3. PRUEBAS DE SIGNIFICANCIA CON DATOS MULTIVARIADOS

Introducción

Objetivo:

Describir algunas **pruebas o tests** disponibles para constatar si existe alguna evidencia respecto a que **dos o más muestras pertenecen a poblaciones con diferentes medias o diferentes varianzas.**

Comparación de valores medios de dos muestras: Caso univariado

Ejemplo 1

Longitud total de los gorriones. Consideremos la **media** de esta variable. **¿Es la misma para los gorriones s y no-s a la tormenta?**

Disponemos de dos muestras (supuestamente aleatorias): M1) 21 sobrevivientes y M2) 28 no sobrevivientes. **¿Son las medias de estas muestras significativamente diferentes?**

Comparación de valores medios de dos muestras: Caso univariado

Prueba t

- Sea una variable aleatoria X y dos muestras aleatorias de la misma variable, pertenecientes a diferentes poblaciones.

$$x_{i1} \quad i = 1,2,.....,n_1$$
$$x_{i2} \quad i = 1,2,.....,n_2$$

- La media y la varianza de la j-ésima muestra son:

$$\bar{x}_j = \sum_{i=1}^{n_j} \frac{x_{ij}}{n_j} \quad y \quad s_j^2 = \sum_{i=1}^{n_j} \frac{(x_{ij} - \bar{x}_j)^2}{n_j - 1}$$

Comparación de valores medios de dos muestras: Caso univariado

Asumiendo que:

- X tiene una distribución normal en ambas muestras, y
- La hipótesis H_0 (las muestras provienen de una misma distribución),

El **estadígrafo t**:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad con \quad s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

sigue una **ley de distribución de Student**, con n_1+n_2-2 grados de libertad.

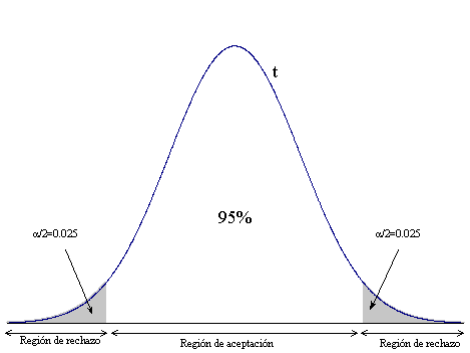
Comparación de valores medios de dos muestras: Caso univariado

- ✓ Esta prueba es bastante robusta a la asunción de normalidad, por lo que si las distribuciones de población de X no son muy diferentes de lo normal, debe ser satisfactorio, en particular para tamaños de muestra de aproximadamente 20 o más.
- ✓ El supuesto de igualdad de varianzas poblacionales tampoco es demasiado crucial si la proporción de las varianzas está en los límites de 0,4 a 2,5.
- ✓ La prueba es particularmente robusta si los dos tamaños de muestra son iguales, o casi.

Comparación de valores medios de dos muestras: Caso univariado

Distribución t de Student

De ser así, el valor t obtenido debería estar dentro del rango de mayor probabilidad según esta distribución. Usualmente se toma como referencia el rango de datos en el que se concentra el 95% de la probabilidad.



Comparación de valores medios de dos muestras: Caso univariado

En otras palabras, la hipótesis H_0 ($\bar{x}_1 = \bar{x}_2$) se rechaza si t no está dentro del rango de mayor probabilidad de esta distribución (95%).

O sea, si t $\longrightarrow$ p_v <0,025

Valores t de Student y probabilidad P asociada en función de los grados de libertad gl.

	P (de una cola)										
gl	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.596
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.706
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
Infinito	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Comparación de valores medios de dos muestras: Caso univariado

- **p-valor**

La elección del nivel de significación **(α=5% o α=1%)**, es en cierta forma arbitraria. El *p-valor* que corresponde al **nivel de significación más pequeño posible** que puede escogerse, para el cual todavía se **aceptaría la hipótesis alternativa H1**.

