ΚΕΦΑΛΑΙΟ 7

ΜΕΡΙΚΕΣ ΕΙΔΙΚΕΣ ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ

Στο κεφάλαιο αυτό, θα μελετήσουμε μερικές ειδικές διακριτές κατανομές που παρουσιάζουν ιδιαίτερο ενδιαφέρον λόγω, κυρίως των πολλών εφαρμογών τους. Πριν όμως ορίσουμε και μελετήσουμε τις κατανομές αυτές, θα ορίσουμε μια μορφή πειράματος που έχει άμεση σχέση με πολλές από αυτές.

Ορισμός: Ένα τυχαίο πείραμα θα λέγεται *πείραμα Bernoulli* αν είναι δυνατόν να καταλήξει σε ένα από δύο μόνο, ξένα μεταξύ τους, ενδεχόμενα που η ένωσή τους αποτελεί τον δειγματικό χώρο. Συνήθως χρησιμοποιούμε τους όρους "επιτυχία" (success) (S) και "αποτυχία" (failure) (F) για τα ενδεχόμενα αυτά και συμβολίζουμε με

$$p = P(S) \quad \text{και} \quad q = 1-p = P(F)\,, \quad 0 \le p \le 1$$

Ο δειγματικός χώρος, δηλαδή, σ' ένα πείραμα Bernoulli αποτελείται από δύο μόνο σημεία, τα S και F.

Παραδείγματα πειραμάτων Bernoulli αποτελούν το στρίψιμο ενός νομίσματος (κεφάλι-γράμματα), η ποιότητα ενός βιομηχανικού προϊόντος (ελαττωματικό-μη ελαττωματικό), η γέννηση ενός παιδιού (αγόρι - κορίτσι) κ.λ.π.

Ορισμός: Θα λέμε ότι έχουμε μια *ακολουθία* n *δοκιμών Bernoulli* όταν έχουμε n ανεξάρτητες επαναλήψεις ενός πειράματος Bernoulli με τέτοιο τρόπο ώστε η πιθανότητα επιτυχίας p να μένει ίδια από δοκιμή σε δοκιμή.

Ο δειγματικός χώρος σε μια ακολουθία n δοκιμών Bernoulli αποτελειται από 2^n σημεία της μορφής S και F.

Παραδείγματα ακολουθίας δοκιμών Bernoulli έχουμε στο στρίψιμο ενός νομίσματος n φορές, στο ρίξιμο ενός ζαριού n φορές κ.λ.π. Στην περίπτωση ακολουθίας δοκιμών Bernoulli, επειδή οι δοκιμές είναι ανεξάρτητες, οι πιθανότητες πολλαπλασιάζονται. Με άλλα λόγια, η πιθανότητα οποιασδήποτε συγκεκριμένης ακολουθίας

είναι το γινόμενο που παίρνουμε αντικαθιστώντας τα σύμβολα S και F με p και q αντίστοιχα.

Έτσι $P(SSFFS...SFS) = ppqqp...pqp$

Η ΚΑΤΑΝΟΜΗ BERNOULLI

Ορισμός: Έστω X μια τυχαία μεταβλητή με πεδίο τιμών το {0, 1} και κατανομή πιθανότητας

$$P(X=x) = \begin{cases} p & \text{αν } x=1 \\ 1-p & \text{αν } x=0 \end{cases}, \qquad 0 \le p \le 1$$

Η τυχαία μεταβλητή X λέγεται *τυχαία μεταβλητή Bernoulli (Bernoulli random variable)* και η κατανομή πιθανότητας της X λέγεται *κατανομή Bernoulli (Bernoulli distribution).*

Είναι προφανές ότι η κατανομή Bernoulli είναι μια καλώς ορισμένη κατανομή ($\sum_x P(x) = 1,\ P(x)\ge 0$).

Μοντέλα που οδηγούν στην κατανομή Bernoulli

Αν έχουμε ένα πείραμα Bernoulli και ορίσουμε την τυχαία μεταβλητή X με τιμές X(S)=1 και X(F)=0 και κατανομή πιθανότητας

$$P(X=x) = \begin{cases} p & \text{αν } x=1 \\ 1-p & \text{αν } x=0 \end{cases}, \qquad 0 \le p \le 1\ ,\quad q = 1-p$$

λέμε ότι η X ακολουθεί την κατανομή Bernoulli.

Πρόταση: Αν X ακολουθεί την κατανομή Bernoulli, τότε

$$E(X) = p,\ \Delta(X) = pq$$

Απόδειξη: Προφανής.

Η ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Ορισμός: Έστω X μια διακριτή τυχαία μεταβλητή με

$$P(X=x) = \binom{n}{x} p^x q^{n-x}$$

$$x=0,1,...,n\ ,\quad n=1,2,...\ ,\ 0 \le p \le 1\ ,\quad q = 1-p$$

Θα λέμε ότι η τυχαία μεταβλητή X ακολουθεί την *διωνυμική κατανομή (Binomial distribution) με παραμέτρους* n *και* p και θα συμβολίζουμε X ~ b(x;n,p).

Παρατήρηση: Η διωνυμική κατανομή είναι μια καλά ορισμένη κατανομή γιατί P(x) ≥ 0 και από το διωνυμικό ανάπτυγμα έχουμε ότι

$$\sum_{x=0}^{\infty} P(x) = \sum_{x=0}^{n} \binom{n}{x} p^x q^{n-x} = (p+q)^n = 1$$

Μοντέλα που οδηγούν στην διωνυμική κατανομή

i) Έστω X ο αριθμός των επιτυχιών σε μια ακολουθία n δοκιμών Bernoulli. Τότε X ~ b(x;n,p).

Απόδειξη: Το ενδεχόμενο {x επιτυχίες σε n δοκιμές} είναι η ένωση ενδεχομένων που αποτελούν αποτελέσματα n δοκιμών Bernoulli, x από τις οποίες έχουν καταλήξει σε επιτυχία και n-x σε αποτυχία (μια και μας ενδιαφέρει ο αριθμός των επιτυχιών στις n δοκιμές και όχι η σειρά με την οποία εμφανίζονται). Δηλαδή {x επιτυχίες σε n δοκιμές} $= \bigcup_{i=1}^{X} A_i$ όπου A_i είναι μια ακολουθία δοκιμών Bernoulli από x "S" και n-x "F". Τα A_i , όμως είναι ξένα μεταξύ τους και ισοπίθανα αφού

$$P(A_i) = p^x q^{n-x} \quad , i=1,2,...,n$$

Ο αριθμός των A_i είναι $\binom{n}{x}$ (όσοι και οι δυνατοί συνδυασμοί των x επιτυχιών στις n δοκιμές). Επομένως

$$P(X = x) = \sum_{i=1}^{\binom{n}{x}} P(A_i) = \binom{n}{x} p^x q^{n-x} \quad , \quad x=0,1,...,n$$

ii) Αν X_1 ,X_2 ,..., X_n είναι μια ακολουθία ανεξάρτητων τυχαίων μεταβλητών Bernoulli με παράμετρο p, τότε μπορεί να αποδειχθεί ότι η τυχαία μεταβλητή

$$X = X_1 + X_2 + ... + X_n$$

ακολουθεί την διωνυμική κατανομή με παραμέτρους n και p.

Δηλαδή, $X \sim b(x;n,p)$.

Παρατήρηση: Είναι προφανές ότι η κατανομή Bernoulli μπορεί να θεωρηθεί ως ειδική περίπτωση της διωνυμικής κατανομής για n=1.

Πρόταση: Αν $X \sim b(x;n,p)$ τότε

$$E(X) = np \quad \text{και} \quad \Delta(X) = npq$$

Απόδειξη: Κάνοντας χρήση του μοντέλου (ii) που οδηγεί στη διωνυμική κατανομή έχουμε ότι

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = np$$

και

$$\Delta(X) = \Delta(\Sigma X_i) = \sum_{i=1}^{n} \Delta(X_i) = \sum_{i=1}^{n} pq = npq$$

(μια και οι X_i είναι ανεξάρτητες).

Σημείωση: Αν $X \sim b(x;n,p)$, η συνάρτηση κατανομής του X στο σημείο x δίνεται από τον τύπο $P(X \le x) = \sum_{i=0}^{x} \binom{n}{i} p^i q^{n-i}$.

Τιμές της συνάρτησης κατανομής της διωνυμικής κατανομής για διάφορες του n και του p δίνονται στον πίνακα 1 του παραρτήματος.

Παράδειγμα: Ένα διαγώνισμα πολλαπλής επιλογής αποτελείται από 15 ερωτήσεις. Για κάθε ερώτηση, υπάρχουν 5 πιθανές απαντήσεις μια μόνο από τις οποίες είναι σωστή. Η βαθμολογία είναι 1 για κάθε σωστή απάντηση και 0 για κάθε λάθος απάντηση. Ένας φοιτητής διαλέγει την απάντηση σε κάθε ερώτηση στην τύχη. Να υπολογισθεί η πιθανότητα

α) Ο παραπάνω φοιτητής να πάρει το πολύ οκτώ.
β) Ο φοιτητής αυτός να βαθμολογηθεί με οκτώ.
γ) Να πάρει βαθμό μεγαλύτερο από 3 και μικρότερο από οκτώ.

Λύση: Έστω X η βαθμολογία του φοιτητή και p η πιθανότητα σωστής απάντησης. Είναι p=1/5=0.2 και n=15.

Επομένως κάνοντας χρήση των πινάκων

α) $P(X\leq 8|n=15, p=0.2) = \sum_{x=0}^{8} b(x;15, 0.2) = 0.9992$

β) $P(X=8) = \sum_{x=0}^{8} b(x;15, 0.2) - \sum_{x=0}^{8} b(x;15, 0.2)$

$= 0.9992 - 0.9958 = 0.0034$

γ) $P(3<X<8) = P(4\leq X\leq 7) = 0.9958 - 0.6482 = 0.3476$

Παράδειγμα: Στο προηγούμενο παράδειγμα, να βρεθεί
α) ο μέσος αναμενόμενος βαθμός των φοιτητών που απαντούν στην τύχη.
β) Ποιά θα είναι η βάση που θα πρέπει να καθορίσει ο καθηγητής έτσι ώστε ένας φοιτητής που απαντά μόνο στην τύχη να έχει πιθανότητα το πολύ ίση με 0.05 να περάσει;

Λύση: α) $E(X) = np = 3$.

β) Έστω α η ζητούμενη βάση. Τότε θα πρέπει

$P(X\geq\alpha) \leq 0.05$

ή ισοδύναμα

$1-P(X\leq \alpha-1) \leq 0.05 \Leftrightarrow P(X\leq \alpha-1) \geq 1-0.05 \Leftrightarrow$

$\Leftrightarrow P(X\leq \alpha-1) \geq 0.95$.

Από τους πίνακες, βρίσκουμε ότι $\alpha-1 = 6 \Rightarrow \alpha=7$.

