Lecture 12: Statistical Error and Model Selection (Cross Validation)

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

11/07/2017

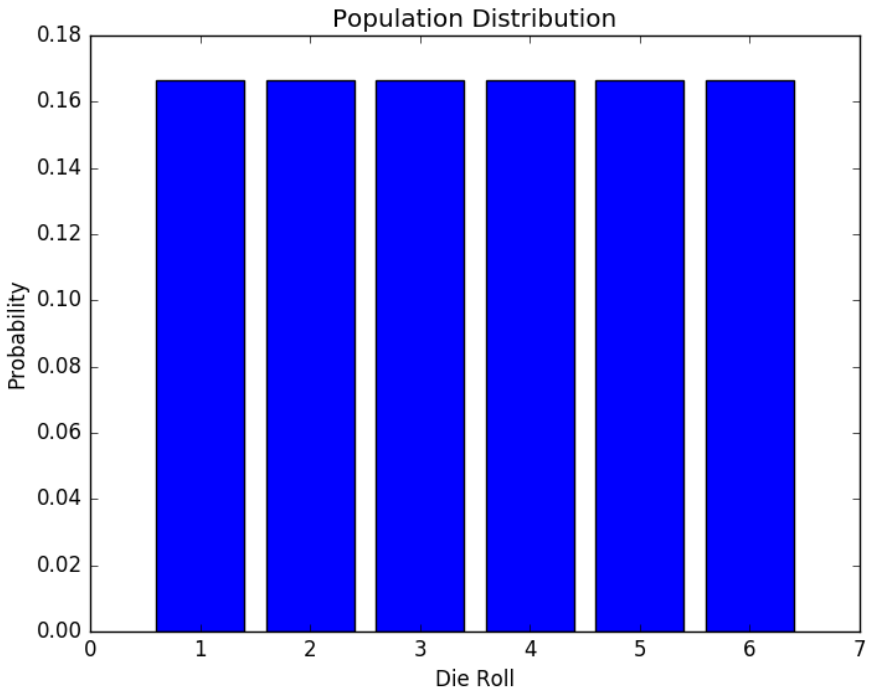
Overview

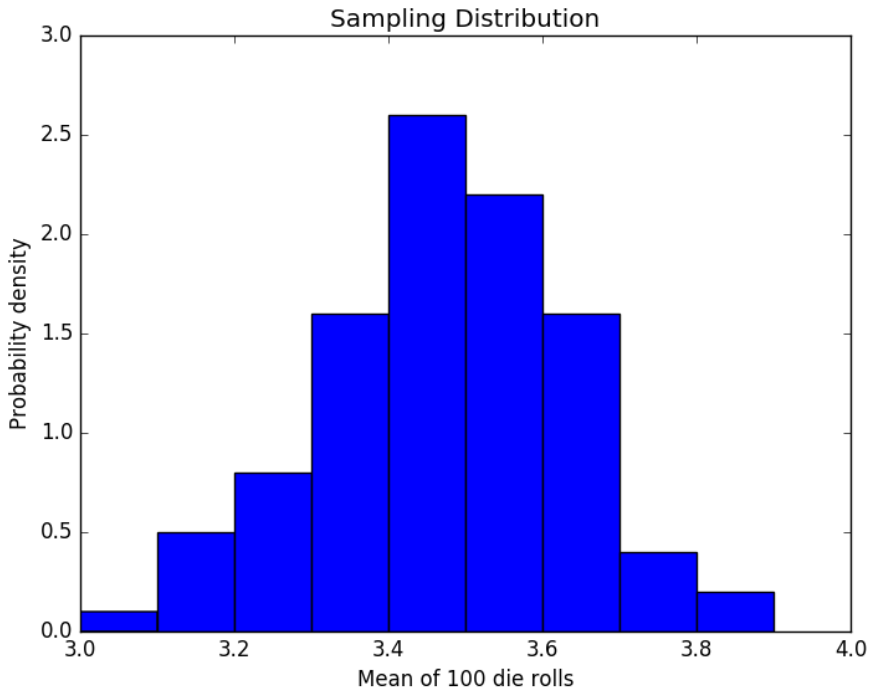
Sampling

- A **parameter** is a number that describes a population (e.g. μ and σ in normal distribution.) It is impossible to know without measuring the whole population.
- A **sample statistic** is a number computed from a sample. We use this as a **point estimate** for the unknown population parameter of interest.
- Statistical inference provides a way to estimate the population parameter from the sample statistics and characterize the uncertainty.
- Quantifying the variability of sample statistics from one sample to another is how we estimate the **margin of error** associated with the point estimate.

