Computer Science

# CSC380: Principles of Data Science

## Probability 4

Xinchen Yu

# Recap

- PDF of a transformation of a continuous RV
  - $X + b$ has a PDF that is a translation of $X$'s PDF by $b$ units
  - $aX$'s PDF is $X$'s PDF stretched by a factor of $a$ horizontally
- Mean
  - $\mathrm{E}[\mathrm{X}] = \int x\, f(x)\, dx$
  - $\mathrm{E}[r(X)] = \int r(x) f(x)\, dx$
- Variance
  - $\mathrm{Var}(X) = \sigma^2 = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - (\mathrm{E}[X])^2$
- Properties
  - $\mathrm{E}[aX] = a\,\mathrm{E}[X]$
  - $\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$
  - $\mathrm{E}[X + b] = \mathrm{E}[X] + b$
  - $\mathrm{Var}(X + b) = \mathrm{Var}(X)$

# Outline

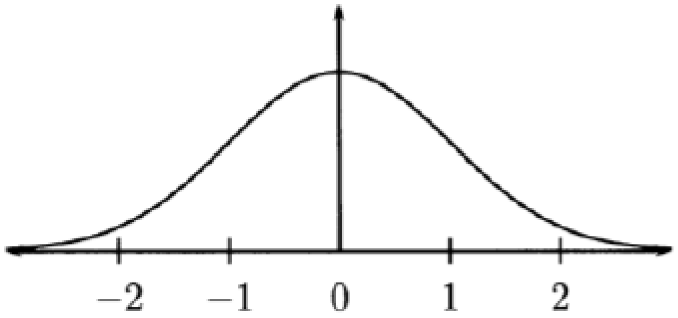


- Calculating probabilities about Gaussians

- Multivariate Random Variables
  - Joint distribution
  - Marginal distribution

# The standard Gaussian distribution

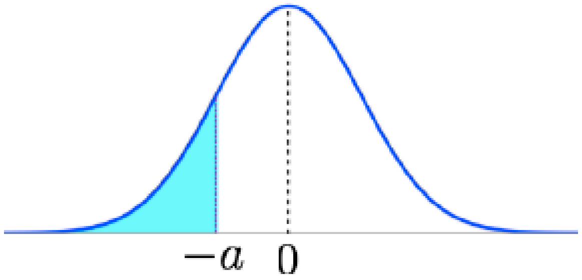
- Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$

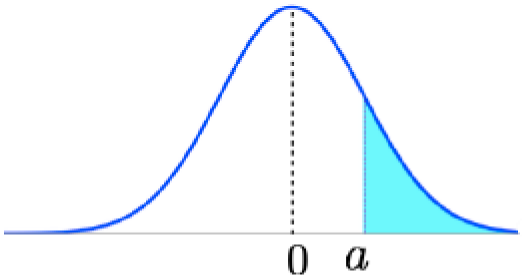


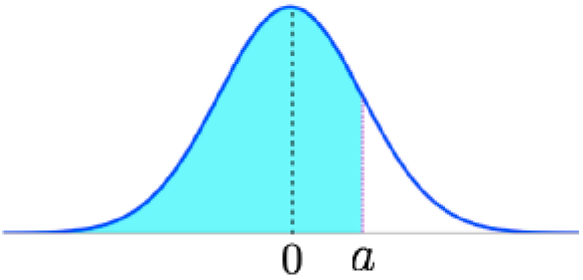

- Denoted by $Z \sim N(0,1)$
- Its PDF denoted by $\phi(z)$, and CDF denoted by $\Phi(z)$

# Calculating probabilities about Gaussians

- Symmetry of $\phi$ => $\Phi(-a) = 1 - \Phi(a)$




$\Phi(-a) = P(Z \leq -a)$ $= P(Z \geq a)$ $= 1 - P(Z \leq a) = 1 - \Phi(a)$

# Calculating probabilities about Gaussians

- Suppose $X \sim N(5, 2^2)$, how can I calculate $P(1 < X < 8)$?

- Transform $X$ into another variable:
  - $X \sim N(\mu, \sigma^2)$: $E[X] = \mu, Var[X] = \sigma^2$

- What is mean and variance for the following transformations of X?

