Lecture 5 Binomial and Poisson Distributions

10/14/2020

Derivations of Binomial properties

$$X \sim Binomial(n, p)$$

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for} \quad x = 0,1,\ldots,n$$

We can write

$$X = \sum_{i=1}^{n} Y_i \quad \text{where} \quad Y_i \underset{iid}{\sim} Bernoulli(p)$$

Remark:

- "iid": identically and independently distributed

  - "identically distributed": Y_i and Y_j $(i \neq j)$ have the same distributoion.
  - "identically distributed": Y_i and Y_j $(i \neq j)$ are independent such that

    $$\Pr(Y_i = y_i, Y_j = y_j) = \Pr(Y_i = y_i)\Pr(Y_j = y_j)$$

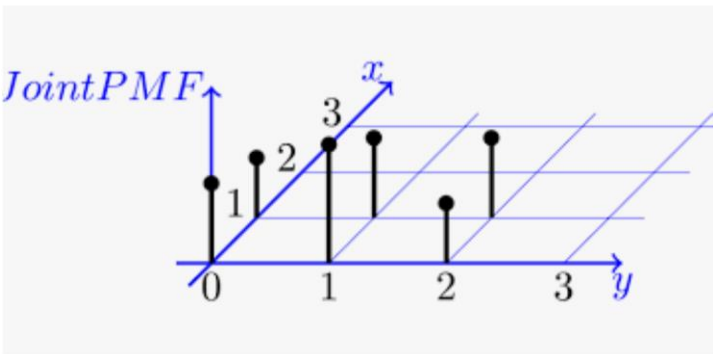
  - $\Pr(Y_i = y_i, Y_j = y_j)$ is the "joint probability function" of (Y_i, Y_j)

Properties of two random variables (will NOT be on the exam)

- 聯合機率函數: "Joint" probability function for (X, Y)

  $\Pr(X = x, Y = y)$ satisfies

  $$\sum_{\forall x} \sum_{\forall y} \Pr(X = x, Y = y) = 1$$



Joint Probability Mass Function ...

- Marginal probability function (邊際機率函數)

  $$\Pr(X=x)=\sum_{\forall y}\Pr(X=x,Y=y)$$

  $$\Pr(Y=y)=\sum_{\forall x}\Pr(X=x,Y=y)$$

  $E(X)=\mu_x$, $E(Y)=\mu_Y$, $Var(X)=\sigma_X$, $Var(Y)=\sigma_Y$

- X,Y are <u>independent</u> if and only if

  $\Pr(X=x,Y=y)=\Pr(X=x)\Pr(Y=y)$ for <u>all</u> (x,y)

Note:

- We've learned how to check the independence between two sets:

  $\Pr(A\cap B)=\Pr(A)\Pr(B)$

- For checking the independence between two random variables , if there exists (x,y) such that $\Pr(X=x,Y=y)\neq\Pr(X=x)\Pr(Y=y)$, then (X,Y) can not be independent

  → To prove dependence is <u>easier</u> than to prove independence.

Q: How to measure the degree of association if (X,Y) are dependent

Covariance (共變數)

$$Cov(X,Y)=E[(X-\mu_X)(Y-\mu_Y)]$$

$$=\sum_{\forall x}\sum_{\forall y}(x-\mu_x)(y-\mu_Y)\Pr(X=x,Y=y)=E(XY)-\mu_X\mu_Y$$

Correlation coefficient (相關係數)

$$Corr(X,Y)=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}\in[-1,1]$$

Remarks:

- If X,Y are independent , then

  $Cov(X,Y)=E(XY)-E(X)E(Y)=\mu_X\mu_Y-\mu_X\mu_Y=0$;

  but not vice versa. (獨立一定無相關; 無相關不一定獨立)

Review of Useful Properties

- $E(c) = c$, $Var(c) = 0$ where c is a constant
- $E(aX + bY) = aE(X) + bE(Y)$ (a,b are constant)
- $Var(aX + bY) = a^2 Var(X) + 2abCov(X,Y) + b^2 Var(Y)$
    - If (X,Y) are independent, then $Cov(X,Y) = 0$.
- "Independence" is a much stronger condition than "zero correlation".