Παράδειγμα: Αυτοκίνητα φθάνουν σε μια διασταύρωση όπου θα πρέπει υποχρεωτικά να στρίψουν δεξιά ή αριστερά. Έστω ότι τα αυτοκίνητα που φθάνουν στην διασταύρωση διαλέγουν την κατεύθυνση που θα στρίψουν ανεξάρτητα το ένα απο το άλλο. Έστω ότι η πιθανότητα p να στρίψει ένα αυτοκίνητο αριστερά είναι 0.7. Να υπολογισθεί η πιθανότητα
α) Τουλάχιστον 10 από τα επόμενα 15 αυτοκίνητα να στρίψουν αριστερά.

β) Μέσα στα επόμενα 15 αυτοκίνητα τουλάχιστον 10 να στρίψουν στην ίδια κατεύθυνση.

Λύση: Έστω X ο αριθμός των αυτοκινήτων, μεταξύ των 15, που στρίβουν αριστερά και Y ο αριθμός αυτών που στρίβουν δεξιά. Προφανώς, $X \sim b(x;15, 0.7)$ και $Y \sim b(y;15, 0.3)$.
Επομένως

α) P(τουλάχιστον 10 στρίβουν αριστερά) = $P(X\geq 10)$

$=P(X=10)+P(X=11)+P(X=12)+P(X=13)+P(X=14)+P(X=15)$

$=P(Y=5)+P(Y=4)+P(Y=3)+P(Y=2)+P(Y=1)+P(Y=0)$

$=\sum_{y=0}^{5} b(y;15, 0.3) = 0.7216.$

β) P(τουλάχιστον 10 στρίβουν στην ίδια κατεύθυνση)

=P(τουλάχιστον 10 στρίβουν δεξιά ή τουλάχιστον 10 στρίβουν αριστερά)

$=P(X\geq 10)+P(Y\geq 10)$

$=P(X\geq 10)+1-P(Y\leq 9)$

$=\sum_{y=0}^{5} (y;15, 0.3) + 1 - \sum_{y=0}^{9} (y;15, 0.3)$

$= 0.7216 + 1 - 0.9963 = 0.726.$

Σημείωση: Η λύση της άσκησης βασίσθηκε σε μια τεχνική που δίνει τη δυνατότητα χρησιμοποίησης των πινάκων της συνάρτησης κατανομής της διωνυμικής κατανομής για $p>0.5$. Συγκεκριμένα, στηρίζεται στην ιδιότητα που εύκολα μπορεί να αποδείξει κανείς ότι

$$b(x;n,p) = b(n-x;n,1-p)$$

Σημείωση: Πολλά βιβλία δίνουν πίνακες της κατανομής πιθανότητας και όχι της συνάρτησης κατανομής της διωνυμικής. Οι πίνακες αυτοί απαιτούν συνήθως περισσότερο χρόνο για τον υπολογισμό πιθανοτήτων.

Παράδειγμα: Ρίχνουμε ένα αμερόληπτο ζάρι 300 φορές. Να υπολογισθεί η πιθανότητα του ενδεχομένου να εμφανισθεί 1 ή 2 λιγότερες από 70 φορές.

Λύση: Στην προηγούμενη ενότητα βρήκαμε ένα άνω φράγμα της πιθανότητας αυτής. Με την χρήση της διωνυμικής κατανομής μπορούμε να την υπολογίσουμε ακριβώς. Είχαμε δει εκεί ότι αν X είναι ο αριθμός των 1 και 2 στις 300 δοκιμές τότε

$$X = X_1 + X_2 + ... + X_{300}$$

όπου X_i , i=1,2,...,300 είναι τυχαίες μεταβλητές Bernoulli με p=1/3. Επομένως, X ~ b(x;300,1/3) και P(το 1 και το 2 εμφανίζονται λιγότερες από 70 φορές) = $P(X<70) = \sum_{x=0}^{69} b(x;300,1/3)$.

Παράδειγμα: (Συνέχεια του παραδείγματος της δίκης του Collins).
Στο παράδειγμα εκείνο, είχαμε δει ότι το ζευγάρι που είχε κατηγορηθεί για τη ληστεία καταδικάσθηκε γιατί είχε όλα τα χαρακτηριστικά των ληστών και η πιθανότητα ένα ζευγάρι να είχε όλα αυτά τα χαρακτηριστικά ήταν 1 στα 12 εκατομμύρια. Το ζευγάρι έκανε έφεση και αθωώθηκε, χρησιμοποιώντας τον εξής μαθηματικό συλλογισμό: Έστω ότι η πιθανότητα να έχει ένα ζευγάρι όλα τα χαρακτηριστικά των ληστών είναι πραγματικά p=1/12.000.000 (μία παραδοχή που δεν την έκαναν και τόσο εύκολα). Έστω ότι υπάρχει ένα τέτοιο ζευγάρι σε ένα πληθυσμό N ζευγαριών. Ποιά είναι η πιθανότητα να υπάρχουν και άλλα τέτοια ζευγάρια;

Θεωρούμε κάθε ένα από τα N ζευγάρια του πληθυσμού σαν μια δοκιμή Bernoulli με "επιτυχία" το ενδεχόμενο το ζευγάρι να έχει τα χαρακτηριστικά των ληστών. Είναι p=1/12.000.000. Έστω X ο αριθμός των ζευγαριών που έχουν τα χαρακτηριστικά αυτά. Χρησιμοποιώντας τη διωνυμική κατανομή και τις ιδιότητες της δεσμευμένης πιθανότητας βλέπουμε ότι

P(να υπάρχουν περισσότερα από ένα τέτοια ζευγάρια | υπάρχει τουλάχιστον ένα τέτοιο ζευγάρι)

$= P(X \geq 2 \mid X \geq 1)$

$$= \frac{P(X \geq 2, X \geq 1)}{P(X \geq 1)} = \frac{P(X \geq 2)}{P(X \geq 1)} = \frac{1 - P(X < 2)}{1 - P(X < 1)}$$

$$= \frac{1 - b(0, N, p) - b(1, N, p)}{1 - b(0, N, p)}$$

$$= \frac{1 - (1 - p)^N - Np(1 - p)^{N-1}}{1 - (1 - p)^N}$$

Ο πίνακας που ακολουθεί δίνει την πιθανότητα αυτή για διάφορες τιμές του N.

Πίνακας

Αριθμός ζευγαριών N (Σε εκατομμύρια)	Πιθανότητες ύπαρξης ζευγαριού με τα συγκεκριμένα χαρακτηριστικά
1	0.0402
2	0.0786
3	0.1160
4	0.1522
5	0.1875
6	0.2216
7	0.2547
8	0.2868
9	0.3179
10	0.3479
15	0.4835
20	0.5959
25	0.6875
30	0.7610
40	0.8644
50	0.9256
75	0.9852
100	0.9973

Η υπεράσπιση χρησιμοποίησε την τιμή N=12.000.000 και κατέληξε στο συμπέρασμα ότι η πιθανότητα να υπάρχουν και άλλα ζευγάρια με τα χαρακτηριστικά των ληστών είναι περίπου 0.40. Το δικαστήριο δέχθηκε το επιχείρημα και έκανε δεκτή την προσφυγή.
(Περισσότερες πληροφορίες για τη δίκη αυτή μπορεί να βρει κανείς στο περιοδικό TIME, της 26/4/68 σε άρθρο με τίτλο "Δίκη με την

χρήση μαθηματικών" (*Trial by Mathematics*). Επίσης, βλέπε Fairly & Mosteller (1974).

Παράδειγμα: ***Διαγωνίσματα πολλαπλής επιλογής.*** Μια πρακτική εφαρμογή των δοκιμών Bernoulli και της διωνυμικής κατανομής συναντάται στα διαγωνίσματα πολλαπλής επιλογής (multiple choice). Συνήθως, στα διαγωνίσματα αυτά, υπάρχουν τέσσερις δυνατές απαντήσεις (Α, Β, Γ, Δ), από τις οποίες μία μόνο είναι σωστή. Κάποιος που επιλέγει απάντηση στην τύχη έχει πιθανότητα p = 0.25 επιλογής της σωστής απάντησης. Αν υπάρχουν n = 20 ερωτήσεις σε κάποιο τέστ, τότε, σύμφωνα με την μέση τιμή της διωνυμικής κατανομής, ο αναμενόμενος αριθμός των σωστών απαντήσεων που θα βασίζονται σε τυχαίες επιλογές θα είναι $p_n = (0.25)(20) = 5$. Επομένως, σε επανάληψη τέτοιας μορφής διαγωνισμάτων ένας φοιτητής που απαντατά πάντοτε στην τύχη θα έχει κατά μέσο όρο πέντε σωστές απαντήσεις ανά διαγώνισμα.

Είναι προφανές ότι, σε ένα συγκεκριμένο διαγώνισμα αυτής της μορφής, ο φοιτητής που επιλέγει τυχαία τις απαντήσεις μπορεί να μαντέψει λιγότερες ή περισσότερες από πέντε σωστές απαντήσεις. Οι ακριβείς πιθανότητες δίνονται από την διωνυμική κατανομή για p=0.25 και n = 20. Ο πίνακας που ακολουθεί δείχνει τις πιθανότητες που αντιστοιχούν σε μια σειρά από σωστές απαντήσεις.

Αριθμός ορθών απαντήσεων	Πιθανότητα
< 3	0.0913
3	0.1339
4	0.1896
5	0.2024
6	0.1686
7	0.1124
> 7	0.1018

Πολλά διαγωνίσματα πολλαπλής επιλογής βαθμολογούνται με τέτοιο τρόπο, ώστε αυτός που απαντά στην τύχη να έχει αναμενόμενο βαθμό μηδέν, τον ίδιο βαθμό, δηλαδή, με κάποιο που δεν γνωρίζει τις ερωτήσεις και δεν τις απαντά. Δεδομένου ότι κάποιος που απαντά

στην τύχη είναι δυνατόν να επιλέξει σωστές απαντήσεις - και προκειμένου η αναμενόμενη βαθμολογία να είναι μηδέν - θα πρέπει να αφαιρούνται βαθμοί για τις λανθασμένες απαντήσεις. Όπως προαναφέρθηκε, σε ένα διαγώνισμα 20 ερωτήσεων, ο αναμενόμενος αριθμός σωστών απαντήσεων για κάποιον που απαντά στην τύχη είναι 5 και, επομένως, ο αναμενόμενος αριθμός λανθασμένων απαντήσεων για το άτομο αυτό είναι 15. Είναι προφανές ότι για να έχει ένας τέτοιος φοιτητής μέσο αναμενόμενο βαθμό μηδέν, θα πρέπει από κάθε λανθασμένη απάντηση να αφαιρείται 1/3 του βαθμού για κάθε λανθασμένη απάντηση.

Γενικότερα, αν X είναι ο αριθμός των σωστών απαντήσεων σε ένα διαγώνισμα n ερωτήσεων όπου ο υποψήφιος απαντά με τυχαίο τρόπο και δίνεται ένας βαθμός για κάθε σωστή ερώτηση, τότε, αν ω είναι οι βαθμοί που αφαιρούνται για κάθε λανθασμένη απάντηση, θα πρέπει, χρησιμοποιώντας τις ιδιότητες της μέσης τιμής, να έχουμε ότι

$$E\{(X)(1)+(n-X)(-\omega)\} = E\{(-\omega)(n)+(1+\omega)(X)\} = (-\omega)(n)+(1+\omega)pn$$

Προκειμένου ο μέσος βαθμός σε τέτοια διαγωνίσματα για κάποιον που απαντά στην τύχη να είναι μηδέν, θα πρέπει να έχουμε ω=p/(1-p).

