IMS 17 (ish): Inference for comparing two proportions

Libraries needed

```
library(tidyverse)
library(openintro)
library(infer)
```

CPR Data

CPR is a procedure used on individuals suffering a heart attack when other emergency resources are unavailable. The procedure is helpful, but can cause internal bleeding. Blood thinners may also be used, but can make internal bleeding much worse. Experts are not sure if the additional blood thinner treatment is worth it.

Here we consider an experiment with patients who underwent CPR for a heart attack and were subsequently admitted to a hospital. Each patient was randomly assigned to either receive a blood thinner (treatment group) or not receive a blood thinner (control group).

The outcome variable of interest was whether the patient survived for at least 24 hours.

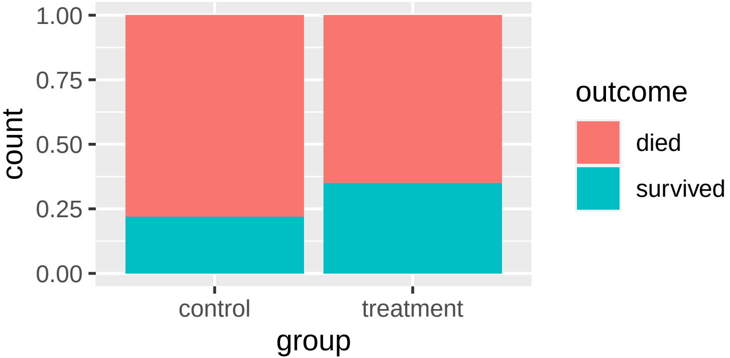
CPR Data

- Heart attack patients that received CPR
- Control: no blood thinner
- Treatment: blood thinner

	Died	**Survived**	**Total**
Control	39	11	50
Treatment	26	14	40
Total	65	25	90

CPR Data

```
ggplot(cpr, aes(x = group, fill = outcome))+
  geom_bar(position = "fill")
```



Calculate the survival rate for each group, p_C and p_T. What are your initial thoughts?

Inference for Two Proportions using a Randomization Test

Step 1: Frame the research question in terms of hypotheses

Under H_0 we think that blood thinners do not have an overall survival effect, i.e., the survival proportions are the same in each group.

$$H_0 : p_T - p_C = 0$$
$$H_A : p_T - p_C \neq 0$$

Step 2: Collect Data (or Check Conditions)

- **Independence** (extended). The data are independent *within* and *between* the two groups. Generally this is satisfied if the data come from two independent random samples, or if the data come from a randomized experiment.
- We technically do not need the success-failure condition for a randomization test.

Example

For the CPR example the grouping was randomly assigned and the patients are independent from each other. It is safe to assume we have independence from patient to patient (within groups), and independence between control and treatment (between groups).

Step 3: Model the randomness that would occur if the null hypothesis was true (randomization test)

If there truly was no difference between the groups then any subject could have just as easily came from either group (control/no-blood-thinner or treatment/blood-thinner).

We will do a *what if* computer experiment:

- Take all the patients and shuffle them.
- Pick 50 at random to be from the "control group", and the remaining 40 to be in the "treatment group".
- Calculate $\hat{p}_T^* - \hat{p}_C^*$ using the shuffled data
- Repeat this process $B \geq 1000$ times.
- Graph the all the B shuffled values in a histogram.
- Check where your observed $\hat{p}_T - \hat{p}_C$ is.

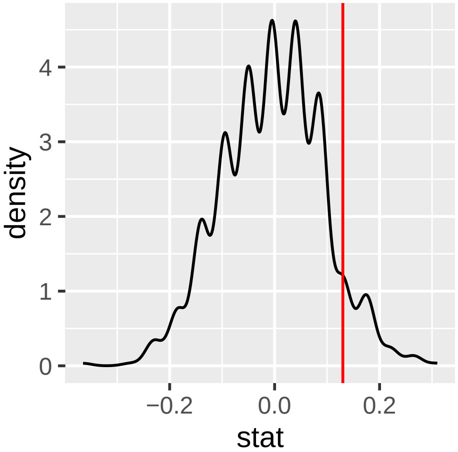
Step 3: Model the randomness that would occur if the null hypothesis was true (randomization test)

```r
set.seed(62)
null <- cpr |>
  specify(outcome ~ group, success = "survived") |>
  hypothesize(null = "independence") |>
  generate(reps = 750, type = "permute") |>
  calculate(stat = "diff in props", order = c("treatment", "control"))


d_hat <- 14/40 - 11/50

ggplot(null, aes(x = stat)) +
  geom_density() +
  geom_vline(xintercept = d_hat, color = "red")
```

