CS 5135/6035 Learning Probabilistic Models

Lecture 4: Discrete Probability Distributions[1]

Gowtham Atluri

September 2, 2018

[1] These slides cover material from Chapter 4 of the book, Scheaffer, Richard L., and Linda Young. Introduction to Probability and its Applications, 2009.

Part a. Characterizing Discrete Probability Distributions

- Discrete Probability Distribution
- Cumulative distribution function
- Expectation
- Variance / Standard deviation
- Julia examples

Definitions

- A random variable x is said to be **discrete** if it can take on only a finite number – or a countably infinite number – of possible values.
- The probability distribution of a discrete random variable is called a **probability mass function**.
  - Mass of probability associated with discrete states in the domain
- Cumulative distribution function $cdf(b)$ for a random variable x is

$$cdf(b) = p(x \leq b) = \sum_{x=-\infty}^{b} p(x)$$

Example

- Consider the probability distribution

$$\begin{aligned} p(x=0) &= 0.04 \\ p(x=1) &= 0.32 \\ p(x=2) &= 0.64 \end{aligned}$$

- Compute the cumulative distribution function

Example

- Consider the probability distribution

$$\begin{aligned} p(x=0) &= 0.04 \\ p(x=1) &= 0.32 \\ p(x=2) &= 0.64 \end{aligned}$$

- Compute the cumulative distribution function

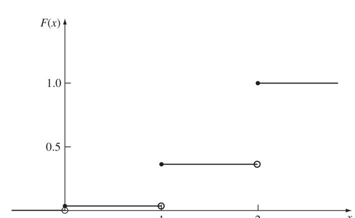
$$cdf(x) = \begin{cases} 0, & x < 0 \\ 0.04, & 0 \leq x < 1 \\ 0.36, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

CDF plot

$$cdf(x) = \begin{cases} 0, & x < 0 \\ 0.04, & 0 \leq x < 1 \\ 0.36, & 1 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

- cdf is discontinuous
  - at points of +ve probability
- The cdf is right-hand continuous
  - not left-hand continuous

$$\lim_{h \to 0+} cdf(1+h) = 0.36 \qquad \lim_{h \to 0-} cdf(1+h) = 0.04 \neq 0.36$$



Expected value (Example: matching coins)

Scenario 1: A and B are matching fair coins. When both A and B flip their coins, if both see the same face, A pays B \$1. Otherwise, B pays A \$1.

The relative frequency distribution of A's wins:

$$p(A = -\$1) = 0.5 \qquad p(A = \$1) = 0.5$$

Question: On an average how much does A win in the long-run?

Expected value (Example: matching coins)

Scenario 1: A and B are matching fair coins. When both A and B flip their coins, if both see the same face, A pays B \$1. Otherwise, B pays A \$1.

The relative frequency distribution of A's wins:

$$p(A = -\$1) = 0.5 \qquad p(A = \$1) = 0.5$$

Question: On an average how much does A win in the long-run?

$$(-1)(\frac{1}{2}) + (1)(\frac{1}{2}) = 0$$

This is the expected wins per game. Is this a fair game?

Expected value

- Discrete probability distribution shows the long-run relative frequency of occurrences for numerical outcomes
- Expectation or Expected value is a one number summary of the distribution

$$E(x) = \sum_x xp(x)$$

- Think of it as the average of a large number of "draws" $x_1, x_2, \dots, x_n$ from the distribution

$$E(x) \approx \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{(Law of large numbers)}$$

- Expectation is also referred to as a mean: $E(x) = \mu$

Expectation of a function of x

- A and B raise the stakes to \$10 per game of matching coins.
- The probability distribution is

$$p(A = -\$10) = 0.5 \qquad p(A = \$10) = 0.5$$

- The expected amount A wins per game is:

$$(-10)(\frac{1}{2}) + (10)(\frac{1}{2}) = 0$$

Expectation of g(x)

If x is a discrete random variable with prob. dist. p(x) and if g(x) is any real-valued function of x, then

$$E(g(x)) = \sum_x g(x)p(x)$$

Properties of Expectations

Property 1

If X is a random variable and a is a constant,

$$E(aX) = aE(X)$$

Property 2

If $X_1, X_2, \dots, X_n$ are random variables, $a_1, a_2, \dots, a_n$ are constant values,

$$E(\sum_i a_i X_i) = \sum_i a_i E(X_i)$$

Property 3

If $X_1, X_2, \dots, X_n$ are **independent** random variables,

$$E(\prod_i X_i) = \sum_i E(X_i)$$

Variance (Example)

Scenario 2: A and B are matching fair coins.

