Unidad 1

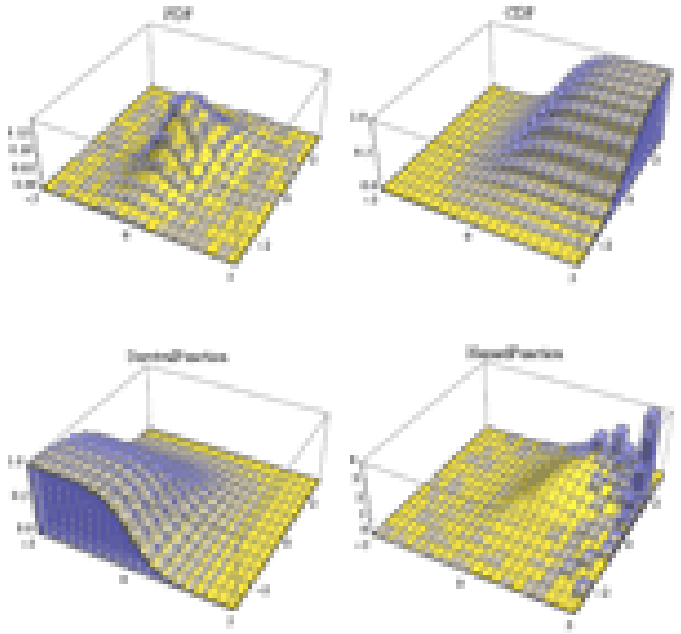
b. Introducción a Estadística Multivariada

CLASE N° 7

IN3401

SEMESTRE OTOÑO, 2012

Distribuciones Multivariadas



$$X = \left\{ x_{ij} \right\} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{bmatrix}$$

Función de Probabilidad(1)

Si $X_1, X_2, \ldots, X_p$ son variables aleatorias *discretas*, definiremos la *función de probabilidad conjunta* de $\mathbf{X}$ como

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p)$$

Propiedades:

- $p(\mathbf{x}) \geq 0$ para todo $\mathbf{x}$.
- $\sum p(\mathbf{x}) = 1$, donde la suma se toma sobre todos los posibles valores de $\mathbf{x}$.

Función de Probabilidad (2)

La *función de probabilidad marginal* para la variable X_i se define como

$$p_i(x_i) = P(X_i = x_i) = \sum p(x_1, \ldots, x_i, \ldots, x_p)$$

donde la suma se realiza sobre todos x tales que su i-ésima componente es x_i.

Diremos que $X_1, X_2, \ldots, X_p$ son *independientes* si

$$p(\mathbf{x}) = \prod_{i=1}^{p} p_i(x_i)$$

Función de Probabilidad (3)

Recordemos que la probabilidad condicional del evento A dado que ocurre el evento B se define como:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Análogamente, si X_1 y X_2 son variables aleatorias discretas, definiremos la *función de probabilidad condicional* de X_1 dado $X_2 = x_2$ como

$$p(x_1|x_2) = P(X_1 = x_1|X_2 = x_2) = \frac{p(x_1, x_2)}{p(x_2)}$$

Con mayor generalidad, puede definirse

$$p(x_1, \ldots, x_k|x_{k+1}, \ldots, x_p) = \frac{p(\mathbf{x})}{p(x_{k+1}, \ldots, x_p)}$$

Función de Distribución Conjunta

En el caso de variables aleatorias continuas, definiremos análogos para la función de densidad y la función de distribución. La *función de distribución conjunta* de $X_1, X_2, \ldots, X_p$ se definirá como

$$F(\mathbf{x}) = F(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p)$$

(Nótese que esta definición es válida también para variables discretas)

Función de Densidad Conjunta

La *función de densidad conjunta* es la función $f(\mathbf{x}) = f(x_1, \ldots, x_p)$ tal que

$$F(\mathbf{x}) = \int_{-\infty}^{x_1} \ldots \int_{-\infty}^{x_p} f(u_1, \ldots, u_p) du_1 \ldots du_p$$

En este caso,

$$f(\mathbf{x}) = \frac{\partial^p F(x_1, \ldots, x_p)}{\partial x_1 \partial x_2 \ldots \partial x_p}$$

La función de densidad conjunta satisface las siguientes propiedades:

- $f(\mathbf{x}) \geq 0$ para todo $\mathbf{x}$.
- $\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(\mathbf{x}) dx_1 \ldots dx_p = 1$.

