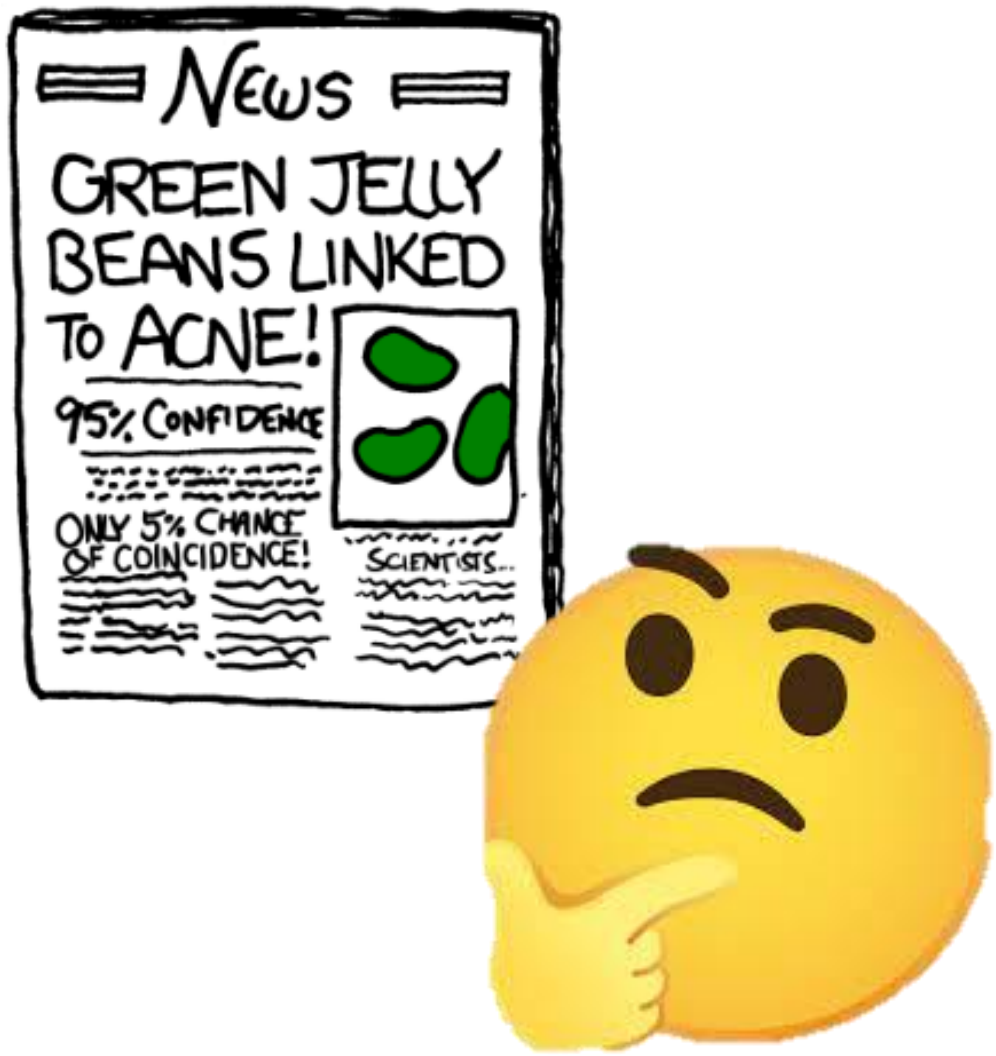


P-values and Alternatives

P&C Model Review Team

10-25-2022

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

P-Value Primer

Dorothy Andrews

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

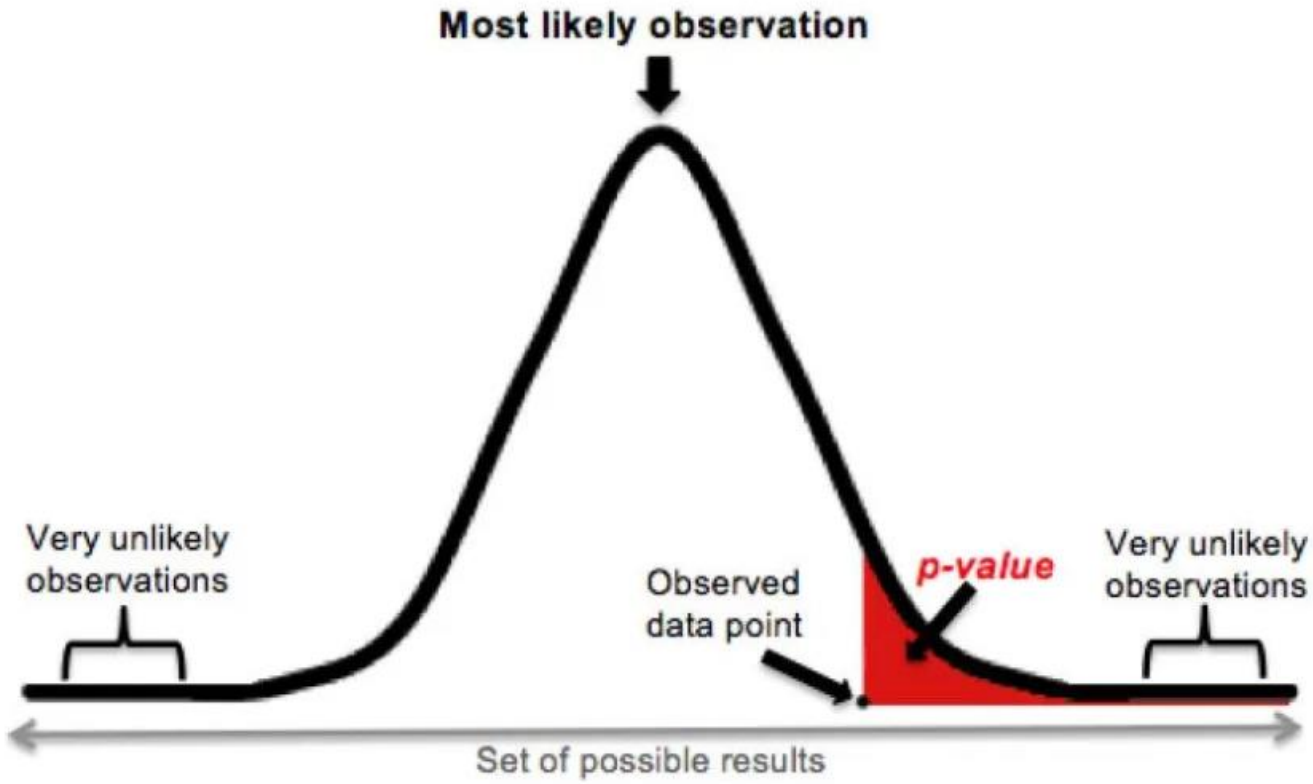
P-Value

For a given statistical model when the null hypothesis is true, the P – value is the probability the model test statistic is equal to or more extreme than the actual observed results.

A p-value is NOT the probability that the null hypothesis is true.

For regression analysis, we test

1.) H_0: $\beta_i = 0$
2.) H_0: σ_i are equal



A *p-value* (shaded red area) is the probability of an observed (or more extreme) result arising by chance

P-Values and Degrees of Freedom

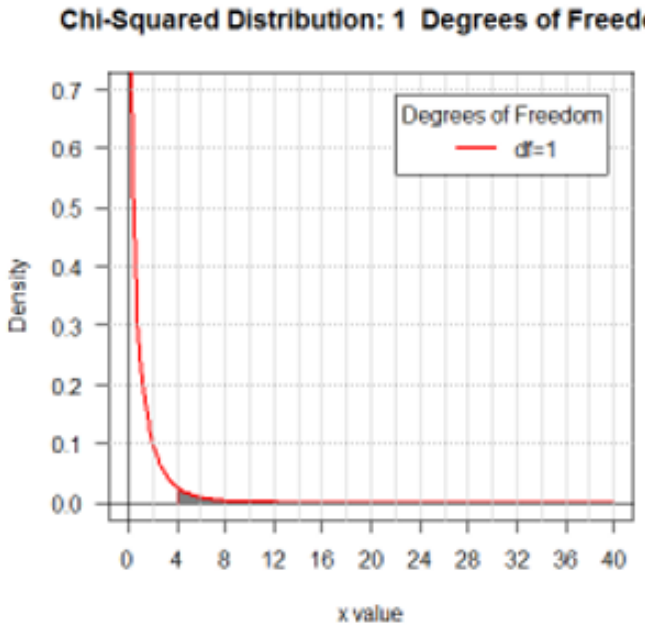
The main distributions used in hypothesis testing are:

1. **Chi-squared:** Testing hypotheses involving count data
2. **Fisher's *F*:** Comparing two variances – ANOVA F test
3. **Student's *t*:** Comparing two parameter estimates – t test

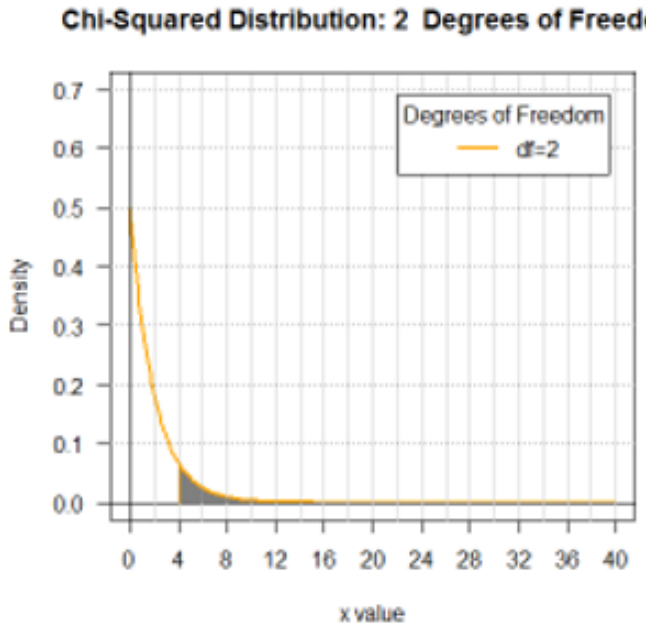
R. A. Fisher (1890–1962)

The F distribution is named after R.A. Fisher, the father of analysis of variance & modern-day statistics, and principal developer of quantitative genetics.

