MATH 323 - DunderMifflin Golden Tickets

Tyrel Stokes

October 2020

Email tyrel.stokes@mail.mcgill.ca for any questions about these problems.

1. a) Let $Y_1 \sim Geometric(p)$. Find the Moment Generating function of Y_1.

Solution:

2 ways. 1st exploits geometric series.

$$\begin{aligned}
M_{Y_1}(t) &= E[e^{tY_1}] && (1)\\
&= \sum_{y=1}^{\infty} e^{ty_1}(1-p)^{y_1-1}p && (2)\\
&= \frac{e^t}{e^t}\sum_{y=1}^{\infty} e^{ty_1}(1-p)^{y_1-1}p && (3)\\
&= p\frac{e^t}{e^t}\sum_{y=1}^{\infty} e^{ty_1}(1-p)^{y_1-1} && (4)\\
&= pe^t\sum_{y=1}^{\infty} e^{t(y_1-1)}(1-p)^{y_1-1} && (5)\\
&= pe^t\sum_{y=1}^{\infty} (e^t(1-p))^{y_1-1} && (6)\\
&= \frac{pe^t}{1-e^t(1-p)} && (7)
\end{aligned}$$

Where the geometric sum only converges if $e^t(1-p) < 1$, which is equivalent to $t < -\ln(1-p)$. Since $-ln(1-p) > 0$, the moment generating function exists in a neighbourhood of 0.

The second way is the same until line 6 above and then instead of exploiting the geometric series, we exploit the fact that we know that valid pmfs sum to 1. So picking up from line 6 above:

$$\begin{aligned} pe^t\sum_{y=1}^{\infty}(e^t(1-p))^{y_1-1} &= \frac{1-e^t(1-p)}{1-e^t(1-p)}pe^t\sum_{y=1}^{\infty}(e^t(1-p))^{y_1-1} \\ &= \frac{pe^t}{1-e^t(1-p)}\sum_{y=1}^{\infty}(e^t(1-p))^{y_1-1}(1-e^t(1-p)) \\ &= \frac{pe^t}{1-e^t(1-p)} \end{aligned}$$

Where we used the fact that $\sum_{y=1}^{\infty}(e^t(1-p))^{y_1-1}(1-e^t(1-p)) = \sum_{y=1}^{\infty}(1-p^\star)^{y_1-1}(p^\star) = 1$, if $1-e^t(1-p) = p^\star$ is a valid probability since it would be the sum of a particular geometric distribution. Once again this is equivalent to $t < -ln(1-p)$.

b) Suppose that $Y_1, Y_2, \ldots, Y_k \sim Geometric(p)$ are identically and independently distributed geometric variables with parameter p. Show that X, $X = \sum_{i=1}^{k} Y_i$ has a negative binomial distribution.

Use the fact that if $X \sim NegBin(r,p)$ is a negative binomial, the it has the following moment generating function

$$M_X(t) = \left(\frac{pe^t}{(1-(1-p)e^t}\right)^r, t < -\ln(1-p)$$

.

In this parametrization of the negative binomial, X models the number of trials until the rth success, so that $P(X=x) = \binom{x-1}{r-1}p^r(1-p)^{x-r}, x \in r, r+1, r+2, \ldots$.

(If you want more MGF practice, derive the MFG above yourself. I will include a solution to get this result when I release the solutions)

Solution:

$$\begin{aligned} M_X(t) &= E[e^{tX}] \\ &= E[\exp(t\sum_{i=1}^{k}Y_1)] \\ &= E[e^{tY_1}e^{tY_2}\ldots e^{tY_k}] \\ &\overset{independent}{=} \prod_{-1}^{k}E[e^{tY_i}] \\ &\overset{ident}{=} (E[e^{Y_1}])^k \\ &= (\frac{pe^t}{1-e^t(1-p)})^k, t < -\ln(1-p) \end{aligned}$$

Where we used the Gamma MGF we derived in part a. Thus X follows a negative binomial with parameters p and k by the uniquness theorem (If the MGF exists in a neighbourhood of zero, then the MGF is sufficient to characterize the distribution. i.e, if the MGF is that of a negative binomial, the variable must be negative binomially distributed).

