349:Machine Learning

Fall 2024

Hypothesis Testing

How do you tell something is better?

Assume we have an error measure….

- How do we tell if it measures something useful?
- To measure intelligence, which is a better? {grades, IQ, salary}
- If it is useful, how precise/unbiased/noisy is it?
- How much of a difference in the measure is required to say things two things are truly “different”?
- Maria’s IQ is 103. Bob’s is 101. Does that make her “smarter”?

What's a useful measure for a…

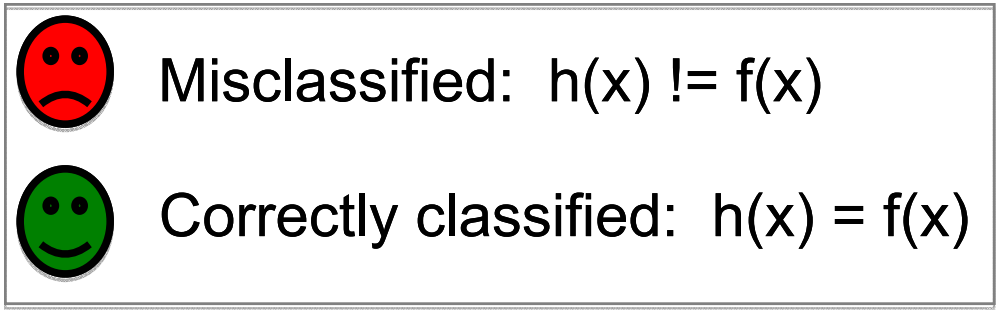
- Classifier?
  - How many predictions did we get correct?
- Regression?
  - Distance between predictions and labels
- What about…
  - Unsupervised clustering method?
  - Search engine results list?
  - Automated translation system?
  - Large language model chatbot?
  - etc.

Definitions of Error

- $error_D(h)$ is the *true error*
  - Hypothesis ***h*** versus true function ***f*** on data from distribution ***D***
  - Probability ***h*** misclassifies a random instance from ***D***

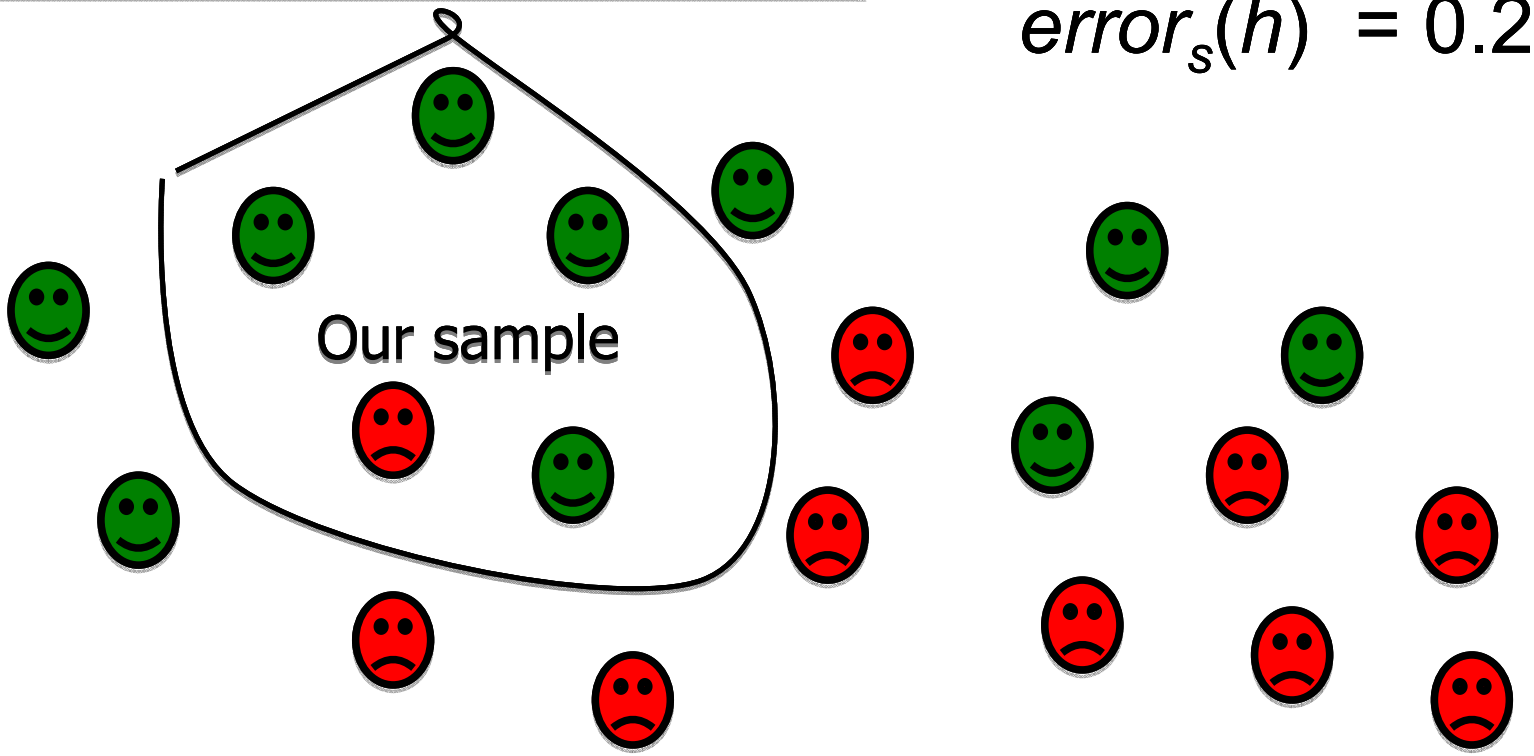
- $error_s(h)$ is the *sample error*
  - Hypothesis ***h*** versus true function ***f*** on data from sample ***S***
  - Proportion of examples in ***S*** that ***h*** misclassifies

