.11 Testing the USA Mean with the p-value

Previously we computed the 95% confidence interval for the true mean return of the USA porfolio using the data from the file conret.csv.

We also looked at testing $H_0 : \mu = .1/12$.

Let's revisit the testing problem with the p-value.

(a)

Test the hypothesis $H_0 : \mu = .1/12$.

What is associated the p-value?
What does it tell you about the hypothesis test?

(b)

Test the hypothesis $H_0 : \mu = 0$.

What is the associated p-value?
What does it tell you about the hypothesis test?

Solution

(a)

```
temp = read.csv('http://www.rob-mcculloch.org/data/conret.csv',header=T)$usa
n = length(temp)
cat("sample size is: ",n,"\n")
```

```
## sample size is:  107
```

```
usam = mean(temp)
usasd = sd(temp)
cat("mean and sd of usa are: ",usam,usasd,"\n")
```

```
## mean and sd of usa are:  0.01345794 0.03328275
```

```
se = usasd/sqrt(n)
muo = .1/12
tval = (usam-muo)/se
cat("t test stat is: ",tval,"\n")
```

```
## t test stat is:  1.592699
```

The pvalue is

```
2*pnorm(-1.6)
```

```
## [1] 0.1095986
```

which is not small (e.g. not less than .05 for a .05 level test) so we fail to reject.

Let's check that we get the same numbers from the R t.test function.

```
t.test(temp,mu=.1/12)
```

```
##
##  One Sample t-test
##
## data:  temp
## t = 1.5927, df = 106, p-value = 0.1142
## alternative hypothesis: true mean is not equal to 0.008333333
## 95 percent confidence interval:
##  0.007078809 0.019837078
## sample estimates:
##  mean of x
## 0.01345794
```

(b)

```
t.test(temp)
```

```
##
##  One Sample t-test
##
## data:  temp
## t = 4.1826, df = 106, p-value = 5.954e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.007078809 0.019837078
## sample estimates:
##  mean of x
```

```
## 0.01345794
```

For testing $\mu = 0$, the p-value is tiny so we have a strong reject.

1.12 Testing Equality of Proportions

We have a random sample of 1000 women voters.
550 of them say they will vote Democrat.

We also have a random sample of 1000 male voters.
450 of them say they will vote Democrat.

Do you think the true proportion of women voters that would vote Democrat is equal to the true proportion for men?

Let p_{men} be the true population proportion for men and p_{women} be the true population proportion for women.

Here is the R output for testing $H_0 : p_{men} = p_{women}$.

The confidence interval (-.14, -.055) is for $p_{men} - p_{women}$.
What does the p-value tell you about the null hypothesis ?

```
prop.test(c(450,550),n=c(1000,1000))
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(450, 550) out of c(1000, 1000)
## X-squared = 19.602, df = 1, p-value = 9.537e-06
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.14460645 -0.05539355
## sample estimates:
## prop 1 prop 2
##   0.45   0.55
```

Solution

The p-value is tiny suggesting a strong reject of the null hypothesis.

1.13 AB Testing

In Data Science estimating the difference in means or proportions in two different situations is often called "A/B Testing''.

What Is A/B Testing?

I cut this from a webpage (https://www.optimizely.com/optimization-glossary/ab-testing/):
`A/B testing (also known as split testing or bucket testing) is a method of comparing two versions of a webpage or app against each other to determine which one performs better. AB testing is essentially an experiment where two or more variants of a page are shown to users at random, and statistical analysis is used to determine which variation performs better for a given conversion goal.`

Suppose for example you are considering two different versions of a webpage and you want to see how the "click-through'' depends on the version.

In one version an add is placed near the top of the page and in the other version it is placed near the bottom.

A click-through means someone who visits the page clicks on the add.

p_T:
Probability of a "click-through'' if you place your ad at the top of the web-page.

p_B:
Probability of a "click-through'' if you place your ad at the bottom of the web-page.

Things like Google analytics allow you to randomly move the ad around the page so that you can estimate the difference $p_T - p_B$!

In the internet age, the opportunity to experiment has exploded.

Suppose you randomly assigned the page position of your ad and 138 out of 5009 times the ad was on top you got a click-through and 97 out of 5116 times the ad was on the bottom you got a click through.

- estimated click-through rate on top: $\hat{p}_T = 138/5009 = 0.02755041$.
- estimated click-through rate on bottom: $\hat{p}_T = 97/5116 = 0.01896013$.

So a reasonable estimate of the difference is $\hat{p}_T - \hat{p}_B = 0.02755041 - 0.01896013 = 0.00859028$

So, it looks like you are getting a much better click-through rate with the ad on top, but, what is the uncertainty?

Here is the R output for prop.test applied to the data:

```
x=c(138 , 97)
n = c(5009, 5116)
prop.test(x,n)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
```

```
## data:  x out of n
## X-squared = 7.8636, df = 1, p-value = 0.005044
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.002517929 0.014662640
## sample estimates:
##     prop 1     prop 2
## 0.02755041 0.01896013
```

Note that:

```
(0.002517929+ 0.014662640)/2
```

```
## [1] 0.008590284
```

so the confidence interval is centered at the simple estimate of the difference we computed above.

What does the ouput of prop.test tell us about the difference in click-through rates?

Are they "significantly'' different?

Solution

The estimate of the treatment effect $p_T - p_B$ is .0086 and the 95% confidence interval for $p_T - p_B$ is (.0025,.0146)

The uncertainty is big, but there is a suggestion of a real difference.

Note that in this business small differences in click-through rates can translate into big differences in profits.

The test rejects the hypothesis that there is no difference in click-through rates.