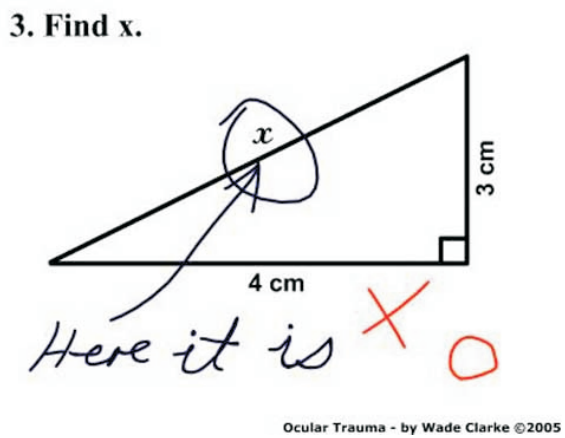
Biometrics (CSE 40537/60537)

Lecture 9: Statistical evaluation of biometrics

Adam Czajka

Biometrics and Machine Learning Group
Warsaw University of Technology, Poland



Fall 2014
University of Notre Dame, IN, USA

Lecture 9: Statistical evaluation of biometrics

Evaluation of a biometric system

We have designed and built a great (we hope so!) biometric system (including the sensor, different algorithms, technical and administrative processes, etc.). We are proud of it :-)

We have found a customer willing to pay ample amount of money. What questions should we answer before delivering our solutions?

Evaluation of a biometric system

Questions

Evaluation of a biometric system

Questions

1. Measurement of living subjects → uncertainty: how to describe uncertainty?

Evaluation of a biometric system

Questions

1. Measurement of living subjects $\rightarrow$ uncertainty: how to describe uncertainty?
2. Performance analysis is based on the experiment: what experiment(s) should we carry out?

Evaluation of a biometric system

Questions

1. Measurement of living subjects $\rightarrow$ uncertainty:
   how to describe uncertainty?
2. Performance analysis is based on the experiment:
   what experiment(s) should we carry out?
3. Authentication decisions depend on acceptance thresholds:
   how to set these thresholds?

Evaluation of a biometric system

Questions

1. Measurement of living subjects $\rightarrow$ uncertainty:
   how to describe uncertainty?
2. Performance analysis is based on the experiment:
   what experiment(s) should we carry out?
3. Authentication decisions depend on acceptance thresholds:
   how to set these thresholds?
4. Experiments need a database (data sample):
   how large should be our data sample?

Evaluation of a biometric system

Questions

1. Measurement of living subjects → uncertainty:
   how to describe uncertainty?
2. Performance analysis is based on the experiment:
   what experiment(s) should we carry out?
3. Authentication decisions depend on acceptance thresholds:
   how to set these thresholds?
4. Experiments need a database (data sample):
   how large should be our data sample?
5. We have calculated basic system parameters, e.g., EER=0:
   what error rates can we guarantee?

ISO/IEC 19795 biometric evaluation

1. Technology evaluation: off-line evaluation of algorithm(s) for a single biometric modality with the use of existing or specially prepared data sample
2. Scenario evaluation: off-line or on-line evaluation performed in a specific scenario (modeling target application) with the use of specific subjects
3. Operational evaluation: on-line evaluation of the system in real environment

ISO/IEC 19795 biometric evaluation

1. Technology evaluation: off-line evaluation of algorithm(s) for a single biometric modality with the use of existing or specially prepared data sample
   - typically: evaluation of a single, selected component (algorithm, sensor, procedure, etc.)
   - fixed data sample, acquired by a 'universal' sensor
   - results depend on environment and population used to generate a data sample
   - data sample must not be known to the authors of the evaluated component
   - repeated evaluations provide exactly the same results
   - significant underestimation of error rates expected in the operational environment

   Example: evaluation of your iris recognition algorithm developed during the third assignment

ISO/IEC 19795 biometric evaluation

2. Scenario evaluation: off-line or on-line evaluation performed in a specific scenario (modeling target application) with the use of specific subjects
   - evaluation of the complete system
   - evaluation environment close to what we expect in reality
   - each system has its own sensor
   - we can repeat the evaluation if we can control the environment and the population
   - some underestimation of error rates expected in the operational environment

   Example: evaluation of iris recognition system on simulated border control

ISO/IEC 19795 biometric evaluation

3. Operational evaluation: on-line evaluation of the system in real environment
   - authentic systems, service, environment, users, etc.
   - environment may be uncontrolled and may change unexpectedly
   - evaluation results are not repeatable
   - the best evaluation of the system, but often costly and difficult to be organized

   Example: evaluation of iris recognition system for real frequent-flyers

Lecture 9: Statistical evaluation of biometrics

Random sample

1. General population X
   - random variable (one- or multi-dimensional), which delivers all interesting information about the examined phenomenon
2. Random sample $X_1, \ldots, X_n$ of the random variable X
   - result of sampling of the random variable X
   - if $X_1, \ldots, X_n$ are independent, we say that the random sample is simple (or i.i.d. = independent and identically-distributed)

In biometrics:

- general population = all the data related to all potential users of our biometric system
- random sample = subset of data used in our evaluation

Building the random sample in biometrics

1. Presentation of biometric characteristics
2. Attempt consisting of one or multiple presentations (to select the best one)
3. Transaction consisting of one or multiple attempts (to apply some decision policy)

