Fundamental Concepts of Statistics

IDA Spring School 2001

European School on Intelligent Data Analysis

Palermo, March 26-30, 2001

Ad Feelders

Institute of Information
and Computing Sciences

Utrecht University
The Netherlands
ad@cs.uu.nl

Sampling

"You don't have to eat the whole ox to know that the meat is tough."

Basic idea of sampling: to gain information about the whole by examining only a part.

Important terms

- Population: the entire group of objects about which information is wanted.
- Unit: any individual member of the population.
- Sample: a subset of the population to gain information about the whole.
- Variable: a characteristic of a unit, to be measured for those units in the sample.

Types of Samples

A sample should be representative of the larger population to which we want to generalize.

- Convenience sample.

- Judgement sample.

- Probability sample.

  - Simple random sample.

Probability sample

Because we cannot draw conclusions about the population from a sample without error, it is important to know how large these errors may be, and how often incorrect conclusions may occur.

An objective assessment of these errors is only possible for a *probability sample*.

For a probability sample, the probability of inclusion in the sample is *known* and *positive* for each unit in the population.

Sampling: random experiment

Drawing a probability sample of size n from a population consisting of N units, may be a quite complex random experiment.

The experiment is simplified considerably by subdividing it into n experiments, consisting of drawing the n consecutive units.

In a *simple random sample* the n consecutive units are drawn with equal probabilities from the units concerned. In random sampling *with replacement* the subexperiments (drawing of one unit) are all identical and independent: n times a unit is randomly selected from the entire population.

This property simplifies the ensuing analysis considerably!

Sample: random experiment

For units in the sample we observe one or more population variables. For probability samples, each draw is a random experiment. Every observation may therefore be viewed as a random variable. The observation of a population variable $\mathcal{X}$ from the unit drawn in the i^{th} trial, yields a random variable X_i. Observation of the complete sample yields n random variables $X_1, ..., X_n$. Likewise, if we observe for each unit the pair of population variables $(\mathcal{X}, \mathcal{Y})$, we obtain pairs of random variables (X_i, Y_i) with outcomes (x_i, y_i). Consider the population of size $N = 6$:

Unit	1	2	3	4	5	6
$\mathcal{X}$	1	1	2	2	2	3

Sample: random experiment

A random sample of size $n = 2$ is drawn *with replacement* from this population.

For each unit drawn we observe the value of $\mathcal{X}$. This yields two random variables X_1 and X_2, with identical probability distribution.

Unit	1	2	3	4	5	6
$\mathcal{X}$	1	1	2	2	2	3

x	1	2	3
$p_1(x) = p_2(x)$	1/3	1/2	1/6

Furthermore X_1 and X_2 are independent, so their joint distribution equals the product of their individual distributions,

$$p(x_1, x_2) = \prod_{i=1}^{2} p_i(x_i) = [p(x)]^2$$

Distribution of the sample

Probability distribution on each trial

x	1	2	3
$p_1(x) = p_2(x)$	1/3	1/2	1/6

The distribution of the sample

(x_1, x_2)	$p(x_1, x_2)$
(1,1)	1/9
(2,2)	1/4
(3,3)	1/36
(1,2)	1/6
(1,3)	1/18
(2,1)	1/6
(2,3)	1/12
(3,1)	1/18
(3,2)	1/12

Sample statistic

Usually we are not really interested in the individual outcomes of the sample, but rather in some sample statistic.

A statistic is a function of the sample observations $X_1, ..., X_n$, and therefore is itself also a random variable.

Some important sample statistics are the sample mean

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

sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

and sample fraction

$$Fr = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ (for 0-1 variable } \mathcal{X}\text{).}$$

Sample statistics

The distribution of the sample with sample statistics. We listed the values of sample statistics $\bar{x}$ and s^2, for all possible samples of size 2.

(x_1, x_2)	$p(x_1, x_2)$	$\bar{x}$	s^2
(1,1)	1/9	1	0
(2,2)	1/4	2	0
(3,3)	1/36	3	0
(1,2)	1/6	1.5	0.5
(1,3)	1/18	2	2
(2,1)	1/6	1.5	0.5
(2,3)	1/12	2.5	0.5
(3,1)	1/18	2	2
(3,2)	1/12	2.5	0.5

Sampling distribution

The probability distribution of a sample statistic is called its *sampling distribution*. The sampling distribution of $\bar{X}$ and S^2:

$\bar{x}$	$p(\bar{x})$
1	1/9
1.5	1/3
2	13/36
2.5	1/6
3	1/36

s^2	$p(s^2)$
0	14/36
0.5	1/2
2	1/9

Expected value of sample statistics

For a discrete random variable Y, the *expected value* or mean is defined as

$$\mathsf{E}(Y) = \sum_y y\, p(y)$$

Note that

$$\mathsf{E}(\bar{X}) = \frac{11}{6} = \mu$$

where μ denotes the population mean, and

$$\mathsf{E}(S^2) = \frac{17}{36} = \sigma^2$$

where σ^2 denotes the population variance.

Determining sampling distributions

When the sample is of realistic size, and $\mathcal{X}$ takes on many distinct values, complete enumeration is not possible. Nevertheless, we would like to be able to infer something about the shape of the sampling distribution of a sample statistic. We consider here two options to make such inferences.

1. The distribution of X has some standard form that allows the mathematical derivation of the exact sampling distribution.

2. We use a limiting distribution to approximate the sampling distribution of interest. The limiting distribution may be derived from some characteristics of the distribution of X.

Exact sampling distribution

The exact sampling distribution of a sample statistic is often hard to derive analytically, even if the population distribution of $\mathcal{X}$ is known.

As an example we consider the sample statistic $\bar{X}$. The mean and variance of the sampling distribution of $\bar{X}$ are $\mathsf{E}(\bar{X}) = \mu$ and $\mathsf{V}(\bar{X}) = \sigma^2/n$, but its exact shape can only be derived in a few special cases.

For example, if the distribution of $\mathcal{X}$ is $\mathcal{N}(\mu, \sigma^2)$ then the distribution of $\bar{X}$ is $\mathcal{N}(\mu, \sigma^2/n)$.

Limiting distributions

A sequence $\{X_n\}$ of random variables *converges in distribution* to a random variable X with distribution function $F(X)$ if for every point x

$$\lim_{n\to\infty} |\, F_n(x) - F(x)\, | = 0$$

where $F_n(x)$ denotes the distribution function of x_n.

Example: Consider a sequence of random variables $\{X_n\}$ with probability distributions

$$P(x_n = 1) = \frac{1}{2} + \frac{1}{n+1}$$

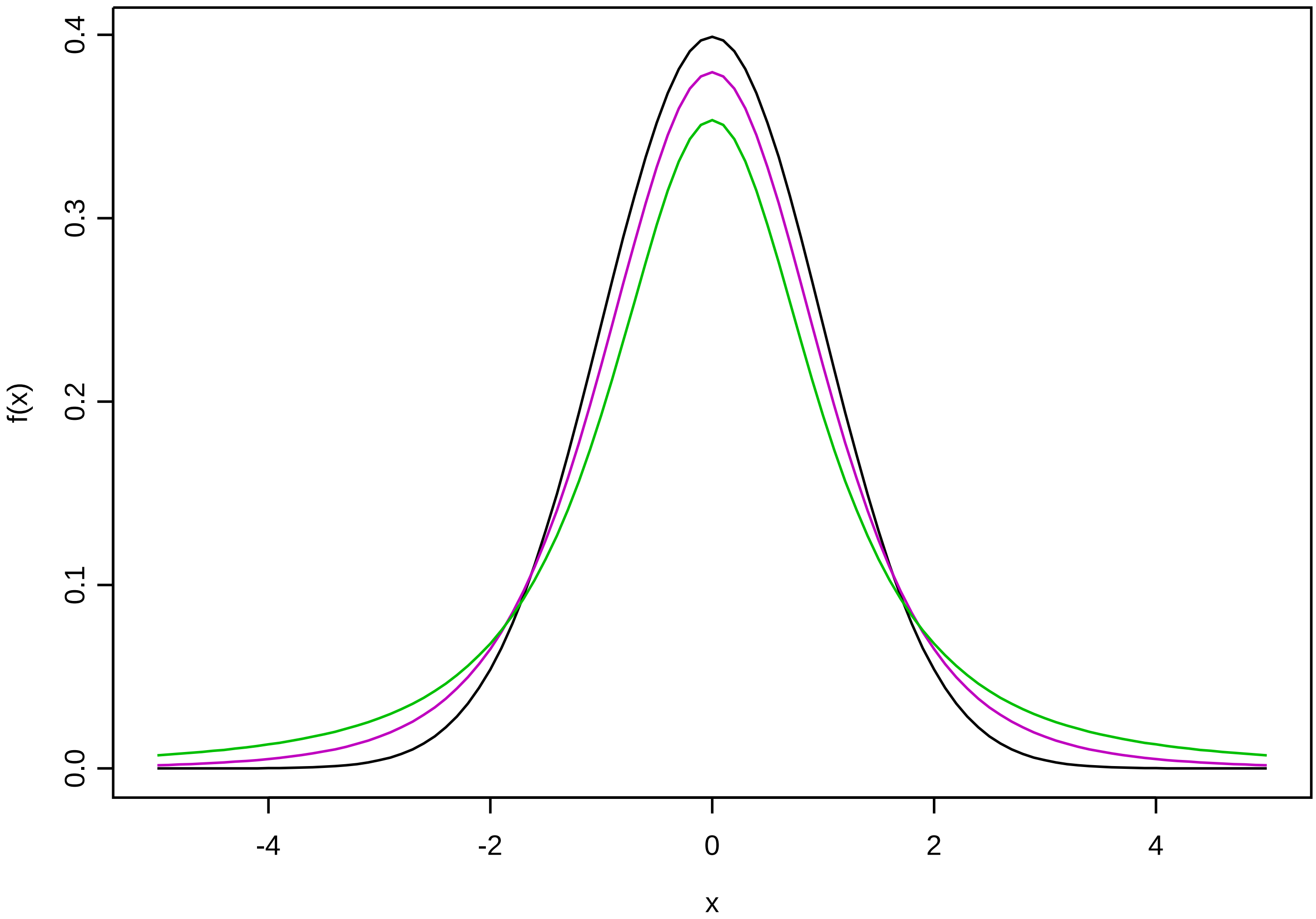
$$P(x_n = 2) = \frac{1}{2} - \frac{1}{n+1},$$

$n = 1, 2, \ldots$
As n increases without bound, the two probabilities converge to 1/2, and

$$P(X = 1) = \frac{1}{2}, P(X = 2) = \frac{1}{2}$$

is called the *limiting distribution* of $\{X_n\}$.

Limiting distribution of t_ν



Central limit theorem

If $(x_1, ..., x_n)$ is a random sample from any probability distribution with finite mean μ and finite variance σ^2, and $\bar{x} = 1/n \sum x_i$ then

$$\frac{(\bar{x} - \mu)}{\sigma/\sqrt{n}} \xrightarrow{D} \mathcal{N}(0, 1)$$

regardless of the form of the parent distribution. In this expression, $\xrightarrow{D}$ denotes convergence in distribution.

This property explains (in part) the importance of the normal distribution in statistical inference.

For extremely skewed distributions $n = 100$ may be required for the normal approximation to be acceptable.

Example: sample fraction

A well-known application of the central limit theorem is the approximation of the distribution of the sample proportion of successes Fr by a normal distribution. Since a success is coded as 1, and failure as 0, the fraction of successes is indeed a mean.

This means the central limit theorem is applicable and as a rule of thumb

$$Fr \approx \mathcal{N}(\pi, \pi(1-\pi)/n)$$

if $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

Even though the exact sampling distribution can be determined in this case, as n becomes large it becomes prohibitively time-consuming to actually calculate this distribution.

