Statistics and Data Analysis
Lecture 2

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2016-11-15
Last edited 17:56:35 2016-12-28

1 Review

$$E_P(y) = \sum yP(Y = y)$$

2 Important Distributions

2.1 Geometric Distribution

Number of Bernoulli trials needed for the first success:

- $P(Y = 1) = p$
- $P(Y = 2) = (1 - p)p$
- $P(Y = k) = (1 - p)^{k-1}p$

Proof of correctness We want to show that this must happen at some point, eventually, so we want to show that the sum of all the probabilities is 1.

$$q = 1 - p$$

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

Setting $y^* = y - 1$ and noting that $y = 1, 2, ...,$ then $y^* = 0, 1, ...,$ and therefore:

$$\sum_{y=1}^{\infty} p(y) = p\sum_{y=1}^{\infty} q^{y^*} = p[\frac{1}{1-q}] = \frac{p}{p} = 1$$

2.1.1 Geometric Expectation

$$\sum_{i=1}^{\infty} iq^{i-1}p = p\sum_{i=1}^{\infty} iq^{i-1}$$

$$\sum_{i=1}^{\infty} q^i = \frac{1}{1-q}(takederivative)$$

$$\sum_{i=1}^{\infty} iq^{i-1} = \frac{1}{(1-q)^2}$$

And if $Y \sim GEO(p)$:

$$E(Y) = \frac{p}{(1-q)^2} = \frac{1}{2}$$

2.1.2 Example: Click through rate

The click through rate (CTR) of an ad is 0.03. We are investigating independent sessions.

What is the prob that the first CT occurs at the 5th entry? At the 56th?

Let G be a geometric random variable with p=0.03. We are interested in P(G=5) and P(G=56).

$$P(G = 5) = 0.97^4 \cdot 0.03 \approx 0.026$$

$$P(G \geq 10) = 0.97^9 \approx 0.76$$

In 7 out of 30 one hour periods, a CT occurred on or before the 9th entry. Does this make sense? Yes!

2.2 Negative Binomial Distribution

In successive Bernoulli trials, what is the distribution of the number of trials needed until the r^th success.

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

$$Y = \sum_{i=1}^{r} X_i$$

Y NegB(r,p)

$$E_P(y) = \sum yP(Y = y)$$

$$E(Y) = \sum_{i=1}^{r} E(X_i asGeo(P))$$

Fill in variance from slide 6

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

2.2.1 Example: Basketball

Players shoot simultaneously.

Player A: p $< 1/2$ Shoots until r success N(A) is the attempt when that happened

$N(A) \quad NegB(r, p)$

Player B: $p = m \cdot p$ for some integer $1 < m$ such that $m \cdot p < 1$ shoots until $m \cdot r$ success $N(B)$ is the attempt when that happened

$N(B) \quad NegB(mr, mp)$

Questions $E(N(A)) = E(N(B))$ They will more or less wait the same amount...

$V(A) > V(B)$

Placing bets on $N(A)$ or $N(B)$? Which is better? Variance?

Placing a bet on $N(A) - N(B)$ is not worthwhile due to spread/commission.

Placing a bet on $N(A)^2 - N(B)^2$:

$$V(N(A)) = E(N(A)^2) - (E(N(A)))^2$$

$$V(N(B)) = E(N(B)^2) - (E(N(B)))^2$$

$$E(N(A)^2) > E(N(B)^2)$$

which is worthwhile and profitable (to within kelley)

2.3 Poisson Distribution

One of the most useful distributions.

- Arrivals of customers to a store within one hour
- Number of flaws in a roll of fabric of a given length
- Number of visitors to a website within an hour
- Number of incoming calls to a service center in an hour
- Defects per meter of electrical cable

A limit of binomials with an increasing n and a fixed mean.

Repeated trials with decreasing chance of success:

$$X_n \sim Binom(n, \frac{\lambda}{n})$$

$$P(X_n = k) = \binom{n}{k}\binom{\lambda}{n}^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

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

$$P(X_n = k) = X_n \sim Poisson(\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$$

2.3.1 Example: Website

Website gets Poisson(0.5) visitors per second

What is the prob of no visits in any given second window? E(-0.5)

$$P(X = 0) = e^{-0.5}\frac{1}{1}$$

What is the prob of no visits in a 10 second stretch?

Poisson(5) gives E(-5)

NOTE: a sum of poissons is poisson

X_1 = Number of visits in 1 sec $\sim Poi(0.5)$ X_10 = Number of visits in 10 sec $\sim Poi(5)$

$$P(X_1 = 0) = e^{-0.5}\frac{\lambda^0}{0!} = e^{-0.5}$$

$$P(X_10 = 0) = E(0.5)^{10} = e^{-5}$$

2.3.2 Correctness

Expectation

$$f(y) = \frac{e^{-\lambda}\lambda^y}{y!}$$

$$E(Y) = \sum_{y=0}^{\infty}\left[\frac{e^{-\lambda}\lambda^y}{y!}\right] = \sum_{y=1}^{\infty} y... = \lambda$$

$$E(Y^2) = \lambda^2 + \lambda$$

$$V(Y) = \lambda$$

Variance

Coefficient of Variation $CV(X) = \delta(X)/E(X) = \frac{1}{\sqrt{\lambda}}$

2.3.3 Example: Droplet encapsulation

microfluidics by controlling the size of the droplets, we can control the number of particles.