True Error vs Sample Error



$error_D(h) = 0.5$

$error_s(h) = 0.2$



Sample Error: It's all we have

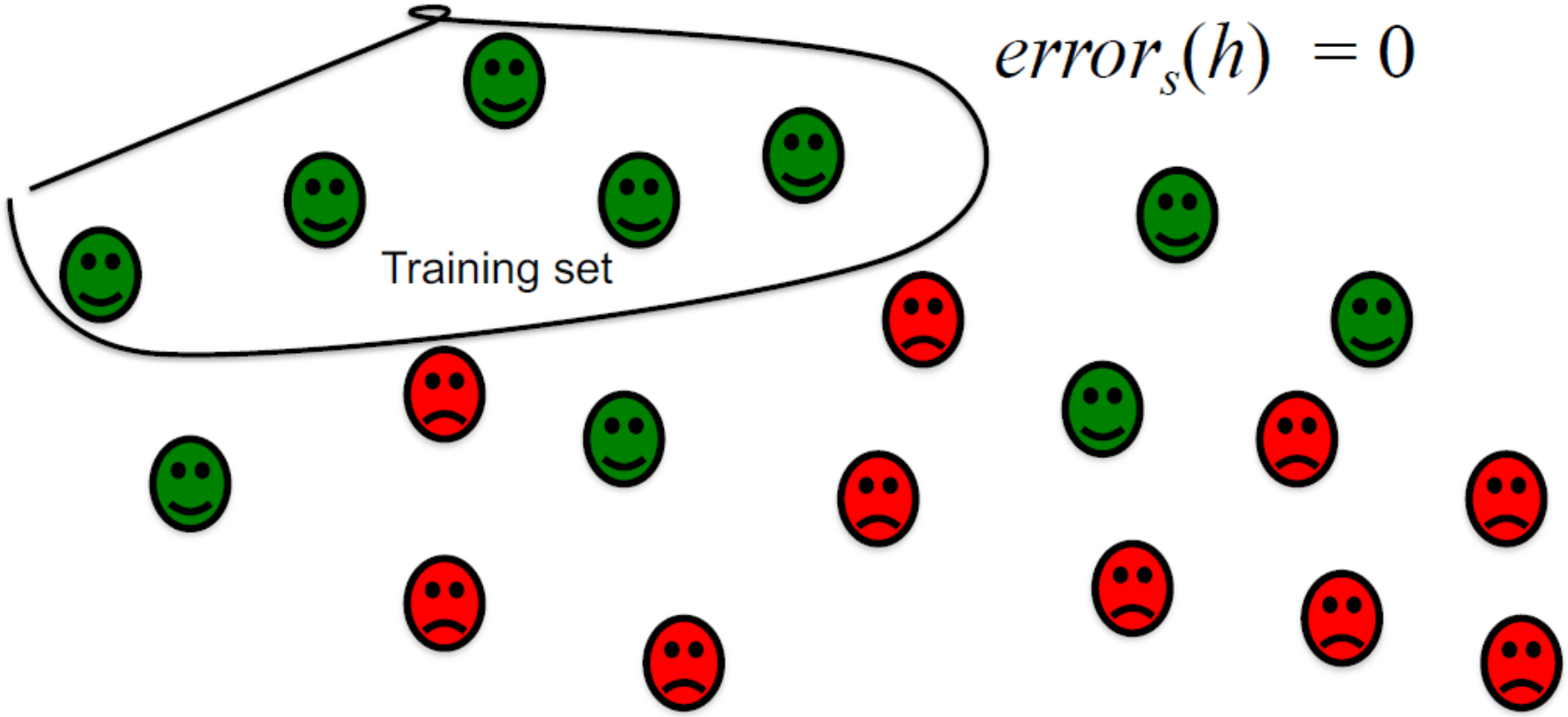
Generally, we never know the true error $error_D(h)$. We only get to see the sample error $error_S(h)$.

How well does the sample error estimate true error?

Can we set conditions for our experiment so that we can get an estimate that is good enough for our needs?