Sampling distribution

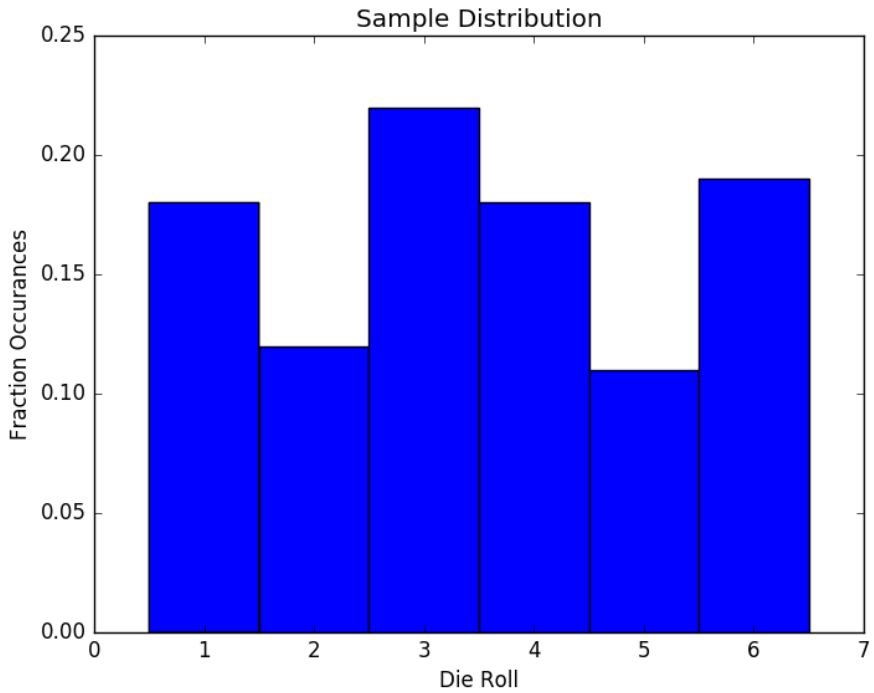
A **sampling distribution** is the distribution of a sample statistic based on a random sample.

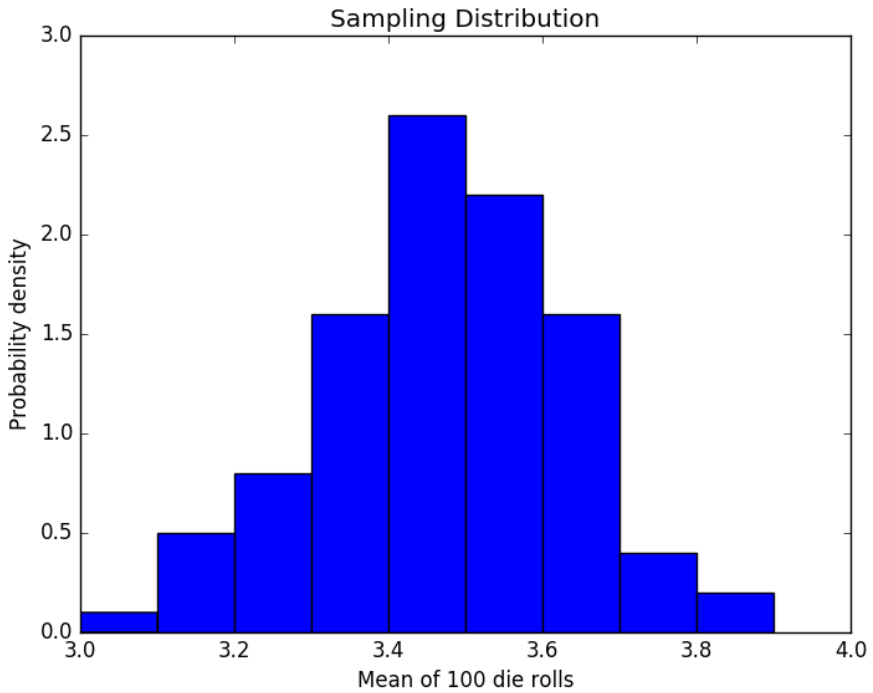




Sample distribution

While the *sampling distribution* is the distribution of a *sample statistic*, the **sample distribution** is the distribution of the sample. Note the difference. The sample distribution shown below is the distribution of values on 100 simulated die rolls. The sampling distribution is the distribution of 100 means of 100 die rolls each.





