Hypothesis testing

James Scott (UT-Austin)

Reference: “Data Science” Chapter 7

Outline

- An introductory example
- The four steps of every hypothesis test
- Two approaches: Fisher vs. Neyman-Pearson

The Patriots

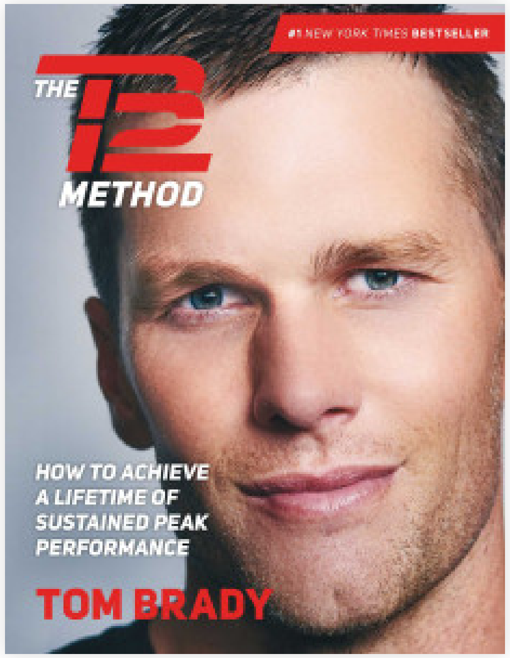
Unless you're from a narrow strip of land from Connecticut to Maine, you probably dislike the New England Patriots.

The Patriots

First of all, they win too much.

The Patriots

Then there's Tom Brady, their star quarterback…



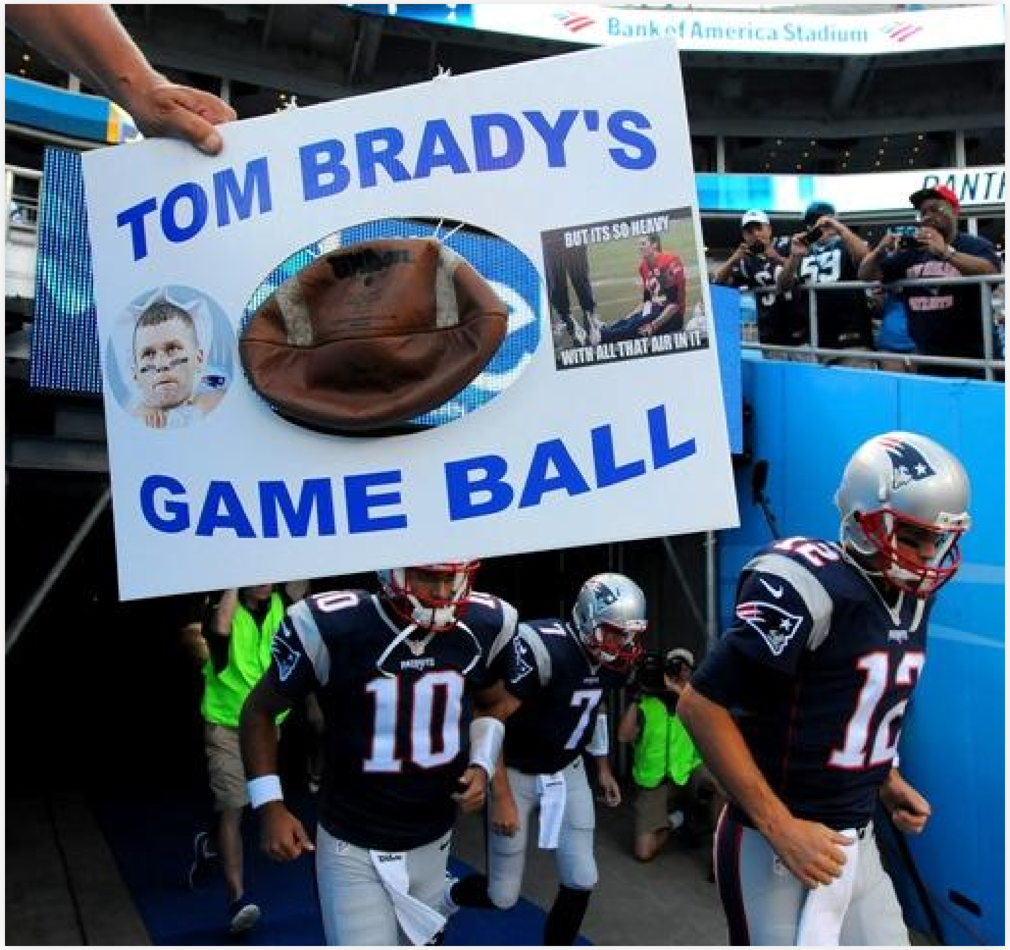
"The more hydrated I am, the less likely I am to get sunburned." – TB12

