Stat 110 Strategic Practice 5, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Poisson Distribution and Poisson Paradigm

1. Raindrops are falling at an average rate of 20 drops per square inch per minute. What would be a reasonable distribution to use for the number of raindrops hitting a particular region measuring 5 inches^2 in t minutes? Why? Using your chosen distribution, compute the probability that the region has no rain drops in a given 3 second time interval.

2. Harvard Law School courses often have assigned seating to facilitate the "Socratic method." Suppose that there are 100 first year Harvard Law students, and each takes two courses: Torts and Contracts. Both are held in the same lecture hall (which has 100 seats), and the seating is uniformly random and independent for the two courses.

   (a) Find the probability that no one has the same seat for both courses (exactly; you should leave your answer as a sum).

   (b) Find a simple but accurate approximation to the probability that no one has the same seat for both courses.

   (c) Find a simple but accurate approximation to the probability that at least two students have the same seat for both courses.

3. Let X be a Pois(λ) random variable, where λ is fixed but unknown. Let $\theta = e^{-3\lambda}$, and suppose that we are interested in estimating θ based on the data. Since X is what we observe, our estimator is a function of X, call it $g(X)$. The *bias* of the estimator $g(X)$ is defined to be $E(g(X)) - \theta$, i.e., how far off the estimate is on average; the estimator is *unbiased* if its bias is 0.

   (a) For estimating λ, the r.v. X itself is an unbiased estimator. Compute the bias of the estimator $T = e^{-3X}$. Is it unbiased for estimating θ?

   (b) Show that $g(X) = (-2)^X$ is an unbiased estimator for θ. (In fact, it is the best unbiased estimator, in the sense of minimizing the average squared error.)

(c) Explain intuitively why $g(X)$ is a silly choice for estimating θ, despite (b), and show how to improve it by finding an estimator $h(X)$ for θ that is always at least as good as $g(X)$ and sometimes strictly better than $g(X)$. That is,

$$|h(X) - \theta| \leq |g(X) - \theta|,$$

with the inequality sometimes strict.

2 Seeking Sublime Symmetry

1. Let $Z \sim \mathcal{N}(0,1)$ and let S be a "random sign" independent of Z, i.e., S is 1 with probability 1/2 and -1 with probability 1/2. Show that $SZ \sim \mathcal{N}(0,1)$.

2. Explain why $P(X < Y) = P(Y < X)$ if X and Y are i.i.d. Does it follow that $P(X < Y) = 1/2$? Is it still always true that $P(X < Y) = P(Y < X)$ if X and Y have the same distribution but are not independent?

3. Explain why if $X \sim \text{Bin}(n,p)$, then $n - X \sim \text{Bin}(n, 1-p)$.

4. There are 100 passengers lined up to board an airplane with 100 seats (with each seat assigned to one of the passengers). The first passenger in line crazily decides to sit in a randomly chosen seat (with all seats equally likely). Each subsequent passenger takes his or her assigned seat if available, and otherwise sits in a random available seat. What is the probability that the last passenger in line gets to sit in his or her assigned seat? (This is another common interview problem, and a beautiful example of the power of symmetry.)

   Hint: call the seat assigned to the jth passenger in line "Seat j" (regardless of whether the airline calls it seat 23A or whatever). What are the possibilities for which seats are available to the last passenger in line, and what is the probability of each of these possibilities?

3 Continuous Distributions

1. Let $Y = e^X$, where $X \sim \mathcal{N}(\mu, \sigma^2)$. Then Y is said to have a *LogNormal* distribution; this distribution is of great importance in economics, finance, and elsewhere. Find the CDF and PDF of Y (the CDF should be in terms of Φ).

2. Let $U \sim \text{Unif}(0,1)$. Using U, construct a r.v. X whose PDF is $\lambda e^{-\lambda x}$ for $x > 0$ (and 0 otherwise), where $\lambda > 0$ is a constant. Then X is said to have a *Exponential* distribution; this distribution is of great importance in engineering, chemistry, survival analysis, and elsewhere.

3. Let $Z \sim \mathcal{N}(0,1)$. Find $E(\Phi(Z))$ *without* using LOTUS, where Φ is the CDF of Z.

4. A stick is broken into two pieces, at a uniformly random chosen break point. Find the CDF and average of the length of the longer piece.

