Lecture 3 - Expectation, moments and inequalities

Jan Bouda

FI MU

March 21, 2012

Part I

Moments and Deviations

Moments

- Let us suppose we have a random variable X and a random variable $Y = \Phi(X)$ for some function Φ. The expected value of Y is

$$E(Y) = \sum_i \Phi(x_i) p_X(x_i).$$

- Especially interesting is the power function $\Phi(X) = X^k$. $E(X^k)$ is known as the kth moment of X. For $k = 1$ we get the expectation of X.
- If X and Y are random variables with matching corresponding moments of all orders, i.e. $\forall k \;\; E(X^k) = E(Y^k)$, then X and Y have the same distributions.
- Usually we center the expected value to 0 – we use moments of $\Phi(X) = X - E(X)$.
- We define the kth central moment of X as

$$\mu_k = E\left([X - E(X)]^k \right).$$

Variance

Definition

The second central moment is known as the **variance** of X and defined as

$$\mu_2 = E\left([X - E(X)]^2\right).$$

Explicitly written,

$$\mu_2 = \sum_i [x_i - E(X)]^2 p(x_i).$$

The variance is usually denoted as σ_X^2 or $Var(X)$.

Definition

The square root of σ_X^2 is known as the **standard deviation** $\sigma_X = \sqrt{\sigma_X^2}$.

If variance is small, then X takes values close to $E(X)$ with high probability. If the variance is large, then the distribution is more 'diffused'.

Variance

Theorem

Let σ_X^2 be the variance of the random variable X. Then

$$\sigma_X^2 = E(X^2) - [E(X)]^2.$$

Proof.

$$\begin{aligned}\sigma_X^2 =& E\left([X - E(X)]^2\right) = E\left(X^2 - 2XE(X) + [E(X)]^2\right) = \\ =& E(X^2) - E[2XE(X)] + [E(X)]^2 = \\ =& E(X^2) - 2E(X)E(X) + [E(X)]^2.\end{aligned}$$

□

Covariance

Definition

The quantity

$$E\big([X - E(X)][Y - E(Y)]\big) = \sum_{i,j} p_{x_i,y_j} \left[x_i - E(X)\right] \left[y_j - E(Y)\right]$$

is called the **covariance** of X and Y and denoted $Cov(X,Y)$.

Theorem

Let X and Y be independent random variables. Then the covariance of X and Y $Cov(X,Y) = 0$.

Covariance

Proof.

$$\begin{aligned} Cov(X,Y) =& E\big([X-E(X)][Y-E(Y)]\big) = \\ =& E[XY - YE(X) - XE(Y) + E(X)E(Y)] = \\ =& E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) = \\ =& \underbrace{E(X)E(Y)}_{\text{independence}} - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) = 0 \end{aligned}$$

□

- Covariance measures linear (!) dependence between two random variables. It is positive if the variables are "correlated", and negative when "anticorrelated".
- E.g. when $X = aY$, $a \neq 0$, using $E(X) = aE(Y)$ we have

$$Cov(X,Y) = aVar(Y) = \frac{1}{a}Var(X).$$

Covariance

In general it holds that

$$0 \leq Cov^2(X, Y) \leq Var(X)Var(Y).$$

Definition

We define the **correlation coefficient** $\rho(X, Y)$ as the normalized covariance, i.e.

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

It holds that $-1 \leq \rho(X, Y) \leq 1$.

Covariance

It may happen that X is completely dependent on Y and yet the covariance is 0, e.g. for $X = Y^2$ and a suitably chosen Y.

Variance of Independent Variables

Theorem

If X and Y are independent random variables, then

$$Var(X+Y) = Var(X) + Var(Y).$$

Proof.

$$\begin{aligned}
&Var(X+Y) = E\big([(X+Y) - E(X+Y)]^2\big) = \\
=&E\big([(X+Y) - E(X) - E(Y)]^2\big) = E\big([(X - E(X)) + (Y - E(Y))]^2\big) = \\
=&E\big([X - E(X)]^2 + [Y - E(Y)]^2 + 2[X - E(X)][Y - E(Y)]\big) = \\
=&E\big([X - E(X)]^2\big) + E\big([Y - E(Y)]^2\big) + 2E\big([X - E(X)][Y - E(Y)]\big) = \\
=&Var(X) + Var(Y) + 2E\big([X - E(X)][Y - E(Y)]\big) = \\
=&Var(X) + Var(Y) + 2Cov(X,Y) = Var(X) + Var(Y).
\end{aligned}$$

□

Variance

- If X and Y are not independent, we obtain (see proof on the previous transparency)

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

- The additivity of variance can be generalized to a set $X_1, X_2, \dots X_n$ of mutually independent variables and constants $a_1, a_2, \dots a_n \in \mathbb{R}$ as

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

Proof is left as a home exercise :-).

