Unit 3: Inference for Categorical and Numerical Data

3. Difference of two means

(Chapter 4.3)

3/14/2022

Recap

1. When our samples are too small, we shouldn’t use the Normal distribution. We use the t distribution to make up for uncertainty in our sample statistics

2. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)

3. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values

Key ideas

1. We can use the t-distribution to estimate the probability of a difference between unpaired values.

2. Degrees of freedom depends on the size of both samples

3. The right test depends on where you think variance comes from

The price of diamonds

The mass of diamonds is measured in units called *carats.*
(1 carat ~200 milligrams)

The difference in size between a .99 carat diamond and a 1 carat diamond is undetectable to the human eye.

But is a 1 carat diamond more expensive?

Let's compare the mean prices of .99 and 1.00 carat diamonds

.85 carat

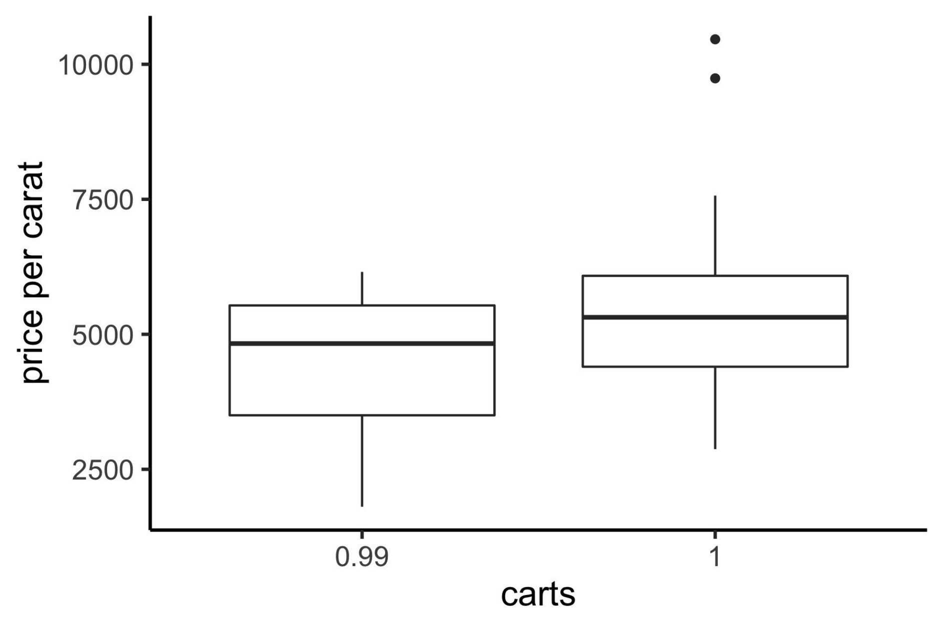
1.00 carat

http://www.zales.com/jewelry101/index.jsp?page=diamonds_Carat

Let's look at some data

I divided the price of each diamond by the number of carats to get a price per carat. **Why?**

	.99c	**1 c**
$\bar{x}$	4451	5486
s	1332	1671
n	23	30



Data are a random sample from the diamonds data set in the ggplot2 package

Parameter and point estimate

Parameter of interest: Difference between the average price per carat of all .99 carat and 1 carat diamonds.

$$\mu_{.99} - \mu_{1}$$

Point estimate: Difference between the average price of sampled .99 carat and 1 carat diamonds.

$$\bar{x}_{.99} - \bar{x}_{1}$$

Practice Question 1

Which is the correct set of hypotheses to test if the average price of 1 carat diamonds is higher than the average price of 0.99 carat diamonds?

a) $H_0: \mu_{.99} = \mu_1$
   $H_A: \mu_{.99} \neq \mu_1$

b) $H_0: \mu_{.99} = \mu_1$
   $H_A: \mu_{.99} > \mu_1$

c) $H_0: \mu_{.99} = \mu_1$
   $H_A: \mu_{.99} < \mu_1$

d) $H_0: \bar{x}_{.99} = \bar{x}_1$
   $H_A: \bar{x}_{.99} < \bar{x}_1$

Practice Question 1

Which is the correct set of hypotheses to test if the average price of 1 carat diamonds is higher than the average price of 0.99 carat diamonds?

a) $H_0: \mu_{.99} = \mu_1$
$H_A: \mu_{.99} \neq \mu_1$

b) $H_0: \mu_{.99} = \mu_1$
$H_A: \mu_{.99} > \mu_1$

c) $\boldsymbol{H_0: \mu_{.99} = \mu_1}$
$\boldsymbol{H_A: \mu_{.99} < \mu_1}$

d) $H_0: \bar{x}_{.99} = \bar{x}_1$
$H_A: \bar{x}_{.99} < \bar{x}_1$

Practice Question 2

Which of the following does <u>not</u> need to be satisfied to conduct using the hypothesis test using t-tests?

a) Per-carat rice of one 0.99 carat diamond in the sample should be independent of another, and the per-carat price of one 1 carat diamond should independent of another as well.

b) Per-carat prices of 0.99 carat and 1 carat diamonds in the sample should be independent.

c) Distributions of per-carat prices of 0.99 and 1 carat diamonds should not be extremely skewed.

d) Both sample sizes should be at least 30.

