STAT 450/460

Handout 1: Discrete Random Variables

Fall 2016

Contents

Mathematical statistics: the gyst

The primary idea behind statistics is to understand a **population** using **data from a sample**, often denoted using y_1, y_2, ..., y_n. These data arise from some **data generating mechanism** (DGM) which is often governed by parameters such as a mean μ or standard deviation σ some other parameter. Stat 450/460 is concerned with the study of DGMs, so that we can understand sampling properties of **statistics** (such as $\bar{y}$ or s^2 or $y_{(n)}$). Understanding the sampling distributions of statistics are imperative for being able to carry out the following critical components of statistical inference:

1. Estimation (e.g. properly calibrated confidence intervals)
2. Hypothesis testing

To begin, we will start off investigating different kinds of random variables.

- Examples:

Chapter 3 - Discrete random variables

Definition: A random variable $Y(\cdot)$ or just simply Y is a function from sample space S into the real numbers. A random variable Y is a ***Discrete random variable*** if it can assume only a finite or countably infinite number of distinct values.

We will often use uppercase such as Y to denote a *random variable* and a lowercase letter such as y to denote a specific value of Y. With this in mind, it makes sense to talk about for example $P(Y = y)$. Think of this as "the probability that the random variable Y takes on the specific value y."

An example is to let Y denote the number of cars that pass through a fast food drive through window on any single day. Y could take on any value between 0, 1, ... We might be interested in $P(Y = 20)$, which would read (not very succinctly) as "The probability that the random variable denoting the number of cars passing through the drive through window takes on the value of 20".

Example 1: Flip a coin 3 times. Let Y denote the number of heads out of 3 flips.

What is the sample space, S?

We might be interested in talking about $P(Y = y)$. Here, what are the values that Y could take on?

Example 2: Roll two die, one white one red.

What is the sample space, S?

Some possible different random variables:

- $Y_1 \equiv$ sum of red/white total. What values could Y_1 take on?

- $Y_2 \equiv$ max of two rolls. What values could Y_2 take on?

- $Y_3 \equiv$ min of two rolls. What values could Y_3 take on?

- $Y_4 \equiv$ absolute difference between two rolls. What values could Y_4 take on?

- $Y_5 \equiv$ ratio of two rolls, taking largest number over smallest number. What values could Y_5 take on? Is Y_5 a discrete random variable?

Probability mass functions (pmf's)

Definition: Suppose Y is a discrete valued random variable that takes on possible values y_1, y_2, y_3,... (countable/possibly infinite, number of values). Then the pmf is defined as:

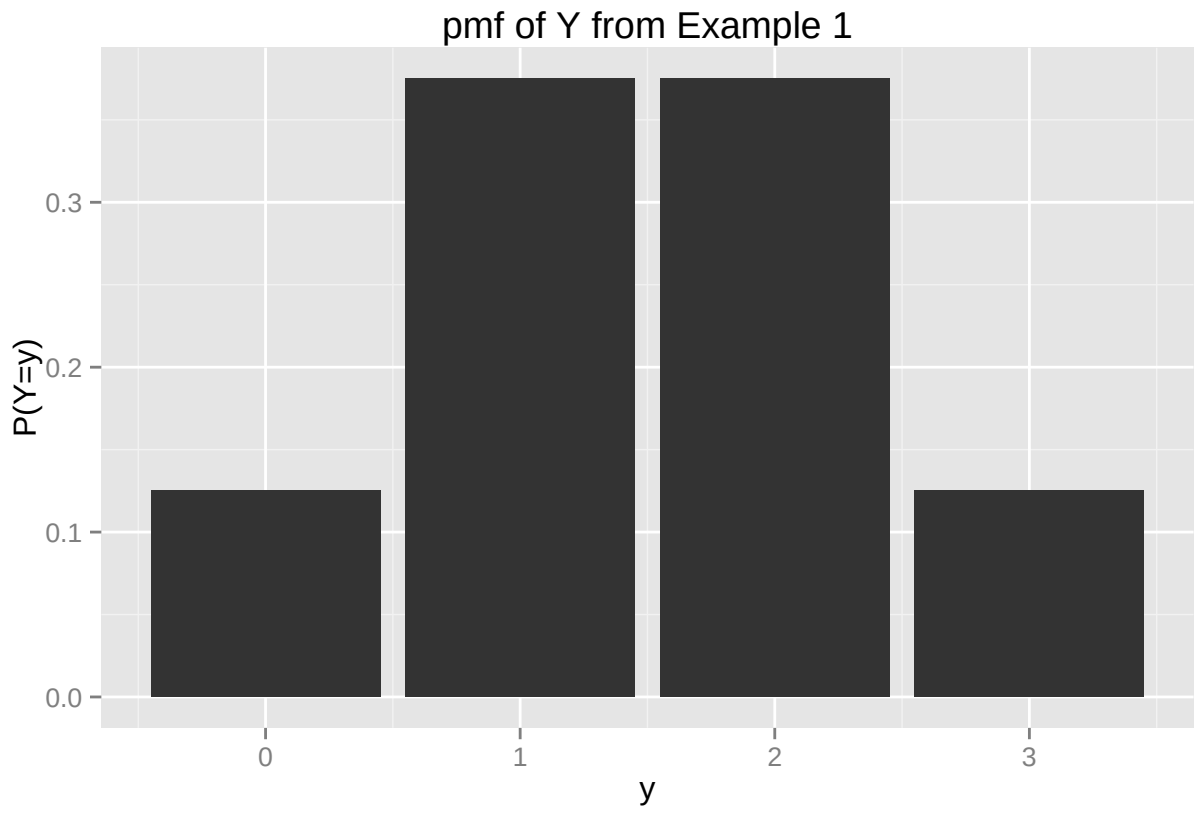
1. $p(y_i) = P(Y = y_i)$
2. $0 \leq p(y_i) \leq 1$ for all i
3. $\sum_{all\ i} p(y_i) = 1$

Often if there are a finite number of y_i, the pmf can be defined in table format.

Tasks: Define the pmf's for Y from Example 1, and Y_1 and Y_4 from Example 2.

It is often helpful to visualize pmfs using histograms. Let's visualize the pmf of Y from Example 1 in R:

```
library(ggplot2)
y <- 0:3
py <- c(1/8,3/8,3/8,1/8)
dd <- data.frame(y=y,probs = py)
ggplot(aes(x=y,y=probs),data=dd) + geom_bar(stat='identity') +
  ylab('P(Y=y)') + ggtitle('pmf of Y from Example 1')
```




```
#Breakdown of above code:
#   ggplot() specifies the basics.
#     What variable is on the x axis and which is on the y axis?  Where to look for data?
# The rest of the commands add LAYERS to the plot.
#  --geom_bar() says to use bars to visualize the data.
# stat='identity' argument needed to allow us to specify the height
#of each bar; see ?geom_bar
```

Expectation of a discrete random variable and functions thereof

It is often very important to understand tendencies of a random variable. One measure of tendency is the **expectation**, or the **mean** of a random variable.

Definition: Let Y be a discrete random variable with probability function $p(y)$. Then the *expected value* of Y, $E(Y)$, is defined to be:

$$\mu = E(Y) = \sum_y yp(y)$$

Similarly, let $g(Y)$ be a function of a discrete random variable. Then the expectation of $g(Y)$ is:

$$E(g(Y)) = \sum_y g(y)p(y)$$

- Some common $g(y)$ include:

Properties of expectation

1. Let $c \in \mathcal{R}$ (c is some constant), then $E(c) = c$.
   Proof:

2. Let $c \in \mathcal{R}$ (c is some constant), then $E(cg(Y)) = cE(g(Y))$.
   Proof: HW

3. (Linearity of expectation) Let $a, b \in \mathcal{R}$, then $E(ag(Y) + b) = aE(g(Y)) + b$. Note that this implies $E(aY + b) = aE(Y) + b$. (VERY IMPORTANT/USEFUL FACT)
   Proof: HW

Variance of a random variable

From the above, we can now define another important characteristic of a random variable: its variance.

Definition: If Y is a random variable with mean $E(Y) = \mu$, then the variance is defined to be the expected value of $(Y - \mu)^2$, i.e.:

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

Properties of variance

1. $Var(Y) = E(Y^2) - E(Y)^2 = E(Y^2) - \mu^2$ (VERY IMPORTANT/USEFUL FACT)
   Proof:

2. $Var(aY + b) = a^2 Var(Y)$
   Proof: HW

Examples

Reconsider Example 1. Find $E(Y)$ and $Var(Y)$.