  $\Rightarrow$ $X - \mu$

  $\Rightarrow$ $\frac{X-\mu}{\sigma}$

$$\mathrm{E}[aX] = a\,\mathrm{E}[X]$$
$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$
$$\mathrm{E}[X + b] = \mathrm{E}[X] + b$$
$$\mathrm{Var}(X + b) = \mathrm{Var}(X)$$

# Calculating probabilities about Gaussians

- Suppose $X \sim N(5, 2^2)$, how can I calculate $P(1 < X < 8)$?

- Transform $X$ into standard normal $Z$:
  - $X \sim N(\mu, \sigma^2)$

    $\Rightarrow X - \mu \sim N(0, \sigma^2)$

    $\Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$

$$\mathrm{E}[aX] = a\,\mathrm{E}[X]$$
$$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$$
$$\mathrm{E}[X + b] = \mathrm{E}[X] + b$$
$$\mathrm{Var}(X + b) = \mathrm{Var}(X)$$

- We can write $\mathrm{P}(a < X < b)$ using $\mathrm{P}(c < Z < d)$, which in turn can be written in $\Phi$.

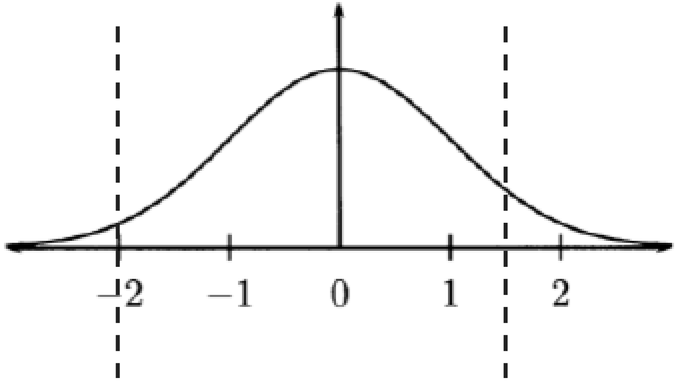
# Calculating probabilities about Gaussians

- $\mathrm{P}(a < X < b)$

$$= P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$

$$= P\left(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$




**Example** Suppose $X \sim N(5, 2^2)$, calculate $P(1 < X < 8)$

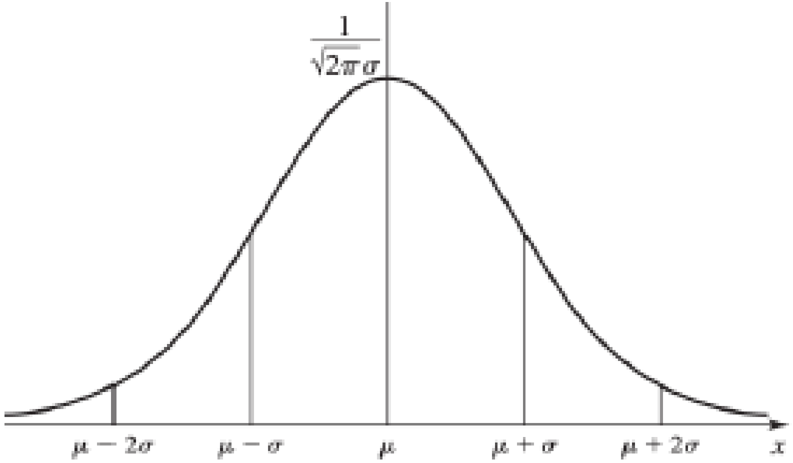
This is $\Phi\left(\frac{8-5}{2}\right) - \Phi\left(\frac{1-5}{2}\right) = \Phi(1.5) - \Phi(-2) = \Phi(1.5) - (1 - \Phi(2))$

```
from scipy.stats import norm
print(norm.cdf(1.5)-(1-norm.cdf(2)))
```

```
0.9104426667829627
```

# Calculating probabilities about Gaussians

- What is the probability that a Gaussian RV $X$ is within $k$ ($k = 1, 2, \ldots$) std of its mean?




