Sta 111 - Summer II 2017
Probability and Statistical Inference

9. Inference using the *t* distribution

Lu Wang

Duke University, Department of Statistical Science

July 18, 2017

Outline

1. T corrects for uncertainty introduced by plugging in s for σ
2. Introducing the t distribution
3. Evaluating hypotheses using the t distribution
4. Hypothesis tests and confidence intervals at equivalent significance/confidence levels should agree
5. Difference of two means
    1. Test statistic for the difference of two means
    2. Confidence intervals for the difference of two means
6. Summary
7. Homework 3

Friday the 13^{th}

Between 1990 - 1992 researchers in the UK collected data on traffic flow, accidents, and hospital admissions on Friday 13^{th} and the previous Friday, Friday 6^{th}. Below is an excerpt from this data set on traffic flow. We can assume that traffic flow on given day at locations 1 and 2 are independent.

	type	date	6^{th}	13^{th}	diff	location
1	traffic	1990, July	139246	138548	698	loc 1
2	traffic	1990, July	134012	132908	1104	loc 2
3	traffic	1991, September	137055	136018	1037	loc 1
4	traffic	1991, September	133732	131843	1889	loc 2
5	traffic	1991, December	123552	121641	1911	loc 1
6	traffic	1991, December	121139	118723	2416	loc 2
7	traffic	1992, March	128293	125532	2761	loc 1
8	traffic	1992, March	124631	120249	4382	loc 2
9	traffic	1992, November	124609	122770	1839	loc 1
10	traffic	1992, November	117584	117263	321	loc 2

Scanlon, T.J., Luben, R.N., Scanlon, F.L., Singleton, N. (1993), “Is Friday the 13th Bad For Your Health?,” BMJ, 307, 1584-1586.

Friday the 13^{th}

- We want to investigate if people's behavior is different on Friday 13^{th} compared to Friday 6^{th}.
- One approach is to compare the traffic flow on these two days.
- H_0 : Average traffic flow on Friday 6^{th} and 13^{th} are equal.
  H_A : Average traffic flow on Friday 6^{th} and 13^{th} are different.

Each case in the data set represents traffic flow recorded at the same location in the same month of the same year: one count from Friday 6^{th} and the other Friday 13^{th}. Are these two counts independent?

- When two sets of observations have this special correspondence (not independent), they are said to be *paired*.
- To analyze paired data, it is often useful to look at the difference in outcomes for each pair of observations; subtract using a consistent order.

What are the hypotheses for testing for a difference between the average traffic flow between Friday 6^{th} and 13^{th}?

(a) $H_0 : \mu_{6th} = \mu_{13th}$
$H_A : \mu_{6th} \neq \mu_{13th}$

(b) $H_0 : p_{6th} = p_{13th}$
$H_A : p_{6th} \neq p_{13th}$

(c) $H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$

(d) $H_0 : \bar{x}_{diff} = 0$
$H_A : \bar{x}_{diff} = 0$

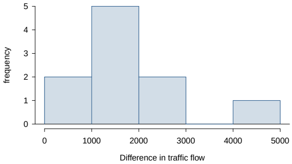
Conditions

- *Independence:* We are told to assume that cases (rows) are independent.

- *Sample size / skew:*

  – The sample distribution does not appear to be extremely skewed, but it's very difficult to assess whether the population distribution to be skewed or not with such a small sample size.

  – We do not know σ and n is too small to assume s is a reliable estimate for σ.



So what do we do when the sample size is small?

The normality condition

As long as observations are independent, and the population distribution is not extremely skewed, a large sample would ensure that...

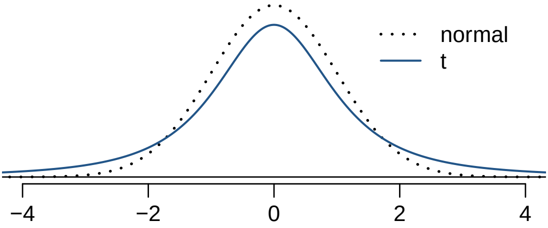
- the sampling distribution of the mean is nearly normal
- the estimate of the standard error, as $\frac{s}{\sqrt{n}}$, is reliable

However, it's inherently difficult to verify normality in small data sets.