Probabilistic theory vs. statistics

1. Probabilistic theory
   - distributions of random variables are known, hence ...
   - ... we calculate probabilities based on known distributions
2. Statistics
   - statistic = any function of a random variable
   - statistic is also a random variable
   - distributions of random variables are not fully known, hence ...
   - ... we make an inference upon interesting quantities related to the general population $\rightarrow$ estimation and hypothesis testing

In biometrics: statistic = any function using biometric data (for instance matching scores) to estimate some quantity (for instance average matching score)

Statistical inference (part I): point estimation

1. Assessment of unknown distribution properties
   based on statistic values (i.e., numbers)
2. Estimator: any statistic $\hat{\theta}_n = g(X_1, \ldots, X_n)$
   used to estimate unknown parameter θ
3. We especially like estimators that are:
   - unbiased, i.e. $\mathcal{E}\hat{\theta} = \theta$
   - consistent, i.e. $\bigwedge_{\epsilon>0} \lim_{n\to\infty} \mathcal{P}(|\hat{\theta_n} - \theta| < \epsilon) = 1$

Example estimators in biometrics:

- mean $\hat{m}$ of the matching scores $m_1, m_2, ..., m_M$
  (estimator of the expected value m of the random variable X)
- FNMR (False Non-Match Rate)
- FMR (False Match Rate)

Lecture 9: Statistical evaluation of biometrics

Failure to Acquire (FTA)

1. Common reasons of failure to acquire
   - biometric characteristic could not be presented (due to illness, non-conformant presentation, etc.)
   - sensor's failure
   - data processing failure (e.g. segmentation failed)
   - low quality of data
2. Estimator of the failure to acquire probability (FTA) calculated for attempts:

$$\text{FTA} = \frac{\text{number of failed attempts}}{\text{number of all attempts}}$$

Failure to Enroll (FTE)

1. Common reasons of failure to enroll
   - as for FTA and:
   - biometric features could not be calculated
   - test verification failed
2. Estimator of the failure to enroll probability (FTE):

$$\text{FTE} = \frac{\text{number of failed enrollments}}{\text{number of all enrollments}}$$

3. NOTE: restrictive enrollment procedures increase FTE, but decrease FMR/FNMR

Lecture 9: Statistical evaluation of biometrics

Types of biometric samples

1. Genuine: originating from the same class (the same eye, finger, signature, etc.)
2. Impostor or random forgeries or zero-effort attempts: originating from different classes in a form as they were acquired (different eye, finger, signature, etc.)
3. Skilled forgeries: prepared (with some effort) to imitate the sample of a given class (iris printout, gummy finger, skilled forgery in signature recognition, etc.)

Matching and decision making

1. Matching
   - calculation of matching score or similarity/dissimilarity score

     similarity $= 0 \rightarrow$ samples are totally unalike
     dissimilarity $= 0 \rightarrow$ samples are identical
   - NOTE: 'similarity' and 'dissimilarity' are not always complementary
2. Possible decisions
   - match/non-match based on comparison the similarity/dissimilarity score with decision threshold

Statistical inference (part II): hypothesis testing

1. Hypotheses
   - null hypothesis (H_0)
     - the hypothesis that is tested (assumed to be true)
     - we try to find reasons to reject H_0
     - acceptance of H_0 does not mean that H_0 is true; it only means that we did not find reasons to reject it
   - alternative hypothesis (H_1)
     - we are leaning towards accepting H_1 if H_0 is rejected
2. We consider the simplest case
   - H_0 and H_1 are simple hypotheses, namely:
     - completely specify the population distribution
     - we assume that we know this distribution

