Statistics and Confidence

- *Αναμενόμενη τιμή και τυπική απόκλιση*

Ας υποθέσουμε ότι ποντάρουμε 1 € κάθε φορά στην ρουλέτα. Πόσα θα έχουμε χάσει μετά από 10000 παιχνίδια;

Σύμφωνα με το κεντρικό οριακό θεώρημα, εάν λάβουμε *n* δείγματα με αντικατάσταση από ένα πληθυσμό, το άθροισμα τους θα ακολουθεί την κανονική κατανομή.

- Η αναμενόμενη τιμή του αθροίσματος θα είναι *n* φορές η αναμενόμενη τιμή κάθε δείγματος.
- Η τυπική απόκλιση του αθροίσματος θα είναι $\sqrt{n}$ φορές η τυπική απόκλιση κάθε δείγματος.

Κατά συνέπεια θα έχουμε:

Το καζίνο κερδίζει 20 στις 38 φορές στην ρουλέτα (μαύρο-κόκκινο). Κατά συνέπεια για κάθε ευρώ που θα παίζουμε, χάνουμε (20-18)/38 = 0,052 δηλαδή περίπου 5 λεπτά.

Στα 10000 παιχνίδια αναμένεται να χάσουμε 10000*0,05 δηλαδή περίπου 500 ευρώ.

Η τυπική απόκλιση για ένα παιχνίδι είναι √(20/38 * 1^2 + 18/38 * (-1)^2 - (-0,05)^2) δηλαδή περίπου ίση με 1 ευρώ.

Η τυπική απόκλιση για τα 10000 παιχνίδια θα είναι √10000 * 1 = 100 ευρώ.

Γνωρίζουμε ότι στο 95% των περιπτώσεων το ποσό που θα έχουμε χάσει θα περικλείεται μέσα σε δύο τυπικές αποκλείσεις από την αναμενόμενη τιμή. Κατά συνέπεια στο 95% των περιπτώσεων, μετά από 10000 παιχνίδια, θα έχουμε χάσει από 300 έως 700 ευρώ.

- *Κατανομές ποσοστών*

Αν υποθέσουμε ότι διενεργούμε δειγματοληψία από έναν πληθυσμό του οποίου ένα ποσοστό *p* έχει ένα συγκεκριμένο χαρακτηριστικό. Τότε στο δείγμα μας *n* στοιχείων (για *n* σχετικά μεγάλο > 30), το ποσοστό αυτών που θα φέρει το συγκεκριμένο χαρακτηριστικό ακολουθεί την κανονική κατανομή με αναμενόμενη τιμή p και τυπική απόκλιση √(p*(1-p)/n).

Εάν ρίξουμε ένα δίκαιο νόμισμα 100 φορές αναμένουμε το 50% αυτών να έρθει Heads με τυπική απόκλιση √(0,5*(1-0,5)/100) = 5%.

Φυσικά αν δεν γνωρίζουμε το ποσοστό p του πληθυσμού που φέρει το συγκεκριμένο χαρακτηριστικό τότε μπορούμε να εκτιμήσουμε το διάστημα που αυτό είναι πιθανό να κυμαίνεται από την αναμενόμενη τιμή και την τυπική απόκλιση του δείγματος.

© 2003 - 2006 gbosmis, web page: http://www.geocities.com/gbosmis

- *Κατανομές μέσων τιμών*

Έστω ότι θέλουμε να υπολογίσουμε την μέση τιμή του χαρακτηριστικού ενός πληθυσμού με βάση ένα δείγμα του *n*. Η μέση τιμή του χαρακτηριστικού αυτού στο δείγμα είναι η εκτιμήτρια που έχουμε για την μέση τιμή του χαρακτηριστικού στον πληθυσμό. Είναι η ίδια μια τυχαία μεταβλητή που ακολουθεί την κανονική κατανομή με τυπική απόκλιση $\sigma/\sqrt{n}$.