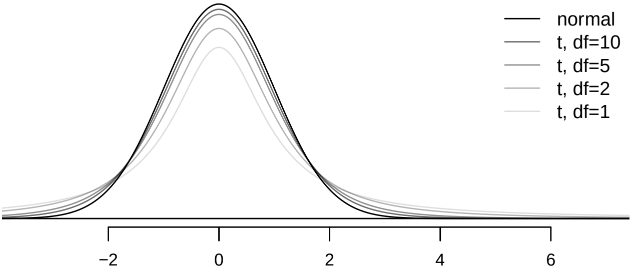
- We should exercise caution when verifying the normality condition for small samples. It is important to not only examine the data but also think about where the data come from.
  - For example, ask: would I expect this distribution to be symmetric, and am I confident that outliers are rare?
- *When the population standard deviation is unknown (almost always), the uncertainty of the standard error estimate is addressed by using a new distribution: the t distribution.*

The *t* distribution

- The *t* distribution also has a bell shape, but its tails are *thicker* than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are exactly the correction we need to resolve the problem of a poorly estimated standard error. (since *n* is small)



- Always centered at zero, like the standard normal (z) distribution.
- Has a single parameter: *degrees of freedom* (*df*).



What happens to shape of the *t* distribution as *df* increases?

Back to Friday the 13th

Test statistic for inference on a small-sample mean

The test statistic for inference on a small-sample mean is the T statistic with $df = n - 1$.

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

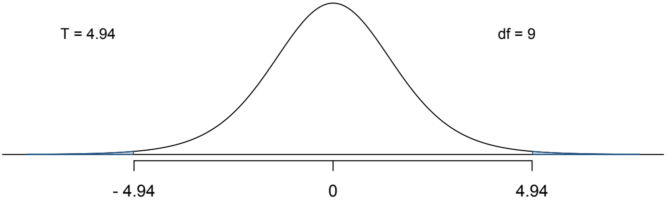
in context...

$$\begin{aligned}
\textit{point estimate} &= \bar{x}_{diff} = 1836 \\
SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \\
T &= \frac{1836 - 0}{372} = 4.94 \\
df &= 10 - 1 = 9
\end{aligned}$$

Note: Null value is 0 because in the null hypothesis we set $\mu_{diff} = 0$.

Finding the p-value

- The p-value is, once again, calculated as the tail area under the *t* distribution.



- Using R:

```
> 2 * pt(-4.94, df = 9)

[1] 0.0008022394
```

- What is the conclusion of the hypothesis test?

-

Constructing confidence intervals using the t distribution

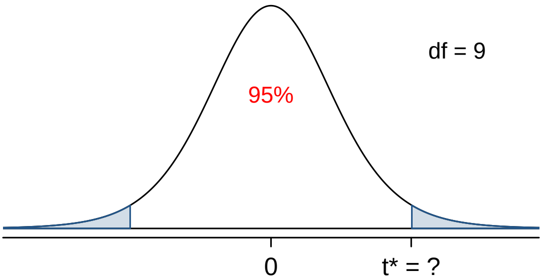
- We concluded that there is a difference in the traffic flow between Friday 6$^{\text{th}}$ and 13$^{\text{th}}$.
- But it would be more interesting to find out what exactly this difference is.
- We can use a confidence interval to estimate this difference.
- Confidence intervals are always of the form

$$\text{point estimate} \pm ME$$

- ME is always calculated as the product of a critical value and SE.
- Since small-sample means follow a t distribution (not a normal distribution), the critical value is a $t^\star$ (as opposed to a $z^\star$).

$$\text{point estimate} \pm t^\star \times SE$$

Finding the critical t (t^*)



