Unit 5 - Probability Models for Counts / Discrete Quantities

Adam Petrie
Department of Business Analytics
University of Tennessee

June 12, 2021

Motivation - Probabilistic Modeling

Averages Aren't Good Enough

Overview

Customer lifetime value is a formula that helps a marketing manager arrive at the dollar value associated with the long-term relationship with any given customer, revealing how much a customer relationship is worth over a period of time.

It is useful for considering the customer acquisition process and selecting optimal service levels to different customer groups.

Scenario

Michelle is the director of marketing for M & J Jewelry, a direct-mail jewelry business that sells high-end costume jewelry via monthly catalogs.

Scenario

She wants to better understand the lifetime value of the customers who buy jewelry from M & J so that she can determine whether she should implement specific marketing initiatives aimed at increasing the retention rate of some portion of her customer base.

- How many items per order?
- How many purchases per year?
- How many total purchases before "churning"?

She asks the head of her accounting department for some data to support the analysis.

Data

Accounting reports back the following averages:

- Average catalog customer spends $250 per purchase
- Gross margin associated with these sales is 40%, so average profit per customer purchase is $100
- Average customer buys from the catalog 4 times per year
- Average annual cost of mailing the catalogs is $9.60 per customer per year
- Cost of acquiring a new customer is $480 (cost of sending catalog divided by response rate)
- Average customer retention rate is 65% from one year to next
- A discount rate of 10% should be used in analysis to adjust future dollars to their value today

CLV Formula

The following equation can be used to calculate lifetime value:

$$-AcquisitionCost + \sum_{i=1}^{years} \frac{GrossMargin \times SpendPerYear - MailingCost}{(1 + DiscountRate)^i}$$

- AcquisitionCost is the one-time cost of acquiring a customer (fixed)
- years is the number of years the customer stays with company (random)
- GrossMargin is the percent that company makes off a purchase (fixed)
- SpendPerYear is the amount of money a customer spends on purchases during a year (random)
- MailingCost is the cost to mail out catalogs to customers over the course of a year (fixed)
- DiscountRate is fixed, often taken to be 10%

"Summation" notation

The Σ notation is just shorthand for "sum over ...". It's very much like a `for` loop in programming, just with math! The name of the "looping variable" is specified below the symbol (i in this case), and the "looping vector" is the integer sequence starting with value specified below the symbol going up to the value specified above the symbol.

$$\sum_{i=1}^{4} i(i+1) = 1 \cdot (1+1) + 2 \cdot (2+1) + 3 \cdot (3+1) + 4 \cdot (4+1) = 40$$

```
s <- 0
for (i in 1:4) { s <- s + i*(i+1) }
s
## [1] 40
```

CLV Formula Example for "average" customer

Let's use customer-wide averages for any random quantities.

$$-AcquisitionCost + \sum_{i=1}^{years} \frac{GrossMargin \times SpendPerYear - MailingCost}{(1 + DiscountRate)^i}$$

- AcquisitionCost is \$480 (fixed)
- years is 3 (average value)
- GrossMargin is 40% (fixed)
- SpendPerYear is \$1000 (average value)
- MailingCost is \$10 (fixed)
- DiscountRate is 10% (fixed)

$$-480 + \frac{1000 \times 0.40 - 10}{(1 + 0.1)^1} + \frac{1000 \times 0.40 - 10}{(1 + 0.1)^2} + \frac{1000 \times 0.40 - 10}{(1 + 0.1)^3}$$

```
-480 + sum( (1000*.4-10)/(1+.1)^(1:3) )
## [1] 489.8723
```

CLV Formula Example for "average" customer

So the average customer is "worth" about \$490 right?

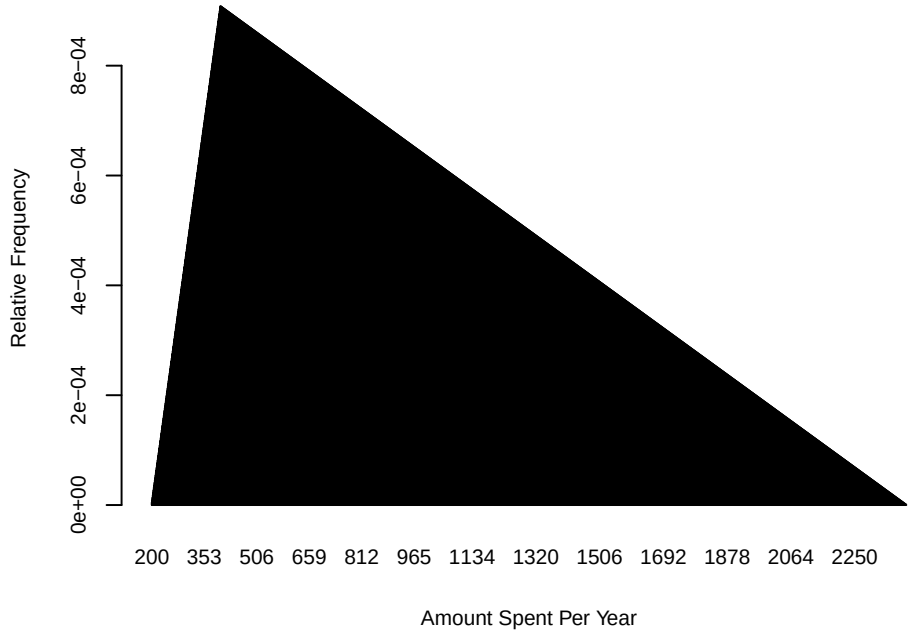
Not necessarily

Lifetime value is the sum of a random number of random amounts that have been spent. It turns out that the average lifetime value in complex cases like this is *not* what you get simply by plugging in the averages for each quantity into the lifetime value formula.

Can't always plug in average values of random quantities into formula

Imagine the # years a customer shops is equally likely to be 1-5 (so that the average is 3).

Imagine that the amount spent per year has an average of $1000, but has the following shape (between 200 and 2400, peaking at 400).



Can't always plug in average values of random quantities into formula

What's the actual average lifetime value of customers under this model? Monte Carlo it. Note: `dtriangle` is not native to R but the code to define it is in the associate .R file.

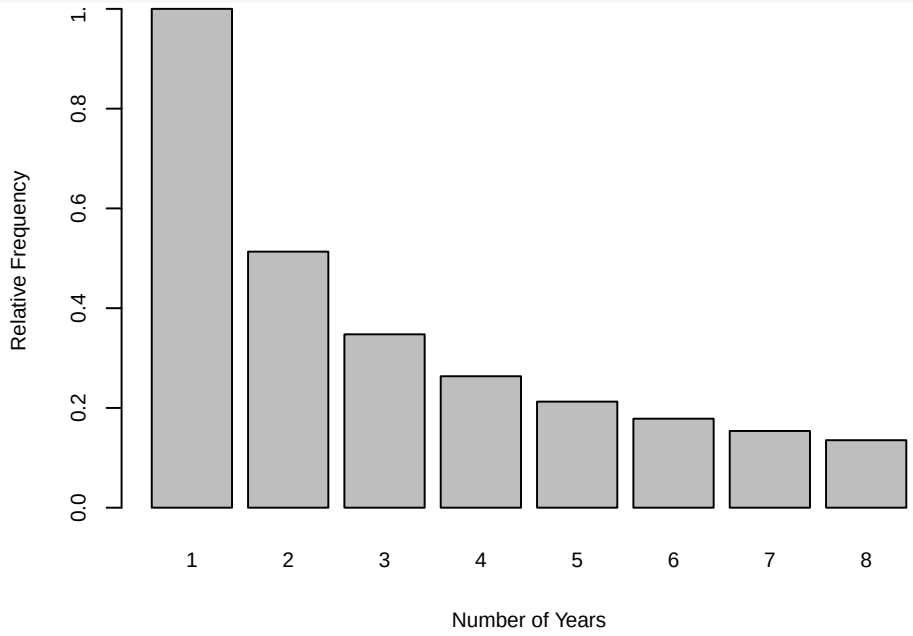
```
values <- 200:2400
probs <- dtriangle(values,min=min(values),peak=400,max=max(values))
simCLV <- rep(0,1e5)
for (i in 1:1e5) {
   years <- sample(1:5,size=1)  #randomly pick 1-5 for number of years
   spent <- sample(values,size=years,prob=probs,replace=TRUE)   #pick spent amount for each
   simCLV[i] <- -480 + sum((spent * 0.4 - 10)/(1 + 0.1)^(1:years))
 }
 mean(simCLV)
## [1] 463.2743
```

Incorporating variation around the average shows the actual average is more like \$460.

Can't always plug in average values of random quantities into formula

What if we used different distributions? How about years is between 1-8 and larger values are less likely but with an average of 3? How about yearly spent amount is either \$250 (40% probability) or \$1500 (60% probability), so the average is \$1000?

```
mean( sample(1:8,size=1e6,replace=TRUE, prob=1/(1:8)^.962 ) ) #Average number of years
## [1] 3.002097
```



Can't always plug in average values of random quantities into formula

The average customer lifetime value with this model is around $430. That's *really* far below $490.

```
simCLV <- rep(0,1e5)
for (i in 1:1e5) {
   years <- sample(1:8,size=1,prob=1/(1:8)^.962 )
   spent <- sample(c(250,1500),size=years,replace = TRUE,prob=c(.4,.6))
   simCLV[i] <- -480 + sum((spent * 0.4 - 10)/(1 + 0.1)^(1:years))
 }
 mean(simCLV)
```

```
## [1] 433.0724
```

Lessons

Most processes worth studying in business analytics are complex with multiple random components.

> Summarizing each random component in a process by its average can yield a very misleading impression of the average outcome of that process.

For complex processes like finding the customer lifetime value, it behooves us to understand the full range of variability of each component – what values might be observed, and how often?

Questions We Need To Answer

- The numbers quoted are averages, but not all customers are the same and these averages probably are different for various *types* of customers
- Can we model the amount spent per purchase by some distribution? The # purchases? The # years? The retention rate?
- Once we are comfortable with the procedure we can change the analysis to account for different customer segments.

Understanding the underlying *distributions* of each quantity allow us to come up with a more accurate estimate of the average customer lifetime value and to explore how much variation we could expect to see (e.g., worst and best case scenarios).

Related question - catalog sending

When should the company stop sending catalogs to customers who haven't been making purchases?

- Need a distribution of the number of quarters without purchases to get a handle over what is typical vs. atypical.
- Maybe need a (conditional) distribution of the purchase amounts after a certain number of quarters with no purchases.

The decision is a cost/benefit analysis.

Refinery optimization under uncertainty

Consider linear programming under uncertainty. A refinery produces gas and jet fuel. It has to satisfy demand constraints (at least 4000 barrels of gas, 5000 barrels of jet fuel), but has finite truck capacity for delivery (at most 50,0000 barrel-miles) and finite product capacity (at most 14,000 barrels). The objective is to maximize profit.

- $G \geq 4000$ and $J \geq 5000$
- $G + J \leq 14,000$
- $10G + 20J \leq 200,000$

But what if constraints are random (demand might be less/more, refinery/trucks subject to breakdowns to capacity is less than expected. It may be more informative to run a Monte Carlo simulation exploring optimal solutions.

More motivation - Kroger Customers

Imagine we are studying the number of days between sequential shoppers visits to Kroger, but our dataset is small.

We might be able to estimate the probability this is 0, 1, 2, 3, ..., 15 pretty well (perhaps we saw each value a few hundred times), but what do we do about numbers larger than this that we haven't observed often enough to be comfortable with the estimated probabilities?

Are there patterns in the probabilities we *can* estimate that allow us to "guess" the ones we couldn't estimate?

Goal of Modeling

We *could* set up Monte Carlo simulation to estimate a specific probability regarding a random process. But we might have a *lot* of questions.

Is there a table of possible values for each outcome that can be observed along with "reasonable" probabilities for each?

Is there a mathematical equation that does a "reasonable" job of capturing the probability of observing each possible value?

If so, we could use those probabilities to answer a slew of questions about the process and that would save us a lot of time!

Why we model

Why is this unit necessary? Why is it still important to learn about probability models when modern business analytics datasets are so big that we have a reasonable idea of how often each possible observed value might be?

- The data isn't always large enough, and some possible values haven't been observed. A probability model can take us beyond what has has happened to what plausibly *might* happen.
- Using data, *each* estimated probability is always subject to error. A probability model often consolidates this collection of uncertainties into an uncertainty about one or two numbers.
- Many random processes commonly studied in business analytics have known probability distributions! Using the model saves time.

Goals for this Unit

By the end of this unit, you should be able to:

- Develop a custom probability model for any "discrete quantity" (usually a count, e.g., 0, 1, 2, 3, ...)
- Use a probability model to calculate the probability of an event (one or more possible outcomes)
- Propose a "notorious distribution" to model counts; of keen interest in business analytics
- Use the probability model to calculate the probability of various events (one or more outcomes)
- Calculate the average and standard deviation of a random process from a probability model and be able to interpret each in layman's
- Check to see if proposed model provides a "reasonable reflection of reality"

Discrete Probability Models

Types of Random Quantities

Discrete quantities

- Categorical variables (red/green/blue)
- Counts (0, 1, 2, ...).
- How can you tell if a numerical variable is discrete? There is some finite minimum spacing between possible values (e.g., amount of change in someone's pocket; increments are in 1 cent).

Continuous quantities

- Always numerical.
- Possible values are a "near-continuum" of numbers (no finite minimum spacing between possible values).
- Examples: distance, mass, time, temperature, etc., assuming you had a measurement device with infinite precision.

The lines between discrete and continuous often get blurred

It is possible to treat a continuous distribution as discrete by lumping values into intervals/categories.

- Time could be short (0-10), medium (10-30), and long (30+)
- Temperatures could be rounded to the nearest degree

Discrete distributions are sometimes treated as continuous when spacings between possible values are small with respect to the range of typical values themselves.

- Populations of cities. Spacing between possible values (1) is small compared to range of typical values (1000 to a few million).
- Money. Spacing between possible values (0.01 dollars) may be small compared to range of typical values (a few dollars up to hundreds, thousands, millions?)

Random Variables

A **random variable** is a quantity that can take on a variety of different values, each having their own probability. By convention, a random variable is denoted with an upper-case letter (usually X; called "big X") and a particular value it may take is denoted with a lower-case letter (usually x; called "little x"). The possible values of x are always mutually exclusive.

- X = color of stoplight (x = red, green, yellow)
- X = # of students absent from class (x = 0, 1, ..., all)
- X = number out of 6 promotions mailed to a person that they eventually use (x = 1-6)
- X = number of purchases in a year (x = 0, 1, 2, ...)

Probability Mass Function (PMF)

The probability mass function of a random variable X is a list of its possible outcomes along with their associated probabilities. It may be given as a table or by a formula.

