THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Model evaluation: problems with *p*-values

Lecture 20

STA 371G

A quick refresher on *p*-values

- To do a study, we select a sample from a population, and collect a statistic of interest in that sample (e.g., a mean or proportion)

A quick refresher on *p*-values

- To do a study, we select a sample from a population, and collect a statistic of interest in that sample (e.g., a mean or proportion)
- For example, in marketing, we might A/B test two potential web online store designs.

A quick refresher on *p*-values

- To do a study, we select a sample from a population, and collect a statistic of interest in that sample (e.g., a mean or proportion)
- For example, in marketing, we might A/B test two potential web online store designs.
- We randomly show design A (the existing design) or B (the new design) to customers, and then track how much money they spend.

A quick refresher on *p*-values

- To do a study, we select a sample from a population, and collect a statistic of interest in that sample (e.g., a mean or proportion)
- For example, in marketing, we might A/B test two potential web online store designs.
- We randomly show design A (the existing design) or B (the new design) to customers, and then track how much money they spend.
- A *t*-test would be used to test the null hypothesis that the average amount spent with design A is the same as the amount spent with design B.

A quick refresher on *p*-values

- To do a study, we select a sample from a population, and collect a statistic of interest in that sample (e.g., a mean or proportion)
- For example, in marketing, we might A/B test two potential web online store designs.
- We randomly show design A (the existing design) or B (the new design) to customers, and then track how much money they spend.
- A *t*-test would be used to test the null hypothesis that the average amount spent with design A is the same as the amount spent with design B.
- If we find $p < .05$ (and that the new design results in significantly higher revenue than the old design) then we redo our store with the new design.

The hypotheses we are testing:

$$H_0 : \mu_A = \mu_B \qquad \text{vs} \qquad H_A : \mu_A < \mu B$$

And:

$$p\text{-value} = P(\text{seeing a difference as large as our sample} \mid H_0 \text{ is true})$$

The hypotheses we are testing:

$$H_0 : \mu_A = \mu_B \qquad \text{vs} \qquad H_A : \mu_A < \mu B$$

And:

$$p\text{-value} = P(\text{seeing a difference as large as our sample} \mid H_0 \text{ is true})$$

So a small p-value means that null hypothesis H_0 is not a tenable assumption—we'll reject it and assume instead that the alternative hypothesis H_A is true.

Remember: a *p*-value is a conditional probability that represents how likely it is that we would see a difference as extreme as the results we saw in the sample, under the assumption that H_0 is true.

As a result, it is sensitive to:

- Sample size: Seeing the same difference in a larger sample will tend to result in a smaller *p*-value.
- Sampling variation: More variation in the population will tend to result in a larger *p*-value.

1. The consequence of running too many hypothesis tests
2. Changing your analysis strategy based on a hypothesis test
3. Stopping an experiment based on a hypothesis test
4. Using *p*-values responsibly
5. Survivorship bias
6. Practical versus statistical significance

The mystery data set

- The "mystery" data set from Lecture 10 contained a Y variable and 20 X variables that were completely random.
- We built a regression predicting Y from the X's, and some of the slope coefficients end up being statistically significant!

```
...
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0048932  0.0315499   0.155  0.87678
X1          -0.0242680  0.0311808  -0.778  0.43658
X2           0.0094610  0.0325268   0.291  0.77121
X3          -0.0164687  0.0316452  -0.520  0.60289
X4           0.0332419  0.0315227   1.055  0.29190
X5          -0.0003434  0.0320555  -0.011  0.99146
X6          -0.0235806  0.0311449  -0.757  0.44916
X7           0.0239592  0.0314237   0.762  0.44597
X8          -0.0153691  0.0308775  -0.498  0.61878
X9           0.0188475  0.0301545   0.625  0.53210
X10          0.0816945  0.0310096   2.634  0.00856 **
X11          0.0300185  0.0314272   0.955  0.33972
X12         -0.0189656  0.0307947  -0.616  0.53812
X13         -0.0657305  0.0311022  -2.113  0.03482 *
X14          0.0076949  0.0318495   0.242  0.80914
X15          0.0227637  0.0316354   0.720  0.47197
X16          0.0689893  0.0332165   2.077  0.03807 *
X17         -0.0239150  0.0312032  -0.766  0.44361
X18         -0.0002470  0.0316023  -0.008  0.99376
X19         -0.0502899  0.0305908  -1.644  0.10051
X20         -0.0628928  0.0325095  -1.935  0.05333 .
...
```

- Suppose that the null hypothesis is always true.

