

Στατιστική
Μέρος Πρώτο: Δειγματικές Κατανομές

Διδάσκουσα: Λουκία Μελιγκοτσίδου
Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών
Τμήμα Μαθηματικών

October 23, 2017

Εισαγωγή

[Τυχαίο Δείγμα]: μεγέθους n από έναν πληθυσμό X ή μια κατανομή, είναι n ανεξάρτητες και ισόνομες τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$.

[Ο]: Άν $X_1, X_2, \dots, X_n$ τυχαίο δείγμα, τότε

- δειγματικός μέσος:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- δειγματική διασπορά:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Η από κοινού συνάρτηση πυκνότητας ή πιθανότητας των $X_1, X_2, \dots, X_n$ είναι

$$f(X_1, X_2, \dots, X_n) = f(X_1)f(X_2)\dots f(X_n) = \prod_{i=1}^n f(X_i).$$

[Στατιστικό ή Στατιστική συνάρτηση]: Μια συνάρτηση των μεταβλητών $(X_1, X_2, \dots, X_n)$ ενός τυχαίου δείγματος η οποία δεν περιέχει άγνωστες παραμέτρους λέγεται *στατιστικό ή στατιστική συνάρτηση*. Π.Χ. τα αριθμητικά χαρακτηριστικά ενός δείγματος όπως ο δειγματικός μέσος $\bar{X}$, δειγματική διασπορά S^2 κ.λ.π. είναι όλα στατιστικές συναρτήσεις.

Τα στατιστικά συνήθως συμβολίζονται με

$$T = T(X) = T(X_1, X_2, \dots, X_n),$$

$$S = S(X) = S(X_1, X_2, \dots, X_n)$$

κ.λ.π. Ός συναρτήσεις τυχαίων μεταβλητών (τ.μ.) είναι και αυτά τ.μ. και παίζουν έναν διπλό ρόλο

$\left\{ \begin{array}{l} \text{από την μια μεριά είναι γνωστοί αριθμοί; σε συγκεκριμένο δείγμα κάθε φορά} \\ \text{από την άλλη είναι τυχαίες μεταβλητές.} \end{array} \right.$

Παίρνουν τις διάφορες τιμές τους ύστερα από νοερές επαναλήψεις της δειγματοληψίας, η οποία γεννά τις τιμές του δείγματος. Οι αριθμητικές τιμές συμβολίζονται με μικρά γράμματα $(\bar{x}, s^2, t)$, ενώ οι τ.μ. συμβολίζονται με κεφαλαία $(\bar{X}, S^2, T)$. Οι κατανομές των στατιστικών λέγονται *δειγματικές κατανομές*.

[Δειγματική Μεταβλητότητα]: Η μεταβλητοτητα (variability) μεταξύ τυχαίων δειγμάτων από τον ίδιο πληθυσμό ονομάζεται *δειγματική μεταβλητότητα* (sampling variability)

[Δειγματική Κατανομή]: Η κατανομή πιθανότητας που χαρακτηρίζει κάποια στοιχεία της δειγματικής μεταβλητότητας ονομάζεται *δειγματική κατανομή* (sampling distribution)

(*) Υποθέστε ότι ενδιαφερόμαστε να περιγράψουμε τα χαρακτηριστικά ενός στατιστικού με το να παίρνουμε συνεχώς δείγματα από συγκεκριμένο πληθυσμό και να τα περιγράφουμε με τον υπολογισμό στατιστικών συναρτήσεων, ιστογράμματος κλπ. Αυτή η κατανομή θα έχει σχήμα, μέσο, τυπική απόκλιση ...κλπ.

Κατανομές χ^2 , t_ν και F_{ν_1, ν_2}

Οι μονοδιάστατες αυτές κατανομές απορρέουν από την κανονική κατανομή και παίζουν πρωταρχικό ρόλο στην στατιστική συμπερασματολογία.

[Κατανομή χ^2]: Έστω $X_1, X_2, \dots, X_\nu$ ν ανεξάρτητες και ισόνομες τ.μ. με τυπική κανονική κατανομή $N(0, 1)$. Η κατανομή της τ.μ.

$$U = X_1^2 + X_2^2 + \dots + X_\nu^2$$

λέγεται χ^2 (χι-τετράγωνο) με ν -βαθμούς ελευθερίας (β.ε.). Συμβολικά

$$\chi_\nu^2 = \sum_1^\nu N(0, 1)^2.$$

Η χ^2 κατανομή είναι λοξή προς τα δεξιά, και καθώς οι β.ε. αυξάνουν η λοξότητα μειώνεται. Έχει