- $X =$ color of stoplight

x	$P(X = x)$
green	0.35
red	0.60
yellow	0.05

- $X =$ number found after rolling a fair die

$$P(X = x) = 1/6 \quad \text{for } x = 1, 2, \ldots, 6$$

- $X =$ number out of 4 promotions person used (0-4)

$$P(X = x) = 0.34 \times (0.75)^x \text{ for } x = 0, 1, \ldots, 4$$

Requirements and properties of PMF

For a table or equation to give a valid PMF:

- The probability for each outcome x must be between 0 and 1 (all probabilities have this property).
- The sum of the probabilities over all outcomes must equal 1 (i.e., the PMF must assign probabilities to every possible outcome).
- Since outcomes x are mutually exclusive (and with obvious extensions to more than 2 events)

$$P(X = x_1 \text{ or } x_2) = P(X = x_1) + P(X = x_2)$$

- If x is *not* a possible outcome, $P(X = x) = 0$, though this is usually omitted when specifying the PMF.

Designer PMFs

You can use these properties to design a custom PMF to model a quantity of interest. Let $X = \#$ of customer service calls someone places in a month.

x	$P(X = x)$
0	0.92
1	0.03
2	0.02
3	?

It must be that $P(X = 3)$ is 0.03 for the probabilities to sum to 1. Note that this model implicitly assumes the probability of getting 4 or more service calls is 0.

Shape to PMF

Often, we'll have idea of what we want the *shape* of a PMF to look like (basically the relative frequencies or weights of each value).

A key task is converting a shape into a PMF. To do so, we ensure the set of relative frequencies that describe the shape add to 1 so that they represent a collection of valid probabilities.

Shape to PMF example

Imagine modeling people bidding on an auction by picking random numbers between 1-100, with the condition that larger numbers were increasingly more likely.

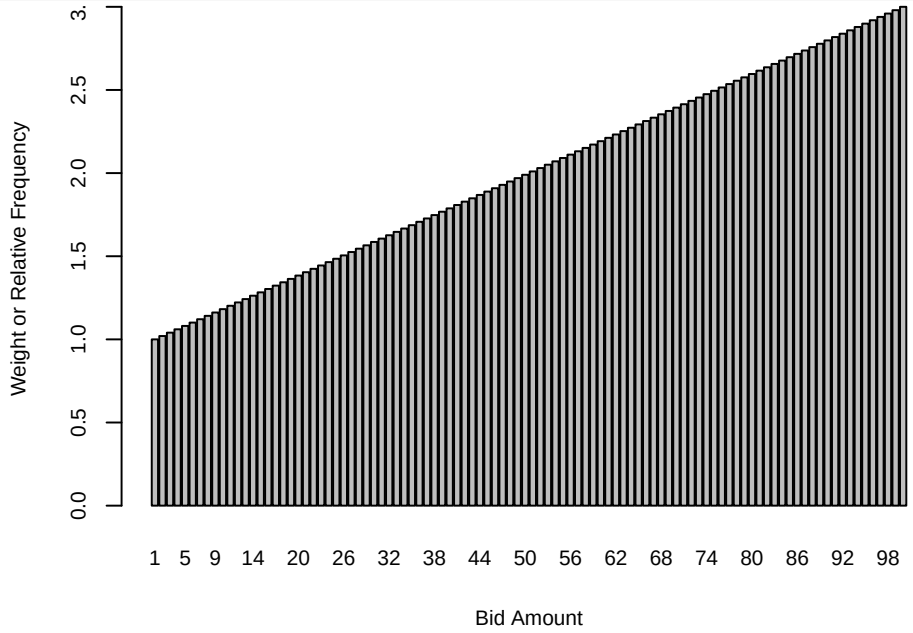
- A 100 was 3 times more likely than a 1.
- The increase in relative frequencies increases linearly with the number under consideration.

The `sample` command was able to easily incorporate these requirements by assigning weights to each outcome 1-100 via: `seq(from=1,to=3,length=100)`.

Shape to PMF example

The sequence of weights assigned to each outcome gives us the "shape" we want, but those numbers aren't probabilities since they do not sum to 1!

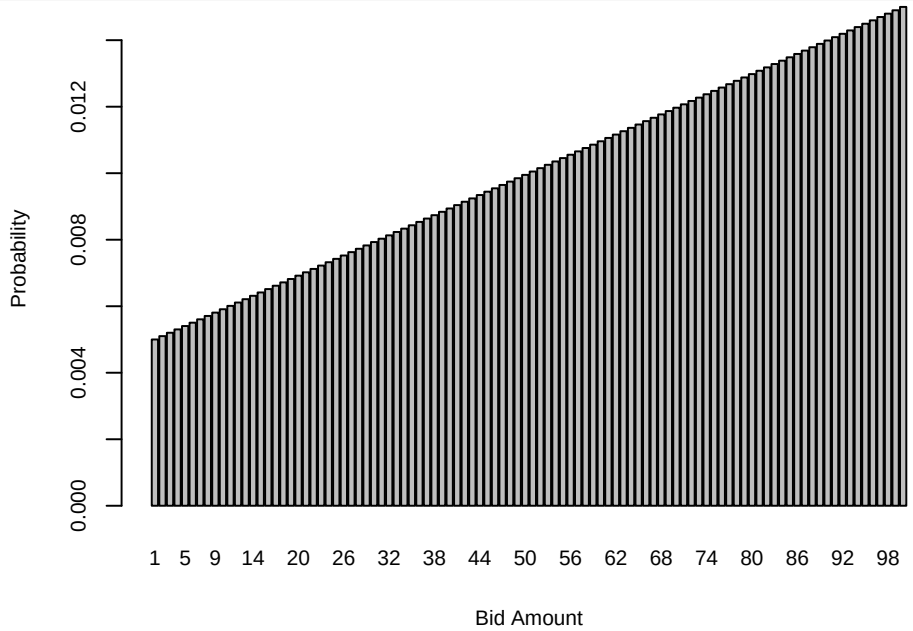
```
weights <- seq(from=1,to=3,length=100)
barplot( weights, names.arg=1:100,xlab="Bid Amount",ylab="Weight or Relative Frequency")
sum( weights )
## [1] 200
```



Shape to PMF example

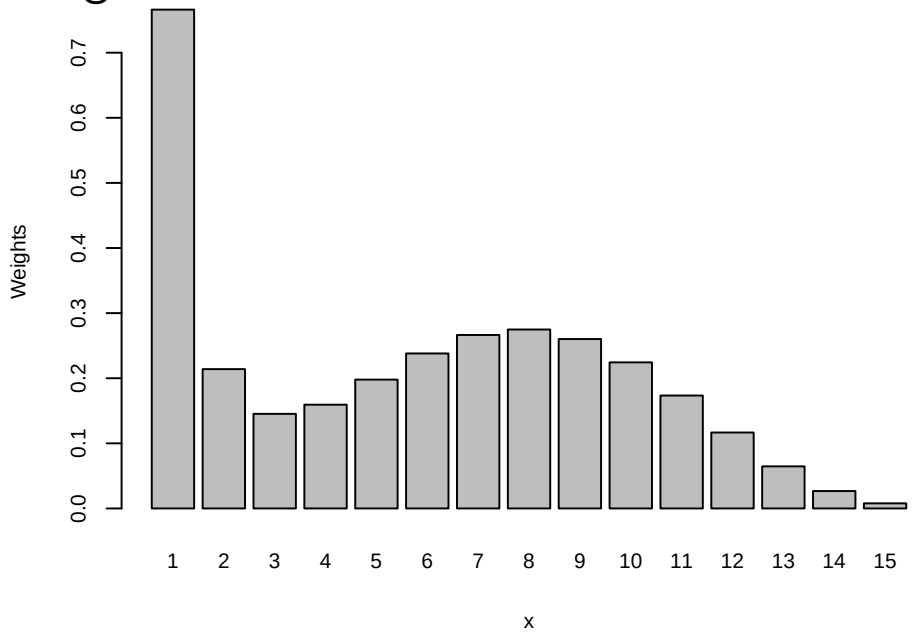
Solution: "rescale" the weights so that they *do* sum to 1! This is always possible by dividing each weight by the total sum of weights. This keeps the shape the same while turning the values in a set of probabilities.

```
p <- weights/sum(weights)
barplot( p , names.arg=1:100,xlab="Bid Amount",ylab="Probability")
sum( p )
## [1] 1
```



Another Shape to PMF example

The following shape does a reasonable job of capturing the relative frequencies of the number of purchases customers make in a year. This shape is NOT a PMF because the weights of each value don't sum to 1.



Another Shape to PMF example

In this case, we have an equation that gives the shape of the PMF.

$$Shape = \frac{1}{x^{0.2}(16-x)^{0.1}} + \frac{x^3(16-x)^3}{1000000} \quad x = 1, 2, 3, \ldots, 15$$

However, since the values don't sum to 1, these are not a valid set of probabilities.

```
x <- 1:15
shape <- 1/(x^2*(16-x)^0.1) + x^3*(16-x)^3/1000000
sum(shape)
## [1] 3.135683
```

Another Shape to PMF example

It's easy to convert a shape to a PMF – just divide each weight by 3.135683 (the total sum of all the original weights).

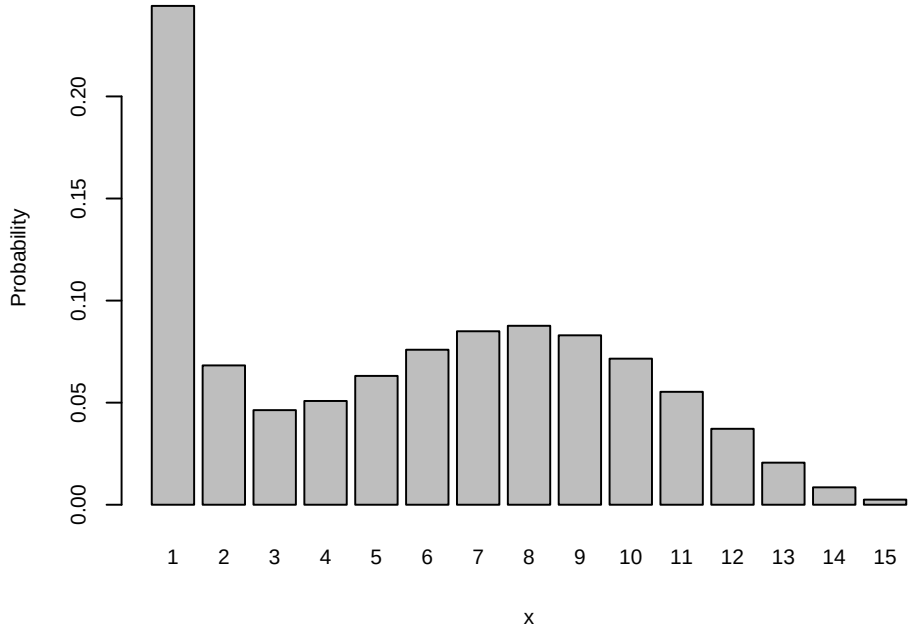
$$P(X = x) = \frac{1}{3.135683}\left(\frac{1}{x^{0.2}(16-x)^{0.1}} + \frac{x^3(16-x)^3}{1000000}\right) \quad x = 1, 2, 3, \ldots, 15$$

```
x <- 1:15
shape <- 1/(x^2*(16-x)^0.1) + x^3*(16-x)^3/1000000
p <- shape/sum(shape)
sum(p)
```

```
## [1] 1
```

Another Shape to PMF example

```
barplot( p, names.arg=x,xlab="x",ylab="Probability")
```



Converting a shape equation into a PMF procedure

- Define a vector `x` to contain the possible values from the distribution.
- Define a vector `shape` to contain the weights / relative frequencies for each value in `x` (usually this just involves typing out the equation for the weights in terms of `x`).
- Define a vector `p` to be `shape/sum(shape)`.
- Voila! The vectors `x` and `p` define a valid PMF!

Streamlined Conversion of Shape to PMF via equations

If you have an equation giving the shape of the PMF, and you want the equation for the PMF itself, no problem.

- Define a vector `x` to contain the possible values that might be observed.
- Define a vector `shape` to contain the weights / relative frequencies for each possible value using the equation for the shape involving `x`.
- Find `sum(shape)`, call that a.

$$P(X = x) = \frac{1}{a} \cdot \textit{equation for shape} \qquad x = \text{some set of values}$$

Streamlined Conversion of Shape to PMF via equations

Example:

$Shape = x^2(20 - x) + 1$ and $x = 0, 1, 2, \ldots, 20$

```
x <- 0:20
shape <- x^2*(20-x) + 1
sum(shape)
```

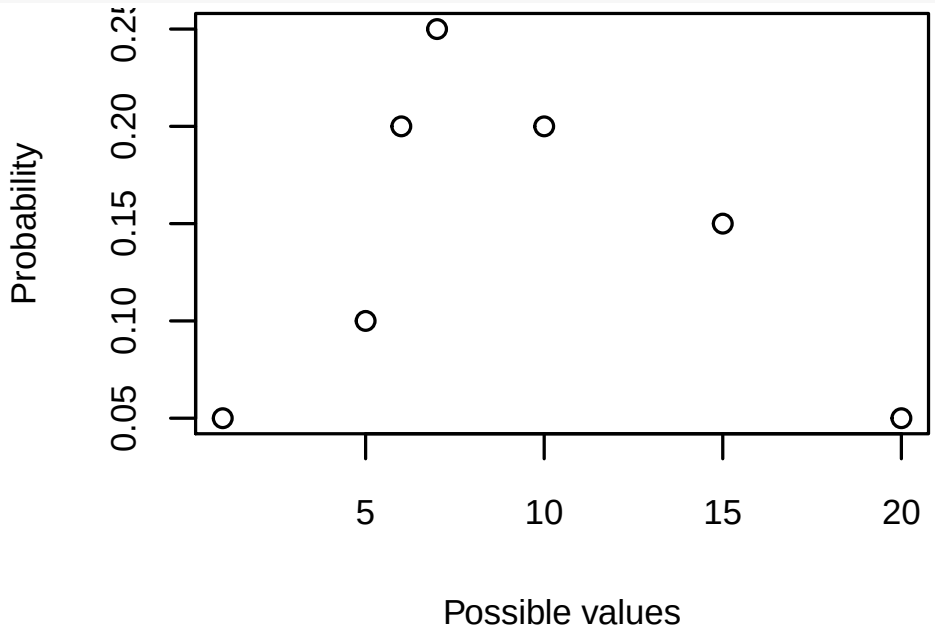
```
## [1] 13321
```

$$P(X = x) = \frac{1}{13321} \cdot \left(x^2(20 - x) + 1\right) \qquad x = 0, 1, 2, \ldots, 20$$

Plotting a PMF

One way to to visualize a PMF is to use `plot()`. This is a great way to look at the distribution if the values of `x` are not a continuous integer sequence.

