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 penalty) Abel and Bertha are working in different weather forecasting companies. They have both calculated a predictive distribution for X, the Celsius temperature tomorrow at noon, and they agree that it has the triangular-shaped density function $f(x) = 0.08\,x$ over the interval $[0, 5]$.

However, their companies and the general public won't hear anything about distributions. They want plain and simple *point predictions*. Abel has to pick a single point $a \in [0, 5]$ as his temperature prediction. Likewise Bertha has to pick a single point $b \in [0, 5]$. They can choose different points if they like.

(a) Verify by integrating that f is really a continuous density function.

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

(c) Abel's company wants to encourage good predictions, so his salary is reduced by a *quadratic penalty* $(X - a)^2$, in some convenient units of money; where X is the temperature observed tomorrow, and a is the point Abel chose. Since X is not known yet, Abel is interested in his *expected penalty* $q(a) = \mathrm{E}\left((X - a)^2\right)$. Simplify this to obtain q as a simple polynomial function of a, not containing any E signs.

Hint: One method is to start by expanding the square of the binomial. Another way is to start by writing the E as an integral.

Draw $q(a)$ over the interval $a \in [0, 5]$ to understand its shape.

Abel chooses his prediction a so that $q(a)$ is minimized. Find his prediction (actual numerical value). Is it one of the values μ or m?

(d) Bertha's company is using a *linear penalty*, $|X - b|$, in some convenient units of money, where X is the observed temperature and b is her prediction.

Bertha's expected penalty is $\ell(b) = \mathrm{E}(\,|X - b|\,)$. Write the expected value as an integral and simplify, to obtain her expected penalty as a simple function of b. Draw the function and find what Bertha should give as her temperature prediction b, so as to minimize $\ell(b)$. Is it one of the values μ or m?

Hint: Split the integral into two parts, one for $x \in [0, b)$ and one for $x \in [b, 5]$. Caution: Bertha's problem leads to more laborious integrals than Abel's, so be persistent and careful with them.

(e) If Abel and Bertha predicted different temperatures, why? Try to explain by common sense. How do the different penalties affect their choices?

(f) (Optional Challenge, More Difficult) If you identified Abel's and Bertha's predictions numerically matching something in the distributions, try to prove that the same would be true for *any* density function f.

This is an example of *facility location* (`https://en.wikipedia.org/wiki/Facility_location_problem`) where one has to choose a point so as to minimize some expected penalty, or "cost function". The same mathematical form appears with concrete facilities, such as warehouses to be placed near the users that are distributed over the space. One might want to minimize the *average* distance of all users to the warehouse. Optional challenge: Try to see how this is mathematically equivalent to Bertha's problem. Think why one might instead use a different cost function that matches Abel's problem.