Karlsruher Institut für Technologie

Modern Methods of Statistical Data Analysis

From parameter estimation to deep learning — A guided tour of probability

Lecture 2

Fundamental Concepts II

P.-D. Dr. Roger Wolf

roger.wolf@kit.edu

Dr. Pablo Goldenzweig

pablo.goldenzweig@kit.edu

Dr. Jan Kieseler

jan.kieseler@kit.edu

Dr. Slavomira Stefkova

slavomira.stefkova@kit.edu

Program today

- Recap of lecture 1
- Review 5 warmup questions
- (Conditional) independence
- Random variables
- Expectation values, Variance
- Answers to Quiz 1

5’ break

- Tour of important probability mass (density) functions
- Quiz 2

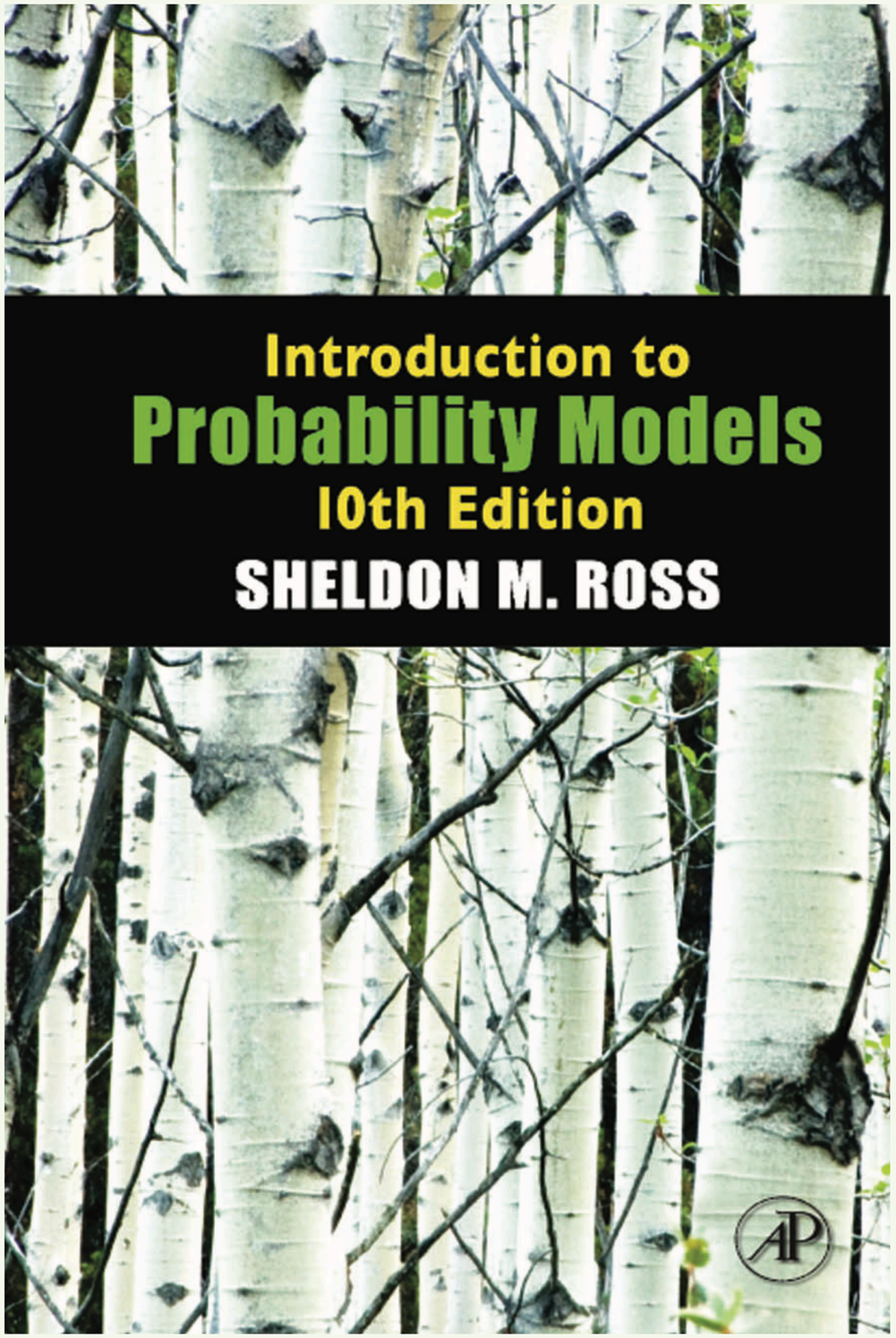
Textbook by S. Ross

ILIAS:

/Reading material / Textbooks / IntroProbModels_SRoss.PDF

Contains many worked out examples and proofs which we will discuss today and next week

(See required reading slide at the end of lecture)



Interpretation of probability

- Although any function satisfying the *3 axioms of Kolmogorov* is a probability, still need to specify interpretation of **elements** in sample space S**.**
  - **Most important interpretations:** *relative frequency* and *subjective probability.*

- **Probability as a relative frequency:** (Frequentist Interpretation)
  - Elements of S correspond to possible outcomes of a repeatable measurement.
  - Subset E of S: occurrence of any of the outcomes in the subset.
    - Often E is called an **event**.
    - An event is said to occur if the outcome of a measurement is in the respective subset.
      - Assign probability for E as:

$$P(E) = \lim_{n\to\infty} \frac{\text{number of occurrences of outcome } E \text{ in } n \text{ measurements}}{n}$$

Interpretation of probability

- **Subjective probability:** (Bayesian Interpretation)
  - Here, elements of S correspond to **hypotheses** or **propositions**, i.e. statements that are either true or false.
  - One interprets the probability associated with a hypothesis as a measure of degree of belief:

$P(E) =$ degree of belief that hypothesis E is true

Interpretation of probability

- **Subjective probability:** (Bayesian Interpretation)
  - Sample space S needs to be constructed such that the elementary hypotheses are mutually exclusive.
    - Otherwise not only one of them can be true.
  - The use of subjective probability is closely related to Bayes' theorem.
    - Subset A appearing therein can be interpreted as the hypothesis that a certain theory is true, and subset B that an experiment will yield a particular result.

$$P(\text{theory} \mid \text{data}) \propto P(\text{data} \mid \text{theory}) \cdot P(\text{theory})$$

Posterior prob. theory is correct given the data

Likelihood or prob. to observe the measured data given theory is true

Prior probability that theory is true

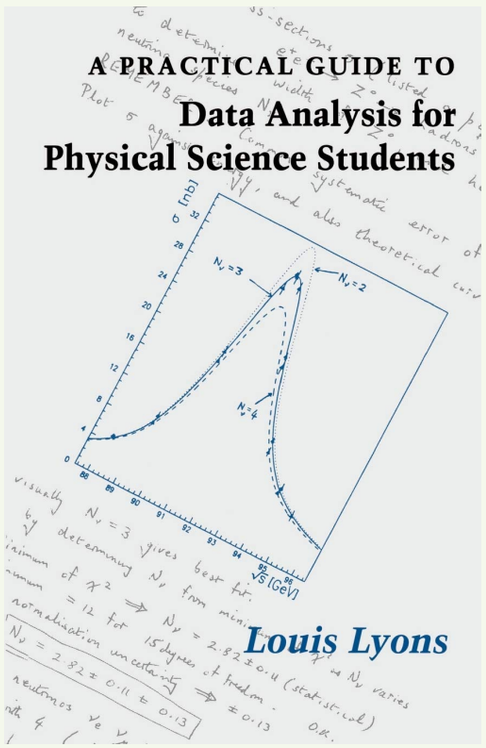
Criticisms of the probability interpretations

- Criticisms of the **frequency** interpretation:
  - $n \to \infty$ can never be achieved in practice. When is n large enough?
  - We want to talk about the probability of events that are not repeatable.
    - Ex. 1: $P(\text{it will rain tomorrow})$, but there is only *one* tomorrow.
    - Ex. 2: $P(\text{universe started with a Big Bang})$, but only *one* universe.
  - P is not an intrinsic property of A, it depends on how the ensemble of possible outcomes was constructed.
    - Ex.: $P(\text{person I talk to is a physicist})$ depends on whether I am in a football stadium or at a scientific conference.

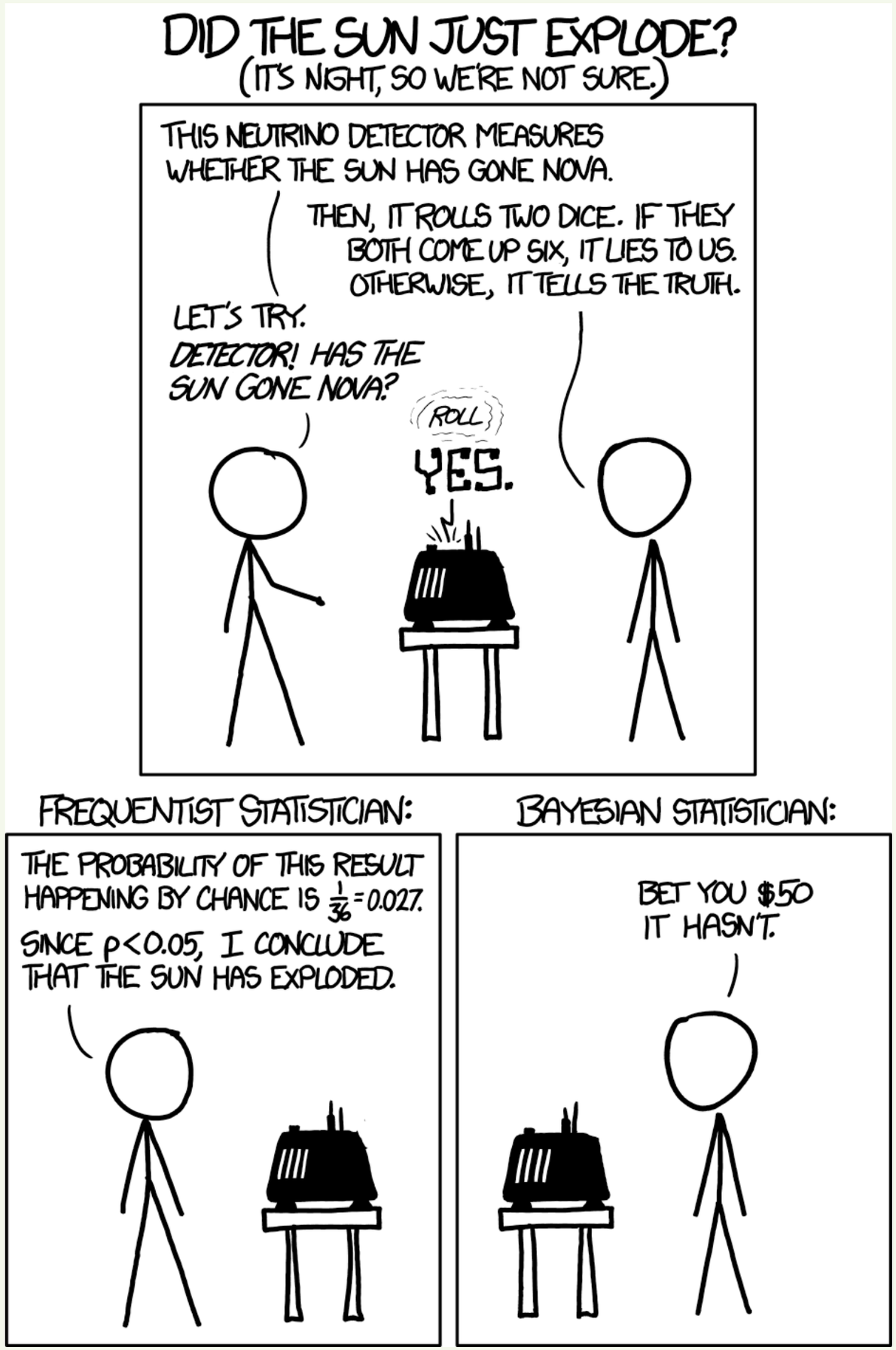
- Criticisms of the **subjective** interpretation:
  - "Subjective" estimates have no place in science.
  - How to quantify the prior state of our knowledge upon which we case our probability estimate?

