M.Sc. Practical MT2005 week 3

Plots, tests and simulation

1 Robustness to non-normality of the 't tools'

How fast does the central limit theorem work?: slower than many people imagine!

```
M <- 40; p <- 0.2
z <- rbinom(1000, size=M, p=p)
cdf.compare(z, distribution = "binom", size=M, p=p)
cdf.compare(z, distribution = "norm", mean=M*p, sd=sqrt(M*p*(1-p)))
```

Repeat for larger values of M and more extreme values of p.

How does this affect the distribution of the 't tools'?

```
x <- rexp(25); y <- rexp(25)
t.test(x, y)
t.test(x, y)$statistic
## let's assemble results from 1000 trials
res <- numeric(1000)
for(i in 1:1000) {
  x <- rexp(25); y <- rexp(25)
  res[i] <- t.test(x, y)$statistic
}
library(MASS)
truehist(res)
lines(density(res, width="sj"))
xx <- seq(-4, 4, len=100)
lines(xx, dt(xx, df=48), col=4)

cdf.compare(res, distribution="t", df=48)

quantile(res, p=c(0.025, 0.975))
qt(c(0.025, 0.975), df=48)
table(cut(res, breaks=c(-Inf, qt(0.025, df=48), qt(0.975, df=48), Inf)))
```

The last calculation allows us to see how close the test is to its nominal size.

Try this again for other sample sizes (including different sample sizes for the two groups), and for samples from the Cauchy and t_4 distributions. Also try unequal variances (and the appropriate test).

2 Heights and Weights of 200 People

J. Fox (1997) reports measurements on 200 people from C. Davis, Departments of Physical Education and Psychology, York University, Canada. The data are in data frame `Davis` in library `car`. For a quick look try

```
library(car)
?Davis

summary(Davis)
pairs(Davis)
```

There is obviously a problem with one observation, a female whose measured height and weight appear to have been interchanged.

```
attach(Davis)
plot(height, weight)
identify(height, weight)
detach("Davis")
```

We can fix up the dataset: suppose it was observation 31 that was in error (it isn't)

```
nDavis <- Davis # make a copy
nDavis[31, 2:3] <- Davis[31, 3:2]
pairs(nDavis)
```

It will make more sense to look at males and female separately. There are many ways to do this: here are a few:

```
sex <- nDavis[,1]
pairs(nDavis[-1],
      panel = function(x, y) {
         points(x[sex=="M"], y[sex=="M"], pch=25)
         points(x[sex=="F"], y[sex=="F"], pch=26)
      })

pairs(nDavis[-1],
      panel = function(x, y) {
         points(x[sex=="M"], y[sex=="M"], col=6)
         points(x[sex=="F"], y[sex=="F"], col=4)
      })

## if no point show up, quit and start a new session
trellis.device()
splom(~nDavis[-1])
splom(~nDavis[-1], groups = sex, panel=panel.superpose)
splom(~nDavis[-1] | sex, strip=function(...) strip.default(style=1, ...))
dev.off()
```

Now suppose we are interested in the correlation between height and weight, and between measured and reported weight.

```
cor(nDavis[2:5], na.method="available")
```

Note that we have to do something with the missing values in the self-reported quantities. The result is somewhat misleading, as the correlations within males and females are rather smaller.

```
cor(nDavis[sex == "M", 2:5], na.method="available")
cor(nDavis[sex == "F", 2:5], na.method="available")
```

Although it is easy to do each of two sexes separately, there are general ways to get results by group, for example

```
by(nDavis[2:5], sex, function(x) cor(x, na.method="available"))
```

(By the way, why are females listed first?)

