Lecture 25: Paired T-tests

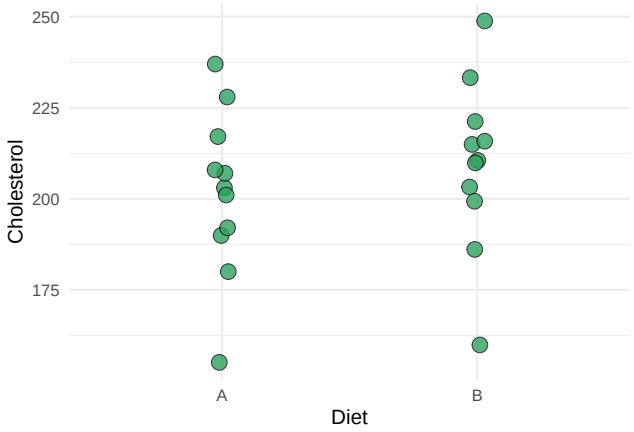
Part of Ch 17

Corinne Riddell (Instructor: Tomer Altman)

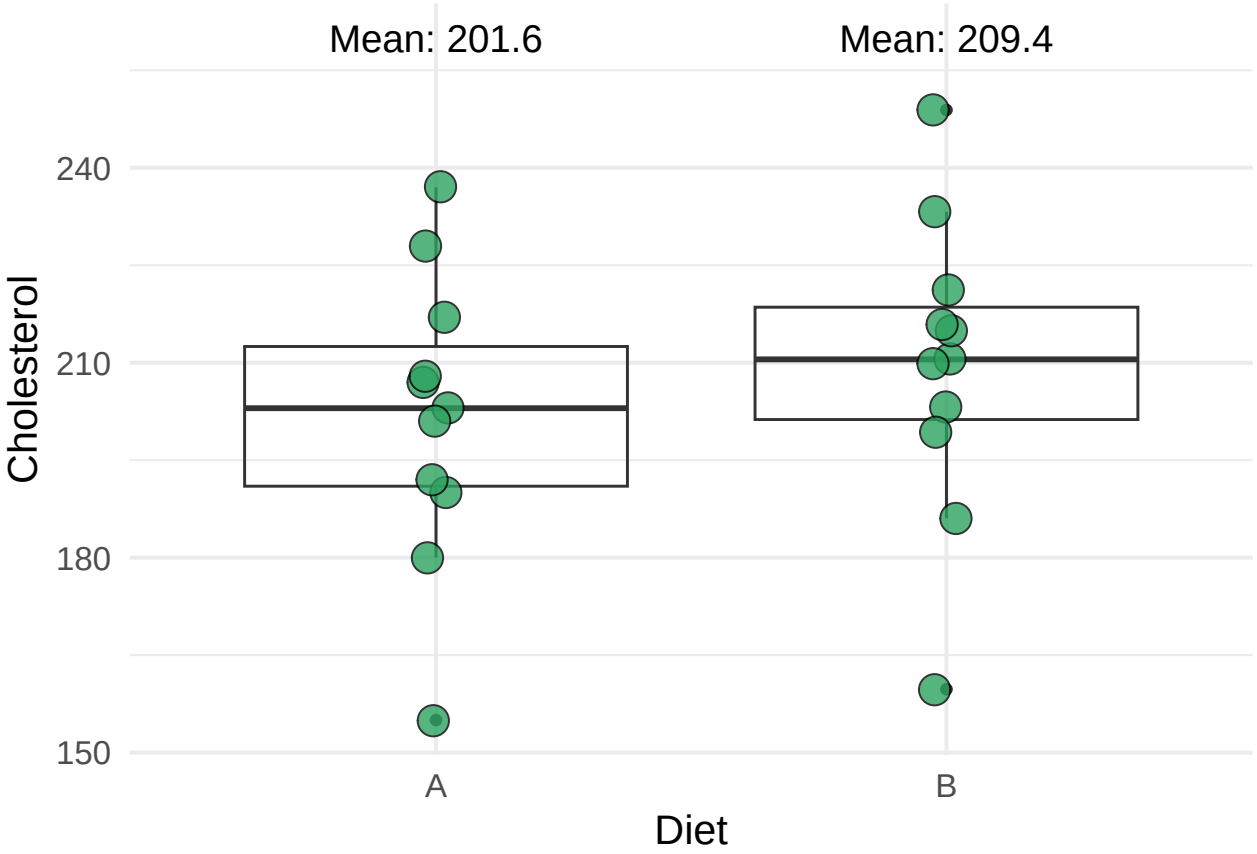
October 27, 2025

Motivation

Suppose you received the following graphic illustrating cholesterol measurements following two alternate diets. What do you think about these data?