We distribute the number of particles (N) and the number of droplets (M) uniformly and independently.

The number of particles per droplet is X:

$X \sim Binom(N, 1/M)$. Set $M = N/\lambda$. If N is large enough then we can use the above Poisson approximation and say that $X \sim Poisson(\lambda)$

The carrier fluid is expensive, and the particles don't work if there are more than 3 per droplet. As such, we want to optimize the number of useful droplets. TOOD: fill in from slide 15

For $\lambda = 1$, we get:

$P(X = 0) = e^{-1}\frac{1^0}{0!} = e^{-1}$ $P(X = 1) = e^{-1}\frac{1^1}{1!} = e^{-1}$ $P(X = 2) = e^{-1}\frac{1^2}{2!} = 0.5e^{-1}$ $P(X \geq 3) = ...$

This gives us $1.5e^{-1}$

Optimize:

$$P(useful) = e^{-\lambda}\left(\lambda + \frac{\lambda^2}{2}\right)$$

$$\frac{d}{d\lambda}P(useful) = e^{-\lambda}\left(1 - \frac{\lambda^2}{2}\right)$$

2.3.4 Example: Delivery Room staffing

NOTE: this appears in HW1

A maternity ward needs to make staffing decisions (midnight to 8am shifts). They want to know how many births take place at night.

They had 3300 deliveries last year, and if it was uniform, there would have been 1100 deliveries during the night shift.

The average number of deliveries per night is 3.

Let's say that the number of babies is $Poisson(3)$.

$$P(0) = 3^0 e^{-3}/0! \approx 0.05$$

$$P(2) = 3^2 e^{-3}/2! \approx 0.23$$

How many days per year expect more than 5? (68)...

$$P(X \leq 4) = e^{-3}(1 + 3 + 3^2/2 + ...)$$

CDF for poisson distribution?

What is the most babies we expect to see? Find k such that $P(k) \geq 1/365$. Answer is 9.

Note: $Binom(365, 1/365) \sim Poisson(1)$

Mgmt decided to have sufficient staff to reduce expected to be called from home to under 10. How many teams will be present in every night shift?

find minimal k such that $CDF(k) \geq 1 - 10/365 = 0.9635$. Answer is 6.

3 Statistical Independence

Value of one independent random variable does **not** affect others and is **not** affected by others.

Two events $A, B \in \Omega$ are said to be independent if the occurrence of one doesn't affect the other.

$P(A|B) = P(A)$, where $P(A|B) = P(A \cap B) \setminus P(B)$ $P(A \cap B) = P(A) \cdot P(B)$

Example: Dice fill in from slide 20

Example: Playing Cards fill in from slide 21

3.1 Independent Random Variables

$P(X = x \text{ and } Y = y) = P(X = x) \cdot P(Y = y)$

Note: $E(XY) = E(X) \cdot E(Y)$ Practice: prove this.... (sums...), and opposite?

4 Linearity of Expected Values

E(X+Y)=E(X)+E(Y)

True for all variables.

NOTE: this will appear in HW1

4.1 Example: Coupon Collection

Website is collecting a sample of users from each of 10 countries. Users are uniformly and independently distributed.

Let X_1 be the number of visits until the first country is in ($X_1 = 1$) Let X_i be the number of visits until the ith country is in, after the $i - 1$th country is in.

$$T = X_1 + ... + X_{10}$$

$$E(T) = E(X_1) + ... + E(X_{10})$$

$X_i \sim Geom((10 - i + 1)/10)$

General m and unequal probabilities is much more complicated.

Harmonic Sum

$$\sum_{k=1}^{n} \frac{1}{k} \sim \log n$$

$$E(T) = n \cdot H_n \sim_{n \to \infty} n \log n$$

$$E(T_m) = n \log n + (m - 1)n \log \log n + O(n)$$

5 Variances and independences

NOTE: Will upload Excel sample

Empirical variance from a sample is $\sigma^2 = \frac{1}{n} \sum_{k=0}^{n} (x_i - \mu)^2$

Note that $Var(L) + VAR(D) \neq Var(L + D)$....

If they are independent, then VAR(X+Y)=VAR(X)+VAR(Y)?

6 Sums of Independent Random Variables

$$P(Z = z) = \sum_{i=-\infty}^{\infty} P(X = i)P(Y = z - i)$$

$$h(z) = \int_{-\infty}^{\infty} f(t)g(z - t)dt$$

6.1 Sum of Poissons is Poissons

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

$$P(Y = k) = e^{-\mu}\frac{\mu^k}{k!}$$

X and Y are independent. Let $Z = X + Y$

$$P(Z = k) = \sum_{i=-\infty}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} + e^{-\mu}\frac{\mu^{k-i}}{(k-i)!}$$

When the denominator is negative, the term is 0.

$$P(Z = k) = e^{-(\lambda+\mu)} \cdot \frac{1}{k!}\sum_{i=0}^{k}\binom{k}{i}\lambda^i\mu^{k-i}$$

7 Histogram visualizations

Histograms: bin data into ranges

It's important to pick good bin sizes.

8 Convulation Kernel

slide 28. will be covered next week as well

9 Covariances

On slide 30

10 Homework Assignment

Due in 2 weeks. Python/Matlab/Octave will work. fmin family for optimizations?

input is two vectors: values and the probability of each value

process is for loop and evaluate code.