More flavors of T: paired tests

Flavors of T

Last lecture we introduced the T test for two independent samples

In this lecture we will extend our t-testing framework to consider what happens when those samples are NOT independent.

Example: Weight by gender

Imagine for example that we want to show that weight is different among males and females in the United States. Imagine we have data from 100 randomly sampled males and 100 randomly sampled females in the United States.

We would test the null hypothesis that there is no difference between the mean weight of men and women in the united states

$$\bar{X}_{(group_a)} = \bar{X}_{(group_b)}$$

Example: Weight by gender

Would we consider these samples independent?

standard two-sample t-test

Two sample t-test

We want to compare the mean weight for each group, and use the standard error of the weights of these groups to calculate a t-test. This helps us to understand if the difference in the means is larger than we might see due to the variability of the weights in our observations.

we would calculate

$$t = \frac{x_{groupa} - x_{groupb}}{\sqrt{\frac{S^2_{groupa}}{100} + \frac{S^2_{groupb}}{100}}}$$

and compare this to a t-distribution at our chosen critical point with appropriate degrees of freedom

Two sample t-test

To illustrate this example, I have simulated data for males and females using the mean and standard deviation of weights in the United States taken from the CDC NHANES data

```
weights %>% group_by(sex) %>% summarise(sample_mean = mean(weight1),
                                        sample_sd = sd(weight1),
                                        length = length(weight1))
```

```
## # A tibble: 2 x 4
##   sex   sample_mean sample_sd length
##   <chr>       <dbl>     <dbl>  <int>
## 1 F            171.      29.1    100
## 2 M            191.      28.9    100
```

Two sample t-test

I can overlay the histograms for these data with this code:

ggplot(weights,aes(x=weight1)) +

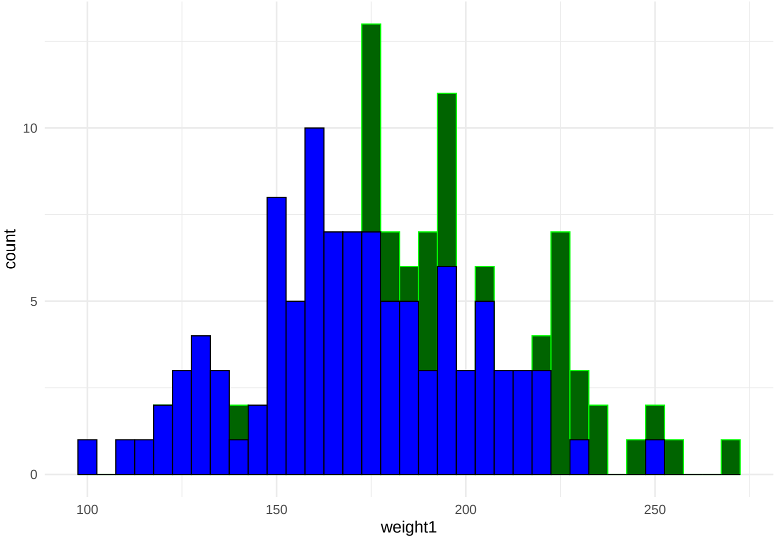
```
geom_histogram(data=subset(weights,sex == 'M'),binwidth=5,fill="dark green",
```

```
geom_histogram(data=subset(weights,sex == 'F'),binwidth=5,fill = "blue", col=
```

theme_minimal(base_size = 15)

Notice that I am using two geom_histogram statements to lay the histograms on top of one another rather than using a "fill" statement in one geom_histogram.

Two sample t-test



Two sample t-test

And a Student's T test will show that this difference is statistically significant - notice the syntax here

```
t.test(weights$weight1~weights$sex, alternative="two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  weights$weight1 by weights$sex
## t = -5.0723, df = 197.99, p-value = 9.015e-07
## alternative hypothesis: true difference in means between group F and group
## 95 percent confidence interval:
##  -28.91051 -12.72374
## sample estimates:
## mean in group F mean in group M
##        170.5458        191.3629
```

Independent vs non-independent samples

In this example we have measured randomly selected males and females and we have no reason to believe their measurements are correlated. So a two-sample simple t-test is a reasonable approach.

