Lecture 12
Probability models

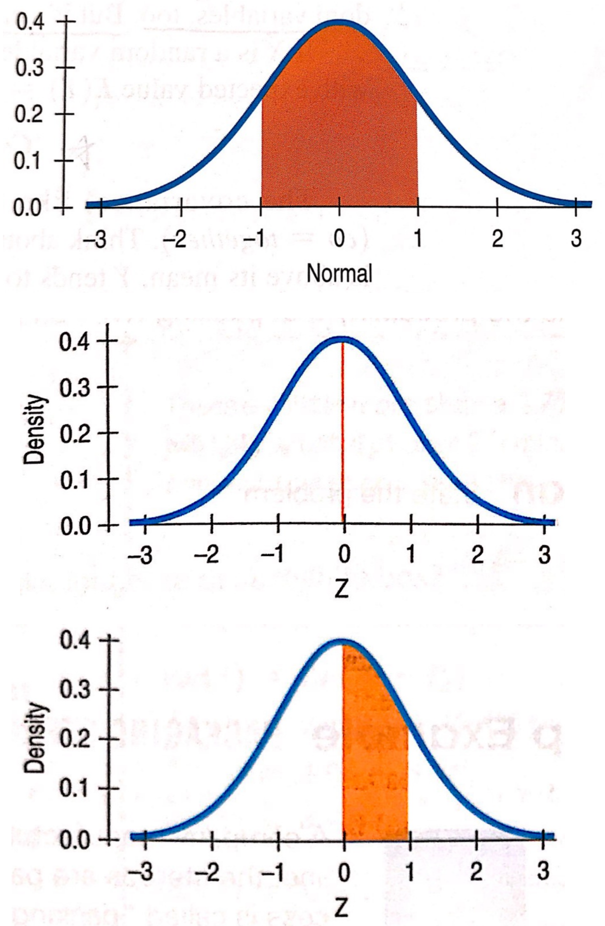
CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Review: Key Points

Chapter 16: Random variables

- Discrete vs. continuous random variables
- Expected values (mean): $\mu = E(X) = \sum xP(x)$
- Here, *probability* conveys the information about population assuming a large number of repeats
- Spread: $\sigma^2 = Var(X) = \sum (x - \mu)^2 P(x)$
  $\sigma = SD(X) = \sqrt{Var(X)}$
- $E(X \pm c) = E(X) \pm c,\ Var(X \pm c) = Var(X)$
- $E(aX) = aE(X),\ Var(aX) = a^2 Var(X)$
- $E(X \pm Y) = E(X) \pm E(Y)$
- $Cov(X, Y) = E\big((X - \mu)(Y - \nu)\big)$
- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$
- $Corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Probability models for continuous random variables



- For a continuous random variable, the probability is defined as the area under the curve over an interval.
- There is no area for a point, thus $P(X = x) = 0$
  - $P(a < X < b)$
  - $P(X < x)$: cumulative probability

Multiple probability models for random variables

- Normal model is just one model among many.
- In this chapter, we will go over the following models:
  - Geometric model
  - Binomial model
  - Poisson model
  - Uniform model
  - Exponential models

Bernoulli Trials

- In a snack box:

Success

10%

Failure

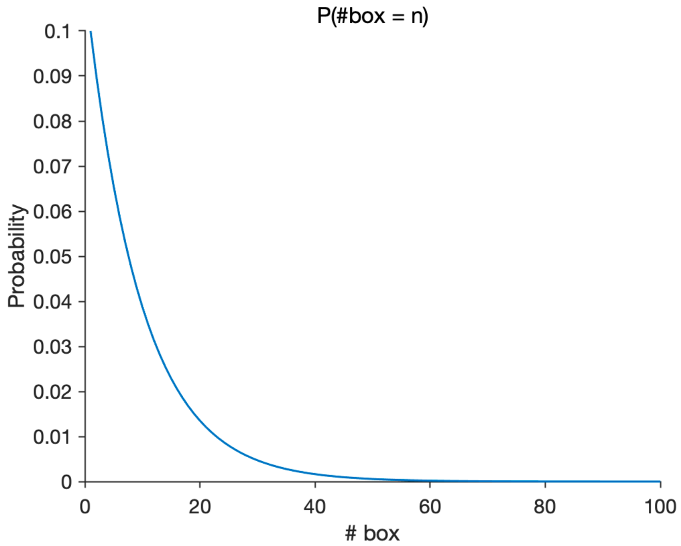
40%

50%

- Success rate = 10%, the probability of success, p = 0.10
- Two possible outcomes (success vs. failure)
- Trials (purchasing snack boxes) are independent
- Other **Bernoulli** trials: tossing a coin, shooting free throws, etc.