Let's calculate these quantities in R, and demonstrate them via simulation:

```
y <- 0:3
py <- c(1/8,3/8,3/8,1/8)
mu <- sum(y*py)
mu
```

```
## [1] 1.5
```

```
EY2 <- sum(y^2*py)
EY2
```

```
## [1] 3
```

```
sigma2 <- EY2-mu^2
sigma2
```

```
## [1] 0.75
```

```
#Now let's simulate lots of sets of 3 flips.
#First, we'll write a function that returns
#the realization of one Y:
one.Y <- function() {
  sample.space <- c('H','T')
  three.flips <- sample(sample.space,3,replace=TRUE)
  num.heads <- sum(three.flips=='H')
  return(num.heads)
}
#Test it out:
one.Y()
```

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

```
set.seed(2222)
many.Y <- replicate(1000,one.Y())
mean(many.Y)
```

```
## [1] 1.488
```

```
mean(many.Y^2)
```

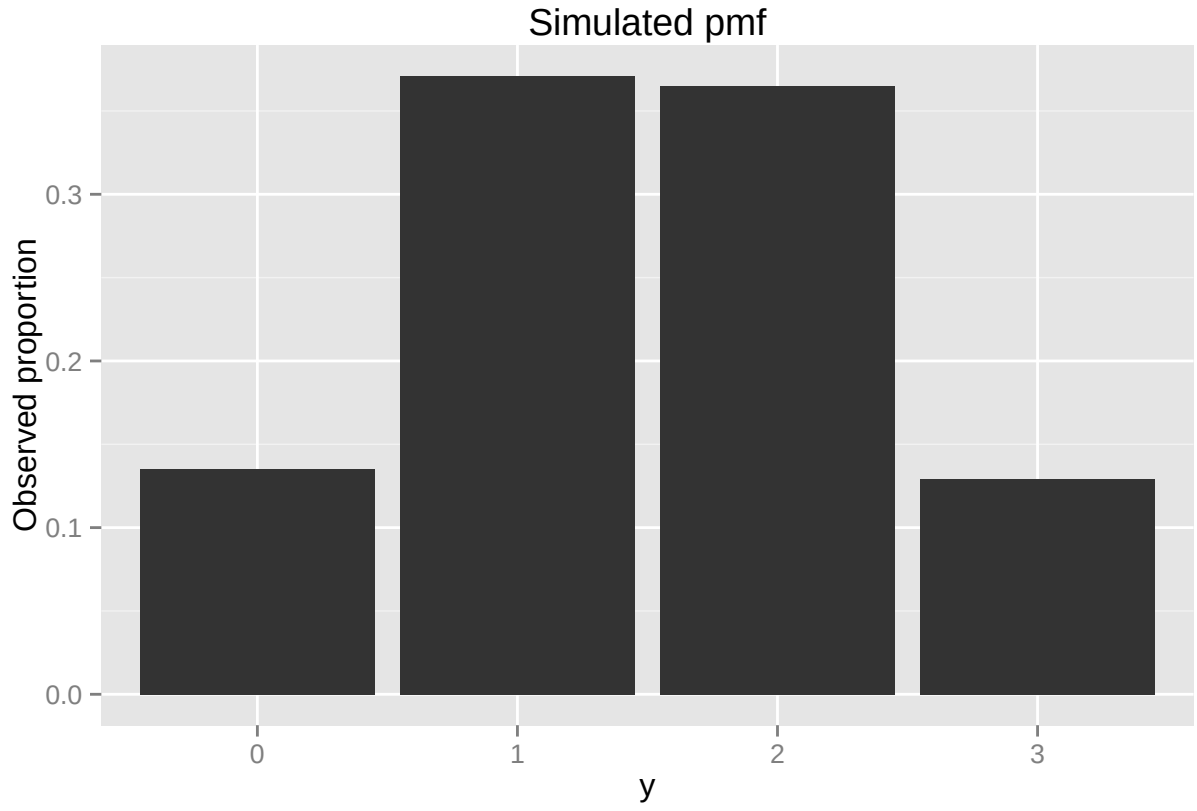
```
## [1] 2.992
```

```
var(many.Y)
```

```
## [1] 0.7786346
```

```r
#Hmmm...these look familiar!

#put results into data frame:
df <- data.frame(x=as.factor(many.Y))
ggplot(aes(x=as.factor(many.Y)),data=df) + geom_bar(aes(y=(..count..)/(sum(..count..)))) +
  ylab('Observed proportion') + xlab('y') + ggtitle('Simulated pmf')
```




```r
#Hmmm...this looks familiar too!
```

Moment generating functions

Although the mean μ and variance σ^2 are important characteristics of a random variable, they do not fully characterize its distribution. May different random variables might have the same mean and variances but completely different distributions.

The moment generating function (MGF), however, completely and uniquely characterizes a random variable's distribution. It can also be used to obtain moments, which in turn can be used to obtain μ and σ^2:

Definition: The k^{th} *moment of a random variable taken about the origin* is defined to be $E(Y^k)$.
Definition: The k^{th} *moment of a random variable taken about its mean*, or the k^{th} *central moment*, is defined to be $E((Y-\mu)^k)$.

Examples:

- The mean μ is the 1^{st} moment centered around the origin: $\mu = E(Y^1)$
- The variance σ^2 is the 2^{nd} moment centered around the mean: $\sigma = E[(Y-\mu)^2]$

Definition: The *moment-generating function* (MGF) $M_Y(t)$ for a random variable Y is defined to be $M_Y(t) = E(e^{tY})$. We say that a moment generating function fo Y exists provided the expectation exists for $t \in (-h, h)$.

MGFs are motivated by Maclaurin series, which it bears to review.

Aside: Taylor and Maclaurin Series (special case of Taylor series with $a = 0$)

Taylor Series: Given a differentiable function $f(x)$, we can approximate $f(x)$ near a point a as follows:

$$f(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + ... + \frac{1}{k!}f^{(k)}(a)(x-a)^k + ...$$

The *Maclaurin series* is just a Taylor series with $a = 0$, i.e. it approximates $f(x)$ near 0:

$$f(x) = f(0) + f'(0)(x) + \frac{1}{2}f''(0)(x)^2 + ... + \frac{1}{k!}f^{(k)}(0)(x)^k + ...$$

Example: Find the Maclaurin series for $f(x) = e^x$.

IMPORTANT CONSEQUENCE: $e^{anything} = \sum_{k=0}^{\infty} \frac{anything^k}{k!}$

So what does this have to do with the MGF? Well, why is it *called* a "MGF"?

Theorem: Let $M_Y(t)$ be the MGF for a random variable. Then $E(Y^k) = M_Y^{(k)}(0) = \frac{d^k}{dt^k} M_Y(t)|_{t=0}$.

Proof: Start by using the Maclaurin series with $f(t) = e^{tY}$.

Given these fundamental aspects of random variables such as mean and variance, moments, pmfs, and moment generating functions, we will now begin studying these for specific, well-known distributions of discrete random variables.

Bernoulli distribution

$$Y \sim BERN(p)$$

or

$$Y \sim BIN(1, p)$$

A random variable Y has a Bernoulli distribution if it arises from an experiement where there are two possible outcomes ("success" or "failure"). We denote $p = P(\text{“}success\text{”})$ and $q = 1 - p = P(\text{“}Failure\text{”})$.

Example of a Bernoulli random variable: Flip a coin once. Let $Y = 1$ if the flip lands HEADS, and $Y = 0$ if the flip lands TAILS. Then $Y \sim BERN(p)$, with $p = 1/2$.

The random variable Y takes on the following values:

$$Y = \begin{cases} 1 & with\ probability\ p \\ 0 & with\ probability\ q \\ any\ other\ value & with\ probability\ 0 \end{cases}$$

This implies the Bernoulli pmf is:

$$p(y) = P(Y = y) = \begin{cases} p^y(1-p)^{(1-y)} & for\ y \in \{0, 1\} \\ 0 & otherwise \end{cases}$$

- Verify $p(1)$ and $p(0)$

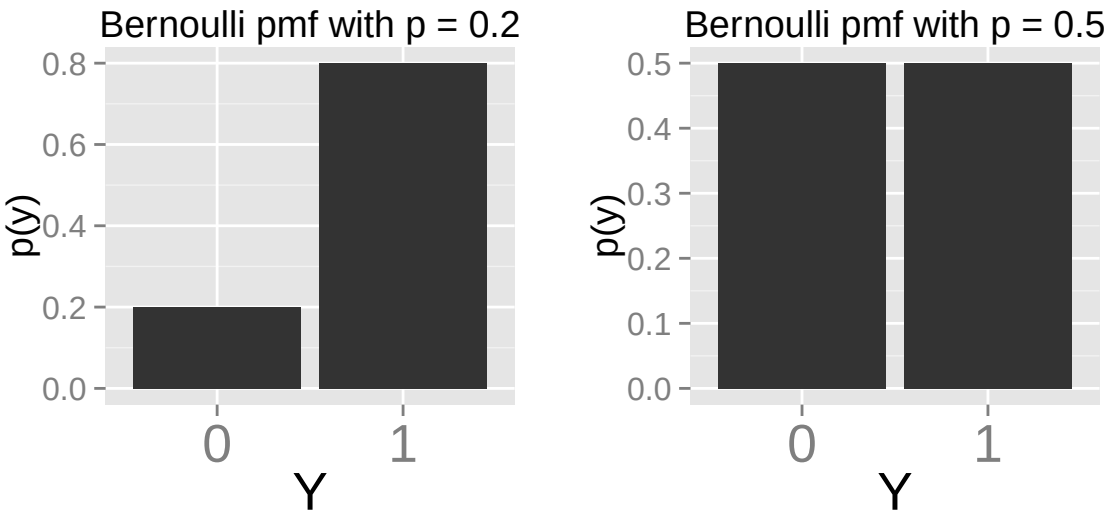
We can create a function to graph the pmf for any parameter p:

```
library(ggplot2)
bern.pmf <- function(p) {
yvalues <- as.factor(c(0,1))
probabilities <- c(p,1-p)
mydata <- data.frame(y=yvalues,prob=probabilities)
mytitle <- paste('Bernoulli pmf with p =',p)
ggplot(aes(x=y,y=prob),data=mydata) + geom_bar(stat='identity') + ylab('p(y)') + xlab('Y') +
  theme(axis.text.x = element_text(size=20), axis.text.y = element_text(size=12),
        axis.title.x = element_text(size=20),axis.title.y = element_text(size=14))  +
        ggtitle(mytitle)
}
```

Notice the `theme()` function above. This controls all aspects of the plot that are *not* mapped to data, for example legend and axis text. See `?theme()`.

