Inferență statistică în ML

Cap 2. Expected values. Variabilitate.

February 26, 2024

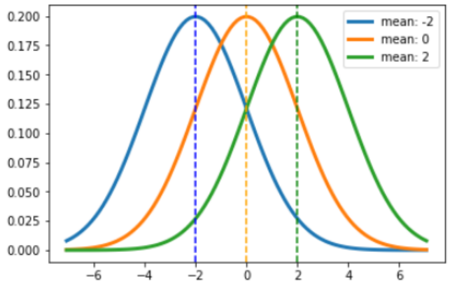
1 Expected values

2 Variabilitate

3 Anexă

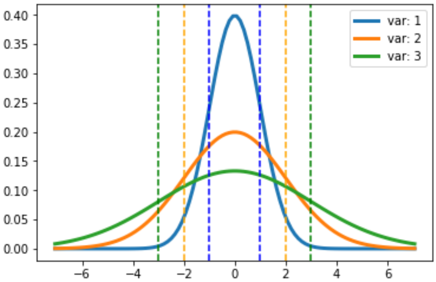
Expected values

- statistical inference: procesul de a genera concluzii despre populații folosind sample-uri noisy extrase din populație (eșantionate)
- gradul de randomness este descris de funcția densitate de probabilitate (PDF sau PMF)
- nu interesează forma analitică (exactă) a funcției ci caracteristicile acesteia: expected value, sample quantiles
- expected value (valoare așteptată, sau **media**) este centrul distribuției



- dacă distribuția se deplasează atunci se deplasează și media sa

Expected values (2)



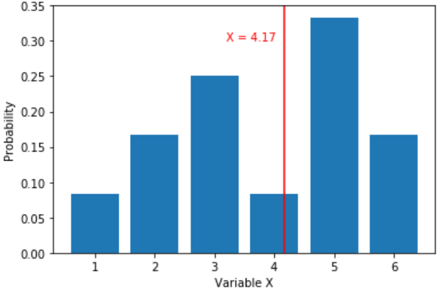
- **varince** și respectiv standard deviation descriu cât de împrăștiată este distribuția respectivă
- la fel cum sample quantiles estimează population quantiles, expected values pentru sample vor estima expected values pentru întreaga populație

- **sample mean** - estimare pentru population mean
- **sample variance** - estimare pentru population variance

Population mean

- **expected value**[1] sau **mean**[2] a unei variabile aleatoare este centrul distribuției sale
- p(x) este probability mass function a variabilei aleatoare X:

$$E[X] = \sum_x x \cdot p(x)$$



- $E[X]$ reprezintă centrul de masă (fizic) a unei colecții de puncte respectiv ponderi {x, p(x)}

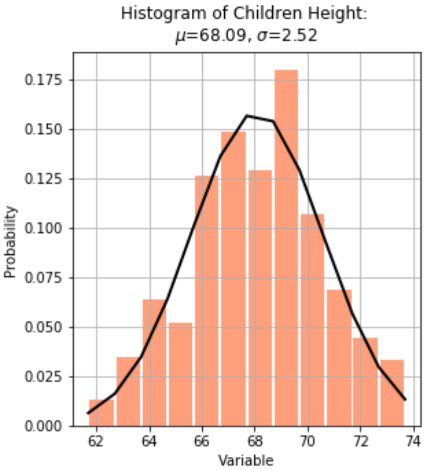
[1] valoarea așteptată
[2] media

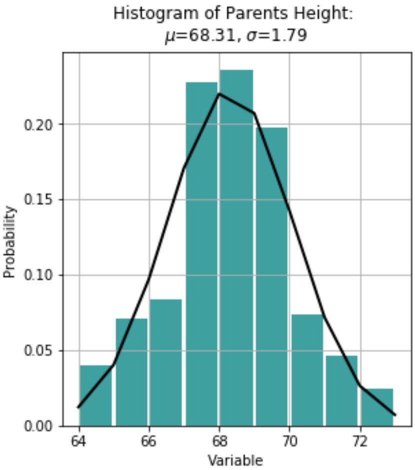
Sample mean

- centrul de masă pentru o distribuție echiprobabilă[3] este chiar media aritmetică
$$\bar{X} = \sum_{i=1}^{n} x_i \cdot p(x_i)$$
- din punct de vedere al notației, vom face distincție între valoarea așteptată (**population mean**) $E[X]$ și **sample mean** $\bar{X}$

