COMP2610 / COMP6261 - Information Theory

Lecture 9: Probabilistic Inequalities

Robert C. Williamson

Research School of Computer Science

Australian National University

20 August, 2018

Last time

Mutual information chain rule

Jensen's inequality

"Information cannot hurt"

Data processing inequality

Review: Data-Processing Inequality

Theorem

if $X \rightarrow Y \rightarrow Z$ then $I(X;Y) \geq I(X;Z)$

- X is the state of the world, Y is the data gathered and Z is the processed data

- No “clever” manipulation of the data can improve the inferences that can be made from the data

- No processing of Y, deterministic or random, can increase the information that Y contains about X

This time

- Markov's inequality

- Chebyshev's inequality

- Law of large numbers

Outline

Expectation and Variance

Let X be a random variable over $\mathcal{X}$, with probability distribution p

Expected value:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot p(x).$$

Variance:

$$\begin{aligned} \mathbb{V}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2. \end{aligned}$$

Standard deviation is $\sqrt{\mathbb{V}[X]}$

Properties of expectation

A key property of expectations is linearity:

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]$$

$$\text{LHS} = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_n \in \mathcal{X}_n} \left(p(x_1, \ldots, x_n) \cdot \sum_{i=1}^{n} x_i \right)$$

This holds even if the variables are dependent!

We have for any $a \in \mathbb{R}$,

$$\mathbb{E}[aX] = a \cdot \mathbb{E}[X].$$

Properties of variance

We have linearity of variance for independent random variables:

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}\left[X_i\right].$$

Does not hold if the variables are dependent

(prove this: expand the definition of variance and rely upon $\mathbb{E}(X_i X_j) = \mathbb{E}(X_i)\mathbb{E}(X_j)$ when $X_i \perp X_j$)

We have for any $a \in \mathbb{R}$,

$$\mathbb{V}[aX] = a^2 \cdot \mathbb{V}[X].$$

1000 school students sit an examination

The busy principal is only told that the average score is 40 (out of 100).

The principal wants to estimate the maximum possible number of students who scored more than 80.

- A question about the *minimum* number of students is trivial to answer. Why?

Markov's Inequality

Motivation

Call x the number of students who score > 80

Call S is the total score of students who score ≤ 80

We know:

$$40 \cdot 1000 - S = \{\text{total score of students who score above } 80\} > 80x$$

Exam scores are nonnegative, so certainly $S \geq 0$

Thus, $80x < 40 \cdot 1000$, or

$$x < 500.$$

Can we formalise this more generally?

Markov's Inequality

Theorem

Let X be a nonnegative random variable. Then, for any $\lambda > 0$,

$$p(X \geq \lambda) \leq \frac{\mathbb{E}[X]}{\lambda}.$$

Bounds probability of observing a large outcome

Vacuous if $\lambda < \mathbb{E}[X]$

Markov's Inequality

Alternate Statement

Corollary

Let X be a nonnegative random variable. Then, for any $\lambda > 0$,

$$p(X \geq \lambda \cdot \mathbb{E}[X]) \leq \frac{1}{\lambda}.$$

Observations of nonnegative random variable unlikely to be much larger than expected value

Vacuous if $\lambda < 1$

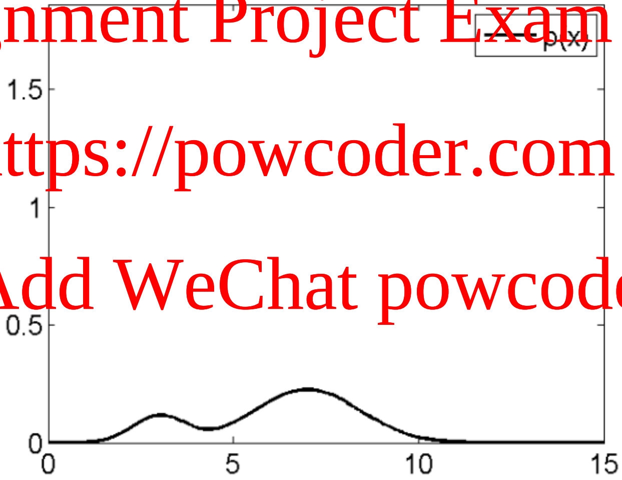
Markov's Inequality

Proof

$$
\begin{aligned}
\mathbb{E}[X] &= \sum_{x \in \mathcal{X}} x \cdot p(x) \\
&= \sum_{x < \lambda} x \cdot p(x) + \sum_{x \geq \lambda} x \cdot p(x) \\
&\geq \sum_{x \geq \lambda} x \cdot p(x) \quad \text{nonneg. of random variable} \\
&\geq \sum_{x \geq \lambda} \lambda \cdot p(x) \\
&= \lambda \cdot p(X \geq \lambda).
\end{aligned}
$$