Let's try the function out:

```
bern.pmf(0.2)
bern.pmf(0.5)
```



Let's derive the following important quantities when $Y \sim BERN(p)$:

1. $E(Y)$
2. $Var(Y)$
3. $M_Y(t)$

1. $E(Y) = p$.
   Proof::

2. $Var(Y) = p(1-p) = pq$.
   Proof::

3. MGF:

Binomial distribution

The binomial distribution describes the distribution of Y when Y is the number of "successes" in n independent Bernoulli runs. Specifically, the following conditions must be satisfied:

1. Independent trials
2. 2 outcomes per trial ("success" "failure")
3. $p(success) \equiv p$ constant
4. n constant

If a random variable Y occurs under these conditions, then $Y \sim BIN(n, p)$. Note that the Bernoulli distribution is a special case of the Binomial, with $n = 1$; hence $Y \sim BIN(1, p)$ is equivalent to $Y \sim BERN(p)$.

PMF:

$$P(Y = y) = p(y) = \begin{cases} \binom{n}{y} p^y (1-p)^{(n-y)} & 0 \leq y \leq n \\ 0 & otherwise \end{cases}$$

Example: LeBron James is a lifetime .744 free-throw shooter.

- If LeBron takes $n = 5$ free throws, how would we define the random variable Y?
- Does Y have a binomial distribution? If so, $Y \sim BIN(?, ?)$
- What is $P(Y = 4)$?
- What is $P(Y = 5)$?
- What is $P(Y \leq 3)$?

The first part of the binomial pmf, $\binom{n}{y}$, is the number of ways there are to arrange y successes among n trials.

- In the previous example, how many ways are there to arrange 4 successes among 5 total? Give the possible arrangements below:

- What is the probability of any one of these arrangements?
- How does $P(Y = 4)$ follow?

Does the PMF sum to 1?

Recall that to be a valid pmf, $\sum_{all\ y} p(y)$ must equal 1. Does that hold for the binomial pmf? More specifically, does:

$$\sum_{y=0}^{n} \binom{n}{y} p^y q^{n-y} = 1?$$

Question: Why must we only sum from $y = 0$ to n?

To show that the binomial pmf is indeed a valid pmf, we will appeal the Binomial Theorem:

$$\begin{aligned}(a+b)^n &= a^n + a^{(n-1)b}\binom{n}{2}a^{n-2}b^2 + ... + \binom{n}{2}a^2b^{n-2} + ab^{n-1} + b^n \\ &= \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i}\end{aligned}$$

- Use the Binomial Theorem to prove that $\sum_{y=0}^{n} \binom{n}{y} p^y q^{n-y} = 1$.

We now turn to the mean, variance, and MGF.

1. If $Y \sim BIN(n, p)$, then $E(Y) = np$.
   Proof:

2. If $Y \sim BIN(n, p)$, then $Var(Y) = npq$.
   Proof:

3. $M_Y(t) = (pe^t + q)^n$
   Proof and usage:

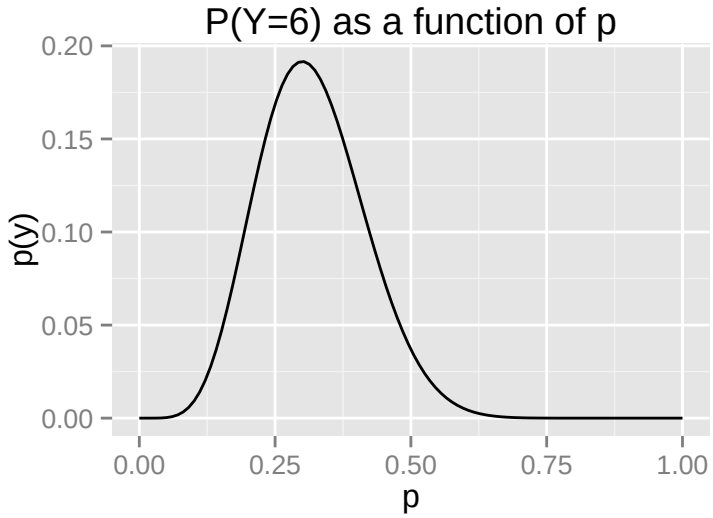
Example 3.10 in book: estimation

Suppose management of a large company is interested in the percent of employees who favor implementation of a new retirement policy. So they survey 20 people and ask their opinion, 6 of which favor the new policy. What information does this sample provide about p, the **true** percent of employees in favor?

- What is the random variable Y here?
- $Y \sim BIN(?, ?)$
- Give an expression for $P(Y = 6)$:

Note we can graph this probability as a function of p:

```
p <- seq(0,1,l=100)
p.y <-choose(20,6)*p^6*(1-p)^14
mydata <- data.frame(p=p,py = p.y)
ggplot(aes(x=p,y=p.y),data=mydata) + geom_line() + ylab('p(y)') + xlab('p') +
  ggtitle('P(Y=6) as a function of p')
```



Where does this graph have its maximum? We can find this easily via calculus, differentiating $P(Y = 6)$ with respect to p. It turns out that it's much easier to differentiate $\ln(P(Y = 6))$. Since $\ln(\cdot)$ is a monotone function, finding the p that maximizes $\ln(P(Y = 6))$ is the same p that maximizes $P(Y = 6)$.

- Differentiate $\ln(P(Y = 6))$ with respect to p, and find the p that maximizes this function (and hence maximizes $P(Y = 6)$).
- What is the value of p that maximizes $P(Y = y)$ for general n and y? Why does this value of p make sense?

This is called a *maximum likelihood estimate* of p, because it is the value that *maximizes* the *likelihood* of seeing the observed data. Much more on maximum likelihood to come!

Binomial distribution in R

The functions `dbinom()`, `pbinom()`, and `rbinom()` are respectively used to find exact binomial probabilities, cumulative binomial probabilities, and to generate binomially-distributed random variables.

Reconsider the free throw example, where $Y \sim BIN(5, .744)$:

```
y <- 0:5
dbinom(y,size=5,prob=0.744)
```

```
## [1] 0.001099512 0.015977278 0.092867930 0.269897423 0.392194692 0.227963165
```

```
pbinom(y,size=5,prob=0.744)
```

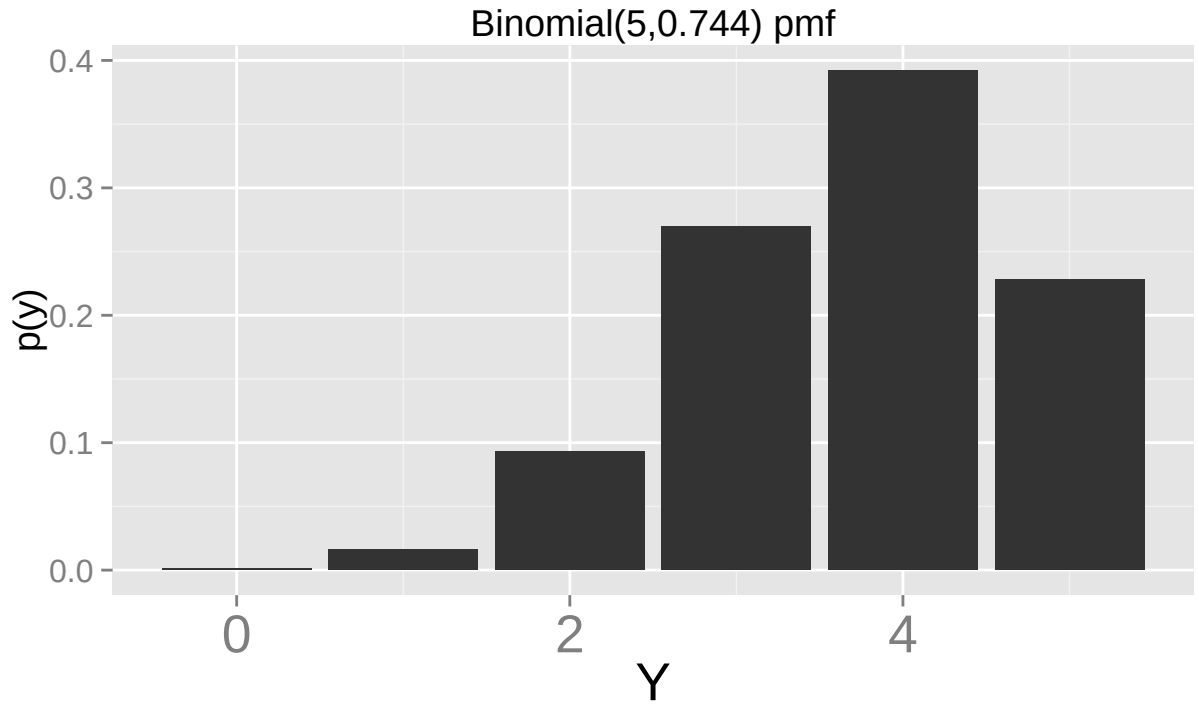
```
## [1] 0.001099512 0.017076790 0.109944720 0.379842143 0.772036835 1.000000000
```

```
#Generate 10 realizations of Y when Y~BIN(5,0.744):
rbinom(10,size=5,prob=0.744)
```

