

Randomized experiments, A/B tests and sequential monitoring

Steve Howard

April 26, 2018

Guess the winner!

Doctor FootCare™ 🇺🇸 Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | 1-888-211-9733

Shop With Confidence

- ✓ Satisfaction Guaranteed
- ✓ 30-day, hassle-free Returns
- ✓ 100% Safe, **Secured** shopping
- ✓ We assure your Privacy

100% Secured Checkout Continue Shopping Proceed To Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	<input type="text" value="1"/>		\$0.00	\$0.00
					Total: \$0.00

Select Shipping Method

100% Secured Checkout Continue Shopping Proceed To Checkout

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | [Shipping Cart](#)

Doctor FootCare™ 🇺🇸 Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | 1-888-211-9733

Shop With Confidence

- ✓ Satisfaction Guaranteed
- ✓ 30-day, hassle-free Returns
- ✓ 100% Safe, **Secured** shopping
- ✓ We assure your Privacy

100% Secured Checkout Continue Shopping Proceed To Checkout

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	<input type="text" value="1"/>		\$0.00	\$0.00
					Discount: \$0.00
					Total: \$0.00

Enter Coupon Code

Select Shipping Method

100% Secured Checkout Recalculate Continue Shopping Proceed To Checkout

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | [Shipping Cart](#)

From Kohavi, Longbotham, et al. (2009).

Guess the winner!

Doctor FootCare™ Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | 888-211-8733

Shop With Confidence

- ✓ Satisfaction Guaranteed
- ✓ 30-day, hassle-free Returns
- ✓ 100% Safe, **Secured** shopping
- ✓ We assure your Privacy

100% Secured Checkout [Continue Shopping](#) [Proceed To Checkout](#)

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	<input type="text" value="1"/>		\$0.00	\$0.00
					Total: \$0.00

Select Shipping Method:

100% Secured Checkout [Continue Shopping](#) [Proceed To Checkout](#)

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | [Shopping Cart](#)

Doctor FootCare™ Shopping Cart

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | 888-211-8733

Shop With Confidence

- ✓ Satisfaction Guaranteed
- ✓ 30-day, hassle-free Returns
- ✓ 100% Safe, **Secured** shopping
- ✓ We assure your Privacy

100% Secured Checkout [Proceed To Checkout](#)

Item Name	Item Number	Quantity	Remove	Unit Price	Subtotal
Trial Kit	FFCS	<input type="text" value="1"/>		\$0.00	\$0.00
					Discount: \$0.00
					Total: \$0.00

Enter Coupon Code:

Select Shipping Method:

100% Secured Checkout [Continue Shopping](#) [Proceed To Checkout](#)

Home | Products | Learn More | Tips | Testimonials | FAQ | About Us | Contact Us | [Shopping Cart](#)

10x revenue!

From Kohavi, Longbotham, et al. (2009).

Guess the winner!

Find a new home or apartment

Existing Homes
From REALTOR.com®

Foreclosures
From RealtyTrac.com™

New Homes
From Move.com™

Rentals
From Move.com™

Price Range: \$0 to No Maximum

Enter City Select a State

Or Enter ZIP

Senior Living Home Plans

Control

Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

 Enter City State

or

Enter Zip

Treatment1

Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

 Enter City State

or

Enter Zip

Treatment 2

What are you looking for?

Existing Homes

New Construction

Rentals

Foreclosures

Senior Living

Home Valuation

Professional Services

Enter City State

Enter Zip

\$0 to No Max

Condos/Townhouse Single Family Home

Treatment 3

Find a new Home or Apartment

Existing Homes

New Construction

Foreclosures

Rentals

Enter Zip or Enter City State

Treatment 4

Find Your Dream Home or Apartment

City, State or ZIP

Existing homes

Foreclosures

New construction

Rentals

Treatment 5

Guess the winner!

Find a new home or apartment

Existing Homes from REALTOR.com®

Foreclosures from RealtyTrac.com™

New Homes from Move.com™

Rentals from Move.com™

Price Range: \$0 to No Maximum

Enter City Select a State

Or Enter ZIP

• Senior Living • Home Plans

Control

Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

 Enter City State

or

Enter Zip

Treatment 1

Existing Homes Foreclosures New Construction Rentals

Find Existing Homes for Sale

 Enter City State

or

Enter Zip

Treatment 2

What are you looking for?

Existing Homes

New Construction

Rentals

Foreclosures

Senior Living

Home Valuation

Professional Services

Enter City State

Enter Zip

\$0 to No Max

Condo/Townhouse Single Family Home

Treatment 3

Find a new Home or Apartment

 Existing Homes

 New Construction

 Foreclosures

 Rentals

Enter Zip or Enter City State

Treatment 4

Find Your Dream Home or Apartment

City, State or ZIP

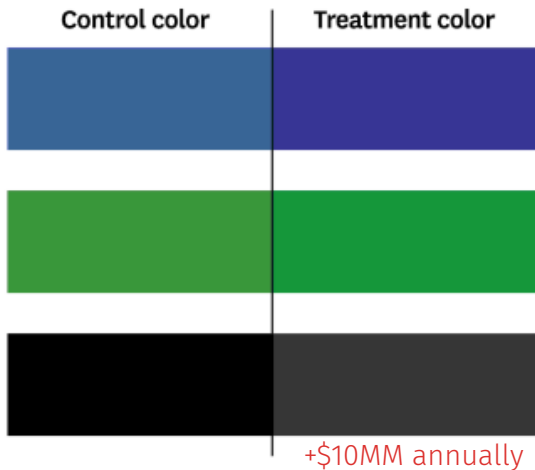
Existing homes New construction

Foreclosures Rentals

Treatment 5

+10% revenue

Yes, the color thing is real.



FROM "THE SURPRISING POWER OF ONLINE
EXPERIMENTS," SEPTEMBER-OCTOBER 2017,
BY RON KOHAVI AND STEFAN THOMKE

© HBR.ORG

For years, Microsoft, like many other companies, had relied on expert designers—rather than the behavior of actual users—to define corporate style guides and colors.

From Kohavi and Thomke (2017).

“Which version is better?” is a thorny question.

- What I'd really like to know:
 1. What would happen if I were to show everyone version A?
 2. What would happen if I were to show everyone version B?

“Which version is better?” is a thorny question.