Motivation



- What do you notice about the variability between participants under each diet?
- What is the mean difference?

Motivation

A two sample t-test assuming all measurements are independent reveals no evidence against the null hypothesis of no difference between the diets:

```
##
##  Two Sample t-test
##
## data:  chol_dat %>% pull(B) and chol_dat %>% pull(A)
## t = 0.78557, df = 20, p-value = 0.4413
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.77356  28.20653
## sample estimates:
## mean of x mean of y
##  209.3529  201.6364
```

. . . Oh my! I forgot to tell you that there were only eleven individuals in the study, and each of them tried out both diets. Let me update my visualization to reflect this a little bit better. . .

Motivation

Now, what do you notice about the paired data?

Check your understanding!

The paired t-test

- We use a paired t-test when the data is **matched by design**
  - Two observations on each participant/patient
- A paired t-test is a test of the mean **differences** for each individual. It only uses the variation within individuals.
- To perform a paired t-test, we actually perform a one-sample t-test of the differences
- The test statistic is: $t = \frac{\bar{x}_d - 0}{\frac{s_d}{\sqrt{n}}}$
- Here, $\bar{x}_d$ is the average of the differences (e.g., Diet B - Diet A) across the eleven individuals in the study, and s_d is the standard deviation of the differences
- It can be compared to a critical value from the t-distribution with $n - 1$ degrees of freedom, where n is the number of unique individuals in the study

Calculate the test statistic, p-value, and 95% confidence interval

- First let's have a look at the dataset as-is:

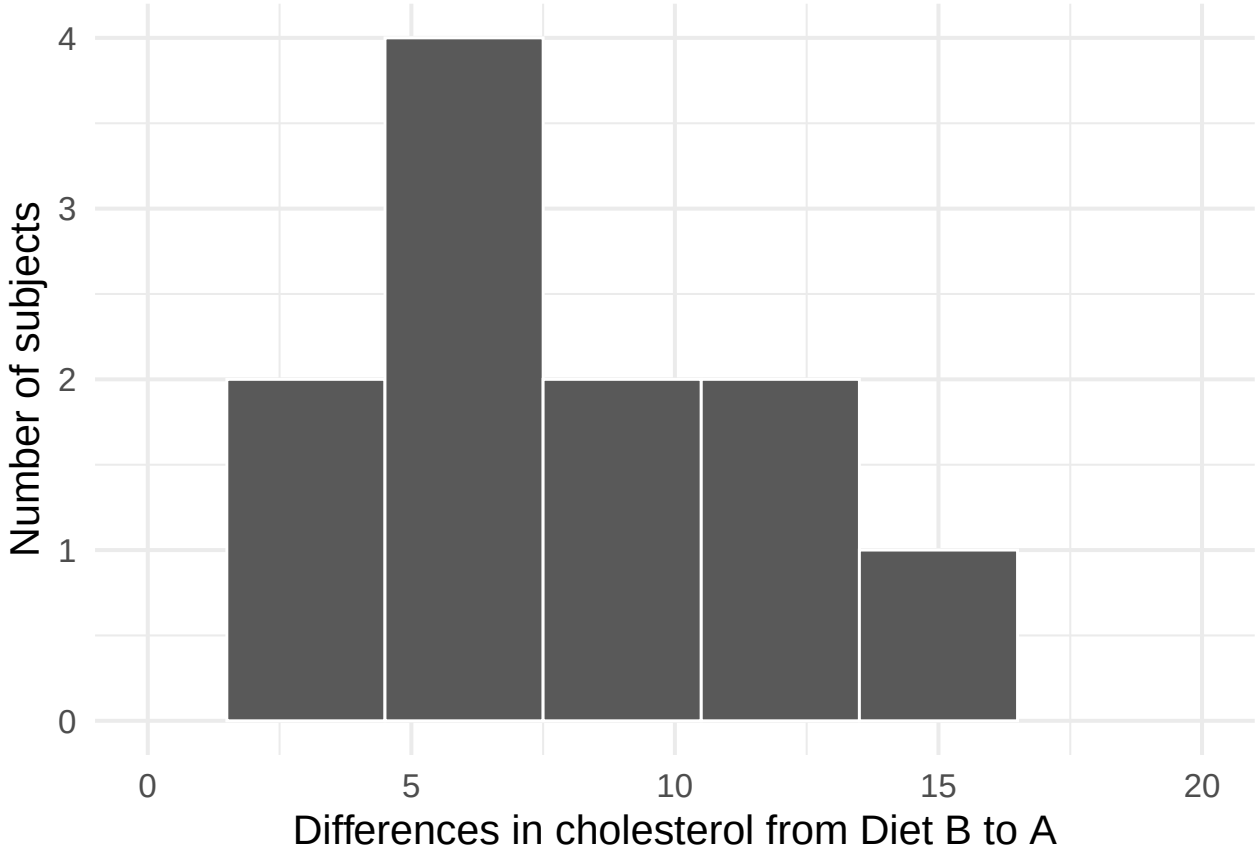
```
head(chol_dat)
```

```
##     A        B id
## 1 155 159.7581  1
## 2 180 186.0793  2
## 3 190 203.2348  3
## 4 192 199.2820  4
## 5 203 210.5172  5
## 6 201 214.8603  6
```

