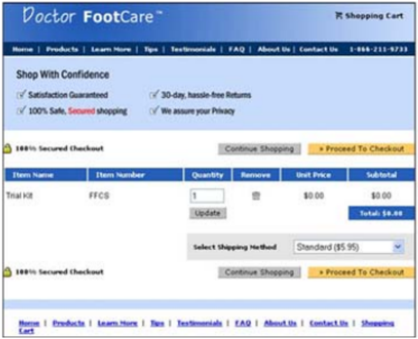
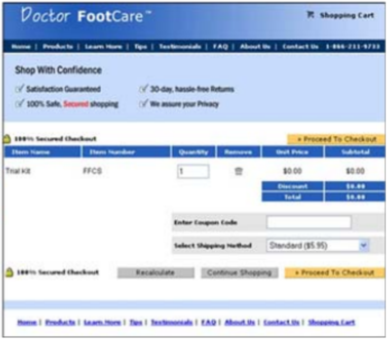
Randomized experiments, A/B tests and sequential monitoring

Steve Howard

April 26, 2018

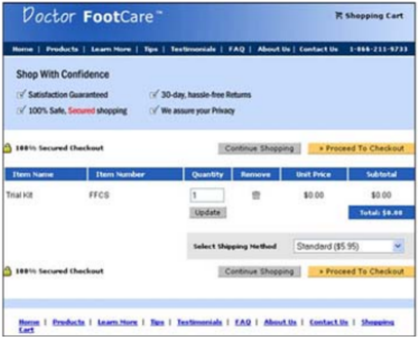
Guess the winner!

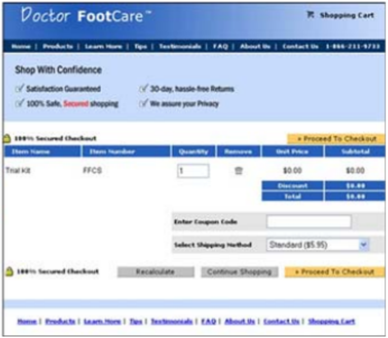




From Kohavi, Longbotham, et al. (2009).

Guess the winner!

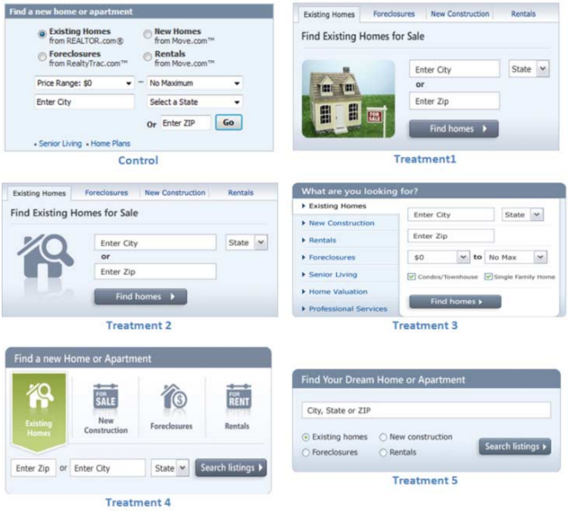




10x revenue!

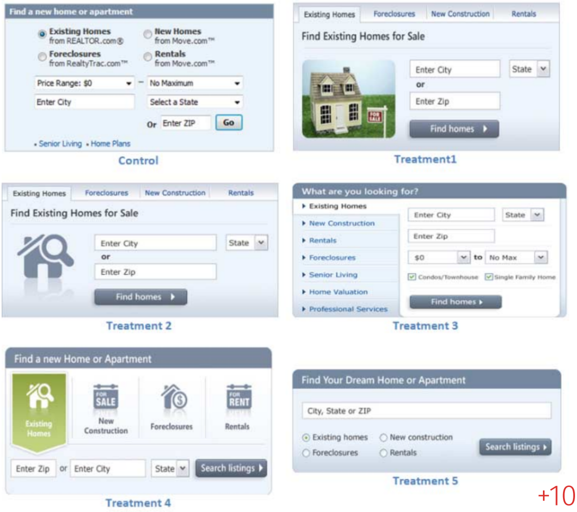
From Kohavi, Longbotham, et al. (2009).