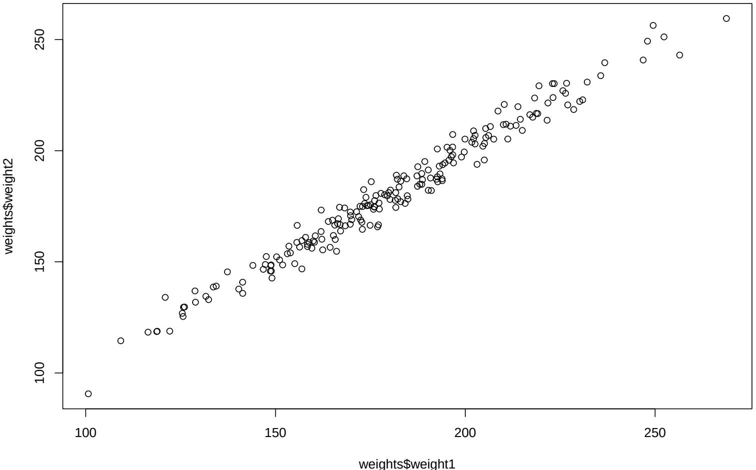
What happens if we imagine that these 200 individuals are all invited to participate in a weight loss trial.

We have their baseline weight, and after 6 months of participation in the trial they are weighted again.

What would we assume about the independence of our measures now?

Independent vs. non-independent samples

Using r to graph the pre and post-trial weights we can see that these are correlated



Independent vs. non-independent samples

For each individual in this study, we will will compare their weight after 6 months in the trial to their weight at baseline. Now we have broken our assumption (needed for the Student's t-test) that the measurements in the two groups (pre and post) are independent of each other.

We would expect that each person's weight at 6 month follow up will be closely related to their own weight at baseline. We would also expect that the variation in weight within one person will be much less than the variation in weight between people.

In this case, because I have simulated the data, I know that this hypothetical weight loss program results in an average weight loss of 5 pounds with a standard deviation of 5 pounds.

Independent vs. non-independent samples

Let's take a look at what happens when we use this Student's T test to compare weights before and after the intervention without taking into account the relationship of these measurements:

If we do not take into account the paired structure of the data, we are testing the null hypothesis

$$\bar{X}_{(weightpretrial)} = \bar{X}_{(weightposttrial)}$$

and our t-test would be based on

$$t = \frac{\overline{X_{(weightpretrial)}} - \overline{X_{(weightposttrial)}}}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

Independent vs. non-independent samples

```r
t.test(weights$weight1, weights$weight3, data=weights)
```

```
##
##  Welch Two Sample t-test
##
## data:  weights$weight1 and weights$weight3
## t = 1.6059, df = 397.72, p-value = 0.1091
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.122791 11.139438
## sample estimates:
## mean of x mean of y
##  180.9544  175.9460
```

Independent vs. non-independent samples

We see that the estimated difference in weight is close to 5 pounds, but the results are not statistically significant. If we do not account for the relatedness of these measurements there is too much "noise" or variation between the measurements to see the "signal" or the true difference in means.

The solution to this problem is to look at the measurements in pairs and base our statistical testing on the variability in the difference between the pre and post intervention measures of weight.

Paired t-test

Paired t-test

In this case we are now testing the null hypothesis that the difference is 0

This is called a paired t-test.

$$t = \frac{\bar{d}_{(weightpost - weightpre)}}{\frac{S_d}{\sqrt{n}}}$$

Paired t-test

```
weights %>% summarise(dif_mean = mean(dif2), dif_sd = sd(dif2),
                      wt1_mean=mean(weight1),wt1_sd=sd(weight1),
                      wt3_mean=mean(weight3),wt3_sd=sd(weight3))
```

```
##    dif_mean   dif_sd wt1_mean   wt1_sd wt3_mean   wt3_sd
## 1 -5.008323 4.854787 180.9544 30.77072  175.946 31.59696
```

Paired t-test

Notice the syntax here:

```
t.test(weights$weight1, weights$weight3,data=weights, paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  weights$weight1 and weights$weight3
## t = 14.589, df = 199, p-value < 2.2e-16
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  4.331380 5.685267
## sample estimates:
## mean difference
##        5.008323
```

Paired test results

Here we see that the estimate of difference is unchanged, but the t-test is now using the standard deviation of the difference (4.85) rather than the standard deviation of weights between people at each time point(30.77 and 31.6) to determine whether this difference is statistically significant.

With the paired test, our value of t is much higher and our results are statistically significant.

