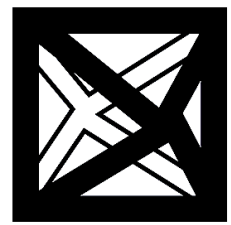
Lecture 3
The law of selection

22 April 2020

Aldo Solari
University of Milano-Bicocca
Statistical Inference II
PhD in Economics and Statistics



The law of selection

You can make things as likely as you want if you choose after the event

David J. Hand

Abraham Lincoln and John F. Kennedy

1. both were assassinated
2. both were killed on a Friday
3. both in the presence of their wives
4. both were shot in the head from behind
5. Lincoln was killed in Ford's Theatre, while Kennedy was killed in a car made by the Ford Motor Company
6. both had four children
7. both had a son who died while they were president
8. Lincoln had a personal secretary named John, and Kennedy had one named Lincoln
9. Lincoln became president in 1861 and Kennedy in 1961
10. John Wilkes Booth (Lincoln's assassin) was born in 1839 Lee Harvey Oswald (Kennedy's assassin) was born in 1939
11. Both Lincoln and Kennedy were succeeded by presidents named Johnson who, wait for it, were born in 1808 and 1908

What is *your* explanation?

- ☐ I don't know
- ☐ It's serendipity
- ☐ It's synchronicity
- ☐ It's a coincidence. The 11 matches (both..both..both..) were selected from a large number of potential pairs, most of which did not match: the names of their mothers, the birth dates of their mothers, the heights of their wives, the dates on which they got married, whether they were bearded or not, the precise nature of their religious beliefs, and so on endlessly. Also, being assassinated, shot in the head and shot in the presence of wife are three highly correlated events

The Baltimore stockbroker scam

1st mail

Tomorrow Bitcoin stock goes UP

2nd mail

I've told you. And tomorrow Bitcoin stock goes UP again.

3rd mail

Be careful. Tomorrow Bitcoin stock goes DOWN

...

11th mail

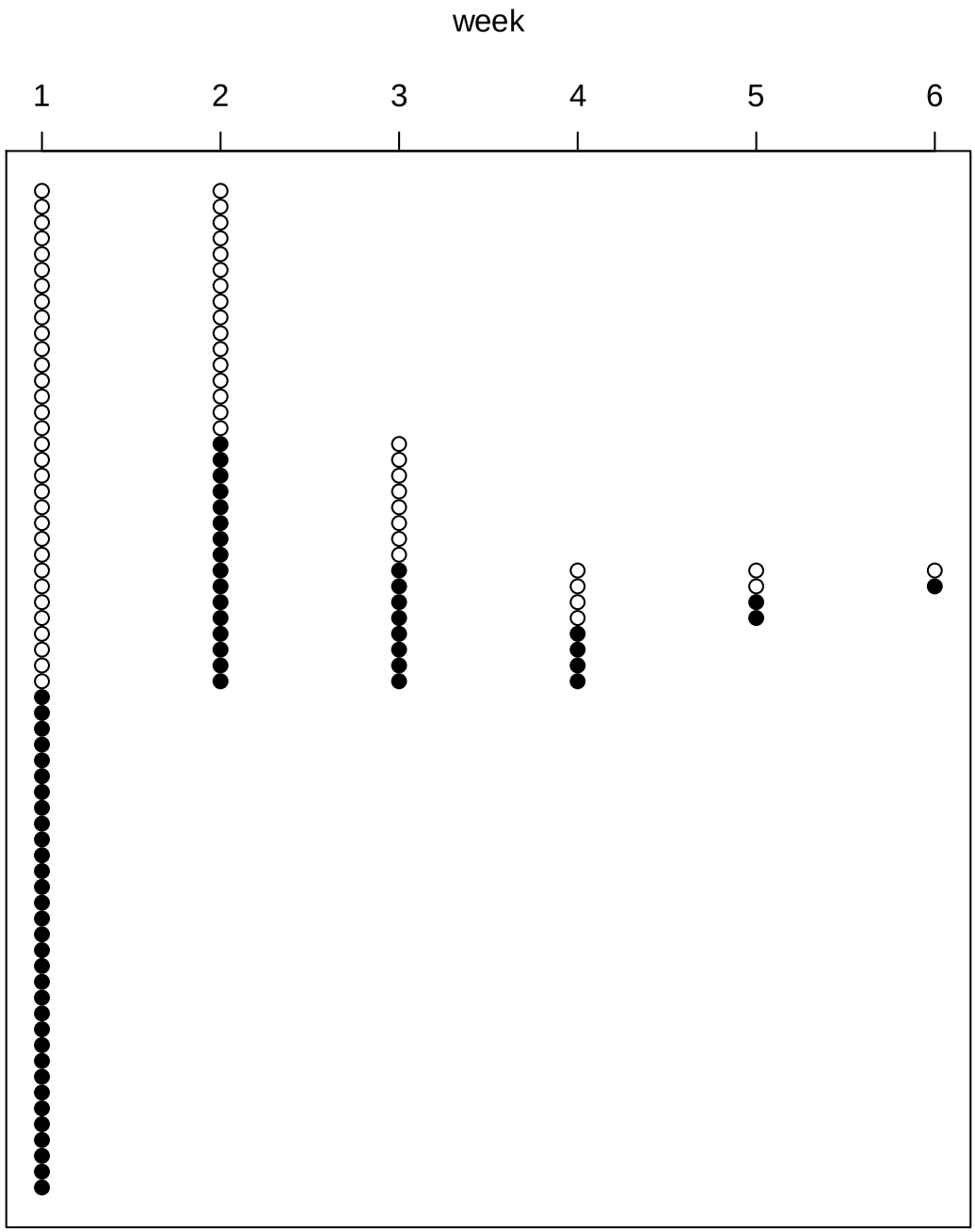
If you want to know if Bitcoin will go UP/DOWN tomorrow, pay me a FEE. Cheers, Anonymous Baltimore stockbroker

How he/she guessed right 10 times in a row?

