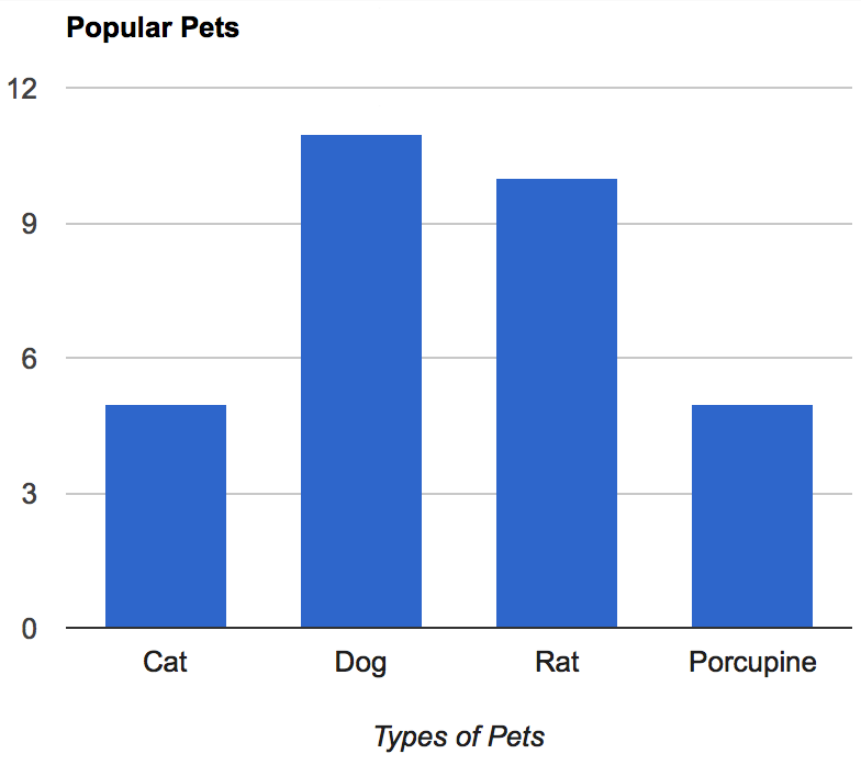
BACK TO THE BASICS

After six rigorous ~~weeks~~ of training in data science, we should all be able to answer the following with ease.

1. When you are asked "**what is a typical *x* value?**", you compute...
2. When you are asked "**what is the spread of *x* values?**", you compute...
3. When you are asked "**how good is your model?**", you compute...
4. When you are asked "**which predictors are most significant?**", you compute...

THE PROBLEM WITH "AVERAGE"

for samples of categorical variables, neither mean or median make sense.



The **mode** might be a better way to find the most "representative" value.

THE PROBLEM WITH "VARIANCE"

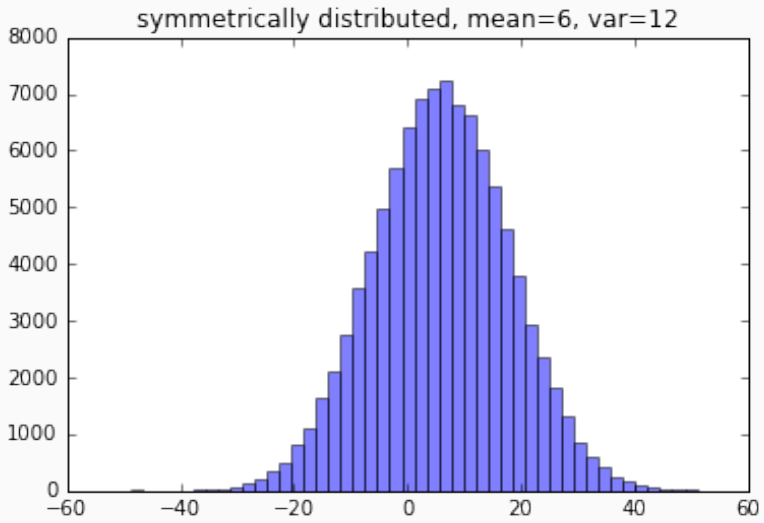
Given two sets of observations, X_1 and X_2, suppose I tell you that