Distribution of differences

If we graph the mean values and distribution of the difference between pre and post trial weight, and the overall weights post trial we can see that the variability is much smaller for the difference.

ggplot() +

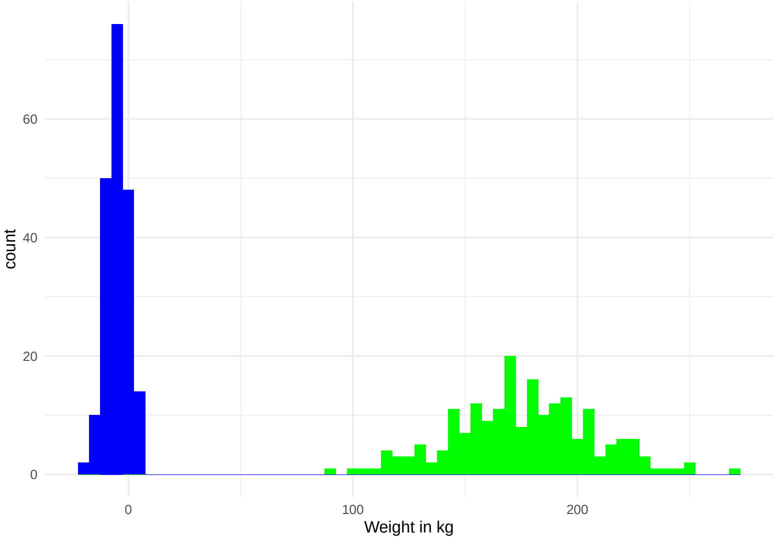
geom_histogram(data = weights, aes(x = weight3), binwidth=5, fill="green") +

geom_histogram(data = weights, aes(x = dif2), binwidth=5, fill="blue") +

labs(x = "Weight in kg") +

theme_minimal(base_size = 15)

Distribution of differences



t-test: More juice per squeeze?

t-test: More juice per squeeze?

When we have

- The standard error was much lower using the paired test. Why?
- Only variation within a subject was used to calculate the SE of the mean difference
- there was much less variation within a subject than between subjects

The Statistical Method

Problem

Plan

Data

Analysis

Conclusion

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.
- When data are paired a paired test greatly maximizes the efficiency by removing the noise introduced by between-subject variability.

When is a paired design the appropriate design?

- Studies with multiple measures on the same units of observation
- Studies with inherently related observations
- Studies that match units of observation to reduce variability

Studies with mutliple measures on the same units

Cross -over or before and after studies - in our weigh-loss example we were looking at measures before and after participation. . .

- When "the treatment alleviates a condition rather than affects a cure." (Hills and Armitrage, 1979)
- The effect of treatment is short-term. After x amount of time, participants return to baseline.
- The x 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.

Studies with mutliple measures on the same units

Considerations for before/after or cross-over studies - 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. - Thus, there is a balance between waiting too long or not waiting long enough.

If we wanted to look at changes in individual related to a treatment what other type of design might we consider?

Inherently related observations

- Matched body parts
- Studies in identical twins
- Studies of diet or health behaviors in couples or family members

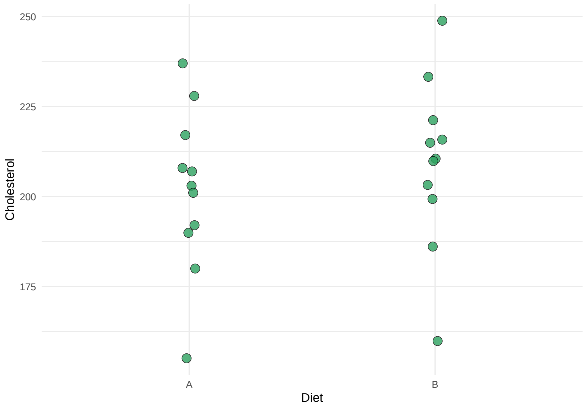
Studies that match to reduce variability

