

## 4.2 TEST ESTADÍSTICS

### COMPARACIÓ I AVALUACIÓ D'UNA POBLACIÓ DE DADES

- **Test estadístic (hipòtesi):** procediment estadístic que fa servir un nivell de significació per decidir entre dues hipòtesis.
- **Criteri de significació:** és un criteri estadístic que s'utilitza per a contrastar si les diferències entre dos valors (o grups de valors) són estadísticament significants a un cert nivell de probabilitat (**nivell de significació,  $\alpha$** ), o poden ser justificats com a variacions aleatòries. ex: si ( $\alpha = 0,05$ ) és la cua, la part que rebutjo. **Nivell de confiança  $1-\alpha$**  Ex: 95%
  - Atribueixo que la diferència de 2 valors es deguda a errors aleatoris amb un nivell de confiança del 95%.
- **Hipòtesi nul·la  $H_0$ :** no hi ha diferencia significativa entre els valors que he comparat.  **$H_0 = a$**
- **Hipòtesi alternativa  $H_1$ :** si que hi ha diferències significatives entre els valors que comparo. Supera els errors aleatoris.  **$a \neq b$  ( $< o >$ )** → Són diferents o és major o és menor.

Exemples de comparació de dades:

1. **Comparació de mitjanes amb un valor veritable o de referència**
2. **Comparació de variàncies de dos sets de dades**
3. **Comparació de mitjanes de dos sets de dades**
4. **Comparació de dos sets amb valors aparellats**
5. **Comprovació del tipus de distribució de dades**
6. **Detecció d'anòmals**

### TIPUS DE TESTS:

Hi ha molts test estadístics basats en conèixer la distribució de dades. S'aplica aquell test que descriu aquells paràmetres determinats. Puc aplicar els test quan se com és la distribució de dades.

- **Test paramètric:** Normalment suposem que segueix una distribució normal o gaussiana, per tal de poder fer servir **t test, F Fisher, Grubbs'test**. Els valors venen d'una escala d'interval iguals.
- **Test no paramètrics:** no es basen en suposicions paramètriques:
  - Poden substituir test t en alguns casos (és útil per a distribucions no conegudes)
  - Poden ser útils per identificar la funció de la distribució de la població de dades
  - Ex: Box-whiskers plots; chi-square test, Kolmogorov-Smirnov tests (one or two sided); Wilcoxon tests (signed-rank; rank-sum (Mann-Whitney U test);...)

### 1. COMPARACIÓ DE MITJANES AMB UN VALOR DE REFERÈNCIA( $\mu$ )

Fem la hipòtesi nul·la sempre.

Escollim l'estadístic segons les condicions. La distància entre mitjana i valor de referència dividit/normalitzat per la desviació estàndard.

A partir de la **z o de la t tabulada** calcularé l'alfa amb el nivell de significació que jo he preestablert a priori.

A partir de la alfa puc calcular el valor de l'estadística crític:

- **Calculat < crític:** accepto la hipòtesi nul·la

Null hypothesis

$$H_0: \bar{X} = \mu$$

Alternative hypothesis

$$H_1: \bar{X} \neq \mu$$

$$H_1: \bar{X} > \mu \text{ or } \bar{X} < \mu$$

Statistics to be calculated:

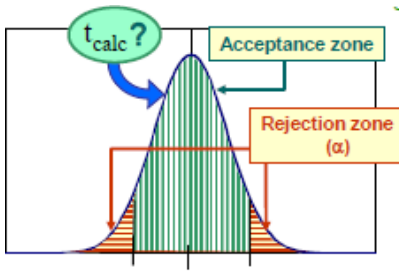
$n > 30$  (and/or  $\sigma$  is known)

$$z_{\text{calc}} = \frac{|\bar{X} - \mu|}{\sigma / \sqrt{n}}$$

$n < 30$  ( $\sigma$  is unknown)

$$t_{\text{calc}} = \frac{|\bar{X} - \mu|}{s / \sqrt{n}}$$

- **Calculat > crític:** accepto la hipòtesi alternativa.



**Test de dues cues:** és la mitjana significament diferent del valor de referència? Si tenim hipòtesi alternativa és diferent de  $\mu$  i pot estar per sobre o per sota.  $H_0 : x = \mu$        $H_1 : x \neq \mu$

**Test d'una cua:** és la mitjana significament major o menor que el valor de referència? Només tenim una àrea de rebuig. Si tenim hipòtesi alternativa  $x$  només pot estar per sobre o per sota.

