Notes: MS 204 Chapter 4.3

Overview

- Differences in two population means

An example

Is there a difference between the prices of .99 carat diamonds and 1 carat diamonds? They look the same, but does the extra rounding lead towards retailers charging more? We'll investigate using a random sample of the `diamonds` data set.

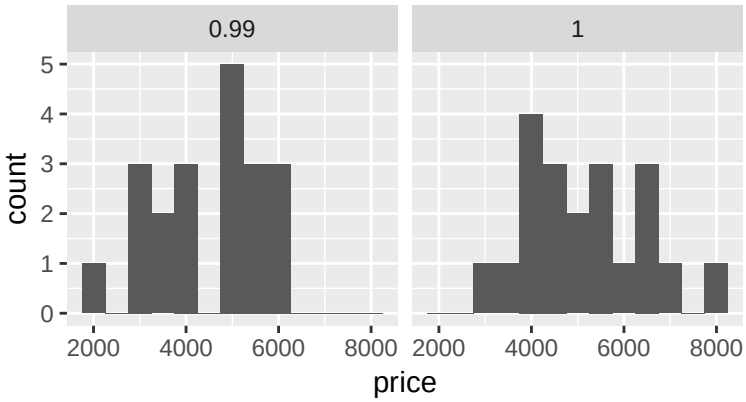
```
library(tidyverse)
library(oilabs)
library(mosaic)
set.seed(0)
diamonds.sample <- diamonds %>% filter(carat == 0.99 | carat == 1.00) %>%
  group_by(carat) %>%
  sample_n(20)
head(diamonds.sample)
```

```
## # A tibble: 6 x 10
## # Groups:   carat [1]
##   carat       cut color clarity depth table price     x     y     z
##   <dbl>     <ord> <ord>   <ord> <dbl> <dbl> <int> <dbl> <dbl> <dbl>
## 1  0.99 Very Good     F     VS2  61.8    57  6094  6.37  6.39  3.94
## 2  0.99 Very Good     J     SI1  60.3    57  4002  6.44  6.49  3.90
## 3  0.99   Premium     F     SI2  60.6    61  4075  6.45  6.38  3.89
## 4  0.99 Very Good     G     SI1  62.8    56  4863  6.34  6.36  3.99
## 5  0.99   Premium     F     VS2  62.6    55  5893  6.50  6.35  4.02
## 6  0.99      Fair     I     SI1  60.7    66  3337  6.42  6.34  3.87
```

```
diamonds.sample %>%
  group_by(carat) %>%
  summarise(ave.price = mean(price), sd.price = sd(price), sample.size = n())
```

```
## # A tibble: 2 x 4
##   carat ave.price sd.price sample.size
##   <dbl>     <dbl>    <dbl>       <int>
## 1  0.99   4420.50 1235.160          20
## 2  1.00   5158.35 1296.157          20
```

```
qplot(price, data = diamonds.sample, binwidth = 500) + facet_wrap(~carat)
```



Inference for differences in two population means

parameter

point estimate

population

Inference

Central limit theorem for differences in two population means

Assumptions?

Hypothesis test

Do these data provide convincing evidence that there is a difference between the average price of 0.99 carat diamonds and 1.00 carat diamonds?

Confidence interval

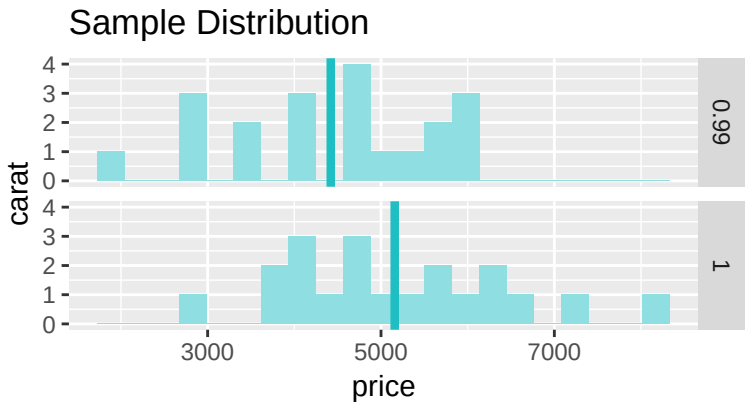
```
qt(0.025, df = 19)
```

```
## [1] -2.093024
```

Code

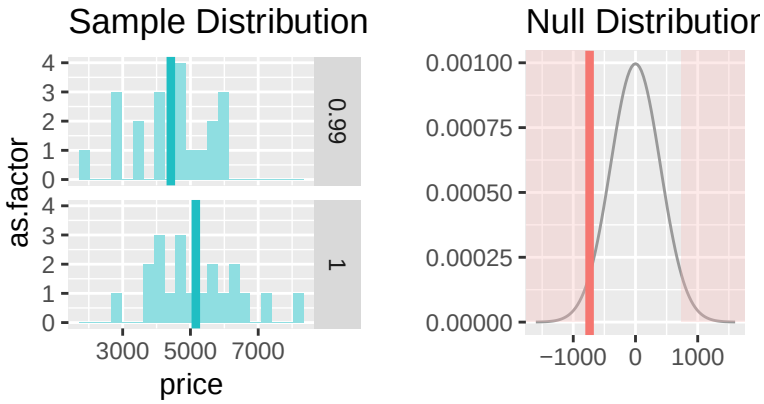
```
inference(y = price, x = carat, data = diamonds.sample, statistic = "mean",
          type = "ci", method = "theoretical")
```

```
## Response variable: numerical, Explanatory variable: categorical (2 levels)
## n_0.99 = 20, y_bar_0.99 = 4420.5, s_0.99 = 1235.1605
## n_1 = 20, y_bar_1 = 5158.35, s_1 = 1296.1569
## 95% CI (0.99 - 1): (-1575.7976 , 100.0976)
```




```
inference(y = price, x = as.factor(carat), data = diamonds.sample, statistic = "mean",
          type = "ht", alternative = "twosided", method = "theoretical", null = 0)
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_0.99 = 20, y_bar_0.99 = 4420.5, s_0.99 = 1235.1605
## n_1 = 20, y_bar_1 = 5158.35, s_1 = 1296.1569
## H0: mu_0.99 =  mu_1
## HA: mu_0.99 != mu_1
## t = -1.843, df = 19
## p_value = 0.081
```




```
inference(y = price, x = as.factor(carat), data = diamonds.sample, statistic = "mean",
          type = "ht", alternative = "less", method = "theoretical", null = 0)
```

```
## Response variable: numerical
## Explanatory variable: categorical (2 levels)
## n_0.99 = 20, y_bar_0.99 = 4420.5, s_0.99 = 1235.1605
## n_1 = 20, y_bar_1 = 5158.35, s_1 = 1296.1569
## H0: mu_0.99 =  mu_1
## HA: mu_0.99 < mu_1
## t = -1.843, df = 19
## p_value = 0.0405
```