20/12/2022

Αναλυτικές Μέθοδοι στη Γεωπληροφορική

Χρήση του R για συναρτήσεις κατανομής πιθανότητας τυχαίων μεταβλητών

Επόμενες ενότητες μαθήματος

- Σύντομη αναφοράς σε θεμελιώδεις πτυχές της Θεωρίας Πιθανοτήτων και της Στατιστικής – *Τυχαίες μεταβλητές* και ***κατανομές πιθανοτήτων***
  - Γιατί είναι σημαντικό να κατανοηθούν και γιατί πρέπει να μας ενδιαφέρουν εξαρχής
- Βασικές έννοιες και ορισμοί
- Πως να περιγράφετε μαθηματικά τυχαίες μεταβλητές και κατανομές πιθανοτήτων στο R
  - Χρήση εργαλείων του R για την ανάλυση δεδομένων που προκύπτουν από τυχαίες διεργασίες

Η απαρχές της Θεωρίας Πιθανοτήτων

- **1564: *Gerolamo Cardano,*** ***Liber de Ludo Aleae*** *(Βιβλίο των τυχερών παιγνίων)*
  - Πρώτη συστηματική μελέτη της θεωρίας πιθανοτήτων

- **1654: Pierre de Fermat & Blaise Pascal**
  - Οι πατέρες των πιθανοτήτων

Το ξεκίνημα της Θεωρίας Πιθανοτήτων ως λύση στο "πρόβλημα των πόντων"

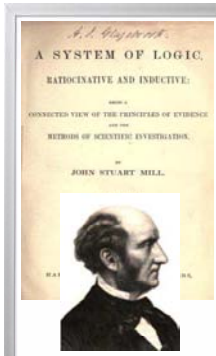
- Πώς μπορούν να μοιραστούν δίκαια, μεταξύ δύο παικτών, τα στοιχήματα μιας παρτίδας τζόγου, η οποία διακόπτεται από εξωτερικές συνθήκες πριν κάποιος παίκτης πετύχει συγκεκριμένο αριθμό επιτυχημένων γύρων;

Pierre-Simon Laplace *Théorie analytique des probabilités, 1812*

" La théorie des probabilités n'est au fond, que le bon sens réduit au calcul … "

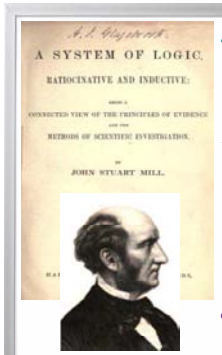
Η Θεωρία Πιθανοτήτων δεν είναι τίποτα άλλο παρά η κοινή λογική που ανάγεται σε υπολογισμούς …

Η πιθανότητα εκφράζει την αβεβαιότητα για την **έκβαση *τυχαίων φαινομένων***



John Stuart Mill (1806-1873)

Το 1843, ο Άγγλος φιλόσοφος διατύπωνε με οξυδέρκεια τη ριζοσπαστική για την εποχή του, αλλά *σήμερα ευρέως αποδεκτή αντίληψη*, ότι η τυχαιότητα στην επιστημονική παρατήρηση δεν είναι συνώνυμο της διαταραχής. Είναι 'κανονικότητα' διαφορετικού είδους.

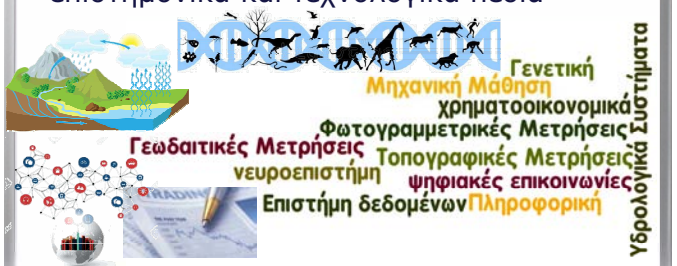


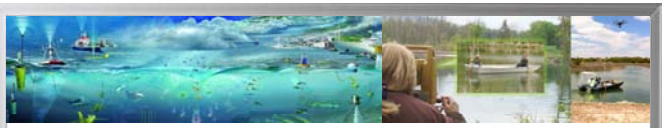
John Stuart Mill (1806-1873)

"…τα ίδια τα γεγονότα που από τη φύση τους φαίνονται πιο ιδιότροπα και αβέβαια, και για τα οποία, σε κάθε μεμονωμένη περίπτωση, κανένας εφικτός βαθμός γνώσης δεν θα μας επέτρεπε να τα προβλέψουμε, όταν μελετηθούν κατ' επανάληψη, κάτω από τις ίδιες συνθήκες, συμβαίνουν με έναν βαθμό κανονικότητας που προσεγγίζει τη μαθηματική περιγραφή τους"

Οι έννοιες των πιθανοτήτων και της τυχαιότητας …

- Παίζουν κεντρικό ρόλο στην ανάλυση παρατηρήσεων/μετρήσεων στα σύγχρονα επιστημονικά και τεχνολογικά πεδία





- Στην επιστημονική έρευνα, συχνά ασχολούμαστε με δεδομένα που επηρεάζονται κατά κάποιο τρόπο από τυχαιότητα ή αβεβαιότητα:

  π.χ., τα δεδομένα μπορεί …
  - Να προέρχονται από πειράματα τύχης ή τυχαία δείγματα (ή υποκείμενα έρευνας) → *ελλιπής παρατηρησιμότητα*,
  - Να επηρεάζονται από τυχαία σφάλματα των μετρήσεων, ή
  - Να μετρούν κάποιο αποτέλεσμα παρατήρησης ενός φυσικού φαινομένου ή διεργασίας.

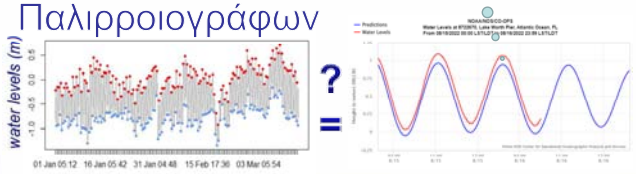
Για να μελετήσουμε ένα φυσικό φαινόμενο συνήθως χτίζουμε ένα μαθηματικό μοντέλο που περιγράφει το φαινόμενο αυτό

Το μοντέλο οφείλει να απλοποιεί τα πράγματα και να αγνοεί τις ασήμαντες λεπτομέρειες

Για να εξετάσουμε την εγκυρότητα του μοντέλου, μπορούμε να συγκρίνουμε τα αποτελέσματα πρόβλεψης με βάση το μοντέλο με τις πραγματικές παρατηρήσεις του φαινομένου

Παρατηρήσεις Παλιρροιογράφων

Προβλέψεις



Στατιστική

- Οι στατιστικές μέθοδοι επιτρέπουν να συνάγουμε συμπεράσματα *(inference)* από την ανάλυση δεδομένων με βάση ένα περιορισμένο αριθμό μετρήσεων ή παρατηρήσεων

Θεωρία Πιθανοτήτων

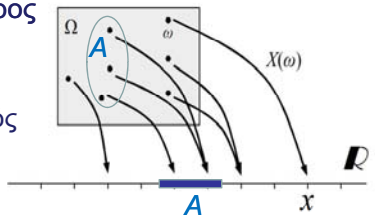
- Η Θεωρία Πιθανοτήτων επιτρέπει να υπολογίσουμε το βαθμό βεβαιότητας για τα συμπεράσματά μας → μέσα από την εξέταση των δεδομένων *(στατιστικά μέτρα)*, ελέγχους υποθέσεων και αξιοπιστίας

- Στα (γεω)επιστημονικά πεδία, εφαρμογές της Θ.Π. και της Στατιστικής μας δίνουν τη δυνατότητα να εκτιμήσουμε με σαφήνεια και ακρίβεια «εννοιολογικά ή μαθηματικά μοντέλα» που χρησιμεύουν ως γέφυρα μεταξύ οποιασδήποτε κατάστασης του πραγματικού κόσμου και της ανάλυσης των εκδηλώσεων τους

Τυχαίες μεταβλητές

- Αποτελούν τον ακρογωνιαίο λίθο πολλών βασικών εννοιών στη Θ.Π. και τη Στατιστική
  - **Με απλά λόγια:** μια **τυχαία μεταβλητή** είναι ένας μαθηματικός φορμαλισμός για να περιγραφούν τα αποτελέσματα ενός πειράματος τύχης, μιας τυχαίας διαδικασίας ή ενός φυσικού φαινομένου.
    - *→Μια μεταβλητή που η τιμή της υπόκειται σε τυχαίες διακυμάνσεις, …*
    - *→Μια και μόνο μια τιμή εκχωρείται σε κάθε σημείο ενός δείγματος, μιας μέτρησης ή μιας παρατήρησης*

- Στην μαθηματική γλώσσα της θεωρίας μετρήσεων, μια **τ.μ.** *δεν είναι ούτε τυχαία ούτε μεταβλητή, …*
- Είναι μια μετρήσιμη συνάρτηση ***X()*** από έναν χώρο μέτρησης πιθανότητας (*χώρος δείγματος*) σε έναν μετρήσιμο χώρο (κάποιο αριθμό/μέτρηση ***x***)
- Ω: Δειγματικός χώρος
- ω: Έκβαση/ γεγονός
- Ο συνολικός αριθμός των εκβάσεων ενός τυχαίου γεγονότος είναι ένα συμβάν *A*



Δειγματικός χώρος S: με ενδεχόμενα συμβάντα Κ ή/και Γ

- Θα μπορούσαμε να ορίσουμε μια τυχαία μεταβλητή ***X(s)*** ως τη συνάρτηση που επιστρέφει τον αριθμό ενδείξεων 'Κ: κεφάλι' για κάθε στοιχείο του δειγματικού χώρου ***S*** :

$$x = \begin{cases} 'K' \to 1 \\ 'Γ' \to 0 \end{cases}$$

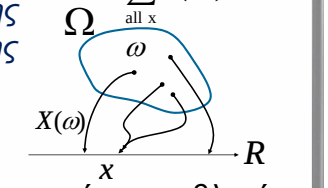
- Μια τ.μ. συμβολίζεται πάντα με κεφαλαία γράμματα, και η τιμή της με πεζά
- Θα μπορούσαμε επίσης να γράψουμε ***P(X = x)*** ως η πιθανότητα ότι ο αριθμός των ενδείξεων *K* ισούται με ***x***.

- X({Γ, K}) = 1
- X({Γ, Γ}) = 0
- X({K, K}) = 2
- X({K, Γ}) = 1

- *Κάθε τ.μ. **X** συνδέεται με μια κατανομή που καθορίζει ένα μέτρο της πιθανότητας εμφάνισης **P(X=x_i)** κάθε ενδεχόμενης τιμής **x_i** της* στο σύνολο όλων των πιθανών τιμών της τυχαίας μεταβλητής

$$0 \le P(X = x) \le 1$$

$$\sum_{\text{all } x} P(x) = 1$$



- Είναι δυνατόν δύο τ.μ. να έχουν πανομοιότυπες κατανομές αλλά να διαφέρουν σημαντικά (π.χ., να είναι ανεξάρτητες)

Συνάρτηση πυκνότητας πιθ., $f(x)$ ↔ Αθροιστική συνάρτηση πιθ., $F(x)$

$$f(x) = \frac{dF(x)}{dx} \iff F(x_i) = \int_{-\infty}^{x_i} f(x)dx$$

- **$f(x) = P(X = x)$:** *σ.π. πιθανότητας* ***(probability density function)*** → *δίνει την πιθανότητα η τ.μ. X να πάρει μια ορισμένη τιμή **x** από το πεδίο τιμών της*
- **$F(x) = P(X \le x)$:** *αθροιστική σ.κ. πιθανότητας* ***(cumulative distribution function)*** → *εκφράζει πόσο πιθανό είναι η παρατηρούμενη τιμή μιας τ.μ. **X** να είναι μικρότερη ή ίση με μια τιμή **x***

- Τυχαίες μεταβλητές κατηγοριοποιούνται
  - σε ***διακριτές*** που έχουν έναν μετρήσιμο εύρος (πιθανών) αποτελεσμάτων, και
  - σε ***συνεχείς*** που μπορούν να πάρουν οποιαδήποτε πιθανή τιμή σε ένα δεδομένο διάστημα των πραγματικών αριθμών

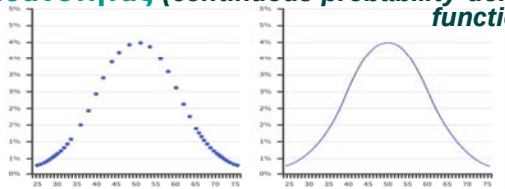

- Η κύρια διαφορά μεταξύ των δύο κατηγοριών είναι ο τύπος των πιθανών τιμών που μπορεί να λάβει κάθε τ.μ.
- Επιπλέον, ο τύπος της (τυχαίας) μεταβλητής υποδηλώνει τη συγκεκριμένη μέθοδο εύρεσης μιας συνάρτησης κατανομής πιθανότητας

Οι τιμές των διακριτών και συνεχών τυχαίων μεταβλητών μπορεί να είναι διφορούμενες:

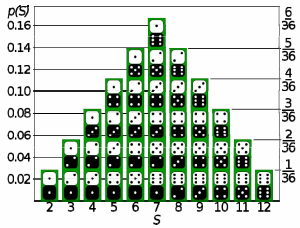
- ***X():=O** αριθμός των σελίδων βιβλίων* → X() *είναι μια διακριτή τυχαία μεταβλητή*
- ***X():=**Το βάρος βιβλίων* → X() *είναι μια συνεχής τυχαία μεταβλητή*
- ***X():=**Το χρονικό διάστημα μεταξύ εξάρσεων του **Old Faithful*** → X() *είναι μια **???** τυχαία μεταβλητή*

- ***X():=** Ο αριθμός των εξάρσεων του **Old Faithful** κατά τη διάρκεια μιας μέρας* → X() *είναι μια **???** τυχαία μεταβλητή*

- Ο διαχωρισμός των τυχαίων μεταβλητών σε ***διακριτές*** και ***συνεχείς*** οδηγεί σε αντίστοιχη διάκριση των συναρτήσεων πιθανότητάς τους σε
  - **διακριτές συναρτήσεις κατανομής πιθανότητας *(discrete probability distribution functions)***, και
  - **συνεχείς συναρτήσεις κατανομής πιθανότητας *(continuous probability density functions)***



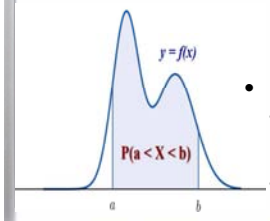
Διακριτές **Σ**υνάρτησεις.**Κ**ατανομής.**Π**ιθανότητας.



- *Με μια διακριτή κατανομή πιθανότητας, κάθε πιθανή τιμή μιας διακριτής τ.μ. μπορεί να συσχετιστεί με μια μη μηδενική πιθανότητα, **P(X=x)***
- Πιθανότητες για άλλα γενικά γεγονότα, π.χ. $P(X \le x)$ μπορεί να ληφθούν αθροίζοντας επιμέρους πιθανότητες $P(X=x_i)$ στο εύρος τιμών ενδιαφέροντος
- Το άθροισμα των πιθανοτήτων στο εύρος των τιμών της τ.μ. πρέπει να ισούται με 1.

Συνεχείς **Σ**υνάρτησεις.**Κ**ατανομής.**Π**ιθανότητας.