Using R:

```
> qt(p=0.975, df=9)

[1] 2.262157
```

Which of the following is the correct calculation of a 95% confidence interval for the difference between the traffic flow between Friday 6th and 13th?

$$\bar{x}_{diff} = 1836 \qquad s_{diff} = 1176 \qquad n = 10 \qquad SE = 372$$

(a) $1836 \pm 1.96 \times 372$

(b) $1836 \pm 2.26 \times 372$

(c) $1836 \pm -2.26 \times 372$

(d) $1836 \pm 2.26 \times 1176$

Which of the following is the *best* interpretation for the confidence interval we just calculated?

$$\mu_{diff:6th-13th} = (995, 2677)$$

We are 95% confident that ...

(a) the difference between the average number of cars on the road on Friday 6^{th} and 13^{th} is between 995 and 2,677.

(b) on Friday 6^{th} there are 995 to 2,677 fewer cars on the road than on the Friday 13^{th}, on average.

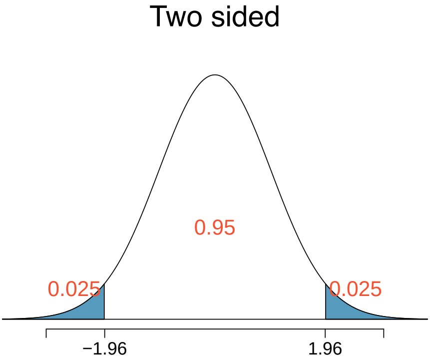
(c) on Friday 6^{th} there are 995 fewer to 2,677 more cars on the road than on the Friday 13^{th}, on average.

(d) on Friday 13^{th} there are 995 to 2,677 fewer cars on the road than on the Friday 6^{th}, on average.

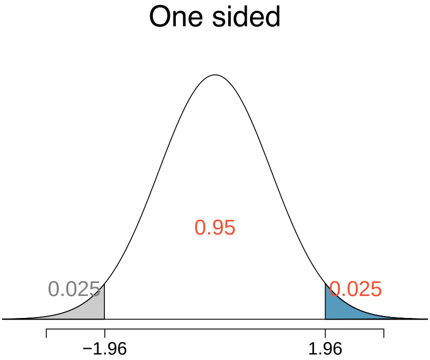
Does the conclusion from the hypothesis test agree with the findings of the confidence interval?

Do you think the findings of this study suggests that people believe Friday 13th is a day of bad luck?

Hypothesis tests and confidence intervals at equivalent significance/confidence levels should agree



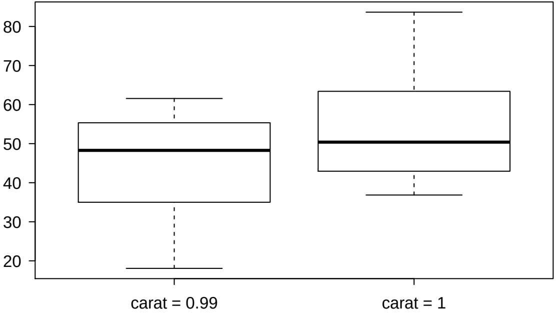
95% confidence level
is equivalent to
two sided HT with $\alpha = 0.05$



95% confidence level
corresponds to
one sided HT with $\alpha = 0.025$

- Weights of diamonds are measured in carats.
- 1 carat = 100 points, 0.99 carats = 99 points, etc.
- The difference between the size of a 0.99 carat diamond and a 1 carat diamond is undetectable to the naked human eye, but does the price of a 1 carat diamond tend to be higher than the price of a 0.99 diamond?
- We are going to test if there is a difference between the average prices of 0.99 and 1 carat diamonds.
- In order to be able to compare equivalent units, we divide the prices of 0.99 carat diamonds by 99 and 1 carat diamonds by 100, and compare the average point prices.

Data



	0.99 carat pt99	*1 carat* pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

These data are a small random sample from the `diamonds` data set in `ggplot2` R package.

- *Parameter of interest:* Average difference between the point prices of *all* 0.99 carat and 1 carat diamonds.

