6. THE BINOMIAL DISTRIBUTION

**Eg:** For 1000 borrowers in the lowest risk category (FICO score between 800 and 850), what is the probability that at least 250 of them will default on their loan?

**Eg:** Ana interviews with 3 firms. The probability that she will get an offer from a given firm is 0.7. The decision of each firm is independent of the others. What is the probability that Ana will get exactly two offers? At least one offer? No offers?
Eg: Suppose you play five rounds of the “Let’s Make a Deal” game, switching each time. What is the probability that you lose all five rounds? How does the answer change if you stick with your original choice each time?

Eg: In a large batch of computer chips, 0.5% are defective. You randomly select 100 for inspection. How many bad chips will you get? How likely are you to find at least one bad chip?
**Eg:** In how many of the next 10 years will NYC experience a hurricane at least as damaging as Hurricane Sandy?

**Eg:** What is the probability that an analyst will correctly forecast the direction of the Dow for at least 7 of the next 10 trading days, assuming that the Dow is a random walk, and therefore changes in the Dow are actually not forecastable?
In each of these examples, we have an experiment consisting of a fixed number, \( n \), of “trials,” or repetitions.

There are two possible outcomes at each trial, which may be denoted by “success” and “failure”.

The results of different trials are independent.

The probability \((p)\) of success on a given trial remains constant for all trials.
We are interested in the distribution of the random variable
\[ X = \text{Total number of successes.} \]

The possible values for \( X \) are \( x = 0, 1, \cdots, n \).

Under the above conditions, \( X \) has a \textit{binomial distribution}, and the probability of \( x \) successes in \( n \) trials is

\[
p(x) = \binom{n}{x} p^x q^{n-x}
\]

\( x = 0, 1, \cdots, n \),

where \( q = 1 - p \), and \( \binom{n}{x} \) is a \textit{binomial coefficient}. (Next slide.)
• **Factorial Notation:** If \( n \) is a positive integer, define \( n! \) ("\( n \) factorial") by

\[
n! = n (n-1) (n-2) \cdots (2) (1).
\]

Define \( 0! = 1 \).

Eg: \( 3! = 3 \cdot 2 \cdot 1 = 6 \).

• **Binomial Coefficient:** The number of different ways of choosing \( x \) objects from a total of \( n \) objects is

\[
\binom{n}{x} = \frac{n!}{x!(n-x)!} = \frac{n(n-1)\cdots(n-x+1)}{x!}
\]

("\( n \) choose \( x \") if the order does not matter.
• For a set of 1000 stocks on the NYSE there are more than 8.25 trillion mutual funds consisting of an equal dollar amount of 5 stocks.

**Eg:** In the Ana example, we have $n = 3$, $p = .7$, so that

\[
p(0) = (0.3)^3 = 0.027, \quad p(1) = 3(0.7)^1(0.3)^2 = 0.189, \quad p(2) = 3(0.7)^2(0.3)^1 = 0.441, \quad p(3) = (0.7)^3 = 0.343.
\]

There is a 44.1% chance that Ana gets two offers.

She gets at least one offer with probability $p(1) + p(2) + p(3) = 0.973$.

There is just a 2.7% chance that she will get no offers.
**Eg:** In tossing a coin 10 times, you are much more likely to get 5 heads than 0 heads. The reason is that there are many more ways to get 5 heads than 0 heads (252, compared to 1).

Note: \( \binom{n}{x} = \frac{(10)(9)(8)(7)(6)}{[(5)(4)(3)(2)(1)]} = 252. \)

- Each pair of values \((n, p)\) determines a distinct binomial distribution.

The shape of a binomial distribution:

- \( p < 0.5 \) skewed to the right
- \( p = 0.5 \) symmetric
- \( p > 0.5 \) skewed to the left

[R Demo: BinomialBarPlots]
We can use tables giving the cumulative probabilities for the binomial distribution,

\[
\Pr(X \leq k) = \sum_{x=0}^{k} p(x).
\]

<table>
<thead>
<tr>
<th>(n = 10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>9.044</td>
</tr>
<tr>
<td>0.9957</td>
</tr>
<tr>
<td>0.9999</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
<tr>
<td>1.0000</td>
</tr>
</tbody>
</table>
**Eg 1:** Compute the probability of getting at least 7 correct forecasts out of 10 for the direction of the Dow, assuming that the Dow is a random walk (efficient market), so that the direction is actually not predictable.
**Eg 2:** A basketball foul shooter has been averaging 80% from the line. Assuming his skill stays the same, what is the probability that at least 10 of his next 15 foul shots will be good?

