

Welcome to Week #11!

**Linear Regression → Multiple Linear
Regression & Logistic Regression**

SLR → MLR & LR

MLR: For multiple linear parameters

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n$$

The diagram shows the equation $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n$ in blue. Below the equation, there are four arrows pointing upwards to the coefficients: β_0 , β_1 , β_2 , and β_3 . The arrow from β_0 points to the word "predicted y" in red. The arrow from β_1 points to the word "intercept" in red. The arrows from β_2 , β_3 , and β_n point to the phrase "slopes along different parameters" in red.

predicted y

intercept

"slopes along different parameters"

one response

many explanatory variables

Visualization in 3D for 2 explanatory variables: <http://miabellaai.net/>

SLR to MLR: R^2 to R_{adj}^2

The strength of the fit of a linear model is most commonly evaluated using R^2 .

R^2 is calculated as the square of the correlation coefficient.

It tells us what percent of variability in the response variable is explained by the model.

$n = \#$ of data points

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{Var(e_i)}{Var(y_i)}$$

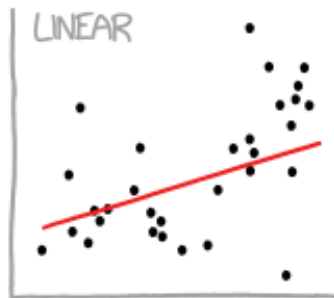
$k = \#$ of parameters in our model

$$R_{adj}^2 = 1 - \frac{Var(e_i)/(n - k - 1)}{Var(y_i)/(n - 1)} = 1 - \frac{Var(e_i)}{Var(y_i)} \times \frac{n - 1}{n - k - 1}$$

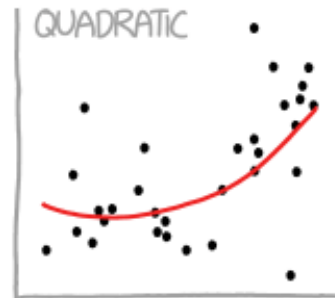
So like R^2 , but takes into account the number of degrees of freedom

More parameters (higher k) means worse R_{adj}^2 - adjusts for the fact that we can fit anything if we have a large enough number of parameters!

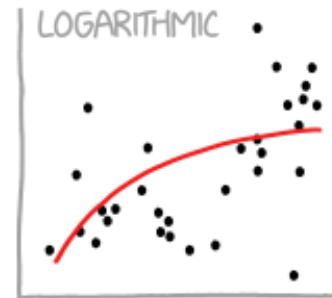
CURVE-FITTING METHODS AND THE MESSAGES THEY SEND



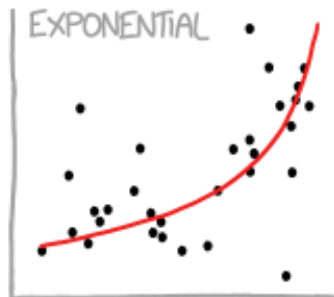
"HEY, I DID A
REGRESSION."



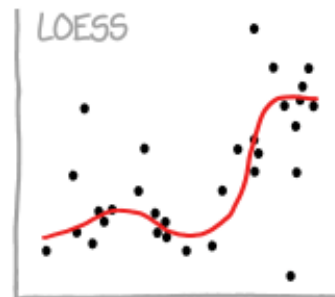
"I WANTED A CURVED
LINE, SO I MADE ONE
WITH MATH."



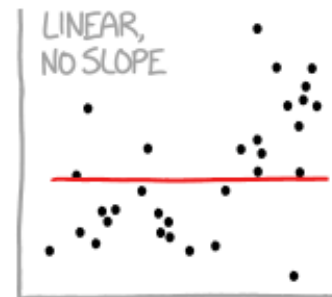
"LOOK, IT'S
TAPERING OFF!"



