Verification and Testing

Chirayu Wongchokprasitti, PhD

University of Pittsburgh

Center for Causal Discovery

Department of Biomedical Informatics

chw20@pitt.edu

http://www.pitt.edu/~chw20

Overview

- Introduction
- Statistical significance testing, confidence intervals
- Sensitivity and specificity
- ROC curves
- AUC
- Cross-validation

Introduction

The task

"I can prove it or disprove it! What do you want me to do?"

Introduction: The need for verification

The fundamental question:

How do we know that the knowledge extracted from data is worth anything?

http://www.ehow.com/how_7897502_evaluate-higher-order-questions-answers.html

Introduction: The need for verification

Many possible (often problem-dependent) answers, e.g.,

- ***"I'm just reporting what I see in the data"***
- ***"My model performs well in practice"***
- ***"I can provide a measure of reliability of the extracted knowledge"***
- ***"I have confirmed the discovery independently"***

Some history: Foundations of Western scientific thought

Introduction: Philosophical foundations

Verificationism

"A statement or question is only legitimate if there is some way to determine whether the statement is true or false, or what the answer to the question is."

http://en.wikipedia.org/wiki/Verification_principle

Introduction: Philosophical foundations

Empiricism

"Experience is our only source of knowledge, i.e., knowledge comes only (or primarily) from sensory experience"

John Locke (1632-1704), a leading philosopher of British empiricism

http://en.wikipedia.org/wiki/Empiricism
http://en.wikipedia.org/wiki/John_Locke

Introduction: Philosophical foundations

Positivism

"Considering unverifiable sentences is pointless, as they cannot be verified"

August Comte (actually Isidore Auguste Marie François Xavier Comte) (1798-1857), a leading French philosopher, credited for founding the doctrine of positivism

http://en.wikipedia.org/wiki/Positivism
http://en.wikipedia.org/wiki/Auguste_Comte

Introduction: Philosophical foundations

Logical positivism

"To be meaningful, a non-analytic sentence has to be empirically verifiable"

Selected Members of the Vienna Circle (from left to right: Moritz Schlick, Rudolf Carnap, Otto Neurath, Hans Hahn and Philipp Frank), roughly 1920s

http://en.wikipedia.org/wiki/Logical_positivism
http://en.wikipedia.org/wiki/Vienna_Circle
http://payingattentiontothesky.com/2011/02/02/the-vienna-circle-verification-falsification-and-god-%E2%80%93-brian-davies/

Introduction: Philosophical foundations

Pragmatism

"There is no difference that doesn't make a difference"

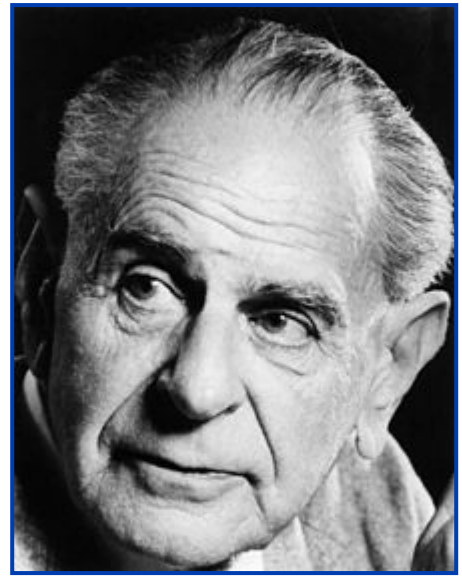
Charles Sanders Peirce (1839-1914), sometimes known as "the father of pragmatism"

http://en.wikipedia.org/wiki/Pragmatism
http://en.wikipedia.org/wiki/Charles_Sanders_Peirce

Introduction: Philosophical foundations

Falsificationism

“Meaningful sentences are falsifiable rather than verifiable”



Sir Karl Popper (1902-1994), regarded as one of the greatest philosophers of the 20th century, known best for his attempt to repudiate the classical observationalist/inductivist form of scientific method in favor of empirical falsification

http://en.wikipedia.org/wiki/Falsifiability
http://en.wikipedia.org/wiki/Karl_Popper

Statistical Significance Testing

Scientific inference

- You cannot ever be sure about truth or falsity of a hypothesis.
- You can get to the truth only with some probability.

Scientific inference: Classical hypothesis testing

Classical hypothesis testing:

- No magic associated with classical hypothesis testing
- Just a tool for decision making under uncertainty
- Why do we do it this and no other way?

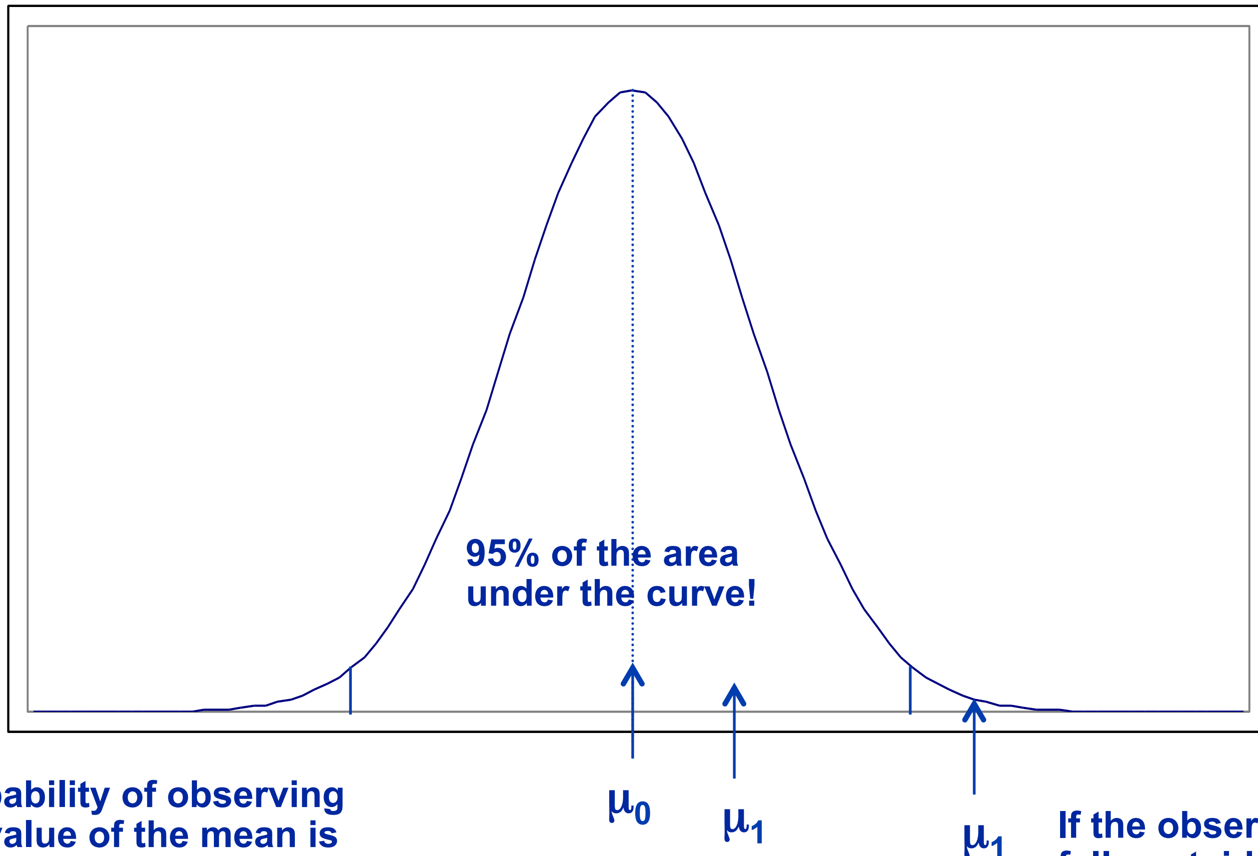
Elements of classical hypothesis testing

Elements of classical significance testing:

- Null hypothesis (H_0) and its complement (H_1).
- Significance level (α, *p* value).
- Statistical power ($1- \beta$), probability of rejecting H_0 given that it should be rejected.
- Sample size *n*.
- Effect size.

H_0 usually says something like “no effect” and is the more conservative one.

Example: Comparing the means



The probability of observing a given value of the mean is proportional to the area under the curve.

True mean

If the observed mean falls inside the 95% range, we keep H_0

If the observed mean falls outside the 95% range, we reject H_0

Risks of classical hypothesis testing

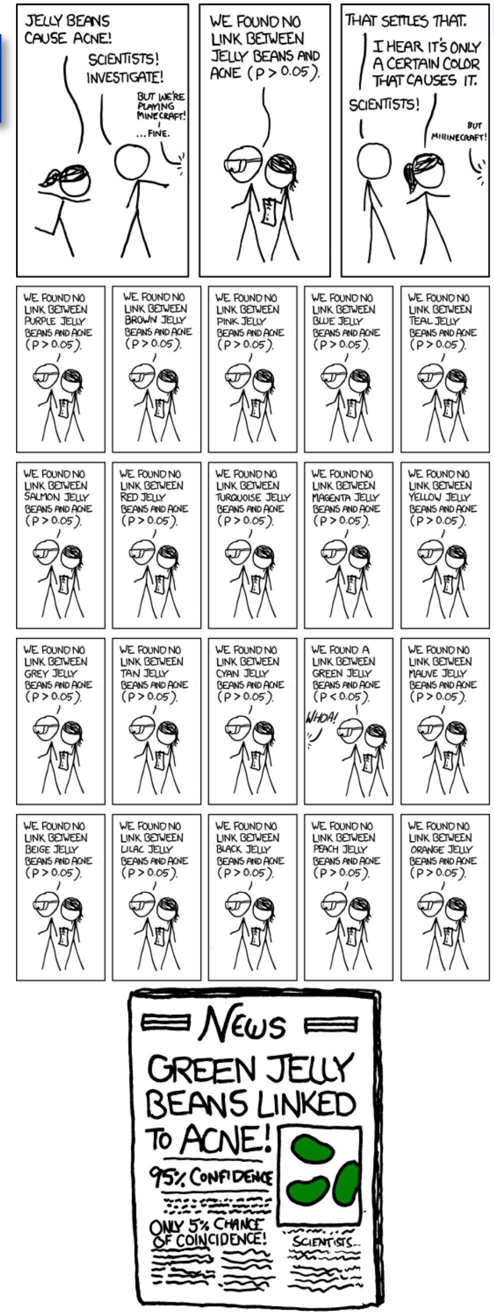
Risks related to classical significance testing (and to any decision making under uncertainty):

- Type I errors
- Type II errors

How do we deal with this risk?

