UNIVERSITATEA DE MEDICINĂ,
FARMACIE, ȘTIINȚE ȘI TEHNOLOGIE
„GEORGE EMIL PALADE"
DIN TÂRGU MUREȘ

PROBABILITĂȚI ȘI STATISTICĂ ÎN SISTEME MEDICALE

Cursul 10, 22.10.2020

INFERENȚA STATISTICĂ PENTRU DATE DIN SISTEME MEDICALE

prof. univ. dr. habil Manuela Rozalia GABOR

STRUCTURA CURSULUI

1. Testarea ipotezelor statistice – Curs 9
2. Alegerea unui test statistic – Curs 9
3. **Normalitatea datelor – Curs 10**
4. Compararea variabiellor cantitative – Curs 11
5. Compararea distribuților variabilelor calitative – Curs 12

1. Utilitatea evaluării normalității datelor

Metodele statistice care utilizează inferența (teste statistice, intervale de încredere, ...), se pot folosi dacă sunt îndeplinite anumite condiții. Întotdeauna înainte de utilizarea unui test statistic sau pentru orice analiză statistică, trebuie căutate care sunt condițiile de aplicare (în engleză: assumptions). O astfel de condiție este normalitatea datelor cantitative.

Condițiile trebuie verificate și în caz că sunt satisfăcute, testele se pot utiliza. Dacă nu sunt satisfăcute condițiile de aplicare, se caută alte metode statistice care nu necesită condiții atât de stricte.

În cazul datelor cantitative, una din condițiile de aplicare pentru multe teste statistice este normalitatea datelor. În statistică se spune că datele sunt normal distribuite, sau că urmează o distribuție normală dacă distribuția de probabilitate a unui set de date este asemănătoare cu distribuția normală (gausiană – curba lui Gauss).

În caz contrar, spunem că datele nu sunt normal distribuite, sau că nu urmează o distribuție normală. În domeniul medical, dentar, biologic, date care nu au o distribuție normală sunt frecvente. Faptul că datele nu urmează o distribuție normală nu e ceva negativ, ci, pur și simplu, trebuie ținut cont de acest aspect când descriem și analizăm aceste date.

2. Evaluarea normalității datelor

Literatura de specielitate descrie diverse metode de a realiza evaluarea normalității datelor, fiecare cu avantaje și dezavantaje. Nu există un consens între statisticieni care e varianta perfectă pentru a evalua normalitatea.

Ceea ce interesează nu este de fapt distribuția datelor din eșantion, ci distribuția datelor din populația din care a fost extras eșantionul. Întrucât nu știm în mod uzual distribuția datelor în populație, frecvent evaluăm normalitatea datelor pentru variabilele din eșantion și presupunem că în populație distribuția e asemănătoare (evident această variantă este supusă riscului de eroare).

Niciuna dintre metode nu garantează că vom afla corect dacă distribuția reală în populație e normală sau nu. Din distribuții perfect normale în populație putem extrage eșantioane a căror distribuție sugerează deviere marcată de la normalitate și viceversa.

Pentru evaluarea normalității se pot folosi metode grafice (cele mai bune), metode utilizând indicatori de statistică descriptivă (mai puțin bune), respectiv teste statistice care evaluează normalitatea (considerate de unii statisticieni puțin folositoare).

2.1. Evaluarea grafică

Grafic putem evalua normalitatea cu ajutorul histogramelor, graficelor cutie cu mustăți, graficelor de quantile, dar și prin alte diagrame.