- Suppose that the null hypothesis is always true.
- If you use $\alpha = .05$ as your cutoff for statistical significance, then you'll get a significant result 5% of the time just by chance!

- Suppose that the null hypothesis is always true.
- If you use $\alpha = .05$ as your cutoff for statistical significance, then you'll get a significant result 5% of the time just by chance!
- If we have 20 predictor variables, we are running 20 hypothesis tests, and we'd expect to see one of them be significant, just by chance.

- Suppose that the null hypothesis is always true.
- If you use $\alpha = .05$ as your cutoff for statistical significance, then you'll get a significant result 5% of the time just by chance!
- If we have 20 predictor variables, we are running 20 hypothesis tests, and we'd expect to see one of them be significant, just by chance.
- That means that when you are building a model from a large data set with many variables, you should expect to see some significant coefficients no matter what.

1. The consequence of running too many hypothesis tests
2. Changing your analysis strategy based on a hypothesis test
3. Stopping an experiment based on a hypothesis test
4. Using *p*-values responsibly
5. Survivorship bias
6. Practical versus statistical significance

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?
  - Nope—still nonsignificant.

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Texas?

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Texas?
  - Nope—still nonsignificant.

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Texas?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Oklahoma?

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Texas?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Oklahoma?
  - Yes—$p = 0.03$!

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Texas?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Oklahoma?
  - Yes—$p = 0.03$!
- So it must be true that there is a significant gender difference in propensity to buy among customers in Oklahoma...

- Suppose you have a strong belief that there is a gender difference in propensity to buy a certain product.
- You build a logistic regression model and find that the gender coefficient is nonsignificant ($p > .05$).
- Maybe it would be significant for your older, 65+ customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for your white customers?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Texas?
  - Nope—still nonsignificant.
- Maybe it would be significant for customers in Oklahoma?
  - Yes—$p = 0.03$!
- So it must be true that there is a significant gender difference in propensity to buy among customers in Oklahoma...
- ...or, you just got a false positive by running a lot of tests!

p-hacking

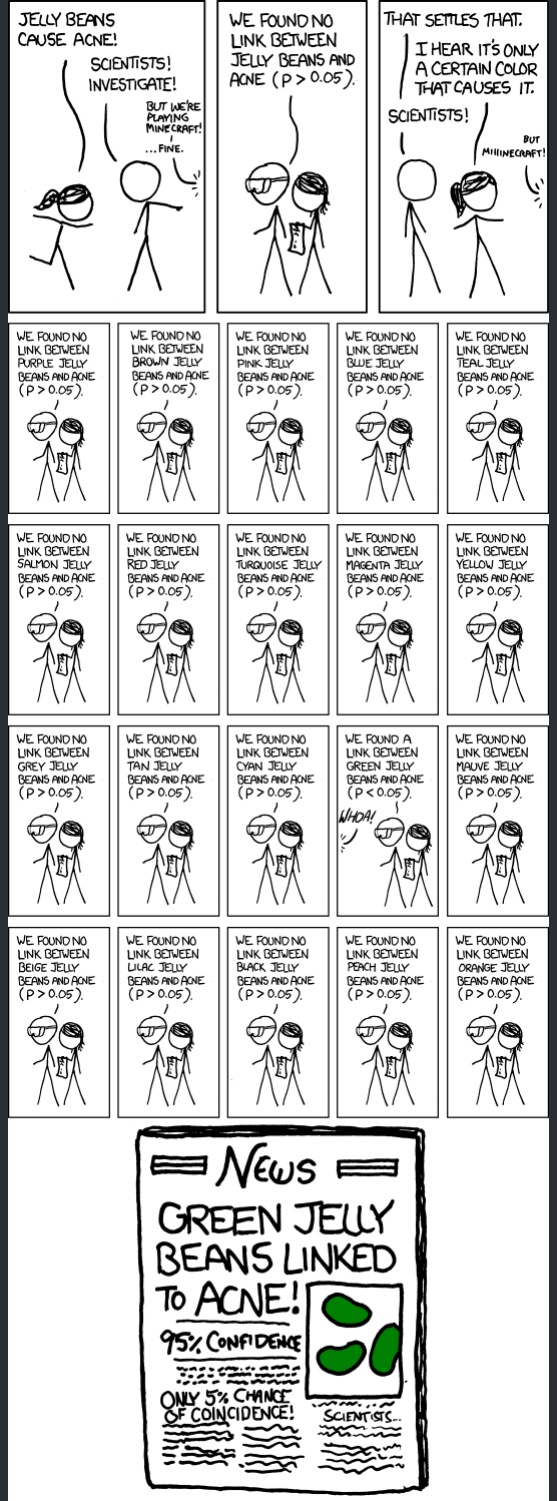
- This process is known as *p*-hacking.

