

# Περιεχόμενα

0.1	Ευχαριστίες - Αφιερώσεις . . . . .	5
<b>1</b>	<b>Εισαγωγή στην Στατιστική κατά Bayes</b> . . . . .	<b>1</b>
1.1	Μπευζιανή συμπερασματολογία . . . . .	1
1.2	Εκ των προτέρων κατανομές . . . . .	2
1.2.1	Συζυγείς κατανομές . . . . .	3
1.2.2	Η μη πληροφοριακή εκ των προτέρων του <b>Jeffreys</b> . . . . .	8
1.2.3	Υπέρ εκ των προτέρων κατανομές . . . . .	11
1.2.4	Εκ των προτέρων κατανομές βασισμένες σε δυνάμεις της πιθανοφάνειας . . . . .	13
<b>2</b>	<b>Έλεγχος Υποθέσεων για Ένα Δείγμα</b> . . . . .	<b>17</b>
2.1	Μειονεκτήματα της Κλασικής προσέγγισης . . . . .	17
2.2	Παράγοντας <b>Bayes</b> . . . . .	19
2.3	Το παράδοξο του <b>Lindley</b> . . . . .	23
2.3.1	Παράδειγμα με προσομοιωμένα δεδομένα . . . . .	24
2.4	Έλεγχος υποθέσεων για ένα δείγμα από κατανομή <b>Poisson</b> . . . . .	26
2.4.1	Παράδειγμα με προσομοιωμένα δεδομένα . . . . .	28
2.4.2	Παράδειγμα με πραγματικά δεδομένα . . . . .	31
2.5	Έλεγχος υποθέσεων για ένα δείγμα από κατανομή <b>Binomial</b> . . . . .	34
2.5.1	Παράδειγμα με προσομοιωμένα δεδομένα . . . . .	36
2.5.2	Παράδειγμα με πραγματικά δεδομένα . . . . .	40
2.6	Επίλογος . . . . .	44
<b>3</b>	<b>Έλεγχος Υποθέσεων για Δύο Δείγματα</b> . . . . .	<b>45</b>
3.1	Έλεγχος υποθέσεων για δύο ανεξάρτητα δείγματα από κατανομή <b>Poisson</b> . . . . .	45
3.1.1	Παράδειγμα (1) με προσομοιωμένα δεδομένα . . . . .	48
3.1.2	Παράδειγμα (2) με προσομοιωμένα δεδομένα . . . . .	49
3.1.3	Παράδειγμα με πραγματικά δεδομένα . . . . .	51
3.2	Έλεγχος υποθέσεων για δύο ανεξάρτητα δείγματα από κατανομή <b>Binomial</b> . . . . .	53
3.2.1	Παράδειγμα (1) με προσομοιωμένα δεδομένα . . . . .	56
3.2.2	Παράδειγμα (2) με προσομοιωμένα δεδομένα . . . . .	58
3.2.3	Παράδειγμα με πραγματικά δεδομένα . . . . .	59
3.3	Έλεγχος υποθέσεων για δύο συσχετιζόμενα δείγματα από κατανομή <b>Poisson</b> . . . . .	61
3.4	Έλεγχος υποθέσεων για δύο συσχετιζόμενες αναλογίες . . . . .	64
3.4.1	Παράδειγμα με πραγματικά δεδομένα . . . . .	68
3.5	Επίλογος . . . . .	71

<b>4 Έλεγχος Υποθέσεων για Πολλαπλά Δείγματα</b>	<b>73</b>
4.1 Έλεγχος υποθέσεων για $k$ ανεξάρτητα δείγματα από κατανομή <b>Poisson</b>	73
4.1.1 Παράδειγμα (1) με προσομοιωμένα δεδομένα	75
4.1.2 Παράδειγμα (2) με προσομοιωμένα δεδομένα	77
4.2 Έλεγχος υποθέσεων για $k$ ανεξάρτητα δείγματα από κατανομή <b>Binomial</b>	79
4.2.1 Παράδειγμα (1) με προσομοιωμένα δεδομένα	81
4.2.2 Παράδειγμα (2) με προσομοιωμένα δεδομένα	83
4.2.3 Παράδειγμα με πραγματικά δεδομένα	85
4.3 Έλεγχος υποθέσεων για πίνακες συνάφειας	85
4.3.1 Εισαγωγή στους πίνακες συνάφειας	85
4.3.2 Έλεγχος υποθέσεων για $r \times c$ πίνακες συνάφειας	86
4.3.3 Παράδειγμα με πραγματικά δεδομένα	88
<b>5 Συζήτηση</b>	<b>91</b>
<b>A' Κατανομές</b>	<b>97</b>
<b>B' Περίληψη</b>	<b>101</b>
<b>Γ' Abstract</b>	<b>103</b>

# Κατάλογος Σχημάτων

2.1	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από $N(4, 0.81)$ . . . . .	25
2.2	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του $\log C$ , ( $1 < C < 10^6$ ) (Παράδειγμα 2.3.1). . . . .	26
2.3	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από $Poisson(0.2)$ . . . . .	28
2.4	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του $\lambda_0$ με εκ των προτέρων κατανομή $(\alpha)$ την $Gamma(1, 1)$ και $(\beta)$ την $Gamma(\lambda_0, 1)$ (Παράδειγμα 2.4.1). . . . .	31
2.5	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές $(\alpha)$ της παραμέτρου $\alpha$ της εκ των προτέρων κατανομής $Gamma(\alpha, 1)$ και $(\beta)$ της παραμέτρου $\delta \in [0, 1]$ της εκ των προτέρων κατανομής $Gamma(\delta n \lambda_0, \delta n)$ (Παράδειγμα 2.4.1). . . . .	31
2.6	Ραβδόγραμμα του πλήθους των φαρμακευτικών αγωγών που χορηγούνται με ιατρική συ-νταγή ή μη, τις τελευταίες 2 μέρες. . . . .	32
2.7	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του $\lambda_0$ με εκ των προτέρων κατανομή $(\alpha)$ την $Gamma(1, 1)$ και $(\beta)$ την $Gamma(\lambda_0, 1)$ (Παράδειγμα 2.4.2). . . . .	33
2.8	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές $(\alpha)$ της παραμέτρου $\alpha$ της εκ των προτέρων κατανομής $Gamma(\alpha, 1)$ και $(\beta)$ της παραμέτρου $\delta \in [0, 1]$ της εκ των προτέρων κατανομής $Gamma(\delta n \lambda_0, \delta n)$ (Παράδειγμα 2.4.2). . . . .	34
2.9	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από $Binomial(10, 0.5)$ . . . . .	36
2.10	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του $\pi_0$ με εκ των προτέρων κατανομές $(\alpha)$ $Beta(0.5, 0.5)$ , $(\beta)$ $Beta(1, 1)$ και $(\gamma)$ $Beta(\pi_0, 1 - \pi_0)$ (Παράδειγμα 2.5.1). . . . .	38
2.11	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes της εκ των προτέρων κατανομής $Beta(\frac{c}{2}, \frac{c}{2})$ ως προς $(\alpha)$ τις διάφορες τιμές του $c$ και $(\beta)$ της διασποράς της $Var(\pi) = \frac{1}{4(c+1)}$ (Παράδειγμα 2.5.1). . . . .	39
2.12	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes της εκ των προτέρων κατανομής $Beta(\alpha, 1 - \alpha)$ ως προς $(\alpha)$ τις διάφορες τιμές του $\alpha \in [0, 1]$ και $(\beta)$ της διασποράς της $Var(\pi) = \frac{\alpha(1-\alpha)}{2}$ (Παράδειγμα 2.5.1). . . . .	40
2.13	Επιοτbar με τα ποσοστά των ατόμων με νεφρική τοξικότητα και σχετικές παρενέργειες για κάθε αντιβιοτικό. . . . .	42
2.14	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές $(\alpha)$ της διασποράς της εκ των προτέρων κατανομής $Beta(\frac{c}{2}, \frac{c}{2})$ και $(\beta)$ της διασποράς της εκ των προτέρων κατανομής $Beta(\alpha, 1 - \alpha)$ , για το αντιβιοτικό Netilmicin (Παράδειγμα 2.5.2). . . . .	43

2.15	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές (α) της παραμέτρου $c$ και (β) της διασποράς $(Var(\pi) = \frac{\pi_o(1-\pi_o)}{c+1})$ , της εκ των προτέρων κατανομής $Beta(c\pi_o, c(1-\pi_o))$ για το αντιβιοτικό Netilmicin (υπό την δειγματική αναλογία $\pi_o = 0.082$ ) (Παράδειγμα 2.5.2). . . . .	43
3.1	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από (α) $Poisson(5)$ και (β) $Poisson(2)$ . . . . .	48
3.2	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes ως προς τις διάφορες τιμές της παραμέτρου $\alpha$ για την γενική μορφή των δύο συνδυασμών, δηλαδή $\lambda \sim Gamma(\alpha, 1)$ , $\lambda_1, \lambda_2 \sim Gamma(\frac{\alpha}{2}, \frac{1}{2})$ (Παράδειγμα 3.1.1). . . . .	49
3.3	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από (α) $Poisson(5)$ και (β) $Poisson(4.9)$ . . . . .	50
3.4	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για διαφορετικό εύρος τιμών της παραμέτρου $\alpha$ για την γενική μορφή των δύο συνδυασμών: $\lambda \sim Gamma(\alpha, 1)$ , $\lambda_1, \lambda_2 \sim Gamma(\frac{\alpha}{2}, \frac{1}{2})$ , όπου (α) $\alpha \in [0, 1000]$ και (β) $\alpha \in [10^3, 10^5]$ (Παράδειγμα 3.1.2). . . . .	51
3.5	Ραβδόγραμμα των παρατηρήσεων για (α) τους άνδρες και (β) τις γυναίκες. . . . .	52
3.6	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου $\alpha$ για την γενική μορφή των δύο συνδυασμών: $\lambda \sim Gamma(\alpha, 1)$ , $\lambda_1, \lambda_2 \sim Gamma(\frac{\alpha}{2}, \frac{1}{2})$ (Παράδειγμα 3.1.3). . . . .	53
3.7	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από $Binomial(0.5)$ . . . . .	56
3.8	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων $\alpha$ & $c$ για (α) την γενική μορφή 1: $\pi \sim Beta(\alpha, 1-\alpha)$ , $\pi_j \sim Beta(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1-\alpha))$ και (β) την γενική μορφή 2: $\pi \sim Beta(\frac{c}{2}, \frac{c}{2})$ , $\pi_j \sim Beta(\frac{c}{2}, \frac{c}{2})$ (Παράδειγμα 3.2.1). . . . .	58
3.9	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από (α) $Binomial(0.3)$ και (β) $Binomial(0.7)$ . . . . .	59
3.10	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων $\alpha$ & $c$ για (α) την γενική μορφή 1: $\pi \sim Beta(\alpha, 1-\alpha)$ , $\pi_j \sim Beta(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1-\alpha))$ και (β) την γενική μορφή 2: $\pi \sim Beta(\frac{c}{2}, \frac{c}{2})$ , $\pi_j \sim Beta(\frac{c}{2}, \frac{c}{2})$ (Παράδειγμα 3.2.2). . . . .	60
3.11	Errorbar με τα ποσοστά των ατόμων με νεφρική τοξικότητα και σχετικές παρενέργειες για κάθε αντιβιοτικό (Παράδειγμα 3.2.3). . . . .	61
3.12	Πίνακας (α) συχνοτήτων και (β) πιθανοτήτων των παραγόντων A & B. . . . .	65
3.13	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές (α) του $q$ και (β) του $k$ (Παράδειγμα 3.4.1). . . . .	70
3.14	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes της εκ των προτέρων κατανομής $Beta(\alpha, 1-\alpha)$ ως προς (α) τις διάφορες τιμές της $\alpha \in [0, 1]$ και (β) της διασποράς της (Παράδειγμα 3.4.1). . . . .	71
4.1	Ραβδογράμματα των προσομοιωμένων παρατηρήσεων από κατανομή $Poisson$ . . . . .	76
4.2	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του $\alpha$ για την γενική μορφή: $\lambda \sim Gamma(\alpha, 1)$ , $\lambda_1, \lambda_2, \lambda_3 \sim Gamma(\frac{\alpha}{3}, \frac{1}{3})$ (Παράδειγμα 4.1.1). . . . .	77
4.3	Ραβδογράμματα των προσομοιωμένων παρατηρήσεων από κατανομή $Poisson$ . . . . .	77
4.4	Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του $\alpha$ για την γενική μορφή: $\lambda \sim Gamma(\alpha, 1)$ , $\lambda_1, \lambda_2, \lambda_3 \sim Gamma(\frac{\alpha}{3}, \frac{1}{3})$ (Παράδειγμα 4.1.2). . . . .	78
4.5	Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από κατανομή $Binomial$ . . . . .	82

- 4.6 Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\alpha$  για (α) την γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$  και (β) την γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$  (Παράδειγμα 4.2.1). . . . . 83
- 4.7 Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από κατανομή *Binomial*. . . . . 83
- 4.8 Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\alpha$  για (α) την γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$  και (β) την γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$  (Παράδειγμα 4.2.2). . . . . 84



# Κατάλογος Πινάκων

2.1	Ερμηνεία του παράγοντα Bayes σύμφωνα με τον Jeffrey (1961).	22
2.2	Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass & Raftery (1995).	22
2.3	Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την $Gamma(1, 1)$ .	29
2.4	Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την $Gamma(\lambda_0, 1)$ .	29
2.5	Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass & Raftery για την $Gamma(1, 1)$ .	30
2.6	Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass & Raftery για την $Gamma(\lambda_0, 1)$ .	30
2.7	Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την $Gamma(1, 1)$ .	32
2.8	Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την $Gamma(\lambda_0, 1)$ .	33
2.9	Αποτελέσματα του παράγοντα Bayes για εκ των προτέρων κατανομή $Beta(0.5, 0.5)$ .	37
2.10	Αποτελέσματα του παράγοντα Bayes για εκ των προτέρων κατανομή $Beta(1, 1)$ .	37
2.11	Αποτελέσματα του παράγοντα Bayes για εκ των προτέρων κατανομή $Beta(\pi_0, 1 - \pi_0)$ .	38
2.12	Αποτελέσματα του παράγοντα Bayes για κάθε αντιβιοτικό με εκ των προτέρων κατανομή $Beta(1, 1)$ .	41
2.13	Αποτελέσματα του παράγοντα Bayes για κάθε αντιβιοτικό με εκ των προτέρων κατανομή $Beta(0.1, 0.9)$ .	41
3.1	Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.	48
3.2	Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.	50
3.3	Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.	52
3.4	Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.	57
3.5	Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.	59
3.6	Αποτελέσματα του παράγοντα Bayes για τις ανά δύο συγκρίσεις υπό την $Beta(\frac{1}{2}, \frac{1}{2})$ εκ των προτέρων κατανομή.	60
3.7	Αποτελέσματα του παράγοντα Bayes για τις ανά δύο συγκρίσεις υπό την $Beta(\frac{1}{2}, \frac{1}{2})$ εκ των προτέρων κατανομή.	61
3.8	Πίνακας συχνοτήτων και πιθανοτήτων.	68
3.9	Μεγάλη αβεβαιότητα ότι $\pi_{1.} > \pi_{.1}$ αντιπροσωπεύεται από την σχέση $a_{12} = a_{21} = q$ .	69
3.10	Ισχυρή πίστη ότι $\pi_{1.} = \pi_{.1}$ αντιπροσωπεύεται από την σχέση $a_{12} = a_{21} = k \geq 1$ .	69
3.11	Εκ των προτέρων κατανομή η power εκ των προτέρων $Beta(\alpha, 1 - \alpha)$ , όπου $\alpha \in [0, 1]$ .	69

4.1	Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων. . . . .	76
4.2	Αποτελέσματα του παράγοντα Bayes για τους τρεις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων. . . . .	78
4.3	Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων. . . . .	82
4.4	Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων. . . . .	84
4.5	Πίνακας πιθανοτήτων των πληθυσμών A & B. . . . .	86
4.6	Πλήθος ασθενών ανά ηλικία που έλαβαν ( $E^+$ ) ή δεν έλαβαν ( $E^-$ ) θεραπεία λιδοκαΐνης για οξύ έμφραγμα του μυοκαρδίου. . . . .	89
4.7	Πλήθος ασθενών ανά ηλικία που έλαβαν ( $E^+$ ) ή δεν έλαβαν ( $E^-$ ) θεραπεία beta blocker για οξύ έμφραγμα του μυοκαρδίου. . . . .	89





# ΠΜΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ

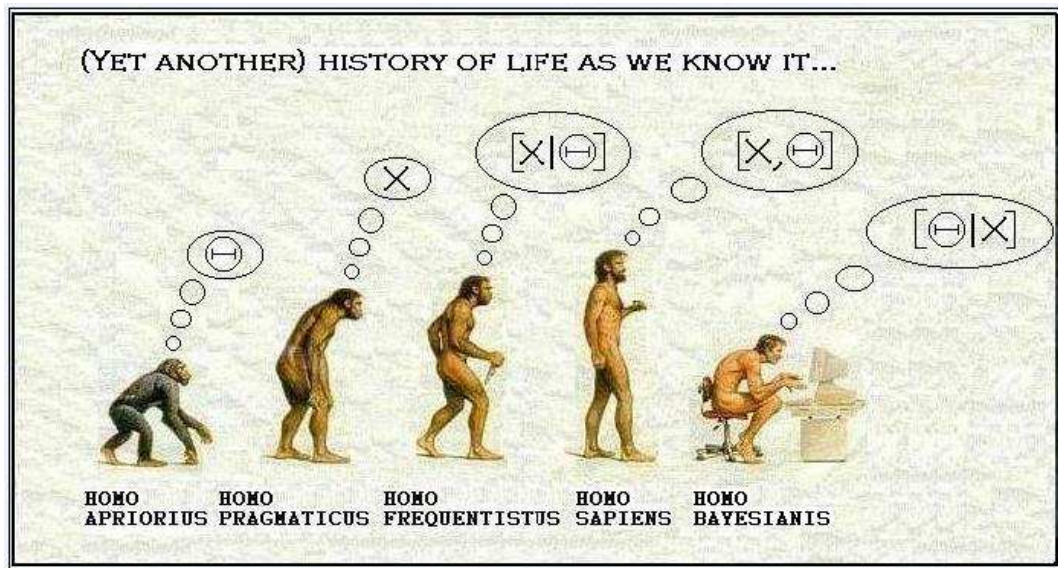
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

## ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΑΛΕΞΑΝΔΡΑ Γ. ΜΑΝΩΛΕΣΟΥ

*e-mail:* [amanoless@gmail.com](mailto:amanoless@gmail.com)

ΜΠΕΥΖΙΑΝΟΙ ΕΛΕΓΧΟΙ ΥΠΟΘΕΣΕΩΝ ΣΕ ΔΙΑΚΡΙΤΑ ΔΕΔΟΜΕΝΑ



Αθήνα, 2007



Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

## ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το Τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών και το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την ..... από την εξεταστική επιτροπή:

ΝΤΖΟΥΦΡΑΣ ΙΩΑΝΝΗΣ (ΕΠΙΒΛΕΠΩΝ)

ΕΠΙΚΟΤΡΟΣ ΚΑΘΗΓΗΤΗΣ

ΔΑΜΙΑΝΟΥ ΧΑΡΑΛΑΜΠΟΣ

ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ

ΤΣΙΑΜΥΡΤΖΗΣ ΠΑΝΑΓΙΩΤΗΣ

ΛΕΚΤΟΡΑΣ

## 0.1 Ευχαριστίες - Αφιερώσεις

Τα Μαθηματικά, και κατ' επέκταση η Στατιστική, είναι μια μαγική επιστήμη. Ο κάθε ερευνητής πολλές φορές ασχολείται με διάφορους τομείς της επιστήμης, μέχρι να έρθει η στιγμή που θα καταφέρει να τους συσχετίσει και να καταλάβει τι συμβαίνει και γιατί συμβαίνει. Εκεί κρύβεται η μαγεία και η ομορφιά τους!

Στο σημείο αυτό θα ήθελα να ευχαριστήσω τον καθηγητή μου κ. Ντζούφρα, που από το πρώτο έτος των προπτυχιακών μου σπουδών μου δίδαξε την μαγεία των πιθανοτήτων. Ακόμα να τον ευχαριστήσω για την πολύτιμη καθοδήγηση και υποστήριξή του μέχρι το πέρας αυτής της εργασίας.

Η παρούσα διπλωματική εργασία είναι αφιερωμένη στην γιαγιά και στον παππού μου. Τέλος, να ευχαριστήσω τον φίλο μου Βενέδικτο που με μύησε στην τέχνη του  $\text{\LaTeX}$ , ώστε να έχετε τώρα στα χέρια σας αυτό το κομψό αποτέλεσμα!

# Κεφάλαιο 1

## Εισαγωγή στην Στατιστική κατά Bayes

### 1.1 Μπευζιανή συμπερασματολογία

Η θεμελιώδης διαφορά της Κλασσικής με την Μπευζιανή στατιστική, έγκειται στο γεγονός ότι η δεύτερη θεωρεί την παράμετρο  $\theta$  ως τυχαία μεταβλητή που ακολουθεί κάποια κατανομή και όχι ως άγνωστη και σταθερή ποσότητα την οποία πρέπει να υπολογίσουμε. Η κατανομή αυτή καλείται εκ των προτέρων κατανομή πιθανότητας  $f(\theta)$  (prior probability distribution) και εκφράζει την εκ των προτέρων γνώση και πεποίθηση μας για την παράμετρο, πριν λάβουμε υπόψη τα δεδομένα  $\mathbf{y}$ . Έτσι η συμπερασματολογία δεν προκύπτει μόνο από την μελέτη των δεδομένων, δηλαδή την πιθανοφάνεια  $f(\mathbf{y}|\theta)$ , αλλά τελικά από την συνάρτηση πιθανότητας της κατανομής της παραμέτρου δοθέντος των δεδομένων  $f(\theta|\mathbf{y})$ , η οποία καλείται εκ των υστέρων κατανομή (posterior distribution).

Ο υπολογισμός της δίνεται από τον τύπο:

$$f(\theta|\mathbf{y}) = \frac{f(\theta)f(\mathbf{y}|\theta)}{\sum_{\theta_i \in \Theta} f(\theta_i)f(\mathbf{y}|\theta_i)}$$

που αποτελεί το θεώρημα του Bayes.

Όταν η παράμετρος  $\theta$  είναι συνεχής, τότε η σταθερά κανονικοποίησης  $\sum_{\theta_i \in \Theta} f(\theta_i)f(\mathbf{y}|\theta_i)$  αντικαθίσταται από τον όρο  $\int_{\theta \in \Theta} f(\theta)f(\mathbf{y}|\theta)d\theta$ .

Η σταθερά κανονικοποίησης είναι μια συνάρτηση μόνο των δεδομένων  $\mathbf{y}$  και δεν εξαρτάται καθόλου από την παράμετρο  $\theta$ . Για τον λόγο αυτό συχνά στην βιβλιογραφία το θεώρημα του Bayes παρουσιάζεται ως :

$$f(\theta|\mathbf{y}) \propto f(\theta)f(\mathbf{y}|\theta)$$

δηλαδή η εκ των υστέρων κατανομή είναι ανάλογη της εκ των προτέρων κατανομής πολλαπλασιαζόμενης με την συνάρτηση πιθανοφάνειας.

### Παράδειγμα Poisson

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από την Poisson κατανομή με συνάρτηση πιθανοφάνειας

$$L(\theta|\mathbf{y}) = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod_{i=1}^n y_i!} \propto e^{-n\theta}\theta^{\sum y_i}, \quad i = 1, \dots, n$$

και με εκ των προτέρων κατανομή την:

$$f(\theta) = e^{-\theta}, \quad \theta > 0$$

Εφαρμόζοντας το θεώρημα του Bayes παίρνουμε

$$f(\theta|\mathbf{y}) \propto f(\theta)L(\theta|\mathbf{y})$$

$$\propto e^{-\theta} \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod_{i=1}^n y_i!}$$

$$\propto e^{-(n+1)\theta}\theta^{\sum y_i+1-1}$$

Άρα

$$f(\theta|\mathbf{y}) \sim \text{Gamma}(\sum y_i + 1, n + 1)$$

## 1.2 Εκ των προτέρων κατανομές

Η επιλογή της εκ των προτέρων κατανομής πρέπει να γίνεται με πολύ προσοχή διότι καθιστά την ανάλυση υποκειμενική (subjective Bayesian analysis).

Διαφορετική εκ των προτέρων κατανομή οδηγεί σε διαφορετικά αποτελέσματα. Ωστόσο μια «λογική» εκ των προτέρων κατανομή χάνει την επίδραση της καθώς συγκεντρώνονται όλο και περισσότερα δεδομένα.

Η επιλογή του κατάλληλου τύπου της εκ των προτέρων κατανομής είναι μια δύσκολη αποστολή. Τυπικά αυτές οι κατανομές προσδιορίζονται βασιζόμενες στην πληροφορία που συσσωρεύεται από προηγούμενες μελέτες ή από γνώμες ειδικών. Συνήθως διαλέγουμε κατανομές που απλοποιούν τους υπολογισμούς οι οποίες όμως περιορίζουν την επιλογή της  $f(\theta)$  σε κάποια από τις γνωστές οικογένειες κατανομών. Σε πολλές περιπτώσεις η διαθέσιμη εκ των προτέρων πληροφορία είναι περιορισμένη. Στην περίπτωση αυτή επιθυμούμε η πληροφορία που προέρχεται από τα δεδομένα, να κυριαρχήσει στον τελικό υπολογισμό της εκ των υστέρων κατανομής.

Γενικά οι κατανομές διακρίνονται σε δύο μεγάλες κατηγορίες, τις πληροφοριακές (informative) και τις μη-πληροφοριακές (non-informative). Πέρα όμως από αυτή την διαφοροποίηση υπάρχουν διάφορες κατηγορίες εκ των προτέρων κατανομών.

Μερικές από τις κατηγορίες εκ των προτέρων κατανομών είναι οι ακόλουθες:

- Συζυγείς κατανομές (*Conjugate Priors*)
- Η μη πληροφοριακή εκ των προτέρων του Jeffrey (*Jeffreys Prior*)
- Υπέρ εκ των προτέρων κατανομές (*Hyperpriors*)
- Εκ των προτέρων κατανομές βασισμένες σε δυνάμεις της πιθανοφάνειας (*Power Priors*)

### 1.2.1 Συζυγείς κατανομές

Η δυσκολία στον υπολογισμό της σταθεράς κανονικοποίησης στον τύπο του Bayes οδήγησε στην χρήση των συζυγών κατανομών.



### Ορισμός συζυγίας

Εάν  $F$  είναι μια οικογένεια κατανομών  $f(y|\theta)$  και  $P$  είναι μια οικογένεια εκ των προτέρων κατανομών για το  $\theta$ , τότε η κατηγορία  $P$  είναι συζυγής για την  $F$  εάν

$$f(\theta|y) \in P \quad \forall \quad \{ f(\cdot|\theta) \in F \quad \text{και} \quad f(\cdot) \in P \}.$$

Εμάς μας ενδιαφέρουν οι φυσικές (natural) συζυγείς εκ των προτέρων οικογένειες που προκύπτουν ορίζοντας το  $P$  να είναι το σύνολο όλων των πυκνοτήτων με τον ίδιο συναρτησιακό τύπο πιθανοφάνειας (Gelman et al. 1995, σελ. 37)[23].

Οι συζυγείς εκ των προτέρων κατανομές παίζουν σημαντικό ρόλο στην Μπευζιανή συμπερασματολογία, αφού είναι επιθυμητό να έχουμε εκ των υστέρων κατανομές με την ίδια συναρτησιακή μορφή και παρόμοιες ιδιότητες με την εκ των προτέρων. Οι συζυγείς συχνά έχουν επιθυμητά χαρακτηριστικά σημαντικά για την ερμηνεία, την ανάλυση δεδομένων και τους υπολογισμούς, αναφέρουν οι Chen και Ibrahim (2003)[15], και συνεχίζουν λέγοντας ότι οι συζυγείς είναι απλές και άμεσες στην κατασκευή για πολλά μοντέλα υπό την προϋπόθεση ότι το δείγμα που παρατηρούμε προέρχεται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές. Είναι γνωστό ότι συζυγείς κατανομές είναι διαθέσιμες για την εκθετική οικογένεια. Συνδυασμοί πιθανοφάνειας - εκ των προτέρων κατανομών περιλαμβάνουν τις κανονική-κανονική, binomial-beta, Poisson-gamma και gamma-gamma. Οι Diaconis & Ylvisaker (1979)[19], και Morris (1982, 1983)[42],[43] εξέτασαν γενικές κατηγορίες συζυγών εκ των προτέρων κατανομών για μοντέλα εκθετικής οικογένειας. Ο Morris (1983)[43] έδειξε ότι κατανομές που ανήκουν στην εκθετική οικογένεια έχουν φυσικές συζυγείς εκ των προτέρων κατανομές.

### Παράδειγμα Poisson

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από την Poisson κατανομή με εκ των προτέρων κατανομή μια  $Gamma(p, q)$  :

$$f(\theta) = \frac{q^p}{\Gamma(p)} \theta^{p-1} \exp\{-q\theta\}$$

Εφαρμόζουμε το θεώρημα του Bayes

$$\begin{aligned} f(\theta|\mathbf{y}) &\propto f(\theta)L(\theta|\mathbf{y}) \\ &\propto \theta^{(p+\sum y_i-1)} \exp\{-(q+n)\theta\} \\ &\equiv \text{Gamma}(p + \sum y_i, q + n) \end{aligned}$$

που δεν είναι άλλη από μια *Gamma* με παραμέτρους που εξαρτώνται από τα δεδομένα. Από την εκ των υστέρων κατανομή έχουμε

$$\begin{aligned} E(\theta|\mathbf{y}) &= \frac{p + \sum y_i}{q + n} \\ &= \frac{q}{(q+n)} \frac{p}{q} + \frac{n}{(q+n)} \bar{y} \\ &= (1-w)E(\theta) + w\bar{y} \end{aligned}$$

που είναι ένας σταθμισμένος μέσος του εκ των προτέρων μέσου και του δειγματικού μέσου. Όπου  $w = \frac{n}{q+n}$ , οπότε καθώς  $n \rightarrow \infty$  το  $w \rightarrow 1$  και άρα  $E(\theta|y) \rightarrow \bar{y}$

$$\begin{aligned} V(\theta|\mathbf{y}) &= \frac{p + \sum y_i}{(q+n)^2} \\ &= \frac{q^2}{(q+n)^2} \frac{p}{q^2} + \frac{n^2}{(q+n)^2} \frac{\sum y_i}{n^2} \\ &= (1-w)^2 V(\theta) + w^2 \frac{\sum y_i}{n^2} \end{aligned}$$

Όπου  $w = \frac{n}{q+n}$ .

## Παράδειγμα Binomial

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από την  $Binomial(N_i, \theta)$ .

$$f(\mathbf{y}|\theta) = \prod_{i=1}^n \binom{N_i}{y_i} \theta^{\sum y_i} (1 - \theta)^{N - \sum y_i}$$

όπου  $N = \sum_{i=1}^n N_i$ .

Παραμετροποιούμε την εκ των προτέρων κατανομή επιλέγοντας μια *Beta* κατανομή.

$$f(\theta) \propto \theta^{p-1} (1 - \theta)^{q-1}$$

Η εκ των υστέρων κατανομή που προκύπτει είναι επίσης μια *Beta*:

$$\begin{aligned} f(\theta|\mathbf{y}) &\propto \theta^{\sum y_i} (1 - \theta)^{N - \sum y_i} \theta^{p-1} (1 - \theta)^{q-1} \\ &= \theta^{\sum y_i + p - 1} (1 - \theta)^{N - \sum y_i + q - 1} \\ &\equiv Beta\left(p + \sum y_i, q + N - \sum y_i\right) \end{aligned}$$

Από την εκ των υστέρων κατανομή έχουμε

$$\begin{aligned} E(\theta|\mathbf{y}) &= \frac{p + \sum y_i}{p + q + N} \\ &= \frac{p + q}{p + q + N} \frac{p}{p + q} + \frac{N}{p + q + N} \frac{\sum y_i}{N} \\ &= (1 - w)E(\theta) + w\hat{\theta} \end{aligned}$$

που έγκειται μεταξύ της δειγματικής αναλογίας  $E(\mathbf{y}) = \frac{\sum y_i}{N} = \hat{\theta}$  και του εκ των προτέρων μέσου  $E(\theta) = \frac{p}{p+q}$ , όπου  $w = \frac{N}{p+q+N}$ .

$$V(\theta|\mathbf{y}) = \frac{(p + \sum y_i)(q + N - \sum y_i)}{(p + q + N)^2(p + q + N + 1)} = \frac{E(\theta|\mathbf{y})[1 - E(\theta|\mathbf{y})]}{p + q + N + 1}$$

Για δοθέν  $p$  και  $q$ , καθώς τα  $\sum y_i$  και  $N - \sum y_i$  μεγαλώνουν, έχουμε  $E(\theta|\mathbf{y}) \rightarrow \hat{\theta}$

και  $V(\theta|\mathbf{y}) \rightarrow \frac{\hat{\theta}(1-\hat{\theta})}{N}$ . Δηλαδή οριακά οι παράμετροι της εκ των προτέρων δεν έχουν καμιά επιρροή στην εκ των υστέρων κατανομή εφόσον και οι εκ των υστέρων παράμετροι συγκλίνουν στους δειγματικούς εκτιμητές (Gelman et al. 1995, σελ. 36)[23].

### Μείξεις συζυγών κατανομών (Mixtures Conjugate Priors)

Μια συζυγής εκ των προτέρων μπορεί να μην απεικονίζει με ακρίβεια την εκ των προτέρων γνώση μας. Έτσι ένα πεπερασμένο μείγμα συζυγών κατανομών μπορούν επίσης να απλοποιήσουν τους υπολογισμούς.

Έστω ότι έχουμε την  $f(y|\theta)$  για την παράμετρο  $\theta$  και υποθέτουμε ένα μείγμα δύο εκ των προτέρων κατανομών.

$$\pi(\theta) = \alpha\pi_1(\theta) + (1 - \alpha)\pi_2(\theta), \quad 0 \leq \alpha \leq 1$$

όπου οι  $\pi_1$  και  $\pi_2$  είναι δυνατόν να ανήκουν στην οικογένεια των συζυγών κατανομών της  $f(y|\theta)$ . Οπότε:

$$\begin{aligned} f(\theta|y) &= \frac{f(y|\theta)\pi_1(\theta)\alpha + f(y|\theta)\pi_2(\theta)(1 - \alpha)}{\int [f(y|\theta)\pi_1(\theta)\alpha + f(y|\theta)\pi_2(\theta)(1 - \alpha)]d\theta} \\ &= \frac{\frac{f(y|\theta)\pi_1(\theta)}{m_1(y)}m_1(y)\alpha + \frac{f(y|\theta)\pi_2(\theta)}{m_2(y)}m_2(y)(1 - \alpha)}{m_1(y)\alpha + m_2(y)(1 - \alpha)} \\ &= \frac{f(y|\theta)\pi_1(\theta)}{m_1(y)}W_1 + \frac{f(y|\theta)\pi_2(\theta)}{m_2(y)}(1 - W_1) \\ &= f_1(\theta|y)W_1 + f_2(\theta|y)(1 - W_1) \end{aligned}$$

όπου

$$m_1(y) = \int f(y|\theta)\pi_1(\theta)d\theta$$

$$m_2(y) = \int f(y|\theta)\pi_2(\theta)d\theta$$

$$W_1 = \frac{m_1(y)\alpha}{m_1(y)\alpha + m_2(y)(1 - \alpha)}$$

Συνεπώς ένα μείγμα εκ των προτέρων συζυγών κατανομών οδηγεί σε μείγμα εκ των υστέρων συζυγών κατανομών (Carlin & Louis 1996, σελ. 33)[12].

Δεν πρέπει όμως να ξεχνάμε ότι βασικός λόγος χρήσης των συζυγών κατανομών είναι ότι απλουστεύουν τους μαθηματικούς υπολογισμούς και ότι είναι εύκολο να ερμηνευτούν τα αποτελέσματα. Γι' αυτό η χρήση τους πρέπει να είναι προσεκτική και να αποφεύγεται όταν δεν εκφράζουν σωστά τις πεποιθήσεις μας. Συχνά αποτελούν μια καλή προσέγγιση, όμως για σύνθετα μοντέλα η συζυγία μπορεί να μην είναι καν δυνατή (Gelman et al. 1995, σελ. 37)[23].

### 1.2.2 Η μη πληροφοριακή εκ των προτέρων του Jeffreys

Όταν οι εκ των προτέρων κατανομές δεν έχουν πληθυσμιακή βάση, δηλαδή όταν δεν έχουμε επαρκή εκ των προτέρων πληροφορία, τότε είναι δύσκολο να κατασκευαστούν. Έτσι δημιουργούμε εκ των προτέρων κατανομές με μικρή ακρίβεια ώστε να διαφυλάξουμε ότι η εκ των προτέρων θα έχει μηδενική επίδραση στον σχηματισμό της εκ των υστέρων κατανομής. Αυτές οι κατανομές καλούνται επίπεδες (flat) ή μη-πληροφοριακές (non-informative). Η πιο γνωστή μη-πληροφοριακή εκ των προτέρων κατανομή είναι η ομοιόμορφη (Uniform).

Δίνοντας όμως στις κατανομές αυτές μικρή ακρίβεια [ $\tau = \frac{1}{\sigma_0^2} \rightarrow 0$  στην περίπτωση της κανονικής κατανομής, όπου  $\mu \sim N(\mu_0, \sigma_0^2)$ ] λαμβάνουμε μια εκ των προτέρων κατανομή με διασπορά που τείνει στο άπειρο, η οποία δεν είναι μια σωστά ορισμένη κατανομή. Η κατανομή γίνεται επίπεδη και το ολοκλήρωμα  $\int \pi(\theta)d\theta = \infty$ , όπου η  $\pi(\theta)$  είναι μια μη γνήσια (improper) κατανομή. Ωστόσο οι μη γνήσιες

κατανομές δεν είναι απαγορευτικές στην Μπευζιανή θεωρία αφού μπορούν να μας δώσουν γνήσιες (proper) εκ των υστέρων κατανομές. Εντούτοις θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην ερμηνεία των εκ των υστέρων που προέρχονται από μη γνήσιες κατανομές και πάντα να ελέγχουμε ότι η εκ των υστέρων έχει πεπερασμένο ολοκλήρωμα (Gelman et al. 1995, σελ. 52-54)[23].

Μια προσέγγιση που χρησιμοποιείται ευρέως στον καθορισμό μη-πληροφοριακών μη γνήσιων κατανομών βασίζεται στην πληροφορία του Fisher, και εισηγήθηκε από τον Jeffreys ως λύση στο πρόβλημα ότι η ομοιόμορφη δεν αποδίδει μια σταθερή ανάλυση όταν οι παράμετροι μετασχηματίζονται.

Η εκ των προτέρων κατανομή του Jeffreys ορίζεται ως :

$$\pi_o(\theta) \propto |I(\theta)|^{1/2}$$

όπου  $I(\theta)$  είναι η πληροφορία του Fisher που δίνεται από τον τύπο

$$\begin{aligned} I(\theta) &= E_{y|\theta} \left[ \left( \frac{d \log f(y|\theta)}{d\theta} \right)^2 \right] \\ &= -E_{y|\theta} \left[ \frac{d^2 \log f(y|\theta)}{d\theta^2} \right] \end{aligned}$$

### Παράδειγμα Poisson

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από την Poisson με παράμετρο  $\lambda$ .