Statistical inference (part II): hypothesis testing

Examples

Null and simple hypothesis H_0

- iris X is the iris of Adam Czajka

Alternative and simple hypothesis H_1

- iris X is the iris of John Doe
- iris X is one of the N known irides

Alternative and composite hypothesis H_1

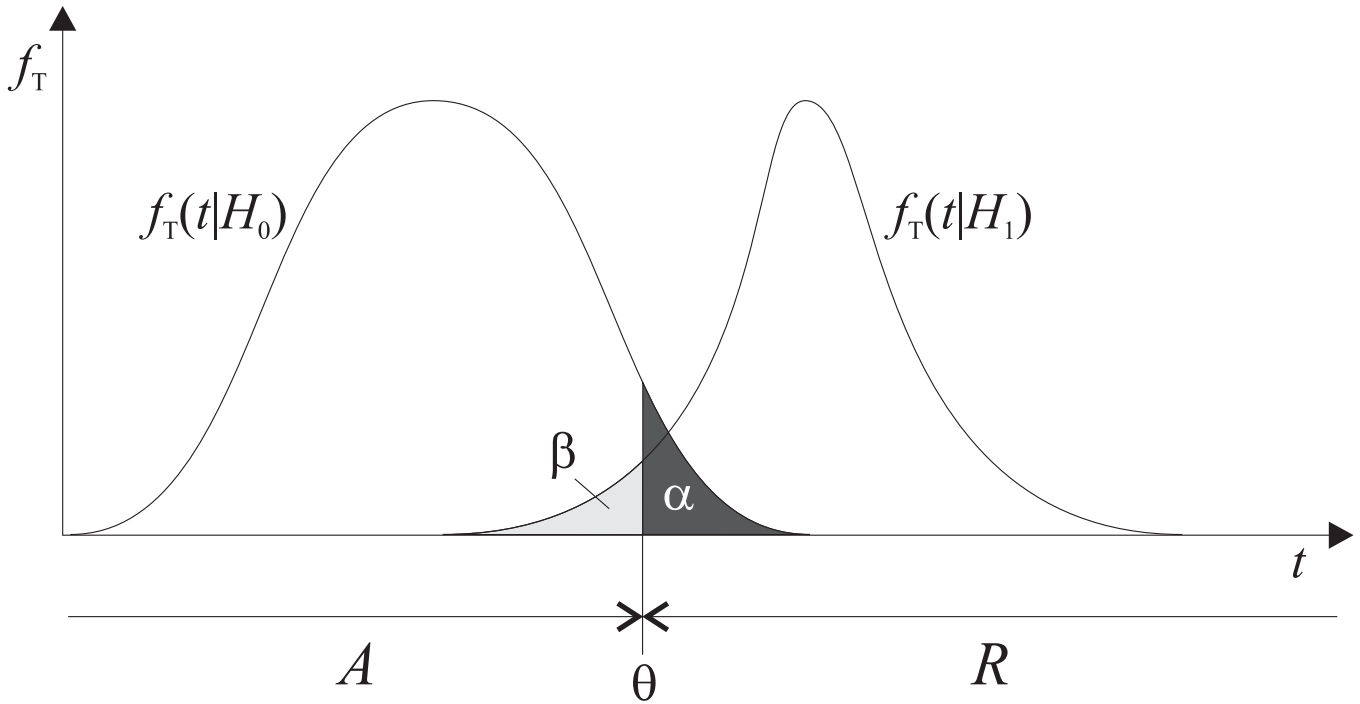
- iris X is not the iris of Adam Czajka

Statistical inference (part II): hypothesis testing

3. Statistical test
   - method that transforms samples into decisions to accept or reject a tested hypothesis (with a fixed probability)
4. Test statistic T
   - statistic (function) used to test H_0
   - we divide the T outputs into two complementary subsets:
     - R: critical (or rejection) region (H_0 is rejected if $T \in R$)
     - A: acceptance region (H_0 is accepted if $T \in A$)

Example of T in biometrics: normalized Hamming distance between two binary iris codes

Statistical inference (part II): hypothesis testing



H_0: sample comes from a distribution "0" $\rightarrow$ T statistic has a distribution $f_T(t|H_0)$

H_1: sample comes from a distribution "1" $\rightarrow$ T statistic has a distribution $f_T(t|H_1)$

Statistical inference (part II): hypothesis testing

When always make mistakes ...

		decision	
		no reason to reject	reject
hypothesis	true	OK $(1-\alpha)$	type I error (α)
	false	type II error (β)	OK $(1-\beta)$

where α: significance level, $1-\beta$: power of the test

Statistical inference (part II): hypothesis testing

Relation to biometric error estimators ...

		decision	
		match	non-match
sample	matches the reference	OK (1-FNMR)	FNM (FNMR)
	does not match the reference	FM (FMR)	OK (1-FMR)

where FNMR: False Non-Match Rate, FMR: False Match Rate

Lecture 9: Statistical evaluation of biometrics

Evaluation of components vs. systems

1. System evaluation employs transactions not attempts
2. Evaluation deals with the entire system, hence it must include:
   - failure to acquire (FTA)
   - system decision policy, in particular the number of allowed attempts in one transaction
3. We have genuine and impostor transactions

System evaluation

False Rejection Rate (FRR)

1. FR (false rejection): an error committed when the genuine transaction is rejected
2. FRR (false rejection rate): estimate of the FR probability

$$\text{FRR} = \frac{\text{number of rejected GENUINE transactions}}{\text{number of all GENUINE transactions}}$$

3. We include FTA when calculating FRR
   (Note: FNMR does not include FTA)
4. FRR is a function of decision policy, for instance when single attempt is allowed in one transaction:

$$\text{FRR} = \text{FTA} + (1\text{-FTA})\ \text{FNMR}$$

System evaluation

False Acceptance Rate (FAR)

1. FA (false acceptance): an error committed when the impostor transaction is accepted
2. FAR (false acceptance rate): estimate of the FA probability

$$\text{FAR} = \frac{\text{number of accepted IMPOSTOR transactions}}{\text{number of all IMPOSTOR transactions}}$$