"**Bayesians** address the questions everyone is interested in by using assumptions that no one believes. **Frequentists** use impeccable logic to deal with an issue that is of no interest to anyone."

—Louis Lyons



Author of:
Data Analysis for Physical Science Students



https://xkcd.com/1132/

Answer Time: 5 quick questions

5 quick questions

- What's the probability to not toss five heads in a row in coin toss?
  - $P = (1 - 0.5^5)$
- Name the following distributions:

Binomial

$$Pr(k; n, p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Poisson

$$f(k; \lambda) = \Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},$$

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]}$$

Breit-Wigner/Lorentzian/Cauchy

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$$

Multi-D Gaussian

- Mean and variance of an independent random variables x_i

Mean: $\bar{x} = \frac{1}{n} \sum_i x_i$

Variance: $\sigma^2 = \frac{1}{n} \sum_i (x_i - \bar{x})^2$ *biased* $\quad \sigma^2 = \frac{1}{n-1} \sum_i (x_i - \bar{x})^2$ *without bias*

- I know what Bayes' theorem is about (Yes/No). → *Now hopefully yes!*
- Write down the definition of a χ^2 function. $z = \sum_{i=1}^{N} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$ $\quad z = (\boldsymbol{x} - \boldsymbol{\mu})^T V^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$

Independence

Independence

Two events E and F are defined as independent if: $P(EF) = P(E)P(F)$

Otherwise E and F are called dependent events

If E and F are independent, then: $P(E \mid F) = P(E)$

Intuition through proof:

$$P(E \mid F) = \frac{P(EF)}{P(F)}$$

def of conditional probability *(Lecture 1, s48)*

$$= \frac{P(E)P(F)}{P(F)}$$

Independence of E and F

$$= P(E)$$

Knowing that F happened does not change our belief that E happened

Generalizing independence

Three events E, F, and G are independent if:

$$\begin{cases} P(EFG) = P(E)P(F)P(G) \text{ and} \\ P(EF) = P(E)P(F) \text{ and} \\ P(EG) = P(E)P(G) \text{ and} \\ P(FG) = P(F)P(G) \end{cases}$$

n events $E_1, E_2, \ldots, E_n$ are independent if:

$$\begin{cases} \text{For } r = 1,\ldots,n: \\ \quad \text{for every subset } E_1, E_2, \ldots, E_r: \\ \quad\quad P(E_1, E_2, \ldots, E_r) = P(E_1), P(E_2), \ldots, P(E_r) \end{cases}$$

Often interested in experiments consisting of independent trials:

- n trials, each with the same set of possible outcomes
- n-way independence: an event in one subset of trials is independent of events in other subsets of trials

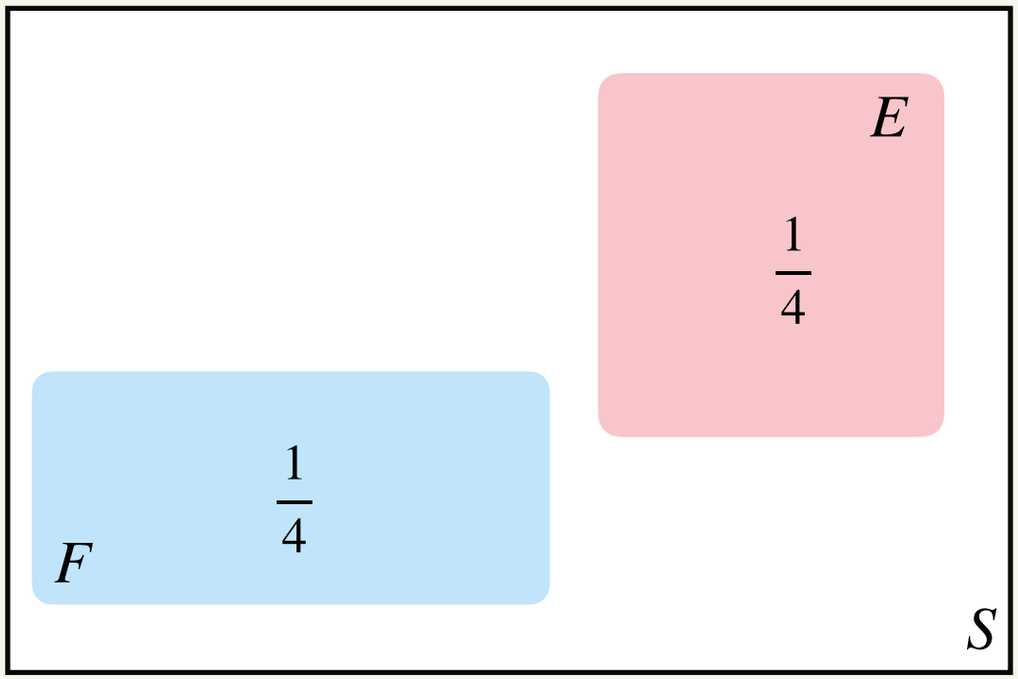
e.g., flip a coin n times, roll a die n times, send a multiple choice survey to n people, send n web requests to k different servers

Independent trials: are when we observe independent sub-experiments, each of which has the same set of possible outcomes

Independence?

Independent events E and F $\Longleftrightarrow$ $P(EF) = P(E)P(F)$, $P(E|F) = P(E)$

Are E and F independent in the following diagrams?

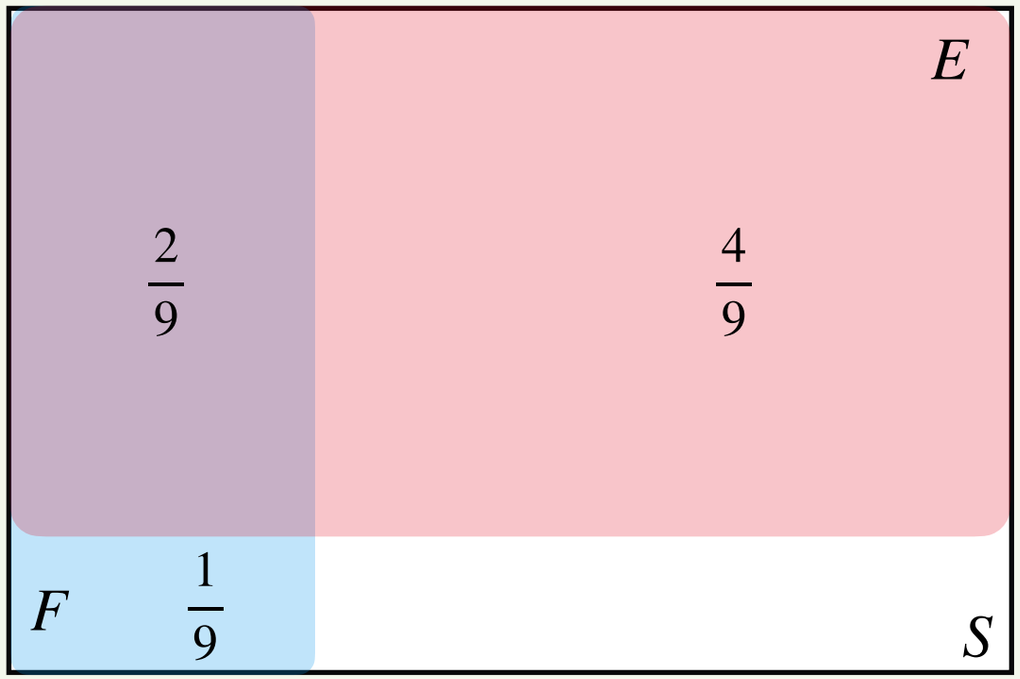


$P(E) = \frac{1}{4}$

$P(F) = \frac{1}{4}$

$P(E)P(F) = \frac{1}{16} \neq P(EF) = \varnothing$

No



$P(E) = \frac{2}{9} + \frac{4}{9} = \frac{2}{3}$

$P(F) = \frac{1}{9} + \frac{2}{9} = \frac{1}{3}$

$P(EF) \overset{?}{=} P(E)P(F)$

$\frac{2}{9} \overset{\checkmark}{=} \frac{1}{3} \cdot \frac{2}{3}$ **Yes**

- Independence is not mutual exclusion!

- Independence is difficult to visualize graphically

De Morgan's law

E or F
$P(E \cup F)$

De Morgan's

E and F
$P(E \cap F)$

Mutually Exclusive?

Just add!

$P(E) + P(F)$

Inclusion-Exclusion principle

$P(E) + P(F) - P(E \cap F)$

Corollary 3
(Lecture 1, s44)

Independent?

Just multiply!

$P(E)P(F)$

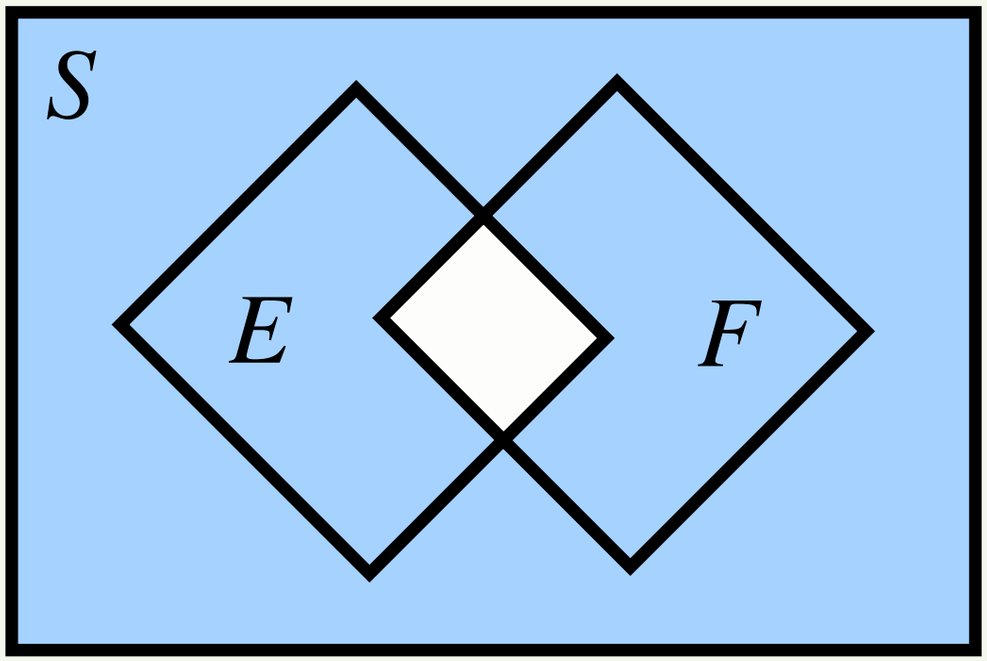
Chain Rule

$P(E)P(F \mid E)$
or
$P(F)P(E \mid F)$

(Lecture 1, s48)

