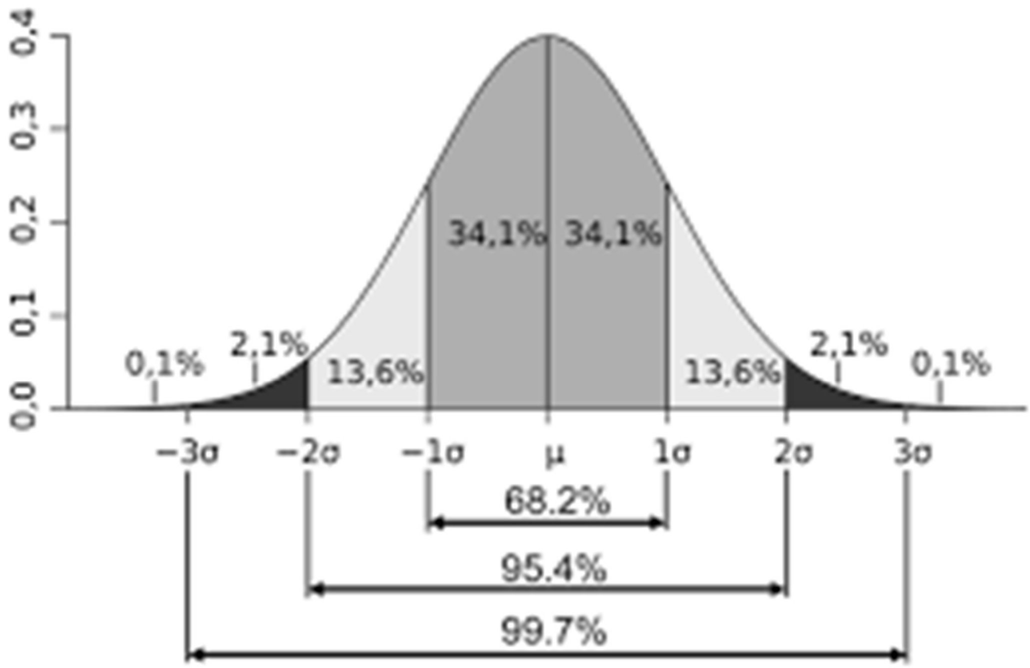
Επικ. καθηγητής Κων/νος Λιαγκούρας
Τμήμα Πληροφορικής, Παν/μιου Πειραιώς

Κανονική κατανομή

Η κανονική κατανομή είναι η πιο ευρέως γνωστή και χρησιμοποιούμενη από όλες τις κατανομές. Επειδή η κανονική κατανομή προσεγγίζει πολλά φυσικά φαινόμενα τόσο καλά, έχει εξελιχθεί σε πρότυπο αναφοράς για πολλά προβλήματα πιθανοτήτων.



I. Χαρακτηριστικά της Κανονικής κατανομής

• Συμμετρικό, σε σχήμα καμπάνας

• Συνεχές για όλες τις τιμές του Χ μεταξύ -∞ και ∞ έτσι ώστε κάθε νοητό διάστημα πραγματικών αριθμών να έχει πιθανότητα διαφορετική από το μηδέν $-\infty \le X \le \infty$

• Δύο παράμετροι, μ και σ. Σημειώστε ότι η κανονική κατανομή είναι στην πραγματικότητα μια οικογένεια κατανομών, αφού τα μ και σ καθορίζουν το σχήμα της κατανομής.

• Ο κανόνας για μια κανονική συνάρτηση πυκνότητας είναι

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

• Ο συμβολισμός N(μ, σ^2) σημαίνει κανονικά κατανεμημένος με μέση μ και διακύμανση σ^2. Αν πούμε X ~ N(μ, σ^2) εννοούμε ότι το X κατανέμεται N(μ, σ^2).

• Περίπου τα 2/3 όλων των περιπτώσεων εμπίπτουν σε μία τυπική απόκλιση του μέσου όρου, δηλαδή $P(\mu - \sigma \le X \le \mu + \sigma) = 0{,}6826$

• Περίπου το 95% των περιπτώσεων βρίσκονται εντός 2 τυπικών αποκλίσεων από τον μέσο όρο, δηλαδή $P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0{,}9544$

I. Γιατί είναι χρήσιμη η κανονική κατανομή;

• Πολλά μεταβλητές κατανέμονται κανονικά ή πολύ κοντά σε αυτό. Για παράδειγμα, το ύψος ενός πληθυσμού κατανέμεται περίπου κανονικά. Τα σφάλματα μέτρησης έχουν επίσης συχνά κανονική κατανομή

• Η κανονική κατανομή είναι εύκολο να εργαστεί μαθηματικά. Σε πολλές πρακτικές περιπτώσεις, οι μέθοδοι που αναπτύχθηκαν χρησιμοποιώντας την κανονική θεωρία λειτουργούν αρκετά καλά ακόμα και όταν η κατανομή δεν είναι κανονική.

• Υπάρχει μια πολύ ισχυρή σύνδεση μεταξύ του μεγέθους ενός δείγματος N και του βαθμού στον οποίο μια κατανομή δειγματοληψίας προσεγγίζει την κανονική μορφή. Πολλές κατανομές δειγματοληψίας που βασίζονται σε μεγάλο N μπορούν να προσεγγιστούν από την κανονική κατανομή, παρόλο που η ίδια η κατανομή πληθυσμού δεν είναι σίγουρα κανονική.

I. Η τυποποιημένη κανονική κατανομή.

1. Γενική Διαδικασία. Όπως ίσως υποψιάζεστε από τη φόρμουλα για της κανονικής συνάρτησης πυκνότητας, θα ήταν δύσκολο και κουραστικό να κάνουμε τον λογισμό κάθε φορά που είχαμε ένα νέο σύνολο παραμέτρων για τα μ και σ. Αντίθετα, συνήθως εργαζόμαστε με την τυποποιημένη κανονική κατανομή, όπου μ = 0 και σ = 1, δηλαδή N(0,1). Δηλαδή, αντί να λυθεί άμεσα ένα πρόβλημα που περιλαμβάνει μια κανονικά κατανεμημένη μεταβλητή X με μέση μ και τυπική απόκλιση σ, χρησιμοποιείται μια έμμεση προσέγγιση.
2. Πρώτα μετατρέπουμε το πρόβλημα σε ισοδύναμο που μετράται σε τυποποιημένες μονάδες απόκλισης, που ονομάζεται τυποποιημένη κανονική μεταβλητή. Για να γίνει αυτό, αν X ~ N(μ, σ^2), τότε