- Need to consider consequences of these errors.
- What is the "correct" significance level? $\alpha = 0.05$? 0.05 is a customary value, said to be proposed by Ronald Fisher at a party.
- Traditionally scientists believe that it is worse to risk being gullible than it is to be blind to a relationship – philosophers of science characterize it as "healthy skepticism" of the scientific outlook.
- Statistical power $(1 - \beta)$
- Power curve is the plot of the power of a test as we vary one of its parameters (α, β, variance, sample size, effect size).

Testing multiple hypotheses



Caution!

If you test many hypothesis using the classical significance testing, you run an increasing risk of accidental errors.

Scientific decision making: Consequences

- Research may involve issues as difficult and important as smoking, cholesterol, etc.
- Then the congressmen or senators will pick up our results, stand up in the congress and enact laws that will either save lives or make our lives unnecessarily uncomfortable.
- Statistical inference in science is a decision process and most books will make you aware of the importance to consider consequences.
- The elements of decision theory that you get here will help you in understanding what this is about.

Elements of decision theory

The theoretically sound way of making decisions under uncertainty

- Decision making: we need to consider uncertainty and preferences. These are measured in terms of probability and utility respectively.
- Some special cases are easy:
  - it is better to be rich and healthy than poor and sick;
  - it's better to start a project that has a high chance of succeeding and a high payoff than to start a project that has a low chance of succeeding and a low payoff.

  We can reason qualitatively but it is possible to do it within this framework.
- Probability is a measure of uncertainty.
- Utility is a measure of preference that combines with probability as mathematical expectation.

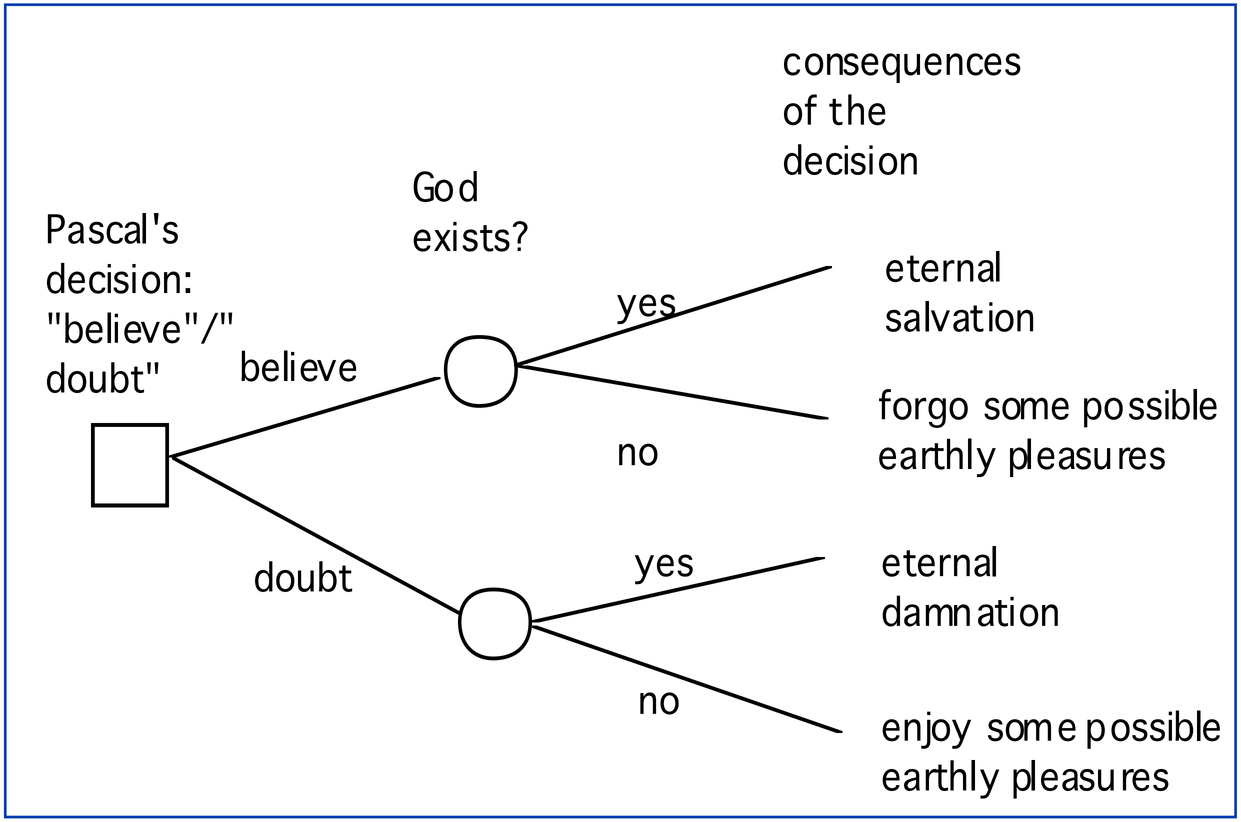
Pascal's wager

Pascal's wager: Should we believe in God or not?

	God exists	God does not exist
believe	eternal salvation	forgo some earthly pleasures in your life
doubt	eternal damnation	enjoy some earthly pleasures in your life

Pascal's wager

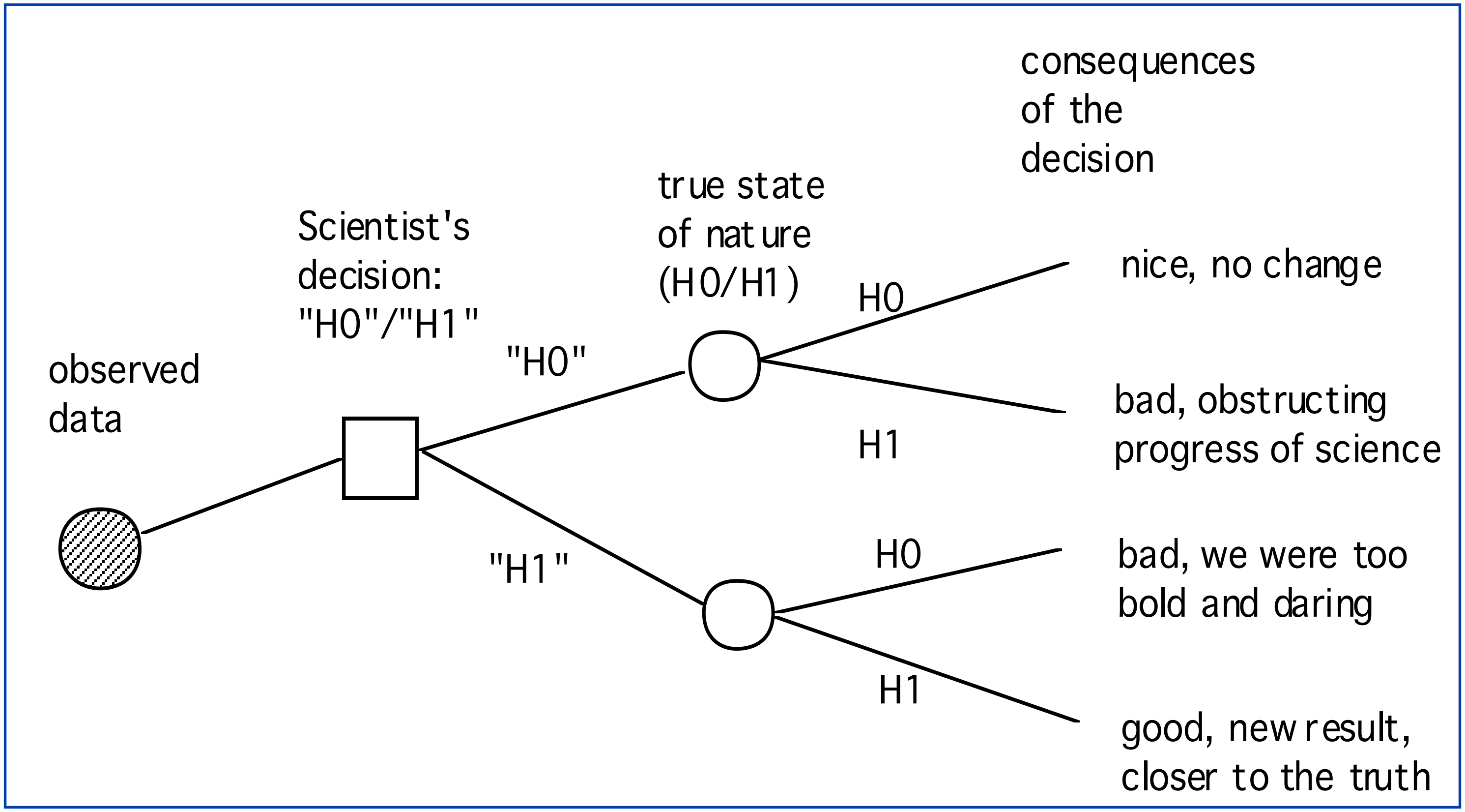
Pascal's wager: Decision tree



EU(believe) = p ∞ + (1-p)(-ε) = ∞
EU(doubt) = p (- ∞) + (1-p) ε = - ∞

The only rational thing is to believe ☺!