De Morgan's law

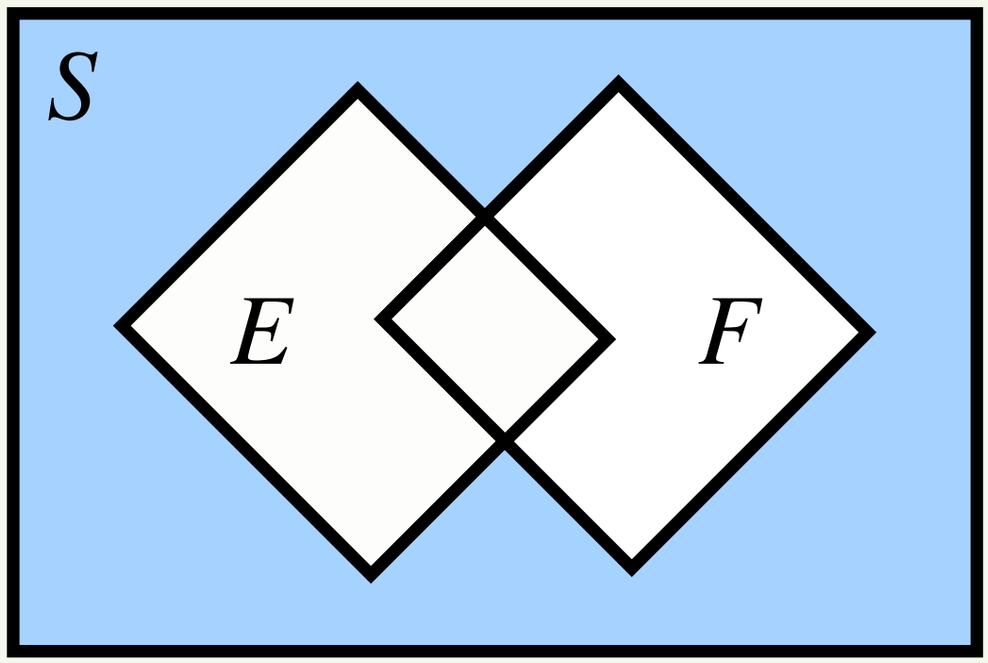
Lets you switch between AND and OR



$$P(E \cap F)^C) = P(E^C \cup F^C)$$

$$\begin{aligned} &P(E_1 E_2 \cdots E_n) \\ &\quad = 1 - P\left((E_1 E_2 \cdots E_n)^C\right) \\ &\quad = 1 - P\left(E_1^C \cup E_2^C \cup \cdots \cup E_n^C\right) \end{aligned}$$

Great if E_i^C mutually exclusive



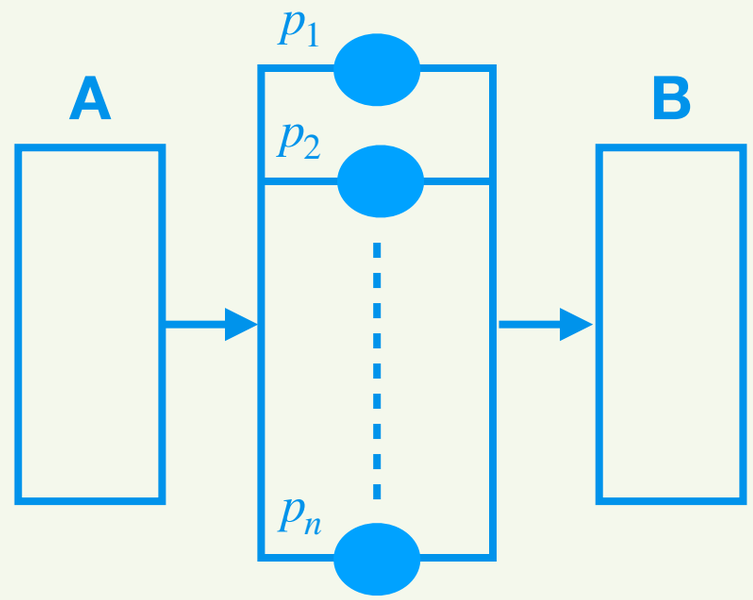
$$P(E \cup F)^C) = P(E^C \cap F^C)$$

$$\begin{aligned} &P(E_1 \cup E_2 \cup \cdots \cup E_n) \\ &\quad = 1 - P\left((E_1 \cup E_2 \cup \cdots \cup E_n)^C\right) \\ &\quad = 1 - P\left(E_1^C E_2^C \cdots E_n^C\right) \end{aligned}$$

Great if E_i^C independent

Example: Network reliability

- Consider a parallel network with:
  - n independent routers, each with probability p_i of functioning, where $(1 \leq i \leq n)$
  - $E =$ functional path from A to B exists
- What is $P(E)$?



Let $F_i =$ event of router i functioning $(i = 1,2,..,n)$, where $P(F_i) = p_i$

$$P(E) = P(F_1 \cup F_2 \cup \ldots \cup F_n)$$ [i.e., $P(\geq 1$ router works)]

Application of De Morgan's Law

$$= 1 - P(F_1^C \cap F_2^C \cap \ldots \cap F_n^C)$$ [i.e., $1 - P($all routers fail$)$]

$$= 1 - (1 - P_1)(1 - P_2) \cdots (1 - P_n) = 1 - \prod_{i=1}^{n} (1 - p_i)$$

Independence!

Conditional Independence

Conditional paradigm

For any events A, B, and E, you can condition consistently on E, and all formulas still hold

Axiom 1 $$0 \leq P(A \mid E) \leq 1$$

Corollary 1 $$P(A \mid E) = 1 - P(A^C \mid E)$$

Transitivity $$P(AB \mid E) = P(BA \mid E)$$

Chain rule $$P(AB \mid E) = P(B \mid E)P(A \mid BE)$$

Bayes' Thm $$P(A \mid BE) = \frac{P(B \mid AE)P(A \mid E)}{P(B \mid E)}$$

Conditional independence

Two events A and B are defined as conditionally independent given E if

$$P(AB \mid E) = P(A \mid E)P(B \mid E)$$

Independence relations can change with conditioning

A and B independent	does NOT always mean	A and B independent given E

Art and condition

Likes:

E_1

E_2

E_3

Will like?

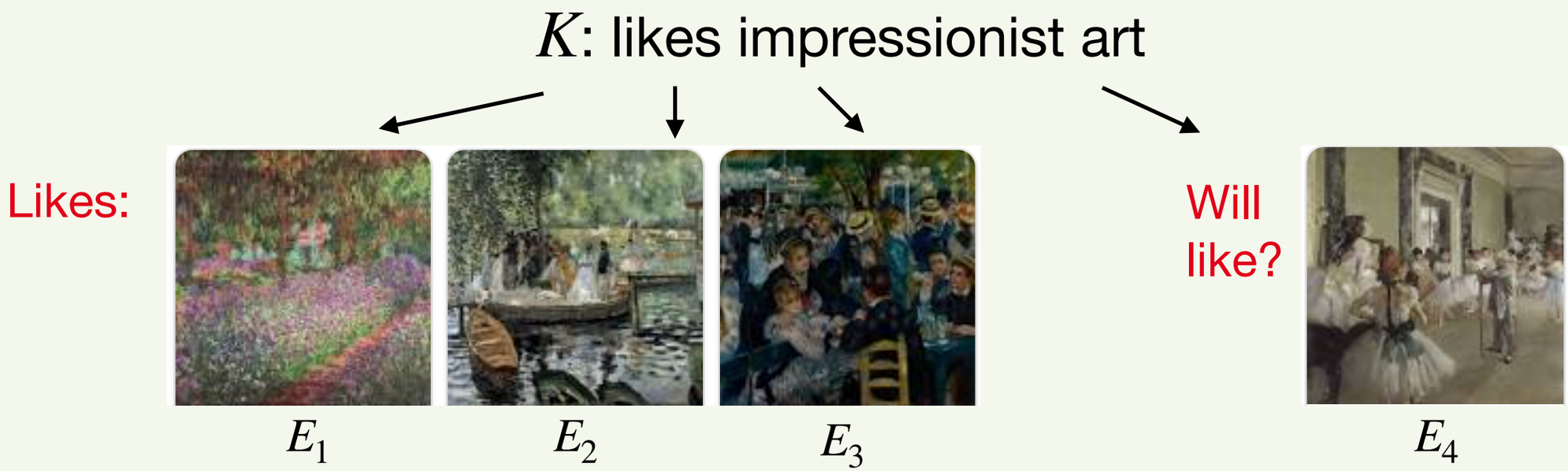
E_4

What if $E_1E_2E_3E_4$ are not independent? (e.g. all impressionist paintings)

$$P(E_4 \mid E_1E_2E_3) = \frac{P(E_1E_2E_3E_4)}{P(E_1E_2E_3)} = \frac{\text{\# of people who like all 4}}{\text{\# of people who like the first 3}}$$

Need to keep track of an exponential # of statistics!

Art and condition



Assume: $E_1E_2E_3E_4$ are *conditionally independent* given K

$$P(E_4 \mid E_1E_2E_3) = \frac{P(E_1E_2E_3E_4)}{P(E_1E_2E_3)}$$

$$P(E_4 \mid E_1E_2E_3K) = \underbrace{P(E_4 \mid K)}$$

An *easier probability* to store and compute

Independence is very important in ML and probabilistic modeling.

Knowing the joint probability of many events requires exponential amounts of data. By making (conditional) **independence** claims, computers can decompose how to calculate the joint probability.

⇒ *Faster to compute and requires less data to learn probabilities.*

“Exploiting conditional independence to generate fast probabilistic computations is one of the main contributions computer science has made to probability theory”

-Judea Pearl 2011 Turing Award,
“For fundamental contributions to AI through the development of a calculus for probabilistic and causal reasoning”

Art and condition

K: likes impressionist art

E_1 E_2 E_3 E_4

$E_1E_2E_3E_4$ are
dependent

$E_1E_2E_3E_4$ are
conditionally independent
given K

Dependent events can become conditionally independent. ⟷ Independent events can become conditionally dependent

Knowing exactly *when conditioning breaks or creates independence* is a big part of building complex probabilistic models

Random Variables

***Note*: Random Variables also called distributions**

Random Variables

A random variable is a real-valued function defined on a sample space

Random variables are NOT events!

An event is a particular *assignment of a random variable*

Example:

3 coins are flipped

Let X = # of heads

X is a random variable

Small x

$X = 2$

event

$P(X = 2)$

probability
(number between 0 & 1)

What would be a *useful function* to define?

The probability of the event that a random variable X takes on the value x

For discrete RVs, this is a *probability mass function*

↳ continuous RVs ↳ *probability density function*

We'll get there in a bit…

Discrete RVs and PMFs

- A random variable X is discrete if it can take on countably many values
  - $X = x$, where $x \in (x_1, x_2, x_3, \dots)$ is the *support* of X
- The probability mass function (PMF) of a discrete random variable is:
  - $P(X = x) = p(x) = p_X(x)$

shorthand notation

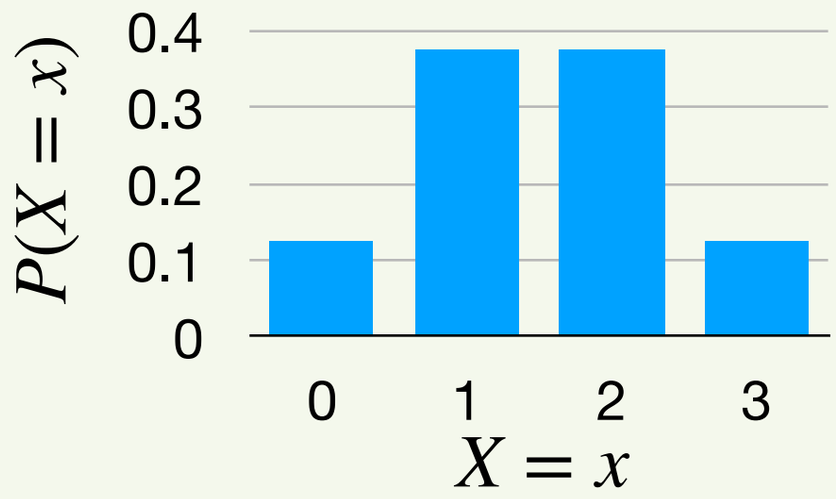
$$\sum_{i=1}^{\infty} p(x_i) = 1$$

Probabilities must sum to 1

Verify any PMF you create

Example:
3 coins are flipped
Let X = # of heads
X is a random variable

$X = x$	$P(X = x)$	Set of outcomes
$X = 0$	$1/8$	{T,T,T}
$X = 1$	$3/8$	{H,T,T}, {T,H,T},{T,T,H}
$X = 2$	$3/8$	{H,H,T}, {H,T,H},{T,H,H}
$X = 3$	$1/8$	{H,H,H}
$X \geq 4$	0	{}



