Quesito 3 – ordinaria. PNI 1995

3. Nella tabella seguente sono riportati i dati di un'indagine campionaria, relativamente ad alcune regioni e al 1990, sulla distribuzione delle abitazioni secondo la superficie abitata (area espressa in metri quadrati):

Superficie / Regione	50-95 mq	96-110 mq	111-130 mq	131-200 mq
Liguria	130	11	6	5
Campania	362	1805	105	122
Sicilia	1068	430	203	149

Il candidato:

a. stimi la superficie media abitata nelle tre regioni e la deviazione standard delle stime, assumendo come valore rappresentativo di ogni classe il valore medio;
b. rappresenti mediante diagrammi opportuni le distribuzioni marginali, rispettivamente per regione e per superficie;
c. verifichi l'ipotesi: H^o: non c'è differenza significativa (5%) tra le medie delle superfici nelle diverse regioni;
d. verifichi l'ipotesi: H^o: non c'è differenza significativa (5%) tra le distribuzioni relative alle diverse regioni.

Soluzione

a. Indici statistici

Nella tabella a doppia entrata si legge il numero delle abitazioni (frequenze assolute) distribuite secondo due caratteri

- La regione di appartenenza (carattere qualitativo)
- La superficie misurata in metri quadrati (carattere quantitativo del quale si chiede la media e la deviazione standard

Come suggerito dal testo, sostituiamo**,** all'intervallo dei valori di superficie, la loro semisomma e aggiungiamo le distribuzioni marginali

	72,5 mq	103 mq	120,5mq	165,5mq	Marginale
Liguria	130	11	6	5	**152**
Campania	362	1805	105	122	**2394**
Sicilia	1068	430	203	149	**1850**
Marginale	1560	2246	314	276	**4396**

Per ciascuna regione, la media aritmetica è la **media ponderata**

$$\bar{X} = \frac{\sum_{i=1}^{4} x_i f_i}{\sum_{i=1}^{4} f_i}$$

Pertanto,

$$M_{Liguria} = \frac{72{,}5\cdot 130+103\cdot 11+120{,}5\cdot 6+165{,}5\cdot 5}{152} = \frac{12108{,}5}{152} \approx \mathbf{79{,}7\ mq}$$

$$M_{Campania} = \frac{72{,}5\cdot 362+103\cdot 1805+120{,}5\cdot 105+165{,}5\cdot 122}{2394} = \frac{12108{,}5}{2394} \approx \mathbf{102{,}3\ mq}$$

$$M_{Sicilia} = \frac{72{,}5\cdot 1068+103\cdot 430+120{,}5\cdot 203+165{,}5\cdot 149}{1850} = \frac{170841}{1850} \approx \mathbf{92{,}3\ mq}$$

La deviazione standard σ (o scarto quadratico medio) è la radice quadrata della varianza

Entrambi sono indici di dispersione atti a misurare l'addensamento maggiore o minore dei dati intorno alla media aritmetica.

Per ciascuna regione

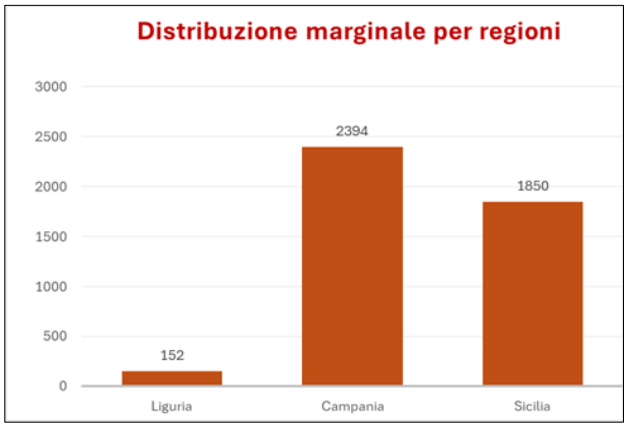
$\sigma = \sqrt{\frac{\sum_{i=1}^{4} f_i (x_i - \bar{X})^2}{\sum_{i=1}^{4} f_i}}$ **Formula alternativa** $\sigma = \sqrt{\frac{\sum_{i=1}^{4} f_i \cdot {x_i}^2}{\sum_{i=1}^{4} f_i} - \bar{X}^2}$

