

Στατιστικός έλεγχος υποθέσεων



ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Νικόλαος Πλαστήρας
7110112200218

Εργασία στα πλαίσια του μαθήματος:
Υπολογιστική Φυσική

Διδάσκοντες:
Φ. Διάκονος, Δ. Φασουλιώτης

2023

Περιεχόμενα

1	Εισαγωγή	1
2	Στατιστικός Έλεγχος Υποθέσεων	1
2.1	Βασικές Έννοιες	1
2.2	Στατιστική Ελέγχου	3
2.3	Λήμμα Neyman-Pearson	5
2.4	Θεώρημα Wilks	5
2.5	Λόγος Πιθανοφανειών για Αναζήτηση Σήματος	6
2.6	p-value	7
2.7	Σημαντικότητα	8
2.8	Υπολογισμός Σημαντικότητας μέσω Monte Carlo	9
3	Υπολογιστική Προσομοίωση	9
3.1	Τεχνικά στοιχεία	10
3.2	Αναζήτηση Σήματος	10
3.2.1	Υπόβαθρο	10
3.2.2	Σήμα	12
3.2.3	Μοντέλο	14
3.2.4	Έλεγχος υπόθεσης	19
3.3	Διαστήματα Εμπιστοσύνης	20
3.3.1	Δημιουργία μοντέλου προσομοίωσης	20
3.3.2	Έλεγχος υπόθεσης	23
4	Βιβλιογραφία	30

Κατάλογος Σχημάτων

1	Σφάλματα α και β για τον έλεγχο της μηδενικής υπόθεσης H_0 κατά της εναλλακτικής H_1 [2]	2
2	Η συνάρτηση πυκνότητας πιθανότητας της στατιστικής ελέγχου q με σχηματική την περιοχή απόρριψης για τις τρεις διαφορετικές περιπτώσεις. . .	4
3	Σχηματική αναπαράσταση της σχέσης μεταξύ της p-value και της σημαντικότητας [9]	8
4	Συνάρτηση πυκνότητας πιθανότητας του υποβάθρου	11
5	Συνάρτηση πυκνότητας πιθανότητας του υποβάθρου σε μορφή ιστογράμματος	12
6	Συνάρτηση πυκνότητας πιθανότητας του σήματος	13
7	Συνάρτηση πυκνότητας πιθανότητας σε μορφή ιστογράμματος	14
8	Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα για την εναλλακτική υπόθεση	16
9	Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα για την μηδενική υπόθεση	17
10	Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα για την εναλλακτική υπόθεση στην μορφή ιστογράμματος	18
11	Αποτελέσματα ελέγχου υπόθεσης	20
12	Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα	22
13	Αποτελέσματα ελέγχου υπόθεσης (ProfileLikelihoodCalculator)	24
14	Κατανομή της στατιστικής ελέγχου συναρτήσει της παραμέτρου ενδιαφέροντος	25
15	Αποτελέσματα ελέγχου υπόθεσης (AsymptoticCalculator)	25
16	Διαστήματα εμπιστοσύνης (AsymptoticCalculator)	26
17	Αποτελέσματα ελέγχου υπόθεσης (FrequentistCalculator)	27
18	Κατανομή της στατιστικής ελέγχου για τις δύο υποθέσεις.	28
19	Διαστήματα εμπιστοσύνης (FrequentistCalculator)	29

1 Εισαγωγή

Η Φυσική Υψηλών Ενεργειών βασίζεται σε μία “υπόθεση”, το Καθιερωμένο Πρότυπο (Standard Model). Αν και σε μεγάλο βαθμό επιτυχές, το Καθιερωμένο Πρότυπο αδυνατεί να εξηγήσει κάποια σημαντικά ερωτήματα, π.χ. το πρόβλημα ιεραρχίας. Για να απαντηθούν αυτά, έχουν προταθεί νέες θεωρίες οι οποίες περιλαμβάνουν καινούργια σωματίδια τα οποία όμως δεν έχουν παρατηρηθεί ακόμα. Ο στόχος της Φυσικής Υψηλών Ενεργειών είναι η δημιουργία τεράστιου όγκου δεδομένων και η ανάλυση αυτού έτσι ώστε να διαπιστωθεί εάν αυτά περιέχουν όντως ενδείξεις για νέα σωματίδια.

Η στατιστική πρόκληση είναι προφανής, πρέπει με τον καλύτερο δυνατό τρόπο και με βάση τις τωρινές γνώσεις μας, να διαπιστώσουμε εάν στα δεδομένα μας υπάρχει “Νέα Φυσική”, δηλαδή Φυσική πέραν του Καθιερωμένου Προτύπου. Υπό αυτή την έννοια, αυτό που είναι ήδη γνωστό είναι το υπόβαθρο (background). Βέβαια, η πολυπλοκότητα της ίδιας της Φυσικής αλλά και των ανιχνευτικών συστημάτων γεννούν μεγάλα συστηματικά σφάλματα, τόσο στο σήμα (signal) όσο και στο υπόβαθρο, τα οποία πρέπει επίσης να ληφθούν υπόψη στην στατιστική ανάλυση.

Σκοπός της παρούσας εργασίας είναι η παρουσίαση των μεθόδων που χρησιμοποιούνται για την στατιστική ανάλυση δεδομένων στην περίπτωση της Σωματιδιακής Φυσικής καθώς και η εφαρμογή αυτών σε ένα απλό υπολογιστικό πρόγραμμα προσομοίωσης.

2 Στατιστικός Έλεγχος Υποθέσεων

Ο στατιστικός έλεγχος υποθέσεων (hypothesis testing) είναι μία συμπερασματική μέθοδος που προσφέρει η Στατιστική Συμπερασματολογία η οποία βρίσκει εφαρμογή σε στοχαστικά προβλήματα απόφασης μεταξύ δύο εναλλακτικών υποθέσεων. Η μία υπόθεση έχει επικρατήσει να συμβολίζεται με H_0 και ονομάζεται μηδενική υπόθεση (null hypothesis) και η άλλη με H_1 και ονομάζεται εναλλακτική υπόθεση (alternative hypothesis).

Στην Σωματιδιακή Φυσική συναντάται πολύ συχνά το πρόβλημα του ελέγχου υποθέσεων όπου ζητείται η αποδοχή ή απόρριψη μιας στατιστικής εικασίας. Χαρακτηριστικό παράδειγμα αποτελεί η διερεύνηση των πειραματικών δεδομένων για το εάν αυτά περιλαμβάνουν γεγονότα που αποτελούνται μόνο από υπόβαθρο ή εάν περιέχουν ένα μίγμα σήματος και υποβάθρου. Με αυτό τον τρόπο μπορούμε να προσδιορίσουμε την ύπαρξη ή μη σήματος και κατ’ επέκταση να οδηγηθούμε σε μία νέα ανακάλυψη [1].

2.1 Βασικές Έννοιες

Η γενική ιδέα της διαδικασίας στατιστικού ελέγχου υποθέσεων είναι η εξής: θέτουμε ως μηδενική υπόθεση (H_0) αυτή για την οποία αμφιβάλλουμε, δηλαδή αυτή που αμφισβητείται, και εξετάζουμε εάν σε ένα τυχαίο δείγμα ενός πληθυσμού προκύπτουν αποδείξεις υπέρ της απόρριψής της έναντι της εναλλακτικής (H_1). Με άλλα λόγια η H_0 απορρίπτεται ή δεν απορρίπτεται με βάση το τι παρατηρείται στο τυχαίο δείγμα που πήραμε από τον πληθυσμό. Πιο συγκεκριμένα, υποθέτοντας ότι η H_0 είναι αληθής, αν αυτό που παρατηρείται στο δείγμα είναι ακραίο, δηλαδή, αν έχει πολύ μικρή πιθανότητα να συμβεί τότε απορρίπτουμε την H_0 . Σε αντίθετη περίπτωση, αν αυτό που παρατηρείται στο δείγμα δεν είναι ακραίο-σπάνιο (όταν είναι αληθής η H_0), τότε το δείγμα που πήραμε δεν μας δίνει αρκετές ενδείξεις για την απόρριψη της H_0 και “αποτυγχάνουμε να την απορρίψουμε”.

Προφανώς, ακολουθώντας αυτή την στρατηγική παίρνουμε “ρίσκο” γιατί και τα ακραία, έστω και με πολύ μικρή πιθανότητα μπορεί να συμβούν. Ειδικότερα, αν υποθέσουμε ότι η H_0 είναι αληθής αλλά την απορρίψουμε γιατί από το τυχαίο δείγμα προκύπτει ότι αυτό που παρατηρείται είναι ακραίο, τότε ακριβώς ένα από τα παρακάτω μπορεί να συνέβη:

1. είτε η H_0 πράγματι δεν είναι αληθής και το συμπέρασμα μας είναι σωστό
2. είτε η H_0 είναι αληθής και αυτό που συνέβη είναι κάτι σπάνιο.

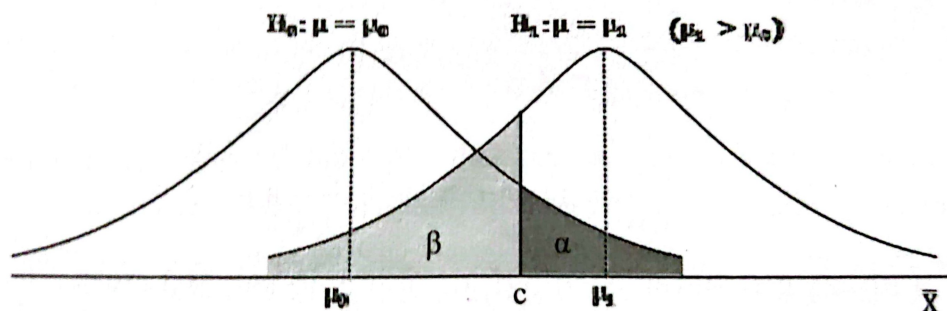
Στην δεύτερη περίπτωση απορρίψαμε λανθασμένα την H_0 . Αυτό ονομάζεται και ως σφάλμα τύπου I (type I error). Αντίστοιχα είναι δυνατή και η αντίθετη περίπτωση, λανθασμένα να μην απορρίψουμε την H_0 . Δηλαδή να αποτύχουμε να απορρίψουμε την H_0 , ενώ είναι αληθής η H_1 . Αυτό ονομάζεται σφάλμα τύπου II (type II error). Συνεπώς, υπάρχουν τέσσερις δυνατές περιπτώσεις στην λήψη μιας απόφασης:

	Ισχύει η H_0	Ισχύει η H_1
Η H_0 γίνεται δεκτή	Ορθή αποδοχή της H_0	Σφάλμα τύπου II
Η H_0 απορρίπτεται	Σφάλμα τύπου I	Ορθή απόρριψη της H_0

Επομένως, το “ρίσκο” είναι διπλό με πιθανότητα:

1. λανθασμένης απόρριψης της H_0 ,
 $\text{Prob}(\text{σφάλμα τύπου I}) = \text{Prob}(\text{απόρριψη της } H_0 \mid \text{αληθής η } H_0) \equiv \alpha$
2. λανθασμένης μη απόρριψης της H_0 ,
 $\text{Prob}(\text{σφάλμα τύπου II}) = \text{Prob}(\text{μη απόρριψη της } H_0 \mid \text{αληθής η } H_1) \equiv \beta$

Η πιθανότητα λάθους τύπου I, δηλαδή η πιθανότητα απόρριψης της υπόθεσης H_0 ενώ αυτή είναι αληθής, ορίζει το επίπεδο σημαντικότητας (significance level), το οποίο συμβολίζεται με α . Αντίστοιχα, η πιθανότητα λάθους τύπου II, η οποία ονομάζεται και ως πιθανότητα εσφαλμένης ταυτοποίησης (misidentification probability) ορίζεται να είναι ίση με β . Προφανώς, οι συμπληρωματικές πιθανότητες θα είναι $1 - \alpha$ και $1 - \beta$ αντίστοιχα, με την πρώτη να ονομάζεται και ως απόδοση επιλογής σήματος (signal selection efficiency) και η άλλη ως ισχύς του ελέγχου (power of the test).

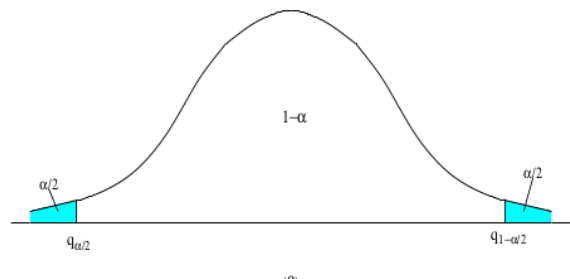


Σχήμα 1: Σφάλματα α και β για τον έλεγχο της μηδενικής υπόθεσης H_0 κατά της εναλλακτικής H_1 [2]

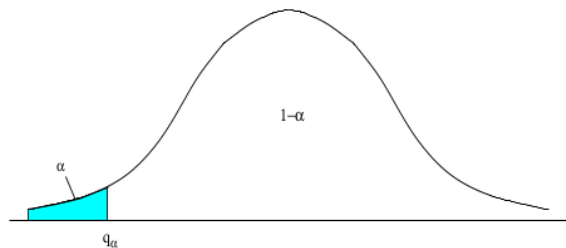
2.2 Στατιστική Ελέγχου

Ο έλεγχος υποθέσεων συνήθως προσδιορίζεται στα πλαίσια μίας στατιστικής ελέγχου (test statistic) q , η οποία θεωρείται ως μία αριθμητική σύνοψη ενός συνόλου δεδομένων σε μία μεταβλητή η οποία μπορεί να χρησιμοποιηθεί για τον έλεγχο της στατιστικής υπόθεσης. Γενικά μια στατιστική ελέγχου επιλέγεται ή ορίζεται με τέτοιο τρόπο ώστε να ποσοτικοποιεί συμπεριφορές που διακρίνουν την μηδενική από την εναλλακτική υπόθεση.

