SESSIONS 4

Inferential Statistics: Hypothesis Testing Part 01

Data Science Program

Purwadhika
Startup and Coding School

Outline

- What is hypothesis
- Use Case : A/B Testing
- Hypothesis testing component
  - Hypothesis
  - Assumption
  - Test Statistics
  - P-value
  - Conclusion
- Different Kind of Hypothesis
  - Numerical (mean)
  - Categorical (proportion)

Purwadhika
Startup and Coding School

Components in Hypothesis Testing

- Hypothesis
- Assumption
- Test Statistics
- P-value
- Draw Conclusion

Illustration

When tossing a coin, we have **presumption of innocent** that the probability of each side would be 0.5 (face-tail).

- Scenario 1, if we throw 20 times, is it still reasonable face show 9 times ?
- Scenario 1, if we throw 20 times and we repeat it let's say 1000 times, is it still acceptable to say that the probability is 0.5 if we get face < 5 for 30 of 1000 trials ? how about 400 of 1000 trials ?
- How do you quantifying each scenario into numerical value ?
- Can we prove our presumption of innocent ?

Purwadhika
Startup and Coding School

What is Hypothesis?

A **Hypothesis** is a **statement about population**.

Ex. we have a hypothesis that each probability for each side of the coin is 0.5. Ho : **P** = 0.5 (the coin is balanced) vs Ha : **P** ≠ 0.5 (the coin is not balanced)

A Hypothesis are consist of Null Hypothesis (Ho) and Alternative Hypothesis (Ha).

Characteristic of Ho:

- Ho is **presumption of innocent**.
- Usually contain a particular value.

  ex. Ho : **μ** = 50, Ho : **P** ≥ 0.5
- Usually takes no effect.

  ex. Ho : Beta = 0, Ho : dice is balance

Characteristic of Ha:

- Usually takes no certain value/ interval.

  ex. Ha : **μ** ≠ 50, Ha : **P** < 0.5
- Usually takes any effect.

  ex. Ha : Beta < 0, Ha : Beta > 0, Ha : dice is not balance.

	Layout A	Layout B
Visitors	122	118
Customers	22	25
Conversion %	18.0%	21.2%

A/B Testing is hypothesis testing in disguise

Principles of A/B Testing:

- Hypothesis : P.I.C.O.T
- Randomization : bias and covariate
- Sample size : minimal sample size needed
- Method of measurement : categorical data or numerical data

Purwadhika
Startup and Coding School

P.I.C.O.T

Population, **I**ntervention, **C**omparison, **O**utcome, **T**ime

Good Example

- **H0**: Amazon.com visitors that receive Layout B will not have higher end-of-visit conversion rates compares to visitors that receive Layout A
- **H1**: Amazon.com visitors that receive Layout B will have higher end-of-visit conversion rates compared to visitors that receive layout A

Bad Example

- **H0:** Banks with nicer colors will not affect loan repayment
- **H1:** Banks with nicer colors will affect loan repayment

Purwadhika
Startup and Coding School

- **H0**: Amazon.com visitors that receive Layout B will not have higher end-of-visit conversion rates compares to visitors that receive Layout A
- **H1**: Amazon.com visitors that receive Layout B will have higher end-of-visit conversion rates compared to visitors that receive layout A

- **P**opulation : visitors of Amazon.com
- **I**ntervention : new layout (layout B)
- **C**omparison : visitors receiving Layout A
- **O**utcome : Conversion rate
- **T**ime : End of visit to Amazon.com

- **H0:** Banks with nicer colors will not affect loan repayment
- **H1:** Banks with nicer colors will affect loan repayment

- Poor **I**ntervention definition : no clear definition of nicer colors
- Poor Population definition : What banks, where and what level ? bank at certain city
- Poor Outcome definition : what effect are you measured, loan default rates, days past due, total branch loss ?

Purwadhika
Startup and Coding School

Randomness

Let’s analyze data gathered from a hospital.