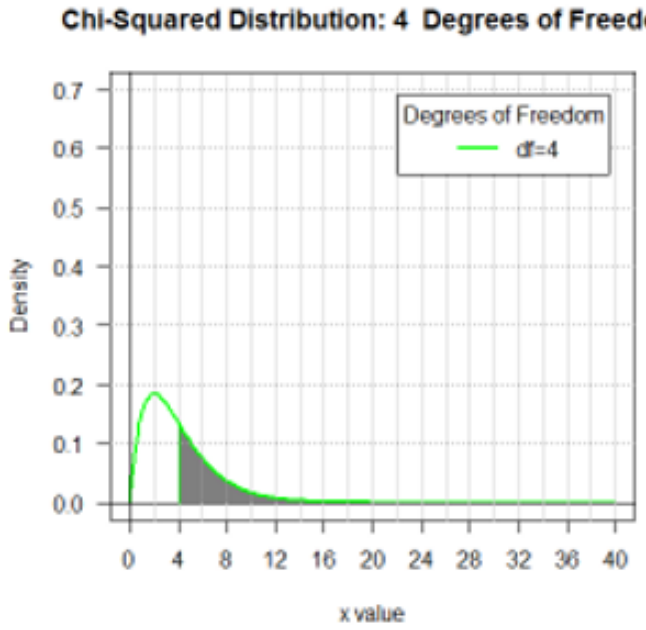
Chi-squared P-Values and Degrees of Freedom



$P(x > 4) \approx 0.04550026$

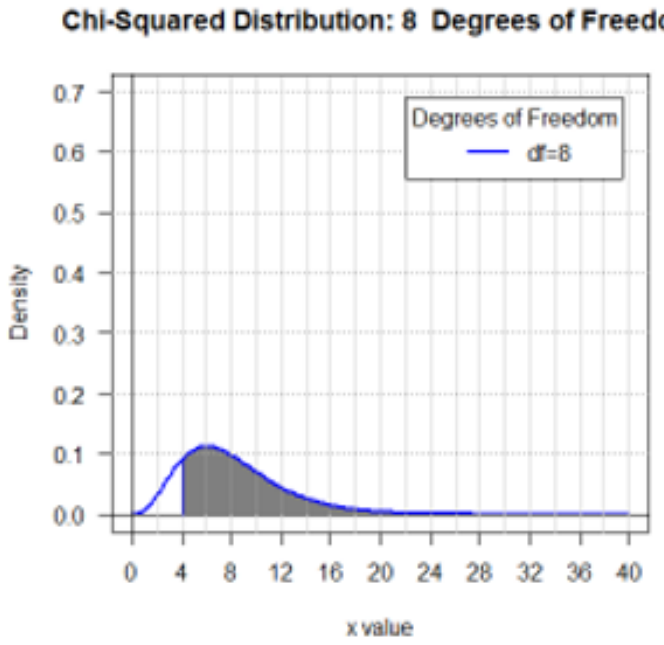


$P(x > 4) \approx 0.1353353$

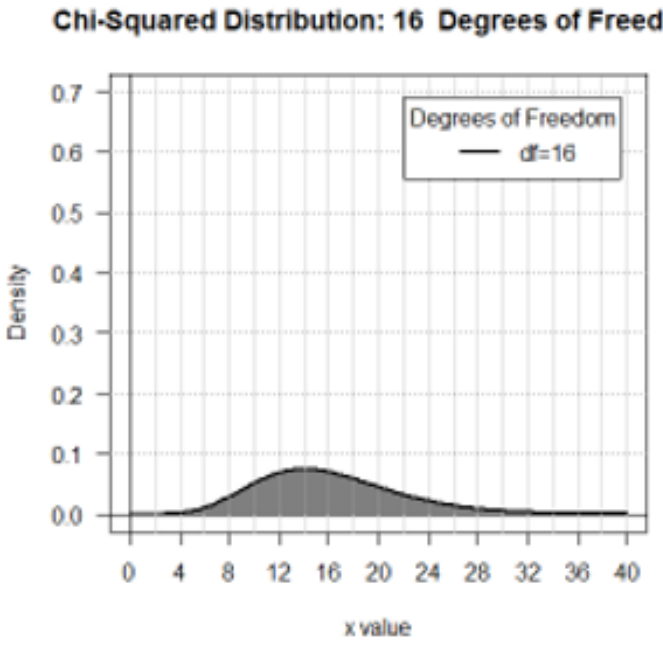


$P(x > 4) \approx 0.4060058$

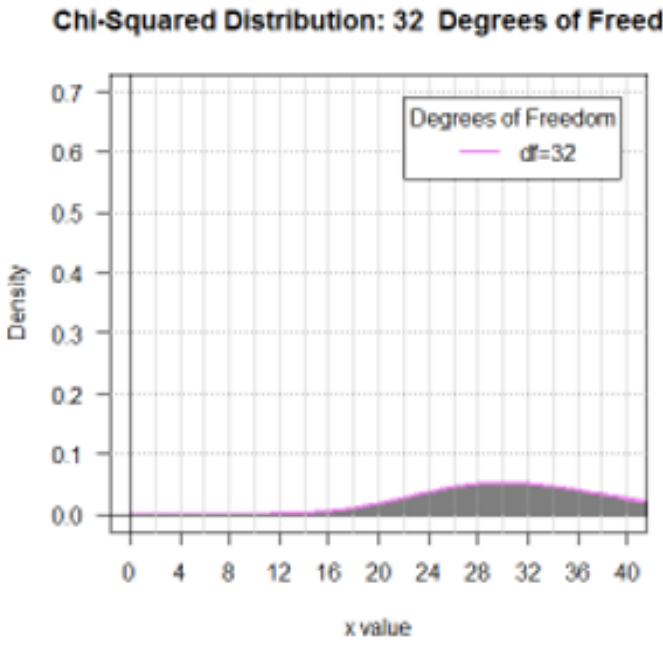
Chi-squared P-Values and Degrees of Freedom



