STATISTICA SOCIALE
Corso di laurea in Scienze Turistiche, a.a. 2007/2008
Esercizi 16 novembre2007

Esercizio 1

La Tabella 1 contiene alcuni dati relativi a 16 lavoratori delle aziende Alfa e Beta.

Tabella 1 – Lavoratori delle aziende Alfa e Beta

Lavoratore	Azienda	Posizione	Stipendio (euro)	Precedenti occupazioni
1	Alfa	Dirigente	5030	3
2	Alfa	Impiegato	1020	0
3	Alfa	Impiegato	1050	1
4	Alfa	Quadro	1900	0
5	Alfa	Impiegato	1070	2
6	Alfa	Impiegato	950	1
7	Alfa	Quadro	2100	2
8	Alfa	Quadro	2200	2
9	Alfa	Impiegato	1050	1
10	Beta	Impiegato	1020	3
11	Beta	Impiegato	960	0
12	Beta	Quadro	1950	2
13	Beta	Dirigente	4800	2
14	Beta	Impiegato	920	1
15	Beta	Impiegato	1010	2
16	Beta	Quadro	1800	0

a) Calcolare un indice di variabilità adeguato per il carattere *Posizione* e per il carattere *Precedenti Occupazioni*.
b) Relativamente al carattere *Azienda*
   - Costruire la distribuzione di frequenza relativa osservata
   - calcolare l'indice di entropia;
   - costruire la distribuzione relativa di massima eterogeneità
   - costruire una delle distribuzioni relative di minima eterogeneità.
c) Relativamente al carattere *Stipendio*
   - Calcolare campo di variazione e scarto interquartile
   - Disegnare il boxplot
   - Calcolare varianza, deviazione standard e coefficiente di variazione
   - Indicare se la distribuzione è simmetrica. In caso di asimmetria, indicare il tipo di asimmetria (senza fare calcoli).

Esercizio 2

In Tabella 2 è riportata la matrice dei dati relativa a 20 film.

Tabella 2 – Dati relativi a 20 film

REPLICHE_TV	NAZIONE	DURATA (in min.)	GIUDIZIO
0	ITA	105	SUFF
0	ITA	114	SUFF
0	ITA	110	SUFF
0	ITA	89	SUFF
0	MES	144	OTTIMO
0	ITA	100	BUONO
0	SPA	100	SUFF
0	ITA	101	SUFF
1	ITA	89	BUONO
1	ITA	97	SUFF
1	ITA	115	OTTIMO
1	USA	90	SUFF
1	USA	93	SUFF
1	USA	91	INSUFF
2	ITA	90	SUFF
2	USA	90	SUFF
2	ITA	90	SUFF
2	USA	100	INSUFF
2	USA	100	SUFF
2	USA	100	INSUFF

(Nota: il carattere REPLICHE_TV è il numero di volte in cui il film è stato trasmesso da una certa emittente televisiva).

a) Calcolare un indice di variabilità adeguato per il carattere NAZIONE;
b) Calcolare un indice di variabilità adeguato per il carattere GIUDIZIO;
c) Calcolare campo di variazione e scarto interquartile del carattere DURATA;
d) Disegnare il BOX-PLOT relativo alla variabile DURATA;
e) Calcolare il coefficiente di variazione per i caratteri REPLICHE_TV e DURATA. Quale delle due variabili presenta la variabilità più elevata?

Esercizio 3

Rispondere brevemente ai seguenti quesiti:

a) Supponiamo che una distribuzione espressa in euro venga trasformata in centesimi di euro: indicare come si trasformano la varianza, lo scarto quadratico medio e il coefficiente di variazione.
b) Come si caratterizza la situazione di massima eterogeneità?
c) Quali sono i valori minimo e massimo assumibili dalla varianza?
d) Se l'unità di misura di un carattere passa da metri a centimetri, come cambiano i seguenti indici: varianza, scarto quadratico medio, coefficiente di variazione.
e) Come si definisce e per quali tipi di carattere si calcola l'indice di entropia?

Soluzione

Esercizio 1

a) Calcolare un indice di variabilità adeguato per il carattere *Posizione* e per il carattere *Precedenti Occupazioni*.

Posizione è una variabile qualitativa ordinale, possiamo valutare l'eterogeneità della distribuzione utilizzando l'indice di Entropia. Di seguito si riporta la tabella contenente i conti necessari a costruire l'indice di Shannon H e l'indice normalizzato H', calcolato utilizzando i logaritmi naturali delle frequenze relative ln(fj)[1].

posizione	*nj*	*fj*	*ln(fj)*	*fj ln(fj)*
Impiegato	9	0.5625	-0.57536	-0.32364
Quadro	5	0.3125	-1.16315	-0.36348
Dirigente	2	0.125	-2.07944	-0.25993
Totale	16	1		-0.94706
H				0.947057
ln(3)				1.098612
H norm				0.862049

$$\sum_{j=1}^{3} f_j \ln f_j = 0.947057$$

$$H' = H / \ln 3 = 0.862049$$

Precedenti Occupazioni è una variabile quantitativa discreta e possiamo utilizzare la deviazione standard per misurarne la variabilità.