- ☐ I don't know
- ☐ A lucky strike
- ☐ He/she is an hacker
- ☐ He/she knows the market
- ☐ The correct prediction was selected from a large number of predictions

The strategy

- The scammer begins with a large pool of people (2^x)
- The scammer divides the pool into two halves, and sends all the people a prediction about the future outcome of a binary event (stock price rising/falling)
- One half receives a prediction that the stock price will rise, and the other half receives the opposite prediction
- After the event occurs, the scammer repeats the process with the group that received a correct prediction
- After x iterations, the "surviving" person has received a sequence of x correct predictions, whereupon the scammer then offers another prediction, this time for a fee


week
1
2
3
4
5
6

Survivorship bias

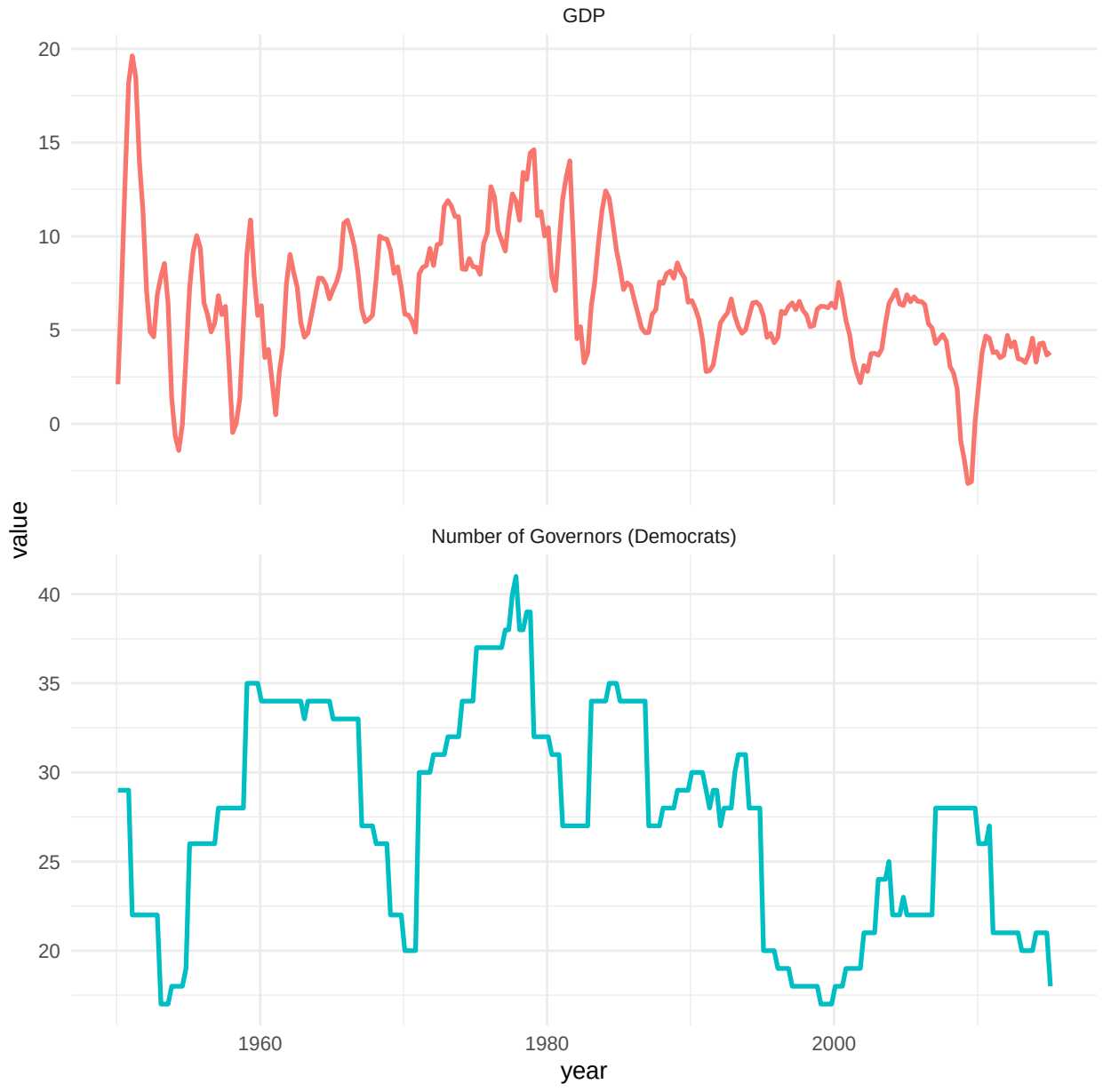
Logical error of concentrating on the people or things that made it past some selection process and overlooking those that did not, typically because of their lack of visibility.

References
Hand, David J. (2014)
The improbability principle: Why coincidences, miracles, and rare events happen every day.
Scientific American/Farrar, Straus and Giroux

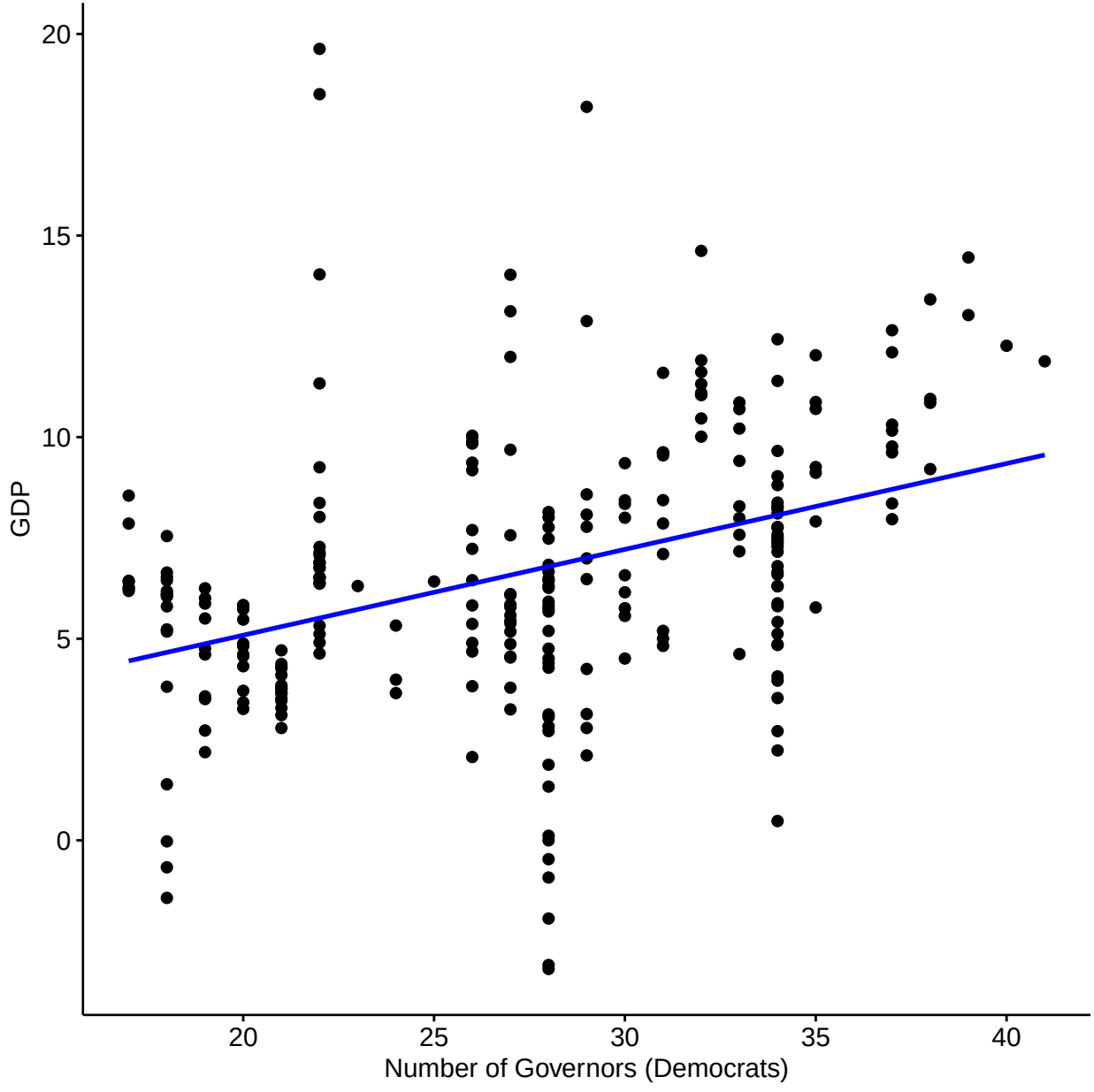
Democrats vs. Republicans: which is better for the U.S. Economy?