	Heart Disease	No Heart Disease	Total
Age > 50	57.8 %	42.2 %	100%
Age < 50	88.8 %	11.2 %	100%

We are talking about this example again

- Notice that young people have higher proportion suffer from heart disease 88.8%.
- Is that make sense that younger people have tend to have higher risk of suffering from heart disease ?

This data is not representative and not gathered randomly:

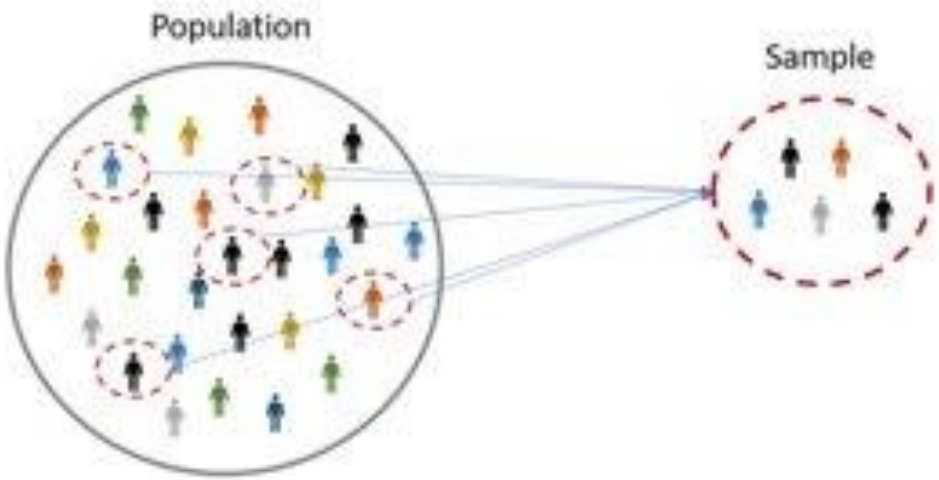
1. This sample is **biased**
2. There is an unmeasured variable (**covariates**) which is self awareness for young people lower than old people
3. self awareness (covariate) affect the outcome not the one we want to know (Young/Old)

Purwadhika
Startup and Coding School

Which one is Random?

1. Participants are allowed to decide whether to be in a treatment or control group
2. Any participant with a national ID number ending in an odd number is assigned to treatment, any participant with an ID ending in an even number is assigned to control.
3. Participants east of my office are assigned to control, participants west of my office are assigned to treatment
4. We flip a coin to decide whether a participant is control or treatment

Sample



Why need sample?

- Resource
- Time
- Cost

Let N the population size and the margin of error e denotes the allowed probability of committing an error in selecting a small representative of the population. The sample size n can be obtained by the formula

$$n = \frac{N}{1 + Ne^2}$$

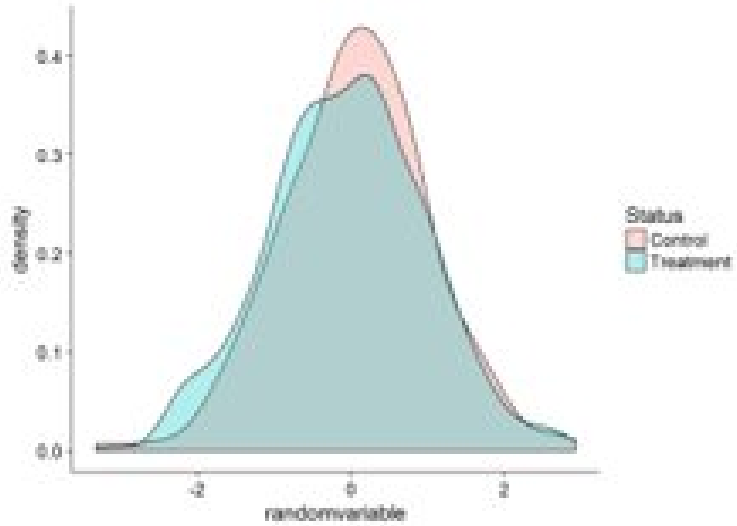
Purwadhika
Startup and Coding School

Measurements

CATEGORICAL VARIABLE

Conversion rate or proportion / sum or count

NUMERICAL VARIABLE



Distribution (mean, median, percentile)
/ difference in ratios

Purwadhika
Startup and Coding School

Which Method to Use?