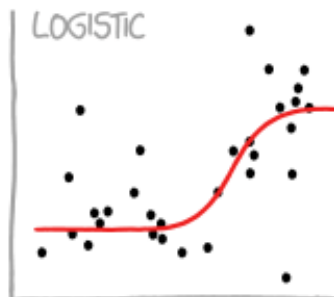
"LOOK, IT'S GROWING
UNCONTROLLABLY!"



"I'M SOPHISTICATED, NOT
LIKE THOSE BUMBLING
POLYNOMIAL PEOPLE."



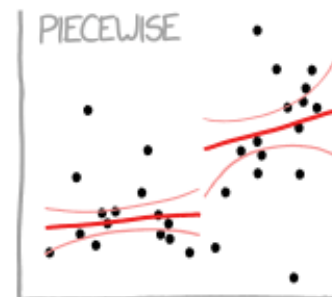
"I'M MAKING A
SCATTER PLOT BUT
I DON'T WANT TO."



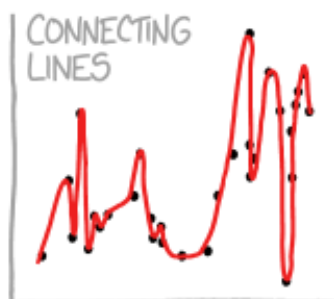
"I NEED TO CONNECT THESE
TWO LINES, BUT MY FIRST IDEA
DIDN'T HAVE ENOUGH MATH."



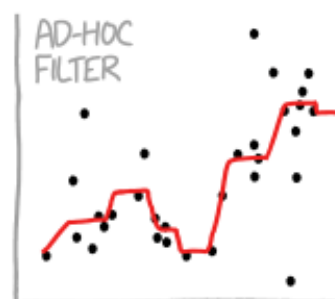
"LISTEN, SCIENCE IS HARD.
BUT I'M A SERIOUS
PERSON DOING MY BEST."



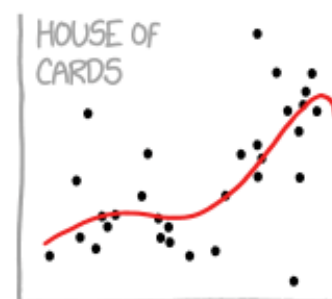
"I HAVE A THEORY,
AND THIS IS THE ONLY
DATA I COULD FIND."



"I CLICKED 'SMOOTH
LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW
TO CLEAN UP THE DATA.
WHAT DO YOU THINK?"



"AS YOU CAN SEE, THIS
MODEL SMOOTHLY FITS
THE— WAIT NO NO DON'T
EXTEND IT AAAAAA!!!"

**And in general: we should
avoid making things overly
complicated**

Enough chit-chat: Lets do an R example!

Intercepts in MLR don't usually make a lot of sense

```
Call:
lm(formula = Snow.Depth ~ Elevation + Min.Temp + Max.Temp, data = snotel2)

Residuals:
    Min       1Q   Median       3Q      Max
-29.508  -7.679  -3.139   9.627  26.394

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.506529   99.616286  -0.105   0.9170
Elevation     0.012332    0.006536   1.887   0.0731 .
Min.Temp     -0.504970    2.042614  -0.247   0.8071
Max.Temp     -0.561892    0.673219  -0.835   0.4133
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.6 on 21 degrees of freedom
Multiple R-squared:  0.6485,    Adjusted R-squared:  0.5983
F-statistic: 12.91 on 3 and 21 DF,  p-value: 5.328e-05
```

“at 0 feet elevation, 0F minimum and maximum daily temperature the snowfall should be -10.5 inches”

Conditions to use MLR

1. Independence of observations of responses
2. Linearity of **all** variables - linear relationship between response variable and each of the explanatory variables
3. Multicollinearity checked for - does not mean we cannot use MLR, but we should be aware of how predictor/explanatory variables are related when quoting our results
4. Constant variance
5. Normality of Residuals
6. No influential points (outliers with strong leverage)

Lets check in R!

