t-tests

CSU, Chico Math 314

2018-10-24

outline

outline

Recap: Inference

Statistical inference revolves around a common theme: assume Y is a random variable such that $E(Y) = \mu$ and $Var(Y) = \sigma^2$. We estimate and/or make statements about μ with

$$\bar{Y} \pm t^*_{df} \cdot s_{\bar{Y}}, \quad \textit{and} \qquad t_{df} = \frac{\bar{Y} - \mu_0}{s_{\bar{Y}}}.$$

outline

One Sample t-test

We have already been doing one sample t-tests and confidence intervals.

```
## pseudocode
## 95% CI for mu
xbar + qt(c(0.025, 0.975), df)*s/sqrt(n)

## test statistic for hypothesis test
t <- (xbar - mu0)/(s/sqrt(n))

## p-value for two sided test, H_1: mu != m
2*(1-pt(abs(t), df))
```

Paired Data, definition

Paired data are somehow intimately connected.

Paired Data

Two sets of observations are paired if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Paired Data, by example

Tell me whether or not these data are paired.

- two websites' prices for the same book
- eye sight ratings by person
- lines of code by program
- upper versus lower bird beak lengths
- weights of male and female babies

Paired Data, t-test

If the data are paired, their difference has direct and interpretable meaning both in english and in statistics; $X_{i,diff} = X_{i,a} - X_{i,b}$ has meaning. Therefore

$$\bar{X}_{diff} \quad \text{and} \quad s_{\bar{X}_{diff}}$$

are simply fancy ways to write new random variables.

Paired Data, one sample t-test example

Are textbooks actually cheaper online? Compare the price of textbooks at the University of California, Los Angeles' (UCLA's) bookstore and prices at Amazon.com. Seventy-three UCLA courses were randomly sampled in Spring 2010.

```
books <- read.csv("https://roualdes.us/data/books.csv")
## look at data in RStudio
## what plot should we make
```

Paired Data, one sample t-test example

Plot the data!

```
suppressMessages(library(ggplot2))
qplot(1, uclaNew - amazNew, data=books, geom="boxplot")
## or
## qplot(uclaNew, amazNew, data=books) +
##     geom_abline(intercept=0, slope=1)
```

Paired Data, one sample t-test example

Calculate and interpret a 95% confidence interval of the difference in Amazon.com versus UCLA's book prices.

Paired Data, one sample t-test example

```
suppressMessages(library(dplyr))
bs <- mutate(books,
             diff=uclaNew-amazNew)$diff # unpack this
anyNA(bs) # any NAs?

## [1] FALSE

n <- length(bs)
mean(bs) + qt(c(0.025, 0.975), n-1)*sd(bs)/sqrt(n)

## [1]  9.435636 16.087652
```

Paired Data, one sample t-test example

We are 95% confident that UCLA's new book prices are greater than Amazon.com's books by between \$9.44 and \$16.09.

Paired Data, one sample t-test example

Set up, evaluate, and conclude in context a hypothesis test at $\alpha = 0.05$.

Paired Data, one sample t-test example

The natural hypotheses are

$$H_0 : \mu_{diff} = 0 \text{ versus } H_1 : \mu_{diff} \neq 0.$$

```
t <- (mean(bs) - 0) / (sd(bs)/sqrt(n)) # test statistic
2*(1-pt(abs(t), n-1)) # p-value

## [1] 6.92757e-11
```

Paired Data, one sample t-test example

Becuase p-value$< 0.0001 < \alpha = 0.05$, we reject H_0. There is insufficient evidence to claim that Amazon.com and UCLA's book prices are the same.

outline

Two Sample t-test

Two sample t-tests estimate the difference between two population means from two independent samples of data. We estimate $\mu_a - \mu_b$ with the point estimator $\bar{X}_a - \bar{X}_b$.