[3] $p(x_i) = c,\ i = 1 \dots n$

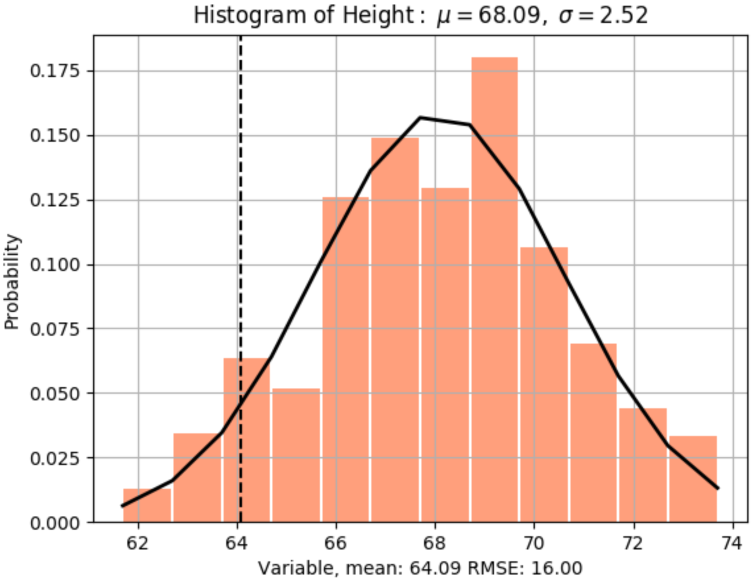
Example: Galton dataset





- centrul de masă pentru distribuția înălțimilor pentru copii și părinți

Children heights

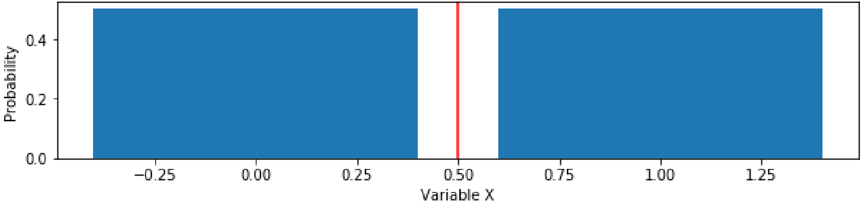


Exemplu: population mean

- pentru o monedă, la fiecare aruncare variabila aleatoare X poate lua valorile 1 dacă se obține Head respectiv 0 dacă iese Tail
- care este valoarea așteptată pentru X?
- valoarea așteptată este o proprietate a **populației**
- dacă moneda este ideală:

$$E[X] = .5 \times 0 + .5 \times 1 = .5$$

- valoarea așteptată este o valoare pe care moneda NU o poate lua



Biased coin

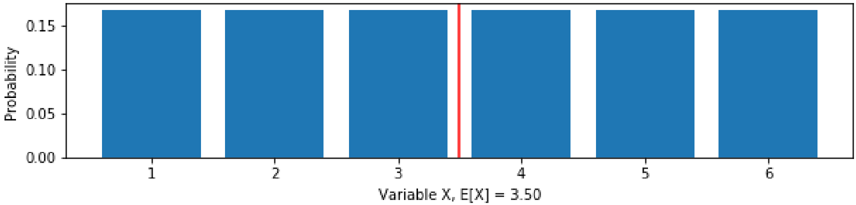
- variabila aleatoare X ce caracterizează moneda dă $P(X = 1) = p$ și $P(X = 0) = (1 - p)$

$$E[X] = 0 \cdot (1 - p) + 1 \cdot p = p$$

- valoarea așteptată a lui X este exact proporția de Heads care ar ieși după un număr infinit de aruncări