Can we give a confidence interval for the correlation?

```
attach(nDavis)
cor.test(height, weight)
```

From the t statistic (where did that come from?) we can construct a 95% confidence interval by Fisher's $\tanh$ transform. Since we will need this again, it is helpful to write a short function to compute it.

```
cor.CI <- function(x, y) {
  r <- cor.test(x, y)$estimate
  n <- length(x)
  c(tanh(c(atanh(r) + qnorm(0.025)/sqrt(n-3),
           atanh(r) + qnorm(0.975)/sqrt(n-3))))
}
cor.CI(height, weight)
```

Can we believe the confidence interval? Let us simulate the problem. We have 200 measurements of (height, weight) pairs, with correlation 0.771.

```
mu <- sapply(nDavis[2:3], mean)
Sigma <- var(nDavis[2:3])

library(MASS)  # for mvrnorm
?mvrnorm
sim <- mvrnorm(200, mu=mu, Sigma=Sigma)
plot(sim)
cor.CI(sim[, "height"], sim[, "weight"])
```

That simulates the problem once: Now let's do it 1000 times.

```
set.seed(some_value)  # so you can repeat this
res <- matrix(, 1000, 2)
for(i in 1:1000) {
  sim <- mvrnorm(200, mu=mu, Sigma=Sigma)
  res[i, ] <- cor.CI(sim[, "height"], sim[, "weight"])
}
```

Now let's see how well we did

```
covered <- (res[, 1] < 0.771) & (res[, 2] > 0.771)
table(covered)
```

and take a look at some of the confidence intervals

```
plot(c(0.6, 0.9), c(1, 100), xlab="confidence interval", ylab="", type="n")
abline(v = 0.771, lty=2)
res100 <- res[1:100, ]
res100 <- res100[sort.list(res100[, 1]), ]  # sort on left end
segments(res100[, 1], 1:100, res100[, 2], 1:100)
```

Why are the intervals shorter at the top?

Is there a significant difference in measured and reported height?

```
nDavis0 <- na.omit(nDavis)  #  to get rid of the missing values
attach(nDavis0)
t.test(height, repht)
```

(Is that the right test?) What about weight?

Suppose we had a smaller sample: try this for various values of M

```
M <- 20
(nos <- sample(seq(along=height), size = 20, replace = F))
t.test(height[nos], repht[nos])
```

Let's repeat the experiment 250 times and look at the results.

```
tstat <- numeric(250)
for(i in 1:250) {
  nos <- sample(seq(along=height), size = M, replace = F)
  tstat[i] <- t.test(height[nos], repht[nos])$statistic
}
truehist(tstat)
lines(density(tstat, width="sj"))
```

Now add a theoretical density for comparison.

3 Salaries of Bank Clerks

Roberts (1979) reported the starting salaries of all 32 male and all 61 female entry-level skilled clerical employees at a bank between 1969 and 1977.

```
library(Sleuth)
attach(case0102)
bankF <- as.vector(SALARY[SEX=="FEMALE"])
bankM <- as.vector(SALARY[SEX=="MALE"])
detach()
bankM
bankF
```

The issue is if female employees are paid systematically less than males. It is simple to test that

```
var.test(bankM, bankF)
t.test(bankM, bankF, alternative = "greater")
t.test(bankM, bankF, alternative = "greater", var.equal = F)

wilcox.test(bankM, bankF, alternative="greater")
```

and the difference can be shown graphically

```
stem(bankF)
stem(bankM)
par(mfrow=c(2,1))
truehist(bankF, xlim=c(3000, 7000))
rug(jitter(bankF))
lines(density(bankF, n=500, width="sj"))
truehist(bankM, xlim=c(4000, 9000))
rug(jitter(bankM))
lines(density(bankM, n=500, width="sj"))
par(mfrow=c(1,1))
```

which shows that perhaps we should worry about the discreteness of the data (and even the Mann–Whitney test assumes continuous distributions).

We can do a randomization test, randomly allocating gender to the $32 + 61$ salaries

```
salary <- case0102[, "SALARY"]
M <- 1000
tstat <- numeric(M)
for(i in 1:M) {
   male <- sample(32+61, 32)
   tstat[i] <- t.test(salary[male], salary[-male],
     alternative = "greater", var.equal = F)$statistic
}
summary(tstat)
par(mfrow=c(1,2))
truehist(tstat)
xx <- seq(-4, 4, len=100); lines(xx, dt(xx, df=51))
cdf.compare(tstat, distribution = "t", df = 51)
par(mfrow=c(1,1))
```