Geometric (등비, 等比) model

- *"How many snack boxes we will need to open to find the*  *card?"*
- $P(\#box=1) = 0.1$
- $P(\#box=2) = (0.9) \times (0.1)$
- … $P(\#box = 5) = (0.9)^4 \times (0.1)$



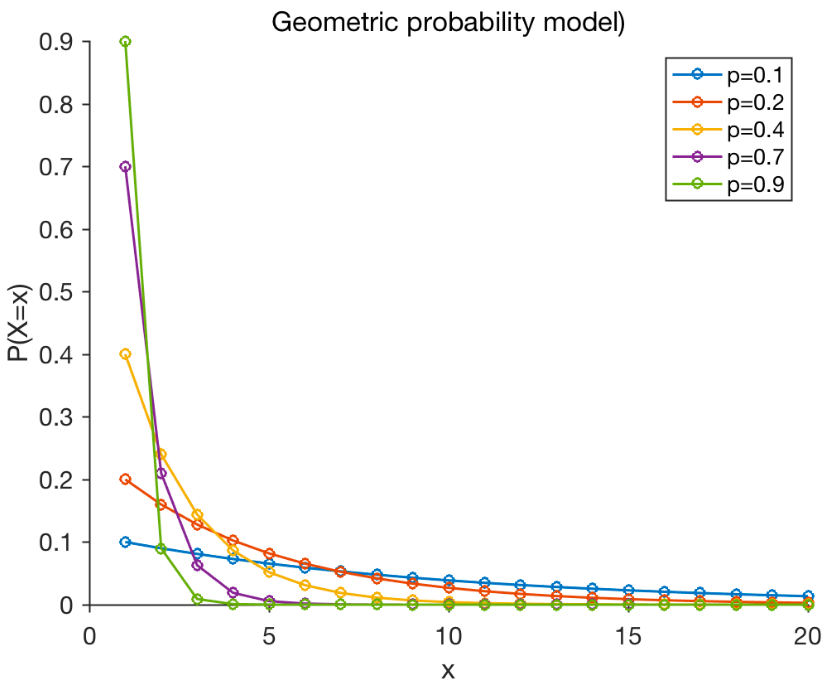

```
p = 0.1;
q = 1-p;
n = 1:100;
p_n_box = q.^(n-1)*p;
plot(p_n_box, 'linewidth', 1.5);
title('P(#box = n)')
xlabel('# box')
ylabel('Probability');
set(gca, 'linewidth', 1, 'fontsize', 15, 'tickdir', 'out');
set(gcf, 'color', 'w');
box off;

str = 'The expected number of boxes to open to find the charizard card is';
fprintf('\n%s %d\n', str, ceil(sum(n.*p_n_box)));
```

```
The expected number of boxes to open to find the charizard card is 10.
```

- $E(\#box) = 10 = 1/0.1$

Geometric model

- Geometric probability model, Geom(p)