Guess the winner!



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

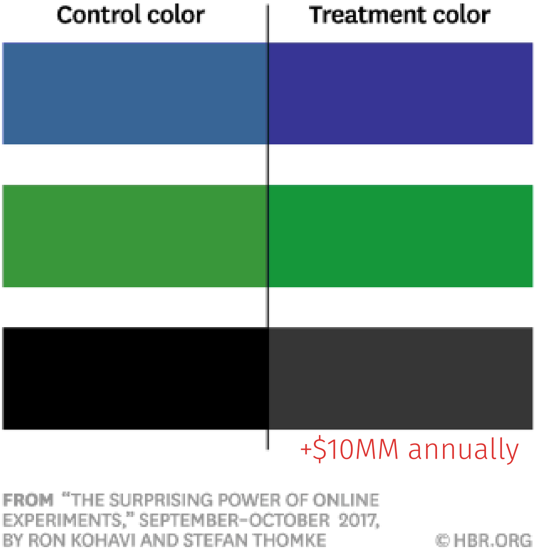
Guess the winner!



+10% revenue

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

Yes, the color thing is real.



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 hosptital, 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 hosptital, 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 hosptital, 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 hosptital, 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 hosptital, 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.

	Degree of enthusiasm		
Design	*Marked*	*Moderate*	*None*
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 (−).

Therapy	*Randomized controlled*		*Historically controlled*	
	+	−	+	−
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

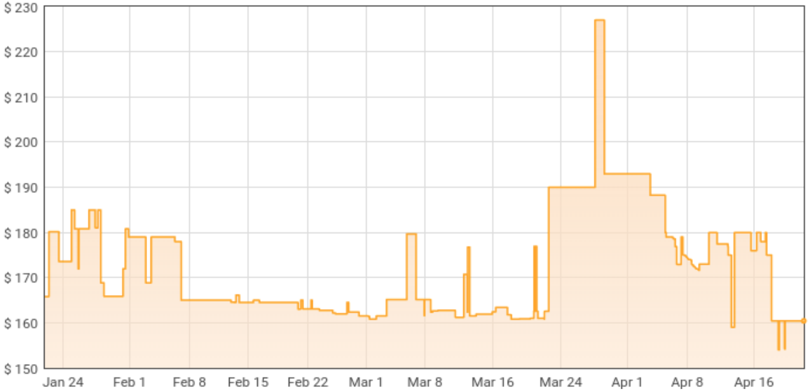
2. Design choices

- 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 keepa.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 $\rightarrow$ 1% advance to checkout $\rightarrow$ 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)
```

$\rightarrow$ 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)
```

$\rightarrow$ need 2,200 visitors.

Idea #2: enroll only those with two items.

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

$\rightarrow$ 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.

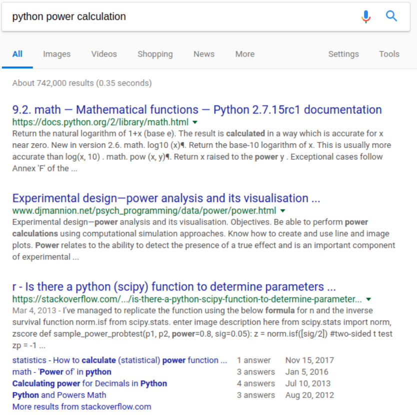
Internal validity: are my conclusions valid for enrolled subjects?

