CAAP Statistics - Lec15

R Session6

Jul 29, 2022

Review

- One sample mean t-test
  - When sample size is small, t-distribution is better choice than normal!
- Paired t-test
- Difference in two means

Learning Objectives

- Learn how to use `` `t.test` `` command and interpret the result appropriately
- Understand the t-distribution: `pt(), qt()`
  - Recall `pnorm(), qnorm()`

Load packages

```r
library(openintro)
library(tidyverse)
library(ggplot2)
```

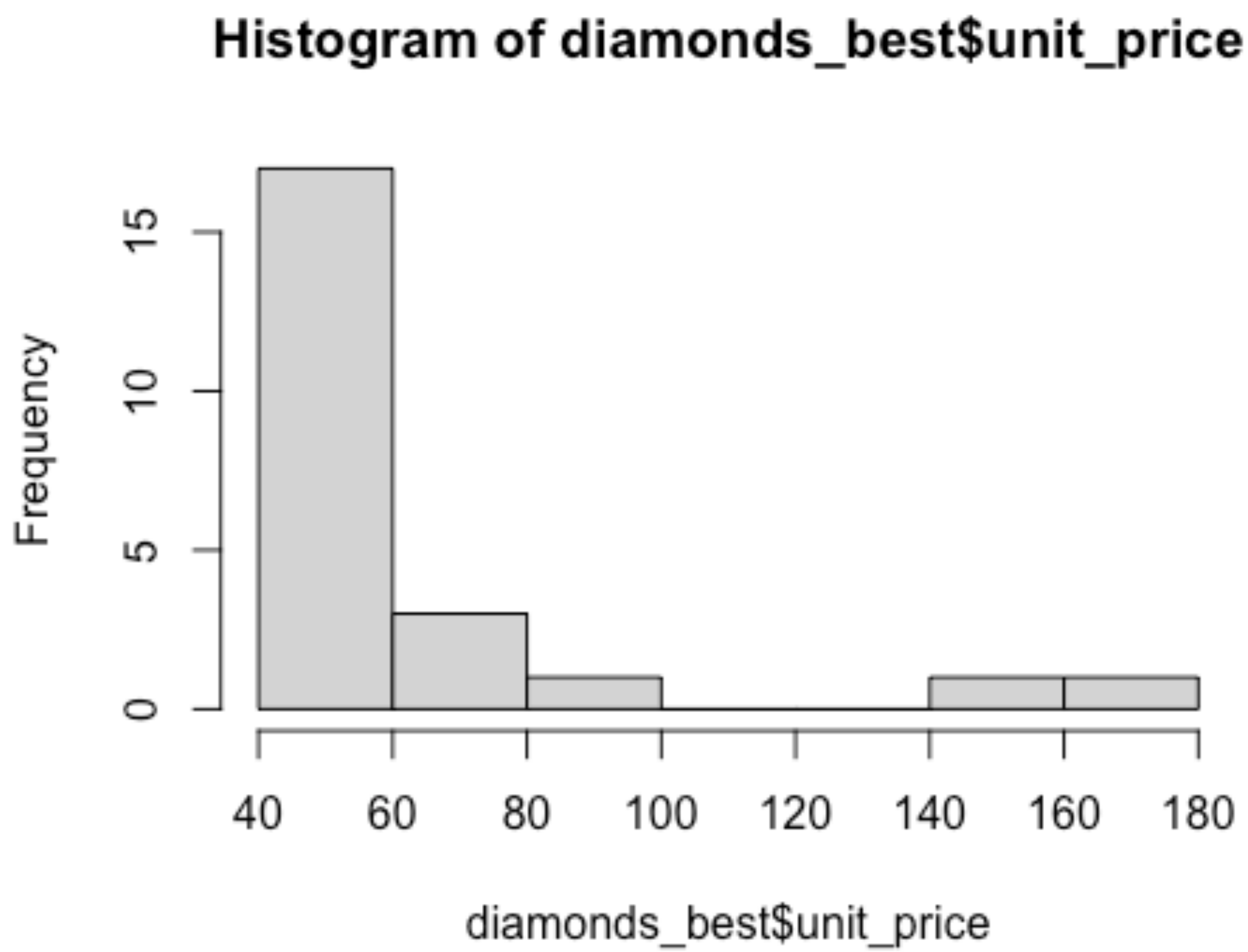
Diamonds dataset

```
head(diamonds)
## # A tibble: 6 × 10
##   carat cut       color clarity depth table price     x     y     z
##   <dbl> <ord>     <ord> <ord>   <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.23 Ideal     E     SI2      61.5    55   326  3.95  3.98  2.43
## 2  0.21 Premium   E     SI1      59.8    61   326  3.89  3.84  2.31
## 3  0.23 Good      E     VS1      56.9    65   327  4.05  4.07  2.31
## 4  0.29 Premium   I     VS2      62.4    58   334  4.2   4.23  2.63
## 5  0.31 Good      J     SI2      63.3    58   335  4.34  4.35  2.75
## 6  0.24 Very Good J     VVS2     62.8    57   336  3.94  3.96  2.48
```

