Multivariate Random Variables

Joint Laws, Independence, Dependence, and the Multivariate Normal

Multivariate Random Variables

Definition — Random Vector

A **multivariate random variable** (or random vector) is a function

$$\mathbf{X} : \Omega \to \mathbb{R}^n$$

where Ω is the sample space and $n \geq 2$.

We write $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$, where each X_i is a univariate random variable.

Examples

- (X, Y) — height and weight of a person
- (X_1, X_2, X_3) — temperature, humidity, pressure
- The stock prices of several companies at the close of trading

Joint Cumulative Distribution Function

Definition — Joint CDF

The **joint cumulative distribution function** of $\mathbf{X} = (X_1, \ldots, X_n)$ is

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

Properties:

(1) $0 \leq F_{\mathbf{X}}(x_1, \ldots, x_n) \leq 1$

(2) $F_{\mathbf{X}}$ is non-decreasing in **each** argument

(3) $F_{\mathbf{X}}(-\infty, \ldots, -\infty) = 0$ and $F_{\mathbf{X}}(\infty, \ldots, \infty) = 1$

(4) Right-continuous in each argument

Why the joint CDF matters

Exactly as in one dimension, it is the one object that always exists — discrete, continuous, or mixed.

Joint Probability Mass Function

For **discrete** random vectors $\mathbf{X} = (X_1, \ldots, X_n)$:

Definition — Joint PMF

$$p_{\mathbf{X}}(x_1, \ldots, x_n) = P(X_1 = x_1, \ldots, X_n = x_n)$$

Properties:

(1) $p_{\mathbf{X}}(x_1, \ldots, x_n) \geq 0$ for all $(x_1, \ldots, x_n)$

(2) $\sum_{x_1} \cdots \sum_{x_n} p_{\mathbf{X}}(x_1, \ldots, x_n) = 1$

(3) $F_{\mathbf{X}}(x_1, \ldots, x_n) = \sum_{t_1 \leq x_1} \cdots \sum_{t_n \leq x_n} p_{\mathbf{X}}(t_1, \ldots, t_n)$

Joint Probability Density Function

For **continuous** random vectors $\mathbf{X} = (X_1, \ldots, X_n)$:

Definition — Joint PDF

The joint density $f_{\mathbf{X}}$ satisfies

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{\mathbf{X}}(t_1, \ldots, t_n)\, dt_n \cdots dt_1$$

Properties:

(1) $f_{\mathbf{X}}(x_1, \ldots, x_n) \geq 0$ for all $(x_1, \ldots, x_n)$

(2) $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \ldots, x_n)\, dx_n \cdots dx_1 = 1$

(3) $$f_{\mathbf{X}}(x_1, \ldots, x_n) = \frac{\partial^n F_{\mathbf{X}}(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}$$

Marginals and Conditionals

Marginal Distributions

Definition — Marginal Distribution

The **marginal distribution** of X_i is the distribution of X_i alone, obtained by summing or integrating the others away.

Discrete case (PMF):

$$p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} p_{\mathbf{X}}(x_1, \ldots, x_n)$$

Continuous case (PDF):

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{X}}(x_1, \ldots, x_n)\, dx_1 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_n$$

Joint Marginal Distributions

For any **subset** of the variables, e.g. (X_1, X_2) from (X_1, X_2, X_3):

Discrete:

$$p_{X_1,X_2}(x_1, x_2) = \sum_{x_3} p_{X_1,X_2,X_3}(x_1, x_2, x_3)$$

Continuous:

$$f_{X_1,X_2}(x_1, x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1, x_2, x_3)\, dx_3$$

One-way street

The joint distribution determines **all** marginals. The marginals do **not** determine the joint — as the live demo two slides on will show.

Conditional Distributions

Given $\mathbf{X} = (X_1, \ldots, X_k)$ and $\mathbf{Y} = (Y_1, \ldots, Y_m)$:

Conditional Densities

If $\mathbf{X}$ and $\mathbf{Y}$ are dependent, learning $\mathbf{Y} = \mathbf{y}$ updates the uncertainty about $\mathbf{X}$:

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{p_{\mathbf{Y}}(\mathbf{y})} \quad \text{or} \quad f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x}, \mathbf{y})}{f_{\mathbf{Y}}(\mathbf{y})}$$

Properties:

- Every property of a probability distribution still holds
- If $\mathbf{X}$ and $\mathbf{Y}$ are independent, $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x})$ — conditioning tells you nothing

Joint, Marginal, Conditional, Live

The joint sits in the middle; the margins are what you get by *integrating out*, the slice is what you get by *conditioning*.

