Type I and Type II error

In hypothesis testing, we have two hypotheses, the null $\mathbf{H}_0$ and the alternative $\mathbf{H}_a$. We get a random sample, use it to obtain a test statistic, construct a null distribution (or when, as usual, we are unable to do that, we approximate it with a randomization distribution), locate our test statistic on this (approximate) null distribution, then find a P-value which measures how likely one is to obtain a result at least as extreme as our test statistic in a world where the null hypothesis is true. Examples of this process for various settings have been provided at `https://rstudio.calvin.edu:3939/connect/#/apps/55/`. I carry out one below involving the single quantitative variable `pulse` (in beats per minute).

Example 1:

I wish to test the hypothesis that mean resting pulse rate is 72 beats per minute vs. a 2-sided alternative. The hypotheses are

$$\mathbf{H}_0: \ \mu = 72, \qquad \mathbf{H}_a: \ \mu \neq 72.$$

In the Lock5withR dataset **BodyTemp50**, there is a sample (we will suppose it is a random sample) of 50 subjects, with one measurement variable being `Pulse`. The mean for this sample, our **test statistic**, is $\overline{x} = 74.4$:

```
mean(~Pulse, data=BodyTemp50)


[1] 74.4
```

If we were bootstrapping to find a confidence interval for μ, the population mean pulse, we would do something like this:

```
manybstrapXbars <- do(5000) * mean(~Pulse, data=resample(BodyTemp50))
```

Constructing a randomization distribution for a single mean is quite similar. The previous command produces a *bootstrap distribution*, one centered on the mean of the original sample, 74.4. That is the one thing we must change in order to call our result a *randomization distribution*, as a randomization distribution should be centered on the null value, 72, in this case. This amounts to sliding all of our bootstrapped statistics over the appropriate amount, in this case adding

$$72 - 74.4 = -2.4$$

to each one.

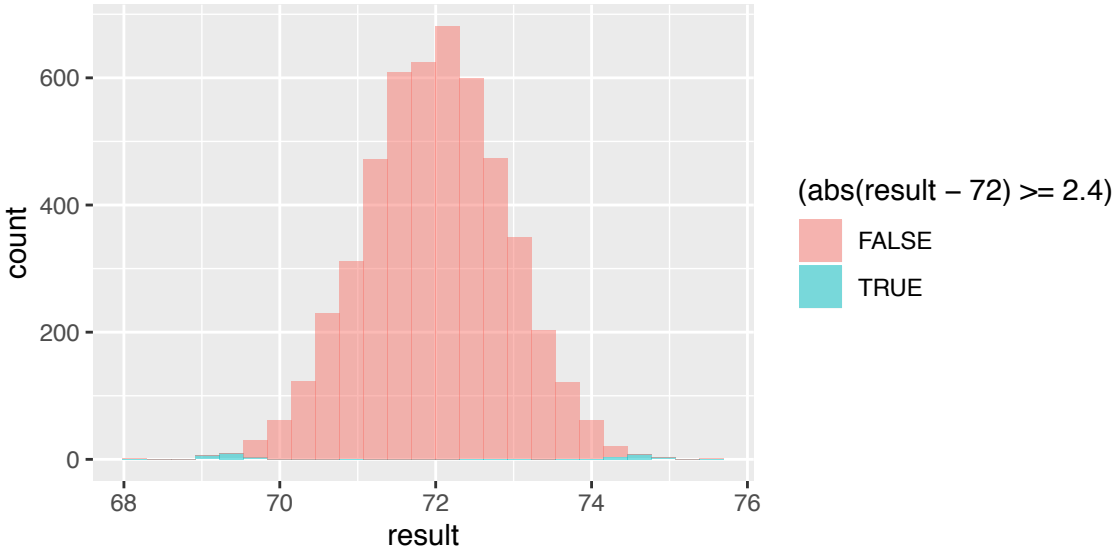
```
manyRandomizationXbars <- do(5000) * (mean(~Pulse, data=resample(BodyTemp50)) - 2.4)
head(manyRandomizationXbars)


  result
1  72.38
2  70.96
```

```
3  71.96
4  72.26
5  71.56
6  71.70
```

These results can be plotted to help us visualize our P-value:

```
gf_histogram(~result, data=manyRandomizationXbars, fill=~(abs(result-72) >= 2.4))
2 * nrow( filter(manyRandomizationXbars, result >= 74.4) ) / 5000
```

```
[1] 0.0064
```



If we are using significance level $\alpha = 0.05$, this P-value of 0.0064 represents a statistically significant result (since it is smaller that α), and we reject $\mathbf{H}_0$ in favor of the alternative.