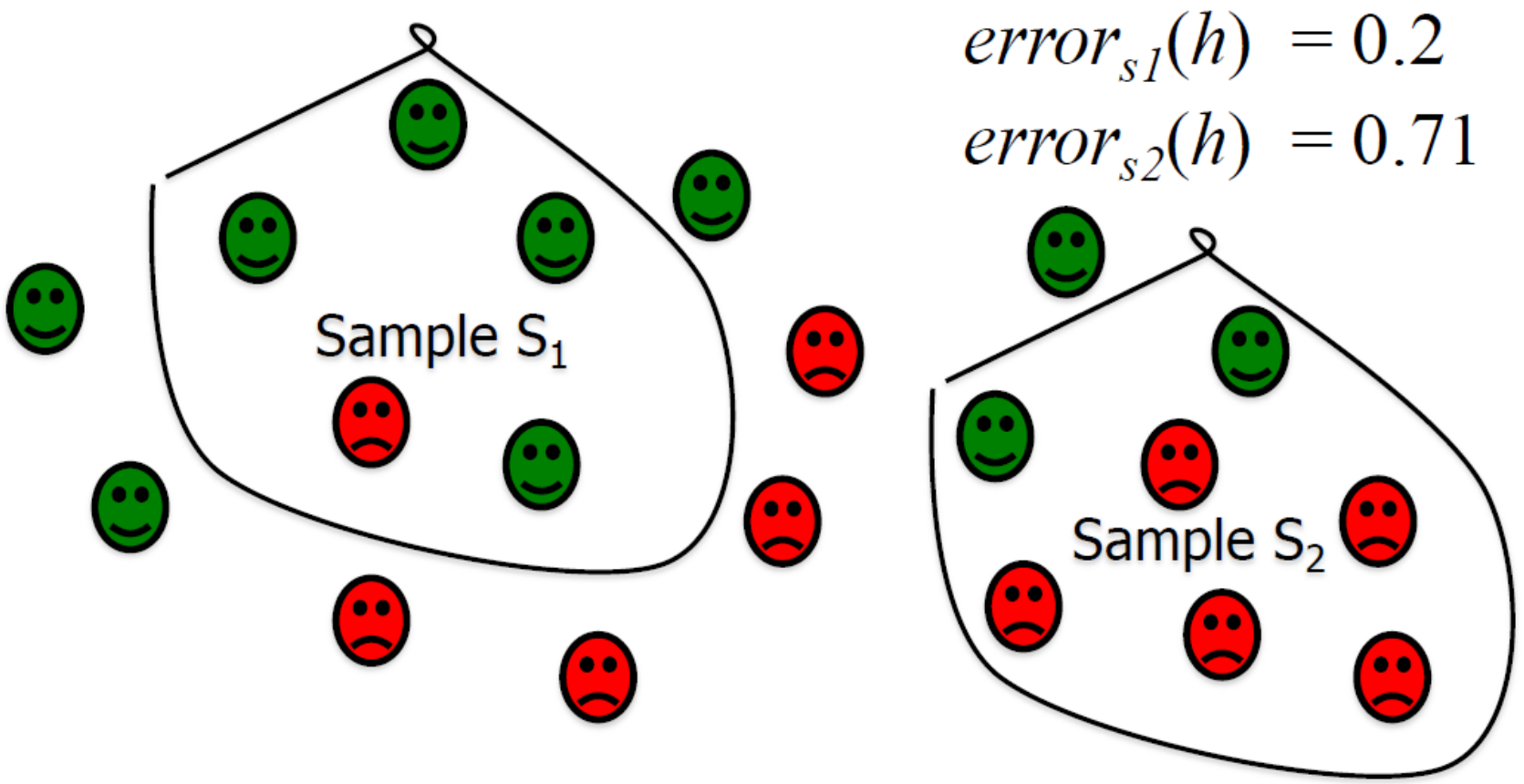
Problems Estimating Error

- Bias: If ***S*** is the training set, $error_s(\boldsymbol{h})$ is optimistically biased. For an unbiased estimate we need a validation set that was not used in training.



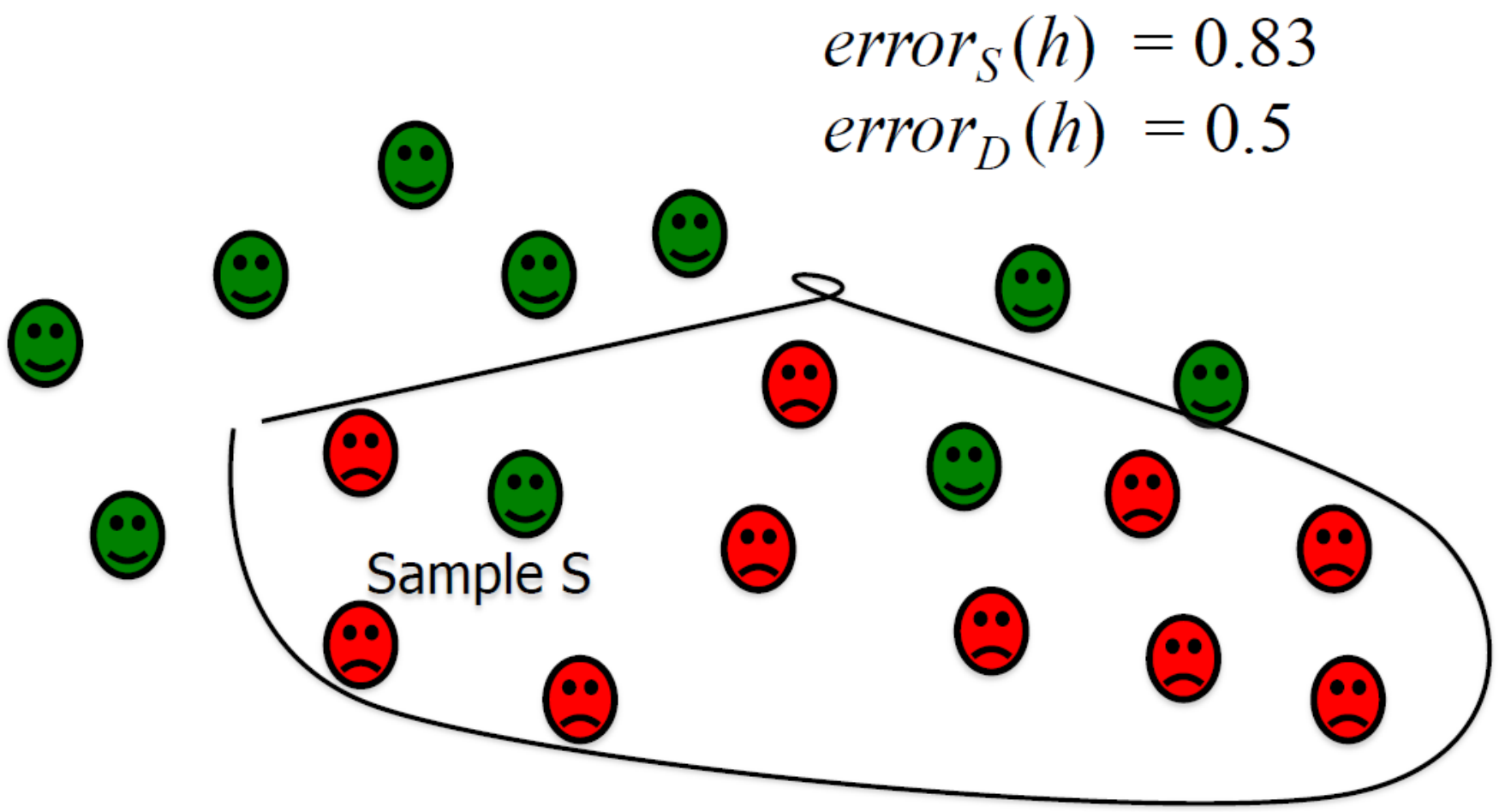
Problems Estimating Error

- Variance: Even without bias, $error_s(\boldsymbol{h})$ may still vary from $error_D(\boldsymbol{h})$



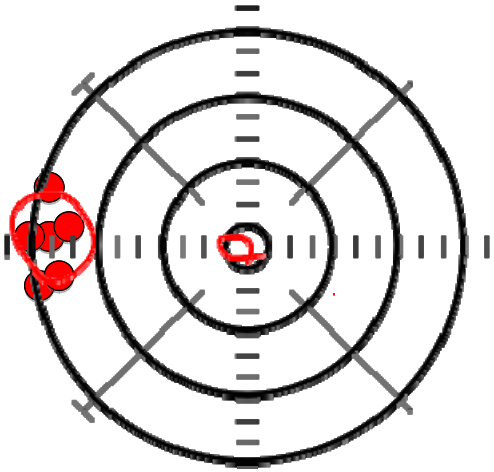
Q: Why not just take one bigger sample?