$P(x > 4) \approx 0.8571235$

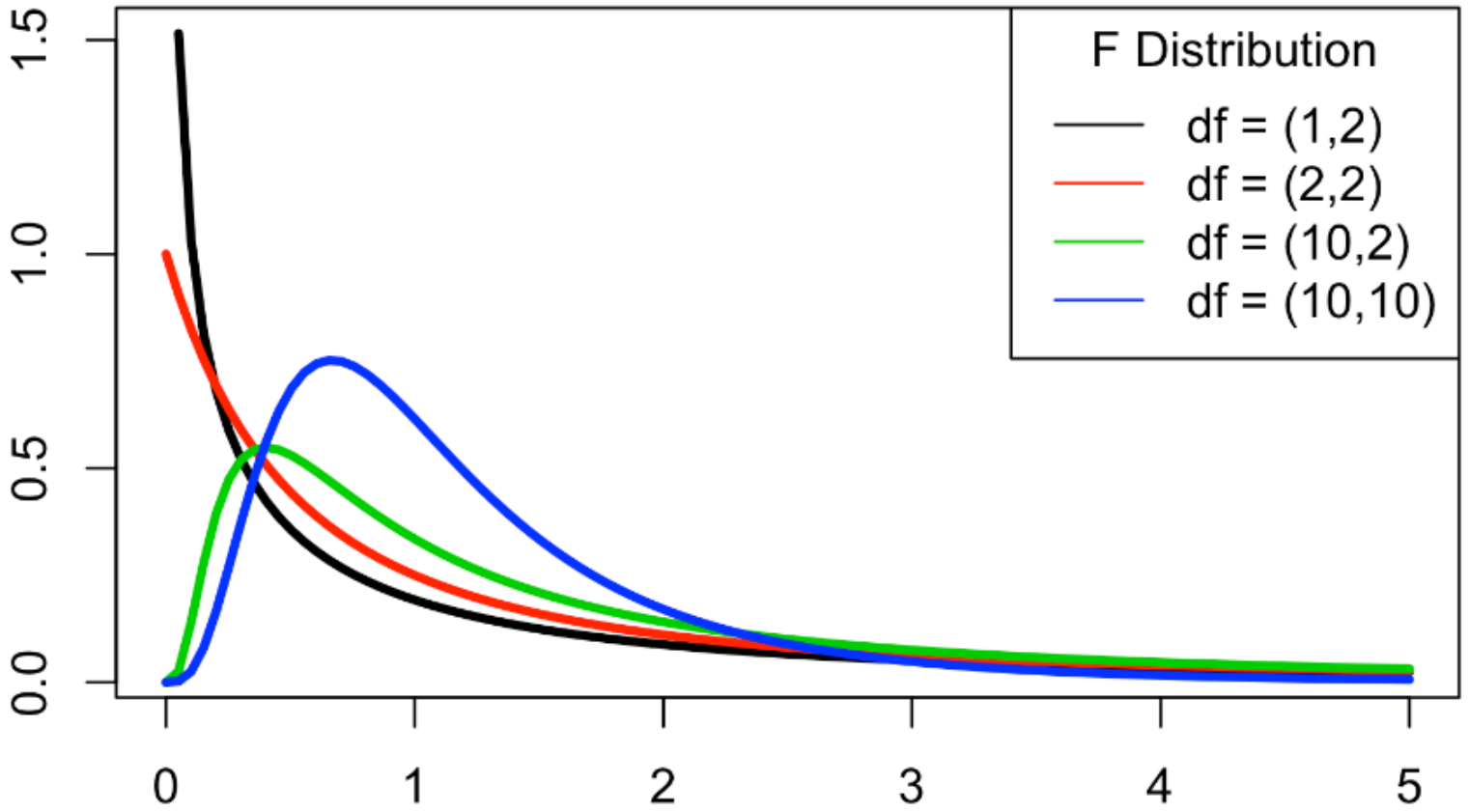


$P(x > 4) \approx 0.9989033$

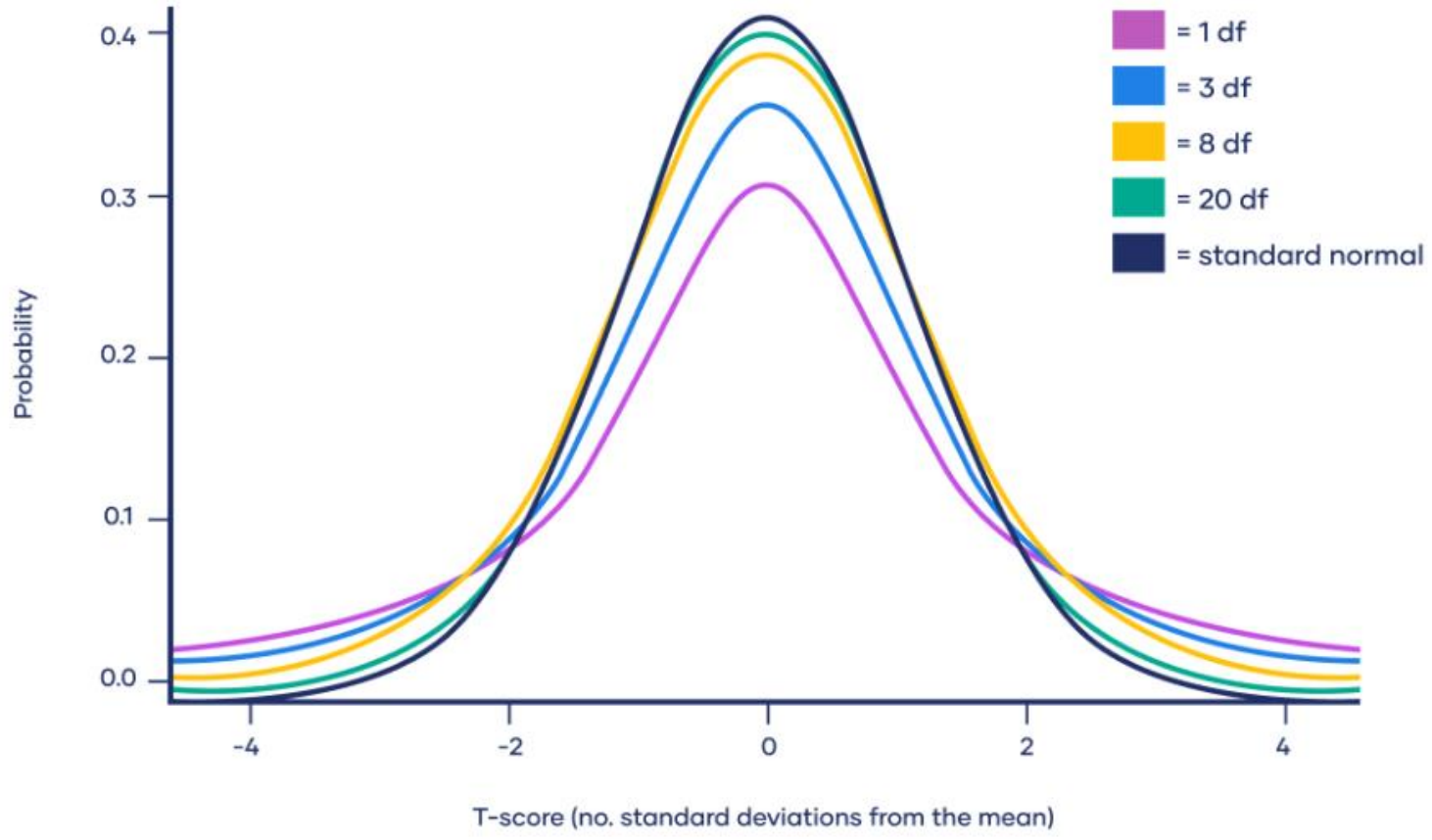


$P(x > 4) \approx 1$

F Distribution P-Values and Degrees of Freedom



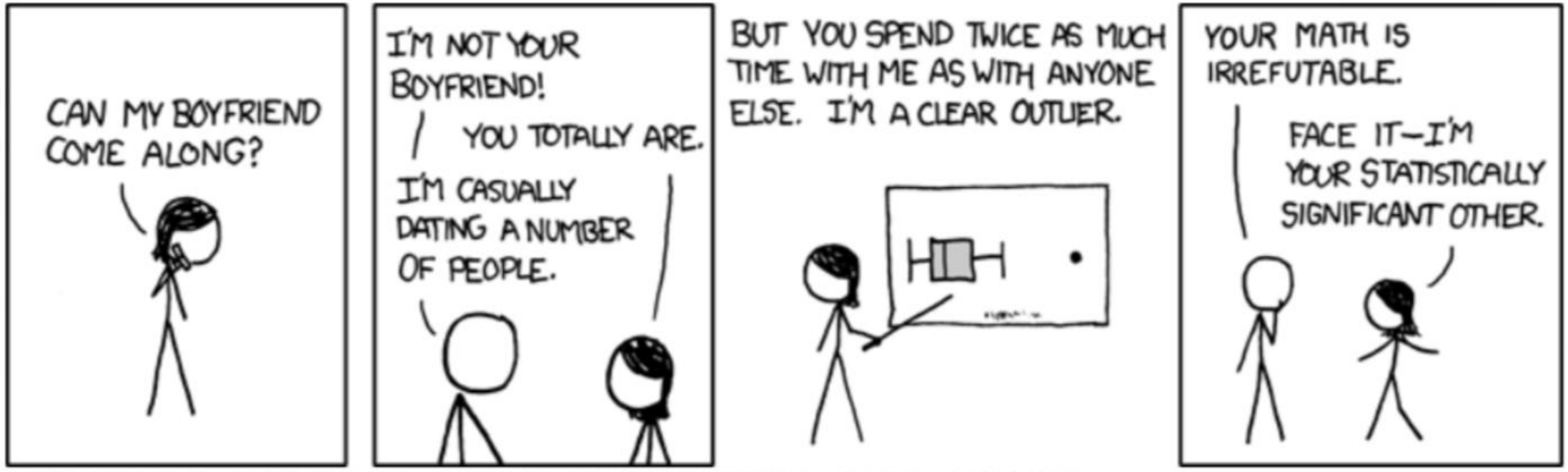
t Distribution P-Values and Degrees of Freedom



American Statistical Association (ASA) Statement on P-Values

- P-values can indicate how incompatible the data are with a specified statistical model.
- 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.
- Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
- Proper inference requires full reporting and transparency
- A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.
- By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Confidence Intervals and Connection to P-Values



I'm Your Statistically Significant Other

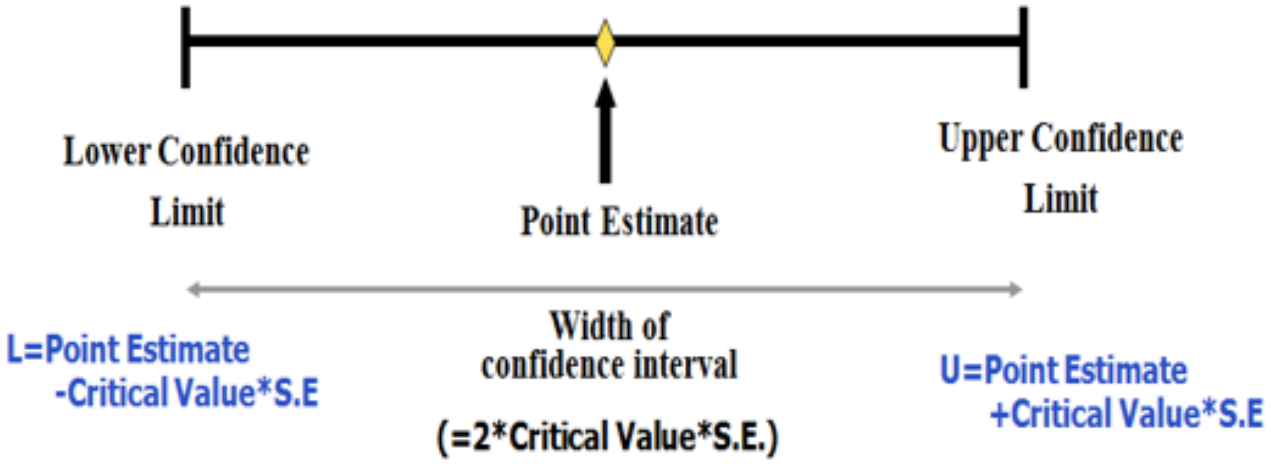
Confidence Intervals and Connection to P-Values

100(1-α)% Two-sided confidence interval for a parameter θ is an interval (L, U) such that

