# 431 Class 23

thomaselove.github.io/431

2021-11-16

# Today's Agenda

On Contingency Tables (Chapter 24)

- Building a $J \times K$ Table
- $\chi^2$ Tests of Independence
  - Cochran Conditions and Checking Assumptions

Replicable Research and the Crisis in Science

- ASA 2016 Statement on P values (Context, Process, Purpose)
- Is changing the $p$ value cutoff the right strategy?
- Second-generation $p$ values: A next step?
- ASA 2019 Statement on Statistical Inference in the 21st Century

# Today's Setup

```
library(janitor)
library(magrittr)
library(patchwork)
library(vcd)
library(tidyverse)

theme_set(theme_bw())
```

# Working with Larger Cross-Tabulations

# A $2 \times 3$ contingency table

This table displays the count of patients who show *complete*, *partial*, or *no response* after treatment with either **active** medication or a **placebo** in a study of 100 patients. . .

| Group | None | Partial | Complete |
|---|---|---|---|
| Active | 8 | 24 | 20 |
| Placebo | 12 | 26 | 10 |

Is there a statistically detectable association here, at $\alpha = 0.10$?

- $H_0$: Response Distribution is the same, regardless of Treatment.
- $H_A$: There is an association between Treatment and Response.

# The Pearson Chi-Square Test

The Pearson $\chi^2$ test assumes the null hypothesis is true (rows and columns are independent.) That is a model for our data. How does it work? Here's the table, with marginal totals added.

| – | None | Partial | Complete | **TOTAL** |
|---|---|---|---|---|
| Active | 8 | 24 | 20 | **52** |
| Placebo | 12 | 26 | 10 | **48** |
| **TOTAL** | **20** | **50** | **30** | **100** |

The test needs to estimate the expected frequency in each of the six cells under the assumption of independence. If the rows and columns are in fact independent of each other, then what is the expected number of subjects that will fall in the Active/None cell?

# The Independence Model

The independence model means the overall rate of "Response = None" or "Partial" or "Complete" applies to both "Active" and "Placebo" subjects.

| – | None | Partial | Complete | **TOTAL** |
|---|---|---|---|---|
| Active | – | – | – | **52** |
| Placebo | – | – | – | **48** |
| **TOTAL** | **20** | **50** | **30** | **100** |

If the rows and columns were independent, then:

- 20/100 or 20% of subjects would have the response "None"
  - That means 20% of the 52 Active, and 20% of the 48 Placebo subjects.
- 50% would have a "Partial" response in each exposure group, and
- 30% would have a "Complete" response in each group.

So, can we fill in the expected frequencies under our independence model?

# Observed (*Expected*) Cell Counts

| – | None | Partial | Complete | **TOTAL** |
|---|---|---|---|---|
| Active | 8 (*10.4*) | 24 (*26.0*) | 20 (*15.6*) | **52** |
| Placebo | 12 (*9.6*) | 26 (*24.0*) | 10 (*14.4*) | **48** |
| **TOTAL** | **20** | **50** | **30** | **100** |

**General Formula for Expected Frequencies under Independence**

$$\text{Expected Frequency} = \frac{\text{Row total} \times \text{Column total}}{\text{Grand Total}}$$

This assumes that the independence model holds: the probability of being in a particular column is exactly the same in each row, and vice versa.

# Chi-Square Assumptions

- Expected Frequencies: We assume that the expected frequency, under the null hypothesized model of independence, will be **at least 5** (and ideally at least 10) in each cell. If that is not the case, then the $\chi^2$ test is likely to give unreliable results. The *Cochran conditions* require us to have no cells with zero and at least 80% of the cells in our table with expected counts of 5 or higher. That's what R uses to warn you of trouble.
- Don't meet the standards? Consider collapsing categories.

## Observed (Expected) Cell Counts

| – | None | Partial | Complete | **TOTAL** |
|---|---|---|---|---|
| Active | 8 (**10.4**) | 24 (**26.0**) | 20 (**15.6**) | 52 |
| Placebo | 12 (**9.6**) | 26 (**24.0**) | 10 (**14.4**) | 48 |
| TOTAL | 20 | 50 | 30 | 100 |

# Getting the Table into R

We'll put the table into a matrix in R. Here's one approach. . .

```
T1 <- matrix(c(8, 24, 20, 12, 26, 10),
             ncol=3, nrow=2, byrow=TRUE)
rownames(T1) <- c("Active", "Placebo")
colnames(T1) <- c("None", "Partial", "Complete")
T1
```

```
        None Partial Complete
Active     8      24       20
Placebo   12      26       10
```

# Chi-Square Test Results in R

- $H_0$: Response Distribution is the same, regardless of Treatment.
  - Rows and Columns of the table are *independent*
- $H_A$: There is an association between Treatment and Response.
  - Rows and Columns of the table are *associated*.

```
chisq.test(T1)
```

```
    Pearson's Chi-squared test

data:  T1
X-squared = 4.0598, df = 2, p-value = 0.1313
```

What is the conclusion?

# Does Sample Size Affect The $\chi^2$ Test?

- T1 results were: $\chi^2 = 4.0598$, df $= 2$, $p = 0.1313$
- What if we had the same pattern, but twice as much data?

```
T1_doubled <- T1*2
T1_doubled
```

```
        None Partial Complete
Active    16      48       40
Placebo   24      52       20
```

```
chisq.test(T1_doubled)
```

```
    Pearson's Chi-squared test

data:  T1_doubled
X-squared = 8.1197, df = 2, p-value = 0.01725
```

# Can we run Fisher's exact test instead?

Yes, but . . . if the Pearson assumptions don't hold, then the Fisher's test is not generally an improvement.

```
fisher.test(T1)
```

```
    Fisher's Exact Test for Count Data

data:  T1
p-value = 0.1358
alternative hypothesis: two.sided
```

- It's also really meant more for square tables, with the same number of rows as columns, and relatively modest sample sizes.

# OK. Back to dm1000

```
dm1000 <- read_rds("data/dm_1000.Rds") %>%
    select(subject, tobacco, insurance) %>%
    filter(complete.cases(.))

head(dm1000)
```

```
# A tibble: 6 x 3
  subject tobacco insurance
  <chr>   <fct>   <fct>
1 M-0001  Current Medicaid
2 M-0002  Never   Commercial
3 M-0003  Former  Medicare
4 M-0004  Never   Medicaid
5 M-0005  Never   Medicare
6 M-0006  Current Medicaid
```

# Arrange the Factors in a Useful Order

```
dm1000 <- dm1000 %>%
    mutate(tobacco =
                fct_relevel(tobacco, "Current", "Former"),
           insurance =
                fct_relevel(insurance, "Medicare",
                            "Commercial", "Medicaid"))

dm1000 %>% tabyl(tobacco, insurance) %>%
    adorn_totals(where = c("row", "col"))
```

| tobacco | Medicare | Commercial | Medicaid | Uninsured | Total |
|---|---|---|---|---|---|
| Current | 99 | 44 | 118 | 13 | 274 |
| Former | 183 | 70 | 103 | 11 | 367 |
| Never | 140 | 80 | 105 | 18 | 343 |
| Total | 422 | 194 | 326 | 42 | 984 |

