t Tests about a Population Mean

Previously:

μ is the mean of a variable, across all observational units in the population.

Our hypotheses are of the form:

H_0: $\mu = \mu_0$

H_A: $\mu \neq \mu_0$ (or $\mu < \mu_0$, or $\mu > \mu_0$)

Two options for how we can think about the sample statistic and what its sampling distribution would be if the null hypothesis was true (relevant for calculating p-values):

1. $\bar{Y} \sim \text{Normal}(\mu_0, \sigma/\sqrt{n})$
2. $z = \frac{\bar{Y}-\mu_0}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$

Note:

- σ is the standard deviation of values in the population
- $\sigma/\sqrt{n}$ is the standard deviation of values you could get for the sample mean based on a sample of size n.

Problem:

We usually don't know the population standard deviation σ

- We can't calculate the standard deviation of the sampling distribution for $\bar{Y}$, $\sigma/\sqrt{n}$
- We can't calculate $z = \frac{\bar{y}-\mu_0}{\sigma/\sqrt{n}}$

Solution:

Estimate the population standard deviation with the sample standard deviation.

New test statistic (final answer):

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

,

where $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$ is the standard deviation of the data in the sample.

Definition:

- A **standard error** is an estimate of the standard deviation of something.
  - The **true** standard deviation of $\bar{Y}$ is $\sigma/\sqrt{n}$, but we don't know σ
  - The **estimated** standard deviation (i.e., standard error) of $\bar{Y}$ is $s/\sqrt{n}$

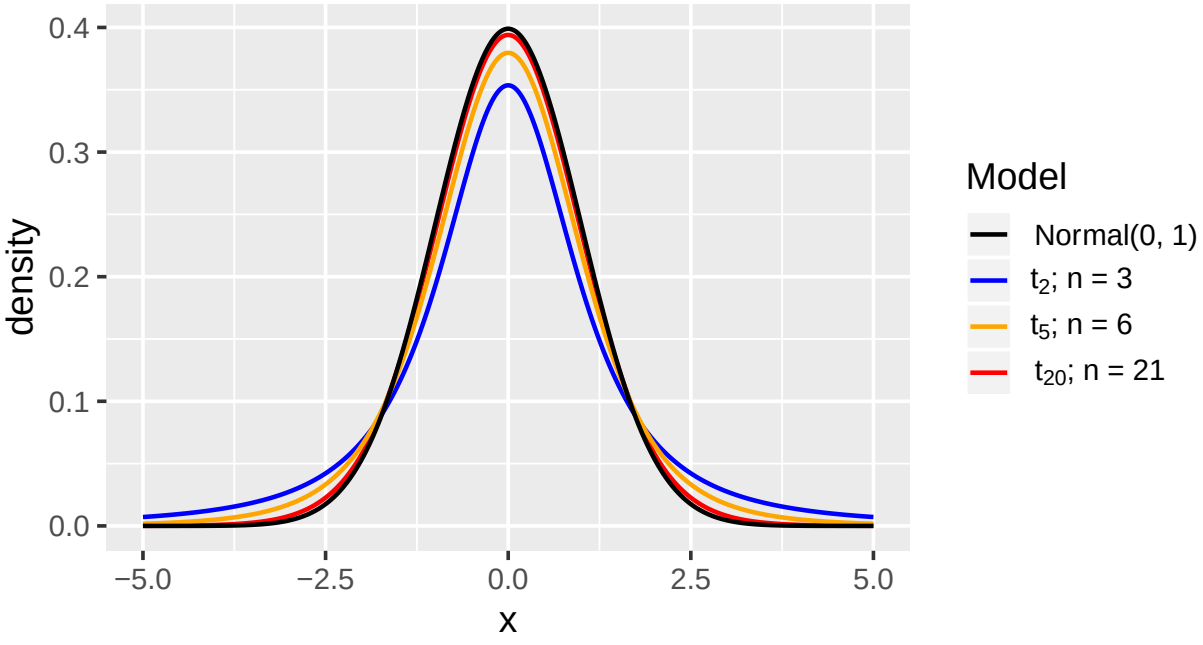
Interpretation of t Statistic:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

- How many standard errors away from the hypothesized population mean μ_0 is our sample mean $\bar{y}$?
- A larger difference between μ_0 and $\bar{y}$ casts more doubt on the null hypothesis.
- But distance between μ_0 and $\bar{y}$ is measured in units of (estimated) standard deviations of $\bar{y}$.