1. if both see the same face and if it is a tail, B pays A \$1,
2. if both see the same face and if it is a head, B pays A \$2,
3. if the faces don't match, A pays B \$1.5.

The relative frequency distribution of A's wins:

$$p(A = -\$1.5) = 0.5 \qquad p(A = \$1) = 0.25 \qquad p(A = \$2) = 0.25$$

The expected amount A wins per game is:

$$(-1.5)(\frac{1}{2}) + (1)(\frac{1}{4}) + (2)(\frac{1}{4}) = 0$$

What is the difference between this and the previous scenario 1?

$$p(A = -\$1) = 0.5 \qquad p(A = \$1) = 0.5$$

Variance (Example)

Scenario 2: A and B are matching fair coins.

1. if both see the same face and if it is a tail, B pays A \$1,
2. if both see the same face and if it is a head, B pays A \$2,
3. if the faces don't match, A pays B \$1.5.

The relative frequency distribution of A's wins:

$$p(A = -\$1.5) = 0.5 \qquad p(A = \$1) = 0.25 \qquad p(A = \$2) = 0.25$$

The expected amount A wins per game is:

$$(-1.5)(\frac{1}{2}) + (1)(\frac{1}{4}) + (2)(\frac{1}{4}) = 0$$

What is the difference between this and the previous scenario 1?

$$p(A = -\$1) = 0.5 \qquad p(A = \$1) = 0.5$$

A has a chance of winning more and also losing more, i.e., **increased variability of winnings**.

Variance and Standard deviation

Variance (σ^2)

The variance of a random variable x with expected value μ is given as

$$\sigma^2 = E[(x-\mu)^2]$$

- Variance can be thought of as the average squared distance between the values of x and the expected value μ.
- $\sigma^2 = 0$, when all the probability is concentrated at a single point.
- σ^2 becomes large as the points with non-zero probability spread out more.

Standard deviation

Standard of a random variable x is the square root of variance

$$\sigma = \sqrt{\sigma^2} = \sqrt{E[(x-\mu)^2]}$$

Example: Variance for matching coins

For scenario 1:

$$\sigma^2 = E[(x-\mu)^2]$$

$$\sigma^2 = (-1)^2(\frac{1}{2}) + (1)^2(\frac{1}{2}) = 1 \qquad \sigma = 1$$

For scenario 2:

$$(-1.5)^2(\frac{1}{2}) + (1)^2(\frac{1}{4}) + (2)^2(\frac{1}{4}) = 2.375 \qquad \sigma = 1.54$$

Observations

- In scenario 1, each outcome (-1 or +1) deviates by 1 unit from the expected value.
- In scenario 2, +ve values (as do -ve values) deviates by 1.5 units from the expected value on average.

Julia Example

```
## 0×4 DataFrames.DataFrame
```

```
## 10×4 DataFrames.DataFrame
##  Row   AgeInterval  Age   Y2000   Y2100
##
##  1     Under 5      3     0.069   0.063
##  2     5-9          8     0.073   0.062
##  3     10-19        15    0.144   0.128
##  4     20-29        25    0.133   0.123
##  5     30-39        35    0.155   0.12
##  6     40-49        45    0.153   0.116
##  7     50-59        55    0.108   0.108
##  8     60-69        65    0.073   0.098
##  9     70-79        75    0.059   0.083
##  10    80 and over  90    0.033   0.099
```


Julia Example

- Mean and std yield a useful summary of the distribution.

```
# computing expectation for Y2000 census and Y2100 projections
[sum(df[:Age].*df[:Y2000]), sum(df[:Age].*df[:Y2100])]'
```

```
## 1×2 RowVector{Float64,Array{Float64,1}}:
##  36.666  42.545
```

```
#computing std for Y2000 census and Y2100 projections
[sqrt(sum((df[:Age]-36.6).^2.*df[:Y2000])),
          sqrt(sum((df[:Age]-42.55).^2.*df[:Y2100]))]'
```

```
## 1×2 RowVector{Float64,Array{Float64,1}}:
##  22.5926  26.2975
```


Part b. Standard Discrete Probability Distributions

- **Bernoulli Distribution**
- **Binomial Distribution**
- **Categorical Distribution**
- **Multinomial Distribution**
- Geometric Distribution
- Negative Binomial Distribution
- Poisson Distribution

Questions:

- What scenarios are these distributions suitable for?
- What is the probability mass function?
- How to compute the probability analytically and using Julia?
- What is the mean and variance?

