**Tutorial 13.**

# 1 Conditional Expectation

We can define *conditional expectation*, or the expectation of a random variable, conditioned on a particular event as follows:

**Definition 1** *The conditional expectation of a random variable $R$ given an event $A$ is defined as*

$$E[R \mid A] = \sum_{r \in Range(R)} r \cdot Pr(R = r \mid A).$$

*Example:* Let $R$ be the outcome of a die. What is the expected value of $R$ given that $R \leq 4$ ?

$$E[R \mid R \leq 4] = \sum_{r=1}^{4} r Pr(R = r \mid R \leq 4)$$

But $Pr(R = r \mid R \leq 4) = \frac{1}{4}$. So,

$$E[R \mid R \leq 4] = \frac{1}{4} \sum_{r=1}^{4} r = 2.5$$

One use of conditional expectation is to compute an expectation.
The following theorem breaks down the expectation into conditional expectations.

**Theorem 1** *Let $A_1, A_2, \ldots, A_n$ be a set of disjoint events whose union is the the sample space. Then for any random variable $R$,*

$$E[R] = \sum_{i=1}^{n} E[R \mid A_i] Pr(A_i)$$

Notice that this is analogous to the theorem of total probability:

**Theorem 2** *Let $A_1, A_2, \ldots, A_n$ be a set of disjoint events whose union is the the sample space. Then for any event $B$,*

$$Pr(B) = \sum_{i=1}^{n} Pr(B \cap A_i).$$

*Proof of Theorem 1.*

$$\begin{aligned}
E[R] &= \sum_r r \cdot Pr(R = r) \\
&= \sum_r r \cdot \left( \sum_{i=1}^{n} Pr(R = r \cap A_i) \right) \\
&= \sum_r r \cdot \left( \sum_{i=1}^{n} Pr(R = r \mid A_i) Pr(A_i) \right) \\
&= \sum_{i=1}^{n} \left( \sum_r r \cdot Pr(R = r \mid A_i) \right) Pr(A_i) \\
&= \sum_i E[R \mid A_i] Pr(A_i)
\end{aligned}$$

The first step uses the theorem of total probability with $B$ being the event $R = r$. The second step is by the definition of conditional probability. ■

*Example:* (Assembly plants)

Suppose there are two computer assembly plants $F$ and $S$. Plant $F$ is fast but sloppy: It makes 70% of the computers, but a computer from $F$ has a 2% probability of breaking down on any given day of use.

Plant $S$ is slow and careful: It makes 30% of the computers, but the breakdown probability is only 1%.

These computers get randomly mixed, so I buy one at random from the store. What is the expected number of days my computer works?

Let $D$ be the random variable giving the number of days my computer works. Let $I_S$ be the event that I bought the computer from $S$, and let $I_F$ be that I bought it from $F$. We solve this problem using conditional expectations. We calculate separately the conditional expectation of the number of working days conditioned on $I_S$ and on $I_F$, then combine using the conditional expectation theorem.

The expected number of working days for a computer from $S$ with probability of failure $p = 0.01$ is just

$$E[D \mid I_S] = \frac{1}{p} - 1 = 99$$

Note that mean time to failure is $\frac{1}{p}$, hence the number of working days is $\frac{1}{p} - 1$. Similarly, the expected number of working days for a computer from $F$ is

$$E[D \mid I_F] = \frac{1}{.02} - 1 = 49$$

.

Now to calculate the expected number of working days for the computer when we don't know which factory it came from, we use the conditional expectation theorem:

$$E[D] = E[D|I_S]Pr(I_S) + E[D|I_F]Pr(I_F) = 99(.30) + 49(.70) = 64$$

## 2 Covariance

We saw in class that the variance of a random variable is a measure of how it deviates from the mean. *Covariance* is another statistical measure. It quantifies the similarity of the distributions of two random variables.

**Definition 2** *The covariance of random variables $R_1$ and $R_2$ is*

$$Cov(R_1, R_2) = E(R_1 R_2) - E(R_1)E(R_2).$$

If $R_1$ and $R_2$ are positively correlated then their covariance is positive and if they are negatively correlated their covariance is negative. Their covariance is zero if they are uncorrelated.

*Examples:* Let $R_1$ and $R_2$ be uniform in the range $[1, 6]$.

1. $R_1 = R_2$. In this case,

$$Cov(R_1, R_2) = E(R_1^2) - E(R_1)^2 = Var(R_1) = 2\frac{11}{12}.$$

   The variables are positively correlated. This is not a surprise — they are always equal.