Deviazione standard delle stime $\sigma_x = \frac{\sigma}{\sqrt{\sum_{i=1}^{4} f_i}}$

	$f_1(x_1-\bar{X})^2$	$f_2(x_2-\bar{X})^2$	$f_3(x_3-\bar{X})^2$	$f_4(x_4-\bar{X})^2$	$\sum_{i=1}^{4} f_i(x_i \cdot -\bar{X})^2$	**Deviazione standard σ**	
						σ	σ_x
Liguria	6666,73	5991,70	10006,85	36841,51	59506,80	**19,8**	**1,6**
Campania	322347,96	784,73	34625,03	486670,73	844428,46	**18,8**	**0.4**
Sicilia	420667,07	48803,86	160901,93	797364,04	1427736,90	**27,8**	**0,6**

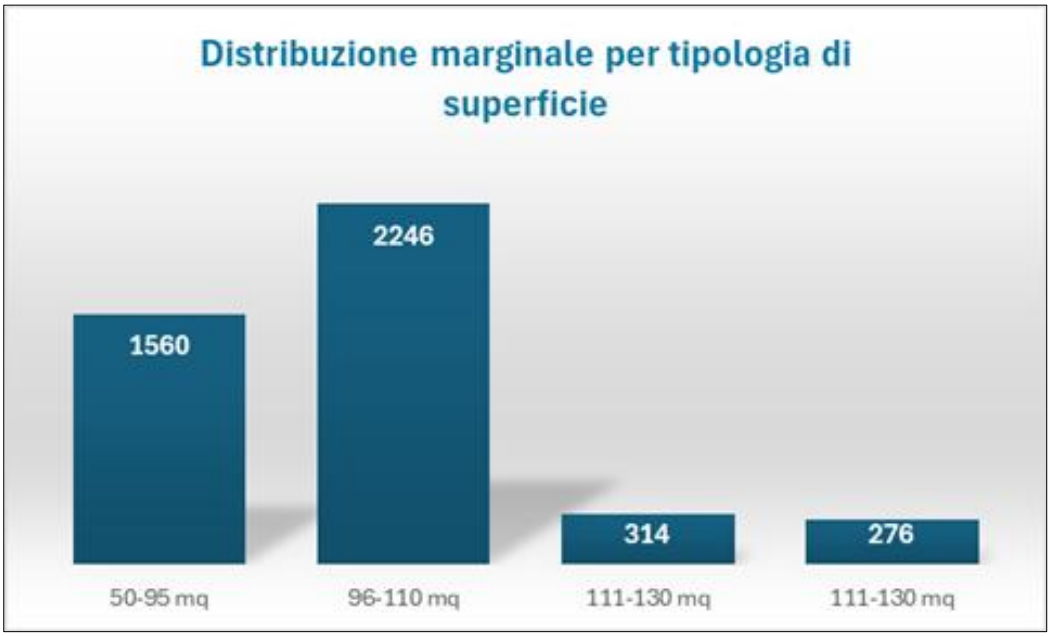
b. Rappresentazioni grafiche

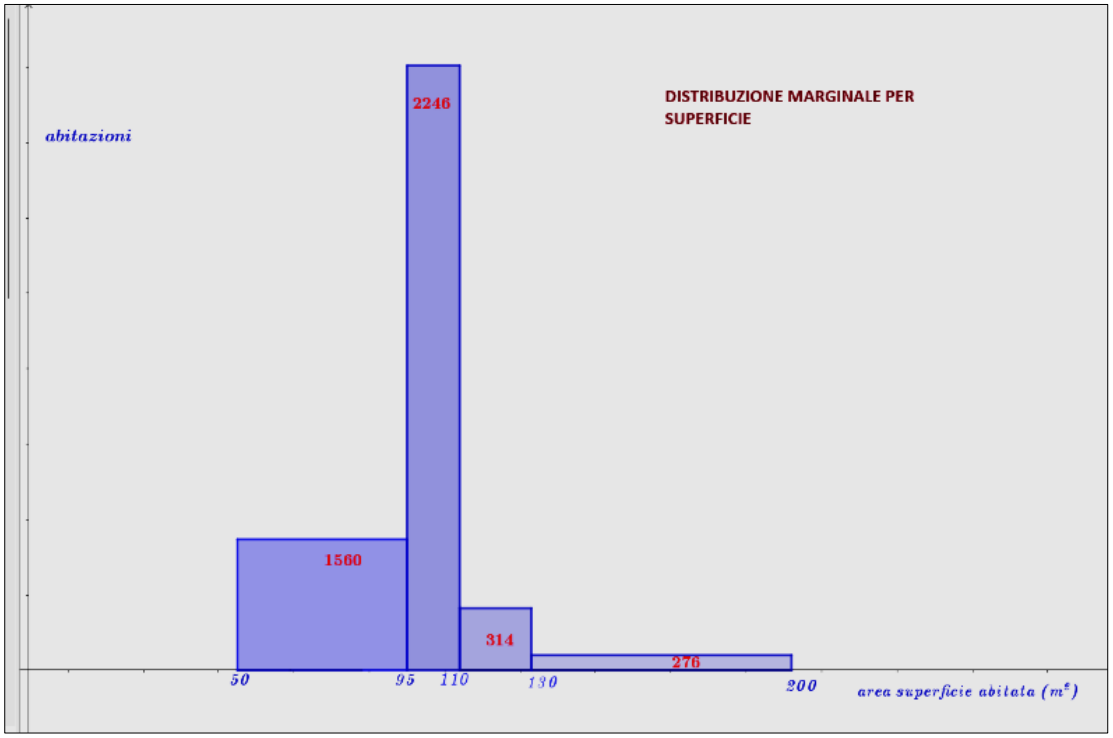
La distribuzione marginale, relativa al numero di abitazioni osservate per ciascuna regione, può essere rappresentata con un diagramma a colonne ,



Per la distribuzione marginale per superficie, la scelta dipende dagli obiettivi dell'indagine statistica.

- **Se si considera irrilevante la misura delle superfici ma serve solo una scala di tipo ordinale (abitazioni piccole, medie, grandi, molto grandi) ,si può utilizzare il diagramma a barre (o colonne)**
- **Se è importante l'informazione di tipo metrico , è preferibile una rappresentazione con istogrammi in quanto gli intervalli in cui è stato suddiviso il carattere "area" non hanno uguale ampiezza.(Nel grafico a istogrammi le aree dei rettangoli sono proporzionali alle frequenze assolute).**





c. Confronto delle medie

Riassumiamo le informazioni sulle tre distribuzioni statistiche

Regione	Numerosità	Media (mq)	Deviazione standard
Liguria	152	79,7	19,8
Campania	2394	102,3	18,8
Sicilia	1850	92,3	27,8

Si chiede se la differenza delle medie, tra le 3 regioni, possa essere considerata non significativa (livello 5%), cioè se la suddetta differenza non sia dovuta a un fattore specifico ma sia frutto di fluttuazioni casuali.

Occorre , pertanto, effettuare un confronto tra le regioni, a due a due:

- **se la differenza delle medie risulta <u>significativa</u> <u>almeno in un confronto</u> , l'ipotesi H^0 viene rifiutata**
- **se la differenza delle medie risulta <u>non significativa</u> <u>in tutti i confronti</u> , l'ipotesi H^0 viene accettata.**

La variabile aleatoria "differenza delle medie" può essere approssimata da una **variabile aleatoria a distribuzione normale**, con media uguale a 0 e varianza pari alla somma $\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}$, dove ${\sigma_1}^2 e\ {\sigma_2}^2$ sono le varianze dei campioni considerati , n_1 e n_2 le numerosità.

