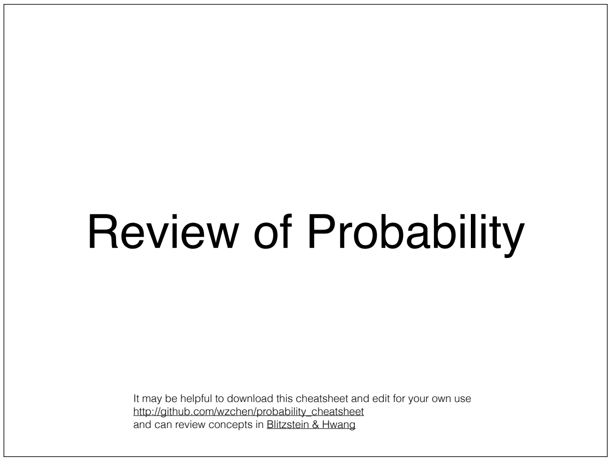
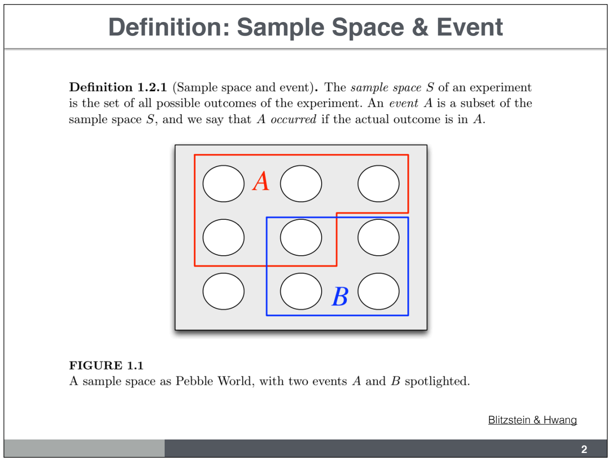


Review of Probability

It may be helpful to download this cheatsheet and edit for your own use
http://github.com/wzchen/probability_cheatsheet
and can review concepts in Blitzstein & Hwang

Today, we will review some concepts in probability, establish a common terminology, and go over basic examples in statistical inference using maximum likelihood estimations and Bayesian inference.



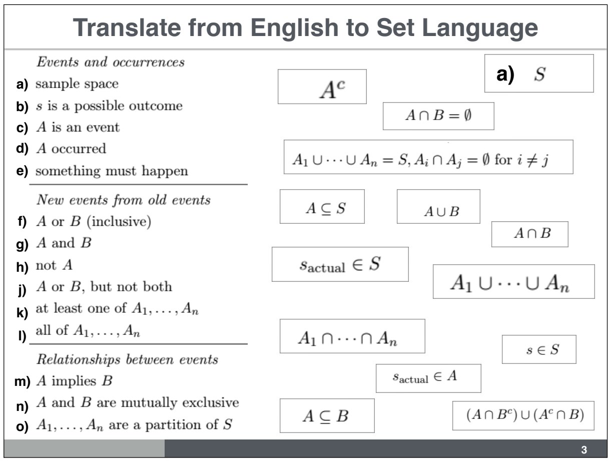
Definition: Sample Space & Event

Definition 1.2.1 (Sample space and event)**.** The *sample space* S of an experiment is the set of all possible outcomes of the experiment. An *event* A is a subset of the sample space S, and we say that A *occurred* if the actual outcome is in A.

FIGURE 1.1
A sample space as Pebble World, with two events A and B spotlighted.

Blitzstein & Hwang

A sample space is the set of all possible outcomes of the experiment.
An event A is a subset of the sample space.
We will say that A occurred if the actual outcome is in A.
Graphically, a sample space is represented in Figure 1.1. as pebbles. Events A and B are subsets of the sample space.



Translate from English to Set Language

Events and occurrences

a) sample space
b) s is a possible outcome
c) A is an event
d) A occurred
e) something must happen

New events from old events

f) A or B (inclusive)
g) A and B
h) not A
j) A or B, but not both
k) at least one of $A_1, \dots, A_n$
l) all of $A_1, \dots, A_n$