Different hypothesis testing usually refer to:

- Type of Data used
- Sample Size
- Assumption
  - The form of population distribution
  - Variance
  - Sampling method
  - Randomization
  - etc

Different Assumption will leads to different method

Test Statistics

Test Statistics follow certain distribution and measure how far point estimate fall from the parameter in the Ho.

Test Statistics depend on what method we used to test the hypothesis

Some Popular test statistic:

- t score, used in t-test for mean population test
- z score, used in z-test for proportion population test
- F score, used in one way ANOVA when we test mean from > 2 population

In General Test Statistics are consist of ratio between magnitude and variability

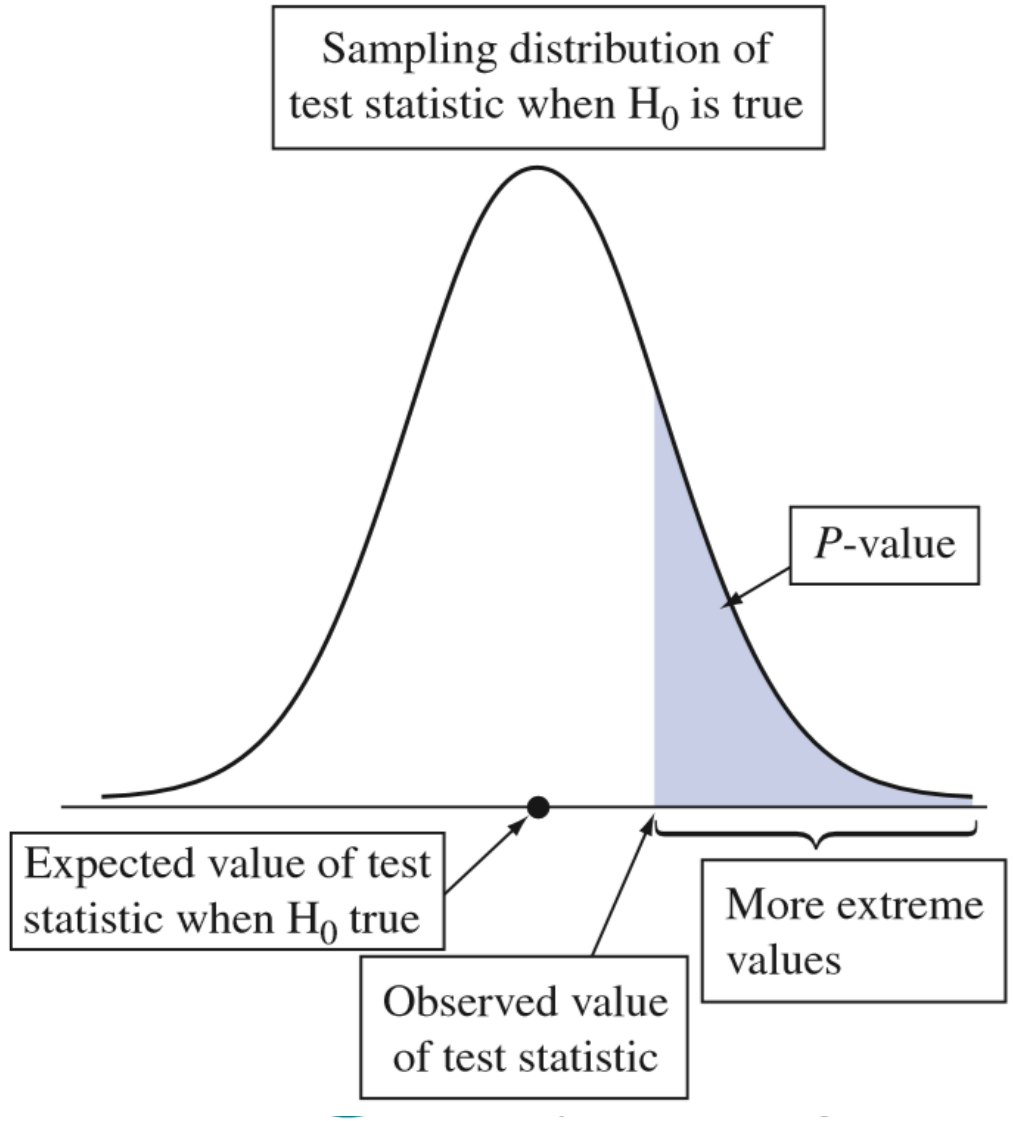
- magnitude (ex. mean different, proportion different)
- variability (ex. standard error)