Classical hypothesis testing: A decision-theoretic view



Classical hypothesis testing: A decision-theoretic view

- The main problem (of course, after determining what the value of the outcomes are) is to determine the prior probability of the hypothesis. To see that, start with $Pr(H_0|D)$ and the derive everything using Bayes theorem in terms of $Pr(H_0)$, $Pr(D|H_0)$, and $Pr(D|H_1)$.
- Recall the possible errors are: (type I and type II) - α and β are probabilities of these errors.
- Decision-theoretic (Bayesian) view allows to explore the exact relation between the significance level and the decision.

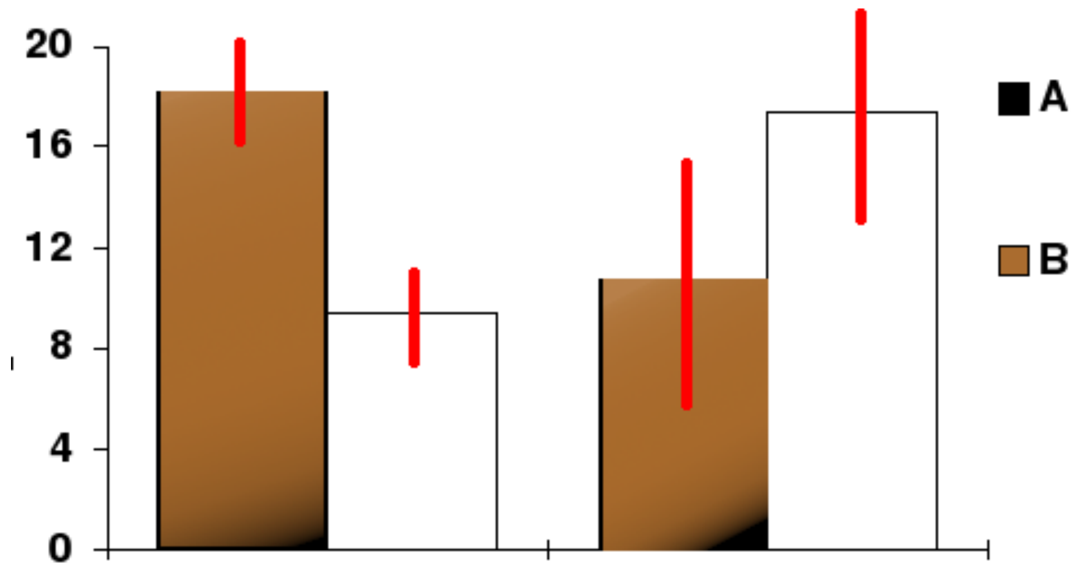
$$Pr(H_0|D)=Pr(D|H_0)Pr(H_0) / (Pr(D|H_0)Pr(H_0)+Pr(D|H_1)Pr(H_1))$$
$$= \alpha\, P(H_0)/(\alpha\, Pr(H_0) + (1-\beta)(1-Pr(H_0)))$$
$$Pr(H_1|D)=Pr(D|H_1)Pr(H_1) / (Pr(D|H_0)Pr(H_0)+Pr(D|H_1)Pr(H_1))$$
$$= (1-\beta)(1-P(H_0))/(\alpha\, Pr(H_0) + (1-\beta)(1-Pr(H_0)))$$

Confidence Intervals

Confidence intervals

- The thing is reversed here (from the point of view of the classical hypothesis testing).
- What are the boundaries of the interval of *x* values such that the interval has 95% chance of being hit.
- Watch out the proper interpretation: "I'm 95% sure that the true mean is inside this interval" and not "I'm 95% sure about the true value of the mean."

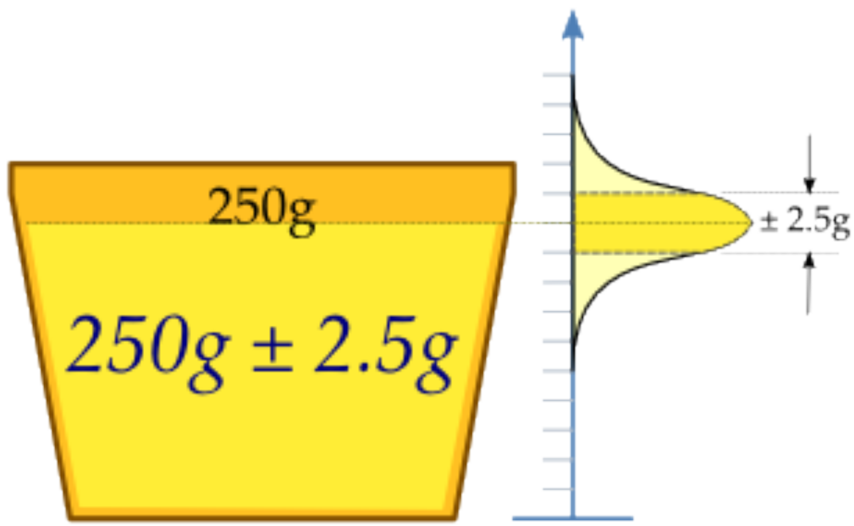
Confidence intervals



In this bar chart, the top ends of the bars indicate observation means and the red line segments represent the confidence intervals surrounding them. Although the bars are shown as symmetric in this chart, they do not have to be symmetric.

http://en.wikipedia.org/wiki/Confidence_interval

Confidence intervals: Example



A machine fills cups with margarine, and is supposed to be adjusted so that the content of the cups is 250g of margarine. As the machine cannot fill every cup with exactly 250g, the content added to individual cups shows some variation, and is considered a random variable X. This variation is assumed to be normally distributed around the desired average of 250g, with a standard deviation of 2.5g.

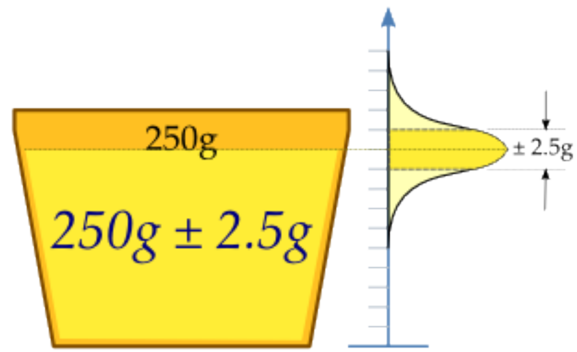
To determine if the machine is adequately calibrated, a sample of n = 25 cups of margarine are chosen at random and the cups are weighed. The resulting measured masses of margarine are $X_1, \ldots, X_{25}$, a random sample from X.

http://en.wikipedia.org/wiki/Confidence_interval

Confidence intervals: Example

To get an impression of the expectation μ, it is sufficient to give an estimate. The appropriate estimator is the sample mean:

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$



The sample shows actual weights $x_1, \ldots, x_{25}$, with mean:

$$\bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i = 250.2\,\text{grams}.$$

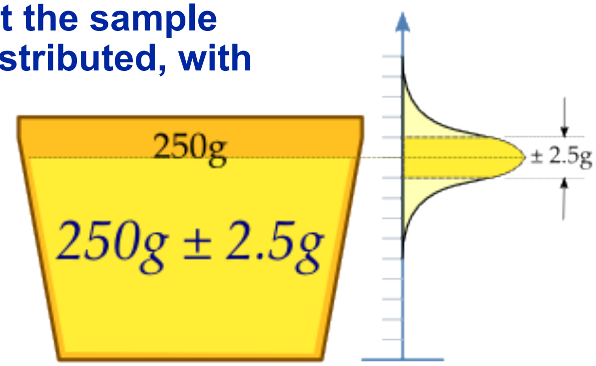
If we take another sample of 25 cups, we could easily expect to find mass values like 250.4 or 251.1 grams. A sample mean value of 280 grams however would be extremely rare if the mean content of the cups is in fact close to 250 grams. There is a whole interval around the observed value 250.2 grams of the sample mean within which, if the whole population mean actually takes a value in this range, the observed data would not be considered particularly unusual. Such an interval is called a confidence interval for the parameter μ. How do we calculate such an interval? The endpoints of the interval have to be calculated from the sample, so they are statistics, functions of the sample $X_1, \ldots, X_{25}$ and, hence, random variables themselves.

http://en.wikipedia.org/wiki/Confidence_interval

Confidence intervals: Example

In our case, we may determine the endpoints by considering that the sample mean *X* from a normally distributed sample is also normally distributed, with the same expectation *μ*, but with a standard error of:

$$\frac{\sigma}{\sqrt{n}} = 0.5 \text{ grams}$$



By standardizing, we get a random variable:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{0.5}$$

dependent on the parameter μ to be estimated, but with a standard normal distribution independent of the parameter *μ*. Hence it is possible to find numbers −*z* and *z*, independent of *μ*, between which *Z* lies with probability 1 − α, a measure of how confident we want to be.