p-hacking

- This process is known as *p*-hacking.
- In some sense, this is a natural part of the data analysis process—you don't necessarily know what effects are out there unless you look for them.

p-hacking

- This process is known as *p*-hacking.
- In some sense, this is a natural part of the data analysis process—you don't necessarily know what effects are out there unless you look for them.
- But running too many tests can result in false positives!


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | Democrats

2 DEFINE TERMS

Which politicians do you want to include?

- [ ] Presidents
- [x] Governors
- [x] Senators
- [x] Representatives

How do you want to measure economic performance?

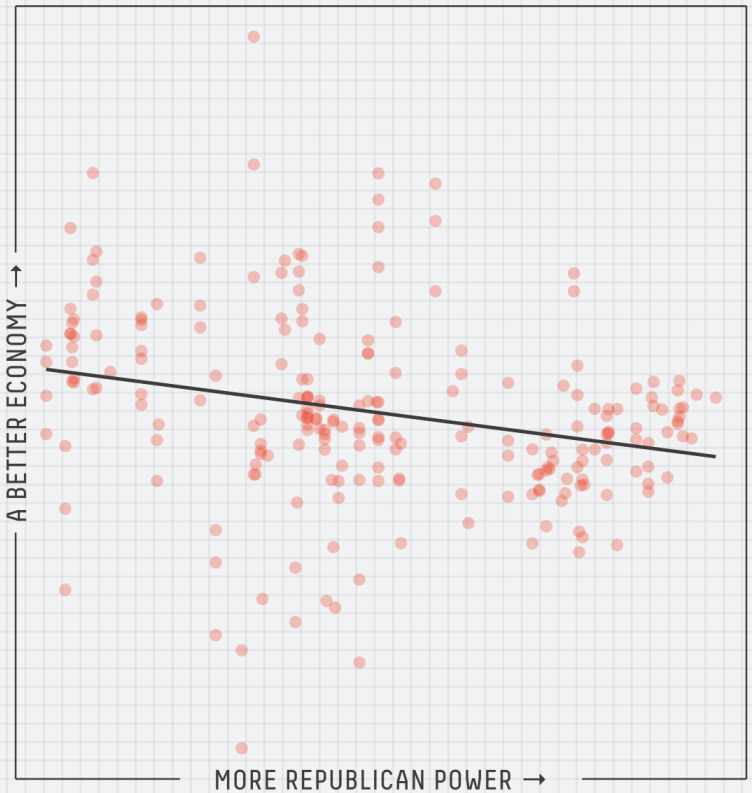
- [ ] Employment
- [x] Inflation
- [x] GDP
- [ ] Stock prices

Other options

- [x] **Factor in power**
  Weight more powerful positions more heavily
- [x] **Exclude recessions**
  Don't include economic recessions

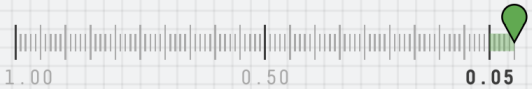
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Republicans are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Publishable

You achieved a p-value of **less than 0.01** and showed that **Republicans** have a **negative** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | **Democrats**

2 DEFINE TERMS

Which politicians do you want to include?

- [x] **Presidents**
- [x] **Governors**
- [x] **Senators**
- [x] **Representatives**

How do you want to measure economic performance?