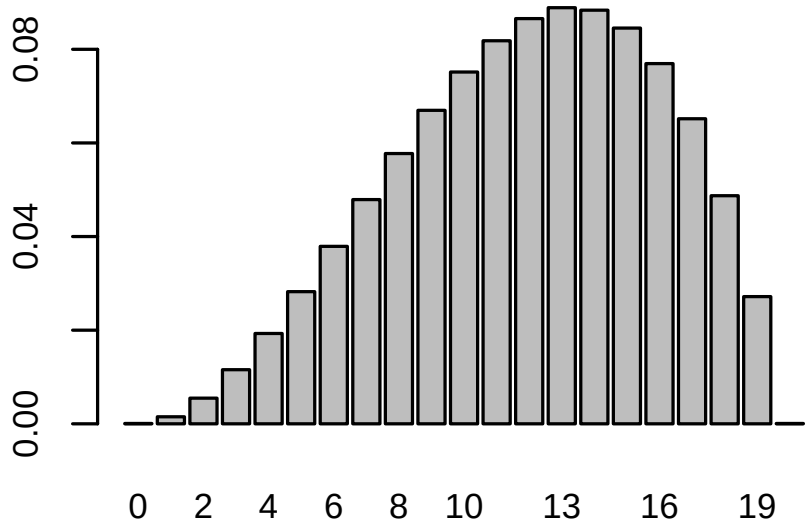
```
x <- c(1, 5, 6, 7, 10, 15, 20); p <- c(.05, .1, .2, .25, 0.2, 0.15, 0.05)
plot(p~x,xlab="Possible values",ylab="Probability")
```



Using barplot to plot a PMF

If `x` is a continuous integer sequence (or a set of levels of a categorical variable), a nicer way to visualize the PMF is by using `barplot`. If `x` and `p` are the vector of names of possible outcomes and associated probabilities (respectively) try:

```
x <- 0:20; shape <- x^2*(20-x) + 1; p <- shape/sum(shape)
barplot(p,names.arg=x)
```



Cumulative Distribution Function (CDF)

The CDF of X, denoted $F(x)$ is the probability that X achieves a value of *at most* x. If x_k represents the set of possible outcomes of X (e.g., x_1 might be 10, x_2 might be 15, etc.), then

$$F(x) = P(X \leq x) = \sum_{\text{all } x_k \leq x} P(X = x_k)$$

$$F(x) = \text{Probability of observing a value of at most } x$$

$$F(x) = \text{Probability of observing a value of } x \text{ or less}$$

Example - number of people seated at a 4-top in a restaurant

x	$P(X = x)$	$F(x) = P(X \leq x)$
1	0.10	0.10
2	0.35	0.45
3	0.15	0.60
4	0.40	1.0

Example calculation:

$$F(3) = P(X \leq 3) = P(X = 1) + P(X = 2) + P(X = 3) = 0.10 + 0.35 + 0.15 = 0.6$$

Uses of the CDF

Knowing the values of the CDF (as opposed to the PMF) can be valuable in analytics.

- Stockout probabilities. If you know the PMF of weekly demand and wanted the probability of a stockout to be no more than 93%, we could find the smallest x (starting week inventory) so that $F(x) \geq 0.93$.
- Auctions. If you know the PMF of the highest competing bid and you wanted the probability of winning (or tying for highest bid) to be at least 45%, you could find the smallest x (your bid) so that $F(x) \geq 0.45$.
- Waiting time. If you know the PMF of the number of trials it takes for a rare event to occur, you might want to calculate the probability it occurs by trial 1000; this equals $F(1000)$.

Be very mindful of $\leq$ vs. $<$

When working with discrete distributions, pay particular attention to $\leq$ vs. $<$!

If the possible values of X are 0, 1, 2, 3, ..., then there is a BIG difference between $P(X \leq 2)$ (which includes 0, 1, and 2) and $P(X < 2)$ (which includes only 0 and 1).

Likewise, there is a big difference between $\geq$ vs. $>$. When we are dealing with counts, if you are asked $P(X > 3)$, that is equivalent for $P(X \geq 4)$

Note on CDF definition

The PMF $P(X = x)$ is non-zero only at particular values (e.g., 1, 2, 3, 4 for the people seated at a four-top model).

The CDF $F(x)$ is non-zero even for values of x that do not correspond to possible values that X can take.

For example, consider the seating PMF and $x = 3.5$. We have that $P(X = 3.5) = 0$ since 3.5 people can't sit at a table. However, $F(3.5) = P(X \leq 3.5)$ equals 0.60 since the event $X \leq 3.5$ corresponds to x being 1, 2, or 3.

Plotting a CDF (code here is more of an FYI)

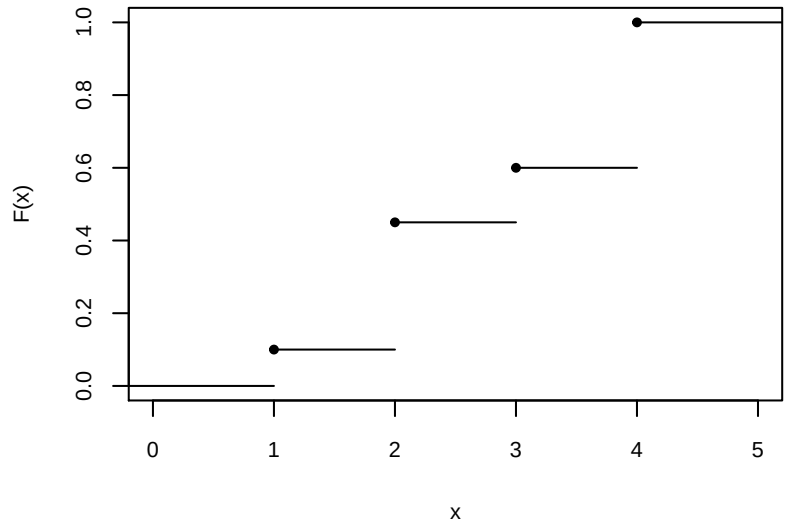
To plot a CDF, we need to write a function to evaluate its value for *all* values of x, even ones that cannot be observed. The function `cumsum` is useful here since we want to take a cumulative sum of the vector containing probabilities.

```
#Let x be a vector of possible values (smallest to largest)
#Let p be a vector of the corresponding probabilities
plot.cdf <- function(x,p) {
  plotted.values <- c(min(x)-100,x,max(x)+100)
  Fx <- c(0,cumsum(p),1)
  plot(plotted.values,Fx,xlim=c(min(x)-1,max(x)+1),pch=20,xlab="x",ylab="F(x)")
  for(i in 2:length(plotted.values)) {
    lines(c(plotted.values[i-1],plotted.values[i]),c(Fx[i-1],Fx[i-1]))
  }
}
```

Plotting a CDF example

The CDF is a line at 0 up to the smallest possible value of x. It jumps an amount equal to the probability of observing the smallest possible value of x, then continues at this level until the 2nd smallest possible value, etc., then is 1 forever beyond the largest possible value of x. The dots on the plot show that the segment includes the left endpoint but not the right.

```
#Plotting the CDF for the number of people seated at a 4-top
x <- 1:4; p <- c(.1,.35,.15,.40); plot.cdf(x,p)
```



Properties of the CDF

- $F(x) = 0$ for any x smaller than the smallest possible outcome
- $F(x) = 1$ for any x at least as large as the largest possible outcome
- In general, while writing the PMF $P(X = x)$ you only need to define what it equals at possible values of x, a written definition of $F(x)$ needs to have an expression for all values of x.
- If x is not a possible value of X, then $P(X = x) = 0$. However, it may be the case that $F(x) \neq 0$. For example, consider the 4-top PMF on the last slide

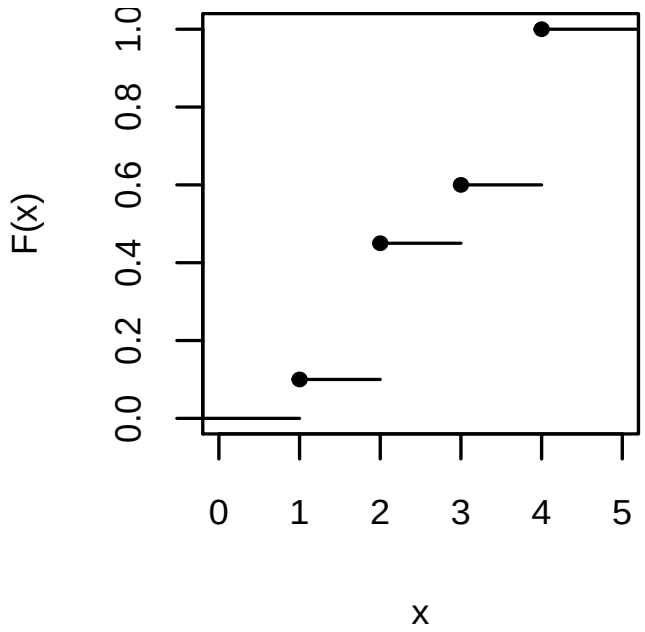
$$F(2.5) = P(X \leq 2.5) = P(X = 1) + P(X = 2) = 0.45$$

$$P(X = 2.5) = 0$$

CDF definition example (4-top) completed

$$F(x) = \begin{cases} 0 & x < 1 \\ 0.10 & 1 \leq x < 2 \\ 0.45 & 2 \leq x < 3 \\ 0.60 & 3 \leq x < 4 \\ 1 & x \geq 4 \end{cases}$$

The CDF has jumps at the possible values of X (the heights of which equal $P(X = x)$). The line includes the left endpoint but not the right since $F(x) = P(X \leq x)$.



Another CDF example

Telemarketers can be relentless. Consider X, the call number on which a customer finally makes a purchase. Let's assume that the decision to buy during each call is independent with the same probability p of buying.

The probability of having the customer's first purchase take place on the xth call equals the probability of not buying on the first $x-1$ calls followed by buying on call x.

$$P(X = x) = \underbrace{(1-p) \cdot (1-p) \cdot (1-p) \cdot \ldots \cdot (1-p)}_{\text{a total of } x-1 \text{ times}} \cdot p$$

$$P(X = x) = (1-p)^{x-1} p \qquad x = 1, 2, 3, \ldots$$

Another CDF example

$$P(X = x) = (1-p)^{x-1}p \qquad x = 1, 2, 3, \ldots$$

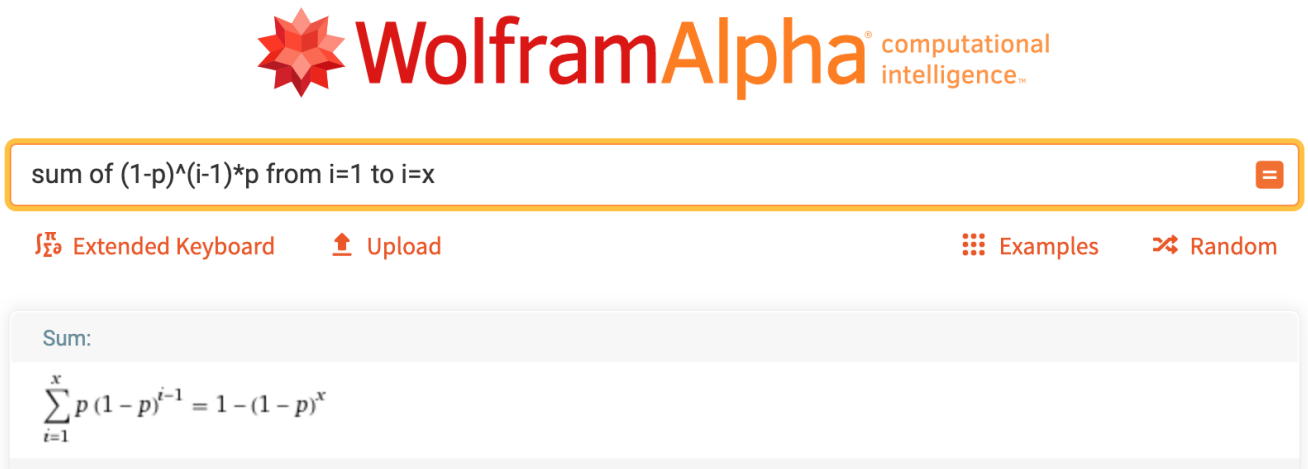
$$F(x) = P(X \leq x) = \sum_{i=1}^{x} (1-p)^{i-1}p$$

$$= (\textit{math tricks}) \begin{cases} 1 - (1-p)^{\lfloor x \rfloor} & x \geq 1 \\ 0 & \textit{otherwise} \end{cases}$$

Here $\lfloor x \rfloor$ is pronounced "floor of x" and is found by lopping off the decimal part of x, e.g., $3.14 \to 3$, $5.98 \to 5$.

Wolfram-Alpha for math tricks

For advanced math in this class (including algebra and calculus), we'll use `https://www.wolframalpha.com/`



Calculating Probabilities

To calculate the probability of observing an event, e.g. x_2, x_5, or x_6, simply sum of the values of the PMF for the relevant outcomes.

$$P(X = x_2 \text{ or } x_5 \text{ or } x_6) = P(X = x_2) + P(X = x_5) + P(X = x_6)$$

Brute forcing the sum by hand is one method, but I recommend using R using `which` and various logical conditions.

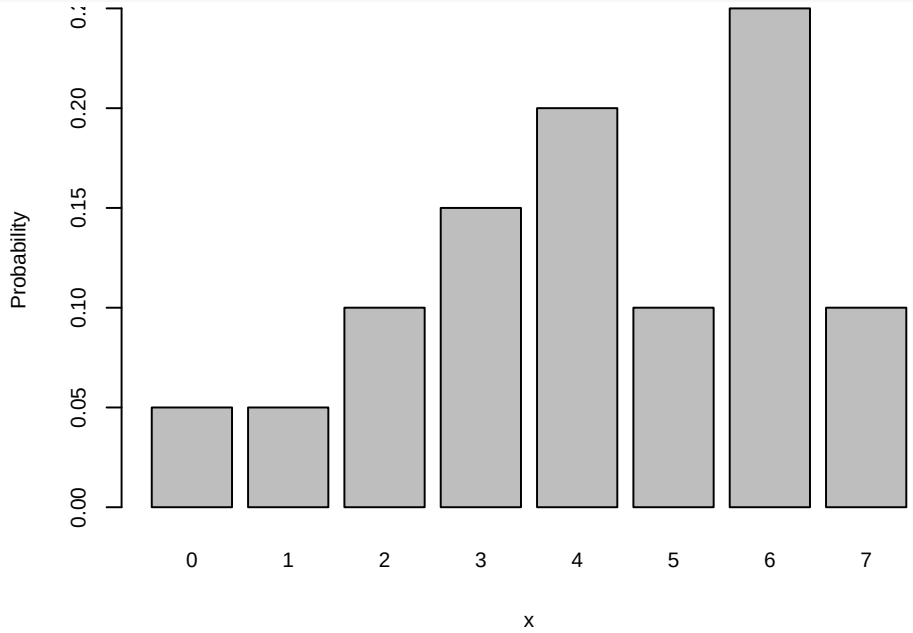
Calculating Probabilities

In R:

- Define a vector `x` that contains the possible values for X
- Define a vector `p` that contains the corresponding probabilities (always do `sum(p)` to make sure it's a valid PMF.
- Construct a command that looks like `sum( p[ which( x ...) ] )`, where you replace the ... with an appropriate logical condition.