A: From one sample mean, you can't tell how $error_s(\boldsymbol{h})$ varies from $error_D(\boldsymbol{h})$

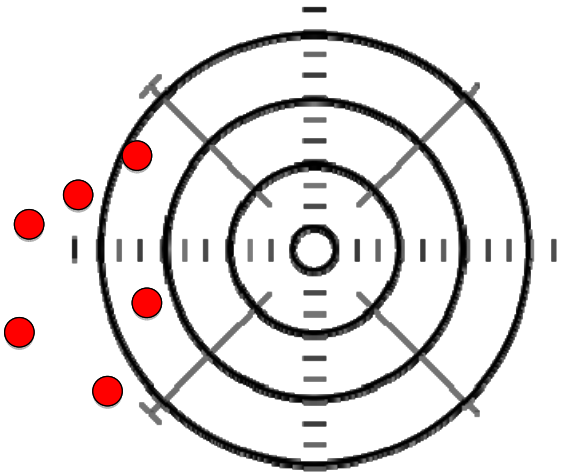


Random Forests -- Why They Work

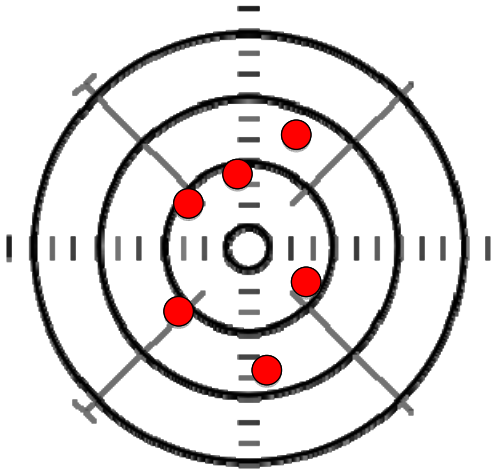
high bias, low variance



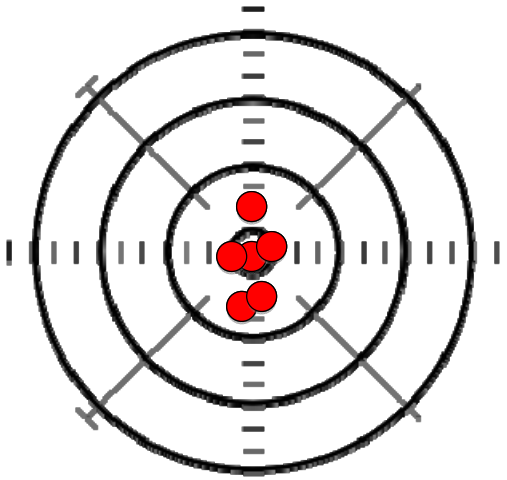
high bias, high variance



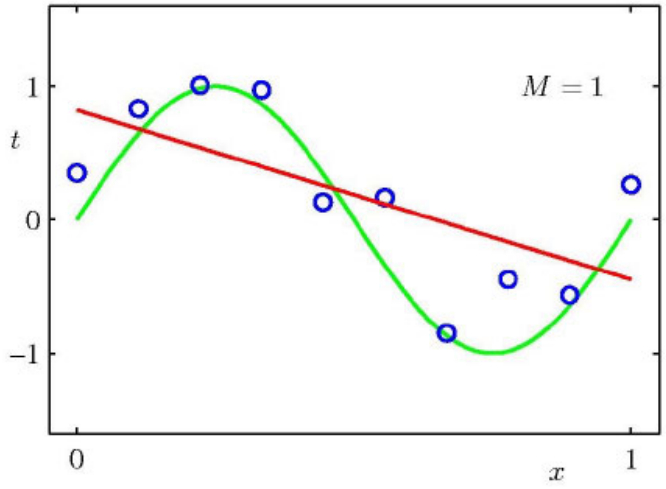
low bias, high variance

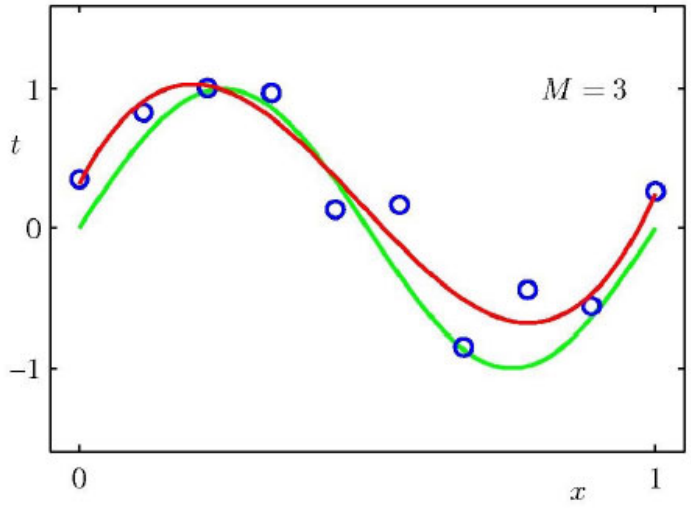


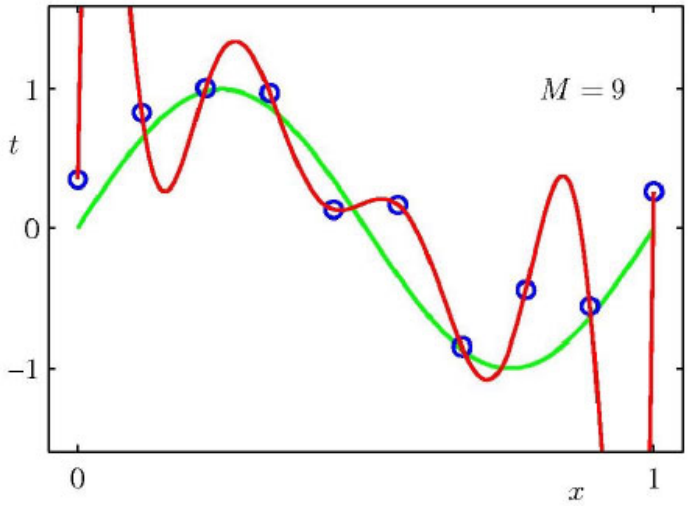
low bias, low variance



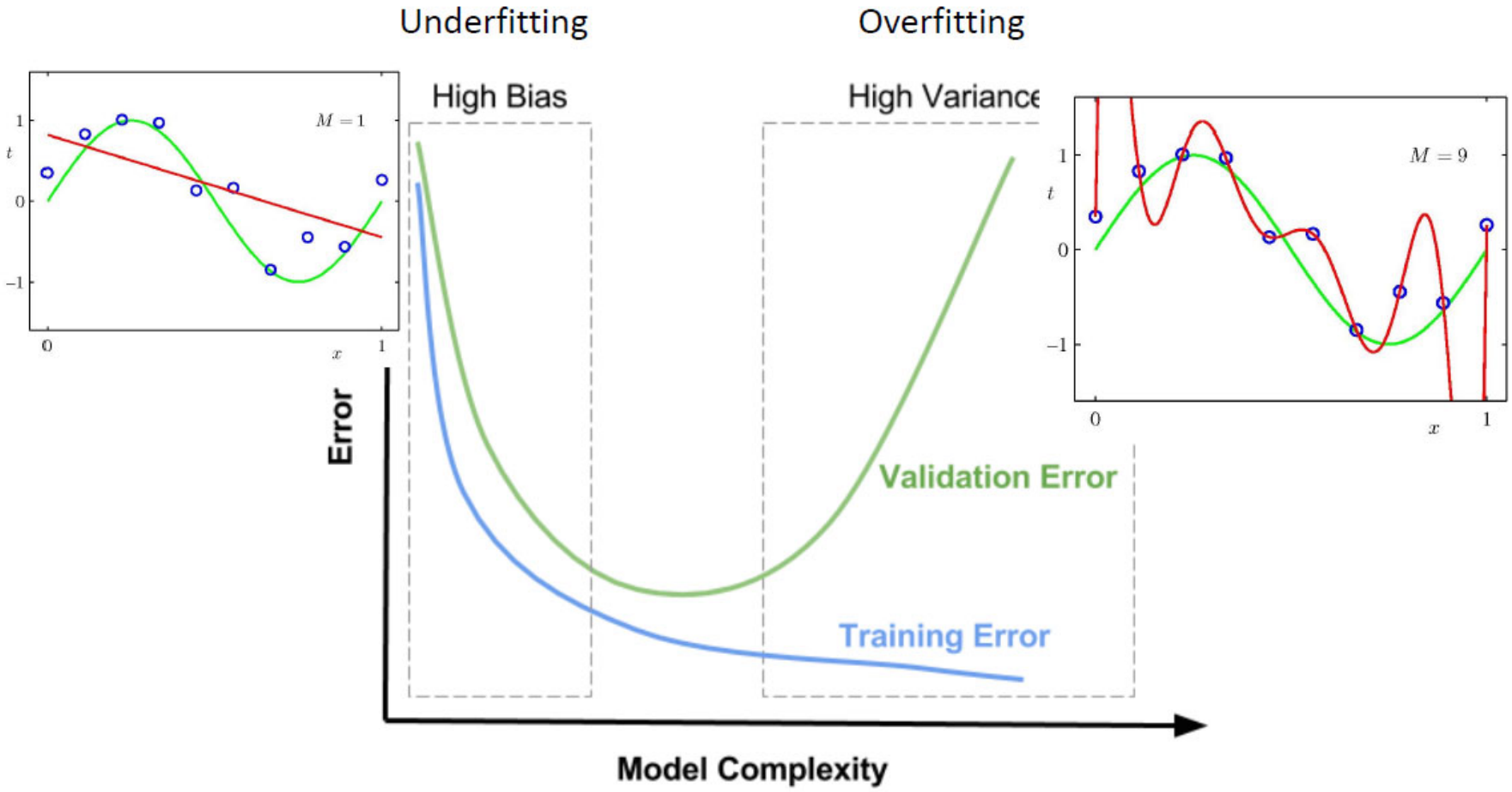