Relationships between events

m) A implies B
n) A and B are mutually exclusive
o) $A_1, \dots, A_n$ are a partition of S

a) S

A^c

$A \cap B = \emptyset$

$A_1 \cup \dots \cup A_n = S, A_i \cap A_j = \emptyset$ for $i \neq j$

$A \subseteq S$

$A \cup B$

$A \cap B$

$s_{\text{actual}} \in S$

$A_1 \cup \dots \cup A_n$

$A_1 \cap \dots \cap A_n$

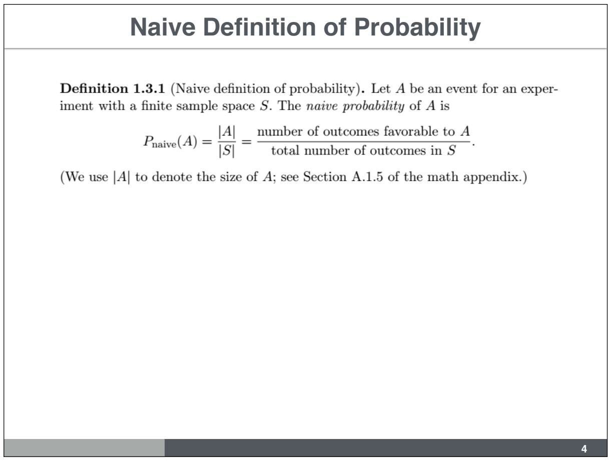
$s \in S$

$s_{\text{actual}} \in A$

$A \subseteq B$

$(A \cap B^c) \cup (A^c \cap B)$

Set theory notation is very useful for manipulating probabilistic statements. To establish a common terminology, let's match the english expressions on the left to the set notation on the right.



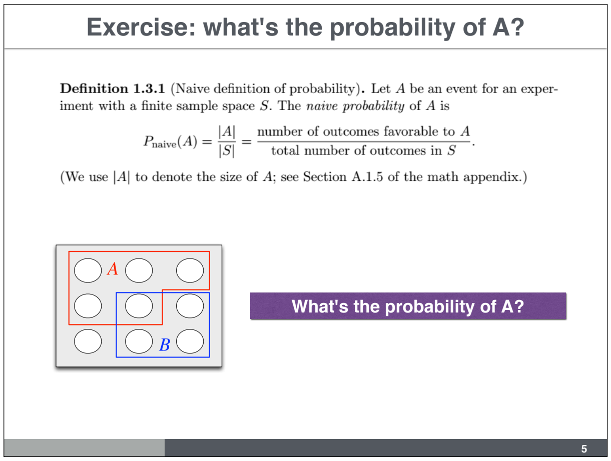
Naive Definition of Probability

Definition 1.3.1 (Naive definition of probability)**.** Let A be an event for an experiment with a finite sample space S. The *naive probability* of A is

$$P_{\text{naive}}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to } A}{\text{total number of outcomes in } S}.$$

(We use $|A|$ to denote the size of A; see Section A.1.5 of the math appendix.)

A naive definition of probability can be done in terms of the size (area, mass) of events. P(A) = |A| / |S| = # of outcomes in A / total number of outcomes in S



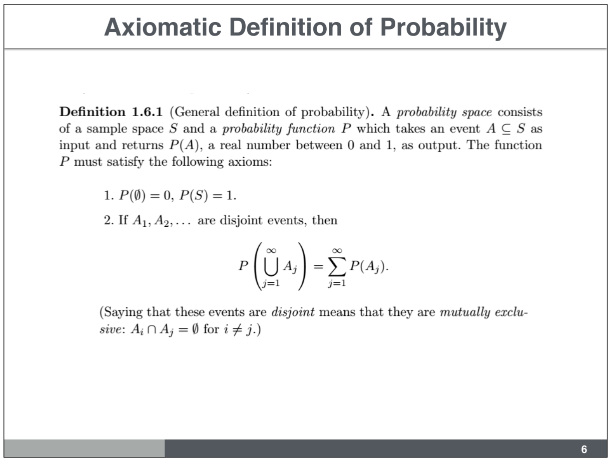
Exercise: what's the probability of A?

Definition 1.3.1 (Naive definition of probability). Let A be an event for an experiment with a finite sample space S. The *naive probability* of A is

$$P_{\text{naive}}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to } A}{\text{total number of outcomes in } S}.$$

(We use $|A|$ to denote the size of A; see Section A.1.5 of the math appendix.)

What's the probability of A?



Axiomatic Definition of Probability

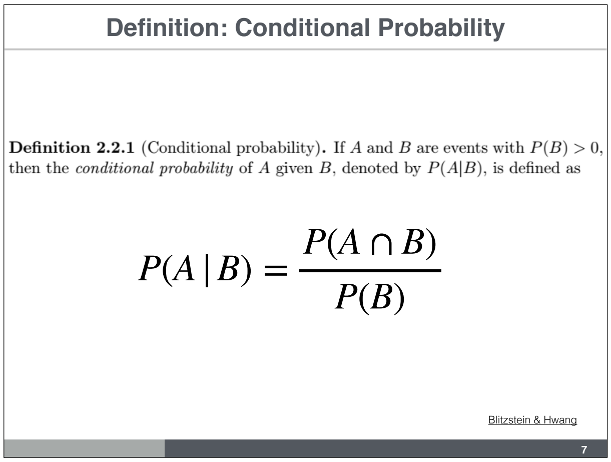
Definition 1.6.1 (General definition of probability). A *probability space* consists of a sample space S and a *probability function* P which takes an event $A \subseteq S$ as input and returns $P(A)$, a real number between 0 and 1, as output. The function P must satisfy the following axioms:

1. $P(\emptyset) = 0$, $P(S) = 1$.
2. If $A_1, A_2, \dots$ are disjoint events, then

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P(A_j).$$

(Saying that these events are *disjoint* means that they are *mutually exclusive*: $A_i \cap A_j = \emptyset$ for $i \neq j$.)

