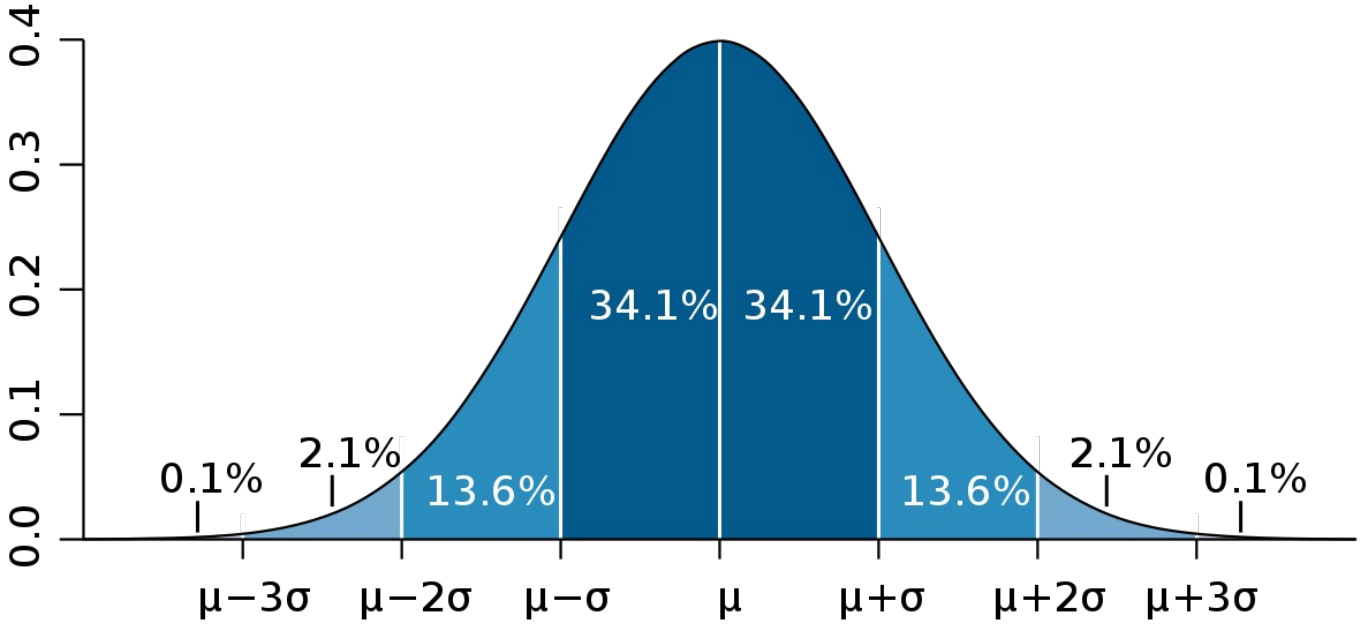
Θεωρία Πιθανοτήτων και Στατιστική



Διδάσκων: Επαμεινώνδας Διαμαντόπουλος
Επικοινωνία: epdiaman@ee.duth.gr

Περιεχόμενα 8ου μαθήματος

- Εκτιμητές μέγιστης πιθανοφάνειας.
- Διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού.
- Διάστημα εμπιστοσύνης για την αναλογία ενός χαρακτηριστικού στον πληθυσμό.
- Υπολογισμός μεγέθους δείγματος για μέση τιμή ή ποσοστό.
- Βήματα μιας στατιστικής έρευνας.
- Δοκιμασία χι-τετράγωνο ως έλεγχος ομοιογένειας.
- Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας.

Γνωστικοί στόχοι 8ου μαθήματος

Στο τέλος αυτού του μαθήματος, ο φοιτητής πρέπει να είναι σε θέση :

- Να υπολογίζει τον εκτιμητή μέγιστης πιθανοφάνειας για μία άγνωστη παράμετρο.
- Να υπολογίζει διάστημα εμπιστοσύνης για τη μέση τιμή του πληθυσμού.
- Να υπολογίζει διάστημα εμπιστοσύνης για την αναλογία του πληθυσμού.
- Να μπορεί να υπολογίζει το απαραίτητο μέγεθος δείγματος για μία μελέτη.
- Να γνωρίζει τα βήματα μίας στατιστικής έρευνας και να αντιλαμβάνεται την ηθική σειρά των βημάτων που πρέπει να ακολουθήσει.
- Να γνωρίζει τις προϋποθέσεις της δοκιμασίας χι-τετράγωνο.
- Να μπορεί να καταγράψει την στατιστική και την ερευνητική υπόθεση μίας δοκιμασίας χι-τετράγωνο
- Να μπορεί να υλοποιήσει τους υπολογισμούς μίας δοκιμασίας.

Εκτιμητές μέγιστης πιθανοφάνειας

Εκτιμητές μέγιστης πιθανοφάνειας

Στην Στατιστική, η εκτίμηση μέγιστης πιθανότητας (MLE) είναι μια μέθοδος εκτίμησης των παραμέτρων μιας άγνωστης κατανομής πιθανότητας, από ένα σύνολο παρατηρούμενων δεδομένων.

Αυτό επιτυγχάνεται με τη μεγιστοποίηση μιας συνάρτησης πιθανότητας, έτσι ώστε, σύμφωνα με το υποτιθέμενο στατιστικό μοντέλο, τα παρατηρούμενα δεδομένα να είναι πιο πιθανά να εμφανιστούν.

Το σημείο στο χώρο των παραμέτρων που μεγιστοποιεί τη συνάρτηση πιθανότητας ονομάζεται εκτιμητής μέγιστης πιθανότητας.

Η διαδικασία είναι εφικτή τόσο για συνεχείς όσο και για διακριτές τμ. Η μέθοδος προτάθηκε από τον Fisher το 1912.

X, Y ανεξ. τ.μ. $f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y)$

Εκτιμητές μέγιστης πιθανοφάνειας

Έστω ότι τα δεδομένα $x = (x_1, \ldots, x_n)$, προέρχονται από μία κατανομή της οποίας γνωρίζουμε το είδος της κατανομής αλλά όχι τις παραμέτρους αυτής. Έστω επίσης πως οι παρατηρήσεις είναι ανεξάρτητες.

Αν f η σ.μ.π. ή σ.π.π και θ είναι μία άγνωστη παράμετρος του πληθυσμού τότε

$$x_i \text{ ανεξάρτητες} \rightarrow f_X(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

$f_X(x_1, x_2, \ldots, x_n; \theta) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$

Ορίζουμε τη συνάρτηση πιθανοφάνειας L ως εξής :

$$L(\theta) = L(X, \theta) = f_X(x_1, \ldots, x_n; \theta)$$

Αναζητούμε την τιμή της παραμέτρου θ που μεγιστοποιεί τη συνάρτηση L. Η τιμή που βρίσκουμε με την παραπάνω διαδικασία ονομάζεται

Εκτιμητής Μέγιστης Πιθανοφάνειας.

Εκτιμητές μέγιστης πιθανοφάνειας

Οι παρατηρήσεις θεωρούνται ανεξάρτητες, άρα

$$f_X(x_1, \ldots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

Η συνάρτηση πιθανοφάνειας L γράφεται:

$$L(\theta) = L(X, \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdot \ldots \cdot f(x_n; \theta)$$

Για να βρούμε την τιμή της παραμέτρου θ που μεγιστοποιεί τη συνάρτηση L, αρκεί να βρούμε την τιμή που μεγιστοποιεί την

$$l(\theta) = \ln L(\theta) = \Sigma \ \ln f(x_1; \theta).$$

Ο **εκτιμητής μέγιστης πιθανοφάνειας** προκύπτει από την επίλυση της εξίσωσης

$$d/d\theta \ l(\theta) = 0 \text{ ή } \ln L(X, \theta) = 0.$$

$P(X=a, Y=\beta) = P(X=a) \cdot P(Y=\beta)$

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 1. Έστω X μία ΤΜ με κατανομή

x	0	1	2	3
$f_X(x)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

Για να εκτιμήσουμε την άγνωστη παράμετρο θ, υλοποιούμε δειγματοληψία 10 τιμών και παίρνουμε το δείγμα (3, 0, 2, 1, 3, 2, 1, 0, 2, 1). Να βρεθεί η τιμή της παραμέτρου θ με τη μέθοδο του Εκτιμητή Μέγιστης Πιθανοφάνειας.

$$L(\theta) = f_X((3,0,2,1,3,2,1,0,2,1)\,|\,\theta) = f_X(3|\theta) \cdot f_X(0|\theta) \cdots f_X(1|\theta) =$$

$$= f_X(0|\theta)^2 \cdot f_X(1|\theta)^3 \cdot f_X(2|\theta)^3 \cdot f_X(3|\theta)^2 =$$

$$= \left(\frac{2\theta}{3}\right)^2 \cdot \left(\frac{\theta}{3}\right)^3 \cdot \left(\frac{2(1-\theta)}{3}\right)^3 \cdot \left(\frac{1-\theta}{3}\right)^2 = \frac{2^5}{3^{10}} \cdot \theta^5 \cdot (1-\theta)^5 .$$

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 1.

$$L(\theta) = \frac{2^5}{3^{10}} \cdot \theta^5 \cdot (1-\theta)^5$$

$$\ell(\theta) = \ln L(\theta) = \ln \frac{2^5}{3^{10}} + 5\ln\theta + 5\ln(1-\theta)$$

$$\ell'(\theta) = \frac{5}{\theta} - \frac{5}{1-\theta} \quad . \quad \ell'(\theta) = 0 \Leftrightarrow \frac{5}{\theta} = \frac{5}{1-\theta} \Leftrightarrow \theta = \frac{1}{2}$$

θ	0		$\frac{1}{2}$		1
$\ell'(\theta)$		$+$	0	$-$	
$\ell(\theta)$		$\nearrow$	M	$\searrow$	

Άρα, $\theta = \frac{1}{2}$ Εκτιμητής Μέγιστης Πιθανοφάνειας.

$X \sim Exp(\lambda)$, $x_1, x_2, \dots, x_n$ $\hat{\lambda} = \frac{n}{x_1 + x_2 + \dots + x_n}$

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 2. Έστω ότι η τ.μ. X μετράει το χρόνο μεταξύ αφίξεων σε μία διαδικασία Poisson και πως X ~ Exp(λ). Παρακολουθούμε δείγμα 4 τιμών της X και βρίσκουμε τις τιμές 1.23, 3.32, 1.98, 2.12. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου λ.