■

Now, a null hypothesis is either true or it isn't. In deciding, we have computed a P-value which assesses our test result against the list of possibilities in a world where the null hypothesis is true. Our null distribution shows that a result like ours can happen in a world such as that, so we rejected the null with some uncertainty in our conclusion.

That's the reality of things. A null hypothesis is true or it is not. But only God can know for certain. We draw a conclusion to reject it or not, but under undertainty. Our conclusion may be mistaken.

Here is what can happen:

	We reject $\mathbf{H}_0$	We do not reject $\mathbf{H}_0$
$\mathbf{H}_0$ is true	Mistake (Type I error)	Success
$\mathbf{H}_0$ is false	Success	Mistake (Type II error)

In the analogy we have made between hypothesis tests and trials by jury, a Type I error is like

convicting an innocent defendent, whereas a Type II error is like failing to convict a guilty person. No one wants to commit this sort of mistake, and by adjusting the line between what we consider *reasonable doubt* and *beyond reasonable doubt*, we can affect how likely we are to make these mistakes. But when you demand stronger evidence to convict, you both

- make it less likely you'll commit Type I error, and
- more likely you'll commit Type II error.

It is like this with setting α. When you set $\alpha = 0.05$, you are declaring you are willing to live with rejecting a true null hypothesis in 5% of cases. Think that's too frequent? Then go ahead and set it lower (i.e., so as to require stronger evidence), maybe at $\alpha = 0.01$. But doing so comes at a cost: the likelihood that you will commit Type II error increases in the process. There is no getting around this. Many people have come to feel that $\alpha = 0.05$ is the right compromise.

Two independent samples vs. matched pairs

Consider this research question: Is it better to fish a certain lake from shore, or from a boat?

Our response variable will be quantitative, the ratio of fishing hours to fish caught. Here is some data.

month	Apr.	May	June	July	Aug.	Sept.	Oct.
shore	3.3	3.6	3.9	3.2	3.0	1.8	1.6
boat	3.8	3.0	3.3	2.2	1.6	1.4	1.5

We have a binary categorical explanatory variable: "Where fishing from?", with values "shore" and "boat". We have a quantitative response variable. We have bootstrapped and tested hypotheses for the difference $\mu_1 - \mu_2$, but the methods I've discussed have presumed *independent samples*. The data collected to investigate the question do not represent independent samples. The responses in the different months are naturally related: when one goes up, the other seems more likely to go up, both being related to the population of fish in the lake during that month. This data is **matched pairs** data, and we should:

- use the months as *cases*, and produce for each case a single difference:

month	Apr.	May	June	July	Aug.	Sept.	Oct.
shore	3.3	3.6	3.9	3.2	3.0	1.8	1.6
boat	3.8	3.0	3.3	2.2	1.6	1.4	1.5
difference	-0.5	0.6	0.6	1.0	1.4	0.4	0.1

- Proceed as if in a "single mean" setting. A confidence interval would be for the purpose of estimating the mean difference μ_{diff}. An hypothesis test would focus on hypotheses:

$$\mathbf{H}_0 \colon \mu_{\text{diff}} = 0 \qquad \text{vs.} \qquad \mathbf{H}_0 \colon \mu_{\text{diff}} \neq 0.$$

  Either way, the slips of paper we would insert into a bag for bootstrapping or randomization would contain only the last set of numbers, the differences.

Practice: Does the data suggest independent samples, warranting analysis on the difference of two means $\mu_1 - \mu_2$, or is it matched pairs?

1. A study was conducted to investigate the effectiveness of hypnotism in reducing pain. Results for randomly selected subjects are shown in the table below. A lower score indicates less pain.

subject	A	B	C	D	E	F	G	H
before	6.6	6.5	9.0	10.3	11.3	8.1	6.3	11.6
after	6.8	2.4	7.4	8.5	8.1	6.1	3.4	2.0

2. To study the effects of a drug on blood pressure, patients had a base reading taken of their diastolic blood pressure. After 3 weeks on the medication, new readings of their diastolic blood pressures were taken.

3. A collection of statistics students is randomly assigned to two groups. One group is given a study regimen that includes listening to recordings of classical music by Mozart, while the other group must study in silence. The response variable is student scores on an exam.