Al proporcionar el *p-valor* obtenido con la muestra actual, la decisión se hará de acuerdo a la regla siguiente:

si $p_v \leq \alpha$, aceptar H_1

si $p_v > \alpha$, aceptar H_0

Comparación de valores medios de dos muestras: Caso univariado

- Otro modo de obtener la misma información es mediante el cálculo de **intervalos de confianza para la diferencia de las medias en ambos grupos.**

- El intervalo de confianza constituye una medida de la incertidumbre con la que se estima esa diferencia a partir de la muestra, permitiendo valorar tanto la significancia estadística como la magnitud de esa diferencia. El intervalo de confianza vendrá dado como:

Comparación de valores medios de dos muestras: Caso univariado

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975}^{n_1+n_2-2} \otimes s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

- donde t denota el valor que según la distribución t de Student con n_1+n_2-2 grados de libertad deja a su derecha el 2.5% de los datos.

- El intervalo de confianza expresa, en definitiva, un rango de valores entre los que se puede encontrar el valor real de la diferencia entre ambos grupos. Si el valor cero pertenece al intervalo indica que no se dispone evidencia para concluir que las medias son diferentes en ambos grupos.

Comparación de valores medios de dos muestras: Caso multivariado

- **Ejemplo 1**
  La prueba descrita en la sesión anterior puede emplearse para cada una de las 5 medidas de los gorriones s y no-s, para **decidir cuál de estas variables, si alguna, muestran diferentes valores medios según se trate de gorriones s y no-s.**

- Sin embargo, también puede tener algún interés saber si las 5 variables, consideradas conjuntamente, sugieren una diferencia entre gorriones s y no-s. En otras palabras:
  ¿La evidencia total apunta a que las medias, en su conjunto, son diferentes según se trate de gorriones s y no-s?

Comparación de valores medios de dos muestras: Caso multivariado

Prueba multivariada o prueba T^2 de Hotelling

- La prueba T^2 de Hotelling es una generalización de la prueba t de Student. O, para ser más precisos, el cuadrado de la prueba t.
- Caso general: p variables X_1, X_2,...X_p y 2 muestras de longitud n_1 y n_2. Por lo tanto, se tiene: dos vectores de medias:

$$\bar{x}_1 = \begin{vmatrix} \bar{x}_1 \\ \bar{x}_2 \\ . \\ . \\ \bar{x}_p \end{vmatrix} \quad con \quad \bar{x}_j = \sum_{i=1}^{n_1} x_j / n_1 \qquad \bar{x}_2 = \begin{vmatrix} \bar{x}_1 \\ \bar{x}_2 \\ . \\ . \\ \bar{x}_p \end{vmatrix} \quad con \quad \bar{x}_j = \sum_{i=1}^{n_2} x_j / n_2$$

Comparación de valores medios de dos muestras: Caso multivariado

- y dos matrices de covarianza, C_1 y C_2:

$$C = \begin{bmatrix} c_{11} & c_{12} & .. & c_{1p} \\ c_{21} & c_{22} & .. & c_{2p} \\ . & . & .. & . \\ c_{p1} & c_{p2} & .. & c_{pp.} \end{bmatrix} \qquad c_{jj} = s_j^2 = \sum_{i=1}^{n} \frac{(x_{ij} - \bar{x}_j)^2}{n-1} \qquad c_{jk} = \sum_{i=1}^{n} \frac{(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{n-1}$$

- Asumiendo que las matrices de covarianza pertenecen a la misma población, la covarianza ponderada C surge de:

$$C = \frac{(n_1 - 1)C_1 + (n_2 - 1)C_2}{n_1 + n_2 - 2}$$

Comparación de valores medios de dos muestras: Caso multivariado

- El estadígrafo de Hotelling se define como:

$$T^2 = \frac{n_1 n_2 (\bar{x}_1 - \bar{x}_2)' C^{-1} (\bar{x}_1 - \bar{x}_2)}{n_1 + n_2}$$

- Un valor significativamente grande de este estadígrafo evidencia que los vectores de medias, de las dos poblaciones muestreadas, son diferentes.
- Puesto que T^2 tiene forma cuadrática es un escalar. Alternativamente:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{i=1}^{p} \sum_{k=1}^{p} (\bar{x}_{1i} - \bar{x}_{2i}) c^{ik} (\bar{x}_{1k} - \bar{x}_{2k})$$

