

# Statistics 101

Paul G. Plöger  
SEE



Hochschule  
Bonn-Rhein-Sieg

Fachbereich  
Informatik

Prof. Dr.  
Paul G. Plöger

Vers 0.4/April 15

# Overview

- 1) Various distributions of random Variables and their display
- 2) Characterizing distributions; central tendency and dispersion
- 3) Some theorems on distributions (with / without models)
- 4) Standard error of empirical mean and empirical variance
- 5) Error propagation in regression for measurements with precision



# Ex: The Quetelet Curve

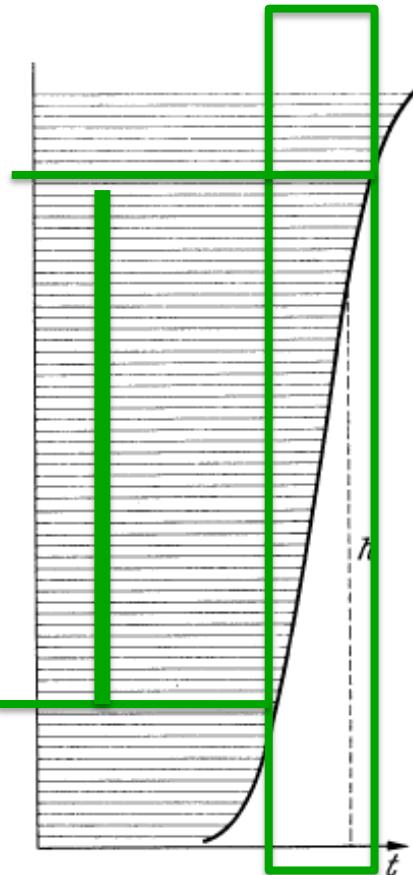


Fig. 11. Quetelet's curve

„I still remember vividly how my father took me one day as a boy to the outskirts of the city, where the willows stood on the bank and had me pick a hundred willow leaves at random. After throwing out those with damaged tips, 89 were left which we took home. We arranged them very carefully according to decreasing size like soldiers in rank and file.

Then my father drew a curved line through the tips and said, "This is the curve of Quetelet. From it you see how the mediocre always form the large majority and only a few stand out or are left behind."



# Two views on Statistics

Statistical methods to describe data in form of diagrams, tables or individual parameters is called **descriptive statistics**.

If we infer, based on these qualitative analysis techniques, from these given empirical data the validity of hypothesizes we talk about **deductive statistics**



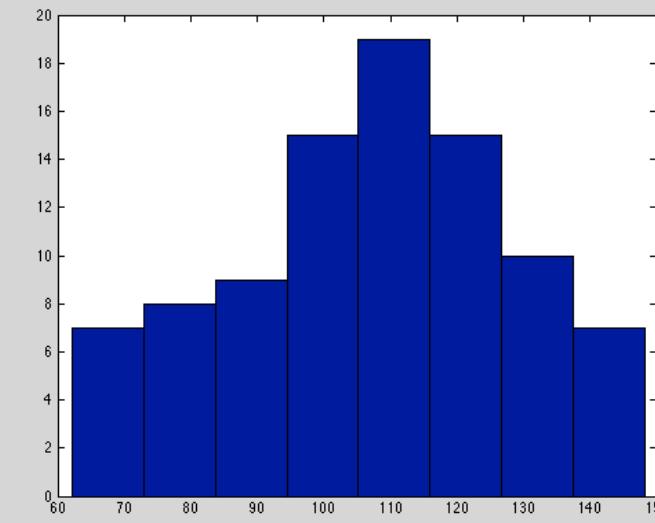
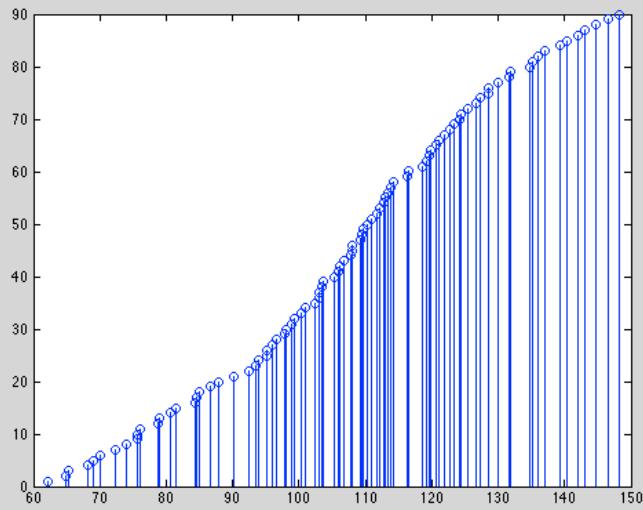
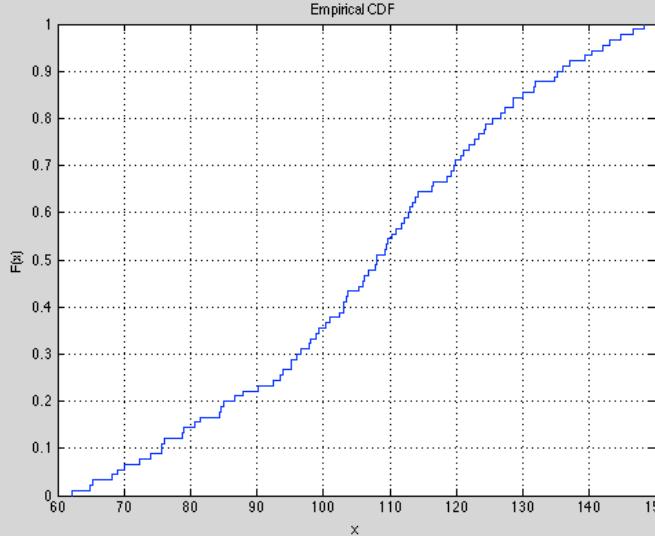
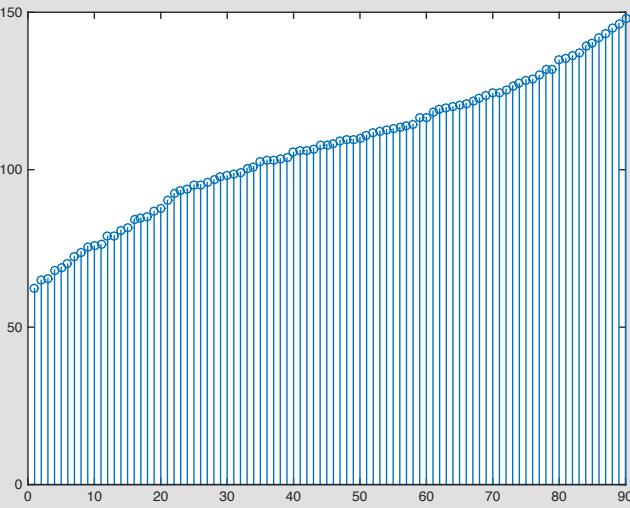
# Ex: Raw data table, Master list

|       |       |       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 131.8 | 106.7 | 116.4 | 84.3  | 118.5 | 93.4  | 65.3  | 113.8 | 140.3 |
| 119.2 | 129.9 | 75.7  | 105.4 | 123.4 | 64.9  | 80.7  | 124.2 | 110.9 |
| 86.7  | 112.7 | 96.7  | 110.2 | 135.2 | 134.7 | 146.5 | 144.8 | 113.4 |
| 128.6 | 142.0 | 106.0 | 98.0  | 148.2 | 106.2 | 122.7 | 70.0  | 73.9  |
| 78.8  | 103.4 | 112.9 | 126.6 | 119.9 | 62.2  | 116.6 | 84.6  | 101.0 |
| 68.1  | 95.9  | 119.7 | 122.0 | 127.3 | 109.3 | 95.1  | 103.1 | 92.4  |
| 103.0 | 90.2  | 136.1 | 109.6 | 99.2  | 76.1  | 93.9  | 81.5  | 100.4 |
| 114.3 | 125.5 | 121.0 | 137.0 | 107.7 | 69.0  | 79.0  | 111.7 | 98.8  |
| 124.3 | 84.9  | 108.1 | 128.5 | 87.9  | 102.4 | 103.7 | 131.7 | 139.4 |
| 108.0 | 109.4 | 97.8  | 112.2 | 75.6  | 143.1 | 72.4  | 120.6 | 95.2  |

**Urdata( : )**



# Some different views / plot types in MATLAB



```
hist(Urdata(:,8));
stem(1:length(Urdata(:,8)),sort(Urdata(:,8)));
stem(sort(Urdata(:,8)),1:length(Urdata(:,8)));
cdfplot(Urdata(:,8))
```

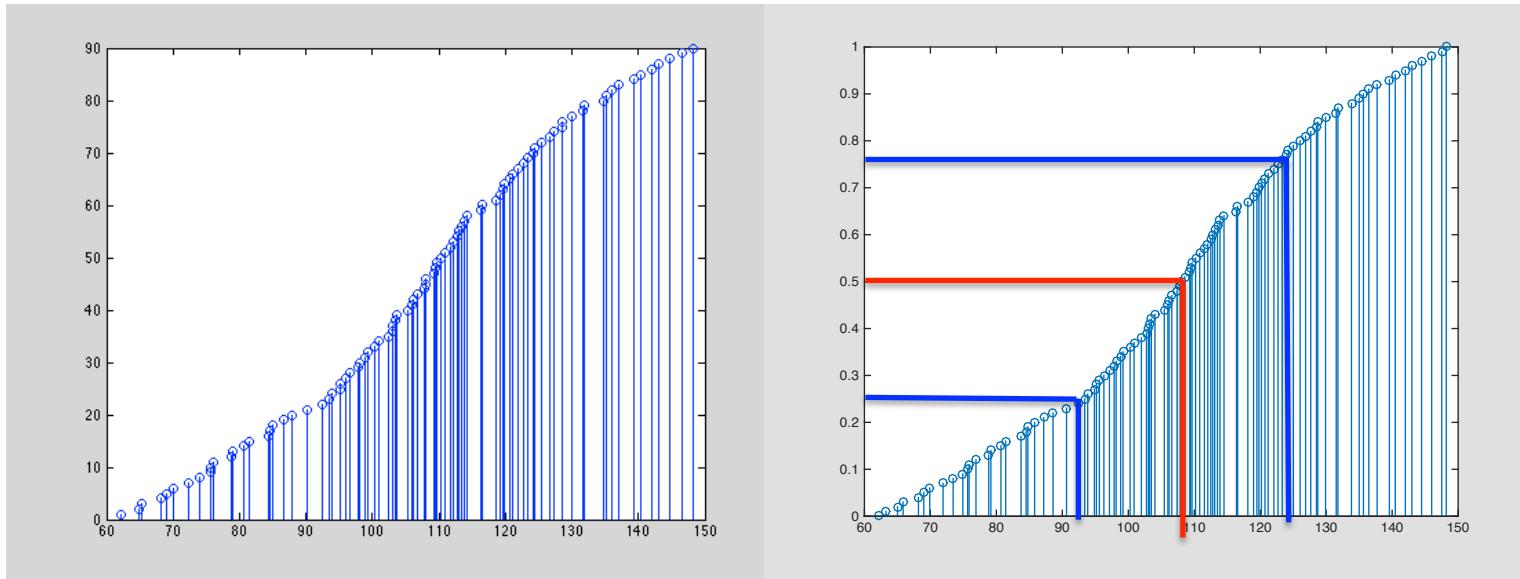
```
import numpy as np
import statsmodels.api as sm
import matplotlib.pyplot as plt

sample = np.random.uniform(0, 1, 50)
ecdf = sm.distributions.ECDF(sample)

x = np.linspace(min(sample), max(sample))
y = ecdf(x)
plt.step(x, y)
plt.show()
```

# Quantile is a scaled version of CDF

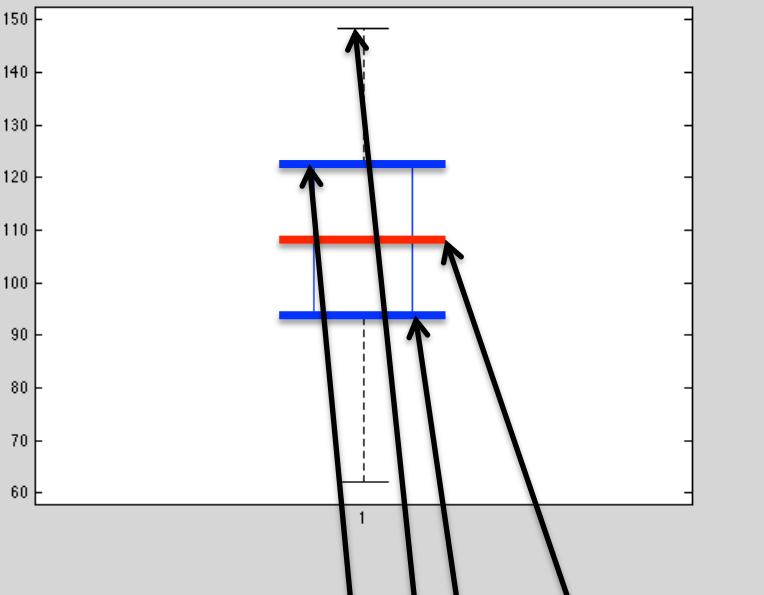
## Quantile is inverse of CDF



```
stem(sort(Urdata(:)),1:length(Urdata(:)));
stem(quantile(Urdata(:),0:0.01:1),0:0.01:1)
```



# One of the easiest: boxplot



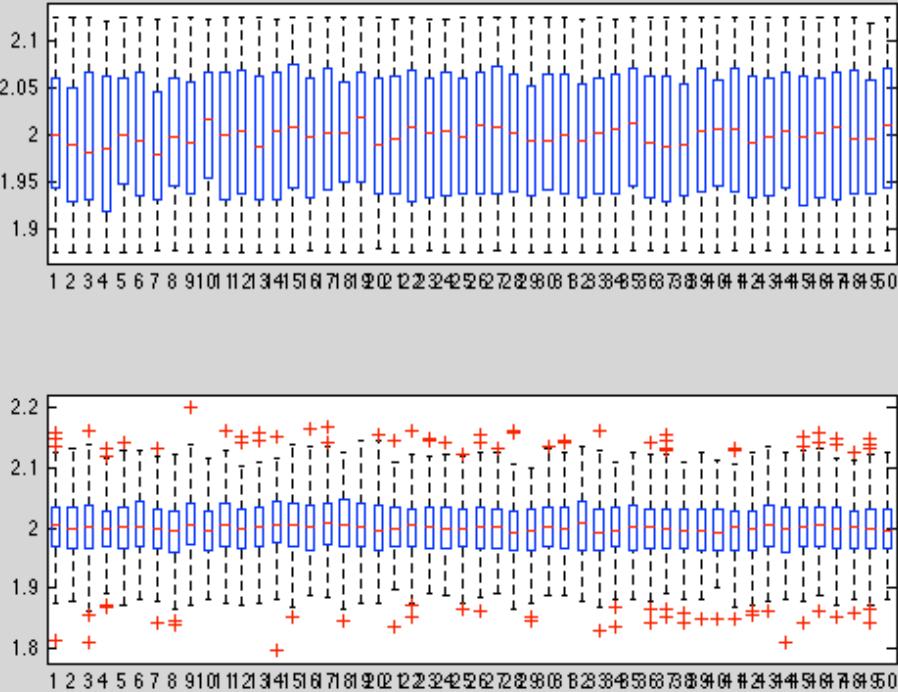
```
boxplot(Urdata(:))
```

It shows:

- the median (red) (50% quantile)
- the lower 25% quantile (lower box line: lbl)
- upper 75% quantile (upper box line: ubl)
- whiskers == median +/- 1.5\*(ubl-lbl)
- outliers (data beyond whiskers == 99.3 % coverage for normal distribution)



# Good for overview spaghetti data



```
spaghetti_50 =(rand(250,50)-0.5)*0.25+2;  
spaghetti_50n = random('normal',2,0.05,250,50);  
subplot(2,1,1);boxplot(spagetti_50);  
subplot(2,1,2);boxplot(spagetti_50n)  
%another example  
PX=[ones(1,30)*0.2 ones(1,40)*1.2 ones(1,40)*2.4];
```



A random variable  $X$  is a real-valued function :

$$X : \Omega \rightarrow I\mathbb{R},$$

$$\omega \mapsto X(\omega)$$

where  $\omega$  is the result of a random experiment like e.g. measured runtime of an algorithm on a CPU with a cache.  $X$  needs to be measurable.

