Week 4 Tutorial

Wray Buntine, Kelvin Li, Hourieh Khalajzadeh and Caroline Gao

18 March 2019

Contents

%

% show answers \newcommand{\answer}[2]{#2} %hide answers

% TODO: % add precision and recall examples

This week, we will review what we learned in Week 4 and work with a few questions about expectations, entropy, Chebyshev's Inequality and week law of large numbers.

1 Expectations

Given a discrete distribution, the expected value of the RV:

$$E(X) = \sum_{x \in \mathcal{X}} xp(x)$$

For continuous RVs, replace the sum with an integral:

$$E(X) = \int x\, f(x) dx$$

Also, the variance is the expectation of the squared deviation of a random variable from its mean. It's normally calculated using:

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

1.1 Tubes

The lifetime in hours of electronic tubes is a random variable have a probability density function given by:

$$f(x) = a^2 x e^{-ax}\,, x \geq 0$$

Compute the expected lifetime of such a tube (Question 28 from Ross, Chap 4) (Hint: if you are having trouble with integral, calculate using WolframAlpha)

Also, compute the expected variance. Likewise, use WolframAlpha.

Answer:

$$E[X] = a^2 \int_0^\infty x^2 e^{-ax} dx = a^2 \Big[- \frac{e^{-ax}(2 + 2ax + a^2x^2)}{a^3} \Big]_0^\infty = \frac{2}{a}$$

$$E[X^2] = \frac{6}{a^2}$$

$$E[X^2] - E[X]^2 = \frac{6-4}{a^2} = \frac{2}{a^2}$$

1.2 Expected values

If E[X] = 2 and $E[X^2] = 8$, calculate (a) $E[(2 + 4X)^2]$ and (b) $E[X^2 + (X + 1)^2]$. (Question 32 from Ross, Chap 4)

Answer:

$$\begin{aligned} (a) &= E[4 + 16X + 16X^2] = 164 \\ (b) &= E[X^2 + X^2 + 2X + 1] = 21 \end{aligned}$$

The last step in the above uses linearity of expect values.

1.3 Expected value puzzle

Argue that for any random variable X, $E[X^2] \geq (E[X])^2$. When does one have equality? Also, verify this yourself using R by generating random numbers using the uniform distribution (Question 42 from Ross, Chap 4).

Answer:

$$0 \leq Var(X) = E[X]^2 - (E[X])^2.$$

Equality when the variance is 0 (that is, when X is constant with probability 1).

1.4 Twin triangle distribution

Optional question: You have a twin triangle distribution given by:

$$p(x) = \begin{cases} 0 & |x| > 2 \\ \frac{1}{2}(1 - ||x| - 1|) & |x| \leq 2 \end{cases}$$

Compute the standard deviation and mean absolute deviation of x. Why are they different?

Answer:

$$p(x) = \begin{cases} 0 & |x| > 2 \\ \frac{x}{2} + 1 & -2 \leq x \leq -1 \\ \frac{-x}{2} & -1 \leq x \leq 0 \\ \frac{x}{2} & 0 \leq x \leq 1 \\ \frac{-x}{2} + 1 & 1 \leq x \leq 2 \end{cases}$$

If you plot it, you will see it is symmetric about 0, so the mean is 0. No math required! Also, by symmetry, to calculate variance or mean deviation, we only need to do for one side, so we do the integral for positive x only, and then double it.

To calculate the variance, we calculate the following first:

$$\begin{aligned} E[X^2] &= 2\left(\int_0^1 x^2 \cdot \frac{x}{2} dx + \int_1^2 x^2 \left(\frac{-x}{2} + 1\right) dx\right) \\ &= \frac{7}{6} \\ Var(X) &= \frac{7}{6} \\ SD(X) &= 1.080 \\ MAD &= E[|X - E[X]|] \\ &= E[|X|] \\ &= 2\left(\int_0^1 x \cdot \frac{x}{2} dx + \int_1^2 x \left(\frac{-x}{2} + 1\right) dx\right) \\ &= 1 \end{aligned}$$

Again, the integrals could be done with WolframAlpha.

Why are the StdDev and MAD different? Well averaging the square makes the larger (more outlying) values contribute more to the average, so the StdDev measure is a bit larger than MAD. But it's only a bit larger because the bulk of the values are at a distance of 1 from the mean anyway.

1.5 Discrete case

Let X be a random variable such that

$$P(X=0)=\frac{1}{8},\ P(X=1)=\frac{1}{4},\ P(X=2)=\frac{3}{8},\ P(X=3)=\frac{1}{4}.$$

Find $E(X)$ and $Var(X)$.

Answer:

According to LOTUS ("law of the unconscious statistician"):

$$E(X)=\sum_{x=0}^{3} xP(X=x)=0\times\frac{1}{8}+1\times\frac{1}{4}+2\times\frac{3}{8}+3\times\frac{1}{4}=1.75$$

Also

$$E(X^2)=\sum_{x=0}^{3} x^2P(X=x)=0^2\times\frac{1}{8}+1^2\times\frac{1}{4}+2^2\times\frac{3}{8}+3^2\times\frac{1}{4}=4$$

Therefore, the variance if X can be calculated as:

$$Var(X)=Var(X^2)-Var(X)^2=4-1.75^2=0.9375$$

2 Chebyshev's Inequality

Question: Suppose that X is a random variable with mean and variance both equal to 20. What can be said about $P(0 <= X <= 40)$? (Question 55 from Ross, Chap 4)

Answer:

$$P(0 <= X <= 40)=1-P(|X-20|>20) >= 1-\frac{1}{20}$$

by Chebyshev's inequality.

