Ph3 Mathematica Homework:
Week 7

Eric D. Black
California Institute of Technology

v2.0

Mathematica contains a number of functions and objects for doing statistics. This week you are going to learn how to use some of them.

1 Distributions

Probability distributions are objects in *Mathematica* that are meant to be passed to functions. There are many, and the ones we have studied go by the following names.

1. `BinomialDistribution[n,p]` - Binomial distribution for n trials, each with a success probability of p.

2. `PoissonDistribution[m]` - Poisson distribution with mean m.

3. `NormalDistribution[a,s]` - Normal distribution with mean a and standard deviation s.

4. `UniformDistribution[min,max]` - Uniform distribution between min and max.

Functions that you use to get numbers from these distributions are, among others,

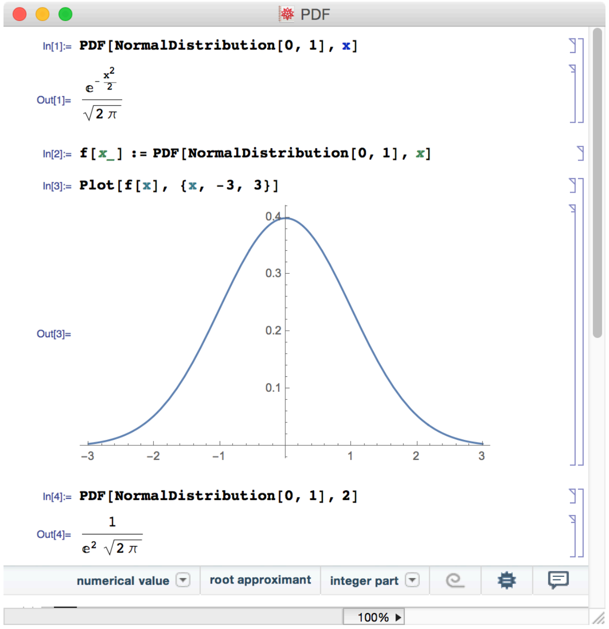
1. `Mean[dist]` - The mean value of the distribution *dist*.

2. `StandardDeviation[dist] -` Just as it says, the standard deviation, what we have been calling σ.

3. `PDF[dist,x]` - Probability Density Function of the distribution *dist* evaluated at x. See Figure 1 for how this works.

4. `CDF[dist,x]` - Cumulative Distribution Function, or the integral of the Probability Distribution Function from $-\infty$ up to x.

5. `Random[dist]` - Pseudorandom number with the specified distribution.



```
In[1]:= PDF[NormalDistribution[0, 1], x]
```

Out[1]= $\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$

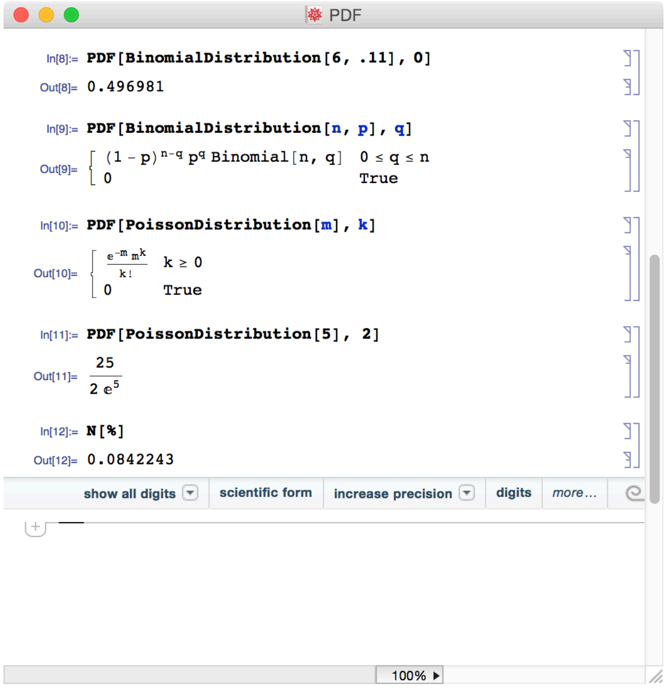
```
In[2]:= f[x_] := PDF[NormalDistribution[0, 1], x]

In[3]:= Plot[f[x], {x, -3, 3}]
```

Out[3]=

```
In[4]:= PDF[NormalDistribution[0, 1], 2]
```

Out[4]= $\frac{1}{e^2\sqrt{2\pi}}$

Figure 1: Getting the functional form of a probability distribution by passing that distribution to the function PDF. If the argument is a variable, in this case x, then the function PDF returns a formula. If the argument is a number, PDF returns the value of the probability density function at that point.