$X \sim Exp(\lambda)$, $f_X(x) = \lambda e^{-\lambda x}$, $x \geq 0$

$L(\lambda) = f_X((\overset{x_1}{1.23}, \overset{x_2}{3.32}, \overset{x_3}{1.98}, \overset{x_4}{2.12}) | \lambda) = f_X(1.23|\lambda) \cdot f_X(3.32|\lambda) \cdot f_X(1.98|\lambda) \cdot f(2.12|\lambda)$

$= \lambda \cdot e^{-1.23\lambda} \cdot \lambda e^{-3.32\lambda} \cdot \lambda e^{-1.98\lambda} \cdot \lambda e^{-2.12\lambda} = \lambda^4 e^{-8.65\lambda}$

$\ell(\lambda) = \ln L(\lambda) = 4 \ln \lambda - 8.65\lambda$. $\ell'(\lambda) = \frac{4}{\lambda} - 8.65 = 0 \Leftrightarrow \lambda = \frac{4}{8.65} = 0,462$.

Μέγιστη Πιθανοφάνεια $\not\rightarrow$ Αμεροληψία

Παράδειγμα 1 (με αφορμή την άσκηση 2)
Έστω X μία μεταβλητή που γνωρίζουμε ότι ακολουθεί την Exp(λ) αλλά δεν γνωρίζουμε το λ (= πλήθος αφίξεων / μονάδα χρόνου). Για το λόγο αυτό παρατηρούμε n διαφορετικές διάρκειες μεταξύ αφίξεων $X_1, X_2, \ldots, X_n$ και εκτιμούμε τον ρυθμό λ από τον τύπο:

$$\hat{\lambda} = \frac{n}{X_1 + X_2 + \ldots + X_n}$$

Ο εκτιμητής αυτός **δεν είναι αμερόληπτος**. Πράγματι, για n = 1, είναι

$$E(\hat{\lambda}) = E\left(\frac{1}{X_1}\right) = \int_0^{+\infty} \frac{\lambda}{x} e^{-\lambda/x} dx = +\infty$$

ενώ για n > 1 είναι $X_1 + X_2 + \ldots + X_n \sim$ Γάμμα(n, λ), άρα $1/(X_1 + X_2 + \ldots + X_n) \sim$ Inverse Γάμμα(n, λ) και

$$E(\hat{\lambda}) = E\left(\frac{n}{X_1 + X_2 + \ldots + X_n}\right) = n\frac{\lambda}{n-1} \neq \lambda .$$

Ιδιαίτερα, συμπεραίνουμε, ότι ο $\hat{\lambda} = \frac{n-1}{X_1 + X_2 + \ldots + X_n}$ είναι ένας αμερόληπτος εκτιμητής του λ.

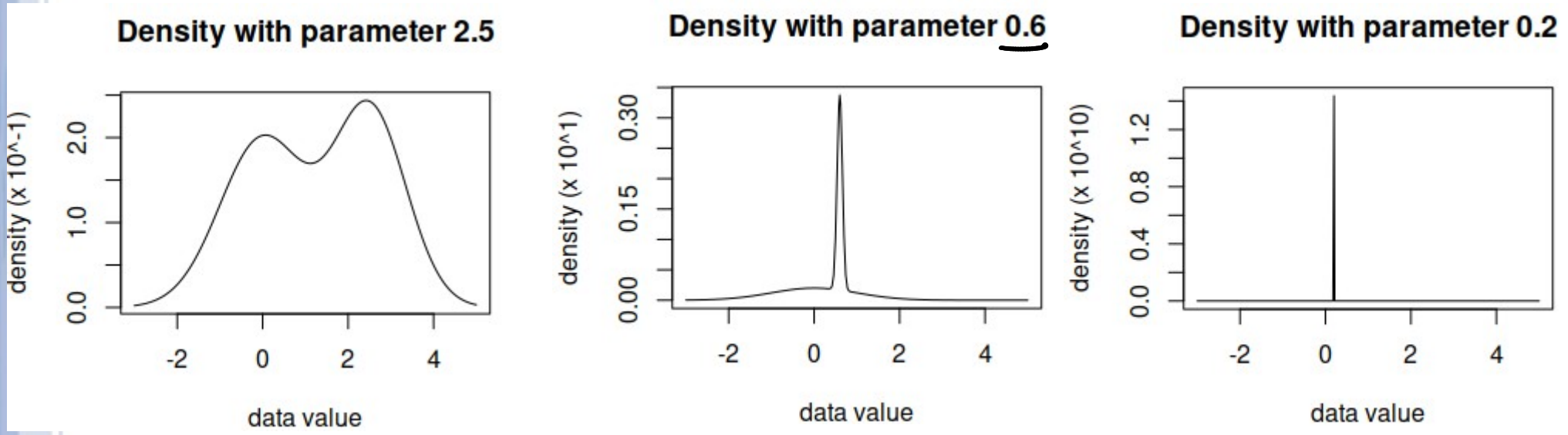
Σημείωση: Αν X ~ Inverse Γαμμα(α, β) τότε $E(X^k) = \beta^k \Gamma(\alpha - k) / \Gamma(\alpha)$.

Μέγιστη Πιθανοφάνεια $\not\rightarrow$ Συνέπεια

Παράδειγμα 2

Έστω X τμ για την οποία γνωρίζουμε ότι $X \sim 1/2N(0,1) + 1/2N\left(t, e^{-\frac{2}{t^2}}\right)$.

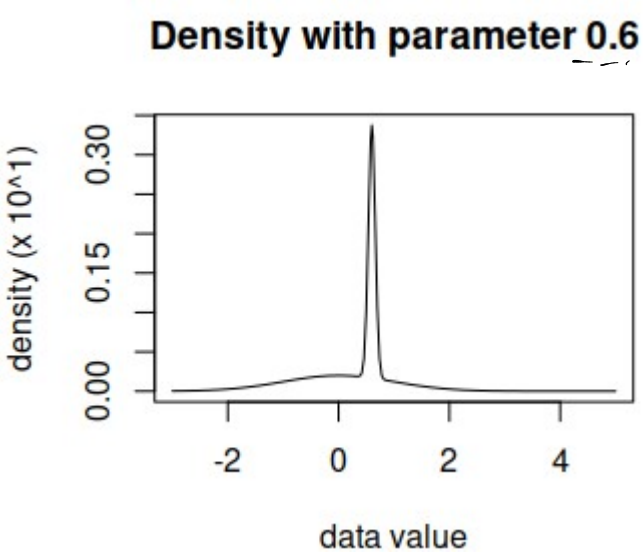
$$f_X(x) = \frac{1}{2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} + \frac{1}{2}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-t)^2}{2\sigma^2}},\quad \sigma = e^{-\frac{2}{t^2}}.$$

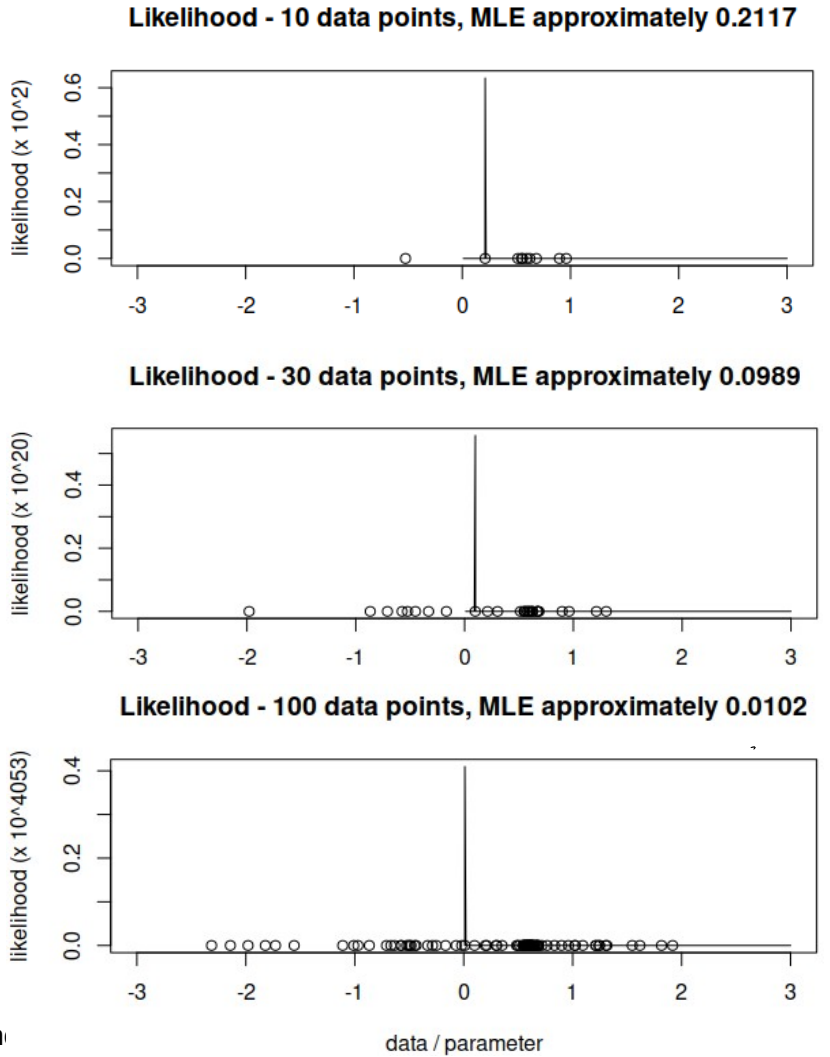


Πηγή: https://radfordneal.wordpress.com/2008/08/09/inconsistent-maximum-likelihood-estimation-an-ordinary-example/

Μέγιστη Πιθανοφάνεια ↛ Συνέπεια

Παράδειγμα 2
Για t = 0,6, και $\{x_i\}_{i\le n}$, n = 10 ή 30 ή 100 παρατηρήσεις της X, ο Ε.Μ.Π. της παραπάνω κατανομής **δεν είναι συνεπής**.





