CSE 312

Foundations of Computing II

18: Finish Polling, Joint Distributions

Agenda

- CLT and Polling
- Joint Distributions
  - Cartesian Products
  - Joint PMFs and Joint Range
  - Marginal Distribution
  - Analogues for continuous distributions
  - LOTUS for joint distns

Central Limit Theorem

$\mathbb{E}[S_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n\mu$

$\mathrm{Var}(S_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\sigma^2$

$X_1, \dots, X_n$ i.i.d., each with expectation μ and variance σ^2

Define $S_n = X_1 + \cdots + X_n$,

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Then distribution of $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to that of a normal distribution with mean 0 and variance 1 as $n \to \infty$.

Central Limit Theorem

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

Theorem. (Central Limit Theorem) The CDF of Y_n converges to the CDF of the standard normal $\mathcal{N}(0,1)$, i.e.,

$$\lim_{n\to\infty} P(Y_n \le y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2} \mathrm{d}x$$

Central Limit Theorem

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

Theorem. (Central Limit Theorem) Suppose that $X_1, \ldots, X_n$ are i.i.d. with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$ The CDF of Y_n converges to the CDF of the standard normal $\mathcal{N}(0,1)$, i.e.,

$$\lim_{n\to\infty} P(Y_n \leq y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2} \mathrm{d}x$$

Other versions:

- $X_1 + \cdots + X_n$ approximately $\mathcal{N}(n\mu, n\sigma^2)$
- $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Outline of CLT steps

- Write the event you are interested in, in terms of a sum of i.i.d. random variables.
- Normalize RV to have mean 0 and standard deviation 1.
- Write event in terms of Φ, the CDF of a $\mathcal{N}(0,1)$.
- Look up in table.

Outline of CLT steps – extra step if random variables are discrete.

- Write the event you are interested in, in terms of a sum of i.i.d. random variables.
- Apply continuity correction if RVs are discrete (see Feb 15 section and Tsun Section 5.7.4)
- Normalize RV to have mean 0 and standard deviation 1.
- Write event in terms of Φ, the CDF of a $\mathcal{N}(0,1)$.
- Look up in table.

Agenda

- Central Limit Theorem (CLT)
- Polling

Formalizing Polls

Population size N, true fraction of voting in favor p, sample size n.

Problem: We don't know p, want to estimate it

Polling Procedure

for $i = 1, \ldots, n$:

1. Pick uniformly random person to call (prob: $1/N$)
2. Ask them how they will vote

$$X_i = \begin{cases} 1, & \text{voting in favor} \\ 0, & \text{otherwise} \end{cases}$$

Report our estimate of p: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Random Variables

Each X_i is Bernoulli (p)

Type	$\mathbb{E}[X_i]$	$\mathrm{Var}(X_i)$
Bernoulli	p	$p(1-p)$

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has expectation. p and variance. $p(1-p)/n$

Therefore, by the Central Limit Theorem, as $n \to \infty$,

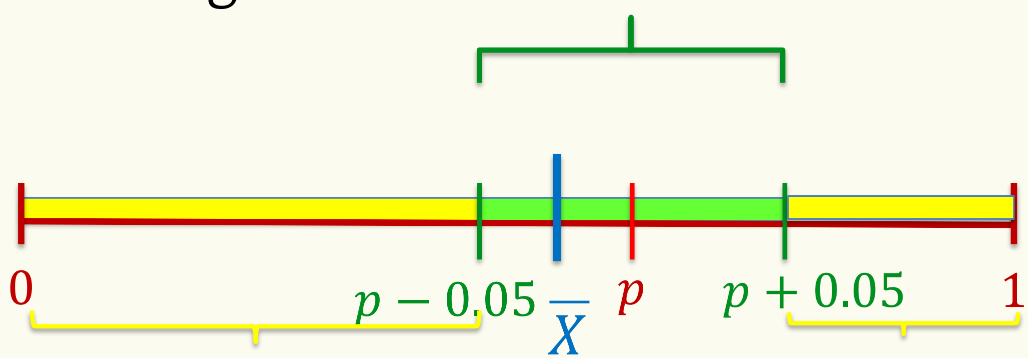
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Therefore, as n gets larger and larger, more and more probability mass concentrates around p