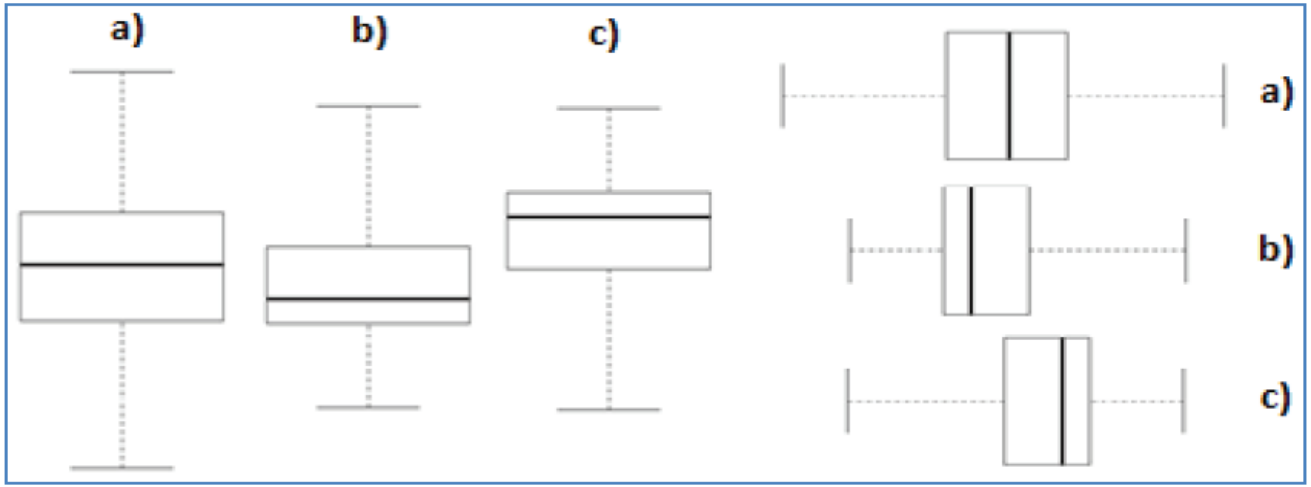

Fig. 6.1 Evaluarea normalității datelor utilizând graficul cutie cu mustăți. a) sugerează date normal distribuite, b) și c) sugerează date care nu urmează o distribuție normală, b) sugerează asimetrie la dreapta, c) sugerează asimetrie la stânga.

2.1. Evaluarea grafică - continuare

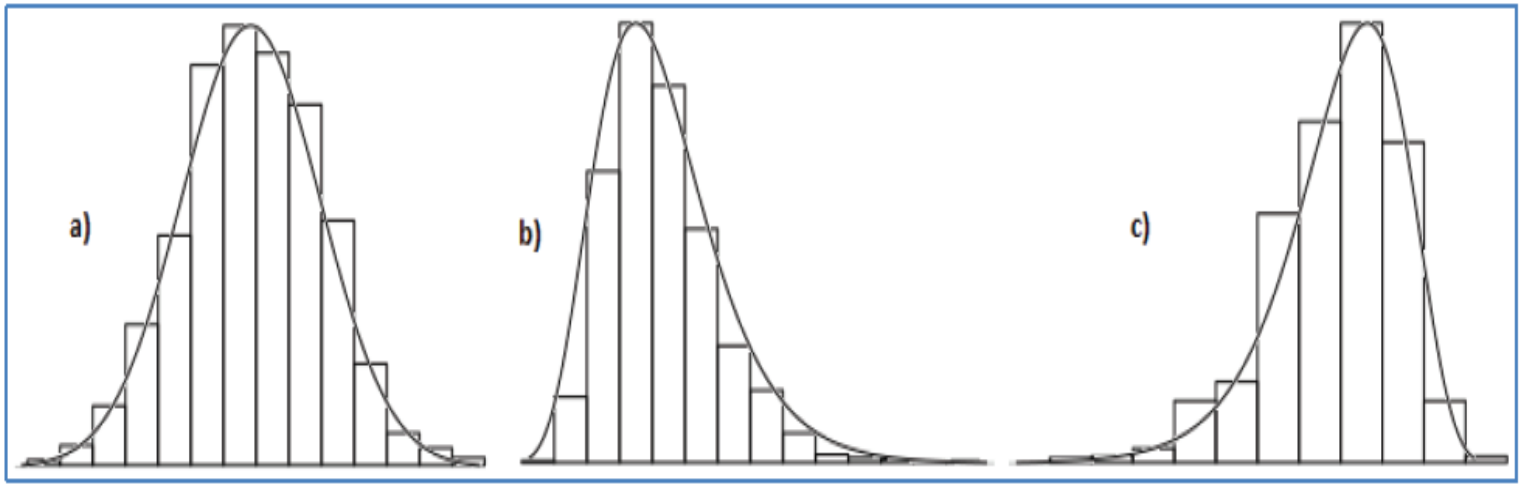

Fig. 6.2 Evaluarea normalității datelor utilizând histograma. a) sugerează date normal distribuite, b) și c) sugerează date care nu urmează o distribuție normală, b) sugerază asimetrie la dreapta, c) sugerează asimetrie la stânga.

2.1. Evaluarea grafică - continuare

Graficul cuantilă-cuantilă, permite cel mai bine evaluarea normalității întrucât observăm toate valorile din seria de date (reprezentate prin puncte sau cercuri). Linia diagonală reprezintă o distribuție normală. Cu cât punctele sunt mai apropiate de această linie, putem presupune o distribuție normală (figura 6.3.a), respectiv o tendință a punctelor de a se îndepărta de linia diagonală sugerează absența unei distribuții normale (figura 6.3.b).

