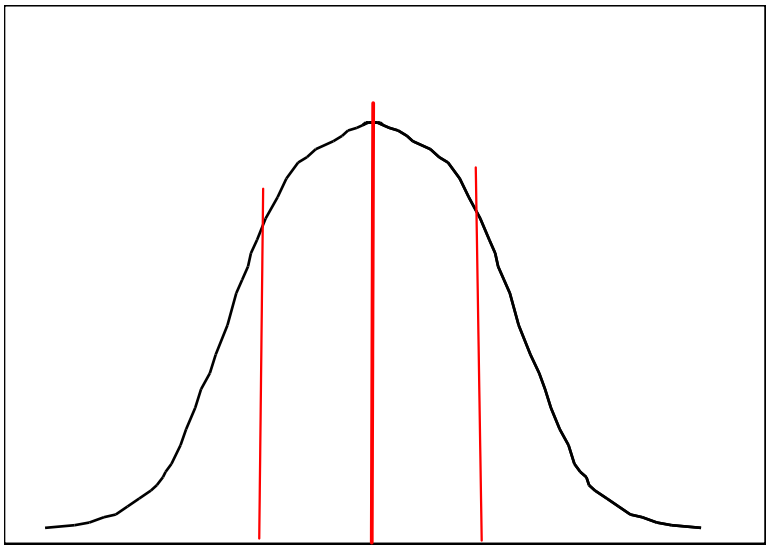
Statistics and Bioinformatics -- Problem Set 4
Due in class Tuesday, November 9

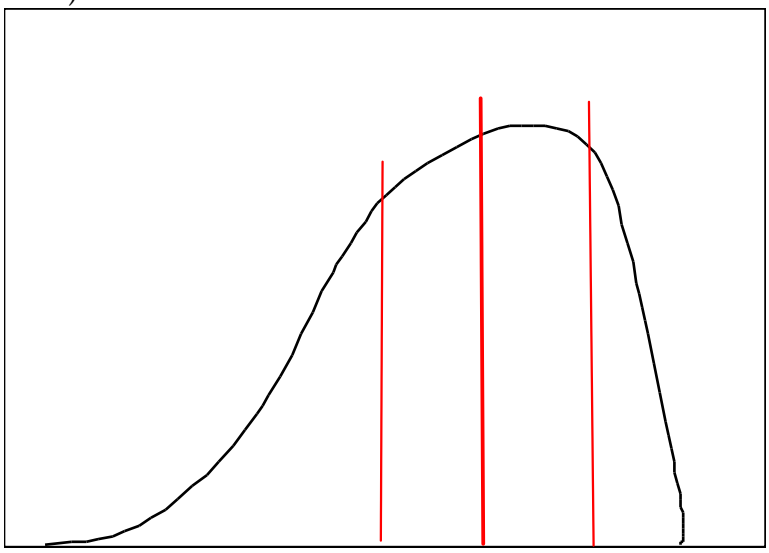
Exercises

Probability Distributions

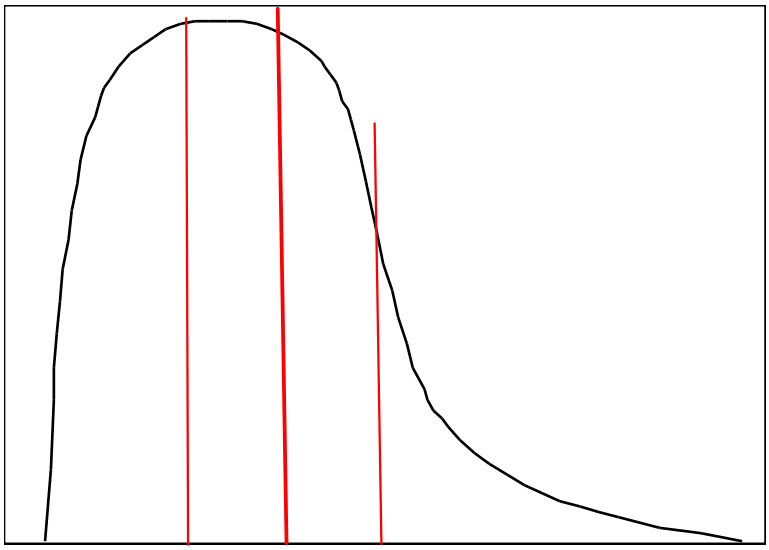
1) For each of the probability distribution graphs below [horizontal axis x, vertical axis y or $f(x)$], indicate the approximate mean and standard deviation on the x axis. Remember the mean is the balance point for the entire distribution.



