Ravi Dayabhai

Stat 110 Strategic Practice 5, Fall 2011

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

1 Poisson Distribution and Poisson Paradigm

P(0 raindrops in given 5 in² spot in 3 sec interval) = e^{-5}

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. Let $X \sim \text{Pois}\left(\frac{20 \cdot 5}{20}\right)$ since there are 20 3 second intervals per minute and the area is 5 in^2.

(Derangements!) ↓ easier to solve when treating as probabilities and unions of events

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). (SEE BACK)

   (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.) (SEE BACK)

③A. $E(e^{-3X}) - e^{-3\lambda} \overset{?}{=} 0$. By LOTUS, $E(e^{-3X}) = \sum_{x=0}^{\infty} e^{-3x}\left(\frac{e^{-\lambda}\lambda^x}{x!}\right) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^{-3})^x}{x!}$

For equality to hold, $\sum_{x=0}^{\infty}\frac{(\lambda e^{-3})^x}{x!} \overset{?}{=} e^{-2\lambda}$. $e^{(\lambda e^{-3})} \neq e^{-2\lambda}$, so it is a $\boxed{\text{biased}}$ estimator.

Bias $= E(e^{-3X}) - e^{-3\lambda} = e^{-\lambda}e^{\lambda e^{-3}} - e^{-3\lambda} = \boxed{e^{-3\lambda}\left(e^{(2+e^{-3})\lambda} - 1\right) \neq 0}$.

② A. Let A_k be the event that j^{th} student is in the same seat. By PRINC. of INCL-EXCLU.,

$$P(\text{no student in same seat}) = 1 - \sum_{k=1}^{100} (-1)^{k-1} \binom{100}{k} \frac{(100-k)!}{100!}$$

$$= 1 - \sum_{k=1}^{100} (-1)^{k-1} \cdot \frac{1}{k!} = 1 - \sum_{k=1}^{100} \frac{(-1)^{k-1}}{k!}$$

(bring 1 into sum) $\boxed{= \sum_{k=0}^{100} \frac{(-1)^k}{k!}}$

② B. By POISSON PARADIGM, let A_j be event that j^{th} student is in the same seat. I_j is indicator r.v. that j^{th} student is in the same seat. $X = \sum_{j=1}^{100} I_j$ counts A_j's occurring and $E(I_j)$ is $\frac{1}{100} = P(A_j)$, so $X \sim \text{Pois}(1)$ because $\lambda = \sum_{j=1}^{100} P(A_j) = 1$.

$$P(\text{no student in same seat}) = P(X=0) = \frac{e^{-1}(1)^0}{0!} = \boxed{\frac{1}{e} \approx 0.368}$$

② C. Using similar approach to ② B:

$$1 - P(0 \text{ or } 1 \text{ students have same seats}) = 1 - P(X \le 1) = 1 - \frac{e^{-1}(1)^0}{0!} - \frac{e^{-1}(1)^1}{1!}$$

$$= \boxed{1 - \frac{2}{e} \approx 0.264}$$

③ B. $g(X) = -2^X$; Show $E(g(X)) - \theta = 0 \iff E(-2^X) = \boxed{e^{-3\lambda}}$

By LOTUS, $E(g(X)) = \sum_{x=0}^{\infty} -2^x \left(\frac{e^{-\lambda} \lambda^x}{x!}\right) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(-2\lambda)^x}{x!} = \boxed{e^{-3\lambda}}$

③ C. It's silly to use $g(X) = (-2)^X$ to estimate $e^{-3\lambda}$ since for large X, $g(X)$ strays very far from $e^{-3\lambda}$ (bounded $(0,1]$); $e^{-3\lambda}$ is always positive (and close to zero for large X). $\boxed{h(X) = 1}$ is biased but a better estimator.