Correlation ρ

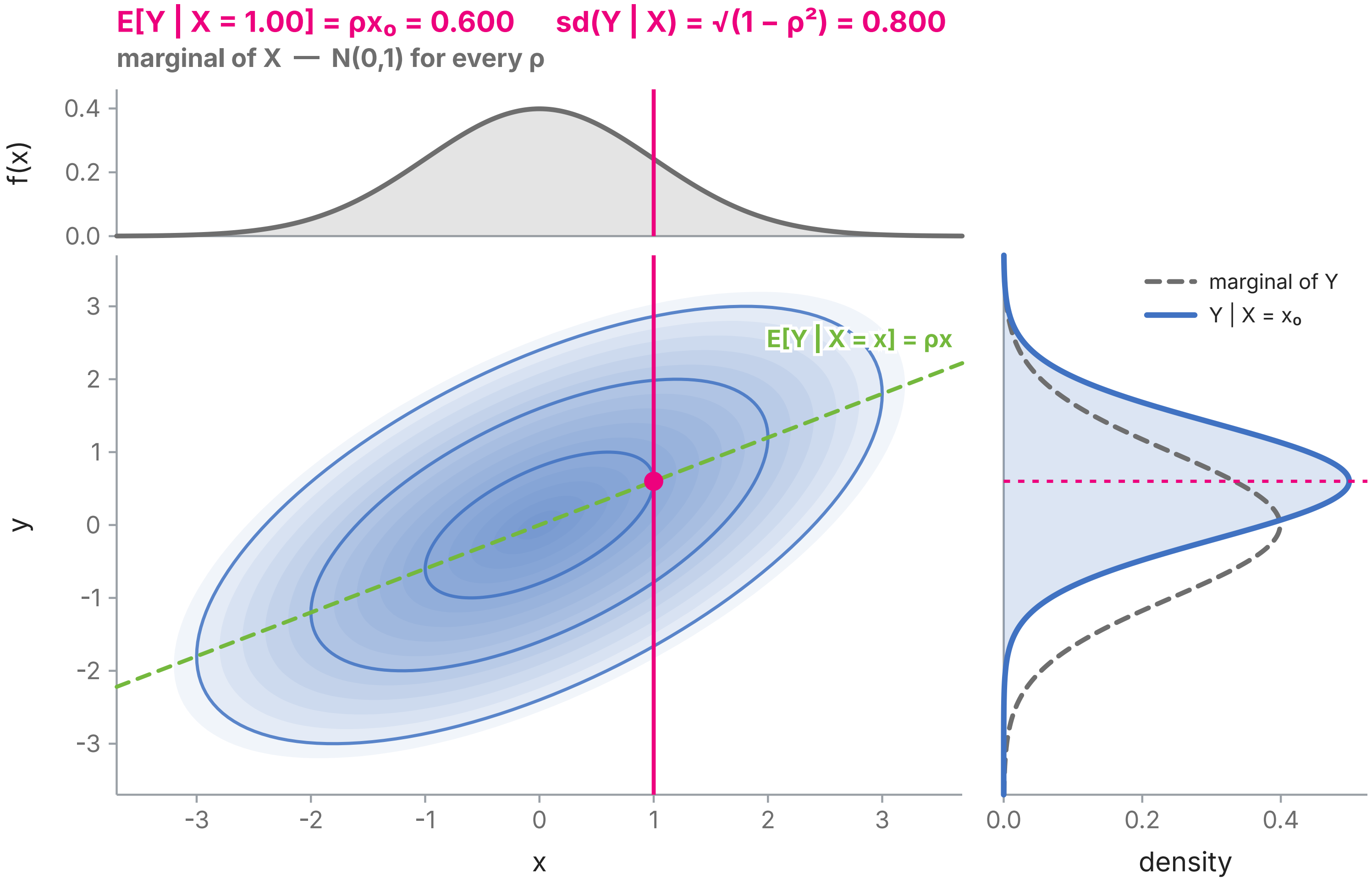
0.6

Condition on X = x_0

1

Standard bivariate normal. The top curve is the marginal of X; the right panel shows the marginal of Y (grey) against the conditional $Y \mid X = x_0$ (blue).

The marginal of Y is $N(0,1)$ **whatever** ρ is — only the joint changes.



Conditional Moments

Conditional Moments

$$E[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}] = \int \mathbf{x}\, f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}$$

$$\mathrm{Var}[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}] = E[\mathbf{X}\mathbf{X}^T \mid \mathbf{Y} = \mathbf{y}] - E[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}]\, E[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}]^T$$

If $\mathbf{X}$ and $\mathbf{Y}$ are independent, then for every $\mathbf{y}$

- $E[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}] = E[\mathbf{X}]$ and $\mathrm{Var}[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}] = \mathrm{Var}[\mathbf{X}]$
- more generally $E[g(\mathbf{X}) \mid \mathbf{Y} = \mathbf{y}] = E[g(\mathbf{X})]$ for any g

Law of Total Expectation

$$E[\mathbf{X}] = E_{\mathbf{Y}}\left[E_{\mathbf{X}|\mathbf{Y}}[\mathbf{X} \mid \mathbf{Y}]\right] = \int_{\mathbf{Y}} \left[\int_{\mathbf{X}|\mathbf{Y}} \mathbf{x}\, f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})\, d\mathbf{x}\right] f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}$$

Independence

Independence of Random Variables

Definition — Mutual Independence

$X_1, \ldots, X_n$ are **mutually independent** if

$$F_{\mathbf{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdots F_{X_n}(x_n)$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$.

Equivalent conditions:

- **Discrete:** $p_{\mathbf{X}}(x_1, \ldots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n)$
- **Continuous:** $f_{\mathbf{X}}(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$

Under independence the joint **is** the product of the marginals — the one case where the marginals do determine the joint.