precedenti occupazioni	nj	fj
0	4	0.250
1	4	0.250
2	6	0.375
3	2	0.125
TOTALE	16	1.000

- Media= $\sum_{j=1}^{4} x_j f_j = 1.4$
- Varianza $\sum x_j^2 f_j - (1.4)^2 = 1$
- Deviazione standard= $\sqrt{\sum x_j^2 f_j - (1.4)^2} = 1$

[1] In generale si può utilizzare il logaritmo in una base qualsiasi.(si veda http://it.wikipedia.org/wiki/Logaritmo). Una breve spiegazione su cos'è il logaritmo naturale di un numero si può trovare su WIKIPEDIA al seguente indirizzo WEB: http://it.wikipedia.org/wiki/Logaritmo_naturale. In Appendice del libro di Pasetti sono spiegati i logaritmi.

b) Relativamente al carattere *Azienda*
- costruire la distribuzione di frequenza relativa osservata
- calcolare l'indice di entropia

AZIENDA	*nj*	*fj*	*ln(fj)*	*fj ln(fj)*
alfa	9	0.5625	-0.57536	-0.32364
beta	7	0.4375	-0.82668	-0.36167
TOTALE	16	1		-0.68531
H				0.685314
ln(2)				0.693147
H norm				0.988699

$$\sum_{j=1}^{4} f_j \ln f_j = 0.685314$$

$$H' = H / \ln 2 = 0.988699$$

- costruire la distribuzione relativa di massima eterogeneità
- costruire una delle distribuzioni relative di minima eterogeneità.

massima eterogeneità

Azienda	*n. dipendenti*	f_j
Alfa	8	0.5
Beta	8	0.5
Totale	16	1.0

minima eterogeneità

Azienda	*n. dipendenti*	f_j
Alfa	16	1.0
Beta	0	0.0
Totale	16	1.0

c) per il carattere *Stipendio*

campo di variazione
- x_{max}-x_{min}=5030-920 =4110

scarto interquartile: Q3-Q1
- Ordinare i valori di stipendio dal più piccolo al più grande

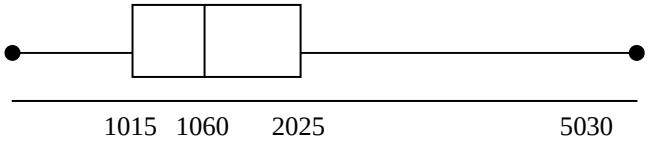
Obs	Stipendio	Lavoratore
1	920	14
2	950	6
3	960	11
4	1010	15
5	1020	2
6	1020	10
7	1050	3
8	1050	9
9	1070	5
10	1800	16
11	1900	4
12	1950	12
13	2100	7
14	2200	8
15	4800	13
16	5030	1

- Trovare la posizione occupata da Q1: 16*0.25=4
- Q1 in [x_4;x_5] ➔ Q1 in [1010;1020], convenzionalmente Q1=(1010+1020)/2=1015
- Trovare la posizione occupata da Q3: 16*0.75=12
- Q3 in [x_{12};x_{13}] ➔ Q3 in [1950;2100], convenzionalmente Q3=(1950+2100)/2=2025
- Scarto interquartile: Q3-Q1=2025-1015=**1010**

Disegnare il boxplot

Per disegnare il box-plot occorre calcolare, oltre a Q1 e Q3, la mediana e trovare il valore massimo e minimo assunto da X

- posizione mediana 16*0.5=8, Me in $[x_8;x_9]$➔ Me in [1050;1070], convenzionalmente Me=(1050+1070)/2=1060
- x_{min}=920, x_{max}=5030



b) Calcolare il coefficiente di variazione

Per calcolare il coefficiente di variazione bisogna calcolare media e deviazione standard

- Media=$\frac{1}{16}\sum x_i = 1801.9$
- Deviazione standard=$\sqrt{\frac{1}{16}\sum x_i^2 - \overline{x}^2}$ =1260.4
- CV=$100\times\frac{\sigma}{\overline{x}} = 100\times\frac{1260.4}{1801.9}$ =69.9

c) Indicare se la distribuzione è simmetrica. In caso di asimmetria, indicare il tipo di asimmetria (senza fare calcoli).

La distribuzione non è simmetrica: poiché la media aritmetica è più grande della mediana, si tratta di asimmetria positiva: addensamento delle frequenze nei valori più bassi di X.

Esercizio 2

a) Calcolare un indice di variabilità adeguato per il carattere NAZIONE

NAZIONE è un carattere qualitativo sconnesso. Come indice di variabilità possiamo calcolare l'indice di entropia di Shannon, che misura il grado di eterogeneità della distribuzione.

