Example - Rent in Manhattan

20 Manhattan apartments were randomly sampled and their rents obtained. The dot plot below shows the distribution of the rents of these apartments. Can we apply the methods we have learned so far to construct a confidence interval using these data. Why or why not?
An alternative approach to constructing confidence intervals is *bootstrapping*. This term comes from the phrase “pulling oneself up by one’s bootstraps”, which is a metaphor for accomplishing an impossible task without any outside help. In this case the impossible task is estimating a population parameter, and we’ll accomplish it using data from only the given sample.
Bootstrapping works as follows:

1. take a bootstrap sample - a random sample taken with replacement from the original sample, of the same size as the original sample
2. calculate the bootstrap statistic - a statistic such as mean, median, proportion, etc. computed on the bootstrap samples
3. repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics

The 95% bootstrap confidence interval is estimated by the cutoff values for the middle 95% of the bootstrap distribution.
Example - Rent in Manhattan - Bootstrap interval

The dot plot below shows the distribution of means of 100 bootstrap samples from the original sample. Estimate the 95% bootstrap confidence interval based on this bootstrap distribution.
Example - Rent in Manhattan - Bootstrap interval

The dot plot below shows the distribution of means of 100 bootstrap samples from the original sample. Estimate the 90% bootstrap confidence interval based on this bootstrap distribution.
Randomization testing for a mean

- We can also use a simulation method to conduct the same test.
- This is very similar to bootstrapping, i.e. we randomly sample with replacement from the sample, but this time we shift the bootstrap distribution to be centered at the null value.
- The p-value is then defined as the proportion of simulations that yield a sample mean at least as favorable to the alternative hypothesis as the observed sample mean.
Example - Rent in Manhattan - Randomization testing

You read an article claiming that the average rent for an apartment in Manhattan is $3,100. Your random sample had a mean of $3156.5. Does this sample provide convincing evidence that the article's estimate is an underestimate?

\[ H_0 : \mu = $3100 \]
\[ H_A : \mu > $3100 \]

p-value: proportion of simulations where the simulated sample mean is at least as extreme as the one observed.
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.

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

We want to investigate if people’s behavior is different on Friday 13\textsuperscript{th} compared to Friday 6\textsuperscript{th}.

One approach is to compare the traffic flow on these two days.

$H_0$: Average traffic flow on Friday 6\textsuperscript{th} and 13\textsuperscript{th} are equal.

$H_A$: Average traffic flow on Friday 6\textsuperscript{th} and 13\textsuperscript{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\textsuperscript{th} and the other Friday 13\textsuperscript{th}. Are these two counts independent?
Inference for Means with Small Samples

Conditions

\[
H_0 : \mu_{\text{diff}} = 0 \\
H_A : \mu_{\text{diff}} \neq 0
\]

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

\[
\begin{array}{c|ccccc}
\text{Difference in traffic flow} & 0 & 1000 & 2000 & 3000 & 4000 & 5000 \\
\hline
\text{frequency} & 1 & 2 & 5 & 2 & 1 & 0 \\
\end{array}
\]

\( n < 30 \) - the sample distribution does not appear to be extremely skewed, but it’s very difficult to assess with such a small sample size.

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

We can use simulation, but there is also a theoretical approach we can use when working with small sample means.
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 ensures that... we can use the CLT and $s$ is a reasonable approximation for $\sigma$.

- the sampling distribution of the mean is nearly normal with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$
- the estimate of the standard error using $\frac{s}{\sqrt{n}}$ is reliable
The normality condition

- The CLT, states that sampling distributions will be nearly normal, holds true for \textit{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?
The $t$ distribution

- When working with small samples, and the population standard deviation is unknown (this is almost always the case), the uncertainty of the standard error estimate is addressed by using a new distribution - the $t$ distribution.

- This distribution also is also 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 helpful for resolving our problem with a less reliable estimate the standard error (since $n$ is small)
History of the $t$ distribution

Discovered by William Gosset ...

- Oxford Graduate with a degree in Chemistry and Mathematics
- Hired as a brewer by the Guinness Brewery in 1999
- Spent 1906 - 1907 studying with Karl Pearson
- Published “The probable error of a mean” in 1908 under the pseudonym “Student”
- Much of his work was promoted by R.A. Fisher
Properties of the $t$ distribution

The $t$ distribution ... 

- is always centered at zero, like the standard normal ($z$) distribution.
- has a single parameter: *degrees of freedom* ($df$).

![Graph showing the $t$ distribution with different degrees of freedom](image)

- as $df$ increases the $t$ distribution approaches normal.
Inference for Means with Small Samples
Evaluating hypotheses using the \( t \) distribution

Back to Friday the 13\textsuperscript{th}

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

\[ \bar{x}_{diff} = 1836 \]
\[ s_{diff} = 1176 \]
\[ n = 10 \]
Finding the test statistic

Test statistic for inference on a small sample mean

The test statistic for inference on a small sample (n < 30) mean is the $T$ statistic with $df = n - 1$.