A probability measure  $P$  is a map from an algebra of sets  $\mathcal{A}$

$$P : \mathcal{A} \rightarrow I\mathbb{R},$$

$A \mapsto P(A)$  which obeys : (Kolmogorov Axioms) :

$$(1) P(A) \geq 0$$

$$(2) P(\Omega) = 1$$

$$(3) P(A \cup B) = P(A) + P(B) \text{ if } A \cap B = \emptyset$$

The cumulative distribution function (CDF) is defined as :

$$F_X : I\mathbb{R} \rightarrow [0,1],$$

$$x \mapsto F_X(x) := P(\{\omega \mid X(\omega) \leq x\})$$

if  $F_X$  is a CDF then the function

$$Q_X : ]0,1[ \rightarrow I\mathbb{R},$$

$$p \mapsto Q_X(p) = \min\{x \mid F_X(x) \geq p\}$$

is called quantile function of  $F_X$ .

$Q_X$  inverse of  $F_X$  in following sense :

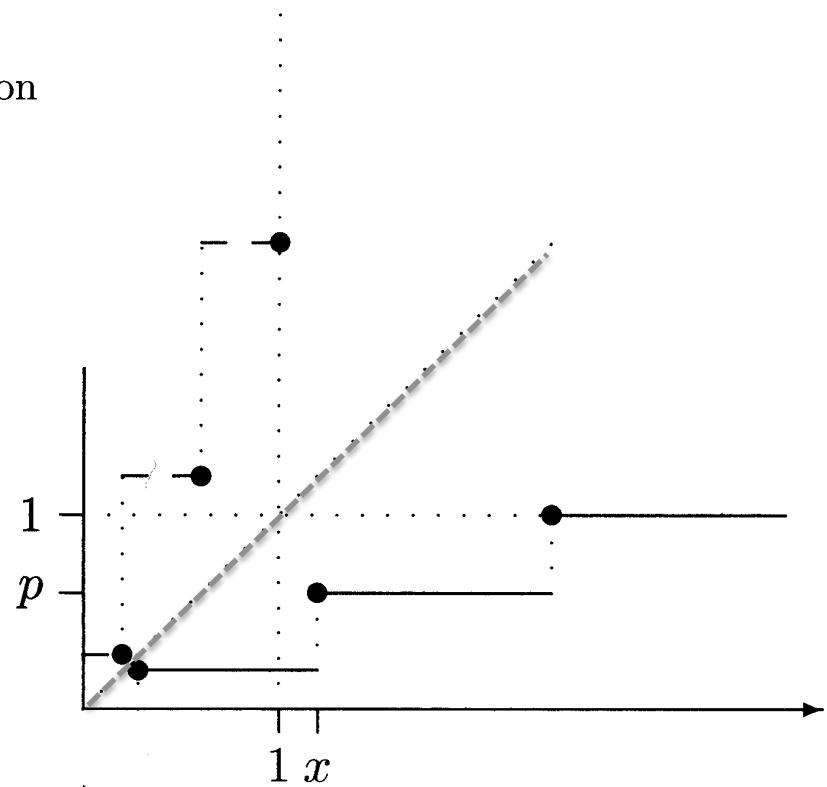
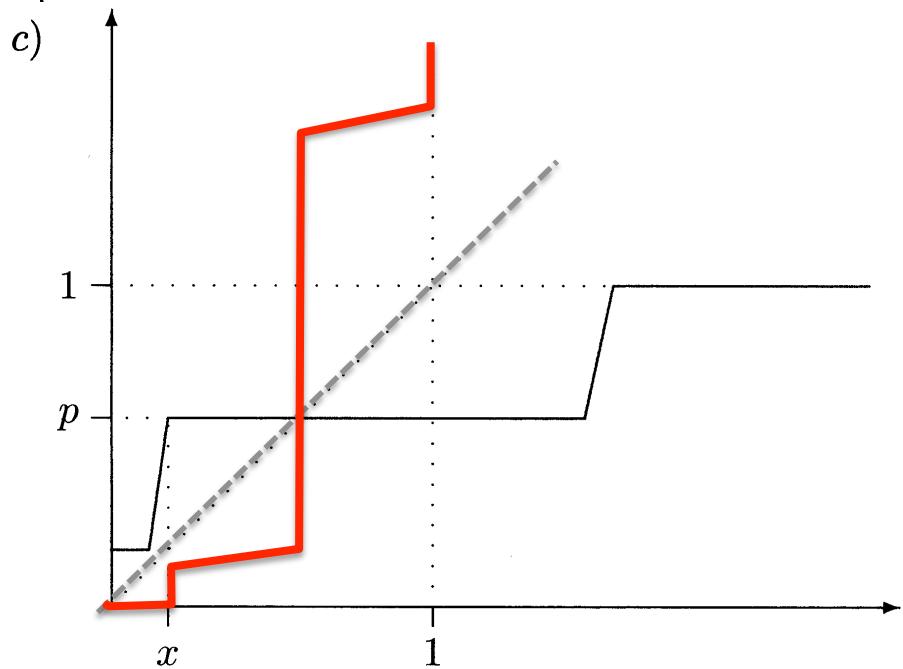
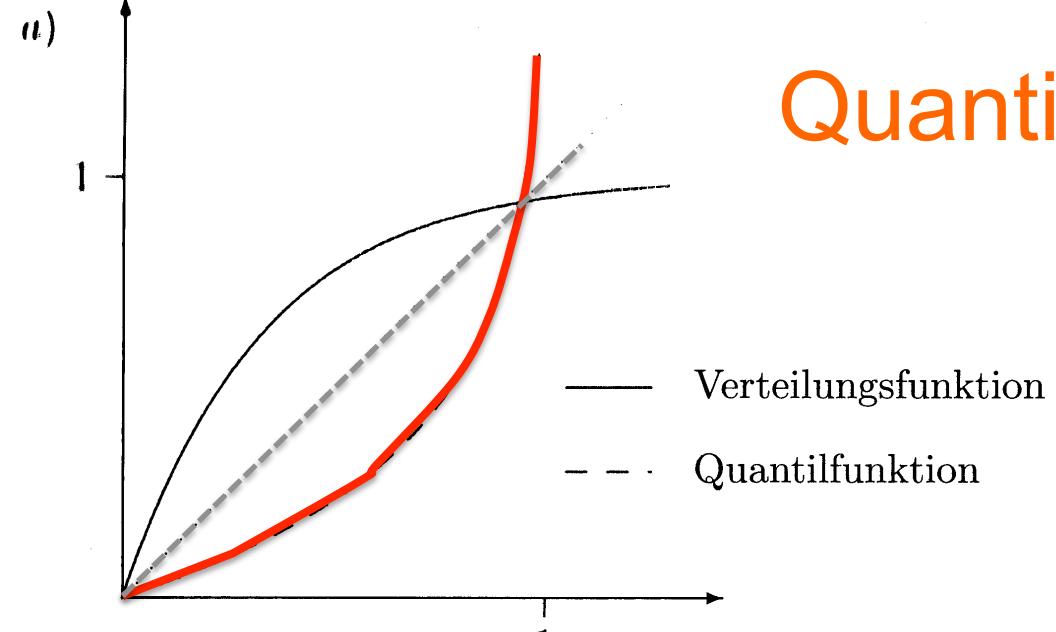
$$F_X(Q_X(p)) \geq p \text{ for } p \in ]0,1[ \text{ and}$$

$$Q_X(F_X(x)) \leq x \text{ for } x \in I\mathbb{R}$$

# Random Variable, Cumulative distribution function, Quantile fct



Quantil == inverse CDF



# Histogram, How to ?

With increasing size of the examined collective smaller category widths can be selected.

The larger the variation of the measured values (the difference between the largest and the smallest value), the wider the categories may be.

According to a rule of thumb of Sturges (1926) the number of categories  $m$  ( $n$  = sample size) should be determined according to the relation  $m \approx 1 + 3.32 \lg(n)$  where  $n$  is the sample size.

The maximum number of categories should not exceed 20 for reasons of clarity.

All categories should have the same width in general.



# Overview

- 1) Various distributions of random Variables and their display
- 2) **Characterizing distributions; central tendency and dispersion**
- 3) Some theorems on distributions (with / without models)
- 4) Standard error of empirical mean and empirical variance
- 5) Error propagation in regression for measurements with precision



# Graph versus Parameters

While the graph of a distribution stresses the **WHOLE** distribution (the whole collection of data is represented), statistical parameters should stress special, global, summarizing features of the underlying distribution.

It should represent „typical“ attributes, like the general tendency or variability i.e. the dispersion of the distribution.



# Which single value is most characteristic?

**Unimodal :=  
one peak**

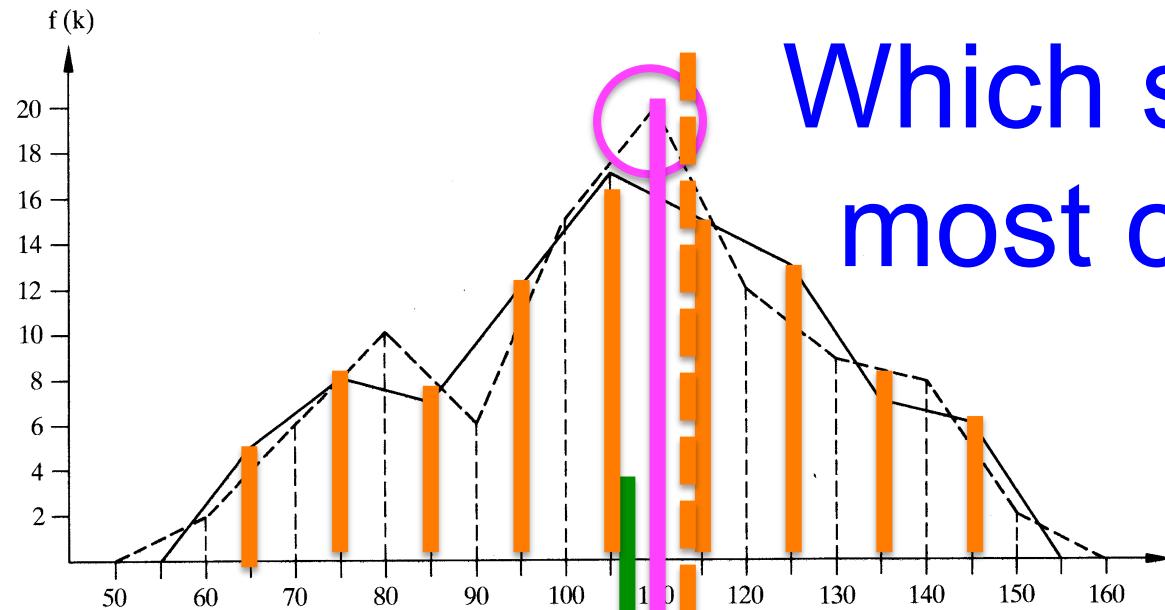


Abb. 1.1. Polygon der Häufigkeiten in Tab. 1.4 für 2 Reduktionslagen

**Modal value:** most frequent value of distribution:

```
[count,values]=hist(Urdata(:,1),11);
values(find(count == max(count)))
```

**Median value:** e.g. 50% are below this value:

```
quantile(Urdata(:,1),1/2) %or [1/4 ½ 3/4])
```

**(emp.) Mean value:** on average closest to the data:

```
1/length(Urdata(:,1))*sum(Urdata(:,1))
```

# Modal value of frequency distribution

**Tabelle 1.6.** Modalwert einer Häufigkeitsverteilung

| Meßwert (x)      | Häufigkeit (f(x)) |
|------------------|-------------------|
| 11               | 2                 |
| 12               | 8                 |
| 13               | 18                |
| 14               | 17                |
| 15               | 22                |
| <b>Modalwert</b> | <b>16</b>         |
|                  | 28                |
| 17               | 21                |
| 18               | 11                |
| 19               | 3                 |



# Median for grouped data

**Tabelle 1.7.** Medianwertbestimmung bei gruppierten Daten

| Fehleranzahl (k) | Häufigkeit (f) | $f_{\text{kum}}(k)$ |
|------------------|----------------|---------------------|
| 1– 20            | 3              | 3                   |
| 21– 40           | 16             | 19                  |
| 41– 60           | 12             | 31                  |
| 61– 80           | 7              | 38                  |
| 81–100           | 5              | 43                  |
| 101–120          | 4              | 47                  |
| 121–140          | 3              | 50                  |

$u$ =lower limit of category containing hit

$Kb$ =interval width

$F$ =count of cases in hit category, but below median

$f_{MD}$ =case count in hit category

$$Md = u + \frac{\frac{n}{2} - F}{f_{MD}} Kb$$



# EM = empirical mean = $\bar{x}$

It lies optimally „in middle“ i.e minimizing summed distance.

The mean equilibrates the difference to all elements  
BUT also  
mean is minimal for the sum of squared differences i.e.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

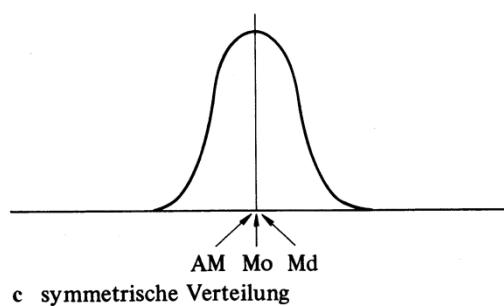
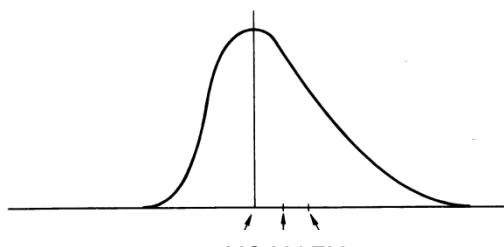
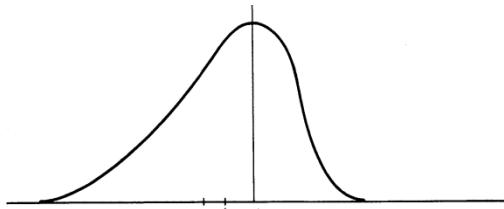
$$\sum (x_i - \bar{x}) = \sum x_i - \sum \bar{x} =$$

$$\sum x_i - n\bar{x} = \sum x_i - n \frac{1}{n} \sum x_i = 0$$

$$\min_c \sum (x_i - c)^2 = \bar{x}$$



# Compare Modal val., Median, EM



In case of symmetric distributions they are the same.