- *Έλεγχος υποθέσεων*

Θέλουμε να ελέγξουμε κατά πόσο ένα νευρωνικό δίκτυο που έχουμε κατασκευάσει, προβλέπει με ακρίβεια την κατεύθυνση της αγοράς. Το δίκτυο έχει διάφορα inputs και σαν output μας δίνει ποια τις 4 (υποθέτουμε για ευκολία ισοπίθανες) συμπεριφορές αναμένεται να εμφανίσει η αγορά στο επόμενο διάστημα

.

- Ισχυρά ανοδική
- Μέτρια ανοδική
- Μέτρια καθοδική
- Ισχυρά καθοδική

Το σύστημα προέβλεψε σωστά στις 30 από τις 100 περιπτώσεις που το δοκιμάσαμε. Μπορεί να οφείλεται αυτό καθαρά σε τύχη;

Ας πάρουμε την null hypothesis ότι η πρόβλεψη είναι τυχαία. Υπολογίζουμε το *z* που μας λέει πόσο μακριά είναι η παρατηρούμενη συμπεριφορά του δείγματος σε σχέση με την αναμενόμενη αν ίσχυε η null hypothesis. Το *z* ακολουθεί την κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1 ενώ ορίζεται ως εξής:

$$z = (\text{Observed} - \text{Expected}) / \text{Standard\_Error}$$

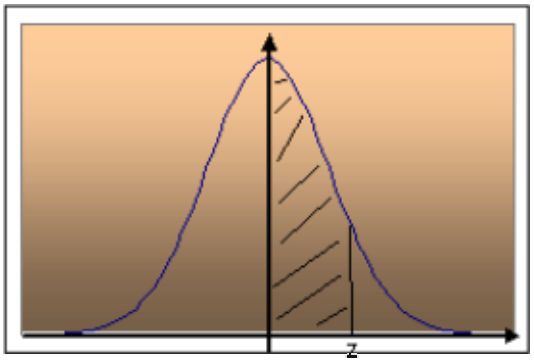
Στην περίπτωση μας αναμένουμε τυπική απόκλιση $\sqrt{(0{,}25*(1-0{,}25)/100)} = 4{,}3\%$. Κατά συνέπεια:

$$z = (30 - 25)/4{,}3 = 1{,}16$$

Η πιθανότητα το z να έχει τιμές μεγαλύτερες από 1,16 (σύμφωνα με τους πίνακες της κανονικής κατανομής) είναι περίπου 12,5%. Κατά συνέπεια δεν μπορούμε να απορρίψουμε την υπόθεση ότι το νευρωνικό μας δίκτυο στάθηκε απλά τυχερό.

© 2003 - 2006 gbosmis, web page: http://www.geocities.com/gbosmis

Appendix: Areas Under the Standard Normal Curve from 0 to z



z	Area
0	0,0000
0,1	0,0398
0,2	0,0793
0,3	0,1179
0,4	0,1554
0,5	0,1915
0,6	0,2257
0,7	0,2580
0,8	0,2881
0,9	0,3159
1	0,3413
1,1	0,3643
1,2	0,3849
1,3	0,4032
1,4	0,4192
1,5	0,4332
1,6	0,4452
1,7	0,4554
1,8	0,4641
1,9	0,4713
2	0,4772
2,1	0,4821
2,2	0,4861
2,3	0,4893
2,4	0,4918
2,5	0,4938
2,6	0,4953
2,7	0,4965
2,8	0,4974
2,9	0,4981
3	0,4987

Παραδείγματα:

1. Η πιθανότητα η τυχαία μεταβλητή να βρίσκεται στην ζώνη μεταξύ 0 και 1 τυπική απόκλιση είναι σύμφωνα με τον πίνακα 0,3413 ή αλλιώς 34,13%.

2. Η πιθανότητα η τυχαία μεταβλητή να βρίσκεται στην ζώνη μεταξύ (-1) και 1 τυπικής απόκλισης είναι 2*0,3413 = 0,6826 ή 68,26%.

3. Η πιθανότητα η τυχαία μεταβλητή να βρίσκεται στην ζώνη μεταξύ (-2) και 2 τυπικές αποκλίσεις είναι 2*0,4772 = 0,9544 ή 95,44%.

