Stats 401 Lab 8

401 GSI team

3/8/2018 and 3/9/2018

Outline

- Quick Reminder: If you are thinking about withdrawing from the course, the deadline is **March 19th**!
- Motivation for using matrices
- Brief review of hypothesis tests and confidence intervals
- Constructing CIs in R
- Lab Ticket

Matrices Motivation

- Matrices are hidden everywhere:
  - In video game rendering (how do you create a reflection?)
  - Airplanes use them to operate
  - MRI's and CAT scans use them
  - Seismic Survey's in geology
  - Optimization problems in economics
  - ... and of course, statistics!
- Many complex calculations require matrices to carry out.

Hypothesis Tests

- Why do you think we care about hypothesis testing?

Hypothesis Tests: T-test

- Recall from STATS 250 the t-distribution
  - degrees of freedom: n-1
  - standard deviation of population unknown and estimated using the data
- We used this distribution to:
  - test a population mean,
  - test a population mean difference (paired data),
  - and test a difference in population means (unpaired data),
  - and construct CIs for all of these
- See STATS 250 lecture notes 7-9 for additional details

Confidence Intervals

- We will be focusing on CIs in this lab
- Why?
  - CIs essentially perform a two-sided hypothesis test and provide you with a estimate the true population value
- There are several natural uses for confidence intervals in regression:
  - estimating population coefficients (β)
  - comparing means of different populations
  - predicting future values (prediction interval)
  - predicting mean future values (confidence interval)
- (We will touch on these last two.)

Confidence Intervals (cont.)

- Recall that a $100(1-\alpha)\%$ confidence interval for a value is given by
  - $x \pm z_{\frac{\alpha}{2}} s.e(x)$ (population s.d. is known) or
  - $x \pm t_{(\frac{\alpha}{2}, n-2)} s.e(x)$ (population s.d. is unknown)
- Recall that a $100(1-\alpha)\%$ confidence interval for a population mean difference is given by
  - $\bar{d} \pm z_{\frac{\alpha}{2}} s.e(\bar{d})$ (population s.d. is known) or
  - $\bar{d} \pm t_{(\frac{\alpha}{2}, n-2)} s.e(\bar{d})$ (population s.d. is unknown)

Constructing CIs in R

- Why do we make you construct the matrix to calculate the value we are interested in?
  - Often the problems we do in class are much simpler than problems you'll encounter at a job.
  - We aim to not only give you the tools necessary to handle realistic problems that you could encounter, but also have you develop an understanding of why the built-in functions work. (Blindly wielding a hammer is not the same as hitting the nail.)

Constructing CIs in R

A Basic Exercise:

Suppose we're interested in determining the differences in the body depth of crabs from two different species (blue and orange).

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

# add indicator variable to data for crab species
crabs$mu1 <- (crabs$sp == "B")*1
crabs$mu2 <- (crabs$sp == "O")*1
```

Constructing CIs in R

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

```
##
## Call:
## lm(formula = BD ~ mu1 + mu2 - 1, data = crabs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0780 -2.1830  0.0695  2.3170  7.4170
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## mu1   12.583      0.311   40.46   <2e-16 ***
## mu2   15.478      0.311   49.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 3.11 on 198 degrees of freedom
```

Constructing a 95% confidence interval for the mean of Blue crabs

- note: I will be using a normal approximation
  - why can I do this?

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

$$12.583 \pm 1.64(0.311)$$

$$(12.072, 13.093)$$

Difference in Means

```r
crabs$mu3 <- 1
crabs$mu_diff <- crabs$mu2

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

```
##
## Call:
## lm(formula = BD ~ mu3 + mu_diff - 1, data = crabs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0780 -2.1830  0.0695  2.3170  7.4170
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## mu3      12.5830     0.3110  40.460  < 2e-16 ***
## mu_diff   2.8950     0.4398   6.582 4.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Constructing a 95% confidence interval for the difference in means

- (note: I am using the normal approximation)

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

$$3.021 \pm 1.64(1.470)$$

$$(0.6102, 5.4318)$$

- Are my data considered to be paired or unpaired?

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
```

Lab activity

Compare the carapace length of between male and female crabs.

1. Construct a design matrix to find the mean carapace length of male and female crabs.
2. Find a 99% CI (using the normal approximation) of the male and female mean carapace length.
3. Construct a design matrix to find the mean difference in carapace length between male and female crabs.
4. Find a 99% CI (using the normal approximation) of the mean difference in carapace length between male and female crabs.
5. (With time) Try constructing the 95% confidence interval and prediction interval for body depth by hand in R.

Lab Ticket

Write down a test of the null hypothesis that $\mu 1 = \mu 2$, obtaining a p-value and drawing a conclusion at a suitable significance level.