# Statistica

Corso di Fisica ed Elementi di Laboratorio ed Informatica

CdL Scienze Biologiche

AA 2016/2017

# Errori casuali e sistematici

Abbiamo definito due tipi di errori che si evidenziano quando effettuiamo una misura

**Errori sistematici**

- Si ripetono sistematicamente ad ogni misura effettuata
- Sono solitamente legati a cause di errore intrinseche nel processo di misurazione
- Di solito non possono essere eliminati ma se sono noti possono essere trattati opportunamente.

**Errori casuali**

- Variano per ogni misurazione in modo non prevedibile
- Sono generati da imprecisioni legate alla singola misurazione (es. rumore di fondo)
- Con uguale probabilità causano sottostime e sovrastime
- Possono essere trattati con un'analisi statistica

# Media, Deviazione standard

La stima migliore per la misura è la **media** delle misurazioni

$$\bar{x} = \sum \frac{x_i}{n}$$

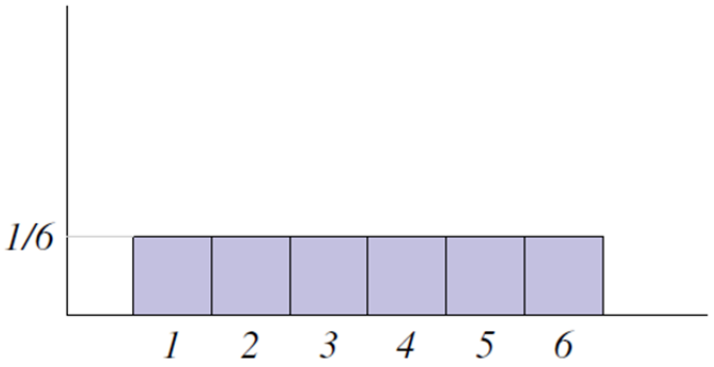
Per capire quanto le misurazioni si discostano dalla media calcoliamo la **deviazione standard**

$$\sigma_x = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$$

# Distribuzione di Gauss

La distribuzione che descrive molti fenomeni casuali è la **distribuzione di Gauss** o normale.

**Lancio di 1 dado**: la probabilità di ottenere un valore compreso tra 1 e 6 è uniforme ed è pari ad 1/6




**Lancio di 2 dadi e somma dei valori ottenuti**: la probabilità di ottenere un certo valore non è più uniforme perché abbiamo la probabilità 1/6 di ottenere il valore medio 7 e solo 1/36 di ottenere uno dei valori agli estremi (2 o 12).

Lancio_di_due_dadi.xls

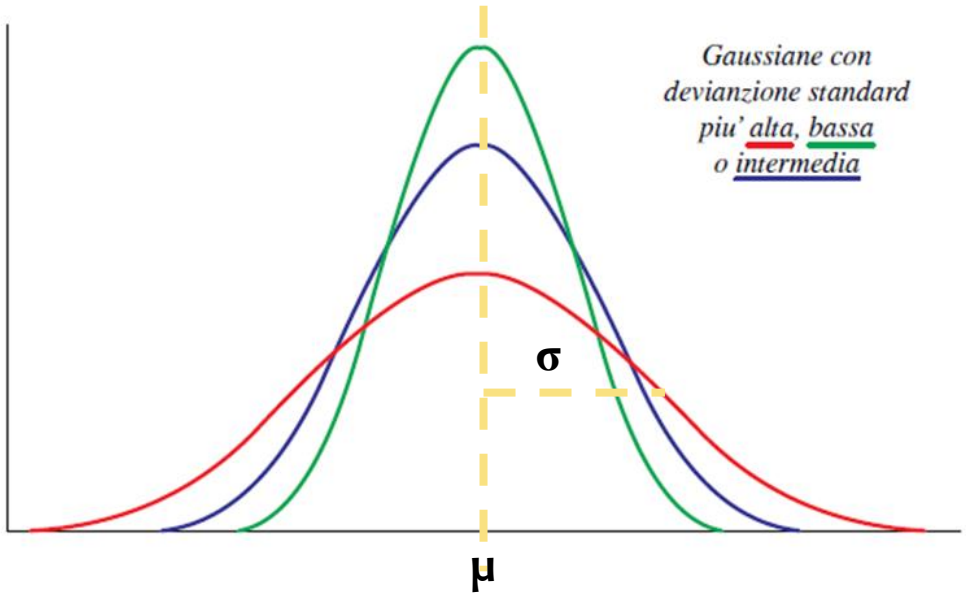
# Distribuzione di Gauss

**Distribuzione di Gauss**

Il suo grafico è a forma di campana, simmetrica rispetto al suo valore medio μ.