4. Η πιθανότητα η τυχαία μεταβλητή να βρίσκεται στην ζώνη μεταξύ 2 τυπικών αποκλίσεων και +∞ είναι 1/2 - 0,4772 = 0,0228 ή 2,28%.

5. Η πιθανότητα η τυχαία μεταβλητή να βρίσκεται στην ζώνη είτε μεταξύ -∞ και (-1) τυπική απόκλιση είτε στην ζώνη μεταξύ 1 τυπική απόκλιση και +∞ είναι 1 - 2*0,3413 = 0,3174 ή 31,74%.

© 2003 - 2006 gbosmis, web page: http://www.geocities.com/gbosmis

Appendix: Standard Error for Sampling Distributions

Sampling Distribution	*Standard Error*	*Remarks*
Means	$\sigma/\sqrt{N}$	* for large values of N (>30) the sampling distribution of means is approximately a normal distribution ** in case the population is normally distributed, the sampling distribution of means is also normally distributed even for small values of N
Standard Deviations	$\sigma/\sqrt{(2N)}$ (i) $\sqrt{\left((m_4 - m_2\text{^}2)/(4Nm_2) \right)}$ (ii)	* use (i) only if the population is normal or approximately normal ** for N>100 the sampling distibution of standard deviations is very nearly normal
Correlations	*Test of Hypothesis* $\rho = \rho_0 \neq 0$ $z = \frac{1}{2} \ln[(1+\rho)/(1-\rho)] = 1.1513 \log[(1+\rho)/(1-\rho)]$ $\mu_z = 1.1513 \log[(1+\rho_0)/(1-\rho_0)]$ $\sigma_z = 1/\sqrt{(N-3)}$	* Η μεταβλητή z ακολουθεί κατά προσέγγιση την κανονική κατανομή με μέση τιμή και τυπική απόκλιση αντιστοίχως μ_z και σ_z .

© 2003 - 2006 gbosmis, web page: http://www.geocities.com/gbosmis

APPENDIX - Probabilities

1.1 Δειγματοχώρος – Γεγονότα

Πείραμα τύχης: Είναι κάθε πείραμα που σε πρακτικό επίπεδο υπάρχει αβεβαιότητα ως προς την έκβαση του.

Δειγματοχώρος Ω: Είναι το σύνολο των δυνατών αποτελεσμάτων ενός πειράματος τύχης.

Ενδεχόμενο: Είναι κάθε συνδυασμός δυνατών αποτελεσμάτων ενός πειράματος τύχης.

Είναι πολύ σπουδαίο στην θεωρία πιθανοτήτων να ορίζουμε σωστά σε κάθε πείραμα και τον αντίστοιχο δειγματοχώρο, αλλιώς μπορεί να κάνουμε λάθη και να περιπέσουμε σε παραδοξότητες.

Ο δειγματοχώρος μπορεί να περιέχει
α. πεπερασμένο πλήθος δυνατών αποτελεσμάτων
β. άπειρο αλλά αριθμήσιμο πλήθος δυνατών αποτελεσμάτων
γ. άπειρο αλλά μη αριθμήσιμο πλήθος δυνατών αποτελεσμάτων

Οι δειγματοχώροι που έχουν πεπερασμένο ή άπειρο αλλά αριθμήσιμο πλήθος δυνατών αποτελεσμάτων ονομάζονται απαριθμητοί ή διακριτοί. Επίσης χρησιμοποιούμε τη λέξη πεπερασμένος για το δειγματοχώρο που έχει πεπερασμένο αριθμό δυνατών αποτελεσμάτων.

Θα λέμε ότι ένα ενδεχόμενο συμβαίνει ή πραγματοποιείται μόνο όταν το αποτέλεσμα του πειράματος περιέχεται στο ενδεχόμενο αυτό.