Calculate the test statistic, p-value, and 95% confidence interval

- We can use functions from the library `dplyr` to calculate the test statistic
- Use `mutate` to calculate each participant's difference:

```
##     A        B id      diff
## 1 155 159.7581  1  4.758097
## 2 180 186.0793  2  6.079290
## 3 190 203.2348  3 13.234833
## 4 192 199.2820  4  7.282034
## 5 203 210.5172  5  7.517151
## 6 201 214.8603  6 13.860260
```



Calculate the test statistic, p-value, and 95% confidence interval

- Then use `summarize` to calculate the mean difference ($\bar{x}_d$), its standard error ($\frac{s_d}{\sqrt{n}}$), and the t-statistic:

```
summary_stats <- chol_dat %>%
  summarize(mean_diff = mean(diff),  # mean difference
            std_err_diff = sd(diff)/sqrt(n()),  # SE of the mean
            t_stat = mean_diff/std_err_diff)   # test statistic

summary_stats
```

```
##   mean_diff std_err_diff   t_stat
## 1  7.716487     1.168587 6.603262
```

Sidebar: The `pull()` function

What does `pull()` do in R and how does it compare to `select()`?

Examine the structure of the output after `pull()`ing vs. `select()`ing `t_stat` from the `summary_stat` data frame:

```
summary_stats %>% pull(t_stat) %>% str()
```

```
##  num 6.6
```

```
summary_stats %>% select(t_stat) %>% str()
```

```
## 'data.frame':    1 obs. of  1 variable:
##  $ t_stat: num 6.6
```

