Statistical Inference week3 again

T confidence intervals

Est +- TQ * SEest

Student t distribution

X - mu / (S/sqrt(n))

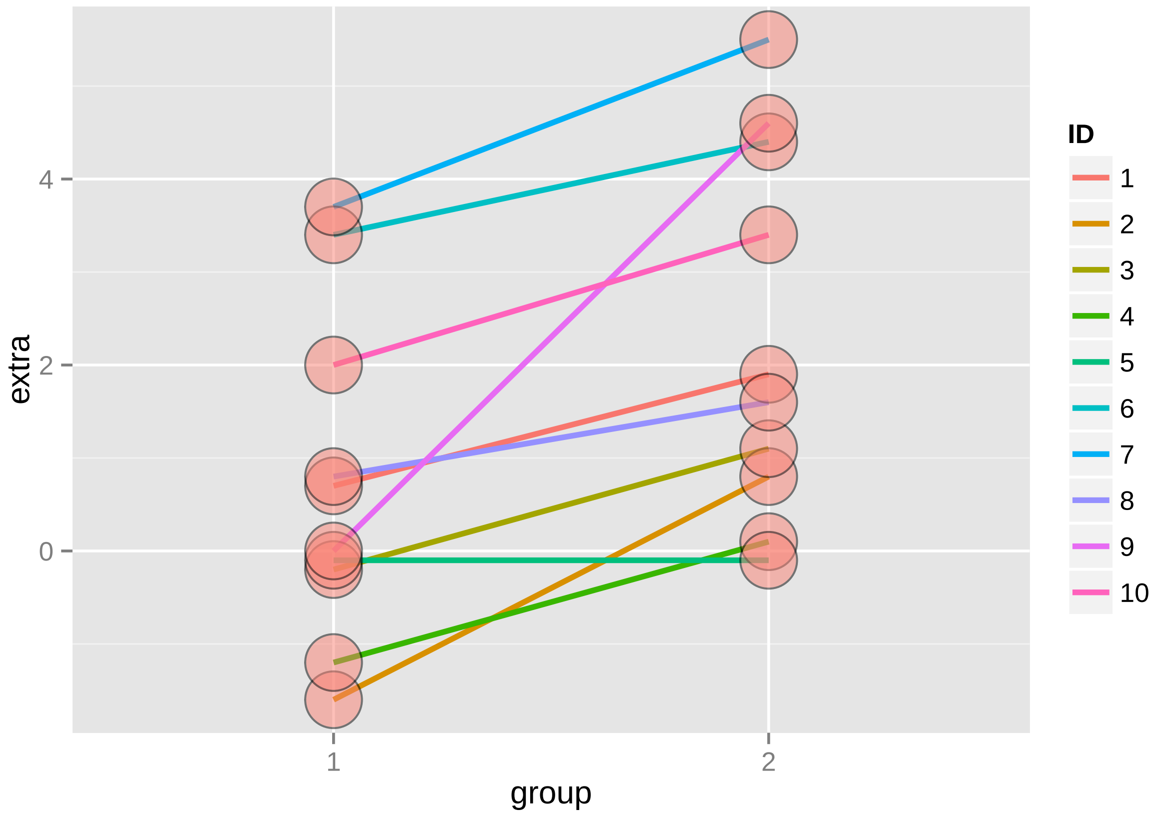
Sleep data

Comes from Gosset's Biometrika paper.
Shows the increase hours in patients sleeping time on two soporic drugs.

```
data(sleep)
head(sleep)
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

```
library(ggplot2)
g <- ggplot(sleep, aes(x = group, y = extra, group = factor(ID)))
g <- g + geom_line(size = 1, aes(colour = ID)) + geom_point(size =10, pch = 21, fill = "salmon", alpha =
g
```



Plotting the data

The difference comparing group 1 and group 2, versus comparing the differnce of each subject (as it should be)

```
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]
difference <- g2 - g1
mn <- mean(difference); s <- sd(difference); n <- 10
```

```
mn + c(-1,1) * qt(0.975,n-1)*s/sqrt(n)
```

Results

```
## [1] 0.7001142 2.4598858
```

```
t.test(difference)
```

```
##
##  One Sample t-test
##
## data:  difference
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  0.7001142 2.4598858
## sample estimates:
## mean of x
##      1.58
```

```r
t.test(g2,g1, paired = T)
```

```
##
##  Paired t-test
##
## data:  g2 and g1
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7001142 2.4598858
## sample estimates:
## mean of the differences
##                    1.58
```

```r
t.test(extra ~ I(relevel(group,2)), paired = T, data = sleep)
```

```
##
##  Paired t-test
##
## data:  extra by I(relevel(group, 2))
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7001142 2.4598858
## sample estimates:
## mean of the differences
##                    1.58
```

Independent group t intervals

Suppose one group recived the drug, the other the placebo.

This isn't a paired test, as the subjets are different in each group.

In this case,

$$\overline{Y} - \overline{X} \pm t_{n_x+n_y-2,1-2/\alpha} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}$$

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2)$$