La simmetria della distribuzione normale da vita ad un'importante proprietà che lega tra loro la probabilità degli eventi e la deviazione standard: l'area (la probabilità) compresa nell'intervallo [μ - σ , μ + σ ] rappresenta il 68,3% della popolazione

$$G(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$



# Distribuzione di Gauss

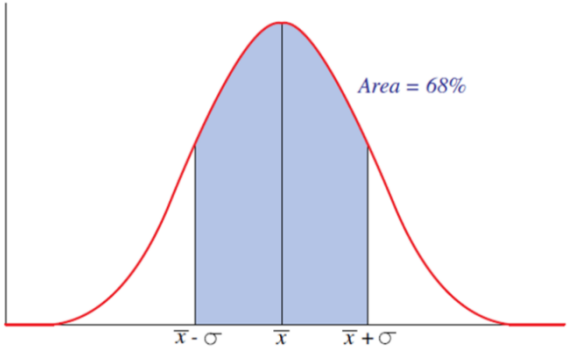
La distribuzione di Gauss è un buon modello per l'errore casuale su una misurazione.

- Il valore medio della distribuzione rappresenterà la migliore stima della misura stessa (ovviamente a meno di non avere errori sistematici)
- La deviazione standard rappresenterà la precisione con cui eseguiamo la misura.

Con questa distribuzione, la deviazione standard può essere utilizzata per definire un intervallo nel quale la misura esatta sarà presente con una certa probabilità (intervalli di **confidenza**).

# Intervalli di confidenza

In particolare




$\bar{x} \pm \bar{\sigma}_x$ corrisponde ad una confidenza del 68%
$\bar{x} \pm 2\bar{\sigma}_x$ corrisponde ad una confidenza del 95.4%
$\bar{x} \pm 3\bar{\sigma}_x$ corrisponde ad una confidenza del 99.7%
$\bar{x} \pm 5\bar{\sigma}_x$ corrisponde ad una confidenza del 99.99997%

**Attenzione che l'intervallo di confidenza non è l'intervallo in cui cadono i valori della variabile, o la media del campione, ma gli intervalli che con una certa probabilità conterranno la media della popolazione!**

(Informalmente, anche se non correttamente, si dice anche che la media della popolazione cadrà con una probabilità 1 - $\alpha$ all'interno dell'intervallo di confidenza calcolato. Ma definito un intervallo, la media della popolazione o è interna o è esterna a questo intervallo, non ha senso parlare di probabilità della media vera di cadere o no nell'intervallo calcolato)

# Esercizio

1. Misurare (almeno 100 volte) il diametro dei pesetti sferici con il calibro Palmer.

2. Rappresentare i dati ottenuti in un istogramma. Calcolare la media e deviazione standard.

3. Dividere il set di dati in sottocampioni con la seguente numerosità

N=2 N=5 N=10
Per ogni sottocampione ottenuto calcolare la media e la deviazione standard.
Riportare i risultati in grafici diversi.

4. Disegnare la gaussiana di riferimento per ogni sottocampione individuato

La distribuzione della media campionaria
- È centrata sulla media della variabile nella popolazione (quindi è una stima corretta)
- Ha minore ampiezza al crescere di n

