Table of Contents

16: Continuous Joint Distributions

Jerry Cain

❤️ February 14, 2024 ❤️

[Lecture Discussion on Ed]

Continuous Joint Distributions

Stanford logo with darts

The Stanford letterhead logo was created by throwing 500,000 darts according to a joint distribution.

not really, but let's pretend

If we throw another dart according to the same distribution, what is

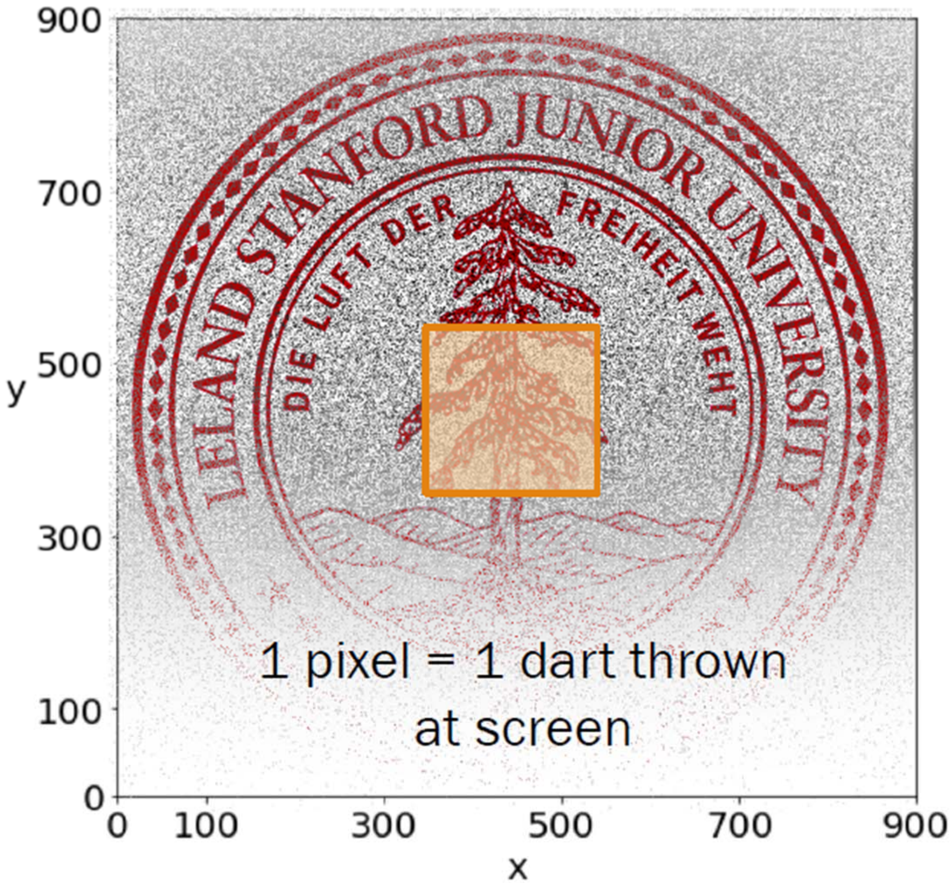
P(dart hits within r pixels of center)?

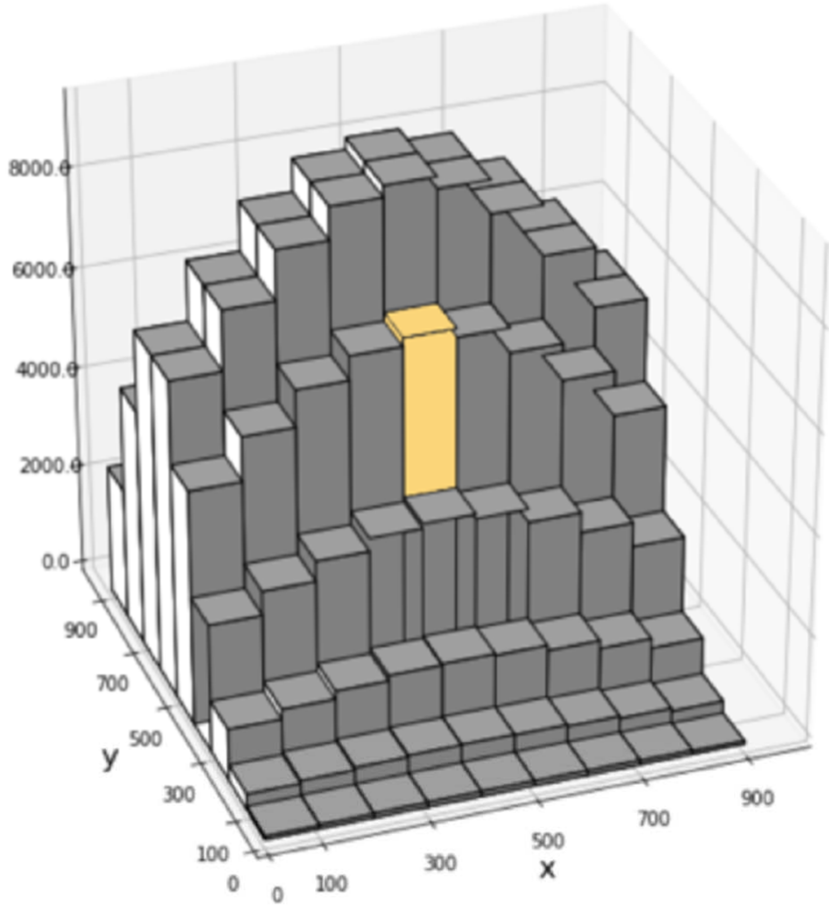
Quick check: What is the probability that a dart hits at (456.2344132343, 532.1865739012)?

so small it's essentially zero

CS109 logo with darts

P(dart hits within r pixels of center)?

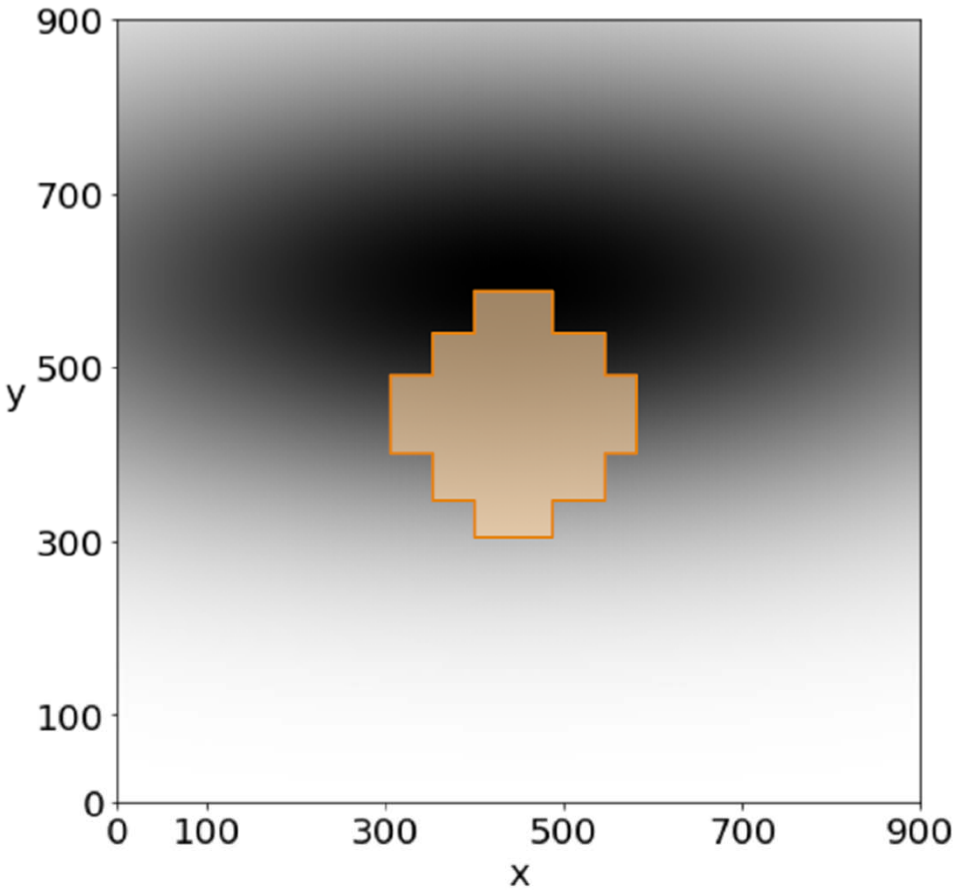


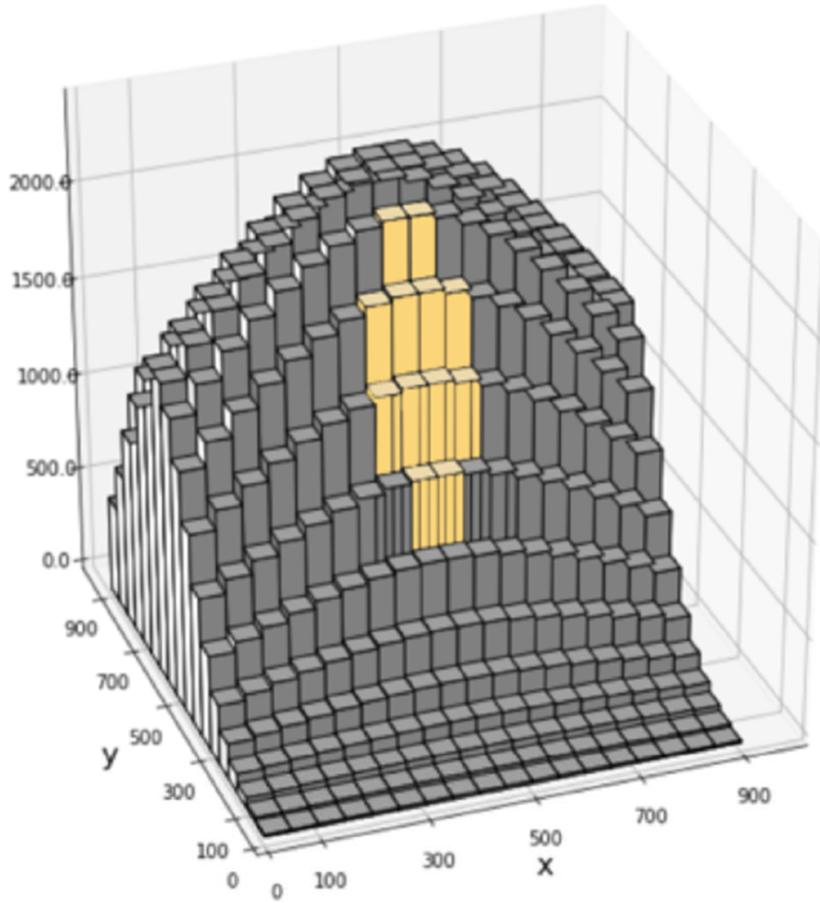


Possible dart counts (in 100x100 boxes)

CS109 logo with darts

P(dart hits within r pixels of center)?