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1)$$

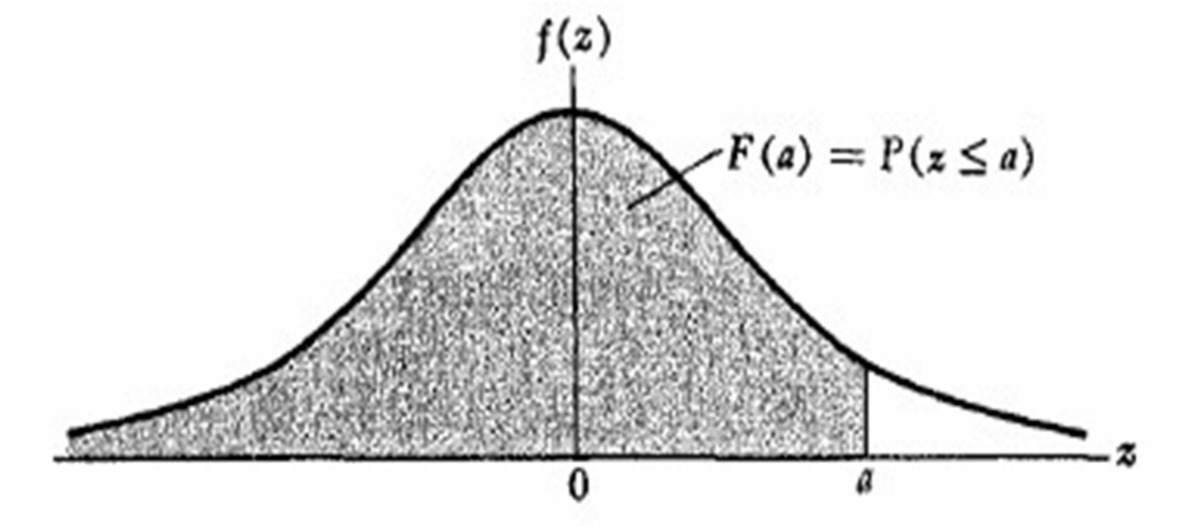
Στη συνέχεια χρησιμοποιείται ένας πίνακας τυποποιημένων κανονικών τιμών (βλέπε παράρτημα πινάκων) για την απόκτηση της απάντησης όσον αφορά το πρόβλημα μετατροπής.

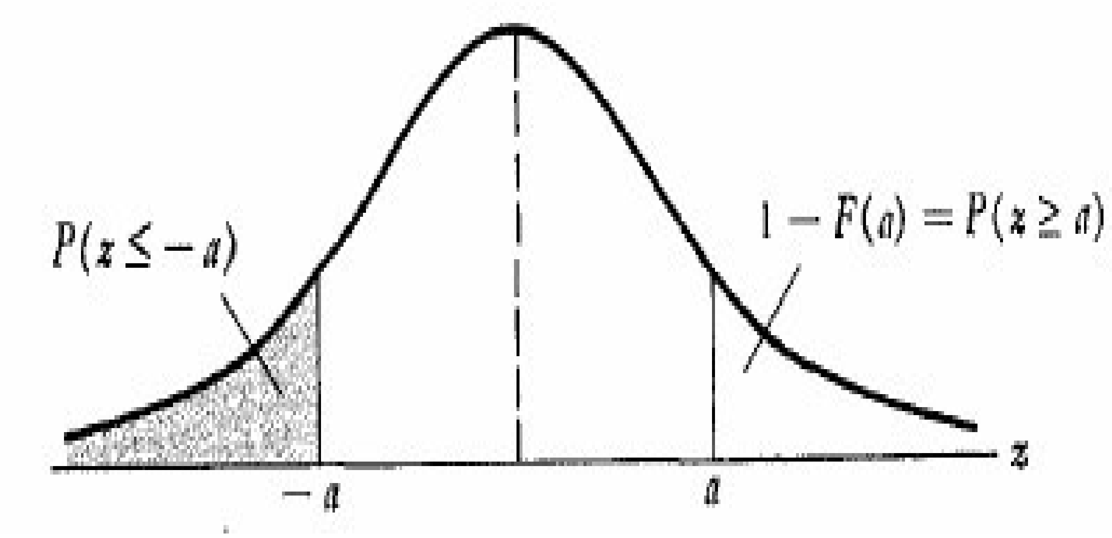
2. Εάν είναι απαραίτητο, μπορούμε στη συνέχεια να μετατρέψουμε ξανά στις αρχικές μονάδες μέτρησης. Για να το κάνετε αυτό, απλά σημειώστε ότι, αν πάρουμε τον τύπο για το Z, πολλαπλασιάσουμε και τις δύο πλευρές με σ και στη συνέχεια προσθέσουμε μ και στις δύο πλευρές, παίρνουμε $\mathbf{X = Z\sigma + \mu}$

3. Κανόνες για τη χρήση της τυποποιημένης κανονικής κατανομής.Είναι πολύ σημαντικό να κατανοήστε πώς λειτουργεί η τυποποιημένη κανονική διανομή. Θυμηθείτε ότι, για μια τυχαία μεταβλητή X, ισχύει:

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

$$P(Z \leq a) = F(a) = 1 - F(-a)$$





Παραδείγματα

$P(Z \leq 1.65) = F(1.65) = 0.95$

$P(Z \leq -1.65) = F(-1.65) = 1 - F(1.65) = 0.05$

$P(Z \leq 1.0) = F(1.0) = 0.84$

$P(Z \leq -1.0) = F(-1.0) = 1 - F(1.0) = 0.16$

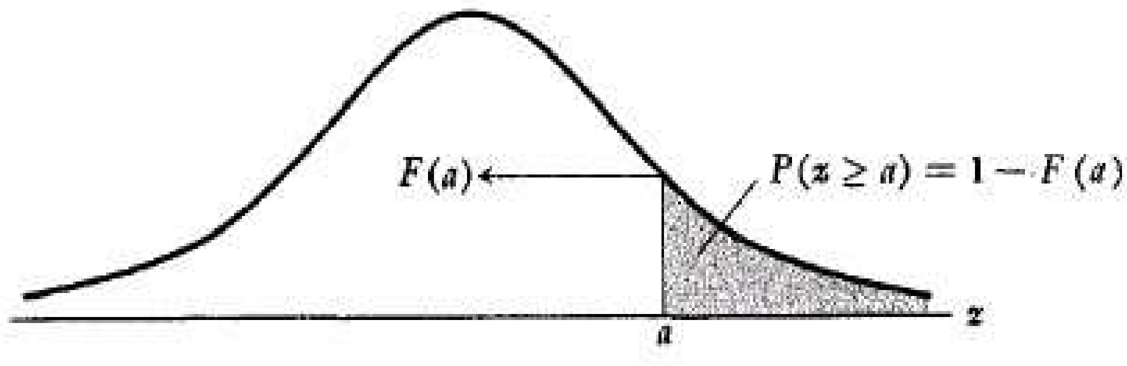
Είναι χρήσιμο να έχετε κατά νου ότι $F(a) + F(-a) = 1$.