Calculating Probabilities in R

```
x <- 0:7; p <- c(.05,.05,.1,.15,.2,.1,.25,.1)
sum(p)
## [1] 1
barplot(p,names.arg=x,xlab="x",ylab="Probability")
```



Calculating Probabilities in R

```
#P(X=3)
sum( p[which(x==3)] )  #technically sum() not needed since only 1 value
```

```
## [1] 0.15
```

```
#P(X = at most 2) = P(X <= 2) = P(X=0) + P(X=1) + P(X=2)
sum( p[which(x <= 2)] )
```

```
## [1] 0.2
```

```
#P(X = odd) = P(X=1) + P(X=3) + P(X=5) + P(X=7)
sum( p[which(x %in% seq(from=1,to=7,by=2))] )
```

```
## [1] 0.4
```

```
#P( X > 3) = P(X=4) + P(X=5) + P(X=6) + P(X=7)
sum( p[which(x > 3)] )
```

```
## [1] 0.65
```

```
#P( 3 < X <= 6 ) = P(X=4) + P(X=5) + P(X=6)
sum( p[which( x > 3 & x <= 6)])
```

```
## [1] 0.55
```

```
#P(not 4) = 1 - P(X=4) by the complement rule
1-p[which(x==4)]
```

```
## [1] 0.8
```

Another Example - solicitations

X = call on which first purchase occurs; each call has $p = 0.10$ of success.

$$P(X = x) = (1 - p)^{x-1} p \qquad x = 1, 2, \ldots$$

Note: infinite number of possible outcomes. We can't have a vector go from 1 to infinity, so what do we do?

A reasonable workaround is to take the maximum x to be a very large number. As long as `sum(p)` is reported back as 1 and `1-sum(p)` is a *very* small number, your answers should be "close enough" to be basically treated as exact.

```
#Solicitation on which customer makes first purchase, 10% success rate
x <- 1:1000; p <- 0.1*0.9^(x-1)
sum(p) #Make sure sum of considered probs is essentially 1
## [1] 1
1-sum(p) #Make sure the probability not captured by our vector is VERY small
## [1] -2.220446e-16
```

Another Example - solicitations

```
##P(5)
sum( p[which(x==5)] )
## [1] 0.06561
##P(4 or more) = 1 - P(3 or fewer) = 1 - P(1)+P(2)+P(3)
1 - sum( p[which(x %in% 1:3)] )
## [1] 0.729
sum( p[which(x>=4)] )
## [1] 0.729
##P(multiple of 5) = P(5) + P(10) + P(15) + ...
sum( p[which(x %in% seq(from=5,to=max(x),by=5))] )
## [1] 0.1602159
```

Calculating Probabilities (visual)

Visualizing the PMF as a bar chart provides a nice tie-in to how we originally defined probability:

$$P(event) = \text{fraction of sample space that contains the event}$$

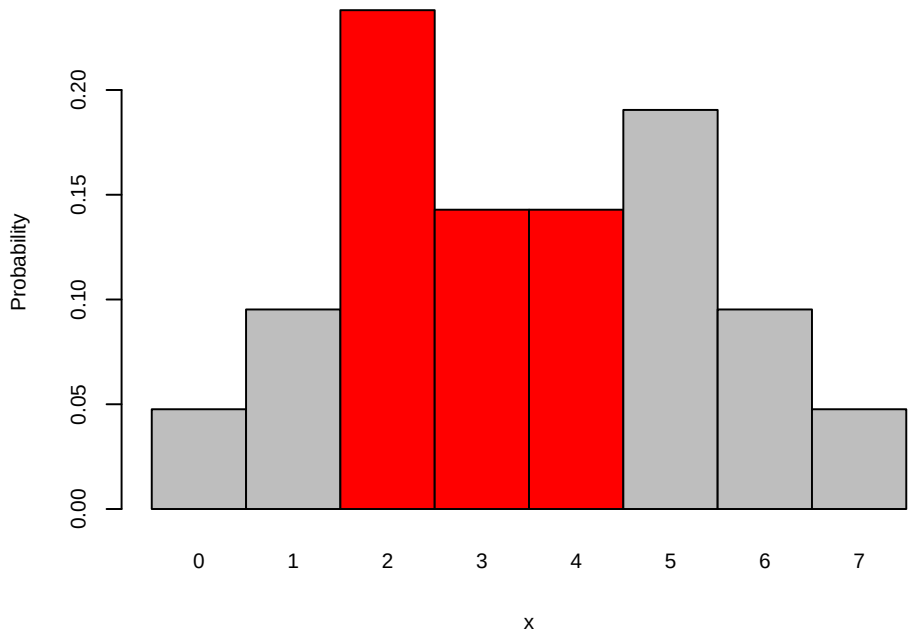
Now, the total area of all bars in the bar chart will equal 1:

- the width of each bar equals 1 and the height of the bar equals the probability of that outcome.
- the area of the bar (width times height) gives the probability of that outcome.
- the sum of the probabilities over all outcomes equals 1, so the sum of all the areas of the bars equals 1

Thus, we can treat the area traced out by the shape of the PMF as a "sample space" (examples in previous units used a simple rectangle). Just as we've done before, we can treat the probability of some event as the fraction of the sample space that contains our event.

Calculating Probabilities (visual)

The probability of some event (e.g., $P(2 \leq X \leq 4)$) equals to the total area of the bars of the outcomes composing the event.



Notorious Discrete Distributions for Counts

Section Goals

There's a relatively small set of distributions that have proven to be unusually effective at modeling random quantities in science, engineering, and business. Your goal here is:

- Based on the description of the random process, select the most appropriate notorious distribution (and a backup)
- Understand the shapes that the notorious PMFs can provide: see Shiny app on Canvas to play around!
- Be able to determine the parameters of the notorious distribution based on data (e.g. an average and standard deviation, or individual data values themselves)
- Use build-in R function to calculate probabilities of notorious distributions
- Understand that "all models are wrong . . . some are useful".

Bernoulli Distribution

The Bernoulli distribution works for any random quantity that takes on two possible values. These values are coded as 1 or 0, with a 1 typically referred to as a "success" (the outcome of interest) and "failure".

	x	P(X=x)
"Success"	1	p
"Failure"	0	1-p

$$P(X = x) = p^x(1-p)^{x-1} \qquad x = 0, 1$$

- If studying clicking on an ad, 1 = click ("success") and 0 = no click ("failure")
- If studying the chance of a stockout, 1 = stockout ("success") and 0 = no stockout ("failure")

$$\mu = E[X] = p \qquad \sigma = \sqrt{p(1-p)}$$

Binomial Distribution

The binomial distribution is a model that provides the probability of observing exactly k "successes" (some event of interest) in n independent trials.

Another way to think about it is the binomial distribution is the PMF of the sum of n independent "Bernoulli trails".

Whenever your question involves counting up the number of times something happens "out of ..." (and that "out of" number is known and fixed), the binomial is your go-to distribution.

- Probability that at least 40 out of 1000 people respond to an email blitz.
- Probability that a customer will redeem exactly 2 out of 5 coupons.
- Probability that you illegally park 30 consecutive days without getting a ticket.

Binomial Distribution

If p is the probability of "success" (your event of interest, the one you are counting up), then

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

The average value is $\mu = E[X] = np$ with standard deviation $\sigma = \sqrt{np(1-p)}$.

Binomial Notation

The quantity:

$$\binom{n}{x}$$

is pronounced "n choose x", and it's just shorthand for the following:

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Where $n!$ is "n factorial", or n times $n-1$ times $n-2$ times ... all the way down to one. For example, $5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$. Mathematicians have agreed to define 0! as equal to 1.

```
factorial(5)
## [1] 120
choose(13,5)
## [1] 1287
factorial(13)/( factorial(5)*factorial(8) )
## [1] 1287
```

Binomial Derivation I

Where does this formula come from? You could have discovered it on your own! Let's illustrate. On your way to work you pass through 5 lights. What's probability that exactly 2 of them will be green if $P(green) = 0.2$?

One possible way to observe 2 of 5 green lights is to find the sequence GGRRR. If sequential light colors are independent, then the multiplication rule says

$$P(GGRRR) = P(G) \times P(G) \times P(R) \times P(R) \times P(R)$$

$$P(GGRRR) = 0.2 \times 0.2 \times (1 - 0.2) \times (1 - 0.2) \times (1 - 0.2) = 0.2^2 \times (1 - 0.2)^3$$

However, many other sequences have exactly two green lights, e.g. GRRGR. The good news is that each of these sequences has the same probability $0.2^2 \times (1 - 0.2)^3$ of being observed. How many such sequences exist? It turns out "n choose x" do, or "5 choose 2" $= 10$ in this case.

Binomial Derivation II

Thus there are 10 sequences that have exactly 2 greens, and each sequence has a probability of $0.2^2 \times (1-0.2)^3$, so the overall probability is $10 \cdot 0.2^2 \times (1-0.2)^3 = 0.2048$.

The terms in the binomial distribution are now clear

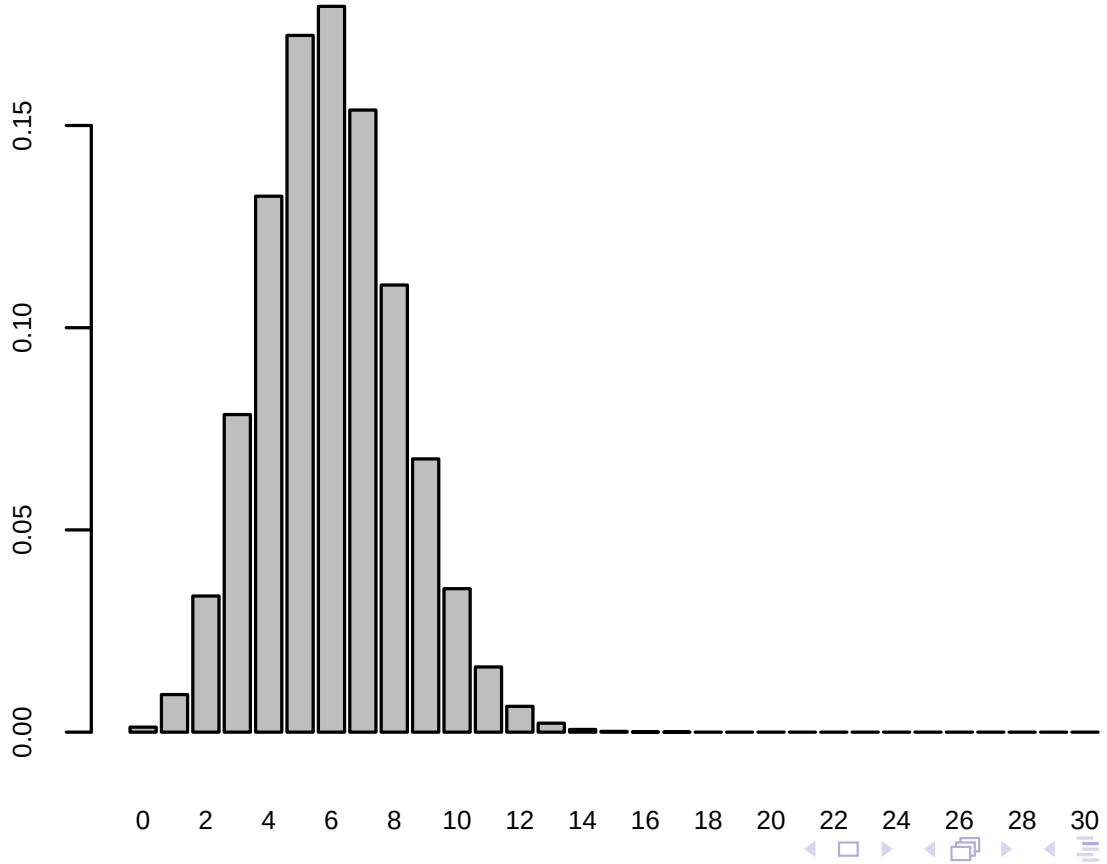
$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n$$

- $\binom{n}{x}$ is the number of possible sequences with x successes and $n - x$ failures
- $p^x(1-p)^{n-x}$ is the probability of any of these sequences

Visualizing the shape of the binomial distribution

Here is the binomial distribution when $n = 30$ and $p = 0.2$, i.e., the probability of having x successes in 30 trials when $P(success) = 0.20$.

```
x <- 0:30;  p <- dbinom(x,size=30,p=0.2)
barplot(p,names.arg=x,cex.axis=.6,cex.lab=.6,cex=.6)
```



Binomial with R (binom)

The function `dbinom(x,n,p)` gives you the probability of having exactly x successes in n trials when the probability of success is p.

- What is the probability that 3 of 48 people are left-handed, when P(left-handed) $= 0.10$?

```
dbinom(3,48,.10)
## [1] 0.1509589
```

- What is the probability that at most 3 out of 100 churn at the end of the contract, if P(churn)=0.02?

```
sum( dbinom(0:3,100,.02) ) #P(X=0) + P(X=1) + P(X=2) + P(X=3) = P(X<=3)
## [1] 0.8589616
```

The function `pbinom(x,n,p)` gives you the CDF, i.e. $P(X \leq x)$. The previous churn probability, $P(X \leq 3)$ is more efficiently calculated using the CDF.

```
pbinom(3,100,.02)  #P(X <= 3)
## [1] 0.8589616
```

Extension to binomial when there's more than 2 outcomes: Multinomial

Multinomial distribution

- Generalization of binomial when there are k outcomes instead of two (e.g. green vs. red vs. yellow).
- Gives probability that outcome 1 occurs n_1 times, outcomes 2 occurs n_2 times, ..., outcome k occurs n_k times in a total of $n = n_1 + n_2 + \ldots + n_k$ trials.
- Let `x` be a vector of frequencies of each outcome and `p` be a vector of corresponding probabilities for the outcomes

Extension to binomial when there's more than 2 outcomes: Multinomial

The probability outcome x_1 occurs n_1 times, etc., where $n = \sum n_i$ is:

$$P(X_1 = n_1, X_2 = n_2, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \cdots n_k!} p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}$$

Derivation: the factorials in front give the number of unique ways of re-sequencing the specific number of each outcome while the product the of p's gives the probability of any particular sequence.

```
#Code for the PMF of a multinomial.  CDF is very tricky
dmultinomial <- function(x,p) { factorial(sum(x))/prod(factorial(x))*prod(p^x) }
```

Multinomial example

A Halloween sized pack of M & Ms is opened and has 24 candies. The probability that a candy is green, orange, and blue is each 20%; red and yellow are 15%; brown is 10%. What is the probability of getting exactly four of each color?

```
x <- c(4,4,4,4,4,4)  #elements of x give number of times a color appeared
p <- c(.2,.2,.2,.15,.15,.1) #probability that a candy is any particular color
dmultinomial(x,p)
```

```
## [1] 0.0003408225
```

Geometric distribution (dgeom, pgeom)

The geometric distribution models the number of "failures" (not your event of interest) that occur *before* the first "success" (your event of interest) in a sequence of independent Bernoulli trials, each of which has a probability of "success" equal to p.

$$P(X = x) = (1 - p)^x p \qquad x = 0, 1, \ldots$$

Note: $x = 0$ corresponds to no failures before the first success, i.e., a success on the first trial.

The expected value is $\mu = E[X] = 1/p - 1$.