Pairwise versus Mutual Independence

Pairwise Independence

$X_1, \ldots, X_n$ are **pairwise independent** if X_i and X_j are independent for every $i \neq j$.

Mutual Independence

$X_1, \ldots, X_n$ are **mutually independent** if **every** subset $\{X_{i_1}, \ldots, X_{i_k}\}$ consists of independent variables.

Key Point

Mutual independence is a **strictly stronger** condition.

$$\text{Mutual independence} \implies \text{Pairwise independence}$$

$$\text{Pairwise independence} \not\Longrightarrow \text{Mutual independence}$$

Classic Example: Three Coin Tosses

Toss a fair coin three times and define

$$
\begin{aligned}
A &: \text{the first coin is heads} \\
B &: \text{the second coin is heads} \\
C &: \text{the first two coins agree}
\end{aligned}
$$

Sample space: $\{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$, all equally likely.

- $A = \{HHH, HHT, HTH, HTT\}$, so $P(A) = 1/2$
- $B = \{HHH, HHT, THH, THT\}$, so $P(B) = 1/2$
- $C = \{HHH, HHT, TTH, TTT\}$, so $P(C) = 1/2$

Every **pair** multiplies correctly: $P(A \cap B) = P(A \cap C) = P(B \cap C) = 1/4$.

But Not Mutually Independent

All three at once:

$$A \cap B \cap C = \{HHH, HHT\}, \qquad P(A \cap B \cap C) = \frac{2}{8} = \frac{1}{4}$$

But mutual independence would require

$$P(A \cap B \cap C) = P(A)P(B)P(C) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$$

Conclusion

Since $\frac{1}{4} \neq \frac{1}{8}$, the three events are **pairwise independent** but **not mutually independent**.

Pairwise versus Mutual, Live

Every pair checks out. The triple does not. Step through them.

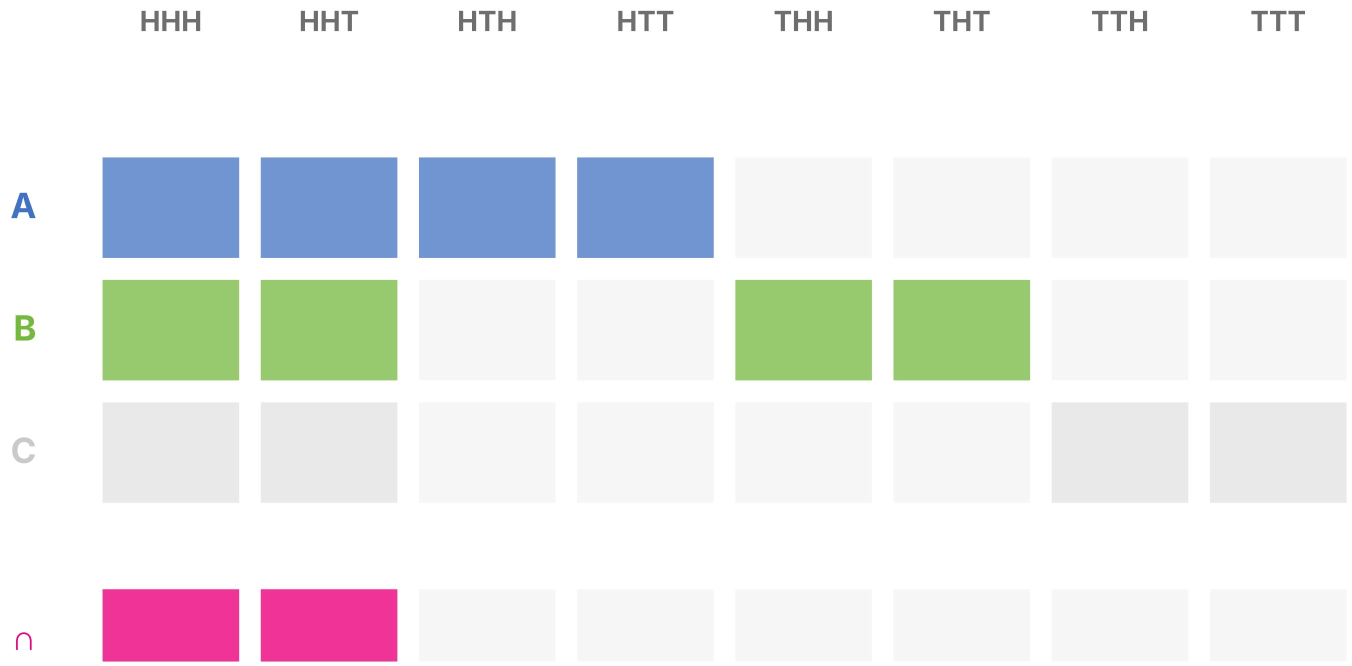
Check

A ∩ B

Eight equally likely outcomes. A column is shaded in a row when that outcome belongs to the event. The bottom row is the intersection of everything you selected — count its squares and compare with the product of the probabilities.

