Statistical Inference Week 3 Notes

Gary R Seamans

October 19, 2017

```r
library(datasets)
library(prettydoc)
library(ggplot2)
library(reshape2)
library(manipulate)
library(dplyr)
library(UsingR)
set.seed(500)
```

T Confidence Intervals

Confidence intervals from the standard normal distribution $Est \pm ZQ \times SE_{est}$ or the estimate plus/minus a quantile from the SDN times the standard error of the estimate.

Gosset's t distribution and t confidence intervals for small samples

$$Est \pm TQ \times SE_{est}$$

- Assumes that the underlying data are **iid** Gaussian with the result that $\frac{\bar{X}-\mu}{S/\sqrt{n}}$
- Follows Gosset's t distribution with $n-1$ degrees of freedom
- Interval is $\bar{X} \pm t_{n-1}S/\sqrt{n}$ where t_{n-1} is the relevant qunatile

Example (Manipulate code, must be run in RStudio):

```r
k <- 1000
xvals <- seq(-5, 5, length = k)
myplot <- function(df){
 d <- data.frame(y = c(dnorm(xvals), dt(xvals, df)),
 x = xvals,
 dist = factor(rep(c("Normal", "T"), c(k,k))))
 g <- ggplot(d, aes(x = x, y = y))
 g <- g + geom_line(size = 2, aes(colour = dist))
 g
}
manipulate(myplot(mu), mu = slider(1, 20, step = 1))
```

```r
pvals <- seq(.5, .99, by = .01)What is penetration testing?

A security-minded form of unit testing that applies early in the development process

A procedure for testing libraries or other program components for vulnerabilities

All of the above

Whole-system testing for security flaws and bugsWhat is penetration testing?

A security-minded form of unit testing that applies early in the development process
```

```
A procedure for testing libraries or other program components for vulnerabilities

All of the above

Whole-system testing for security flaws and bugs
myplot2 <- function(df){
 d <- data.frame(n= qnorm(pvals),t=qt(pvals, df),
 p = pvals)
 g <- ggplot(d, aes(x= n, y = t))
 g <- g + geom_abline(size = 2, col = "lightblue")
 g <- g + geom_line(size = 2, col = "black")
 g <- g + geom_vline(xintercept = qnorm(0.975))
 g <- g + geom_hline(yintercept = qt(0.975, df))
 g
}
manipulate(myplot2(df), df = slider(1, 20, step = 1))
```

- For skewed distributions, the spirit of the t interval assumptions are violated, don't use them for skewed distributions.

Example using the **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
```

```
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 = TRUE)
```

```
##
##  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
```

```
t.test(extra ~ I(relevel(group, 2)), paired = TRUE, 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
```

Independently grouped t confindence intervals

- Cannot use a paired t test because the groups are independent and may have different sample sizes
- $(1 - \alpha) \times 100\%$ confidence interval for $\mu_y - \mu_x$ is:

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

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

  S_p^2 is the pooled variance.

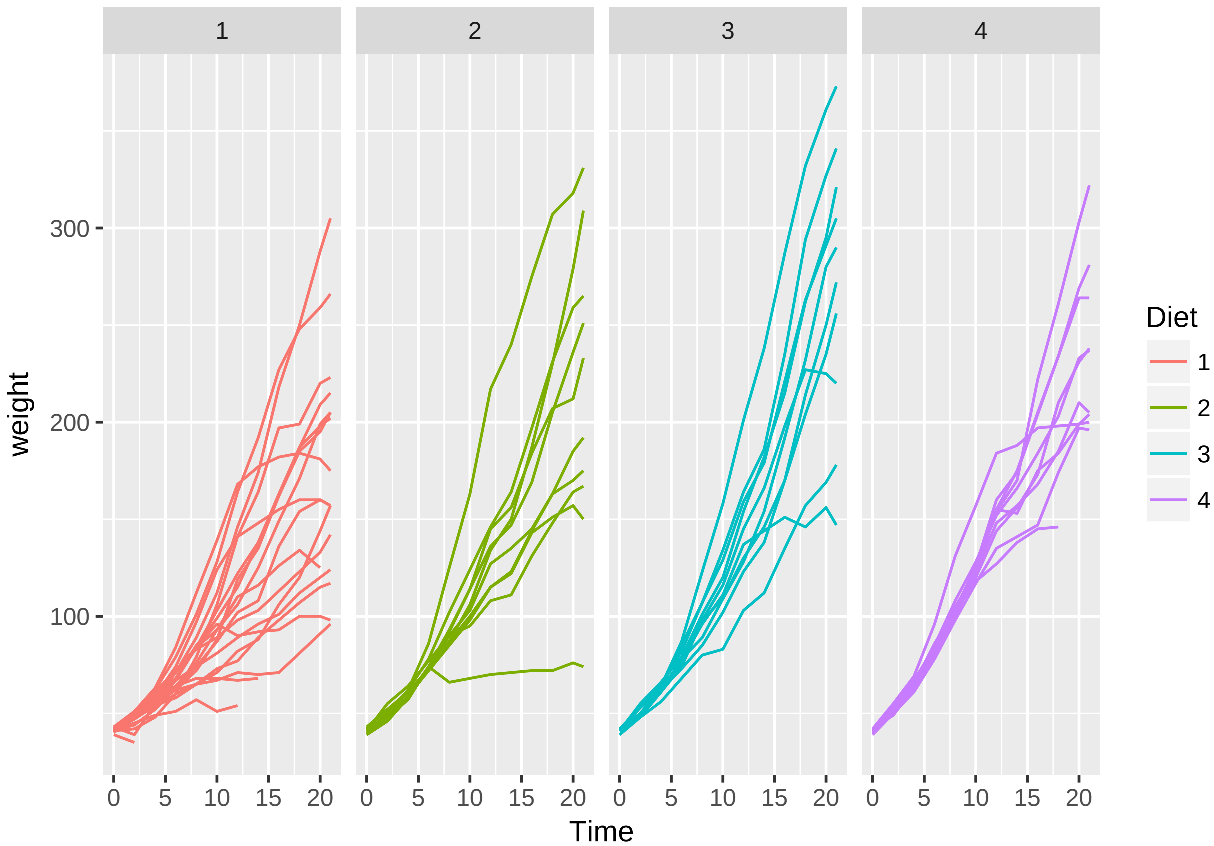
- This interval assumes the same variance between groups.

Example using the **ChickWeight** data:

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

A plot of the raw ChickData by diet:

```
ggplot(ChickWeight, aes(x=Time, y=weight, colour=Diet, group=Chick)) +
    geom_line() + facet_grid(. ~ Diet)
```



t interval for 1 and 4 (Diet)

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

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

Unequal variances

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

Where t_{df} is calculated with degrees of freedom:

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

- Approximately 95% interval
- When in doubt, assume unequal variances
- Use **t.test(..., var.equal = FALSE)** in R

Hypothesis testing

- H_0 is the null hypothesis, it is assumed to be true until statistical evidence proves otherwise
- α is used for the Type I error rate, typically 0.05
- Type I error, probability of rejecting the null hypothesis when it is true
- Type II error, probability of accepting the null hypothesis when it is false

T test

```
data(father.son)
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
```

Unequal variance T test, using ChickWeight, coparing diets 1 and 4

```
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
t.test(gain ~ Diet, paired = FALSE, var.equal = TRUE, data = wideCW14)
```

```
##
##  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
```

P-values

- A P-value is the probability of getting data as, or more extreme, than the observed data in favor of the alternative.
- Most common measure of statistical significance

What is the probability of getting a T statistic as large as 0.8?

```
pt(0.8, 15, lower.tail = FALSE)
```

```
## [1] 0.218099
```

- If the P-value is less than α

- Suppose a friend has 8 children, 7 of which are girls and none are twins

- If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?

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

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

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

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

- Reject above at a 5% level or 4% but not at 3%

Poisson example

- Suppose that a hospital has an infection rate of 10 infections per 100 person/days at risk (rate of 0.1) during the last monitoring period.
- Assume that an infection rate of 0.05 is an important benchmark.
- Given the model, could the observed rate being larger than 0.05 be attributed to chance?
- Under $H_0 : \lambda = 0.05$ so that $\lambda_0 100 = 5$
- Consider $H_0 : \lambda > 0.05$

```
ppois(9, 5, lower.tail = FALSE)
```

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

Quiz

Question 1

In a population of interest, a sample of 9 men yielded a sample average brain volume of 1,100cc and a standard deviation of 30cc. What is a 95% Student's T confidence interval for the mean brain volume in this new population?

- μ is 1100
- σ is 30
- n is 9
- $error$ = qt(0.975, df = n -1)* $\frac{\sigma}{sqrt(n)}$

```
mu <- 1100
sigma <- 30
n <- 9
result <- mu + c(-1,1)*qt(0.975, df = n-1)*sigma/sqrt(n)
result
```

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

Question 2

- μ is -2
- $\sigma = 0 - \mu \times \sqrt(n)/qt(.95 + (1 - .95)/2), df = n - 1)$
- n is 9

A diet pill is given to 9 subjects over six weeks. The average difference in weight (follow up - baseline) is -2 pounds. What would the standard deviation of the difference in weight have to be for the upper endpoint of the 95% T confidence interval to touch 0?