Introduction to Inference

Make a statement about something that is *not observed*, and characterize uncertainty about that statement. Before making an inference:

1. Identify and describe the population.
2. Describe the sampling process.
3. Describe a model for the population, complete with assumptions.

Example: A simple linear model

$$y = \beta_0 + \beta_1 x + \epsilon$$

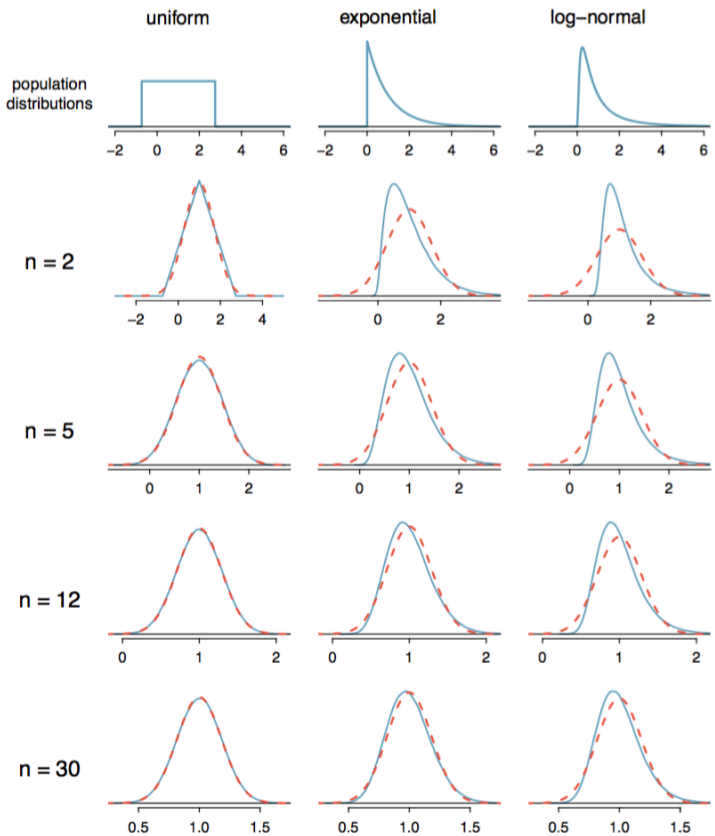
x, y are features of population; β_0 , β_1 describe the relationship,

ϵ is random, making this a statistical model

Central Limit Theorem

Central Limit Theorem

The mean of a large number (> 30) of independent, identically distributed variables will be approximately normal, for all underlying distributions.



Graphic from [Diez, 2016]

Central Limit Theorem Applied to Sample Mean

Normal Model of Sample Mean

The distribution of the sample mean is approximated by the normal distribution centered about the population mean (μ) with standard deviation equal to the **standard error**: $SE = \frac{\sigma}{\sqrt{n}}$

Conditions

1. **Independence:** Sampled observations are independent. Necessary but not sufficient checks:
   - Random sampling/assignment
   - For sampling without replacement, $n < 10\%$ of the population
2. **Sample size/skew:** The limit of "large" n depends on the underlying distribution. For moderately skewed distributions, $n > 30$ is widely used to be large enough to use the CLT. To check the skew, the sample distribution is assumed to mirror the population distribution.

Standard Error

Standard error of an estimate

The standard deviation associated with an estimate. It describes the uncertainty associated with the estimate.

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is:

$$SE = \frac{\sigma}{\sqrt{n}}$$

Since we do not generally have the population standard deviation σ, we use the sample standard deviation s to estimate the standard error.

$$SE \approx \frac{s}{\sqrt{n}}$$

Statistical Inference

1. Determine which point estimate or test statistic is useful.
2. Identify an appropriate distribution for the point estimate or test statistic.
3. Create a confidence interval or hypothesis test using the chosen distribution.