Two Sample t-test, standard error

We estimate the variance of the point estimator $\bar{X}_a - \bar{X}_b$, namely $Var(\bar{X}_a - \bar{X}_b) = \sigma^2_{\bar{X}_a - \bar{X}_b}$, with

$$s^2_{\bar{X}_a - \bar{X}_b} = \frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}.$$

Two Sample t-test, test statistics

$$t_{df} = \frac{(\bar{X}_a - \bar{X}_b) - (\mu_a - \mu_b)}{s^2_{\bar{X}_a - \bar{X}_b}}$$

Two Sample t-test, confidence interval

$$(\bar{X}_a - \bar{X}_b) \pm t^*_{df} s^2_{\bar{X}_a - \bar{X}_b}$$

Two Sample t-test, degrees of freedom

It's common to use a consertive approximation for the degrees of freedom for a t-test of the difference of two independent means,

$$df = \min\left(n_a - 1, n_b - 1\right).$$

Two Sample t-test, example

Consider the data set `ape::carnivora`. Calculate a 98% confidence interval for the difference in mean longevity between the two SuperFamilies Caniformia and Feliformia.

```
suppressMessages(library(ape))
data(carnivora)
```

Tangent on NAs

Some helpful functions to help you avoid NAs in R are

```
anyNA()
is.na()
na.omit()
```

Two Sample t-test, example

A 98% CI, difference in longevity by Caniformia and Feliformia.

```
(d <- carnivora %>%
    select(SuperFamily, LY) %>% # only two columns
    na.omit() %>% # be careful with placement
    group_by(SuperFamily) %>%
    summarise(n=n(), v=var(LY), xbar=mean(LY)))

## # A tibble: 2 x 4
##   SuperFamily     n     v  xbar
##   <fct>       <int> <dbl> <dbl>
## 1 Caniformia     24 7139.  192.
## 2 Feliformia     25 2576.  172.
```

Two Sample t-test, example

A 98% CI, difference in longevity by Caniformia and Feliformia.

```
xbarC <- 192.4583; xbarF <- 171.96
nC <- 24; nF <- 25
vC <- 7139.216; vF <- 2576.04
tstar <- qt(c(0.01, 0.99), min(c(nC, nF)-1))
(xbarC - xbarF) + tstar*sqrt(vC/nC + vF/nF)

## [1] -29.53083  70.52743
```

Two Sample t-test, example

We are 98% confident that the population difference in mean longevity between the SuperFamlies Caniformia and Feliformia is between -29.53 and 70.53 months.

Two Sample t-test, example

Set up, evaluate, and conclude in context a hypothesis test at $\alpha = 0.02$.

Two Sample t-test, example

The natural hypotheses are

$$H_0 : \mu_C = \mu_F \text{ versus } H_1 : \mu_C \neq \mu_F.$$

```
t <- ((xbarC - xbarF) - 0) / sqrt(vC/nC + vF/nF) # test st
2*(1-pt(abs(t), min(c(nC, nF)-1))) # p-value

## [1] 0.3163636
```

Two Sample t-test, example

Because p-value $= 0.32 > \alpha = 0.02$, we fail to reject H_0. There is insufficient evidence to claim that the true difference in mean longevity between Caniformia and Feliformia is different.

Tangent on R code

The above calculations were done for clarity, not cleanliness of code. Here's what I would have done.

```
# Confidence Interval
with(d, { # d is the summaraised data frame
    tstar <- qt(c(0.01, 0.99), min(n-1))
     -diff(xbar) + tstar*sqrt(sum(v/n))})

## [1] -29.53080  70.52746

# p-value
with(d, {t <- -diff(xbar)/sqrt(sum(v/n))
    2*(1-pt(abs(t), min(n-1)))})

## [1] 0.3163629
```

outline

Two Sample t-test, pooled standard deviation

Sometimes it is reasonable to assume that the two populations of interest share a common variance. In this case, we can estimate $\sigma^2_{\bar{X}_a - \bar{X}_b}$ with (something close to the average of the variances)

$$s^2_{pooled} = \frac{s_a^2(n_a - 1) + s_b^2(n_b - 1)}{n_a + n_b - 2},$$

with $df = n_a + n_b - 2$.

outline

Take Away

We're just playing a game: find parameters of interest, insert point estimates, and calculate the standard error of the estimators.

- Paired data
  - Two variables are intimately connected $\Rightarrow$ their difference has meaning
  - create one variable from the two $\Rightarrow$ one sample t-test
- Two sample data
  - Two variables are independent
  - Point estimate is difference of means
  - Standard error follows from independence
- CLT via standardization is the real workhorse.

outline

references I

David M Diez, Christopher D Barr, and Mine Cetinkaya-Rundel. *OpenIntro Statistics*. CreateSpace independent publishing platform, third edition, 2015.

Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer Science and Business Media, 2009.