Stats 401 Lab 9

Naomi

11/2/2018

Outline

- Quick Reminder: If you are thinking about withdrawing from the course, the deadline is **November 9th**!
- Quiz 2: November 16th!
- Review of hypothesis tests, confidence interval for the expected value of a new outcome, and the prediction interval for a new outcome
- Practice Problems

Review: Hypothesis Tests

"Steps of a Hypothesis Tests"

1. Ask a question
2. Write question in terms of null hypothesis H_0 and alternative hypothesis H_a
3. Collect Data
4. Determine the test statistic
5. Calculate the p-value
6. Write appropriate conclusion

Hypothesis Test Example 1

We have been recruited by a California university to explore the relationship between water salinity, water oxygen, and water temperature. We have been given 60 years of oceanographic data collected from the California Current by the California Cooperative Oceanic Fisheries Investigations. Below is a snapshot of the data. (Source: https://www.kaggle.com/sohier/calcofi)

- Depthm: Depth in meters
- T_degC: Water temperture in degrees Celsius
- Salnty: Water Salinity in g of salt per kg of water
- 02ml_L: O_2 mixing ratio in ml/L

```
##   X Depthm T_degC  Salnty O2ml_L Year
## 1 1      0 13.342 32.7369  6.189 2016
## 2 2      2 13.342 32.7369  6.189 2016
## 3 3      6 13.358 32.8685  6.179 2016
## 4 4     10 13.362 33.0125  6.057 2016
## 5 5     11 13.313 33.0972  6.048 2016
## 6 6     20 13.390 33.2259  5.992 2016
```

Hypothesis Test Example 1: Cont.

1. We are interested in knowing there is a positive relationship between water temperature and water salinity, while controlling for oxygen levels and depth.
2. $H_0 : \beta_1 = 0 \quad H_a : \beta_1 > 0$
3. Our test statistic is b_1
4. We want to calculate $P(\hat{\beta}_1 > b_1)$

Hypothesis Test Example 1: Cont.

To obtain our test statistic b_1, we fit a linear model to our data.

```r
lm1 <- lm(T_degC ~ Depthm + Salnty + O2ml_L,
          data = bottle_2016)
summary(lm1)$coefficients[, c("Estimate", "Std. Error")]
```

```
##                  Estimate   Std. Error
## (Intercept) -78.591784806 3.6966365630
## Depthm       -0.003844847 0.0001576979
## Salnty        2.481605712 0.1077711182
## O2ml_L        1.955793588 0.0240747199
```

Hypothesis Test Example 1: Cont.

Recall: under the null hypothesis $\hat{\beta}_1 \sim N(0, SD(\hat{\beta}_1))$. We estimate the $SD(\hat{\beta}_1)$ using $SE(b_1) = s\sqrt{[(\mathbb{X}^T\mathbb{X})^{-1}]_{11}}$. Therefore, $\hat{\beta}_1 \sim N(0, 0.1077)$ (approx).

We can calculate $P(\hat{\beta}_1 > b_1)$ using $pnorm(b_1, SE(b_1))$.

```r
pnorm(summary(lm1)$coefficients["Salnty", "Estimate"],
      mean = 0,
      sd = summary(lm1)$coefficients["Salnty", "Std. Error"])
```

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

What is our conclusion?

Confidence Intervals and Prediction Intervals for new outcome

- A confidence interval is contructed for the expected value of a new outcome
  - [**x*b** - 1.96SE(**x*b**), **x*b** $+$ 1.96SE(**x*b**)]
  - where SE(**x*b**) $= s\sqrt{\mathbf{x*}(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x*}^T}$
- A prediction interval is contructed for the observed value of a new outcome
  - $[\mathbf{x*b} - 1.96SE_{pred}(\mathbf{x*b}), \mathbf{x*b} + 1.96SE_{pred}(\mathbf{x*b})]$
  - where SE(**x*b**) $= s\sqrt{1 + \mathbf{x*}(\mathbb{X}^T\mathbb{X})^{-1}\mathbf{x*}^T}$

Confidence Intervals Example

- Construct a 95% confidence interval for the expected value of a new observation obtained at a depth of 100 meters, a salinity of 33.5, and an oxygen level of 4.

```r
x <- c(1, 100, 33.5, 4)
pred <- x %*%coef(lm1)
V <- summary(lm1)$cov.unscaled
s <- summary(lm1)$sigma
SE <- s*sqrt(x%*%V%*%x)
c <- qnorm(0.975)
cat("CI = [", round(pred-c*SE,3),
", ", round(pred+c*SE,3), "]", sep = "")
```

```
## CI = [11.94, 12.021]
```

Prediction Intervals Example

- Construct a 95% prediction interval for this new observation.

```r
SE_pred <-s*sqrt(1 + x%*%V%*%x)
cat("CI = [", round(pred-c*SE_pred,3),
", ", round(pred+c*SE_pred,3), "]", sep = "")
```

```
## CI = [8.852, 15.109]
```

Hypothesis Test Example 2

We can think of a die as a random variable and the realization (a random draw from the distribution) is one roll of the die.

1. We're planning on playing a game of Yahtzee with a friend. Before we begin, we want to test if one of the dice (selected at random) is a fair die. To do this, we roll the die 100 times and calculate the proportion each value appears.
2. $H_0 : p_1 = p_2 = \cdots = p_6 = \frac{1}{6}$ $\qquad H_a$: at least one of the probabilities is not equal to $\frac{1}{6}$

Note: What type of test are we performing?

Hypothesis Test Example 2: Cont.

Simulate Rolling a Die

```r
# simulate rolling a die 100 times
die_roll <- function(x){
  r <- runif(x, min = 0, max = 6)
  value <- ceiling(r)
  return(value)
}

die_rolls <- replicate(100, die_roll(1))

# alternative (simpler) solution
sides <- c(1:6)
die_rolls2 <- sample(sides, 100, replace = TRUE)
```

Hypothesis Test Example 2: Cont.

Calculate our test statistic

```r
# get the count of rolls for each value
die_prop <- aggregate(data.frame(count = die_rolls),
                      list(value = die_rolls), FUN = length)

# calculate our test statistic
die_prop$obs_exp <- die_prop$count - 100*(1/6)
chi_sqr <- sum(die_prop$obs_exp^2/(100*(1/6)))

# calculate our p-value
# chi-squared has 5 degrees of freedom
pchisq(chi_sqr, 5)
```

```
## [1] 0.6264584
```

What is our conclusion?

In Lab Activity Part 1

- Construct a 95% confidence interval for the association between water temperature and oxygen level. Hint: Go back to last lab.

In Lab Activity Part 2

Download the crime dataset from the 401 webpage.

Construct a hypothesis test to determine if there is a relationship between the total amount of crime and the percent of 25 year olds with a high school degree.

In Lab Activity Part 2 cont.

Construct a 95% confidence interval **and** a 95% prediction interval for the expected value of a town with an annual police funding of 38, 50% of people over 25 with a high school diploma, 19% of teenagers not in high school, 15% of college-aged children in college, and 12% of adults with a college degree.

In Lab Activity Part 3

Conduct a hypothesis test to check if a die is specifically weighted such that 6 is twice as likely as all the other values. Hint: You may need to change the limits of the uniform distribution used in lab.

Exit Ticket

- Conduct a hypothesis test to check if a coin is weighted towards heads.
- Construct a 99% confidence interval for the relationship between the total amount of crime and the percent of 25 year olds with a high school degree using the t-distribution.
    - Comment on why the t-distribution may be more appropriate in this case.