Two-sample inference

Prof. Maria Tackett

Click for PDF of slides

Recap

So far, we've talked about performing interval estimation and hypothesis testing for means using

- simulation-based methods, such as bootstrap or direct simulation, and
- the Central Limit Theorem

In all cases so far, we've only compared one sample against a hypothesized value.

But what if we wanted to compare two samples against *each other*?

Two-sample inference for means

Suppose we have two (representative) samples, and wanted to either

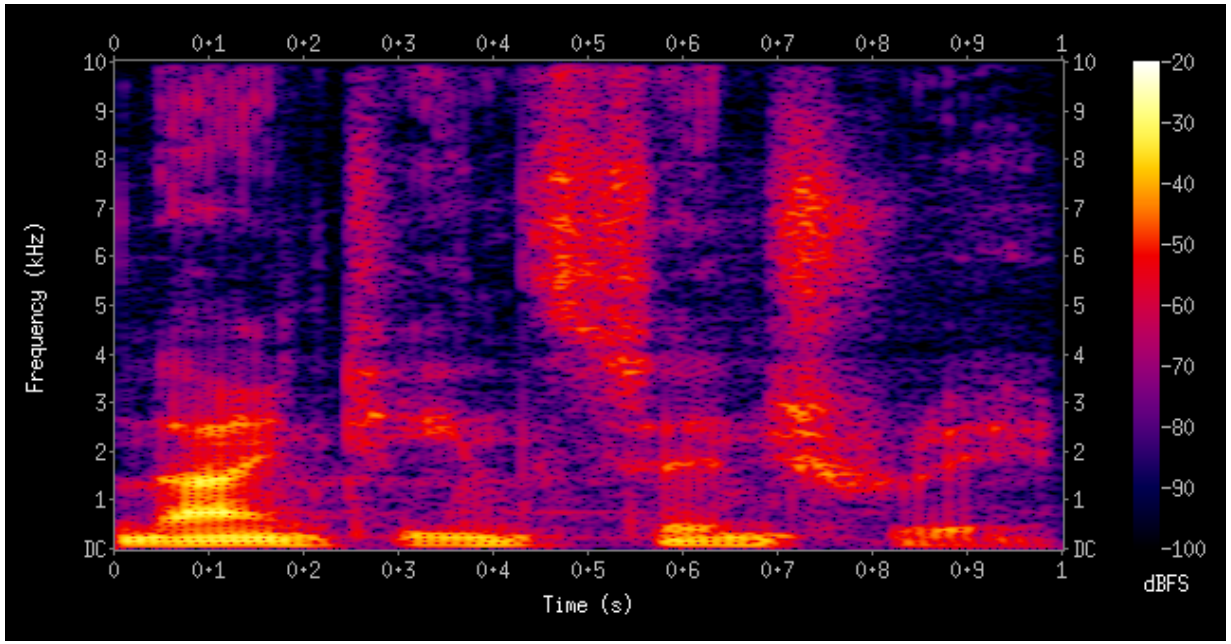
- estimate the difference in means in the two populations
  - confidence interval for $\mu_1 - \mu_2$
- Test the hypotheses

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 \neq \mu_2,$$

where μ_1 and μ_2 are the population means in groups 1 and 2.