P-Value

In Python and any statistical software, every test statistics will be converted into p-value (risk/probability). So, P-value is used to interpret any test statistics

P-Value is risk/probability to reject Ho (**μ** = 50) while the population actually Ho (**μ** = 50) the based on sample data.

We reject Ho when P-value is equal to or below certain level of risk (Type of error I or **alpha**, α). Level of risk usually 0.1, 0.05, 0.01.



Type of Error

Type I error

- Rejects a null hypothesis when it is true.
- The probability of committing a Type I error is called the **significance level** (**alpha**, α)

Type II error.

- Can't happened in real life.
- Fails to reject a null hypothesis that is false.
- The probability of committing a Type II error is called **Beta**, β
- The probability of *not* committing a Type II error is called the **Power** of the test.

Population	Sample	
	Accept Ho	Reject Ho
Ho True	Correct Conclusion	**Type I error** Reject True Ho
Ho False	**Type II error** Accept False Ho	Correct Conclusion

Purwadhika
Startup and Coding School

Drawing Conclusion

Final decision of hypothesis testing is whether we reject or not reject Ho.

- When we **reject Ho** we conclude that we have **enough evidence** that Ho is wrong below certain level of risk. P-Value and test statistic are the quantification of evidence.
- When we **do not reject Ho** it doesn't mean that Ho is true because **Ho** is **a presumption of innocent**. we just don't have **enough evidence** to proof that Ho is wrong. As an alternative for interpretation you can use interval estimate.

Let's say we have hypothesis about mean population of Bogor citizen's age is below 33 or not.

