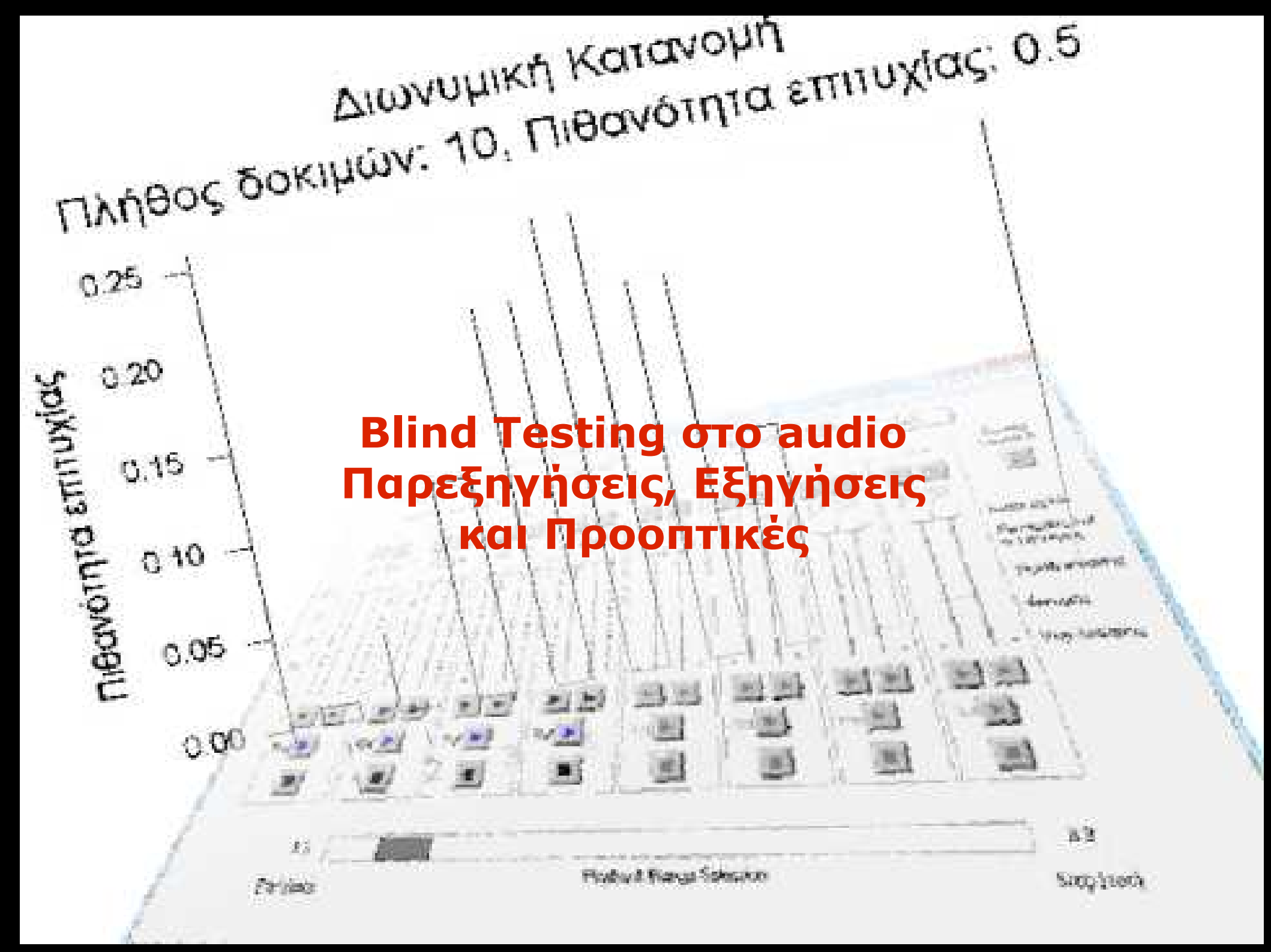


Blind Testing στο audio
Παρεξηγήσεις, Εξηγήσεις
και Προοπτικές

Σήμερα:

Κριτική ακρόαση
Ορισμός
Η ιστορία των Κριτικών ακροάσεων
Το τελικό κριτήριο είναι το αυτί...
Κρίνοντες, κρινόμενοι και η άγνωστη έννοια της προβολής
Γιατί μας είναι χρήσιμη μια τυφλή δοκιμή, τελικώς;

Τυφλές δοκιμές
Ορισμός
Μεθοδολογίες
Μερικές βασικές απαιτήσεις
Πως αξιολογείται μια τυφλή δοκιμή

Προοπτικές
Εργαλεία για την πραγματοποίηση τυφλών δοκιμών

Κριτική ακρόαση:

Εξοπλισμός
Πρόγραμμα
Ακροατές
Διαδικασίες

Στόχος: Αξιολόγηση τεχνολογιών-συσκευών-συστημάτων

Βασικές απαιτήσεις:
Ελαχιστοποίηση πόλωσης
Μεγιστοποίηση Αξιοπιστίας
Επαναληψιμότητα

Η Κριτική ακρόαση έχει ιστορία!

1931: Snow (εύρος συχνοτήτων)
1941: Chinn/Eisenberg (εύρος συχνοτήτων), Olson (περί bias)
1947: Le Bel (πρώτες προδιαγραφές)

1970: Η υψηλή πιστότητα αλλάζει την κριτική ακρόαση
Southeastern Michigan Woofer and Tweeter Marching Society (SMWTMS) - Η πρώτη δοκιμή A/B/X
Daniel Shanefield/Bell-Boston Acoustics Society – Το πρόβλημα των "όμοιων" ενισχυτών

1981: Lipshitz/Vandenrcooy (JAES) Great Debate – BoP
1982: Floyd Toole – Turning Opinion into Fact (JAES)

Από τότε η κουβέντα συνεχίζεται...

Το αυτί είναι το τελειότερο όργανο!

Σωστά...
Αλλά υπάρχει ένα πρόβλημα...

Δεν ακούμε με τα αυτιά...!

Όταν ακούμε...

Ακούμε -κυρίως- με τον εγκέφαλο
Το τελικό αίσθημα είναι συνδυασμός ερεθίσματος-προσδοκίας-εμπειρίας
Οι ακουστικές ψευδαισθήσεις είναι ισχυρότερες των οπτικών
Diana Deutsch – Scale illussion
Jean-Claude Risset - Shepherds Ascending Tones

1994: Toole/Olive: παράγοντες πόλωσης
1994: Thomas Nousaine: Null Testing

Στην πραγματικότητα, φαίνεται ότι:

Με ελάχιστες εξαιρέσεις, η “ελεύθερη” ακρόαση οδηγεί σε σφάλματα
Η κριτική ακρόαση πρέπει να είναι αυστηρά ελεγχόμενη
Ο μόνος τρόπος για να ελεγχθεί μια ακρόαση είναι να είναι μετρήσιμη
Ο μόνος τρόπος για να είναι μια ακρόαση μετρήσιμη είναι να είναι τυφλή!