Πηγή: https://radfordneal.wordpress.com/2008/08/09/inconsistent-maximum-likelih

Μέγιστη Πιθανοφάνεια $\not\rightarrow$ Μοναδικότητα

Παράδειγμα 3

Αν X ~ Laplace(μ, β) με $f(x) = \frac{1}{2\beta}\exp\left(-\frac{|x-\mu|}{\beta}\right)$, με άγνωστες παραμέτρους μ, β και $x_1, x_2, \ldots, x_n$, ένα δείγμα τιμών, τότε αποδεικνύεται ότι ο εκτιμητής μέγιστης πιθανοφάνειας για την παράμετρο μ είναι η διάμεσος του δείγματος(1).

Καθώς η διάμεση τιμή δεν είναι μοναδική όταν το n είναι άρτιος, καταλαβαίνουμε ότι ο εκτιμητής δεν ορίζεται με μοναδικό τρόπο.

(1) Μία απόδειξη είναι διαθέσιμη εδώ:
https://math.stackexchange.com/questions/240496/finding-the-maximum-likelihood-estimator

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 3. Έστω X μία ΤΜ με κατανομή

$$f(x) = \frac{1}{2\beta} e^{-\frac{|x|}{\beta}}$$

Επιλέγουμε δείγμα n τιμών από την κατανομή της X. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου β.

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 3.

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 4. Μία ΤΜ Χ, ακολουθεί την κατανομή Pareto με σ.π.π.

$$f(x) = \theta x^{-\theta-1},\ x \geq 1,\ \theta > 1.$$

Επιλέγουμε δείγμα n τιμών από την κατανομή της Χ. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου θ.

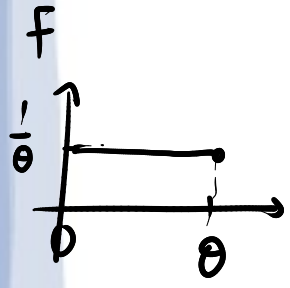
Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 4.

Εκτιμητές μέγιστης πιθανοφάνειας

Άσκηση 5. Για μία ΤΜ Χ, γνωρίζουμε ότι Χ ~ U(0, θ). Επιλέγουμε δείγμα n τιμών από την κατανομή της Χ. Να βρεθεί ο Εκτιμητής Μέγιστης Πιθανοφάνειας της παραμέτρου θ.

$X \sim U(0,\theta)$, $f_X(x) = \frac{1}{\theta}$, $0 \le x \le \theta$, και 0 αλλού.



$0 \le x_1, x_2, .., x_n \le \theta \Rightarrow \theta \ge \max\{x_1, x_2, .., x_n\}$

$f((x_1, x_2, .., x_n)|\theta) = f(x_1|\theta) .. f(x_n|\theta) = \frac{1}{\theta} \cdot \frac{1}{\theta} .. \frac{1}{\theta} = \frac{1}{\theta^n}$.

$L(\theta) = \frac{1}{\theta^n} \searrow (0, \infty)$ και $\theta \ge \max\{x_1, x_2.., x_n\} \Rightarrow \theta = \max\{x_1, x_2.., x_n\}$.

Διάστημα εμπιστοσύνης (confidence interval)

Ποσοστιαία σημεία κανονικής κατανομής

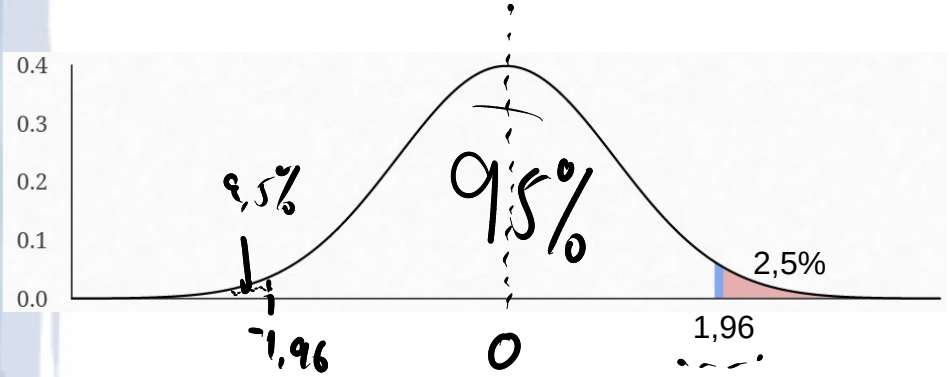
Έστω $Z \sim N(0, 1)$, $0 < \alpha < 1$, και z_α ο αριθμός για τον οποίο

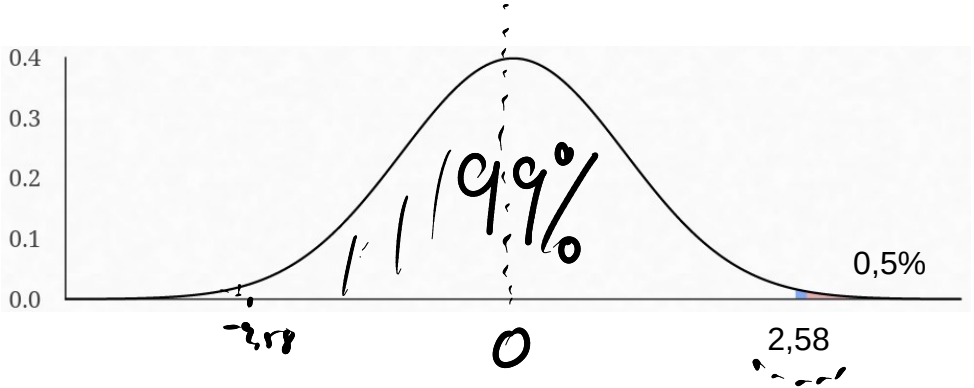
$$\mathbf{P(Z > z_\alpha) = \alpha \text{ ή } \Phi(z_\alpha) = 1 - \alpha.}$$

Το z_α ονομάζεται α – ποσοστιαίο σημείο της κανονικής κατανομής. Ενδεικτικά:

Αν $\alpha = 0{,}025$, τότε $z_{0{,}025}$ είναι ο αριθμός για τον οποίον $\Phi(z_{0{,}025}) = 0{,}975$ ή $\mathbf{z_{0{,}025} = 1{,}96.}$

Αν $\alpha = 0{,}005$, τότε $z_{0{,}005}$ είναι ο αριθμός για τον οποίον $\Phi(z_{0{,}005}) = 0{,}995$ ή $\mathbf{z_{0{,}005} = 2{,}58.}$





Κατανομή αριθμητικού μέσου

Έστω ότι μετρούμε n > 30, τιμές $x_1, x_2, \ldots, x_n$, από έναν πληθυσμό με άγνωστη αναμενόμενη τιμή μ και τυπική απόκλιση σ.

Ο αριθμητικός μέσος των τιμών είναι ο αριθμός $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$ και είναι **μία σημειακή εκτίμηση** για την άγνωστη μέση τιμή του πληθυσμού.

Κάθε μία από τις τιμές που παρατηρήσαμε είναι το αποτέλεσμα ενός ανεξάρτητου πειράματος δειγματοληψίας στον ίδιο πληθυσμό (ανεξάρτητο = καλής ποιότητας δειγματοληψία). Πιο συγκεκριμένα, η τιμή x_i που παρατηρήσαμε είναι μία από τις πιθανές τιμές που παίρνει μία τυχαία μεταβλητή, την οποία ας την ονομάσουμε X_i, i = 1, 2, …, n. Με τον τρόπο αυτό ο αριθμητικός μέσος που υπολογίσαμε αποτελεί μία από τις πιθανές τιμές που μπορεί να πάρει η τυχαία μεταβλητή

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Αναδεικνύεται με φυσικό τρόπο το ερώτημα:

Ποια είναι η κατανομή της συνεχούς μεταβλητής $\bar{X}$;

Κατανομή αριθμητικού μέσου

Καθώς, $n > 30$ και οι τ.μ. X_i, $i = 1, 2, \ldots, n$, είναι ισόνομες (αφού λαμβάνονται από τον ίδιο πληθυσμό) και ανεξάρτητες, το Κεντρικό Οριακό Θεώρημα μας εξασφαλίζει ότι: $X_1 + X_2 + \ldots + X_n \sim N(n \cdot \mu, n \cdot \sigma^2)$, ή ισοδύναμα ότι:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Δηλαδή προκύπτει το αξιοσημείωτο συμπέρασμα, πως:

Ανεξάρτητα από την κατανομή που ακολουθεί ο πληθυσμός μέσα από τον οποίο λαμβάνουν τιμές οι $X_1, X_2, \ldots, X_n$, η τυχαία μεταβλητή του αριθμητικού μέσου, για n αρκετά μεγάλο, ακολουθεί την κανονική κατανομή $N(\mu, \sigma^2 / n)$,

Δραστηριότητα
Επαληθεύστε ότι αν $X \sim N(\mu, \sigma^2)$, και $Z = (X - \mu) / \sigma$, τότε $Z \sim N(0, 1)$.

Αναπαράσταση: https://www.zoology.ubc.ca/~whitlock/Kingfisher/CLT.htm

$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

Διάστημα Εμπιστοσύνης: σ γνωστό

Το συμπέρασμα που καταλήξαμε: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ δίνει τη δυνατότητα να κάνουμε μία **εκτίμηση σε διάστημα για την (άγνωστη) μέση τιμή μ του πληθυσμού.** Πράγματι, από τη θεωρία που συνοδεύει την κανονική κατανομή, γνωρίζουμε ότι:

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \Leftrightarrow \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0, 1).$$

Αν $\alpha > 0$, θεωρούμε τον αριθμό $z_{\alpha/2}$, για τον οποίο $P(Z > z_{\alpha/2}) = \alpha/2$. Τότε, θα είναι:

$$P\left(-z_{\alpha/2} < \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} < z_{\alpha/2}\right) = 1-\alpha \Leftrightarrow P\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} < \mu < \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) = 1-\alpha.$$

Τι σημαίνει η τελευταία ισότητα; Η απάντηση είναι:
Εάν E(X) = μ και η διαδικασία λήψης δείγματος μεγέθους n, επαναληφθεί πολλές φορές, τότε σε κάθε 100 διαστήματα της μορφής

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right)$$

στα (1 – α)·100 από αυτά θα βρίσκεται μέσα η πραγματική μέση τιμή μ του πληθυσμού.

Διάστημα Εμπιστοσύνης: σ γνωστό

Το διάστημα

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

ονομάζεται **διάστημα εμπιστοσύνης** (confidence interval) για την άγνωστη μέση τιμή του πληθυσμού.