3 Weak Law of Large Numbers

3.1 Demonstration

Question: The weak law of large numbers was first proven by Jakob Bernoulli in 1713. Interested readers can see Bernoulli's original proof. Here we would like you to test this theory with a simple example. Suppose that you have a random variable X sampled from the Uniform distribution between 0 and 1. Plot the mean of X, given different sample sizes (from 1 to 10^5). In this case, when would the mean of X become very close to the expected value?

Answer: We will use "ggplot2" to demonstrate the results of the simulation, please install it via following code: install.packages(ggplot2) You would also need to install the "scale" package to make our code work. If it is not installed on your computer, please install it using: install.packages(scale).

Note this is a sophisticated use of plotting tools in R. Wouldn't expect you to come up with this sort of thing yourself. You may go fishing on the internet though to get this sort of thing.

3.2 Proof

Optional question: Argue why the weak law of large numbers is true. HINT: use Chebyshev's inequality.

Answer: Proof using Chebyshev's inequality, assuming finite mean $E(X_i) = \mu$ and variance $Var(X_i) = \sigma^2$ for all i.i.d. random variables. Hence

$$Var(\bar{X_n}) = Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}.$$

The expected value $E(\bar{X_n}) = \mu$. By Chebyshev's inequality we get

$$P(\left|\overline{X}_n - \mu\right| \geq \frac{\varepsilon}{\sigma}\sigma) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

so the above probability equals to 0 as $n \to \infty$.

4 Code Lengths and Entropy

You have a probability distribution on six symbols, the letters 'A' to 'F'.

symbol	'A'	'B'	'C'	'D'	'E'	'F'
probability	0.025	0.22	0.095	0.31	0.15	0.20

4.1 Find a code

Question:

Note that with the Kraft inequality it was shown that a binary prefix code can be constructed with code lengths

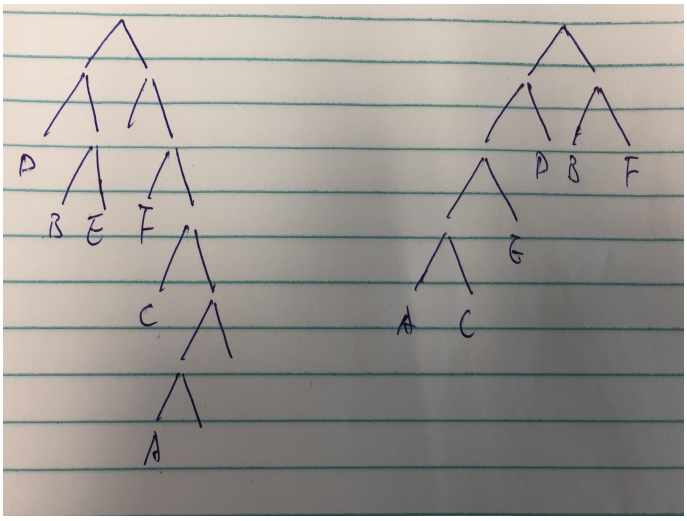
$$l_k = \left\lceil (log(\frac{1}{P_k}) \right\rceil$$

Draw the tree for such a prefix code.
Is this a shortest possible code? Can you shorten it?
Compute the expected code length for your shorter code and compare it with $H(letter)$, where *letter* is the event of getting one of 6 letters.

Answer: Given the probabilities, the length for each symbol is $6, 3, 4, 2, 3, 3$ respectively. A binary tree can be constructed as shown on the left in the following figure.

This is not the shortest code, as not all leaves are occupied. The shortest code is shown on the right of the figure. You may have something similar, but you should have at least tried to get rid of the empty leaves in the left figure.

The expected code length is $E[L] = 0.025 \times 4 + 0.22 \times 2 + 0.095 \times 4 + 0.31 \times 2 + 0.15 \times 3 + 0.2 \times 2 = 2.39$.
The entropy $H(p) = \sum_{k=1}^{6} p_i * \log_2(1/p_i) \approx 2.33$. So $E[L] < H(p) + 1$.



4.2 Conditional entropy

Let *vowel* be the event of receiving letter 'A' or 'E'. Compute $H(letter)$, $H(vowel)$ and $H(letter|vowel)$ and compare them. What do these tell you? Also, what is $H(vowel|letter)$?

Answer: $H(letter)$ was done before as 2.33. $p(vowel) = 0.175$ so $H(vowel) = 0.175 * \log_2 1/0.175 + 0.825 * \log_2 1/0.825 \approx 0.67$. $p(letter|vowel) = 0$ when the letter is not a vowel. Now $p('A'|vowel) = 0.025/0.175 = 0.143$ so $p('E'|vowel) = 0.857$. Thus $H(letter|vowel) \approx 0.59$. The issue is that $H(letter|vowel) \ll H(letter)$. This means being told a letter is a vowel informs you a lot about the letter.

$H(vowel|letter)$ must be 0 because *letter* determines *vowel*.