DESIGN *of* EXPERIMENTS

Statistical foundations for causal inference

Lukas Vermeer is Senior Product Owner for Experimentation at Booking.com

@lukasvermeer (http://twitter.com/lukasvermeer) — lukasvermeer.github.io/ab-stats (http://lukasvermeer.github.io/ab-stats)

Making good decisions requires

CAUSAL INFERENCE

Umbrellas often appear just before it pours, but banning them will not stop the rain; it will just make everyone **more wet**

RUBIN CAUSAL MODEL (WHAT WE WANT TO KNOW)

	Under A	Under B	Treatment Effect
Alice	0.4	0.9	0.5
Bob	0.3	0.8	0.5
Charlie	0.5	1	0.5
Dave	0.2	0.7	0.5
Eve	0.4	0.9	0.5
Frank	0.1	0.6	0.5
(Everyone)	...	...	...
Average	0.3	0.8	**0.5**

[1] Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” Journal of Educational Psychology 66 (5): 688–701.

The fundamental problem of

CAUSAL INFERENCE

1. We cannot sample the entire **population**
2. We cannot expose units to both treatments **exclusively**
3. We cannot directly observe **underlying** probabilities

RUBIN CAUSAL MODEL (WHAT WE CAN MEASURE)

	Under A	Under B	Treatment Effect
Alice	No	?	?
Bob	?	Yes	?
Charlie	?	Yes	?
Dave	No	?	?
Eve	?	Yes	?
Frank	No	?	?
(Sample)	...	...	...
Average	?	?	?

[1] Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” Journal of Educational Psychology 66 (5): 688–701.

We will try to answer two

KEY QUESTIONS

1. **Is there** any causal effect?
2. What is the **size** of the causal effect?

Our answers will be correct

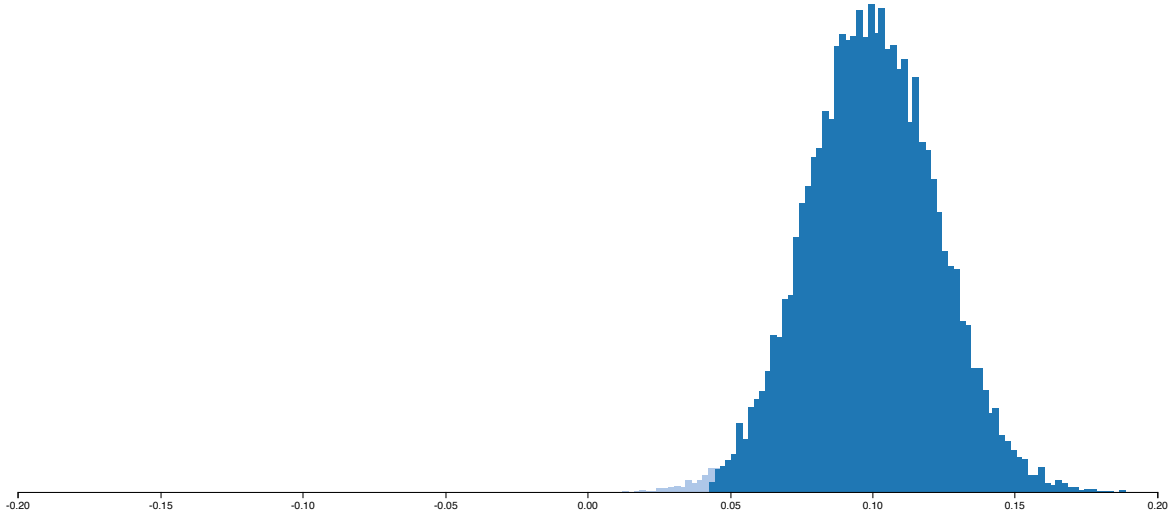
IN EXPECTATION

If we show A and B to random samples of the population, then the fraction of yes in both groups converges to the true underlying means, and the difference between the means converges to the **average treatment effect**

RUBIN CAUSAL MODEL (WHAT RANDOMIZATION GIVES US)

	Under A	Under B	Treatment Effect
Alice	No	?	?
Bob	?	Yes	?
Charlie	?	Yes	?
Dave	No	?	?
Eve	No	?	?
Frank	?	No	?
(Sample)	...	...	...
Average	0.0	0.7	**0.7**

[1] Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” Journal of Educational Psychology 66 (5): 688–701.



simulated effect	average observed effect
0.1	0.100

n = 1000, base conversion rate = 0.1, effect of treatment = 0.1

Randomization ensures only three things can

EXPLAIN A DIFFERENCE

1. **Causation** resulted in people behaving differently when treatment was applied
2. **Pure chance** resulted in a difference between the two groups unrelated to the treatment
3. **Measurement error** resulted in an unintended difference in results unrelated to user behaviour

We want to reject the

NULL HYPOTHESIS

The **null hypothesis** assumes no difference between treatment and control; any difference we observe is simply due to chance.

