Security Analytics

Topic 3: Review of Probability

Purdue University
Prof. Ninghui Li
Based on slides by Prof. Jenifer Neville and Chris Clifton

Readings

- Chapter 4 Data Analysis and Uncertainty
  - Sections 4.1 to 4.3
- Handout: "Background on Probability and Statistics"
- Stefan Axelsson. The Base-Rate Fallacy and the Difficulty of Intrusion Detection. In ACM TISSEC 2000.

Quiz

1. A standard normal distribution has:
   (a) mean equal to the variance
   (b) mean equal 1 and variance equal 1
   (c) mean equal 0 and variance equal 1
   (d) mean equal 0 and standard deviation equal 0
   (e) none of these
2. **True or False**: P(A and B) = P(A|B)P(B|A).
3. **True or False**: If P(A|B) = P(A) then A and B are independent.
4. A card is drawn at random from a deck of playing cards. If it is red, the player wins 1 dollar; if it is black, the player loses 2 dollars. Find the expected value of the game.
5. An urn contains eight balls, two are red and six white. Two balls are drawn at random with replacement. What is the probability that at least one of the balls drawn is red?
6. What if two balls are drawn without replacement?

Quiz

1. A standard normal distribution has:
   (a) mean equal to the variance
   (b) mean equal 1 and variance equal 1
   (c) mean equal 0 and variance equal 1
   (d) mean equal 0 and standard deviation equal 0
   (e) none of these
2. **True or False**: P(A and B) = P(A|B)P(B|A).
3. **True or False**: If P(A|B) = P(A) then A and B are independent.
4. A card is drawn at random from a deck of playing cards. If it is red, the player wins 1 dollar; if it is black, the player loses 2 dollars. Find the expected value of the game.
   E = P(red)*1 + P(black)*(-2) = 1/2 - 1 = -1/2
5. An urn contains eight balls, two are red and six white. Two balls are drawn at random with replacement. What is the probability that at least one of the balls drawn is red?
   1-P(two white) = 1 - (6/8)^2 = 1 - 9/16 = 7/16
6. Without replacement, **1-P(two white) = 1 - (6/8)(5/7) = 1 - 15/28 = 13/28**

Probability basics

- Basic notion: **Random variable (RV)**
  - A variable that can take one of a set of possible values
  - *X* refers to random variable; *x* refers to a value of that random variable
- Types of random variables
  - Discrete RV has a finite set of possible values; Continuous RV can take any value within an interval
  - **Boolean**: e.g., Warning (will there be a storm warning? = <yes, no>)
  - **Discrete**: e.g., The weather tomorrow is one of <sunny,rainy,cloudy,snow>
  - **Continuous**: e.g., The average temperature tomorow

Probability basics

- **Sample space (S)**
  - Set of all possible outcomes of an experiment
- **Event**
  - Any subset of *outcomes* contained in the sample space S
  - When events ***A*** and ***B*** have no outcomes in common they are said to be *mutually exclusive*

Examples

Random variable(s)	**Sample space**
One coin toss	H, T
Two coin tosses	HH, HT, TH, TT
Select one card	2♥, 2♠, ..., A♣ (52)
Result of a chess game	Win, Lose, Draw
Inspect a part	Defective, OK
Cavity and toothache	TT, TF, FT, FF