4 LOTUS

1. For $X \sim \text{Pois}(\lambda)$, find $E(X!)$ (the average factorial of X), if it is finite.

2. Let $Z \sim \mathcal{N}(0,1)$. Find $E|Z|$.

3. Let $X \sim \text{Geom}(p)$ and let t be a constant. Find $E(e^{tX})$, as a function of t (this is known as the *moment generating function*; we will see later how this function is useful).

Stat 110 Strategic Practice 5 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1 Poisson Distribution and Poisson Paradigm

1. Raindrops are falling at an average rate of 20 drops per square inch per minute. What would be a reasonable distribution to use for the number of raindrops hitting a particular region measuring 5 inches^2 in t minutes? Why? Using your chosen distribution, compute the probability that the region has no rain drops in a given 3 second time interval.

   A reasonable choice of distribution is Poisson(λt), where $\lambda = 20 \cdot 5 = 100$ (the average number of raindrops per minute hitting the region). Assuming this distribution,

   $$P(\text{no raindrops in } 1/20 \text{ of a minute}) = e^{-100/20}(100/20)^0/0! = e^{-5}.$$

2. Harvard Law School courses often have assigned seating to facilitate the "Socratic method." Suppose that there are 100 first year Harvard Law students, and each takes two courses: Torts and Contracts. Both are held in the same lecture hall (which has 100 seats), and the seating is uniformly random and independent for the two courses.

   (a) Find the probability that no one has the same seat for both courses (exactly; you should leave your answer as a sum).

   Let N be the number of students in the same seat for both classes. The problem has the same structure as the *de Montmort matching problem* from lecture. Let E_j be the event that the j^{th} student sits in the same seat in both classes. Then

   $$P(N = 0) \;=\; 1 - P\left(\bigcup_{j=1}^{100} E_j\right)$$

   By symmetry, inclusion-exclusion gives

   $$P\left(\bigcup_{j=1}^{100} E_j\right) \;=\; \sum_{j=1}^{100} (-1)^{j-1} \binom{100}{j} P\left(\bigcap_{k=1}^{j} E_k\right)$$

The j-term intersection event represents j particular students sitting pat throughout the two lectures, which occurs with probability $(100-j)!/100!$. So

$$\begin{aligned} P\left(\bigcup_{j=1}^{100} E_j\right) &= \sum_{j=1}^{100}(-1)^{j-1}\binom{100}{j}\frac{(100-j)!}{100!} = \sum_{j=1}^{100}(-1)^{j-1}/i! \\ P(N=0) &= 1-\sum_{j=1}^{100}\frac{(-1)^{j-1}}{j!} = \sum_{j=0}^{100}\frac{(-1)^j}{j!}. \end{aligned}$$

(b) Find a simple but accurate approximation to the probability that no one has the same seat for both courses.

Define I_i to be the indicator for student i having the same seat in both courses, so that $N = \sum_{i=1}^{100} I_i$. Then $P(I_i = 1) = 1/100$, and the I_i are weakly dependent because

$$P((I_i=1)\cap(I_j=1)) = \left(\frac{1}{100}\right)\left(\frac{1}{99}\right) \approx \left(\frac{1}{100}\right)^2 = P(I_i=1)P(I_j=1)$$

So N is close to Pois(λ) in distribution, where $\lambda = E(N) = 100EI_1 = 1$. Thus,

$$P(N=0) \approx e^{-1}1^0/0! = e^{-1}.$$

This agrees with the result of (a), which we recognize as the Taylor series for e^x, evaluated at $x=-1$.

(c) Find a simple but accurate approximation to the probability that at least two students have the same seat for both courses.

Using a Poisson approximation, we have

$$P(N\geq 2) = 1-P(N=0)-P(N=1) \approx 1-e^{-1}-e^{-1} = 1-2e^{-1}.$$

3. Let X be a Pois(λ) random variable, where λ is fixed but unknown. Let $\theta = e^{-3\lambda}$, and suppose that we are interested in estimating θ based on the data. Since X is what we observe, our estimator is a function of X, call it $g(X)$. The *bias* of the estimator $g(X)$ is defined to be $E(g(X)) - \theta$, i.e., how far off the estimate is on average; the estimator is *unbiased* if its bias is 0.