Part II

Conditional Distribution and Expectation

Conditional probability

Using the derivation of conditional probability of two events we can derive conditional probability of (a pair of) random variables.

Definition

The **conditional probability distribution** of random variable Y given random variable X (their joint distribution is $p_{X,Y}(x,y)$) is

$$p_{Y|X}(y|x) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \\ = \frac{p_{X,Y}(x,y)}{p_X(x)} \tag{1}$$

provided $p_X(x) \neq 0$.

Conditional expectation

We may consider $Y|(X = x)$ to be a new random variable that is given by the conditional probability distribution $p_{Y|X}$. Therefore, we can define its mean and moments.

Definition

The **conditional expectation** of Y given $X = x$ is defined

$$E(Y|X = x) = \sum_y yP(Y = y|X = x) = \sum_y yp_{Y|X}(y|x). \tag{2}$$

Analogously can be defined conditional expectation of a transformed random variable $\Phi(Y)$, namely the conditional kth moment of Y: $E(Y^k|X = x)$. Of special interest will be the conditional variance

$$Var(Y|X = x) = E(Y^2|X = x) - [E(Y|X = x)]^2.$$

Conditional expectation

We can derive the expectation of Y from the conditional expectations. The following equation is known as the **theorem of total expectation**:

$$E(Y) = \sum_x E(Y|X = x)p_X(x). \tag{3}$$

Analogously, the **theorem of total moments** is

$$E(Y^k) = \sum_x E(Y^k|X = x)p_X(x). \tag{4}$$

Example: Random sums

Let $N, X_1, X_2, \ldots$ be mutually independent random variables. Let us suppose that $X_1, X_2, \ldots$ have identical probability distribution $p_X(x)$, mean $E(X)$, and variance $Var(X)$. We also know the values $E(N)$ and $Var(N)$. Let us consider the random variable defined as a sum

$$T = X_1 + X_2 + \cdots + X_N.$$

In what follows we would like to calculate $E(T)$ and $Var(T)$. For a fixed value $N = n$ we can easily derive the conditional expectation of T by

$$E(T|N = n) = \sum_{i=1}^{n} E(X_i) = nE(X). \tag{5}$$

Using the theorem of total expectation we get

$$E(T) = \sum_n nE(X)p_N(n) = E(X)\sum_n np_N(n) = E(X)E(N). \tag{6}$$

Example: Random sums

It remains to derive the variance of T. Let us first compute $E(T^2)$. We obtain

$$E(T^2|N=n) = Var(T|N=n) + [E(T|N=n)]^2 \tag{7}$$

and

$$Var(T|N=n) = \sum_{i=1}^{n} Var(X_i) = nVar(X) \tag{8}$$

since $(T|N=n) = X_1 + X_2 + \cdots + X_n$ and $X_1, \ldots, X_n$ are mutually independent.
We substitute (5) and (8) into (7) to get

$$E(T^2|N=n) = nVar(X) + n^2E(X)^2. \tag{9}$$

Example: Random sums

Using the theorem of total moments we get

$$
\begin{aligned}
E(T^2) &= \sum_n \left(nVar(X) + n^2[E(X)]^2\right) p_N(n) \\
&= \left(Var(X)\sum_n np_N(n)\right) + \left([E(X)]^2 \sum_n p_N(n)n^2\right) \\
&= Var(X)E(N) + E(N^2)[E(X)]^2.
\end{aligned} \tag{10}
$$

Finally, we obtain

$$
\begin{aligned}
Var(T) &= E(T^2) - [E(T)]^2 = \\
&= Var(X)E(N) + E(N^2)[E(X)]^2 - [E(X)]^2[E(N)]^2 = \\
&= Var(X)E(N) + [E(X)]^2 Var(N).
\end{aligned} \tag{11}
$$

Part III

Markov and Chebyshev Inequality

Markov Inequality

It is important to derive as much information as possible even from a partial description of random variable. The mean value already gives more information than one might expect, as captured by Markov inequality.

