Funzione di verosimiglianza

Definizione

La funzione di verosimiglianza è una funzione di probabilità condizionata, ovvero la probabilità di osservare un certo campione di dati, dato un certo valore dei parametri del modello. La funzione di verosimiglianza è definita come:

$$L(\theta|x) = P(X = x|\theta)$$

dove:

- $L(\theta|x)$ è la funzione di verosimiglianza
- θ sono i parametri del modello
- x è il campione di dati
- $P(X = x|\theta)$ è la probabilità di osservare il campione di dati x dato il valore dei parametri θ.

Esempio

Supponiamo di avere un campione di dati $x = \{x_1, x_2, x_3, x_4\}$ e di voler calcolare la funzione di verosimiglianza per una Bernulli. La funzione di verosimiglianza sarà:

$$L(\theta|x) = P(X = x|\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \cdot ... \cdot \theta^{x_4}(1-\theta)^{1-x_4}$$

Funzione di verosimiglianza per una Bernulli

In maniera sintetica la forula della funzione di verosimiglianza per una Bernulli è:

$$L(\theta|x) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$$

Stima di massima verosimiglianza

La stima di massima verosimiglianza è un metodo per stimare i parametri di un modello statistico. La stima di massima verosimiglianza consiste nel trovare i valori dei parametri che massimizzano la funzione di verosimiglianza. Formalmente, la stima di massima verosimiglianza è definita come:

$$\hat{\theta} = argmax_\theta L(\theta|x)$$

Vogliamo massimizzare la funzione perchè vogliamo trovare i parametri che rendono più probabile l'osservazione dei dati che abbiamo a disposizione

Noi vogliamo massimizzare questa funzione perche cerchiamo quel parametro θ che rende più probabile l'osservazione dei dati che abbiamo a disposizione, ovviamente se abbiamo un numero infinito di dati riusciremo a stimarlo con una precisione maggiore. In altre parole vogliamo trovare quel parametro che rende i dati osservati più verosimili. Qindi ad esempio se abbiamo dei dati osservati che sono 1, 0, 1, 1, 0 e vogliamo trovare il parametro p che rende più probabile l'osservazione di questi dati, allora la stima di massima verosimiglianza ci dirà che il parametro p che massimizza la probabilità di osservare questi dati è 0.6.

Nell'esempio di una Bernulli il parametro θ si calcola come:

$$\hat{\theta} = \frac{\sum x_i}{n}$$

Questo coincide con la media campionaria. Il che ha senso perchè se lanciamo una moneta 10 volte e otteniamo 5 volte testa e 5 volte croce, la probabilità di ottenere testa è 0.5. Infatti se eseguiamo i calcoli con la formula sopra otteniamo:

$$\hat{\theta} = \frac{1}{n}\sum_{i=0}^{10} x_i$$

dove se la realizzazione di x_i è 1 allora otteniamo testa, altrimenti 0, croce.

Bontà di uno stimatore

Come scegliamo uno stimatore $T = T(X_1, \dots, X_n)$ per θ? Come ne valutiamo la bontà? P
Vogliamo minimizzare la deviazione dal vaalore reale del parametro e per farlo ci basiamo sui valori di $\mathbb{E}(T)$ **e** $Var(T)$, quindi il valore atteso dello stimatore e la sua varianza.

BIAS

Il bias è la differenza tra il valore atteso dello stimatore e il valore reale del parametro. Formalmente il bias è definito come:

$$b(T) = \mathbb{E}(T) - \theta$$

Se il bias è nullo allora lo stimatore è non distorto, altrimenti è distorto.

Esempio di bias

Sappiamo che la media campionaria $\bar{X}$ è uno stimatore non distorto per la media μ della popolazione.
Infatti:

$$b(\bar{X}) = \mathbb{E}(\bar{X}) - \mu = \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}(X_i)) - \mu = \frac{1}{n}n\mu - \mu = 0$$

Errore quadratico medio

L'errore quadratico medio è una misura della bontà di uno stimatore. L'errore quadratico medio è definito come:

$$MSE(T) = \mathbb{E}((T-\theta)^2) = Var(T) + b(T)^2$$

Dove:

- $MSE(T)$ è l'errore quadratico medio
- $Var(T)$ è la varianza dello stimatore
- $b(T)$ è il bias dello stimatore

L'errore quadratico medio è la somma della varianza dello stimatore e del suo bias al quadrato.