(a) For estimating λ, the r.v. X itself is an unbiased estimator. Compute the bias of the estimator $T = e^{-3X}$. Is it unbiased for estimating θ?

The estimator is biased, with bias given by

$$E(e^{-3X}) - \theta = \sum_{k=0}^{\infty} e^{-3k} \frac{\lambda^k}{k!} e^{-\lambda} - e^{-3\lambda}$$

$$= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^{-3}\lambda)^k}{k!} - e^{-3\lambda}$$

$$= e^{-\lambda} e^{e^{-3}\lambda} - e^{-3\lambda}$$

$$= e^{-3\lambda}(e^{(2+e^{-3})\lambda} - 1) \neq 0.$$

(b) Show that $g(X) = (-2)^X$ is an unbiased estimator for θ. (In fact, it is the best unbiased estimator, in the sense of minimizing the average squared error.)

The estimator $g(X) = (-2)^X$ is unbiased since

$$E(-2)^X - \theta = \sum_{k=0}^{\infty} (-2)^k \frac{\lambda^k}{k!} e^{-\lambda} - e^{-2\lambda}$$

$$= e^{-\lambda} e^{-2\lambda} - e^{-3\lambda} = 0.$$

(c) Explain intuitively why $g(X)$ is a silly choice for estimating θ, despite (b), and show how to improve it by finding an estimator $h(X)$ for θ that is always at least as good as $g(X)$ and sometimes strictly better than $g(X)$. That is,

$$|h(X) - \theta| \leq |g(X) - \theta|,$$

with the inequality sometimes strict.

The estimator $g(X)$ is silly in the sense that it is sometimes negative, whereas $e^{-3\lambda}$ is positive. One simple way to get a better estimator is to modify $g(X)$ to make it nonnegative, by letting $h(X) = 0$ if $g(X) < 0$ and $h(X) = g(X)$ otherwise.

Better yet, note that $e^{-3\lambda}$ is between 0 and 1 since $\lambda > 0$, so letting $h(X) = 0$ if $g(X) < 0$ and $h(X) = 1$ if $g(X) > 0$ is clearly more sensible than using $g(X)$. It turns out that the answer to this part *must* be biased: it can be shown that $g(X)$ is the *only* unbiased estimator for θ, even though it would be silly to use $g(X)$ in practice.

2 Seeking Sublime Symmetry

1. Let $Z \sim \mathcal{N}(0,1)$ and let S be a "random sign" independent of Z, i.e., S is 1 with probability $1/2$ and -1 with probability $1/2$. Show that $SZ \sim \mathcal{N}(0,1)$.

   Condition on S to find the CDF of SZ:

   $$\begin{aligned} P(SZ \leq x) &= P(SZ \leq x|S=1)\frac{1}{2} + P(SZ \leq x|S=-1)\frac{1}{2} \\ &= P(Z \leq x)\frac{1}{2} + P(Z \geq -x)\frac{1}{2} \\ &= P(Z \leq x)\frac{1}{2} + P(Z \leq x)\frac{1}{2} \\ &= \Phi(x), \end{aligned}$$

   where the penultimate equality is by symmetry of the Normal.

2. Explain why $P(X < Y) = P(Y < X)$ if X and Y are i.i.d. Does it follow that $P(X < Y) = 1/2$? Is it still always true that $P(X < Y) = P(Y < X)$ if X and Y have the same distribution but are not independent?

   If X and Y are i.i.d., then $P(X < Y) = P(Y < X)$ by symmetry: we can interchange X and Y since both are the probability of one draw from the distribution being less than another, independent draw. In the discrete case, $P(X < Y) < 1/2$ since $P(X = Y) > 0$. In the continuous case, $P(X < Y) = 1/2$ since $P(X = Y) = 0$. If X and Y are not independent, then it is not necessarily true that $P(X < Y) = P(Y < X)$ since they may be structured in a way that tends to make X less than Y.

3. Explain why if $X \sim \text{Bin}(n,p)$, then $n - X \sim \text{Bin}(n, 1-p)$.

   This follows immediately from the story of the Binomial, by redefining "success" as "failure" and vice versa.

