Statistical Inference week3

T distruibution

$$\frac{\overline{X}-\mu}{S/\sqrt{n}}$$

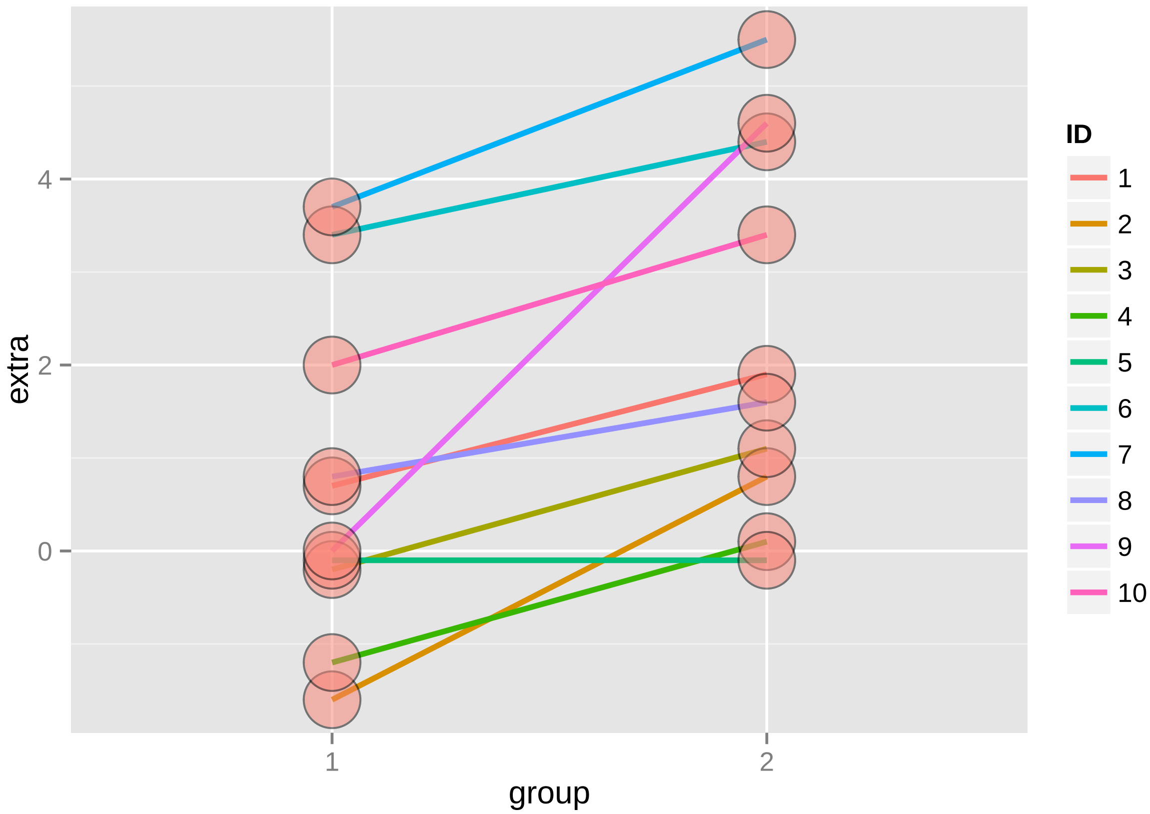
- will be a t distribtion, not a regular distruibution
- t always has a wider bit at the edges.
- In large scale, it's nearly the same
- for poisson, other stuff are available

Sleep data

```
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
```




```
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(.975, n-1)* s / sqrt(n)
```

```
## [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
```

```
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
```

Checking the 95% confidence interval for the mean of the paried test.
The t test in 4 sentences.