$P(A \cap B) = 2/8 = 0.250$ $P(A) \cdot P(B) = 0.250$

the two agree — this collection is independent



Independence: Summary

- **Pairwise independence:** every *pair* of variables is independent
- **Mutual independence:** every *subset* of variables is independent
- Mutual independence is strictly stronger than pairwise independence
- Knowing that variables are pairwise independent is **not** enough to conclude mutual independence

Where this bites

The distinction matters wherever a product of probabilities is taken over more than two things at once — likelihoods, hashing arguments, and independence assumptions in machine learning models.

Dependence

Dependence of Random Variables

Mean Dependence

$$E[Y \mid X = x] \neq E[Y]$$

The conditional expectation of Y depends on the value of X.

Tail Dependence

Extreme values occur together. For the upper tail,

$$\lambda_U = \lim_{\alpha \to 1^-} P\big(Y > Q_Y(\alpha) \mid X > Q_X(\alpha)\big)$$

Generally: Copulas

Sklar's Theorem. Any joint distribution can be written as

$$F_{X,Y}(x, y) = C\big(F_X(x), F_Y(y)\big)$$

where the copula C carries the entire dependence structure — separately from the marginals.

Types of Dependence

Types of dependence:

- **Linear:** $Y = aX + b + \varepsilon$ — captured by correlation
- **Monotonic:** Y increases (or decreases) with X, but not linearly
- **Nonlinear:** $Y = X^2$, $Y = \sin(X)$, U-shaped relationships
- **Conditional:** the dependence itself varies across the support

Popular alternative measures:

- **Pearson's** r — correlation coefficient, captures *linear* dependence
- **Spearman's** ρ — correlation of *ranks*, captures *monotonic* dependence
- **Kendall's** τ — based on concordance of pairs, also monotonic
- **Mutual information** — $I(X,Y) = 0$ if and only if independence

Same Marginals, Different Joints, Live

Every panel below has **exactly the same** marginals. Only the copula changes.