Stima intervallare

Scopo della stima intervallare

La funzione di likelihood è una funzione che ci restituisce un valore puntuale del parametro θ ma non dobbiamo aspettarci che sia il valore effettivao. Per questo motivo calcoliamo l'intervallo di confidenza. **L'intervallo di confidenza ci dice che abbiamo una certa fiducia che il parametro θ si trovi al suo interno.**

Definizione

L'intervallo di confidenza è un intervallo che contiene il valore del parametro θ con una certa probabilità. Formalmente, l'intervallo di confidenza è definito come:

$$IC(\theta) = [L, U]$$

dove:

- $IC(\theta)$ è l'intervallo di confidenza
- L è il limite inferiore dell'intervallo di confidenza
- U è il limite superiore dell'intervallo di confidenza

Sia $X_1, .., X_n$ un campione casuale da una distribuzione di probabilità $f(x|\theta)$, dove θ è il parametro da stimare. Sia $L(X_1, .., X_n)$ e $U(X_1, .., X_n)$ due statistiche tali che:

$$\mathbb{P}(L \le \theta \le U) = 1 - \alpha$$

dove:

- α è il livello di confidenza che decidiamo noi in base alla confidenza che vogliamo avere
- L'intervallo $[L, U]$ si chiama **STIMATORE INTERVALLARE** del parametro θ

Quando effettuiamo le realizzazioni delle statistiche L e U, che indichiamo come $\hat{l}$ e $\hat{u}$, otteniamo l'intervallo di confidenza $[\hat{l}, \hat{u}]$ di livello $1 - \alpha$ del parametro θ.

Notazioni

La **Stima Intervallare** è una variabile aleatoria perchè dipede dalle variabili aleatorie L, U.
Quando effettuiamo le realizzazioni delle statistiche L e U, che indichiamo come $\hat{l}$ e $\hat{u}$, otteniamo l'**Intervallo di Confidenza** $(\hat{l}, \hat{u}) \in \mathbb{R}$.

Nota

Per riuscire a costruire l'intervallo di confidenza dobbiamo conoscere la distribuzione di probabilità dei dati, come d'altronde anche per la stima di massima verosimiglianza.

Distribuzioni delle statistiche campionarie

Le statistiche campionarie sono quei valori che ricaviamo dal campione di dati.
Le distribuzioni delle statistiche campionarie sono le distribuzioni di probabilità delle statistiche calcolate su un campione casuale. Le distribuzioni delle statistiche campionarie sono utili per calcolare gli intervalli di confidenza.

Sia $X_1, .., X_n$ un campione estratto da una popolazione normale con media $\mu \in \mathbb{R}$ e varianza $\sigma^2 > 0$.
Siamo interessati alla distribuzione delle statistiche campionarie $\bar{X}$ e S^2.
Quando si dice "siamo interessati alla distribuzione delle statistiche campionarie $\bar{X}$ e S^2", significa che si vuole studiare come queste statistiche variano se si prendono diversi campioni dalla stessa popolazione. Questo può aiutare a capire quanto si può fidare delle stime basate su un singolo campione.
Questo perchè, come sappiamo, $\mathbb{E}[\bar{X}] = \mu$ e $\mathbb{E}[S^2] = \sigma^2$.

Distribuzione della media campionaria

La media campionaria è una variabile aleatoria che è distribuita secondo una distribuzione normale con media μ e varianza $\frac{\sigma^2}{n}$.
La sua distribuzione è:

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Distribuzione della varianza campionaria

Inoltre sappiamo che ha distribuzione chi-quadro con $n - 1$ gradi di libertà:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

dove χ^2_{n-1} è la distribuzione chi-quadro che ha una pdf tutta positiva.

Intervallo di confidenza per la media di una popolazione normale, con varianza nota

Vogliamo ricavare gli intervalli di confidenza, ad un livello $1 - \alpha$, per la media μ. Quindi avere:

$$\mathbb{P}(L_1 < \mu < L_2) = 1 - \alpha$$

Sapendo la distribuzione della media campionaria possiamo normalizzarla per poi trovare gli estremi dell'intervallo di confidenza.
Quindi una distribuzione normalizzata è calcolata come:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Essendo che vogliamo che la probabilità sia $1-\alpha$ possiamo dire che:

$$\mathbb{P}(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}) = 1-\alpha$$

Qindi risolvendo per μ otteniamo la formula per calcolare gli intervalli di confidenza per la media con varianza nota:

$$\mathbb{P}(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Per calcolare gli intervalli **UNILATERALI** si calcola:

$$(\bar{X} - z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty) \quad \text{Unilaterle destro}$$

$$(-\infty, \bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}) \quad \text{Unilaterle sinistro}$$

NB: occhio al segno di z_α nel calcolo unilaterale.