2. $R_2 = 7 - R_1$. Now,

$$\begin{aligned} Cov(R_1, R_2) = E(R_1(7-R_1)) - E(R_1)E(7-R_1) &= 7E(R_1) - E(R_1^2) - 7E(R_1) + E(R_1)^2 \\ &= -E(R_1^2) + E(R_1)^2 \\ &= -Var(R_1) = -2\frac{11}{12} \end{aligned}$$

   The variables are negatively correlated, i.e., when one is high the other is low. This is reflected by the negative covariance.

3. $R_1$ are $R_2$ are independent (E.g. consecutive coin tosses). Then

$$Cov(R_1, R_2) = E(R_1 R_2) - E(R_1)E(R_2) = E(R_1)E(R_2) - E(R_1)E(R_2) = 0.$$

   The variables are uncorrelated.

So if the variables are independent then, indeed their covariance is zero. What about the converse? If the covariance is zero, are the variables independent? The answer is no. E.g. let $R_1, R_2$ be in the range $[1, 3]$ such that $Pr(R_1 = i \cap R_2 = j) = \frac{1}{8}$ for all $(i, j)$ except $(2, 2)$ and $Pr(R_1 = R_2 = 2) = 0$. We can check that this gives valid probability distributions for $R_1$ and $R_2$.

$$Pr(R_1 = 1) = \frac{3}{8}$$

$$Pr(R_1 = 2) = \frac{2}{8}$$
$$Pr(R_1 = 3) = \frac{3}{8}$$

Which add up to 1.

The distribution of $R_1$ is the same as the distribution of $R_2$ but since $Pr(R_1 = 2 \mid R_2 = 2) = 0 \neq Pr(R_1 = 2)$, $R_1$ and $R_2$ are not independent. However,

$$\begin{aligned}
Cov(R_1, R_2) &= E(R_1 R_2) - E(R_1)E(R_2) \\
&= \sum_{i,j \in [1,3]} ij Pr(R_1 = i \cap R_2 = j) - (\sum_{i \in [1,3]} i Pr(R_1 = i))^2 \\
&= \frac{1}{8}(1 + 2 + 3 + 2 + 6 + 3 + 6 + 9) - (\frac{3}{8}(1 + 3) + \frac{2}{8}2)^2 \\
&= 4 - 4 = 0
\end{aligned}$$

**Theorem 3** *For random variables $R_1$ and $R_2$,*

$$Var(R_1 + R_2) = Var(R_1) + Var(R_2) + 2Cov(R_1, R_2)$$

*Proof.*

$$\begin{aligned}
Var(R_1 + R_2) &= E((R_1 + R_2 - E(R_1 + R_2))^2) \\
&= E(R_1^2 + R_2^2 + 2R_1R_2 - 2(R_1 + R_2)E(R_1 + R_2) + E(R_1 + R_2)^2) \\
&= E(R_1^2 + R_2^2 + 2R_1R_2 - 2(R_1 + R_2)(E(R_1) + E(R_2)) \\
&+ E(R_1)^2 + E(R_2)^2 + 2E(R_1)E(R_2)) \\
&= E(R_1^2) + E(R_2^2) + 2E(R_1R_2) - 2E(R_1)^2 - E(R_2)^2 - 4E(R_1)E(R_2) \\
&+ E(R_1)^2 + E(R_2)^2 + 2E(R_1)E(R_2) \\
&= E(R_1^2) - E(R_1)^2 + E(R_2^2) - E(R_2)^2 + 2E(R_1R_2) - 2E(R_1)E(R_2) \\
&= Var(R_1) + Var(R_2) + 2Cov(R_1, R_2)
\end{aligned}$$

■

# 3 Variance of Sums

As a corollary to the previous theorem we have:

**Corollary 4** *If $R_1$ and $R_2$ are independent, then*

$$Var(R_1 + R_2) = Var(R_1) + Var(R_2)$$

*Proof.* If $R_1, R_2$ are independent, then $Cov(R_1, R_2) = 0$. ■

A similar result holds for the sum of many independent random variables — but only needs that the variables are *pairwise* independent.

**Theorem 5** *If $R_1, \ldots, R_n$ are pairwise independent random variables, then*

$$Var(R_1 + \ldots + R_n) = \sum_{i=1}^{n} Var(R_i).$$

The proof is simple and follows from the definition of variance. We won't do it here.