In[8]:= `PDF[BinomialDistribution[6, .11], 0]`

Out[8]= `0.496981`

In[9]:= `PDF[BinomialDistribution[n, p], q]`

Out[9]= $\begin{cases} (1-p)^{n-q}\, p^{q}\, \text{Binomial}[n, q] & 0 \le q \le n \\ 0 & \text{True} \end{cases}$

In[10]:= `PDF[PoissonDistribution[m], k]`

Out[10]= $\begin{cases} \frac{e^{-m}\, m^{k}}{k!} & k \ge 0 \\ 0 & \text{True} \end{cases}$

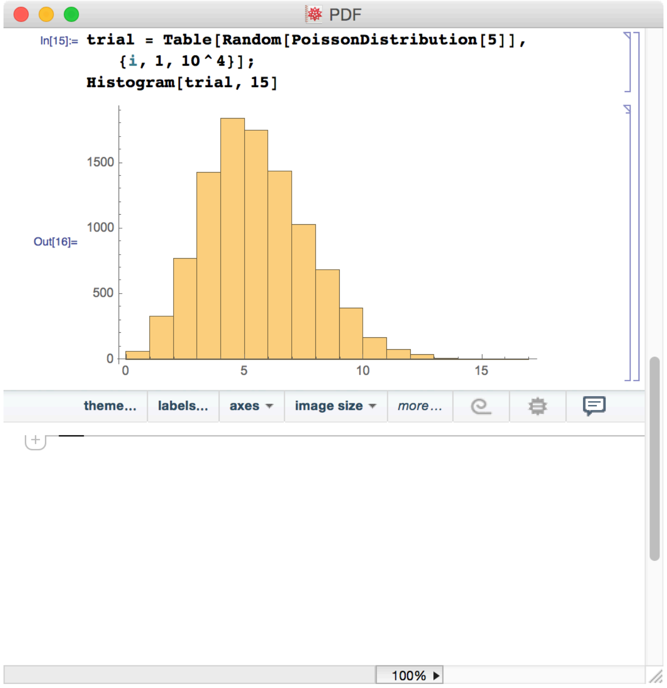
In[11]:= `PDF[PoissonDistribution[5], 2]`

Out[11]= $\frac{25}{2\, e^{5}}$

In[12]:= `N[%]`

Out[12]= `0.0842243`

Figure 2: PDF can be used with discrete distributions as well. Here the first cell evaluates the chance of having no successes in six tries, if each individual trial has an eleven percent chance of succeeding. Note in the second cell how the binomial coefficient is written in *Mathematica*.



```
In[15]:= trial = Table[Random[PoissonDistribution[5]],
             {i, 1, 10^4}];
         Histogram[trial, 15]
```

Figure 3: Passing a distribution to the function Random yields a pseudorandom number that, over many iterations, follows that distribution.

2 Normal Error Integral

In *Taylor*'s Appendix A you have been looking up values for the integral

$$Prob(\text{within } t\sigma) = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-z^2/2} dz,$$

which is the probability that a normally-distributed measurement will fall *within* $t\sigma$ of the mean. *Mathematica*'s built-in Cumulative Distribution Function gives a slightly-different integral,

$$CDF(\text{NormalDistribution[0,1]},\ t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-z^2/2} dz,$$

which is the probability that your measurement will fall *below* $t\sigma$ of the mean. One can be got from the other by noting

$$\int_{-t}^{t} e^{-z^2/2} dz = \int_{-\infty}^{t} e^{-z^2/2} dz - \int_{-\infty}^{-t} e^{-z^2/2} dz$$