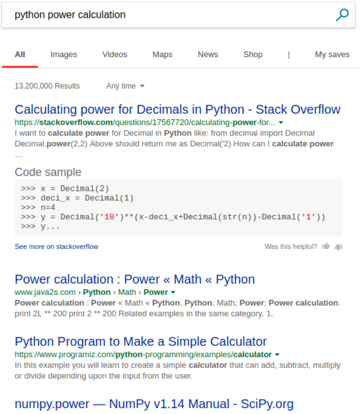
External validity: do my conclusions generalize to other groups?

Sometimes a difficult tradeoff.

Choosing an outcome metric

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





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 $\rightarrow$ issue more queries
- Bad organic results $\rightarrow$ 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 $\rightarrow$ 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

The coin-flipping game: long leads

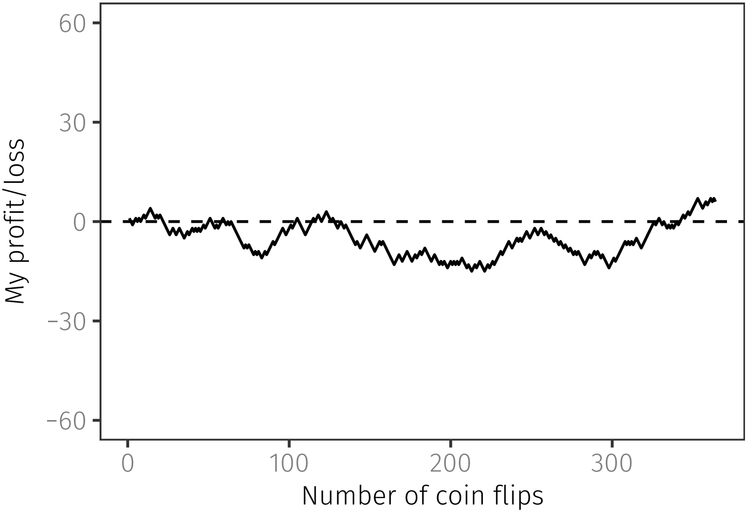
Every day for one year, I flip a fair coin.

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

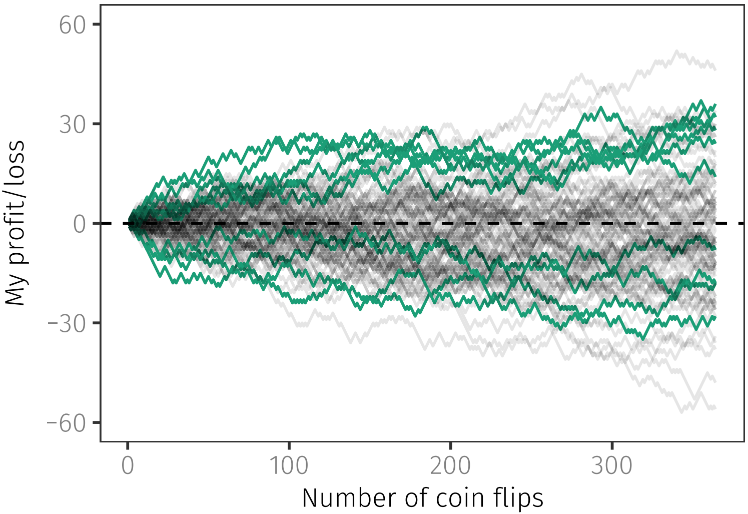
What's the chance that, after the first eight days,
one of us stays in the lead the *entire rest of the year*?