Sidebar: The `pull()` function

Data frame containing many variables

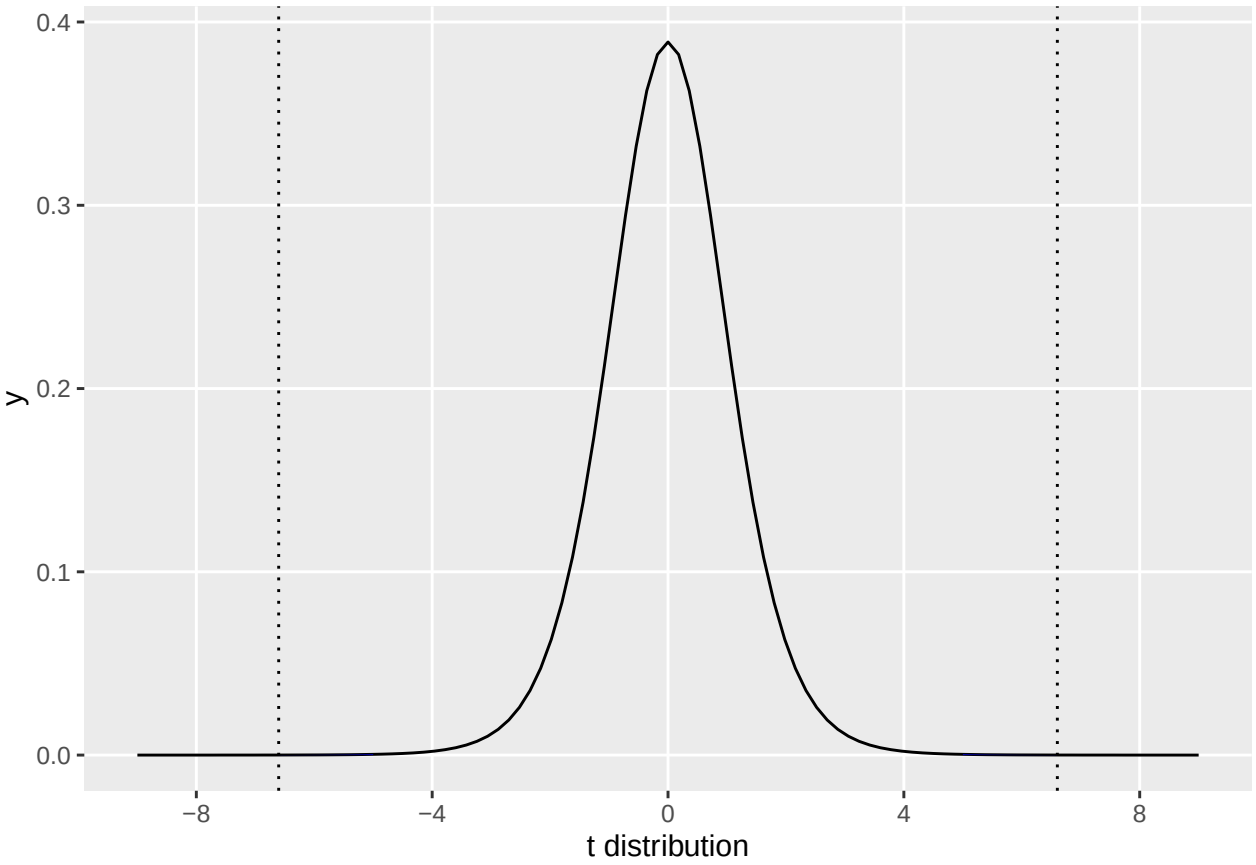
`dat %>% select(pepper)`

`dat %>% pull(pepper)`

Image credit and original idea: https://twitter.com/hadleywickham/status/643381054758363136?s=20

Calculate the test statistic, p-value, and 95% confidence interval

What is the probability of observing a t-statistic $\geq$ 6.6 or $\leq$ -6.6 using the `pt` command?




```
pt(q = summary_stats %>% pull(t_stat), lower.tail = F, df = 10) * 2
```

```
## [1] 6.053362e-05
```

Calculate the test statistic, p-value, and 95% confidence interval

- To calculate the 95% confidence interval, we need to know the quantile of the t-distribution such that 2.5% of the data lies above or below it.
- Ask R: What is the quantile such that 97.5% of the t-distribution is below it with 10 degrees of freedom using the `qt` command?

```
q <- qt(p = 0.975, lower.tail = T, df = 10)
q
```

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

```
mean_of_diffs <- summary_stats %>% pull(mean_diff)
moe <- q * summary_stats %>% pull(std_err_diff)
ucl <- mean_of_diffs + moe
lcl <- mean_of_diffs - moe
c(lcl, ucl) %>% round(2)
```

```
## [1]  5.11 10.32
```

The confidence interval is 5.11 to 10.32.

Does this 95% confidence interval contain the null hypothesized value if there were no difference between Diet A and Diet B in their effects on cholesterol?

Calculate the test statistic, p-value, and 95% confidence interval

- Or, have R do the work for you! Just be sure to specify that `paired = T`.

```
paired_t <- t.test(chol_dat %>% pull(B), chol_dat %>% pull(A),
                   alternative = "two.sided", mu = 0, paired = T)
paired_t
```

```
##
##  Paired t-test
##
## data:  chol_dat %>% pull(B) and chol_dat %>% pull(A)
## t = 6.6033, df = 10, p-value = 6.053e-05
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   5.112712 10.320261
## sample estimates:
## mean difference
##        7.716487
```

Compare the outputs from the independent and paired tests

- Below, we re-run the code, but this time we do not specify `paired=T`.
- Note this is the wrong test because these data are **paired**. We print it here to see what the results look like when we run the data as if it were not paired, that is, as if it were **independent**:

```
indep_t
```

```
##
##  Two Sample t-test
##
## data:  chol_dat %>% pull(B) and chol_dat %>% pull(A)
## t = 0.78557, df = 20, p-value = 0.4413
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.77356  28.20653
## sample estimates:
## mean of x mean of y
##  209.3529  201.6364
```