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

_In context..._

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

$$SE = \frac{s_{\text{diff}}}{\sqrt{n}} = \frac{1176}{\sqrt{10}} = 372$$

$$T = \frac{1836 - 0}{372} = 4.94$$

$$df = 10 - 1 = 9$$

Null value is 0 because in the null hypothesis we set $\mu_{\text{diff}} = 0$. 
Finding the p-value

- The p-value is, once again, calculated as the area tail area under the $t$ distribution.
- Using R:
  ```r
  > 2 * pt(4.94, df = 9, lower.tail = FALSE)
  [1] 0.0008022394
  ```
- Using a web applet: [http://www.socr.ucla.edu/htmls/SOCR_Distributions.html](http://www.socr.ucla.edu/htmls/SOCR_Distributions.html)
- Or when these aren’t available, we can use a $t$ table.
# Finding the p-value

Locate the calculated $T$ statistic on the appropriate $df$ row, obtain the p-value from the corresponding column heading (one or two tail, depending on the alternative hypothesis).

<table>
<thead>
<tr>
<th>df</th>
<th>0.100</th>
<th>0.050</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.08</td>
<td>6.31</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
</tr>
<tr>
<td>2</td>
<td>1.89</td>
<td>2.92</td>
<td>4.30</td>
<td>6.96</td>
<td>9.92</td>
</tr>
<tr>
<td>3</td>
<td>1.64</td>
<td>2.35</td>
<td>3.18</td>
<td>4.54</td>
<td>5.84</td>
</tr>
<tr>
<td>17</td>
<td>1.33</td>
<td>1.74</td>
<td>2.11</td>
<td>2.57</td>
<td>2.90</td>
</tr>
<tr>
<td>18</td>
<td>1.33</td>
<td>1.73</td>
<td>2.10</td>
<td>2.55</td>
<td>2.88</td>
</tr>
<tr>
<td>19</td>
<td>1.33</td>
<td>1.73</td>
<td>2.09</td>
<td>2.54</td>
<td>2.86</td>
</tr>
<tr>
<td>20</td>
<td>1.33</td>
<td>1.72</td>
<td>2.09</td>
<td>2.53</td>
<td>2.85</td>
</tr>
<tr>
<td>400</td>
<td>1.28</td>
<td>1.65</td>
<td>1.97</td>
<td>2.34</td>
<td>2.59</td>
</tr>
<tr>
<td>500</td>
<td>1.28</td>
<td>1.65</td>
<td>1.96</td>
<td>2.33</td>
<td>2.59</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1.28</td>
<td>1.64</td>
<td>1.96</td>
<td>2.33</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Finding the p-value (cont.)

<table>
<thead>
<tr>
<th>df</th>
<th>0.100</th>
<th>0.050</th>
<th>0.025</th>
<th>0.010</th>
<th>0.005</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1.44</td>
<td>1.94</td>
<td>2.45</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>7</td>
<td>1.41</td>
<td>1.89</td>
<td>2.36</td>
<td>3.00</td>
<td>3.50</td>
</tr>
<tr>
<td>8</td>
<td>1.40</td>
<td>1.86</td>
<td>2.31</td>
<td>2.90</td>
<td>3.36</td>
</tr>
<tr>
<td>9</td>
<td>1.38</td>
<td>1.83</td>
<td>2.26</td>
<td>2.82</td>
<td>3.25</td>
</tr>
<tr>
<td>10</td>
<td>1.37</td>
<td>1.81</td>
<td>2.23</td>
<td>2.76</td>
<td>3.17</td>
</tr>
</tbody>
</table>

\[ T = 4.94; \ p-value < 0.005 \]

What is the conclusion of the hypothesis test?
What is the difference?

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

- 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 (and not a \( z \) distribution), the critical value is a \( t^* \) (as opposed to a \( z^* \)).

  \[ \text{point estimate} \pm t^*_d \times SE \]
Finding the critical $t$ ($t^*$)

$n = 10$, $df = 10 - 1 = 9$

t* is at the intersection of row $df = 9$ and two tail probability 0.05.
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?

$$\bar{x}_{\text{diff}} = 1836 \quad s_{\text{diff}} = 1176 \quad n = 10 \quad SE = 372$$
Interpreting the CI

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

\[ \mu_{\text{diff} \ 6\text{th} - 13\text{th}} = (995, 2677) \]

We are 95% confident that ...

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

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

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

(d) on Friday 13\text{th} there are 995 to 2,677 fewer cars on the road than on the Friday 6\text{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 13\textsuperscript{th} is a day of bad luck?
Recap: Inference using a small sample mean

- If $n < 30$, sample means follow a $t$ 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)
  - $n < 30$ and no extreme skew
- Hypothesis testing:
  $$T_{df} = \frac{\text{point estimate} - \text{null value}}{SE}, \text{ where } df = n - 1$$
- Confidence interval:
  $$\text{point estimate} \pm t^*_{df} \times SE$$

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, therefore the mechanics are the same as when we are working with just one sample. We'll discuss this type of problem in more detail after the midterm.