Hypothesis Tests and Confidence Intervals for Means with the t-distribution

2/11/2024

Delta has a reputation for being a very reliable airline. After a bad experience with United this past summer I started flying Delta and was impressed to find that each of the domestic Delta flights I took arrived at my destination not just on time but early! After talking to some other people I found that they had had similar experiences flying Delta. I'm curious to see if this is typical for Delta; *on average do domestic Delta flights arrive early at their destinations?* We will narrow down our investigation to flights leaving from the JFK since that is where I almost always am flying out of, and JFK is a Delta hub.

###The data We will be using a data set containing "on-time data for a random sample of domestic flights that departed NYC (JFK, LGA or EWR) in 2013." https://www.openintro.org/data/index.php?data=nycflights

dep_delay,arr_delay: Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.

###Read in data:

```
nyc.flights <- read.csv("nycflights.csv")
head(nyc.flights)
```

```
##   year month day dep_time dep_delay arr_time arr_delay carrier tailnum flight
## 1 2013     6  30      940        15     1216        -4      VX  N626VA    407
## 2 2013     5   7     1657        -3     2104        10      DL  N3760C    329
## 3 2013    12   8      859        -1     1238        11      DL  N712TW    422
## 4 2013     5  14     1841        -4     2122       -34      DL  N914DL   2391
## 5 2013     7  21     1102        -3     1230        -8      9E  N823AY   3652
## 6 2013     1   1     1817        -3     2008         3      AA  N3AXAA    353
##   origin dest air_time distance hour minute
## 1    JFK  LAX      313     2475    9     40
## 2    JFK  SJU      216     1598   16     57
## 3    JFK  LAX      376     2475    8     59
## 4    JFK  TPA      135     1005   18     41
## 5    LGA  ORF       50      296   11      2
## 6    LGA  ORD      138      733   18     17
```

###Subset data: We want to look at just *Delta flights that departed from JFK*

First we will subset for all Delta flights, and then we will subset the Delta flights for just those flights departing from JFK:

```
delta.flights <- nyc.flights[nyc.flights$carrier == "DL",]

delta.jfk <- delta.flights[delta.flights$origin == "JFK",]
```

###State hypotheses and alpha mu = mean arrival delay * H0: mu = 0 (on time) * Ha: mu < 0 (early) * alpha = .05

###Explore data and collect sample statistics

```r
#define your sample statistics; use R functions to define these when possible
#consider printing out the values to make sure everything is working
mu <- 0
xbar <- mean(delta.jfk$arr_delay)
sx <- sd(delta.jfk$arr_delay)
n <- nrow(delta.jfk)
SE.xbar <- sx/sqrt(n)
```

###Check assumptions and conditions

n > 30, n <10% of pop? Yes

```r
n
```

```
## [1] 2070
```

###Calculate t statistic then perform the test Perform the hypothesis test using pt, and state the p-value OUTSIDE OF CODE CHUNK!

```r
t <- (xbar - mu)/SE.xbar
t
```

```
## [1] -3.072685
```

```r
pt(t, n-1)
```

```
## [1] 0.001074537
```

###Make a conclusion using alpha Reject H0, accept HA. Delta flights out of JFK arrive early on average.

###Confidence Interval For proportions we found a z* value. For 95%, that was around 2, and we calculated using qnorm(.975)

For means, we find a t* value: qnorm(.975, n-1)

Then, just use the formula for CIs for means that you should know already!

```r
xbar - qt(.975,n-1)*SE.xbar
```

```
## [1] -4.646429
```

```r
xbar + qt(.975,n-1)*SE.xbar
```

```
## [1] -1.026035
```

95% CI: (-4.64, -1.03)

We are 95% confident that Delta flights leaving JFK on average arrive at destinations 1 to 4.6 minutes EARLY (negative = early!)

Your Turn:

Flying in the winter is always a bit of a gamble, especially when you are flying from the Northeast. You assume that even Delta will have departure delay issues, but your friend is a Delta fanatic and refuses to believe that Delta flights could be late, even by 5 minutes! Perform a hypothesis test to see if the average *departure* delay for Delta flights out of JFK in December is more than 5 minutes.

First run this code to subset the nyc flight data for just December flights:

```r
delta.winter <- delta.jfk[delta.jfk$month == 12,]
```

###State hypotheses and alpha

mu = mean arrival delay * H0: mu = 5 (5 minutes late) * Ha: mu > 5 (more than 5 minutes late) * alpha = .05

###Explore data and collect sample statistics

```r
#define your sample statistics; use R functions to define these when possible
#consider printing out the values to make sure everything is working

mu <- 5
xbar <- mean(delta.winter$dep_delay)
sx <- sd(delta.winter$dep_delay)
n <- nrow(delta.winter)
SE.xbar <- sx/sqrt(n)
```

###Check assumptions and conditions

n > 30: Yes, 160 > 30 n < 10% of population: Yes, 1600 < total Delta flights departed from JFK in December

###Calculate t statistic then perform the test Perform the hypothesis test, and state the p-value OUTSIDE OF CODE CHUNK!

```r
t <- (xbar - mu)/SE.xbar
1 - pt(t, n-1)
```

```
## [1] 0.2552535
```

p-value: 0.25525

###Make a conclusion using alpha

p-value of 0.25525 is greater than alpha of 0.05, so we failed to reject H0. We failed to find sufficient evidence that the mean departure delay of Delta flights from JFK in December is greater than 5 minutes.