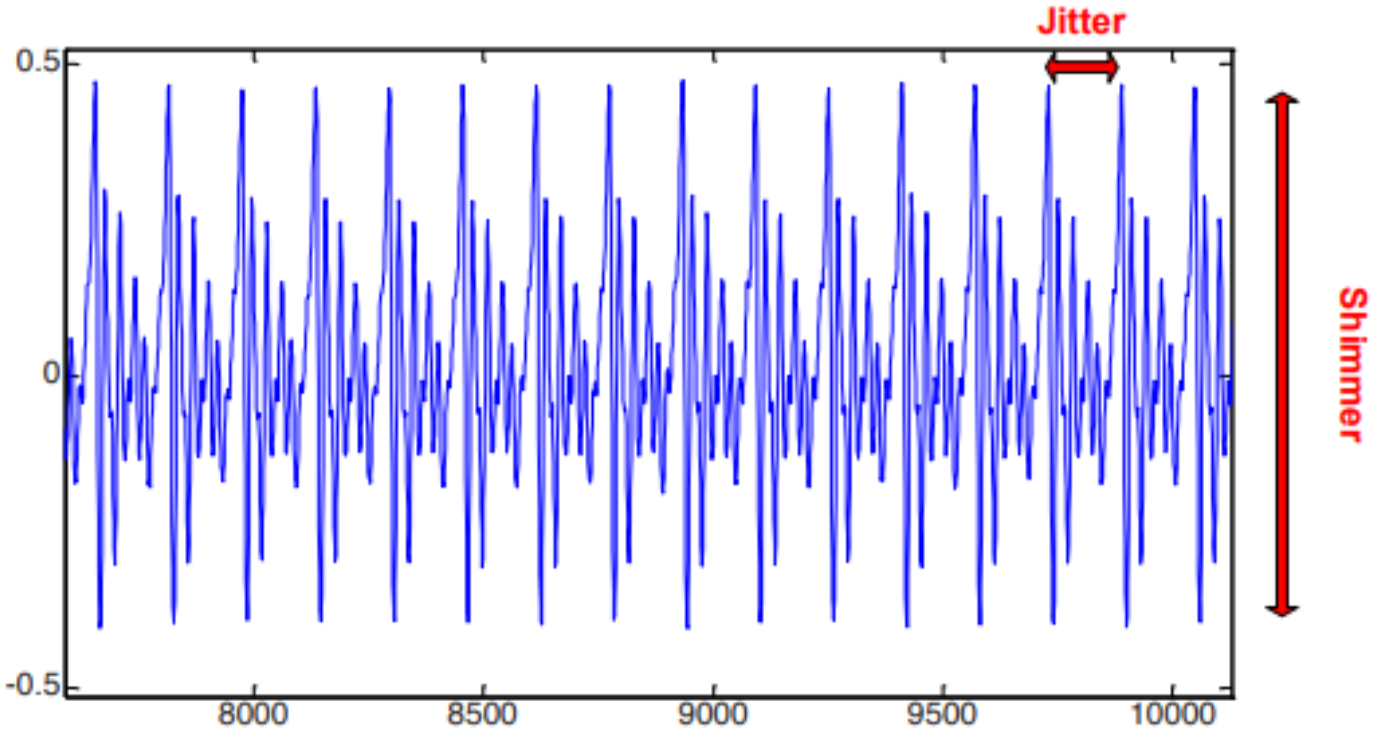
How might you calculate a confidence interval and address the above hypothesis test using simulation-based methods? How about the CLT?

Today's data



Adapted from Erdogdu Sakar, B., et al. *Collection and Analysis of a Parkinson Speech Dataset with Multiple Types of Sound Recordings*, IEEE Journal of Biomedical and Health Informatics, vol. 17(4), pp. 828-834, 2013 (image from **Wikipedia**)

Some voice analysis terminology



- Jitter: frequency variation from cycle to cycle
- Shimmer: amplitude variation of the sound wave

Question of interest

Is there a difference in average voice jitter between patients with Parkinson's disease (PD) and those who don't have Parkinson's disease (control group)?

`parkinsons.csv` contains repeated voice recordings from a number of patients, some with PD and some serving as non-PD controls (Erdogdu B et al.). For now, **assume that all samples were taken independently from each other** (this is not actually the case, but we'll make this assumption).

Jitter is given in milliseconds (ms), and shimmer is given in decibels (dB).

Bootstrap estimation

Let's construct the bootstrap distribution for the **difference in means**.

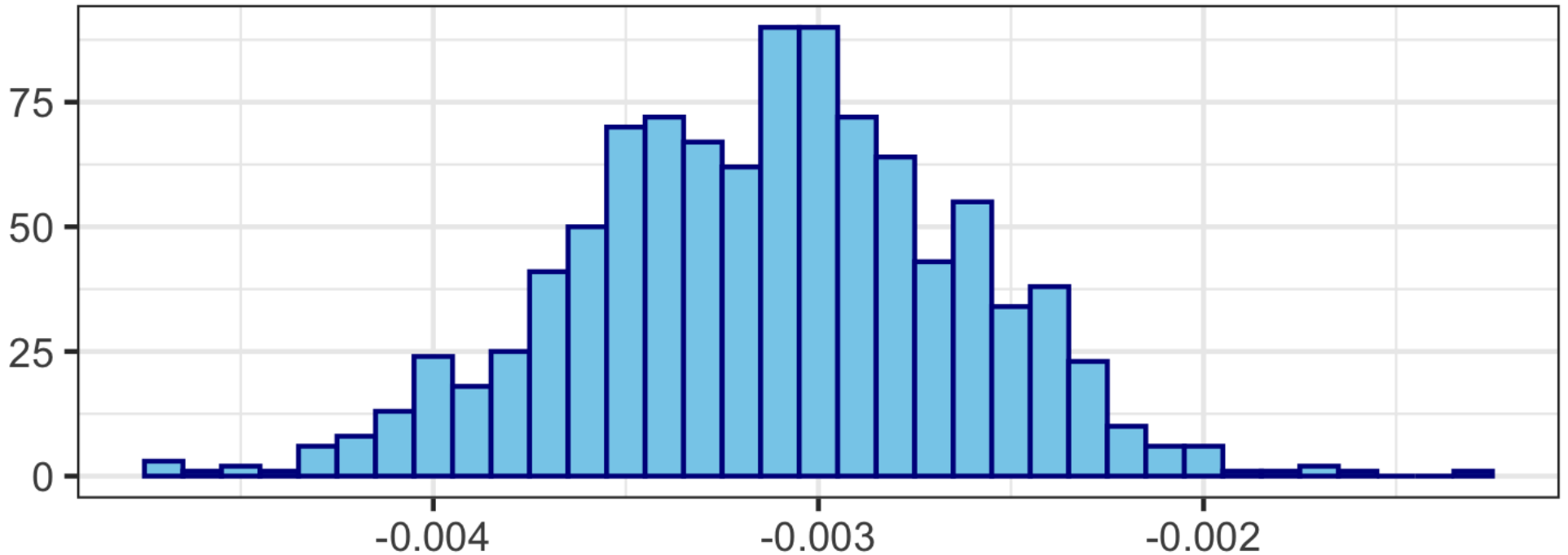
```
set.seed(2020)
parkinsons <- read_csv("data/parkinsons.csv")

library(infer)

boot_diffs <- parkinsons %>%
  specify(jitter ~ status) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means",
            order = c("Healthy", "PD"))
```