Distribuciones Marginales y Condicionales

Distribuciones marginales y condicionales pueden ser definidas en el caso continuo de manera análoga al caso discreto.

$$f_i(x_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\mathbf{x}) dx_1 \dots dx_{i-1} dx_{i+1} \dots dx_p = 1$$

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f(x_2)}$$

o, más generalmente,

$$f(x_1, \dots, x_k|x_{k+1}, \dots, x_p) = \frac{f(\mathbf{x})}{f(x_{k+1}, \dots, x_p)}$$

Media, Varianza y Covarianza

El valor esperado de $\mathbf{X}$ es el vector $\boldsymbol{\mu}^T = [\mu_1, \ldots, \mu_p]$ tal que

$$\mu_i = E(X_i) = \int_{-\infty}^{\infty} x f_i(x) dx$$

es la media de la i-ésima componente de X. (Si X_i es discreta, $\mu_i = E(X_i) = \sum x p_i(x)$)

La varianza de la i-ésima componente de $\mathbf{X}$ viene dada por

$$\sigma_i^2 = Var(X_i) = E(X_i - \mu_i)^2 = EX_i^2 - \mu_i^2$$

La covarianza de dos variables X_i y X_j se define como

$$\sigma_{ij} = Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i X_j) - \mu_i \mu_j$$

Nótese que $\sigma_{ii} = \sigma_i^2$

Matriz de Varianzas-Covarianzas (1)

Cuando se trabaja con p variables, se tienen p varianzas y $\frac{1}{2}p(p-1)$ covarianzas. En este caso, es útil presentar todas estas cantidades en una matriz $p \times p$ de la siguiente manera:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & & & \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_p^2 \end{bmatrix}$$

Esta matriz se denomina *matriz de dispersión*, *matriz de variancia y covarianza* o simplemente *matriz de covarianza*. Nótese que esta matriz es simétrica.

Matriz de Varianzas-Covarianzas (2)

Σ puede también escribirse en alguna de las siguientes formas:

$$\begin{aligned}\Sigma &= E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T] \\ &= E[\mathbf{X}\mathbf{X}^T]-\boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

Propiedades:

Consideremos ahora la variable aleatoria univariada Y definida como una combinación lineal de $X_1,\ldots,X_p$. En ese caso

$$Y=\mathbf{a}^T\mathbf{X}$$

donde $\mathbf{a}^T=[a_1,a_2,\ldots,a_p]$ es un vector de constantes. Entonces

$$E(Y)=\mathbf{a}^T\boldsymbol{\mu}$$

$$Var(Y)=\mathbf{a}^T\Sigma\mathbf{a}$$

Matriz de Varianzas-Covarianzas (3)

Estos resultados pueden generalizarse al caso en el cual A es una matriz de constantes $p \times m$. En ese caso $A^T\mathbf{X}$ es un vector $m \times 1$, y tiene vector de medias y matriz de covarianza dados por las siguientes expresiones:

$$E(A^T\mathbf{X}) = A^T\boldsymbol{\mu}$$

$$Var(A^T\mathbf{X}) = A^T\Sigma A$$

Matriz de Correlaciones (1)

La *correlación* entre dos variables aleatorias X_i y X_j se define como

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

- Es una medida de dependencia *lineal* entre variables.
- $|\rho_{ij}| < 1$
- Si X_i y X_j son independientes, $\rho_{ij} = 0$. El recíproco no es cierto.

Cuando se tienen p variables $X_1, X_2, \ldots, X_p$, es en general conveniente presentar las $p(p-1)/2$ correlaciones en una matriz simétrica cuyo elemento ij es ρ_{ij} (nótese que todos los elementos de la diagonal son 1). Esta matriz se denomina *matriz de correlación* y se denotará por P.