- Περιγράφουν τις πιθανότητες των πιθανών τιμών μιας συνεχούς τυχαίας μεταβλητής (τ.μ.) με ένα σύνολο πιθανών τιμών (το *'εύρος'*) που είναι απεριόριστο και μετρήσιμο



- Συνεχείς τ.μ. μπορούν να έχουν μη μηδενική πιθανότητα μόνο σε διαστήματα τιμών
- Η πιθανότητα ότι μια συνεχής τυχαία μεταβλητή ισούται με κάποια συγκεκριμένη τιμή είναι πάντα μηδέν: ***P(X=c) = 0***

Κατανομές Πιθανότητας

Πολλές διαφορετικές συναρτήσεις
... εξυπηρετούν διαφορετικούς σκοπούς και αντιπροσωπεύουν διαφορετικές διαδικασίες παραγωγής δεδομένων

Διακριτές κατανομές

- Bernoulli → Μόνο δύο πιθανά συμβάντα
- **Binomial** → Το άθροισμα των επιτυχιών σε Bernoulli συμβάντα
- Negative Binomial → Αριθμός των επιτυχιών σε μια ακολουθία ανεξάρτητων και πανομοιότυπων κατανεμημένων δοκιμών Bernoulli πριν εμφανιστεί ένας καθορισμένος (μη τυχαίος) αριθμός αστοχιών
- **Poisson** → Πόσο πιθανό είναι ένα συμβάν κατά την διάρκεια μιας περιόδου παρατήρησης
- Geometric → Αριθμός δοκιμών μέχρι το πρώτο επιτυχές συμβάν
- Multinomial → Πρόβλεψη σειράς επαναλήψεων ανεξάρτητων τυχαίων ενδεχομένων

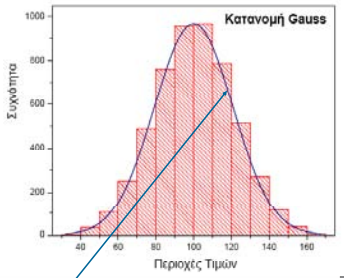
Συνεχείς κατανομές

- Uniform → Γεγονότα που είναι εξίσου πιθανό να συμβούν σε ένα δεδομένο διάστημα
- **Normal → Η πιο σημαντική κατανομή της στατιστικής μεθοδολογίας**
- Exponential → Χρόνος αναμονής μέχρι την πραγματοποίηση ενός γεγονότος (σε μια διαδικασία όπου γεγονότα συμβαίνουν συνεχώς και ανεξάρτητα με ένα σταθερό μέσο ρυθμό)
  - Gamma & Beta → Γενικεύσεις της εκθετικής
- **Chi Square**
- **F**
- **t (Student)**

  Για τη διενέργεια στατιστικών δοκιμών και ελέγχων υποθέσεων
- Weibull → Πρόβλεψη σειράς επαναλήψεων ανεξάρτητων τυχαίων ενδεχομένων

Συνηθισμένες μονοδιάστατες κατανομές

- Σε πολλά τυχαία μεγέθη που μελετάμε παρατηρούμε ότι *οι τιμές τους* x_i *'συγκεντρώνονται' συμμετρικά γύρω από μια κεντρική τιμή* x_0 και 'αραιώνουν' καθώς απομακρύνονται από αυτήν.



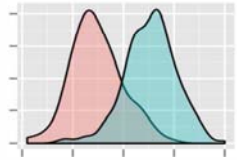
Η κατανομή των μετρήσεων μπορεί να αναπαρασταθεί και από την αντίστοιχη καμπύλη κατανομής

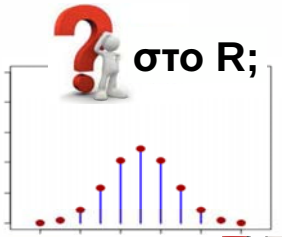
Συνηθισμένες μονοδιάστατες κατανομές

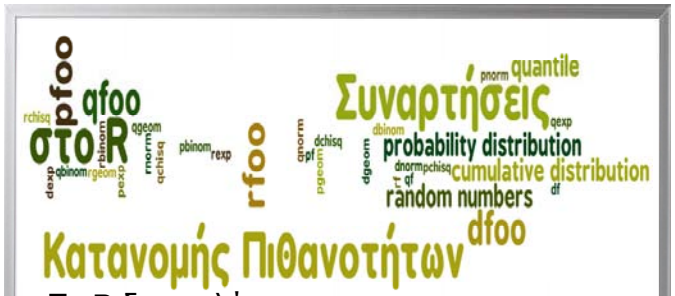
- Σε άλλες περιπτώσεις, παρατηρούνται **ασύμμετρες κατανομές (*skewed distribution*)**
  - συγκέντρωση των περισσότερων τιμών δεξιά ή αριστερά από τη μέση τιμή στον άξονα των τιμών

Σύντομη επισκόπηση

- Συνήθεις συνεχείς κατανομές τυχαίων μεταβλητών
  - Η κανονική και η τυποποιημένη κατανομή
  - Κατανομές Student, χ^2, Fisher, …
  - …
- Διακριτές κατανομές πιθανότητας τυχαίων μεταβλητών
  - Διωνυμική & Poisson







- Το **R** διευκολύνει τους απαραίτητους υπολογισμούς για στατιστικές αναλύσεις με τη χρήση κατάλληλων ενσωματωμένων συναρτήσεων κατανομής πιθανοτήτων

Κατανομές πιθανότητας στο R

υπολογίζονται από 4 βασικές συναρτήσεις.

- Για καθεμία από αυτές, υπάρχει μια *ονομασία ρίζας* ***(root name)***, συμβολικά ***foo()***, που υποδηλώνει τον τύπο της εκάστοτε συνάρτησης κατανομής πιθανότητας
- Στην ονομασία ρίζας, προστίθεται ένα πρόθεμα, με ένα από τα γράμματα ***"p", "q", "d", "r"***, π.χ. ***pfoo(), qfoo(), dfoo(), rfoo(),***

αλλάζοντας έτσι τη συμπεριφορά για τη συνάρτηση κατανομής πιθανότητας (και τον τρόπο χρήσης της) με διαφορετικούς τρόπους


```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)
```

- **Παράδειγμα: Διωνυμική κατανομή**

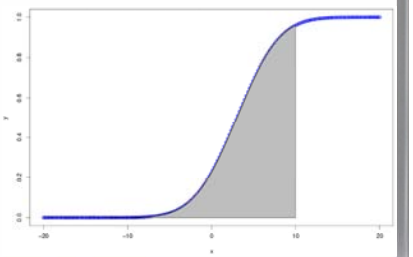
- Κάθε μια από τις συναρτήσεις ***" p..., q..., d..., r... "*** έχει το δικό της σύνολο παραμέτρων εισόδου **(ορίσματα).**
  - Η πρώτη παράμετρος είναι ένα μοναδικό (φαινομενικό) όρισμα στο οποίο υπολογίζεται η εκάστοτε συνάρτηση. Στη συνέχεια ακολουθούν οι παράμετροι κατανομής και εκάστοτε άλλες επιλογές (optional arguments).

- ***d*** , από το ***"density"*** / πυκνότητα, για τη συνάρτηση πυκνότητας πιθανότητας (probability density function ***pdf***).
  - *Η συνάρτηση* ***dfoo( )*** επιστρέφει την τιμή στον άξονα **y** μιας κατανομής πυκνότητας **(**ή μάζας**)** πιθανότητας για μια **διακριτή τιμή x** μιας τ.μ. X**,** δηλ**.**
    - στο σημείο **x: P(X = x)**

$$f(x) = F'(x) = \lim_{\epsilon \to 0} \frac{\Pr(x < X \le x + \epsilon)}{\epsilon}$$

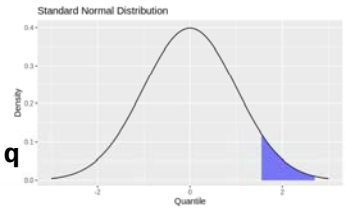
    - υποδηλώνει την πιθανότητα παρατήρησης μιας τ.μ. με μια συγκεκριμένη τιμή

- **p ,** από το ***"probability"*****,** υποδηλώνει την αθροιστική κατανομή πιθανότητας ***(probability cumulative distribution function, CDF***
  - Η συνάρτηση **pfoo( )** παρέχει τη σωρευτική τιμή από το αρνητικό άπειρο μέχρι την τιμή ***x*** μιας τυχαίας μεταβλητής ***X*** **: F *(x)* = P *(X* ≤ *x),***
  - ή την περιοχή στα αριστερά μιας τιμής **x** κάτω από την καμπύλη της αθροιστικής κατανομής πιθανότητας

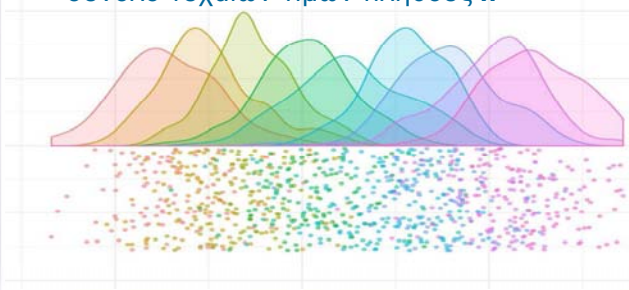


- ***q*** **,** από το ***"quantile*** / ποσοστημόριο"), υποδηλώνει το αντίστροφο της CDF**.**
  - Η συνάρτηση q***foo( )*** εκτελεί τον υπολογισμό ποσοστιαίων σημείων ή ισοδύναμα το αντίστροφο της αντίστοιχης ***pfoo( )*** συνάρτησης, δηλαδή

για μια δεδομένη πιθανότητα ***x*** επιστρέφει την πιθανότητα **P(X≤ x) > q,** για καθορισμένη τιμή **q (**ποσοστημόριο**)**



- ***r*** , από το ***"random"***, για να αντλήσουμε/ προσομοιώσουμε τυχαία δείγματα που προέρχονται από την καθορισμένη κατανομή.
  - *Η συνάρτηση* ***rfoo(n,…)*** επιστρέφει ένα σύνολο τυχαίων τιμών πλήθους ***n***

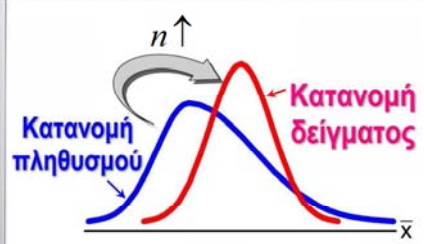


Κανονική κατανομή

- Μελετήθηκε αρχικά από τον *De Moivre (1667-1754)* στην προσπάθειά του να διαμορφώσει το μαθηματικό υπόβαθρο για να εξηγήσει την τυχαιότητα
- και αργότερα από τον ***Laplace (1749-1827)*** με την εξαγωγή του θεμελιώδους θεωρήματος του κεντρικού ορίου *(Central Limit Theorem)*

"Κανονική" κατανομή

- Ο όρος "κανονική" χρησιμοποιείται με την έννοια του **'συνήθους'** ή **αυτού *που θα συνέβαινε*** μακροπρόθεσμα υπό ορισμένες συνθήκες
  - σύμφωνα με ένα πρότυπο, με ένα υπόδειγμα, που δεν παρουσιάζει αποκλίσεις από αυτό, που δεν αποτελεί εξαίρεση σε ότι ισχύει συνήθως
- Η "κανονική" κατανομή έχει τεράστιο εύρος εφαρμογών γιατί περιγράφει με ακρίβεια την κατανομή των τιμών για τα χαρακτηριστικά πολλών φυσικών φαινομένων, τα οποία είναι το άθροισμα πολλών ανεξάρτητων διεργασιών



Το Θ.Κ.Ο. συνδέει την κανονική κατανομή με οποιαδήποτε άλλη κατανομή

- **Το Θεώρημα του Κεντρικού Ορίου *(Central Limit Theorem)***: Η μέση τιμή, μεγάλου αριθμού ανεξάρτητων παρατηρήσεων, ακολουθεί κατά προσέγγιση κανονική κατανομή, ανεξάρτητα από το ποια κατανομή ακολουθούν οι παρατηρήσεις

- Το 1809 ο ***Gauss (1777-1855)*** παρατήρησε ότι οι κατανομές των σφαλμάτων αστρονομικών παρατηρήσεων μπορούσαν να προσεγγιστούν ικανοποιητικά από μια συνεχή καμπύλη …

→ ***«κατανομή των σφαλμάτων»*** που αποδίδονταν στην τυχαιότητα → *τρόπος εκλογίκευσης της Μεθόδου των Ελαχίστων Τετραγώνων*

Κανονική ή Gauss κατανομή

- Απεικονίζεται, μαζί με προσωπογραφία του Gauss και τον μαθηματικό τύπο της, στο προ-ευρώ χαρτονόμισμα των 10 μάρκων της Γερμανίας

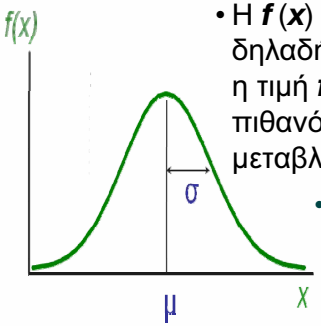
Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής

- Έχει σχήμα 'τομής καμπάνας'
- Συμβολίζεται ως ***N(μ, σ²)*** και ορίζεται από τη *συνάρτηση πυκνότητας πιθανότητας*

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

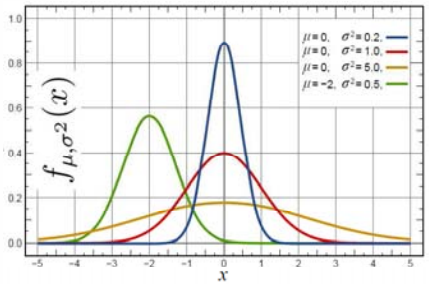
με χαρακτηριστικές παραμέτρους: την *τυπική απόκλιση σ* (ή *τη διασπορά σ²*) και την *μέση τιμή μ*

Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής



- Η ***f*** **(*x*)** εκφράζει *πυκνότητα*, δηλαδή, όσο μεγαλύτερη είναι η τιμή ***f*** **(*x*)** τόσο περισσότερο πιθανό είναι να πάρει η μεταβλητή ***X*** τιμές κοντά στο ***x***
- Η καμπύλη της σ.π.π. είναι συμμετρική, και **ασυμπτωτική** προς τον οριζόντιο άξονα

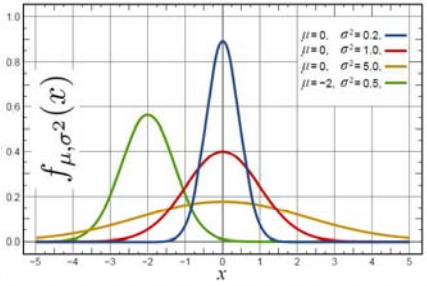
Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής



- Δεν ορίζει μια συγκεκριμένη κανονική καμπύλη αλλά μια οικογένεια κανονικών καμπύλων

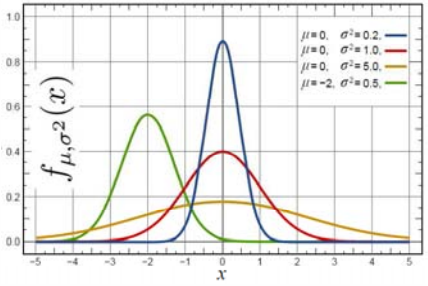
→ $N(\mu,\sigma^2)$ για διαφορετικές τιμές μ και σ

Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής



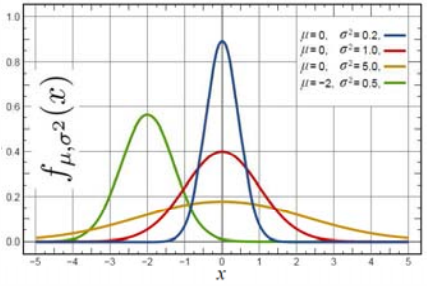
- Στην κορυφή της καμπύλης (στην περιοχή που παρουσιάζει τη μεγαλύτερη πυκνότητα), η μέση τιμή και η διάμεσος ταυτίζονται.

Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής



- Αλλαγή της μέσης τιμής προκαλεί μόνο μετατόπιση της κανονικής καμπύλης σε μια νέα θέση

Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής

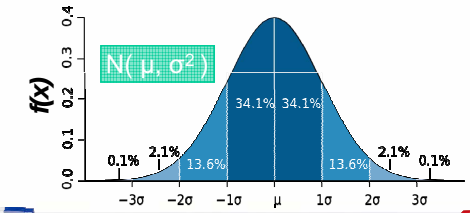


- Αλλαγή, της τυπικής απόκλισης, όμως, προκαλεί αλλαγή στο εύρος της κανονικής καμπύλης

Όσο πιο οξεία είναι η καμπύλη κοντά στον κατακόρυφο άξονα , τόσο πιο ακριβής είναι η σειρά μετρήσεων (μικρή τιμή **σ**).

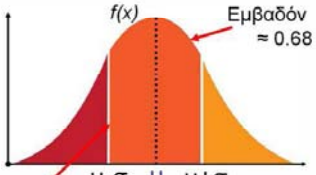
Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής

- Η καμπύλη παρουσιάζει μέγιστη τιμή, ίση με $\mathbf{1/(\sigma\sqrt{2\pi}) = 0.399/\sigma}$, στη θέση $\mathbf{x = \mu}$ και στις θέσεις $\mathbf{x = \mu-\sigma}$ και $\mathbf{x = \mu+\sigma}$ παρουσιάζει **σημεία καμπής *(inflection points)***



- Το εμβαδόν του σκιαγραφημένου χωρίου στο σχήμα, εκφράζει την πιθανότητα η τ.μ. X να πάρει κάποια τιμή μεταξύ των τιμών [-∞, +∞]

- π.χ., το εμβαδόν (μεταξύ μ±σ) είναι περίπου 68% του συνολικού εμβαδού κάτω από ολόκληρη την καμπύλη (το οποίο, εξ ορισμού, είναι ίσο με 1).



$$P(-\sigma,+\sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)\,dx$$

Συνάρτηση πυκνότητας πιθανότητας της κανονικής κατανομής

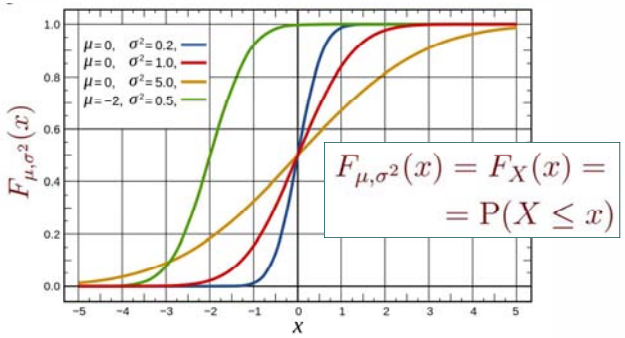
$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} \frac{1}{\sigma\sqrt{2\pi}} \bullet e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx = .997$$

- Παρέχει διαστήματα εμπιστοσύνης περί τη μέση τιμή μ όταν είναι γνωστή η μεταβλητότητα σ^2 του πληθυσμού

Αθροιστική συνάρτηση πιθανότητας της κανονικής κατανομής



Αθροιστική συνάρτηση πιθανότητας της κανονικής κατανομής

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} f_X(t)\,dt =$$

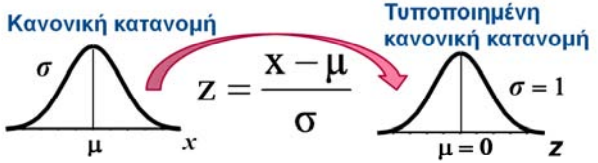
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,dt$$

- Το ολοκλήρωμα που εκφράζει την $F(x)$ δεν μπορεί να γραφεί σε απλούστερη μορφή → γι' αυτό συνήθως υπολογίζεται με προσεγγιστικές μεθόδους

Standardized Normal Distribution

Τυποποιημένη Κανονική Κατανομή

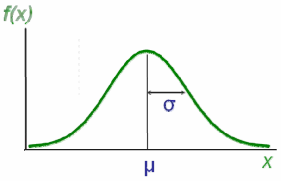
Η Τυποποιημένη ή τυπική κανονική κατανομή



- Μια τ.μ. που ακολουθεί την τυποποιημένη κανονική κατανομή, έχει επικρατήσει να συμβολίζεται με *Z* και η συνάρτηση πυκνότητάς της με $f_Z(\,)$ ή συνήθως $\varphi_Z(\,)$ ή απλά $\varphi(\,)$

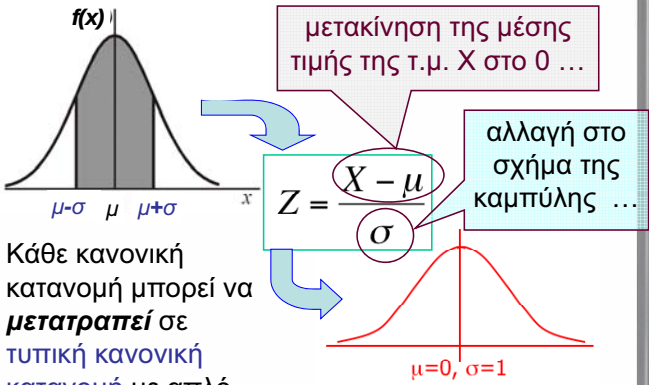
$$\phi(x) = \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}$$

$$-\infty < z < \infty$$



- και η αντίστοιχη αθροιστική συνάρτηση κατανομής Φ(x)

$$\Phi(x) = \int_{-\infty}^{x} \phi_X(t)\,dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\,dt$$



Κάθε κανονική κατανομή μπορεί να ***μετατραπεί*** σε τυπική κανονική κατανομή με απλό αλγεβρικό τρόπο

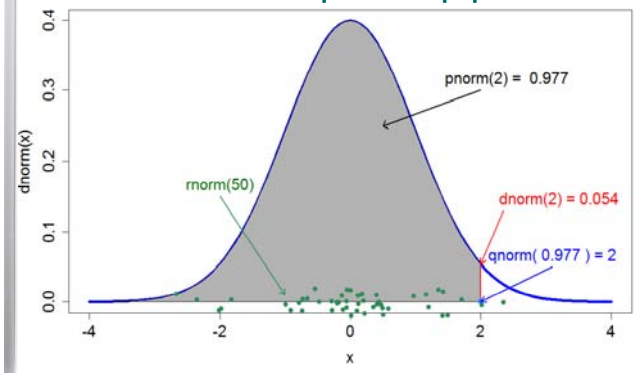
Κανονική → Τυποποιημένη κανονική κατανομή

- Ο απλός μετασχηματισμός …

$$X \sim N(\mu, \sigma^2) \quad \Longrightarrow \quad Z \equiv \frac{X-\mu}{\sigma} \sim N(0,1)$$

επιτρέπει να αποφεύγεται ο υπολογισμός των περίπλοκων ολοκληρωμάτων των συναρτήσεων $f_X(x)$ και $F_X(x)$ της κανονικής κατανομής, χρησιμοποιώντας κατάλληλους πίνακες της τυποποιημένης κανονικής κατανομής $f_Z(x):=\varphi(x)$ και $F_Z(x):=\Phi(x)$

Οι 4 συναρτήσεις πιθανότητας για την κανονική κατανομή




```
KANONIKH KATANOMH - Συναρτήσεις κατανομής πιθανότητας
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)

x, q   vector of quantiles.
p      vector of probabilities.
n      number of observations.
       If length(n) > 1,
       the length is taken to be the number required.
mean   vector of means.
sd     vector of standard deviations.
log, log.p    logical; if TRUE,
              probabilities p are given as log(p).
lower.tail    logical; if TRUE (default),
              probabilities are P[X≤x]
              otherwise, P[X>x].
```

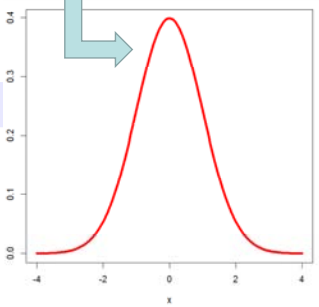
Σύνταξη εντολών


```
x=seq(-4,4,length=200)
y=1/sqrt(2*pi)*exp(-x^2/2)
plot(x,y,type="l",
     lwd=4,col="red")
```

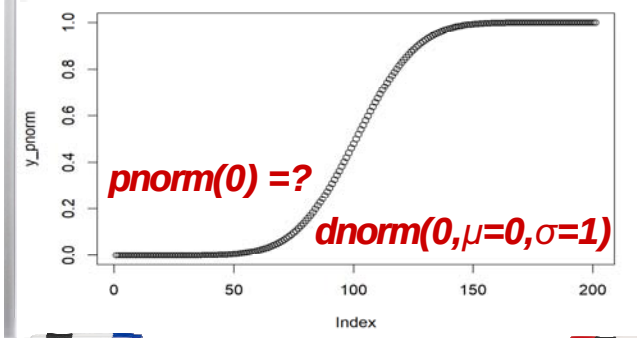
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Το ίδιο αποτέλεσμα με τη χρήση της dnorm()

```
x=seq(-4,4,length=200)
y=dnorm(x,mean=0,sd=1)
plot(x,y,type="l",
     lwd=2,col="red")
```

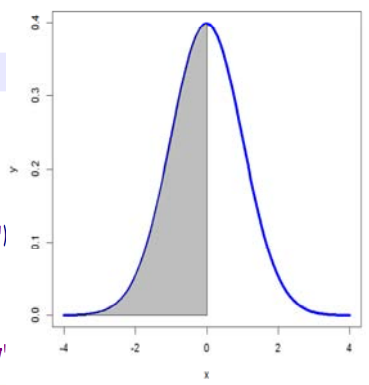



```
x_pnorm <- seq(- 5, 5, by = 0.05)   # Specify x-values for pnorm function
y_pnorm <- pnorm(x_pnorm)           # Apply pnorm function
plot(y_pnorm)                       # Plot pnorm values
```

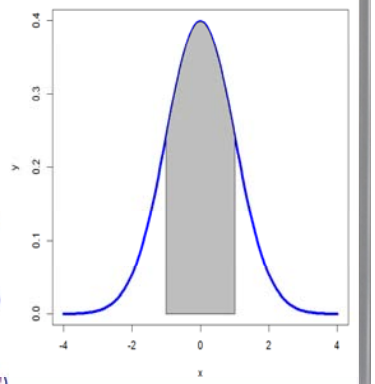


- Η ***pnorm()*** είναι χρήσιμη και για υπολογισμούς με την τυποποιημένη κανονική κατανομή, Z~N(0,1).
- Επιστρέφει το ολοκλήρωμα (από -∞ σε q) της σ.π.π. (pdf) της κανονικής κατανομής, όπου ***q*** είναι το Z-score ή τυποποιημένο σκορ $z = (x - \mu) / (\sigma / \sqrt{n})$, δηλ. είναι ο αριθμός των τυπικών αποκλίσεων από τον μέσο όρο που απέχει ένα σημείο των δεδομένων.
- Η ***pnorm()*** είναι η συνάρτηση που αντικαθιστά τους γνώριμους πίνακες πιθανοτήτων βάσει τιμών Z σε εγχειρίδια Στατιστικής.

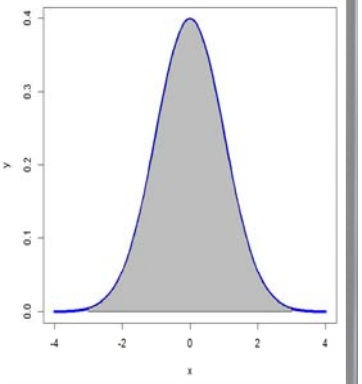
```
pnorm(0, mean=0, sd=1)
[1] 0.5
The area under
  the standard normal curve
  to the left of x = 0
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=4, col="blue")
x=seq(-4,0,length=100)
y=dnorm(x)
polygon(c(-4,x,0),c(0,y,0),col="gray")
```




```
pnorm(1, mean=0, sd=1)
[1] 0.8413447
pnorm(-1, mean=0, sd=1)
[1] 0.1586553
{pnorm(1, mean=0, sd=1) -
    pnorm(-1, mean=0, sd=1)}
[1] 0.6826895
The area under the standard
  normal curve that falls within
  one standard deviation of the mean
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l", lwd=4, col="blue")
x=seq(-1,1,length=100)
y=dnorm(x)
polygon(c(-1,x,1),c(0,y,0),col="gray")
```




```
pnorm(3,mean=0,sd=1)-pnorm(-3,mean=0,sd=1)
[1] 0.9973002
The area under the standard
  normal curve that falls within
  3 standard deviations of the mean
x=seq(-4,4,length=200)
y=dnorm(x)
plot(x,y,type="l",lwd=2,col="blue")
x=seq(-3,3,length=200)
y=dnorm(x)
polygon(c(-3,x,3),c(0,y,0),col="gray")
```



Η συνάρτηση πιθανότητας ποσοστημορίων: *qnorm*

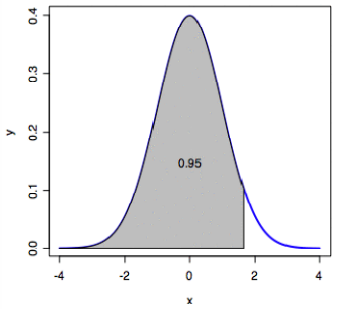
- Είναι απλά το αντίστροφο της αθροιστικής σ.κ.π. (CDF), ***Q(p)=F^{-1}(p)*** : ***inverse look-up*** → επίσης μπορεί να θεωρηθεί ως το αντίστροφο της ***pnorm (direct look-up) : p = F(x) & x = F^{-1}(p)***

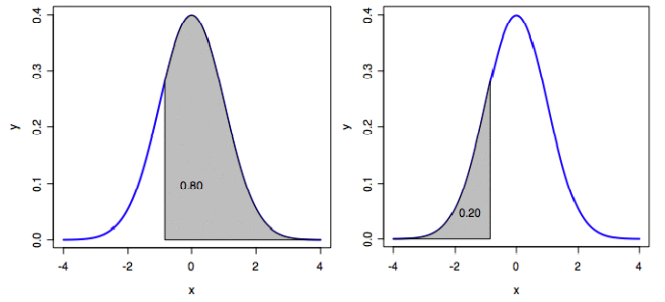
Απαντά σε ερωτήματα, όπως:

- Αν είναι γνωστή η περιοχή κάτω από την καμπύλη της σ.π.π. προς τα αριστερά ενός άγνωστου αριθμού: Ποιος είναι ο άγνωστος αριθμός (δηλ. η οριακή τιμή που καθορίζει αυτή την περιοχή);
- Ή, ποιο είναι το Z-σκορ για το ***p*** ποσοστημόριο της κανονικής κατανομής;

- π.χ., αν η περιοχή κάτω από την καμπύλη της σ.π.π. στα αριστερά κάποιας άγνωστης τιμής x είναι 0.95 → x = ?
- Η ***qnorm*** (όπως και η *pnorm*) χρησιμοποιούνται για την περιοχή κάτω από την καμπύλη στα αριστερά. *Εάν δοθεί μια περιοχή προς τα δεξιά, τότε πρέπει να προηγηθεί μια απλή προσαρμογή πριν εφαρμοστεί η* ***qnorm***.

```
qnorm(0.95,mean=0,sd=1)
[1] 1.644854
```