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^n f(\mathbf{y}|\lambda) = \frac{\exp\{-n\lambda\} \lambda^{\sum y_i}}{\prod_{i=1}^n y_i!}$$

$$\log L(\lambda|\mathbf{y}) = -n\lambda + \sum y_i \log \lambda + C$$

$$\frac{d \log L(\lambda|\mathbf{y})}{d\lambda} = -n + \frac{\sum y_i}{\lambda}$$

$$\begin{aligned}\frac{d^2 \log L(\lambda|\mathbf{y})}{d\lambda^2} &= -\frac{\sum y_i}{\lambda^2} \\ I(\lambda) &= -E \left[ \frac{d^2 \log L(\lambda|\mathbf{y})}{d\lambda^2} \right] \\ &= E \left[ \frac{\sum y_i}{\lambda^2} \right] \\ &= \frac{n}{\lambda}\end{aligned}$$

$$f_o(\lambda) \propto |I(\lambda)|^{1/2} = \frac{\sqrt{n}}{\sqrt{\lambda}} \propto \lambda^{-\frac{1}{2}}$$

### Παράδειγμα Binomial

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από την  $Binomial(N_i, \theta)$ .

$$L(\theta|\mathbf{y}) = f(\mathbf{y}|\theta) = \prod_{i=1}^n \binom{N_i}{y_i} \theta^{\sum y_i} (1-\theta)^{N-\sum y_i}$$

όπου  $N = \sum_{i=1}^n N_i$ .

$$\log L(\theta|\mathbf{y}) = \sum y_i \log \theta + (N - \sum y_i) \log(1-\theta) + C$$

$$\frac{d \log L(\theta|\mathbf{y})}{d\theta} = \frac{\sum y_i}{\theta} - \frac{N - \sum y_i}{1-\theta}$$

$$\frac{d^2 \log L(\theta|\mathbf{y})}{d\theta^2} = -\frac{\sum y_i}{\theta^2} - \frac{N - \sum y_i}{(1-\theta)^2}$$

$$\begin{aligned}I(\theta) &= -E \left[ \frac{d^2 \log L(\theta|\mathbf{y})}{d\theta^2} \right] \\ &= \frac{E(\sum y_i)}{\theta^2} + \frac{N - E(\sum y_i)}{(1-\theta)^2} \\ &= \frac{N}{\theta(1-\theta)}\end{aligned}$$

$$f_o(\theta) \propto |I(\theta)|^{1/2} = \theta^{-1/2}(1-\theta)^{-1/2}$$

η οποία είναι μια  $Beta(\frac{1}{2}, \frac{1}{2})$  κατανομή.

### 1.2.3 Υπέρ εκ των προτέρων κατανομές

Έστω ότι έχουμε ένα δείγμα από κατανομή  $f(y|\theta)$  και την εκ των προτέρων της  $\theta$ ,  $f(\theta|\phi)$ . Όμως το  $\phi$  δεν είναι γνωστό και έτσι έχει την δική του εκ των προτέρων κατανομή  $f(\phi)$ . Από τα παραπάνω προκύπτει τελικά η εκ των υστέρων κατανομή η οποία δίνεται από τον τύπο (Carlin & Louis 1996, σελ. 23)[12].

$$\begin{aligned} f(\theta|y) &= \frac{f(y, \theta)}{f(y)} \\ &= \frac{\int f(y, \theta, \phi) d\phi}{\int \int f(y, \theta, \phi) d\phi d\theta} \\ &= \frac{\int f(y|\theta) f(\theta|\phi) f(\phi) d\phi}{\int \int f(y|\theta) f(\theta|\phi) f(\phi) d\phi d\theta} \end{aligned}$$

Μια δεύτερη εκ των υστέρων κατανομή είναι η  $f(\phi|y)$  η οποία δίνεται από τον τύπο (Gelman et al. 1995, σελ. 129 [23] & Bernardo et al. 1994, σελ. 372) [11]

$$\begin{aligned} p(\phi|y) &= \frac{f(y|\phi) f(\phi)}{f(y)} = \frac{[\int f(y|\theta) f(\theta|\phi) d\theta] f(\phi)}{f(y)} \\ &= \left[ \frac{f(y|\theta) f(\theta|\phi)}{f(\theta|y)} \right] \frac{f(\phi)}{f(y)} = \frac{f(y|\phi, \theta) f(\theta|\phi) f(\phi)}{f(\theta|\phi, y) f(y)} \\ &= \frac{f(y|\phi, \theta) f(\theta, \phi)}{f(\theta|\phi, y) f(y)} \\ &= \frac{f(\phi, \theta|y)}{f(\theta|\phi, y)} \end{aligned}$$

Οι δύο αυτές εκ των προτέρων κατανομές συνδέονται με την σχέση (Bernardo et al. 1994, σελ. 372)[11]

$$f(\theta|y) = \int f(\theta, \phi|y) d\phi = \int f(\theta|\phi, y) f(\phi|y) d\phi$$

Εναλλακτικά μπορούμε να αντικαταστήσουμε την παράμετρο  $\phi$  με μια εκτιμήτρια της  $\hat{\phi}$ , που θα προκύψει μεγιστοποιώντας την κατανομή  $f(y|\phi)$  ως προς  $\phi$ .



Συνεπώς η συμπερασματολογία θα βασίζεται στην «εκτιμώμενη» εκ των υστέρων  $f(\theta|y, \hat{\phi})$ . Η παραπάνω προσέγγιση αναφέρεται συχνά ως Εμπειρική Μπευζιανή ανάλυση (Empirical Bayes analysis) λόγω του ότι χρησιμοποιούμε τα δεδομένα για να υπολογίσουμε την εκτιμήτρια της εκ των προτέρων παραμέτρου  $\phi$  (Carlin & Louis 1996, σελ. 24)[12].

Θα μπορούσε επίσης η υπερπαραμέτρος  $\phi$ , από μόνη της να εξαρτάται από μια άλλη παράμετρο, έστω  $\psi$ . Κατ' επέκταση ορίζεται μια δευτέρου επιπέδου εκ των προτέρων κατανομή  $f(\phi|\psi)$  και μια τρίτου επιπέδου εκ των προτέρων κατανομή  $f(\psi)$ . Το εγχείρημα αυτό καλείται Ιεραρχική μοντελοποίηση (Hierarchical modeling, Carlin & Louis 1996, σελ. 24)[12].

### Παράδειγμα Poisson

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από  $Poisson(\lambda_i)$ . Υποθέτουμε ότι η παράμετρος  $\lambda_i$  είναι ένα ανεξάρτητο δείγμα από κατανομή  $Gamma(\lambda_o, 1)$  (power prior). Επίσης θέτουμε την υπερπαραμέτρο  $\lambda_o \sim Gamma(1, 1)$ . Από τα παραπάνω λαμβάνουμε την εκ των υστέρων κατανομή

$$\begin{aligned} f(\lambda_o|\mathbf{y}) &\propto \frac{\prod_{i=1}^n e^{-\lambda_o} \prod_{i=1}^n \frac{1}{\Gamma(\lambda_o)} \lambda_i^{\lambda_o-1} e^{-\lambda_i} \prod_{i=1}^n \frac{1}{y_i!} \lambda_i^{y_i} e^{-\lambda_i}}{\prod_{i=1}^n \frac{2^{\lambda_o+y_i}}{\Gamma(\lambda_o+y_i)} \lambda_i^{\lambda_o+y_i-1} e^{-2\lambda_i}} \\ &\propto \prod_{i=1}^n \frac{\Gamma(\lambda_o + y_i)}{\Gamma(\lambda_o) y_i! 2^{\lambda_o+y_i}} e^{-\lambda_o} \end{aligned}$$

### Παράδειγμα Binomial

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από  $Binomial(N_i, \theta_i)$ . Υποθέτουμε ότι η παράμετρος  $\theta_i$  είναι ένα ανεξάρτητο δείγμα από κατανομή  $Beta(\alpha, \beta)$ . Ακόμα, για ευκολία στους υπολογισμούς, θεωρούμε ότι οι υπερπαραμέτροι είναι μεταξύ τους ανεξάρτητοι και ότι  $\alpha \sim Beta(1, 1)$  και  $\beta \sim Beta(1, 1)$ . Αυτό συνεπάγεται ότι η από κοινού συνάρτηση κατανομής τους είναι  $f(\alpha, \beta) = f(\alpha)f(\beta) = 1$ . Από τα

παραπάνω λαμβάνουμε την εκ των υστέρων κατανομή

$$\begin{aligned}
 f(\alpha, \beta | \mathbf{y}) &\propto \frac{\prod_{i=1}^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta_i^{\alpha-1} (1-\theta_i)^{\beta-1} \prod_{i=1}^n \theta_i^{y_i} (1-\theta_i)^{N_i-y_i}}{\prod_{i=1}^n \frac{\Gamma(\alpha+\beta+N_i)}{\Gamma(\alpha+y_i)\Gamma(\beta+N_i-y_i)} \theta_i^{\alpha+y_i-1} (1-\theta_i)^{\beta+N_i-y_i-1}} \\
 &\propto \prod_{i=1}^n \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+y_i)\Gamma(\beta+N_i-y_i)}{\Gamma(\alpha+\beta+N_i)} \\
 &\propto \prod_{i=1}^n \frac{B(\alpha+y_i, \beta+N_i-y_i)}{B(\alpha, \beta)}
 \end{aligned}$$

#### 1.2.4 Εκ των προτέρων κατανομές βασισμένες σε δυνάμεις της πιθανοφάνειας

Αν και οι μη-πληροφοριακές και οι μη-γνήσιες εκ των προτέρων κατανομές είναι εύκολες να οριστούν για κάποια προβλήματα, δεν μπορούν να χρησιμοποιηθούν παντού όπως για παράδειγμα στην επιλογή μοντέλων ή στη σύγκριση μοντέλων, όπου απαιτούνται κατάλληλες εκ των προτέρων για τον υπολογισμό του παράγοντα Bayes. Έτσι οι πληροφοριακές εκ των προτέρων είναι απαραίτητες αφού επιτρέπουν την εφαρμογή σημαντικής πληροφορίας από προηγούμενες παρόμοιες μελέτες. Τα δεδομένα προηγούμενων μελετών καλούνται ιστορικά δεδομένα (historical data).

Η βασική ιδέα της power prior είναι η χρήση της μιας παραμέτρου  $\delta$ , ( $0 \leq \delta \leq 1$ ) μέσω της οποίας ελέγχουμε την επίδραση των ιστορικών δεδομένων στην τρέχουσα μελέτη. Υποθέτουμε ότι για δοθέν  $\theta$  τα ιστορικά δεδομένα  $\mathbf{y}_o$  και τα τρέχοντα δεδομένα  $\mathbf{y}$  είναι ανεξάρτητα τυχαία δείγματα. Ορίζουμε  $L(\theta | \mathbf{y}_o)$  τη συνάρτηση πιθανοφάνειας της παραμέτρου που βασίζεται στα ιστορικά δεδομένα και  $\pi(\theta)$  να είναι η αρχική (initial) εκ των προτέρων κατανομή της παραμέτρου, δηλαδή πριν από κάθε ιστορική πληροφορία. Αυτή είναι συνήθως μια μη - πληροφοριακή εκ των προτέρων κατανομή. Για δοθέν  $\delta$ , οι Chen και Ibrahim (2000, 2006) [28], [14], ορίσανε την power prior του  $\theta$  για μια τρέχουσα μελέτη να είναι:

$$\pi(\theta | \mathbf{y}_o, \delta) \propto [L(\theta | \mathbf{y}_o)]^\delta \pi(\theta)$$

Η παράμετρος  $\delta$  ελέγχει το επιθυμητό βάρος που τα ιστορικά δεδομένα θα έχουν στην εκ των υστέρων κατανομή. Στις περιπτώσεις που

- $\delta = 0$  σημαίνει ότι τα ιστορικά δεδομένα δεν θα χρησιμοποιηθούν (δηλ. θα έχουν βάρος ίσο με μηδέν).
- $\delta = 1$  σημαίνει ότι δίνουμε το ίδιο βάρος στις παρατηρήσεις των ιστορικών δεδομένων  $\mathbf{y}_o$  και των των παρατηρήσεων  $\mathbf{y}$ .

Η παράμετρος  $\delta$  μπορεί επίσης να ερμηνευτεί σαν παράμετρος ακρίβειας (precision parameter). Για παράδειγμα, έστω ότι έχουμε την περίπτωση ενός κανονικού δείγματος με γνωστή διασπορά. Υποθέτουμε ότι τα  $\mathbf{y}_o$  αποτελούνται από  $n_o$  παρατηρήσεις από κανονικό πληθυσμό με άγνωστο μέσο  $\theta$  και γνωστή διασπορά  $\sigma^2$ . Εάν η εκ των προτέρων κατανομή  $\pi(\theta)$  υποθέσουμε ότι είναι μια ομοιόμορφη (μη-πληροφοριακή) κατανομή τότε συνεπάγεται μια εκ των προτέρων κατανομή του  $\theta$  για τα τρέχοντα δεδομένα  $\mathbf{y}$ ,  $[\theta|\mathbf{y}_o, \delta] \sim N(\bar{y}_o, \frac{\sigma^2}{\delta n_o})$ , όπου  $\bar{y}_o$  είναι ο δειγματικός μέσος των ιστορικών δεδομένων. Έτσι, το  $\delta$  αποτελεί μέρος της ακρίβειας αφού για μικρές τιμές του  $\delta$  έχουμε μεγάλη power prior διασπορά, ενώ για μεγάλες τιμές του  $\delta$  έχουμε μικρή power prior διασπορά.

Η power prior  $\pi(\theta|\mathbf{y}_o, \delta)$  αρχικά κατασκευάστηκε για σταθερό  $\delta$ . Οι Ibrahim και Chen (2000)[28] επεκτάθηκαν στην περίπτωση όπου το  $\delta$  είναι τυχαίο, ισχυριζόμενοι ότι ένα τυχαίο  $\delta$  δίνει στον ερευνητή περισσότερη ελαστικότητα στην χρήση των ιστορικών δεδομένων. Πρότειναν μια από κοινού power prior κατανομή για τα  $(\theta, \delta)$  ως,

$$\pi(\theta, \delta|\mathbf{y}_o, \delta) \propto [L(\theta|\mathbf{y}_o)]^\delta \pi(\theta) \pi(\delta|\gamma_o)$$

όπου το  $\gamma_o$  είναι μια καθορισμένη υπερπαράμετρος. Μια φυσική εκ των προτέρων για το  $\delta$  μπορεί να είναι μια  $Beta(\alpha, b)$  κατανομή ή απλά μια ομοιόμορφη κατανομή αφού  $0 \leq \delta \leq 1$ .

Στη συνέχεια, γενίκευσαν την power prior για πολλαπλά σετ ιστορικών δεδομένων. Υποθέτουμε ότι υπάρχουν  $k$  ανεξάρτητες ιστορικές μελέτες,  $\mathbf{y}_{oj}$  τα ιστορικά δεδομένα της  $j$ -μελέτης, με  $j = 1, \dots, k$  και  $\mathbf{y}_o = (y_{o1}, \dots, y_{ok})$ . Ακόμα πρότειναν

μια διαφορετική παράμετρο βαρύτητας  $\delta_j$  για κάθε  $j$ -ιστορική μελέτη και πήραν το  $\delta_j$  να ακολουθεί ανεξάρτητα και ισόνομα μια  $Beta(\alpha, b)$  κατανομή. Έστω  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_k)$  τότε η power prior για πολλαπλά ιστορικά δεδομένα παίρνει την μορφή

$$\pi(\boldsymbol{\theta}, \boldsymbol{\delta} | \mathbf{y}_o) \propto \left[ \prod_{j=1}^k [L(\boldsymbol{\theta} | \mathbf{y}_{oj})]^{\delta_j} \pi(\delta_j | a, b) \right] \pi(\boldsymbol{\theta}) .$$

Αυτή η προσέγγιση διευθετεί ενδεχόμενη ετερογένεια μεταξύ των διάφορων ιστορικών δεδομένων και έτσι ο ρόλος των ιστορικών δεδομένων μπορεί να αποτιμηθεί με περισσότερη ακρίβεια. Για περισσότερες πληροφορίες παραπέμπουμε στην διδακτορική εργασία της Yuyan Duan (2005) [49].

Για την παράμετρο  $\delta$  μπορούμε να χρησιμοποιήσουμε μια εκ των προτέρων κατανομή με μέσο  $\frac{1}{n}$ , έτσι ώστε τα ιστορικά δεδομένα να έχουν εκ των προτέρων βαρύτητα κατά μέσο όρο ίση με ένα σημείο δεδομένων.

### Παράδειγμα Poisson

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από την Poisson με παράμετρο  $\lambda$ . Για τον υπολογισμό μιας power εκ των προτέρων κατανομής για την Poisson κατανομή, θέτουμε ως αρχική εκ των προτέρων  $\pi(\lambda) \sim Gamma(\alpha_o, \beta_o)$ . Από τα παραπάνω προκύπτει

$$\begin{aligned} \pi(\lambda | \mathbf{y}_o, \delta) &\propto [L(\lambda | \mathbf{y}_o)]^\delta \pi(\lambda) \\ &\propto \left[ \prod_{i=1}^n \frac{e^{-\lambda} \lambda^{y_{oi}}}{y_{oi}!} \right]^\delta \frac{\beta_o^{\alpha_o}}{\Gamma(\alpha_o)} \lambda^{\alpha_o-1} e^{-\beta_o \lambda} \\ &\propto \left[ e^{-n\lambda} \lambda^{\sum y_{oi}} \right]^\delta \lambda^{\alpha_o-1} e^{-\beta_o \lambda} \\ &= \lambda^{\delta n \bar{y}_o + \alpha_o - 1} e^{-(\delta n + \beta_o) \lambda} \end{aligned}$$

Που είναι μια  $Gamma(\delta n \bar{y}_o + \alpha_o, \delta n + \beta_o)$ . Εάν θέσουμε  $\alpha_o = \beta_o = 0$  και  $\delta = \frac{1}{n}$  παίρνουμε την υπό περίπτωση όπου η power prior θα είναι μια  $Gamma(\bar{y}_o, 1)$ .

### Παράδειγμα Binomial

Έστω  $y_1, \dots, y_n$  είναι τυχαίο δείγμα από την  $Binomial(N_i, \theta)$ . Για τον υπολογισμό μιας power εκ των προτέρων κατανομής για την Binomial κατανομή, θέτουμε ως αρχική εκ των προτέρων  $\pi(\theta) \sim Beta(\alpha_o, \beta_o)$ . Από τα παραπάνω προκύπτει

$$\begin{aligned} \pi(\theta | \mathbf{y}_o, \delta) &\propto [L(\theta | \mathbf{y}_o)]^\delta \pi(\theta) \\ &\propto \left[ \theta^{\sum y_{oi}} (1 - \theta)^{\sum N_{oi} - \sum y_{oi}} \right]^\delta \frac{\Gamma(\alpha_o + \beta_o)}{\Gamma(\alpha_o) \Gamma(\beta_o)} \theta^{\alpha_o - 1} (1 - \theta)^{\beta_o - 1} \\ &\propto \theta^{N_o \delta \bar{y}_o + \alpha_o - 1} (1 - \theta)^{N_o \delta - N_o \delta \bar{y}_o + \beta_o - 1} \end{aligned}$$

Που είναι μια

$$Beta(N_o \delta \bar{y}_o + \alpha_o, N_o \delta (1 - \bar{y}_o) + \beta_o)$$

όπου  $N_o = \sum N_{oi}$  και  $\bar{y}_o = \frac{1}{N_o} \sum y_{oi}$  η δειγματική διωνυμική αναλογία.

Θέτοντας  $\alpha_o = \beta_o = 0$  και  $\delta = \frac{c}{N_o}$  παίρνουμε την υπό περίπτωση όπου η power prior θα είναι μια  $Beta(c \bar{y}_o, c(1 - \bar{y}_o))$ .

Στην παραπάνω εκ των προτέρων αν επιπλέον θέσουμε  $\bar{y}_o = \frac{1}{2}$  τότε θα έχουμε μια  $Beta(\frac{c}{2}, \frac{c}{2})$  τις οποίες γνωστή υπό περίπτωση είναι η  $Beta(\frac{1}{2}, \frac{1}{2})$ , που είναι η *Jefferys* εκ των προτέρων κατανομή στην συγκεκριμένη περίπτωση.

## Κεφάλαιο 2

# Έλεγχος Υποθέσεων για Ένα Δείγμα

### 2.1 Μειονεκτήματα της Κλασικής προσέγγισης

Στο κεφάλαιο αυτό θα ασχοληθούμε με τον έλεγχο υποθέσεων. Δηλαδή αναζητάμε έναν τρόπο για να αποφασίσουμε το κατά πόσο μια καθορισμένη τιμή της παραμέτρου είναι αποδεκτή με βάση τις παρατηρήσεις και την εκ των προτέρων πεποίθησή μας για την παράμετρο.

Στην Κλασική στατιστική ο έλεγχος υποθέσεων βασίζεται στα θεμελιώδη λήμματα των Neyman - Pearson και Fisher (Carlin & Louis, 1996, σελ. 45)[12]. Τα στοιχεία ενός κλασσικού ελέγχου είναι τα εξής (Ρούσσας, 1976, σελ. 11-15)[2]:

- (1) Ορίζεται η μηδενική υπόθεση ( $H_0$ ), δηλαδή καθορίζεται ένα υποσύνολο  $\omega$  που μπορεί να περιέχει την αληθινή τιμή για την παράμετρο  $\theta$ .  $H_0 : \theta \in \omega$
- (2) Ορίζεται η εναλλακτική υπόθεση ( $H_1$ ), δηλαδή καθορίζεται ένα υποσύνολο  $\omega^c$  (συμπληρωματικό του  $\omega$  σε σχέση με το  $\Omega$ , όπου  $\Omega$  ο παραμετρικός χώρος) ως το υποσύνολο που μπορεί να περιέχει την αληθινή τιμή για την παράμετρο  $\theta$ .  $H_1 : \theta \in \omega^c$
- (3) Ορίζεται ο στατιστικός έλεγχος από το δείγμα ( $Y$ ), δηλαδή μια ελεγχο-συνάρτηση  $T(Y)$  για τον έλεγχο της μηδενικής έναντι της εναλλακτικής υπόθεσης.

- (4) Ορίζεται η περιοχή απόρριψης  $R$  της μηδενικής υπόθεσης  $H_0$ .
- (5) Υπολογίζεται το παρατηρούμενο επίπεδο σημαντικότητας (observed level of significance or p-value) που είναι η πιθανότητα η στατιστική συνάρτηση  $T(\mathbf{Y})$  του ελέγχου να πάρει μια τιμή τόσο ακραία ως προς την  $H_0$  ή περισσότερο ακραία από αυτήν που πήρε για το συγκεκριμένο δείγμα, υπό την  $H_0$ .

$$p - value = P[T(\mathbf{Y}) \text{ πιο ακραία από την } T(\mathbf{Y}_{obs}) | \theta, H_0]$$

- (6) Γίνεται σύγκριση με το επίπεδο στατιστικής σημαντικότητας που ορίζει ο ερευνητής και εξάγονται συμπεράσματα. Εάν το p-value είναι μικρότερο από το καθορισμένο σφάλμα Τύπου I τότε η  $H_0$  απορρίπτεται, αλλιώς όχι.

Παρ' όλο όμως που ο κλασικός έλεγχος υποθέσεων είναι ευρέως διαδεδομένος στην στατιστική βιβλιογραφία, παρουσιάζει και αρκετά μειονεκτήματα (Carlin & Louis 1996, σελ.45-46)[12].

Πρώτον, μπορεί μόνο να εφαρμοστεί όταν οι ζητούμενες υποθέσεις είναι «ένθετες» (nested), δηλαδή η  $H_0$  απαρτίζει μια απλούστευση της  $H_1$  (για παράδειγμα  $H_0 : \theta = 0$  vs  $H_1 : \theta \neq 0$ ). Φυσικά τα περισσότερα προβλήματα περιλαμβάνουν περιπτώσεις «μη-ένθετων» υποθέσεων .

Δεύτερη δυσκολία των ελέγχων αυτών είναι ότι μπορούν μόνο να μας παρέχουν μια απόδειξη εναντίον της μηδενικής υπόθεσης και όχι της εναλλακτικής. Εάν το p-value είναι αρκετά μικρό μπορούμε να απορρίψουμε την  $H_0$  ενώ αν ισχυεί το αντίθετο δεν μπορούμε να πούμε ότι δεχόμαστε την  $H_0$  παρά μόνο ότι αποτυγχάνουμε να την απορρίψουμε.

Τρίτον, το p-value από μόνο του δεν προσφέρει μια άμεση ερμηνεία σαν έναν βαθμό της απόδειξης παρά μόνο ως αναλογία - πιθανότητα · έπειτα από συνεχόμενη και μακροχρόνια επανάληψη του ίδιου πειράματος.

Τέταρτο σημαντικό μειονέκτημα των κλασικών ελέγχων σημαντικότητας είναι ότι όχι μόνο εξαρτάται από τα παρατηρούμενα δεδομένα μας, αλλά και από την

συνολική δειγματοληπτική πιθανότητα μη-παρατηρήσιμων δεδομένων. Αυτό έχει σαν αποτέλεσμα ότι δύο πειράματα με πανομοιότυπες πιθανοφάνειες μπορούν να οδηγήσουν σε διαφορετικά αποτελέσματα (βλ. παράδειγμα 1.2, Carlin & Louis 1996, σελ.3)[12] γεγονός που έρχεται σε αντίθεση με την αρχή της πιθανοφάνειας (βλ. παράρτημα Α).

Φυσικά δεν πρέπει να ξεχνάμε ότι το  $p$ -value επηρεάζεται από το μέγεθος του δείγματος. Όσο μεγαλύτερο δείγμα, τόσο μικρότερη η τιμή του και άρα απορρίπτουμε την  $H_0$ . Ενώ όσο μεγαλύτερη η τιμή του, δεν μπορούμε να απορρίψουμε την  $H_0$ . Επίσης οι ποσότητες  $\alpha$  και  $\beta$  (σφάλματα τύπου I & II αντίστοιχα) αυξομειώνονται σε σχέση με το μέγεθος του δείγματος  $n$  (Πανάρετος & Ξεκαλάκη 2000, σελ. 369-372)[1].

- Για δεδομένο  $\alpha$ , το  $\beta$  μπορεί να ελαττώνεται με αύξηση του  $n$ .
- Για δεδομένο  $n$ , όσο μεγαλώνουμε το μέγεθος της περιοχής απόρριψης (και άρα το  $\alpha$ ) θα ελαττώνεται το  $\beta$  (αντιστρόφος ανάλογα).
- Η τιμή του  $\beta$  εξαρτάται από την  $H_1$ .

Γενικά δεν υπάρχει αποδεκτός κανόνας που να οδηγεί στην επιλογή επίπεδου στατιστικής σημαντικότητας στο πρόβλημα ελέγχου στατιστικών υποθέσεων στην κλασική στατιστική.

## 2.2 Παράγοντας Bayes

Όλα τα προηγούμενα μειονεκτήματα οδηγούν στη διερεύνηση των δυνατοτήτων της στατιστικής κατά Bayes. Η Μπευζιανή προσέγγιση στον έλεγχο υποθέσεων, που αναπτύχθηκε από τον Jeffreys (1935, 1961)[30],[31] είναι πιο απλή, συμβατή με την αρχή της πιθανοφάνειας και αποφεύγονται τα προβλήματα που αναφέρθηκαν στην προηγούμενη παράγραφο. Ένα από τα πλεονέκτηματα της μεθόδου αυτής είναι ότι δεν υπάρχει περιορισμός στο πλήθος των υποθέσεων που μπορούν να εξεταστούν ταυτόχρονα. Επιπλέον δεν είναι υποχρεωτικό τα υπό εξέταση μοντέλα



να είναι «ένθετα».

Η βασική ιδέα είναι ότι η εκ των προτέρων και η εκ των υστέρων πληροφορία συνδυάζεται σ'ένα σχετικό λόγο πιθανοτήτων που ονομάζεται παράγοντας Bayes (Bayes Factor). Ο παράγοντας αυτός μπορεί να χρησιμοποιηθεί ως σχετικό μέτρο για την σύγκριση δύο μοντέλων.

## Ορισμός

Έστω ότι έχουμε παρατηρήσει κάποια δεδομένα  $\mathbf{y}$ . Θέλουμε να εξετάσουμε δύο υποψήφια παραμετρικά μοντέλα  $M_1$  και  $M_2$  (υπό τις αντίστοιχες υποθέσεις  $H_1$  και  $H_2$ ) με παραμέτρους  $\boldsymbol{\theta}_1$  και  $\boldsymbol{\theta}_2$  αντίστοιχα. Υπό τις εκ των προτέρων πυκνότητες  $\pi_i(\boldsymbol{\theta}_i)$ , η περιθώρια κατανομή των  $y$  είναι

$$p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \quad (2.1)$$

με  $i = 1, 2$ . Από το θεώρημα του Bayes αποκτάμε τις εκ των υστέρων πιθανότητες  $P(M_1|\mathbf{y})$  και  $P(M_2|\mathbf{y}) = 1 - P(M_1|\mathbf{y})$  των δύο μοντέλων. Οπότε για να εξάγουμε συμπεράσματα υπέρ ενός μοντέλου έναντι ενός άλλου, χρησιμοποιούμε είτε τον λόγο των εκ των υστέρων σχετικών πιθανοτήτων (posterior odds)

$$PO_{12} = \frac{P(M_1|\mathbf{y})}{P(M_2|\mathbf{y})},$$

είτε υπολογίζουμε τον παράγοντα Bayes

$$\begin{aligned} BF_{12} &= \frac{\text{posterior odds}}{\text{prior odds}} = \frac{\frac{P(M_1|\mathbf{y})}{P(M_2|\mathbf{y})}}{\frac{P(M_1)}{P(M_2)}} \\ &= \frac{\left[ \frac{p(\mathbf{y}|M_1)P(M_1)}{p(\mathbf{y})} \right] / \left[ \frac{p(\mathbf{y}|M_2)P(M_2)}{p(\mathbf{y})} \right]}{P(M_1)/P(M_2)} \\ &= \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \end{aligned}$$

Η προσοχή μας επικεντρώνεται στον παράγοντα Bayes επειδή αποτελεί ένα μέτρο για την βαρύτητα της πληροφορίας η οποία περιλαμβάνεται στα δεδομένα υπέρ ενός μοντέλου έναντι ενός άλλου (Carlin & Louis 1996, σελ.47)[12].

### Παρατηρήσεις

- Στην βιβλιογραφία ο παράγοντας Bayes συμβολίζεται είτε με  $B$  ή  $BF$ , είτε φέροντας τους δείκτες των υποθέσεων που εξετάζει δηλαδή  $B_{ij}$  ή  $BF_{ij}$ .
- Εάν τα δύο μοντέλα έχουν την ίδια εκ των προτέρων πιθανότητα  $P(M_1) = P(M_2)$  τότε

$$PO_{12} = \frac{P(M_1|\mathbf{y})}{P(M_2|\mathbf{y})} = BF_{12}$$

- Στην περίπτωση που τα δύο μοντέλα μοιράζονται την ίδια παραμετροποίηση και οι δύο υποθέσεις είναι απλές, δηλ. αφορούν μεμονωμένα σημεία (π.χ.  $M_1 : \theta = \theta_1$  και  $M_2 : \theta = \theta_2$ ) τότε

$$BF_{12} = \frac{f(\mathbf{y}|\theta = \theta_1)}{f(\mathbf{y}|\theta = \theta_2)}$$

που δεν είναι παρά ο λόγος των πιθανοφανειών των δύο μοντέλων.

- Στην περίπτωση που το ένα μοντέλο είναι μια απλή υπόθεση ενώ το άλλο μοντέλο μια σύνθετη υπόθεση (π.χ.  $M_1 : \theta = \theta_0$  και  $M_2 : \theta \neq \theta_0$ ) τότε

$$BF_{12} = \frac{f(\mathbf{y}|\theta_0)}{\int_{\Theta_2} f(\mathbf{y}|\theta)\pi(\theta)d\theta}$$

- Στην περίπτωση που και τα δύο μοντέλα αποτελούνται από σύνθετες υποθέσεις (π.χ.  $M_1 : \theta \in \Theta_1$  και  $M_2 : \theta \in \Theta_2$ ), τότε έχουμε

$$BF_{12} = \frac{\int_{\Theta_1} f(\mathbf{y}|\theta)\pi_1(\theta)d\theta}{\int_{\Theta_2} f(\mathbf{y}|\theta)\pi_2(\theta)d\theta}$$

που είναι ο λόγος των «σταθμισμένων» πιθανοφανειών υπό τις υποθέσεις  $\Theta_1$  και  $\Theta_2$  (Bernardo & Smith 1995, σελ. 392)[11].

## Ερμηνεία

Ο παράγοντας Bayes είναι ένα συνοπτικό στατιστικό μέτρο που καταδεικνύει μια ένδειξη υπέρ ενός μοντέλου, έναντι κάποιου άλλου. Ο Jeffreys (1961) πρότεινε μια πιθανή ερμηνεία του  $BF$  στην  $\log$  κλίμακα. Ο παρακάτω πίνακας μας δίνει τα αποτελέσματα.

$\log(BF_{10})$	$BF_{10}$	Ένδειξη υπέρ της $H_1$ , έναντι της $H_0$
0 - 1/2	1 - 3.2	Όχι άξια αναφοράς
1/2 - 1	3.2 - 10	Ουσιώδης
1 - 2	10 - 100	Ισχυρή
> 2	> 100	Οριστικά αμετάκλητη

Πίνακας 2.1: Ερμηνεία του παράγοντα Bayes σύμφωνα με τον Jeffrey (1961).

Όμως οι Kass & Raftery (1995) [34] έδωσαν άλλη ερμηνεία στον παράγοντα Bayes πολλαπλασιάζοντας τον επι δύο, ώστε να έχει την ίδια κλίμακα με τις γνωστές στατιστικές συναρτήσεις : την συνάρτηση απόκλισης (deviance) και τον έλεγχο των λόγων πιθανοφάνειας (likelihood ratio test).

$2\log_e(BF_{10})$	$BF_{10}$	Ένδειξη υπέρ της $H_1$ , έναντι της $H_0$
0 - 2	1 - 3	Όχι άξια αναφοράς
2 - 6	3 - 20	Θετική
6 - 10	20 - 150	Ισχυρή
> 10	> 150	Πολύ ισχυρή

Πίνακας 2.2: Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass & Raftery (1995).

Παρ' όλο που ο παράγοντας Bayes αποτελεί σημαντικό εργαλείο για την στατιστική ανάλυση, η χρήση του δεν είναι ελεύθερη αμφισβήτησης διότι επηρεάζεται από την μεταβλητότητα της εκ των προτέρων κατανομής, όπως θα δούμε αναλυτικά στην επόμενη παράγραφο. Επίσης παρουσιάζονται αρκετές δυσκολίες στους υπολογισμούς, όταν χρησιμοποιούνται ακατάλληλες εκ των προτέρων κατανομές, όπως υπέδειξε ο O'Hagan (1995)[45]. Συχνά χρησιμοποιούνται διάφορες παραλλαγές του παράγοντα Bayes. Από τις πιο γνωστές είναι ο ψευδοπαράγοντας Bayes που προτάθηκε από τους Geisser & Eddy (1979)[20], Gelfand et al.(1992)[22] και Gelfand & Dey (1994)[21], ο εκ των υστέρων παράγοντας Bayes που προτά-

θηκε από τον Aitkin (1991)[5], ο ενδογενής (intrinsic) παράγοντας Bayes που προτάθηκε από τους Berger & Pericchi (1996)[10] και ο κλασματικός (fractional) παράγοντας Bayes που προτάθηκε από τους O'Hagan (1995)[45] και De Santis & Spezzaferri (1997)[18]. Ωστόσο αυτές οι τεχνικές είναι ασταθείς όταν το μέγεθος των δειγμάτων είναι μικρό (Giampaoli, 2004)[24].

### 2.3 Το παράδοξο του Lindley

Το παράδοξο του Lindley (Lindley's Paradox) αναφέρεται στην κατάσταση όπου η Μπευζιανή και η Κλασσική προσέγγιση δίνουν αντίθετα αποτελέσματα στους ελέγχους υποθέσεων. Αρχικά ο Lindley (1957) [39] αναφέρθηκε στην επίδραση του μεγέθους του δείγματος  $n$ , ενώ ο Bartlett (1957) [9] παρατήρησε το ίδιο για προκαθορισμένες επιλογές της εκ των προτέρων κατανομής.

Έστω ότι έχουμε ένα τυχαίο δείγμα  $y_i \sim N(\theta, \sigma^2)$  με άγνωστο μέσο  $\theta$  και γνωστή διασπορά  $\sigma^2$ . Θέλουμε να ελέγξουμε την υπόθεση ότι  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta \neq \theta_0$ . Επιλέγουμε μια μη-μηδενική εκ των προτέρων πιθανότητα  $P(M_0) = \pi_0$  για την  $H_0$  (αντίστοιχα  $P(M_1) = 1 - \pi_0$  για την  $H_1$ ) και μια σχεδόν επίπεδη πυκνότητα πιθανότητας (fairly flat propability density  $\pi(\theta)$ ) ως εκ των προτέρων κατανομή της παραμέτρου  $\theta$ . Θέτουμε  $\pi(\theta) \sim N(\mu_\theta, \tau^2)$ , όπου η διασπορά  $\tau^2 \gg \sigma^2$ . Έστω ότι παρατηρούμε μια τιμή  $Y = y$ , όπου το  $y$  είναι αρκετά  $\sigma$  μακριά από το  $\theta_0$ . Μια τιμή τόσο μακριά από το  $\theta_0$  είναι πολύ απίθανη υπό την  $H_0$  και έτσι η  $H_0$  απορρίπτεται εάν χρησιμοποιήσουμε έναν τυπικό έλεγχο βασιζόμενο στο στατιστικό έλεγχο  $(y - \theta_0)/\sigma$ , δηλ. το t-test με  $\alpha = 5\%$ . Ακόμα θέτουμε  $\tau^2 = \frac{C}{n}\sigma^2$ . Αλλά από την στιγμή που η  $\tau^2 \gg \sigma^2$ , οι τιμές του  $\theta$  με υψηλή πιθανοφάνεια (αυτές εντός αρκετών  $\sigma$  μακριά από το  $y$ ) θα δώσουν πολύ μικρές εκ των προτέρων πιθανότητες μέσω της  $\pi(\theta)$  και επομένως η συνολική πιθανοφάνεια για το εναλλακτικό μοντέλο θα είναι:

$$f(y|M_1) = \exp\left\{+\frac{(y - \bar{y})^2}{2\sigma^2}\right\} \exp\left\{+\frac{(\bar{y} - \mu_\theta)^2}{\frac{2\sigma^2}{n}(1 + C)}\right\} (\sqrt{1 + C})^{-1}(\sigma\sqrt{2\pi})^{-1}$$

που είναι πολύ μικρότερη από την πιθανοφάνεια της  $H_0$

$$f(y|M_0) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(y - \theta_0)^2}{2\sigma^2} \right\}$$

Ο παράγοντας Bayes δίνεται από την σχέση

$$\begin{aligned} BF_{10} &= \frac{f(y|M_1)}{f(y|M_0)} = \\ &= \exp \left\{ \frac{n}{2\sigma^2} \left[ \frac{(y - \theta_0)^2}{n} + \frac{(y - \bar{y})^2}{n} + \frac{(\bar{y} - \mu_0)^2}{1 + C} \right] \right\} (\sqrt{1 + C})^{-1} \end{aligned}$$

Επειδή  $\tau^2 \gg \sigma^2$  και  $\tau^2 = \frac{C}{n}\sigma^2$  τότε το  $C$  θα είναι πολύ μεγάλο και κατ' επέκταση ο  $BF$  θα μικραίνει, υποστηρίζοντας την  $H_0$ . Αν πάρουμε και την οριακή περίπτωση όπου  $C \rightarrow \infty$ , τότε  $BF \rightarrow 0$ . Δηλαδή όταν η εκ των προτέρων διασπορά  $\tau^2$  είναι πολύ μεγαλύτερη από αυτήν του δείγματος  $\sigma^2$ , η Μπευζιανή ανάλυση οδηγεί σε ένδειξη υπέρ της  $H_0$  σε αντίθεση με την Κλασσική στατιστική, που το στατιστικό  $(y - \theta_0)/\sigma$  δείχνει ισχυρή ένδειξη εναντίον της  $H_0$  (Shafer 1982)[48].

