Confidence Intervals for Population Means

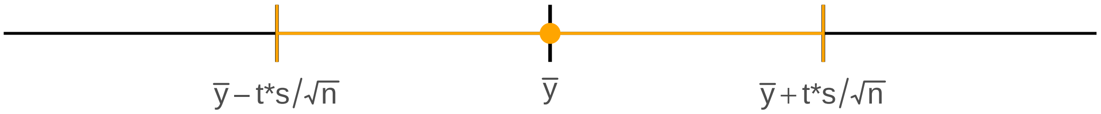
Reminder of Notation: $\bar{y}$ is the sample mean, s is the sample standard deviation, n is the sample size

Goal: A $(1-\alpha) \times 100$ % confidence interval for μ.

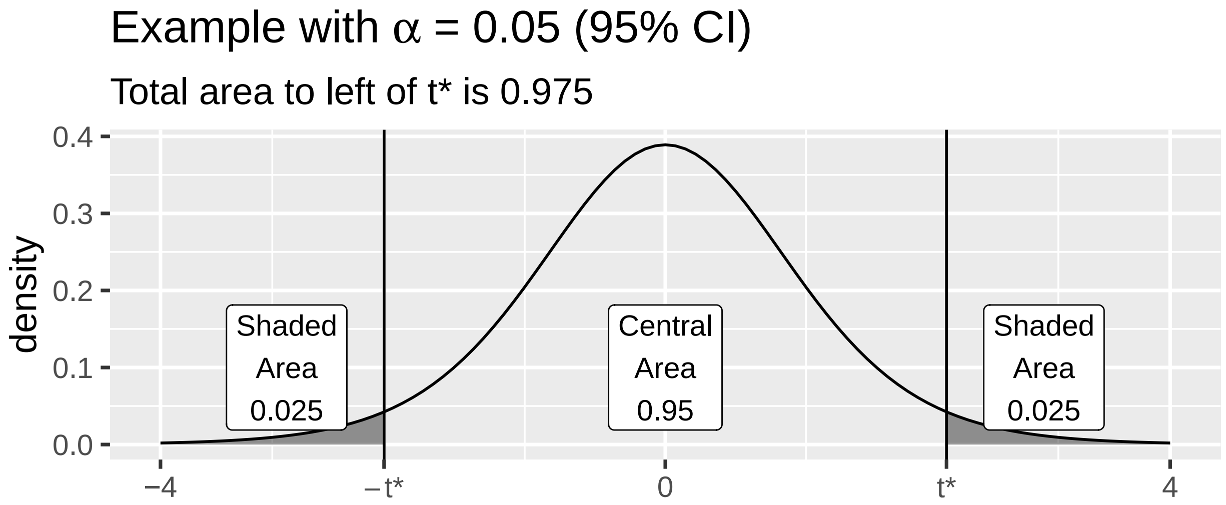
Example: $\alpha = 0.05 \Leftrightarrow 95\%$ confidence interval.

Interval: $\bar{y} \pm t^* SE(\bar{y})$

$SE(\bar{y}) = \frac{s}{\sqrt{n}}$ is the *standard error* (estimated standard deviation) of $\bar{y}$



- The **margin of error** is $t^* SE(\bar{y})$: the amount we add and subtract from $\bar{y}$.
- t^* is the $(1-\frac{\alpha}{2})$ quantile of the t_{n-1} distribution. Examples:
  - $\alpha = 0.05 \Leftrightarrow 95\%$ CI $\Leftrightarrow t^* = 0.975$th Quantile of t_{n-1} distribution.



In R, to look up t^*:

```r
qt(0.975, df = 10) # For a 95% CI, sample size is n = 11
```

```
## [1] 2.228139
```

Important things:

- For a 95% CI, the first argument to `qt` is 0.975, not 0.95!
- The second argument to `qt` is $n-1$!