We take 1 − α = 0.95, for example. So we have:

$$P(-z \le Z \le z) = 1 - \alpha = 0.95.$$

The number *z* follows from the cumulative distribution function, in this case the cumulative normal distribution function:

$$\Phi(z) = P(Z \le z) = 1 - \tfrac{\alpha}{2} = 0.975,$$

$$z = \Phi^{-1}(\Phi(z)) = \Phi^{-1}(0.975) = 1.96,$$

and we get:

$$0.95 = 1 - \alpha = P(-z \le Z \le z) = P\left(-1.96 \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le 1.96\right).$$

$$= P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

http://en.wikipedia.org/wiki/Confidence_interval

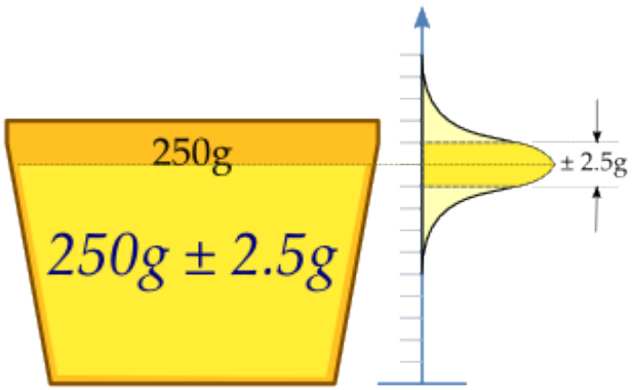
Confidence intervals: Example

In other words, the lower endpoint of the 95% confidence interval is:

$$Lower\ endpoint = \bar{X} - 1.96\frac{\sigma}{\sqrt{n}},$$

and the upper endpoint of the 95% confidence interval is:

$$Upper\ endpoint = \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}.$$



With the values in this example, the confidence interval is:

$$0.95 = P\left(\bar{X} - 1.96 \times 0.5 \le \mu \le \bar{X} + 1.96 \times 0.5\right)$$

$$= P\left(\bar{X} - 0.98 \le \mu \le \bar{X} + 0.98\right).$$

This might be interpreted as: with probability 0.95 we will find a confidence interval in which we will meet the parameter μ between the stochastic endpoints

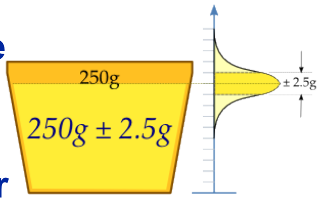
$$\bar{X} - 0.98$$

and

$$\bar{X} + 0.98.$$

http://en.wikipedia.org/wiki/Confidence_interval

Confidence intervals: Example



This does not mean that there is 0.95 probability of meeting the parameter μ in the interval obtained by using the currently computed value of the sample mean,

$$(\bar{x} - 0.98,\ \bar{x} + 0.98).$$

Instead, every time the measurements are repeated, there will be another value for the mean X of the sample. In 95% of the cases μ will be between the endpoints calculated from this mean, but in 5% of the cases it will not be. The actual confidence interval is calculated by entering the measured masses in the formula. Our 0.95 confidence interval becomes:

$$(\bar{x} - 0.98; \bar{x} + 0.98) = (250.2 - 0.98; 250.2 + 0.98) = (249.22; 251.18).$$

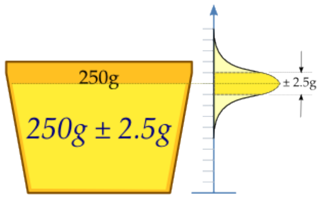
In other words, the 95% confidence interval is between the lower endpoint 249.22g and the upper endpoint 251.18g.

As the desired value 250 of μ is within the resulted confidence interval, there is no reason to believe the machine is wrongly calibrated.

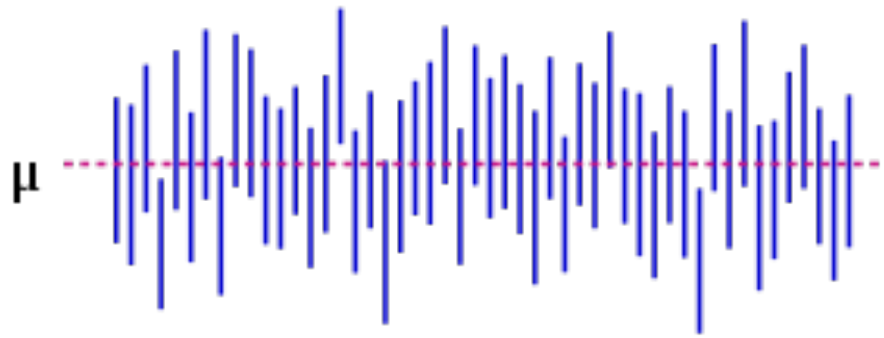
The calculated interval has fixed endpoints, where μ might be in between (or not). Thus this event has probability either 0 or 1. One cannot say: "with probability (1−α) the parameter μ lies in the confidence interval." One only knows that by repetition in 100(1 − α) % of the cases, μ will be in the calculated interval. In 100α % of the cases however it does not. And unfortunately one does not know in which of the cases this happens. That is why one can say: "with confidence level 100(1 − α) %, μ lies in the confidence interval."

http://en.wikipedia.org/wiki/Confidence_interval

Confidence intervals: Example



The figure below shows 50 realizations of a confidence interval for a given population mean μ. If we randomly choose one realization, the probability is 95% we end up having chosen an interval that contains the parameter; however we may be unlucky and have picked the wrong one. We will never know; we are stuck with our interval.



http://en.wikipedia.org/wiki/Confidence_interval

Confidence intervals: Simulation

- If there are enough samples, you can estimate the confidence intervals from the sample.
- If there are not enough samples but they can be assumed that they are representative of the population, you can "bootstrap" the sample.
- Generally, parametric methods, like those in the example, can be replaced by computer-intensive methods (read "simulation").

Sensitivity and Specificity

Sensitivity and specificity

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function.

Sensitivity (also called recall rate in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition).

Specificity measures the proportion of negatives which are correctly identified (e.g. the percentage of healthy people who are correctly identified as not having the condition).

These two measures are closely related to the concepts of type I and type II errors. A perfect predictor would be described as 100% sensitivity (i.e., predict all people from the sick group as sick) and 100% specificity (i.e., not predict anyone from the healthy group as sick).

In practice, however, there are no perfect predictors.

http://en.wikipedia.org/wiki/Specificity_(statistics)

Sensitivity and specificity: Definitions

Sensitivity

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \text{probability of a positive test given that the patient is ill}$$

Specificity

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$= \text{probability of a negative test given that the patient is well}$$

http://en.wikipedia.org/wiki/Specificity_(statistics)

Sensitivity and specificity: Relationship among terms

		Condition (as determined by "Gold standard")		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive	False Positive (Type I error)	Positive predictive value = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test Outcome Positive}}$
	Test Outcome Negative	False Negative (Type II error)	True Negative	Negative predictive value = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test Outcome Negative}}$
		Sensitivity = $\frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	Specificity = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

http://en.wikipedia.org/wiki/Specificity_(statistics)

Sensitivity and specificity: Example

The fecal occult blood (FOB) screen test used in 2030 people to look for bowel cancer

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = 10%
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ 99.5%
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ 67%	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = 91%	

http://en.wikipedia.org/wiki/Specificity_(statistics)

Sensitivity and specificity: Example

Related calculations

- **False positive rate** (α) = type I error = 1 − specificity = FP / (FP + TN) = 180 / (180 + 1820) = 9%
- **False negative rate** (β) = type II error = 1 − sensitivity = FN / (TP + FN) = 10 / (20 + 10) = 33%
- **Power** = sensitivity = 1 − β
- **Likelihood ratio positive** = sensitivity / (1 − specificity) = 66.67% / (1 − 91%) = 7.4
- **Likelihood ratio negative** = (1 − sensitivity) / specificity = (1 − 66.67%) / 91% = 0.37

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	True Positive (TP) = 20	False Positive (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = 10%
	Test Outcome Negative	False Negative (FN) = 10	True Negative (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ 99.5%
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ 67%	Specificity = TN / (FP + TN) = 1820 / (180 + 1820) = 91%	

Hence, with large numbers of false positives and few false negatives, a positive FOB screen test is in itself poor at confirming cancer (PPV = 10%) and further investigations must be undertaken; it did, however, correctly identify 66.7% of all cancers (the sensitivity). However as a screening test, a negative result is very good at reassuring that a patient does not have cancer (NPV = 99.5%) and at this initial screen correctly identifies 91% of those who do not have cancer (the specificity).

http://en.wikipedia.org/wiki/Specificity_(statistics)

Confusion Matrix

Confusion matrix

The same thing as what we saw before but used with reference to the model's predictions and the true state of the World.