Under/Over Fitting





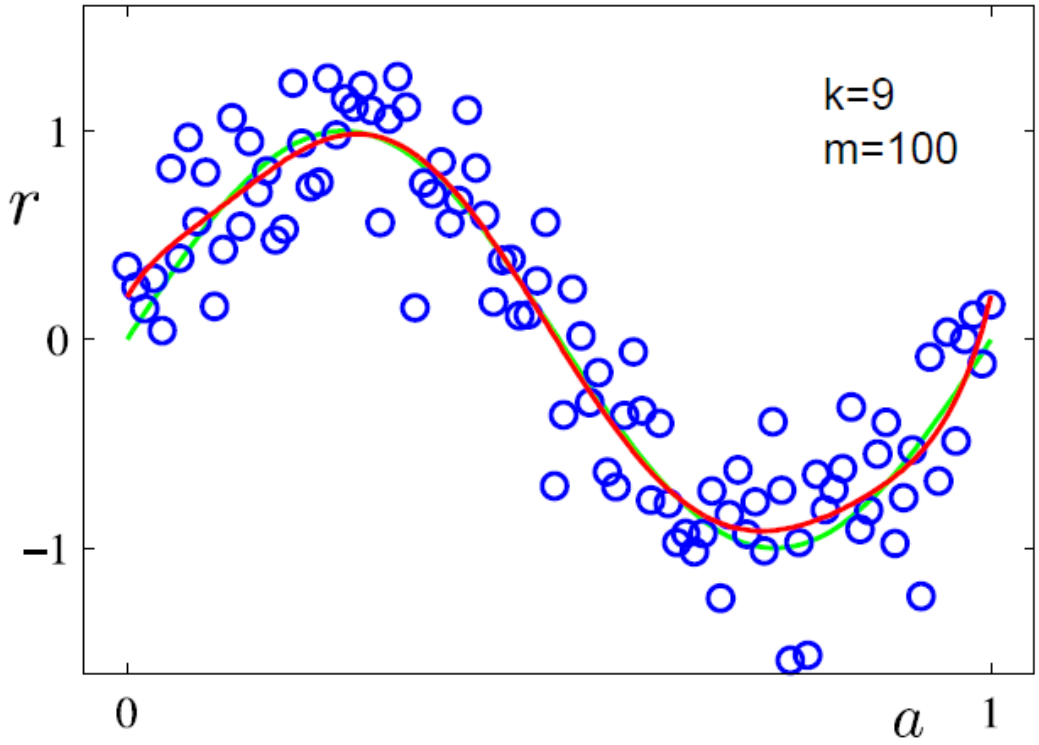


Bias/Variance Trade-off



What happens if we add more data?

kth order polynomial regression



Coin flips

- Assume an unbiased coin ***X*** that takes two values {0,1}.
- Let ***K*** be the number we get if we flip the coin ***n*** times and add up the values of all the flips.
- What is the expected value of ***K***?
- Assume ***n*** = 5
  How likely is ***K*** to be 0?
  How likely is ***K*** to be *n*/2?
- What distribution models this?