See the end of this document for a proof of the Moment generating function for the negative binomial.

c) Michael Scott from the Dundermifflin paper company has organized a promotional sale. He has hidden 200 golden tickets in different boxes of paper in the warehouse. A Dundermifflin customer who finds a golden ticket is entitled to a 0.5% discount on all purchases for next year for each ticket they find. Darryl, the warehouse manager, is worried that Michael forgot to randomly distribute the tickets across the warehouse and that a single customer might end up with several or many of the tickets and the company will go bankrupt. Darryl sends Creed (the quality assurance representative) to perform a check to see if the tickets are well-dispersed. Creed starts randomly opening boxes in the southwest corner of the warehouse checking whether there is a golden ticket in it or not. He keeps track of how many tickets he finds, but forgets to remember which boxes he opens (and thus when he randomly picks a new box he may be opening a box he has already checked). Creed finds his 100th ticket on exactly the 800th try. If Michael had truly randomly dispersed the tickets, what is the probability of creed finding his 100th ticket on his 800th time. Suppose that there are 2,000 boxes in the warehouse. Do not find the numerical value, but an expression which evaluates correctly the desired probability.

Solution:

If Michael had randomly distributed the tickets, given Creeds selection method, the number of tries it takes to reach 100 golden tickets found is negative binomially distributed with parameters $p = \frac{200}{2000} = 0.1$ and $r = 100$. (Notice that if Creed had done the search more efficiently and kept track of which boxes he opened we would get a hypergeometric distribution). Therefore:

$$\begin{aligned} P(X = 800) &= \binom{(800-1)}{(100-1)}(1-p)^{700}p^{100} \\ &= \binom{799}{99}(0.9)^{700}(0.1)^{100} \end{aligned}$$

d) If Michael had truly randomly distributed the golden tickets, what is the probability that we would expect to find Creed to have found his 100th in 800 or more attempts (using his less than perfect inspection method). Find an approximation to this probability using the information from the previous parts. Use a table or calculator to find the value of the approximation.

Solution:

We want to find an approximation to the probability that $P(X \geq 800)$. In part b we showed that a negative binomial random variable can be thought of as a sum of k independent and identical geometric random variables, where k is the number of successes each with parameter p. In this context, we have that $k = 100$, since creed finds 100 golden tickets and $p = 0.1$ from above.

$$P(X \geq 800) = P(\sum_{i=1}^{100} Y_i \geq 800)$$

Where $Y_i \sim Geometric(0.1)$.

In class we have learned about the Central Limit Theorem (CLT). The CLT is useful whenever we have a sum of a sufficiently large number of independent variables (when the mean and variance of those variables is finite). Since we can express a negative binomial as a sum of independent and identical geometric variables we can think about whether or not this approximation makes sense. The number of terms in the sum depends on the number of success, which in this case is large (100), thus the normal approximation should be reasonably accurate.

$$\begin{aligned} P(\sum_{i=1}^{100} Y_i \geq 800) &= P(\frac{\sum_{i=1}^{100} Y_i - E[\sum_{i=1}^{100} Y_i]}{\sqrt{Var(\sum_{i=1}^{100} Y_i)}} \geq \frac{800 - E[\sum_{i=1}^{100} Y_i]}{\sqrt{Var(\sum_{i=1}^{100} Y_i)}}) \\ &\overset{CLT}{\approx} P(Z \geq \frac{800 - E[\sum_{i=1}^{100} Y_i]}{\sqrt{Var(\sum_{i=1}^{100} Y_i)}} \end{aligned}$$