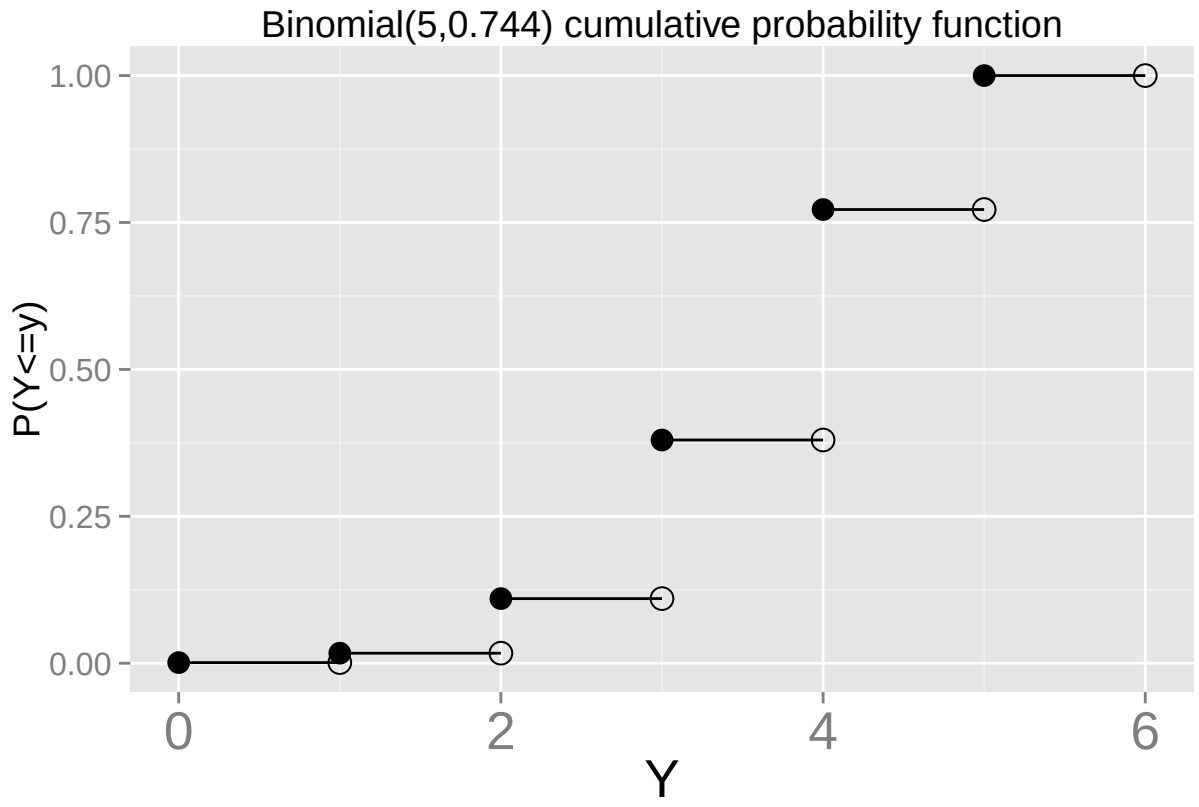
```
##  [1] 2 5 4 4 3 4 4 5 4 4
```

We can use these functions to graph the $BIN(5, 0.744)$ pmf, and the cumulative probability function:

```
mydata <- data.frame(y=y,probs = dbinom(y,size=5,prob=0.744),cumprobs = pbinom(y,size=5,prob=0.744))
ggplot(aes(x=y,y=probs),data=mydata) + geom_bar(stat='identity')+ ylab('p(y)') + xlab('Y') +
  theme(axis.text.x = element_text(size=20), axis.text.y = element_text(size=12),
        axis.title.x = element_text(size=20),axis.title.y = element_text(size=14))  +
        ggtitle('Binomial(5,0.744) pmf')
```



```
##Note the additional aesthetics xend and yend needed by geom_segment();
##see ?geom_segment
##There are two geom_point() layers; one for the filled circles and one for the empty
ggplot(data=mydata) + geom_segment(aes(x=y,xend=y+1,y=cumprobs,yend=cumprobs))+
  geom_point(aes(x=y,y=cumprobs),size=4)+ geom_point(aes(x=y+1,y=cumprobs),size=4,shape=1) +
  ylab('P(Y<=y)') + xlab('Y') +
  theme(axis.text.x = element_text(size=20), axis.text.y = element_text(size=12),
        axis.title.x = element_text(size=20),axis.title.y = element_text(size=14))  +
        ggtitle('Binomial(5,0.744) cumulative probability function')
```



Note that the above plot is slightly misleading since $P(Y \leq y) = 1$ for all $y \geq 5$; it doesn't stop at $y = 6$.

Geometric distribution

The geometric distribution is related to the binomial in that it involves independent Bernoulli trials.

Let p be the probability of success for a single Bernoulli trial. The random variable Y of the number of Bernoulli trials until the 1^{st} success has a geometric distribution, and we say $Y \sim GEO(p)$.

PMF:

$$P(Y = y) = p(y) = pq^{y-1}, \ y = 1, 2, 3, ...$$

Here, as for the binomial, p is the probability of "success" while q is the probability of "failure".

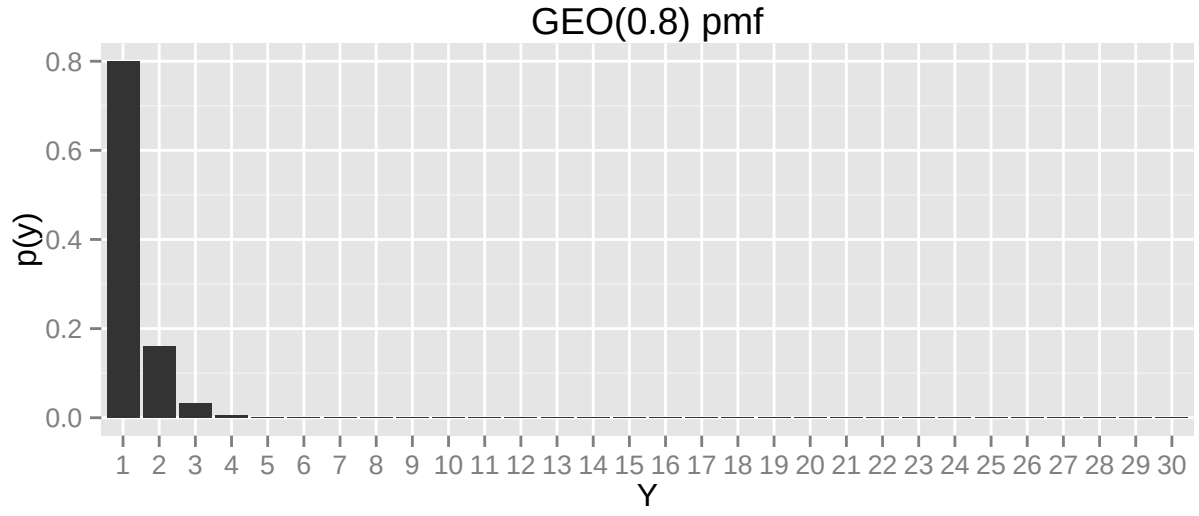
If we reconsider the free-throw example, Y would be the distribution of the number of shots LeBron would need to take before making his first one.

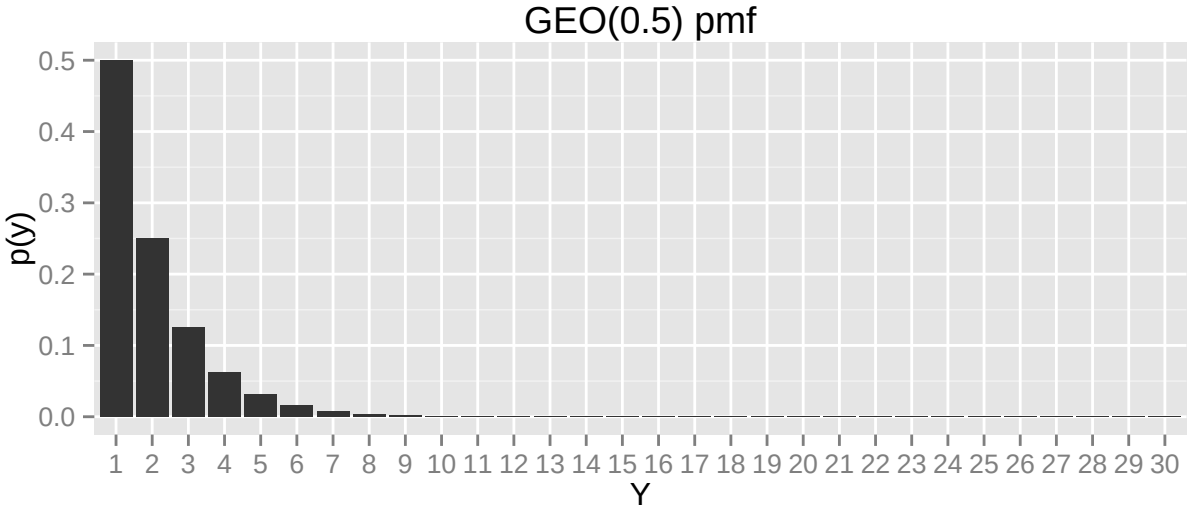
- Reconsider the LeBron example; What is the probability that $Y = 1$? $Y = 2$?