1.2 Αξιωματικός ορισμός της πιθανότητας

Αξιώματα της συνάρτησης πιθανότητας

1. $P(\Omega) = 1$
2. Εάν Α ενδεχόμενο τότε $P(A) \geq 0$
3. Εάν δύο ενδεχόμενα Α και Β είναι ασυμβίβαστα μεταξύ τους τότε $P(A+B) = P(A) + P(B)$

1.3 Βασικές ιδιότητες των πιθανοτήτων

Αποδεικνύονται οι εξής ιδιότητες:

1. $0 \leq P(A) \leq 1$
2. $P(AB') = P(A) - P(AB)$
3. $P(A+B) = P(A) + P(B) - P(AB)$
4. $P(A') = 1 - P(A)$

1.4 Δειγματοληψία

Α. Διατεταγμένο δείγμα χωρίς επανάθεση

Από ένα πληθυσμό n διαφορετικών στοιχείων παίρνουμε r, ένα - ένα χωρίς επανάθεση και σημειώνοντας την σειρά εμφάνισης. Το πλήθος των διαφορετικών δειγμάτων που μπορούμε να πάρουμε είναι:

$(n)_r = n*(n-1)* \ldots *(n-r+1)$

Στην περίπτωση που r = n τότε:

$(n)_n = n*(n-1)* \ldots *1$

Β. Διατεταγμένο δείγμα με επανάθεση

Από ένα πληθυσμό n διαφορετικών στοιχείων παίρνουμε r, ένα - ένα με επανάθεση και σημειώνοντας την σειρά εμφάνισης. Το πλήθος των διαφορετικών δειγμάτων που μπορούμε να πάρουμε είναι:

$(n)_r = n^r$

Γ. Δείγμα χωρίς επανάθεση χωρίς διάταξη

Από ένα πληθυσμό n διαφορετικών στοιχείων παίρνουμε r στοιχεία μαζί, χωρίς επανάθεση και χωρίς να μας ενδιαφέρει η σειρά εμφάνισης. Το πλήθος των διαφορετικών δειγμάτων που μπορούμε να πάρουμε είναι:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Δ. Δείγμα με επανάθεση χωρίς διάταξη

Από ένα πληθυσμό n διαφορετικών στοιχείων παίρνουμε r στοιχεία, με επανάθεση και χωρίς να μας ενδιαφέρει η σειρά εμφάνισης. Το πλήθος των διαφορετικών δειγμάτων που μπορούμε να πάρουμε είναι:

$$\binom{n+r-1}{r} = \frac{(n+r-1)!}{r!(n-1)!}$$

Ε. Διαδοχική δειγματοληψία

© 2003 - 2006 gbosmis, web page: http://www.geocities.com/gbosmis

Αν έχουμε ένα πληθυσμό μεγέθους n και πάρουμε ένα δείγμα μεγέθους n (όλα τα στοιχεία) σε k διαφορετικά στάδια , δηλαδή πρώτα πάρουμε δείγμα μεγέθους r_1,μετά δείγμα μεγέθους r_2 από τα n - r_1 εναπομένοντα στοιχεία και συνεχίσουμε έως ότου να εξαντλήσουμε στην k δειγματοληψία μεγέθους r_k τα r_k εναπομένοντα στοιχεία, τότε το πλήθος των διαφορετικών δειγμάτων είναι:

$$\frac{n!}{r_1!r_2!.....r_k!}$$

1.5 Δεσμευμένη (υπό συνθήκη) πιθανότητα

Εάν Α και Β είναι δύο ενδεχόμενα με P(A) > 0 τότε η δεσμευμένη πιθανότητα του Α δεδομένου ότι συνέβη (ή ότι θα συμβεί) το Β είναι:

$$P(A/B)=\frac{P(AB)}{P(B)}$$

Ιδιότητες των δεσμευμένων πιθανοτήτων

1. $P(B/A) \geq 0$ για κάθε Β
2. $P(A/A) = 1$
3. $P(UB_n/A) = \Sigma_n P(B_n/A)$ εάν τα B_n είναι ασυμβίβαστα
4. $P(\Omega/A) = 1$
5. $P(\emptyset/A) = 0$
6. Εάν $A \subset B$ τότε $P(B/A) = 1$
7. $P(A' / B) = 1 – P(A/B)$ αλλά $P(A/B') \neq 1 – P(A/B)$
8. $P(A+B/C) = P(A/C) + P(B/C) – P(AB/C)$

Παρατηρούμε ότι η δεσμευμένη πιθανότητα ικανοποιεί όλα τα αξιώματα μιας συνάρτησης πιθανότητας.