For a more general definition of probability, we need a probability space, which consists of a sample space, and a probability function P. The P function assigns a number between 0 and 1 to events and must comply with the following rules: the probability of the empty set is 0, the probability of the sample space is 1, and the union of disjoint events is equal to the sum of the probability of each event.



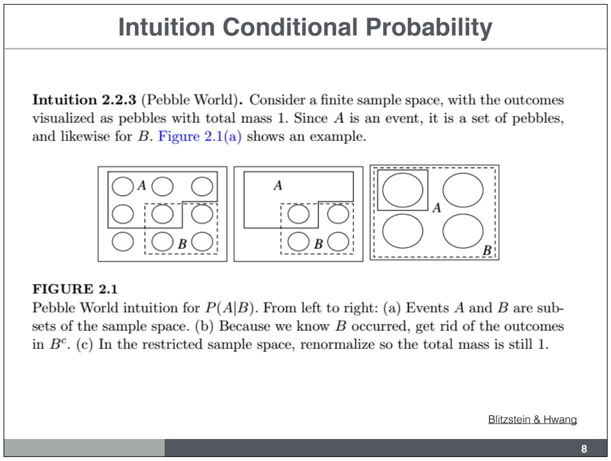
Definition: Conditional Probability

Definition 2.2.1 (Conditional probability). If A and B are events with $P(B) > 0$, then the *conditional probability* of A given B, denoted by $P(A|B)$, is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Blitzstein & Hwang

Conditional probability of an even A given B is defined as the ratio of the probability of the intersection between A and B, divided by the probability of B.



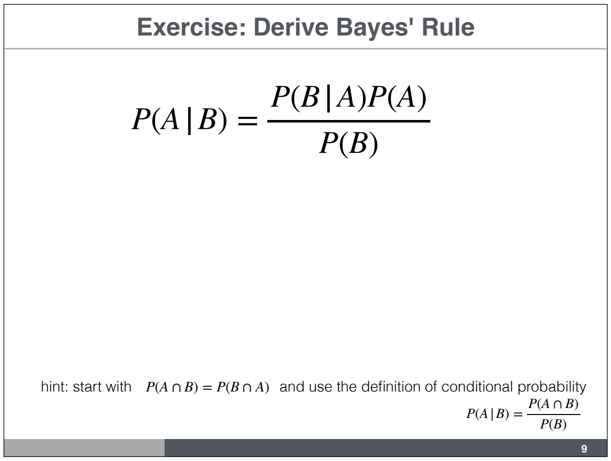
Intuition Conditional Probability

Intuition 2.2.3 (Pebble World). Consider a finite sample space, with the outcomes visualized as pebbles with total mass 1. Since A is an event, it is a set of pebbles, and likewise for B. Figure 2.1(a) shows an example.

FIGURE 2.1
Pebble World intuition for $P(A|B)$. From left to right: (a) Events A and B are subsets of the sample space. (b) Because we know B occurred, get rid of the outcomes in B^c. (c) In the restricted sample space, renormalize so the total mass is still 1.

Blitzstein & Hwang

The intuition behind the conditional probability is illustrated in Figure 2.1. The rectangle with 9 pebbles is the sample space. Event A is shown as a subset of S. When we condition on B, we consider that B has occurred. Thus we need to exclude all the pebbles outside of B, which is accomplished by taking the intersection with B. The conditional probability of B|B has to be one. This is accomplished by dividing by P(B).



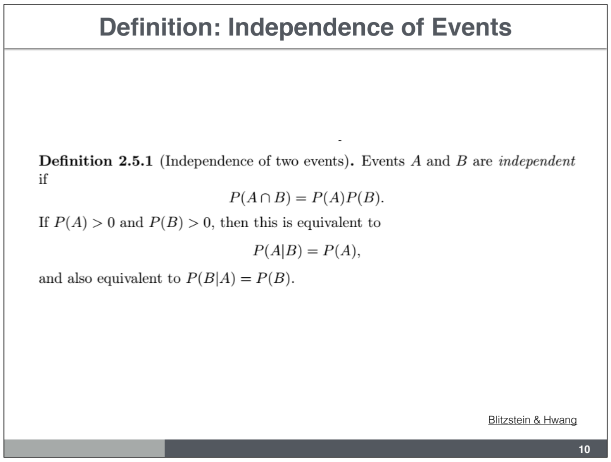
Exercise: Derive Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

hint: start with $P(A \cap B) = P(B \cap A)$ and use the definition of conditional probability $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Derive the Bayes' rule.