Συνεπώς σε οποιαδήποτε σύγκριση μοντέλων, ο παράγοντας εξαρτάται από την εκ των προτέρων κατανομή καθοριζόμενη από τις παραμέτρους κάθε μοντέλου. Αυτό είναι το παράδοξο το οποίο συζητήθηκε λεπτομερώς από τον Bartlett (1957)[9] ενώ πιο πρόσφατες αναφορές έχουν γίνει από τους Smith (1965), Bernardo (1980), Shafer (1982)[48] κ.α.

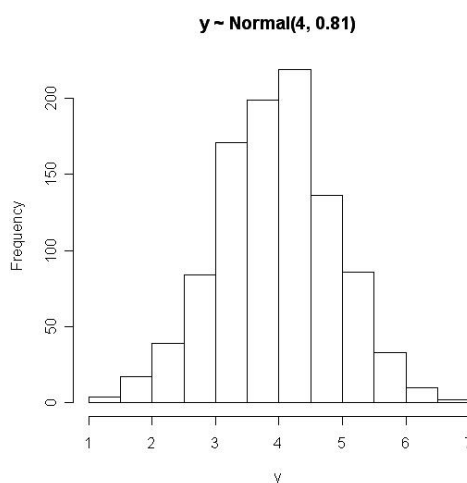
### 2.3.1 Παράδειγμα με προσομοιωμένα δεδομένα

Προσομοιώνουμε 1000 παρατηρήσεις ( $y_i$ ) που προέρχονται από κανονική κατανομή με μέση τιμή 4 και τυπική απόκλιση 0.9 ( $\sigma^2 = 0.81$ ). Στο σχήμα (2.1) βλέπουμε το ιστόγραμμα των παρατηρήσεων. Ελέγχουμε αν η μέση τιμή του δείγματος διαφέρει σημαντικά από την τιμή  $\mu_0 = 4$  δηλ.

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0$$

που εκφράζεται από τα μοντέλα

$$M_0 : y_i \sim \text{Normal}(\mu_0, \sigma^2) \quad \text{και} \quad M_1 : y_i \sim \text{Normal}(\mu, \sigma^2)$$

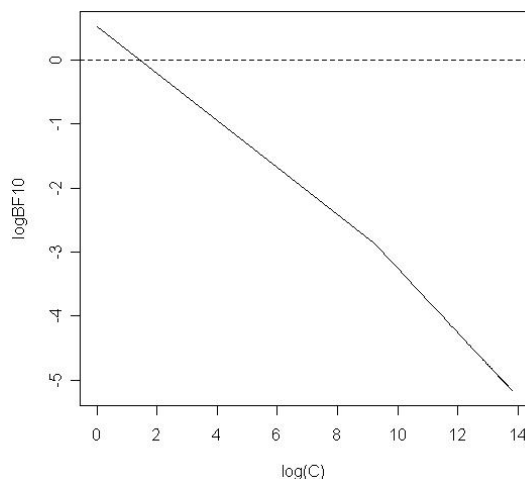


Σχήμα 2.1: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από  $N(4, 0.81)$

- Για την άγνωστη παράμετρο  $\mu$  επιλέγουμε την εκ των προτέρων κατανομή  $\mu \sim N(\mu_\theta, \frac{C}{n}\sigma^2)$ . Θεωρούμε ότι  $\mu_\theta = 4$ , διότι εφόσον πιστεύουμε ότι η μηδενική υπόθεση αξίζει να ελεγχθεί, τότε υπάρχει και εκ των προτέρων πεποίθηση ότι αυτή είναι η πραγματική της τιμή. Ακόμα θέτουμε  $C = 1000$  (ίσο με το μέγεθος του δείγματος) ως αρκετά μεγάλη τιμή, ώστε να εξασφαλίσουμε μικρή εκ των προτέρων πληροφορία (μεγάλη εκ των προτέρων διασπορά της παραμέτρου  $\mu$ ). Συνεπώς  $\mu \sim N(4, 0.81)$  (Empirical Bayes approach). Με την βοήθεια της  $R$  βρίσκουμε  $BF_{10} = 0.18 < 1$ , δηλαδή υποστηρίζεται η μηδενική υπόθεση, που σημαίνει ότι η μέση τιμή δεν διαφέρει από το 4.

- Στο σχήμα (2.2) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\log C$ . Από το γράφημα επιβεβαιώνεται το παράδοξο του Lindley, αφού καθώς το  $C$  αυξάνεται, διαπιστώνεται ότι ο παράγοντας Bayes μειώνεται, υποστηρίζοντας την  $H_0$ . Οι οριζόντιες γραμμές των γραφημάτων, δηλώνουν τα επίπεδα ερμηνείας σύμφωνα με τους Kass & Raftery. Για θετικές τιμές του λογαρίθμου το μηδενικό μοντέλο απορρίπτεται, ενώ για αρνητικές τιμές απορρίπτεται η εναλλακτική υπόθεση. Για τιμές του λογαρίθμου

μεταξύ  $(-1, 1)$  η διαφορά δεν είναι άξια αναφοράς. Από τα γραφήματα συμπεραίνουμε ότι η εναλλακτική υπόθεση απορρίπτεται για όλες τις τιμές του  $C$ .



Σχήμα 2.2: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\log C$ , ( $1 < C < 10^6$ ) (Παράδειγμα 2.3.1).

## 2.4 Έλεγχος υποθέσεων για ένα δείγμα από κατανομή Poisson

Θεωρούμε τυχαίο δείγμα από την κατανομή  $Poisson(\lambda)$ . Θέλουμε να ελέγξουμε την ισότητα του ρυθμού του δείγματος με μια συγκεκριμένη τιμή  $\lambda_0$ .

$$H_0 : \lambda = \lambda_0 \quad vs \quad H_1 : \lambda \neq \lambda_0$$

Τα μοντέλα που αντιπροσωπεύουν τις παραπάνω υποθέσεις είναι αντίστοιχα

$$M_0 : y_i \sim Poisson(\lambda_0) \quad \text{και} \quad M_1 : y_i \sim Poisson(\lambda)$$

Για το μοντέλο  $M_0$ , όπου η παράμετρος  $\lambda_0$  είναι γνωστή και όχι προς εκτίμηση, το ολοκλήρωμα (2.1) (δηλ. η περιθώρια κατανομή των  $y_i$ ) ισούται με την συνάρτηση πιθανοφάνειας

$$f(\mathbf{y}|M_0) = L(\lambda_0; \mathbf{y}) \propto \exp\{-n\lambda_0\} \lambda_0^{\sum y_i}$$

Για το μοντέλο  $M_1$ , επειδή η παράμετρος  $\lambda$  είναι άγνωστη και άρα προς εκτίμηση, υπολογίζεται αρχικά η συνάρτηση πιθανοφάνειας  $L(\lambda; \mathbf{y}) \propto \exp\{-n\lambda\} \lambda^{\sum y_i}$ .

Στη συνέχεια υποθέτουμε ότι η εκ των προτέρων πεποίθησή μας για την παράμετρο  $\lambda$  μπορεί να εκφραστεί με την βοήθεια μιας Gamma κατανομής. Από το θεώρημα του Bayes υπολογίζεται αρχικά η εκ των υστέρων κατανομή και στη συνέχεια υπολογίζεται το ολοκλήρωμα (2.1) που ουσιαστικά αποτελεί την σταθερά κανονικοποίησης του θεωρήματος του Bayes. Οπότε :

$$f(\mathbf{y}|M_1) = \int f(\mathbf{y}|\lambda)f(\lambda)d(\lambda) = \frac{f(\lambda)f(\mathbf{y}|\lambda)}{f(\lambda|\mathbf{y})} = \frac{\Gamma(p+n\bar{y})q^p}{\Gamma(p)(q+n)^{p+n\bar{y}}} \frac{1}{\prod y_i!}$$

### Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$\begin{aligned} BF_{10} &= \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_0)} = \frac{\sum f(\mathbf{y}|\lambda)f(\lambda)}{L(\lambda_o; \mathbf{y})} \\ &= \frac{\Gamma(p+n\bar{y})q^p / \Gamma(p)(q+n)^{p+n\bar{y}} \left[ \frac{1}{\prod y_i!} \right]}{\exp\{-n\lambda_o\} \lambda_o^{\sum y_i} / \left[ \frac{1}{\prod y_i!} \right]} \\ &= \frac{\Gamma(p+n\bar{y})q^p}{\Gamma(p)(q+n)^{p+n\bar{y}} \exp\{-n\lambda_o\} \lambda_o^{n\bar{y}}} \end{aligned}$$

Ο παράγοντας Bayes εξαρτάται από τις εκ των προτέρων παραμέτρους, το μέγεθος του δείγματος και τον μέσο των δεδομένων. Στις περιπτώσεις όπου το μέγεθος του δείγματος είναι πολύ μεγάλο ( $n \rightarrow \infty$ ) τότε ο παράγοντας Bayes τείνει στο μηδέν, ( $BF \rightarrow 0$  επειδή ο παρονομαστής αυξάνεται πολύ πιο γρήγορα από τον αριθμητή) δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_0$ . Ακόμα όταν η εκ των προτέρων διασπορά  $Var(\lambda) = \frac{p}{q^2} = \frac{E(\lambda)}{q} \rightarrow \infty$ , δηλαδή όταν η παράμετρος  $q \rightarrow 0$ , τότε ο  $BF$  ελλατώνεται δίνοντας επίσης πολύ ισχυρή ένδειξη υπέρ της  $H_0$ .



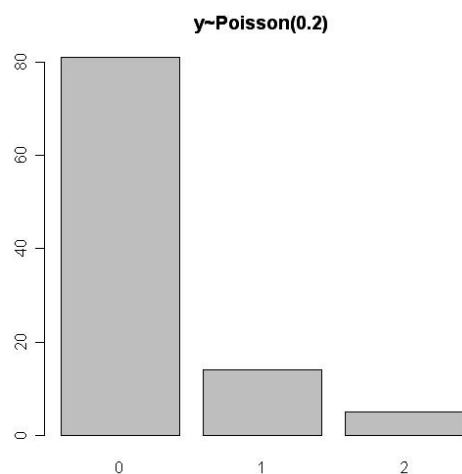
### 2.4.1 Παράδειγμα με προσομοιωμένα δεδομένα

Προσομειώνουμε 100 παρατηρήσεις ( $y_i$ ) από κατανομή  $Poisson(0.2)$ . Στο σχήμα (2.3) βλέπουμε το ραβδόγραμμα των παρατηρήσεων. Ελέγχουμε αν η μέση τιμή του δείγματος διαφέρει σημαντικά από τις τιμές  $\lambda_o = 0.1$ ,  $\lambda_o = 0.2$ ,  $\lambda_o = 0.3$  και  $\lambda_o = 0.4$  δηλ.

$$H_o : \lambda = \lambda_o \quad vs \quad H_1 : \lambda \neq \lambda_o$$

που εκφράζεται από τα μοντέλα

$$M_o : y_i \sim Poisson(\lambda_o) \quad και \quad M_1 : y_i \sim Poisson(\lambda)$$



Σχήμα 2.3: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από  $Poisson(0.2)$

- Βασιζόμενοι στην power εκ των προτέρων κατανομή  $Gamma(\bar{y}_o, 1)$ , θα επιλέξουμε δύο υπό περιπτώσεις της για την άγνωστη παράμετρο  $\lambda$ .

- ◇ Για  $\bar{y}_o = 1$  προκύπτει η  $Gamma(1, 1)$  με  $E(\lambda) = Var(\lambda) = 1$ .

- ◇ Για  $\bar{y}_o = \lambda_o^*$  προκύπτει η  $Gamma(\lambda_o^*, 1)$  με  $E(\lambda) = Var(\lambda) = \lambda_o^*$ . Θέτουμε  $\bar{y}_o = \lambda_o^* = \lambda_o$ , διότι εφόσον πιστεύουμε ότι η μηδενική υπόθεση αξίζει να ελεγχθεί, τότε υπάρχει εκ των προτέρων πεποίθηση ότι αυτή είναι η πραγματική της τιμή.

- Με την βοήθεια της  $R$  προκύπτουν οι πίνακες (2.3) και (2.4). Από τους πίνακες διαπιστώνουμε ότι η εκ των προτέρων κατανομή  $Gamma(\lambda_o, 1)$  δίνει μικρότερες

τιμές στον παράγοντα Bayes απ' ότι η εκ των προτέρων κατανομή  $Gamma(1, 1)$ . Όπως θα δούμε παρακάτω αυτό οφείλεται στην μεγαλύτερη διασπορά ( $\lambda_0 > 1$ ). Ωστόσο και οι δύο εκ των προτέρων κατανομές δίνουν παρόμοια αποτελέσματα.

$\lambda_0$	0.1	0.2	0.3	0.4
$BF_{10}$	106.57	0.14	0.18	4.046
$\log(BF)_{10}$	4.669	-1.967	-1.698	1.398
Ένδειξη εναντίον της	$H_0$	$H_1$	$H_1$	$H_0$
	Ισχυρή	Θετική	Θετική	Θετική

Πίνακας 2.3: Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την  $Gamma(1, 1)$ .

$\lambda_0$	0.1	0.2	0.3	0.4
$BF_{10}$	40.75	0.096	0.167	4.298
$\log(BF)_{10}$	3.708	-2.345	-1.79	1.46
Ένδειξη εναντίον της	$H_0$	$H_1$	$H_1$	$H_0$
	Ισχυρή	Θετική	Θετική	Θετική

Πίνακας 2.4: Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την  $Gamma(\lambda_0, 1)$ .

- Στο σχήμα (2.4) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\lambda_0$ . Οι οριζόντιες γραμμές των σχημάτων, δηλώνουν τα επίπεδα ερμηνείας σύμφωνα με τους Kass & Raftery. Για αρνητικές τιμές του λογαρίθμου το εναλλακτικό μοντέλο απορρίπτεται, ενώ για θετικές τιμές μεγαλύτερες του ένα η μηδενική υπόθεση απορρίπτεται. Για τιμές μεταξύ  $(-1, 1)$  η διαφορά δεν είναι άξια αναφοράς. Έχοντας επιλέξει εκ των προτέρων κατανομή την  $Gamma(1, 1)$  αποδεχόμαστε την μηδενική υπόθεση για τιμές του  $\lambda_0 \in (0.149, 0.3622)$  ενώ έχοντας επιλέξει εκ των προτέρων κατανομή την  $Gamma(\lambda_0, 1)$ , αποδεχόμαστε την μηδενική υπόθεση για  $\lambda_0 \in (0.149, 0.3622)$ . Αναλυτικότερα βλ. πίνακες (2.5) και (2.6).

- Στο σχήμα (2.5α) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές τις παραμέτρου  $\alpha$ , στην γενική μορφή της εκ των προτέρων κατανομής  $Gamma(\alpha, 1)$ , υπό την πραγματική τιμή του  $\lambda_0 = 0.2$ . Για την κατανομή  $Gamma(\alpha, 1)$ , με  $E(\lambda) = Var(\lambda) = \alpha$ , καθώς το  $\alpha \rightarrow \infty$ , ο  $BF$  ελλατώνεται, δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_0$ . Υπό την πραγματική

$\log_e(BF_{10})$	$\lambda_o \in$	Ένδειξη υπέρ της $H_1$ , έναντι της $H_0$
αρνητικές τιμές	(0.149 , 0.3622)	Αποδεχόμαστε την $H_0$
0 - 1	(0.134, 0.149) $\cup$ (0.3622, 0.389)	Όχι άξια αναφοράς
1 - 3	(0.1134, 0.134) $\cup$ (0.389, 0.4375)	Θετική
3 - 5	(0.0977, 0.1134) $\cup$ (0.4375, 0.4795)	Ισχυρή
> 5	$(-\infty, 0.0977) \cup (0.4795, +\infty)$	Πολύ ισχυρή

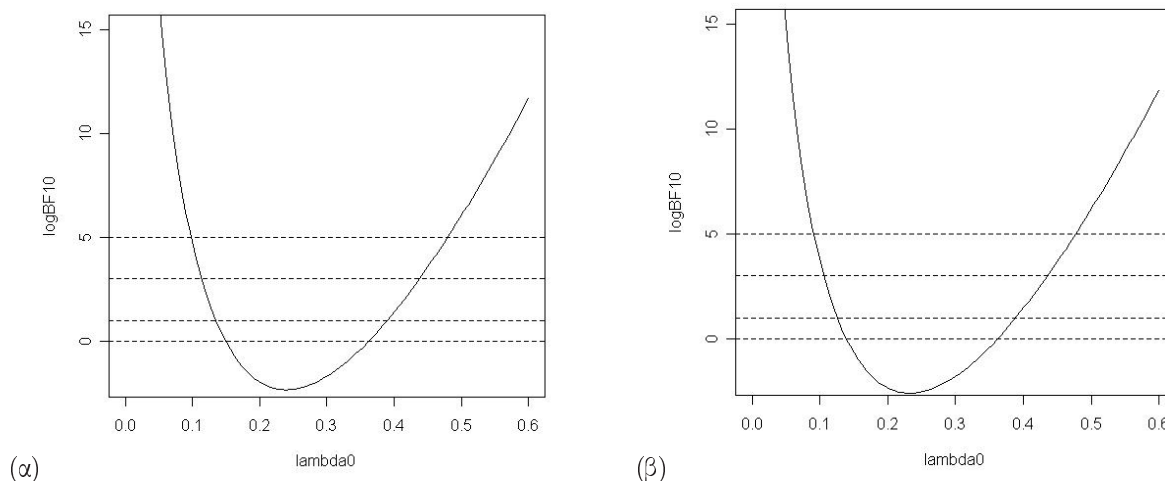
Πίνακας 2.5: Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass & Raftery για την  $Gamma(1, 1)$ .

$\log_e(BF_{10})$	$\lambda_o \in$	Ένδειξη υπέρ της $H_1$ , έναντι της $H_0$
αρνητικές τιμές	(0.1389, 0.3618)	Αποδεχόμαστε την $H_0$
0 - 1	(0.1256, 0.1389) $\cup$ (0.3618, 0.3886)	Όχι άξια αναφοράς
1 - 3	(0.1056, 0.1256) $\cup$ (0.3886, 0.4354)	Θετική
3 - 5	(0.091, 0.1056) $\cup$ (0.4354, 0.477)	Ισχυρή
> 5	$(-\infty, 0.091) \cup (0.477, +\infty)$	Πολύ ισχυρή

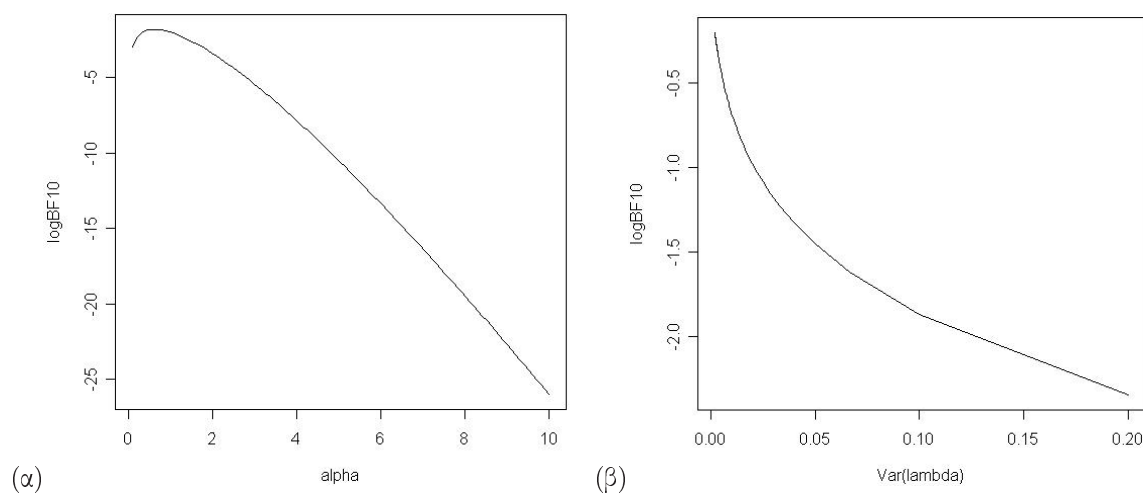
Πίνακας 2.6: Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass & Raftery για την  $Gamma(\lambda_o, 1)$ .

τιμή του  $\lambda_o = 0.2$ , δεχόμαστε την μηδενική υπόθεση για όλες τις τιμές του  $\alpha$ .

- Στο σχήμα (2.5β) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\delta \in [0, 1]$ , στην γενική μορφή της εκ των προτέρων κατανομής  $Gamma(\delta n \lambda_o, \delta n)$ , υπό την πραγματική τιμή του  $\lambda_o = 0.2$ . Για την κατανομή  $Gamma(\delta n \lambda_o, \delta n)$ , με  $E(\lambda) = \lambda_o$  και  $Var(\lambda) = \frac{\lambda_o}{\delta n}$ , καθώς το  $\delta \rightarrow 1$ , ο  $BF$  αυξάνεται, δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_1$ . Υπό την πραγματική τιμή του  $\lambda_o = 0.2$ , δεχόμαστε την μηδενική υπόθεση για όλες τις τιμές του  $\delta$ .



Σχήμα 2.4: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\lambda_0$  με εκ των προτέρων κατανομή (α) την  $Gamma(1,1)$  και (β) την  $Gamma(\lambda_0, 1)$  (Παράδειγμα 2.4.1).



Σχήμα 2.5: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές (α) της παραμέτρου  $\alpha$  της εκ των προτέρων κατανομής  $Gamma(\alpha, 1)$  και (β) της παραμέτρου  $\delta \in [0, 1]$  της εκ των προτέρων κατανομής  $Gamma(\delta n \lambda_0, \delta n)$  (Παράδειγμα 2.4.1).

### 2.4.2 Παράδειγμα με πραγματικά δεδομένα

Από την Αυστραλιανή αποτίμηση υγείας (Australian Health survey) με δεδομένα από το 1977-1978, μεγέθους ίσο με 5190, επιλέγουμε τις 300 πρώτες παρατηρήσεις για να μελετήσουμε το πλήθος των φαρμακευτικών αγωγών που χορηγούνται με ιατρική συνταγή ή μη, τις τελευταίες 2 μέρες. Στο σχήμα (2.6) βλέπουμε το ραβδόγραμμα των παρατηρήσεων. Τα δεδομένα έχουν φορτωθεί από την ιστοσελίδα του βιβλίου των Cameron & Trivedi (1998)

[<http://www.econ.ucdavis.edu/faculty/cameron/racd/racddata.html>]

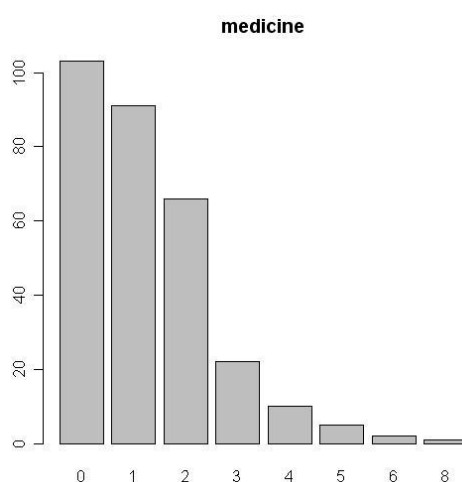
Ελέγχουμε αν η μέση τιμή του δείγματος διαφέρει σημαντικά από τις τιμές

$$\lambda_o = 0.01, \quad \lambda_o = 1 \quad \text{και} \quad \lambda_o = 2$$

• Βασιζόμενοι στην power εκ των προτέρων κατανομή επιλέγουμε δύο υπό περιπτώσεις της για την άγνωστη παράμετρο  $\lambda$ .

◇ Για  $\bar{y}_o = 1$  προκύπτει η  $Gamma(1, 1)$  με  $E(\lambda) = Var(\lambda) = 1$ .

◇ Για  $\bar{y}_o = \lambda_o^* = \lambda_o$  προκύπτει η  $Gamma(\lambda_o, 1)$  με  $E(\lambda) = Var(\lambda) = \lambda_o$ .



Σχήμα 2.6: Ραβδόγραμμα του πλήθους των φαρμακευτικών αγωγών που χορηγούνται με ιατρική συνταγή ή μη, τις τελευταίες 2 μέρες.

• Με την βοήθεια της  $R$  προκύπτουν οι πίνακες (2.7) και (2.8). Ομοίως με το προηγούμενο παράδειγμα, διαπιστώνουμε ότι οι εκ των προτέρων κατανομές  $Gamma(\lambda_o, 1)$  και  $Gamma(1, 1)$  δίνουν παρόμοια αποτελέσματα για τον παράγοντα Bayes. Το πλήθος των φαρμακευτικών αγωγών που χορηγούνται με ιατρική συνταγή ή μη, τις τελευταίες 2 μέρες διαφέρει από τις τιμές 0.01, 1 και 2.

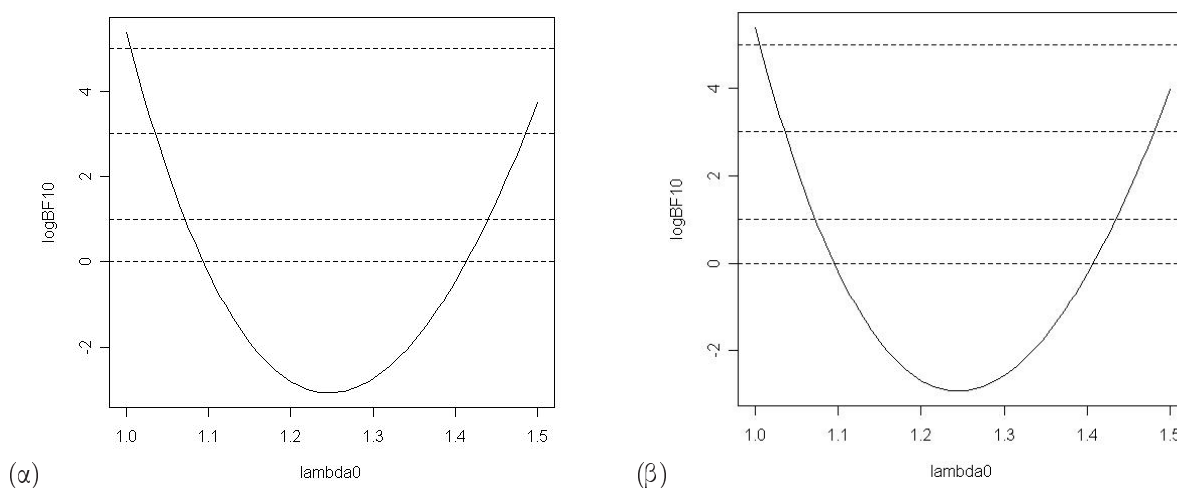
$\lambda_o$	0.01	1	2
$BF_{10}$	$+\infty$	218.46	$1.103 \times 10^{20}$
$\log(BF)_{10}$	$+\infty$	5.39	46.15
Ένδειξη εναντίον της	$H_0$	$H_0$	$H_0$
	Πολύ Ισχυρή	Πολύ Ισχυρή	Πολύ Ισχυρή

Πίνακας 2.7: Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την  $Gamma(1, 1)$ .

$\lambda_0$	0.01	1	2
$BF_{10}$	$+\infty$	218.46	$1.37 \times 10^{20}$
$\log(BF)_{10}$	$+\infty$	5.39	46.37
Ένδειξη εναντίον της	$H_0$	$H_0$	$H_0$
	Πολύ Ισχυρή	Πολύ Ισχυρή	Πολύ Ισχυρή

Πίνακας 2.8: Αποτελέσματα του παράγοντα Bayes με εκ των προτέρων κατανομή την  $Gamma(\lambda_0, 1)$ .

- Στο σχήμα (2.7) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\lambda_0$ . Έχοντας επιλέξει εκ των προτέρων κατανομή την  $Gamma(1, 1)$  αποδεχόμαστε την μηδενική υπόθεση για τιμές του  $\lambda_0 \in (1.094, 1.414)$  ενώ έχοντας επιλέξει εκ των προτέρων κατανομή την  $Gamma(\lambda_0, 1)$ , αποδεχόμαστε την μηδενική υπόθεση για  $\lambda_0 \in (1.096, 1.407)$ .

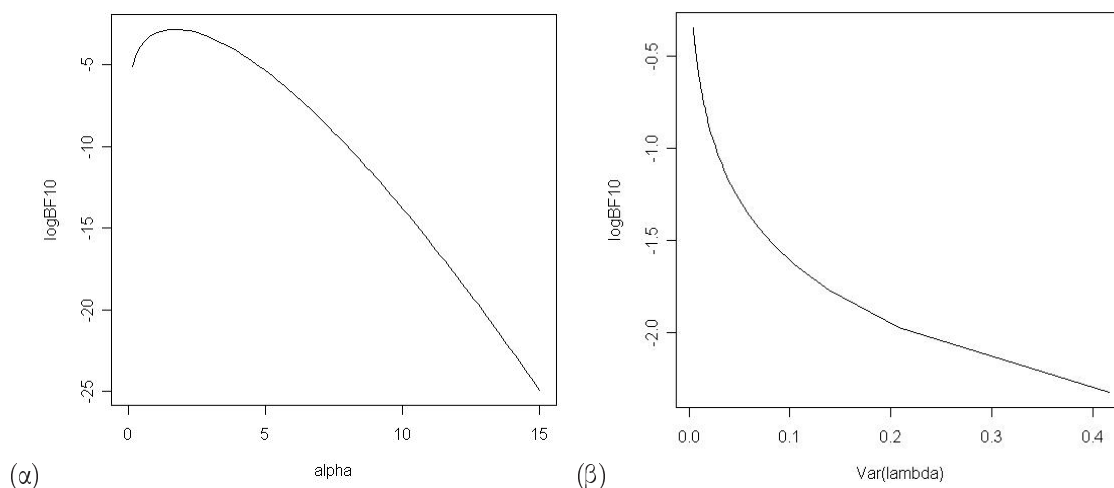


Σχήμα 2.7: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\lambda_0$  με εκ των προτέρων κατανομή (α) την  $Gamma(1, 1)$  και (β) την  $Gamma(\lambda_0, 1)$  (Παράδειγμα 2.4.2).

- Στο σχήμα (2.8α) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$ , στην γενική μορφή της εκ των προτέρων κατανομής  $Gamma(\alpha, 1)$ , υπό τον δειγματικό μέσο  $\lambda_0 = 1.25$ . Για την κατανομή  $Gamma(\alpha, 1)$ , καθώς το  $\alpha \rightarrow \infty$ , ο  $BF$  ελαττώνεται, δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_0$ . Υπό τον δειγματικό μέσο  $\lambda_0 = 1.25$ , δεχόμαστε την μηδενική υπόθεση για όλες τις τιμές του  $\alpha$ .

- Στο σχήμα (2.8β) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγο-

να Bayes για τις διάφορες τιμές της παραμέτρου  $\delta \in [0, 1]$ , στην γενική μορφή της εκ των προτέρων κατανομής  $Gamma(\delta n \lambda_o, \delta n)$ , υπό τον δειγματικό μέσο  $\lambda_o = 1.25$ . Για την κατανομή  $Gamma(\delta n \lambda_o, \delta n)$ , με  $E(\lambda) = \lambda_o$  και  $Var(\lambda) = \frac{\lambda_o}{\delta n}$ , καθώς το  $\delta \rightarrow 1$ , ο  $BF$  αυξάνεται, δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_1$ . Υπό τον δειγματικό μέσο  $\lambda_o = 1.25$ , δεχόμαστε την μηδενική υπόθεση για όλες τις τιμές του  $\delta$ .



Σχήμα 2.8: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές (α) της παραμέτρου  $\alpha$  της εκ των προτέρων κατανομής  $Gamma(\alpha, 1)$  και (β) της παραμέτρου  $\delta \in [0, 1]$  της εκ των προτέρων κατανομής  $Gamma(\delta n \lambda_o, \delta n)$  (Παράδειγμα 2.4.2).

## 2.5 Έλεγχος υποθέσεων για ένα δείγμα από κατανομή Binomial

Θεωρούμε τυχαίο δείγμα μεγέθους  $n$  ( $i = 1, \dots, n$ ) από την κατανομή Binomial με παράμετρο  $\pi$ . Θέλουμε να ελέγξουμε την ισότητα μιας αναλογίας στοιχείων ενός πληθυσμού με μια συγκεκριμένη τιμή  $\pi_o$ .

$$H_0 : \pi = \pi_o \quad vs \quad H_1 : \pi \neq \pi_o$$

Τα μοντέλα που αντιπροσωπεύουν τις παραπάνω υποθέσεις είναι αντίστοιχα

$$M_0 : y_i \sim Binomial(N_i, \pi_o) \quad \text{και} \quad M_1 : y_i \sim Binomial(N_i, \pi)$$

Για το μοντέλο  $M_0$  όπου η παράμετρος  $\pi_0$  είναι γνωστή και όχι προς εκτίμηση, το ολοκλήρωμα (2.1) ισούται με την συνάρτηση πιθανοφάνειας

$$f(\mathbf{y}|M_0) = L(\pi_0; \mathbf{y}) \propto \pi_0^{\sum y_i} (1 - \pi_0)^{N - \sum y_i}, \quad \text{όπου } N = \sum_{i=1}^n N_i$$

Για το μοντέλο  $M_1$ , επειδή η παράμετρος  $\pi$  είναι άγνωστη και άρα προς εκτίμηση, υπολογίζεται αρχικά η συνάρτηση πιθανοφάνειας  $L(\pi; \mathbf{y}) \propto \pi^{\sum y_i} (1 - \pi)^{N - \sum y_i}$ .

Στη συνέχεια υποθέτουμε ότι η εκ των προτέρων πεποίθησή μας για την παράμετρο  $\pi$  μπορεί να εκφραστεί με την βοήθεια μιας Beta κατανομής. Από το θεώρημα του Bayes υπολογίζεται αρχικά η εκ των υστέρων κατανομή και στη συνέχεια υπολογίζεται το ολοκλήρωμα (2.1).

## Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$\begin{aligned} BF_{10} &= \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_0)} = \frac{\int f(\mathbf{y}|\pi) f(\pi) d\pi}{L(\pi_0; \mathbf{y})} \\ &= \frac{\left[ \prod_{i=1}^n \binom{N_i}{y_i} \right] \frac{\Gamma(p+q)\Gamma(p+\sum y_i)\Gamma(q+N-\sum y_i)}{\Gamma(p)\Gamma(q)\Gamma(p+q+N)}}{\left[ \prod_{i=1}^n \binom{N_i}{y_i} \right] \pi_0^{\sum y_i} (1 - \pi_0)^{N - \sum y_i}} \\ &= \frac{B(p + \sum y_i, q + N - \sum y_i)}{B(p, q) \pi_0^{\sum y_i} (1 - \pi_0)^{N - \sum y_i}} \end{aligned}$$

Ο παράγοντας Bayes εξαρτάται από τις εκ των προτέρων παραμέτρους, το μέγεθος του δείγματος και τα δεδομένα. Στις περιπτώσεις όπου το μέγεθος του δείγματος είναι πολύ μεγάλο ( $n \rightarrow \infty$ ) τότε ο παράγοντας Bayes τείνει στο μηδέν ( $BF \rightarrow 0$ ) δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_0$ . Ακόμα όταν οι παράμετροι  $p \rightarrow \infty$  και  $q \rightarrow \infty$  τότε η εκ των προτέρων διασπορά  $Var(\pi) = \frac{pq}{(p+q)^2(p+q+1)} \rightarrow 0$ , συνεπώς ο  $BF \rightarrow \infty$  δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_1$ .



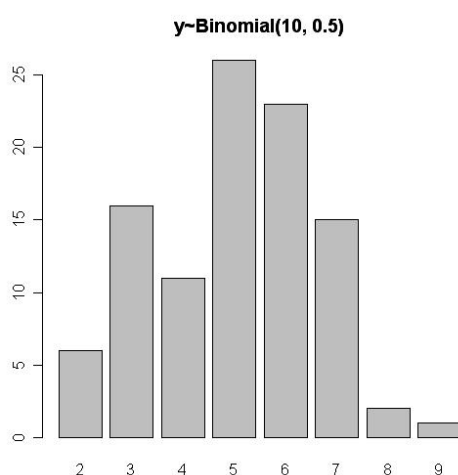
### 2.5.1 Παράδειγμα με προσομοιωμένα δεδομένα

Προσομοιώνουμε 100 παρατηρήσεις ( $y_i$ ) από κατανομή  $Binomial(10, 0.5)$ . Στο σχήμα (2.9) βλέπουμε το ραβδόγραμμα των παρατηρήσεων. Ελέγχουμε αν η αναλογία του δείγματος διαφέρει σημαντικά από τις τιμές  $\pi_o = 0.45$ ,  $\pi_o = 0.47$ ,  $\pi_o = 0.49$ ,  $\pi_o = 0.5$ ,  $\pi_o = 0.51$ ,  $\pi_o = 0.53$  και  $\pi_o = 0.55$  δηλ.

$$H_0 : \pi = \pi_o \quad vs \quad H_1 : \pi \neq \pi_o$$

που εκφράζεται από τα μοντέλα

$$M_0 : y_i \sim Binomial(10, \pi_o) \quad \text{και} \quad M_1 : y_i \sim Binomial(10, \pi)$$



Σχήμα 2.9: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από  $Binomial(10, 0.5)$ .

- Έχουμε ήδη δείξει ότι η power εκ των προτέρων κατανομή για τα Binomial μοντέλα θα έχει την μορφή  $Beta(c\bar{y}_o, c(1 - \bar{y}_o))$ . Για την άγνωστη παράμετρο  $\pi$  θα επιλέξουμε τρεις υπό περιπτώσεις της.

- (1) Για  $c = 1$  και  $\bar{y}_o = \frac{1}{2}$  προκύπτει η  $Beta(0.5, 0.5)$  με  $E(\pi) = \frac{1}{2}$  και  $Var(\pi) = \frac{1}{8}$ , που είναι η κατανομή του Jeffreys, όπως είδαμε στην παράγραφο 1.2.2.
- (2) Για  $c = 2$  και  $\bar{y}_o = \frac{1}{2}$  προκύπτει η  $Beta(1, 1)$  με  $E(\pi) = \frac{1}{2}$  και  $Var(\pi) = \frac{1}{12}$ , που είναι η ομοιόμορφη κατανομή (επίπεδη κατανομή).
- (3) Για  $c = 1$  και  $\bar{y}_o = \pi_o^* = \pi_o$  προκύπτει μια πιο γενική κατανομή, η  $Beta(\pi_o, 1 - \pi_o)$  με  $E(\pi) = \pi_o$  και  $Var(\pi) = \frac{\pi_o(1-\pi_o)}{2}$ . Θέτουμε  $\bar{y}_o = \pi_o$ ,

διότι εφόσον πιστεύουμε ότι η μηδενική υπόθεση αξίζει να ελεγχθεί, τότε υπάρχει εκ των προτέρων πεποίθηση ότι αυτή είναι η πραγματική της τιμή την οποία ζυγίζουμε ως ένα επιπλέον σημείο.