Step 3: Model the randomness that would occur if the null hypothesis was true (randomization test)



Step 4: Analyze the data.

TWO-SIDED TEST:

To calculate the p-value we look at how many observations are in the smaller tail, and multiply it by two.

ONE-SIDED TEST:

You only need to consider one tail.

```
null |>
  summarize(pval = 2 * mean(stat >= d_hat))
```

```
# A tibble: 1 x 1
   pval
  <dbl>
1 0.197
```

Step 5: Form a conclusion.

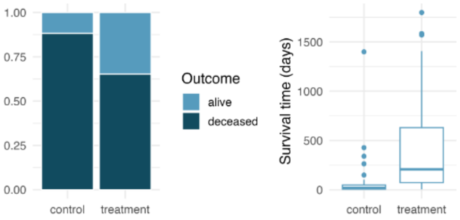
Conclusions follow same format as before.

Example

Our p-value is 0.197 which is larger than $\alpha = 0.05$. We therefore conclude in favor of H_0. It seems reasonable to believe that the survival proportions could be the same in each group.

Practice Problem

[IMS 11.8 Adjusted] Heart transplants. The Stanford University Heart Transplant Study was conducted to determine whether an experimental heart transplant program increased lifespan for those who were gravely ill and would likely benefit from a new heart. Patients in the treatment group got a transplant and those in the control group did not. Of the 34 patients in the control group, 30 died. Of the 69 people in the treatment group, 45 died. Another variable called survived was used to indicate whether the patient was alive at the end of the study.



a) Does the stacked bar plot indicate that survival is independent of whether the patient got a transplant? Explain your reasoning.

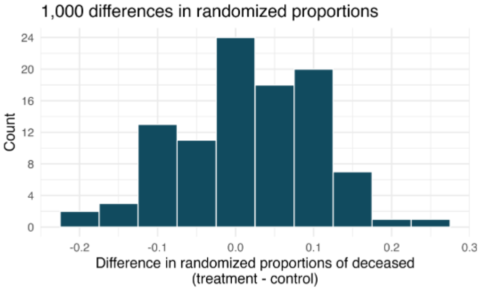
b) What do the box plots above suggest about the efficacy (effectiveness) of the heart transplant treatment.

Practice Problem

[IMS 11.8 Adjusted] Heart transplants.

c) What proportion of patients in the treatment group and what proportion of patients in the control group died?

d) What do the simulation results shown below suggest about the effectiveness of the transplant program?



Inference for Two Proportions using Math

Pooled Proportion

- When testing if the difference between two groups is 0, i.e. $H_0 : p_T - p_C = 0$, this is equivalent to saying the groups (treatment and control) do not matter .
- If the groups do not matter, then we can get an estimate as a "what-if" for what the proportion of success we'd expect for the full data set.

$$\hat{p}_{pool} = \frac{\text{number of succeses}}{\text{number of cases}}$$

- A "success" is a case that meets a condition.
- $\hat{p}_T$ and $\hat{p}_C$ are probabilities of "success", or in our case *survival*

Pooled Proportion

For our example: $\hat{p}_{pool} = \frac{25}{90}$

	Died	**Survived**	**Total**
Control	39	11	50
Treatment	26	14	40
Total	65	25	90

Step 2: Collect Data (or Check Conditions)

- **Independence** (extended). The data are independent *within* and *between* the two groups. Generally this is satisfied if the data come from two independent random samples, or if the data come from a randomized experiment.
- **Success-Failure Condition**. The success-failure condition needs to hold for both groups, where we check successes and failures in each group separately.