Πολλαπλασιαστικός τύπος πιθανοτήτων:

$P(AB) = P(A)P(B/A) = P(B)P(A/B)$

Αν δύο ενδεχόμενα είναι ανεξάρτητα θα πρέπει $P(AB) = P(A)P(B)$

Για να είναι n ενδεχόμενα τελείως ανεξάρτητα θα πρέπει να ισχύει για κάθε υποσύνολο τους $P(A_1A_2…A_r) = P(A_1)P(A_2)…P(A_r)$

Ανεξάρτητα πειράματα

Αν έχουμε n πειράματα τύχης τότε λέμε ότι αυτά είναι ανεξάρτητα, εάν για οποιαδήποτε ενδεχόμενα A_1,A_2,A_n ώστε το ενδεχόμενο A_i να σχετίζεται μόνο με το πείραμα i, ισχύει η σχέση $P(A_1A_2…A_n) = P(A_1)P(A_2)…P(A_n)$

Θεώρημα της ολικής πιθανότητας

Έστω ένα ενδεχόμενο Α το οποίο μπορεί να πραγματοποιηθεί μόνο σε συνδυασμό με οποιοδήποτε των n ενδεχομένων $B_1,B_2,...,B_n$ τα οποία είναι ξένα μεταξύ τους, δηλαδή $B_iB_j = \emptyset$, $i \neq j$. Τότε:

$P(A) = P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + … + P(B_n)P(A/B_n)$

Θεώρημα του Bayes

Έστω τα ενδεχόμενα $B_1,B_2,...,B_n$ όπως και στο θεώρημα ολικής πιθανότητας. Τότε ισχύει:

$$P(B_i / A)=\frac{P(B_i)P(A / B_i)}{P(B_1)P(A / B_1)+...+P(B_n)P(A / B_n)}$$

Η $P(B_i)$ καλείται 'εκ των προτέρων' ή 'a priori' ενώ η $P(B_i/A)$ καλείται 'εκ των υστέρων' ή 'a posteriori' πιθανότητα του B_i.

1.6 Τυχαίες μεταβλητές και συνάρτηση κατανομής

Τυχαία μεταβλητή είναι μια συνάρτηση που απεικονίζει τα απλά ενδεχόμενα του δειγματοχώρου Ω στην ευθεία των πραγματικών αριθμών. Οι τυχαίες μεταβλητές μπορούν να παίρνουν συνεχείς ή διακριτές τιμές.

Εάν Χ είναι μια τυχαία μεταβλητή τότε εξ ορισμού η συνάρτηση κατανομής της είναι:

$F(x) = P(X \leq x)$

Στην περίπτωση διακριτών τυχαίων μεταβλητών ισχύει:

$$F(x)=\sum_{x_j \leq x} p(x_j)$$

Στην περίπτωση συνεχών τυχαίων μεταβλητών ισχύει:

$$F(x)=\int_{-\infty}^{x} f(z)dz$$

όπου $f(x)=\frac{dF(x)}{dx}$ η συνάρτηση πυκνότητας πιθανότητας.

Ιδιότητες της συνάρτησης κατανομής

1. Η F(x) είναι αύξουσα συνάρτηση
2. $0 \le F(x) \le 1$
3. $F(-\infty) = 0$
4. $F(\infty) = 1$
5. Η F(x) είναι δεξιά συνεχής
6. $P(x_1 < x \le x_2) = F(x_2) - F(x_1)$

1.7 Μέση Τιμή – Ροπές

Η μέση τιμή μιας τυχαίας μεταβλητής ορίζεται ως:

$E[X] = \int_{-\infty}^{\infty} x f(x) dx$ για συνεχή τυχαία μεταβλητή

$E[X] = \sum_j x_j p(x_j)$ για διακριτή τυχαία μεταβλ.