$H_0 : x = \mu$     $H_1 : x < \mu$    o    $H_1 : x > \mu$

**ERRORS ASSOCIATS:**

Qualsevol test estadístic està subjecte a un error. Amb un 95% de confiança accepto o rebutjo una hipòtesi. Ens marquem un nivell de significació. La resolució de la hipòtesi es fa amb un nivell de probabilitat. La veritat no la sabem; podem cometre un error.

- **TIPUS I ( $\alpha$ ):** quan rebutgem la hipòtesi nul·la quan és veritat. Concloem que la mitjana és significament diferent del valor de referència quan no ho és.
- **TIPUS II ( $\beta$ ):** quan acceptem la hipòtesi nul·la quan és falsa. Concloem que la mitjana no és significament diferent del valor de referència quan ho és.

**Quan rebutgem la  $H_0$ , podem cometre error tipus I**  
**Quan acceptem la  $H_0$  podem cometre error tipus II**

No els puc cometre els dos alhora.

**ERROR DE TIPUS I (ERROR  $\alpha$ )**

**Error alfa:** error de la zona que rebutjo. L'àrea alfa: determina la probabilitat de cometre un error de tipus I. Ex: del 5% Quan jo marco aquests nivells de probabilitat ja estic marcant l'error alfa.

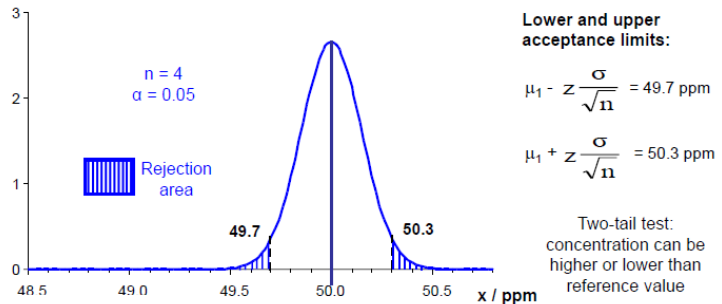
Alternativa: puc calcular tcalculada i recalculer alfa del 2/3% del meu estadístic t calculat, sabent les dades que tinc calculo la probabilitat d'aquestes dades. Per no cometre l'error de tipus alfa. És la probabilitat associada a l'estadístic calculat. Serà l'àrea de rebuig nova. I no tindrem error tipus alfa.

Un increment en el nivell de significació comporta un increment en la probabilitat de tenir error de tipus I.

**Si  $t_{calc} < t_{crit} \rightarrow H_0$  acceptada**  
 (no puc cometre error  $\alpha$ )  
**Si  $t_{calc} > t_{crit} \rightarrow H_0$  rebutjada**  
 (puc cometre error  $\alpha$ )

**ERROR DE TIPUS II (ERROR  $\beta$ )**

El valor real és igual a  $\mu = 50$  ppm. D'acord amb aquest criteri, una mitjana inferior a 49.7 ppm o superior a 50,3 ppm fàcilment podria portar a rebutjar  $H_0$ , i un error de tipus I es podria cometre.

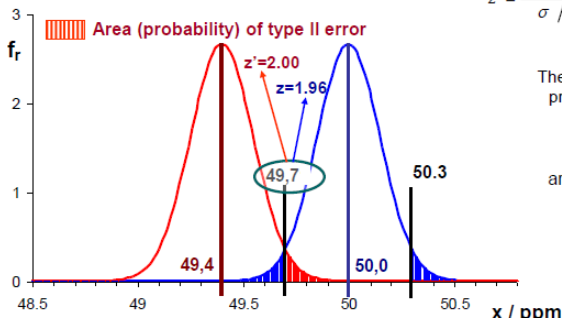


$$z' = \frac{|x_{\text{mín}} - \bar{x}_2|}{\sigma / \sqrt{n}} = \frac{|49,7 - 49,4|}{0,3 / \sqrt{4}} = 2,00$$

The area corresponding to the probability of type II error is quantified by  $z'$ :

area ( $z' = 2.00$ ; 1 tail) = 0.023

**Probability = 2.3 %**



No obstant això, si considerem que tenim una població sospitosament dolenta ja que la mitjana dóna 49,4 ppm. Hi ha una probabilitat que, en un set d'anàlisi, una valors de la població defectuosa puguin caure dins la regió d'acceptació de la població correcta, que portaria a acceptar  $H_0$ .