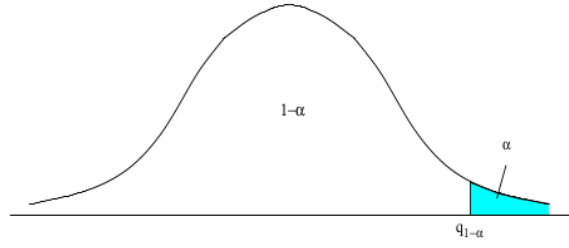
Ο έλεγχος ξεκινάει υποθέτοντας ότι η μηδενική υπόθεση H_0 είναι σωστή. Με βάση την H_0 προσδιορίζουμε την στατιστική ελέγχου q και την κατανομή της. Η κατανομή της δίνει την πιθανότητα η q να πάρει κάποια τιμή (αν η q είναι διακριτή τυχαία μεταβλητή) ή να βρίσκεται σε ένα διάστημα τιμών (αν η q είναι συνεχής τυχαία μεταβλητή) όταν η H_0 είναι αληθής. Αντίστροφα μπορούμε να πούμε πως με βάση αυτή την κατανομή, αν παρατηρήσουμε τιμές της q που αντιστοιχούν σε μεγάλες πιθανότητες αυτό δείχνει πως η H_0 είναι αληθής, ενώ αν παρατηρήσουμε τιμές της q που αντιστοιχούν σε μικρές πιθανότητες αυτό υποδηλώνει αμφιβολία για την ισχύ της H_0 . Άρα μη πιθανές τιμές της q συνιστούν την απόρριψη της H_0 . Η οριακή πιθανότητα για την αποδοχή ή απόρριψη της H_0 όπως έχουμε πει είναι το επίπεδο σημαντικότητας α κι αυτό καθορίζει την κρίσιμη τιμή (critical value), ή τις κρίσιμες τιμές, της q για τον προσδιορισμό της περιοχής αποδοχής και της περιοχής απόρριψης R της H_0 . Κατά κανόνα η περιοχή απόρριψης σχηματίζεται από τα άκρα της κατανομής της στατιστικής ελέγχου q όπως αυτά ορίζονται από τις κρίσιμες τιμές. Αν ο έλεγχος είναι δίπλευρος τότε οι κρίσιμες τιμές $q_{\alpha/2}$ και $q_{1-\alpha/2}$ ορίζουν την περιοχή απόρριψης της H_0 ως $R = \{q | q < q_{\alpha/2} \vee q > q_{1-\alpha/2}\}$, δηλαδή σχηματίζεται από τις δύο ουρές της κατανομής της q με συνολική πιθανότητα α . Αν ο έλεγχος είναι μονόπλευρος τότε υπάρχει μόνο μία κρίσιμη τιμή, q_α ή $q_{1-\alpha}$, που ορίζει την περιοχή απόρριψης της H_0 , $R = \{q | q < q_\alpha\}$ για την αριστερή πλευρά και $R = \{q | q > q_{1-\alpha}\}$ για τη δεξιά πλευρά.



(a) Δίπλευρος έλεγχος



(b) Μονόπλευρος έλεγχος για την αριστερή πλευρά



(c) Μονόπλευρος έλεγχος για την δεξιά πλευρά

Σχήμα 2: Η συνάρτηση πυκνότητας πιθανότητας της στατιστικής ελέγχου q με σκιαγραφημένη την περιοχή απόρριψης για τις τρεις διαφορετικές περιπτώσεις.

Στην Σωματιδιακή Φυσική συνήθως χρησιμοποιείται ο μονόπλευρος έλεγχος και η κρίσιμη τιμή ονομάζεται και ως “cut” η οποία και επιλέγεται εκ των προτέρων. Προφανώς, ανάλογα με την επιλογή του cut θα μεταβληθεί και η πιθανότητα λάθους. Στόχος μας αποτελεί η επιλογή του cut να είναι τέτοια ώστε να ελαχιστοποιεί την πιθανότητα εσφαλμένης ταυτοποίησης για μία επιθυμητή τιμή της απόδοσης επιλογής σήματος.

Υπάρχουν πολλοί τρόποι ορισμού μίας στατιστικής ελέγχου ανάλογα με την φύση της απαιτούμενης υπόθεσης. Στην Σωματιδιακή Φυσική, για την περίπτωση ανακάλυψης ή ορίων αποκλεισμού, συνήθως χρησιμοποιούνται στατιστικές ελέγχου που βασίζονται σε λόγους πιθανοφάνειας (Likelihood ratios). Να σημειώσουμε ότι η πιθανοφάνεια (likelihood) είναι μία συνάρτηση των δεδομένων,

$$L(H_0) = \text{Prob}(\vec{x}|H_0) \quad (1)$$

όπου $\vec{x} = \{x_1, \dots, x_n\}$ είναι τα δεδομένα. Ειδικότερα, η πιθανοφάνεια είναι μία συνάρτηση η οποία για δεδομένες τιμές των αγνώστων παραμέτρων, έστω $\vec{\theta} = (\theta_1, \dots, \theta_m)$, επιστρέφει την τιμή της συνάρτησης πυκνότητας πιθανότητας (Probability Density Function - PDF), f , υπολογισμένη από το σύνολο δεδομένων,

$$L(x_1, \dots, x_n; \theta_1, \dots, \theta_m) = f(x_1, \dots, x_n; \theta_1, \dots, \theta_m) \quad (2)$$

Εάν έχουμε N επανειλημμένες μετρήσεις όπου κάθε μία περιέχει n τιμές για τις τυχαίες μεταβλητές, x_1, \dots, x_n , τότε η συνάρτηση πιθανοφάνειας είναι η πυκνότητα πιθανότητας που αντιστοιχεί στο συνολικό δείγμα $\vec{x} = \{(x_1^1, \dots, x_n^1), \dots, (x_1^N, \dots, x_n^N)\}$. Εάν οι μετρήσεις είναι ανεξάρτητες μεταξύ τους τότε η συνάρτηση πιθανοφάνειας του δείγματος που αποτελείται από N γεγονότα, μπορεί να γραφεί ως το γινόμενο των συναρτήσεων πυκνότητας πιθανότητας που αντιστοιχούν σε ένα μόνο γεγονός,

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^N f(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m) \quad (3)$$

Συνήθως υπολογίζουμε τον φυσικό λογάριθμο της συνάρτησης πιθανοφάνειας καθώς έτσι το γινόμενο που εμφανίζεται στον παραπάνω ορισμό μετατρέπεται σε άθροισμα λογαρίθμων,

$$-\ln L(\vec{x}; \vec{\theta}) = \sum_{n=1}^N \ln f(x_1^i, \dots, x_n^i; \theta_1, \dots, \theta_m) \quad (4)$$

2.3 Λήμμα Neyman-Pearson

Μία στατιστική ελέγχου η οποία εξασφαλίζει την βέλτιστη απόδοση στην ελαχιστοποίηση της πιθανότητας εσφαλμένης ταυτοποίησης, παρέχεται από το λήμμα Neyman-Pearson [3]. Σύμφωνα με αυτό το λήμμα, η στατιστική ελέγχου ορίζεται ως ο λόγος των πιθανοφανειών υπολογισμένες κάτω από τις δύο υποθέσεις H_0 και H_1 ,

$$q^{NP} = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)} \quad (5)$$

Για μία καθορισμένη τιμή της απόδοσης επιλογής σήματος, $\epsilon = 1 - \alpha$ η λύση που καταλήγει στην χαμηλότερη δυνατή πιθανότητα εσφαλμένης ταυτοποίησης β δίνεται από,

$$q^{NP} = \frac{L(\vec{x}|H_1)}{L(\vec{x}|H_0)} \geq k_\alpha \quad (6)$$

όπου η τιμή του cut k_α πρέπει να επιλεχθεί έτσι ώστε να επιτευχθεί η απαιτούμενη τιμή για το α .

Εάν οι πολυδιάστατες συναρτήσεις πυκνότητας πιθανότητας που χαρακτηρίζουν το πρόβλημα είναι γνωστές, στην προκειμένη περίπτωση $f(q^{NP}|H_0)$, $f(q^{NP}|H_1)$, τότε το λήμμα Neyman-Pearson παρέχει μία διαδικασία για την υλοποίηση μίας επιλογής που επιτυγχάνει την βέλτιστη απόδοση.

Σε πολλές ρεαλιστικές περιπτώσεις δεν είναι εύκολο να καθορίσει κανείς το σωστό μοντέλο για αυτές τις πολυδιάστατες PDFs και μπορεί να χρησιμοποιηθούν προσεγγιστικές λύσεις. Υπάρχουν αριθμητικές μέθοδοι και αλγόριθμοι για την εύρεση επιλογών στο χώρο των μεταβλητών που έχουν επιδόσεις όσον αφορά την αποτελεσματικότητα και την πιθανότητα εσφαλμένης αναγνώρισης κοντά στο βέλτιστο όριο που δίνεται από το λήμμα Neyman-Pearson. Μερικοί από αυτούς τους αλγόριθμους όμως έχουν πολύ μεγάλη πολυπλοκότητα. Μεταξύ τέτοιων μεθόδων μερικές από τις πιο συχνά χρησιμοποιούμενες στην Φυσική Υψηλών Ενέργειών είναι τα νευρωνικά δίκτυα (Artificial Neural Networks) και τα ενισχυμένα δέντρα απόφασης (Boosted Decision Trees) [4, 5].

2.4 Θεώρημα Wilks

Όταν ένας μεγάλος αριθμός μετρήσεων είναι διαθέσιμος, το θεώρημα του Wilks [6] επιτρέπει την εύρεση μίας προσεγγιστικής ασυμπτωτικής έκφρασης για την στατιστική ελέγχου βασισζόμενο στον λόγο πιθανοφανειών από το λήμμα Neyman-Pearson [7].

Υποθέτουμε ότι έχουμε δύο υποθέσεις H_0 και H_1 οι οποίες μπορούν να εκφραστούν συναρτήσει κάποιων παραμέτρων $\vec{\theta} = (\theta_1, \dots, \theta_m)$ οι οποίες εμφανίζονται στον ορισμό της συνάρτησης πιθανοφάνειας. Όταν η υπόθεση H_1 είναι αληθής τότε η παράμετρος $\vec{\theta}$ ανήκει σε ένα συγκεκριμένο υποσύνολο του παραμετρικού χώρου Θ , έστω Θ_1 . Όταν η H_0 είναι αληθής τότε αντίστοιχα έχουμε $\vec{\theta} \in \Theta_0$. Επίσης, υποθέτουμε ότι $\Theta_0 \subseteq \Theta_1$ (nested hypotheses). Το θεώρημα του Wilks εξασφαλίζει ότι η ποσότητα,

$$\chi_r^2 = -2 \ln \frac{\sup_{\vec{\theta} \in \Theta_1} \prod_{i=1}^N L(\vec{x}_i; \vec{\theta})}{\sup_{\vec{\theta} \in \Theta_0} \prod_{i=1}^N L(\vec{x}_i; \vec{\theta})} \quad (7)$$

που αντιστοιχεί σε ένα παρατηρούμενο δείγμα δεδομένων $(\vec{x}_1, \dots, \vec{x}_N)$, έχει μία κατανομή η οποία μπορεί να προσεγγιστεί για $N \rightarrow \infty$, εάν η H_0 είναι αληθής, με μία κατανομή χ^2 με βαθμούς ελευθερίας όσους είναι η διαφορά μεταξύ των διαστάσεων των Θ_1 και Θ_0 .

Ειδικότερα, έστω ότι χωρίζουμε τις παραμέτρους σε μία παράμετρο ενδιαφέροντος μ και τις υπόλοιπες τις παίρνουμε ως “nuisance”, $\vec{\theta} = (\theta_1, \dots, \theta_{m-1})$. Για παράδειγμα, η παράμετρος μ μπορεί να είναι το σθένος του σήματος ($\mu = \frac{\sigma_{obs}}{\sigma_{the}}$) όπου για $\mu = 1$ έχουμε την παρατηρούμενη ενεργό διατομή ίση με την θεωρητικά προβλεπόμενη. Εάν παρόυμε ως την μηδενική υπόθεση να είναι $\mu = \mu_0$, ενώ η εναλλακτική να είναι το μ να έχει οποιαδήποτε άλλη τιμή μεγαλύτερη ή ίση του μηδενός, το θεώρημα Wilks εξασφαλίζει ότι η ποσότητα

$$\chi_r^2(\mu_0) = -2 \ln \frac{\sup_{\vec{\theta}} \prod_{i=1}^N L(\vec{x}_i; \mu_0, \vec{\theta})}{\sup_{\mu, \vec{\theta}} \prod_{i=1}^N L(\vec{x}_i; \mu, \vec{\theta})} \quad (8)$$

είναι ασυμπτωτικά κατανεμημένη ως χ^2 με k βαθμούς ελευθερίας.