Bernoulli Distribution

- Numerouns experiments have two possible outcomes:
  - A coin toss may result in a head or a tail.
  - An item from an assembly line is defective or not defective
  - A piece of fruit is either damaged or not damaged.
  - A child is either male or female
- Such experiments are called **Bernoulli trials**, after Swiss mathematician Jacob Bernoulli
- One outcome of a Bernoulli trial is identified as a *success* and the other is identified as a *failure*.
- Probability of observing a success is p, and probability of observing failure is $1 - p$.

Probability distribution of x is

$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Julia example of Bernoulli Distribution

```
using Distributions;
d = Bernoulli(0.1);
data = rand(d,10)'
```

```
## 1×10 RowVector{Int64,Array{Int64,1}}:
##  0  0  0  0  0  0  0  1  0  0
```

```
d = Bernoulli(0.5);
data = rand(d,10)'
```

```
## 1×10 RowVector{Int64,Array{Int64,1}}:
##  0  1  0  0  0  0  1  0  0  1
```

```
d = Bernoulli(0.9);
data = rand(d,10)'
```

```
## 1×10 RowVector{Int64,Array{Int64,1}}:
##  1  1  1  1  1  1  1  1  1  1
```


Expectation of Bernoulli Distribution

Distribution is:

$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Expectation:

$$E(x) = \textstyle\sum_x xp(x)$$

Expectation of Bernoulli Distribution

Distribution is:

$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Expectation:

$$\begin{aligned} E(x) &= \textstyle\sum_x xp(x) \\ &= 0p(0) + 1p(1) \\ &= 0(1-p) + 1(p) \\ &= p \end{aligned}$$

Variance of Bernoulli Distribution

Distribution is:

$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Expectation: p

Variance:

$$\sigma^2 = E(X^2) - [E(X)]^2$$

Variance of Bernoulli Distribution

Distribution is:

$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Expectation: p

Variance:

$$\begin{aligned} \sigma^2 &= E(X^2) - [E(X)]^2 \\ &= \textstyle\sum_x x^2 p(x) - p^2 \\ &= (0)^2(1-p) + 1^2(p) - p^2 \\ &= p - p^2 \\ &= p(1-p) \end{aligned}$$

Binomial Distribution

- **Setup:** n **independent Bernoulli** trials, **each** with a probability p of success
- **random variable:** x is a random variable of success in the n trials

Probability of x success in n trials is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n$$

This probability can be derived from Bernoulli distribution.

Binomial distribution example in Julia

```
d = Binomial(10,0.1);
data = rand(d,10)'
```

```
## 1×10 RowVector{Int64,Array{Int64,1}}:
##  2  1  0  0  2  0  0  1  3  0
```

```
d = Binomial(10,0.5);
data = rand(d,10)'
```

```
## 1×10 RowVector{Int64,Array{Int64,1}}:
##  7  2  7  3  5  4  5  3  4  6
```

```
d = Binomial(10,0.9);
data = rand(d,10)'
```

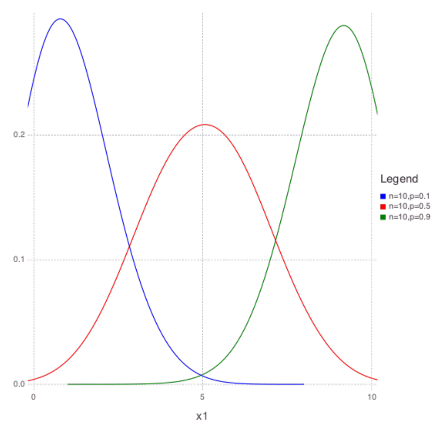
```
## 1×10 RowVector{Int64,Array{Int64,1}}:
##  10  8  8  9  10  10  9  9  10  10
```


Binomial distribution example in Julia

```
#draw samples from three binomial dist.
d1 = Binomial(10,0.1);
data1 = rand(d1,1000);
d2 = Binomial(10,0.5);
data2 = rand(d2,1000);
d3 = Binomial(10,0.9);
data3 = rand(d3,1000);
df = convert(DataFrame, hcat(data1, data2, data3));

#plot the density of the resultant samples
using Gadfly;
myplot = plot(layer(df,x="x1",Geom.density(bandwidth=1),
                Theme(default_color=colorant"blue")),
    layer(df,x="x2",Geom.density(bandwidth=1),
                Theme(default_color=colorant"red")),
    layer(df,x="x3",Geom.density(bandwidth=1),
                Theme(default_color=colorant"green")),
    Guide.manual_color_key("Legend",
                ["n=10,p=0.1", "n=10,p=0.5", "n=10,p=0.9"],
                ["blue","red","green"]),
    Coord.Cartesian(xmin=0, xmax=10, ymin=0));
```


Binomial distribution example in Julia



Binomial distribution example

Problem:

- 10% of a large lot of apples are damaged.
- If four apples are randomly sampled from the lot, find the probability that at least one apple in the sample of four is defective.