Έτσι, με τέσσερις δυνατές απαντήσεις για κάθε ερώτηση όπου p = 0.25, η "ποινή" για κάθε λάθος απάντηση θα πρέπει να είναι $\omega = \frac{0.25}{0.75} = \frac{1}{3}$. Με πέντε δυνατές απαντήσεις, p = 0.2 και η "τιμή" θα πρέπει να είναι $\omega = \frac{0.2}{0.8} = \frac{1}{4}$. (Αυτές είναι και οι "ποινές" που χρησιμοποιούνται στο SAT τέστ και στο GMAT τέστ σε άλλες γνωστές διεθνείς εξετάσεις πολλαπλής επιλογής).

Για ένα διαγώνισμα της μορφής αυτής τίθεται το ερώτημα αν "συμφέρει" ένα φοιτητή που δεν γνωρίζει την σωστή απάντηση να απαντά στην τύχη ή όχι. Στις περίπτωσεις αυτές, είναι φυσικά προτιμότερο να χρησιμοποιεί κανείς την κρίση του. Αν μπορεί να αποκλείσει μία από τις δυνατές απαντήσεις, η τυχαία επιλογή θα αυξήσει τον αναμενόμενο βαθμό που είναι καλύτερος από το μηδέν

που θα πάρει, αν αφήσει την ερώτηση αναπάντητη. Για παράδειγμα, ας υποθέσουμε ότι υπάρχει μια δύσκολη ερώτηση με τέσερις δυνατές απαντήσεις (Α, Β, Γ, Δ). Ας υποθέσουμε ότι ο υποψήφιος μπορεί να αποκλείσει μία από τις απαντήσεις, έστω την Α, αλλά δεν έχει ιδέα για το ποιά από τις υπόλοιπες είναι σωστή. Αν επιλέξει στην τύχη έχει πιθανότητα 1/3 να επιλέξει την σωστή απάντηση (και να κερδίσει τον ένα βαθμό) και 2/3 να επιλέξει λανθασμένη απάντηση (και να "τιμωρηθεί" με 1/3 βαθμού). Η αναμενόμενη βαθμολογία σε αυτή την ερώτηση θα είναι

$$(1)(1/3) + (-1/3)(2/3) = 1/9$$

η οποία, έστω και αν είναι μικρή, είναι θετική. Κατά μέσο όρο, μακροπρόθεσμα, αυτή η στρατηγική που στηρίζεται στην κριτική επιλογή με τυχαίο τρόπο θα αυξήσει την βαθμολογία του φοιτητή κατά 1/9 βαθμών ανά ερώτηση.

Τί συμβαίνει στην περίπτωση που οι απαντήσεις δίνονται με εντελώς τυχαίο τρόπο επειδή ο υποψήφιος δεν έχει την δυνατότητα να επιλέξει με κριτικό τρόπο; Αν για παράδειγμα κάποιος δεν έχει χρόνο και του έχουν μείνει μια σειρά από αναπάντητες ερωτήσεις, είναι καλύτερα να τις απαντήσει στην τύχη ή να τις αφήσει αναπάντητες; Συνήθως, οι οδηγίες τέτοιων διαγωνισμάτων πολλαπλής επιλογής αναφέρουν ότι *"η τυχαία επιλογή είναι πιθανόν να οδηγήσει σε μηδενική βαθμολογία και, επομένως, δεν αποτελεί μια σωστή στρατηγική"*. Είναι βέβαια σωστό ότι απαντήσεις στην τύχη θα έχουν ένα αναμενόμενο μέσο βαθμό μηδέν και, επομένως, μια τέτοια στρατηγική κατά μέσο όρο δεν θα αποδώσει μακροπρόθεσμα. Δεδομένου όμως ότι αυτός που παίρνει το διαγώνισμα ενδιαφέρεται για το συγκεκριμένο διαγώνισμα, ο υποθετικός μακροχρόνιος μέσος δεν έχει σημασία για αυτόν. Σε ένα συγκεκριμένο διαγώνισμα, η τυχαία επιλογή απαντήσεων σε κάποιες ερωτήσεις ίσως αυξήσει αλλά ίσως και μειώσει την συνολική βαθμολογία.

Όπως έχουμε ήδη δει, εάν απαντήσει κανείς τυχαία σε ένα διαγώνισμα 20 ερωτήσεων, υπάρχει πιθανότητα μόνο 0.2 σωστής επιλογής. Επομένως, υπάρχει πιθανότητα 0.8 ότι αυτός που απαντά στην τύχη θα απαντήσει περισσότερες ή λιγότερες από τον αναμενόμενο αριθμό των σωστών απαντήσεων. Επομένως, σε ένα

δεδομένο διαγώνισμα όπου ο υποψήφιος απαντά στην τύχη, η πιθανότητα ότι θα πάρει βαθμό μεγαλύτερο ή μικρότερο από τον αναμενόμενο βαθμό είναι 0.8.

Μια από τις πιο γνωστές περιπτώσεις όπου χρησιμοποιούνται οι εξετάσεις πολλαπλής επιλογής είναι το Graduate Management Admission Test (GMAT) που χρησιμοποιείται από πολλά πανεπιστήμια του εξωτερικού, αλλά και από το Οικονομικό Πανεπιστήμιο Αθηνών, προκειμένου να επιλεγούν μεταπτυχιακοί φοιτητές για σπουδές στην Διοίκηση Επιχειρήσεων. Το διαγώνισμα αυτό διαρκεί 3 ½ ώρες και αποτελείται από περίπου 200 ερωτήσεις πολλαπλής επιλογής που καλύπτουν την ικανότητα χρήσης πληροφοριών από κείμενο που έχει αναγνωσθεί (reading recall), προφορική ικανότητα (verbal aptitude), μαθηματικά (mathematics), ικανότητα διαχείρισης δεδομένων (data sufficiency) και επιχειρηματική κρίση (business judgement). Κάθε ερώτηση έχει πέντε δυνατές απαντήσεις και για κάθε σωστή απάντηση δίνεται ένας βαθμός, ενώ για κάθε λανθασμένη απάντηση αφαιρείται 1/4 του βαθμού.

Η ΥΠΕΡΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Ορισμός: Έστω X μια διακριτή τυχαία μεταβλητή με

$$P(X=x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

n=1,2,..., N=1,2,..., m=0,1,2,...,N, x=0,1,2,...,min(m,n).
Η τυχαία μεταβλητή X λέγεται ότι ακολουθεί την *υπεργεωμετρική κατανομή με παραμέτρους* N, n *και* m. (Συμβολικά X ~ h(x;N,n,m)).

Παρατήρηση: Η υπεργεωμετρική κατανομή είναι μια καλά ορισμένη κατανομή γιατί P(x)≥0 και

$$\sum_{x=0}^{\min(m,n)} \binom{m}{x}\binom{N-m}{n-x} = \binom{N}{n}$$

Μοντέλα που οδηγούν στην υπεργεωμετρική κατανομή

1) *Δειγματοληψία απο υδρία*. Σε ένα δοχείο υπάρχουν N σφαιρίδια. Από αυτά m είναι μαύρα και τα υπόλοιπα N-m άσπρα. Διαλέγουμε στην τύχη ένα σύνολο από n σφαιρίδια. Εστω X ο αριθμός των μαύρων σφαιριδίων στο σύνολο αυτό. Με τις προϋποθέσεις αυτές

$$X \sim h(x;N,n,m)$$

Απόδειξη: Το σύνολο των δυνατών περιπτώσεων του πειράματος είναι $\binom{N}{n}$. Για τον καθορισμό των ευνοϊκών περιπτώσεων παρατηρούμε ότι στο σύνολο των n σφαιριδίων πρέπει να έχουμε x μαύρα και n-x άσπρα. Τα x μαύρα μπορούν να επιλεγούν με $\binom{m}{x}$ τρόπους και τα n-x άσπρα με $\binom{N-m}{n-x}$ τρόπους. Κάθε όμως επιλογή x μαύρων μπορεί να συνδυασθεί με οποιαδήποτε από τις επιλογές n-x άσπρων. Επομένως, κάτω από την υπόθεση ότι όλα τα υποσύνολα μεγέθους n έχουν την ίδια πιθανότητα να επιλεγούν, καταλήγουμε στο ζητούμενο.

Σημείωση: Οι πιθανότητες P_x ορίζονται, προφανώς, για $x \le \min(m,n)$. Επειδή όμως $\binom{\alpha}{\beta}=0$ για κάθε $\beta>\alpha$ ο τύπος της υπεργεωμετρικής κατανομής ισχύει για όλα τα $x \ge 0$ με την προϋπόθεση ότι $P_x = 0$ υποδηλώνει ότι η τιμή x είναι ανέφικτη για την τυχαία μεταβλητή X. Από την παρατήρηση αυτή προκύπτει ότι η υπεργεωμετρική κατανομή είναι μια πεπερασμένη κατανομή (Υπάρχει δηλαδή ένα πεπερασμένο πλήθος τιμών της τυχαίας μεταβλητής με μη μηδενική πιθανότητα).

Σημείωση: Το προηγούμενο πρόβλημα μπορεί να παρουσιασθεί εναλλακτικά ως εξής: Έστω ότι η επιλογή των n σφαιριδίων του προηγούμενου παραδείγματος γίνεται ένα προς ένα με τρόπο ώστε σε κάθε επιλογή σημειώνουμε το χρώμα του σφαιριδίου και το απομακρύνουμε από το δοχείο. (Δειγματοληψία χωρίς επανάθεση). Έστω A_i το ενδεχόμενο ότι το i σφαιρίδιο είναι μαύρο και έστω

$X_{A_i}(x_i) \equiv X_i$ η αντίστοιχη μεταβλητή-δείκτης (μεταβλητή Bernoulli) του ενδεχομένου A_i (i=1,2,...,n). (X_i =1 αν το i σφαιρίδιο είναι μαύρο και X_i =0 αν το i σφαιρίδιο είναι λευκό). Ο συνολικός αριθμός X των μαύρων σφαιριδίων σε ένα σύνολο n δοκιμών της μορφής αυτής είναι, προφανώς

$$X = X_1 + X_2 + ... + X_n$$

Η τυχαία μεταβλητή X ακολουθεί την υπεργεωμετρική κατανομή με παραμέτρους N, n, m.

Πριν αποδείξουμε τον παραπάνω ισχυρισμό αποδεικνύουμε το ακόλουθο λήμμα.

Λήμμα: $P(X_k = 1) = \frac{m}{n} \equiv p$ για κάθε k=1,2,...,n.

(Δηλαδή η πιθανότητα "επιτυχίας" (μαύρου σφαιριδίου) στην i δοκιμή είναι σταθερή (ανεξάρτητη από την σειρά της δοκιμής).

