Lab #4: Distributions:

A. Objectives:

1. Compute probabilities for the Binomial Distribution
2. Compute probabilities for the Poisson Distribution
3. Explore Continuous Distributions- The Normal Distribution

The Binomial distribution in SAS:

As we learned in class, one of the fundamental discrete distributions is the Binomial distribution, which is built on repeating n trials with two outcomes, one of which is called success and the other failure, with P(success)=p. Suppose we want to simulate Milgram's experiment that we talked about in class with 20 people. Thus our random variable X follows the binomial distribution with n=20 trials and probability of "success" p=0.35. Let's answer the following questions:

1. What is the probability of 10 or fewer successes? $P(X \leq 10)$
2. What is the probability of 9 or fewer successes? $P(X \leq 9)$
3. What is the probability of exactly 10 successes? $P(X = 10)$
4. What is the probability of more than 10 successes? $P(X > 10)$
5. What is the probability of 8, 9 or 10 successes? $P(8 \leq X \leq 10)$

SAS has a built in function to compute the probabilities above, but instead of using the simple function for the binomial: $P(k\ successes\ in\ n\ trials) = \binom{n}{k} p^k (1-p)^{n-k}$ it uses what is called the ***cumulative probability function***. The cumulative probability function is nothing more than the sum (or integral for continuous distributions) of all the outcomes up to k. For example the cumulative function at 3 for our binomial distribution is:

$$F(3) = P(0\ successes\ in\ 20) + P(1\ success\ in\ 20) + P(2\ successes\ in\ 20) + P(3\ successes\ in\ 20)$$

The function that SAS uses is called **PROBBNML(p,n,k)** where p is the probability of success in one trial, n is the total number of trials and k is the number of successes.

So let's try to answer all our questions, one by one and store the outputs in a nice data table. The following code will answer question 1.

```
DATA binomial;
F_10 = PROBBNML(0.35, 20, 10);
RUN;
PROC PRINT data=binomial;
TITLE 'Binomial Probabilities';
RUN;
```

This code basically creates a dataset called "**binomial**" in which we store the results, and then computes the first variable using the function **PROBBNML** with parameters p=0.35, n=20 and k=10. It saves that variable as F_10 (to denote cumulative probability up to 10)

Similarly the code:

```
DATA binomial;
F_10 = PROBBNML(0.35, 20, 10);
F_9 = PROBBNML(0.35, 20, 9);
RUN;
PROC PRINT data=binomial;
TITLE 'Binomial Probabilities';
RUN;
```

will add the computation of the second probability we want, i.e. $P(X \leq 9)$ as a second variable named F_9 (to denote cumulative probability up to 9).
To compute the 3rd probability we need to notice that:

$$\boldsymbol{P(X = 10) = P(X \leq 10) - P(X \leq 9) = F(10) - F(9)}$$

so all we need to do is subtract the two numbers we just computed. We let SAS do that for us by adding the line: **P_10= PROBBNML(0.35, 20, 10)- PROBBNML(0.35, 20, 9);** which will compute the wanted difference and save it as variable P_10 (to denote actual probability).

```
DATA binomial;
F_10 = PROBBNML(0.35, 20, 10);
F_9 = PROBBNML(0.35, 20, 9);
P_10 = PROBBNML(0.35, 20, 10)- PROBBNML(0.35, 20, 9);
RUN;
PROC PRINT data=binomial;
TITLE 'Binomial Probabilities';
RUN;
```

To answer question 4 we need to remember that if two events are complementary then

$$\boldsymbol{P(A) + P(A^c) = 1 \Leftrightarrow P(A^c) = 1 - P(A)}$$

And notice that the events $X \leq 10$ and $X > 10$ are indeed complementary. So we just need to subtract the probability of $\boldsymbol{P(X \leq 10)}$ from 1. Again we will let SAS do that for us with the following line:
FC_10 =1- PROBBNML(0.35, 20, 10); which will compute the corresponding probability and save it as variable FC_10 (to denote the complement of the cumulative probability)

```
DATA binomial;
F_10 = PROBBNML(0.35, 20, 10);
F_9 = PROBBNML(0.35, 20, 9);
P_10 = PROBBNML(0.35, 20, 10)- PROBBNML(0.35, 20, 9);
FC_10 = 1- PROBBNML(0.35, 20, 10);
RUN;
PROC PRINT data=binomial;
TITLE 'Binomial Probabilities';
RUN;
```

Using similar logic to question 4, if you want to find the probability $P(8 \leq X \leq 10)$ you need to remember that:

$$\boldsymbol{P(8 \leq X \leq 10) = P(X \leq 10) - P(X < 8) = P(X \leq 10) - P(X \leq 7) = F(10) - F(7)}$$

Adding the following line will do the computation and save it as variable **P_7to10**:

P_7to10= PROBBNML(0.35, 20, 10)- PROBBNML(0.35, 20, 7);

```
DATA binomial;
F_10 = PROBBNML(0.35, 20, 10);
F_9 = PROBBNML(0.35, 20, 9);
P_10 = PROBBNML(0.35, 20, 10)- PROBBNML(0.35, 20, 9);
FC_10 = 1- PROBBNML(0.35, 20, 10);
P_7to10= PROBBNML(0.35, 20, 10)- PROBBNML(0.35, 20, 7);
RUN;
PROC PRINT data=binomial;
TITLE 'Binomial Probabilities';
RUN;
```

For more information on the binomial distribution is SAS follow this link:

http://support.sas.com/documentation/cdl/en/qcug/63964/HTML/default/viewer.htm#qcug_functions_sect013.htm

The Poisson distribution in SAS:

Another useful discrete distribution that we learned about is the Poisson distribution. Again, SAS has a built in function to compute the cumulative probabilities of a Poisson distribution given the parameter λ, the average occurrences in a certain time period. As we learned in class the probability function for the Poisson is:

$$\boldsymbol{P(k\ occurences) = \frac{\lambda^{k} * e^{-\lambda}}{k!}}$$

Remember that SAS is computing F(k) which is the cumulative probability for up to k occurrences in a certain amount of time using the function **POISSON**(*λ,k*).