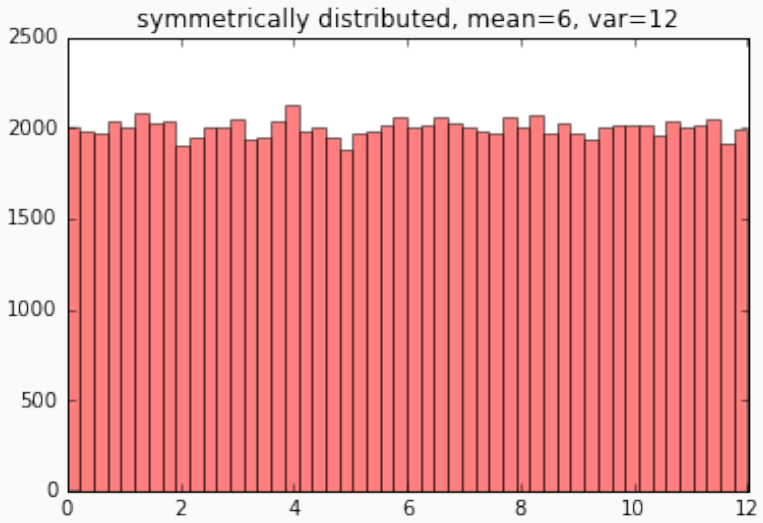
1. $\mathrm{Var}(\mathsf{X}_1) = \mathrm{Var}(\mathsf{X}_2)$
2. $\overline{\mathsf{X}_1} = \overline{\mathsf{X}_2}$
3. X_1 and X_2 are symmetrically distributed about their means

What can you conclude about the shapes of the distributions of these data sets?

THE PROBLEM WITH "VARIANCE"

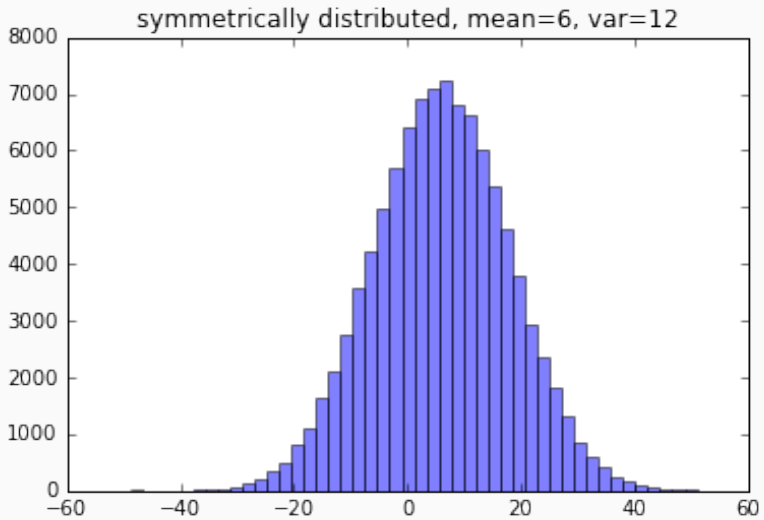
What can you conclude about the shapes of the distributions of these data sets?

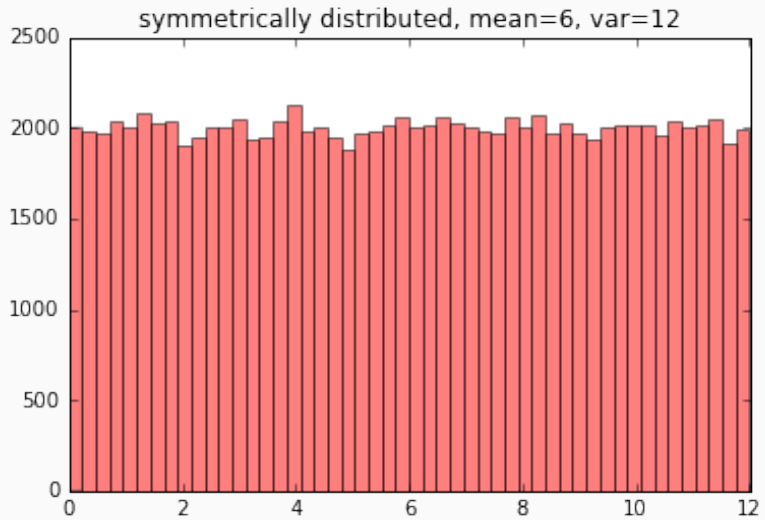




THE PROBLEM WITH "VARIANCE"

What can you conclude about the shapes of the distributions of these data sets?