La **variabile standardizzata** relativa a ciascuna coppia di regioni è $t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$

Liguria- Campania

$$t = \frac{79{,}7 - 102{,}3}{\sqrt{\frac{19{,}8^2}{152} + \frac{18{,}8^2}{2394}}} \approx \mathbf{-12{,}7}$$

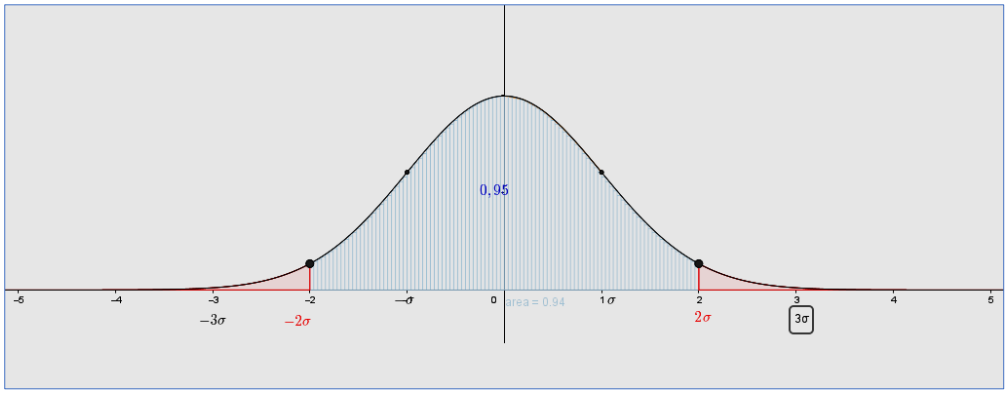
Liguria- Sicilia

$$t = \frac{79{,}7 - 92{,}3}{\sqrt{\frac{19{,}8^2}{152} + \frac{27{,}8^2}{1850}}} \approx \mathbf{-7{,}3}$$

Campania – Sicilia

$$t = \frac{102{,}3 - 92{,}3}{\sqrt{\frac{18{,}8^2}{2394} + \frac{27{,}8^2}{1850}}} \approx \mathbf{13{,}3}$$

Si verifica facilmente che tutti e tre i valori sono significativi e cadono nella zona di rifiuto dell'ipotesi nulla, con livello di significatività del 5%.



Infatti, per una variabile X aleatoria a distribuzione normale

$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 95\%$ *mentre* $P(X < \mu - 2\sigma) \cup (X > \mu + 2\sigma) \approx 5\%$

Poiché nella gaussiana standardizzata $\mu = 0$ $e\,\sigma = 1$

la differenza delle medie è

- Liguria -Campania $-12{,}7 < \mu - 2\sigma = -2$
- Liguria- Sicilia $-7{,}3 < \mu - 2\sigma = -2$
- Campania – Sicilia $13{,}3 > \mu + 2\sigma = 2$

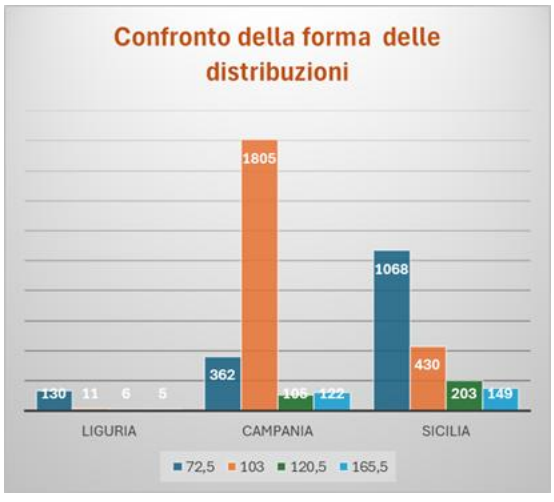
Precisamente, il valore critico letto nelle tavole della distribuzione normale standardizzata, con significatività del 5%, *è uguale a* ***1.96*** *(al quale va associato il valore simmetrico -1.96), valori per i quali le disuguaglianze sono ancora valide*

d. Test di verifica delle ipotesi sulle distribuzioni

Formulazione delle Ipotesi

- **Ipotesi Nulla** (H^0): Non c'è differenza significativa (al livello di significatività del 5%) tra le distribuzioni delle superfici abitate nelle diverse regioni (Liguria, Campania e Sicilia).
- **Ipotesi Alternativa (H^1):** Esiste una differenza significativa tra le distribuzioni.