Οι δύο πιο συνηθισμένες επιλογές για το α είναι α = 0,05 και α = 0,01.

Αν α = 0,05 τότε $z_{0,025} = 1,96$ και βρίσκουμε το

95% διάστημα εμπιστοσύνης $\left(\overline{X} - 1,96 \frac{\sigma}{\sqrt{n}}, \overline{X} + 1,96 \frac{\sigma}{\sqrt{n}} \right)$

Αν α = 0,01 τότε $z_{0,005} = 2,58$ και βρίσκουμε το

99% διάστημα εμπιστοσύνης $\left(\overline{X} - 2,58 \frac{\sigma}{\sqrt{n}}, \overline{X} + 2,58 \frac{\sigma}{\sqrt{n}} \right)$

Αναπαράσταση: https://www.zoology.ubc.ca/~whitlock/Kingfisher/CIMean.htm

Διάστημα Εμπιστοσύνης: Συνήθη σφάλματα

Λανθασμένες ερμηνείες για το 95% διάστημα εμπιστοσύνης $\left(\overline{X} - 1{,}96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1{,}96\frac{\sigma}{\sqrt{n}}\right)$

- Επίπεδο εμπιστοσύνης 95% δεν σημαίνει ότι στο ένα διάστημα που υπολογίζουμε από το δείγμα μας, υπάρχει 95% πιθανότητα το διάστημα να καλύπτει την παράμετρο πληθυσμού. Σε αυτό το ένα που υπολογίζουμε είτε θα είναι είτε δεν θα είναι. Αν όμως υπολογίσουμε 100 τότε περιμένουμε τα 95 από αυτά να καλύπτουν την άγνωστη μέση τιμή μ.

- Ένα 95% διάστημα εμπιστοσύνης δεν περιέχει το 95% των τιμών του δείγματος.

Διάστημα Εμπιστοσύνης

Άσκηση 1

Υποθέτουμε ότι οι βαθμολογίες στις εξετάσεις κατανέμονται κανονικά. Δεν γνωρίζουμε τη μέση τιμή αλλά γνωρίζουμε πως η τυπική απόκλιση είναι σ = 3 μονάδες. Σε δείγμα 36 γραπτών βρέθηκε μέση βαθμολογία ίση με 68 μονάδες. Να βρεθεί το 95% και το 99% διάστημα εμπιστοσύνης για την άγνωστη μέση βαθμολογία όλων των φοιτητών.

95% Δ.Ε. $[67.02, 68.98]$.

$X \sim N(\mu, 3^2)$, $\mu = ;$

$n = 36 \quad \bar{x} = 68$

95% : $\bar{x} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96 \frac{\sigma}{\sqrt{n}}$

$68 - 1{,}96 \frac{3}{\sqrt{36}} \leq \mu \leq 68 + 1{,}96 \cdot \frac{3}{\sqrt{36}}$ ή $67{,}02 \leq \mu \leq 68{,}98$

Διάστημα Εμπιστοσύνης: σ άγνωστο

Στην πράξη, συνήθως δεν γνωρίζουμε την τυπική απόκλιση σ του πληθυσμού. Ως εκ τούτου, δεν είναι δυνατόν να υπολογίσουμε το διάστημα εμπιστοσύνης ως

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \right)$$

Ωστόσο, στη θέση της άγνωστης τυπικής απόκλισης του πληθυσμού μπορούμε να χρησιμοποιήσουμε τη δειγματική μέση απόκλιση s. Στην περίπτωση αυτή, αποδεικνύεται ότι:

$$\frac{\sqrt{n}(\overline{X} - \mu)}{s} \sim t(n-1),$$

και ανάλογα προκύπτει ότι το 95% διάστημα εμπιστοσύνης είναι το

$$\left(\overline{X} - \frac{s}{\sqrt{n}} t_{n-1,\, \alpha/2}, \overline{X} + \frac{s}{\sqrt{n}} t_{n-1,\, \alpha/2} \right)$$

Όπου t(n – 1) είναι η κατανομή Student με n – 1 βαθμούς ελευθερίας και $t_{n-1;\, \alpha/2}$, το αντίστοιχο ποσοστιαίο σημείο.

Ποσοστιαία σημεία t(n) κατανομής

Έστω $X \sim t(n)$, $0 < \alpha < 1$, και $t_{n,\alpha}$ ο αριθμός για τον οποίο

$$\mathbf{P(X > t_{n,\alpha}) = \alpha \text{ ή } P(X < t_{n,\alpha}) = 1 - \alpha.}$$

Το $t_{n,\alpha}$ ονομάζεται **α – ποσοστιαίο σημείο της t(n) κατανομής**. Οι ακριβείς τιμές $\mathbf{t_{n,\alpha}}$ μπορούν να βρεθούν από πίνακες είτε με χρήση υπολογιστή.

Αν n > 30, τότε αρκεί η χρήση της κανονικής κατανομής γιατί αποδεικνύεται ότι:

$$\mathbf{\text{αν } n > 30, \text{ τότε } t_{n,\alpha} \approx z_{\alpha}.}$$

Σημείωση
Μία απόδειξη για τον ισχυρισμό $t_{n,\alpha} \approx z_{\alpha}$ είναι διαθέσιμη εδώ:
https://math.stackexchange.com/questions/3240536/convergence-of-students-t-distribution-to-a-standard-normal

T-Table Distribution

df/alpha	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.706200	31.820520	63.656740	636.619200
2	0.288675	0.816497	1.885618	2.919986	4.302650	6.964560	9.924840	31.599100
3	0.276671	0.764892	1.637744	2.353363	3.182450	4.540700	5.840910	12.924000
4	0.270722	0.740697	1.533206	2.131847	2.776450	3.746950	4.604090	8.610300
5	0.267181	0.726687	1.475884	2.015048	2.570580	3.364930	4.032140	6.868800
6	0.264835	0.717558	1.439756	1.943180	2.446910	3.142670	3.707430	5.958800
7	0.263167	0.711142	1.414924	1.894579	2.364620	2.997950	3.499480	5.407900
8	0.261921	0.706387	1.396815	1.859548	2.306000	2.896460	3.355390	5.041300
9	0.260955	0.702722	1.383029	1.833113	2.262160	2.821440	3.249840	4.780900
10	0.260185	0.699812	1.372184	1.812461	2.228140	2.763770	3.169270	4.586900
11	0.259556	0.697445	1.363430	1.795885	2.200990	2.718080	3.105810	4.437000
12	0.259033	0.695483	1.356217	1.782288	2.178810	2.681000	3.054540	4.317800
13	0.258591	0.693829	1.350171	1.770933	2.160370	2.650310	3.012280	4.220800
14	0.258213	0.692417	1.345030	1.761310	2.144790	2.624490	2.976840	4.140500
15	0.257885	0.691197	1.340606	1.753050	2.131450	2.602480	2.946710	4.072800
16	0.257599	0.690132	1.336757	1.745884	2.119910	2.583490	2.920780	4.015000
17	0.257347	0.689195	1.333379	1.739607	2.109820	2.566930	2.898230	3.965100
18	0.257123	0.688364	1.330391	1.734064	2.100920	2.552380	2.878440	3.921600
19	0.256923	0.687621	1.327728	1.729133	2.093020	2.539480	2.860930	3.883400
20	0.256743	0.686954	1.325341	1.724718	2.085960	2.527980	2.845340	3.849500
21	0.256580	0.686352	1.323188	1.720743	2.079610	2.517650	2.831360	3.819300
22	0.256432	0.685805	1.321237	1.717144	2.073870	2.508320	2.818760	3.792100
23	0.256297	0.685306	1.319460	1.713872	2.068660	2.499870	2.807340	3.767600
24	0.256173	0.684850	1.317836	1.710882	2.063900	2.492160	2.796940	3.745400
25	0.256060	0.684430	1.316345	1.708141	2.059540	2.485110	2.787440	3.725100
26	0.255955	0.684043	1.314972	1.705618	2.055530	2.478630	2.778710	3.706600
27	0.255858	0.683685	1.313703	1.703288	2.051830	2.472660	2.770680	3.689600
28	0.255768	0.683353	1.312527	1.701131	2.048410	2.467140	2.763260	3.673900
29	0.255684	0.683044	1.311434	1.699127	2.045230	2.462020	2.756390	3.659400
30	0.255605	0.682756	1.310415	1.697261	2.042270	2.457260	2.750000	3.646000

Διάστημα Εμπιστοσύνης

Άσκηση 2

Τον Οκτώβριο του 2008, εξετάστηκαν 20 νεογέννητα βρέφη στις Ηνωμένες Πολιτείες ως προς το πλήθος βιομηχανικών ενώσεων, ρύπων και άλλων χημικών ουσίες που συνδέονται με την τοξικότητα του εγκεφάλου και του νευρικού συστήματος, την τοξικότητα του ανοσοποιητικού συστήματος, την αναπαραγωγική τοξικότητα και προβλήματα γονιμότητας. Συνολικά εξετάστηκαν 430 ουσίες. Στο αίμα ομφάλιου λώρου των 20 βρεφών βρέθηκαν οι εξής ουσίες (σε πλήθος):

79 145 147 160 116 100 159 151 156 126 137 83 156 94 121 144 123 114 139 99

Να βρεθεί 95% διάστημα εμπιστοσύνης για το μέσο πλήθος ουσιών του συνόλου των νεογέννητων βρεφών.

Δίνεται ότι η μέση τιμή του δείγματος είναι 127,45, η τυπ. απόκλιση s = 25,965 και ότι $t_{19;0.025} = 2,093$

$n = 20 < 30 \Rightarrow 95\%\ \Delta.E. \left[\bar{x} - t_{19;0.025}\cdot\frac{s}{\sqrt{n}}\ ,\ \bar{x} + t_{19;0.025}\cdot\frac{s}{\sqrt{n}}\right]$

$\bar{x} = \frac{79+145+\cdots+99}{20} = 127.45 \Rightarrow \left[127.45 - 2.093\cdot\frac{25.965}{\sqrt{20}}\ ,\ 127.45 + 2.093\frac{25.965}{\sqrt{20}}\right]$

$\Rightarrow [121.64,\ 133.26]$

$s^2 = \frac{1}{19}\left[(79-127.45)^2 + \cdots + (99-127.45)^2\right]\ ,\ s = \sqrt{s^2} = 25.965$

Διάστημα Εμπιστοσύνης

Άσκηση 3

Δεκαπέντε εθελοντές συμμετείχαν σε μία μελέτη για την ανακούφιση του πόνου που οφείλεται στο βελονισμό, Τα αποτελέσματα ήταν τα εξής:

8,6; 9,4; 7,9; 6,8; 8,3; 7,3; 9,2; 9,6; 8,7; 11,4; 10,3; 5,4; 8,1; 5,5; 6,9

Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για τη μέση ανακούφιση που θα δηλωθεί στο σύνολο του πληθυσμού στον οποίο θα εφαρμοστεί αυτή η μέθοδος.