Per això, trobem el punt de tall amb el límit inferior  $z=2$  anem a la taula i busquem d'una cua

i ens dona que la probabilitat és de 2, 3%. La mitjana és inferior al límit inferior del valor real. Els valors es solapen amb el límit inferior i em portà a acceptar valors que no hauria.

Un increment en el nivell de significació comporta una disminució en la probabilitat de tenir error de tipus II. La probabilitat de tenir error de tipus II augmenta quan el número de replicats disminueix.

## 2. TEST FISHER PER COMPARAR VARIÀNCIES DE DOS SETS DE DADES

La població de variàncies i la relació de variàncies no segueixen una distribució normal, sinó una de **Fisher**.

**Statistic:** L'estadístic es calcula entre **variàncies**: ha de ser més gran que 1 obligatori i cal que al

$$F_{\text{calc}} = \frac{S_1^2}{S_2^2}$$

where  $S_1^2 > S_2^2$   
( $F_{\text{calc}} \geq 1$ )

numerador estigui el número més gran matemàticament parlant.

- També es fa servir un nivell de significació:  $\alpha = 0,05$  normalment
- Graus de llibertat dels dos sets de dades 1 i 2

Podem tenir de **dues cues** (diferent) i **d'una cua** (és més gran o més petit)

### Two-tail test:

Null hypothesis  $H_0: s_1^2 = s_2^2 \leftarrow F_{\text{calc}} < F_{\text{crit}}$

Alternative hypothesis  $H_1: s_1^2 \neq s_2^2 \leftarrow F_{\text{calc}} > F_{\text{crit}}$

### One-tail test:

Null hypothesis  $H_0: s_1^2 = s_2^2 \leftarrow F_{\text{calc}} < F_{\text{crit}}$

Alternative hypothesis  $H_1: \begin{cases} s_1^2 > s_2^2 \\ s_1^2 < s_2^2 \end{cases} \leftarrow F_{\text{calc}} > F_{\text{crit}}$

Distribució de Fisher depèn molt dels graus de llibertat. Quan es dispara ( $\geq 100$  aprox) la funció de Fisher pren una forma gairebé simètrica. Distribució amb una freqüència màxima desplaçada cap a l'esquerra (esbiaixada, *skewed*).

Cada tipus de cua ve marcada per un nivell de significació determinat.  $F(\alpha, n_1, n_2)$  Taula:

Horitzontal: d.o.f. numerador

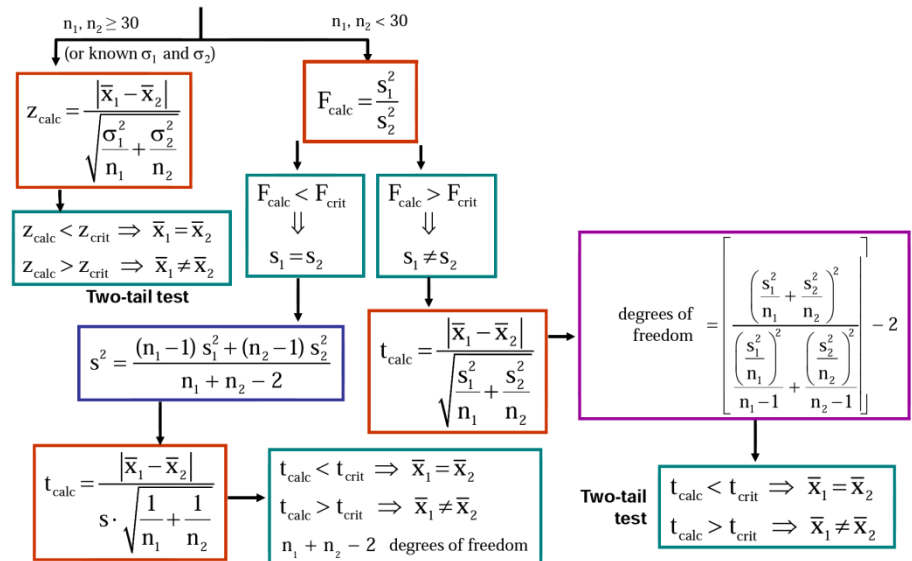
Vertical: d.o.f. denominador

Problema: no conec les desviacions estàndard de cada població.

## 3. COMPARACIÓ DE MITJANES DE DOS SETS DE DADES

Si conec les sigmes de cada població puc aplicar el **test z**. La desviació estàndard que tinc en compte es la  $\sigma$  de cada població. Les variàncies estan corregides pels graus de llibertat.