Using this result we can now derive the variance of a binomial distribution. Let $R$ have an $n, p$ binomial distribution, i.e.,

$$Pr(R = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

We can think of $R$ as the sum of $n$ Bernoulli independent random variables. That is, the number of heads in $n$ tosses with the probability of heads equal to $p$.

$$R = R_1 + R_2 + \ldots + R_n$$

And $Pr(R_i = 1) = p$, $Pr(R_i = 0) = 1 - p$. Now

$$Var(R_i) = E(R_i^2) - E(R_i)^2 = p - p^2$$

Hence from the theorem,

$$Var(R) = \sum_{i=1}^{n} Var(R_i) = nVar(R_i) = p(1-p)n.$$

# 4 Review of Markov, Chebyshev, Chernoff and Binomial bounds

Let us review the methods we have for bounding deviation from the mean via the following example. Assume that IQ is made up of thinkatons; each thinkaton fires independently with a 10% chance. We have 1000 thinkatons in all, and IQ is the number of thinkatons that fire. What is the probability of having Marilyn's IQ of 228?

So the IQ is a Binomial distribution with $n = 1000, p = 0.1$. Hence $E(IQ) = 100, \sigma_{IQ} = \sqrt{0.09 \times 1000} = 9.48$.

An IQ of 228 is $128/9.48 > 13.5$ standard deviations away.

Let us compare the methods we have for bounding the probability of this IQ.

1. Markov:
$$Pr(IQ \geq 228) \leq \frac{100}{228} < 0.44$$
2. Chebyshev:
$$Pr(IQ - 100 \geq 128) \leq \frac{1}{13.5^2 + 1} < \frac{1}{183}$$
3. Chernoff:
$$Pr(IQ \geq 2.28 \times 100) \leq e^{-(\ln 2.28 + 1/2.28 - 1)228} \leq e^{-59.9}$$
4. Binomial tails:
$$Pr(IQ \geq 228) = Pr(1000 - IQ \leq 772) = F_{0.9,1000}(772) \leq e^{-72.5}$$
   Here we used the formula for the binomial tail from Lecture 21 with $p = 0.9, n = 1000, \alpha = 0.772$.

Note that the more we know about the distribution the better are the bounds we can obtain on the probability.

# 5 Chernoff bounds and binomial tails

Here we derive a bound on the deviation of a binomial random variable from its mean in terms of standard deviations.

**Theorem 6** *If $R$ has an unbiased binomial distribution, then for $c \geq 1$,*
$$Pr(R - E(R) \geq c\sigma_R) \leq e^{-c^2/2}$$

¿From Chebyshev we know that this probability is at most $\frac{1}{1+c^2}$. We can actually show that this probability is at most $e^{-c^2/2}$, which is much smaller as $c$ increases. The proof can be derived from the tail probabilities of the binomial distribution covered in Lecture 21. Alternatively, we can use the following Chernoff-bound-like result:

**Theorem 7** *Let $X_1, X_2, \ldots, X_n$ be independent random variables with $Pr(X_i = 1) = Pr(X_i = -1) = \frac{1}{2}$. Let $X = \sum_{i=1}^n X_i$. Then,*
$$Pr(X \geq c\sqrt{n}) \leq e^{-c^2/2}$$

*Proof of Theorem* **??**. For $i = 1, \ldots, n$, let $R_i = \frac{X_i+1}{2}$. Then $Pr(R_i = 0) = Pr(R_i = 1) = \frac{1}{2}$. That is the $R_i$'s are unbiased Bernoulli trials. Then $R = \sum_{i=1}^n R_i$, has the unbiased binomial distribution.

Now

$$\begin{aligned} Pr(R - \frac{n}{2} \geq c\frac{\sqrt{n}}{2}) &= Pr(2R - n \geq c\sqrt{n}) \\ &= Pr(X \geq c\sqrt{n}) \\ &= \leq e^{-c^2/2} \end{aligned}$$

But $E(R) = \frac{n}{2}$ and $\sigma_R = \frac{\sqrt{n}}{2}$. Hence,

$$Pr(R - E(R) \geq c\sigma_R) \leq e^{-c^2/2}$$

∎

*Proof of Theorem* **??**.
$Pr(X \geq c\sqrt{n}) = Pr(e^{tX} \geq e^{tc\sqrt{n}})$ for any $t > 0$. By Markov's inequality,

$$Pr(e^{tX} \geq e^{tc\sqrt{n}}) \leq \frac{E(e^{tX})}{e^{tc\sqrt{n}}}$$

Since the $X_i's$ are independent,

$$E(e^{tX}) = E(e^{t\sum_{i=1}^{n} X_i}) = [E(e^{tX_i})]^n$$

Now,

$$E(e^{tX_i}) = \frac{e^t + e^{-t}}{2} \leq e^{t^2/2}$$

from the Taylor expansions of $e^t$ and $e^{-t}$. Hence,

$$Pr(X \geq c\sqrt{n}) \leq \frac{e^{\frac{t^2 n}{2}}}{e^{tc\sqrt{n}}} = e^{\frac{t^2 n}{2} - tc\sqrt{n}}$$

Since we can choose any positive $t$, use $t = \frac{c}{\sqrt{n}}$ to get

$$Pr(X \geq c\sqrt{n}) \leq e^{-c^2/2}$$

∎

# 6 Generalized Chernoff

The theorem from the previous section can be generalized as follows. We don't give the proof here, but it is similar to the previous one. Note that this result is not as good as those derived previously for binomial distributions but it is more general in what it covers.