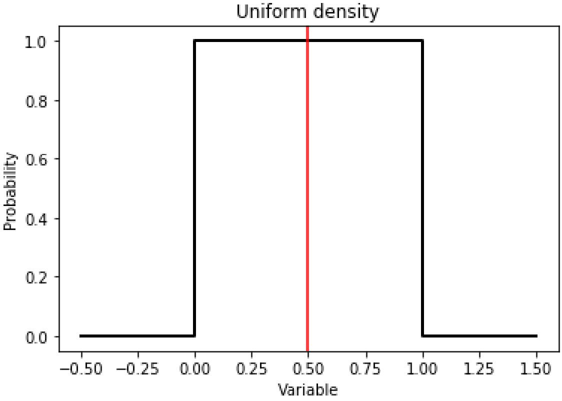
Zar ideal

$$E[X] = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5$$



Variabile aleatoare continue

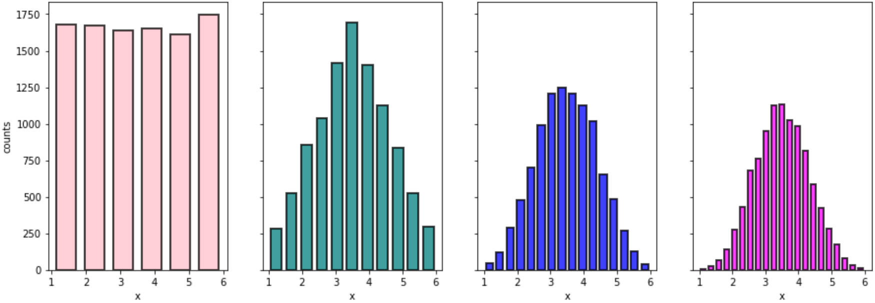
- expected value: centrul de masă poate fi asimimat cu centrul de echilibru (fizic) dacă densitatea de probabiliate ar fi o bucată de lemn



Expected values (3)

- expected values sunt proprietăți ale distribuției (centru de masă)
- media unei variabile aleatoare este ea însăși o variabilă aleatoare
  - dăm cu 10 zaruri și calculăm media
  - noua valoare este ea însăși o variabilă aleatoare
- deoarece este o variabilă aleatoare, are o anumită distribuție iar acea distribuție are o valoare așteptată
- centrul de masă al acestei distribuții este același cu cel al distribuției originale
- valoarea așteptată a mediei sample-ului este **exact media populației** pe care încearcă să o estimeze
- dacă estimatorul are această proprietate el este **unbiased**, pentru că distribuția sa e centrată pe valoarea pe care încearcă să o estimeze

Mediile a n aruncări cu zarul



- (1) distribuția aruncărilor cu zarul, din 10.000 de aruncări
- (2) distribuția mediei a 2 aruncări cu zarul;
- mai gaussiană; mai concentrată; centrată la 3.5 (population mean)
- population mean pentru media a două aruncări este population mean a aruncărilor
- valabil și pentru aruncări cu moneda (media pe 1, 10, 20 și 30 aruncări)

Sumar

- valoarea așteptată este o proprietate a distribuției
- media populației este centrul de masă al distribuției
- sample mean este centrul de masă al datelor observate (sample)
- sample mean e un estimator al population mean
- sample mean este unbiased (distribuția sample means este chiar media populației pe care încearcă să o estimeze)
- cu cât avem un sample mai mare pentru care calculăm media, cu atât mai concentrată este distribuția

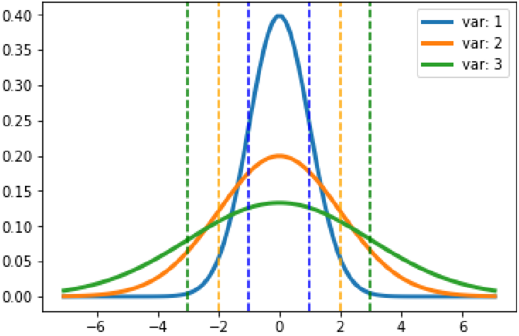
1 Expected values

2 Variabilitate

3 Anexă

Variability

- o caracteristică importantă a populației este cât de împrăștiată este
- variabilitatea se măsoară cu sample variance σ^2 sau deviația standard - sqrt(variance)
- deviația standard σ are aceeași unitate de măsură ca și populația



Variance

- **variance** pentru o variabilă aleatoare este o măsură a împrăștierii sale în jurul mediei

$$Var(X) = E[(X - \mu)^2]$$

- dispersia (variance) este valoarea așteptată[4] a pătratului distanței la care variabila aleatoare este față de medie

$$Var(X) = E[(X - \mu)^2] = \sum_i \left[p_i x_i^2 - 2p_i x_i \mu + p_i \mu^2\right]$$

$$= E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - E[X]^2$$

- rădăcina pătrată a dispersiei este denumită deviația standard[5]