$P(Z \leq 0.26) = 0.6026$

$P(Z \leq 1.96) = 0.9750$

$P(Z \leq -0.40) = F(-0.40) = 1 - F(0.40) = 1 - 0.6554 = 0.3446$

$P(Z \geq a) = 1 - P(Z \leq a) = 1 - F(a) = F(-a)$

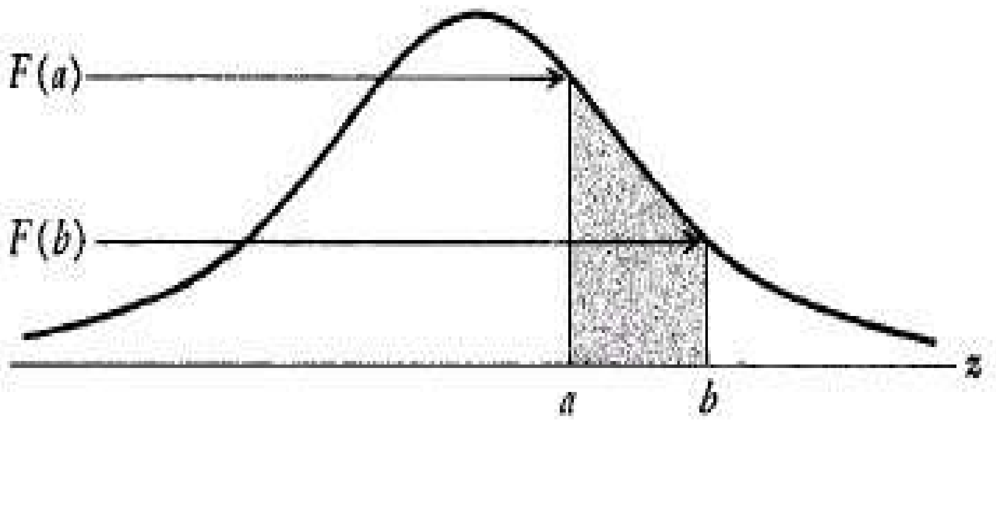


Παραδείγματα

$P(Z \geq 1.5) = 1 - F(1.5) = 1 - .9332 = .0668$

$P(Z \geq -1.5) = F(1.5) = .9332$

$P(a \leq Z \leq b) = P(Z \leq b) - P(Z \leq a) = = F(b) - F(a)$



Παραδείγματα

$P(-1 \leq Z \leq 1.5) = F(1.5) - F(-1) = F(1.5) - (1 - F(1)) = .9332 - 1 + .8413 = .7745$

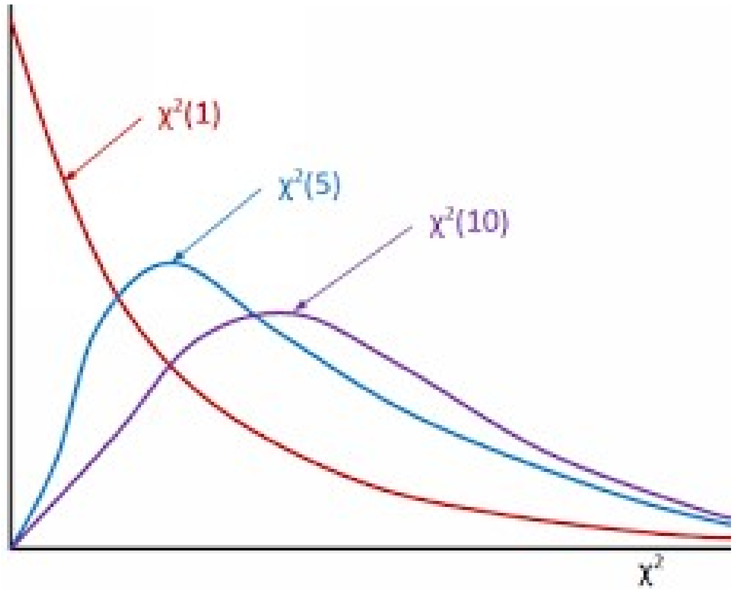
Chi-square (X^2) distribution, t-distribution, και Fisher distribution

Η κατανομή Chi-square, η κατανομή t και η κατανομή Fisher είναι όλες σημαντικές κατανομές πιθανοτήτων που χρησιμοποιούνται στη στατιστική, καθεμία με συγκεκριμένες εφαρμογές ανάλογα με το πλαίσιο των δεδομένων και της υπόθεσης που ελέγχονται. Ακολουθεί μια ανάλυση για το καθένα και πότε θα τα χρησιμοποιήσετε:

Κατανομή Chi-Square (X^2 Κατανομή)

• Ορισμός: Η κατανομή χ-τετράγωνο είναι η κατανομή ενός αθροίσματος των τετραγώνων ανεξάρτητων τυπικών κανονικών τυχαίων μεταβλητών. Είναι μια ειδική περίπτωση της κατανομής γάμμα και είναι πάντα θετική (δεν μπορεί να πάρει αρνητικές τιμές).

• Βαθμοί ελευθερίας: Το σχήμα της κατανομής χι-τετράγωνο εξαρτάται από τον αριθμό των βαθμών ελευθερίας (df), ο οποίος τυπικά αντιστοιχεί στον αριθμό των ανεξάρτητων τυπικών κανονικών μεταβλητών στο τετράγωνο.



• Πότε να χρησιμοποιείται:

ο Δοκιμές καλής προσαρμογής: Χρησιμοποιούνται σε δοκιμές υποθέσεων για να προσδιοριστεί εάν μια παρατηρούμενη κατανομή συχνότητας ταιριάζει με την αναμενόμενη κατανομή

ο Test for Variance: Κατά τη σύγκριση της διακύμανσης του δείγματος με τη διακύμανση πληθυσμού σε περιπτώσεις όπου τα δεδομένα διανέμονται κανονικά.

ο Διαστήματα εμπιστοσύνης για τη διακύμανση: Εκτίμηση της διακύμανσης του πληθυσμού χρησιμοποιώντας τη διακύμανση του δείγματος σε κανονικά κατανεμημένα δεδομένα.

• Παράδειγμα: Ένα τεστ Χ-τετράγωνο θα μπορούσε να χρησιμοποιηθεί για να προσδιοριστεί εάν η παρατηρούμενη κατανομή των χρωμάτων σε ένα σακουλάκι με καραμέλες διαφέρει από την αναμενόμενη κατανομή.