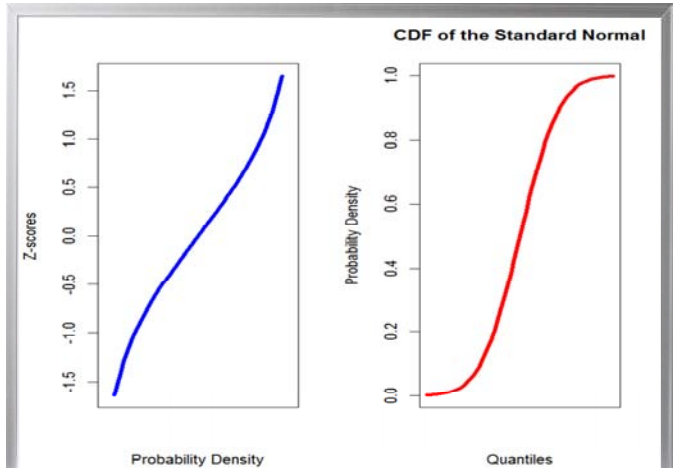
```
qnorm(0.20,mean=0,sd=1)
[1] -0.8416212
```



```
Η τιμή x που χωρίζει το κατώτατο k%
των τιμών από μια κανονική κατανομή N~(μ,σ)
Pk=qnorm(k (in decimal form), mean =μ, sd =σ, lower.tail=TRUE)
P25=qnorm(.25, mean =μ, sd =σ, lower.tail=TRUE)
P90=qnorm(.90, mean =μ, sd =σ, lower.tail=TRUE)

Usage for the standard normal (z) distribution (μ= 0andσ= 1)
Pk=qnorm(k (in decimal form))
P25=qnorm(.25) = -0.67449→-0.67
P90=qnorm(.90) = 1.28155→1.28

# pnorm() και qnorm() είναι 'αντίστροφες' συναρτήσεις !
pnorm(qnorm(0))
[1] 0
qnorm(pnorm(0))
[1] 0
```



```
# Get two graphs next to each other
oldpar <- par()
par(mfrow=c(1,2))

# Make a vector of quantiles: from 0 to 1 by increments of .05
quantiles <- seq(0, 1, by = .05)
quantiles

# Next find the Z-score at each quantile
qvalues <- qnorm(quantiles)
qvalues

# Plot the z_scores
plot(qvalues,
     type = "l", lwd=4, col="blue", # Make a line graph
     xaxt = "n", # No x-axis
     xlab="Probability Density",
     ylab="Z-scores")

# Same pnorm plot from before
plot(pvalues, # Plot where y = values and x = index of the value in the vector
     xaxt = "n", # Don't label the x-axis
     type = "l", lwd=4, col="red", # Make it a line plot
     main = "CDF of the Standard Normal",
     xlab= "Quantiles",
     ylab="Probability Density")
```

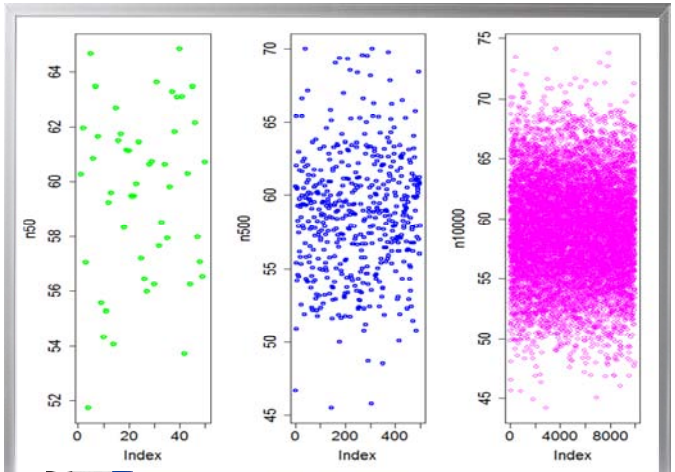
Η συνάρτηση πιθανότητας: *rnorm*

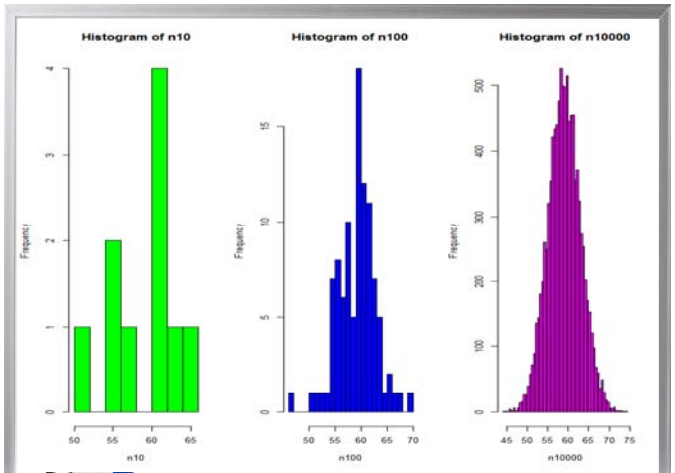
- Χρησιμοποιείται για την προσομοίωση τιμών (δειγμάτων) από μια κανονική κατανομή, η οποία χαρακτηρίζεται από μια δεδομένη μέση τιμή μ και τυπική απόκλιση σ (και όχι τη διακύμανση σ^2)

rnorm (n, mean = μ, sd = σ)

- Το μόνο απαιτούμενο όρισμα είναι ένας αριθμός, το ***n***, που να προσδιορίζει τον αριθμό των πραγματοποιήσεων (τιμών) της κανονικής τυχαίας μεταβλητής που παράγει.


```
set.seed(10-1-2020)
# Generate 3 vectors of random numbers from normal distribution
n50 <- rnorm(50, mean = 59, sd = 4)
n500 <- rnorm(500, mean = 59, sd = 4)
n10000 <-  rnorm(10000, mean = 59, sd = 4)

# This is for getting 3 graphs next to each other
oldpar <- par() ; par(mfrow=c(1,3))
plot(n50, col="green",lwd=3,xlab="Index",ylab="n50",
          cex.lab=1.5,cex.axis=1.5)
plot(n500, col="blue", lwd = 2,xlab= "Index",ylab="n500",
          cex.lab=1.5,cex.axis=1.5)
plot(n10000, col="magenta", xlab= "Index",ylab="n10000",
          cex.lab=1.5,cex.axis=1.5)

# The breaks argument specifies how many bars are in the histogram
hist(n50, breaks = 5, col="green")
hist(n500, breaks = 20, col="blue")
hist(n10000, breaks = 100, col="magenta")
```



Άλλες συχνά χρησιμοποιούμενες κατανομές πιθανότητας

- ***Κατανομή χ²***
- ***Κατανομή* Student (t)**
- ***Κατανομή* Fisher (F)**

✓ Πρόκειται για συνεχείς συναρτήσεις κατανομής *ανεξάρτητων* τυχαίων μεταβλητών

✓ Είναι αξιοσημείωτο, ότι και οι τρεις έχουν ως αφετηρία την *κανονική κατανομή*

- *Κατανομή* ***χ2*** – παρέχει διαστήματα εμπιστοσύνης της μεταβλητότητας του πληθυσμού ή της εκτίμησής της από ένα δείγμα
- *Κατανομή* ***Student (t)*** – παρέχει διαστήματα εμπιστοσύνης μέσης τιμής ή της ακριβούς τιμής όταν δεν είναι γνωστή η μεταβλητότητα του πληθυσμού
- *Κατανομή* ***Fisher (F)*** – παρέχει διαστήματα εμπιστοσύνης της αναλογίας των μεταβλητοτήτων δύο δειγμάτων

Κατανομή χ2

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Κατανομή χ^2

(chi-squared distribution)

- Την εισήγαγε το 1876 ο Γερμανός γεωδαίτης ***Friedrich Robert Helmert***
- Η ***χ^2-κατανομή με k βαθμούς ελευθερίας*** χρησιμοποιείται για να περιγράψει την κατανομή του αθροίσματος των τετραγώνων, $X=Z_1^2+Z_2^2+Z_3^2+...+Z_k^2$, ***k*** ανεξάρτητων τυχαίων μεταβλητών $Z_1, Z_2, ..., Z_k$ που ακολουθούν την τυπική κανονική κατανομή, δηλ. $Z_i \sim N(0, 1), i = 1,2,...,k$

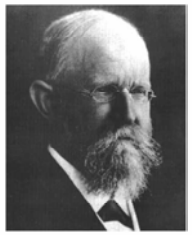
Κατανομή χ2

(chi-squared distribution)

*Για κάθε τιμή του **k** → διαφορετική κατανομή*

Προφανώς πρόκειται για οικογένεια κατανομών

$$X=\sum_{i=1}^{k} Z_i^2 \rightarrow X \sim \chi^2(k)$$



F.R. Helmert (1843–1917)

Κατανομή χ^2
(chi-squared distribution)

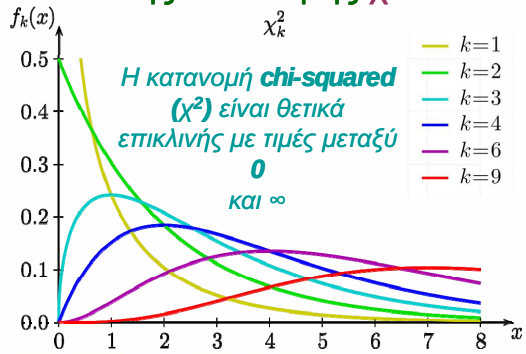
Επιτρέπει την αναζήτηση λύσεων στο ακόλουθο πρόβλημα:

- Έστω ένα δείγμα n-στοιχείων που ακολουθούν την τ.κ.κ. N(0,1) → η κατανομή της μέσης τιμής του δείγματος είναι N(0,1/n). ***Ποια όμως είναι η κατανομή από την οποία προέρχεται η διακύμανση S^2 του δείγματος;***

Κατανομή χ^2
(chi-squared distribution)

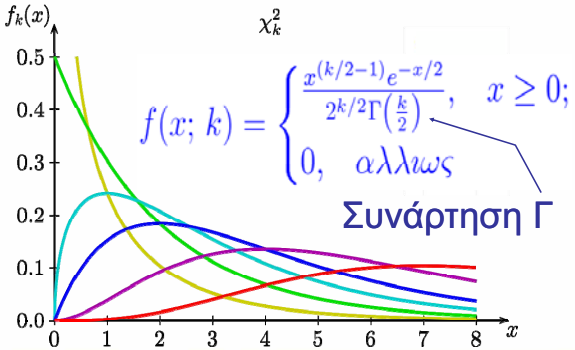
- Στον πραγματικό κόσμο πολύ λίγες παρατηρήσεις ακολουθούν μια κατανομή χ^2.
- Ο κύριος σκοπός των κατανομών χ^2 είναι ο έλεγχος υποθέσεων ***(π.χ. της ανεξαρτησίας δειγμάτων ή της καλής προσαρμογής ενός μοντέλου δεδομένων)*** και όχι η περιγραφή των κατανομών του πραγματικού κόσμου.
  - Αυτό λόγω της στενής σχέσης τους με την τυπική κανονική κατανομή

Συνάρτηση πυκνότητας πιθανότητας της κατανομής χ^2

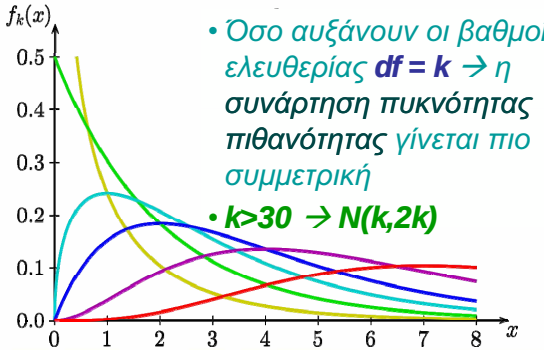


*Η κατανομή **chi-squared** (χ^2) είναι θετικά επικλινής με τιμές μεταξύ **0** και ∞*

Συνάρτηση πυκνότητας πιθανότητας της κατανομής χ^2

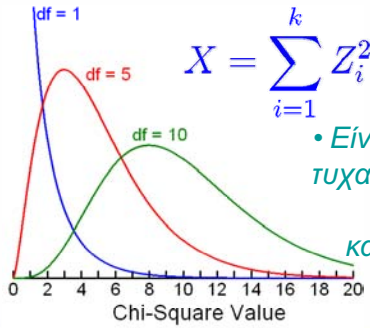


Συνάρτηση πυκνότητας πιθανότητας της κατανομής χ^2



- *Όσο αυξάνουν οι βαθμοί ελευθερίας **df = k** → η συνάρτηση πυκνότητας πιθανότητας γίνεται πιο συμμετρική*
- ***k>30 → N(k,2k)***

Συνάρτηση πυκνότητας πιθανότητας της κατανομής χ^2



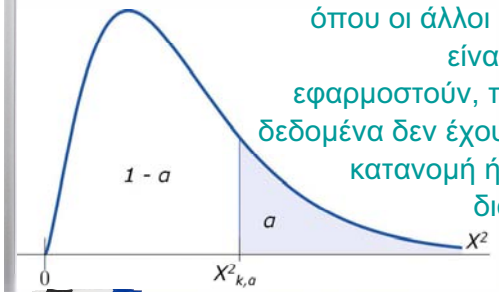
$$X=\sum_{i=1}^{k} Z_i^2,\quad X \sim \chi^2_k$$

- *Είναι προφανές ότι μια τυχαία μεταβλητή X που ακολουθεί μια χ^2_n κατανομή δεν παίρνει αρνητικές τιμές.*

Συνάρτηση πυκνότητας πιθανότητας της κατανομής χ^2

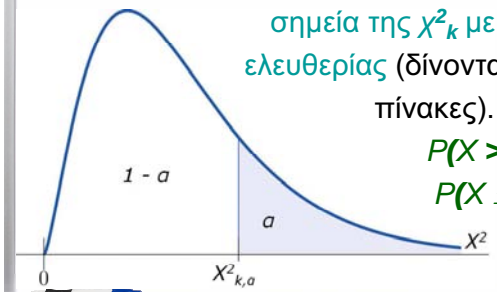
- *Η τιμή $\chi^2_{k,\alpha}$* μας λέει τι πιθανότητα έχουμε να πάρουμε ένα συγκεκριμένο δείγμα από ένα πληθυσμό που ακολουθεί τη συγκεκριμένη κατανομή

Συνάρτηση πυκνότητας πιθανότητας της κατανομής χ^2



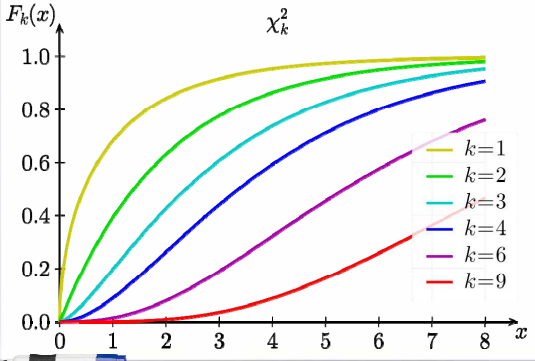
- Χρησιμεύει σε περιπτώσεις όπου οι άλλοι έλεγχοι δεν είναι εύκολο να εφαρμοστούν, π.χ. όταν τα δεδομένα δεν έχουν κανονική κατανομή ή έχουν ίσες διακυμάνσεις

Συνάρτηση πυκνότητας πιθανότητας της κατανομής χ^2

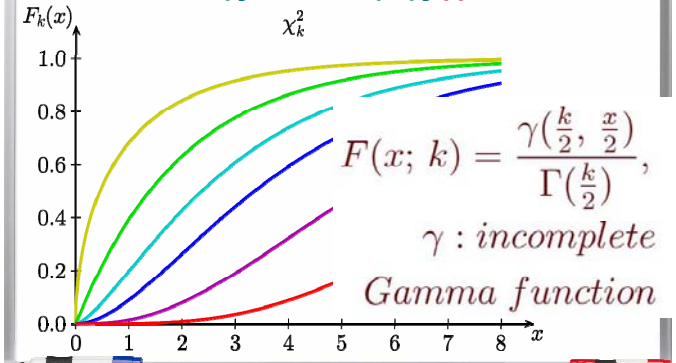


- $\chi^2_{k,\alpha}$ ή $\chi^2_k(\alpha)$: α-ποσοστιαία σημεία της χ^2_k με **k** βαθμούς ελευθερίας (δίνονται π.χ., από πίνακες). Αν $X \sim \chi^2_k$, $P(X > \chi^2_{k,\alpha})=\alpha$, ή $P(X \leq \chi^2_{k,\alpha})=1-\alpha$