Statistical Inference

The relation between sample data and population may be used for reasoning in two directions:

1. From known population to yet to be observed sample data (as discussed so far), and

2. From observed data to (partially) unknown population.

This last direction of reasoning is of inductive nature and is addressed in statistical inference. It is the form of reasoning most relevant to data analysis, since one typically has available one set of sample data from which one intends to draw conclusions about the unknown population.

Statistical Inference Paradigms

Inference should be based on

Frequentist: Sampling distribution.

Likelihood: Likelihood function.

Bayesian: Posterior distribution.

Point Estimation

Provide an estimate for an unknown population parameter, denoted by θ, with *one number*: the point estimate.

If G denotes the estimator of θ, then the estimation error is a random variable $G - \theta$, which should preferably be close to zero.

An important quality measure is the bias

$$\mathsf{B}_\theta = \mathsf{E}_\theta(G - \theta) = \mathsf{E}_\theta(G) - \theta,$$

where expectation is taken with respect to repeated samples from the population. If

$$\mathsf{E}_\theta(G) = \theta$$

then the estimator G is called *unbiased*.

Variance of an estimator

Another important quality measure of an estimator is its variance

$$\mathsf{V}_\theta(G) = \mathsf{E}_\theta(G - \mathsf{E}_\theta(G))^2$$

which measures how much individual estimates g tend to differ from $\mathsf{E}_\theta(G)$, the average value of g over a large number of samples.

Mean Squared Error of an Estimator

An overall quality measure that combines bias and variance is the *mean squared error*

$$\mathsf{M}_\theta(G) = \mathsf{E}_\theta(G - \theta)^2$$

where low values indicate a good estimator. After some algebraic manipulation, we can decompose mean squared error into

$$\mathsf{M}_\theta(G) = \mathsf{B}_\theta^2(G) + \mathsf{V}_\theta(G)$$

that is mean squared error equals squared bias plus variance. It follows that if an estimator is unbiased, then its mean squared error equals its variance.

Interval Estimation

An interval estimator for population parameter θ is an interval of type (G_L, G_U). Two important quality measures for interval estimates are:

$$\mathsf{E}_\theta(G_U - G_L),$$

i.e. the expected width of the interval, and

$$P_\theta(G_L < \theta < G_U),$$

i.e. the probability that the interval contains the true parameter value.

Clearly there is a trade-off between these quality measures.

Confidence intervals

It is customary to choose a confidence level $(1 - \alpha)$ and use an interval estimator such that

$$P_\theta(G_L < \theta < G_U) \geq 1 - \alpha$$

for all possible values of θ.

A realisation (g_L, g_U) of such an interval estimator is called a

$100(1 - \alpha)\%$ *confidence interval*.

Confidence Interval

Example

Estimation of the mean of a normal population with variance σ^2 known, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$.

The distribution of the sample mean for random samples of size n from this population is known to be $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. First $\bar{X}$ is standardized to obtain

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

which allows us to use a table for the standardnormal distribution $Z \sim \mathcal{N}(0, 1)$ to find the relevant probabilities. The probability that $\bar{X}$ is more than one standard error larger than unknown μ is

$$\begin{aligned} P(\bar{X} > \mu + \frac{\sigma}{\sqrt{n}}) &= P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > 1) \\ &= P(Z > 1) = 0.1587 \end{aligned}$$

Confidence interval

But we can *reverse* this reasoning by observing that

$$P(\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu) = 1 - 0.1587 = 0.8413$$

because

$$\bar{X} - \frac{\sigma}{\sqrt{n}} < \mu$$

holds unless

$$\bar{X} > \mu + \frac{\sigma}{\sqrt{n}}$$

Therefore, the probability that the interval

$$(\bar{X} - \sigma/\sqrt{n}, \infty)$$

will contain the true value of μ equals 0.8413. This is called a left-sided confidence interval because it only states a lower bound for μ.

Confidence Interval

In general a $100(1-\alpha)\%$ left-sided confidence interval for μ reads

$$(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty),$$

where $P(Z > z_\alpha) = \alpha$.

Likewise, we may construct a right-sided confidence interval

$$(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{n}})$$

and a two-sided confidence interval

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

Confidence Interval for μ: distribution unknown

If the distribution of X is unknown, i.e. $X \sim \mu, \sigma^2$, then for sufficiently large n we may invoke the central limit theorem and use $\bar{X} \approx \mathcal{N}(\mu, \sigma^2/n)$, and proceed as above.

In most practical estimation problems we don't know the value of σ^2, and we have to estimate it from the data as well. A rather obvious estimator is the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})$$

Now we may use

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} denotes the t-distribution with $n-1$ degrees of freedom.

Hypothesis testing: deterministic case

Hypothesis: Two objects of different weights will fall at the same speed.

Test: Drop two canon balls, one large and one small, from the tower of Pisa.

Outcome: The two canon balls did indeed strike at nearly the same instant.

General structure of this method:

- According to the hypothesis an observable quantity x should have the value x_0.
- If we observe a value of x different from x_0, we must reject the hypothesis.
- The observation that $x = x_0$ serves to support the hypothesis.

Hypothesis Testing: statistical

A statistical test is a procedure to make a choice between two hypotheses concerning the value of a population parameter θ.

One of these, called the *null hypothesis* and denoted by H_0, gets the "benefit of the doubt".

The two possible conclusions are to reject or not to reject H_0. H_0 is only rejected if the sample data contains strong evidence that it is not true.

We use a test statistic G to make a decision. Usually H_0 will assign positive probability to all possible values of G.

Type I error and Type II error

The null hypothesis is rejected if realisation g of test statistic G is in the *critical region* denoted by C.

In doing so we can make two kinds of errors

Type I error: Reject H_0 when it is true.

Type II error: Accept H_0 when it is false.

Type I errors are considered to be more serious than Type II errors.

Test statistic G is usually a point estimator for θ, e.g. if we test a hypothesis concerning the value of population mean μ, then $\bar{X}$ is an obvious choice of test statistic.

Example: left one-sided test

$$H_0 : \theta \geq \theta_0 \ , \ H_a : \theta < \theta_0$$

The highest value of G that leads to the rejection of H_0 is called the critical value c_u, it is the upper bound of the so-called critical region $C = (-\infty, c_u]$. All values of G to the left of c_u lead to the rejection of H_0, so this is called a left one-sided test. An overall quality measure for a test is its power β

$$\beta(\theta) = P_\theta(\text{Reject } H_0) = P_\theta(G \in C)$$

Because we would like a low probability of Type I and Type II errors, we like to have $\beta(\theta)$ small for $\theta \in H_0$ and $\beta(\theta)$ large for $\theta \in H_a$. It is common practice in hypothesis testing to restrict the probability of a Type I error to a maximum called the *significance level* α of the test, i.e.

$$\max_{\theta \in H_0} \beta(\theta) \leq \alpha$$

Example: left one-sided test

Since the maximum is reached for $\theta = \theta_0$ this reduces to the restriction $\beta(\theta_0) \leq \alpha$. If possible the test is performed in such a way that $\beta(\theta_0) = \alpha$ (This may not be possible for discrete sampling distributions). Common levels for α are 0.1, 0.05 and 0.01. If in a specific application of the test, the conclusion is that H_0 should be rejected, then the result is called *significant*.

Test on μ with σ^2 known

Consider a left one-sided test on population mean μ with $X \sim \mathcal{N}(\mu, \sigma^2)$ and the value of σ^2 known. That is

$$H_0 : \mu \geq \mu_0 \ , \ H_a : \mu < \mu_0$$

We determine the sampling distribution of the test statistic $\bar{X}$ under the assumption that the $\mu = \mu_0$, i.e. $\bar{X} \sim \mathcal{N}(\mu_0, \sigma^2/n)$. Now

$$\begin{aligned} \alpha = P_{\mu_0}(\bar{X} \leq c_u) &= P(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{c_u - \mu_0}{\sigma/\sqrt{n}}) \\ &= P(Z \leq \frac{c_u - \mu_0}{\sigma/\sqrt{n}}) \end{aligned}$$

and since $P(Z \leq -z_\alpha) = \alpha$, we obtain

$\dfrac{c_u - \mu_0}{\sigma/\sqrt{n}} = -z_\alpha$, and therefore $c_u = \mu_0 - z_\alpha \dfrac{\sigma}{\sqrt{n}}$

Example

Consider a random sample of size $n = 25$ from a normal population with known $\sigma = 5.4$ and unknown mean μ. The observed sample mean is $\bar{x} = 128$. We want to test the hypothesis

$$H_0 : \mu \geq 130, \text{ against } \; H_a : \mu < 130$$

i.e. $\mu_0 = 130$. The significance level of the test is set to $\alpha = 0.05$. We compute the critical value

$$c_u = \mu_0 - z_{0.05} \frac{\sigma}{\sqrt{n}} = 130 - 1.645 \frac{5.4}{\sqrt{25}} = 128.22$$

where $z_{0.05} = 1.645$ was determined using a statistical package.

So the critical region is $(-\infty, 128.22]$ and since $\bar{x} = 128$ is in the critical region, we reject H_0.

Testing: unknown population distribution and variance

We may invoke the central limit theorem if

$X \sim \mu, \sigma^2$ and n is large.

Furthermore, if σ^2 is unknown, we have to estimate it from the data and use a t_{n-1} distribution rather than the standardnormal distribution to determine the critical region.

Observed significance

Sometimes one doesn't want to specify the significance level α of the test in advance. In that case it us customary to report so-called p-values, indicating the *observed significance*.

Example:

The p-value of the observed outcome $\bar{x} = 128$ is

$$\begin{aligned} P_{\mu_0}(\bar{X} \leq 128) &= P(Z \leq \frac{128 - \mu_0}{\sigma/\sqrt{n}}) \\ &= P(Z \leq -1.852) = 0.0322 \end{aligned}$$

Since the p-value is 0.0322, we would reject H_0 at $\alpha = 0.05$, but we would accept H_0 at $\alpha = 0.01$.

Probability and Likelihood

y	π				
	0.1	0.3	0.5	0.7	0.9
0	.349	.028	.001		
1	.387	.121	.01		
2	.194	.234	.044	.002	
3	.057	.267	.117	.009	
4	.011	.2	.205	.036	
5	.002	.103	.246	.103	.002
6		.036	.205	.2	.011
7		.009	.117	.267	.057
8		.002	.044	.234	.194
9			.01	.121	.387
10			.001	.028	.349
	1	1	1	1	1

Probability distribution and likelihood function

Binomial distribution with parameters n and π:

$$P(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}$$

Binomial distribution with $n = 10$ and $\pi = 0.6$:

$$P(y) = \binom{10}{y} 0.6^y \, 0.4^{10-y}$$

Probability of observing $y = 7$:

$$\binom{10}{7} 0.6^7 0.4^3 \approx 0.215$$

Likelihood function if we observe 7 heads:

$$L(\pi \mid y = 7) = \binom{10}{7} \pi^7 (1-\pi)^3$$

Likelihood function

Let

$$\mathbf{y} = (y_1, \ldots, y_n)$$

be n independent observations, all from the same probability density

$$f(y \,|\, \theta),$$

where θ is the parameter vector of f, then

$$L(\theta \,|\, \mathbf{y}) \propto \prod_{i=1}^{n} f(y_i | \, \theta)$$

is the likelihood function for $\mathbf{y}$.

Maximum likelihood estimation:

Find that particular $\widehat{\theta}$ which maximizes L, i.e. that $\widehat{\theta}$ such that the observed $\mathbf{y}$ are more likely to have come from $f(y \,|\, \widehat{\theta})$ than from $f(y \,|\, \theta)$ for any other value of θ.