Solution:

We need to find $p(x \geq 1)$

$$p(x \geq 1) = 1 - p(x = 0)$$

Binomial distribution example

Problem:

- 10% of a large lot of apples are damaged.
- If four apples are randomly sampled from the lot, find the probability that at least one apple in the sample of four is defective.

Solution:

We need to find $p(x \geq 1)$

$$\begin{aligned} p(x \geq 1) &= 1 - p(x = 0) \\ &= 1 - p(0) \\ &= 1 - \tbinom{4}{0}(0.1)^0(0.9)^4 \\ &= 1 - (0.9)^4 \\ &= 0.3439 \end{aligned}$$

Binomial distribution example

Problem:

- 10% of a large lot of apples are damaged.
- If four apples are randomly sampled from the lot, find the probability that at least one apple in the sample of four is defective.

Solution (Julia):

```
d = Binomial(4,0.1);
1-pdf(d,0)
```

```
## 0.3439
```


Binomial distribution: mean and variance

$$E(x) = \sum_x xp(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

This is not straightforward to solve.

Using Bernoulli random variables $y_1, y_2, \ldots, y_n$,

$$E(x) = E(\sum_{i=1}^{n} y_i) = \sum_{i=1}^{n} E(y_i) = \sum_{i=1}^{n} p = np$$

Similarly,

$$\sigma^2(x) = \sum_{i=1}^{n} \sigma^2(y_i) = \sum_{i=1}^{n} p(1-p) = np(1-p)$$

Categorical distribution

- There are scenarios where there are more than two possible outcomes
  - roll of a dice (6 possible outcomes)
- This is a generalized version of Bernoulli distribution
- While Bernoulli distribution deals with two outcomes, categorical distribution can allow multiple outcomes (say k)
- **Setup:**
  - k mutually exclusive outcomes for a trial
  - p_i is the probability associated with outcome i
  - $\sum_i p_i = 1$
- **random variable:** x is a random variable of seeing one of the k outcomes

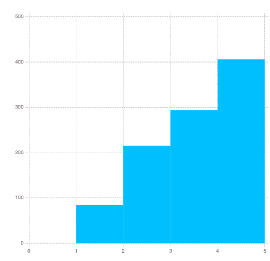
Probability of x is

$$p(x=i) = p_i, x = i \ldots k \ (or) \ p(x) = p_1^{[x=1]} p_2^{[x=2]} \ldots p_k^{[x=k]}$$

where $[x = i]$ evalutes to 1 if outcome is i, 0 otherwise.

Categorical distribution: Julia examples

```
d = Categorical([0.1, 0.2, 0.3, 0.4]);
data = rand(d,1000);
myplot = plot(x=data,
        Geom.histogram(bincount=length(unique(data))));
draw(PNG("./figs/categ1.png", 6inch, 6inch), myplot);
```



Multinomial distribution

- **Setup:**
  - n **independent 'Categorical'** trials,
  - probability of each of the k outcomes $p_1, p_2, \ldots, p_k$ ($\sum_i p_i = 1$)
    - these probabilities are same across trials
- **random variable:** x is a random variable of counts of each of the k possible outcomes over n trials
- Example: When a fair 6-sided dice is rolled 10 times, what is the probability of seeing four 2s, three 4s, and three 6s?