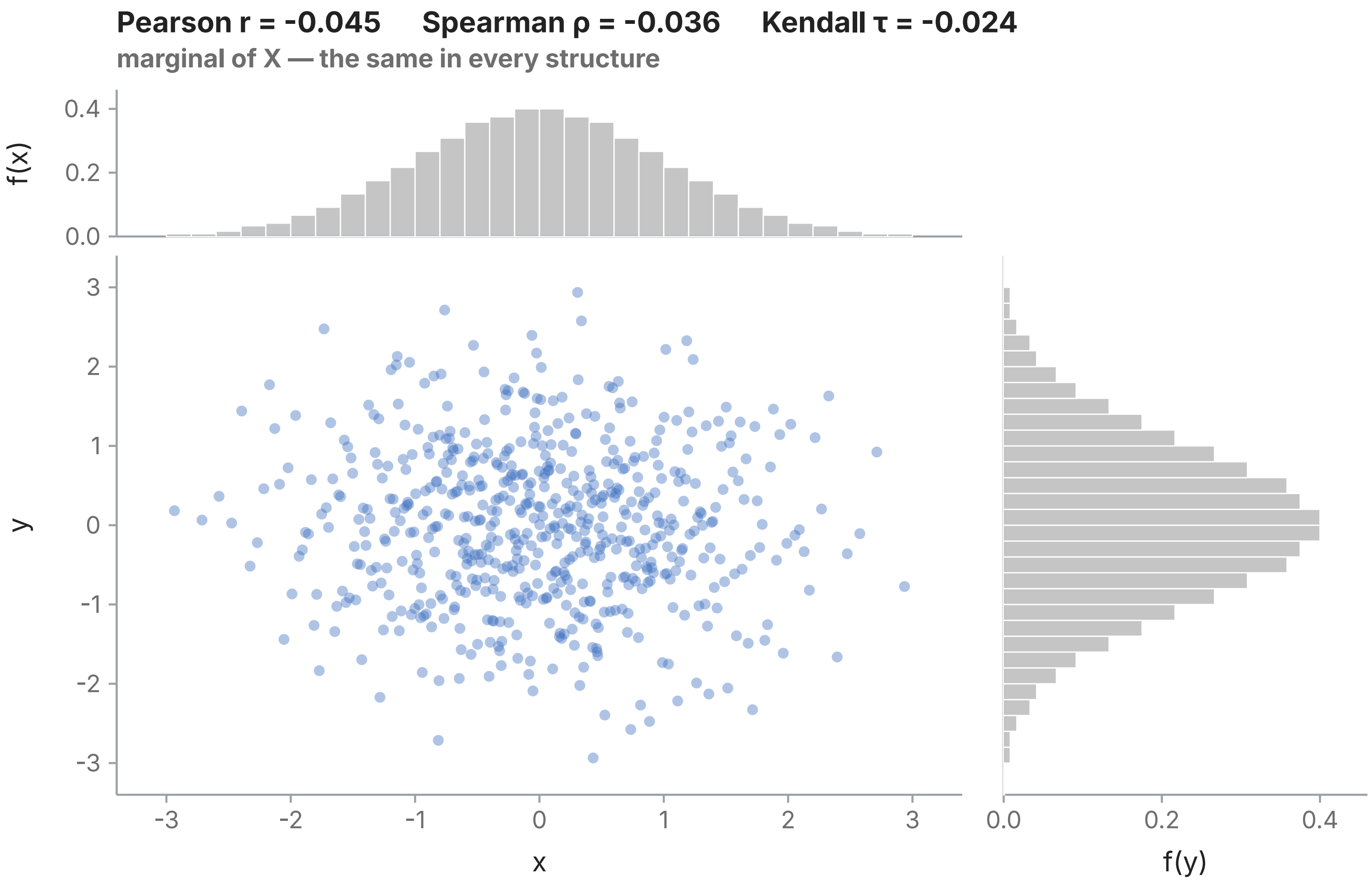
Dependence structure

Independent

New sample

Both coordinates are pushed through their own rank transform, so each margin is exactly the same set of normal scores in every panel. Anything you see change is dependence, not marginals.

Watch the **U-shape**: Pearson, Spearman and Kendall are all near zero, yet Y is almost a function of X.



Covariance and Correlation

Covariance

Definition — Covariance

$$\mathrm{Cov}(X,Y) = E\big[(X - E[X])(Y - E[Y])\big] = E[XY] - E[X]E[Y]$$

Properties:

- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$ and $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(aX+b, cY+d) = ac \cdot \mathrm{Cov}(X,Y)$ — it is **not** scale-free
- If X and Y are independent then $\mathrm{Cov}(X,Y) = 0$
- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$

The converse fails

$\mathrm{Cov}(X,Y) = 0$ does **not** imply independence — the U-shaped panel of the last demo is a counterexample.

Correlation

Definition — Correlation Coefficient

$$\rho(X,Y) = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

provided $\mathrm{Var}(X), \mathrm{Var}(Y) > 0.$

Properties:

- $-1 \le \rho(X,Y) \le 1$
- $|\rho(X,Y)| = 1$ **if and only if** $Y = aX + b$ for constants $a \neq 0, b$
- $\rho(X,Y) = 0$ if X and Y are independent
- $\rho(aX+b, cY+d) = \mathrm{sign}(ac) \cdot \rho(X,Y)$ — invariant up to sign

Covariance Matrix

For a random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$:

Definition — Covariance Matrix

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) = E\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right], \qquad \boldsymbol{\mu} = E[\mathbf{X}]$$

with (i, j)-th element $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$.

Properties:

- $\boldsymbol{\Sigma}$ is symmetric: $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$
- $\boldsymbol{\Sigma}$ is positive semidefinite
- Diagonal elements are variances: $\Sigma_{ii} = \mathrm{Var}(X_i)$

The Covariance Ellipse, Live

Σ *is* an ellipse. Its axes are the eigenvectors; the axis lengths are $\sqrt{\lambda_i}$.