Ο παρονομαστής, $\sup_{\mu, \vec{\theta}} \prod_{i=1}^N L(\vec{x}_i; \mu, \vec{\theta})$ είναι η πιθανότητα υπολογισμένη σε σημείο μέγιστης πιθανοφάνειας, δηλαδή στις τιμές των παραμέτρων $\mu = \hat{\mu}$ και $\vec{\theta} = \hat{\vec{\theta}}$ οι οποίες μεγιστοποιούν την συνάρτηση πιθανοφάνειας,

$$\prod_{i=1}^N L(\vec{x}_i; \hat{\mu}, \hat{\vec{\theta}}) \quad (9)$$

Για τον αριθμητή μόνο οι παράμετροι $\vec{\theta}$ είναι ελεύθερες αφού το μ είναι σταθερό και ίσο με $\mu = \mu_0$. Εάν οι τιμές για τις $\vec{\theta}$ που μεγιστοποιούν την συνάρτηση πιθανοφάνειας με $\mu = \mu_0$, είναι $\vec{\theta} = \hat{\vec{\theta}}(\mu_0)$ τότε η Εξίσωση 8 μπορεί να γραφεί ως,

$$\chi_r^2 = \frac{L(\vec{x}|\mu, \hat{\vec{\theta}}(\mu_0))}{L(\vec{x}|\hat{\mu}, \hat{\vec{\theta}})} \quad (10)$$

Αυτή η στατιστική ελέγχου ονομάζεται στην βιβλιογραφία και ως “profile likelihood” και είναι πολύ χρήσιμη για τον έλεγχο ύπαρξης σήματος.

2.5 Λόγος Πιθανοφανειών για Αναζήτηση Σήματος

Σε προηγούμενη ενότητα θεωρήσαμε ότι η συνάρτηση πιθανοφάνειας για ένα σετ μετρήσεων $\{\vec{x}_1, \dots, \vec{x}_N\}$ με ένα σετ παραμέτρων $\vec{\theta}$, ότι είναι,

$$L(\vec{x}_1, \dots, \vec{x}_N; \vec{\theta}) = \prod_{i=1}^N f(\vec{x}_i; \vec{\theta}) \quad (11)$$

Οι δύο υποθέσεις H_1 και H_0 αναπαριστώνται ως δύο πιθανά σετ τιμών Θ_1 και Θ_0 των παραμέτρων $\vec{\theta} = (\theta_1, \dots, \theta_m)$ που χαρακτηρίζουν τις συναρτήσεις πυκνότητας πιθανότητας.

Συνήθως θέλουμε να χρησιμοποιήσουμε τον αριθμό των γεγονότων N ως πληροφορία για τον ορισμό της πιθανοφάνειας, για αυτό χρησιμοποιούμε την εκτεταμένη συνάρτηση πιθανοφάνειας,

$$L = p(N; \theta_1, \dots, \theta_m) \prod_{i=1}^N f(\vec{x}_i; \vec{\theta}) \quad (12)$$

όπου στις περισσότερες περιπτώσεις ενδιαφέροντος της φυσικής η $p(N; \theta_1, \dots, \theta_m)$ είναι μία κατανομή Poisson, οπότε έχουμε τελικά,

$$L(\vec{x}_1, \dots, \vec{x}_N; \vec{\theta}) = \frac{e^{-\nu(\vec{\theta})} \nu(\vec{\theta})^N}{N!} \prod_{i=1}^N f(\vec{x}_i; \vec{\theta}) \quad (13)$$

όπου ο αναμενόμενος αριθμός γεγονότων ν μπορεί επίσης να εξαρτάται από τις παραμέτρους $\vec{\theta}$: $\nu = \nu(\vec{\theta})$. Συνήθως θέλουμε να κάνουμε διάκριση μεταξύ δύο υποθέσεων, η παρουσία μόνο γεγονότων υποβάθρου στο δείγμα μας, $\nu = b$ ενάντια στην παρουσία και σήματος και υποβάθρου $\nu = \mu s + b$. Ο παράγοντας μ είναι το σθένος σήματος που εισάγαμε στην προηγούμενη ενότητα. Η μηδενική υπόθεση H_0 που αντιστοιχεί στην παρουσία μόνο υποβάθρου είναι ισοδύναμη με $\mu = 0$, ενώ η εναλλακτική υπόθεση H_1 που αντιστοιχεί στην ύπαρξης σήματος συν υποβάθρου επιτρέπει οποιαδήποτε μη μηδενική, θετική τιμή για το μ .

Η συνάρτηση πυκνότητας πιθανότητας μπορεί να γραφεί ως υπέρθεση των δύο συνιστωσών, μία συνάρτηση για το σήμα και μία για το υπόβαθρο, σταθμισμένες ως προς τις αναμενόμενες τιμές τους,

$$f(\vec{x}; \vec{\theta}) = \frac{\mu s}{\mu s + b} f_s(\vec{x}; \vec{\theta}) + \frac{b}{\mu s + b} f_b(\vec{x}; \vec{\theta}) \quad (14)$$

Σε αυτή την περίπτωση, η εκτεταμένη συνάρτηση πιθανοφάνειας γράφεται,

$$L_{s+b}(\vec{x}_1, \dots, \vec{x}_N; \mu, \vec{\theta}) = \frac{e^{-(\mu s(\vec{\theta}) + b(\vec{\theta}))}}{N!} \prod_{i=1}^N (\mu s f_s(\vec{x}_i; \vec{\theta}) + b f_b(\vec{x}_i; \vec{\theta})) \quad (15)$$

Υπό την μηδενική υπόθεση H_0 όπου $\mu = 0$, η συνάρτηση πιθανοφάνειας γράφεται ως:

$$L_b(\vec{x}_1, \dots, \vec{x}_N; \vec{\theta}) = \frac{e^{-b(\vec{\theta})}}{N!} \prod_{i=1}^N (b f_b(\vec{x}_i; \vec{\theta})) \quad (16)$$

Παίρνοντας το λόγο των πιθανοφανειών σύμφωνα με το λήμμα Neyman-Pearson (Εξίσωση 5) προκύπτει,

$$q(\mu, \vec{\theta}) = \frac{L_{s+b}(\vec{x}_1, \dots, \vec{x}_N; \mu, \vec{\theta})}{L_b(\vec{x}_1, \dots, \vec{x}_N; \vec{\theta})} = e^{-\mu s(\vec{\theta})} \prod_{i=1}^N \left(\frac{\mu s f_s(\vec{x}_i; \vec{\theta})}{b f_b(\vec{x}_i; \vec{\theta})} + 1 \right) \quad (17)$$

2.6 p-value

Μία σημαντική ιδιότητα της στατιστικής ελέγχου είναι ότι η κατανομή της δειγματοληψίας κάτω από την μηδενική υπόθεση πρέπει να είναι υπολογίσιμη, είτε ακριβώς είτε κατά προσέγγιση, γεγονός που επιτρέπει τον υπολογισμό των p-values.

Η p-value είναι μία συνάρτηση που ποσοτικοποιεί πόσο συχνά, εάν ένα πείραμα επαναλαμβάνονταν πολλές φορές, θα λάμβανε κανείς δεδομένα τόσο μακριά από την μηδενική υπόθεση έως τα παρατηρούμενα δεδομένα, υποθέτοντας ότι η μηδενική υπόθεση είναι αληθής. Η p-value είναι μέτρηση του παρατηρούμενου επιπέδου σημαντικότητας. Αποτελεί συνάρτηση των δεδομένων συνεπώς πρόκειται για μία τυχαία μεταβλητή και δεν θα έπρεπε να συγχέεται με το επίπεδο σημαντικότητας α το οποίο είναι μία προκαθορισμένη σταθερά. Το επίπεδο σημαντικότητας εξαρτάται άμεσα από το cut το οποίο ορίζει την κρίσιμη περιοχή σε αντίθεση με την p-value η οποία εξαρτάται από την παρατηρούμενη τιμή της στατιστικής ελέγχου q_{obs} . Ειδικότερα, η p-value υπολογίζεται από,

$$p = \int_{q_{obs}}^{\infty} f(q|H_0) dq \quad (18)$$

Υπολογίζοντας την p-value του δείγματος, γνωρίζουμε πόσο πιθανή είναι η εμφάνιση του δείγματος που πήραμε με την υπόθεση ότι η H_0 είναι αληθής. Επομένως, όσο πιο μικρή είναι

η p-value τόσο ισχυρότερες ενδείξεις εναντίον της H_0 προκύπτουν από το συγκεκριμένο τυχαίο δείγμα ή αλλιώς τόσο πιο σημαντική είναι η τιμή της στατιστικής ελέγχου που δίνει το δείγμα [8].

Έτσι, αν θέλουμε να κάνουμε έλεγχο σε κάποιο επίπεδο σημαντικότητας α τότε υπολογίζοντας την p-value μπορούμε άμεσα να την συγκρίνουμε με αυτό το α και να αποφασίσουμε για την απόρριψη ή όχι της H_0 .

2.7 Σημαντικότητα

Συνήθως, αντί για την p-value, αναφερόμαστε σε ένα ισοδύναμο μέγεθος που ονομάζεται σημαντικότητα (significance), Z . Η σημαντικότητα ορίζεται να είναι ο αριθμός των τυπικών αποκλίσεων που αντιστοιχεί στην περιοχή της ουράς της κανονικής κατανομής η οποία έχει πιθανότητα ίση με την p-value (βλ. Σχήμα 3). Ειδικότερα, μπορούμε να εξάγουμε την σημαντικότητα από τον ακόλουθο μετασχηματισμό,

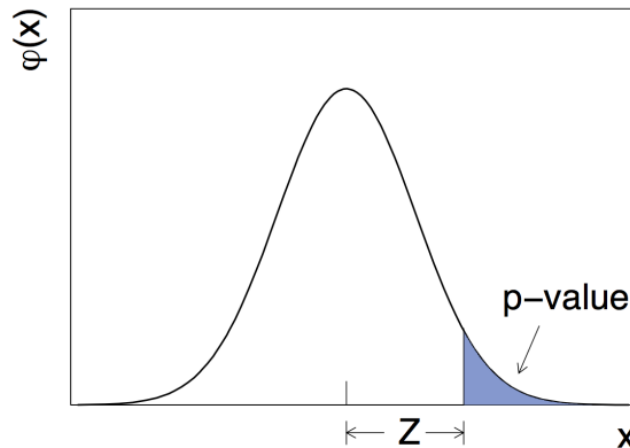
$$p = \int_Z^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1 - \Phi(Z) = \Phi(-Z) = \frac{1}{2} \left[1 - \operatorname{erf}\left(\frac{Z}{\sqrt{2}}\right) \right] \quad (19)$$

ή ανάποδα,

$$Z = \Phi^{-1}(1 - p) \quad (20)$$

όπου Φ είναι η συνάρτηση κατανομής (Cumulative Distribution Function) της Γκαουσιανής κατανομής:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{x'^2}{2}} dx' = \frac{1}{2} \left[\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right) + 1 \right] \quad (21)$$



Σχήμα 3: Σχηματική αναπαράσταση της σχέσης μεταξύ της p-value και της σημαντικότητας [9]

Κατά σύμβαση λέμε ότι έχουμε “παρατήρηση” (“observation”) του σήματος υπό εξέταση όταν η σημαντικότητα είναι τουλάχιστον 3σ ($Z = 3$), το οποίο αντιστοιχεί στην p-value να είναι ίση με 1.35×10^{-3} . Αντίστοιχα, λέμε ότι έχουμε “ανακάλυψη” (“discovery”) στην περίπτωση που η σημαντικότητα είναι τουλάχιστον 5σ ($Z = 5$) το οποίο είναι ισοδύναμο με την τιμή της p-value να είναι 2.87×10^{-7} [7].

Στην ενότητα 2.5 ορίσαμε την ακόλουθη στατιστική ελέγχου,

$$q(\mu, \vec{\theta}) = \frac{L_{s+b}(\vec{x}_1, \dots, \vec{x}_N; \mu, \vec{\theta})}{L_b(\vec{x}_1, \dots, \vec{x}_N; \vec{\theta})} \quad (22)$$

Στην περίπτωση όπου μας ενδιαφέρει μόνο μία παράμετρος μ , είναι δυνατό να παραστήσουμε γραφικά το $-\ln q(\mu)$ συναρτήσει του μ . Τότε, η παρουσία ενός ελάχιστου στο $\mu = \hat{\mu}$ είναι μία ένδειξη πιθανής παρουσίας σήματος με σθένος σήματος ίσο με $\hat{\mu}$ εντός κάποιας αβεβαιότητας. Για να προσδιορίσουμε την σημαντικότητα από το μετρούμενο σήμα, μπορούμε να εφαρμόσουμε το θεώρημα του Wilks, εάν η συνάρτηση της πιθανοφάνειας είναι επαρκώς ομαλή. Σε αυτή την περίπτωση, η κατανομή της $-2 \ln q(\mu)$ μπορεί να προσεγγιστεί από μία χ^2 κατανομή με έναν βαθμό ελευθερίας και η τετραγωνική ρίζα της τιμής του στο ελάχιστο δίνει προσεγγιστικά μία εκτίμηση για την σημαντικότητα, δηλαδή,

$$Z = \sqrt{-2 \ln(\hat{\mu})} \quad (23)$$

2.8 Υπολογισμός Σημαντικότητας μέσω Monte Carlo

Μια καλύτερη εκτίμηση της σημαντικότητας μπορεί να επιτευχθεί χρησιμοποιώντας ως στατιστική ελέγχου την ποσότητα $-2 \ln q$ παράγοντας έναν μεγάλο αριθμό ψευδοπειραμάτων (toy Monte Carlo) που αντιπροσωπεύουν τυχαία εξαγόμενα δείγματα τα οποία υποθέτουν ότι δεν υπάρχει σήμα ($\mu = 0$) έτσι ώστε να αποκτήσουν την αναμενόμενη κατανομή της $-2 \ln q$.

Η κατανομή της παραγόμενης συνάρτησης $-2 \ln q$ μπορεί να χρησιμοποιηθεί μαζί με την παρατηρούμενη τιμή $q = \hat{q}$ ώστε να καθοριστεί η p-value. Συγκεκριμένα η p-value είναι ίση με την πιθανότητα η q να είναι μικρότερη ή ίση από την παρατηρούμενη τιμή της στατιστικής ελέγχου \hat{q} .

$$p = P_b(q \leq \hat{q}) \quad (24)$$

η οποία με την σειρά της είναι ίση με το κλάσμα των παραγόμενων ψευδοπειραμάτων για τα οποία $q \leq \hat{q}$.