Distributions

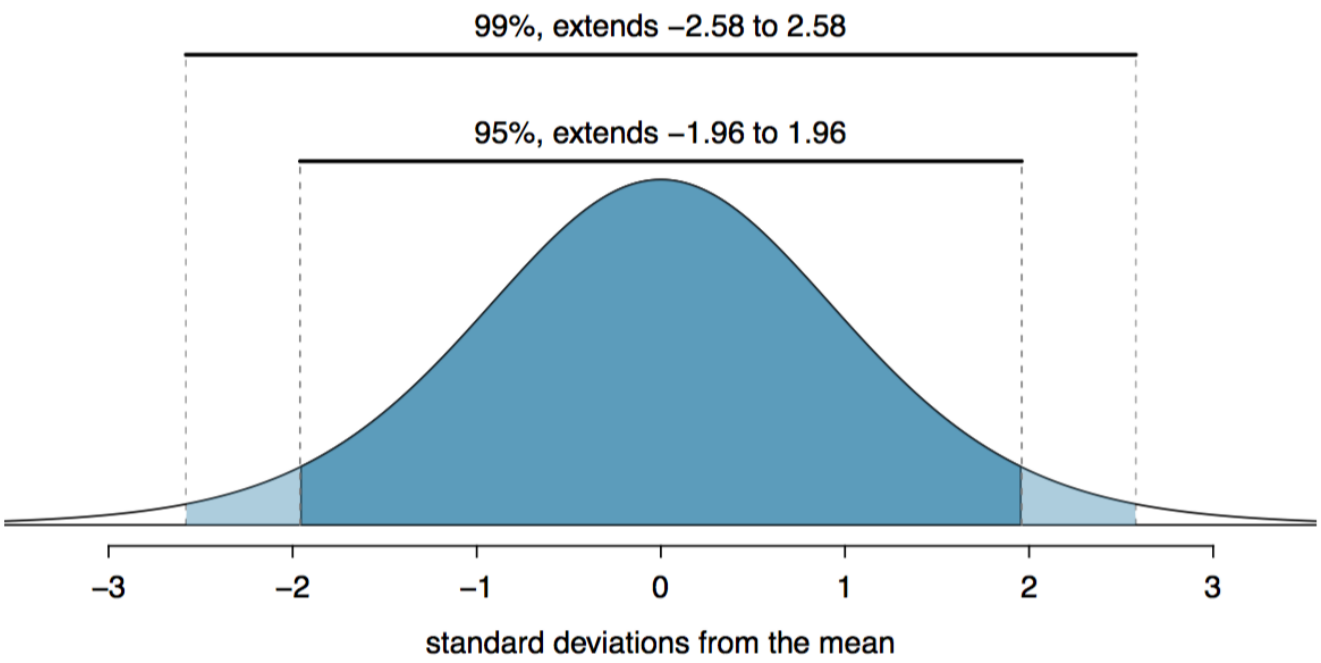
- Normal distribution: large sample, independent observations
- Student's t-distribution: small sample, independent observations, observations come from a nearly normal distribution
- F-distribution: Compare means of more than two groups using ANOVA
- χ^2 distribution: categorical data

Confidence Intervals

- A confidence interval gives a range of possible values of a **population parameter** with a given level of confidence that the parameter is in the range.
- To use the normal distribution in defining a confidence interval, the sample distribution must be nearly normal:
  - The sample observations are independent (a simple random sample consisting of under 10% of the population can be assumed to be independent).
  - The sample size is large (≥ 30 is a good rule of thumb).
  - The population distribution is note strongly skewed (the larger the sample size, the more skew is okay).
- In a confidence interval, $z^* \times SE$ is the **margin of error**.

Confidence Intervals

95% Confidence Interval: point estimate $\pm 1.96 \times SE$
99% Confidence Interval: point estimate $\pm 2.58 \times SE$
Generally, z^* chosen such that the area between $-z^*$ and z^* corresponds to the confidence level.



Graphic 4.10 in [Diez, 2016]

Hypothesis Testing

- Specify the null (H_0) and the alternate (H_A) hypothesis.
- Choose a sample.
- Assess the evidence.
- Draw conclusions.

p-value

p-value provides an estimate of how often the obtained result would occur by chance, if in fact the null hypothesis is true.

A result is statistically significant if it is unlikely to have occurred by chance alone.

Significance Level of a Test

1. The cut-off of what we consider to be "unlikely".
2. Commonly chosen to be $\alpha = 0.05$.
3. If p-value $< \alpha$, we reject the null hypothesis and accept the alternate hypothesis. If p-value $> \alpha$, we fail to reject the null hypothesis.

		Test Conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	okay	Type 1 Error
	H_A true	Type 2 Error	okay

Hypothesis Testing Example

A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a simple random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. The mean is 128 minutes. The following year, she repeats the test and finds that the average wait time is 135 minutes with standard deviation 39 minutes.