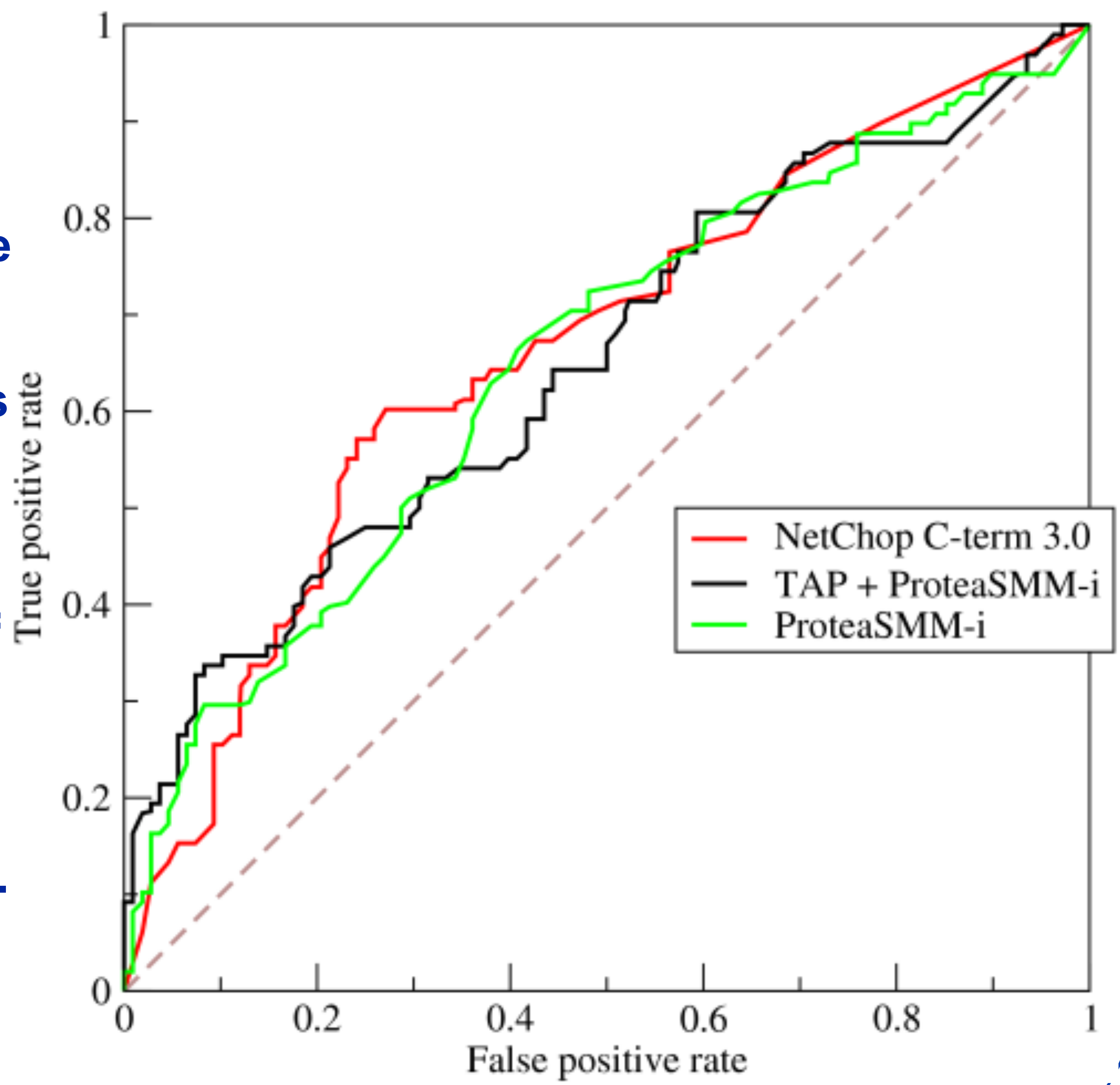
		Patients with bowel cancer **(as confirmed on endoscopy)**		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

http://en.wikipedia.org/wiki/Specificity_(statistics)

ROC (Receiver Operating Characteristic) Curves

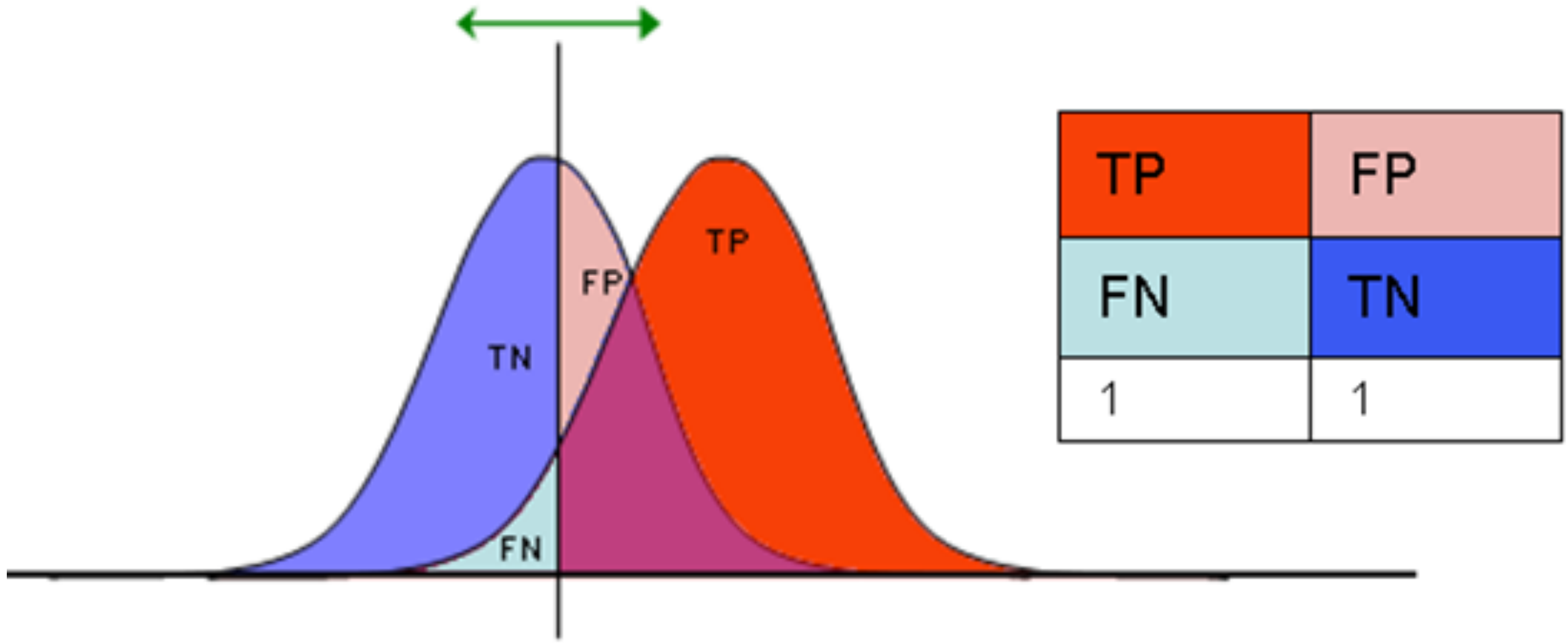
ROC curve

- Originates from the signal detection theory
- A graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.
- Created by plotting the fraction of true positives out of the positives (TPR = true positive rate) vs. the fraction of false positives out of the negatives (FPR = false positive rate), at various threshold settings. TPR is also known as sensitivity, and FPR is one minus the specificity or true negative rate.



http://en.wikipedia.org/wiki/Receiver_operating_characteristic

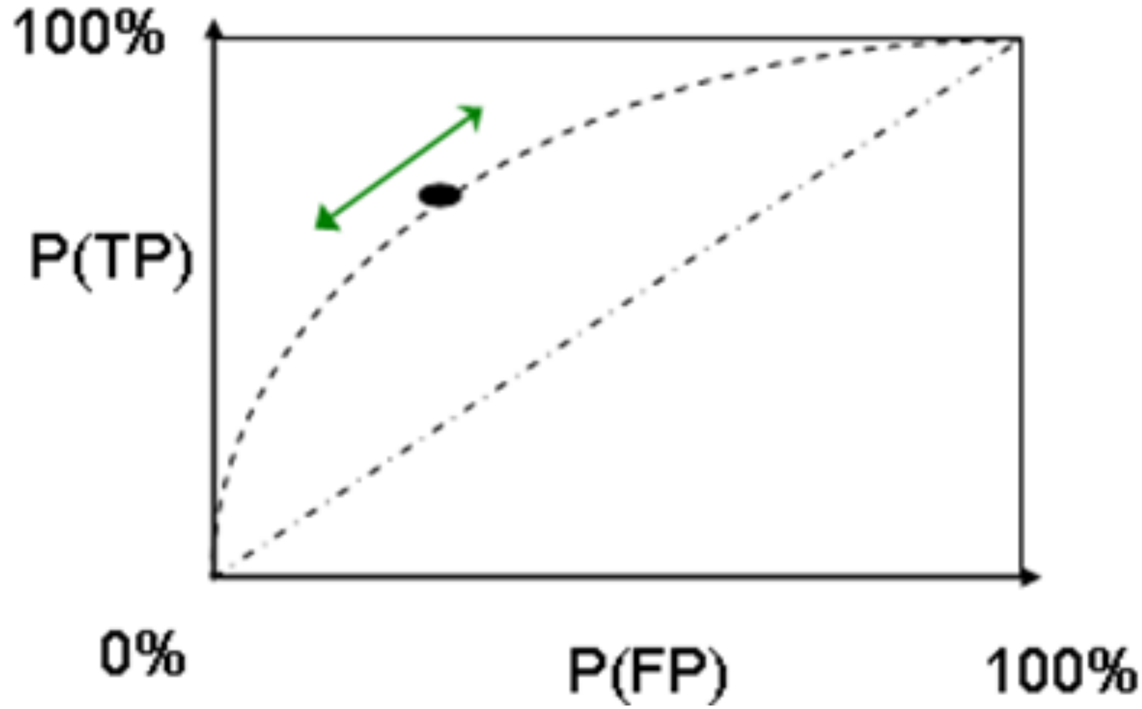
ROC curve (the common sense)