For biased ones (tilted in one direction) see left.



# Dispersion I

**AD = average deviation**

$$AD = \frac{\sum_{i=1}^n (|x_i - \bar{x}|)}{n}.$$

**Example**

**Tabelle 1.8.** Berechnung einer AD-Streuung

| Noten (x) | ( x <sub>i</sub> - $\bar{x}$  ) |
|-----------|---------------------------------|
| 3,3       | 0,8                             |
| 1,7       | 0,8                             |
| 2,0       | 0,5                             |
| 4,0       | 1,5                             |
| 1,3       | 1,2                             |
| 2,0       | 0,5                             |
| 3,0       | 0,5                             |
| 2,7       | 0,2                             |
| 3,7       | 1,2                             |
| 2,3       | 0,2                             |
| 1,7       | 0,8                             |
| 2,3       | 0,2                             |

$$\sum_{i=1}^n x_i = 30$$

$$\bar{x} = 2,5$$

$$\sum_{i=1}^n (|x_i - \bar{x}|) = 8,4$$

$$AD = \frac{8,4}{12} = 0,70$$



# Dispersion II

Empirical variance

Empirical standard deviation

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

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

Remark:

For small n it's better to divide by  $(n-1)!$   
(Unbiased estimator!)



**Tabelle 1.9.** Berechnung einer Varianz

| Noten (x) | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|-----------|-----------------|---------------------|
| 3,3       | 0,8             | 0,64                |
| 1,7       | -0,8            | 0,64                |
| 2,0       | -0,5            | 0,25                |
| 4,0       | 1,5             | 2,25                |
| 1,3       | -1,2            | 1,44                |
| 2,0       | -0,5            | 0,25                |
| 3,0       | 0,5             | 0,25                |
| 2,7       | 0,2             | 0,04                |
| 3,7       | 1,2             | 1,44                |
| 2,3       | -0,2            | 0,04                |
| 1,7       | -0,8            | 0,64                |
| 2,3       | -0,2            | 0,04                |

$$\sum_{i=1}^n x_i = 30$$

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^n (x_i - \bar{x})^2 = 7,92$$

$$\bar{x} = 2,5$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{7,92}{12} = 0,66$$

$$s = \sqrt{0,66} = 0,81$$



# Overview

- 1) Various distributions of random Variables and their display
- 2) Characterizing distributions; central tendency and dispersion
- 3) **Some theorems on distributions (with / without models)**
- 4) Standard error of empirical mean and empirical variance
- 5) Error propagation in regression for measurements with precision



# Tschebyscheff inequality

If  $X$  is a random variable with  $E(X)=\mu$  and standard deviation  $\sigma$ ,  $k > 0$  then it holds that:

$$P(|X - \mu| \leq k\sigma) \geq \left(1 - \frac{1}{k^2}\right)$$

This determines a lower bound of that part of data that falls within  $k$  number of std.-deviations from the mean. But taking complements (focus on data away from the mean, i.e. outliers):

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$



# Patient in Hospital

Let  $X$  be the days of patients in hospital, expected value  $\bar{x} = 10$ , std variance  $\sigma = 4$ .

How likely is it, that a patient stays more then five and less then 10 days?

$$P(5 < X < 15) = P(|X - 10| < 5) \geq 1 - \frac{4*4}{5*5} = \frac{9}{25}$$



# Estimate X inside 2 or 3 $\sigma$ bounds

Just Tschebyscheff on P :

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \geq \frac{3}{4} = 0.75$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) \geq \frac{8}{9} = 0.88$$

e.g. additional knowledge: P is normal =>

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 2\Phi(2) - 1 = 0.9544$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 2\Phi(3) - 1 = 0.9974$$

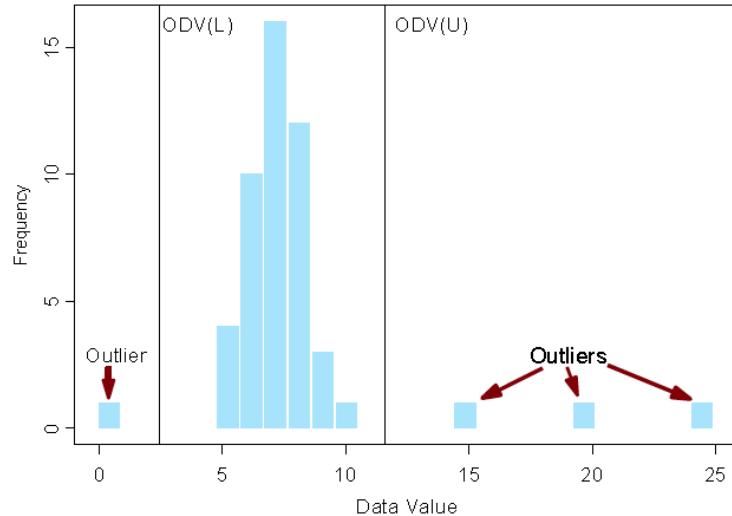


# In practice we have to use empirical values

$$\mu \approx \bar{x} = \frac{1}{n} \sum x_i$$

$$\sigma^2 \approx s_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

Example for outliers



# Ex: Tscheb. Outlier Algorithm

```

function [Rdata] = PGPremoveOutlier(PData,Pp1,Pp2)
%function [Rdata] = PGPremoveOutlier(Pdata,Pp1,Pp2)
%Pdata discrete, continuous
%Pp1 likelihood for expected outliers e.g. like 0.10, 0.05 , 0.01
%this will sort out bad data to get better mu and sigma
%Pp2 final likelihood for real outliers e.g. like 0.01, 0.001 , 0.0001
%all according to publ: Data Outlier Detection using the Chebyshev Theorem
%Brett G. Amidan, Thomas A. Ferryman, and Scott K. Cooley

%TODO check if its unimodal bcz then sharper bounds result

%%% phase 1
mu1= mean(PData);
sigma1 = std(PData);
k = inv(sqrt(Pp1));
ODV1U = mu1 + k*sigma1;
ODV1L = mu1 - k*sigma1;
NewData=PData(find(PData<=ODV1U));
NewData=NewData(find(NewData>=ODV1L));

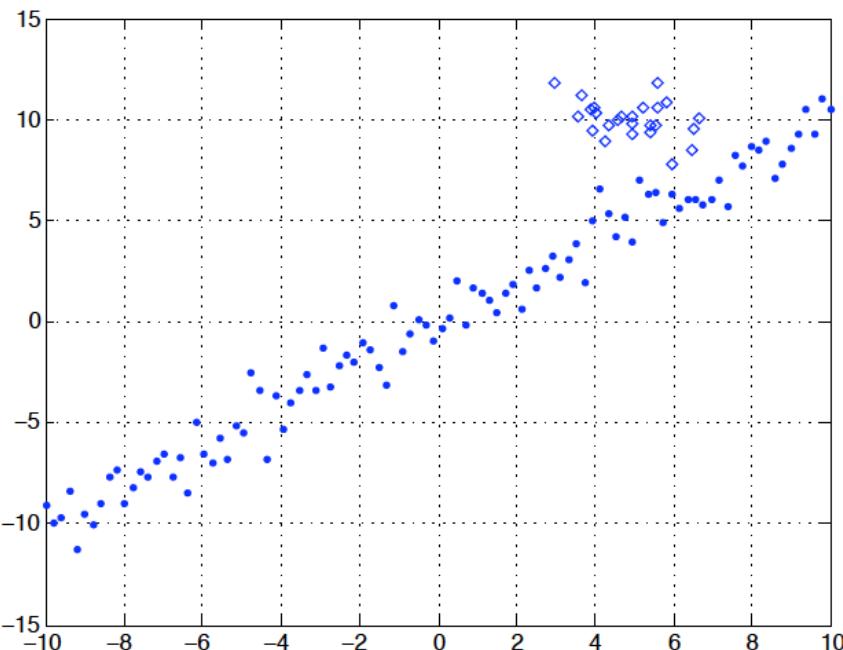
%%% phase 2: calculate new u
%%% unbiased mu and sigma
mu2 = mean(NewData);
sigma2 = std(NewData);
k = inv(sqrt(Pp2));
ODV2U = mu2 + k*sigma2;
ODV2L = mu2 - k*sigma2;
NewNewData=NewData(find(NewData<=ODV2U));
Rdata=NewNewData(find(NewNewData>=ODV2L));
;

return

```



# Outlier removal



Entropy 2009, 11, 560-585; doi:10.3390/e11040560

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

## An Entropy-Like Estimator for Robust Parameter Identification

Giovanni Indiveri

Dipartimento Ingegneria Innovazione, University of Salento, Via Monteroni s.n., 73100 Lecce, Italy;  
E-Mail: giovanni.indiveri@unisalento.it; Tel.: +39 0832 29 7220; Fax +39 0832 29 7733

Received: 7 September 2009 / Accepted: 23 September 2009 / Published: 12 October 2009

**Abstract:** This paper describes the basic ideas behind a novel prediction error parameter identification algorithm exhibiting high robustness with respect to outlying data. Given the low sensitivity to outliers, these can be more easily identified by analysing the residuals of the fit. The devised cost function is inspired by the definition of entropy, although the method in itself does not exploit the stochastic meaning of entropy in its usual sense. After describing the most common alternative approaches for robust identification, the novel method is presented together with numerical examples for validation.

**Keywords:** system identification; model fitting; data processing



Hochschule  
Bonn-Rhein-Sieg

Fachbereich  
Informatik

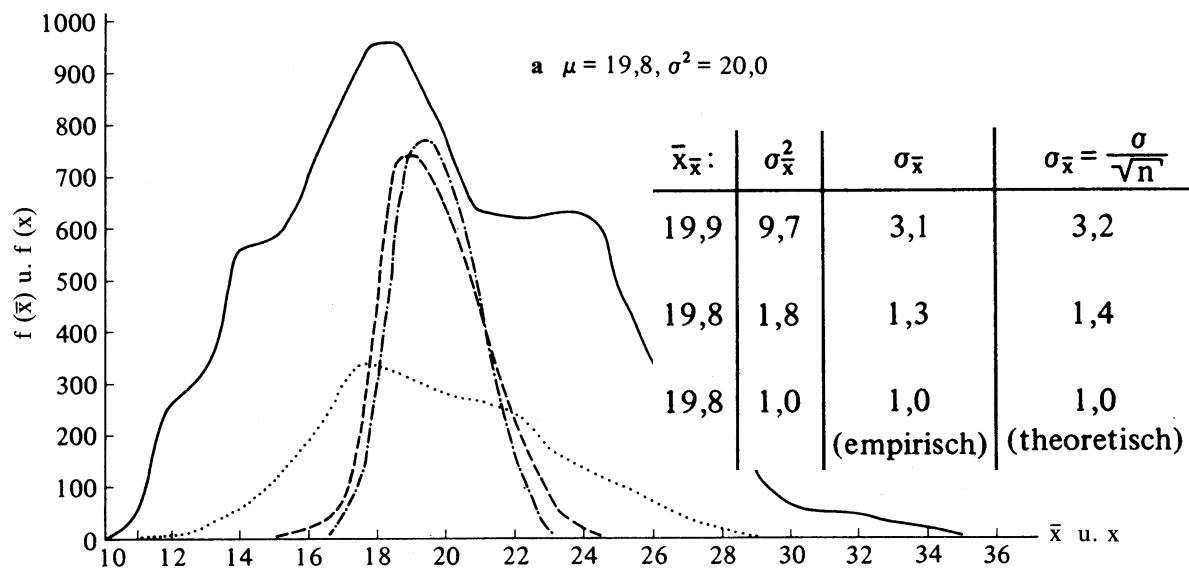
Prof. Dr.  
Paul G. Plöger

# Overview

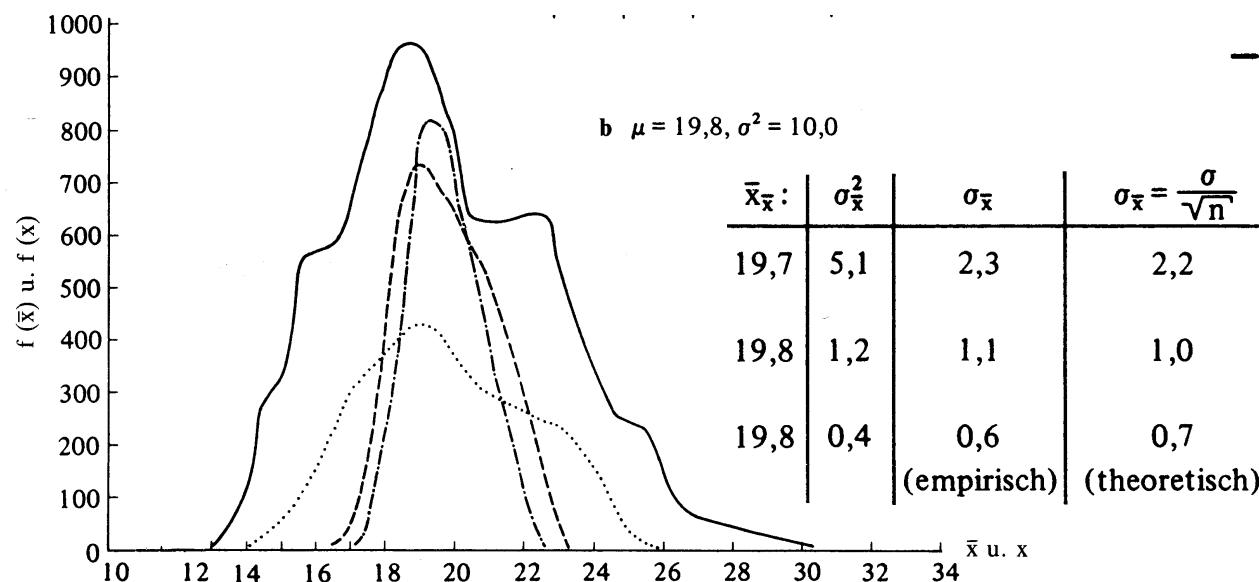
- 1) Various distributions of random Variables and their display
- 2) Characterizing distributions; central tendency and dispersion
- 3) Some theorems on distributions (with / without models)
- 4) **Standard error of empirical mean and empirical variance**
- 5) Error propagation in regression for measurements with precision



# Example Std. error



- Population
- Mittelwerteverteilung von 200 Stichproben ( $n = 2$ )
- - - Mittelwerteverteilung von 200 Stichproben ( $n = 10$ )
- - - - Mittelwerteverteilung von 200 Stichproben ( $n = 20$ )



Is dependent on  $\sigma$  of population  
Is dependent on sample size



# Standard Errors of mean / deviation

$\sigma_{\bar{x}}$  is defined as "standard error of the mean".