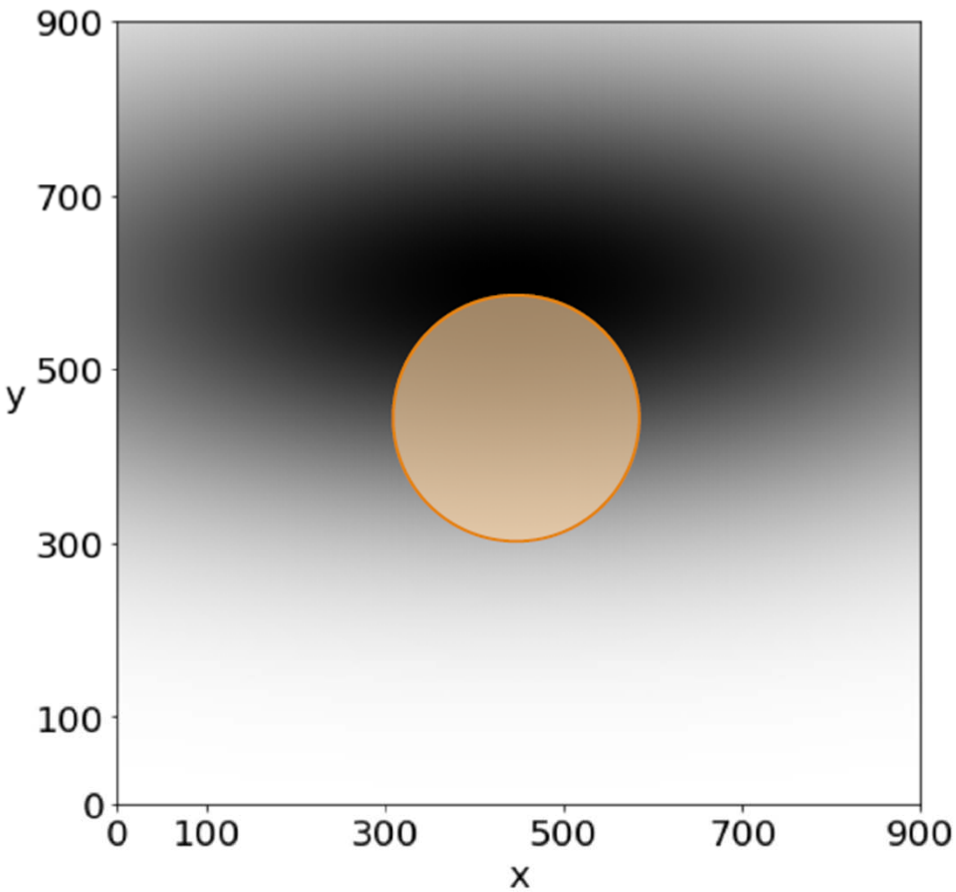


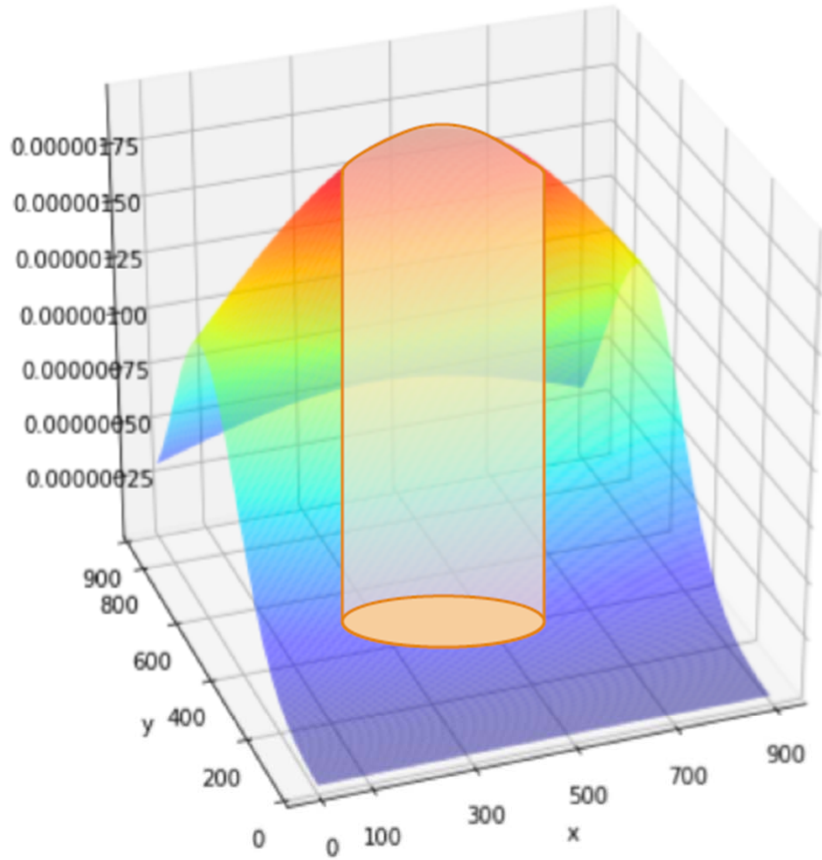


Possible dart counts (in 50x50 boxes)

CS109 logo with darts

P(dart hits within r pixels of center)?





Possible dart counts
(in infinitesimally small boxes)

Continuous joint probability density functions

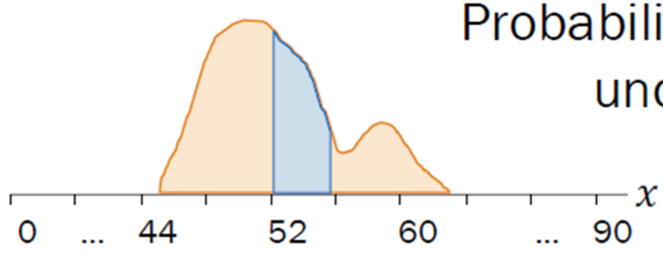
If two random variables X and Y are jointly continuous, then there exists a joint probability density function $f_{X,Y}$ defined over $-\infty < x, y < \infty$ such that:

$$P\left(a_1 \leq X \leq a_{2,} \;\; b_1 \leq Y \leq b_2\right) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x, y) dy\, dx$$

From one continuous RV to jointly continuous RVs

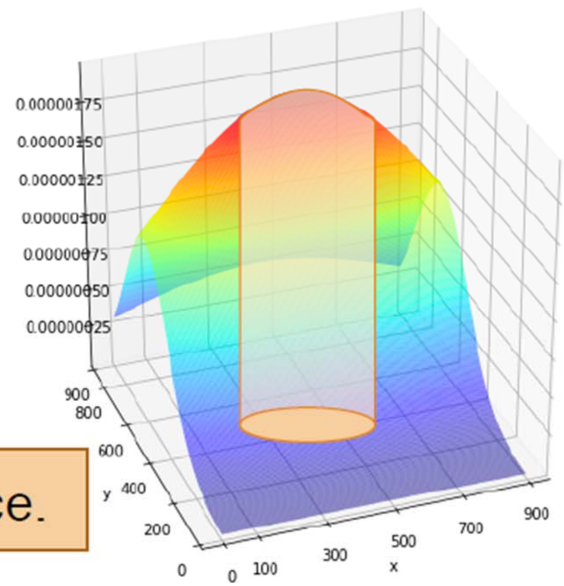
Single continuous RV X

- PDF f_X such that $\int_{-\infty}^{\infty} f_X(x)dx = 1$
- Integrate to get probabilities



Jointly continuous RVs X and Y

- PDF $f_{X,Y}$ such that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\,dx = 1$
- Double integrate to get probabilities



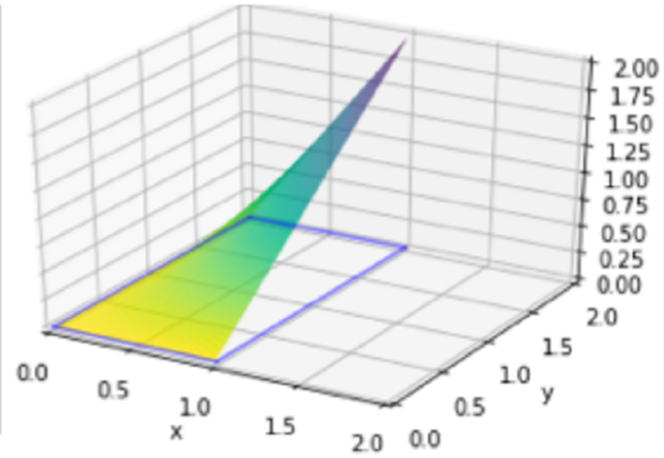
Probability for jointly continuous RVs is **volume** under a surface.

Double integrals without tears

Let X and Y be two continuous random variables.

- Support: $0 \leq X \leq 1$, $0 \leq Y \leq 2$.

Is $g(x, y) = xy$ a valid joint PDF over X and Y?

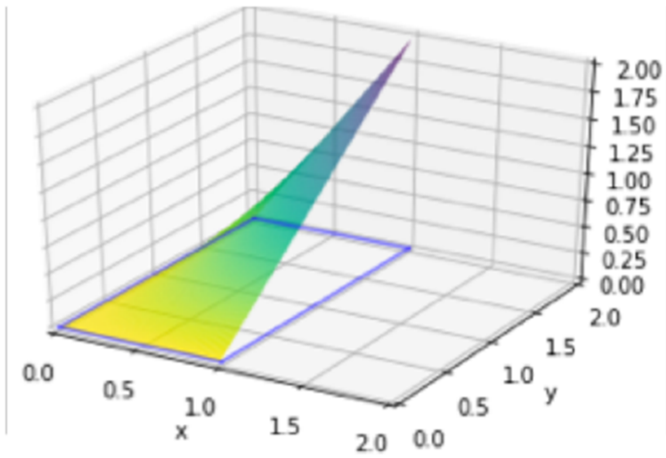


Write down the definite double integral that must integrate to 1:

Double integrals without tears

Let X and Y be two continuous random variables.