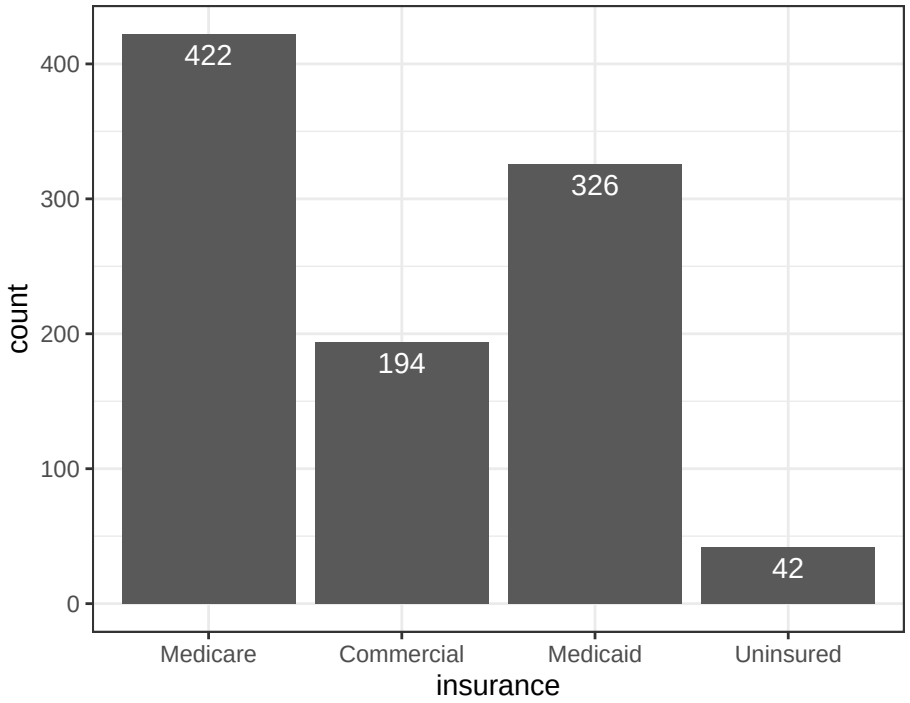
# What am I plotting here?

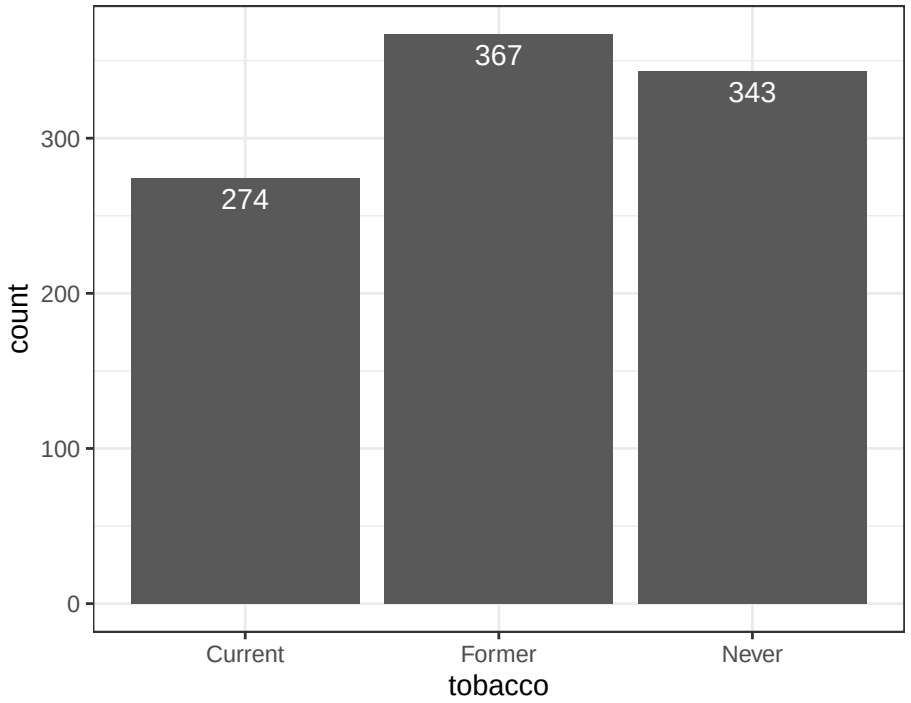
```
p1 <- ggplot(dm1000, aes(x = insurance)) + geom_bar() +
    theme_bw() +
    geom_text(aes(label = ..count..), stat = "count",
              vjust = 1.5, col = "white")

p2 <- ggplot(dm1000, aes(x = tobacco)) + geom_bar() +
    theme_bw() +
    geom_text(aes(label = ..count..), stat = "count",
              vjust = 1.5, col = "white")

p1 + p2
```

# dm1000: Two Categorical Variables of interest



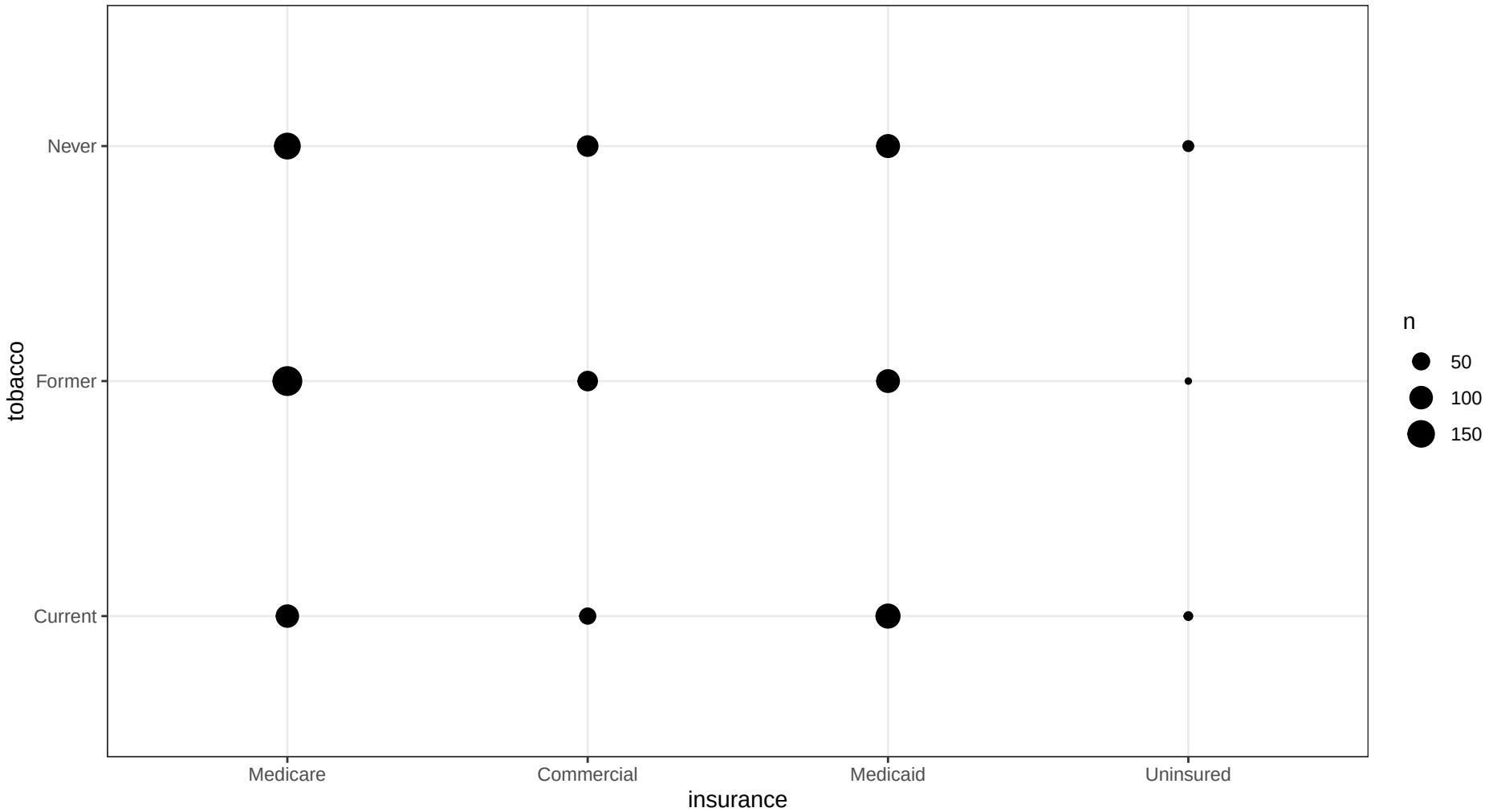
# A 4 × 3 table with the dm1000 data

```
dm1000 %>%
    tabyl(insurance, tobacco) %>%
    adorn_totals(where = c("row", "col"))
```

| insurance | Current | Former | Never | Total |
|---|---|---|---|---|
| Medicare | 99 | 183 | 140 | 422 |
| Commercial | 44 | 70 | 80 | 194 |
| Medicaid | 118 | 103 | 105 | 326 |
| Uninsured | 13 | 11 | 18 | 42 |
| Total | 274 | 367 | 343 | 984 |