The Patriots

And Bill Belichick, their coach…

The Patriots

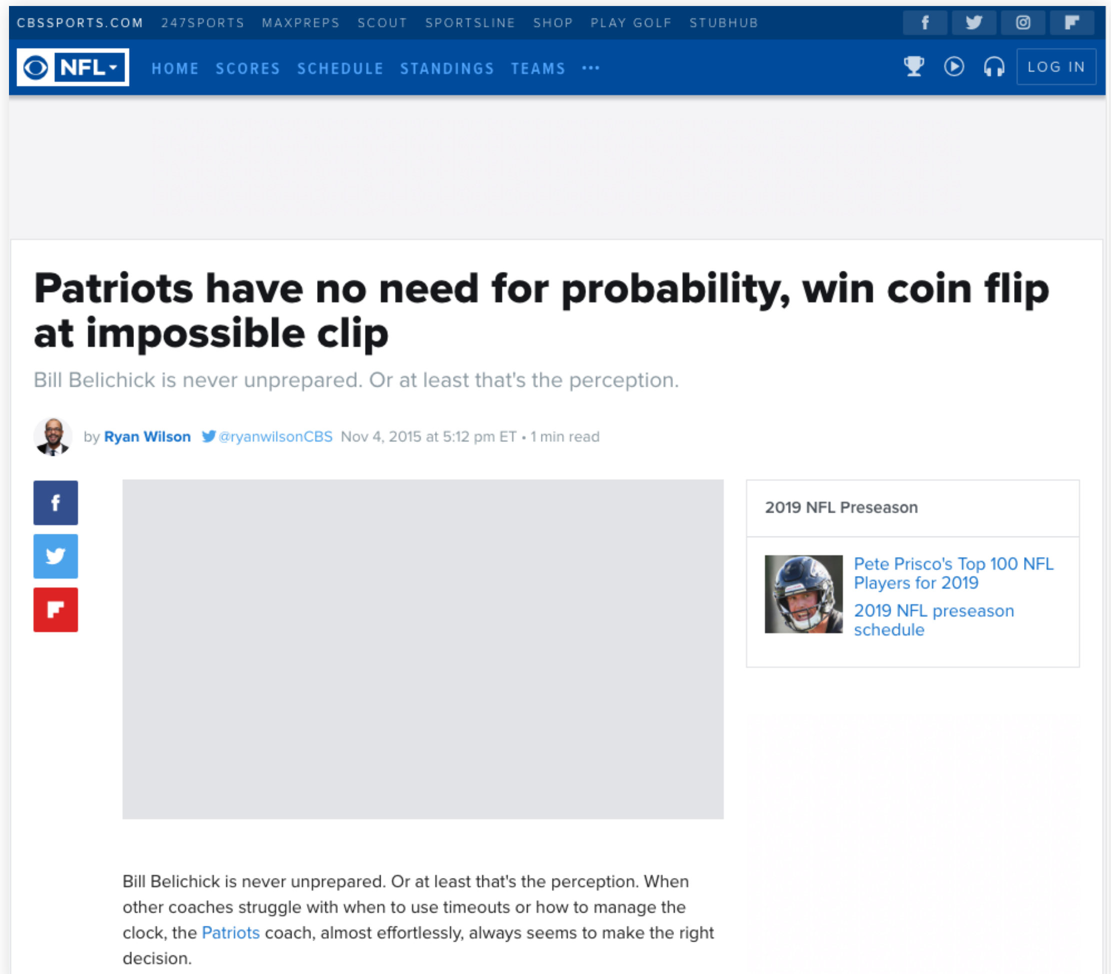
And of course, the cheating!



The Patriots

But could even the Patriots cheat at the *pre-game coin toss?*

The Patriots



CBSSPORTS.COM 247SPORTS MAXPREPS SCOUT SPORTSLINE SHOP PLAY GOLF STUBHUB

NFL HOME SCORES SCHEDULE STANDINGS TEAMS ··· LOG IN

Patriots have no need for probability, win coin flip at impossible clip

Bill Belichick is never unprepared. Or at least that's the perception.

by Ryan Wilson @ryanwilsonCBS Nov 4, 2015 at 5:12 pm ET • 1 min read

2019 NFL Preseason

Pete Prisco's Top 100 NFL Players for 2019

2019 NFL preseason schedule

Bill Belichick is never unprepared. Or at least that's the perception. When other coaches struggle with when to use timeouts or how to manage the clock, the Patriots coach, almost effortlessly, always seems to make the right decision.

The Patriots

For a 25-game stretch during the 2014 and 2015 NFL seasons, the Patriots won the pre-game coin toss 19 out of 25 times, for a suspiciously high winning percentage of 76%.

The Patriots

"Use the Force…"

The Patriots

But before we invoke religion or the Force to explain this fact, let's consider the innocent explanation first: blind luck.

- If you toss a coin over and over again, you'll see some long streaks with more heads, and some with more tails, just by luck.
- Is it plausible that the Patriots just went on a lucky 25-game streak?