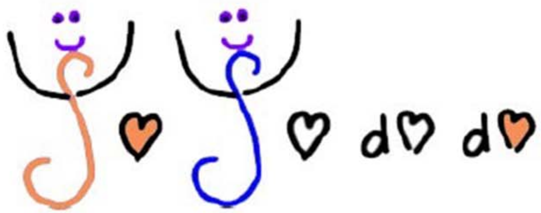
- Support: $0 \leq X \leq 1$, $0 \leq Y \leq 2$.



Is $g(x, y) = xy$ a valid joint PDF over X and Y?

Write down the definite double integral that must integrate to 1:

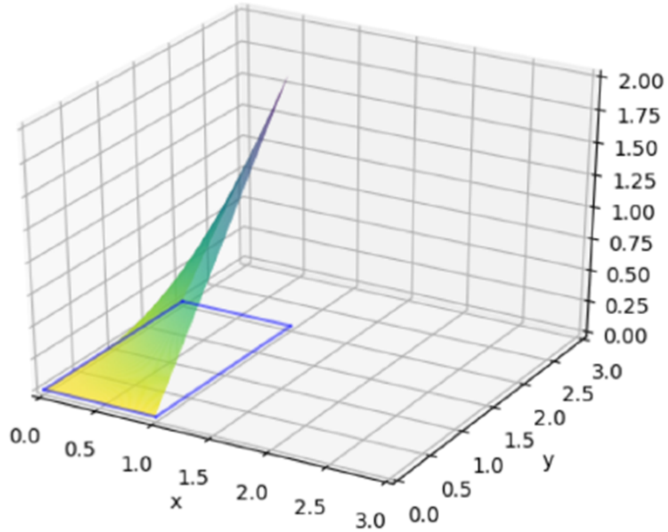
$$\underbrace{\int_{y=0}^{2} \int_{x=0}^{1} xy \, dx \, dy = 1}_{\text{(used in next slide)}} \quad \text{or} \quad \int_{x=0}^{1} \int_{y=0}^{2} xy \, dy \, dx = 1$$



Double integrals without tears

Let X and Y be two continuous random variables.

- Support: $0 \le X \le 1$, $0 \le Y \le 2$.



Is $g(x, y) = xy$ a valid joint PDF over X and Y?

0. Set up integral:

$$1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)dx\, dy = \int_{y=0}^{2}\int_{x=0}^{1} xy\, dx\, dy$$

1. Evaluate inside integral by treating y as a constant:

$$\int_{y=0}^{2}\left(\int_{x=0}^{1} xy\, dx\right)dy = \int_{y=0}^{2} y\left(\int_{x=0}^{1} x\, dx\right)dy = \int_{y=0}^{2} y\left[\frac{x^2}{2}\right]_0^1 dy = \int_{y=0}^{2} y\frac{1}{2}dy$$

2. Evaluate remaining (single) integral:

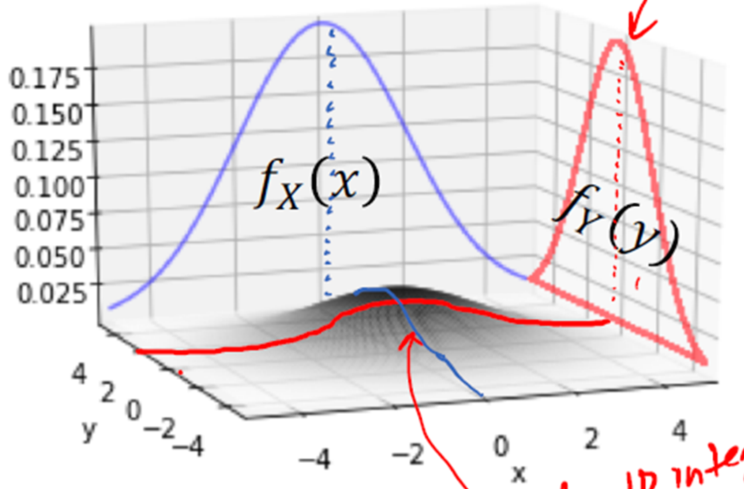
$$\int_{y=0}^{2} y\frac{1}{2}dy = \left[\frac{y^2}{4}\right]_{y=0}^{2} = 1 - 0 = 1$$

Yes, $g(x, y)$ is a valid joint PDF because it integrates to 1.

Marginal distributions

Suppose X and Y are continuous random variables with joint PDF:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy\,dx = 1$$



the 1D integral of density along red line produces height of density in marginal

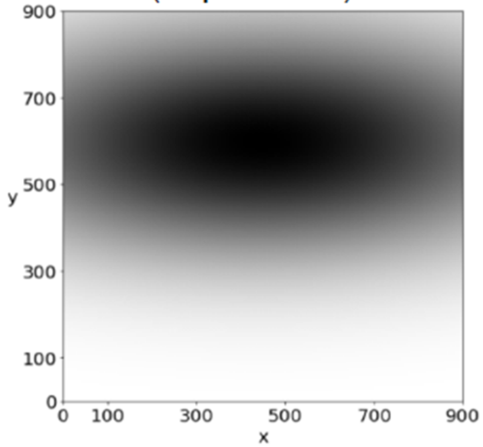
The marginal density functions (marginal PDFs) are therefore:

$$f_X(a) = \int_{-\infty}^{\infty} f_{X,Y}(a,y)\,dy \qquad f_Y(b) = \int_{-\infty}^{\infty} f_{X,Y}(x,b)\,dx$$

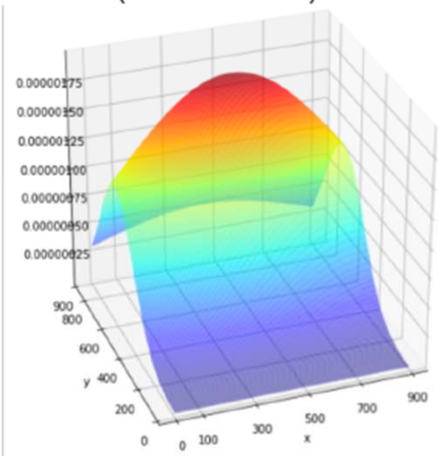
similar story for blue..

Back to darts!

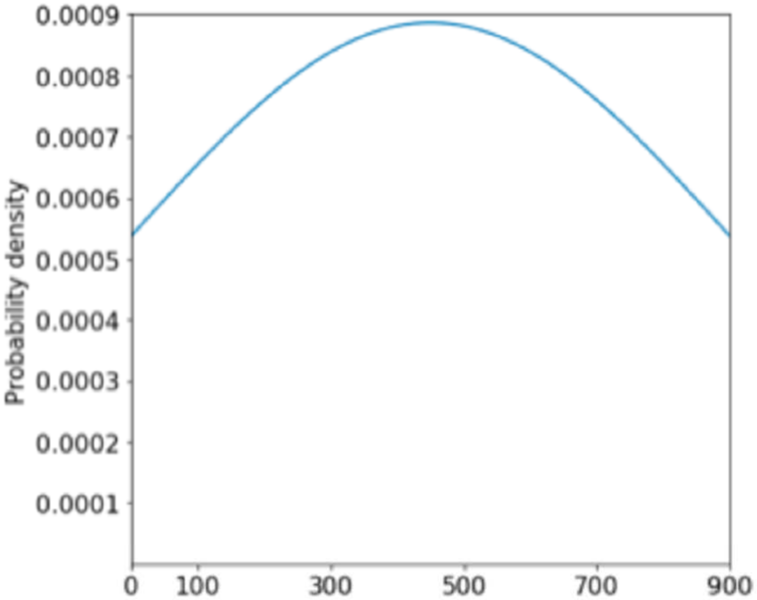
(top-down)

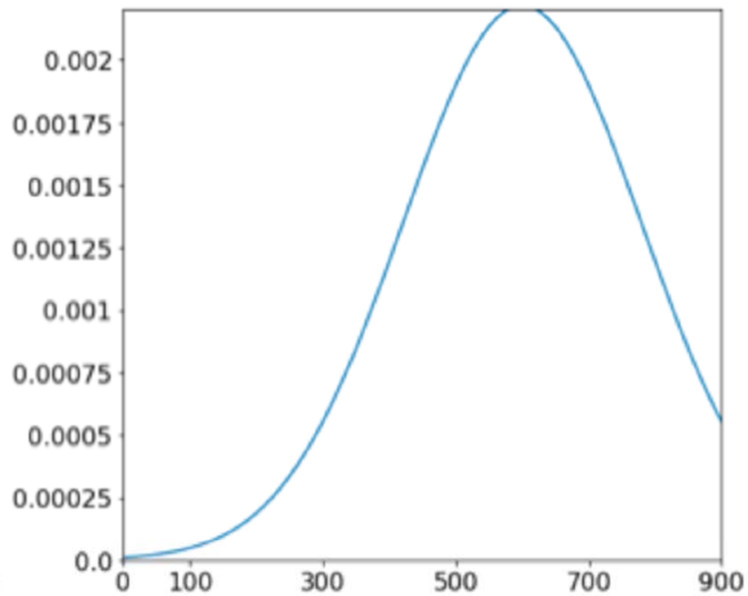


(side view)



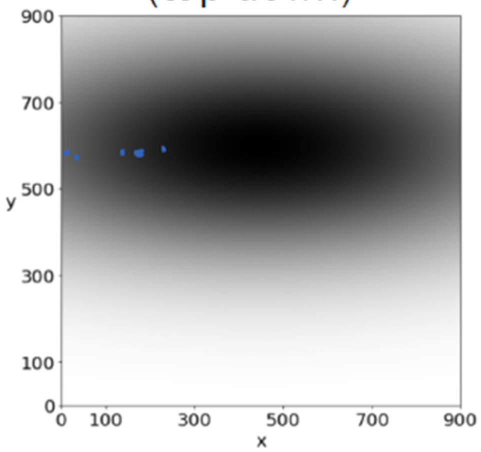
Match X and Y to their respective marginal PDFs:



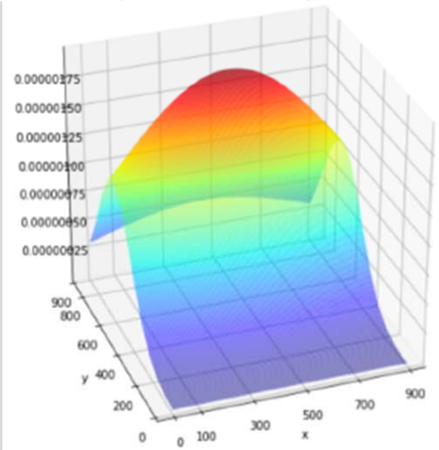


Back to darts!