Αθροιστική συνάρτηση πιθανότητας της κατανομής χ^2



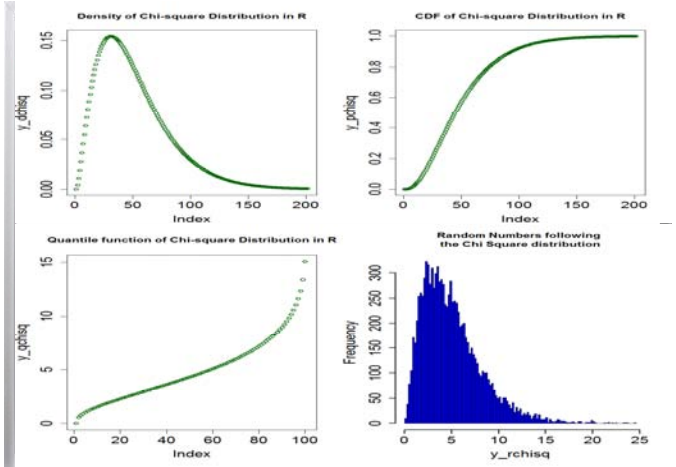
Αθροιστική συνάρτηση πιθανότητας της κατανομής χ^2



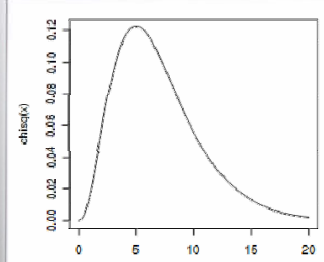
Κατανομή χ^2 *(chi-squared distribution)*

- ***Χαρακτηριστικές παράμετροι***
  - *Μέση τιμή* $\mu = k$ *(=k, βαθμοί ελευθερίας)*
  - *Διακύμανση,* $\sigma^2 = 2k$
  - *Διάμεσος,* $M \approx k\,[\,1 - (2/9k)\,]^3$
  - *Επικρατούσα τιμή,* $\max(k-2,\,0)$
  - *Λοξότητα ή ασυμμετρία,* $\gamma = \sqrt{8/k}$
  - *Κύρτωση,* $\beta = 12/k$


```
Usage of probability functions of Chisq (χ2) distribution in R
dchisq(x, df, ncp = 0, log = FALSE)
pchisq(q, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
qchisq(p, df, ncp = 0, lower.tail = TRUE, log.p = FALSE)
rchisq(n, df, ncp = 0)

x, q   vector of quantiles.
p      vector of probabilities.
n      number of observations. If length(n) > 1,
       the length is taken to be the number required.
df     degrees of freedom (non-negative, but can be non-integer).
ncp    non-centrality parameter (non-negative).
log, log.p   logical; if TRUE, probabilities p are given as log(p).
lower.tail   logical; if TRUE (default), probabilities are P[X≤x],
             otherwise, P[X>x].
```


```
# Calculate the density for the integer values 4 to 8 of
#     a χ2-curve with df=7
dchisq(4:8, df = 7)
[1] 0.11518073 0.12204152 0.11676522 0.10411977 0.08817914

# Calculate the area under the curve for the interval [0,6]
#     and the interval [6,∞) of a χ2-curve with df=7.
#     Further, we ask R if the sum of the intervals [0,6] and [6,∞) sums up to 1.

# interval $[0,6]
pchisq(6, df = 7,lower.tail = TRUE)
[1] 0.4602506

# interval $[6,inf]
pchisq(6, df = 7,lower.tail = FALSE)
[1] 0.5397494
pchisq(6, df = 7,lower.tail = TRUE) +
  pchisq(6, df = 7,lower.tail = FALSE) == 1
[1] TRUE
```

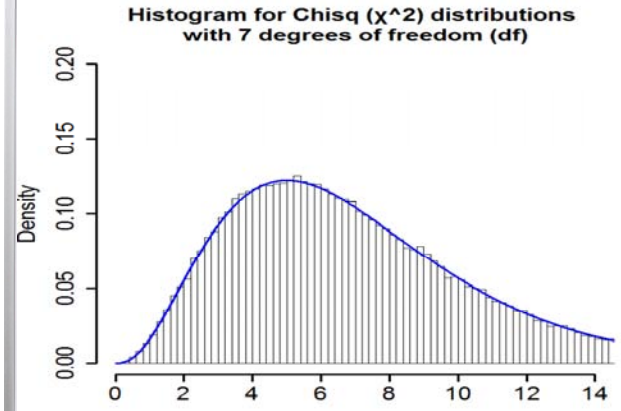
Στο R, εφαρμόζουμε τη συνάρτηση (***quantile***) ***qchisq*** της κατανομής Chi-Squared

```
Ποιό είναι το 95ο εκατοστημόριο της κατανομής
Chi-Squared με 7 βαθμούς ελευθερίας

qchisq (.95, df = 7) # 7 βαθμοί ελευθερίας
[1] 14.067
```



```
# Calculate the quantile for a given area (= probability) under
# the curve for a χ2-curve with df=7 that corresponds
# to q=0.25,0.5,0.75 and 0.999. We set lower.tail = TRUE
# in order the get the area for the interval [0,q]
qchisq(0.25, df=7,lower.tail = TRUE)
[1] 4.254852
qchisq(0.5, df=7,lower.tail = TRUE)
[1] 6.345811
qchisq(0.75, df=7,lower.tail = TRUE)
[1] 9.037148
qchisq(0.999, df=7,lower.tail = TRUE)
[1] 24.32189
```

```
x <- rchisq(100000, df = 7)
hist(x,
     breaks = 'Scott', freq = FALSE, xlim = c(0,14),
     ylim = c(0,0.2), xlab = '',
     main = "Histogram for Chisq (χ^2) distributions
     with 7 degrees of freedom (df)", cex.main=1.5,
     cex.lab=1.5, cex.axis=1.5, lwd=3)
curve(dchisq(x, df = 7), from = 0, to = 15, n = 5000,
     col='blue', lwd=3, add = T)
```

Κατανομή Student (t)

- Την εισήγαγε, το 1908, ο ***W. S. Gosset*** *(1876-1937)*
- Ως επικεφαλής εμπειρογνώμονας στα αγροκτήματα της ζυθοποιίας *Guinness*, στο Δουβλίνο, εφάρμοσε τις στατιστικές γνώσεις του, για την επιλογή των καλύτερων αποδόσεων των διαφόρων ποικιλιών κριθαριού για την παρασκευή μπύρας

- Ονομάζεται *κατανομή Student*, ή *Student t-κατανομή*, από το ψευδώνυμο με το οποίο ο *Gosset* δημοσίευσε τη σχετική εργασία του '*The probable error of a mean*' στο περιοδικό *Biometrika* (http://www.york.ac.uk/depts/maths/histstat/student.pdf) προκειμένου να αποφύγει την απαγόρευση της Guinness για τη δημοσιοποίηση ευαίσθητων στοιχείων, ενδεχομένως χρήσιμων στους ανταγωνιστές της
- *Ορθότερα θα έπρεπε να είχε ονομαστεί Gosset t-κατανομή*

Κατανομή Student (t)

- Έστω $X \sim N(\mu, \sigma^2)$ μια τυχαία μεταβλητή η οποία ακολουθεί την *κανονική κατανομή*, με μέση τιμή *μ*, και διακύμανση σ^2
- Η μέση τιμή $x \sim N(\mu, \sigma^2/n)$ ενός δείγματος με *n*-παρατηρήσεις της τυχαίας μεταβλητής ακολουθεί επίσης την *κανονική κατανομή*, με μέση τιμή *μ*, και διακύμανση σ^2/n

Κατανομή Student (t)

$$Z = \frac{\overline{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}.(\overline{x} - \mu)}{\sigma}$$

- Η τυποποιημένη μεταβλητή *Z* εκφράζει τη διαφορά του μέσου του δείγματος από την πραγματική μέση τιμή *μ* της κατανομής, και έχει τυποποιηθεί με τη διακύμανση της δικής της κατανομής, ώστε $Z \sim N(0, 1)$

Κατανομή Student (t)

- Εάν η **διακύμανση σ^2 *της κανονικής κατανομής είναι γνωστή***, *η μέση τιμή του δείγματος μετατρέπεται εύκολα σε μια τυποποιημένη κανονική μεταβλητή*
- *Εάν όμως η διακύμανση σ^2 είναι άγνωστη, αλλά είναι μόνο γνωστή η διακύμανση του δείγματος S^2 (αμερόληπτη εκτίμηση του σ^2) …*

$$S^2 = \frac{1}{n-1}\sum_{1}^{n}(x_i - \overline{x})^2$$

Κατανομή Student (t)

… λαμβάνεται έτσι η μεταβλητή

$$T = \frac{(\overline{x} - \mu)}{\frac{S}{\sqrt{n}}} = \frac{\sqrt{n}.(\overline{x} - \mu)}{S}$$

όπου όμως η *S* είναι μια τυχαία μεταβλητή που εμποδίζει την κατανομή της *T* να είναι κανονική

⇓

ποια είναι η κατανομή που ακολουθεί η μεταβλητή T ?

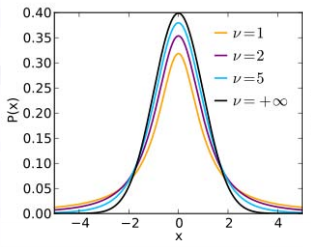
Κατανομή Student (t)

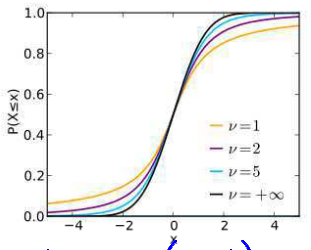
- Έστω μια τυχαία μεταβλητή $Z \sim N(0, 1)$, και
- S_ν μια τ.μ. *ανεξάρτητη* από την *Z*, η οποία ακολουθεί την κατανομή χ_ν^2 , δηλ. $S_\nu = (S^2/\sigma^2) \sim \chi_\nu^2$.
- Τότε, η κατανομή t_ν της τυχαίας μεταβλητής

$$T = \frac{\sqrt{n}.(\overline{x} - \mu)}{S} = \frac{\sqrt{n}.(\overline{x} - \mu)/\sigma}{S/\sigma}$$ ή $T = Z / (S_\nu/\nu)^{1/2}$

ονομάζεται ***t κατανομή του Student*** *ή απλά* ***t κατανομή***, $T \sim t_\nu$, *με ν βαθμούς ελευθερίας (ν=n-1). Το γράμμα T χρησιμοποιείται για να αναχωρίσει την τυχαία μεταβλητή από το όνομα της κατανομής*

- *Η κατανομή Student είναι μια οικογένεια συνεχών κατανομών που προκύπτει κατά τον υπολογισμό της τιμής (& των διαστημάτων εμπιστοσύνης) του μέσου μιας κανονικής κατανομής ενός πληθυσμού σε καταστάσεις όπου το μέγεθος του δείγματος είναι μικρό και δεν είναι γνωστή η μεταβλητότητα του πληθυσμού*
- *Βασικό χαρακτηριστικό της είναι ότι δεν εξαρτάται από τη διακύμανση σ^2 του αρχικού πληθυσμού της τυχαίας μεταβλητής από την οποία προέρχεται*





$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \quad \frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$$

Σ.π.π. (PDF) της κατανομής Student

$$f_\nu(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

όπου *v* είναι οι βαθμοί ελευθερίας και η *συνάρτηση Γ (γάμμα)* δίνεται ως ***Γ(k) = (k-1)(k-2)…(2)(1)=(k-1)!***

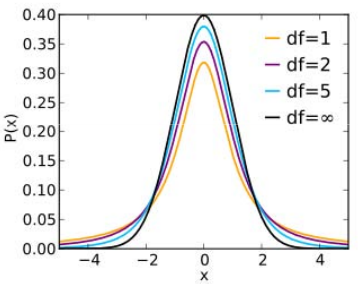
$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots 5\cdot 3}{2\sqrt{\nu}(\nu-2)(\nu-4)\cdots 4\cdot 2}$$

v: ζυγός αριθμός

$$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} = \frac{(\nu-1)(\nu-3)\cdots 4\cdot 2}{\pi\sqrt{\nu}(\nu-2)(\nu-4)\cdots 5\cdot 3}$$

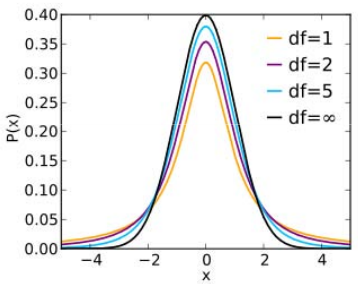
v: μονός αριθμός

Σ.π.π. (PDF) της κατανομής Student



- Όπως και στην τυπική κανονική κατανομή, η συνάρτηση της πυκνότητας πιθανότητας για την *t-κατανομή* έχει το σχήμα «καμπάνας» και είναι συμμετρική ως προς το μηδέν

Σ.π.π. (PDF) της t-κατανομής

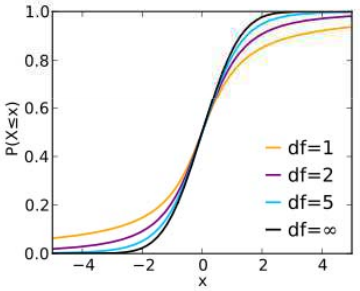


- *E(T) = μ = 0, v ≥ 1*
- *Var(T) = σ = v/(v-2), v > 2*
- Όσο ο αριθμός *v* των βαθμών ελευθερίας αυξάνει, η ***t*** κατανομή προσεγγίζει την τυπική κανονική κατανομή

Αθροιστική συνάρτηση κατανομής πιθανότητας της t-κατανομής

- όπου *F1* είναι η λεγόμενη *υπεργεωμετρική συνάρτηση*

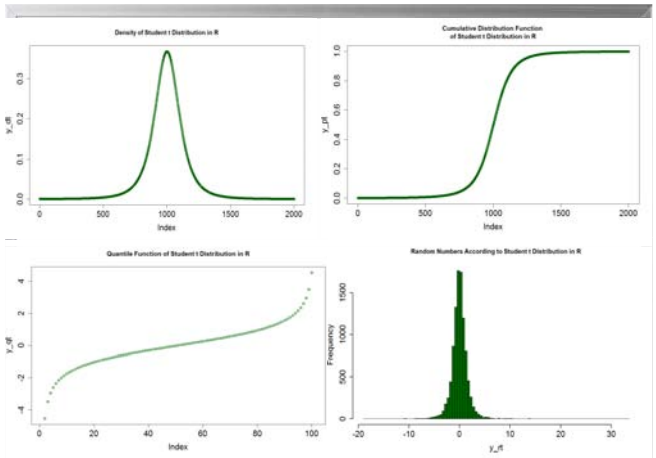
$$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2},\frac{\nu+1}{2};\frac{3}{2};-\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}$$




```
Usage of probability functions for the Student distribution
dt(x, df, ncp, log = FALSE)
pt(q, df, ncp, lower.tail = TRUE, log.p = FALSE)
qt(p, df, ncp, lower.tail = TRUE, log.p = FALSE)
rt(n, df, ncp)

x, q   vector of quantiles.
p      vector of probabilities.
n      number of observations. If length(n) > 1,
       the length is taken to be the number required.
df     degrees of freedom (> 0, maybe non-integer).
       df = Inf is allowed.
ncp    non-centrality parameter delta;
       currently except for rt(), only for abs(ncp) <= 37.62.
       If omitted, use the central t distribution.
log, log.p   logical; if TRUE, probabilities p are given
                      as log(p).
lower.tail   logical; if TRUE (default),
             probabilities are P[X ≤ x], otherwise, P[X > x].
```



```
x_dt <- seq(- 10, 10, by = 0.01)  # Specify x-values for dt function
y_dt <- dt(x_dt, df = 3)          # Apply the dt function
plot(y_dt, main="Density of Student t Distribution in R",
   col='darkgreen',cex.lab=1.5,cex.axis=1.5)        # Plot dt values

x_pt <- seq(- 10, 10, by = 0.01)  # Specify x-values for pt function
y_pt <- pt(x_pt, df = 3)          # Apply pt function
plot(y_pt, main="Cumulative Distribution Function
   of Student t Distribution in R",
   col='darkgreen',cex.lab=1.5,cex.axis=1.5)        # Plot dt values

x_qt <- seq(0, 1, by = 0.01)      # Specify x-values for qt function
y_qt <- qt(x_qt, df = 3)          # Apply qt function
plot(y_qt, main="Quantile Function of
   Student t Distribution in R",
   col='darkgreen',cex.lab=1.5,cex.axis=1.5, lwd=2) # Plot dt values
set.seed(15-12-2019)              # Set seed for reproducibility
N <- 10000                        # Specify the sample size
y_rt <- rt(N, df = 3)             # Draw N log normally distributed values
y_rt                              # Print values to RStudio console
hist(y_rt,                        # Plot of randomly drawn student t density
     breaks = 100,
     main="Random Numbers According to Student t Distribution in R",
   col='darkgreen',cex.lab=1.5,cex.axis=1.5, lwd=2)
```