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

One random walk path

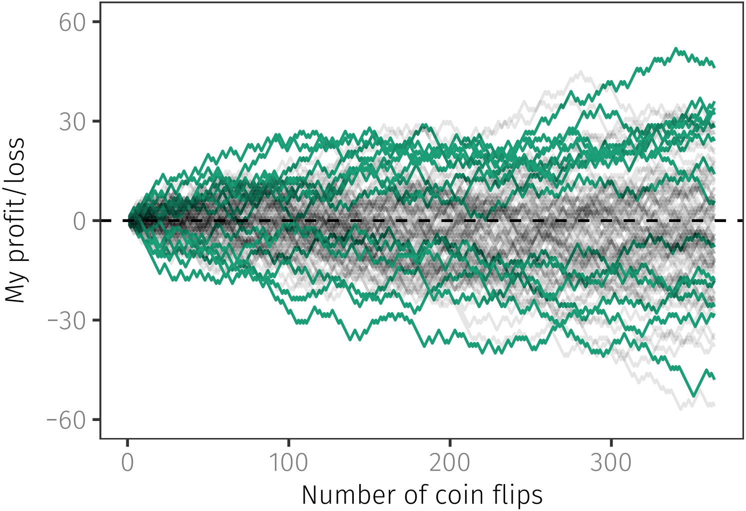


Random walks with long leads



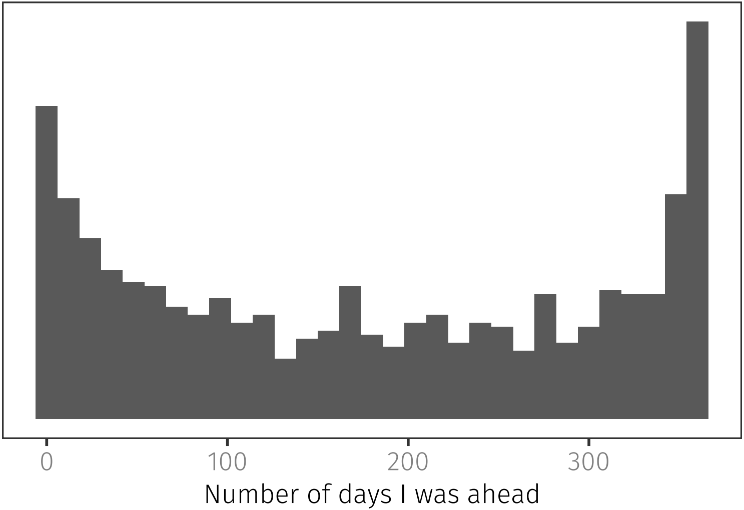
10 / 100 walks have one leader for the last 357 / 365 days.

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

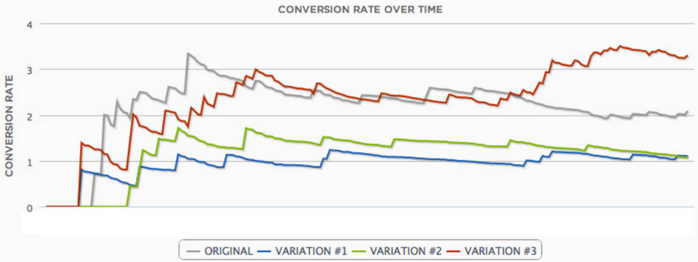
Sequential monitoring of A/B tests is desirable but problematic.

Click on the button

The percentage of visitors who clicked on a tracked element.

Variation #3 is beating Original by +58.0%.

VARIATIONS	VISITORS	CONVERSIONS	CONVERSION RATE	IMPROVEMENT	CHANCE TO BEAT BASELINE
Variation #3	970	32	3.3% (±1.12%)	+58.0%	95.2%
Original BASELINE	1,006	21	2.1% (±0.88%)	---	---
Variation #1	999	11	1.1% (±0.65%)	-47.3%	3.9%
Variation #2	1,027	11	1.1% (±0.63%)	-48.7%	3.3%



From https://crozdesk.com/analytics-intelligence/a-b-testing-software/optimizely

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:

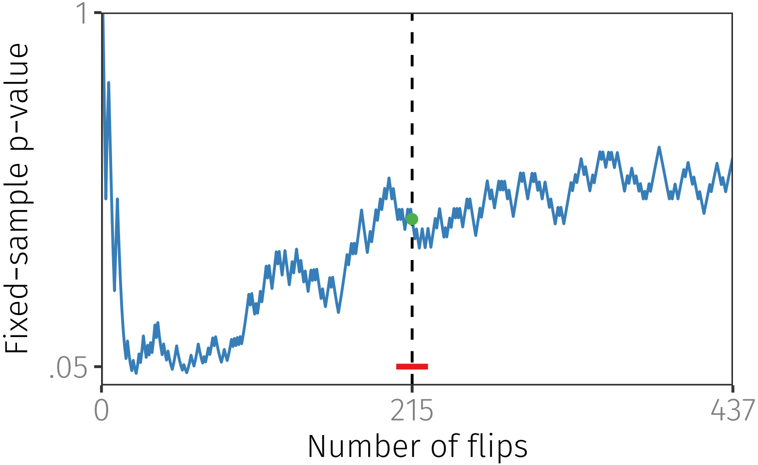
"If the treatment has no effect,
the chance of false discovery is at most 5%."

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

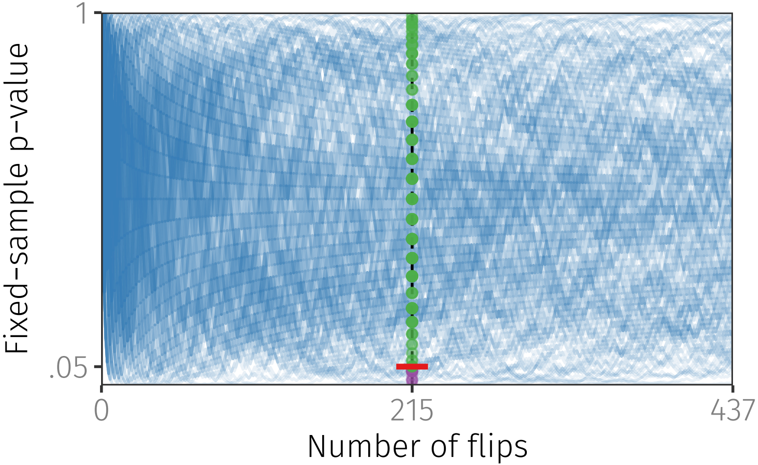
$$\mathbb{P}(\text{p-value} \leq 0.05) \leq 0.05.$$

Declare a discovery when p-value ≤ 0.05
$\rightarrow$ chance of false discovery is at most 5%.

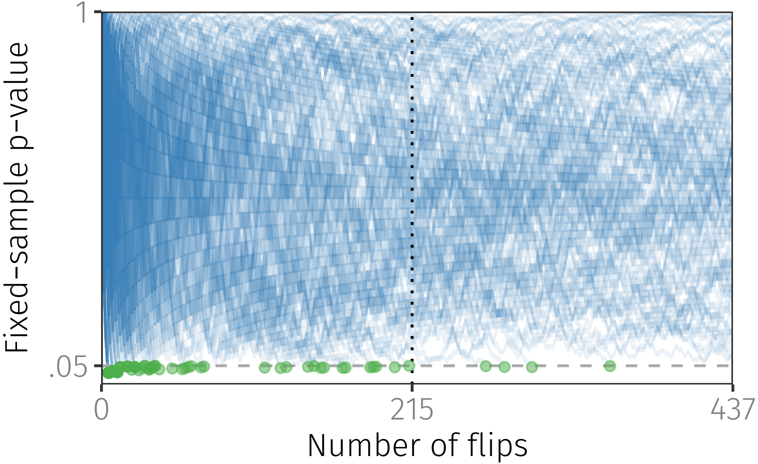
One path of p-values from a fair coin.