Οι κατανομές (1) και (2) αποτελούν υπό περιπτώσεις της  $Beta(\frac{c}{2}, \frac{c}{2})$ , με  $E(\pi) = \frac{1}{2}$  και  $Var(\pi) = \frac{1}{4(c+1)}$ .

• Με την βοήθεια της  $R$  προκύπτουν οι πίνακες (2.9), (2.10) και (2.11). Από τα αποτελέσματα διαπιστώνουμε ότι οι εκ των προτέρων κατανομές  $Beta(0.5, 0.5)$  και  $Beta(\pi_o, 1-\pi_o)$  δίνουν σχεδόν τα ίδια αποτελέσματα, ενώ η κατανομή  $Beta(1, 1)$  (ομοιόμορφη κατανομή) μας δίνει τις μεγαλύτερες τιμές για τον παράγοντα Bayes. Επίσης για την κατανομή  $Beta(\pi_o, 1-\pi_o)$ , παρατηρούμε ότι καθώς το  $\pi_o \rightarrow 0.5$  ο παράγοντας Bayes μειώνεται, ενώ καθώς το  $\pi_o$  συνεχίζει να αυξάνεται, αυξάνεται και ο παράγοντας Bayes.

$\pi_o$	0.45	0.47	0.49	0.5	0.51	0.53	0.55
$BF_{10}$	5.78	0.19	0.034	0.025	0.029	0.12	2.59
$\log(BF)_{10}$	1.75	-1.63	-3.39	-3.67	-3.55	-2.11	0.95
Ένδειξη εναντίον της	$H_0$ Θετική	$H_1$ Θετική	$H_1$ Ισχυρή	$H_1$ Ισχυρή	$H_1$ Ισχυρή	$H_1$ Θετική	$H_0$ Αδιάφορη

Πίνακας 2.9: Αποτελέσματα του παράγοντα Bayes για εκ των προτέρων κατανομή  $Beta(0.5, 0.5)$ .

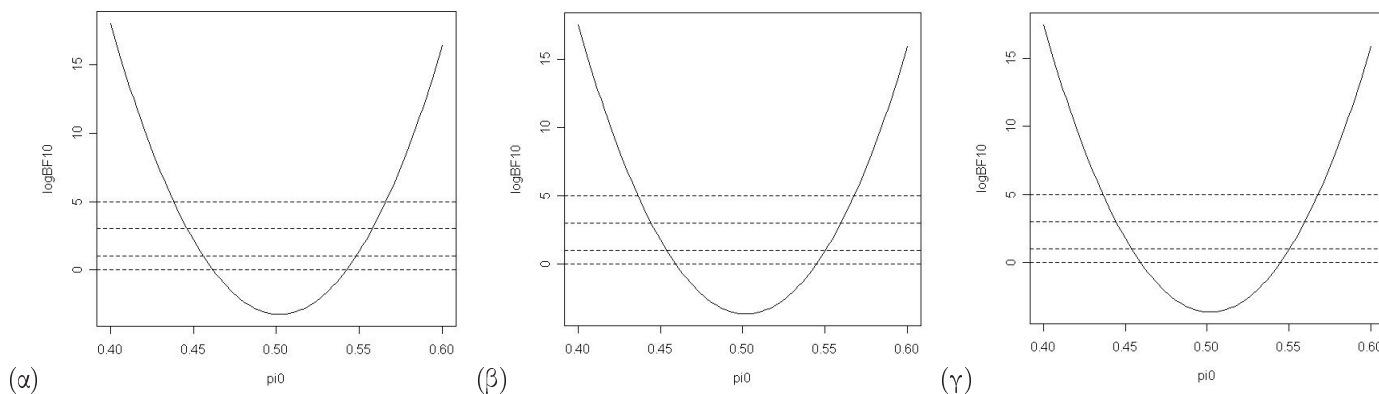
$\pi_o$	0.45	0.47	0.49	0.5	0.51	0.53	0.55
$BF_{10}$	9.08	0.308	0.05	0.04	0.045	0.19	4.07
$\log(BF)_{10}$	2.206	-1.18	-2.94	-3.22	-3.1	-1.66	1.4
Ένδειξη εναντίον της	$H_0$ Θετική	$H_1$ Θετική	$H_1$ Θετική	$H_1$ Ισχυρή	$H_1$ Ισχυρή	$H_1$ Θετική	$H_0$ Θετική

Πίνακας 2.10: Αποτελέσματα του παράγοντα Bayes για εκ των προτέρων κατανομή  $Beta(1, 1)$ .

• Στο σχήμα (2.10) εξετάζεται γραφικά η ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\pi_o$ . Οι οριζόντιες γραμμές των σχημάτων, δηλώνουν τα επίπεδα ερμηνείας σύμφωνα με τους Kass & Raftery.

$\pi_o$	0.45	0.47	0.49	0.5	0.51	0.53	0.55
$BF_{10}$	5.71	0.195	0.034	0.025	0.03	0.12	2.56
$\log(BF)_{10}$	1.74	-1.6	-3.39	-3.672	-3.55	-2.113	0.94
Ένδειξη εναντίον της	$H_0$	$H_1$	$H_1$	$H_1$	$H_1$	$H_1$	$H_0$
	Θετική	Θετική	Ισχυρή	Ισχυρή	Ισχυρή	Θετική	Αδιάφορη

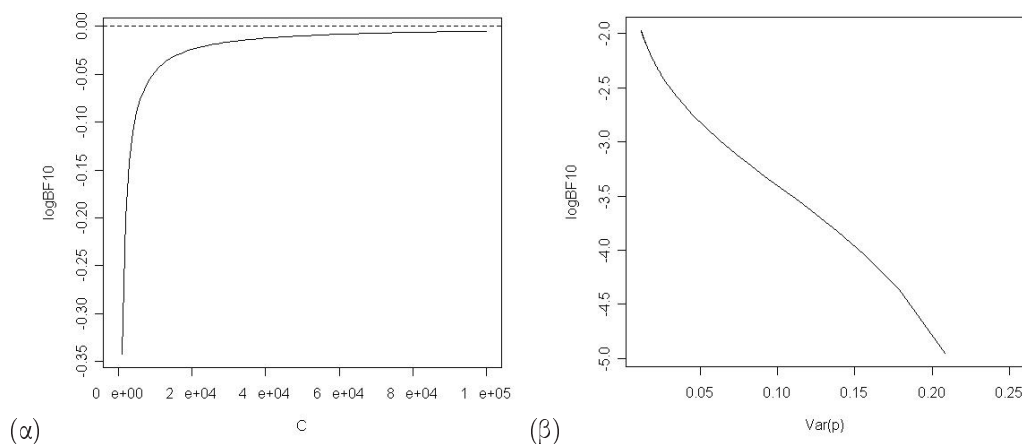
Πίνακας 2.11: Αποτελέσματα του παράγοντα Bayes για εκ των προτέρων κατανομή  $Beta(\pi_o, 1 - \pi_o)$ .



Σχήμα 2.10: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\pi_o$  με εκ των προτέρων κατανομές (α)  $Beta(0.5, 0.5)$ , (β)  $Beta(1, 1)$  και (γ)  $Beta(\pi_o, 1 - \pi_o)$  (Παράδειγμα 2.5.1).

- ◇ Υπό την  $Beta(0.5, 0.5)$  αποδεχόμαστε την μηδενική υπόθεση για τιμές του  $\pi_o \in (0.4592, 0.5448)$ . Για τιμές από  $(0.4537, 0.4592)$  ή  $(0.5448, 0.5503)$  η διαφορά δεν είναι άξια αναφοράς.
- ◇ Υπό την  $Beta(1, 1)$  αποδεχόμαστε την μηδενική υπόθεση για τιμές του  $\pi_o \in (0.4619, 0.5421)$ . Για τιμές από  $(0.4561, 0.4619)$  ή  $(0.5421, 0.5479)$  η διαφορά δεν είναι άξια αναφοράς.
- ◇ Υπό την  $Beta(\pi_o, 1 - \pi_o)$  αποδεχόμαστε την μηδενική υπόθεση για τιμές του  $\pi_o \in (0.4591, 0.5449)$ . Για τιμές από  $(0.4537, 0.4591)$  ή  $(0.5449, 0.5503)$  η διαφορά δεν είναι άξια αναφοράς.
- Στο σχήμα (2.11) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $c$  στη γενική μορφή της εκ των προτέρων κατανομής  $Beta(\frac{c}{2}, \frac{c}{2})$ , υπό την πραγματική τιμή του  $\pi_o = 0.5$ . Για εκ των προτέρων παραμέτρους  $p = q = \frac{c}{2}$  έχουμε  $Var(\pi) = \frac{1}{4(c+1)}$ , οπότε καθώς  $c \rightarrow \infty$ , το  $Var(\pi) \rightarrow 0$  και κατ' επέκταση ο  $BF$  αυξάνεται. Ωστόσο για

όλες τις τιμές της παραμέτρου  $c$ , που απεικονίζονται στο διάγραμμα, δεχόμαστε την μηδενική υπόθεση.

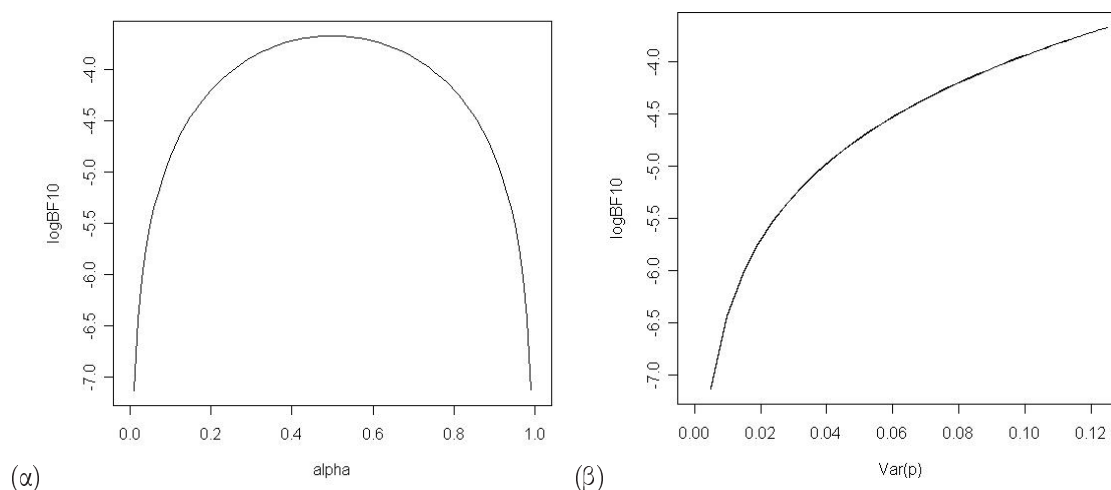


Σχήμα 2.11: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes της εκ των προτέρων κατανομής  $Beta(\frac{c}{2}, \frac{c}{2})$  ως προς (α) τις διάφορες τιμές του  $c$  και (β) της διασποράς της  $Var(\pi) = \frac{1}{4(c+1)}$  (Παράδειγμα 2.5.1).

- Στο σχήμα (2.12) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha \in [0, 1]$  στη γενική μορφή της εκ των προτέρων κατανομής  $Beta(\alpha, 1 - \alpha)$ , υπό την πραγματική τιμή του  $\pi_0 = 0.5$ .

$$Var(\pi) = \frac{\alpha(1 - \alpha)}{2} \in \begin{cases} 0 - 0.125 & , \text{για } 0 \leq \alpha \leq 0.5 \\ 0.125 - 0 & , \text{για } 0.5 \leq \alpha \leq 1 \end{cases}$$

Για τις τιμές της παραμέτρου  $0 \leq \alpha \leq 0.5$  η διασπορά αυξάνεται, ενώ ο παράγοντας Bayes ελαττώνεται. Για τις τιμές της παραμέτρου  $0.5 \leq \alpha \leq 1$  η διασπορά ελαττώνεται, ενώ ο παράγοντας Bayes αυξάνεται. Ωστόσο για όλες τις τιμές της παραμέτρου  $\alpha$  δεχόμαστε την μηδενική υπόθεση. Ακόμα παρατηρούμε ότι το διάγραμμα είναι συμμετρικό διότι η παράμετρος  $\pi_0 = 0.5$ .



Σχήμα 2.12: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes της εκ των προτέρων κατανομής  $Beta(\alpha, 1 - \alpha)$  ως προς (α) τις διάφορες τιμές του  $\alpha \in [0, 1]$  και (β) της διασποράς της  $Var(\pi) = \frac{\alpha(1-\alpha)}{2}$  (Παράδειγμα 2.5.1).

### 2.5.2 Παράδειγμα με πραγματικά δεδομένα

Τα αμινοξυγλυκοσιδικά (aminoglycoside) αντιβιοτικά, παρά την ισχυρή τοξικότητά τους, είναι ευρέως διαδεδομένα στην θεραπεία βαρυσήμαντων αρνητικών κατά gram ραβδοειδών μολύνσεων (gram-negative bacillary infections) μεταξύ των νοσηλευομένων ασθενών. Πολλές τυχαιοποιημένες κλινικές δοκιμές έχουν δημοσιευτεί για να συγκρίνουν διαφορετικά αμινοξυγλυκοσιδικά αντιβιοτικά σε σχέση με την δραστηριότητα (efficacy), την νεφρική τοξικότητα (nephrotoxicity) και την οξυγονοτοξικότητα (oxotoxicity) τους. Επειδή οι ατομικές μελέτες δεν έχουν δώσει ακλόνητα συμπεράσματα σχετικά με τα αντιβιοτικά (Amikacin, Gentamicin, Netilmicin, Sisomicin & Tobramycin), εφαρμόζεται μια μετά-ανάλυση (meta-analysis) όλων των δεδομένων των τυχαιοποιημένων δοκιμών από το 1975 έως τον Σεπτέμβρη του 1985.

Τα δεδομένα προέρχονται από το βιβλίο του Rosner [47].

Στο παράδειγμα μας έχουμε μια ομάδα από 72 τυχαιοποιημένες ατομικές κλινικές δοκιμές σχετικά με την νεφρική τοξικότητα. Για κάθε μελέτη και αντιβιοτικό καταγράφεται το ποσοστό των ασθενών που εμφανίζουν νεφρική τοξικότητα και σχετικές παρενέργειες.

• Για κάθε αντιβιοτικό ελέγχουμε αν η αναλογία του δείγματος διαφέρει σημαντικά από την τιμή  $\pi_o = 0.1$ . Για να αποφανθούμε για το ποιό αντιβιοτικό παρουσιάζει τις λιγότερες παρενέργειες θέτουμε ως «αποδεχτό» ποσοστό παρενεργειών το 10%. Για την άγνωστη παράμετρο  $\pi$  επιλέγουμε μια  $Beta(1, 1)$ , ώστε να εκφράσουμε την εκ των προτέρων άγνοια μας για τα δεδομένα και μια  $Beta(0.1, 0.9)$ , διότι εφόσον πιστεύουμε ότι η μηδενική υπόθεση αξίζει να ελεγχθεί, τότε πιθανόν να είναι αυτή η πραγματική τιμή. Με την βοήθεια της  $R$  προκύπτουν οι πίνακες πίνακας (2.12) και (2.13). Και οι δύο κατανομές μας δίνουν την ίδια συμπερασματολογία.

$\pi_o = 0.1$	<i>Amikacin</i>	<i>Gentamicin</i>	<i>Netilmicin</i>	<i>Sisomicin</i>	<i>Tobramycin</i>
$BF_{10}$	0.029	1.7	4588.634	326.2	3.44
$\log(BF)_{10}$	-3.54	0.5	8.43	5.79	1.24
Ένδειξη	$H_1$	$H_0$	$H_0$	$H_0$	$H_0$
εναντίον της	Ισχυρή	Αδιάφορη	Πολύ Ισχυρή	Πολύ Ισχυρή	Θετική

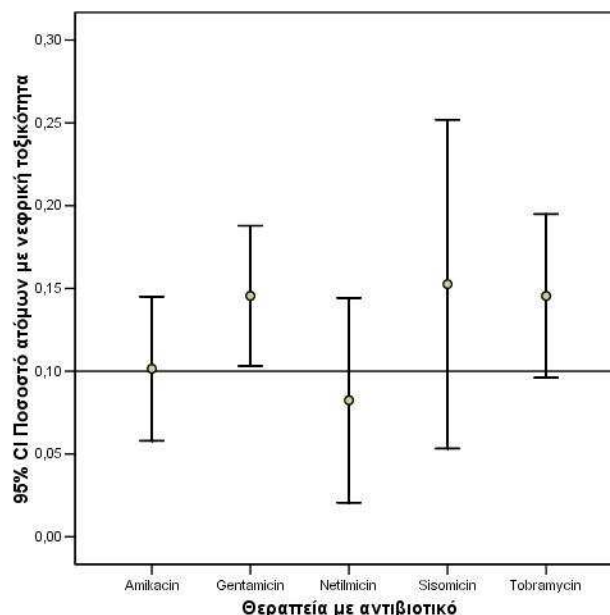
Πίνακας 2.12: Αποτελέσματα του παράγοντα Bayes για κάθε αντιβιοτικό με εκ των προτέρων κατανομή  $Beta(1, 1)$ .

$\pi_o = 0.1$	<i>Amikacin</i>	<i>Gentamicin</i>	<i>Netilmicin</i>	<i>Sisomicin</i>	<i>Tobramycin</i>
$BF_{10}$	0.024	1.122	6351.49	148.89	2.11
$\log(BF)_{10}$	-3.75	0.115	8.756	5.003	0.75
Ένδειξη	$H_1$	$H_0$	$H_0$	$H_0$	$H_0$
εναντίον της	Ισχυρή	Αδιάφορη	Πολύ Ισχυρή	Πολύ Ισχυρή	Αδιάφορη

Πίνακας 2.13: Αποτελέσματα του παράγοντα Bayes για κάθε αντιβιοτικό με εκ των προτέρων κατανομή  $Beta(0.1, 0.9)$ .

• Στο σχήμα (2.13) βλέπουμε γραφικά τα 95% διαστήματα εμπιστοσύνης (κλαστική προσέγγιση) των ποσοστών των ατόμων με νεφρική τοξικότητα και σχετικές παρενέργειες για κάθε αντιβιοτικό. Η οριζόντια ευθεία αντιπροσωπεύει την υπό έλεγχο τιμή  $\pi_o = 0.1$ . Από τους παραπάνω πίνακες και το σχήμα διαπιστώνουμε ότι το ποσοστό παρενεργειών του αντιβιοτικού *Amikacin* είναι το μόνο που δεν διαφέρει από το 10%. Όμως το αντιβιοτικό *Netilmicin* παρουσιάζει σημαντικά μικρότερο ποσοστό παρενεργειών, γεγονός που το καθιστά καλύτερο σε σχέση με τα υπόλοιπα τέσσερα. Ακόμα από το σχήμα (2.13) διαπιστώνουμε ότι τα αν-

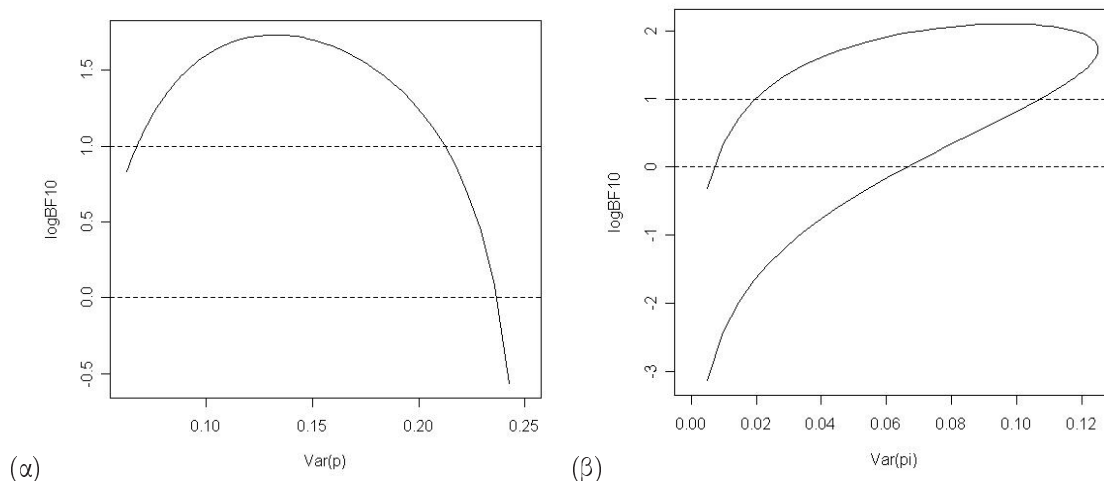
τιβιοτικά Gentamicin, Sisomicin και Tobramycin έχουν περίπου το ίδιο ποσοστό εμφάνισης νεφρικής τοξικότητας, όμως με τις μεταξύ τους συγκρίσεις θα ασχοληθούμε στο επόμενο κεφάλαιο.



Σχήμα 2.13: Errorbar με τα ποσοστά των ατόμων με νεφρική τοξικότητα και σχετικές παρενέργειες για κάθε αντιβιοτικό.

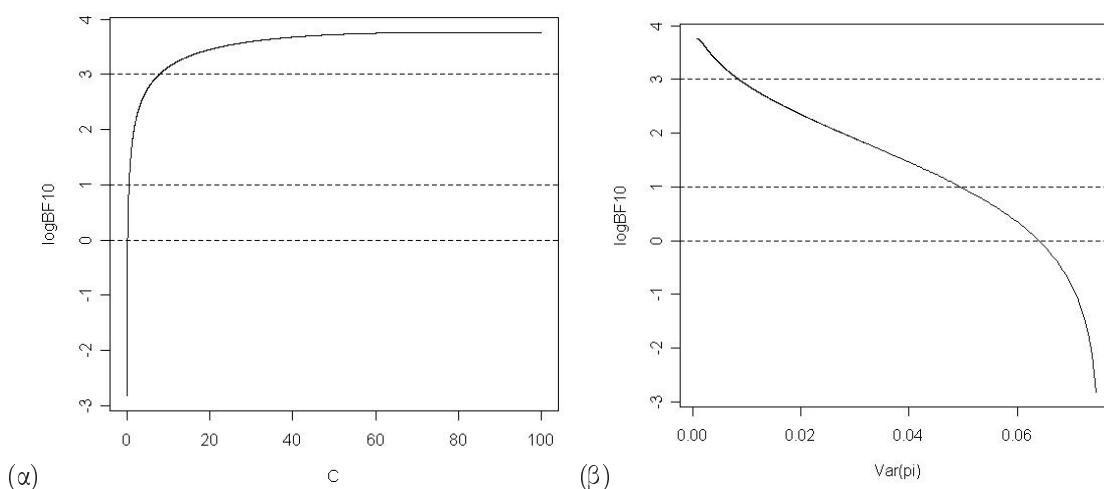
- Στο σχήμα (2.14α) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes ως προς τις τιμές της διασποράς  $\left(Var(\pi) = \frac{1}{4(c+1)}\right)$  της εκ των προτέρων κατανομής  $Beta\left(\frac{c}{2}, \frac{c}{2}\right)$ , για το αντιβιοτικό Netilmicin (υπό την δειγματική αναλογία  $\pi_o = 0.082$ ). Για τιμές της παραμέτρου  $c \in [0.54, 1.37]$  η μηδενική υπόθεση απορρίπτεται. Στο σχήμα (2.14β) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes ως προς τις τιμές της διασποράς  $\left(Var(\pi) = \frac{\alpha(1-\alpha)}{2}\right)$  της εκ των προτέρων κατανομής  $Beta(\alpha, 1 - \alpha)$ , για το αντιβιοτικό Netilmicin (υπό την δειγματική αναλογία  $\pi_o = 0.082$ ). Για τιμές της παραμέτρου  $\alpha \in [0.042, 0.688]$  η μηδενική υπόθεση απορρίπτεται. Παρατηρούμε ότι το διάγραμμα ως προς την διασπορά δεν είναι συμμετρικό όπως είχαμε δει στο σχήμα (2.12), γιατί εδώ η δειγματική αναλογία  $\pi_o = 0.082 \neq 0.5$ .

- Στο σχήμα (2.15) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $c$  και της διασποράς



Σχήμα 2.14: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές (α) της διασποράς της εκ των προτέρων κατανομής  $Beta(\frac{\xi}{2}, \frac{\xi}{2})$  και (β) της διασποράς της εκ των προτέρων κατανομής  $Beta(\alpha, 1 - \alpha)$ , για το αντιβιοτικό Netilmicin (Παράδειγμα 2.5.2).

$(Var(\pi) = \frac{\pi_o(1-\pi_o)}{c+1})$ , της εκ των προτέρων κατανομής  $Beta(c\pi_o, c(1 - \pi_o))$  για το αντιβιοτικό Netilmicin (υπό την δειγματική αναλογία  $\pi_o = 0.082$ ). Για τιμές της παραμέτρου  $c > 0.6$  η μηδενική υπόθεση απορρίπτεται.



Σχήμα 2.15: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές (α) της παραμέτρου  $c$  και (β) της διασποράς  $(Var(\pi) = \frac{\pi_o(1-\pi_o)}{c+1})$ , της εκ των προτέρων κατανομής  $Beta(c\pi_o, c(1 - \pi_o))$  για το αντιβιοτικό Netilmicin (υπό την δειγματική αναλογία  $\pi_o = 0.082$ ) (Παράδειγμα 2.5.2).



## 2.6 Έπιλογος

Στο δεύτερο κεφάλαιο ορίσαμε και υπολογίσαμε τον παράγοντα Bayes στον έλεγχο υποθέσεων ενός δείγματος από τις κατανομές Poisson και Binomial. Αρχικά ασχοληθήκαμε με προσομειωμένα δεδομένα με σκοπό να μελετήσουμε την ευαισθησία του παράγοντα Bayes σε σχέση με τις υπό έλεγχο τιμές της παραμέτρου και τις διασπορές των εκ των προτέρων κατανομών. Προτείναμε εκ των προτέρων κατανομές που είτε εκφράζουν την άγνοια μας για τα δεδομένα, είτε υποστηρίζουν την μηδενική υπόθεση που μελετάμε, είτε τίποτα από τα δύο. Διαπιστώσαμε ότι η επιλογή των εκ των προτέρων κατανομών είναι το σημαντικότερο και δυσκολότερο στάδιο της ανάλυσης, κάθε Μπευζιανού στατιστικού, λόγω του παραδόξου του Lindley. Όταν η διασπορά αυξανόταν τότε ο παράγοντας Bayes μειωνόταν και αντίστροφα. Βέβαια σε σχέση με τους κλασσικούς ελέγχους ( $p - value$ ), ο παράγοντας Bayes είναι πιο απλός και μας δίνει μια άμεση ερμηνεία του κατά πόσο ισχύει η μηδενική υπόθεση.

## Κεφάλαιο 3

# Έλεγχος Υποθέσεων για Δύο Δείγματα

### 3.1 Έλεγχος υποθέσεων για δύο ανεξάρτητα δείγματα από κατανομή Poisson

Θεωρούμε δύο τυχαία ανεξάρτητα δείγματα

$$\mathbf{y}_1 = (y_{11}, y_{21}, \dots, y_{n_11}) \quad \& \quad \mathbf{y}_2 = (y_{12}, y_{22}, \dots, y_{n_22})$$

μεγέθους  $n_1$  &  $n_2$  αντίστοιχα από κατανομή Poisson. Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0 : \lambda_1 = \lambda_2 = \lambda \quad vs \quad H_1 : \lambda_1 \neq \lambda_2$$

Τα αντίστοιχα μοντέλα που αντιπροσωπεύουν οι παραπάνω υποθέσεις μας είναι :

$$\begin{aligned} M_0 : \mathbf{y}_1 &\sim Poisson(\lambda) & M_1 : \mathbf{y}_1 &\sim Poisson(\lambda_1) \\ \mathbf{y}_2 &\sim Poisson(\lambda) & \mathbf{y}_2 &\sim Poisson(\lambda_2) \end{aligned}$$

Για το μοντέλο  $M_0$  ορίζουμε εκ των προτέρων κατανομή για την παράμετρο  $\lambda \sim Gamma(p, q)$ , από την οποία προκύπτει η εκ των υστέρων  $f(\lambda|\mathbf{y}) \sim Gamma(p + N\bar{y}, q + N)$ , όπου  $N = n_1 + n_2$ ,  $\bar{y}$  είναι ο μέσος όλων των δεδομένων και  $\mathbf{y} = \{y_{11}, \dots, y_{n_11}, y_{12}, \dots, y_{n_22}\} = \{\mathbf{y}_1, \mathbf{y}_2\}$ .

$$f(\mathbf{y}|M_0) = \int_0^\infty f(\mathbf{y}|\lambda)f(\lambda)d\lambda = \frac{\Gamma(p + \sum y_i)q^p}{\Gamma(p)(q + N)^{p+\sum y_i} \prod_{i=1}^N y_i!}$$

Για το μοντέλο  $M_1$  ορίζουμε τις εκ των προτέρων κατανομές των παραμέτρων  $\lambda_1 \sim \text{Gamma}(p_1, q_1)$  και  $\lambda_2 \sim \text{Gamma}(p_2, q_2)$ .

$$\begin{aligned}
f(\mathbf{y}|M_1) &= \int_0^\infty \int_0^\infty f(\mathbf{y}|\lambda_1, \lambda_2, M_1) f(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2 \\
&= \int_0^\infty \int_0^\infty f(\mathbf{y}_1|\lambda_1) f(\mathbf{y}_2|\lambda_2) f(\mathbf{y}_1) f(\mathbf{y}_2) d\lambda_1 d\lambda_2 \\
&= \int_0^\infty f(\mathbf{y}_1|\lambda_1) f(\mathbf{y}_1) d\lambda_1 \times \int_0^\infty f(\mathbf{y}_2|\lambda_2) f(\mathbf{y}_2) d\lambda_2 \\
&= \frac{\Gamma(p_1 + \sum y_{1i}) q_1^{p_1} \Gamma(p_2 + \sum y_{2i}) q_2^{p_2}}{\Gamma(p_1) (q_1 + n_1)^{p_1 + \sum y_{1i}} \prod_{i=1}^{n_1} y_{1i}! \Gamma(p_2) (q_2 + n_2)^{p_2 + \sum y_{2i}} \prod_{i=1}^{n_2} y_{2i}!}
\end{aligned}$$

### Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$BF_{10} = \frac{f(\mathbf{y}|M_1)}{f(\mathbf{y}|M_0)} = \frac{\Gamma(p)(q + N)^{p + N\bar{y}} \Gamma(p_1 + n_1\bar{y}_1) q_1^{p_1} \Gamma(p_2 + n_2\bar{y}_2) q_2^{p_2}}{\Gamma(p + N\bar{y}) q^p \Gamma(p_1) (q_1 + n_1)^{p_1 + n_1\bar{y}_1} \Gamma(p_2) (q_2 + n_2)^{p_2 + n_2\bar{y}_2}}$$

Ο παράγοντας Bayes εξαρτάται από τις εκ των προτέρων παραμέτρους, το μέγεθος των δειγμάτων και τους μέσους τους. Στις περιπτώσεις όπου το άθροισμα των μεγεθών των δειγμάτων είναι πολύ μεγάλο ( $N \rightarrow \infty$ ) τότε ο παράγοντας Bayes τείνει στο μηδέν ( $BF \rightarrow 0$ ) δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_0$ . Ακόμα όταν οι εκ των προτέρων διασπορές αυξάνονται, δηλαδή

$$\text{Var}(\lambda) = \frac{p}{q^2} \rightarrow \infty, \quad \text{Var}(\lambda_1) = \frac{p_1}{q_1^2} \rightarrow \infty, \quad \text{Var}(\lambda_2) = \frac{p_2}{q_2^2} \rightarrow \infty$$

που σημαίνει ότι  $q, q_1, q_2 \rightarrow 0$  τότε ο  $BF \rightarrow 0$  δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_0$ .

### Επιλογή της εκ των προτέρων κατανομής

Για να εξάγουμε συμπεράσματα θα βασιστούμε στην power εκ των προτέρων κατανομή,  $Gamma(\delta n \bar{y}_o, \delta n)$ . Οπότε για  $\delta = \frac{1}{N}$  οι κατανομές που θα πάρουμε θα είναι για την μηδενική και την εναλλακτική υπόθεση αντίστοιχα οι :

$$H_0 : \lambda \sim Gamma(\bar{y}, 1) \quad vs \quad H_1 : \lambda_j \sim Gamma\left(\frac{n_j}{N} \bar{y}_j, \frac{n_j}{N}\right), \quad \text{όπου } j = 1, 2$$

Η παραπάνω σχέση μπορεί πιο γενικά να πάρει την μορφή

$$H_0 : \lambda \sim Gamma(\alpha, 1) \quad vs \quad H_1 : \lambda_j \sim Gamma(w_j \alpha_j, w_j)$$

όπου

$$\begin{aligned} - \alpha &= w_1 \alpha_1 + w_2 \alpha_2 & - w_j &= \frac{n_j}{N} \\ - N &= n_1 + n_2 & - j &= 1, 2 \end{aligned}$$

Από την γενική μορφή της εκ των προτέρων κατανομής διακρίνουμε τις εξής περιπτώσεις :

- ◇ Αν  $\alpha_1 = \alpha_2 = \alpha$  τότε για τα  $\lambda$  και  $\lambda_j$ , θα πάρουμε τις εκ των προτέρων κατανομές  $\lambda \sim Gamma(\alpha, 1)$  και  $\lambda_j \sim Gamma\left(\frac{n_j}{N} \alpha, \frac{n_j}{N}\right)$ .
- ◇ Αν  $\alpha_1 = \alpha_2 = \alpha$  και επιπλέον  $n_1 = n_2$  τότε για τα  $\lambda$  και  $\lambda_j$ , θα πάρουμε τις εκ των προτέρων κατανομές  $\lambda \sim Gamma(\alpha, 1)$ , και  $\lambda_j \sim Gamma\left(\frac{\alpha}{2}, \frac{1}{2}\right)$ .
- Για  $\alpha = 1$  προκύπτουν οι εκ των προτέρων κατανομές

$$\lambda \sim Gamma(1, 1) \quad \& \quad \lambda_j \sim Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\text{με } E(\lambda) = Var(\lambda) = 1 \text{ και } E(\lambda_j) = 1 \quad \& \quad Var(\lambda_j) = 2.$$

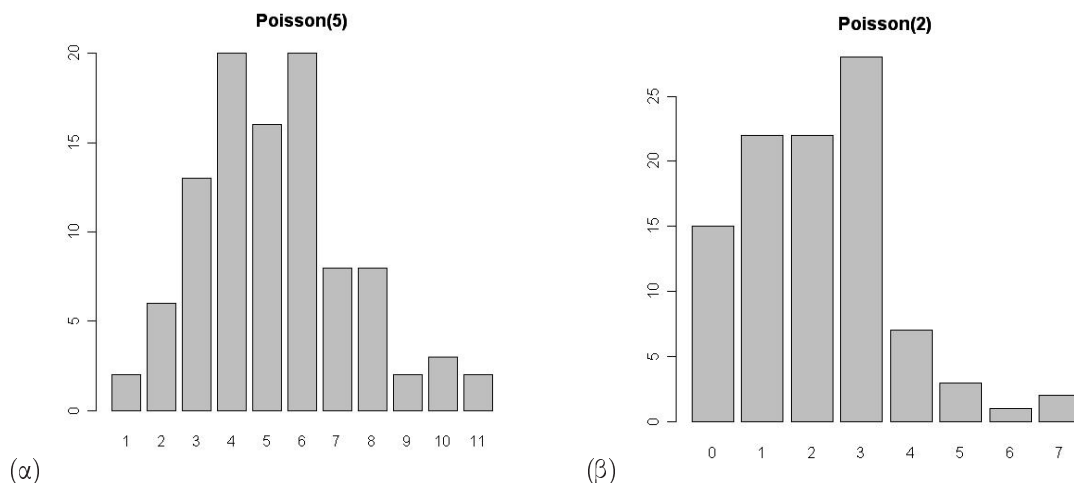
- Για  $\alpha = \bar{y}$  (δειγματικός μέσος) προκύπτουν οι εκ των προτέρων κατανομές

$$\lambda \sim Gamma(\bar{y}, 1) \quad \& \quad \lambda_j \sim Gamma\left(\frac{\bar{y}}{2}, \frac{1}{2}\right)$$

με  $E(\lambda) = Var(\lambda) = \bar{y}$  και  $E(\lambda_j) = \bar{y} \quad \& \quad Var(\lambda_j) = 2\bar{y}$ , που είναι Empirical Bayes Priors διότι βασίζονται στα παρατηρούμενα δεδομένα.

### 3.1.1 Παράδειγμα (1) με προσομοιωμένα δεδομένα

Προσομοιώνουμε 100 παρατηρήσεις από κατανομή  $Poisson(5)$  και 100 παρατηρήσεις από κατανομή  $Poisson(2)$ . Στο σχήμα (3.1) βλέπουμε τα ραβδογράμματα των παρατηρήσεων.



Σχήμα 3.1: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από (α)  $Poisson(5)$  και (β)  $Poisson(2)$ .

- Για τις παραμέτρους  $\lambda$ ,  $\lambda_1$  και  $\lambda_2$  αντίστοιχα, επιλέγουμε τις εκ των προτέρων κατανομές

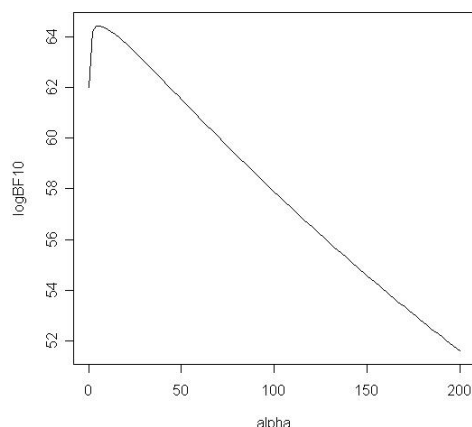
$$\diamond \text{prior 1} : \lambda \sim \text{Gamma}(1, 1), \quad \lambda_1, \lambda_2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\diamond \text{prior 2} : \lambda \sim \text{Gamma}(\bar{y}, 1), \quad \lambda_1, \lambda_2 \sim \text{Gamma}\left(\frac{\bar{y}}{2}, \frac{1}{2}\right)$$

- Με την βοήθεια της  $R$  προκύπτει ο πίνακας (3.1). Από τον πίνακα συμπεραίνουμε ότι ο συνδυασμός *prior 1* μας δίνει την μικρότερη τιμή για τον παράγοντα Bayes. Όμως και οι δύο συνδυασμοί απορρίπτουν την μηδενική υπόθεση, δηλαδή τα δύο δείγματα έχουν διαφορετικούς μέσους.