Derivations for $X \sim Binomial(n, p)$

$$X = \sum_{i=1}^{n} Y_i \quad \text{where} \quad Y_i \underset{iid}{\sim} Bernoulli(p)$$

We know that

- $E(Y_i) = p$
- $Var(Y_i) = p(1-p)$
- $Cov(Y_i, Y_j) = 0$ (because of independence)

$$E(X) = E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} p = np = \mu$$

$$Var(X) = Var(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} Var(Y_i) + \sum_{i \neq j} Cov(Y_i, Y_j) = np(1-p) + 0$$

n^2 terms $\quad$ n terms $\quad$ $n(n-1)$ terms

Example (Lecture 4)

XAMPLE 12.1 Blood types

Genetics says that children receive genes from their parents independently. Each child of a particular pair of parents has probability 0.25 of having type O blood. If these parents have 5 children, the number who have type O blood is the count X of successes in 5 independent trials with probability 0.25 of a success on each trial. So X has the binomial distribution with $n = 5$ and $p = 0.25$.

Pr(a child has blood type O) = 1/4

Pr(None of the children have blood type O) $= \left(\frac{3}{4}\right)^5$

Pr(Exactly one child has blood type O) = $\binom{5}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^4$

Pr(<u>At least </u>two children have blood type O) = $1-\mathrm{P}(X=0)-P(X=1)$

$$\mathrm{P}(X=0)=\binom{5}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^5;\quad \mathrm{P}(X=1)=\binom{5}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^4$$

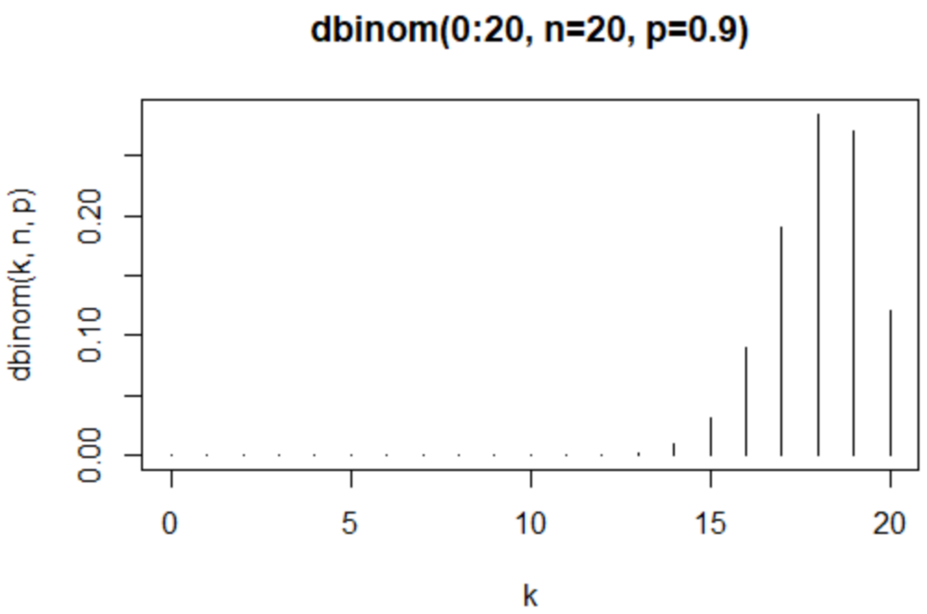
Pr(<u>At most</u> two children have blood type O) = $\mathrm{P}(X=0)+P(X=1)+P(X=2)$

$$\mathrm{P}(X=2)=\binom{5}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^3$$

Example: Germination of Seeds

- Germination rate = 90% ← claimed by the store manager
- If you bought n = 20 seeds, what is the expected number of seeds to germinate?
- If you planted the 20 seeds and there were 15 seeds which eventually germinate, can you say that the store manager lied to you?
  - ■ A naïve method:

    15/20 = 75% < < 90% (沒有學過統計的人會這麼算)




```
> n=20; p=0.9; k=seq(0,n)
> plot(k, dbinom(k,n,p), type='h', main='dbinom(0:20, n=20, p=0.9)', x
lab='k')
```

Solution:

Let X = # of seeds that germinate → calculate $\Pr(X \le 15)$

→ this value is related to "p-value" which provides evidence to check whether $p = 0.9$ is correct.