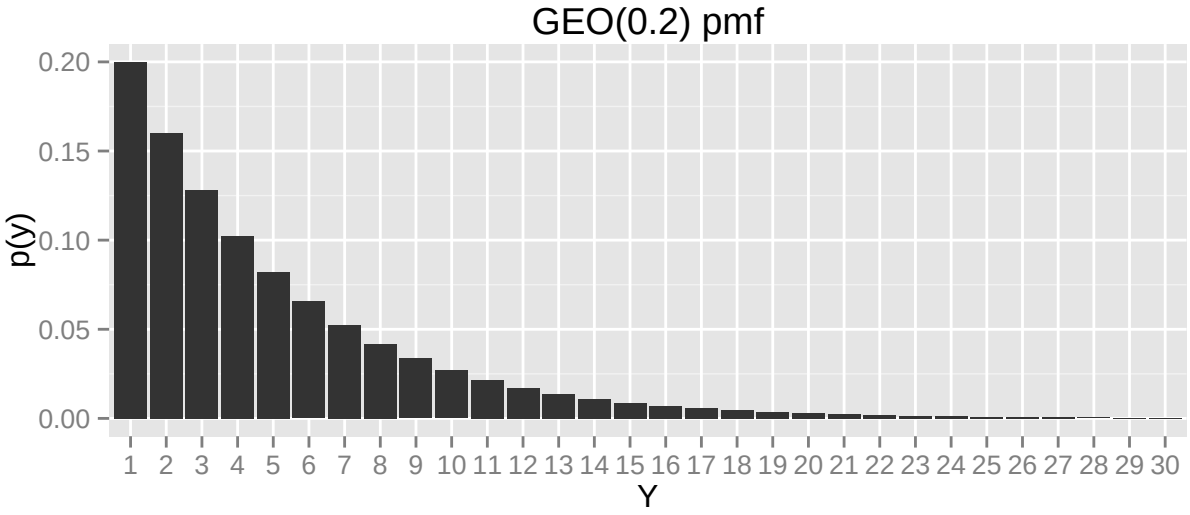
Below is a function that creates graphs of geometric pmf, in `R`, for various values of p.

- Why does their shape make sense?

```
graph.geom.pmf <- function(p) {
  y <- seq(1,30) #Note that y goes to infinity, but we can't graph that obviously!
  p.y <- p*(1-p)^(y-1)
  mytitle <- paste('GEO(',p,') pmf',sep='')
  mydata <- data.frame(y=as.factor(y),probs=p.y)
  ggplot(aes(x=y,y=probs),data=mydata) + geom_bar(stat='identity')+ ylab('p(y)') + xlab('Y') +
        ggtitle(mytitle)
}
graph.geom.pmf(.8)
graph.geom.pmf(.5)
graph.geom.pmf(.2)
```







Aside: geometric series

To study the pmf/expectation/variance/MGF, recall a fundamental concept from Calculus II regarding geometric series:

If $|a| < 1$, then:

$$\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$$

$$\sum_{i=r}^{\infty} a^i = \frac{a^r}{1-a}$$

We will use these facts to investigate properties of the $GEO(p)$ distribution, noting that $q = 1 - p \leq 1$ (equality makes analysis trivial).

Use these facts about geometric series to show:

1. $\sum_{y=1}^{\infty} p(y) = 1$.

2. $E(Y) = \frac{1}{p}$

3. $Var(Y) = \frac{1-p}{p^2}$ **HOMEWORK**

4. MGF: $M_Y(t) = E(e^{tY}) = \frac{pe^t}{1-qe^t}$. Verify that $\frac{d}{dt} M_Y(t)|_{t=0} = E(Y) = \frac{1}{p}$

Geometric distribution in R

Note that the `dgeom()`, `pgeom()`, and `rgeom()` functions in `R` refer to a slightly *different* version of a geometric random variable, namely, letting X be the number of *failures* before the first success. With this version, $X = 0, 1, 2, ...$ and $p(x) = pq^x$.

- Recall the LeBron free-throw example. Suppose we want to simulate several Y, where $Y \sim GEO(0.744)$. How would we do this (easily) in R?

A from-scratch way to generate geometric data would be to use a `while()` loop. Given an initial state (`success.status='failure'`), the loop will continue to run, updating `y` each time, until the first "success" is observed. At this point, the loop is broken and the most recent version of `y` is returned:

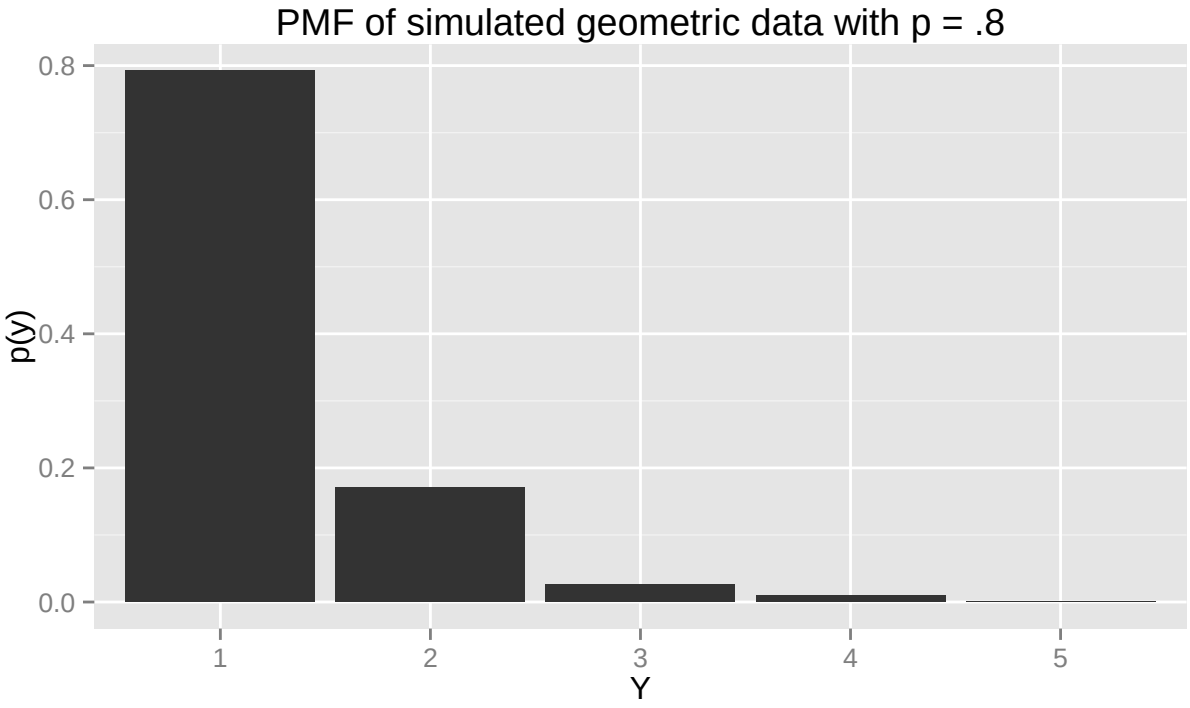
```
one.geo.Y <- function(p) {
  success.status <- 'failure'
  y <- 0
  while(success.status=='failure') {
    success.status <- sample(c('success','failure'),1,prob=c(p,1-p))
    y <- y + 1 #y is at least 1
  }
return(y)
}
sim.data <- data.frame(Y.8 = replicate(1000,one.geo.Y(.8)),
                       Y.5 = replicate(1000,one.geo.Y(.5)),
                       Y.2 = replicate(1000,one.geo.Y(.2)))
apply(sim.data,2,mean) ##What should these equal?
```

```
##   Y.8   Y.5   Y.2
## 1.257 1.991 4.955
```