	<i>prior 1</i>	<i>prior 2</i>
$BF_{10}$	$5.32 \times 10^{27}$	$9.19 \times 10^{27}$
$\log(BF)_{10}$	63.84	64.39
Ένδειξη εναντίον της	$H_0$	$H_0$
	Πολύ ισχυρή	Πολύ ισχυρή

Πίνακας 3.1: Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.



Σχήμα 3.2: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes ως προς τις διάφορες τιμές της παραμέτρου  $\alpha$  για την γενική μορφή των δύο συνδυασμών, δηλαδή  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2 \sim \text{Gamma}(\frac{\alpha}{2}, \frac{1}{2})$  (Παράδειγμα 3.1.1).

• Στο σχήμα (3.2) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\lambda$ ,  $\lambda_1$  και  $\lambda_2$  αντίστοιχα.

[◇] Γενική μορφή:  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2 \sim \text{Gamma}(\frac{\alpha}{2}, \frac{1}{2})$

Υπό την γενική μορφή των συνδυασμών, η μηδενική υπόθεση απορρίπτεται για τιμές της παραμέτρου  $\alpha \in (0, 4 \times 10^4)$ , για τιμές  $\alpha \in (4 \times 10^4, 1.7 \times 10^{14})$  η διαφορά δεν είναι άξια αναφοράς, ενώ για τιμές της παραμέτρου  $\alpha > 1.7 \times 10^{14}$  αποδεχόμαστε την μηδενική υπόθεση.

### 3.1.2 Παράδειγμα (2) με προσομοιωμένα δεδομένα

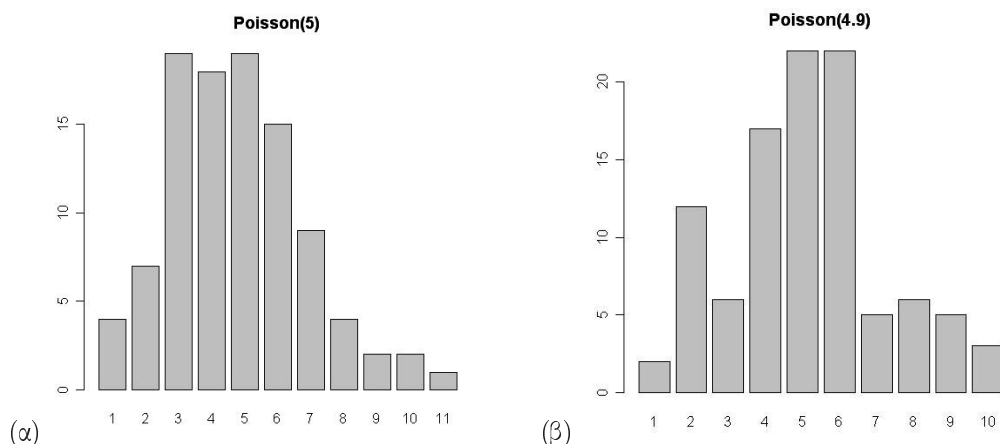
Προσομειώνουμε 100 παρατηρήσεις από κατανομή  $Poisson(5)$  και 100 παρατηρήσεις από κατανομή  $Poisson(4.9)$ . Στο σχήμα (3.3) βλέπουμε τα ραβδογράμματα των παρατηρήσεων.

• Για τις παραμέτρους  $\lambda$ ,  $\lambda_1$  και  $\lambda_2$  αντίστοιχα, επιλέγουμε τις εκ των προτέρων κατανομές

◇ *prior 1* :  $\lambda \sim \text{Gamma}(1, 1)$ ,  $\lambda_1, \lambda_2 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$

◇ *prior 2* :  $\lambda \sim \text{Gamma}(\bar{y}, 1)$ ,  $\lambda_1, \lambda_2 \sim \text{Gamma}(\frac{\bar{y}}{2}, \frac{1}{2})$

• Με την βοήθεια της  $R$  προκύπτει ο πίνακας (3.2). Από τον πίνακα συμπεραί-



Σχήμα 3.3: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από (α)  $Poisson(5)$  και (β)  $Poisson(4.9)$ .

νομε ότι ο συνδυασμός *prior* 1 μας δίνει την μικρότερη τιμή για τον παράγοντα Bayes. Όμως, και οι δύο συνδυασμοί απορρίπτουν την εναλλακτική υπόθεση, δηλαδή τα δύο δείγματα έχουν ίσους μέσους.

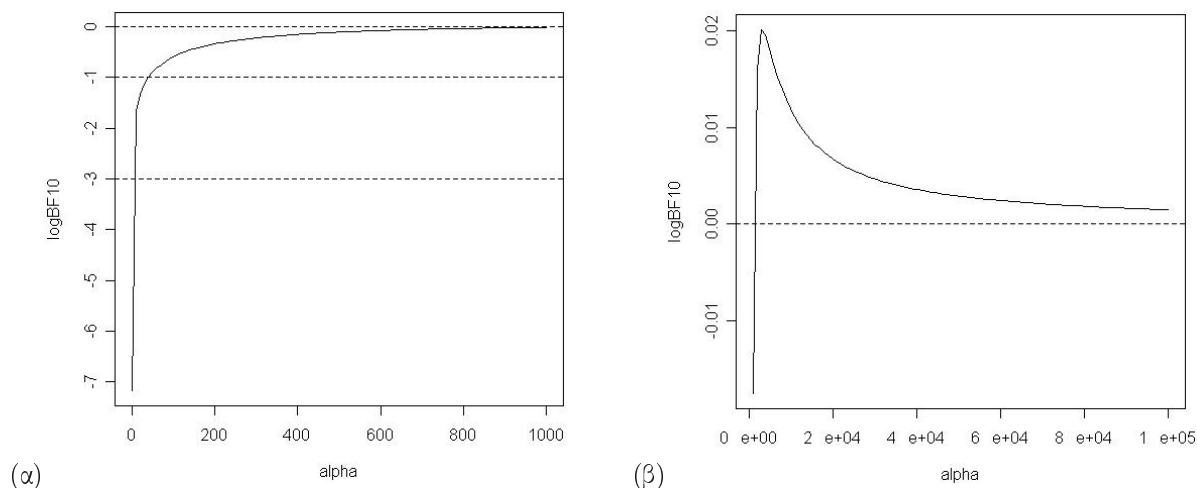
	<i>prior</i> 1	<i>prior</i> 2
$BF_{10}$	0.049	0.129
$\log(BF)_{10}$	-3.018	-2.048
Ένδειξη εναντίον της	$H_1$ Ισχυρή	$H_1$ Θετική

Πίνακας 3.2: Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.

- Στο σχήμα (3.4α) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\lambda$ ,  $\lambda_1$  και  $\lambda_2$ .

[◊] Γενική μορφή:  $\lambda \sim Gamma(\alpha, 1)$ ,  $\lambda_1, \lambda_2 \sim Gamma\left(\frac{\alpha}{2}, \frac{1}{2}\right)$

Υπό την γενική μορφή των συνδυασμών, φαίνεται ότι για ένα μεγάλο εύρος τιμών της παραμέτρου  $\alpha$  αποδεχόμαστε την μηδενική υπόθεση. Από το σχήμα (3.4α) προκύπτει το συμπέρασμα ότι δεν ισχύει το παράδοξο του Lindley επειδή καθώς το  $\alpha$  αυξάνεται, αυξάνεται και ο παράγοντας Bayes. Αυτή η αίσθηση είναι ψευδής διότι για μεγαλύτερες τιμές της παραμέτρου  $\alpha$  ο παράγοντας Bayes ελατώνεται (βλ. σχήμα (3.4β)). Επίσης το σχήμα (3.4β) δίνει την εντύπωση ότι για κάποιες



Σχήμα 3.4: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για διαφορετικό εύρος τιμών της παραμέτρου  $\alpha$  για την γενική μορφή των δύο συνδυασμών :  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2 \sim \text{Gamma}(\frac{\alpha}{2}, \frac{1}{2})$ , όπου (α)  $\alpha \in [0, 1000]$  και (β)  $\alpha \in [10^3, 10^5]$  (Παράδειγμα 3.1.2).

τιμές της παραμέτρου  $\alpha$  απορρίπτουμε την μηδενική υπόθεση. Ούτε αυτό όμως ισχύει διότι οι τιμές αυτές δίνουν τιμές στον παράγοντα Bayes λίγο πάνω από το μηδέν, όπου όμως η διαφορά δεν είναι άξια αναφοράς. Οπότε αποδεχόμαστε την μηδενική υπόθεση.

### 3.1.3 Παράδειγμα με πραγματικά δεδομένα

Θα συνεχίσουμε το παράδειγμα της παραγράφου 2.4.2. Θα εξετάσουμε το πλήθος των φαρμακευτικών αγωγών που χορηγούνται με ιατρική συνταγή ή μη, τις τελευταίες 2 μέρες σε σχέση με το φύλο (133 άνδρες και 167 γυναίκες). Στο σχήμα (3.5) βλέπουμε τα ραβδογράμματα των παρατηρήσεων.

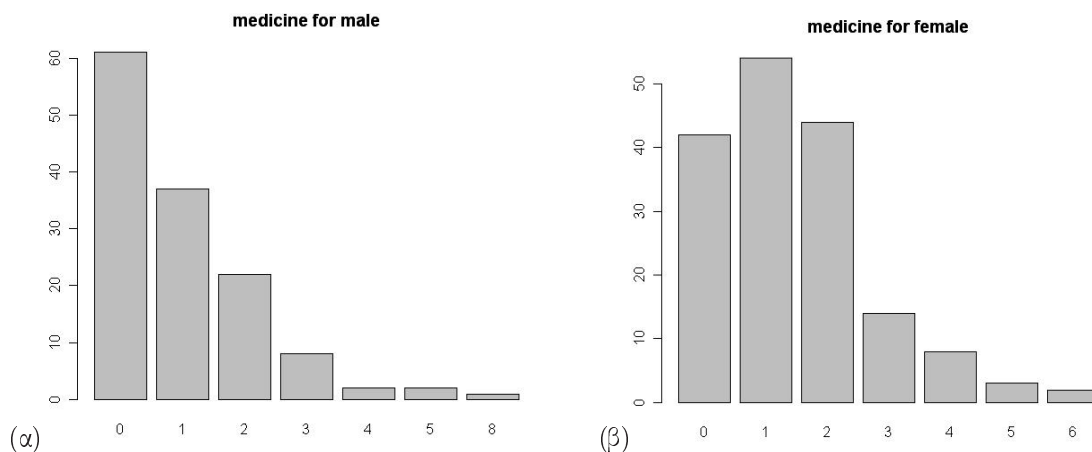
- Για τις παραμέτρους  $\lambda$ ,  $\lambda_1$  και  $\lambda_2$  αντίστοιχα, επιλέγουμε τις εκ των προτέρων κατανομές

$$\diamond \text{prior 1} : \lambda \sim \text{Gamma}(1, 1), \quad \lambda_1, \lambda_2 \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\diamond \text{prior 2} : \lambda \sim \text{Gamma}(\bar{y}, 1), \quad \lambda_1, \lambda_2 \sim \text{Gamma}\left(\frac{\bar{y}}{2}, \frac{1}{2}\right)$$

- Με την βοήθεια της  $R$  προκύπτει ο πίνακας (3.3). Από τον πίνακα συμπεραίνουμε ότι ο συνδυασμός *prior 1* μας δίνει την μικρότερη τιμή για τον παράγοντα Bayes. Όμως και οι δύο συνδυασμοί απορρίπτουν την μηδενική υπόθεση, δηλαδή τα δύο δείγματα έχουν διαφορετικούς μέσους, με τις γυναίκες να ξεπερνούν τους





Σχήμα 3.5: Ραβδόγραμμα των παρατηρήσεων για (α) τους άνδρες και (β) τις γυναίκες.

άνδρες στη λήψη φαρμακευτικών αγωγών. Δεν θα μπορούσαμε όμως να ισχυριστούμε ότι αυτό σημαίνει πως οι γυναίκες αρρωσταίνουν περισσότερο από τους άνδρες, αλλά πιθανόν η διαφορά να οφείλεται σε άλλους παράγοντες όπως ότι:

- οι γυναίκες επισκέφτονται πιο συχνά τον γιατρό τους ή
- οι γυναίκες χρησιμοποιούν περισσότερα φαρμακευτικά καλλυντικά, συμπληρώματα διατροφής, φάρμακα για την ακμή κ.α.

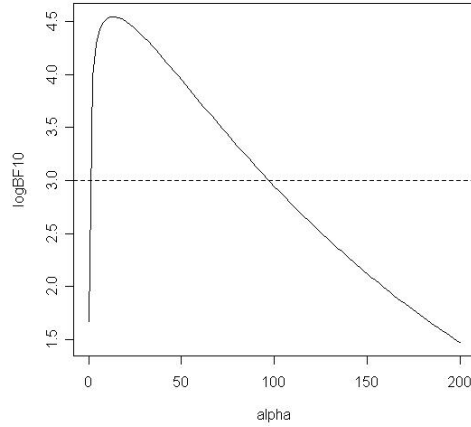
	<i>prior 1</i>	<i>prior 2</i>
$BF_{10}$	34.6	39.78
$\log(BF)_{10}$	3.54	3.68
Ένδειξη εναντίον της	$H_0$ Ισχυρή	$H_0$ Ισχυρή

Πίνακας 3.3: Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.

• Στο σχήμα (3.6) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\lambda$ ,  $\lambda_1$  και  $\lambda_2$  αντίστοιχα.

[◇] *prior 1 & 2*:  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2 \sim \text{Gamma}(\frac{\alpha}{2}, \frac{1}{2})$

Υπό την γενική μορφή των συνδυασμών, για τις τιμές που απεικονίζονται στο σχήμα (3.6) η μηδενική υπόθεση απορρίπτεται.



Σχήμα 3.6: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$  για την γενική μορφή των δύο συνδυασμών :  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2 \sim \text{Gamma}(\frac{\alpha}{2}, \frac{1}{2})$  (Παράδειγμα 3.1.3).

### 3.2 Έλεγχος υποθέσεων για δύο ανεξάρτητα δείγματα από κατανομή Binomial

Θεωρούμε δύο ( $j = 1, 2$ ) τυχαία ανεξάρτητα δείγματα.

$$\mathbf{y}_1 = (y_{11}, y_{21}, \dots, y_{n_11}) \ \& \ \mathbf{y}_2 = (y_{12}, y_{22}, \dots, y_{n_22})$$

μεγέθους  $n_1$  &  $n_2$  ( $n = n_1 + n_2$ ) αντίστοιχα από Binomial κατανομή.

Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0 : \pi_1 = \pi_2 = \pi \quad \text{vs} \quad H_1 : \pi_1 \neq \pi_2$$

Τα αντίστοιχα μοντέλα που αντιπροσωπεύουν οι παραπάνω υποθέσεις μας είναι :

$$\begin{aligned} M_0 : \mathbf{y}_1 &\sim \text{Binomial}(N_{i1}, \pi) & M_1 : \mathbf{y}_1 &\sim \text{Binomial}(N_{i1}, \pi_1) \\ \mathbf{y}_2 &\sim \text{Binomial}(N_{i2}, \pi) & \mathbf{y}_2 &\sim \text{Binomial}(N_{i2}, \pi_2) \end{aligned}$$

Για το μηδενικό μοντέλο  $M_0$  ορίζουμε την εκ των προτέρων  $\pi \sim \text{Beta}(\alpha_o, \beta_o)$ ,  
οπότε η εκ των υστέρων κατανομή θα είναι η

$$f(\pi|\mathbf{y}) \sim \text{Beta}(\alpha_o + \sum_{i=1}^n y_i, \beta_o + N - \sum_{i=1}^n y_i)$$

όπου  $N = N_{.1} + N_{.2}$ ,  $N_{.1} = \sum_{i=1}^{n_1} N_{i1}$ ,  $N_{.2} = \sum_{i=1}^{n_2} N_{i2}$  και  $\mathbf{y} = \{\mathbf{y}_1, \mathbf{y}_2\}$ .

$$f(\mathbf{y}|M_0) = \frac{\Gamma(\alpha_o + \beta_o)\Gamma(\alpha_o + \sum y_i)\Gamma(\beta_o + N - \sum y_i)}{\Gamma(\alpha_o + \beta_o + N)\Gamma(\alpha_o)\Gamma(\beta_o)} \prod_{i=1}^n \prod_{j=1}^2 \binom{N_{ij}}{y_{ij}}$$

Για το εναλλακτικό μοντέλο  $M_1$  ορίζουμε τις εκ των προτέρων κατανομές των παραμέτρων  $\pi_1 \sim Beta(\alpha_1, \beta_1)$  και  $\pi_2 \sim Beta(\alpha_2, \beta_2)$  οπότε παίρνουμε

$$\begin{aligned} f(\mathbf{y}|M_1) &= \int_0^1 \int_0^1 f(\mathbf{y}|\pi_1, \pi_2, M_1) f(\pi_1, \pi_2) d\pi_1 d\pi_2 \\ &= \frac{\Gamma(\alpha_1 + \beta_1)\Gamma(\alpha_1 + \sum_{i=1}^{n_1} y_{i1})\Gamma(\beta_1 + N_{.1} - \sum_{i=1}^{n_1} y_{i1})}{\Gamma(\alpha_1 + \beta_1 + N_{.1})\Gamma(\alpha_1)\Gamma(\beta_1)} \prod_{i=1}^{n_1} \binom{N_{i1}}{y_{i1}} \\ &\times \frac{\Gamma(\alpha_2 + \beta_2)\Gamma(\alpha_2 + \sum_{i=1}^{n_2} y_{i2})\Gamma(\beta_2 + N_{.2} - \sum_{i=1}^{n_2} y_{i2})}{\Gamma(\alpha_2 + \beta_2 + N_{.2})\Gamma(\alpha_2)\Gamma(\beta_2)} \prod_{i=1}^{n_2} \binom{N_{i2}}{y_{i2}} \end{aligned}$$

### Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$\begin{aligned} BF_{10} &= \frac{B(\alpha_o, \beta_o)}{B(\alpha_o + N \bar{y}, \beta_o + N - N \bar{y})} \\ &\times \frac{B(\alpha_1 + N_{.1} \bar{y}_1, \beta_1 + N_{.1} - N_{.1} \bar{y}_1)}{B(\alpha_1, \beta_1)} \\ &\times \frac{B(\alpha_2 + N_{.2} \bar{y}_2, \beta_2 + N_{.2} - N_{.2} \bar{y}_2)}{B(\alpha_2, \beta_2)} \end{aligned}$$

Ο παράγοντας Bayes εξαρτάται από τις εκ των προτέρων παραμέτρους, το μέγεθος των δειγμάτων και τους μέσους τους. Στις περιπτώσεις όπου το άθροισμα των μεγεθών των δειγμάτων είναι πολύ μεγάλο ( $n \rightarrow \infty$ ) τότε ο παράγοντας Bayes ελλατώνεται δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_0$ . Ακόμα όταν οι εκ των προτέρων παράμετροι τείνουν στο άπειρο, τότε οι εκ των προτέρων διασπορές

τείνουν στο μηδέν (π.χ.  $Var(\pi) = \frac{pq}{(p+q)^2(p+q+1)} \rightarrow 0$ ) και ο παράγοντας Bayes τείνει στο άπειρο ( $BF \rightarrow \infty$ ), δίνοντας πολύ ισχυρή ένδειξη υπέρ της  $H_1$ .

### Επιλογή της εκ των προτέρων κατανομής

Για να εξάγουμε συμπεράσματα θα βασιστούμε στην power εκ των προτέρων κατανομή,  $Beta(N_o\delta\bar{y}_o, N_o\delta(1 - \bar{y}_o))$ . Οπότε για  $\delta = \frac{1}{N_o}$  οι κατανομές που θα πάρουμε θα είναι για την μηδενική και την εναλλακτική υπόθεση αντίστοιχα οι :

$$H_0 : \pi \sim Beta(\bar{y}, (1 - \bar{y})) \quad vs \quad H_1 : \pi_j \sim Beta\left(\frac{N_j\bar{y}_j}{N}, \frac{N_j(1 - \bar{y}_j)}{N}\right)$$

Όπου  $\bar{y}$  και  $\bar{y}_j$  είναι οι δειγματικές διωνυμικές αναλογίες. Η παραπάνω σχέση μπορεί πιο γενικά να πάρει την μορφή

$$H_0 : \pi \sim Beta(\alpha, 1 - \alpha) \quad vs \quad H_1 : \pi_j \sim Beta(w_j\alpha_j, w_j(1 - \alpha_j))$$

όπου

$$\begin{aligned} - \alpha &= w_1\alpha_1 + w_2\alpha_2 & - w_j &= \frac{N_j}{N} \\ - N &= N_1 + N_2 & - j &= 1, 2 \end{aligned}$$

Από την γενική μορφή της εκ των προτέρων κατανομής διακρίνουμε τις εξής περιπτώσεις :

◇ Αν  $\alpha_1 = \alpha_2 = \alpha$  :  $\alpha \in (0, 1)$  τότε θα έχουμε τις εκ των προτέρων κατανομές  $\pi \sim Beta(\alpha, 1 - \alpha)$  και  $\pi_j \sim Beta\left(\frac{N_j\alpha}{N}, \frac{N_j(1 - \alpha)}{N}\right)$ .

- Για  $\alpha = \frac{1}{2}$  προκύπτουν οι εκ των προτέρων κατανομές

$$\pi \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right) \quad \& \quad \pi_j \sim Beta\left(\frac{N_j}{2N}, \frac{N_j}{2N}\right)$$

όπου θα έχουμε μια Jeffreys priors για την μηδενική υπόθεση. Θέτουμε  $\alpha = \frac{1}{2}$  διότι είναι το κέντρο της κατανομής και θεωρούμε ότι το κέντρο είναι η πιο ουδέτερη τιμή.

- Για  $\alpha = \bar{y}$  προκύπτουν οι εκ των προτέρων κατανομές

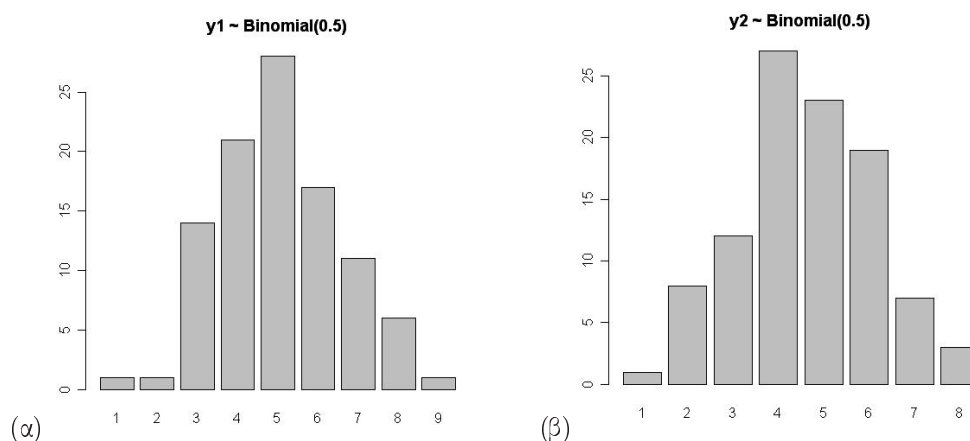
$$\pi \sim Beta(\bar{y}, 1 - \bar{y}) \quad \& \quad \pi_j \sim Beta\left(\frac{N_j\bar{y}_j}{N}, \frac{N_j(1 - \bar{y}_j)}{N}\right)$$

με  $E(\pi) = \bar{y}$ ,  $Var(\pi) = \frac{\bar{y}(1-\bar{y})}{2}$  και  $E(\pi_j) = \bar{y}$  &  $Var(\pi_j) = \frac{\bar{y}(1-\bar{y})}{\frac{N_j}{N}+1}$ , που είναι Empirical Bayes Priors.

- ◇ Αν  $\alpha_1 = \alpha_2 = \alpha$  και επιπλέον  $N_1 = N_2$  τότε θα πάρουμε τις εκ των προτέρων κατανομές  $\pi \sim Beta(\alpha, 1 - \alpha)$  και  $\pi_j \sim Beta\left(\frac{\alpha}{2}, \frac{1}{2}(1 - \alpha)\right)$ .
- ◇ Ενδιαφέρον έχει επίσης να μελετήσουμε την  $Beta\left(\frac{c}{2}, \frac{c}{2}\right)$  εκ των προτέρων κατανομή  $\left[E(\pi) = \frac{1}{2} \text{ \& } Var(\pi) = \frac{1}{4(c+1)}\right]$ , που είναι υπό περίπτωση της power εκ των προτέρων για  $\bar{y}_o = \frac{1}{2}$ , και ιδιαίτερα τις υπό περιπτώσεις :
  - Για  $c = 1$ , οι εκ των προτέρων κατανομές που θα προκύψουν για τα  $\pi$  και  $\pi_j$ , θα είναι όλες  $Beta(1, 1)$  (Ομοιόμορφη κατανομή).
  - Για  $c = \frac{1}{2}$ , οι εκ των προτέρων κατανομές που θα προκύψουν για τα  $\pi$  και  $\pi_j$ , θα είναι όλες  $Beta\left(\frac{1}{2}, \frac{1}{2}\right)$  (Jeffreys κατανομή).

### 3.2.1 Παράδειγμα (1) με προσομοιωμένα δεδομένα

Προσομειώνουμε 100 παρατηρήσεις από κατανομή  $Binomial(10, 0.5)$  και 100 παρατηρήσεις από κατανομή  $Binomial(10, 0.5)$ . Στο σχήμα (3.7) βλέπουμε τα ραβδόγραμματα των παρατηρήσεων.



Σχήμα 3.7: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από  $Binomial(0.5)$ .

- Για τις παραμέτρους  $\pi$ ,  $\pi_1$  και  $\pi_2$  αντίστοιχα, επιλέγουμε τις εκ των προτέρων κατανομές

$$\diamond \text{prior 1 : } \pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right), \quad \pi_j \sim \text{Beta}\left(\frac{N_j}{2N}, \frac{N_j}{2N}\right)$$

$$\diamond \text{prior 2 : } \pi \sim \text{Beta}(\bar{y}, 1 - \bar{y}), \quad \pi_j \sim \text{Beta}\left(\frac{N_j \bar{y}}{N}, \frac{N_j(1 - \bar{y})}{N}\right)$$

$$\diamond \text{prior 3 : } \pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right), \quad \pi_j \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\diamond \text{prior 4 : } \pi \sim \text{Beta}(1, 1), \quad \pi_j \sim \text{Beta}(1, 1), \quad \text{όπου } j = 1, 2$$

• Με την βοήθεια της  $R$  προκύπτει ο πίνακας (3.4). Διαπιστώνουμε ότι οι συνδυασμοί *prior 1* και *prior 2* που είναι υπό περιστώσεις της ίδιας γενικής μορφής, δίνουν σχεδόν τις ίδιες τιμές για τον παράγοντα Bayes. Το ίδιο συμβαίνει και για τους συνδυασμούς *prior 3* και *prior 4*. Ωστόσο και οι τέσσερις συνδυασμοί δίνουν την ίδια συμπερασματολογία.

	<i>prior 1</i>	<i>prior 2</i>	<i>prior 3</i>	<i>prior 4</i>
$BF_{10}$	0.075	0.0749	0.209	0.326
$\log(BF)_{10}$	-2.589	-2.59	-1.567	-1.119
Ένδειξη	$H_1$	$H_1$	$H_1$	$H_1$
εναντίον της	Θετική	Θετική	Θετική	Θετική

Πίνακας 3.4: Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.

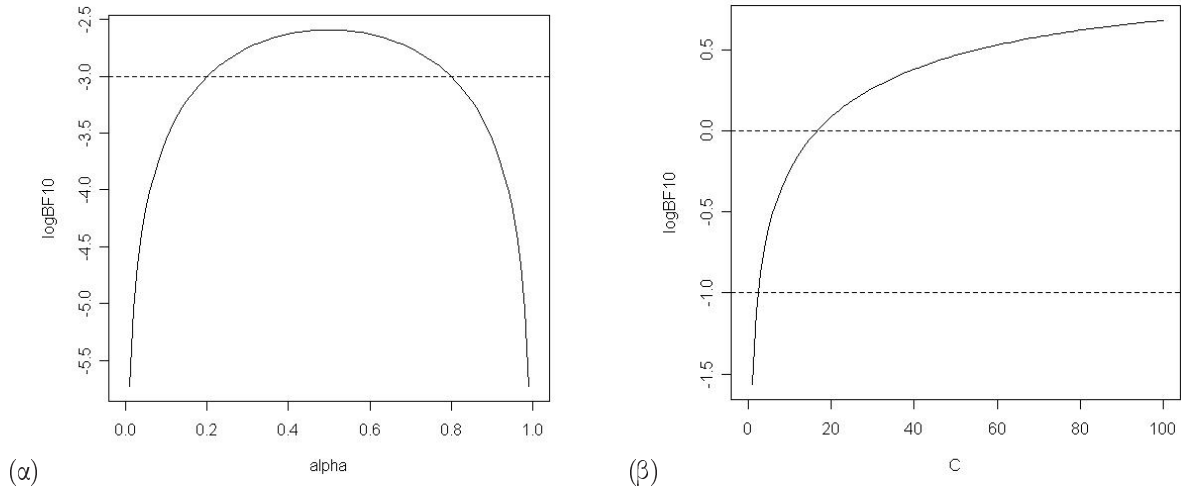
• Στο σχήμα (3.8) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων  $\alpha$  &  $c$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\pi$ ,  $\pi_1$  και  $\pi_2$  αντίστοιχα.

$$\diamond \text{Γενική μορφή 1: } \pi \sim \text{Beta}(\alpha, 1 - \alpha), \quad \pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$$

Υπό γενική μορφή των δύο πρώτων συνδυασμών, φαίνεται ότι για όλες τις τιμές της παραμέτρου  $\alpha$  αποδεχόμαστε την μηδενική υπόθεση.

$$\diamond \text{Γενική μορφή 2: } \pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right), \quad \pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$$

Υπό την γενική μορφή του τρίτου και τέταρτου συνδυασμού, φαίνεται επίσης ότι για όλες τις τιμές της παραμέτρου  $c$  αποδεχόμαστε την μηδενική υπόθεση.



Σχήμα 3.8: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων  $\alpha$  &  $c$  για (α) την γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$  και (β) την γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$  (Παράδειγμα 3.2.1).

### 3.2.2 Παράδειγμα (2) με προσομοιωμένα δεδομένα

Προσομοιώνουμε 100 παρατηρήσεις από κατανομή  $\text{Binomial}(10, 0.3)$  και 100 παρατηρήσεις από κατανομή  $\text{Binomial}(10, 0.7)$ . Στο σχήμα (3.9) βλέπουμε τα ραβδογράμματα των παρατηρήσεων.

- Για τις παραμέτρους  $\pi$ ,  $\pi_1$  και  $\pi_2$  αντίστοιχα, επιλέγουμε τις εκ των προτέρων κατανομές

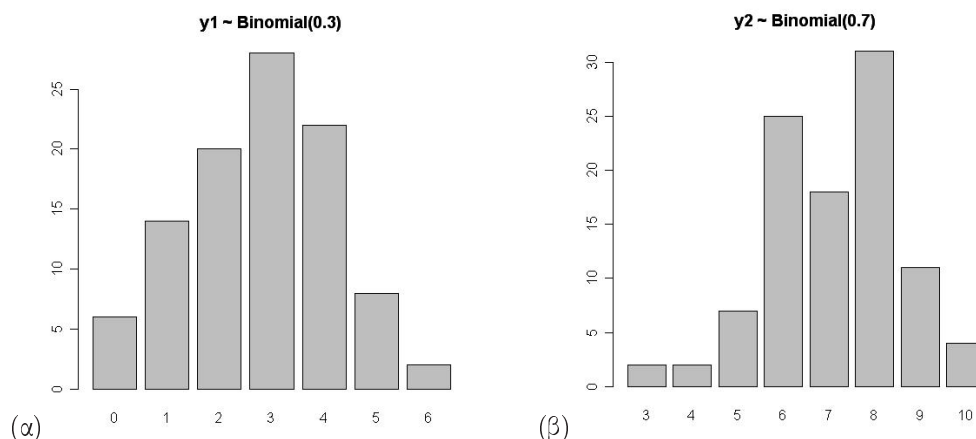
$$\diamond \text{prior 1} : \pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right), \quad \pi_j \sim \text{Beta}\left(\frac{N_j}{2N}, \frac{N_j}{2N}\right)$$

$$\diamond \text{prior 2} : \pi \sim \text{Beta}(\bar{y}, 1 - \bar{y}), \quad \pi_j \sim \text{Beta}\left(\frac{N_j}{N}\bar{y}, \frac{N_j}{N}(1 - \bar{y})\right)$$

$$\diamond \text{prior 3} : \pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right), \quad \pi_j \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$\diamond \text{prior 4} : \pi \sim \text{Beta}(1, 1), \quad \pi_j \sim \text{Beta}(1, 1), \quad \text{όπου } j = 1, 2$$

- Με την βοήθεια της  $R$  προκύπτει ο πίνακας (3.5). Όπως και στο προηγούμενο παράδειγμα, οι συνδυασμοί *prior 1* και *prior 2* δίνουν σχεδόν τις ίδιες τιμές για τον παράγοντα Bayes και το ίδιο συμβαίνει για τους συνδυασμούς *prior 3* και *prior 4*. Ωστόσο και οι τέσσερις συνδυασμοί δίνουν την ίδια συμπερασματολογία.
- Στο σχήμα (3.10) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων  $\alpha$  &  $c$  στην γενική



Σχήμα 3.9: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από (α)  $Binomial(0.3)$  και (β)  $Binomial(0.7)$ .

<i>prior</i>	<i>prior 1</i>	<i>prior 2</i>	<i>prior 3</i>	<i>prior 4</i>
$BF_{10}$	$5.604 \times 10^{82}$	$5.599 \times 10^{82}$	$1.4 \times 10^{83}$	$1.79 \times 10^{83}$
$\log(BF)_{10}$	$1.905 \times 10^2$	$1.9 \times 10^2$	$1.92 \times 10^2$	$1.92 \times 10^2$
Ένδειξη εναντίον της	$H_0$ Πολύ Ισχυρή	$H_0$ Πολύ Ισχυρή	$H_0$ Πολύ Ισχυρή	$H_0$ Πολύ Ισχυρή

Πίνακας 3.5: Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.

μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\pi$ ,  $\pi_1$  και  $\pi_2$  αντίστοιχα.

◇ Γενική μορφή 1:  $\pi \sim Beta(\alpha, 1 - \alpha)$ ,  $\pi_j \sim Beta\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$

Υπό γενική μορφή των δύο πρώτων συνδυασμών, για όλες τις τιμές της παραμέτρου  $\alpha \in (0, 1)$  απορρίπτουμε την μηδενική υπόθεση.

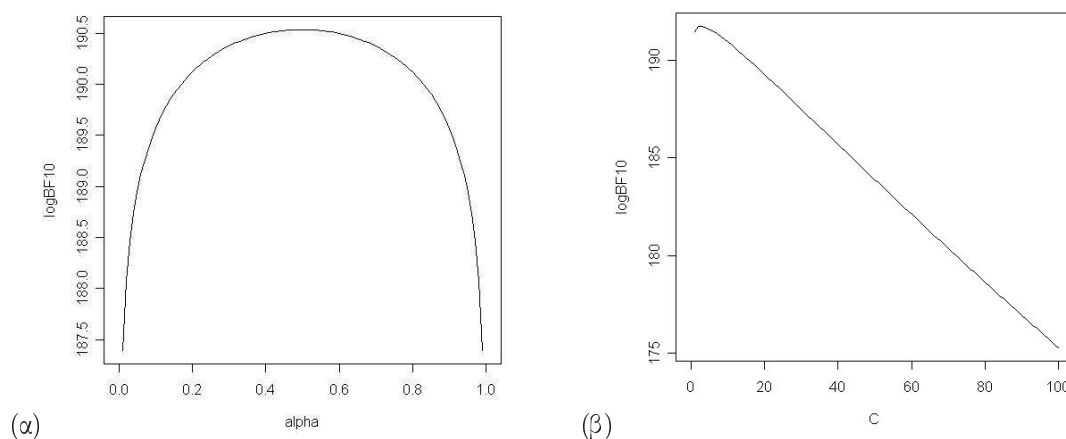
◇ Γενική μορφή 2:  $\pi \sim Beta\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim Beta\left(\frac{c}{2}, \frac{c}{2}\right)$

Υπό την γενική μορφή του τρίτου και τέταρτου συνδυασμού, η μηδενική υπόθεση απορρίπτεται για τιμές της παραμέτρου  $c \in (0, 10^5)$ .

### 3.2.3 Παράδειγμα με πραγματικά δεδομένα

Συνέχεια από το παράδειγμα της παραγράφου 2.5.2. Συγκρίνουμε ανά δύο τα πέντε αντιβιοτικά (Amikacin[1], Gentamicin[2], Netilmicin[3], Sisomicin[4], Tobramycin[5]) επιλέγοντας την εκ των προτέρων κατανομή  $Beta\left(\frac{1}{2}, \frac{1}{2}\right)$ , που είναι μια αρκετά πληροφοριακή εκ των προτέρων με μέσο  $E(\pi) = \frac{1}{2}$  και  $Var(\pi) = \frac{1}{8}$ .





Σχήμα 3.10: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων  $\alpha$  &  $c$  για (α) την γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$  και (β) την γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$  (Παράδειγμα 3.2.2).

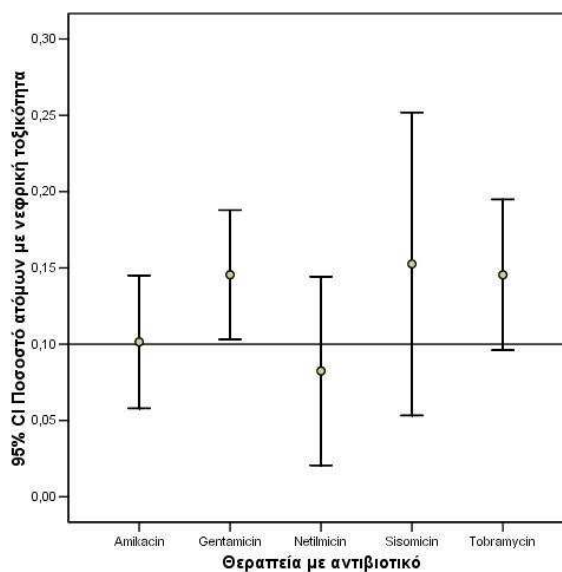
• Με την βοήθεια της  $R$  προκύπτουν οι πίνακες (3.6) και (3.7). Από τους πίνακες (3.6) και (3.7) καθώς και από το σχήμα (3.11) μπορούμε να συμπεράνουμε ότι τα αντιβιοτικά Gentamicin[2], Sisomicin[4] και Tobramycin[5] έχουν περίπου το ίδιο ποσοστό εμφάνισης νεφρικής τοξικότητας. Το Amikacin[1] δίνει περίπου το ίδιο ποσοστό με το Gentamicin[2], όμως το Netilmicin[3] είναι εκείνο που διαφέρει σημαντικά με όλα τα υπόλοιπα και δίνει το μικρότερο ποσοστό εμφάνισης νεφρικής τοξικότητας.