It is the standard deviation of equally sized samples taken from one common population.

if  $\sigma$  of population is known :

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

but in general  $\sigma$  itself must be estimated :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad s.t.$$

$$\hat{\sigma}_{\bar{x}} = \sqrt{\frac{\hat{\sigma}^2}{n}} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n(n-1)}}$$

the smaller  $\hat{\sigma}_{\bar{x}}$  is the better has  $\mu$  been estimated

standard error of the standard deviation :

$$\hat{\sigma}_s = \sqrt{\frac{\hat{\sigma}^2}{2n}}$$



```

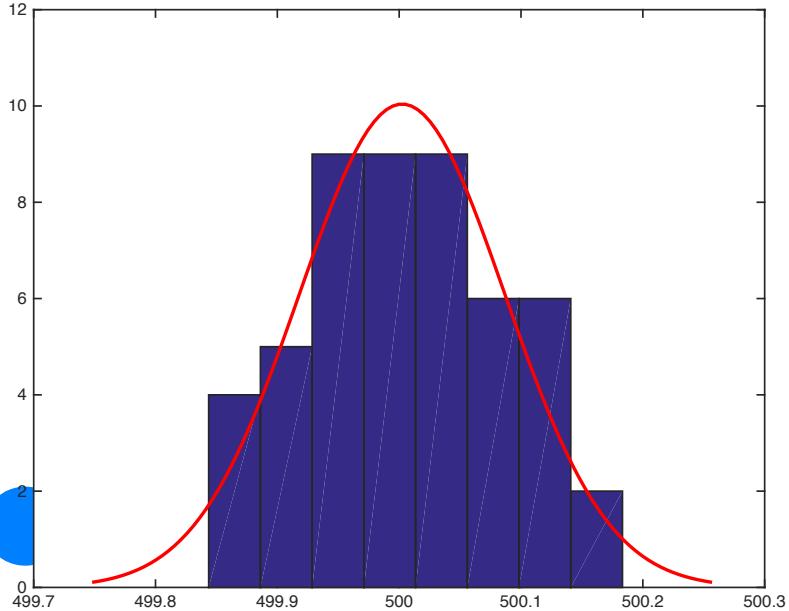
spaghetti_10000n = random('normal',2,0.05,250,10000);
spaghetti_10000 =(rand(250,10000)-0.5)*0.25+2;
%weight distrb. of 10k boxes with 250 noodles summed
box_10000n = sum(spagetti_10000n);
box_10000 = sum(spagetti_10000);
histfit(PGPStdErrorOfMeanv2(box_10000,200,50))
histfit(PGPStdErrorOfMeanv2(box_10000n,200,50))

```

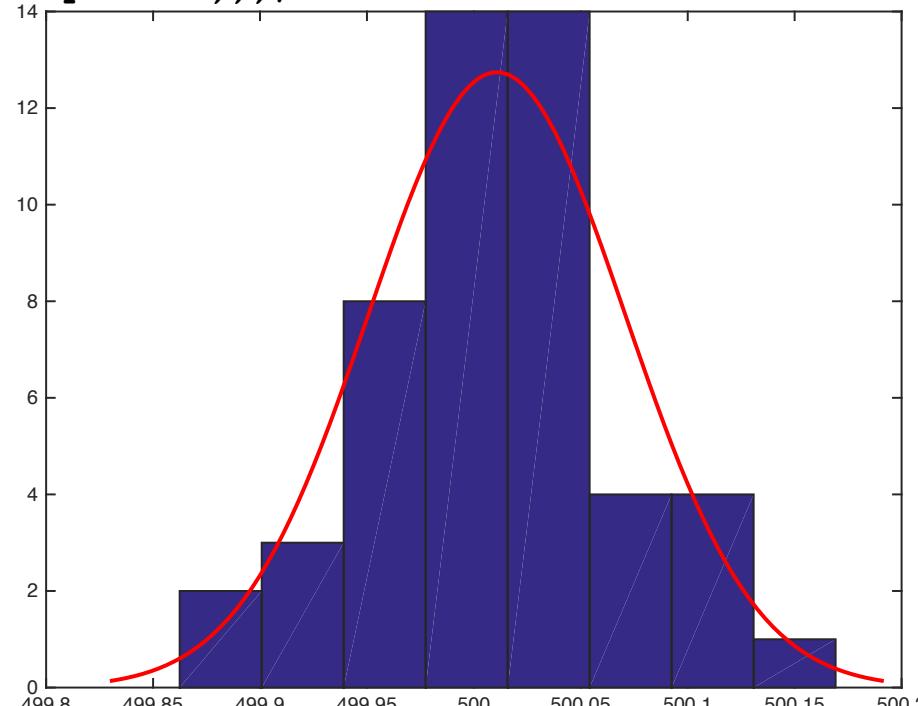
```

function[RDist]=PGPStdErrorOfMeanv2(Pdist,PStichSampleSize,PStichRepeatCnt)
%PDist : 1 dim == Grundgesamtheit
%PStichSamplesize: each sample drawn from Pdist has this size, e.g. M=200
%PStichRepeatCnt : how many samples used to calc. the mean zB n=50
boxNum=length(Pdist);
RDist = zeros(1,PStichRepeatCnt);
for k=1:PStichRepeatCnt % for clarity programmed as for loop
%grab a random sample from distribution via randperm and sum it
    cur_sum = sum(Pdist(randperm(boxNum,PStichSampleSize)));
    RDist(k)= cur_sum/ PStichSampleSize;
end
return

```



## Which spaghetti factory makes better 500g packages?



# Overview

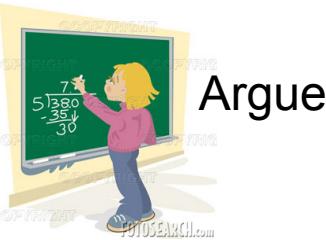
- 1) Various distributions of random Variables and their display
- 2) Characterizing distributions; central tendency and dispersion
- 3) Some theorems on distributions (with / without models)
- 4) Standard error of empirical mean and empirical variance
- 5) **Error propagation in regression for measurements with precision**



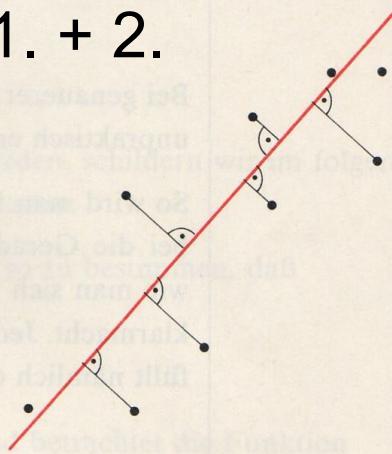
## Observation

1. Sum of distances above == sum of distances below
2. Sum of orthogonal distances is minimal
3. Sum of squares is minimal
4. sum of straight line displacements is zero
5. sum of absolute values of line displacements is minimal
6. Sum of squares of straight line displacements is minimal

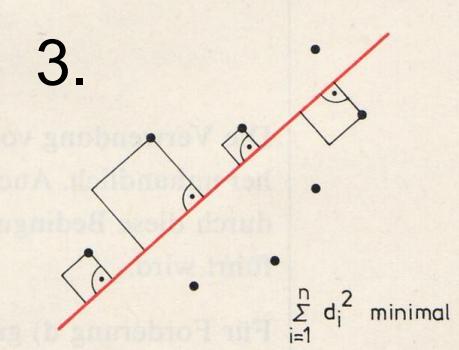
# Many metrics



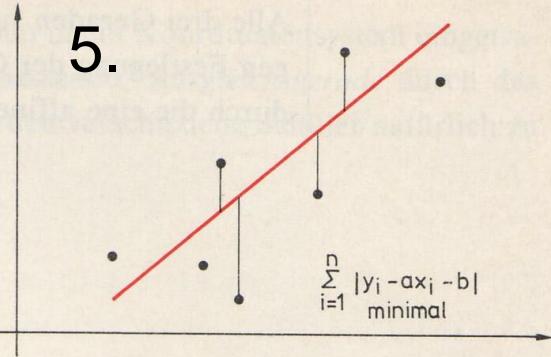
1. + 2.



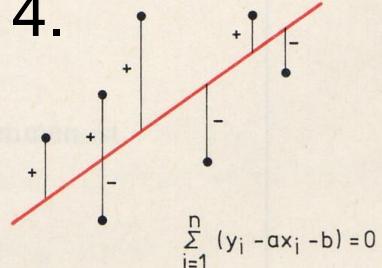
3.



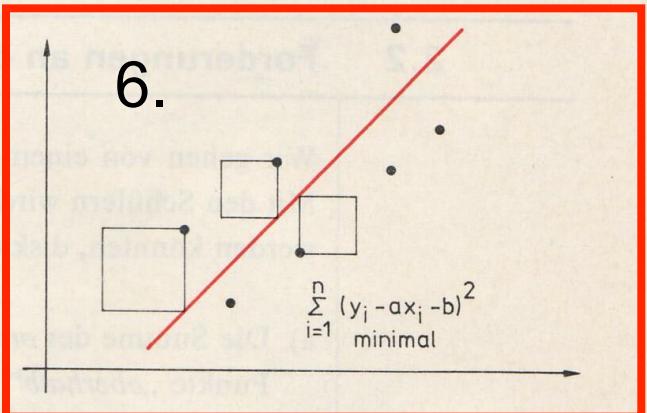
5.



4.



6.



$$F(a,b) = \sum_i (y_i - (ax_i + b))^2 \stackrel{!}{=} \text{minimum}$$

$$\frac{\partial}{\partial b} F(a,b) = -2 \sum_i (y_i - ax_i - b) = 0$$

$$\Rightarrow nb = \sum_i y_i - a \sum_i x_i$$

$$\Rightarrow b = \frac{1}{n} \sum_i y_i - a \frac{1}{n} \sum_i x_i = E[Y] - aE[X]$$

$$\frac{\partial}{\partial a} F(a,b) = \frac{\partial}{\partial a} \sum_i (y_i - (ax_i + E[Y] - aE[X]))^2 =$$

$$\frac{\partial}{\partial a} \sum_i [(y_i - E[Y]) - a(x_i - E[X])]^2 =$$

$$\frac{\partial}{\partial a} \sum_i (y_i - E[Y])^2 - 2a(y_i - E[Y])(x_i - E[X]) + a^2(x_i - E[X])^2 =$$

$$- 2 \sum_i (y_i - E[Y])(x_i - E[X]) + 2a \sum_i (x_i - E[X])^2 = 0$$

$$\Rightarrow a = \frac{\sum_i (y_i - E[Y])(x_i - E[X])}{\sum_i (x_i - E[X])^2}$$

Hochschule  
Bonn-Rhein-Sieg

Fachbereich  
Informatik

# Derivation

data model:

$$y = f(x) = ax + b$$

## Covariance

What happens if  
 $y_i$  are measured  
with uncertainties

Use the following  
abbrev.:

$$S = \sum_i \frac{1}{\sigma_i^2}$$

$$S_y = \sum_i \frac{y_i}{\sigma_i^2}$$

$$S_x = \sum_i \frac{x_i}{\sigma_i^2}$$

$$S_{xy} = \sum_i \frac{y_i x_i}{\sigma_i^2}$$

$$S_{xx} = \sum_i \frac{x_i x_i}{\sigma_i^2}$$

$$\Delta = SS_{xx} - S_x^2$$

$y_i$  with variances  $\sigma_i$

Assume the measurements of  $y_i$  have different  $\sigma_i$ .  
weight them according to their variance.

*using the new statistics :*

$$a = \frac{S_y S_{xx} - S_{xy} S_x}{\Delta} \quad b = \frac{SS_{xy} - S_x S_y}{\Delta}$$

# How do Errors in a and b depend on errors in $y_i$ ?

Observe the general law to propagate variances:

$$\sigma_F^2 = \sum_i \sigma_i^2 \left( \frac{\partial F}{\partial y_i} \right)^2$$

using the new abbreviations:

$$a = \frac{S_y S_{xx} - S_{xy} S_x}{\Delta} \quad b = \frac{SS_{xy} - S_x S_y}{\Delta}$$

Now:

$$\sigma_a^2 = \sum_i \sigma_{y_i}^2 \left( \frac{\partial a}{\partial y_i} \right)^2 \quad \frac{\partial a}{\partial y_i} = \frac{S_{xx} - S_x x_i}{\sigma_{y_i}^2 \Delta} \Rightarrow \dots \Rightarrow \sigma_a^2 = \frac{S}{\Delta}$$

$$\sigma_b^2 = \sum_i \sigma_{y_i}^2 \left( \frac{\partial b}{\partial y_i} \right)^2 \quad \frac{\partial b}{\partial y_i} = \frac{S_x x_i - S_x x_i}{\sigma_{y_i}^2 \Delta} \Rightarrow \dots \Rightarrow \sigma_b^2 = \frac{S_{xx}}{\Delta}$$

How do a and b vary if  $y_i$ 's vary?



# Reverse Argument

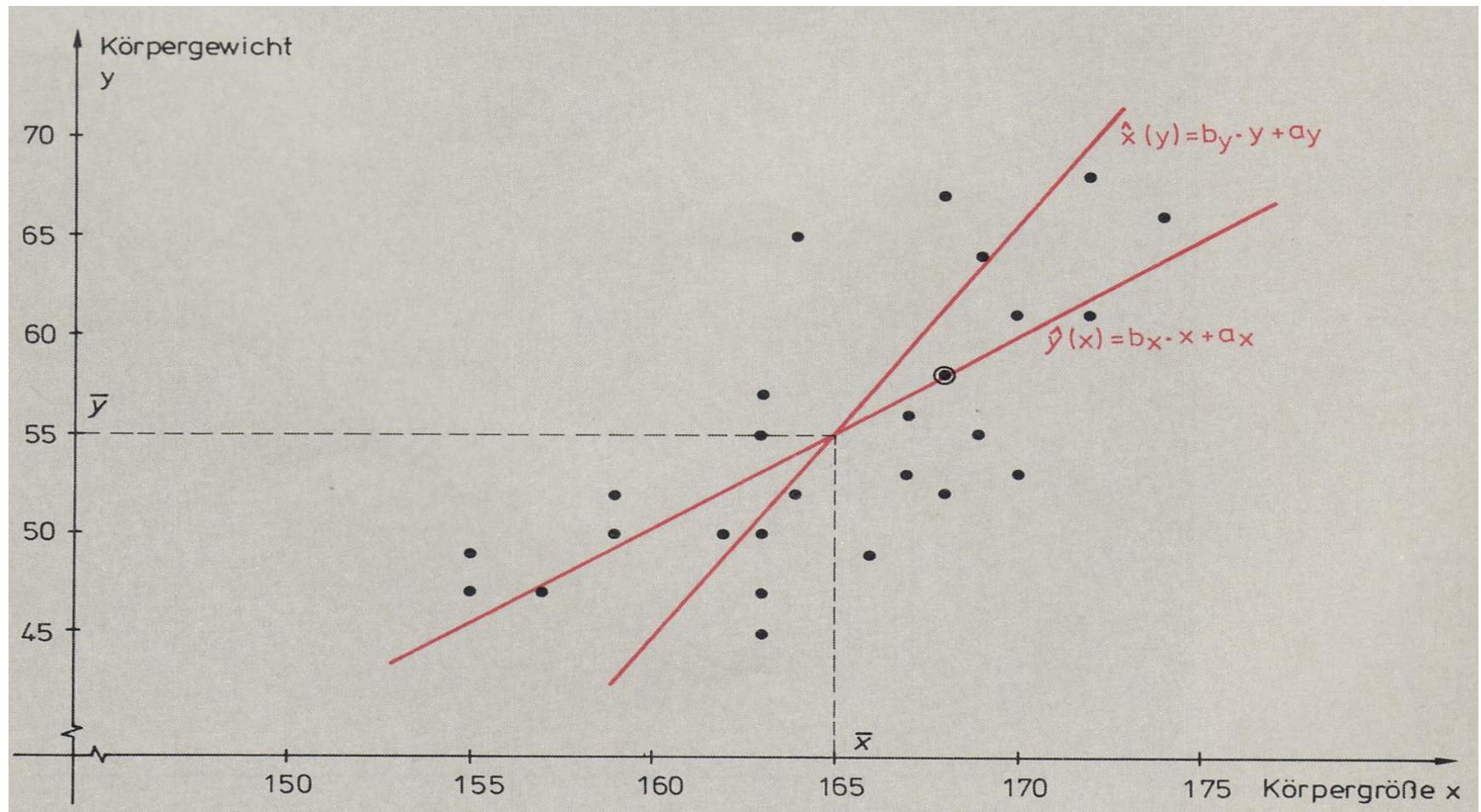
by the very same reasoning we can build:  
 $x(y) = by + a$  (dependency of feature X on  
feature Y)  
solution