NAZIONE	Freq	fj	ln(fj)
ITA	11	0.55	-0.59784
MES	1	0.05	-2.99573
SPA	1	0.05	-2.99573
USA	7	0.35	-1.04982

$$H = -\sum_{j=1}^{4} f_j \ln f_j = 0.99582$$

$$H' = H / \ln 4 = 0.718333$$

b) Calcolare un indice di variabilità adeguato per il carattere GIUDIZIO;

GIUDIZIO è un carattere qualitativo ordinale. Come indice di variabilità possiamo calcolare l'indice di entropia di Shannon, che misura il grado di eterogeneità della distribuzione.

GIUDIZIO	Freq	fj	ln(fj)
BUONO	2	0.10	-2.30259
INSUFF	3	0.15	-1.89712
OTTIMO	2	0.10	-2.30259
SUFF	13	0.65	-0.43078

$$H = -\sum_{j=1}^{4} f_j \ln f_j = 1.02509$$

$$H' = H / \ln 4 = 0.739449$$

c) *Calcolare campo di variazione e scarto interquartile del carattere DURATA*

Per calcolare il coefficiente di variazione bisogna calcolare media e deviazione standard

- Media=$\frac{1}{20}\sum x_i = 100.4$
- Deviazione standard=$\sqrt{\frac{1}{20}\sum x_i^2 - \overline{x}^2}$ =12.8
- CV=$100 \times \frac{\sigma}{\overline{x}} = 100 \times \frac{12.8}{100.4}$ =12.7

Calcolo dello scarto interquartile

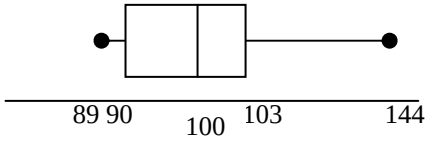
- Ordinare i valori di DURATA dal più piccolo al più grande

Obs	DURATA
1	89
2	89
3	90
4	90
5	90
6	90
7	91
8	93
9	97
10	100
11	100
12	100
13	100
14	100
15	101
16	105
17	110
18	114
19	115
20	144

- Trovare la posizione occupata da Q1: 20*0.25=5
- Q1 in $[x_5;x_6]$ ➔ Q1 in [90;90], Q1=90
- Trovare la posizione occupata da Q3: 20*0.75=15
- Q3 in $[x_{15};x_{16}]$ ➔ Q3 in [101;105], convenzionalmente Q3=(101+105)/2=103
- Scarto interquartile: Q3-Q1=103-90=**13**

d) *Disegnare il BOX-PLOT relativo alla variabile DURATA*
Per disegnare il box-plot occorre calcolare, oltre a Q1 e Q3, la mediana e trovare il valore massimo e minimo assunto da X

- posizione mediana 20*0.5=10, Me in $[x_{10};x_{11}]$➔ Me in [100;100], Me=100
- x_{min}=89, x_{max}=144



e) Calcolare il coefficiente di variazione del carattere REPLICHE_TV e DURATA. Quale delle due variabili presenta la variabilità più elevata?

Per calcolare il coefficiente di variazione bisogna calcolare media e deviazione standard.
REPLICHE_TV

- Media=$\frac{1}{20}\sum x_i = 0.9$
- Deviazione standard=$\sqrt{\frac{1}{20}\sum x_i^2 - \overline{x}^2}$ =0.8
- CV=$100 \times \frac{\sigma}{\overline{x}} = 100 \times \frac{0.8}{0.9}$ =92.3

DURATA

- Media=$\frac{1}{20}\sum x_i = 100.4$
- Deviazione standard=$\sqrt{\frac{1}{20}\sum x_i^2 - \overline{x}^2}$ =12.8
- CV=$100 \times \frac{\sigma}{\overline{x}} = 100 \times \frac{12.8}{100.4}$ =12.7

REPLICHE_TV è più variabile di DURATA perché il suo coefficiente di variazione è più elevato.

Esercizio 3

a) Supponiamo che una distribuzione espressa in euro venga trasformata in centesimi di euro: indicare come si trasformano la varianza, lo scarto quadratico medio e il coefficiente di variazione

Se Y=100*X ➔ Var(Y)=100^2*Var(X)=10000*Var(X), σ(Y)=100*σ(X), CV(Y)=CV(X)

b) Come si caratterizza la situazione di massima eterogeneità?

C'è massima eterogeneità quando le frequenze sono uniformemente distribuite tra le modalità del carattere.

c) Quali sono i valori minimo e massimo assumibili dalla varianza?

Il valore minimo è pari a 0, mentre il massimo è infinito!

d) Se l'unità di misura di un carattere passa da metri a centimetri, come cambiano i seguenti indici: varianza, scarto quadratico medio, coefficiente di variazione

Se Y=100*X ➔ Var(Y)=100^2*Var(X)=10000*Var(X), σ(Y)=100*σ(X), CV(Y)=CV(X)

e) Come si definisce e per quali tipi di carattere si calcola l'indice di entropia?

L'indice di entropia è un indice di eterogeneità e si ottiene come media ponderata del contenuto di informazione di ciascuna delle modalità osservate con pesi pari a f_j, dove per contenuto di informazione si intende $\log(1/f_j)=-\log f_j$. Questo indice si può calcolare per tutti i tipi di carattere.