Ζεύγη συγκρίσεων	[1] – [2]	[1] – [3]	[1] – [4]	[1] – [5]	[4] – [5]
$BF_{10}$	0.196	8.989	41.89	0.52	0.47
$\log(BF)_{10}$	-1.629	2.196	3.735	-0.65	-0.75
Ένδειξη εναντίον της	$H_1$ Θετική	$H_0$ Θετική	$H_0$ Ισχυρή	$H_1$ Αδιάφορη	$H_1$ Αδιάφορη

Πίνακας 3.6: Αποτελέσματα του παράγοντα Bayes για τις ανά δύο συγκρίσεις υπό την  $\text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$  εκ των προτέρων κατανομή.

Ζεύγη συγκρίσεων	[2] – [3]	[2] – [4]	[2] – [5]	[3] – [4]	[3] – [5]
$BF_{10}$	$6.646 \times 10^5$	1.968	0.0448	$1.2 \times 10^7$	$6.68 \times 10^5$
$\log(BF)_{10}$	$1.34 \times 10$	0.677	-3.105	$1.63 \times 10$	$1.34 \times 10$
Ένδειξη εναντίον της	$H_0$	$H_0$	$H_1$	$H_0$	$H_0$
	Πολύ Ισχυρή	Αδιάφορη	Ισχυρή	Πολύ Ισχυρή	Πολύ Ισχυρή

Πίνακας 3.7: Αποτελέσματα του παράγοντα Bayes για τις ανά δύο συγκρίσεις υπό την  $Beta(\frac{1}{2}, \frac{1}{2})$  εκ των προτέρων κατανομή.



Σχήμα 3.11: Errorbar με τα ποσοστά των ατόμων με νεφρική τοξικότητα και σχετικές παρενέργειες για κάθε αντιβιοτικό (Παράδειγμα 3.2.3).

### 3.3 Έλεγχος υποθέσεων για δύο συσχετιζόμενα δείγματα από κατανομή Poisson

Έστω δύο τυχαία συσχετιζόμενα δείγματα μεγέθους  $n$  από κατανομές Poisson. Γράφοντας  $Y_1 = V_1 + V_3$  και  $Y_2 = V_2 + V_3$  με  $V_k \sim Poisson(\lambda_k)$  τότε

$$Y_1 \sim Poisson(\lambda_1 + \lambda_3) \text{ \& } Y_2 \sim Poisson(\lambda_2 + \lambda_3), \text{ και } Cov(Y_1, Y_2) = \lambda_3$$

Είναι προφανές ότι στην περίπτωση που  $\lambda_3 = 0$  τότε τα δείγματα είναι ανεξάρτητα. Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0 : \lambda_1 = \lambda_2 = \lambda \quad vs \quad H_1 : \lambda_1 \neq \lambda_2$$

Όπως και στην κλασσική στατιστική, υπολογίζουμε τις διαφορές των παρατηρήσεων, δηλαδή  $Z = Y_1 - Y_2$ , οπότε το πρόβλημα ανάγεται σε πρόβλημα ενός δείγματος μεγέθους  $n$ . Η συνάρτηση πιθανότητας των  $z_i$ ,  $i = 1, \dots, n$ , είναι επίσης μια διακριτή κατανομή, που ορίζεται ως το σύνολο των ακέραιων αριθμών ( $\mathbb{Z}$ ).

Υποθέτουμε ότι οι μεταβλητές  $Y_1$  και  $Y_2$ , από κοινού, ακολουθούν μια Διμεταβλητή Poisson κατανομή με συνάρτηση πιθανότητας

$$f_{Y_1, Y_2}(y_1, y_2 | \lambda_1, \lambda_2, \lambda_3) = e^{-(\lambda_1 + \lambda_2 + \lambda_3)} \frac{\lambda_1^{y_1} \lambda_2^{y_2}}{y_1! y_2!} \sum_{i=0}^{\min(y_1, y_2)} \binom{y_1}{i} \binom{y_2}{i} i! \left( \frac{\lambda_3}{\lambda_1 \lambda_2} \right)^i$$

Οπότε η τυχαία μεταβλητή  $Z$  θα ακολουθεί την Skellam κατανομή με συνάρτηση πιθανότητας

$$f_Z(z | \lambda_1, \lambda_2) = e^{-(\lambda_1 + \lambda_2)} \left( \frac{\lambda_1}{\lambda_2} \right)^{z/2} I_{|z|}(2\sqrt{\lambda_1 \lambda_2}), \quad z \in \mathbb{Z}$$

όπου  $I_r(x)$  είναι μια Bessel συνάρτηση  $r$  τάξης που ορίζεται από τον τύπο

$$I_r(x) = \left( \frac{x}{2} \right)^r \sum_{k=0}^{\infty} \frac{\left( \frac{x^2}{4} \right)^k}{k! \Gamma(r + k + 1)}$$

Η Skellam κατανομή είναι η κατανομή πιθανότητας της διαφοράς δύο συσχετιζόμενων ή ανεξάρτητων τυχαίων μεταβλητών  $Y_1$  και  $Y_2$  από κατανομή Poisson με διαφορετικούς μέσους. Παρ' όλο που η κατανομή της διαφοράς των συσχετιζόμενων ή μη τυχαίων μεταβλητών είναι ίδια, η ερμηνεία των παραμέτρων διαφέρει.

Παρατηρούμε ότι οι διαφορές  $z_i$  δεν εξαρτώνται από την συνδιακύμανση  $\lambda_3$  οπότε ορίζουμε τις εκ των προτέρων κατανομές μόνο για τις παραμέτρους  $\lambda_1$  και  $\lambda_2$ .

$$\lambda_1 \sim \text{Gamma}(\alpha_1, \beta_1) \quad \lambda_2 \sim \text{Gamma}(\alpha_2, \beta_2)$$

Η συνάρτηση πιθανοφάνειας των διαφορών  $z_i$  δίνεται από τον τύπο

$$f(\mathbf{z}|\lambda_1, \lambda_2) = e^{-n(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{(1/2)\sum z_i} \prod_{i=1}^n I_{|z_i|}(2\sqrt{\lambda_1\lambda_2})$$

Από τα παραπάνω υπολογίζεται η εκ των υστέρων κατανομή των διαφορών  $z_i$ .

$$f(\lambda_1, \lambda_2|\mathbf{z}) \propto f(\mathbf{z}|\lambda_1, \lambda_2)f(\lambda_1)f(\lambda_2)$$

Επειδή ο υπολογισμός δεν είναι εύκολος, μετασχηματίζουμε τα δεδομένα υπό την προϋπόθεση ότι  $z_i = v_i - u_i$ , εισάγοντας τα λανθάνοντα δεδομένα :

$$v_i \sim \text{Poisson}(\lambda_1) \quad \& \quad u_i \sim \text{Poisson}(\lambda_2)$$

όπου  $\mathbf{v} = (v_1, \dots, v_n)$  και  $\mathbf{u} = (u_1, \dots, u_n)$ .

Η από κοινού εκ των υστέρων συνάρτηση πιθανότητας των παραμέτρων ( $\lambda_1$  &  $\lambda_2$ ) και των λανθάνοντων δεδομένων ( $\mathbf{v}$  &  $\mathbf{u}$ ) δίνεται από τον τύπο:

$$f(\mathbf{v}, \mathbf{u}, \lambda_1, \lambda_2|\mathbf{z}) \propto f(\mathbf{z}|\mathbf{v}, \mathbf{u}, \lambda_1, \lambda_2)f(\lambda_1, \lambda_2)$$

$$\propto e^{-n(\lambda_1+\lambda_2)} \lambda_1^{\sum v_i} \lambda_2^{\sum u_i} \left\{ \prod_{i=1}^n \frac{g(z_i = v_i - u_i)}{v_i!u_i!} \right\} f(\lambda_1, \lambda_2)$$

$$\propto e^{-(n+\beta_1)\lambda_1 - (n+\beta_2)\lambda_2} \lambda_1^{n\bar{v}+\alpha_1-1} \lambda_2^{n\bar{u}+\alpha_2-1} \left\{ \prod_{i=1}^n \frac{g(z_i = v_i - u_i)}{v_i!u_i!} \right\}$$

όπου  $\bar{v} = \sum_{i=1}^n v_i/n$  και  $\bar{u} = \sum_{i=1}^n u_i/n$ . Χρησιμοποιώντας την παραπάνω κατανομή ως κατανομή στόχο (target distribution) κατασκευάζεται ένας *MCMC* (Markov Chain Monte Carlo) αλγόριθμος ώστε να προσομοιώσουμε τυχαίες τιμές από αυτήν.

1. Δειγματοληπτούμε  $(v_i, u_i)$  από την συνάρτηση

$$f(v_i, u_i|z_i = v_i - u_i, \lambda_1, \lambda_2) \propto \frac{\lambda_1^{v_i} \lambda_2^{u_i}}{v_i!u_i!} g(z_i = v_i - u_i), \quad i = 1, \dots, n.$$

2. Δειγματοληπτούμε  $\lambda_1$  από την συνάρτηση

$$f(\lambda_1|\lambda_2, \mathbf{v}, \mathbf{u}) \sim \text{Gamma}(n\bar{v} + \alpha_1, n + \beta_1).$$

3. Δειγματοληπούμε  $\lambda_2$  από την συνάρτηση

$$f(\lambda_2 | \lambda_1, \mathbf{v}, \mathbf{u}) \sim \text{Gamma}(n\bar{u} + \alpha_2, n + \beta_2).$$

Για να ενημερώσουμε τα προσαυξημένα δεδομένα  $(v_i, u_i)$  του πρώτου βήματος, χρησιμοποιείται ο ακόλουθος αλγόριθμος Metropolis.

- Εάν  $z_i < 0$  και  $(v_i, u_i)$  οι τρέχουσες τιμές των προσαυξημένων δεδομένων τότε
  - Προτείνεται  $v'_i \sim \text{Poisson}(\lambda_1)$  και  $u'_i = v'_i - z_i$ .
  - Δεχόμαστε την παραπάνω πρόταση με πιθανότητα  $\alpha = \min \left\{ 1, \lambda_2^{\frac{(v'_i - v_i)(v_i - z_i)!}{(v'_i - z_i)!}} \right\}$ .
- Εάν  $z_i \geq 0$  και  $(v_i, u_i)$  οι τρέχουσες τιμές των προσαυξημένων δεδομένων τότε
  - Προτείνεται  $u'_i \sim \text{Poisson}(\lambda_2)$  και  $v'_i = u'_i + z_i$ .
  - Δεχόμαστε την παραπάνω πρόταση με πιθανότητα  $\alpha = \min \left\{ 1, \lambda_1^{\frac{(u'_i - u_i)(u_i - z_i)!}{(u'_i - z_i)!}} \right\}$ .

Έχοντας πλέον «υπολογίσει» την εκ των υστέρων κατανομή των διαφορών  $z_i$  προχωράμε στον υπολογισμό του παράγοντα Bayes. Επειδή όμως η εκτίμηση του παράγοντα Bayes είναι πολύ πιο σύνθετη και προϋποθέτει την χρήση πολύπλοκων μεθόδων *MCMC*, για περισσότερες λεπτομέρειες παραπέμπουμε στην εργασία των Karlis & Ntzoufras (2006) [33].

### 3.4 Έλεγχος υποθέσεων για δύο συσχετιζόμενες αναλογίες

Για να εξετάσουμε δύο συσχετιζόμενες αναλογίες, μπορούμε να εκφράσουμε τα δεδομένα με την μορφή ενός  $2 \times 2$  πίνακα συνάφειας και κατόπιν να γίνει ο έλεγχος της ισότητας των αναλογιών. Στην κλασσική στατιστική ο αντίστοιχος έλεγχος επιτυγχάνεται με την χρήση του McNemar test. Ακολουθώντας την ίδια ιδεολογία με το McNemar test τα κελιά που μας δίνουν πληροφορία είναι αυτά της μη-διαγωνίου όπου οι δύο παράγοντες διαφωνούν. Ο έλεγχος μπορεί

να γίνει με δύο προσεγγίσεις (υπό συνθήκη ή χωρίς συνθήκη - conditional or unconditional) που όμως καταλήγουν στον ίδιο Bayes Factor συμφωνώντας με την θεωρία των μεγάλων δειγμάτων (Kateri et al. 2001)[35].

### Χωρίς συνθήκη προσέγγιση (unconditional)

Έστω  $n_{ij}$  ένας  $2 \times 2$  πίνακας συνάφειας με τις παρατηρούμενες συχνότητες των δύο συσχετιζόμενων αναλογιών  $\pi_1$  &  $\pi_{\cdot 1}$ .

	$B_1$	$B_2$	
$A_1$	$n_{11}$	$n_{12}$	$n_{1\cdot}$
$A_2$	$n_{21}$	$n_{22}$	
	$n_{\cdot 1}$		

(α)

	$B_1$	$B_2$	
$A_1$	$\pi_{11}$	$\pi_{12}$	$\pi_{1\cdot}$
$A_2$	$\pi_{21}$	$\pi_{22}$	
	$\pi_{\cdot 1}$		

(β)

Σχήμα 3.12: Πίνακας (α) συχνοτήτων και (β) πιθανοτήτων των παραγόντων A & B.

Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0 : \pi_1 = \pi_{\cdot 1} \quad vs \quad H_1 : \pi_1 \neq \pi_{\cdot 1}$$

Η οποία μεταφράζεται ισοδύναμα στον έλεγχο της υπόθεσης ότι

$$H_0 : \pi_{12} = \pi_{21} = \pi^* \quad vs \quad H_1 : \pi_{12} \neq \pi_{21}$$

Υποθέτουμε ότι τα δεδομένα ακολουθούν Multinomial κατανομή οπότε

$$\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22}) \sim Mult(N, \pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}) : \sum_{ij} \pi_{ij} = 1$$

Στην συμπερασματολογία των πινάκων συνάφειας οι πιθανότητες  $\pi_{ij}$  εκτιμώνται από τις αντίστοιχες δειγματικές πιθανότητες  $p_{ij}$ . Κάτω από την υπόθεση της συμμετρίας, οι μη-διαγώνιες πιθανότητες ( $\pi_{12}$  &  $\pi_{21}$ ) εκτιμώνται από την σχέση  $(\pi_{12} + \pi_{21})/2$ . Συνεπώς οι συσχετιζόμενες αναλογίες  $\pi_1$  &  $\pi_{\cdot 1}$  εκτιμώνται από

την σχέση  $\pi_{11} + (\pi_{12} + \pi_{21})/2$ .

Χρησιμοποιώντας μια συζηγή εκ των προτέρων κατανομή για τις πιθανότητες, υποθέτουμε ότι

$$(\pi_{11}, \pi_{12}, \pi_{21}) \sim \text{Dirichlet}(a_{11}, a_{12}, a_{21}, a_{22}), \quad a_{ij} > 0$$

και καταλήγουμε στην εκ των υστέρων

$$(\pi_{11}, \pi_{12}, \pi_{21}) | \mathbf{n} \sim \text{Dirichlet}(n_{11} + a_{11}, n_{12} + a_{12}, n_{21} + a_{21}, n_{22} + a_{22})$$

Έτσι για το μηδενικό μοντέλο ( $M_0$ ) έχουμε ότι  $(\pi_{11}, \pi_{12} + \pi_{21})$  θα ακολουθεί μια Dirichlet κατανομή με παραμέτρους  $a_{11}, a_{12} + a_{21}, a_{22}$  ή αλλιώς

$$(\pi_{11}, 2\pi^*) \sim \text{Dirichlet}(a_{11}, a_{12} + a_{21}, a_{22})$$

από όπου θα πάρουμε την περιθώρια πιθανοφάνεια

$$\begin{aligned} f(\mathbf{n} | M_0) &= \int_0^1 \int_0^1 f(\mathbf{n} | \pi_{11}, \pi^*) f(\pi_{11}, \pi^*) d\pi_{11} d\pi^* \\ &= C \frac{\Gamma(a_o) \Gamma(n_{11} + a_{11}) \Gamma(n + a_{12} + a_{21}) \Gamma(n_{22} + a_{22})}{2^n \Gamma(a_{11}) \Gamma(a_{12} + a_{21}) \Gamma(a_{22}) \Gamma(N + a_o)} \end{aligned}$$

Ενώ για το εναλλακτικό μοντέλο ( $M_1$ ) θα πάρουμε την περιθώρια πιθανοφάνεια

$$\begin{aligned} f(\mathbf{n} | M_1) &= \int_0^1 \int_0^1 \int_0^1 f(\mathbf{n} | \pi_{11}, \pi_{12}, \pi_{21}) f(\pi_{11}, \pi_{12}, \pi_{21}) d\pi_{11} d\pi_{12} d\pi_{21} \\ &= C \frac{\Gamma(a_o) \Gamma(n_{11} + a_{11}) \Gamma(n_{12} + a_{12}) \Gamma(n_{21} + a_{21}) \Gamma(n_{22} + a_{22})}{\Gamma(a_{11}) \Gamma(a_{12}) \Gamma(a_{21}) \Gamma(a_{22}) \Gamma(N + a_o)} \end{aligned}$$

όπου

- $C = \binom{N}{n_{11} n_{12} n_{21} n_{22}} = \frac{N!}{n_{11}! n_{12}! n_{21}! n_{22}!}$ , ο πολυωνυμικός συντελεστής
- $a_o = \sum a_{ij}$
- $n = n_{12} + n_{21}$

## Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$BF_{10} = \frac{2^n \Gamma(a_{12} + a_{21}) \Gamma(n_{12} + a_{12}) \Gamma(n_{21} + a_{21})}{\Gamma(a_{12}) \Gamma(a_{21}) \Gamma(n + a_{12} + a_{21})} \quad (3.1)$$

Διαπιστώνουμε ότι ο παράγοντας Bayes δεν επηρεάζεται από τα διαγώνια στοιχεία ( $n_{11}$  &  $n_{22}$ ) αλλά ούτε και από τις αντίστοιχες παραμέτρους ( $a_{11}$  &  $a_{22}$ ), όπως συμβαίνει και στο McNemar test.

### Υπό συνθήκη προσέγγιση (conditional)

Στην υπό συνθήκη προσέγγιση, θέτουμε  $\pi^* = \pi_{12}/(\pi_{12} + \pi_{21})$  να είναι η πιθανότητα μιας παρατήρησης να βρεθεί στο κελί (1, 2) δοθέντος ότι θα κατανεμηθεί σε ένα από τα μη-διαγώνια κελιά. Οπότε η υπόθεση μπορεί ισοδύναμα να πάρει την μορφή

$$H_0 : \pi^* = \frac{1}{2} \quad vs \quad H_1 : \pi^* \neq \frac{1}{2}$$

Από τα παραπάνω έχουμε ότι  $n_{12} \sim \text{Binomial}(n, \pi^*)$ , όπου  $n = n_{12} + n_{21}$ , και εκ των προτέρων κατανομή, για το εναλλακτικό μοντέλο,  $\pi^* \sim \text{Beta}(a_{12}, a_{21})$  που είναι υπό περίπτωση της Dirichlet κατανομής. Υπολογίζοντας τον παράγοντα Bayes θα καταλήξουμε στην σχέση (3.1).

### Επιλογή της εκ των προτέρων κατανομής

Ανάλογα με την εκ των προτέρων πεποίθησή μας μπορούμε να διακρίνουμε τις ακόλουθες περιπτώσεις.

- Ισχυρή πίστη ότι  $\pi_{1.} > \pi_{.1}$  αντιπροσωπεύεται από την σχέση  $a_{12} \gg a_{21}$
- Μεγάλη αβεβαιότητα ότι  $\pi_{1.} > \pi_{.1}$  αντιπροσωπεύεται από την σχέση

$$a_{12} = a_{21} = q \quad : \quad q \in [0, 1]$$



Όταν  $q \in [0, 1]$ , η πιθανότητα  $P\left(\frac{\pi_{12}}{(\pi_{12} + \pi_{21})} > \frac{1}{2}\right) \Rightarrow P(\pi_{1.} > \pi_{.1})$  είναι μικρότερη από από το αντίστοιχο p-value της κλασικής προσέγγισης και η διαφορά των δύο «επιπέδων σημαντικότητας» δεν είναι μεγαλύτερη από την πιθανότητα των δεδομένων υπό την μηδενική υπόθεση, Altman (1969) [8].

- Ισχυρή πίστη ότι  $\pi_{1.} = \pi_{.1}$  αντιπροσωπεύεται από την σχέση  $a_{12} = a_{21} = k \geq 1$

Αν για παράδειγμα πάρουμε την περίπτωση όπου  $a_{12} = a_{21} = 1$  (δηλ. μια Uniform εκ των προτέρων κατανομή για το  $\pi^*$ ) τότε η σχέση (3.1) μειώνεται στην

$$BF_{10} = \frac{2^n \Gamma(n_{12} + 1) \Gamma(n_{21} + 1)}{\Gamma(n + 2)} \quad (3.2)$$

που είναι ο παράγοντας Bayes όταν ελέγχεται η μηδενική υπόθεση ότι μια διωνυμική παράμετρος ισούται με 1/2 (Bernardo & Smith 1994, σελ. 414)[11].

### 3.4.1 Παράδειγμα με πραγματικά δεδομένα

Σε 18 ασθενείς χορηγούνται δύο φάρμακα (A & B). Μετά την λήψη των φαρμάκων μετρήθηκαν πόσοι ασθενείς ανάρρωσαν και πόσοι όχι (Leslie et al. 1991, σελ. 130-132)[38]. Ο πίνακας (3.8) δίνει τα αποτελέσματα.

		Φάρμακο A		
		Επιτυχία	Αποτυχία	
Φάρμακο B	Επιτυχία	1 (0.055)	3 (0.16)	4 (0.22)
	Αποτυχία	9 (0.5)	5 (0.27)	14 (0.77)
Σύνολο		10 (0.55)	8 (0.44)	18 (1)

Πίνακας 3.8: Πίνακας συχνότητων και πιθανοτήτων.

- Ανάλογα με την εκ των προτέρων πεποίθησή μας θα πάρουμε διάφορες περιπτώσεις εκ των προτέρων κατανομών και από την σχέση (3.1) θα προκύψουν οι πίνακες (3.9), (3.10) και (3.11).

$\alpha_{12} = \alpha_{21} = q : q \in [0, 1]$	0.1	0.3	0.5	0.7	0.9	1
$BF_{10}$	0.35	0.8	1.08	1.26	1.38	1.43
$\log(BF)_{10}$	-1.05	-0.22	0.08	0.23	0.33	0.36
Ένδειξη εναντίον της	$H_1$ Θετική	$H_1$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη

Πίνακας 3.9: Μεγάλη αβεβαιότητα ότι  $\pi_1 > \pi_1$  αντιπροσωπεύεται από την σχέση  $\alpha_{12} = \alpha_{21} = q$ .

$\alpha_{12} = \alpha_{21} = k$	1	3	5	7	9	15
$BF_{10}$	1.43	1.65	1.59	1.5	1.44	1.3
$\log(BF)_{10}$	0.36	0.5	0.46	0.4	0.36	0.26
Ένδειξη εναντίον της	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη

Πίνακας 3.10: Ισχυρή πίστη ότι  $\pi_1 = \pi_1$  αντιπροσωπεύεται από την σχέση  $\alpha_{12} = \alpha_{21} = k \geq 1$ 

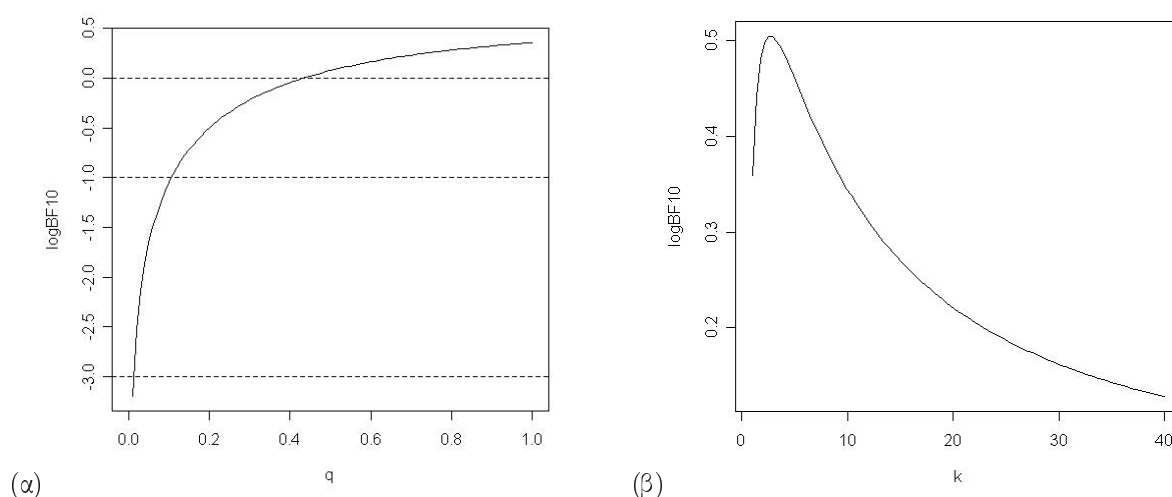
$Beta(\alpha, 1 - \alpha)$	$Beta(0.1, 0.9)$	$Beta(0.3, 0.7)$	$Beta(0.4, 0.6)$	$Beta(0.7, 0.3)$	$Beta(0.9, 0.1)$
$BF_{10}$	0.536	1.096	1.148	0.708	0.223
$\log(BF)_{10}$	-0.624	0.09	0.138	-0.346	-1.5
Ένδειξη εναντίον της	$H_1$ Αδιάφορη	$H_0$ Αδιάφορη	$H_0$ Αδιάφορη	$H_1$ Αδιάφορη	$H_1$ Θετική

Πίνακας 3.11: Εκ των προτέρων κατανομή η power εκ των προτέρων  $Beta(\alpha, 1 - \alpha)$ , όπου  $\alpha \in [0, 1]$ .

- Από τους παραπάνω πίνακες έχουμε ότι
  - ◇ Με τον πίνακα (3.9) εκφράζεται η αβεβαιότητα μας στο γεγονός ότι το φάρμακο A έχει μεγαλύτερη αποτελεσματικότητα από το φάρμακο B. Συνεπώς θέτουμε  $\alpha_{12} = \alpha_{21} = q : q \in [0, 1]$  και καταλήγουμε στην αποδοχή της μηδενικής υπόθεσης για όλες τις τιμές  $q \in [0, 1]$ .
  - ◇ Με τον πίνακα (3.10) εκφράζεται η ισχυρή πεποίθηση ότι και τα δύο φάρμακα έχουν την ίδια αποτελεσματικότητα. Συνεπώς θέτουμε  $\alpha_{12} = \alpha_{21} = k : k \geq 1$  και διαπιστώνουμε ότι η διαφορά των δύο φαρμάκων δεν είναι άξια αναφοράς.
  - ◇ Με τον πίνακα (3.11) χρησιμοποιούμε μια power εκ των προτέρων κατανομή διότι πιστεύουμε ότι τα φάρμακα A και B δεν έχουν την ίδια αποτελεσματικότητα. Παρ' όλα αυτά τα δεδομένα υπερσχύουν και πάλι διαπιστώνουμε ότι ισχύει η μηδενική υπόθεση.

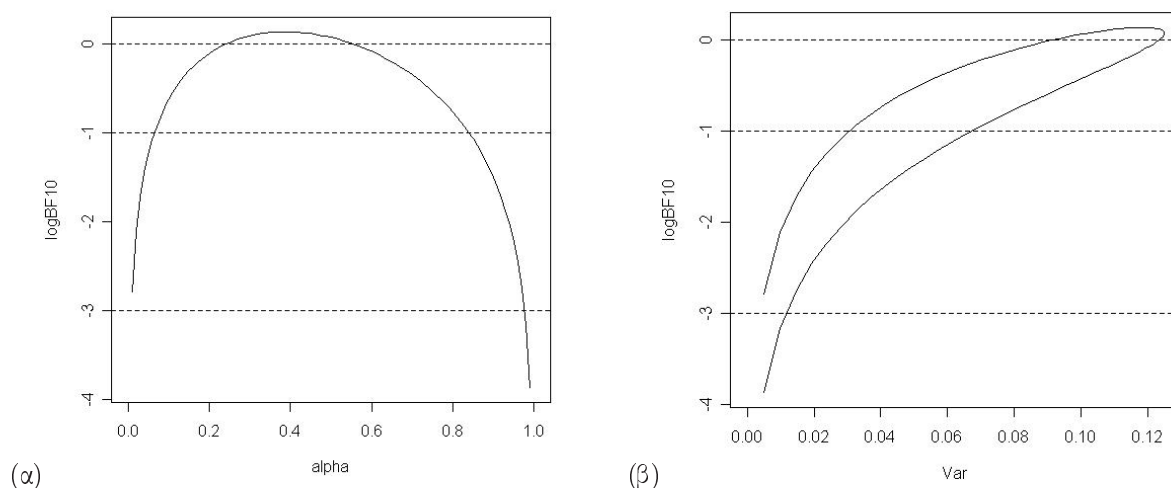
Και στις τρεις περιπτώσεις εξάγουμε την ίδια συμπερασματολογία ως προς την αποτελεσματικότητα των δύο φαρμάκων. Ωστόσο ο βαθμός της πληροφορίας που περιέχεται στα δεδομένα υπέρ της διαφοράς έναντι της μη διαφοράς των δύο φαρμάκων, αλλάζει ανάλογα με την εκ των προτέρων πεποίθησή μας.

- Στο σχήμα (3.13) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των  $q$  και  $k$ . Για τις διάφορες τιμές των παραμέτρων  $q$  και  $k$  δεχόμαστε την μηδενική υπόθεση, δηλ. ότι τα δύο φάρμακα παρουσιάζουν την ίδια αποτελεσματικότητα.



Σχήμα 3.13: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές (α) του  $q$  και (β) του  $k$  (Παράδειγμα 3.4.1).

- Στο σχήμα (3.14) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$  :  $\alpha \in (0, 1)$  της εκ των προτέρων κατανομής  $Beta(\alpha, 1 - \alpha)$ . Από το σχήμα διαπιστώνουμε ότι για τις διάφορες τιμές της παραμέτρου  $\alpha$ , αποδεχόμαστε την μηδενική υπόθεση.



Σχήμα 3.14: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes της εκ των προτέρων κατανομής  $Beta(\alpha, 1 - \alpha)$  ως προς (α) τις διάφορες τιμές της  $\alpha \in [0, 1]$  και (β) της διασποράς της (Παράδειγμα 3.4.1).

### 3.5 Έπιλογος

Στο τρίτο κεφάλαιο υπολογίσαμε τον παράγοντα Bayes για τον έλεγχο υποθέσεων δύο ανεξάρτητων ή συσχετισμένων δειγμάτων από κατανομές Poisson και Binomial. Διαπιστώσαμε ότι ο υπολογισμός του παράγοντα Bayes απαιτεί την επίλυση υψηλόβαθμων διαστάσεων ολοκληρωμάτων. Έχοντας επιλέξει συζυγείς εκ των προτέρων κατανομές, στην περίπτωση των ανεξάρτητων δειγμάτων η περιθώρεια κατανομή των δεδομένων ισούται με το γινόμενο των επί μέρους ολοκληρωμάτων. Στη περίπτωση όμως των συσχετισμένων δειγμάτων ο υπολογισμός του παράγοντα Bayes είναι μια πολύπλοκη διαδικασία και απαιτείται η χρήση υπολογιστικών αλγορίθμων *MCMC*. Για τα δείγματα από Binomial κατανομή προτείναμε την εκ των προτέρων  $Beta(\alpha, 1 - \alpha)$ , όπου  $\alpha \in (0, 1)$ . Σε αυτήν την κατανομή, η διασπορά αλλάζει μονοτονία καθώς η παράμετρος  $\alpha$  αυξάνεται με αποτέλεσμα να αλλάζει μονοτονία και ο παράγοντας Bayes. Για το παράδειγμα 3.4.1 αυτή η ιδιότητα δεν μας εξυπηρετεί πουθενά, οπότε η επιλογή της συγκεκριμένης εκ των προτέρων ήταν λανθασμένη.



## Κεφάλαιο 4

# Έλεγχος Υποθέσεων για Πολλαπλά Δείγματα

### 4.1 Έλεγχος υποθέσεων για $k$ ανεξάρτητα δείγματα από κατανομή Poisson

Θεωρούμε  $k$  τυχαία ανεξάρτητα δείγματα

$$y_{ij} = (y_{i1}, y_{i2}, \dots, y_{in_j})$$

μεγέθους  $n_j$ ,  $j = 1, \dots, k$  από κατανομή Poisson. Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0 : \lambda_1 = \dots = \lambda_k = \lambda \quad vs \quad H_1 : \lambda_i \neq \lambda_j, \quad \forall (i, j)$$

Τα αντίστοιχα μοντέλα που αντιπροσωπεύουν οι παραπάνω υποθέσεις μας είναι :

$$M_0 : y_{ij} \sim Poisson(\lambda) \quad M_1 : y_{ij} \sim Poisson(\lambda_j)$$

Για το μηδενικό μοντέλο  $M_0$  ορίζουμε  $\lambda \sim Gamma(p, q)$  εκ των προτέρων κατανομή για την κοινή παράμετρο  $\lambda$ . Η εκ των υστέρων που προκύπτει είναι  $f(\lambda|\mathbf{y}) \sim Gamma(p + N\bar{y}, q + N)$ , όπου  $N = \sum_{j=1}^k n_j$ ,  $\bar{y}$  είναι ο μέσος όλων των δεδομένων και  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ . Οπότε

$$f(\mathbf{y}|M_0) = \frac{\Gamma(p + N\bar{y})q^p}{\Gamma(p)(q + N)^{p+N\bar{y}} \prod_{i=1}^N y_i!}$$

Για το εναλλακτικό μοντέλο  $M_1$  ορίζουμε  $\lambda_j \sim \text{Gamma}(p_j, q_j)$  εκ των προτέρων κατανομές για τις παραμέτρους  $\lambda_j$ , όπου  $j = 1, \dots, k$ . Οπότε

$$\begin{aligned} f(\mathbf{y}|M_1) &= \int_0^\infty \dots \int_0^\infty f(\mathbf{y}|\lambda_1, \dots, \lambda_k, M_1) f(\lambda_1, \dots, \lambda_k) d\lambda_1 \dots d\lambda_k \\ &= \int_0^\infty f(\mathbf{y}_1|\lambda_1) f(\lambda_1) d\lambda_1 \times \dots \times \int_0^\infty f(\mathbf{y}_k|\lambda_k) f(\lambda_k) d\lambda_k \\ &= \prod_{j=1}^k \frac{\Gamma(p_j + n_j \bar{y}_j) q_j^{p_j}}{\Gamma(p_j) (q_j + n_j)^{p_j + n_j \bar{y}_j} \prod_{i=1}^{n_j} y_{ij}!} \end{aligned}$$

### Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$BF_{10} = \frac{\Gamma(p)(q + N)^{p + N\bar{y}}}{\Gamma(p + N\bar{y})q^p} \prod_{j=1}^k \frac{\Gamma(p_j + n_j \bar{y}_j) q_j^{p_j}}{\Gamma(p_j) (q_j + n_j)^{p_j + n_j \bar{y}_j}}$$

### Επιλογή της εκ των προτέρων κατανομής

Για να εξάγουμε συμπεράσματα θα βασιστούμε στην power εκ των προτέρων κατανομή με βάρος ίσο με ένα σημείο δεδομένων,  $\text{Gamma}(\delta n \bar{y}_0, \delta n)$ . Οπότε για  $\delta = \frac{1}{N}$  οι κατανομές που θα πάρουμε θα είναι για την μηδενική και την εναλλακτική υπόθεση αντίστοιχα οι :

$$H_0 : \lambda \sim \text{Gamma}(\bar{y}, 1) \quad \text{vs} \quad H_1 : \lambda_j \sim \text{Gamma}\left(\frac{n_j}{N} \bar{y}_j, \frac{n_j}{N}\right)$$

Η παραπάνω σχέση μπορεί πιο γενικά να πάρει την μορφή

$$H_0 : \lambda \sim \text{Gamma}(\alpha, 1) \quad \text{vs} \quad H_1 : \lambda_j \sim \text{Gamma}(w_j \alpha_j, w_j)$$

όπου

$$\begin{aligned} - \alpha &= w_1\alpha_1 + \dots + w_k\alpha_k & - w_j &= \frac{n_j}{N} \\ - N &= n_1 + \dots + n_k & - j &= 1, \dots, k \end{aligned}$$

Από την γενική μορφή της εκ των προτέρων κατανομής διακρίνουμε τις εξής περιπτώσεις :

◇ Αν  $\alpha_j = \alpha$  τότε  $\lambda \sim \text{Gamma}(\alpha, 1)$  και  $\lambda_j \sim \text{Gamma}\left(\frac{n_j}{N}\alpha, \frac{n_j}{N}\right)$ .

- Για  $\alpha = 1$  προκύπτουν οι εκ των προτέρων κατανομές

$$\lambda \sim \text{Gamma}(1, 1) \quad \& \quad \lambda_j \sim \text{Gamma}\left(\frac{n_j}{N}, \frac{n_j}{N}\right)$$

◇ Αν  $\alpha_j = \alpha$  και επιπλέον  $n_1 = \dots = n_k$  τότε  $\lambda \sim \text{Gamma}(\alpha, 1)$  και  $\lambda_j \sim \text{Gamma}\left(\frac{\alpha}{k}, \frac{1}{k}\right)$ .

- Για  $\alpha = 1$  προκύπτουν οι εκ των προτέρων κατανομές

$$\lambda \sim \text{Gamma}(1, 1) \quad \text{και} \quad \lambda_j \sim \text{Gamma}\left(\frac{1}{k}, \frac{1}{k}\right)$$

με  $E(\lambda) = \text{Var}(\lambda) = 1$  και  $E(\lambda_j) = 1 \quad \& \quad \text{Var}(\lambda_j) = k$ .

- Για  $\alpha = \bar{y}$  προκύπτουν οι εκ των προτέρων κατανομές

$$\lambda \sim \text{Gamma}(\bar{y}, 1), \quad \text{και} \quad \lambda_j \sim \text{Gamma}\left(\frac{\bar{y}}{k}, \frac{1}{k}\right)$$

με  $E(\lambda) = \text{Var}(\lambda) = \bar{y}$  και  $E(\lambda_j) = \bar{y} \quad \& \quad \text{Var}(\lambda_j) = k\bar{y}$ , που είναι Empirical Bayes Priors διότι βασίζονται στα παρατηρούμενα δεδομένα.

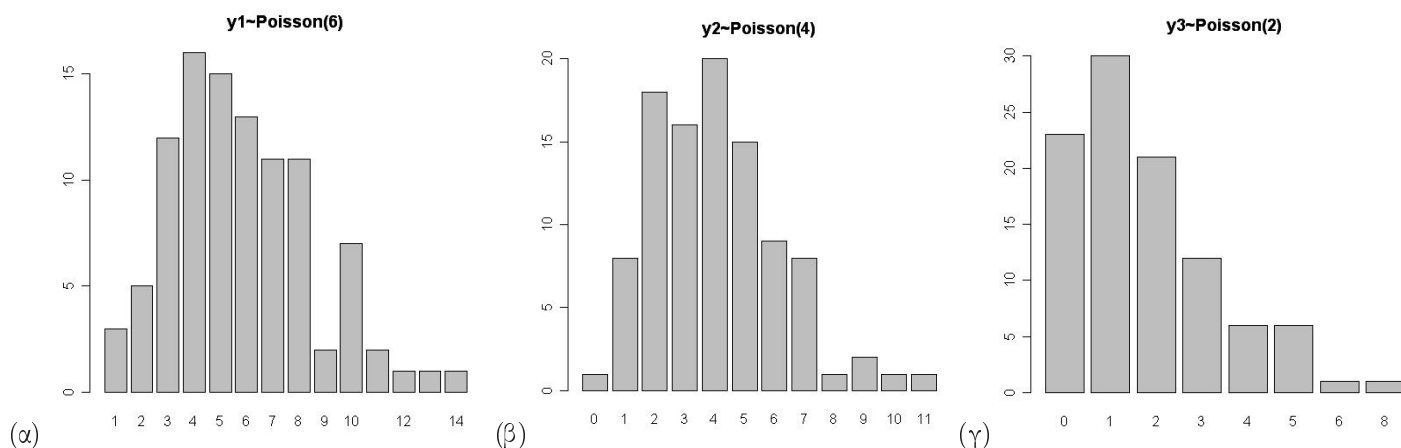
#### 4.1.1 Παράδειγμα (1) με προσομοιωμένα δεδομένα

Προσομειώνουμε τρία δείγματα των 100 παρατηρήσεων από κατανομή *Poisson* όπου  $\mathbf{y}_1 \sim \text{Poisson}(6)$ ,  $\mathbf{y}_2 \sim \text{Poisson}(4)$ ,  $\mathbf{y}_3 \sim \text{Poisson}(2)$  (σχήμα 4.1).