Bounding Error

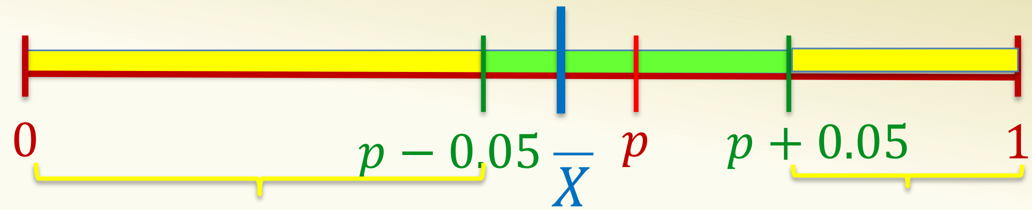
Goal: Find the value of n such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p

Get good estimate if $\overline{X}$ lands in this region



Question: for what n is $P\left(\left|\overline{X} - p\right| > 0.05\right) \leq 0.02$

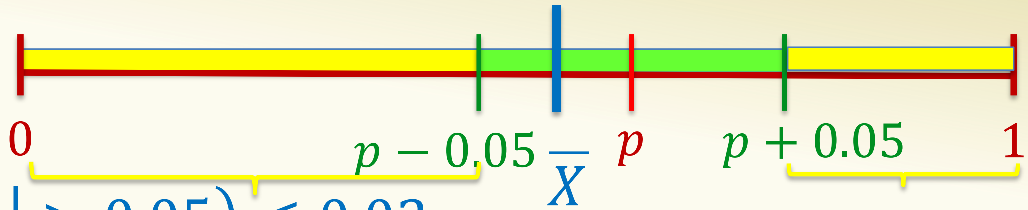
Recap



Goal: Find the value of n such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p

1. Define question. For what n is $P(|\overline{X} - p| > 0.05) \leq 0.02$
2. Apply CLT: By CLT $\overline{X} \rightarrow \mathcal{N}(\mu, \sigma^2)$ where $\mu = p$ and $\sigma^2 = p(1-p)/n$
3. Convert to a standard normal. Specifically, define $Z = \frac{\overline{X}-\mu}{\sigma} = \frac{\overline{X}-p}{\sigma}$. Then, by the CLT $Z \rightarrow \mathcal{N}(0, 1)$
4. Solve for n

In more detail



1. The question: for what n is $P\left(\left|\overline{X}-p\right| > 0.05\right) \le 0.02$

2. By CLT $\overline{X} \to \mathcal{N}(\mu, \sigma^2)$ where $\mu = p$ and $\sigma^2 = p(1-p)/n$

3. $P\left(\left|\overline{X}-p\right| > 0.05\right) = P\left(\overline{X}-p > 0.05\right) + P\left(\overline{X}-p < -0.05\right)$

4. $P\left(\left|\overline{X}-p\right| > 0.05\right) \le 0.02$ equivalent to $P\left(\overline{X}-p > 0.05\right) \le 0.01$

5. Equivalent to $P\left({}^{\overline{X}-p}\Big/_{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} > {}^{0.05}\Big/_{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} \right) \le 0.01$

6. Equivalent to $P\left(Z > \frac{0.05}{\sqrt{p(1-p)}/\sqrt{n}}\right) \le 0.01$

Recall!!

For what a is

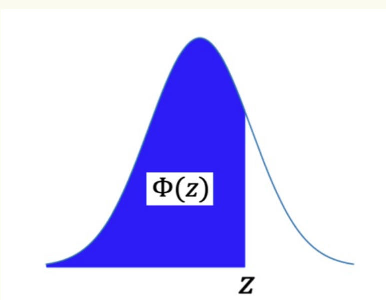
$$P(Z > a) \leq 0.01\ ?$$

Equivalently

$$P(Z \leq a) \geq 0.99\ ?$$

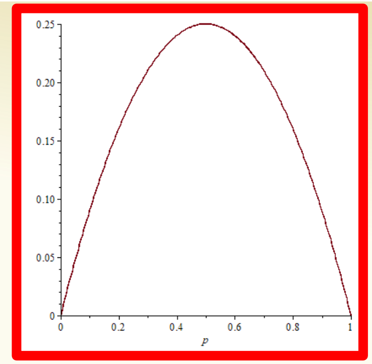
For any $a \geq 2.33$

$$P(Z > a) \leq 0.01.$$



Φ Table: $\mathbb{P}(Z \leq z)$ when $Z \sim \mathcal{N}(0,1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999



So we know that if $\frac{0.05}{\sqrt{p(1-p)}/\sqrt{n}} \geq 2.33$

Then $P\left(Z > \frac{0.05}{\sqrt{p(1-p)}/\sqrt{n}}\right) \leq 0.01$.