4. There are 100 passengers lined up to board an airplane with 100 seats (with each seat assigned to one of the passengers). The first passenger in line crazily decides to sit in a randomly chosen seat (with all seats equally likely). Each subsequent passenger takes his or her assigned seat if available, and otherwise sits in a random available seat. What is the probability that the last passenger in line gets to sit in his or her assigned seat? (This is another common interview problem, and a beautiful example of the power of symmetry.)

Hint: call the seat assigned to the jth passenger in line "Seat j" (regardless of whether the airline calls it seat 23A or whatever). What are the possibilities for which seats are available to the last passenger in line, and what is the probability of each of these possibilities?

The seat for the last passenger is either Seat 1 or Seat 100; for example, Seat 42 can't be available to the last passenger since the 42nd passenger in line would have sat there if possible. Seat 1 and Seat 100 are equally likely to be available to the last passenger, since the previous 99 passengers view these two seats symmetrically. So the probability that the last passenger gets Seat 100 is 1/2.

3 Continuous Distributions

1. Let $Y = e^X$, where $X \sim \mathcal{N}(\mu, \sigma^2)$. Then Y is said to have a *LogNormal* distribution; this distribution is of great importance in economics, finance, and elsewhere. Find the CDF and PDF of Y (the CDF should be in terms of Φ).

   The CDF of Y is

   $$P(Y \leq y) = P(X \leq \ln(y)) = P\left(\frac{X-\mu}{\sigma} \leq \frac{\ln(y)-\mu}{\sigma}\right) = \Phi\left(\frac{\ln(y)-\mu}{\sigma}\right),$$

   for $y > 0$ (and is 0 for $y \leq 0$). Taking the derivative using the chain rule, the PDF of Y is

   $$f(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-(\ln(y)-\mu)^2/(2\sigma^2)}.$$

2. Let $U \sim \text{Unif}(0,1)$. Using U, construct a r.v. X whose PDF is $\lambda e^{-\lambda x}$ for $x > 0$ (and 0 otherwise), where $\lambda > 0$ is a constant. Then X is said to have a *Exponential* distribution; this distribution is of great importance in engineering, chemistry, survival analysis, and elsewhere.

   The CDF of X is given by $F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$, for $x > 0$ (and 0 otherwise). The inverse function is $F^{-1}(u) = -\lambda^{-1}\ln(1-u)$. So by Universality of the Uniform, $-\lambda^{-1}\ln(1-U)$ has CDF F.

3. Let $Z \sim \mathcal{N}(0,1)$. Find $E(\Phi(Z))$ *without* using LOTUS, where Φ is the CDF of Z.

   By Universality of the Uniform, $F(X) \sim \text{Unif}(0,1)$ for any continuous random variable X with CDF F. Therefore, $E(\Phi(Z)) = 1/2$.

4. A stick is broken into two pieces, at a uniformly random chosen break point. Find the CDF and average of the length of the longer piece.

   We can assume the units are chosen so that the stick has length 1. Let L be the length of the longer piece, and let the break point be $U \sim \text{Unif}(0,1)$. For any $l \in [1/2, 1]$, observe that $L < l$ is equivalent to $U < l, 1-U < l$, which can be written as $1 - l < U < l$. We can thus obtain L's CDF as

$$F_L(l) \;=\; P(L < l) = P(1 - l < U < l) = 2l - 1,$$

   so $L \sim \text{Unif}(1/2, 1)$ and $E(L) = 3/4$.

4 LOTUS

1. For $X \sim \text{Pois}(\lambda)$, find $E(X!)$ (the average factorial of X), if it is finite.

   By LOTUS,

$$E(X!) = e^{-\lambda} \sum_{k=0}^{\infty} k! \frac{\lambda^k}{k!} = \frac{e^{-\lambda}}{1-\lambda},$$

   for $0 < \lambda < 1$ since this is a geometric series (and $E(X!)$ is infinite if $\lambda \geq 1$).