Where $Z \sim N(0,1)$

$$\begin{aligned} E[\sum_{i=1}^{100} Y_i] &\overset{\text{ind., id.}}{=} 100 * E[Y_1] \\ &\overset{\text{geometric dist}}{=} 100 * \frac{1}{p} \\ &= \frac{100}{0.1} \\ &= 1000 \end{aligned}$$

$$\begin{aligned} Var(\sum_{i=1}^{100} Y_i) &\overset{\text{ind., id.}}{=} 100 * Var(Y_1) \\ &\overset{\text{geometric dist}}{=} 100 * \frac{(1-p}{p^2} \\ &= 100 * \frac{0.9}{.01} \\ &= 100 * 90 \\ &= 9000 \end{aligned}$$

Putting it altogether

$$\begin{aligned}
P(Z \geq \frac{800 - E[\sum_{i=1}^{100} Y_i]}{\sqrt{Var(\sum_{i=1}^{100} Y_i)}} &= P(Z \geq \frac{800 - 1000}{\sqrt{9000}} \\
&= P(Z \geq \frac{-200}{\sqrt{9000}} \\
&= 1 - \Phi(\frac{-200}{\sqrt{9000}}) \\
&\approx .982
\end{aligned}$$

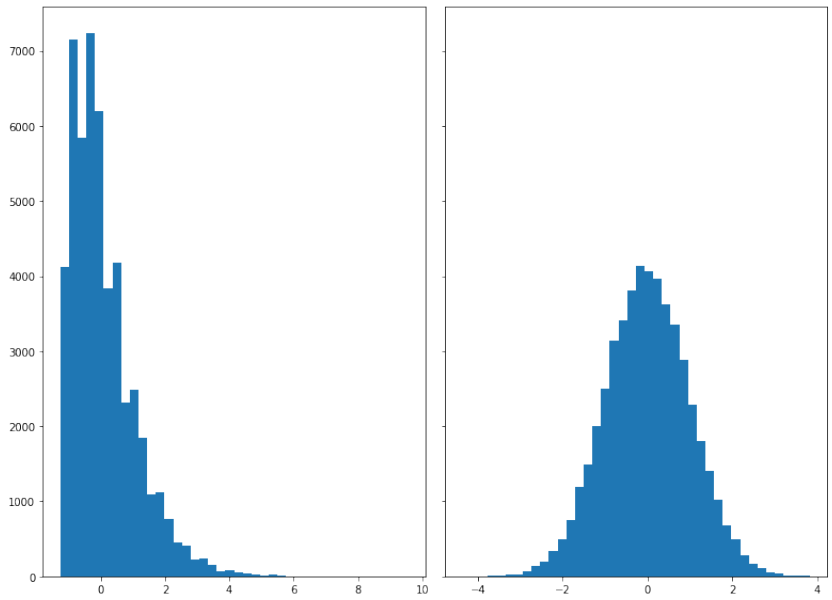
This turns out to be a reasonable approximation. Using the negative binomial directly (and some fancier computer code) the true probability is $\approx .987$.

These numbers should cast some doubt on whether or not Michael actually spread the golden tickets out evenly. Either that or Creeds check was even more dubious than it appeared.

Below is some visualization of the how the CLT gives us better and better approximations as k increases. NB: The CLT itself doesn't actually tell us how fast it becomes a good approximation, it just says that for large n the approximation is good. However, in simple cases we have other results which give us a notion of the rate at which the standardized quantities become normal.