<table>
<thead>
<tr>
<th>k</th>
<th>0.01</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>0.95</th>
<th>0.99</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.8601</td>
<td>0.4633</td>
<td>0.2059</td>
<td>0.0352</td>
<td>0.0047</td>
<td>0.0005</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>0.9904</td>
<td>0.8290</td>
<td>0.5490</td>
<td>0.1671</td>
<td>0.0353</td>
<td>0.0052</td>
<td>0.0005</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>0.9996</td>
<td>0.9638</td>
<td>0.8159</td>
<td>0.3980</td>
<td>0.1268</td>
<td>0.0271</td>
<td>0.0037</td>
<td>0.0003</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>1.0000</td>
<td>0.9945</td>
<td>0.9444</td>
<td>0.6482</td>
<td>0.2969</td>
<td>0.0905</td>
<td>0.0176</td>
<td>0.0019</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>1.0000</td>
<td>0.9994</td>
<td>0.9873</td>
<td>0.8358</td>
<td>0.5155</td>
<td>0.2173</td>
<td>0.0592</td>
<td>0.0093</td>
<td>0.0007</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9978</td>
<td>0.9389</td>
<td>0.7216</td>
<td>0.4032</td>
<td>0.1509</td>
<td>0.0338</td>
<td>0.0037</td>
<td>0.0001</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>6</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9977</td>
<td>0.9819</td>
<td>0.8689</td>
<td>0.6098</td>
<td>0.3036</td>
<td>0.0950</td>
<td>0.0152</td>
<td>0.0008</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>7</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9958</td>
<td>0.9500</td>
<td>0.7869</td>
<td>0.5000</td>
<td>0.2131</td>
<td>0.0500</td>
<td>0.0042</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>8</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9992</td>
<td>0.9848</td>
<td>0.9050</td>
<td>0.6964</td>
<td>0.3902</td>
<td>0.1311</td>
<td>0.0181</td>
<td>0.0003</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>9</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9963</td>
<td>0.9962</td>
<td>0.8491</td>
<td>0.5968</td>
<td>0.2784</td>
<td>0.0611</td>
<td>0.0022</td>
<td>0.0001</td>
<td>0.0000</td>
</tr>
<tr>
<td>10</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9993</td>
<td>0.9907</td>
<td>0.9408</td>
<td>0.7827</td>
<td>0.4845</td>
<td>0.1642</td>
<td>0.0127</td>
<td>0.0006</td>
<td>0.0000</td>
</tr>
<tr>
<td>11</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9999</td>
<td>0.9981</td>
<td>0.9824</td>
<td>0.9095</td>
<td>0.7031</td>
<td>0.3518</td>
<td>0.0556</td>
<td>0.0055</td>
<td>0.0000</td>
</tr>
<tr>
<td>12</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9997</td>
<td>0.9963</td>
<td>0.9729</td>
<td>0.8732</td>
<td>0.6020</td>
<td>0.1841</td>
<td>0.0362</td>
<td>0.0004</td>
</tr>
<tr>
<td>13</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9995</td>
<td>0.9948</td>
<td>0.9647</td>
<td>0.8329</td>
<td>0.4510</td>
<td>0.1710</td>
<td>0.0096</td>
<td>0.0000</td>
</tr>
<tr>
<td>14</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9995</td>
<td>0.9953</td>
<td>0.9648</td>
<td>0.7941</td>
<td>0.5367</td>
<td>0.1399</td>
<td>0.0000</td>
</tr>
</tbody>
</table>
Eg: Logic analyzers come off the assembly line with a 3% defective rate. You must ship 17 of these analyzers tomorrow. How many analyzers should you schedule for production today in order to be reasonably sure that 17 or more of the scheduled machines will work?

Sol: If you schedule $n$ machines, then $X = \#\text{Working Machines}$ has a binomial distribution with $p = 0.97$. 
• If you schedule 17 machines (no margin for error), you might think that the high (97%) rate would help you, but in fact the probability that all 17 machines will work is just 0.596. So by leaving no margin for error, you are taking a 40.4% chance that you will fail to ship the entire order in working condition.

• If you schedule 18 machines, then
\[ P(\text{At least 17 machines work}) = 0.900. \]

• If you schedule 19 machines, then
\[ P(\text{At least 17 machines work}) = 0.982. \]

• Conclusion: You’d better schedule at least 19 machines to get 17 good ones.

Note: Since the table doesn’t cover these values of \( n \), I computed the binomial probabilities by hand. Try it yourself!
If $X$ is binomial $(n, p)$ then

$$E(X) = \mu = np,$$

$$\text{Var}(X) = \sigma^2 = npq.$$  

**Eg:** For $n = 17$, $p = 0.97$, we get $\mu = 16.49$, $\sigma = 0.703$. The relatively large value of $\sigma$ helps to “explain” why Prob(17 successes) is so low (0.596), despite the large value of $p$. 

**Mean and Variance of Binomial Random Variables**
Eg 3: The manufacturer of peanut M&Ms claims that only one bag out of 10 will contain any defectives. We inspect 5 bags, and 3 contain defectives. Do we reject the manufacturer’s claim? How many defective bags should be expected? With what variation?
The Sample Proportion

If X is binomial \((n, p)\) then the sample proportion is \(\hat{p} = \frac{X}{n}\). The mean and standard deviation of \(\hat{p}\) are

\[
E(\hat{p}) = p \\
SD(\hat{p}) = \sqrt{\frac{pq}{n}}
\]

If \(p = q = 1/2\), we get the worst-case value, \(SD(\hat{p}) \leq \frac{1}{\sqrt{4n}}\).

The "margin of error" often given in opinion polls is

\[
MOE = \pm 1.96 / \sqrt{4n}
\]

This is based on the empirical rule: There is a 95% chance that the true proportion \(p\) will be within 1.96 standard deviations of the random variable \(\hat{p}\).
Eg: “Goldman Sachs surveyed 752 Millennials, asking whether they prefer to pay for purchases using PayPal. The margin of error in the poll was 3.5 percent, according to the firm.”

We can compute $\text{MOE} = \frac{1.96}{\sqrt{4(752)}} = .035$. 