Απόδειξη: Ας υποθέσουμε προς στιγμήν ότι τα m μαύρα σφαιρίδια είναι σημειωμένα με τα στοιχεία 1,2,...,m και έστω S_{jk} το ενδεχόμενο της επιλογής του j μαύρου σφαιριδίου στην k δοκιμή. Το ενδεχόμενο $\{X_k = 1\} = \{A_k\}$ ισοδυναμεί με το ενδεχόμενο $[S_{1k} \cup S_{2k} \cup ... \cup S_{mk}]$ (μια και οποιαδήποτε επιλογή από τα m μαύρα σφαιρίδια στην k δοκιμή συνεπάγεται πραγματοποίηση του ενδεχομένου $\{A_k\}$). Τα ενδεχόμενα S_{jk} , j=1,2,...,m είναι προφανώς ξένα μεταξύ τους.

Επομένως,

$$P(X_k = 1) = P(S_{1k} \cup S_{2k} \cup ... \cup S_{mk}) = P(S_{1k}) + P(S_{2k}) + ... + P(S_{mk})$$

Ομοίως,

$$P(S_{ik}) = \frac{N-1}{N}\frac{N-2}{N-1}...\frac{N-(k-1)}{N-(k-2)}\frac{1}{N-(k-1)} = \frac{1}{N}$$

που δεν είναι τίποτε άλλο από το γινόμενο των πιθανοτήτων της μη επιλογής του συγκεκριμένου αυτού σφαιριδίου στις πρώτες k-1 δοκιμές και της πιθανότητας της επιλογής αυτού του συγκεκριμένου σφαιριδίου στην k δοκιμή. Δηλαδή, τελικά

$$P(X_k = 1) = m\frac{1}{N} = \frac{m}{N}$$

Προσδιορισμός της κατανομής του X. Μια και η επιλογή των σφαιριδίων γίνεται χωρίς επανάθεση, ο συνολικός αριθμός των διαφορετικών διατεταγμένων υποσυνόλων μεγέθους n των N σφαιριδίων είναι ο αριθμός των διατάξεων των N ανά n δηλαδή $N(N-1) \ldots (N-n+1)= N_{(n)}$. Ευνοϊκές περιπτώσεις είναι τα διατεταγμένα σύνολα μεγέθους n στα οποία κανένα σφαιρίδιο δεν εμφανίζεται περισσότερο από μια φορά (δειγματοληψία χωρίς επανάθεση) και στα οποία υπάρχουν x μαύρα και n-x άσπρα σφαιρίδια. Σύνολα της μορφής αυτής μπορούν να κατασκευασθούν σε 3 βήματα.

i) Επιλογή, χωρίς επανάθεση και χωρίς να ενδιαφέρει η διάταξη, των x μαύρων από τα m μαύρα σφαιρίδια (αυτό μπορεί να γίνει με $\binom{m}{x}$ τρόπους).

ii) Επιλογή, χωρίς επανάθεση και χωρίς να ενδιαφέρει η διάταξη, των n-x άσπρων από τα N-m άσπρα σφαιρίδια. (αυτό μπορεί να γίνει με $\binom{N-m}{n-x}$ τρόπους).

iii) Συνδυασμός των παραπάνω επιλογών και επιλογή ενός από τα n! $\binom{m}{x}\binom{N-m}{n-x}$ δυνατά διατεταγμένα σύνολα με x μαύρα και n-x άσπρα σφαιρίδια.

Επομένως,

$$P(X = x) = \frac{n!\binom{m}{x}\binom{N-m}{n-x}}{N_{(n)}} = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

δηλαδή $X \sim h(x;N,n,m)$.

Παρατήρηση: Το προηγούμενο μοντέλο της υπεργεωμετρικής κατανομής οδηγεί στο συμπέρασμα ότι μια τυχαία μεταβλητή που ακολουθεί την υπεργεωμετρική κατανομή μπορεί να παρασταθεί σαν άθροισμα m ισόνομων τυχαίων μεταβλητών Bernoulli κάθε μια από τις οποίες παίρνει τις τιμές 1 και 0 με πιθανότητες

$$p = \frac{m}{N} \text{ και } q = \frac{N - m}{N}$$

αντίστοιχα. Πράγματι, για $i \neq j$ ισχύει

$$P(X_i = 1, X_j = 1) = P(X_i = 1 \mid X_j = 1)\, P(X_j = 1) = \frac{m - 1}{N - 1}\frac{m}{N}$$

ενώ

$$P(X_i = 1)\, P(X_j = 1) = \frac{m^2}{N^2}$$

Είναι χρήσιμο να παρατηρήσει κανείς την διαφορά της υπεργεωμετρικής κατανομής από την διωνυμική κατανομή όπου, όπως είδαμε στο μοντέλο (ii) της διωνυμικής κατανομής τα X_i είναι ανεξάρτητες και *ισόνομες* τυχαίες μεταβλητές Bernoulli. Στην θεώρηση αυτών των δύο κατανομών ως αποτελέσματος τυχαίας δειγματοληψίας, η παρατήρηση αυτή συνεπάγεται ότι η δειγματοληψία χωρίς επανάθεση έχει σαν αποτέλεσμα οι παρατηρήσεις (τυχαίες μεταβλητές) να είναι εξαρτημένες. Το γεγονός αυτό καθιστά αδύνατη την χρησιμοποίηση της μεθόδου των γεννητριών συναρτήσεων πιθανοτήτων για τον καθορισμό της κατανομής πιθανότητας μιας υπεργεωμετρικής τυχαίας μεταβλητής ως αθροίσματος τυχαίων μεταβλητών Bernoulli.

Παρατήρηση: Εκ πρώτης όψεως ίσως φαίνεται παράδοξο το γεγονός ότι στην υπεργεωμετρική κατανομή η πιθανότητα "επιτυχίας" (μαύρου σφαιριδίου) σε μια δοκιμή είναι σταθερή (ανεξάρτητη από την συγκεκριμένη δοκιμή). Αυτό οφείλεται στο γεγονός ότι αναφερόμαστε στην περιθώρια πιθανότητα "επιτυχίας" στην k δοκιμή και όχι στην δεσμευμένη πιθανότητα. Δεν μας ενδιαφέρει δηλαδή τι έγινε στις προηγούμενες k-1 δοκιμές. Αν μας ενδιέφερε η πιθανότητα "επιτυχίας" στην k δοκιμή δοθέντων των αποτελεσμάτων των προηγουμένων k-1 δοκιμών η κατάσταση θα ήταν, όπως είναι φυσικό, διαφορετική.

Εφαρμογές της Υπεργεωμετρικής Κατανομής

Στατιστικός Έλεγχος Ποιότητας *(Statistical Quality Control)*

Στην βιομηχανία, όπως είναι γνωστό, ενδιαφέρει ο καθορισμός του ποσοστού των ελαττωματικών αντικειμένων που φθάνουν στην αγορά, και ο περιορισμός του σε προκαθορισμένα "ανεκτά" επίπεδα. Σε πολλές περιπτώσεις δεν είναι δυνατόν να ελεγχθεί κάθε αντικείμενο που παράγεται, είτε διότι ο έλεγχος είναι πολύ δαπανηρός, είτε διότι συνεπάγεται καταστροφή του ελεγχόμενου προϊόντος. Σε τέτοιες περιπτώσεις χρησιμοποιείται μια μέθοδος δειγματικού ελέγχου που χρησιμοποιεί την έννοια της υπεργεωμετρικής κατανομής.

Έστω ότι τα παραγόμενα αντικείμενα είναι ή ελαττωματικά ή μη ελαττωματικά και ότι φθάνουν στο τμήμα ελέγχου σε πακέτα μεγέθους N αντικειμένων. Από κάθε πακέτο επιλέγεται τυχαία για έλεγχο ένα δείγμα μεγέθους n. Έστω X ο αριθμός των ελαττωματικών αντικειμένων στο δείγμα. Εάν το X είναι μεγάλο, είναι πιθανόν το πακέτο να περιέχει πολλά ελαττωματικά οπότε θα πρέπει να απορριφθεί. Αντίστροφα, αν το X ειναι μικρό, είναι πιθανόν το πακέτο να περιέχει λίγα ελαττωματικά οπότε θα πρέπει να θεωρηθεί αποδεκτό. Αυτό οδηγεί σε ένα κανόνα "δεκτό το πακέτο αν $X<c$, απορριπτέο αν $X\geq c$ (ή ανάγκη για παραπέρα έλεγχο)". Ο καθορισμός του n και του c εξαρτάται από το ανεκτό περιθώριο λαθών και το κόστος ελέγχου.

Παράδειγμα: Έστω ότι τα νέα πρϊόντα μιας παραγωγής φθάνουν για έλεγχο σε πακέτα των N=50 αντικειμένων και ότι ένα τυχαίο δείγμα n=10 από αυτά περνά από έλεγχο. Το πακέτο θεωρείται αποδεκτό αν το δείγμα περιέχει το πολύ ένα ελαττωματικό αντικείμενο. Να καθορισθούν οι πιθανότητες αποδοχής του πακέτου ως συναρτήσεις του αριθμού m των ελαττωματικών αντικειμένων σ' αυτό.

Λύση: Η πιθανότητα ότι ένα τυχαίο δείγμα μεγέθους 10 περιέχει ακριβώς x ελαττωματικά δίνεται από τον τύπο της υπεργεωμετρικής κατανομής

$$P(X=x) = \frac{\binom{m}{x}\binom{50-m}{10-x}}{\binom{50}{10}}, \qquad x=0,1,2,...$$

(μια και φυσικά πρόκειται για δειγματοληψία χωρίς επανάθεση). Η πιθανότητα αποδοχής του πακέτου είναι

$$P(X\le 1) = P(X=0) + P(X=1) = \left\{\binom{50-m}{10} + \binom{m}{1}\binom{50-m}{9} \Big/ \binom{50}{10}\right\}$$

Η πιθανότητα αποδοχής του πακέτου είναι 1 για m=0 και m=1 ενώ είναι δυνατόν να προσδιορισθεί επαγωγικά για m>1. Ο πίνακας που ακολουθεί δίνει τις πιθανότητες αποδοχής p για διαφορετικές τιμές του m.

Πίνακας
Πιθανότητες αποδοχής ως συναρτήσεις του m

m	4	8	12	16	20	24
p	0.826	0.491	0.236	0.094	0.031	0.008

Σε πολλά προβλήματα, όπως αυτό του προηγούμενου παραδείγματος, είναι αναγκαίο να υπολογισθεί η P(x) για αρκετές διαδοχικές τιμές του x. Μια απλή μέθοδος για τον υπολογισμό των πιθανοτήτων αυτών συνίσταται στον υπολογισμό της P(x) από τον τύπο για την μικρότερη από τις τιμές του x που μας ενδιαφέρει και κατόπιν στον υπολογισμό των υπολοίπων πιθανοτήτων μέσω του αναγωγικού τύπου

$$P(x) = r(x)\, P(x-1)$$

όπου r(x) είναι ο λόγος δύο διαδοχικών όρων:

$$r(x) = \frac{P(x)}{P(x-1)} = \frac{(m-x+1)(n-x+1)}{x(N-m-n+x)}$$

Μέθοδος Σύλληψης και Επανασύλληψης
(Capture-recapture method)

Η διαδικασία που ακολουθεί χρησιμοποιείται αρκετές φορές για τον κατα προσέγγιση υπολογισμό (εκτίμηση) του μεγέθους ενός πληθυσμού ζώων όπως για παράδειγμα τον αριθμό N των ψαριών σε

μια λίμνη. Αρχικά, ψαρεύονται μερικά ψάρια, έστω m, σημειώνονται, έτσι ώστε να είναι δυνατόν να αναγνωρισθούν στο μέλλον, και επιστρέφονται στην λίμνη. Η λίμνη τότε περιέχει m σημειωμένα ψάρια και N-m μη σημειωμένα. Στην συνέχεια, συγκεντρώνεται ένα δεύτερο δείγμα απο n ψάρια. Υποθέτοντας ότι το δείγμα αυτό είναι τυχαίο, έχουμε ότι η πιθανότητα να περιέχει x σημειωμένα ψάρια θα δίνεται από την υπεργεωμετρική κατανομή. Μια λογική εκτίμηση για τον συνολικό αριθμό ψαριών στην λίμνη είναι στην περίπτωση αυτή η τιμή $\left(n \frac{m}{x} \right)$.