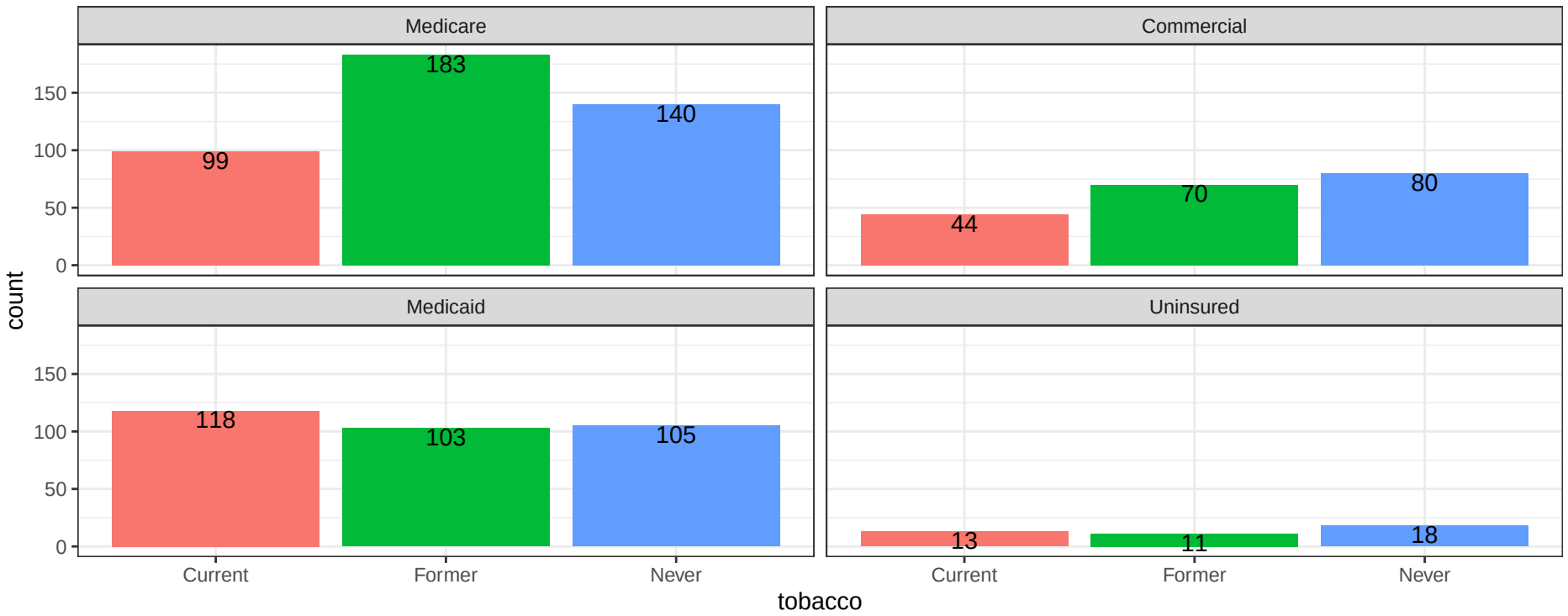
# Plotting a Cross-Tabulation?

```
ggplot(dm1000, aes(x = insurance, y = tobacco)) +
    geom_count() + theme_bw()
```



# Tobacco Bar Chart faceted by Insurance

```
ggplot(dm1000, aes(x = tobacco, fill = tobacco)) +
    geom_bar() + theme_bw() + facet_wrap(~ insurance) +
    guides(fill = "none") +
    geom_text(aes(label = ..count..), stat = "count",
              vjust = 1, col = "black")
```



# Tobacco Status and Insurance in dm1000

- $H_0$: Insurance type and Tobacco status are independent
- $H_A$: Insurance type and Tobacco status have a detectable association

Pearson $\chi^2$ results?

```
dm1000 %>% tabyl(insurance, tobacco) %>% chisq.test()
```

```
    Pearson's Chi-squared test

data:  .
X-squared = 25.592, df = 6, p-value = 0.0002651
```

Can we check our expected frequencies?

# Checking Expected Frequencies

```
res <- dm1000 %>% tabyl(insurance, tobacco) %>% chisq.test()

res$observed
```

```
  insurance Current Former Never
   Medicare      99    183   140
 Commercial      44     70    80
   Medicaid     118    103   105
  Uninsured      13     11    18
```

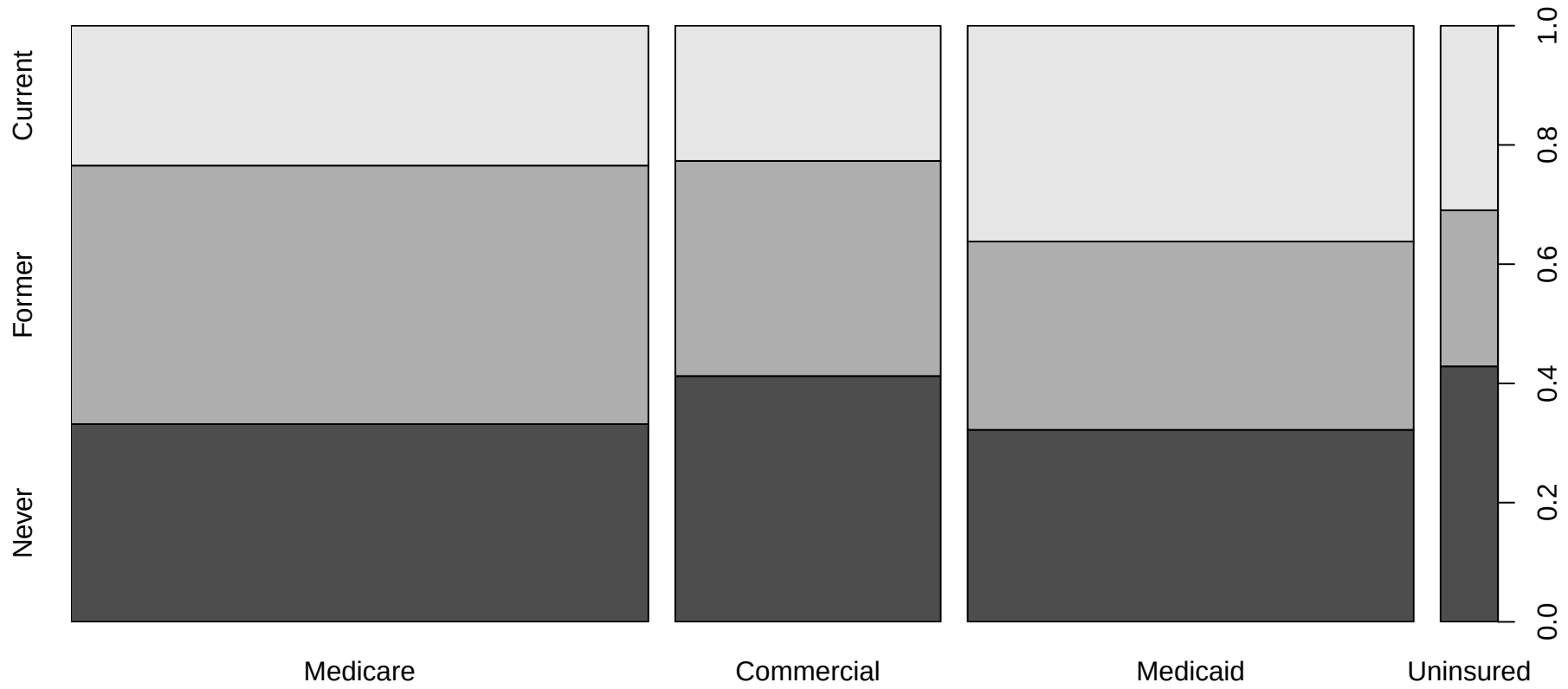
```
res$expected
```

```
  insurance   Current    Former     Never
   Medicare 117.50813 157.39228 147.09959
 Commercial  54.02033  72.35569  67.62398
   Medicaid  90.77642 121.58740 113.63618
  Uninsured  11.69512  15.66463  14.64024
```