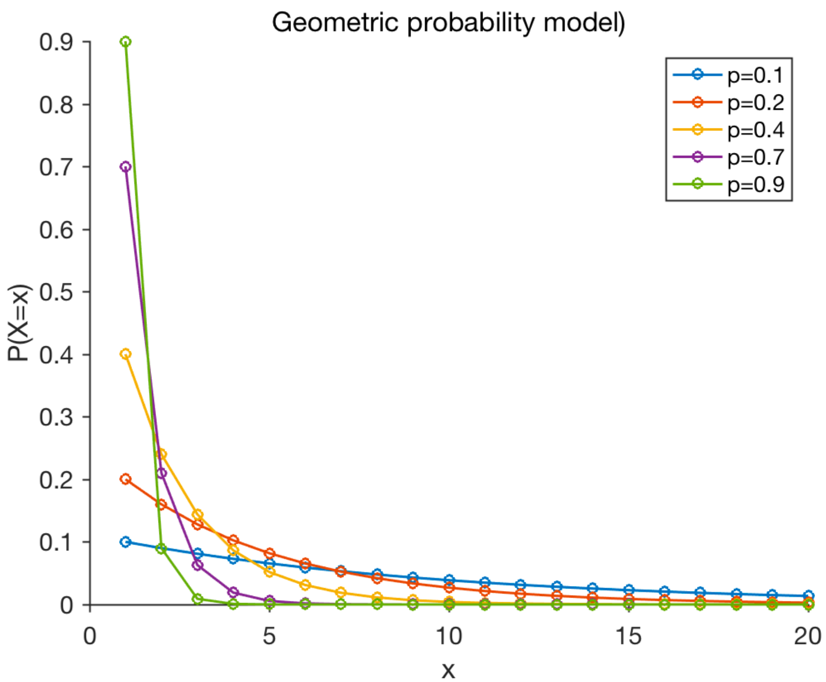


Plot geometric probability

```
close all;
p_all = [.1 .2 .4 .7 .9];
legend_cell = cell(numel(p_all),1);

figure;
for i = 1:numel(p_all)
    p = p_all(i);
    q = 1-p;
    n = 1:20;
    p_n_box = q.^(n-1)*p;
    hold on;
    plot(p_n_box, 'o-', 'linewidth', 1.5);
    legend_cell{i} = sprintf('p=%0.1f', p);
end

title('Geometric probability model')
xlabel('x')
ylabel('P(X=x)');

legend(legend_cell);

set(gca, 'linewidth', 1, 'fontsize', 15, 'tickdir', 'out');
set(gcf, 'color', 'w');
box off;
```

Geometric model

- Geometric probability model, Geom(p)



- p = probability of success (and q = 1 - p, p of failure)
- $P(X = x) = q^{x-1}p$
- Expected value: $E(X) = \mu = \frac{1}{p}$
- Standard deviation: $\sigma = \sqrt{\frac{q}{p^2}}$

Geometric model

- geopdf.m and geocdf.m in MATLAB
- geom.pmf, cdf, etc. in Scipy (Python)

```
>> help geopdf
geopdf - Geometric probability density function

    This MATLAB function returns the probability density function (pdf) of the
    geometric distribution at each value in x using the corresponding probabilities
    in p.

    y = geopdf(x,p)

    참고 항목 geocdf, geoinv, geornd, geostat, mle, pdf
```

Geometric model

- Geometric probability model

참고: **무한등비급수** [편집]

무한등비급수는 등비수열의 각 항을 무한히 더한 것이며, 그 합은 다음과 같다.

$$\sum_{k=0}^{\infty} ar^k = \lim_{n\to\infty}\sum_{k=0}^{n-1} ar^k = \lim_{n\to\infty}\frac{a(1-r^n)}{1-r} = \frac{a}{1-r} (\text{ 단, } |r|<1)$$

MATH BOX

We want to find the mean (expected value) of random variable X, using a geometric model with probability of success p.

First, write the probabilities:

x	1	2	3	4	$\cdots$
$P(X=x)$	p	qp	q^2p	q^3p	$\cdots$

The expected value is: $E(X) = 1p + 2qp + 3q^2p + 4q^3p + \cdots$

Let $p = 1 - q$: $= (1-q) + 2q(1-q) + 3q^2(1-q) + 4q^3(1-q) + \cdots$

Simplify: $= 1 - q + 2q - 2q^2 + 3q^2 - 3q^3 + 4q^3 - 4q^4 + \cdots$

That's an infinite geometric series, with first term 1 and common ratio q:

$$= 1 + q + q^2 + q^3 + \cdots$$

$$= \frac{1}{1-q}$$

So, finally . . . $E(X) = \dfrac{1}{p}.$

Binomial model

- *"Suppose you bought 5 boxes of snack, and what is the probability that you get exactly two*  *cards?"*
- Still Bernoulli trials, but different question: "the number of *successes* in the 5 trials"

