The Variance, the Standard Deviation and the Mean Square

Probability and Statistics for Data Science

Carlos Fernandez-Granda

These slides are based on the book Probability and Statistics for Data Science by Carlos Fernandez-Granda, available for purchase here. A free preprint, videos, code, slides and solutions to exercises are available at https://www.ps4ds.net

Discrete random variable

The mean of a discrete random variable $\tilde{a}$ with range A is

$$\mathrm{E}\left[\tilde{a}\right] := \sum_{a \in A} a \, p_{\tilde{a}}\left(a\right)$$

if the sum converges

Continuous random variable

The mean of a continuous random variable $\tilde{a}$ is

$$\mathrm{E}\left[\tilde{a}\right] := \int_{a=-\infty}^{\infty} a f_{\tilde{a}}\left(a\right)\,\mathrm{d}a$$

if the integral converges

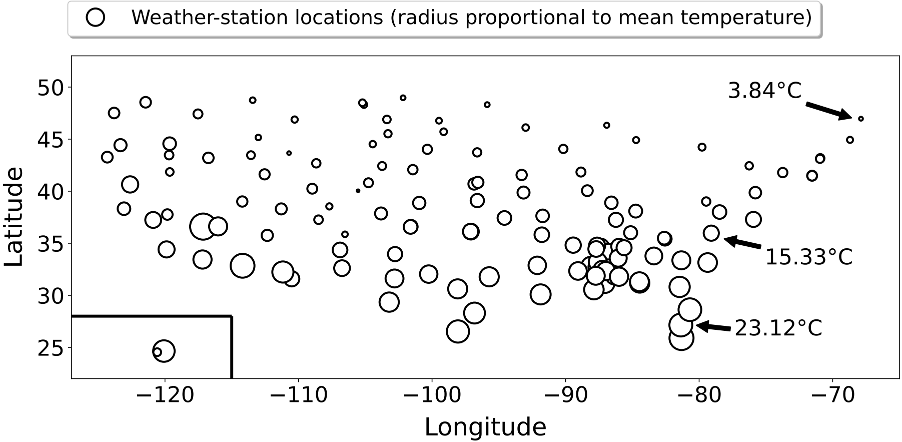
Sample mean

The sample mean of $X := \{x_1, x_2, \ldots, x_n\}$ is

$$m(X) := \frac{\sum_{i=1}^{n} x_i}{n}$$

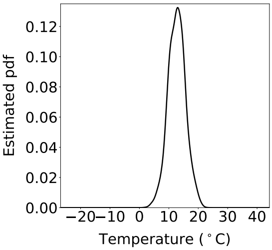
Temperature dataset

Hourly temperatures at 134 weather stations in the US

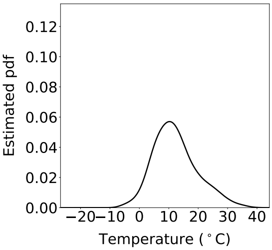


Same mean

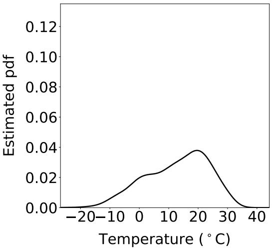
$m(X) = 12.7^{\circ}\mathrm{C}$



$m(X) = 12.3^{\circ}\mathrm{C}$



$m(X) = 12.7^{\circ}\mathrm{C}$



Challenge

Quantifying *magnitude* of deviation from the mean

Magnitude of a random variable?

Magnitude of real number a: $|a| = \sqrt{a^2}$

Euclidean length of vector x: $||x||_2 = \sqrt{\sum_{i=1}^{d} x[i]^2}$

Magnitude/energy of random variable $\tilde{a}$?

Mean square or second moment $\mathrm{E}\left[\tilde{a}^2\right]$

Mean squared error

The mean squared error (MSE) between an estimate $\tilde{e}$ and a random variable $\tilde{a}$ is

$$\mathrm{E}\left[(\tilde{a} - \tilde{e})^2\right]$$

Minimum MSE constant estimate

Best constant estimate of $\tilde{a}$?