Theorem (Markov inequality)

Let X be a nonnegative random variable with finite mean value $E(X)$. Then for all $t > 0$ it holds that

$$P(X \geq t) \leq \frac{E(X)}{t}$$

Markov Inequality

Proof.

Let us define the random variable Y_t (for fixed t) as

$$Y_t = \begin{cases} 0 & \text{if } X < t \\ t & X \geq t. \end{cases}$$

Then Y_t is a discrete random variable with probability distribution $p_{Y_t}(0) = P(X < t)$, $p_{Y_t}(t) = P(X \geq t)$. We have

$$E(Y_t) = tP(X \geq t).$$

The observation $X \geq Y_t$ gives

$$E(X) \geq E(Y_t) = tP(X \geq t),$$

what is the Markov inequality. □

Markov Inequality: Example

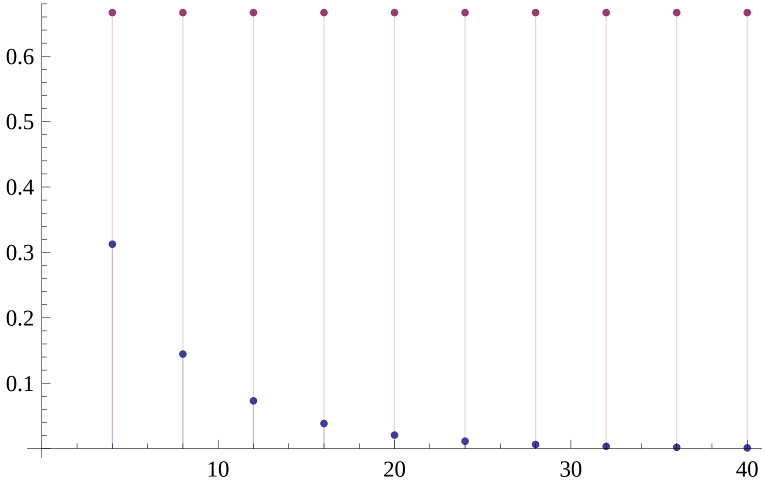
Assume that we want to bound the probability of obtaining more that $3n/4$ heads in a sequence of n fair coin flips. Let

$$X_i = \begin{cases} 1 & \text{if the } i\text{th coin flip is head} \\ 0 & \text{otherwise,} \end{cases}$$

and let $X = \sum_{i=1}^{n} X_i$ be the number of heads in n coin flips. Note that $E(X_i) = 1/2$, and $E(X) = n/2$.
Using the Markov inequality we get

$$P(X \geq 3n/4) \leq \frac{E(X)}{3n/4} = \frac{n/2}{3n/4} = \frac{2}{3}.$$

Markov Inequality: Example



Chebyshev Inequality

In case we know both mean value and variance of a random variable, we can use much more accurate estimation

Theorem (Chebyshev inequality)

Let X be a random variable with finite variance. Then

$$P\left[|X - E(X)| \geq t\right] \leq \frac{Var(X)}{t^2}, \ t > 0$$

or, alternatively, substituting $X' = X - E(X)$

$$P(|X'| \geq t) \leq \frac{E(X'^2)}{t^2}, \ t > 0.$$

We can see that this theorem is in agreement with our interpretation of variance. If σ^2 is small, then there is a large probability of getting outcome close to $E(X)$. If σ^2 is large, then there is a large probability of getting outcomes farther from the mean.

Chebyshev Inequality

Proof.

We apply the Markov inequality to the nonnegative variable $[X - E(X)]^2$ and we replace t by t^2 to get

$$P\left[(X - E(X))^2 \geq t^2\right] \leq \frac{E\left([X - E(X)]^2\right)}{t^2} = \frac{\sigma^2}{t^2}.$$

We obtain the Chebyshev inequality using the fact that the events $[(X - E(X))^2 \geq t^2] = [|X - E(X)| \geq t]$ are the same. $\square$