σ_x

1

σ_y

0.7

ρ

0.6

Re-scale X by c

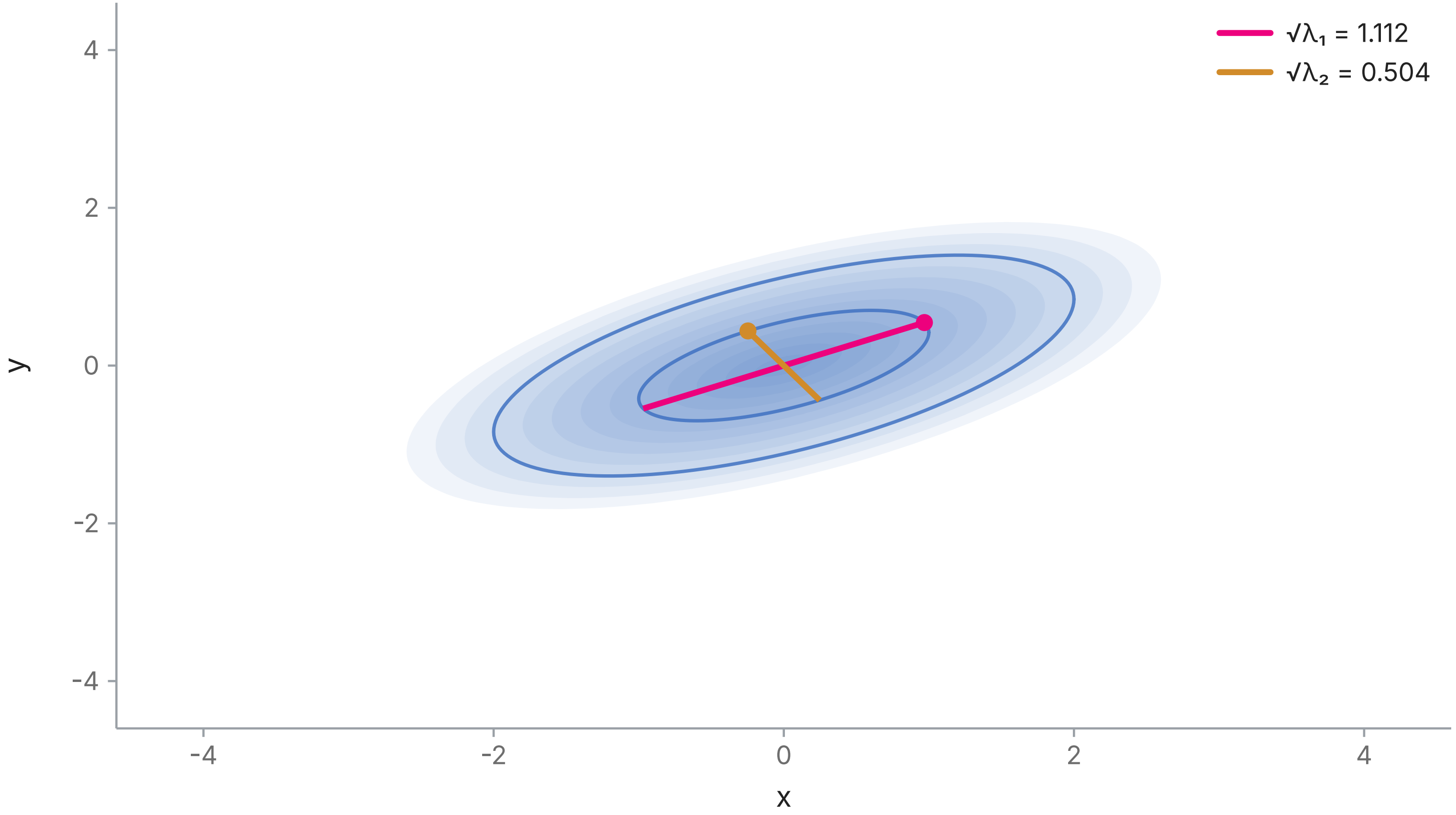
1

Changing the *units* of X multiplies Cov by c but leaves ρ alone — that is exactly why correlation, not covariance, is the comparable number.

Push $|\rho| \to 1$ and the ellipse collapses onto a line: $\det \Sigma \to 0$, the boundary of positive semidefiniteness.

Σ = [1.000, 0.420 ; 0.420, 0.490]   det Σ = 0.3136

Cov(X,Y) = 0.420 — Corr(X,Y) = 0.600 (unchanged by c)



The Multivariate Normal

Multivariate Normal Distribution

Definition — Multivariate Normal

$\mathbf{X} = (X_1, \ldots, X_n)^T$ follows $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if its density is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}$ positive definite.

Parameters: $\boldsymbol{\mu} = E[\mathbf{X}]$ (mean vector) and $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X})$ (covariance matrix).

Notation: $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

The exponent is a quadratic form, so the level sets of the density are exactly the ellipses of the previous demo.

Properties of the Multivariate Normal

(1) **Linear combinations:** if $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$\mathbf{AX} + \mathbf{b} \sim \mathcal{N}_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

(2) **Marginals:** every marginal distribution is normal

(3) **Conditionals:** every conditional distribution is normal

(4) **Independence:** for the multivariate normal, uncorrelated $\Rightarrow$ independent

(5) **Quadratic forms:** $(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_n$

Property 4 is special

In general zero correlation says nothing about independence. Inside the multivariate normal family it says everything.

Bivariate Normal Distribution

For $(X, Y) \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}$$

the density is

$$f_{X,Y}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{Q}{2(1-\rho^2)}\right)$$

Transformations

Transformations of Random Vectors

Given $\mathbf{X}$ with a known distribution, find the distribution of $\mathbf{Y} = \mathbf{g}(\mathbf{X})$, where $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$.

Method 1 — CDF Method

$$F_{\mathbf{Y}}(\mathbf{y}) = P(\mathbf{g}(\mathbf{X}) \leq \mathbf{y}) = P\big(\mathbf{X} \in \{\mathbf{x} : \mathbf{g}(\mathbf{x}) \leq \mathbf{y}\}\big)$$

Method 2 — Jacobian Method (continuous)

If $\mathbf{g}$ is one-to-one with inverse $\mathbf{h}$, then

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\big(\mathbf{h}(\mathbf{y})\big) \cdot \big|J_{\mathbf{h}}(\mathbf{y})\big|$$

where $J_{\mathbf{h}}$ is the Jacobian determinant of $\mathbf{h}$.

Example: Linear Transformation

Problem. $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$, with $\mathbf{A}$ of size $m \times n$ and $\mathbf{b} \in \mathbb{R}^m$.

Solution.

$$\mathbf{Y} \sim \mathcal{N}_m\big(\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\ \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T\big)$$

Special cases:

- **Sum:** $X + Y$ where (X, Y) is bivariate normal
- **Difference:** $X - Y$ where (X, Y) is bivariate normal
- **Sample mean:** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ for i.i.d. normal X_i

Linear Transformations, Live

Rotate and stretch a normal cloud. It stays normal — only $\mathbf{A\Sigma A}^T$ changes.