Compare the outputs from the independent and paired tests

- Now we can compare the output on the previous slide to the output when we specify `paired=T`
- This is the correct test. It accounts for the **paired** nature of the data. How does the test statistic (t), the p-value, and the degrees of freedom change from the previous output?

```
paired_t
```

```
##
##  Paired t-test
##
## data:  chol_dat %>% pull(B) and chol_dat %>% pull(A)
## t = 6.6033, df = 10, p-value = 6.053e-05
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##   5.112712 10.320261
## sample estimates:
```

```
## mean difference
##        7.716487
```

Compare the outputs from the independent and paired tests

- What is the same?
- What is different?
- What is the estimate of the standard error of the mean for the independent vs. paired t-test?
  - To get the answer, rearrange the formulas for the test statistics to solve for the SEs
  - You know the value of the test statistics and the numerator of the test statistics from the R output on the previous two slides

Compare the outputs from the independent and paired tests

- What is the same?
- What is different?
- What is the estimate of the standard error of the mean for each test?

```
SE_indep <- (indep_t$estimate[1] - indep_t$estimate[2]) / indep_t$statistic
SE_paired <- paired_t$estimate / paired_t$statistic

c(unname(SE_indep), unname(SE_paired))
```

```
## [1] 9.822822 1.168587
```

The first value is the SE when the data was treated as if it were independent. The second value is the SE when the data was treated as paired.

Paired t-test: More juice per squeeze

- Question:The standard error was much lower using the paired test. Why?
- Answer:
  - Only variation within an individual was used to calculate the SE of the mean difference
  - There is much less variation within an individual (across diets) than between individuals

The Statistical Method

Problem **P**lan **D**ata **A**nalysis **C**onclusion

The Statistical Method

Problem **P**lan **D**ata **A**nalysis **C**onclusion

Plan, a.k.a. experimental design

- Once the **problem** has been stated, the next step is to determine a **plan** to best answer the question
- One of the tenets of design is to maximize **efficiency**
- In our toy example, a paired test greatly maximized the efficiency by removing the noise introduced by between-subject variability

When is a paired design the appropriate design?

1. When "the treatment alleviates a condition rather than affects a cure." (Hills and Armitrage, 1979)
   - The effect of treatment is short-term. After t amount of time, participants return to baseline.
   - The t above refers to the **wash-out** period. Before applying the second treatment, participants should have enough time to reach their baseline level. Otherwise there may be a **carry over** effect.

- For example, could you run this study if Diet A led to a permanent change in cholesterol levels?
  - No, because if individuals took Diet A first, then their cholesterol measures after Diet B would equal those after Diet A because the effect of Diet A on the outcome is persisting

When is a paired design the appropriate design?

2. The time between the alternative treatments isn't so long as to introduce confounding by other factors
   - For example, if you waited a year between applying treatments, other things may have changed in the world or in a person's life that affects the outcome
   - You need to strike a balance between waiting too long and not waiting long enough

Example: Canned soup, fresh soup, and urinary bisphenol A (BPA)

Reference: Carwile JL, Ye X, Zhou X, Calafat AM, Michels KB. **Canned soup consumption and urinary bisphenol A: a randomized crossover trial**. *JAMA*. 2011. 306(20):2218-20.

