SWARTHMORE

# STAT011 Statistical Methods I

## Lecture 26 Advanced Topics

Lu Chen
Swarthmore College
5/2/2019

# Final Exam

## Final Exam

- Tuesday 5/14 9am-12pm **SCI 101**
- Practice problems available on Thursday 5/2

## Important dates

- Office hours
  - Thursday 5/9, 9:30 - 11:30am
- Muses sessions
  - Monday 5/13, 7 - 9pm
- Stat Clinics
  - Friday 5/10, 3 - 6pm
  - Saturday 5/11, 3 - 6pm
  - Sunday 5/12, 6 - 9pm

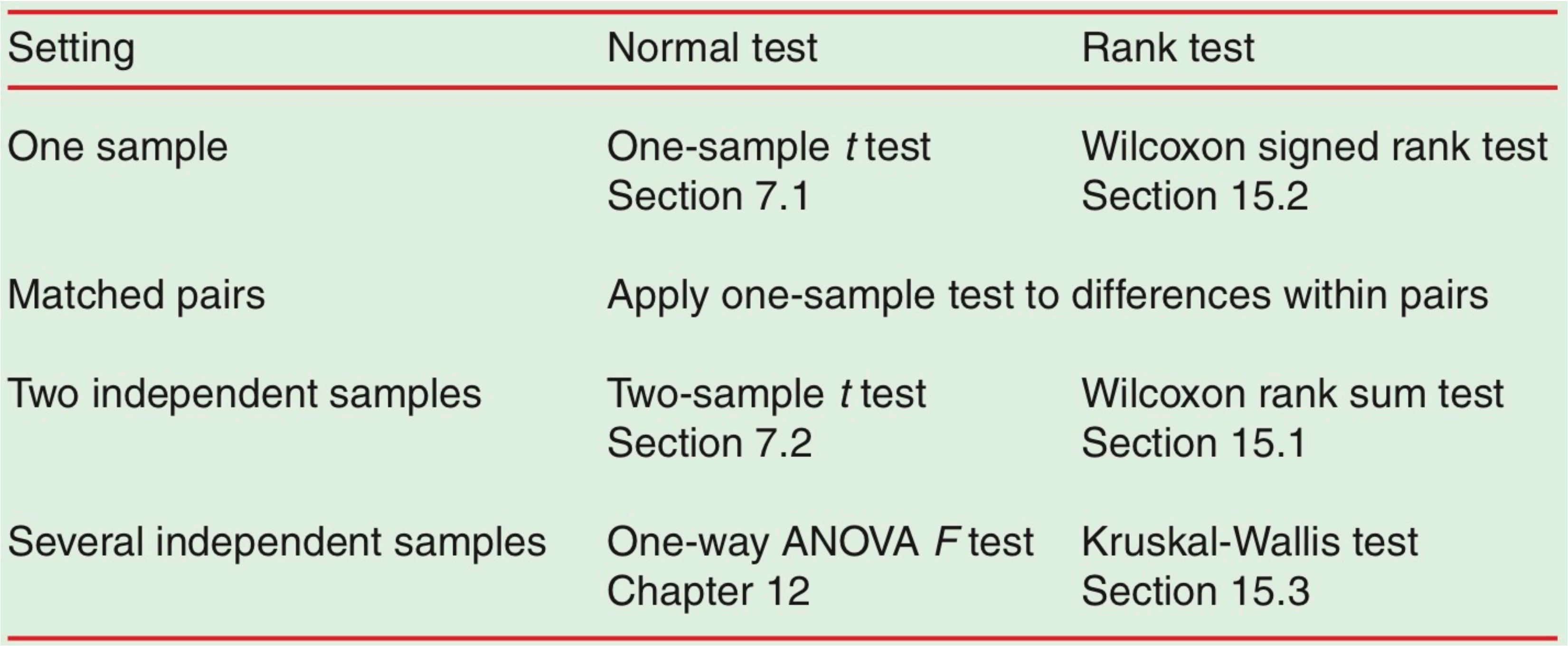
# Outline

- Non-parametric methods
  - Example: the Wilcoxon rank sum test
- Boostrapping
- Permutation test
- $P$-value

# Non-parametric methods

- All the methods we learned in this semester assume certain distributions of the population data and thus involve **population parameters**. We use sample data to make inferences about these population parameters.
- For example, one-sample $t$ test assumes the populations are Normally distributed with mean $\mu$ and SD $\sigma$. We make inference about the **population mean $\mu$** and estimate population SD $\sigma$ using sample SD $s$. If the population distribution is not Normal, it requires a large sample size to use the method.
- These methods are called ***parametric* methods** for we make inferences about the population ***parameters*** in the distributions.
- **Non-parametric methods**, however, have no assumptions about the population distributions thus no parameters involved.
- There are also **semi-parametric methods** available.

# Non-parametric methods



| Setting | Normal test | Rank test |
|---|---|---|
| One sample | One-sample *t* test<br>Section 7.1 | Wilcoxon signed rank test<br>Section 15.2 |
| Matched pairs | Apply one-sample test to differences within pairs | |
| Two independent samples | Two-sample *t* test<br>Section 7.2 | Wilcoxon rank sum test<br>Section 15.1 |
| Several independent samples | One-way ANOVA *F* test<br>Chapter 12 | Kruskal-Wallis test<br>Section 15.3 |

- Non-parametric methods are used when we do not want to assume population distribution or we know the data with a small sample size is definitely not Normally distributed.

# Example: *Coffee ~ ClassYear*

| Class Year | Coffee | Class Year | Coffee |
|---|---|---|---|
| **Fr** | 1 | **Sr** | 5 |
| **Fr** | 5 | **Sr** | 10 |
| **Fr** | 2 | **Sr** | 0 |
| **Fr** | 0 | **Sr** | 0 |

- Do freshmen and seniors drink similar amount of coffee?
- On average, freshmen drink 2 cups of coffee a week; seniors drink 3.75 cups.
- Population distribution is not Normal (usually count data are right skewed).
- Sample size is small ($n_1 = n_2 = 4$).
- Two-sample $t$ procedure is not appropriate.