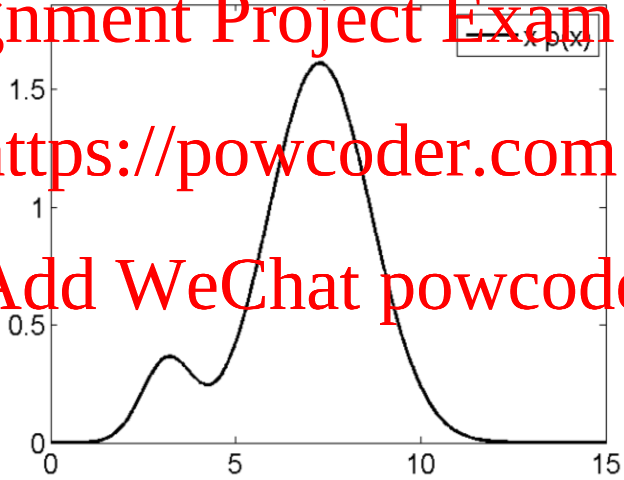
Markov's Inequality

Illustration from
https://justindomke.wordpress.com/2008/06/19/markovs-inequality/



Markov's Inequality

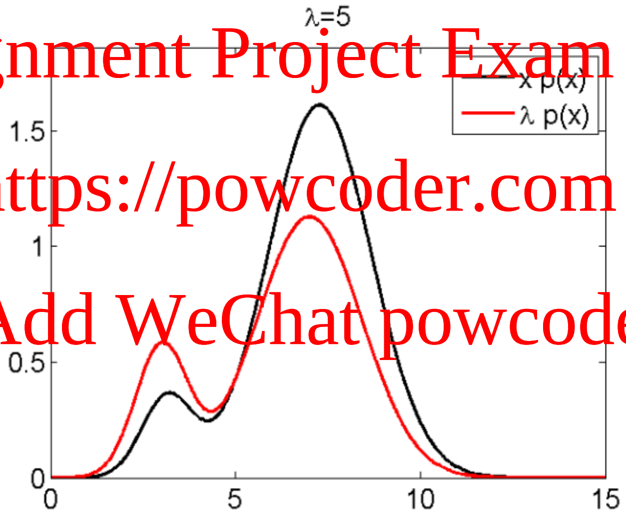
Illustration from
https://justindomke.wordpress.com/2008/06/19/markovs-inequality/



Markov's Inequality

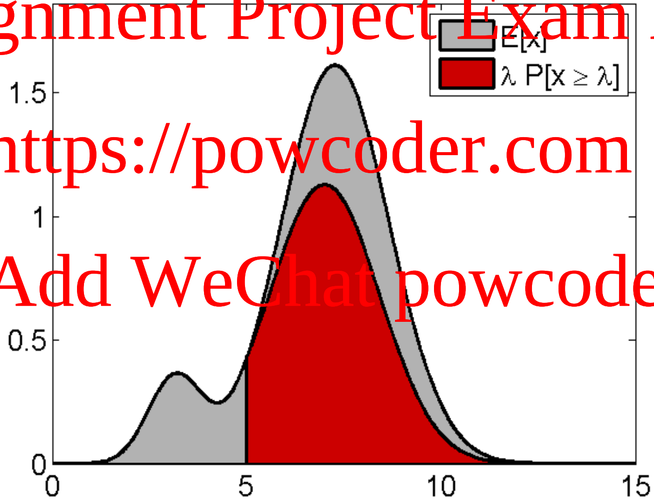
Illustration from

https://justindomke.wordpress.com/2008/06/19/markovs-inequality/



Markov's Inequality

Illustration from
`https://justindomke.wordpress.com/2008/06/19/markovs-inequality/`

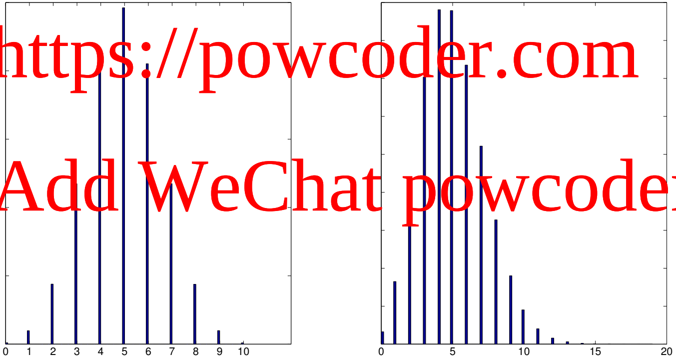


Chebyshev's Inequality

Motivation

Markov's inequality only uses the mean of the distribution

What about the spread of the distribution (variance)?



