CSM 70 Discrete Mathematics and Probability Theory

Spring 2016 Confidence Int., Covar, Cond. Expectation, LLSE, Continuous **Worksheet 12 Solutions**

Please finish this by tonight (I will print them tomorrow morning)

I. Confidence Intervals -- Ni(c)k(hil)

Walk through proof of 95% confidence interval

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%:

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

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

Give the 99% confidence interval for μ:

$\left[A_n - 10\frac{\sigma}{\sqrt{n}}, A_n + 10\frac{\sigma}{\sqrt{n}}\right]$, because

$$Pr\left[|A_n - \mu| \geq 10\frac{\sigma}{\sqrt{n}}\right] \leq \frac{\text{Var}(A_n)}{(10\sigma/\sqrt{n})^2} \approx \frac{\sigma^2/n}{100\sigma^2/n} = \frac{1}{100} = 1\%$$

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*.

Nikhil is writing the solution for this.

II. Covariance -- Anwar

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

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

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

cov(X, X) = E(XX) - E(X)E(X) = $E(X^2) - E(X)^2$

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

cov(X, Y) = E(XY) - E(X)E(Y)

Remember that a property of expectation is that if X and Y are independent, then E(XY) = E(X)E(Y), so we get 0 when we subtract

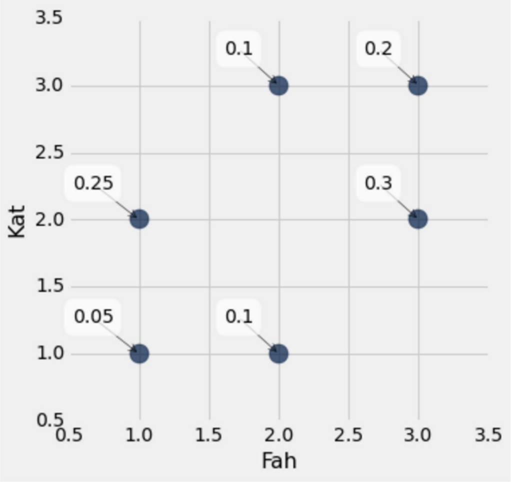
Prove that cov(X + Y, Z) = cov(X, Z) + cov(Y, Z):

cov(X + Y, Z) = E((X + Y)Z) - E(X + Y)E(Z)
= E(XZ) + E(YZ) - (E(X)E(Z) + E(Y)E(Z))
= E(XZ) - E(X)E(Z) + E(YZ) - E(Y)E(Z)
= cov(X, Z) + cov(Y, Z)

Problems:

1. Roll 2 dice. Let A be the number of 6s you get, and B be the number of 5s, find cov(A, B)
   E(A) = ⅓, E(B) = ⅓
   E(AB) = 1/18
   cov(AB) = -1/18

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



Find the covariance of Fah and Kat.

E(Fah) = 1 * .3 + 2 * .2 + 3 * .5 = 1.9

E(Kat) = 1 * .15 + 2 * .55 + 3 * .3 = 2.15

E(KatFah) = 1*1 * .05 + 1*2 * .25 + 2*1 * .1 + 2*3 * .1 + 3*2 * .3 + 3*3 * .2 = 4.95

cov(Kat, Fah) = 4.95 - 1.9 * 2.15

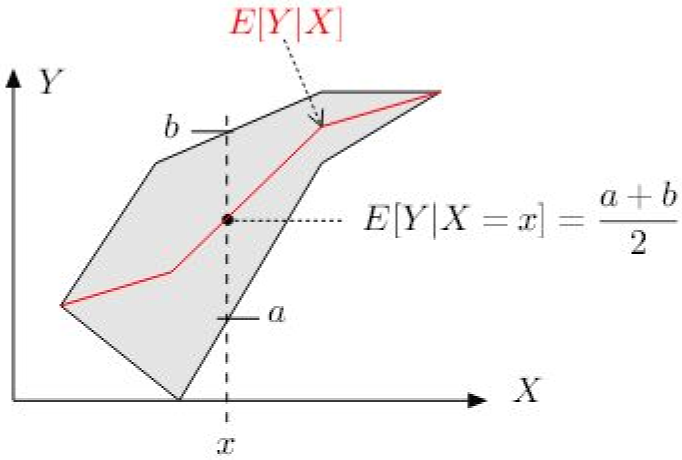
= 0.865

III. Conditional Expectation -- Albert

The conditional expectation of Y given X is defined by

$$E[Y|X=x] = \sum_y yP[Y=y|X=x] = \sum_y y\frac{P(X=x,Y=y)}{P(X=x)}.$$

Draw a picture to show that conditioning creates a new random variable with a new distribution. Figure 9 of note 26 does so.



