Statistica Applicata
Corso di Laurea in Scienze Naturali
a. a. 2017/2018

prof. Federico Plazzi

5 Settembre 2018

Nome: _____

Cognome: _____

Matricola: _____

Alcune indicazioni:

- La prova è constituita da cinque esercizi; dopo ogni esercizio c'è lo spazio in cui scrivere la risposta o le risposte. In caso questo spazio non sia sufficiente, si può continuare a rispondere sul retro del foglio, avendo cura di indicare il numero dell'esercizio a fianco della continuazione della risposta.
- Alcuni esercizi richiedono semplici calcoli, per i quali è consentito l'uso di una calcolatrice ed eventualmente la consultazione di una o più delle tabelle allegate.
- Altri esercizi richiedono invece la lettura dei dati: verrà valutata in questo caso l'argomentazione che giustifica l'interpretazione fornita.
- La durata massima della prova è di 60 minuti.
- Si prega di non scrivere nulla sulle tabelle allegate.

1 Dati

Di seguito (tabella 1) sono elencati i principali dati demografici delle 12 prefetture più popolose del Giappone al 2008[1]. Per confronto, la tabella successiva (tabella 2) riporta gli stessi dati per le 12 regioni più popolose d'Italia al 2017[2].

Tabella 1: Giappone: prime 12 prefetture per popolazione.

Prefettura	Capoluogo	Popolazione (ab.)	Superficie (km^2)	Densità (ab./km^2)	Comuni
Tokyo	Shinjuku	12577000	2188	5788	49
Ōsaka	Ōsaka	8817000	1898	4647	44
Kanagawa	Yokohama	8792000	2416	3655	37
Aichi	Nagoya	7255000	5165	1415	88
Saitama	Saitama	7054000	3797	1862	90
Chiba	Chiba	6056000	5156	1178	80
Hokkaido	Sapporo	5628000	83456	71	222
Hyogo	Kōbe	5591000	8396	666	60
Fukuoka	Fukuoka	5050000	4977	1016	91
Shizuoka	Shizuoka	3792000	7780	488	74
Ibaraki	Mito	2975000	6096	488	61
Hiroshima	Hiroshima	2877000	8479	339	37

Tabella 2: Italia: prime 12 regioni per popolazione.

Regione	Capoluogo	Popolazione (ab.)	Superficie (km^2)	Densità (ab./km^2)	Comuni
Lombardia	Milano	10023876	23861	419	1516
Lazio	Roma	5897723	17236	341	378
Campania	Napoli	5829936	13590	429	550
Sicilia	Palermo	5039041	25711	197	390
Veneto	Venezia	4903445	18399	267	571
Emilia-Romagna	Bologna	4446567	22446	198	331
Piemonte	Torino	4380502	25402	174	1197
Puglia	Bari	4055213	19358	209	258
Toscana	Firenze	3739759	22994	163	274
Calabria	Catanzaro	1960101	15081	130	405
Sardegna	Cagliari	1650058	24100	69	377
Liguria	Genova	1560090	5422	292	234

[1] Fonte: https://web.archive.org/web/20081207140448/http://www.stat.go.jp/english/data/handbook/pdf/ap_1.pdf.
[2] Fonte: http://demo.istat.it/bilmens2017gen/index.html.

2 Esercizi

2.1 Statistiche di base

Calcolare media, devianza, varianza e deviazione standard della densità delle 12 regioni italiane mostrate in tabella 2.

Media	*Devianza*	*Varianza*	*Deviazione standard*
$240,6667$	$137590,7$	$11465,89$	$107,0789$