Ναι, αλλά...
... η ελεγχόμενη ακρόαση αγχώνει...

Αυτό είναι μια κοινοτοπία!

1991: Robert Harley (Stereophile) - AES
Δεν βασίζεται σε κάποια έρευνα
Αντιθέτως έχει αποδειχθεί, μάλλον, το αντίθετο

1991: David Clark – AES
Πείραμα Clark/Greenhill (περιοδικό Audio, 1984)
TAS Vs SMWTMS
2.5% thd Vs Flat
Ελεύθεροι Vs Τυφλοί (A/B/X)
Τελικό αποτέλεσμα: 1/4 (μόνο ο συνδυασμός SMWTMS-A/B/X)

Κρίνοντες, Κρινόμενοι, Προβολή

Ακαδημία-Ερευνητές:
Πειράματα ψυχοακουστικής - Τεχνολογίες
Βιομηχανία-Σχεδιαστές:
Τεχνολογίες - Προϊόντα - Marketing
Ειδικός Τύπος-Κριτικοί:
Τεχνολογίες – Προϊόντα
Αγορά – Product Managers – Audiophiles:
Προϊόντα – Τεχνολογίες

Κριτική Ακρόαση = Πείραμα
Πείραμα = Αποτέλεσμα
Αποτέλεσμα --> Ανάγκη προβολής στον γενικό πληθυσμό

Δυνατότητα προβολής <=> Μέτρο αξιοπιστίας πειράματος

Στο audio: Δυνατότητα προβολής <=> Blind Testing

Οι τυφλές δοκιμές είναι απαραίτητες επειδή:

Το αποτέλεσμά τους μπορεί να προβληθεί στο γενικό πληθυσμό

Αλλά, γενικώς, μας είναι χρήσιμες επειδή:

Μας απαλλάσσουν από πολλούς παράγοντες πόλωσης
Διαθέτουν μηχανισμούς αυτοπροστασίας
Παράγουν αξιοποιήσιμα/μετρήσιμα/επαναλαμβανόμενα αποτελέσματα
Μπορούν να ελεγχθούν (μέσω της επαναληψιμότητας)
Βασίζονται σε καθορισμένες διαδικασίες (AES, IEEE, ITU)
Δοθέντος ενός αποτελέσματος το BoP περνάει στην άλλη πλευρά!

Τυφλές Δοκιμές – Το πλαίσιο

Τυφλή:
Κάθε δοκιμή κατά την οποία:
Δεν είναι γνωστό το/τα αντικείμενο/αντικείμενα
Δεν είναι δυνατή η αναγνώριση των DUTs παρά μόνο μέσω ακρόασης
Έχουν εξασφαλιστεί όροι ομοιότητας μεταξύ των DUTs
Δίνεται η δυνατότητα σύγκρισης τουλάχιστον μεταξύ δύο εκδοχών

Διπλά Τυφλή (Double Blind):
Όταν, επιπροσθέτως, οι δύο πρώτοι όροι εξασφαλίζονται και για τον συντονιστή της δοκιμής.

Είδη:
Σύγκριση ζευγών (Pair Comparison, A/B)
Πολλαπλή σύγκριση με κρυμμένη αναφορά (ABC HR)
Δοκιμή MUSHRA (MUltiple Stimuli, Hidden Reference, Anchors)
Δοκιμή A/B/X

Βασικές τεχνικές απαιτήσεις:

Σταθερότητα στάθμης (ενίοτε, loudness)
Σταθερότητα φάσης
Ισοστάθμιση γραμμικών παραμορφώσεων (συζητείται!)
Συνθήκες ελαχιστοποίησης μη γραμμικότητας (συζητείται!)
Συγκεκριμένη στάθμη αναπαραγωγής (80-90dBSPL)
Συγκεκριμένες προδιαγραφές ακουστικής του χώρου
Ελεγχόμενοι χρόνοι ακουστικού ερεθίσματος (15-20 sec)
Ελεγχόμενοι χρόνοι δοκιμής (15-20 min)
Screening ακροατών (ακοόγραμμα)
Ελεγχόμενο μείγμα έμπειρων/μη έμπειρων ακροατών
Έλεγχος της παραμέτρου "εκπαιδευμένος ακροατής"
Έλεγχος της παραμέτρου "εξειδικευμένος ακροατής"

Οι βασικές τεχνικές απαιτήσεις καθορίζουν τις συνθήκες της δοκιμής και αποτελούν προσάρτημα της δοκιμής και των αποτελεσμάτων της!

Τι ψάχνουμε σε μια Τυφλή δοκιμή;

Κύριος ΑνΣκ:
Να απαντηθεί το ερώτημα:
"Είναι ακουστές διαφορές μεταξύ των εκδοχών της δοκιμής;"

Δευτερεύων ΑνΣκ:
Με την προϋπόθεση της καταφατικής απάντησης στον κύριο ΑνΣκ, να καταγραφεί ή άποψη των μετεχόντων για το κάθε ερέθισμα.

Αυτονόητο:
Δεν είναι αποδεκτή η διατύπωση άποψης/κριτικής για τις εκδοχές/ερεθίσματα, αν προηγουμένως δεν έχει αποδειχθεί η δυνατότητα αναγνώρισης διαφορών.

Και, εδώ, το αποδειχθεί, σημαίνει *αποδειχθεί!*

Πώς αποδεικνύεται ότι υπάρχουν διαφορές;

Προβολή <=> Στατιστική Ανάλυση

Τα μαθηματικά της Τυφλής Δοκιμής:

Δοκιμές Beroulli, Κορώνες, Γράμματα και Ηχεία...
Πιθανότητες αποτελέσματος και Διωνυμικές Κατανομές
Ο στατιστικός έλεγχος μιας υπόθεσης
Σφάλματα του ελέγχου
Στην πράξη: Πότε είμαστε ικανοποιημένοι

Τι είναι μια δοκιμή Bernoulli;

Ρίψη "τίμιου" νομίσματος:
p=0.5=> 50/100 γράμματα (τελικώς...)
Ρίψη "ανισοβαρούς" νομίσματος:
p>0.5 => >50/100 γράμματα (τελικώς)

Στον χώρο του audio:
Ακρόαση χωρίς δυνατότητα αναγνώρισης: p=0.5
Ακρόαση με δυνατότητα αναγνώρισης: p>0.5

Σε μια τυφλή δοκιμή, προσπαθούμε πάντα να αποδείξουμε ότι:
P=0.5 => Ο ακροατής δεν μπορεί να ακούσει διαφορές.