- $\mathrm{P}(\mu - k\sigma \leq X < \mu + k\sigma)$

# Calculating probabilities about Gaussians

- $p_k = \mathrm{P}(\mu - k\sigma \le X < \mu + k\sigma)$

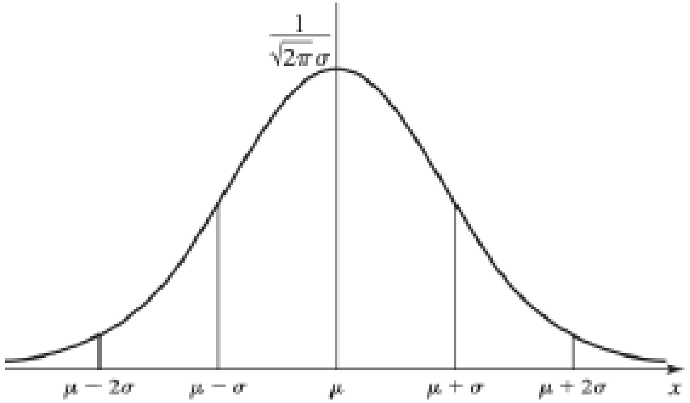
$= P\left(-k < \frac{X-\mu}{\sigma} < k\right)$

$= P(-k < Z < k)$

$= \Phi(k) - (1 - (\Phi(k))$

$= 2\Phi(k) - 1$

| $k$ | $p_k$ |
|---|---|
| 1 | 0.6826 |
| 2 | 0.9544 |
| 3 | 0.9974 |
| 4 | 0.99994 |

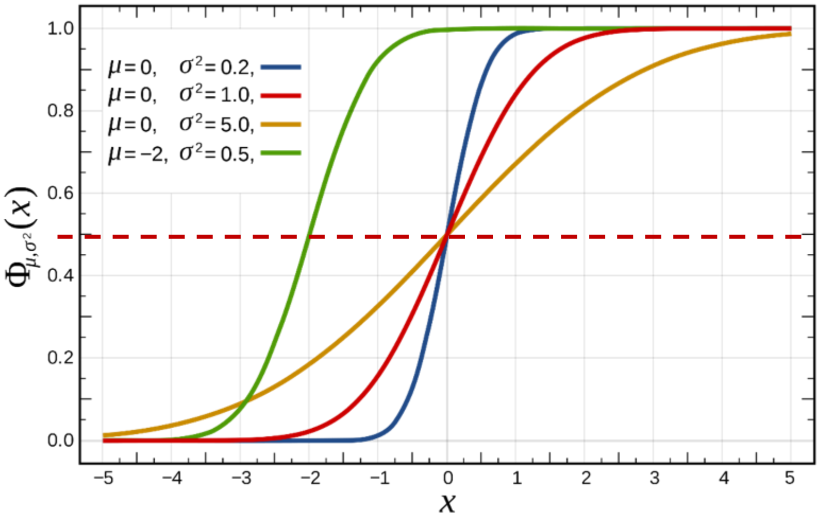



In words,

- With probability about 95%, $X$ is within 2 std of its mean
- With overwhelming prob. (99.7%), $X$ within 3 std of mean

# CDF of Gaussian Distributions

- $F$: CDF of Gaussian $N(\mu, \sigma^2)$



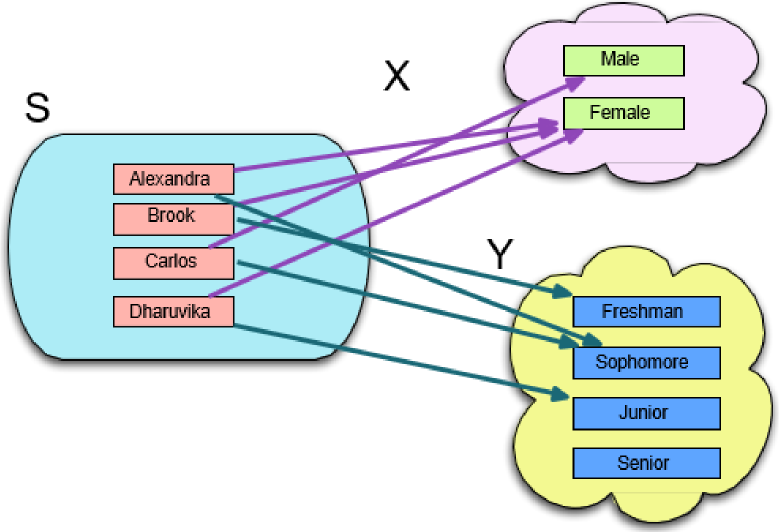

Q: what is $F(\mu)$?