2.2 Distribuzione dei dati

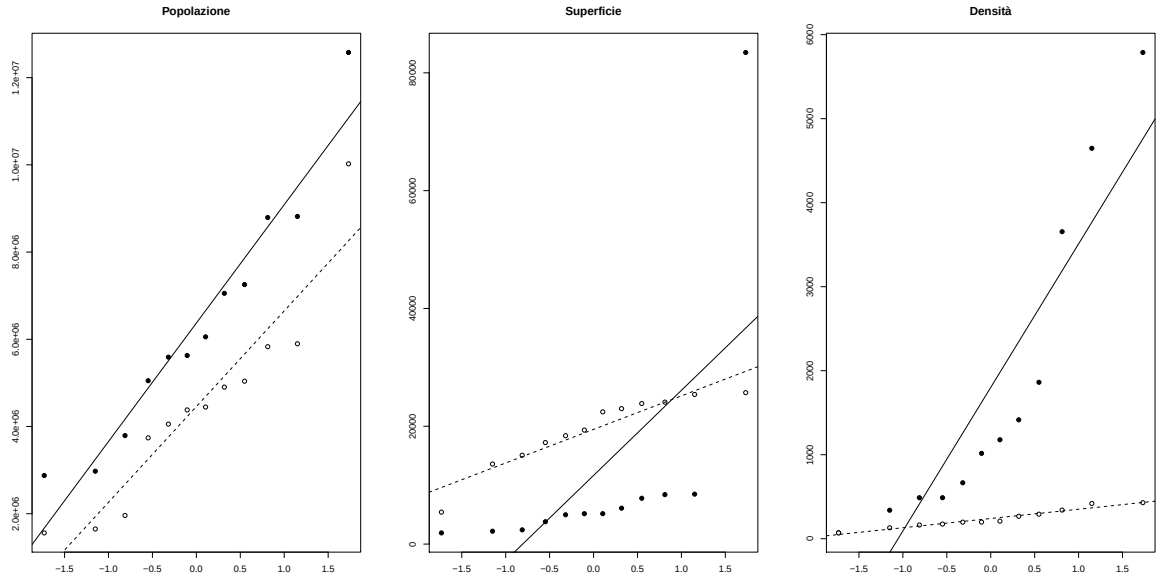


Figura 1: QQ-plot

La figura 1 mostra i QQ-plot ottenuti da tutte e tre le variabili numeriche in esame: popolazione, superficie e densità. Per ogni grafico, i cerchi neri sono i punti ottenuti dalle prefetture del Giappone, mentre i cerchi vuoti indicano i punti ottenuti dalle regioni italiane. Le linee rappresentano la regressione lineare dei punti: la linea continua si riferisce ai cerchi neri (prefetture), mentre quella tratteggiata ai cerchi vuoti (regioni). Alla pagina successiva sono elencati i risultati del test di Shapiro e Wilk eseguito per prefetture e regioni su ciascuna delle tre variabili: quale potrebbe essere la coppia di test di Shapiro e Wilk relativa a ciascuna delle tre variabili? Perché? Cosa si dovrebbe concludere?

Tabella 3: Test di Shapiro e Wilk.

Variabile	Dati	W	p-value
1	Prefetture	$0,42041$	$4,848 \times 10^{-6}$
	Regioni	$0,89079$	$0,1206$
2	Prefetture	$0,81477$	$0,01386$
	Regioni	$0,94448$	$0,5582$
3	Prefetture	$0,93692$	$0,4593$
	Regioni	$0,89346$	$0,1306$

I punti nel QQ-plot tendono a disporsi lungo una retta se la distribuzione è approssimativamente normale. L'unico grafico in cui entrambe le serie di punti si dispongono abbastanza bene lungo la rispettiva retta è il primo: corrisponderà dunque alla variabile numero 3, l'unica per cui il p-value del test di Shapiro e Wilk non è mai significativo, indicando che l'ipotesi nulla di normalità del test non può essere rifiutata. Negli altri due casi, il test risulta significativo (cioé minore di $0,05$*) solo per le prefetture, così come soprattutto i cerchi vuoti non si allineano bene lungo la retta di regressione. È difficile dire quindi a quale variabile corrispondano le altre due coppie di test: la coppia 1 corrisponde alla variabile "Superficie" e la coppia 2 alla variabile "Densità". La popolazione risulta dunque distribuita sempre in modo normale, mentre per superficie e densità questo vale solo nel caso delle regioni italiane.*

2.3 Differenze in popolazione

Vengono proposti di seguito diversi approcci per confrontare la popolazione di prefetture e regioni in esame. Qual è quello corretto per capire se le prime 12 prefetture giapponesi per popolazione siano globalmente più o meno popolose delle prime 12 regioni italiane per popolazione? Cosa si può concludere?

- Test t a campioni appaiati:

```
Paired t-test

t = 10.465, df = 11, p-value = 4.687e-07
alternative hypothesis:  true difference in means is not equal
to 0
```

- Test t a campioni non appaiati:

```
Two Sample t-test

t = 1.8353, df = 21.285, p-value = 0.08047
alternative hypothesis:  true difference in means is not equal
to 0
```

- Test di Wilcoxon:

```
Wilcoxon signed rank test

V = 78, p-value = 0.0004883
alternative hypothesis:  true location shift is not equal to 0
```

- Test di Mann e Whitney:

```
Wilcoxon rank sum test

W = 104, p-value = 0.06836
alternative hypothesis:  true location shift is not equal to 0
```

I campioni non sono appaiati (non c'è ragione di confrontare la prefettura più popolosa con la regione più popolosa, la seconda prefettura per popolazione con la seconda regione per popolazione e così via) e, dall'esercizio precedente, si sa che la distribuzione della popolazione è normale: il test corretto è perciò il secondo, il test t a campioni non appaiati, da cui risulta che non ci sono differenze significative tra prefetture e regioni in esame per quanto riguarda la popolazione ($p > 0,05$).