Few notes on Collinearity & Model Selection

- Two predictor variables are said to be collinear when they are correlated, and this **collinearity** complicates model estimation.
Remember: Predictors are also called explanatory or independent variables. Ideally, they would be independent of each other.
- We don't like adding predictors that are associated with each other to the model, because often times the addition of such variable brings nothing to the table. Instead, we prefer the simplest best model, i.e. **parsimonious** model.
It is OK to consider multiple reasons to select a model but it is dangerous to “shop” for a model across many possible models – a practice which is sometimes called **data-dredging** and leads to a high chance of spurious results from a single model that is usually reported based on this type of exploration.
- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

vif(our model) =

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R_{X_j|X_{-j}}^2},$$

Few notes on Collinearity & Model Selection

- Two “general” methods for model selection:
 - ★ Forward: choose lowest p-value/highest R_{adj}^2 parameters first
 - ★ Backward: subtract parameters to minimize/maximize p-value/ R_{adj}^2

Example: For a forward model

```
Call:
lm(formula = Snow.Depth ~ Elevation + Min.Temp + Max.Temp, data = snotel2)

Residuals:
    Min       1Q   Median       3Q      Max
-29.508  -7.679  -3.139   9.627  26.394

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.506529   99.616286  -0.105   0.9170
Elevation     0.012332    0.006536   1.887   0.0731 1
Min.Temp     -0.504970    2.042614  -0.247   0.8071 3
Max.Temp     -0.561892    0.673219  -0.835   0.4133 2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.6 on 21 degrees of freedom
Multiple R-squared:  0.6485,    Adjusted R-squared:  0.5983
F-statistic: 12.91 on 3 and 21 DF,  p-value: 5.328e-05
```

Play with this in R!