The Patriots

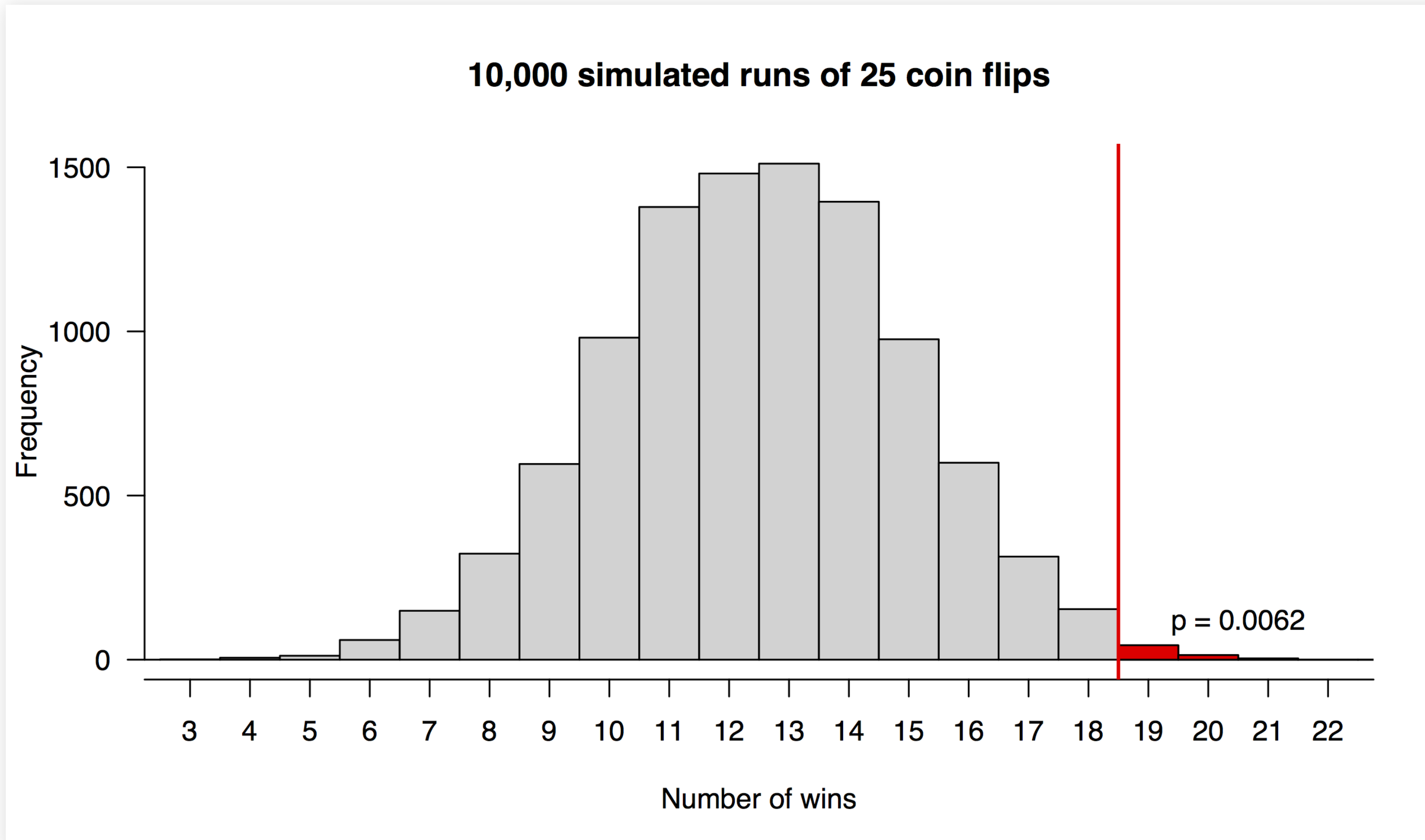
To the code in `patriots.R`! Let's simulate some coin flips.

Summary

This simple example has all the major elements of *hypothesis testing*:

1. We have a *null hypothesis*, that the pre-game coin toss in the Patriots' games was truly random.
2. We use a *test statistic*, number of Patriots' coin-toss wins, to measure the evidence against the null hypothesis.
3. We calculated the probability distribution of the test statistic, assuming that the null hypothesis is true. Here, we just ran a Monte Carlo simulation of coin flips, assuming an unbiased coin.
4. Finally, we used this probability distribution to assess whether the null hypothesis looked believable in light of the data.

Summary



10,000 simulated runs of 25 coin flips
Frequency
1500
1000
500
0
p = 0.0062
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Number of wins

Summary

All hypothesis testing problems have these same four elements.

1. A null hypothesis H_0.
2. A test statistic $T \in \mathcal{T}$ that summarizes the data and measures the evidence against the null hypothesis. Larger values of T mean stronger evidence.
3. $P(T \mid H_0)$: the sampling distribution of the test statistic, assuming that the null hypothesis is true. *This provides context for our measurement in step 2.*
4. An assessment: in light of what we see in step 3, does our test statistic look plausible under the null hypothesis?

Two schools of thought

Within this basic framework, there are two schools of thought about how to proceed.

1. The Fisherian approach: step 4 is about **summarizing the evidence after the fact.** Key terms: p-*value*.
2. The Neyman-Pearson approach: step 4 is about **making a decision with ex-ante performance guarantees.** Key terms: rejection region, α level, power curve.