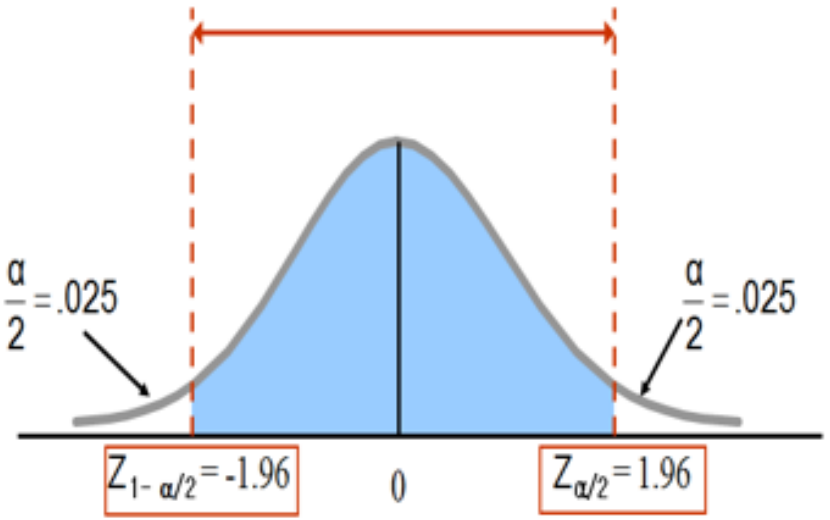
$$P(L \leq \theta \leq U) = 1 - \alpha \quad \Rightarrow \quad Confidence\ level = 1 - \alpha$$

•The general formula for all confidence intervals is:

Point Estimate ± (Critical Value) (Standard Error)



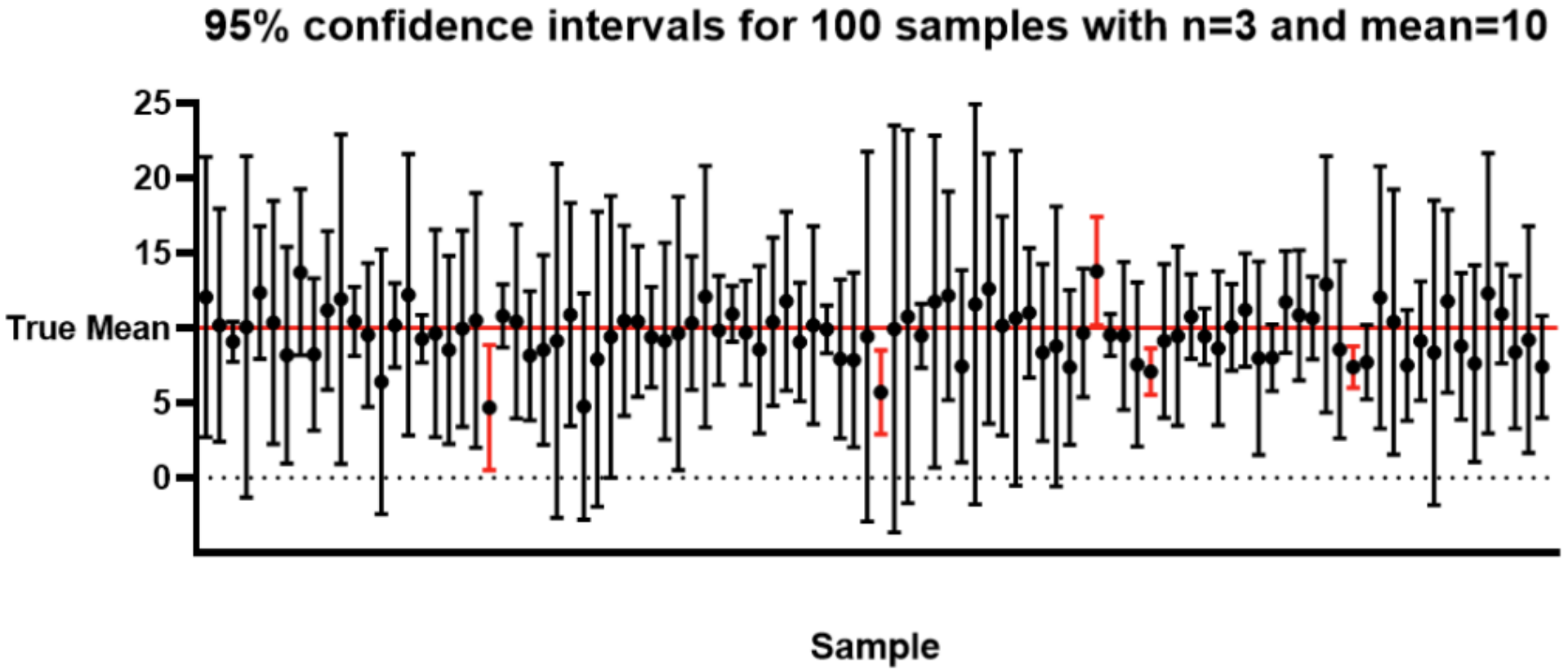
Consider a 95% confidence interval: 1−α =0.95



Source: https://medium.com/@ArtisOne/data-science-interview-questions-statistics-4a731ec0be59

Confidence Intervals and Connection to P-Values

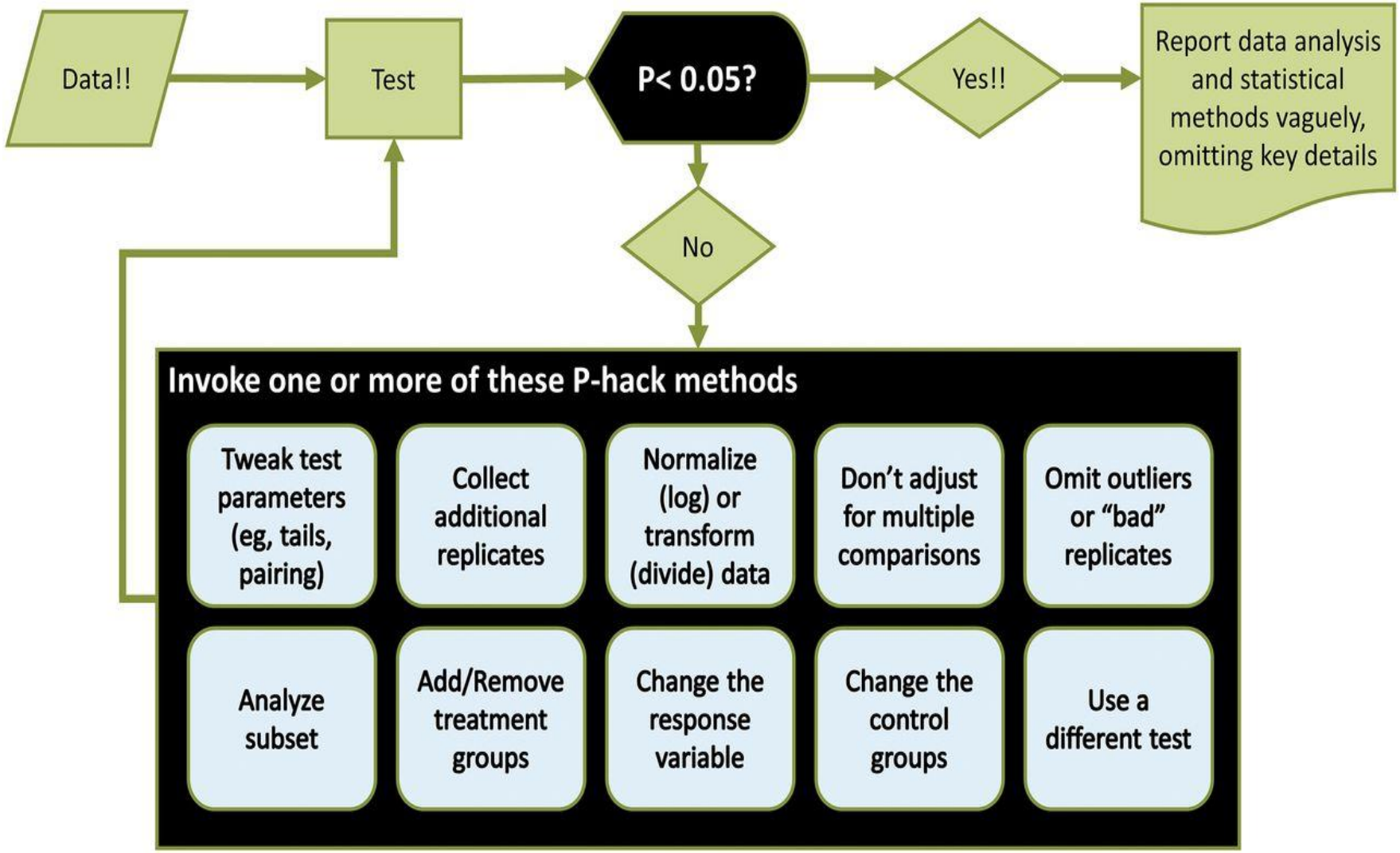
Equivalent Explanation



Source: https://www.graphpad.com/support/faq/the-distinction-between-confidence-intervals-prediction-intervals-and-tolerance-intervals/

What is P-Hacking?

P-value hacking, also known as **data dredging**, **data fishing**, **data snooping** or **data butchery**, is an exploitation of data analysis in order to discover patterns which would be presented as statistically significant, when in reality, there is no underlying effect.