So just need to solve this: $\sqrt{n} \geq \frac{2.33\sqrt{p(1-p)}}{0.05}$

Since we don't know p, we will just take n so it works for all p.
RHS is largest with p = 0.5. So we take

$\sqrt{n} \geq \frac{2.33 \cdot 0.5}{0.05}$ or $\sqrt{n} \geq 23.3$

Then $n \geq 543 \geq (23.3)^2$ would be good enough.

- Since only have $Z \to \mathcal{N}(0, 1)$ so there is some loss due to approximation error.
- So should increase n a bit more to be safe.

Zooming back out: we found an approximate``confidence interval”

We are trying to estimate some parameter (e.g. p). We output an estimator $\overline{X}$ such that $P\left(\left|\overline{X} - p\right| > \epsilon\right) \leq \delta$ for some (ϵ, δ).

- Often found using CLT

- We say that we are $(1 - \delta)$*100% confident that the result of our poll ($\overline{X}$) is an accurate estimate of p to within ϵ*100% percent.

- In our example, $(\epsilon = 0.05, \delta = 0.02)$.

Idealized Polling

So far, we have been discussing "idealized polling". Real life is normally not so nice ☹

Assumed we can sample people uniformly at random, not really possible in practice

- Not everyone responds
- Response rates might differ in different groups
- Will people respond truthfully?

Makes polling in real life much more complex than this idealized model!

Agenda

- CLT and Polling
- Joint Distributions
  - Cartesian Products
  - Joint PMFs and Joint Range
  - Marginal Distribution
  - Analogues for continuous distributions
  - LOTUS for joint distns

Why joint distributions?

- Given all of its user's ratings for different movies, and any preferences you have expressed, Netflix wants to recommend a new movie for you.
- Given a large amount of medical data correlating symptoms and personal history with diseases, predict what is ailing a person with a particular medical history and set of symptoms.
- Given current traffic, pedestrian locations, weather, lights, etc. decide whether a self-driving car should slow down or come to a stop

Review Cartesian Product

Definition. Let A and B be sets. The **Cartesian product** of A and B is denoted

$$A \times B = \{(a, b) : a \in A, b \in B\}$$

Example.

$$\{1,2,3\} \times \{4, 5\} = \{(1, 4), (1,5), (2,4), (2,5), (3, 4), (3,5)\}$$

If A and B are finite sets, then $|A \times B| = |A| \cdot |B|$.

The sets don't need to be finite! You can have $\mathbb{R} \times \mathbb{R}$ (often denoted $\mathbb{R}^2$)

Joint PMFs and Joint Range

Definition. Let X and Y be discrete random variables. The **Joint PMF** of X and Y is

$$p_{X,Y}(a,b) = P(X = a, Y = b)$$

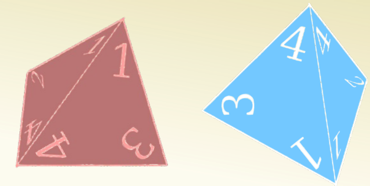
Definition. The **joint range** of $p_{X,Y}$ is

$$\Omega_{X,Y} = \{(c,d) : p_{X,Y}(c,d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that

$$\sum_{(s,t)\in\Omega_{X,Y}} p_{X,Y}(s,t) = 1$$

Example – Weird Dice



Suppose I roll two fair 4-sided die independently. Let X be the value of the first die, and Y be the value of the second die.