Να σημειώσουμε ότι προκειμένου να εκτιμηθούν μεγάλες τιμές της σημαντικότητας, ο αριθμός των παραγόμενων toy Monte Carlo ο οποίος απαιτείται για μία επαρκή ακρίβεια μπορεί να είναι πολύ μεγάλος, αφού η απαιτούμενη τιμή για την p-value είναι πάρα πολύ μικρή.

3 Υπολογιστική Προσομοίωση

Έχοντας κατανοήσει τα βασικά σημεία της θεωρίας της μεθόδου του στατιστικού ελέγχου υποθέσεων από την προηγούμενη ενότητα, είμαστε έτοιμοι να την χρησιμοποιήσουμε σε ένα απλό υπολογιστικό πρόβλημα. Ειδικότερα, θα την εφαρμόσουμε σε ένα από τα πιο συχνά προβλήματα της Φυσικής Υψηλών Ενεργειών, στην διαπίστωση της ύπαρξης ή μη, σήματος σε ένα σύνολο δεδομένων. Για τους σκοπούς της παρούσας εργασίας υποθέτουμε ότι αναζητούμε το μποζόνιο Higgs.

Το πρώτο βήμα που πρέπει να κάνουμε στην ανάλυση μας είναι να ταυτοποιήσουμε τις δύο υποθέσεις. Στην προκειμένη περίπτωση ορίζουμε ως την μηδενική υπόθεση να είναι η ύπαρξη μόνο υποβάθρου (background only hypothesis). Σκοπός μας είναι να ελέγξουμε αυτήν την υπόθεση και είτε να την αποδεχτούμε ως σωστή είτε να την απορρίψουμε υπέρ της εναλλακτικής υπόθεσης η οποία και είναι η ύπαρξη σήματος.

3.1 Τεχνικά στοιχεία

Για την υπολογιστική προσομοίωση του στατιστικού ελέγχου χρησιμοποιήσαμε το υλικολογισμικό ROOT [10] που έχει δημιουργηθεί και χρησιμοποιείται στο CERN. Ειδικότερα, χρησιμοποιήσαμε την βιβλιοθήκη RooFit [11] για την δημιουργία των συναρτήσεων πυκνότητας πιθανότητας και την βιβλιοθήκη RooStats [12] για τον στατιστικό έλεγχο του μοντέλου που δημιουργήσαμε.

Στην συνέχεια θα παρουσιάσουμε αναλυτικά την διαδικασία για την δημιουργία και τον στατιστικό έλεγχο του μοντέλου ενώ ο πλήρης κώδικας που χρησιμοποιήσαμε για την προσομοίωση βρίσκεται στο εξής repository: https://github.com/nplastir/Hypothesis_test.

3.2 Αναζήτηση Σήματος

Ως πρώτο παράδειγμα θα μελετήσουμε το κανάλι διάσπασης του Higgs σε δύο φωτόνια. Η επιλογή αυτής της διαδικασίας είναι ιδανική για τους σκοπούς της παρούσας εργασίας καθώς μπορεί να προσομοιωθεί πολύ εύκολα. Το κυρίαρχο υπόβαθρο αυτής της διαδικασίας αποτελεί προφανώς η παραγωγή δύο φωτονίων($\gamma\gamma$) ενώ επίσης έχουμε μικρές συνεισφορές από την παραγωγή $\gamma + \text{jet}$ και $\text{jet}+\text{jet}$ όπου ένα ή αντίστοιχα δύο jet έχουν ταυτοποιηθεί λανθασμένα ως φωτόνια καθώς και από την διαδικασία Drell-Yan [13]. Αυτό το υπόβαθρο, όπως θα δούμε στην συνέχεια μπορεί να προσομοιωθεί απλοϊκά ως μία εκθετική κατανομή. Η αναζήτηση του σήματος πραγματοποιείται στο εύρος μαζών 80 – 200 GeV:

```
// Set range of observable
Double_t low = 80, high = 200;

// Create a variable for the observable (invariant mass)
RooRealVar invMass("invMass", "M_{inv}", low, high, "GeV");
```

3.2.1 Υπόβαθρο

Όπως αναφέραμε, προσομοιώνουμε το υπόβαθρο ως μία εκθετική κατανομή. Συγκεκριμένα έχουμε:

```
// Create background dataset (exponential)
RooRealVar alpha("alpha", "#alpha", -0.01,-0.2,0.01);
RooExponential background("background", "Background PDF", invMass, alpha);
```

όπου έχουμε χρησιμοποιήσει την κλάση “RooExponential”[14] του πακέτου RooFit σύμφωνα με την οποία μπορούμε να δημιουργήσουμε μία εκθετική συνάρτηση πυκνότητας πιθανότητας:

$$RooExponential(x, c) = N \cdot \exp(c \cdot x) \quad (25)$$

όπου N είναι η σταθερά κανονικοποίησης η οποία εξαρτάται από το εύρος και τις τιμές των ορισμάτων.

Ο συντελεστής της εκθετικής συνάρτησης τον οποίο έχουμε ορίσει ως α τον έχουμε αφήσει ως ελεύθερο παράμετρο της κατανομής και η τιμή του προσδιορίζεται από την προσαρμογή που θα κάνουμε στα δεδομένα.

Έχοντας ορίσει την κατανομή για το υπόβαθρο μπορούμε πολύ εύκολα να την σχεδιάσουμε σε ένα διάγραμμα. Μάλιστα, μπορούμε να δημιουργήσουμε και κάποια MonteCarlo

δεδομένα με βάση αυτή την κατανομή και να την κάνουμε προσαρμογή σε αυτά:

```
//Toy MC generation for background - DataSet (unbinned)
RooDataSet *bkgData = background.generate(RooArgSet(invMass), 10000);

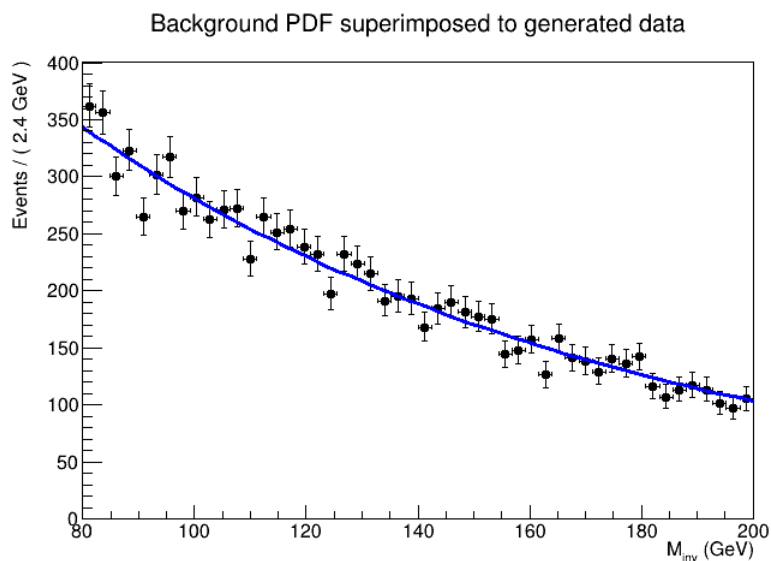
//Fit background model to data
background.fitTo(*bkgData);

//Create a ROOT Canvas
TCanvas *bkg_can = new TCanvas();

//Create a ROOT Canvas
TCanvas *bkg_can = new TCanvas();
//Create a plot of generated data superimposed to the function for
background
RooPlot *bkg_plot = invMass.frame();
bkgData->plotOn(bkg_plot);
background.plotOn(bkg_plot, RooFit::LineColor(kBlue));
//background.paramOn(bkg_plot);

//Draw components and save as a .png
bkg_plot->SetTitle("Background PDF superimposed to generated data");
bkg_plot->Draw();
bkg_can->Draw();
bkg_can->SaveAs("background.png");
```

Όπως βλέπουμε από τον παραπάνω κώδικα δημιουργούμε ένα σύνολο δεδομένων με βάση την κατανομή μας και την αποθηκεύουμε σε ένα container μέσω της κλάσης RooDataSet η οποία και επιτρέπει την μέγιστη ακρίβεια καθώς αποθηκεύει κάθε γεγονός (unbinned). Στην συνέχεια, κάνουμε την προσαρμογή των δεδομένων πάνω στην κατανομή αν και στην προκειμένη περίπτωση είναι περιττό καθώς τα δεδομένα έχουν δημιουργηθεί με βάση την ίδια κατανομή. Τέλος, δημιουργούμε και αποθηκεύουμε το διάγραμμα. Εν τέλει, το διάγραμμα που προκύπτει είναι το ακόλουθο:



Σχήμα 4: Συνάρτηση πυκνότητας πιθανότητας του υποβάθρου

Βέβαια, μπορεί τα δεδομένα μας να είναι αποθηκευμένα σε bins. Σε αυτή την περίπτωση έχουμε:

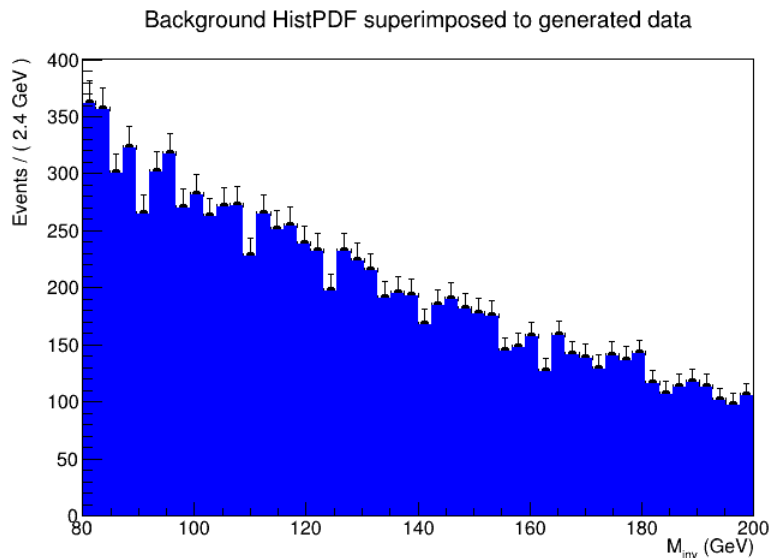
```
//Toy MC generation for background - DataHist (binned)
RooDataHist* bkgDataHist = bkgData->binnedClone() ;

//Hist PDF from toy MC
RooHistPdf bkgHistPdf("bkg", "bkg", invMass, *bkgDataHist, 0);

//Create a ROOT Canvas
TCanvas *bkg_binned_can = new TCanvas();

//Create a plot of the generated data superimposed to the background model
RooPlot *bkg_binned_plot = invMass.frame();
bkgDataHist->plotOn(bkg_binned_plot);
bkgHistPdf.plotOn(bkg_binned_plot, RooFit::DrawOption("F"),
    RooFit::FillColor(kBlue), RooFit::FillStyle(1001));
bkgHistPdf.paramOn(bkg_binned_plot);

//Draw components and save as a .png
bkg_binned_plot->SetTitle("Background HistPDF superimposed to generated
    data");
bkg_binned_plot->Draw();
bkg_binned_can->Draw();
bkg_binned_can->SaveAs("background_binned.png");
```



Σχήμα 5: Συνάρτηση πυκνότητας πιθανότητας του υποβάθρου σε μορφή ιστογράμματος

3.2.2 Σήμα

Όσον αφορά το σήμα στην συγκεκριμένη προσομοίωση χρησιμοποιήσαμε απλώς μία Γκαουσιανή κατανομή:

```
// Create signal dataset (gaussian)
RooRealVar mean("mean", "Signal Mean", 125, 90, 160);
RooRealVar sigma("sigma", "Signal Sigma", 10, 0, 20);
RooGaussian signal("signal", "Signal PDF", invMass, mean, sigma);
```

όπου έχουμε χρησιμοποιήσει την κλάση “RooGaussian” [15] του πακέτου RooFit.