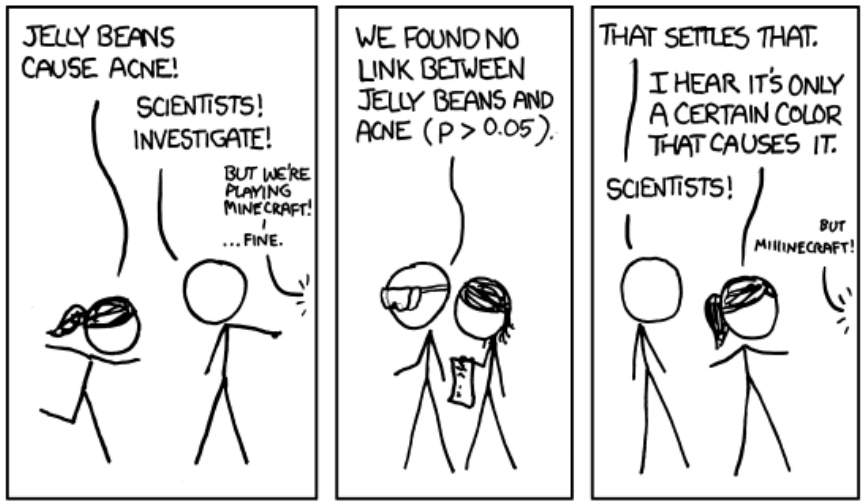
Source: https://jpet.aspetjournals.org/content/372/1/136/tab-figures-data

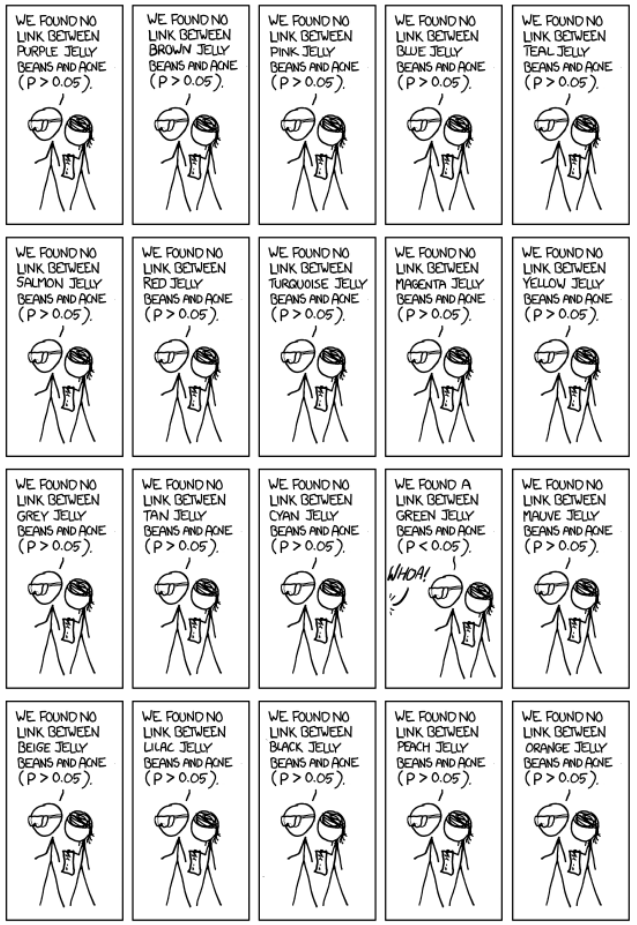
Availability in Packages

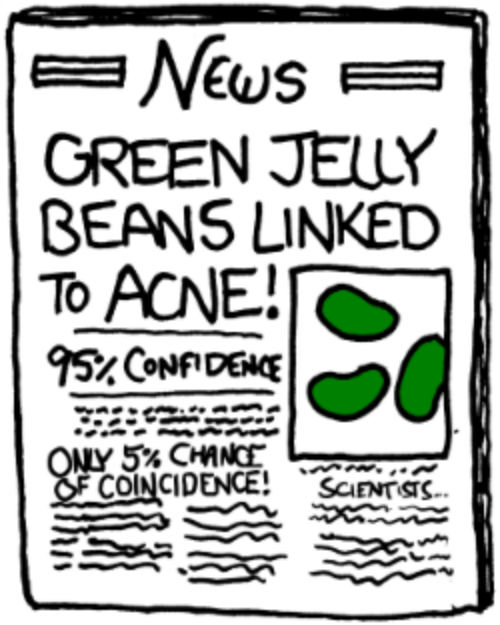
Roberto Perez

NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

For p-hacking slide!





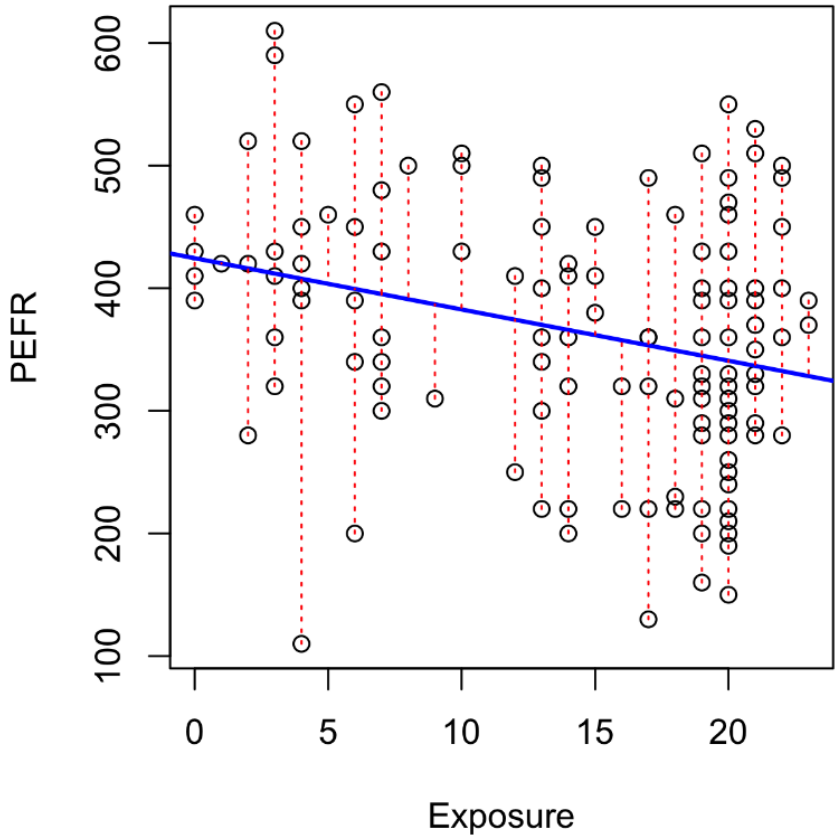


Traditional Regression

$$Y_i = f(X_i, \beta) + e_i$$

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta x_i)^2 = \underset{\beta}{\operatorname{argmin}} \|y - \beta x\|_2^2.$$

We get unbiased estimates of the coefficients!



Practical Statistics for Data Scientists by Peter Bruce, Andrew Bruce

Regular vs. Penalized Regression

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie

Stanford University, USA

[Received December 2003. Final revision September 2004]

Summary. We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

Keywords: Grouping effect; LARS algorithm; Lasso; Penalization; $p \gg n$ problem; Variable selection

Regular vs. Penalized Regression

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

Traditional Regression

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie
Stanford University, USA

[Received December 2003. Final revision September 2004]

Summary. We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

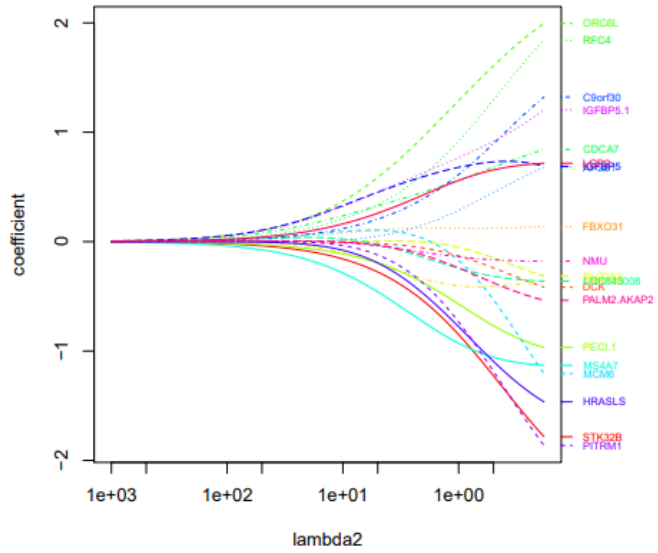
Keywords: Grouping effect; LARS algorithm; Lasso; Penalization; $p \gg n$ problem; Variable selection

Regular vs. Penalized Regression

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}\left(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1\right).$$

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Ridge Regression



Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie

Stanford University, USA

[Received December 2003. Final revision September 2004]

Summary. We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

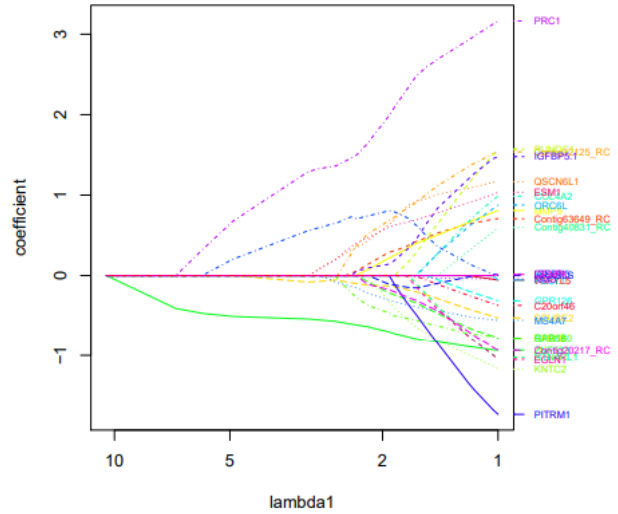
Keywords: Grouping effect; LARS algorithm; Lasso; Penalization; $p \gg n$ problem; Variable selection

Regular vs. Penalized Regression

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}\left(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1\right).$$

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

Lasso Regression



Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie

Stanford University, USA

[Received December 2003. Final revision September 2004]

Summary. We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

Keywords: Grouping effect; LARS algorithm; Lasso; Penalization; $p \gg n$ problem; Variable selection

Regular vs. Penalized Regression

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

J. R. Statist. Soc. B (2005)
67, *Part* 2, *pp.* 301–320

"standard errors are not very meaningful for strongly biased estimates such as arise from penalized estimation methods."