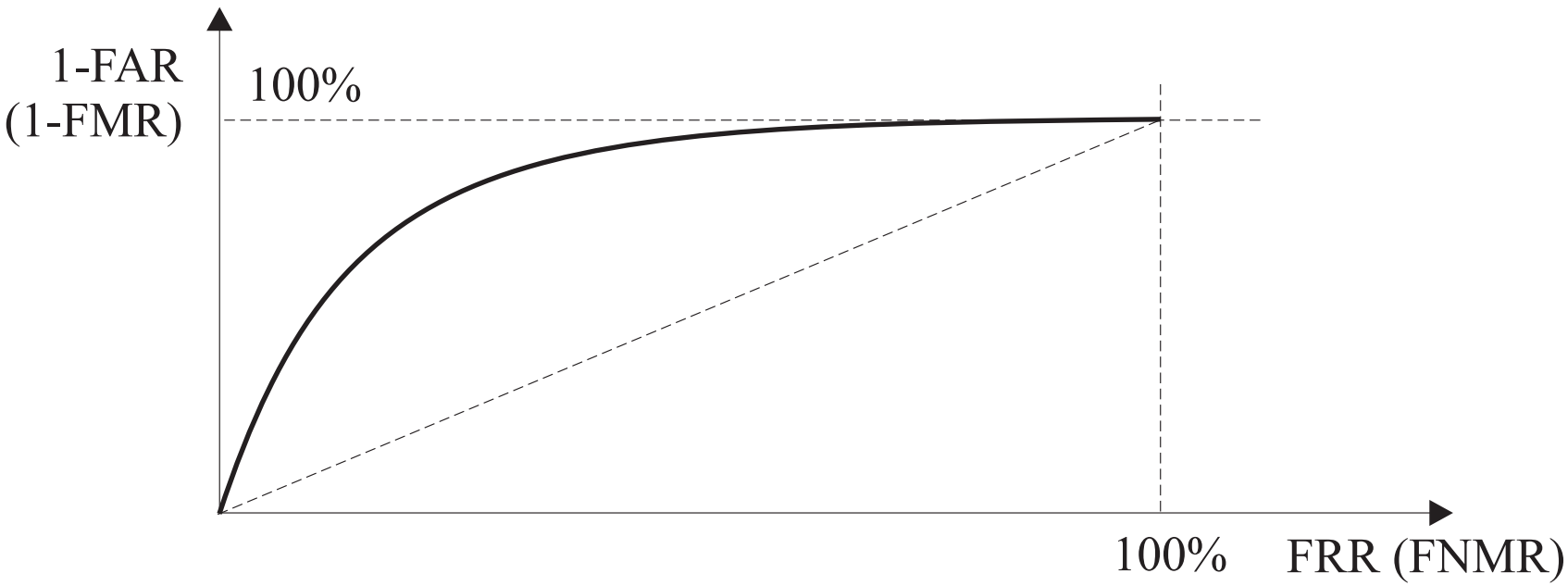
3. We do not include FTA when calculating FAR
4. FAR (like FRR) is a function of decision policy, for instance when single attempt is allowed in one transaction:

$$\text{FAR} = \text{(1-FTA) FMR}$$

Verification system evaluation

Receiver Operating Characteristic (ROC)

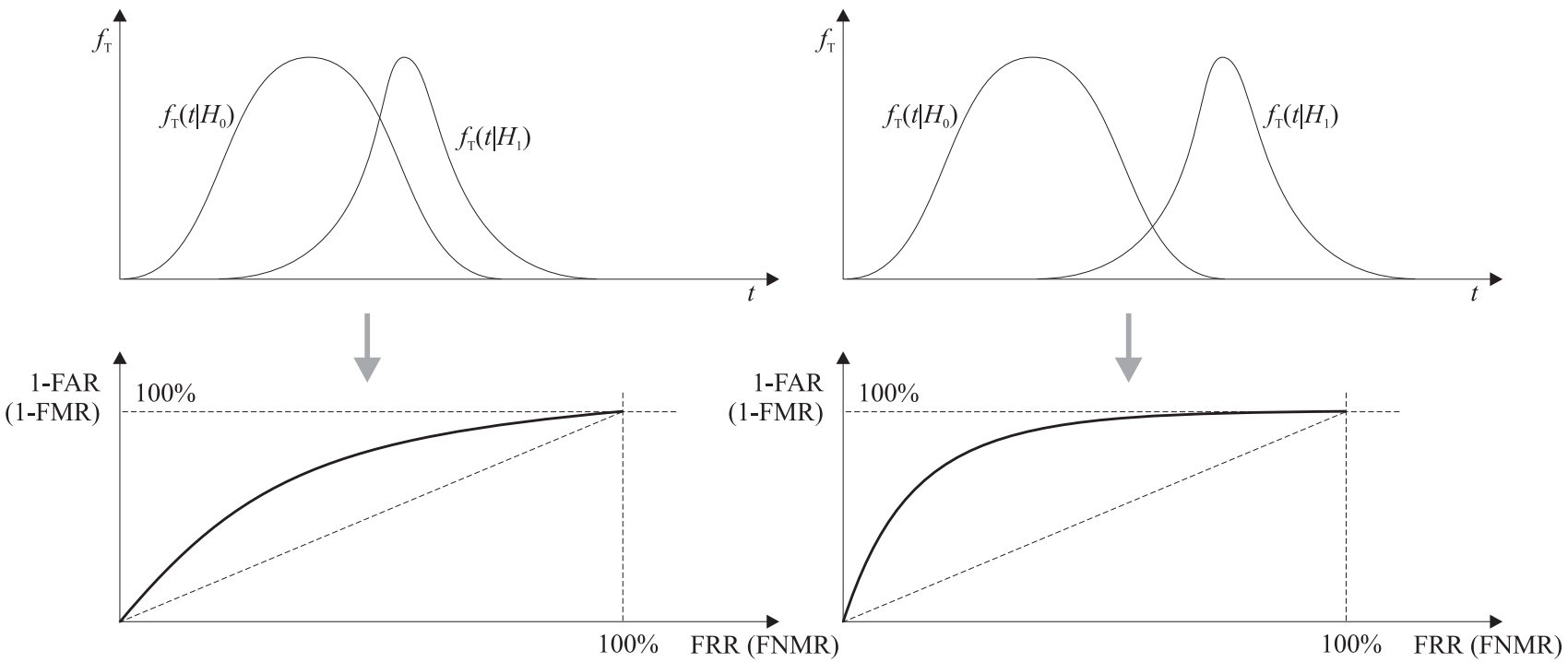
Parametric curve joining FAR and FRR, or FMR and FNMR.
Acceptance threshold is the parameter.