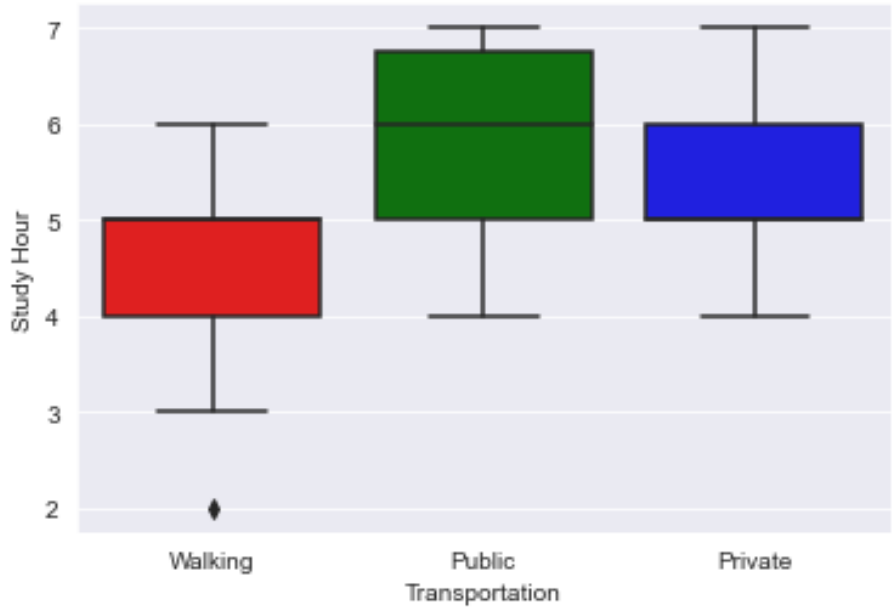
Ho : $\mu = 33$ vs Ha : $\mu < 33$

If we have P-value 0.0005. It means that the probability we reject Ho (age mean population equal to 33) while the Ho True (age mean population equal to 33) is 0.0005. very low risk and we have enough evidence

If we have P-value 0.5. We accept Ho or fail to reject Ho. We can't say age mean population is equal to 33 Because we never know the actual value of age mean population and a hypothesis is a statement about population.

Why Do We Need Hypothesis Testing?

For example, we use hypothesis testing to compare mean between populations. Let's say we want to compare means of study hour between student who walking, using public transport, and using private transportation. here is the graphical summary.



The problem is **descriptive statistics** like this is **vulnerable to subjectivity**. I may say "no different" between student who use private and public. what about everyone else ?

Purwadhika
Startup and Coding School

Why do we need A/B testing?

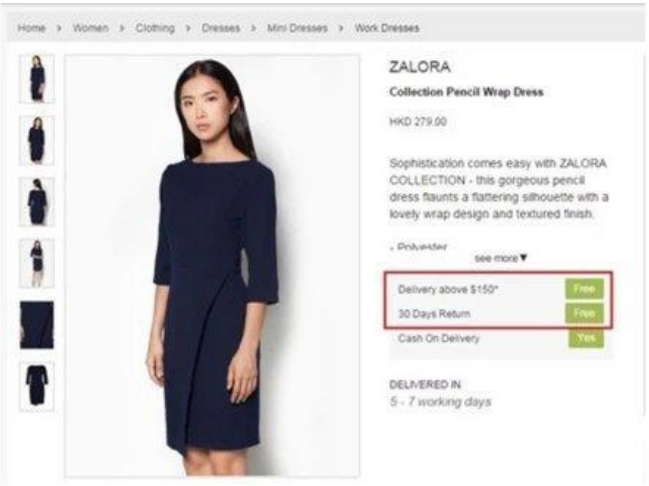
With causality we can finally lay to rest the “correlation vs causation” argument , and prove that our new product actually works. As we know that correlation doesn’t imply causation. with A/B testing we proof statistically.

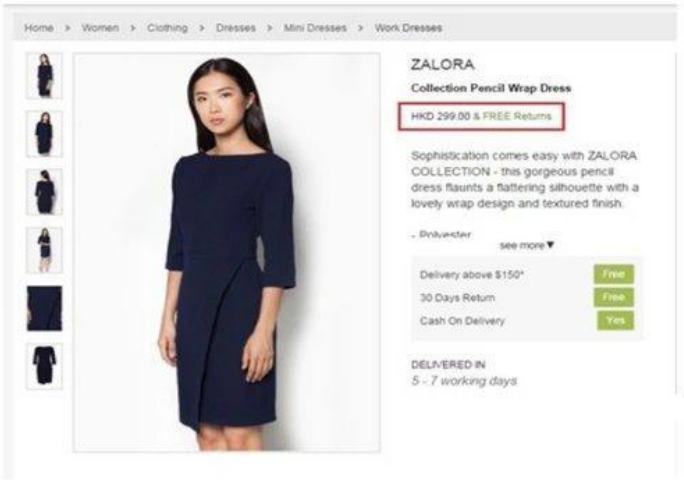
	Layout A	Layout B
Visitors	122	118
Customers	22	25
Conversion %	18.0%	21.2%

“Based on our testing, we infer that **new layout** cause **increase** in **conversion rate**”