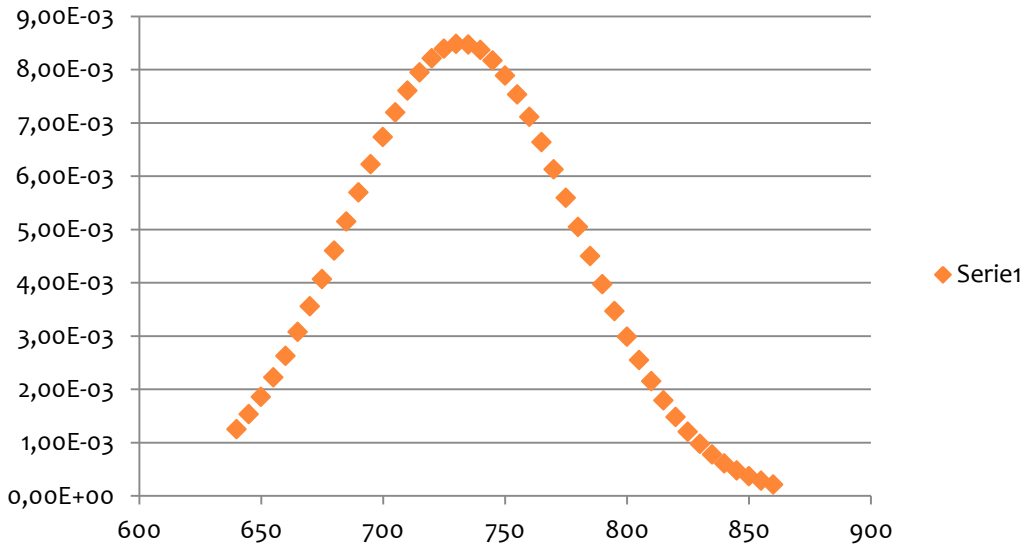
# Esercizio

**Definiamo la gaussiana in Excel**

=DISTRIB.NORM.N(x;media;dev.st;FALSO)

Restituisce la distribuzione normale per la media e la distribuzione standard specificate.

$$G(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$$



# Dev. Standard della media

Quando si esprime un risultato come

$$\bar{x} \pm \sigma_x$$

si intende dire che c'è una probabilità del 68.3% che, ripetendo una singola misura, essa cada nell'intervallo

$$(\bar{x} - \sigma_x) < x < (\bar{x} + \sigma_x)$$

Quindi la deviazione standard rappresenta l'incertezza media sulle singole misure, non l'incertezza sulla media. Si dimostra che la migliore stima per l'errore sulla media è la **deviazione standard della media** (o errore standard) che tiene conto del numero di misure effettuate

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$$

Il risultato di una misura dovrà quindi essere espresso come: $\bar{x} \pm \bar{\sigma}_x$

# Test di ipotesi

Le assunzioni fatte sulla distribuzione di probabilità di una v.a. associata ad un fenomeno reale sono dette **ipotesi** e riguardano parametri incogniti.

La **verifica statistica delle ipotesi** appura se tali ipotesi possono ritenersi compatibili con le osservazioni campionarie.

Nella prova di ipotesi distinguiamo:
- **ipotesi nulla $H_0$,** che contempla la situazione prima dell'osservazione campionaria,
- **ipotesi alternativa $H_1$,** che contempla una situazione differentemente specificata.

Le ipotesi statistiche si dicono
- **semplici** se specificano in modo univoco la distribuzione della popolazione in oggetto,
- **composte** se specificano diversi valori del parametro.

# Test di ipotesi

***La verifica di un'ipotesi significa la sua accettazione o il rifiuto, ad un prestabilito livello di probabilità.***

A questo scopo si utilizza un test: una funzione delle osservazioni che ha distribuzione nota con la condizione che l'ipotesi enunciata sia vera.
Il test è una procedura inferenziale atta a valutare la conformità probabilistica fra un campione e la popolazione.

E' possibile tramite il test valutare l'attendibilità delle osservazione campionarie, allo scopo di stabilire se le differenze rispetto alla popolazione sono casuali, dovute ad errore campionario, o significative.

**PROCEDURA DI TEST**

1. si formula l'ipotesi nulla $H_0$ ed un'ipotesi alternativa $H_1$ sulla popolazione
2. attraverso i risultati campionari ed un'opportuna valutazione statistica si decide se accettare o rigettare l'ipotesi $H_0$.