Δίνεται ότι η μέση τιμή του δείγματος είναι 8,23, η τυπ, απόκλιση s = 1,67 και ότι $t_{14;0,025} = 2,140$

$$95\%\ \Delta.\Sigma. \left[\bar{x} - t_{14;0.025}\cdot\frac{s}{\sqrt{n}},\ \bar{x} + t_{14;0.025}\cdot\frac{s}{\sqrt{n}}\right]$$

$$\text{η } \left[8.23 - 2.140\cdot\frac{1.67}{\sqrt{15}},\ 8.23 + 2.140\cdot\frac{1.67}{\sqrt{15}}\right]$$

$$\text{η } [7.31,\ 9.15]$$

$\frac{X_1 + X_2 + \cdots + X_n}{n}$ $\bar{X}$: μέσος όρος

Διάστημα Εμπιστοσύνης για Ποσοστό

Διάστημα Εμπιστοσύνης για Ποσοστό

Έστω ότι ένα πείραμα διεξάγεται με πιθανότητα επιτυχίας p την οποία δεν γνωρίζουμε. Εκτελούμε n φορές το πείραμα και βρίσκουμε X επιτυχίες. Μία σημειακή εκτίμηση του άγνωστου ποσοστού είναι

$$\hat{p} = \frac{X}{n}$$

Γνωρίζουμε όμως, ότι αν X = {πλήθος από επιτυχίες στις n επαναλήψεις ενός πειράματος} τότε $X \sim B(n,p)$. Επιπλέον, αν $n > 30$, $np > 5$, $nq > 5$, τότε $X \sim N(np, npq)$ ή $X/n \sim N(p, pq/n)$ ή

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right)$$

Το τελευταίο μας δίνει τη δυνατότητα να υπολογίσουμε με ανάλογο τρόπο, το α – διάστημα εμπιστοσύνης για το άγνωστο ποσοστό.

$$\hat{p} \sim N\left(p, \frac{pq}{n}\right) \Leftrightarrow \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{pq}} \sim N(0,1).$$

Αν $\alpha > 0$, θεωρούμε τον αριθμό $z_{\alpha/2}$, για τον οποίο $P(Z > z_{\alpha/2}) = \alpha/2$. Τότε, θα είναι:

$$P\left(-z_{\alpha/2} < \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{pq}} < z_{\alpha/2}\right) = 1-\alpha \Leftrightarrow P\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2} < p < \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2}\right) = 1-\alpha.$$

Διάστημα Εμπιστοσύνης για Ποσοστό

Το διάστημα $\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2},\ \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2}\right)$ ονομάζεται (1 – α) – διάστημα εμπιστοσύνης για το άγνωστο ποσοστό του πληθυσμού. Δηλαδή:

Εάν η διαδικασία λήψης δείγματος μεγέθους n, επαναληφθεί πολλές φορές, τότε σε κάθε 100 διαστήματα της μορφής

$$\left(\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2},\ \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2}\right)$$

στα (1 – α)·100 από αυτά θα βρίσκεται μέσα το πραγματικό ποσοστό p του πληθυσμού.

Οι δύο περιπτώσεις που εμφανίζονται συνήθως είναι αυτές του 95% και του 99% δ.ε.

$$95\%\ \delta.\varepsilon.: \left(\hat{p} - 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$99\%\ \delta.\varepsilon.: \left(\hat{p} - 2{,}58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 2{,}58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Διάστημα Εμπιστοσύνης για Ποσοστό

Άσκηση 3

Μία εταιρεία έρευνας αγοράς θέλει να υπολογίσει το ποσοστό των ενηλίκων που ζουν σε μια μεγάλη πόλη και έχουν κινητά τηλέφωνα. Πεντακόσιοι τυχαία επιλεγμένοι ενήλικες κάτοικοι αυτής της πόλης ερευνώνται για να διαπιστωθεί εάν έχουν κινητά τηλέφωνα. Από τα 500 άτομα που συμμετείχαν στην έρευνα, τα 421 απάντησαν ναι - είναι κάτοχοι κινητών τηλεφώνων. Να βρεθεί ένα 95% διάστημα εμπιστοσύνης για την πραγματική αναλογία των ενηλίκων κατοίκων αυτής της πόλης που έχουν κινητά τηλέφωνα.

$n = 500 \quad \hat{p} = \frac{421}{500} = 0,842$

95% Δ.Ε. $\left(\hat{p} - 1,96 \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{500}} \; , \; \hat{p} + 1,96 \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{500}} \right)$

$\approx \quad (0.81 \; , \; 0.874)$

Διάστημα Εμπιστοσύνης για Ποσοστό

Άσκηση 4

Σε δείγμα 250 τυχαία επιλεγμένων ατόμων, οι 98 ανέφεραν ότι κατέχουν tablet. Χρησιμοποιώντας ένα επίπεδο εμπιστοσύνης 95%, υπολογίστε ένα διάστημα εμπιστοσύνης για το πραγματικό ποσοστό των ατόμων που κατέχουν tablet.

$n = 250, \quad \hat{p} = \frac{98}{250} = 0.394$

95% Δ.Ε. $\left[\hat{p} - 1.96 \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{250}} \; , \; \hat{p} + 1.96 \sqrt{\frac{\hat{p}\cdot(1-\hat{p})}{250}} \right]$

ή $[0.333, 0.455]$.

Μέγεθος Δείγματος

Μέγεθος Δείγματος για εκτίμηση μέσης τιμής

Παρατηρούμε ότι το πιθανοθεωρητικό σφάλμα του διαστήματος εμπιστοσύνης

$$\left(\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2},\ \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right) \qquad \overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \mu \le \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$$

είναι ίσο με $E = \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$. Από τον τύπο αυτό, αν θεωρήσουμε δεδομένο το σφάλμα E,

μπορούμε να λύσουμε ως προς το μέγεθος του δείγματος και να βρούμε

$$n = \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2 .$$

Ο τύπος αυτός μπορεί να χρησιμοποιηθεί για τον εντοπισμό του μεγέθους δείγματος που χρειάζεται ώστε το μέγιστο σφάλμα μεταξύ του αριθμητικού μέσου του δείγματος και της (άγνωστης) μέσης τιμής να μην ξεπερνάει το E με (1 – α)% πιθανότητα (όπως αυτή εκφράζεται με το (1 – α) δ.ε.).

Μέγεθος Δείγματος για εκτίμηση αναλογίας

Αντίστοιχα, αν γίνεται προσπάθεια εκτίμησης αναλογίας στον πληθυσμό, υπολογίζουμε

$$E = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_{\alpha/2} \Leftrightarrow n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}).$$

Καθώς, για $x \in [0, 1]$, είναι $x(1-x) \leq \frac{1}{4}$, συμπεραίνουμε ότι $n \leq \left(\frac{z_{\alpha/2}}{2E}\right)^2$.

Συνεπώς, αν το μέγιστο σφάλμα εκτίμησης είναι E:
(α) Αν υπάρχει μία αξιόπιστη εκτίμηση σχετικά την άγνωστη τιμή της αναλογίας, τότε μπορούμε να υπολογίσουμε το μέγεθος δείγματος από τον τύπο

$$n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}).$$

(β) Αν δεν υπάρχει αξιόπιστη εκτίμηση σχετικά με την άγνωστη αναλογία, τότε η επιλέγεται η πιο συντηρητική επιλογή:

$$n = \left(\frac{z_{\alpha/2}}{2E}\right)^2.$$

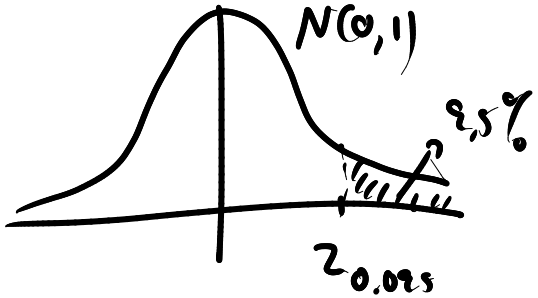
Μέγεθος Δείγματος

Άσκηση 5

Θέλουμε να εκτιμήσουμε το ποσοστό καπνιστών στους κατοίκους της πόλης της Ξάνθης. Πόσο μεγάλο πρέπει να είναι το δείγμα ώστε το σφάλμα να μην ξεπερνάει το 4% με πιθανότητα 95%;

$n = ?$ $\quad E = 0.04$, $\alpha = 0.05$

$$n = \left(\frac{z_{0.025}}{2 \cdot E} \right)^2 = \left(\frac{1.96}{2 \cdot 0.04} \right)^2 = 600,25.$$



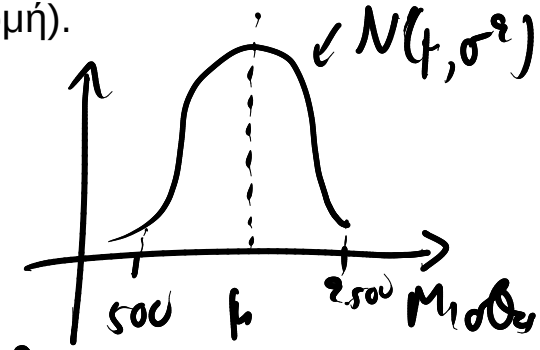
$z_{0.025} = 1.96.$

Μέγεθος Δείγματος

Άσκηση 6

Ένας οικονομολόγος επιθυμεί να εκτιμήσει, με εμπιστοσύνη 95%, το ετήσιο εισόδημα των ιδιωτικών υπαλλήλων του Νομού Ξάνθης με μέγιστο σφάλμα 200 ευρώ. Ποιο πρέπει να είναι το μέγεθος του δείγματος που πρέπει να μετρήσει, αν γνωρίζει ότι ο ελάχιστος μισθός είναι 500 ευρώ και ο μέγιστος 2500; (Υποθέτουμε ότι ο μισθός των υπαλλήλων ακολουθεί κανονική κατανομή).