- You are a scientist with a question: Is the U.S. economy 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.

Time series


GDP
20
15
10
5
0
value
Number of Governors (Democrats)
40
35
30
25
20
1960
1980
2000
year

Scatter plot



Statistical significance

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.8299784	0.91074910	0.911314	3.629773e-01
gov_dem	0.2128750	0.03255596	6.538742	3.286125e-10

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

Selection

Y: How do you want to measure economic performance?

☐ Employment ☐ Inflation ☒ **GDP** ☐ Stock prices

X: Which politicians do you want to include?

☐ Presidents ☒ **Governors** ☐ Senators ☐ Representatives

References
Christie Aschwanden (2015)
Science Isn't Broken. It's just a hell of a lot harder than we give it credit for.
FiveThirtyEight
`https://fivethirtyeight.com/features/science-isnt-broken/`

Try to support the Republicans

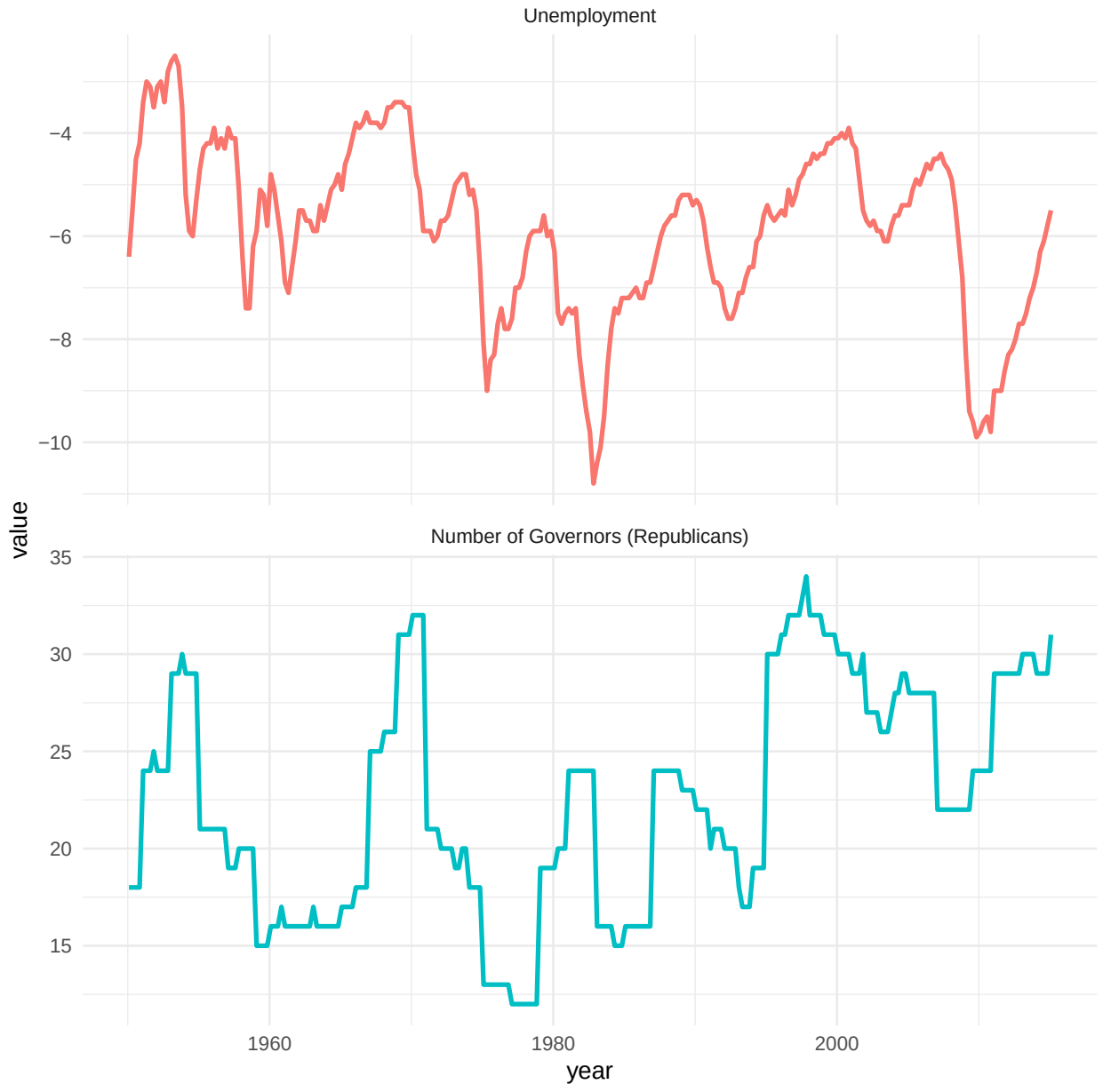
Y: How do you want to measure economic performance?