```r
sp <- sqrt((7*15.34^2 + 20*18.23^2)/(8+21-2))
132.86 - 127.44 + c(-1,1) * qt(0.975, 8+21-2) * sp * (1/8+1/21)^.5
```

```
## [1] -9.521097 20.361097
```

Act as if the sleep group were independent trials

```r
n1 <- length(g1); n2 <- length(g2)
sp <- sqrt(((n1-1)*sd(g1)^2 + (n2-1)*sd(g2)^2)/(n1+n2-2))
md <- mean(g2) - mean(g1)
semd <- sp * sqrt(1/n1 + 1/n2)
md + c(-1,1)*qt(.975, n1 + n2 -2) * semd
```

```
## [1] -0.203874  3.363874
```

```r
t.test(g2,g1,paired = T)
```

```
##
##  Paired t-test
##
## data:  g2 and g1
## t = 4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7001142 2.4598858
## sample estimates:
## mean of the differences
##                     1.58
```

```r
t.test(g2,g1,paired = F,var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  g2 and g1
## t = 1.8608, df = 18, p-value = 0.07919
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.203874  3.363874
## sample estimates:
## mean of x mean of y
##      2.33      0.75
```

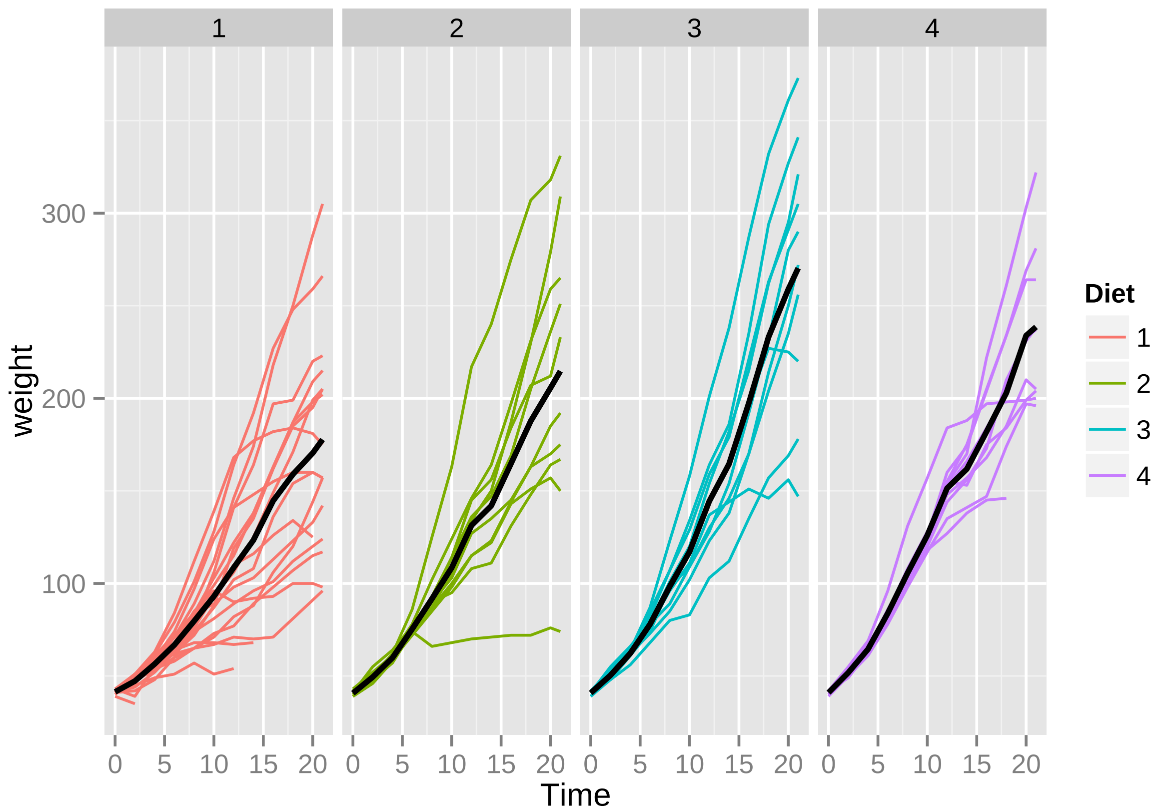
```r
library(datasets); data(ChickWeight); library(reshape2)
#head(ChickWeight)
#table(ChickWeight)
#summary(ChickWeight)
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1:2)] <- paste("time", names(wideCW[-(1:2)]), sep="")
#head(wideCW)
library(dplyr)
```

Chickweight

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
wideCW <- mutate(wideCW, gain = time21 - time0)
```

```
g <- ggplot(ChickWeight, aes(x = Time, y = weight,
                             colour = Diet, group = Chick))
g <- g + geom_line()
g <- g + stat_summary(aes(group = 1), geom = "line", fun.y = mean, size = 1, col = "black")
g <- g + facet_grid(. ~ Diet)
g
```