Si no conec sigmes o valors petits, faig un test Fisher per veure si les sigma o les variàncies son iguals o no. Si ho son calculo la desviació estàndard ponderada calculo paràmetre t. Si són diferents calculo estadístic t treballant de manera diferent. Formula de dof (la donen).



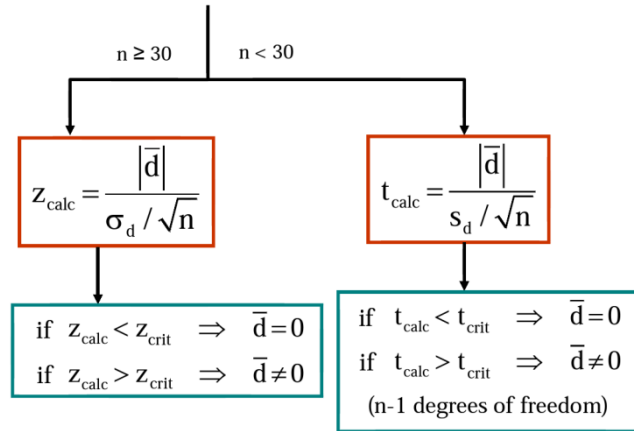
One-tail tests ( $H_1: \bar{x}_1 > \bar{x}_2$ ) are also possible

## 4. TEST T DE DIFERENCIES DE DOS SETS DE DADES AMB VALORS APARELLATS

Comparo valors que no venen de la mateixa rèplica, sinó que venen de diferents mostres amb de mètodes diferent. **Test t o z:** és un test de diferències entre valors aparellats: suposem que si jo fes rèpliques tindria una distribució normal (no fem la hipòtesi).

1. Calculo la **diferència** de cada valor per a cada mostra. Així puc calcular la mitjana de les diferències i la desviació estàndard de les diferències.
2. **Hipòtesi nul·la:** no hi ha diferència entre els valors obtinguts per tant la diferència tendeix a 0. No puc comparar mitjanes sinó **diferències** entre elles. + la població de diferències es normal: puc calcular mitjana aritmètica i desviació estàndard.

**Example:** comparison of two methods applied to n samples



Si comparés mitjanes (tipus 3.) puc compara dues que tinguin diferent número de rèpliques. Aquí no: perquè comparem la diferència i hem de fer la resta de cada valor → igual número de rèpliques. Hi ha d'una o dues cues.

### 5. TEST DE RANGS AMB SIGNE DE WILCOXON:

Resultats	Rang (Signe)
-1	1(-)
-4	2(-)
+7	3(+)
+11	4(+)
<b>W=4</b>	

És el que es fa servir ja que es un test robust no paramètric. Test de rangs amb signe de Wilcoxon. Test es idèntic a l'altre amb igual objectiu: establir si dues sèries o parelles són iguals. Calculem la diferència de valors ordenats, i miro les diferències que té. Les diferències s'ordenen en valor absolut de menor a major però mantenint el signe.

**Hipòtesi** la suma de rangs ha de tendir a 0 degut a que els rangs + i rangs - han

distribuïts. No es basa en desviació.

En valor absolut sempre.

**Si n > 10:** la probabilitat distribució normal.

If n > 10: the probability of having +/- ranks is normally distributed

$$H_0: \mu_w = 0 \quad \sigma_w = \sqrt{n(n+1)(2n+1)/6}$$

Statistic:

$$z_{\text{calc}} = \frac{W \pm 0.5}{\sigma_w}$$

If n < 10: W is compared to  $W_{\text{crit}}$  →

n	±W <sub>crit</sub> for small populations			
	Level of significance			
	0.05	0.025	0.01	0.005
	One-tail			
	0.05	0.025	0.01	0.005
	Two-tail			
5	15	-	-	-
6	17	21	-	-
7	22	24	28	-
8	26	30	34	36
9	29	35	39	43

d'estar igualment fer la mitjana aritmètica i

**Suma de rangs = 4 positiu.** de tenir +/- rangs té una

## 6. DETECCIÓ D'ANÒMALS:

Fins ara: hem comparat mitjanes i hem fet la hipòtesi de quedar-nos totes les dades.

- Els anòmals distorsionen la distribució de dades i afecten a l'estimació de la mitjana i la desviació estàndard.
- La detecció dels valors anòmals i supressió ha de ser un pas previ a la aplicació del test estadístic.
- Test paramètric i no paramètrics es poden fer servir per a detectar anòmals.
- Diagrames de caixes i bigotis.

Si estan a l'extrem 95% són de la població encara que siguin molt grans.