Maximum Likelihood

Take the derivatives of L with respect to the components of θ and equate them to zero (*normal equations*)

$$\frac{\partial L}{\partial \theta_j} = 0$$

Solve for the θ_j.

Maximizing the log-likelihood function is often easier

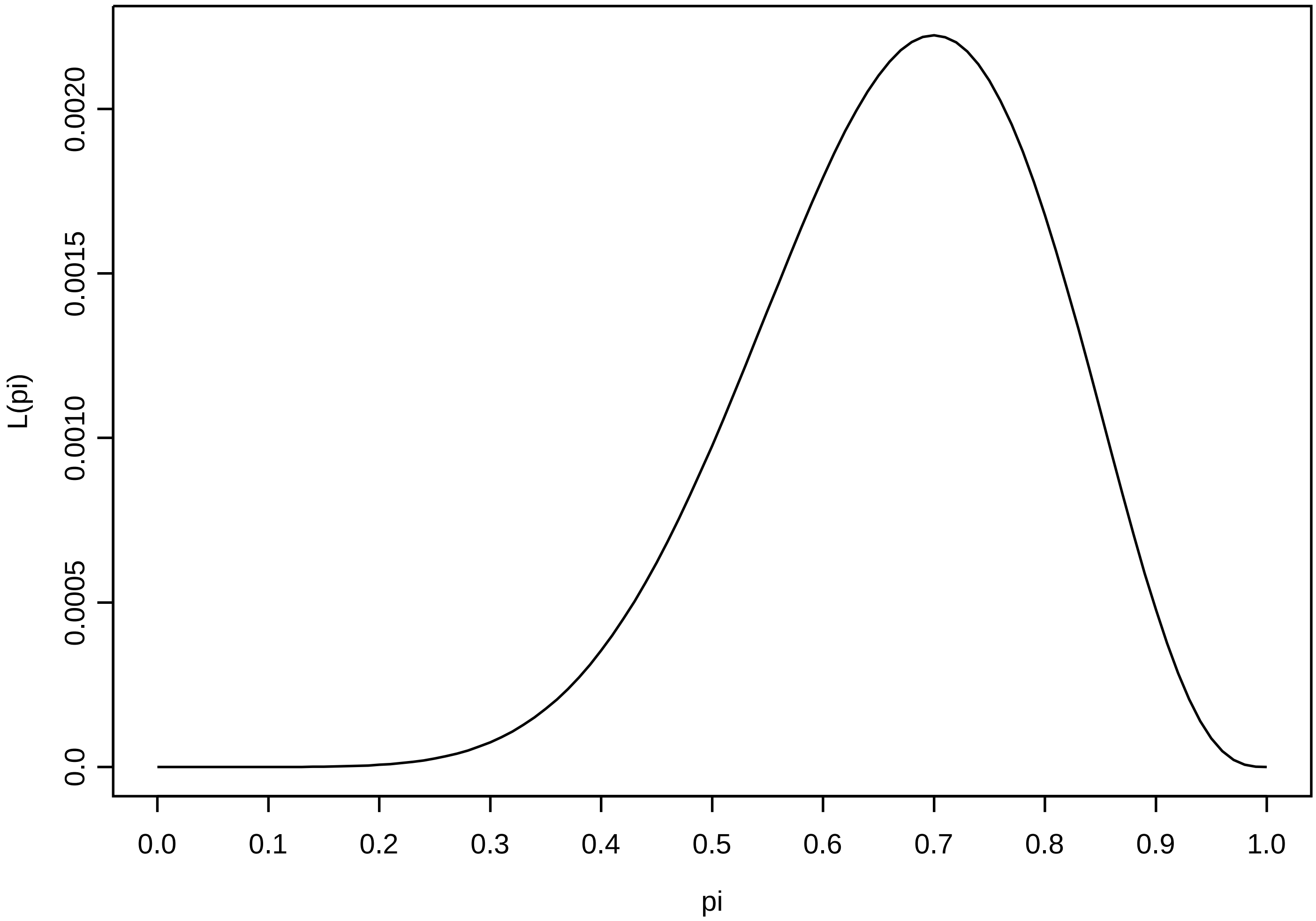
$$\begin{aligned} l(\theta \mid \mathbf{y}) = \log(L(\theta \mid \mathbf{y})) &= \log(\prod_{i=1}^{n} f(y_i \mid \theta)) \\ &= \sum_{i=1}^{n} \log f(y_i \mid \theta) \end{aligned}$$

since $\log ab = \log a + \log b$.

This is possible because the log function is strictly increasing on $(0, \infty)$.

Likelihood function

Likelihood function for 7 heads out of 10



Example: coin flipping

Random variable Y with $y = 1$ if heads comes up, and $y = 0$ if tails comes up. Probability distribution for one coin flip

$$f(y) = \pi^y (1 - \pi)^{1-y}$$

Sequence of n coin flips

$$f(\mathbf{y}) = f(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} \pi^{y_i} (1 - \pi)^{1-y_i}$$

which defines the likelihood when viewed as a function of π. The log-likelihood consequently becomes

$$l(\pi \mid \mathbf{y}) = \sum_{i=1}^{n} y_i \ln(\pi) + (1 - y_i) \ln(1 - \pi)$$

since $\log a^b = b \log a$.

Example: coin flipping (continued)

In a sequence of 10 coin flips with seven times heads coming up, we obtain

$$l(\pi) = \ln(\pi^7(1-\pi)^3) = 7\ln\pi + 3\ln(1-\pi)$$

To determine the maximum we take the derivative with respect to π

$$\frac{dl}{d\pi} = \frac{7}{\pi} - \frac{3}{1-\pi} = 0$$

which yields maximum likelihood estimate $\widehat{\pi} = 0.7$.

Note: the derivative of $\ln x = \frac{1}{x}$.

Information

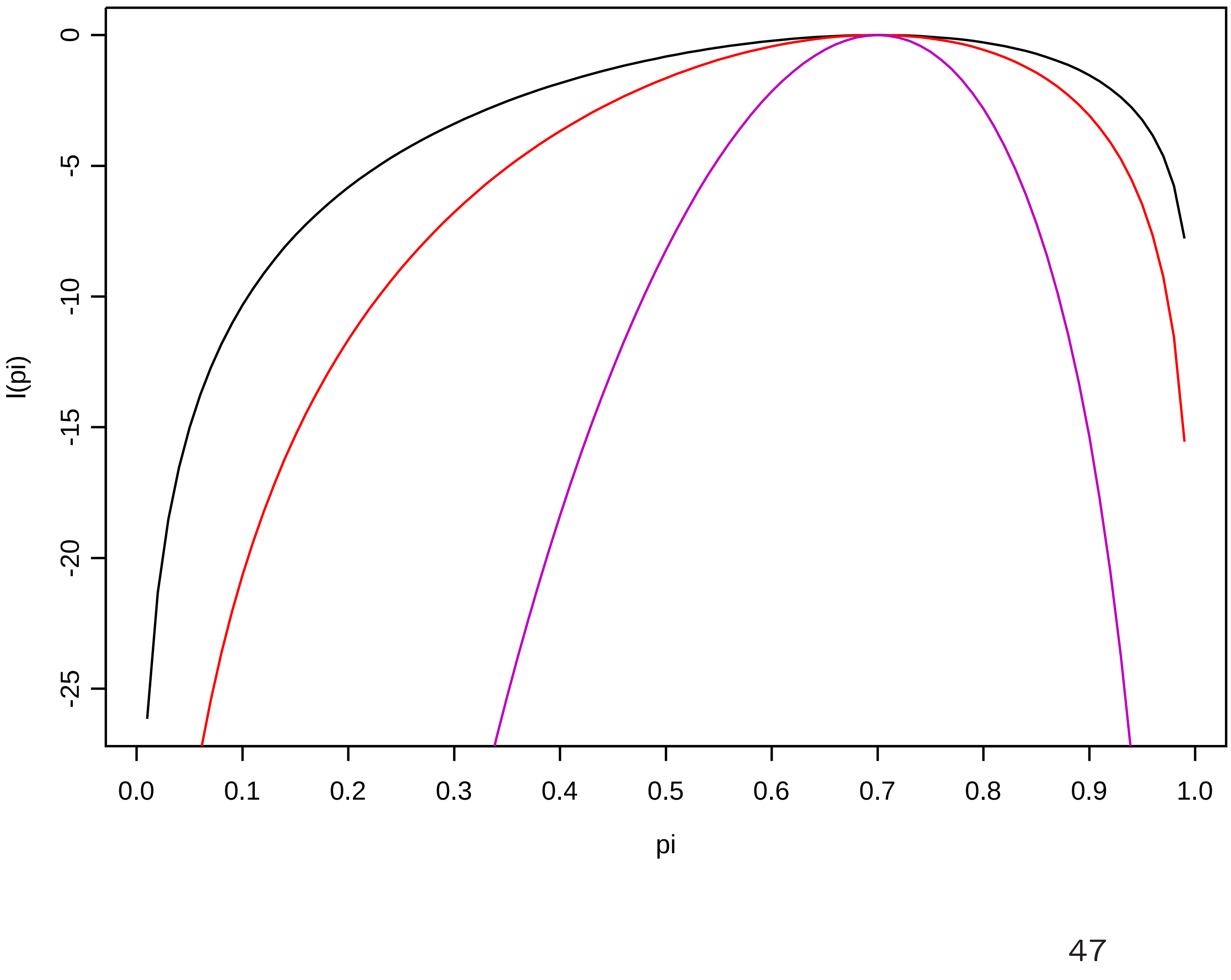
Another important aspect of the log-likelihood function is its shape in the region near the maximum.

If it is rather flat, one could say that the likelihood contains little information in the sense that there are many values of θ with log-likelihood near that of $\widehat{\theta}$.

If, on the other hand, it is rather steep, one could say that the log-likelihood contains much information about $\widehat{\theta}$.

Information

Log-likelihood function for 7 heads out of 10, 14 heads out of 20 and 70 heads out of 100.



Observed information

Minus the second derivative of the log-likelihood function is known as the *information*. When evaluated at $\widehat{\theta}$ (the maximum likelihood estimate of θ) it is called the *observed information*.