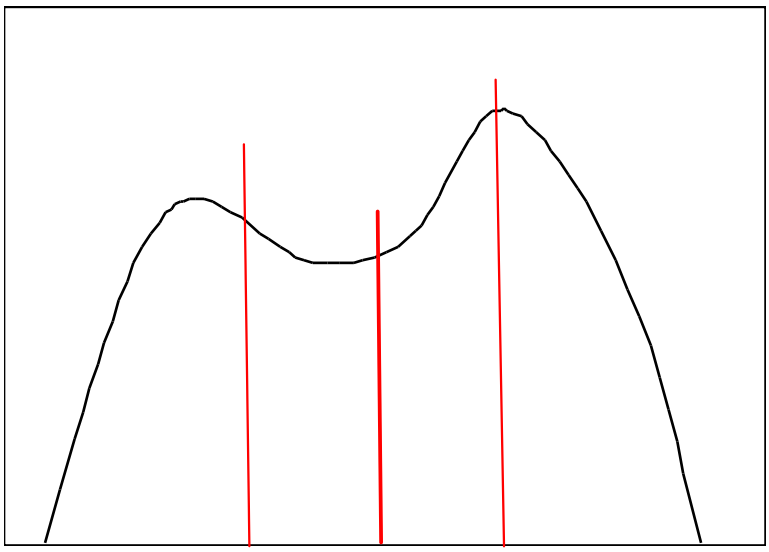
a)



b)



c)



d)

2) The table below contains x and $P(x)$ for all possible values of x. From these values, calculate:

x	$P(x)$
0	0.30
1	0.40
2	0.20
3	0.07
4	0.03

a) the mean of x

$\bar{x} = (0.3)(0) + (0.4)(1) + (0.2)(2) + (0.07)(3) + (0.03)(4)$

$= 0.4 + 0.4 + 0.21 + 0.12 = 1.13$

b) the variance of x

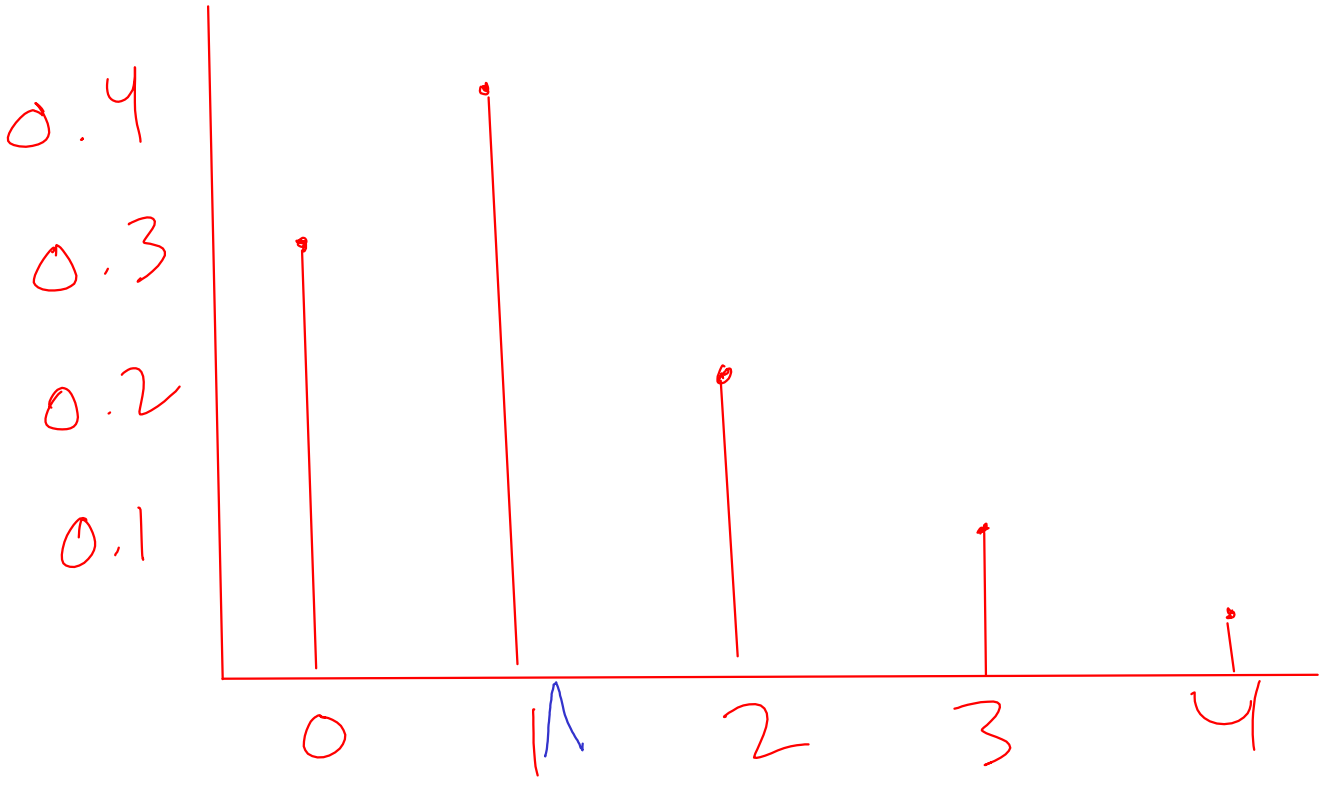
$Var(x) = (0-1.13)^2(0.3) + (1-1.13)^2(0.4) + (2-1.13)^2(0.2) + (3-1.13)^2(0.07) + (4-1.13)^2(0.03)$

$= 0.383 + 0.00676 + 0.151 + 0.245 + 0.247 = 1.0331$

c) the standard deviation of x

$\sigma(x) = \sqrt{Var(x)} = \sqrt{1.0331} = 1.016$

d) Make a plot of the probability distribution (by hand or using R, your choice), and verify that the mean and standard deviation are close to what you would predict by visual judging of the balance points.



Poisson Distribution

3) Describe the following, using the term "probability":
   a) the independence assumption of the Poisson distribution.

The occurrence of an event has no effect on the probability of its occurrence at any other place or time.

   b) the constant mean assumption of the Poisson distribution.

Each increment of space and time has the same probability of occurrence of the event.

4) For each example below, state whether it is an example of a Binomial variable, Poisson variable, or neither. Assume the assumptions of independence and constant probability or mean are met in all cases. If you think the variable is neither, briefly explain why.
   a) the number of females in litters of six kittens B
   b) the number of females in litters of six kittens divided by the number of males neither, not a count
   c) the number of times a person is stung by a bee during the summer P
   d) the number of yellow pea seeds produced by a cross between green and yellow pea plants B assuming only 2 colors
   e) the number of sedge seedlings per square meter of salt marsh P
   f) the number of buras last year in Starigrad P
   g) the number of buras last year in Starigrad, divided by the total number of buras during the last decade neither, not a count
   h) the number of times an individual catches a cold per year P
   i) the number of times a person has broken a bone during their lifespan P

5) Now do *not* assume that the assumptions of independence or constant probability or mean are met. Which of the following examples do you think are Poisson distributed, and which are not? If you think an example is not, briefly explain which assumption(s) you think are violated.

a) the number of colds caught per individual U Zadar student last year independence
b) the number of car accidents within the Zadar city limits, over the last 50 years. constant probability
c) the number of rattlesnakes within a particular hectare of land in Zadar county at any given instant of time. both could be violated
d) the number of tuna individuals within a particular volume of water in the Adriatic at any given instant of time. both could be violated
e) the number of aphids attacking a tomato plant in a field at any given time. both
f) the total number of mutations per genome per generation in seeds of a corn plant of a particular variety. likely poisson

6) Briefly describe an example of the Poisson distribution from a biological area of your interest. Specifically explain why you think the independence and constant mean assumptions are met.

The number of mutations within a defined region of the genome usually satisfies the assumptions; mutations at one region of the genome generally don't affect the mechanism of mutation at other regions, and the same mechanisms operate in all individuals.

The Normal Distribution

7) Which of the following variables follows a normal distribution (exactly or approximately)? Hint: in R you could try plot(dbinom()) or plot(dpois()) depending on the distribution below.
   a) The number of heads in 1000 fair coin flips. almost exactly
   b) The number of heads in 2 fair coin flips. not at all
   c) The number of colds per year per individual, where the mean is 5. approx. if poisson
   d) The number of car accidents per year per individual, where the mean is 0.02. not at all
   e) The number of bone-break incidents over a 20-year period per individual, where the mean is 0.66. not at all
   f) The number of mutations per genome per generation in Arabidopsis plants, where the mean is 100. yes if poisson
   g) The number of yellow seeds in a count of 8000 seeds total, where the probability that one seed is yellow is 1/4. yes if binomial
   h) The weight of a single U Zadar student maybe approximately
   i) The mean weight of 100 U Zadar students yes almost exactly
   j) The sum of all weights of 100 U Zadar students yes almost exactly

Problems (Poisson and normal distributions)

8) A famous example of the Poisson distribution is data by von Bortkiewicz (1898) showing the number of Prussian cavalrymen killed by a horse-kick per corps during one year. A total of 200 corps was studied, and the number of corps experiencing 0, 1, 2, 3 ... horse-kick deaths in the year was tallied, as shown in the table below.
   a) What is the observed mean number of deaths per corps per year?

$$\bar{x} = \sum x f(x) = 0.61$$

   b) What is the observed variance in the above?

$$Var(x) = \sum (x - \bar{x})^2 f(x) = 0.608$$

c) Fill in the column for "Expected Frequency" based on the Poisson distribution using the mean calculated in (a). This can be calculated from dpois() in R, or you can use the formula P(k) = lambda^k exp(-lambda)/k!, where lambda is the mean obtained in a) above and k is the number of deaths per year whose probability you want to calculate.

Number of deaths per corps per year	Observed Frequency	Expected Frequency
0	109	109
1	65	66.3
2	22	20.2
3	3	4.1
4	1	0.6
5	0	0.1
6	0	0.01

dpois(0:6, 0.61)(200)

or

0.54
0.33
0.10
0.021
0.0031
0.00038
0.000039

dpois(0:6, 0.61)

d) Do you think this variable follows the Poisson distribution? Briefly explain your answer.

Looks almost exactly a Poisson distribution since the mean and variance are almost equal

e) Describe two reasons why one might have predicted (before seeing the numbers) that this variable is not Poisson distributed.

Constant probability assumption might be violated; some corps might have a murderous horse, others not. Or some corps might be less experienced with horses than others, in a way that influences kicking accidents. Or the independence assumption might be violated: one kicking horse may cause other horses to kick.

f) A new corps is formed that experiences 3 horse-kick deaths in one year. Is this observation "unusual"? Why or why not?

This is unusual because its probability is 0.02, less than 0.05

9) Assume the probability that a newborn baby in a particular inner city hospital is HIV positive is 0.008.

a) If 500 babies from this hospital are randomly sampled, what is the binomial probability that exactly 5 will be HIV positive? (Hint: you can use the dbinom() function in R to calculate this).

dbinom(5, 500, 0.008) = 0.157

b) What is the Poisson approximation of the probability in (a)? (Hint: use the dpois() function in R, or the formula in c) in the last problem, using lambda = 0.008 * 500.)

0.008*500 = 4

dpois(5, 4) = 0.157

c) Is the Poisson approximation a good one for these data? Is the normal approximation a good one? Explain briefly why or why not.

Clearly the Poisson approximation is very good, we don't need to do the complex binomial calculations to get the probability (but don't need to anyway, using R). The normal will not be a very good approximation because the mean (4) is less than 5. Also, calculating a probability for 5 is difficult with the normal because it's a continuous distribution, not a discrete distribution.

pnorm(5.5, 4, sqrt(4)) - pnorm(4.6, 4, sqrt(4)) = 0.155

d) If 100 newborns are screened in this hospital, what are the expected mean, variance, and standard deviation of number HIV positive?

mean = 100 x 0.008 = 0.8
variance = 100 x 0.008 = 0.8
sd = sqrt(0.8) = 0.894

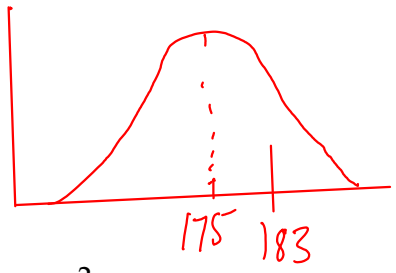
10) A particular group of 100 newborns in this hospital is screened based on the mother's socio-economic status, and 2 are found to be HIV positive. Is this observation "unusual" for this hospital? Why or why not?

1 - ppois(2, 0.8) = 0.047
This is unusual in a one-tailed test for an alpha of 0.05.

11) Calculate the *Z* score for your height, assuming a mean and standard deviation of 163 cm and 8 cm for women and 175 cm and 8 cm for men.

a) Draw a normal curve and indicate on this curve where your height lies.

$$Z = \frac{183 - 175}{8} = \frac{8}{8} = 1$$



b) What proportion of the population of your sex is taller than you?

1 - pnorm(1) = 0.16

c) What proportion is shorter?

$\text{pnorm}(1) = 0.84$

d) What proportion is closer to the mean than your height?

$1 - (2)(0.16) = 0.68$

e) What proportion of the opposite sex is taller than you?

$1 - \text{pnorm}(183, 163, 8) = 0.0062$

f) What proportion of the opposite sex is shorter than you?

$1 - 0.0062 = 0.9938$