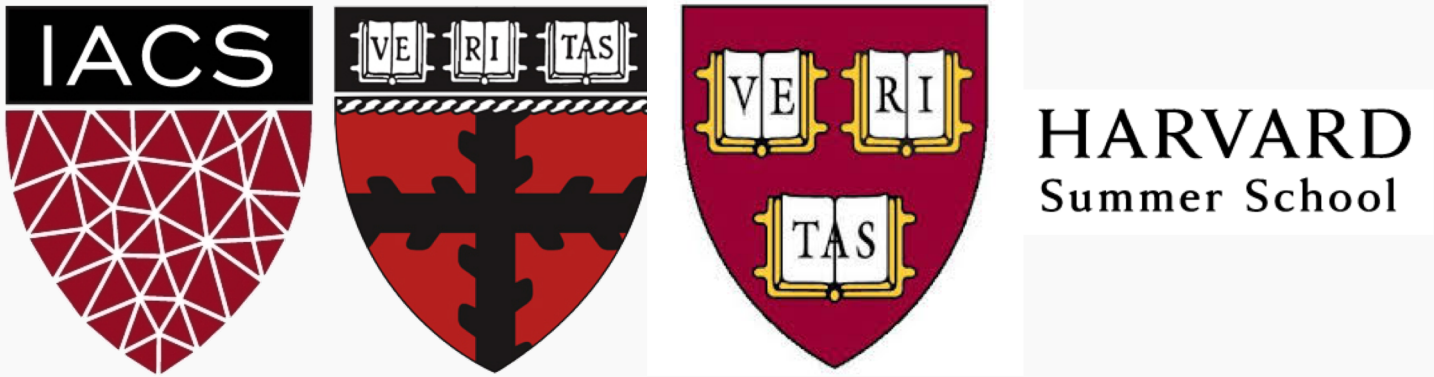
Lecture #10: Visualizing the Black Box and AB Testing

CS-S109A: Introduction to Data Science

Kevin Rader



Outline

Variable Importance

Permutation Importance

Interpretation through Predictions

Adding Uncertainty

LIME

Causal Effects and AB Testing

Variable Importance

Variable Importance for Tree-Based Models

How does sklearn determine **variable importance** (feature_importance) from a tree-based model?

- It determines the improvement in the loss function every time a predictor is involved in a split.

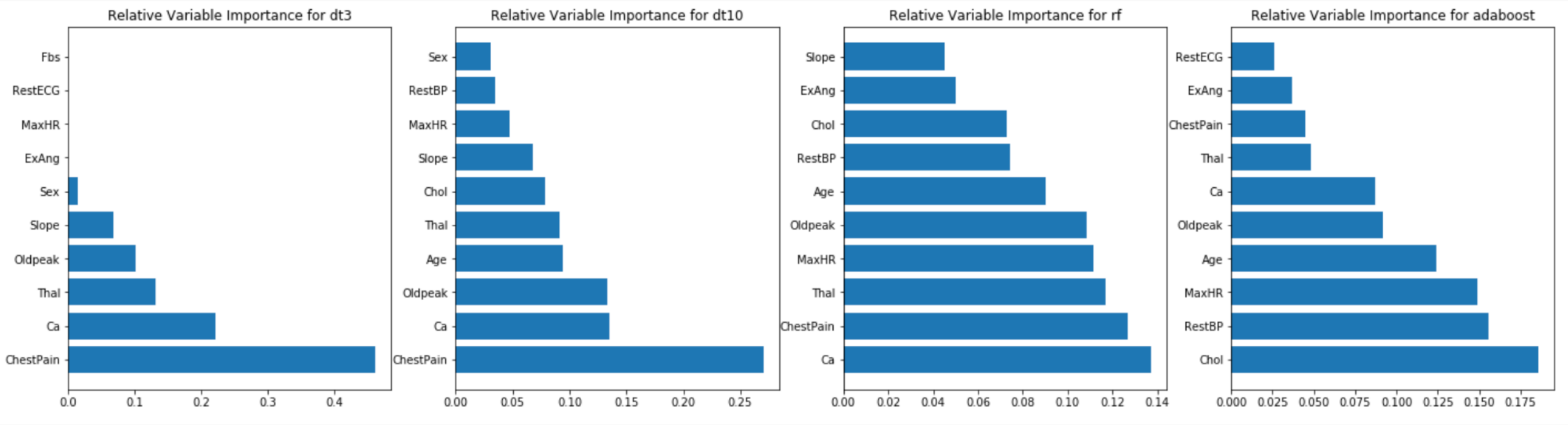
- More specifically, it calculate the total amount that the SSE (for regression) or Gini index (for classification) is improved (decreased) due to splits over a given predictor (averaged over all B trees if a bagged/random forest method).

How should variable importance compare across the various different tree models we've considered (trees, random forests/bagging, and boosting)?

A picture is worth a thousand words…

Variable Importance for trees, bags, and boosts

Below are the variable importance plots for the top 10 predictors for each of a (i) decision tree with maxdepth=3, (ii) decision tree with maxdepth=10, (iii) a random forest, and (iv) an adaboost classifier.



Compare them? Are the differences surprising?

Other Variable Importance Measures

What other approaches can be taken to measure variable importance?

Alternative:

- Record the prediction accuracy on the *oob* samples for each tree.
- Randomly permute the data for column j in the *oob* samples, then record the accuracy again.
- The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of variable j in the random forest.

Permutation Importance

Permutation Variable Importance

This idea of re-permuting a variable and refitting a model to see how much more poorly it performs is called **permutation feature importance.**

It is sometimes preferred to the standard feature importance, why?

When two features are correlated and one of the features is permuted, the model will still have access to the feature through its correlated feature.

What is the one glaring disadvantage to this?

Computational time, and things may not ‘add up’.

Permutation Variable Importance in sklearn

To perform permutation variable importance in sklearn (or keras), one can use the ELI5 library:

https://eli5.readthedocs.io/en/latest/index.html

This is easy to do with sklearn models.

An example is worth a thousand words:

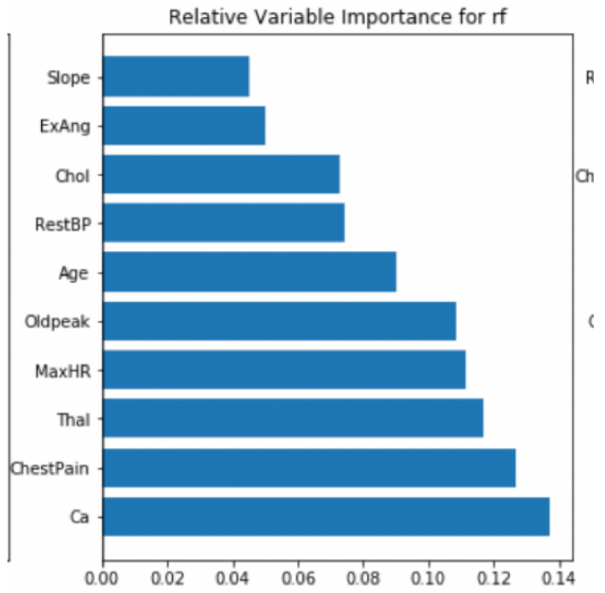
Permutation Variable Importance Example

```
#permutation importance
from eli5.sklearn import PermutationImportance
from eli5.permutation_importance import get_score_importances


perm = PermutationImportance(randomforest).fit(X_test, y_test)
#eli5.show_weights(perm,feature_names=X.columns)
print(X.columns)
eli5.explain_weights(perm)
```

```
Index(['Age', 'Sex', 'ChestPain', 'RestBP', 'Chol', 'Fbs', 'RestECG', 'MaxHR',
       'ExAng', 'Oldpeak', 'Slope', 'Ca', 'Thal'],
      dtype='object')
```