Chebyshev's Inequality

Theorem

Let X be a random variable with $\mathbb{E}[X] < \infty$. Then, for any $\lambda > 0$,

$$p(|X - \mathbb{E}[X]| \geq \lambda) \leq \frac{\mathbb{V}[X]}{\lambda^2}.$$

Bounds the probability of observing an "unexpected" outcome

Does not require non negativity

Two-sided bound

Chebyshev's Inequality

Alternate Statement

Corollary

Let X be a random variable with $\mathbb{E}[X] < \infty$. Then, for any $\lambda > 0$,

$$p(|X - \mathbb{E}[X]| \geq \lambda \cdot \sqrt{\mathbb{V}[X]}) \leq \frac{1}{\lambda^2}.$$

Observations are unlikely to occur several standard deviations away from the mean

Chebyshev's Inequality

Proof

Define

$$Y = (X - \mathbb{E}[X])^2.$$

Then, by Markov's inequality, for any $\nu > 0$,

$$p(Y \geq \nu) \leq \frac{\mathbb{E}[Y]}{\nu}.$$

But,

$$\mathbb{E}[Y] = \mathbb{V}[X].$$

Also,

$$Y \geq \nu \iff |X - \mathbb{E}[X]| \geq \sqrt{\nu}.$$

Thus, setting $\lambda = \sqrt{\nu}$,

$$p(|X - \mathbb{E}[X]| \geq \lambda) \leq \frac{\mathbb{V}[X]}{\lambda^2}.$$

Chebyshev's Inequality

Illustration

For a binomial X with N trials and success probability θ, we have e.g.

$$p\left(|X - N\theta| \geq \sqrt{2N\theta(1-\theta)}\right) \leq \frac{1}{2}.$$

Chebyshev's Inequality

Example

Suppose we have a coin with bias θ, i.e. $p(X = 1) = \theta$

Say we flip the coin n times, and observe $x_1, \ldots, x_n \in \{0, 1\}$

We use the maximum likelihood estimator of θ:

$$\hat{\theta}_n = \frac{x_1 + \ldots + x_n}{n}$$

Estimate how large n should be such that

$$p(|\hat{\theta}_n - \theta| \geq 0.05) \leq 0.01?$$

1% probability of a 5% error

(Aside: the need for two parameters here is generic: "Probabably Approximately Correct")

Chebyshev's Inequality
Example

Observe that

$$\mathbb{E}[\hat{\theta}_n] = \frac{\sum_{i=1}^{n} \mathbb{E}[x_i]}{n} = \theta$$

$$\mathbb{V}[\hat{\theta}_n] = \frac{\sum_{i=1}^{n} \mathbb{V}[x_i]}{n^2} = \frac{\theta(1-\theta)}{n}.$$

Thus, applying Chebyshev's inequality to $\hat{\theta}_n$,

$$p(|\hat{\theta}_n - \theta| > 0.05) \leq \frac{\theta(1-\theta)}{(0.05)^2 \cdot n}.$$

We are guaranteed this is less than 0.01 if

$$n \geq \frac{\theta(1-\theta)}{(0.05)^2(0.01)}.$$

When $\theta = 0.5$, $n \geq 10,000$ (!)

