DIS 13A

1 Chebyshev's Inequality vs. Central Limit Theorem

Let n be a positive integer. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with the following distribution:

$$\mathbb{P}[X_1 = -1] = \frac{1}{12}; \qquad \mathbb{P}[X_1 = 1] = \frac{9}{12}; \qquad \mathbb{P}[X_1 = 2] = \frac{2}{12}.$$

(a) Calculate the expectations and variances of X_1, $\sum_{i=1}^n X_i$, $\sum_{i=1}^n (X_i - \mathbb{E}[X_i])$, and

$$Z_n = \frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{n/2}}.$$

(b) Use Chebyshev's Inequality to find an upper bound b for $\mathbb{P}[|Z_n| \geq 2]$.

(c) Can you use b to bound $\mathbb{P}[Z_n \geq 2]$ and $\mathbb{P}[Z_n \leq -2]$?

(d) As $n \to \infty$, what is the distribution of Z_n?

(e) We know that if $Z \sim \mathcal{N}(0,1)$, then $\mathbb{P}[|Z| \leq 2] = \Phi(2) - \Phi(-2) \approx 0.9545$. As $n \to \infty$, can you provide approximations for $\mathbb{P}[Z_n \geq 2]$ and $\mathbb{P}[Z_n \leq -2]$?

Solution:

(a) $\mathbb{E}[X_1] = -1/12 + 9/12 + 4/12 = 1$, and

$$\operatorname{var} X_1 = \frac{1}{12} \cdot 2^2 + \frac{9}{12} \cdot 0^2 + \frac{2}{12} \cdot 1^2 = \frac{1}{2}.$$

Using linearity of expectation and variance (since $X_1, \ldots, X_n$ are independent), we find that $\mathbb{E}[\sum_{i=1}^n X_i] = n$ and $\operatorname{var}(\sum_{i=1}^n X_i) = n/2$.

Again, by linearity of expectation, $\mathbb{E}[\sum_{i=1}^n X_i - n] = n - n = 0$. Subtracting a constant does not change the variance, so $\operatorname{var}(\sum_{i=1}^n X_i - n) = n/2$, as before.

Using the scaling properties of the expectation and variance, $\mathbb{E}[Z_n] = 0/\sqrt{n/2} = 0$ and $\operatorname{var} Z_n = (n/2)/(n/2) = 1$.

(b)

$$\mathbb{P}[|Z_n| \geq 2] \leq \frac{\operatorname{var} Z_n}{2^2} = \frac{1}{4}$$

(c) $1/4$ for both, since $\mathbb{P}[Z_n \geq 2] \leq \mathbb{P}[|Z_n| \geq 2]$ and $\mathbb{P}[Z_n \leq -2] \leq \mathbb{P}[|Z_n| \geq 2]$.

(d) By the Central Limit Theorem, we know that $Z_n \to \mathcal{N}(0,1)$, the standard normal distribution.

(e) Since $Z_n \to \mathcal{N}(0,1)$, we can approximate $\mathbb{P}[|Z_n| \geq 2] \approx 1 - 0.9545 = 0.0455$. By the symmetry of the normal distribution, $\mathbb{P}[Z_n \geq 2] = \mathbb{P}[Z_n \leq -2] \approx 0.0455/2 = 0.02275$.

2 Binomial Concentration

Here, we will prove that the binomial distribution is *concentrated* about its mean as the number of trials tends to ∞. Suppose we have i.i.d. trials, each with a probability of success $1/2$. Let S_n be the number of successes in the first n trials (n is a positive integer), and define

$$Z_n := \frac{S_n - n/2}{\sqrt{n}/2}.$$

(a) What are the mean and variance of Z_n?

(b) What is the distribution of Z_n as $n \to \infty$?

(c) Use the bound $\mathbb{P}[Z > z] \leq (\sqrt{2\pi}z)^{-1} \mathrm{e}^{-z^2/2}$ when Z is normally distributed in order to bound $\mathbb{P}[S_n/n > 1/2 + \delta]$, where $\delta > 0$.

Solution:

(a) 0 and 1, respectively. We made them so, in order to apply the CLT. Here are the computations.

$$\mathbb{E}[Z_n] = \frac{1}{\sqrt{n}/2} \mathbb{E}\left[S_n - \frac{n}{2}\right] = \frac{1}{\sqrt{n}/2}\left(\mathbb{E}[S_n] - \frac{n}{2}\right) = 0,$$
$$\operatorname{var} Z_n = \frac{1}{n/4} \operatorname{var}\left(S_n - \frac{n}{2}\right) = \frac{1}{n/4} \operatorname{var} S_n = 1,$$

since $S_n \sim \text{Binomial}(n, 1/2)$.