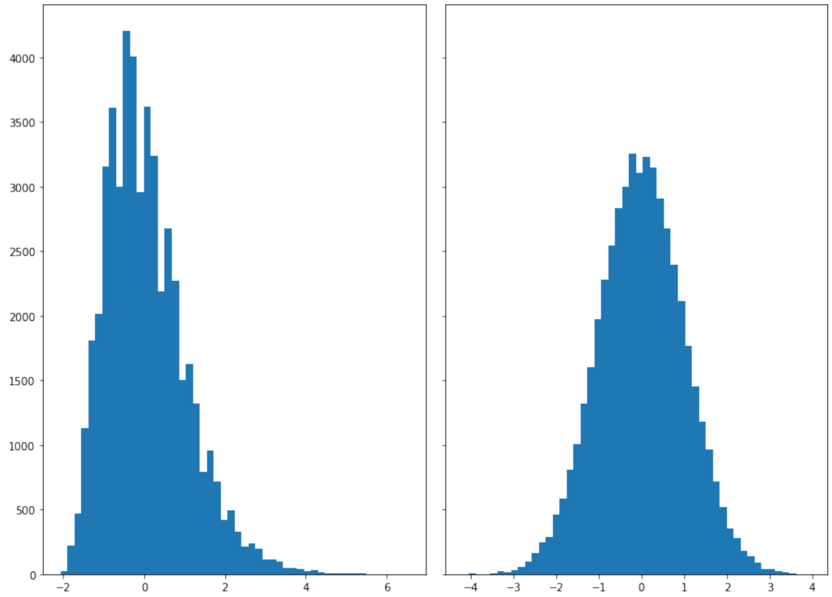
With this in mind, the negative binomial will increasingly be well approximated by a standard normal as the parameter for the number of successes grow, the large this parameter is the more indendent geometrics we can think of the variable being generated from. I will visualize the negative binomial with parameter $p = 0.1$ with varying number of success parameter once it is standarded.

Let $k = 2$

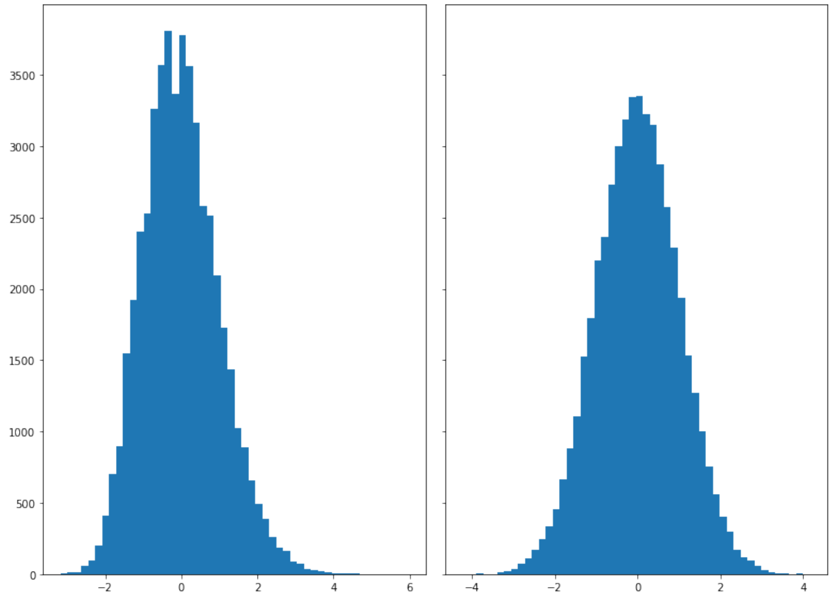


The standardized negative binomial is on the left, a standard normal on the right. They negative binomial is very skewed and clearly not well approximated by a standard normal