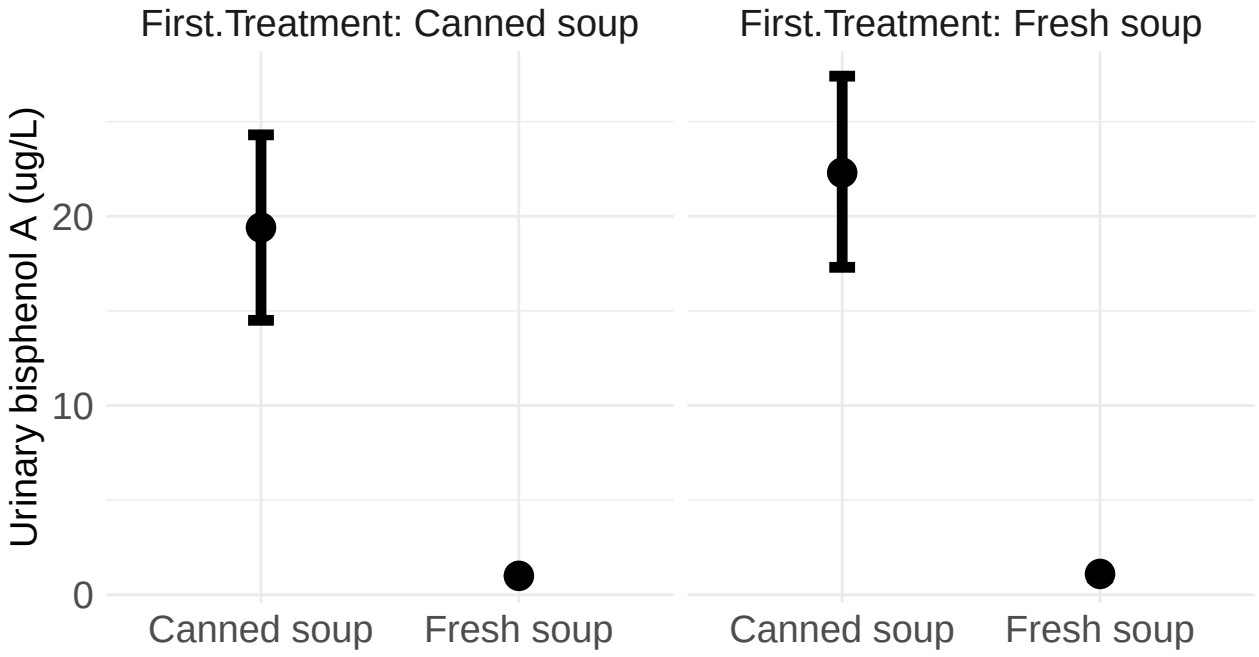
- One group randomized to consuming canned soup for five days at lunch
- Second group randomized to consuming fresh soup
- After two day washout, treatment assignment were reversed
- On the 4th and 5th day of each treatment, urine samples were collected and pooled for each subject to measure BPA

Example: Canned soup, fresh soup, and urinary bisphenol A (BPA)

- Paired t-test used to measure the difference
- After canned soup consumption, urinary BPA concentrations were 22.5 μg/L higher (95% CI: 19.6, 25.5) than those measured after fresh soup consumption ($p < .001$)

Example: Canned soup, fresh soup, and urinary bisphenol A (BPA)

- Treatment sequence did not appear to affect the results



Statistical assumptions of the paired t-test

- The sample is randomly chosen
  - How was the sample selected?
- The distribution of differences is Normal
  - How could you examine this assumption?

Observational analogues

- Recall from our earlier example the primary benefit of the paired t-test was the total removal of between-subject variability
- Observational settings at high risk of confounding due to differences between subjects can also benefit from this design
- In observational studies, we're less interested in hypothesis testing, and more interested in estimating the difference and its confidence interval

Example: Vaccination during infancy and healthcare utilization

- While there are many disproven concerns about vaccination, one potentially valid concern is that vaccination may trigger an immune response and lead to inflammation and associated hospital visits (Wilson et al, 2011)
- For non-live vaccines, any adverse inflammation event would occur within 3 days of vaccination
- How might you study the effect of vaccination on healthcare utilization?

