Un **eşantion** este o submulţime a populaţiei. Dintr-un punct de vedere teoretic fiecare individ are aceleaşi şanse de a aparţine eşantionului, şi orice grup particular de indivizi este ales în mod independent pentru a face parte din eşantion. Dacă aceste condiţii sunt îndeplinite atunci avem un **eşantion aleator simplu**.

O **variabilă discretă** este o variabilă care are un număr finit sau infinit dar numărabil de valori; o astfel de variabilă poate avea valori corespunzând unor puncte izolate de pe un interval real.

O **variabilă contină** este o variabilă care are un număr infinit şi nenumărabil de valori; o astfel de variabilă poate avea, de obicei, orice valoare dintr-un interval real, incluzând orice valoare posibilă dintre orice două valori.

Datele sunt valorie variabilei colectate de la fiecare individ din eşantion.

O **variabilă** sau un **atribut** este o caracteristică a indivizilor din populaţie sau eşantion.

O **variabilă calitativă** (sau **categorică**) este o variabilă care descrie un individ dintr-o populaţie (conform unor categorii).

O **variabilă cantitativă** este o variabilă care măsoară un individ dintr-o populaţie.

O **populaţie** este o mulţime de obiecte (numite şi indivizi) ale căror proprietăţi vor fi analizate.

Sumarul celor cinci numere este compus din

1. min, cea mai mică valoare din eşantion;
2. Q_1, prima cvartilă;
3. Me, mediana;
4. Q_3, a treia cvartilă;
5. max, cea mai mare valoare din eşantion.

Cvartila medie este valoarea de mijloc dintre prima şi cea de-a treia cvartilă:

$$midq = \frac{Q_1 + Q_3}{2}.$$

Domeniu intercvartilic este diferenţa dintre prima şi cea de-a treia cvartilă:

$$IQR = Q_3 - Q_1.$$

O **variabilă nominală** este o variabilă care numeşte sau descrie un individ dintr-o populaţie fără a putea asigna o ordine naturală acestor valori..

O **variabilă ordinală** este o variabilă ale cărei valori pot fi ordonate în mod natural.

Deviaţia standard a eşantionului, s, este rădăcina pătrată a dispersiei eşantionului.

Dispersia eşantionului, s^2, n fiind dimensiunea eşantionului, este:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}{n-1},$$

Media populaţiei se notează cu μ.

Cvartilele sunt valori care împart domeniul (ordonat al) observaţiilor în patru segmente egale.

Median (Me) este valoarea din mijloc când datele din eşantion sunt sortate.

Mòdul este observaţia cea mai frecventă din eşantion.

Momentul central de ordin k al populaţiei este

$$\mu_k = M\left[(X-\mu)^k\right].$$

Momentul central de ordin k al eşantionului X este

$$m_k = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)^k}{n}.$$

Asimetria[a] (**skewness**) al unei distribuţii X este

$$\gamma_1 = \frac{M\left[(X-\mu)^3\right]}{\left(M\left[(X-\mu)^2\right]\right)^{3/2}}.$$

Turtirea (**kurtosis**) unei distribuţii X este

$$kurt[X] = \frac{M\left[(X-\mu)^4\right]}{\left(M\left[(X-\mu)^2\right]\right)^2}.$$

Asimetria eşantionului este $a_3 = \frac{m_3}{s^3}$.

Turtirea eşantionului este $a_4 = \frac{m_4}{s^4}$.

Estimarea cu un interval a unui parametru constă în determinarea unui interval ale cărui limite sunt statistici calculate din eşantioane.

Nivelul de încredere $(1-\alpha)$ este proporţia acelor intervale (de încredere) care conţin parametrul estimat.

Un interval de încredere este un interval care are un anumit nivelul de încredere (prescris).

Valoarea P este probabilitatea de a obţine un rezultat cel puţin la fel de neobişnuit (extrem) ca rezultatul obţinut din eşantion. Când valoarea P este "mică" putem respinge H_0.

(Legea tare numerelor mari) Fie $(X_n)_{n\geqslant 1}$ un şir de variabile aleatoare independente şi identic distribuite cu media μ şi dispersia σ^2. Atunci

$$P\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i = \mu\right) = 1.$$

Estimarea punctuală a unui parametru constă în determinarea unui număr, de obicei valoarea unei statistici corespunzătoare, desemnat sa estimeze acel parametru.