Sidebar: *Notation*

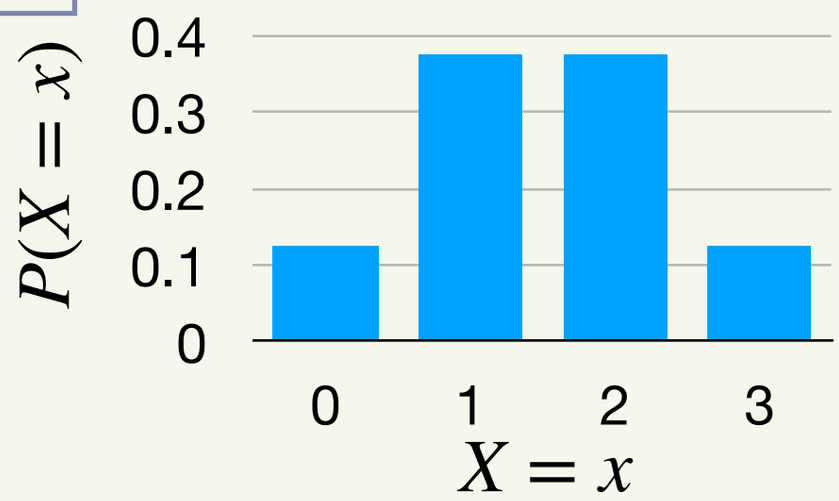
- Careful with the notation in the Cowan textbook
  - He uses x (i.e. small x) for both the random variable and the value it can assume
    - See this footnote on page 7:

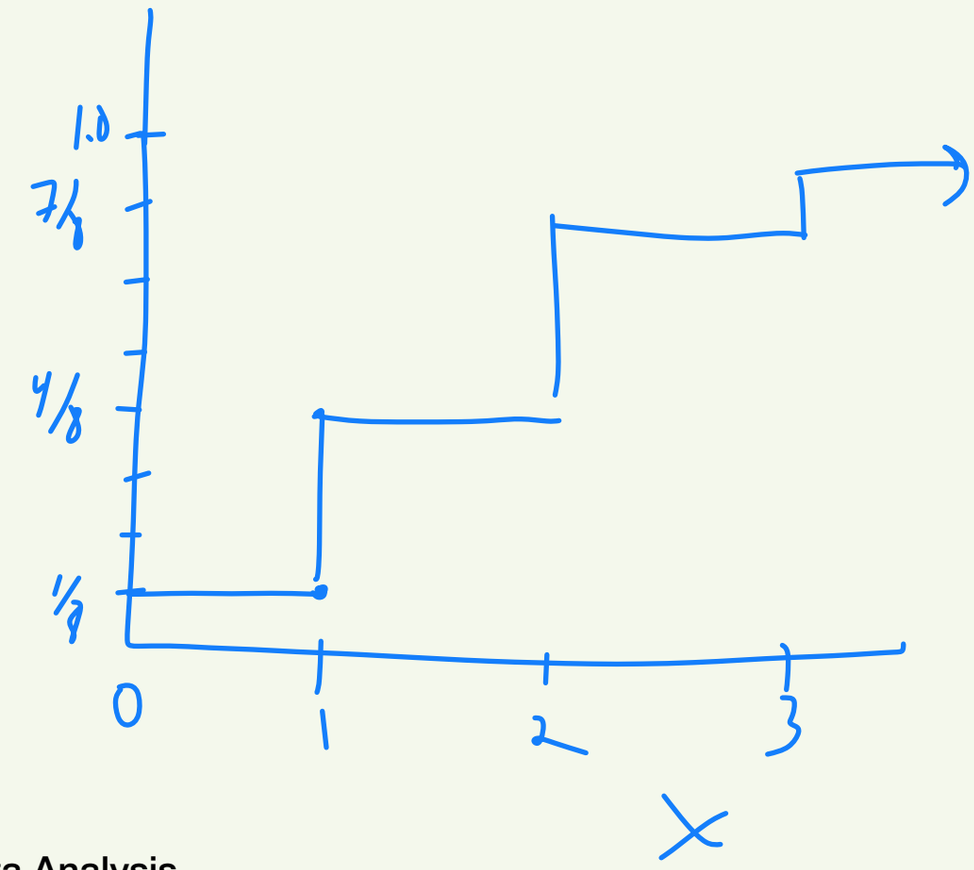
[2] A possible confusion can arise from the notation used here, since x refers both to the random variable and also to a value that can be assumed by the variable. Many authors use upper case for the random variable, and lower case for the value, i.e. one speaks of X taking on a value in the interval $[x, x + dx]$. This notation is avoided here for simplicity; the distinction between variables and their values should be clear from context.

Cumulative distribution function

- For a random variable X, the cumulative distribution function (CDF) is defined as
  - $F(a) = F_X(a) = P(X \leq a)$ where $-\infty < a < \infty$
- For a discrete RV X, the CDF is
  - $F(a) = F_X(a) = \sum_{\text{all } x \leq a} p(x)$

Example:
3 coins are flipped
Let X = # of heads
X is a random variable





Expectation

- The expectation of a discrete random variable X is
  - $E[X] = \sum_{x:p(x)>0} p(x) \cdot x$

Sum over all values of $X = x$ that have non-zero probability

- Important properties of expectation
  - *Linearity:*
    - $E[aX + b] = aE[X] + b$
  - *Expectation of a sum = sum of expectation*
    - $E[X + Y] = E[X] + E[Y]$
  - *Unconscious statistician:*
    - $E[g(X)] = \sum_x g(x)p(x)$

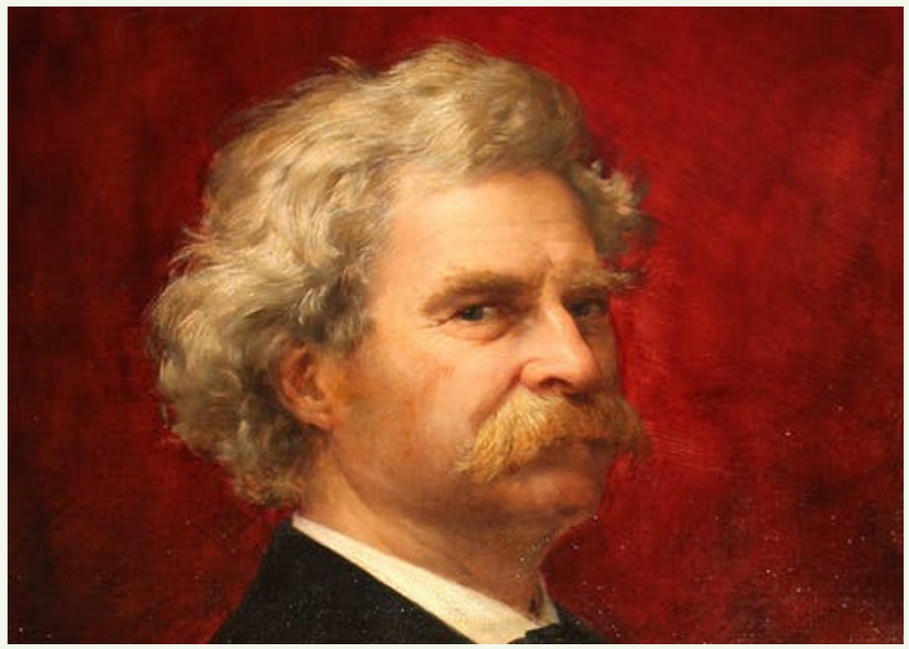
Prove them!

Can look this one up

Known as the law of the unconscious statistician (LOTUS) because of a purported tendency to use the identity without realizing that it must be treated as the result of a rigorously proved theorem, not merely a definition (wikipedia)

Lying with statistics

“There are three kinds of lies:
lies, damned lies, and statistics”



Popularized by Mark Twain (1906)

Generally attributed to Sir Charles Dilke (1891)

Lying with statistics

A school has 3 classes with 5, 10, and 150 students

What is the average class size?

Interpretation #1

- Randomly choose a class with equal probability
- $X =$ size of chosen class

$$E[X] = 5\left(\frac{1}{3}\right) + 10\left(\frac{1}{3}\right) + 150\left(\frac{1}{3}\right)$$

$$= \frac{165}{3} = 55$$

What universities usually report

Interpretation #2

- Randomly choose a student with equal probability
- $Y =$ size of chosen class

$$E[Y] = 5\left(\frac{5}{165}\right) + 10\left(\frac{10}{165}\right) + 150\left(\frac{150}{165}\right)$$

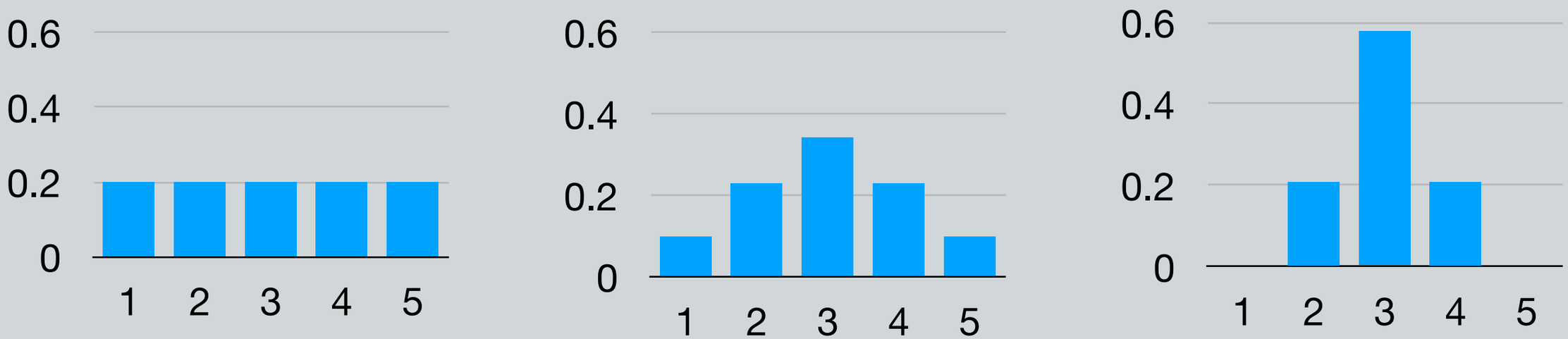
$$= \frac{22635}{165} \approx 137$$

Average student perception of class size

Variance: *A formal quantification of "spread"*

Is $E[X]$ enough?

Consider these 3 distributions



- $E[X] = 3$ for all distributions
- But the spread in the distributions is different

def: $\mathrm{Var}[X] = E[(X - E[X])^2]$ *Units of* X^2

$\mathrm{SD}[X] = \sqrt{\mathrm{Var}[X]}$ *Units of* X

Properties of variance

- Property 1: $\mathrm{Var}[X] = E[X^2] - (E[X])^2$ → *Often easier to compute than the definition*

- Property 2: $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$ → *Unlike expectation, variance is not linear*

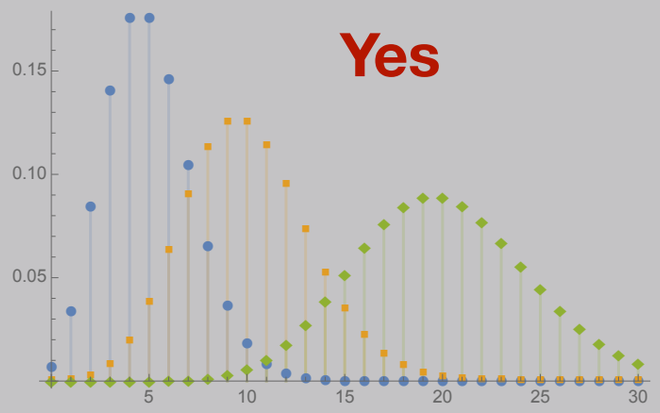
Prove them!