This example was presented in Rice (2007):

> To study the effect of cigarette smoking on platelet aggregation, Levine (1973) drew blood samples from 11 individuals before and after they smoked a cigarette and measured the extent to which the blood platelets aggregated. Platelets are involved in the formation of blood clots, and it is known that smokers suffer more often from disorders involving blood clots than do nonsmokers. The data are shown in the following table, which gives the maximum percentage of all the platelets that aggregated afer being exposed to a stimulus.

```r
head(platelets)
```

```
## # A tibble: 6 x 2
##   before after
##    <int> <int>
## 1     25    27
## 2     25    29
## 3     27    37
## 4     44    56
## 5     30    46
## 6     67    82
```

This is an example of **paired data**:

- We have two measurements on each person (these are **not independent**!)
- We are interested in the **difference** between these measurements
- These **differences are independent** across different people

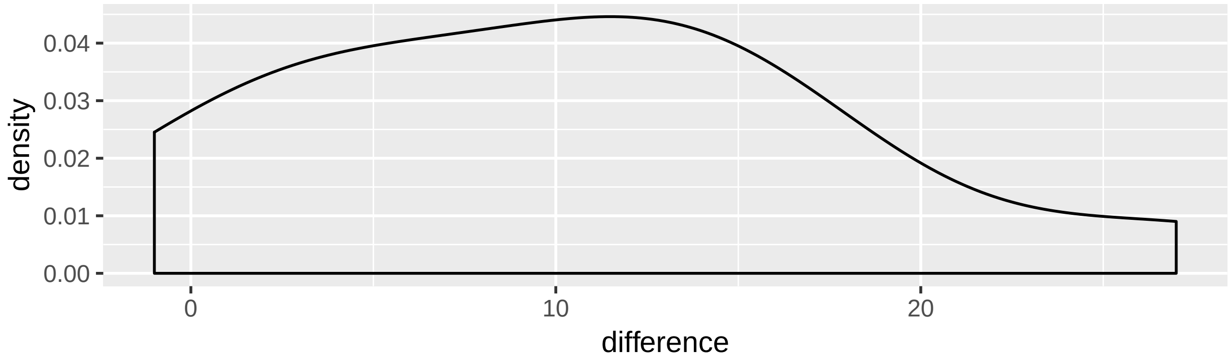
```r
platelets <- platelets %>%
  mutate(difference = after - before)
head(platelets)
```

```
## # A tibble: 6 x 3
##   before after difference
##    <int> <int>      <int>
## 1     25    27          2
## 2     25    29          4
## 3     27    37         10
## 4     44    56         12
## 5     30    46         16
## 6     67    82         15
```

(a) What is the population parameter of interest?

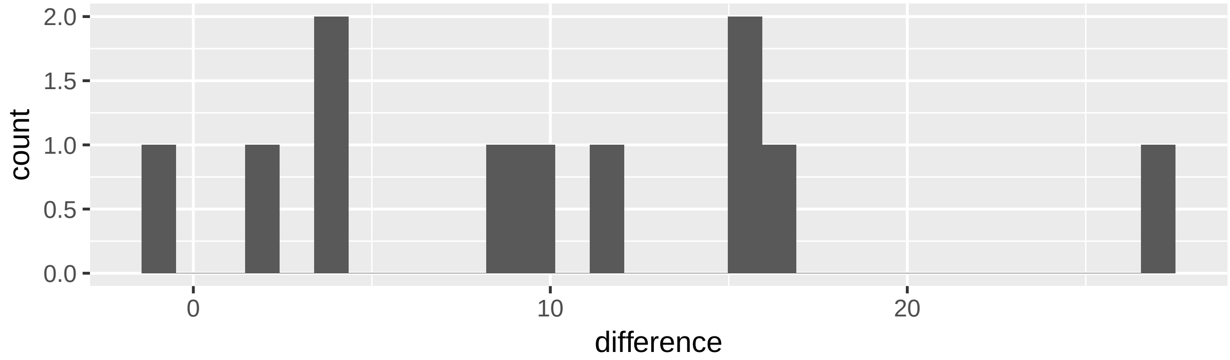
(b) Check the conditions for inference with these data

```
ggplot(data = platelets, mapping = aes(x = difference)) +
  geom_density()
```



```
ggplot(data = platelets, mapping = aes(x = difference)) +
  geom_histogram()
```