$$a = E[X] - bE[Y]$$

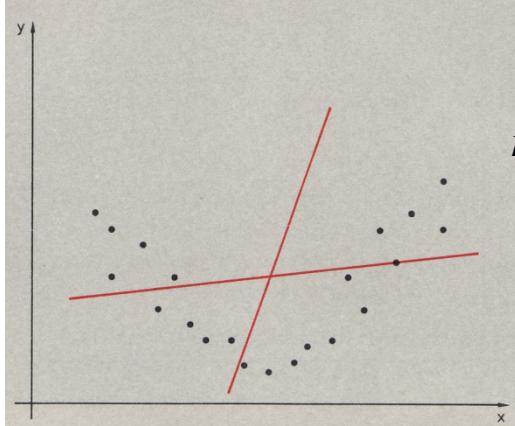
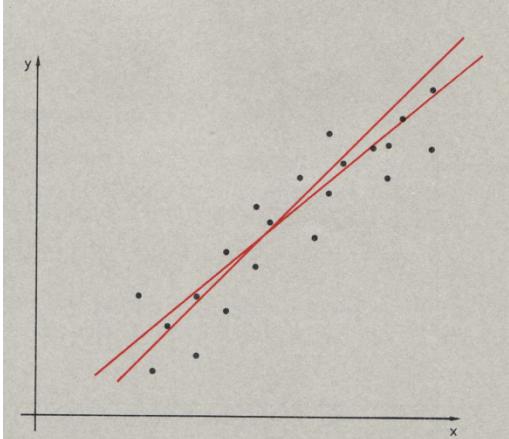
$$b = \frac{\sum_i (y_i - E[Y])(x_i - E[X])}{\sum_i (y_i - E[Y])^2}$$



# Both Regression Lines

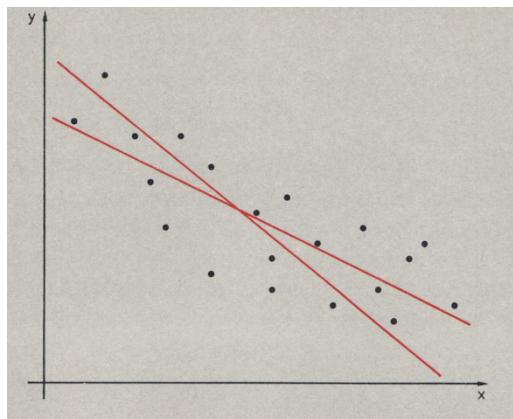
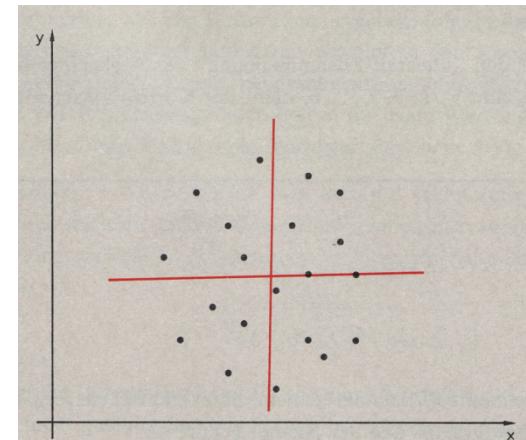


# Possible Intersections



$$r = \frac{\sum_i (x_i - E[X])(y_i - E[Y])}{\sqrt{\sum_i (x_i - E[X])^2} \sqrt{\sum_i (y_i - E[Y])^2}}$$

correlation coefficient r



unit vectors  
scalar product

cos of enclosed angle  
 $-1 \leq r \leq 1$



# Summary

the regression line measures a linear trend  
in the given data

the covariance is an important part in the  
regression line

the correlation coefficient indicates, how  
much these trends agree in the given two  
data series.



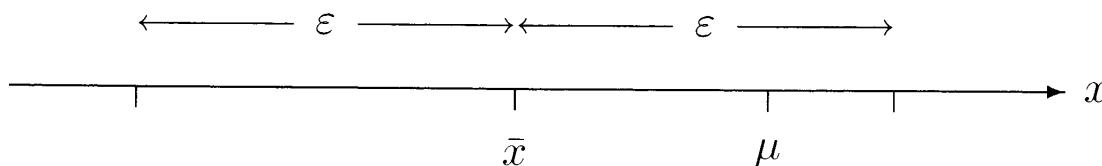
# Testing Hypothesis

- 1) How to construct tests?
- 2) How confident can we be?
- 3) How many samples needed? (sample size)

## Intervall estimation

$$P(\bar{X} - \varepsilon \leq \mu \leq \bar{X} + \varepsilon) = 0.95$$

gilt. Mit den realisierten Werten der Stichprobe ist der Punktschätzwert  $\bar{x}$  eine Zahl und das Schätzintervall  $[\bar{x} - \varepsilon, \bar{x} + \varepsilon]$  ein gewöhnliches Intervall.



Let  $X$  be a random variable,

$X_1, \dots, X_N$  a sample from  $X$ ,  
 $\theta$  a parameter (unknown).

Then – given  $\alpha$  – find (confidence)

interval  $[\hat{\theta}_L, \hat{\theta}_U]$  s.t.:

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$

(confidence level  $1 - \alpha$ )

observe:

if  $x_1, x_2, \dots, x_n$  are values of  $X_1, \dots, X_N$

then  $\hat{\theta}_L = g_l(x_1, x_2, \dots, x_n)$  and

$\hat{\theta}_U = g_u(x_1, x_2, \dots, x_n)$  are bounds

of the confidence interval.

Abbildung 5.1: Schätzintervall für  $\mu$  im Beispiel 5.7

# Required sample size to find $\mu$

Given:

a required width  $b$  of confidence interval at confidence level  $(1-\alpha)$ .  
a distribution  $X$  first normal, then relaxed.

let  $X$  be a random variable.

$X_1, \dots, X_n$  a sample thereof,  $\sigma^2 = V[X] = E[(X - \mu)^2]$

to find  $\mu = E[X]$  at confidence level  $(1-\alpha)$

and in confidence interval  $\hat{\theta}_U - \hat{\theta}_L \leq 2b$

distinguish two cases :

1) assume  $X \sim N(\mu, \sigma^2)$ ,  $\sigma^2$  known. Then every  
 $n \geq n_{\min} = \frac{u_{1-\alpha/2}^2 \sigma^2}{b^2}$  gives a confidence interval of  
at most  $2b$  width.  $u()$  is quantile of  $N(0,1)$   
at  $(1 - \alpha / 2)$ .

2) now  $X$  arbitrary,  $\sigma^2$  known, then choose  
 $n \geq n_{\min}^* = \max\{40, n_{\min}\}$ . If  $\sigma^2$  unknown replace  
 $\sigma^2$  by  $s^2$  gotten from different estimator.



# Spice bags

Assume:

$X_i \sim N(\mu, 25)$ ,  $\sigma^2 = 25$  known

- a) A sample of size  $n=20$  finds  $\bar{x} = 99.35$  gr  
 Find confidence interval at level  $(1-0.01)=0.99$

- b) Find  $n_{\min}$  using the maximum width of 1 for the confidence intervall

$$a) \hat{\theta}_{L,U} = \bar{X} \mp u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$$\text{using } n = 20, \sigma^2 = 25, u_{1-\alpha/2} = u_{0.995} = 2.58$$

$$\bar{x} = 99.35$$

$$\hat{\theta}_{L,U} = 99.35 \mp 2.58 \frac{5}{\sqrt{20}} = 99.35 \mp 2.88$$

$$\Rightarrow \text{interval is :} [96.47, 102.23]$$

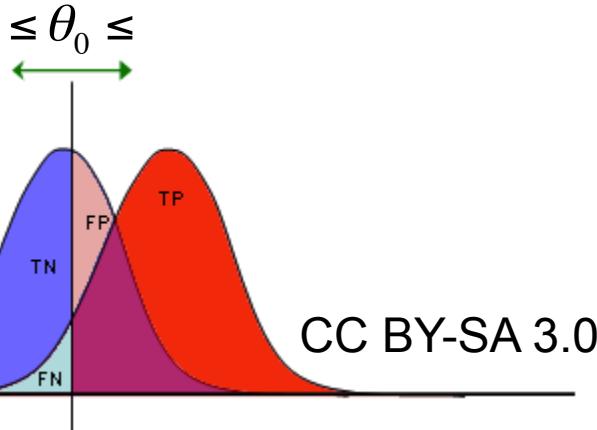
$$b) 2b = 1 \Rightarrow b = 0.5, u_{1-\alpha/2} = 2.58, \sigma^2 = 25,$$

$$n_{\min} = \frac{u_{1-\alpha/2}^2 \sigma^2}{b^2} = \frac{2.58^2 \cdot 25}{0.25} = 665.64$$

$$n \geq 666$$



# ROC Continuous Example: Fever Bound => flue yes / no



|   |      | Evidence<br>(True Class) |     |  |
|---|------|--------------------------|-----|--|
|   |      | pos                      | neg |  |
| Classifier<br>prediction<br>(Hypothesized<br>class) | Hit  | TP                       | FP  |  |
|   | Miss | FN                       | TN  |  |
| Column totals                                       |      | P                        | N   |  |

Classifier: Decide on disease by varying diagnosis temperature  $\theta_0$ :

What is the “best” classifier value to indicate: you have the flue?

**Red:** true positive (**TP**)

high temperature and will have flue  
[if a pos. instance is classified as Hit.]

**Light blue:** false negative (**FN**)

too low temp, but will get the flue  
[if a pos. instance is classified as Miss.]

**Strong blue:** true negative (**TN**)

too low temp, and no flue  
[if a neg. instance is classified as Miss.]

**Light red:** false positives (**FP**)

high temperature but never get it  
[if a neg. instance is classified as Hit.]



# From Confusion table to ROC graph

|                    |                         | True class                               |                 |
|--------------------|-------------------------|--|-----------------|
|                    |                         | P  | N               |
| Hypothesized class | Y                       | True Positives                           | False Positives |
|                    | N                       | False Negatives                          | True Negatives  |
| Column totals:     |                         | P  | N               |
| FP rate =          | $\frac{FP}{N}$          | TP rate = $\frac{TP}{P} = \text{Recall}$ |                 |
| Precision =        | $\frac{TP}{TP + FP}$    | F-score = Precision x Recall             |                 |
| Accuracy =         | $\frac{TP + TN}{P + N}$ |  |                 |

From:  
 ROC Graphs: Notes and Practical Considerations for Data Mining Researchers  
 Tom Fawcett, Intelligent Enterprise Technologies Laboratory, HP Laboratories Palo Alto,  
 slides from GIC 09

ROC graph uses for:  
 x axis == FP rate  
 y axis == TP rate

|                 |                             | Condition<br>(as determined by "Gold standard")                                     |   | Precision =<br>$\frac{\sum \text{True Positive}}{\sum \text{Test Outcome Positive}}$ |
|-----------------|-----------------------------|---|---|--|
|                 |                             | Condition Positive  |   |  |
| Test<br>Outcome | Test<br>Outcome<br>Positive | True Positive   | False Positive<br>(Type I error)  | Precision =<br>$\frac{\sum \text{True Positive}}{\sum \text{Test Outcome Positive}}$ |
|                 | Test<br>Outcome<br>Negative | False Negative<br>(Type II error)   | True Negative   |  |
|                 |                             | Sensitivity =<br>$\frac{\sum \text{True Positive}}{\sum \text{Condition Positive}}$ | Specificity =<br>$\frac{\sum \text{True Negative}}{\sum \text{Condition Negative}}$ | Accuracy   |

Co Wikipedia (en):  
 „ROC“, CC BY-SA 3.0



# ROC graphs

**ROC graphs are two-dimensional graphs in which TP rate is plotted on the Y axis and FP rate is plotted on the X axis.**

A ROC graph depicts relative trade-offs between benefits (true positives) and costs (false positives).

A discrete classifier is one that outputs only a class label. Each discrete classifier produces a pair (FP rate, TP rate), which corresponds to a single point in ROC space.

From:

ROC Graphs: Notes and Practical Considerations for Data Mining Researchers

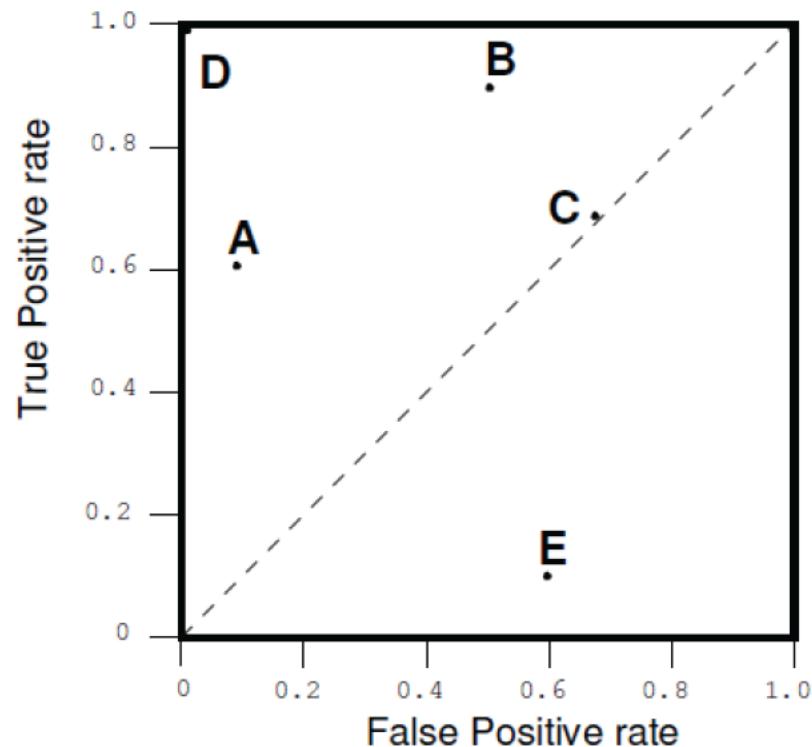
Tom Fawcett, Intelligent Enterprise Technologies

Laboratory, HP Laboratories Palo Alto,

slides from GIC 09

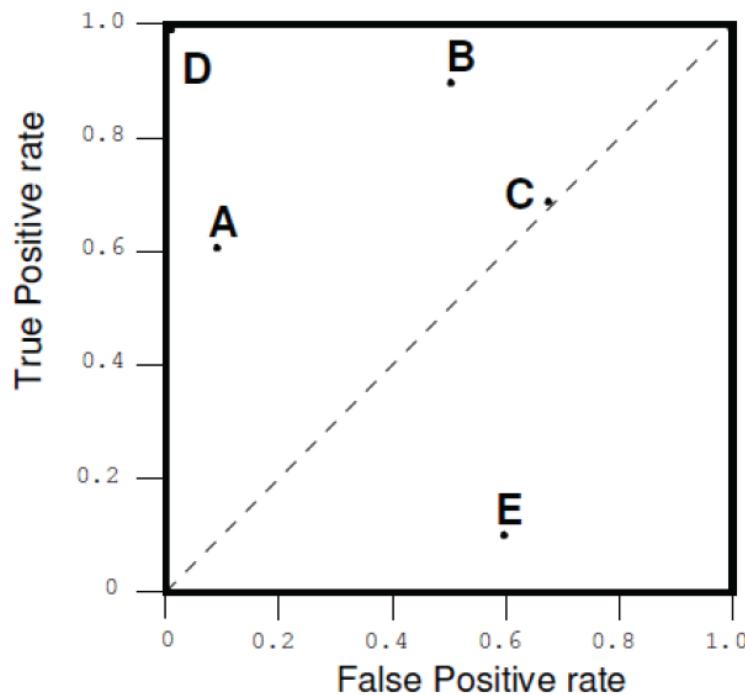
Hochschule  
Bonn-Rhein-Sieg

Fachbereich  
Informatik



A basic ROC graph showing five discrete classifiers.