Facts:

- $t \sim t_{n-1}$
- Read as "t follows a t distribution with $n-1$ degrees of freedom"
- The degrees of freedom of $n-1$ matches the denominator of $n-1$ in the sample standard deviation
- The t distribution is similar to the Normal$(0,1)$, but t has more probability in the tails (estimating σ with s introduces more variability)
- As n increases, the t becomes more like a Normal$(0,1)$



Example 1: Friday the 13th

The *British Medical Journal* published an article titled "Is Friday the 13th Bad for Your Health?" The article examined the number of people admitted to emergency rooms for vehicular accidents on 12 Friday evenings (6 each on the 6th and 13th of a given month). The following R code reads the data in:

```
## Parsed with column specification:
## cols(
##   `Year and Month` = col_character(),
##   `6th` = col_integer(),
##   `13th` = col_integer()
## )
```

```
## # A tibble: 6 x 3
##   year_month accidents_6th accidents_13th
##   <chr>              <int>          <int>
## 1 Oct-89                 9             13
## 2 Jul-90                 6             12
## 3 Sep-91                11             14
## 4 Dec-91                11             10
## 5 Mar-92                 3              4
## 6 Nov-92                 5             12
```

```
## [1] 6 3
```

Let's treat these as **paired data**: there may be some connection between the number of accidents on Friday the 13th in a particular month and the number of accidents that occurred one week earlier (for example, maybe it was a winter month with bad weather). We will therefore consider the differences in the number of observed accidents between the 13th and the 6th of a given month:

```
## # A tibble: 6 x 4
##   year_month accidents_6th accidents_13th accidents_difference
##   <chr>              <int>          <int>                <int>
## 1 Oct-89                 9             13                    4
## 2 Jul-90                 6             12                    6
## 3 Sep-91                11             14                    3
## 4 Dec-91                11             10                   -1
## 5 Mar-92                 3              4                    1
## 6 Nov-92                 5             12                    7
```

Step 1: Identify population parameter of interest

Step 2: State hypotheses

- **Null Hypothesis** (H_0):

- **Alternative Hypothesis** (H_A):

Step 3: Identify the sample statistic and the sampling distribution of the sample statistic, assuming H_0 is true. Check all necessary conditions.

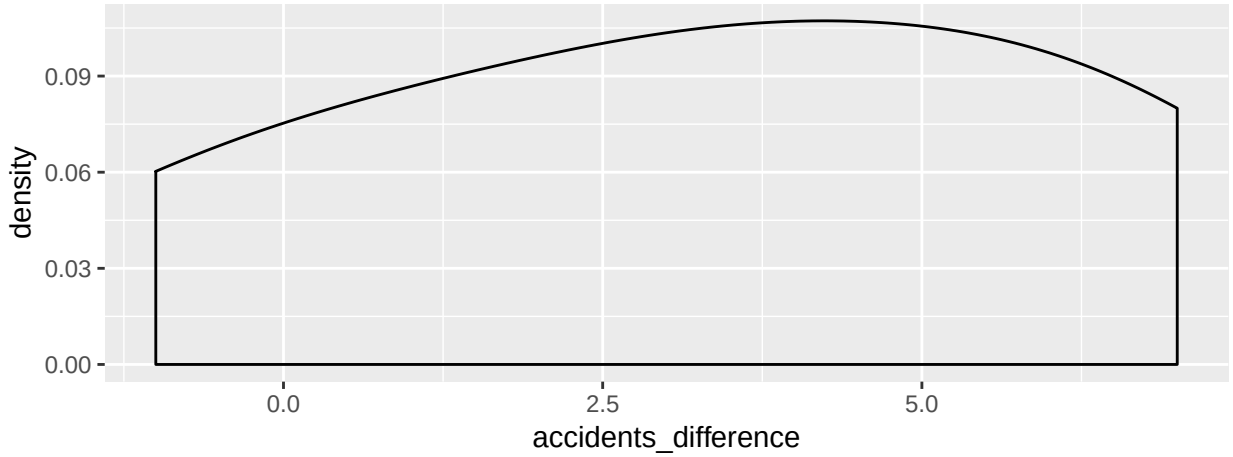
Check conditions:

- Were there any forms of bias in our sampling procedure?