2. Let $Z \sim \mathcal{N}(0,1)$. Find $E|Z|$.

   Let $f(z) = \Phi'(z)$ be the PDF of the $\mathcal{N}(0,1)$ distribution. By LOTUS and the substitution $u = z^2$ and since $|z|f(z)$ is an even function, we have

$$E|Z| \;=\; \int_{-\infty}^{\infty} |z| f(z) dz = 2\int_0^{\infty} z f(z) dz = \int_0^{\infty} \frac{2ze^{-z^2/2}}{\sqrt{2\pi}} dz = \int_0^{\infty} \frac{e^{-u/2}}{\sqrt{2\pi}} du = \sqrt{2/\pi}.$$

3. Let $X \sim \text{Geom}(p)$ and let t be a constant. Find $E(e^{tX})$, as a function of t (this is known as the *moment generating function*; we will see later how this function is useful).

   Letting $q = 1 - p$, we have

$$E(e^{tX}) = p \sum_{k=0}^{\infty} e^{tk} q^k = p \sum_{k=0}^{\infty} (qe^t)^k = \frac{p}{1 - qe^t},$$

   for $qe^t < 1$ (while for $qe^t \geq 1$, the series diverges).

Stat 110 Homework 5, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. A group of $n \geq 4$ people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.).

(a) Let I_{ij} be the indicator r.v. of i and j having the same birthday (for $i < j$). Is I_{12} independent of I_{34}? Is I_{12} independent of I_{13}? Are the I_{ij}'s independent?

(b) Explain why the Poisson Paradigm is applicable here even for moderate n, and use it to get a good approximation to the probability of at least 1 match when $n = 23$.

(c) About how many people are needed so that there is a 50% chance (or better) that two either have the same birthday or are only 1 day apart? (Note that this is much harder than the birthday problem to do exactly, but the Poisson Paradigm makes it possible to get fairly accurate approximations quickly.)

2. Joe is waiting in continuous time for a book called *The Winds of Winter* to be released. Suppose that the waiting time T until news of the book's release is posted, measured in years relative to some starting point, has PDF $\frac{1}{5}e^{-t/5}$ for $t > 0$ (and 0 otherwise); this is known as the *Exponential distribution* with parameter $1/5$. The news of the book's release will be posted on a certain website.

Joe is not so obsessive as to check multiple times a day; instead, he checks the website *once* at the end of each day. Therefore, he observes the day on which the news was posted, rather than the exact time T. Let X be this measurement, where $X = 0$ means that the news was posted within the first day (after the starting point), $X = 1$ means it was posted on the second day, etc. (assume that there are 365 days in a year). Find the PMF of X. Is this a distribution we have studied?

3. Let U be a Uniform r.v. on the interval $(-1, 1)$ (be careful about minus signs).

(a) Compute $E(U), \text{Var}(U)$, and $E(U^4)$.

(b) Find the CDF and PDF of U^2. Is the distribution of U^2 Uniform on $(0, 1)$?

4. Let F be a CDF which is continuous and strictly increasing. The inverse function, F^{-1}, is known as the *quantile function*, and has many applications in statistics and econometrics. Find the area under the curve of the quantile function from 0 to 1, in terms of the mean μ of the distribution F. Hint: Universality.

5. Let $Z \sim \mathcal{N}(0, 1)$. A measuring device is used to observe Z, but the device can only handle positive values, and gives a reading of 0 if $Z \leq 0$; this is an example of *censored data*. So assume that $X = ZI_{Z>0}$ is observed rather than Z, where $I_{Z>0}$ is the indicator of $Z > 0$. Find $E(X)$ and $\text{Var}(X)$.

Stat 110 Homework 5 Solutions, Fall 2011

Prof. Joe Blitzstein (Department of Statistics, Harvard University)

1. A group of $n \geq 4$ people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.).

(a) Let I_{ij} be the indicator r.v. of i and j having the same birthday (for $i < j$). Is I_{12} independent of I_{34}? Is I_{12} independent of I_{13}? Are the I_{ij}'s independent?

The indicator I_{12} is independent of the indicator I_{34} since knowing the birthdays of persons 1 and 2 gives us no information about the birthdays of persons 3 and 4. Also, I_{12} is independent of I_{13} since even though both of these indicators involve person 1, knowing that persons 1 and 2 have the same birthday gives us no information about whether persons 1 and 3 have the same birthday (this relies on the assumption that the 365 days are equally likely). In general, the indicator r.v.s here are pairwise independent. But they are *not* independent since, for example, if person 1 has the same birthday as person 2 and person 1 has the same birthday as person 3, then persons 2 and 3 must have the same birthday.