If we could reasonably rule out chance, we might reject the null and consider this to be evidence for the **alternative hypothesis**.

We compute a

P-VALUE

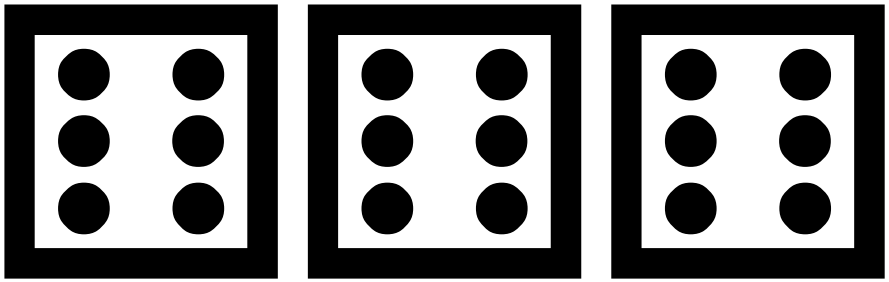
We assume the null hypothesis is true and compute the p-value.

Assuming there is no effect, the p-value is the probability of seeing a particular result or more extreme by chance.

How likely is this result assuming the null is true?

Is this die

FAIR?



p = 0.00462962

(Not fair; I cheated)

We need to pick a

THRESHOLD

One swallow does not a summer make, nor one fine day, but how many swallows do we count before we pack away our umbrellas?

Scientific standard for significance: $p < 0.05$

p-values are often[2]

MISINTERPRETED

Some examples of **incorrect** interpretations

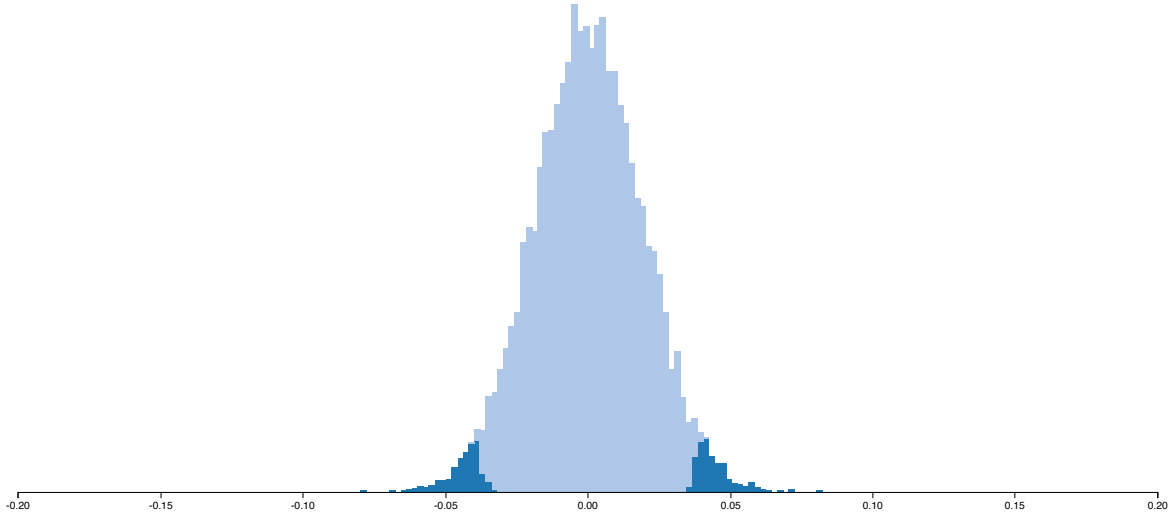
1. p = .05 means the null hypothesis has only a 5% chance of being true
2. A non-significant difference (e.g. p > .05) means there is no difference between groups
3. p = .05 means that we have observed data that would occur only 5% of the time under the null

[2] Goodman, Steve 2008. “A dirty dozen: twelve P-value misconceptions.” Seminars in Hematology, 45 (2008), pp. 135-140.