Some definitions

- A **Bernoulli Trial** is a random experiment with two outcomes (e.g., heads or tails). If the probability of a "success" outcome is p (and thus failure is 1 - p), then we call this a Bernoulli Random Variable with parameter p.

- A set of random $\{X_1, X_2, \ldots X_n\}$ variables is **independent and identically distributed (IID)** if all variables in the set are mutually independent*, and all are governed by the same probability distribution ***D***.

- A sum of n IID Bernoulli trials each with probability p is modeled by the **Binomial Distribution** with parameters n, p

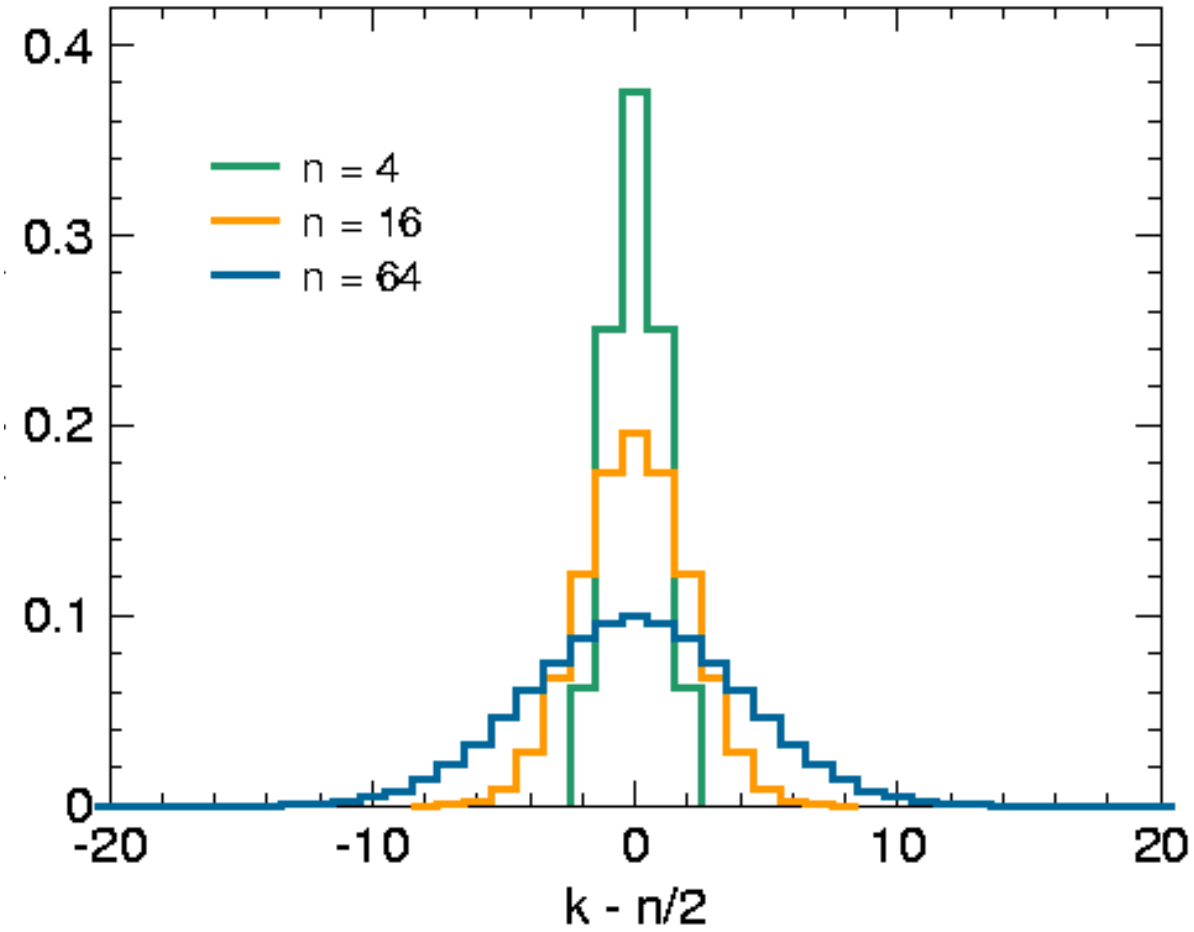
Back to the coin example…

- Assume an unbiased coin ***X*** that takes two values {0,1}
- Let ***K*** be the number we get if we flip the coin ***n*** times and add up the values of all the flips
- What is the expected value of ***K***?
- If n is 6, what's the probability that K = 3?

$$P(K = k) = B(n, k, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\text{Where } \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The binomial distribution as *n* grows



The Normal Distribution

As ***n*** goes to infinity, the Normal distribution approximates the Binomial distribution.