[Seeking Sublime Symmetry]

① $Z \sim N(0,1)$ and S is sign independent where $S \in \{1,-1\}$ and $P(S=1) = P(S=-1) = \frac{1}{2}$. $SZ \sim N(0,1)$ because when $S=1$, $Z \sim N(0,1)$. When $S=-1$, $-Z \sim N(0,1)$ by the fact that the PDF of Z is even function $\Rightarrow f(-x) = f(x)$. Both cases yield same outcome.

[Seeking sublime symmetry]

② $P(X<Y) = P(Y<X)$ if $X, Y \sim$ i.i.d because of symmetry, i.e., the ranking of $\{X>Y\}$ and $\{Y>X\}$ are equally likely. If cont. r.v.s, then $P(X<Y) = P(Y<X) = \frac{1}{2!}$. If disc. r.v. $P(X<Y) = P(Y<X) < \frac{1}{2}$ because of possibility of $X=Y$.

(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,

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

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 $n - X \sim \text{Bin}(n, 1-p)$. (SEE BELOW)

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, CDF of std $\mathcal{N}$ (SEE BACK)

③ $X \sim \text{Bin}(n,p)$ is a r.v. counting the number of successes in n trials. $n-X$ counts failures $\Leftrightarrow$ counting failures in n trials is $\text{Bin}(n, 1-p)$. Algebraically

$P(X=k) = \binom{n}{k} p^k q^{n-k}$ is PMF for X. PMF for $n-X$ is $P(n-X = n-k)$.

$P(n-X = n-k) = \binom{n}{n-k} p^{n-k} q^k \Rightarrow \text{Bin}(n,q) = \text{Bin}(n, 1-p)$

① Φ is CDF for standard normal. $X \sim N(\mu, \sigma^2)$, so

$$X = \mu + \sigma Z,\ Z \sim N(0,1) \Longrightarrow \frac{X-\mu}{\sigma} = Z \Longrightarrow \frac{X-\mu}{\sigma} \sim N(0,1).$$

So CDF of X is $F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$ and PDF is

$f(x) = \frac{d}{dx} F(x) = \varphi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}$ (by chain rule). Given $Y = e^X$,

$\ln Y = X \sim N(\mu, \sigma^2) \Rightarrow$ $\boxed{\text{CDF of } Y \text{ is } F(y) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right) \text{ and PDF is } f(y) = \frac{d}{dy}F(y) = \varphi\left(\frac{\ln y - \mu}{\sigma}\right)\frac{1}{\sigma}\frac{1}{y}}$, where Φ and φ

are $\varphi(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-t^2/2}$ and $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$

② $U \sim \text{Unif}(0,1)$. PDF of X is $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ (0 otherwise) where $\lambda > 0$ and constant. By UNIVERSALITY of UNIFORM, r.v.

$X = F^{-1}(U)$. $F(x) = \int_0^x \lambda e^{-\lambda t}\, dt = -e^{-\lambda t}\Big|_0^x = -e^{-\lambda x} + 1$. Therefore,

to get inverse, $x = 1 - e^{-\lambda F(x)} \Rightarrow \frac{\ln(1-x)}{-\lambda} = F^{-1}(x)$, so

$\boxed{\frac{\ln(1-U)}{-\lambda}}$ is CDF of X

③ $E(\Phi(Z)) = E(U)$ where $U \sim \text{Unif}(0,1)$ by UNIVERSALITY of UNIFORM.

$\boxed{E(U) = \frac{1}{2}}$ $= \int_0^1 u\, du = \frac{1}{2}u^2\Big|_0^1$. When you plug a r.v. into it's own CDF, you get $U \sim \text{Unif}(0,1)$ or the quantile for any value of r.v.