$$\arg\min_{c \in \mathbb{R}} \mathrm{E}\left[(c - \tilde{a})^2\right] = \mathrm{E}[\tilde{a}]$$

$$\mathrm{MSE}(c) := \mathrm{E}\left[(c - \tilde{a})^2\right] = c^2 - 2c\mathrm{E}\left[\tilde{a}\right] + \mathrm{E}\left[\tilde{a}^2\right]$$

$$\mathrm{MSE}'(c) = 2(c - \mathrm{E}[\tilde{a}])$$

$$\mathrm{MSE}''(c) = 2$$

The mean $\mathrm{E}[\tilde{a}]$

Variance

Mean squared distance of a random variable to its mean

$$\begin{aligned}
\mathrm{Var}\left[\tilde{a}\right] &:= \mathrm{E}\left[\left(\tilde{a} - \mathrm{E}\left[\tilde{a}\right]\right)^2\right] \\
&= \mathrm{E}\left[\tilde{a}^2 - 2\tilde{a}\mathrm{E}\left[\tilde{a}\right] + \mathrm{E}\left[\tilde{a}\right]^2\right] \\
&= \mathrm{E}\left[\tilde{a}^2\right] - 2\mathrm{E}\left[\tilde{a}\right]\mathrm{E}\left[\tilde{a}\right] + \mathrm{E}\left[\tilde{a}\right]^2 \\
&= \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}(\tilde{a})^2
\end{aligned}$$

Standard deviation

The standard deviation $\sigma_{\tilde{a}}$ of $\tilde{a}$ is

$$\sigma_{\tilde{a}} := \sqrt{\mathrm{Var}\left[\tilde{a}\right]}$$

Sample variance

Dataset: $x_1, x_2, \ldots, x_n$

The sample variance is the average squared deviation from the sample mean

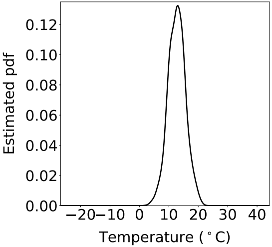
$$v(X) := \frac{\sum_{i=1}^{n}(x_i - m(X))^2}{n-1}$$

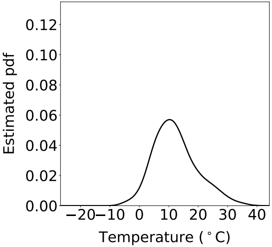
The sample standard deviation σ_X is the square root of the sample variance

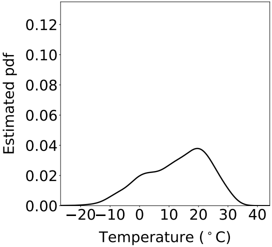
Same mean

$$m(X) = 12.7^\circ\mathrm{C} \qquad m(X) = 12.3^\circ\mathrm{C} \qquad m(X) = 12.7^\circ\mathrm{C}$$

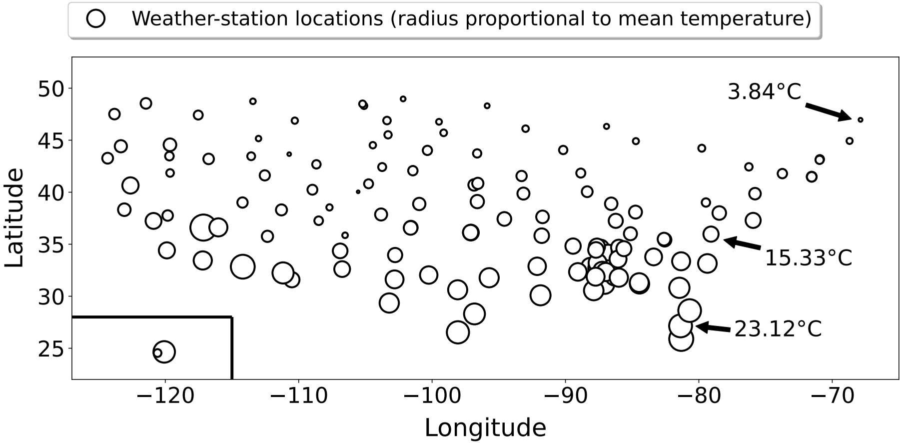
$$\sqrt{v(X)} = 2.9^\circ\mathrm{C} \qquad \sqrt{v(X)} = 7.5^\circ\mathrm{C} \qquad \sqrt{v(X)} = 10.6^\circ\mathrm{C}$$