$$\begin{aligned} n_T\hat{p}_{pool} &\geq 10 \\ n_T(1-\hat{p}_{pool}) &\geq 10 \\ n_C\hat{p}_{pool} &\geq 10 \\ n_C(1-\hat{p}_{pool}) &\geq 10 \end{aligned}$$

Step 2: Collect Data (or Check Conditions)

Checking the Success-Failure Condition with the CPR data

$$40\left(\frac{25}{90}\right) \geq 10 \qquad \checkmark$$

$$40\left(1-\frac{25}{90}\right) \geq 10 \qquad \checkmark$$

$$50\left(\frac{25}{90}\right) \geq 10 \qquad \checkmark$$

$$50\left(1-\frac{25}{90}\right) \geq 10 \qquad \checkmark$$

Step 3: Model the randomness that would occur if the null hypothesis was true (math)

We need a new estimate for standard error because now we have two groups.

Standard Error under H_0 for Two Proportions

The standard error under H_0 for testing if the difference in two propotions is 0 is

$$SE_{H_0} = \sqrt{\hat{p}_{pool}(1 - \hat{p}_{pool})\left(\frac{1}{n_T} + \frac{1}{n_C}\right)}$$

Step 3: Model the randomness that would occur if the null hypothesis was true (math)

Example

The standard error for the CPR data is

$$SE_{H_0} = \sqrt{\frac{25}{90}\left(1 - \frac{25}{90}\right)\left(\frac{1}{50} + \frac{1}{40}\right)} \approx .0950$$

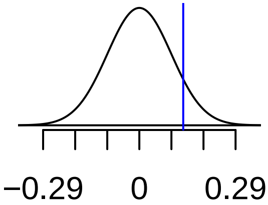
Step 3: Model the randomness that would occur if the null hypothesis was true (math)

Under H_0, we believe

$$\hat{p}_T - \hat{p}_C \overset{H_0}{\sim} N(0, 0.0950)$$

Vertical line is at $\hat{p}_T - \hat{p}_C$, the data we observed

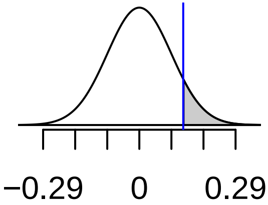
```
p_hat_diff <- 14/40 - 11/50
SE = .0950
normTail(m = 0, s = SE)
abline(v = p_hat_diff, col = "blue")
```



Step 4: Analyze the data.

Create a two-sided p-value

```
p_hat_diff <- 14/40 - 11/50
SE = .0950
normTail(m = 0, s = SE, U = p_hat_diff)
abline(v = p_hat_diff, col = "blue")
```




```
p_value <- 2*(1 - pnorm(p_hat_diff, mean = 0, sd = SE))
p_value
```

```
[1] 0.1711803
```

The p-value is approximately 0.1712

Step 5: Form a conclusion

Example

Our p-value is 0.1712 which is larger than $\alpha = 0.05$. We therefore conclude in favor of H_0. It seems reasonable to believe that the survival proportions could be the same in each group.

Practice Problem

[IMS 17.9 Adjusted] National Health Plan. A Kaiser Family Foundation poll for US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic 'National Health Plan.' There were 347 Democrats, 298 Republicans, and 617 Independents surveyed. 79% of 347 Democrats and 55% of 617 Independents support a National Health Plan.

a) Set up the hypotheses for testing if the difference between the proportion of Democrats and Independents who support a National Health Plan is 0.
b) Check the conditions for doing a hypothesis test with the mathematical model.
c) Calculate your p-value using the mathematical model, and make a conclusion using $\alpha = 0.05$.
d) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the National Health Plan than the Independent.

Practice Problem

[IMS 17.22 Adjusted] Diabetes and unemployment. A Gallup poll surveyed Americans about their employment status and whether they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.

a) Create a two-way table presenting the results of this study.
b) State appropriate hypotheses to test for difference in proportions of diabetes between employed and unemployed Americans.
c) The sample difference is about 1%. If we completed the hypothesis test, we would find that the p-value is very small (about 0), meaning the difference is statistically significant. Use this result to explain the difference between statistically significant and practically significant findings.