This seemingly simple consequence of the definition of conditional distribution is the basis of the powerful Bayesian inference.



Definition: Independence of Events

Definition 2.5.1 (Independence of two events). Events A and B are *independent* if

$$P(A \cap B) = P(A)P(B).$$

If $P(A) > 0$ and $P(B) > 0$, then this is equivalent to

$$P(A|B) = P(A),$$

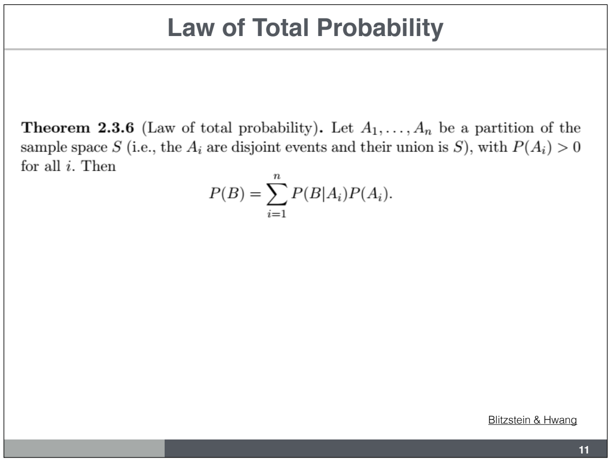
and also equivalent to $P(B|A) = P(B)$.

Blitzstein & Hwang

Two events are said to be independent if the probability of their intersection (joint probability) is the product of their (marginal) probabilities. One can show that this definition is equivalent to the more intuitive one: that the probability of A does not change knowing that B has occurred. Similarly, that the probability of B does not depend on knowing that A has occurred.

Why do we require P(A) and P(B) different from 0 for the last two conditions and not for the first?

Exercise: prove that the three definitions are equivalent.



Law of Total Probability

Theorem 2.3.6 (Law of total probability). Let $A_1, \dots, A_n$ be a partition of the sample space S (i.e., the A_i are disjoint events and their union is S), with $P(A_i) > 0$ for all i. Then

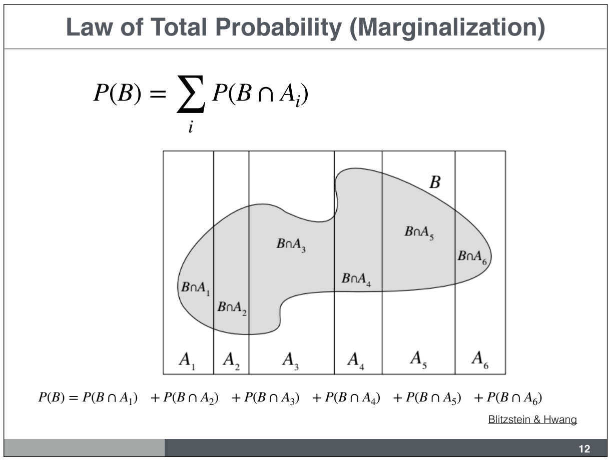
$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i).$$

Blitzstein & Hwang

The law of total probability states that given a partition of the sample space, the probability of an event B can be calculated as the sum of conditional probabilities of B within each slice in the partition, weighted by the probability of the slice.

Can you think of reasons why this law could this be useful?

If we partition the sample space wisely, conditional probabilities can be much easier to calculate the the total marginal probability. Intuitively, this can happen because conditional probabilities have additional information.



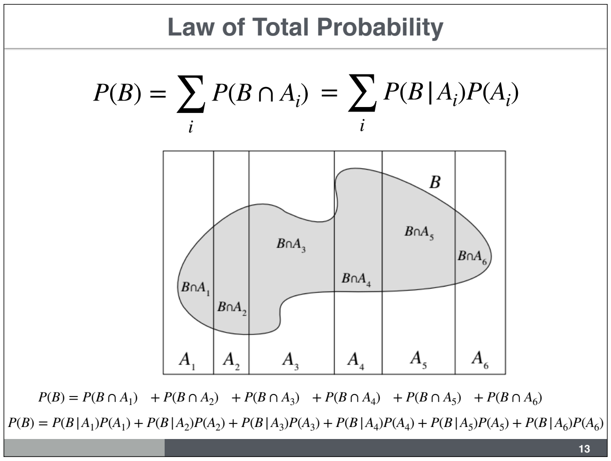
Law of Total Probability (Marginalization)

$$P(B) = \sum_i P(B \cap A_i)$$

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4) + P(B \cap A_5) + P(B \cap A_6)$$

Blitzstein & Hwang

It is quite intuitive that the probability of B can be calculated directly or as the sum of the probabilities of the intersection of B with disjoint slices of the sample space. This is sometimes known as marginalization.