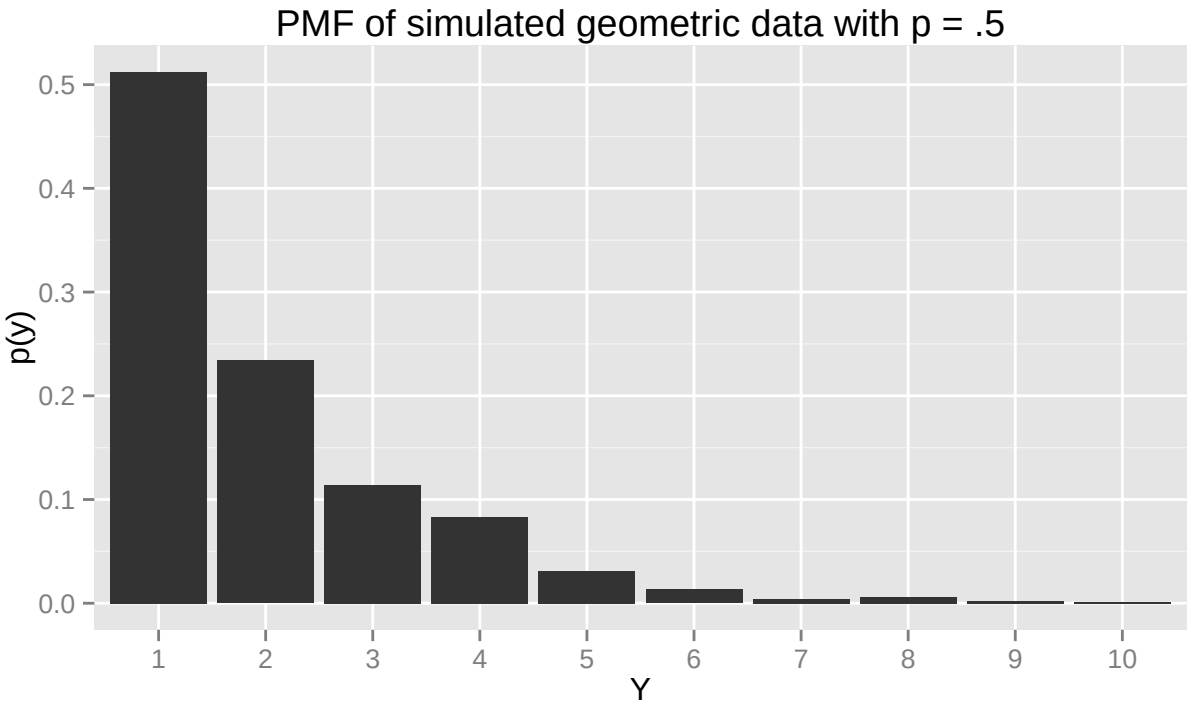
```
apply(sim.data,2,var)  ##What should these equal?
```

```
##        Y.8        Y.5        Y.2
##  0.3152663  1.9248438 20.2452202
```

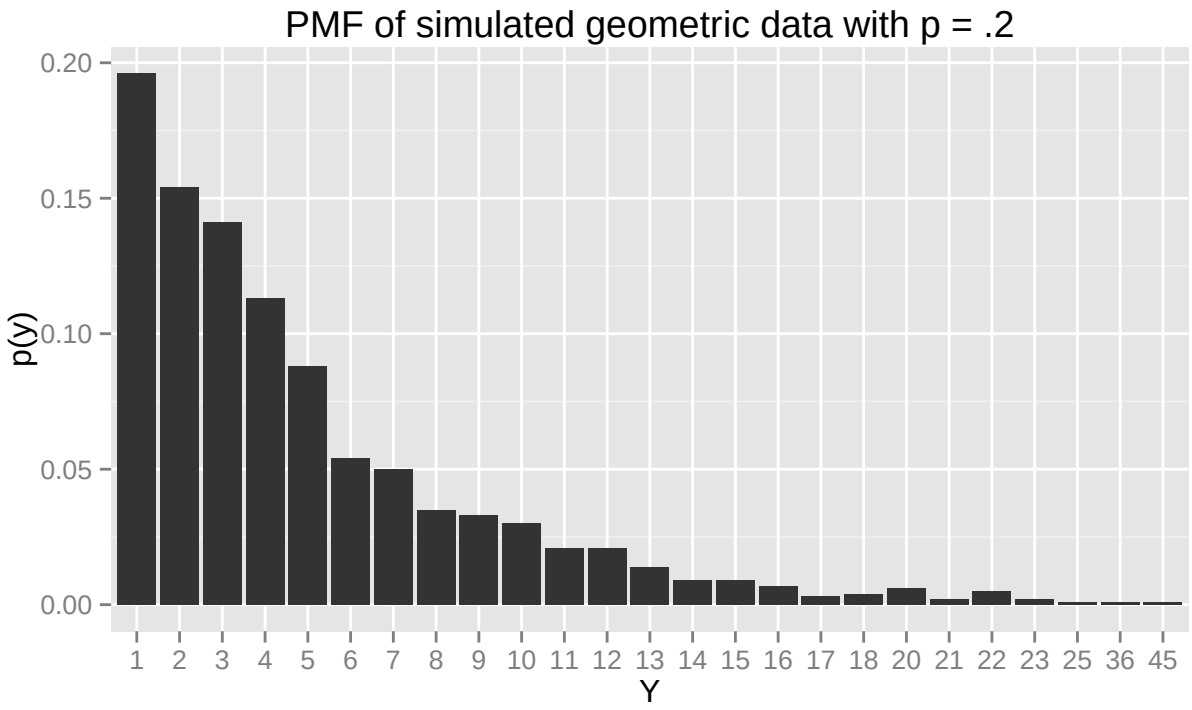
```
ggplot(aes(x=as.factor(Y.8)),data=sim.data) + geom_bar(aes(y=(..count..)/sum(..count..))) +
    ylab('p(y)') + xlab('Y') + ggtitle('PMF of simulated geometric data with p = .8')
```




```
ggplot(aes(x=as.factor(Y.5)),data=sim.data) + geom_bar(aes(y=(..count..)/sum(..count..))) +
    ylab('p(y)') + xlab('Y') + ggtitle('PMF of simulated geometric data with p = .5')
```



```
ggplot(aes(x=as.factor(Y.2)),data=sim.data) + geom_bar(aes(y=(..count..)/sum(..count..))) +
    ylab('p(y)') + xlab('Y') + ggtitle('PMF of simulated geometric data with p = .2')
```



It's nice to see that the `while()` loop method of generating $GEO(p)$ random variables does indeed give us simulated data with the properties we would expect!

Negative binomial distribution

Related to the geometric distribution is the negative binomial distribution. Rather than the number of trials needed to achieve the 1st successes, it refers to the distribution of the number of trials needed to obtain any r^{th} success.

If Y = number of Bernoulli trials needed to observe the r^{th} success, then $Y \sim NB(r,p)$.

PMF:

$$P(Y=y) = p(y) = \binom{y-1}{r-1} p^r q^{y-r}, \ y = r, r+1, ...$$

Gist:

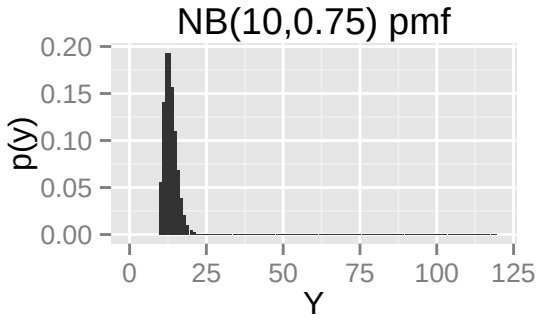
$$F \ S \ S \ ... \ S \ F \ \mathbf{S} \leftarrow r^{th} \ success; y \ trials \ needed$$

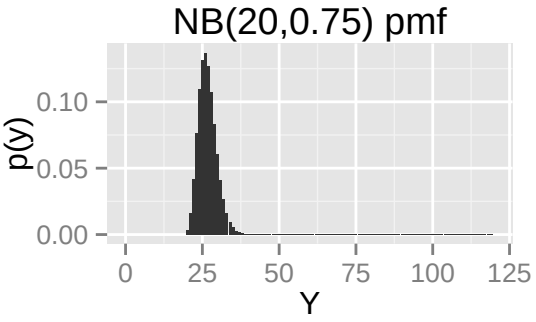
There are $y-1$ trials before the r^{th} success; out of which we choose $r-1$ spots for the successes (equivalently, choose $y-r$ of the $y-1$ spots for the failures).

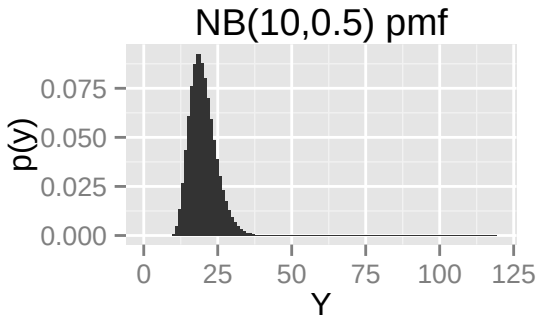
- The $GEO(p)$ distribution is a specific case of the $NB(r,p)$ distribution: which case?