- Two parameters are needed to define the binomial model, Binom(n, p)
  - n: the number of trials
  - p: the probability of *success*

- 2 successes in 5 trials means, 2 successes and 3 failures, p = $(0.1)^2 \times (0.9)^3$
- Many combinations of 2 successes and 3 failures:
  - $\binom{n}{k}$ or ${}_nC_k$: "n choose k"
  - ${}_nC_k = \frac{n!}{k!(n-k)!}$

Binomial model

- *"Suppose you bought 5 boxes of snack, and what is the probability that you get exactly two*  *cards?"*

BINOMIAL PROBABILITY MODEL FOR BERNOULLI TRIALS: BINOM(*n*, *p*)

n = number of trials
p = probability of success (and $q = 1 - p$ = probability of failure)
X = number of successes in n trials

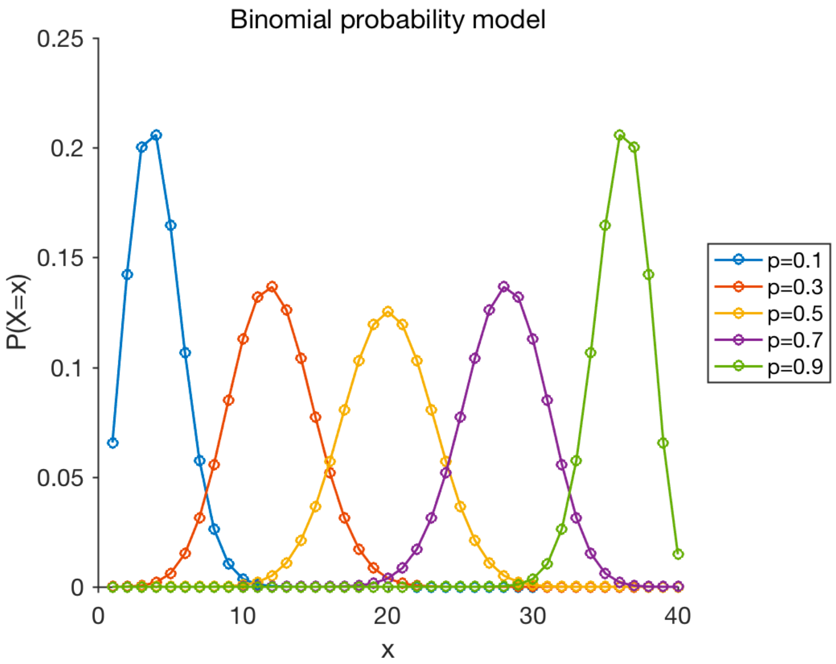
$$P(X = x) = {}_nC_x\, p^x q^{n-x}, \text{ where } {}_nC_x = \frac{n!}{x!(n-x)!}$$