Υπάρχει όμως ένα πρόβλημα: Το *"τελικώς"*!

Πώς εξασφαλίζουμε ότι η τύχη δεν επηρεάζει *"τελικώς"*;

Μπορούμε να υπολογίσουμε...

Την πιθανότητα εμφάνισης ενός αποτελέσματος:

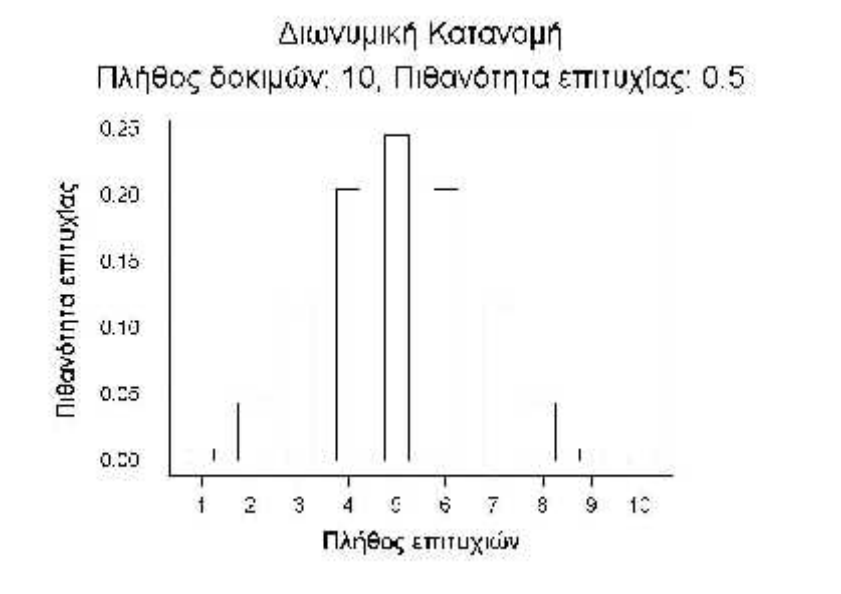
$$f(k;n,p) = \Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

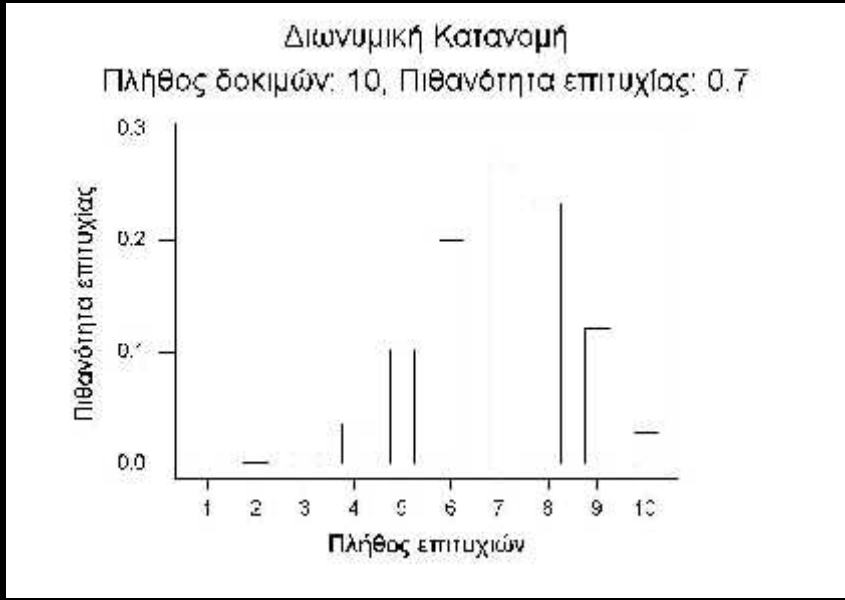
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

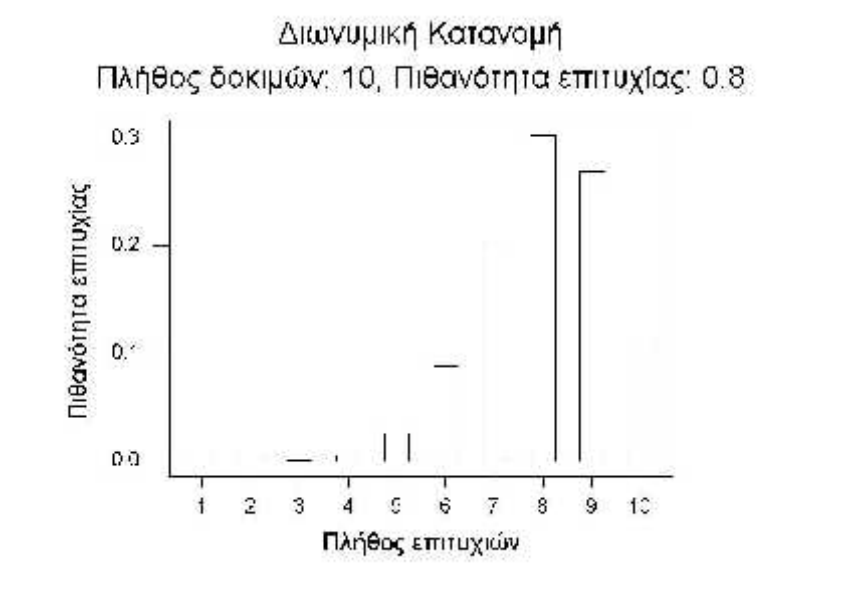
k: αριθμός επιτυχών αναγνωρίσεων
n: πλήθος δοκιμών (ακροατών ή ακροατών x tracks)
p: Η πιθανότητα του πειράματος Bernoulli