Geometric distribution (dgeom, pgeom)

The geometric distribution is our go-to when our question is about how long (in terms of trials) it takes an event to first occur:

- Each game of GEMS has a 4% chance of unlocking the bonus feature. What is the probability that it takes at least 50 games before you get to play it? $P(X \geq 50)$ (number of failures before first success is at least 50)
- A customer churns when they have their first bad experience. If the probability a customer has a bad experience on any given visit is 10%, what is the probability that the total number of visits is at most 8 before they churn? $P(X \leq 8)$.

Geometric distribution (dgeom, pgeom)

The derivation of the formula is something you could do! For x failures to occur before the first success, we must have x failures in a row (each with probability $1-p$) followed by a success (which has probability p).

$$P(X=x) = \underbrace{(1-p)\cdot(1-p)\cdot\ldots\cdot(1-p)}_{\text{a total of } x \text{ times since } x \text{ failures}} \cdot p$$

$$P(X=x) = (1-p)^x \cdot p \qquad x = 0, 1, 2, \ldots$$

Memorylessness property of geometric distribution

Allegedly, 20% of M & Ms are orange. You consider getting an orange M & M out of a bag a "success". What is the probability that you will get an orange at some point within the first 5 candies you take?

You pull 10 candies out of the bag, but *none* are orange! Bad luck! But you must be "due" for one now right? Now (given you've opened 10 non-oranges in a row), what is the probability that you will get an orange candy at some point within the next 5 candies?

How does this probability (of getting an orange candy in the *next* 5 candies) change if the first 10 candies weren't orange? First 50?

Memorylessness property of geometric distribution

When a random process is well-modeled by a geometric distribution, the answer to the question "what is the probability that the next success comes within the next t trials" is the *same*, regardless of the historical sequence of successes and failures.

With the M & Ms, it doesn't matter if we just opened the bag, if we drew 10 non-oranges in a row, or if we drew 50 non-oranges in a row, the probability that we'll get an orange in the next 5 candies is always about $1 - 0.8^5 = 2/3$!

In other words, the geometric distribution is *memoryless*. For all intents and purposes, the process "resets itself" from a probabilistic standpoint after each trial. Regardless of the past sequence of successes or failures, the next trial will always have a probability p of success and $1 - p$ of failure.

Memorylessness property of geometric distribution

First, let's calculate the general probability of getting (at least one) orange candy in the next 5 trials. This is the complement of the event "no orange candies in next 5 trails".

Since a non-orange candy occurs with probability $1 - p$, the probability of 5 non-orange candies in a row is $(1 - p)^5$.

By the complement rule, the probability of getting an orange candy in the next 5 trails is:

$$1 - (1 - p)^5 = 1 - 0.8^5 = 0.67232$$

Memorylessness property of geometric distribution

To establish memorylessness, remember our rule for conditional probability:

$$P(A|B) = P(A \text{ and } B)/P(B)$$

$$P(\text{succeed in next 5 trials}|\text{first } n \text{ trials were failures})$$

$$= \frac{P(\text{succeed in next 5 trials } \textit{and} \text{ first } n \text{ trials were failures})}{P(\text{first } n \text{ trials were failures})}$$

Since each trial is independent, we can write out the probability of n failures followed by at least 1 success in the next 5 trials ("at least one success" is the complement of "all failures") as the product of these two events.

$$= \frac{(1-(1-p)^5) \times (1-p)^n}{(1-p)^n}$$

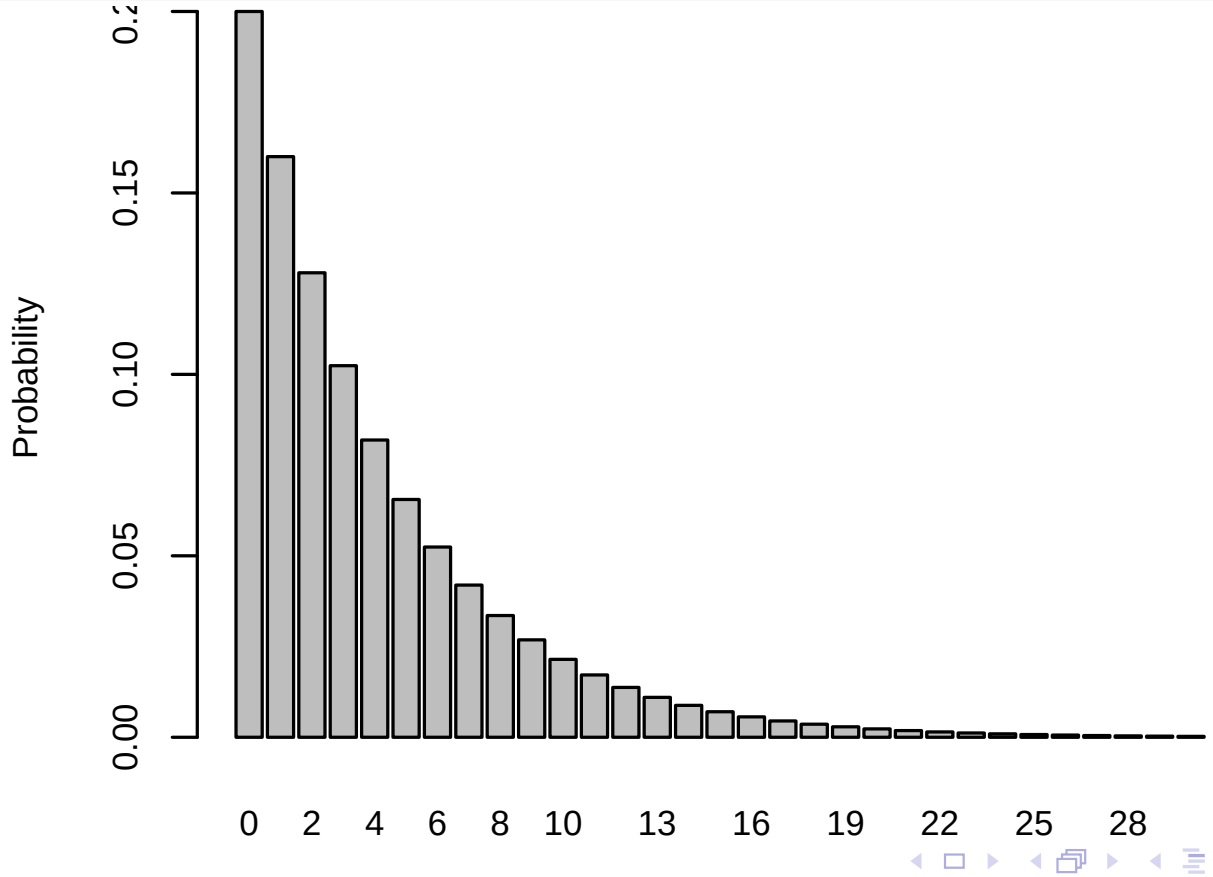
$$= 1-(1-p)^5$$

$$= P(\text{succeed in next 5 trials with no information about previous trials})$$

Visualizing geometric distribution

Here is the PMF of the geometric when the probability of success is 20%. All geometric distribution peak at $x = 0$!

```
x <- 0:30; p<- dgeom(x,p=0.2)
barplot(p,names.arg=x,xlab="x",ylab="Probability")
```



Negative Binomial (dnbinom,pnbinom)

The negative binomial distribution is the "generalization" of the geometric distribution and gives the *number of failures* (not number of trials) which occur *before* a target number of successes r.

$$P(X = x) = \binom{x + r - 1}{x} p^r (1 - p)^x \qquad x = 0, 1, \ldots$$

Derivation: to have x failures before the r-th success, the sequence must have some ordering of x failures and $r - 1$ successes, followed by a success number r on the final trial.

Each of these sequences has probability $\left((1 - p)^x \times p^{r-1}\right) \times p = (1 - p)^x p^r$.

There are a total "$x + r - 1$ choose x" possible sequences with x failures and $r - 1$ successes.

Negative Binomial (dnbinom,pnbinom)

Whenever we ask a question about the number of trials it takes for a certain number of "successes" to occur, our go-to is the negative binomial distribution.

Normally, you'll have to recast the question so that it refers to the number of failures before the target number of successes (notoriously tricky).

If your question is about the trial (N) on which the r-th success occurs, the x you need to use for the negative binomial distribution will equal $N - r$. For the r-th success to occur on trial N, $x = N - r$ total failures (and $r - 1$ successes) must occur in the first $N - 1$ trials.

For example, for the 4th success (r) to occur on trial 9 (N), the number of failures (x) before the 4th success needs to equal 5 ($N - r$). One possible sequence is SFFFSSFFS.

Negative Binomial (dnbinom,pnbinom)

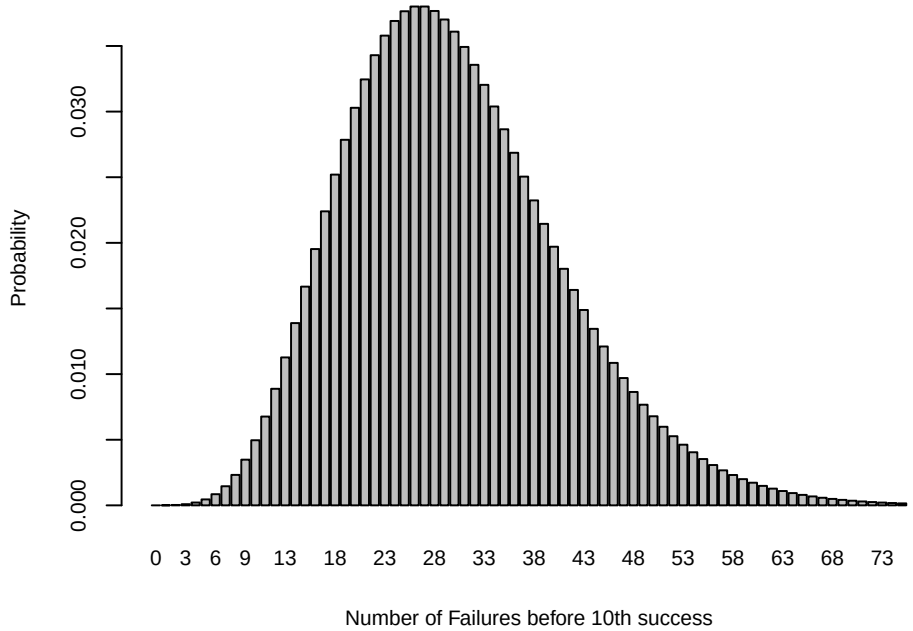
Example questions answered with the negative binomial:

- You need to hand out 15 flyers; the probability that someone accepts a flyer is 10%. What is the probability you have to ask at least 165 people in order to give away all your flyers? To get rid of all 15 flyers on person 165 or later, the number of "failures" has to be at least 150 (since an additional 15 have to take the flyers to make 165). $P(X \geq 150)$.
- A customer churns when they have their 3rd bad experience. If the probability a customer has a bad experience on any given visit is 10%, what is the probability that the total number of visits is at most 8 before they churn? To have 8 or fewer visits, at most 5 "failures" (good experiences) have to occur before the 3rd "success" (bad experience), e.g. GGGGGBBB. $P(X \leq 5)$.

Visualizing Negative Binomial

A telemarketer has a quota of 10 sales before he or she can take a lunch break. If each call has a 25% chance of success, what is the distribution of the number of *failures* that will occur before a lunch break. Here's the shape of the negative binomial PMF for $p = 0.25$ and $r = 10$.

```
x <- 0:75; p <- dnbinom(x,p=0.25,size=10) #p = probability of success; size=target number of successes
barplot(p,names.arg=x,xlab="Number of Failures before 10th success",ylab="Probability")
```



Note: the total number of *calls* made will be 10 plus the numbers in this distribution, since this probability describes the number of failures before r successes.

Properties of the Negative Binomial

The average (or expected value) is:

$$\mu = \frac{r(1-p)}{p}$$

The standard deviation is:

$$\sigma = \frac{\sqrt{r(1-p)}}{p}$$

Poisson Distribution (dpois,ppois)

The Poisson distribution is the go-to model for "generic" counts, i.e., "counts without a context" (no upper bound, not "out of . . .", not "before something happens").

- The number of books someone has checked out from the library
- The number of customers that visit BestBuy between 1-2pm
- The number of items someone has in their shopping cart at Kroger
- The number of people in line (being served and waiting) at checkout

Poisson Distribution (dpois,ppois)

The shape of a Poisson distribution is determined by its average value λ.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, 2, \ldots$$

The average value is $\mu = \lambda$ and the standard deviation is $\sigma = \sqrt{\lambda}$.

Poisson example using R

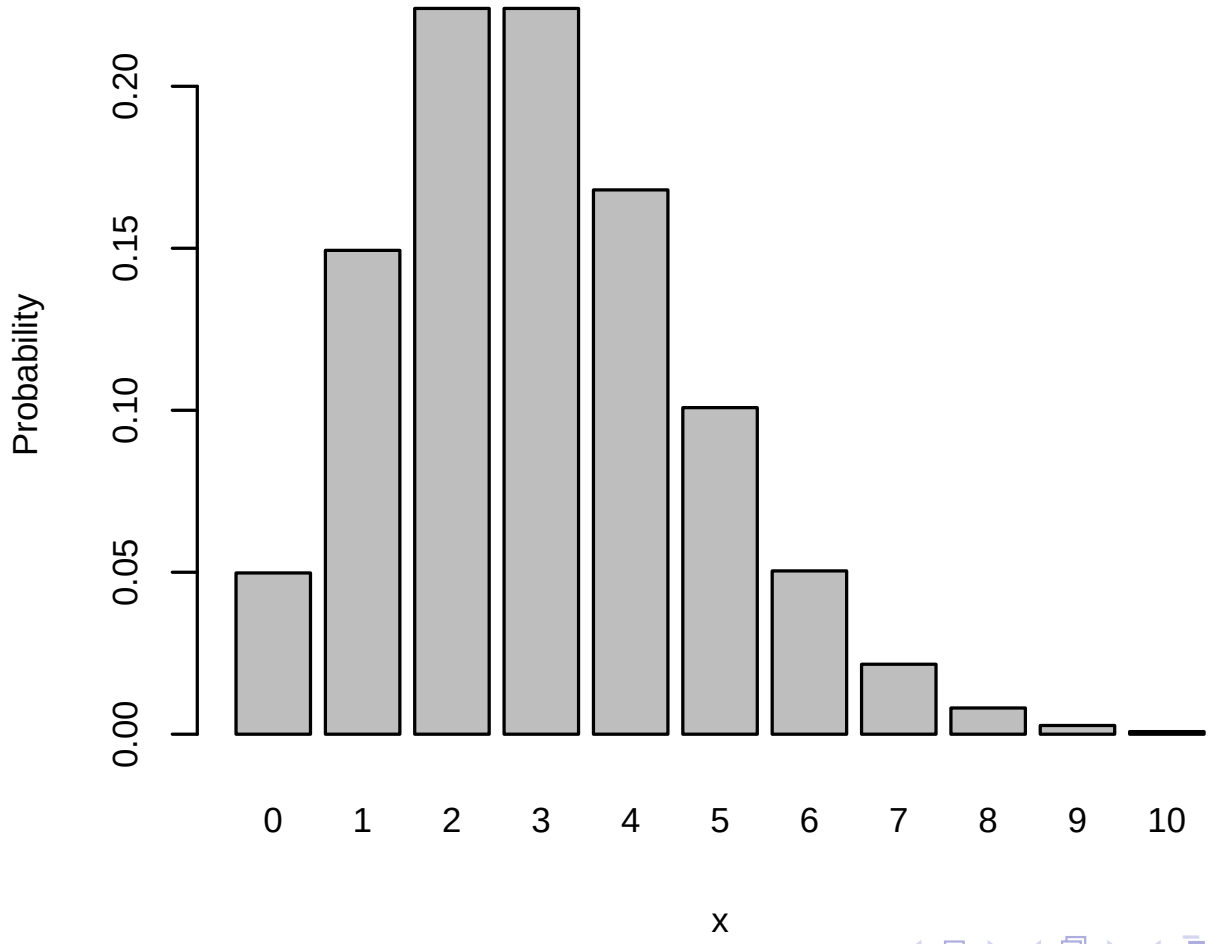
The number of items in a cart averages 4.2. What is the probability that a customer has exactly 3? 7 or more? Between 2 and 6?

```
4.2^3*exp(-4.2)/factorial(3)  #P(X=3) by hand
## [1] 0.1851654
dpois(3,4.2) #dpois gives PMF and P(X=3) directly
## [1] 0.1851654
1-ppois(6,4.2) #ppois gives CDF, "7 or more" is the complement of 6 or fewer.
## [1] 0.132536
1-sum(dpois(0:6,4.2)) #Another way to get P(X>=7)
## [1] 0.132536
1-sum( 4.2^(0:6)*exp(-4.2)/factorial(0:6) )  #P(X>=7) 'by hand'
## [1] 0.132536
sum( dpois(2:6,4.2) ) #Between 2 and 6
## [1] 0.789487
```