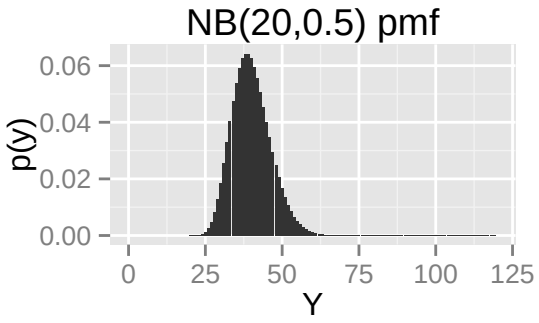
Let's look at some NB pmfs:

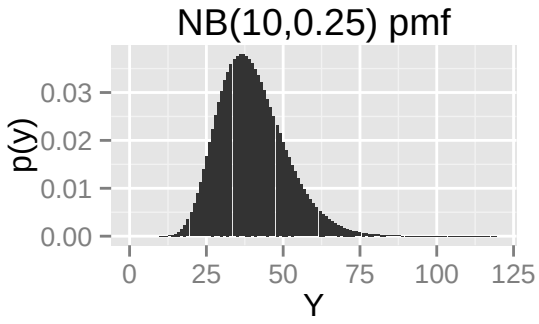
```
plot.nb.pmf <- function(r,p) {
  y <- seq(r,r+120,by=1)
  p.y <- choose(y-1,r-1)*p^r*(1-p)^(y-r)
  mydata <- data.frame(y=y,probs=p.y)
  ggplot(aes(x=y,y=probs),data=mydata) + geom_bar(stat='identity')+ ylab('p(y)') + xlab('Y') +
        ggtitle(paste('NB(',r,',',p,') pmf',sep=''))+xlim(c(0,120))
}
plot.nb.pmf(10,.75)
plot.nb.pmf(20,.75)
plot.nb.pmf(10,.5)
plot.nb.pmf(20,.5)
plot.nb.pmf(10,.25)
plot.nb.pmf(20,.25)
```

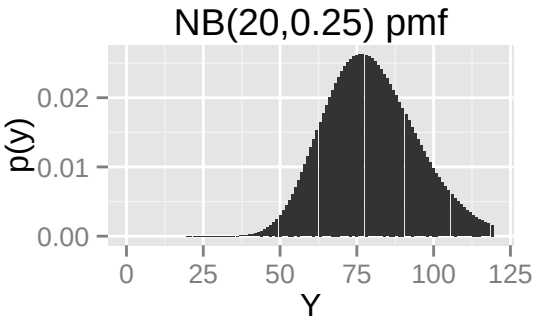












Mean, variance

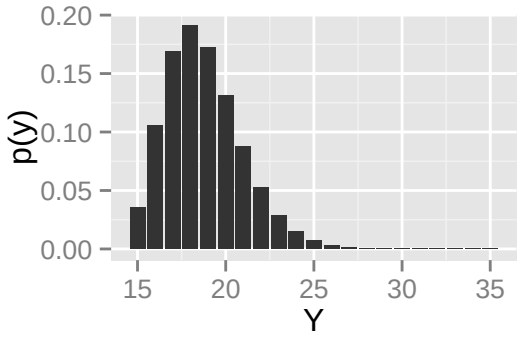
1. $E(Y) = \frac{r}{p}$

This makes sense: if LeBron is a 75% free-throw shooter, how many shots would you expect him to need to take in order to make $r = 15$ of them?

2. $Var(Y) = \frac{rq}{p^2}$

- How does the variance depend on r and p? How is this corroborated in the PMFs graphed above?

In class activity: For what value of y is $p(y)$ maximum? I.e., if $Y \sim NB(r, p)$ for what y is the PMF bar tallest, *in general*?



The gist: find the largest y such that $p(y)/p(y-1) > 1$, by following these steps:

1. Show that:

$$\frac{p(y)}{p(y-1)} = \left(\frac{y-1}{y-r}\right) q.$$

2. Show that:

$$\frac{p(y)}{p(y-1)} > 1 \text{ as long as } y < \frac{r-q}{1-q}$$

.

3. Use the above facts to show that $p(y)$ is maximized when:

$$y = \left\lfloor \frac{r-q}{p} \right\rfloor$$

4. Use facts 1-3 to find $max_y p(y)$ when $Y \sim NB(5, 0.3)$. Graph $p(y)$ for this case and verify your result.

Negative binomial in R

`R` again uses a different version of the negative binomial; specifically, if $X \sim NB(r,p)$ then:

$$P(X=x) = p(x) = \binom{x+r-1}{x} p^r q^x.$$

In this context, X is the number of *failures* needed before observing the r^{th} success.

- What is the relationship between X and Y?

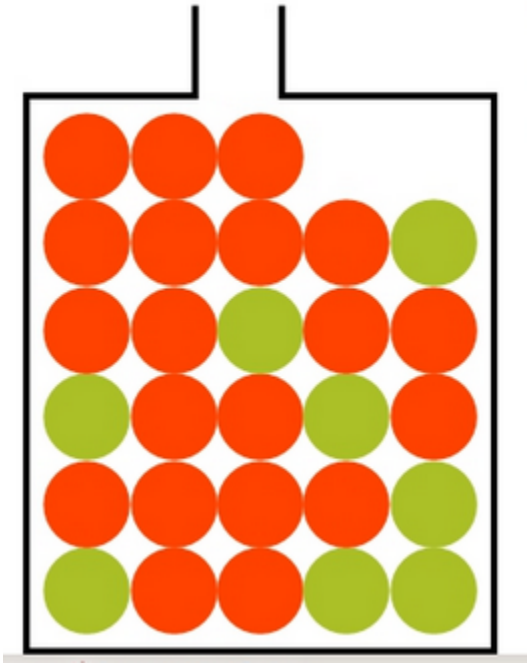
The functions `dnbinom()`, `pnbinom()`, and `rnbinom()` can be used to calculate exact and cumulative probabilities and generate $NB(r,p)$ random variables.

- (**Homework**) How could we use a `while()` loop to generate $NB(r,p)$ random variables? Getting started:

```
one.NB.Y <- function(r,p) {
  num.trials <- 0
  num.successes <- 0
  while(num.successes < r){
    ##generate 1 Bernoulli trial
    ##update num.trials
    ##update num.successes (hint: ?ifelse())
  }
  return(what should you return here?)
}
```

Hypergeometric distribution

This distribution describes a random variable obtained when sampling *without replacement*. It is often motivated using "balls in urns" to describe the distribution of the number of balls of a certain type sampled when drawing n total.



Example: From the figure above, suppose that Y is the number of red balls out of a sample of 5 drawn.

A random variable with a hypergeometric distribution, $Y \sim HYPERGEO(r, n, N)$, is parameterized by the following:

- r: the number of balls of a certain type. In our example, $r =$_____ .
- n: the sample size. In our example, $n =$_____.
- N: the "population size", i.e. the total number of balls in the urn. In the graph above, $N =$_____.

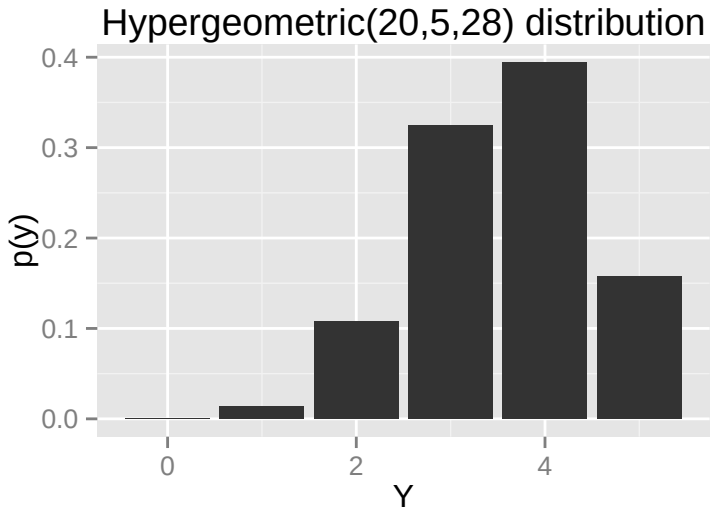
PMF:

$$P(Y = y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}, \quad min(0, n - N + r) \leq y \leq min(r, n)$$

- What are the limits on y for this example?

Graphing $p(y)$:

```
y <- 0:5
py <- choose(20,y)*choose(8,5-y)/choose(28,5)
mydata <- data.frame(y=y,probs=py)
  ggplot(aes(x=y,y=probs),data=mydata) + geom_bar(stat='identity') +
  ylab('p(y)') + xlab('Y') +
  ggtitle('Hypergeometric(20,5,28) distribution')
```



Example: capture/recapture

The hypergeometric distribution has a very interesting application in capture/recapture studies. In these studies, the intent is to estimate the total population size, N, for example the number of a type of animal species on an island. This is done as follows:

1. Capture r animals and tag them, then release them.
2. At a later date, capture n animals
3. Let \$y = \$ the number of tagged animals in second sample of size n.

In this context, $Y \sim HYPERGEO(r, n, N)$, where the parameters r and n are known, but not the parameter N. Interest lies in estimating N. How does this work?

First, write out the hypergeometric PMF:

$$p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}},$$

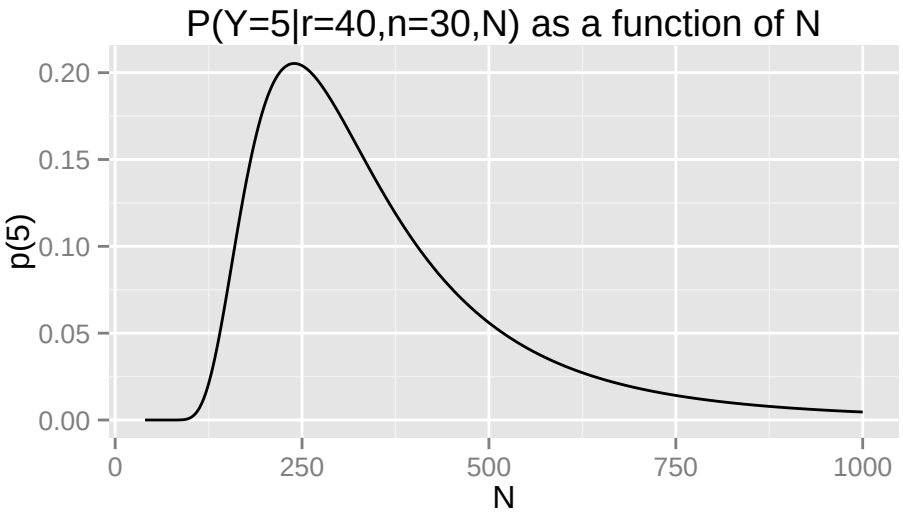
where the only unknown quantity is N. We want to find N that maximizes this probability, given the y we observed and the known r and n. In other words, we want to find the *maximum likelihood estimate* (MLE) of the unknown population size N given the data.