- Are conditions for inference met?
- What is the hypothesis of this situation?
- If we plan to collect a sample of size $n = 64$, what values could $\bar{x}$ take so that we reject H_o?
- What is the probability of a type 2 error?

Example from OpenIntro Stats, Second Ed., end of chapter 4 exercises 11, 23, 51.

Hypothesis Testing Example

Are conditions for inference met?

- **Independence**: The sample is random and 64 patients would almost certainly make up less than 10% of the ER residents.
- **Sample size/skew**: The sample size is at least 30. No information is provided about the skew. In practice, we would ask to see the data to check this condition, but here we will make the assumption that the skew is not very strong.

Hypothesis Testing Example

What is the hypothesis of this situation?
The null hypothesis is that the mean this year is the same as last year, 128 minutes. The alternative hypothesis is that the mean is different from 128 minutes.

$$H_0 : \bar{x} = 128$$

$$H_A : \bar{x} \neq 128$$

Hypothesis Testing Example

If we plan to collect a sample of size $n = 64$, what values could $\bar{x}$ take so that we reject H_0 at a significance level of 0.05?
Calculate the standard error assuming that the population standard deviation σ is equal to the sample standard deviation $s = 39$ minutes.

$$SE = \frac{s}{\sqrt{n}} = \frac{39}{\sqrt{64}} = 4.875$$

Next identify the Z-scores that would result in rejecting H_0:
$Z_{lower} = -1.96$, $Z_{upper} = 1.96$, so:

$$\bar{x}_{lower} = \bar{x}_{H_0} + Z_{lower} \times SE = 128 - 1.96 \times 4.87 = 118.445$$

$$\bar{x}_{upper} = \bar{x}_{H_0} + Z_{upper} \times SE = 128 + 1.96 \times 4.87 = 137.555$$

The null hypothesis will be rejected if $\bar{x} < 118.445$ or $\bar{x} > 137.555$.

Hypothesis Testing Example

What is the probability of a type 2 error?

I.e., what is the chance that we will not reject the null hypothesis, even if the average wait time really has changed?

Assume that the point estimate $\bar{x} = 135$ is the true population mean, rather than the null hypothesis. The probability of correctly rejecting the null hypothesis would be the total probability of all values outside the range $(\bar{x}_{lower}, \bar{x}_{upper}) = (118.455, 137.555)$.

$$Z_{lower} = \frac{118.445 - 135}{4.875} = -3.396$$

$$Z_{upper} = \frac{137.555 - 135}{4.875} = 0.524$$

$$CDF_{norm}(-3.396) = 0.0003 \quad 1 - CDF_{norm}(0.524) = 0.3001$$

The **power** is $0.0003 + 0.3001 = 0.3004$. The probability of a type 2 error is 1-power $= 0.6996$.

Exercise: Statistical Inference notebook

- Explore how parameter estimates vary with different samples
- Visualize confidence intervals to understand how to interpret them.