### Z-score

$$Z_i = \frac{X_i - \bar{X}}{S}$$

Es basa en com de gran és la distància d'un resultat donat ( $x_i$ ) a la mitjana, respecte la desviació estàndard. Quan  $z_i > 2,5$  →  $x_i$ : es considerat un valor anòmal. → 97/98%  
És útil per a poblacions grans. Si tenim molts anòmals per una població petita no val la pena.

### Test de Grubbs

Basat en el concepte de z-score però s'aplica per a poblacions petites.

Compara el valor mínim i el màxim amb la mitjana i la desviació estàndard de la població (aproximació d'una cua).

- $H_0$ : no hi ha anòmals a la població
- $H_1$ : hi ha un anòmal en la població
- Si  $G_{\text{calc}} > G_{\text{crit}}$  el valor examinat és un anòmal.
- Fiable quan només un anòmal esta present.

$$G_{\text{crit}} = \left( \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2}{n-2+t^2}} \right); \text{ d.o.f.} = n-2$$

Statistics:

$$G_{\text{calc min}} = \frac{\bar{X} - X_{\text{min}}}{S}$$

$$G_{\text{calc max}} = \frac{X_{\text{max}} - \bar{X}}{S}$$

**Residu estandarditzat robust:** Residual = diferència

Es basa en com de gran es la distància d'un resultat al valor central **mediana**, respecte la desviació absoluta mediana **MAD**.

$$\text{MAD} = \text{mediana}(x_i - \text{median}(x_i))$$

Es pot corregir amb un factor de consistència si sabem com es la distribució. Normalment suposem que es gaussiane.  $F(DN) = 1,428$

$$\text{Robust standardized residual}_i = \frac{x_i - \text{median}(x_i)}{\text{MAD} \cdot \text{Consistency factor}}$$

(Consistency factor is 1.428 if distribution is assumed to be normal)

Mirem quantes vegades respecte desviació estàndard m'allunyo del valor central.

Si **residu estandarditzat robust**  $> 2,5$ ,  $x_i$  es considera com un anòmal.

- Més d'un anòmal pot estar present a la població, degut a que la mediana i la MAD es veuen menys afectades per la presència d'anòmals.

## Diagrames de caixa i bigotis: *Box and whisker plots*

Molt senzills de construir, hi ha de diferents tipus.

Son útils per **descriure/dibuixar** la població de dades, per indicar si una distribució es esbiaixada (skewed) i si hi ha observacions potencialment inusuals (anòmals) en el set de dades. Es basen en calcular la mediana, el q1 i q3 → IQR.

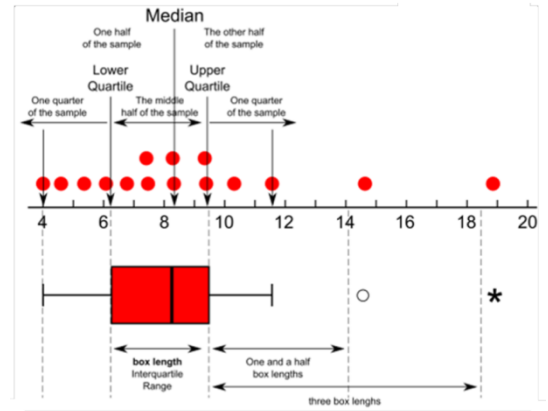
Com construir-ho:

- Els **extrems** de la caixa són els **quartil superior i inferior**, així que la **caixa** és el interval **interquartil**.
- La **mediana** ve marcada per una **línia vertical** dins de la caixa.
- **Bigotis**: dues línies fora de la caixa que s'estenen a l'observació més gran i la més petita.

Com més **ample** sigui la caixa millor perquè més gran serà IQ i així tinc menys probabilitat de tenir anòmals.

- Qualsevol valor és considerat un **anòmal lleuger** si cau **1,5xIQR** per sobre del tercer quartil o per sota del primer quartil.
- Qualsevol valor és considerat un **anòmal sever** si cau **3xIQR** per sobre del tercer quartil o per sota del primer quartil.
- Quan detecto els anòmals, els elimino i **recalculo els bigotis**: box and whiskers modificat per identificar anòmals. La caixa s'acaba al valor que no està dins de ser candidat de ser anòmal ni lleuger ni sever!!

Exemple: La mediana està situada cap a la l'esquerra. Potser l'extrem màxim es anòmal. → recalcular.



No esta a l'Excel però hi ha un *template*: Només representa els mínims o màxims anòmals. Però jo se fins on arriba bigoti així ja sé que puc eliminar.

