ISTA 421 + INFO 521
Machine Learning

Probability Review

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A, 621-6609

12 September 2016

References for probability

Recommend:

- (lvl 1) Doing Bayesian Data Analysis (**DBDA**)
  - Ch 2, 4, 5
- (lvl 2) First Course in Machine Learning (**FCML**)
  - Ch 2.2 (foundations),
  - Ch 2.3 (Discrete),
  - Ch 2.4-2.5 (Continuous)
  - Ch 2.6-2.7 (Expectation and Maximum Likelihood)
  - Ch 3 (Bayesian)
- (lvl 3) Pattern Recognition and Machine Learning (**PRML**)
  - Ch 1.2 (foundations),
  - Ch 2.1-2.2 (Discrete),
  - Ch 2.3 (Continuous)

Google (and WikiPedia) for unfamiliar terms and alternative explanations.

Wisdom from tea dipper handle

Probability semantics

Two broad interpretations of probability
(variants exist for both)

1) Representation of expected frequency (“frequentist”)

2) Degree of belief (“Bayesian”)

There is a 20% chance of rain tomorrow.

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A “generic” outcome is often denoted by ω
and we can say things like, e.g., “for each $\omega \in \Omega$…”

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A "generic" outcome is often denoted by ω
and we can say things like, e.g., "for each $\omega \in \Omega$…"

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A "generic" outcome is often denoted by ω
and we can say things like, e.g., "for each $\omega \in \Omega$…"

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Semantics of Set Operations

Equivalence between "set" and "proposition" representations.

1. Set E: outcomes s.t. proposition E is true.
2. Union, $E \cup F$: logical OR between propositions E and F.
3. Intersection, $E \cap F$: logical AND
4. Complement, E^C: logical negation

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A "generic" outcome is often denoted by ω
and we can say things like, e.g., "for each $\omega \in \Omega$…"

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Denote the **collection of measurable events** (ones we want to assign probabilities to) by S.

S must include $\varnothing$ and Ω

These special events are represent the cases where "nothing" among all the choices happens (impossible), and "something" happens (certain).

Reason for being technical: It is important to be tuned into **what** a particular probability is **about** (precisely!).

Basic terminology and rules

Sample Space of *outcomes* (often denoted by Ω)

{H, T}

{1, 2, 3, 4, 5, 6}

An outcome is just ONE element of the sample space
A "generic" outcome is often denoted by ω
and we can say things like, e.g., "for each $\omega \in \Omega$…"

Event (subset of Ω) …does or does not contain (is true or false for) a particular outcome

odd {1, 3, 5}, even {2, 4, 6}, prime {2, 3, 5}

Denote the **collection of measurable events** (ones we want to assign probabilities to) by S.

S must include $\varnothing$ and Ω

S is *closed* under set operations

…aka: σ-algebra

$$\alpha, \beta \in S \;\Rightarrow\; \alpha \cup \beta \in S\;,\; \alpha \cap \beta \in S\;,\; \alpha^C = \Omega - \alpha \in S\;, \text{etc.}$$

Translation: We need to be able to deal with concepts such as "either A or B" happens, or "both A and B" happen.

E.g., I'll accept either an even or prime number

E.g., If I roll a 3, it is both odd and prime

Basic terminology and rules

Probability Space

A **probability space** is a sample space augmented with a function, P, that assigns a **probability** to each event, $E \subset S$.

Kolmogorov Axioms

1. $0 \leq P(E) \leq 1$ for all $E \subset S$.
2. $P(\Omega) = 1$.
3. If $E \cap F = \emptyset$ then $P(E \cup F) = P(E) + P(F)$.

Important Consequences

1. $P(\emptyset) = 0$.
2. $P(E^{C}) = 1 - P(E)$
3. In general, $P(E \cup F) = P(E) + P(F) - P(E \cap F)$.

Random Variables

Random variables

- Defined by **functions** mapping **outcomes** (ω) to **values**
- A random variable is a way of reporting an attribute of an outcome
- Typically r.v. are denoted by uppercase letters (e.g., X)
- Generic values are corresponding lower case letters (e.g., x)
- Shorthand: P(x) = P(X=x)
- Value “type” is arbitrary (typically categorical or real)

Example (from K&F)

- Outcomes are student grades (A,B,C)
- Random variable $G=f_{GRADE}(\text{student})$

$$P('A') = P(G='A') = P(\{w \in \Omega : f_{GRADE}(w) = 'A'\})$$

We sometimes use sets, but usually R.Vs.: $P(\overbrace{A \cap B \cap C}^{\text{Sets}}) \equiv P(\overbrace{A, B, C}^{\text{R. Vs.}})$

Random Variables

Random Variable

- Formally, a **random variable** is a function, X that assigns a number to each outcome in S (e.g., dead $\to 0$, alive $\to 1$).
- Key consequence: a random variable divides the sample space into **equivalence classes**: sets of outcomes that share some property (differ only in ways irrelevant to X)

Example

- Let $S =$ all sequences of 3 coin tosses.
- We can define a r.v. X that counts number of heads.
- Then HHT and HTH are equivalent in the eyes of X:

$$X(HHT) = X(HTH) = 2$$

Random Variables

Distribution of a Random Variable

- The expression $P(X = x)$ refers to the probability of the event $E = \{\omega \in S : X(\omega) = x\}$.
- Sometimes we can obtain it by breaking it down into simpler, mutually exclusive events and adding their probabilities (Kolmogorov axiom 3)

Example

- $S =$ all sequences of 3 coin tosses.
- $X(\omega) =$ # of heads in ω.

$$\begin{aligned}\{X = 2\} &= \{HHT\} \cup \{HTH\} \cup \{THH\} \\ P(X = 2) &= P(HHT) + P(HTH) + P(THH) \\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\end{aligned}$$

Random Variables

Distribution of a Random Variable

- Similarly, $P(X < x)$ is the probability of the event $E = \{\omega \in S : X(\omega) < x\}$.
- Can sometimes obtain it the same way as we did above.

Example

- $S =$ all sequences of 3 coin tosses.
- $X(\omega) =$ # of heads in ω.

$$
\begin{aligned}
\{X < 2\} &= \{TTT\} \cup \{TTH\} \cup \{THT\} \cup \{HTT\} \\
P(X < 2) &= P(TTT) + P(TTH) + P(THT) + P(HTT) \\
&= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}
\end{aligned}
$$

Random Variables

Distribution of a Random Variable

Example, continued

- Notice that in this example we could also have written

$$\{X < 2\} = \{X = 0\} \cup \{X = 1\}$$
$$P(X < 2) = P(X = 0) + P(X = 1)$$

  which is useful if we have already calculated $P(X = x)$ for each value of x.
- This always works if X is always an integer.

Joint Probability

Joint Probability

- We have already seen the concept of *intersecting events*: $A \cap B$ is the event that occurs when *both* A and B are true *at the same time*.
- $P(A \cap B)$ is called the **joint probability** of A and B.
- If A is $\{X = x\}$ and B is $\{Y = y\}$, then $A \cap B$ means $X = x$ and $Y = y$ *at the same time*.
- If X and Y are discrete, $P(X = x, Y = y)$, for different combinations of x and y, characterize the **joint distribution** of X and Y.

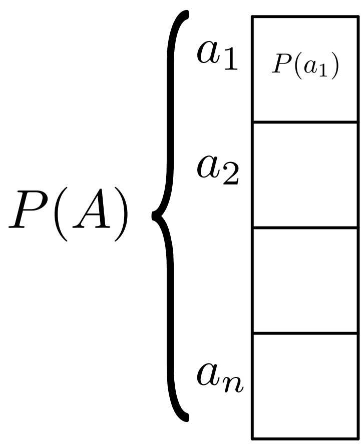
We write $P(x,y)$ for $P(\{w \in \Omega : X(w) = x \text{ and } Y(w) = y\})$

Alternatively, $P((X = x) \cap (Y = y))$

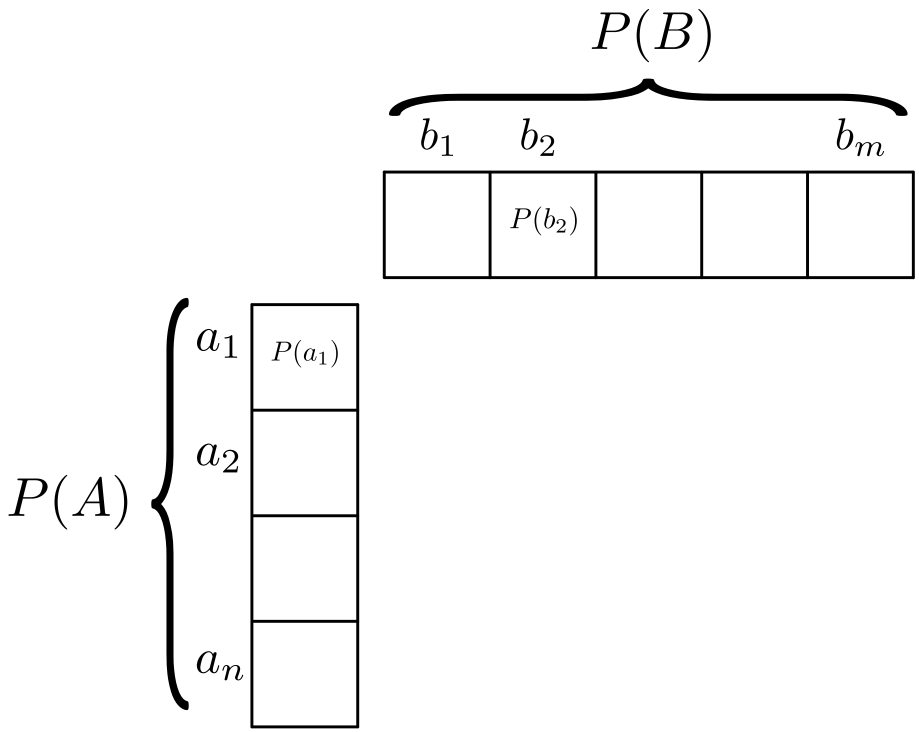
Note that the comma in the usual form, $P(x,y)$, is read as "and".

Here events are being defined by assignments of random variables

Joint Probability

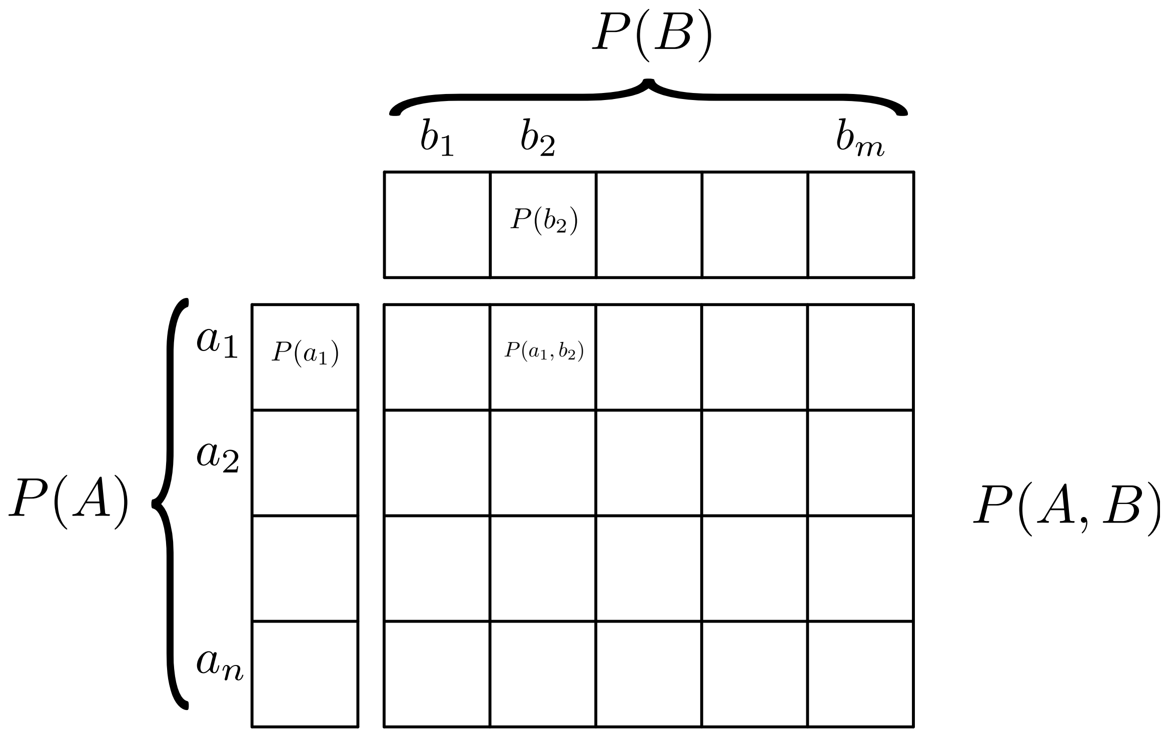


Joint Probability



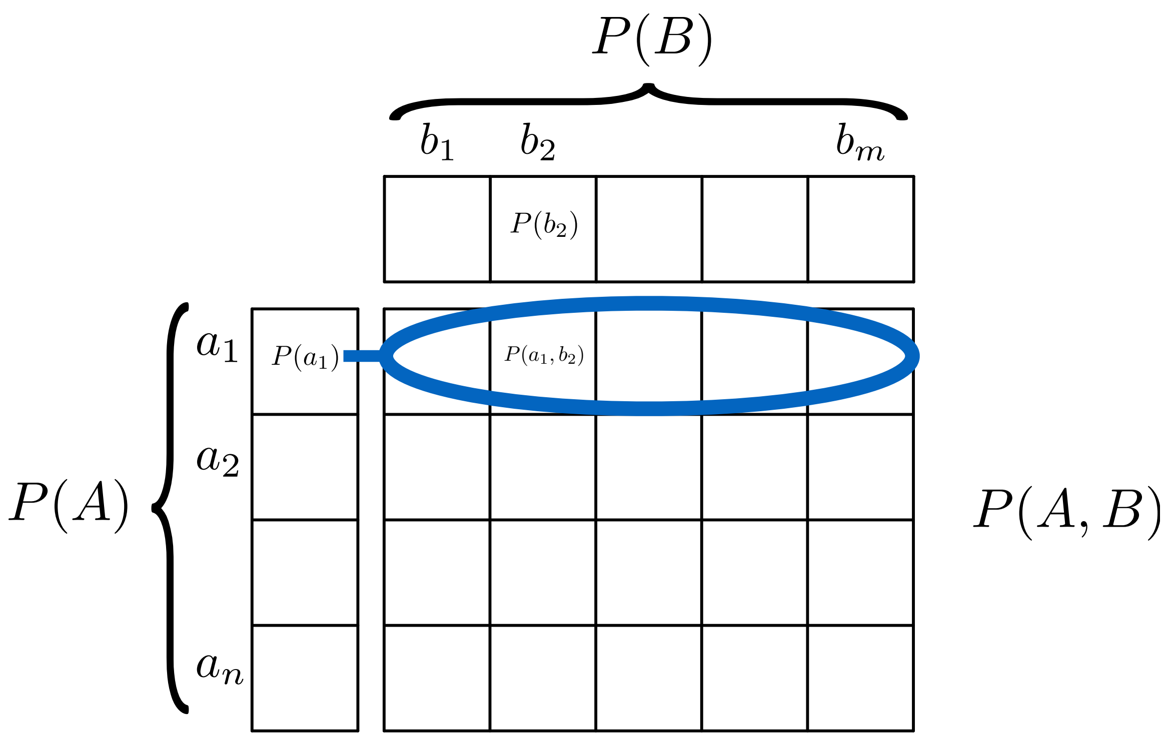
Joint Probability

Joint Probability



Joint Probability

Joint Probability



Marginalization: $$P(A) = \sum_{b \in B} P(A, B) \quad *$$

Formulas that you should be comfortable with are marked by * .

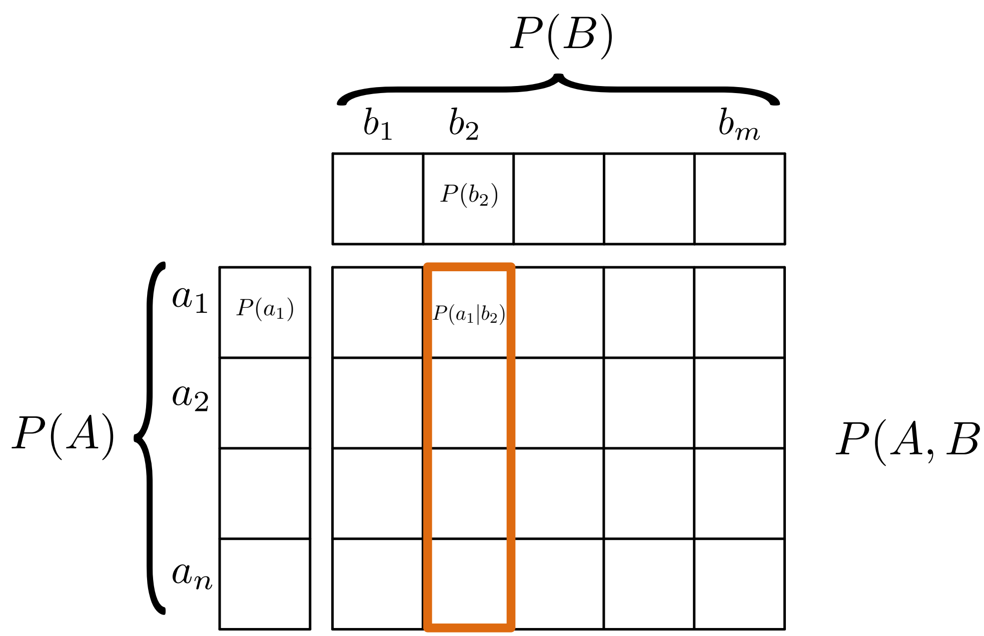
Conditional Probability

"probability in context"

Conditional probability (definition)

$$P\left(A\middle|B\right) \equiv \frac{P(A \cap B)}{P(B)}$$

*



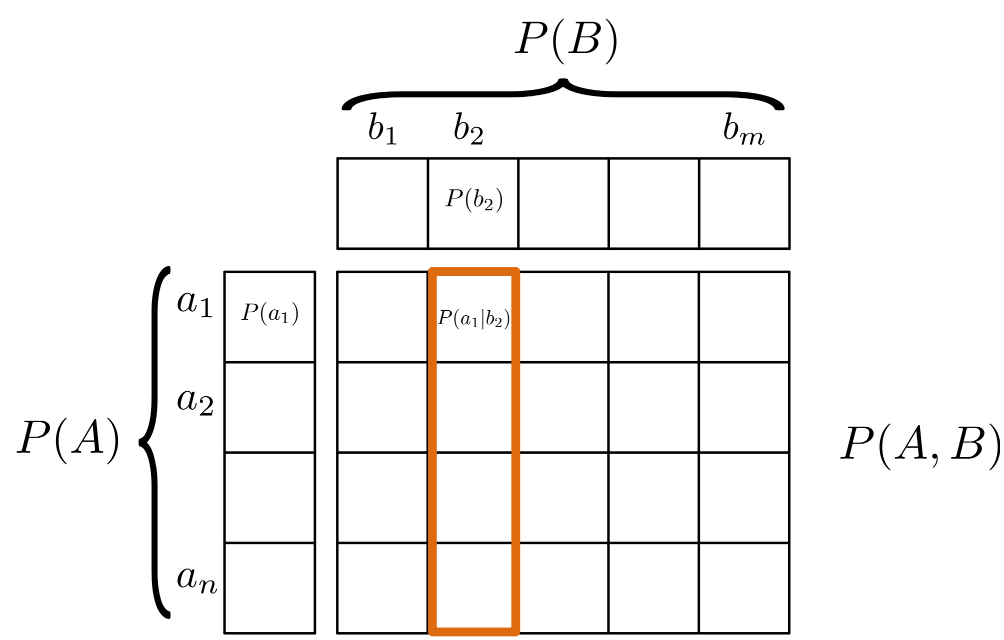
Conditional Probability

“probability in context”

Conditional probability (definition)

$$P\left(A \middle| B\right) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Example: what is the probability that you roll 2 (on a six sided die), given that you know you have rolled a prime number?



Product Rule

“probability in context”

Conditional probability (definition)

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)} \quad *$$

Applying a bit of algebra,

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

Chain (Product) Rule

"probability in context"

Conditional probability (definition)

$$P\left(A|B\right) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Applying a bit of algebra,

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

In general, we have the **chain** (**product**) rule:

Product $$P\left(A_1 \cap A_2\right) = P(A_1)P(A_2|A_1)$$

*

Chain $$P\left(A_1 \cap A_2 \cap \;\; .... \;\; A_N\right) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \;\; .... \;\; P(A_N|A_1 \cap A_2 \cap \;\; .... \;\; A_{N-1})$$

Bayes Rule

Going back to the definition of conditional probability

$$P\left(A \middle| B\right) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a little bit more algebra,

$$P\left(A \cap B\right) = P(A)P(B|A)$$

and $$P\left(A \cap B\right) = P(B)P(A|B)$$

and thus $$P(B)P(A|B) = P(A)P(B|A)$$

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

Bayes rule *

Bayes Rule

Going back to the definition of conditional probability

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

Applying a little bit more algebra,

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

$$\text{and} \quad P(A \cap B) = P(B)P(A|B)$$

$$\text{and thus} \quad P(B)P(A|B) = P(A)P(B|A)$$

$$\text{and we get} \quad P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayes rule *

<u>Pro tip!</u>: Common to represent denominator as marginalization of numerator:

$$P(B) = \sum_{a \in A} P(A, B)$$
$$= \sum_{a \in A} P(A)P(B|A)$$

Expectation

The **expected value** of a function of a random variable *X* that is distributed according to P(X) is:

$$\mathbf{E}_{P(x)}\{f(X)\} = \sum_x f(x)P(x)$$

The expected value of a (function of a) random variable is the **weighted (by probability) average** of all possible values of that variable (through that function).

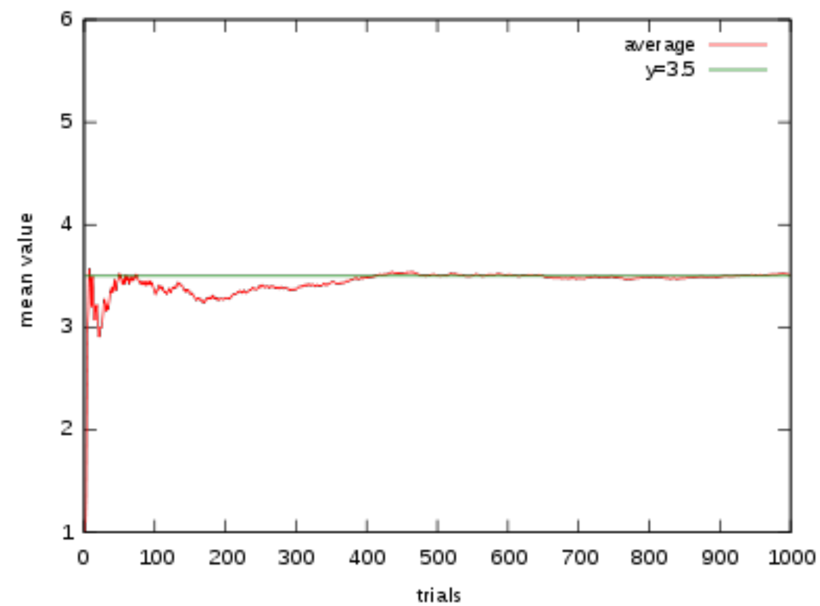
The expected value of the random variable *X* itself: the **mean**

$$\mathbf{E}_{P(x)}\{X\} = \sum_x xP(x)$$

What is the relationship of the *arithmetic mean* to the expected value?

$$= \frac{1}{N}\sum_{i=1}^{N} x_i$$

(cumulative) average dice value against number of rolls



Expectation

$$\mathbf{E}_{P(x)}\left\{f(X)\right\} \quad = \quad \sum_x f(x)P(x)$$

The expectation of the value of *X* if *X* is a fair die:

$$\mathbf{E}_{P(x)}\left\{X\right\} = \sum_x x\frac{1}{6} = \frac{1}{6} + \frac{2}{6} + \ldots + \frac{6}{6} = \frac{21}{6} = (3.5)^2 = 12.25$$

$$\mathbf{E}_{P(x)}\left\{X^2\right\} = \sum_x x^2\frac{1}{6} = \frac{1}{6} + \frac{4}{6} + \ldots + \frac{36}{6} = \frac{91}{6} \quad \approx 15.17$$

$$12.25 \quad \neq \quad 15.17$$

$$\left(\mathbf{E}_{P(x)}\left\{X\right\}\right)^2 \quad \neq \quad \mathbf{E}_{P(x)}\left\{X^2\right\}$$

Expectation

$$\mathbf{E}_{P(x)}\{f(X)\} \quad = \quad \sum_x f(x)P(x)$$

In *general*: the expected value of a function of X is ***not equal*** to the function evaluated at the expected value of X!

$$f(\mathbf{E}_{P(x)}\{X\}) \overset{\textit{usually}}{\neq} \mathbf{E}_{P(x)}\{f(X)\}$$

BUT! These cases *do* hold:

$f(X) = a \quad : \ \mathbf{E}_{P(x)}\{X\} = a$

$f(X) = aX \ : \ \mathbf{E}_{P(x)}\{f(aX)\} = a\mathbf{E}_{P(x)}\{f(X)\}$

$\mathbf{E}_{P(x)}\{f(X) + g(X)\} = \mathbf{E}_{P(x)}\{f(X)\} + \mathbf{E}_{P(x)}\{g(X)\}$

Expectation: Variance

$$\mathbf{E}_{P(x)}\{f(X)\} \quad = \quad \sum_x f(x)P(x)$$

Variance:

$$\mathrm{var}\{X\} = \mathbf{E}_{P(x)}\left\{\left(X - \mathbf{E}_{P(x)}\{x\}\right)^2\right\}$$

$$\begin{aligned}
\mathrm{var}\{X\} &= \mathbf{E}_{P(x)}\left\{\left(X - \mathbf{E}_{P(x)}\{x\}\right)^2\right\} \\
&= \mathbf{E}_{P(x)}\left\{X^2 - 2X\mathbf{E}_{P(x)}\{X\} + \mathbf{E}_{P(x)}\{x\}^2\right\} \\
&= \mathbf{E}_{P(x)}\left\{X^2\right\} - 2\mathbf{E}_{P(x)}\{X\}\,\mathbf{E}_{P(x)}\{X\} + \mathbf{E}_{P(x)}\{X\}^2
\end{aligned}$$

$$\mathrm{var}\{X\} = \mathbf{E}_{P(x)}\left\{X^2\right\} - \mathbf{E}_{P(x)}\{X\}^2$$

Basic rules (so far)

Marginalization

$$P(A) = \sum_{b \in B} P(A, B)$$ *

Conditional probability (definition)

$$P\left(A \middle| B\right) \equiv \frac{P(A \cap B)}{P(B)}$$ *

Chain (Product) Rule

$$P\left(A_1 \cap A_2\right) = P(A_1)P(A_2 \mid A_1)$$ *

$$P\left(A_1 \cap A_2 \cap \quad .... \quad A_N\right) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \; .... \; P(A_N \mid A_1 \cap A_2 \cap \quad .... \quad A_{N-1})$$

Bayes Rule

$$P\left(A \middle| B\right) = \frac{P(B \mid A)P(A)}{P(B)}$$ *