$\bar{x}_{ji}$ = media de la variable X_i en la j-ésima muestra.

c^{ik} = elemento en la i-ésima fila y k-ésima columna de la matriz inversa C^{-1}

Comparación de valores medios de dos muestras: Caso multivariado

- La significancia o la falta de significancia de T^2 se determina más simplemente utilizando que, en el caso de la hipótesis nula (las medias de las poblaciones son iguales), esto significa que el estadígrafo transformado F sigue una distribución F con p y (n_1+n_2-p-1) g.d.l.

$$F = \frac{(n_1+n_2-p-1)T^2}{(n_1+n_2-2)p}$$

Comparación de valores medios de dos muestras: Caso multivariado

El estadígrafo T^2 de Hotelling se basa en la hipótesis de normalidad e igual variabilidad dentro de la muestra. Más precisamente, se asume que dos muestras que se comparan utilizando el estadígrafo T^2 provienen de distribuciones multinormales* con igual matriz de covarianza.

* La distribución multinormal es similar a la distribución normal o Gaussiana para la variable multinormal. En contraste al caso univariado, la distribución multivariada es definida por la media del vector y la matriz de covarianza.

EJEMPLO DE APLICACIÓN DE LAS PRUEBAS UNI- Y MULTIVARIADA

Ej. 3.1 Gorriones sobrevivientes a una tormenta

¿Existe alguna diferencia entre los gorriones s y no-s, con respecto a los valores medios de las cinco características morfológicas?

a) Caso univariado – Ejemplo: X_1 = longitud total

21 sobrevivientes: $\bar{x}_1 = 157{,}38 \quad s_1^2 = 11{,}05$

28 no-sobrevivientes: $\bar{x}_2 = 158{,}43 \quad s_2^2 = 15{,}07$

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{20 \otimes 11{,}05 + 27 \otimes 15{,}07}{21+28-2} = 13{,}36$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{157{,}38-158{,}43}{\sqrt{13{,}36\left(\frac{1}{21}+\frac{1}{28}\right)}} = -0{,}99 \quad con \quad n_1+n_2-2 = 47 \quad g.d.l.$$

EJEMPLO DE APLICACIÓN DE LAS PRUEBAS UNI- Y MULTIVARIADA

- O bien

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975}^{n_1+n_2-2} \otimes s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = (157{,}38 - 158{,}43) \pm 2{,}017 \otimes \sqrt{13{,}36\left(\frac{1}{21}+\frac{1}{28}\right)} = [-3{,}178\,;\,1{,}078]$$

- **Conclusión:**

  Para un nivel de significancia del 5%, se tiene que:

$$t = 0{,}99 \quad y \quad 47\,g.d.l. \rightarrow p - valor = 0{,}163 > 0{,}025(\alpha/2) \rightarrow se\ acepta\ HO$$

$$-3{,}178 < 0\ (\bar{x}_1 = \bar{x}_2) < 1{,}078$$

Esto indica que no se dispone de evidencia para concluir que las medias sean distintas en ambos grupos (s y no-s).

EJEMPLO DE APLICACIÓN DE LAS PRUEBAS UNI- Y MULTIVARIADA

Table 4.1 Comparison of Mean Values for Survivors and Nonsurvivors for Bumpus's Female Sparrows with Variables Taken One at a Time

Variable	Survivors $\bar{x}_1$	Survivors s_1^2	Nonsurvivors $\bar{x}_2$	Nonsurvivors s_2^2	t (47 df)	P-value[a]
Total length	157.38	11.05	158.43	15.07	−0.99	0.327
Alar extent	241.00	17.50	241.57	32.55	−0.39	0.698
Length of beak and head	31.43	0.53	31.48	0.73	−0.20	0.842
Length of humerus	18.50	0.18	18.45	0.43	0.33	0.743
Length of keel of sternum	20.81	0.58	20.84	1.32	−0.10	0.921

[a] Probability of obtaining a t-value as far from zero as the observed value if the null hypothesis of no population mean difference is true.