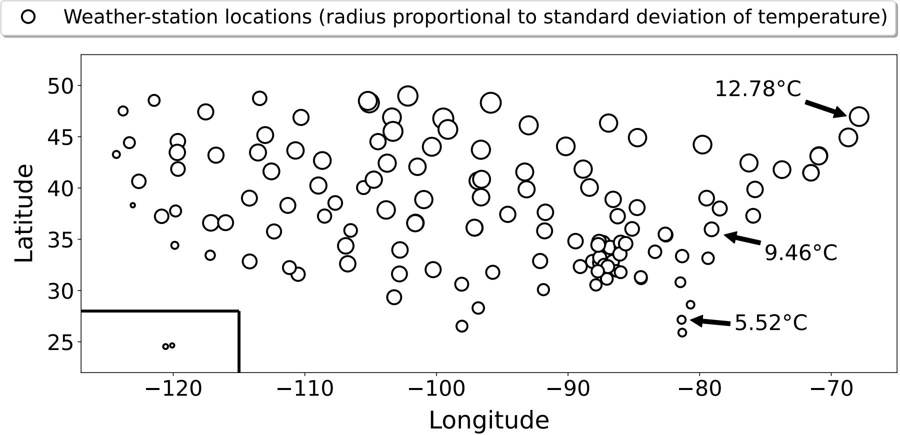


Means



Weather-station locations (radius proportional to mean temperature)
3.84°C
15.33°C
23.12°C
50
45
40
35
30
25
Latitude
−120
−110
−100
−90
−80
−70
Longitude

Standard deviations



Scaled, shifted variable

We scale a random variable by c_1 and shift it by c_2

Variance?

$$
\begin{aligned}
\operatorname{Var}\left[c_1\,\tilde{a}+c_2\right] &= \operatorname{E}\left[\left(c_1\,\tilde{a}+c_2-\operatorname{E}\left[c_1\,\tilde{a}+c_2\right]\right)^2\right] \\
&= \operatorname{E}\left[\left(c_1\,\tilde{a}+c_2-c_1\operatorname{E}\left[\tilde{a}\right]-c_2\right)^2\right] \\
&= c_1^2\operatorname{E}\left[\left(\tilde{a}-\operatorname{E}\left[\tilde{a}\right]\right)^2\right] \\
&= c_1^2\operatorname{Var}\left[\tilde{a}\right]
\end{aligned}
$$

Bernoulli random variable

$$\mathrm{E}\left[\tilde{a}\right] = \theta$$

$$\begin{aligned}\mathrm{E}\left[\tilde{a}^2\right] &= 0 \cdot p_{\tilde{a}}\left(0\right) + 1 \cdot p_{\tilde{a}}\left(1\right) \\ &= \theta\end{aligned}$$

$$\begin{aligned}\mathrm{Var}\left[\tilde{a}\right] &= \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}\left[\tilde{a}\right]^2 \\ &= \theta(1-\theta)\end{aligned}$$

Geometric random variable

$$\sum_{k=1}^{\infty} k\,\alpha^k = \frac{\alpha}{(1-\alpha)^2}$$

$$\sum_{k=1}^{\infty} k^2 \alpha^{k-1} = \frac{1+\alpha}{(1-\alpha)^3}$$

$$\begin{aligned}
\mathrm{E}\left[\tilde{a}^2\right] &= \sum_{k=1}^{\infty} k^2\, p_{\tilde{a}}(k) \\
&= \sum_{k=1}^{\infty} k^2\, \theta\, (1-\theta)^{k-1} \\
&= \frac{2-\theta}{\theta^2}
\end{aligned}$$

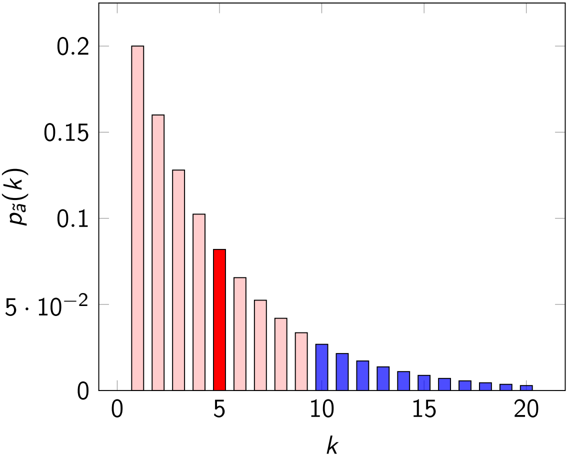
Geometric random variable

$$\mathrm{E}\left[\tilde{a}\right] = \frac{1}{\theta}$$

$$\mathrm{E}\left[\tilde{a}^2\right] = \frac{2-\theta}{\theta^2}$$

$$\begin{aligned}\mathrm{Var}\left[\tilde{a}\right] &= \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}\left[\tilde{a}\right]^2 \\ &= \frac{1-\theta}{\theta^2}\end{aligned}$$