$E = 200 \quad n = ; \quad Min = 500 \;,\; Max = 2.500$



$$n = \left(\frac{z_{0.025} \cdot \sigma}{E}\right)^2 = \left(\frac{1.96 \cdot \sigma}{200}\right)^2 = \left(\frac{1.96 \cdot 333.3}{200}\right)^2 = 10,7 \approx \underline{\underline{11}}$$

$$\sigma \approx \frac{Max - Min}{6} = \frac{2.500 - 500}{6} = \frac{2.000}{6} = 333,3$$

Βήματα μιας στατιστικής έρευνας

Βήματα μιας στατιστικής έρευνας

Περιληπτικά, τα βασικά βήματα μιας στατιστικής έρευνας είναι τα παρακάτω (Neil, Kelly & McNeil, 1978):

1. Προσδιορισμός και περιγραφή πληθυσμού ο οποίος μελετάται.

2. Δήλωση της **ερευνητικής υπόθεσης H_1**,. Ενδεικτικά, όταν αυτή αφορά ποσοτικές μεταβλητές η ερευνητική υπόθεση είναι ο ισχυρισμός πως υπάρχει διαφοροποίηση στις μέσες τιμές ή στις διασπορές ανάμεσα σε δύο ή περισσοτέρους πληθυσμούς και όταν αφορά ποιοτικές μεταβλητές είναι ο ισχυρισμός της εξάρτησης μεταξύ δύο ποιοτικών μεταβλητών.

3. Δήλωση της **στατιστικής υπόθεσης** η οποία συμβολίζεται με **H_0** και δηλώνει ισότητα ποσοτικών μεταβλητών ή ανεξαρτησία ποιοτικών μεταβλητών.

4. Προσδιορισμός του κατάλληλου στατιστικού ελέγχου που απαιτείται για την επιστημονική υποστήριξη ή αναίρεση της ερευνητικής υπόθεσης.

5. **Ορισμός του α** – της πιθανότητας ο ερευνητής να απορρίψει μια αληθής στατιστική υπόθεση.

6. Συλλογή των δεδομένων από ένα αντιπροσωπευτικό δείγμα του πληθυσμού.

7. Εφαρμογή του κατάλληλου στατιστικού ελέγχου και ερμηνεία των αποτελεσμάτων.

McNeil, K. A., Kelly F. J. & McNeil, J. T. (1978) Testing Research Hypotheses Using Multiple Linear Regression, Southern Illinois University Press.

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Κάθε μία στατιστική δοκιμασία συνδέεται με μία στατιστική (ή μηδενική υπόθεση) H_0, και με μία εναλλακτική (ή ερευνητική) H_1.

Σύμφωνα με το πρωτόκολλο που έχει καθιερωθεί στην επιστημονική κοινότητα ως προς την διεξαγωγή μίας στατιστικής έρευνας, η υπόθεση H_0 εκφράζεται ως μία αλγεβρική ισότητα (όταν αφορά συνεχείς μεταβλητές) ή σαν δήλωση στοχαστικής ανεξαρτησίας (όταν αφορά ζεύγη ποιοτικών μεταβλητών) ενώ η υπόθεση H_1 εκφράζεται ως η αντίστοιχη μονόπλευρη ή δίπλευρη αλγεβρική διαφοροποίηση (για συνεχείς μεταβλητές) ή η στοχαστική εξάρτηση (για ζεύγη ποιοτικών μεταβλητών).

Μετά την λεκτική περιγραφή των δύο υποθέσεων H_0, H_1, ακολουθεί η συλλογή του δείγματος και ο υπολογισμός ενός κατάλληλου στατιστικού μεγέθους το οποίο είναι ανάλογο με τη διαφοροποίηση του δείγματος από τη μηδενική υπόθεση.

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Η διαφοροποίηση του δείγματος από την μηδενική υπόθεση κρίνεται ως *εξαιρετικά μεγάλη* όταν η πιθανότητα p να υπολογιστεί ένα στατιστικό με τιμή περισσότερο ακραία από αυτή που υπολογίστηκε στο δείγμα μας, υποθέτοντας πως η μηδενική υπόθεση είναι αληθής είναι *εξαιρετικά μικρή*. Στο σημείο αυτό ανακύπτει με φυσιολογικό τρόπο το ερώτημα:

Πότε μία πιθανότητα είναι «εξαιρετικά» μικρή;

$\binom{10}{9} = \frac{10!}{9!\cdot 1!} = 10$ $\binom{10}{8} = \frac{10!}{8!\cdot 2!} = \frac{10\cdot 9}{2} = 45$ $\binom{10}{7} = \frac{10!}{7!\cdot 3!} = \frac{10\cdot 9\cdot 8}{3\cdot 2} = 120$

Πότε μία πιθανότητα είναι «εξαιρετικά» μικρή;

Δραστηριότητα

Θέλετε να αγοράσετε ένα αμερόληπτο κέρμα (ίδια πιθανότητα να έρθει κορώνα και γράμματα). Ένας έμπορος ισχυρίζεται πως έχει να σας πουλήσει ένα τέτοιο κέρμα. Εσείς, πριν το αγοράσετε το ρίχνετε 10 φορές και παρατηρείτε το αποτέλεσμα. Πόσες φορές πρέπει να έρθει κορώνα ώστε να πειστείτε να το αγοράσετε;

Διατυπώνουμε τις υποθέσεις **H_0: $p_{κορώνα}$ = 0,5**, έναντι της: **H_1: $p_{κορώνα}$ ≠ 0,5**. Αν η H_0 δεν απορριφθεί τότε θα πειστούμε πως $p_{κορώνα}$ = 0,5.

Αν X = {πλήθος Κ στις 10 ρίψεις} και η H_0 ισχύει, θα είναι X ~ B(10, 0.5). Υπολογίζουμε:

P(X = 10) = $\binom{10}{10} \cdot 0,5^{10} \cdot (1-0,5)^0 \simeq 0,001$

P(X = 9) = $\binom{10}{9} \cdot 0,5^9 \cdot (1-0,5)^1 \simeq 0,01$

P(X = 8) = $\binom{10}{8} \cdot 0,5^8 \cdot (1-0,5)^2 \simeq 0,044$

P(X = 7) = $\binom{10}{7} \cdot 0,5^{10} \simeq 0,117$

P(X = 6) = $\binom{10}{6} \cdot 0,5^{10} \simeq 0,205$

$\binom{10}{6} = \frac{10!}{6!\,4!} = \frac{10\cdot 9\cdot 8\cdot 7}{4\cdot 3\cdot 2} = 210$

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Το όριο απόρριψης της στατιστικής υπόθεσης συμβολίζεται με **a** ή **α** και ονομάζεται **σφάλμα τύπου Ι.**

Στην πλειοψηφία των δημοσιευμένων εργασιών επιλέγεται όριο απόρριψης a για τη μηδενική (στατιστική) υπόθεση το 0,05.

Ο βασικότερος λόγος για αυτήν την επιλογή του ορίου απόρριψης της H_0 στο 0,05, φαίνεται να είναι πως αυτή η τιμή επιλέχθηκε από τον Ronald Fisher (1890 – 1962) στις πρώτες δημοσιεύσεις του το 1925 (Fisher, 1925, σελ 47).

Ο Fisher, σχετικά με την απόφασή του να επιλέξει το λόγο 1/20 (= 0,05) ως όριο απόρριψης επιχειρηματολογεί ως εξής:

«It is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant».

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Από το κείμενο του προκύπτει και η γεωμετρική σημασία της συγκεκριμένης επιλογής: στην κανονική κατανομή το ποσοστό των παρατηρήσεων που βρίσκεται σε απόσταση μεγαλύτερη από 2 τυπικές αποκλίσεις μακρύτερα από τη μέση τιμή είναι περίπου ίσο με 5% (= 0,05).

Αργότερα, πρότεινε και τα ποσοστά 0,01 και 0,02 ως όρια απόρριψης που κατά περίπτωση χρησιμοποιούνται από τους ερευνητές σήμερα (Fisher 1926, σελ. 504):

«If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance».

Fisher, R. A. (1925) Statistical Methods for Research Workers, ISBN 0-05-002170-2.
Fisher, R. A. (1926) The arrangement of field experiments, Journal of the Ministry of Agriculture of Great Britain, 33: 503-513

Αποκτώντας ένα ερευνητικό αποτέλεσμα από μία πιθανότητα

Σήμερα, στους περισσότερους επιστημονικούς τομείς τα παραπάνω όρια εξακολουθούν να είναι κοινώς αποδεκτά.

Ωστόσο, σε κάποιες εφαρμογές της στατιστικής είτε λόγω του εξαιρετικά μεγάλου δείγματος (πειραματική φυσική) είτε λόγω του μεγάλου πλήθους παράλληλων ελέγχων (γονιδιακή έρευνα) έχουν επικρατήσει μικρότερα όρια απόρριψης.

Ενδεικτικές αναφορές: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC170937/

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Παράδειγμα

Ένα έντομο εμφανίζεται σε τρεις χρωματισμούς: Κίτρινο, Πράσινο και Κόκκινο.

Ένας βιολόγος ισχυρίζεται πως στον πληθυσμό του εντόμου, ισχύει ότι:

Κίτρινο: 50% Πράσινο: 25% Κόκκινο: 25%

Ισοδύναμα: H_0: $p_{κίτρινων} = 0,5$, $p_{πράσινων} = 0,25$, $p_{κόκκινων} = 0,25$,

Ο βιολόγος σκοπεύει να συλλέξει δείγμα 20 εντόμων και από την παρατήρηση του δείγματος να ελέγξει αν απορρίπτεται η H_0 έναντι της (ερευνητικής) υπόθεσης:

H_1: όχι η H_0.

(α) Προσομοίωση για τα πιθανά αποτελέσματα της δειγματοληψίας.
(β) Διερεύνηση του στατιστικού που αντανακλά την “απόσταση” του δείγματος από την υπόθεση H_0.

