ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ Ι

Συμπληρωματικές Σημειώσεις

Δημήτριος Παντελής

ΣΤΑΤΙΣΤΙΚΕΣ ΕΚΤΙΜΗΣΕΙΣ

Οι συναρτήσεις πιθανότητας ή πυκνότητας πιθανότητας των διαφόρων τυχαίων μεταβλητών χαρακτηρίζονται από κάποιες παραμέτρους. Παραδείγματος χάριν, η μέση τιμή και η διασπορά (μ, σ^2) χαρακτηρίζουν τη συνάρτηση πυκνότητας πιθανότητας της κανονικής τυχαίας μεταβλητής, η μέση τιμή λ τη συνάρτηση πιθανότητας της Poisson, και η πιθανότητα επιτυχίας p τη συνάρτηση πιθανότητας της διωνυμικής. Όταν αυτές οι παράμετροι είναι άγνωστες, εκτιμώνται από δειγματοληπτικές μετρήσεις, π.χ. η μέση τιμή του βάρους μιας συσκευασίας εκτιμάται από το μέσο βάρος της συσκευασίας σε ένα πεπερασμένο δείγμα. Η ανάλυση που ακολουθεί υποθέτει ότι η δειγματοληψία γίνεται με τυχαίο τρόπο, δηλαδή οι μετρήσεις του δείγματος είναι ανεξάρτητες μεταξύ τους.

1. Σημειακή εκτίμηση

Έστω θ άγνωστη παράμετρος που χαρακτηρίζει μια τυχαία μεταβλητή X. Η παράμετρος αυτή εκτιμάται από τυχαίο δείγμα που αποτελείται από μετρήσεις $X_1, X_2, \ldots, X_n$. Οι τιμές του δείγματος είναι ανεξάρτητες τυχαίες μεταβλητές με συνάρτηση (πυκνότητας) πιθανότητας ίδια με της X. Από τις τιμές αυτές προκύπτει μια εκτιμήτρια συνάρτηση $\hat{\theta}$ η τιμή της οποίας δίνει μια εκτίμηση για την παράμετρο θ. Ακολουθούν ορισμένα παραδείγματα εκτιμητριών.

Μέση τιμή. Εκτιμάται από τη μέση τιμή του δείγματος $\bar{X}$, όπου

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

Διασπορά. Εκτιμάται από τη διασπορά του δείγματος S^2, όπου

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1} = \frac{X_1^2 + X_2^2 + \cdots + X_n^2 - n\bar{X}^2}{n-1}.$$

Ποσοστό. Έστω p το ποσοστό του πληθυσμού με μια συγκεκριμένη ιδιότητα (π.χ. ποσοστό ελαττωματικών σε παραγωγική διαδικασία). Το ποσοστό αυτό εκτιμάται

από το αντίστοιχο ποσοστό στο δείγμα, το οποίο ισούται με $\hat{p} = A_n / n$, όπου A_n ο αριθμός των μελών του δείγματος με αυτή την ιδιότητα.

Ιδιότητες εκτιμητριών

Η εκτίμηση βασίζεται σε μετρήσεις από ένα υποσύνολο του πληθυσμού (δείγμα), άρα είναι φυσικό να αποκλίνει από την πραγματική τιμή της παραμέτρου θ με συνέπεια την ύπαρξη σφάλματος εκτίμησης $|\hat{\theta} - \theta|$. Η μέση τιμή του τετραγώνου του σφάλματος εκτίμησης $\mathrm{E}[(\hat{\theta} - \theta)^2]$ ονομάζεται μέσο τετραγωνικό σφάλμα (mean squared error, MSE) και χαρακτηρίζει την ποιότητα της εκτιμήτριας: όσο μικρότερο το MSE, τόσο καλύτερη η εκτίμηση.

Μια άλλη επιθυμητή ιδιότητα είναι η εκτιμήτρια να δίνει κατά μέσο την πραγματική τιμή της άγνωστης παραμέτρου, δηλαδή $\mathrm{E}(\hat{\theta}) = \theta$. Μια τέτοια εκτιμήτρια ονομάζεται <u>αμερόληπτη.</u> Είναι φανερό από τον ορισμό του μέσου τετραγωνικού σφάλματος ότι για αμερόληπτες εκτιμήτριες ισχύει $\mathrm{MSE}=\mathrm{var}(\hat{\theta})$. Ακολουθούν παραδείγματα αμερόληπτων εκτιμητριών.

<u>Μέση τιμή δείγματος.</u> Έστω μ η μέση τιμή μιας τυχαίας μεταβλητής X.

$$\mathrm{E}(\bar{X}) = \frac{\mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)}{n} = \frac{\mu + \mu + \cdots + \mu}{n} = \frac{n\mu}{n} = \mu,$$

άρα η $\bar{X}$ είναι αμερόληπτη εκτιμήτρια της μ.

