Unit 3: Inference for Categorical and Numerical Data

3. Difference of two means

(Chapter 4.3)

11/08/2017

Recap from last time

1. We can use the t-distribution either to estimate the probability of either a single value, or the difference between two paired values
2. We can keep using the t-distribution even when the number of samples is large (it asymptotically approaches the normal)
3. All of our statistical theory still holds, we are just plugging in different distributions

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

Questions Statisticians Answer...

	Means	Proportions
	When we measure a **number** for each individual (e.g. height, weight, SAT score). We use the **mean** to summarize the data for a group. *-Example: The average weight of a sample of 100 granny smith apples from my orchard was 72 grams.*	When we measure which of **one of two options** is true for each individual (e.g. organ donor or not?, lived or died? red or blue?) we use the **proportion** to summarise the data for a group. ***-****Example: The proportion of people at this fertility clinic who have live births using ART is 0.4 (i.e. 40%).*
One Sample	Is the mean in my group different from some value I care about? *-Example: Are the apples in my orchard heavier than typical granny smith apples (which have a known average weight of 70 grams)?* *-Null: Mean in my orchard = 70g* *-Alternative: Mean in my orchard > 70g*	Is the proportion in my group different from some value I care about? *-Example: Is the probability of a having a live birth at my clinic higher than the known US probability of a live birth using ART (which is 0.3)?* *-Null: Probability in my clinic = 0.3* *-Alternative: Probability in my clinic > 0.3*
Two Samples	Are the means of two groups different? *-Example: Does my new drug lower cholesterol more than a placebo?* *-Null: Mean cholesterol of drug takers = Mean cholesterol of placebo takers* *-Alternative: Mean cholesterol of drug takers < Mean cholesterol of placebo takers*	Are the proportions in two groups different? *-Example: Is Curry's probability of making a shot higher for "Hot Shots" than for "Not Shots"* *-Null: Probability for hot shots = Probability for not shots* *-Alternative: Probability for hot shots > Probability for not shots*

The price of diamonds

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

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

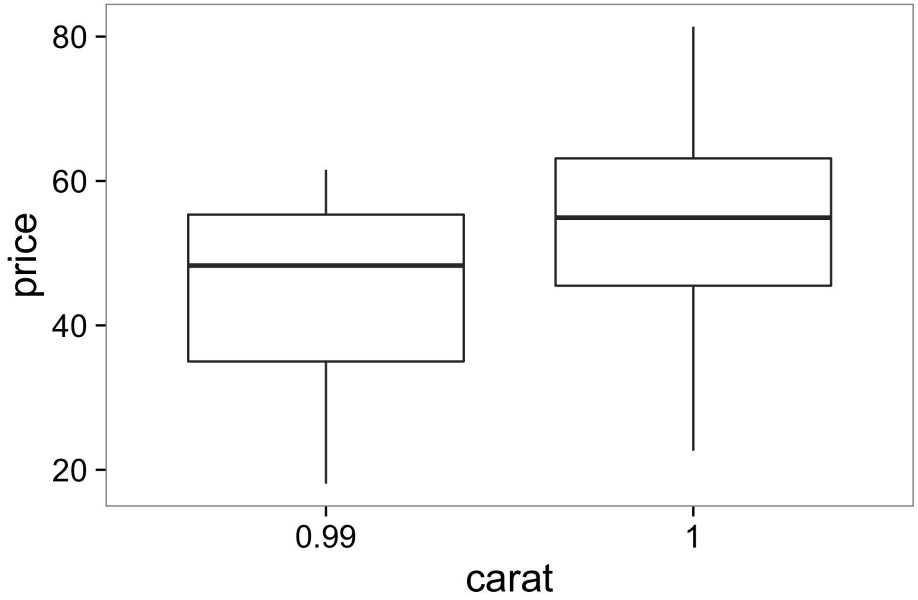
1.00 carat

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

Let's look at some data

I divided the price of each diamond by 100*carat to get a price per .01 carat (pt) just for ease of comparison

	.99c	**1 c**
$\bar{x}$	44.50	53.43
s	13.32	12.22
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 of <u>all</u> .99 carat and 1 carat diamonds.

$$\mu_{pt99} - \mu_{pt100}$$

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

$$\bar{x}_{99} - \bar{x}_{pt100}$$

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) $H_0: \mu_{pt99} = \mu_{pt100}$
$H_A: \mu_{pt99} \neq \mu_{pt100}$

b) $H_0: \mu_{pt99} = \mu_{pt100}$
$H_A: \mu_{pt99} > \mu_{pt100}$

c) $H_0: \mu_{pt99} = \mu_{pt100}$
$H_A: \mu_{pt99} < \mu_{pt100}$

d) $H_0: \bar{x}_{pt99} = \bar{x}_{pt100}$
$H_A: \bar{x}_{pt99} < \bar{x}_{pt100}$