Purwadhika
Startup and Coding School

Some Proof in Real Cases: Zalora



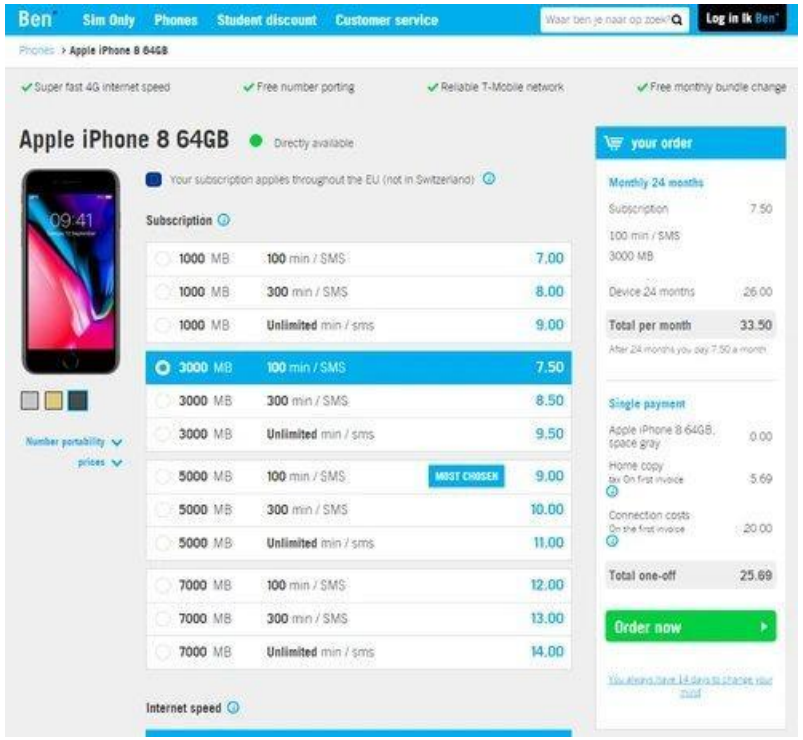


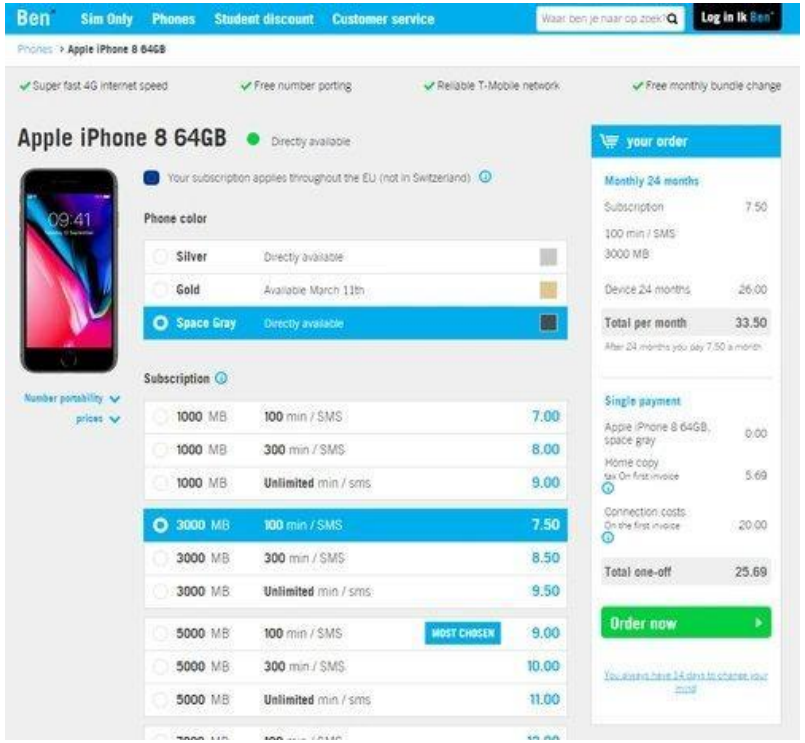
Zalora found that Variation 1 outperformed the control and its counter variation 2. By simply bringing uniformity in Zalora's call to action button, the e-Commerce giant saw an increase of 12.3% in its checkout rate.