**$\chi^2$  (CHI-SQUARE)**: test per a comprovar la funció de distribució d'una població de dades:

- Comprova si totes les dades examinades pertanyen a la mateixa població de dades (Especialment per a variables no contínues). Si no pertany podria ser degut a l'existència d'anòmals.
- Comprova la funció de distribució que segueix una població de dades.
- Les dades s'organitzen en classes
- S'ha de fer una hipòtesi en la funció de distribució potencial (Ex: normal, lognormal, triangular, rectangular...)
- Cal un mínim de dades (ex: per una distribució normal:  $n > 40$ )
- Els graus de llibertat:  $n-1$

$$\chi^2_{\text{calc}} = \sum_i \frac{(f_i^{\text{exp}} - f_i^{\text{theor}})^2}{f_i^{\text{theor}}}$$

$F_i^{\text{exp}}$ = freqüència experimental absoluta: nombre de dades (àrea) que comprèn cada classe

$F_i^{\text{teo}}$ = freqüència teòrica absoluta: nombre de dades (àrea) que compren cada classe d'acord amb la funció de distribució teòrica. Ex: si és rectangular dividir el nombre d'observacions/  $n^{\circ}$  de classes!!

Treballa per freqüències absolutes: he de pensar com agrupo les dades en classes → quin **criteri**? En funció de com cregui les classes puc rebutjar hipòtesi.

$\chi^2_{\text{calc}} < \chi^2_{\text{crit}}$  **H<sub>0</sub>**: La dada segueix la distribució: no hi ha diferència significativa entre la teòrica i l'experimental.  
 $\chi^2_{\text{calc}} > \chi^2_{\text{crit}}$  **H<sub>1</sub>**: La dada no segueix la distribució

Taula: amb alfa i graus de llibertat. → si augmenta el número de classes arriba a distribució normal. En canvi a k petites està esbiaixada cap a la esquerra.

**KOLMOGOROV-SMIRNOV. TEST ROBUST:** serveix per veure quin tipus de distribució segueix la població de dades. Test de comparació entre freqüències experimentals i teòriques no **absolutes**, sinó **acumulades** fins a un valor real. No cal definir classes.

Es calcula per posició relativa. Posició 3 → freqüència 3. Es un **rang!!!**

La freqüència màxima és:  $D^+ = i/n$  i la mínima  $D^- = (i-1)/n$

$D_{crit}$ : depèn de la distribució teòrica,  $n$  i el nivell de significació.

$$D^+ = \max [i/n - F_o(x_i)] \quad D^- = \max [F_o(x_i) - (i-1)/n]$$

$$D_{calc} < D_{crit} \quad H_0: S'accepta$$

$$D_{calc} > D_{crit} \quad H_1: Rebutgem H_0$$

- $D_{calc}$  is the highest absolute difference between observed and theoretical frequencies.

**Statistics:**  $D_{calc} = \max [D^+, D^-]$

- $D_{crit}$ :  $c_\alpha/k_{(n)}$  (depends on the theoretical distribution function,  $n$  and  $\alpha$ ).

**TEST WILCOXON:** Test robust, aquestes dades són simètriques o no al voltant d'un valor central, estan centrades?? El valor central el podem decidir nosaltres: rara vegada és la mitjana aritmètica. Perquè si l'agafem estem assumint una certa normalitat en les dades. Normalment agafem la mediana o un valor que nosaltres decidim.  $N > 6$  → la mitjana és igual a la mediana? Així tindríem una distribució normal.

1. Es defineix un valor central
2. El valor central es resta a cada valor
3. Les diferències s'ordenen d'acord al seu valor absolut (de menor a major) però mantenint el signe.
4. Sumem els **RANGS** positius per una banda i els negatius per l'altra.
5.  $T_{CALC}$  = més petit del les dues  $T$

$$T_+ = \sum i_+ \quad T_- = |\sum i_-|$$

**Statistic:**

$$T_{calc} = \text{lowest } (T_+, T_-)$$

$$H_0 \quad T_+ = T_- \quad T_{calc} > T_{crit}(\alpha, n)$$

$$H_1 \quad T_+ \neq T_- \quad T_{calc} < T_{crit}(\alpha, n)$$

- Preguntem estàs centrada o no: DUES CUES simetria o no
- Si preguntes estàs esbiaixada cap a la dreta/esquerra una cua.

Una dispersió molt alta (gran DS) no vol dir població no simètrica.