- Matched communities
- Matched individuals

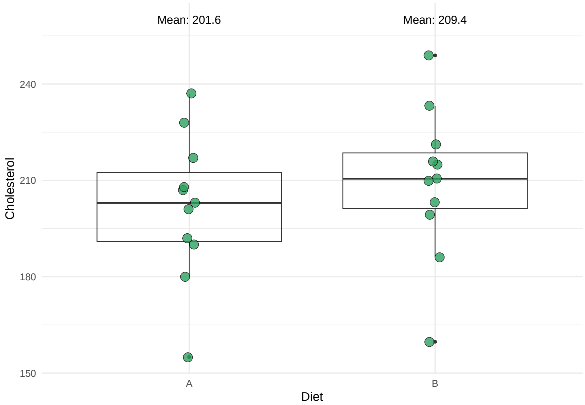
Example small study of diet

Cholestorol measurements following two alternative diets -

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



Cholestorol measurements following two alternative diets -



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

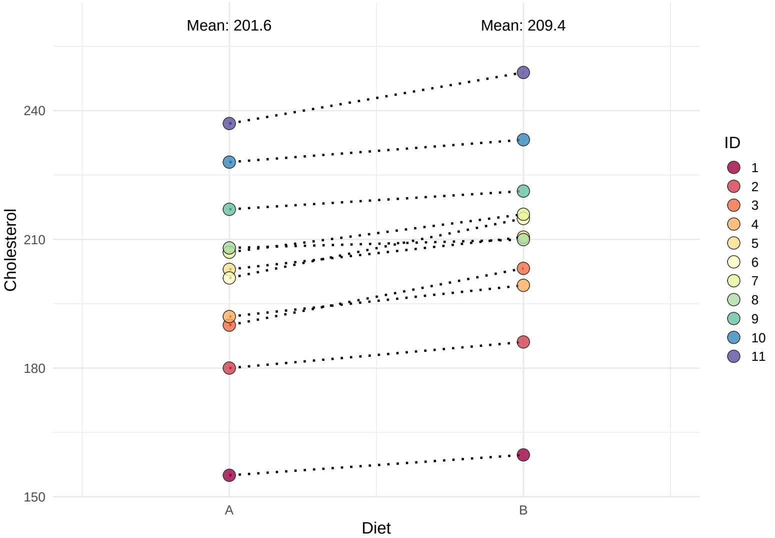
Cholestorol measurements following two alternative diets -

An independent t-test reveals no evidence against the null hypothesis of no difference between the diets:

```
##
##  Welch Two Sample t-test
##
## data:  chol_dat %>% pull(A) and chol_dat %>% pull(B)
## t = -0.78557, df = 19.976, p-value = 0.4413
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28.20808  12.77511
## sample estimates:
## mean of x mean of y
##  201.6364  209.3529
```

Better visualization for a very small study

Now, what do you notice about the paired data?



apply a paired t-test

- The observed value of the test statistic is: $t = \frac{\bar{x}_d - 0}{s_d/\sqrt{n}}$
- It can be compared to a critical value from the t distribution with $n-1$ degrees of freedom

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

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

```
##     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:

```
chol_dat <- chol_dat %>%mutate(diff = B - A)
head(chol_dat)
```

```
##     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

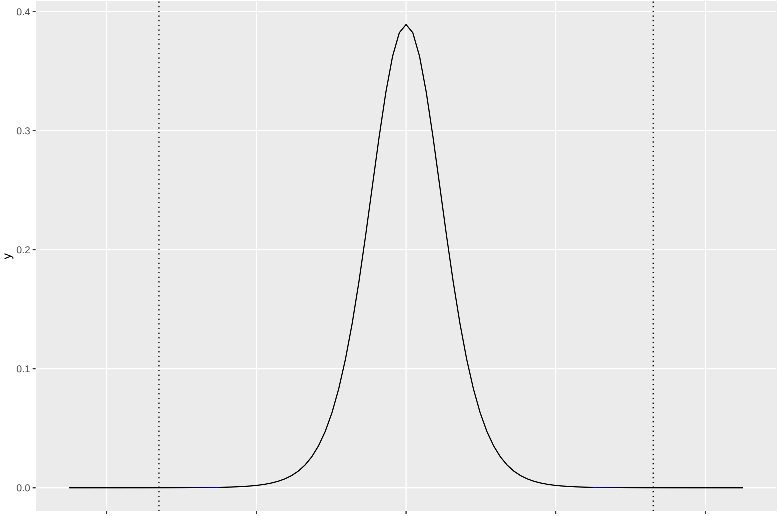
- Then use `summarize` to calculate the mean difference ($\hat{\mu}_d$), its standard error ($\hat{s}_d/\sqrt{n}$), and the observed t-statistic:

```
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
```

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

What is the probability of observing a t-stat $\geq$ 6.6 or $\leq$ -6.6 using the pt command.



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 on 10 degrees of freedom using the `qt` command.