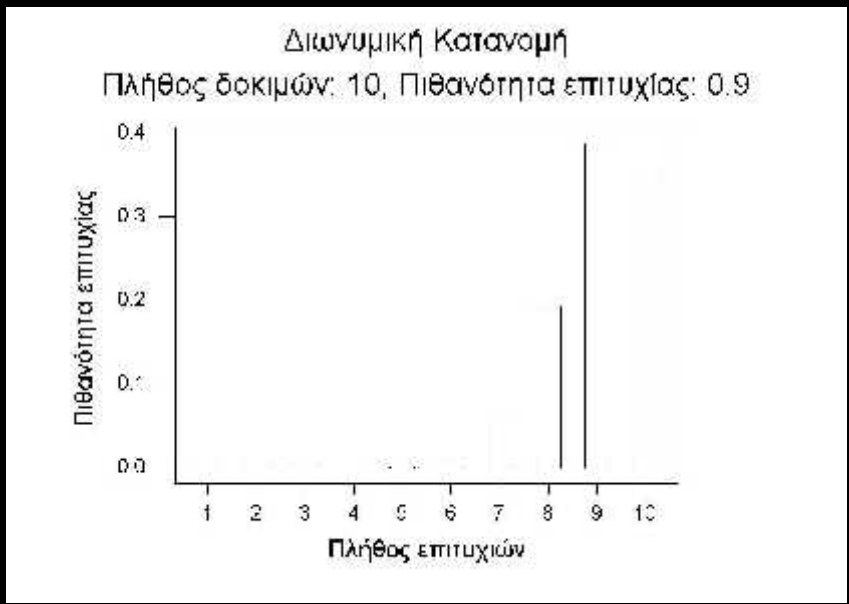
Οι πιθανότητες για ένα τίμιο νόμισμα να έλθει 10/10 φορές γράμματα υπάρχει αλλά είναι 0.001. Πρέπει να κάνουμε το πείραμα 1000 φορές!

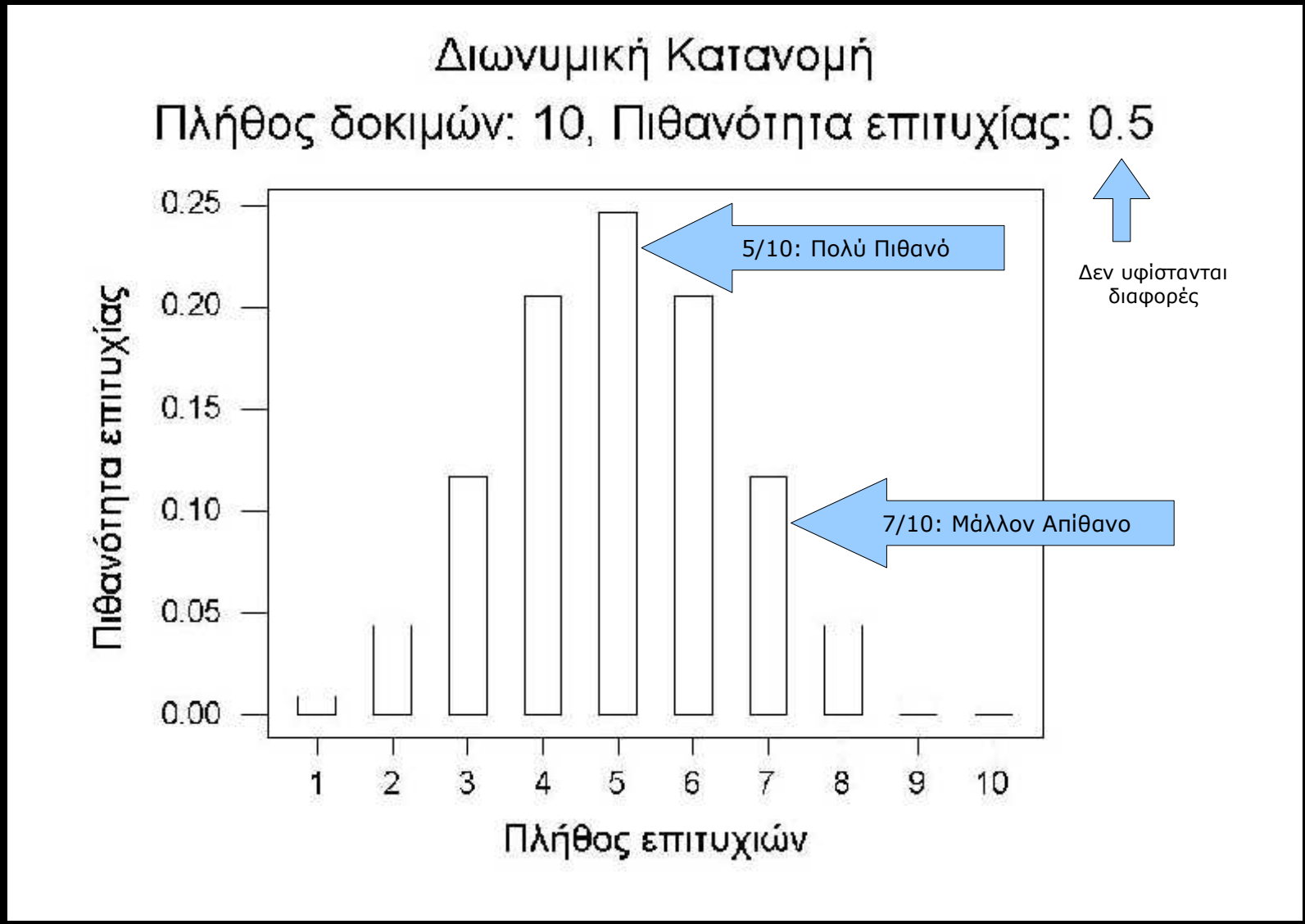
Διωνυμικές κατανομές:

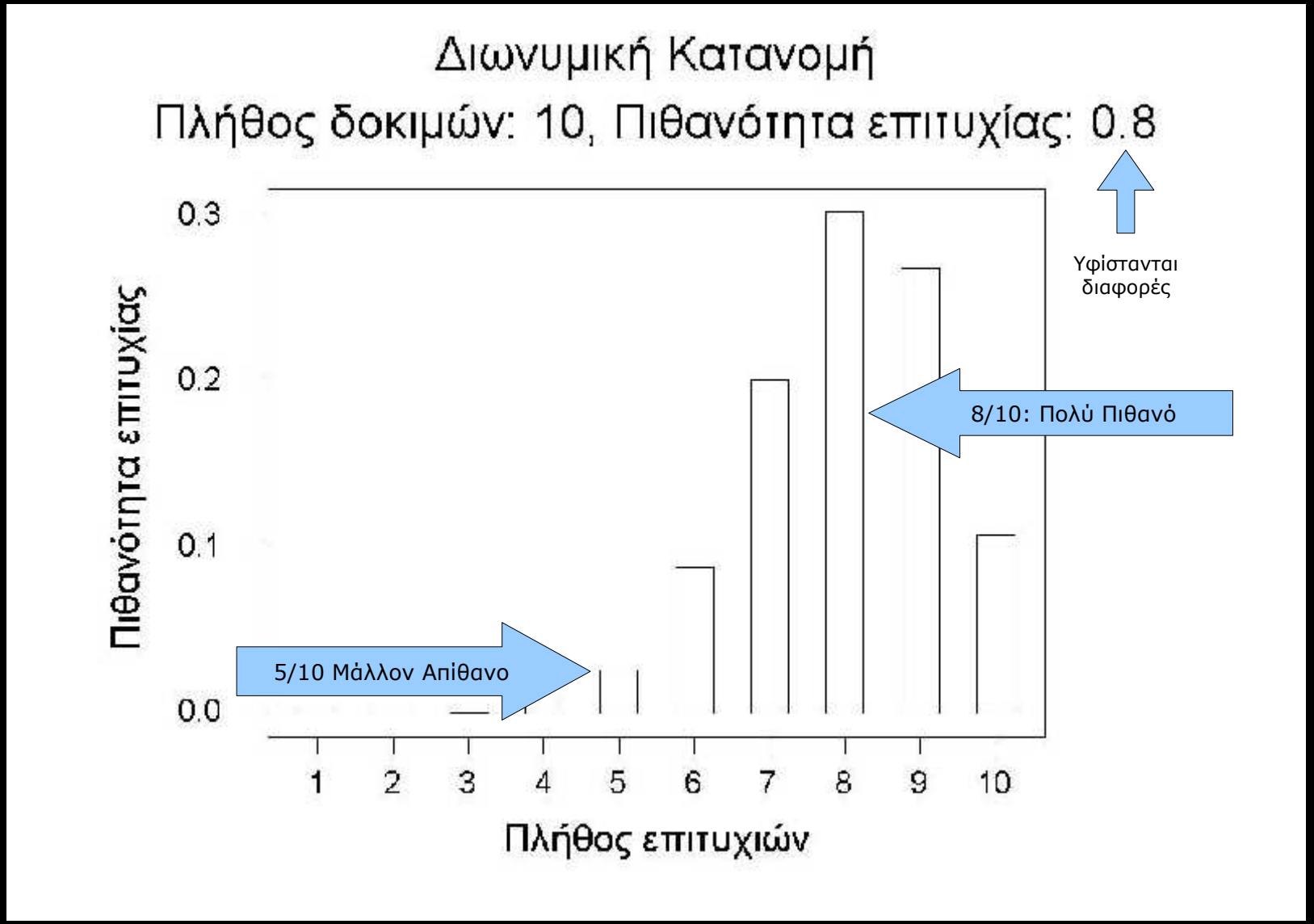












Ο έλεγχος της υπόθεσης στην Τυφλή δοκιμή:

Η Τυφλή δοκιμή μεταξύ δύο εκδοχών διατυπώνεται ως εξής:

Ho: *"Δεν υπάρχει διαφορά μεταξύ των δύο εκδοχών"*
H1: *"Υπάρχει διαφορά μεταξύ των δύο εκδοχών"*

Το αποτέλεσμα της δοκιμής είναι ένας αριθμός επιτυχών αναγνωρίσεων στο σύνολο των προσπαθειών (το k στην εξίσωση)

Μπορούμε να υπολογίσουμε την πιθανότητα να εμφανιστεί το συγκεκριμένο k, με δεδομένο ότι p=0.5 (υπόθεση Ho)

Αν η πιθανότητα αυτή είναι πολύ μικρή, είναι δηλαδή πολύ απίθανο να εμφανιστεί ενώ οι ακροατές επιλέγουν στην τύχη, τότε η Ho απορρίπτεται, υπέρ της H1.

Πόσο απίθανο δηλαδή;

Το ποσοστό απιθανότητας είναι κάτι που επιλέγεται αυθαίρετα

Όμως:
Στη διεθνή πρακτική επιλέγουμε την τιμή 0.05 (5%)

Και σε ευαίσθητες δοκιμές (φαρμάκων) την τιμή 0.01 (1%)

Στην στατιστική αυτό ονομάζεται "επίπεδο σημαντικότητας" (Significance Level) της δοκιμής.

Το πρόβλημα σε όλα αυτά, είναι ότι

Υπάρχει πάντα η πιθανότητα να κάνουμε λάθος!

Τα σφάλματα στον έλεγχο της υπόθεσης:

Σφάλμα τύπου I:
Η Ηο απορρίπτεται υπέρ της Η1 ενώ η Η1 δεν ισχύει
Αυτό σημαίνει ότι:
Η δοκιμή αποφαίνεται πως υπάρχουν διαφορές ενώ δεν υπάρχουν.
Η πιθανότητα του σφάλματος τύπου I είναι ίση με το Ε.Σ.

Σφάλμα τύπου II:
Η Ηο δεν απορρίπτεται υπέρ της Η1 ενώ η Ηο δεν ισχύει
Αυτό σημαίνει ότι:
Η δοκιμή αποφαίνεται πως δεν υπάρχουν διαφορές ενώ υπάρχουν.
Η πιθανότητα του σφάλματος τύπου II εξαρτάται από το n και σε "μικρά" πειράματα υπάρχει ο κίνδυνος να είναι σημαντική.

Αποτέλεσμα Πειράματος με περιθώριο σφάλματος α	Πραγματικότητα	
	Ισχύει η υπόθεση H_0	Ισχύει η υπόθεση H_1
Γίνεται δεκτή η υπόθεση H_0	✓	Σφάλμα Τύπου II (β)
Γίνεται δεκτή η υπόθεση H_1	Σφάλμα Τύπου I (α)	✓

Στην πράξη:

Πάνελ 10 επιλογών (n=10)
Επίπεδο σημαντικότητας: 0.05
Απαραίτητες επιτυχίες για απόρριψη της Ho: k>=8/10
Σφάλμα Τύπου I: <0.0547 (k=8)
Σφάλμα Τύπου II: <0.3222 (k=8)

Πάνελ 25 επιλογών (n=25)
Επίπεδο σημαντικότητας: 0.05
Απαραίτητες επιτυχίες για απόρριψη της Ho: k>=17/25
Σφάλμα Τύπου I: <0.0539 (k=17)
Σφάλμα Τύπου II: <0.4882 (0.3231 για k=17)

Πάνελ 50 επιλογών (n=50)
Επίπεδο σημαντικότητας: 0.05
Απαραίτητες επιτυχίες για απόρριψη της Ho: k>31/50
Σφάλμα Τύπου I: <0.0594 (k=31)
Σφάλμα Τύπου II: <0.5535 (k=31)

Επομένως:

Οι μικρές σε "n" τυφλές δοκιμές έχουν:
Δυνατότητα για μικρά σφάλματα Τύπου I
Γενικώς πολύ μεγαλύτερα σφάλματα Τύπου II

Για το λόγο αυτό:
Είναι περισσότερο αξιόπιστες όταν απορρίπτουν την Ho
Είναι περισσότερο πολωμένες όταν απορρίπτουν την H1

Τι μπορούμε να κάνουμε;
Αύξηση του "n" (γινόμενο ακροατών x tracks)
Προσπάθεια εξομοίωσης των δύο Τύπων σφάλματος (Leventhal)
Περισσότερες δοκιμές με μικρά "n"

Στην πράξη (μέρος β'):

Αριθμός Ακροατών: 3
Αριθμός Tracks ανά ακροατή: 10
Δεχόμαστε: N=30

Με βάση τους Πίνακες Leventhal:

Για περιθώριο σφάλματος (επίπεδο σημαντικότητας): 0.05:
Απαιτούνται το λιγότερο 20 ορθές επιλογές στις 30 (δηλαδή επιτυχία 0.7)

Αυτό σημαίνει: α=0.0494, Pwr=0.7304, β=0.2696
Ωστόσο η ελάχιστη επίδοση, οδηγεί σε μεγάλο β!