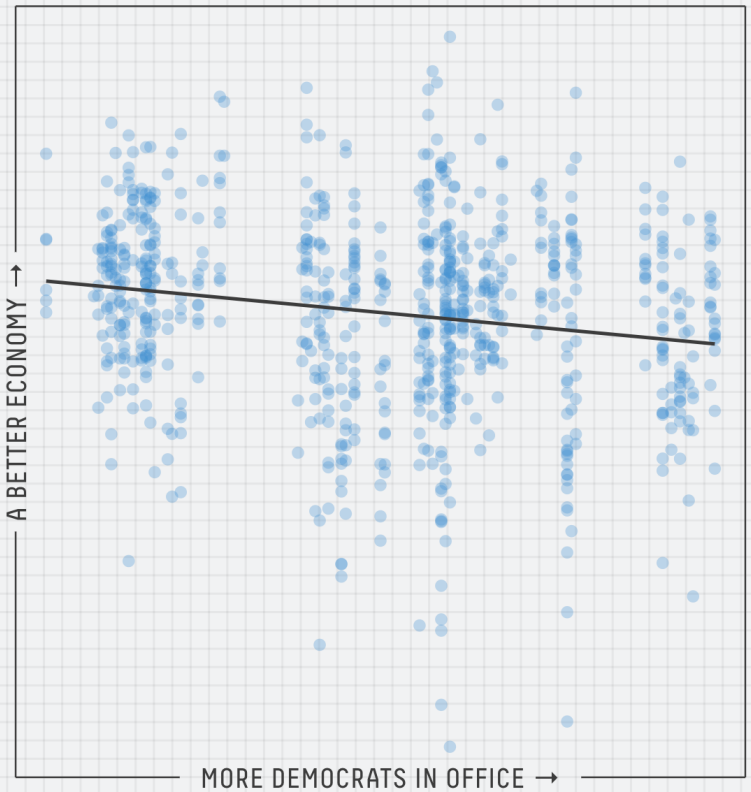
- [ ] **Employment**
- [x] **Inflation**
- [ ] **GDP**
- [x] **Stock prices**

Other options

- [ ] **Factor in power**
  Weight more powerful positions more heavily
- [ ] **Exclude recessions**
  Don't include economic recessions

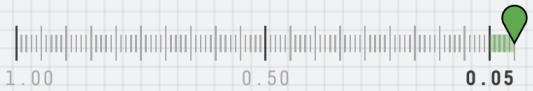
3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in office? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.



Result: Publishable

You achieved a p-value of **less than 0.01** and showed that **Democrats** have a **negative** effect on the economy. Get ready to be published!

If you're interested in reading real (and more rigorous) studies on the connection between politics and the economy, see the work of Larry Bartels and Alan Blinder and Mark Watson.

Data from The @unitedstates Project, National Governors Association, Bureau of Labor Statistics, Federal Reserve Bank of St. Louis and Yahoo Finance.

Journal of Personality and Social Psychology
2011, Vol. 100, No. 3, 407–425

0022-3514/11/$12.00 DOI: 10.1037/a0021524

Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect

Daryl J. Bem
Cornell University

The term *psi* denotes anomalous processes of information or energy transfer that are currently unexplained in terms of known physical or biological mechanisms. Two variants of psi are *precognition* (conscious cognitive awareness) and *premonition* (affective apprehension) of a future event that could not otherwise be anticipated through any known inferential process. Precognition and premonition are themselves special cases of a more general phenomenon: the anomalous retroactive influence of some future event on an individual's current responses, whether those responses are conscious or nonconscious, cognitive or affective. This article reports 9 experiments, involving more than 1,000 participants, that test for retroactive influence by "time-reversing" well-established psychological effects so that the individual's responses are obtained before the putatively causal stimulus events occur. Data are presented for 4 time-reversed effects: precognitive approach to erotic stimuli and precognitive avoidance of negative stimuli; retroactive priming; retroactive habituation; and retroactive facilitation of recall. The mean effect size (*d*) in psi performance across all 9 experiments was 0.22, and all but one of the experiments yielded statistically significant results. The individual-difference variable of stimulus seeking, a component of extraversion, was significantly correlated with psi performance in 5 of the experiments, with participants who scored above the midpoint on a scale of stimulus seeking achieving a mean effect size of 0.43. Skepticism about psi, issues of replication, and theories of psi are also discussed.

Keywords: psi, parapsychology, ESP, precognition, retrocausation

What did the Bem study do?

- Reported on 9 different experiments of precognition (ESP)

What did the Bem study do?