# Mosaic Plot for Cross-Tabulation

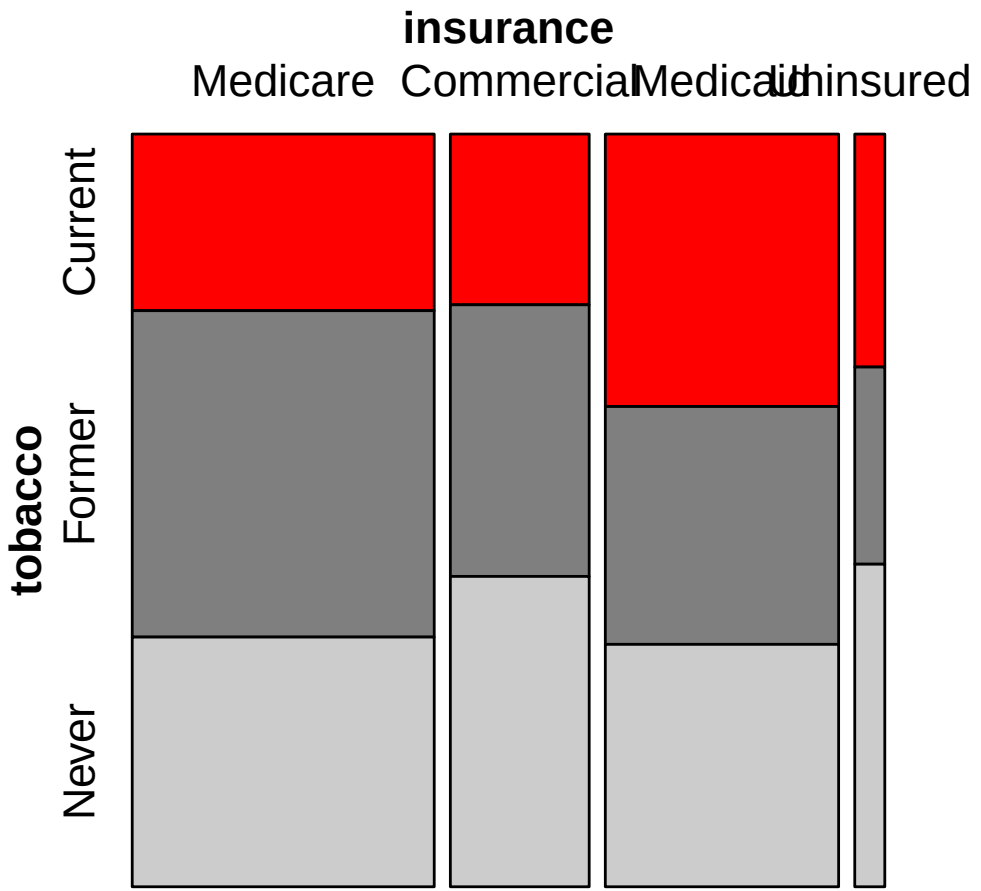
Each rectangle's area is proportional to the number of cases in that cell.

```
dm1000 %$% plot(insurance, tobacco, ylab = "", xlab = "")
```



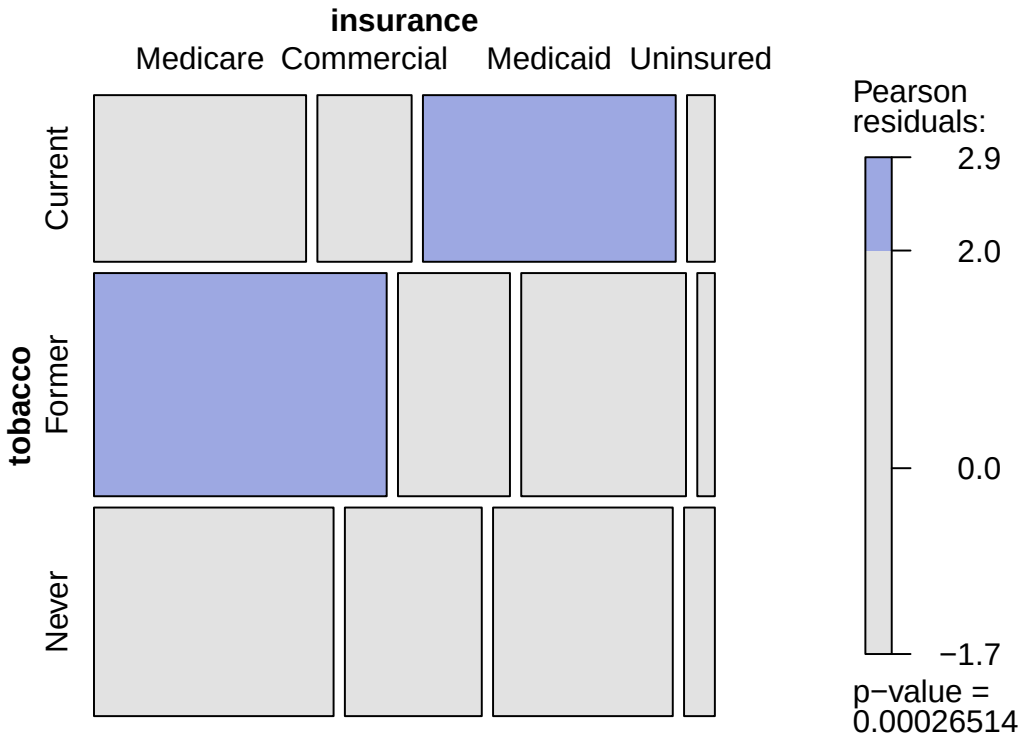
# Mosaic Plot from the vcd package (highlighting)

```
mosaic(~ tobacco + insurance, data = dm1000,
       highlighting = "tobacco",
       highlighting_fill = c("red", "gray50", "gray80"))
```