Για πιο ισορροπημένα αποτελέσματα:

Επιλέγουμε (από τον πίνακα), p=0.8 δηλαδή: 24/30
Αυτό σημαίνει: Pwr=0.9744, β=0.0256

Με άλλα λόγια (αν όλα πήγαν καλά):

Τρεις ακροατές, άκουσαν δύο εκδοχές με 10 tracks ο καθένας με τη μέθοδο A/B/X, δηλαδή συνολικά έκαναν 30 επιλογές.

Από αυτές, οι 24 ήταν επιτυχείς (βρήκαν το X)

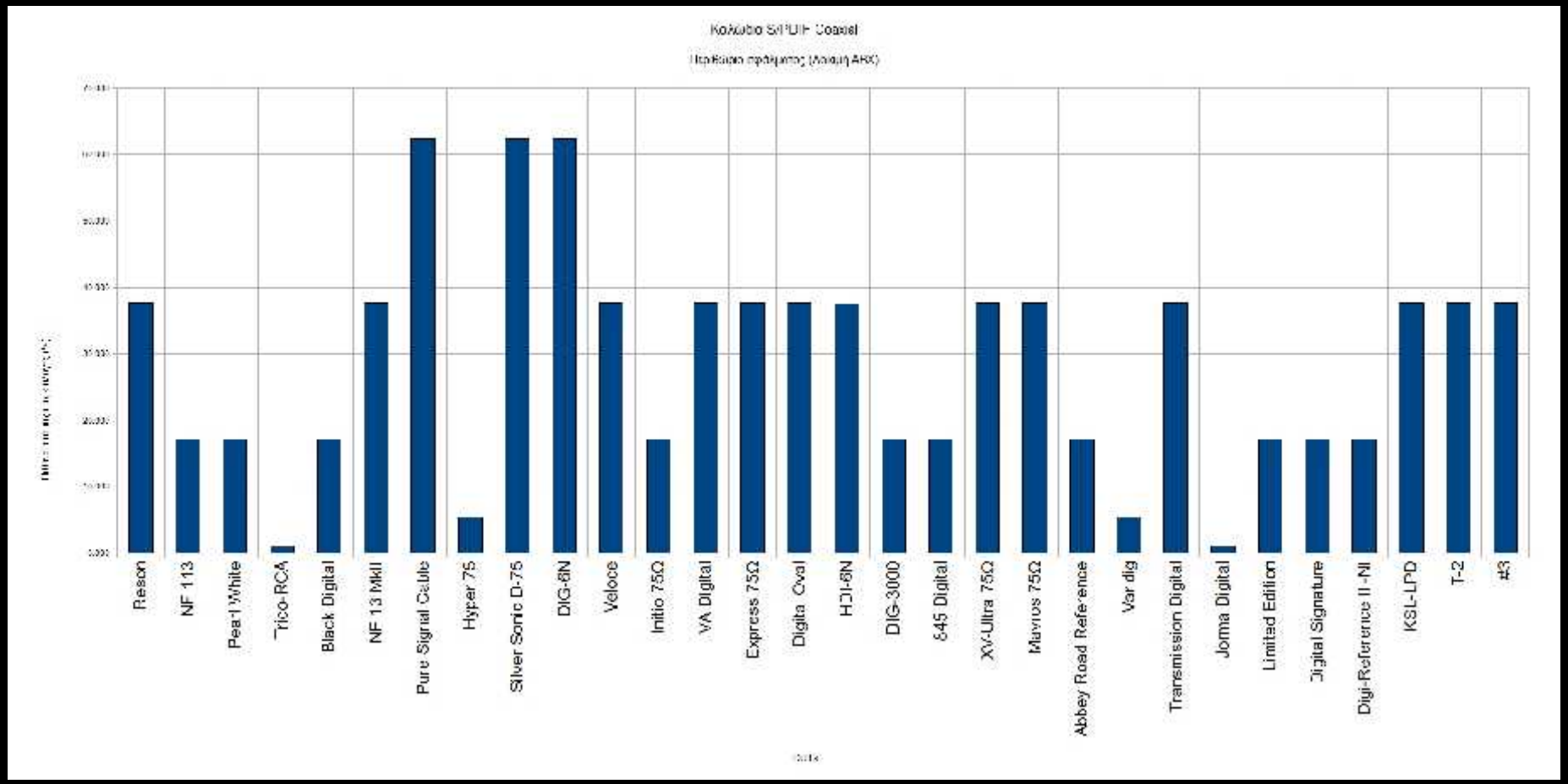
Το αποτέλεσμα του πειράματος είναι:

Υπάρχουν αρκετές αποδείξεις για να απορριφθεί η Ho (τα A και B δεν έχουν ακουστικές διαφορές) υπέρ της H1 (έχουν ακουστικές διαφορές), με περιθώριο σφάλματος 5%.
Τα δεδομένα του πειράματος δείχνουν ότι η πιθανότητα σωστής επιλογής ήταν 0.8 και αυτό σημαίνει ότι:

Η ισχύς του πειράματος (H1 έναντι Ho) είναι 97.44%
Η πιθανότητα σφάλματος τύπου β (Ho έναντι H1) είναι 2.56%

ΔΟΚΙΜΗ: ΚΑΛΩΔΙΑ S/PDIF (avmentor.gr/reviews/group_test_spdif_coaxial_cables_2011_00.htm)

Πείραμα: 1 Ακροατής, 10 Tracks (N=10)
για α=0.0547, β=0.3222 p=0.8 και για β=0.0702: p=0.9)
Ho: Το καλώδιο δεν έχει διαφορές σε σχέση με ένα RG59A/U
H1: Το καλώδιο έχει διαφορές σε σχέση με ένα RG59A/U