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

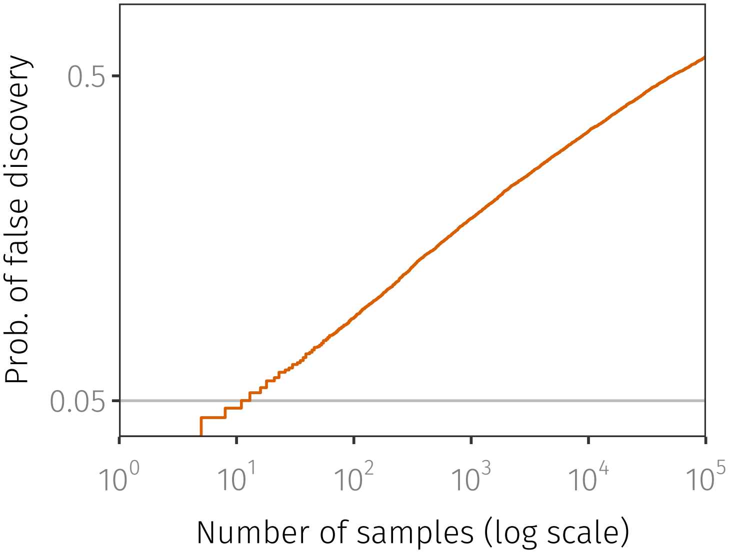


Continuous monitoring of fixed-sample p-values breaks the guarantee.



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 experments.
- …

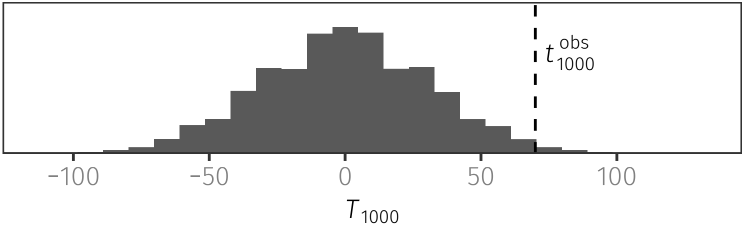
Simulation-based sequential p-values

Reminder: fixed-sample p-values by simulation

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

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

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

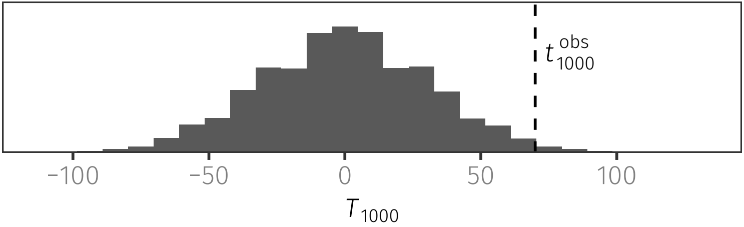


Reminder: fixed-sample p-values by simulation

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

- Flip the coin 1,000 times.
- Compute $t_{1000}^{\text{obs}} = \#$ heads - # 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: $t_{1000}^{\text{obs}} = 70 \to p \approx 0.029$.

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

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

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

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

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

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

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

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

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.

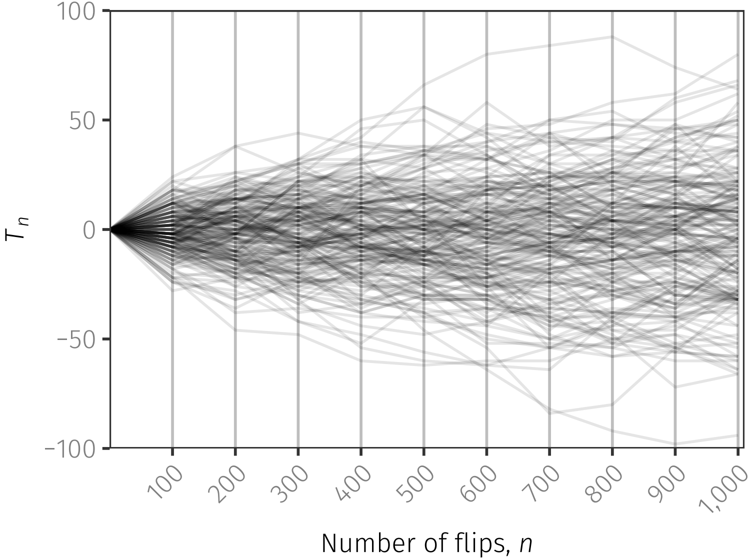
Now say we want to compute a p-value after every 100 flips.