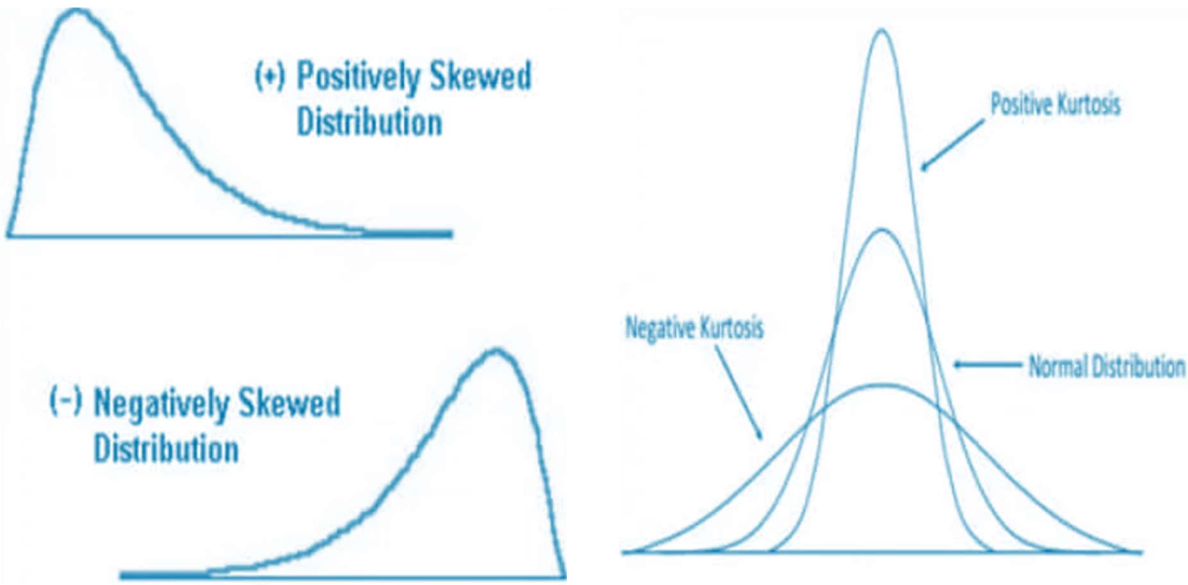
Η κατανομή **Χ-τετράγωνο (χ^2)** είναι μια ειδική περίπτωση της κατανομής γάμμα και χρησιμοποιείται ευρέως στη στατιστική ανάλυση, ιδιαίτερα στον έλεγχο υποθέσεων. Τα βασικά χαρακτηριστικά είναι:

Σχήμα

• Η κατανομή του χι-τετράγωνου είναι λοξή προς τα δεξιά και γίνεται πιο συμμετρική όσο αυξάνονται οι βαθμοί ελευθερίας (df).

• Για μικρές τιμές df (π.χ. 1 ή 2), η κατανομή είναι πολύ λοξή προς τα δεξιά.

• Καθώς αυξάνεται η df, η κατανομή πλησιάζει περισσότερο σε μια κανονική κατανομή. Για μεγάλους βαθμούς ελευθερίας (συνήθως df > 30), η κατανομή χ-τετράγωνο φαίνεται χονδρικά συμμετρική, όπως μια κανονική κατανομή.



Μη αρνητικές τιμές

• Η κατανομή χ-τετράγωνο παίρνει μόνο θετικές τιμές (δηλαδή, κυμαίνεται από 0 έως άπειρο). Δεν μπορεί να έχει αρνητικές τιμές.

• Αυτό είναι λογικό γιατί η στατιστική chi-square είναι το άθροισμα των τετραγώνων ανεξάρτητων τυπικών κανονικών μεταβλητών, οι οποίες είναι πάντα μη αρνητικές.

Βαθμοί Ελευθερίας (df)

• Οι βαθμοί ελευθερίας (df) για την κατανομή χι-τετράγωνο καθορίζουν το σχήμα της. Για ένα δεδομένο δείγμα, οι βαθμοί ελευθερίας συνήθως σχετίζονται με τον αριθμό των ανεξάρτητων τιμών στο δείγμα.

• Στον έλεγχο υποθέσεων, οι βαθμοί ελευθερίας συνήθως εξαρτώνται από τον αριθμό των κατηγοριών ή των παραμέτρων που εκτιμώνται.

• Καθώς αυξάνεται το df, η κατανομή γίνεται πιο διασκορπισμένη και πλησιάζει σε μια κανονική κατανομή.

Μέσος

• Ο μέσος όρος της κατανομής χ-τετράγωνο είναι ίσος με τους βαθμούς ελευθερίας της (df).

• Mean = df

• Παράδειγμα: Αν df = 5, ο μέσος της κατανομής είναι 5.

Διακύμανση

• Η διακύμανση της κατανομής χι-τετράγωνο είναι ίση με τους διπλάσιους βαθμούς ελευθερίας.

• Variance =2 * df

• Παράδειγμα: Αν df = 5, η διακύμανση είναι 10.

Στρεβλότητα

• Η λοξότητα της κατανομής χι-τετράγωνο είναι θετική, δηλαδή είναι λοξή προς τα δεξιά.

• Καθώς το df αυξάνεται, η λοξότητα μειώνεται και η κατανομή πλησιάζει την κανονικότητα.

• Ο τύπος λοξότητας είναι: **Skewness =** $\sqrt{\frac{\mathbf{8}}{\boldsymbol{df}}}$

Κύρτωση

• Η κύρτωση (ένα μέτρο της «ουράς» της κατανομής) είναι υψηλότερη από αυτή μιας κανονικής κατανομής, ειδικά για μικρά df. Η κατανομή έχει βαρύτερες ουρές.

• Ο τύπος κύρτωσης είναι:

$$\text{Kurtosis} = \frac{12}{df} + 3$$

Συνάρτηση πυκνότητας πιθανότητας - Probability Density Function (PDF)

$$f(x; df) = \frac{x^{(df/2)-1} e^{-x/2}}{2^{df/2}\Gamma(df/2)}$$

όπου:

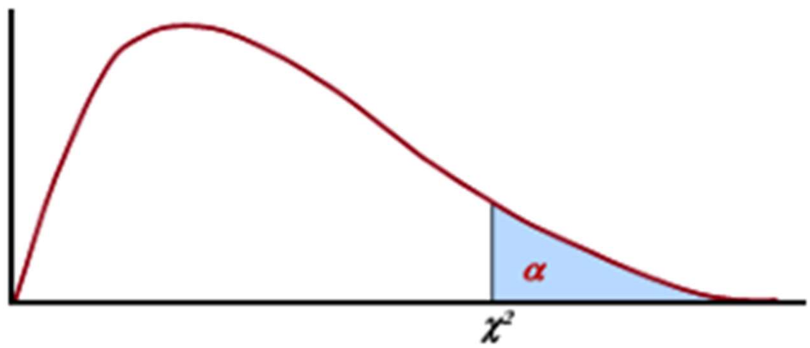
• x είναι η τιμή της μεταβλητής chi-square,

• Γ(·) είναι η συνάρτηση γάμα,

• df είναι οι βαθμοί ελευθερίας.