Law of Total Probability

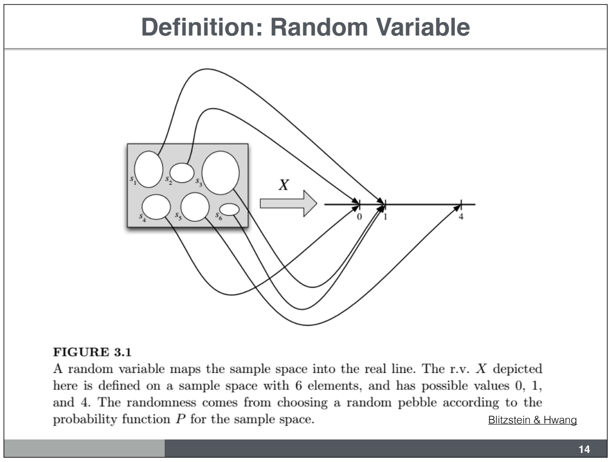
$$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B|A_i)P(A_i)$$

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4) + P(B \cap A_5) + P(B \cap A_6)$$

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4) + P(B|A_5)P(A_5) + P(B|A_6)P(A_6)$$

The law of total probability can be shown by simple application of the definition of conditional probability.

If you haven't seen an application of this law in practice, it may not look very promising. However, it happens that computations can be simplified a lot if we choose the partition judiciously.

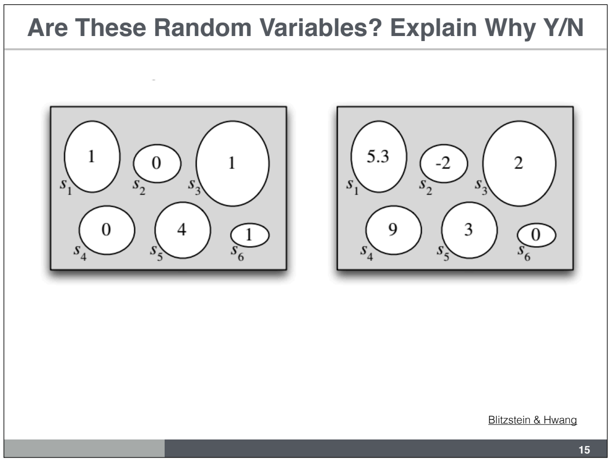


Definition: Random Variable

FIGURE 3.1
A random variable maps the sample space into the real line. The r.v. X depicted here is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The randomness comes from choosing a random pebble according to the probability function P for the sample space.

Blitzstein & Hwang

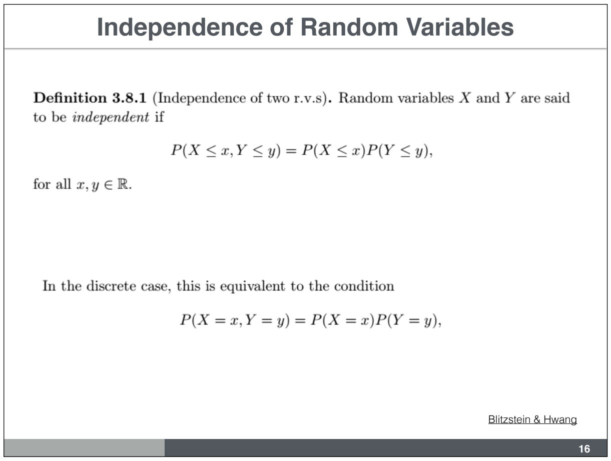
Random variables map outcomes in the sample space into the real line. In this example, each of the 6 elements in the sample space map into the real line.



Are These Random Variables? Explain Why Y/N

Blitzstein & Hwang

Here each rectangle is a sample space with the numbers in each pebble representing the value of the random variable. Does the rectangle on the left represent a random variable? How about the one on the right? Explain why.



Independence of Random Variables

Definition 3.8.1 (Independence of two r.v.s). Random variables X and Y are said to be *independent* if

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

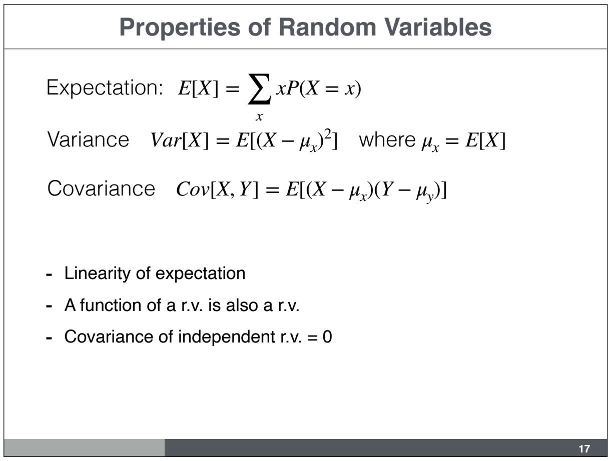
for all $x, y \in \mathbb{R}$.

In the discrete case, this is equivalent to the condition

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

Blitzstein & Hwang

Random variables X and Y are said to be independent if the joint probability of X being no greater than x and Y being no greater than y is the product of their marginal probabilities. In the discrete case, this is equivalent to the probability that X=x and Y=y is the product of their marginal probabilities.



Properties of Random Variables

Expectation: $E[X] = \sum_x xP(X = x)$

Variance $Var[X] = E[(X - \mu_x)^2]$ where $\mu_x = E[X]$

Covariance $Cov[X, Y] = E[(X - \mu_x)(Y - \mu_y)]$

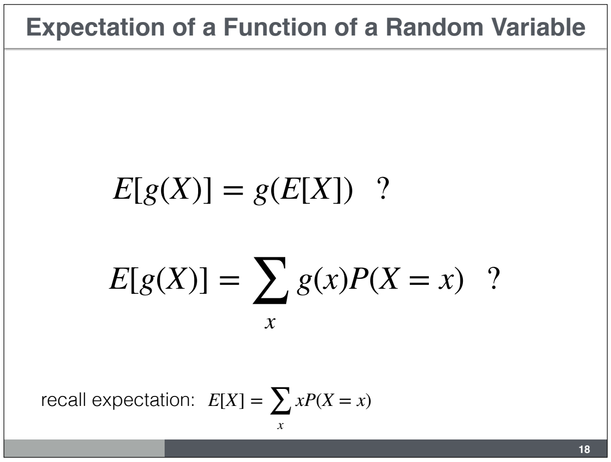
- Linearity of expectation
- A function of a r.v. is also a r.v.
- Covariance of independent r.v. = 0

The expected value (also called mean) of a discrete random variable is the sum over all values x that the r.v. can take, weighted by the probability of each x. The variance is the expected value of the squared difference between the r.v. and its mean.

Linearity of the expectation: E[aX + bY] = aE[X] + bE[Y]

if X is a r.v., a real valued function g(X) is also a r.v.

if X and Y are independent, then Cov(X,Y) = 0. How about the converse, does Cov(X,Y)=0 implies that X and Y are independent? If not, when would the converse hold?



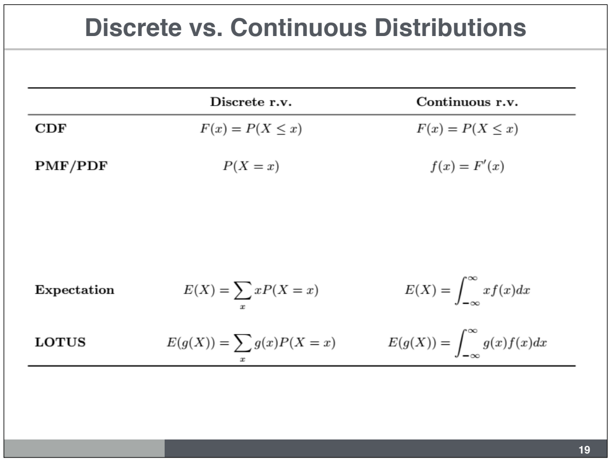
Expectation of a Function of a Random Variable

$$E[g(X)] = g(E[X]) \quad ?$$

$$E[g(X)] = \sum_x g(x)P(X = x) \quad ?$$

recall expectation: $E[X] = \sum_x xP(X = x)$

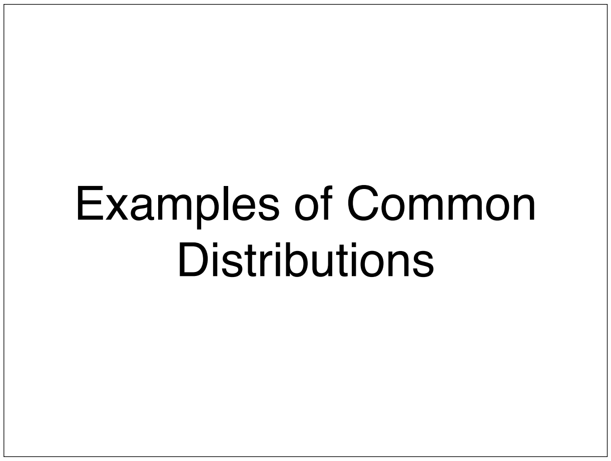
Are these equalities true in general? If not, when do they hold?