Example

Consider a sequence of n coin tosses, with heads coming up y times. We are interested in the probability of heads π. We have seen that

$$l(\pi) = y \ln(\pi) + (n-y)\ln(1-\pi)$$

Setting the first derivative to zero and solving for π yields $\widehat{\pi} = y/n$. The information is

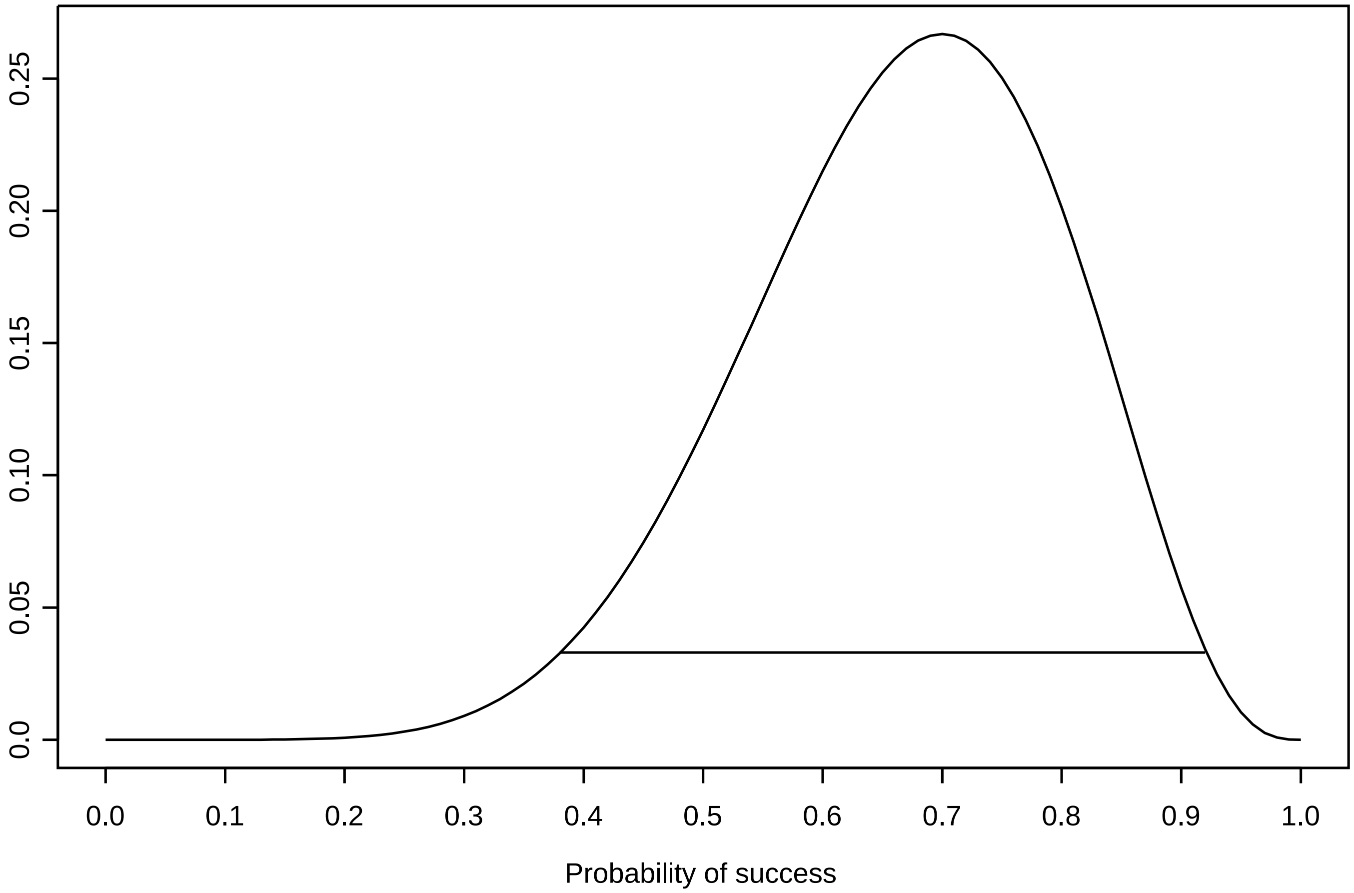
$$-\frac{d^2 l}{d\pi^2} = \frac{y}{\pi^2} + \frac{(n-y)}{(1-\pi)^2}$$

Evaluating this expression at $\widehat{\pi} = y/n$ we obtain the observed information

$$\frac{n}{\widehat{\pi}(1-\widehat{\pi})}.$$

Likelihood interval

The horizontal line indicates the range of values of π for which the ratio of $L(\pi)$ to the maximum $L(0.7)$ is greater than 1/8. The 1/8 likelihood interval is approximately $(0.38, 0.92)$.



Bayesian Inference

Consider these three experiments

Tea-drinking lady: claims to be able to tell whether milk or tea was first poured into the cup.

Music expert: claims he can distinguish a page of Haydn score from a page of Mozart.

A drunk friend: claims he can predict the outcome when a coin is spun.

Each of them makes ten correct predictions in ten trials.

Who would you believe?

Bayesian Inference

Consider again the coin tossing experiment. We stated that the probability of heads, denoted by π, is a fixed yet unknown quantity. From a relative frequency viewpoint, it makes no sense to talk about the probability distribution of π since it is not a random variable. In Bayesian inference one departs from this strict interpretation of probability. We may express prior, yet incomplete, knowledge concerning the value of π through the construction of a *prior distribution*.

This prior distribution is then combined with sample data to obtain a posterior distribution. The posterior distribution expresses the new state of knowledge, in light of the sample data.

Bayes' rule

Let $B_1, B_2, \ldots, B_s$ be a partition of Ω. Bayes' rule follows from the axioms of probability

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

In Bayesian inference Bayes' rule is used as follows:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_j P(D|M_j)P(M_j)}$$

where the M_i specify different models for the data D. The models are hypotheses concerning the parameter value(s) of the probability distribution from which the data were drawn.

Example: Bayes' rule

Two hypotheses concerning the probability of heads of a particular coin are entertained, namely

1. M_1: $\pi = 0.8$
2. M_2: $\pi = 0.4$.

We observe 5 heads in a sequence of 10 coin flips.

	Prior $P(M_i)$	Likelihood $P(y = 5 \mid M_i)$	Posterior $P(M_i \mid y = 5)$
M_1: $\pi = 0.8$	0.7	0.027	0.239
M_2: $\pi = 0.4$	0.3	0.201	0.761

Bayes' rule: density version

Replace probabilities by probability densities and summation by integration:

$$f(\theta \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \theta)\ f(\theta)}{\int_\Omega f(\mathbf{y} \,|\, \theta)\ f(\theta)\ d\theta}$$

where $\mathbf{y}$ denotes the observed data and Ω denotes the parameter space, i.e. the space of possible values of θ.

Prior ignorance

Consider the case where we have no prior knowledge whatsoever concerning the probability of heads π. How should this be reflected in the prior distribution? One way of reasoning is to say that all values of π are considered equally likely, which can be expressed by a uniform distribution over $\Omega = [0, 1]$: the range of possible values of π.

Posterior distribution:

$$f(\pi \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \pi) f(\pi)}{\int_0^1 f(\mathbf{y} \,|\, \pi) f(\pi) d\pi}$$

and since $f(\pi) = 1$ (uniform on $[0, 1]$)

$$f(\pi \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \pi)}{\int_0^1 f(\mathbf{y} \,|\, \pi) d\pi}$$

Prior ignorance:example

If we observe once again 7 times heads in a sequence of 10 coin flips, then

$$f(\mathbf{y} \mid \pi) = \pi^7 (1-\pi)^3$$

Since $f(\pi) = 1$, the denominator becomes

$$\int_0^1 \pi^7 (1-\pi)^3 d\pi = \frac{1}{1320}$$

and so the posterior density becomes

$$f(\pi \mid \mathbf{y}) = 1320\, \pi^7 (1-\pi)^3$$

It is reassuring to see that in case of prior ignorance the posterior distribution is proportional to the likelihood function of the observed sample.

Conjugate priors

In general, the computationally most difficult part of obtaining the posterior distribution is the evaluation of the (multiple) integral in the denominator of

$$f(\theta \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \theta)\ f(\theta)}{\int_\Omega f(\mathbf{y} \,|\, \theta)\ f(\theta)\ d\theta}$$

For this reason, a particular class of priors, called conjugate priors, have received special attention in Bayesian statistics.

Conjugate priors allow for simple rules to update the prior with sample data to arrive at the posterior distribution.

Furthermore, the posterior distribution belongs to the same family as the prior distribution.

Conjugate priors: example

Assume our prior knowledge concerning the value of π may be expressed by a Beta(4,6) distribution

$$f(\pi) = \frac{\pi^3(1-\pi)^5}{\int_0^1 \pi^3(1-\pi)^5 d\pi}$$

Since $\int_0^1 \pi^3(1-\pi)^5 d\pi = \frac{1}{504}$, we get $f(\pi) = 504\,\pi^3(1-\pi)^5$.

Multiplied with the likelihood this results in $504\,\pi^3(1-\pi)^5\pi^7(1-\pi)^3 = 504\pi^{10}(1-\pi)^8$, so the denominator becomes

$$\int_0^1 504\,\pi^{10}(1-\pi)^8 = \frac{28}{46189}$$

and the posterior density becomes

$$f(\pi \mid \mathbf{y}) = 831402\,\pi^{10}(1-\pi)^8$$

which is a Beta(11,9) distribution.

Update rule: Beta prior and Binomial sample

In general, when we have a binomial sample of size n with r successes ($n - r$ failures), and we combine that with a prior distribution

$$\text{Beta}(l, k),$$

then the posterior distribution is

$$\text{Beta}(l + r, k + n - r).$$

Since the uniform distribution over the interval $[0, 1]$ is the same as a Beta(1,1) distribution, we could have used this simple update rule in the "prior ignorance" case as well: combining a Beta(1,1) prior with a binomial sample of size 10 with 7 successes yields a Beta(8,4) posterior distribution.

Extracting information from the posterior distribution

Once we have calculated the posterior distribution, we can extract all kinds of information from it.

We may for example determine the mode of the posterior distribution which represents the value of π for which the posterior density is maximal. When asked to give a point estimate for π, it makes sense to report this value.

When asked for a range of plausible values for π we may use the posterior distribution to determine a so-called $100(1-\alpha)\%$ probability interval, which is an interval $[g_l, g_u]$ such that $P(\pi < g_l) = \alpha/2$ and $P(\pi > g_u) = \alpha/2$ where the relevant probabilities are based on the posterior distribution for π.

Prediction

The value of a random variable Y depends on the outcome of a random experiment. Before the experiment is performed, the value of the random variable is unknown. In many cases, we would like to *predict* the value of a yet to be observed random variable. The usual assumption is that the distribution of Y depends in some way on some random vector $\mathbf{X} = (X_1, ..., X_p)$

For example, in the *simple linear regression model*, the assumption is that

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma_\epsilon^2),$$

i.e. Y is normally distributed with mean a linear function of x, and constant variance σ_ϵ^2.

Prediction

Often, the goal of data analysis is to gain information from a training sample

$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), ..., (\mathbf{x}_n, y_n)\}$$

in order to estimate the parameters of such a model.

Once the model is estimated, we can use it to make predictions about random variable Y.

If Y is numeric, we speak about *regression*.

If Y takes its values from a discrete unordered set we speak about *classification*.

Model Specification

How does the distribution of Y (or $E(Y)$), depend on $\mathbf{x}$?

Linear regression model is quite restrictive. If the true relationship is not linear, the estimated model will produce some prediction error due to the unwarranted assumption of linearity.

If we make very mild assumptions, basically allowing the relation between E(Y) and $\mathbf{x}$ to be chosen from a very large class of functions, the estimated function is capable to adapt very well to the data. In fact so well that it will also model the peculiarities that are due to random variations. The estimate becomes highly sensitive to the particular sample drawn, that is upon repeated sampling it will exhibit high variance.

Model Specification: example

Suppose that $Y_i \sim \mathcal{N}(\mu = 2.0 + 0.5x_i, \sigma_\varepsilon^2 = 1)$, i.e. the true relation between E(Y) and x is

$$\mathsf{E}(Y) = 2.0 + 0.5x.$$

We have a sample T of ten (x, y) observations. Note that x is not a random variable but its values are chosen by us to be $1, 2, \ldots, 10$.

Model Specification

Although $\mathsf{E}(Y)$ is a linear function of x, the observations do not lie on a straight line due to the inherent variability of Y. We pretend we don't know the relation between x and y, but only have T at our disposal, as would be the case in most data analysis settings. We consider three classes of models to describe the relation between x and y.

