One-sample mean with the *t*-distribution

Friday the 13th

Between 1990 - 1992 researchers in the UK collected data on traffic flow, accidents, and hospital admissions on Friday 13th and the previous Friday, Friday 6th. 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	6th	13th	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

Friday the 13th

- We want to investigate if people’s behavior is different on Friday the 13th compared to Friday 6th.

Friday the 13th

- We want to investigate if people’s behavior is different on Friday the 13th compared to Friday 6th.
- One approach is to compare the traffic flow on these two days.

Friday the 13^{th}

- We want to investigate if people's behavior is different on Friday the 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.

Friday the 13^{th}

- We want to investigate if people's behavior is different on Friday the 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?

Friday the 13^{th}

- We want to investigate if people's behavior is different on Friday the 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?

No!

Hypotheses

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

A. $H_0 : \boldsymbol{\mu}_{6th} = \boldsymbol{\mu}_{13th}$

   $H_A : \boldsymbol{\mu}_{6th} \neq \boldsymbol{\mu}_{13th}$

B. $H_0 : \boldsymbol{p}_{6th} = \boldsymbol{p}_{13th}$

   $H_A : \boldsymbol{p}_{6th} \neq \boldsymbol{p}_{13th}$

C. $H_0 : \boldsymbol{\mu}_{diff} = 0$

   $H_A : \boldsymbol{\mu}_{diff} \neq 0$

D. $H_0 : \bar{x}_{diff} = 0$

   $H_A : \bar{x}_{diff} \neq 0$

Hypotheses

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

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. $H_0 : \bar{x}_{diff} = 0$

$H_A : \bar{x}_{diff} \neq 0$

Conditions

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

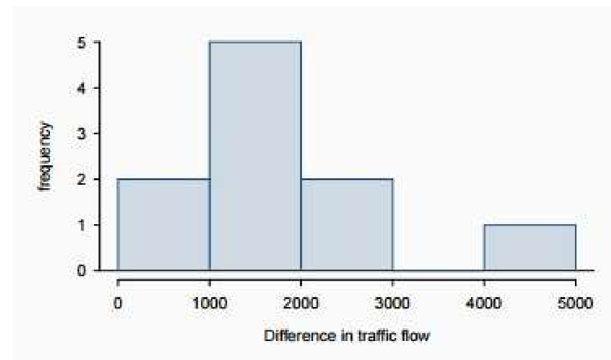
Conditions

- *Independence*: We are told to assume that cases (rows) are independent
- Sample size / skew:

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 with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.

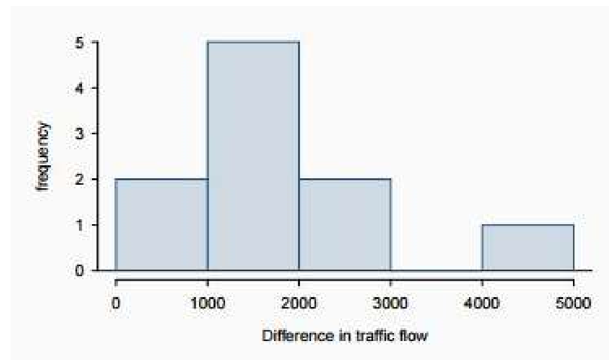


- We do not know $\boldsymbol{\sigma}$ and n is too small to assume s is reliable estimate for $\boldsymbol{\sigma}$

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 with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not probably not, it should be equally likely to have days with lower than average traffic and higher than average traffic.



- We do not know $\boldsymbol{\sigma}$ and n is too small to assume s is reliable estimate for $\boldsymbol{\sigma}$

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

Review:
what purpose does a large sample serve?

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

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal
- While this is a helpful special case, it's inherently difficult to verify normality in small data sets

The normality condition

- The CLT, which states that sampling distributions will be nearly normal, hold true for *any* sample size as long as the population distribution is nearly normal
- While this is a helpful special case, 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?

Back to Friday the 13th

	type	date	6th	13th	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

$\bar{x}_{diff} = 1836$

$s_{diff} = 1176$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample ($n < 50$) mean is the T statistic with $df = n - 1$

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

Note: Null value is 0 because in the null hypothesis we set $\boldsymbol{\mu}_{diff} = 0$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample (n < 50) mean is the T statistic with $df = n - 1$

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

in context...

$$point\ estimate = \bar{x}_{diff} = 1836$$

Note: Null value is 0 because in the null hypothesis we set $\boldsymbol{\mu}_{\text{diff}} = 0$

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample (n < 50) mean is the T statistic with $df = n - 1$

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

in context...

$$\begin{aligned} \textit{point estimate} &= \bar{x}_{diff} = 1836 \\ SE &= \frac{s_{diff}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372 \end{aligned}$$

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

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample (n < 50) mean is the T statistic with $df = n - 1$

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

in context...