☒ **Employment** ☐ Inflation ☐ GDP ☐ Stock prices

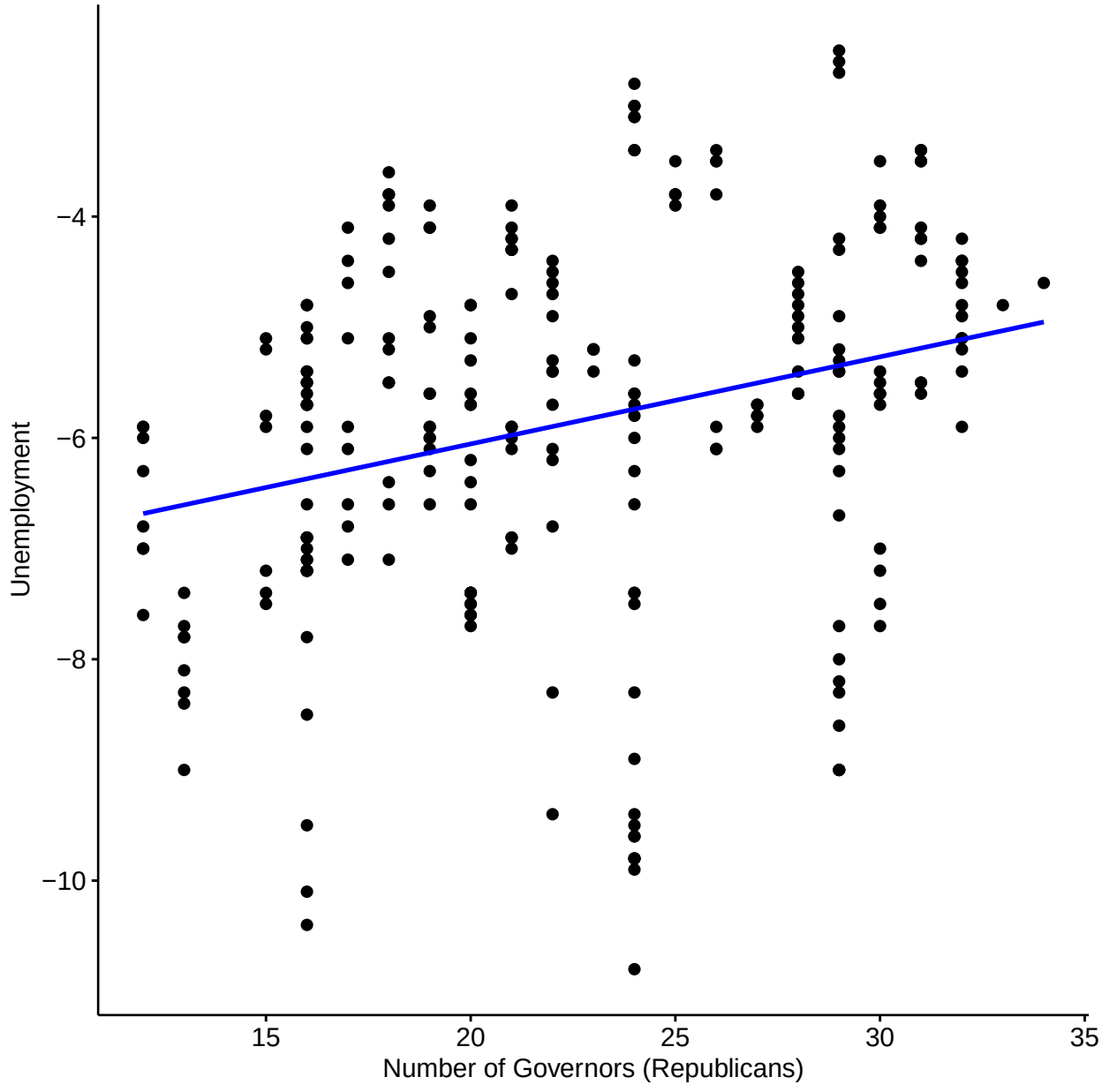
X: Which politicians do you want to include?

☐ Presidents ☒ **Governors** ☐ Senators ☐ Representatives

Time series


Unemployment
−4
−6
−8
−10
Number of Governors (Republicans)
35
30
25
20
15
value
1960
1980
2000
year

Scatter plot



Opposite conclusion

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -7.62868955 0.39902797 -19.11818 2.043845e-51
gov_gop      0.07870434 0.01717247   4.58317 7.128718e-06
```

Now you achieved a p-value of less than 0.01 and showed that Republicans have a positive effect on the economy.

P-hacking & HARKing

- Which political party is best for the economy seems like a pretty straightforward question. But as you saw, it's much easier to get a *result* than it is to get an *answer*
- The data can be selected (*p*-hacked) to make either hypothesis appear correct (Hypothesizing After the Results are Known). This was possible because of the large number of possible combinations
- Answering even a simple scientific question requires lots of choices that can shape the results. This doesn't mean that science is unreliable. It just means that it's more challenging than we sometimes give it credit for

The “quiet scandal” in the statistical community

In a regression problem with many explanatory variables, it is common practice to

1. select the variables to be used in the model
2. evaluating the selected model (as if it had been given a priori)

all with the *same* data

Leo Breiman (1993) referred to this practice as to the “quiet scandal” in the statistical community

Hitters data

We wish to predict a baseball player's Salary by using 19 variables associated with performance in the previous year:

```
"AtBat"     "Hits"      "HmRun"     "Runs"
"RBI"       "Walks"     "Years"     "CAtBat"
"CHits"     "CHmRun"    "CRuns"     "CRBI"
"CWalks"    "League"    "Division"  "PutOuts"
"Assists"   "Errors"    "NewLeague"
```

The variable selection algorithm (best AIC) selects 5 variables:

```
"HmRun"   "Runs"   "Walks"   "CRuns"    "CWalks"
```

out of $2^{19} = 524288$ possible selections

Post-selection inference

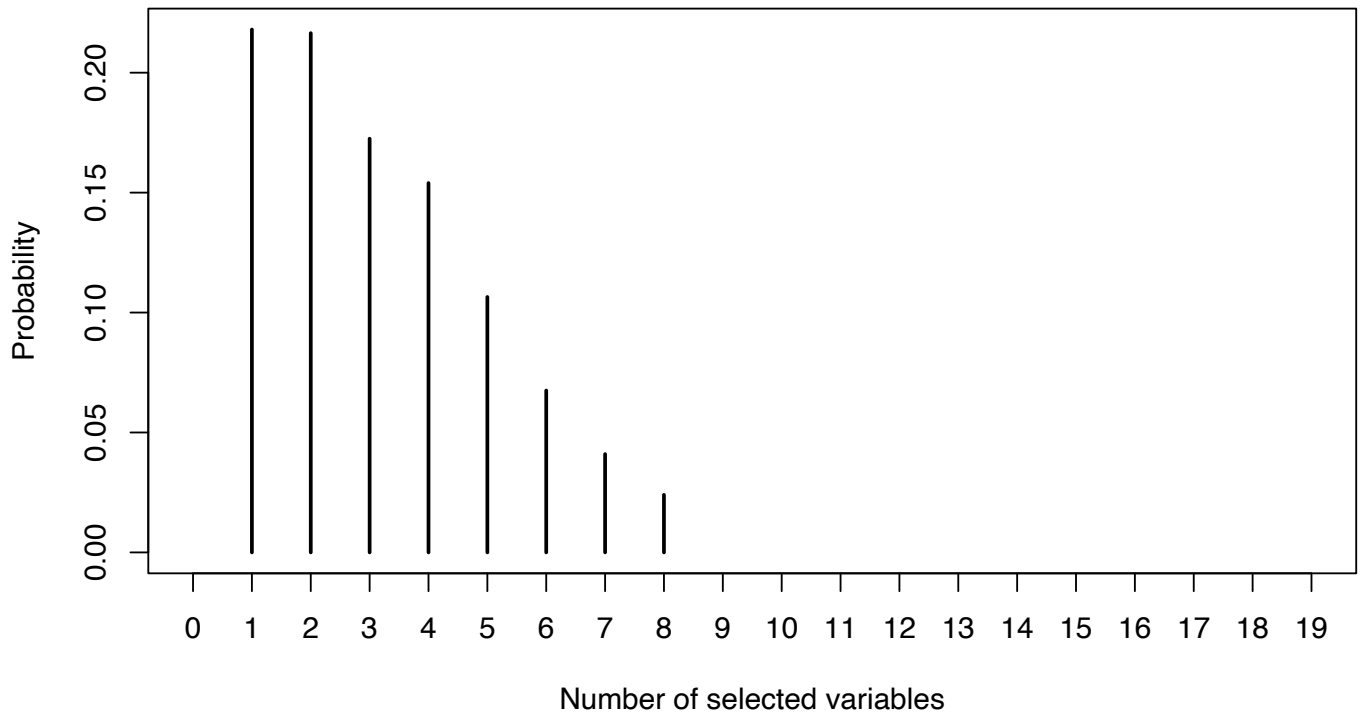
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 631.9910    85.0090   7.434  1.5e-11 ***
HmRun       -13.1438     5.2434  -2.507  0.01348 *
Runs          8.7373     2.7076   3.227  0.00160 **
Walks       -11.2012     3.0252  -3.703  0.00032 ***
CRuns        -0.9092     0.3507  -2.592  0.01068 *
CWalks        1.2993     0.4613   2.816  0.00565 **
---
Multiple R-squared:  0.1361
F-test p-value: 0.002537
```