Weight	Feature
0.1082 ± 0.0161	x12
0.0689 ± 0.0435	x7
0.0393 ± 0.0675	x11
0.0361 ± 0.0636	x2
0.0328 ± 0.0359	x8
0.0295 ± 0.0245	x9
0.0295 ± 0.0131	x1
0.0197 ± 0.0245	x3
0.0164 ± 0.0688	x10
0.0131 ± 0.0131	x6
0.0098 ± 0.0161	x4
0 ± 0.0000	x5
-0.0098 ± 0.0334	x0



The problem with Variable Importance

Variable Importance is great! It tells you what features are important in shaping the model.

But what is missing?

- It does not give any measure for how the predictors are related to the response (positive, negative, quasi-linear, curved, interactions, etc.).
- This is where the parametric model wins out! Inference and interpretations are much easier and the whole point in these models.

What can we do to measure these relationships in a machine learning or non-parametric model? Think: what did we do with *k*-NN?

What needs to be done algorithmically to put this in practice?

Interpretation through Predictions

Parametric vs. Nonparametric models

In a machine learning model (like ensemble methods), the association between predictors and the response are not measured directly as these models are 'black box' models:

Inputs (*X*, predictors). → black box (sklearn, etc.) → Outputs (*Y*, response)

What if we care about how the predictors relate to the response? This is where we need to figure out what the black box is doing to transform the inputs into the outputs.

Simplest Approach: observed $\hat{Y}$ vs. X_j

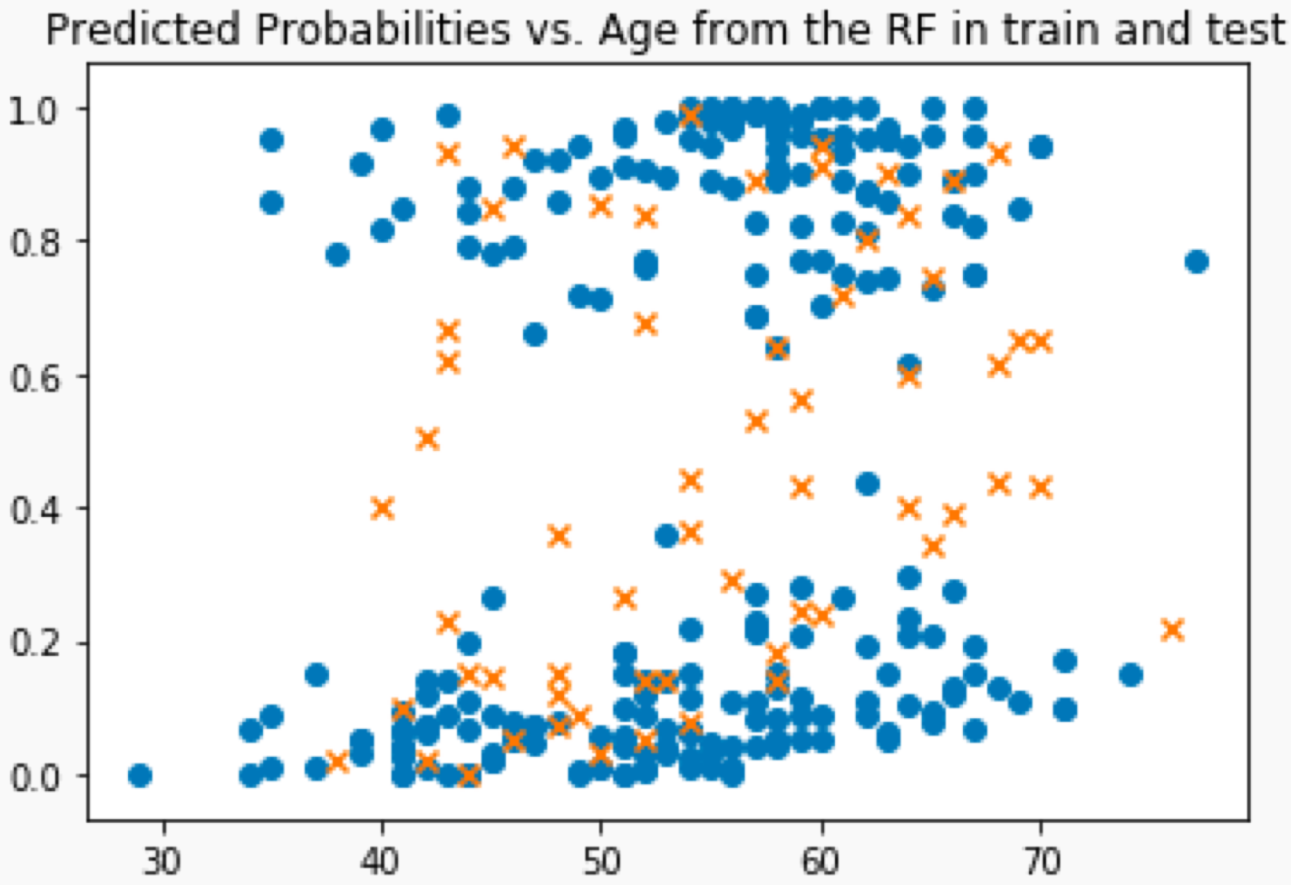
Use predict (or better yet, predict_proba) to plot the observed predicted values vs. the observed values for X_j.

What is a problem with this approach how can we fix it?

The fix is not so easy. We cannot just fit a logistic regression model so easily to the predicted probabilities. Why not?

An example is worth a thousand words…

Example: observed $\hat{Y}$ vs. X_j



Unboxing the black box

Inputs (*X*, predictors). → black box (sklearn, etc.) → Outputs (*Y*, response)

This gives us our approach: vary the inputs (the predictors) and see what happens to the response.

If we care about the 'marginal' or 'conditional' effect of how a specific *X* relates to *Y*, then we should vary only one predictor at a time.

How should we handle the other predictors? That is to say, what value should we *hold* them at?

Unboxing the black box (cont.)

There are two general approaches to interpreting the machine learning model through predictions:

The approach is just like in multiple regression: what is the marginal effect of a unit change in X_j holding all the other predictors constant.

So at what values should we hold the other predictors?

There are two general approaches **holding the other predictors constant**:

1. Predict $\hat{Y}$ at the **mean (or most common) value** for each of the other predictors, vary only the predictor you care about, X_j, and plot the predictions $\hat{Y}$ vs. X_j.
2. Predict $\hat{Y}$ at the **observed values of** for all the other predictors, vary only the predictor you care about, X_j, and plot the predictions $\hat{Y}$ vs. X_j. Essentially this means creating a new data frame for each observation, and imputing all reasonable values of X_j in.

