J Kohonen

3A Standard deviation and correlation

Class problems

Remember that *mean* is another name for expected value.

3A1 (Correlation versus dependence) Discrete random variables X and Y have the following joint distribution.

	Y		
X	-1	0	1
-1	0	$\frac{1}{6}$	$\frac{1}{6}$
0	$\frac{1}{3}$	0	0
1	0	$\frac{1}{6}$	$\frac{1}{6}$

(a) Determine the distribution, mean and standard deviation of X.

(b) Determine the distribution, mean and standard deviation of Y.

(c) Calculate the correlation between X and Y.

(d) Determine whether X and Y are (stochastically) dependent or independent.

3A2 (Average of dice) An ordinary die is rolled many times. The individual results are denoted $X_1, X_2, \ldots,$ and they are independent. The average of the first n results is denoted A_n.

(a) Find mean and standard deviation of X_1.

(b) Find the distribution of $A_2 = \frac{1}{2}(X_1 + X_2)$.

(c) Find mean and standard deviation of A_2.

(d) Find mean and standard deviation of

$$A_{100} = \frac{1}{100}(X_1 + X_2 + \cdots + X_{100}).$$

Hint: In (c), you can either calculate directly from the distribution of A_2, or you can apply linearity of expectation and covariance. In particular, observe that

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Cov}(X_1{+}X_2,\ X_1{+}X_2) = \mathrm{Var}(X_1) + 2 \cdot \mathrm{Cov}(X_1, X_2) + \mathrm{Var}(X_2).$$

How does this simplify when X_1 and X_2 are independent? In (d) it would be very laborious to find the exact distribution of A_{100}, so the previous formula is very convenient. How does it work for a sum of many random variables?

Home problems

3A3 (Predicting temperatures) A meteorologist is modelling the relation between today's temperature T_0 and tomorrow's temperature T_1 with the equation

$$T_1 \;=\; T_0 + \Delta T$$

where ΔT is a random variable indicating the change in temperature. The random variables T_0 and ΔT are assumed independent. Moreover, we know that $\mathrm{E}(T_0) = \mu$, $\mathrm{Var}(T_0) = \sigma^2$, $\mathrm{E}(\Delta T) = 0$, and $\mathrm{Var}(\Delta T) = \theta^2$. The model parameters μ, σ and θ are known (and $\sigma > 0$ and $\theta \geq 0$).

(a) Find $\mathrm{E}(T_1)$.

(b) Find $\mathrm{SD}(T_1)$. Recall from exercise 3A2 how you can calculate the variance of a sum.

(c) Find $\mathrm{Cov}(T_1, T_0)$. Use the linearity of covariance.

(d) Find $\mathrm{Cor}(T_1, T_0)$. Before you calculate it, try to guess, by thinking about the meaning of correlation, how the correlation should be if θ is small or zero, and if it is very large (much larger than σ).

The results should be formulas expressed in terms of the model parameters (μ, σ, θ).

3A4 (Minimizing loss functions) Eastham is a town that stretches along a straight road, two kilometers from west to east, so we model it as a line segment of length 2. There are more people living in the east end than in the west end; the location of a randomly chosen inhabitant is a random variable X that has density function $f(x) = x/2$, when $0 \leq x \leq 2$. (Although in reality the population would be finite, here we think there are so many inhabitants that we can treat them as a continuous mass.)

(a) Find the mean $\mu = \mathrm{E}(X)$, and the median m, which is a point such that $\mathrm{P}(X \leq m) = \frac{1}{2}$.

(b) Find $\mathrm{E}(X^2)$ and $\mathrm{SD}(X)$.

(c) Abel is a planner who wants to choose a location c for the town hall, somewhere in town. He tries to minimize the *quadratic loss* $q(c) = \mathrm{E}\left((X-c)^2\right)$. In other words, he tries to minimize the average of all inhabitants' *squared distances* to the town hall. Express q as a simple function of c. What is the shape of this function? Find the value of c where $q(c)$ is minimized. Is it one of the values μ and m?

(d) Bertha is another planner that wants to choose a location c for the library, somewhere in town. She tries to minimize the *linear loss* $\ell(c) = \mathrm{E}(\,|X-c|\,)$. In other words, she wants to minimize the average of all inhabitants' *distances* to the library. Express $\ell(c)$ as a simple function of c. Find the value of c where $\ell(c)$ is minimized. Is it one of the values μ and m? Compare the locations that Abel and Bertha chose, and try to give an intuitive explanation. Hint: Express the linear loss as an integral over the whole town. Then break it into two integrals, one for $x < c$ and one for $x \geq c$.

For concreteness this problem is stated as a problem of physical placement and distances. But there is a more statistical interpretation too; more about this in the model solution.