# The Wilcoxon rank sum test

- Rank all 8 observations together from the smallest to the largest.

| Class Year | Fr | Sr | Sr | Fr | Fr | Fr | Sr | Sr |
|---|---|---|---|---|---|---|---|---|
| Coffee | 0 | 0 | 0 | 1 | 2 | 5 | 5 | 10 |
| Rank | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |

- Assign all tied values the average of the ranks they occupy.

| Class Year | Fr | Sr | Sr | Fr | Fr | Fr | Sr | Sr |
|---|---|---|---|---|---|---|---|---|
| Coffee | 0 | 0 | 0 | 1 | 2 | 5 | 5 | 10 |
| Rank | 2 | 2 | 2 | 4 | 5 | 6.5 | 6.5 | 8 |

# The Wilcoxon rank sum test

- Sum the ranks for each class year

| Class Year | Sum of ranks |
|---|---|
| **Fr** | 17.5 |
| **Sr** | 18.5 |

- Since the total of the ranks is 36, if the two class years of students drink similar amount of coffee in a week, we expect they both have sum of ranks 18.
- The Wilcoxon Rank Sum test statistic measures how far away the observed sum of ranks is from the expected sum of ranks.

# The Wilcoxon rank sum test

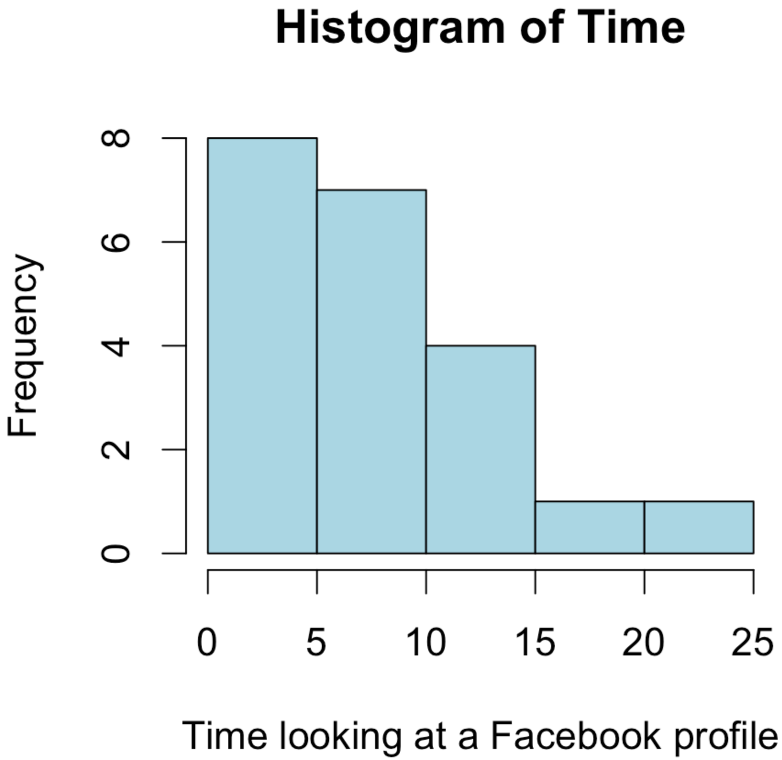
```r
wilcox.test(c(1, 5, 2, 0), c(5, 10, 0, 0))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  c(1, 5, 2, 0) and c(5, 10, 0, 0)
## W = 7.5, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

- When there are ties, the test computes an approximated $P$-value.
- $P = 1 > 0.05$. We cannot reject $H_0$ that the two class years of students drink similar amount of coffee.

# Example: Time looking at a Facebook profile

```
hist(Time, xlab="Time looking at a Facebook profile", col="lightblue")
```



- $n = 21$
- $\bar{x} = 7.87$ mins, $s = 5.65$ mins
- What's the 95% CI for average time looking at a Facebook profile?
- Skewed distribution; small sample size. The one-sample $t$ statistic $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ does not necessarily have a $t$ distribution.
- The cirtical value $t^*$ for confidence interval and the $P$-value based on $t$ distribution for significance test are probably inappropriate.

# Bootstrapping

- We need to know the distribution of $\bar{x}$ since CLT may not hold for this example.
- If we have the population data, we can repeatedly draw samples from the population, calculate the mean $\bar{x}$ for each sample and then examine the distribution of $\bar{x}$.
- However, in reality, we do not have the population data but only a single sample.
- The Boostrap idea is to use the current sample as the population and resamples from the sample.

# Bootstrapping

**The Boostrap Idea**

The original sample is representative of the population from which it was drawn. Thus, resamples from this original sample represent what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on the resamples, represents the sampling distribution of the statistic.

- **Step 1: Resampling**. Create many resamples by repeatedly sampling with replacement from this one random sample. Each resample is the same size as the original random sample.
- **Step 2: Bootstrap distribution**. Compute the statistic of interest (eg., $\bar{x}$) from each sample and examine the bootstrap distribution of the statistic. The bootstrap distribution gives information (that is, shape and spread) about the sampling distribution.

# Bootstrapping

```r
Time # Time looking at a Facebook profile
```

```
##  [1]  8.18 11.22  4.38  3.77 17.77 23.52  0.23  7.95  3.03  5.08  8.07  4.35
## [13]  4.02 13.23  1.32  8.48  1.40  8.33  8.60 12.25 10.10
```

```r
set.seed(2019)
s1 <- sample(Time, size = 21, replace = TRUE); s1 # Resample 1
```

```
##  [1]  1.40  1.32  0.23  4.02 11.22  8.18  8.33  8.18  4.38  4.02  1.40 13.23
## [13] 17.77 17.77  1.32 13.23 11.22  1.32  7.95  5.08  3.03
```

```r
s2 <- sample(Time, size = 21, replace = TRUE); s2 # Resample 2
```

```
##  [1]  4.02 12.25  3.77  4.38  8.18  4.02 12.25 13.23  5.08 13.23  8.18  8.33
## [13]  3.77  8.18 11.22  7.95  5.08  8.18  1.40  4.02 23.52
```