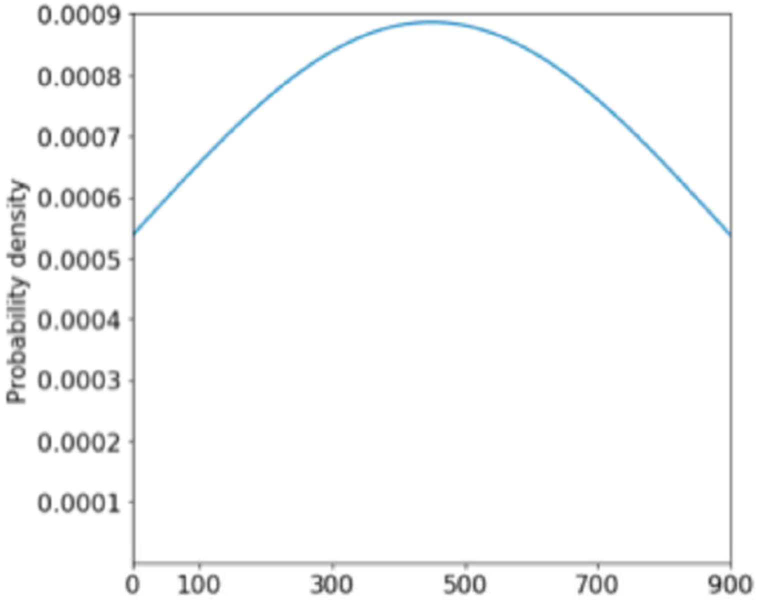
(top-down)



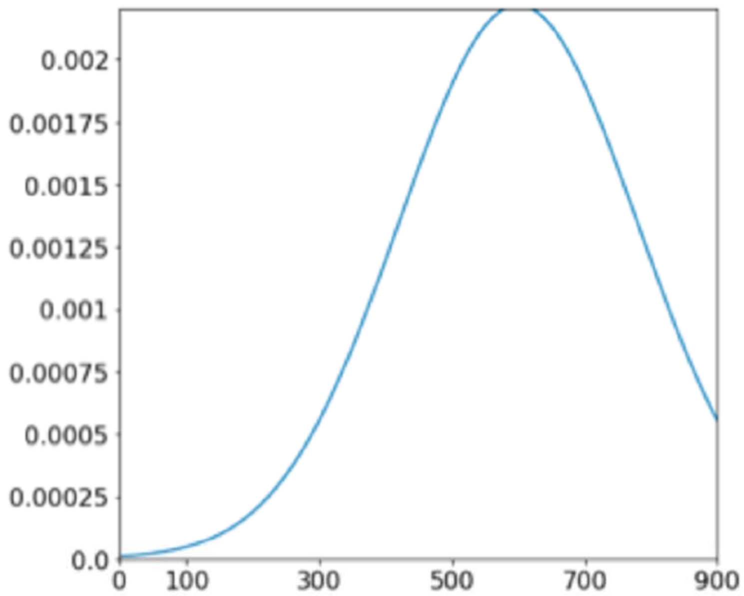
(side view)



Match X and Y to their respective marginal PDFs:



pixel x



pixel y

Joint CDFs

An observation: Connecting CDF to PDF

For a continuous random variable X with PDF f, the CDF (cumulative distribution function) is

$$F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

The density f is therefore the derivative of the CDF, F:

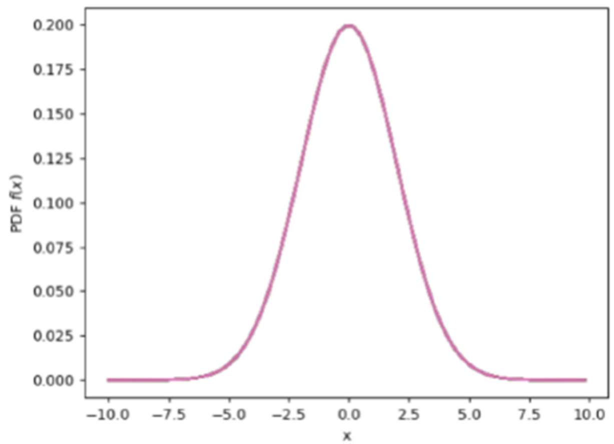
$$\boxed{f(x) = \frac{d}{dx}F(x)}$$

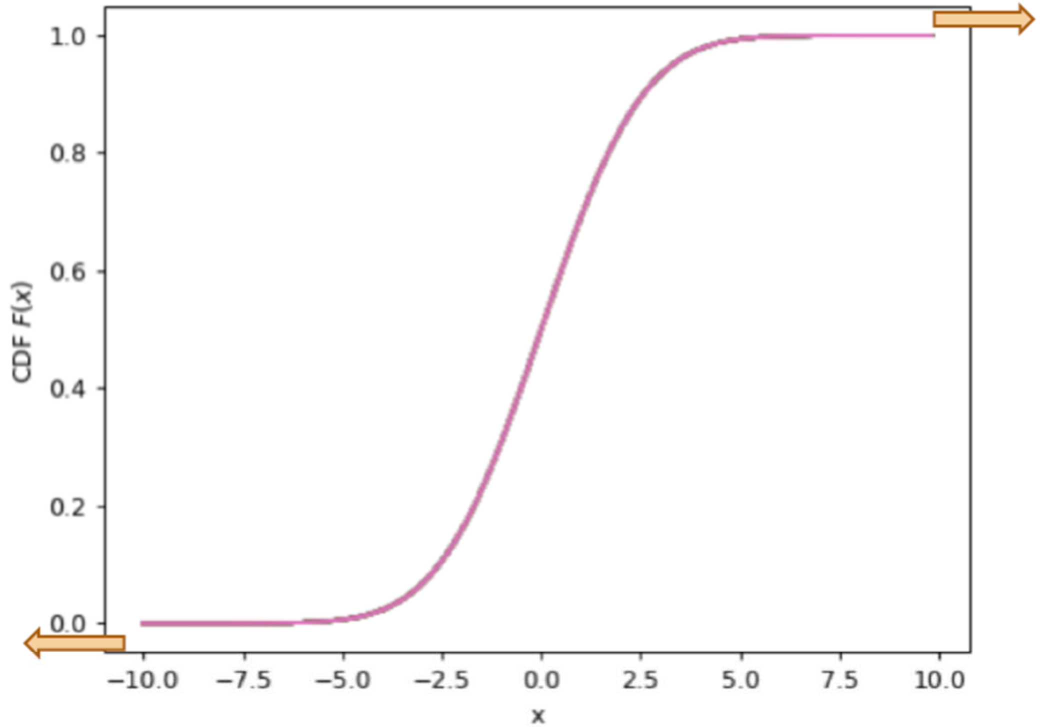
Fundamental Theorem of Calculus

Single variable CDF, graphically

Review

$$\lim_{x \to +\infty} F_X(x) = 1$$





$$\lim_{x \to -\infty} F_X(x) = 0$$

$f_X(x)$

$F_X(x) = P(X \leq x)$

Joint cumulative distribution function

For two random variables X and Y, there can be a joint cumulative distribution function $F_{X,Y}$:

$$F_{X,Y}(a, b) = P(X \leq a, Y \leq b)$$

For discrete X and Y:

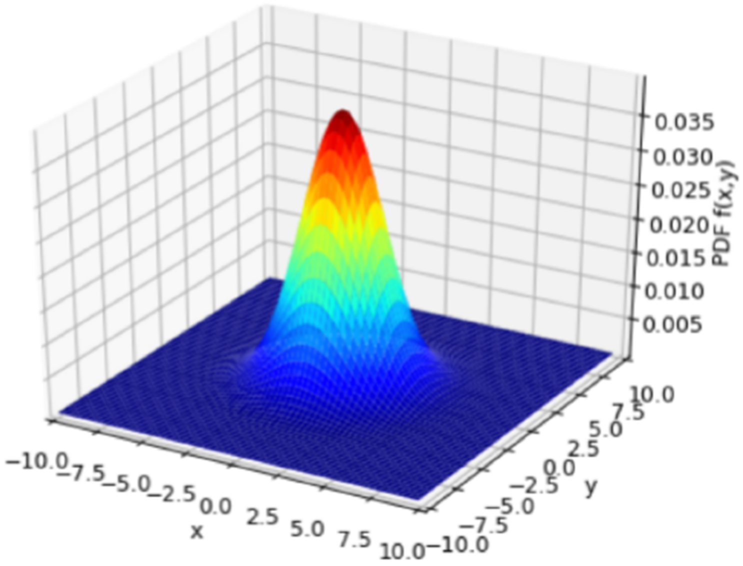
$$F_{X,Y}(a, b) = \sum_{x \leq a} \sum_{y \leq b} p_{X,Y}(x, y)$$

For continuous X and Y:

$$F_{X,Y}(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} f_{X,Y}(x, y) dy\, dx$$