- What I'd really like to know:
 1. What would happen if I were to show everyone version A?
 2. What would happen if I were to show everyone version B?
- A simpler version:
 1. What would happen if I were to show *you* version A?
 2. What would happen if I were to show *you* version B?

“Which version is better?” is a thorny question.

- What I'd really like to know:
 1. What would happen if I were to show everyone version A?
 2. What would happen if I were to show everyone version B?
- A simpler version:
 1. What would happen if I were to show *you* version A?
 2. What would happen if I were to show *you* version B?
- I can never (reliably) answer this question!

“Which version is better?” is a thorny question.

- What I'd really like to know:
 1. What would happen if I were to show everyone version A?
 2. What would happen if I were to show everyone version B?
- A simpler version:
 1. What would happen if I were to show *you* version A?
 2. What would happen if I were to show *you* version B?
- I can never (reliably) answer this question!

“The fundamental problem of causal inference”

Carefully designed,
randomized controlled experiments are the
only reliable way to learn what works best.

Sometimes the only thing you can do with a poorly designed experiment is to try to find out what it died of. (Fisher)

From Box, J. S. Hunter, and W. G. Hunter (2005).

1. The need for randomized experiments

- Prediction, estimation, and causal inference
- The two benefits of randomization

2. Design choices

3. Sequential experimentation

Prediction, estimation, and causal inference

Prediction, estimation, and causal inference: a coarse classification of statistical problems

Suppose I have data on birth weights at a certain hospital, and whether each mother smoked.

Prediction, estimation, and causal inference: a coarse classification of statistical problems

Suppose I have data on birth weights at a certain hospital, and whether each mother smoked.

A prediction problem:

Can you predict the birth weight of the next baby, given the mother's smoking status and other info?

Use any algorithm we want, check accuracy on held-out data.

Prediction, estimation, and causal inference: a coarse classification of statistical problems

Suppose I have data on birth weights at a certain hospital, and whether each mother smoked.

An estimation problem:

What is the (adjusted) difference in birth weight between smokers and nonsmokers, in the population?

How precise is that estimate?

Need a probability model.

Prediction, estimation, and causal inference: a coarse classification of statistical problems

Suppose I have data on birth weights at a certain hospital, and whether each mother smoked.

A causal inference problem:

What will be the effect on birth weight of telling mothers to stop smoking?

Need two groups of mothers **similar except for treatment**.

Prediction, estimation, and causal inference: a coarse classification of statistical problems

Suppose I have data on birth weights at a certain hospital, and whether each mother smoked.

A causal inference problem:

What will be the effect on birth weight of telling mothers to stop smoking?

Need two groups of mothers **similar except for treatment**.

Randomized assignment yields such groups.

The two benefits of randomization

First benefit of randomization: freedom from bias

Designing an experiment is like gambling with the devil: Only a random strategy can defeat all his betting systems. (Fisher)

From Box, J. S. Hunter, and W. G. Hunter (2005).

Table 2. A study of 51 studies on the portacaval shunt. The well-designed studies show the surgery to have little or no value. The poorly-designed studies exaggerate the value of the surgery.

<i>Design</i>	<i>Degree of enthusiasm</i>		
	<i>Marked</i>	<i>Moderate</i>	<i>None</i>
No controls	24	7	1
Controls, but not randomized	10	3	2
Randomized controlled	0	1	3

Source: N. D. Grace, H. Muench, and T. C. Chalmers, "The present status of shunts for portal hypertension in cirrhosis," *Gastroenterology* vol. 50 (1966) pp. 684–91.

From Freedman, Pisani, and Purves (2007).

Table 4. A study of studies. Four therapies were evaluated both by randomized controlled trials and by trials using historical controls. Conclusions of trials were summarized as positive (+) about the value of the therapy, or negative (-).

<i>Therapy</i>	<i>Randomized controlled</i>		<i>Historically controlled</i>	
	+	-	+	-
Coronary bypass surgery	1	7	16	5
5-FU	0	5	2	0
BCG	2	2	4	0
DES	0	3	5	0

Note: 5-FU is used in chemotherapy for colon cancer; BCG is used to treat melanoma; DES, to prevent miscarriage.

Source: H. Sacks, T. C. Chalmers, and H. Smith, "Randomized versus historical controls for clinical trials," *American Journal of Medicine* vol. 72 (1982) pp. 233-40.⁷

From Freedman, Pisani, and Purves (2007).

Second benefit of randomization: a “reasoned basis for inference”

By putting known randomness into the world,
we justify probability calculation *by design*.

An idea due to Fisher. Also known as
“putting a rabbit into the hat”. (Freedman)

Without randomization, probabilities are justified
purely *by a model*.

Don't fall in love with a model.

From Box, J. S. Hunter, and W. G. Hunter (2005).

1. The need for randomized experiments

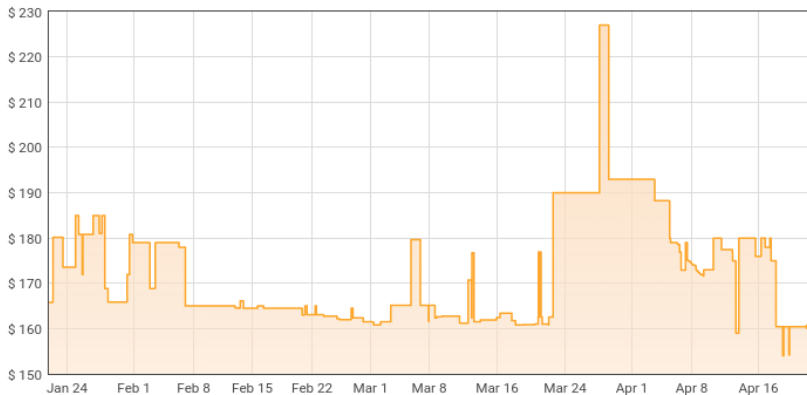
2. Design choices

- Choosing the unit of randomization
- Choosing who to enroll
- Choosing an outcome metric

3. Sequential experimentation

Choosing the unit of randomization

Pricing is a tricky thing to experiment on.



From [keeper.com](https://www.keeper.com)

How would you run a pricing experiment?

- Randomize by session?

How would you run a pricing experiment?

- Randomize by session?
- Randomize by user?

How would you run a pricing experiment?

- Randomize by session?
- Randomize by user?
- Randomize by product?

How would you run a pricing experiment?

- Randomize by session?
- Randomize by user?
- Randomize by product?
- Randomize by product category?