Ιδιότητα προσθετικότητας

• Εάν αθροίσετε δύο ανεξάρτητες τυχαίες μεταβλητές Χ-τετράγωνο με βαθμούς ελευθερίας df_1 και df_2, η μεταβλητή που προκύπτει ακολουθεί μια κατανομή χ-τετράγωνο με βαθμούς ελευθερίας $df_1 + df_2$



Έστω X_α^2 ώστε το εμβαδόν κάτω από την κατανομή chi square στο δεξιό της μέρος να είναι ίσο με a

$$P(\chi^2(\nu) \geq \chi^2_\alpha) = \alpha.$$

Η τυχαία μεταβλητή X^2 ακολουθεί μία chi square κατανομή με 19 βαθμούς ελευθερίας. Να βρείτε τα $X_{0.05}^2$ και $X_{0.01}^2$

$$\chi^2_{0.05} = 30.14, \quad \chi^2_{0.01} = 36.19.$$

Υπολογίστε την πιθανότητα

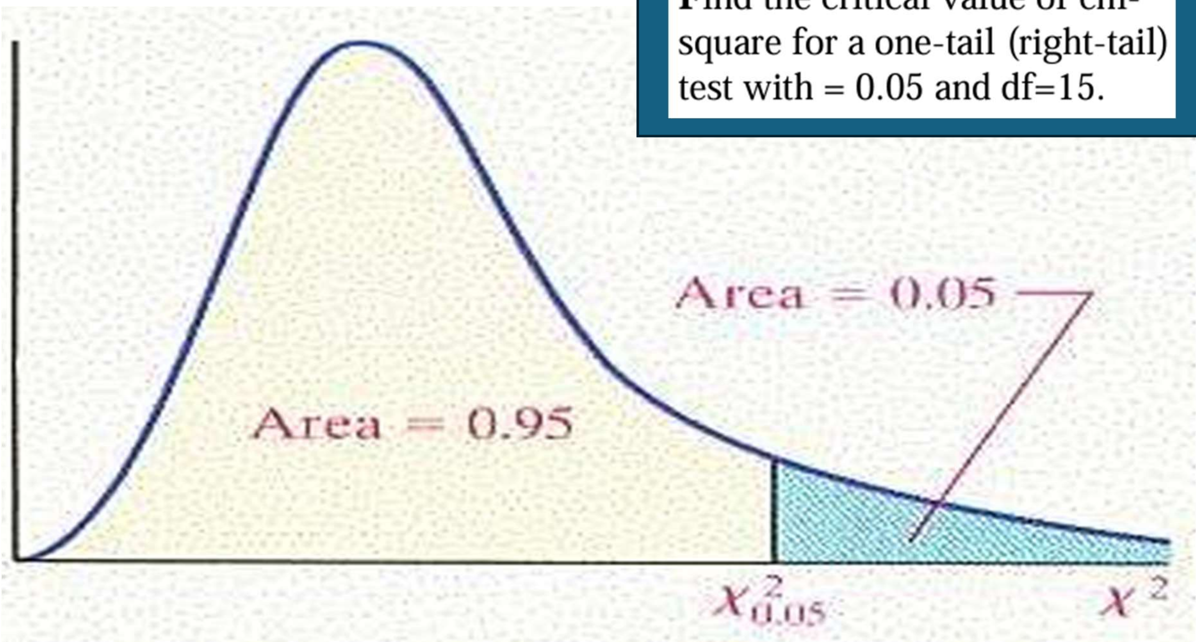
$$P(\chi^2_{1-\alpha/2} < \chi^2(\nu) < \chi^2_{\alpha/2}).$$

Λύση

$$P(\chi^2(\nu) \geq \chi^2_{\alpha/2}) = \frac{\alpha}{2}, \quad P(\chi^2(\nu) \geq \chi^2_{1-\alpha/2}) = 1 - \frac{\alpha}{2},$$

$$P(\chi^2_{1-\alpha/2} < \chi^2(\nu) < \chi^2_{\alpha/2}) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - 0.05 = 0.95.$$

Find the critical value of chi-square for a one-tail (right-tail) test with = 0.05 and df=15.



Degrees of Freedom	Area to the Right of the Critical Value									
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.299
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.365	7.215	8.231	9.288	10.265	25.200	28.601	31.595	24.265	36.456

t-Distribution (Student's t-Distribution)

• Ορισμός: Η κατανομή t είναι μια οικογένεια κατανομών που είναι συμμετρικές και σε σχήμα καμπάνας, παρόμοια με την κανονική κατανομή αλλά με βαρύτερες ουρές. Χρησιμοποιείται κυρίως όταν το μέγεθος του δείγματος είναι μικρό και η τυπική απόκλιση πληθυσμού είναι άγνωστη. Πλησιάζει την κανονική κατανομή καθώς αυξάνεται το μέγεθος του δείγματος.

• Βαθμοί ελευθερίας: Η κατανομή t εξαρτάται από το μέγεθος του δείγματος, το οποίο σχετίζεται με τους βαθμούς ελευθερίας, τυπικά df = n - 1 για ένα τεστ t ενός δείγματος, όπου n είναι το μέγεθος του δείγματος.

Πότε να χρησιμοποιείται:

- Έλεγχος υποθέσεων για μέσο όρο: Χρησιμοποιείται όταν το μέγεθος του δείγματος είναι μικρό και η τυπική απόκλιση του πληθυσμού είναι άγνωστη (π.χ., t-test για μέσους όρους).
- Διαστήματα εμπιστοσύνης για τα μέσα: Κατά την εκτίμηση του μέσου όρου του πληθυσμού από ένα μικρό δείγμα.
- Σύγκριση δύο μέσων: Όταν συγκρίνονται οι μέσοι όροι δύο ομάδων (π.χ. ανεξάρτητες ή ζευγαρωμένες t-test).
- Παράδειγμα: Ένα t-test μπορεί να χρησιμοποιηθεί για να συγκριθούν οι μέσες βαθμολογίες τεστ δύο διαφορετικών μεθόδων διδασκαλίας, με ένα μικρό μέγεθος δείγματος.
- Η κατανομή t είναι συμμετρική ως προς το 0, όπως και η κανονική κατανομή.
- Έχει βαρύτερες ουρές από την κανονική κατανομή, που σημαίνει ότι είναι πιο επιρρεπές στο να παράγει τιμές μακριά από τη μέση τιμή του. Αυτό οφείλεται στο γεγονός ότι ευθύνεται για την αυξημένη αβεβαιότητα κατά την εργασία με μικρότερα μεγέθη δειγμάτων.
- Καθώς οι βαθμοί ελευθερίας (df) αυξάνονται, η κατανομή t πλησιάζει το σχήμα της κανονικής κατανομής.

Βαθμοί Ελευθερίας (df)

Οι βαθμοί ελευθερίας (df) μιας t-κατανομής σχετίζονται με το μέγεθος του δείγματος. Για μια δοκιμή t ενός δείγματος, df = n−1, όπου n είναι το μέγεθος του δείγματος.