```r
s3 <- sample(Time, size = 21, replace = TRUE); s3 # Resample 3
```

```
##  [1]  8.33  1.32  1.40  8.07 12.25 12.25 17.77 17.77  0.23 11.22  8.18  7.95
## [13]  8.07  7.95 10.10 12.25 11.22 11.22  3.77 17.77  8.48
```

# Bootstrapping

```
mean(Time) # mean of the original sample
```

```
## [1] 7.870476
```

```
mean(s1) # mean of resample 1
```

```
## [1] 6.885714
```

```
mean(s2) # mean of resample 2
```

```
## [1] 8.106667
```

```
mean(s3) # mean of resample 3
```

```
## [1] 9.408095
```

# Bootstrapping

```
set.seed(2019)
mean_Time <- NULL
for(i in 1:1000){
  resample <- sample(Time, 21, replace=TRUE)
  mean_Time[i] <- mean(resample)
}
hist(mean_Time, col="lightblue",
     main="Sampling Distribution")
```

```
c(mean(Time), sd(Time))
```

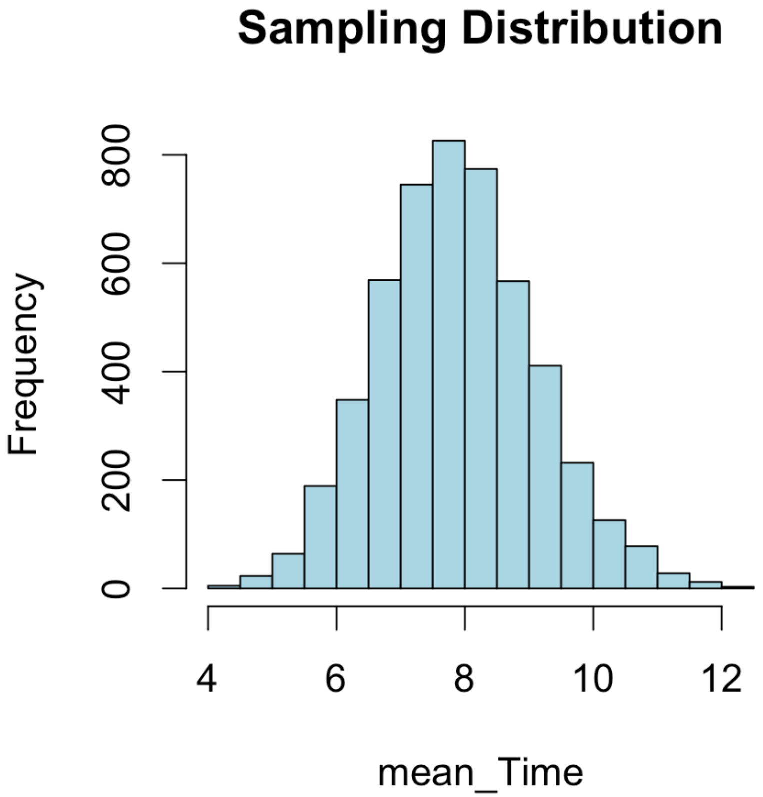
```
## [1] 7.870476 5.651755
```

```
c(mean(mean_Time), sd(mean_Time))
```

```
## [1] 7.881743 1.222393
```

```
quantile(mean_Time, prob=c(0.025, 0.975))
```

```
##    2.5%   97.5%
##  5.6190 10.4672
```



# Bootstrapping

### Bootstrap Percentile Confidence Interval

The interval between the 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% **bootstrap percentile confidence interval** for the corresponding parameter. Use this method when the bootstrap estimate of bias is small.

- The bootstrap percentile confidence interval can be applied not only on mean but also easily on other statistics like median, quartiles, SD, etc.

| Statistic | Original sample | Bootstrap percentile confidence interval |
|---|---|---|
| **Mean** | 7.9 | [5.6, 10.2] |
| **Median** | 8.1 | [4.4, 8.6] |
| **SD** | 5.7 | [3.3, 7.4] |

# Example: Rreading ability

Do new "directed reading activities" improve the reading ability of elementary school students, as measured by their Degree of Reading Power (DRP) scores? A study assigns students at random to either the new method (treatment group, 4 students) or traditional teaching methods (control group, 5 students).

```
Treatment
```

```
## [1] 61 44 67 49
```

```
Control
```

```
## [1] 42 33 46 37 42
```

```
mean(Treatment)
```

```
## [1] 55.25
```

```
mean(Control)
```

```
## [1] 40
```

# Permutation test

Analyzing using regular two-sample $t$ test:

```r
t.test(Treatment, Control, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  Treatment and Control
## t = 2.6482, df = 4.0882, p-value = 0.02791
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  3.049094      Inf
## sample estimates:
## mean of x mean of y
##     55.25     40.00
```

- However, sample size is too small to use the $t$ test. Results may not be trustworthy.

# Permutation test

- **Step 1**. Combine the two groups into one.
- **Step 2**. Randomly choose 4 students out of the 9 (without replacement) to be the treatment group. The rest 5 students form the control group. This is called a **permutation resample**.
- **Step 3**. Calculate the difference in mean score between the treatment and control group.
- **Step 4**. Do Step2 and 3 many times and obtain a series of mean differences.
- **Step 5**. Compare the actual difference observed in the original sample to the distribution of the mean differences.

# Permutation test

```
combined <- c(Treatment, Control) # combine the two groups
diff <- mean(Treatment) - mean(Control) # calculate the actual difference in mean
diff
```