```r
mu <- -2
n <- 9
-mu*sqrt(n)/qt(.95+(1-.95)/2,n-1)
```

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

Question 3

In an effort to improve running performance, 5 runners were either given a protein supplement or placebo. Then, after a suitable washout period, they were given the opposite treatment. Their mile times were recorded under both the treatment and placebo, yielding 10 measurements with 2 per subject. The researchers intend to use a T test and interval to investigate the treatment. Should they use a paired or independent group T test and interval?

They should use a paired interval

Question 4

In a study of emergency room waiting times, investigators consider a new and the standard triage systems. To test the systems, administrators selected 20 nights and randomly assigned the new triage system to be used on 10 nights and the standard system on the remaining 10 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 3 hours with a variance of 0.60 while the average MWT for the old system was 5 hours with a variance of 0.68. Consider the 95% confidence interval estimate for the differences of the mean MWT associated with the new system. Assume a constant variance. What is the interval? Subtract in this order (New System - Old System).

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

- S_p^2 is the pooled variance
- n_x is the new triage system sample and S_x^2 is the new variance
- n_y is the old triage system sample and S_y^2 is the old variance

The confidence interval is given by:

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

Doing the substitutions:

```r
xhat <- 3
yhat <- 5
new_n <- 10
old_n <- 10
new_var <- 0.6
old_var <- 0.68
sp <- sqrt(((new_n -1) * new_var + (old_n -1) * old_var)/(new_n + old_n -2))
ci <- (xhat - yhat + c(-1,1) *qt(0.975, new_n + old_n -2) * sp * sqrt(1/new_n +1/old_n))
sp
```

```
## [1] 0.8
```

```r
ci
```

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

Question #5

Suppose that you create a 95% T confidence interval. You then create a 90% interval using the same data. What can be said about the 90% interval with respect to the 95% interval?

The interval will be narrower

Question 6

To further test the hospital triage system, administrators selected 200 nights and randomly assigned a new triage system to be used on 100 nights and a standard system on the remaining 100 nights. They calculated the nightly median waiting time (MWT) to see a physician. The average MWT for the new system was 4 hours with a standard deviation of 0.5 hours while the average MWT for the old system was 6 hours with a standard deviation of 2 hours. Consider the hypothesis of a decrease in the mean MWT associated with the new treatment.

What does the 95% independent group confidence interval with unequal variances suggest vis a vis this hypothesis? (Because there's so many observations per group, just use the Z quantile instead of the T.)

- μ_{old} 6
- σ_{old} 2
- n_{old} 100
- σ^2_{old} 4
- μ_{new} 4
- σ_{new} 0.5
- n_{new} 100
- σ^2_{new} 0.25

Use the Z quantile:

```
old <- qnorm(0.975, 6, 2)
new <- qnorm(0.975, 4, 0.5)
old
```

```
## [1] 9.919928
```

```
new
```

```
## [1] 4.979982
```

```
old - new
```

```
## [1] 4.939946
```

When subtracting (old -new) the interval is entirely above zero. The new system appears to be effective

Question 7

Suppose that 18 obese subjects were randomized, 9 each, to a new diet pill and a placebo. Subjects' body mass indices (BMIs) were measured at a baseline and again after having received the treatment or placebo for four weeks. The average difference from follow-up to the baseline (followup - baseline) was -3 kg/m2 for the treated group and 1 kg/m2 for the placebo group. The corresponding standard deviations of the differences was 1.5 kg/m2 for the treatment group and 1.8 kg/m2 for the placebo group. Does the change in BMI over the four week period appear to differ between the treated and placebo groups? Assuming normality of the underlying data and a common population variance, calculate the relevant *90%* t confidence interval. Subtract in the order of (Treated - Placebo) with the smaller (more negative) number first.

- n_n 9

- μ_n -3
- σ_n 1.5
- n_p 9
- μ_p 1
- σ_p 1.8

Substituing, see question 4:

```
n_n <- 9
mu_n <- -3
sigma_n <- 1.5
var_n <- sigma_n^2

n_p <- 9
mu_p <- 1
sigma_p <- 1.8
var_p <- sigma_p^2

sp <- sqrt(((n_p -1) * var_p + (n_n -1) * var_n)/(n_p + n_n -2))
ci <- (mu_n - mu_p) + c(1,-1) * qt(1 - (1 - .9)/2, n_p + n_n -2) * sp * sqrt(1/n_n + 1/n_p)
ci
```

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