What if we want the joint relationship with two predictors?

How can we look at how the response relates to two predictors at once (to directly interpret interactions)?

This is a bit trickier to do. Why?

We can go back to our friend: the classification boundary!

What can we do in a regression problem?

Should I use PCA components?

What if we want an overall picture of how the response relates to the predictors?

Well, this is where PCA components sometimes come in to play.

But this is no Bueno for interpretability. Why?

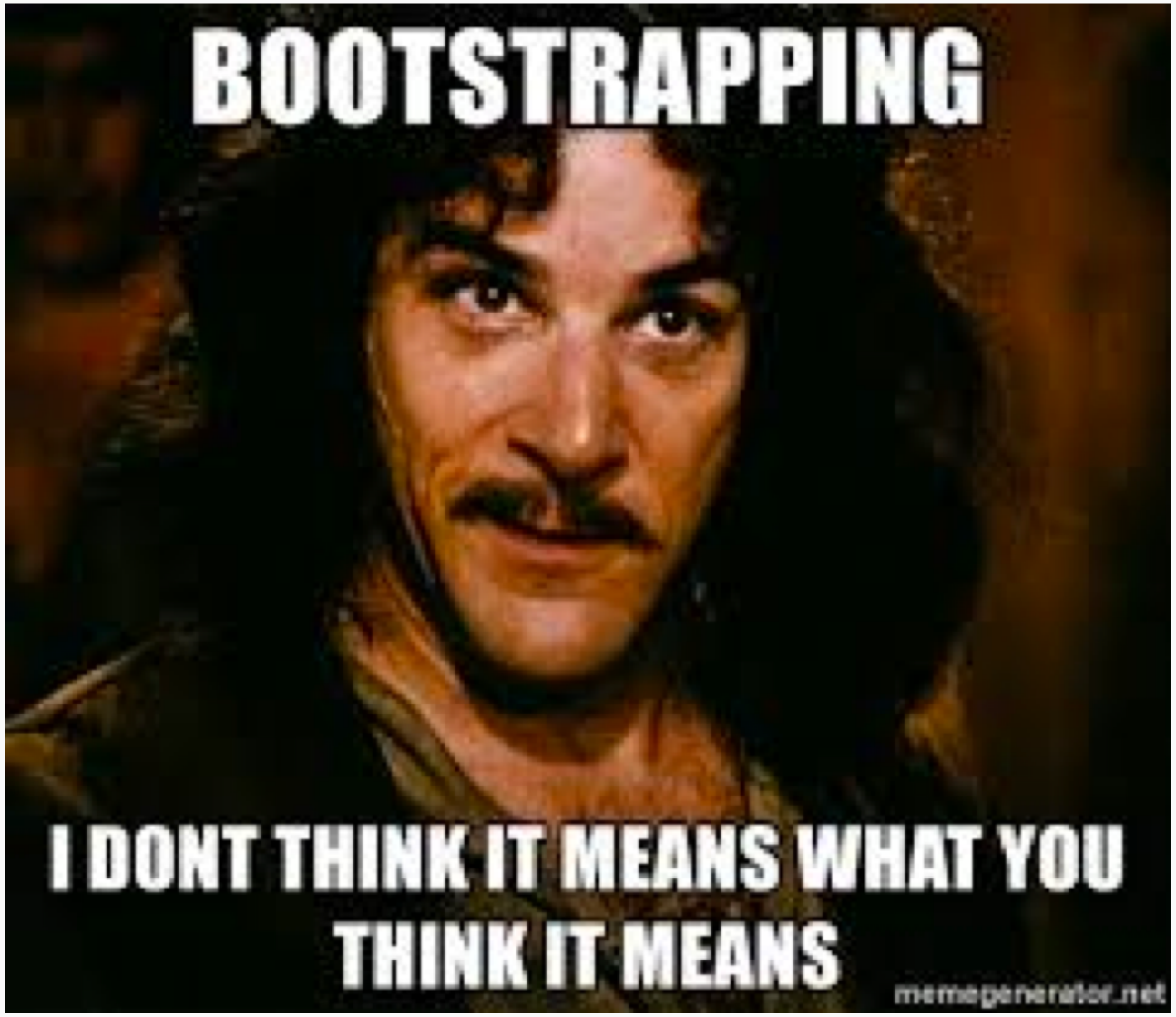
Adding Uncertainty

Adding Uncertainty

Not only should we plot predictions, but we can (should) also include the uncertainty of these predictions.

How can we add uncertainty measures to these predictions? How do we do it for linear/logistic regression?

Bootstrapping!


BOOTSTRAPPING
I DONT THINK IT MEANS WHAT YOU THINK IT MEANS
memegenerator.net

Bootstrapping Basics, review

How do we perform a bootstrap? What are the key steps?

The bootstrap algorithm is an approach to mimic the uncertainty of taking the observed sample from a population. So we:

1. Sample the same number of observations as was observed (in train).
2. Sample with replacement (so we get some randomness).

Why does this work?

Bootstrapping Basics, review

When is a bootstrapping approach used?

1. When a probabilistic closed form solution is not known or not easy to get at.
2. When the assumptions for a parametric model break down.
3. When there is no 'formula' at all to add the uncertainty into a calculation.
4. Or we are just being lazy.

Bootstrapping Basics, review

How do we use the results of a bootstrap technique?

It provides us a re-created **estimate** every time we bootstrap.

This estimate could be a mean, a beta, a plot of average predictions, etc.

But they will almost always be for a summary, not a single observation, and thus the sample size is already taken care of (the *n* in the formula). Aka, do not divide by $\sqrt{n}$ again!

Bootstrapping for Predictions

So we can refit a model on a bootstrap sample of data, and look at the predicted probabilities each time.

Then we can take the top 97.5^{th} percentile and bottom 2.5^{th} percentile to build a 95% interval for the predictions!

But how does this work for random forests and ensemble models? Yikes!

LIME

LIME

One package that can be very useful to help interpret our machine learning models is **LIME**:

Local

Interpretable

Model-agnostic

Explanations

What do each of those contributions represent?

LIME (cont.)

Local fidelity: The explanation should be able to explain how the model behaves for individual observations.

Interpretable: The explanation must be easy to understand by humans (but may depend on the target audience).

Model-agnostic: The method should be able to explain any model.

Explanations: self-explanatory ☺. This is the whole goal.