- $E(\chi_\nu^2) = \nu$
- $Var(\chi_\nu^2) = 2\nu$

και επιπλέον η χ_ν^2 προσεγγίζεται από την κανονική όταν $\nu \rightarrow \infty$. Οι συνήθεις πίνακες της χ_ν^2 δίνουντα αντίστροφα εκατοστιαία σημεία $\chi_{\alpha, \nu}^2$ για β.ε. $\nu < 30$. Τα σημεία αυτά ορίζονται από τη σχέση

$$P(\chi_\nu^2 \geq \chi_{\alpha, \nu}^2) = \alpha.$$

[Κατανομή t]: Έστω $X \sim N(0, 1)$ και $Y \sim \chi_\nu^2$, με X και Y ανεξάρτητες τ.μ. Η κατανομή της τ.μ.

$$T = \frac{X}{\sqrt{Y/\nu}}$$

λέγεται t -κατανομή με ν βαθμούς ελευθερίας. Συμβολικά

$$t_\nu = \frac{N(0, 1)}{\sqrt{\chi_\nu^2/\nu}}.$$

Η t κατανομή είναι συμμετρική γύρω από τον άξονα $X = 0$ και μοιάζει πολύ με την κανονική. Έχει

- $E(t_\nu) = 0$
- $Var(t_\nu) = \frac{\nu}{\nu-2}$

και επιπλέον $t_\infty = N(0, 1)$. Οι συνήθεις πίνακες της t δίνονται αντίστροφα εκατοστιαία σημεία $t_{\alpha, \nu}$ που ορίζονται από τη σχέση

$$P(T_\nu \geq t_{\alpha, \nu}) = \alpha.$$

Λόγω συμμετρίας $t_{1-\alpha, \nu} = -t_{\alpha, \nu}$.

[Κατανομή F]: Έστω $X_1 \sim \chi_{\nu_1}^2$ και $X_2 \sim \chi_{\nu_2}^2$, με X_1 και X_2 ανεξάρτητες τ.μ. Η κατανομή της τ.μ.

$$T = \frac{X_1/\nu_1}{X_2/\nu_2}$$

λέγεται F -κατανομή με ν_1, ν_2 βαθμούς ελευθερίας. Συμβολικά

$$F_{\nu_1, \nu_2} = \frac{\chi_{\nu_1}^2/\nu_1}{\chi_{\nu_2}^2/\nu_2}.$$

Η F κατανομή είναι ασυμμετρική. Έχει

- $E(F_{\nu_1, \nu_2}) = \frac{\nu_2}{\nu_2-2}, \quad \nu_2 > 2$
- $Var(F_{\nu_1, \nu_2}) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}, \quad \nu_2 > 4.$

Τα αντίστροφα εκατοστιαία σημεία της F ορίζονται από τη σχέση

$$P(F_{\nu_1, \nu_2} \geq F_{\alpha, \nu_1, \nu_2}) = \alpha.$$

Με βάση τους ορισμούς μπορούμε να δείξουμε ότι

$$t_\nu^2 = F_{1, \nu}.$$

Πράγματι, αν $X \sim N(0, 1)$ και $Y \sim \chi_\nu^2$ (ανεξάρτητες), τότε $T = \frac{X}{\sqrt{Y/\nu}} \sim t_\nu$.

Είναι $X^2 \sim \chi_1^2$, άρα $T^2 = \frac{X^2}{Y/\nu} \sim F_{1, \nu}$.

Δειγματοληψία από κανονικούς πληθυσμούς

[Θεώρημα]: Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα από ένα πληθυσμό X , με μέση τιμή μ και διασπορά σ^2 . Τότε

1. $E(\bar{X}) = \mu$
2. $Var(\bar{X}) = \frac{\sigma^2}{n}$
3. $E(S^2) = \sigma^2$.

[Απόδειξη]:

1. $E(\bar{X}) = E\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n E(X_i)}{n} = \frac{n\mu}{n} = \mu$.
2. $Var(\bar{X}) = Var\left[\frac{\sum_{i=1}^n X_i}{n}\right] = \frac{\sum_{i=1}^n Var(X_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$.
- 3.