- Reported on 9 different experiments of precognition (ESP)
- A typical example: participants see two images and pick one; the computer than randomly picks one of the images as the "correct" image, and the participant sees a "positive" image as a reward (or a "negative" image as punishment) if they happened to select the right one in advance.

What did the Bem study do?

- Reported on 9 different experiments of precognition (ESP)
- A typical example: participants see two images and pick one; the computer than randomly picks one of the images as the "correct" image, and the participant sees a "positive" image as a reward (or a "negative" image as punishment) if they happened to select the right one in advance.
- Participants picked the right image 51.7% of the time ($t = 2.39$, $p = .009$)

How did he prove that ESP exists?!

- Every time you do a hypothesis test, there is a 5% chance that you'll get $p < .05$ just by chance!

How did he prove that ESP exists?!

- Every time you do a hypothesis test, there is a 5% chance that you'll get $p < .05$ just by chance!
- Bem reports on 9 experiments that came out with $p < .05$, but he doesn't report on all of the failed attempts! This is called the file-drawer problem.

How did he prove that ESP exists?!

- Every time you do a hypothesis test, there is a 5% chance that you'll get $p < .05$ just by chance!
- Bem reports on 9 experiments that came out with $p < .05$, but he doesn't report on all of the failed attempts! This is called the file-drawer problem.
- Many researchers later tried to replicate Bem's results, and failed.

1. The consequence of running too many hypothesis tests
2. Changing your analysis strategy based on a hypothesis test
3. Stopping an experiment based on a hypothesis test
4. Using *p*-values responsibly
5. Survivorship bias
6. Practical versus statistical significance

Running A/B testing until we get significance

- Consider a similar situation: we run an A/B test where a customer is presented with two different offers.

Running A/B testing until we get significance

- Consider a similar situation: we run an A/B test where a customer is presented with two different offers.
- After each customer, we test the null hypothesis that the proportion of customers that prefer offer A is the same as the proportion that prefer offer B.

Running A/B testing until we get significance

- Consider a similar situation: we run an A/B test where a customer is presented with two different offers.
- After each customer, we test the null hypothesis that the proportion of customers that prefer offer A is the same as the proportion that prefer offer B.
- We try this with up to 10,000 customers, but stop if we get a *p*-value less than .05.

Let’s simulate a situation where the customers in the population are just selecting randomly between offers A and B.

Let's simulate a situation where the customers in the population are just selecting randomly between offers A and B.

In theory, the proportion of customers choosing offer A should stay near 0.5 and we should get through all 10,000 customers with $p > .05$ each time.

The pink line shows the *p*-value after each customer, and the blue line shows the proportion of customers that chose offer A.

- Running the hypothesis test after each customer provides too many opportunities for false positives.
- Better: decide in advance how long you'll run the test.

1. The consequence of running too many hypothesis tests
2. Changing your analysis strategy based on a hypothesis test
3. Stopping an experiment based on a hypothesis test
4. Using *p*-values responsibly
5. Survivorship bias
6. Practical versus statistical significance

Two phases of analysis

- Think of any analysis you do as having two phases, rather than one:
  - The exploratory phase where you try out different conditions, different subgroups, different measures of success, etc. until you find something that you *think* works.
  - The confirmatory phase where you try to replicate your results in a new sample.

The exploratory phase in regression analysis

- Try looking at a large number of variables.
- Try looking at different subsets—e.g. only women, or only men; only large companies, or only small companies; etc.
- Use low *p*-values as a guide to what *might* generalize to the larger population, but take them with a grain (cannister?!) of salt.

The confirmatory phase in regression analysis

- Using what you think is the best model from the exploratory phase, see if your results generalize to a completely new sample.
- Now you can trust that your *p*-values are not misleading, since you are only running a single test on this sample!

1. The consequence of running too many hypothesis tests
2. Changing your analysis strategy based on a hypothesis test
3. Stopping an experiment based on a hypothesis test
4. Using *p*-values responsibly
5. Survivorship bias
6. Practical versus statistical significance

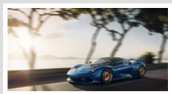
This is how many fund managers actually beat index funds

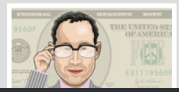
Published: May 13, 2017 10:46 a.m. ET

Roughly 1 in 20 actively managed domestic funds beat index funds

MOST POPULAR


Baby boomers commit the '7 deadly sins' of retirement planning


Move over, Ferrari and Lamborghini — this Italian electric car's 1,900 horsepower makes it faster than a Formula 1 racer


My dying father gave his girlfriend his debit card to buy food — and she

Which funds beat the market?