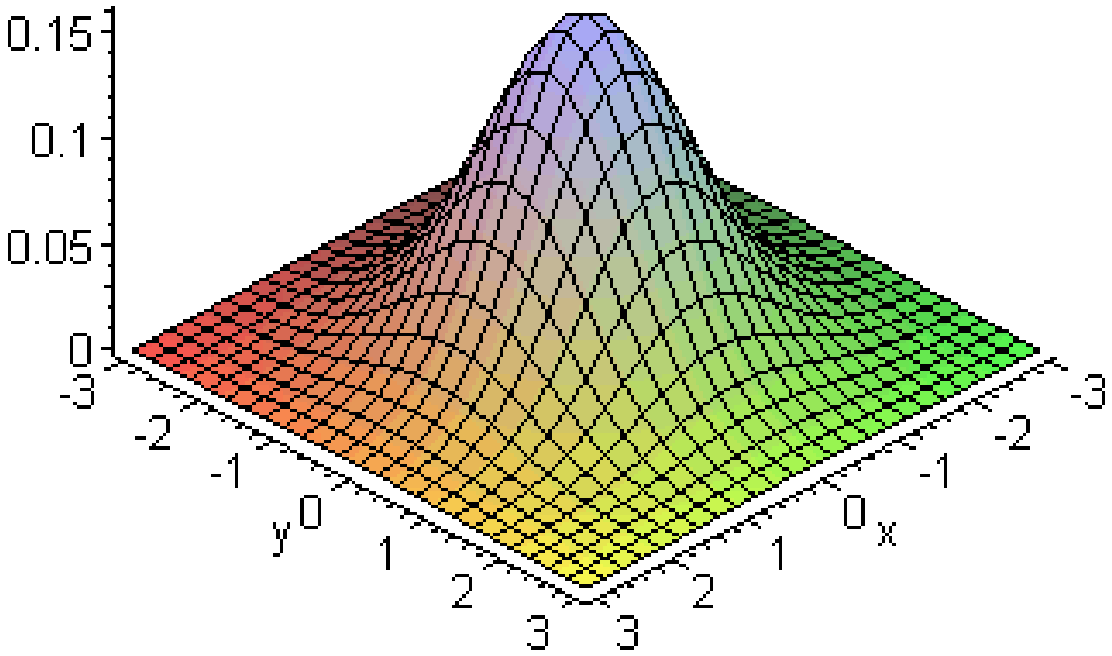
Matriz de Correlaciones (2)

Si definimos $D = diag(\sigma_1, \sigma_2, \ldots, \sigma_p)$, puede verse que:

$$\begin{aligned} \Sigma &= DPD \\ P &= D^{-1}\Sigma D^{-1} \end{aligned}$$

Tanto Σ como P son matrices semipositivo definidas, y tienen el mismo rango. Si $\text{rango}(\Sigma) = \text{rango}(P) = p$, las matrices son positivo definidas.

Normal Multivariada



Normal Multivariada

En el caso univariado, decimos que una variable aleatoria X tiene distribución normal con media μ y varianza σ^2 ($X \sim N(\mu, \sigma^2)$) si su función de densidad tiene la forma

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-(x-\mu)^2/2\sigma^2]$$

En el caso multivariado, decimos que el vector $\mathbf{X}$ sigue una distribución normal multivariada si su función de densidad conjunta es de la forma

$$f(\mathrm{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathrm{x}-\boldsymbol{\mu})]$$

donde Σ es una matriz $p \times p$ no singular, simétrica y positiva definida.

Si $|\Sigma| = 0$, x tienen una distribución degenerada.

Normal Bivariada

En este caso $X = (X_1, X_2)$; la media es $\theta = (\mu_1, \mu_2)$ y la matriz de Varianza-Covarianza tienen la forma:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

donde $|\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2)$; ρ es la correlación entre X_1 y X_2 y $\sigma_{12} = \rho\sigma_1\sigma_2$ es la covarianza entre X_1 y X_2.

- Si $\rho = 0$ X_1 y X_2 no están correlacionadas y además son independientes.
- Si X_1 y X_2 son independientes esto implica que X_1 y X_2 no están correlacionadas. Esto se cumple para todas las distribuciones bivariadas.
- Si $\rho = 0$ esto no implica independencia. Sin embargo esta afirmación si es cierta para la distribución normal.