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} E\left(\sum (X_i - \bar{X})^2\right) \\ &= \frac{1}{n-1} E\left(\sum (X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right) \\ &= \frac{1}{n-1} E\left(\sum (X_i^2) - \sum (2X_i\bar{X}) + n\bar{X}^2\right) \\ &= \frac{1}{n-1} E\left(\sum (X_i^2) - 2n\bar{X}^2 + n\bar{X}^2\right) \\ &= \frac{1}{n-1} \left(\sum E(X_i^2) - nE(\bar{X}^2)\right) \\ &= \frac{1}{n-1} (nE(X^2) - nE(\bar{X}^2)) \\ &= \frac{n}{n-1} (\sigma^2 + \mu^2 - Var(\bar{X}) - E(\bar{X})^2) \\ &= \frac{n}{n-1} \left(\sigma^2 - \frac{\sigma^2}{n}\right) \\ &= \frac{n}{n-1} \left(\frac{n-1}{n}\sigma^2\right) \\ &= \sigma^2. \end{aligned}$$

[Πόρισμα]: Έστω $(X_{11}, X_{12}, \dots, X_{1v_1})$ τυχαίο δείγμα μεγέθους v_1 από πληθυσμό με μέση τιμή μ_1 και διασπορά σ_1^2 , και $(X_{21}, X_{22}, \dots, X_{2v_2})$ τυχαίο δείγμα μεγέθους v_2 από πληθυσμό με μέση τιμή μ_2 και διασπορά σ_2^2 . Τότε

1. $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$
2. $Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{v_1} + \frac{\sigma_2^2}{v_2}$.

[Θεώρημα]: Αν $X_1, X_2, \dots, X_v$ είναι ανεξάρτητες τ.μ. που έχουν χ^2 κατανομή με β.ε. $\kappa_1, \kappa_2, \dots, \kappa_v$, τότε η τ.μ.

$$Y = \sum_{i=1}^v X_i$$

έχει χ^2 κατανομή με $\kappa_1 + \kappa_2 + \dots + \kappa_v$ β.ε.

[Θεώρημα]: Έστω $X_1, X_2, \dots, X_v$ τυχαίο δείγμα από ένα κανονικό πληθυσμό, με μέση τιμή μ και διασπορά σ^2 . Τότε

- $\bar{X} \sim N(\mu, \sigma^2/v)$
- τα στατιστικά $\bar{X}$ και S^2 είναι ανεξάρτητα
- η δειγματική κατανομή του

$$\frac{(v-1)S^2}{\sigma^2} \sim \chi_{v-1}^2$$

[Θεώρημα]: Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα από ένα κανονικό πληθυσμό, με μέση τιμή μ και διασπορά σ^2 . Τότε η στατιστική συνάρτηση $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ ακολουθεί t-κατανομή με $n-1$ β.ε. Συμβολικά

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

[Απόδειξη]: Γνωρίζουμε ότι

$$\bar{X} \sim N(\mu, \sigma^2/n) \quad \text{ή} \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

και

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Επιπλέον, οι μεταβλητές $\bar{X}$ και S^2 είναι στατιστικώς ανεξάρτητες. Έτσι έχουμε

$$\frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

[Θεώρημα]: Έστω οι διακυμάνσεις δύο τ.δ. μεγέθους n_1 και n_2 από δύο ανεξαρτητους κανονικούς πληθυσμούς με μέσες τιμές και διακυμάνσεις (μ_1, σ_1^2) και (μ_2, σ_2^2) αντίστοιχα. Τότε

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}.$$

[Απόδειξη]: Γνωρίζουμε ότι

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2 \quad \text{και} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi_{n_2-1}^2.$$

Εφαρμόζοντας τον ορισμό για την F-κατανομή έχουμε

$$\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n_1-1, n_2-1}.$$

[Κ.Ο.Θ.]: Έστω $X_1, X_2, \dots, X_n$ τυχαίο δείγμα από ένα άπειρο πληθυσμό, με μέση τιμή μ και διασπορά σ^2 . Τότε όταν $n \rightarrow \infty$ η κατανομή του

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

τείνει στην $N(0, 1)$. (Πρακτικά το εφαρμόζουμε για $n > 30$).

[Θεώρημα]: Άν $\bar{X}$ είναι ο δειγματικός μέσος από ένα τ.δ. μεγέθους n από κανονικά πληθυσμό με μέσο μ και διασπορά σ^2 , η δειγματική κατανομή του τότε είναι η κανονική κατανομή με μέσο μ και διασπορά σ^2/n . Δηλαδή,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{ή} \quad \sum_{i=1}^n X_i = n\bar{X} \sim N(n\mu, n\sigma^2)$$

[Παράδειγμα 1]: Έστω ένα μηχάνημα που βγάζει αναψυκτικά. Η ποσότητα αναψυκτικού που βγαίνει είναι μια τ.μ. με μέση τιμή 200ml και τυπική απόκλιση 15ml. Ποιά η πιθανότητα η μέση ποσότητα αναψυκτικού σε ένα τ.δ. μεγέθους 36 να είναι τουλάχιστον 204ml;

