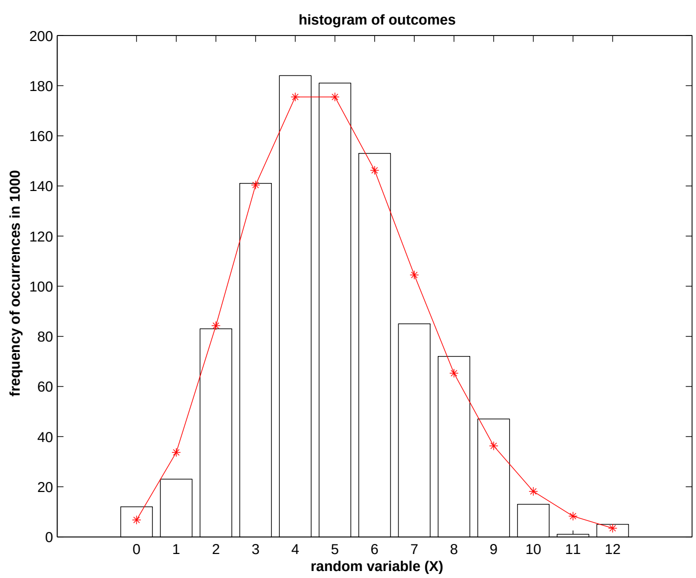
Homework #02 • MATH495/STAT490 • Probability Functions

- submit your write-up Wednesday 24 September.
- participation in webct discussions is encouraged.
- please respect page limits & remember to follow the *Guidelines for Reports*.
- highlight major results.

A) Down with Spam (2 pages, 10 pts) A simple spam filter might be based on the occurrence of certain features found in e-mail messages. Say that two types of e-mails hoped to be filtered out are virus-generated messages [V] and commercial spam [C] which typically comprise 10% and 60% of all messages. An e-mail census determines that the "$" character [$] appears in 20% of viral and 80% of commercial e-mails, in contrast to 10% for all other e-mail types. Likewise, the "W32.Vybab" string [W] appears in 90% of viral e-mails, yet never in any other types. Event labels are indicated by [·] — the probabilities in this problem are fictional and should not be construed as an accurate portrayal of reality.

- What is the meaning of P(W)? Find its value.
- What is the probability that a message containing "$" should be filtered?
- What is the probability that a message containing "W32.Vybab" should be filtered?



B) Conditional Simulation (3 pages, 10 pts) Modify the Matlab script *w03distr.m* to simulate the following random process. A discrete random vector $\vec{X} = (X, Y)$ has components which are independent and uniformly distributed on integers 1 through n. Based on a simulation, compute an empirical (discrete) probability function (EPF) for the second component Y conditioned on $X > Y$. That is, compute $EPF(b) = P(Y = b \,|\, X > Y)$.

Derive the theoretical $DPF(b)$ using a Bayes argument. Design appropriate graphical presentation. Comment on your experimental design.

C) Group Testing (2 pages, 10 pts) A large number of people N are subject to a blood test. The test can be administered in one of two ways:

(i) Each person is tested separately (N tests are needed in total).

(ii) A blood sample which is a mixture of k people's samples is tested. If the result is positive, each of the k samples is then retested separately ($k+1$ tests for this group). This procedure is followed N/k times so that all the blood has been tested.

Assume the probability p that the result is positive is the same for all people, and that the results have 100% repeatability. Also,the test results are independent for different people (no hereditary family issues, for example).

- What is the probability that the result for a mixed sample of k people is positive?
- What is the expected number of tests necessary under plan (ii)? (Hint: you can check against the attached plot.)
- Find the equation for the value of k_{min} which will minimize the expected number of tests, $ET(k)$, under plan (ii). (Minimization over real values. Hint: use $a^k = \exp(k \ln a)$.)
- Show that, when p is very small, an approximate solution for k_{min} is $1/\sqrt{p}$. By first Taylor expanding the exponential and logarithm (in small $\sqrt{p}$), show that the minimum expected number of tests, $ET(k_{min})$, is on average about $2N\sqrt{p}$. (This is a bit tricky – a webct discussion will likely be useful.)