Chebyshev Inequality: Example

Let us again consider the coin flipping example and try to bound the probability that we obtain more than $3n/4$ heads. Again, $X_i = 1$ if the ith outcome is head and 0 otherwise, and $X = \sum_{i=1}^{n} X_i$. Let us calculate the variance of X:

$$E(X_i^2) = E(X_i) = \frac{1}{2}.$$

Then

$$Var(X_i) = E(X_i^2) - [E(X_i)]^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

and using the independence we have

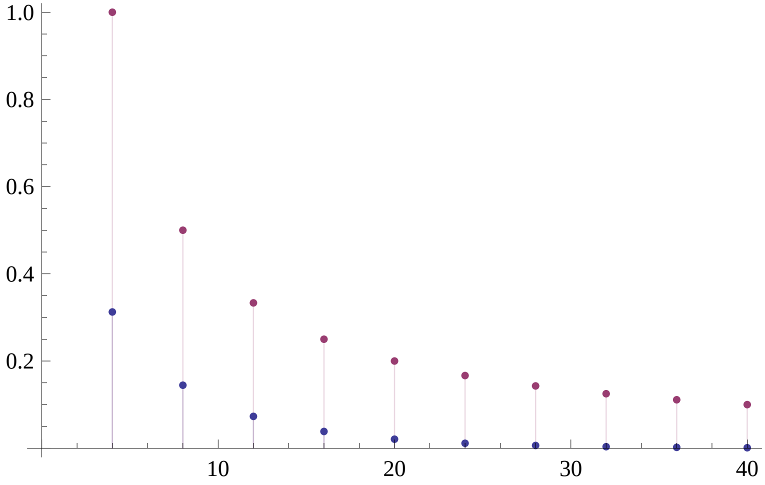
$$Var(X) = \frac{n}{4}.$$

Chebyshev Inequality: Example

We apply the Chebyshev bound to get

$$\begin{aligned}P(X \geq 3n/4) \leq & P(|X - E(X)| \geq n/4)\\ \leq & \frac{Var(X)}{(n/4)^2}\\ = & \frac{n/4}{(n/4)^2}\\ = & \frac{4}{n}.\end{aligned}$$

Chebyshev Inequality: Example



Part IV

Moment Generating Functions and Chernoff Bounds

Moment Generating Function

Definition

The **moment generating function** of a random variable X is

$$M_X(t) = E(e^{tX}).$$

We will be interested mainly in the properties of this function around $t = 0$.

Moment Generating Function and Moments

The moment generating function captures all moments:

Theorem

Let $M_X(t)$ be a moment generating function of X. Assuming that exchanging the expectation and differentiation operands is legitimate, for all $n > 1$ we have

$$E(X^n) = M_X^{(n)}(0),$$

where $M_X^{(n)}(0)$ is the nth derivative of $M_X(t)$ evaluated at 0.

The assumption that expectation and differentiation can be exchanged holds whenever the moment generating function exists in a neighborhood of 0.

Moment Generating Function and Moments

Proof.

Assuming that exchanging the expectation and differentiation operands is legitimate, we have

$$M_X^{(n)}(t) = E(X^n e^{tX}). \quad (12)$$

Computing at $t = 0$ we get

$$M_X^{(n)}(0) = E(X^n). \quad (13)$$

□

Moment Generating Function and Distributions

Moment generating functions uniquely define the probability distribution:

Theorem

Let X and Y be two random variables, then

$$M_X(t) = M_Y(t) \tag{14}$$

for some $\delta > 0$ and all $-\delta < t < \delta$

This allows us e.g. to calculate probability distribution of sum of independent random variables:

Theorem

If X and Y are independent random variables, then

$$M_{X+Y}(t) = M_X(t)M_Y(t). \tag{15}$$

Moment Generating Function and Distributions

Proof.

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) \overbrace{= E(e^{tX})E(e^{tY})}^{\text{using independence}} = M_X(t)M_Y(t).$$

□

Chernoff Bound

The Chernoff bound for random variable X is obtained by applying the Markov inequality to e^{tX} for some suitably chosen t. For any $t > 0$

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}}. \tag{16}$$

Similarly, for any $t < 0$

$$P(X \leq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E(e^{tX})}{e^{ta}}. \tag{17}$$

While the value of t that minimizes $\frac{E(e^{tX})}{e^{ta}}$ gives the best bound, in practice we usually use the value of t that gives a convenient form. Bounds derived using this approach are called the **Chernoff bounds**.