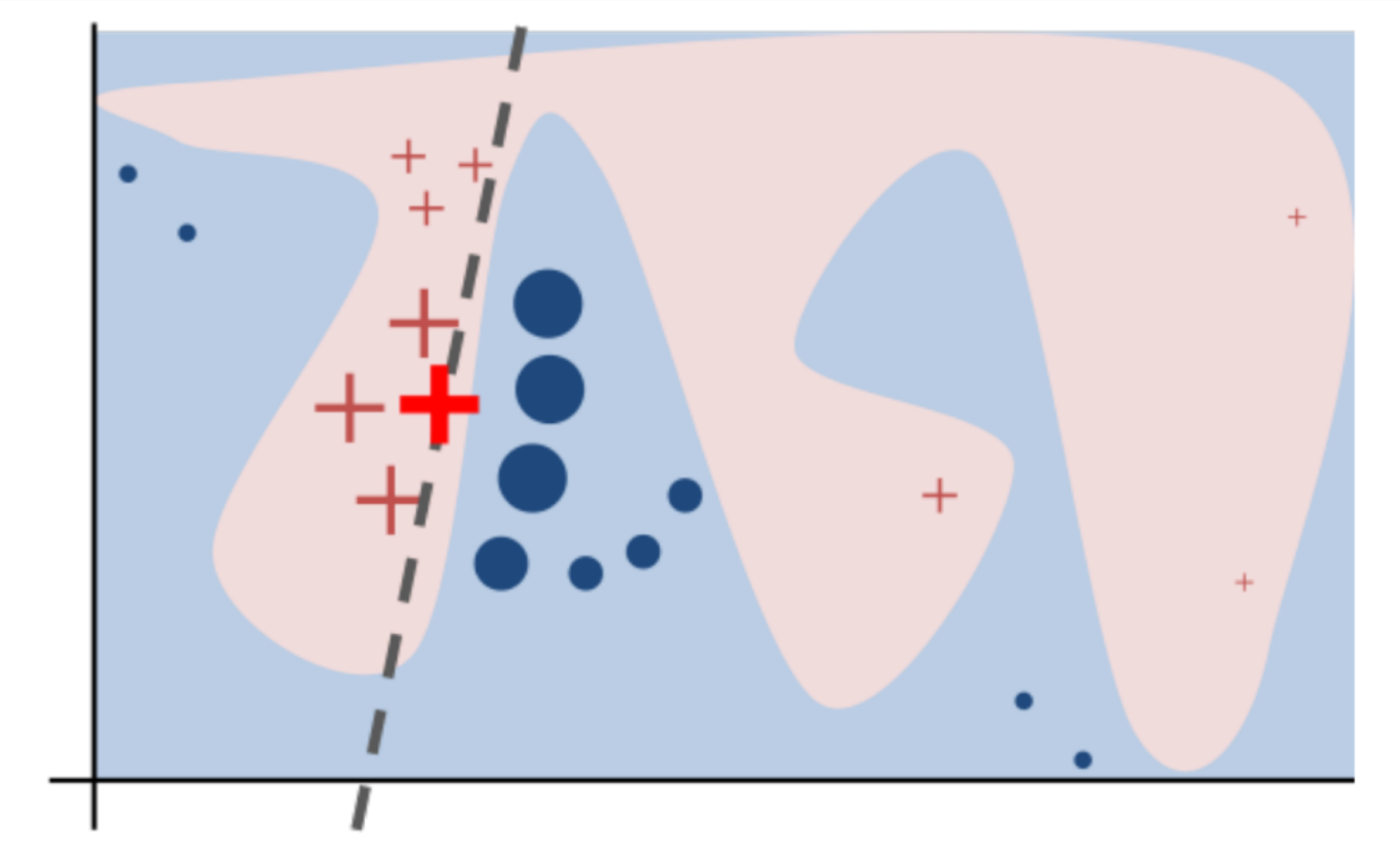
LIME: in a nutshell

The approach is to use a directly interpretable model (aka, linear regression) to help explain a model that is not directly interpretable.

To interpret how predictors are related to the response for a specific observation, synthetic data are created by perturbing that observations' *X*s. Then the original model is used to create new *Y*s.

The interpretable model is then fit to the synthetic data, and the results are saved.

LIME: an illustrative picture



LIME: the algorithm

The 5-steps:

- **Select your observation of interest** for which you want to have an explanation of its black box prediction.
- **Perturb your observation** (rx: normal dist. perturbations for numerical predictors) and get the black box **predictions** for these new points.
- **Weight the new observations** based on their proximity to the observation of interest.
- **Train an interpretable local model** (decision tree or linear regression with Ridge penalty) on the weighted, perturbed observations.
- Explain the prediction by **interpreting the local model**.

LIME: Pros and Cons

Advantages:

- Easy to apply LIME to tabular data, text data, and images
- They make **human-friendly** explanations.
- Comes with a **fidelity measure** that gives us a good idea of how reliable the interpretable model is in explaining the black box predictions.

Disadvantages:

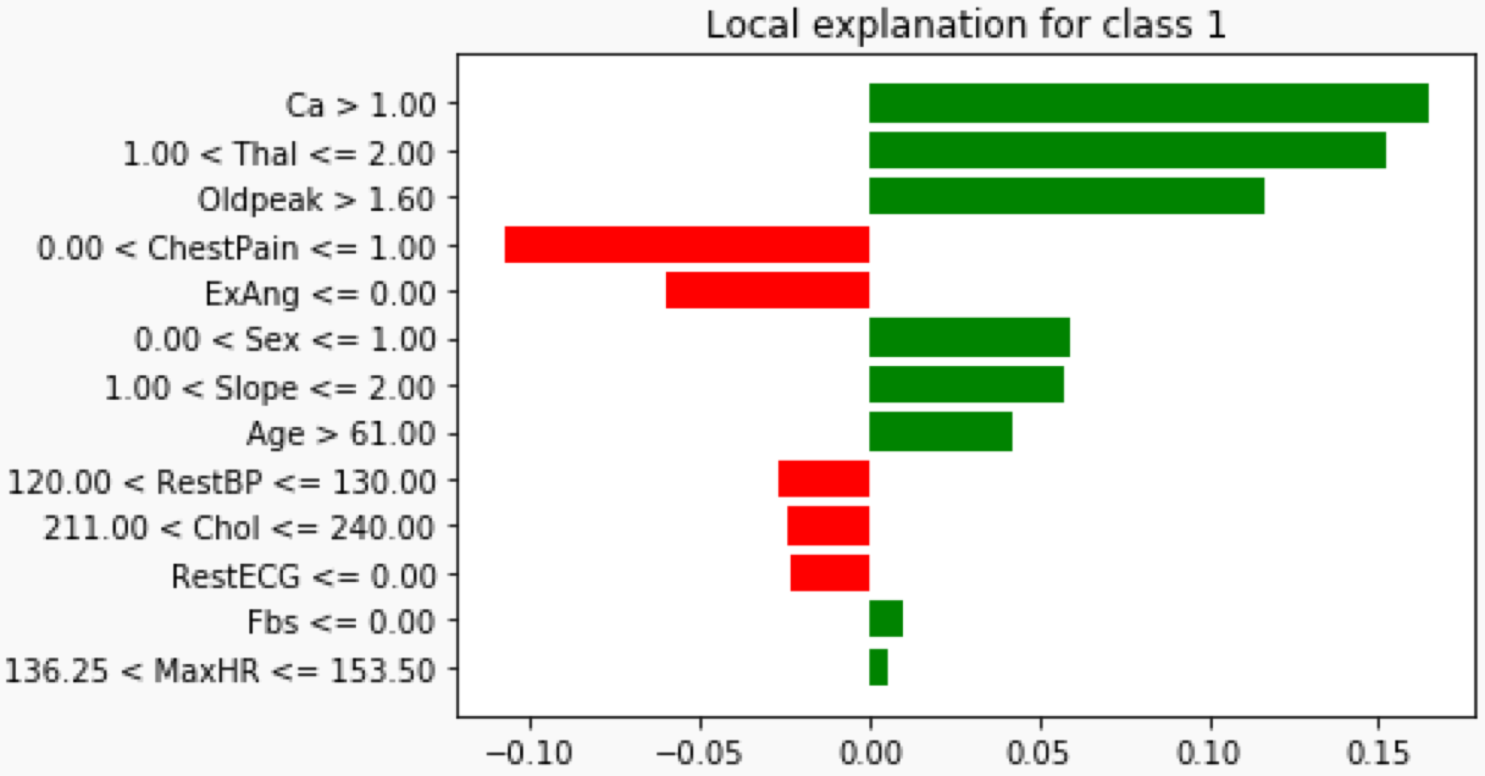
- The correct definition of the neighborhood is a very big, unsolved problem when using LIME with tabular data leading to **instability**.
- Interpretations are single **observation-specific**. Do not get easy generalities as to how *X* relates to *Y*.