$$P(K = k) = B(n, k, p) \approx N(\mu, \sigma^2)$$

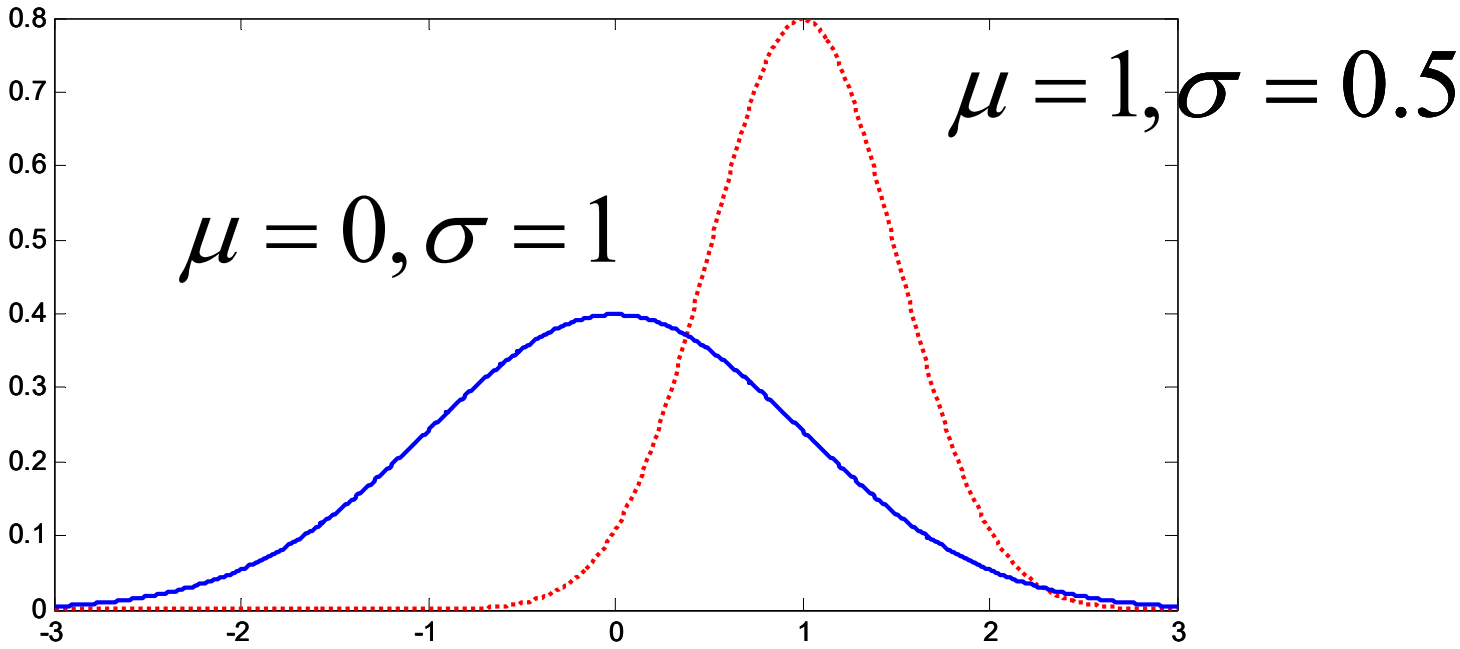
$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The Normal (Gaussian) Distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

mean

variance



Central Limit Theorem

- Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size ***n***…i.e. a set of IID discrete random variables from some distribution ***D*** with expected value μ and variance σ.
- Define the sample average $\bar{x}$ as…

$$\bar{x} = \frac{1}{n}K = \frac{1}{n}\sum_{1}^{n} X_n$$

- For large *n*, the distribution of $\bar{x}$ is approximated by the normal distribution.
- **Important:** The distribution for the sample average approaches normality regardless of the shape of the distribution *D* governing our random samples X_i.

Why the previous slides matter

- Classification is like a coin flip, you're either right or wrong.
- If classification is independent, then the number of correct classifications ***K*** is governed by a Binomial distribution
- If the Binomial distribution is approximated by the Normal distribution we can use what we know about the Normal distribution.
- The Normal distribution lets us estimate how close the TRUE error is to the SAMPLE error.

How many samples do I need…

…before my sample's distribution is approximately normal?

More is always better. The more samples you have, the closer it gets to a normal distribution.

Rule of thumb: have at least 30 IID trials.

(let's look)

Confidence Intervals: Estimating a value

1. Pick a parameter to estimate

   $error_D(h)$

2. Choose an estimator

   $error_s(h)$

3. Determine the probability distribution governing the estimator

   $error_s(h)$ governed by Binomial, approximated by Normal
   when $n > 30$ (and bigger values for n are always better)

4. Find the interval such that N% of the probability mass falls in that interval

   Use your favorite statistics software, or look up z_N values.

How many samples do I need…

…to give me good confidence intervals (assuming we already have a normal distribution)?

The standard deviation of the sample mean is related to the standard deviation of the population σ and the size of the sample, ***n*** by the following:

$$SD_{\bar{X}} = \sigma / \sqrt{n}$$

Practical result: to decrease uncertainty in a mean estimate by a factor of *n* requires n^2 observations.

Setting 95% confidence interval size

Recall that:

$$SD_{\bar{X}} = \sigma / \sqrt{n}$$

For a normal distribution, 95% of the mass is within 2 standard deviations of the mean.

For estimating a sample mean, an approximate 95% confidence interval has the form…

$$(\bar{x} - 2\sigma / \sqrt{n}, \bar{x} + 2\sigma / \sqrt{n})$$

So, the 95% confidence interval width is

$$W = 4\sigma / \sqrt{n}$$

Student's t-test Facts

- One of the most commonly used statistical tests
- Assumes normally distributed data
- Different variants for different questions….

one sample t-test: Is a known population mean μ different from the mean of a sample population?

$$t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

paired samples t-test: Is 0 the mean difference between paired responses measured on the same data?

independent samples t-test: Are the means of two normally distributed populations equal?

Student's t-test Fact(oid)s

- The t-test was devised by William Gosset in 1908
- It was used to monitor the quality of Guinness Stout (beer).
- Gosset published the t-statistic under the name "student" because Guinness considered it a trade secret

One sample t-test

Abstract question:

Is a known population mean μ different from the mean of a sample population?

Example:

We know $\mu = 0.3$ is the error rate ID3 has on categorizing a given data set.

I trained 30 neural nets to categorize the same data set and the mean error rate was $\overline{x} = 0.2$. Are neural nets better on this data set? Or was that a fluke?

Use a one-sample t-test to find out.

$$t = \frac{0.2 - 0.3}{0.5 / \sqrt{30}}$$
$$= 1.09$$

One sample t-test

Null Hypothesis: There is no significant difference between the sample mean and the population mean

Neural nets perform no better than ID3 on this data.

Alternate Hypothesis: There is a significant difference between the sample mean and the population mean.