Πρόταση: Αν $X \sim h(x;N,n,m)$, τότε

$$E(X) = np,\ \Delta(X) = np(1-p)\ \frac{N-n}{N-1} \quad \text{όπου } p = \frac{m}{N}$$

Απόδειξη: Έχουμε ήδη δει ότι $X=X_1+X_2+...+X_m$ όπου X_i, i=1,2,...,m είναι τυχαίες μεταβλητές Bernoulli.
Ισχύει ότι

$$E(X_i) = E(X_i^2) = p \quad \text{και } \Delta(X_i) = p(1-p)$$

Επομένως,

$$E(X) = \sum_{i=1}^{n} E(X_i) = np$$

Για τον υπολογισμό της $\Delta(X)$ απαιτείται προηγουμένως ο καθορισμός της $Cov(X_i, X_j)$ μια και οι τυχαίες μεταβλητές $X_1, X_2, ..., X_m$ είναι εξαρτημένες. Επειδή

$$P(X_i = 1, X_j = 1) = \frac{m(m-1)}{N(N-1)}$$

έχουμε

$$E(X_i X_j) = \sum\sum P(x_i, x_j)\, x_i\, x_j = P(X_i = 1, X_j = 1) = \frac{m(m-1)}{N(N-1)}$$

Επομένως,

$$Cov(X_i, X_j) = E(X_i X_j) - E(X_i)\, E(X_j) = \frac{m(m-1)}{N(N-1)} - \frac{m^2}{N^2}$$

$$= \frac{m}{N}\left[\frac{m-1}{N-1} - \frac{m}{N}\right] = \frac{m}{N}\left[\frac{Nm - N - Nm + m}{(N-1)N}\right]$$

$$= \frac{m(m-N)}{N^2(N-1)}$$

$$= -\frac{p(1-p)}{N-1}$$

(Το αρνητικό πρόσημο της συνδιασποράς εξηγείται από το γεγονός ότι εκλογή μαύρου σφαιριδίου στην j δοκιμή ελαττώνει την πιθανότητα επιλογής μαύρου σφαιριδίου στην i επιλογή).

Τελικά

$$\Delta(X) = \Delta\left(\sum X_i\right) = \sum \Delta(X_i) + 2\sum_{j<i} \mathrm{Cov}\left(X_i\, X_j\right) =$$

$$= np(1 - p) + 2\binom{n}{2}\frac{p\,(1-p)}{N-1}$$

$$= np(1-p)\frac{N-n}{N-1}$$

Παρατήρηση: Σύγκριση της ιδιότητας αυτής της υπεργεωμετρικής κατανομής με την αντίστοιχη ιδιότητα της διωνυμικής κατανομής οδηγεί στο συμπέρασμα ότι η διωνυμική και η υπεργεωμετρική κατανομή έχουν την ίδια μέση τιμή ενώ η διασπορά της υπεργεωμετρικής είναι μικρότερη της αντίστοιχης διασποράς της διωνυμικής κατά ένα παράγοντα $\frac{N-n}{N-1}$. Ο παράγοντας αυτός λέγεται *παράγοντας διόρθωσης (correction factor) ή διόρθωση πεπερασμένου πληθυσμού (finite population correction)*.

Επειδή $\frac{N-n}{N-1} \underset{N\to\infty}{\longrightarrow} 1$ είναι προφανές ότι η διασπορά της υπεργεωμετρικής κατανομής συγκλίνει στην διασπορά της διωνυμικής κατανομής όταν το N αυξάνει. Αναμένεται λοιπόν να υπάρχει κάποια προσέγγιση της υπεργεωμετρικής κατανομής από την διωνυμική κατανομή στην περίπτωση αυτή. Πράγματι, ισχύει το παρακάτω θεώρημα.

Θεώρημα: $h(x;N,n,m) \underset{N\to\infty}{\longrightarrow} b(x;n,p)$

για σταθερά x,n και σταθερό $\frac{m}{N} = p$.

Απόδειξη:

$$h(x;N,n,m) = P(X=x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

$$= \frac{(m(m-1)...(m-x+1)(N-m)(N-m-1)...(N-m-n+x+1)}{N(N-1)...(N-x+1)(N-x)(N-x-1)...(N-n+1)} \frac{n!}{x!(n-x)!}$$

Αλλά, για σταθερό x

$$\frac{m-k}{N-k} \underset{N\to\infty}{\longrightarrow} \frac{m}{N} = p, \quad k \text{ σταθερό, } k=0,1,2,...,x-1$$

Επίσης,

$$\frac{N-m-k}{N-x-k} \underset{N\to\infty}{\longrightarrow} 1-p=q, \quad k \text{ σταθερό, } k=0,1,2,...,n-x-1$$

Επομένως,

$$h(x;N,n,m) \underset{N\to\infty}{\longrightarrow} p^x q^{n-x} \qquad \binom{n}{r} \sim b(x;n,p)$$

Το προηγούμενο θεώρημα αποδεικνύει ότι, για αρκετά μεγάλο N και m, η υπεργεωμετρική κατανομή μπορεί να προσεγγισθεί από την διωνυμική κατανομή. Αυτό είναι λογικό μια και στα μοντέλα δειγματοληψίας, όταν η δειγματοληψία γίνεται χωρίς επανάθεση και το N είναι αρκετά μεγάλο, δεν περιμένει κανείς σημαντική διαφοροποίηση αν κάθε στοιχείο που επιλέγεται επανατοποθετείται πριν την επόμενη επιλογή.

Είναι επίσης ενδιαφέρον να παρατηρηθεί ότι η σχέση της διασποράς της υπεργεωμετρικής με την διασπορά της διωνυμικής κατανομής οδηγεί στο συμπέρασμα ότι η δειγματοληψία χωρίς

επανάθεση δίνει ακριβέστερα αποτελέσματα (μικρότερη διασπορά) από ότι η δειγματοληψία με επανάθεση.

Παράδειγμα: Ο αριθμός των ενηλίκων κατοίκων μιας πόλης είναι 75.000, από τους οποίους 500 είναι οικονομολόγοι. Σε μια δειγματοληπτική έρευνα γίνεται μια τυχαία επιλογή 25 ενηλίκων χωρίς επανάθεση. Να υπολογισθεί η πιθανότητα το δείγμα αυτό να περιλαμβάνει το πολύ ένα οικονομολόγο.

Λύση: Έστω X ο αριθμός των οικονομολόγων στο δείγμα. Τότε

$$X \sim h(x;75000, 25, ,500)$$

Επομένως,

$$P(X=x) = \binom{500}{x}\binom{74500}{25-x}\Big/\binom{75000}{25}, \qquad x=0,1,2,...$$

Η ζητούμενη πιθανότητα είναι

$$P(X\le 1) = P(X=0)+P(X=1) = \left[\binom{74500}{25} + 500\binom{74500}{24}\right]\Big/\binom{74500}{25}$$

$$= 0.98798$$

Επειδή η τιμή του N είναι πολύ μεγάλη σε σχέση με την τιμή του n μπορούμε να χρησιμοποίησουμε και την διωνυμική προσέγγιση της υπεργεωμετρικής κατανομής.

Έτσι $p = \frac{m}{N} = \frac{500}{75000} = \frac{1}{150}$ και επομένως

$$P(X=x) \approx \binom{25}{x}\left(\frac{1}{150}\right)^{x}\left(\frac{149}{150}\right)^{25-x} \quad , \qquad x=0,1,2,....$$

και

$$P(X\le 1) \approx b\left(0;25;\frac{1}{150}\right)+b\left(1;25;\frac{1}{150}\right)$$

$$=\left(\frac{149}{150}\right)^{25}+\left(\frac{25}{150}\right)\left(\frac{149}{150}\right)^{24}=0.98796$$

Παρατηρούμε δηλαδή ότι η προσέγγιση είναι πάρα πολύ ικανοποιητική.

Η ΚΑΤΑΝΟΜΗ POISSON

Ορισμός: Έστω X μία διακριτή τυχαία μεταβλητή με τιμές 0,1,2,... θα λέμε ότι η τυχαία μεταβλητή X ακολουθεί την *κατανομή Poisson με παράμετρο* λ και θα γράφουμε X~P(x;λ) αν

$$P(X=x) = e^{-\lambda}\frac{\lambda^{x}}{x!}, \qquad x = 0,1,2,... \ , \qquad \lambda>0$$

Παρατήρηση: Η κατανομή Poisson είναι μία καλά ορισμένη κατανομή μια και P(x)≥0 και

$$\sum_{x=0}^{\infty}P(x)=e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^{x}}{x!}=e^{-\lambda}e^{\lambda}=1$$

Ιδιότητες:

α) $E(X) = \lambda$

β) $\Delta(X) = \lambda$

Μοντέλα που οδηγούν στην κατανομή Poisson

Η κατανομή Poisson ως νόμος των σπανίων γεγονότων

Η κατανομή Poisson είναι η πιο συχνά χρησιμοποιούμενη κατανομή για την περιγραφή του αριθμού των "γεγονότων" ή "σημείων" για κάποια περιοχή χώρου ή χρόνου όταν η κατανομή και η "εμφάνιση" τέτοιων γεγονότων γίνεται με τυχαίο τρόπο. Για

παράδειγμα, ας πάρουμε την απλούστερη περίπτωση όπου τα "γεγονότα" κατανέμονται τυχαία στο διάστημα $(-\infty,+\infty)$ με τρόπο ώστε:

1) Οι αριθμοί των "γεγονότων" που συμβαίνουν σε δύο ξένα μεταξύ τους διαστήματα κατανέμονται ανεξάρτητα ο ένας από τον άλλο.
2) Ο αναμενόμενος αριθμός "γεγονότων" σε ένα πεπερασμένο διάστημα I είναι πεπερασμένος και ανάλογος του μήκους του διαστήματος (έστω $\lambda|I|$, $\lambda>0$).
3) Η πιθανότητα να συμβούν περισσότερα από ένα "γεγονότα" στο I τείνει στο 0 ταχύτερα από ότι το $|I|$ όταν $|I|\to 0$. (Εναλλακτικά, η συνθήκη αυτή καθορίζει ότι η πιθανότητα περισσότερων από ένα "γεγονότων" σε κάποιο διάστημα που το μήκος του τείνει στο μηδέν είναι μηδέν). Όταν οι παραπάνω συνθήκες ικανοποιούνται λέμε ότι το υπό συζήτηση φαινόμενο ακολουθεί την *στοχαστική ανέλιξη Poisson με παράμετρο λ*.