A new world with RVs

- **Event-driven probability**
  - Relate only binary events
    - Either happens (E)
    - or doesn't happen (E^C)
  - Can only report probability
  - Lots of combinatorics
- **Random variables**
  - Link multiple similar events together ($X = 1$, $X = 2$, $\ldots$)
  - Can compute statistics: report the "average" outcome
  - Once we have the PMF (PDF), can do regular math

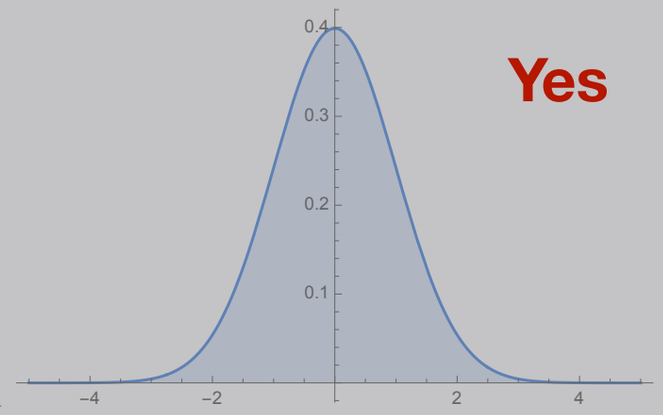
Answer Time: Quiz 1

1. Which of the following functions are good PDFs?



Yes

(a) $\frac{e^{-\lambda}\lambda^k}{k!}$

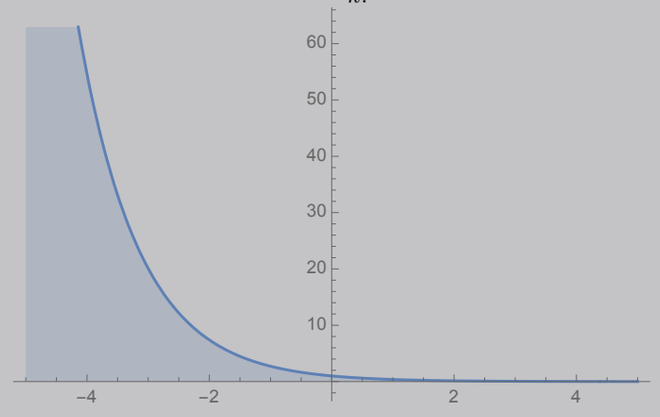


Yes

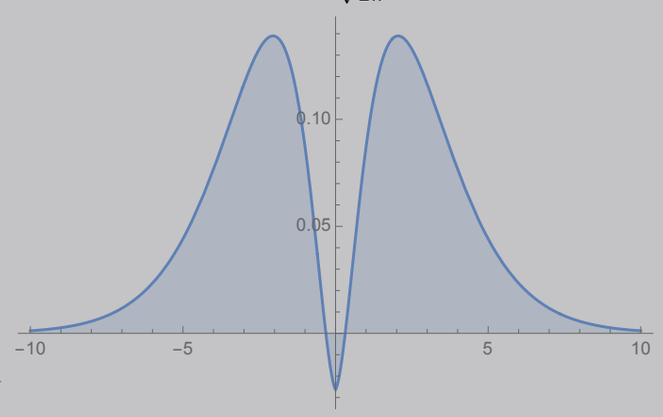
(b) $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$

No

integral over sample space not finite



(c) e^{-x}



(d) $(x^2 - 0.1) \times \exp(-|x|)$ with $x \in \mathbb{R}$

No

Negative density

Yes

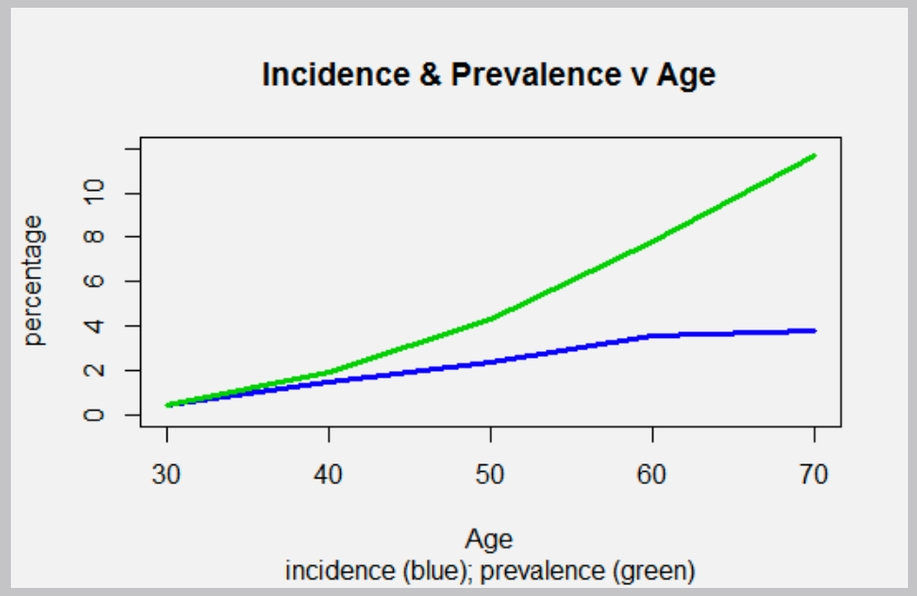
since sample space: [0,10]

(e) $\Theta(5-x)\left(\frac{1}{2} + \frac{\arctan(x-1)}{\pi}\right) + x\,\Theta(x-5)$ with Θ denoting the step function.

2. Are mammographies useful?

At age 30:

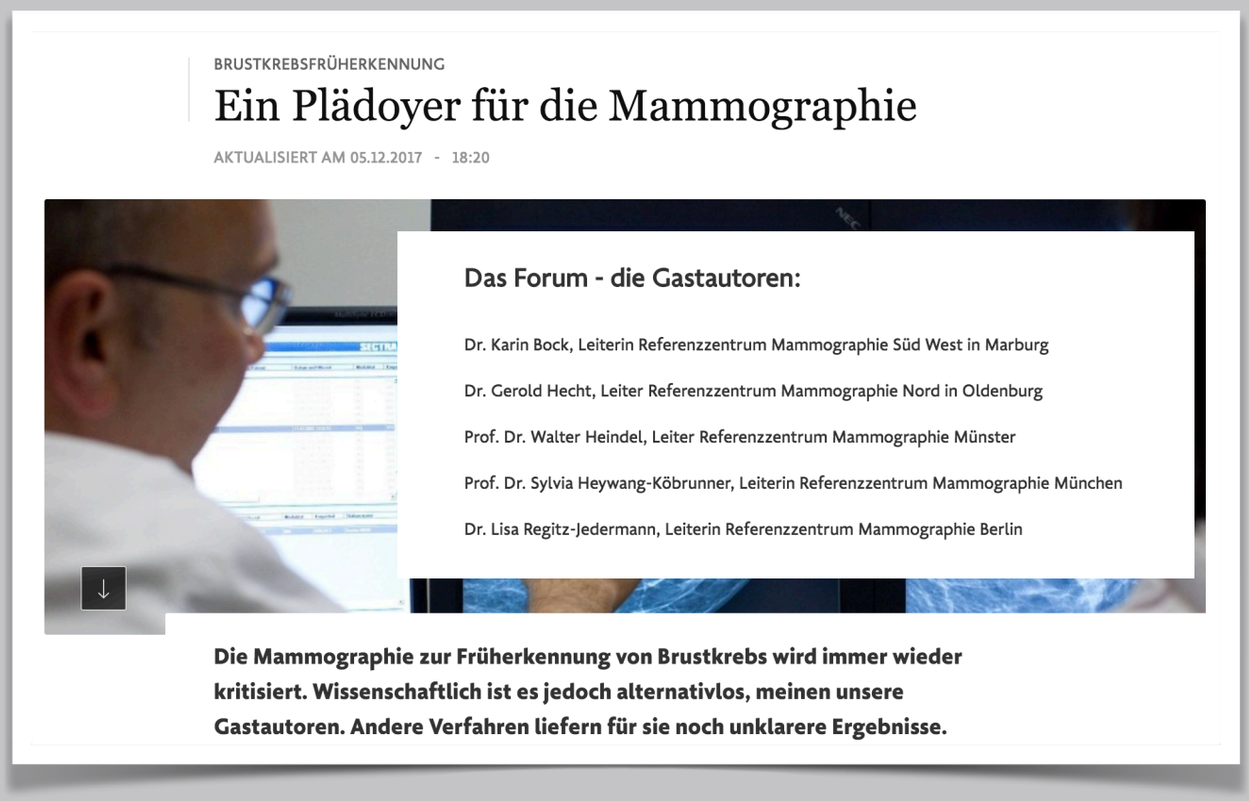
$$p(C|+) = \frac{p(+|C)}{p(+|C) * p(C) + p(+|\bar{C}) * p(\bar{C})} * p(C) = \frac{0.8}{0.8*0.015 + 0.07*0.985} * 0.015 = 0.148.$$



At age 70:

$$p(C|+) = \frac{0.8}{0.8*0.04 + 0.07*0.96} * 0.04 = 0.32$$

But careful: Reality is more complicated!



BRUSTKREBSFRÜHERKENNUNG

Ein Plädoyer für die Mammographie

AKTUALISIERT AM 05.12.2017 - 18:20

Das Forum - die Gastautoren:

Dr. Karin Bock, Leiterin Referenzzentrum Mammographie Süd West in Marburg

Dr. Gerold Hecht, Leiter Referenzzentrum Mammographie Nord in Oldenburg

Prof. Dr. Walter Heindel, Leiter Referenzzentrum Mammographie Münster

Prof. Dr. Sylvia Heywang-Köbrunner, Leiterin Referenzzentrum Mammographie München

Dr. Lisa Regitz-Jedermann, Leiterin Referenzzentrum Mammographie Berlin

Die Mammographie zur Früherkennung von Brustkrebs wird immer wieder kritisiert. Wissenschaftlich ist es jedoch alternativlos, meinen unsere Gastautoren. Andere Verfahren liefern für sie noch unklarere Ergebnisse.

http://www.faz.net/aktuell/wissen/medizin-ernaehrung/keine-alternativen-zur-mammographie-15313597.html
https://stats.stackexchange.com/questions/185817/interpretation-of-bayes-theorem-applied-to-positive-mammography-results

Take 5

Ex. Discrete RVs

Bernoulli RV → *success / failure*

Consider an experiment with two outcomes: *"success" and "failure"*

def: A Bernoulli random variable X maps "success" to 1 and "failure" to 0

The probability for success is some constant value p

4) with parameter

1) The RV → $X \sim \text{Ber}(p)$

2) is distributed as a

Support: $\{0,1\}$

3) Bernoulli

PMF $\quad P(X=1) = p(1) = p$

$\quad P(X=0) = p(0) = 1-p$

Expectation $\quad E[X] = p$

Variance $\quad \text{Var}[X] = p(1-p)$

S. Ross textbook notation

Bernoulli RV → *success / failure*

- **Run a program**
  - Crashes w.p. p
  - Works w.p. $(1-p)$
- Let X: 1 if crash
  - $X \sim \text{Ber}(p)$
    - $P(X=1) = p$
    - $P(X=0) = 1-p$

- **Serve an ad**
  - User clicks w.p. 0.2
  - Ignores otherwise
- Let X: 1 if clicked
  - $X \sim \text{Ber}(0.2)$
    - $P(X=1) = 0.2$
    - $P(X=0) = 0.8$

- **Roll two dice**
  - Success: roll two 6's
  - Failure: anything else
- Let X: 1 if success
  - $X \sim \text{Ber}(\frac{1}{36})$
    - $E[X] = \frac{1}{36}$

Binomial RV → *# of successes in N trials*

Consider an experiment: N independent trials of $\mathrm{Ber}(p)$ random variables (1,0)

def: A Binomial random variable n is the number of successes in N trials

The probability for success is some constant value p

Switch to Cowan notation

$$n \sim \mathrm{Bin}(N, p)$$

# of trials

$P(\mathit{success})$

Support: $\{0,1,...,N\}$

PMF	$f(n;N,p) = \binom{N}{n} p^n (1-p)^{N-n}$
Expectation	$E[n] = Np$
Variance	$\mathrm{Var}[n] = Np(1-p)$