First things first: let's consider some specific values:

- $r = 40$ animals tagged and released
- $n = 30$ animals resampled
- $y = 5$ animals in new sample with tags

Let's graph the hypergeometric pmf in this case as a function of N. Note that it would be silly to consider any values of $N < 40$!

```
N <- seq(40,1000,by=1)
py <- choose(40,5)*choose(N-40,30-5)/choose(N,30)
mydata <- data.frame(N=N,probs=py)
  ggplot(aes(x=N,y=probs),data=mydata) + geom_line() +
  ylab('p(5)') + xlab('N') +
  ggtitle('P(Y=5|r=40,n=30,N) as a function of N')
```




```
N[which.max(py)]
```

```
## [1] 240
```

Thus given the number initially tagged, the number resampled, and the number of tagged found in the second sample, it would be reasonable to estimate a population size of $N = 240$.

- Does this make sense intuitively?

Let's verify this analytically, and for any given r, n, and y.

We will proceed in a similar manner to how we proceeded in the negative binomial setting, by finding the largest N such that:

$$\frac{P(Y = y|r, n, N)}{P(Y = y|r, n, N - 1)} \geq 1.$$

Finding the MLE of N

The Poisson Distribution

The Poisson distribution is frequently used to model Y when Y is the number of events per space/time unit. Some examples:

1. Number of arrivals at a drive-thru window per hour
2. Number of microorganisms per mL of water
3. Number of highway entrances per minute during rush hour
4. Number of frogs per cubic foot in a pond

One motivation for the Poisson distribution is that is arises as a limit of the binomial distribution as $n \to \infty$ and for small p.

Example from Section 3.8

Let Y be the random variable that denotes the number of accidents that occur during a 1-week time period at a particular intersection. Consider splitting the 1 week up into n time intervals. For n large enough, the intervals will be sufficiently small that, at most, only 1 accident could occur per time subinterval.

- Visualization:

$$P(\text{no accidents occur in a subinterval}) = 1 - p$$

$$P(1 \text{ accident occurs in subinterval}) = p$$

$$P(>1 \text{ accident occurs in subinterval}) = 0$$

From here, it is obvious that as n increases, p decreases. It also follows that Y can be re-represented as the number of subintervals that contain an accident, i.e. the number of "successes" out of n Bernoulli trials. Thus:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

However, we want to consider this probability as $n \to \infty$ with corresponding $p \to 0$. Suppose this happens at a constant rate, such that $\lambda = np$. We can show that under these conditions,

$$p(y) \to \frac{e^{-\lambda}\lambda^y}{y!},\ y = 0, 1, 2, ...$$

This PMF is the *Poisson pmf*, and we say that $Y \sim POI(\lambda)$.

Before we walk through the proof, recall some calc facts:

1. $\lim_{x\to\infty} f(x) = L \iff \lim_{x\to\infty} \ln(f(x)) = \ln(L)$
2. L'Hopital's Rule: if $\lim_{x\to\infty} \frac{f(x)}{g(x)} = \frac{0}{0}$ or $\frac{\infty}{\infty}$, then $\lim_{x\to\infty} \frac{f'(x)}{g'(x)}$
3. The previous two bullets imply that for any constant c:

$$\lim_{n\to\infty} \left(1 + \frac{c}{n}\right)^n = e^c$$

PROOF:

PROOF (continued)

It is quite easy to show that $p(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}$ is a valid PMF by showing that it sums to 1:

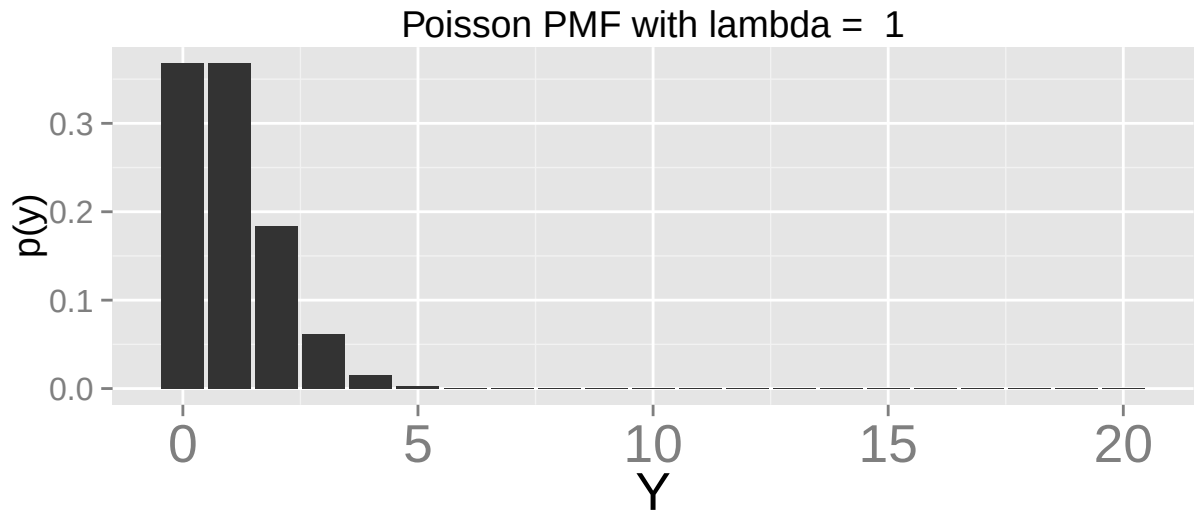
Mean, variance, MGF
If $Y \sim POI(\lambda)$, then:

1. $E(Y) = \lambda$
   Proof:

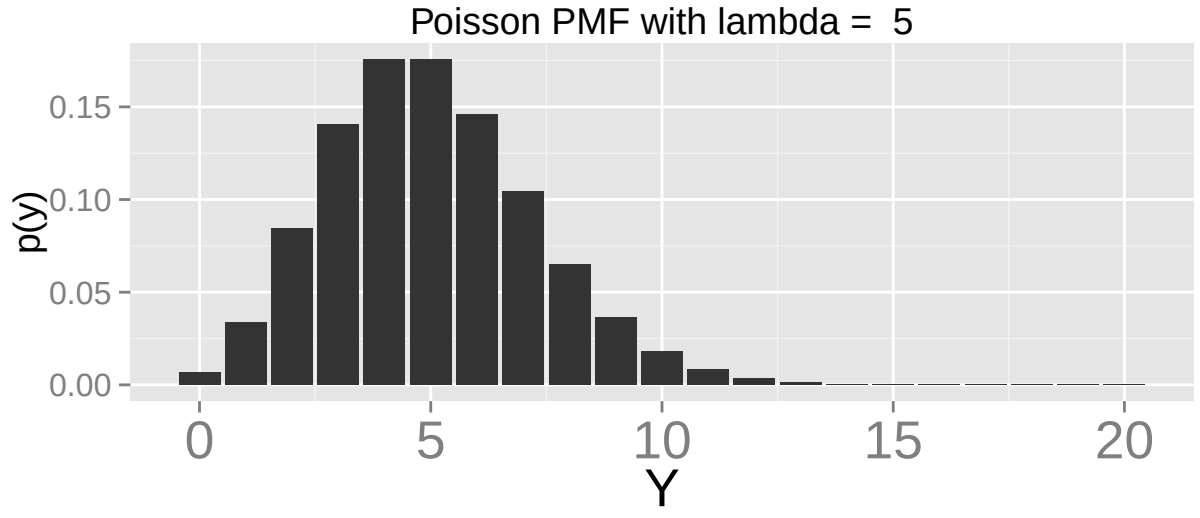
2. $Var(Y) = \lambda$; *pertinent feature of the Poisson distribution*
   Proof: HOMEWORK
3. MGF: $M_Y(t) = e^{\lambda e^t - \lambda}$
   Proof:

Let's visualize the Poisson PMF:

```
poisson.pmf <- function(lambda) {
yvalues <- seq(0,20,by=1)
probabilities <- (exp(-lambda)*lambda^yvalues)/factorial(yvalues)
mydata <- data.frame(y=yvalues,prob=probabilities)
mytitle <- paste('Poisson PMF with lambda = ',lambda)
ggplot(aes(x=y,y=prob),data=mydata) + geom_bar(stat='identity') + ylab('p(y)') + xlab('Y') +
  theme(axis.text.x = element_text(size=20), axis.text.y = element_text(size=12),
        axis.title.x = element_text(size=20),axis.title.y = element_text(size=14))  +
        ggtitle(mytitle)
}
poisson.pmf(1)
```




```
poisson.pmf(5)
```



```
poisson.pmf(10)
```