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

Solution:

$$\begin{aligned}
E[E[Y|X]] &= \sum_x E[Y|X=x]Pr[X=x] \\
&= \sum_x (\sum_y yPr[Y=y|X=x])Pr[X=x] \\
&= \sum_x \sum_y yPr[Y=y|X=x]Pr[X=x] \\
&= \sum_y y \sum_x Pr[X=x|Y=y]Pr[Y=y] \\
&= \sum_y yPr[Y=y] \sum_x Pr[X=x|Y=y] \\
&= \sum_y yPr[Y=y] = E[Y]
\end{aligned}$$

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

Solution:

$$\begin{aligned}
E[h(X)Y|X] &= \sum_y h(X)yPr[Y=y|X] \\
&= h(X)\sum_y yPr[Y=y|X] \\
&= h(X)E[Y|X]
\end{aligned}$$

IV. Linear Least Squares -- Erik:

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))]

L(Y|Y+2X) = E(Y) + (cov(Y+2X, Y)/var(Y+2X))*(Y+2X-E(Y+2X))

(E(Y) = 0)

= 0 + E((Y+2X)*Y) - (E(Y+2X)E(Y)/(E($(Y+2X)^2$)-0)(Y+2X-E(Y+2X))

= (E(Y^2+2XY)/E(Y^2+4XY + 4X^2))(Y+2X)

= E(Y^2)/(E(Y^2)+4E(X^2))

= $_{-1}S^1y^2f_y(y)dy/(_{-1}S^1y^2f_y(y)dy + 4\,_{-1}S^1x^2f_x(x)dx)$

= $y^3/3\ |^1_{-1}$ = (⅓ + ⅓)/((⅓ + ⅓) + 4(⅔) = ⅕ * (Y+2X)

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

L[V|W] = E(V) + ((cov(V, W)/Var(W))(W-E(W))

E(V) = E(2X + 3Y + 4Z) = 2E(X) + 3E(Y) + 4E(Z) = 0 + 0 + 0

E(W) = E(X+Y+Z) = 0+0+0 = 0

Cov(V,W) = E(VW) - E(V)E(W)

= E(VW) = E((2X+3Y+4Z)(X+Y+Z))

= E(2X^2+2XY+2XZ + 3YX + 3Y^2 + 3YZ + 4ZX + 4ZY + 4Z^2) = E(2X^2 + 3Y^2 + 4Z^2) = 2E(X^2) + 3E(Y^2) + 4E(Z^2)

[E(X^2)=E(Y^2)=E(Z^2)= Var(N(0,1)) + E(N(0,1)) = 1 + 0 = 1]

= 2*1+3*1+4*1 = 9

Var(W) = Var(X+Y+Z) = (by independence) Var(X) + Var(Y) + Var(Z) = 1+1+1 = 3

E(V) + ((cov(V, W)/Var(W))(W-E(W)) = 0 + (9/3)(W-0)= 3W

V. Continuous Probability -- Katya

Given some different density functions, are these valid RVs? If yes, find expectation and Variance. If not, what rules does it violate?
Given some expectation and the bounds, find the bounds of the variable

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) = ¼ on (½, 9/2), = 0 elsewhere
        i. Expectation: 5/2
        ii. Variance: 4/3
    b. f(x) = x - ½ on (0, $$\infty$$)
        i. Has negative values on (0, ½)
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?

For a continuous RV with probability density function f(x), the probability that X takes on a value between a and b is the area under the pdf from a to b, which is the integral from a to b of f(x).

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

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

No. $f(x) >= 0$ must be true.

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

No. $S^b_a\ 0 = 0$ for all a,b.