Bootstrap estimation

Let's construct the bootstrap distribution for the *difference* in means.



CI for difference in means

Let's construct the bootstrap distribution for the **difference in means**.

```
boot_diffs %>%
  summarize(lower = quantile(stat, 0.025),
            upper = quantile(stat, 0.975))
```

```
## # A tibble: 1 x 2
##       lower    upper
##       <dbl>    <dbl>
## 1 -0.00413 -0.00220
```

CI for difference in means

```
## # A tibble: 1 x 2
##       lower    upper
##       <dbl>    <dbl>
## 1 -0.00413 -0.00220
```

Interpretation: We are 95% confident that the mean voice jitter for people without Parkinson's disease is about 0.002 to 0.004 ms less than the mean voice jitter for those with Parkinson's disease.

CI for difference in means

```
## # A tibble: 1 x 2
##       lower    upper
##       <dbl>    <dbl>
## 1 -0.00413 -0.00220
```

Interpretation: We are 95% confident that the mean voice jitter for people without Parkinson's disease is about 0.002 to 0.004 ms less than the mean voice jitter for those with Parkinson's disease.

Is there evidence that there is a difference in mean voice jitter between PD patients and healthy patients?

Hypothesis testing

Let μ_P be the mean voice jitter among PD patients, and μ_H be the mean voice jitter among healthy patients. Let's test

$$H_0 : \mu_P = \mu_H$$
$$H_a : \mu_P \neq \mu_H$$

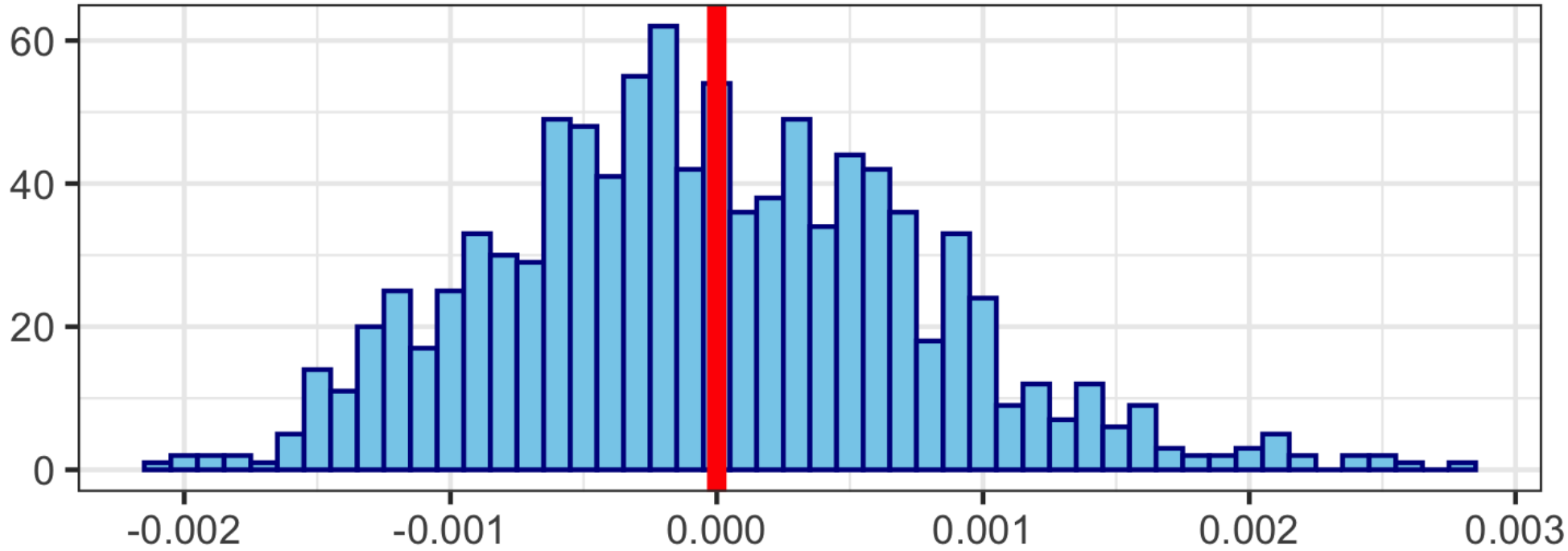
If the two means are truly equal (i.e., if H_0 is true), then the difference, $\mu_H - \mu_P$, should be **zero**.