Observation: Variance isn't the final word on the "spread" of the data.

THE PROBLEM WITH MODEL "GOODNESS"

"goodness" we should choose to evaluate a model:

1. Only look at the largest error the model makes on a data set $\mathbf{X}, \mathbf{Y}$
2. Take the sum of error the model makes on a data set $\mathbf{X}, \mathbf{Y}$
3. Take the sum of squared errors the model makes on a data set $\mathbf{X}, \mathbf{Y}$
4. Compute R^2

Question: Which notion or overall "goodness" is generally better? Is there a notion that is generally better?

THE PROBLEM WITH MODEL "GOODNESS"

When we built *probabilistic models* for the observed data (linear mode plus i.i.d. normal noise), we felt compelled to minimize sum of squared errors (MLE).

When we built *probabilistic models* for the observed data and prior beliefs about the model, we found it obvious to minimize the ridge regression loss function (MAP).

Question: Does this mean that we didn't make any choices in our modeling process?

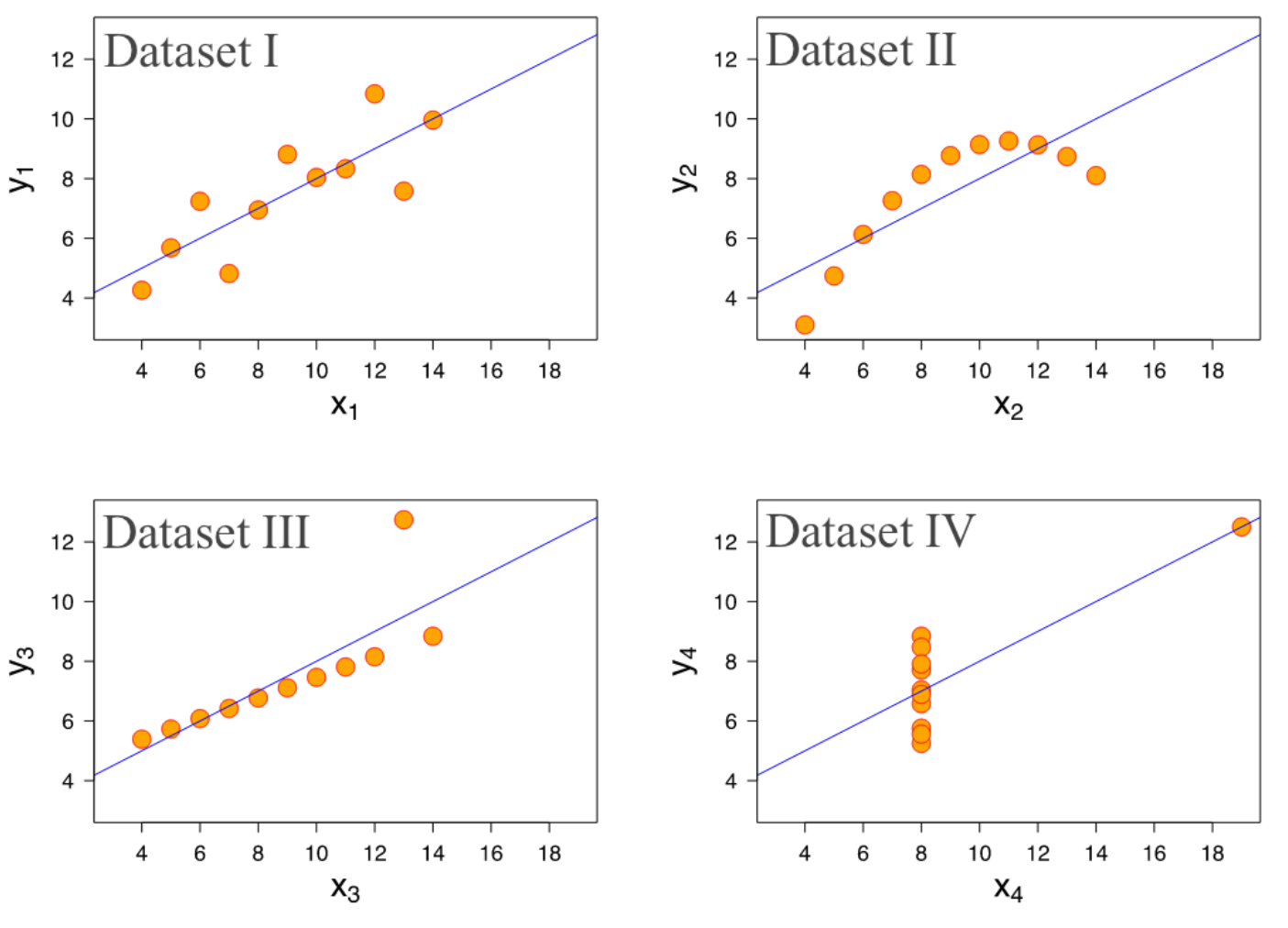
THE PROBLEM WITH R^2

The following data sets comprise the Anscombe's Quartet; which model fits the data the best?

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
Reg. Line:	$y = 3 + 0.5x$		$y = 3 + 0.5x$		$y = 3 + 0.5x$		$y = 3 + 0.5x$	
R^2 :		0.816		0.816		0.816		0.816

THE PROBLEM WITH R^2

The following data sets comprise the Anscombe's Quartet; which model fits the data the best?



THE PROBLEM WITH "SIGNIFICANCE"

Question: What's a p-value again?

The p-value for the value of particular stat $T = t$ is the probability of observing a "similarly extreme" value $T = t$, assuming some (null) model, M_0:

$$p(T \geq t | M_0) \text{ or } p(T \leq t | M_0) \text{ or combo}$$

If the p-value is small, we tend to think that observing a value "similar to" $T = t$ is unlikely assuming the model M_0, and consider this as evidence for **rejecting** the null model.

THE PROBLEM WITH "SIGNIFICANCE"

Question: Why use 0.05 or 0.01 as the cut off for p-values?

For decades, the convention have been to interpret $p < .05$ as "significant", and $p < .01$ as "highly significant.

But when p-values where introduced by Sir Ronald Fisher in 1925, he adopted .05 as a **reference point** for rejecting a null hypothesis. But 0.5 was not a sharp cutoff and should be considered in the context of other results and tests.