Fie θ un anumit parametru al unei populaţii, x un eşantion al acestei populaţii şi $\hat{\theta}_n = \hat{\theta}_n(x)$ un estimator punctual al lui θ.

$\hat{\theta}_n$ este o **statistică nedeplasată** dacă $M[\hat{\theta}_n] = \theta$. Altfel $\hat{\theta}_n$ este numită statistică **deplasată**.

$\hat{\theta}_n$ este o **statistică consistentă** dacă, pentru orice $\epsilon > 0$

$$\lim_{n\to\infty}P(|\hat{\theta}_n - \theta| < \epsilon) = 1.$$

Un estimator nedeplasat cu dispersie minimă, dacă există, se numeşte **statistică eficientă**.

Un interval de încredere cu nivelul de încredere $(1-\alpha)$ pentru pentru media unei populaţii cu dispersia cunoscută este

$$\left(\overline{x}_n - z^*\frac{\sigma}{\sqrt{n}}, \overline{x}_n + z^*\frac{\sigma}{\sqrt{n}}\right),$$

unde z^* este valoarea critică asociată cu $\alpha/2$. Mai mult, acest interval este exact pentru o populaţie distribuită normal şi aproximativ altfel, când eşantionul este suficient de mare ($n \geqslant 30$).

Fie $(X_n)_{n\geqslant 1}$ un şir de variabile aleatoare independente având dispersii finite, uniform mărginite, i. e. $D^2[X_n] \leqslant c$, pentru orice $n \geqslant 1$. Atunci

$$\lim_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\sum_{i=1}^{n}M[X_i]\right| < \epsilon\right) = 1.$$

(Legea slabă numerelor mari , legea lui Khintchine) Fie $(X_n)_{n\geqslant 1}$ un şir de variabile aleatoare independente şi identic distribuite cu media μ şi dispersia σ^2. Atunci

$$\lim_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right| < \epsilon\right) = 1 \text{ sau } \lim_{n\to\infty}P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right| \geqslant \epsilon\right) = 0$$

Un interval de încredere cu nivelul de încredere $(1-\alpha)$ pentru pentru media unei populaţii normale cu dispersia necunoscută este

$$\left(\overline{x}_n - t^*\frac{s}{\sqrt{n}}, \overline{x}_n + t^*\frac{s}{\sqrt{n}}\right).$$

unde t^* este valoarea critică asociată cu $\alpha/2$.

Testarea ipotezelor statistice este un proces prin care se ia o decizie între două ipoteze opuse.

Ipotezele statistice sunt formulate în aşa fel încât întotdeauna una din ele este falsă iar cealaltă adevărată.

Una dintre ipoteze este testată sperând că se poate arăta că este puţin probabil ca ea să fie adevărată, ceea ce implică faptul că cealaltă ipoteză este probabil adevărată.

Ipoteza nulă, H_0, este ipoteza status-quo-ului în ceea ce priveşte populaţia; formal este o afirmaţie legată de populaţie: spre exemplu că are o anumită medie sau dispersie, sau o anumită distribuţie etc.

Ipoteza alternativă, H_a, este ipoteza de cercetare, şi susţine un lucru diferit despre obiectul ipotezei nule.

Scorul testului este statistica ce corespunde valorii P; **valoarea critică**, pe de altă parte, coresponde nivelului de semnificaţie. Concluzia testului depinde de rezultatul comparării celor două valori.

Ecuaţia acestei drepte (linia SD) este

$$y - \overline{y}_n = m(x - \overline{x}_n),$$

unde $\overline{x}_n$ şi $\overline{y}_n$ sunt mediile.

Paşii care trebuie urmaţi când întreprindem un test de semnificaţie:

1-2. Se formulează cele două ipoteze: H_0 şi H_a: H_a va fi acceptată dacă H_0 este respinsă.
3. Se alege un nivel de semnificaţie α - cât de semnificative trebuie să fie evidenţele pentru a respinge H_0.
4. Se calculează statistica sau scorul testului.
5. Se determină valoarea critică.
6. Se compară scorul cu valoarea critică şi, dacă este cazul, se respinge H_0 şi se acceptă H_a, altfel nu se acceptă H_a.