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

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

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

```
ucl <- summary_stats %>% pull(mean_diff) + (q * summary_stats %>% pull(std_er
lcl <- summary_stats %>% pull(mean_diff) - (q * summary_stats %>% pull(std_er
c(lcl, ucl)
```

```
## [1]  5.112712 10.320261
```

The confidence interval is (5.1127122, 10.3202611).

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`.

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

```
##
##  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

	Independent	Paired
T statistic	-0.78557	6.6033
df	19.976	10
pvalue	0.4413	6.053e-05
mean	201.67 vs 209.35	7.72
95% CI	-28.21 to 12.78	5.11 to 10.32
SE	9.823	1.169

- What is the same?
- What is different?

Examples - which flavor of T?

Statistics is everywhere

The "sleepy" mocktail

You may have seen "sleepy girl mocktail" recipes making the rounds on social media, promising deep and restful sleep. Usually, these non-alcoholic beverages include a combination of tart cherry juice with other ingredients, like a magnesium supplement powder.

We spoke with wellness dietitian Lindsey Wohlford to understand more about how these ingredients impact sleep and what cancer patients should know.

Image from an MD Anderson health website article

RCT of magnesium

Abbasi B, Kimiagar M, Sadeghniiat K, Shirazi MM, Hedayati M, Rashidkhani B. The effect of magnesium supplementation on primary insomnia in elderly: A double-blind placebo-controlled clinical trial. J Res Med Sci. 2012 Dec;17(12):1161-9. PMID: 23853635; PMCID: PMC3703169.

Recruited elderly individuals with diagnosed insomnia, screened out other medical conditions including sleep apnea. Intervention was 8 weeks.

RCT of magnesium

Table 2.

Comparison of sleep indices in magnesium supplementation and placebo groups before and after intervention

Variable	Magnesium supplementation (n=21)			$P1^*$	Placebo (n=22)			$P2^\dagger$	$P3^\ddagger$
	Before intervention	After intervention	Difference (CI=95%)		Before intervention	After intervention	Difference (CI=95%)		
Insomnia severity index	16.52±2.01	14.14±2.68	−2.38±2.24	<0.001	16.27±1.69	15.77±1.92	−0.5±1.71	0.2	0.006
Total sleep time (h)	7.8±1.1	7.9±0.6	0.1±0.7	0.4	7.6±0.9	7.6±0.8	−0.03±0.3	0.6	0.3
Sleep time (h)	5.1±0.8	5.7±0.9	0.6±0.7	0.002	5.0±0.5	5.0±0.6	−0.02±0.3	0.7	0.002
Sleep onset latency (h)	1.3±0.2	1.1±0.4	−0.2±0.4	0.04	1.4±0.2	1.4±0.2	0.04±0.1	0.1	0.02
Early morning awakening (h)	1.04±0.02	1.01±0.05	−0.03±0.05	0.05	1.03±0.02	1.03±0.02	−0.01±0.01	0.09	0.08
Sleep efficiency (h)	0.67±0.07	0.73±0.1	0.06±0.1	0.02	0.66±0.04	0.66±0.07	0.00±0.05	0.2	0.006

$P1^*$ P value of differences in magnesium group compared via paired t-test; $P2^\dagger$ P value of differences in placebo group compared via paired t-test; $P3^\ddagger$ P value of differences between magnesium and placebo groups compared via independent samples t-test

Which flavor of T?

1) You want to see if there is a difference in blood pressure among men and women. You randomly sample 10 households from each census tract in a city and measure blood pressure of a man and woman living in each household.
2) You are interested in the efficacy of a medication for rheumatoid arthritis. You measure severity of symptoms among individuals randomized to treatment or control.
3) You are interested in family size and hyperactivity. You measure hyperactive behavior among only children vs children with siblings.

Which flavor of T?

4) You are testing a new medication for glaucoma. You randomize individuals with glaucoma in both eyes to put active medication in their right or left eye.
5) You are interested in educational attainment in charter schools. You measure scores on a standardized test among students and a charter school and compare the scores to the state average for public schools.

coding notes

A one sample t- test will take the form:

t.test(x = x variable, alternative = greater, less or two.sided, mu = null hypothesis value)

A two sample t-test will take the form:

t.test(first sample data, second sample data, alternative = greater, less or two.sided)

A paired t-test will take the form:

t.test(first data points, second datapoints, alternative = greater, less or two.sided, paired=TRUE)

parting humor