Dove $z_{\frac{\alpha}{2}}$ è il valore tale che calcolando la probabilità di una normale con tale valore sarà $\frac{\alpha}{2}$. Ad esempio se $\alpha = 0.05$ allora $z_{\frac{\alpha}{2}} = 1.96$, infatti se calcoliamo la probabilità di una normale con valore 1.96 otteniamo 0.025 che è proprio $\frac{\alpha}{2}$.
Dobbiamo quindi calcolare l'inversa della probabilità della normal:

$$\Phi(z_{\frac{\alpha}{2}}) = 0.025 \rightarrow z_{\frac{\alpha}{2}} = \Phi^{-1}(0.025)$$

Come calcolare l'inversa della probabilità della normale in R

Per calcolare l'inversa della probabilità della normale in R si usa la funzione `qnorm` :

```
qnorm(p=0.05/2, mean = 0, sd=1) = 1.96
```

Esempio

Supponiamo di avere un campione di dati:

$$345\ 389\ 363\ 417\ 476$$

distribuiti secondo una distribuzione normale cone varianza 50^2. Si costruisca l'intervallo di confidenza al 95% per la media della popolazione.

Quindi dobbiamo effettuare:

- Calcolare la media campionaria
- Calcolare i valori di $z_{\frac{\alpha}{2}}$
- Calcolare l'intervallo di confidenza

1. Media campionaria:

$$\bar{X} = \frac{345+389+363+417+476}{5} = 398$$

2. Calcolare i valori di $z_{\frac{\alpha}{2}}$:

```
qnorm(p=0.05/2, mean = 0, sd=1) = 1.96
```

3. Calcolare l'intervallo di confidenza:

$$\mathbb{P}(398 - 1.96\frac{50}{\sqrt{5}} < \mu < 398 + 1.96\frac{50}{\sqrt{5}}) = (354.05 < \mu < 442.95)$$

Intervalli di confidenza per la media di una popolazione normale, con varianza sconosciuta

Sia $X_1, .., X_n$ un campione casuale da una popolazione normale con media $\mu \in \mathbb{R}$ e varianza $\sigma^2 > 0$, **entrambe ignote**.
Ricaviamo gli intervalli di confidenza per la media μ al livello $1 - \alpha$.
Ricordando che la media campionaria è distribuita con un t-student con $n - 1$ gradi di libertà:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

dove:

- $\bar{X}$ è la media campionaria.
- S che è la deviazione standard campionaria. calcolata come:
  - $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$
- t_{n-1} è la distribuzione t-student con $n - 1$ gradi di libertà.
  - dove n è il numero di osservazioni del campione.

Se $X \sim t_n$ allora indichiamo con $t_{\alpha,n} \in \mathbb{R}$ il valore tale che $\mathbb{P}(X > t_{\alpha,n}) = \alpha$.
Sia $\hat{X}(x_1, .., x_n) = \hat{x}$ la media campionaria e $\hat{S}(x_1, .., x_n) = \hat{s}$ la deviazione standard campionaria a livello di confidenza **BILATERALE** $1 - \alpha$ otteniamo gli intervalli:

$$\mathbb{P}(\hat{x} - t_{\frac{\alpha}{2},n-1} \frac{\hat{s}}{\sqrt{n}} < \mu < \hat{x} + t_{\frac{\alpha}{2},n-1} \frac{\hat{s}}{\sqrt{n}}) = 1 - \alpha$$

Se invece vogliamo calcolare l'intervallo di confidenza **UNILATERALE** al livello $1 - \alpha$ otteniamo:

$$(\hat{x} - t_{\alpha,n-1} \frac{\hat{s}}{\sqrt{n}}, \infty) \quad \text{Unilaterle destro}$$

$$(-\infty, \hat{x} + t_{\alpha,n-1} \frac{\hat{s}}{\sqrt{n}}) \quad \text{Unilaterle sinistro}$$

NB: occhio al segno di $t_{\alpha,n-1}$ nel calcolo unilaterale.

In R per calcolare il valore di $t_{\frac{\alpha}{2},n-1}$ si usa la funzione `qt` :

```
qt(p=0.05, df=4) = 2.776
```

Esempio calcolato

Supponiamo di avere un campione di dati:

$$11.1 \ 10.5 \ 11.4 \ 10.7 \ 11.4$$

Vogliamo calcolare l'intervallo di confidenza *unilaterale destro* all 99% per la media della popolazione.
Dobbiamo calcolare:

- La media campionaria
- La deviazione standard campionaria
- Il valore di $t_{\alpha,5-1}$
- Calcolare l'intervallo di confidenza unilaterale destro