Για να υπάρχει η μέση τιμή θα πρέπει το ολοκλήρωμα (ή αντίστοιχα το άθροισμα) να συγκλίνει απολύτως. Δηλαδή:

$$\int_{-\infty}^{\infty} |x| f(x) dx < \infty$$

Ιδιότητες της μέσης τιμής

1. $E[aX+b] = aE[X] + b$, (a,b σταθερές)
2. $E[aX] = aE[X]$
3. $E[b] = b$

Η $E[X^r]$ λέγεται ροπή r-τάξης ως προς την αρχή της τυχαίας μεταβλητής X. Η $E[(X - E[X])^r]$ λέγεται ροπή r-τάξης ως προς την μέση τιμή της τυχαίας μεταβλητής. Για την προσέγγιση μιας κατανομής χρησιμοποιούμε συνήθως τις ροπές μέχρι τέταρτη τάξη.

Η ροπή δεύτερης τάξης ως προς την μέση τιμή ονομάζεται διακύμανση (variance) και αποτελεί μέτρο της συγκέντρωσης των τιμών της τυχαίας μεταβλητής γύρω από την μέση τιμή της. Ορίζεται ως εξής:

$Var(X) = \sigma^2 = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

Η τετραγωνική ρίζα της διακύμανσης (σ) ονομάζεται τυπική απόκλιση (standard deviation).

Ιδιότητες της διακύμανσης

1. $Var(X + b) = Var(X)$
2. $Var(aX) = a^2 Var(X)$, (a,b σταθερές)

Γενικά οι υψηλότερης τάξης ροπές εκφράζονται ως συνάρτηση της κανονικοποιημένης τ.μ.

$\frac{X - E[X]}{std[X]} = \frac{X - \mu}{\sigma}$, η οποία έχει μέση τιμή 0 και τυπική απόκλιση 1.

Η κανονικοποιημένη ροπή $3^{ης}$ τάξης ονομάζεται skewness και ορίζεται ως $E[((X-\mu)/\sigma)^3]$. Όπως και κάθε ροπή περιττής τάξης, θα είναι μηδέν για κατανομές που είναι συμμετρικές γύρω από την μέση τιμή τους. Η skewness θα είναι αρνητική αν η κατανομή σκεβρώνει προς τα αριστερά του μέσου της, και θετική αν η κατανομή σκεβρώνει προς τα δεξιά του μέσου της.

Η κανονικοποιημένη ροπή $4^{ης}$ τάξης ονομάζεται kurtosis και ορίζεται ως $E[((X-\mu)/\sigma)^4]$. Η kurtosis μετρά το σχετικό πάχος στις ουρές της κατανομής.

1.8 Πολυδιάστατες τυχαίες μεταβλητές

Στην περίπτωση που σε κάθε αποτέλεσμα ενός πειράματος τύχης μετρούμε n διαφορετικά μεγέθη τότε ορίζουμε μια απεικόνιση από τον δειγματοχώρο Ω στον R_n η οποία καλείται n-διάστατη τυχαία μεταβλητή. Η από κοινού συνάρτηση κατανομής μιας n-διάστατης τυχαίας μεταβλητής ορίζεται ως εξής:

$F(x_1,x_2,\ldots,x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$

Η συνάρτηση πυκνότητας πιθανότητας μιας n-διάστατης τυχαίας μεταβλητής ορίζετια ως εξής:

$$f(x_1, x_2, ..., x_n) = \frac{\partial^n F(x_1, x_2, ..., x_n)}{\partial x_1 \partial x_2 ... \partial x_n}$$

Ιδιότητες δυσδιάστατων τυχαίων μεταβλητών

1. $F_{XY}(+\infty,+\infty) = 1$
2. $F_{XY}(-\infty,y) = F_{XY}(x,-\infty) = 0$
3. $F_{XY}(+\infty,y) = F_Y(y)$
4. $F_{XY}(x,+\infty) = F_X(x)$
5. Η $F_{XY}(x,y)$ είναι συνεχής εκ δεξιών και εκ των άνω

1.9 Συνδιακύμανση – Συσχέτιση

Η συνδιακύμανση δύο τυχαίων μεταβλητών X και Y ορίζεται ως εξής:

$Cov(X,Y) = \sigma_{XY} = E[(X - E[X])(Y - E[Y])]$

Ισχύει ότι

$Cov(X,Y) = E[XY] - E[X]E[Y]$

Εάν η συνδιακύμανση δύο μεταβλητών είναι θετική τότε όταν η μία μεταβλητή αυξάνει, κατά μέσο όρο και η άλλη μεταβλητή αυξάνει. Αντιθέτως εάν η συνδιακύμανση δύο μεταβλητών είναι αρνητική τότε όταν η μια μεταβλητή αυξάνει, κατά μέσο όρο η άλλη μεταβλητή ελαττώνεται.