- Penalized methods introduce bias when estimating coefficients, which becomes a major component of MSE.
- Confidence statement based on variance can be misleading.

Regularization and variable selection via the elastic net

Hui Zou and Trevor Hastie

Stanford University, USA

[Received December 2003. Final revision September 2004]

Summary. We propose the elastic net, a new regularization and variable selection method. Real world data and a simulation study show that the elastic net often outperforms the lasso, while enjoying a similar sparsity of representation. In addition, the elastic net encourages a grouping effect, where strongly correlated predictors tend to be in or out of the model together. The elastic net is particularly useful when the number of predictors (p) is much bigger than the number of observations (n). By contrast, the lasso is not a very satisfactory variable selection method in the $p \gg n$ case. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso.

Keywords: Grouping effect; LARS algorithm; Lasso; Penalization; $p \gg n$ problem; Variable selection

Regularized Regression Packages

R Packages:

- No p-values
  - Glmnet
  - HDM
  - BigLasso
  - lars
  - Caret
  - h2o
- Lassopv
  - Uses the regularization strength when each predictor enters the active set of regularization path for the first time as the statistic. (Only for Lasso)

Generalized Additive Models (GAMs)

$$Y_i = \beta_0 + \sum_{j=1}^{p} s_j(x) + \epsilon_i$$

- We set the null hypothesis as s(x) = 0
- MGCV R package
  - Approximate p-values
  - "All p-values are computed without considering uncertainty in the smoothing parameter estimates"
  - The p-values are likely to "be somewhat too low when smoothing parameter estimates are highly uncertain."

Statistical Science
1986, Vol. 1, No. 3, 297–318

Generalized Additive Models

Trevor Hastie and Robert Tibshirani

Abstract. Likelihood-based regression models such as the normal linear regression model and the linear logistic model, assume a linear (or some other parametric) form for the covariates $X_1, X_2, \cdots, X_p$. We introduce the class of *generalized additive models* which replaces the linear form $\sum \beta_j X_j$ by a sum of smooth functions $\sum s_j(X_j)$. The $s_j(\cdot)$'s are unspecified functions that are estimated using a scatterplot smoother, in an iterative procedure we call the *local scoring* algorithm. The technique is applicable to any likelihood-based regression model: the class of *generalized linear models* contains many of these. In this class the linear predictor $\eta = \sum \beta_j X_j$ is replaced by the additive predictor $\sum s_j(X_j)$; hence, the name generalized additive models. We illustrate the technique with binary response and survival data. In both cases, the method proves to be useful in uncovering nonlinear covariate effects. It has the advantage of being completely automatic, i.e., no "detective work" is needed on the part of the statistician. As a theoretical underpinning, the technique is viewed as an empirical method of maximizing the *expected log likelihood*, or equivalently, of minimizing the *Kullback–Leibler distance* to the true model.

Key words and phrases: Generalized linear models, smoothing, nonparametric regression, partial residuals, nonlinearity.

https://cran.r-project.org/web/packages/mgcv/mgcv.pdf

https://projecteuclid.org/journals/statistical-science/volume-1/issue-3/Generalized-Additive-Models/10.1214/ss/1177013604.full

P-value "Alternatives"

Sam Kloese

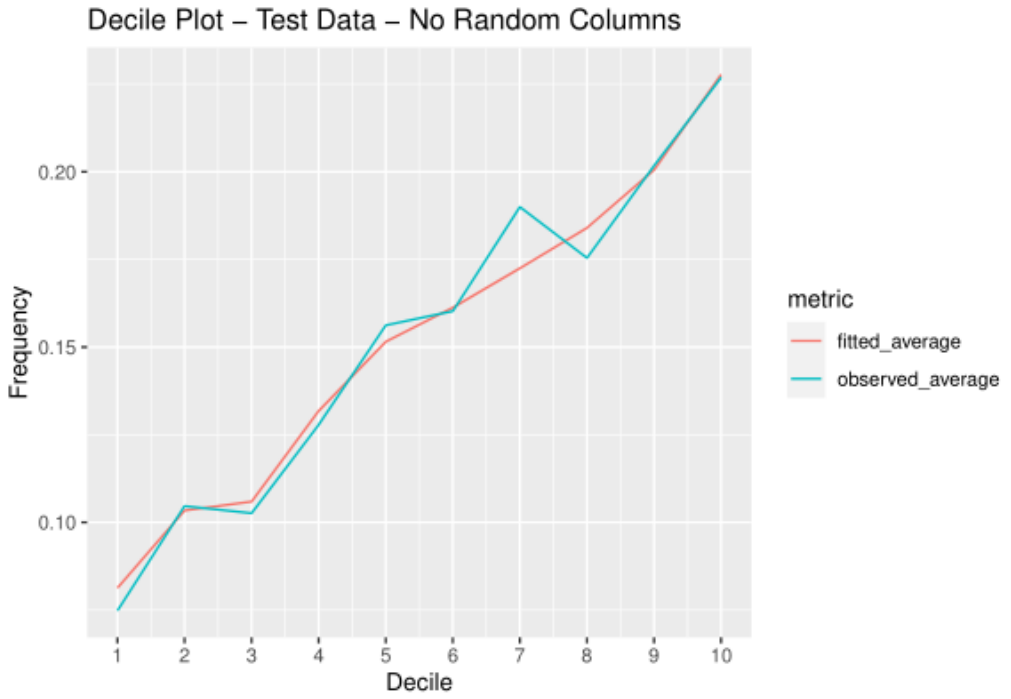
NAIC NATIONAL ASSOCIATION OF INSURANCE COMMISSIONERS

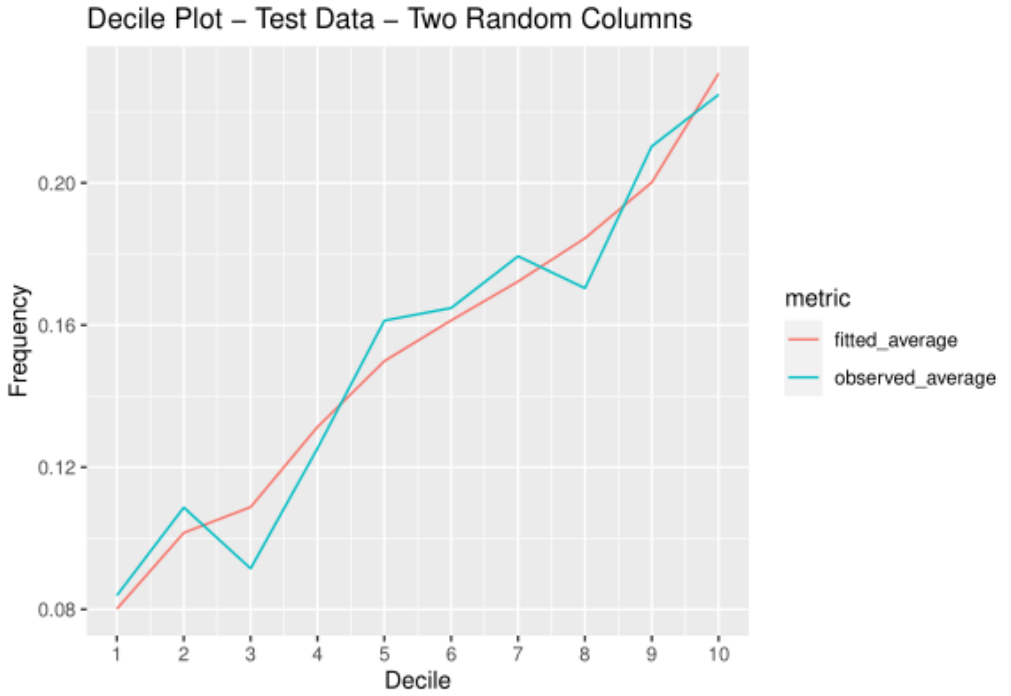
Final Model Lift Chart Is Not Enough

- Occasionally companies will reply that the overall model generalizes well to new data when asked about the significance of a specific variable
- A lift chart on holdout data may not look that bad if there is an insignificant variable included.

Final Model Lift Chart Is Not Enough

- Example:
  - GLM was built, data included 100 columns with random #'s 1-5
  - 7 random # columns had P-value < 0.05
  - Model A was built excluding all random #s
  - Model B was built including 2 random # columns with lowest p-values
  - The decile plot for Model B doesn't look that bad!





What can I ask for if p-values are unavailable?