The probability of seeing $[x_1 x_2 ...x_k] outcomes in n trials is

$$p(x_1, x_2, \ldots, x_k) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

Multinomial distribution: Mean and Variance

Mean

For $i \in \{1, 2, \ldots, k\}$, the mean of x_i is:

$$E(x_i) = np_i$$

Variance

For $i \in \{1, 2, \ldots, k\}$, the variance of x_i is:

$$\sigma^2(x_i) = np_i(1 - p_i)$$

Covariance

In general,

$$cov(x, y) = E[(x - E(x))(y - E(y))]$$

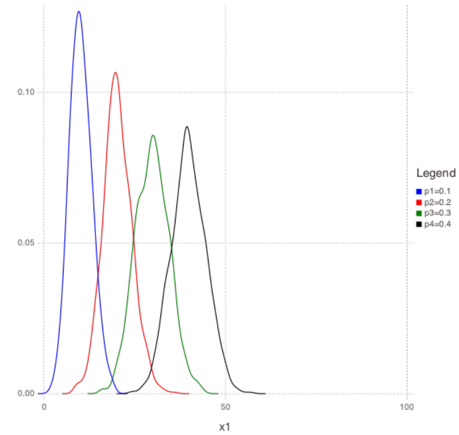
For distinct $i, j \in \{1, 2, \ldots, k\}$, the co-variance of x_i and x_j is:

Multinomial distribution: Julia examples

```
d = Multinomial(100,[0.1, 0.2, 0.3, 0.4]);
data = rand(d,1000);
df = convert(DataFrame,data');

myplot = plot(layer(df,x="x1",Geom.density(bandwidth=1),
                Theme(default_color=colorant"blue")),
    layer(df,x="x2",Geom.density(bandwidth=1),
                Theme(default_color=colorant"red")),
    layer(df,x="x3",Geom.density(bandwidth=1),
                Theme(default_color=colorant"green")),
    layer(df,x="x4",Geom.density,
                Theme(default_color=colorant"black")),
    Guide.manual_color_key("Legend",
                ["p1=0.1", "p2=0.2", "p3=0.3", "p4=0.4"],
                ["blue","red","green","black"]),
    Coord.Cartesian(xmin=0, xmax=100, ymin=0));
draw(PNG("./figs/mulinomial_plot.png", 6inch, 6inch), myplot)
```


Multinomial distribution: Julia examples



Part c. Standard Discrete Probability Distributions

- Bernoulli Distribution
- Binomial Distribution
- Categorical Distribution
- Multinomial Distribution
- **Geometric Distribution**
- **Negative Binomial Distribution**
- **Poisson Distribution**

Questions:

- What scenarios are these distributions suitable for?
- What is the probability mass function?
- How to compute the probability analytically and using Julia?
- What is the mean and variance?

Geometric Distribution

- Suited for modeling the number of failures before a first success
  - when each experiment is a Bernoulli trial with a constant probability of success
- Example: Number of failed rocket engine firings before a successful firing.

$$\begin{aligned} p(x) &= p(y_1 = 0, y_2 = 0, \ldots, y_x = 0, y_{x+1} = 1) \\ &= p(y_1 = 0)p(y_2 = 0)\ldots p(y_x = 0)p(y_{x+1} = 1)) \\ &= (1-p)(1-p)\ldots(1-p)p \\ &= (1-p)^x p \\ &= q^x p \quad x = 0, 1, 2, \ldots \end{aligned}$$

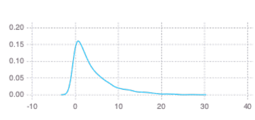
- This is the geometric probability distribution $p(x) = q^x p$.
  - This r.v. can take on a countably infinite number of possible values

Geometric Distribution: Example

- A recruitment firm finds that 20% of the applicants have the necessary skills. Applicants are randomly selected and interviewed. What is the probability that five applicants are interviewed before finding the first suitable applicant?

$p(x = 5) = (0.8)^5(0.2) = 0.066$

```
d = Geometric(0.2);
data = rand(d,1000);
myplot = plot(x=data,Geom.density);
draw(PNG("./figs/geometric.png", 4inch, 2inch), myplot);
```



Mean and variance of a Geometric distribution

Mean

$$E(x) = \sum_x x p(x) = \sum_{x=0}^{\infty} x p q^x$$

Using geometric progression, this will be $E(x) = \frac{q}{p}$.

Variance

$$\sigma^2 = \frac{q}{p^2}$$

Memoryless property: only discrete distribution

$$p(x \geq j + k | X \geq j) = p(x \geq k)$$

If we observed j straight failures, then the probability of observing at least k more failures ($j + k$ total failures) before a success is the same as if we were just beginning and wanted to determine the probability of observing at least k failures prior to a success.