Geometric, $\theta := 0.2$

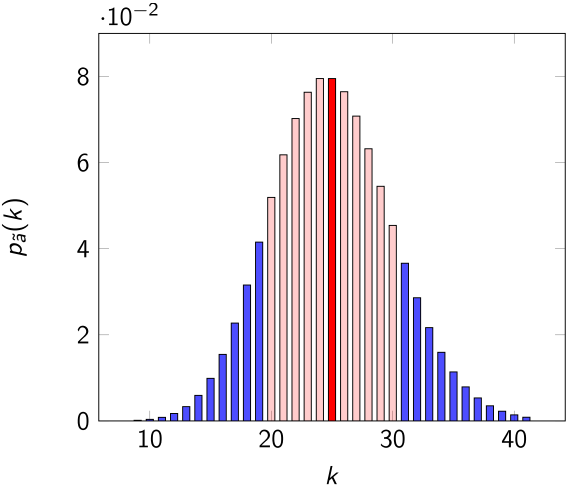


Poisson random variable

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$$

$$\begin{aligned}
\mathrm{E}\left[\tilde{a}^2\right] &= \sum_{k=1}^{\infty} k^2\, p_{\tilde{a}}\left(k\right) = \sum_{k=1}^{\infty} \frac{k^2 \lambda^k e^{-\lambda}}{k!} \\
&= \sum_{k=1}^{\infty} \frac{k \lambda^k e^{-\lambda}}{(k-1)!} \\
&= e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{(k-1)\,\lambda^k}{(k-1)!} + \frac{\lambda^k}{(k-1)!} \right) \\
&= e^{-\lambda} \left(\sum_{m=0}^{\infty} \frac{\lambda^{m+2}}{m!} + \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!} \right) \\
&= \lambda^2 + \lambda \\
\mathrm{Var}\left[\tilde{a}\right] &= \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}\left[\tilde{a}\right]^2 \; = \lambda
\end{aligned}$$

Poisson, $\lambda := 25$

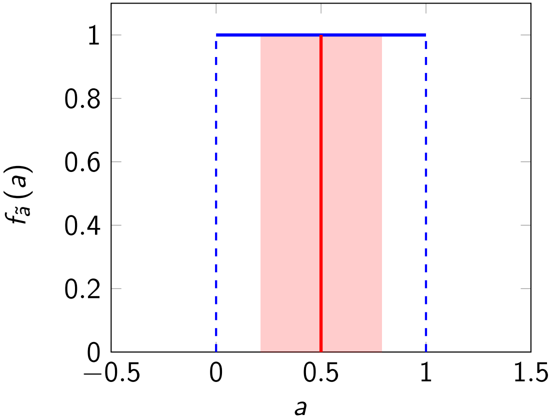


Uniform random variable

$$\begin{aligned}
\mathrm{E}\left[\tilde{u}\right] &= \frac{a+b}{2} \\
\mathrm{E}\left[\tilde{u}^2\right] &= \int_{u=-\infty}^{\infty} u^2 f_{\tilde{a}}\left(u\right)\, \mathrm{d}u \\
&= \int_{u=a}^{b} \frac{u^2}{b-a}\, \mathrm{d}u \\
&= \frac{b^3-a^3}{3\left(b-a\right)} \\
&= \frac{a^2+ab+b^2}{3}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}\left[\tilde{u}\right] &= \mathrm{E}\left[\tilde{u}^2\right] - \mathrm{E}\left[\tilde{u}\right]^2 \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