Hypothesis testing

Let's construct the simulated **null distribution** for the difference in means, $\mu_H - \mu_P$. If the two means are truly equal (i.e., if H_0 is true), then this difference should be zero.

```
null_dist <- parkinsons %>%
  specify(jitter ~ status) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means",
            order = c("Healthy", "PD"))
```

Hypothesis testing



Hypothesis testing

```
obs_diff <- parkinsons %>%
  specify(jitter ~ status) %>%
  calculate(stat = "diff in means", order = c("Healthy", "PD")) %>%
  pull()
obs_diff
```

```
## [1] -0.00312321
```

Hypothesis testing

```
obs_diff <- parkinsons %>%
  specify(jitter ~ status) %>%
  calculate(stat = "diff in means", order = c("Healthy", "PD")) %>%
  pull()
obs_diff
```

```
## [1] -0.00312321
```

```
null_dist %>%
  filter(abs(stat) >= abs(obs_diff)) %>%
  summarise(p_val = n() / nrow(null_dist))
```

```
## # A tibble: 1 x 1
##   p_val
##   <dbl>
## 1     0
```

Conclusion

The p-value is very small, so we reject H_0. The data provide sufficient evidence that there is a difference in the mean voice jitter between patients who have Parkinson's disease and those who don't have the disease.

Difference in means using CLT

Difference in means using CLT

CLT-based inference for a difference in means relies on the two-sample t-test for independent samples. Like the t-test we've seen before, the **test statistic** takes on the following form:

Difference in means using CLT

CLT-based inference for a difference in means relies on the two-sample t-test for independent samples. Like the t-test we've seen before, the **test statistic** takes on the following form:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\widehat{SE}_{diff}}$$

Difference in means using CLT

CLT-based inference for a difference in means relies on the two-sample t-test for independent samples. Like the t-test we've seen before, the **test statistic** takes on the following form:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\widehat{SE}_{diff}}$$

The test statistic depends on whether we can assume that the two groups have the same underlying variability in their observations.

Difference in means using CLT

CLT-based inference for a difference in means relies on the two-sample t-test for independent samples. Like the t-test we've seen before, the **test statistic** takes on the following form:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\widehat{SE}_{diff}}$$

The test statistic depends on whether we can assume that the two groups have the same underlying variability in their observations.

The exact form of the test statistic under the null hypothesis, including the degrees of freedom, are a complicated fraction that no one calculates by hand. Let's let R handle this!

CLT: Difference in means

```
parkinsons %>%
  t_test(jitter ~ status,
         mu = 0,
         order = c("Healthy", "PD"),
         alternative = "two-sided",
         conf_int = TRUE, conf_level = 0.95)
```

```
## # A tibble: 1 x 6
##   statistic  t_df      p_value alternative lower_ci upper_ci
##       <dbl> <dbl>        <dbl> <chr>          <dbl>    <dbl>
## 1     -5.96  187. 0.0000000124 two.sided   -0.00416 -0.00209
```

CLT: Difference in means

```
## # A tibble: 1 x 6
##   statistic  t_df      p_value alternative lower_ci upper_ci
##       <dbl> <dbl>        <dbl> <chr>           <dbl>    <dbl>
## 1     -5.96  187. 0.0000000124 two.sided   -0.00416 -0.00209
```

Comprehensively evaluate the research question by specifying the hypotheses, the test statistic and its the distribution under the null, the p-value, and decision at the $\alpha = 0.05$ significance level. Interpret the conclusions from your hypothesis test in context of the original research question.