Media şi dispersia mediei de selecţie, $\overline{x}_n$, sunt μ şi σ^2/n:

$$M[\overline{x}_n] = \mu,\ D^2[\overline{x}_n] = \frac{\sigma^2}{n}.$$

În plus, pentru valori mari ale lui n ($\geqslant 30$), distribuţia mediei de selecţie este normală, i. e.

$$\overline{x}_n \sim N(\mu, \sigma^2/n).$$

Variabila aleatoare care are drept valori toate mediile de selecţie posibile, $\overline{x}_n^{(k)}$, pentru eşantioane de dimensiune n, se numeşte **distribuţia mediei de selecţie**.

Un **test parametric** presupune că populaţia urmează o anumită distribuţie şi inferează asupra parametrilor acelei distribuţii.

Un **parametru** este o valoare numerică care priveşte întreaga populaţie.

O **statistică** este un parametru calculat pentru un eşantion în locul întregii populaţii.

(Teorema limită centrală, Lindeberg-Lévy) Fie $(X_n)_{n\geq 1}$ un şir de variabile aleatoare independenteşi identic distribuite cu media μ şi dispersia σ^2. Atunci

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1) \text{ sau}$$

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leqslant a\right) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a} \exp -t^2/2dt.$$

Dimensiunea efectului este magnitudinea diferenţei descoperite în eşantionul aleator (dacă există).

Nivelul de semnificaţie, α, este probabilitatea (condiţionată) maximă pe care ne-o asumăm drept risc de face o eroare de tipul I.

Puterea testului, este 1 minus probabilitatea de a face o eroare de tipul II.

Un interval de încredere pentru un parametru θ cu $(1-\alpha)$ nivel de încredere este definit cu ajutorul a două statistici L şi U astfel ca

$$P(L \leqslant \theta \leqslant U) = (\geqslant) 1-\alpha.$$

Un **test neparametric** numit şi **distribution-free** sau **parameter-free** se bazează pe puţine fapte - de obicei distribuţia şi parametrii săi (medie, dispersie) nu sunt cunoscuţi.

		validitatea ipotezei H_0	
		adevărată	falsă
Decizia asupra H_0	se respinge	Eroare de tip I (fals pozitiv)	Corect (adevărat pozitiv)
	nu se respinge	Corect (adevărat negativ)	Eroare de tip II (fals negativ)

Testul F - Inferenţe asupra raportului dintre dispersii

- Testul se desfăşoară astfel:

1. Formulăm *ipoteza nulă*:
$$\boxed{H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1}$$
2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantion. Putem avea două[1] tipuri de ipoteză alternativă
$\boxed{H_a: \frac{\sigma_1}{\sigma_2} > 1}$ (*asimetrică la dreapta*) pentru un *un test one-tail*
$\boxed{H_a: \frac{\sigma_1}{\sigma_2} \neq 1}$ (*simetrică*) pentru un *un test two-tailed*.
3. Alegem nivelul de semnificaţie $\alpha \in \{1\%, 5\%\}$.
4. Calculăm *scorul F* (*statistica* testului)
$$\boxed{F = \frac{s_1^2}{s_2^2}}$$
5. Determinăm valoarea critică corespunzătoare pentru α
$\boxed{F^* = qf(1-\alpha, n_1 - 1, n_2 - 1)}$ pentru H_a asimetrică la dreapta
$\boxed{F_s^* = qf(\alpha/2, n_1 - 1, n_2 - 1), F_d^* = qf(1-\alpha/2, n_1 - 1, n_2 - 1)}$ pentru H_a simetrică.
6. Comparăm valoarea critică cu scorul F; dacă scorul F aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:
$[F^*, +\infty)$ pentru H_a asimetrică la dreapta,
$(0, F_s^*] \cup [F_d^*, +\infty)$ pentru H_a simetrică.
Dacă scorul F nu aparţine zonei de respingere spunem că *nu există suficiente dovezi cu nivelul de semnificaţie α* pentru a respinge ipoteza nulă (*încercarea de a respinge H_0 eşuează*).