Neural nets DO perform better than ID3 on this data.

Paired samples t-test

Abstract question:

Is 0 the mean difference between paired responses measured on the same data ?

Example:

Does eating carrots make you see better? Take 1000 people. Administer a vision test. Feed each person carrots. Administer a second vision test.

A paired-samples t-test is appropriate. Why?

$$t = \frac{\frac{1}{N}\sum_{N}(x_n^{(a)} - x_n^{(b)})}{\sigma / \sqrt{N}}$$

Paired samples t-test

Null Hypothesis: There is no significant difference between the two sample means.

Carrots do not make you see better.

Alternate Hypothesis: There is a significant difference between the two sample means.

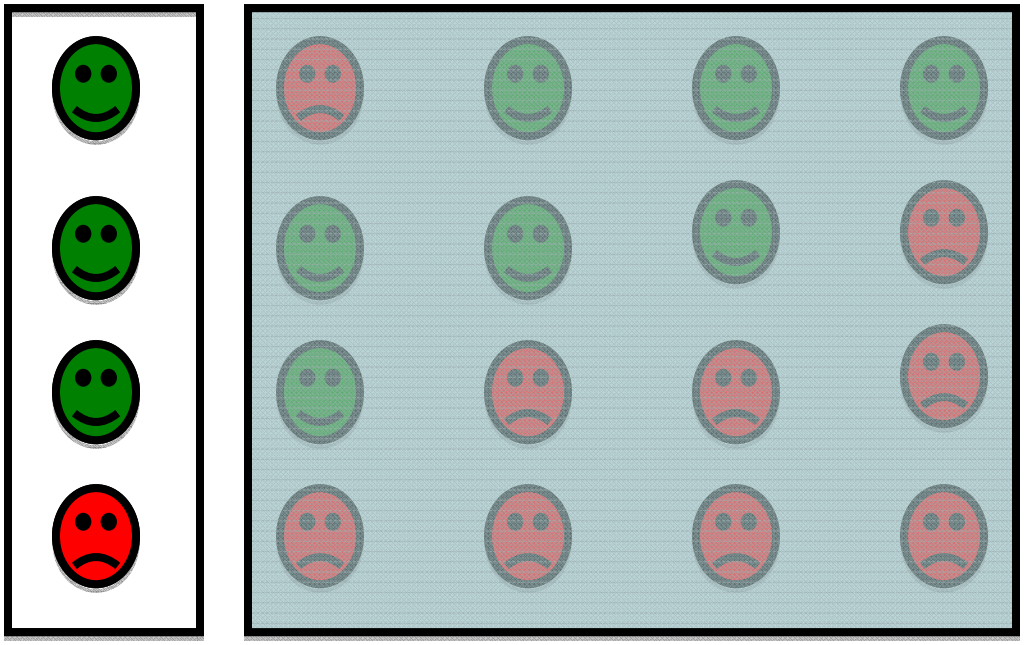
Carrots make you see better. Or they make you see worse.
We didn't actually check which way the difference goes.

Conclusions (and Pitfalls)

- The error measure should captures what you really want to know….not what is easy to measure.
- Your measure may have variance/bias/noise. Therefore…
- Results are more meaningful when a statistical significance test is done.
- Many tests depend on the data being normally distributed
- By taking the sample average of a large set of IID trials, you can ensure normal-like data
- The t-test is a good, easy test to use…if you know when to use it and how
- Common Pitfalls:
  - Data is not normally distributed (can't use a t-test)
  - Not enough sample points
  - Using a paired-samples t-test on data where the samples aren't paired (use independent samples t-test, instead)

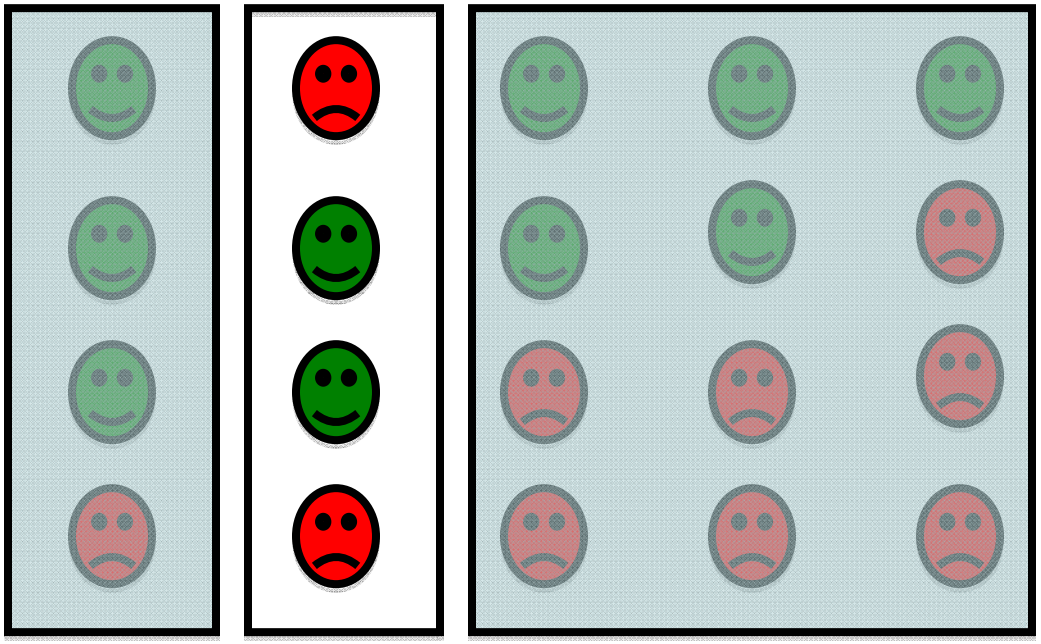
N-fold cross validation

- Spilt data into N groups.
- Train on 2:N groups.
- Validate on the 1st.
- Rotate, repeat.



N-fold cross validation

- Spilt data into N groups.
- Train on 1 + 3:N groups.
- Validate on the 2nd.
- Rotate, repeat.



N-fold cross validation

- Spilt data into N groups.
- Train on 1:2 and 4:N groups.
- Validate on the 3rd.
- Rotate, repeat.