How would you run a pricing experiment?

- Randomize by session?
- Randomize by user?
- Randomize by product?
- Randomize by product category?
- Randomize by day?

How would you run a pricing experiment?

- Randomize by session?
- Randomize by user?
- Randomize by product?
- Randomize by product category?
- Randomize by day?

The right unit of randomization is sometimes not obvious!

The unit of analysis should be the same as the unit of randomization.

Whatever your unit of randomization,

- compute one summary outcome per unit, and
- analyze results with these outcomes.

The unit of analysis should be the same as the unit of randomization.

Whatever your unit of randomization,

- compute one summary outcome per unit, and
- analyze results with these outcomes.

Sample size = number of randomized units!

Making the unit of analysis differ from the unit of randomization is dangerous.

Be wary of finer-grained analysis, e.g.,

- randomizing by city, analyzing by user.
- randomizing by category, analyzing by product.

It can be done, but requires delicate modeling assumptions.

Making the unit of analysis differ from the unit of randomization is dangerous.

Be wary of finer-grained analysis, e.g.,

- randomizing by city, analyzing by user.
- randomizing by category, analyzing by product.

It can be done, but requires delicate modeling assumptions.

An extreme example: imagine we have just two groups, say San Francisco and Los Angeles.

There are only two possible randomizations.

The randomization implies only two possible outcomes.

Choosing who to enroll

An aside: sample size planning / power calculations

The guarantee of a hypothesis test:

“If the treatment has no effect,
the chance of false discovery is at most 5%.”

An aside: sample size planning / power calculations

The guarantee of a hypothesis test:

“If the treatment has no effect,
the chance of false discovery is at most 5%.”

What if the treatment *does* have an effect?

“If the treatment has at least a 20% lift,
the chance of detecting it is at least 80%.”

An aside: sample size planning / power calculations

The guarantee of a hypothesis test:

“If the treatment has no effect,
the chance of false discovery is at most 5%.”

What if the treatment *does* have an effect?

“If the treatment has at least a 20% lift,
the chance of detecting it is at least 80%.”

How to guarantee the second statement?

Sample size planning.

An aside: sample size planning / power calculations

The guarantee of a hypothesis test:

“If the treatment has no effect,
the chance of false discovery is at most 5%.”

Type I error rate

What if the treatment *does* have an effect?

Minimum planned-for effect

“If the treatment has at least a 20% lift,
the chance of detecting it is at least 80%.”

Power

How to guarantee the second statement?

Sample size planning.

An aside: sample size planning / power calculations

- Type I error rate: 5%
- Minimum planned-for effect: 20% lift
- Power: 80%

```
>>> power_prop_test(p1=0.1, p2=0.1 * 1.2,  
                    significance_level=0.05,  
                    power=0.8).n_per_group  
3840.847482436278
```

Randomize as lazily as possible.

50,000 visitors/week → 1% advance to checkout → 20% complete

Say our variation increases conversion rate from 20% to 25%.

Randomize as lazily as possible.

50,000 visitors/week → 1% advance to checkout → 20% complete

Say our variation increases conversion rate from 20% to 25%.

Bad idea: enroll everyone.

```
power_prop_test(p1=.20 * .01, p2=.25 * .01, power=.8)
```

→ need 280,000 visitors.

Randomize as lazily as possible.

50,000 visitors/week → 1% advance to checkout → 20% complete

Say our variation increases conversion rate from 20% to 25%.

Bad idea: enroll everyone.

```
power_prop_test(p1=.20 * .01, p2=.25 * .01, power=.8)
```

→ need 280,000 visitors.

Good idea: enroll only those get to the checkout page.

```
power_prop_test(p1=.20, p2=.25, power=.8)
```

→ need 2,200 visitors at checkout

Randomize as lazily as possible.

50,000 visitors/week → 1% advance to checkout → 20% complete

Say our variation increases conversion rate from 20% to 25%.

Bad idea: enroll everyone.

```
power_prop_test(p1=.20 * .01, p2=.25 * .01, power=.8)
```

→ need 280,000 visitors.

Good idea: enroll only those get to the checkout page.

```
power_prop_test(p1=.20, p2=.25, power=.8)
```

→ need 2,200 visitors at checkout

→ need 220,000 visitors total (*20% decrease in sample size*)

Sometimes it helps to enroll a highly-affected subgroup.

Suppose treatment is expensive, and

- among all users at checkout, 20% complete the purchase;
- Among users with at least two items in their cart, 40% complete the purchase.

Sometimes it helps to enroll a highly-affected subgroup.

Suppose treatment is expensive, and

- among all users at checkout, 20% complete the purchase;
- Among users with at least two items in their cart, 40% complete the purchase.

Idea #1: enroll everyone.

```
power_prop_test(p1=.20, p2=.25, power=.8)
```

→ need 2,200 visitors.

Sometimes it helps to enroll a highly-affected subgroup.

Suppose treatment is expensive, and

- among all users at checkout, 20% complete the purchase;
- Among users with at least two items in their cart, 40% complete the purchase.

Idea #1: enroll everyone.

```
power_prop_test(p1=.20, p2=.25, power=.8)
```

→ need 2,200 visitors.

Idea #2: enroll only those with two items.

```
power_prop_test(p1=.40, p2=.50, power=.8)
```

→ need 770 visitors at checkout

(65% decrease in enrolled sample size)

Focusing on a subgroup may help internal validity, but hurt external validity.

Internal validity: are my conclusions valid for enrolled subjects?

Focusing on a subgroup may help internal validity, but hurt external validity.

Internal validity: are my conclusions valid for enrolled subjects?

External validity: do my conclusions generalize to other groups?

Focusing on a subgroup may help internal validity, but hurt external validity.

Internal validity: are my conclusions valid for enrolled subjects?

External validity: do my conclusions generalize to other groups?

Sometimes a difficult tradeoff.

Choosing an outcome metric

Imagine we're testing changes to a search ranking algorithm.

python power calculation

All Images Videos Shopping News More Settings Tools

About 742,000 results (0.35 seconds)

9.2. math — Mathematical functions — Python 2.7.15rc1 documentation
<https://docs.python.org/2/library/math.html> ▼
Return the natural logarithm of 1+x (base e). The result is **calculated** in a way which is accurate for x near zero. New in version 2.6. **math.log10(x)**¶. Return the base-10 logarithm of x. This is usually more accurate than **log(x, 10)**. **math.pow(x, y)**¶. Return x raised to the **power y**. Exceptional cases follow Annex F of the ...