# Il test del $\chi^2$

E' un test di verifica d'ipotesi che utilizza la distribuzione della variabile casuale $\chi^2$ per decidere se rifiutare o non rifiutare l'ipotesi nulla:

Supponiamo che in un particolare campione si sia osservato che un insieme di possibili eventi $\mathbf{E_1}, \mathbf{E_2}, \ldots, \mathbf{E_k}$ si presenta con frequenze $\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_k}$ dette frequenze osservate, e che, secondo le regole della probabilità, ci si attenda che si presenti con frequenze $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_k}$ dette frequenze teoriche o attese.

La variabile test $\chi^2$ si ottiene sommando, per ogni evento $\mathbf{E_i}$ il quadrato degli scarti tra le frequenze teoriche e quelle osservate pesato sulle frequenze teoriche

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

http://dctf.uniroma1.it/galenotech/campion5.htm

# Il test del $\chi^2$

Il valore del $\chi^2$ è un indicatore ragionevole dellìaccordo tra la distribuzione osservata e quella attesa. Se $\chi^2$ =0 l'accordo è perfetto. In generale per sostenere che l'accordo si abuono occorre che $\chi^2$ <n, dove n è il numero di termini della somma

Procedura del test:

1) Dividere il campione in intervalli
2) Calcolare la media e la deviazione standard del campione
3) Calcolare le probabilità che la misura cada negli intervalli scelti utilizzando la gaussiana che ha con parametri uguali a quelli calcolati
4) Calcolare il numero di eventi osservati moltiplicando la probabilità per il numero di misurazioni effettuate
5) Calcolare la variabile Chi2

NB: La variabile $\chi^2$ è legata alla distribuzione gaussiana in una maniera abbastanza esplicita: se infatti si parte dalla distribuzione normale e la si eleva al quadrato si ottiene la distribuzione del chi quadro ad un grado di libertà

# Esercizio

Abbiamo fatto 40 misure della gittata di un proiettile ottenendo i seguenti risultati (L in cm)

| | |
|---|---|
| 810 | 731 |
| 830 | 772 |
| 760 | 771 |
| 805 | 681 |
| 725 | 722 |
| 748 | 688 |
| 778 | 653 |
| 672 | 757 |
| 764 | 733 |
| 738 | 742 |
| 687 | 739 |
| 753 | 678 |
| 638 | 698 |
| 766 | 780 |
| 709 | 748 |
| 787 | 770 |
| 645 | 709 |
| 675 | 689 |
| 712 | 754 |

Si ottiene

**Media = 727,11 cm**

**Deviazione Std = 37,46 cm**

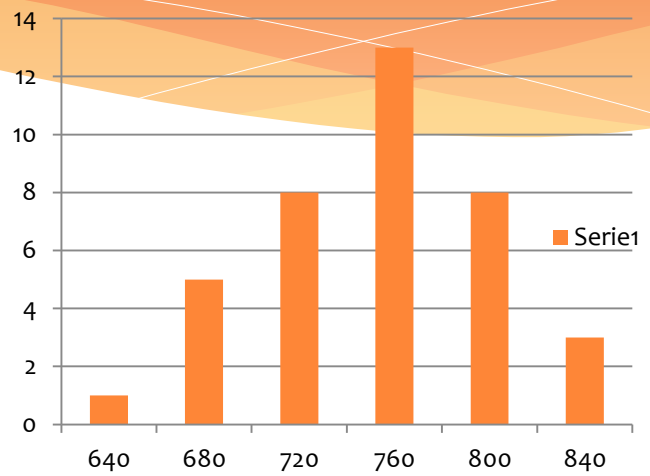
- **La distribuzione reale dei dati è consistente con la nostra ipotesi che le misure siano governate dalla distribuzione di Gauss con media e sigma stimati?**

Dobbiamo applicare il test del $\chi^2$ !