###Confidence Interval

```r
xbar - qt(.975,n-1)*SE.xbar
```

```
## [1] 2.693822
```

```r
xbar + qt(.975,n-1)*SE.xbar
```

```
## [1] 9.618678
```

95% CI: (2.69, 9.62)

We are 95% confident that Delta flights leaving JFK in December on average depart 2.69 to 9.62 minutes late.

Practice

1. Doctors in a North Carolina hospital are trying to figure out if the mean length of pregnancies in their hospital is low compared to the national average of 40 weeks. They take a sample `births.csv` https://www.openintro.org/data/index.php?data=births and a perform hypothesis test to see if there is cause for concern or further research into this issue. The duration of pregnancies is stored as the variable `weeks` in this data set.

```r
births <- read.csv("births.csv")
head(births)
```

```
##   f_age m_age weeks premature visits gained weight sex_baby     smoke
## 1    31    30    39 full term     13      1   6.88     male    smoker
## 2    34    36    39 full term      5     35   7.69     male nonsmoker
## 3    36    35    40 full term     12     29   8.88     male nonsmoker
## 4    41    40    40 full term     13     30   9.00   female nonsmoker
## 5    42    37    40 full term     NA     10   7.94     male nonsmoker
## 6    37    28    40 full term     12     35   8.25     male    smoker
```

###State hypotheses and alpha

mu = mean length of pregnancies * H0: mu = 40 (40 weeks) * Ha: mu < 40 (less than 40 weeks) * alpha = .05

###Explore data and collect sample statistics

```
#define your sample statistics; use R functions to define these when possible
#consider printing out the values to make sure everything is working

mu <- 40
xbar <- mean(births$weeks)
sx <- sd(births$weeks)
n <- nrow(births)
SE.xbar <- sx/sqrt(n)
```

###Check assumptions and conditions

n > 30: Yes, 150 > 30 n < 10% of population: Yes, 1500 < total births in this hospital

###Calculate t statistic then perform the test Perform the hypothesis test using, and state the p-value OUTSIDE OF CODE CHUNK!

```
t <- (xbar - mu)/SE.xbar
pt(t, n-1)
```

```
## [1] 6.396096e-10
```

p-value: 6.396096e-10

###Make a conclusion using alpha

p-value of 6.396096e-10 is less than alpha of 0.05, so we reject the null hypothesis and accept the alternative. The mean length of pregnancies delivered in this hospital is lower than the national average of 40 weeks.

###Confidence Interval

```
xbar - qt(.975,n-1)*SE.xbar
```

```
## [1] 38.1033
```

```
xbar + qt(.975,n-1)*SE.xbar
```

```
## [1] 38.99003
```

95% CI: (38.10, 38.99)

We are 95% confident that the mean length of pregnancies delivered in this hospital is between 38.10 and 38.99 weeks. This supports the results of our hypothesis test as the national average of 40 weeks is not in and is above the CI.

2. About 12.5% of babies are born preterm in the United States and thus classified as premies. The same North Carolina hospital is trying to figure out if last year in their hospital there was a greater percent of premies (babies born early) compared what is typical for the United States. Use the same data set as problem 1 (births.csv) to perform a hypothesis test and construct a confidence interval to determine if this hospital had a greater proportion of premie births than the national parameter. You will be working with the variable called `premature`, and it is up to you to figure out what the different responses in the data set are so you can tabulate the sample proportion.

NOTE THAT THIS IS A PROPORTION PROBLEM!!!

###State hypotheses and alpha

p = true proportion of babies born preterm in this hospital last year * H0: p = 12.5% (12.5% of babies were born preterm in this hospital last year) * Ha: p > 12.5% (more than 12.5%) * alpha = .05

###Explore data and collect sample statistics

```
#define your sample statistics; use R functions to define these when possible
#consider printing out the values to make sure everything is working
p <- 0.125
x <- sum(births$premature == "premie")
n <- nrow(births)
q <- 1 - p
SD.p.hat <- sqrt(p*q/n)
p.hat <- x/n
q.hat <- 1 - p.hat
```

###Check assumptions and conditions

```
n*p
```

```
## [1] 18.75
```

```
n*q
```

```
## [1] 131.25
```

```
10*n
```

```
## [1] 1500
```

np > 10: Yes, 18.75 > 10 nq > 10: Yes, 131.25 > 10 1500 is less than all babies born in this hospital last year

###Perform the test Perform the hypothesis test using pnorm, and state the p-value OUTSIDE OF CODE CHUNK!

```
1 - pnorm(p.hat, p, SD.p.hat)
```

```
## [1] 0.2892791
```

p-value: 0.28928

###Make a conclusion using alpha

p-value of 0.28928 is greater than alpha of 0.05, so we failed to reject the null hypothesis. Therefore, we do not have sufficient evidence to conclude that in this hospital last year, the proportion of babies born prematurely was greater than the national average.

###Confidence Interval

```
SE.p.hat <- sqrt(p.hat*q.hat/n)

z.star <- qnorm(0.975)

n*p.hat
```

```
## [1] 21
```

```
n*q.hat
```

```
## [1] 129
```

21 > 10 129 > 10

```
p.hat - z.star*SE.p.hat
```

```
## [1] 0.08447153
```

```r
p.hat + z.star*SE.p.hat
```

```
## [1] 0.1955285
```

95% CI: (0.08447, 0.19553)

We are 95% confident that the proportion of babies born prematurely in this hospital last year is between 8.447 and 19.553%. This supports our hypothesis test because the national average of 12.5% is within our CI.