Ακολουθώντας την αντίστοιχη διαδικασία με το υποβαθρο, μπορούμε να προσομοιώσουμε δεδομένα με βάση την κατανομή του σήματος μας και να δημιουργήσουμε το αντίστοιχο διάγραμμα:

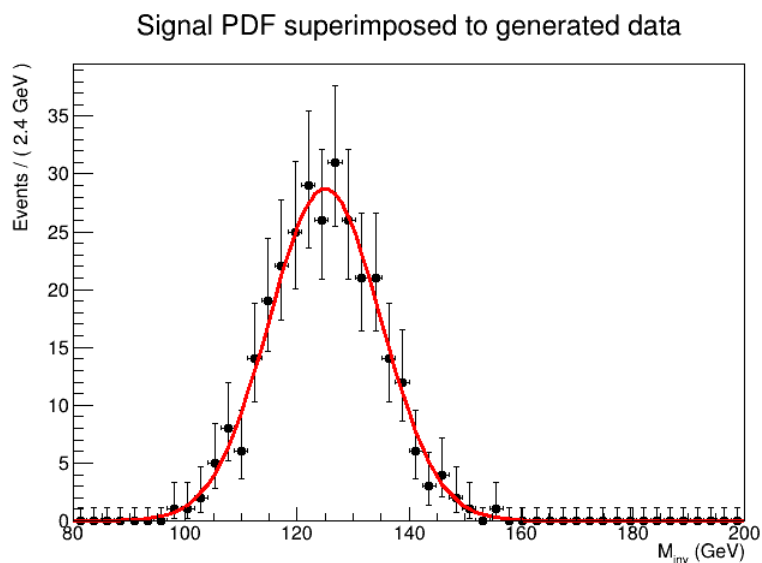
```
//Toy MC generation for signal - DataHist (unbinned)
RooDataSet *sigData = signal.generate(RooArgSet(x), 300);

//Fit signal model to generated data
signal.fitTo(*sigData);

//Create a ROOT Canvas
TCanvas *sig_can = new TCanvas();

//Create a plot of generated data superimposed to the function for signal
RooPlot *sig_plot = invMass.frame();
sigData->plotOn(sig_plot);
signal.plotOn(sig_plot, RooFit::LineColor(kRed));
//signal.paramOn(sig_plot);

//Draw components and save as a .png
sig_plot->SetTitle("Signal PDF superimposed to generated data");
sig_plot->Draw();
sig_can->Draw();
sig_can->SaveAs("signal.png");
```



Σχήμα 6: Συνάρτηση πυκνότητας πιθανότητας του σήματος

Αντίστοιχα, για την περίπτωση που έχουμε τα δεδομένα μας αποθηκευμένα σε bins:

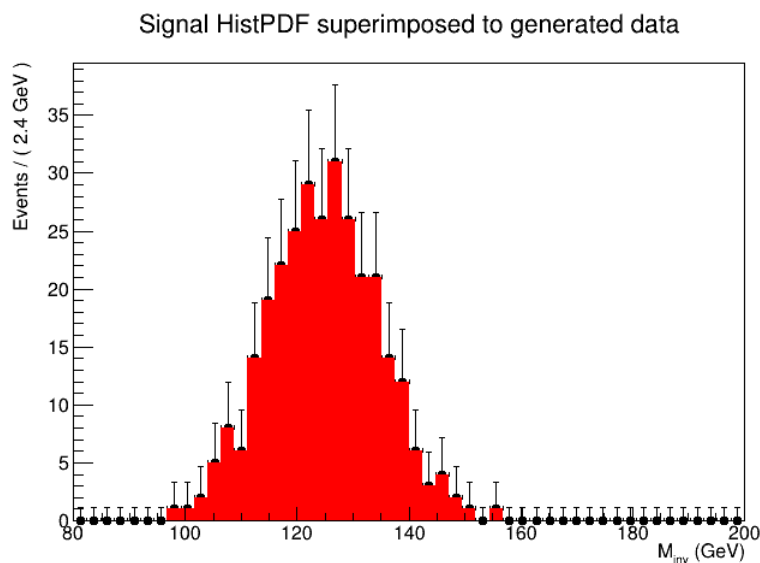
```
//Toy MC generation for background - DataHist (binned)
RooDataHist* sigDataHist = sigData->binnedClone() ;

//Hist PDF from toy MC
RooHistPdf sigHistPdf("sig","sig",invMass,*sigDataHist,0);

//Create a ROOT Canvas
TCanvas *sig_binned_can = new TCanvas();

//Create a plot of the generated data superimposed to the background model
RooPlot *sig_binned_plot = invMass.frame();
sigDataHist->plotOn(sig_binned_plot);
sigHistPdf.plotOn(sig_binned_plot, RooFit::DrawOption("F"),
    RooFit::FillColor(kRed), RooFit::FillStyle(1001));
//sigHistPdf.paramOn(sig_binned_plot);

//Draw components and save as a .png
sig_binned_plot->SetTitle("Signal HistPDF superimposed to generated data");
sig_binned_plot->Draw();
sig_binned_can->Draw();
sig_binned_can->SaveAs("signal_binned.png");
```



Σχήμα 7: Συνάρτηση πυκνότητας πιθανότητας σε μορφή ιστογράμματος

3.2.3 Μοντέλο

Έχοντας δημιουργήσει το σήμα και το υπόβαθρο, είμαστε σε θέση να φτιάξουμε το μοντέλο προσομοίωσης μας. Αρχικά, εισάγουμε μία ακόμη μεταβλητή, το σθένος του σήματος, $\mu = \frac{\sigma_{obs}}{\sigma_{SM}}$.

```
// Introduce mu
RooRealVar mu("mu", "signal strength in units of SM expectation",1, 0., 2);
```

Προφανώς, όταν έχουμε ότι $\mu = 1$ τότε η παρατηρούμενη ενεργός διατομή είναι ίση με την προβλεπόμενη από το Καθιερωμένο Πρότυπο, ενώ προφανώς, για $\mu = 0$ δεν έχουμε σήμα.

Επίσης εισάγουμε τις μεταβλητές που σχετίζονται με τον αναμενόμενο αριθμό γεγονότων σήματος καθώς και την απόδοση μέτρησης τους.

```
// Introduce signal parameters
RooRealVar fsigExpected("fsigExpected", "expected fraction of signal
    events", 0.05, 0., 1);

RooRealVar ratioSigEff("ratioSigEff", "ratio of signal efficiency to
    nominal signal efficiency", 0.5, 0., 2);
```

Για την προκειμένη ανάλυση θεωρούμε ότι η παράμετρος ενδιαφέροντος είναι το σθένος σήματος μ για αυτό και τις υπόλοιπες τις ορίζουμε να είναι σταθερές.

```
// Use mu as main parameter, so fix the other variables.
mean.setConstant();
sigma.setConstant();
fsigExpected.setConstant();
ratioSigEff.setConstant();
```

Για ευκολία, ορίζουμε μία νέα μεταβλητή η οποία πρόκειται για τον συνολικό συντελεστή της κατανομής του σήματος και η οποία είναι απλά το γινόμενο των παραπάνω παραμέτρων.

```
//The signal parameter
RooProduct fsig("fsig", "fraction of signal events", RooArgSet(mu,
    ratioSigEff, fsigExpected));
```

Οπότε, μπορούμε να κατασκευάσουμε το μοντέλο προσομοίωσης:

```
//Create model = fsig*signal + (1-fsig)*background
RooAddPdf model("model", "Data", RooArgList(signal, background), fsig);
```

Με τον ίδιο τρόπο που χρησιμοποιήσαμε για να κατασκευάσουμε το σήμα και το υπόβραθρο, δημιουργούμε μερικά δεδομένα με βάση το μοντέλο μας και το κάνουμε προσαρμογή σε αυτά:

```
//Toy MC generation - DataSet (unbinned)
RooDataSet *Data = model.generate(RooArgSet(invMass), 5000);

model.fitTo(*Data, Save(kTRUE), Minos(kFALSE), Hesse(kFALSE),
    PrintLevel(-1));

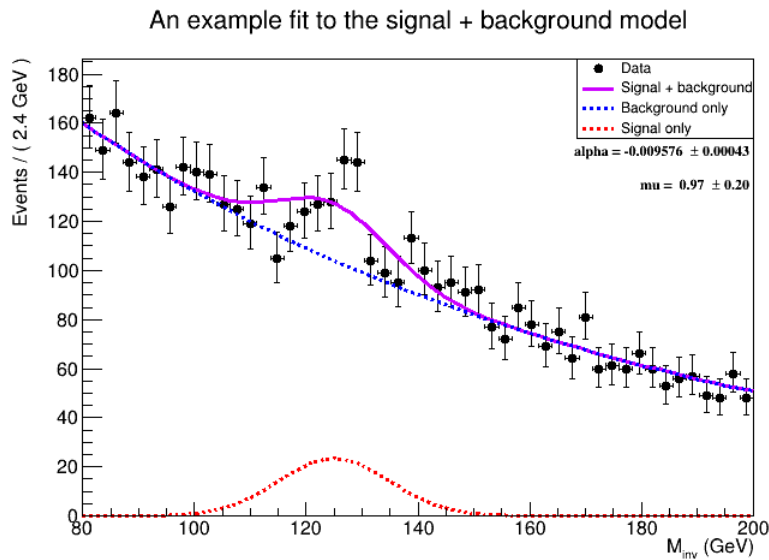
// plot sig candidates, full model, and individual components
TCanvas *sbmodel_can =new TCanvas();
RooPlot *sbmodel_plot = invMass.frame();
Data->plotOn(sbmodel_plot, Name("data"));
model.plotOn(sbmodel_plot, Name("model"), LineColor(kViolet));
model.plotOn(sbmodel_plot, Name("signal only"), Components(signal),
    LineStyle(kDashed), LineColor(kRed));
model.plotOn(sbmodel_plot, Name("background"), Components(background),
    LineStyle(kDashed), LineColor(kBlue));
model.paramOn(sbmodel_plot);
```

```

//Create Legend
TLegend *leg1 = new TLegend(0.65,0.73,0.86,0.87);
leg1->SetFillColor(kWhite);
leg1->SetLineColor(kBlack);
leg1->AddEntry(sbmodel_plot->findObject("data"), "Data", "P");
leg1->AddEntry(sbmodel_plot->findObject("model"),"Signal +
    background","L");
leg1->AddEntry(sbmodel_plot->findObject("background"), "Background only",
    "L");
leg1->AddEntry(sbmodel_plot->findObject("signal only"), "Signal only",
    "L");

//Draw components and save as .png
sbmodel_plot->SetTitle("An example fit to the signal + background model");
sbmodel_plot->Draw();
sbmodel_can->Draw();
leg1->Draw();
sbmodel_can->SaveAs("signal+background_model.png");

```



Σχήμα 8: Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα για την εναλλακτική υπόθεση

Όπως φαίνεται από το Σχήμα 8 η καλύτερη προσαρμογή στα δεδομένα μας είναι αυτή που δίνει σθένος σήματος ίσο με $\mu = 0.97 \pm 0.20$. Προφανώς, αυτό το διάγραμμα αντιστοιχεί στην εναλλακτική υπόθεση. Για την μηδενική υπόθεση, δηλαδή στην περίπτωση όπου έχουμε μόνο υπόβαθρο, δημιουργούμε με τον ίδιο τρόπο με μόνη εξαίρεση ότι τώρα θέτουμε το μ να είναι σταθερό και ίσο με το μηδέν.

```

// Set signal fraction to be a constant and specifically 0
mu.setVal(0);
mu.setConstant(kTRUE);

model.fitTo(*Data, Save(kTRUE), Minos(kFALSE), Hesse(kFALSE),
    PrintLevel(-1));

```

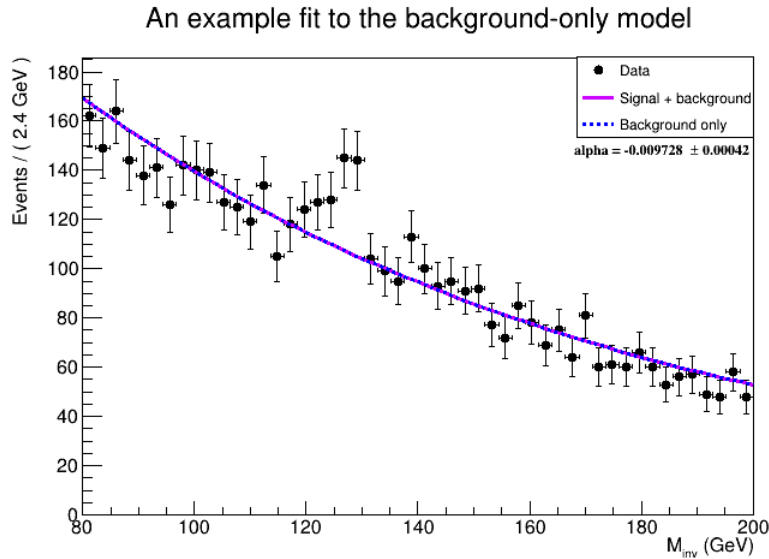
```

// Plot signal candidates with background model and components
TCanvas *bmodel_can = new TCanvas();
RooPlot *bmodel_plot = invMass.frame();
Data->plotOn(bmodel_plot, Name("data"), DataError(RooAbsData::SumW2));
model.plotOn(bmodel_plot, Name("model"), LineColor(kViolet));
model.plotOn(bmodel_plot, Name("background"), Components(background),
    LineStyle(kDashed), LineColor(kBlue));
model.paramOn(bmodel_plot);

//Create Legend
TLegend *leg2 = new TLegend(0.65,0.73,0.86,0.87);
leg2->SetFillColor(kWhite);
leg2->SetLineColor(kBlack);
leg2->AddEntry(bmodel_plot->findObject("data"), "Data", "P");
leg2->AddEntry(bmodel_plot->findObject("model"), "Signal + background", "L");
leg2->AddEntry(bmodel_plot->findObject("background"), "Background only",
    "L");

//Draw components and save as .png
bmodel_plot->SetTitle("An example fit to the background-only model");
bmodel_plot->Draw();
bmodel_can->Draw();
leg2->Draw();
bmodel_can->SaveAs("backgroundonly.png");

```



Σχήμα 9: Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα για την μηδενική υπόθεση

Είναι εμφανές ότι με βάση τα δεδομένα που έχουμε δημιουργήσει θα πρέπει να απορρίψουμε την μηδενική υπόθεση. Προφανώς το αποτέλεσμα αλλάζει ανάλογα με τις παραμέτρους που εισάγουμε. Θα δούμε στην συνέχεια, κάνοντας τον στατιστικό έλεγχο αν το συμπέρασμα που βγάζουμε από τα διαγράμματα είναι σωστό και κυρίως ποια είναι η p-value.