[Λύση]: Γνωρίζουμε ότι η κατανομή του $\bar{X}$ έχει μέση τιμή $\mu_{\bar{X}} = 200$ και $\sigma_{\bar{X}} = \frac{15}{\sqrt{36}} = 2.5$. Σύμφωνα με το Κ.Ο.Θ. η κατανομή θα είναι κατα προσέγγιση κανονική. Επομένως

$$P(\bar{X} \geq 204) = P\left(Z \geq \frac{204 - 200}{2.5}\right) = P(Z \geq 1.6) = 0.0548.$$

[Παράδειγμα 2]: Θεωρούμε ότι κατά την κατασκευή ενός εξαρτήματος για ένα μηχάνημα, το πάχος του είναι η πιο σημαντική του διάσταση. Η διαδικασία κατασκευής θεωρείται "υπό έλεγχο" εάν η πραγματική μεταβλητότητα για το πάχος του εξαρτήματος είναι $\sigma = 0.60$ χιλιοστά της ίντσας. Για να κάνουμε ελέγχους στη παραγωγή, τ.δ. μεγέθους $n=20$ λαμβάνονται περιοδικά και η παραγωγή κρίνεται "εκτός ελέγχου" αν η πιθανότητα το S^2 να πάρει τιμές μεγαλύτερες ή ίσες από την παρατηρούμενη δειγματική τιμή είναι 0.01 και κάτω. Τι κρίνουμε για την παραγωγή αν σε ένα δείγμα έχουμε $s = 0.84$ χιλιοστά της ίντσας?

[Λύση]: Η διαδικασία θα είναι "εκτός ελέγχου" αν $\frac{(n-1)s^2}{\sigma^2}$, με $n=20$ και $\sigma = 0.60$ ξεπερνά την τιμή $\chi^2_{0.01,19} = 36.191$. Οπότε

$$\frac{(n-1)s^2}{\sigma^2} = \frac{19(0.84)^2}{(0.60)^2} = 37.24$$

Συνεπώς, είναι "εκτός ελέγχου".

[Παράδειγμα 3]: Σε 16 1-ώρες ελέγχους, ένας κινητήρας κατανάλωσε κατα μέσο όρο 16.4 λίτρα βενζίνης με τυπική απόκλιση 2.1 λίτρα. Ελέγξτε την υπόθεση ότι η μέση κατανάλωση του κινητήρα είναι 12.0 λίτρα την ώρα.

[Λύση]: Έχουμε $n=16$, $\mu = 12.0$, $\bar{x} = 16.4$ και $s = 2.1$. Τότε

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{16.4 - 12.0}{2.1/\sqrt{16}} = 8.38.$$

Από την t-κατανομή με $n-1=15$ β.ε. βλέπουμε ότι για $T > 2.947$ η πιθανότητα είναι 0.005. Συνεπώς για τιμές μεγαλύτερες του 8 η πιθανότητα θα πρέπει να είναι αμελητέα. Συνεπώς συμπεραινουμε ότι η μέση κατανάλωση θα πρέπει να είναι μεγαλύτερη των 12 λίτρων/ώρα.

[Παράδειγμα 4]: Έστω 100 παρατηρήσεις από την $U(-1, 1)$, και έστω

$$S_{100} = X_1 + X_2 + \dots + X_{100}.$$

Να υπολογίσετε την πιθανότητα

$$P(|S_{100}| \leq 10). \quad [\Delta\text{νετα}\Phi(1.73) = 0.96]$$

[Λύση]: Γνωρίζουμε ότι:

$$\bar{X} \sim N(\mu, \sigma^2/\nu) \Rightarrow S_{100} \sim N(100\mu, 100\sigma^2),$$

όπου $\sigma^2 = \frac{(1-(-1))^2}{12} = \frac{1}{3}$ και $\mu = \frac{1+(-1)}{2} = 0$. Άρα

$$S_{100} \sim N\left(0, \frac{100}{3}\right).$$

Επομένως:

$$\begin{aligned} P(|S_{100}| \leq 10) &= P(-10 \leq S_{100} \leq 10) = \\ P\left(\frac{-10}{\sqrt{100/3}} \leq \frac{S_{100}}{\sqrt{100/3}} \leq \frac{10}{\sqrt{100/3}}\right) &= P\left(-\sqrt{3} \leq z \leq \sqrt{3}\right) = \\ \Phi(1.73) - \Phi(-1.73) &= 2\Phi(1.73) - 1 = \\ 2(0.96) - 1 &= 0.92. \end{aligned}$$