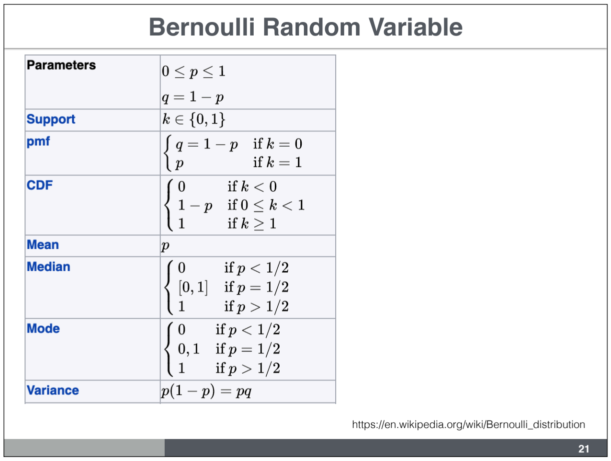
Discrete vs. Continuous Distributions

	Discrete r.v.	Continuous r.v.
CDF	$F(x) = P(X \leq x)$	$F(x) = P(X \leq x)$
PMF/PDF	$P(X = x)$	$f(x) = F'(x)$
Expectation	$E(X) = \sum_x xP(X = x)$	$E(X) = \int_{-\infty}^{\infty} xf(x)dx$
LOTUS	$E(g(X)) = \sum_x g(x)P(X = x)$	$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$

For convenience, we have used the definitions for discrete r.v. For continuous r.v., we need to use integrals instead of sums and instead of the probability mass function, we need to use the probability density function.



Examples of Common Distributions

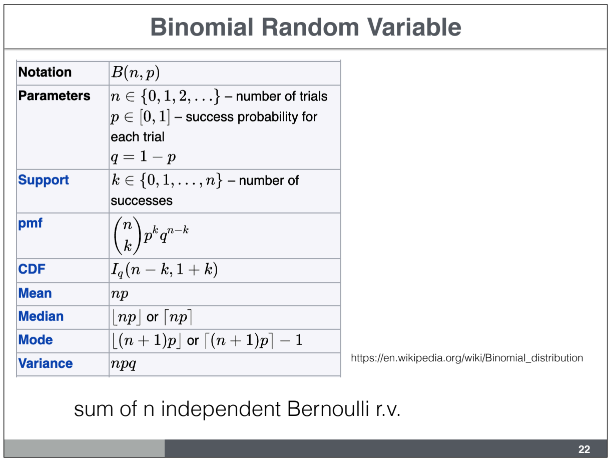


Bernoulli Random Variable

Parameters	$0 \leq p \leq 1$ $q = 1 - p$
Support	$k \in \{0, 1\}$
pmf	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
CDF	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$
Mean	p
Median	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0, 1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Mode	$\begin{cases} 0 & \text{if } p < 1/2 \\ 0, 1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
Variance	$p(1 - p) = pq$

https://en.wikipedia.org/wiki/Bernoulli_distribution

Can you think of an experiment that results in a bernoulli r.v?



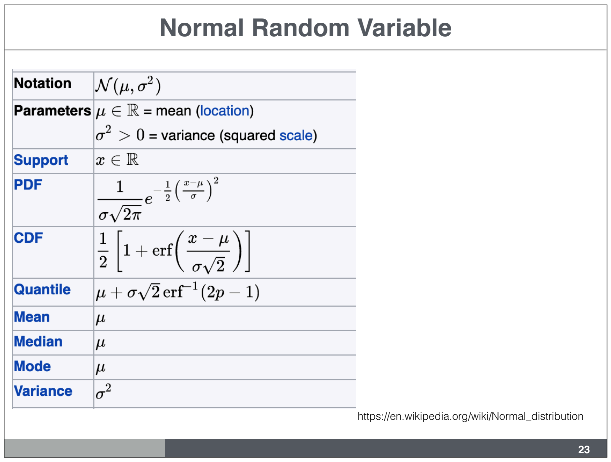
Binomial Random Variable

Notation	$B(n, p)$
Parameters	$n \in \{0, 1, 2, \ldots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial $q = 1 - p$
Support	$k \in \{0, 1, \ldots, n\}$ – number of successes
pmf	$\binom{n}{k} p^k q^{n-k}$
CDF	$I_q(n - k, 1 + k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n + 1)p \rfloor$ or $\lceil (n + 1)p \rceil - 1$
Variance	npq

https://en.wikipedia.org/wiki/Binomial_distribution

sum of n independent Bernoulli r.v.