Source ρ

0.7

Rotate A by (°)

0

Stretch first axis

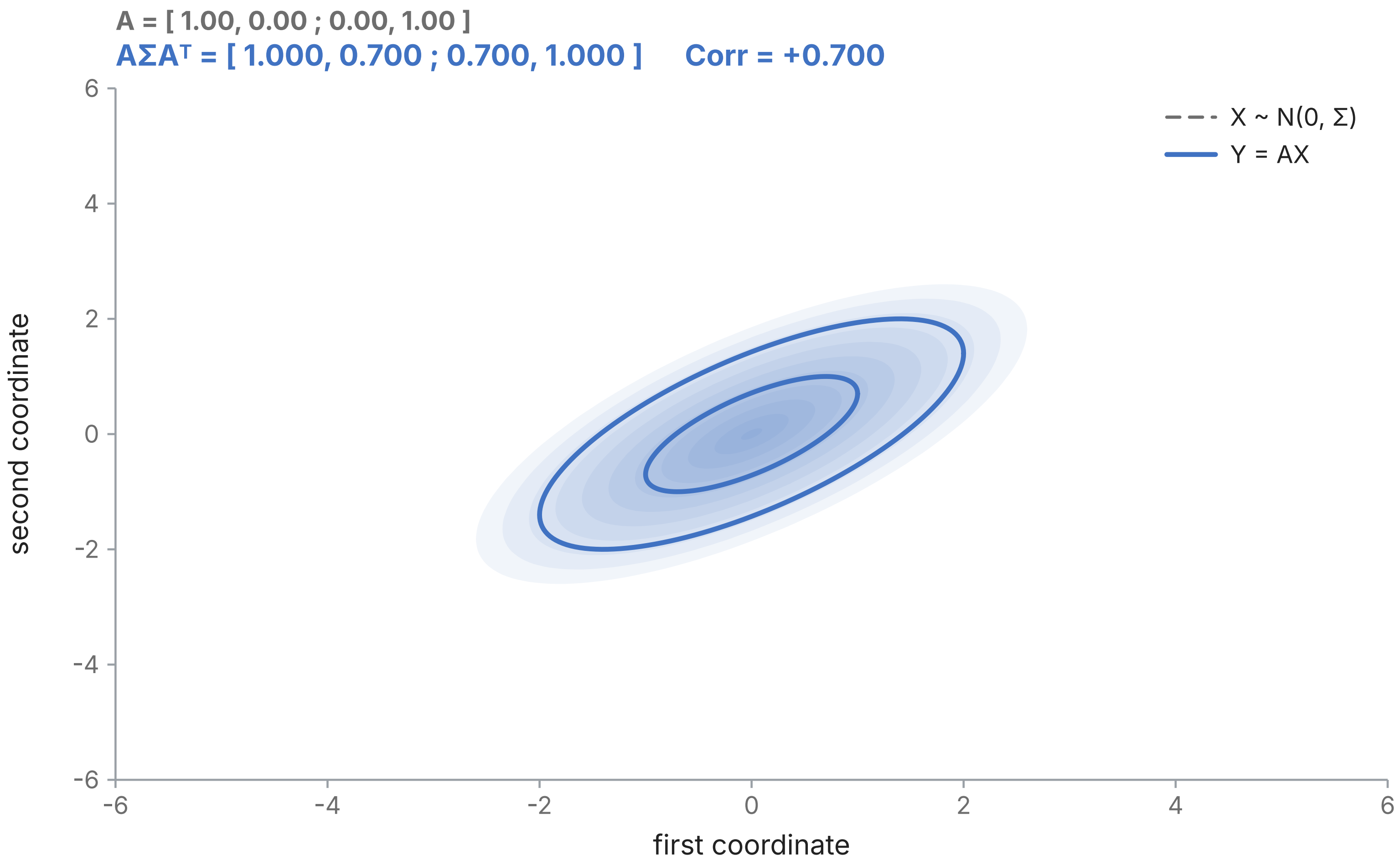
1

Stretch second axis

1

$\mathbf{A} = R(\theta)\,\mathrm{diag}(a, b)$ applied to $\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma})$ with unit variances and correlation ρ.

Set the rotation to 45° with equal stretches to see a sum-and-difference transform: $X + Y$ and $X - Y$ become **uncorrelated**.



Order Statistics

Order Statistics

Given n i.i.d. random variables $X_1, \ldots, X_n$ with CDF F and density f:

> **Definition — Order Statistics**
>
> The **order statistics** $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the sorted values of $X_1, \ldots, X_n$.

Joint density of all order statistics:

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n!\, f(x_1) \cdots f(x_n), \qquad x_1 \leq \cdots \leq x_n$$

Density of the k-th order statistic:

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,(n-k)!} \left[F(x)\right]^{k-1} \left[1 - F(x)\right]^{n-k} f(x)$$

Special Order Statistics

Minimum

$$X_{(1)} = \min\{X_1, \ldots, X_n\}$$

$$F_{X_{(1)}}(x) = 1 - [1 - F(x)]^n$$

$$f_{X_{(1)}}(x) = n[1 - F(x)]^{n-1} f(x)$$

Maximum

$$X_{(n)} = \max\{X_1, \ldots, X_n\}$$

$$F_{X_{(n)}}(x) = [F(x)]^n$$

$$f_{X_{(n)}}(x) = n[F(x)]^{n-1} f(x)$$

Range: $R = X_{(n)} - X_{(1)}$ — the spread of the whole sample, and itself a random variable.

Order Statistics, Live

n draws, sorted. Each faint curve is one order statistic; the bold one is the k-th.