However, the response values were randomly *permuted*.
This means that the true model contains only the intercept term

On new data

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 593.2505   105.6424   5.616 1.18e-07 ***
HmRun         3.7647     6.3749   0.591   0.5559
Runs         -3.6964     3.1006  -1.192   0.2354
Walks         3.1364     3.4645   0.905   0.3670
CRuns         0.7261     0.4033   1.800   0.0742
CWalks       -1.0194     0.5587  -1.825   0.0704
---
Multiple R-squared:  0.02983
F-test p-value: 0.5651
```

Number of (wrongly) selected variables



Based on 2000 resamples

The problem of post-selection inference

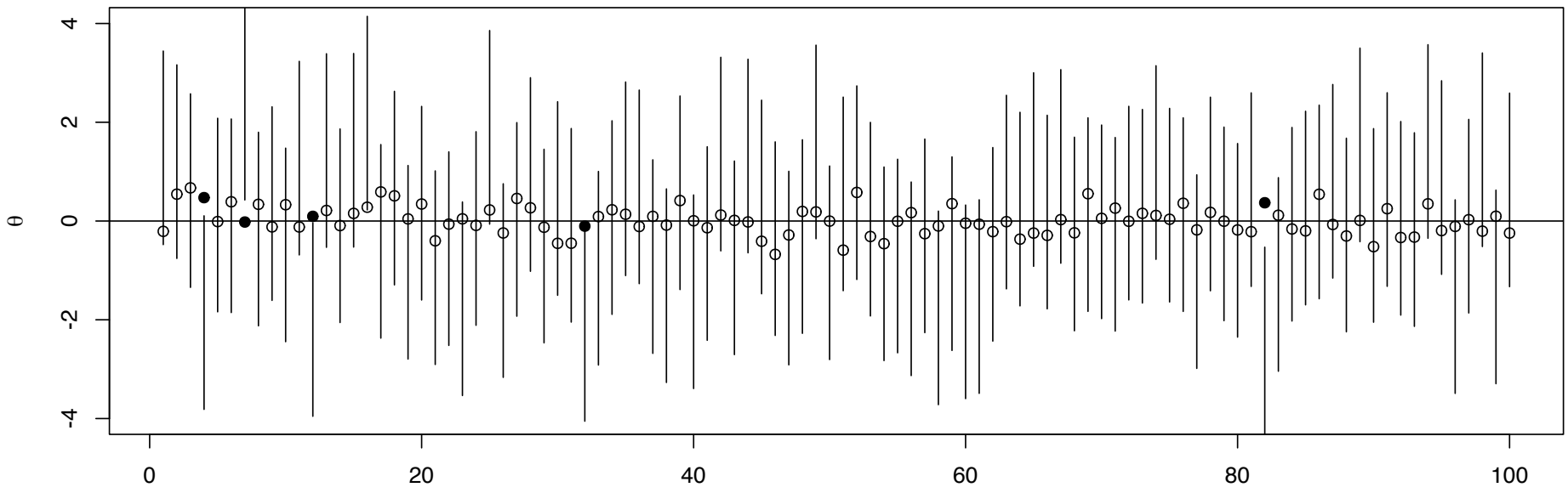
- Classical inference assumes the model is chosen independently of the data (the model is fixed a priori)
- Using the data to select the model introduces additional uncertainty (the model is itself a random variable)
- Need to correct the over-optimism in inference due to data-driven selection

“In a large number of 95% confidence intervals, 95% of them contain the population parameter [...]
but it would be wrong to imagine that the same rule also applies to a large number of 95% **interesting** confidence intervals”

Branko Sorić (1989)
Statistical “Discoveries” and Effect-Size Estimation
Journal of the American Statistical Association

Confidence intervals

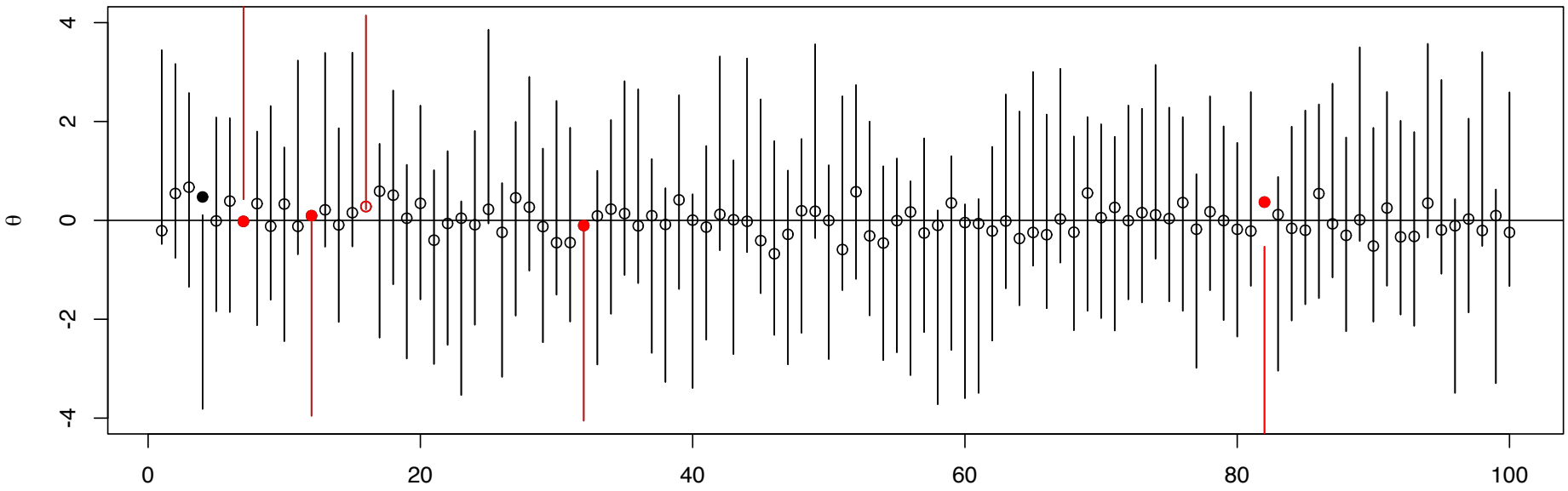
Suppose we have 100 parameters and we construct 95% confidence intervals (CIs)



As expected, 95/100 = 95% CIs cover the respective parameters

Selected confidence intervals

Suppose we select the *interesting* CIs (e.g. CIs not including zero)



Only 1/5 = 20% CIs covers the selected parameters

Eat pizza and avoid prostate cancer

Headlines in newspapers in 1995 announced that eating

tomato sauce, *tomatoes*, and *pizza*

would decrease the risk of prostate cancer

Giovannucci et al.

- The source of the news was a publication by a group led by Edward Giovannucci from Harvard