[4] tot o medie

[5] standard deviation

Aruncarea cu zarul

$$E[X] = 3.5$$

$$E[X^2] = \frac{1}{6} \cdot 1^2 + \frac{1}{6} \cdot 2^2 + \frac{1}{6} \cdot 3^2 + \frac{1}{6} \cdot 4^2 + \frac{1}{6} \cdot 5^2 + \frac{1}{6} \cdot 6^2 = 15.17$$

$$Var(X) = E[X^2] - E[X]^2 = 2.92$$

Aruncarea cu o monedă

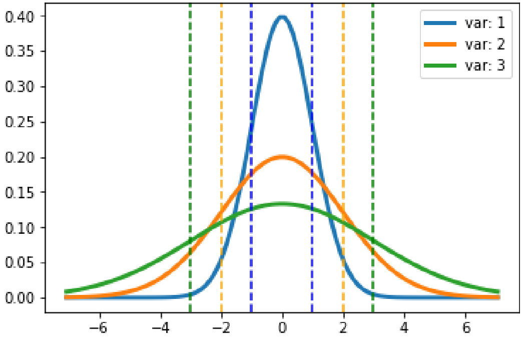
$$E[X] = 0 \times (1 - p) + 1 \times p = p$$

$$E[X^2] = 0^2 \times (1 - p) + 1^2 \times p = E[X] = p$$

$$Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1 - p)$$

- dispersia populației aruncărilor cu o monedă biased

Distribuții cu dispersii crescătoare



- dispersia crește, 'turtește' densitatea și distribuie mai multă masă către părți (tails)

Sample variance

- population mean μ este centrul de masă al populației
- sample mean $\bar{X}$ este centrul de masă al datelor observate (sample)
- population variance (dispersia, σ^2) este valoarea așteptată a pătratului distanței dintre variabila populației și media populației:

$$Var(X) = E[(X-\mu)^2] = E[X^2] - E[X]^2$$

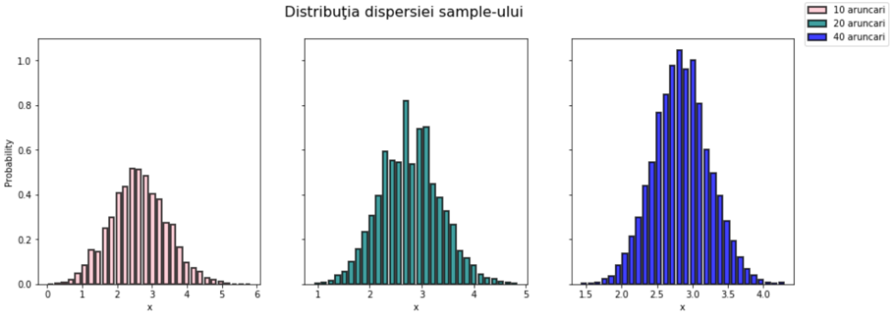
- sample variance este media pătratelor distanțelor dintre valoarea observată și sample mean:

$$S^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$$

Variance of sample variance

- variance (dispersia) este o caracteristică a datelor, este tot o variabilă aleatoare, are deci o distribuție a populației
- acea populație a dispersiilor are și ea o valoare așteptată (medie), iar
- acea valoare așteptată este este population variance σ^2 pe care sample-ul încearcă să o estimeze
- pe măsură ce se acumulează mai multe date, populația de dispersii va fi din ce în ce mai concentrată către population variance pe care încearcă să o estimeze
- rădăcina pătrată este deviația standard a sample-ului

Variance of sample variance (2)



- population distribution pentru dispersia aruncărilor cu zarul de 10, 20, respectiv 40 de ori
- distribuția (pentru toate cazurile) este centrată în jurul valorii $Var(X) = 2.92$
- distribuția devine mai concentrată în jurul a ceea ce încearcă să estimeze (population variance) - unbiased (de aici vine n-1)