- Please note that very often we need to make a compromise between sensitivity and specificity: Higher sensitivity means lower specificity and vice versa.
- Setting the threshold is a matter of decision.
- The threshold that we decide to adopt will determine the parameters of our test (i.e., true/false positive and true/false negative rates).

http://en.wikipedia.org/wiki/Receiver_operating_characteristic

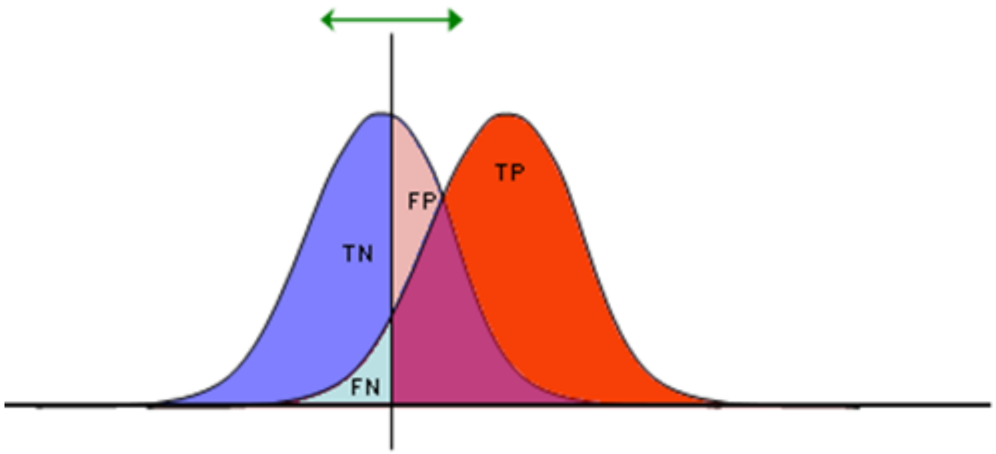
ROC curve (the common sense)



- As we move the threshold, we change the values of sensitivity and specificity. The plot of all possible values of these two parameters gives us an interesting characterization of the test (classification system, receiver, etc.)

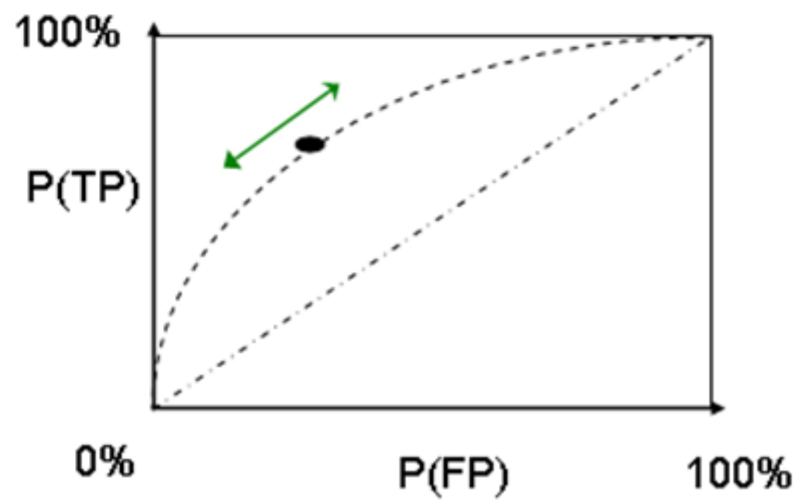
http://en.wikipedia.org/wiki/Receiver_operating_characteristic

ROC curve (the common sense)



TP	FP
FN	TN
1	1

- Plots like the one on the right-hand side are called ROC (Receiver Operating Characteristics) Curves
- They are a way of characterizing the quality of the detection system

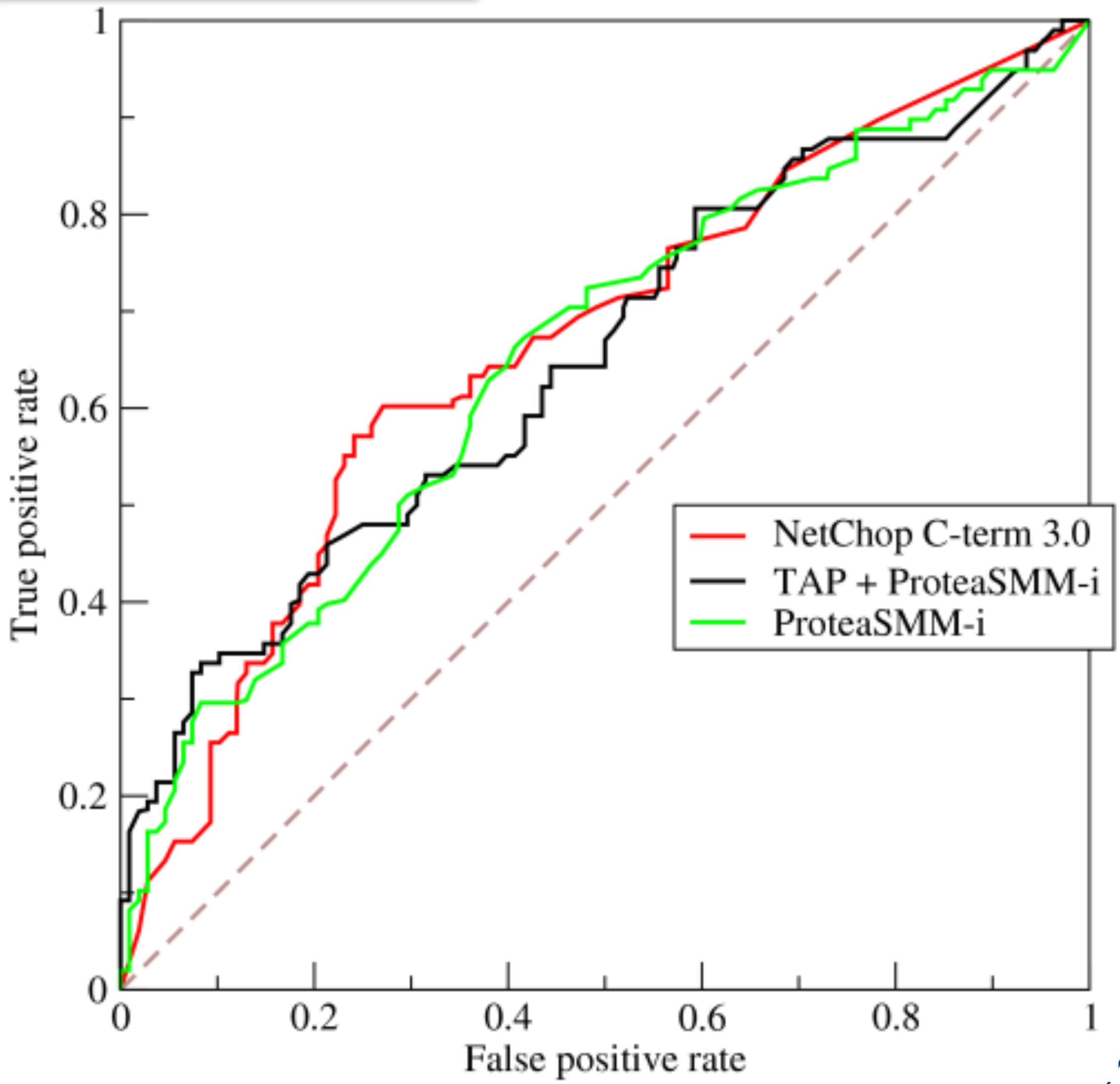


http://en.wikipedia.org/wiki/Receiver_operating_characteristic

AUC: Area Under the (ROC) Curve

AUC: Area Under the (ROC) Curve

- AUC is a way of characterizing a "receiver" by means of one number
- It captures the intuitive idea that an ROC that is higher is better.
- A perfect ROC curve will go through the point (0,1). The area under it will be 1.0.
- Criticized as an unreliable measure of performance.



http://en.wikipedia.org/wiki/Receiver_operating_characteristic

Cross-Validation

Cross-validation: The idea

- Imagine that I design a (closed book) exam consisting of 20 questions in such a way that I reveal all 20 questions and answers before the exam.
- What will the most successful approach be like?

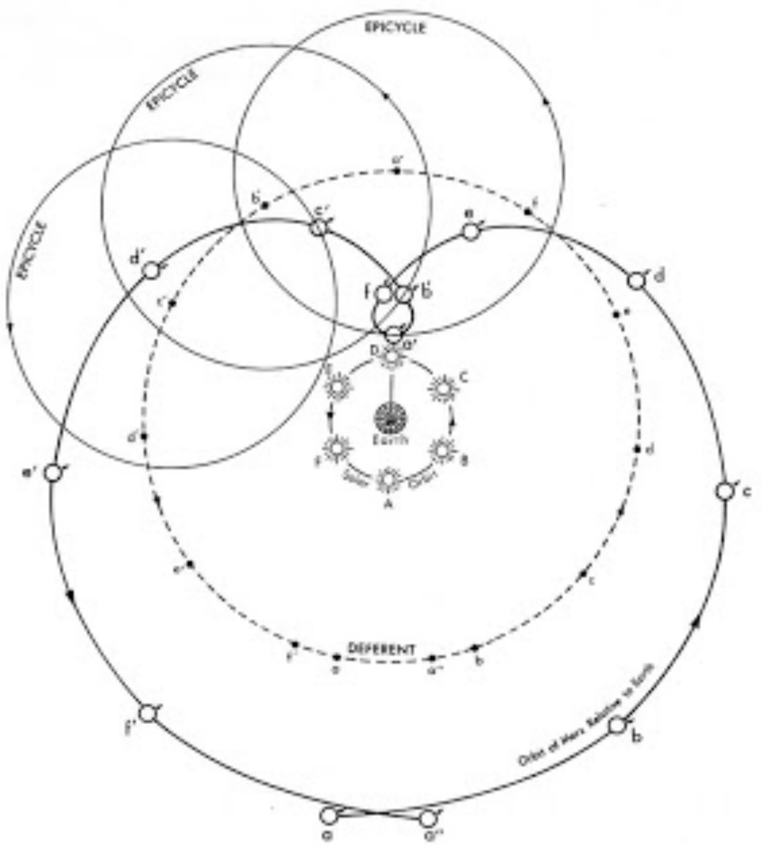
Cross-validation: The idea

- Imagine that the Kaggle competition releases all data to the participants.
- What will the most successful approach be like?