Some “non-linear” linear models

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_n x_n$$

predicted y

intercept

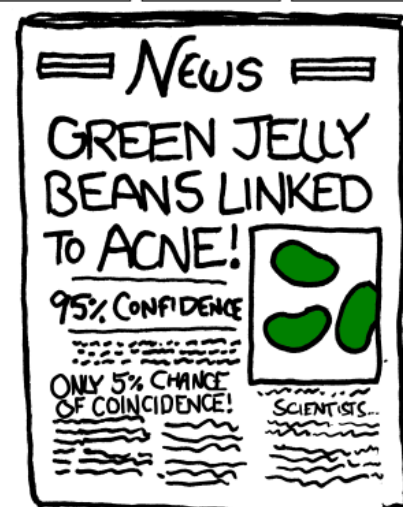
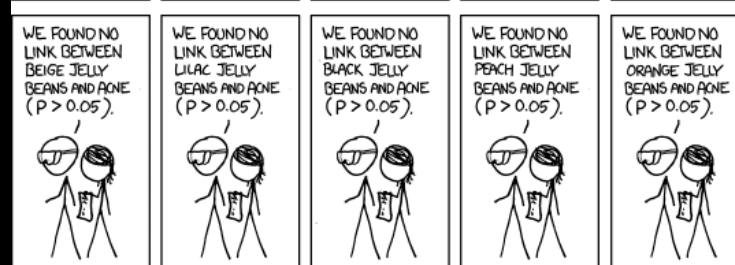
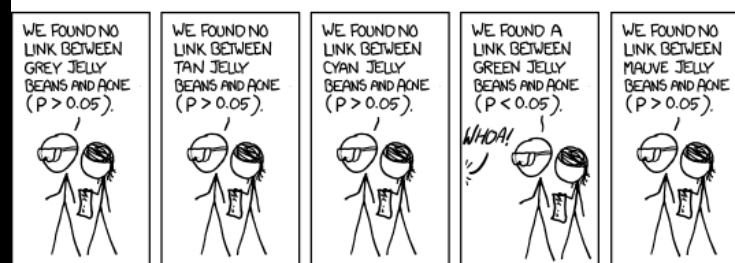
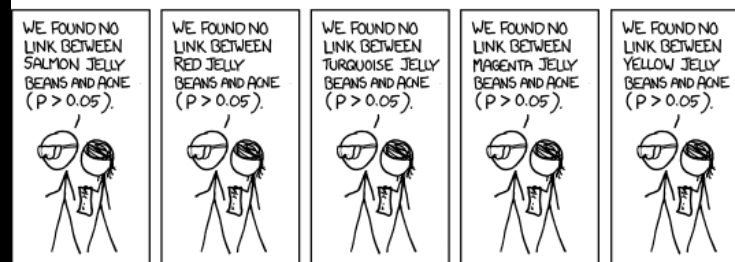
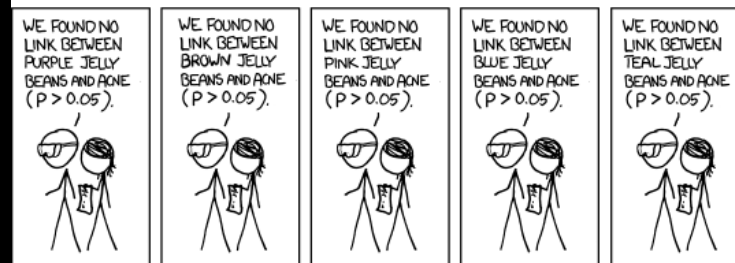
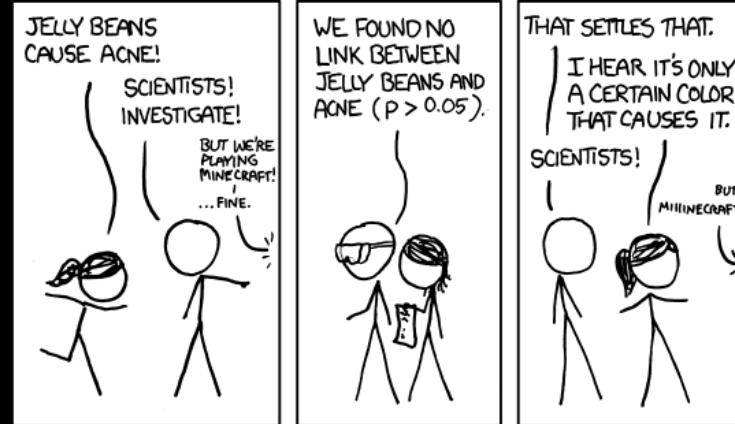
“slopes along different parameters”

what can x_n be here? For example, I can say $x_4 = x_3^2$?