Για την συνδιακύμανση ισχύουν

1. $Cov(X+Y,Z) = Cov(X,Z) + Cov(Y,Z)$
2. $Cov(X,X) = Var(X)$
3. $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
4. $Cov(aX+b,cY+d) = acCov(X,Y)$

Συχνά στην πράξη χρησιμοποιούμε τον λεγόμενο συντελεστή συσχέτισης μεταξύ δύο μεταβλητών, ο οποίος είναι καθαρός αριθμός (χωρίς διαστάσεις). Αυτός ορίζεται ως εξής:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

όπου τα σ_X και σ_Y είναι οι τυπικές αποκλίσεις των X και Y αντιστοίχως.

Για τον συντελεστή συσχέτισης ισχύουν

1. Παραμένει αναλλοίωτος κάτω από γραμμικούς μετασχηματισμούς των μεταβλητών (δηλ. δεν μεταβάλλεται εάν αλλάξει η αρχή των αξόνων και η μονάδα μέτρησης).
2. $-1 \le \rho \le 1$

Εάν η συσχέτιση δύο μεταβλητών είναι θετική τότε όταν η μια μεταβλητή αυξάνει, κατά μέσο όρο και η άλλη μεταβλητή αυξάνει. Αντιθέτως εάν η συσχέτιση δύο μεταβλητών είναι αρνητική τότε όταν η μια μεταβλητή αυξάνει, κατά μέσο όρο η άλλη μεταβλητή ελαττώνεται. Τέλεια θετική γραμμική συσχέτιση δύο μεταβλητών έχουμε όταν $\rho=1$.

Δύο ανεξάρτητες μεταβλητές έχουν συντελεστή συσχέτισης $\rho=0$. Το ανάποδο δεν ισχύει, δηλαδή συσχέτιση $\rho=0$ δεν σημαίνει κατά ανάγκη ότι οι δύο μεταβλητές είναι ανεξάρτητες. Για τον λόγο αυτό είναι συχνά χρήσιμο να βλέπουμε και το scattered διάγραμμα των μεταβλητών X και Y ώστε να ελέγχουμε αν υπάρχει άλλη (μη γραμμική) μεταξύ τους συσχέτιση.

1.10Διωνυμική κατανομή

Ας θεωρήσουμε ένα πείραμα τύχης του οποίου το αποτέλεσμα μπορεί να διακριθεί σε δύο κατηγορίες «επιτυχία» ή «αποτυχία». Εάν η πιθανότητα επιτυχίας είναι p και σταθερή από πείραμα σε πείραμα, τότε η πιθανότητα να παρατηρήσουμε x επιτυχίες σε n πειράματα καθορίζεται από την διωνυμική κατανομή. Η σειρά των αποτελεσμάτων ονομάζεται διαδικασία Bernoulli.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

1.11Κανονική κατανομή

Κάτω από πολύ γενικές συνθήκες, το άθροισμα ενός μεγάλου αριθμού τυχαίων μεταβλητών ακολουθεί την λεγόμενη κανονική κατανομή. Η συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής δίδεται από την:

$$f_X(x) = \frac{1}{\sqrt{2 \cdot \pi} \cdot \sigma} \cdot e^{-(x-\mu)^2/(2\sigma^2)}$$

Η παράμετρος μ είναι η μέση τιμή της κατανομής ενώ η παράμετρος σ είναι η τυπική απόκλιση της κατανομής.

© 2003 - 2006 gbosmis, web page: http://www.geocities.com/gbosmis