EJEMPLO DE APLICACIÓN DE LAS PRUEBAS UNI- Y MULTIVARIADA

b) Caso multivariado

- Para los 21 sobrevivientes:

$$\bar{x}_1 = \begin{bmatrix} 157{,}381 \\ 241{,}000 \\ 31{,}433 \\ 18{,}500 \\ 20{,}810 \end{bmatrix} \quad y \quad C_1 = \begin{bmatrix} 11{,}048 & 9{,}100 & 1{,}557 & 0{,}870 & 1{,}286 \\ 9{,}100 & 17{,}500 & 1{,}910 & 1{,}310 & 0{,}880 \\ 1{,}557 & 1{,}910 & 0{,}531 & 0{,}189 & 0{,}240 \\ 0{,}870 & 1{,}310 & 0{,}189 & 0{,}176 & 0{,}133 \\ 1{,}286 & 0{,}880 & 0{,}240 & 0{,}133 & 0{,}575 \end{bmatrix}$$

- Para los 28 no-sobrevivientes:

$$\bar{x}_2 = \begin{bmatrix} 158{,}429 \\ 241{,}571 \\ 31{,}479 \\ 18{,}446 \\ 20{,}839 \end{bmatrix} \quad y \quad C_2 = \begin{bmatrix} 15{,}069 & 17{,}190 & 2{,}243 & 1{,}746 & 2{,}931 \\ 17{,}190 & 32{,}550 & 3{,}398 & 2{,}950 & 4{,}066 \\ 2{,}243 & 3{,}398 & 0{,}728 & 0{,}470 & 0{,}559 \\ 1{,}743 & 2{,}950 & 0{,}470 & 0{,}434 & 0{,}506 \\ 2{,}931 & 4{,}066 & 0{,}559 & 0{,}506 & 1{,}321 \end{bmatrix}$$

EJEMPLO DE APLICACIÓN DE LAS PRUEBAS UNI- Y MULTIVARIADA

- La matriz ponderada de covarianza es:

$$C = (20C_1 + 27C_2)/47 = \begin{bmatrix} 13{,}358 & 13{,}748 & 1{,}951 & 1{,}373 & 2{,}231 \\ 13{,}748 & 26{,}146 & 2{,}765 & 2{,}252 & 2{,}710 \\ 1{,}951 & 2{,}765 & 0{,}645 & 0{,}350 & 0{,}423 \\ 1{,}373 & 2{,}252 & 0{,}350 & 0{,}324 & 0{,}347 \\ 2{,}231 & 2{,}710 & 0{,}423 & 0{,}347 & 1{,}004 \end{bmatrix}$$

- La matriz inversa de C es:

$$C^{-1} = \begin{bmatrix} 0{,}2061 & -0{,}0694 & -0{,}2395 & 0{,}0785 & -0{,}1969 \\ -0{,}0694 & 0{,}1234 & -0{,}0376 & -0{,}5517 & 0{,}0277 \\ -0{,}2395 & -0{,}0376 & 4{,}2219 & -3{,}2624 & -0{,}0181 \\ 0{,}0785 & -0{,}5517 & -3{,}2624 & 11{,}4610 & -1{,}2720 \\ -0{,}1969 & 0{,}0277 & -0{,}0181 & -1{,}2720 & 1{,}8068 \end{bmatrix}$$

EJEMPLO DE APLICACIÓN DE LAS PRUEBAS UNI- Y MULTIVARIADA

- Con lo que resulta el estadígrafo de Hotelling con 5 g.d.l.:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} \sum_{i=1}^{p} \sum_{k=1}^{p} (\bar{x}_{1i} - \bar{x}_{2i}) c^{ik} (\bar{x}_{1k} - \bar{x}_{2k}) = 2{,}824$$

- y el estadígrafo F con 43 g.d.l.:

$$F = \frac{(n_1 + n_2 - p - 1) T^2}{(n_1 + n_2 - 2) p} = \frac{(21 + 28 - 5 - 1) \otimes 2{,}824}{(21 + 28 - 2) \otimes 5} = 0{,}517$$