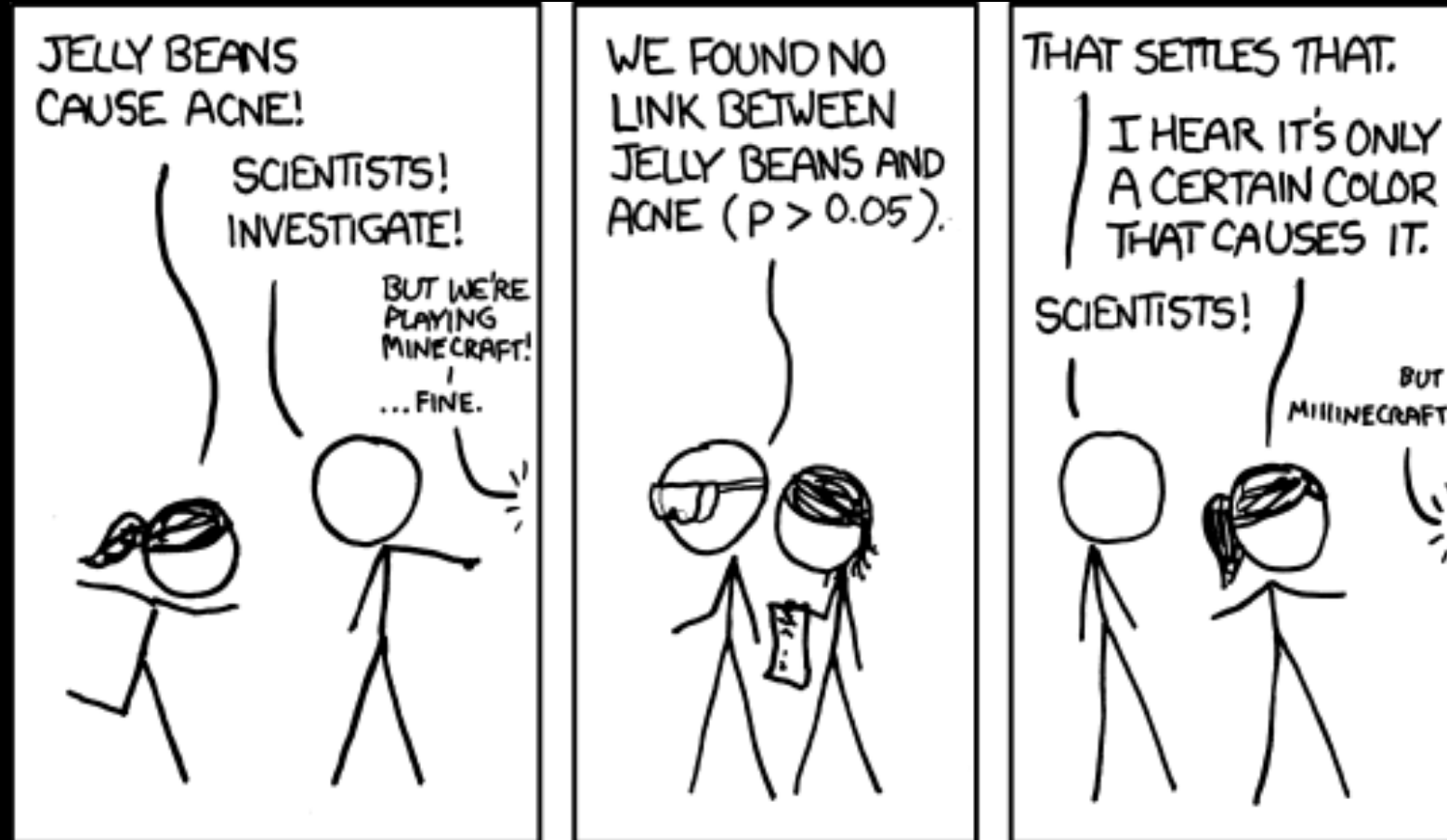
Issues with p-values

Issues with p-values

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



Issues with p-values



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
($P < 0.05$).



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
($P > 0.05$).







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...



Issues with p-values

<https://www.amstat.org//asa/files/pdfs/P-ValueStatement.pdf>



AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • [www.twitter.com/AmstatNews](https://twitter.com/AmstatNews)

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

*Provides Principles to Improve the Conduct and Interpretation of Quantitative
Science*

March 7, 2016

The American Statistical Association (ASA) has released a “Statement on Statistical Significance and *P*-Values” with six principles underlying the proper use and interpretation of the *p*-value [<http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN>]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice “emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.”

“The *p*-value was never intended to be a substitute for scientific reasoning,” said Ron Wasserstein, the ASA’s executive director. “Well-reasoned statistical arguments contain much

Issues with p-values

<https://www.amstat.org//asa/files/pdfs/P-ValueStatement.pdf>

Good statistical practice is an essential component of good scientific practice, the statement observes, and such practice “emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean.”

“The p-value was never intended to be a substitute for scientific reasoning,” said Ron Wasserstein, the ASA’s executive director. “Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a ‘post $p < 0.05$ era.’”

“Over time it appears the p -value has become a gatekeeper for whether work is publishable, at least in some fields,” said Jessica Utts, ASA president. “This apparent editorial bias leads to the ‘file-drawer effect,’ in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as ‘p-hacking’ and ‘data dredging’ that emphasize the search for small p -values over other statistical and scientific reasoning.”

Issues with p-values

<https://www.amstat.org//asa/files/pdfs/P-ValueStatement.pdf>

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. **SCIENCE!***