Φυσικά, όπως και με τις μεμονομένες κατανομές, μπορούμε να δημιουργήσουμε ένα μοντέλο προσομοίωσης και για την περίπτωση όπου έχουμε binned δεδομένα.

```

//Toy MC generation - DataHist (binned)
RooDataHist* DataHist = Data->binnedClone() ;

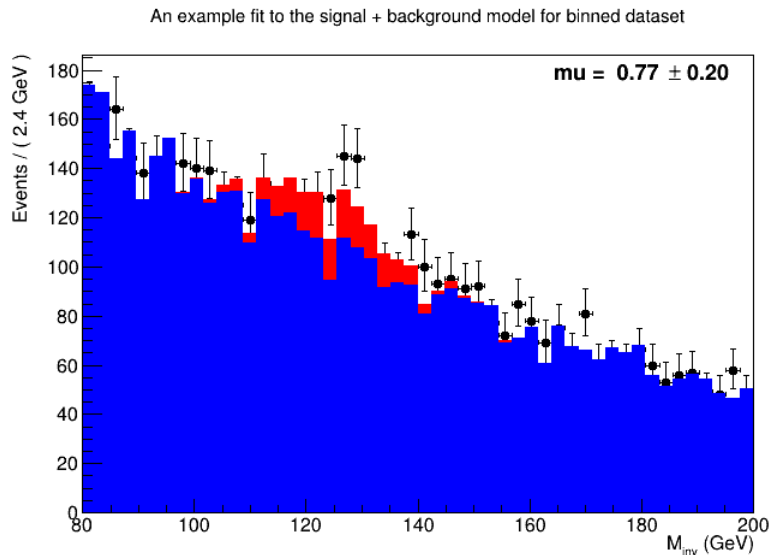
//Create binned_model
RooAddPdf binned_model("binned model","binned
    model",RooArgList(sigHistPdf,bkgHistPdf),fsig);

//Create a ROOT Canvas
TCanvas *model_binned_can = new TCanvas();

//Create a plot of the generated data superimposed to the model
RooPlot *model_binned_plot = invMass.frame();
binned_model.plotOn(model_binned_plot, RooFit::DrawOption("F"),
    RooFit::FillColor(kRed), RooFit::FillStyle(1001));
binned_model.plotOn(model_binned_plot, RooFit::Components(bkgHistPdf),
    RooFit::DrawOption("F"), RooFit::LineStyle(2) ,
    RooFit::FillStyle(1001), RooFit::FillColor(kBlue));
binned_model.paramOn(model_binned_plot);

//Draw components and save as a .png
model_binned_plot->Draw();
model_binned_can->Draw();
model_binned_can->SaveAs("model_binned.png");

```



Σχήμα 10: Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα για την εναλλακτική υπόθεση στην μορφή ιστογράμματος

Τέλος εισάγουμε το μοντέλο προσομοίωσης και τα δεδομένα σε ένα RooWorkspace έτσι ώστε να μπορούμε να τα χρησιμοποιήσουμε στην συνέχεια στον έλεγχο υποθέσης.

```

wks->import(model);
wks->import(*Data, Rename("data"));

```

3.2.4 Έλεγχος υπόθεσης

Έχοντας δημιουργήσει τα δεδομένα και το μοντέλο μας, μπορούμε να κάνουμε τον έλεγχο της υπόθεσης. Αρχικά, πρέπει να εισάγουμε από το workspace το μοντέλο προσομοίωσης μας με το οποίο θα κάνουμε τον έλεγχο.

```
//Set up ModelConfig
ModelConfig model;
model.SetWorkspace(*wks);
model.SetPdf("model");
```

Για την πραγματοποίηση του ελέγχου υπόθεσης θα χρησιμοποιήσουμε την κλάση “ProfileLikelihoodCalculator” [16] της βιβλιοθήκης RooStats. Αυτό το εργαλείο όπως μαρτυρά και το όνομά του χρησιμοποιεί την “profile likelihood” ως την στατιστική ελέγχου και υποθέτει ότι ισχύει το θεώρημα του Wilks. Προφανώς, πέρα από το μοντέλο προσομοίωσης, χρειάζεται να εισάγουμε και τα δεδομένα στα οποία θα κάνουμε τον έλεγχο.

```
// Use a RooStats ProfileLikelihoodCalculator to do the hypothesis test.
ProfileLikelihoodCalculator plc;
plc.SetData(*(wks->data("data")));
plc.SetModel(model);
```

Έχοντας εισάγει το μοντέλο και τα δεδομένα αυτό που ακόμα απαιτείται είναι ο προσδιορισμός της παραμέτρου ενδιαφέροντος. Στο συγκεκριμένο παράδειγμα η παράμετρος που μας ενδιαφέρει, όπως έχουμε ήδη αναφέρει, είναι το σθένος σήματος, μ . Προφανώς, επειδή κάνουμε έλεγχο της μηδενικής υπόθεσης στην οποία δεν έχουμε σήμα, πρέπει να θέσουμε ρητά την τιμή της παραμέτρου να είναι ίση με το μηδέν. Οι υπόλοιπες παράμετροι θέτονται να είναι nuisance.

```
// Load from workspace the mu and set it as poi
RooRealVar *mu = wks->var("mu");
RooArgSet poi(*mu);

// Here we explicitly set the value of the parameters for the null.
// We want no signal contribution, mu = 0
RooArgSet *nullParams = (RooArgSet *)poi.snapshot();
nullParams->setRealValue("mu", 0);

// Set the other parameters as nuisance
plc.SetNullParameters(*nullParams);
```

Τέλος, χρησιμοποιώντας την κλάση “HypoTestResult” [17], μπορούμε πολύ εύκολα να πάρουμε τα αποτελέσματα από το εργαλείο “ProfileLikelihoodCalculator” και να υπολογίσουμε την σημαντικότητα και την p-value.

```
// Get the result of the hypothesis test from the calculator.
HypoTestResult *htr = plc.GetHypoTest();

//Calculate p-value and significance
double p-value = htr->NullPValue()
double Significance = htr->Significance();
```

Τελικά, τα αποτελέσματα που παίρνουμε από τον έλεγχο υπόθεσης είναι:

```
-----  
The p-value for the null hypothesis is 3.95092e-07  
Which corresponds to a significance of 4.93778 sigma  
-----
```

Σχήμα 11: Αποτελέσματα ελέγχου υπόθεσης

Όπως βλέπουμε από τα αποτελέσματα, πρέπει να απορρίψουμε την μηδενική υπόθεση, δηλαδή την υπόθεση ότι έχουμε μόνο υπόβαθρο και να δεχτούμε την εναλλακτική. Μάλιστα, η σημαντικότητα είναι τόσο μεγάλη που οριακά έχουμε “ανακάλυψη”.

3.3 Διαστήματα Εμπιστοσύνης

Ως δεύτερο παράδειγμα θα μελετήσουμε και πάλι το κανάλι διάσπασης του Higgs σε δύο φωτόνια αλλά με διαφορετικό τρόπο. Συγκεκριμένα, τώρα θα δώσουμε έμφαση περισσότερο στον έλεγχο υποθέσεων όπου και θα χρησιμοποιήσουμε διάφορες κλάσεις που μας παρέχει η βιβλιοθήκη RooStats με σκοπό να βγάλουμε τα διαγράμματα για τα διαστήματα εμπιστοσύνης (Confidence Levels).

3.3.1 Δημιουργία μοντέλου προσομοίωσης

Όπως και στο προηγούμενο παράδειγμα έτσι και εδώ θα δημιουργήσουμε το μοντέλο μας με βάση μία Γκαουσιανή κατανομή ως το σήμα και μία εκθετική κατανομή ως το υπόβαθρο. Βέβαια, τώρα θα φτιάξουμε τις κατανομές κατευθείαν στο RooWorkspace έτσι ώστε να μπορούμε να τα χρησιμοποιήσουμε στην συνέχεια για τον έλεγχο της υπόθεσης. Μάλιστα, διακριτοποιούμε το πρόβλημα σε δύο διαφορετικά αρχεία, ένα για την δημιουργία των δεδομένων και ένα για τον έλεγχο υπόθεσης για μεγαλύτερη ευκολία και γενίκευση.

Ειδικότερα για το παρόν παράδειγμα θα θεωρήσουμε ότι ο αριθμός των καταγεγραμμένων γεγονότων αποτελεί και αυτός μία τυχαία μεταβλητή. Αυτό έχει ως συνέπεια την δημιουργία του “εκτεταμένου” (“extended”) μοντέλου. Μάλιστα, θα κάνουμε την ανάλυση μας θεωρώντας ως παράμετρο ενδιαφέροντος τον αριθμό γεγονότων σήματος. Επομένως, αρχικά, για την δημιουργία του μοντέλου μας θα καθορίσουμε τον αριθμό γεγονότων σήματος και υποβάθρου.

```
//Set the number of signal and background events  
int nsig = 100;  
int nbkg = 1000;
```

Στην συνέχεια, ορίζουμε το RooWorkspace που θα χρησιμοποιήσουμε και δημιουργούμε σε αυτό τις κατανομές σήματος και υποβάθρου καθώς και του μοντέλου.

```
RooWorkspace w("w");  
w.factory("Exponential:bkg_pdf(x[80,200], a[-0.01,-0.2,0.01])");  
w.factory("Gaussian:sig_pdf(x, mass[125], sigma[10])");  
  
//Create extended model  
w.factory("SUM:model(nsig[0,10000]*sig_pdf, nbkg[0,10000]*bkg_pdf)");
```

Προφανώς, έχουμε χρησιμοποιήσει ακριβώς τις ίδιες τιμές για τις διάφορες παραμέτρους των κατανομών. Επίσης, έχουμε ορίσει αυστηρά την μέση τιμή του σήματος και την διασπορά του

να είναι σταθερές ενώ όπως και στο προηγούμενο παράδειγμα έχουμε αφήσει την παράμετρο α του υποβάθρου να είναι ελεύθερη, εντός ενός διαστήματος.

Στην συνέχεια εξάγουμε από το workspace την κατανομή του μοντέλου και την παρατηρούμενη μεταβλητή έτσι ώστε να δημιουργήσουμε τα δεδομένα.

```
//Extract from workspace the pdf and the observable variable
RooAbsPdf * pdf = w.pdf("model");
RooRealVar * x = w.var("x");

// Fix number of bins to 50
x->setBins(50);

//Generate the data
RooDataSet * data = pdf->generate( *x);
```

Όπως μπορεί κανείς να παρατηρήσει, σε αντίθεση με το πρώτο παράδειγμα, κατά την δημιουργία των δεδομένων δεν καθορίζουμε τον αριθμό αυτών καθώς τώρα θα παραχθούν με βάση τον συνολικό αριθμό των γεγονότων σήματος και υποβάθρου που ορίσαμε στην αρχή.

Έχοντας δημιουργήσει τα δεδομένα, τα εισάγουμε στο RooWorkspace:

```
data->SetName("data");
w.import(*data);
```

Έτσι, έχουμε όλα τα απαραίτητα στοιχεία που απαιτούνται για τον έλεγχο υπόθεσης. Βέβαια, για να μπορέσουμε να χρησιμοποιήσουμε αυτό το μοντέλο για τον έλεγχο υπόθεσης, πρέπει να τα εισάγουμε όλα σε ένα μοντέλο χρησιμοποιώντας την κλάση ModelConfig [18] της βιβλιοθήκης RooStats. Συγκεκριμένα, αφού ορίσουμε το ModelConfig, εισάγουμε την συνάρτηση πυκνότητας πιθανότητας του μοντέλου, την παραμέτρο ενδιαφέροντος, την παρατηρούμενη μεταβλητή καθώς και ορίζουμε τις υπόλοιπες μεταβλητές ως nuisance.

```
//Create the ModelConfig in order to use later
RooStats::ModelConfig mc("ModelConfig",&w);
mc.SetPdf(*pdf);
mc.SetParametersOfInterest(*w.var("nsig"));
mc.SetObservables(*w.var("x"));
// Define set of nuisance parameters
w.defineSet("nuisParams", "a,nbkg");

mc.SetNuisanceParameters(*w.set("nuisParams"));
```

Τέλος, επειδή όπως αναφέραμε έχουμε διακριτοποιήσει το πρόβλημα σε δύο μέρη, πρέπει να εισάγουμε το ModelConfig που δημιουργήσαμε στο RooWorkspace και αυτό να το αποθηκεύσουμε σε ένα “.root” αρχείο ώστε να μπορέσουμε να το χρησιμοποιήσουμε στην συνέχεια.

```
// Import model in the workspace
w.import(mc);

// Write the workspace in the file
TString fileName = "GausExpModel.root";
w.writeToFile(fileName,true);
cout << "model written to file " << fileName << endl;
```

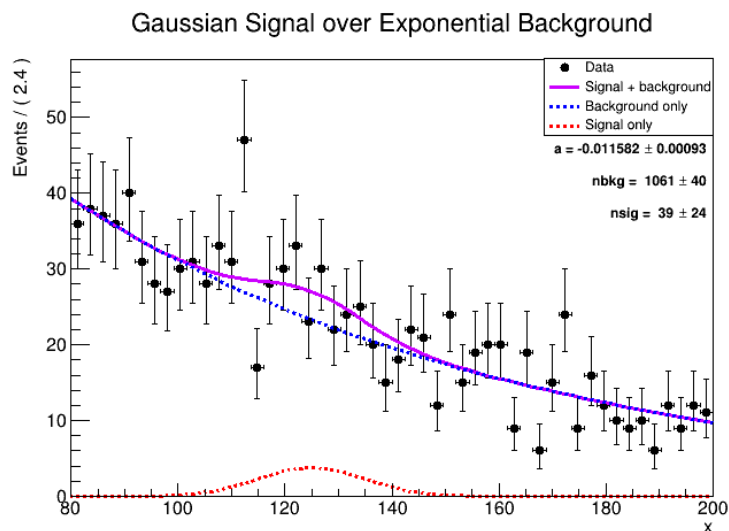

Φυσικά, μπορούμε να δημιουργήσουμε ένα διάγραμμα για να δούμε την προσαρμογή του μοντέλου στα δεδομένα που παράγαμε ώστε να κάνουμε έναν οπτικό έλεγχο πριν προχωρήσουμε στον έλεγχο της υπόθεσης.

```
//Create Canvas
TCanvas *can = new TCanvas();
RooPlot *plot = x->frame(Title("Gaussian Signal over Exponential
    Background"));
data->plotOn(plot, Name("data"));
plot->Draw();

RooFitResult * r = pdf->fitTo(*data, RooFit::Save(true),
    RooFit::Minimizer("Minuit2","Migrad"));

pdf->plotOn(plot, Name("model"), RooFit::LineColor(kViolet));
pdf->plotOn(plot, Name("background"), RooFit::Components("bkg_pdf"),
    RooFit::LineColor(kBlue), RooFit::LineStyle(kDashed) );
pdf->plotOn(plot, Name("signal only"), RooFit::Components("sig_pdf"),
    RooFit::LineColor(kRed), RooFit::LineStyle(kDashed) );
pdf->paramOn(plot,Layout(0.5,0.9,0.85));
//Create Legend
TLegend *leg = new TLegend(0.65,0.73,0.86,0.87);
leg->SetFillColor(kWhite);
leg->SetLineColor(kBlack);
leg->AddEntry(plot->findObject("data"), "Data", "P");
leg->AddEntry(plot->findObject("model"),"Signal + background","L");
leg->AddEntry(plot->findObject("background"), "Background only", "L");
leg->AddEntry(plot->findObject("signal only"), "Signal only", "L");

plot->Draw();
can->Draw();
leg->Draw();
can->SaveAs("Signal+background_model.png");
```