We need to consider the sequence of test statistics

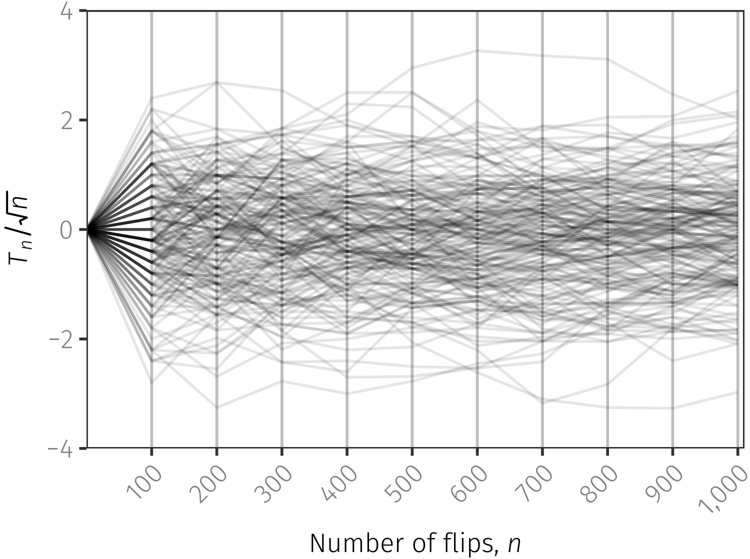
$$T_{100}, T_{200}, \ldots, T_{900}, T_{1000}.$$

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 $\sqrt{n}$).
$\Rightarrow$ The variance of $T_n/\sqrt{n}$ is 1.



100
50
0
−50
−100
T_n
100
200
300
400
500
600
700
800
900
1,000
Number of flips, n



4
2
0
−2
−4
$T_n/\sqrt{n}$
100
200
300
400
500
600
700
800
900
1,000
Number of flips, n

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

Now we'll simulate the test statistic

$$T^{\star}_{1000} = \max\left\{ \frac{T_{100}}{\sqrt{100}}, \frac{T_{200}}{\sqrt{200}}, \ldots, \frac{T_{900}}{\sqrt{900}}, \frac{T_{1000}}{\sqrt{1000}} \right\}.$$

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

Now we'll simulate the test statistic

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

Our sequential procedure:

- After every 100 flips, compute

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

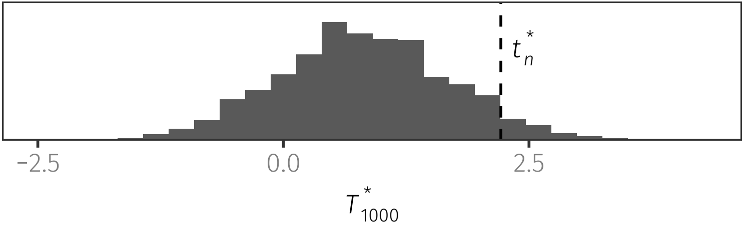
$$t_n^\star = \max\left\{ t_{100}^{\text{obs}}, t_{200}^{\text{obs}}, \dots, t_n^{\text{obs}} \right\}.$$

- Simulate $T^\star_{1000}$ many times and estimate

$$\text{p-value} = \mathbb{P}(|T^\star_{1000}| \geq t_n^\star).$$

Example: $t^{\star}_{1000} = 70/\sqrt{1000} \to p \approx 0.054$.

(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

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

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

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

if the coin is fair.

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

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

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

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.