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Αν $p_{κίτρινων}$, $p_{πράσινων}$, $p_{κόκκινων}$ τα ποσοστό στον πληθυσμό των κίτρινων, πράσινων και κόκκινων εντόμων τότε επιθυμούμε να βρούμε αν από τα δεδομένα του δείγματος απορρίπτεται ή όχι η στατιστική υπόθεση:

$$H_0: p_{κίτρινων} = 0,5,\ p_{πράσινων} = 0,25,\ p_{κόκκινων} = 0,25,$$

έναντι της ερευνητικής υπόθεσης:

$$H_1: \text{όχι η } H_0.$$

Στο δείγμα των 20 εντόμων ο βιολόγος βρήκε 8 κίτρινα, 7 Πράσινα και 5 Κόκκινα

	Χρώμα			
	Κίτρινο	Πράσινο	Κόκκινο	Σύνολο
Υπόθεση H_0	10	5	5	20
Δείγμα	8	7	5	20
Διαφορά	-2	2	0	0

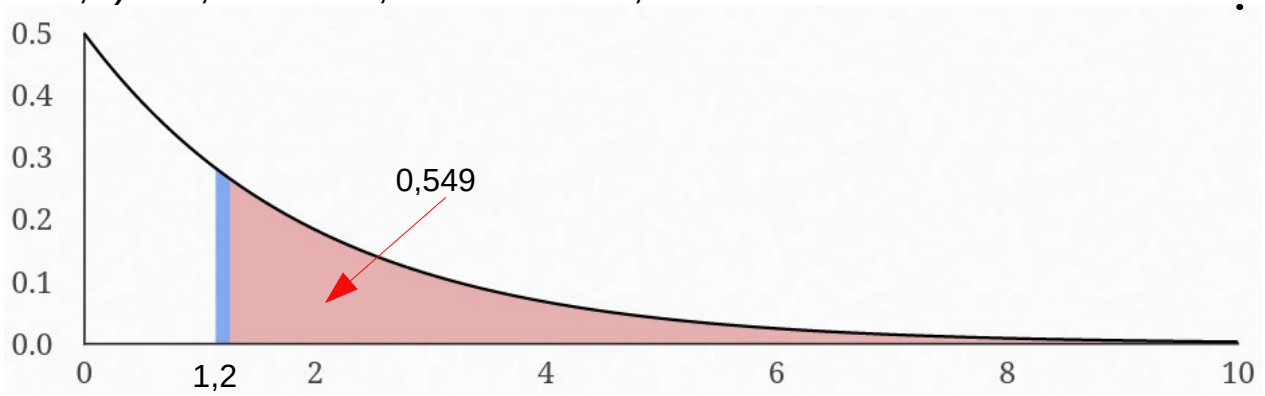
$X \sim N(\lambda, \lambda) \Rightarrow \frac{X-\lambda}{\sqrt{\lambda}} \sim N(0,1)$ $\lambda = p_i \cdot N$ $\frac{X - p_i \cdot N}{\sqrt{p_i \cdot N}} \sim N(0,1) \Rightarrow \frac{(X - p_i \cdot N)^2}{p_i \cdot N} \sim \chi^2(1)$

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Χρώμα	Κίτρινο	Πράσινο	Κόκκινο	Σύνολο
Υπόθεση H_0	10	5	5	20
Δείγμα	8	7	5	20
Διαφορά	-2	2	0	

$$\chi_0^2 = \sum_{i=1}^{3} \frac{(n_i - p_{i,0} \cdot N)^2}{p_{i,0} \cdot N} = \sum_{i=1}^{3} \frac{\delta_i^2}{p_{i,0} \cdot N} = \frac{(-2)^2}{10} + \frac{2^2}{5} + \frac{0^2}{5} = 0{,}4 + 0{,}8 + 0 = 1{,}2$$

$p = P(\chi^2 > \chi^2_0) = P(\chi^2 > 1{,}2) = 0{,}549 = 54{,}9\% > 5\% = 0{,}05$



Η υπόθεση H_0 δεν απορρίπτεται σε επίπεδο σημαντικότητας 0,05 ($\chi^2(2) = 1{,}2$, $p = 0{,}549$).

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Έστω ότι υπάρχει μία ποιοτική μεταβλητή με k επίπεδα τιμών και έστω, p_i το ποσοστό εμφάνισης στον πληθυσμό της τιμής i, i = 1, 2, …., k. Λαμβάνουμε ένα δείγμα μεγέθους N και βρίσκουμε τις συχνότητες n_1, n_2, …, n_k, ($n_1 + n_2 + \ldots + n_k = N$).

Για να ελέγξουμε τη στατιστική υπόθεση **H_0: $p_1 = p_{1,0}$, $p_2 = p_{2,0}$, …, $p_k = p_{k,0}$,** έναντι της **H_1: όχι η H_0.**

(α) Συμπληρώνουμε τον πίνακα

Συχνότητες \ Τιμή	1	2	...	k
Αναμενόμενες σύμφωνα με την H_0	$p_{1,0}\cdot N$	$p_{2,0}\cdot N$	...	$p_{k,0}\cdot N$
Παρατηρούμενες στο Δείγμα	n_1	n_2	...	n_k

(β) Υπολογίζουμε το στατιστικό $\chi_0^2 = \sum_{i=1}^{k} \frac{(n_i - p_{i,0}\cdot N)^2}{p_{i,0}\cdot N} = \sum_{i=1}^{k} \frac{\delta_i^2}{p_{i,0}\cdot N}$.

(γ) Βρίσκουμε την πιθανότητα $p = P(\chi^2 > \chi_0^2)$.

- Αν $p < 0{,}05$ λέμε (και γράφουμε) ότι η H_0 απορρίπτεται ($\chi^2(k-1) = \chi_0^2$, p = …)
- Αν $p \geq 0{,}05$ λέμε (και γράφουμε) ότι η H_0 δεν απορρίπτεται ($\chi^2(k-1) = \chi_0^2$, p = …)

(δ) Το όριο απόρριψης α = 0,05 ορίζεται στην αρχή της διαδικασίας.

⊗ $p = 0{,}018 < 0{,}05 \Rightarrow$ η H_0 απορρίπτεται.

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 1

Ένα ζάρι ρίχνεται 120 φορές και έρχονται τα αποτελέσματα του παρακάτω πίνακα. Να βρείτε αν τα αποτελέσματα διαφέρουν σημαντικά από αυτά που θα περιμέναμε να είχαμε αν σε ένα αμερόληπτο ζάρι.

Αποτέλεσμα	1	2	3	4	5	6	Σύνολο
Συχνότητα	15	29	16	15	30	15	120
Αναμενόμενες	20	20	20	20	20	20	120
Διαφορά	5	-9	4	5	-10	5	

$H_0: p_1 = p_2 = \dots = p_6 = \frac{1}{6}$

$H_1:$ όχι η H_0.

$$X_0^2 = \frac{5^2}{20} + \frac{(-9)^2}{20} + \frac{4^2}{20} + \frac{5^2}{20} + \frac{(-10)^2}{20} + \frac{5^2}{20} =$$

$$= \frac{272}{20} = 13{,}6 \quad X_0^2 \sim \chi^2(5) \Rightarrow p = P(X_0^2 > 13{,}6) = 0.018$$

⊗

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 1

Σύμφωνα με μία θεωρία κληρονομικότητας, ύστερα από διασταύρωση ενός είδους ζώων, πρέπει να προκύπτουν απόγονοι που να ανήκουν στα είδη Α, Β, Γ σε αναλογία 9:3:4. Σε ένα σχετικό πείραμα, από 64 απογόνους που προέκυψαν, 34 βρέθηκαν να είναι τύπου Α, 10 τύπου Β, και 20 τύπου Γ. Να βρεθεί αν απορρίπτεται η θεωρία σε επίπεδο σημαντικότητας 1%.

	9/16	3/16	4/16	
Τύπος	Α	Β	Γ	Σύνολο
Παρατηρούμενη	34	10	20	64
Αναμενόμενη	36	12	16	64
Διαφορά	-2	-2	4	-

$$X_0^2 = \frac{2^2}{36} + \frac{2^2}{12} + \frac{(-4)^2}{16} =$$

$$= \frac{1}{9} + \frac{1}{3} + 1 = \frac{13}{9} = 1,44.$$

$$X_0^2 \sim \chi^2(2)$$

$p = P(\chi^2 > 1.44) = 0.487 > 0,05 \Rightarrow$ η $H_0: p_A = \frac{9}{16}, p_B = \frac{3}{16}, p_\Gamma = \frac{4}{16}$ Δεν απορρίπτ.

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 2

Από τα δεδομένα της Στατιστικής Υπηρεσίας γνωρίζουμε ότι το 42% του πληθυσμού είναι παντρεμένοι/ες, το 38% ελεύθεροι/ες και το 20% διαζευγμένοι/χήροι. Σε ένα αντιπροσωπευτικό δείγμα 400 ατόμων στην πόλη της Ξάνθης, βρέθηκαν 250 παντρεμένοι, 120 ελεύθεροι και 30 διαζευγμένοι/χήροι. Να βρείτε αν η οικογενειακή κατάσταση των κατοίκων της Ξάνθης διαφέρει σημαντικά σε σχέση με το σύνολο του πληθυσμού (α = 0,05).

$H_0: p_{\Pi} = 0.42,\ p_{E} = 0.38,\ p_{\Delta} = 0.2$

Οικογ. Κατ.	Παντρ.	Ελευθ.	Δ. ή Χ	Σύνολο
Observed	250	120	30	400
Expected	168	152	80	406
Δ	-82	32	50	-

$p = P(\chi^2 > 78.01) < 0,001$

άρα ν H_0: απορρίπτεται

$\chi_0^2 = \frac{(-82)^2}{168} + \frac{32^2}{152} + \frac{50^2}{80} = 78,01.\quad \chi_0^2 \sim \chi^2(2)$

Δοκιμασία χι-τετράγωνο ως έλεγχος προσαρμογής

Άσκηση 3

Αξιοποιώντας τα δεδομένα του πίνακα σχετικά με τις γεννήσεις ανά μήνα, βρείτε αν οι γεννήσεις κατανέμονται ομοιόμορφα στους 12 μήνες του χρόνου

Μήνας	1	2	3	4	5	6	7	8	9	10	11	12
Πλήθος γεννήσεων	60	44	45	50	49	56	46	41	69	49	44	47

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Έστω δύο μεταβλητές με m και k κατηγορίες αντίστοιχα τις οποίες παρατηρήσαμε στα ίδια υποκείμενα. Θέλουμε να ελέγξουμε αν οι δύο μεταβλητές είναι στατιστικώς εξαρτημένες, δηλαδή αν η απόκριση στην μία συσχετίζεται με την απόκριση στην άλλη.