Verification system evaluation

Receiver Operating Characteristic (ROC)

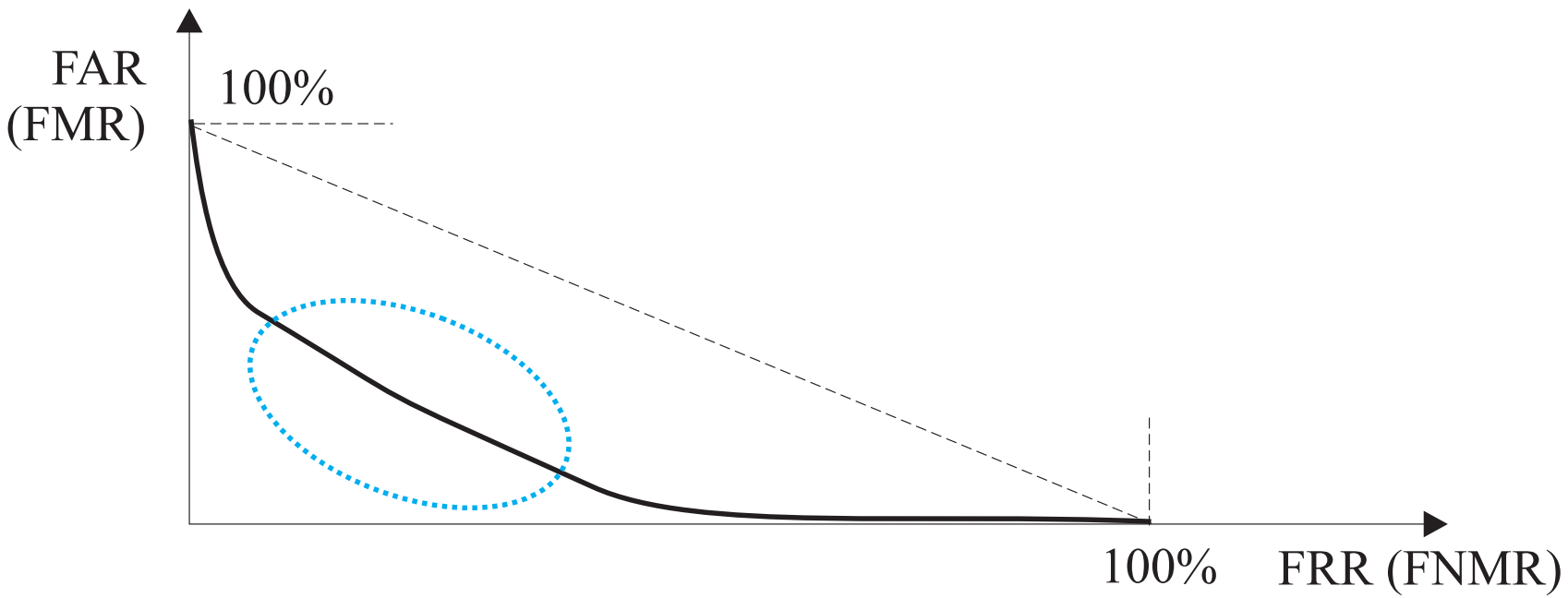
Area under the ROC curve estimates the quality of the classifier, in particular: 1.0 for ideal classification, 0.5 for random classification



Verification system evaluation

Detection Error Tradeoff (DET)

Use of the inverse normal CDF (cumulative probability distribution) for expressing the estimator values (makes the curve in the analysis area to be roughly a straight line)



Verification system evaluation

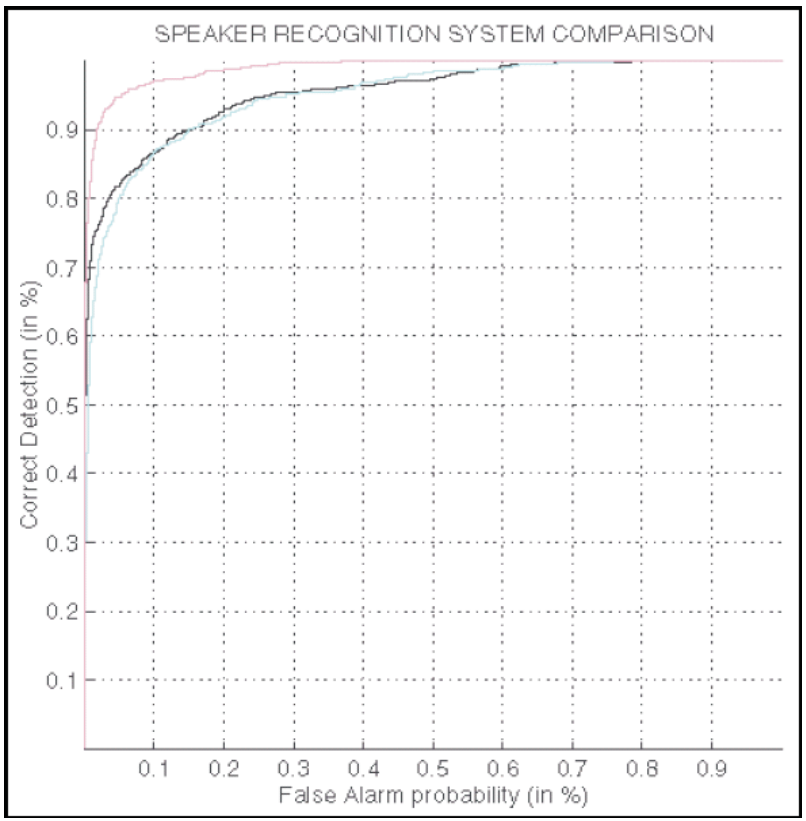
DET: properties that simplify system comparison