1. Media campionaria $\bar{x} = \frac{11.1+10.5+11.4+10.7+11.4}{5} = 11.02$
2. Deviazione standard campionaria $\hat{s} = \sqrt{\frac{1}{5-1} \sum_{i=1}^{5} (x_i - \bar{x})^2} = 0.41$
3. Calcolare il valore di $t_{\alpha,5-1}$:

```
qt(p=0.01, df=4) = 3.747
```

4. Calcolare l'intervallo di confidenza unilaterale destro:

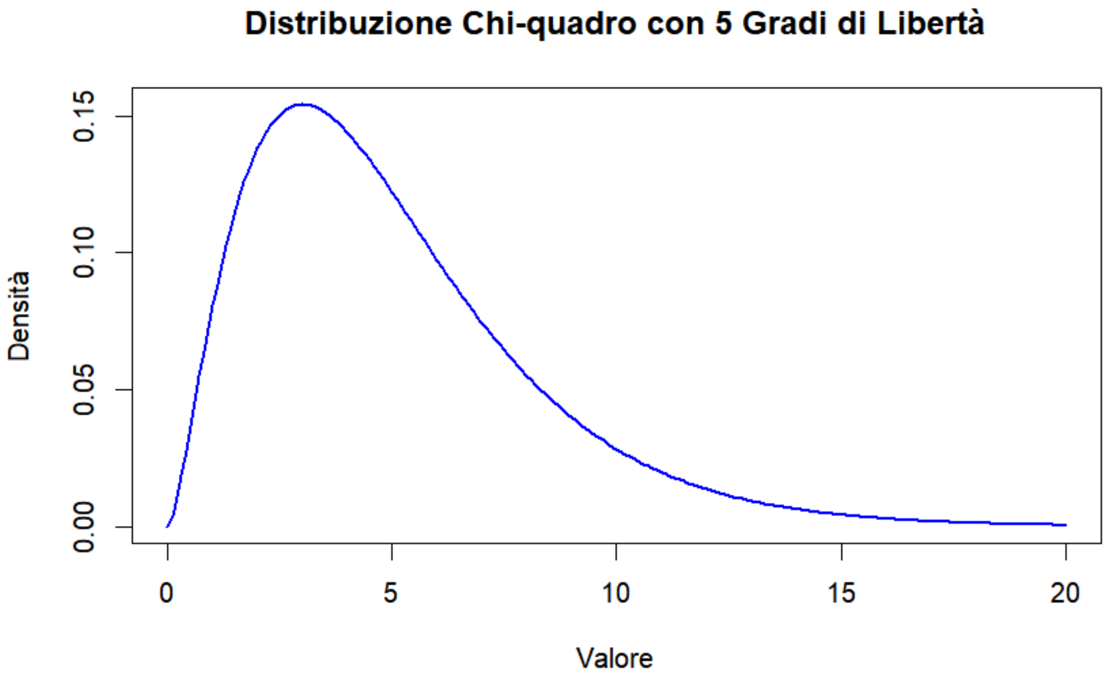
$$\mathbb{P}(11.02 - 3.747\frac{0.41}{\sqrt{5}}, \infty) = (10.3, \infty)$$

Intervalli di confidenza per la varianza di una popolazione normale

Sia $X_1, .., X_n$ un campione casuale da una popolazione normale con media $\mu \in \mathbb{R}$ e varianza $\sigma^2 > 0$, **entrambe ignote**. Possiamo costruire gli intervalli di confidenza basandoci sul fatto che:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

**Come la t-student anche il chi-quadro è in funzione del numero di campioni Inoltre distribuzione chi-quadro ha una pdf tutta positiva.*



Quindi possiamo costruire l'intervallo **BILATERALE** di confidenza per la varianza al livello $1 - \alpha$ come:

$$\mathbb{P}(\frac{(n-1)\hat{s}^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}) = 1 - \alpha$$

dove:

- $\hat{s}^2$ è la varianza campionaria calcolata dai dati

In questo caso usiamo $\frac{\alpha}{2}$ e $1 - \frac{\alpha}{2}$ perchè la distribuzione chi-quadro è tutta positiva e noi ricerchiamo quell'area che vale $1 - \alpha$.