Testul T - inferenţe asupra mediilor a două populaţii (σ_1^2, σ_2^2 necunoscute)

Testul decurge astfel:

1. Formulăm *ipoteza nulă*, care susţine că diferenţa mediilor celor două populaţii ia o anumită valoare:
$$\boxed{H_0: \mu_1 - \mu_2 = m_0}$$
2. Formulăm *ipoteza alternativă* conform datelor din eşantioane. Putem avea trei tipuri de ipoteză alternativă
$\boxed{H_a: \mu_1 - \mu_2 < m_0}$ (*asimetrică la stânga*) sau
$\boxed{H_a: \mu_1 - \mu_2 > m_0}$ (*asimetrică la dreapta*) sau
$\boxed{H_a: \mu_1 - \mu_2 \neq m_0}$ (*simetrică*).
Ipotezele asimetrice se mai numesc *one-tailed*, iar cea simetrică *two-tailed*.
3. Alegem nivelul de semnificaţie $\alpha \in \{1\%, 5\%\}$.
4. Calculăm *scorul t* (*statistica* testului)
   a) dacă dispersiile sunt egale:
$$\boxed{t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}}$$
unde $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, iar numărul de grade de libertate este $df = n_1 + n_2 - 2$.
   b) dacă dispersiile sunt diferite:
$$\boxed{t = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}}$$
numărul de grade de libertate fiind $df = \min(n_1 - 1, n_2 - 1)$.
5. Determinăm valoarea critică corespunzătoare
$\boxed{t^* = qt(\alpha, df)}$ pentru H_a asimetrică la stânga ($t^* < 0$),
$\boxed{t^* = qt(1-\alpha, df)}$ pentru H_a asimetrică la dreapta ($t^* > 0$),
$\boxed{t^* = -qt(\alpha/2, df) = qt(1-\alpha/2, df)}$ pt. H_a simetrică ($t^* > 0$).
6. Comparăm valoarea critică cu scorul z; dacă scorul z aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:
$(-\infty, t^*]$ pentru H_a asimetrică la stânga,
$[t^*, +\infty)$ pentru H_a asimetrică la dreapta,
$(-\infty, -|t^*|] \cup [|t^*|, +\infty)$ pentru H_a simetrică.

- Putem utiliza testul Z chiar dacă cele două populaţii sunt doar aproximativ normal distribuite, dacă cele două eşantioane sunt suficient de mari ($n_1, n_2 \geqslant 30$).
- Testul se desfăşoară astfel:

1. Formulăm *ipoteza nulă*, care susţine că diferenţa mediilor celor două populaţii ia o valoare fixată:
$$\boxed{H_0: \mu_1 - \mu_2 = m_0}$$
2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantioane.
Putem avea trei tipuri de ipoteză alternativă
$\boxed{H_a: \mu_1 - \mu_2 < m_0}$ (*asimetrică la stânga*) sau
$\boxed{H_a: \mu_1 - \mu_2 > m_0}$ (*asimetrică la dreapta*) sau
$\boxed{H_a: \mu_1 - \mu_2 \neq m_0}$ (*simetrică*).
Ipotezele asimetrice se mai numesc *one-tailed*, iar cea simetrică *two-tailed*.
3. Alegem un nivel de semnificaţie $\alpha \in \{1\%, 5\%\}$.
4. Calculăm *scorul z* (*statistica* testului)
$$\boxed{z = \frac{(\overline{x}_{n_1} - \overline{x}_{n_2}) - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}$$
5. Determinăm valoarea critică corespunzătoare lui α
$\boxed{z^* = qnorm(\alpha)}$ pentru H_a asimetrică la stânga ($z^* < 0$),
$\boxed{z^* = qnorm(1-\alpha)}$ pentru H_a asimetrică la dreapta ($z^* > 0$),
$\boxed{z^* = -qnorm(\alpha/2) = qnorm(1-\alpha/2)}$ pt. H_a simetrică ($z^* > 0$).
6. Comparăm valoarea critică cu scorul z; dacă scorul z aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:
$(-\infty, z^*]$ pentru H_a asimetrică la stânga,
$[z^*, +\infty)$ pentru H_a asimetrică la dreapta,
$(-\infty, -|z^*|] \cup [|z^*|, +\infty)$ pentru H_a simetrică.
Dacă scorul z nu aparţine zonei de respingere spunem că *nu există suficiente dovezi cu nivelul de semnificaţie α* pentru a respinge ipoteza nulă (*încercarea de a respinge H_0 eşuează*).