Mean: $\mu = np$
Standard Deviation: $\sigma = \sqrt{npq}$

- $P(\#\text{success} = 2) = 10 \times (0.1)^2 \times (0.9)^3 = 0.0729$

Binomial model

- Plot for Binom(n, p), where n = 40

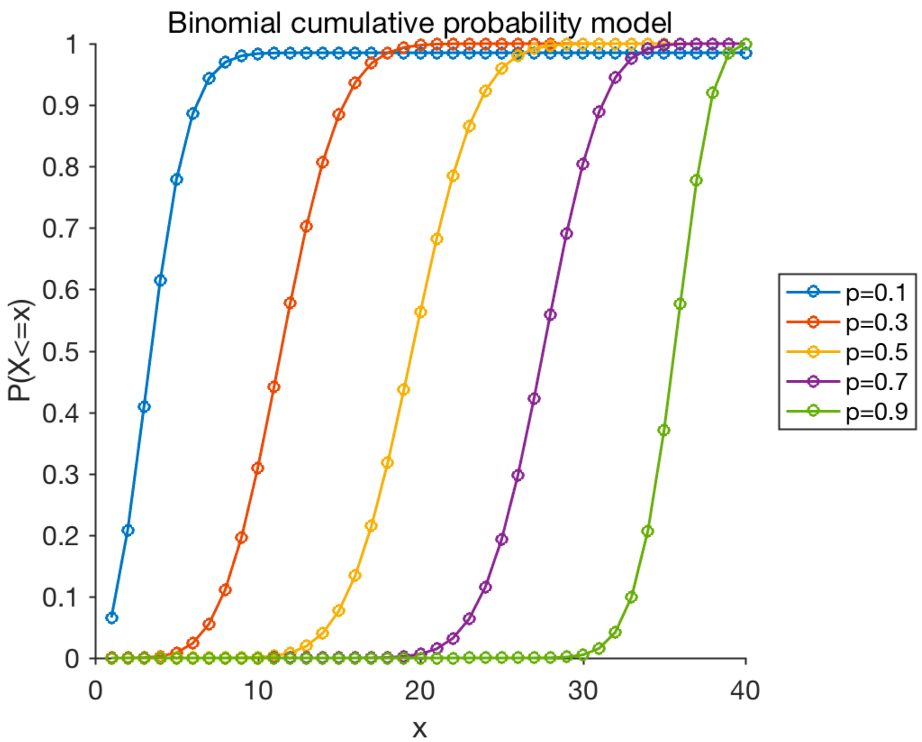


Binom(n,p) with n = 40

```
% multiple p
p_all = 0.1:0.2:1;
n = 40;
x = 1:n;
legend_cell = cell(numel(p_all),1);
cprob = cell(numel(p_all),1);

close all;

figure;
for i = 1:numel(p_all)
    p = p_all(i);
    q = 1-p;
    % Binomial probability X = x
    prob = factorial(n)./(factorial(x).*factorial(n-x)) .* (p.^x) .* (q.^(n-x));

    % Calculating cumulative probability, P(X <= x)
    cprob{i}(1) = 0;
    for j = 1:numel(prob)
        cprob{i}(j+1) = cprob{i}(j)+prob(j);
    end
    hold on;
    plot(prob, 'o-', 'linewidth', 1.5);
    legend_cell{i} = sprintf('p=%0.1f', p);
end

title('Binomial probability model')
xlabel('x')
ylabel('P(X=x)');

legend(legend_cell, 'location', 'eastoutside');

set(gca, 'linewidth', 1, 'fontsize', 15, 'tickdir', 'out');
set(gcf, 'color', 'w');
box off;
```

Binomial model

- Cumulative probability density function



Cumulative probability function

```
figure;
for i = 1:numel(p_all)
    p = p_all(i);
    hold on;
    % plot cumulative probability
    plot(cprob{i}(2:end), 'o-', 'linewidth', 1.5);
    legend_cell{i} = sprintf('p=%0.1f', p);
end

title('Binomial cumulative probability model')
xlabel('x')
ylabel('P(X<=x)');

legend(legend_cell, 'location', 'eastoutside');

set(gca, 'linewidth', 1, 'fontsize', 15, 'tickdir', 'out');
set(gcf, 'color', 'w');
box off;
```

Binomial model

- binopdf.m and binocdf.m in MATLAB
- binom in scipy

```
>> help binopdf
binopdf - Binomial probability density function

    This MATLAB function computes the binomial pdf at each of the values in X using
    the corresponding number of trials in N and probability of success for each
    trial in P.

    Y = binopdf(X,N,P)

    참고 항목 binocdf, binofit, binoinv, binornd, binostat, pdf
```

Binomial model

MATH BOX

To derive the formulas for the mean and standard deviation of a Binomial model we start with the most basic situation.

Consider a single Bernoulli trial with probability of success p. Let's find the mean and variance of the number of successes.

Here's the probability model for the number of successes:

x	0	1
$P(X = x)$	q	p

Find the expected value:

$$E(X) = 0q + 1p$$
$$E(X) = p$$

And now the variance:

$$\begin{aligned} Var(X) &= (0 - p)^2 q + (1 - p)^2 p \\ &= p^2 q + q^2 p \\ &= pq(p + q) \\ &= pq(1) \\ Var(X) &= pq \end{aligned}$$

Binomial model

MATH BOX

What happens when there is more than one trial, though? A Binomial model simply counts the number of successes in a series of n independent Bernoulli trials. That makes it easy to find the mean and standard deviation of a binomial random variable, Y.

$$\begin{aligned}
\text{Let } Y &= X_1 + X_2 + X_3 + \cdots + X_n \\
E(Y) &= E(X_1 + X_2 + X_3 + \cdots + X_n) \\
&= E(X_1) + E(X_2) + E(X_3) + \cdots + E(X_n) \\
&= p + p + p + \cdots + p \text{ (There are } n \text{ terms.)}
\end{aligned}$$

So, as we thought, the mean is $E(Y) = np$.