One-sample t-test

Q: Is the unit price of the "best" diamonds higher than $10,000?

```
diamonds_best = diamonds %>%
  filter(cut=="Ideal", carat == 1.00, color =="D")%>%
  mutate(unit_price = price/100)
hist(diamonds_best$unit_price)
```



One-sample t-test: p-value, critical value

```
mu = mean(diamonds_best$unit_price)
sd = sd(diamonds_best$unit_price)
(t.stat = (mu-100)/(sd/sqrt(23)))
## [1] -5.539674
## p-value
2*pt(-abs(t.stat), 23-1) # H_a: mu != 100
## [1] 1.441455e-05
pt(t.stat, 23-1) # H_a: mu < 100
## [1] 7.207275e-06
1-pt(t.stat, 23-1) # H_a: mu > 100
## [1] 0.9999928
## critical value
qt(0.95, 23-1)
## [1] 1.717144
```

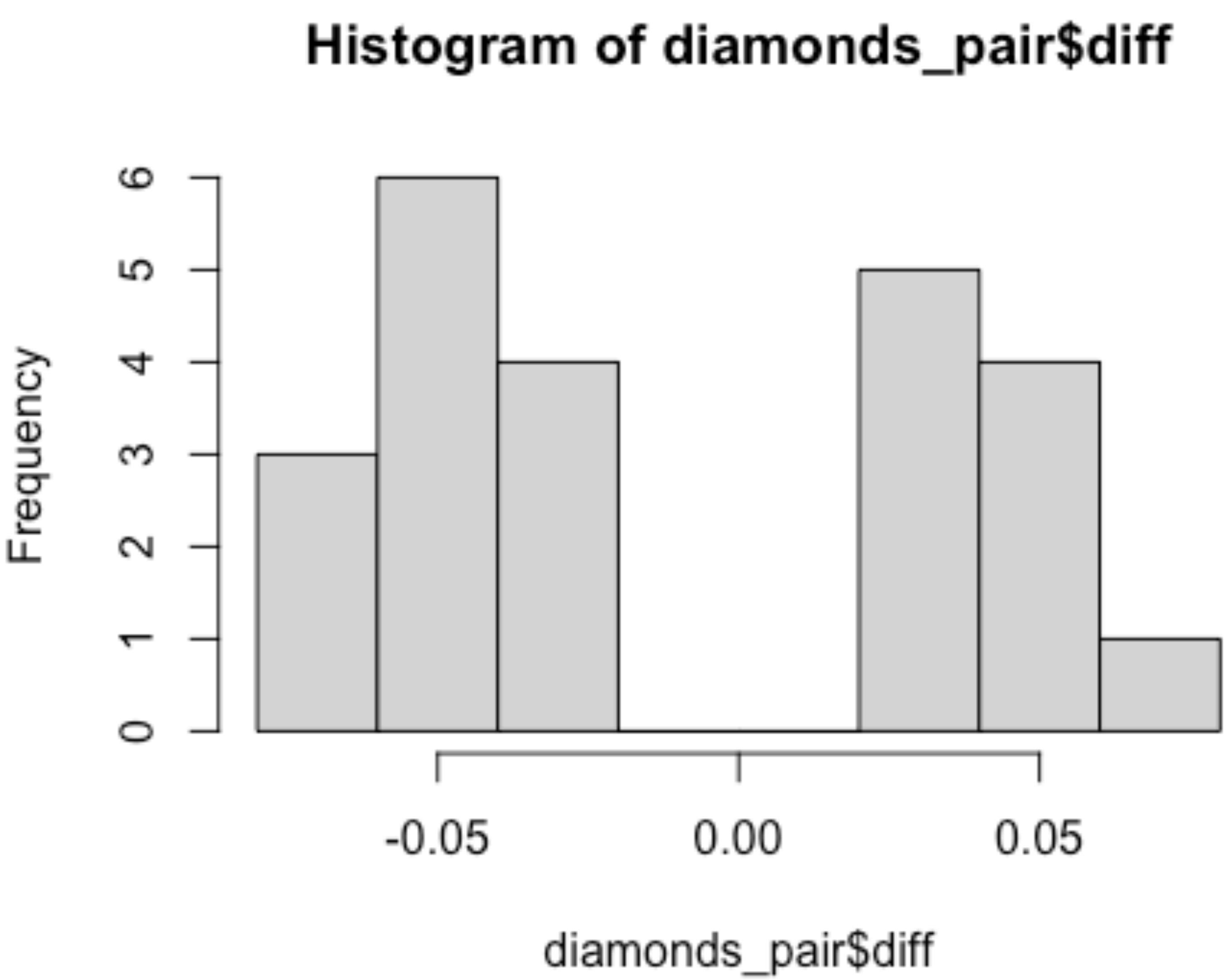
One-sample t-test: `t.test` and confidence interval

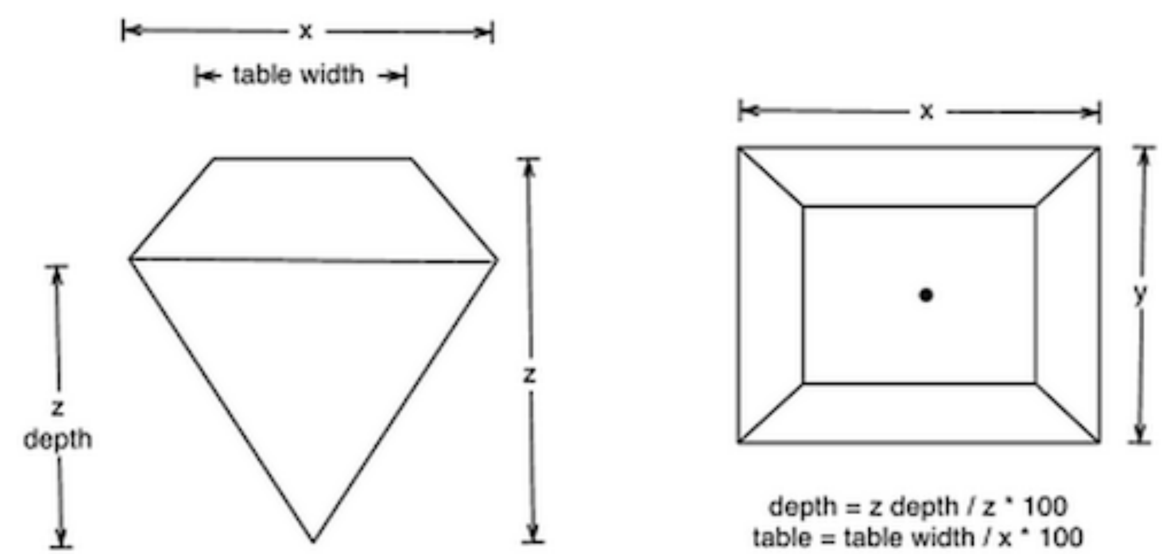
```
t.test(diamonds_best$unit_price, mu = 100, conf.level = 0.05, alternative =
"greater")
##
##  One Sample t-test
##
## data:  diamonds_best$unit_price
## t = -5.5397, df = 22, p-value = 1
## alternative hypothesis: true mean is greater than 100
## 5 percent confidence interval:
##  76.37944      Inf
## sample estimates:
## mean of x
##   65.7687
```