1. For normal distributions: straight lines on DET diagram
2. Line slope: quotient of the distribution variances
3. Distance between lines: difference in distribution means
4. Example: for a random classifier applied to normal variables DET is the line FAR=1-FRR (or FMR=1-FNMR)

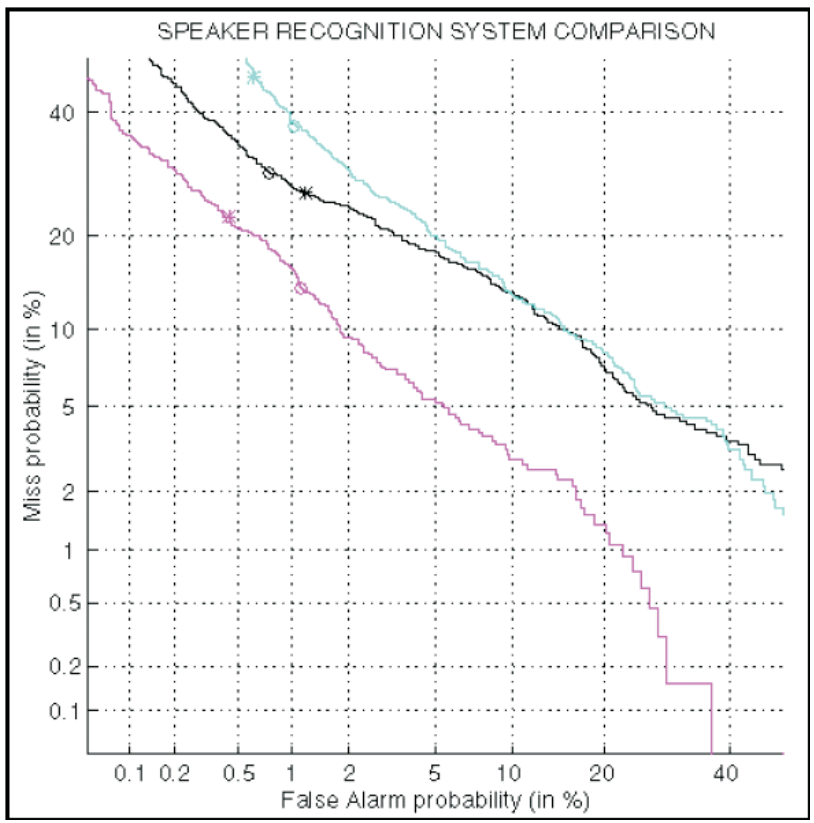
Verification system evaluation

Comparison of DET and ROC in the same evaluation

ROC



DET

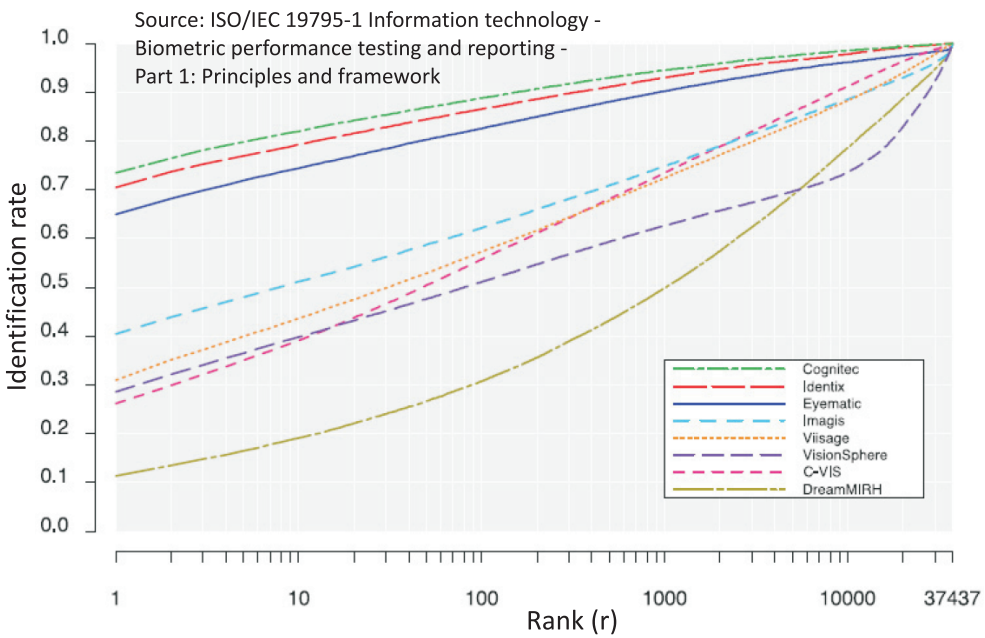


Source: A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, „The DET curve in assessment of detection task performance", EuroSpeech 1997, Proceedings Volume #4, pp. 1895-1898, NIST 1997