describe an experiment that would yield a binomial random variable



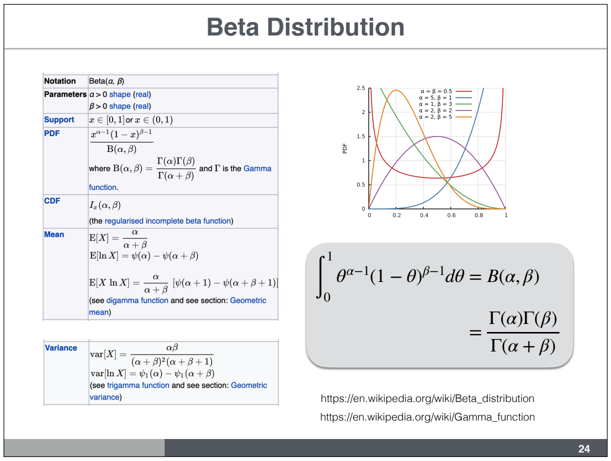
Normal Random Variable

Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ = mean (location) $\sigma^2 > 0$ = variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
CDF	$\frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x - \mu}{\sigma\sqrt{2}}\right)\right]$
Quantile	$\mu + \sigma\sqrt{2}\operatorname{erf}^{-1}(2p - 1)$
Mean	μ
Median	μ
Mode	μ
Variance	σ^2

https://en.wikipedia.org/wiki/Normal_distribution

Normal r.v. is the most commonly used continuous r.v.

Processes that are an accumulation of multiple small effects are modeled well as normal r.v.. Can you think of why?



Beta Distribution

Notation	Beta(α, β)
Parameters	$\alpha > 0$ shape (real) $\beta > 0$ shape (real)
Support	$x \in [0, 1]$ or $x \in (0, 1)$
PDF	$\frac{x^{\alpha-1}(1 - x)^{\beta-1}}{\mathrm{B}(\alpha, \beta)}$ where $\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$ and Γ is the Gamma function.
CDF	$I_x(\alpha, \beta)$ (the regularised incomplete beta function)
Mean	$\mathrm{E}[X] = \frac{\alpha}{\alpha + \beta}$ $\mathrm{E}[\ln X] = \psi(\alpha) - \psi(\alpha + \beta)$ $\mathrm{E}[X \ln X] = \frac{\alpha}{\alpha + \beta}\,[\psi(\alpha + 1) - \psi(\alpha + \beta + 1)]$ (see digamma function and see section: Geometric mean)

Variance	$\operatorname{var}[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$ $\operatorname{var}[\ln X] = \psi_1(\alpha) - \psi_1(\alpha + \beta)$ (see trigamma function and see section: Geometric variance)

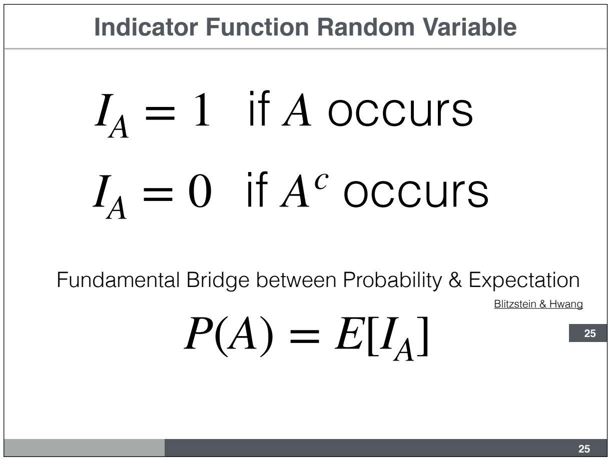
$$\int_0^1 \theta^{\alpha-1}(1 - \theta)^{\beta-1} d\theta = B(\alpha, \beta)$$

$$= \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

https://en.wikipedia.org/wiki/Beta_distribution

https://en.wikipedia.org/wiki/Gamma_function

Since the integral of the pdf of any continuous random variable over its support has = 1. This means that the integral in the gray box has to be B(α,β). This knowledge will come in handy later when we try to find the P(data) for Bayesian inference.



Indicator Function Random Variable

$$I_A = 1 \quad \text{if } A \text{ occurs}$$

$$I_A = 0 \quad \text{if } A^c \text{ occurs}$$

Fundamental Bridge between Probability & Expectation

Blitzstein & Hwang

$$P(A) = E[I_A]$$

25

The indicator function is a deceivingly simple r.v. that can make calculations easier. The probability of an event can be calculated as the expected value of the indicator function. Blitzstein and Hwang thought that this was so important that they called it the "Fundamental bridge between probability and expectation".