Confronto tra i grafici delle distribuzioni delle singole regioni



Il confronto qualitativo suggerisce una differenza significativa tra le distribuzioni, inoltre il risultato del punto c induce a non accettare l'ipotesi nulla del punto d (se non ci fosse differenza significativa tra le distribuzioni non ci sarebbe neppure differenza significativa tra le medie)
Per un confronto quantitativo ricorriamo a un test "$chi\ quadrato$ ".

A tal fine costruiamo una tabella delle frequenze teoriche o frequenze "attese "nel caso in cui non ci fosse alcuna relazione tra regione e superficie abitata . I valori marginali di riga (o di colonna) saranno, pertanto, ridistribuiti proporzionalmente ai valori marginali di colonna (o di riga)

$$\frac{x_{i,j}}{totale\ colonna\ j} = \frac{totale\ riga\ i}{totale\ generale} \qquad \frac{x_{i,j}}{totale\ \ riga\ i} = \frac{totale\ colonna\ j}{totale\ generale}$$

Nella tabella delle frequenze attese (o teoriche),il generico termine $x_{i,j}$, appartenente alla $i-sima$ riga e alle $j-sima$ colonna deve essere $x_{i,j} = \frac{totale\ di\ riga\ i \cdot totale\ colonna\ j}{totale\ generale}$

La Statistica χ^2

Per misurare quanto le frequenze osservate si discostano dalle frequenze attese si calcola, per ogni cella, il valore $\frac{(Osservato - Atteso)^2}{Atteso}$ e si sommano tutti i valori.

Il numero ottenuto, che si indica con χ^2

- **vale 0 in caso di indipendenza assoluta**
- **è maggiore di 0 se le frequenze osservate differiscono da quelle teoriche**

Più grande è il valore di χ^2, più grande è la discrepanza tra le frequenze osservate e quelle teoriche

Tabella osservata

	72,5 mq	103 mq	120,5mq	165,5mq	Marginale
Liguria	130	11	6	5	152
Campania	362	1805	105	122	2394
Sicilia	1068	430	203	149	1850
Marginale	1560	2246	314	276	4396

Tabella teorica

	72,5 mq	103 mq	120,5mq	165,5mq	Marginale
Liguria	53,9	77,7	10,9	9,5	152
Campania	849,6	1223,1	171,0	150,3	2394
Sicilia	656,5	945,2	132,1	116,2	1850
Marginale	1560	2246	314	276	4396

Tabella per il calcolo del χ^2 Gradi di libertà = (numero colonne-1) (numero righe -1) = 6

	72,5	103	120,5	165,5	TOTALE
Liguria	107,25	57,22	2,17	2,16	168,80
Campania	279,80	276,80	25,47	5,33	587,41
Sicilia	257,92	280,82	37,99	9,29	586,03
					χ^2 =1342,24