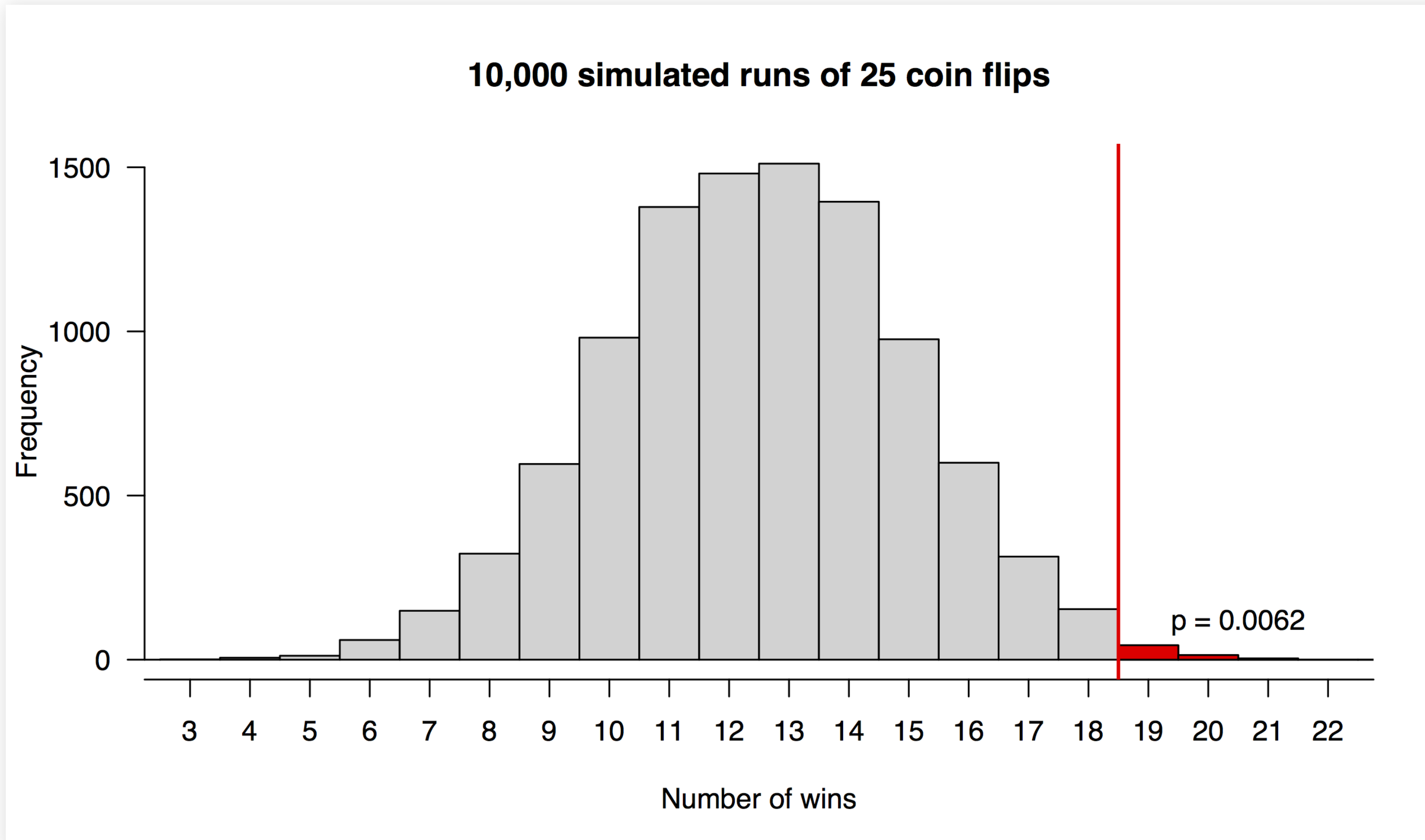
Fisher's approach

Suppose our observed test statistic is t_{ob}. In step 4, we should report the quantity

$$p = P(T \geq t_{ob} \mid H_0)$$

Fisher called this the p-value: the probability that, if the null hypothesis were true, we would observe a test statistic T at least as large as the value we actually observed (t_{ob}).

Recall the Patriots' example



Fisher's approach

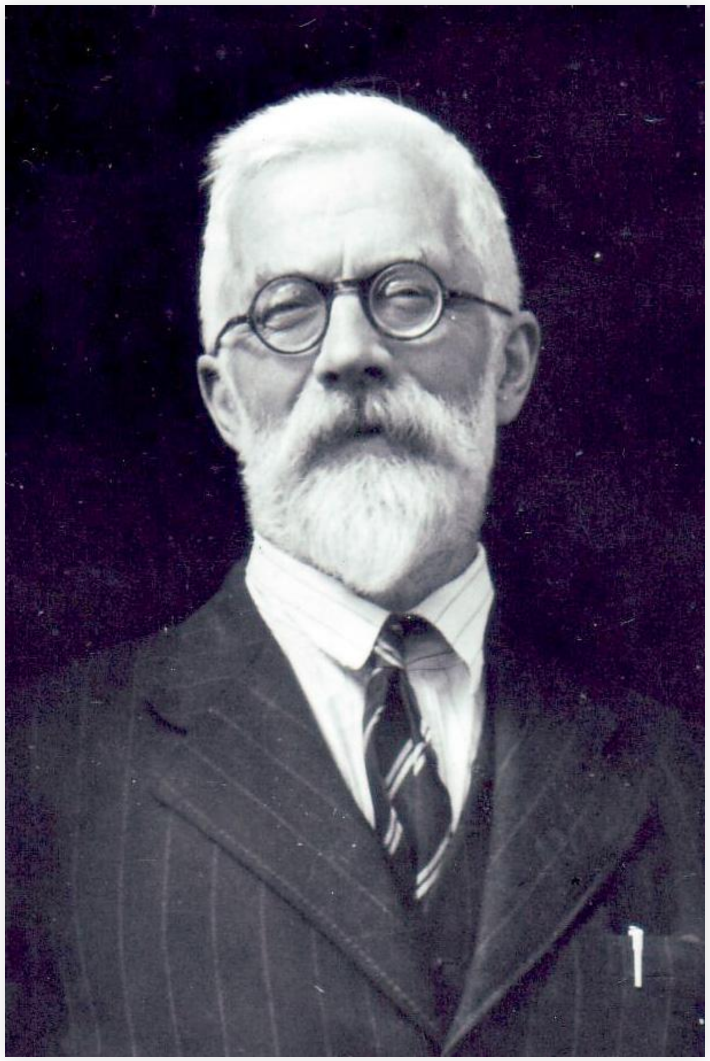
The p-value summarizes the strength of evidence provided by the data against the null hypothesis.