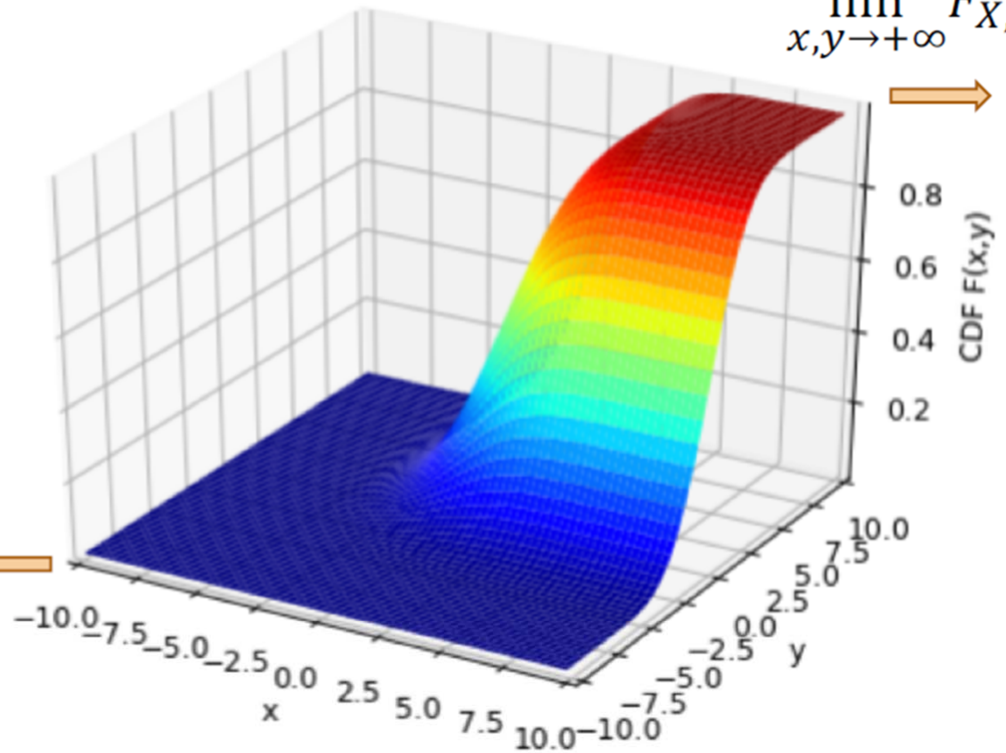
$$f_{X,Y}(a, b) = \frac{\partial^2}{\partial a\, \partial b} F_{X,Y}(a, b)$$

Joint CDF, graphically



$$\lim_{x,y\to+\infty} F_{X,Y}(x,y) = 1$$

$$\lim_{x,y\to-\infty} F_{X,Y}(x,y) = 0$$



$$f_{X,Y}(x,y)$$

$$F_{X,Y}(x,y) = P(X \le x, Y \le y)$$

Independent Continuous RVs

Independent continuous RVs

Two continuous random variables X and Y are independent if:

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \qquad \forall x, y$$

Equivalently:

$$\boxed{\begin{aligned} F_{X,Y}(x,y) &= F_X(x)F_Y(y) \\ f_{X,Y}(x,y) &= f_X(x)f_Y(y) \end{aligned}} \qquad \forall x, y$$

Proof of PDF:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x\,\partial y} F_X(x)F_Y(y)$$

$$= \frac{\partial}{\partial x}\frac{\partial}{\partial y} \underline{F_X(x)}F_Y(y) = \frac{\partial}{\partial x}F_X(x)\,\frac{\partial}{\partial y}F_Y(y)$$

this is a constant with respect to y!

$$= f_X(x)f_Y(y)$$

Independent continuous RVs

Two continuous random variables X and Y are **independent** if:

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) \qquad \forall x, y$$

Equivalently:

$$F_{X,Y}(x,y) = F_X(x)F_Y(y) \qquad \forall x, y$$
$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

More generally, X and Y are independent if joint density on factors into two, single-variable marginal probability densities:

$$f_{X,Y}(x,y) = g(x)h(y), \text{ where } -\infty < \underbrace{x, y} < \infty$$

states that x and y can freely roam over their full support!

Pop quiz! (more kidding)

$f_{X,Y}(x, y) = g(x)h(y)$, where $-\infty < x, y < \infty$ ⇒ independent X and Y

Are X and Y independent in the following cases?

✅ 1. $f_{X,Y}(x, y) = 6e^{-3x}e^{-2y}$ where $0 < x, y < \infty$

Separable functions: $g(x) = 3e^{-3x}$, $h(y) = 2e^{-2y}$

$\int_0^1\int_0^1 4xy = 1 = \int_0^1 c\, 4x\, dx \cdot \int_0^1 \frac{1}{c} y\, dy$

Constant is dictated by requirement that yellow integrate to 1 and blue integrate to 1. $c = 1/2$

✅ 2. $f_{X,Y}(x, y) = 4xy$ where $0 < x, y < 1$

Separable functions: $g(x) = 2x$, $h(y) = 2y$

3. $f_{X,Y}(x, y) = 24xy$ where $0 < x + y < 1$

Cannot capture constraint on $x + y$!

If you can factor densities over the entire support, you have independence.

More pop quiz! (more kidding)

X and Y have the following joint PDF:

$$f_{X,Y}(x,y) = 3e^{-3x}$$
where $0 < x < \infty, 1 < y < 2$

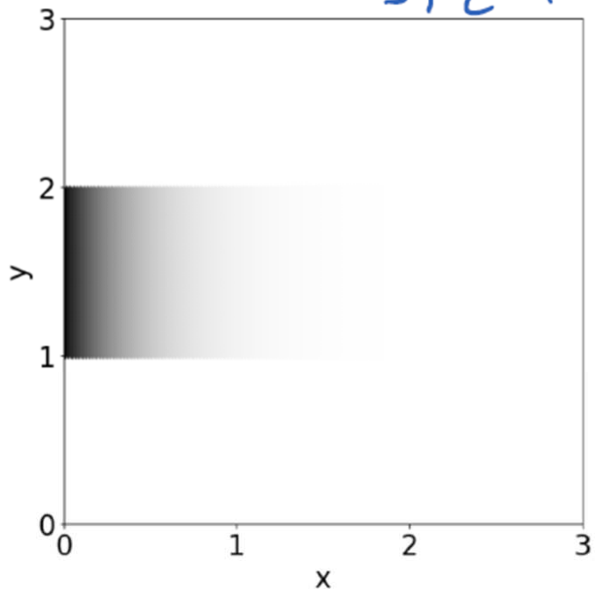
1. Are X and Y independent?

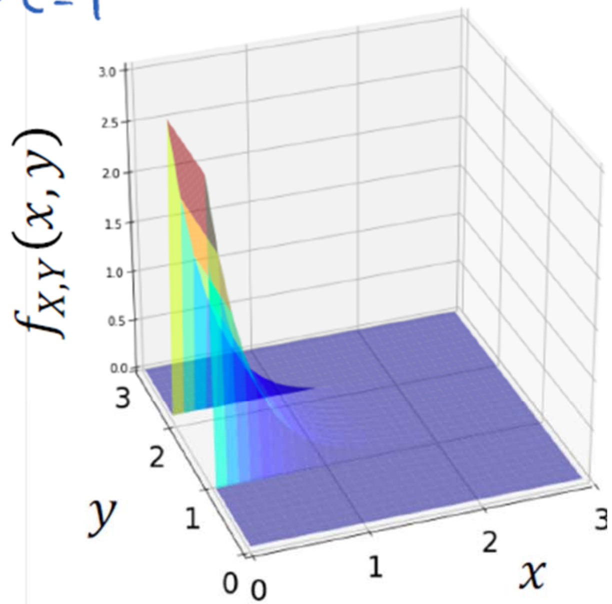
$g(x) = 3Ce^{-3x}, 0 < x < \infty$ C is a constant
$h(y) = 1/C, \quad 1 < y < 2$

$\int_1^2 \frac{1}{C} dy = 1 \Rightarrow C = 1$

2. What is the marginal PDF of X? Of Y?

3. What is $E[X+Y]$?





More pop quiz! (more kidding)

X and Y have the following joint PDF:

$$f_{X,Y}(x,y) = 3e^{-3x}$$
where $0 < x < \infty, 1 < y < 2$

1. Are X and Y independent?

$$g(x) = 3e^{-3x}, 0 < x < \infty$$
$$h(y) = 1, \qquad 1 < y < 2$$

2. What is the marginal PDF of X? Of Y?

$$\begin{cases} f_Y(y) = h(y) = 1, \; 1<y<2 \iff Y \sim \text{Uni}(1,2) \\ f_X(x) = g(x) = 3e^{-3x}, \; 0<x<\infty \iff X \sim \text{Exp}(3) \end{cases}$$

3. What is $E[X+Y]$?

Strategy 1: $E[g(x,y)] = \iint g(x,y) f_{X,Y}(x,y)\,dy\,dx$
$= \iint (x+y) f_{X,Y}(x,y)\,dy\,dx$ LOTUS

Strategy 2: $E[X+Y] = E[X] + E[Y]$
$= 1/3 + 3/2 = 11/6$ linearity of expectation

The joy of meetings

Two people set up a meeting time. Each arrives independently at a time uniformly distributed between 12pm and 12:30pm.

Define X = # minutes past 12pm that person 1 arrives. $X \sim \text{Uni}(0, 30)$
Y = # minutes past 12pm that person 2 arrives. $Y \sim \text{Uni}(0, 30)$

What is the probability that the first to arrive waits >10 mins for the other?

Compute: $P(X + 10 < Y) + P(Y + 10 < X) = 2P(X + 10 < Y)$ (by symmetry)

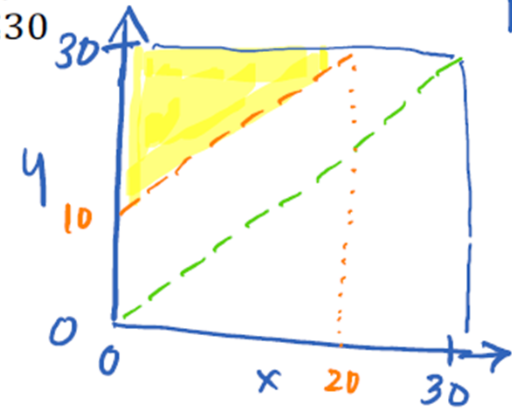
1. What is symmetry here?
2. How do we integrate to compute this probability?

$$f_{X,Y}(x, y) = \left(\frac{1}{30}\right)^2$$

$$\iint\limits_{\substack{0 \le x,y \le 30 \\ x+10<y}} \left(\frac{1}{30}\right)^2 dx\,dy + \iint\limits_{\substack{0 \le x,y \le 30 \\ y+10<x}} \left(\frac{1}{30}\right)^2 dx\,dy$$

Double integrals: A guide

From last slide: $2P(X + 10 < Y) = 2 \iint_{\substack{x+10<y, \\ 0 \le x,y, \le 30}} (1/30)^2 dxdy$ (by symmetry, independence)



blue box delineates full range of 2D PDF

green dashed line is $y=x$

orange dashed line is $y=x+10$

yellow is everywhere $y > x+10$

<u>Steps</u>:

1. Draw a picture.
2. Set bounds "from outside in".
   - Outer integral bounds should be full range possible
   - Inner integral can depend on integration variable of outer integral

$$\boxed{= \frac{2}{30^2} \int_{10}^{30} \int_{0}^{y-10} dxdy}$$

$$= \frac{2}{30^2} \int_{10}^{30} (y - 10)dy \qquad = \cdots = \frac{4}{9}$$

Bivariate Normal Distribution

(랜덤변수 2개)

Bivariate Normal Distribution

recall that for $X \sim N(\mu, \sigma^2) \Rightarrow f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

X_1 and X_2 follow a bivariate normal distribution if their joint PDF f is

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)}$$

Can show that $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ (Ross chapter 6, example 5d)

Often written as: $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

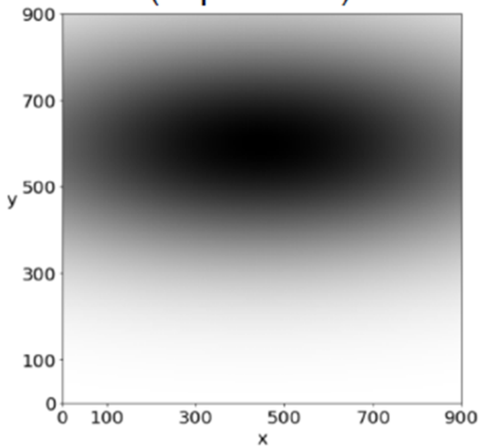
- Vector $\boldsymbol{X} = (X_1, X_2)$
- Mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2)$, Covariance matrix: $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) \end{bmatrix}$

Recall correlation: $\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_1\sigma_2}$

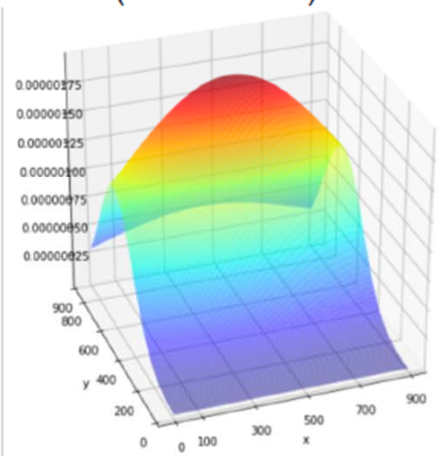
We will focus on understanding the **shape** of a bivariate Normal RV.

Back to darts

(top-down)



(side view)



Darts were thrown according to a bivariate normal distribution:

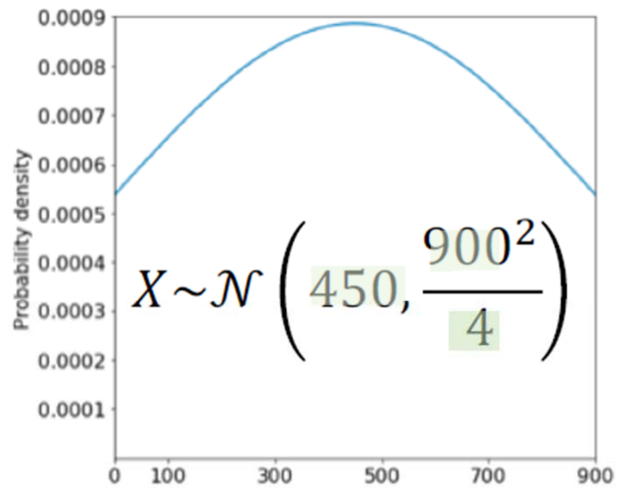
$$(X, Y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\boldsymbol{\mu} = (450, 600)$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 900^2/4 & 0 \\ 0 & 900^2/25 \end{bmatrix}$$

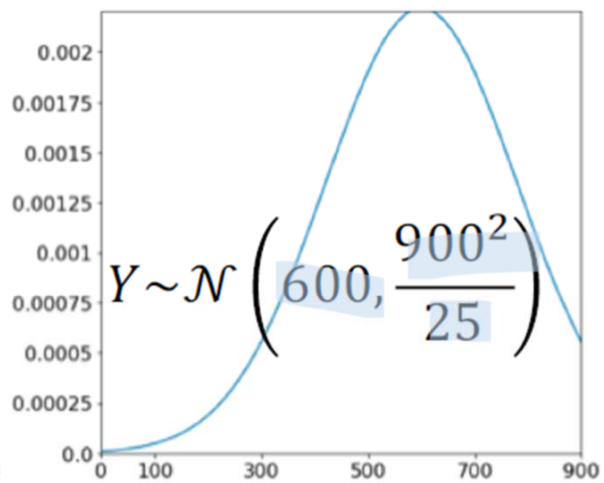
says that $Cov(X_1, X_2) = 0$

Marginal PDFs:



$$X \sim \mathcal{N}\left(450, \frac{900^2}{4}\right)$$

pixel x



$$Y \sim \mathcal{N}\left(600, \frac{900^2}{25}\right)$$

pixel y

A diagonal covariance matrix

Let $\boldsymbol{X} = (X_1, X_2)$ follow a bivariate normal distribution $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = (\mu_1, \mu_2), \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

still says $\mathrm{Cov}(X_1, X_2) = 0$

Are X_1 and X_2 independent?

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\underbrace{\rho^2}_{0})}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \underbrace{\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}}_{\text{goes to zero}} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)}$$

$$= \frac{1}{2\pi\sigma_1\sigma_2} e^{-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)}$$

(Note covariance: $\rho\sigma_1\sigma_2 = 0$)

$$= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-(x_1-\mu_1)^2/2\sigma_1^2} \frac{1}{\sigma_2\sqrt{2\pi}} e^{-(x_2-\mu_2)^2/2\sigma_2^2}$$ ✅

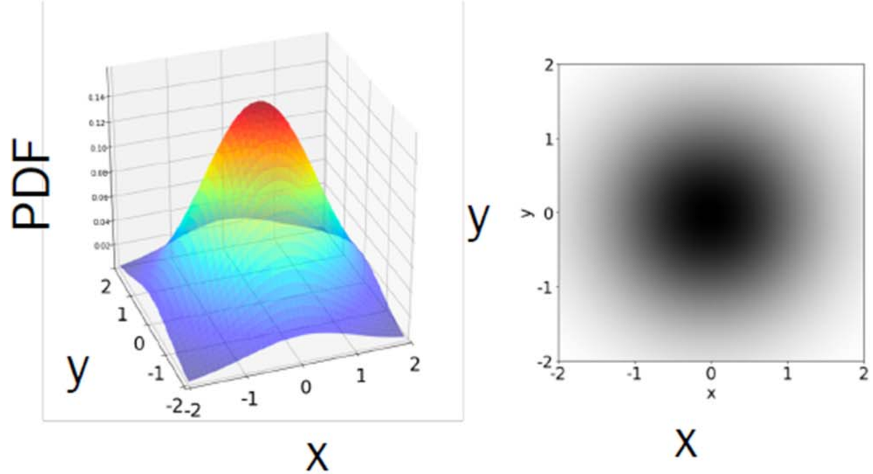
X_1 and X_2 are **independent** with marginal distributions $X_1 \sim \mathcal{N}(\mu_1\ \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2\ \sigma_2^2)$

(X, Y) Matching (all have $\boldsymbol{\mu} = (0, 0)$)

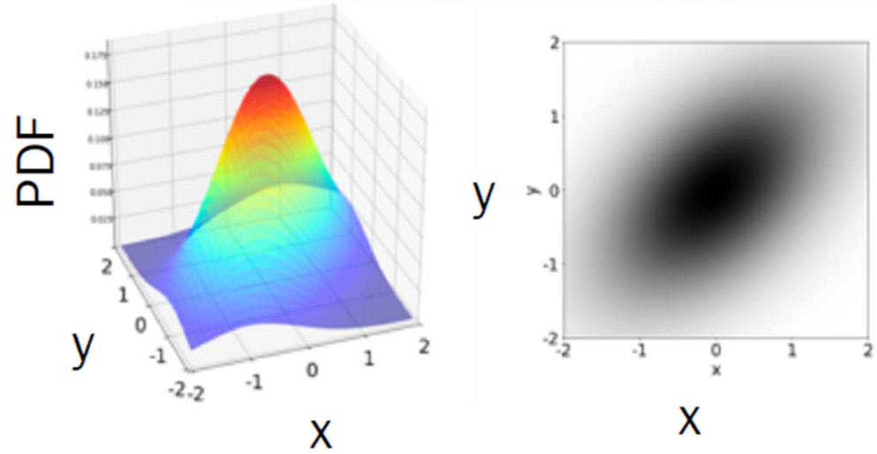
A. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ B. $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

C. $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

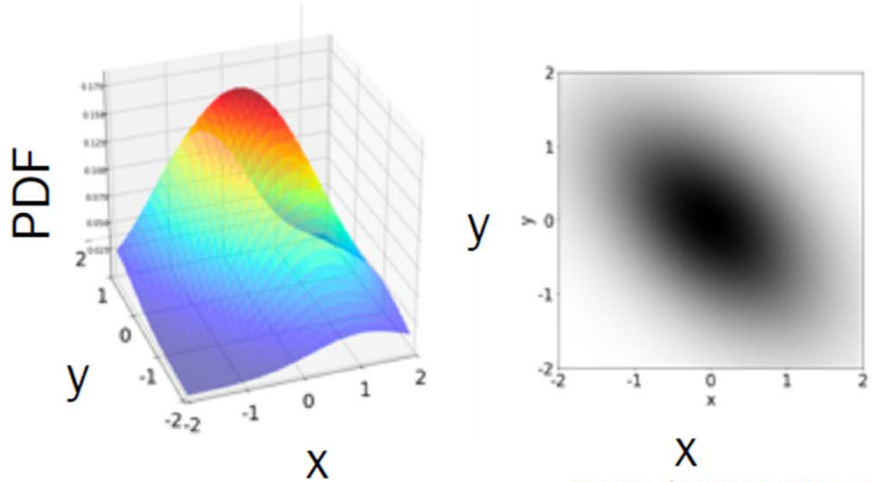
1.