Visualizing the Poisson Distribution

Here is a visualization for the Poisson distribution when $\lambda = 3$.

```
x <- 0:10; p <- dpois(x,lambda=3)
barplot(p,names.arg=x,xlab="x",ylab="Probability")
```



Zero-truncated Poisson

While the Poisson distribution is the go-to for modeling "counts without a context", if we wanted to model the number of items in a shopping cart *at checkout* with a Poisson we run into difficulty since carts at checkout have *at least* one time.

When 0 is not a possible value, but otherwise we're studying otherwise generic counts, we can use a **zero-truncated Poisson** (ZTP).

You can derive the formula for its PMF! If X is Poisson, then a ZTP is the conditional distribution of X given that $X > 0$.

Zero-truncated Poisson

Let X be the number of items in a cart.

$$P(X = x|X > 0) = \frac{P(X = x \text{ and } X > 0)}{P(X > 0)} = \frac{P(X = x)}{1 - P(X = 0)}$$

$$P(X = x|X > 0) = \frac{\lambda^x e^{-\lambda}/x!}{1 - e^{-\lambda}} \qquad x = 1, 2, 3, \ldots$$

$$P(X = x|X > 0) = \frac{\texttt{dpois(x,lambda)}}{\texttt{1 - exp(-lambda)}} \qquad x = 1, 2, 3, \ldots$$

Since our model allows only $x > 0$, the expression $P(X = x \text{ and } X > 0)$ becomes $P(X = x)$ since $X > 0$ is redundant.

Zero-truncated Poisson Example

Let X (zero-truncated Poisson) be the number of items in a cart at checkout and let $\lambda = 8$. What is the probability that the number of items equals 7? Is 5 or fewer? Is 12 or more?

```
#P(X=7)
dpois(7,8)/(1-exp(-8))
## [1] 0.1396334
#P(X<=5); careful not to include 0
sum( dpois(1:5,8)/(1-exp(-8)) )
## [1] 0.1909647
#P(X>=12); use complement rule 1-P(X<=11) and careful not to include 0
1 - sum( dpois(1:11,8)/(1-exp(-8)) )
## [1] 0.1119616
```

Negative Binomial as an Alternative to Poisson

The Poisson distribution is the go-to model for "counts without a context" (i.e., no upper bound, not "out of", not "before something happens"). Often, in real-world data, the Poisson doesn't *quite* capture the observed variability in the collected values.

If data values are a bit more spread out than what a Poisson can produce, the go-to alternative is the negative binomial distribution (choosing p and r so that $r(1-p)/p = \lambda$). Even though the negative binomial is designed to count up the number of failures before a target number of successes occur, it's remarkably useful for modeling "counts without a context" as well.

Aside - there's actually a deep and meaningful connection between the Poisson and Negative Binomial that we will flesh out mathematically later, so it's actually no accident that the Negative Binomial often provides a more realistic alternative than the Poisson.

Uniform

Discrete Uniform Distribution:

- Features: all integers in a range $[a, b]$ are equally likely, i.e., the result is picked "at random".
- Use: typically when no other information available or logic dictates each value in a range has equal probability
- The PMF is simply $P(X = x) = 1/(b - a + 1)$ at each possible outcome (i.e., 1 over the number of possible outcomes). The CDF (at possible values) is $F(x) = \frac{x-a}{b-a}$

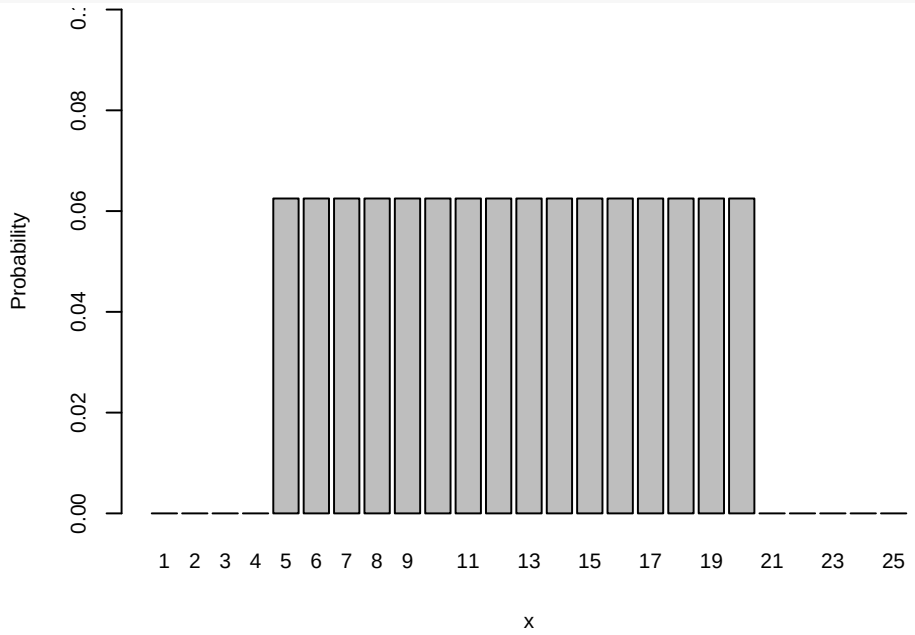
Example quantities suitable to be modeled with a uniform distribution:

- The number of minutes displayed by a clock if you pick a moment at random to check the time (uniform 0-59)
- The month of a randomly picked person's birthday (1-12)

Uniform

Shape when values are uniformly distributed between 5-20.

```
x <- 1:25; p <- c(rep(0,4),rep(1/16,16),rep(0,5)) #Uniform between 5 and 20
barplot(p,names.arg=x,xlab="x",ylab="Probability",ylim=c(0,0.10))
```



Using Notorious Distributions in R

Most notorious distribution have an abbreviation in R. Once you know it, you can preface the abbreviation with a d, p, q, or r to study various properties of the distribution.

Example: R knows the binomial distribution as the abbreviation `binom`. If we wanted to study properties of the distribution when $n = 20$ and $p = 0.5$:

- `dbinom(x,size=20,prob=0.5)` - The value of the PMF at the value x
- `pbinom(x,size=20,prob=0.5)` - The CDF, i.e., the probability of obtaining a value `x` or less, $P(value \leq x)$
- `qbinom(q,size=20,prob=0.5)` - Quantiles. q must be between 0 and 1. Returns the percentiles of the distribution, i.e., returns the value of x such that $P(value \leq x) = q$
- `rbinom(n,size=20,prob=0.5)` - Gives n random numbers from the distribution.

Distributions in R

- Binomial - `binom`
- Poisson - `pois`
- Negative Binomial - `nbinom`
- Geometric - `geom`

There are no built-in functions for the zero-truncated Poisson, Bernoulli, multinomial, or uniform PMFs in R's default installation, though you can find them in other libraries.

Expected Value and Standard Deviation

Overview

The PMF of a discrete random variable fully describes the possible values it may take along with the respective probabilities.

What if we wanted to summarize the distribution with a single number, i.e. a "typical" value?

What if we wanted to summarize how spread out numbers from that distribution might be?

For example, if we have the PMF for the demand for an item, how do we fill in the blanks?

"The typical demand is about _____, give or take _____ or so."

Expected value and standard deviation

The **expected value** of the random variable X, abbreviated μ, is the *average* or *mean* value of X and often provides a good summary of a typical value you may see from the distribution.

The **standard deviation** of X, abbreviated σ, provides a measure of how spread out values from the distribution may be. Specifically, it gives the "typical difference" between the possible values and the overall average μ.

Instead of giving the standard deviation of distributions, most resources give the **variance**, abbreviated σ^2. It's easy to get the standard deviation though:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

$$\sigma = \sqrt{\sigma^2}$$

Summaries of Notorious Distributions

The means, variances, and standard deviations of the notorious distributions are (usually) complicated functions of the parameters of the distributions. When you need these, google it or use Wikipedia:

`http://en.wikipedia.org/wiki/Binomial_distribution`
`http://en.wikipedia.org/wiki/Poisson_distribution`
`http://en.wikipedia.org/wiki/Uniform_distribution_(discrete)`
`http://en.wikipedia.org/wiki/Negative_binomial_distribution`

Note: be very careful with the negative binomial. There are two very different ways to write it.

List of "all" distributions:
`http://en.wikipedia.org/wiki/List_of_probability_distributions`

Expected Value (Average)

The expected value of a random variable X, abbreviated μ ("mu") or $E[X]$, is just a fancy word for the average or mean value of X.

$$\mu = E[X] = \sum_{\text{all possible } x} x \cdot P(X = x) \qquad \text{Discrete}$$

The average value of a random variable is a weighted sum of its possible values, with weights equal to the probability that each value would occur.

Example of finding the expected value

Let X to be the number of calls a telemarketer makes during the course of a day. Take the PMF to be:

x	$P(X = x)$
20	0.14
21	0.18
22	0.21
23	0.30
24	0.17

Example (discrete)

x	20	21	22	23	24
$P(X = x)$	0.14	0.18	0.21	0.30	0.17

$$\mu = E[X] = \sum_{all\ x} x \cdot P(X = x)$$

$$\mu = E[X] = 20 \times 0.14 + 21 \times 0.18 + 22 \times 0.21 + 23 \times 0.30 + 24 \times 0.17 = 22.18$$

The typical (or average) number of calls the telemarketer makes is 22.18.

Calculating the average of a PMF in R

Finding the expected value of a random variable in R is straightforward once you know its PMF. Let `x` be a vector of possible outcomes and `p` be a vector with their probabilities. Then `sum(x*p)` gives the expected value.

```
x <- 20:24
p <- c(.14,.18,.21,.30,.17)
sum(p) #Always check to see it sums to one so you know you didn't typo!
## [1] 1
sum(x*p)
## [1] 22.18
```

Average of PMF vs. average of data

Note that the calculation of the average for a PMF proceeds *very* differently than the calculation of the average for a set of data.

- If `x` is a vector of data values, then `mean(x)` tells you the average of those values.
- If `x` is a vector containing the possible values of a random variable, and `p` contains their respective probabilities, then `sum(x*p)` is what you need.

```
x <- 0:40; p <- dbinom(x,size=40,prob=0.1) #Random variable X has binomial distribution with n=40 and p=0.
sum(x*p) #Average of PMF
## [1] 4
mean(x)  #NOT the average of the PMF (just the numerical average of the possible values of X)
## [1] 20
x <- rbinom(500,size=40,prob=0.1)  #500 random numbers from a binomial distribution
mean(x)  #Average of the data values contained in x
## [1] 4.054
```

Expected values of functions of a random variable

The expected value of $f(X)$ (some function of X) is

$$E[f(X)] = \sum_{all\ x} f(x) \cdot P(X = x) \qquad \text{Discrete}$$

In English, evaluate $f(x)$ for every possible value of x that might be observed. Then take a weighted sum of these values, with weights equal to the probability of observing the corresponding value of x.

Note: $E[f(X)] \neq f(E[X])$ in general, so you can't just find the expected value and calculate the relevant function.

Expected values of function example

Imagine the telemarketer is paid by the call.

$$Pay = Calls \log(Calls + 1)$$

The following gives the PMF of *Calls* (the expected value is 22.18). What is the expected value of *Pay*, i.e. of the function $Calls \log(Calls + 1)$?

x (calls)	20	21	22	23	24
$P(X = x)$	0.14	0.18	0.21	0.30	0.17

$$E[f(X)] = \sum_{all\ x} f(x) \cdot P(X = x)$$

$$E[X \log(X + 1)] = \sum_{all\ x} x \log(x + 1) \cdot P(X = x)$$

$$E[X \log(X + 1)] = 20 \log 21 \times 0.14 + 21 \log 22 \times 0.18 + (22 \log 23) \times 0.21$$

$$+(23 \log 24) \times 0.30 + (24 \log 25) \times 0.17 = 69.75637$$

Calculating expected values of function in R

The expected pay is 69.76 This is a little different than the value that emerges if you just put the average number of calls of the number of calls (22.18) into the equation for pay ($22.18 \cdot \log(1 + 22.18) = 69.72$).

```
x <- 20:24
p <- c(.14,.18,.21,.30,.17)
sum(p) #Always check to see it sums to one so you know you didn't typo!
## [1] 1
mu <- sum(x*p)  #Expected value of X
mu
## [1] 22.18
sum(x*log(x+1)*p) #Expected value of f(X)
## [1] 69.75637
mu*log(mu+1)  #E[f(x)] is not f(E[X])!
## [1] 69.71817
```

Calculating expected values of function in R

Although the difference between the two wasn't large in this case, it can be for a different PMF!

```
x <- 0:4
p <- c(.8,.14,.03,.02,.01)
sum(x*log(x+1)*p)
```

```
## [1] 0.3105125
```

```
mu <- sum(x*p)
mu*log(mu+1)
```

```
## [1] 0.07870928
```

Note how the expected value of the function $X\log(X+1)$ is almost **four times larger** than the function evaluated at the expected value of X. These differences can be large and meaningful (which ties back to how using the averages in the Jewelry catalog example is *not* good enough).

Special note on expected value of a function

This point cannot be understated.

If what you want is the average value of some function of X, you cannot in general just calculate the average value of X and plug that into your function.

$$E\left[f(X)\right] \neq f\left(E[X]\right)$$

This had profound consequences for our motivating customer lifetime value example. Plugging in the averages for each random quantity in the formula will most likely give us the *wrong* number for the actual average lifetime value!

