1. Inequalities for random variables

The first things you learn, and probably one of the few things you remember, are the words expected value and standard variation. Mathematical inequalities relating the two, make them so useful. In the mathematics literature the name Markov’s and Chebychev’s inequality are often used for Markov’s inequality

\[ a \Pr(X > a) \leq E X \]

for any positive valued random variable. The proof in the discrete setting are, of course the same:

Discrete Setting: Let \(X\) have the values \(v_1 < ... < v_k\). Let \(j\) be such that \(v_j \leq a < v_k\). Then

\[
EX = \sum_k v_k \Pr(X = v_k) \geq \sum_{k>v} a \Pr(X = v_k) = a \Pr(X > a).
\]

Continuous setting: Now we assume that \(\Omega \subset \mathbb{R}^d\) is some region \(\mu(x)\) is a density function (as in physics for an inhomogenous material). Let \(X : \Omega \to \mathbb{R}_+\) be a positive valued function. Then

\[
EX = \int_{\Omega} X(x)\mu(x)dx \geq \int_{x \in \Omega : X(x) > a} X(x)\mu(x)dx \\
\geq \int_{x \in \Omega : X(x) > a} a\mu(x)dx = a \Pr(X > a)
\]

Recall here that \(\Pr(A) = \int 1_A(x)\mu(x)dx = \int 1_A(x)d\mu(x) = \mathbb{E}1_A\) is the measure of lying in a set. We call this probability if \(\int_{\Omega} \mu(x)dx = 1\).

\[\textbf{Corollary 1.1.} \text{ Let } X \text{ be a random variable. Then} \]

\[a \Pr(X^2 > a) \leq EX^2\]

Moreover,

\[a^p \Pr(|X| > a) \leq E|X|^p.\]

\[\text{Proof. We apply Markov to } Y = X^2 \text{ or } Y = |X|^p.\]

\[\textbf{Corollary 1.2.} \text{ (Chebychev) } \Pr(|X - EX| > a) < \frac{\text{Var}X}{a^2}.\]
Proof. Let $Y = |X - EX|$. We apply Markov’s inequality in the form of \[a^2 \text{Prob}(Y^2 > a) \leq EY^2 = E|X - EX|^2 = \text{Var} \] and get
\[a^2 \text{Prob}(Y^2 > a) \leq EY^2 = E|X - EX|^2 = \text{Var}.\]

We are done. □

Problem 1.3. (Try G2b) You reach into a data set $X$ with $EX = 81.5$ and $\text{Var}(X) = 6.7$ and pull out a 92.7. Should you be surprised?

Indeed, we have $a = |92.7 - 81.5| = 11.2$ and $\text{Var}(X)/a = 6.7/(11.2)^2 = 0.05...$ Yes surprised.

If you were to pull 101.5, then $6.7/20^2 = 6.7/400 = 0.0167$ you are even more surprised. Whereas pulling 82.5 gives $6.7/1 = 6.7$ tells you nothing about the probability, and hence you are not surprised, at all.

1.1. Weak law of large numbers.

Definition 1.4. A family $(f_k)$ of random variables are called independent if
\[\text{Prob}(X_1 \in A_1, ..., X_n \in A_n) = \prod_j \text{Prob}(X_j \in A_j)\]

holds for all $n \in \mathbb{N}$ and intervals $A_j \subset \mathbb{R}$.

Lemma 1.5. Let $X_k$ be independent. Then
\[\text{Var}\left(\sum_k X_k\right) = \sum_k \text{Var}(x_k).\]

Proof. We note that $X_k$ independent implies $X_1$ is independent from $\sum_{k=2}^n X_k$. Thus we get the result by induction by observing that independence implies $E XY = E X E Y$ and hence
\[\text{Var}(X + Y) = E(X + Y)^2 - (E X + E Y)^2\]
\[= EX^2 + EY^2 + 2EXY - (EX)^2 - (EY)^2 - 2EXEY\]
\[= \text{Var}(X) + \text{Var}(Y).\]

Theorem 1.6. Let $X_k$ be independent random variables. Then
\[\text{Prob}\left(\frac{1}{n} \sum_{k=1}^n E(X_k - EX_k) > a\right) \leq \frac{\sum_{k=1}^n \text{Var}(X_k)}{n^2a^2} = \frac{1}{n} \frac{\sum_{k=1}^n \text{Var}(X_k)}{na^2}.\]
In particular, if the $X_k$ have all the same variation, then
\[
\operatorname{Prob}(\left| \frac{1}{n} \sum_{k=1}^{n} (X_k - \operatorname{E}X_k) \right| > a) \leq \frac{\operatorname{Var}(X_1)}{na^2}.
\]

Proof. We apply Chebychev for $b = na$ and $Y = \sum_{k=1}^{n} X_k - \operatorname{E}X_k$. Note that $\operatorname{E}Y = 0$ and hence Chebychev implies
\[
\operatorname{Prob}(\left| Y \right| > b) \leq \frac{\operatorname{Var}(Y)}{b^2} = \frac{\sum_{k=1}^{n} \operatorname{Var}(X_k)}{n^2a}.
\]
Multiplying by $n$ inside the probability gives the result.