- Was a quantitative variable recorded for each observational unit?

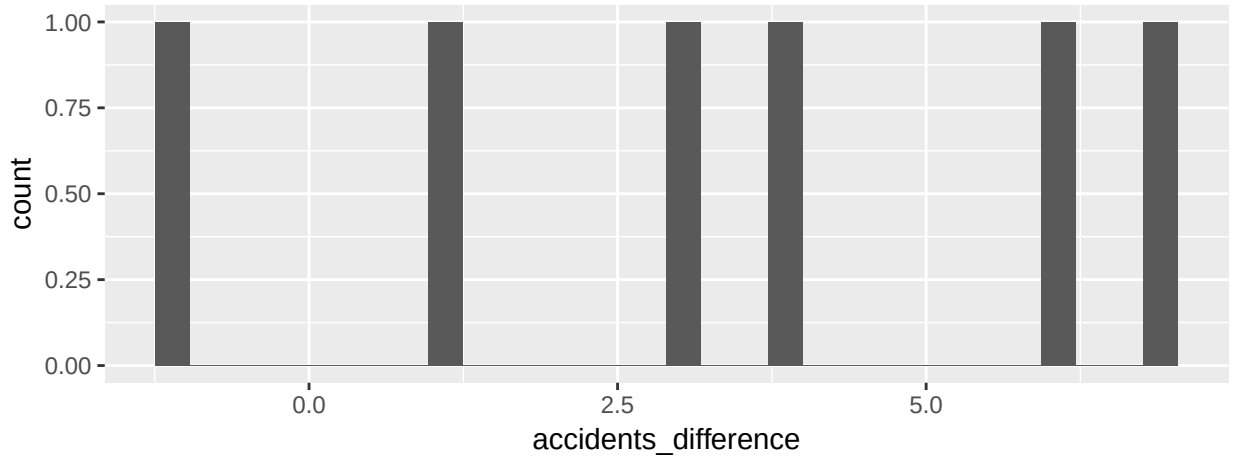
- Is the mean a reasonable summary of the center, and is the sample size large enough that the Central Limit Theorem applies? (See output on next page.)

```
ggplot(data = fridays, mapping = aes(x = accidents_difference)) +
    geom_density()
```




```
ggplot(data = fridays, mapping = aes(x = accidents_difference)) +
    geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
fridays %>%
  summarize(
    mean = mean(accidents_difference),
    median = median(accidents_difference)
  )
```

```
## # A tibble: 1 x 2
##        mean median
##       <dbl>  <dbl>
## 1 3.3333333    3.5
```

- Are different observational units in our sample independent?

State the test statistic and its sampling distribution, assuming H_0 is true:

Step 4: Calculate the p-value for the test

```
fridays %>%
    summarize(
        mean_diff = mean(accidents_difference),
        sd_diff = sd(accidents_difference)
    )
```

```
## # A tibble: 1 x 2
##   mean_diff   sd_diff
##       <dbl>     <dbl>
## 1 3.3333333 3.0110906
```

```
(3.333333 - 0)/(3.011091 / sqrt(6))
```

```
## [1] 2.71163
```

```
pt(2.71163, df = 5, lower.tail = FALSE)
```

```
## [1] 0.02109702
```

...or...

```
pt(-2.71163, df = 5)
```

```
## [1] 0.02109702
```

...or...

```r
t.test(~ accidents_difference, data = fridays, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  accidents_difference
## t = 2.7116, df = 5, p-value = 0.0211
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  0.8562896       Inf
## sample estimates:
## mean of x
##  3.333333
```

Step 5: Draw a conclusion

What is the strength of evidence against the null hypothesis provided by the data?

Make a decision using a significance level of $\alpha = 0.05$.