- Η συνάρτηση *dt()* επιστρέφει την τιμή της συνάρτησης πυκνότητας πιθανότητας (pdf) της κατανομής Student t για μια τυχαία μεταβλητή *x* και βαθμούς ελευθερίας *df*

```
Παραδείγματα χρήσης της συνάρτησης dt()
dt(x = 0, df = 20)   # find the value of the Student t distribution pdf at x = 0
                     # with 20 degrees of freedom
[1] 0.3939886
dt(0, 20)            # by default, R assumes the first argument is x
                     # and the second argument is df
[1] 0.3939886

dt(1, 30)            # find the value of the Student t distribution pdf at x = 1
                     # with 30 degrees of freedom
[1] 0.2379933
```



```
# Find the area to the left of a t-statistic with a value of -0.785
#    # and 14 degrees of freedom
pt(-0.785, 14)
[1] 0.2227675

# the following approaches give equivalent results
# 1 - area to the left
1 - pt(-0.785, 14)
[1] 0.7772325

# area to the right
pt(-0.785, 14, lower.tail = FALSE)
[1] 0.7772325

# Find the total area in a Student t distribution with
#  14 degrees of freedom that lies to the left of -0.785
#  or to the right of 0.785.
pt(-0.785, 14) + pt(0.785, 14, lower.tail = FALSE)
[1] 0.4455351
```

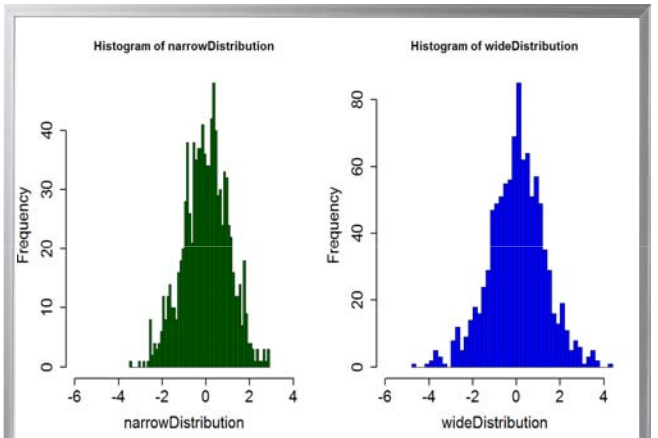
- Η συνάρτηση *pt()* επιστρέφει την τιμή της αθροιστικής σ.π.π. (cdf) της κατανομής Student για μια τυχαία μεταβλητή *x* και βαθμούς ελευθερίας *df*.

```
# Find the t-score of the 99th quantile
#      of the Student t distribution with df = 20
qt(.99, df = 20)
[1] 2.527977

# Find the t-score of the 95th quantile
#      of the Student t distribution with df = 20
qt(.95, df = 20)
[1] 1.724718

# Find the t-score of the 90th quantile
#      of the Student t distribution with df = 20
qt(.9, df = 20)
[1] 1.325341
```

- Η *qt* επιστρέφει την τιμή της αντίστροφης αθροιστικής σ.π.π. (cdf) της κατανομής Student για μια συγκεκριμένη τ.μ. *x* και βαθμούς ελευθερίας *df*.


```
ΠΑΡΑΔΕΙΓΜΑ ΧΡΗΣΗΣ ΤΗΣ ΣΥΝΑΡΤΗΣΗΣ dr()
# Generate a vector of 5 random variables that follow
#    a Student t distribution with df = 20
rt(n = 5, df = 20)

[1] -1.7422445  0.9560782  0.6635823  1.2122289 -0.7052825

# Ditto for a vector of 1000 random variables that follow
#    a Student t distribution with df = 40, and then with df = 5
narrowDistribution <- rt(1000, 40)
wideDistribution <- rt(1000, 5)

# Generate two histograms to view these two distributions side by side,
#    using e.g. 50 bars in the histogram,
par(mfrow=c(1, 2)) #one row, two columns
hist(narrowDistribution, breaks=50, xlim = c(-6, 4),
  col='darkgreen',cex.lab=1.5,cex.axis=1.5, lwd=2)
hist(wideDistribution, breaks=50, xlim = c(-6, 4),
  col='blue',cex.lab=1.5,cex.axis=1.5, lwd=2)
```

- Η συνάρτηση *rt()* παράγει ένα διάνυσμα τυχαίων μεταβλητών που ακολουθούν μια κατανομή Student δεδομένου ενός μήκους διανύσματος *n* και βαθμών ελευθερίας *df*.



Κατανομή Fisher

- Την εισήγαγε το 1934 ο *George W. Snedecor*, *δίνοντάς της την ονομασία Fisher τιμώντας τον διακεκριμένο στατιστικολόγο και γενετιστή R.A. Fisher*
- *Συναντάται με τις ονομασίες*
  - *κατανομή Fisher, ή*
  - *Snedecor F κατανομή, ή*
  - *κατανομή Snedecor-Fisher*

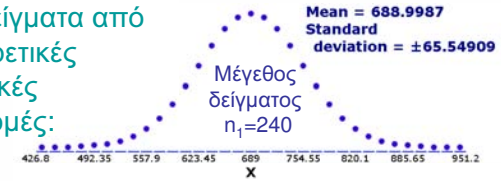
Roland Aylmer Fisher (1890–1962)

Κατανομή Fisher

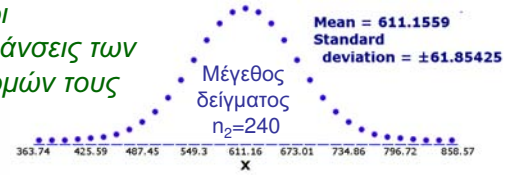
- Παρέχει διαστήματα εμπιστοσύνης του λόγου της διασποράς δύο δειγμάτων
- Εάν δύο τυχαίες μεταβλητές ακολουθούν την κατανομή χ^2: $S_n \sim \chi^2_n$ και $S_d \sim \chi^2_d$
  → η κατανομή της τ.μ.
  $F = (S_n/n)/(S_d/d) = F_{n,d}$
  ονομάζεται ***F κατανομή*** με ***δύο βαθμούς ελευθερίας***
  - *n (nominator)* ***ή*** d_1 για τον αριθμητή ***και***
  - *d (denominator)* ***ή*** d_2 για τον παρανομαστή

Σκεπτικό της κατανομής Fisher

- Δύο δείγματα από διαφορετικές κανονικές κατανομές:

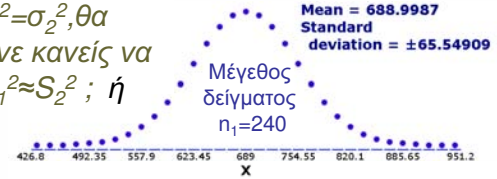


- *Είναι οι διακυμάνσεις των κατανομών τους ίσες ;*

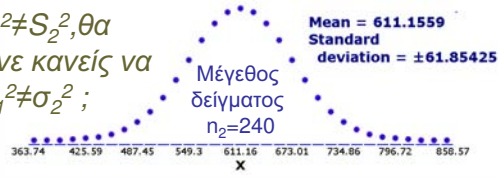


Σκεπτικό της κατανομής Fisher

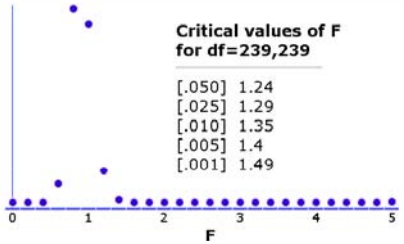
- *Εάν $\sigma_1^2=\sigma_2^2$,θα περίμενε κανείς να είναι $S_1^2 \approx S_2^2$; ή*



- *Εάν $S_1^2 \neq S_2^2$,θα περίμενε κανείς να είναι $\sigma_1^2 \neq \sigma_2^2$;*

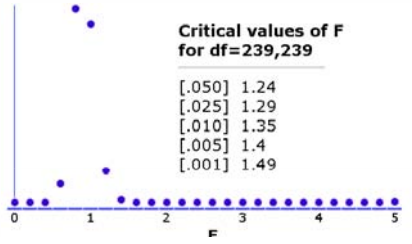


Χαρακτηριστικά της κατανομής Fisher



- Είναι μια ασύμμετρη κατανομή που έχει μια ελάχιστη τιμή από 0, αλλά όχι μέγιστη τιμή
- Πλησιάζει, αλλά ποτέ δεν αγγίζει αρκετά τον οριζόντιο άξονα

Χαρακτηριστικά της κατανομής Fisher



- Η καμπύλη της κατανομή F είναι πιο απλωμένη όταν οι βαθμοί ελευθερίας είναι μικροί → όσο αυξάνονται οι βαθμοί ελευθερίας, έχει μικρότερη διασπορά

Συνάρτηση π.π. κατανομής Fisher

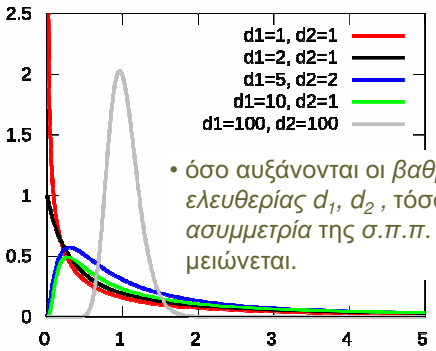
- Πρόκειται για οικογένεια κατανομών: Για κάθε συνδυασμό των βαθμών ελευθερίας d_1, d_2 υπάρχει μια διαφορετική κατανομή F

Συνάρτηση 'Βήτα'

$$f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} \, d_2^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x \, \mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)}$$

$$= \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2} x\right)^{-\frac{d_1+d_2}{2}}$$

Συνάρτηση π.π. κατανομής Fisher



- όσο αυξάνονται οι *βαθμοί ελευθερίας* d_1, d_2 , τόσο η (*θετική*) *ασυμμετρία* της *σ.π.π.* της $F_{n,d}$ μειώνεται.

Αθροιστική σ.π. κατανομής Fisher

$$F(x; d_1, d_2) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$$

Ελλιπής συνάρτηση 'Βήτα'

$$B(x, a, b) = \int_0^x t^{a-1} (1-t)^{b-1} \, \mathrm{d}t$$

Κανονικοποιημένη ελλιπής συνάρτηση 'Βήτα'

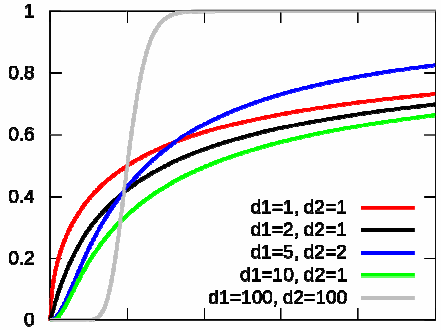
$$I_x(a, b) = \frac{B(x, a, b)}{B(a, b)}$$

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} \, \mathrm{d}t =$$

$$= \frac{\Gamma(x)\,\Gamma(y)}{\Gamma(x+y)} = B(y, x)$$

Συνάρτηση 'Βήτα'

Αθροιστική σ.π. κατανομής Fisher



Στατιστικά μέτρα κατανομής Fisher

- Μια τυχαία μεταβλητή $X \sim F_{n,d}$ δεν παίρνει αρνητικές τιμές
- μ – μέση τιμή
- σ – τυπική απόκλιση
- γ - ασυμμετρία

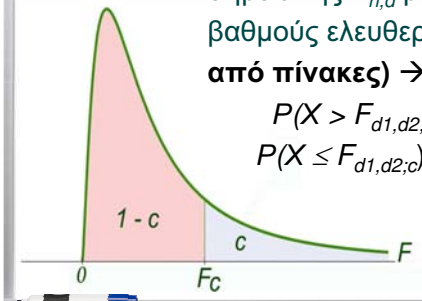
$$\mu = \frac{d_2}{d_2 - 2}$$

$$\sigma^2 = \frac{2\,d_2^2\,(d_1 + d_2 - 2)}{d_1 (d_2 - 2)^2 (d_2 - 4)}, \quad d_2 > 4$$

$$\gamma = \frac{(2d_1 + d_2 - 2)\sqrt{8(d_2 - 4)}}{(d_2 - 6)\sqrt{d_1(d_1 + d_2 - 2)}}, \quad d_2 > 6$$

Περιοχές (επίπεδα) σημαντικότητας

- $F_{d1,d2;c}$ ή $F_{d1,d2}(c)$: α-ποσοστιαία σημεία της $F_{n,d}$ με *d1 και d2* βαθμούς ελευθερίας **(δίνονται από πίνακες)** → εάν $X \sim F_{d1,d2}$, $P(X > F_{d1,d2;c}) = c$, ή $P(X \le F_{d1,d2;c}) = 1\text{-}c$



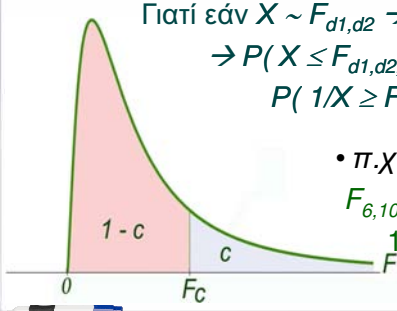
Περιοχές (επίπεδα) σημαντικότητας

- *Μπορεί να αποδειχθεί ότι* $F_{d1,d2,\,1-c} = 1/F_{d1,d2,\,c}$

Γιατί εάν $X \sim F_{d1,d2}$ → $1/X \sim F_{d1,d2,}$

→ $P(X \le F_{d1,d2,\,1-c}) = c \Leftrightarrow$

$P(1/X \ge F_{d1,d2,\,1-c}) = c$



- *π.χ.* $F_{6,10;0.95}$ = ????