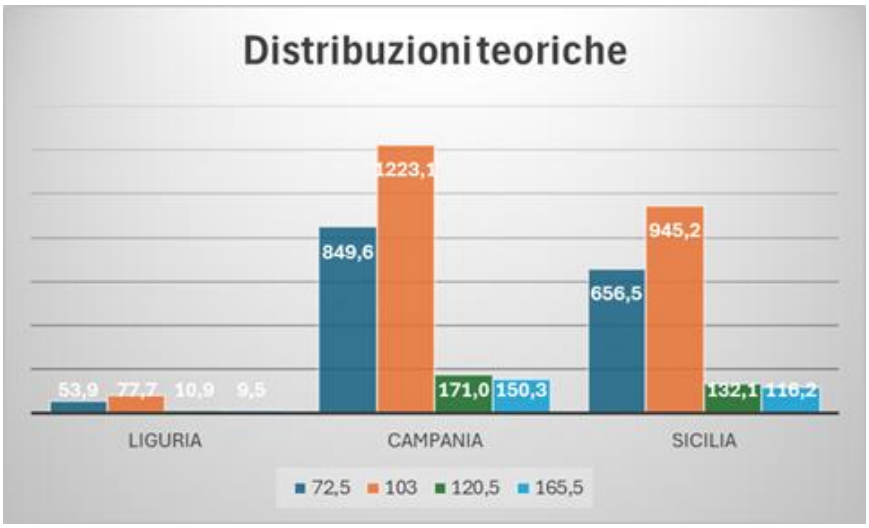
Confronto col valore critico

Consultando la tavola di distribuzione del **Chi Quadrato** $[allegata\ alla\ fine\ delle\ soluzioni\ ]$ si osserva che il valore critico corrispondente al livello di significatività del 5% (α = 0,05) e a 6 gradi di libertà è uguale a **12,592.**

Il suo valore può essere stimato, senza far uso delle tavole, come suggerito nel caso precedente, approssimando la variabile aleatoria χ^2 con una distribuzione normale, avente media μ uguale al numero g di gradi di libertà e varianza $\sigma^2 = 2g$ (teorema del limite centrale).

$$1{,}65 < \frac{x-\mu}{\sigma} < 1{,}96 \qquad 1{,}65 < \frac{x-6}{\sqrt{12}} < 1{,}96$$

Essendo il valore di χ^2 maggiore del valore critico, si rifiuta l'ipotesi nulla H^0e si conclude che esiste una differenza significativa tra le distribuzioni delle superfici abitate nelle diverse regioni.



Soluzione alternativa- Si confrontano le distribuzioni delle tre regioni a due a due.

Questa interpretazione dell'ultima richiesta del problema sembra avere maggiore coerenza con le richieste del punto c (confronto di due campioni) , ha il vantaggio di semplificare di molto i calcoli e, soprattutto, rendere meno problematico il confronto coi valori critici in assenza di tavole da consultare.

Si prende come riferimento, fra le due regioni, quella di maggiore numerosità. Stimiamo ogni volta la distribuzione teorica sulla base del campione più numeroso e calcoliamo la distribuzione attesa corrispondente al campione meno numeroso. I valori di quest'ultima dovranno risultare proporzionali ai valori del campione di riferimento, lasciando invariato il valore marginale

ESEMPIO					
Superficie	72,5	103	120,5	165,5	Marginale
Liguria -valori attesi	23,0	114,6	6,7	7,7	152
Campania	362	1805	105	122	2394
Rapporto	15,75	15,75	15,75	15,75	15,75

Le tabelle da utilizzare hanno 2 righe e 4 colonne, pertanto, si hanno 3 gradi di libertà.

In presenza di valori abbastanza alti del χ^2 sarà sufficiente ricordare l'ordine di grandezza dei valori critici delle prime righe delle apposite tavole. **In tutti e tre i casi il valore del χ^2 è maggiore del valore critico (α = 0,05 e 3 gradi di libertà = 7,15) si rifiuta l'ipotesi nulla H^0**

CONFRONTO LIGURIA -CAMPANIA					
Superficie	**72,5**	**103**	**120,5**	**165,5**	**Totale**
$\boldsymbol{Distribuzione\ osservata\text{-}Liguria}$	**130**	**11**	**6**	**5**	**152**
$\boldsymbol{Distribuzione\ attesa\text{-}Liguria}$	**23**	**114,6**	**6,7**	**7,7**	**152**
$\dfrac{(Val.osservato - Val.atteso)^2}{Val.atteso}$	**498,3**	**93,7**	**0,1**	**1,0**	χ^2 **593**

CONFRONTO LIGURIA -SICILIA					
Superficie	**72,5**	**103**	**120,5**	**165,5**	**Totale**
$Distribuzione\ osservata\text{-}Liguria$	**130**	**11**	**6**	**5**	**152**
$Distribuzione\ attesa$-Liguria	**87,7**	**35,3**	**16,7**	**12,3**	**152**
$\frac{(Val.\,osservato - Val.\,a\,tteso)^2}{Val.\,atteso}$	**20,3**	**16,8**	**6,8**	**12,3**	χ^2 **48,2**