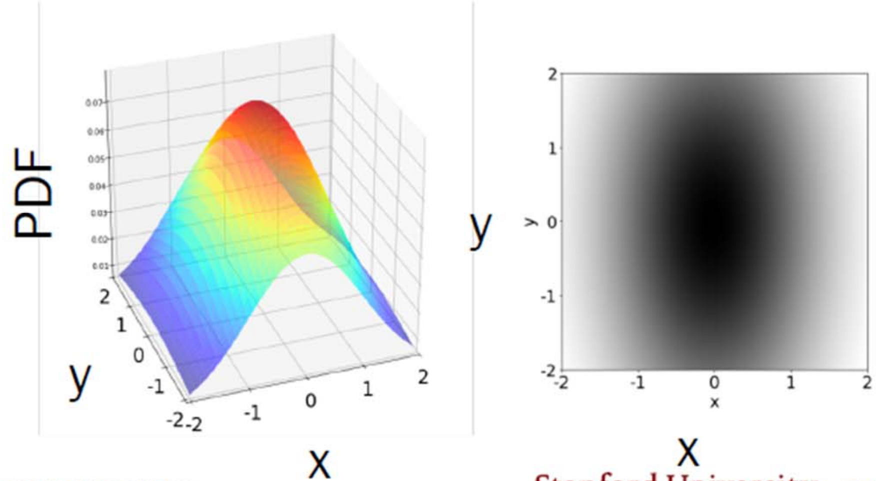
2.



3.



4.



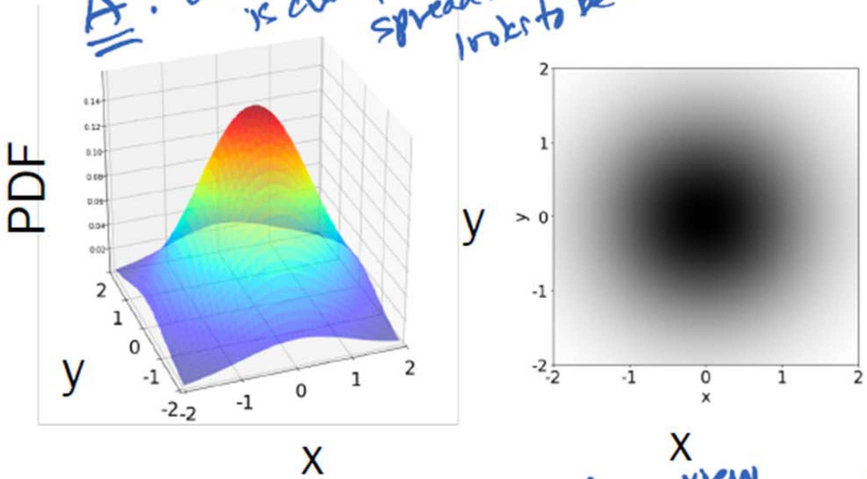
(X, Y) Matching (all have $\boldsymbol{\mu} = (0, 0)$)

A. $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ B. $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$

C. $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ D. $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

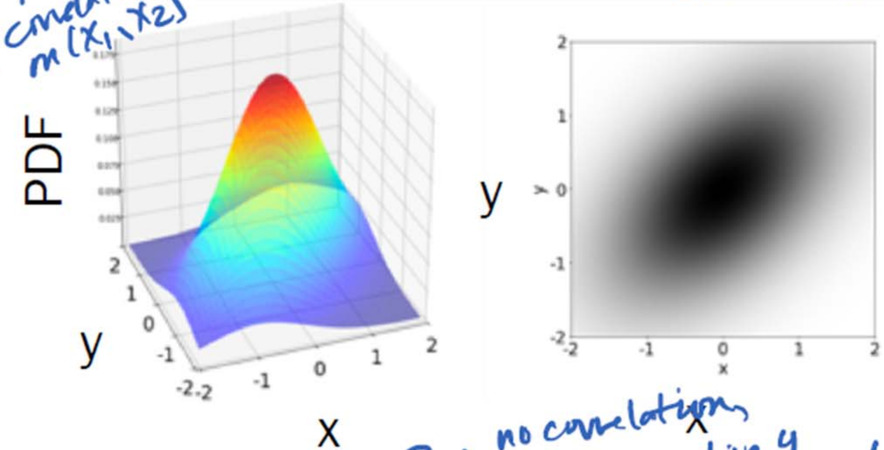
A: covariance is clearly zero, spread in x and y looks to be the same.

1.



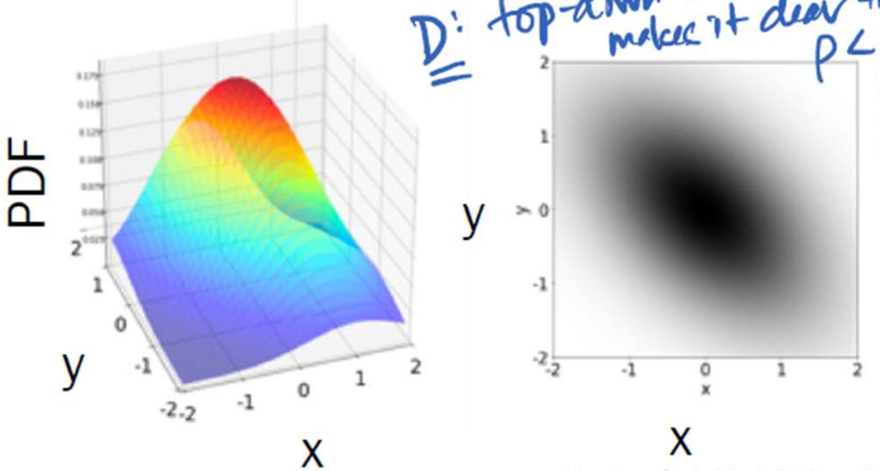
C top-down view suggests positive correlation in (X_1, X_2)

2.



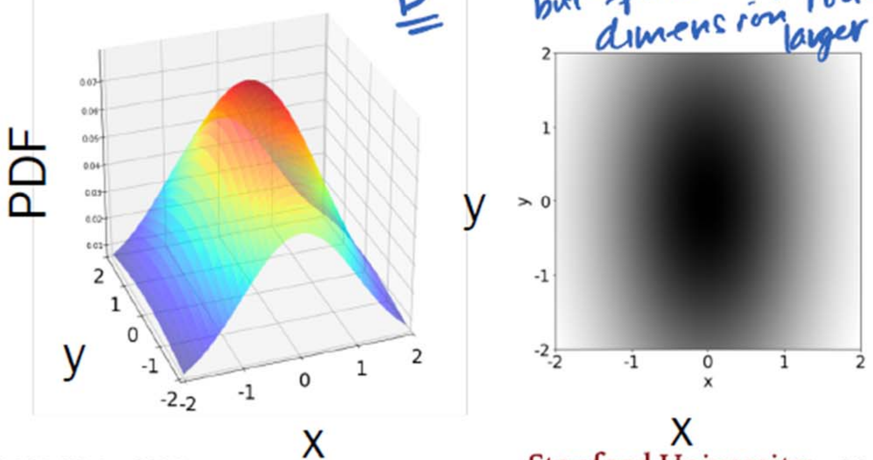
D: top-down view makes it clear that $\rho < 0$, because y decreases as x increases

3.

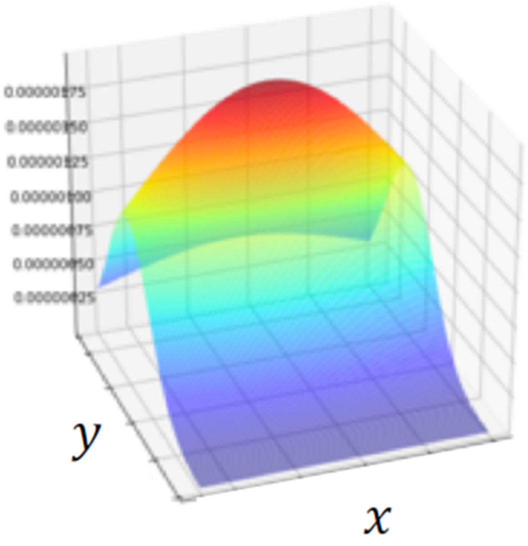


B: no correlation, but spread in y dimension looks larger

4.



Why are joint PDFs useful?



Independence
2-D support
Joint PDF
Joint CDF
Marginal PDF
(Friday) Conditional PDF

- How 2 continuous RVs vary with each other

  $P(X < Y)$, $\text{Cov}(X, Y), \rho(X, Y)$

- How continuous RV is distributed given evidence (more on Friday)

  Given $Y = y$, the distribution of X

- How a continuous RV can be decomposed into 2 RVs (or vice versa)

  Distribution of $Z = X + Y$ (which is a <u>1-D</u> RV!)

Sum of Independent Gaussians

Sum of independent Gaussians

$E[X+Y] = E[X] + E[Y] = \mu_1 + \mu_2$

$Var(X+Y) = Var(X) + Var(Y) = \sigma_1^2 + \sigma_2^2$

because of independence

$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$

$X \sim \mathcal{N}(\mu_1, \sigma_1^2)$,
$Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$
X, Y independent

$$X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

(proof left to Wikipedia)

Holds in general case:

$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
X_i independent for $i = 1, \ldots, n$

$$\sum_{i=1}^{n} X_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

Back for another playoffs game

What is the probability that the Warriors win?
How do you model zero-sum games?

$$P(A_W > A_B)$$

This is a probability of an event involving **two** random variables!

We will compute:

$$P(\underbrace{A_W - A_B}_{\text{A sum of Normals!}} > 0)$$

Motivating idea: Zero sum games

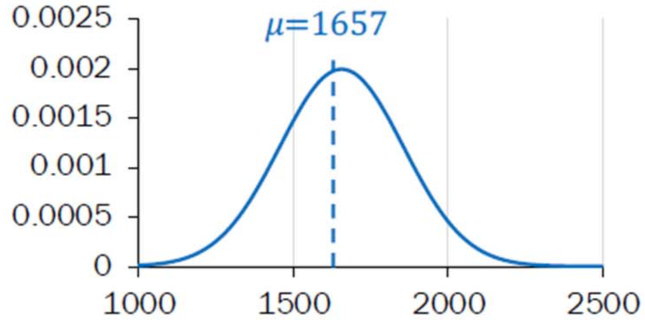
Want: $P(\text{Warriors win}) = P(A_W - A_B > 0)$

Assume A_W, A_B are independent.
Let $D = A_W - A_B$.