The only times you are allowed to do is when $f(x) = a + bx$. It does work out that the expected value of $a + bX$ is $a + b\mu$.

Further properties of expectation

Let a and b be constants and X and Y be two random variables.

- $E[a] = a$ (the expected value of number is itself)
- $E[aX + b] = aE[X] + b$
- $E[aX + bY] = aE[X] + bE[Y]$ (expectation is a "linear operator")

The above identities are still valid if X is replaced by some function $f(X)$, e.g., $E[af(X) + bg(Y)] = aE[f(X)] + bE[g(X)]$

Variance and Standard Deviation

The **variance** of a random variable X is denoted by σ^2 ("sigma squared") and is equal to the expected value of the function $(X-\mu)^2$, with μ being the average value of X. The standard deviation σ is the square root of the variance.

$$\text{Variance} = Var(X) = \sigma^2 = E[(X-\mu)^2]$$

$$\text{Standard Deviation} = \sigma = \sqrt{\sigma^2}$$

The variance of the random variable has no immediate interpretation (it's just easier to work with mathematically).

> The standard deviation tells you the typical difference between possible values of X and the expected value (average) of X.

Summarizing a PMF with its expected value and standard deviation

Back in intro stats, you summarized a collection of data values with its mean and standard deviation.

> If `x` is a data vector, then a good way to summarize the collection of values is "the observed values are typically about `mean(x)` give or take `sd(x)` or so".

Let's us a similar summarization for a PMF.

> "A typical value we might observe from the random variable X is μ, give or take σ or so." Or, if you like: "Values from the PMF are typically about μ give or take σ or so".

Example calculation and interpretation

The number of email queries a company receives in its "Contact Us" link seems to be always at least 100 and is capped at 200 for a day. Larger numbers seem to be increasingly more likely, so let's model it by

$$P(X = x) = \frac{x^2}{2358350} \qquad x = 100, 101, \ldots, 200$$

$$\mu = E[X] = \sum x \cdot P(X = x) = \sum_{x=100}^{200} x \cdot \frac{x^2}{2358350} = 160.9208$$

$$\texttt{Variance} = \sigma^2 = \sum_{x=100}^{200} (x - \mu)^2 \cdot P(X = x) = \sum_{x=100}^{200} (x - 160.9208)^2 \cdot \frac{x^2}{2358350} = 755.4832$$

$$\texttt{Standard Deviation} = \sigma = \sqrt{\sigma^2} = 27.48$$

Interpretation: the number of email queries is typically about 161 give or take 27 or so.

Example calculation of standard deviation in R

The calculation proceeds nicely when using R. Define `x` and `p`, then `sum(x*p)` is the expected value and `sqrt( sum( (x-sum(x*p))^2*p ) )` is the standard deviation.

$$P(X = x) = \frac{x^2}{2358350} \qquad x = 100, 101, \ldots, 200$$

```
x <- 100:200; p<- x^2/2358350;
mu <- sum(x*p); mu
## [1] 160.9208
V <- sum( (x-mu)^2*p ); V
## [1] 755.4832
sigma <- sqrt(V); sigma
## [1] 27.48606
sqrt( sum( (x-sum(x*p))^2*p ) )  #getting sigma with fewer steps
## [1] 27.48606
```

Alternative formula for variance

There is an alternative formulation for the variance that is sometimes easier for working with PMFs mathematically, but if we do everything in R the previous definition more than suffices.

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

```
x <- 1:100; p <- abs( 50-x )/sum(abs( 50-x ) )
sum(p)
## [1] 1
mu <- sum( x*p ); mu
## [1] 51
V <- sum( x^2*p) - mu^2  #E[X^2] - mu^2; alternative formulation
sqrt(V)
## [1] 35.34827
sqrt( sum( (x-mu)^2 * p ) )  #same answer as above!
## [1] 35.34827
```

Alternative formula for variance

Where does that come from? You can find it with a little algebra and using the properties of the expected value.

$$\begin{aligned} Var(X) = \sigma^2 = E[(X-\mu)^2] = & \quad E[X^2 - 2\mu X + \mu^2] \\ = & \quad E[X^2] - 2\mu E[X] + E[\mu^2] \\ = & \quad E[X^2] - 2\mu \cdot \mu + \mu^2 \\ Var(X) = \sigma^2 = & \quad E[X^2] - \mu^2 \end{aligned}$$

Properties of the Variance

Let a and b be constants and X and Y be random variables.

- $Var(a) = 0$ (the variance of a number is 0)
- $Var(X + b) = Var(X)$ (adding a constant to a random variable does not change how spread out the values are around the average)
- $Var(aX) = a^2 Var(X)$ but $SD(aX) = aSD(X)$
- $Var(X + Y) = Var(X) + Var(Y)$ only if X and Y are independent
- $SD(X + Y) = \sqrt{Var(X) + Var(Y)}$ only if X and Y are independent
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + abCov(X, Y)$ in general, where $Cov(X, Y)$ is the covariance (to be discussed at a later time)
- $Var(f(X)) = E[(f(X))^2] - (E[f(X)])^2$ (the variance of a function of X can't be broken down into a simple function involving the variance of X)

Examples

The high temperature in July has an average of 87 Fahrenheit with a standard deviation of 5. What is its expected value, variance, and standard deviation in Celsius? $C = 5F/9 - 160/9$.

Answer: $E[a + bX] = a + bE[X]$ so average in Celsius is $5 \cdot 87/9 - 160/9 = 30.6$. The variance in Fahrenheit is 25, so the variance in Celsius is $(5/9)^2 \cdot 25 = 7.7$. The standard deviation is $\sqrt{7.7}$, or since $SD(a + bX) = bSD(X)$, it is $25/9$.

Examples

The number of purchases at Walmart and Target per person per year averages 12 and 8, respectively, with standard deviations of 10 and 6. What are the mean and standard deviation of the total amount of purchases across both stores?

Answer: The average is 20, since $E[X+Y] = E[X] + E[Y]$. The standard deviation cannot be computed. We have that $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$. $Cov(X,Y)$ is a measure how how correlated the two quantities are. Since this information is not given, and we cannot assume they are uncorrelated, we cannot answer.

Checking a PMF for reasonableness

Overview

Often, a reasonable model can be assumed for a process based on general features of the distribution (symmetric vs skewed, unimodal vs. not) or the context (counting up successes in a fixed number of trials). Business analytics is about data-driven decision making, so it is imperative to check to see whether the model provides a reasonable reflection of reality when data is available.

At a later time, we will discuss estimating parameters (p, λ, r, etc.) of the models from the data itself, but there is a standard process for checking to see whether a given model is consistent with the data.

Overview

We will focus on the graphical procedure for checking. While there are formal statistical tests (you may run into them in another class), they are often too strict because "All models are wrong (some are useful)", and in business analytics we just need models to be "good enough".

Empirical Cumulative Distribution Function (ECDF)

If X is our random variable, then $F(x) = P(X \leq x)$ is the cumulative distribution function CDF. The *empirical* cumulative distribution function ECDF is defined based on observed data

$$ECDF(x) = F_n(x) = \text{fraction of observations who values are} \leq \text{x}$$

The subscript n is to make it clear that this is the empirical distribution as observed from a sample of size n.

Checking whether a probability model makes sense involves comparing the theoretical CDF to the observed ECDF. If the model is reasonable, the two should closely resemble each other.

ECDF example

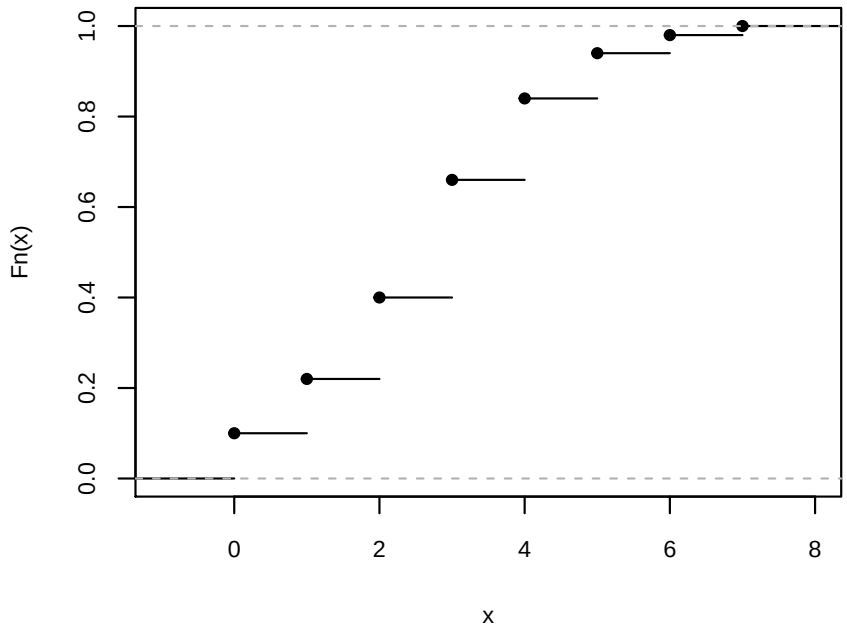
Consider a sample consisting of 50 random numbers from a Poisson distribution with a mean of 3. The ECDF can tell us what fraction of observations are at most 1, at most 2, etc.

```
set.seed(533)
x <- rpois(50,3)
sort(x)
##  [1] 0 0 0 0 0 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 4 4
## [36] 4 4 4 4 4 4 5 5 5 5 5 6 6 7
mean(x <= 0)  #Fraction of values at most 0
## [1] 0.1
mean(x <= 1)  #Fraction of values at most 1
## [1] 0.22
mean(x <= 2)  #Fraction of values at most 2
## [1] 0.4
```

ECDF example

Vertical axis gives the fraction of data values that are less than or equal to x. It "jumps" at integer values of x because possible values here are integers (the fraction less than or equal to 2.5 is the same as the fraction less than or equal to 2, etc.).

```
plot(ecdf(x),main="")
```



Is the model reasonable?

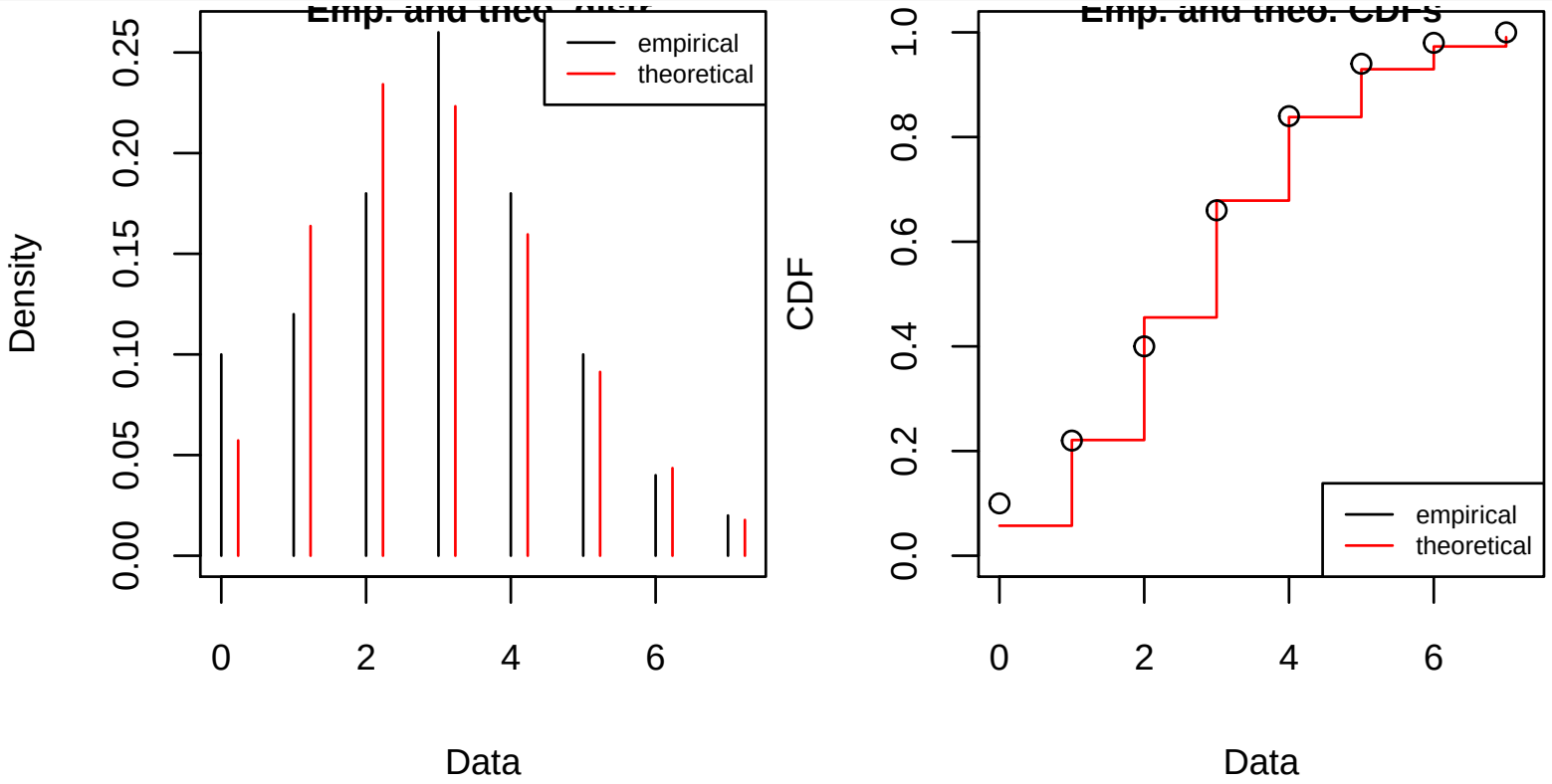
Let's compare the PMF of the proposed model to the observed frequencies of each value and the CDF of the proposed model to the ECDF. We will use the `fitdistrplus` package.

```
library(fitdistrplus)  #Must install this package ahead of time
FIT <- fitdist(x,"pois") #Fit the data contained in x to a Poisson distribution
```

Does the model fit?

So what do we look for?

```
plot(FIT)
```



Does the model fit?