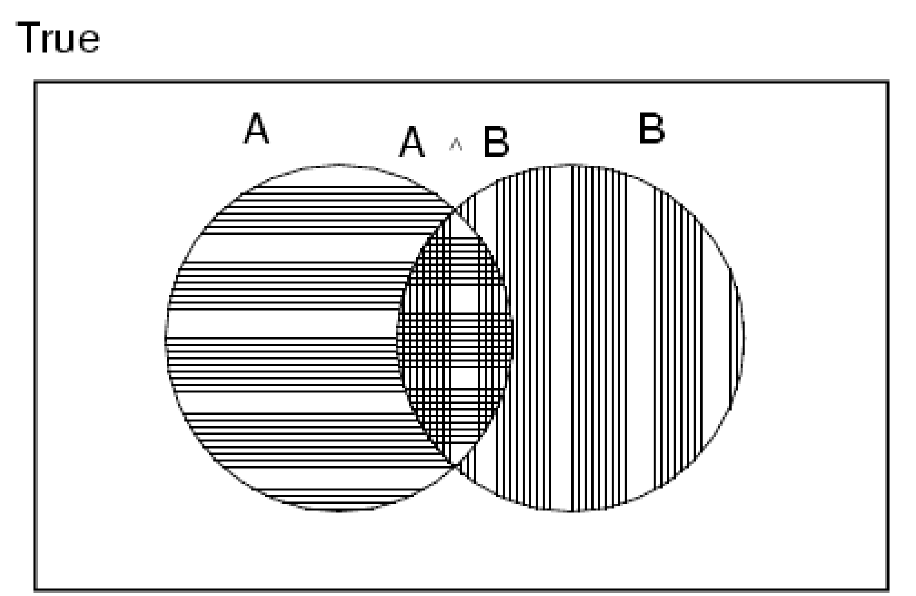
Axioms of probability

- For a sample space S with possible events, a function that associates real values with each event A is called a ***probability function*** if the following properties are satisfied:
  1. $0 \leq P(A) \leq 1$ for every A
  2. $P(S) = 1$
  3. $P(A_1 \vee A_2 \ldots \vee A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$

     if $A_1, A_2, \ldots, A_n$ *are pairwise mutually exclusive events*

Implications of axioms

- For any events ***A, B***
  - P(*A*) = 1 - P(¬*A*)
  - P(true) = 1   and   P(false) = 0
  - If A and B are mutually exclusive then P(*A* ∧ *B*) = 0
  - P(*A* ∨ *B*) = P(*A*) + P(*B*) - P(*A* ∧ *B*)



Permutations and combinations

- An **ordered** sequence of k objects taken from a set of n distinct objects without replacement, is called a **permutation** of size k
  - The number of permutations of size k that can be constructed from the n objects is:

$$P_{k,n} = \frac{n!}{(n-k)!}$$

- An **unordered** sequence of k objects taken from a set of n distinct objects without replacement, is called a **combination** of size k
  - The number of combinations of size k that can be constructed from the n objects is:

$$C_{k,n} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

Example

- An urn contains ten balls, six of which are red and four of which are white.
  Five balls are drawn at random (without replacement).
  What is the probability of drawing three red and two white balls?

$$\frac{C_{3,6} \cdot C_{2,4}}{C_{5,10}} = \frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{5}} = \frac{6!}{3!3!}\frac{4!}{2!2!}\frac{5!5!}{10!}$$

- An urn contains five balls, numbered from 1 to 5. Three balls are drawn at random. What is the probability that we draw the sequence 3, 4, 1?

$$\frac{1}{P_{3,5}} = \frac{(5-3)!}{5!}$$

Joint probability

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables:
  E.g., P(Weather, Warning) = a 4 × 2 matrix of values:

	sunny	rainy	cloudy	snow
$\text{warning} = Y$	0.005	0.08	0.02	0.02
$\text{warning} = N$	0.415	0.12	0.31	0.03

- Every question about events can be answered by the joint distribution

Under what weather condition, is there most likely to have a warning?

If there is a warning, what is the most likely weather?

Conditional probability

- **Conditional** (or posterior) probability:
  - e.g., P(warning=Y | snow=T) = 0.4
  - Complete conditional distributions specify conditional probability for all possible combinations of a set of RVs:
    P(warning | snow) =
    { P(warning = Y | snow = T),
    P(warning = N | snow = T), }
    { P(warning = Y | snow = F),
    P(warning = N | snow = F) }
- If we know more, then we can update the probability by conditioning on more evidence
  - e.g., if Windy is also given then P(warning | snow, windy) = 0.5

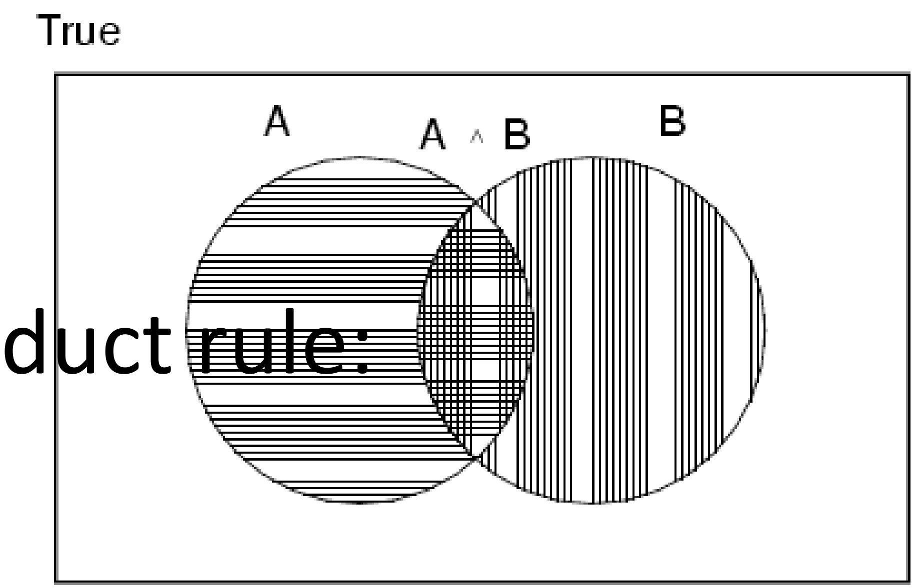
Conditional probability

- Definition of conditional probability:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \quad \text{if } P(B) > 0$$

- **Product rule** gives an alternative formulation:

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



- **Bayes rule** uses the product rule:

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

Example

- Conditional probability:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \quad \text{if } P(B) > 0$$

- Example: What is P(sunny | warning = Y)?

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Conditional probability

- **Chain rule** is derived by successive application of product rule:

$$
\begin{aligned}
P(X_1,\ldots,X_n) &= P(X_n|X_1,\ldots,X_{n-1})P(X_1,\ldots,X_{n-1}) \\
&= P(X_n|X_1,\ldots,X_{n-1})P(X_{n-1}|X_1,\ldots,X_{n-2})P(X_1,\ldots,X_{n-2}) \\
&= \ldots \\
&= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1})
\end{aligned}
$$

The Monty Hall Problem (1)

A popular game show is played as follows. A stage has three curtains. Behind one curtain (chosen at random by the show's host) is a brand-new car. Behind each of the other two curtains is a goat. The contestant chooses one of the curtains. The host then opens one of the other curtains, exposing a goat. The contestant is now given the opportunity to switch to the other unopened curtain, or keep the one that he originally chose. Should he switch?

The Monty Hall Problem (2)

- Answer: Yes.
  - Probability of winning if not switching: 1/3
  - Probability of winning if switching: 1-1/3=2/3
- Variant: Suppose there are 10 curtains with 1 car and 9 goats. After the initial selection, the host reveals 5 of the goats. What's the probability of winning if the contestant switches?

The Monty Hall Problem (3)

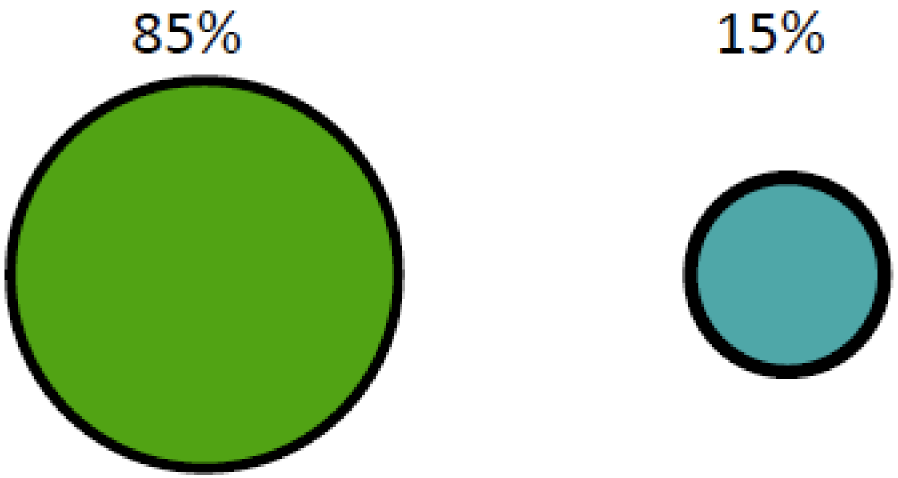
- Variant: Suppose there are 10 curtains with 1 car and 9 goats. After the initial selection, the host reveals 5 of the goats. What's the probability of winning if the contestant switches?
  - Now switching has winning the probability 1/10.
  - The total probability of winning of switching to any of the 4 curtains is 9/10.
  - These 4 curtains are equally likely, thus each has probability 9/40.

The Base Rate Fallacy

- Taxi-cab problem *(Tversky & Kahneman '72)*
  - 85% of the cabs are Green
  - 15% of the cabs are Blue
  - An accident eyewitness reports a Blue cab
  - But she is wrong 20% of the time.
- What is the probability that the cab is Blue?
  - Participants tend to overestimate probability, most answer 80%
  - They ignore baseline prior probability of blue cabs.

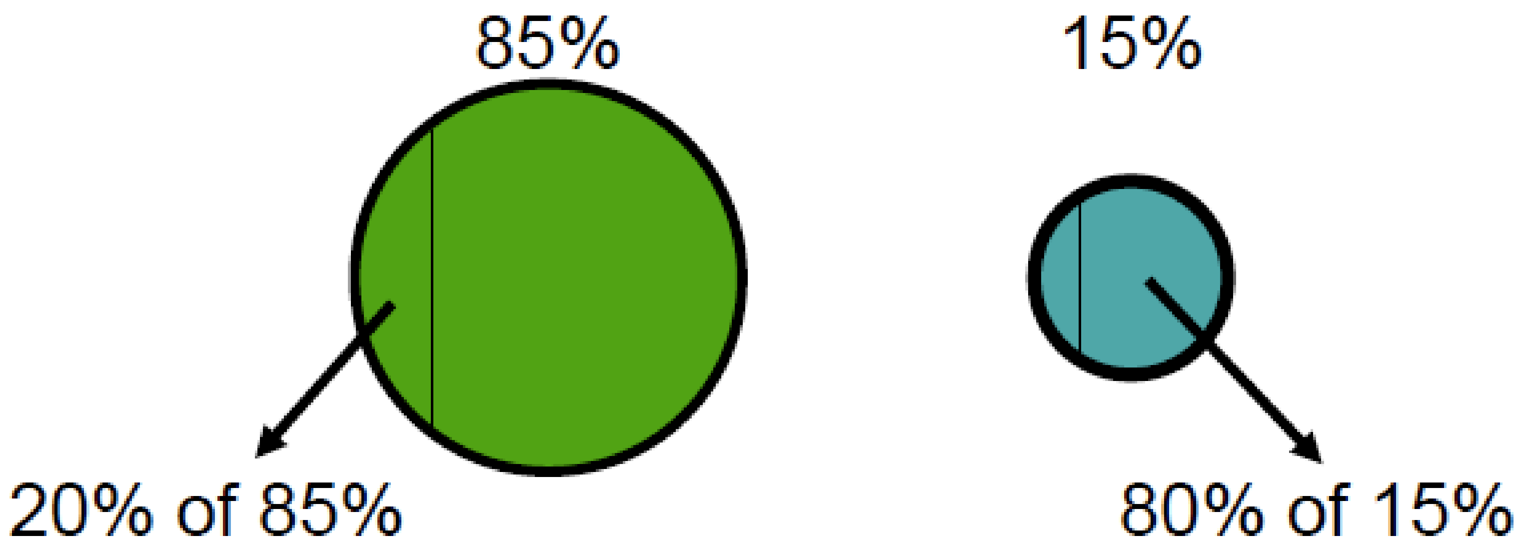
A priori (beforehand)

$$P(green) = 0.85$$
$$P(blue) = 0.15$$



$$P(seeBlue|blue) = 0.80$$
$$P(seeBlue|green) = 0.20$$

After accident (only cars reported as being blue)



Base Rate Fallacy

- How to compute probability

$$
\begin{aligned}
P(blue|seeBlue) &= \frac{P(blue \wedge seeBlue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue)} \\
&= \frac{P(seeBlue|blue)P(blue)}{P(seeBlue|blue)P(blue) + P(seeBlue|green)P(green)} \\
&= \frac{0.80 \cdot 0.15}{(0.80 \cdot 0.15) + (0.20 \cdot 0.85)} \\
&= 0.41
\end{aligned}
$$

Most people answered 80%

Medical Test

- In the 1980's in the US, a HIV test was used that had the following properties:
  There were 4% false positives
  There were 100% true positives
- About 0.4% of the male population was HIV positive
- If a man tested HIV positive, what is the probability he is actually HIV positive?

Representation

- P(positive | no HIV) = .04 (4% false positives)
- P(positive | HIV) = 1 (100% true positives)
- P(HIV) = .004 (0.4% HIV positive rate)

- want: P(HIV | positive) = ?

	HIV	no HIV
Positive	P(positive | HIV)P(HIV)	P(positive | noHIV)P(noHIV)
Negative	P(negative | HIV)P(HIV)	P(negative | noHIV)P(noHIV)

Representation

- P(positive | no HIV) = .04 (4% false positives)
- P(positive | HIV) = 1 (100% true positives)
- P(HIV) = .004 (0.4% HIV positive rate)

- want: P(HIV | positive) = ?

	HIV	**no HIV**
Positive	P(positive | HIV)P(HIV)= (1)(.004) = .004	P(positive | noHIV)P(noHIV)= (.04)(.996) = .03984
Negative	P(negative | HIV)P(HIV)= (0)(.004) = 0	P(negative | noHIV)P(noHIV)= (.96)(.996) = .95616

Solution

- P(HIV | positive) = .004 / (.004 + .03984)
  = .091

	HIV	no HIV
Positive	P(positive | HIV)P(HIV)= (1)(.004) = .004	P(positive | noHIV)P(noHIV)= (.04)(.996) = .03984
Negative	P(negative | HIV)P(HIV)= (0)(.004) = 0	P(negative | noHIV)P(noHIV)= (.96)(.996) = .95616

The Prosecutor’s Fallacy

- A person's DNA matches that of a sample found at a crime scene. The chances of a DNA match are just one in two million, so the person must be guilty beyond a reasonable doubt, right?

The Prosecutor’s Fallacy (2)

- It depends on whether DNA is the only evidence and why the DNA is collected. Consider two cases:
- There are 5 suspects, and their DNAs are collected and tested, this person is found to be a match.
- The crime scene DNA is compared with a database of DNA’s and a match is found.

The Sally Clark Case (1)

- Sally Clark, a British woman, was accused in 1998 of having killed her first child at 11 weeks of age and then her second child at 8 weeks of age. The prosecution had expert witness Sir Roy Meadow, a professor and consultant pediatrician, testify that the probability of two children in the same family dying from Sudden Infant Death Syndrome (SIDS) is about 1 in 73 million.
  - Meadow's (now discredited) law: "one sudden infant death in a family is a tragedy, two is suspicious and three is murder unless proven otherwise"

Do you think Sally Clark is guilty? What question should you ask?

The Sally Clark Case (2)

- Meadow had arrived at the 1 in 73 million figure erroneously by squaring 1 in 8500, as being the likelihood of a cot death in similar circumstances. The Royal Statistical Society later issued a statement arguing that there was "no statistical basis" for Meadow's claim, and expressing its concern at the "misuse of statistics in the courts“
- Suppose that the probability is 1 in 5 million, what else do we need to know to help judge whether Sally Clark is likely to be guity?

The Sally Clark Case (3)

- The number 1 in 5 million should not be interpreted as the probability that Sally is innocent.
- Another statistics, in the UK more than 200 babies dies of SIDS every year.
- It is found that “After a first cot death the chances of a second become greatly increased", by a dependency factor of between 5 and 10.
- Would another statistics necessary/helpful?
- Would be helpful to know the base rate of infant death due to guilty parent. The probability that double infant murder is very low.
- Prof. Ray Hill of Stanford calculated the odds ratio for double SIDS to double homicide at between 4.5:1 and 9:1

What Happened to Clark?

Clark was convicted in 1999 and sentenced to life . The convictions were upheld on appeal in 2000, but overturned in a second appeal in January 2003, after it emerged that Dr Alan Williams, the prosecution forensic pathologist who examined both of her babies, had incompetently failed to disclose microbiological reports that suggested the second of her sons had died of natural causes. She was released from prison having served more than three years of her sentence.

What Happened to Clark?

Journalist Geoffrey Wansell called Clark's experience "one of the great miscarriages of justice in modern British legal history". As a result of her case, the Attorney-General ordered a review of hundreds of other cases, and two other women had their convictions overturned.

Base Rate Fallacy in Intrusion Detection

- Assumptions in the hypothesized system:
  - Few tens of workstations running UNIX
  - Few servers running UNIX
  - Couple of dozen users
  - Capable of generating 1,000,000 audit records per day (with C2 compliant logging)
  - Single site security officer (SSO)
  - 10 audit records affected in the average intrusion
  - 2 intrusions per day => 20 records per 1,000,000 account to actual intrusions

Stefan Axelsson. The Base-Rate Fallacy and the Difficulty of Intrusion Detection. ACM TISSEC. 2000.

Base Rate Fallacy in Intrusion Detection (Continued)

- Let us use the following notation
  - I: an audit record is due to Intrusive behavior
  - A: an audit record triggers an alarm
- With the assumptions on previous slide, what is P(I) and P(¬I)?
  - $P(I) = 2 \cdot 10^{-5}$; $P(\neg I) = 1 - P(I) = 0.99998$
- We typically use the following to measure the degree of correctness of an intrusion detection system
  - Detection rate or True positive rate: $P(A|I)$
  - False alarm rate: $P(A|\neg I)$
- Suppose $P(A|I)=0.99$ and $P(A|\neg I)=0.01$. Is this good enough? What do we need to compute?
- We need to compute the Bayesian Detection Rate:
  - $P(I|A)$, $P(\neg I|\neg A)$
- What are they under the above assumption?

Base Rate Fallacy in Intrusion Detection (Continued)

- For P(A|I)=0.99, P(A| ¬I)=0.01, we get $\Pr[I|A] = \frac{\Pr[I]\Pr[A|I]}{\Pr[A]} = \frac{2\times10^{-5}\times0.99}{2\times10^{-5}\times0.99+(1-2\times10^{-5})\times0.01} \approx \frac{1.98\times10^{-5}}{1.98\times10^{-5}+0.01} \approx 0.00198$
- What values do we need For P(A|I) and P(A| ¬I) to be to have effective intrusion detection?
- For P(A|I)=1, P(A| ¬I)=$1\cdot10^{-5}$, we get P(I|A) as 0.66
- For P(A|I)=0.7, P(A| ¬I)=$1\cdot10^{-5}$, we get P(I|A) as 0.58
- Even for large detection rate, viz. P(A|I), Bayesian detection rate is dominated by the factor of false alarm rate, viz. factor of P(A| ¬I)
- P(I|A) close to 50% will induce SSO to ignore all (or most) of the alarms generated

Base Rate Fallacy in Intrusion Detection (Lessons)

- Intrusion detection is difficult in real world
- The “effectiveness” of an intrusion detection system depends not just on its ability to detect intrusive behavior but on its ability to suppress false alarms
- Comparison shows anomaly-based detection methods have larger false alarm rates than signature-based detection, but signature-based detection methods cannot provide protection against novel intrusions

Marginal probability

- **Marginal** (or unconditional) probability corresponds to belief that event will occur regardless of conditioning events
- Marginalization: $$P(A) = \sum_{b \in B} P(A, b) = \sum_{b \in B} P(A|b) P(b)$$
- Example: What is P(cloudy)?

	sunny	rainy	cloudy	snow
warning $= Y$	0.005	0.08	0.02	0.02
warning $= N$	0.415	0.12	0.31	0.03

Independence

- **A and B are independent iff:**
  - $P(A|B) = P(A)$ or $P(B|A) = P(B)$
    or $P(A, B) = P(A)\ P(B)$
  - *Knowing B tells you nothing about A*
- **Examples**
  - Coin flip 1 and coin flip 2?
  - Weather and storm warning?
  - Weather and coin flip=H?
  - Weather and election?

Conditional independence

- Two variables *A* and *B* are **conditionally** independent given *Z*
  iff for all values of *A, B, Z:*

$$P(A, B \mid Z) = P(A \mid Z)\, P(B \mid Z)$$

- ***Note***: *independence does not imply conditional independence or vice versa*

Example 1

- **Conditional independence does not imply independence**
- Gender and lung cancer are not independent

$$P(C \mid G) \neq P(C)$$

- Gender and lung cancer are conditionally independent given smoking

$$P(C \mid G, S) = P(C \mid S)$$

- Why? Because gender indicates likelihood of smoking, and smoking causes cancer

Example 2

- **Independence does not imply conditional independence**
- Sprinkler-on and raining are independent
  $$P(S \mid R) = P(S)$$
- Sprinkler-on and raining are not conditionally independent given
  grass is wet
  $$P(S \mid R, W) \neq P(S \mid R)$$
- Why? Because once we know the grass is wet, if it's not raining, then the explanation for the grass being wet has to be the sprinkler

Example

- You flip a fair coin twice
    1. The first flip is heads
    2. The second flip is tails
    3. The two flips are not the same
- Are (1) and (2): *independent*? *Conditionally independent given (3)*? *Neither*?

Probability distribution

- **Probability distribution** *(i.e., probability mass function or probability density function)* specifies the probability of observing every possible value of a random variable
- Discrete (*probability mass function)*
  - Denotes probability that *X* will take on a particular value:

$$P(X = x)$$

- Continuous (*probability density function*)
  - Probability of any particular point is 0, have to consider probability within an interval:

$$P(a < X < b) = \int_a^b p(x)dx$$

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.
- X = {0, 1, 2, 3}
- P(X=0) = 1/8; P(X=1) = 3/8; P(X=2) = 3/8; P(X=3) = 1/8
- What is the expected value of X, E[X]?

$$E[X] = (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8})$$
$$= \frac{3}{2}$$

Expectation

- Denotes the expected value or mean value of a random variable X
- Discrete

$$E[X] = \sum_x x \cdot p(x)$$

- Continuous

$$E[X] = \int_x x \cdot p(x) dx$$

- Expectation of a function

$$E[h(X)] = \sum_x h(x) \cdot p(x)$$

$$E[aX + b] = a \cdot E[X] + b$$

$$E[X + Y] = E[X] + E[Y]$$

This holds whether X and Y are independent or not!

Proof of Linearity of Expectation

We'll explicitly prove this theorem for discrete random variables X and Y. By the basic definition of expected value,

$$\begin{aligned}
E[X+Y] &= \sum_x \sum_y [(x+y) \cdot P(X=x, Y=y)] \\
&= \sum_x \sum_y [x \cdot P(X=x, Y=y)] + \sum_x \sum_y [y \cdot P(X=x, Y=y)] \\
&= \sum_x x \underbrace{\sum_y P(X=x, Y=y)}_{P(X=x)} + \sum_y y \underbrace{\sum_x P(X=x, Y=y)}_{P(Y=y)} \\
&= \sum_x x \cdot P(X=x) + \sum_y y \cdot P(Y=y) \\
&= E[X] + E[Y].
\end{aligned}$$

This result can be extended for n variables using induction.

An Example Problem

- I have 12 addressed letters to mail, and 12 corresponding pre-addressed envelopes. For some wacky reason, I decide to put the letters into the envelopes at random, one letter per envelope. What is the expected number of letters that get placed into their proper envelopes?
  - Let X_i denote the event that the i-th letter is put in the right envelope. Then $\Pr[X_i = 1] = {}^{1}/_{12}$. Thus $E\left[\sum_{i=1}^{12} X_i\right] = \sum_{i=1}^{12} E[X_i] = 12 \times {}^{1}/_{12} = 1$

Variance

- Denotes the expectation of the squared deviation of X from its mean
- Variance

$$Var(X) = E[(x - E[X])^2] \\ = E[X^2] - (E[X])^2$$

- Standard deviation

$$\sigma = \sqrt{Var(X)}$$

- Variance of a function

$$Var(aX + b) = a^2 \cdot Var(X)$$

$$Var(h(X)) = \sum_x (h(x) - E[h(x)])^2 \cdot p(x)$$

Example

- Let X be a random variable that represents the number of heads which appear when a fair coin is tossed three times.
- X = {0, 1, 2, 3}

$$E[X] = (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8})$$
$$= \frac{3}{2}$$

- What is the variance of X, Var(X)?

$$Var(X) = \left(\left[0 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right) + \left(\left[1 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[2 - \frac{3}{2}\right]^2 \cdot \frac{3}{8}\right) + \left(\left[3 - \frac{3}{2}\right]^2 \cdot \frac{1}{8}\right)$$
$$= \left(\frac{9}{4} \cdot \frac{1}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{1}{4} \cdot \frac{3}{8}\right) + \left(\frac{9}{4} \cdot \frac{1}{8}\right)$$
$$= \frac{3}{4}$$

Common distributions

- Bernoulli
- Binomial
- Multinomial
- Poisson
- Normal

Bernoulli

- Binary variable (0/1) that takes the value of 1 with probability p
  - E.g., Outcome of a fair coin toss is Bernoulli with *p=0.5*

$$P(x) = p^x(1-p)^{1-x}$$

$$E[X] = 1(p) + 0(1-p) = p$$

$$\begin{aligned} Var(X) &= E[X]^2 - (E[X])^2 \\ &= 1^2(p) + 0^2(1-p) - p^2 \\ &= p(1-p) \end{aligned}$$

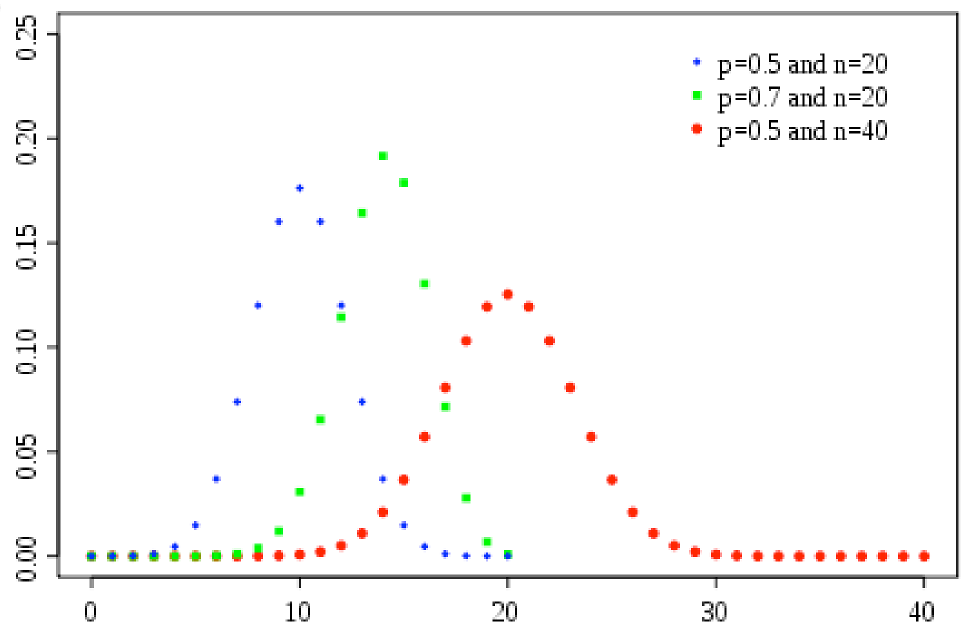
Binomial

- Describes the number of successful outcomes in *n* independent Bernoulli(p) trials
  - E.g., Number of heads in a sequence of 10 tosses of a fair coin is Binomial with *n=10* and *p=0.5*