$F_{6,10;0.95} = F_{6,10;1-0.05} = 1/F_{6,10;0.95} = 0.246$

ΟΙ ΣΥΝΑΡΤΗΣΕΙΣ ΠΙΘΑΝΟΤΗΤΑΣ ΓΙΑ ΤΗΝ ΚΑΤΑΝΟΜΗ FISHER

```
df(x, df1, df2, ncp, log = FALSE)
pf(q, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
qf(p, df1, df2, ncp, lower.tail = TRUE, log.p = FALSE)
rf(n, df1, df2, ncp)

Arguments
x, q        vector of quantiles.
p           vector of probabilities.
n           number of observations. If length(n) > 1,
            the length is taken to be the number required.
df1, df2    degrees of freedom. Inf is allowed.
ncp         non-centrality parameter. If omitted the central F
            is assumed.
log, log.p    logical; if TRUE, probabilities p are given as log(p).
lower.tail    logical; if TRUE (default), probabilities are P[X ≤ x],
            otherwise, P[X > x].
```

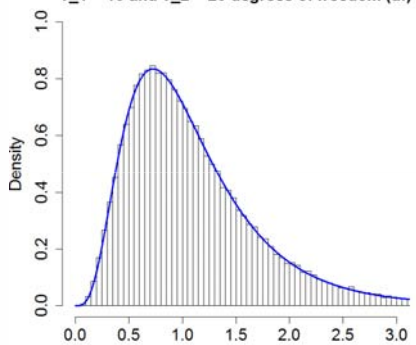


```
# Calculate the density at the value of 1.2 of a F-curve with v1=10 and v2=20
df(1.2, df1 = 10, df2 = 20)
[1] 0.5626125

# Calculate the area under the curve for the interval [0,1.5] and the interval
# [1.5,+∞) of a F-curve with with v1=10 and v2=20. Further ask R if the
# sum of the intervals [0,1.5] and [1.5,+∞) sums up to 1.
x = 1.5
v1 = 10
v2 = 20
# interval $[0,1.5]
pf(x, df = v1, df2 = v2, lower.tail = TRUE)
[1] 0.7890535
# interval $[1.5,+inf)
pf(x, df = v1, df2 = v2, lower.tail = FALSE)
[1] 0.2109465
pf(x, df = v1, df2 = v2, lower.tail = TRUE) +
  pf(x, df = v1, df2 = v2, lower.tail = FALSE) == 1
[1] TRUE
```

```r
# Calculate the quantile for a given area (= probability) under the curve
# for a F-curve with v1=10 and v2=20 that corresponds to q=0.25,0.5,0.75
# and 0.999. By setting lower.tail = TRUE, we get the area for the interval [0,q]
q <- c(0.25, 0.5, 0.75, 0.999)
v1=10
v2=20

qf(q[1], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 0.6563936

qf(q[2], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 0.9662639

qf(q[3], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 1.399487

qf(q[4], df1 = v1, df2 = v2, lower.tail = TRUE)
[1] 5.075246
```

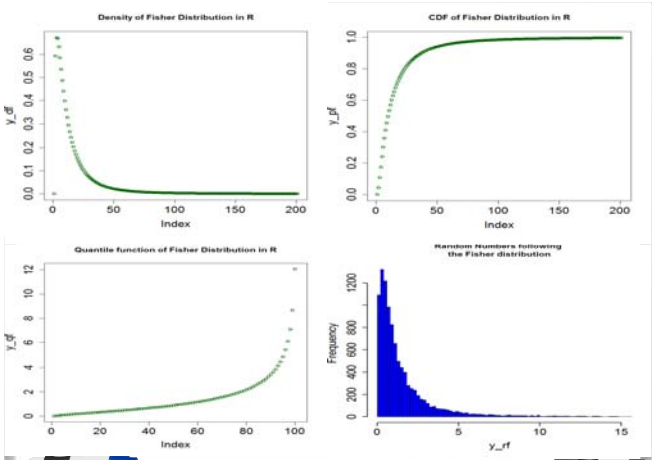


```r
# Use the rf() function to generate 100000 random values from
# the F-distribution with v1=10 and v2=20. Then plot a histogram and
# compare it to the probability density function of the F-distribution
# with v1=10 and v2=20 (blue line).

x <- rf(100000, df1 = 10, df2 = 20)
hist(x,
     breaks = 'Scott',
     freq = FALSE,
     xlim = c(0,3),
     ylim = c(0,1),
     xlab = '',
     main = "Histogram for a F-distribution with
             v_1 = 10 and v_2 = 20 degrees of freedom (df)",
             cex.lab=1.5, cex.axis=1.5, cex.main=1.5)

curve(df(x, df1 = 10, df2 = 20), from = 0, to = 4,
      n = 5000, col= 'blue', lwd=3, add = T)
```


```r
x_df <- seq(0, 20, by = 0.1)          # Specify x-values for df function
y_df <- df(x_df, df1 = 3, df2 = 5)    # Apply df function
plot(y_df, main="Density of Fisher Distribution in R",
  ylab= "", cex.lab=1.5 cex.axis=1.5  lwd=2)

x_pf <- seq(0, 20, by = 0.1)          # Specify x-values for pf function
y_pf <- pf(x_pf, df1 = 3, df2 = 5)    # Apply pf function
plot(y_pf, main="CDF of Fisher Distribution in R",
  ylab= "", cex.lab=1.5 cex.axis=1.5  lwd=2)

x_qf <- seq(0, 1, by = 0.01)          # Specify x-values for qf function
y_qf <- qf(x_qf, df1 = 3, df2 = 5)    # Apply qf function
plot(y_qf, main="Quantile function of Fisher Distribution in R",
  ylab= "", cex.lab=1.5 cex.axis=1.5  lwd=2)  # Plot qf values
```

```r
set.seed(55555)                       # Set seed for reproducibility
N <- 10000                            # Specify sample size
y_rf <- rf(N, df1 = 3, df2 = 5)       # Draw N F-distributed values
y_rf                                  # Print values to R console
hist(y_rf,                            # Plot of randomly drawn f density
     breaks = 500,
     main = "Random Numbers following
     the Fisher distribution",
     col='blue',cex.lab=1.5,cex.axis=1.5, lwd=2, xlim = c(0, 15))
```

Δύο σημαντικές διακριτές τ.μ.

- Η **διωνυμική τυχαία μεταβλητή** και
- η **τυχαία μεταβλητή Poisson**

και *οι αντίστοιχες κατανομές που αυτές ακολουθούν*.

- Θεωρούνται ιδιαίτερα σημαντικές γιατί έχουν ως οριακή περίπτωση την **κανονική κατανομή**, η χρήση της οποίας κυριαρχεί στα προβλήματα όπου εφαρμόζεται η μέθοδος των ελαχίστων τετραγώνων

- Διωνυμική κατανομή
  - *n* στατιστικά ανεξάρτητες δοκιμές, η κάθε μια από τις οποίες έχει μόνο δυο δυνατά και αντίθετα αποτελέσματα (→ σταθερές πιθανότητες σε όλες τις δοκιμές)
- Κατανομή Poisson
  - Αφορούν ακολουθίες τυχαίων ανεξάρτητων ενδεχομένων που πραγματοποιούνται με ένα σταθερό μέσο ρυθμό *λ* ανά μονάδα χρόνου
    - χρησιμοποιείται για να περιγράψει τον αριθμό των εμφανίσεων «σπάνιων» ενδεχόμενων.

Διωνυμική κατανομή

- **Διωνυμική τ.μ.** *(binomial random variable)* ορίζεται ως μια τυχαία μεταβλητή *X* που αναπαριστά τον αριθμό των επιτυχιών σε *n* ανεξάρτητες επαναλήψεις ενός πειράματος τύχης, της παρατήρησης ενός φαινομένου ή μιας μετρητικής διαδικασίας, με δυο πιθανά αποτελέσματα (επιτυχία - αποτυχία) και την *πιθανότητα* της "επιτυχίας" να είναι ίση με *p* και την πιθανότητα της "αποτυχίας" ίση με *q=1-p*

Πρακτικά παραδείγματα καταστάσεων ή συμβάντων που ακολουθούν τη διωνυμική κατανομή

- Ο αριθμός των όψεων κεφαλή / γράμματα σε μια σειρά από ρίψεις ενός νομίσματος
- Η ψηφοφορία για δύο διαφορετικούς υποψηφίους στις πρυτανικές εκλογές του ΕΜΠ
- Ο αριθμός των ανδρών / γυναικών εργαζομένων σε μια εταιρεία
- Ο αριθμός των ελαττωματικών προϊόντων σε ένα στάδιο παραγωγής
- Ο αριθμός των ημερών του μήνα που ένα δίκτυο Η/Υ αντιμετωπίζει κάποιο πρόβλημα
- ...

Διωνυμική κατανομή - Ιδιότητες

1. Το δείγμα αποτελείται από ένα σταθερό αριθμό παρατηρήσεων, *n*.
2. Κάθε παρατήρηση κατατάσσεται σε μία από τις δύο αλληλοαποκλειόμενες και συλλογικά εξαντλούμενες κατηγορίες.
3. Η πιθανότητα *p*, μια παρατήρηση να ταξινομηθεί ως έκβαση (γεγονός) ενδιαφέροντος, είναι σταθερή από παρατήρηση σε παρατήρηση.
4. Το αποτέλεσμα της κάθε παρατήρησης είναι ανεξάρτητο από την έκβαση οποιασδήποτε άλλης παρατήρησης

- Συμβολικά *X ~ Bin(n,p)* και λέμε ότι η *X* ακολουθεί τη **διωνυμική κατανομή** *(binomial distribution)*. Οι ποσότητες *n, p (ή q=1-p)* καλούνται **παράμετροι** της διωνυμικής κατανομής

$$f(k;n,p) = P(X=k) =$$

$$= \frac{k!}{k!(n-k)!}\, p^k (1-p)^{n-k}$$

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

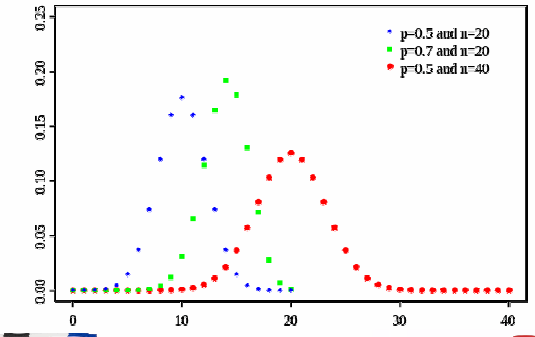
Mean

$$\mu = n\pi$$

$$\sigma = \sqrt{n\pi(1-\pi)}$$

Standard deviation

Μορφή της διωνυμικής κατανομής για διαφορετικές πιθανότητες επιτυχίας p που επαναλαμβάνεται n φορές



- Μέση τιμή *μ* και διακύμανση *σ²* *(τυπική απόκλιση σ)* μιας διωνυμικής κατανομής με παραμέτρους *p* και *n*

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

$$k = 0, 1, 2, \ldots, \quad p \in (0,1), \quad n \in N$$

$$\mu = np\,, \quad \sigma^2 = np(1-p)$$

- Για *n=1* *λαμβάνεται η λεγόμενη κατανομή Bernoulli που επίσης περιγράφει πειράματα τύχης με μόνο δύο, αμοιβαίως αποκλειόμενα, δυνατά αποτελέσματα*

ΣΥΝΑΡΤΗΣΕΙΣ ΠΙΘΑΝΟΤΗΤΑΣ ΓΙΑ ΤΗ ΔΙΩΝΥΜΙΚΗ ΚΑΤΑΝΟΜΗ ΣΤΟ R

```
dbinom(x, size, prob, log = FALSE)
pbinom(q, size, prob, lower.tail = TRUE, log.p = FALSE)
qbinom(p, size, prob, lower.tail = TRUE, log.p = FALSE)
rbinom(n, size, prob)

x, q   numeric vectors in the range [0, size] that specify
       the quantiles.
p      a numeric vector in the range [0, 1] that specifies
       the probabilities.
n      an integer value in the range [0, Inf) that specifies
       the number of random samples requested. If the input value
       is not an integer, it is truncated. If length(n) is greater
       than 1, the random function returns length(n) random samples.
```



```
size   an integer vector in the range [0, Inf) that specifies
       the number of Bernoulli trials (success/failure).
prob   a numeric vector in the range [0, 1] that specifies
       the probability of a success in a Bernoulli trial.
log    a logical value. If FALSE (default), the density function
       returns the density itself. If TRUE, it returns
       the log of the density.
lower.tail a logical value. If TRUE (default), the probability
           supplied to the quantile function or returned by
           the probability function is P[X <= x].
           If FALSE, it is P[X > x].
log.p      a logical value. If FALSE (default), the probability
           supplied to the quantile function or returned by
           the probability function is the probability itself.
           If TRUE, it is the log of the probability.
```


- Η συνάρτηση ***dbinom()*** δίνει την πιθανότητα ενός αριθμού θετικών δοκιμών (x), από μια σειρά από συνολικές δοκιμές *(μέγεθος, size, n)*, με βάση την πιθανότητα του θετικού αποτελέσματος *(prob)*.
- Για παράδειγμα: Εάν ενδιαφέρει η πιθανότητα 5 ενδείξεων Κ *(x = 5)*, από 20 συνολικά ρίψεις ενός νομίσματος (μέγεθος/size= 20), και η πιθανότητα του ενδεχόμενου 'Κ' είναι 0.5 (prob = 0.5) → *dbinom (5, 20, 0.5) ... [1] 0.0147*

- Η συνάρτηση ***pbinom()*** δίνει την πιθανότητα ≤ ή ≥ με τις θετικές δοκιμές (q), από έναν αριθμό συνολικών δοκιμών (μέγεθος/size) και πιθανότητα θετικής έκβασης (prob). Σημειώστε την παράμετρο *lower.tail* στην εντολή.
- Για παράδειγμα: Ποια θα είναι η πιθανότητα για 5 ή λιγότερες ενδείξεις 'Κ' (q = 5) από συνολικά 20 ρίψεις (μέγεθος = 20) ενός μη πειραγμένου (δίκαιου) κέρματος (prob = 0.5);

  → pbinom (5, 20, 0.5, lower.tail = TRUE) *... [1] 0.02069473*

- Η συνάρτηση ***qbinom()*** είναι το αντίθετο της pbinom. Δίνει τον αριθμό θετικών δοκιμών ≤ της πιθανότητας *(p)* θετικών αποτελεσμάτων, για συνολικό αριθμό δοκιμών *(μέγεθος / size)* και την πιθανότητα μιας θετικής δοκιμής *(prob)*. Σημειώστε τη μεταβλητή lower.tail έτσι ώστε να έχουμε μεγαλύτερο ή ίσο αριθμό θετικών δοκιμών.
- Παράδειγμα: εάν θέλουμε να γνωρίζουμε τον αριθμό των ενδείξεων 'Κ' που θα πάρουμε λιγότερο από ή ίσο με το 10% των ρίψεων (p = 0.1), από ένα σύνολο 20 δοκιμών (μέγεθος size = 20), με ένα δίκαιο κέρμα (prob = 0.5)

  → *qbinom (0.1, 20, 0.5) ... [1] 7*

- Η συνάρτηση ***rbinom()*** θα δημιουργήσει μια διωνυμική κατανομή με ένα ορισμένο πλήθος στοιχείων *(n)*, από δείγματα μεγέθους *(size)* και πιθανότητα *(prob)* θετικού αποτελέσματος για κάθε δοκιμή.
- Για παράδειγμα, μπορούμε να δημιουργήσουμε ένα διάνυσμα 1000 στοιχείων, με τις ενδείξεις των ρίψεων ενός δίκαιου νομίσματος:

  → *rbinom (1000,1,0.5) …*

  Ή τα πιθανά αποτελέσματα για 1000 δοκιμές ενός φαρμάκου που έχει ποσοστό επιτυχίας 75%, και για 20 ασθενείς ανά δοκιμή: → *rbinom (1000, 20, 0.75)* .