Read more at: https://vwo.com/blog/ab-testing-examples/?utm_campaign=tof_ugc

Purwadhika
Startup and Coding School

Some Proof in Real Cases: Ben



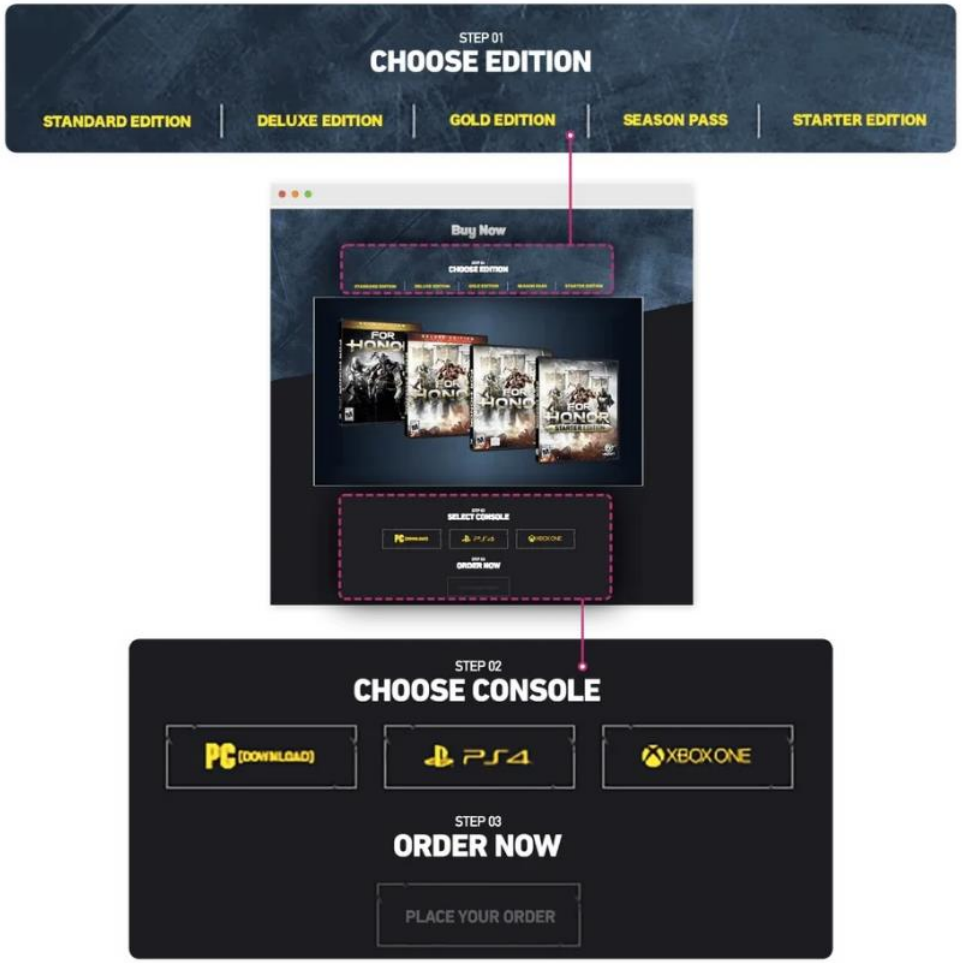


Ben ran the experiment for about two weeks and found that by simply making their color palette more prominently visible, it's conversions went up by 17.63% and the number of customer calls to change device colors dropped significantly.