Experimental design—power analysis and its visualisation ...
www.djmannon.net/psych_programming/data/power/power.html ▼
Experimental design—power analysis and its visualisation. Objectives. Be able to perform **power calculations** using computational simulation approaches. Know how to create and use line and image plots. **Power** relates to the ability to detect the presence of a true effect and is an important component of experimental ...

r - Is there a python (scipy) function to determine parameters ...
<https://stackoverflow.com/.../is-there-a-python-scipy-function-to-determine-parameter...> ▼
Mar 4, 2013 - I've managed to replicate the function using the below **formula** for n and the inverse survival function **norm.isf** from **scipy.stats**. enter image description here from **scipy.stats** import norm, zscore def **sample_power_probttest**(p1, p2, **power**=0.8, sig=0.05): z = norm.isf([sig/2]) #two-sided 1 test zp = -1 ...

statistics - How to calculate (statistical) power function ...	1 answer	Nov 15, 2017
math - Power of in python	3 answers	Jan 5, 2016
Calculating power for Decimals in Python	4 answers	Jul 10, 2013
Python and Powers Math	3 answers	Aug 20, 2012

More results from [stackoverflow.com](#)

python power calculation

All Images Videos Maps News Shop | My saves

13,200,000 Results Any time ▼

Calculating power for Decimals in Python - Stack Overflow
<https://stackoverflow.com/questions/1756720/calculating-power-for...> ▼
I want to **calculate power** for Decimal in Python like: from decimal import Decimal
Decimal.**power**(2,2) Above should return me as Decimal(2) How can I **calculate power** ...

Code sample

```
>>> x = Decimal(2)
>>> deci_x = Decimal(1)
>>> n=4
>>> y = Decimal('10')**(x-deci_x+Decimal(str(n))-Decimal('1'))
>>> y...
```

[See more on stackoverflow](#) Was this helpful? #1

Power calculation : Power « Math « Python
www.java2s.com » Python » Math » **Power** ▼
Power calculation : Power = Math = Python. Python; Math; Power; **Power calculation**.
print 2L ** 200 print 2 ** 200 Related examples in the same category: 1.

Python Program to Make a Simple Calculator
<https://www.programiz.com/python-programming/examples/calculator> ▼
In this example you will learn to create a simple **calculator** that can add, subtract, multiply or divide depending upon the input from the user.

numpy.power — NumPy v1.14 Manual - SciPy.org

How should we evaluate search quality?

We need a metric to run an experiment.

How should we evaluate search quality?

We need a metric to run an experiment.

We'd like to improve market share:

$$\frac{\# \text{ queries to our search engine}}{\# \text{ queries to all search engines}}$$

How should we evaluate search quality?

We need a metric to run an experiment.

We'd like to improve market share:

$$\frac{\text{\# queries to our search engine}}{\text{\# queries to all search engines}}$$

We can't measure the denominator. So just use the numerator?

Finding the right metric isn't easy!

Kohavi, Deng, et al. (2012): a buggy experiment showing very poor search results caused

- 10% lift in queries per user, and
- 30% lift in revenue per user!

What happened here?

Finding the right metric isn't easy!

Kohavi, Deng, et al. (2012): a buggy experiment showing very poor search results caused

- 10% lift in queries per user, and
- 30% lift in revenue per user!

What happened here?

- Bad results → issue more queries
- Bad organic results → more clicks on ads

$$\frac{\text{Queries}}{\text{Month}} = \frac{\text{Users}}{\text{Month}} \times \frac{\text{Sessions}}{\text{User}} \times \frac{\text{Queries}}{\text{Session}}$$

From Kohavi, Deng, et al. (2012).

$$\frac{\text{Queries}}{\text{Month}} = \frac{\text{Users}}{\text{Month}} \times \frac{\text{Sessions}}{\text{User}} \times \frac{\text{Queries}}{\text{Session}}$$

- # Users is fixed by design.

From Kohavi, Deng, et al. (2012).

$$\frac{\text{Queries}}{\text{Month}} = \frac{\text{Users}}{\text{Month}} \times \frac{\text{Sessions}}{\text{User}} \times \frac{\text{Queries}}{\text{Session}}$$

- # Users is fixed by design.
- Queries / Session is difficult to interpret.

From Kohavi, Deng, et al. (2012).

$$\frac{\text{Queries}}{\text{Month}} = \frac{\text{Users}}{\text{Month}} \times \frac{\text{Sessions}}{\text{User}} \times \frac{\text{Queries}}{\text{Session}}$$

- # Users is fixed by design.
- **Sessions / User seems the best metric.**
- Queries / Session is difficult to interpret.

From Kohavi, Deng, et al. (2012).

What if I want to look at multiple outcome metrics?

My suggestion for multiple outcome metrics:

- **Pick one primary metric**—a “key performance indicator” or KPI.
- If the others are important, correct for multiple testing.
- Otherwise, look at them, but educate yourself and others about multiplicity.

1. The need for randomized experiments

2. Design issues

3. Sequential experimentation

- A lesson about random walks
- Repeated looks inflate error
- Simulation-based sequential p-values

A lesson about random walks

The coin-flipping game: long leads

Every day for one year, I flip a fair coin.

- Heads \rightarrow you pay me \$1.
- Tails \rightarrow I pay you \$1.

The coin-flipping game: long leads

Every day for one year, I flip a fair coin.

- Heads → you pay me \$1.
- Tails → I pay you \$1.

What's the chance that, after the first eight days, one of us stays in the lead the *entire rest of the year*?