```
## [1] 15.25
```

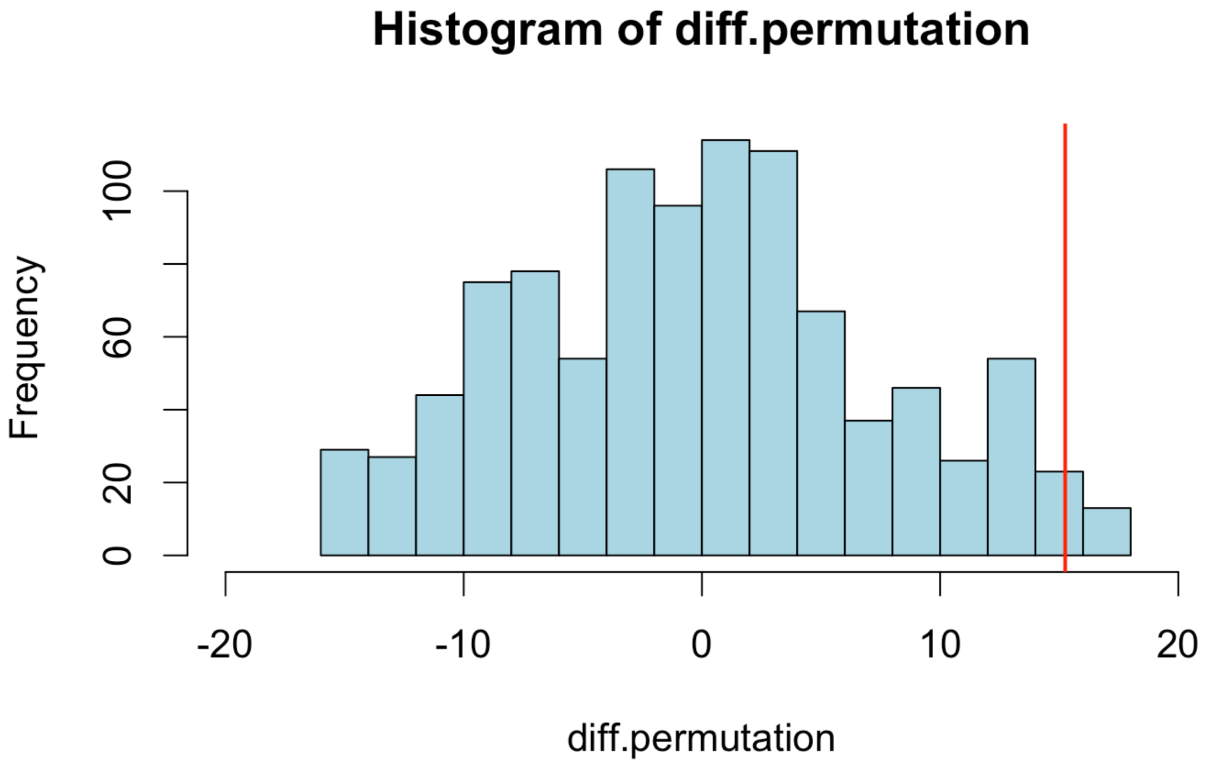
```
diff.permutation <- NULL
set.seed(2017)
for(i in 1:1000){
  index <- sample(1:9, 4) # choose 4 from 1:9
  trt <- combined[index] # assign the 4 students to the trt group
  crl <- combined[-index] # assign the rest 5 students to the crl group
  diff.permutation[i] <- mean(trt) - mean(crl) # calculate the differnece in mean
}
# number of permutated differences that are greater then the actual difference
sum(diff.permutation > diff)
```

```
## [1] 13
```

```
# permutation test p-value: 13/1000 = 0.013
```

# Permutation test

```r
hist(diff.permutation, col="lightblue", breaks=20, xlim=c(-20, 20))
abline(v = diff, col="red", lwd=2)
```



# Permutation test

