BBM 205 Discrete Mathematics
Hacettepe University
http://web.cs.hacettepe.edu.tr/∼bbm205

Lecture 10: Bayes' Theorem, Expected Value and Variance

Lecturer: Lale Özkahya

Resources:
Kenneth Rosen, "Discrete Mathematics and App."
http://www.inf.ed.ac.uk/teaching/courses/dmmr

Reverend Thomas Bayes (1701-1761),
studied logic and theology as an undergraduate student
at the University of Edinburgh from 1719-1722.

Bayes' Theorem

Bayes Theorem

Let A and B be two events from a (countable) sample space Ω, and $P : \Omega \rightarrow [0, 1]$ a probability distribution on Ω, such that $0 < P(A) < 1$, and $P(B) > 0$. Then

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

This may at first look like an obscure equation,
but as we shall see, it is useful....

Proof of Bayes' Theorem:

Let A and B be events such that $0 < P(A) < 1$ and $P(B) > 0$.

By definition, $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$. So: $P(A \cap B) = P(A \mid B)P(B)$.

Likewise, $P(B \cap A) = P(B \mid A)P(A)$.

Likewise, $P(B \cap \overline{A}) = P(B \mid \overline{A})P(\overline{A})$. (Note that $P(\overline{A}) > 0$.)

Note that $P(A \mid B)P(B) = P(A \cap B) = P(B \mid A)P(A)$. So,

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

Furthermore,

$$\begin{aligned} P(B) &= P((B \cap A) \cup (B \cap \overline{A})) = P(B \cap A) + P(B \cap \overline{A}) \\ &= P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A}) \end{aligned}$$

So: $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})}. \quad \square$$

Using Bayes' Theorem

Problem: There are two boxes, Box B_1 and Box B_2.

Box B_1 contains 2 red balls and 8 blue balls.
Box B_2 contains 7 red balls and 3 blue balls.

Suppose Jane first randomly chooses one of two boxes B_1 and B_2, with equal probability, $1/2$, of choosing each.

Suppose Jane then randomly picks one ball out of the box she has chosen (without telling you which box she had chosen), and shows you the ball she picked.

Suppose you only see that the ball Jane picked is red.

Question: Given this information, what is the probability that Jane chose box B_1?

Using Bayes' Theorem, continued

Answer: The underlying sample space, Ω, is:

$$\Omega = \{(a, b) \mid a \in \{1, 2\}, b \in \{\text{red}, \text{blue}\}\}$$

Let $F = \{(a, b) \in \Omega \mid a = 1\}$ be the event that box B_1 was chosen. Thus, $\overline{F} = \Omega - F$ is the event that box B_2 was chosen.

Let $E = \{(a, b) \in \Omega \mid b = \text{red}\}$ be the event that a red ball was picked. Thus, $\overline{E}$ is the event that a blue ball was picked.

We are interested in computing the probability $P(F \mid E)$.

We know that $P(E \mid F) = \frac{2}{10}$ and $P(E \mid \overline{F}) = \frac{7}{10}$.

We also know that: $P(F) = 1/2$ and $P(\overline{F}) = 1/2$.

Can we compute $P(F \mid E)$ based on this? Yes, using Bayes'.

Using Bayes' Theorem, continued

Note that, $0 < P(F) < 1$, and $P(E) > 0$.

By Bayes' Theorem:

$$
\begin{aligned}
P(F \mid E) &= \frac{P(E \mid F)P(F)}{P(E \mid F)P(F) + P(E \mid \overline{F})P(\overline{F})} \\
&= \frac{(2/10) * (1/2)}{(2/10) * (1/2) + (7/10) * (1/2)} \\
&= \frac{2/20}{2/20 + 7/20} = \frac{2}{9}. \quad \square
\end{aligned}
$$

Note that, without the information that a red ball was picked, the probability that Jane chose Box B_1 is $P(F) = 1/2$.

But given the information, E, that a red ball was picked, the probability becomes much less, changing to $P(F \mid E) = 2/9$.

More on using Bayes' Theorem: Baysian Spam Filters

Problem: Suppose it has been observed empirically that the word "Congratulations" occurs in 1 out of 10 spam emails, but that "Congratulations" only occurs in 1 out of 1000 non-spam emails. Suppose it has also been observed empirically that about 4 out of 10 emails are spam.

In Bayesian Spam Fitering, these empirical probabilities are interpreted as genuine probabilities in order to help estimate the probability that a incoming email is spam.

Suppose we get a new email that contains "Congratulations".

Let C be the event that a new email contains "Congratulations". Let S be the event that a new email is spam.

We have observed C. We want to know $P(S \mid C)$.

Bayesian spam filtering example, continued

Bayesian solution: By Bayes' Theorem:

$$P(S \mid C) = \frac{P(C \mid S)P(S)}{P(C \mid S)P(S) + P(C \mid \overline{S})P(\overline{S})}$$

From the "empirical probabilities", we get the estimates:

$P(C \mid S) \approx 1/10; \quad P(C \mid \overline{S}) \approx 1/1000;$

$P(S) \approx 4/10; \quad P(\overline{S}) \approx 6/10.$

So, we estimate that:

$$\begin{aligned} P(S \mid C) &\approx \frac{(1/10)(4/10)}{(1/10)(4/10) + (1/1000) * (6/10)} \\ &\approx \frac{.04}{.0406} \approx 0.985 \end{aligned}$$

So, with "high probability", such an email is spam. (However, much caution is needed when interpreting such "probabilities".)

Generalized Bayes' Theorem

Suppose that $E, F_1, \ldots, F_n$ are events from sample space Ω, and that $P : \Omega \to [0,1]$ is a probability distribution on Ω. Suppose that $\cup_{i=1}^{n} F_j = \Omega$, and that $F_i \cap F_j = \emptyset$ for all $i \neq j$.
Suppose $P(E) > 0$, and $P(F_j) > 0$ for all j. Then for all j:

$$P(F_j \mid E) = \frac{P(E \mid F_j)P(F_j)}{\sum_{i=1}^{n} P(E \mid F_i)P(F_i)}$$

Suppose Jane first randomly chooses a box from among n different boxes, $B_1, \ldots, B_n$, and then randomly picks a coloured ball out of the box she chose. (Each Box may have different numbers of balls of each colour.)
We can use the *Generalized Bayes' Theorem* to calculate the probability that Jane chose box B_j (event F_j), given that the colour of the ball that Jane picked is red (event E).

Proof of Generalized Bayes' Theorem: Very similar to the proof of Bayes' Theorem. Observe that:

$$P(F_j \mid E) = \frac{P(F_j \cap E)}{P(E)} = \frac{P(E \mid F_j)P(F_j)}{P(E)}$$

So, we only need to show that $P(E) = \sum_{i=1}^{n} P(E \mid F_i)P(F_i)$.
But since $\bigcup_i F_i = \Omega$, and since $F_i \cap F_j = \emptyset$ for all $i \neq j$:

$$\begin{aligned} P(E) &= P(\bigcup_i (E \cap F_i)) \\ &= \sum_{i=1}^{n} P(E \cap F_i) \quad \text{(because } F_i\text{'s are disjoint)} \\ &= \sum_{i=1}^{n} P(E \mid F_i)P(F_i). \quad \square \end{aligned}$$

Expected Value (Expectation) of a Random Variable

Recall: A **random variable** (**r.v.**), is a function $X : \Omega \to \mathbb{R}$, that assigns a real value to each outcome in a sample space Ω.

The **expected value**, or **expectation**, or mean, of a random variable $X : \Omega \to \mathbb{R}$, denoted by $E(X)$, is defined by:

$$E(X) = \sum_{s \in \Omega} P(s)X(s)$$

Here $P : \Omega \to [0, 1]$ is the underlying probability distribution on Ω.

Question: Let X be the r.v. outputing the number that comes up when a fair die is rolled. What is the expected value, $E(X)$, of X?

Expected Value (Expectation) of a Random Variable

Recall: A **random variable** (**r.v.**), is a function $X : \Omega \to \mathbb{R}$, that assigns a real value to each outcome in a sample space Ω.

The **expected value**, or **expectation**, or mean, of a random variable $X : \Omega \to \mathbb{R}$, denoted by $E(X)$, is defined by:

$$E(X) = \sum_{s \in \Omega} P(s)X(s)$$

Here $P : \Omega \to [0, 1]$ is the underlying probability distribution on Ω.

Question: Let X be the r.v. outputing the number that comes up when a fair die is rolled. What is the expected value, $E(X)$, of X?

Answer:

$$E(X) = \sum_{i=1}^{6} \frac{1}{6} \cdot i = \frac{21}{6} = \frac{7}{2}. \quad \square$$

A bad way to calculate expectation

The definition of expectation, $E(X) = \sum_{s \in \Omega} P(s)X(s)$, can be used directly to calculate $E(X)$. But sometimes this is horribly inefficient.

Example: Suppose that a biased coin, which comes up heads with probability p each time, is flipped 11 times consecutively.

Question: What is the expected # of heads?

A bad way to calculate expectation

The definition of expectation, $E(X) = \sum_{s \in \Omega} P(s)X(s)$, can be used directly to calculate $E(X)$. But sometimes this is horribly inefficient.

Example: Suppose that a biased coin, which comes up heads with probability p each time, is flipped 11 times consecutively.

Question: What is the expected # of heads?

Bad way to answer this: Let's try to use the definition of $E(X)$ directly, with $\Omega = \{H, T\}^{11}$. Note that $|\Omega| = 2^{11} = 2048$.
So, the sum $\sum_{s \in \Omega} P(s)X(s)$ has 2048 terms!

This is clearly not a practical way to compute $E(X)$.

Is there a better way? Yes.

Better expression for the expectation

Recall $P(X = r)$ denotes the probability $P(\{s \in \Omega \mid X(s) = r\})$.
Recall that for a function $X : \Omega \to \mathbb{R}$,

$$range(X) = \{r \in \mathbb{R} \mid \exists s \in \Omega \text{ such that } X(s) = r\}$$

Theorem: For a random variable $X : \Omega \to \mathbb{R}$,

$$E(X) = \sum_{r \in range(X)} P(X = r) \cdot r$$

Better expression for the expectation

Recall $P(X = r)$ denotes the probability $P(\{s \in \Omega \mid X(s) = r\})$.
Recall that for a function $X : \Omega \to \mathbb{R}$,

$$range(X) = \{r \in \mathbb{R} \mid \exists s \in \Omega \text{ such that } X(s) = r\}$$

Theorem: For a random variable $X : \Omega \to \mathbb{R}$,

$$E(X) = \sum_{r \in range(X)} P(X = r) \cdot r$$

Proof: $E(X) = \sum_{s \in \Omega} P(s)X(s)$, but for each $r \in range(X)$, if we sum all terms $P(s)X(s)$ such that $X(s) = r$, we get $P(X = r) \cdot r$ as their sum. So, summing over all $r \in range(X)$ we get $E(X) = \sum_{r \in range(X)} P(X = r) \cdot r$. $\square$

So, if $|range(X)|$ is small, and if we can compute $P(X = r)$, then we need to sum a lot fewer terms to calculate $E(X)$.

Expected # of successes in n Bernoulli trials

Theorem: The expected # of successes in n (independent) Bernoulli trials, with probability p of success in each, is np.

Note: We'll see later that we do not need independence for this.

First, a proof which uses mutual independence: For $\Omega = \{H, T\}^n$, let $X : \Omega \to \mathbb{N}$ count the number of successes in n Bernoulli trials. Let $q = (1 - p)$. Then...

$$\begin{aligned} E(X) &= \sum_{k=0}^{n} P(X = k) \cdot k \\ &= \sum_{k=1}^{n} \binom{n}{k} p^k q^{n-k} \cdot k \end{aligned}$$

The second equality holds because, assuming mutual independence, $P(X = k)$ is the binomial distribution $b(k; n, p)$.

first proof continued

$$
\begin{aligned}
E(X) = \sum_{k=0}^{n} P(X=k)\cdot k = \sum_{k=1}^{n} \binom{n}{k} p^k q^{n-k} \cdot k = \\
&= \sum_{k=1}^{n} \frac{n!}{k!(n-k)!} p^k q^{n-k} \cdot k = \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k q^{n-k} \\
&= \sum_{k=1}^{n} n \cdot \frac{(n-1)!}{(k-1)!(n-k)!} p^k q^{n-k} = n \sum_{k=1}^{n} \binom{n-1}{k-1} p^k q^{n-k} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{n-k} = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j q^{n-1-j} \\
&= np(p+q)^{n-1} \\
&= np\,. \quad \square
\end{aligned}
$$

We will soon see this was an unnecessarily complicated proof.

Expectation of a geometrically distributed r.v.

Question: A coin comes up heads with probability $p > 0$ each time it is flipped. The coin is flipped repeatedly until it comes up heads. What is the expected number of times it is flipped?

Note: This simply asks: "What is the expected value $E(X)$ of a geometrically distributed random variable with parameter p?"

Answer: $\Omega = \{H, TH, TTH, \ldots\}$, and $P(T^{k-1}H) = (1-p)^{k-1}p$. And clearly $X(T^{k-1}H) = k$. Thus $E(X) = \sum_{s \in \Omega} P(s)X(s) =$

$$E(X) = \sum_{k=1}^{\infty}(1-p)^{k-1}p \cdot k = p\sum_{k=1}^{\infty}k(1-p)^{k-1} = p \cdot \frac{1}{p^2} = \frac{1}{p}.$$

This is because: $\sum_{k=1}^{\infty} k \cdot x^{k-1} = \frac{1}{(1-x)^2}$, for $|x| < 1$. $\square$

Example: If $p = 1/4$, then the expected number of coin tosses before we see Heads for the first time is 4.

Linearity of Expectation (VERY IMPORTANT)

Theorem (Linearity of Expectation): For any random variables $X, X_1, \ldots, X_n$ on Ω, $$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + \ldots + E(X_n).$$

Furthermore, for any $a, b \in \mathbb{R}$,
$$E(a\,X + b) = a\,E(X) + b.$$

(In other words, the expectation function is a **linear function**.)

Linearity of Expectation (VERY IMPORTANT)

Theorem (Linearity of Expectation): For any random variables $X, X_1, \ldots, X_n$ on Ω,
$$E(X_1 + X_2 + \ldots + X_n) = E(X_1) + \ldots + E(X_n).$$
Furthermore, for any $a, b \in \mathbb{R}$,
$$E(a\,X + b) = a\,E(X) + b.$$
(In other words, the expectation function is a **linear function**.)

Proof:

$$E(\sum_{i=1}^{n} X_i) = \sum_{s\in\Omega} P(s) \sum_{i=1}^{n} X_i(s) = \sum_{i=1}^{n} \sum_{s\in\Omega} P(s) X_i(s) = \sum_{i=1}^{n} E(X_i).$$

$$E(aX + b) = \sum_{s\in\Omega} P(s)(aX(s) + b) = (a \sum_{s\in\Omega} P(s)X(s)) + b \sum_{s\in\Omega} P(s)$$
$$= aE(X) + b. \quad \square$$

Using linearity of expectation

Theorem: The expected # of successes in n (not necessarily independent) Bernoulli trials, with probability p of success in each trial, is np.

Using linearity of expectation

Theorem: The expected # of successes in n (not necessarily independent) Bernoulli trials, with probability p of success in each trial, is np.

Easy proof, via linearity of expectation: For $\Omega = \{H, T\}^n$, let X be the r.v. counting the expected number of successes, and for each i, let $X_i : \Omega \to \mathbb{R}$ be the binary r.v. defined by:

$$X_i((s_1, \ldots, s_n)) = \begin{cases} 1 & \text{if } s_i = H \\ 0 & \text{if } s_i = T \end{cases}$$

Note that $E(X_i) = p \cdot 1 + (1-p) \cdot 0 = p$, for all $i \in \{1, \ldots, n\}$.

Also, clearly, $X = X_1 + X_2 + \ldots + X_n$, so:

$$E(X) = E(X_1 + \ldots + X_n) = \sum_{i=1}^{n} E(X_i) = np. \quad \square$$

Note: this holds even if the n coin tosses are totally correlated.

Using linearity of expectation, continued

Hatcheck problem: At a restaurant, the hat-check person forgets to put claim numbers on hats.
n customers check their hats in, and they each get a random hat back when they leave the restuarant.
What is the expected number, $E(X)$, of people who get their correct hat back?

Answer: Let X_i be the r.v. that is 1 if the i'th customer gets their hat back, and 0 otherwise.
Clearly, $E(X) = E(\sum_i X_i)$.
Furthermore, $E(X_i) = P(i\text{'th person gets its hat back}) = 1/n$.
Thus, $E(X) = n \cdot (1/n) = 1$. $\square$

This would be much harder to prove without using the linearity of expectation.
Note: $E(X)$ doesn't even depend on n in this case.

Independence of Random Variables

Definition: Two random variables, X and Y, are called **independent** if for all $r_1, r_2 \in \mathbb{R}$:

$$P(X = r_1 \text{ and } Y = r_2) = P(X = r_1) \cdot P(Y = r_2)$$

Example: Two die are rolled. Let X_1 be the number that comes up on die 1, and let X_2 be the number that comes up on die 2. Then X_1 and X_2 are independent r.v.'s.

Theorem: If X and Y are independent random variables on the same space Ω. Then

$$E(XY) = E(X)E(Y)$$

We will not prove this in class. (The proof is a simple re-arrangement of the sums in the definition of expectation. See Rosen's book for a proof.)

Variance

The "variance" and "standard deviation" of a r.v., X, give us ways to measure (roughly) *"on average, how far off the value of the r.v. is from its expectation"*.

Variance and Standard Deviation

Definition: For a random variable X on a sample space Ω, the **variance** of X, denoted by $V(X)$, is defined by:

$$V(X) = E((X - E(X))^2) = \sum_{s \in \Omega} (X(s) - E(X))^2 P(s)$$

The **standard deviation** of X, denoted $\sigma(X)$, is defined by

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

Example, and a useful identity for variance

Example: Consider the r.v., X, such that $P(X = 0) = 1$, and the r.v. Y, such that $P(Y = -10) = P(Y = 10) = 1/2$.
Then $E(X) = E(Y) = 0$, but $V(X) = 0 = \sigma(X)$, whereas $V(Y) = 100$ and $\sigma(Y) = 10$. $\square$

Theorem: For any random variable X,

$$V(X) = E(X^2) - E(X)^2$$

Example, and a useful identity for variance

Example: Consider the r.v., X, such that $P(X = 0) = 1$, and the r.v. Y, such that $P(Y = -10) = P(Y = 10) = 1/2$.
Then $E(X) = E(Y) = 0$, but $V(X) = 0 = \sigma(X)$, whereas $V(Y) = 100$ and $\sigma(Y) = 10$. $\square$

Theorem: For any random variable X,

$$V(X) = E(X^2) - E(X)^2$$

Proof:

$$\begin{aligned} V(X) &= E((X - E(X))^2) \\ &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \quad \square \end{aligned}$$