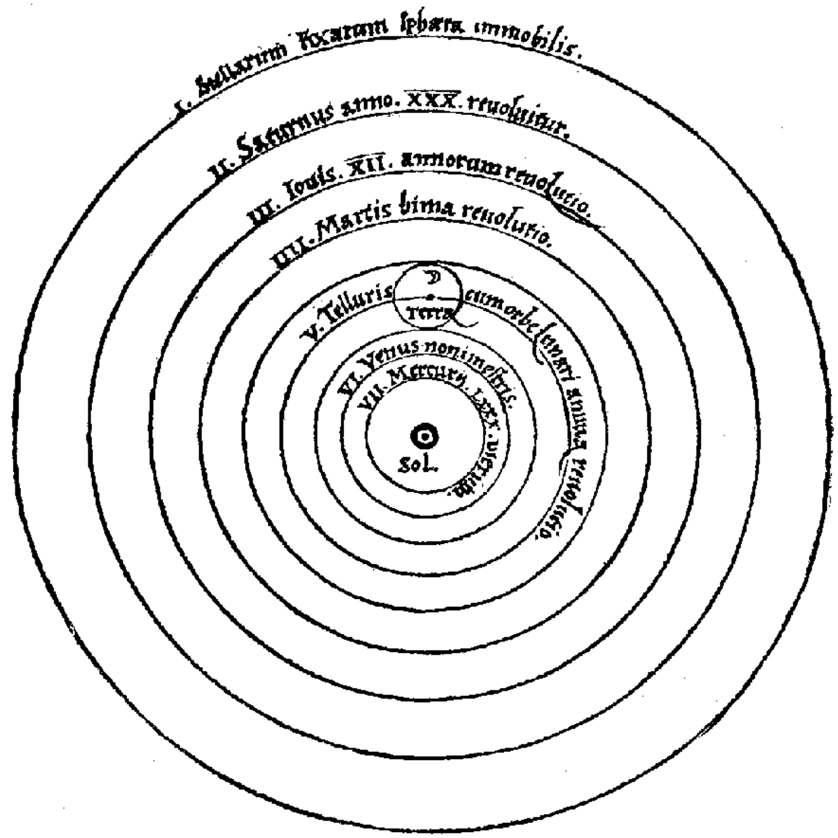
Cross-validation: The idea

- Testing a model on the same data that we used for training it does not seem fair.
- It will favor most complex models that fit the data best.
- What about simplicity?
- Simpler model may actually fit future instances of data better than complex models.

Simple vs. complex models of reality



Ptolemy's model



Copernicus' model

Cross-validation: Test set method

- **Cross-validation** is a technique for assessing how the results of a statistical analysis will generalize to an independent data set.
- It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.
- In its simplest form, you divide the data into two disjoint sets: (1) training set and (2) test set (a.k.a. validation set).
- You perform the analysis on the training set and validate the results on the test set.
- This is simple and effective.

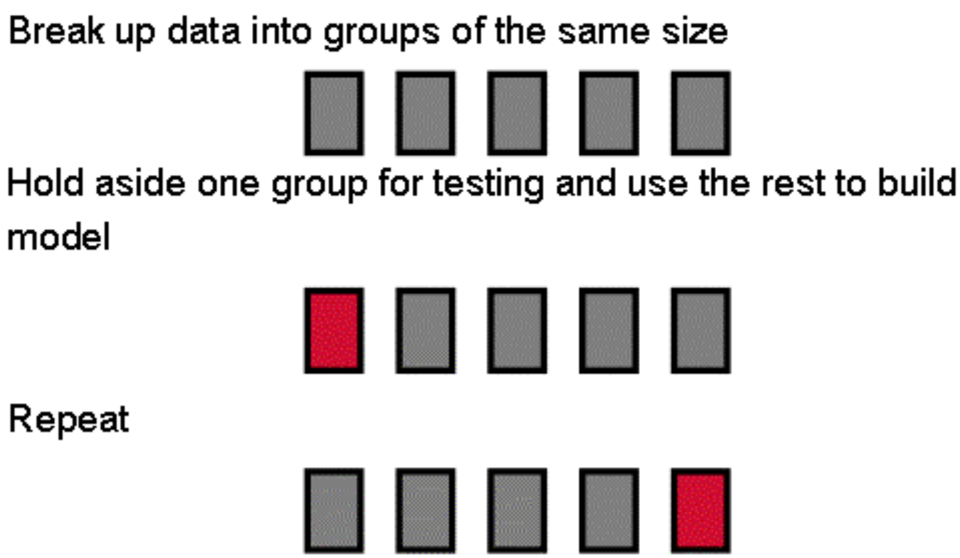
http://en.wikipedia.org/wiki/Cross-validation_(statistics)

k-Fold cross-validation

- To reduce variability, you may perform multiple rounds of cross-validation, using a different partition of the data in each round.
- The validation results can be then averaged over the rounds.
- The cost of this is more computation, as you have to repeat the procedure of learning multiple times.

k-Fold cross-validation

The idea:



http://blog.weisu.org/2011/05/cross-validation.html

k-Fold cross-validation

Procedure: k-fold cross validation

1. Shuffle the items in the training set
2. Divide the training set into *k* equal parts of size *n* (e.g., 10 sets of size n=50 instances)
3. Do *i* = 1 to *k* times:
    a. Call the *i*th set of n sentences the test set; set it aside
    b. Train the system on the remaining *k*-1 sets; test the system on the test set; record performance.
    c. Clear memory: forget everything learned during training
4. Calculate average performance from the *k* test sets

Now every instance is used for both training and testing.

Why do we need to clear memory in Step 3-c?

Leave-One-Out cross-validation

A special case of *k*-fold cross-validation taking it to the extreme (*k*=*n*).

Very efficient in terms of <u>maximizing the size of the training set</u>.

Procedure: Leave-One-Out

1. Do *i* = 1 to *n* times:
   a. Set aside instance *i*
   b. Train the system on the remaining *n*-1 instances;
      test the system on the instance *i* that was set aside;
      record performance.
   c. Clear memory: forget everything learned during training
4. Calculate average performance on the *n* test cases

Now we have effectively used n-1 instances for training and tested the model on all n instances.

Bootstrap cross-validation

Cross-validation is a form of re-sampling, just like bootstrap. The bootstrap, however, involves re-sampling with replacement from a sample *M* of *n* (e.g., *n*=500 instances)

Procedure: bootstrap cross validation

1. Do *k* times:
   a. Draw *n* items from *M* with replacement, call this sample *R*
   b. Find the items in *M* that do not appear in *R*, call these the test set
   c. Train the system on the items in *R*; test it on the items in the test set; and record the performance
   d. Clear memory, that is, forget everything learned during training
2. Calculate the average performance from the *k* test sets

Because *k* should be 200 or more, this involves a lot more computation than, say, 10-fold cross validation.

It can outperform cross-validation in some cases.

Cross-validation to measure learning

Cross-validation is easily adapted to estimate performance after 1, 2, ..., *M* training sentences:

Procedure: Incremental *k*-fold cross validation

1. Shuffle the items in the training set
2. Divide the training set into *k* equal parts of size *n*, say, ten sets of size *n*=50 sentences
3. Do i = 1 to *k* times:
   a. Call the *i*-th set of *n* sentences the test set; call the remaining *k*-1 sets of sentences the training set
   b. Repeat 10 times:
   Select one sentence at random from the training set; train the system on this sentence; test the system on the test set; record the performance.
   c. Repeat 9 times:
   Select ten sentences at random from the training set; train the system on these sentences; test the system on the test set; record the performance.
   …
   d. Clear memory, that is, forget everything learned during training
4. Calculate the performance for each level of training averaged over the the k test sets

Trains and tests on 1, 2, 3, ...,10, 20, 30, ..., 100 sentences in each fold.

Cross-validation as a guide in model selection

- Learn different models, perform k-fold cross-validation, choose the model with the lowest error
- Train the best model on all data and use it as a predictive model that you will apply to future cases.

Cross-validation as a guide in algorithm choice

- Apply different algorithms, perform k-fold cross-validation.
- Choose the algorithm with the lowest error.
- This is, in a way, what Kaggle competitions are doing.