# Important Points in ROC space



The lower left point (0,0) represents the strategy of never issuing a positive classification.

The opposite strategy is represented by the upper right point (1,1).

The point (0,1) represents perfect classification.

From:  
ROC Graphs: Notes and Practical Considerations for Data Mining Researchers

Tom Fawcett, Intelligent Enterprise Technologies

Laboratory, HP Laboratories Palo Alto,

slides from GIC 09  
Hochschule  
Bonn-Rhein-Sieg

Fachbereich  
Informatik

Prof. Dr.  
Paul G. Plöger

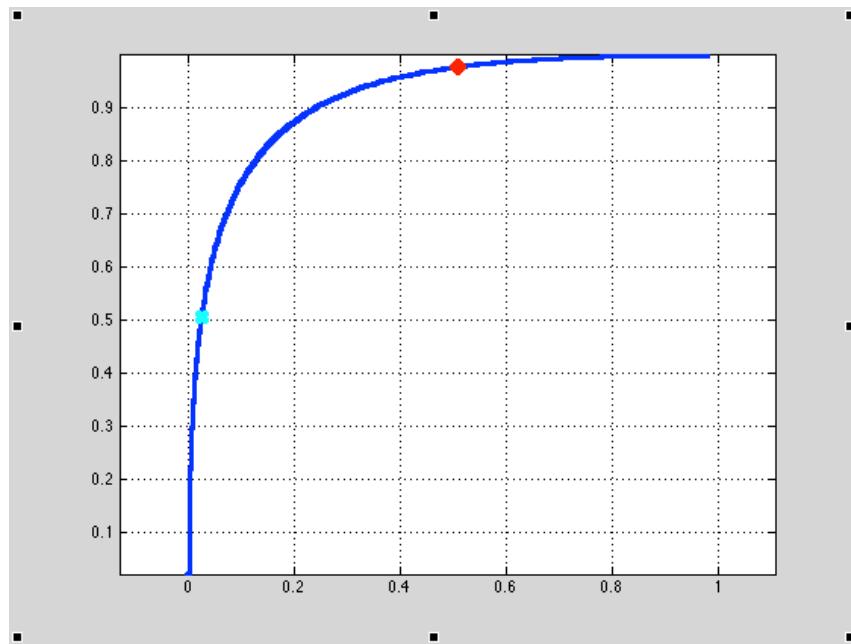
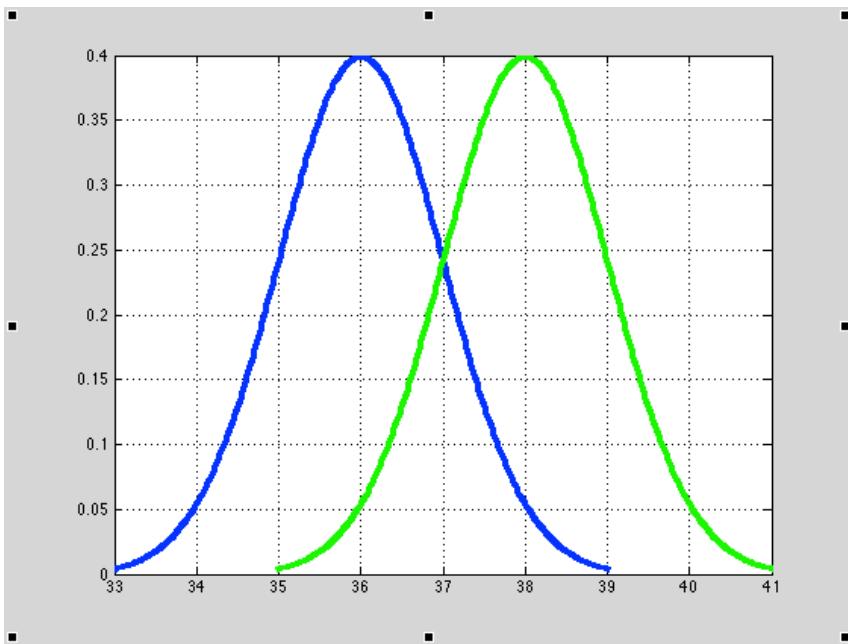
```

%two temperature distributions, healthy and sick
mu=36;
mu2=38;
sigma=1.0;
sigma2=1.0;
%sample it arround ist modes == peaks
X=linspace(mu-3,mu+3,201);
Y=linspace(mu2-3,mu2+3,201);
plot(X,normpdf(X,mu,sigma), 'b' ,Y,normpdf(Y,mu2,sigma2), 'g' );grid on;
pause;
clf;
bounds=[34:0.02:40];%these are the values for boundaries
TN=normcdf(bounds,mu,sigma);% BLUE area
FP=1-TN; % ROSE area
FN=normcdf(bounds,mu2,sigma2);% CYAN area
TP=1-FN; % RED area
TPrate=TP./(TP+FN);% build percentage or rates
FPrate=FP./(FP+TN);%
plot(FPrate,TPrate);axis equal;grid on;%MY first ROC diagram
pause;
hold on
plot(FPrate(100),TPrate(100),'rd','LineWidth',3);% right in Raw, is left in ROC
plot(FPrate(200),TPrate(200),'cx','LineWidth',3);%
Pause;

```



# Results: normcdf, ROC by hand



# Important Points in ROC space

Informally, one point in ROC space is **better** than another if it is to the **northwest** (TP rate is higher, FP rate is lower, or both) **of the first**.

Classifiers appearing on **the left hand-side of an ROC graph**, near the X axis, may be thought of as **conservative**: they make **positive classifications only with strong evidence** so they make few false positive errors, but they often have low true positive rates as well.

Classifiers on the **upper right-hand** side of an ROC graph may be thought of as **liberal**.

From:  
 ROC Graphs: Notes and Practical Considerations for Data Mining Researchers  
 Tom Fawcett, Intelligent Enterprise Technologies  
 Laboratory, HP Laboratories Palo Alto, Fachbereich Informatik  
 slides from GIO Bonn-Rhein-Sieg

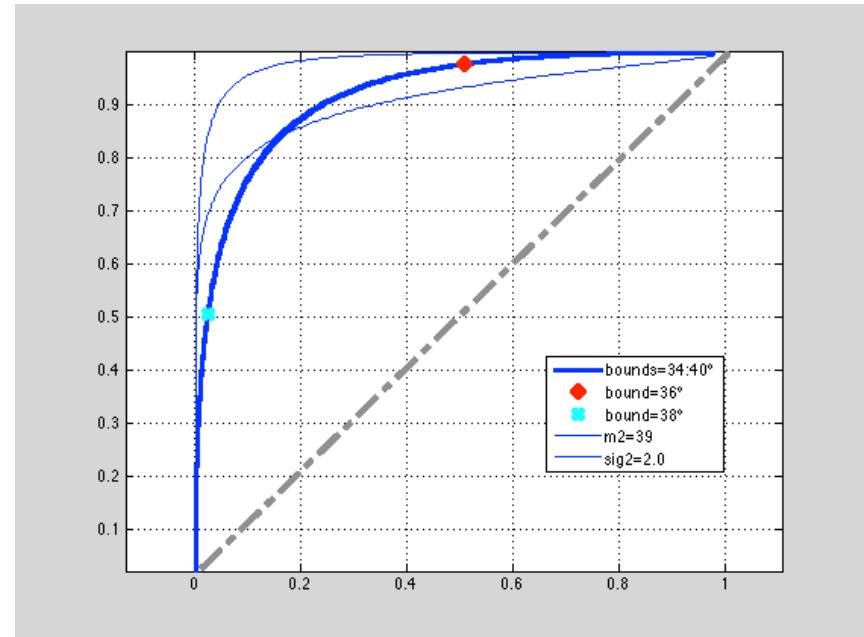
The **diagonal line  $y = x$**  represents the strategy of **randomly guessing** a class.

A random classifier will produce an ROC point that slides back and forth on the diagonal based on the frequency with which it guesses the positive class. **In order to get away from this diagonal into the upper triangular region, the classifier must exploit some information in the data.**

Any classifier that appears in the **lower right triangle performs worse than random guessing**. This triangle is therefore usually empty in ROC graphs.

# Results: 2nd ROC for mu2=39°

## 3rd ROC for sigma2=2.0



# Eg.: Application to signals

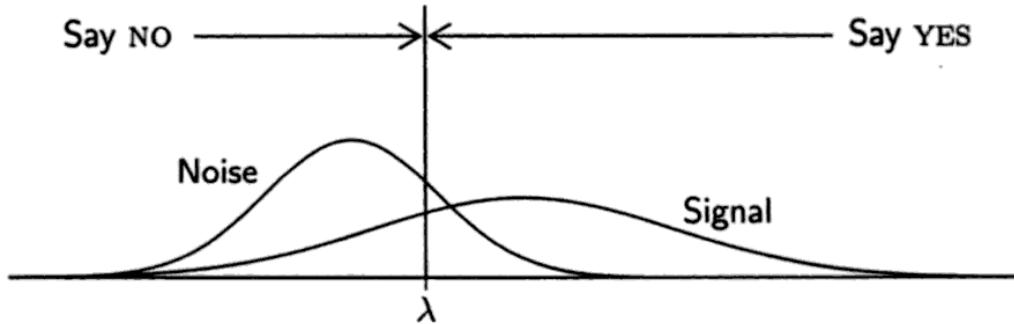


Figure 1.2: The signal and noise distributions of Figure 1.1 shown on a single axis with a decision criterion at the value  $\lambda$ .

$X_n$  : random variable for noise trials, density  $f_n(x)$

$X_s$  : random variable for signal trials, density  $f_s(x)$

false-alarm rate :  $P_F = P(YES | noise) =$

$$P(X > \lambda | noise) = P(X_n > \lambda) = \int_{\lambda}^{\infty} f_n(x) dx$$

$= 1 - F_n(\lambda)$  where  $F_n$  is Cummulative Distribution

Function of  $f_n$

# Test for two simple Hyp's

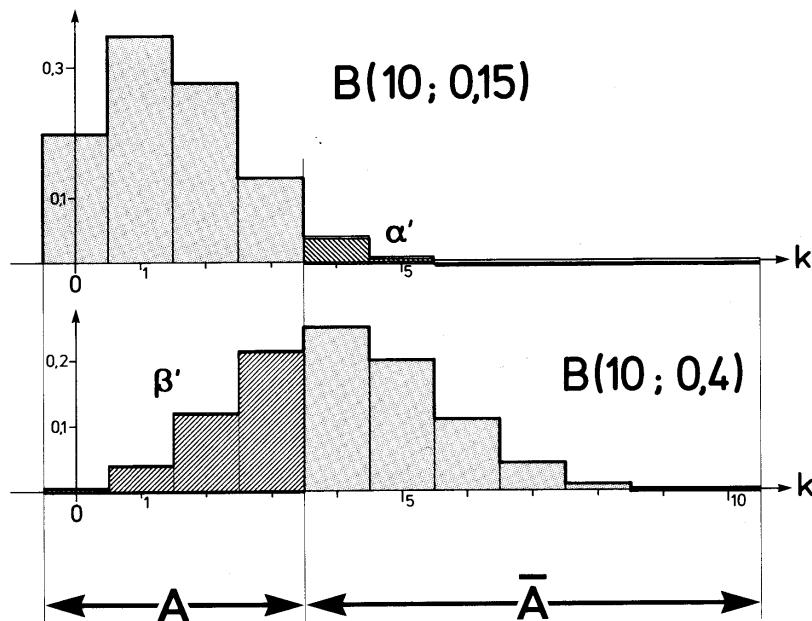


Fig. 339.1 Wahrscheinlichkeitsverteilung zur Hypothese 1 (Erste Qualität, oben) und Hypothese 2 (Zweite Qualität, unten) von Beispiel 1.  
grau: Wahrscheinlichkeit für ein richtiges Urteil  
rot: Wahrscheinlichkeit für einen Fehler 1. Art bzw. 2. Art.