Independent group t confidence trials

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

This is what happens for 2 groups

The variance

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

- this assumes constant variance among groups

calculate from a book
8 oral contraceptive users, 21 controlls
Xoc = 132.86, soc = 15.34 Xc = 127.44, sc = 18.23

```r
sp <- sqrt((7 * 15.34^2 + 20*18.23^2) / (8 + 21 - 2))
132.86 - 127.44 + c(-1,1) * qt(.975, 27) * sp * (1/8 + 1/21)^.5
```

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

The answer is [-9.52, 20.4]
This contains 0, you can't rule out that they're the same

```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)
rbind(
        md + c(-1,1) * qt(.975, n1 + n2 -2) * semd,
        t.test(g2, g1, paired = F, var.equal = T)$conf,
        t.test(g2, g1, paired = T)$conf
)
```

Sleep patients again

```
##             [,1]     [,2]
## [1,] -0.2038740 3.363874
## [2,] -0.2038740 3.363874
## [3,]  0.7001142 2.459886
```

The answer
[,1] [,2]
[1,] -0.2038740 3.363874
[2,] -0.2038740 3.363874
[3,] 0.7001142 2.459886

In a paired test, if you take g2 - g1, the answer is the 3rd row.

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

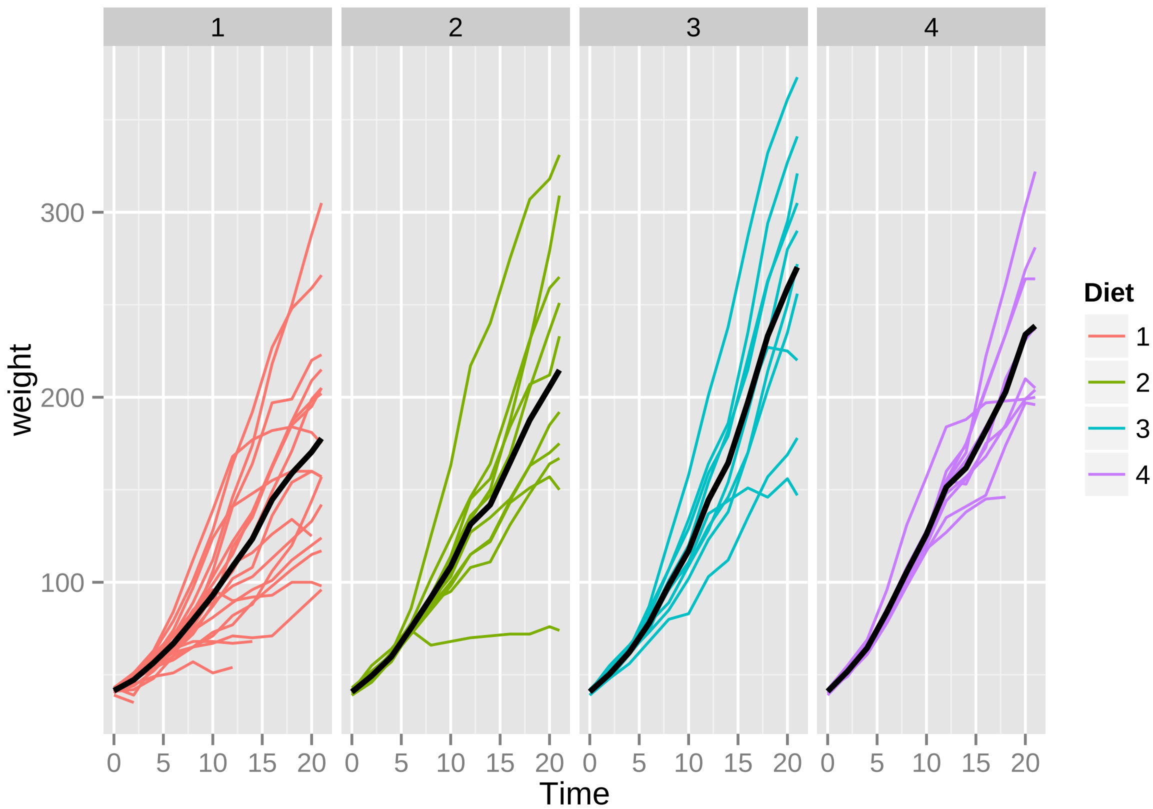
```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

Plotting the raw data