- (a) One in 10,000
- (b) One in 1,000
- (c) One in 100
- (d) One in 10

The coin-flipping game: long leads

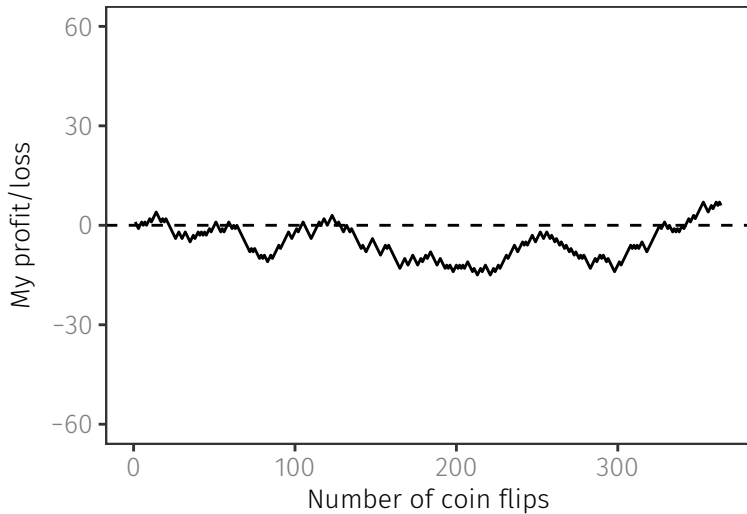
Every day for one year, I flip a fair coin.

- Heads \rightarrow you pay me \$1.
- Tails \rightarrow I pay you \$1.

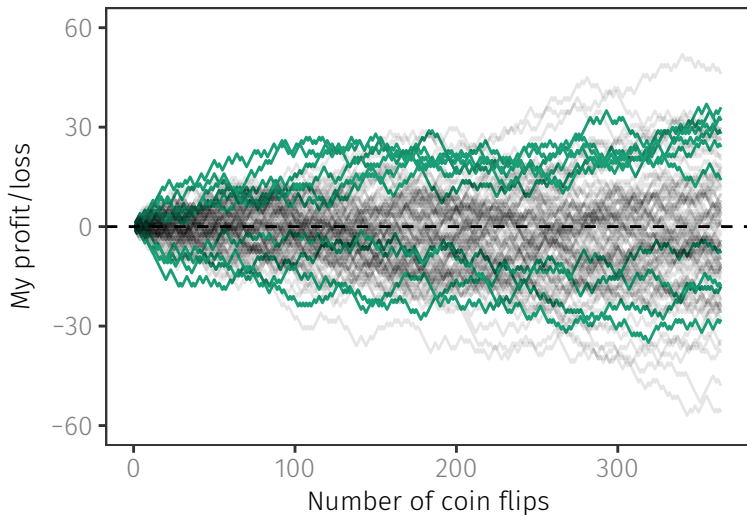
What's the chance that, after the first eight days, one of us stays in the lead the *entire rest of the year*?

- (a) One in 10,000
- (b) One in 1,000
- (c) One in 100
- (d) One in 10**

One random walk path

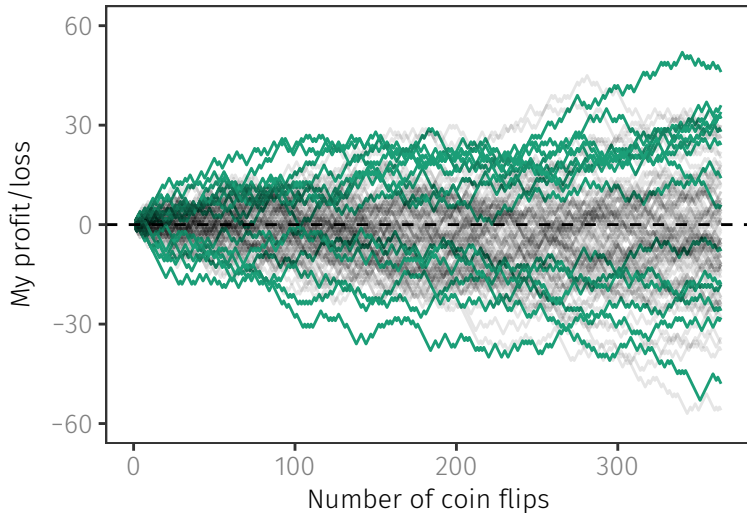


Random walks with long leads



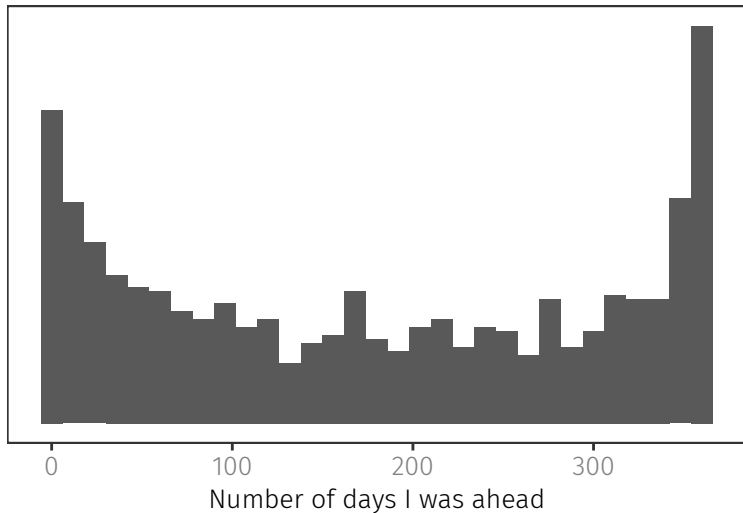
10 / 100 walks have one leader for the last 357 / 365 days.

Random walks with a dominant leader



15 / 100 walks have one player ahead any 357 / 365 days.

A highly uneven outcome is the norm, not the exception.



(These are called arcsine laws.)

What's going on here?

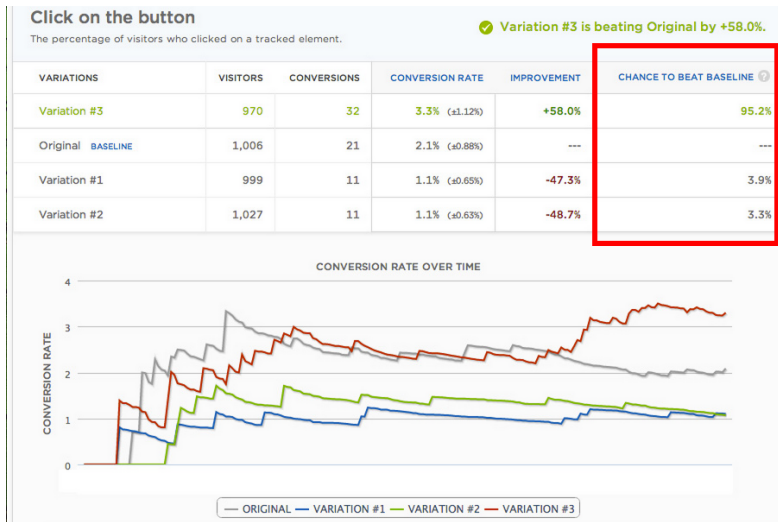
In a random walk, outcomes at different times are *highly correlated*.