Η δοκιμασία χ^2 ως έλεγχος ανεξαρτησίας είναι ο τρόπος για να ελέγξουμε αν οι παρατηρούμενες συχνότητες $\{n_{ij}\}_{i=1,\ldots,m,\ j=1,\ldots,k}$, διαφέρουν σημαντικά από τις θεωρητικώς αναμενόμενες $\{n_{ij,\,0}\}_{i=1,\ldots,m,\ j=1,\ldots,k}$, κάτω από την υπόθεση της ανεξαρτησίας.

Το στατιστικό που ενσωματώνει το σύνολο των αποκλίσεων είναι το

$$\chi_0^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - n_{ij,\,0}\right)^2}{n_{ij,\,0}}$$

Αποδεικνύεται ότι $\chi^2 \sim \chi^2((m-1)\cdot(k-1))$. Η πιθανότητα με την οποία κρίνεται η απόρριψη ή μη της μηδενικής υπόθεσης υπολογίζεται ως $p = P(\chi^2 > \chi^2_0)$.

$P(X = \text{"Πολ. Αρν."}, Y = \text{"Απογ."}) = P(X = \text{"Πολ. Αρν."}) \cdot P(Y = \text{"Απογ."}) = \frac{16}{102} \cdot \frac{42}{102}$

Συχνότητες υπό την υπόθεση της ανεξαρτησίας

Καταγράφηκε η άποψη 102 φοιτητών για τον καθηγητή ενός μαθήματος σε κλίμακα 1 = Πολύ αρνητική έως 5 = Πολύ θετική. Επιπλέον, καταγράφηκε η επιτυχία τους στις τελικές εξετάσεις. Οι παρατηρήσεις συγκεντρώθηκαν στον παρακάτω διμεταβλητό πίνακα συχνοτήτων.

	X = Εντύπωση για τον καθηγητή					
Αποτέλεσμα Y	Πολύ αρνητική	Αρνητική	Ουδέτερη	Θετική	Πολύ θετική	Σύνολο
Αποτυχία	10	11	5	8	8	42
Επιτυχία	6	8	6	15	25	60
Σύνολο	16	19	11	23	33	102

Επιθυμούμε να βρούμε αν η εντύπωση συσχετίζεται με την επιτυχία στις εξετάσεις. Για να κρίνουμε την εξάρτηση μεταξύ των δύο μεταβλητών “Εντύπωση” και “Αποτέλεσμα” αρκεί να συγκρίνουμε τις παραπάνω συχνότητες με αυτές που θα περιμέναμε να είχαμε αν οι δύο μεταβλητές ήταν στοχαστικά ανεξάρτητες.

Συχνότητες υπό την υπόθεση της ανεξαρτησίας

Αν A_i = {Εντύπωση = i}, i = 1, 2, 3, 4, 5 και B_j = {Αποτέλεσμα = j}, j = 1 (Αποτυχία), 2 (Επιτυχία), τότε έστω

E_{ij} = {Αναμενόμενη συχνότητα Εντύπωση = i και Αποτέλεσμα = j}.

Είναι $E_{ij} = 102 \cdot P(A_i \cdot B_j)$.
Όμως, αν A_i, B_j στοχαστικά ανεξάρτητα τότε

$$P(A_i \cdot B_j) = P(A_i) \cdot P(B_j) = \frac{\text{Άθροισμα i στήλης}}{102} \cdot \frac{\text{Άθροισμα j γραμμής}}{102.}$$

Συμπεραίνουμε, ότι

$$E_{ij} = \frac{\text{Άθροισμα i στήλης} \cdot \text{Άθροισμα j γραμμής}}{102}.$$

Συχνότητες υπό την υπόθεση της ανεξαρτησίας

Μετά τις απαραίτητες πράξεις βρίσκουμε τις αναμενόμενες συχνότητες για κάθε ένα κελί:

Αποτέλεσμα	Πολύ αρνητική	Αρνητική	Ουδέτερη	Θετική	Πολύ θετική	Σύνολο
Αποτυχία	10 (6,6)	11 (7,8)	5 (4,5)	8 (9,5)	8 (13,6)	42
Επιτυχία	6 (9,4)	8 (11,2)	6 (6,5)	15 (13,5)	25 (19,4)	60
Σύνολο	16	19	11	23	33	102

$$\chi_0^2 = \frac{(10-6{,}6)^2}{6{,}6} + \dots + \frac{(25-19{,}4)^2}{19{,}4} \sim \chi^2\left((5-1)\cdot(2-1)\right)$$

$$\chi^2\left((k-1)\cdot(l-1)\right)$$

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

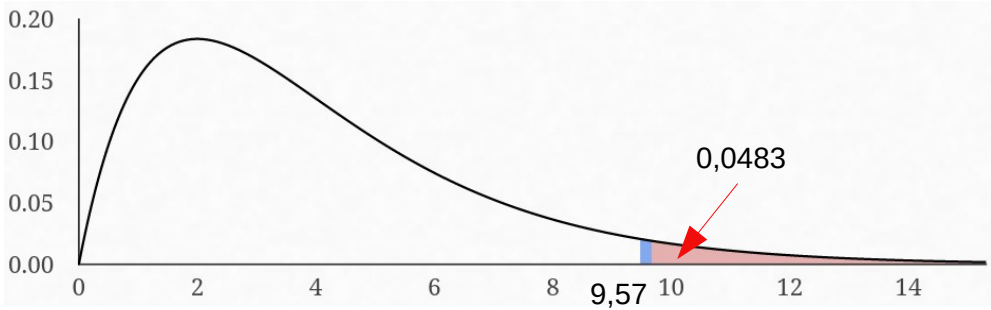
Υπολογίζουμε τις αναμενόμενες συχνότητες σε κάθε ένα κελί, αν η υπόθεση H_0 ήταν αληθής.

Αποτέλεσμα	Πολύ αρνητική	Αρνητική	Ουδέτερη	Θετική	Πολύ θετική	Σύνολο
Αποτυχία	10 (6,6)	11 (7,8)	5 (4,5)	8 (9,5)	8 (13,6)	42
Επιτυχία	6 (9,4)	8 (11,2)	6 (6,5)	15 (13,5)	25 (19,4)	60
Σύνολο	16	19	11	23	33	102

Είναι:

$$\chi_0^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{(n_{ij} - n_{ij,\,0})^2}{n_{ij,\,0}} = 9{,}57 \quad \text{και} \quad \chi^2 \sim \chi^2(4).$$

Υπολογίζουμε, $p = P(\chi^2 > \chi^2_0) = 0{,}0483 = 4{,}83\%$.



Συμπέρασμα

Η υπόθεση H_0 απορρίπτεται, δηλαδή, η γνώμη των φοιτητών είναι στατιστικά εξαρτημένη με την επιτυχία τους στις εξετάσεις σε επίπεδο σημαντικότητας 0,05 ($\chi^2(4) = 9{,}57$, $p = 0{,}048$).

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Υλοποίηση στην R

```
my.table <- matrix(c(10, 11, 5, 8, 8, 6, 8, 6, 15, 25), ncol=5, byrow=TRUE)
chisq.test(my.table)
```

Output

```
Pearson's Chi-squared test

data:  my.table
X-squared = 9.5743, df = 4, p-value = 0.04824

Warning message:
In stats::chisq.test(x, y, ...) :
  Chi-squared approximation may be incorrect
```

(εμφανίζεται γιατί υπάρχει αναμενόμενη τιμή < 5)

⊛ $p = P(\chi^2 > 5.395) = 0.070 > 0.05$ άρα η H_0 ΔΕΝ απορρίπτεται.

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Άσκηση 1

Τρία κέρματα Α, Β, Γ, ρίχνονται από 200 φορές και καταγράφεται το πλήθος Κ (Κορώνα) και Γ (Γράμματα). Χρησιμοποιώντας τα δεδομένα του πειράματος, αποφασίστε σε επίπεδο σημαντικότητας 0,05 αν τα κέρματα ρίχνονται Κ ή Γ με την ίδια πιθανότητα.

Κέρμα	Α	Β	Γ	Σύνολ.
Κορώνα	88 (97)	93 (97)	110 (97)	291
Γράμματα	112 (103)	107 (103)	90 (103)	309
Σύνολο	200	200	200	600

H_0: Κ και Γ με ίδια πιθανότητα στα 3 κέρματα.

$$\chi_0^2 = \frac{(88-97)^2}{97} + \frac{(93-97)^2}{97} + \frac{(110-97)^2}{97} + \frac{(112-103)^2}{103} + \frac{(107-103)^2}{103} + \frac{(90-103)^2}{103} =$$

$$= \frac{81+16+169}{97} + \frac{81+16+169}{103} = 5,395,\ \chi_0^2 \sim \chi^2((3-1)\cdot(2-1)) = \chi^2(2)$$ ⊛

Δοκιμασία χι-τετράγωνο ως έλεγχος ανεξαρτησίας

Άσκηση 2

Χρησιμοποιώντας τα παραπάνω δεδομένα ελέγξτε την υπόθεση πως οι άνδρες και οι γυναίκες έχουν ανάλογες προτιμήσεις στα χρώματα.

Χρώμα	Κόκκινο	Κίτρινο	Μπλε	Σύνολο
Άνδρες	21 (28,5)	34 (33,5)	45 (38)	100
Γυναίκες	36 (28,5)	33 (33,5)	31 (38)	100
Σύνολο	57	67	76	200

H_0: Χρώμα ανεξάρτητο από φύλο. H_1: όχι η H_0.

$$\chi_0^2 = \frac{7{,}5^2}{28{,}5} + \frac{0{,}5^2}{33{,}5} + \frac{7^2}{38} + \frac{7{,}5^2}{28{,}5} + \frac{0{,}5^2}{33{,}5} + \frac{7^2}{38} = 6{,}541$$

$\chi_0^2 \sim \chi^2(2) \Rightarrow p = P(\chi^2 > 6.541) = 0.04 < 0.05 \Rightarrow H_0$: απορρίπτεται.