- Este no es un valor significativamente grande puesto que los valores significativos de F son mayores que 1. Por lo tanto, **no existe evidencia de una diferencia en las medias de los gorriones s y no-s, tomando en forma conjunta las 5 variables.**

Pruebas multivariadas vs univariadas

Situaciones posibles:

Pruebas	Estadísticamente	
Univariadas	no-significativas	significativas (algunas)
Multivariadas	significativas	no-significativas

Puede ocurrir debido a la acumulación de evidencia de las variables individuales en la prueba global.
Recíprocamente, la prueba multivariada puede ser estadísticamente no significativa, mientras que algunas pruebas univariadas pueden ser estadísticamente significativas, debido a que la evidencia que presentan las variables significativas se amortigua por la evidencia de no diferencias suministrada por las restantes variables.

Pruebas multivariadas vs univariadas

Error tipo I: Se obtiene un resultado estadísticamente significativo cuando, en realidad, las dos muestras que se comparan provienen de la misma población.

- Prueba univariada, nivel de significancia 5%
  P(no significativa) = 0,95 cuando las medias de las poblaciones son idénticas.
- P pruebas independientes
  P(no significativa) = $0,95^p$
- P(por lo menos un resultado significativo) = $1-0,95^p$

Pruebas multivariadas vs univariadas

EJEMPLO:

- Si p=5 P = $1-0,95^5$ = 0,23
  Cuando se trabaja con datos multivariados, las variables usualmente no son independientes.
  Por lo tanto, P = $1-0,95^5$ no representa la probabilidad correcta si las variables se analizan una a una utilizando la prueba t univariada.
- En cambio, en una prueba multivariada (Hotelling T^2) con un 5% de nivel de signficancia:
  P(cometer un error de tipo I)= 0,05 independientemente del número de variables.

Comparación de la varianza de dos muestras: Caso univariado

Prueba F de Fisher

Si s_j^2 representa la varianza de la j-ésima muestra, luego la relación s_1^2 / s_2^2 se compara con la distribución F con (n_1-1) y (n_2-1) g.d.l. Un valor significativamente diferente de 1 evidencia que las muestras pertenecen a dos poblaciones con diferentes varianzas.

Desventajas: La prueba F es sensible a la hipótesis de normalidad. Un resultado estadísticamente significativo puede deberse a que la variable no se ajusta a una distribución normal y no a que las varianzas sean diferentes.

Comparación de la varianza de dos muestras: Caso univariado

- **Prueba de Levene (1960)**
  Propuso una alternativa más robusta, que consiste en transformar los datos originales en desviaciones absolutas respecto a la media de las muestras y, seguidamente, examinar si existe una diferencia significativa entre las medias de las desviaciones de las dos muestras, utilizando para ello la prueba t.
- **Prueba de Schultz (1983)**
  Propuso una alternativa aún más robusta, utilizando las desviaciones absolutas respecto a la mediana de las muestras.

Comparación de la varianza de dos muestras: Caso multivariado

- **La prueba M de Box**

  Utilizada por muchos paquetes computacionales, compara la varianza de dos **o más** muestras multivariadas. (se analizará cuando veamos el caso de "varias muestras").

  Esta prueba es muy sensible respecto a la suposición en cuanto a que las muestras provienen de distribuciones multinormales. ➡ Existe, por lo tanto, siempre la posibilidad de que un resultado significativo se deba a la no-normalidad, más que a la desigualdad de las matrices de covarianza poblacional.

Comparación de la varianza de dos muestras: Caso multivariado

- ***Procedimiento alternativo con base en Levene:***

  Los datos se transforman en desviaciones absolutas de las medias o las medianas. Por lo tanto, la interrogante en cuanto a si dos muestras presentan varianzas significativamente diferentes se transforma en una interrogante en cuanto a las diferencias entre los vectores de las medias de los datos transformados. Esto se analiza utilizando la prueba T^2 de Hotelling.