- $F(\mu) = \frac{1}{2}$

- $F(x)$ changes fast near $\mu$
- $F$'s "sensitive range" is about $[\mu - 3\sigma, \mu + 3\sigma]$

# Multivariate Random Variables

# Multivariate RVs: example




- X: people -> their genders
- Y: people -> their class year
- We'd like to answer questions such as: does X and Y have a correlation?
  - I.e., is a student in higher class year more likely to be male?
- We call (X, Y) a multivariate RV, and will study its *joint* distribution

# Joint distribution of discrete RVs

- The joint PMF (probability mass function) of discrete random variables X, Y:

$$f(x, y) = P(X = x, Y = y)$$

**Examples**

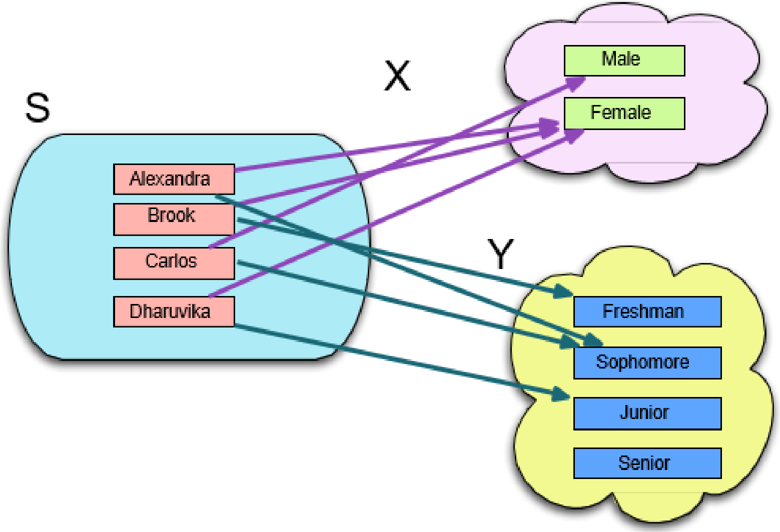
Alexandra

$$P(X = \mathrm{Fem}, Y = \mathrm{Soph}) = \frac{1}{4}$$

Dharuvika

$$P(X = \mathrm{Fem}, Y = \mathrm{Jun}) = \frac{1}{4}$$

…



# Joint distribution of discrete RVs

- X: # of cars owned by a randomly selected household
- Y: # of computers owned by the same household
- Joint pmf shown with a table

| | $y$ | | | |
|---|---|---|---|---|
| $x$ | 1 | 2 | 3 | 4 |
| 1 | 0.1 | 0 | 0.1 | 0 |
| 2 | 0.3 | 0 | 0.1 | 0.2 |
| 3 | 0 | 0.2 | 0 | 0 |

- Probability that a randomly selected household has $\geq 2$ cars and $\geq 2$ computers?
  - $P(X \geq 2, Y \geq 2) = 0.5$

# Marginal distributions

Given joint distribution of $(X, Y)$, need distribution of one of them, say $X$. Named the ***marginal distribution*** of $X$.

- How to find $P(X = x)$?
- Using law of total probability:

$$f_1(x) = \sum_y f(x, y)$$

| | $y$ | | | |
|---|---|---|---|---|
| $x$ | 1 | 2 | 3 | 4 |
| 1 | 0.1 | 0 | 0.1 | 0 |
| 2 | 0.3 | 0 | 0.1 | 0.2 |
| 3 | 0 | 0.2 | 0 | 0 |

- This operation is called *marginalization* ('marginalizing out variable Y', or variable elimination)

# Marginal distributions

| $x$ | $y$ = 1 | 2 | 3 | 4 | Total |
|---|---|---|---|---|---|
| 1 | 0.1 | 0 | 0.1 | 0 | 0.2 |
| 2 | 0.3 | 0 | 0.1 | 0.2 | 0.6 |
| 3 | 0 | 0.2 | 0 | 0 | 0.2 |
| Total | 0.4 | 0.2 | 0.2 | 0.2 | 1.0 |