# Esercizio

**Dividiamo il campione in intervalli**

| Classi | Frequenza |
|---|---|
| 640 | 1 |
| 680 | 5 |
| 720 | 8 |
| 760 | 13 |
| 800 | 8 |
| 840 | 3 |




**Definiamo la gaussiana**

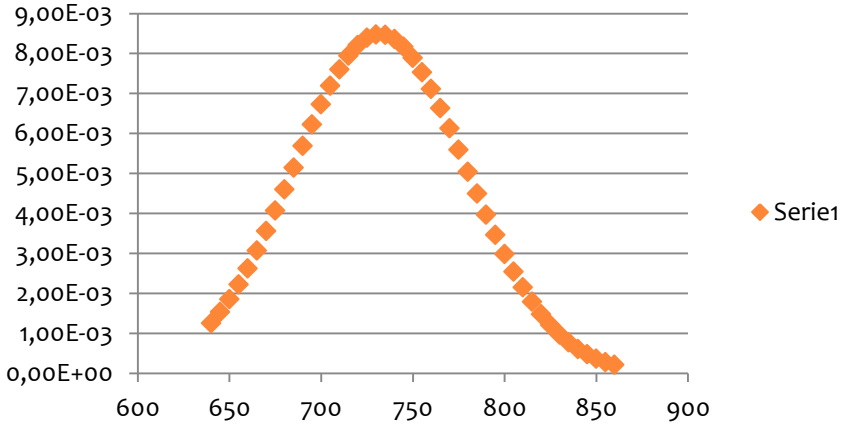
$$G(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}} = \frac{e^{-\frac{(x-727{,}11)^2}{2*37{,}46^2}}}{37{,}46\sqrt{2\pi}}$$

*Excel*

DISTRIB.NORM.N(x;media;dev.st;FALSO)

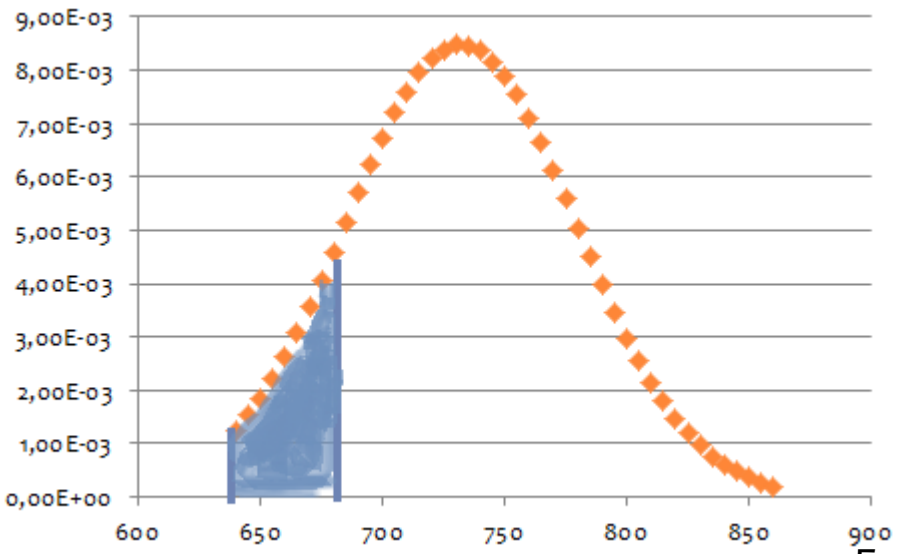
*LibreOffice*

=NORMDIST(x;media;dev.st;0)



# Esercizio

**Calcoliamo le probabilità e i valori attesi**




Calcoliamo l'integrale tra gli estremi dell'intervallo considerato.
Posso usare la funzione cumulativa nei due punti e farne la differenza

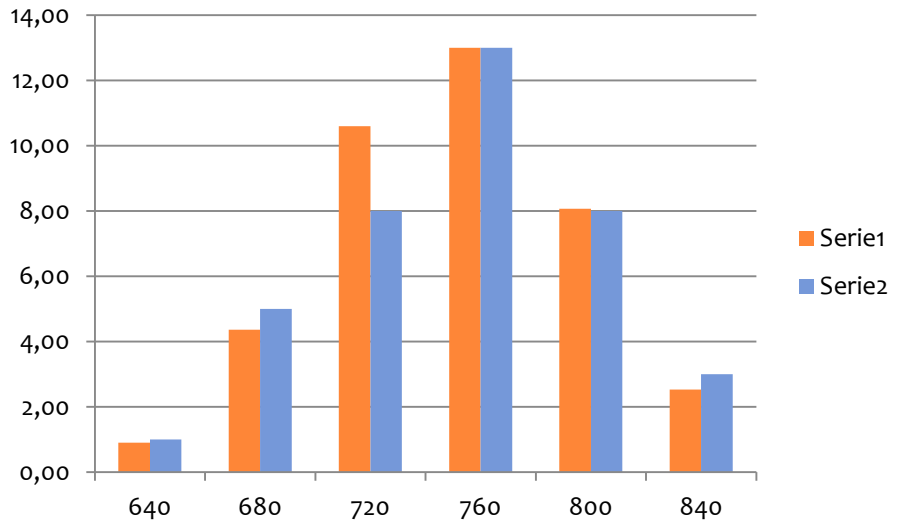
*Excel*
DISTRIB.NORM.N(x;media;dev.st;VERO)
*LibreOffice*
=NORMDIST(x;media;dev.st;1)