Καθώς οι βαθμοί ελευθερίας αυξάνονται, η κατανομή t γίνεται πιο κοντά στην κανονική κατανομή. Όταν το df είναι μεγάλο (συνήθως μεγαλύτερο από 30), η κατανομή t είναι σχεδόν ίδια με την τυπική κανονική κατανομή (κατανομή Z).

Μέσος

Ο μέσος όρος της t-κατανομής είναι 0, όπως και η κανονική κατανομή. Αυτό ισχύει για όλες τις τιμές των βαθμών ελευθερίας.

Διακύμανση

Η διακύμανση της κατανομής t είναι μεγαλύτερη από 1 και εξαρτάται από τους βαθμούς ελευθερίας.

Για df > 2, η διακύμανση δίνεται από:

$$\text{Variance} = \frac{df}{df - 2}$$

Καθώς **df → ∞**, η διακύμανση πλησιάζει το 1, που είναι η διακύμανση της τυπικής κανονικής κατανομής.

Για **df ≤ 2**, η διακύμανση είναι απροσδιόριστη.

Λοξότητα - Skewness

Η κατανομή t είναι συμμετρική, άρα έχει μηδενική λοξότητα. Όπως η κανονική κατανομή, έχει ένα συμμετρικό σχήμα καμπάνας με την ίδια πιθανότητα τιμών και στις δύο πλευρές του μέσου όρου.

Κύρτωση - Kurtosis

Η κύρτωση της κατανομής t είναι μεγαλύτερη από την κύρτωση της κανονικής κατανομής (που είναι 3). Αυτό σημαίνει ότι η κατανομή t έχει βαρύτερες ουρές και είναι πιο επιρρεπής στην παραγωγή ακραίων τιμών.

Η περίσσεια κύρτωση (η κύρτωση μείον 3) για μια κατανομή t δίνεται από:

$$\text{Excess Kurtosis} = \frac{6}{df - 4}$$

Καθώς το df αυξάνεται, η υπερβολική κύρτωση μειώνεται και η κατανομή t πλησιάζει περισσότερο σε μια κανονική κατανομή.

Συνάρτηση πυκνότητας πιθανότητας - Probability Density Function (PDF)

Η συνάρτηση πυκνότητας πιθανότητας (PDF) για την κατανομή t με dfβαθμούς ελευθερίας:

$$f(x; df) = \frac{\Gamma\left(\frac{df+1}{2}\right)}{\sqrt{df\pi}\,\Gamma\left(\frac{df}{2}\right)} \left(1 + \frac{x^2}{df}\right)^{-\frac{df+1}{2}}$$

x είναι η τιμή μίας μεταβλητής που ακολουθεί την της t-κατανομή

Γ είναι η συνάρτηση γάμα,

df είναι οι βαθμοί ελευθερίας.

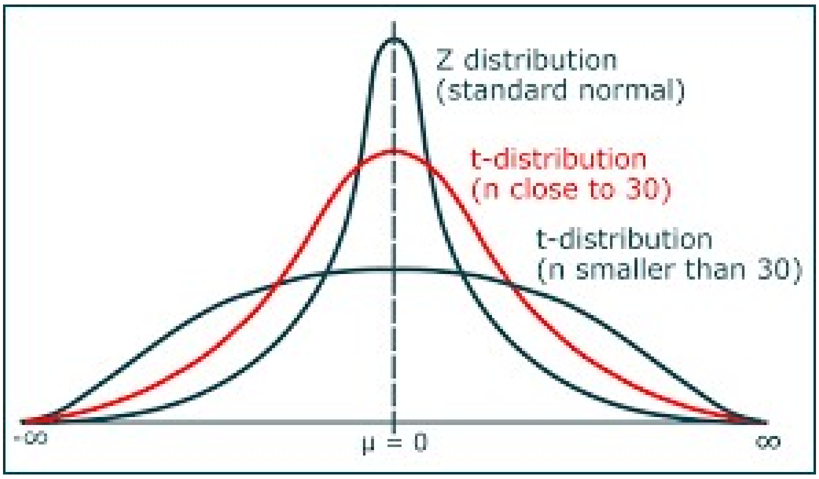
Εφαρμογές στον Έλεγχο Υποθέσεων

- Δοκιμή t ενός δείγματος: Χρησιμοποιείται για τον έλεγχο του μέσου όρου ενός μόνο δείγματος όταν η τυπική απόκλιση πληθυσμού είναι άγνωστη.
- Δοκιμή t δύο δειγμάτων: Χρησιμοποιείται για τη σύγκριση των μέσων δύο ανεξάρτητων δειγμάτων, ιδιαίτερα όταν τα μεγέθη του δείγματος είναι μικρά.
- Paired t-test: Χρησιμοποιείται για τη σύγκριση των μέσων δύο σχετικών ομάδων ή μετρήσεων.
- Διαστήματα εμπιστοσύνης για μέσους όρους: Όταν η τυπική απόκλιση πληθυσμού (όπως και η διακύμανση που είναι το τετράγωνο της std) είναι άγνωστη, η κατανομή t χρησιμοποιείται για την κατασκευή διαστημάτων εμπιστοσύνης για τη μέση τιμή.

Σχέση με την Κανονική Κατανομή

- Η κατανομή t προσεγγίζει την κανονική κατανομή καθώς αυξάνονται οι βαθμοί ελευθερίας.

- Για μεγάλους βαθμούς ελευθερίας, η κατανομή t είναι ουσιαστικά πανομοιότυπη με την τυπική κανονική κατανομή, πράγμα που σημαίνει ότι για αρκετά μεγάλα δείγματα, μπορούμε να χρησιμοποιήσουμε την κατανομή Z (κανονική κατανομή) αντί για την κατανομή t.



Παράδειγμα: Εύρεση της κρίσιμης τιμής του t στον πίνακα t

Χρησιμοποιώντας τον πίνακα t, διαπιστώνεται ότι για μια δοκιμή <u>δύο ουρών</u> με df = 29 και α = 0,05 η κρίσιμη τιμή του t είναι 2,045.

Critical Values for Student's t-Distribution.