Chernoff Bound and a Sum of Poisson Trials

Poisson trials (do not confuse with Poisson random variables!!) are a sequence of independent coin flips, but the probability of respective coin flips differs. Bernoulli trials are a special case of the Poisson trials.
Let $X_1, \ldots, X_n$ be independent Poisson trials with $P(X_i = 1) = p_i$, and $X = \sum_{i=1}^{n} X_i$ their sum. Note that the expected value is

$$E(X) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p_i.$$

We want to bound the probabilities $P(X \geq (1+\delta)E(X))$ and $P(X \leq (1-\delta)E(X))$

Chernoff Bound and a Sum of Poisson Trials

We derive a bound on the moment generating function

$$
\begin{aligned}
M_{X_i}(t) =& E(e^{tX_i}) = p_i e^t + (1 - p_i) \\
=& 1 + p_i(e^t - 1) \leq e^{p_i(e^t - 1)}
\end{aligned}
$$

using that for any y, $1 + y \leq e^y$.
The generating function of X is

$$
\begin{aligned}
M_X(t) =& \prod_{i=1}^{n} M_{x_i}(t) \leq \prod_{i=1}^{n} e^{p_i(e^t - 1)} \\
=& exp \left\{ \sum_{i=1}^{n} p_i(e^t - 1) \right\} = e^{(e^t - 1)E(X)}.
\end{aligned}
$$

Chernoff Bound and a Sum of Poisson Trials

Theorem

Let $X_1, \ldots, X_n$ be independent Poisson trials with $P(X_i = 1) = p_i$, $X = \sum_{i=1}^{n} X_i$ their sum and $\mu = E(X)$. Then the following Chernoff bounds hold:

1. *for any $\delta > 0$*

$$P(X \geq (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

2. *for $0 < \delta \leq 1$*

$$P(X \geq (1+\delta)\mu) \leq e^{-\mu\delta^2/3}$$

Chernoff Bound and a Sum of Poisson Trials

Proof.

1. Using Markov inequality we have that for any $t > 0$

$$\begin{aligned} P(X \geq (1+\delta)\mu) =& P(e^{tX} \geq e^{t(1+\delta)\mu}) \\ \leq& \frac{E(e^{tX})}{e^{t(1+\delta)\mu}} \\ \leq& \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}}. \end{aligned}$$

For any $\delta > 0$ we can set $t = \ln(1+\delta)$ to get

$$P(X \geq (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

□

Chernoff Bound and a Sum of Poisson Trials

Proof.

2. We want to show that for any $0 < \delta \leq 1$

$$\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}} \leq e^{-\delta^2/3},$$

what will give us the result immediately. Taking the natural logarithm of both sides we obtain the equivalent condition

$$f(\delta) \stackrel{def}{=} \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3} \leq 0.$$

□

Chernoff Bound and a Sum of Poisson Trials

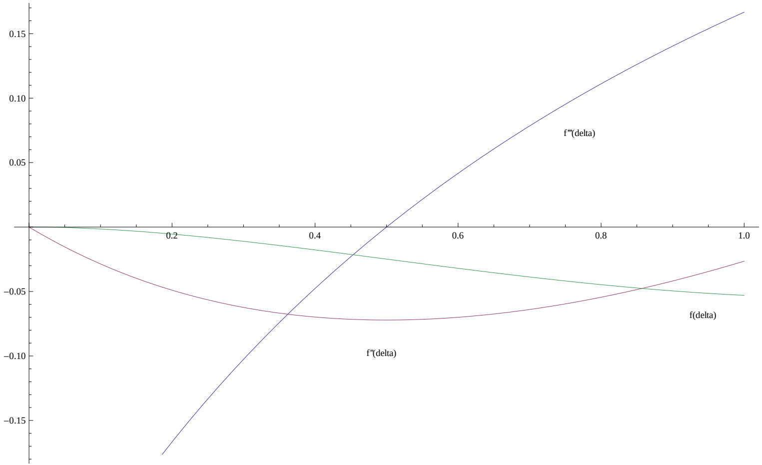
Proof.

We calculate the first and second derivative of $f(\delta)$

$$f'(\delta) = 1 - \frac{1+\delta}{1+\delta} - \ln(1+\delta) + \frac{2}{3}\delta = -ln(1+\delta) + \frac{2}{3}\delta$$
$$f''(\delta) = -\frac{1}{1+\delta} + \frac{2}{3}.$$

We see that $f''(\delta) < 0$ for $0 \leq \delta < 1/2$ and $f''(\delta) > 0$ for $\delta > 1/2$. Hence, $f'(\delta)$ first decreases and then increases on $[0, 1]$. Since $f'(0) = 0$ and $f'(1) < 0$, we see that $f'(t) \leq 0$ on $[0, 1]$. Since $f(0) = 0$, it follows that $f(t) \leq 0$ on $[0, 1]$ as well, what completes the proof. □