  Giovannucci, Ascherio, Rimm, Stampfer, Colditz, Willett (1995)
  Intake of carotenoids and retinol in relation to risk of prostate cancer
  Journal of the National Cancer Institute

  Cited by 1878 on March 2020

- They claim that the *lycopene*, a carotenoid hydrocarbon found in tomatoes, was related to lower risk of prostate cancer

Experimental results

- 47894 subjects (free of diagnosed cancer)
- Semiquantitative food-frequency questionnaire on years 1986, 1988, 1990, and 1992
- Between 1986 and 1992, 812 new cases of prostate cancer
- 4 significant associations with cancer out of 46
- 3 significant associations (tomato sauce, tomatoes, pizza) contain lycopene
- tomato sauce $p = 0.001$ tomatoes $p = 0.03$ pizza $p = 0.05$

How likely is this result?

- 4 p-values < 0.05 out of 46
- Suppose that none of the 46 types of food is associated with prostate cancer. Then we can expect an average of 2.3 p-values < 0.05

$$\mathrm{E}\Big(\sum_{i=1}^{46} \mathbb{1}\{p_i \leq 0.05\}\Big) = 46 \cdot 0.05 = 2.3$$

  where $\mathbb{1}\{\cdot\}$ is the indicator function
- If we assume that the p-values are i.i.d. Uniform(0,1), $V = \sum_{i=1}^{46} \mathbb{1}\{p_i \leq 0.05\} \sim \text{Binomial}(46, 0.05)$ and

$$\mathrm{P}\Big(V \geq 4\Big) = 0.196$$

  having at least 4 p-values < 0.05 is not an unlikely event
- It seems that Giovannucci et al. did not have a theory about lycopene when they started the study

Confirmatory vs exploratory

Confirmatory inference

1. Define hypotheses/models/questions
2. Collect data
3. Perform inference

Exploratory inference

1. Collect data
2. Select hypotheses/models/questions
3. Perform inference

Modern practice is highly exploratory (big dataset)
Elementary statistical textbooks often ignore selection issues
As a consequence, inference may be wrong and misleading

AIDSVAX study

- In late February 2003, the world heard preliminary results of the first large scale human trial of a vaccine designed to prevent HIV infection
- A clinical trial was conducted among 5009 people. Of those, 1/3 got a placebo, and 2/3 got the AIDSVAX B/B vaccine:
- Scientists then followed the two groups over three years to see whether the vaccinated group had fewer HIV infections than the placebo group

$$H_0 : \pi_{\text{placebo}} = \pi_{\text{treatment}}$$

Placebo (Infections/Total):	98/1679	(5.8%)
Vaccine (Infections/Total):	191/3330	(5.7%)

Subgroup analysis

- The original study was not designed to determine whether the vaccine was efficacious in any subgroup
- VaxGen, the vaccine's maker, described the result as disappointing but presented an analysis of subgroup data, and claimed "a statistically significant reduction of HIV infection in certain racial groups"

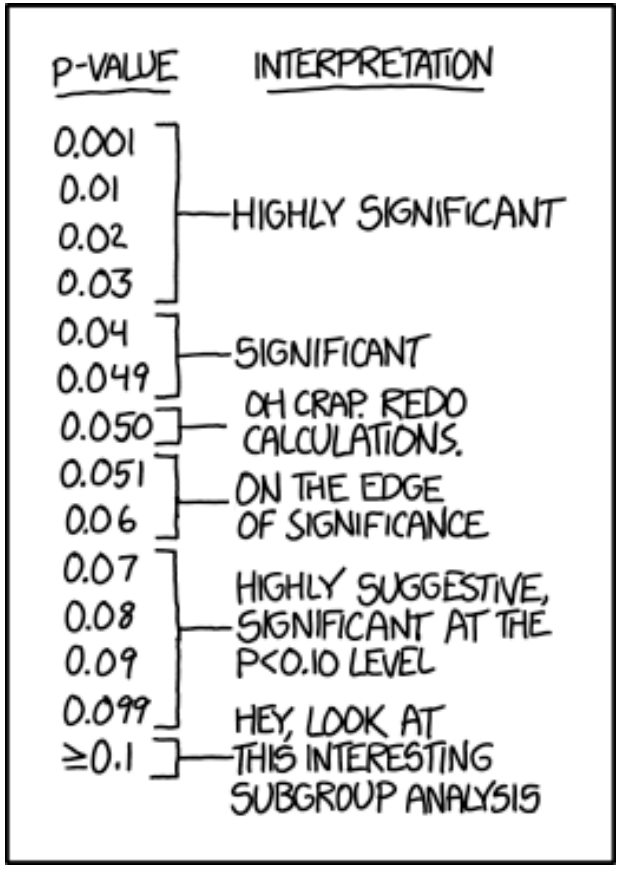
	White/Hispanic	Black	Asian/Pacific Islanders	Other
Placebo	81/1508	9/111	2/20	6/40
Vaccine	179/3003	4/203	2/53	6/71
p-value	0.456	0.015	0.663	0.345

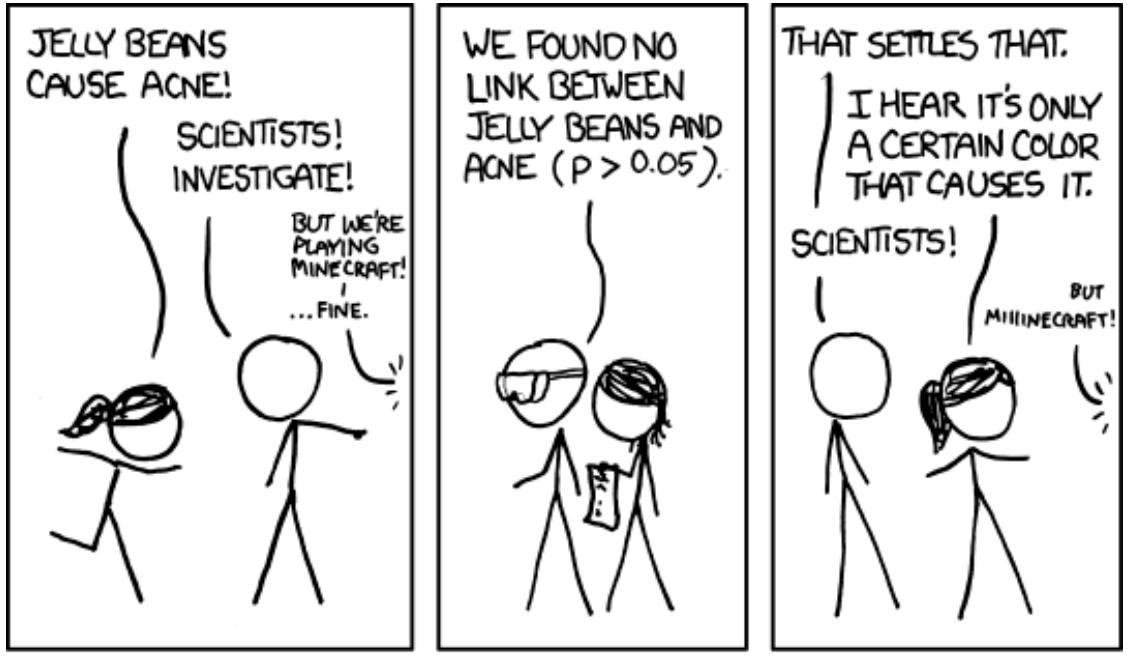
Critical thinking

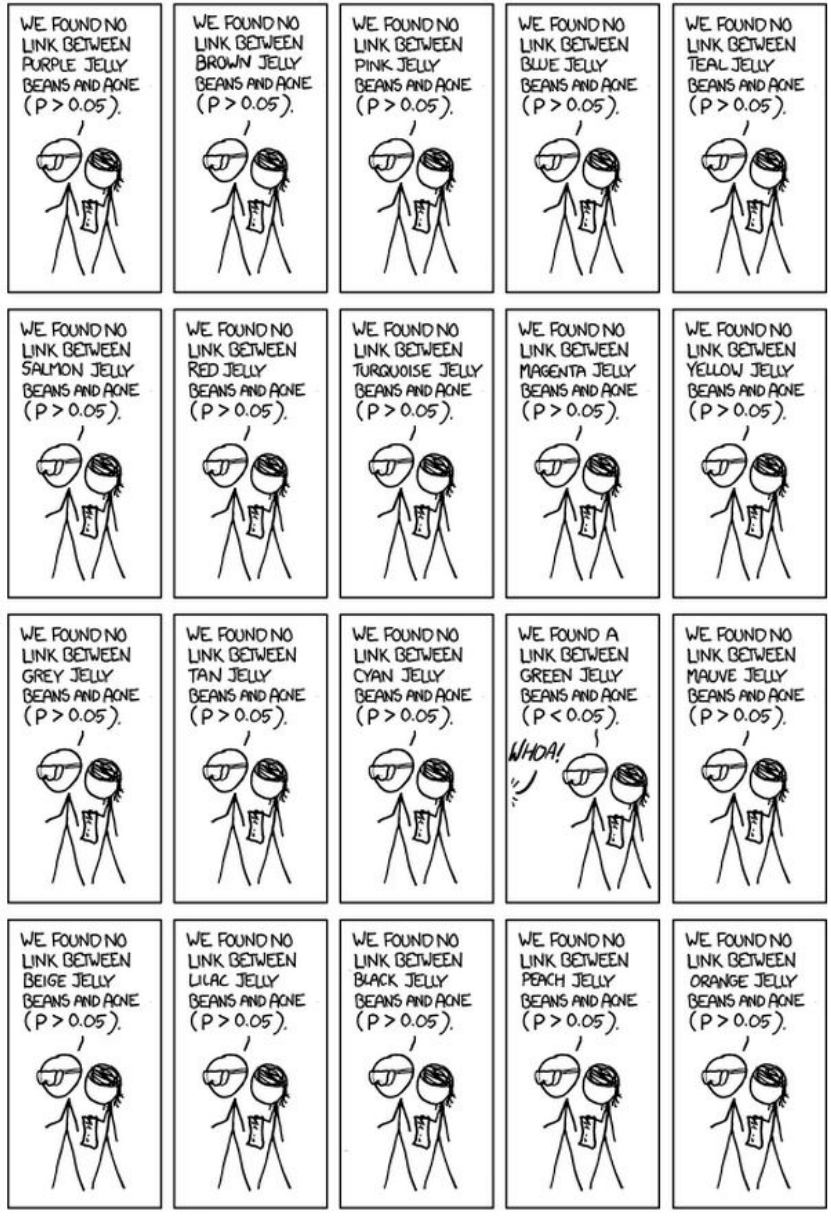
- Dr. Anthony Fauci cautioned against jumping to conclusions without further analysis:
- "Professional statisticians warn us that one must be very careful in doing subset analyses when the primary endpoint of a given study shows no effect"
- "Therefore, one really cannot say at all that the vaccine is effective in Blacks without a very careful scrutiny of the data and the statistical analysis of the data"
- "In this context, most statisticians say that with the penalties that one must apply in this type of subset analysis, the results in the Black subset would not be statistically significant"

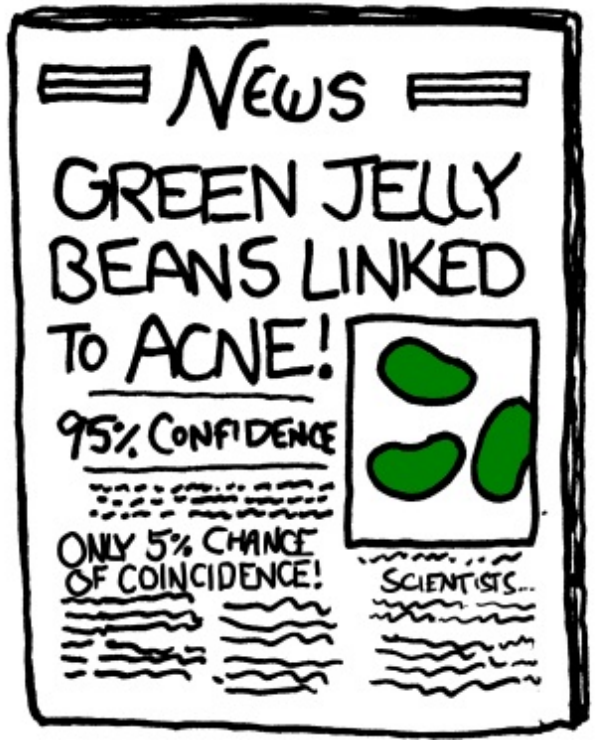
p-value	0.456	0.015	0.663	0.345
(Sidak) adjusted p-value	0.912	0.059	0.987	0.816

xkcd


P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
0.099
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS


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 MIIIINECRAFT!


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).

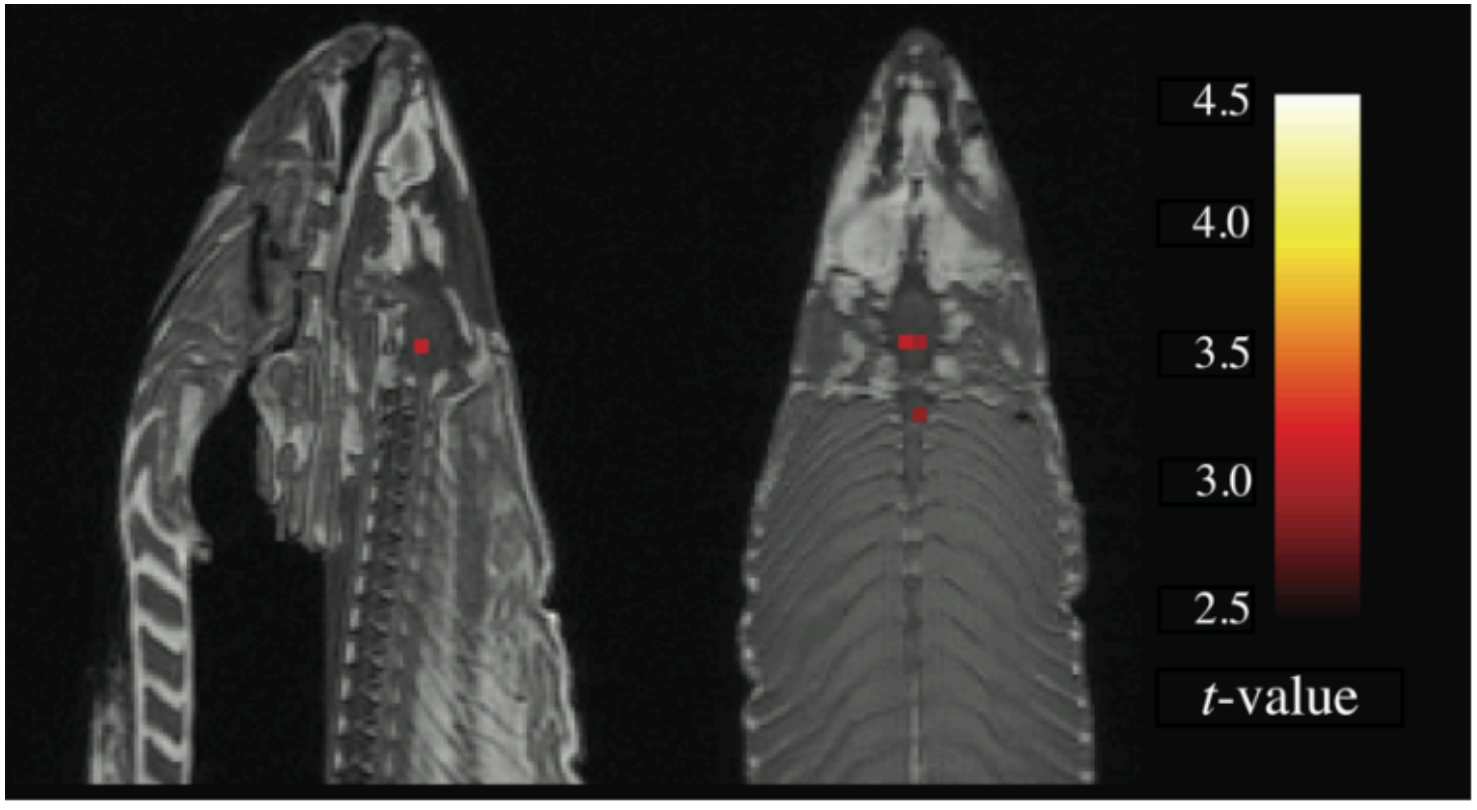

News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

The dead salmon study

- In 2009, a highly remarkable scientific experiment was performed by four American brain researchers
- They used functional magnetic resonance imaging (fMRI) to determine which brain areas respond to emotional stimuli
- The researchers were able to clearly identify a brain area that showed a response to the stimulus offered

Bennett, Baird, Miller, Wolford (2010)
Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction
Journal of Serendipitous and Unexpected Results

(Cited by 401)


4.5
4.0
3.5
3.0
2.5
t-value

Large-scale testing

- However, the subject was a dead salmon
- To find brain regions, 8064 t-tests were performed at $\alpha = 0.001$, and 16 were found significant
- Apparently, standard imaging techniques with standard analysis methods could produce clearly nonsensical results
- How likely is this result?

$$\mathbb{E}\Big(\sum_{i=1}^{8064} \mathbb{1}\{p_i \leq 0.001\}\Big) = 8.064$$

95% confidence interval (assuming independence): $[3, 14]$

Tukey's higher criticism

- Tukey introduced the notion of the higher criticism by means of a story
- A young psychologist administers many hypothesis tests as part of a research project, and finds that, of 250 tests 11 were significant at the 5% level
- The young researcher feels very proud of this fact and is ready to make a big deal about it, until a senior researcher (Tukey himself?) suggests that one would expect 12.5 significant tests even in the purely null case, merely by chance
- In that sense, finding only 11 significant results is actually somewhat disappointing!

What most surprised you?

- ☐ Abraham Lincoln and John F. Kennedy
- ☐ The Baltimore stockbroker scam
- ☐ Democrats vs. Republicans: which is better for the U.S. Economy?
- ☐ The "quiet scandal" in the statistical community
- ☐ Eat pizza and avoid prostate cancer
- ☐ AIDSVAX study
- ☐ The dead salmon study