I. Markov and Chebyshev

Markov's Inequality
For a non-negative random variable X with expectation E(X) = μ, and any α > 0:

$$\Pr[X \geq \alpha] \leq \frac{\mathrm{E}(X)}{\alpha}$$

Chebyshev's Inequality
For a random variable X with expectation E(X) = μ, and any α > 0:

$$\Pr[|X - \mu| \geq \alpha] \leq \frac{\mathrm{Var}(X)}{\alpha^2}$$

Use Markov's to prove Chebyshev's Inequality:

A random variable X always takes on values greater than -60. Find the best bound possible for Pr[X>= -10] when E[X] = -35.

Tossing Coins

Consider a coin that comes up with head with probability 0.2 . Let us toss it n times. Use Markov's to bound the probability of getting 80 percent heads.

Squirrel Standard Deviation

As we all know, Berkeley squirrels are extremely fat and cute. The average squirrel is 40% body fat. The standard deviation of body fat is 5%. Provide an upper bound on the probability that a randomly trapped squirrel is either too skinny or too fat? A skinny squirrel has less than 27.5% body fat, and a fat squirrel has more than 52.5% body fat?

Bound It!!!

A random variable X is always strictly larger than -100. You know that $E[X] = -60$. Give the best upper bound you can on $P(X \geq -20)$.

Give a distribution for a random variable where the expectation is 1,000,000 and the probability that the random variable is zero is 99%.

Consider a random variable Y with expectation μ whose maximum value is $3\mu/2$, prove that the probability that Y is 0 is at most 1/3.

II. Confidence Intervals

We know that the variables X_i, for *i* from 1 to *n*, are i.i.d. random variables and have variance . We also have a value (an observation) of $A_n = \frac{X_1 + \ldots + X_n}{n}$. We want to guess the mean, μ, of each X_i.

We prove that we have 95% confidence μ lies in the interval

$$\left[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}\right]$$

That is,

$$Pr\left[\mu \in \left[A_n - 4.5\frac{\sigma}{\sqrt{n}}, A_n + 4.5\frac{\sigma}{\sqrt{n}}\right]\right] \geq 95\%$$

To do this, we use Chebyshev's. Because $E[A_n] = \mu$ (A_n is the average of the X_i's), we bound the probability that $|A_n - \mu|$ is *more* than the interval size at 5%:

$$Pr\left[|A_n - \mu| \geq 4.5\frac{\sigma}{\sqrt{n}}\right] \leq \frac{\mathrm{Var}(A_n)}{(4.5\sigma/\sqrt{n})^2} \approx \frac{\sigma^2/n}{20\sigma^2/n} = \frac{1}{20} = 5\%$$

Thus, the probability that μ is *in* the interval is 95%.

Give the 99% confidence interval for μ:

We have a die whose 6 faces are values of consecutive integers, but we don't know where it starts (it is shifted over by some value *k*; for example, if *k* = 6, the die faces would take on the values 7, 8, 9, 10, 11, 12). If we observe that the average of the n samples (n is large enough) is 15.5, develop a 99% confidence interval for the value of *k*.

III. Covariance

The covariance of two random variables X and Y is defined as

$$\mathrm{Cov}(X,Y) := E((X - E(X))(Y - E(Y)))$$

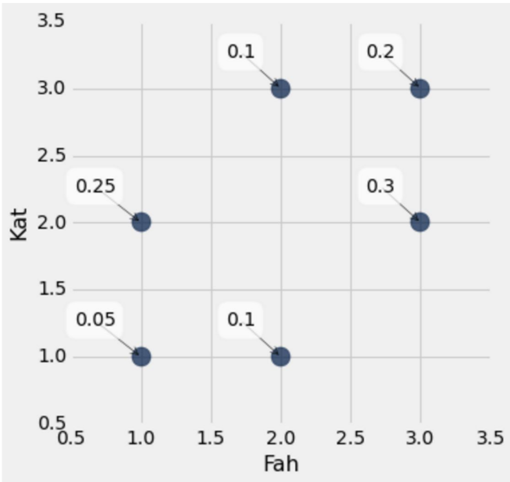
Prove that cov(X, X) = var(X):

Prove that if X and Y are independent, then cov(X, Y) = 0:

Prove that cov(X + Y, Z) = cov(X, Z) + cov(Y, Z):

1. Roll 2 dice. Let A be the number of 6s you get, and B be the number of 5s, find cov(A, B)

2. Consider the following distribution with random variables Fah and Kat:



Find the covariance of Fah and Kat.