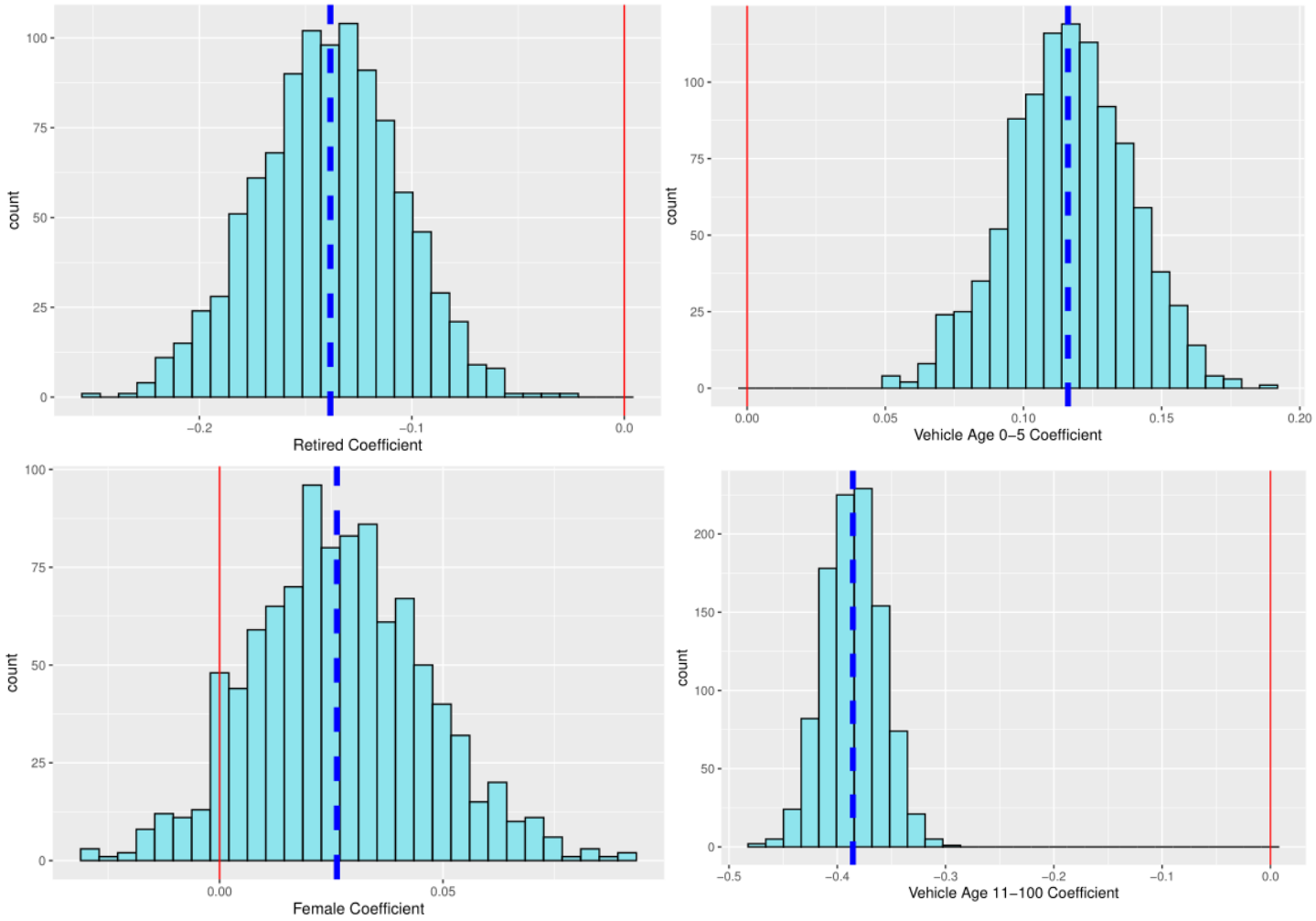
- P-values are a common metric for variable significance
- Other tests that may help address the question of significance
  - Bootstrapping: Do variations to the data result in radically different coefficients?
  - Cross Fold Validation: Are the coefficients consistent across folds?
  - GLM Reference Model: What are the p-values from a similar GLM?

1. Bootstrapping

- The model could be run several times times on bootstrapped samples
  - Bootstrapping involves sampling from replacement from the original dataset
  - The bootstrap samples have the same number of records
  - Each model run would result in different coefficients, since the dataset is different
- Evaluating the coefficients
  - The range of coefficients can be evaluated by variable
  - If the range of coefficients is narrow, it raises our confidence in statistical significance
  - If the range of coefficients is quite wide, it is a sign of model instability
  - Histograms can help visualize the range and distribution of coefficients
    - Narrower histograms with tall peaks are preferable
    - Variables where the histogram crosses over the 0 line should be further scrutinized

1. Bootstrapping

- Example:
  - Elastic Net model was built
    - Glmnet package in R does not produce p-values
    - Instead, the same model was run 1,000 times on bootstrapped data samples
    - Histograms were analyzed to determine variability of coefficients by variable



2. Cross Fold Validation

- K fold validation is a common cross fold validation type
- Training data is broken up into k folds
- Ideally, the modeler still has a true holdout dataset for final model validation

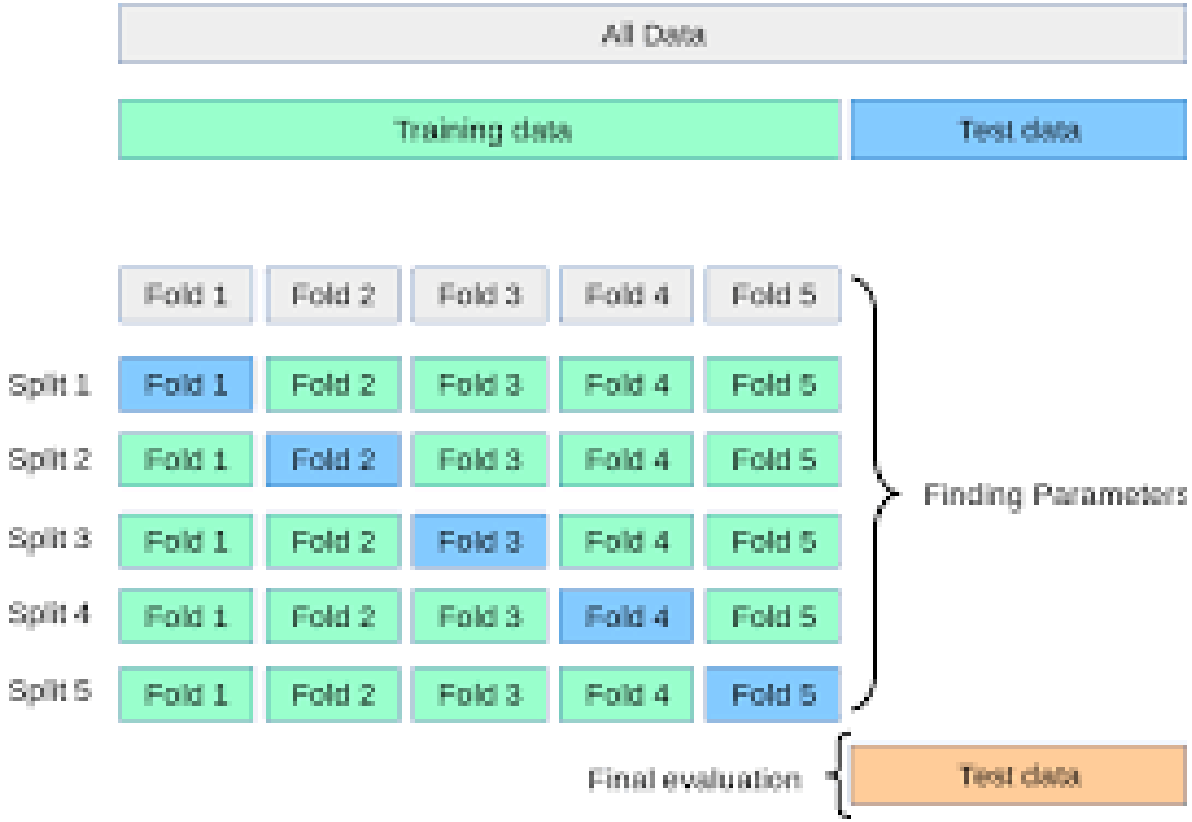


Image from https://scikit-learn.org/stable/modules/cross_validation.html

2. Cross Fold Validation

- The model is trained k times
- The predictions for a specific fold are based on a model trained with all other folds
- Each time the model is trained, a set of coefficients is determined
- The modeler may need to specify that they want each fold's coefficients to be saved