THE PROBLEM WITH "SIGNIFICANCE"

Question: Do significant p-values indicate the presence of an effect (a real relationship between predictor and response)?

Example

Since the p-value variable selection method **only gathers evidence against the null hypothesis**, the more tests you do, the higher the likelihood of falsely rejecting the null hypothesis.

When you have a large number of predictors, performing step-wise selection using p-values may result in predictor subsets that are not truly significant.

Question: Do significant p-values indicate the significance of an effect (does a "statistically" significant effect indicate a meaningful relationship)?

Example

Our regression model is:

$$\textbf{Price of House} \text{ (in \$)} = \underbrace{10,000}_{\beta_2} * \textbf{Area} \text{ (in SqFt)} + \underbrace{0.00001}_{\beta_1} * \textbf{Number of Blue Bathroom Tiles} + \underbrace{10,000}_{\beta_0}$$

The p-values for the coefficients are:

	β_2	β_1	β_0
p-value	0.0005	0.001	0.0001

Which predictors is are significant?

THE PROBLEM WITH "SIGNIFICANCE"

"Ultimately the problem is not with *p*-values but with null-hypothesis significance testing, that parody of falsificationism in which straw-man null hypothesis A is rejected and this is taken as evidence in favor of preferred alternative B.

Whenever this sort of reasoning is being done, the problems discussed above will arise. Confidence intervals, credible intervals, Bayes factors, cross-validation: you name the method, it can and will be twisted, even if inadvertently, to create the appearance of strong evidence where none exists."

- Andrew Gelman

(comments on the ASA statement regard the misuse of *p*-values)

THE PROBLEM WITH CHOICE

The problem problem with commanding an arsenal of sophisticated tools and theory is that it's easy to make choices based on habit, convenience, or intimidation/glamour factor. Or you're overwhelmed and are unable to make a choice.

Tips for staying grounded:

1. Do what works (for the task, for the context of the problem at hand, for time/resources available)
2. Be absolutely honest and vigilant about our assumptions and choices (justify them and identify their draw backs)
3. Be accountable to the data, to the problem (not to our favorite technique or pet theory)