Our usual notions about long-run behavior don't apply.

These examples are based on Feller (1971, §III.4).

Repeated looks inflate error

Sequential monitoring of A/B tests is desirable but problematic.



Testing a coin for fairness

How to determine if a coin is fair? Start flipping it!

T H T

Testing a coin for fairness

How to determine if a coin is fair? Start flipping it!

T H T H H T H H H

Testing a coin for fairness

How to determine if a coin is fair? Start flipping it!

T H T H H T H H H T H H T T H T H T H H H T H H ...

Testing a coin for fairness

How to determine if a coin is fair? Start flipping it!

T H T H H T H H H T H H T T H T H T H H H T H H ...

Is 15 / 24 heads surprising?

Testing a coin for fairness

How to determine if a coin is fair? Start flipping it!

T H T H H T H H H T H H T T H T H T H H H T H H ...

Is 15 / 24 heads surprising? This is what p-values are for.

p-values control the chance of false discovery

The guarantee of a hypothesis test:

“If the treatment has no effect,
the chance of false discovery is at most 5%.”

p-values control the chance of false discovery

The guarantee of a hypothesis test:

“If the treatment has no effect,
the chance of false discovery is at most 5%.”

The key property of p-values: if the treatment has no effect,

$$\mathbb{P}(\text{p-value} \leq 0.05) \leq 0.05.$$

p-values control the chance of false discovery

The guarantee of a hypothesis test:

“If the treatment has no effect,
the chance of false discovery is at most 5%.”

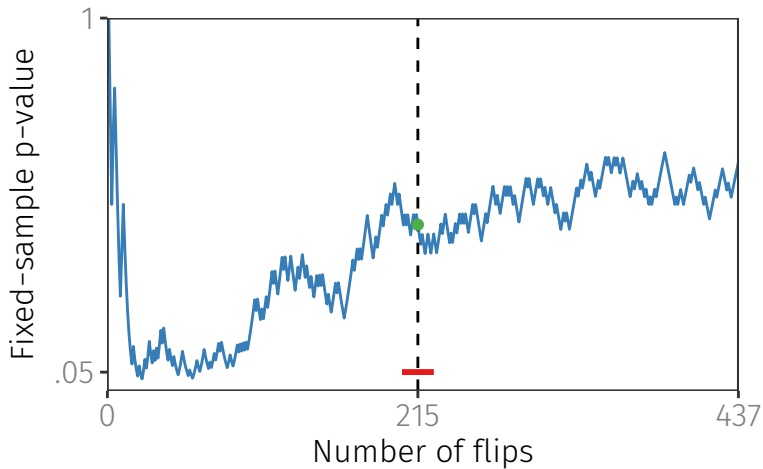
The key property of p-values: if the treatment has no effect,

$$\mathbb{P}(\text{p-value} \leq 0.05) \leq 0.05.$$

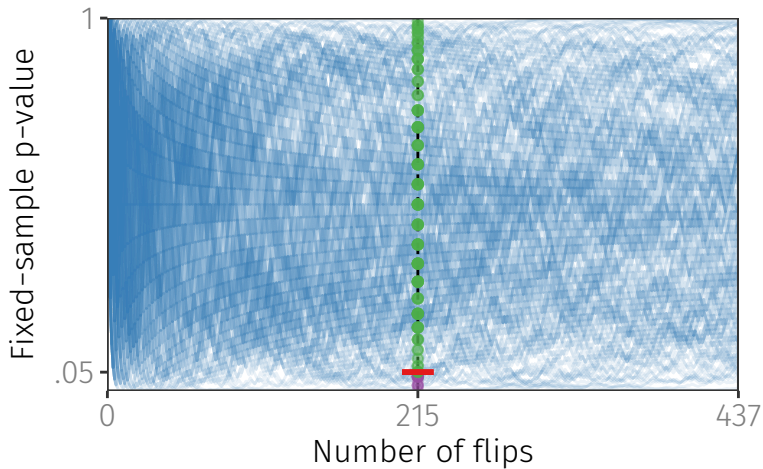
Declare a discovery when $\text{p-value} \leq 0.05$

→ chance of false discovery is at most 5%.

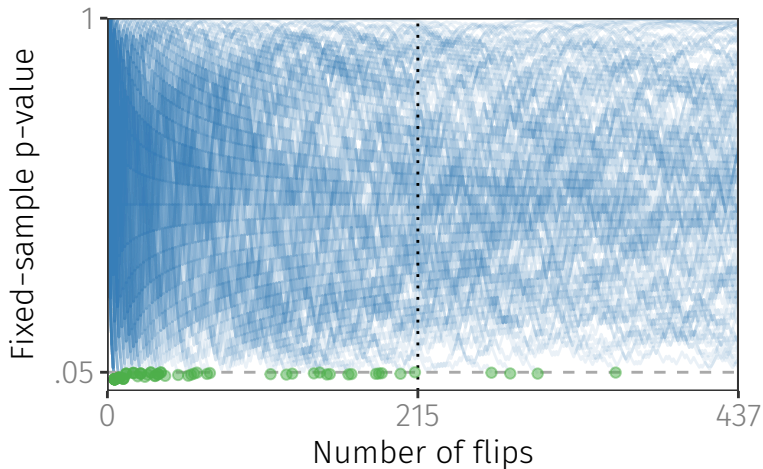
One path of p-values from a fair coin.



With no bias, we only rarely conclude the coin is biased.