	Upper Tail Probability: $\Pr(T > t)$									
df	0.2	0.1	0.05	0.04	0.03	0.025	0.02	0.01	0.005	0.0005
1	1.376	3.078	6.314	7.916	10.579	12.706	15.895	31.821	63.657	636.619
2	1.061	1.886	2.920	3.320	3.896	4.303	4.849	6.965	9.925	31.599
3	0.978	1.638	2.353	2.605	2.951	3.182	3.482	4.541	5.841	12.924
4	0.941	1.533	2.132	2.333	2.601	2.776	2.999	3.747	4.604	8.610
5	0.920	1.476	2.015	2.191	2.422	2.571	2.757	3.365	4.032	6.869
6	0.906	1.440	1.943	2.104	2.313	2.447	2.612	3.143	3.707	5.959
7	0.896	1.415	1.895	2.046	2.241	2.365	2.517	2.998	3.499	5.408
8	0.889	1.397	1.860	2.004	2.189	2.306	2.449	2.896	3.355	5.041
9	0.883	1.383	1.833	1.973	2.150	2.262	2.398	2.821	3.250	4.781
10	0.879	1.372	1.812	1.948	2.120	2.228	2.359	2.764	3.169	4.587
11	0.876	1.363	1.796	1.928	2.096	2.201	2.328	2.718	3.106	4.437
12	0.873	1.356	1.782	1.912	2.076	2.179	2.303	2.681	3.055	4.318
13	0.870	1.350	1.771	1.899	2.060	2.160	2.282	2.650	3.012	4.221
14	0.868	1.345	1.761	1.887	2.046	2.145	2.264	2.624	2.977	4.140
15	0.866	1.341	1.753	1.878	2.034	2.131	2.249	2.602	2.947	4.073
16	0.865	1.337	1.746	1.869	2.024	2.120	2.235	2.583	2.921	4.015
17	0.863	1.333	1.740	1.862	2.015	2.110	2.224	2.567	2.898	3.965
18	0.862	1.330	1.734	1.855	2.007	2.101	2.214	2.552	2.878	3.922
19	0.861	1.328	1.729	1.850	2.000	2.093	2.205	2.539	2.861	3.883
20	0.860	1.325	1.725	1.844	1.994	2.086	2.197	2.528	2.845	3.850
21	0.859	1.323	1.721	1.840	1.988	2.080	2.189	2.518	2.831	3.819
22	0.858	1.321	1.717	1.835	1.983	2.074	2.183	2.508	2.819	3.792
23	0.858	1.319	1.714	1.832	1.978	2.069	2.177	2.500	2.807	3.768
24	0.857	1.318	1.711	1.828	1.974	2.064	2.172	2.492	2.797	3.745
25	0.856	1.316	1.708	1.825	1.970	2.060	2.167	2.485	2.787	3.725
26	0.856	1.315	1.706	1.822	1.967	2.056	2.162	2.479	2.779	3.707
27	0.855	1.314	1.703	1.819	1.963	2.052	2.158	2.473	2.771	3.690
28	0.855	1.313	1.701	1.817	1.960	2.048	2.154	2.467	2.763	3.674
29	0.854	1.311	1.699	1.814	1.957	2.045	2.150	2.462	2.756	3.659
30	0.854	1.310	1.697	1.812	1.955	2.042	2.147	2.457	2.750	3.646
31	0.853	1.309	1.696	1.810	1.952	2.040	2.144	2.453	2.744	3.633

Παράδειγμα

Υποθέστε δείγμα n = 5 . Υπολογίστε ένα διάστημα εμπιστοσύνης 95% με δύο ουρές.

Υπολογίστε το t-score από τον πίνακα της t-distribution. Ισχύει df = n - 1

$$t_{4,0.025} = 2.776$$

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	[illegible]
2	0.000	0.816	1.061	1.386	1.886	2.920	[illegible]
3	0.000	0.765	0.978	1.250	1.638	2.353	[illegible]
4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365

Fisher (F) Distribution

• **Ορισμός:** Η κατανομή F είναι η αναλογία δύο ανεξάρτητων μεταβλητών χ-τετράγωνο, η καθεμία διαιρούμενη με τους αντίστοιχους βαθμούς ελευθερίας της. Χρησιμοποιείται κυρίως για τη σύγκριση διακυμάνσεων μεταξύ δύο ή περισσότερων ομάδων. Είναι λοξή προς τα δεξιά και χρησιμοποιείται για τον έλεγχο υποθέσεων σχετικά με τις διακυμάνσεις του πληθυσμού.

• **Βαθμοί ελευθερίας:** Η κατανομή F έχει δύο σύνολα βαθμών ελευθερίας: ένα για τον αριθμητή (df1) και ένα για τον παρονομαστή (df2). Αυτοί αντιστοιχούν στους βαθμούς ελευθερίας των κατανομών Χ-τετράγωνο που συγκρίνονται

• **Πότε να χρησιμοποιείται:**

•Ανάλυση Διακύμανσης (ANOVA): Χρησιμοποιείται για τη σύγκριση των διακυμάνσεων μεταξύ πολλαπλών ομάδων για να προσδιοριστεί εάν τουλάχιστον η μέση τιμή μιας ομάδας είναι διαφορετική από τις άλλες.

• Δοκιμή για ισότητα διακυμάνσεων: Χρησιμοποιείται κατά τη σύγκριση των διακυμάνσεων δύο πληθυσμών, όπως σε ένα τεστ για ίσες διακυμάνσεις.

• Παράδειγμα: Ένα τεστ F μπορεί να χρησιμοποιηθεί για τη σύγκριση των αποκλίσεων των βαθμολογιών των εξετάσεων μεταξύ δύο μεθόδων διδασκαλίας για να προσδιοριστεί εάν μία μέθοδος οδηγεί σε μεγαλύτερη μεταβλητότητα στα αποτελέσματα.

• Η κατανομή Fisher, γνωστή και ως κατανομή F, είναι μια συνεχής κατανομή πιθανοτήτων που προκύπτει στο πλαίσιο της σύγκρισης διακυμάνσεων και της διεξαγωγής της ανάλυσης διακύμανσης (ANOVA). Χρησιμοποιείται κυρίως στον έλεγχο υποθέσεων, ειδικά κατά τη σύγκριση δύο διακυμάνσεων δειγμάτων ή την αξιολόγηση του λόγου των διακυμάνσεων.

Σχήμα

• Η κατανομή F είναι λοξή προς τα δεξιά, που σημαίνει ότι έχει μια μακριά ουρά στη δεξιά πλευρά.

• Η κατανομή είναι ασύμμετρη, με τη λοξότητα να μειώνεται όσο αυξάνονται οι βαθμοί ελευθερίας (df).

• Καθώς οι βαθμοί ελευθερίας τόσο του αριθμητή όσο και του παρονομαστή αυξάνονται, η κατανομή γίνεται πιο συμμετρική και προσεγγίζει μια κανονική κατανομή, αν και ποτέ δεν γίνεται απόλυτα συμμετρική.