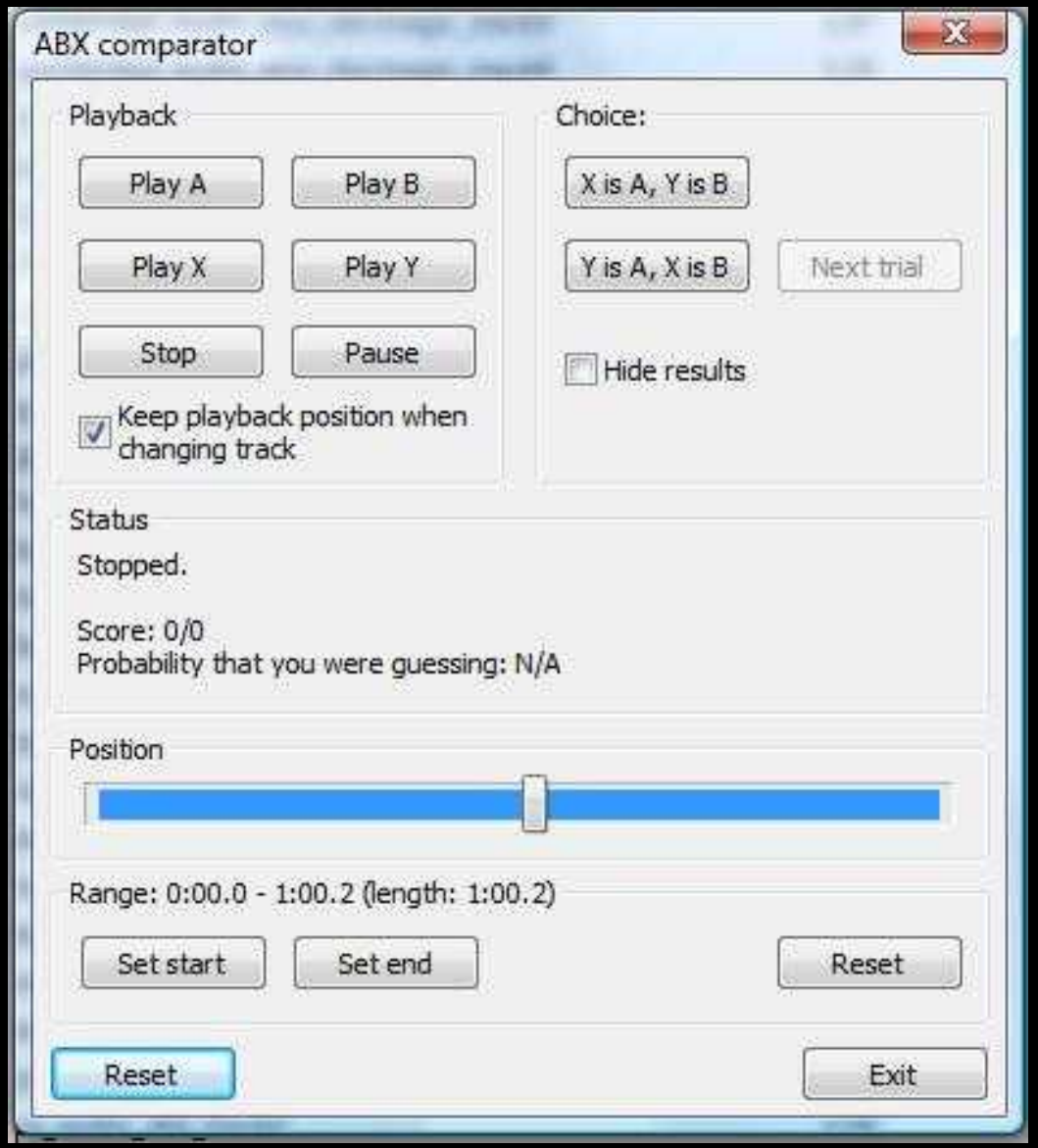


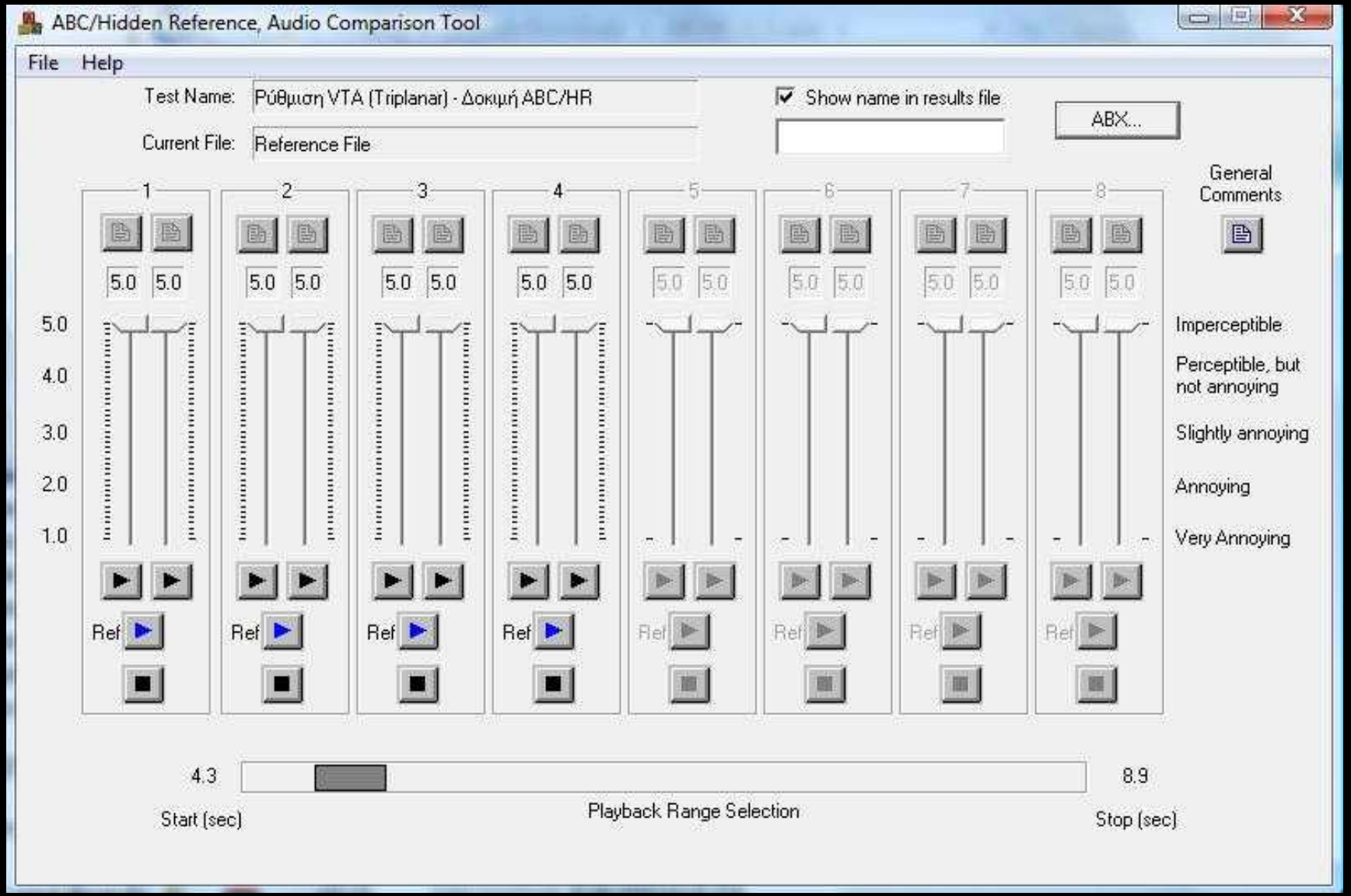
Εργαλεία για την πραγματοποίηση Τυφλών Δοκιμών