Distribuciones Normales Condicionales

Distribuciones Normales Condicionales

Distribuciones Normales Condicionales

Sea $X : p \times 1$ un vector de variables aleatorias tal que:

$$\mathbf{X} \sim N(\theta, \Sigma)$$

Entonces la distribución condicional:

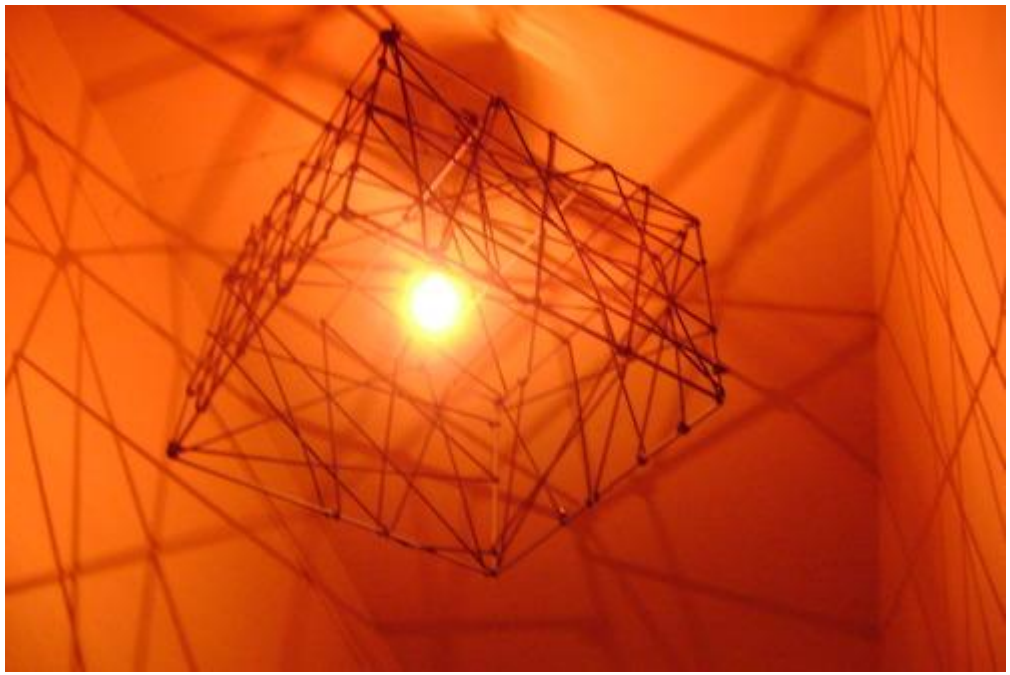
$$Y|Z = z \sim N(\theta_Y + \Sigma_{12}\Sigma_{22}^{-1}(z - \theta_Z), \Sigma_{11,2})$$

donde $\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

En el caso bi-dimensional: $X = (X_1, X_2)'$; $\theta = (\theta_1, \theta_2)$

$\Sigma_{11} = \sigma_1^2$, $\Sigma_{12} = \rho\sigma_1\sigma_2$, $\Sigma_{22} = \sigma_2^2$ y $\Sigma_{11,2} = \sigma_1^2(1 - \rho^2)$

Muestra Aleatoria



Media y Matriz de Covarianza Muestrales (1)

Sean $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ observaciones de una variable aleatoria $\mathbf{X}_{p\times 1}$. Entonces el vector de medias muestrales se calcula como

$$\bar{\mathbf{X}} = \begin{bmatrix} \bar{X}_{,1} \\ \bar{X}_{,2} \\ \vdots \\ \bar{X}_{.p} \end{bmatrix}$$

donde $\bar{X}_{.j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$.