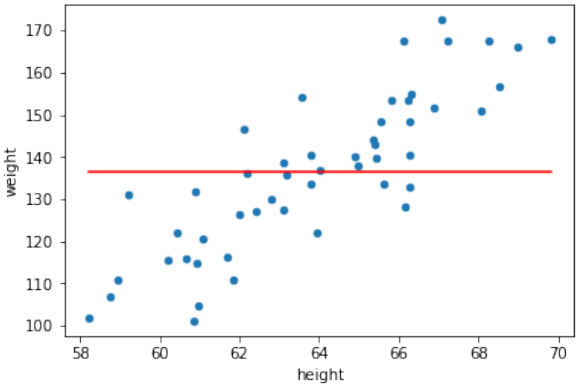
Model Selection and Validation

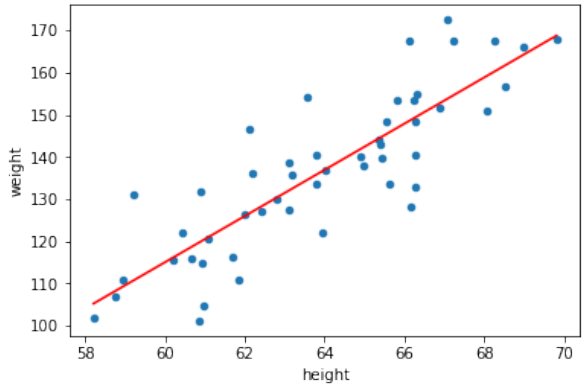
Two Goals

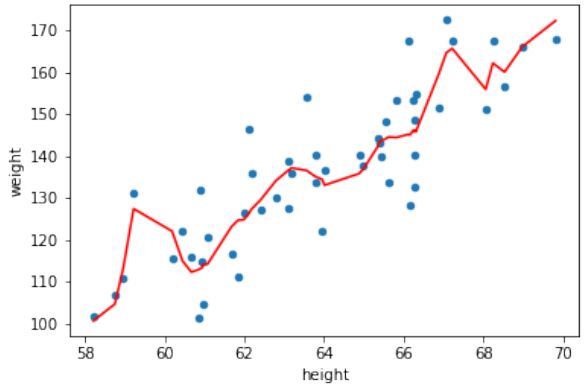
1. Select the best model
2. Assess model's predictive capabilities

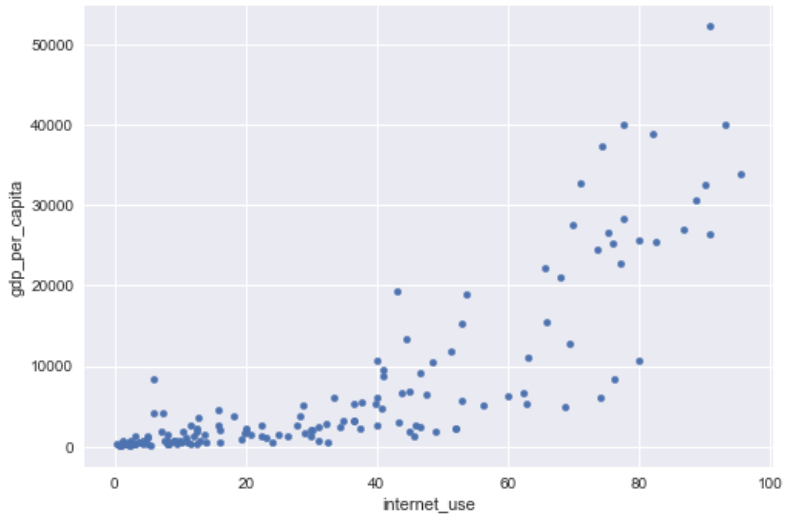
Imagine we had all the data. How would we reach these goals? What is different when we only have a small sample of data?

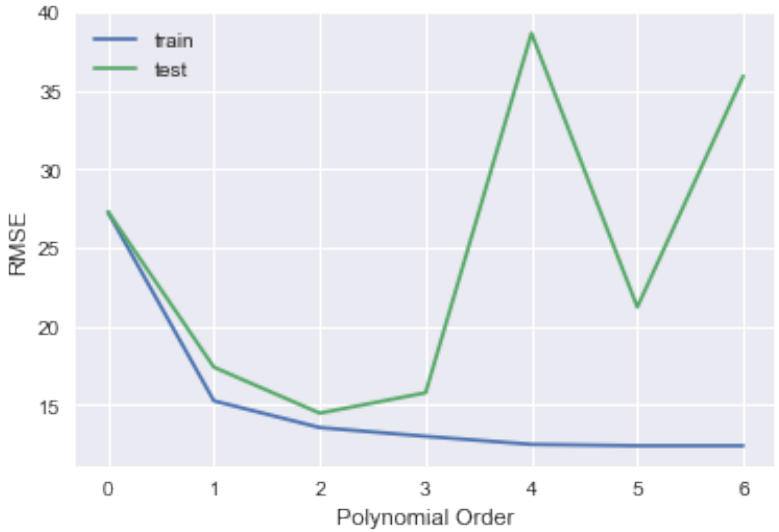
Overfitting











Cross Validation

- Three-way split
- K-fold
- Leave-one-out (LOOCV)
- Leave p-out

References

David Diez, Christopher Barr, & Mine Çetinkaya-Rundel (2015)

OpenIntro Statistics, OpenIntro

Trevor Hastie, Robert Tibshirani & Jerome Friedman (2008)

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Ed.

Recommended Reading

OpenIntro Statistics, Chapter 4
Data Science from Scratch, Chapter 7
Art of Data Science, Chapter 6

Cross-Validation in Machine Learning

Articles about *p*-values and *p*-hacking:

Statisticians Found One Thing They Can Agree On: Its Time To Stop Misusing P-Values
Statisticians issue warning over misuse of P values
I Fooled Millions into Thinking Chocolate Helps Weight Loss. Here's How.
You can't trust what you read about nutrition
Science Isn't Broken
Not Even Scientists Can Easily Explain P-values