- p closer to 0: data less likely under the null, so the null is more likely to be wrong. **Stronger evidence against H_0.**
- p further from 0: data more likely under the null. **Weaker evidence against H_0.**

According to Fisher: job done! *Report the p-value and let your readers make whatever they will of it.*

What do you mean by "close to 0"?

What do you mean by "close to 0"?



“Mwa-ha-ha-ha-ha-ha!”

“You're on your own, suckahs!”

p-values in the real world

The biggest advantage of p-values is that they provide a sliding scale of evidence against the null hypothesis: small p-values mean stronger evidence.

Problem 1: they are hard to interpret.

"I got a p-value of 0.02, so there's a 2% chance that the null hypothesis is right."

Wrong:

- $p = P(T \geq t_{ob} \mid H_0)$
- $p \neq P(H_0 \mid t_{ob})$

Remember: conditional probabilities aren't symmetric!

Problem 1: they are hard to interpret

"Bob got a p-value of 0.1, but I got a p-value of 0.01. My null hypothesis is ten times less likely to be true than Bob's is."

Wrong:

- the p-value does not directly measure the likelihood that the null hypothesis is true.
- p-values are bizarre like that: they're numbers, but you can't really compare them like numbers!
- A p-value that's ten times smaller doesn't mean that you have ten times stronger evidence that the null is false.

Problem 1: they are hard to interpret.

"I got a p-value of 0.02. There's only a 2% chance I would have observed my test statistic if the null hypothesis were true."

Wrong:

- $p = P(T \geq t_{ob} \mid H_0)$
- $p \neq P(T = t_{ob} \mid H_0)$

Remember: the p-value is the probability of observing the test statistic you actually observed, **or any more extreme test statistic**, assuming H_0 is true.

Problem 2: people oversimplify them

Because p-values are hard to interpret, people tend to impose arbitrary cut-offs for what counts as a "significant" p-value.

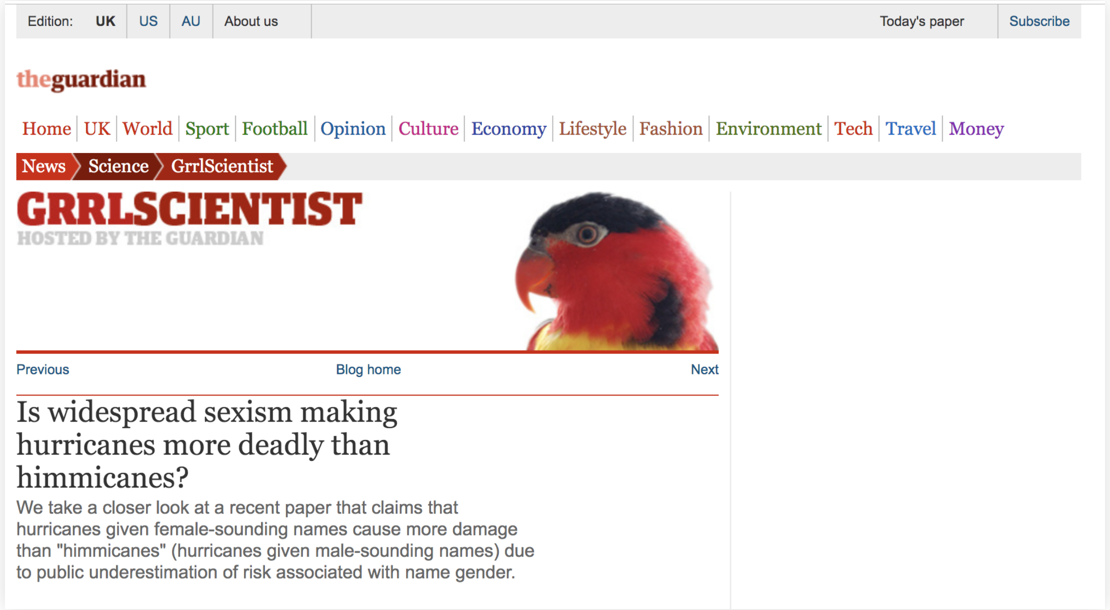
Psychologists, for example, will generally publish research findings for which $p < 0.05$.

Problem 2: people oversimplify them

Then again, psychologists will believe anything.

Problem 2: people oversimplify them

Then again, psychologists will believe anything.


Edition: UK US AU About us Today's paper Subscribe

theguardian

Home | UK | World | Sport | Football | Opinion | Culture | Economy | Lifestyle | Fashion | Environment | Tech | Travel | Money

News > Science > GrrlScientist

GRRLSCIENTIST

HOSTED BY THE GUARDIAN

Previous Blog home Next

Is widespread sexism making hurricanes more deadly than himmicanes?

We take a closer look at a recent paper that claims that hurricanes given female-sounding names cause more damage than "himmicanes" (hurricanes given male-sounding names) due to public underestimation of risk associated with name gender.

Problem 2: people oversimplify them

Then again, psychologists will believe anything.

