Lab Solutions

401 GSI team

3/8/2018 and 3/9/2018

```r
# install.packages("MASS")
#Load library MASS
library(MASS)
#Load data crabs
data('crabs')

# add indicator variable to data for crab species
crabs$mu1 <- (crabs$sex == "M" )*1
crabs$mu2 <- (crabs$sex == "F")*1
```

Q1) Constructing CIs in R

```r
# Obtain estimate of population mean
cl_crabs <- lm(CL~mu1+mu2-1, data = crabs)
summary(cl_crabs)
```

```
##
## Call:
## lm(formula = CL ~ mu1 + mu2 - 1, data = crabs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.751  -5.178   0.240   4.974  14.840
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## mu1  32.8510     0.7097   46.28   <2e-16 ***
## mu2  31.3600     0.7097   44.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 7.097 on 198 degrees of freedom
```

Q2) Constructing a 99% confidence interval for the mean of Male crabs

$$\bar{y} \pm z_{\frac{\alpha}{2}} s.e(\bar{y})$$

$$32.8510 \pm 2.56(0.7097)$$

$$(31.1974, 34.5046)$$

Constructing a 99% confidence interval for the mean of Female crabs

$$\bar{y} \pm z_{\frac{\alpha}{2}} s.e(\bar{y})$$

$$31.3600 \pm 2.56(0.7097)$$

$$(29.7064, 33.0136)$$

Q3) Difference in Means

```r
crabs$mu3 <- 1
crabs$mu_diff <- crabs$mu1

bd_crabs2 <- lm(CL ~ mu3 + mu_diff - 1, data = crabs)
summary(bd_crabs2)
```

```
##
## Call:
## lm(formula = CL ~ mu3 + mu_diff - 1, data = crabs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.751  -5.178   0.240   4.974  14.840
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## mu3      31.3600     0.7097  44.185   <2e-16 ***
## mu_diff   1.4910     1.0037   1.485    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Q4) Constructing a 95% confidence interval for the difference in means

- (note: I am using the normal approximation)

$$\bar{d} \pm z_{\frac{\alpha}{2}} s.e(\bar{d})$$

$$1.4910 \pm 2.56(1.0037)$$

$$(-0.847621, 3.829621)$$

Confidence Intervals for Future Values

- Motivating Question: What's the point of performing a regression?

Confidence Intervals for Future Values

- A $100(1-\alpha)\%$ **Confidence Interval** for a mean future value (or the regression line at) $\tilde{y}$ given values $\tilde{x}$:
  - $\hat{y} \pm t_{(\frac{\alpha}{2}, n-2)} s \sqrt{\frac{1}{n} + \frac{(\tilde{x}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
- A $100(1-\alpha)\%$ **Prediction Interval** for a future value $\tilde{y}$ given values $\tilde{x}$:
  - $\hat{y} \pm t_{(\frac{\alpha}{2}, n-2)} s \sqrt{1 + \frac{1}{n} + \frac{(\tilde{x}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$
- It is important to note that the confidence interval is narrower than the prediction interval
  - Why is this? (Hint: What do we know about means from 250?)
- Details can be found in sections 2.3 and 2.4 of Sheather

Confidence Intervals for Future Values in R

Construct a 95% confidence interval and a 95% prediction interval for the crab's body depth given it is a blue crab with a carapace length of 45.

```
crab_bd_reg <- lm(BD ~ sp + CL, data = crabs)
summary(crab_bd_reg)
```

```
##
## Call:
## lm(formula = BD ~ sp + CL, data = crabs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31623 -0.22544  0.00332  0.27120  1.08043
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.996643   0.123044   -8.10 5.65e-14 ***
## spO          1.044956   0.055373   18.87  < 2e-16 ***
## CL           0.451781   0.003899  115.87  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 
```

Confidence Intervals for Future Values in R

```
x_star <- data.frame(sp = "B", CL = 45)

# confidence interval
predict(crab_bd_reg, x_star, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 19.33352 19.19689 19.47014
```

```
# prediction interval
predict(crab_bd_reg, x_star, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 19.33352 18.58163 20.08541
```