Practice Question 2

Which of the following does <u>not</u> need to be satisfied to conduct using the hypothesis test using t-tests?

a) Per-carat rice of one 0.99 carat diamond in the sample should be independent of another, and the per-carat price of one 1 carat diamond should independent of another as well.

b) Per-carat prices of 0.99 carat and 1 carat diamonds in the sample should be independent.

c) Distributions of per-carat prices of 0.99 and 1 carat diamonds should not be extremely skewed.

d) Both sample sizes should be at least 30.

Defining the test statistic

The test statistic for inference on the difference of two small sample means ($n_1 < 30$ and/or $n_2 < 30$) mean is the *T* statistic.

$$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$\text{point estimate} = \bar{x}_1 - \bar{x}_2$$

$$\text{null value} = 0$$

where $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $df = min(n_1 - 1, n_2 - 1)$

Note: the true *df* is actually different and more complex to calculate (it involves the variance in each estimate relative to its size). But this is close.

Computing the test statistic

So…

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

$$= \frac{(4451 - 5486) - 0}{}$$

$$= \frac{-1035}{413}$$

$$= -2.51$$

	.99c	1 c
$\bar{x}$	4451	5486
s	1332	1671
n	23	30

Practice Question 3

What is the correct degrees of freedom for this test?

a) 22

b) 23

c) 29

d) 30

e) 50

	.99c	**1 c**
$\bar{x}$	4451	5486
s	1332	1671
n	23	30

Practice Question 3

What is the correct degrees of freedom for this test?

a) **22**

b) 23

c) 29

d) 30

e) 50

$$df = \min(n_{.99} - 1, n_1 - 1)$$

$$= \min(23 - 1, 30 - 1)$$

$$= \min(22,29)$$

$$= 22$$

Computing the p-value

```
> qt(.05, 22) = -1.72
```

(Compare to our t-value -2.51)

Why not qt(.025, 22)?

What is the conclusion of the hypothesis test? How (if at all) would this conclusion change your behavior if you went diamond shopping?

- p-value is small so reject H_0. The data provide convincing evidence to suggest that the per-carat price of 0.99 carat diamonds is lower than the per-carat price of 1 carat diamonds.

- Maybe buy a 0.99 carat diamond? It looks like a 1 carat, but is significantly cheaper.

Practice Question 4

What is the equivalent confidence interval for a one-sided hypothesis test with α = 0.05?

a) 90%

b) 92.5%

c) 95%

d) 97.5%

Practice Question 4

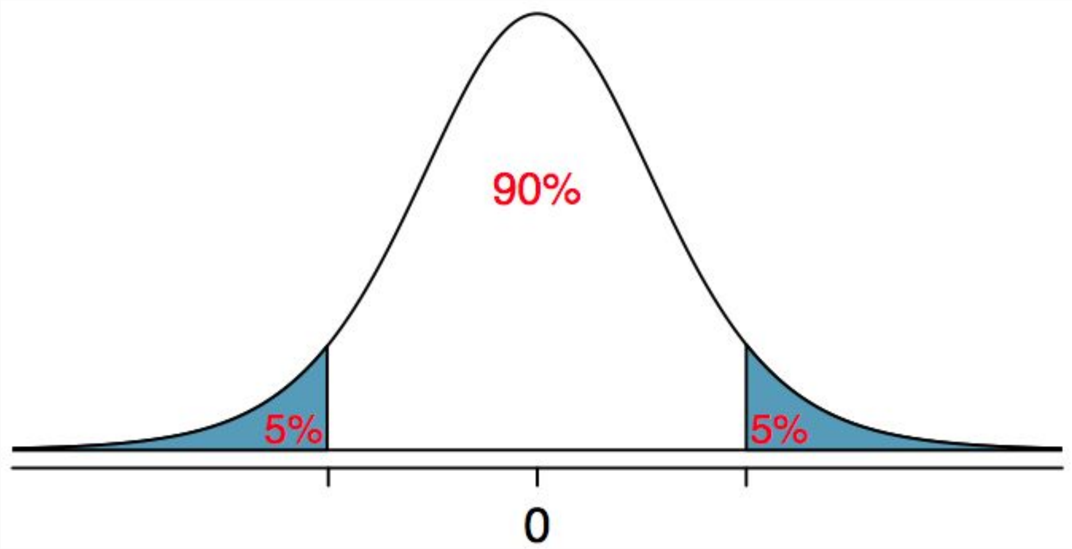
What is the equivalent confidence interval for a one-sided hypothesis test with α = 0.05?

a) **90%**

b) 92.5%

c) 95%

d) 97.5%



Practice Question 4

Ok so let's compute the confidence interval:

```
> qt(.05, 22) = -1.72
```

Same value!

$$(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^{\star}_{df} \times SE = (4451 - 5486) \pm 1.72 \times 413$$

$$= -1035 \pm 710$$

$$= (-1745, -325)$$

We are 90% confident that the average per-carat of a .99 carat diamond is $1745 to $325 lower than the average per-carat price of a 1 carat diamond.

Key ideas

1. We can use the t-distribution to estimate the probability of a difference between unpaired values.

2. Degrees of freedom depends on the size of both samples

3. The right test depends on where you think variance comes from