$\Omega_X = \{1,2,3,4\}$ and $\Omega_Y = \{1,2,3,4\}$

In this problem, the joint PMF is if

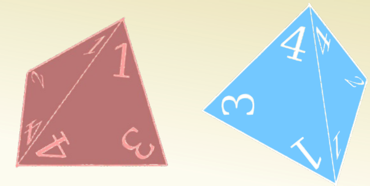
$$p_{X,Y}(x,y) = \begin{cases} 1/16 & \text{if } x,y \in \Omega_{X,Y} \\ 0 & \text{otherwise} \end{cases}$$

X\Y	1	2	3	4
1	1/16	1/16	1/16	1/16
2	1/16	1/16	1/16	1/16
3	1/16	1/16	1/16	1/16
4	1/16	1/16	1/16	1/16

and the joint range is (since all combinations have non-zero probability)

$\Omega_{X,Y} = \Omega_X \times \Omega_Y$

Example – Weirder Dice



Suppose I roll two fair 4-sided die independently. Let X be the value of the first die, and Y be the value of the second die. Let $U = \min(X, Y)$ and $W = \max(X, Y)$

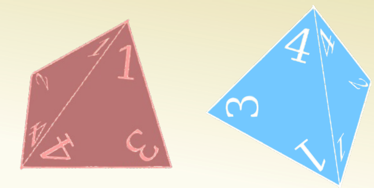
$\Omega_U = \{1,2,3,4\}$ and $\Omega_W = \{1,2,3,4\}$

$\Omega_{U,W} = \{(u, w) \in \Omega_U \times \Omega_W : u \leq w\} \neq \Omega_U \times \Omega_W$

What is $p_{U,W}(1, 3) = P(U = 1, W = 3)$?

U\W	1	2	3	4
1				
2				
3				
4				

Example – Weirder Dice



Suppose I roll two fair 4-sided die independently. Let X be the value of the first die, and Y be the value of the second die. Let $U = \min(X, Y)$ and $W = \max(X, Y)$

$\Omega_U = \{1,2,3,4\}$ and $\Omega_W = \{1,2,3,4\}$

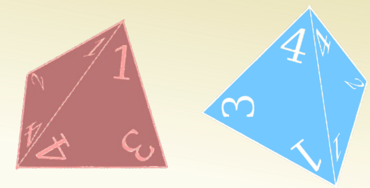
$\Omega_{U,W} = \{(u, w) \in \Omega_U \times \Omega_W : u \leq w\} \neq \Omega_U \times \Omega_W$

The joint PMF $p_{U,W}(u, w) = P(U = u, W = w)$ is

$$p_{U,W}(u,w) = \begin{cases} 2/16 & \text{if } (u,w) \in \Omega_U \times \Omega_W \text{ where } w > u \\ 1/16 & \text{if } (u,w) \in \Omega_U \times \Omega_W \text{ where } w = u \\ 0 & \text{otherwise} \end{cases}$$

U\W	1	2	3	4
1	1/16	2/16	2/16	2/16
2	0	1/16	2/16	2/16
3	0	0	1/16	2/16
4	0	0	0	1/16

Example – Weirder Dice

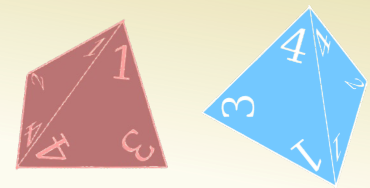


Suppose I roll two fair 4-sided die independently. Let X be the value of the first die, and Y be the value of the second die. Let $U = \min(X, Y)$ and $W = \max(X, Y)$

Suppose we didn't know how to compute $P(U = u)$ directly. Can we figure it out if we know $p_{U,W}(u, w)$?

U\W	1	2	3	4
1	1/16	2/16	2/16	2/16
2	0	1/16	2/16	2/16
3	0	0	1/16	2/16
4	0	0	0	1/16

Example – Weirder Dice



Suppose I roll two fair 4-sided die independently. Let X be the value of the first die, and Y be the value of the second die. Let $U = \min(X, Y)$ and $W = \max(X, Y)$

Suppose we didn't know how to compute $P(U = u)$ directly. Can we figure it out if we know $p_{U,W}(u, w)$?