Example: Vaccination during infancy and healthcare utilization

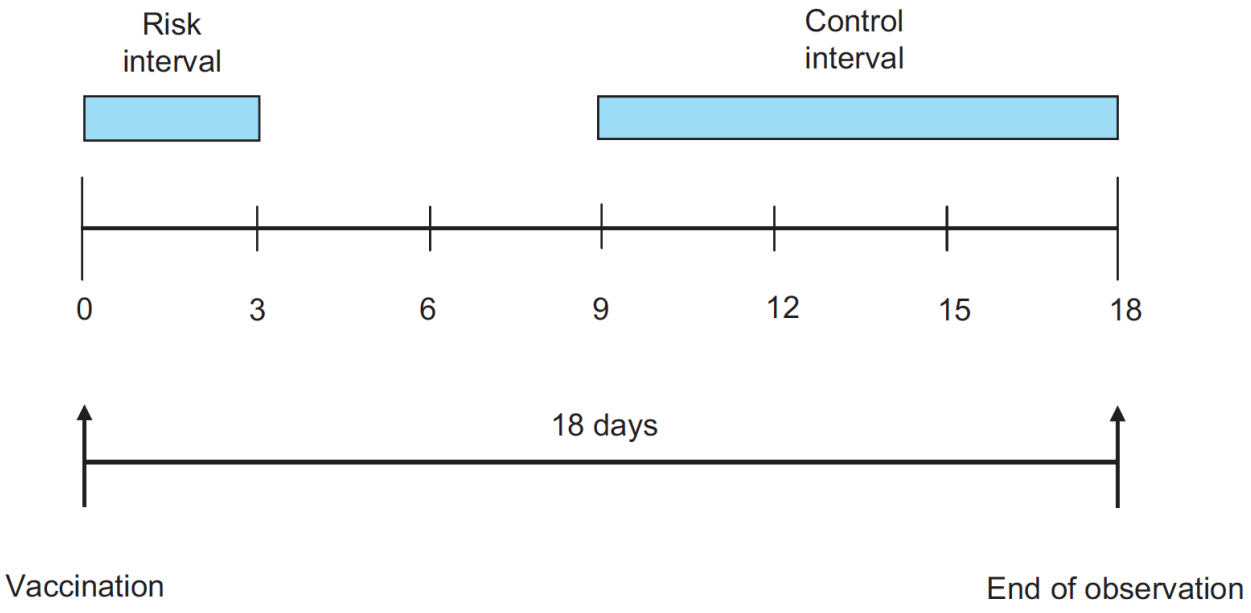
- To fully remove confounding from differences between children, you could conduct a paired design in an observational setting
- You can compare a period of time after vaccination to a control period of time *for each child*

Example: Vaccination during infancy and healthcare utilization

- When should the control period be?
- How long should the washout period be?

Example: Vaccination during infancy and healthcare utilization

Reference: Wilson K, Hawken S, et al. *Vaccine.* 2011;29(21):3746-52.



Example: Vaccination during infancy and healthcare utilization

- "333,244 children received a vaccination at 62 days of age ±30 days, 86% of which were within a ±10 days window. Of these, 1,388 experienced one of the combined end-points during the immediate 3 days post vaccination, compared to 4,893 in the 9-day control period for our primary analysis. The relative incidence of an event was 0.85 (0.80–0.90)."

Example: Vaccination during infancy and healthcare utilization

- 1,388 events in 3 days after vaccination among 333,244 children
  - This implies 462.7 events per day for these children
- 4,893 events in 9 days after vaccination among 333,244 children
  - This implies 543.7 events per day for the same number of children
- The relative incidence is: $462.7/543.7 = 0.85$, as stated in the paper
- The relative incidence is < 1 which indicates more events in the 9-18 day block after vaccination vs. the 0-3 day block

Check your understanding!

Summary

1. Use a paired t-test in an experimental setting where the data is paired by design
   - Pairing can greatly improve efficiency of the estimator
2. Incorporate pairing/clustering into observational analyses when the data is paired/clustered by design or in nature
   - Pairing eliminates confounding from factors that differ between subjects

Resources

- Gerald Dallal write-up on paired t-tests: http://www.jerrydallal.com/LHSP/paired.htm
  - The cholesterol example was derived from this work
- Mike Marin lecture on the t-distribution (4.18): https://www.youtube.com/watch?v=ETd-jPhI_tE, and on conducting paired t-tests (4.19): https://www.youtube.com/watch?v=yD6aU0fY2lo
  - Useful if you want a short introduction on conducting t-tests in R
- Hills M, Armitage P. The Two-Period Cross-over Clinical Trial. *Br J Clin Pharmac.* 1979. 8:7-20.

  - https://www.ncbi.nlm.nih.gov/pubmed/15595959
  - A bit advanced for this lecture. A really nice description of these trial designs
- Carwile JL, Ye X, Zhou X, Calafat AM, Michels KB. **Canned soup consumption and urinary bisphenol A: a randomized crossover trial**. *JAMA*. 2011. 306(20):2218-20.
  - https://www.ncbi.nlm.nih.gov/pubmed/22110104
  - A simple and short example of an experimental crossover trial
- Wilson K, Hawken S, et al. Patterns of emergency room visits, admissions and death following recommended pediatric vaccinations—a population based study of 969,519 vaccination events. *Vaccine*. 2011;29(21):3746-52.
  - https://www.ncbi.nlm.nih.gov/pubmed/21443964
  - Observational example of paired data (emphasis on estimation, not testing)