2.4 Legame tra popolazione e densità

Tabella 4: Regressione lineare tra densità e popolazione - Prefetture giapponesi

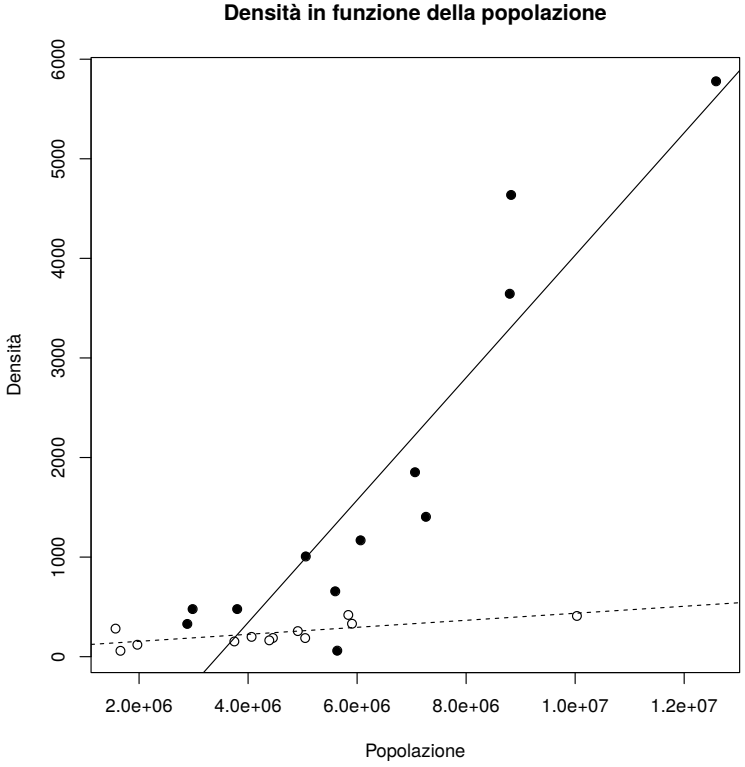
	Stima	p-value	r	R^2
Intercetta	$-2,116 \times 10^3$			
Pendenza	$6,147 \times 10^{-4}$	$3,22 \times 10^{-5}$	$0,9138809$	$0,8352$

Tabella 5: Regressione lineare tra densità e popolazione - Regioni italiane

	Stima	p-value	r	R^2
Intercetta	$8,375 \times 10^1$			
Pendenza	$3,520 \times 10^{-5}$	$0,00739$	$0,727002$	$0,5285$

La figura 2.4 alla pagina successiva mostra la regressione lineare tra densità abitativa e popolazione residente: come nel caso della figura 1, i cerchi neri interpolati dalla linea continua si riferiscono ai dati delle prefetture giapponesi, mentre i cerchi vuoti interpolati dalla linea tratteggiata si riferiscono alle regioni italiane.

Le tabelle 4 e 5 elencano i principali parametri delle correlazioni mostrate in figura 2.4. Cosa si può concludere dall'analisi dei modelli lineari? Ci sono differenze tra prefetture e regioni?



In entrambi i casi c'è una correlazione (positiva) significativa tra densità e popolazione ($p > 0,05$); nel caso delle prefetture giapponesi questa correlazione lineare appare più forte (con un R^2 pari a $0,83$ contro lo $0,52$ delle regioni italiane).

2.5 Fanalino di coda

L'ultima regione italiana per popolazione è la Valle d'Aosta, con soli 126388 abitanti al 2017; la densità abitativa della Valle d'Aosta risulta di 39 abitanti per kilometro quadrato[3]. Si può affermare che la densità abitativa della Valle d'Aosta è significativamente più bassa di quella delle 12 regioni italiane più popolose di cui alla tabella 2? Quale test statistico è necessario eseguire per dare la risposta con rigore? Cosa si ottiene?

[3]Fonte: http://demo.istat.it/bilmens2017gen/index.html.

Per rispondere con rigore alla domanda è necessario calcolare la deviata normale, perché si sa che la densità delle regioni italiane si distribuisce in modo normale:

$$Z = \frac{X - \mu}{\sigma} = \frac{39 - 240,6667}{107,0789} = -1,88$$

Dalla tabella si ottiene che questo valore di Z è associato a un'area sotto la curva normale di $1,0000 - 0,9699 = 0,0301$*, il che significa che valori così o più piccoli corrispondono a una quota di* $3,01\%$ *della distributione totale. Questa quota è inferiore alla soglia standard del* 5% *se il test si considera a una coda (cioé se la domanda è se la densità della Valle d'Aosta sia significativamente* minore *delle altre 12), ma non è inferiore alla soglia standard del* $2,5\%$ *se il test si considera a due code (cioè se la domanda è se la densità della Valle d'Aosta sia significativamente* diversa *dalle altre 12).*
Visto che si prende la Valle d'Aosta in quanto regione meno popolosa d'Italia e che si sa che la densità è correlata positivamente con la popolazione, per cui ci si aspetta che questa regione possa essere casomai meno *densamente abitata delle altre, il test più adatto potrebbe essere quello a una coda: questo porterebbe a concludere che la densità abitativa della Valle d'Aosta è significativamente più bassa di quella delle altre 12 regioni, anche considerando la sua ridotta popolazione.*