Let $k = 5$

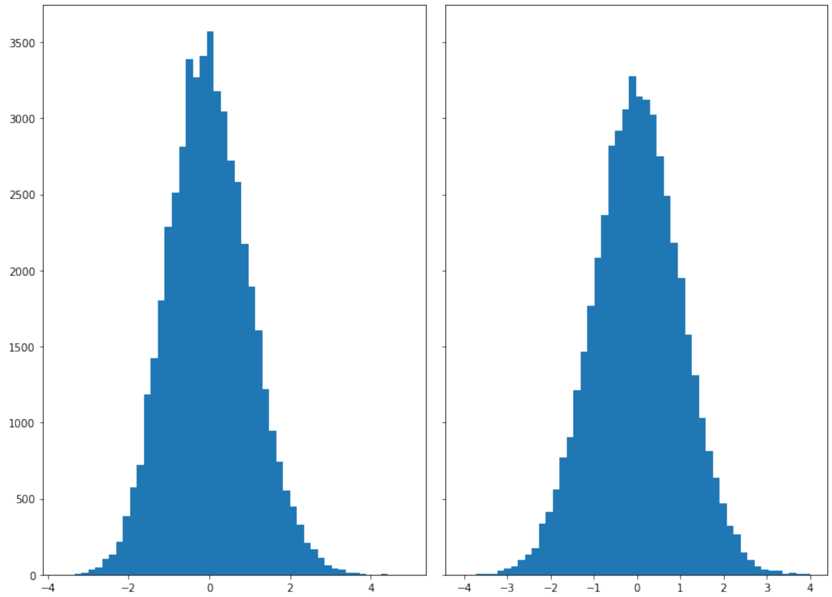


Let $k = 25$

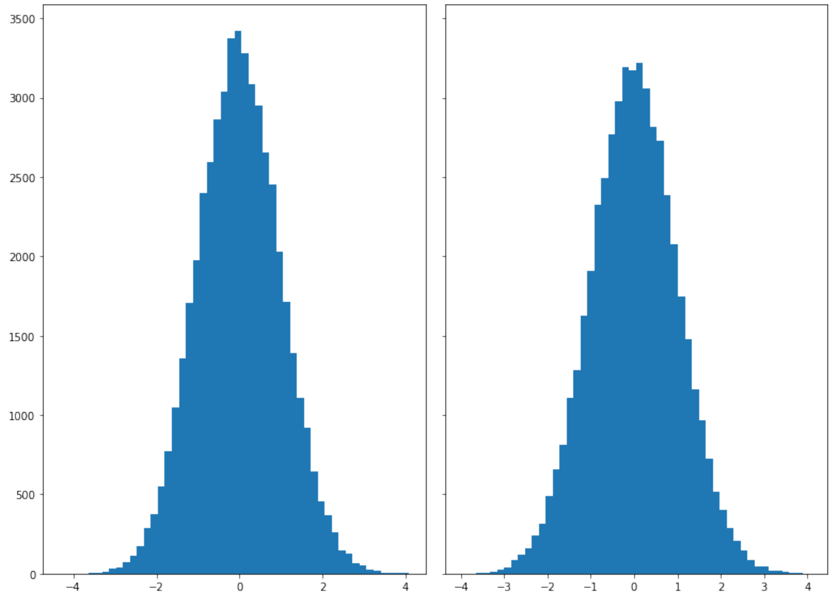


Already with $K = 25$ the standardized normal is beginning to be less skewed and resemble the standard normal more closely.

Let $k = 100$



Let $k = 1000$



At 1000, the negative binomial is very well approximated by a standard normal.

e) Supposing again that Michael truly randomly distributed the golden tickets, what is the expected number of boxes for Creed to open to find 100 golden tickets? Using what you have learned in class bound the probability of Creed finding 100 golden tickets within 2 standard deviations of this mean.

Solution:

$$\begin{aligned}
E[X] &= E[\sum_{i=1}^{100} Y_i] && (8)\\
&= \sum_{i=1}^{100} E[Y_i] && (9)\\
&= 100 * \frac{1}{p} && (10)\\
&= 1000 && (11)
\end{aligned}$$

We used this in the last part already.

Now we would like to bound the probability that the number of boxes to reach 100 golden tickets is within or equal to 2 standard deviations away from the mean.

$$\begin{aligned}
P(|X - E[X]| \leq 2 * sd(X)) &= 1 - P(|X - E[X]| > 2 * sd(X))\\
&\overset{Markov/Chebyshev's}{\geq} 1 - \frac{E[(X - E[x])^2]}{(2 * sd(X))^2}\\
&= 1 - \frac{Var(X)}{4 * Var(X)}\\
&= 1 - \frac{1}{4}\\
&= \frac{3}{4}
\end{aligned}$$

The Chebyshev bound tells us that there is a probability of $\frac{3}{4}$ or greater of being within 2 standard deviations of the mean.