Chernoff Bound and a Sum of Poisson Trials



Chernoff Bound and a Sum of Poisson Trials

Theorem

Let $X_1, \ldots, X_n$ be independent Poisson trials with $P(X_i = 1) = p_i$, $X = \sum_{i=1}^{n}$ their sum and $\mu = E(X)$. Then for $0 < \delta \leq 1$

1. $$P(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$

2. $$P(X \leq (1-\delta)\mu) \leq e^{-\mu\delta^2/2}$$

Proof: Analogous to the previous theorem, left as a home exercise. Hint: start with any $t < 0$.

Chernoff Bound and a Sum of Poisson Trials

Corollary

Let $X_1, \ldots, X_n$ *be independent Poisson trials and* $X = \sum_{i=1}^{n} X_i$. *For* $0 < \delta < 1$,

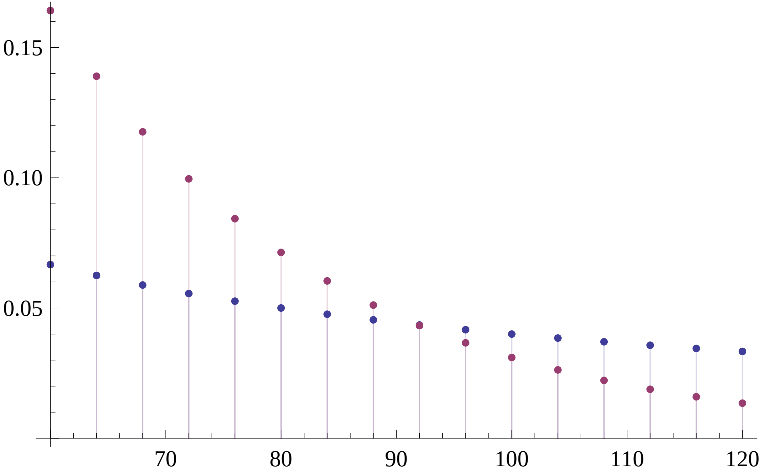
$$P(|X - E(X)| \geq \delta E(X)) \leq 2e^{-E(X)\delta^2/3}$$

Chernoff Bound: Example

Let us once again consider the coin flipping example and try to bound the probability that we obtain more than $3n/4$ heads. Again, $X_i = 1$ if the ith outcome is head and 0 otherwise, and $X = \sum_{i=1}^{n} X_i$.
Using the Chernoff bound for Poisson trials we get

$$\begin{aligned} P(X \geq 3n/4) \leq & P(|X - E(X)| \geq n/4) \\ \leq & 2e^{-\frac{1}{3}\frac{n}{2}\frac{1}{4}} \\ \leq & 2e^{-n/24}. \end{aligned}$$

Chernoff Bound: Example



Part V

Laws of Large Numbers

(Weak) Law of Large Numbers

Theorem ((Weak) Law of Large Numbers)

Let $X_1, X_2, \ldots$ be a sequence of mutually independent random variables with a common probability distribution. If the expectation $\mu = E(X_k)$ exists, then for every $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) = 0.$$

In words, the probability that the average S_n/n differs from the expectation by less then arbitrarily small ϵ goes to 0.

Proof.

WLOG we can assume that $\mu = E(X_k) = 0$, otherwise we simply replace X_k by $X_k - \mu$. This induces only change of notation. □

(Weak) Law of Large Numbers

Proof.

In the special case $Var(X_k)$ exists, the law of large numbers is a direct consequence of the Chebyshev inequality; we substitute $X = X_1 + \cdots + X_n = S_n$ to get

$$P(|S_n - \mu| \geq t) \leq \frac{Var(X_k)n}{t^2}. \tag{18}$$

We substitute $t = \epsilon n$ and observe that with $n \to \infty$ the right-hand side tends to 0 to get the result. However, in case $Var(X_k)$ exists, we can apply the more accurate central limit theorem. The proof without the assumption that $Var(X_k)$ exists follows. □

(Weak) Law of Large Numbers

Proof.

Let δ be a positive constant to be determined later. For each k we define a pair of random variables $(k = 1 \dots n)$

$$U_k = X_k, V_k = 0 \qquad \text{if } |X_k| \leq \delta n \tag{19}$$

$$U_k = 0, V_k = X_k \qquad \text{if } |X_k| > \delta n \tag{20}$$

By this definition

$$X_k = U_k + V_k. \tag{21}$$

□

(Weak) Law of Large Numbers

Proof.