Just apply LTP over the possible values of W:

$$p_U(1) = 7/16$$

$$p_U(2) = 5/16$$

$$p_U(3) = 3/16$$

$$p_U(4) = 1/16$$

U\W	1	2	3	4
1	1/16	2/16	2/16	2/16
2	0	1/16	2/16	2/16
3	0	0	1/16	2/16
4	0	0	0	1/16

Marginal PMF

Definition. Let X and Y be discrete random variables and $p_{X,Y}(a,b)$ their joint PMF. The **marginal PMF** of X

$$p_X(a) = \sum_{b \in \Omega_Y} p_{X,Y}(a,b)$$

Similarly, $p_Y(b) = \sum_{a \in \Omega_X} p_{X,Y}(a,b)$

Continuous distributions on $\mathbb{R}\times\mathbb{R}$

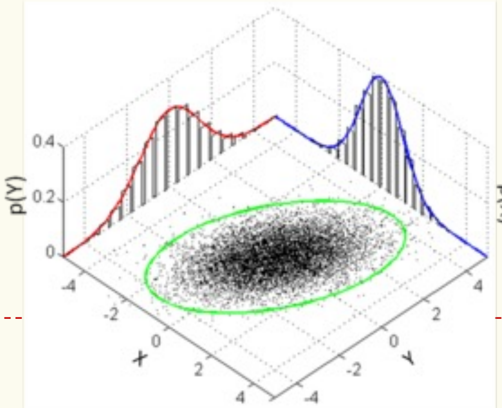
Definition. The **joint probability density function (PDF)** of continuous random variables X and Y is a function $f_{X,Y}$ defined on $\mathbb{R}\times\mathbb{R}$ such that

- $f_{X,Y}(x,y) \geq 0$ for all $x, y \in \mathbb{R}$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x\ \mathrm{d}y = 1$

for $A \subseteq \mathbb{R}\times\mathbb{R}$ the probability that $(X,Y) \in A$ is $\iint_A f_{X,Y}(x,y)\ \mathrm{d}x\mathrm{d}y$

The **(marginal) PDFs** f_X and f_Y are given by

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ \mathrm{d}y$
- $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ \mathrm{d}x$



Independence and joint distributions

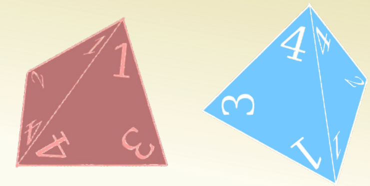
Definition. Discrete random variables X and Y are **independent** iff

- $p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y)$ for all $x \in \Omega_X, y \in \Omega_Y$

Definition. Continuous random variables X and Y are **independent** iff

- $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all $x, y \in \mathbb{R}$

Example – Weirder Dice



Suppose I roll two fair 4-sided die independently. Let X be the value of the first die, and Y be the value of the second die. Let $U = \min(X, Y)$ and $W = \max(X, Y)$

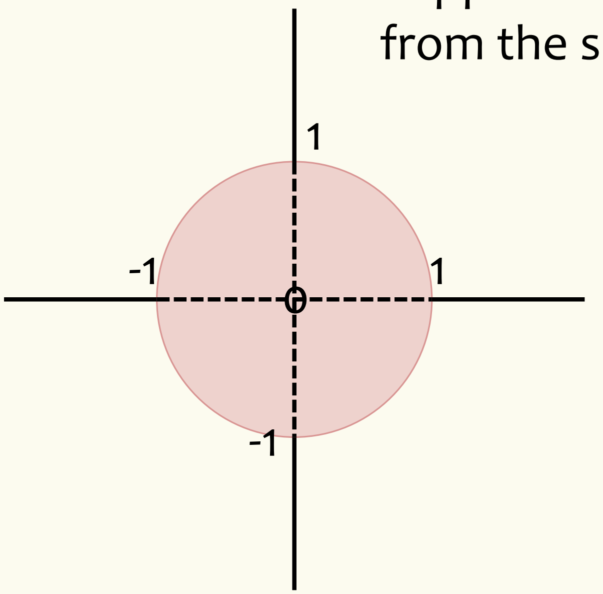
$\Omega_U = \{1,2,3,4\}$ and $\Omega_W = \{1,2,3,4\}$

$$\Omega_{U,W} = \{(u, w) \in \Omega_U \times \Omega_W : u \leq w\} \neq \Omega_U \times \Omega_W$$

Are U and W independent?