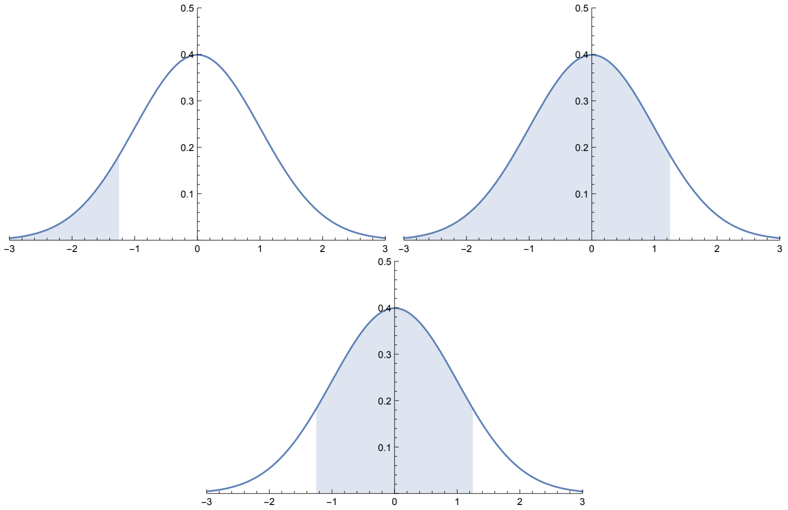


Figure 4: The Normal Error Integral in *Taylor*'s Appendix A (bottom area) is the difference of two Cumulative Distribution Functions (top areas).

Exercise 1: Use *Mathematica*'s Cumulative Distribution Function to generate one row each of the tables in *Taylor*'s Appendix A and B. For Appendix A, generate the row for $t = 1.00$ through $t = 1.09$. For Appendix B, generate the row for $t = 0.50$ through $t = 0.59$. Verify that they agree to the appropriate number of significant figures.

3 Correlation

Last week in your *Taylor* homework you learned how to calculate correlation coefficients, and Appendix C in the back of the book gives the probabilities that these coefficients would arise by random chance. *Mathematica* has built-in functions for both.

There are many methods for evaluating correlation. The one you learned last week is called the *Pearson* method and is accessed in *Mathematica* with the following command.

```
PearsonCorrelationTest[{x1,x2,...}, {y1,y2,...}, "TestDataTable"]
```

The arguments of the command are two one-dimensional lists, the first being all the x-values, and the second all the y-values. This may seem a little bit weird if you are used to entering your data as pairs of values, *e.g.*

```
{{x1,y1}, {x2,y2}, ... }
```

However, you can easily convert between the two orderings using the `Transpose` command.

The option `"TestDataTable"` simply formats the output as an easily-readable table of the correlation coefficient and its associated probability, or P-value. The correlation coefficient, as you may recall, is given by

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}}$$

The P-value is given by the integral

$$Prob_N(|r| \geq |r_0|) = \frac{2\Gamma[(N-1)/2]}{\sqrt{\pi}\Gamma[(N-2)/2]} \int_{|r_0|}^{1} (1 - r^2)^{(N-4)/2} dr,$$

and values for some r_0 and N are are listed in the table in *Taylor*'s Appendix C.

Exercise 2: Do Problem 9.16 in Taylor using *Mathematica*. Do it three ways, using the following options.

```
    "TestDataTable"
    "TestConclusion"
    "TestConclusion", SignificanceLevel->0.01
```

Exercise 3: To get an idea of how many different correlation tests there are, run the following command on your data from Problem 9.16.

```
IndependenceTest[{1,2,3,4,5}, {8,8,5,6,3},
                              {"TestDataTable", All}]
```

Notice how different the results are. Of course, you expect the value in the Statistic column to vary, since each test calculates a different measure of correlation or independence. However, notice how much variation there is in the P-value column. If you are setting your boundary for rejection of the null hypothesis at 5%, for example, it matters a great deal for this data set whether you use the Pearson Correlation test or the Spearman Rank.

It is beyond the scope of this course to cover the differeneces between the various correlation tests. Just be aware that there are several and that the one you have studied is known as the Pearson test.

References

[1] John R. Taylor, *An Introduction to Error Analysis, Second Edition*, University Science Books, (1997).

[2] Mathematica Language & System Documentation Center, http://reference.wolfram.com/language/