Testul T - Inferenţă asupra mediei unei populaţii (σ necunoscută)

1. Formulăm mai întâi *ipoteza nulă*, care susţine că media populaţiei ia o anumită valoare:
$$\boxed{H_0: \mu = \mu_0}$$
2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantion. Putem avea trei tipuri de ipoteză alternativă
$\boxed{H_a: \mu < \mu_0}$ (*asimetrică la stânga*) sau
$\boxed{H_a: \mu > \mu_0}$ (*asimetrică la dreapta*) sau
$\boxed{H_a: \mu \neq \mu_0}$ (*simetrică*).
Ipotezele asimetrice se mai numesc *one-tailed*, iar cea simetrică *two-tailed*.
3. Alegem un nivel de semnificaţie $\alpha \in \{1\%, 5\%\}$.
4. Calculăm *scorul t* (*statistica* testului)
$$\boxed{t = \frac{\overline{x}_n - \mu_0}{s/\sqrt{n}}}$$
5. Determinăm valoarea critică corespunzătoare lui α
$\boxed{t^* = qt(\alpha, n-1)}$ pentru H_a asimetrică la stânga ($t^* < 0$),
$\boxed{t^* = qt(1-\alpha, n-1)}$ pentru H_a asimetrică la dreapta ($t^* > 0$),
$\boxed{t^* = -qt(\alpha/2, n-1) = qt(1-\alpha/2, n-1)}$ pt. H_a simetrică ($t^* > 0$
6. Comparăm valoarea critică cu scorul t; dacă scorul t aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:
$(-\infty, t^*]$ pentru H_a asimetrică la stânga,
$[t^*, +\infty)$ pentru H_a asimetrică la dreapta,
$(-\infty, -|t^*|] \cup [|t^*|, +\infty)$ pentru H_a simetrică.
Dacă scorul t nu aparţine zonei de respingere vom spune că *nu există suficiente dovezi cu nivelul de semnificaţie α* pentru a respinge ipoteza nulă (*încercarea de a respinge H_0 eşuează*).

Testul Z - Inferenţă asupra mediei unei populaţii (σ cunoscută)

1. Formulăm *ipoteza nulă*, care susţine că media populaţiei ia o anumită valoare:
$$\boxed{H_0: \mu = \mu_0}$$
2. Formulăm *ipoteza alternativă* conform informaţiilor obţinute din eşantion. Putem avea trei tipuri de ipoteză alternativă
$\boxed{H_a: \mu < \mu_0}$ (*asimetrică la stânga*) sau
$\boxed{H_a: \mu > \mu_0}$ (*asimetrică la dreapta*) sau
$\boxed{H_a: \mu \neq \mu_0}$ (*simetrică*).
Ipotezele asimetrice se mai numesc *one-tailed*, iar cea simetrică *two-tailed*.
3. Alegem un nivel de semnificaţie $\alpha \in \{1\%, 5\%\}$.
4. Calculăm *scorul z* (*statistica* testului)
$$\boxed{z = \frac{\overline{x}_n - \mu_0}{\sigma/\sqrt{n}}}$$
5. Determinăm valoarea critică corespunzătoare lui α
$\boxed{z^* = qnorm(\alpha)}$ pentru H_a asimetrică la stânga ($z^* < 0$),
$\boxed{z^* = qnorm(1-\alpha)}$ pentru H_a asimetrică la dreapta ($z^* > 0$),
$\boxed{z^* = -qnorm(\alpha/2) = qnorm(1-\alpha/2)}$ pentru H_a simetrică ($z^* >$
6. Comparăm valoarea critică cu scorul z; dacă scorul z aparţine *zonei de respingere*, atunci acceptăm H_a şi respingem H_0. Zonele de respingere sunt:
$(-\infty, z^*]$ pentru H_a asimetrică la stânga,
$[z^*, +\infty)$ pentru H_a asimetrică la dreapta,
$(-\infty, -|z^*|] \cup [|z^*|, +\infty)$ pentru H_a simetrică.
Dacă scorul z nu aparţine zonei de respingere spunem că *nu există suficiente dovezi cu nivelul de semnificaţie α* pentru a respinge ipoteza nulă (*încercarea de a respinge H_0 eşuează*).