Linear Model:

$$\mathsf{E}(Y) = f_1(x) = \beta_0 + \beta_1 x$$

Quadratic Model:

$$\mathsf{E}(Y) = f_2(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Cubic Model:

$$\mathsf{E}(Y) = f_3(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

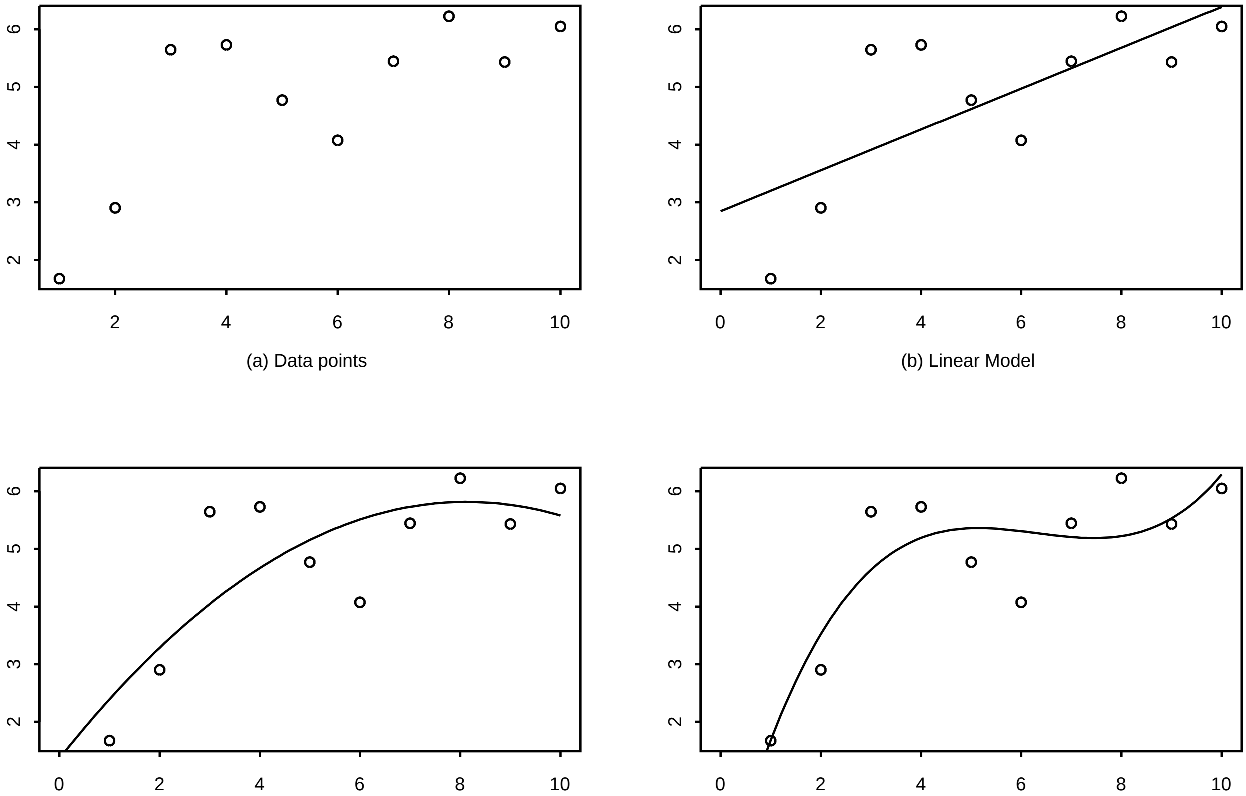
Estimation by least squares

The β_j are the parameters of the model, whose estimates are chosen in such a way that the sum of squared *vertical* distances from the points (x_i, y_i) to the fitted equation is minimized. For example, for the simple linear regression model we choose the estimates of β_0 and β_1 such that

$$\sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 \, x_i)]^2$$

is minimal. The expression $\hat{\beta}_0 + \hat{\beta}_1 \, x_i$ denotes the predicted value for y_i, so one effectively minimizes the sum of squared differences between predicted values and realisations of y. The estimates $\hat{\beta}_j$ of the β_j thus obtained are called the *least squares* estimates.

Training data and fitted equations



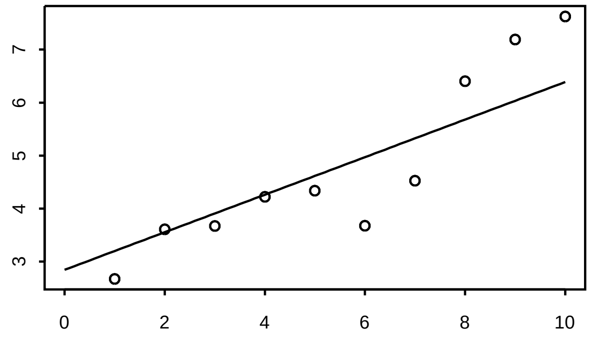
(a) Data points

(b) Linear Model

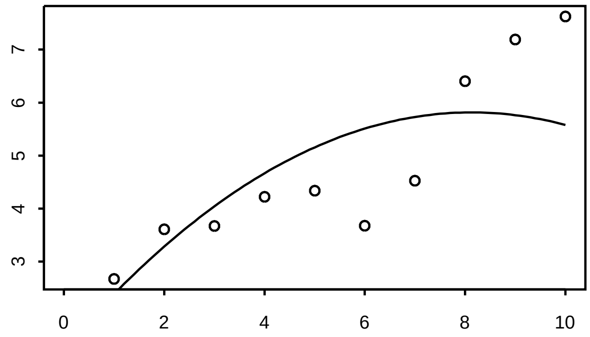
(c) Quadratic Model

(d) Cubic Model

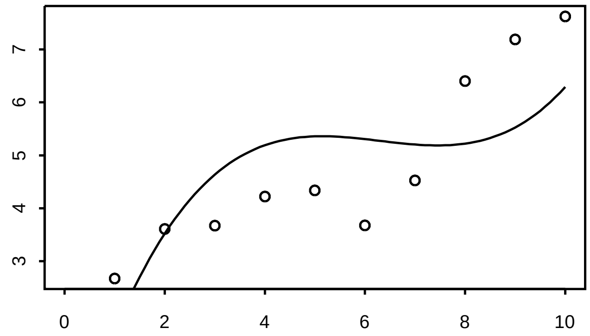
Test data and fitted equations



(a) Linear Model



(b) Quadratic Model



(c) Qubic Model

Prediction error in regression

Once we have obtained estimates $\widehat{\beta}_j$ by estimation from some training set T, we may use the resulting function to make predictions of y when we know the corresponding value of $\mathbf{x}$.

Henceforth we denote this prediction by $\widehat{f}(\mathbf{x})$. The difference between prediction $\widehat{f}(\mathbf{x})$ and realisation y is called prediction error. It should preferrably take values close to zero. A natural quality measure of $\widehat{f}$ as a predictor of Y is the mean squared prediction error.

For fixed T and $\mathbf{x}$

$$\mathsf{E}\left[(Y - \widehat{f}(\mathbf{x} \,|\, T))^2\right]$$

where the expectation is taken with respect to $p(Y \,|\, \mathbf{x})$, the probability distribution of Y at $\mathbf{x}$.

Error decomposition: reducible and irreducible

We may decompose this overall error into a *reducible* part, and an *irreducible* part that is due to the variability of Y at $\mathbf{x}$, as follows

$$\mathsf{E}\,[(Y-\hat{f}(\mathbf{x}\,|\,T))^2] = [f(\mathbf{x})-\hat{f}(\mathbf{x}\,|\,T)]^2+\mathsf{E}[(y-f(\mathbf{x}))^2]$$

where $f(\mathbf{x}) \equiv \mathsf{E}[Y\,|\,\mathbf{x}]$.

The last term in this expression is the mean square error of the best possible (in the mean squared error sense) prediction $\mathsf{E}[Y\,|\,\mathbf{x}]$.

Decomposition of reducible error

To say something about the quality of the estimator $\widehat{f}$, we take its expectation over all possible training samples (of fixed size) and decompose it into its bias and variance components:

$$\begin{aligned} \mathsf{E}_T[(f(\mathbf{x}) - \widehat{f}(\mathbf{x}\,|\,T))^2] &= (f(\mathbf{x}) - \mathsf{E}_T[\widehat{f}(\mathbf{x}\,|\,T)])^2 + \\ &\quad \mathsf{E}_T[(\widehat{f}(\mathbf{x}\,|\,T) - \mathsf{E}_T[\widehat{f}(\mathbf{x}\,|\,T)])^2] \end{aligned}$$

The first component is the squared bias, where bias is the difference between the best prediction $f(\mathbf{x})$ and its average estimate over all possible samples of fixed size. The second component, variance, is the expected squared difference between an estimate obtained for a single training sample and the average estimate obtained over all possible samples.

Bias-Variance decomposition: example

We illustrate these concepts by a simple simulation study using the models introduced in the previous section. The expectations in the above decomposition are taken over *all possible* training samples, but this is a little bit to much to compute. Instead we use the computer to draw a large number of random samples to obtain an estimate of the desired quantities. In the simulation we sampled 1000 times from

$$Y_i \sim \mathcal{N}(\mu = 2 + 0.5x_i, \sigma_\varepsilon^2 = 1)$$

with $x_i = 1, 2, \ldots, 10$. In other words we generated 1000 random samples, $T_1, T_2, \ldots, T_{1000}$ each consisting of 10 (x, y) pairs. For each T_k, the least squares parameter estimates for the three models were computed. Using the estimated models we computed the predicted values $\hat{f}(x)$. From the 1000 predicted values we computed the mean to estimate the expected value $\mathsf{E}(\hat{f}(x))$ and variance to estimate $\mathsf{V}(\hat{f}(x))$.

Results of simulation

Expected value and variance of $\widehat{f}_j$

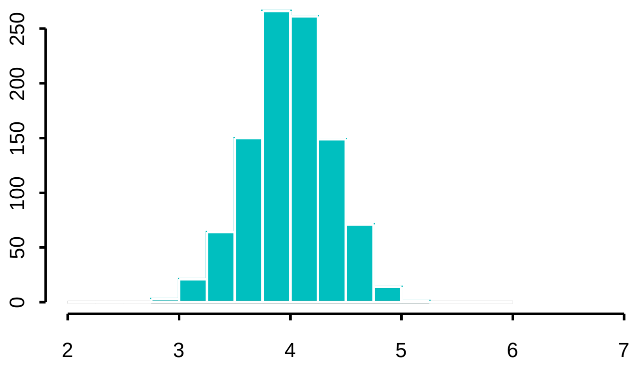
x	$f(x)$	$\mathsf{E}(\widehat{f}_1)$	$\mathsf{E}(\widehat{f}_2)$	$\mathsf{E}(\widehat{f}_3)$	$\mathsf{V}(\widehat{f}_1)$	$\mathsf{V}(\widehat{f}_2)$	$\mathsf{V}(\widehat{f}_3)$
1	2.50	2.48	2.48	2.49	0.34	0.61	0.84
2	3.00	2.99	2.98	2.98	0.25	0.27	0.29
3	3.50	3.49	3.49	3.48	0.18	0.18	0.33
4	4.00	3.99	4.00	3.99	0.13	0.20	0.32
5	4.50	4.50	4.50	4.50	0.10	0.23	0.25
6	5.00	5.00	5.00	5.01	0.10	0.22	0.23
7	5.50	5.50	5.51	5.52	0.13	0.19	0.28
8	6.00	6.01	6.01	6.02	0.17	0.18	0.31
9	6.50	6.51	6.51	6.51	0.24	0.28	0.30
10	7.00	7.01	7.01	7.00	0.33	0.62	0.80

Example

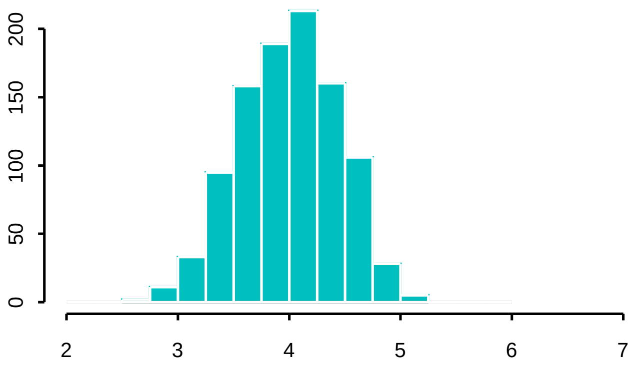
The fourth row contains the simulation results of the predictions of the different models for $x = 4$. The expected value is

$$f(4) = \mathsf{E}(Y|x = 4) = 2 + 0.5 \cdot 4 = 4.$$

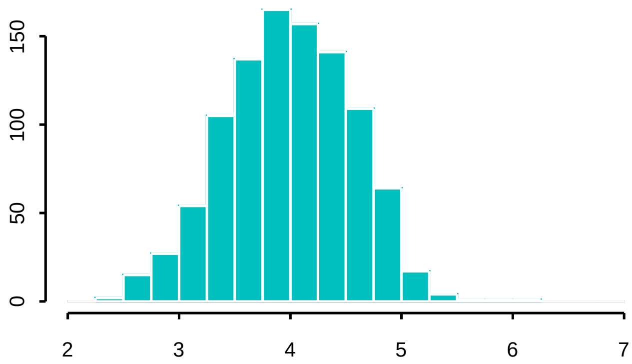
Histograms of $\widehat{f}_j(4)$ based on 1000 samples



(a) Linear Model



(b) Quadratic Model



(c) Cubic Model

Variance goes down with sample size

For a fixed bias, the variance tends to decrease when the training sample gets larger and larger. Consequently, for very large training samples, bias tends to be the most important source of prediction error.