Media y Matriz de Covarianza Muestrales (2)

La matriz de covarianza muestral S contiene los siguientes elementos:

$$
\begin{aligned}
s_{jj} &= \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_{.j})^2}{n-1} \\
s_{jk} &= \frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_{.j})(X_{ik} - \bar{X}_{.k})}{n-1}, \quad i \neq j
\end{aligned}
$$

S puede escribirse con $\frac{1}{n-1} X^{*T} X^{*}$, donde

$$
X^{*} = \begin{bmatrix}
X_{11} - \bar{X}_{,1} & X_{12} - \bar{X}_{,2} & \dots & X_{1p} - \bar{X}_{.p} \\
X_{21} - \bar{X}_{,1} & X_{22} - \bar{X}_{,2} & \dots & X_{2p} - \bar{X}_{.p} \\
\vdots & & & \\
X_{n1} - \bar{X}_{,1} & X_{n2} - \bar{X}_{,2} & \dots & X_{np} - \bar{X}_{.p}
\end{bmatrix}
$$

Media y Matriz de Covarianza Muestrales

La matriz de correlación muestral R tiene elementos

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

(Nótese que si $j = k$, $r_{jk} = 1$)

R puede escribirse como $R = \frac{1}{n-1}\tilde{X}^T\tilde{X}$, donde

$$\tilde{X} = \begin{bmatrix} (X_{11} - \bar{X}_{,1})/s_1 & (X_{12} - \bar{X}_{,2})/s_2 & \dots & (X_{1p} - \bar{X}_{.p})/s_p \\ (X_{21} - \bar{X}_{,1})/s_1 & (X_{22} - \bar{X}_{,2})/s_2 & \dots & (X_{2p} - \bar{X}_{.p})/s_p \\ \vdots & & & \\ (X_{n1} - \bar{X}_{,1})/s_1 & (X_{n2} - \bar{X}_{,2})/s_2 & \dots & (X_{np} - \bar{X}_{.p})/s_p \end{bmatrix}$$

Matriz de Correlaciones Muestrales (1)

La matriz de correlación muestral R tiene elementos

$$r_{jk} = \frac{s_{jk}}{s_j s_k}$$

(Nótese que si $j = k$, $r_{jk} = 1$)

R puede escribirse como $R = \frac{1}{n-1}\tilde{X}^T\tilde{X}$, donde

$$\tilde{X} = \begin{bmatrix} (X_{11} - \bar{X}_{,1})/s_1 & (X_{12} - \bar{X}_{,2})/s_2 & \dots & (X_{1p} - \bar{X}_{.p})/s_p \\ (X_{21} - \bar{X}_{,1})/s_1 & (X_{22} - \bar{X}_{,2})/s_2 & \dots & (X_{2p} - \bar{X}_{.p})/s_p \\ \vdots & & & \\ (X_{n1} - \bar{X}_{,1})/s_1 & (X_{n2} - \bar{X}_{,2})/s_2 & \dots & (X_{np} - \bar{X}_{.p})/s_p \end{bmatrix}$$

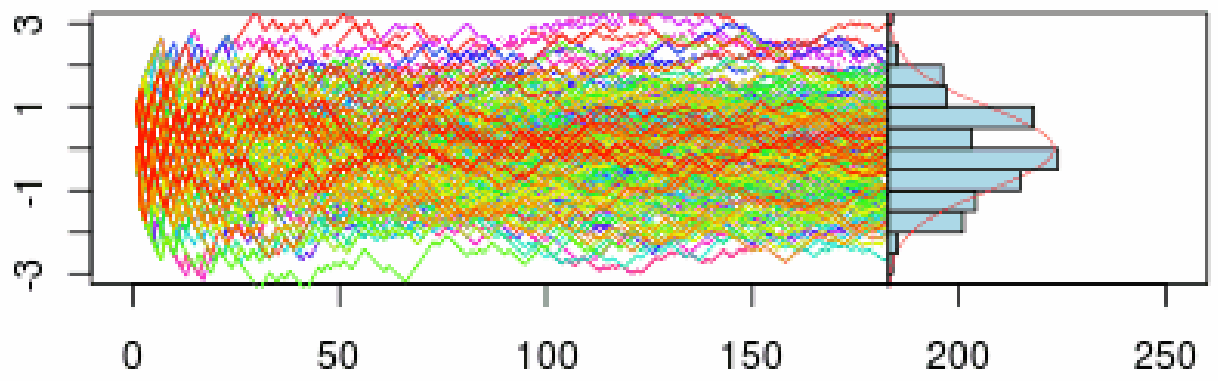
Matriz de Correlaciones Muestrales (2)

La matriz de covarianza muestral y la matriz de correlación muestral se relacionan a través de la expresión

$$R = D^{-1/2} S D^{-1/2}$$

donde $D = diag(s_1^2, s_2^2, \ldots, s_p^2)$.