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

Yes, $S^a_0\ 10000 = 1 = 10000a - 0 = 1 \Rightarrow a = 1/10000$

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

For a function to represent a probability density function, we need to have that the integral of the function from negative infinity to positive infinity to equal 1 and for f(x) to be greater than or equal to 0. So we need integral over (-inf, inf) of f(x) = 1

= integral over (0, 1) of ax

$= (ax^2)/2\ |_0^1 = 1 \Leftrightarrow a/2 - 0 = 1 \Leftrightarrow a = 2$

For RV Y with pdf = f(x),

E(Y) = integral over (-inf, inf) of x*f(x)

=integral over(0,1) of $x*(2x) = 2x^3/3\ |^1_0 = ⅔ - 0 = ⅔$

Var(Y) = integral over (-inf, inf) of $x^2*f(x)$ - E(Y) = integral over (0, 1) of x^2*2x

= integral over(0,1) of $2x^3 = x^4/2\ |^1_0 = ½ - 0 = ½$

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

Again we need $f(x) >= 0$, so here a, b <= 0, so b=0.

Then $S^0_a\ f(x) = 1 = S^0_a\ -2x = -2x^2/2\ |^0_a = 0 - (-2a^2/2) = 2a^2/2 = 1$

$\Leftrightarrow a^2 = 1 \Leftrightarrow a = +- 1 \Rightarrow a = -1$

For RV Y with pdf = f(x),

$E(Y) = S^{inf}_{-inf}\ x*f(x) = S^0_{-1}\ x*(-2x) = -2x^3/3\ |^0_{-1} = 0 - (-2(-1)^3/3) = -⅔$

$Var(Y) = S^{inf}_{-inf}\ x^2*f(x) = S^0_{-1}\ x^2*(-2x) = -x^4/2\ |^0_{-1} = 0 - (-(-1)^4)/2 = ½$

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

We need $S^{inf}_{-inf}\ f(x) = 1$ and $f(x) >= 0$. So $c>=0$.

$S^{inf}_{-inf}\ f(x) = 1 = S^{-20}_{-30}\ c + S^5_{-5}\ c + S^{70}_{60}\ c = cx|^{-20}_{-30} + cx|^5_{-5} + cx|^{70}_{60}$

$= 10c + 10c+ 10c = 30c = 1 \Rightarrow c = 1/30$

For RV Y with pdf = f(x),

Don't worry too much about calculations, but you should be able to set up the equations

$E(Y) = S^{inf}_{-inf}\ x*f(x) = S^{-20}_{-30}\ xc + S^5_{-5}\ xc + S^{70}_{60}\ xc = x^2c/2|^{-20}_{-30} + x^2c/2|^5_{-5} + x^2c/2|^{70}_{60}$

$= (-30)^2c/2 - (-20)^2c/2 + 5^2c/2 - (-5)^2c/2 + 70^2c/2 - 60^2c/2 = 900c = 900/30 = 30$

$Var(Y) = S^{inf}_{-inf}\ x^2f(x) = S^{inf}_{-inf}\ x^2*f(x) = S^{-20}_{-30}\ x^2c + S^5_{-5}\ x^2c + S^{70}_{60}\ x^2c$

$= x^3c/3|^{-20}_{-30} + x^3c/3|^5_{-5} + x^3c/3|^{70}_{60} = (-30)^3c/3 - (-20)^3c/3 + 5^3c/3 - (-5)^3c/3 + 70^3c/3 - 60^3c/3$

$= 108250c/3 = 1202.77…$

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?

   ¼, because the area of the circle with distance between 0 and ½ is ($\pi(½)^2 = \pi/4$), and the area of the entire circle is

b. What is (should be) the probability that R is between *a* and *b*, for any $0 \leq a \leq b \leq 1$?
The area of the region containing these points is the area of the outer circle minus the area of the inner circle, or $\pi b^2 - \pi a^2 = \pi(b^2 - a^2)$. The probability that a point is within this region, rather than the entire circle, is $\pi(b^2 - a^2) / \pi = b^2 - a^2$.

c. What is a function f(*x*), for which $\int_a^b f(x)dx$ satisfies these same probabilities?
$f(x) = 2x$, because $\int_a^b f(x)dx = [x^2]_a^b = b^2 - a^2$

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

$$g(x) = \begin{cases} 2x, & \text{if } x \in [0,1] \\ 0, & \text{otherwise} \end{cases}$$

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.

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

(b) 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*.

P(S = m) = \sum\limits_{i=-\infty}^{\infty}P(X=i)P(Y=m-i)

(c) 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*}.

P(S = m) = \sum\limits_{i=0}^{m}(1/K)(1/K) = m/K^2
We only care if if the probabilities are non-zero

(d) 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*.

h(t) = \int\limits_{-\infty}^{\infty}f(s)g(t-s)ds

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

h(t) = \int\limits_{0}^{t}f(

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