# Mosaic Plot from the vcd package (with $\chi^2$ shading)

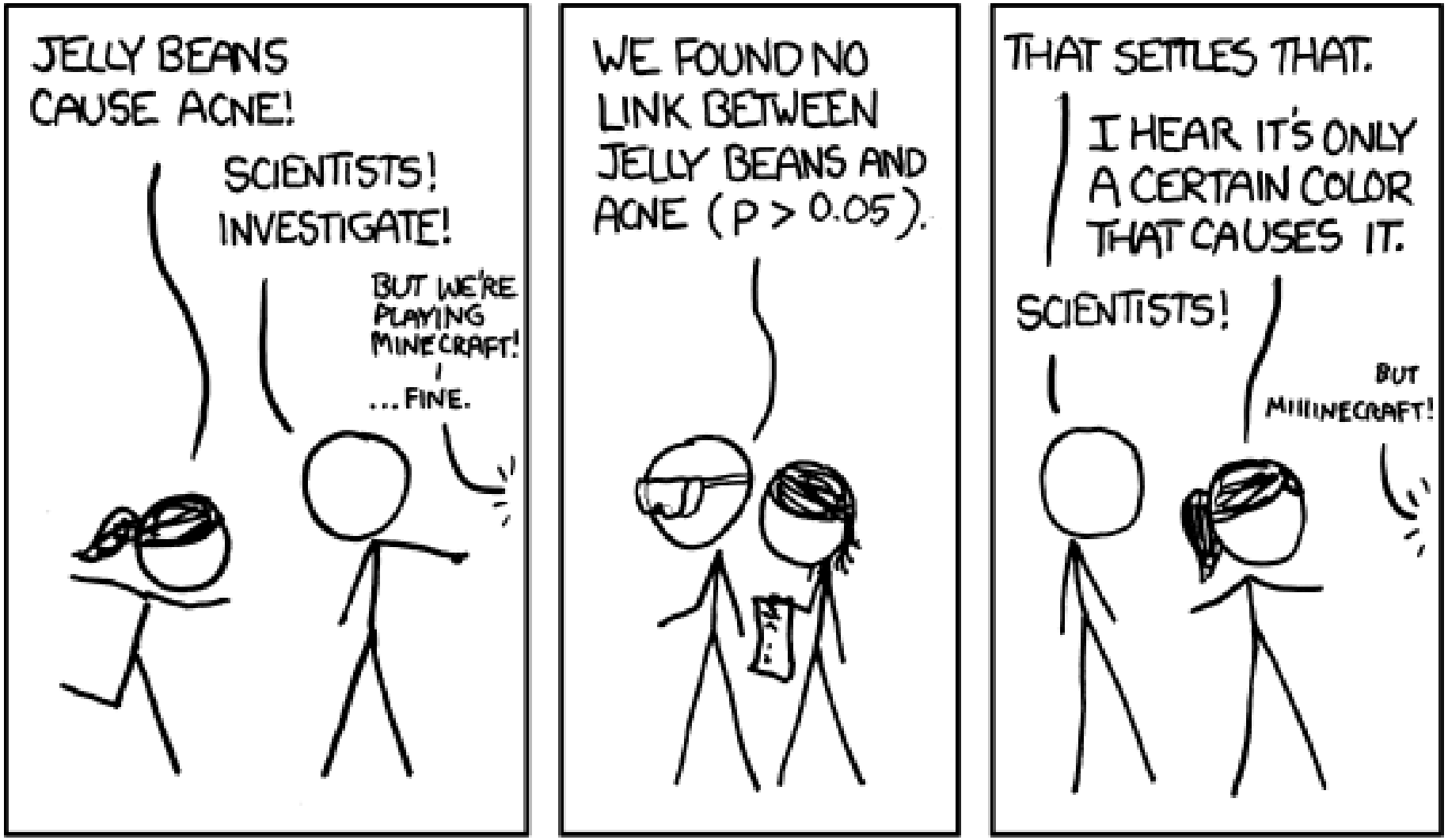
```
mosaic(~ tobacco + insurance, data = dm1000, shade = TRUE)
```