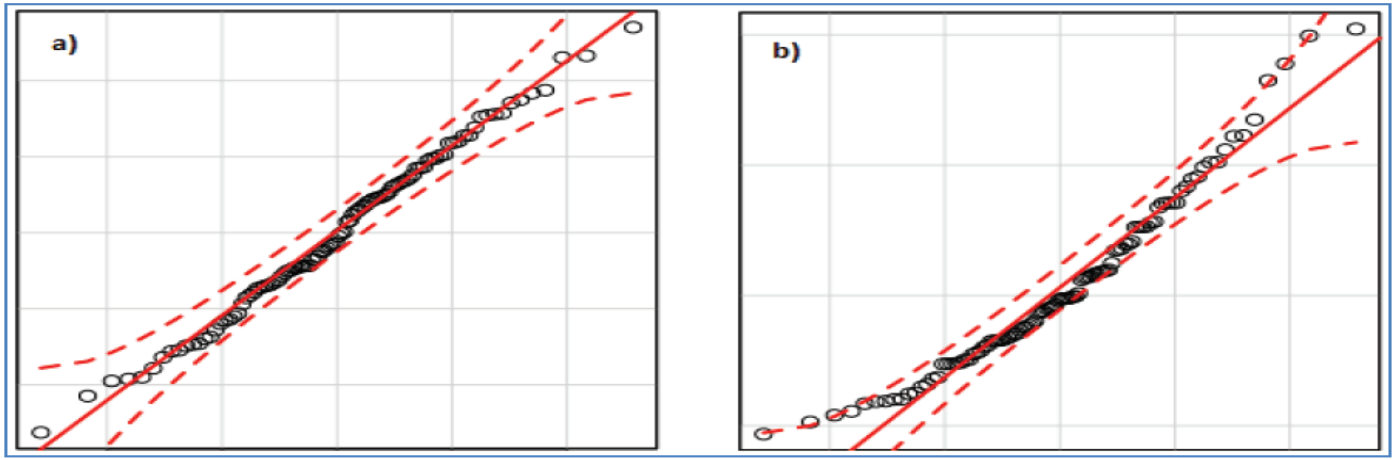

Fig. 6.3 Evaluarea normalității datelor utilizând graficul cuantilă-cuantilă. a) date normal distribuite, b) date care nu urmează o distribuție normală

2.2. Evaluarea prin parametrii de statistică descriptivă

Evaluarea normalității se poate face cu **indicatori de statistică descriptivă**, precum media, modulul, mediana, coeficientul de asimetrie (în engleză skewness) și coeficientul de

boltire (în engleză kurtosis).

Media, **modulul** și **mediana** au valori apropiate, sau chiar identice în cazul distribuției normale. E greu de evaluat ce este apropiat sau nu, întucât depinde de caracteristica studiată, respectiv există întotdeauna și o doză de subiectivism. O metodă simplă care ne sugerează probleme de normalitate a datelor este compararea mediei cu deviațiastandard: în general media este mai mare decât deviația standard în eșantioanele normal distribuite

Coeficientul de asimetrie apropiat de 0 sugerează o distribuție simetrică. Vom considera o valoare în intervalul (-1, 1) ca sugestie de normalitate O valoare în afara intervalului o vom considera ca absență a unei distribuții normale.

Excesul de boltire apropiat de 0 sugerează o distribuție având o formă a curbei similară distribuției normale. Vom considera o valoare în intervalul (-1, 1) ca sugestie de normalitate. O valoare în afara intervalului o vom considera ca absență a unei distribuții normale.

2.3. Evaluare prin teste statistice

O altă metodă de a evalua normalitatea este utilizarea **testelor statistice pentru evaluarea normalității**. Există numeroase teste specifice pentru normalitate:

- Shapiro-Wilk (folosit mai ales pentru eșantioane având sub 50 de subiecți),
- Kolmogorov Smirnov,
- D' Agostino-Pearson și
- testul Chi-pătrat poate fi folosit în anumite condiții pentru a o testa.

Testele specifice compară distribuția observată cu una teoretică normal distribuită.

Interpretarea pentru toate aceste teste se face similar. Ipoteza nulă spune că distribuția este normală, ipoteza alternativă spune că distribuția nu este normal distribuită (de fapt, corect este să spunem că ipoteza nulă reprezintă situația în care nu există diferență statistic semnificativă între distribuția datelor și distribuția normală, repectiv ipoteza alternativă reprezintă situația în care există diferență statistic semnificativă între distribuția datelor și distribuția normală). Dacă valoarea lui p a unui test pentru evaluarea normalității este sub 0,05, atunci se consideră că datele nu urmează o distribuție normală.

3. Teste parametrice/neparametrice

Printre testele statistice cele mai folosite în domeniul medical referitor la comparații între grupuri, pentru date de tip cantitativ, alegerea între teste parametrice și neparametrice echivalente se face în funcție de prezența normalității datelor. Astfel, dacă datele sunt normal distribuite se va merge în principal pe teste parametrice, în caz contrar se vor utiliza teste neparametrice echivalente

Tabel 6.1. Alegerea între teste parametrice și neparametrice echivalente în funcție de normalitatea datelor, pentru comparații între grupuri, pentru variabile cantitative.

Teste parametrice	Teste neparametrice echivalente
Student (t) pentru eșantioane independente	Mann–Whitney U Mann–Whitney–Wilcoxon Wilcoxon–Mann–Whitney Wilcoxon rank-sum Wilcoxon pentru eșantioane independente
Student (t) pentru eșantioane pereche	Wilcoxon signed-rank Wilcoxon pentru eșantioane dependente
ANOVA	Kruskal-Wallis