Restituisce la distribuzione cumulativa normale per la media e la distribuzione standard specificate.

# Esercizio

| x | Gaus | Gaus Cumulativa | Probabilità | Valori attesi |
|---|---|---|---|---|
| 600 | 1,65E-04 | 2,50E-03 | 2,27E-02 | 0,91 |
| 640 | 1,25E-03 | 2,52E-02 | 1,09E-01 | 4,36 |
| 680 | 4,60E-03 | 1,34E-01 | 2,65E-01 | 10,59 |
| 720 | 8,21E-03 | 3,99E-01 | 3,25E-01 | 13,00 |
| 760 | 7,11E-03 | 7,24E-01 | 2,02E-01 | 8,07 |
| 800 | 2,98E-03 | 9,26E-01 | 6,33E-02 | 2,53 |
| 840 | 6,08E-04 | 9,89E-01 | | |

| Chi2 | Probabilità Chi2 |
|---|---|
| 0,009468608 | 0,9756 |
| 0,092506397 | |
| 0,634017101 | |
| 3,0119E-07 | |
| 0,00064508 | |
| 0,086011344 | |
| | |
| 0,822 | |




*Excel*
TEST.CHI.QUAD(osservati;attesi)
*LibreOffice*
CHISQ.TEST(osservati;attesi)

Restituisce il valore dalla distribuzione del chi quadrato (χ2) per un dato statistico e i gradi di libertà appropriati

# Estratto tabella del $\chi^2$

I valori interni alla tabella corrispondono ai valori critici riferiti alla coda di destra, ovvero ai valori alla cui destra cade la frazione della curva riportata nella prima riga. Per esempio, con 2 gradi di libertà, il 5% della distribuzione ha valori superiori a 5.991. Si tratta quindi di una tabella delle aree *a una coda*.

| df | 0.995 | 0.99 | 0.975 | 0.95 | 0.90 | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.000 | 0.000 | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| 2 | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| 3 | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| 4 | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| 5 | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 |
| 10 | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |
| 20 | 7.434 | 8.260 | 9.591 | 10.851 | 12.443 | 28.412 | 31.410 | 34.170 | 37.566 | 39.997 |
| 21 | 8.034 | 8.897 | 10.283 | 11.591 | 13.240 | 29.615 | 32.671 | 35.479 | 38.932 | 41.401 |
| 22 | 8.643 | 9.542 | 10.982 | 12.338 | 14.041 | 30.813 | 33.924 | 36.781 | 40.289 | 42.796 |
| 23 | 9.260 | 10.196 | 11.689 | 13.091 | 14.848 | 32.007 | 35.172 | 38.076 | 41.638 | 44.181 |
| 24 | 9.886 | 10.856 | 12.401 | 13.848 | 15.659 | 33.196 | 36.415 | 39.364 | 42.980 | 45.559 |
| 25 | 10.520 | 11.524 | 13.120 | 14.611 | 16.473 | 34.382 | 37.652 | 40.646 | 44.314 | 46.928 |
| 30 | 13.787 | 14.953 | 16.791 | 18.493 | 20.599 | 40.256 | 43.773 | 46.979 | 50.892 | 53.672 |
| 40 | 20.707 | 22.164 | 24.433 | 26.509 | 29.051 | 51.805 | 55.758 | 59.342 | 63.691 | 66.766 |
| 50 | 27.991 | 29.707 | 32.357 | 34.764 | 37.689 | 63.167 | 67.505 | 71.420 | 76.154 | 79.490 |