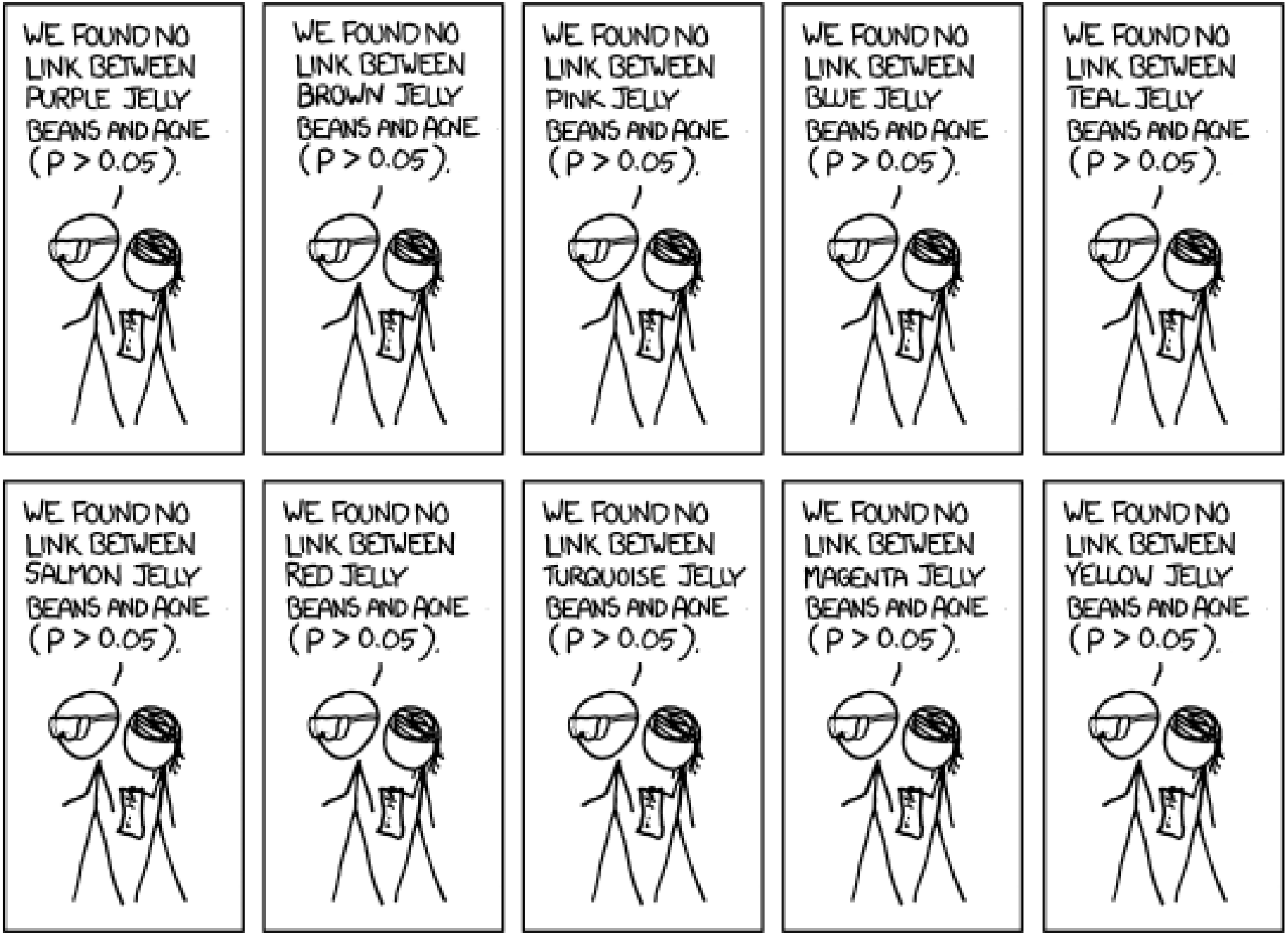


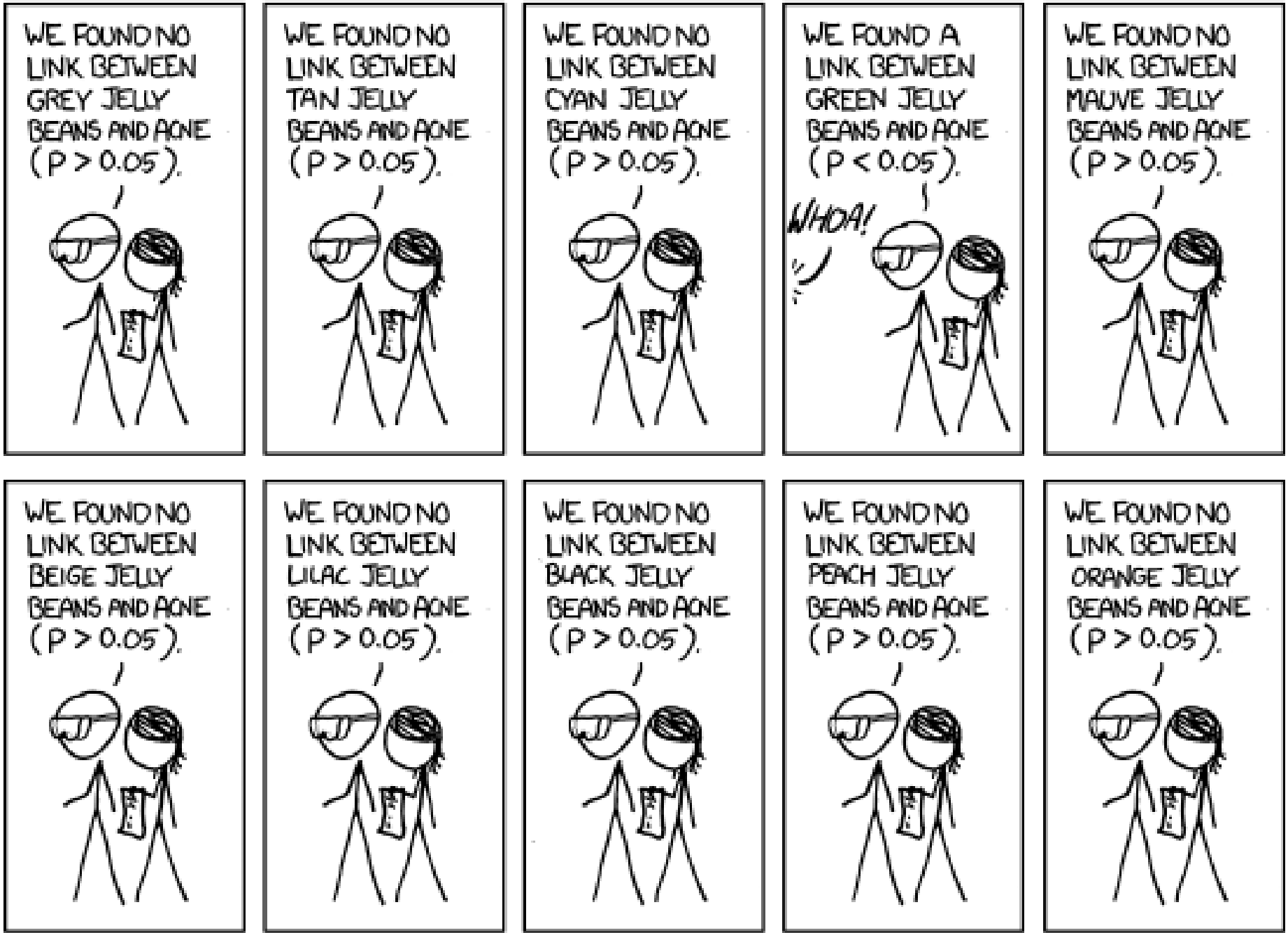
# P values: What's the problem?

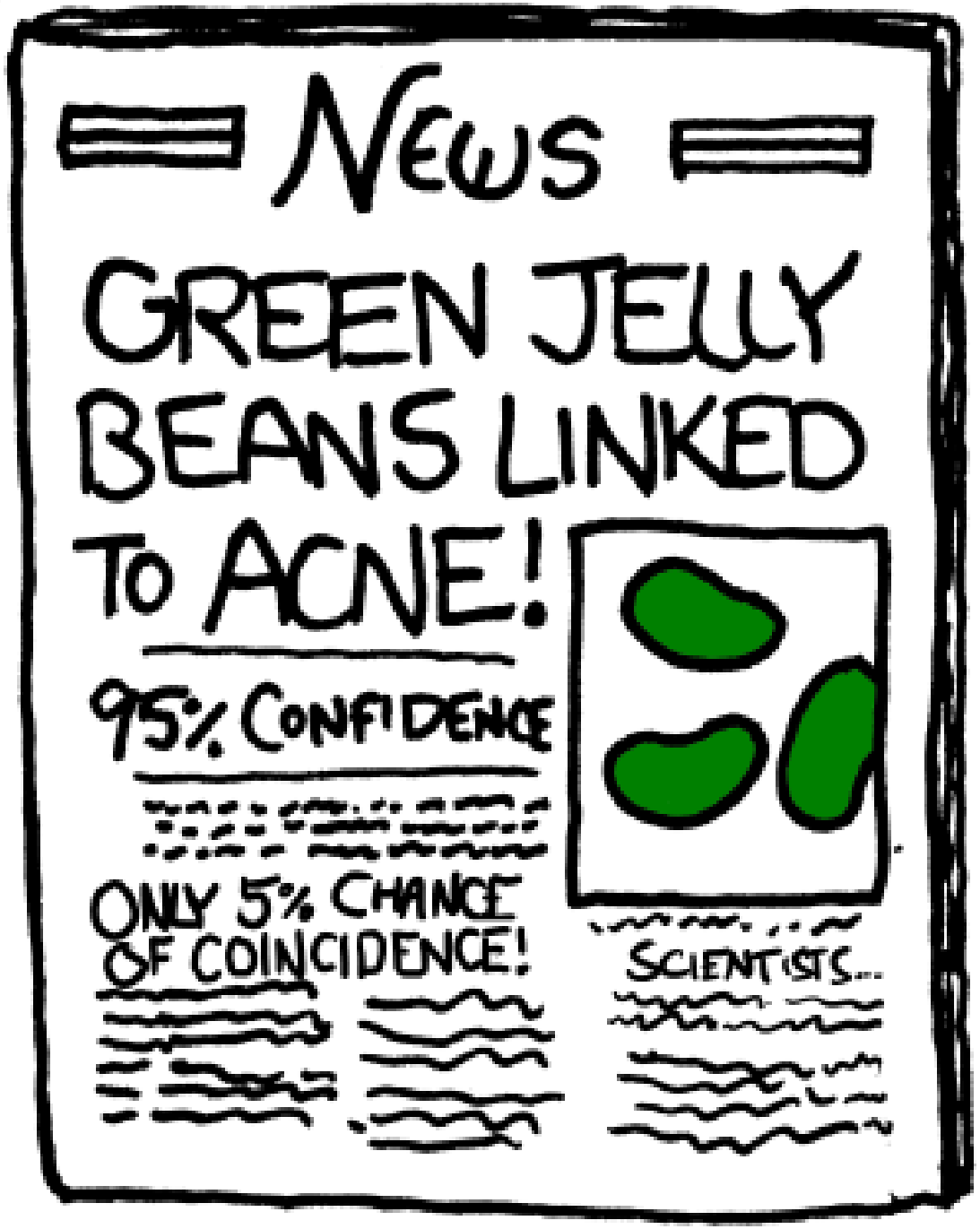
# Replicable Research and the Crisis in Science

- ASA 2016 Statement on P values (Context, Process, Purpose)
- Is changing the $p$ value cutoff the right strategy?
- Second-generation $p$ values: A next step?
- ASA 2019 Statement on Statistical Inference in the 21st Century


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...