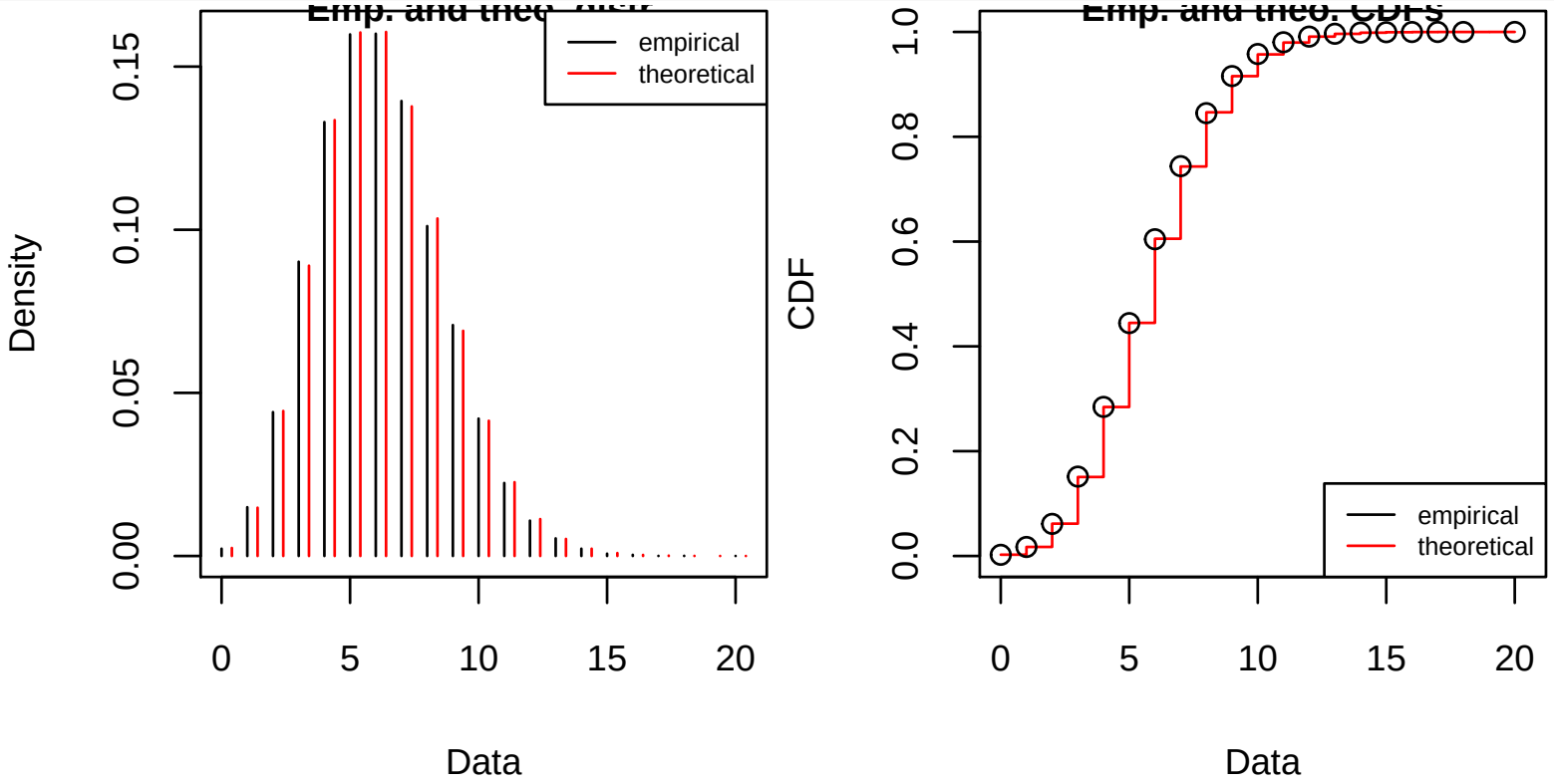
The plot on the left compares probability of observing each value (according to the theoretical PMF) to the observed frequency of those values. The black and red bars should match up "pretty well" for the distribution to be reasonable. However, there is a fair amount of leeway, and sometimes it's hard to do this visual comparison. The other plot is better.

The plot on the right compares the theoretical CDF to the ECDF. The black circles should fall "close" to the outside corners of the red steps to be reasonable.

Good vs. bad fits

Good fit:

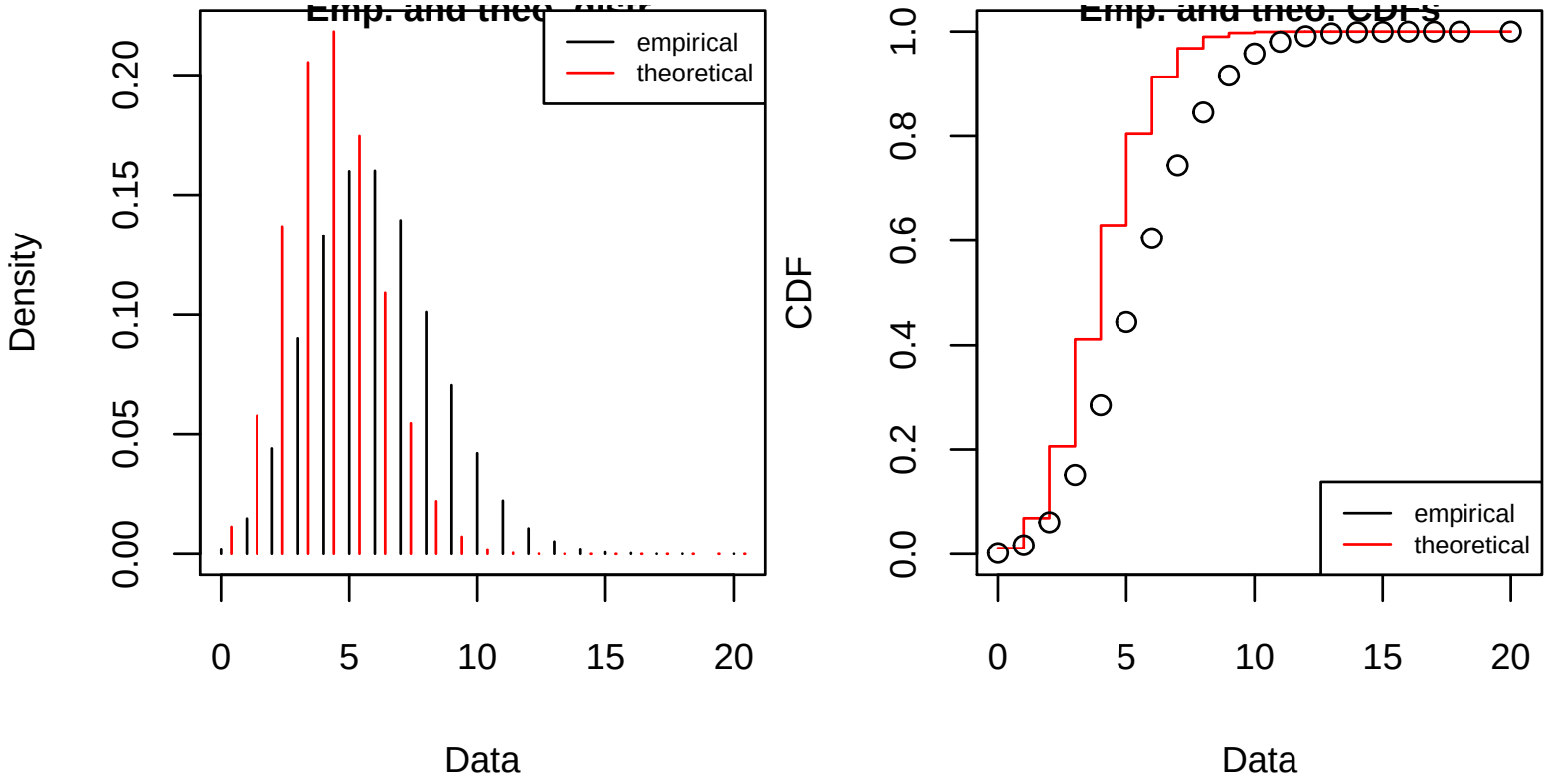
```
set.seed(533); x <- rpois(50000,6)
FIT <- fitdist(x,"pois"); plot(FIT) #good fit since this *is* the distribution that made th
```



Good vs. bad fits

Bad fit:

```
plotdist(x,"binom",para=list(size=20,prob=0.2)) #Plot a fit to a particular binomial
```



Using fitdist

To use fitdist, you pretty much only need to know R's abbreviation for the PMF. Save the result of running fitdist to something, then plot() it! If you get an error (warning messages are common and ok), it's likely that the distribution you're trying to use isn't suitable for that kind of data.

```
x <- sample( 40:150, size=100, replace=TRUE) #just so that x contains data
#Note:  R has no built-in function for a discrete uniform distribution
FIT <- fitdist(x,"pois") #Fit to Poisson
FIT <- fitdist(x,"nbinom") #Fit to Negative Binomial
FIT <- fitdist(x,"geom") #Fit to Geometric
#Fit to binomial - note special syntax.  Change size=150 to size=your number of trials
FIT <- fitdist(x,"binom",start=list(prob=0.5),fix.arg=list(size=150))
plot(FIT) #This will produce the plots after fitting
```

Using a custom distribution

You can check your own custom PMFs as well, but it takes a bit of programming. You need to come up with an abbreviation for it, then define a function with the same name (though prefaced with a `d`) to give the PMF, then define a function with the same name (though prefaced with a `p`) to give the CDF, then finally define an empty function with the same name (though prefaced with a `q`).

Example: zero-truncated Poisson (abbreviated as `ztpois`). The first argument will be `x`, following by arguments giving each parameter that the distribution depends on (just `lambda` for ZTP).

We need to write the `dztpois` and `pztpois` functions.

Using a custom distribution

The d function will require transcribing the formula giving the probability of each possible value of x. The p function will resemble what you see here. The q function has to be defined, but can be blank. The name of the first arguments *must* be x, q, and p for the d, p, and q functions, respectively.

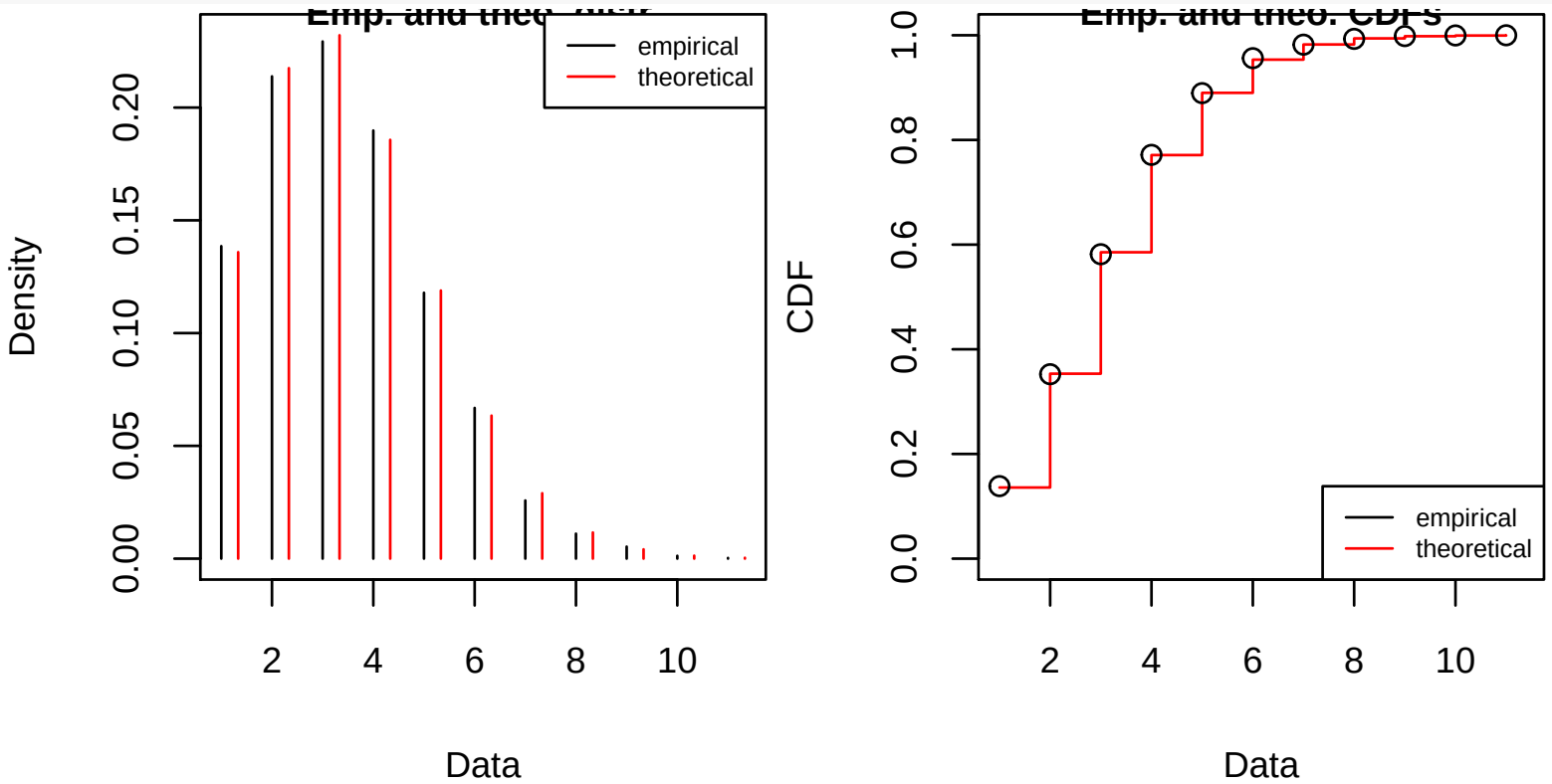
```
dztpois <- function(x,lambda) { ifelse(x<=0,0,dpois(x,lambda)/(1-exp(-lambda))) }
pztpois <- function(q,lambda) {
  ans <- rep(0,length(q))
  for (i in 1:length(q)) { ans[i] <- sum(dztpois(0:q[i],lambda)) }
  ans
}
qztpois <- function(p,lambda) { }
x <- rpois(1e4,3.2); x <- x[x>0]  #Make some data
table(x)
```

```
## x
##    1    2    3    4    5    6    7    8    9   10   11
## 1327 2048 2196 1818 1129  640  247  106   51   12    3
```

Using a custom distribution

When running `fitdist`, you need to add `discrete=TRUE` for a discrete distribution and also give it a starting "guess" for what the parameter of the distribution might be.

```
FIT <- fitdist(x,"ztpois",discrete=TRUE,start=list(lambda=3.2))
plot(FIT)
```



Using a custom distribution

The authors of `fitdistrplus` decided to output a warning to `fitdist` when custom distributions are used unless they are written in an excruciatingly specific and unnecessary (for us) way.

```
#The pztpois function should return a vector of with NA values when input has missing values and not remove
#The pztpois function should return a vector of with NaN values when input has inconsistent values and not
```

Ignore them, it's working just fine.

Discrete Uniform set of functions

Because you might be interested in fitting a uniform distribution to data, here are the functions needed to do so. The nickname is going to be "unifdiscrete" and it will take arguments `lower` and `upper`

```
dunifdiscrete <- function(x,lower,upper) {
  ifelse( x %in% lower:upper, 1/length(lower:upper),0)
}
punifdiscrete <- function(q,lower,upper) {
  ifelse( q %in% lower:upper, (q-lower+1)/(upper-lower+1), ifelse(q<lower,0,1))
}
qunifdiscrete <- function(p,lower,upper) {  }
```

Discrete Uniform fit

Checking a uniform fit requires running `plotdist` instead of `plot()` on a `fitdist` object.

```
x <- sample( 2:14, size=500, replace=TRUE ) #uniform between 2-14
plotdist(x,"unifdiscrete",para=list(lower=min(x),upper=max(x)),discrete=TRUE)
```