Application: We roll a dice with $m$ sides. The expected value of one event is
\[
\operatorname{E}X = \frac{1}{m} \sum_{j=1}^{m} j = \frac{m(m+1)}{2m} = \frac{m+1}{2}.
\]
The second moment is
\[
\operatorname{E}X^2 = \frac{1}{m} \sum_{j=1}^{m} j^2 = \frac{m(m+1)(2m+1)}{6m} = \frac{(m+1)(2m+1)}{6}.
\]
Thus we get
\[
\operatorname{Var}(X) = \frac{(m+1)(2m+1)}{6} - \frac{(m+1)^2}{4} = \frac{2m^2 + 3m + 1}{6} - \frac{m^2 + 2m + 1}{4} = \frac{m^2 + 2m - 1}{12} = \frac{(m+1)^2 - 2}{12}.
\]
In order to test, whether the expected value $m+1/2$ is found frequently we note that
\[
\operatorname{Prob}(\left| \frac{1}{n} \sum_{j=1}^{n} X_j - \frac{(m+1)}{2} \right| > a) \leq \frac{\sum_{j=1}^{n} \operatorname{Var}(X_j)}{n^2a^2} = \frac{\operatorname{Var}(X)}{na^2}.
\]
Let us put $a = 1$, we are one off. Then for $n = K(m+1)^2$ many trials we get
\[
\operatorname{Prob}(\left| \frac{1}{n} \sum_{j=1}^{n} X_j - (m+1)/2 \right| > 1) < \frac{1}{12K}.
\]
Thus with probability $1/12K$ after $n$ many trials we hit a value between $m - 1/2$ and $m + 1/2$, and hence we could guess the number of sides.
Theorem 1.7. (Weak law of large numbers) Let $X_k$ be independent random variables and $a > 0$. Then

$$\lim_{n \to \infty} \text{Prob}\left(\left|\frac{1}{n} \sum_{k=1}^{n} (X_k - E X_k)\right| > a\right) = 0.$$ 

Remark 1.8. This means, if we wait long enough, our statistics will be more and more concentrated around their mean values. We will see this further illustrated later.

Proof. We assume in addition that $\sigma_k = E X_k$ are all upper bounded by some $\sigma_k \leq \sigma$. Then the Chebychev application yields

$$\text{Prob}\left(\left|\frac{1}{n} \sum_{k=1}^{n} (X_k - E X_k)\right| > a\right) \leq \frac{\sigma^2}{na^2}.$$ 

This sending $n$ to $\infty$ gives 0. \hfill \blacksquare

This means, if we wait long enough, our statistics will be more and more concentrated around their mean values. We will see this further illustrated later.

Remark 1.9. This statement is an example of Kolmogorov’s famous 0-1 law: An tail event, i.e. an event which decides an all outcomes of an infinite number of trials, but does not depend of a few finite number of initial events, has probability 0 or 1. Another example of such an event is: In infinitely many trials of rolling a dice I will have found three 6’s one after the other. Of course this event depends on the outcome of an infinite number of trials, but if you not have seen the consecutive after 25 trials, you can not decide the final statement, because the sixes are yet still to come. This indicates that this is a tail event.

Let us calculate the probability: We roll the dice $n$ times. We want three consecutive 6’s. If they occur at the beginning we get

$$\text{Prob}(A_{111}) = a = \frac{1}{6^3}.$$ 

We may partition $n$ into $n/3$ blocks of intervals of length 3. The results in these intervals are independent, and hence

$$\text{Prob}\left(\bigcup_{k=1}^{n/3} A_{3k,3k+1,3k+2}\right) = 1 - \text{Prob}\left(\bigcap_{k=1}^{n/3} A_{3k,3k+1,3k+2}\right)$$

$$= 1 - \prod_{k=1}^{n/3} \text{Prob}(A_{3k,3k+1,3k+2}^c)$$
This is a conservative lower bound, because we may also get consecutive 6’s in position 234. At any rate, as \( n \) tends to \( \infty \), we get \( \lim_n (1 - a)^{n/3} = 0 \). Hence with probability 1 we get three 6’s consecutively.

1.2. Bernoulli random variables. The Bernoulli random variable has outcome 0 or 1, i.e.

\[
\text{Prob}(X = 1) = p \quad \text{Prob}(X = 0) = 1 - p.
\]

We get

\[
E X = p, \quad E X^n = p, \quad \text{Var}(X) = p - p^2 = p(1 - p).
\]

Now we consider the random variable \( X_1, \ldots, X_n \) of \( n \) independent trials and get a new random variable

\[
Y_n = \sum_{k=1}^{n} X_k = \text{card}\{k : X_k = 1\}
\]

is the random size of tails obtained after \( n \) trials. Then

i) \( E Y_n = np \);  
ii) \( \text{Var}(Y) = n(p(1 - p)) \);  
iii) \( \text{Prob}(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k} \).

The random variable is called of Binomial distribution of parameter \((p, n)\).

**Example 1.10.** It is known that a screws produced by a certain company will be defective with prob 0.01. The company sells screws in packages of 10 and offers a money back guarantee if at most one of the 10 is defective. What proportion of packages sold do they expect to replace.

**Solution:** We have to calculate the probability that at least one is defective. Let \( Y \) be \((p, n)\) distributed, where \( p = 0.01 \) and \( n = 10 \). Then \( \{Y \geq 1\}^c = \{Y = 0\} \cup \{Y = 1\} \), and hence

\[
P(Y \geq 1) = 1 - P(Y = 0) - P(Y = 1)
\]

\[
= 1 - \binom{10}{0} (0.01)^0 (0.99)^{10} - \binom{10}{1} (0.01)^1 (0.99)^9 \approx 0.04.
\]

Hence the return rate is 0.4 percent.

**Remark 1.11.** For Bernoulli random variables one can even improve on the estimate in the law of large numbers.