In other words

- Consider N independent trials / observations, each having **two** possible outcomes
  - Call them **“success”** and **“failure”**
  - The **probability** for success is some constant value p
- Set of trials can be summarized by single discrete random variable n, defined as ***# of successes***.
  - Sample space: set of possible values of n **successes given** N ***observations***
- If one were to repeat the entire experiment many times with N ***trials each***, the resulting value of n would occur with **relative frequencies** given by the **binomial distribution**

Derivation of the binomial distribution

- Let's derive the form of the binomial distribution:
  - Probability of success is p
  - Probability of failure $1 - p$
- Since each trial is assumed to be independent, the probability for a series of successes and failures in a particular order equals the product of individual probabilities:
  - E.g. probability of **five** trials to have
    - of **success**, **success**, **failure**, **success**, **failure**

$$= p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3 (1-p)^2$$

  - Generalized: probability for n successes and $N - n$ failures is

$$= p^n (1-p)^{N-n}$$

continued…

Generalized: *probability for n successes and $N - n$ failures* $p^n(1-p)^{N-n}$

- We are not interested in the order of these processes to happen, *only in the final number of successes*
  - The number of sequences having n successes in N events is

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

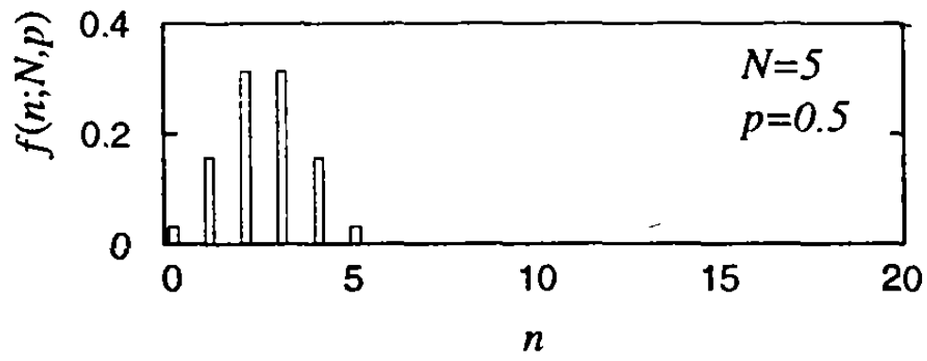
- So the total probability to have n successes in N events is

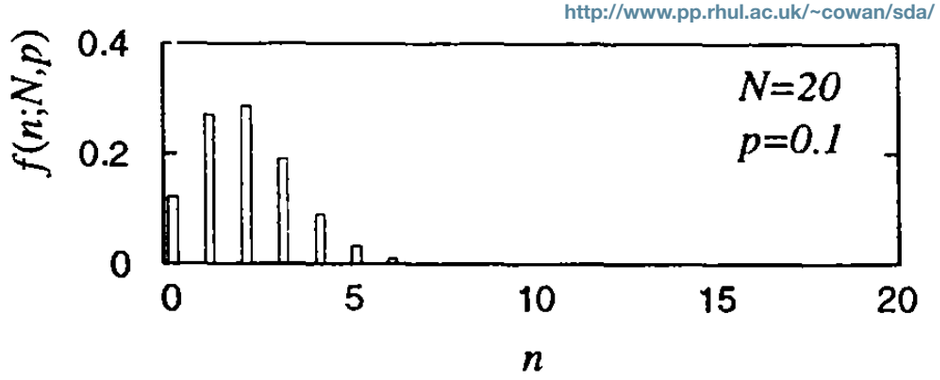
$$f(n; N, p) = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

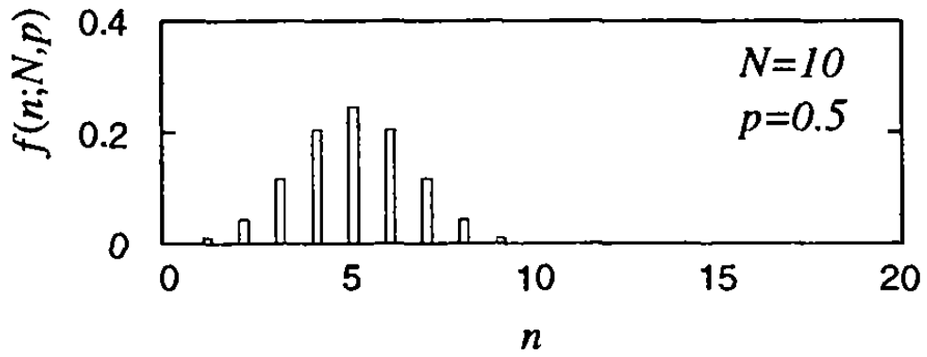
  - with $n = 0, \ldots, N$

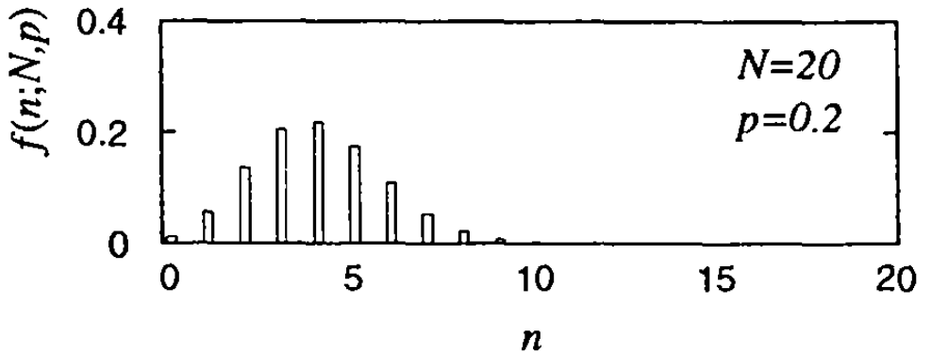
Examples

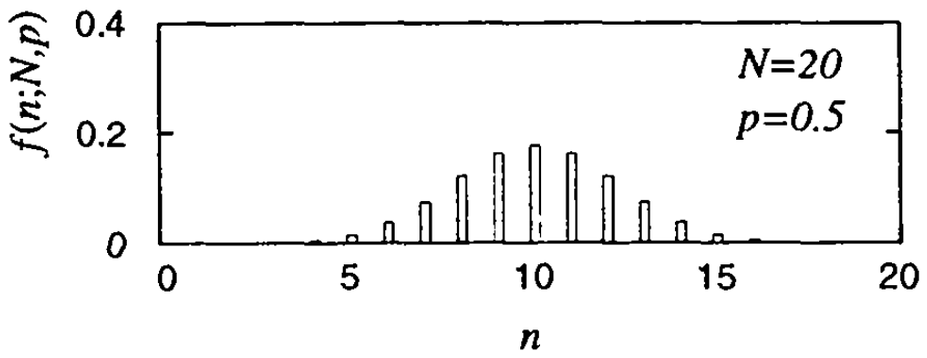
http://www.pp.rhul.ac.uk/~cowan/sda/

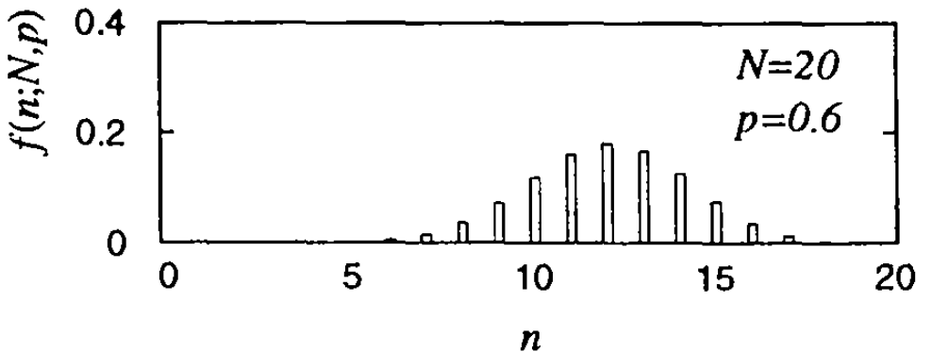












Recall: expectation value $E[n]$ and variance $V[n]$are not functions of the random variable n, but they depend on the parameters of the probability function.

$$E[n] = \sum_{n=0}^{\infty} n \frac{N!}{n!\,(N-n)!}\, p^n\,(1-p)^{N-n} = Np \qquad V[n] = E[n^2] - (E[n])^2 = Np\,(1-p)$$

Multinomial distribution I

- Generalization of binomial distribution to the case where there are **more than two outcomes** (“success” and “failure”)
  - Rather m different possible outcomes are allowed with probabilities p_i

$$\sum_{i=1}^{m} p_i = 1$$

- Now consider a measurement consisting of N independent trials, each which yields one of the possible m outcomes.
- The probability for a particular sequence of outcomes, e.g. i on **first** trial and j on **second** etc. in a particular order is the produce of the N corresponding probabilities,

$$p_i p_j \dots p_k$$

- The number of such sequences that will lead to n_1 outcomes of type 1, n_2 outcomes of type 2, etc. is

$$\frac{N!}{n_1! n_2! \dots n_m!}$$

Multinomial distribution II

- If we are again not interested in the order of the outcomes, but only in the total number of each type, the joint probability for n_1 outcomes of type 1, n_2 outcomes of type 2, etc. is given by the **multinomial distribution**:

$$f(n_1, n_2, \ldots, n_m; N, p_1, p_2, \ldots, p_m) = \frac{N!}{n_1! n_2! \ldots n_m!} p_1^{n_1} p_2^{n_2} \ldots p_m^{n_m}$$

- E.g. for **three** possible outcomes i, j and *everything else*:

$$f(n_i, n_j; N, p_1, p_2) = \frac{N!}{n_i! n_j! (N - n_i - n_j)!} p_i^{n_i} p_j^{n_j} (1 - p_i - p_j)^{N - n_i - n_j}$$

- The covariance $V_{ij} = \mathrm{cov}[n_i, n_j]$ is

$$\begin{aligned} V_{ij} &= E[(n_i - E[n_i])\, E[(n_j - E[n_j]) \\ &= -N p_i p_j \qquad \text{for } i \neq j \end{aligned}$$

otherwise $\quad V_{ii} = \sigma_i^2 = N p_i (1 - p_i)$

Pause for a moment

And recall our dear friend e

$$\lim_{N\to\infty}\left(1-\frac{\nu}{N}\right)^{N}=e^{-\nu}$$

Binomial to the extreme

$$\lim_{N\to\infty}\left(1-\frac{\nu}{N}\right)^{N}=e^{-\nu}$$

$$f(n;N,p)=\frac{N!}{n!(N-n)!}p^n(1-p)^{N-n} \qquad E[n]=Np=\nu \qquad \Rightarrow p=\frac{\nu}{N}$$

$$=\frac{N!}{n!(N-n)!}\left(\frac{\nu}{N}\right)^n\left(1-\frac{\nu}{N}\right)^{N-n}$$

$$=\frac{\overbrace{N(N-1)(N-2)\cdots(N-n+1)}^{n\text{ terms}}}{n!N^n}\nu^n\left(1-\frac{\nu}{N}\right)^{N-n}$$

$$\lim_{N\to\infty}\frac{\overbrace{\left(1-\frac{1}{N}\right)\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{n-1}{N}\right)}^{=1}}{n!}\nu^n\underbrace{\left(1-\frac{\nu}{N}\right)^{N}}_{=e^{-\nu}}\overbrace{\left(1-\frac{\nu}{N}\right)^{-n}}^{=1}$$