④ $0 \quad \times \quad \frac{m}{2} \quad m$ stick of length m. B is breakpoint and $B \sim \text{Unif}(0, m)$

L is r.v. modeling the longer piece. So $L = \begin{cases} B & \text{if } B \geq \frac{m}{2} \\ m - B & \text{if } B < \frac{m}{2} \end{cases}$ (support of B

is $(0, m)$), support of L is $(\frac{m}{2}, m)$ because it models always longer piece. Since B is uniform and L is always the longer piece, the PDF

of $L = \begin{cases} \frac{2}{m} & \text{for } \frac{m}{2} < l < m \\ 0 & \text{otherwise} \end{cases}$. $\boxed{\text{CDF of } L,\ F(l) = \begin{cases} \frac{2}{m} l & \text{for } \frac{m}{2} < l < m \\ 0 & \text{for } l < 0 \\ 1 & \text{for } l > m \end{cases}}$. By def.

$$E(L) = \int_{\frac{m}{2}}^{m} l \cdot \frac{2}{m}\, dl = \frac{1}{m} l^2 \Big|_{\frac{m}{2}}^{m} = m - \frac{m}{4} = \boxed{\frac{3}{4}m}$$

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. (SEE BACK OF PREV. PAGE)

3. Let $Z \sim \mathcal{N}(0,1)$. Find $E(\Phi(Z))$ *without* using LOTUS, where Φ is the CDF of Z. "plug r.v. back into its own CDF" (SEE BACK OF PREV. PAGE)

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 $E|Z|$. (SEE BELOW)

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).

[LOTUS]

① $E(X!)$ for $X \sim \text{Pois}(\lambda) \Rightarrow \sum_{x=0}^{\infty} \cancel{x!}\left(\frac{e^{-\lambda}\lambda^x}{\cancel{x!}}\right) = \boxed{e^{-\lambda}\sum_{x=0}^{\infty}\lambda^x \to \infty}$ for $\lambda > 1$.

$\boxed{\frac{e^{-\lambda}}{1-\lambda}}$ for $0 < \lambda < 1$ (since $\sum_{x=0}^{\infty}\lambda^x \to \frac{1}{1-\lambda}$ for $\lambda \in (0,1)$)

② By definition & LOTUS, $E(|Z|) = \int_{-\infty}^{\infty}|z|\varphi(z)\,dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|z|e^{-z^2/2}\,dz$.
By symmetry, $E|Z| = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} z e^{-z^2/2}\,dz$. Let $u = z^2/2 \Rightarrow du = z\,dz$.
Integration by substitution yields: $\frac{2}{\sqrt{2\pi}}\int_0^{\infty} z e^{-z^2/2}\,dz = \frac{2}{\sqrt{2\pi}}\int_0^{\infty} e^{-u}\,du$,
$-e^{-u}\Big|_0^{\infty} = 1$, so $\boxed{E(|Z|) = \frac{2}{\sqrt{2\pi}}}$.

③ By LOTUS, $E(e^{tX}) = \sum_x e^{tx}q^x p = \boxed{p\sum_x (qe^t)^x = \frac{p}{1-qe^t} \text{ (converges for } qe^t < 1)}$
$(q = 1-p)$

[HW5]

A. I_{12} and I_{34} are independent, but I_{12} and I_{13} are ~~not~~ independent. ~~by FUND. BRIDGE and DEF of IND.~~

$$\cancel{P(A_{12}, A_{13}) = \frac{1}{(365)^3} \neq P(A_{12})P(A_{13}) = \frac{1}{(365)^4}} \Rightarrow I_{ij}\text{'s}$$

are ~~NOT~~ independent ~~because every finite subset of I_{ij}s must be independent~~ Pairwise ind. but not independent.

B. The Poisson Paradigm is still applicable because the I_{ij}'s are only weakly dependent and there are a large number of events $\binom{n}{2}$ each with a small probability of occurring. (By symmetry, each $E(I_{ij}) = p_{ij} = \frac{1}{365}$). For $n = 23$, by Poisson approximation, $\sum p_{ij} = \binom{23}{2}\frac{1}{365} = \frac{253}{365} = \lambda$. $X \sim \text{Pois}(\lambda)$, so

$P(X=0) = e^{-\lambda} \Rightarrow P(X \geq 1) = 1 - e^{-\lambda} \approx .50$

C. From above $\lambda = \binom{m}{2}\frac{3}{365} \Rightarrow X \sim \text{Pois}(\lambda) \Rightarrow P(X \geq 1) = 1 - e^{-\lambda}$

To get $P(X \geq 1) > 50\%$, $m \geq 14$.