For example, let's suppose that we anticipate one traffic accident on I10 every morning. What is the probability that up to 3 accidents happen tomorrow morning? Basically we are looking for

$$\boldsymbol{P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = F(3)}$$

Once again we will create a dataset called **poisson** to store our computations as follows:

```
DATA poisson;
F_3 = POISSON(1, 3);
RUN;
PROC PRINT data=poisson;
TITLE 'Poisson Probabilities';
RUN;
```

If we want then to compute the probability that more than 3 accidents will happen tomorrow we need to remember again that this is the complement of the probability we just computed so

$$\boldsymbol{P(X > 3) = 1 - P(X \leq 3) = 1 - F(3)}$$

The line FC_3 = 1-POISSON(1, 3); will compute the wanted probability and save it as the variable FC_3 to denote complement of the cumulative probability.

```
DATA poisson;
F_3 = POISSON(1, 3);
FC_3 = 1-POISSON(1, 3);
RUN;
PROC PRINT data=poisson;
TITLE 'Poisson Probabilities';
RUN;
```

The Normal Distribution in SAS:

SAS has a very nice way of producing the cumulative probability function for most of the well-known continuous distributions including the normal distribution.

The command for it is **CDF('NORMAL', x, μ, σ)** and it computes the probability $P(X \leq x)$ for a normal distribution N(μ,σ).

The following link explains the syntax for other continuous distributions of interest.

http://support.sas.com/documentation/cdl/en/lrdict/64316/HTML/default/viewer.htm#a000208980.htm

We will explore the Normal distribution in SAS by trying to understand IQ scores. IQ tests are designed in such a way so that the results of IQ test over a large population (let's say the US) creates a normal distribution with mean 100 and standard deviation 15, i.e. N(μ,σ)=N(100,15).

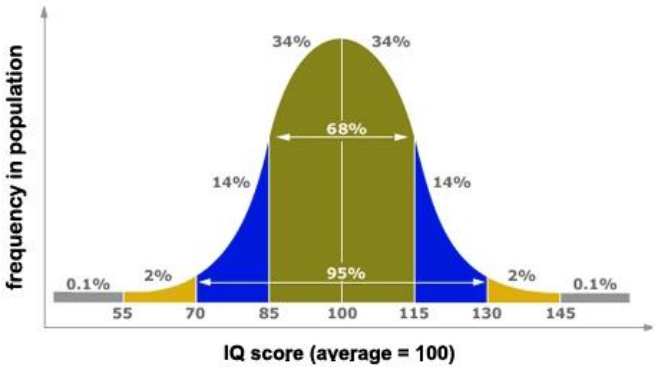

Image retrieved from: http://www.i3mindware.com/highest-iq/what-is-an-iq-score/

If we want for example to find the probability a person has an IQ test of 85 to 115 we must compute:

$$P(85 \leq X \leq 115) = P(X \leq 115) - P(X \leq 85) = F(115) - F(85)$$

```
DATA normal;
P_85to115=CDF('NORMAL',115,100,15)-CDF('NORMAL',85,100,15);
RUN;
PROC PRINT data=normal;
TITLE 'Normal Example 85-115';
Run;
```

We notice that this probability is about 68%. This is actually plus and minus one standard deviation away from the mean.

Similarly if we want to find the probability a person has an IQ test of 70 to 130 we must compute:

$$P(70 \leq X \leq 130) = P(X \leq 130) - P(X \leq 70) = F(130) - F(70)$$

```
DATA normal;
P_70to115=CDF('NORMAL',130,100,15)-CDF('NORMAL',70,100,15);
RUN;
PROC PRINT data=normal;
TITLE 'Normal Example 70-130';
Run;
```

This probability is almost 95% which means that a vast majority of people are between this two scores. Again, this is plus and minus 2 standard deviations away from the mean. The remaining almost 5% are the "exceptional cases" of either too low IQ score or two high, both of which can be thought of as "outliers". Also, mark this 5%, we will need it later when we are doing our hypothesis testing!

Let's now compute the probability a person scores more than 140 on an IQ test. Again remember that

$$P(X > 132) = 1 - P(X \leq 132) = 1 - F(132)$$

```
DATA normal;
```

```
PC_132=1-CDF('NORMAL',132,100,15);
RUN;
PROC PRINT data=normal;
TITLE "Normal Example of more than 132";
Run;
```

The probability is 1.645%. This by the way is the score that Mensa International requires for a person to become a member. (More information on mensa can be found here https://www.mensa.org/

As we learn new continuous distributions we will expand on the command CDF in SAS.