- In any given year, the vast majority of managed funds trail the S&P 500.

Which funds beat the market?

- In any given year, the vast majority of managed funds trail the S&P 500.
- Some managed funds do outperform the S&P 500—but the high performers are different every year.

Which funds beat the market?

- In any given year, the vast majority of managed funds trail the S&P 500.
- Some managed funds do outperform the S&P 500—but the high performers are different every year.
- Why are there no funds that consistently outperform the market?

Survivorship bias

- The worst performers are the ones most likely to be shut down every year.

Survivorship bias

- The worst performers are the ones most likely to be shut down every year.
- So the situation is even worse than it looks!

Survivorship bias

- The worst performers are the ones most likely to be shut down every year.
- So the situation is even worse than it looks!
- This phenomenon is called survivorship bias: the survivors are not necessarily representative of the whole.

Survivorship bias

- The worst performers are the ones most likely to be shut down every year.
- So the situation is even worse than it looks!
- This phenomenon is called survivorship bias: the survivors are not necessarily representative of the whole.
- Survivorship bias is why we should be careful about trying to emulate the characteristics of the most successful people or companies!

1. The consequence of running too many hypothesis tests
2. Changing your analysis strategy based on a hypothesis test
3. Stopping an experiment based on a hypothesis test
4. Using *p*-values responsibly
5. Survivorship bias
6. Practical versus statistical significance

- We show design A to 10,000 customers and design B to 10,000 customers.

- We show design A to 10,000 customers and design B to 10,000 customers.
- Design A results in a mean of $40 spent, with an SD of $20.

- We show design A to 10,000 customers and design B to 10,000 customers.
- Design A results in a mean of $40 spent, with an SD of $20.
- Design B results in a mean of $41 spent, with an SD of $20.

- We show design A to 10,000 customers and design B to 10,000 customers.
- Design A results in a mean of $40 spent, with an SD of $20.
- Design B results in a mean of $41 spent, with an SD of $20.
- A one-sample t-test of the null hypothesis $\mu_A = \mu_B$ against $\mu_A < \mu_B$ gives a p-value of 0.0002.

- We show design A to 10,000 customers and design B to 10,000 customers.
- Design A results in a mean of $40 spent, with an SD of $20.
- Design B results in a mean of $41 spent, with an SD of $20.
- A one-sample t-test of the null hypothesis $\mu_A = \mu_B$ against $\mu_A < \mu_B$ gives a p-value of 0.0002.
- So we should redesign the store with design B, right?

But:

- What if it would cost millions of dollars to do, and we would never recoup our investment?

But:

- What if it would cost millions of dollars to do, and we would never recoup our investment?
- What if there is another design that would cost just as much to implement but would deliver a much higher ROI?

Practical vs statistical significance

- We tend to misuse *p*-values as a proxy for "important," but *p*-values really just tell us whether our sample is consistent with the null hypothesis being true.

Practical vs statistical significance

- We tend to misuse *p*-values as a proxy for "important," but *p*-values really just tell us whether our sample is consistent with the null hypothesis being true.
- A small *p*-value means that we can be confident that designs A and B do not result in exactly the same revenue per user—is the difference statistically significant…

Practical vs statistical significance

- We tend to misuse *p*-values as a proxy for "important," but *p*-values really just tell us whether our sample is consistent with the null hypothesis being true.
- A small *p*-value means that we can be confident that designs A and B do not result in exactly the same revenue per user—is the difference statistically significant…
- …but that's not answer to the question we really have: is the difference large enough for us to care about—is the difference practically significant?

- What “practical” means depends on your business situation: is the benefit worth the cost?

- What “practical” means depends on your business situation: is the benefit worth the cost?
- Design B might be statistically significantly better than design A, but the online store redesign might cost millions to execute.

- What “practical” means depends on your business situation: is the benefit worth the cost?
- Design B might be statistically significantly better than design A, but the online store redesign might cost millions to execute.
- One fund might outperform the market in a statistically significant way, but the difference might be partially (or completely!) offset by the fund’s fees.