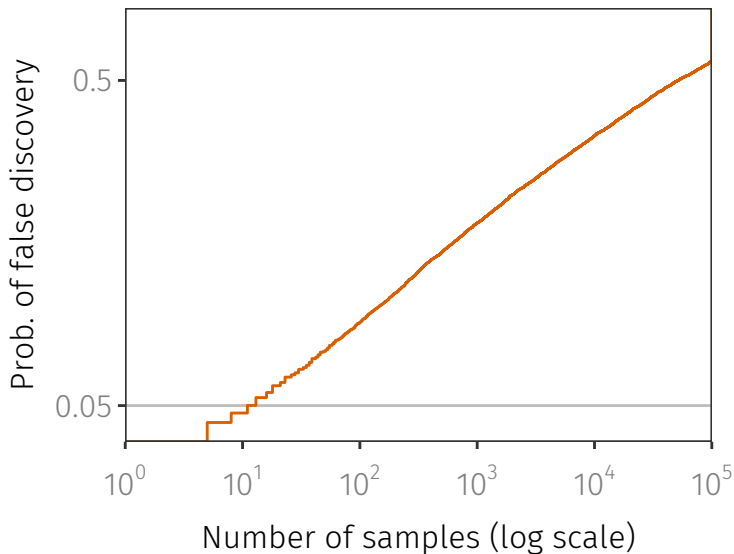


Continuous monitoring of fixed-sample p-values breaks the guarantee.



Here, with a fair coin, 35% of paths reach significance.

For a fair coin, chance of false discovery grows arbitrarily large with enough flips.



Sequential monitoring happens all over the place.

- In A/B testing.
- In clinical trials.
- In lab experiments.
- ...

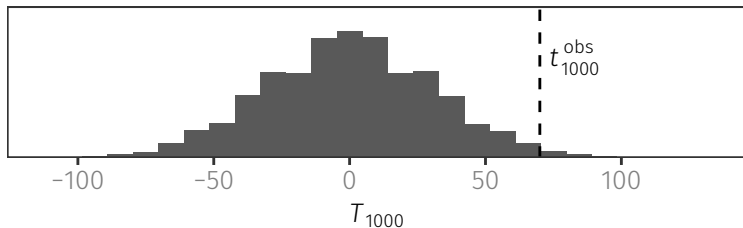
Simulation-based sequential p-values

Reminder: fixed-sample p-values by simulation

Standard p-value to test whether a coin is fair:

- Flip the coin 1,000 times.
- Compute $t_{1000}^{\text{obs}} = \# \text{ heads} - \# \text{ tails}$ after 1,000 flips.
- Simulate T_{1000} many times and estimate

$$\text{p-value} = \mathbb{P}(|T_{10,000}| \geq t_{10,000}^{\text{obs}}).$$

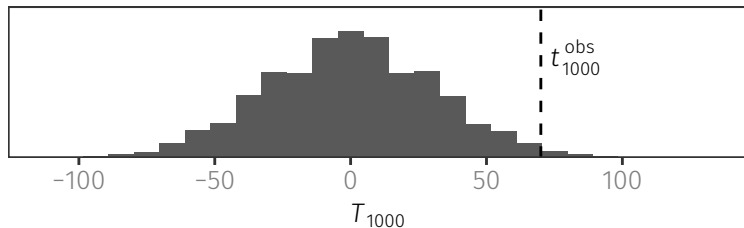


Reminder: fixed-sample p-values by simulation

Standard p-value to test whether a coin is fair:

- Flip the coin 1,000 times.
- Compute $t_{1000}^{\text{obs}} = \# \text{ heads} - \# \text{ tails}$ after 1,000 flips.
- Simulate T_{1000} many times and estimate

$$\text{p-value} = \mathbb{P}(|T_{10,000}| \geq t_{10,000}^{\text{obs}}).$$



Example: $t_{1000}^{\text{obs}} = 70 \rightarrow p \approx 0.029$.

Now we have a sequence of test statistics.

Now say we want to compute a p-value after every 100 flips.

Now we have a sequence of test statistics.

Now say we want to compute a p-value after every 100 flips.

We need to consider the sequence of test statistics

$$T_{100}, T_{200}, \dots, T_{900}, T_{1000}.$$

Now we have a sequence of test statistics.

Now say we want to compute a p-value after every 100 flips.

We need to consider the sequence of test statistics

$$T_{100}, T_{200}, \dots, T_{900}, T_{1000}.$$

For a fair coin, T_n is a sum of n i.i.d. random variables, each taking values ± 1 with probability $1/2$ each.

Now we have a sequence of test statistics.

Now say we want to compute a p-value after every 100 flips.

We need to consider the sequence of test statistics

$$T_{100}, T_{200}, \dots, T_{900}, T_{1000}.$$

For a fair coin, T_n is a sum of n i.i.d. random variables, each taking values ± 1 with probability $1/2$ each.

The variance of this ± 1 random variable is one.

Now we have a sequence of test statistics.

Now say we want to compute a p-value after every 100 flips.

We need to consider the sequence of test statistics

$$T_{100}, T_{200}, \dots, T_{900}, T_{1000}.$$

For a fair coin, T_n is a sum of n i.i.d. random variables, each taking values ± 1 with probability $1/2$ each.

The variance of this ± 1 random variable is one.

\Rightarrow The variance of T_n is n (standard deviation \sqrt{n}).

Now we have a sequence of test statistics.

Now say we want to compute a p-value after every 100 flips.

We need to consider the sequence of test statistics

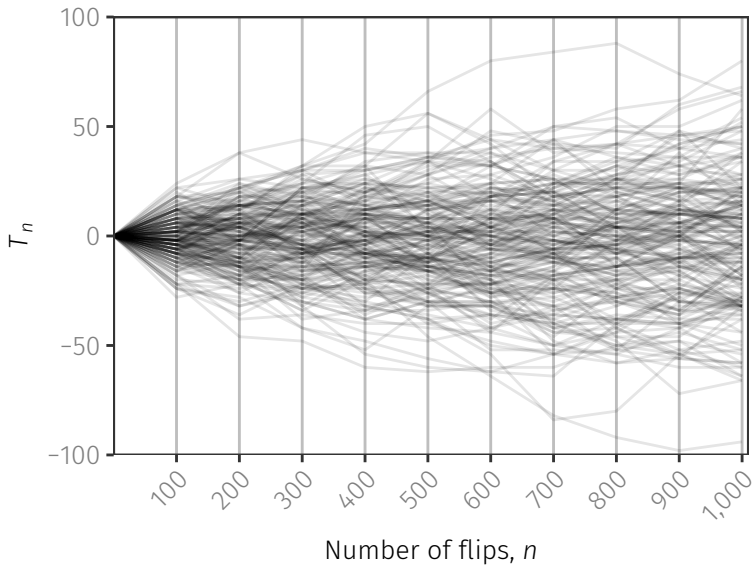
$$T_{100}, T_{200}, \dots, T_{900}, T_{1000}.$$