(b) Explain why the Poisson Paradigm is applicable here even for moderate n, and use it to get a good approximation to the probability of at least 1 match when $n = 23$.

Let X be the number of birthday matches in a group of n people. There are at most $\binom{n}{2}$ matches, so the possible values of X are $0, 1, \ldots, \binom{n}{2}$. The Poisson paradigm says that if we have a large number n of events which are independent (or weakly dependent), each of which has a small probability of occurring, then how many of these events occur will be approximately Poisson.

In the birthday problem, even for moderate n, the number of experiments is much larger: $\binom{n}{2} \approx n^2/2$, the experiments are weakly dependent (in fact they are pairwise independent), and the probabilities for each event to occur are small (each of the events has probability $\frac{1}{365}$ of occurring). Using a Poisson approximation, X is approximately Pois($\frac{n(n-1)}{730}$) in distribution, which is Pois(0.693) for $n = 23$. Then

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) \\ &\approx 1 - e^{-0.693} \\ &\approx 0.500. \end{aligned}$$

(c) About how many people are needed so that there is a 50% chance (or better) that two either have the same birthday or are only 1 day apart? (Note that this is much harder than the birthday problem to do exactly, but the Poisson Paradigm makes it possible to get fairly accurate approximations quickly.)

Now define the r.v. I_{ij} to be the indicator of "persons i and j have the same birthday or are only one day apart". So $P(I_{ij} = 1) = \frac{3}{365}$. Again we can use the Poisson paradigm, letting X be the number of "near-matches". By Poisson approximation, X is approximately a Pois($\frac{3n(n-1)}{730}$). Now set the Poisson approximation for $P(X \geq 1)$ equal to 1/2, and solve for n:

$$\begin{aligned} P(X \geq 1) = 1 - P(X = 0) &\approx 1 - e^{\frac{-3n^2+3n}{730}} = 1/2 \\ \frac{-3n^2+3n}{730} &= -\log 2 \\ 3n^2 - 3n - 730 \log 2 &= 0. \end{aligned}$$

Solving this quadratic equation, we obtain $n \approx 13.5$. So with 14 people, there should be a slightly better than 50-50 chance of two having the same birthday or birthdays one day apart. (The exact answer for the number of people needed turns out to be $n = 14$, through much harder and more tedious calculations; so the Poisson approximation works very well here!)

2. Joe is waiting in continuous time for a book called *The Winds of Winter* to be released. Suppose that the waiting time T until news of the book's release is posted, measured in years relative to some starting point, has PDF $\frac{1}{5}e^{-t/5}$ for $t > 0$ (and 0 otherwise); this is known as the *Exponential distribution* with parameter 1/5. The news of the book's release will be posted on a certain website.

Joe is not so obsessive as to check multiple times a day; instead, he checks the website *once* at the end of each day. Therefore, he observes the day on which the news was posted, rather than the exact time T. Let X be this measurement, where $X = 0$ means that the news was posted within the first day (after the starting point), $X = 1$ means it was posted on the second day, etc. (assume that there are 365 days in a year). Find the PMF of X. Is this a distribution we have studied?

The event $X = k$ is the same as the event $k \leq 365T < k + 1$, i.e., $X = \lfloor 365T \rfloor$, where $\lfloor t \rfloor$ is the floor function of t (the greatest integer less than or equal to t). The CDF of T is $F_T(t) = 1 - e^{-t/5}$ for $t > 0$ (and 0 for $t \leq 0$). So

$$P(X = k) = P\left(\frac{k}{365} \leq T < \frac{k+1}{365}\right) = F_T\left(\frac{k+1}{365}\right) - F_T\left(\frac{k}{365}\right) = e^{-k/1825} - e^{-(k+1)/1825}.$$

This factors as $\left(e^{-1/1825}\right)^k (1 - e^{-1/1825})$, which shows that $X \sim \text{Geom}(1 - e^{-1/1825})$.

Miracle check: A Geometric distribution is plausible for a waiting time, and does take values $0, 1, 2, \ldots$. The parameter $p = 1 - e^{-1/1825} \approx 0.0005$ is very small, which

reflects both the fact that there are a lot of days in a year (so each day is unlikely) and the fact that the author is not known for the celerity of his writing.