We generated the training sets by drawing from

$$Y_i \sim \mathcal{N}(\mu = 2 + 0.5x_i + 0.02x_i^2, \sigma_\varepsilon^2 = 1).$$

	Squared bias			Variance		
n	10	100	1000	10	100	1000
Linear (f_1)	.021	.022	.022	.197	.022	.002
Quadratic (f_2)	.000	.000	.000	.299	.037	.004
Cubic (f_3)	.001	.000	.000	.401	.054	.006

	Mean square error		
n	10	100	1000
Linear (f_1)	.218	.043	.024
Quadratic (f_2)	.299	.037	.004
Cubic (f_3)	.401	.054	.006

Prediction error in classification

In classification y assumes values on an unordered discrete set $y \in \{y_1, ..., y_L\}$. For ease of exposition we assume $L = 2$, and $y \in \{0, 1\}$. We now have

$$\mathsf{E}[y \,|\, \mathbf{x}] \equiv f(\mathbf{x}) = P(y = 1 | \, \mathbf{x}) = 1 - P(y = 0 \,|\, \mathbf{x})$$

The role of a classification procedure is to construct a rule that makes a prediction $\widehat{y}(\mathbf{x}) \in \{0, 1\}$ for the class label y at every input point $\mathbf{x}$.

Training data T are again used to form an estimate $\widehat{f}(\mathbf{x} \,|\, T)$ of $f(\mathbf{x})$ to construct an allocation rule:

$$\widehat{y}(\mathbf{x} \,|\, T) = I(\widehat{f}(\mathbf{x} \,|\, T) \geq 1/2)$$

Prediction error in classification

For classification it is customary to measure prediction error in terms of the error rate $P(\hat{y} \neq y)$. The best possible allocation rule, in terms of error rate, is the so called Bayes rule:

$$y_B(\mathbf{x}) = I(f(\mathbf{x}) \geq 1/2)$$

where $I(\cdot)$ is an indicator function of the truth of its argument. This rule simply states that $\mathbf{x}$ should be allocated to the class that has highest probability at $\mathbf{x}$. The probability of error

$$P(y_B \neq y)$$

of the Bayes rule represents the irreducible error rate, analogous to

$$\mathsf{E}[(y - f(\mathbf{x}))^2]$$

in regression.

Friedman's error decomposition

Friedman proposes the following decomposition of the error rate of an allocation rule into a reducible and irreducible part

$$\begin{aligned} P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) &= |2f(\mathbf{x}) - 1| P(\hat{y}(\mathbf{x}) \neq y_B(\mathbf{x})) + \\ &\quad P(y_B(\mathbf{x}) \neq y(\mathbf{x})) \end{aligned}$$

Note that the reducible error is expressed as a *product* of two terms, rather than a *sum*, in the above equation. The first term of the product denotes the increase in error if $\hat{f}$ is on the wrong side of 1/2, and the second term of the product denotes the probability of this event.

Friedman's error decomposition: example

Suppose

$$P(y = 1|\mathbf{x}) = f(\mathbf{x}) = 0.8, \text{and } \mathrm{E}(\hat{f}(\mathbf{x})) = 0.78.$$

Since $f(\mathbf{x}) \geq 1/2$, Bayes rule allocates $\mathbf{x}$ to class 1, that is $y_B(\mathbf{x}) = 1$. The error rate of Bayes rule is

$$P(y = 0|\mathbf{x}) = 1 - 0.8 = 0.2.$$

Now $\hat{y}(\mathbf{x})$ differs from Bayes rule if it is below 1/2. Assume the distribution $p(\hat{f}(\mathbf{x}))$ of $\hat{f}(\mathbf{x})$ is such that $P(\hat{f}(\mathbf{x}) < 1/2) = 0.1$, and consequently $P(\hat{y}(\mathbf{x}) \neq y_B(\mathbf{x})) = 0.1$. We can now compute the error rate as follows

$$\begin{aligned} P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) &= P(\hat{f}(\mathbf{x}) \geq 1/2)P(y = 0|\mathbf{x}) + \\ &\quad P(\hat{f}(\mathbf{x}) < 1/2)P(y = 1|\mathbf{x}) \\ &= 0.9 \cdot 0.2 + 0.1 \cdot 0.8 = 0.26 \end{aligned}$$

Computation using Friedman's decomposition

Irreducibe error

$$P(y_B(\mathbf{x}) \neq y(\mathbf{x})) = 0.2.$$

We assumed

$$P(\hat{y}(\mathbf{x}) \neq y_B(\mathbf{x})) = 0.1.$$

If $\hat{f}(\mathbf{x}) < 1/2$ the error rate increases from 0.2 to 0.8, an increase of 0.6 that is represented by the term

$$|2f(\mathbf{x}) - 1| = 2 \cdot 0.8 - 1 = 0.6$$

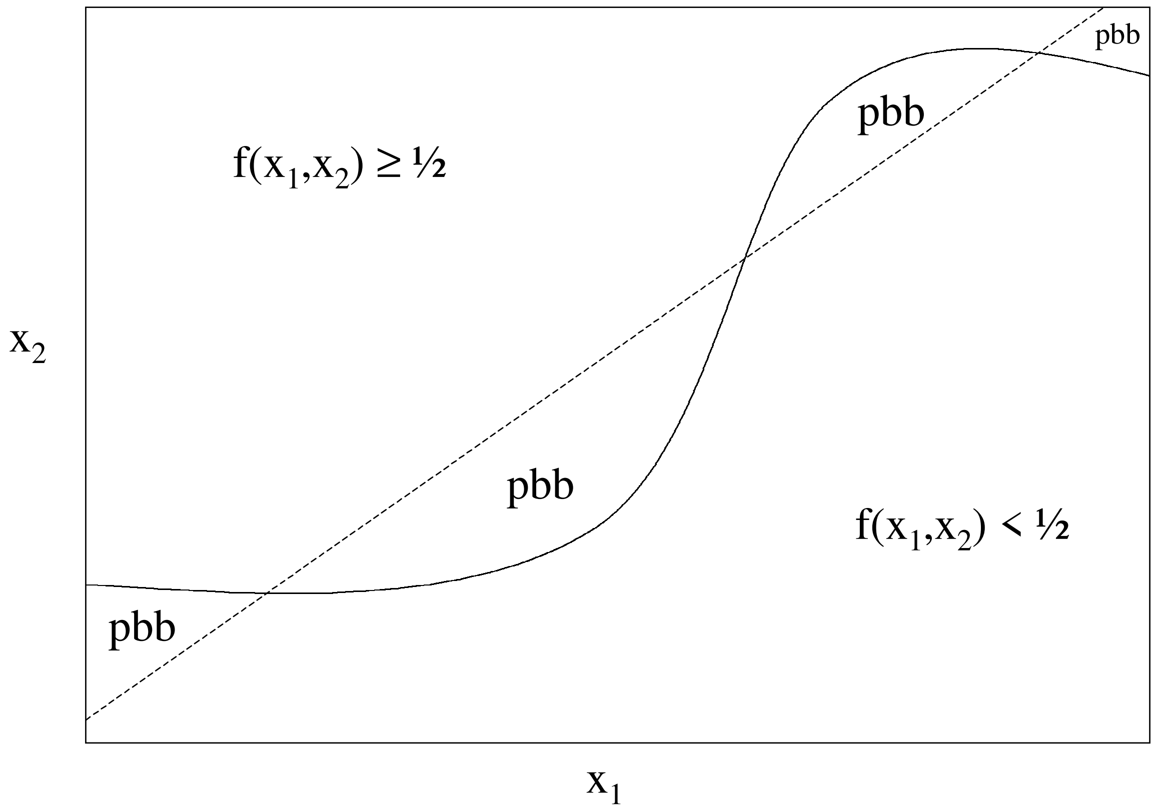
Summarizing Friedman's decomposition gives

$$P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) = 0.6 \cdot 0.1 + 0.2 = 0.26.$$

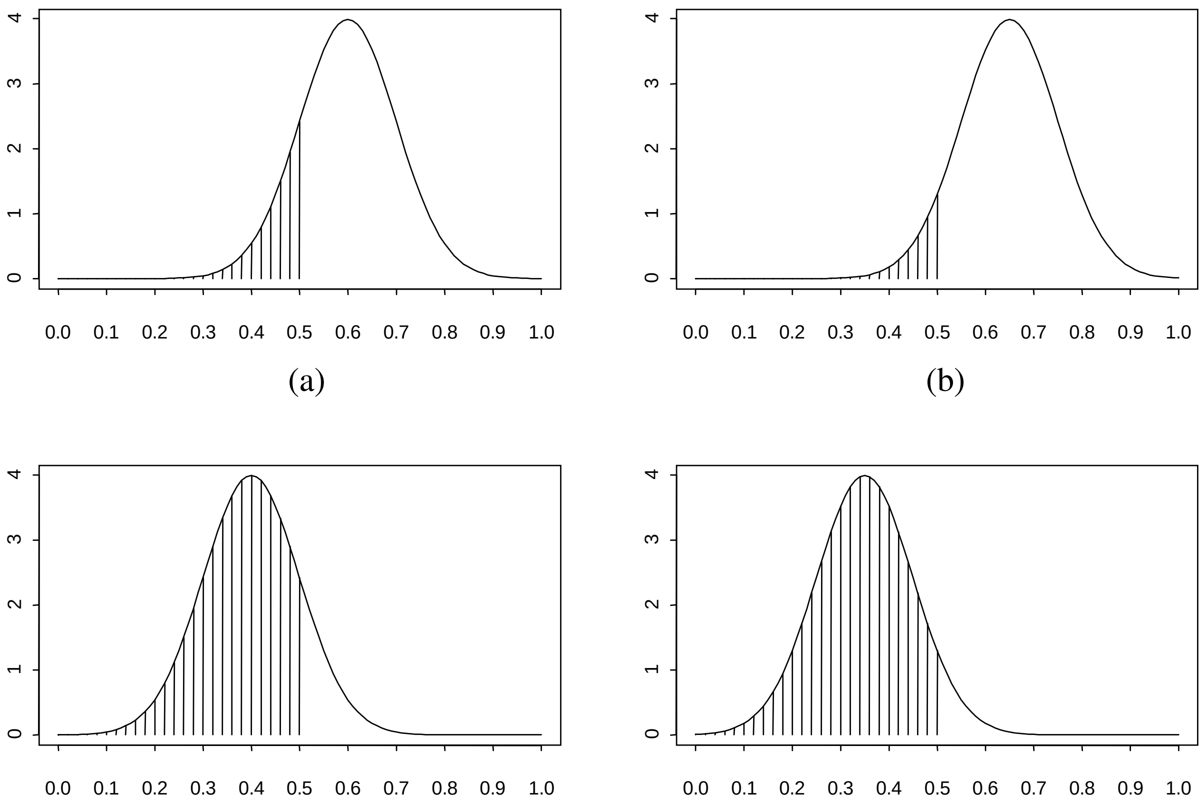
The important difference with regression is that a wrong estimate $\hat{f}(\mathbf{x})$ of $f(\mathbf{x})$ does not necessarily lead to a different allocation as Bayes rule, as long as they are on the same side of 1/2.