Independent and Identically Distributed

Let $X_1, \ldots, X_n$ be random variables such that:

- Each X_i is independent of X_j

- The distribution of X_i is the same as that of X_j

Then, we say that $X_1, \ldots, X_n$ are independent and identically distributed (or iid)

Example: For n independent flips of an unbiased coin, $X_1, \ldots, X_n$ are iid from Bernoulli$\left(\frac{1}{2}\right)$

Law of Large Numbers

Theorem

Let $X_1, \ldots, X_n$ be a sequence of iid random variables, with

$$\mathbb{E}[X_i] = \mu$$

and $\mathbb{V}[X_i] < \infty$. Define

$$\bar{X}_n = \frac{X_1 + \ldots + X_n}{n}.$$

Then, for any $\epsilon > 0$,

$$\lim_{n \to \infty} p(|\bar{X}_n - \mu| > \epsilon) = 0.$$

Given enough trials, the empirical “success frequency” will be close to the expected value

Law of Large Numbers

Proof

Since the X_i's are identically distributed,

$$\mathbb{E}[\bar{X}_n] = \mu.$$

Since the X_i's are independent,

$$\begin{aligned}
\mathbb{V}[\bar{X}_n] &= \mathbb{V}\left[\frac{X_1 + \ldots + X_n}{n}\right] \\
&= \frac{\mathbb{V}\left[X_1 + \ldots + X_n\right]}{n^2} \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Law of Large Numbers

Proof

Applying Chebyshev's inequality to $\bar{X}_n$,

$$
\begin{aligned}
p(|\bar{X}_n - \mu| \geq \epsilon) &\leq \frac{\mathbb{V}[\bar{X}_n]}{\epsilon^2} \\
&= \frac{\sigma^2}{n\epsilon^2}.
\end{aligned}
$$

For any fixed $\epsilon > 0$, as $n \to \infty$, the right hand side $\to 0$.

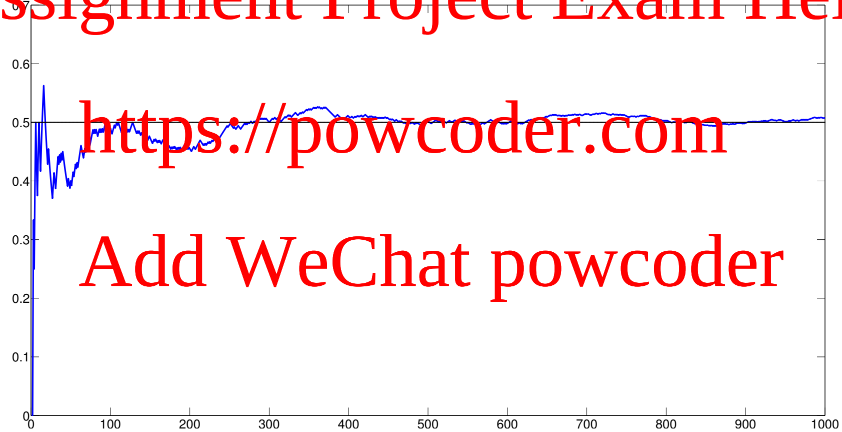
Thus,

$$
p(|\bar{X}_n - \mu| < \epsilon) \to 1.
$$

Law of Large Numbers

Illustration

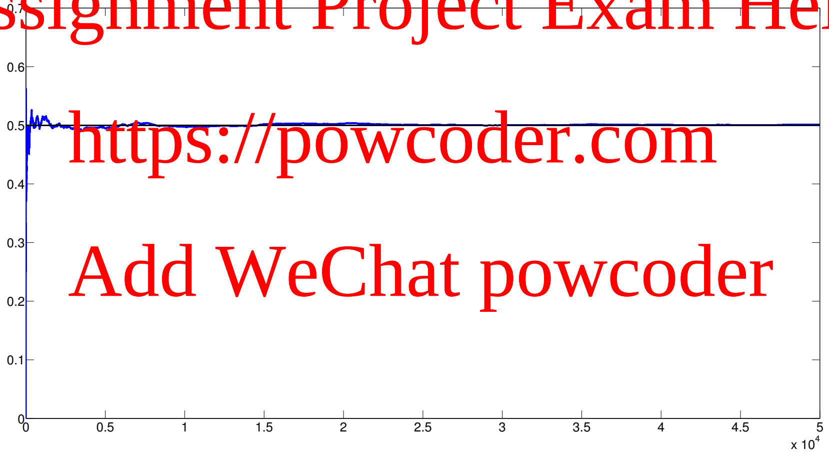
$N = 1000$ trials with Bernoulli random variable with parameter $\frac{1}{2}$



Law of Large Numbers

Illustration

$N = 50000$ trials with Bernoulli random variable with parameter $\frac{1}{2}$



Summary & Conclusions

- Markov's inequality

- Chebyshev's inequality

- Law of large numbers

Next time

- Ensembles and sequences

- Typical sets

- Approximation Equipartition (AEP)