(b) The CLT tells us that $Z_n \to \mathcal{N}(0,1)$.

(c) In order to apply the bound, we must apply it to Z_n.

$$\mathbb{P}\left[\frac{S_n}{n} > \frac{1}{2} + \delta\right] = \mathbb{P}\left[\frac{S_n - n/2}{n} > \delta\right] = \mathbb{P}\left[\frac{S_n - n/2}{\sqrt{n}/2} > 2\delta\sqrt{n}\right] \approx \mathbb{P}[Z_n > 2\delta\sqrt{n}]$$
$$\leq \frac{1}{2^{3/2}\delta\sqrt{\pi n}} \mathrm{e}^{-2\delta^2 n}$$

3 Correlation and Independence

(a) What does it mean for two random variables to be uncorrelated?

(b) What does it mean for two random variables to be independent?

(c) Are all uncorrelated variables independent? Are all independent variables uncorrelated? If your answer is yes, justify your answer; if your answer is no, give a counterexample.

Solution:

(a) Recall that for two random variables X and Y,

$$\text{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Two random variables are uncorrelated iff their covariance is equal to zero. If X and Y are uncorrelated, then there is no linear relationship between them.

(b) Recall that two random variables X and Y are independent if and only if the following criteria are met (the three criteria are equivalent and connected by Bayes rule):

$$\mathbb{P}(X = x \mid Y = y) = \mathbb{P}(X = x)$$

$$\mathbb{P}(Y = y \mid X = x) = \mathbb{P}(Y = y)$$

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$$

for all x, y such that $\mathbb{P}(X = x), \mathbb{P}(Y = y) > 0$.

If X and Y are independent, any information about one variable offers no information whatsoever about the other variable.

(c) Note that if two random variables are independent, they must have no relationship whatsoever, including linear relationships; therefore they must be uncorrelated. The converse, however, is not true: two uncorrelated variables may not be independent. Consider two variables X and Y that follow a uniform joint distribution over the points $(1,0),(0,1),(-1,0),(0,-1)$. See Figure 1. Then

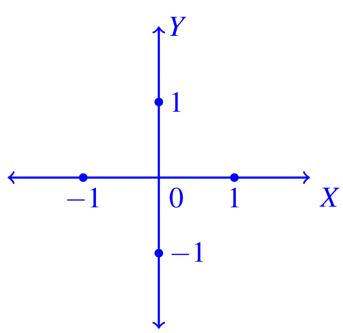

Figure 1: Choose one of the four points shown uniformly at random.

$$\text{cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) = 0.$$

To see why, observe that $XY = 0$ always because at least one of X and Y is always 0, and furthermore $\mathbb{E}[X] = \mathbb{E}[Y] = 0$ because both X and Y are symmetric around 0. So, there is no linear relationship, but X and Y are not independent (for example, $\mathbb{P}(Y = 0) = 1/2$ but $\mathbb{P}(Y = 0 \mid X = 1) = 1$).

4 Covariance

We have a bag of 5 red and 5 blue balls. We take two balls from the bag without replacement. Let X_1 and X_2 be indicator random variables for the first and second ball being red. What is $\text{cov}(X_1,X_2)$? Recall that $\text{cov}(X_1,X_2) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

Solution:

We can use the formula $\text{cov}(X_1,X_2) = \mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\,\mathbb{E}(X_2)$.

$$\begin{aligned}
\mathbb{E}(X_1) &= \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2}, \\
\mathbb{E}(X_2) &= \frac{5}{10} \times 1 + \frac{5}{10} \times 0 = \frac{1}{2}, \\
\mathbb{E}(X_1X_2) &= \frac{5}{10} \cdot \frac{4}{9} \times 1 + \left(1 - \frac{5}{10} \cdot \frac{4}{9}\right) \times 0 = \frac{2}{9}.
\end{aligned}$$

Therefore,

$$\mathbb{E}(X_1X_2) - \mathbb{E}(X_1)(X_2) = \frac{2}{9} - \frac{1}{2} \times \frac{1}{2} = -\frac{1}{36}.$$