Έστω σ^2 η διασπορά της X. Επειδή οι τιμές του δείγματος είναι ανεξάρτητες τυχαίες μεταβλητές έχουμε

$$\mathrm{MSE} = \mathrm{var}(\bar{X}) = \frac{\mathrm{var}(X_1) + \mathrm{var}(X_2) + \cdots + \mathrm{var}(X_n)}{n^2} = \frac{\sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

<u>Διασπορά δείγματος.</u> Η μέση τιμή της διασποράς του δείγματος ισούται με

$$\mathrm{E}(S^2) = \frac{\mathrm{E}(X_1^2) + \mathrm{E}(X_2^2) + \cdots + \mathrm{E}(X_n^2) - n\mathrm{E}(\bar{X}^2)}{n-1},$$

όπου

$$\mathrm{E}(X_i^2) = \mathrm{var}(X_i) + [\mathrm{E}(X_i)]^2 = \sigma^2 + \mu^2, \quad i = 1, 2, \ldots, n,$$

$$\mathrm{E}(\bar{X}^2) = \mathrm{var}(\bar{X}) + [\mathrm{E}(\bar{X})]^2 = \frac{\sigma^2}{n} + \mu^2.$$

Συνδυάζοντας τις παραπάνω σχέσεις παίρνουμε $\mathrm{E}(S^2)=\sigma^2$, άρα η S^2 είναι αμερόληπτη εκτιμήτρια της σ^2.

Ποσοστό σε δείγμα. Έστω p η πιθανότητα ένα μέλος του δείγματος να έχει μια συγκεκριμένη ιδιότητα. Άρα ο αριθμός A_n των μελών του δείγματος με αυτή την ιδιότητα είναι διωνυμική τυχαία μεταβλητή με $\mathrm{E}(A_n)=np$ και $\mathrm{var}(A_n)=np(1-p)$. Επομένως

$$\mathrm{E}(\hat{p})=\frac{\mathrm{E}(A_n)}{n}=\frac{np}{n}=p,$$

άρα το $\hat{p}$ είναι αμερόληπτη εκτιμήτρια του p. Επίσης

$$\mathrm{MSE}=\mathrm{var}(\hat{p})=\frac{\mathrm{var}(A_n)}{n^2}=\frac{np(1-p)}{n^2}=\frac{p(1-p)}{n}.$$

Κατανομές εκτιμητριών

Υπάρχουν περιπτώσεις εκτιμητριών για τις οποίες, εκτός από τη μέση τιμή και τη διασπορά τους, είναι εφικτό να προσδιοριστεί και η συνάρτηση πυκνότητας πιθανότητας.

Μέση τιμή δείγματος. 1) Αν η τυχαία μεταβλητή X είναι κανονική με μέση τιμή μ και διασπορά σ^2, η μέση τιμή του δείγματος, ως γραμμικός συνδυασμός κανονικών τυχαίων μεταβλητών, είναι επίσης κανονική τυχαία μεταβλητή με μέση τιμή μ και διασπορά σ^2/n, δηλαδή $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$. Αν όμως η διασπορά είναι άγνωστη, χρησιμοποιούμε την εκτιμήτριά της S^2 για την οποία ισχύει $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\sim t_{n-1}$, όπου t_{n-1} είναι η κατανομή Student (ή κατανομή t) με $n-1$ βαθμούς ελευθερίας.

2) Αν το δείγμα είναι μεγάλο ($n\geq 30$), τότε λόγω του κεντρικού οριακού θεωρήματος και ανεξαρτήτως της κατανομής του πληθυσμού η μέση τιμή του δείγματος είναι προσεγγιστικά κανονική με μέση τιμή μ και διασπορά σ^2/n. Η προσέγγιση ισχύει ακόμα και όταν η διασπορά σ^2 είναι άγνωστη, οπότε έχουμε $\dfrac{\bar{X}-\mu}{S/\sqrt{n}}\sim N(0,1)$.

Ποσοστό σε δείγμα. Για μεγάλα δείγματα το ποσοστό $\hat{p}$ είναι προσεγγιστικά κανονική τυχαία μεταβλητή με μέση τιμή p και διασπορά $\frac{p(1-p)}{n}$.

2. Διαστήματα εμπιστοσύνης

Τα διαστήματα εμπιστοσύνης αποτελούν μια ευρύτερη μορφή εκτίμησης από τη σημειακή. Αντί η άγνωστη παράμετρος να εκτιμηθεί από μια συγκεκριμένη τιμή, υπολογίζεται ένα διάστημα που περιέχει την τιμή αυτής της παραμέτρου με μια προκαθορισμένη πιθανότητα.

Ορισμός. Το διάστημα $[a,b]$ είναι ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για την παράμετρο θ αν $\mathrm{P}(\theta \in [a,b]) = 1-\alpha$.