Έστω X ο αριθμός των γεγονότων στο διάστημα (0,t). Για να βρούμε την κατανομή του X διαιρούμε το διάστημα (0,t) σε n υποδιαστήματα ίσου μήκους $\frac{t}{m}$.

Σύμφωνα με τα προηγούμενα,

$$P(\text{ένα "γεγονός" στο διάστημα } \frac{t}{n}) =$$

$$= \lambda \frac{t}{n} P(X>1 \text{ στο διάστημα } \frac{t}{n}) = 0$$

Προφανώς ένα "γεγονός" ή συμβαίνει ή δεν συμβαίνει στο διάστημα $\frac{t}{n}$. Επομένως, η πραγματοποίηση ενός "γεγονότος" σε ένα υποδιάστημα μπορεί να χαρακτηρισθεί σαν μία δοκιμή Bernoulli. Λόγω δε της συνθήκης (1) έχουμε μία ακολουθία x δοκιμών Bernoulli με πιθανότητα "επιτυχίας" $\frac{\lambda t}{n}$.

Μας ενδιαφέρει ο καθορισμός της πιθανότητας x επιτυχιών. Επομένως,

$$P(X = x) = \begin{bmatrix} n \\ x \end{bmatrix} \left[\frac{\lambda t}{n}\right]^x \left[1 - \frac{\lambda t}{n}\right]^{n-x}$$

Αν $n \to \infty$ έχουμε

$$\lim_{n\to\infty} \begin{bmatrix} n \\ x \end{bmatrix} \left[\frac{\lambda t}{n}\right]^x \left[1 - \frac{\lambda t}{n}\right]^{n-x}$$

$$= \lim_{n\to\infty} \frac{n(n-1)\ldots(n-x+1)}{n^x} \frac{(\lambda t)^x}{x!} \left[1 - \frac{\lambda t}{n}\right]^n \left[1 - \frac{\lambda t}{n}\right]^{-x}$$

Για καθορισμένο x έχουμε

$$= \lim_{n\to\infty} \frac{n(n-1)\ldots(n-x+1)}{n^x}$$

$$= \lim_{n\to\infty} \left[1\left[1-\frac{1}{n}\right]\left[1-\frac{2}{n}\right]\ldots\left[1-\frac{x-1}{n}\right]\right] = 1$$

Επίσης,

$$\lim_{n\to\infty}\left[1-\frac{\lambda}{n}\right]^n = e^{-\lambda} \quad \text{και} \qquad \lim_{n\to\infty}\left[1-\frac{\lambda}{n}\right]^{-x} = 1$$

Επομένως,

$$\lim_{n\to\infty} P(X = x) = e^{-\lambda t}\frac{(\lambda t)^x}{x!}, \quad x = 0,1,2,\ldots$$

Αν πάρουμε t=1 (δηλαδή θεωρήσουμε ένα διάστημα μήκους 1) τότε

$$P(X = x) \approx e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0,1,2,\ldots$$

Παρατήρηση: Από την προηγούμενη συζήτηση προκύπτει ότι η παράμετρος λ της κατανομής Poisson εκφράζει τον μέσο αριθμό των "γεγονότων" στην μονάδα του χρόνου.

Το προηγούμενο μοντέλο που οδηγεί στην κατανομή Poisson δίνει την δυνατότητα διατύπωσης και του παρακάτω θεωρήματος που αναφέρεται στην χρησιμοποίηση της κατανομής Poisson ως ορίου της διωνυμικής κατανομής.

Θεώρημα: (Η κατανομή Poisson ως προσέγγιση της διωνυμικής κατανομής).
Έστω $X \sim b(x;n,p)$. Αν $n \to \infty$ και $p \to 0$ έτσι ώστε $np=\lambda$, όπου λ σταθερά, τότε

$$P(X = x) \approx e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0,1,2,...$$

Απόδειξη: Η απόδειξη προκύπτει από τα προηγούμενα.

Σημείωση: Για τον υπολογισμό διαδοχικών πιθανοτήτων από την κατανομή Poisson μπορεί να χρησιμοποιηθεί ο αναγωγικός τύπος

$$\frac{P(x)}{P(x-1)} = \frac{\lambda}{x}$$

Παραδείγματα: Η κατανομή Poisson έχει πάρα πολλές εφαρμογές. Ενδεικτικά αναφέρονται τομείς όπου η κατανομή έχει εφαρμοσθεί με επιτυχία.

1) Αριθμός ατυχημάτων σε ένα ορισμένο χρονικό διάστημα σε μία συγκεκριμένη περιοχή.
2) Αριθμός αιτήσεων αποζημίωσης σε ασφαλιστική εταιρία σε κάποιο χρονικό διάστημα.
3) Αριθμός μειοδοτών σε κάποιο μειοδοτικό διαγωνισμό.
4) Αριθμός τηλεφωνημάτων που φθάνουν σε ένα τηλεφωνικό κέντρο σε μία ορισμένη χρονική περίοδο της ημέρας.
5) Αριθμός διασπάσεων χρωμοσωμάτων στα κύτταρα που δημιουργούν οι ακτίνες Χ.
6) Αριθμός των τυπογραφικών λαθών σε μία σελίδα βιβλίου.

Η κατανομή Poisson δεν αναφέρεται μόνο σε αριθμό "γεγονότων" στον χρόνο. Μία άλλη εφαρμογή της είναι στην κατανομή αντικειμένων στο χώρο. Έστω, για παράδειγμα, ότι οι οργανισμοί κατανέμονται τυχαία μέσα σε ένα υγρό όγκο V, έτσι ώστε η πιθανότητα κάποιος οργανισμός να βρίσκεται σε μία συγκεκριμένη σταγόνα όγκου D είναι D/V. Η παρουσία των n οργανισμών μέσα ή έξω από την σταγόνα μπορεί να θεωρηθεί σαν

μία ακολουθία n ανεξάρτητων δοκιμών με σταθερή πιθανότητα "επιτυχίας" D/V σε κάθε δοκιμή. Στην πράξη το n είναι συνήθως πολύ μεγάλο ενώ το D/V είναι πολύ μικρό. Επομένως, η κατανομή του αριθμού X των οργανισμών σε μία σταγόνα όγκου D μπορεί να προσεγγισθεί με την κατανομή Poisson με μέση τιμή λD όπου $\lambda = \frac{n}{V}$ είναι ο μέσος αριθμός οργανισμών ανά μονάδα όγκου του διαλύματος

$$P(x) = e^{-\lambda D} \frac{(\lambda D)^x}{x!}, \; x = 0,1,2,\ldots$$

Παράδειγμα: Τα ελαττώματα κατασκευής σε μεγάλα φύλλα ενός μετάλλου εμφανίζονται τυχαία και με συχνότητα, κατά μέσο όρο, 2.56 ανά 100 τετραγωνικά μέτρα. (α) Να υπολογισθεί η πιθανότητα ανά φύλλο διαστάσεων 4μ x 8μ να μην υπάρχουν καθόλου ελαττώματα. (β) Πόσα φύλλα της μορφής αυτής από μία παρτίδα των 100 αναμένεται να έχουν δύο ή περισσότερα ελαττώματα;

Λύση: Έστω X ο αριθμός των ελαττωμάτων ανά φύλλο διαστάσεων 4μ x 8μ. Μπορεί να θεωρηθεί ότι το X ακολουθεί την κατανομή Poisson με παράμετρο $\lambda = 2.56 \times 0.32 = 0.819$.
Επομένως,
α) $P(X=0) = e^{-0.819} = 0.4408$
β) $P(X \geq 2) = 1-P(X=0) - P(X=1) = 1- e^{-0.819}- 0.819\, e^{-0.819}$.
Επομένως ο αναμενόμενος αριθμός είναι 19.82.

Η ΓΕΩΜΕΤΡΙΚΗ ΚΑΤΑΝΟΜΗ

Στον ορισμό της διωνυμικής κατανομής είδαμε ότι η κατανομή αυτή μπορεί να προέλθει από μια ακολουθία δοκιμών Bernoulli όπου ο αριθμός των δοκιμών n είναι σταθερός και μας ενδιαφέρει ο αριθμός X των επιτυχιών στις n αυτές δοκιμές.

Θα ασχοληθούμε τώρα με μια διαφορετική θεώρηση ακολουθίας δοκιμών Bernoulli. Στην θεώρηση αυτή ο αριθμός των δοκιμών Bernoulli δεν είναι προκαθορισμένος. Αυτό γιατί μας ενδιαφέρει να μελετήσουμε όχι τον αριθμό των επιτυχιών, αλλά τον αριθμό των αποτυχιών X που θα συναντήσουμε μέχρις ότου

φθάσουμε στην k επιτυχία. Η θεώρηση αυτή οδηγεί στην γεωμετρική κατανομή και στην γενίκευσή της, την αρνητική διωνυμική κατανομή.

Ορισμός: Έστω X μια διακριτή τυχαία μεταβλητή. Θα λέμε ότι η X ακολουθεί την *γεωμετρική κατανομή (geometric distribution) με παράμετρο* p και θα συμβολίζουμε με X~G(x;p) αν

$$P(X=x)=pq^x, \quad x=0,1,2,\ldots, \quad 0<p<1, \quad q=1-p$$

Παρατήρηση: Η γεωμετρική κατανομή είναι μια καλά ορισμένη κατανομή γιατί

$$P(x)\geq 0 \text{ και } \sum_x P(x) = p\sum_x q^x = 1$$

Πρόταση: $E(X) = \frac{q}{p}$ $\Delta(X) = \frac{q}{p^2}$.

Απόδειξη: Η απόδειξη είναι εύκολη αν κανείς κάνει χρήση του ορισμού της E(X) και στην συνέχεια της $E(X^2)$ και του γεγονότος ότι $\Delta(X)=E(X^2)-\{E(X)\}^2$.

Παρατήρηση: Ο όρος γεωμετρική κατανομή προέρχεται από το γεγονός ότι οι πιθανότητές της αποτελούν τους όρους μιας γεωμετρικής σειράς με λόγο q=1-p.