(2) T is just continuous time waiting for a success (i.e., book release pt), whereas X is discrete measure of time in days (T in years, assume 365 days per year). Since the PDF of T is $f(t) = \frac{1}{5}e^{-\frac{t}{5}}$, if we integrate in one day "chunks", we will get PMF of X: $P(X=k) = \int_{\frac{k}{365}}^{\frac{k+1}{365}} \frac{1}{5}e^{-\frac{t}{5}}\,dt = -e^{-u}\Big|_{\frac{k}{5\cdot 365}}^{\frac{k+1}{5\cdot 365}}$

which reduces to $\boxed{e^{-k/1825} - e^{-(k+1)/1825}} = \underbrace{e^{-k/1825}\left(1 - e^{-1/1825}\right)}_{\boxed{X \sim \text{Geom}\left(1 - e^{-1/1825}\right)}}$.

3A) $\boxed{E(U) \text{ is } 0}$ (by SYMMETRY).

By DEF. OF VARIANCE

$\text{Var}(U) = E([U - E(U)]^2) = E(U^2 - 2UE(U) + E(U)^2) = E(U^2) - E(UE(U)) - E(UE(U)) + E(U$

$= E(U^2) - E(U)^2 = E(U^2)$ from $E(U) = 0$. By LOTUS, $E(U^2)$ is

$\int_{-1}^{1} u^2 f(u)\,du$ where $f(u)$ is PDF of $\text{Unif}(-1,1)$. $f(u) = \frac{1}{2}$ on $u \in (-1,1)$

$\frac{1}{2}\int_{-1}^{1} u^2\,du = \frac{1}{2}\left(\frac{1}{3}u^3\Big|_{-1}^{1}\right) = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{3}\right) = \boxed{\frac{1}{3}}$. $E(U^4) = \int_{-1}^{1} u^4 f(u)\,du = \frac{1}{2}\left(\frac{1}{5}u^5\Big|_{-1}^{1}\right)$

So $E(U^4) = \frac{1}{2}\left(\frac{1}{5} + \frac{1}{5}\right) = \boxed{\frac{1}{5}}$

Stat 110 Homework 5, Fall 2011

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

COMPLETELY WRONG WORK

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

(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 $E(U^4)$. (SEE BACK OF PREVIOUS PAGE)

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

ALMOST? ?

NEED BETTER GRASP OF OBJECTS

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. (SEE BACK) $\boxed{\mu}$

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 $\text{Var}(X)$. (SEE BACK)

6. Let $X = U^2$ for $u \in [-1, 1]$

5. $X = Z I_{Z>0}$, so $E(X) = \int_{-\infty}^{\infty} x f(x)\,dx \overset{\text{(by LOTUS)}}{=} \cancel{\int_{-\infty}^{0} x f(x)\,dx} + \int_{0}^{\infty} x f(x)\,dx =$

$$\frac{1}{\sqrt{2\pi}} \int_0^{\infty} z e^{-z^2/2}\,dz = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-u}\,du = \left(\frac{1}{\sqrt{2\pi}}\right) -e^{-u}\Big|_0^{\infty} = \frac{1}{\sqrt{2\pi}}(0+1) = \boxed{\frac{1}{\sqrt{2\pi}}}$$

Let $u = z^2/2$
$du = z\,dz$

$\mathrm{Var}(X) = E(X^2) - E(X)^2 = E(X^2) - \frac{1}{2\pi}$ (from above). $E(X^2) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} z^2 e^{-z^2/2}\,dz$ since

$x = zi \Rightarrow x^2 = z^2 i^2 = z^2 i$ which is still 0 for $z<0$. Integration by parts yields $E(X^2) = \frac{1}{2}$, so $\boxed{\mathrm{Var}(X) = \frac{1}{2} - \frac{1}{2\pi}}$