```r
library(perm)
permTS(Treatment, Control, alternative = "greater")
```

```
##
##  Exact Permutation Test (network algorithm)
##
## data:  Treatment and Control
## p-value = 0.01587
## alternative hypothesis: true mean Treatment - mean Control is greater than 0
## sample estimates:
## mean Treatment - mean Control
##                          15.25
```

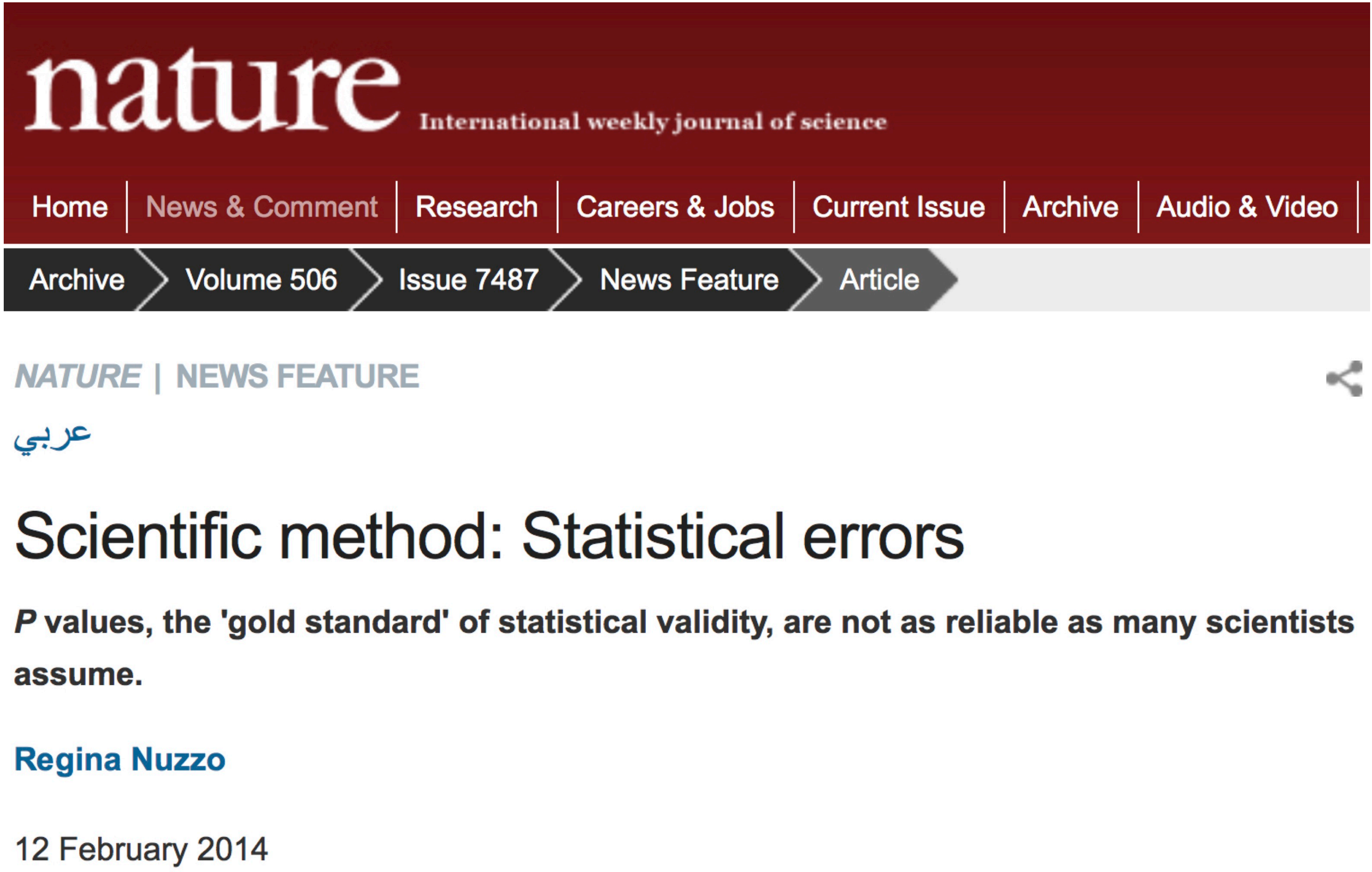
# $P$-value and significance

## Lecture 12

> The probability, assuming $H_0$ is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***P*-value** of the test. **The smaller the $P$-value, the stronger the evidence against $H_0$ provided by the data.**

> Denote $\alpha$ as the **significance level**, which is the decisive value to reject the null hypothesis $H_0$. $\alpha$ is usually 0.05, sometimes 0.01 or 0.1.
>
> If the $P$-value is as small or smaller than $\alpha$, we say that the data are **statistically significant** at level $\alpha$.

# *P*-value



*NATURE* | NEWS FEATURE

عربي

## Scientific method: Statistical errors

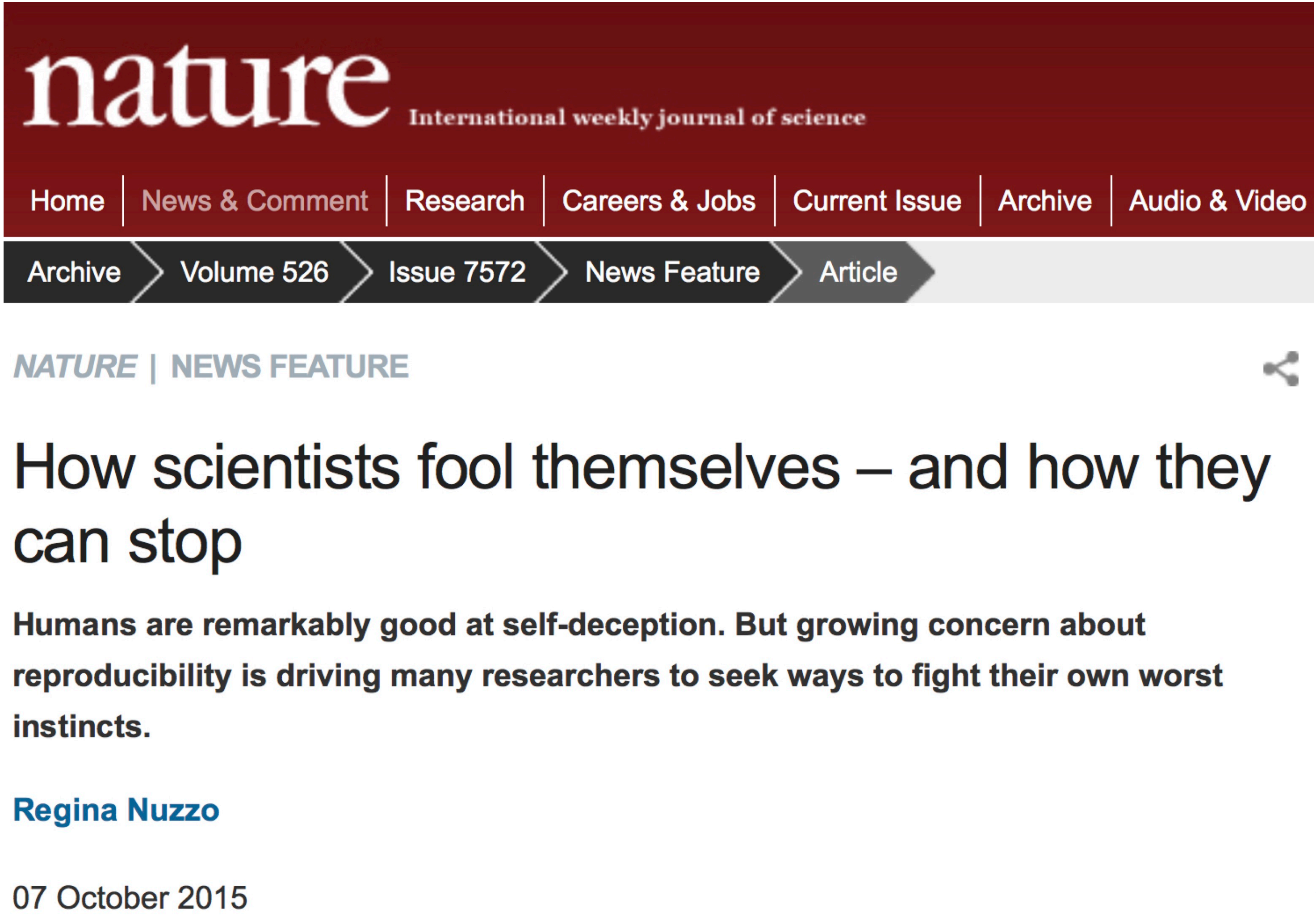
***P* values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.**

**Regina Nuzzo**

12 February 2014

Link.

# *P*-value



*NATURE* | NEWS FEATURE