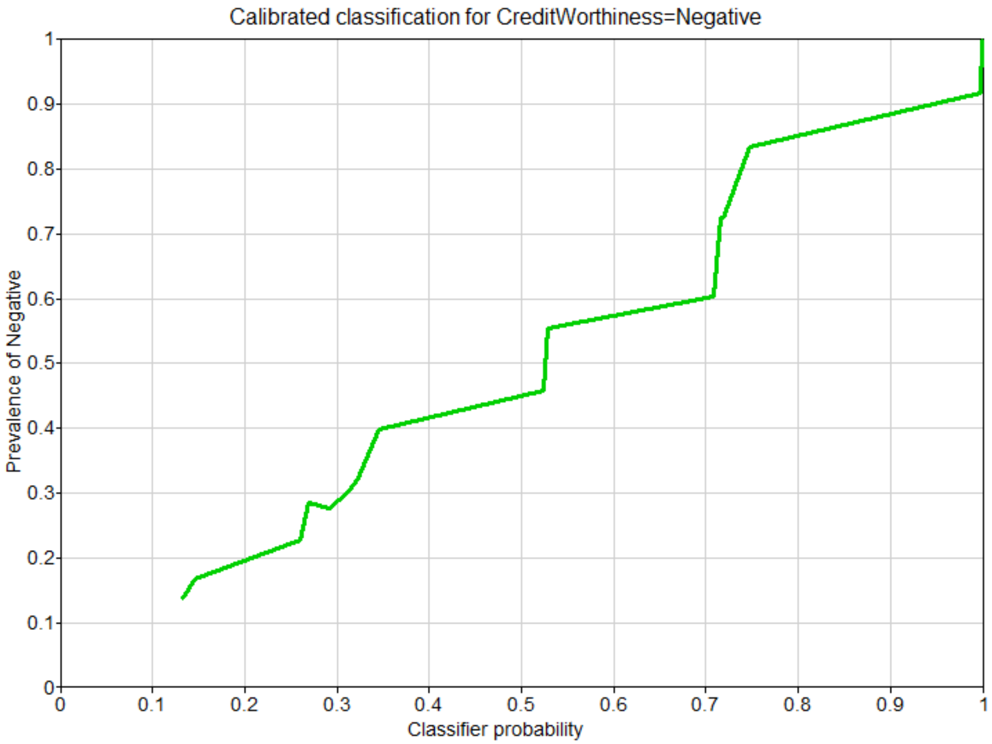
Calibration

Calibration

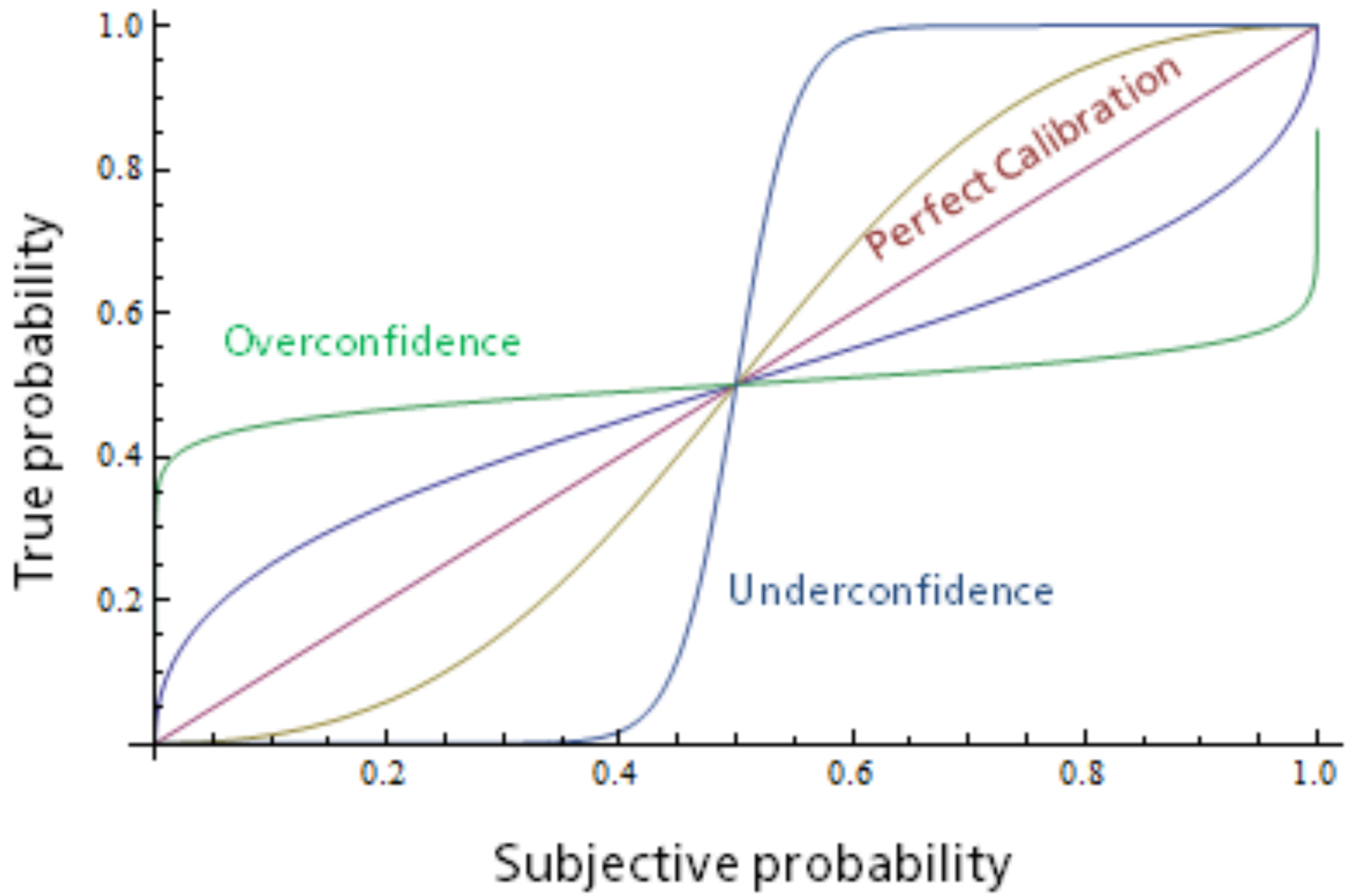
The question here is: Is my model producing accurate probabilities?

Why do you want them accurate? Useful in decision making!

We plot the “true probabilities” (i.e., the frequencies observed in the data) against the probabilities calculated by the system.

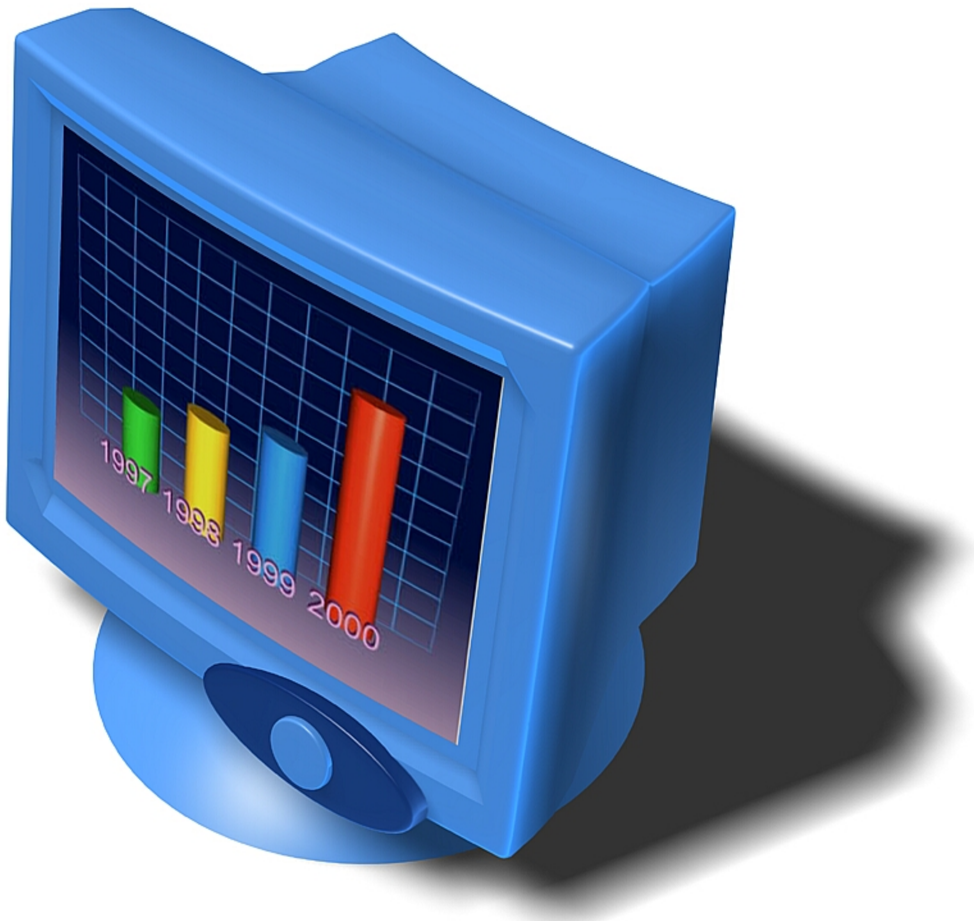


Calibration: Overconfidence and Underconfidence



http://3.bp.blogspot.com/-lmpGS0cqvuw/VDxhem5qTFI/AAAAAAAAACU/qu0hVUn9PBQ/s1600/20141014-Calibration.png

Examples: GeNIe



Concluding remarks

- Reality is a great check on every activity ☺.
- Verification is critical for every model and theory, including models and theories derived from data.
- Statistics is again a guiding light in this respect.
- There are a variety of approaches to verification and testing, cross-validation being the prominent one for data-based analysis.