$H_0$ : screws are defect according to  $B(10; 0.15)$

$H_1$ : screws are defect according to  $B(10; 0.4)$

Till  $A$  :  $H_0$  accepted  
( $A$  == Annahmebereich)

Beyond  $A$ :  $H_0$  assumed wrong,  
i.e.  $H_1$  accepted

$\alpha$  = error of 1.st kind

$\beta$  = error of 2.nd kind

Sample  $Z=(0|1|1|1|0|0|0|1|1|1|0)$

*decision rule :*

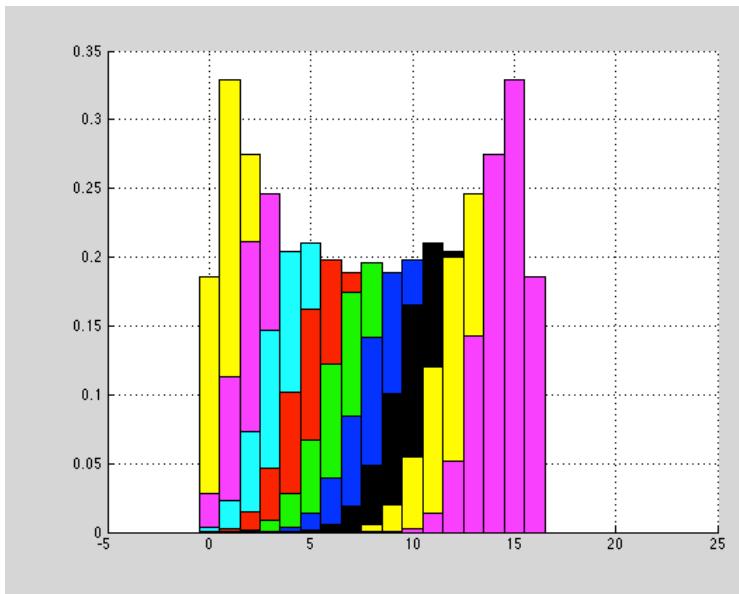
$$\delta_k : \begin{cases} Z \leq k \Rightarrow \text{decide for } H_0 \\ Z > k \Rightarrow \text{decide for } H_1 \end{cases}$$



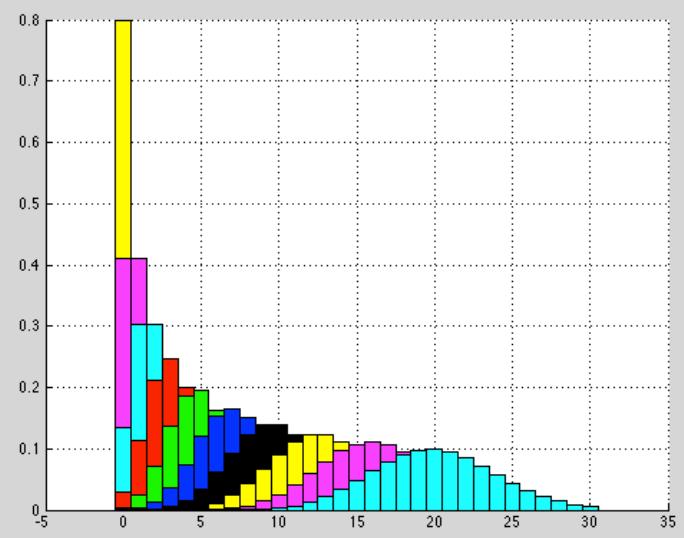
```

mycol=['y' 'm' 'c' 'r' 'g' 'b' 'k' 'y' 'm' 'c' 'r' 'g' 'b' 'k'];
hold on;trials=16;
for i=1:9
    p=0.1*i;
    bar1=bar(0:20,pdf('Binomial',0:20,trials,p));
    set(bar1,'BarWidth',1,'FaceColor',mycol(i));
end

```



# Binomial examples

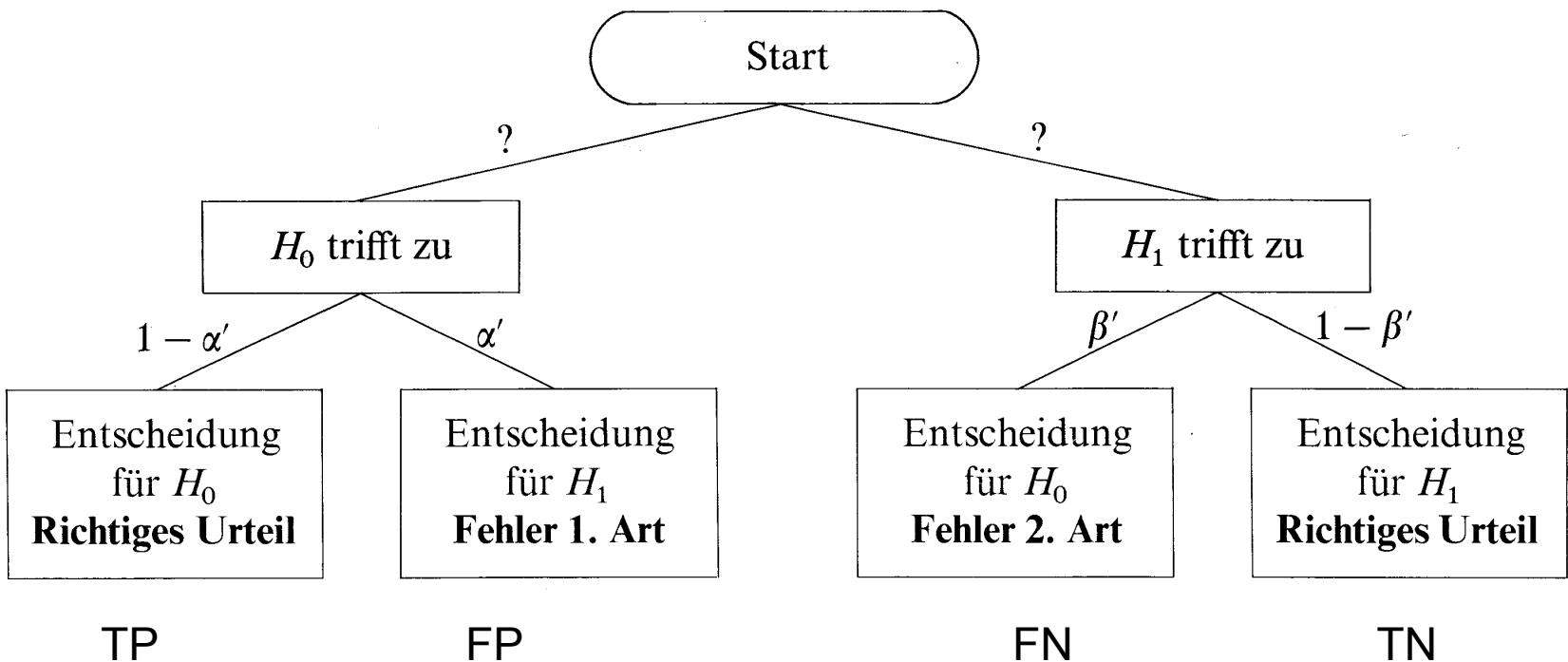


```

p=0.2; hold on;
for i=1:10
    trials=i*i;
    bar1=bar(0:30,pdf('Binomial',0:30,trials,p));
    set(bar1,'BarWidth',1,'FaceColor',mycol(i));
end

```

# $H_0/H_1$ Test with two simple Hyp's



# Outline: Evaluating Hypotheses

[Read Ch. 5] Tom Mitchell: Machine Learning  
[Recommended exercises: 5.2, 5.3, 5.4]

Error of a Sample versus true error  
Confidence intervals for observed hypothesis error  
Estimators  
Binomial distribution, Normal distribution, Central Limit Theorem  
Paired t tests  
Comparing learning methods



# Sampling

Consider learning the target function "people who plan to purchase new skis this year," given a sample of training data collected by surveying people as they arrive at a ski resort.

instance space  $X$  == space of all people (denoted by  $x$ ). Each individual  $x$  may be described by features like e.g.: age, occupation, how many times they skied last year, etc.

Some (unknown) distribution  $D$  specifies for each person  $x$  the probability that  $x$  will be encountered as the next person arriving at the ski resort.

The target function  $f: X \rightarrow \{0, 1\}$  classifies each  $x$  according to whether or not they plan to purchase skis this year.

We are interested in following two questions:

1. Given a hypothesis  $h$  and a data sample  $S$  containing  $n$  examples drawn at random according to the distribution  $D$ , what is the best estimate of the accuracy of  $h$  over future instances drawn from the same distribution?
2. What is the probable error in this accuracy estimate?



# Two Definitions of Error

The **true error** of hypothesis  $h$  with respect to target function  $f$  and distribution  $D$  is the probability that  $h$  will misclassify an instance drawn at random according to  $D$ :

$$\text{error}_D(h) \equiv \Pr_{x \in D}[f(x) \neq h(x)]$$

The **sample error** of  $h$  with respect to target function  $f$  and data sample  $S$  is the proportion of examples that  $h$  misclassifies:

$$\text{error}_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$$

**here  $\delta(f(x) \neq h(x))$  is 1 if  $f(x) \neq h(x)$ , and 0 otherwise.**  
**(decision rule)**

**Q: How well does  $\text{error}_S(h)$  estimate  $\text{error}_D(h)$ ?**



# Estimate a binomial $p$ **IS SAME** as estimate the $error_D(h)$

Estimate  $p$  from random sample of coin tosses

single toss

probability  $p$  that a coin toss yields head up

$r$  heads over  $n$  tosses =  $r/n$

Estimate  $error_D(h)$  from random sample of instances

draw single random instance  $i$  from  $D$ .

$i$  is misclassified by  $h$

$error_S(h)$



# General setting for Binom. distr.

1. There is a base, or underlying, experiment (e.g., toss of the coin) whose outcome can be described by a random variable, say  $Y$ . The random variable  $Y$  can take on only two possible values (e.g.,  $Y = 1$  if heads,  $Y = 0$  if tails).
2. The probability that  $Y = 1$  on any single trial of the underlying experiment is given by some constant  $p$ , independent of the outcome of any other experiment. The probability that  $Y = 0$  is therefore  $(1 - p)$ . Typically,  $p$  is not known in advance, and the problem is to estimate it.
3. A series of  $n$  independent trials of the underlying experiment is performed (e.g.,  $n$  independent coin tosses), producing the sequence of independent, identically distributed random variables  $Y_1, Y_2, \dots, Y_n$ . Let  $R$  denote the number of trials for which  $Y_i = 1$  in this series of  $n$  experiments:

$$R \equiv \sum_{i=1}^n Y_i$$

4. The probability that the random variable  $R$  will take on a specific value  $r$  (e.g., the probability of observing exactly  $r$  heads) is given by the Binomial distribution

$$\Pr(R = r) = \binom{n}{r} p^r (1-p)^{n-r}$$

A plot of this probability distribution is shown on slide 63.



# Problems when Estimating Error

## 1. Bias:

If  $S$  is training set,  $\text{error}_S(h)$  is optimistically biased

$$\text{bias} = E[\text{error}_S(h)] - \text{error}_D(h)$$

For unbiased estimate,  $h$  and  $S$  must be chosen independently

## 2. Variance:

Even with unbiased  $S$ ,  $\text{error}_S(h)$  may still **vary** from  $\text{error}_D(h)$



# Example

Hypothesis  $h$  misclassifies 12 of the 40 examples in  $S$

$$error_S(h) = \frac{12}{40} = 0.30$$

What is  $error_D(h)$ ?



# Estimators

Experiment:

1. choose sample  $S$  of size  $n$  according to distribution  $D$
2. measure  $\text{error}_S(h)$

$\text{error}_S(h)$  is a random variable (why?)

(i.e., the result of a random experiment)

$\text{error}_S(h)$  is an unbiased estimator for  $\text{error}_D(h)$

given the observed  $\text{error}_S(h)$  what can we conclude about  $\text{error}_D(h)$ ?

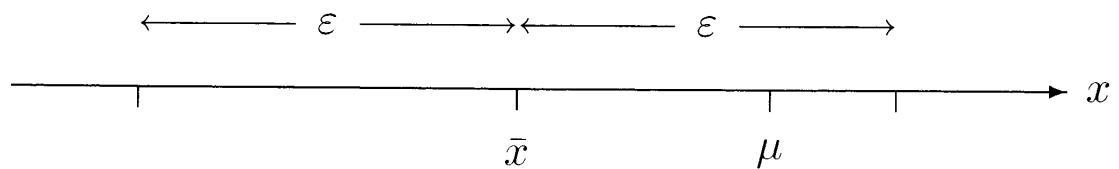


# Confidence Intervals

If

$S$  contains  $n$  examples, drawn independently of  $h$  and each other, and  $n \geq 30$

then



with approximately 95% probability,  
 $error_D(h)$  lies in interval

$$error_S(h) \pm 1.96 \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$



# Confidence Intervals

If

$S$  contains  $n$  examples, drawn independently of  $h$  and each other, and  $n \geq 30$

then

with approximately  $N\%$  probability,  
 $\text{error}_D(h)$  lies in interval

| $N\%$       | $z_N$ |
|-------------|-------|
| 50%         | 0.67  |
| 68%         | 1.00  |
| 80%         | 1.28  |
| 90%         | 1.64  |
| 95%         | 1.96  |
| 99%         | 2.58  |
| (Table 5.1) |       |

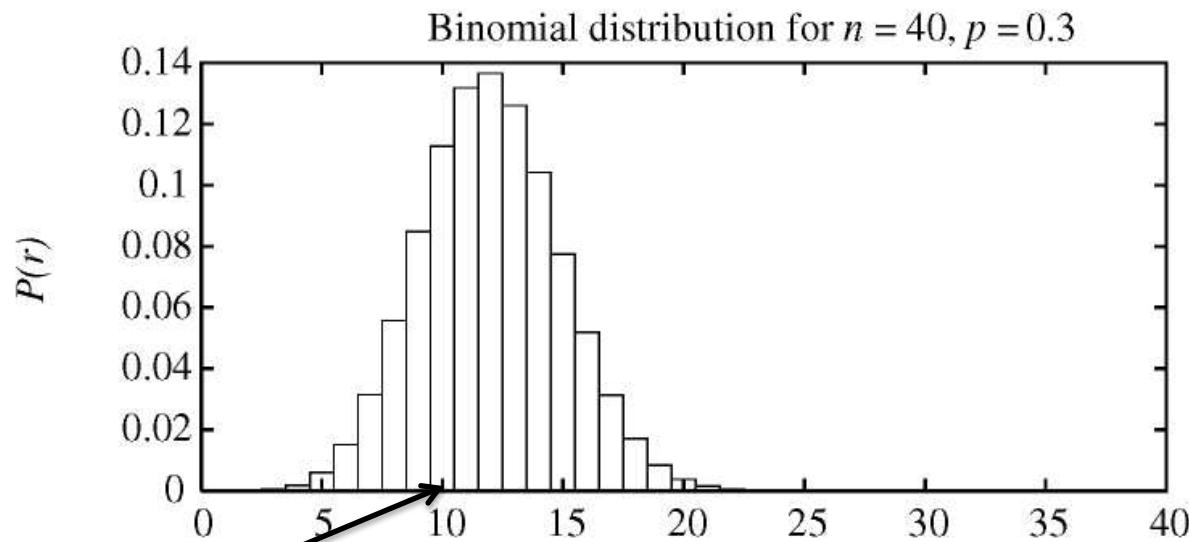
$$\text{error}_S(h) \pm z_N \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}$$



# $\text{error}_S(h)$ is a (binomial) Random Variable

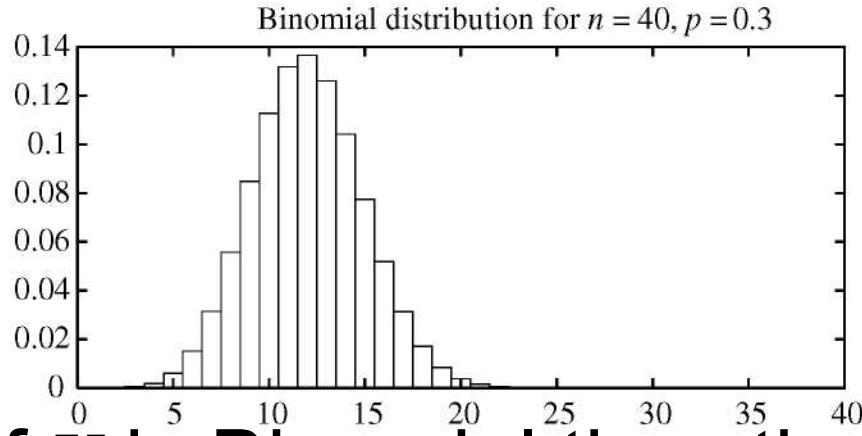
Rerun the experiment with different randomly drawn  $S$   
(of size  $n$ )

Probability of observing  $r$  misclassified examples:



O 
$$P(r) = \frac{n!}{r!(n-r)!} \text{error}_{\mathcal{D}}(h)^r (1 - \text{error}_{\mathcal{D}}(h))^{n-r}$$

# Binomial Probability Distribution



$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

If  $X$  is Binomial then the probability  $Pr(X=r)$  of  $r$  heads in  $n$  coin flips, is given by  $P(r)$

Expected, or mean value of  $X$ ,  $E[X]$ , is :

Variance of  $X$  is :

Standard deviation of  $X$ ,  $\sigma_X$ , is :

$$E[X] \equiv \sum_{i=0}^n iP(i) = np$$

$$\begin{aligned}Var(X) &\equiv E[(X - E[X])^2] \\&= np(1 - p)\end{aligned}$$

$$\begin{aligned}\sigma_X &\equiv \sqrt{E[(X - E[X])^2]} \\&= \sqrt{np(1 - p)}\end{aligned}$$