Paired t-test

Q: Do "good" diamonds tend to be circle or square form?

```
diamonds_pair = diamonds_best %>%
  mutate(diff = x-y)
hist(diamonds_pair$diff)
```





Paired t-test: p-value, critical value

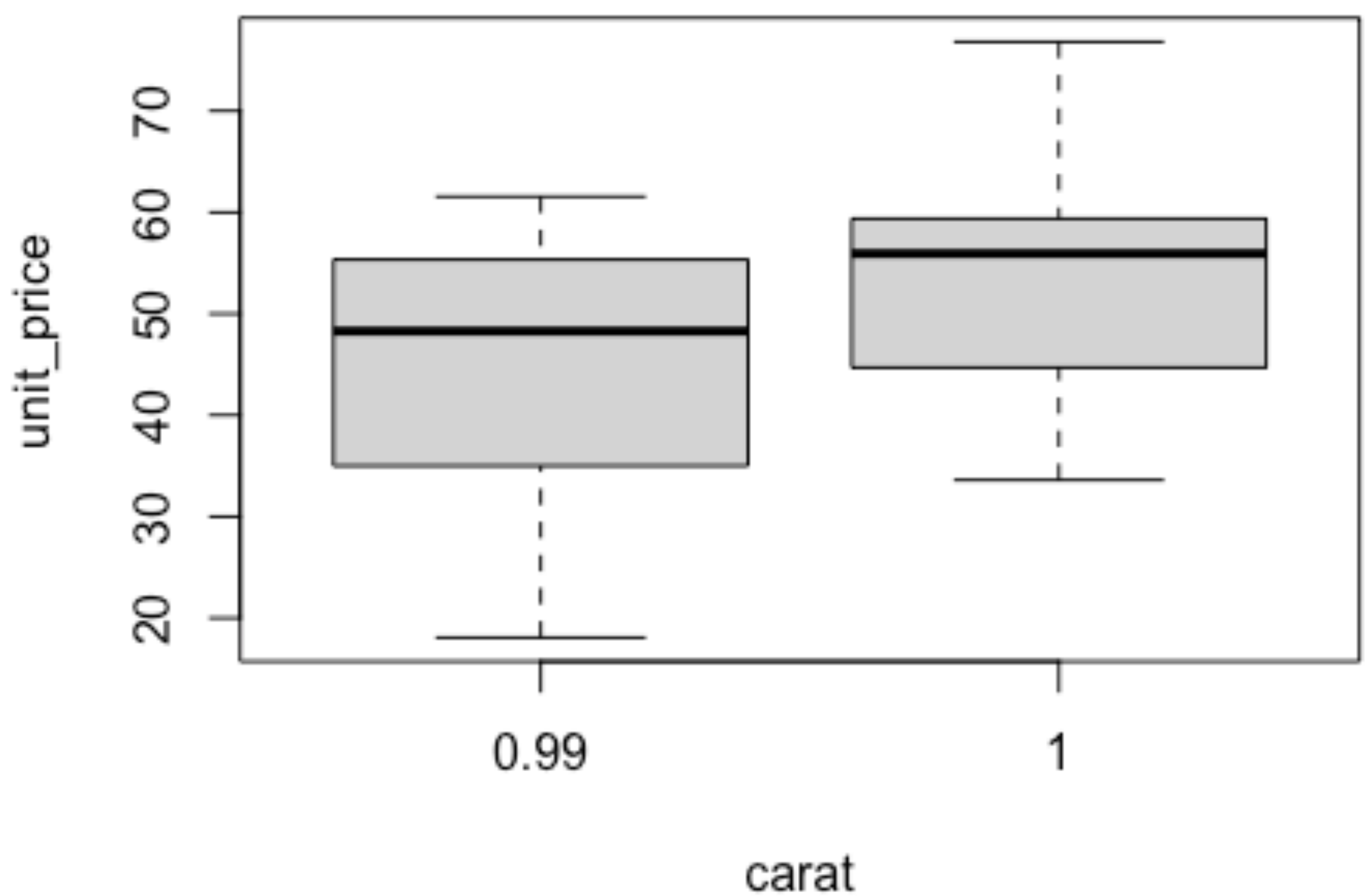
```
mu_pair = mean(diamonds_pair$diff)
sigma_pair = sd(diamonds_pair$diff)
df = nrow(diamonds_pair)-1
(t.stat.pair = mu_pair/(sigma_pair/sqrt(nrow(diamonds_pair))))
## [1] -0.3810935
## p-value
2*pt(-abs(t.stat.pair), df) # H_a: diff != 0
## [1] 0.7067893
pt(t.stat.pair, df) # H_a: diff < 0
## [1] 0.3533946
1-pt(t.stat.pair, df) # H_a: diff > 0
## [1] 0.6466054
## critical value
qt(0.975, df)
## [1] 2.073873
```