LIME: the equation

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

- x is a specific observation at which you want to do interpretations
- f is the machine learning model that is hard to explain
- G is the class of interpretable models you are considering using to explain f, and g is the specific model use estimate
- $\mathcal{L}$ is a measure for how well g approximates f
- Π_x is the locality/neighborhood of the observation being used around x
- $\Omega(g)$ is the restriction placed on g (number of non-zero entries, for example)
- $\xi(x)$ is the resulting interpretation of model f using g at x.

LIME: an output



Causal Effects and AB Testing

- Causal Effects
- Experiments and *AB*-testing
- *t*-tests, binomial *z*-test, randomization tests, oh my!
- Obama 2008

Association vs. Causation

In many of our methods (regression, for example) we often want to measure the association between two variables: the response, *Y*, and the predictor, *X*. For example, this association is modeled by a β coefficient in regression, or amount of increase in R^2 in a regression tree associated with a predictor, etc...

If β is *significantly different* from zero (or amount of R^2 is greater than by chance alone), then there is evidence that the response is associated with the predictor.

How can we determine if β is *significantly different* from zero in a model?

Association vs. Causation (cont.)

But what can we say about a *causal association*? That is, can we manipulate *X* in order to influence *Y*?

Not necessarily. Why not?

There is potential for confounding factors to be the driving force for the observed association.

Controlling for confounding

How can we fix this issue of confounding variables?

There are 2 main approaches:

1. Model all possible confounders by including them into the model (multiple regression, for example). Or use fancy methods ('causal methods') to account for the confounders.
2. An *experiment* can be performed where the scientist manipulates the levels of the predictor (now called the *treatment*) to see how this leads to changes in values of the response.

What are the advantages and disadvantages of each approach?

Controlling for confounding: advantages/disadvantages

1. Modeling the confounders

- Advantages: cheap
- Disadvantages: not all confounders may be measured.

2. Performing an experiment

- Advantages: confounders will be *balanced*, on average, across treatment groups
- Disadvantages: expensive, can be an artificial environment

Experiments and *AB*-testing

Completely Randomized Design

There are many ways to design an experiment, depending on the number of treatment types, number of treatment groups, how the treatment effect may vary across subgroups, etc...

The simplest type of experiment is called a Completely Randomized Design (CRD). If two treatments, call them treatment *A* and treatment *B*, are to be compared across *n* subjects, then *n*/2 subject are randomly assigned to each group.

- If $n = 100$, this is equivalent to putting all 100 names in a hat, and pulling 50 names out and assigning them to treatment *A*.

Experiments and *AB*-testing

In the world of Data Science, performing experiments to determine causation, like the completely randomized design, is called *AB*-testing.

AB-testing is often used in the tech industry to determine which form of website design (the treatment) leads to more ad clicks, purchases, etc... (the response). Or to determine the effect of a new app rollout (treatment) on revenue or usage (the response).

Assigning subject to treatments

In order to balance confounders, the subjects must be properly randomly assigned to the treatment groups, and sufficient enough sample sizes need to be used.

For a CRD with 2 treatment arms, how can this randomization be performed via a computer?

You can just sample $n/2$ numbers from the values 1, 2, ..., n without replacement and assign those individuals (in a list) to treatment group *A*, and the rest to treatments group *B*. This is equivalent to sorting the list of numbers, with the first half going to treatment *A* and the rest going to treatment *B*.

This is just like a 50-50 test-train split!

Beyond just A vs. B

How can an AB test be expanded to include more than two options? What if there are more than just one type of treatment?

The **multivariate experimental design** generalizes this approach. If there are two treatment types (font color, and website layout), then both treatments' effects can (and should) be tested simultaneously. Why?

In a **full factorial experimental** design, each and every combination of treatments are considered different treatment groups. Experiments online are cheap. Full factorial designs are often possible and feasible.

Better than AB…

***Optional: *t*-tests, binomial *z*-test, F-tests, χ^2 tests, oh my!

Analyzing the results

Just like in statistical/machine learning, the analysis of results for any experiment depends on the form of the response variable (categorical vs. quantitative), but also depends on the design of the experiment.

For *AB*-testing (classically called a 2-arm CRD), this ends up just being a 2-group comparison procedure, and depends on the form of the response variable (aka, if *Y* is binary, categorical, or quantitative).

Analyzing the results (cont.)

For those of you who have taken Stat 100/101/102/104/111/139:

If the response is quantitative, what is the classical approach to determining if the means are different in 2 independent groups?

- a 2-sample *t*-test for means

If the proportions of successes are different in 2 independent groups?

- a 2-sample *z*-test for proportions

2-sample *t*-test

Formally, the 2-sample *t*-test for the mean difference between 2 treatment groups is:

$H_0: \mu_A = \mu_B$ vs. $H_A: \mu_A \neq \mu_B$

$$t = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{S_A^2}{n_A} + \frac{S_B^2}{n_B}}}$$

The *p*-value can then be calculated based on a $t_{\min(n_A, n_B)-1}$ distribution.

The assumptions for this test include (i) independent observations and (ii) normally distributed responses within each group (or sufficiently large sample size).

2-sample *z*-test for proportions

Formally, the 2-sample z test for the difference in proportions between 2 treatment groups is:

$H_0: p_A = p_B$ vs. $H_A: p_A \neq p_B$

$$z = \frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}_p(1-\hat{p}_p)\frac{1}{n_A} + \frac{1}{n_B}}}$$

where $\hat{p}_p = \frac{n_A\hat{p}_A + n_B\hat{p}_B}{n_A + n_B}$ is the overall 'pooled' proportion of successes.

The *p*-value can then be calculated based on a standard normal distribution.

Normal approximation to the binomial

The use of the standard normal here is based on the fact that the binomial distribution can be approximated by a normal, which is reliable when $np \geq 10$ and $n(1 - p) \geq 10$.

What is a Binomial distribution? Why can it be approximated well with a Normal distribution?

Summary of analyses for CRD Experiments

Variable Type	# Trt's	Classic Approach	Alternative Approach
Quantitative	2	t-test	Randomization test
	3+	ANOVA	
Binary	2	z-test	Fisher's exact test
	3+	χ^2 test	
Categorical (3+)	2+	χ^2 test	Fisher's exact test

The classical approaches are typically *parametric*, based on some underlying distributional assumptions of the individual data, and work well for large *n* (or if those assumptions are actually true). The alternative approaches are *nonparameteric* in that there is no assumptions of an underlying distribution, but they have slightly less power if assumptions are true and may take more time & care to calculate.