To prove the theorem it suffices to show that both

$$\lim_{n\to\infty} P(|U_1 + \cdots + U_n| > \frac{1}{2}\epsilon n) = 0 \tag{22}$$

and

$$\lim_{n\to\infty} P(|V_1 + \cdots + V_n| > \frac{1}{2}\epsilon n) = 0 \tag{23}$$

hold, because $|X_1 + \cdots + X_n| \leq |U_1 + \cdots + U_n| + |V_1 + \cdots + V_n|$.
Let us denote all possible values of X_k by $x_1, x_2, \ldots$ and the corresponding probabilities $p(x_i)$. We put

$$a = E(|X_k|) = \sum_i |x_i| p(x_i). \tag{24}$$

□

(Weak) Law of Large Numbers

Proof.

The variable U_1 is bounded by δn and $|X_1|$ and therefore

$$U_1^2 \leq |X_1|\delta n.$$

Taking expectation on both sides gives

$$E(U_1^2) \leq a\delta n. \tag{25}$$

Variables $U_1, \ldots U_n$ are mutually independent and have the same probability distribution. Therefore,

$$\begin{aligned} E[(U_1 + \cdots + U_n)^2] - [E(U_1 + \cdots + U_n)]^2 = Var(U_1 + \cdots + U_n) = \\ = nVar(U_1) \leq nE(U_1^2) \leq a\delta n^2. \end{aligned} \tag{26}$$

□

(Weak) Law of Large Numbers

Proof.

On the other hand, $\lim_{n\to\infty} E(U_1) = E(X_1) = 0$ and for sufficiently large n we have

$$[E(U_1 + \cdots U_n)]^2 = n^2[E(U_1)]^2 \leq n^2 a\delta \tag{27}$$

and for sufficiently large n we get from Eq. (26) that

$$E[(U_1 + \cdots + U_n)^2] \leq 2a\delta n^2. \tag{28}$$

Using the Chebyshev inequality we get the result (22) observing that

$$P(|U_1 + \cdots + U_n| > 1/2\epsilon n) \leq \frac{8a\delta}{\epsilon^2}. \tag{29}$$

By choosing sufficiently small δ we can make the right-hand side arbitrarily small to get (22). □

(Weak) Law of Large Numbers

Proof.

In case of (23) note that

$$P(V_1 + V_2 + \cdots V_n \neq 0) \leq \sum_{i=1}^{n} P(V_i \neq 0) = nP(V_1 \neq 0). \qquad (30)$$

For arbitrary $\delta > 0$ we have

$$P(V_1 \neq 0) = P(|X_1| > \delta n) = \sum_{|x_i| > \delta n} p(x_i) \leq \frac{1}{\delta n} \sum_{|x_i| > \delta n} |x_i| p(x_i). \qquad (31)$$

The last sum tends to 0 as $n \to \infty$ and therefore also the left side tends to 0. This statement is even stronger than (23) and it completes the proof. □

Strong Law of Large Numbers

The (weak) law of large number implies that large values $|S_n - m_n|/n$ occur infrequently. In many practical situation we require the stronger statement that $|S_n - m_n|/n$ remains small for all sufficiently large n.

Definition (Strong Law of Large Numbers)

We say that the sequence $X_1, X_2, \ldots$ obeys the strong law of large numbers if to every pair $\epsilon > 0,\ \delta > 0$ there exists an $n \in \mathbb{N}$ such that

$$P\left(\forall r : \frac{|S_n - m_n|}{n} < \epsilon \wedge \frac{|S_{n+1} - m_{n+1}|}{n+1} < \epsilon \wedge \ldots \frac{|S_{n+r} - m_{n+r}|}{n+r} < \epsilon\right) \geq 1 - \delta, \quad (32)$$

where $m_n = E(S_n)$.

It remains to determine the conditions when the strong law of large numbers holds.

Strong Law of Large Numbers

Theorem (Kolmogorov criterion)

Let $X_1, X_2, \ldots$ be a sequence of random variables with corresponding variances $\sigma_1^2, \sigma_2^2, \ldots$. Then the convergence of the series

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} \tag{33}$$

is a sufficient condition for the strong law of large numbers to apply.