ν fixed (ie, $p\to 0$)

$$=\frac{\nu^n}{n!}e^{-\nu}$$

In other words

- The binomial distribution, in the limit that N **becomes very large**, p **becomes very small**, but the product Np **remains** equal to some **finite** value ν, becomes

$$f(n;\nu) = \frac{\nu^n}{n!}e^{-\nu}$$

- This distribution is called the **Poisson distribution** for the integer random variable n, where $n = 0, 1, \ldots, \infty$
  - The PDF has one parameter ν
  - Expectation value and variance:

$$E[n] = \sum_{n=0}^{\infty} n\frac{\nu^n}{n!}e^{-\nu} = \nu$$

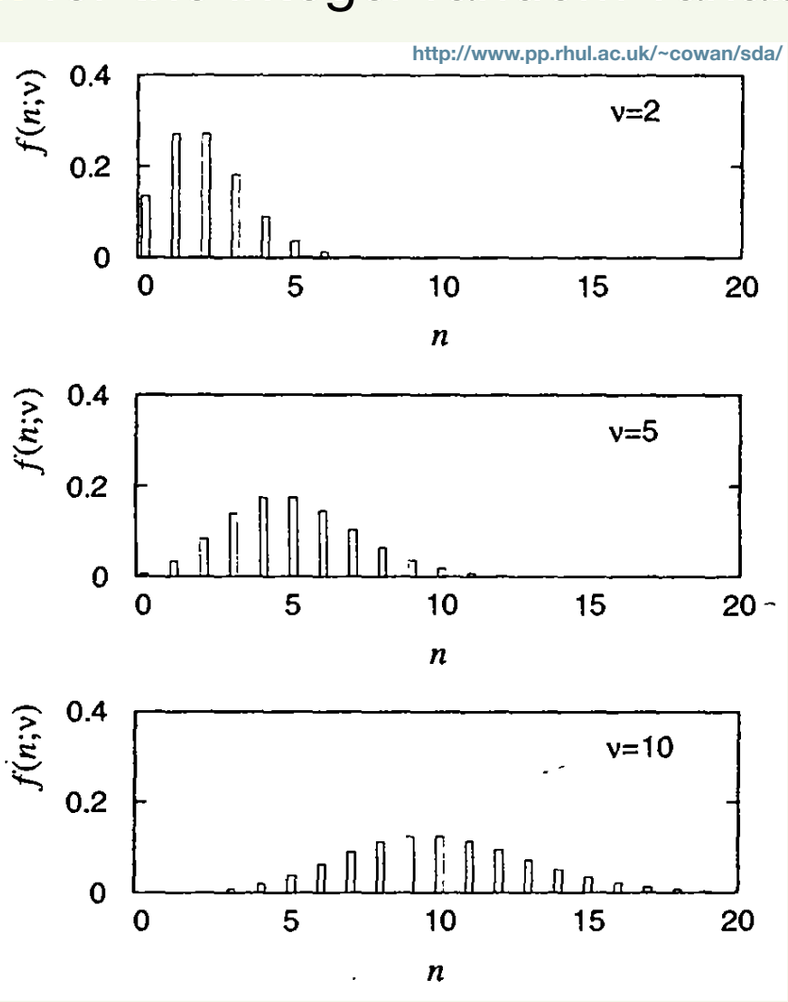
$$V[n] = \sum_{n=0}^{\infty} (n-\nu)^2\frac{\nu^n}{n!}e^{-\nu} = \nu$$

Proof:

$$E[n] = Np = \nu$$ *Recall the binomial!*

$$V[n] = Np(1-p) = \nu(1-0) = \nu$$

http://www.pp.rhul.ac.uk/~cowan/sda/



Examples

Poisson RV → *# of successes in an interval of time at a fixed rate*

Consider an experiment that lasts a fixed interval of time

def: A Poisson random variable n is the # of successes over the experiment duration, assuming the time each success occurs is independent and the average # of successes over time is constant

- Examples:
  - number of decays of radioactive material in a fixed time period in the limit that the total number of decays is large
  - number of events of a certain type observed in a particle scattering experiment with a given integrated luminosity L. The **expectation value of the number of events** is

$$\nu = \sigma L \epsilon$$

Efficiency to observe an event

Cross section

Luminosity

Other discrete RVs *(before we move to continuous RVs…)*

Consider an experiment: independent trials of $\text{Ber}(p)$ random variables

def: A Geometric random variable X is the # of trials until the first success

$$X \sim \text{Geo}(p)$$

Support: $\{1,2,...\}$

PMF	$P(X = x) = (1-p)^{(x-1)}p$
Expectation	$E[X] = 1/p$
Variance	$\text{Var}[X] = (1-p)/p^2$

def: A Negative Binomial random variable X is the # of trials until r successes

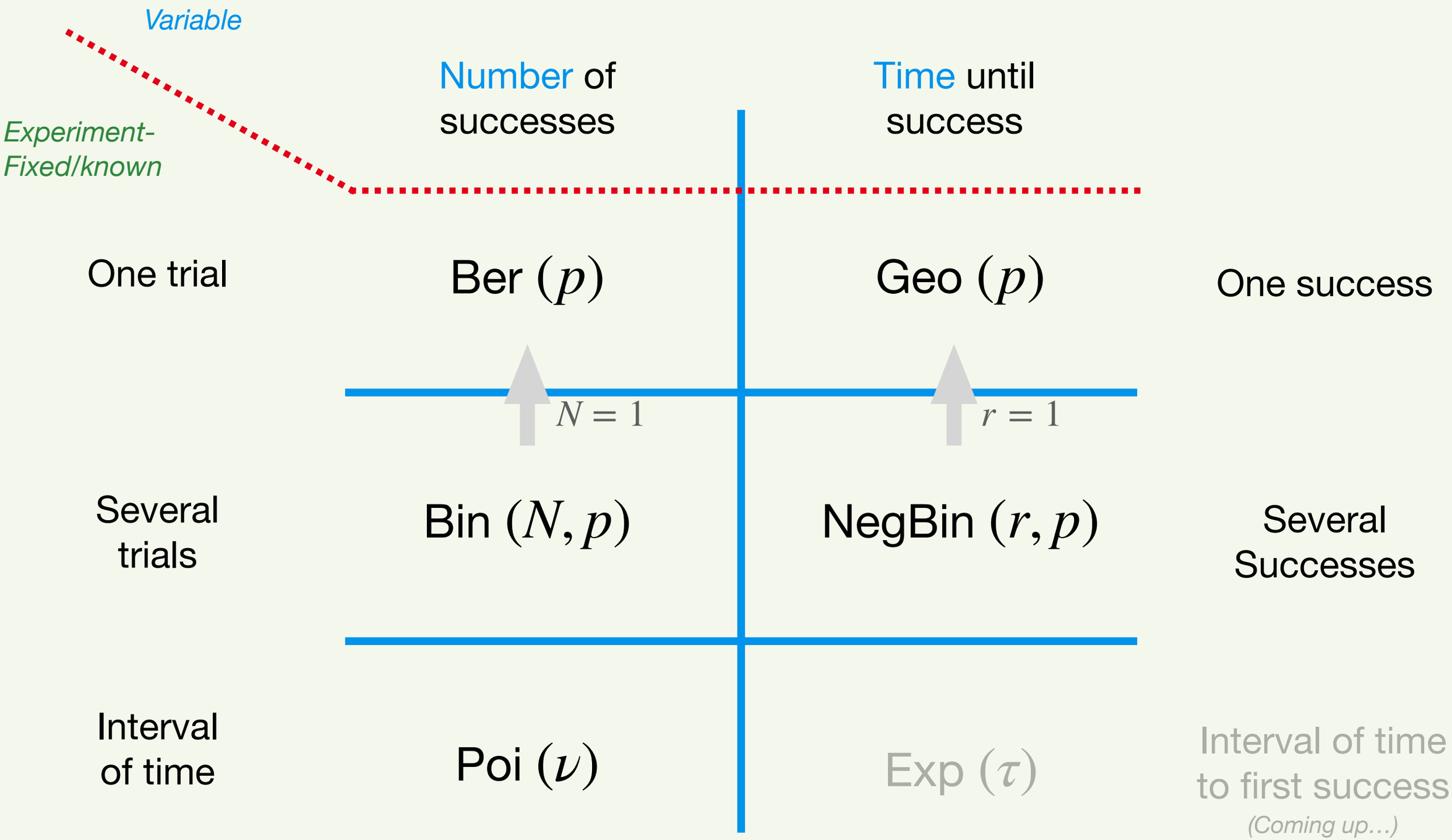
$$X \sim \text{NegBin}(p)$$

Support: $\{r, r+1,...\}$

PMF	$P(X = x) = \binom{k-1}{x-1}(1-p)^{x-r}p^r$
Expectation	$E[X] = r/p$
Variance	$\text{Var}[X] = r(1-p)/p^2$

RV grid (so far)

Translate a problem statement into a RV.
I.e., model real life situations with probability distributions.



Continuous RVs

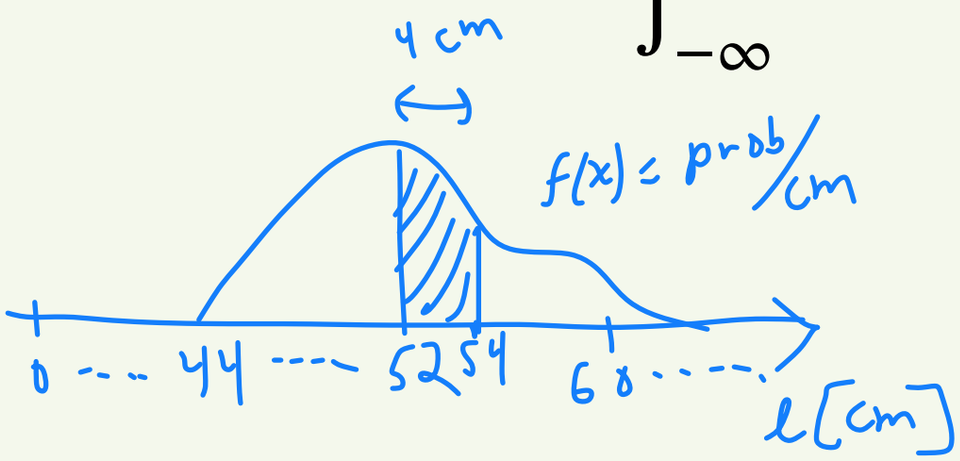
Continuous RV definition

- A random variable X is continuous if there is a probability density function $f(x) \geq 0$ such that for $-\infty < x < \infty$:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- Integrating a PDF must always yield valid probabilities, and therefore the PDF must also satisfy

$$\int_{-\infty}^{\infty} f(x)dx = P(-\infty < X < \infty) = 1$$



$$P(52 \leq X \leq 54) = \int_{52}^{54} f(x)dx$$

Integrate $f(x)$ to get probabilities

PDF Units: probability per units of X

PMF vs PDF

- Discrete random variable X
  - Probability mass function (PMF)
    - To get probability:

$$P(a \leq X \leq b) = \sum_{x=a}^{b} p(x)$$

$$E[X] = \sum_{x} xp(x)$$

$$E[g(X)] = \sum_{x} g(x)p(x)$$

- Continuous random variable X
  - Probability density function (PDF)
    - To get probability:

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

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Both discrete and continuous:

$$E[aX + b] = aE[X] + b$$

Linearity of expectation

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

$$\text{Var}[aX + b] = a^2\text{Var}[X]$$

Properties of variance

PMF vs PDF: Cumulative Distribution Function

- Discrete random variable X
  - CDF

$$F(a) = P(X \leq a) = \sum_{\text{all } x \leq a}^{b} p(x)$$

Recall 3 coin toss example from s30

- Continuous random variable X
  - CDF

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

http://www.pp.rhul.ac.uk/~cowan/sda/

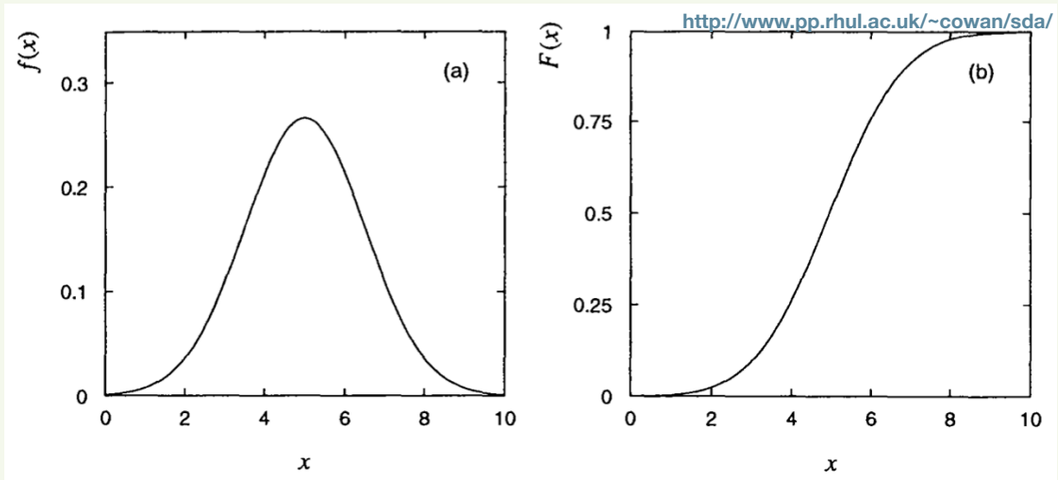


Fig. 1.3 (a) A probability density function $f(x)$. (b) The corresponding cumulative distribution function $F(x)$.