**Theorem 8** *Let $R_1, \ldots, R_n$ be independent random variables with $0 \le R_i \le 1$. Let $R = \sum_{i=1}^{n} R_i$. Then*

$$Pr(R - E(R) \ge c\sqrt{n}) \le e^{-c^2/2}$$

*Example:* Load balancing.

A set of jobs $J_1, \ldots, J_n$ have to be scheduled on a set of processors $P_1, \ldots, P_m$. The length (processing time) of each job is some number $L_i$ in the range $[0, 1]$, unknown to us. We would like to schedule the jobs on the machines so that the load on the machines is balanced. The average load is $L = \frac{\sum_{i=1}^{n} L_i}{m}$. We would like the load on every machine to be not much more than this. Achieving such a schedule can be hard though, since we don't know ahead of time which jobs are long and which are short. By using a random strategy we can be almost certain to do very well, however. For example, suppose we pick a random machine for each job.

Consider this process from a machine's point of view. The probability that job $j$ is assigned to it is $\frac{1}{m}$. Let $R_{ij}$ be the random variable indicating the time spent by the machine $i$ on job $j$. Then

$$Pr(R_{ij} = L_j) = \frac{1}{m}, \quad Pr(R_{ij} = 0) = 1 - \frac{1}{m}.$$

Hence $E(R_{ij}) = \frac{L_j}{m}$. The load on machine $i$ is then $\sum_{j=1}^{n} R_{ij}$.

$$E(\sum_{j=1}^{n} R_{ij}) = \sum_{j=1}^{n} E(R_{ij}) = \sum_{j=1}^{n} \frac{L_j}{m} = L$$

¿From the generalized Chernoff bound,

$$Pr(\text{load on machine i } - L \ge c\sqrt{n}) \le e^{-c^2/2}$$

By Boole's inequality,

$$\begin{aligned} Pr(\text{load on any machine } \ge L + c\sqrt{n}) &\le \sum_{i=1}^{m} Pr(\text{load on machine i } \ge L + c\sqrt{n}) \\ &\le me^{-c^2/2} \end{aligned}$$

If we choose $c = \sqrt{2 \ln m + 6}$, say, then

$$Pr(\text{load on each machine } \le L + c\sqrt{n}) \ge 1 - me^{-\ln m - 18} = 1 - e^{-18} = 0.99999998..$$

Hence we can be fairly sure (i.e., with probability at least $1 - e^{-18}$) that every machine will have load at most $L + \sqrt{2n \ln m + 6}\sqrt{n}$. For many values of $n$ and $m$ this is very close to optimal. For example, if $m = 10$, $n = 5000$, and $L = 300$, the maximum load on any machine will almost surely not exceed 332. (In fact it is likely to be even less — we have been fairly crude in our bounds.)

## 7 Team Problem

Consider a square grid of 100x100 nodes. This is a communication network. Each node sends 1 message per second to a randomly selected destination node. Messages are routed through the row to the correct column and then through the column to the correct destination. It can be assumed that messages are routed instantaneously, provided there is sufficient bandwidth (capacity) on the edges of the network. The problem is to get upper bounds on the edge capacities so that with probability 0.99 all traffic for the next 10 years can be handled without ever exceeding any edge capacity. In other words, we want to be virtually certain that there will never be a second in the next 10 years during which we don't have enough capacity on a link in the network.

1. What is the expected traffic per second through any given edge? [Hint: This is not the same for every edge.]
2. We need a success probability of 99% over 10 years. What success probability do we need per second? Let $P_f$ be the probability of failure in 1 second. That is $1 - P_f$ is the probability that the network succesfully handles the load in one second (i.e. the traffic through every edge is at most capacity of the edge). How small must $P_f$ be to ensure that probability of any failure ever in 10 years is less than 1%.
3. Show that if each edge has capacity $136 > 50e$, then the guarantee can be made. Use the bound obtained on $P_f$. [Hint: Apply Boole's inequality and Chernoff's bound.]
4. Find better bounds for the edge capacities. You can specify a different bound for each edge (depending on its expected traffic).