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



```
platelets %>%
  summarize(
    mean_diff = mean(difference),
    median_diff = median(difference)
  )
```

```
## # A tibble: 1 x 2
##   mean_diff median_diff
##       <dbl>       <int>
## 1 10.272727          10
```

(c) Find a 95% confidence interval for the population parameter, and verify that it agrees with the output from `t.test`.

```r
nrow(platelets)
```

```
## [1] 11
```

```r
platelets %>%
  summarize(
    mean_difference = mean(difference),
    sd_difference = sd(difference))
```

```
## # A tibble: 1 x 2
##   mean_difference sd_difference
##             <dbl>         <dbl>
## 1       10.272727     7.9761007
```

```r
qt(0.975, df = 11 - 1)
```

```
## [1] 2.228139
```

```r
10.27273 -  2.228139 * 7.976101 / sqrt(11)
```

```
## [1] 4.914312
```

```r
10.27273 +  2.228139 * 7.976101 / sqrt(11)
```

```
## [1] 15.63115
```

```r
t.test(~ difference, conf.level = 0.95, alternative = "two.sided",
  data = platelets)
```

```
##
##  One Sample t-test
##
## data:  difference
## t = 4.2716, df = 10, p-value = 0.001633
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   4.91431 15.63114
## sample estimates:
## mean of x
##  10.27273
```

(d) How much did that potential outlier affect our inference?

- Definition (from Wikipedia): “an **influential observation** is an observation for a statistical calculation whose deletion from the dataset would noticeably change the result of the calculation.”

Let’s find out what happens if we remove the outlier.

```r
platelets_no_outlier <- platelets %>% filter(difference < 20)

t.test(~ difference, conf.level = 0.95, alternative = "two.sided",
  data = platelets_no_outlier)
```

```
##
##  One Sample t-test
##
## data:  difference
## t = 4.5021, df = 9, p-value = 0.001484
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   4.278813 12.921187
## sample estimates:
## mean of x
##       8.6
```

- How much does the confidence interval change when that observation is removed?

- Does the conclusion of the analysis change when that observation is removed?

Note:

- I am not removing the outlier from the analysis.
- I am checking to see how sensitive my conclusions are to that one data point.
- If anything, I would discuss results from both tests (e.g., “there was one mild outlier in the data set, but the substantive conclusion of the analysis was the same whether or not that outlier was included.”)

(e) Find a 99% confidence interval for the population parameter, and verify that it agrees with the output from `t.test`.

```
nrow(platelets)
```

```
## [1] 11
```

```
platelets %>%
  summarize(
    mean_difference = mean(difference),
    sd_difference = sd(difference))
```

```
## # A tibble: 1 x 2
##   mean_difference sd_difference
##             <dbl>         <dbl>
## 1       10.272727     7.9761007
```

```
qt(0.995, df = 11 - 1)
```

```
## [1] 3.169273
```

```
10.27273 -  3.169273 * 7.976101 / sqrt(11)
```

```
## [1] 2.650993
```

```
10.27273 +  3.169273 * 7.976101 / sqrt(11)
```

```
## [1] 17.89447
```

```
t.test(~ difference, conf.level = 0.99, alternative = "two.sided",
  data = platelets)
```

```
##
##  One Sample t-test
##
## data:  difference
## t = 4.2716, df = 10, p-value = 0.001633
## alternative hypothesis: true mean is not equal to 0
## 99 percent confidence interval:
##   2.650991 17.894463
## sample estimates:
## mean of x
##  10.27273
```