Μοντέλα που οδηγούν στην γεωμετρική κατανομή

i) Θεωρούμε μία ακολουθία δοκιμών Benoulli. Έστω X ο αριθμός των αποτυχιών πριν εμφανισθεί η πρώτη επιτυχία. Τότε

$$P(X=x) = p\,(1-p)^x, \qquad x = 0,1,2,\ldots$$

Απόδειξη: Προφανής αφού μιλάμε για ακολουθία δοκιμών Bernoulli.

ii) Μία εναλλακτική παρουσίαση στην βιβλιογραφία της γεωμετρικής κατανομής είναι ως κατανομή του αριθμού Y των δοκιμών που απαιτούνται για να επιτευχθεί η πρώτη επιτυχία σε μία ακολουθία δοκιμών Bernoulli. Στην περίπτωση αυτή

$$P(Y=y) = p\,(1-p)^{y-1}, \qquad y = 1,2,\ldots$$

Η ισοδυναμία των μοντέλων (i) και (ii) είναι προφανής μια και Y=X+1.

iii) *Δειγματοληψία από υδρία:* Σε μία υδρία υπάρχουν άσπρα και μαύρα σφαιρίδια σε αναλογία p μαύρα και 1-p άσπρα. Επιλέγουμε στην τύχη σφαιρίδια από την υδρία με επανάθεση. (Σε κάθε επιλογή σημειώνουμε το χρώμα του σφαιριδίου και το επανατοποθετούμε στην υδρία πριν από την επόμενη επιλογή). Ο αριθμός των άσπρων σφαιριδίων που θα επιλεγούν πριν επιλεγεί το πρώτο μαύρο σφαιρίδιο ακολουθεί την γεωμετρική κατανομή με παράμετρο p.

Για την γεωμετρική κατανομή έχουμε, για κάθε θετικό ακέραιο α

$$P(X < \alpha) = \sum_{x=0}^{\alpha-1} P(x) = \sum_{x=0}^{\alpha-1} pq^x = 1 - \sum_{x=\alpha}^{\infty} pq^x = 1 - \frac{pq^\alpha}{1-q} = 1 - q^\alpha$$

Επίσης,

$$P(X \geq \alpha) = 1 - P(X < \alpha) = q^\alpha$$

Παράδειγμα. ***(Ρωσική ρουλέτα):*** Έστω ότι έχουμε ένα εξάσφαιρο περίστροφο το οποίο έχει μόνο μία σφαίρα.

i) Να βρεθεί η πιθανότητα να εκπυρσοκροτήσει το περίστροφο με την πρώτη δοκιμή.

ii) Να βρεθεί η πιθανότητα να εκπυρσοκροτήσει το περίστροφο πριν την τέταρτη δοκιμή.

iii) Να βρεθεί οαριθμός των δοκιμών που απαιτούνται για να εκπυρσοκροτήσει το περίστροφο.

Λύση: Έστω X ο αριθμός των δοκιμών (αποτυχίων) πριν εκπυρσοκροτήσει (επιτυχία) το περίστροφο.
Είναι

i) $P(X = 0) = p(1-p)^0 = \frac{1}{6}.$

ii) $P(X < 3) = 1 - p^3 = 1 - \left[\frac{5}{6}\right]^3.$

iii) $E(X+1) = E(X) + 1 = \frac{q}{p} + 1 = \frac{5/6}{1/6} + 1 = 6.$

Άλλα παραδείγματα

1) **Ιατρική.** Μία τράπεζα αίματος χρειάζεται αίμα ομάδας Β ρέζους αρνητικού και συνεχίζει να αγοράζει αίμα από ιδιώτες μέχρις ότου εμφανισθεί κάποιος με αυτή την ομάδα αίματος. Αν οι αγορές αίματος γίνονται ανεξάρτητα η μία από την άλλη, ο αριθμός Χ των αγορών που θα γίνουν πριν εμφανισθεί κάποιος με την συγκεκριμένη ομάδα αίματος ακολουθεί την γεωμετρική κατανομή.

2) **Τυχερά παιχνίδια.** Ενας παίκτης ρουλέτας στοιχηματίζει το ίδιο ποσόν α στον ίδιο αριθμό μέχρις ότου κερδίσει για πρώτη φορά. Αν οι στροφές της ρουλέτας γίνονται ανεξάρτητα η μία από την άλλη, ο αριθμός Χ των φορών που ο παίκτης θα χάσει πριν κερδίσει για πρώτη φορά ακολουθεί την γεωμετρική κατανομή.

Σημείωση: Στο κεφάλαιο 4 δόθηκε ο ορισμός της μέσης τιμής. Στον ορισμό αυτό τονίστηκε ότι ο ορισμός έχει έννοια εφόσον η σειρά που ορίζει την μέση τιμή συγκλίνει. Η γεωμετρική κατανομή μας δίνει την δυνατότητα παρουσίασης του εξής παραδείγματος όπου η μέση τιμή δεν υπάρχει.

Παράδειγμα: Έστω ότι σε ένα τυχερό παιχνίδι ο παίκτης έχει πιθανότητα $p = \frac{1}{2}$ να κερδίσει μία παρτίδα του παιχνιδιού. Έστω ότι ο παίκτης ακολουθεί ένα "σύστημα" σύμφωνα με το οποίο διπλασιάζει το ποσόν το οποίον στοιχηματίζει μέχρις ότου κερδίσει για πρώτη φορά, ενώ ξεκινά με στοίχημα μιας δραχμής. (Έτσι αν χάσει την πρώτη φορά, την δεύτερη στοιχηματίζει δύο δραχμές, την τρίτη 4 δραχμές κ.ο.κ.). Το σύστημα αυτό δίνει μια σίγουρη μέθοδο για να πάρει ο παίκτης πίσω ότι έχει χάσει και επιπλέον να κερδισει μία δραχμή.
Να βρεθεί το ποσόν που ο παίκτης πρέπει να έχει διαθέσιμο ώστε μα είναι σε θέση να διατηρήσει το σύστημα με το οποίο στοιχηματίζει.

Λύση: Η πιθανότητα να κερδίσει για πρώτη φορά ο παίκτης στην k δοκιμή είναι

$$P(X = k-1) = \left(\frac{1}{2}\right)^k, \qquad k = 1,2,\ldots$$

Έστω Z το ποσόν που ο παίκτης χρειάζεται για να είναι σε θέση να συνεχίσει το παιχνίδι μέχρις ότου κερδίσει για πρώτη φορά. Το ποσόν που χρειάζεται για να είναι σε θέση να παίξει μέχρι και την k παρτίδα είναι $1+2+4+\ldots+2^{k-1} = 2^k - 1$. Επομένως, η Z είναι μια τυχαία μεταβλητή που παίρνει τις τιμές $2^k - 1$, $k = 1,2,\ldots$. Η πιθανότητα να χρειασθεί ο παίκτης $2^k - 1$ δραχμές για να κερδίσει για πρώτη φορά είναι η ίδια με την πιθανότητα να κερδίσει για πρώτη φορά στην k δοκιμή. Δηλαδή,

$$P(Z = 2^k - 1) = P(X = k - 1) = \left(\frac{1}{2}\right)^k, \qquad k = 1,2,\ldots$$

Επομένως, το ποσόν που πρέπει, κατά μέσο όρο, να έχει ο παίκτης στην διάθεση του για να είναι σε θέση να συνεχίσει να παίζει μέχρις ότου κερδίσει είναι

$$E(Z) = \sum_{k=1}^{\infty} (2^k - 1)\frac{1}{2^k} = \frac{1}{2} + \frac{3}{4} + \frac{7}{8} + \frac{15}{16} + \ldots = \infty$$

Δηλαδή, κατά μέσο όρο, δεν υπάρχει πεπερασμένο ποσό χρημάτων αρκετό να υποστηρίξει το σύστημα αυτού του παιχνιδιού.

Η ΑΡΝΗΤΙΚΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Έστω X μια διακριτή τυχαία μεταβλητή. Θα λέμε ότι η X *ακολουθεί την αρνητική διωνυμική κατανομή* (η αλλιώς την κατανομή Pascal) *με παραμέτρους* p *και* r και θα συμβολίζουμε με $X \sim NB(x; r, p)$, αν

$$P(X = x) = \binom{x + r - 1}{x} p^r q^x = \binom{x + r - 1}{r - 1} p^r q^x \qquad x = 0,\ 1,2,\ldots$$

$r = 1,\ 2,\ldots$

$0<p<1,\ q=1-p$

Παρατήρηση: Ο ορισμός αυτός είναι ο κλασσικός ορισμός της αρνητικής διωνυμικής κατανομής. Είναι όμως δυνατόν να επεκταθεί έτσι ώστε να περιλαμβάνει και μη ακέραιες τιμές της παραμέτρου r. Αυτό γίνεται με την χρήση του γενικευμένου ορισμού του διωνυμικού συντελεστή σύμφωνα με τον οποίο για κάθε πραγματικό αριθμό α και για κάθε μη αρνητικό ακέραιο x

$$\binom{\alpha}{x} = \frac{\alpha_{(x)}}{x!} = \frac{\alpha(\alpha-1)...(\alpha-x+1)}{x!}$$

Είναι εύκολο να διαπιστωθεί ότι η αρνητική διωνυμική κατανομή είναι μία καλά ορισμένη κατανομή. Αυτό γιατί

$$P(x) \geq 0, \qquad x = 0,1,2,\ldots \qquad \text{και}$$

$$\sum_{x=0}^{\infty} P(x) = \sum_{x=0}^{\infty} \binom{x+r-1}{x} p^r q^x = p^r \sum_{x=0}^{\infty} \binom{x+r-1}{x} q^x = p^r (1-q)^{-r} = 1$$

Ο παραπάνω τύπος είναι αποτέλεσμα της γενίκευσης του διωνυμικού θεωρήματος όπου για $0 < q < 1$

$$(1-q)^{-r} = \sum_{x=0}^{\infty} \binom{-r}{x} (-q)^x = \sum_{x=0}^{\infty} \binom{x+r-1}{x} q^x$$

μια και από την την γενίκευση του διωνυμικού συντελεστή έχουμε για κάθε πραγματικό r

$$\binom{-r}{x} = \frac{(-r)(-r-1)...(-r-x-1)}{x!} = \frac{(-1)^x r(r+1)...(r+x-1)}{x!}$$

$$= \frac{(-1)^x (r-1)!r(r+1)...(r+x-1)}{x!(r-1)!} = \frac{(-1)^x (r+x-1)!}{x!(r-1)!}$$

$$= (-1)^x \binom{r+x-1}{x}$$

Μοντέλα που οδηγούν στην αρνητική διωνυμική κατανομή

i) Έστω X ο αριθμός των αποτυχιών μέχρις ότου εμφανισθούν r επιτυχίες σε μια ακολουθία δοκιμών Bernoulli. Τότε το X ακολουθεί την αρνητική διωνυμική κατανομή με παραμέτρους r και p.