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$E[X] = np$$

$$Var[X] = np(1-p)$$



Multinomial

- Generalization of binomial to *k* possible outcomes; outcome *i* has probability p_i of occurring
  - E.g., Number of {outs, singles, doubles, triples, homeruns} in a sequence of 10 times at bat is Multinomial
- Let X_i denote the number of times the *i*-th outcome occurs in *n* trials:

$$P(x_1, ...x_k) = \binom{n}{x_1, ...x_k} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}$$

$$E[X_i] = np_i$$

$$Var(X_i) = np_i(1 - p_i)$$

Normal (Gaussian)

- Important distribution gives well-known bell shape

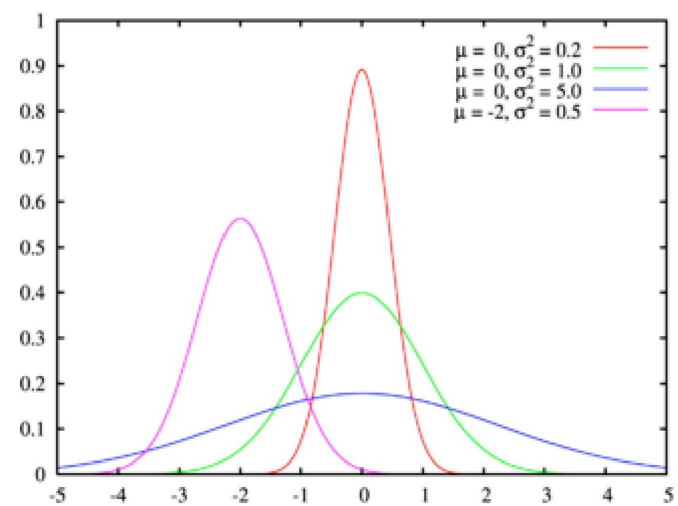
$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E[X] = \mu$$

$$Var(X) = \sigma^2$$

- **Central limit theorem:**

Distribution of the mean of n samples becomes normally distributed as n ↑, regardless of the distribution of the underlying population



Multivariate RV

- A multivariate random variable ***X*** is a set $X_1, X_2, \ldots, X_p$ of random variables
- **Joint** density function: $P(\mathbf{x})=P(x_1, x_2, \ldots, x_p)$
- **Marginal** density function: the density of any subset of the complete set of variables, e.g.,:

$$P(x_1) = \sum_{x_2, x_3} p(x_1, x_2, x_3)$$

Conditional density function: the density of a subset conditioned on particular values of the others, e.g.,:

$$P(x_1 | x_2, x_3) = \frac{p(x_1, x_2, x_3)}{p(x_2, x_3)}$$

Frequentist view of Probability

- Dominant perspective for last century
- Probability is an **objective** concept
  - Defined as the frequency of an event occurring under repeated trials in “same” situation
  - E.g., number of heads in repeated coin tosses
- Restricts application of probability to repeatable events

Bayesian view

- Increasing importance over last decade
  - Due to increase in computational power that facilitates previously intractable calculations
- Probability is a **subjective** concept
  - Defined as individual degree-of-belief that event will occur
  - E.g., belief that we will have another snow storm tomorrow
- Begin with prior belief estimates and update those by conditioning on observed data

Calculating probabilities: Bayesian

- *Begin with prior belief estimates*: P(A)
  - E.g., After the Seahawks won their conference, Vegas casinos believed the Seahawks were likely to win the Superbowl over the Patriots:
    P(S wins)=0.525, P(P wins)=0.475
- Observe data
  - But then Vegas observed a heavy majority of the betters (80%) chose the Patriots, which is unlikely given their current belief
- *Update belief by conditioning on observed data*
  P(A|data) = P(data|A) P(A) / P(data)
  - So they updated their belief to increase the the Patriots's chance of a win:
    P(S wins | betting) = P(betting | S wins) P(S wins) / P(betting) = 0.50
- Even when the same data is observed, if people have different priors, they can end up with different posterior probability estimates P(A|data)