To find the actual probability:

$$\begin{aligned}
P(|X - E[X]| \leq 2 * sd(X)) &= P(X \in [-2 * sd(X) + E[X], 2 * sd(X) + E[X]])\\
&= F_X(2 * sd(X) + E[X]) - P(X < -2 * sd(X) + E[X])\\
&\approx .95
\end{aligned}$$

This bound isn't particularly sharp in this case, but does give usa rough idea of where the true probability lies and it is very easy to compute. In order to numerically evaluate the cdf for the negative binomial required statistical software, whereas the bound required only the simplification of small fractions. Notice also that the bound actually does not depend on the distribution of X, only that X has some variance $Var(X)$. The proof we did is quite general. It tells us that for any random variable X with finite mean and variance, that the probability of that random variable falling within 2 standard deviations is $\frac{3}{4}$.

Negative Binomial MGF Proof

Let $X \sim NegBin(r, p)$

$$\begin{aligned}
MGF_X(t) &= E[e^{tX}] \\
&= \sum_{x=r}^{\infty} e^{tx} \binom{(x-1)}{(r-1)} (1-p)^{x-r} p^r \\
&= p^r \sum_{x=r}^{\infty} e^{tx} \binom{(x-1)}{(r-1)} (1-p)^{x-r} \\
&= \frac{e^{tr}}{e^{tr}} p^r \sum_{x=r}^{\infty} e^{tx} \binom{(x-1)}{(r-1)} (1-p)^{x-r}
\end{aligned}$$

The strategy we are following is to try and rewrite the infinite sum in the form of some other negative binomial pmf, which we know sums to 1. The first step is to pull out the constants (p^r) and multiply by a carefully chosen $1 = \frac{e^{tr}}{e^{tr}}$.

$$\begin{aligned}
&= p^r e^{tr} \sum_{x=r}^{\infty} e^{t(x-r)} \binom{(x-1)}{(r-1)} (1-p)^{x-r} \\
&= (pe^t)^r \sum_{x=r}^{\infty} \binom{(x-1)}{(r-1)} (1-p)^{x-r} (e^t)^{x-r} \\
&= (pe^t)^r \sum_{x=r}^{\infty} \binom{(x-1)}{(r-1)} (e^t(1-p))^{x-r}
\end{aligned}$$

Out carefully chosen one in the first section now allows us to pull the annoying term (e^{tx}) together with $(1-p)$ forming the kernel of some new negative binomial mass function. To complete this, we now just need to get $(1-p^\star)^r, p^\star = e^t(1-p)$ into the summands. Since this doesn't depend on x, we can once again multiply by a carefully chosen 1.

$$\begin{aligned}
&= (pe^t)^r \frac{(1-e^t(1-p))^r}{(1-e^t(1-p))^r} \sum_{x=r}^{\infty} \binom{(x-1)}{(r-1)} (e^t(1-p))^{x-r} \\
&= \frac{(pe^t)^r}{(1-e^t(1-p))^r} \sum_{x=r}^{\infty} \binom{(x-1)}{(r-1)} (e^t(1-p))^{x-r} (1-e^t(1-p))^r \\
&= \frac{(pe^t)^r}{(1-e^t(1-p))^r}, \quad t < -ln(1-p) \\
&= \left(\frac{pe^t}{(1-e^t(1-p))} \right)^r. \quad t < -ln(1-p)
\end{aligned}$$

Where the sum converges to 1 if and only if the summands was a valid negative binomial pmf. The sufficient condition is that $p^\star$ is a valid probability, which is equivalent to $t < -ln(1-p)$.