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

- *Point estimate:* Average difference between the point prices of *sampled* 0.99 carat and 1 carat diamonds.

$$\bar{x}_{pt99} - \bar{x}_{pt100}$$

Hypotheses

Which of the following is the correct set of hypotheses for testing if the average point price of 1 carat diamonds ($_{pt100}$) is higher than the average point price of 0.99 carat diamonds ($_{pt99}$)?

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

Which of the following does not need to be satisfied in order to conduct this hypothesis test using theoretical methods?

(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 be 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.

Sampling distribution of test statistic

Test statistic for inference on the difference of two small sample means

The test statistic for inference on the difference of two means where σ_1 and σ_2 are unknown is the T statistic.

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

where

$$\text{point estimate} = \bar{x}_{pt99} - \bar{x}_{pt100} \quad \text{and} \quad SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

This test statistic has a t distribution with $df = \min(n_1 - 1, n_2 - 1)$

Note: The calculation of the df is actually much more complicated. For simplicity we'll use the above formula to estimate the true df when conducting the analysis by hand.

Test statistic (cont.)

	0.99 carat pt99	*1 carat* pt100
$\bar{x}$	44.50	53.43
s	13.32	12.22
n	23	30

in context...

$$
\begin{aligned}
T &= \frac{\text{point estimate} - \text{null value}}{SE} \\
&= \frac{(44.50 - 53.43) - 0}{\sqrt{\frac{13.32^2}{23} + \frac{12.22^2}{30}}} \\
&= \frac{-8.93}{3.56} \\
&= -2.508
\end{aligned}
$$

Which of the following is the correct *df* for this hypothesis test?

(a) 22
(b) 23
(c) 30
(d) 29
(e) 52

What is the p-value for this hypothesis test?

$$T = -2.508 \qquad df = 22$$

Using R:

```
> pt(q=-2.508, df=22)

[1] 0.0100071
```

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

Equivalent confidence level

What is the corresponding confidence level for a one-sided hypothesis test at $\alpha = 0.05$?

(a) 90%

(b) 92.5%

(c) 95%

(d) 97.5%

Critical value & Confidence interval

- What is the appropriate $t^\star$ for a confidence interval for the average difference between the point prices of 0.99 and 1 carat diamonds?

  Using R:

```
> qt(p=0.95, df=22)

[1] 1.717144
```

- Calculate the interval, and interpret it in context.

$$\text{point estimate} \pm ME$$

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

  - We are 90% confident that the average point price of a 0.99 carat diamond is $15.05 to $2.81 lower than the average point price of a 1 carat diamond.

Recap: Inference using *t* distribution

$$HT: \textit{test statistic} = \frac{\textit{point estimate} - \textit{null}}{SE}$$

$$CI: \textit{point estimate} \pm \textit{critical value} \times SE$$

One mean:

$df = n - 1$

HT:

$H_0 : \mu = \mu_0$

$T_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

CI:

$\bar{x} \pm t^\star_{df} \frac{s}{\sqrt{n}}$

Paired means:

$df = n_{diff} - 1$

HT:

$H_0 : \mu_{diff} = 0$

$T_{df} = \frac{\bar{x}_{diff} - 0}{\frac{s_{diff}}{\sqrt{n_{diff}}}}$

CI:

$\bar{x}_{diff} \pm t^\star_{df} \frac{s_{diff}}{\sqrt{n_{diff}}}$

Independent means:

$df = min(n_1 - 1, n_2 - 1)$

HT:

$H_0 : \mu_1 - \mu_2 = 0$

$T_{df} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

CI:

$\bar{x}_1 - \bar{x}_2 \pm t^\star_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Homework 3

Graded questions:

- Ch 5: 4.32, 5.4, 5.20, 5.34, 5.44

Practice questions:

- t-inference: 5.1, 5.3, 5.5, 5.13, 5.17, 5.19, 5.21
- Power: 5.39
- ANOVA: 5.41, 5.43, 5.45, 5.47, 5.49, 5.51