Corecția Bessel

- calculul sample variance se bazează pe diferențele $(X_i - \bar{X})$
- putem calcula media sample-ului $\bar{X}$, iar atunci nu toate diferențele vor fi independente
- de exemplu, putem considera $(X_i - \bar{X})$ independente pentru $i = 1 \dots n - 1$, dar $(X_n - \bar{X})$ este calculabilă știind primele diferențe și media
- în această configurație, n este lungimea sample-ului, iar $n - 1$ este numărul de grade de libertate[6] al variabilei aleatoare care este sample variance - ultima diferență față de medie e calculabilă din cele $i = 1 \dots n - 1$ valori X_i și sample mean $\bar{X}$

[6] degrees of freedom

Corecția Bessel (2)

- folosim valoarea așteptată a variabilei aleatoare $x_1 x_2$, unde x_1 și x_2 sunt două sample-uri extrase din distribuție, deci variabile aleatoare independente: $E[x_1 x_2] = E[x_1]E[x_2]$

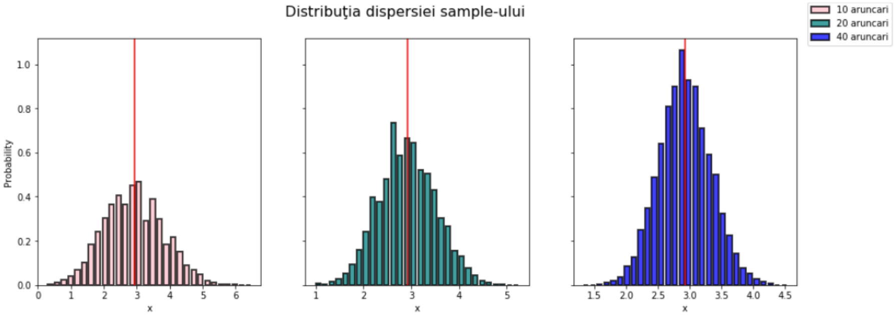
$$\begin{aligned} E[(x_1 - x_2)^2] &= E[x_1^2] - E[2x_1 x_2] + E[x_2^2] \\ &= (\sigma^2 + \mu^2) - 2\mu^2 + (\sigma^2 + \mu^2) \\ &= 2\sigma^2 \end{aligned}$$

- când selectăm uniform x_1, x_2 dintr-un sample de n valori, în $1/n$ cazuri $x_1 = x_2$, diferența va fi 0 independent de distribuția originală
- s-a calculat doar pentru proporția $1 - 1/n$, adică rezultatul trebuie înmulțit cu $\frac{n}{n-1}$
- `https://en.wikipedia.org/wiki/Bessel%27s_correction`

Distribuția dispersiei sample-urilor

```
def roll(n):
    r = np.random.randint(1, high=7, size=(10000, n))

    # dispersia se calculeaza ca mean(abs(x - x.mean())**2), unde
    # mean(x) = x.sum() / N, cu N numarul de sample-uri
    return np.var(r, axis=1, ddof=1)

fig, (ax1, ax2, ax3) = plt.subplots(1, 3, sharey=True, figsize=(15, 5))
kwargs = dict(rwidth=0.7, density=True, alpha=0.75, ec='k', linewidth=2)
[ax_.set_xlabel('x') for ax_ in [ax1, ax2, ax3, ax4]]
[ax_.axvline(2.92, c='r') for ax_ in [ax1, ax2, ax3, ax4]]
ax1.set_ylabel('Probability')

ax1.hist(roll(10), 30, **kwargs, facecolor='pink')
ax2.hist(roll(20), 30, **kwargs, facecolor='teal')
ax3.hist(roll(40), 30, **kwargs, facecolor='blue')

fig.legend(['10 aruncari', '20 aruncari', '40 aruncari'])
fig.suptitle('Distribuţia dispersiei sample-ului', fontsize=16)
plt.show()
```

Distribuția dispersiei sample-urilor (2)



- rulați de mai multe ori, cu și fără corecția Bessel

Distribuția sample means

- media pentru fiecare sample extras din populație este ea însăși o variabilă aleatoare
- această medie dă propria populație caracterizată de (medie, dispersie)
- media sample mean-urilor este aceeași ca media populației: $E[\bar{X}] = \mu$
- dispersia acestor medii este dispersia populației originale împărțită la n: $Var(\bar{X}) = \sigma^2/n$
- dispersia mediilor tine la zero pe măsură ce acumulăm mai multe date (distribuția sample mean-urilor se 'ascute')