Two types of

ERRORS

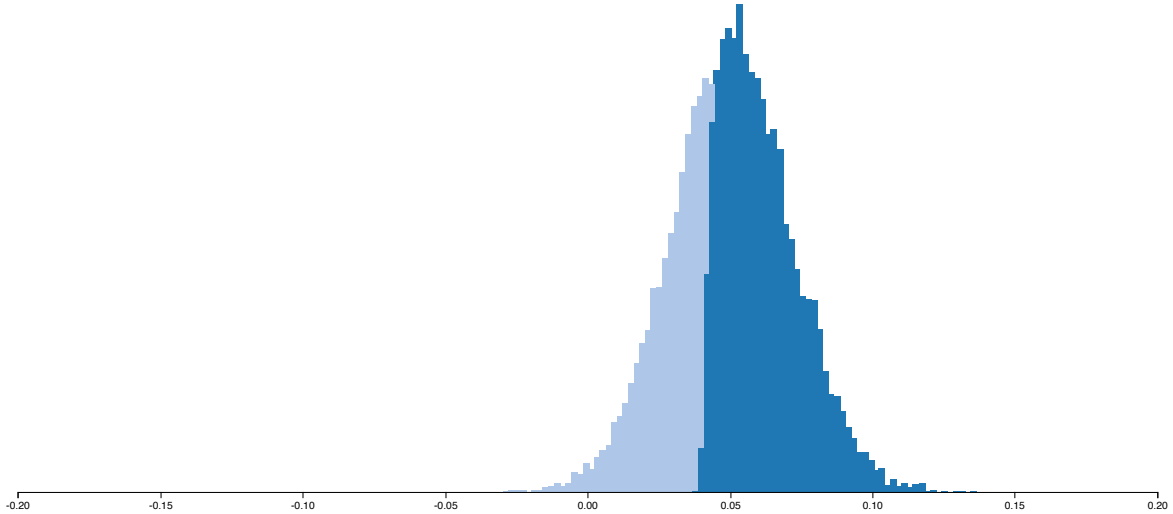
1. **Type-I** is the incorrect rejection of a true null hypothesis; we cried wolf when there was none
2. **Type-II** is the failure to reject a false null hypothesis; we failed to detect a real effect



simulated effect **type-I error rate**

0 5%

n = 1000, base conversion rate = 0.1, effect of treatment = 0



simulated effect	type-II error rate
0.05	33%

n = 1000, base conversion rate = 0.1, effect of treatment = 0.05

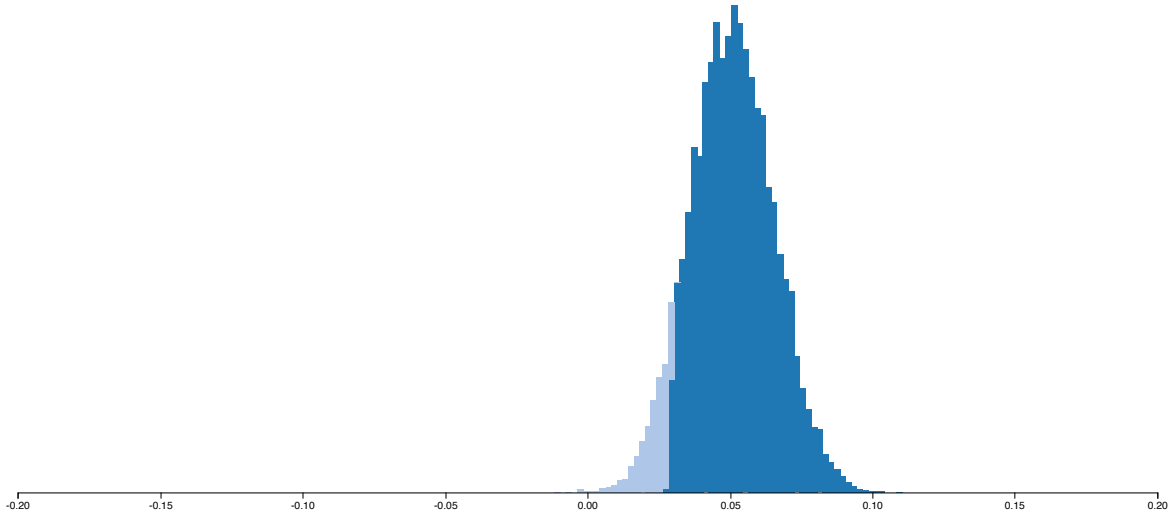
The importance of

STATISTICAL POWER

Statistical power is the probability that the test **correctly rejects** the null hypothesis when the alternative hypothesis is true

Two main things affect statistical power:

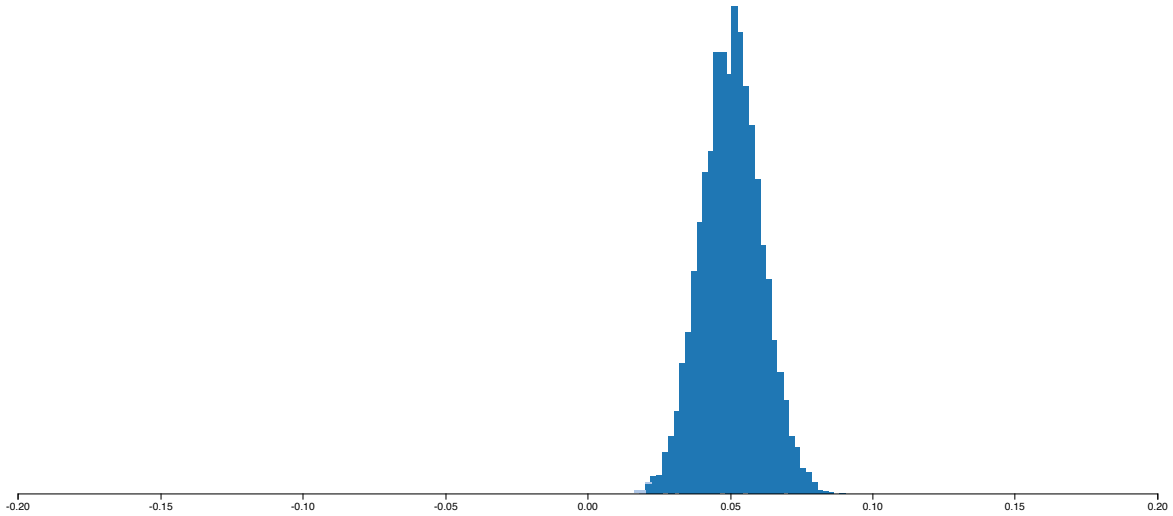
- Sample size (more is better)
- Effect size (more is better)



type-II error rate **observed power**

7% 93%

n = 2000, base conversion rate = 0.1, effect of treatment = 0.05



type-II error rate **observed power**

0% 100%

n = 4000, base conversion rate = 0.1, effect of treatment = 0.05

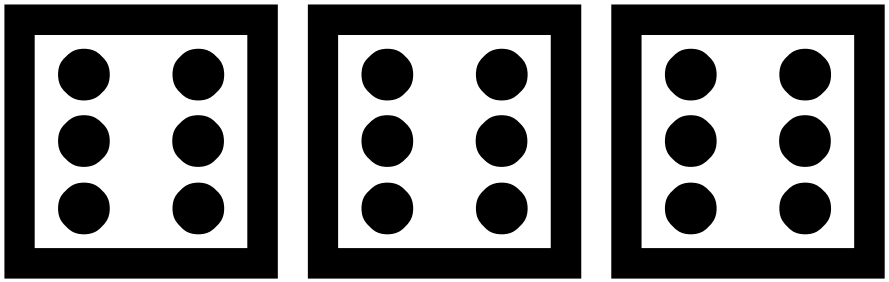
The importance of sticking to

PROTOCOL

The methods described assume strict adherence to protocol; violations of protocol such as **peeking** and **multiple testing** increase the type-I error rate

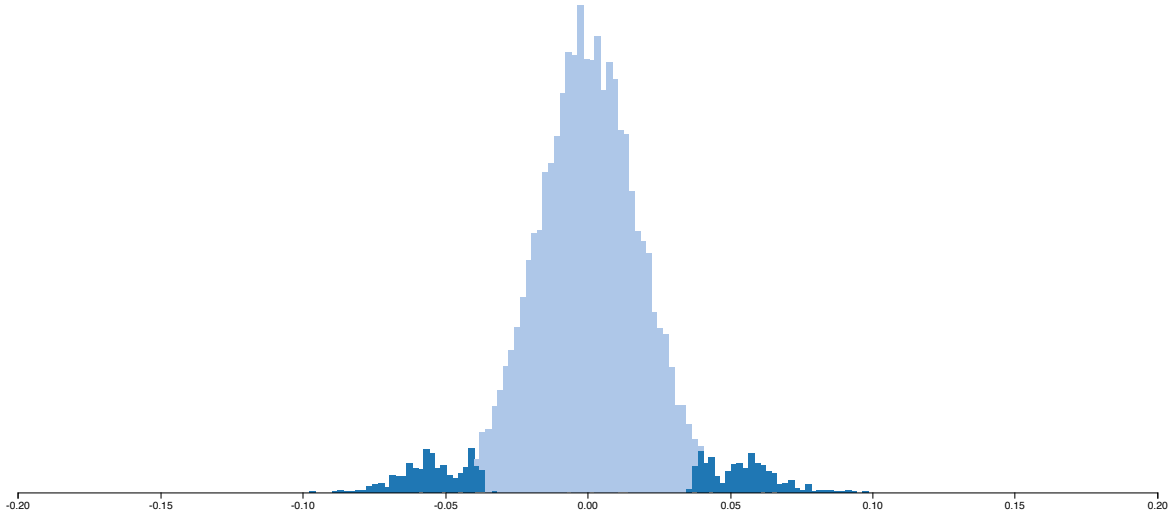
Is this die

FAIR?