Comparación de la varianza de dos muestras: Caso multivariado

- **Prueba de Van Valen (1978)**

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ijk} - \bar{x}_{jk})^2}$$

Donde:

x_{ijk} = valor de la variable X_k para la i-ésima observación de la j-ésima muestra.

$\bar{x}_{jk}$ = media de la variable X_k para la j-ésima muestra.

- Las **medias de las muestras de los valores d_{ij}** se comparan aplicando la **prueba t de Student**.
- Para asegurar que todas las variables tienen igual peso, deben **estandarizarse** antes de calcular los valores d_{ij}

Comparación de la varianza de dos muestras: Caso multivariado

- Una prueba más robusta resulta de considerar las medianas de las muestras:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ijk} - M_{jk})^2}$$

donde:

M_{jk} = mediana de la variable X_k en la j-ésima muestra.

Comparación de la varianza de dos muestras: Caso multivariado

Ambos métodos se basan en una hipótesis implícita: "Si dos muestras difieren, una muestra es más variable que la otra, para todas las variables."

Sin embargo, no se puede esperar un resultado estadísticamente significativo en el caso que:

X_1 y X_2 son más variables en la muestra 1, pero X_3 y X_4 lo son en la muestra 2, puesto que el efecto de las diferentes varianzas tiene a cancelarse en el cálculo de d_{ij}.

Comparación de la varianza de dos muestras: Caso multivariado

La prueba de Van Valen no es apropiada en situaciones en que no se espera que los cambios en el nivel de variación sean consistentes para todas las variables.

Comparación de la varianza de dos muestras: Caso multivariado

Ej. 3.2 Varianza de los gorriones hembra de Bumpus

¿Son los gorriones no-s más variables que los gorriones s? Esto es lo que se espera si es válida la hipótesis de selección.

Caso univariado. X1 (longitud total)

a) Prueba de Levene (datos = desviaciones de la mediana)

$M_1(s) = 157\ mm$

$\bar{x}_1 = 2{,}57 \quad s_1^2 = 4{,}26$

$M_2(no-s) = 159\ mm$

$\bar{x}_2 = 3{,}29 \quad s_2^2 = 4{,}21$

$s^2 = 4{,}231$ Varianza ponderada

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2{,}57 - 3{,}29}{\sqrt{4{,}231\left(\frac{1}{21} + \frac{1}{28}\right)}} = -1{,}21 \quad con \quad 47 \quad g.d.l.$$

Comparación de la varianza de dos muestras: Caso multivariado

Conclusión:

- ✓ Para un nivel de significancia del 5%, se tiene que:

$t = -1{,}21 \rightarrow p_{valor} = 0{,}116 > 0{,}025$ → se acepta Ho

- ✓ Para las restantes variables:

Variable	t	p-valor
X1	1,21	0,116
X2	1,18	0,122
X3	0,81	0,211
X4	1,91	0,031
X5	1,4	0,084

Solamente para la variable X_4 el valor de t es significativamente bajo para un nivel de significancia de 5%

$$(\bar{x}_1 - \bar{x}_2) \pm t_{0,975}^{n_1+n_2-2} \otimes s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = (2{,}57 - 3{,}29) \pm 2{,}017 \otimes \sqrt{4{,}231\left(\frac{1}{21} + \frac{1}{28}\right)} = [-1{,}918\,;\,0{,}478]$$

Comparación de la varianza de dos muestras: Caso multivariado

La Tabla muestra las desviaciones absolutas de las medianas de las muestras, para los datos de Bumpus después de haber sido estandarizados para la prueba de Van Valen y los valores de d:

Pájaro	X1	X2	X3	X4	X5	d
1	0,28	1,00	0,25	0,00	0,10	1,07
2	0,83	0,00	1,27	1,07	1,02	2,12
3	1,11	0,00	0,51	0,18	0,00	1,23
.						
.						
.						

Comparación de la varianza de dos muestras: Caso multivariado

Comparando los vectores de la media de la muestra transformada para las cinco variables:

- ✓ T^2 (Hotelling)=4,75
- ✓ F=0,85 con 5 y 43 g.d.l.