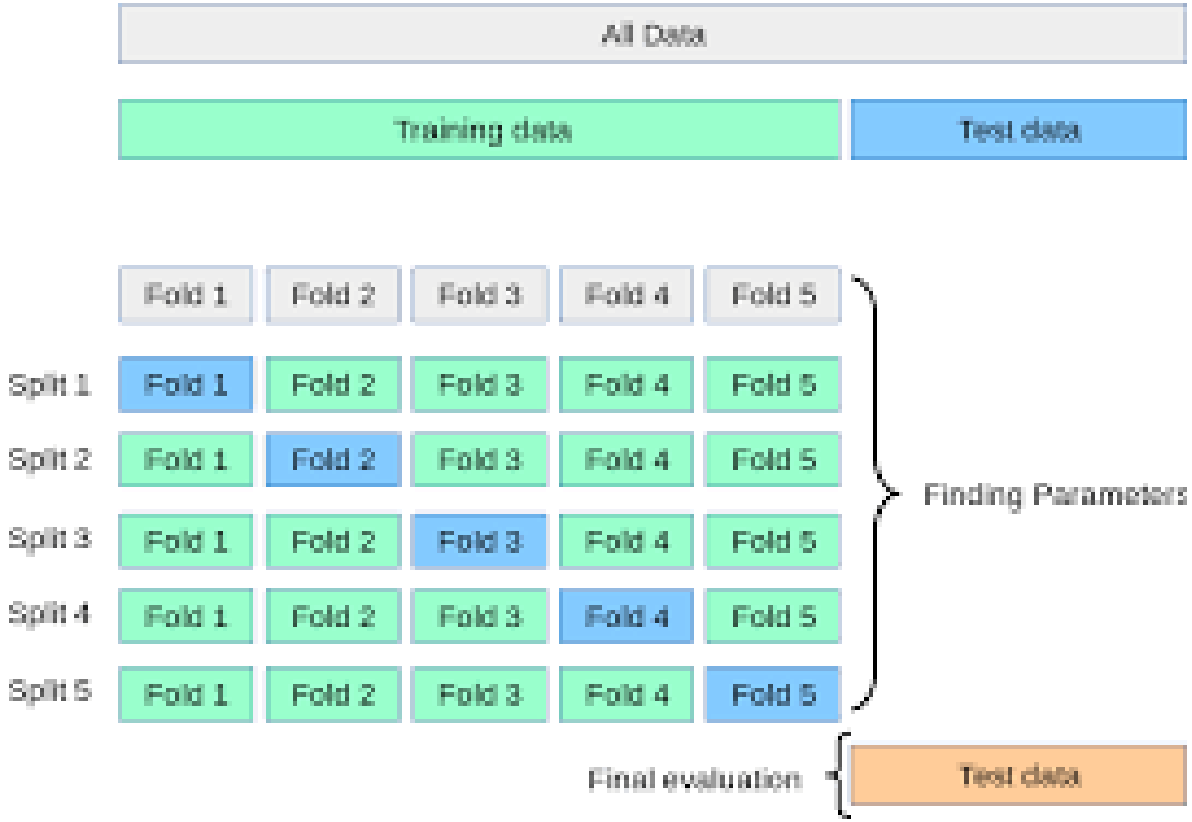


Image from https://scikit-learn.org/stable/modules/cross_validation.html 10/24/2022

2. Cross Fold Validation

- Often the final model is run using 100% of the training data (including all folds)
- Companies often just provide coefficients associated with the final run
- However, reviewing the coefficients from the k folds may be useful

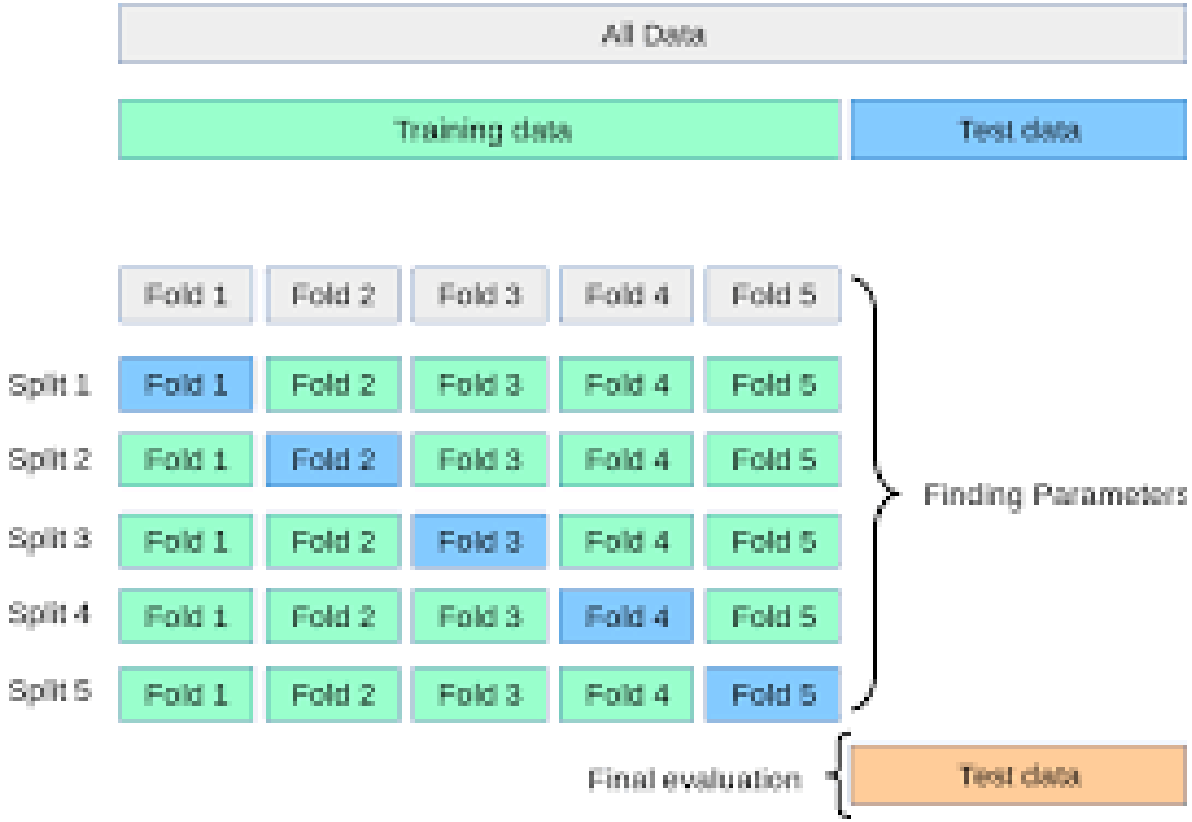


Image from https://scikit-learn.org/stable/modules/cross_validation.html

2. Cross Fold Validation

- The model reviewer can ask for the coefficients from each fold
  - If k fold validation was used, there are k different sets of coefficients
  - Unsure of the ideal k value
    - Small k values mean there are less sets of coefficients to analyze
    - Large k values mean that each model has a larger share of overlapping training data
  - Each model run would result in different coefficients, since the folds in training are different
- Evaluating the coefficients by fold
  - The range of coefficients can be evaluated by variable
  - If the range of coefficients is narrow, it raises our confidence in statistical significance
  - If the range of coefficients is quite wide, it is a sign of model instability
  - Histograms can help visualize the range and distribution of coefficients
    - Narrower histograms with tall peaks are preferable
    - Variables where the histogram crosses over the 0 line should be further scrutinized

2. Cross Fold Validation

- Example:
  - Elastic Net model was built
    - Glmnet package in R does not produce p-values
    - Instead, the same model was run on 5 different folds
    - Consistency across folds can be analyzed

		1	2	3	4	5	Full Dateset
Usage	All Trips	0.530	0.518	0.560	0.755	0.690	0.618
	Professional	0.233	0.240	0.256	0.252	0.213	0.239
	Retired	-0.123	-0.134	-0.133	-0.152	-0.148	(0.138)
	Work Private			*Base*			
Gender	Female	0.039	(0.002)	0.034	0.020	0.041	0.026
	Male			*Base*			
Driver Age	16 - 20	0.132	0.124	0.185	0.080	0.271	0.161
	21 - 30	(0.037)	0.042	0.009	0.001	0.042	0.011
	31 - 40			Base			
	41 - 50	(0.039)	(0.043)	(0.042)	(0.022)	(0.013)	(0.032)
	51 - 60	0.028	0.009	0.005	0.049	0.043	0.027
	61+	0.067	0.063	0.063	0.104	0.112	0.082
Vehicle Age	0 - 5	0.116	0.132	0.109	0.109	0.114	0.116
	6 - 10			*Base*			
	11+	(0.400)	(0.385)	(0.379)	(0.404)	(0.360)	(0.386)
Vehicle Din	0 - 50	(0.606)	(0.607)	(0.534)	(0.631)	(0.635)	(0.601)
	51 - 100			*Base*			
	101 - 150	0.211	0.213	0.198	0.220	0.220	0.212
	151+	0.271	0.227	0.250	0.275	0.241	0.253

3. GLM Reference Model

- GLMs provide p-values in most software
- A GLM could be built which is as similar as possible to the model in question
  - This is probably more appropriate when the model in question is still some type of linear model (Lasso, ridge, elastic net)
- Consider the GLM provided p-values a reasonable approximation for the model in question
  - P-values from the GLM may be a little underestimated
- The modeler should describe why their model type is preferable to a GLM for their modeling purpose.
  - Once they have a similar GLM, they should describe why they favor the other model
  - Why not use Lasso or Elastic Net for variable selection, but run a GLM on the final features?
- If the coefficients are radically different in the reference GLM, the GLM p-values may not be as relevant

3. GLM Reference Model

- Example:
  - Elastic Net model was built
  - GLM model was built with the same variables
  - The coefficients are compared side by side
  - Low p-values from the GLM suggest the variables should be significant

		Elastic Net	Reference GLM	GLM p-value
Usage	All Trips	0.618	0.622	< 0.001
	Professional	0.239	0.239	0.002
	Retired	(0.138)	(0.142)	< 0.001
	Work Private	*Base*	*Base*	
Gender	Female	0.026	0.027	0.157
	Male	*Base*	*Base*	
Driver Age	16 - 20	0.161	0.170	0.398
	21 - 30	0.011	0.014	0.769
	31 - 40	*Base*	*Base*	
	41 - 50	(0.032)	(0.031)	0.320
	51 - 60	0.027	0.029	0.329
	61+	0.082	0.087	0.016
Vehicle Age	0 - 5	0.116	0.116	< 0.001
	6 - 10	*Base*	*Base*	
	11+	(0.386)	(0.386)	< 0.001
Vehicle Din	0 - 50	(0.601)	(0.606)	< 0.001
	51 - 100	*Base*	*Base*	
	101 - 150	0.212	0.213	< 0.001
	151+	0.253	0.255	< 0.001

Comparison of Alternatives

- Bootstrapping
  - Can provide a large distribution of coefficients
  - May be impractical for large datasets due to model run time
- K Fold Validation
  - Typically provides a much smaller distribution of coefficients
  - Often requires the modeler to change programming to save coefficients from each fold
  - Takes less time than the bootstrapping approach since there are less model runs
- GLM Reference Model
  - Less appropriate for non-linear models
  - The p-values may not be relevant if the beta coefficients are radically different from the model in question