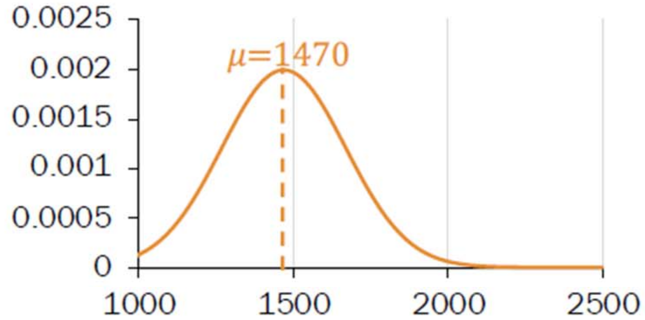
What is the distribution of D?

A. $D \sim \mathcal{N}(1657 - 1470,\ 200^2 - 200^2)$
B. $D \sim \mathcal{N}(1657 - 1470,\ 200^2 + 200^2)$
C. $D \sim \mathcal{N}(1657 + 1470,\ 200^2 + 200^2)$
D. $D \sim \mathcal{N}(1657 + 1470,\ 200^2)$

Warriors $A_W \sim \mathcal{N}(S = 1657, 200^2)$



Opponents $A_B \sim \mathcal{N}(S = 1470, 200^2)$



Motivating idea: Zero sum games

Want: $P(\text{Warriors win}) = P(A_W - A_B > 0)$

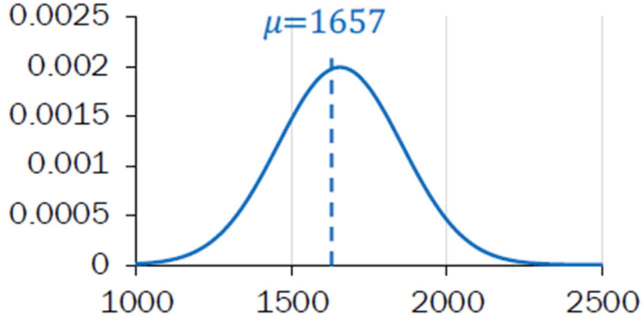
Assume A_W, A_B are independent.
Let $D = A_W - A_B$. } this is really $A_W + (-A_B)$

$-A_B \sim N(-1470, (-1)^2 200^2)$
$\Downarrow$
$N(-1470, 200^2)$

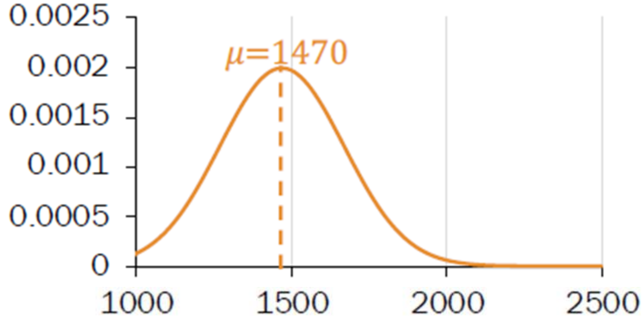
What is the distribution of D?

A. $D \sim \mathcal{N}(1657 - 1470,\ 200^2 - 200^2)$
B. $D \sim \mathcal{N}(1657 - 1470,\ 200^2 + 200^2)$
C. $D \sim \mathcal{N}(1657 + 1470,\ 200^2 + 200^2)$
D. $D \sim \mathcal{N}(1657 + 1470,\ 200^2)$

Warriors $A_W \sim \mathcal{N}(S = 1657, 200^2)$



Opponents $A_B \sim \mathcal{N}(S = 1470, 200^2)$



If $X \sim \mathcal{N}(\mu_1, \sigma^2)$,
then $(-X) \sim \mathcal{N}(-\mu, (-1)^2\sigma^2 = \sigma^2)$.

Motivating idea: Zero sum games

Want: $P(\text{Warriors win}) = P(A_W - A_B > 0)$

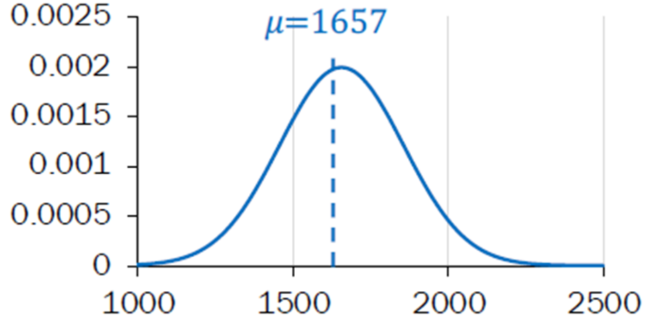
Assume A_W, A_B are independent.
Let $D = A_W - A_B$.

$$D \sim \mathcal{N}(1657 - 1470, \quad 200^2 + 200^2)$$
$$\sim \mathcal{N}(187, 2 \cdot 200^2) \quad \sigma \approx 282.842$$

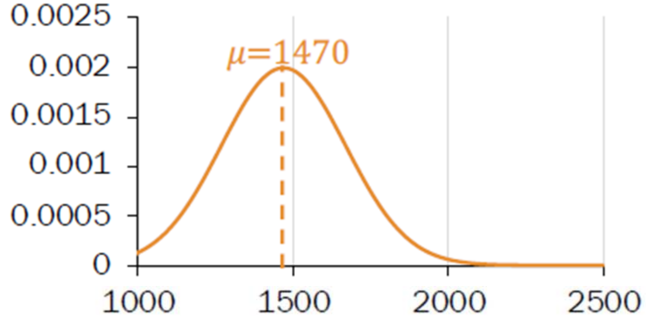
$$P(D > 0) = 1 - F_D(0) = 1 - \Phi\left(\frac{0 - 187}{282.842}\right)$$
$$\approx 0.74574$$

Compare with 0.7488, calculated by sampling!

Warriors $A_W \sim \mathcal{N}(S = 1657, 200^2)$



Opponents $A_B \sim \mathcal{N}(S = 1470, 200^2)$




```
>>> from scipy.stats import norm
>>> 1 - norm(187, 80000 ** 0.5).cdf(0)
0.7457402843526317
>>> 1 - norm(0, 1).cdf(-187 / (80000 ** 0.5))
0.7457402843526317
```

Virus infections

Suppose you are working with the WHO to initiate a response to the onset of a virus. There are two exposed groups:

- G1: 20000 people, each independently infected with $p_1 = 0.1$
- G2: 10000 people, each independently infected with $p_2 = 0.4$

What is $P(\text{people infected} \geq 6100)$? An approximation is okay.

1. Define RVs & state goal

Let A = # infected in G1.
$A \sim \text{Bin}(20000, 0.1)$
B = # infected in G2.
$B \sim \text{Bin}(10000, 0.4)$

Want: $P(A + B \geq 6100)$

Strategy:

A. Sum of independent Binomials — no, $p_1 \neq p_2$:(
B. Sum of independent Poissons — no, because p_1 and p_2 too large
C. Sum of independent Gaussians — yes!!!! :)
D. Sum of independent Exponentials — no, because mean and stdev of B are very different

Virus infections

Suppose you are working with the WHO to initiate a response to the onset of a virus. There are two exposed groups:

- G1: 20000 people, each independently infected with $p_1 = 0.1$
- G2: 10000 people, each independently infected with $p_2 = 0.4$

What is $P(\text{people infected} \geq 6100)$? An approximation is okay.

1. Define RVs & state goal

Let $A = \#$ infected in G1.
$A \sim \text{Bin}(20000, 0.1)$
$B = \#$ infected in G2.
$B \sim \text{Bin}(10000, 0.4)$

Want: $P(A + B \geq 6100)$

2. Approximate as sum of Gaussians

$A \approx X \sim \mathcal{N}(2000, 1800)$ $B \approx Y \sim \mathcal{N}(4000, 2400)$

$P(A + B \geq 6100) \approx P(X + Y \geq 6099.5)$ continuity correction

3. Solve

Virus infections

```
>>> 1 - norm(6000, 4200 ** 0.5).cdf(6099.5)
0.06235282662988528
>>> 1 - norm(0, 1).cdf((6099.5 - 6000)/(4200 ** 0.5))
0.06235282662988528
```

Suppose you are working with the WHO to initiate a response to the onset of a virus. There are two exposed groups:

- G1: 20000 people, each independently infected with $p_1 = 0.1$
- G2: 10000 people, each independently infected with $p_2 = 0.4$

What is $P(\text{people infected} \geq 6100)$? An approximation is okay.

1. Define RVs & state goal

Let $A = \#$ infected in G1.
$A \sim \text{Bin}(20000, 0.1)$
$B = \#$ infected in G2.
$B \sim \text{Bin}(10000, 0.4)$

Want: $P(A + B \geq 6100)$

2. Approximate as sum of Gaussians

$A \approx X \sim \mathcal{N}(2000, 1800)$ $B \approx Y \sim \mathcal{N}(4000, 2400)$

$P(A + B \geq 6100) \approx P(X + Y \geq 6099.5)$ continuity correction

3. Solve

Let $W = X + Y \sim \mathcal{N}(6000, 4200)$

$$P(W \geq 6099.5) = 1 - \Phi\left(\frac{6099.5 - 6000}{\sqrt{4200}}\right)$$

$$\approx 1 - \Phi(1.53531) \approx 0.06235$$

Sum of independent Gaussians

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$,
$X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
X_1, X_2 independent

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Is this related to linear transformations of Gaussians?

Recall:

If $Y = aX + b$, then $Y \sim \mathcal{N}(a\mu_X + b, a^2\sigma_X^2)$

Linear transforms vs. independence ⚠️

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = X + X$. What is the distribution of Y?

- Are both approaches valid?

Independent RVs approach ✖

Let $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
be independent.
Then $Y = X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

$Y = X + X$

$X + X \sim \mathcal{N}(\mu + \mu, \sigma^2 + \sigma^2)$?

$Y \sim \mathcal{N}(\underline{2\mu, 2\sigma^2})$?

X is NOT independent of X!

Linear transform approach ✅

Let $X \sim \mathcal{N}(\mu, \sigma^2)$.
If $Y = aX + b$,
then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

$Y = 2X$

$Y \sim \mathcal{N}(\underline{2\mu, 4\sigma^2})$

For <u>independent</u> $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$,
$aX_1 + bX_2 + c \sim \mathcal{N}(a\mu_1 + b\mu_2 + c, a^2\sigma_1^2 + b^2\sigma_2^2)$