Σχήμα 12: Προσαρμογή του μοντέλου προσομοίωσης στα δεδομένα

Όπως βλέπουμε από το Σχήμα 12, σε αντίθεση με το προηγούμενο παράδειγμα δεν είναι άμεσα εμφανής η ύπαρξη του σήματος. Αυτό οφείλεται κυρίως στον μικρότερο αριθμό γεγονότων που επιλέχθηκε εσκεμμένως για το συγκεκριμένο παράδειγμα με ακριβώς αυτό τον σκοπό. Επομένως, δεν μπορούμε να προβλέψουμε a priori το αποτέλεσμα αλλά θα το δούμε στην συνέχεια πραγματοποιώντας τον στατιστικό έλεγχο.

3.3.2 Έλεγχος υπόθεσης

Αρχικά, για να μπορέσουμε να κάνουμε την ανάλυση μας, χρειάζεται να διαβάσουμε τα απαραίτητα στοιχεία από το αρχείο που δημιουργήσαμε. Για να είναι εύκολη η χρήση του κώδικα και για άλλα παραδείγματα, ορίζουμε ως ορίσματα τις μεταβλητές που χρειαζόμαστε από το αρχείο.

```
void HypothesisTest( const char* filename = "HiggsModel.root",
                    const char* workspaceName = "w",
                    const char* modelConfigName = "ModelConfig",
                    const char* dataName = "data" )
```

Στην συνέχεια, ανοίγουμε το αρχείο με τα δεδομένα μας, δημιουργούμε ένα καινούργιο RooWorkspace με βάση αυτό που υπάρχει στο αρχείο και εξάγουμε και τα δεδομένα.

```
// Open input file
TFile *file = TFile::Open(filename);
if (!file) return;

// Get the workspace out of the file
RooWorkspace* w = (RooWorkspace*) file->Get(workspaceName);

// Get the data out of the file
RooAbsData* data = w->data(dataName);
```

Επίσης, δημιουργούμε ένα καινούργιο ModelConfig με βάση αυτό που ήδη έχουμε. Μάλιστα, ορίζουμε και την παράμετρο ενδιαφέροντος καθώς και καθορίζουμε μία τιμή της.

```
// Get the ModelConfig out of the file
ModelConfig* sbModel = (RooStats::ModelConfig*) w->obj(modelConfigName);
sbModel->SetName("S+B Model");
RooRealVar* poi = (RooRealVar*)
    sbModel->GetParametersOfInterest()->first();
poi->setVal(50); // set POI snapshot in S+B model for expected significance
sbModel->SetSnapshot(*poi);
```

Προφανώς, αυτό το μοντέλο είναι για την εναλλακτική υπόθεση. Θα πρέπει να δημιουργήσουμε και ένα για την μηδενική υπόθεση. Φυσικά το μόνο που χρειάζεται να αλλάξει είναι η τιμή της παραμέτρου ενδιαφέροντος.

```
// Create the Background only model from the S+B model
ModelConfig * bModel = (ModelConfig*) sbModel->Clone();
bModel->SetName("B Model");
poi->setVal(0);
bModel->SetSnapshot( *poi );
```

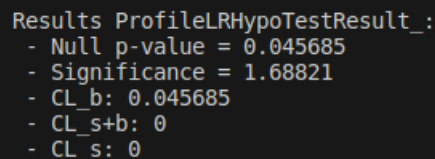
Έχοντας εξάγει όλα τα δεδομένα που χρειαζόμαστε από το αρχείο καθώς και δημιουργήσει τα απαραίτητα μοντέλα είμαστε έτοιμοι να κάνουμε τον έλεγχο υπόθεσης. Αρχικά, θα χρησιμοποιήσουμε το “ProfileLikelihoodCalculator” το οποίο χρησιμοποιήσαμε και στο προηγούμενο παράδειγμα. Επειδή τώρα έχουμε ορίσει από την αρχή το ModelConfig, το μόνο που χρειάζεται για να κάνουμε τον έλεγχο είναι η ακόλουθη γραμμή κώδικα.

```
ProfileLikelihoodCalculator plc(*data, *bModel);
```

Να επισημάνουμε ότι κάνουμε έλεγχο της μηδενικής υπόθεσης για αυτό και στο όρισμα της κλάσης χρησιμοποιούμε το *bModel*. Για να πάρουμε τα αποτελέσματα από το μοντέλο αρκεί το ακόλουθο.

```
HypoTestResult *plcResult = plc.GetHypoTest();  
plcResult->Print();
```

όπου η εντολή *Print()* εκτυπώνει απευθείας τα αποτελέσματα στο τερματικό (terminal). Ειδικότερα, το αποτέλεσμα που παίρνουμε είναι:



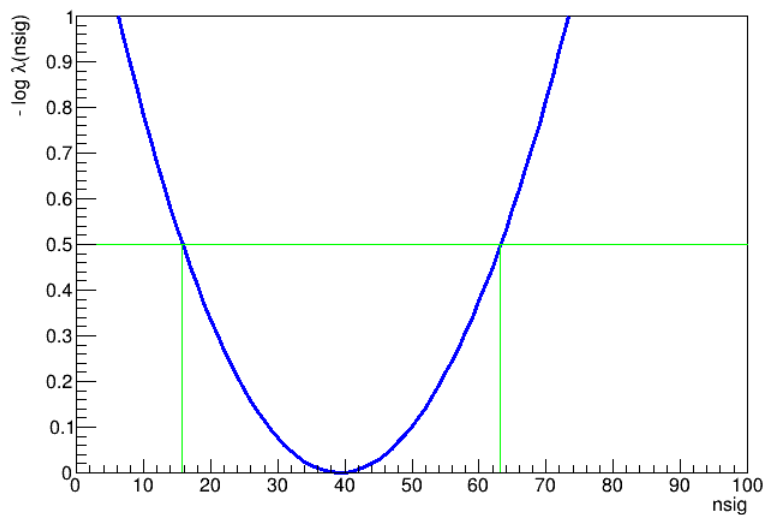
```
Results ProfileLRHypoTestResult_  
- Null p-value = 0.045685  
- Significance = 1.68821  
- CL_b: 0.045685  
- CL_s+b: 0  
- CL_s: 0
```

Σχήμα 13: Αποτελέσματα ελέγχου υπόθεσης (ProfileLikelihoodCalculator)

Όπως αναμέναμε, η τιμή για την p-value είναι μεγαλύτερη από ότι στο πρώτο παράδειγμα και κατ'επέκταση η σημαντικότητα είναι μικρότερη, μόλις 1.6 σ.

Αυτό που μπορούμε να δούμε σχηματικά από αυτήν την κλάση είναι που ελαχιστοποιείται ο αρνητικός λογάριθμος της πιθανοφάνειας για την παράμετρο ενδιαφέροντος. Συγκεκριμένα, έχουμε:

```
plc.SetConfidenceLevel(0.683);  
  
LikelihoodInterval * plcinterval = plc.GetInterval();  
  
TCanvas *plccan = new TCanvas();  
LikelihoodIntervalPlot plcplot(plcinterval);  
plcplot.SetRange(0,100);  
plcplot.Draw();  
plccan->Draw();  
plccan->SaveAs("Negative_logarithm_of_the_profile _likelihood.png");
```



Σχήμα 14: Κατανομή της στατιστικής ελέγχου συναρτήσεως της παραμέτρου ενδιαφέροντος

Παρατηρούμε ότι η στατιστική ελέγχου παρουσιάζει ελάχιστο όταν η παράμετρος ενδιαφέροντος είναι ίση περίπου με 39. Αυτό έρχεται σε απόλυτη συμφωνία με την προσαρμογή του μοντέλου που δημιουργήσαμε στα δεδομένα, όπως φαίνεται από το Σχήμα 12. Να σημειώσουμε ότι με τις πράσινες γραμμές εμφανίζονται τα διαστήματα εμπιστοσύνης τα οποία στην προκειμένη περίπτωση έχουν τεθεί να είναι στο $\pm 1 \sigma$.

Φυσικά η βιβλιοθήκη RooStats έχει και άλλες κλάσεις οι οποίες πραγματοποιούν στατιστικό έλεγχο υπόθεσης πέραν της “ProfileLikelihoodCalculator”. Μία από αυτές είναι η “AsymptoticCalculator” [19] η οποία βασίζεται στην ασυμπτωτική μορφή της στατιστικής ελέγχου profile likelihood. Για να την εφαρμόσουμε στα δεδομένα μας, απλώς γράφουμε:

```
AsymptoticCalculator ac(*data, *sbModel, *bModel);
```

και για να πάρουμε τα αποτελέσματα χρησιμοποιούμε τις ίδιες εντολές με πριν

```
HypoTestResult * asResult = ac.GetHypoTest();
asResult->Print();
```

Το αποτέλεσμα που παίρνουμε είναι:

```
Results HypoTestAsymptotic_result:
- Null p-value = 0.0456852
- Significance = 1.68821
- CL_b: 0.0456852
- CL_s+b: 0.678116
- CL_s: 14.8432
```

Σχήμα 15: Αποτελέσματα ελέγχου υπόθεσης (AsymptoticCalculator)

Όπως είναι εμφανές από την παραπάνω εικόνα, το αποτέλεσμα σε σύγκριση με την κλάση “ProfileLikelihoodCalculator” (Σχήμα 13) είναι πρακτικά το ίδιο. Αυτό είναι πλήρως αναμενόμενο αφού και τα δύο χρησιμοποιούν την ίδια στατιστική ελέγχου καθώς και βασίζονται στο θεώρημα Wilks. Η μόνη διαφορά έγκειται στο γεγονός ότι η κλάση “AsymptoticCalculator” πραγματοποιεί έλεγχο και στην εναλλακτική υπόθεση. Επίσης, σε αυτήν την κλάση

μπορούμε να χρησιμοποιήσουμε ένα άλλο εργαλείο που ονομάζεται “HypoTestInverter” [20]. Πιο συγκεκριμένα έχουμε:

```
//HypoTestInverter
HypoTestInverter acinverter(ac);

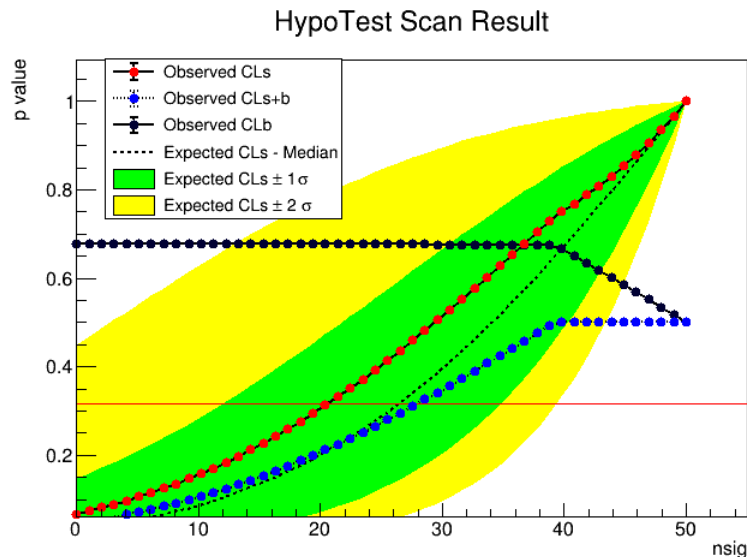
//Statistical configuration of hypothesis test inverter
acinverter.SetConfidenceLevel(0.683);
acinverter.UseCLs(true);

//Technical configuration of hypothesis test inverter
acinverter.SetVerbose(false);
acinverter.SetFixedScan(50,0.0,50.0); // set number of points , xmin and
    xmax

//Calculation of limit
HypoTestInverterResult* acinvresult = acinverter.GetInterval();
```

Αφού τρέξουμε το εργαλείο μας δίνεται η δυνατότητα να δημιουργήσουμε ένα διάγραμμα με τα διαστήματα εμπιστοσύνης για τις διαφορετικές υποθέσεις.

```
//Create a CL plot
TCanvas* acinvcan = new TCanvas();
HypoTestInverterPlot* acinvplot = new
    HypoTestInverterPlot("HTI_Result_Plot","HypoTest Scan
    Result",acinvresult);
acinvplot->Draw("CLb 2CL"); // plot also CLb and CLs+b
acinvcan->SaveAs("Brazil_plot_asymptotic.png");
```