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) $H_0: \mu_{pt99} = \mu_{pt100}$
$H_A: \mu_{pt99} \neq \mu_{pt100}$

b) $H_0: \mu_{pt99} = \mu_{pt100}$
$H_A: \mu_{pt99} > \mu_{pt100}$

c) $\boldsymbol{H_0: \mu_{pt99} = \mu_{pt100}}$
$\boldsymbol{H_A: \mu_{pt99} < \mu_{pt100}}$

d) $H_0: \bar{x}_{pt99} = \bar{x}_{pt100}$
$H_A: \bar{x}_{pt99} < \bar{x}_{pt100}$

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) Point price of one 0.99 carat diamond in the sample should be independent of another, and the point price of one 1 carat diamond should independent of another as well.

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

c) Distributions of point 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) Point price of one 0.99 carat diamond in the sample should be independent of another, and the point price of one 1 carat diamond should independent of another as well.

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

c) Distributions of point 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 it's size). But this is a reasonable approximation.

Computing the test statistic

So...

$$
\begin{aligned}
T &= \frac{\text{point estimate} - \text{null value}}{SE} \\
&= \frac{(44.50 - 53.43) - 0}{} \\
&= \frac{-8.93}{3.56} \\
&= -2.508
\end{aligned}
$$

	.99c	**1 c**
$\bar{x}$	44.50	53.43
s	13.32	12.22
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**
$\bar{x}$	44.50	53.43
s	13.32	12.22
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

$$df = \min(n_{pt99} - 1, n_{pt100} - 1)$$

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

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

$$= 22$$

Computing the p-value

`> qt(.05, 22) = -1.72` (Compare to our t-value -2.508)

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 point price of 0.99 carat diamonds is lower than the point price of 1 carat diamonds.
- Maybe buy a 0.99 carat diamond? It looks like a 1 carat, but is significantly cheaper.

Practice Question 3

What is the correct degrees of freedom for this test?

a) **22**

b) 23

c) 29

d) 30

e) 50

$$df = \min(n_{pt99} - 1, n_{pt100} - 1)$$

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

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

$$= 22$$

Practice Question 4

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

a) 90%

b) 92.5%

c) 95%

d) 97.5%

Practice Question 4

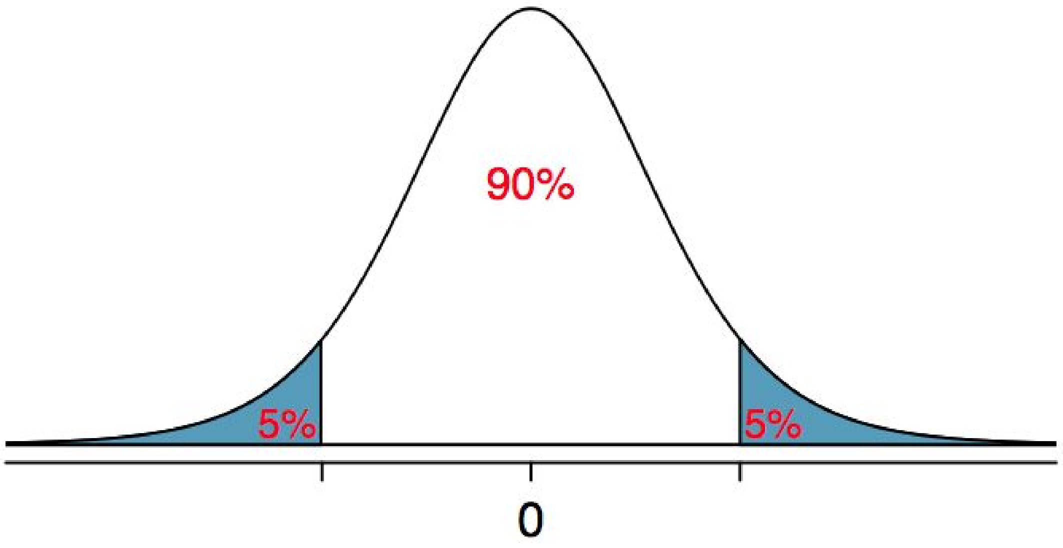
What is the equivalent confidence interval for a one-sided hypothesis test with $\alpha = 0.05$?

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!

$$
\begin{aligned}
(\bar{x}_{pt99} - \bar{x}_{pt1}) \pm t^{\star}_{df} \times SE &= (44.50 - 53.43) \pm 1.72 \times 3.56 \\
&= -8.93 \pm 6.12 \\
&= (-15.05, -2.81)
\end{aligned}
$$

We are 90% confident that the average point price of a .99 carat diamond is $15.05 to $2.81 lower than the average point 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