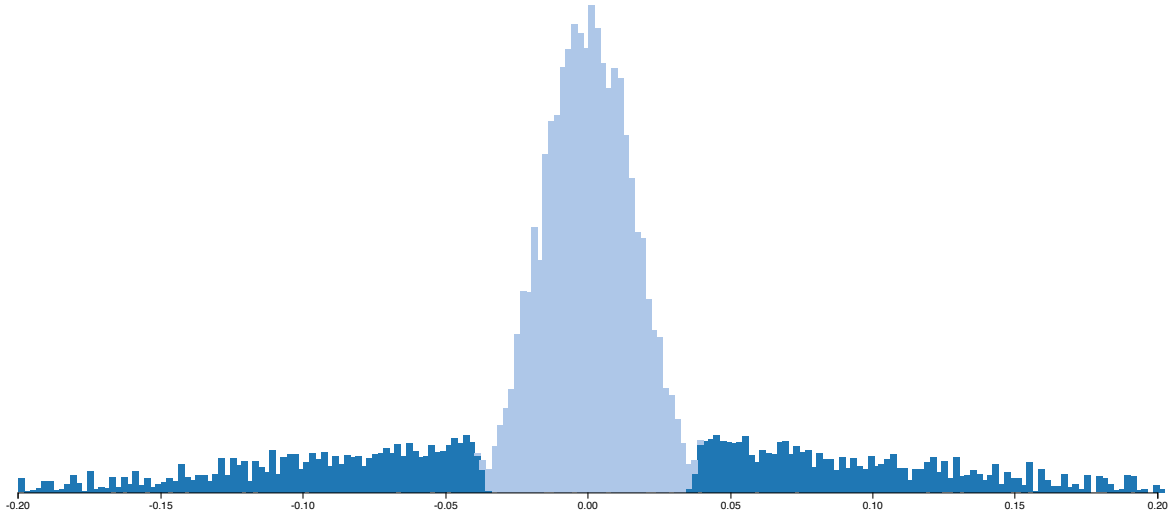
$p = 1$

(Fair die; I still cheated)



simulated effect	type-I error rate
0	9%

n = 1000, base conversion rate = 0.1, effect of treatment = 0, **peeking 2 times**



simulated effect	type-I error rate
0	32%

n = 1000, base conversion rate = 0.1, effect of treatment = 0, **peeking 100 times**

Reasons for violating

PROTOCOL

More flexible protocols may be desirable

- **early stopping rules** to mitigate damage
- **early shipping** to minimize opportunity cost
- **multiple variants** to test several alternatives
- **multiple metrics** to guard business KPIs

All these are possible[4], but require protocol adjustments

[4] Alex Deng, Tianxi Li, Yu Guo 2014 “Statistical Inference in Two-Stage Online Controlled Experiments with Treatment Selection and Validation” WWW '14. 609–618.

DESIGN *of* EXPERIMENTS

Statistical foundations for causal inference

Lukas Vermeer is Senior Product Owner for Experimentation at Booking.com

@lukasvermeer (http://twitter.com/lukasvermeer) — lukasvermeer.github.io/ab-stats (http://lukasvermeer.github.io/ab-stats)

REFERENCES

1. Rubin, Donald B. 1974. “Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies.” Journal of Educational Psychology 66 (5): 688–701. (link (http://dx.doi.org/10.1037/h0037350))
2. Goodman, Steve 2008. “A dirty dozen: twelve P-value misconceptions.” Seminars in Hematology, 45 (2008), pp. 135-140. (link (https://www.researchgate.net/publication/5272766_A_Dirty_Dozen_Twelve_P-Value_Misconceptions))
3. Kohavi, R., Longbotham, R., Sommerfield, D. et al. 2009 “Controlled experiments on the web: survey and practical guide” Data Min Knowl Disc (2009) 18: 140. (link (http://bit.ly/expSurvey))
4. Alex Deng, Tianxi Li, Yu Guo 2014 “Statistical Inference in Two-Stage Online Controlled Experiments with Treatment Selection and Validation” WWW '14. 609–618.(link (http://www.exp-platform.com/Documents/p609-deng.pdf))