Σχήμα 16: Διαστήματα εμπιστοσύνης (AsymptoticCalculator)

Φυσικά, πέρα από τους ελέγχους υπόθεσης βάση την ασυμπτωτική συμπεριφορά της στατιστικής ελέγχου, μπορούμε να πραγματοποιήσουμε έλεγχο δημιουργώντας ψευδοπειράματα (toy MonteCarlo). Η κλάση που πραγματοποιεί ακριβώς αυτό είναι η “Frequentist-

Calculator” [21]. Όπως και στις προηγούμενες περιπτώσεις έτσι και εδώ απλώς καλούμε την κλάση με ορίσματα τα δεδομένα ενώ τώρα πρέπει να καθορίσουμε και τον αριθμό των ψευδοπειραμάτων που θέλουμε να χρησιμοποιήσουμε. Στην προκειμένη περίπτωση επιλέξαμε 500 ψευδοπειράματα για την μηδενική υπόθεση και 500 για την εναλλακτική.

```
FrequentistCalculator fc(*data, *sbModel, *bModel);  
fc.SetToys(500,500);
```

Ως στατιστική ελέγχου επιλέγουμε να είναι πάλι η profile likelihood.

```
// Create the test statistics  
ProfileLikelihoodTestStat profll(*sbModel->GetPdf());  
// Use one-sided profile likelihood  
profll.SetOneSidedDiscovery(true);
```

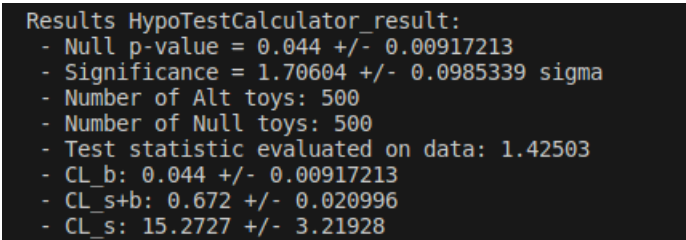
Στην συνέχεια, ρυθμίζουμε τις παραμέτρους για τα ψευδοπειράματα

```
// Configure ToyMCSampler and set the test statistics  
ToyMCSampler *toymcs = (ToyMCSampler*)fc.GetTestStatSampler();  
toymcs->SetTestStatistic(&profll);  
  
if (!sbModel->GetPdf()->canBeExtended())  
    toymcs->SetNEventsPerToy(1);
```

Προφανώς, για να δούμε το αποτέλεσμα του ελέγχου γράφουμε και πάλι.

```
HypoTestResult *fqResult = fc.GetHypoTest();  
fqResult->Print();
```

Το αποτέλεσμα που παίρνουμε είναι:



```
Results HypoTestCalculator_result:  
- Null p-value = 0.044 +/- 0.00917213  
- Significance = 1.70604 +/- 0.0985339 sigma  
- Number of Alt toys: 500  
- Number of Null toys: 500  
- Test statistic evaluated on data: 1.42503  
- CL_b: 0.044 +/- 0.00917213  
- CL_s+b: 0.672 +/- 0.020996  
- CL_s: 15.2727 +/- 3.21928
```

Σχήμα 17: Αποτελέσματα ελέγχου υπόθεσης (FrequentistCalculator)

Όπως βλέπουμε από την εικόνα, το αποτέλεσμα είναι σύμφωνο με πολύ καλή ακρίβεια με τους άλλους δύο τρόπους ανάλυσης (Σχήματα 13,15). Παρατηρούμε, ότι ο τρόπος αυτός είναι λιγότερο ακριβής σε σύγκριση με τους άλλες δύο. Αυτό οφείλεται στην επιλογή του μικρού αριθμού των ψευδοπειραμάτων η οποία έγινε για λόγους οικονομίας υπολογιστικών πόρων.

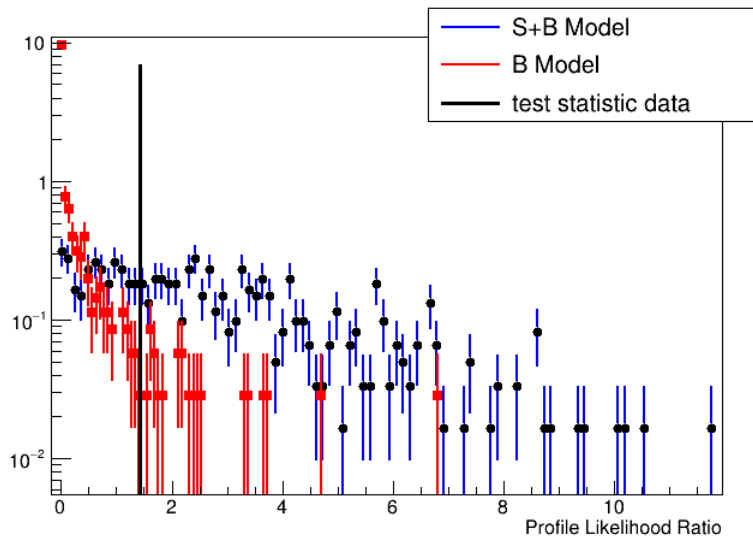
Επιπλέον, σε αυτή την περίπτωση, μπορούμε να σχεδιάσουμε την κατανομή των τιμών του στατιστικού ελέγχου για τις δύο υποθέσεις για κάθε ένα ψευδοπείραμα που πραγματοποιήθηκε.

```
// plot test statistic distributions  
TCanvas *can = new TCanvas();  
HypoTestPlot *plot = new HypoTestPlot(*fqResult);
```

```

plot->SetLogYaxis(true);
plot->Draw();
can->Draw();
can->SaveAs("test_statistic_distributions.png");

```



Σχήμα 18: Κατανομή της στατιστικής ελέγχου για τις δύο υποθέσεις.

Όπως αναφέραμε στην ενότητα 2.8, η p-value και συνεπώς και η σημαντικότητα υπολογίζονται με βάση τις κατανομές που παρουσιάζονται στο Σχήμα 18. Συγκεκριμένα, η p-value υπολογίζεται από την περιοχή που καθορίζεται από την τιμή της στατιστικής ελέγχου από τα δεδομένα, όπου στο παραπάνω σχήμα συμβολίζεται με μία μαύρη κάθετη γραμμή.

Φυσικά, και σε αυτήν την κλάση μπορούμε να χρησιμοποιήσουμε το εργαλείο “HypoTestInverter”.

```

//HypoTestInverter
HypoTestInverter fcinverter(fc);

//Statistical configuration of hypothesis test inverter
fcinverter.SetConfidenceLevel(0.683);
fcinverter.UseCLs(true);

//Technical configuration of hypothesis test inverter
fcinverter.SetVerbose(false);
fcinverter.SetFixedScan(50,0.0,50.0); // set number of points , xmin and
    xmax

//Calculation of limit
HypoTestInverterResult* fcinvresult = fcinverter.GetInterval();

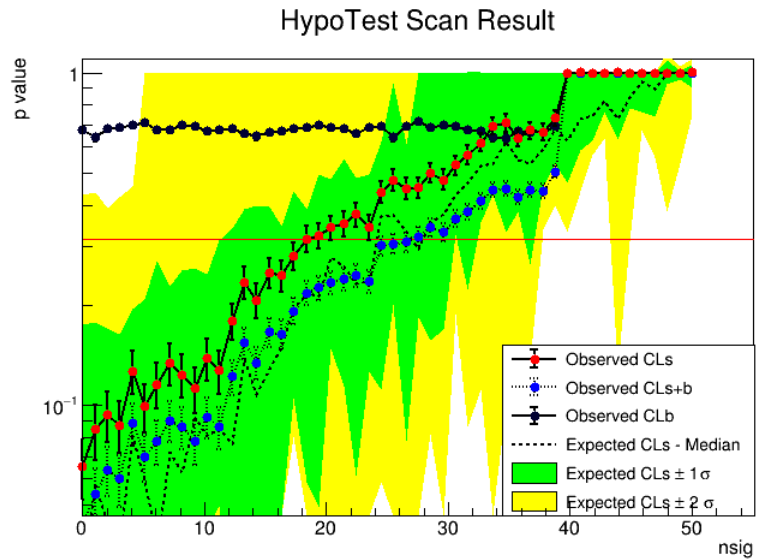
```

Τέλος, μπορούμε και πάλι να παραστήσουμε γραφικά το διάγραμμα με τα διαστήματα εμπιστοσύνης.

```

TCanvas* fcinvcan = new TCanvas();
HypoTestInverterPlot* fcinvplot = new
    HypoTestInverterPlot("HTI_Result_Plot","HypoTest Scan
        Result",fcinvresult);
fcinvplot->Draw("CLb 2CL"); // plot also CLb and CLs+b
fcinvcan->Draw();
fcinvcan->SaveAs("Brazil_plot_frequentist.png");

```



Σχήμα 19: Διαστήματα εμπιστοσύνης (FrequentistCalculator)

4 Βιβλιογραφία

- [1] E. Gross. *Practical statistics for High Energy Physics*. Dec. 2018. URL: <https://doi.org/10.23730/CYRSP-2018-003.199>.
- [2] Χαράλαμπος Δαμιανού, Νικόλαος Παπαδάτος, and Χαράλαμπος Χααραλαμπίδης. *Εισαγωγή στις Πιθανότητες και τη Στατιστική*. Εκδόσεις Συμμετρία, 2010. ISBN: 978-960-266-308-0.
- [3] In: *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231.694-706 (1933), pp. 289–337. DOI: [10.1098/rsta.1933.0009](https://doi.org/10.1098/rsta.1933.0009).
- [4] C Peterson and T S Rönvaldsson. “An introduction to artificial neural networks”. In: (1992). DOI: [10.5170/CERN-1992-002.113](https://doi.org/10.5170/CERN-1992-002.113). URL: <https://cds.cern.ch/record/231036>.
- [5] Byron P. Roe et al. “Boosted decision trees as an alternative to artificial neural networks for particle identification”. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 543.2-3 (May 2005), pp. 577–584. DOI: [10.1016/j.nima.2004.12.018](https://doi.org/10.1016/j.nima.2004.12.018). URL: <https://doi.org/10.1016%2Fj.nima.2004.12.018>.
- [6] S. S. Wilks. “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses”. In: *The Annals of Mathematical Statistics* 9.1 (1938), pp. 60–62. ISSN: 00034851. URL: <http://www.jstor.org/stable/2957648> (visited on 08/04/2023).
- [7] Luca Lista. *Statistical methods for data analysis in particle physics*. 2nd ed. Springer, 2017. ISBN: 978-3-319-62839-4. DOI: <https://doi.org/10.1007/978-3-319-62840-0>.
- [8] Olaf Behnke et al. *Data Analysis in high energy physics: A practical guide to statistical methods*. Wile, 2013. ISBN: 9783527410583.
- [9] Glen Cowan et al. “Asymptotic formulae for likelihood-based tests of new physics”. In: *The European Physical Journal C* 71.2 (Feb. 2011). DOI: [10.1140/epjc/s10052-011-1554-0](https://doi.org/10.1140/epjc/s10052-011-1554-0). URL: <https://doi.org/10.1140%2Fepjc%2Fs10052-011-1554-0>.
- [10] Rene Brun et al. *root-project/root: v6.18/02*. Version v6-18-02. Aug. 2019. DOI: [10.5281/zenodo.3895860](https://doi.org/10.5281/zenodo.3895860). URL: <https://doi.org/10.5281/zenodo.3895860>.
- [11] Wouter Verkerke and David P. Kirkby. “The RooFit toolkit for data modeling”. In: *eConf C0303241* (2003). Ed. by L. Lyons and Muge Karagoz, MOLT007. arXiv: [physics/0306116](https://arxiv.org/abs/physics/0306116).
- [12] Gregory Schott. “RooStats for Searches”. In: (2011). Comments: Contributed to "PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, pp. 199–208. DOI: [10.5170/CERN-2011-006.199](https://doi.org/10.5170/CERN-2011-006.199). arXiv: [1203.1547](https://arxiv.org/abs/1203.1547). URL: <https://cds.cern.ch/record/1430313>.
- [13] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. DOI: [10.1016/j.physletb.2012.08.020](https://doi.org/10.1016/j.physletb.2012.08.020). URL: <https://doi.org/10.1016%2Fj.physletb.2012.08.020>.

- [14] *RooExponential Class Reference*. URL: <https://root.cern.ch/doc/master/classRooExponential.html>.
- [15] *RooGaussian Class Reference*. URL: <https://root.cern.ch/doc/master/classRooGaussian.html>.
- [16] *RooStats::ProfileLikelihoodCalculator Class Reference*. URL: https://root.cern.ch/doc/master/classRooStats_1_1ProfileLikelihoodCalculator.html.
- [17] *RooStats::HypoTestResult Class Reference*. URL: https://root.cern.ch/doc/master/classRooStats_1_1HypoTestResult.html.
- [18] *RooStats::ModelConfig Class Reference*. URL: https://root.cern.ch/doc/master/classRooStats_1_1ModelConfig.html.
- [19] *RooStats::AsymptoticCalculator Class Reference*. URL: https://root.cern.ch/doc/master/classRooStats_1_1AsymptoticCalculator.html.
- [20] *RooStats::HypoTestInverter Class Reference*. URL: https://root.cern.ch/doc/master/classRooStats_1_1HypoTestInverter.html.
- [21] *RooStats::FrequentistCalculator Class Reference*. URL: https://root.cern.ch/doc/master/classRooStats_1_1FrequentistCalculator.html.
- [22] Gerhard Bohm and Günter Zech. *Introduction to statistics and data analysis for physicists*. 3rd ed. DESY, 2010. ISBN: 978-3-945931-13-4. DOI: [10.3204/PUBDB-2017-08987](https://doi.org/10.3204/PUBDB-2017-08987).