Negative Binomial Distribution

- Geometric distribution model the probabilistic behavior of # failures prior to first success in a sequence of Bernoulli trials
- A negative binomial models the # failures prior to the second success, or third success or r^{th} success.

$$\begin{aligned} p(x) &= p(\text{1st } (x+r-1) \text{ trials contain } (r-1) \text{ successes } \ldots \\ &\qquad \text{and the } (x+r)\text{th trial is a success}) \\ &= p(\text{1st } (x+r-1) \text{ trials contain } (r-1) \text{ successes}) \times \\ &\qquad p((x+r)\text{th trial is a success}) \\ &= \binom{x+r-1}{r-1} p^{r-1}(1-p)^x \;\; p \\ &= \binom{x+r-1}{r-1} p^r q^x, \qquad x = 0, 1, \ldots \end{aligned}$$

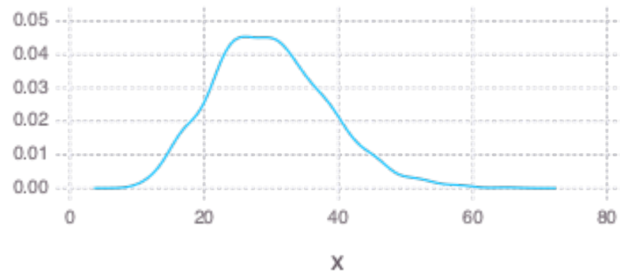
Mean and Variance

$$E(x) = \frac{rq}{p} \qquad \sigma^2 = \frac{rq}{p^2}$$

Negative Binomial Distribution: Julia example

Number of failures before 5th success. Probability of success is 0.4.

```
d = NegativeBinomial(20,0.4);
data = rand(d,1000);
myplot = plot(x=data,Geom.density);
draw(PNG("./figs/neg_binomial.png", 4inch, 2inch), myplot);
```



Poisson Distribution

- Typically used to model counts of rare events in a given area or volume or time.
- Develoment of a probabilistic model for the number of accidents that occur at a given intersection in a period of 1 week.
  - We can think of dividing the time interval into n subintervals such that p(One accident in a subinterval) $= p$ p(One accident in a subinterval) $= 1 - p$
- Assumptions:
  - p($>$1 accidents in a subinterval) $= 0$
  - occurrence of accident in 1 interval is independent of other intervals
  - total # accidents = total # subintervals that contain 1 accident follows Binomial

Poisson Distribution

- No unique way to choose the n and p for the binomial model - as $n \to \infty$, $p \to 0$ - we also restrict the mean to remain a constant λ, $\lambda = np$
- The probability of x number of accidents can be written, using Binomial distribution, as

$$p(x) = \lim_{n\to\infty} \binom{n}{x} \frac{\lambda}{n}^x (1 - \frac{\lambda}{n})^{n-x} = \frac{\lambda^x}{x!} e^{-\lambda}, \qquad x = 0, 1, 2, \ldots for \lambda > 0$$

- λ denotes the mean occurrences of events in one time period

Poisson Distribution: Example

- Example: During business hours, the number of calls passing through a cellular system averages 5 per minute. Find the probability that no call will pass through the system during a given minute.

$$p(x = 0) = \frac{5^0}{0!} e^{-5} = e^{-5} = 0.007$$

- Find the probability that no call will pass through during a 2 minute period.
  - Here mean # calls in 2 minutes $= \lambda = 5 \times 2 = 10$

$$p(x = 0) = \frac{10^0}{0!} e^{-10} = e^{-10} = 0.00005$$

- Find the probability that three calls will pass through during a 2 minute period.

$$p(x = 3) = \frac{10^3}{3!} e^{-10} = 0.0076$$

Poisson Distribution: Julia example

```
#draw samples from three Poisson distributions
d1 = Poisson(0.66);
data1 = rand(d1,1000);
d2 = Poisson(2.11);
data2 = rand(d2,1000);
d3 = Poisson(5.01);
data3 = rand(d3,1000);
df = convert(DataFrame, hcat(data1, data2, data3));

#plot the density of the resultant samples
using Gadfly;
myplot = plot(layer(df,x="x1",Geom.density(bandwidth=1),
                Theme(default_color=colorant"blue")),
    layer(df,x="x2",Geom.density(bandwidth=1),
                Theme(default_color=colorant"red")),
    layer(df,x="x3",Geom.density(bandwidth=1),
                Theme(default_color=colorant"green")),
    Guide.manual_color_key("Legend",
                ["n=10,p=0.1", "n=10,p=0.5", "n=10,p=0.9"],
                ["blue","red","green"]),
    Coord.Cartesian(xmin=0, xmax=10, ymin=0));
```

Poisson Distribution: Julia example