$f_1$: *marginal distribution* of $X$

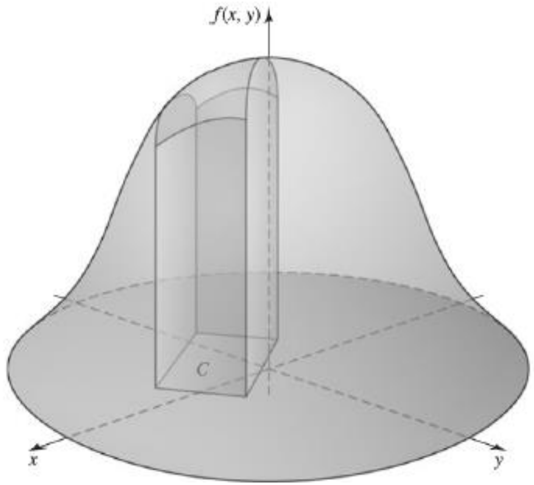
$f_2$: *marginal distribution* of $Y$

$$f_1(X=1) = \sum_y f(1,y) = 0.1 + 0 + 0.1 + 0 = 0.2$$

# Joint distribution of continuous RVs

- Any continuous random vector (X,Y) has a *joint probability density function* (PDF) $f(x, y)$, such that for all $C$,

$$P\big((X, Y) \in C\big) = \iint_C f(x, y)\, dx\, dy$$




$f(x, y)$: represent a 2D surface

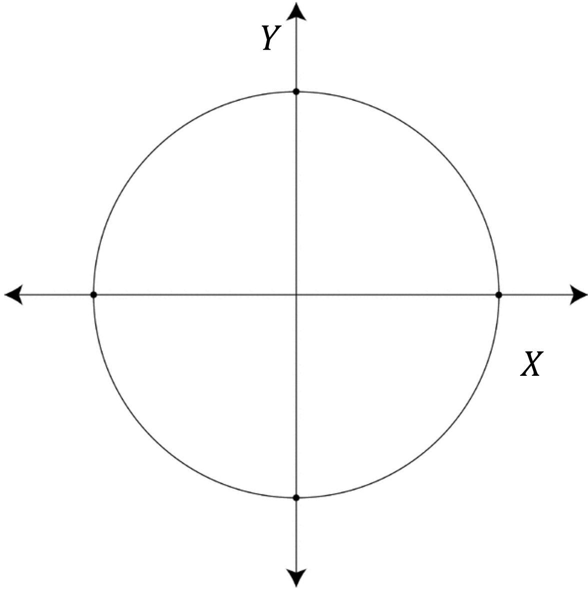
double integral: the *volume* under the surface

Properties:

- $f$ is nonnegative
- $\iint_{R^2} f(x, y)\, dx\, dy = 1$ ($R^2 =$ the whole x-y plane)
    - $P\big((X, Y) \in R^2\big) = 1$

# Example: dartboard

- Dartboard with center (0,0) and radius 1; dart lands uniformly at random on the board




- What is the joint PDF of $(X, Y)$?

- Fact: the PDF is

$$f(x, y) = \begin{cases} c, x^2 + y^2 \leq 1 \\ 0, \text{otherwise} \end{cases}$$

- This is called "the Uniform distribution over the unit disk"

# Example: dartboard

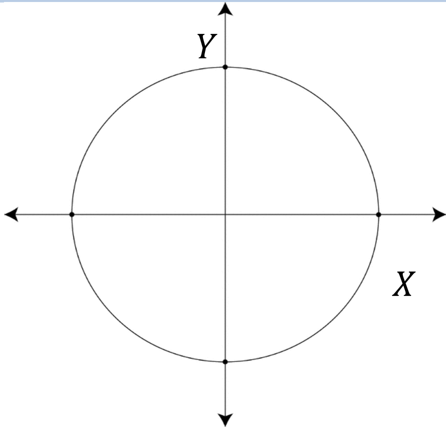
The PDF of $X, Y$ is

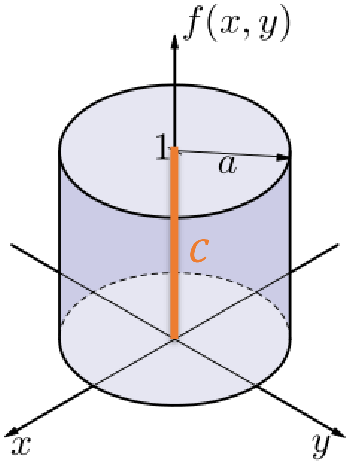
$$f(x,y) = \begin{cases} c, x^2 + y^2 \leq 1 \\ 0, \text{otherwise} \end{cases}$$

Can we find $c$?