Distribuția sample means: media

$$E(\bar{X}) = E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)$$

$$E(\bar{X}) = \frac{1}{n}E(X_1 + X_2 + \cdots + X_n)$$

$$E(\bar{X}) = \frac{1}{n}\left[E(X_1) + E(X_2) + \cdots + E(X_n)\right]$$

$$E(\bar{X}) = \frac{1}{n}\left[\mu + \mu + \cdots + \mu\right]$$

$$E(\bar{X}) = \frac{1}{n}n\mu = \mu$$

Distribuția sample means: dispersia

$$\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n}$$

$$Var(\bar{X}) = Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)$$

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 Var(X_1+X_2+\cdots+X_n)$$

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 [Var(X_1)+Var(X_2)+\cdots+Var(X_n)]$$

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 \left[\sigma^2+\sigma^2+\cdots+\sigma^2\right]$$

$$Var(\bar{X}) = \left(\frac{1}{n}\right)^2 n\sigma^2 = \frac{\sigma^2}{n}$$

Standard error of the mean

$$Var(\bar{X}) = \sigma^2/n$$

- deviația standard a unei variabile statistice e denumită standard error
- în cazul nostru vorbim de 'standard error of the mean'
- descriem în acest fel variabilitatea mediei
- populație cu media μ și dispersia σ^2
- sample variance S^2 estimează dispersia σ^2
- distribuția lui S^2 e centrată în jurul lui σ^2, din ce în ce mai concentrată cu cât numărul de observații crește
- dispersia sample mean este σ^2/n
- estimatorul dispersiei este S^2/n
- atunci estimatorul deviației standard este $S/\sqrt{n}$
- aceasta, 'standard error of the mean', descrie variabilitatea mediilor sample-urilor random de mărime n

Exemplu: distribuția mediilor unei distribuții normale

- într-o distribuție normală, deviația standard este 1, dispersia tot 1
- media a n deviații standard va avea deviația standard $1/\sqrt{n}$

```
>> nosim = 10000
>> n = 10
>> r = np.random.randn(nosim, n)
>> r = np.mean(r, axis=1)
>> print(np.std(r))
>> print(1/np.sqrt(n))

0.3146968499857521
0.31622776601683794
```

Exemplu: distribuția mediilor unei distribuții binomiale (monedă)

- într-o distribuție binomială, dispersia este $p(1-p) = 0.25$
- media a n deviații standard va avea deviația standard $0.5/\sqrt{n}$

```
>> nosim = 10000
>> n = 10
>> r = np.random.randint(low=0, high=2, size=(nosim, n))
>> r = np.mean(r, axis=1)
>> print(np.std(r))
>> print(0.5/np.sqrt(n))

0.1558563120313066
0.15811388300841897
```

Recapitulare

	Population (parametru)	Sample (statistic)
mean	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
variance	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1}$
mean of the sample means	$E[\bar{X}] = \mu$	
standard error of the mean, squared	$Var(\bar{X}) = \frac{\sigma^2}{n}$	

Recapitulare (2)

- sample variance S^2 este un estimator al population variance σ^2
- distribuția sample variance este centrată pe ceea ce estimează (unbiased)
- devine mai concentrată pe măsură ce se colectează date mai multe
- dispersia sample mean este dispersia populației împărțită la n $Var(\bar{X}) = \sigma^2/n$
- standard error of the mean este $\sigma/\sqrt{n}$
- aceste cantități descriu variabilitatea mediilor sample-urilor scoase din populație
- aceste relații sunt informative asupra gradului de reprezentativitate al mediei sample-ului pe care l-am extras

1 Expected values

2 Variabilitate

3 Anexă

Anexă: valoarea așteptată a produsului a două variabile independente

$$
\begin{aligned}
E[xy] &= \sum_{x,y} xyPr[x = X \text{ and } y = Y] \\
&= \sum_{x,y} xyPr[x = X]Pr[y = Y] \\
&= \sum_{x} \sum_{y} xyPr[x = X]Pr[y = Y] \\
&= \sum_{x} \left(xPr[x = X] \sum_{y} yPr[y = Y] \right) \\
&= \left(\sum_{x} xPr[x = X] \right) \left(\sum_{y} yPr[y = Y] \right) \\
&= E[x]E[y]
\end{aligned}
$$