```r
g <- ggplot(ChickWeight, aes(x = Time, y = weight,
                             colour = Diet, group = Chick))
g <- g + geom_line()
```

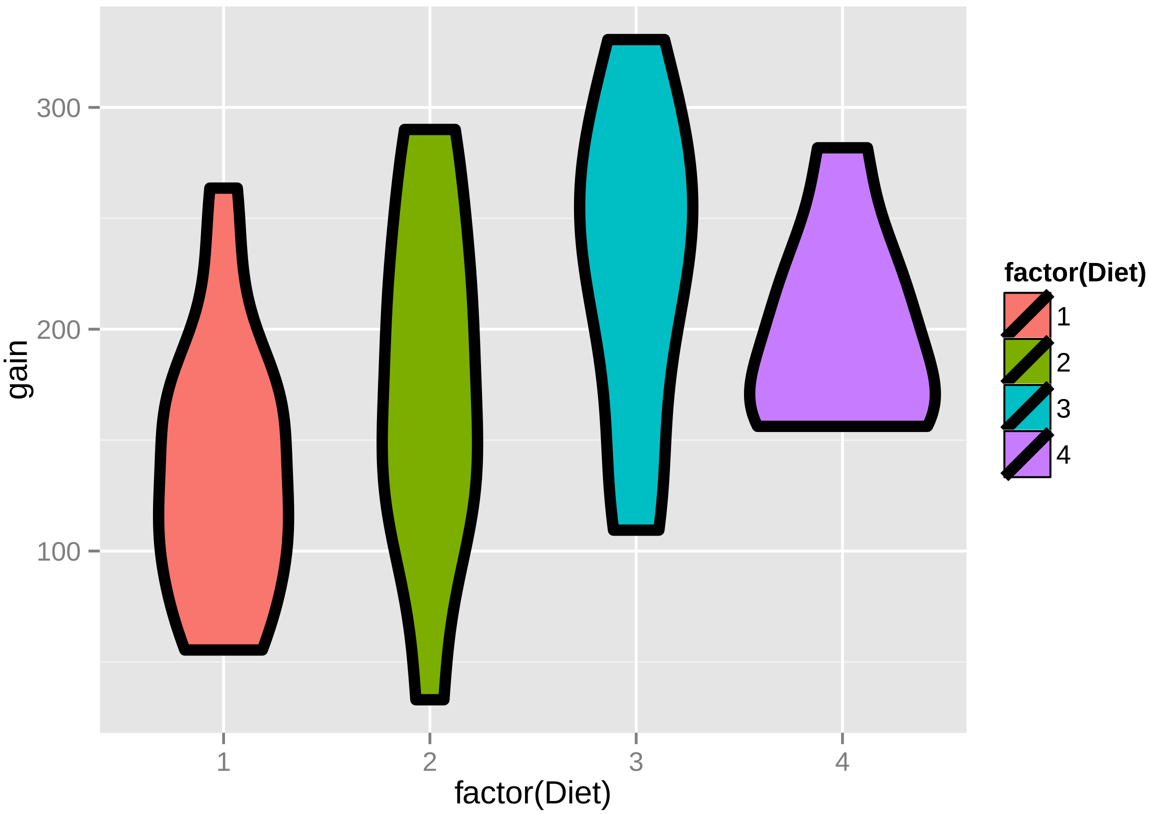
```r
g <- g + stat_summary(aes(group = 1), geom = "line", fun.y = mean, size = 1, col = "black")
g <- g + facet_grid(. ~ Diet)
g
```



The 4 the diet looks faster. Let's check that.

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

```
## Warning: Removed 5 rows containing non-finite values (stat_ydensity).
```



by diet-1.pdf

A t interval

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

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

Unequal Variances

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