Observe: volume under $f(x,y)$ is $\pi c$ (cylinder) which must also be 1

Therefore, $c = 1/\pi$



# Marginal distribution of continuous RV

Given joint distribution of continuous RV $(X, Y)$, how to find $X$'s PDF $f_1$?

**Fact (marginalization)** $f_1(x) = \int_R f(x, y)\, dy$

Replacing summation with integration in the continuous case ('marginalizing / integrating out variable Y')

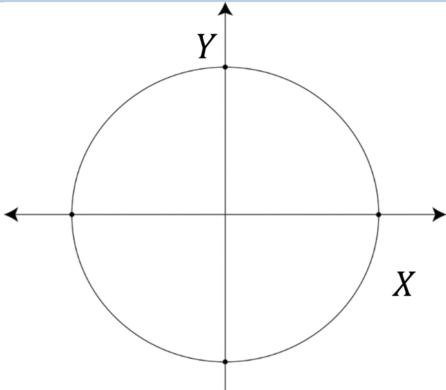
How about $Y$'s PDF $f_2$?

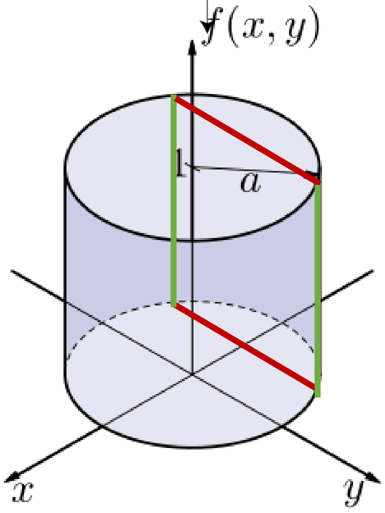
- Marginalize out $X$

# Example: dartboard

The PDF of $X, Y$ is

$$f(x,y) = \begin{cases} \frac{1}{\pi}, x^2 + y^2 \leq 1 \\ 0, \text{otherwise} \end{cases}$$



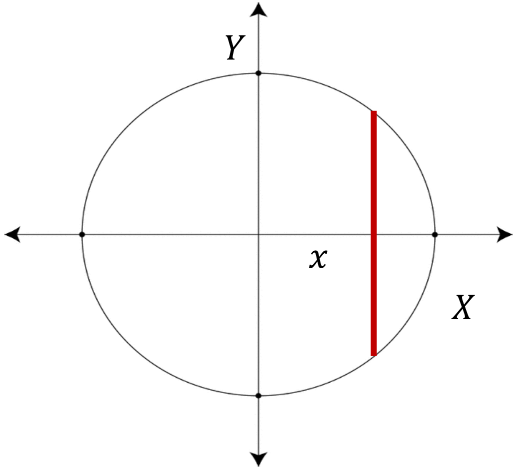



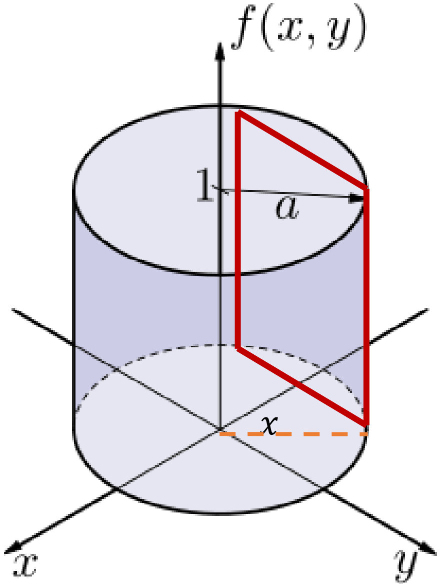

What is the marginal distribution over $X$?