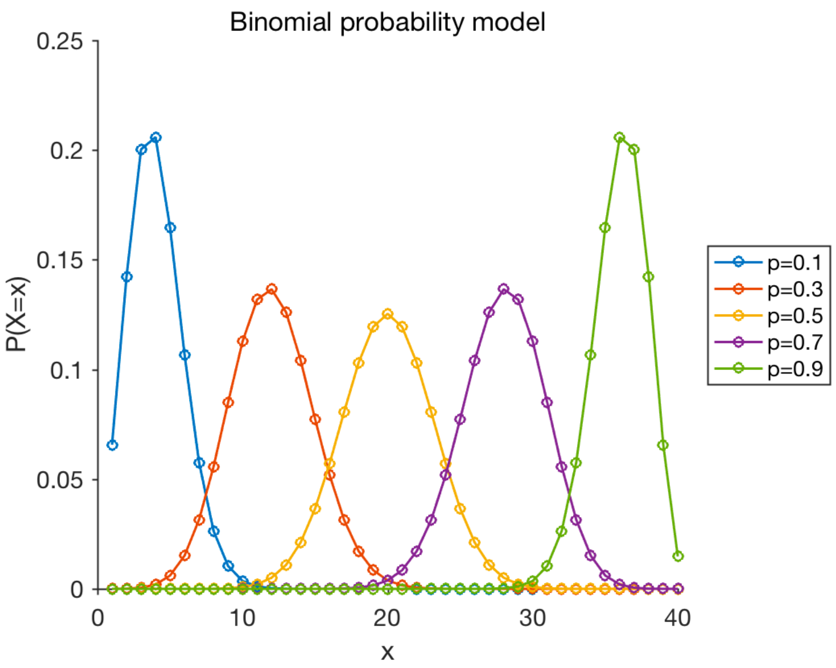
And since the trials are independent, the variances add:

$$\begin{aligned}
Var(Y) &= Var(X_1 + X_2 + X_3 + \cdots + X_n) \\
&= Var(X_1) + Var(X_2) + Var(X_3) + \cdots + Var(X_n) \\
&= pq + pq + pq + \cdots + pq \text{ (Again, } n \text{ terms.)} \\
Var(Y) &= npq
\end{aligned}$$

Voilà! The standard deviation is $SD(Y) = \sqrt{npq}$.

Approximating the binomial with a normal model

- Plot for Binom(n, p), where n = 40



- Some of them look exactly like the normal distribution.
- Usually, when $np \geq 10$ and $nq \geq 10$, the binomial model is approximately Normal, which can simplify the calculation of the probability.

The Poisson model

- When rare events occur together or in clusters, people often want to know if that happened just by chance or whether something else is going on.
- Binomial probability could be difficult to calculate in when n is too big (you should calculate $n!$, for example).
- Simeon Denis Poisson (French mathematician) derived his model to approximate the Binomial model when the probability of a success, p, is very small and the number of trials, n, is very large.

POISSON PROBABILITY MODEL FOR SUCCESSES: *Poisson* (λ)

λ = mean number of successes.
X = number of successes.

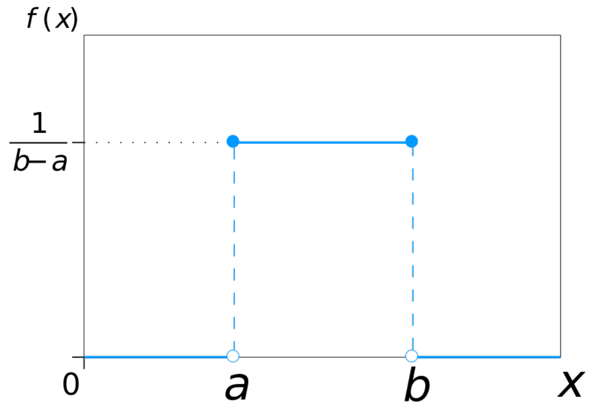
$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

Expected value: $E(X) = \lambda$
Standard deviation: $SD(X) = \sqrt{\lambda}$

The Poisson model is a reasonably good approximation of the Binomial when $n \geq 20$ with $p \leq 0.05$ or $n \geq 100$ with $p \leq 0.10$.