  Por lo tanto, no existe evidencia de una diferencia significativa entre las muestras de esta prueba, puesto que el F-valor < 1.
- ✓ Prueba de Van Valen.

  Considerando los valores de d:

$$s: \quad \bar{x}_d = 1{,}760 \quad s_d^2 = 0{,}411$$
$$no-s: \quad \bar{x}_d = 2{,}265 \quad s_d^2 = 1{,}133$$

t=-1,92 p-valor = 0.030, que es significativamente próximo al valor de $t_{crítico}$ para un nivel de significancia del 5%. Por lo tanto, **la prueba indica mayor variación entre los gorriones no-s que entre los s.**

Comparación de la varianza de dos muestras: Caso multivariado

El **resultado significativo con la prueba de Van Valen y no-significativo con la prueba de Levene**, se explica en que la prueba de T^2 no es direccional y no toma en cuenta la expectativa de que los gorriones s serán, en todo caso, menos variables que los gorriones no-s. Por otra parte, la prueba de Van Valen es específica para menor variación en la muestra 1 que en la muestra 2, para todas las variables. En este caso, todas las variables muestran menor variación en la muestra 1 que en la muestra 2. La prueba de Van Valen enfatiza este hecho, pero la prueba de Levene no.

Comparación de las medias para varias muestras. Una sola variable y varias muestras

- Cuando se tiene **una sola variable y varias muestras**, la generalización de la prueba t es la prueba F de análisis de la varianza con 1-factor (one-factor ANOVA).

Fuente de variación	Suma de cuadrados	g.d.l.	Media de los cuadrados	F
Entre muestras (o Inter-grupos)	B=T-W	m-1	M_B=B/(m-1)	M_B/M_W
Dentro de las muestras (o Intra-grupos)	$W=\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^2$	n-m	M_W=W/(n-m)	
Total	$T=\sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij}-\bar{x})^2$	n-1		

Comparación de las medias para varias muestras. Una sola variable y varias muestras

donde:

n_j	tamaño de la j-ésima muestra
$n = \sum_{j=1}^{m} n_j$	número total de observaciones m muestras
x_{ij}	i-ésima observación en la j-ésima muestra
$\bar{x}_j = \sum_{i=1}^{n_j} \frac{x_{ij}}{n_j}$	media de la j-ésima muestra
$\bar{x} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{x_{ij}}{n}$	media de todas las observaciones

Comparación de las medias para varias muestras. Varias variables y varias muestras

Cuando hay **varias variables y varias muestras**, la situación se complica por el hecho de que hay **cuatro estadísticas alternativas** que se utilizan comúnmente para **probar la hipótesis de que todas las muestras provienen de poblaciones con el mismo vector medio.**

Comparación de las medias para varias muestras. Varias variables y varias muestras

- **Estadígrafo lambda de Wilks**

$$\Lambda = |W|/|T|$$

$|W|$= determinante de la matriz suma de cuadrados y productos cruzados de las muestras

$|T|$= determinante de la matriz global suma total de cuadrados y de productos cruzados

Esencialmente, esto compara la variación dentro de las muestras con la variación dentro y entre las muestras.

Comparación de las medias para varias muestras. Varias variables y varias muestras

Matrices T y W:

Sea $\mathbf{x_{ijk}}$ el valor de la variable $\mathbf{X_k}$ para **la i-ésima observación** en la **j-ésima muestra**, $\bar{x}_{jk}$ la **media de $\mathbf{X_k}$** en la misma muestra, y $\bar{x}_k$ la **media global de $\mathbf{X_k}$** para todo el conjunto de datos. Adicionalmente, **m muestras**, siendo **la j-ésima de tamaño $\mathbf{n_j}$**. Luego, **el elemento en la fila r y columna c** de **T** es:

$$t_{rc} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ijr} - \bar{x}_r\right)\left(x_{ijc} - \bar{x}_c\right)$$ global

y el elemento en la fila r y columna c de **W** es:

$$w_{rc} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ijr} - \bar{x}_{jr}\right)\left(x_{ijc} - \bar{x}_{jc}\right)$$ dentro de las muestras