CONFRONTO CAMPANIA -SICILIA					
Superficie	**72,5**	**103**	**120,5**	**165,5**	**Totale**
$Distribuzione\ osservata_Sicilia$	**1068**	**430**	**203**	**149**	**1850**
$Distribuzione\ attesa_Sicilia$	**279,7**	**1394,8**	**81,2**	**94,3**	**1850**
$\frac{(Val.\,osservato - Val.\,atteso)^2}{Val.\,atteso}$	**2221,2**	**667,4**	**183,0**	**31,8**	χ^2 **3103,4**

Variabile aleatoria chi-quadrato per quantili

α: indice di significatività
g: gradi di libertà

$g\backslash\alpha$	**0,100**	**0,050**	**0,025**	**0,020**	**0,010**	**0,005**	**0,0025**	**0,001**	**0,0005**
1	2,706	3,841	5,024	5,412	6,635	7,879	9,141	10,828	12,116
2	4,605	5,991	7,378	7,824	9,210	10,597	11,983	13,816	15,202
3	6,251	7,815	9,348	9,837	11,345	12,838	14,320	16,266	17,730
4	7,779	9,488	11,143	11,668	13,277	14,860	16,424	18,467	19,997
5	9,236	11,070	12,833	13,388	15,086	16,750	18,386	20,515	22,105
6	10,645	12,592	14,449	15,033	16,812	18,548	20,249	22,458	24,103
7	12,017	14,067	16,013	16,622	18,475	20,278	22,040	24,322	26,018
8	13,362	15,507	17,535	18,168	20,090	21,955	23,774	26,124	27,868
9	14,684	16,919	19,023	19,679	21,666	23,589	25,462	27,877	29,666
10	15,987	18,307	20,483	21,161	23,209	25,188	27,112	29,588	31,420
11	17,275	19,675	21,920	22,618	24,725	26,757	28,729	31,264	33,137
12	18,549	21,026	23,337	24,054	26,217	28,300	30,318	32,909	34,821
13	19,812	22,362	24,736	25,472	27,688	29,819	31,883	34,528	36,478
14	21,064	23,685	26,119	26,873	29,141	31,319	33,426	36,123	38,109
15	22,307	24,996	27,488	28,259	30,578	32,801	34,950	37,697	39,719
16	23,542	26,296	28,845	29,633	32,000	34,267	36,456	39,252	41,308
17	24,769	27,587	30,191	30,995	33,409	35,718	37,946	40,790	42,879
18	25,989	28,869	31,526	32,346	34,805	37,156	39,422	42,312	44,434
19	27,204	30,144	32,852	33,687	36,191	38,582	40,885	43,820	45,973
20	28,412	31,410	34,170	35,020	37,566	39,997	42,336	45,315	47,498
21	29,615	32,671	35,479	36,343	38,932	41,401	43,775	46,797	49,011
22	30,813	33,924	36,781	37,659	40,289	42,796	45,204	48,268	50,511
23	32,007	35,172	38,076	38,968	41,638	44,181	46,623	49,728	52,000
24	33,196	36,415	39,364	40,270	42,980	45,559	48,034	51,179	53,479
25	34,382	37,652	40,646	41,566	44,314	46,928	49,435	52,620	54,947
26	35,563	38,885	41,923	42,856	45,642	48,290	50,829	54,052	56,407
27	36,741	40,113	43,195	44,140	46,963	49,645	52,215	55,476	57,858
28	37,916	41,337	44,461	45,419	48,278	50,993	53,594	56,892	59,300
29	39,087	42,557	45,722	46,693	49,588	52,336	54,967	58,301	60,735
30	40,256	43,773	46,979	47,962	50,892	53,672	56,332	59,703	62,162
40	51,805	55,758	59,342	60,436	63,691	66,766	69,699	73,402	76,095
50	63,167	67,505	71,420	72,613	76,154	79,490	82,664	86,661	89,561
60	74,397	79,082	83,298	84,580	88,379	91,952	95,344	99,607	102,695
70	85,527	90,531	95,023	96,388	100,425	104,215	107,808	112,317	115,578
80	96,578	101,879	106,629	108,069	112,329	116,321	120,102	124,839	128,261
100	118,498	124,342	129,561	131,142	135,807	140,169	144,293	149,449	153,167