Ελέγχουμε την υπόθεση ότι

$$H_0 : \lambda_1 = \lambda_2 = \lambda_3 = \lambda \quad \text{vs} \quad H_1 : \lambda_i \neq \lambda_j, \quad \forall(i, j)$$





Σχήμα 4.1: Ραβδογράμματα των προσομοιωμένων παρατηρήσεων από κατανομή *Poisson*.

- Για τις παραμέτρους  $\lambda$ ,  $\lambda_1$ ,  $\lambda_2$  και  $\lambda_3$ , επιλέγουμε τις εκ των προτέρων κατανομές
  - ◇ *prior 1* :  $\lambda \sim \text{Gamma}(1, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim \text{Gamma}\left(\frac{1}{3}, \frac{1}{3}\right)$
  - ◇ *prior 2* :  $\lambda \sim \text{Gamma}(\bar{y}, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim \text{Gamma}\left(\frac{\bar{y}}{3}, \frac{1}{3}\right)$
- Με την βοήθεια της  $R$  προκύπτει ο πίνακας (4.1). Ο συνδυασμός *prior 1* μας δίνει μικρότερη τιμή για τον παράγοντα Bayes, ωστόσο και οι δύο συνδυασμοί μας δίνουν την ίδια συμπερασματολογία, δηλαδή ότι οι μέσοι των δειγμάτων διαφέρουν.

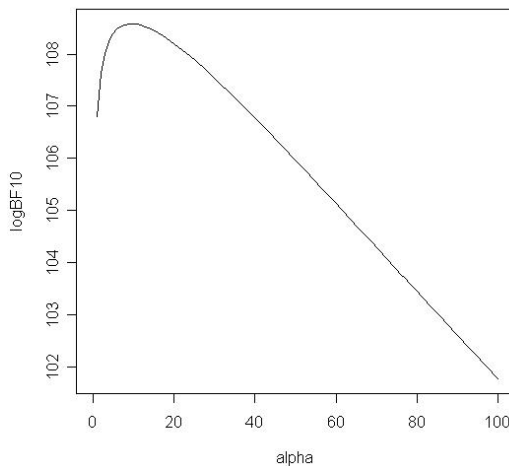
	<i>prior 1</i>	<i>prior 2</i>
$BF_{10}$	$2.39 \times 10^{46}$	$9.97 \times 10^{46}$
$\log(BF)_{10}$	$1.068 \times 10^2$	$1.08 \times 10^2$
Ένδειξη εναντίον της	$H_0$ Πολύ ισχυρή	$H_0$ Πολύ ισχυρή

Πίνακας 4.1: Αποτελέσματα του παράγοντα Bayes για τους συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.

- Στο σχήμα (4.2) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\lambda$ ,  $\lambda_1$ ,  $\lambda_2$  και  $\lambda_3$  αντίστοιχα.

◇ Γενική μορφή :  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim \text{Gamma}\left(\frac{\alpha}{3}, \frac{1}{3}\right)$

Υπό την γενική μορφή των δύο πρώτων συνδυασμών, η μηδενική υπόθεση απορρίπτεται για τιμές  $\alpha \in (0, 1.2 \times 10^5)$ .

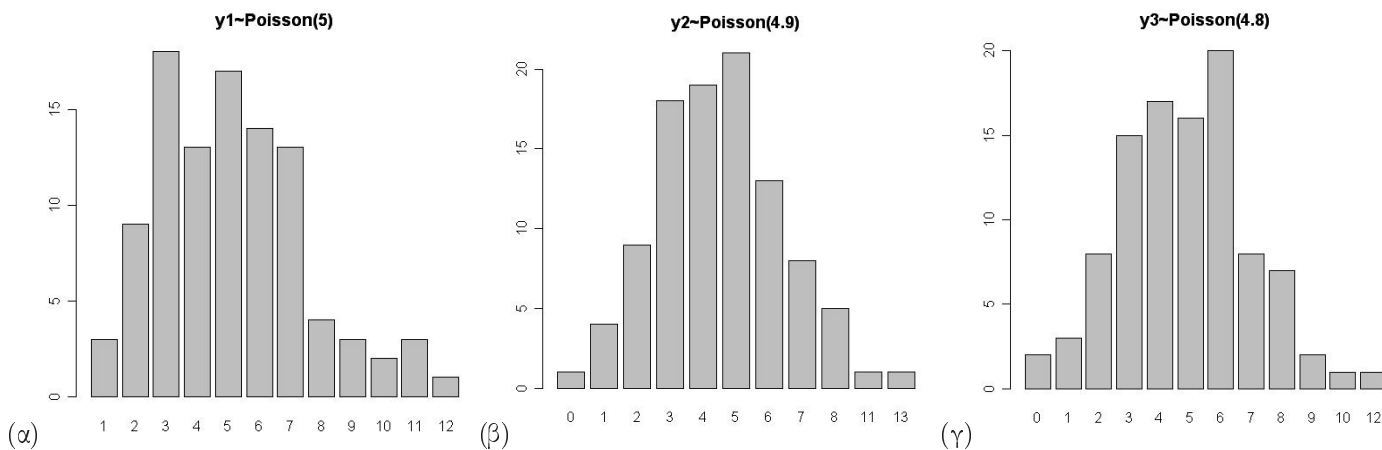


Σχήμα 4.2: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\alpha$  για την γενική μορφή :  $\lambda \sim Gamma(\alpha, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim Gamma\left(\frac{\alpha}{3}, \frac{1}{3}\right)$  (Παράδειγμα 4.1.1).

4.1.2 Παράδειγμα (2) με προσομοιωμένα δεδομένα

Προσομοιώνουμε τρία δείγματα των 100 παρατηρήσεων από κατανομή *Poisson* όπου  $\mathbf{y}_1 \sim Poisson(5)$ ,  $\mathbf{y}_2 \sim Poisson(4.9)$ ,  $\mathbf{y}_3 \sim Poisson(4.8)$  (σχήμα 4.3). Ελέγχουμε την υπόθεση ότι

$$H_0 : \lambda_1 = \lambda_2 = \lambda_3 = \lambda \quad vs \quad H_1 : \lambda_i \neq \lambda_j, \quad \forall(i, j)$$



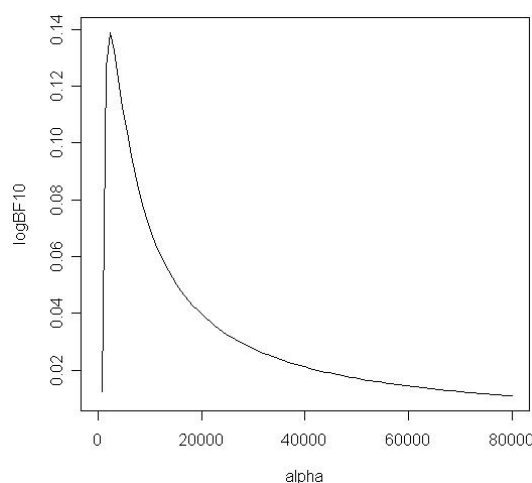
Σχήμα 4.3: Ραβδογράμματα των προσομοιωμένων παρατηρήσεων από κατανομή *Poisson*.

- Για τις παραμέτρους  $\lambda$ ,  $\lambda_1$ ,  $\lambda_2$  και  $\lambda_3$ , επιλέγουμε τις εκ των προτέρων κατανομές
  - ◇ *prior 1* :  $\lambda \sim Gamma(1, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim Gamma\left(\frac{1}{3}, \frac{1}{3}\right)$
  - ◇ *prior 2* :  $\lambda \sim Gamma(\bar{y}, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim Gamma\left(\frac{\bar{y}}{3}, \frac{1}{3}\right)$

• Με την βοήθεια της  $R$  προκύπτει ο πίνακας (4.2). Ομοίως ο συνδυασμός *prior* 1 μας δίνει μικρότερη τιμή για τον παράγοντα Bayes, όμως ενώ οι δύο συνδυασμοί μας δίνουν την ίδια συμπερασματολογία, δηλαδή ότι οι μέσοι των δειγμάτων δεν διαφέρουν, η βαρύτητα της πληροφορίας διαφέρει. Ο συνδυασμός *prior* 1 μας δίνει πιο ισχυρή ένδειξη υπέρ της ισότητας των μέσων των δειγμάτων.

	<i>prior</i> 1	<i>prior</i> 2
$BF_{10}$	0.00198	0.0147
$\log(BF)_{10}$	-6.223	-4.222
Ένδειξη	$H_1$	$H_1$
εναντίον της	Πολύ ισχυρή	Θετική

Πίνακας 4.2: Αποτελέσματα του παράγοντα Bayes για τους τρεις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.



Σχήμα 4.4: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\alpha$  για την γενική μορφή :  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim \text{Gamma}(\frac{\alpha}{3}, \frac{1}{3})$  (Παράδειγμα 4.1.2).

• Στο σχήμα (4.4) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές της παραμέτρου  $\alpha$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\lambda$ ,  $\lambda_1$ ,  $\lambda_2$  και  $\lambda_3$  αντίστοιχα.

◊ Γενική μορφή :  $\lambda \sim \text{Gamma}(\alpha, 1)$ ,  $\lambda_1, \lambda_2, \lambda_3 \sim \text{Gamma}(\frac{\alpha}{3}, \frac{1}{3})$

Υπό την γενική μορφή των δύο πρώτων συνδυασμών, για όλες τις τιμές της παραμέτρου  $\alpha$  δεχόμαστε την μηδενική υπόθεση.

## 4.2 Έλεγχος υποθέσεων για $k$ ανεξάρτητα δείγματα από κατανομή Binomial

Θεωρούμε  $k$  τυχαία ανεξάρτητα δείγματα μεγέθους  $n_j$ , όπου  $j = 1, \dots, k$ ,

$$\mathbf{y}_{ij} = (y_{i1}, y_{i2}, \dots, y_{in_k k})$$

από κατανομή Binomial ( $n = \sum_{j=1}^k n_j$ ). Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0 : \pi_1 = \dots = \pi_k = \pi \quad vs \quad H_1 : \pi_i \neq \pi_j, \quad \forall (i, j)$$

Τα αντίστοιχα μοντέλα που αντιπροσωπεύουν οι παραπάνω υποθέσεις μας είναι :

$$M_0 : \mathbf{y}_j \sim Binomial(N_{ij}, \pi) \quad vs \quad M_1 : \mathbf{y}_j \sim Binomial(N_{ij}, \pi_j)$$

Για το μηδενικό μοντέλο  $M_0$  ορίζουμε  $\pi \sim Beta(\alpha_o, \beta_o)$  εκ των προτέρων κατανομή για την κοινή παράμετρο  $\pi$ . Οπότε η εκ των υστέρων κατανομή που προκύπτει είναι  $f(\pi|\mathbf{y}) \sim Beta(\alpha_o + \sum_{i=1}^n y_i, \beta_o + N - \sum_{i=1}^n y_i)$ , όπου  $N = \sum_{j=1}^k N_{.j}$  με  $N_{.j} = \sum_{i=1}^{n_j} N_{ij}$  και  $\mathbf{y} = \{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ . Οπότε

$$f(\mathbf{y}|M_0) = \frac{B(\alpha_o + N \bar{y}, \beta_o + N - N \bar{y})}{B(\alpha_o, \beta_o)} \prod_{i=1}^n \prod_{j=1}^k \binom{N_{ij}}{y_{ij}}$$

Για το εναλλακτικό μοντέλο  $M_1$  ορίζουμε τις εκ των προτέρων κατανομές των παραμέτρων  $\pi_j \sim Beta(\alpha_j, \beta_j)$ , όπου  $j = 1, \dots, k$ . Οπότε

$$\begin{aligned} f(\mathbf{y}|M_1) &= \int_0^1 \dots \int_0^1 f(\mathbf{y}|\pi_1, \dots, \pi_k, M_1) f(\pi_1, \dots, \pi_k) d\pi_1 \dots d\pi_k \\ &= \prod_{j=1}^k \left[ \frac{B(\alpha_j + N_{.j} \bar{y}_j, \beta_j + N_{.j} - N_{.j} \bar{y}_j)}{B(\alpha_j, \beta_j)} \prod_{i=1}^{n_j} \binom{N_{ij}}{y_{ij}} \right] \end{aligned}$$

## Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$BF_{10} = \frac{B(\alpha_o, \beta_o)}{B(\alpha_o + N \bar{y}, \beta_o + N - N \bar{y})} \prod_{j=1}^k \frac{B(\alpha_j + N_j \bar{y}_j, \beta_j + N_j - N_j \bar{y}_j)}{B(\alpha_j, \beta_j)}$$

## Επιλογή της εκ των προτέρων κατανομής

Για να εξάγουμε συμπεράσματα θα βασιστούμε στην power εκ των προτέρων κατανομή,  $Beta(N_o \delta \bar{y}_o, N_o \delta (1 - \bar{y}_o))$ . Οπότε για  $\delta = \frac{1}{N_o}$  οι κατανομές που θα πάρουμε θα είναι για την μηδενική και την εναλλακτική υπόθεση αντίστοιχα οι :

$$H_0 : \pi \sim Beta(\bar{y}, (1 - \bar{y})) \quad vs \quad H_1 : \pi_j \sim Beta\left(\frac{N_j}{N} \bar{y}_j, \frac{N_j}{N} (1 - \bar{y}_j)\right)$$

Η παραπάνω σχέση μπορεί πιο γενικά να πάρει την μορφή

$$H_0 : \pi \sim Beta(\alpha, 1 - \alpha) \quad vs \quad H_1 : \pi_j \sim Beta(w_j \alpha_j, w_j (1 - \alpha_j))$$

όπου

$$\begin{aligned} - \alpha &= w_1 \alpha_1 + \dots + w_k \alpha_k & - w_j &= \frac{N_j}{N} \\ - N &= N_1 + \dots + N_k & - j &= 1, \dots, k \end{aligned}$$

Από την γενική μορφή της εκ των προτέρων κατανομής διακρίνουμε τις εξής περιπτώσεις :

◇ Αν  $\alpha_1 = \dots = \alpha_k = \alpha$  :  $\alpha \in (0, 1)$  τότε θα έχουμε τις εκ των προτέρων κατανομές  $\pi \sim Beta(\alpha, 1 - \alpha)$  και  $\pi_j \sim Beta\left(\frac{N_j}{N} \alpha, \frac{N_j}{N} (1 - \alpha)\right)$ .

- Για  $\alpha = \frac{1}{2}$  προκύπτουν οι εκ των προτέρων κατανομές

$$\pi \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right) \quad \& \quad \pi_j \sim Beta\left(\frac{N_j}{2N}, \frac{N_j}{2N}\right)$$

όπου θα έχουμε μια Jeffreys priors για την μηδενική υπόθεση. Θέτουμε  $\alpha = \frac{1}{2}$  διότι είναι το κέντρο της κατανομής και θεωρούμε ότι το κέντρο είναι η πιο ουδέτερη τιμή.

- Για  $\alpha = \bar{y}$  προκύπτουν οι εκ των προτέρων κατανομές

$$\pi \sim \text{Beta}(\bar{y}, 1 - \bar{y}) \quad \& \quad \pi_j \sim \text{Beta}\left(\frac{N_j \bar{y}}{N}, \frac{N_j(1 - \bar{y})}{N}\right)$$

με  $E(\lambda) = \bar{y}$ ,  $\text{Var}(\lambda) = \frac{\bar{y}(1-\bar{y})}{2}$  και  $E(\lambda_j) = \bar{y}$  &  $\text{Var}(\lambda_j) = \frac{\bar{y}(1-\bar{y})}{\frac{N_j}{N}+1}$ , που είναι Empirical Bayes Priors.

◇ Αν  $\alpha_1 = \dots = \alpha_k = \alpha$  και επιπλέον  $N_1 = \dots = N_k$  τότε θα πάρουμε τις εκ των προτέρων κατανομές  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$  και  $\pi_j \sim \text{Beta}\left(\frac{\alpha}{k}, \frac{1}{k}(1 - \alpha)\right)$ .

◇ Ενδιαφέρον έχει επίσης να μελετήσουμε την  $\text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$  εκ των προτέρων κατανομή  $\left[E(\pi) = \frac{1}{2} \quad \& \quad \text{Var}(\pi) = \frac{1}{4(c+1)}\right]$ , που είναι υποπερίπτωση της power εκ των προτέρων για  $\bar{y}_o = \frac{1}{2}$ , και ιδιαίτερα τις υποπεριπτώσεις :

- Για  $c = 1$ , οι εκ των προτέρων κατανομές που θα προκύψουν για τα  $\pi$  και  $\pi_j$ , θα είναι όλες  $\text{Beta}(1, 1)$  (Ομοιόμορφη κατανομή).

- Για  $c = \frac{1}{2}$ , οι εκ των προτέρων κατανομές που θα προκύψουν για τα  $\pi$  και  $\pi_j$ , θα είναι όλες  $\text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$  (Jeffreys κατανομή).

#### 4.2.1 Παράδειγμα (1) με προσομοιωμένα δεδομένα

Προσομειώνουμε τρία δείγματα των 100 παρατηρήσεων από *Binomial* κατανομή. Στο σχήμα (4.5) βλέπουμε τα ραβδογράμματα των παρατηρήσεων.

$$\mathbf{y}_1 \sim \text{Binomial}(10, 0.5), \quad \mathbf{y}_2 \sim \text{Binomial}(10, 0.49), \quad \mathbf{y}_3 \sim \text{Binomial}(10, 0.48)$$

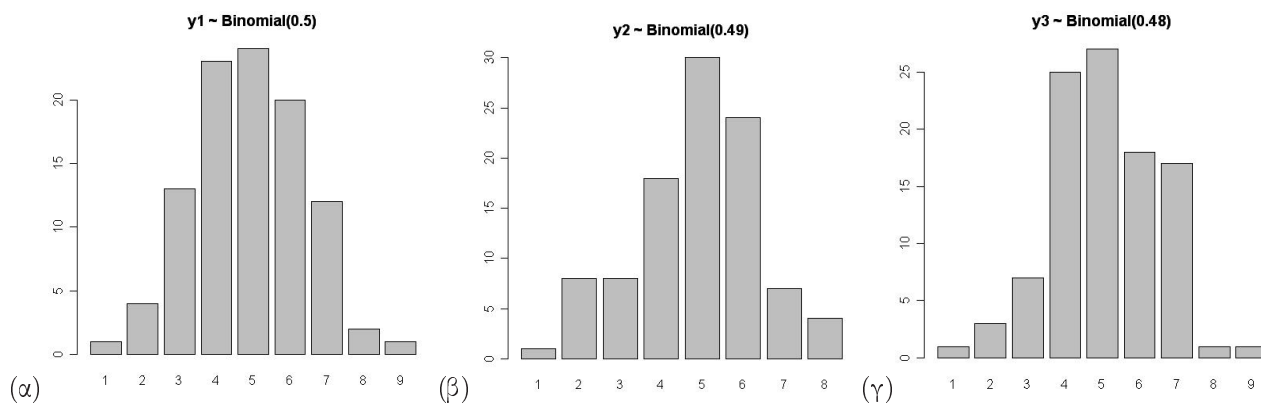
• Για τις παραμέτρους  $\pi$ ,  $\pi_1$ ,  $\pi_2$  και  $\pi_3$ , επιλέγουμε τις εκ των προτέρων κατανομές

◇ *prior 1* :  $\pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{2N}, \frac{N_j}{2N}\right)$

◇ *prior 2* :  $\pi \sim \text{Beta}(\bar{y}, 1 - \bar{y})$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j \bar{y}}{N}, \frac{N_j(1 - \bar{y})}{N}\right)$

◇ *prior 3* :  $\pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$

◇ *prior 4* :  $\pi \sim \text{Beta}(1, 1)$ ,  $\pi_j \sim \text{Beta}(1, 1)$ , όπου  $j = 1, 2, 3$



Σχήμα 4.5: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από κατανομή *Binomial*.

- Με την βοήθεια της  $R$  προκύπτει ο πίνακας (4.3). Διαπιστώνουμε ότι οι συνδυασμοί *prior 1* και *prior 2* που είναι υπό περιστώσεις της ίδιας γενικής μορφής, δίνουν σχεδόν τις ίδιες τιμές για τον παράγοντα Bayes. Το ίδιο συμβαίνει και για τους συνδυασμούς *prior 3* και *prior 4*, όπως είδαμε και στο κεφάλαιο 3. Ωστόσο και οι τέσσερις συνδυασμοί δίνουν την ίδια συμπερασματολογία.

	<i>prior 1</i>	<i>prior 2</i>	<i>prior 3</i>	<i>prior 4</i>
$BF_{10}$	0.00014	0.00014	0.0017	0.0043
$\log(BF)_{10}$	-8.8756	-8.8758	-6.35	-5.45
Ένδειξη	$H_1$	$H_1$	$H_1$	$H_1$
εναντίον της	Πολύ ισχυρή	Πολύ ισχυρή	Πολύ ισχυρή	Πολύ ισχυρή

Πίνακας 4.3: Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.

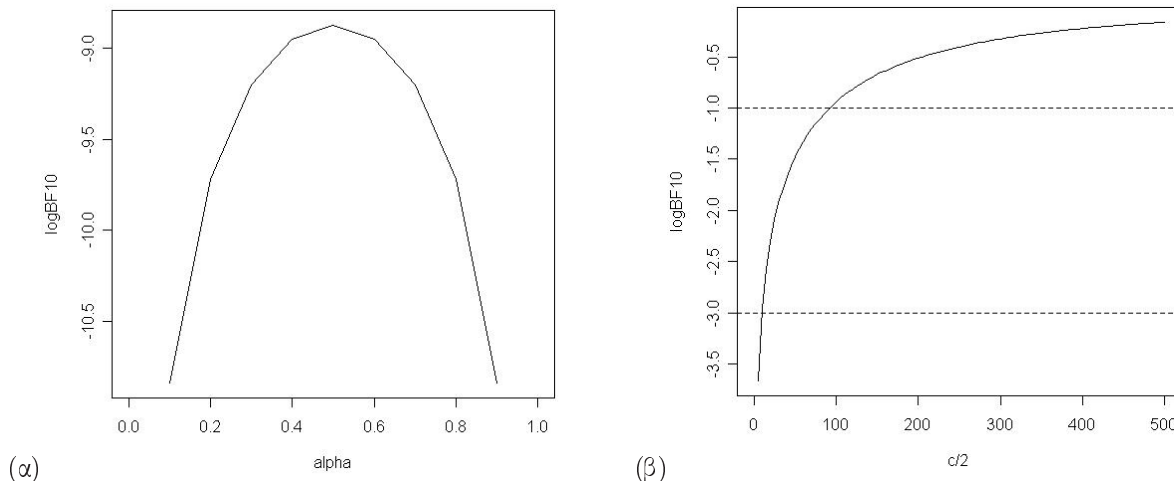
- Στο σχήμα (4.6) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων  $\alpha$  &  $c$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\pi$ ,  $\pi_1$ ,  $\pi_2$  και  $\pi_3$  αντίστοιχα.

◇ Γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$

Υπό γενική μορφή των δύο πρώτων συνδυασμών, για όλες τις τιμές της παραμέτρου  $\alpha \in (0, 1)$  δεχόμαστε την μηδενική υπόθεση.

◇ Γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$

Υπό την γενική μορφή του τρίτου και τέταρτου συνδυασμού, η εναλλακτική υπόθεση απορρίπτεται για τιμές της παραμέτρου  $c \in [0, 10^3]$ .

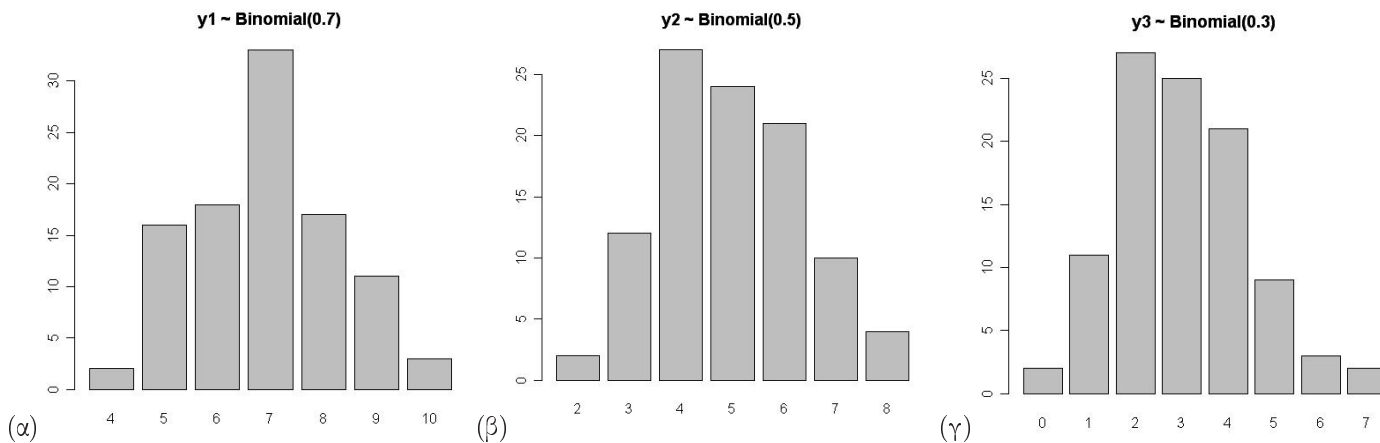


Σχήμα 4.6: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\alpha$  για (α) την γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$  και (β) την γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$  (Παράδειγμα 4.2.1).

### 4.2.2 Παράδειγμα (2) με προσομοιωμένα δεδομένα

Προσομοιώνουμε τρία δείγματα των 100 παρατηρήσεων από *Binomial* κατανομή. Στο σχήμα (4.7) βλέπουμε τα ραβδογράμματα των παρατηρήσεων.

$$\mathbf{y}_1 \sim \text{Binomial}(10, 0.7), \quad \mathbf{y}_2 \sim \text{Binomial}(10, 0.5), \quad \mathbf{y}_3 \sim \text{Binomial}(10, 0.3)$$



Σχήμα 4.7: Ραβδόγραμμα προσομοιωμένων παρατηρήσεων από κατανομή *Binomial*.

- Για τις παραμέτρους  $\pi$ ,  $\pi_1$ ,  $\pi_2$  και  $\pi_3$ , επιλέγουμε τις εκ των προτέρων κατανομές

◇ *prior 1* :  $\pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{2N}, \frac{N_j}{2N}\right)$

◇ *prior 2* :  $\pi \sim \text{Beta}(\bar{y}, 1 - \bar{y})$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\bar{y}, \frac{N_j}{N}(1 - \bar{y})\right)$



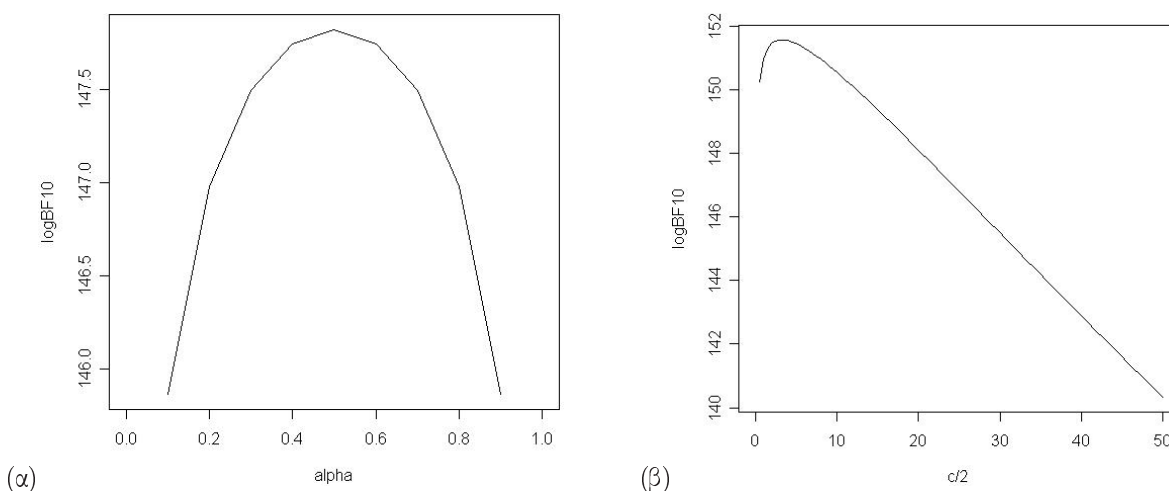
◇ *prior 3* :  $\pi \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$

◇ *prior 4* :  $\pi \sim \text{Beta}(1, 1)$ ,  $\pi_j \sim \text{Beta}(1, 1)$ , όπου  $j = 1, 2, 3$

• Με την βοήθεια της  $R$  προκύπτει ο πίνακας (4.4). Ομοίως διαπιστώνουμε ότι οι συνδυασμοί *prior 1* και *prior 2* που είναι υπό περιστώσεις της ίδιας γενικής μορφής, δίνουν σχεδόν τις ίδιες τιμές για τον παράγοντα Bayes. Το ίδιο συμβαίνει και για τους συνδυασμούς *prior 3* και *prior 4*. Ωστόσο και οι τέσσερις συνδυασμοί δίνουν την ίδια συμπερασματολογία.

	<i>prior 1</i>	<i>prior 2</i>	<i>prior 3</i>	<i>prior 4</i>
$BF_{10}$	$1.5798 \times 10^{64}$	$1.5797 \times 10^{64}$	$1.76 \times 10^{65}$	$3.68 \times 10^{65}$
$\log(BF)_{10}$	$1.478 \times 10^2$	$1.478 \times 10^2$	$1.5 \times 10^2$	$1.51 \times 10^2$
Ένδειξη	$H_0$	$H_0$	$H_0$	$H_0$
εναντίον της	Πολύ ισχυρή	Πολύ ισχυρή	Πολύ ισχυρή	Πολύ ισχυρή

Πίνακας 4.4: Αποτελέσματα του παράγοντα Bayes για τους τέσσερις συνδυασμούς εκ των προτέρων κατανομών των παραμέτρων.



Σχήμα 4.8: Γραφική απεικόνιση της ευαισθησίας του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές του  $\alpha$  για (α) την γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$  και (β) την γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$  (Παράδειγμα 4.2.2).

• Στο σχήμα (4.8) βλέπουμε γραφικά την ευαισθησία του λογαρίθμου του παράγοντα Bayes για τις διάφορες τιμές των παραμέτρων  $\alpha$  &  $c$  στην γενική μορφή των συνδυασμών εκ των προτέρων κατανομών για τις παραμέτρους  $\pi$ ,  $\pi_1$ ,  $\pi_2$  και  $\pi_3$  αντίστοιχα.

- ◇ Γενική μορφή 1:  $\pi \sim \text{Beta}(\alpha, 1 - \alpha)$ ,  $\pi_j \sim \text{Beta}\left(\frac{N_j}{N}\alpha, \frac{N_j}{N}(1 - \alpha)\right)$   
Υπό γενική μορφή των δύο πρώτων συνδυασμών, για όλες τις τιμές της παραμέτρου  $\alpha \in (0, 1)$  απορρίπτουμε την μηδενική υπόθεση.
- ◇ Γενική μορφή 2:  $\pi \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$ ,  $\pi_j \sim \text{Beta}\left(\frac{c}{2}, \frac{c}{2}\right)$   
Υπό την γενική μορφή του τρίτου και τέταρτου συνδυασμού, η μηδενική υπόθεση απορρίπτεται για τιμές της παραμέτρου  $c < 9 \times 10^4$ .

### 4.2.3 Παράδειγμα με πραγματικά δεδομένα

Θα συνεχίσουμε το παράδειγμα της παραγράφου 2.5.2. Έχουμε πέντε αντιβιοτικά (Amikacin[1], Gentamicin[2], Netilmicin[3], Sisomicin[4], Tobramycin[5]) και τα συγκρίνουμε όλα μαζί (5 ομάδες) επιλέγοντας την εκ των προτέρων κατανομή  $\text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$  για όλες τις παραμέτρους.

Με την βοήθεια της  $R$ , λαμβάνουμε  $BF = 1.94 \times 10^6$ , που μας δίνει πολύ ισχυρή ένδειξη εναντίον της μηδενικής υπόθεσης, δηλαδή τα ποσοστά εμφάνισης νεφρικής τοξικότητας των αντιβιοτικών, διαφέρουν μεταξύ τους.

## 4.3 Έλεγχος υποθέσεων για πίνακες συνάφειας

### 4.3.1 Εισαγωγή στους πίνακες συνάφειας

Έστω  $A$  και  $B$  δύο παράγοντες με  $r$  και  $c$  επίπεδα αντίστοιχα. Όταν ταξινομούμε τις παρατηρήσεις μας στους δύο παράγοντες υπάρχουν  $ij$  (όπου  $i = 1, \dots, r$  και  $j = 1, \dots, c$ ) δυνατοί συνδυασμοί ταξινόμησης. Αυτοί αποτυπώνονται σ' έναν πίνακα με  $r$  γραμμές και  $c$  στήλες, που καλείται πίνακας συνάφειας (contingency table), όρος που ειηγήθηκε από τον Karl Pearson (1904) (Agresti, 1990, σελ. 9)[3]. Θέτουμε  $\pi_{ij}$  την πιθανότητα κάποια παρατήρηση να βρεθεί στο  $ij$  κελί. Επίσης  $\sum_{i=1}^r \sum_{j=1}^c \pi_{ij} = 1$ .

Οι δύο παράγοντες είναι ανεξάρτητοι όταν ισχύει  $\pi_{ij} = \pi_{i.} * \pi_{.j}$ , αλλιώς είναι συσχετισμένοι. Παραδοσιακά υπάρχουν δύο τρόποι να ελέγξουμε την ανεξαρτη-

	$B_1$	$B_2$	$B_3$	$\dots$	$B_c$	
$A_1$	$\pi_{11}$	$\pi_{12}$	$\pi_{13}$	$\dots$	$\pi_{1c}$	$\pi_{1.}$
$A_2$	$\pi_{21}$	$\pi_{22}$	$\pi_{23}$	$\dots$	$\pi_{2c}$	$\pi_{2.}$
$A_3$	$\pi_{31}$	$\pi_{32}$	$\pi_{33}$	$\dots$	$\pi_{3c}$	$\pi_{3.}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$A_r$	$\pi_{r1}$	$\pi_{r2}$	$\pi_{r3}$	$\dots$	$\pi_{rc}$	$\pi_{r.}$
	$\pi_{.1}$	$\pi_{.2}$	$\pi_{.3}$	$\dots$	$\pi_{.c}$	$\pi_{..}$

Πίνακας 4.5: Πίνακας πιθανοτήτων των πληθυσμών A &amp; B.

σία σ' έναν  $r \times c$  πίνακα συνάφειας. Πρώτη προσέγγιση είναι η χρήση του ελέγχου  $X^2 - Pearson$  που λειτουργεί για απλή τυχαία δειγματοληψία (πολυωνυμική δειγματοληψία). Η δεύτερη προσέγγιση είναι η χρήση του παράγοντα του Bayes. Από τις δυο αυτές προσεγγίσεις, η πρώτη παρουσιάζει διάφορα μειονεκτήματα. Σε περιπτώσεις όπου οι υποθέσεις πολυωνυμικού δείγματος δεν ισχύουν (δηλαδή όταν υπάρχει συσχέτιση μεταξύ των μελών της ίδιας ομάδας, όπως παιδιά από την ίδια οικογένεια ή ποντίκια από το ίδιο εργαστήριο) τότε ο κλασικός  $X^2$  έλεγχος είναι ακατάλληλος. Η παραπάνω συσχέτιση καλείται intra-class correlation (Nandram & Choi, 2007)[44]. Επίσης είναι ακατάλληλος όταν οι μετρήσεις των κελιών είναι πολύ μικρές ή μηδέν.