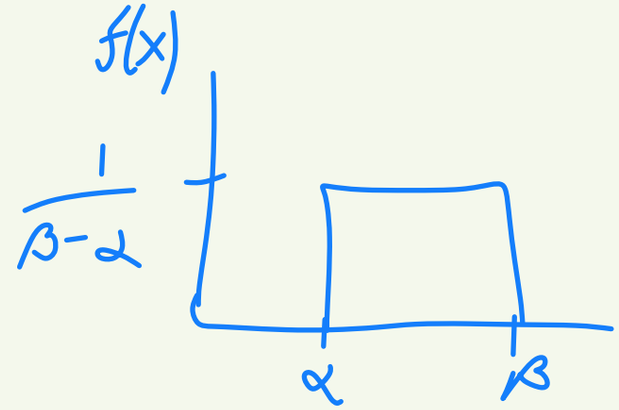
Important points:

- CDF is a probability, though PDF is not
- If you learn to use CDFs, *you can avoid integrating the PDF*

Uniform RV

- The **uniform** PDF for a continuous RV X with support $x \in (-\infty, \infty)$ is defined by

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta-\alpha} & \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$



  - i.e. x is **equally likely** to be found anywhere between α and β
  - Mean and variance:

$$E[X] = \int_\alpha^\beta \frac{x}{\beta-\alpha} dx = \frac{1}{2}(\alpha+\beta)$$

average of start + end

$$V[X] = \int_\alpha^\beta [x - \frac{1}{2}(\alpha+\beta)]^2 \frac{1}{\beta-\alpha} dx = \frac{1}{12}(\beta-\alpha)^2$$

An important feature of the uniform distribution is that any continuous random variable X with PDF $f(x)$ and CDF $F(x)$ can easily be transformed to a new variable Y which is uniformly distributed

Uniform RV: *transformed variable*

- The transformed variable y is simply given by

$$y = F(x)$$ *The CDF of the variable x*

- For any CDF $y = F(x)$, one has

$$\frac{dy}{dx} = \frac{d}{dx}\int_{-\infty}^{x} f(x')dx' = f(x)$$

- Hence one finds for the PDF of y

$$g(y) = f(x)\left|\frac{dx}{dy}\right| = f(x)\left|\frac{dy}{dx}\right|^{-1} = 1 \qquad (0 \le y \le 1)$$

Review: section 1.4 Cowan "functions of RVs"

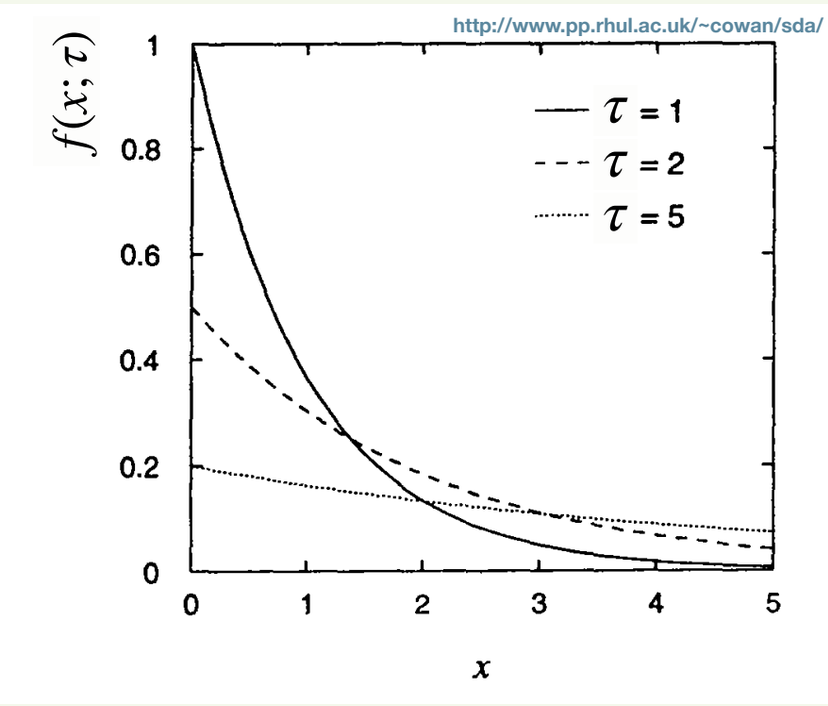
This property of the uniform distribution will be used in the next lecture when we talk about Monte Carlo Methods

Exponential RV: *amount of time* *until first success*

Consider an experiment that lasts a duration of time until success occurs

def: An Exponential random variable X is the amount of time until first success, with probability density defined by:

$$f(x;\tau) = \frac{1}{\tau} e^{-t/\tau} \qquad x \in [0,\infty)$$



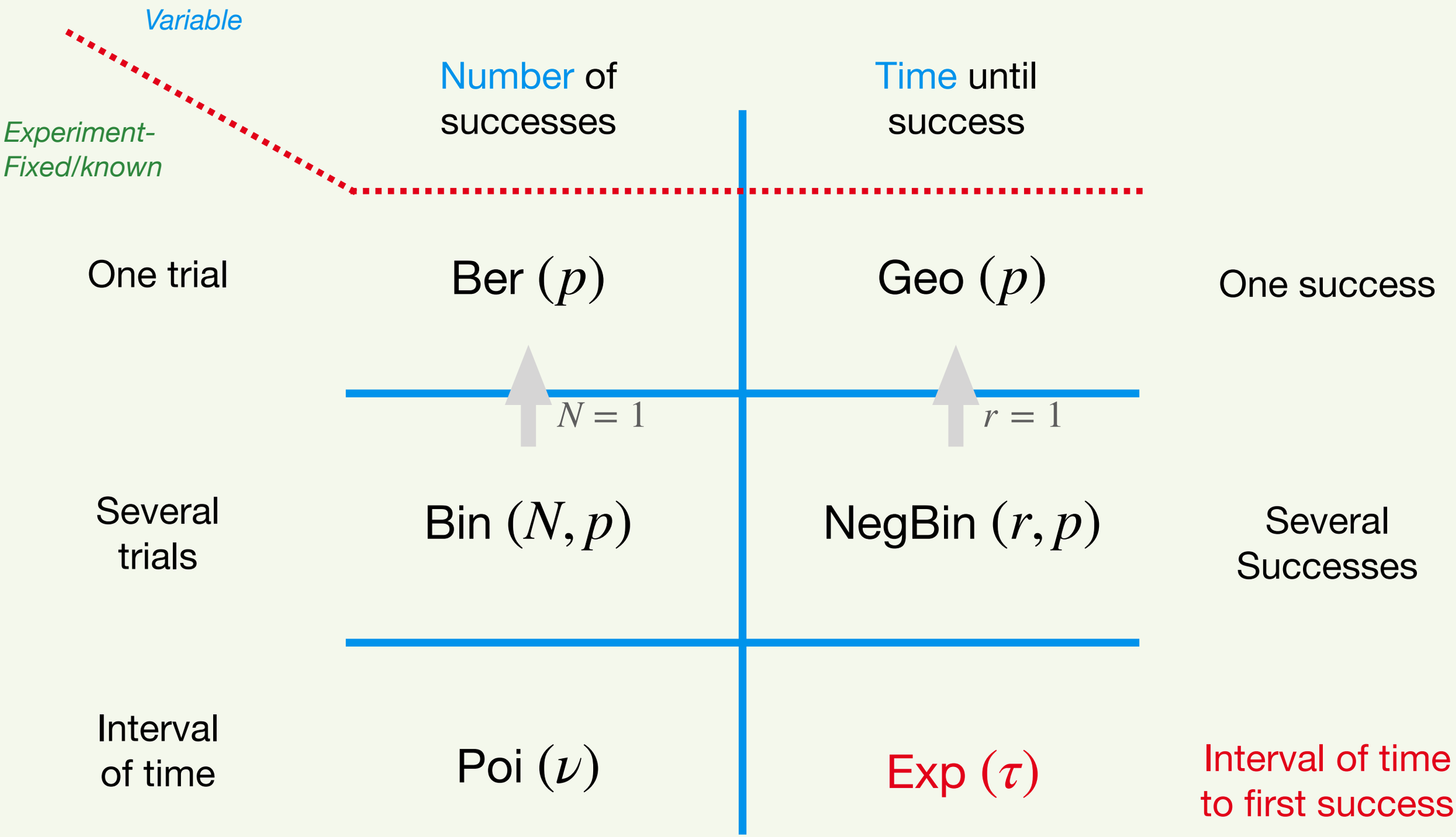
- Characterized by a single parameter τ
- Expectation value and variance:

$$E[x] = \frac{1}{\tau}\int_0^\infty x e^{-x/\tau} dx = \tau \qquad V[x] = \frac{1}{\tau}\int_0^\infty (x-\tau)^2 e^{-x/\tau} dx = \tau^2$$

- **Example**: decay time of unstable particle measured in its rest frame
  - The parameter τ then corresponds to the mean lifetime.

(Hence our choice for using "τ")

RV grid (last piece)



Quiz Time: 2nd Round

1. Error propagation: You have 2 random variables x_1, x_2 with a known covariance matrix

$$C = \begin{pmatrix} 0.5. & 0.2 \\ 0.2 & 0.7 \end{pmatrix},$$

and expectation values $\mu_1 = 6$ and $\mu_2 = 1$. You know want to determine the covariance of two functions f_1 and f_2 of x_1 and x_2 defined as

$$f_1(x_1, x_2) = x_1 + x_2\,,$$
$$f_2(x_1, x_2) = \sqrt{x_1^2 + x_2^2}\,.$$

Calculate the Jacobian $A_{ij} = \left[\frac{\partial f_i}{\partial x_j}\right]_{x_1=\mu_1, x_2=\mu_2}$ and the covariance matrix D between f_1 and f_2. What is the **correlation** between f_1 and f_2? *Hint:* Use $D = ACA^T$.

2. Show that the expectation value and the variance of a Poisson random variable is given by the Poisson parameter ν, i.e. that

$$E[n] = \sum_{n=0}^{\infty} n \frac{\nu^n}{n!} e^{-\nu} \doteq \nu\,,$$
$$V[n] = \sum_{n=0}^{\infty} (n-\nu)^2 \frac{\nu^n}{n!} e^{-\nu} \doteq \nu\,.$$

Hint: Use $V[n] = E[n^2] - (E[n])^2$ and $E[n^2] = E[n\,(n-1) + n] = E[n\,(n-1)] + E[n]$ and the properties of the exponential series ($e^{-x} = \sum_{n=0}^{\infty} \frac{x^n}{n!}$).

For next time

- Required reading
  - Cowan textbook: chapter 2
  - Ross textbook: pages 21-44

Next time

- Complete our tour of important PDFs
- Central Limit Theorem
- i.i.d.

- Monte Carlo
  - *What is it?*
  - *How do you produce it?*
  - *How do you have fun with it?*

Bibliography

- Part of the material presented in this lecture is adapted from the following sources. See the active links (when available) for a complete reference
  - **Probability for CS** (Stanford): slides 13-28, 30-36, 42-44, 55-56
  - **Statistical Data Analysis** textbook by G. Cowan (U. London): slides 45-50, 53-54, 61-63