Απόδειξη: Το ενδεχόμενο {X=x} είναι ισοδύναμο με το ενδεχόμενο {σε x+r δοκιμές r επιτυχίες με τρόπο ώστε η τελευταία επιτυχία να πραγματοποιηθεί στην x+r δοκιμή}. Το ενδεχόμενο αυτό είναι ισοδύναμο με το ενδεχόμενο {r+1 επιτυχίες και x αποτυχίες με οποιαδήποτε σειρά στις πρώτες x+r-1 δοκιμές και επιτυχία στην x+r δοκιμή}. Επειδή οι δοκιμές είναι ανεξάρτητες

$$P(X = x) = \binom{x + r - 1}{r - 1} p^{r-1} q^x p = \binom{x + r - 1}{r - 1} p^r q^x$$

Σημείωση: Από τον τύπο της αρνητικής διωνυμικής κατανομής, αλλά και από το προηγούμενο μοντέλο προκύπτει ότι για r=1 η αρνητική διωνυμική είναι η γεωμετρική κατανομή.

ii) *Δειγματοληψία από υδρία:* Σε μία υδρία υπάρχουν άσπρα και μαύρα σφαρίδια σε αναλογία (ποσοστό) p μαύρα και 1-p=q άσπρα. Ο αριθμός των άσπρων ασφαιριδίων που θα επιλεγούν μέχρις ότου επιλεγούν μαύρα σφαρίδια ακολουθεί την αρνητική διωνυμική κατανομή με παραμέτρους p και r. Το μοντέλο αυτό δικαιολογεί και την ονομασία της κατανομής ως αρνητική διωνυμική. Αυτό γιατί όπως η διωνυμική έτσι και η αρνητική διωνυμική κατανομή προκύπτει από δειγματοληψία με επανάθεση (ακολουθία δοκιμών Bernoulli). Στην διωνυμική όμως ο αριθμός n των δοκιμών είναι καθορισμένος ενώ ο αριθμός r των επιτυχιών είναι τυχαία μεταβλητή, ενώ αντίθετα στην αρνητική διωνυμική ο αριθμός r των επιτυχιών είναι καθορισμένος ενώ ο αριθμός των αποτυχιών (ισοδύναμα των δοκιμών) είναι τυχαία μεταβλητή.

Μια ισοδύναμη παρουσίαση της αρνητική διωνυμικής μπορεί να γίνει μέσω του αριθμού των δοκιμών που απαιτούνται σε μια ακολουθία Bernoulli για να παρατηρηθούν r επιτυχίες. Έστω Y ο αριθμός αυτός. Τότε

$$P(Y = y) = \binom{y-1}{r-1} p^r q^{y-r}, \qquad y=r, r+1, \ldots$$

$r=1,2,\ldots$

$0<p<1$, $q=1-p$

Προφανώς $Y=X+r$.

iii) *Η αρνητική διωνυμική ως άθροισμα γεωμετρικών τυχαίων μεταβλητών*. Έστω, X_1, X_2, …, X_r μια ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλητών που ακολουθούν την γεωμετρική κατανομή με παράμετρο p. Θεωρούμε την τυχαία μεταβλητή

$$X= X_1+X_2+\ldots+ X_r.$$

Η τυχαία μεταβλητή X ακολουθεί την αρνητική διωνυμική κατανομή με παραμέτρους r και p.

Απόδειξη: Η απόδειξη μπορεί να στηριχθεί στον συλλογισμό ότι ο αριθμός των αποτυχιών X_1 μέχρι την πρώτη αποτυχία ακολουθεί την γεωμετρική κατανομή. Το ίδιο συμβαίνει και με τον αριθμό των αποτυχιών X_2 που μεσολαβούν από την πρώτη μέχρι την δεύτερη επιτυχία και γενικά τον αριθμό των αποτυχιών X_i που μεσολαβούν από την i-1 έως την i επιτυχία (i=1,2,…,r). Επομένως, $X_1+X_2+\ldots+ X_r$ είναι ο συνολικός αριθμός των αποτυχιών μέχρις ότου εμφανισθεί η r επιτυχία.

Πρόταση: Αν η τυχαία μεταβλητή X ακολουθεί την αρνητική διωνυμική κατανομή, τότε

$$E(X) = \frac{rq}{p} \qquad \Delta(X) = \frac{rq}{p^2}$$

Απόδειξη: Η απόδειξη είναι εύκολη αν κάνει κανείς χρήση της έκφρασης της X ως $X_1+X_2+\ldots+ X_r$, όπου X_i, i=1,2,…,r ανεξάρτητες και ισόνομες γεωμετρικές τυχαίες μεταβλητές.

Παράδειγμα: Μια γραμματέας κάνει, κατά μέσο όρο, δύο τυπογραφικά λάθη ανά σελίδα. Σελίδες με περισσότερα από δύο τυπογραφικά λάθη πρέπει να ξαναγραφούν. Πόσες σελίδες συνολικά αναμένεται να δακτυλογραφήσει ώστε ένα κείμενο 100 σελίδων, να είναι αποδεκτό;

Λύση: Έστω Z ο αριθμός των σελίδων, X ο αριθμός των αποτυχιών πριν την 100στη αποδεκτή σελίδα και Y ο αριθμός των λαθών ανά σελίδα. Είναι λογικό από τις συνθήκες του προβλήματος να υποθέσει κανείς ότι το X ακολουθεί την κατανομή Poisson με λ=2 και το Y την αρνητική διωνυμική κατανομή με r=100. Έτσι

$$p = P(Y\le 2) = P(Y=0) + P(Y=1) + P(Y=2)$$

$$= e^{-2} + 2e^{-2} + e^{-2}\frac{2^2}{2!} = 5e^{-2}$$

Άρα ο αριθμός των σελίδων που αναμένεται να δακτυλογραφηθούν είναι

$$E(Z) = E(X+100) = E(X)+100 = \frac{100(1-5e^{-2})}{5e^{-2}} = 47.8+100 = 147.8.$$

Η ΠΟΛΥΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ

Ορισμός: Έστω $X_1, X_2, \ldots, X_k$ μια ακολουθία διακριτών τυχαίων μεταβλητών. Θα λέμε ότι το διάνυσμα $\underset{\sim}{X} = (X_1, X_2, \ldots, X_k)$ ακολουθεί την *πολυωνυμική κατανομή με παραμέτρους $p_1, p_2, \ldots, p_k$ και n*, αν

$$P(X_1=x_1, X_2=x_2, \ldots, X_k=x_k) = \frac{n!}{x_1!x_2!\ldots x_k!}p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}$$

με $\sum_{i=1}^{k} x_i = n$, $0<p_i<1$, $i=1,2,\ldots k$, $\sum_{i=1}^{k} p_i = 1$, $x=0,1,2,\ldots$

Παρατήρηση: Η πολυωνυμική κατανομή είναι μια καλά ορισμένη κατανομή μια και $P(\underset{\sim}{X}=\underset{\sim}{x})>0$ και

$$\sum_{x_1}\sum_{x_2}\ldots\sum_{x_k} P(X_1=x_1, X_2=x_2,\ldots,X_k=x_k)$$
$$\sum x_i = n$$

$$= \sum_{x_1}\sum_{x_2}\ldots\sum_{x_k}\binom{n}{x_1, x_2,\ldots, x_k}p_1^{x_1}p_2^{x_2}\ldots p_k^{x_k}$$
$$\sum x_i = n$$

$$=(p_1+p_2+\ldots+p_k)^n = 1.$$

Μοντέλα που οδηγούν στην πολυωνυμική κατανομή

Η πολυωνυμική κατανομή είναι φυσιολογική επέκταση της διωνυμικής κατανομής. Έτσι όλα τα μοντέλα της διωνυμικής μπορούν να επεκταθούν και να δώσουν την πολυωνυμική κατανομή. Για παράδειγμα, έστω ότι έχουμε μια ακολουθία ανεξάρτητων δοκιμών, κάθε μια από τις οποίες μπορεί να καταλήξει σε ένα από k δυνατά ενδεχόμενα $A_1,A_2, \ldots, A_k$ αμοιβαία ξένα μεταξύ τους όπου $A_1, A_2, \ldots, A_k$ μια διαμέριση του δειγματικού χώρου. Έστω $p_i=P(A_i)$, i=1,2,…k η πιθανότητα του ενδεχομένου A_i και έστω ότι το p_i παραμένει σταθερό από δοκιμή σε δοκιμή και $\sum_{i=1}^{n} p_i = 1$. Αν X_i είναι ο αριθμός εμφανίσεων του ενδεχομένου A_i σε n δοκιμές, η από κοινού συνάρτηση κατανομής των $X_1, X_2, \ldots, X_k$ είναι η πολυωνυμική με παραμέτρους $p_1, p_2, \ldots, p_k$ και n. Στα πλαίσια μοντέλων δειγματοληψίας, η πολυωνυμική κατανομή προκύπτει από δειγματοληψία με επανάθεση από υδρία που περιέχει σφαιρίδια k χρωμάτων σε αναλογία $p_1, p_2, \ldots, p_k$.

Περιθώριες κατανομές: Αν στην πολυωνυμική κατανομή ενδιαφερόμαστε μόνο για την πραγματοποίηση ή όχι του ενδεχομένου A_i, μπορούμε να θέσουμε $p_i=P(A_i)$ και $q_i=1-p_i=1-\sum_{j\neq 1} P(A_j)$. Στην περίπτωση αυτή έχουμε ακολουθία δοκιμών Bernoulli οπότε η πιθανότητα x_i επιτυχιών δίνεται από τον τύπο

$$P(X_i = x_i) = \binom{n}{x_i} p_i^{x_i} (1-p_i)^{n-x_i}, \quad i=1,2,\ldots,k, \quad x=0,1,2,\ldots,n$$

Επομένως, η περιθώρια κατανομή μιας μόνο μεταβλητής X_i είναι διωνυμική με παραμέτρους n και p_i. Στο αποτέλεσμα αυτό μπορούμε να φθάσουμε και με αλγεβρικές μεθόδους αθροίζοντας την από κοινού συνάρτηση πιθανότητας ως προς όλες τις άλλες μεταβλητές εκτός της X_i.

Παράδειγμα. ***(Νόμος των Hardy-Weinberg):*** Στον νόμο των Hardy-Weinberg (παράδειγμα κεφ. 4) είχαμε υπολογίσει ότι οι πιθανότητες (ποσοστά) με τις οποίες τα τρία είδη γονοτύπων ΑΑ, Αα και αα συναντώνται σε ένα πληθυσμό είναι

$$P(AA)=p^2,\quad P(A\alpha)=2p(1-p) \text{ και } P(\alpha\alpha)=(1-p)^2$$

όπου p είναι η πιθανότητα μεταφοράς του γονιδίου Α στον απόγονο. Έστω ότι επιλέγουμε τυχαία οκτώ άτομα από ένα πληθυσμό θέλοντας να καθορίσουμε τα γονότυπά τους. Να υπολογισθούν, συναρτήσει του p, οι πιθανότητες ότι

(α) Δεν υπάρχουν γονότυπα της μορφής ΑΑ στο δείγμα.

(β) Υπάρχουν δύο ΑΑ, τέσσερα Αα και δύο αα.

(γ) Ποιά είναι η τιμή του p που δίνει την μέγιστη τιμή στην πιθανότητα του ερωτήματος (β);

Λύση: Έστω X_1 ο αριθμός των ΑΑ, X_2 των Αα και X_3 των αα στο δείγμα. Τότε το τυχαίο διάνυσμα (X_1, X_2, X_3) ακολουθεί την πολυωνυμική κατανομή.

(α) $P(\text{μηδέν ΑΑ στο δείγμα})=P(X_1=0)=\binom{8}{0}(p^2)^0(1-p^2)^8=(1-p^2)^8$

(β) $P(X_1=2, X_2=4, X_3=2)=\frac{8!}{2!4!2!}(p^2)^2(2p(1-p))^4((1-p)^2)^2$

$$=6720p^8(1-p)^8$$

(γ) $\frac{\partial P(X_1=2,X_2=4,X_3=2)}{\partial p}=0 \Leftrightarrow p=\frac{1}{2}.$