#### 4.3.2 Έλεγχος υποθέσεων για $r \times c$ πίνακες συνάφειας

Θέλουμε να ελέγξουμε την υπόθεση ότι

$$H_0 : \nexists \text{ συσχέτιση μεταξύ των } A \text{ \& } B$$

vs

$$H_1 : \exists \text{ συσχέτιση μεταξύ των } A \text{ \& } B$$

Οπότε για έναν  $r \times c$  πίνακα συνάφειας, υποθέτουμε δύο πολυωνυμικά - Dirichlet μοντέλα, ένα με συσχέτιση ( $M_1$ ) και ένα χωρίς ( $M_0$ ) (Nandram & Choi, 2007)[20]. Για το μοντέλο με συσχέτιση ( $M_1$ ), υποθέτουμε ότι τα δεδομένα του πίνακα ακολουθούν μια πολυωνυμική κατανομή με παράμετρο  $\pi_{ij}$ . Έτσι έχουμε:

$$\mathbf{n}|\boldsymbol{\pi} \sim \text{Multinomial}(n, \boldsymbol{\pi})$$

$$\boldsymbol{\pi} \sim \text{Dirichlet}(k, \dots, k)$$

Οπότε λαμβάνουμε

$$\begin{aligned} f(y|M_1) &= \frac{\prod_{i=1}^r \prod_{j=1}^c \Gamma(n_{ij} + k) \Gamma(\sum_{i=1}^r \sum_{j=1}^c k)}{\Gamma(\sum_{i=1}^r \sum_{j=1}^c (n_{ij} + k)) \prod_{i=1}^r \prod_{j=1}^c \Gamma(k)} \frac{n!}{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!} \\ &= \frac{D_{rc}(\mathbf{n} + k)}{D_{rc}(k)} \frac{n!}{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!} \end{aligned}$$

Από τα παραπάνω αν επιλέξουμε την υπό περίπτωση της Dirichlet όπου  $k = 1$  (δηλ. μια Ομοιόμορφη κατανομή), τότε θα προκύψει η περιθώρεια συνάρτηση

$$f(y|M_1) = \frac{(rc - 1)!n!}{(n + rc - 1)!}$$

Για το μοντέλο χωρίς συσχέτιση ( $M_0$ ), υποθέτουμε ότι τα δεδομένα του πίνακα ακολουθούν μια πολυωνυμική κατανομή με παράμετρο  $\pi_{ij}$ , όπου  $\pi_{ij} = \pi_{i.}^{(1)} \times \pi_{.j}^{(2)}$ , ( $i = 1, \dots, r$  και  $j = 1, \dots, c$ ). Έτσι έχουμε :

$$\mathbf{n}|\boldsymbol{\pi}^{(1)}, \boldsymbol{\pi}^{(2)} \sim Multinomial(n, \boldsymbol{\pi}), \quad \text{με } \boldsymbol{\pi} = \boldsymbol{\pi}^{(1)} \times \boldsymbol{\pi}^{(2)\tau}$$

$$\boldsymbol{\pi}^{(1)} \sim Dirichlet(k^{(1)}, \dots, k^{(1)})$$

$$\boldsymbol{\pi}^{(2)} \sim Dirichlet(k^{(2)}, \dots, k^{(2)})$$

όπου τα  $\boldsymbol{\pi}^{(1)}$  και  $\boldsymbol{\pi}^{(2)}$  είναι ανεξάρτητα. Οπότε λαμβάνουμε

$$\begin{aligned} f(y|M_0) &= \frac{\prod_{i=1}^r \Gamma(n_{i.}^{(1)} + k^{(1)}) \Gamma(\sum_{i=1}^r k^{(1)})}{\Gamma(\sum_{i=1}^r (n_{i.}^{(1)} + k^{(1)})) \prod_{i=1}^r \Gamma(k^{(1)})} \\ &\times \frac{\prod_{j=1}^c \Gamma(n_{.j}^{(2)} + k^{(2)}) \Gamma(\sum_{j=1}^c k^{(2)})}{\Gamma(\sum_{j=1}^c (n_{.j}^{(2)} + k^{(2)})) \prod_{j=1}^c \Gamma(k^{(2)})} \frac{n!}{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!} \\ &= \frac{D_r(\mathbf{n}^{(1)} + k^{(1)}) D_c(\mathbf{n}^{(2)} + k^{(2)})}{D_r(k^{(1)}) D_c(k^{(2)})} \frac{n!}{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!} \end{aligned}$$

Από τα παραπάνω αν πάλι επιλέξουμε την υπό περίπτωση της Dirichlet όπου  $k^{(1)} = k^{(2)} = 1$ , τότε θα προκύψει η περιθώρεια συνάρτηση

$$f(y|M_0) = n! \frac{(r-1)!(c-1)!}{(n+r-1)!(n+c-1)!} \frac{\prod_{i=1}^r n_{i.}! \prod_{j=1}^c n_{.j}!}{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!}$$

### Παράγοντας Bayes

Υπολογίζεται ο Bayes Factor για την σύγκριση του μοντέλου  $M_1$  έναντι του  $M_0$ .

$$BF_{10} = \frac{D_{rc}(\mathbf{n} + k)/D_r(\mathbf{n}^{(1)} + k^{(1)})D_c(\mathbf{n}^{(2)} + k^{(2)})}{D_{rc}(k)/D_r(k^{(1)})D_c(k^{(2)})}$$

Όπου  $n_{.}^{(1)} = \sum_{j=1}^c n_{ij}$ ,  $n_{.}^{(2)} = \sum_{i=1}^r n_{ij}$  και  $D_d(\boldsymbol{\delta}) = \frac{\prod_{i=1}^d \Gamma(\delta_i)}{\Gamma(\sum \delta_i)}$

### Επιλογή της εκ των προτέρων κατανομής

Φυσικά στην υπό περίπτωση της Dirichlet όπου  $k = k^{(1)} = k^{(2)} = 1$  (Ομοιόμορφη), ο παράγοντας Bayes μειώνεται στην σχέση

$$BF_{10} = \frac{(rc-1)!}{(r-1)!(c-1)!} \frac{(n+r-1)!(n+c-1)!}{(n+rc-1)!} \frac{\prod_{i=1}^r \prod_{j=1}^c n_{ij}!}{\prod_{i=1}^r n_{i.}! \prod_{j=1}^c n_{.j}!} \quad (4.1)$$

#### 4.3.3 Παράδειγμα με πραγματικά δεδομένα

Στο παράδειγμα αυτό θα χρησιμοποιηθούν δεδομένα από την μελέτη του Worcester της Μασσαχουσέτης για το έμφραγμα του μυοκαρδίου την περίοδο 1975 με 1991 (Goldberg et al. 1988)[25],[27]. Για να εξετάσουμε εάν η ηλικία σχετίζεται με την θεραπευτική αγωγή, θα χρησιμοποιήσουμε τον τύπο του παράγοντα Bayes (4.1) όπου έχουμε επιλέξει την υπό περίπτωση της Ομοιόμορφης κατανομής (ώστε να εκφράσουμε την εκ των προτέρων άγνοια μας για τα δεδομένα).

Για τον πίνακα (4.6) έχουμε ότι η λήψη λιδοκαΐνης δεν σχετίζεται με την ηλικία αφού  $BF = 0.036 < 1$  με ισχυρή ένδειξη εναντίον της εναλλακτικής.

Ηλικία (Χρόνια)	$E^+$ λιδοκαΐνη	$E^-$ λιδοκαΐνη	Σύνολο
< 55	4	9	13
55 – 59	3	5	8
60 – 64	8	9	17
65 – 69	10	9	19
70 – 74	11	23	34
$\geq 75$	24	57	81
Σύνολο	60	112	172

Πίνακας 4.6: Πλήθος ασθενών ανά ηλικία που έλαβαν ( $E^+$ ) ή δεν έλαβαν ( $E^-$ ) θεραπεία λιδοκαΐνης για οξύ έμφραγμα του μυοκαρδίου.

Για τον πίνακα (4.7) έχουμε ότι η λήψη beta blocker επίσης δεν σχετίζεται με την ηλικία αφού  $BF = 0.058 < 1$  με θετική ένδειξη εναντίον της εναλλακτικής.

Ηλικία (Χρόνια)	$E^+$ beta blocker	$E^-$ beta blocker	Σύνολο
< 55	3	3	6
55 – 59	2	7	9
60 – 64	3	9	12
65 – 69	6	10	16
70 – 74	13	18	31
$\geq 75$	23	87	110
Σύνολο	50	134	184

Πίνακας 4.7: Πλήθος ασθενών ανά ηλικία που έλαβαν ( $E^+$ ) ή δεν έλαβαν ( $E^-$ ) θεραπεία beta blocker για οξύ έμφραγμα του μυοκαρδίου.



## Κεφάλαιο 5

### Συζήτηση

Στην παρούσα διπλωματική εργασία αναπτύχθηκε η Μπευζιανή προσέγγιση σε βασικούς ελέγχους υποθέσεων. Στην Μπευζιανή στατιστική οι έλεγχοι υποθέσεων πραγματοποιούνται μέσω ενός σχετικού λόγου πιθανοτήτων της εκ των υστέρων και της εκ των προτέρων πληροφορίας που καλείται παράγοντας Bayes. Οι παράγοντες Bayes που υπολογίστηκαν, συνέκριναν την εναλλακτική υπόθεση  $H_1$  έναντι της μηδενικής υπόθεσης  $H_0$ , έτσι ώστε η ερμηνεία των αποτελεσμάτων να συμβαδίζει με αυτήν των Kass & Raftery. Όπως είδαμε η Μπευζιανή προσέγγιση ξεπερνά κάποια από τα μειονεκτήματα της Κλασικής στατιστικής (παράγραφος 2.1) και συνεπώς είναι μια πιο αξιόπιστη μεθοδολογία. Όμως η Μπευζιανή προσέγγιση είναι μια μεθοδολογία στην οποία αντιμετωπίζουμε κυρίως υπολογιστικά προβλήματα και γι' αυτό τον λόγο δεν είναι ευρέως διαδεδομένη. Ένα σημαντικό πρόβλημα της Μπευζιανής προσέγγισης είναι : η πολυπλοκότητα στον υπολογισμό των εκ των υστέρων κατανομών και της αντίστοιχης περιθώριας πιθανοφάνειας των δεδομένων. Επίσης οι εκ των υστέρων κατανομές, και κυρίως των υπό εξέταση μοντέλων ή υποθέσεων, είναι ευαίσθητες σε αλλαγές των εκ των προτέρων παραμέτρων (π.χ. Lindley's paradox). Στην παρούσα εργασία για να απλοποιηθούν οι υπολογισμοί, χρησιμοποιήθηκαν συζηγείς εκ των προτέρων κατανομές. Πιο συγκεκριμένα χρησιμοποιήθηκαν συζηγείς εκ των προτέρων βασισμένες σε δυνάμεις της πιθανοφάνειας (power prior) οι οποίες είναι ευέλικτες, μας επιτρέπουν την χρήση σημαντικής πληροφορίας από προηγούμενες μελέτες και όταν δεν υπάρχει διαθέσιμη πληροφορία μας δίνουν μια ερμηνεία βασισμένη



σε εικονικά δεδομένα. Μια άλλη προσέγγιση στην επιλογή της εκ των προτέρων κατανομής, προτάθηκε από τον E. T. Jaynes [29] και αφορά την μεγιστοποίηση της πληροφορίας entropy (information entropy) της κατανομής. Η εντροπία είναι ένα μέτρο του κατά πόση πληροφορία στερούμαστε να περιγράψουμε από την κατανομή. Όσο μεγαλύτερη είναι η εντροπία, τόσο λιγότερη είναι η γνώση μας για να ορίσουμε την κατανομή. Ενδιαφέρον για περαιτέρω έρευνα μπορεί να αποτελέσει ο έλεγχος υποθέσεων σε περιπτώσεις μη συζητών εκ των προτέρων κατανομών, όπου οι αριθμητικές πράξεις δεν είναι δυνατόν να επιλυθούν αριθμητικά και απαιτείται η χρήση υπολογιστικών αλγορίθμων Markov Chain Monte Carlo. Μια ιδέα για να εξάγουμε συμπερασματολογία και παράλληλα να ξεπεραστούν οι υπολογιστικές δυσκολίες του παράγοντα Bayes προτάθηκε από τον Valen Johnson (2004) [32]. Ο Johnson προτείνει την μοντελοποίηση της δειγματικής κατανομής του στατιστικού ελέγχου, αντί της δειγματικής κατανομής των παρατηρήσεων. Υποστηρίζει ότι με τον τρόπο αυτό μειώνεται η υποκειμενικότητα από την επιλογή των εκ των προτέρων κατανομών και απλοποιούνται οι υπολογισμοί.

Στον έλεγχο υποθέσεων δεδομένων που προέρχονται από Poisson κατανομή, επιλέξαμε power εκ των προτέρων κατανομές με βάρος ίσο με ένα σημείο δεδομένων. Όταν ελέγχουμε ένα δείγμα είναι εύκολο να διερευνήσουμε την ευαισθησία του παράγοντα Bayes με την διασπορά της εκ των προτέρων κατανομής, ώστε να μελετήσουμε το παράδοξο του Lindley. Για την γενική μορφή της κατανομής της κοινής παραμέτρου, είδαμε ότι καθώς η παράμετρος τείνει στο άπειρο, η διασπορά αυξάνεται και κατ' επέκταση παράγοντας Bayes ελλατώνεται δίνοντας πολύ ισχυρή ένδειξη υπέρ της μηδενικής υπόθεσης. Αντίθετα στην Κλασική στατιστική λόγω της μεγάλης διασποράς θα ήταν απίθανο να παρατηρήσουμε μια τιμή τόσο μακριά από τον μέσο, οπότε θα απορρίπταμε την μηδενική υπόθεση. Ισχύει δηλαδή το παράδοξο του Lindley. Το ίδιο ισχύει και στον έλεγχο υποθέσεων δεδομένων που προέρχονται από Binomial κατανομή. Για την γενική μορφή της κατανομής της κοινής παραμέτρου, όταν η διασπορά αυξάνεται, ο παράγοντας Bayes μειώνεται

και αντίστροφα. Για δύο και περισσότερα δείγματα είναι πιο δύσκολο να μελετήσουμε την ευαισθησία του παράγοντα Bayes σε σχέση με την διασπορά, διότι έχουμε περισσότερες από μία εκ των προτέρων κατανομές. Γνωρίζοντας όμως την συμπεριφορά των διασπορών σε σχέση με τις παραμέτρους, μπορούμε να εξακριβώσουμε την ύπαρξη του παραδόξου του Lindley. Σε κάθε παράδειγμα, με την βοήθεια της  $R$  [46], κατασκευάστηκαν εύχρηστες ρουτίνες για την εξαγωγή των αποτελεσμάτων.

Στα παραδείγματα που γίνεται χρήση πραγματικών δεδομένων, έχει γίνει σιωπηλά η υπόθεση ότι τα δεδομένα προέρχονται από τις αντίστοιχες κατανομές. Για περαιτέρω έρευνα μπορούμε να υπολογίσουμε τον παράγοντα Bayes όταν τα δεδομένα της κατανομής Poisson παρουσιάζουν υπερδιασπορά (overdispersion). Σε αυτήν την περίπτωση μπορούμε να θέσουμε ότι τα δεδομένα να ακολουθούν μια Negative - Binomial κατανομή.

Στην παράγραφο 3.3.3 ελέγχθηκαν ανά δύο πέντε αντιβιοτικά για την καταπολέμηση βαρυσήμαντων μολύνσεων. Στο σημείο αυτό να σχολιάσουμε ότι όταν πραγματοποιούνται πολλαπλές συγκρίσεις η Μπευζιανή προσέγγιση δεν επηρεάζεται. Στην Κλασική στατιστική όταν πολλές μέσες τιμές ελέγχονται ανά δύο, η πιθανότητα να προκύψει σημαντική διαφορά μεταξύ δύο μέσων «από τύχη» αυξάνεται ανάλογα με το πλήθος των συγκρίσεων. Αυτό συμβαίνει διότι στην Κλασική στατιστική η υπό εξέταση παράμετρος αντιμετωπίζεται ως μια σταθερή τιμή που θα περιέχεται ή μη στο διάστημα εμπιστοσύνης. Έτσι όταν πραγματοποιούνται πολλαπλές συγκρίσεις αυτό που λαμβάνει χώρα είναι ένα νέο διωνυμικό πείραμα για το αν η παράμετρος είναι ή δεν είναι εντός του διαστήματος εμπιστοσύνης. Οπότε καθώς το πλήθος των συγκρίσεων αυξάνεται, αυξάνεται και η πιθανότητα να βρεθεί στατιστικά σημαντική διαφορά μεταξύ δύο μέσων. Για περαιτέρω έρευνα έχει ενδιαφέρον να μελετηθεί η περίπτωση όπου μερικές ομάδες ενωποιούνται. Για παράδειγμα σε έναν  $r \times c$  πίνακα συνάφειας κάποιες γραμμές ή στήλες μπορεί να

εκφράζουν το ίδιο πράγμα και να αποτελέσουν μια ομάδα.

Μια πολύ σημαντική διαπίστωση ήταν τα γεγονότα ότι στις περιπτώσεις που θέλουμε να ελέγξουμε δύο δείγματα από κατανομή Binomial το πρόβλημα μπορούσε να εκφραστεί στην μορφή ενός δύο επί δύο πίνακα και αντίστροφα.

Για τον έλεγχο ανεξαρτησίας ενός  $r \times c$  πίνακα συνάφειας υποθέσαμε ότι τα δεδομένα των κελιών ακολουθούν πολυωνυμική κατανομή με εκ των προτέρων κατανομή Dirichlet. Από την βιβλιογραφία γνωρίζουμε ότι υπάρχουν και άλλα δειγματικά μοντέλα για την εξαγωγή συμπερασματολογίας στους πίνακες συνάφειας, όπως για παράδειγμα ο Lindley (1964) [40], χρησιμοποίησε Binomial και Multinomial κατανομές.

Η Altman (1969) [8], που μελέτησε την ακριβή Μπευζιανή ανάλυση ενός  $2 \times 2$  πίνακα συνάφειας, παρήγαγε μια σχέση ανάμεσα στην εκ των υστέρων πιθανότητα μιας αρνητικής σχέσης μεταξύ των γραμμών και των στηλών ενός  $2 \times 2$  πίνακα συνάφειας και την ακριβή πιθανότητα του Fisher όπως δίνεται σε υπάρχων πίνακες ελέγχου της υπόθεσης ανεξαρτησίας μεταξύ γραμμών και στηλών. Ακόμα γενίκευσε τα αποτελέσματα του  $2 \times 2$  πίνακα συνάφειας, για να εξασφαλίσει μια εκ των υστέρων πιθανότητα ότι ένα τυχαίο Dirichlet διάνυσμα είναι στοχαστικά μεγαλύτερο απ' ό,τι το άλλο. Η Altman επίσης μελέτησε το odds ratio για έναν  $2 \times 2$  πίνακα συνάφειας με μια πολυωνυμική κατανομή και μια Dirichlet εκ των προτέρων κατανομή, αντί για δύο ανεξάρτητες Διωνυμικές με Beta εκ των προτέρων. Απέκτησε την εκ των υστέρων πιθανότητα ότι το odds ratio είναι μικρότερο της μονάδας, σαν ένα πεπερασμένο άθροισμα γεωμετρικών πιθανοτήτων. Χρησιμοποιώντας Jeffrey's εκ των προτέρων κατανομή, αυτή η πιθανότητα αποδεικνύεται ότι είναι ίση με το p-value του κλασσικού Fisher exact test κάτω από μια συγκεκριμένη επιλογή εκ των προτέρων κατανοής.

Οι Gunel & Dickey (1974)[26] αναλύουν τρία επιπλέον μοντέλα όπου

- Στο πρώτο μοντέλο οι μετρήσεις των κελιών ακολουθούν ανεξάρτητες Pois-

son κατανομές.

- Στο δεύτερο, δοθέντος τα περιθώρια αθροίσματα των σειρών, οι σειρές είναι ανεξάρτητες και η κάθε μία ακολουθεί πολυωνυμική κατανομή.
- Στο τρίτο, δοθέντος όλα τα περιθώρια σύνολα και την μη ανεξαρτησία, τα κελιά από κοινού ακολουθούν υπεργεωμετρική κατανομή.

Ακολουθούν οι Leonard (1975)[37] και Laird (1978)[36] που έδωσαν μια εναλλακτική Μπευζιανή προσέγγιση, εστιάζοντας στις παραμέτρους του κορεσμένου loglinear μοντέλου. Σε έναν  $2 \times 2$  πίνακα συνάφειας η Laird έθεσε τις παραμέτρους να ακολουθούν ανεξάρτητη κανονική  $N(0, \sigma^2)$  κατανομή. Η ίδια επίσης πρότεινε μια εμπειρική Μπευζιανή ανάλυση. Εκτίμησε το  $\sigma^2$  βρίσκοντας την τιμή που μεγιστοποιεί μια προσέγγιση για την περιθώρια κατανομή των μετρήσεων των κελιών, αποτιμώντας τα παρατηρούμενα δεδομένα.

Οι Crook & Good (1980)[17] εξέτασαν τον έλεγχο ανεξαρτησίας πινάκων συνάφειας, χρησιμοποιώντας συμμετρικές Dirichlet εκ των προτέρων κατανομές και μίξεις αυτών. Επίσης ανέπτυξαν ένα ποσοτικό μέτρο για το βαθμό της απόδειξης της ανεξαρτησίας προερχόμενο από τα περιθώρια σύνολα.

Οι Albert & Gupta (1982)[7] επίσης χρησιμοποίησαν μεικτές Dirichlet κατανομές για να εκτιμήσουν τις παραμέτρους των πινάκων συνάφειας. Ο Albert (1997) [6] γενίκευσε την Μπευζιανή προσέγγιση στον έλεγχο ανεξαρτησίας και εκτίμησε τα odds ratios υπό άλλες προϋποθέσεις, επεκτείνοντας προηγούμενη εργασία του (Albert, 1996).

Στο δεύτερο κεφάλαιο αναφερθήκαμε στον ενδογενή παράγοντα Bayes που προτάθηκε από τους Berger & Pericchi (1996)[10]. Οι Casella & Moreno (2002)[13] παρήγαγαν ενδογενείς εκ των προτέρων κατανομές για τον έλεγχο ανεξαρτησίας  $2 \times 2$  πινάκων συνάφειας υπό διαφορετικά δειγματικά μοντέλα. Επίσης ενδογενείς εκ των προτέρων κατανομές χρησιμοποιούν οι Consonni & La Rocca (2007)[16] για να ελέγξουν την ισότητα δύο συσχετιζόμενων αναλογιών. Κρατάνε μια Dirichlet εκ των προτέρων κατανομή για την μηδενική υπόθεση (Kateri et al., 2001) και θέτουν μια ενδογενή εκ των προτέρων για την εναλλακτική υπόθεση.

Πολλοί γνωστοί στατιστικοί, συμπεριλαμβανομένων των Lindley και Agresti [4], έχουν ασχοληθεί με την Μπευζιανή ανάλυση των πινάκων συνάφειας. Η ραγδαία ανάπτυξη της τεχνολογίας έχει συντελέσει στην αλματώδη διάδοση της Μπευζιανής στατιστικής, που συνεχώς κερδίζει όλο και περισσότερους οπαδούς. Κλείνοντας να θυμηθούμε τον D. V. Lindley [41], ενός από τους βασικότερους ιδρυτές της Μπευζιανής στατιστικής, ο οποίος το 1975 σε μια εργασία του πρόβλεψε ότι ο 21ος αιώνας θα ανήκει στους Μπευζιανούς στατιστικούς. Ο Lindley ήλπιζε αυτή η αλλαγή να έρθει νωρίτερα και στην εργασία του μας προτρέπει να φέρουμε σύντομα την αλλαγή. Άλλωστε η μοναδική οδός της μαθηματικής στατιστικής πρέπει να είναι η Μπευζιανή οδός.

# Παράρτημα Α΄

## Κατανομές

---

### Διακριτές Κατανομές

---

#### Bernoulli

$$\begin{aligned} f(y) &= \theta^y (1 - \theta)^{n-y} & 0 < \theta < 1 & & y = 0, 1 \\ E(y) &= n\theta & \text{Var}(y) &= n\theta(1 - \theta) \end{aligned}$$

---

#### Binomial

$$\begin{aligned} f(y) &= \binom{n}{y} \theta^y (1 - \theta)^{n-y} & 0 < \theta < 1, n = 1, 2, \dots & & y = 0, 1, \dots, n \\ E(y) &= n\theta & \text{Var}(y) &= n\theta(1 - \theta) \end{aligned}$$

---

#### Poisson

$$\begin{aligned} f(y) &= \exp[-\lambda] \frac{\lambda^y}{y!} & \lambda > 0 & & y = 0, 1, 2, \dots \\ E(y) &= \lambda & \text{Var}(y) &= \lambda \end{aligned}$$

---

**Bivariate Poisson**

$$f(y_1, y_2) = e^{-(\lambda_1 + \lambda_2 + \lambda_3)} \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{i=0}^{\min(y_1, y_2)} \binom{y_1}{i} \binom{y_2}{i} i! \left( \frac{\lambda_3}{\lambda_1 \lambda_2} \right)^i \quad \lambda_i > 0$$

$$Y_1 \sim \text{Poisson}(\lambda_1 + \lambda_3) \quad Y_2 \sim \text{Poisson}(\lambda_2 + \lambda_3), \quad \lambda_3 = \text{Cov}(Y_1, Y_2)$$


---

**Multinomial**

$$\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$$

$$\mathbf{y} = (y_1, \dots, y_k)$$

$$0 < \theta_i < 1, \quad \sum_{i=1}^k \theta_i = 1$$

$$y_i = 0, 1, 2, \dots \quad \sum_{i=1}^k y_i = n$$

$$f(\mathbf{y}|n, \boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^k y_i!} \prod_{i=1}^k \theta_i^{y_i}$$

$$E(y_i) = n\theta_i \quad \text{Var}(y_i) = n\theta_i(1 - \theta_i) \quad \text{Cov}(y_i, y_j) = -n\theta_i\theta_j$$


---

**Συνεχείς Κατανομές****Beta**

$$f(y) = \frac{1}{B(\alpha, \beta)} y^{\alpha-1} (1-y)^{\beta-1} \quad \alpha > 0, \quad \beta > 0 \quad 0 < y < 1$$

$$E(y) = \frac{\alpha}{\alpha + \beta} \quad \text{Var}(y) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \quad \frac{1}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}$$


---

**Dirichlet**

$$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_k)$$

$$\mathbf{y} = (y_1, \dots, y_k)$$

$$\alpha_i \geq 0, \quad \alpha_o = \sum_{i=1}^k \alpha_i$$

$$0 \leq y_i \leq 1, \quad \sum_{i=1}^k y_i = 1$$

$$f(\mathbf{y}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_o)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k y_i^{\alpha_i-1}$$

$$\frac{1}{D_k(\boldsymbol{\alpha})} = \frac{\Gamma(\alpha_o)}{\prod_{i=1}^k \Gamma(\alpha_i)}$$

$$E(y_i) = \frac{\alpha_i}{\alpha_o} \quad \text{Var}(y_i) = \frac{\alpha_i(\alpha_o - \alpha_i)}{\alpha_o^2(\alpha_o + 1)} \quad \text{Cov}(y_i, y_j) = -\alpha_i\theta_i\theta_j$$

---

**Gamma**

$$f(y) = \frac{1}{\theta^\kappa \Gamma(\kappa)} y^{\kappa-1} \exp[-\frac{1}{\theta}y] \quad \kappa > 0, \quad \theta > 0 \quad y > 0$$

$$E(y) = \kappa\theta \quad Var(y) = \kappa\theta^2$$

---

**Normal**

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-\frac{(y-\mu)^2}{2\sigma^2}] \quad \mu \in \mathfrak{R}, \quad \sigma > 0 \quad y \in \mathfrak{R}$$

$$E(y) = \mu \quad Var(y) = \sigma^2$$

---

**Ορισμός****Η Αρχή της Πιθανοφάνειας (The Likelihood Principle)**

Αν  $\mathbf{x}$  και  $\mathbf{y}$  δύο δείγματα τ.ω.  $L(\theta; \mathbf{x})$  να είναι ανάλογη της  $L(\theta; \mathbf{y})$  δηλαδή υπάρχει  $c(\mathbf{x}; \mathbf{y})$  τ.ω.  $L(\theta; \mathbf{x}) = c(\mathbf{x}; \mathbf{y})L(\theta; \mathbf{y}) \quad \forall \theta$  τότε τα συμπεράσματα που θα εξάγουμε από τα  $\mathbf{x}$  και  $\mathbf{y}$  για το  $\theta$  οφείλουν να είναι ταυτόσημα.





## Παράρτημα Β΄

### Περίληψη

Το 1975 ο D.V.Lindley δημοσίευσε μια εργασία σχετικά με το μέλλον της Στατιστικής, προβλέποντας ότι ο 21ος αιώνας θα είναι ένας Μπευζιανός αιώνας [41]. Ο Lindley πολύ σωστά υποστηρίζει ότι η μόνη κατάλληλη μέθοδος για να εφαρμόσουμε στατιστική συμπερασματολογία είναι μέσω της Μπευζιανής προσέγγισης. Η Μπευζιανή στατιστική δεν είναι απλά μια ακόμη τεχνική που προστίθεται στο ρεπερτόριο μας, αλλά είναι η μόνη μέθοδος που μπορεί να εξάγει έγκυρη συμπερασματολογία που οδηγεί σε σωστές αποφάσεις. Τέλος επισημαίνει ότι η μοναδική οδός της μαθηματικής στατιστικής πρέπει να είναι η Μπευζιανή οδός.

Η παρούσα εργασία παρουσιάζει την Μπευζιανή προσέγγιση στον έλεγχο υποθέσεων διακριτών και κατηγορικών δεδομένων. Συγκεκριμένα, το πρώτο κεφάλαιο αποτελεί μια εισαγωγή στην Μπευζιανή συμπερασματολογία. Το δεύτερο κεφάλαιο μας φέρνει σε επαφή με τον ορισμό του παράγοντα Bayes και μας παρουσιάζει τους ελέγχους υποθέσεων της ισότητας μιας παραμέτρου με μια συσχετισμένη τιμή. Έλεγχοι υποθέσεων της ισότητας δύο παραμέτρων από εξαρτημένα ή ανεξάρτητα δείγματα εξετάζονται στο τρίτο κεφάλαιο. Τέλος, το τέταρτο κεφάλαιο εξετάζει τον έλεγχο υπόθεσης της ισότητας παραμέτρων από ανεξάρτητα πολλαπλά δείγματα καθώς και τον έλεγχο ανεξαρτησίας ενός  $r \times c$  πίνακα συνάφειας. Επιπλέον, όλες οι στατιστικές αναλύσεις έχουν πραγματοποιηθεί με την χρήση της  $R$ .



## Παράρτημα Γ'

### Abstract

In 1975 professor Dennis Lindley published a paper about the future of Statistics in 21th century where he predicted that it would be a *Bayesian century* [41]. Bayesian statistics is not just another technique to be added to our repertoire, but the only method that can provide valid inference that leads to correct decisions. Finally, he underlined that the Bayesian way is the only way to mathematical statistics.

The current thesis presents the Bayesian approach on hypothesis testing for categorical and count data. More specifically, chapter one is an introduction to Bayesian inference. Chapter two brings us in contact with the definition of Bayes Factor and presents tests on the hypothesis of equality of one parameter with a specific value. Hypothesis testing of equality of two parameters of dependent or independent samples are illustrated on the third chapter. Finally, chapter four examines the hypothesis testing of equality of more than two parameters of independent samples and the hypothesis testing of independence  $r \times c$  contingency tables. All the statistical analyses were performed using *R*.

*Keywords* : Bayesian inference, Poisson, Binomial, Bayesian hypothesis tests, Bayes Factor, power priors, Multinomial, Dirichlet, contingency tables.



# Βιβλιογραφία

- [1] Πανάρετος Ι. και Ξεκαλάκη Ε. , 2000. *Εισαγωγή στη Στατιστική σκέψη*. Αθήνα.
- [2] Ρούσσας Γρ. Γεώργιος, 1976. *Στατιστική Συμπερασματολογία - Έλεγχος Υποθέσεων*. Ζήτη.
- [3] Agresti Alan, 1990. *Categorical Data Analysis*. John Wiley & Sons.
- [4] Agresti Alan and Hitchcock B. David. Bayesian inference for categorical data analysis: A survey. *University of Florida and University of South Carolina, Tech. Rep.*
- [5] Aitkin M. , 1991 Posterior bayes factors. *Journal of the Royal Statistical Society B*, **53**:111–142.
- [6] Albert H. James , 1997. Bayesian testing and estimation of association in a two-way contingency table. *Journal of the American Statistical Association*, **92**:685–693.
- [7] Albert, J. H. and Gupta, A. K , 1982. Mixtures of dirichlet distributions and estimation in contingency tables. *The Annals of Statistics*, **10**:1261–1268.
- [8] Altham M. E. Patricia , 1969. Exact bayesian analysis of a  $2 \times 2$  contingency table, and fisher’s exact significance test. *Journal of the Royal Statistical Society B*, **31**:261–269.
- [9] Barlett M. S. , 1957 A comment on d. v. lindley’s statistical paradox. *Biometrika*, **44**:533–534.

- [10] Berger, J. O. and Pericchi, L. , 1996 The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**:109–122.
- [11] Bernardo Jose M. and Smith Adrian F. M. , 1994 *Bayesian Theory*. John Wiley & Sons.
- [12] Carlin P. Bradley and Louis A. Thomas , 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman and Hall/CRC.
- [13] Casella, G. and Moreno, E. , 2002 Objective bayesian analysis of contingency tables. *Department of Statistics, University of Florida*, **Tech. Rep.**.
- [14] Chen Ming-Hui and Ibrahim G. Joseph , 2006. The relationship between the power prior and hierarchical models. *Bayesian Analysis*, **1**:551–574.
- [15] Chen Ming-Hui and Ibrahim G. Joseph , 2003 Conjugate priors for generalized linear models. *Statistica Sinica*, **13**:461–476.
- [16] Consonni Guido and Luca La Rocca , 2007. Intrinsic tests for the equality of two correlated proportions. *University of Pavia and University of Modena & Reggio Emilia*, **Tech. Rep.**.
- [17] Crook J. F. and Good I. J , 1980. n the application of symmetric dirichlet distributions and their mixtures to contingency tables, part ii. *The Annals of Statistics*, **8**:1198–1218.
- [18] De Santis, F. and Spezzaferri, F. , 1997 Alternative bayes factors for model selection. *The Canadian Journal of Statistics*, **4**:503–515.
- [19] Diaconis, P. and Ylvisaker, D. , 1979 Conjugate priors for exponential families. *The Annals of Statistics*, **7**:269–281.
- [20] Geisser, S. and Eddy, W. , 1979 A predictive approach to model selection. *Journal of the American Statistical Association*, **74**:153–160.

- 
- [21] Gelfand, A. E. and Dey, D. K. , 1994 Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society B*, **56**:501–514.
- [22] Gelfand, A. E., Dey, D. K. and Chang, H. , 1992 *Model determination using predictive distributions with implementation via sampling-based methods*. In *Bayesian Statistics 4*. (Edited by J. Bernardo, J.O. Berger, A. P. Dawid and A. F. M. Smith) Oxford University Press.
- [23] Gelman Andrew, Carlin John B., Stern Hal S., Rubin Donald B. , 1995 *Bayesian Data Analysis*. Chapman and Hall.
- [24] Giampaoli Viviana and Singer M. Julio , 2004. Bayes factors for comparing two restricted means: An example involving hypertense individuals. *Journal of Data Science*, **2**:399–418.
- [25] Goldberg, R., Gore, J., Alpert, J. and Dalen, P. , 1988 Incidence and case fatality rates of acute myocardial infarction (1975-1984): The worcester heart attack study. *American Heart Journal*, **115**:761–767.
- [26] Gunel Erdogan and Dickey James , 1974. Bayes factors for independence in contingency tables. *Biometrika*, **61**:545–557.
- [27] Hashemi Lobat, Nandram Balgobin and Golberg Robert , 1997. Bayesian analysis for a single  $2 \times 2$  table. *Statistics in Medicine*, **16**:1311–1328.
- [28] Ibrahim G. Joseph and Chen Ming-Hui , 2000 Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, **84**:121–137.
- [29] Jaynes T. E. , 1994 *Probability Theory : The Logic of Science*. .
- [30] Jeffreys , 1935. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, **31** :203–222.



- [31] Jeffreys , 1961. *Theory of probability (3rd ed.)*. Oxford University Press.
- [32] Johnson E. Valen , 2005. Bayes factors based on test statistics. *Journal of the Royal Statistical Society B*, **67**:689–701.
- [33] Karlis. D. and Ntzoufras I. , 2006 Bayesian analysis of the differences of count data. *Statistics in Medicine*, **25**:1885–1905.
- [34] Kass Robert E. and Raftery Adrian E. , 1995. Bayes factors. *Journal of the American Statistical Association*, **90**:773–795.
- [35] Kateri M., Papaioannou T. and Dellaportas P. ,2001. Bayesian analysis of correlated proportions. *Sankhya : The Indian Journal of Statistics*, **63**:270–285.
- [36] Laird, N. M , 1978. Empirical bayes methods for two-way contingency tables. *Biometrika*, **65**:581–590.
- [37] Leonard Tom , 1975. Bayesian estimation methods for two-way contingency tables. *Journal of the Royal Statistical Society Series B (Methodological)*, **37**:23–37.
- [38] Leslie E. Dary, Geoffrey J. Bourke and James Mc. Gilvray , 1991. *Interpretation and users of Medical Statistics*. Blackwell Science.
- [39] Lindley D. V. , 1957 A statistical paradox. *Biometrika*, **44**:187–192.
- [40] Lindley D. V. , 1964 The bayesian analysis of contingency tables. *The Annals of Mathematical Statistics*, **35**:1622–1643.
- [41] Lindley D. V. , 1975 The future of statistics : A bayesian 21th century. *Advances in Applied Probability*, **7**:106–115.
- [42] Morris, C. N. ,1982 Natural exponential families with quadratic variance functions. *The Annals of Statistics*, **10**:65–80.
- [43] Morris, C. N. ,1983 Natural exponential families with quadratic variance functions: statistical theory. *The Annals of Statistics*, **11** :515–529.

- 
- [44] Nandram Balgobin and Choi Jai Won , 2007. Alternative tests of independence in two-way categorical tables. *Journal of Data Science*, **5**:217–237.
- [45] O’ Hagan, A. , 1995 Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society B*, **57**:99–138.
- [46] *R* manual. An introduction to R notes on R: A programming environment for data analysis and graphics version 2.2.0.
- [47] Rosner R. Bernard , 2005 . *Fundamentals of Biostatistics*. .
- [48] Shafer Glenn , 1982. Lindley’s paradox. *Journal of American Statistical Association*, **77**:325–334.
- [49] Yuyan Duan, 2005. A modified bayesian power prior approach with applications in water quality evaluation. *Dissertation submitted to the Faculty of Virginia Polytechnic Institute and State University*.

# Ευρετήριο

- Bartlett*, 23  
*Beta* κατανομή, 35  
*Binomial* κατανομή, 34, 53, 79  
*Bivariate Poisson* κατανομή, 62  
*Chen & Ibrahim*, 13  
*Dirichlet* κατανομή, 66, 86  
*Gamma* κατανομή, 27  
*Jeffreys*, 22  
*Jeffreys prior*, 8  
*Karl Pearson*, 85  
*Kass & Raftery*, 22  
*Lindley's paradox*, 23  
*Markov Chain Monte Carlo*, 63  
*McNemar test*, 64  
*Multinomial* κατανομή, 65, 87  
*Neyman & Pearson*, 17  
*Poisson* κατανομή, 26, 45, 73  
*Skellam* κατανομή, 62  
*fractional Bayes factor*, 23  
*hyperprior*, 11  
*intra – class correlation*, 86  
*intrinsic Bayes factor*, 23  
*mixture prior*, 7  
*p – value*, 18  
*power prior*, 13  
*target distribution*, 63  
έλεγχος των λόγων πιθανοφάνειας  
(*likelihood ratio test*), 22  
ένθετο (*nested*), 18  
συνάρτηση απόκλισης (*deviance*), 22  
ψευδοπαράγοντας *Bayes*, 22  
  
αλγόριθμος (*Metropolis*), 64  
αρχή της πιθανοφάνειας (*Likelihood principle*), 19  
  
εκ των προτέρων (*prior*), 1  
εκ των υστέρων (*posterior*), 1  
εκθετική οικογένεια, 4  
ελεγχοσυνάρτηση, 17  
εμπειρική Μπευζιανή (*empirical Bayes*), 12  
εναλλακτική υπόθεση, 17  
επίπεδη (*flat*), 8  
ετερογένεια, 15  
  
φυσική συζυγής, 4  
  
γνήσια (*proper*), 9  
  
ιεραρχική μοντελοποίηση (*Hierarchical modeling*), 12  
ιστορικά δεδομένα (*historical data*), 13  
  
λόγος των εκ των υστέρων σχετικών πιθανοτήτων  
(*posterior odds*), 20  
λανθάνοντα δεδομένα, 63  
  
μετά-ανάλυση (*meta – analysis*), 40  
μη γνήσια (*improper*), 8  
μη πληροφοριακή (*non informative*), 3  
μηδενική υπόθεση, 17  
  
πίνακας συνάφειας (*contingency table*), 64, 85  
παράγοντας *Bayes* (*Bayes Factor*), 20, 27, 35, 46, 54, 67,  
74, 80, 88  
παράμετρος ακρίβειας (*precision parameter*), 14  
περιοχή απόρριψης, 18  
πληροφορία *Fisher*, 9  
πληροφοριακή (*informative*), 3  
  
σφάλμα Τύπου I, 18  
σφάλματα Τύπου II, 19  
στατιστικός έλεγχος, 17  
σχετιζόμενα δείγματα, 61  
συσχετιζόμενα δείγματα, 64  
συζυγής (*conjugate*), 3  
  
τρέχοντα δεδομένα (*current data*), 13  
  
υποκειμενική (*subjective*), 2