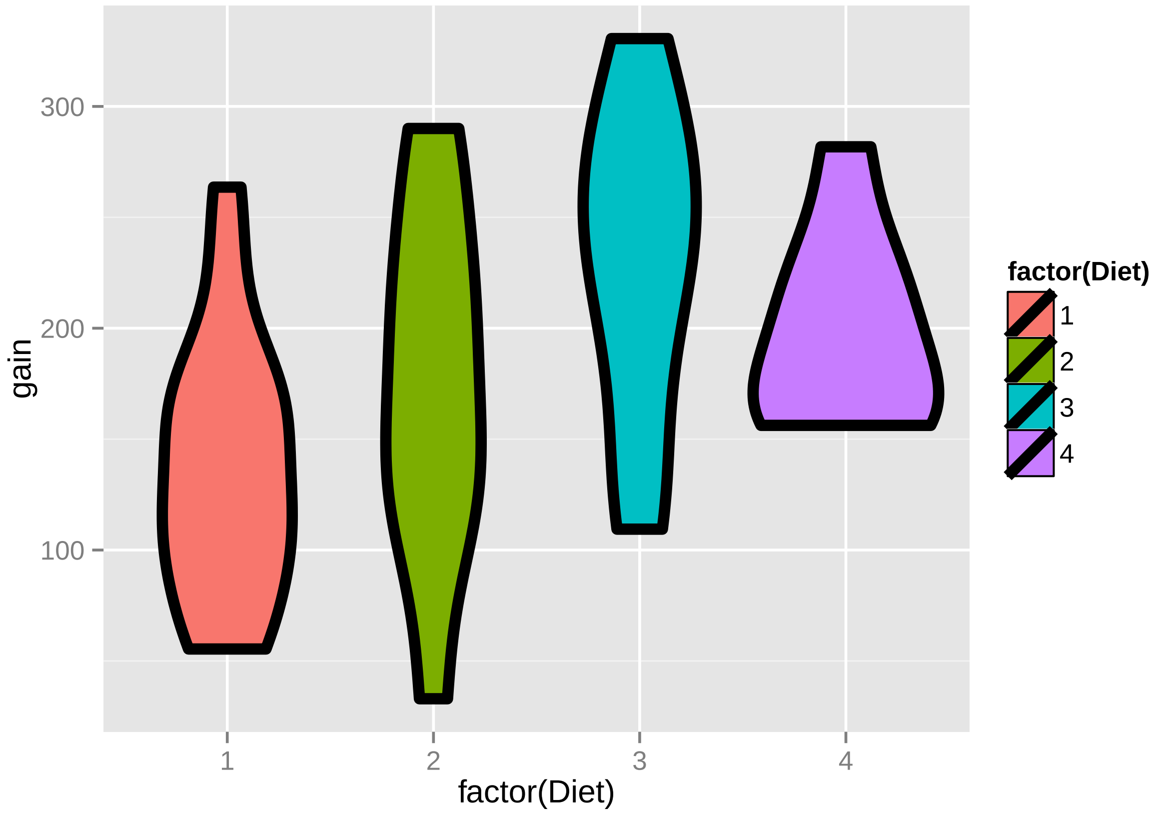


Plotting

Is the 4th diet faster than the 1st?

```
g <- ggplot(wideCW, aes(x = factor(Diet), y = gain, fill = factor(Diet)))
g <- g + geom_violin(col = "black", size = 2)
g
```

```
## Warning in loop_apply(n, do.ply): Removed 5 rows containing non-finite
## values (stat_ydensity).
```




```
wideCW14 <- subset(wideCW, Diet %in% c(1,4))
rbind(
        t.test(gain~Diet, paired = F, var.equal = T, data = wideCW14)$conf,
        t.test(gain~Diet, paired = F, var.equal = F, data = wideCW14)$conf
)
```

```
##             [,1]      [,2]
## [1,] -108.1468 -14.81154
## [2,] -104.6590 -18.29932
```

differnet variance in two groups

$$\overline{Y} - \overline{X} \pm t_{df} \times \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^{1/2}$$

$$df = \frac{(S_x^2/n_x + S_y^2/n_y)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{S_y^2}{n_y}\right)^2/(n_y - 1)}$$

In R

```r
132.86 - 127.44 + c(-1,1) *2.13 * (15.34^2/8 + 18.23^2/21)^(1/2)
```

```
## [1] -8.906499 19.746499
```

```r
# t.test(..., var.equal = F)
```