3. Let U be a Uniform r.v. on the interval $(-1, 1)$ (be careful about minus signs).

(a) Compute $E(U)$, $\text{Var}(U)$, and $E(U^4)$.

We have $E(U) = 0$ since the distribution is symmetric about 0. By LOTUS,

$$E(U^2) = \frac{1}{2}\int_{-1}^{1} u^2 du = \frac{1}{3}.$$

So $\text{Var}(U) = E(U^2) - (EU)^2 = E(U^2) = \frac{1}{3}$. Again by LOTUS,

$$E(U^4) = \frac{1}{2}\int_{-1}^{1} u^4 du = \frac{1}{5}.$$

(b) Find the CDF and PDF of U^2. Is the distribution of U^2 Uniform on $(0, 1)$?

Let $G(t)$ be the CDF of U^2. Clearly $G(t) = 0$ for $t \leq 0$ and $G(t) = 1$ for $t \geq 1$, because $0 \leq U^2 \leq 1$. For $0 < t < 1$,

$$G(t) = P(U^2 \leq t) = P(-\sqrt{t} \leq U \leq \sqrt{t}) = \sqrt{t},$$

since the probability of U being in an interval in $(-1, 1)$ is proportional to its length. The PDF is $G'(t) = \frac{1}{2}t^{-1/2}$ for $0 < t < 1$ (and 0 otherwise). The distribution of U^2 is *not* Uniform on $(0, 1)$ as the PDF is not a constant on this interval (it is an example of a *Beta distribution*, which is another important distribution in statistics and will be discussed later).

4. Let F be a CDF which is continuous and strictly increasing. The inverse function, F^{-1}, is known as the *quantile function*, and has many applications in statistics and econometrics. Find the area under the curve of the quantile function from 0 to 1, in terms of the mean μ of the distribution F. Hint: Universality.

We want to find $\int_0^1 F^{-1}(u)du$. Let $U \sim \text{Unif}(0, 1)$ and $X = F^{-1}(U)$. By Universality of the Uniform, $X \sim F$. By LOTUS,

$$\int_0^1 F^{-1}(u)du = E(F^{-1}(U)) = E(X) = \mu.$$

Equivalently, make the substitution $u = F(x)$, so $du = f(x)dx$, where f is the PDF of the distribution with CDF F. Then the integral becomes

$$\int_{-\infty}^{\infty} F^{-1}(F(x))f(x)dx = \int_{-\infty}^{\infty} xf(x)dx = \mu.$$

Miracle check: For the simple case that F is the Unif$(0,1)$ CDF, which is $F(u) = u$ on $(0,1)$, we have $\int_0^1 F^{-1}(u)du = \int_0^1 udu = 1/2$, which is the mean of a Unif$(0,1)$.

5. Let $Z \sim \mathcal{N}(0,1)$. A measuring device is used to observe Z, but the device can only handle positive values, and gives a reading of 0 if $Z \leq 0$; this is an example of *censored data*. So assume that $X = ZI_{Z>0}$ is observed rather than Z, where $I_{Z>0}$ is the indicator of $Z > 0$. Find $E(X)$ and $\text{Var}(X)$.

By LOTUS,

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} I_{z>0} z e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z e^{-z^2/2} dz.$$

Letting $u = z^2/2$, we have

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-u} du = \frac{1}{\sqrt{2\pi}}.$$

To obtain the variance, note that

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{2},$$

since a $\mathcal{N}(0,1)$ r.v. has variance 1. Thus,

$$\text{Var}(X) = E(X^2) - (EX)^2 = \frac{1}{2} - \frac{1}{2\pi}.$$

Note that X is neither purely discrete nor purely continuous, since $X = 0$ with probability $1/2$ and $P(X = x) = 0$ for $x \neq 0$. So X has neither a PDF nor a PMF; but LOTUS still works, allowing us to work with the PDF of Z to study expected values of functions of Z.

Miracle check: The variance is positive, as it should be. It also makes sense that the variance is substantially less than 1 (which is the variance of Z), since we are reducing variability by making the r.v. 0 half the time, and making it nonnegative rather than roaming over the entire real line.