Foobar2000 + ABX Plug-in

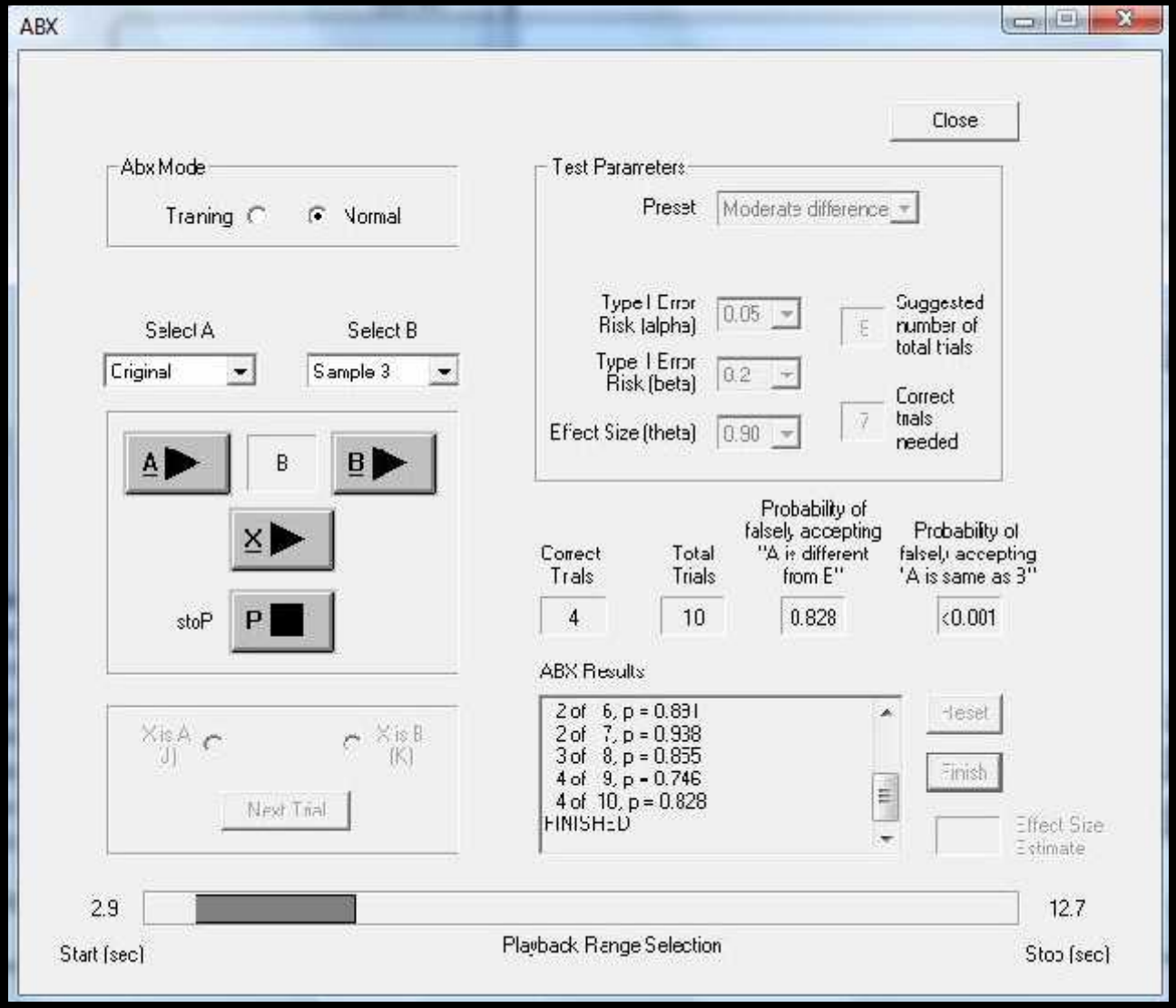
Λογισμικό ABC HR

Blind Testing στο audio





Blind Testing στο audio



Για περισσότερο διάβασμα:

01. "Ten years of ABX Testing", David Clark, AES 91st Convention, 1991
02. "Listening Tests-Turning Opinion into Fact", Floyd Toole, JAES Vol.30, No.6, 1982
03. "The Great Debate - Subjective Evaluation", Stanley Lipshitz, John Vanderkooy, JAES Vol.29, No. 7/8, 1981
04. "This Is Your Brain On Music", Daniel Levitin, Penguin Books, 2006<
05. "Auditory Illusions",Special Issue, JAES Vol.31, No.9
06. "Hearing is Believing vs Believing is Hearing - Blind vs. Sighted Listening Tests and Other Interesting Things", Floyd Toole, Sean Olive, AES 97th Convention, 1994
07. "Can You Trust Your Ears?", Thomas Nousaine, AES 91st Convention, 1991
08. "The Role of Critical Listening in Evaluating Audio Equipment Quality", Robert Harley, AES 91st Convention, 1991
09. "Type 1 and Type 2 Errors in the Statistical Analysis of Listening Tests", Les Leventhal, JAES Vol.34, No.6, 1986
10. "Analyzing Listening Tests with the Directional Two-Tailed Test", Les Leventhal, JAES Vol.44, No.10, 1996
12. "Approximation Formulas for Error Risk and Sample Size in ABX Testing", Herman Burstein, JAES Vol.36, No.11, 198
13. "Perceptual Audio Evaluation: Theory, Method and Application", Soren Bech/Nick Zacharov, Wiley, 2006
14. "Η Διωνυμική κατανομή", Wikipedia, http://en.wikipedia.org/wiki/Binomial_distribution
15. "Diana Deutsch", Wikipedia, http://en.wikipedia.org/wiki/Diana_Deutsch

Ερωτήσεις - Συζήτηση