Parent distribution

Uniform(0,1)

Sample size n

8

Which one, k

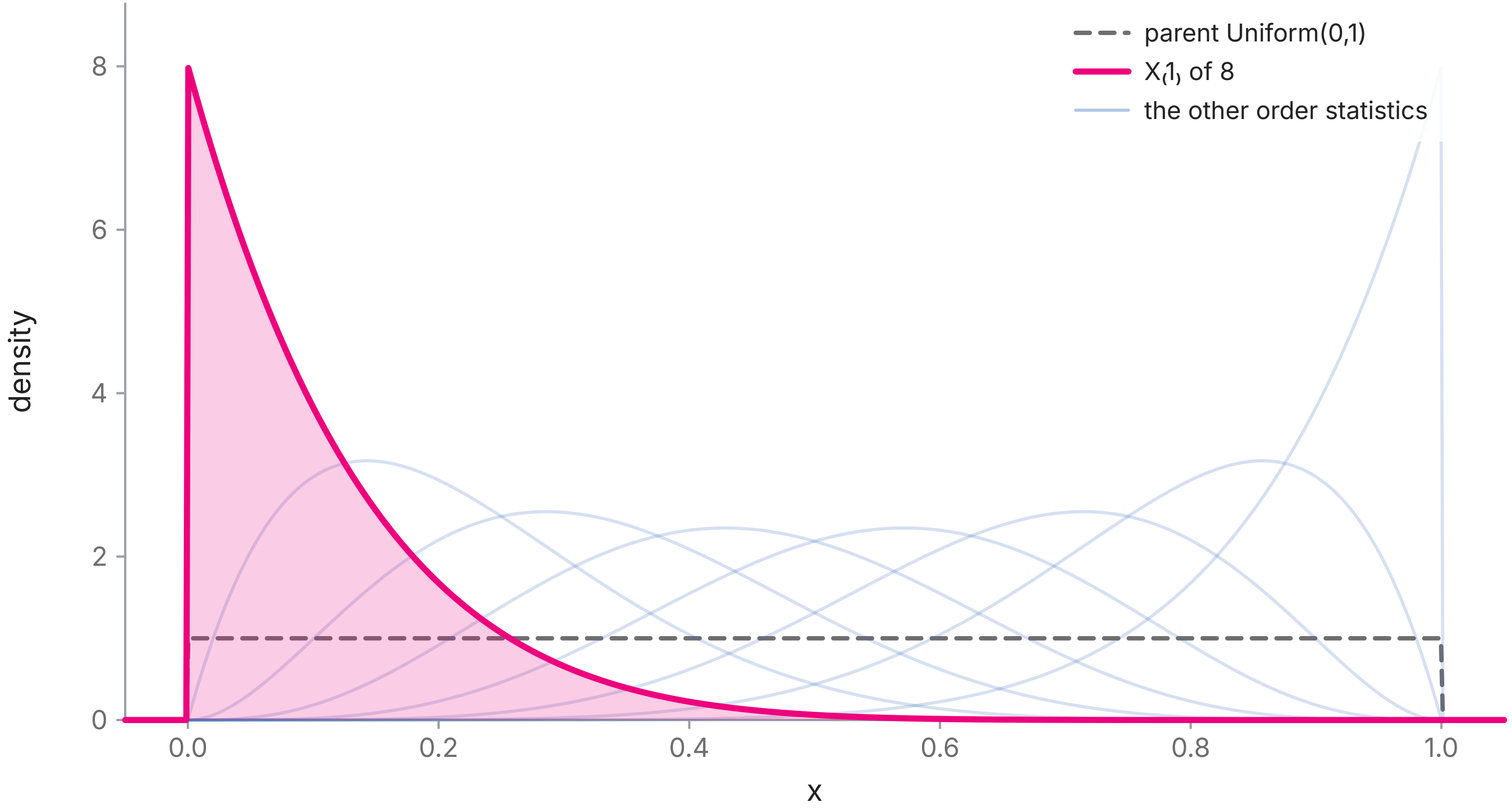
1

$k = 1$ is the minimum, $k = n$ the maximum. Exact densities — no simulation.

Note how the extremes are the most skewed, and how they pull away from the parent as n grows.

k = 1: the minimum

$E[X_{(1)}] \approx 0.111$ exact $k/(n+1) = 0.111$



Summary

Summary

Key Concepts

- Joint, marginal and conditional distributions
- Independence — pairwise versus mutual
- Covariance, correlation and the covariance matrix
- The multivariate normal distribution
- Transformations of random vectors
- Order statistics

Important Relationships

- Independence $\Rightarrow$ zero covariance, **but not conversely**
- For the multivariate normal: zero covariance $\Leftrightarrow$ independence
- Linear transformations preserve normality
- The joint determines the marginals, but not the other way round

The thread

Marginals describe each coordinate on its own. Everything interesting — dependence, conditioning, prediction — lives in what the marginals leave out.

Next Steps

Applications in Statistics

- Sampling distributions
- Maximum likelihood estimation
- Hypothesis testing for multiple parameters
- Regression analysis
- Principal component analysis
- Multivariate analysis of variance (MANOVA)

Further Study

- Characteristic functions
- Copulas
- Extreme value theory
- Time series analysis