Paired t-test: `t.test` and confidence interval

```
t.test(diamonds_pair$x, diamonds_pair$y, paired=TRUE, conf.level = 0.95)
##
##  Paired t-test
##
## data:  diamonds_pair$x and diamonds_pair$y
## t = -0.38109, df = 22, p-value = 0.7068
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  -0.02520744  0.01738135
## sample estimates:
## mean difference
##    -0.003913043
```

Difference in two means

```
diamonds_99 = diamonds %>%
  filter(carat == 0.99) %>%
  mutate(unit_price = price/(carat*100), carat_binary = ifelse(carat == 0.99, 0,1))
diamonds_10 = diamonds %>%
  filter(carat == 1.0) %>%
  mutate(unit_price = price/(carat*100), carat_binary = ifelse(carat == 0.99, 0,1))
set.seed(220729)
diamonds_10 = diamonds_10[sample(1:nrow(diamonds_10),30,replace = FALSE),]
diamonds_new = rbind(diamonds_99, diamonds_10)
boxplot(unit_price~carat, data = diamonds_new)
```



Difference in two means: p-value, critical value

```
mu_99 = mean(diamonds_99$unit_price)
sd_99 = sd(diamonds_99$unit_price)
mu_10 = mean(diamonds_10$unit_price)
sd_10 = sd(diamonds_10$unit_price)
sd_diff = sqrt((sd_99^2/nrow(diamonds_99))+(sd_10^2/
nrow(diamonds_10)))
df_new = min(nrow(diamonds_99)-1, nrow(diamonds_10)-1)
(t.stat.two = (mu_10-mu_99)/sd_diff)
## [1] 2.803798
## p-value
2*pt(-abs(t.stat.two), df_new) # H_a: diff != 0
## [1] 0.01034594
pt(t.stat.two, df_new) # H_a: diff < 0
## [1] 0.994827
1-pt(t.stat.two, df_new) # H_a: diff > 0
## [1] 0.005172969
## critical value
qt(0.975, df_new)
## [1] 2.073873
```

Difference in two means: `t.test` and confidence interval

```
t.test(diamonds_10$unit_price,diamonds_99$unit_price,  conf.level = 0.95)
##
##  Welch Two Sample t-test
##
## data:  diamonds_10$unit_price and diamonds_99$unit_price
## t = 2.8038, df = 42.946, p-value = 0.007552
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.72062 16.66377
## sample estimates:
## mean of x mean of y
##  54.19900  44.50681
```