U\W	1	2	3	4
1	1/16	2/16	2/16	2/16
2	0	1/16	2/16	2/16
3	0	0	1/16	2/16
4	0	0	0	1/16

Example – Uniform distribution on a unit disk



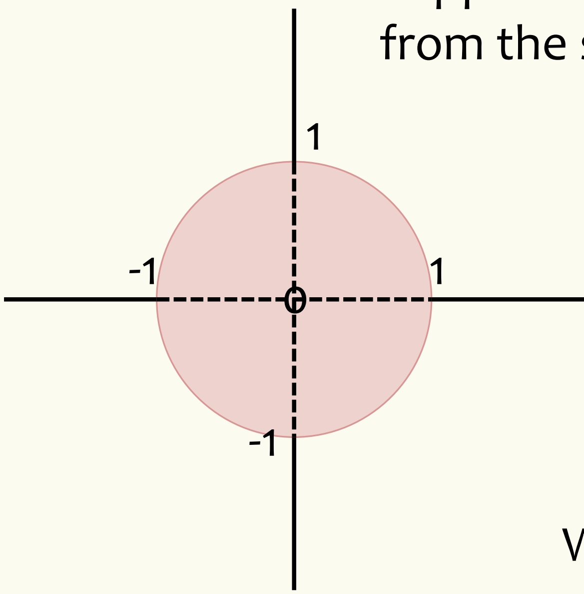
Suppose that a pair of random variables (X, Y) is chosen uniformly from the set of real points (x, y) such that $x^2 + y^2 \leq 1$

This is a disk of radius 1 which has area π

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Are X and Y independent?

Example – Uniform distribution on a unit disk



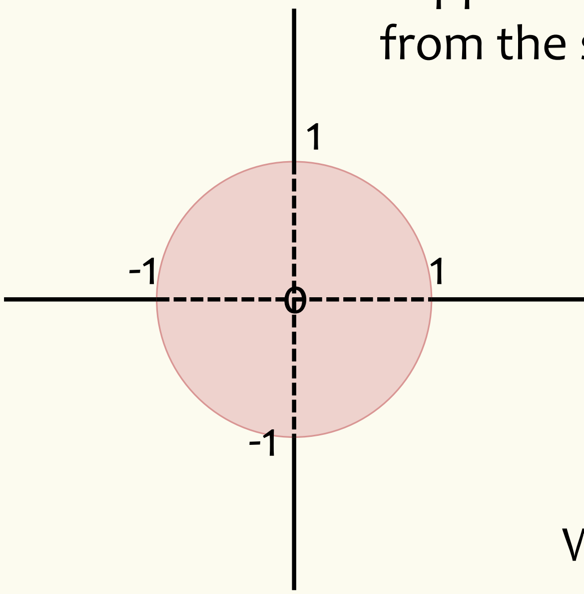
Suppose that a pair of random variabes (X, Y) is chosen uniformly from the set of real points (x, y) such that $x^2 + \mathrm{y}^2 \leq 1$

This is a disk of radius 1 which has area π

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

What is $f_X(x)$?

Example – Uniform distribution on a unit disk



Suppose that a pair of random variabes (X, Y) is chosen uniformly from the set of real points (x, y) such that $x^2 + y^2 \leq 1$

This is a disk of radius 1 which has area π

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

What is $f_X(x)$?

$$f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} \mathrm{d}y$$

$$= 2\sqrt{1-x^2}/\pi$$

Joint Expectation

Definition. Let X and Y be discrete random variables and $p_{X,Y}(a,b)$ their joint PMF. The **expectation** of some function $g(x,y)$ with inputs X and Y

$$\mathbb{E}[g(X,Y)] = \sum_{a \in \Omega_X} \sum_{b \in \Omega_Y} g(a,b) \cdot p_{X,Y}(a,b)$$

Reference Sheet (with continuous RVs)

	Discrete	Continuous
Joint PMF/PDF	$p_{X,Y}(x,y) = P(X = x, Y = y)$	$f_{X,Y}(x,y) \neq P(X = x, Y = y)$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t \le x} \sum_{s \le y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s)\,ds\,dt$
Normalization	$\sum_{x} \sum_{y} p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy = 1$
Marginal PMF/PDF	$p_X(x) = \sum_{y} p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy$
Expectation	$E[g(X,Y)] = \sum_{x} \sum_{y} g(x,y) p_{X,Y}(x,y)$	$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy$
Independence	$\forall x, y, p_{X,Y}(x,y) = p_X(x) p_Y(y)$	$\forall x, y, f_{X,Y}(x,y) = f_X(x) f_Y(y)$