IV. Conditional Expectation

> *The conditional expectation of Y given X is defined by*
>
> $$E[Y|X=x] = \sum_y y P[Y=y|X=x] = \sum_y y \frac{P(X=x, Y=y)}{P(X=x)}.$$

Prove E[E[Y|X]] = E[Y]

Prove E[h(X)Y|X] = h(X)E[Y|X]

V. Linear Least Squares

Linear Least Squares Estimate (LLSE)

$$L[Y|X] = E(Y) + \frac{cov(X,Y)}{var(X)}(X - E(X)).$$

Let X, Y be i.i.d. Uniform(-1,1). Calculate L(Y|Y+2X). [Var(X) = 2Var(U(0,1))]

Let X, Y, Z be i.i.d. N(0,1) and V= 2X + 3Y + 4Z, W= X+Y+Z. Find L[V|W].

VI. Continuous Probability -- Katya

1. Given the following density functions, identify if they are valid random variables. If yes, find the expectation and variance. If no, what rules does the variable violate?
    a. $f(x) = \frac{1}{4}$ on $(\frac{1}{2}, 9/2)$, $= 0$ elsewhere

    b. $f(x) = x - \frac{1}{2}$ on (0, $$\infty$$)

2. For a discrete random variable X we have Pr(X within [a, b]) that we can calculate directly by finding how many points in the probability space fall in the interval and how many total points are in the probability space. How do we find Pr(X within [a, b]) for a continuous random variable?

3. Are there any values of a, b for which we have a valid pdf? If not, why? If yes, what values?

    $f(x) = -1 \quad a<x<b$

    $f(x) = 0 \quad a<x<b$ (Are there any values of a,b for which we have a valid pdf?)

    $f(x) = 10000, \; 0<x<a$ (Are there any values of a for which we have a valid pdf?)

4. For what values of the parameters are the following functions probability density functions? What is the expectation and variance of the random variable that the function represents?

   $f(x) = ax,\ 0<x<1,\ f(x) = 0$ otherwise

   $f(x) = -2x,\ a<x<b,\ (a=0$ OR $b=0),\ f(x) = 0$ otherwise

   $f(x) = c,\ -30<x<-20,\ -5<x<5,\ 60<x<70,\ f(x) = 0$ otherwise

5. Define a continuous random variable R as follows: we pick a random point on a disk of radius 1; the value of R is distance of this point from the center of the disk. We will find the probability density function of this random variable.
   a. What is (should be) the probability that R is between 0 and ½? Why?

   b. What is (should be) the probability that R is between a and b, for any $0 \leq a \leq b \leq 1$?

   c. What is a function f(x), for which $\int_a^b f(x)dx$ satisfies these same probabilities?

   d. Define $g(x)$, the probability density function for R.

VI. Distributions (Uniform, Exponential, Gaussian=normal, Zipf) (longest)-- Alex T, Corrina, Anwar

There are certain jellyfish that don't age called hydra. The chances of them dying is purely due to environmental factors, which we'll call *lambda*. On average, 2 hydras die within 1 day.

What is the probability you have to wait 5 days for a hydra dies?

(a) Let X and Y be two independent *discrete* random variables. Derive a formula for expressing the distribution of the sum $S = X + Y$ in terms of the distributions of X and of Y.

(b) Use your formula in part (a) to compute the distribution of $S = X + Y$ if X and Y are both discrete and uniformly distributed on $\{1,...,K\}$.

(c) Suppose now X and Y are *continuous* random variables with densities f and g respectively (X,Y still independent). Based on part (a) and your understanding of continuous random variables, give an educated guess for the formula of the density of $S = X + Y$ in terms of f and g.

(d) Use your formula in part (c) to compute the density of S if X and Y have both uniform densities on $[0, a]$.

(e) Show that if X and Y are independent normally distributed variables, then $X + Y$ is also a normally distributed variable.