## How scientists fool themselves – and how they can stop

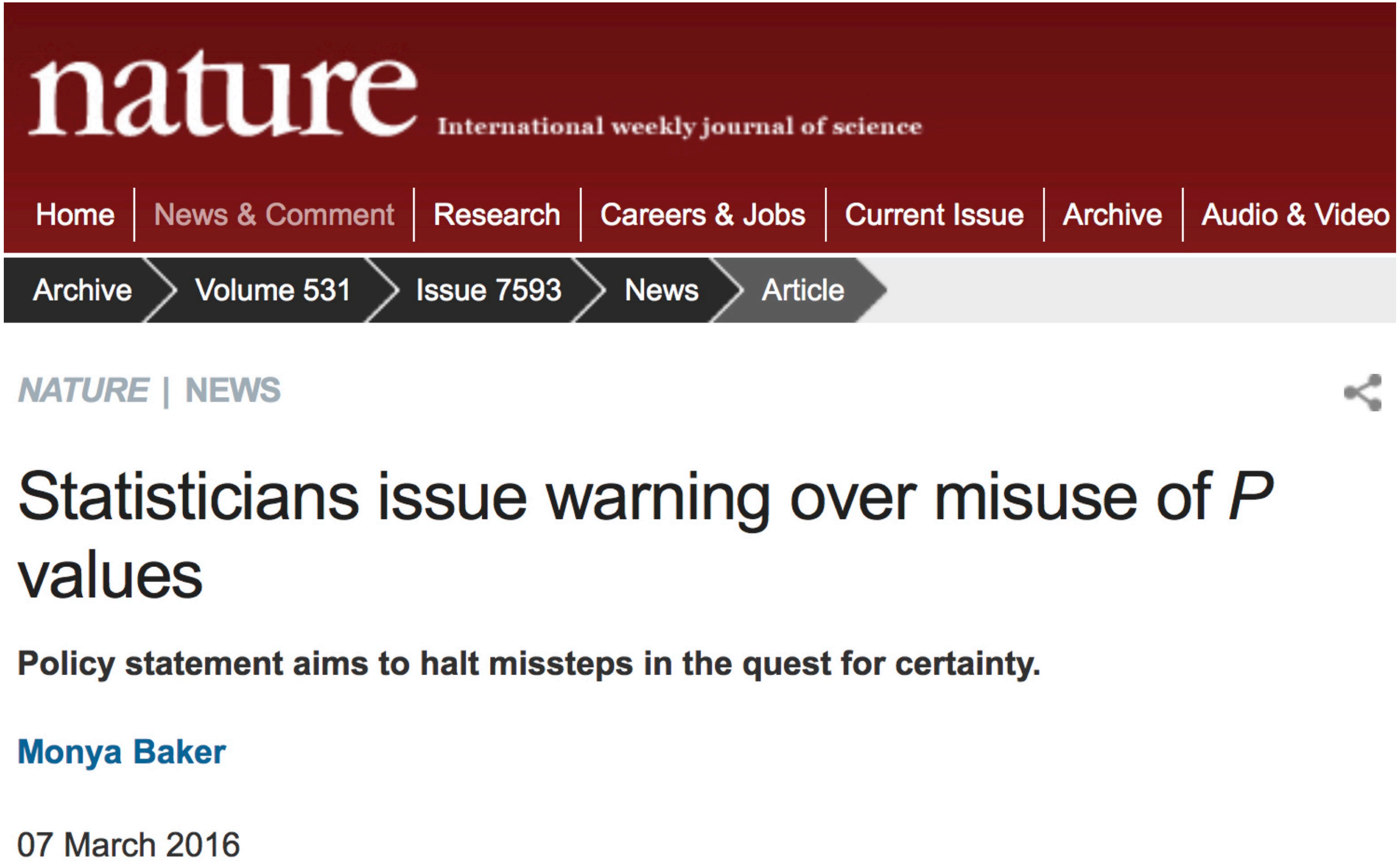
**Humans are remarkably good at self-deception. But growing concern about reproducibility is driving many researchers to seek ways to fight their own worst instincts.**

**Regina Nuzzo**

07 October 2015

Link.

# *P*-value



nature International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video
Archive > Volume 531 > Issue 7593 > News > Article

NATURE | NEWS

Statisticians issue warning over misuse of *P* values

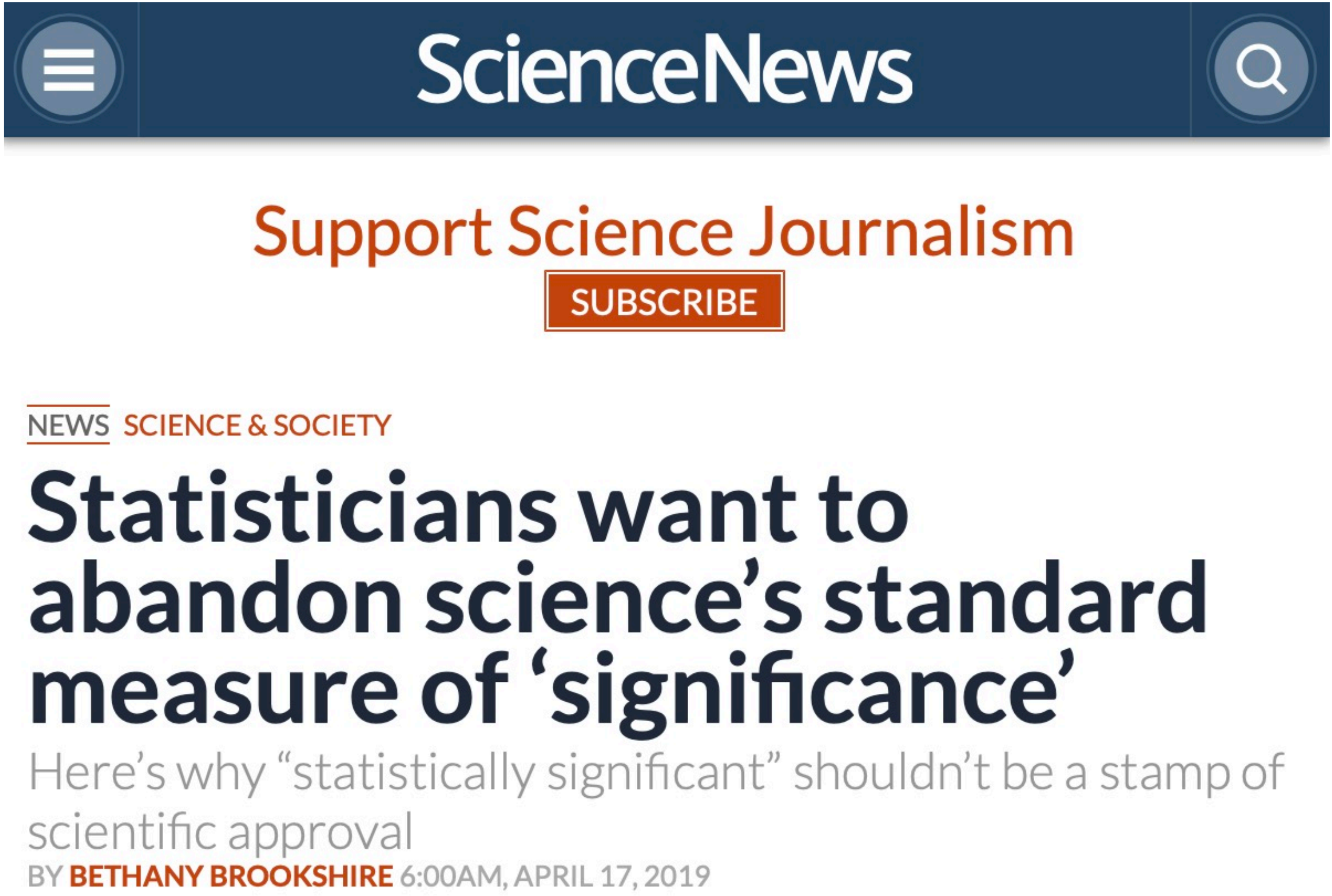
**Policy statement aims to halt missteps in the quest for certainty.**

Monya Baker

07 March 2016

Link.

# *P*-value



NEWS SCIENCE & SOCIETY

## Statisticians want to abandon science's standard measure of 'significance'

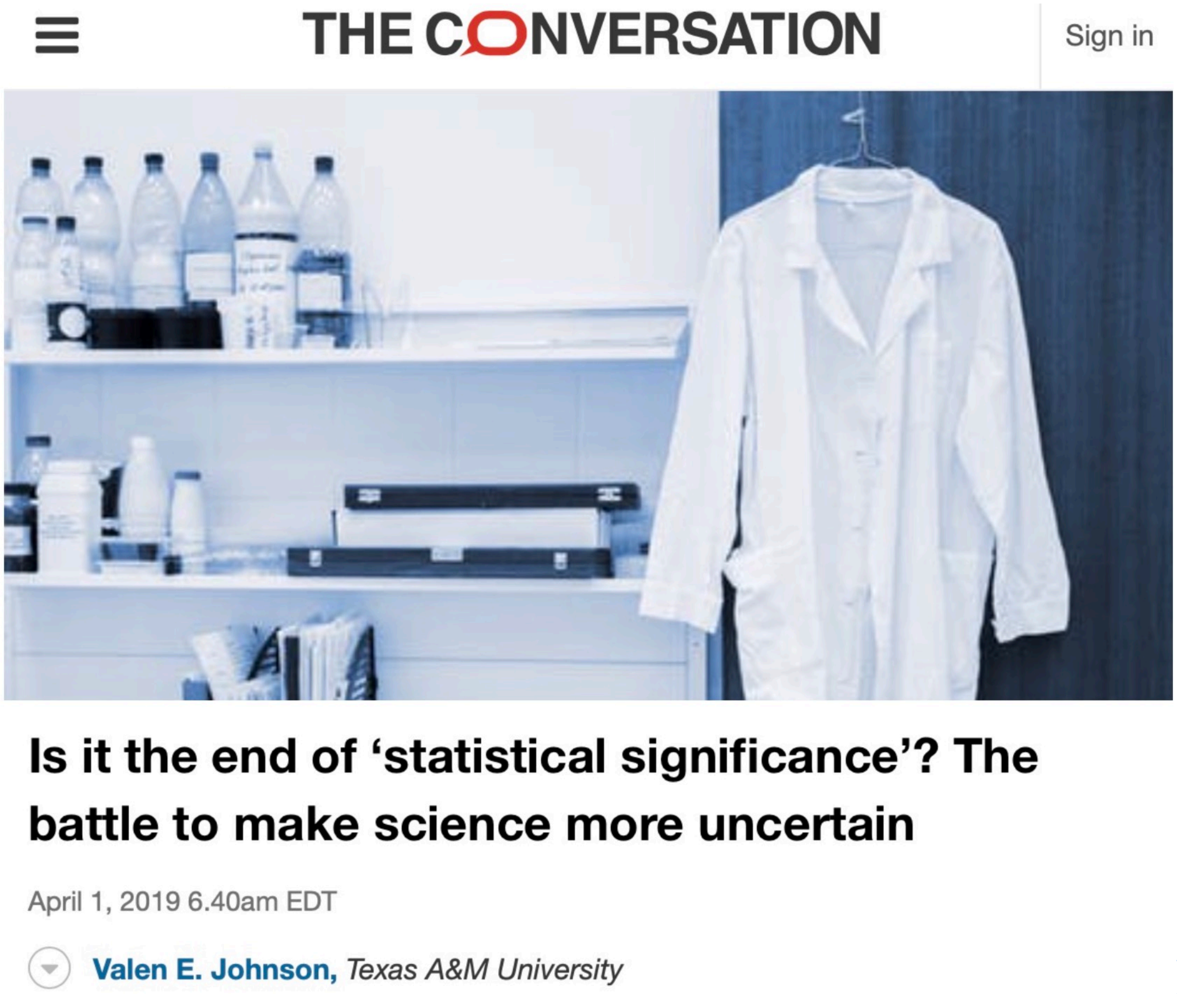
Here's why "statistically significant" shouldn't be a stamp of scientific approval

BY BETHANY BROOKSHIRE 6:00AM, APRIL 17, 2019

Link.

# *P*-value

# *P*-value



THE CONVERSATION

Sign in

**Is it the end of ‘statistical significance’? The battle to make science more uncertain**

April 1, 2019 6.40am EDT

**Valen E. Johnson,** *Texas A&M University*

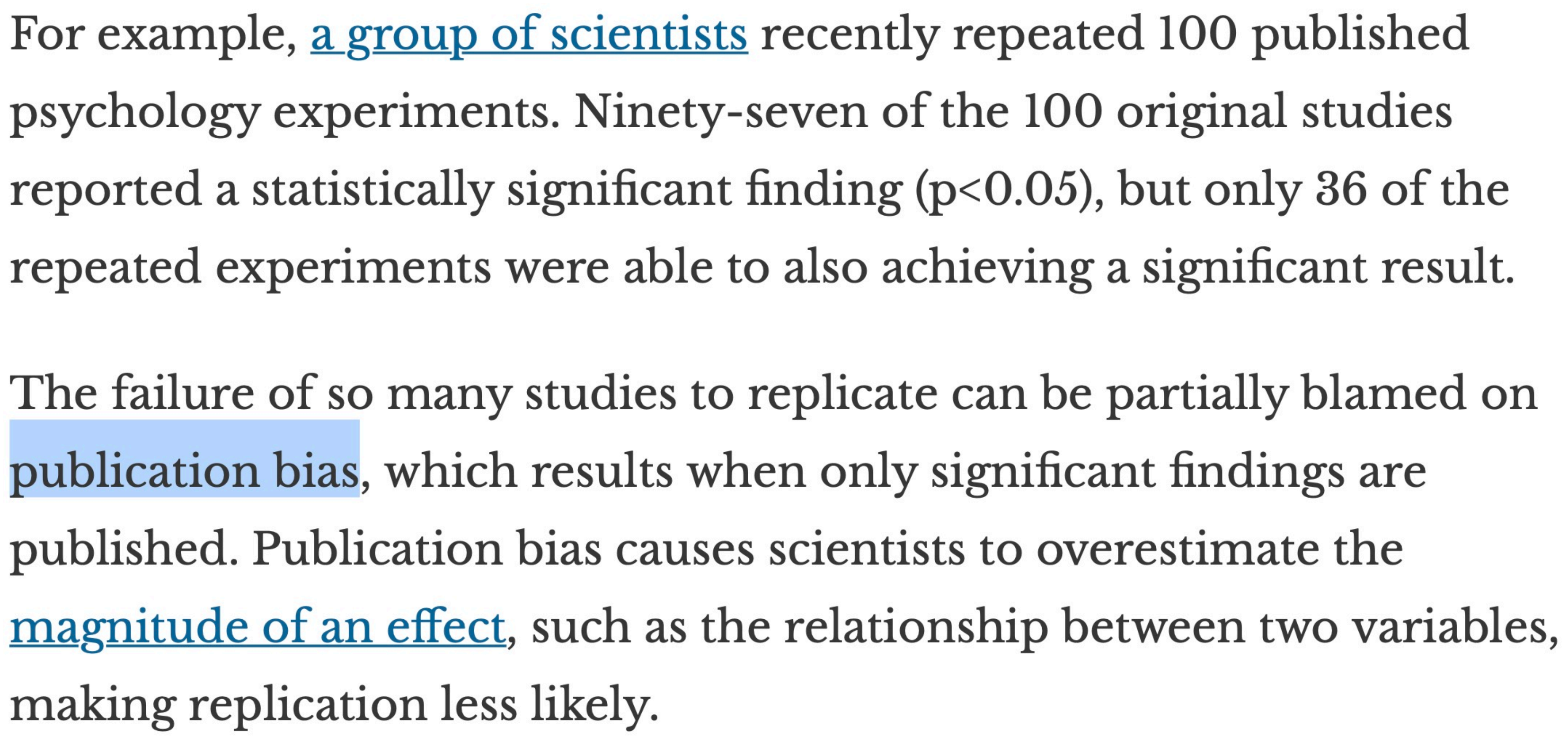
Link.