Example regions of positive boundary bias

Solid line denotes $f(x_1, x_2) = 1/2$ and dotted line denotes $\mathrm{E}[\hat{f}(x_1, x_2)] = 1/2$

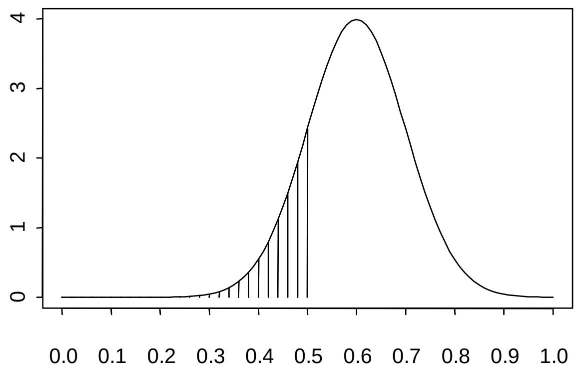


Distributions $p(\hat{f}(\mathbf{x}))$; $f(\mathbf{x}) = 0.6$.

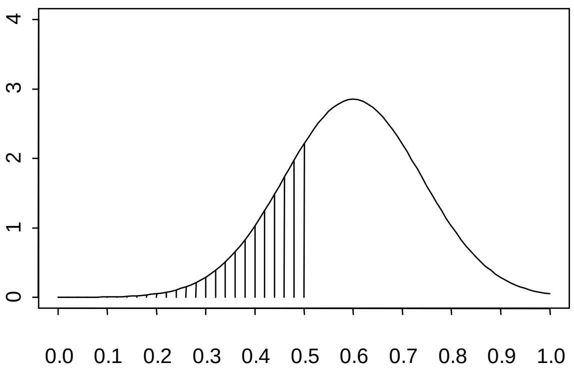


(a) (b) (c) (d)

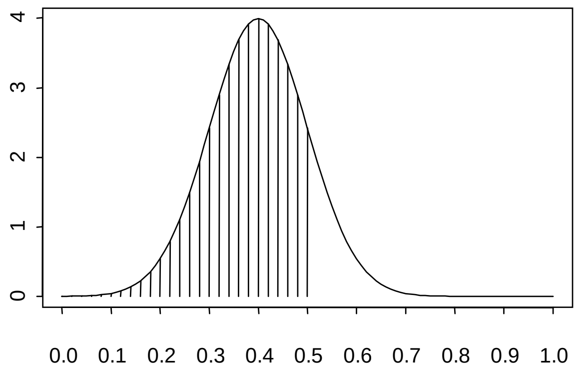
Distributions $p(\hat{f}(\mathbf{x}))$; $f(\mathbf{x}) = 0.6$.



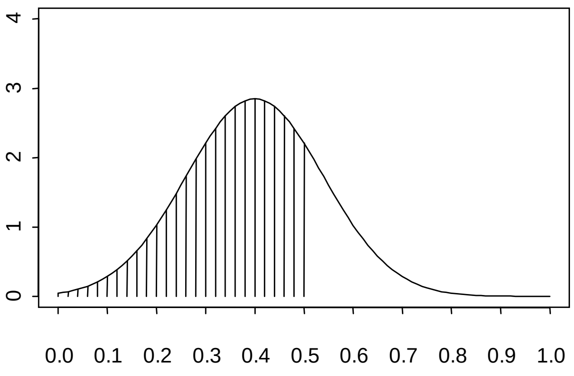
(a)



(b)



(c)



(d)

Bootstrapping

For many statistics that may be of interest to the analyst, such exact or limiting distributions cannot be derived analytically. In yet other cases the asymptotic approximation may not provide a good fit for a finite sample. In such cases an alternative approximation to the sampling distribution of a statistic $t(\mathbf{x})$ may be obtained using just the data at hand, by a technique called *bootstrapping*. To explain the basic idea of the *non-parametric* bootstrap, we first introduce the *empirical distribution function*

$$\widehat{F}(z) = \frac{1}{n}\sum_{i=1}^{n} I\left(x_i \leq z\right) \quad -\infty < z < \infty$$

where I denotes the indicator function and $\mathbf{x} = (x_1, x_2, ..., x_n)$ is a random sample from population distribution function F.

Bootstrap distribution

We now approximate the sampling distribution of $t(\mathbf{x})$ by repeated sampling from $\widehat{F}$. This is achieved by drawing samples $\mathbf{x}^{(r)}$ of size n by sampling independently *with replacement* from $(x_1, x_2, ..., x_n)$. If all observations are distinct, there are $\begin{pmatrix} 2n-1 \\ n \end{pmatrix}$ distinct samples in

$$\mathcal{B} = \{\mathbf{x}^{(r)}, r = 1, ..., \begin{pmatrix} 2n-1 \\ n \end{pmatrix}\}$$

with respective multinomial probabilities

$$P(\mathbf{x}^{(r)}) = \frac{n!}{j_1^{(r)}!\, j_2^{(r)}! \ldots j_n^{(r)}!}(\frac{1}{n})^n$$

where $j_i^{(r)}$ is the number of copies of x_i in $\mathbf{x}^{(r)}$.

Bootstrap distribution

The bootstrap distribution of $t(\mathbf{x})$ is derived by calculating the realisation $t(\mathbf{x}^{(r)})$ for each of the resamples and assigning each one probability $P(\mathbf{x}^{(r)})$.

We can in principle compute all $\begin{pmatrix} 2n-1 \\ n \end{pmatrix}$ values of the statistic to obtain its "ideal" bootstrap distribution, but this is computationally infeasible even for moderate n.

For $n = 15$ there are already 77558760 distinct samples. The usual alternative is to use Monte-Carlo simulation, by drawing a number B of samples and using them to approximate the bootstrap distribution.

Parametric and non-parametric bootstrap

If a *parametric* form is adopted for the underlying distribution, where θ denotes the vector of unknown parameters, then the parametric bootstrap uses an estimate $\widehat{\theta}$ formed from $\mathbf{x}$ in place of θ. If we write F_θ to signify its dependence on θ, then bootstrap samples are generated from $\widehat{F} = F_{\widehat{\theta}}$.

The non-parametric bootstrap makes it unneccesary to make assumptions about the form of the underlying distribution.

The parametric bootstrap may still provide more accurate answers than those provided by limiting distributions, and makes inference possible when no exact or limiting distributions can be derived for a sample statistic.

Example: bootstrapping the correlation coefficient

The linear dependence between population variables $\mathcal{X}$ and $\mathcal{Y}$ is measured by the covariance

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$$

A term $(x_i - \mu_x)(y_i - \mu_y)$ from this sum is positive if x_i and y_i are both above or both below their mean, and negative if x_i and y_i are on opposite sides of their mean. Division by both σ_x and σ_y yields a dimensionless number called the correlation coefficient

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

If $\rho = 0$ there is no linear dependence: the two variables are uncorrelated. The linear dependence increases as $|\rho|$ gets closer to 1. If all pairs (x, y) are on a straight line with positive slope, then $\rho = 1$; if all pairs are on a straight line with negative slope then $\rho = -1$.

Example: bootstrapping the correlation coefficient

To make inferences about ρ we use the sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The sampling distribution of this statistic can't be mathematically derived in general, in fact there is no general expression for the expected value of r_{xy}.

Therefore it makes sense to use the bootstrap to make inferences concerning ρ.

Example: simulation study

We draw 30 (x, y) pairs from a standard binormal distribution with $\rho = 0.7$, i.e.

$(X, Y) \sim \mathcal{N}^2(\mu_x = \mu_y = 0, \sigma_x^2 = \sigma_y^2 = 1, \rho = 0.7)$

Based on this dataset, bootstrapping proceeds as follows

Non-parametric: Draw samples of 30 (x, y) pairs (with replacement) from the data. For each bootstrap sample, compute r, to obtain an empirical sampling distribution.

Parametric: Make appropriate assumptions about the joint distribution of X and Y. In our study we assume

$$(X, Y) \sim \mathcal{N}^2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

Build an empirical sampling distribution by drawing samples of size 30 from

$$\mathcal{N}^2(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$$

Example: simulation study

Draw 1000 samples to generate the empirical sampling distribution of r. To construct $100(1-\alpha)\%$ confidence intervals for ρ, we simply take the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of this distribution.

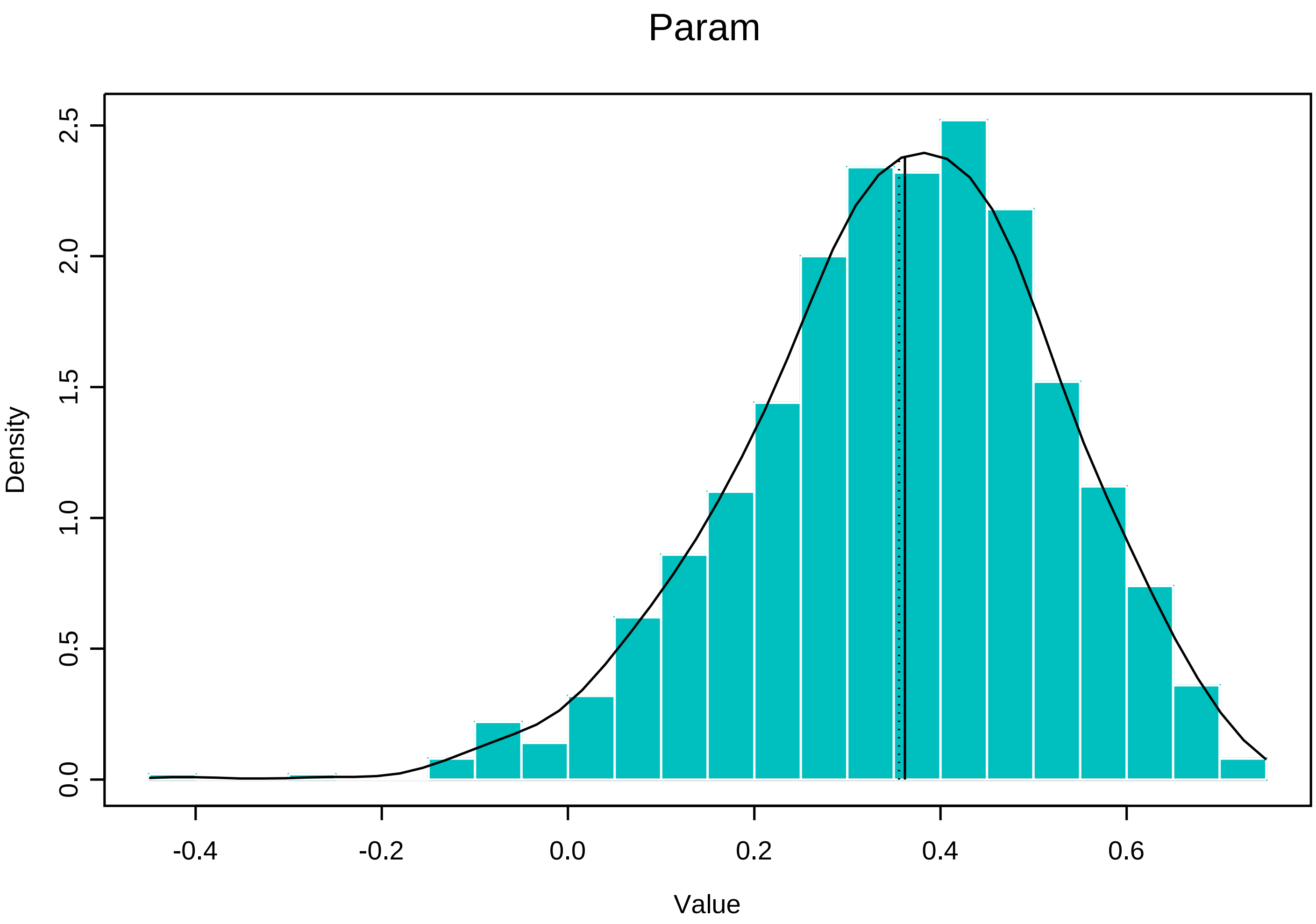
Reliable confidence intervals with the right coverage? We repeated the following procedure 100 times

1. Draw a sample of size 30 from the population.

2. Build a bootstrap distribution for r, and construct 90% confidence intervals for ρ. (both parametric and non-parametric)

3. Determine whether the true value of ρ is inside the confidence interval.

Bootstrap distribution for r. Observed value of r is 0.36.