Per calcolare l'intervallo di confidenza **UNILATERALE** per la varianza al livello $1 - \alpha$ facciamo:

$$(\frac{(n-1)\hat{s}^2}{\chi^2_{\alpha,n-1}}, \infty) \quad \text{Unilaterale destro}$$

$$(0, \frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha,n-1}}) \quad \text{Unilaterale sinistro}$$

In R per calcolare il valore di $\chi^2_{\frac{\alpha}{2},n-1}$ si usa la funzione `qchisq` :

```
qchisq(p=0.05/2, df=4) = 0.484
```

In questo esempio staremo calcolando il valore per un intervallo bilaterale essendo che facciamo $\frac{\alpha}{2}$

Esempio calcolato

L'esercizio come dati ci fornisce:

- $\hat{s}^2 = 0.24$
- $n = 20$

Vogliamo calcolare l'intervallo di confidenza bilaterale al 95% per la varianza della popolazione, quindi $\alpha = 0.05$.
Dobbiamo calcolare:

- Il valore di $\chi^2_{\frac{\alpha}{2},n-1}$
- Il valore di $\chi^2_{1-\frac{\alpha}{2},n-1}$
- Calcolare l'intervallo di confidenza

1. Calcolare il valore di $\chi^2_{\frac{\alpha}{2},n-1}$:

```
qchisq(p=0.05/2, df=19) = 8.907
```

2. Calcolare il valore di $\chi^2_{1-\frac{\alpha}{2},n-1}$:

```
qchisq(p=1-0.05/2, df=19) = 32.852
```

3. Calcolare l'intervallo di confidenza:

$$\mathbb{P}(\frac{(20-1)0.14}{32.852} < \sigma^2 < \frac{(20-1)0.14}{8.907}) = (0.081 < \sigma^2 < 0.299)$$

Intervalli di confidenza per la media di una popolazione di Bernulli

Sia $X_1, .., X_n$ un campione casuale da una popolazione di Bernulli con parametro $p \in (0,1)$, **ignoto**.
Essendo una Bernulli possiamo stimare il parametro p con la media campionaria $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$
Se $n\hat{p} \geq 5$ e $n\hat{p}(1-\hat{p}) \geq 5$ (campione numeroso) si avrà $X \sim N(n\hat{p}, n\hat{p}(1-\hat{p}))$.
Possiamo costruire l'intervallo di confidenza per la **BILATERALE** al livello $1-\alpha$ come:

$$\mathbb{P}(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) = 1-\alpha$$

Esempio calcolato

L'esercizio come dati ci fornisce:

- $n = 100$
- 80 successi

Dobbiamo calcolare:

- Stimare il parametro p
- Il valore di $z_{\frac{\alpha}{2}}$
- Calcolare l'intervallo di confidenza bilaterale al 95%

1. Stimare il parametro p:

$$\hat{p} = \frac{80}{100} = 0.8$$

2. Calcolare il valore di $z_{\frac{\alpha}{2}}$:

```
qnorm(p=0.05/2, mean = 0, sd=1) = 1.96
```

3. Calcolare l'intervallo di confidenza:

$$\mathbb{P}(0.8 - 1.96\sqrt{\frac{0.8(1-0.8)}{100}} < p < 0.8 + 1.96\sqrt{\frac{0.8(1-0.8)}{100}}) \\ = (0.72 < p < 0.88)$$

[^1]

Formulario generale

Descrizione		Unilaterale	Bilaterale dx	Bilaterale sx
Stimare la media con varianza nota	μ	$(\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$	$(\bar{X} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \infty)$	$(-\infty, \bar{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}})$
Stimare la media con varianza sconosciuta	μ	$(\hat{x} - t_{\frac{\alpha}{2},n-1} \frac{\hat{s}}{\sqrt{n}} < \mu < \hat{x} + t_{\frac{\alpha}{2},n-1} \frac{\hat{s}}{\sqrt{n}})$	$(\hat{x} - t_{\alpha,n-1} \frac{\hat{s}}{\sqrt{n}}, \infty)$	$(-\infty, \hat{x} + t_{\alpha,n-1} \frac{\hat{s}}{\sqrt{n}})$
Stimare la varianza	σ^2	$(\frac{(n-1)\hat{s}^2}{\chi^2_{\frac{\alpha}{2},n-1}} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{1-\frac{\alpha}{2},n-1}})$	$(\frac{(n-1)\hat{s}^2}{\chi^2_{\alpha,n-1}}, \infty)$	$(0, \frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha,n-1}})$
Stimare la media di una Bernulli	p	$(\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$		

Formulario R

Descrizione		Funzione R
Calcolare $z_{\frac{\alpha}{2}}$	$z_{\frac{\alpha}{2}}$	`qnorm(p=0.05/2, mean = 0, sd=1)`
Calcolare $t_{\frac{\alpha}{2},n-1}$	$t_{\frac{\alpha}{2},n-1}$	`qt(p=0.05/2, df=4)`
Calcolare $\chi^2_{\frac{\alpha}{2},n-1}$	$\chi^2_{\frac{\alpha}{2},n-1}$	`qchisq(p=0.05/2, df=4)`