Search Sign In

SLATE News & Politics Culture Technology Business Human Interest

PLAYS SECOND BANANA TO NO ONE FOLLOW US

SCIENCE

Too Good to Be True

Statistics may say that women wear red when they're fertile …
but you can't always trust statistics.

By ANDREW GELMAN JULY 24, 2013 • 12:37 PM

Problem 2: people oversimplify them

Physicists are a bit more skeptical; they generally publish results only when $p < 0.000001$.

For example, the p-value in the paper announcing the discovery of the Higgs boson was 1.7×10^{-9}.

They spent a lot of time and money (> $1 billion) collecting more data, even after the evidence was really, really strong.

EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

CERN-PH-EP-2012-218

Accepted by: Physics Letters B

Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC

The ATLAS Collaboration

This paper is dedicated to the memory of our ATLAS colleagues who did not live to see the full impact and significance of their contributions to the experiment.

Abstract

A search for the Standard Model Higgs boson in proton-proton collisions with the ATLAS detector at the LHC is presented. The datasets used correspond to integrated luminosities of approximately 4.8 fb^{-1} collected at $\sqrt{s} = 7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s} = 8$ TeV in 2012. Individual searches in the channels $H \to ZZ^{(*)} \to 4\ell$, $H \to \gamma\gamma$ and $H \to WW^{(*)} \to e\nu\mu\nu$ in the 8 TeV data are combined with previously published results of searches for $H \to ZZ^{(*)}$, $WW^{(*)}$, $b\bar{b}$ and $\tau^+\tau^-$ in the 7 TeV data and results from improved analyses of the $H \to ZZ^{(*)} \to 4\ell$ and $H \to \gamma\gamma$ channels in the 7 TeV data. Clear evidence for the production of a neutral boson with a measured mass of $126.0 \pm 0.4\,(\text{stat}) \pm 0.4\,(\text{sys})$ GeV is presented. This observation, which has a significance of 5.9 standard deviations, corresponding to a background fluctuation probability of 1.7×10^{-9}, is compatible with the production and decay of the Standard Model Higgs boson.

Neyman's criticisms of p-values

1. Nobody except Fisher knows how to interpret them.

2. Rejecting a null hypothesis isn't meaningful unless we have some alternative hypothesis in mind. Since people will inevitably use a p-value to make a binary decision ("null" versus "alternative"), we should formalize that decision process.

Neyman-Pearson testing

The Neyman-Pearson approach is aimed at quantifying (and controlling) the error probabilities associated with a hypothesis test.

- False positive: rejecting H_0 when it is actually true. ("Type I error")
- False negative: retaining H_0 when it is actually false. ("Type II error")

Neyman-Pearson testing

In Neyman Pearson testing, we have a modified sequence of steps:

1. Specify H_0 (the null hypothesis) **and $\mathbf{H_A}$ (an alternative hypothesis.)**
2. Choose your test statistic T taking values in $\mathcal{T}$.
3. Calculate $P(T \mid H_0)$.

Neyman-Pearson testing

Before looking at the observed test statistic t_{ob} *for your actual data,* continue as follows.

4a. Specify a rejection region $R \subset \mathcal{T}$.
4b. Calculate $\alpha = P(T \in R \mid H_0)$. This is called the alpha level or *size* of the test.
4c. Calculate the *power* of your test as $P(T \in R \mid H_A)$.
4d. Check whether your observed test statistic, t_{ob}, falls in R. If so, reject H_0 in favor of H_A. If not, retain H_0.