Teorema Central del Límite



Teorema Central del Límite

Si las filas de la matriz de datos X representan una muestra de una variable aleatoria multivariada $\mathbf{X}$ con $E(\mathbf{X}) = \boldsymbol{\mu}$ y $Cov(\mathbf{X}) = \Sigma$, entonces la distribución asintótica de $\bar{\mathbf{X}}$ es una normal multivariada con vector de medias $\boldsymbol{\mu}$ y matriz de covarianza $\frac{1}{n}\Sigma$.

Región de Confianza

Región de Confianza

- Definition: $P\big(R(\mathbf{X}) \text{ will cover the true } \boldsymbol{\theta}\big) = 1 - \alpha$
- Key ideas: when $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i.i.d.,$

$$P\left(n(\bar{\mathbf{X}} - \boldsymbol{\mu})^{\intercal}\mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)\right) = 1 - \alpha$$

- Conclusion: a 100% confidence region for $\boldsymbol{\mu}$ is the ellipsoid determined by all $\boldsymbol{\mu}$ such that

$$n(\bar{\mathbf{x}} - \boldsymbol{\mu})^{\intercal}\mathbf{s}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p}F_{p,n-p}(\alpha)$$

Intervalos Simultáneos de Confianza

- Definition: $P\big(R(\mathbf{X}) \text{ will cover the true } \boldsymbol{\theta}\big) = 1 - \alpha$
- Key ideas: when $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), i.i.d.,$

$$P\left(n(\bar{\mathbf{X}} - \boldsymbol{\mu})^\top \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right) = 1 - \alpha$$

- Conclusion:

$$\bar{x}_j \pm \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)} \sqrt{\frac{s_{jj}}{n}}, \quad j = 1, \ldots, p$$

- Comparison with one-at-a-time confidence interval:

$$\bar{x}_j \pm t_{n-1}(\alpha/2) \sqrt{\frac{s_{jj}}{n}}, \quad j = 1, \ldots, p$$

Comparación de Medias

Comparación de Medias de dos Poblaciones (1)

Recall univariate two sample t test:

- Setup: $X_{11}, \ldots, X_{1n_1} \sim N(\mu_1, \sigma_1^2)$; $X_{21}, \ldots, X_{2n_2} \sim N(\mu_2, \sigma_2^2)$
- $H_0 : \mu_1 - \mu_2 = \delta$
- Assumptions: independent populations; equal variance $\sigma_1^2 = \sigma_2^2$; normal distributions
- Test statistic:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- Distribution under H_0: $t \sim t_{n_1+n_2-2}$

Comparación de Medias de dos Poblaciones (2)

Multivariate two sample inference:

- Setup: $\mathbf{X}_{11}, \ldots, \mathbf{X}_{1n_1} \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$; $\mathbf{X}_{21}, \ldots, \mathbf{X}_{2n_2} \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$
- $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 = \boldsymbol{\delta}$
- Assumptions: independent populations; equal covariance $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$; normal distributions
- Test statistic:

$$T^2 = \left\{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\delta}\right\}^{\intercal} \left\{S_p \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right\}^{-1} \left\{(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) - \boldsymbol{\delta}\right\}$$

- Distribution under H_0: $T^2 \sim \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1}$
- Confidence region and simultaneous confidence intervals

Comparación de Medias de dos Poblaciones (3)

What if $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$:

- when $n_1 - p$ and $n_2 - p$ are both large, $T^2 \sim \chi^2_p$ approximately
- when n_1 and n_2 are large, $\frac{(n_1+n_2-2)p}{n_1+n_2-p-1} F_{p,n_1+n_2-p-1} \approx \chi^2_p$

So the analysis assuming equal covariance is still valid approximately.