- Η συνάρτηση ***rbinom()*** θα δημιουργήσει μια διωνυμική κατανομή με ένα ορισμένο πλήθος στοιχείων *(n)*, από δείγματα μεγέθους *(size)* και πιθανότητα *(prob)* θετικού αποτελέσματος για κάθε δοκιμή.
- Για παράδειγμα, μπορούμε να δημιουργήσουμε ένα διάνυσμα 1000 στοιχείων, με τις ενδείξεις των ρίψεων ενός δίκαιου νομίσματος: → *rbinom (1000,1,0.5)*

```
• rbinom(20, 10, 0.5) # sample of size 20 with mean 10*0.5=5
  [1] 5 7 4 6 6 4 3 4 5 4 6 7 2 2 4 5 4 7 4 3
• rbinom(11, 10, 0:10/10)  # different values of prob
  [1]  0  1  4  4  4  6  5  8  9  9 10
• rbinom(10, 1:10, .5)       # different values of size
  [1] 1 0 3 3 2 2 1 3 3 5
```

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

Διωνυμική κατανομή – *παράδειγμα από το GPS*

$$L_{m}^{i}{}_{k} = -\lambda_m \phi_{m}^{i}{}_{k} = \varrho_k^i - \frac{\kappa}{f_m^2} + \lambda_m b_{m}^{i}{}_{k}$$

$$P_{m}^{i}{}_{k} = \varrho_k^i + \frac{\kappa}{f_m^2},$$

Στο GPS, και τα άλλα συστήματα GNSS, η χρήση των μετρήσεων φάσης της φέρουσας συχνότητας των ραδιοσημάτων από τους δορυφόρους, απαιτούν τον υπολογισμό ενός αγνώστου ακεραίου αριθμού κύκλων φάσης για κάθε ζεύγος δορυφόρων και δεκτών που χρησιμοποιούνται

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

Διωνυμική κατανομή – *παράδειγμα από το GPS*

Στατιστική περιγραφή του προβλήματος: Έστω ότι μια τυχαία μεταβλητή *X* εκφράζει τη δυνατότητα οι ασάφειες των μετρήσεων φάσης να επιλυθούν σωστά κατά την συνόρθωση μιας σειράς μετρήσεων GPS (***επιτυχία*** *= εάν αυτές προσδιορίζονται ως ακέραιοι αριθμοί,* ***αποτυχία*** *= εάν αυτές προσδιορίζονται μόνο ως πραγματικοί αριθμοί*) → ***X~ B(n, p)***

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

Διωνυμική κατανομή – *παράδειγμα από το GPS*

$$L_{m}^{i}{}_{k} = -\lambda_m \phi_{m}^{i}{}_{k} = \varrho_k^i - \frac{\kappa}{f_m^2} + \lambda_m b_{m}^{i}{}_{k}$$

$$P_{m}^{i}{}_{k} = \varrho_k^i + \frac{\kappa}{f_m^2},$$

Εάν έχουμε να υπολογίσουμε n = 3 ασάφειες φάσης, και η πρότερη εμπειρία έχει δείξει ότι η σωστή επίλυση των ασαφειών φάσης με κάποιο συγκεκριμένο λογισμικό γίνεται τυπικά με πιθανότητα 89% → ποια είναι η πιθανότητα να επιλυθούν σωστά 0, 3 ή ≥2 ασάφειες φάσεις ? → ***ως άσκηση …***

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

Διωνυμική κατανομή – *παράδειγμα από το GPS (ασάφειες φάσης)*

• Δεδομένα του προβλήματος: n=3, p=0.89

Επαληθεύστε ως άσκηση τους υπολογισμούς

• Μέση τιμή της δ. κατανομής, μ = 2.67

• Διακύμανση της δ. κατανομής, $\sigma^2 = 0.2937$ → τυπική απόκλιση σ = 0.542

Οι ζητούμενες πιθανότητες είναι:

• P(X=3) = … = 0.705

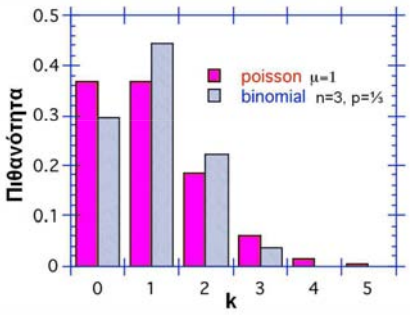
• P(X=0) = … = 0.001

  • ενδιάμεσος υπολογισμός P(X=2) = 0.261

• P(X≥2) = P(X=2) + P(X=3) = … = 0.966

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

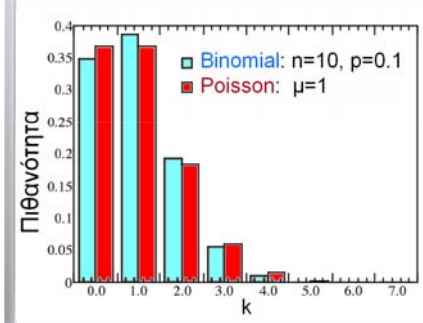
Η διωνυμική κατανομή συγκλίνει στην *κατανομή Poisson*



– Για μικρές τιμές του n και p, η διωνυμική και η *κατανομή Poisson (με την ίδια μέση τιμή μ)* διαφέρουν αρκετά μεταξύ τους

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

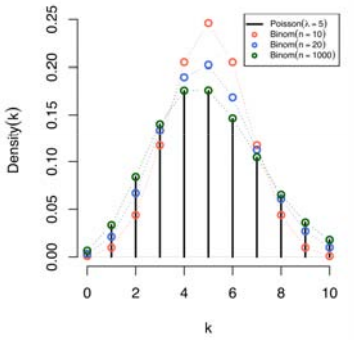
Η διωνυμική κατανομή συγκλίνει στην *κατανομή Poisson*



– Για αυξανόμενες τιμές του n και p→0, η διωνυμική και η *κατανομή Poisson (με την ίδια μέση τιμή μ)* δεν έχουν μεγάλη διαφορά

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

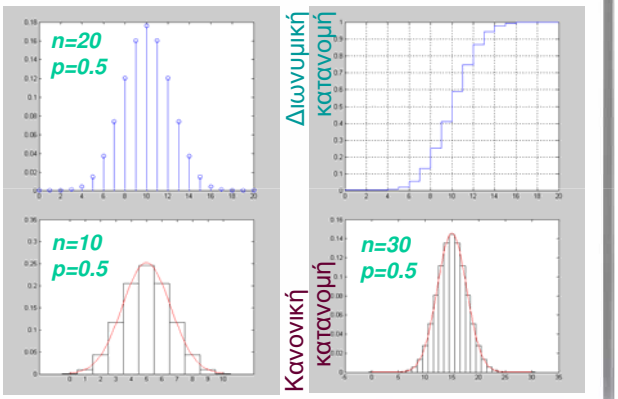
Η διωνυμική κατανομή συγκλίνει στην *κατανομή Poisson*



– Για n→∞ και p→0 έτσι ώστε *np* σταθερό, η διωνυμική κατανομή συγκλίνει στην *κατανομή Poisson* με παράμετρο *np = λ*

Δ. ΔΕΛΗΚΑΡΑΟΓΛΟΥ, ΣΑΤΜ/ΕΜΠ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ "ΓΕΩΠΛΗΡΟΦΟΡΙΚΗ"

Διωνυμική vs. Κανονική κατανομή



Κατανομή Poisson

- *Η διωνυμική κατανομή είναι μια πραγματικά θεμελιώδης κατανομή, αλλά δύσκολη στη χρήση της*
  - *Υπολογισμός παραγοντικών όρων*
  - *Αισθητή η ανάγκη για μια απλούστερη κατανομή → ως ένα (κατά προσέγγιση) υποκατάστατο της διωνυμικής κατανομής για μεγάλες τιμές του αριθμού των επαναλήψεων n*

Κατανομή Poisson

- Ανήκει στην κατηγορία των διακριτών κατανομών πιθανότητας
- *Προτάθηκε,* το 1837, *από τον Γάλλο μαθηματικό* Siméon Denis Poisson
  - στο έργο "*Recherches sur la probabilité des jugements en matière criminelle et en matière civile*" ("Έρευνα σχετικά με την πιθανότητα αποφάσεων σε ποινικές και αστικές υποθέσεις')

Κατανομή Poisson

- Εκφράζει την πιθανότητα ενός δεδομένου αριθμού γεγονότων *λ (>0)* που συμβαίνουν σε ένα σταθερό διάστημα χρόνου ή/και χώρου *(διαστήματα ή περιοχές ευκαιρίας)*, αν αυτά τα γεγονότα συμβαίνουν με ένα γνωστό μέσο ρυθμό και είναι ανεξάρτητα από το τελευταίο γεγονός
- Μπορεί επίσης να χρησιμοποιηθεί για τον αριθμό γεγονότων σε άλλα καθορισμένα διαστήματα όπως η απόσταση, η επιφάνεια, ο όγκος κ.ο.κ.

Πρακτικά παραδείγματα καταστάσεων ή συμβάντων που ακολουθούν την κατανομή Poisson

- X: ο αριθμός των τυπογραφικών λαθών στην εκτυπωμένη σελίδα ενός τεύχους πανεπιστημιακών σημειώσεων
- X: ο αριθμός αυτοκινήτων που διέρχονται την πύλη Κατεχάκη/ΕΜΠ σε διάστημα 5 λεπτών
- X: ο αριθμός ψαριών που αλιεύονται στα δίχτυα ενός ψαρά
- X: ο αριθμός ατόμων που εξυπηρετούνται σε ένα ΑΤΜ σε διαστήματα 10 λεπτών
- X: ο αριθμός των φοιτητών που παρακολουθούν μια διάλεξη
- X: ο αριθμός των ελαττωματικών ηλεκτρικών λαμπτήρων που παράγονται από μια εταιρεία
- X: ο αριθμός των βιομηχανικών ατυχημάτων ανά μήνα σε ένα εργοστάσιο

Κατανομή Poisson

- Εάν η τυχαία μεταβλητή *X* λαμβάνει τιμές x=*k=0,1,2,…* → η πιθανότητα *P(X=k)* μπορεί να προσεγγισθεί ικανοποιητικά από την αποκαλούμενη **κατανομή Poisson** *(Poisson distribution).*

$$f(k;\lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$k = 0, 1, 2, \ldots, \quad \lambda \in R_+$$

$$\mu = \sigma^2 = \lambda$$

Χαρακτηριστικές παράμετροι της κατανομής Poisson

- Η αναμενόμενη τιμή και η διακύμανση μιας τυχαίας μεταβλητής X που ακολουθεί την κατανομή Poisson είναι και οι δύο ίσες με τη παράμετρο λ
  - $\lambda = \mu = E[X] = \sigma^2 = E[(X-\mu)^2]$
- Συντελεστής ασυμμετρίας, $\gamma = \lambda^{-1/2}$
- Συντελεστής κύρτωσης, $\beta = \lambda^{-1}$
- Διάμεσος, $M \approx \lambda + \frac{1}{3} - 0.2/\lambda$ με όρια $\lambda - \ln 2 \le M < \lambda + \frac{1}{3}$

- Η κατανομή Poisson $\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

$$k = 0, 1, 2, \ldots, \quad \lambda \in R_+$$

αποτελεί όριο της διωνυμικής κατανομής $\mu = \sigma^2 = \lambda$

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- υπό τις προϋποθέσεις: $n \to \infty$ και $p \to 0$

Κατανομή Poisson ως προσέγγιση της διωνυμικής κατανομής

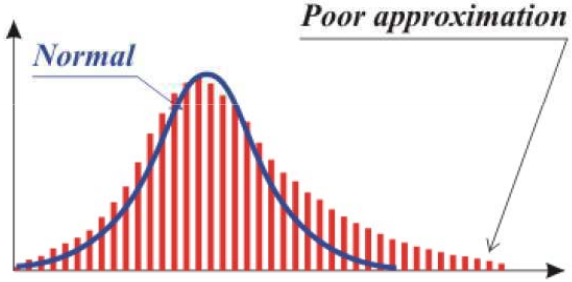
Πρακτικά, η κατανομή Poisson είναι

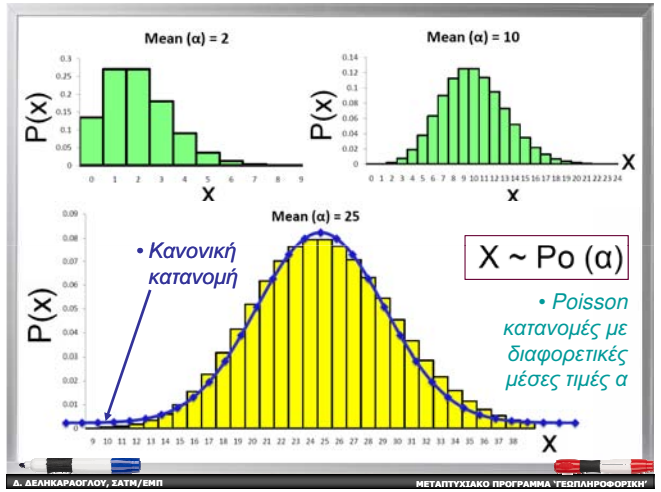
- μια καλή προσέγγιση της διωνυμικής κατανομής αν το n είναι τουλάχιστον 20 και το *p* είναι μικρότερο ή ίσο του 0.05, και
- μια άριστη προσέγγιση, αν $n \ge 100$ και $np \le 10$
  - $F_{Binomial}(k, n, p) \approx F_{Poisson}(k, \lambda = np)$

Κατανομή Poisson ως προσέγγιση της κανονικής κατανομής

- Για μεγάλες τιμές του λ, (π.χ., λ>1000), η κανονική κατανομή με μέση τιμή λ και διασπορά λ (τυπική απόκλιση $\sqrt{\lambda}$) είναι μία άριστη προσέγγιση της κατανομής Poisson
  - $F_{Poisson}(k, \lambda=np) \approx F_{Normal}(x, \mu=\lambda, \sigma^2=\lambda)$
- *Από τον 18ο αι. ο de Moivre έδειξε ότι με τη χρήση κατάλληλης κλίμακας, η κανονική κατανομή θα μπορούσε να παίξει αυτό το ρόλο*

Κατανομή Poisson ως προσέγγιση της κανονικής κατανομής

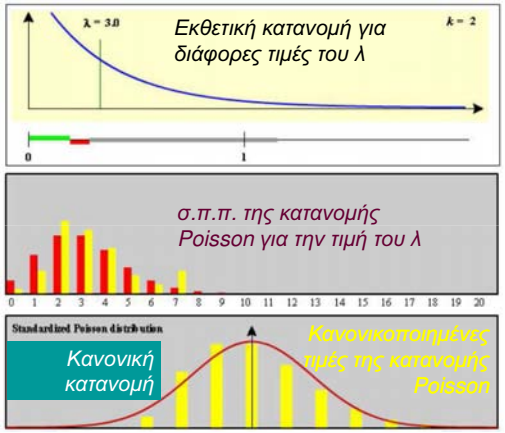




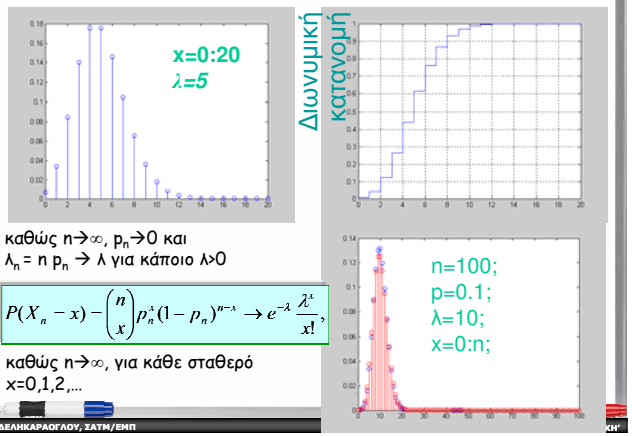
Κατανομή Poisson ως προσέγγιση της κανονικής κατανομής

Ακολουθεί γραφικό παράδειγμα

1. *Δείγμα από παρατηρήσεις που ακολουθούν μια εκθετική κατανομή του λ*
2. *Οι τιμές τους προστίθενται μέχρι που το άθροισμα τους υπερβαίνει το 1*
3. *Εάν το άθροισμα από τις πρώτες k παρατηρήσεις είναι <1, η k+1 παρατήρηση καθιστά το άθροισμα >1 → η τιμή k θεωρείται ότι προέρχεται από μια κατανομή Poisson*



Poisson vs. Διωνυμική κατανομή



Μένει να δούμε ...

Πως εφαρμόζουμε τις διάφορες κατανομές για τη διεξαγωγή στατιστικών ελέγχων;