Uniform in $[0, 1]$

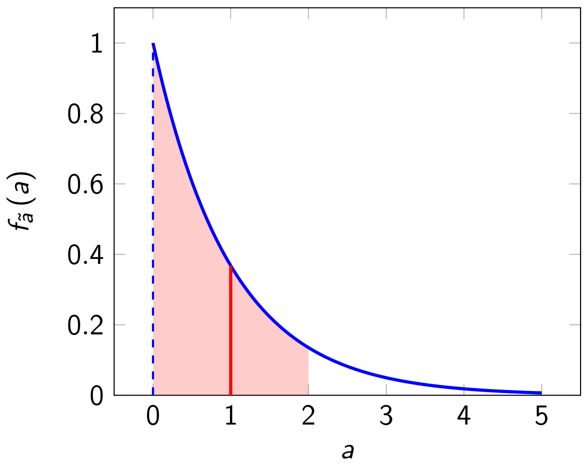


Exponential random variable

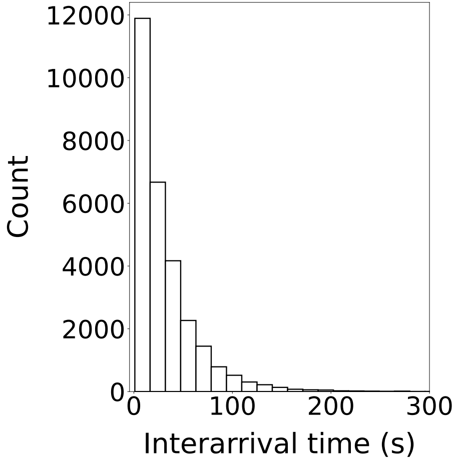
$$
\begin{aligned}
\mathrm{E}\left[\tilde{a}\right] &= \frac{1}{\lambda} \\
\mathrm{E}\left[\tilde{a}^2\right] &= \int_{a=-\infty}^{\infty} a^2 f_{\tilde{a}}\left(a\right)\,\mathrm{d}a \\
&= \int_{a=0}^{\infty} a^2 \lambda e^{-\lambda a}\,\mathrm{d}a \\
&= a^2 e^{-\lambda a}\Big]_0^{\infty} + 2\frac{1}{\lambda}\int_0^{\infty} a\lambda e^{-\lambda a}\,\mathrm{d}a \\
&= \frac{2}{\lambda^2}
\end{aligned}
$$

$$
\mathrm{Var}\left[\tilde{a}\right] = \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}\left[\tilde{a}\right]^2 = \frac{1}{\lambda^2}
$$

Exponential, $\lambda := 1$



Call center data



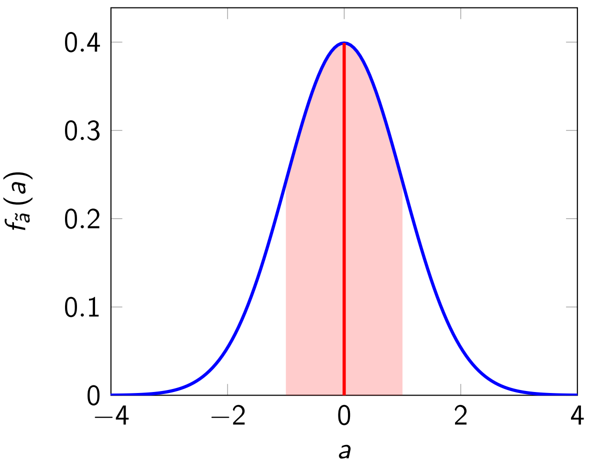
Sample mean = 30.8 Sample standard deviation = 33.6

Gaussian random variable

Change of variables $t = (a - \mu)/\sigma$

$$
\begin{aligned}
\mathrm{E}[\tilde{a}] &= \mu \\
\mathrm{E}\left[\tilde{a}^2\right] &= \int_{a=-\infty}^{\infty} a^2 f_{\tilde{a}}(a)\, \mathrm{d}a \\
&= \int_{a=-\infty}^{\infty} \frac{a^2}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}}\, \mathrm{d}a \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}} \mathrm{d}t + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} t e^{-\frac{t^2}{2}} \mathrm{d}t \\
&\quad + \frac{\mu^2}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t \\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(t^2 e^{-\frac{t^2}{2}}]_{-\infty}^{\infty} + \int_{t=-\infty}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t \right) + \mu^2 \\
&= \sigma^2 + \mu^2 \\
\mathrm{Var}[\tilde{a}] &= \mathrm{E}\left[\tilde{a}^2\right] - \mathrm{E}[\tilde{a}]^2 = \sigma^2
\end{aligned}
$$

Gaussian $\mu := 0$, $\sigma^2 := 1$



What have we learned

Definition of mean square, variance and standard deviation

Variance of popular parametric models