# Now: Normal Distribution approximates Binomial distr. !!!

$\text{error}_S(h)$  follows a **Binomial distribution**,

with a mean that is

$$\mu_{\text{error}_S}(h) = \text{error}_D(h)$$

and standard deviation:

$$\sigma_{\text{error}_S(h)} = \sqrt{\frac{\text{error}_D(h)(1 - \text{error}_D(h))}{n}}$$

approximate this by a **Normal distribution** with estimated mean and variance:

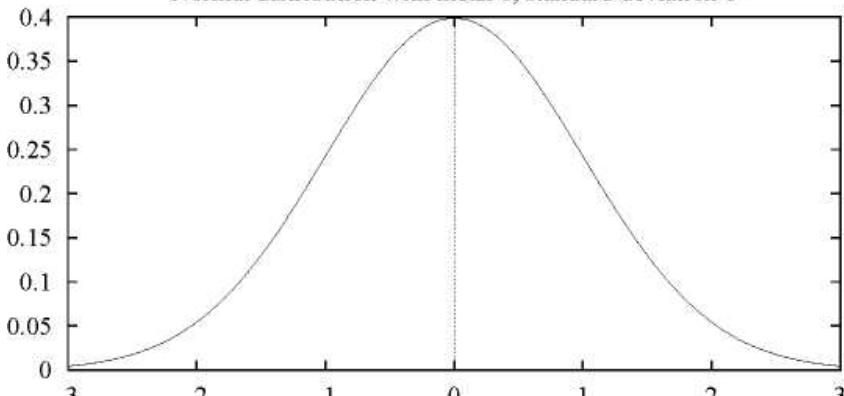
$$\mu_{\text{error}_S}(h)$$

and estimated standard deviation:

$$\sigma_{\text{error}_S(h)} \approx \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}$$



Normal distribution with mean 0, standard deviation 1



# Normal Probability Distribution

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

The probability that  $X$  will fall into the interval  $(a,b)$  is given by =>

$$\int_a^b p(x)dx$$

Expected, or mean value of  $X$ ,  $E[X]$ , is

$$\Rightarrow E[X] = \mu$$

Variance of  $X$  is

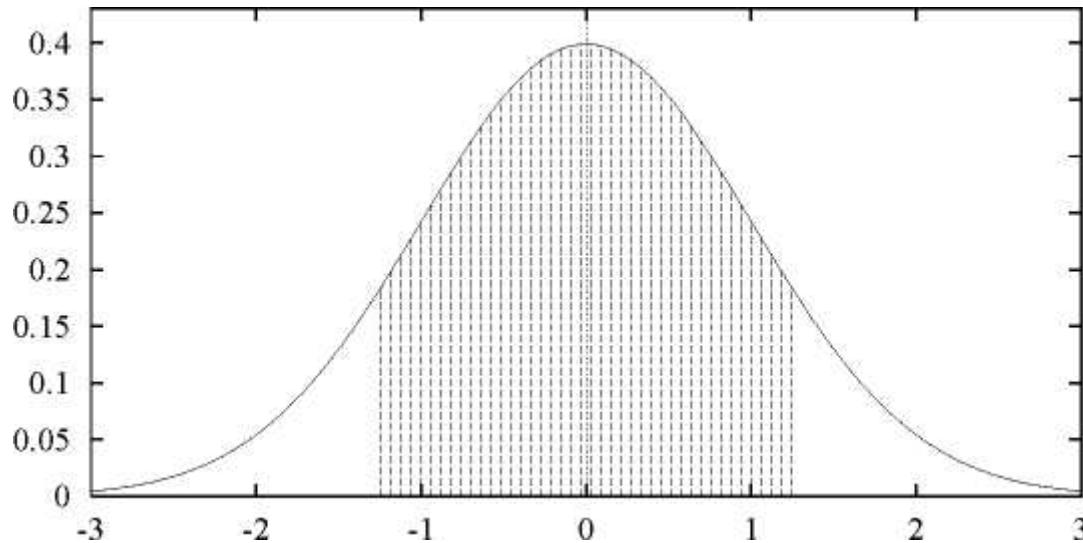
$$\Rightarrow \text{Var}(X) = \sigma^2$$

Standard deviation of  $X$ ,  $\sigma_X$ , is

$$\Rightarrow \sigma_X = \sigma$$



# Normal Probability Distribution



80% of area (probability) lies in  $\mu \pm 1.28\sigma$

N% of area (probability) lies in  $\mu \pm z_N\sigma$

|         |      |      |      |      |      |      |      |
|---------|------|------|------|------|------|------|------|
| $N\%$ : | 50%  | 68%  | 80%  | 90%  | 95%  | 98%  | 99%  |
| $z_N$ : | 0.67 | 1.00 | 1.28 | 1.64 | 1.96 | 2.33 | 2.58 |

# Confidence Intervals, More Correctly

If

$S$  contains  $n$  examples,  
drawn independently of  $h$  and each other  
and  $n > 30$

then

with approximately 95% prob.,  $\text{error}_S(h)$  lies in the interval

$$\text{error}_{\mathcal{D}}(h) \pm 1.96 \sqrt{\frac{\text{error}_{\mathcal{D}}(h)(1 - \text{error}_{\mathcal{D}}(h))}{n}}$$

equivalently,  $\text{error}_{\mathcal{D}}(h)$  lies in interval

$$\text{error}_S(h) \pm 1.96 \sqrt{\frac{\text{error}_{\mathcal{D}}(h)(1 - \text{error}_{\mathcal{D}}(h))}{n}}$$

which is approximately

$$\text{error}_S(h) \pm 1.96 \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}$$



# Central Limit Theorem

Consider a set of independent, identically distributed random variables  $Y_1, \dots, Y_n$  all governed by an arbitrary probability distribution with mean  $\mu$  and finite variance  $\sigma^2$ . Define the sample mean,

$$\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^n Y_i$$

As  $n \rightarrow \infty$ , the distribution governing  $\bar{Y}$  approaches a Normal distribution, with mean  $\mu$  and variance  $\sigma^2/n$ .



# Calculating Confidence Intervals

1. Pick parameter  $p$  to estimate:  
e.g.  $\text{error}_D(h)$
2. Choose an estimator:  
e.g.  $\text{error}_S(h)$
3. Determine probability distribution that governs the estimator:  
 $\text{error}_S(h)$  is governed by Binomial Distribution, which in turn is approximated by a Normal Distribution (if  $n \geq 30$ )
4. Find the interval  $(\theta_L, \theta_U)$ , such that  $N\%$  of probability mass falls in this interval:  
i.e. use table of  $z_N$  values



# Difference Between Hypotheses

Test  $h_1$  on sample  $S_1$ , test  $h_2$  on  $S_2$  (Section 5.5)

1. Pick parameter to estimate

$$d \equiv \text{error}_D(h_1) - \text{error}_D(h_2)$$

2. Choose an estimator

$$\hat{d} \equiv \text{error}_{S_1}(h_1) - \text{error}_{S_2}(h_2)$$

3. Determine probability distribution that governs

$$\text{estimator } \sigma_{\hat{d}} \approx \sqrt{\frac{\text{error}_{S_1}(h_1)(1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2)(1 - \text{error}_{S_2}(h_2))}{n_2}}$$

4. Find interval  $(L, U)$  such that  $N\%$  of probability mass falls in the interval

$$\hat{d} \pm z_N \sqrt{\frac{\text{error}_{S_1}(h_1)(1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2)(1 - \text{error}_{S_2}(h_2))}{n_2}}$$



# Paired $t$ test to compare $h_A, h_B$

Partition data into  $k$  disjoint test sets  $T_1, T_2, \dots, T_k$  of equal size, where this size is at least 30.

For  $i$  from 1 to  $k$ , do

$$\delta_i \leftarrow \text{error}_{T_i}(h_A) - \text{error}_{T_i}(h_B)$$

Return the value  $\delta$ , where

$$\bar{\delta} = \frac{1}{k} \sum_{i=1}^k \delta_i$$

$N\%$  confidence interval estimate for  $d$ :

$$\bar{\delta} \pm t_{N, k-1} s_{\bar{\delta}} \quad (5.17)$$

$$s_{\bar{\delta}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2}$$

Note:  $\delta_i$  are approximately Normally distributed

|                                     | Confidence level $N$ |      |      |      |
|-------------------------------------|----------------------|------|------|------|
|                                     | 90%                  | 95%  | 98%  | 99%  |
| $v = 2$                             | 2.92                 | 4.30 | 6.96 | 9.92 |
| $v = 5$                             | 2.02                 | 2.57 | 3.36 | 4.03 |
| $v = 10$                            | 1.81                 | 2.23 | 2.76 | 3.17 |
| $v = 20$                            | 1.72                 | 2.09 | 2.53 | 2.84 |
| $v = 30$                            | 1.70                 | 2.04 | 2.46 | 2.75 |
| Prof. Dr. 120                       | 1.66                 | 1.98 | 2.36 | 2.62 |
| Paul G. Plöger<br>$v \equiv \infty$ | 1.64                 | 1.96 | 2.33 | 2.58 |



# Comparing learning algorithms $L_A$ and $L_B$

What we'd like to estimate:

$$E_{S \subset D}[\text{error}_D(L_A(S)) - \text{error}_D(L_B(S))]$$

where  $L(S)$  is the hypothesis output by learner  $L$  using training set  $S$

i.e., the expected difference in true error between hypotheses output by learners  $L_A$  and  $L_B$ , when trained using randomly selected training sets  $S$  drawn according to distribution  $D$ .

But, given limited data  $D_0$ , what is a good estimator?

We could partition  $D_0$  into training set  $S$  and training set  $T_0$ , and measure

$$\text{error}_{T_0}(L_A(S_0)) - \text{error}_{T_0}(L_B(S_0))$$

even better, repeat this many times and average the results (next slide)



# Comparing learning algorithms $L_A$ and $L_B$

1. Partition data  $D_0$  into  $k$  disjoint test sets  $T_1, T_2, \dots, T_k$ , of equal size, where this size is at least 30.
2. For  $i$  from 1 to  $k$ , do  
*use  $T_i$  for the test set, and the remaining data for training set  $S_i$*

$$S_i \leftarrow \{D_0 - T_i\}$$

$$h_A \leftarrow L_A(S_i)$$

$$h_B \leftarrow L_B(S_i)$$

$$\delta_i \leftarrow \text{error}_{Ti}(h_A) - \text{error}_{Ti}(h_B)$$

3. Return the value

$$\bar{\delta} = \frac{1}{k} \sum_{i=1}^k \delta_i$$

Prof. Dr.  
Paul G. Pföger



# Comparing learning algorithms $L_A$ and $L_B$

Notice we'd like to use the paired t test on  $\delta$  to obtain a confidence interval

but not really correct, because the training sets in this algorithm are not independent (they overlap!)

more correct to view algorithm as producing an estimate of

$$E_{S \subset D_0} [error_D(L_A(S)) - error_D(L_B(S))] \quad (5.16)$$

instead of

$$E_{S \subset D} [error_D(L_A(S)) - error_D(L_B(S))] \quad (5.14)$$

but even this approximation is better than no comparison



# Summary

Statistical theory provides a basis for estimating the true error ( $\text{error}_D(h)$ ) of a hypothesis  $h$ , based on its observed error ( $\text{error}_S(h)$ ) over a sample  $S$  of data. For example, if  $A$  is a discrete-valued hypothesis and the data sample  $S$  contains  $n > 30$  examples drawn independently of  $h$  and of one another, then the  $N\%$  confidence interval for  $\text{error}_D(h)$  is approximately

$$\text{error}_S(h) \pm z_N \sqrt{\frac{\text{error}_S(h)(1 - \text{error}_S(h))}{n}}$$

where values for  $z_N$  are given in Table 5.1.

In general, the problem of estimating confidence intervals is approached by identifying the parameter to be estimated (e.g.,  $\text{error}_D(h)$ ) and an estimator (e.g.,  $\text{error}_S(h)$ ) for this quantity. Because the estimator is a random variable (e.g.,  $\text{error}_S(h)$  depends on the random sample  $S$ ), it can be characterized by the probability distribution that governs its value. Confidence intervals can then be calculated by determining the interval that contains the desired probability mass under this distribution.

One possible cause of errors in estimating hypothesis accuracy is estimation bias. If  $Y$  is an estimator for some parameter  $p$ , the estimation bias of  $Y$  is the difference between  $p$  and the expected value of  $Y$ . For example, if  $S$  is the training data used to formulate hypothesis  $h$ , then  $\text{error}_S(h)$  gives an optimistically biased estimate of the true error  $\text{error}_D(h)$ .



# Summary

A second cause of estimation error is variance in the estimate. Even with an unbiased estimator, the observed value of the estimator is likely to vary from one experiment to another. The variance  $\sigma^2$  of the distribution governing the estimator characterizes how widely this estimate is likely to vary from the correct value. This variance decreases as the size of the data sample is increased.

Comparing the relative effectiveness of two learning algorithms is an estimation problem that is relatively easy when data and time are unlimited, but more difficult when these resources are limited. One possible approach described in this chapter is to run the learning algorithms on different subsets of the available data, testing the learned hypotheses on the remaining data, then averaging the results of these experiments.

In most cases considered here, deriving confidence intervals involves making a number of assumptions and approximations. For example, the above confidence interval for  $error_D(h)$  involved approximating a Binomial distribution by a Normal distribution, approximating the variance of this distribution, and assuming instances are generated by a fixed, unchanging probability distribution. While intervals based on such approximations are only approximate confidence intervals, they nevertheless provide useful guidance for designing and interpreting experimental results in machine learning.



## EXERCISES (5 out of 8)

1. Suppose you test a hypothesis  $h$  and find that it commits  $r = 300$  errors on a sample  $S$  of  $n = 1000$  randomly drawn test examples. What is the standard deviation is  $\text{error}_S(h)$ ? How does this compare to the standard deviation in the example at the end of Section 5.3.4?
2. Consider a learned hypothesis,  $h$ , for some Boolean concept. When  $h$  is tested on a set of 100 examples, it classifies 83 correctly. What is the standard deviation and the 95% confidence interval for the true error rate for  $\text{Error}_D(h)$ ?
3. Suppose hypothesis  $h$  commits  $r = 10$  errors over a sample of  $n = 65$  independently drawn examples. What is the 90% confidence interval (two-sided) for the true error rate? What is the 95% one-sided interval (i.e., what is the upper bound  $U$  such that  $\text{error}_D(h) \leq U$  with 95% confidence)? What is the 90% one-sided interval?
4. You are about to test a hypothesis  $h$  whose  $\text{error}_D(h)$  is known to be in the range between 0.2 and 0.6. What is the minimum number of examples you must collect to assure that the width of the two-sided 95% confidence interval will be smaller than 0.1?
5. Give general expressions for the upper and lower one-sided  $N\%$  confidence intervals for the difference in errors between two hypotheses tested on different samples of data. Hint: Modify the expression given in Section 5.5.
6. Explain why the confidence interval estimate given in Equation (5.17) applies to estimating the quantity in Equation (5.16), and not the quantity in Equation (5.14).
7. proof  $E[X]$  for binomial ==  $np$     8. Prove: for binomial  $\text{Var}(X) = np(1-p)$