Identification system evaluation

Closed-set identification systems: all potential users are enrolled

1. Identification rate at rank r: proportion of identification transactions by a user enrolled in the system, for which user's true identifier is included in the candidate list returned



2. Cumulative Match Characteristic (CMC): identification rate at rank r as a function of r

Identification system evaluation

Open-set identification systems: not all potential users are enrolled

1. False Negative Identification Rate (FNIR): proportion of identification transactions by users enrolled in the system, for which the user's correct identifier is not included in the candidate list returned

   Example: when single attempt is allowed in one transaction

$$\text{FNIR} = \text{FTA} + (1 - \text{FTA})\text{FNMR}$$

Identification system evaluation

Open-set identification systems: not all potential users are enrolled

2. False Positive Identification Rate (FPIR): proportion of identification transactions by users not enrolled in the system, for which a non-empty list of candidate identifiers is returned

   Example: when single attempt is allowed in one transaction and the database size is N:

$$\text{FPIR} = (1 - \text{FTA})(1 - (1 - \text{FMR})^{\text{N}})$$

Lecture 9: Statistical evaluation of biometrics

Statistical inference (part I): interval estimation

1. Interval estimate is defined by two random variables
2. Typically interval estimation is based on calculation of confidence intervals
3. We use biometric data sample to calculate the confidence intervals related to our experiment

Statistical inference (part I): interval estimation

Confidence intervals

1. Extreme values θ_1, θ_2 of the confidence interval $\langle \theta_1, \theta_2 \rangle$ for unknown θ are functions of a random sample, namely

$$\theta_1 = g_1(X_1, \ldots, X_n), \quad \theta_2 = g_2(X_1, \ldots, X_n)$$

   and they are independent of θ
2. In a single experiment, we get a single estimates of random variables θ_1 and θ_2, hence a single confidence interval $\langle \hat{\theta}_1, \hat{\theta}_2 \rangle$
3. The unknown value θ will be included or will not be included in $\langle \hat{\theta}_1, \hat{\theta}_2 \rangle$ (calculation of the probability does not make sense here)

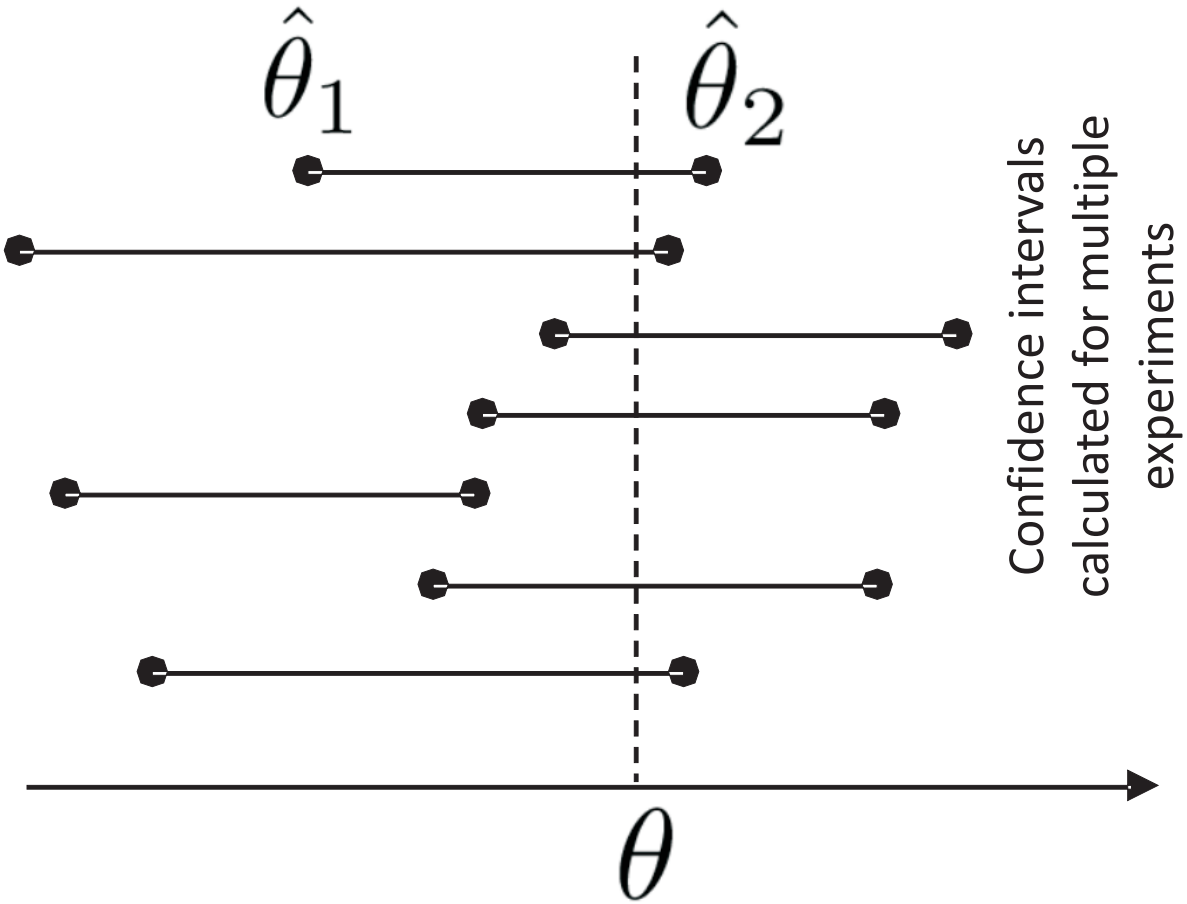
Statistical inference (part I): interval estimation