Analyses for CRD Experiments in Python

- *t*-test:
  `scipy.stats.ttest_ind`
- proportion *z*-test:
  `statsmodels.stats.proportion.proportions_ztest`
- ANOVA *F*-test:
  `scipy.stats.f_oneway`
- χ^2 test for independence:
  `scipy.stats.chi2_contingency`
- Fisher's exact test:
  `scipy.stats.fisher_exact`
- Randomization test: ???

ANOVA procedure

The classic approach to compare 3+ means is through the Analysis of Variance procedure (aka, ANOVA).

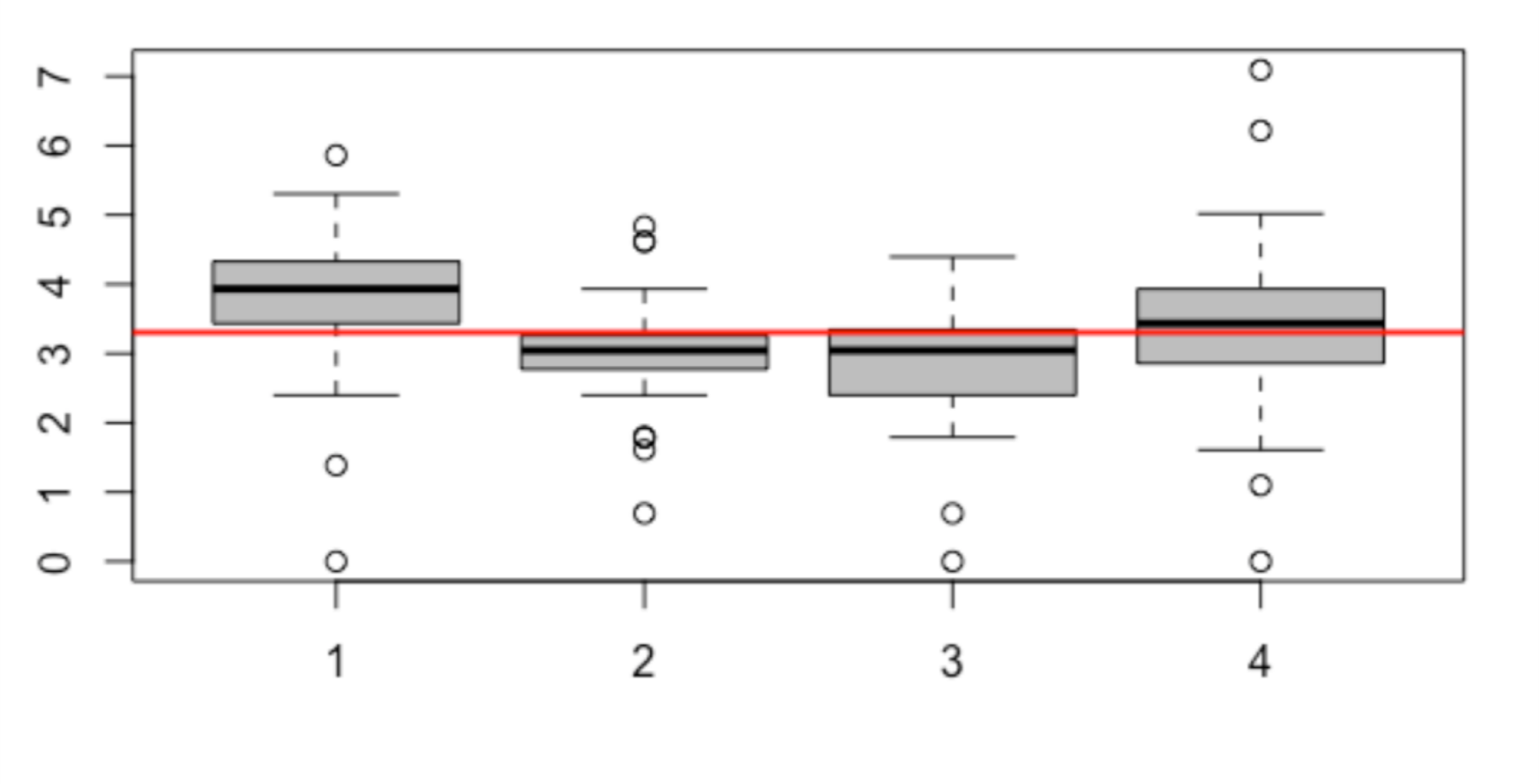
The ANOVA procedure's *F*-test is based on the decomposition of sums of squares in the response variable (which we have indirectly used before when calculating R^2).

$$SST = SSM + SSE$$

In this multi-group problem, it boils down to comparing how far the group means are from the overall grand mean (*SSM*) in comparison to how spread out the observations are from their respective group means (*SSE*).

A picture is worth a thousand words...

Boxplot to illustrate ANOVA



ANOVA *F*-test

Formally, the ANOVA F test for differences in means among 3+ groups can be calculated as follows:

H_0: the mean response is equal in all K treatment groups.
H_A: there is a difference in mean response somewhere among the treatment group.

$$F = \frac{\sum_{k=1}^{K} \frac{n_k(\bar{Y}_k - \bar{Y})^2}{(K-1)}}{\sum_{k=1}^{K} \frac{(n_k - 1)S_k^2}{(n-K)}}$$

where n_k is the sample size in treatment group k, $\bar{Y}_k$ is the mean response in treatment group k, S_k^2 is the variance of responses in treatment group k, $\bar{Y}$ is the overall mean response, and $n = \sum n_k$ is the total sample size.

The p-value can then be calculated based on a $F_{df_1=(K-1),df_2=(n-K)}$ distribution.

Comparing categorical variables

The classic approach to see if a categorical response variable is different between 2 or more groups is the χ^2 test for independence. A contingency table (we called it a confusion matrix) illustrates the idea:

Abortion Should be	Republican	Democrat	total
Legal	166	430	596
Illegal	366	345	711
Total	532	775	1307

If the two variables were independent, then:
$P(Y = 1 \cap X = 1) = P(Y = 1)P(X = 1)$.

How far the inner cell counts are from what they are expected to be under this condition is the basis for the test.

χ^2 test for independence

Formally, the χ^2 test for independence can be calculated as follows:

H_0: the 2 categorical variables are independent

H_A: the 2 categorical variables are not independent (response depends on the treatment).

$$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp}$$

where *Obs* is the observed cell count and *Exp* is the expected cell count:

$Exp = \frac{(row\ total) \times (column\ total)}{n}$.

The *p*-value can then be calculated based on a $\chi^2_{df=(r-1)\times(c-1)}$ distribution (*r* is the # categories for the row var., *c* is the # categories for the column var.).

Fisher's exact test

R.A. Fisher also came up with what is known as Fisher's exact test.

This analysis approach is useful for a contingency table, and does not need to rely on large sample size.

It fixes the row and column totals, and then determines all the ways in which the inner cells can be calculated given those row and column totals.

The probability of any of these filled out tables, given the row and column totals is fixed, is then based on a hypergeometric distribution.

Then the possible filled out tables that are less likely to occur than what was actually observed contribute to the *p*-value (by adding up their probabilities).