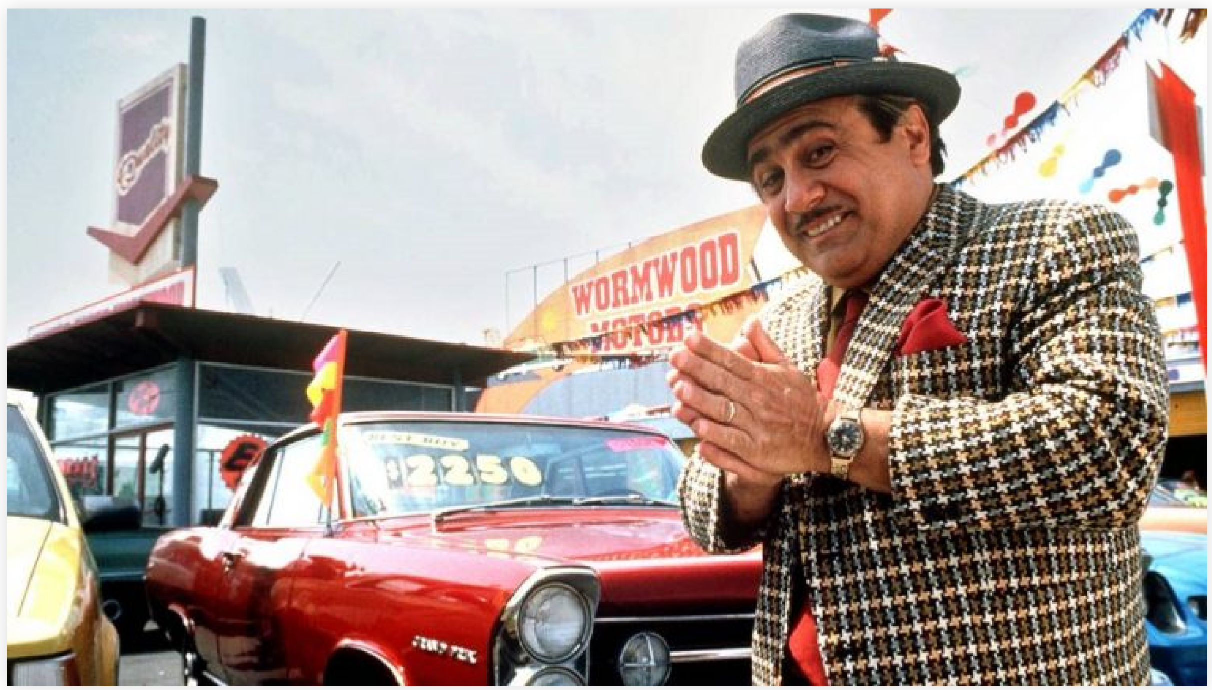
Neyman-Pearson testing

The test is characterized by two properties:

- $\alpha = P(T \in R \mid H_0)$ = size: the probability of falsely rejecting the null hypothesis when it's actually true.
- $\beta = 1-$ Power $= 1 - P(T \in R \mid H_A)$: the probability of retaining the null hypothesis when it's actually false.

Neyman-Pearson testing

Neyman and Pearson are basically asking you: would you buy a car without a warranty?



Then you shouldn't test a hypothesis without one, either!

Neyman-Pearson testing

α and power (or β) serve as the test's "warranty," or specific guarantee of performance:

- Lower α means lower probability of a false positive if the null is actually true.
- High power (low β) means low probability of missing true departures from the null.

These are knowable in advance. As with cars, so too with hypothesis tests: **always check the whole warranty!** If someone only tells you the α level of a test and omits the power, it's like only giving a warranty on part of the car.

Neyman-Pearson testing

At the end of a Neyman-Pearson test, you report two things.

- The warranty: that is, the *size* (α level) and the *power* of the test (or equivalently $\beta = 1-$ power).
- The result of the test: reject or retain ("fail to reject") the null hypothesis.

No p-values! (This was Fisher's criticism: no matter how strong the evidence against the null, an NP test ends up reporting the same thing for any $T \in R$.)

Neyman-Pearson testing

The difficulty of conducting a Neyman-Pearson test depends upon the alternative hypothesis.

- “Simple” alternatives are easy: the power is just a single number.
- “Composite” alternatives are a bit harder: the power is a function of an unknown parameter.

Simple alternative

Let's go back to the Patriots problem. Our test statistic is X, the number of successful coin flips in 25 tries. Suppose that p is the true probability that the Patriots will win the coin toss. Consider testing the two hypotheses:

- $H_0 : p = 1/2$, versus…
- $H_A : p = 2/3$.

Suppose we decide to reject H_0 if $X \geq 17$. In this case the power is easy to calculate: it's just $P(X \geq 17)$ when $X \sim \text{Binom}(N = 25, p = 2/3)$.

Let's look at `power.R` (part 1).

Your turn

Suppose now that we follow the Patriots for a 50-flip stretch and count the number of times X they win the coin toss. As before, our null hypothesis is that $X \sim \mathrm{Binom}(N, p = 0.5)$.

Follow the steps of an NP test:

- pick a rejection region of the form $R = \{X : X \geq c\}$ for some threshold c.
- characterize the α level of the resulting test.
- calculate the power of your test assuming that the Patriots are able to cheat, winning the coin flip with true probability $p = 0.6$.

Composite alternative

Compare this to the more realistic situation where our alternative hypothesis isn't so specific:

- $H_0 : p = 1/2$.
- $H_A : p \neq 1/2$.

This is called a "composite alternative hypothesis." It "hedges its bets," i.e. it doesn't make any specific predictions except that the null hypothesis is wrong.

Composite alternative

Now the power of the test isn't just a single number.

Rather, it's a function, or a **power curve**:

$$\text{Power}(p) = P(X \geq 17 \mid p)$$

where p is the assumed binomial success probability. This is a function of p.

Back to `power.R` (part 2).

Notation for parametric models

Suppose we take data points $X_1, \ldots, X_N$, where each X_i comes from some some parametric probability distribution $p(X \mid \theta)$.

- A null hypothesis usually takes the form $H_0 : \theta = \theta_0$.
- A test statistic $T \in \mathcal{T}$ is a function of the data with range $\mathcal{T}$:

$$T = T(X_1, \ldots, X_N)$$

Notation for parametric models

Suppose our observed test statistic is t_{ob}.

- In a typical one-sided test, the p-value is the probability

$$P(T \geq t_{ob} \mid H_0)$$

- More generically, if we let $\Gamma(t) \subset \mathcal{T}$ denote the set of all possible test statistics that are judged to be "more extreme than" some specific value t, then the p-value is the probability

$$P(T \in \Gamma(t_{ob}) \mid H_0)\,,$$

Notation for parametric models

- A rejection region is a subset of $\mathcal{T}$, the possible outcomes for the test statistic. It generally takes the form

$$R = \{\mathcal{T} : T \geq c\} \quad \text{or} \quad R = \{\mathcal{T} : |T| \geq c\}$$

- We refer to c, the boundary of the rejection region, as the critical value.
- The α level of a test is the probability

$$\alpha = P(T \in R \mid H_0) = P(T \in R \mid \theta = \theta_0)$$

Notation for parametric models

- An alternative hypothesis takes the form $H_A : \theta \in \Theta_A$, where Θ_A is some subset of the parameter space not containing θ_0.

- A simple alternative is where Θ_A contains a single value, whereas composite alternative has multiple possible values. For example, $H_A : \theta \neq 0$ and $H_A : p > 0.5$ are both composite alternatives.

- The power of the test at some specific $\theta_a \in \Theta_A$ is defined as

$$R(\theta) = P(T \in R \mid \theta = \theta_a) = 1 - \beta(\theta)$$

- Calculating the power $R(\theta_a)$ over all values $\theta_a \in \Theta_A$ defines the power curve of the test.

Your turn

Return to our example where we follow the Patriots for a 50-flip stretch and count the number of times X they win the coin toss. Clearly $X \sim \mathrm{Binom}(N, p)$, and our null hypothesis is that $p = 0.5$.

Follow the steps of an NP test for two different rejection regions: $R_1 = \{X : X \geq 30\}$ and $R_2 = \{X : X \geq 34\}$. For each of these two rejection regions, **check the warranty**! That is:

- calculate the α level of the test.
- calculate the power curve of test test for the composite alternative hypothesis $H_A : p > 0.5$.

An important caveat

In Neyman-Pearson testing, it's really important that you specify the rejection region R *in advance*, before seeing the data.

An important caveat

In particular, you **absolutely, positively cannot** do the following:

- look at the data and calculate a p-value. "Hey look, $p = 0.009$: that's small!"
- then retrospectively choose α to be just a little bit bigger than the p-value you found. "I'll choose $\alpha = 0.01$, and claim a warranty against false positives at the 1% level, which is the smallest and most impressive round number that makes my data significant."

An important caveat

In particular, you **absolutely, positively cannot** do the following:

- look at the data and calculate a p-value. “Hey look, $p = 0.009$: that's small!”
- then retrospectively choose α to be just a little bit bigger than the p-value you found. “I'll choose $\alpha = 0.01$, and claim a warranty against false positives at the 1% level, which is the smallest and most impressive round number that makes my data significant.”

That's what cheaters like Tom Brady do. It voids the warranty usually enjoyed by a Neyman-Pearson test.

An important caveat

Why is this cheating?

An important caveat

Why is this cheating?

Because the two key probabilities in NP testing assume that R is fixed and pre-specified, and that only the test statistic T is random from sample to sample:

$$\alpha = P(T \in R \mid H_0) \quad \text{and} \quad \text{Power}(\theta_a) = P(T \in R \mid \theta = \theta_a)$$

If you choose R based on the data, then the *rejection region itself is a random variable*, and all probability statements are off. **This voids the warranty!**

An important caveat

Be like a physicist here.

Physicists (especially particle physicists) are the most honest statisticians in the observable universe. They **always** specify R in advance, and they never change their α level to make their data look maximally impressive.

In my experience, people in most other fields are sloppy, dishonest, or both when it comes to setting α levels.

Testing: best practices

Testing: best practices

Remember the difference between Fisher and Neyman-Pearson, and don't conflate them:

- Fisher: calculate a p-value to summarize the evidence provided by the data against the null hypothesis. **No cutoffs, no alternative hypothesis, no warranty! (no α or β)**
- Neyman-Pearson: set up a pre-defined rejection region for making a decision about whether to retain H_0 or reject it in favor of an alternative H_A. Choose the rejection region by "checking the warranty," i.e. calculating size and power in advance. **No p-values, no distinction among levels of evidence with varying strength.**

(Sidebar)

In fact, this is a great litmus test for probing the depth of someone's statistical knowledge: ask them “What's the difference between the Fisherian and Neyman-Pearson frameworks for hypothesis testing?”

(Sidebar)

In fact, this is a great litmus test for probing the depth of someone's statistical knowledge: ask them "What's the difference between the Fisherian and Neyman-Pearson frameworks for hypothesis testing?"

Most people who can compute a p-value and think they understand statistics will look at you with a blank stare if you ask them this question.

This is like a biologist who can't explain the difference between Darwin's and Lamarck's views on evolution. **Hold yourself to a higher standard.**

Testing: best practices

Be careful about using p-values.

- Most people can't interpret them correctly.
- The p-value does not directly measure how likely it is that the null hypothesis is false. In fact, the p-value *assumes* that the null hypothesis is true!
- The p-value is *not* the probability of a type-I error. **That's the α level!**

Testing: best practices

- You're always on safe ground with a Neyman-Pearson test.
- You get an unambiguous result and a warranty in advance: that is, specific performance guarantees in the form of α and the power curve.
- But remember: the warranty is void if you specify the rejection region **after** seeing the data!
- The big downside is that you don't get a continuous measure of evidence against the null.

Testing: best practices

And finally, the single most important "best practice" of hypothesis testing is:

Testing: best practices

And finally, the single most important "best practice" of hypothesis testing is:

Don't do hypothesis testing.

Testing: best practices

And finally, the single most important “best practice” of hypothesis testing is:

Don't do hypothesis testing.

- Or really: don't do hypothesis testing unless it's really, really obvious you need to.
- In most situations, p-values are lazy and not that useful, and it is more scientifically informative to report a confidence interval for the underlying parameter or difference of interest.

Any time you're about to calculate a p-value, ask yourself: do I really need to? Wouldn't I be better off reporting a confidence interval instead? **Usually the answer is yes!**