Διάστημα εμπιστοσύνης μέσης τιμής

Έστω ότι η τυχαία μεταβλητή X είναι κανονική με γνωστή διασπορά σ^2, άρα $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Έστω επίσης $z_{\alpha/2}$ η τιμή για την οποία ισχύει $\mathrm{P}\left(Y > z_{\alpha/2}\right) = \frac{\alpha}{2}$, όπου Y η τυποποιημένη κανονική τυχαία μεταβλητή. Για δεδομένο α η τιμή αυτή βρίσκεται από τον πίνακα της κανονικής κατανομής. Επομένως

$$\mathrm{P}\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = 1-\alpha \Rightarrow$$

$$\mathrm{P}\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Συγκρίνοντας την τελευταία σχέση με τον ορισμό του διαστήματος εμπιστοσύνης συμπεραίνουμε ότι το διάστημα $\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$ είναι ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη μέση τιμή μ.

Με παρόμοιο τρόπο μπορεί να δειχθεί ότι για κανονικό πληθυσμό με άγνωστη διασπορά ένα $100(1-\alpha)\%$ διάστημα εμπιστοσύνης για τη μέση τιμή είναι το $\left[\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$, όπου $t_{\alpha,\nu}$ είναι η τιμή για την οποία ισχύει $\mathrm{P}\left(t_\nu > t_{\alpha,\nu}\right) = \alpha$ και μπορεί να βρεθεί από τον πίνακα στο τέλος των σημειώσεων.

Τέλος, για μεγάλα δείγματα το αντίστοιχο διάστημα εμπιστοσύνης είναι της μορφής $\left[\bar{X} - z_{\alpha/2} \frac{\sigma \text{ ή } S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma \text{ ή } S}{\sqrt{n}} \right]$.

Το εύρος του διαστήματος εμπιστοσύνης έχει άμεση σχέση με το σφάλμα εκτίμησης $\left|\bar{X} - \mu\right|$. Συγκεκριμένα ισχύει $\mathrm{P}\left(\left|\bar{X} - \mu\right| > z_{\alpha/2} \frac{\sigma \text{ ή } S}{\sqrt{n}} \right) = \alpha$. Αν λοιπόν θέλουμε το σφάλμα εκτίμησης της μέσης τιμής να είναι μεγαλύτερο από d με πιθανότητα το πολύ α, πρέπει το μισό του $100(1-\alpha)\%$ διαστήματος εμπιστοσύνης να είναι το πολύ d. Αυτό σημαίνει ότι το απαιτούμενο μέγεθος δείγματος είναι $\left(\frac{z_{\alpha/2}}{d} \right)^2 \sigma^2$ ή $\left(\frac{z_{\alpha/2}}{d} \right)^2 S^2$ ανάλογα με το αν η διασπορά είναι γνωστή ή άγνωστη. Στη δεύτερη περίπτωση η διασπορά S^2 υπολογίζεται από κάποιο μικρό αρχικό δείγμα.

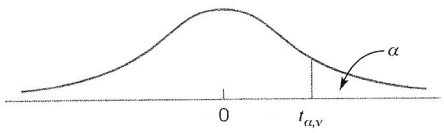
Διάστημα εμπιστοσύνης ποσοστού

Θεωρούμε την περίπτωση μεγάλων δειγμάτων για την οποία ισχύει $\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$. Παρόμοια με την περίπτωση της μέσης τιμής προκύπτει διάστημα εμπιστοσύνης $\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right]$, όπου όμως η παράμετρος p είναι η άγνωστη παράμετρος που θέλουμε να εκτιμήσουμε. Το πρόβλημα παρακάμπτεται αντικαθιστώντας το $p(1-p)$ με το $\hat{p}(1-\hat{p})$ και τελικά παίρνουμε διάστημα εμπιστοσύνης $\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$.

Όσον αφορά το σφάλμα εκτίμησης $\left|\hat{p} - p\right|$, αν θέλουμε να είναι μεγαλύτερο από d με πιθανότητα το πολύ α, πρέπει το μισό του $100(1-\alpha)\%$ διαστήματος εμπιστοσύνης να είναι το πολύ d. Αυτό σημαίνει μέγεθος δείγματος τουλάχιστον ίσο με $\left(\frac{z_{\alpha/2}}{d} \right)^2 \hat{p}(1-\hat{p})$, όπου το ποσοστό $\hat{p}$ υπολογίζεται από κάποιο μικρό αρχικό δείγμα. Εναλλακτικά, επειδή $\hat{p}(1-\hat{p}) \le 1/4$, μπορούμε ως απαιτούμενο μέγεθος δείγματος να πάρουμε το $\frac{1}{4}\left(\frac{z_{\alpha/2}}{d} \right)^2$.

APPENDIX IV

Percentage Points of the t Distribution[a]



	α									
ν	**0.40**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.0025**	**0.001**	**0.0005**
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.727	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.49	4.019	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.20	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.992
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

ν = degrees of freedom.

[a]Adapted with permission from *Biometrika Tables for Statisticians,* Vol. 1, 3rd ed., by E. S. Pearson and H. O. Hartley, Cambridge University Press, Cambridge, 1966.