Βαθμοί Ελευθερίας (df)

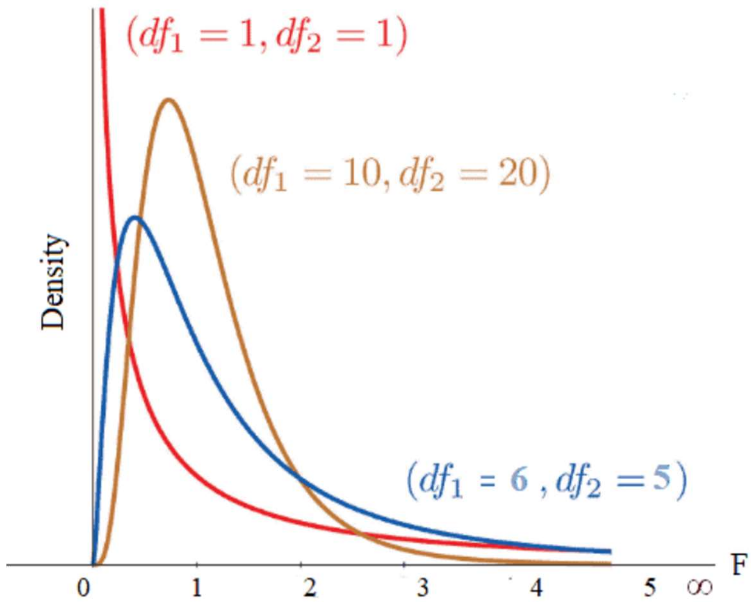
Η κατανομή F ορίζεται από δύο σύνολα βαθμών ελευθερίας:

df1: Βαθμοί ελευθερίας που σχετίζονται με τον αριθμητή

df2: Βαθμοί ελευθερίας που σχετίζονται με τον παρονομαστή

Το σχήμα της κατανομής εξαρτάται σε μεγάλο βαθμό από αυτές τις δύο παραμέτρους και οι βαθμοί ελευθερίας του αριθμητή (df1) και οι βαθμοί ελευθερίας του παρονομαστή (df2) μπορεί να διαφέρουν.

Καθώς τα df1 και df2 αυξάνονται, η κατανομή F προσεγγίζει μια κανονική κατανομή. Για μεγάλους βαθμούς ελευθερίας, η κατανομή F γίνεται παρόμοια με μια κανονική κατανομή με κέντρο το 1, αν και παραμένει λοξή προς τα δεξιά.



Μέσος

Ο μέσος όρος της κατανομής F υπάρχει μόνο εάν οι βαθμοί ελευθερίας για τον παρονομαστή (df_2) είναι μεγαλύτερος από 2.

Για **df_2 > 2**, ο μέσος όρος δίνεται από:

$$\text{Mean} = \frac{df_2}{df_2 - 2}$$

Εάν df2 ≤ 2, ο μέσος όρος είναι απροσδιόριστος (δηλαδή, η κατανομή έχει άπειρη διακύμανση).

Διακύμανση

Η διακύμανση της κατανομής F δίνεται από την σχέση:

$$\text{Variance} = \frac{2 \times (df_2)^2 \times (df_1 + df_2 - 2)}{df_1 \times (df_2 - 2)^2 \times (df_2 - 4)}$$

Αυτός ο τύπος ισχύει για **df_2 > 4.**

Όπως και ο μέσος όρος, η διακύμανση γίνεται απροσδιόριστη εάν **df2 ≤ 4.**

Λοξότητα - Skewness

Η κατανομή F είναι θετικά λοξή (δεξιά-λοξή).

Η λοξότητα μειώνεται καθώς αυξάνονται οι βαθμοί ελευθερίας, αλλά η κατανομή θα έχει πάντα μια μακριά ουρά προς τα δεξιά.

$$\text{Skewness} = \frac{2\,(df_2 - 2)}{\sqrt{2\,(df_1 + df_2 - 2)\,(df_2 - 4)}}$$

Κύρτωση

Η κύρτωση της κατανομής F είναι μεγαλύτερη από 3 (πιο «κορυφωμένη» και με βαρύτερες ουρές από την κανονική κατανομή).

Όπως η λοξότητα, η κύρτωση μειώνεται καθώς αυξάνονται οι βαθμοί ελευθερίας.

$$\text{Excess Kurtosis} = \frac{6 \times (df_2 - 4)}{df_1 \times (df_2 - 2) \times (df_2 - 4)}$$

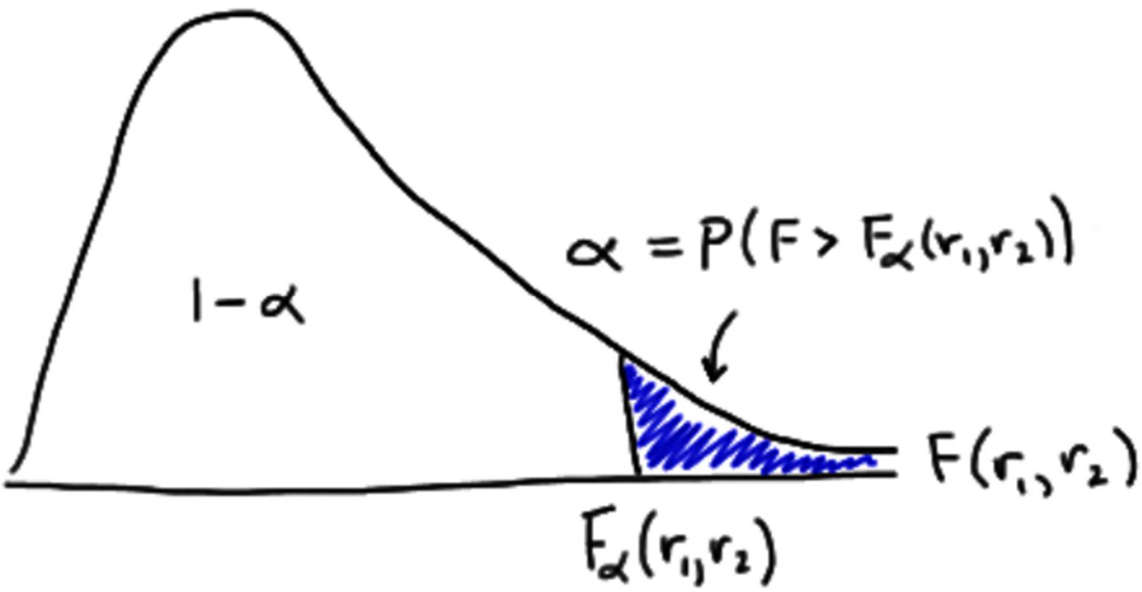
F-κατανομή

Εάν τα U και V είναι ανεξάρτητες τυχαίες μεταβλητές Χ-τετράγωνο με r1 και r2 βαθμούς ελευθερίας, αντίστοιχα, τότε:

$$F = \frac{U/r_1}{V/r_2}$$

ακολουθεί μια κατανομή F με r1 αριθμητή βαθμοί ελευθερίας και r2 παρονομαστή βαθμούς ελευθερίας. Γράφουμε

$$F \sim F(r_1, r_2)$$



- Degrees of freedom θ1 = 8 -1 = 7 (highest variance in the numerator)
- θ2 = 6 – 1 = 5
- it is a one-tail (right) test
- Level of significance α= 0.05

The F - Distribution with α = 0.05

$\nu_2 \backslash \nu_1$	2	3	4	5	6	7	8	9
2	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18