Read more at: https://vwo.com/blog/ab-testing-examples/?utm_campaign=tof_ugc

Purwadhika
Startup and Coding School

Some Proof in Real Cases: Ubisoft



Chose game edition and console (PC, PS4, or X-Box) separately

Purwadhika
Startup and Coding School

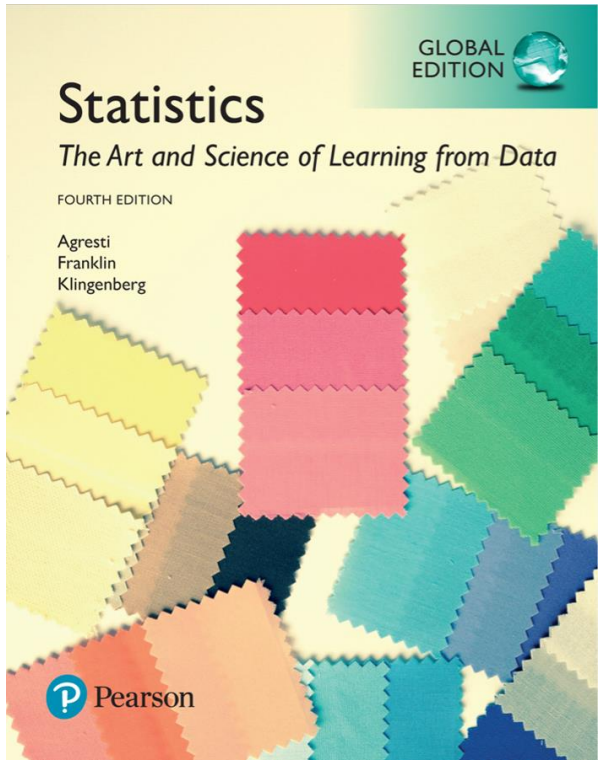
Some Proof in Real Cases: Ubisoft

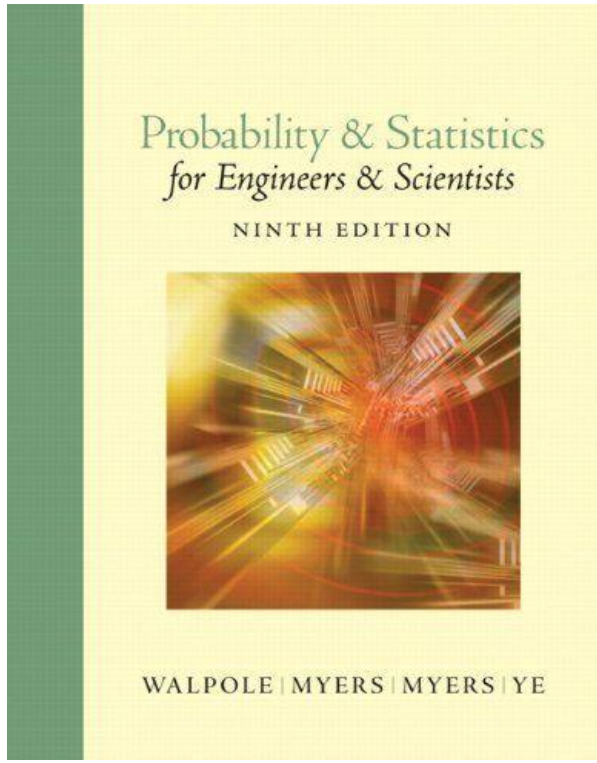
Simplify the buying process

Post running the test for about three months, Ubisoft saw that variation brought about more conversions to the company than the control. Conversions went up from 38% to 50% conversions, and overall lead generation increase by 12%.

Purwadhika
Startup and Coding School

Reference





Purwadhika
Startup and Coding School

Reference

https://towardsdatascience.com/data-science-you-need-to-know-a-b-testing-f2f12aff619a

https://towardsdatascience.com/data-science-fundamentals-a-b-testing-cb371ceecc27

https://www.niagahoster.co.id/blog/ab-testing-adalah/

https://vwo.com/blog/ab-testing-examples/

https://www.scribbr.com/methodology/sampling-methods/

Purwadhika
Startup and Coding School