- To use the Poisson model to approximate the Binomial, we need to set $\lambda = np$

The Uniform Model



- Probability model for the continuous uniform random variable:
  - $f(x)\begin{cases}\frac{1}{b-a} & if\ a \leq x \leq b \\ 0 & otherwise\end{cases}$

- For values c and d ($c \leq d$) both within the interval $[a, b]$
  - $P(c \leq x \leq d) = \frac{(d-c)}{(b-a)}$
  - $E(X) = \frac{a+b}{2}$
  - $Var(X) = \frac{(b-a)^2}{12}$

The Exponential Model

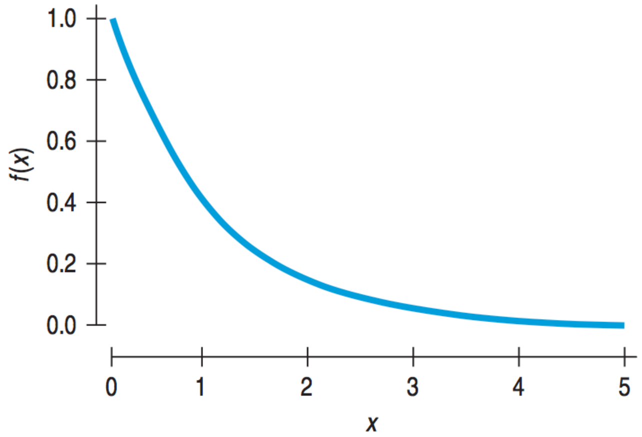
- The Poisson model is a good model for the arrival, or occurrence, of events.
  - E.g., we can use the Poisson model to model the probability of x visits to our website within the next minute.
- Then the exponential model with parameter λ can be used to model the time between the events.

$$f(x) = \lambda e^{-\lambda x} \quad for\ x \geq 0\ and\ \lambda > 0$$

- Mean and standard deviation of the exponential is $1/\lambda$

$$P(s \leq X \leq t) = e^{-\lambda s} - e^{-\lambda t}.$$

$$P(X \leq t) = P(0 \leq X \leq t) = e^{-\lambda 0} - e^{-\lambda t} = 1 - e^{-\lambda t}.$$



The exponential probability model (with $\lambda = 1$). The probability that x lies between any two values corresponds to the area under the curve between the two values.

Key Points I

Chapter 17: Probability models

- Bernoulli trials: two possible outcomes with probability, independent trials
- **Geometric model:** how many trials do we need to get a specific outcome?
  - $P(X = x) = q^{x-1}p,\ \ E(X) = \mu = \frac{1}{p},\ \sigma = \sqrt{\frac{q}{p^2}}$
- **Binomial model:** Among *n* trials, what is the probability of getting a specific outcome *x* times?
  - $P(X = x) = {}_nC_x p^x q^{n-x},\ E(X) = \mu = np,\ \ \sigma = \sqrt{npq}$

Key Points II

Chapter 17: Probability models

- When $np \geq 10$ and $nq \geq 10$, the binomial model is approximately Normal.
- Poisson model: when *p* is very small and *n* is very large.
    - $P(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!},\ E(X) = \lambda,\ SD(X) = \sqrt{\lambda}$, for approximating binomial, $\lambda = np$
- Uniform model:
    - $P(c \leq x \leq d) = \frac{(d-c)}{(b-a)}, E(X) = \frac{a+b}{2}, SD(X) = \sqrt{\frac{(b-a)^2}{12}}$
- Exponential model:
    - $P(s \leq x \leq t) = e^{-\lambda s} - e^{-\lambda t}, E(X) = \frac{1}{\lambda}, SD(X) = \frac{1}{\lambda}$