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

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

Find the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample (n < 50) mean is the T statistic with $df = n - 1$

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

in context...

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

Note: Null value is 0 because in the null hypothesis we set $\boldsymbol{\mu}_{diff} = 0$

Finding the p-value

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

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the t distribution
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)
[1] 0.0008022394
```

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the *t* distribution
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)
[1] 0.0008022394
```

Finding the p-value

- The p-value is, once again, calculated as the area under the tail of the *t* distribution
- Using R:

```
> 2 * pt(4.94, df = 9, lower.tail = FALSE)
[1] 0.0008022394
```

Conclusion of the test

What is the conclusion of this hypothesis test?

Conclusion of the test

What is the conclusion of this hypothesis test?

Since the p-value is quite low, we conclude that the data provide strong evidence of a difference between traffic flow on Friday 6th and 13th.

What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6th and 13th

What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6^{th} and 13^{th}
- But it would be more interesting to find out what exactly this difference is

What is the difference?

- We concluded that there is a difference in the traffic flow between Friday 6th and 13th
- 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 interval for a small sample mean

- Confidence intervals are always of the form

$$point\ estimate \pm ME$$

Confidence interval for a small sample mean

- Confidence intervals are always of the form

$$point\ estimate \pm ME$$

- ME is always calculated as the product of a critical value and SE

Confidence interval for a small sample mean

- Confidence intervals are always of the form

$$point\ estimate \pm ME$$

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

$$point\ estimate \pm t^*\ x\ SE$$

Finding the critical value ($t^\star$)

Using R:

```
> qt(p = 0.975, df = 9)
[1] 2.262157
```

Constructing a CI for a small sample mean

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 ± 1.96 x 372

B. 1836 ± 2.26 x 372

C. 1836 ± -2.26 x 372

D. 1836 ± 2.26 x 1176

Constructing a CI for a small sample mean

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 ± 1.96 x 372

B. 1836 ± 2.26 x 372 → *(995, 2677)*

C. 1836 ± -2.26 x 372

D. 1836 ± 2.26 x 1176

Interpreting the CI

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 6th and 13th is between 995 and 2,677
B. on Friday 6th there are 995 to 2,677 fewer cars on the road than on the Friday 13th, on average
C. on Friday 6th there are 995 fewer to 2,677 more cars on the road than on the Friday 13th, on average
D. on Friday 13th there are 995 to 2,677 fewer cars on the road than on the Friday 6th, on average

Interpreting the CI

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

Synthesis

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?

Synthesis

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

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0.

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

Synthesis

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

Yes, the hypothesis test found a significant difference, and the CI does not contain the null value of 0

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

No, this is an observational study. We have just observed a significant difference between the number of cars on the road on these two days. We have not tested for people's beliefs

Recap: Inference using the *t*-distribution

- If $\boldsymbol{\sigma}$ is unknown, use the *t*-distribution with $SE = \frac{s}{\sqrt{n}}$

Recap: Inference using the t-distribution

- If σ is unknown, use the t-distribution with $SE = \frac{s}{\sqrt{n}}$
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, n < 10% of population)
  - no extreme skew

Recap: Inference using the *t*-distribution

- If σ is unknown, use the *t*-distribution with $SE = \frac{s}{\sqrt{n}}$
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, *n* < 10% of population)
  - no extreme skew
- Hypothesis Testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = n - 1$$

Recap: Inference using the *t*-distribution

- If σ is unknown, use the *t*-distribution with $SE = \frac{s}{\sqrt{n}}$
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, n < 10% of population)
  - no extreme skew
- Hypothesis Testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = n - 1$$

- Confidence interval:

$$point\ estimate \pm t^{*}_{df} \times SE$$

Recap: Inference using the *t*-distribution

- If $\boldsymbol{\sigma}$ is unknown, use the *t*-distribution with $SE = \frac{s}{\sqrt{n}}$
- Conditions:
  - independence of observations (often verified by a random sample, and if sampling without replacement, n < 10% of population)
  - no extreme skew
- Hypothesis Testing:

$$T_{df} = \frac{point\ estimate\ - null\ value}{SE}, where\ df = n - 1$$

- Confidence interval:

$$point\ estimate \pm t^{*}_{df} \times SE$$

Note: The example we used was for paired means (difference between dependent groups). We took the difference between the observations and used only these differences (one sample) in our analysis