# Roger Peng's description of a successful data analysis

> *A data analysis is successful if the audience to which it is presented accepts the results.*

- "What is a Successful Data Analysis?" *simplystatistics.org* (2018-04-17).

So what makes a data analysis more believable / more acceptable?

# The American Statistical Association

2016

- Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, *The American Statistician*, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

2019

- Ronald L. Wasserstein, Allen L. Schirm & Nicole A. Lazar (2019) Moving to a World Beyond "p $< 0.05$", *The American Statistician*, 73:sup1, 1-19, DOI: 10.1080/00031305.2019.1583913.

# Statistical Inference in the 21st Century

*. . . a world learning to venture beyond "$p < 0.05$"*

*This is a world where researchers are free to treat "$p = 0.051$" and "$p = 0.049$" as not being categorically different, where authors no longer find themselves constrained to selectively publish their results based on a single magic number.*

*In this world, where studies with "$p < 0.05$" and studies with "$p > 0.05$" are not automatically in conflict, researchers will see their results more easily replicated – and, even when not, they will better understand why.*

*The 2016 ASA Statement on P-Values and Statistical Significance started moving us toward this world. As of the date of publication of this special issue, the statement has been viewed over 294,000 times and cited over 1700 times-an average of about 11 citations per week since its release. Now we must go further.*

# The American Statistical Association Statement on P values and Statistical Significance

The ASA Statement (2016) was mostly about what **not** to do.

The 2019 effort represents an attempt to explain what to do.

> *Some of you exploring this special issue of The American Statistician might be wondering if it's a scolding from pedantic statisticians lecturing you about what not to dowith p-values, without offering any real ideas of what to do about the very hard problem of separating signal from noise in data and making decisions under uncertainty. Fear not. In this issue, thanks to 43 innovative and thought-provoking papers from forward-looking statisticians, help is on the way.*

# "Don't" is not enough.

*If you're just arriving to the debate, here's a sampling of what not to do.*

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant" (i.e., the *p* value passed some arbitrary threshold such as $p < 0.05$).
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

# One More Don't. . .

The *ASA Statement on P-Values and Statistical Significance* stopped just short of recommending that declarations of "statistical significance" be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term "statistically significant" entirely. Nor should variants such as "significantly different," "$p < 0.05$," and "nonsignificant" survive, whether expressed in words, by asterisks in a table, or in some other way.

Regardless of whether it was ever useful, a declaration of "statistical significance" has today become meaningless. Made

*A label of statistical significance adds nothing to what is already conveyed by the value of p; in fact, this dichotomization of p-values makes matters worse.*

# Problems with *P* Values

1. *P* values are inherently unstable
2. The *p* value, or statistical significance, does not measure the size of an effect or the importance of a result
3. Scientific conclusions should not be based only on whether a *p* value passes a specific threshold
4. Proper inference requires full reporting and transparency
5. By itself, a *p* value does not provide a good measure of evidence regarding a model or hypothesis

*Link*

# Solutions to the *P* Value Problems

1. Estimation of the Size of the Effect
2. Precision of the Estimate (Confidence Intervals)
3. Inference About the Target Population
4. Determination of Whether the Results Are Compatible With a Clinically Meaningful Effect
5. Replication and Steady Accumulation of Knowledge

*Link*

# Importance of Meta-Analytic Thinking

*In JAMA Otolaryngology: Head & Neck Surgery, we look to publish original investigations where the investigators planned the study with sufficient sample size to have adequate power to detect a clinically meaningful effect and report the results with effect sizes and CIs. Authors should interpret the effect sizes in relation to previous research and use CIs to help determine whether the results are compatible with clinically meaningful effects. And finally, we acknowledge that no single study can define truth and that the advancement of medical knowledge and patient care depends on the steady accumulation of reliable clinical information.*

*Link*

# The Value of a *p*-Valueless Paper

Jason T. Connor (2004) *American J of Gastroenterology* 99(9): 1638-40.

Abstract: As is common in current bio-medical research, about 85% of original contributions in *The American Journal of Gastroenterology* in 2004 have reported *p*-values. However, none are reported in this issue's article by Abraham et al. who, instead, rely exclusively on effect size estimates and associated confidence intervals to summarize their findings. **Authors using confidence intervals communicate much more information in a clear and efficient manner than those using *p*-values. This strategy also prevents readers from drawing erroneous conclusions caused by common misunderstandings about *p*-values**. I outline how standard, two-sided confidence intervals can be used to measure whether two treatments differ or test whether they are clinically equivalent.

*Link*

# Editorial from JAMA Cardiology 2016-10-12

Editor's Note

## Do Not Over (*P*) Value Your Research Article

Laine E. Thomas, PhD; Michael J. Pencina, PhD

**P value** is by far the most prevalent statistic in the medical literature but also one attracting considerable controversy. Recently, the American Statistical Association[1] released a policy statement on *P* values, noting that misunderstanding and misuse of *P* values is an important contributing factor to the common problem of scientific conclusions that fail to be reproducible. Furthermore, reliance on *P* values may distract from the good scientific principles that are needed for high-quality research. Mark et al[2] delve deeper into the history and interpretation of the *P* value in this issue of *JAMA Cardiology*. Herein, we take the opportunity to state a few principles to help guide authors in the use and reporting of *P* values in the journal.

Related article

When the limitations surrounding *P* values are emphasized, a common question is, "What should we do instead?" Ron Wasserstein of the American Statistical Association explained: "In the post p<0.05 era, scientific argumentation is not based on whether a p-value is small enough or not. Attention is paid to effect sizes and confidence intervals. Evidence is thought of as being continuous rather than some sort of dichotomy.... Instead, journals [should evaluate] papers based on clear and detailed description of the study design, execution, and analysis, having conclusions that are based on valid

We suggest that researchers submitting manuscripts to *JAMA Cardiology* should also consider the following:

1. Data that are descriptive of the sample (ie, indicating imbalances between observed groups but not making inference to a population) should not be associated with *P* values. Appropriate language, in this case, would describe numerical differences and sample summary statistics and focus on differences of clinical importance.
2. In addition to summary statistics and confidence intervals, standardized differences (rather than *P* values) are a preferred way to exhibit imbalances between groups.
3. *P* values are most meaningful in the context of clear, a priori hypotheses that support the main conclusions of a manuscript.
4. Reporting stand-alone *P* values is discouraged, and preference should be given to presentation and interpretation of effect sizes and their uncertainty (confidence intervals) in the scientific context and in light of other evidence. Crossing a threshold (eg, $P < .05$) by itself constitutes only weak evidence.
5. Researchers should define and interpret effect measures that are clinically relevant. For example, clinical importance is often difficult to establish on the odds ratio scale but is clearer on the risk ratio or absolute risk difference scale.