# $P$-value

"The irony is that when UK statistician Ronald Fisher introduced the $P$ value in the 1920s, he did not mean it to be a definitive test. He intended it simply as an informal way to judge whether evidence was significant in the old-fashioned sense: worthy of a second look."

# $P$-value

"The wake-up call is that so many of our published findings are not true."



For example, a group of scientists recently repeated 100 published psychology experiments. Ninety-seven of the 100 original studies reported a statistically significant finding (p<0.05), but only 36 of the repeated experiments were able to also achieving a significant result.

The failure of so many studies to replicate can be partially blamed on publication bias, which results when only significant findings are published. Publication bias causes scientists to overestimate the magnitude of an effect, such as the relationship between two variables, making replication less likely.

# *P*-value

"….statistical significance has been used to draw a bright line between experimental success and failure. Achieving an experimental result with statistical significance often determines if a scientist's paper gets published or if further research gets funded."

# *P*-value

"There is good reason to want to scrap statistical significance. But with so much research now built around the concept, it's unclear how — or with what other measures - the scientific community could replace it."

"While these are fine qualities, I believe that scientists must not let them obscure the precision and rigor that science demands... If scientists further weaken the already very weak threshold of 0.05, then that would inevitably make scientific findings more difficult to interpret and less likely to be trusted."

# $P$-value

"To avoid the trap of thinking about results as significant or not significant, researchers should always report effect sizes and confidence intervals. These convey what a $P$ value does not: the magnitude and relative importance of an effect."

# *P*-value

"the ASA [American Association of Statistics] advises researchers to avoid drawing scientific conclusions or making policy decisions based on P values alone. Researchers should describe not only the data analyses that produced statistically significant results, but all statistical tests and choices made in calculations. Otherwise, results may seem falsely robust."

"People want something that they can't really get.
They want certainty."
– Andrew Gelman, Columbia University