Hypothesis testing 5% is a benchmark. sd = 10, n = 100, so sd' = 1
if the average is 32, it's larger than 30 + 1.645
usually, you do the

$$\sqrt{n}(\overline{X} - \mu_0)/s > Z_{1-\alpha}$$

```r
qt(.95,15)
```

```
## [1] 1.75305
```

```r
library(UsingR); data(father.son)
```

```
## Loading required package: MASS
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
##
## Loading required package: HistData
## Loading required package: Hmisc
## Loading required package: grid
## Loading required package: lattice
## Loading required package: survival
## Loading required package: Formula
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     combine, src, summarize
##
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
##
##
## Attaching package: 'UsingR'
##
## The following object is masked from 'package:survival':
##
##     cancer
##
## The following object is masked from 'package:ggplot2':
##
##     movies
```

```r
t.test(father.son$sheight - father.son$fheight)
```

```
##
##  One Sample t-test
##
## data:  father.son$sheight - father.son$fheight
## t = 11.789, df = 1077, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8310296 1.1629160
## sample estimates:
## mean of x
## 0.9969728
```

This was a paired test.

```r
library(datasets); data(ChickWeight); library(reshape2)
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-c(1:2)] <- paste("time", names(wideCW)[-(1:2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
                 gain = time21 - time0
)
wideCW14 <- subset(wideCW, Diet %in% c(1,4))
t.test(gain ~ Diet, paired = F, var.equal = T, data = wideCW14)
```

two group t testing

```
##
##  Two Sample t-test
##
## data:  gain by Diet
## t = -2.7252, df = 23, p-value = 0.01207
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -108.14679  -14.81154
## sample estimates:
## mean in group 1 mean in group 4
##        136.1875        197.6667
```

```r
t.test(gain ~ Diet, paired = F, var.equal = F, data = wideCW14)
```

```
##
##  Welch Two Sample t-test
##
## data:  gain by Diet
## t = -2.9615, df = 20.937, p-value = 0.007464
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
```

```
## -104.65901 -18.29932
## sample estimates:
## mean in group 1 mean in group 4
##        136.1875        197.6667
```

Exact binomial test

P values

If the p-value is less than

$$\alpha$$

you reject the null hypothesis

```r
pt(2.5,15,lower.tail = F)
```

```
## [1] 0.0122529
```

```r
choose(8,7) * 0.5^8 + choose(8,8) * 0.5^8
```

```
## [1] 0.03515625
```

```r
pbinom(6, size = 8, prob = 0.5, lower.tail = F)
```

```
## [1] 0.03515625
```

```r
ppois(9,5, lower.tail = F)
```

```
## [1] 0.03182806
```

```r
mean <- 1100
sd <- 30
n <- 9
?t.test
mean + c(-1,1) * sd/sqrt(n) * qt(0.975,n-1)
```

Test

```
## [1] 1076.94 1123.06
```

```r
n <- 9
mean <- -2
sd <- mean * -1 * sqrt(n) / qt(0.975,n-1)
sd
```

```
## [1] 2.601903
```

```r
n <- 10
nh <- 3
nv <- 0.6
oh <- 5
ov <- 0.68

sp <- function(nx, sx, ny, sy){
        sqrt(((nx - 1)*sx^2 + (ny - 1)*sy^2)/(nx + ny -2))
}
sp_here <- sp(10, sqrt(nv), 10, sqrt(ov))

nh - oh + c(-1,1) * qt(0.975, 18) * sp_here * sqrt(1/10 + 1/10)
```

```
## [1] -2.751649 -1.248351
```

```r
nh - oh + c(-1,1) * qt(0.95, 18) * sp_here * sqrt(1/10 + 1/10)
```

```
## [1] -2.620397 -1.379603
```

```r
n <- 9
t <- -3
p <- 1
vt <- 1.5
vp <- 1.8

sp_here <- sp(n, vt, n, vp)
t - p + c(-1,1) * qt(0.95, 16) * sp_here * sqrt(1/9 + 1/9)
```

```
## [1] -5.363579 -2.636421
```

```r
t - p + c(-1,1) * qt(0.9, 16) * sp_here * sqrt(1/9 + 1/9)
```

```
## [1] -5.044041 -2.955959
```

```r
n <- 100
nw <- 4
nsd <- 0.5
ow <- 6
osd <- 2

df <- function(sx, nx, sy, ny){
        (sx^2/nx + sy^2/ny)^2 / {(sx^2/nx)^2/(nx - 1) + (sy^2/ny)^2/(ny - 1)}
}
df_here <- df(nsd, n, osd, n)
ow - nw + c(-1,1) * qt(0.975, df_here) * sqrt(nsd^2/n + osd^2/n)
```

```
## [1] 1.591503 2.408497
```