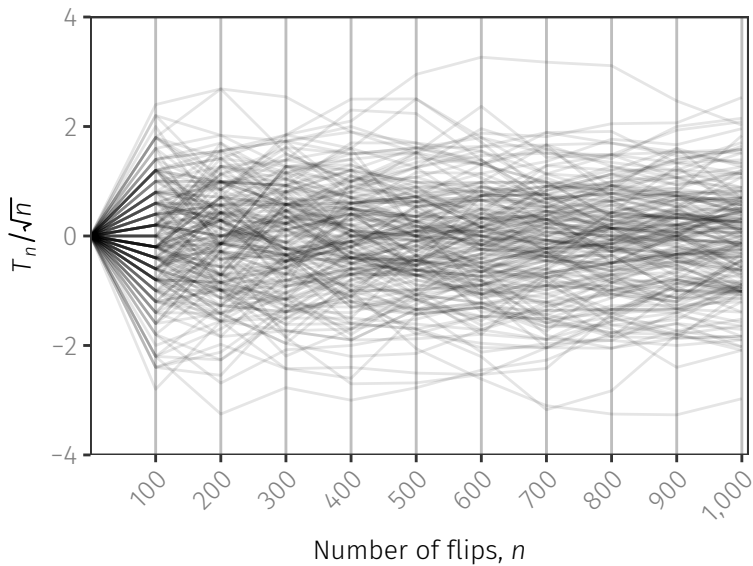
For a fair coin, T_n is a sum of n i.i.d. random variables, each taking values ± 1 with probability $1/2$ each.

The variance of this ± 1 random variable is one.

⇒ The variance of T_n is n (standard deviation \sqrt{n}).

⇒ The variance of T_n/\sqrt{n} is 1.





We'll use a maximal test statistic to compute sequential p-values.

Now we'll simulate the test statistic

$$T_{1000}^* = \max \left\{ \frac{T_{100}}{\sqrt{100}}, \frac{T_{200}}{\sqrt{200}}, \dots, \frac{T_{900}}{\sqrt{900}}, \frac{T_{1000}}{\sqrt{1000}} \right\}.$$

We'll use a maximal test statistic to compute sequential p-values.

Now we'll simulate the test statistic

$$T_{1000}^* = \max \left\{ \frac{T_{100}}{\sqrt{100}}, \frac{T_{200}}{\sqrt{200}}, \dots, \frac{T_{900}}{\sqrt{900}}, \frac{T_{1000}}{\sqrt{1000}} \right\}.$$

Our sequential procedure:

- After every 100 flips, compute

$$t_n^{\text{obs}} = \frac{\# \text{ heads} - \# \text{ tails after } n \text{ flips}}{\sqrt{n}}$$

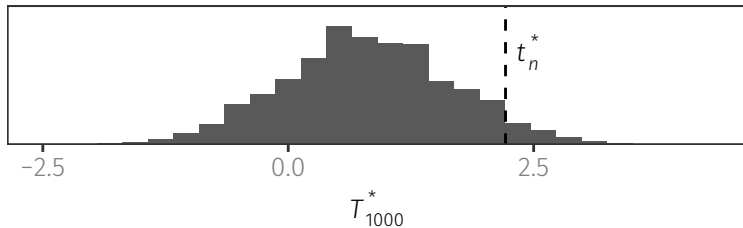
$$t_n^* = \max \left\{ t_{100}^{\text{obs}}, t_{200}^{\text{obs}}, \dots, t_n^{\text{obs}} \right\}.$$

- Simulate T_{1000}^* many times and estimate

$$\text{p-value} = \mathbb{P}(|T_{1000}^*| \geq t_n^*).$$

Example: $t_{1000}^* = 70/\sqrt{1000} \rightarrow p \approx 0.054$.

(Compare to $p \approx 0.029$ earlier.)



We can look at these p-values repeatedly.

Now we can compute a p-value after every 100 flips, stop as soon as $p \leq 0.05$, and still have the guarantee

$$\mathbb{P}(\text{any p-value} \leq 0.05) \leq 0.05$$

if the coin is fair.

We can look at these p-values repeatedly.

Now we can compute a p-value after every 100 flips, stop as soon as $p \leq 0.05$, and still have the guarantee

$$\mathbb{P}(\text{any } p\text{-value} \leq 0.05) \leq 0.05$$

if the coin is fair.

If the coin is biased, we have a chance to stop early.

We can look at these p-values repeatedly.

Now we can compute a p-value after every 100 flips, stop as soon as $p \leq 0.05$, and still have the guarantee

$$\mathbb{P}(\text{any p-value} \leq 0.05) \leq 0.05$$

if the coin is fair.

If the coin is biased, we have a chance to stop early.

Remember:

- We must choose the maximum sample size in advance (here, 1,000).
- We can only look as often as we do in the simulation (here, every 100 flips).

Recap

1. Randomized assignment

- protects from bias, and
- justifies probability calculations.

Recap

1. Randomized assignment







- protects from bias, and
- justifies probability calculations.

2. Before running an experiment, carefully choose

- the unit of randomization (and analysis!),
- the enrolled population, and
- the outcome metric.

Recap

1. Randomized assignment
 - protects from bias, and
 - justifies probability calculations.
2. Before running an experiment, carefully choose
 - the unit of randomization (and analysis!),
 - the enrolled population, and
 - the outcome metric.
3. If you want to monitor sequentially, use sequential methods!

- 
- Box, G. E. P., J. S. Hunter, and W. G. Hunter (2005). *Statistics for experimenters: design, innovation, and discovery*. Wiley-Interscience.
- 
- Feller, W. (1971). *An introduction to probability theory and its applications*. 3rd. Wiley.
- 
- Freedman, D., R. Pisani, and R. Purves (2007). *Statistics*. W.W. Norton & Company.
- 
- Kohavi, R., A. Deng, B. Frasca, R. Longbotham, T. Walker, and Y. Xu (2012). “Trustworthy online controlled experiments: Five puzzling outcomes explained”. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 786–794.
- 
- Kohavi, R., R. Longbotham, D. Sommerfield, and R. M. Henne (2009). “Controlled experiments on the web: survey and practical guide”. *Data Mining and Knowledge Discovery* 18 (1), pp. 140–181.
- 
- Kohavi, R. and S. H. Thomke (2017). “The Surprising Power of Online Experiments”. *Harvard Business Review* 95 (5), pp. 74–82.

Thank you.