Fisher's exact test

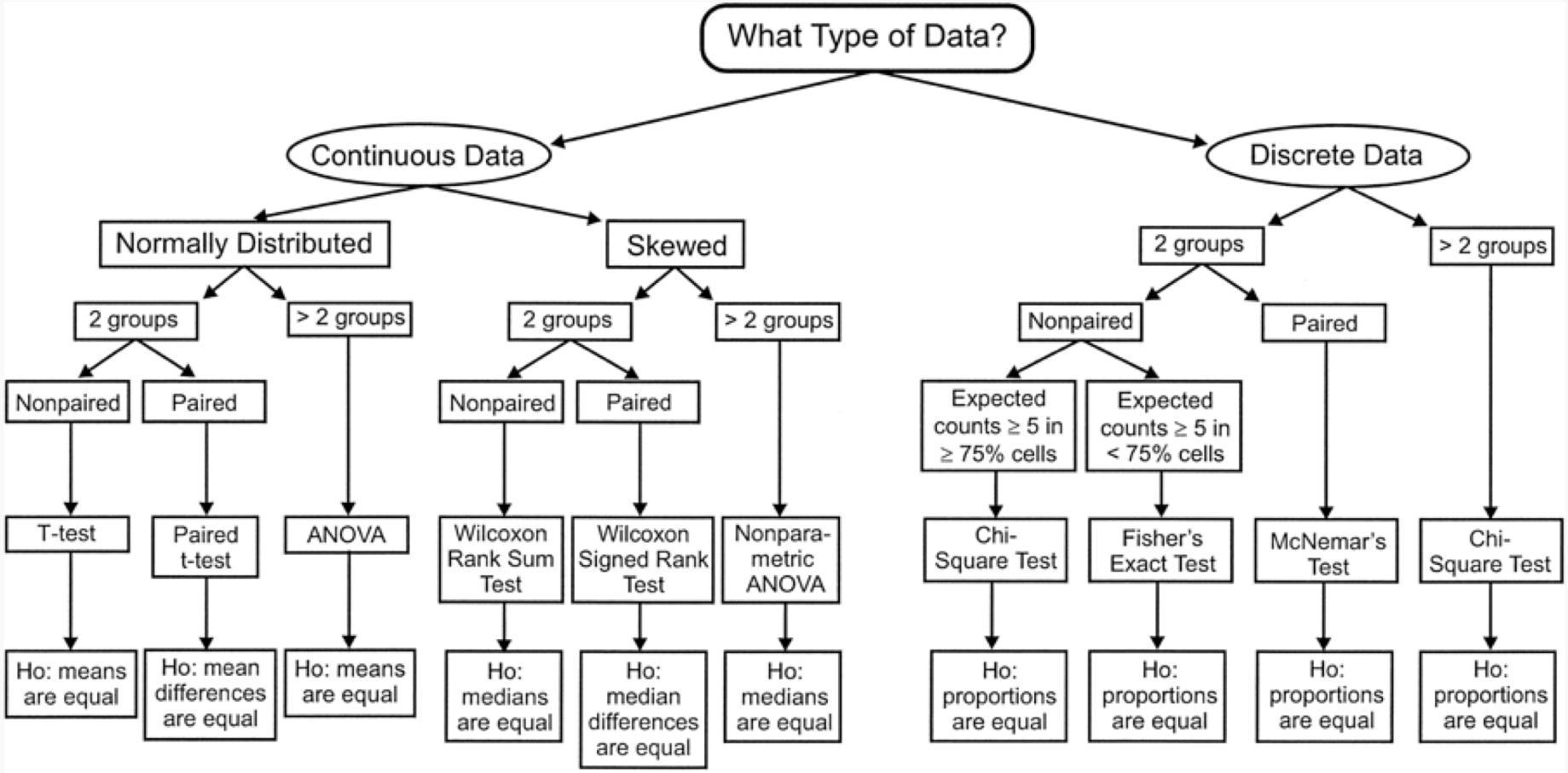
Abortion Should be	Republican	Democrat	total
Legal	166	430	596
Illegal	366	345	711
Total	532	775	1307

$$P(X_1 = 166) = \frac{\binom{596}{166}\binom{711}{366}}{\binom{1307}{532}} = 1.33\times10^{-18}$$

Then a similar calculation is done for all possible values of X_1, and these probabilities are summed up for those cases of X_1 that are not more likely to occur.

A Decision Tree for testing.

Inference



Source: Waning B, Montagne M: *Pharmacoepidemiology: Principles and Practice*: http://www.accesspharmacy.com

Alternative Method: Randomization tests

Randomization test

A randomization test is the non-parametric approach to analyzing quantitative data in an experiment. It is an example of a *resampling* approach (the bootstrap is another resampling approach).

The basic assumption of the randomization test is that if the treatments are truly the same, then the measured response variable, Y_i, for subject *i* would not change if that subject was instead randomly assigned to a different treatment. This is sometimes called *exchangeability*.

Randomization test (cont.)

So to analyze the results, we re-randomize the individuals to treatment through simulation (keeping the sample sizes the same). We then re-calculate the statistic of interest (difference in 2 sample means or sums of squares between 3+ groups) many-many times and build a histogram of the results. This histogram is then used as the reference distribution to determine how extreme our actual observed result is.

This approach is also called a permutation test, since we are re-permuting each of the subjects into the treatment groups (and then assume this has no bearing on the response).

Example #1: The app update roll-out problem

A company is interested in updating their app/program, so they start a 'pilot program' to test the waters to see how this update will affect some important measure (like revenue or usage). How should they do this?

They select a sample of users and ask them to voluntarily update the app on their phones in order to estimate the affect of this update.

Any issues with this design?

Volunteers will always be the most excited, dedicated users: a biased sample from all of their users.

We can potentially check for this bias via a χ^2 test for goodness-of-fit.

χ^2 test for goodness-of-fit

Formally, the χ^2 test for goodness-of-fit can be calculated as follows:

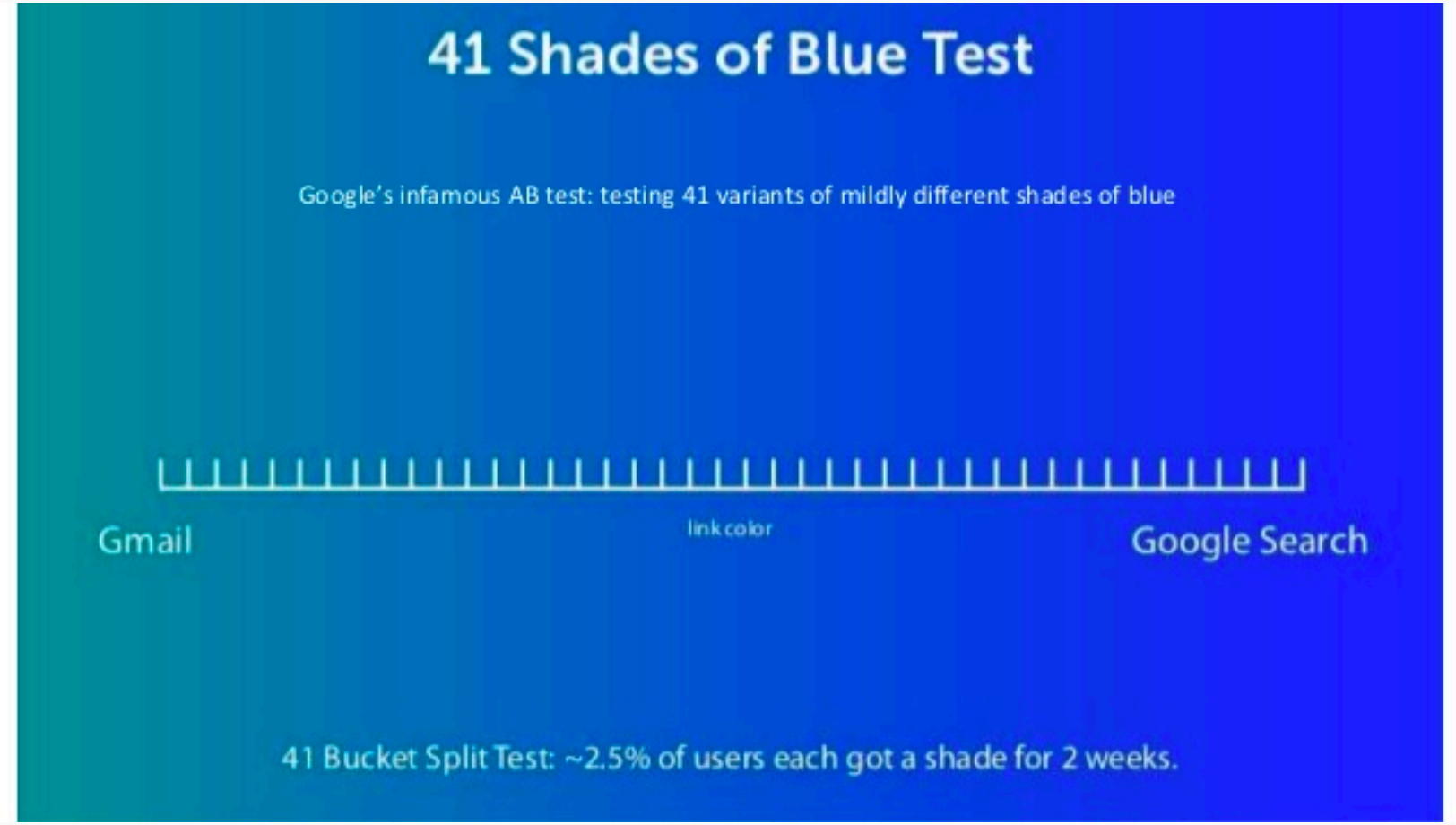
H_0: the variable follows some known distribution in the population
H_A: the variable does not follow this distribution

$$\chi^2 = \sum_{all\ cells} \frac{(Obs - Exp)^2}{Exp}$$

where *Obs* is the observed cell count and *Exp* is the expected cell count: $Exp_i = n\pi_i$ (π_i is the theoretical probability of being in category/bucket *i*).

The *p*-value can then be calculated based on a $\chi^2_{df=(k-1)}$ distribution (*k* is the # categories in the population).

Example #2: An infamous AB Test, 41 Shades of Blue



How should the study proceed? How should the data be analyzed?

Obama's 2008 Campaign

The 2008 Obama Campaign

In 2008, the Obama campaign raised much of its money via online donations through its website.

They wanted to optimize the launch page that visitors saw when they came to the campaign website. They were attempting to maximize the number of visitors that would sign up for their emailing list.

There were 2 treatments they attempted to vary:

- the image or video the user saw.
- the words on the click-through button.

Media choices (6 of them):

- 3 images and 3 videos were possibly shown

Click- through button

- one of 4 choices:

How to design the experiment?

How should this experiment unfold?

1. What was the response variable?
2. What were the treatments? What were the treatment groups?
3. How many observations (sample size) needed to be selected in order to determine which treatment group is most effective?
4. What analysis should be performed?

Obama 2008: the specifics

1. What was the response variable?
   - sign-up rate
2. What were the treatments? What were the treatment groups?
   - 2 treatments (media type and button). 24 treatment groups
3. How many observations (sample size) needed to be selected in order to determine which treatment group is most effective?
   - The campaign decided to run the experiment on 310,382 visitors!!!
4. What analysis should be performed?
   - Classically, a χ^2 test for independence could be performed. Or better yet, a randomization test ☺

The data were overwhelming…

The Results

The results are shown to the right (note: they are from a 3rd party site that runs AB tests for website design: Optimizely).

https://blog.optimizely.com/2010/11/29/how-obama-raised-60-million-by-running-a-simple-experiment/

Combinations (24) | Page Sections (2) — Download: XML CSV TSV | Print

Relevance Rating	Variation	Est. conv. rate	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Button	Original	7.51% ± 0.2%	—	—	5851 / 77858
5 / 5	Learn More	8.91% ± 0.2%	100%	18.6%	6927 / 77729
	Join Us Now	7.62% ± 0.2%	73.5%	1.37%	5915 / 77644
	Sign Up Now	7.34% ± 0.2%	13.7%	-2.38%	5660 / 77151
Media	Original	8.54% ± 0.2%	—	—	4425 / 51794
5 / 5	Family Image	9.66% ± 0.2%	100%	13.1%	4996 / 51696
	Change Image	8.87% ± 0.2%	92.2%	3.85%	4595 / 51790
	Barack's Video	7.76% ± 0.2%	0.04%	-9.14%	3992 / 51427
	Sam's Video	6.29% ± 0.2%	0.00%	-26.4%	3261 / 51864
	Springfield Video	5.95% ± 0.2%	0.00%	-30.3%	3084 / 51811

Combinations (24) | Page Sections (2) — Download: XML CSV TSV | Print

Disable | All Combinations (24) | Key: Winner Inconclusive Loser

Combination	Status	Est. conv. rate	Chance to Beat Orig.	Observed Improvement	Conv./Visitors
Original	Enabled	8.26% ± 0.5%	—	—	1088 / 13167
☆ Top high-confidence winners. Run a follow-up experiment »					
Combination 11	Enabled	11.6% ± 0.6%	100%	40.6%	1504 / 12947
Combination 7	Enabled	10.3% ± 0.6%	100%	24.0%	1340 / 13073
Combination 3	Enabled	9.80% ± 0.6%	99.7%	18.7%	1277 / 13025
Combination 10	Enabled	9.23% ± 0.6%	95.9%	11.7%	1203 / 13031
Combination 8	Enabled	9.03% ± 0.6%	91.6%	9.28%	1178 / 13046
Combination 9	Enabled	8.77% ± 0.6%	81.8%	6.10%	1111 / 12672
Combination 6	Enabled	8.64% ± 0.5%	75.3%	4.58%	1108 / 12822

The winning variation had a sign-up rate of 11.6%. The original page had a sign-up rate of 8.26%. That's an improvement of 40.6% in sign-up rate…[leading to an] additional 2,880,000 email addresses translated into 288,000 more volunteers…and an additional $60 million in donations.

See any issue in this conclusion?

But more importantly, they did learn one lesson: those intimately involved in designing websites (or medical treatments) are too biased to properly make conclusions as to what works best. They like the videos, and they performed the worst.

And the winner was: