Homework 2 – February 8, 2006 **Solution** prepared by Tobin Fricke
Physics 402 – Probability Prof. S.G. Rajeev, Univ. of Rochester

**4.1** *It is known that the O-negative blood type is found in* 7% *of people.* ***(a)*** *What is the probability that in a random sample of 30 people, the number with this type is more than* 3*?* ***(b)*** *What is the probability that there will be at least one person with this blood type in a random sample of* 15 *people?*

The survey may be regarded as a collection of Bernoulli trials, each with probability $p = 0.07$ of 'success'. Let the random variable $\xi$ be the number of successes.

**(a)** The numbers are small enough here that we can use the exact form of the binomial distribution, utilizing the fact that, because the events $\xi > 4$ and $\xi \leq 3$ are mutually exclusive, $P[\xi > 3] = 1 - P[\xi \leq 3]$:

$$p[\xi > 3] = 1 - p[\xi \leq 3] = 1 - p[\xi = 0] - p[\xi = 1] - p[\xi = 2]$$

$$p[\xi > 3] = 1 - \sum_{k=0}^{3} \binom{N}{k} p^k (1-p)^{N-k} = \boxed{0.154981\ldots}$$

**(b)**

$$p[\xi \geq 1] = 1 - p[\xi = 0] = \binom{N}{0}(1-p)^N = 1 - \frac{15!}{15!0!}(0.93)^{15} = 1 - (0.93)^{15} \approx \boxed{0.66}$$

One could have used the Poisson distribution to approximate the binomial distribution here, but there's little reason to do so.

**4.2** *The proportions of males and female in the human population are equal.* ***(a)*** *What is the probability that in a random sample of 100 people, more than 55 will be male?* ***(b)*** *If you find that there are only four women in a group of 15, with what degree of confidence can you conclude that the group was not selected in a gender-neutral way?*

A selection of people to compose a group can be considered a sequence of Bernoulli trials, with $p$ and $q$ being the probabilities of selecting each gender. Tbe combinatorics here, however, require that we use an approximation. Because $Np$ and $Nq$ are both appreciable, the appropriate approximation is the normal distribution. The normal distribution has the additional advantage of being continuous, so we may easily approximate the sum over possible group compositions as an integral.

**(a)** The mean of the binomial distribution is $\mu = Np$ and the variance is $\sigma^2 = Npq$. We approximate this distribution with the normal distribution with the same mean and variance. The probability density function for the normal distribution is:

$$\text{normal pdf}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Integrating this, we get the cumulative density function for the normal distribution:

$$p[\xi < x] = \text{cdf}(x) = \int_{-\infty}^{x} \text{pdf}(x')dx' = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{(x'-\mu)^2}{2\sigma^2}\right\} dx' = \frac{1}{2}\left(1 + \text{erf}\left\{\frac{x-\mu}{\sigma\sqrt{2}}\right\}\right)$$

Evaluating this numerically with $\mu = Np = (100)(0.5) = 50$ and $\sigma^2 = Np(1-p) = 25$ we find $p[\xi \geq 55] \approx p[\xi' > 55] = 1 - p[\xi' < 55] \approx 0.1587$, where $\xi'$ is a *normally distributed* random variable with the same mean and variance as the random variable $\xi$ with binomial distribution. However, there is one more detail: In approximating a discrete random variable with a continuous one, it is necessary to make a *continuity correction*:

$$p[\xi \geq 55] = 1 - p[\xi < 55] \approx 1 - p[\xi' < 54.5] \approx \boxed{0.184}$$

We may also solve this problem with only reference to a chart of the cumulative density function for a normal distribution with zero mean and unit variance by acting on our random variable with a linear transformation as in problem six on this assignment. The transformation is:

$$\xi \mapsto \xi'' = \frac{1}{\sigma}\xi - \frac{\mu}{\sigma} = \frac{1}{\sqrt{Np(1-p)}}\xi - \sqrt{\frac{Np}{1-p}} = \frac{\xi}{5} - 10$$

Then $p[\xi \geq 55] \approx p[\xi' \geq 54.5] \approx p[\xi'' \geq 0.9] \approx 0.184$.

The following Matlab code estimates this via numerical experiment:

```
% The parameters for our experiment
N = 100;
k = 55;
p = 0.5;
experiments = 50000;

% Generate an ensemble of random events
x = rand(N, experiments);

% Turn these into Bernoulli trials
x = x < p;

% Count the number of successes in each experiment
successes = sum(x);

% The fraction of experiments with k or more successes is:
length(find(successes >= k)) / length(successes)  % 0.18482

% Using the normal cumulative density function to approximate this:
1 - normcdf(k - 0.5, N*p, sqrt(N*p*(1-p)))        % 0.18406
```

**(b)** Similarly, the probability that a group of fifteen people, randomly chosen, will contain four or fewer women is $p[\xi \leq 4] \approx \text{normalcdf}(4.5)$ with mean $\mu = Np = 7.5$ and $\sigma^2 = Np(1-p) = 3.75$, approximately $\boxed{6\%}$. So it is very unlikely that this group was randomly selected (if such a group were composed 100 times, this situation would occur only about six times, etc).

**5. Linear random walk.** *Suppose a particle executes a biased random walk on a line...*

Each step in the random walk is a bernoulli trial. Let $\xi$ be the random variable giving the number of successes (=steps to the *left*) after $N$ steps; then the position is $-\xi l + (N-\xi)l = (N-2\xi)l$. If there are $n$ steps per unit time, we may write this as $x(t) = (nt - 2\xi)l = -2l\xi + ntl$. The expected position is $\langle x(t)\rangle = -2l\langle\xi\rangle + ntl = -2l(ntp) + ntl = ntl(1-2p)$. **(a)** Taking the derivative with respect to time, we find the expected velocity $\boxed{\langle v\rangle = nl(1-2p)}$. **(b)** The random variable giving position is a linear transformation of the random variable $\xi$ giving the number of successes. From question six we see that the variance of the random variable $\xi' = a\xi + b$ is simply $(\sigma')^2 = a^2\sigma^2$. We know the variance of $\xi$ is (from the binomial distribution) $Np(1-p)$, so the variance of $x$ is $\boxed{4l^2ntp(1-p)}$. **(c)** As the number of steps per unit time $n$ gets larger, so does the total number of trials for any given time, as $N = nt$. Therefore this is the limit in which the binomial distribution converges to the normal distribution.

**6. Linear transformation of a random variable.** *If a random variable $x$ is changed via the map $x \mapsto ax + b$ (where $a$ and $b$ are constants), how do the mean and variance change? What is the function of $x$ that has zero mean and unit variance?*

Because this is a linear transformation, we can solve the problem using just linearity of expectation. Let's label our new random variable as $x'$. First we find the mean $\mu'$ of $x'$ in terms of the mean $\mu$ of $x$:

$$\mu' = \langle x' \rangle = \langle ax + b \rangle = a\langle x \rangle + b = \boxed{a\mu + b}$$

It turns out that the mean transforms under the same linear transformation as the random variable. Now, the variance $\sigma^{2\,\prime}$ in terms of $\sigma^2$ (using $\langle x^2 \rangle = \sigma^2 + \mu^2$):

$$\sigma^{2\,\prime} = \left\langle x'^{\,2} \right\rangle - \langle x' \rangle^2 = \langle (ax+b)^2 \rangle - (\mu')^2 = \left\langle a^2x^2 + 2abx + b^2 \right\rangle - (a\mu + b)^2$$

$$= a^2 \left\langle x^2 \right\rangle + 2ab\langle x \rangle + b^2 - a^2\mu^2 - 2ab\mu - b^2 = a^2\sigma^2 + \cancel{a^2\mu^2} + \cancel{2ab\mu} + \cancel{b^2} - \cancel{a^2\mu^2} - \cancel{2ab\mu} - \cancel{b^2} = \boxed{a^2\sigma^2}$$

Summarising:

$$\boxed{\begin{array}{l} \mu \mapsto a\mu + b \\ \sigma^2 \mapsto a^2\sigma^2 \end{array}}$$

To find the linear map our random variable $x$ into $x'$ with zero mean and unit variance, we need to solve the system $\{a\mu + b = 0,\ a^2\sigma^2 = 1\}$. Clearly this has solution $\boxed{\{a = 1/\sigma, b = -\mu/\sigma\ \}}$.

We may verify this numerically:

```
% Generate an ensemble of random numbers
x = rand(10000,1);

% Just for fun, make the distribution more complicated
x = x.^3 - 2*x + log(x);

% find the sample mean and standard deviation
mu = mean(x)
sigma = std(x)

% Determine the parameters of our transformation
a = 1/sigma;
b = -mu/sigma;

% Transform to the ensemble of values with the new distribution
y = a*x + b;

% Find the mean and variance; these should be very close to 0 and 1.
mean(y)
std(y)^2
```