Confidence intervals

4. Unknown value θ will be frequently included in this interval if our experiment is repeated
   - probability of inclusion of the unknown value θ in the interval $\langle \theta_1, \theta_2 \rangle$ is $1 - \alpha$ (where α significance level)
   - It is NOT true that θ belongs to $\langle \hat{\theta}_1, \hat{\theta}_2 \rangle$ (calculated for a single experiment) with a probability $1 - \alpha$

Statistical inference (part I): interval estimation

Confidence intervals



Interval estimation in biometrics

1. Independent biometric comparisons can be understood as Bernoulli tries with the failure probability p_e
2. Probability of observing up to N_e failures in N_t tries:

$$\mathcal{P}(\xi \leqslant N_e) = \sum_{n=0}^{N_e} \frac{N_t!}{n!(N_t-n)!} p_e^n (1-p_e)^{N_t-n}$$

3. The upper limit p_e^+ of the confidence interval and the significance level α are interrelated:

$$\sum_{n=0}^{N_e} \frac{N_t!}{n!(N_t-n)!} (p_e^+)^n (1-p_e^+)^{N_t-n} = \alpha$$

Interval estimation in biometrics

We obtained no errors in our experiment (e.g., FMR=0).

Question: What is the minimum error rate
that can be guaranteed for a given α?

Interval estimation in biometrics

For $N_e = 0$ (no errors) we get:

$$(1 - p_e^+)^{N_t} = \alpha \qquad \text{hence} \qquad N_t \ln(1 - p_e^+) = \ln(\alpha)$$

For $\alpha = 0.05$, assuming that

$$\ln(0.05) \approx -3$$

and for small p_e^+

$$\ln(1 - p_e^+) \approx -p_e^+$$

we get

$$p_e^+ = \frac{3}{N_t} \qquad \text{("rule of three")}$$

Interval estimation in biometrics

Example: providing statistical guarantees for FMR=0

- We have samples for $N = 12$ distinct persons

Interval estimation in biometrics

Example: providing statistical guarantees for FMR=0

- We have samples for $N = 12$ distinct persons
- We can generate $N_t = N/2 = 6$ statistically independent tries (each impostor comparison uses a distinct pair of subjects)

Interval estimation in biometrics

Example: providing statistical guarantees for FMR=0

- We have samples for $N = 12$ distinct persons
- We can generate $N_t = N/2 = 6$ statistically independent tries (each impostor comparison uses a distinct pair of subjects)
- We obtained FMR=0 (we are proud of our algorithm). Hence for $N_e = 0$ and $\alpha = 0.05$ we get:

Interval estimation in biometrics

Example: providing statistical guarantees for FMR=0

- We have samples for $N = 12$ distinct persons
- We can generate $N_t = N/2 = 6$ statistically independent tries (each impostor comparison uses a distinct pair of subjects)
- We obtained FMR=0 (we are proud of our algorithm). Hence for $N_e = 0$ and $\alpha = 0.05$ we get:

$$p_e^+ = \frac{3}{N_t} = \frac{3}{6} = 50\%$$

- Interpretation: if we repeat our evaluation many times (independently) we expect that in 95% of experiments our true FMR may be as large as 50% (!)

Interval estimation in biometrics

4. For statistically dependent comparisons or for variable p_e in test population we cannot say that we have Bernoulli tries
5. When the comparisons are statistically dependent?
    - data of the same subjects are used multiple times when estimating FA error (cross-comparisons) or FR error (multiple tries of the same person)

Lecture 9: Statistical evaluation of biometrics

Correct use of biometric databases

1. Always divide your database into disjoint subsets
   - estimation (or training) subset: used to train your method and setting all the parameters
   - testing subset: unknown when training and used to make the evaluation (and presenting your results)
2. There are many ways of dividing the biometric database
   - try to generate statistically independent comparisons
     $\rightarrow$ needs "big" datasets
   - when the database is small it is better to analyze more dependent data than too little independent
   - use appropriate statistical tools for small databases, e.g., k-fold cross-validation

Selection of the acceptance threshold

1. EER is not the most important error estimate
   - helpful in technology evaluation and method comparisons
   - in operational scenarios we are typically interested in FRR at a given FAR (or vice versa)
2. Requirements for FAR/FRR vary depending on the application
   - secure systems:
     low values of FAR at the cost of increasing FRR
   - comfortable systems:
     low values of FRR at the cost of increasing FAR

Selection of the acceptance threshold

3. In evaluation of PAD (Presentation Attack Detection) we typically expect low (or zero) NPCER
   - additional rejections introduced by PAD are not welcome
   - probability of presentation attack (skilled forgery) is relatively lower than zero-effort impostor presentation (random forgery)
   - estimating the statistical distributions related to presentation attacks is difficult or impossible (we cannot guess all possible attacks)