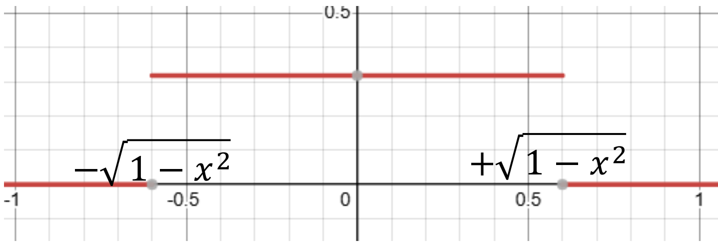
$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy$$

How to find this integral?

# Example: dartboard



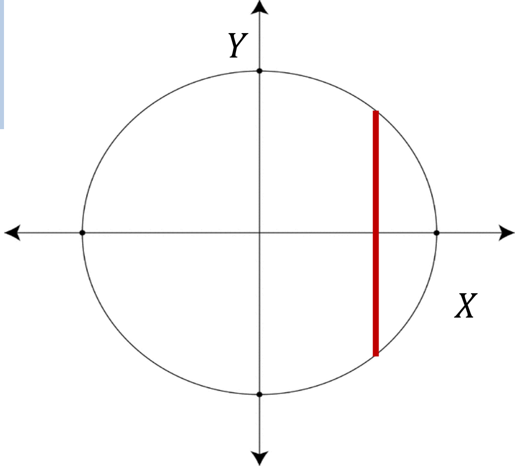




For a fixed $x \in [-1, 1]$, we can think of $f(x)$ is the area of the slice:

- height: $\frac{1}{\pi}$, width: $2 \cdot \sqrt{1-x^2}$
- $f_1(x) = \frac{2}{\pi} \cdot \sqrt{1-x^2}$

# Example: dartboard




- In summary,

$$f(x) = \begin{cases} \frac{2}{\pi} \cdot \sqrt{1-x^2}, x \in [-1,1] \\ 0, \qquad \text{otherwise} \end{cases}$$


$X$'s distribution is NOT $\text{Uniform}([-1,1])$!

Actually makes sense: $X$ closer to 1 is harder to be hit

# Joint distribution of more than 3 RVs

- We can consider the joint distribution of more than 3 random variables,
  - E.g. (A,B,C), A = gender, B = class year, C = blood type
- Discrete RVs: can still define joint PMFs

| $a$ | $b$ | $c$ | $P(A=a, B=b, C=c)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.06 |
| 0 | 0 | 1 | 0.09 |
| 0 | 1 | 0 | 0.08 |
| 0 | 1 | 1 | 0.12 |
| 1 | 0 | 0 | 0.06 |
| 1 | 0 | 1 | 0.24 |
| 1 | 1 | 0 | 0.10 |
| 1 | 1 | 1 | 0.25 |

# Marginalization

| $a$ | $b$ | $c$ | $P(A=a,B=b,C=c)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.06 |
| 0 | 0 | 1 | 0.09 |
| 0 | 1 | 0 | 0.08 |
| 0 | 1 | 1 | 0.12 |
| 1 | 0 | 0 | 0.06 |
| 1 | 0 | 1 | 0.24 |
| 1 | 1 | 0 | 0.10 |
| 1 | 1 | 1 | 0.25 |

Given the joint distribution of $(A, B, C)$

- What is the distribution of $A$?
  - Need to find $P(A = 0)$ and $P(A = 1)$

$$P(A=0) = \sum_{b,c} P(A=0, B=b, C=c)$$

Marginalization: summing over irrelevant variables

- What is the joint distribution of $(A, B)$?
  - Need to find $P(A = 0, B = 0), \ldots ,\ P(A = 1, B = 1)$

$$P(A=0, B=0) = \sum_{c} P(A=0, B=0, C=c)$$

# Marginalization for continuous RVs

Suppose joint PDF of $(A, B, C)$ is $f(a, b, c)$

- What is the PDF of $A$?

$$f_A(a) = \iint_{R^2} f(a, b, c)\ db\, dc$$

- What is the joint PDF of $(A, B)$?

$$f_{A,B}(a, b) = \int_R f(a, b, c) dc$$

Marginalization: summing over irrelevant variables

- These operations generalize to joint PDFs of more RVs..