In summary, following Mark et al,[2] we encourage research-

# Mark, Lee, Harrell JAMA Cardiol 2016-10-12

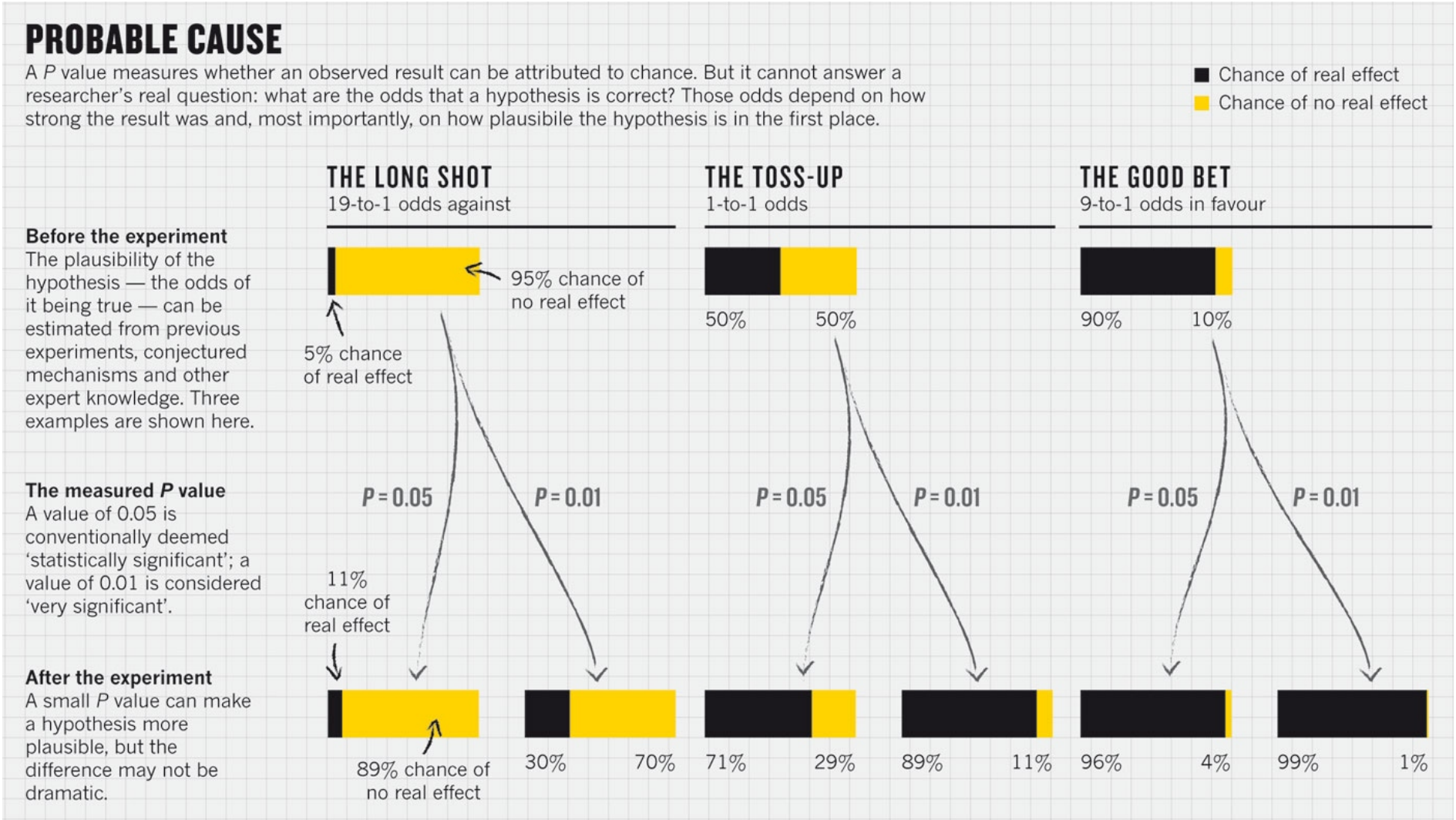
## Abstract

*P* values and hypothesis testing methods are frequently misused in clinical research. Much of this misuse appears to be owing to the widespread, mistaken belief that they provide simple, reliable, and objective triage tools for separating the true and important from the untrue or unimportant. The primary focus in interpreting therapeutic clinical research data should be on the treatment ("oomph") effect, a metaphorical force that moves patients given an effective treatment to a different clinical state relative to their control counterparts. This effect is assessed using 2 complementary types of statistical measures calculated from the data, namely, effect magnitude or size and precision of the effect size. In a randomized trial, effect size is often summarized using constructs, such as odds ratios, hazard ratios, relative risks, or adverse event rate differences. How large a treatment effect has to be to be consequential is a matter for clinical judgment. The precision of the effect size (conceptually related to the amount of spread in the data) is usually addressed with confidence intervals. *P* values (significance tests) were first proposed as an informal heuristic to help assess how "unexpected" the observed effect size was if the true state of nature was no effect or no difference. Hypothesis testing was a modification of the significance test approach that envisioned controlling the false-positive rate of study results over many (hypothetical) repetitions of the experiment of interest. Both can be helpful but, by themselves, provide only a tunnel vision perspective on study results that ignores the clinical effects the study was conducted to measure.

*Link*

# Dividing Data Comparisons into Categories based on p values

# Regina Nuzzo in Nature on Statistical Errors



# Gelman on $p$ values, 1

> *The common practice of dividing data comparisons into categories based on significance levels is terrible, but it happens all the time. . . . so it's worth examining the prevalence of this error.*

Consider, for example, this division:

- "really significant" for $p < .01$,
- "significant" for $p < .05$,
- "marginally significant" for $p < .1$, and
- "not at all significant" otherwise.

Now consider some typical $p$-values in these ranges: say, $p = .005$, $p = .03$, $p = .08$, and $p = .2$.

Translate these two-sided $p$-values back into z-scores. . .

*Gelman* 2016-10-15

# Gelman on *p* values, 2

| Description | really sig. | sig. | marginally sig. | not at all sig. |
|---|---|---|---|---|
| *p* value | 0.005 | 0.03 | 0.08 | 0.20 |
| Z score | 2.8 | 2.2 | 1.8 | 1.3 |

The seemingly yawning gap in p-values comparing the not at all significant *p*-value of .2 to the really significant *p*-value of .005, is only a z score of 1.5.

If you had two independent experiments with z-scores of 2.8 and 1.3 and with equal standard errors and you wanted to compare them, you'd get a difference of 1.5 with a standard error of 1.4, which is completely consistent with noise.

# Gelman on *p* values, 3

From a **statistical** point of view, the trouble with using the p-value as a data summary is that the p-value can only be interpreted in the context of the null hypothesis of zero effect, and (much of the time), nobody's interested in the null hypothesis.

Indeed, once you see comparisons between large, marginal, and small effects, the null hypothesis is irrelevant, as you want to be comparing effect sizes.

From a **psychological** point of view, the trouble with using the p-value as a data summary is that this is a kind of deterministic thinking, an attempt to convert real uncertainty into firm statements that are just not possible (or, as we would say now, just not replicable).

**The key point**: The difference between statistically significant and NOT statistically significant is not, generally, statistically significant.

# Are P values all that bad?

**Grumpy Old Health Stats Dude**
@healthstatsdude

Following

"If you never use another p-value, you will have improved medicine."

-me, to clinicians

#statstwitter #medtwitter #epitwitter

12:36 AM - 4 Mar 2019

**Grumpy Old Health Stats Dude**
@healthstatsdude

Following

Replying to @healthstatsdude @EugeneDayDSc and 2 others

# My main reason for being overtly/in public anti p-values is this:

*P* values
f overall
analyses
partly
atistical
even if
group,
wer, due
l distri-
fference
specific
ll death

mate of a 5% decrease in 10-year survival with watchful waiting, 750 men might have died prematurely as a result.

A mistake in the operating room can threaten the life of one patient; a mistake in statistical analysis or interpretation can lead to hundreds of early deaths. So it is perhaps odd that, while we allow a doctor to conduct surgery only after years of training, we give SPSS® (SPSS, Chicago, IL) to almost anyone. Moreover, whilst only a surgeon would comment on surgical technique, it seems that anybody, regardless of statistical training,

day); a
that on
risk of t
in many
that the
event is

**Comp**
The aut
no com

7:59 PM - 19 Apr 2019

# Where to Go from Here?

1. Be the change you want to see in the world.
2. Frank Harrell's "A Litany of Problems with p-values" blog post
3. William Briggs' "Everything Wrong with P-values under One Roof" article.

These resources are linked on our Class 23 README.