Right coverage?

We observed 13 misses out of 100, so the observed value of our test statistic is $a = 0.13$. Now we may test whether the bootstrap confidence intervals have the right coverage:

$$H_0 : \alpha = 0.1 \quad \text{against} \quad H_a : \alpha \neq 0.1$$

The distribution of $\widehat{\alpha}$ under H_0 (the null-hypothesis) may be approximated by

$$\widehat{\alpha} \approx_{H_0} \mathcal{N}(\mu = \alpha, \sigma^2 = \alpha(1-\alpha)/n)$$

which yields $\widehat{\alpha} \approx \mathcal{N}(0.1, 0.0009)$.

Right coverage?

We may now compute the p-value of the observed value under the null-hypothesis as follows

$$
\begin{aligned}
P_{H_0}(\widehat{\alpha} > a) &= P_{H_0}(\widehat{\alpha} > 0.13) \\
&= P(Z > \frac{0.13 - 0.1}{\sqrt{0.0009}}) \\
&= P(Z > 1) = 0.1587
\end{aligned}
$$

where the value 0.1587 was looked-up in a table for the standardnormal distribution. Since we are performing a two-sided test this probability should be doubled, so we obtain a p-value of $2 \times 0.1587 = 0.3174$. This means we would not reject the null-hypothesis under any conventional significance level.

Bagging Predictors

Bagging is an acronym for **b**ootstrap **agg**regat**ing**, and is used to construct aggregated predictors with the intention of reducing the variance component of prediction error.

Thought experiment:

If we were to have access to the entire population of interest we could in principle compute aggregated regression and classification functions by averaging over all possible training samples T.

Aggregated regression and classification

If y is numerical, we replace $\widehat{f}(\mathbf{x} \,|\, T)$ by the average of $\widehat{f}(\mathbf{x} \,|\, T)$ over all possible samples T, that is by

$$f_A(\mathbf{x}) = \mathsf{E}_T \, \widehat{f}(\mathbf{x} \,|\, T)$$

where the A in f_A denotes aggregation.

This aggregation reduces mean prediction error, since the bias of $f_A(\mathbf{x})$ is the same as that of $\widehat{f}(\mathbf{x} \,|\, T)$, but its variance is 0.

If y is a class label, say $y \in \{0, 1\}$, the aggregated classifier may defined as

$$y_A(\mathbf{x}) = I(f_A \geq 1/2) = I(\mathsf{E}_T \widehat{f}(\mathbf{x} \,|\, T) \geq 1/2)$$

where again the A in y_A denotes aggregation.

Aggregation by Bootstrapping

In practice we only have a single training sample, so we cannot compute f_A.

Instead of sampling repeatedly from the population, we *resample* the single sample T we have, and estimate f_A by

$$f_B(\mathbf{x}) = \hat{f}_A(\mathbf{x}) = \mathsf{E}_{T^{(r)}} \hat{f}(\mathbf{x} \,|\, T^{(r)})$$

where $T^{(r)}$ are bootstrap samples from T.

What do we gain?

Now there are two factors influencing the performance of f_B.

If the predictor has high variance, it can give improvement through aggregation.

If the predictor has low variance, then the unaggregated predictor $\widehat{f}(\mathbf{x} \mid T)$ tends to be close to $\mathsf{E}_T \widehat{f}(\mathbf{x} \mid T)$, and so $f_B = \mathsf{E}_{T^{(r)}} \widehat{f}(\mathbf{x} \mid T^{(r)})$ may be less accurate then $\widehat{f}(\mathbf{x} \mid T)$.

High variance: neural networks, trees

Low variance: linear regression

Example: bagging a neural network

Neural networks can be used as flexible regression methods and can approximate complex non-linear functions. In general, neural networks tend to have a low bias and high variance, so they may benefit from aggregation methods such as bagging.

We generated a simulated data set of size 200 with ten independent predictor variables $x_1, \ldots, x_{10}$ with $x_j \sim U(0,1)$, $j = 1, \ldots, 10$.

The response is given by

$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$

with $\varepsilon \sim \mathcal{N}(0,1)$.

Example: bagging a neural network

Test sample: 1000 observations

Repeat 100 times

1. Draw a training sample of 200 observations.

2. Train a single neural network on the dataset. Make predictions on the test sample.

3. Draw 25 bootstrap samples, and train a neural network on each bootstrap sample. Make averaged predictions for the test sample.

Bias and variance of single and aggregated neural network

We may now decompose the mean square estimation error into its bias and variance components. We would expect the aggregated neural network to show smaller variance and comparable bias, on average leading to smaller prediction error than the unaggregated version.

		unaggregated			bagged		
	$f(\mathbf{x})$	$\mathrm{E}(\hat{f})$	B^2	V	$\mathrm{E}(\hat{f})$	B^2	V
1	6.79	5.62	1.36	98.58	6.75	0.00	23.59
2	14.14	14.27	0.02	1.81	14.50	0.13	0.55
3	14.60	14.31	0.08	2.72	13.93	0.45	5.29
4	6.31	7.85	2.37	77.48	7.80	2.20	1.28
5	20.19	19.80	0.15	1.29	19.24	0.90	1.24
6	15.36	15.07	0.09	83.13	13.42	3.77	2.03
7	7.22	5.99	1.52	105.79	5.83	1.94	36.42
8	18.53	18.36	0.03	14.62	18.35	0.03	1.57
9	5.40	5.38	0.00	1.70	5.77	0.14	7.17
10	23.05	22.04	1.01	12.86	22.12	0.85	2.64

Bagging a neural network: results

Averaged over the 1000 test observations,

	squared bias	variance
single network	1.51	47.64
bagged network	1.97	4.66

This example illustrates that bagging is a variance reduction technique; the bias component of error is not reduced by bagging.

Arcing classifiers

Arcing is an acronym for **a**daptive **r**esampling and **c**ombin**ing,** and has proved to be a succesfull technique in reducing the prediction error of classifiers.

Like bagging, arcing is an aggregation method based on resampling the available data. The main difference with bagging is the *adaptive* resampling scheme.

In bagging bootstrap samples are generated by sampling the original data where each data point has probability $1/n$ (where n is the number of elements in the original sample).

In the arcing algorithm these inclusion probabilities may change.

Arcing: inclusion probabilities

At the start of each construction, there is a probability distribution

$$\{p(i)\}, i = 1, ..., n$$

on the cases in the training set. A bootstrap sample $T^{(r)}$ is constructed by sampling n times (with replacement) from this distribution. Then the probabilities are updated in such a way that the probability of point $\mathbf{x}$ is increased if it is misclassified by the classifier constructed from $T^{(r)}$ and remains the same otherwise

After R classifiers have been constructed in this manner, a (possibly weighted) voting for the class allocation is performed.

Intuitive idea: the points most likely to be selected for the replicate data sets are those most likely to be misclassified. Since these are the troublesome points, focussing on them may do better than the “neutral” bagging approach.

Arcing: arcx4

The algorithm works as follows

1) At the r^{th} step, sample with replacement from T to obtain training set $T^{(r)}$, using inclusion probabilities $\{p^{(r)}(i)\}$. Use $T^{(r)}$ to construct (estimate) $\widehat{f}_r$.

2) Predict the class of each point in T using $\widehat{f}_r$, i.e. determine $\widehat{y}_r = I(\widehat{f}_r \geq 1/2)$ for each point in T. Now let $k(i)$ be the number of misclassifications of the i^{th} point by $\widehat{y}_1, \ldots, \widehat{y}_r$, the classifiers constructed so far.

Arcing: arcx4 (continued)

3) The new $r+1$ step inclusion probabilities are defined by

$$p^{(r+1)}(i) = (1 + k(i)^4) / \sum_{j=1}^{n} (1 + k(j)^4)$$

4) Finally, after R steps the aggregated classifier is determined

$$\hat{y}_A = I(\hat{f}_A \geq 1/2), \quad \text{with} \quad \hat{f}_A = 1/R \sum \hat{f}_r$$

Example: arcing neural networks

Repeat the following steps 100 times

1. Generate a training set T containing 150 observations from class 0 and 150 observations from class 1.

2. Estimate the unaggregated function $\widehat{f}$ and classifier $\widehat{y}$ using T.

3. Apply arcx4 to T to obtain an aggregated function $\widehat{f}_a$ and classifier $\widehat{y}_a$, and apply bagging to T to construct an aggregated function $\widehat{f}_b$ and classifier $\widehat{y}_b$ We chose $R = 25$ so the aggregated classifiers are based on averages of 25 functions.

4. Use the classifiers so obtained to predict the class of 1000 observations in an independent test.

The Data

The class 0 observations were drawn from a 10-dimensional normal distribution with mean 0 and covariance matrix 5 times the identity. The class 1 observations were drawn from a 10-dimensional normal distribution with mean 0.5 and identity covariance matrix.

$$\mathbf{x} \,|\, y = 0 \sim \mathcal{N}^{10}(\mu_{\mathbf{0}} = \mathbf{0}, \Sigma_0 = I^{10})$$

$$\mathbf{x} \,|\, y = 1 \sim \mathcal{N}^{10}(\mu_{\mathbf{1}} = \mathbf{0.5}, \Sigma_1 = 5I^{10})$$

Since the class-conditional distributions are normal with distinct covariance matrices, the optimal decision boundary is a quadratic function of $\mathbf{x}$.

Bayes error at $\mathbf{x}$

We can compute the Bayes error for each point in the test set since we specified the distribution from which the class 0 points were drawn, the distribution from which the class 1 points were drawn, and the prior probabilities of both classes. This allows us the compute the probability of class 1 at point $\mathbf{x}$ using Bayes rule as follows

$$P(y = 1 \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, y = 1)P(y = 1)}{p(\mathbf{x} \,|\, y = 1)P(y = 1) + p(\mathbf{x} \,|\, y = 0)P(y = 0)}$$

where p denotes a probability density.

Arcing and Bagging and its bias-variance decomposition

	Breiman			
	$\bar{e}$	e_B	bias	var
unaggregated	21.6%	3.7%	4.4%	13.6%
arced	10.4%	3.7%	1.8%	4.9%
bagged	14.1%	3.7%	6.1%	4.3%

	Squared error		
	$\mathrm{E}(f-\hat{f})^2$	$(f-\mathrm{E}\hat{f})^2$	$\mathrm{E}(\hat{f}-\mathrm{E}\hat{f})^2$
unaggregated	.138	.068	.070
arced	.095	.089	.006
bagged	.082	.072	.010