Solution

Let X = # of seeds that germinate ~ Binomial(n=20,p)

Based on the manager's argument: $p = 0.9$

E(X) = 20*0.9 = 18

$$\Pr(X <= 15) = \sum_{x=0}^{x=15} \frac{20!}{x!(20-x)!} 0.9^x (0.1)^{20-x} \quad \rightarrow \text{sum of 16 terms}$$

$$= 1 - \sum_{x=16}^{x=20} \frac{20!}{x!(20-x)!} 0.9^x (0.1)^{20-x} \quad \rightarrow \text{sum of 5 terms}$$

$$= 0.0432$$

Interpretation: Since $\Pr(X \le 15) = 0.0432$ is very low (compared with 0.05, the value of significance which will be taught later), $p = 0.9$ is not reasonable.

Remarks

1. You can use the attached table to find the probability for selected values of (n, p).

2. For any (n, p), use the online calculator

https://stattrek.com/online-calculator/binomial.aspx

- Useful techniques

X = number of successes $\sim Binomial(20, p = 0.9)$

$Y = n - X$ = number of failure $\sim Binomial(20, 1 - p = q = 0.1)$

$X \le 15$ = at most 15 seeds germinate

$Y \ge 5$ = at least 5 seeds do not germinate

$\{X \le 15\} = \{Y \ge 5\}$ → the two events are equivalent

$$\Pr(X \le 15) = \sum_{x=0}^{15} \binom{20}{x} \cdot (0.9)^x \cdot (0.1)^{20-x} \quad \rightarrow \text{要算 16 個項的和}$$

$$
\begin{aligned}
\Pr(X \le 15) &= \Pr(-X \ge -15) \\
&= \Pr(n - X \ge 20 - 15) \\
&= \Pr(Y \ge 5) \\
&= 1 - \Pr(Y \le 4)
\end{aligned}
$$

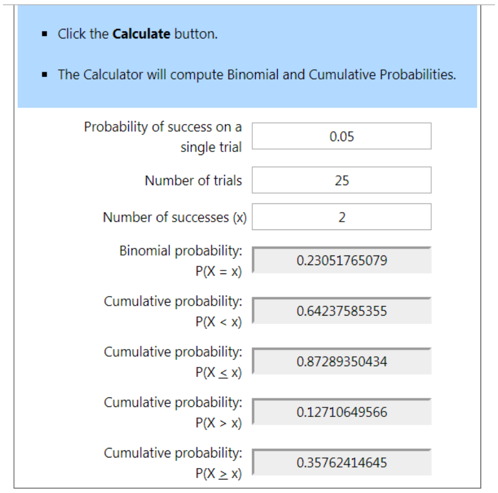
$$\Pr(Y \le 4) = \sum_{y=0}^{y=4} \binom{20}{y} (0.1)^y (0.9)^{20-y} \rightarrow \text{只要算 5 個項的和}$$

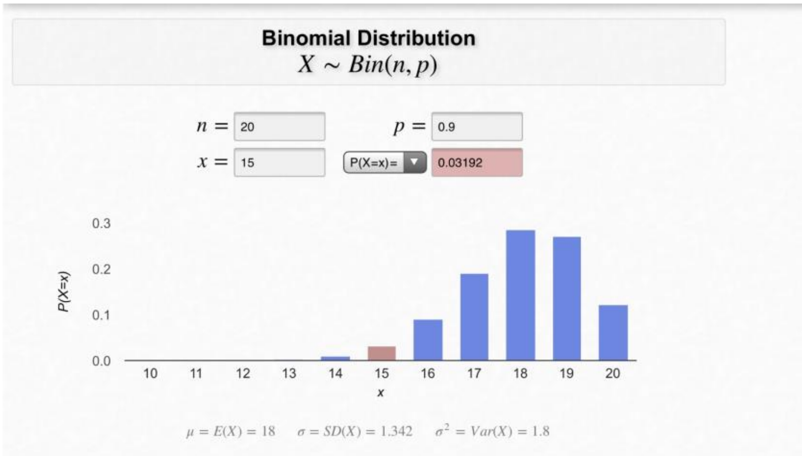
Example: When circuit boards used in the manufacture of compact disc players are tested, the long-run percentage of defectives is p = 5%.

Let X = the number of defective boards in a random sample of size n = 25.

1. Determine $\Pr(X \le 2)$, $\Pr(X \ge 4)$ and $\Pr(1 \le X \le 4)$

$$\Pr(X \le 2) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) = \sum_{x=0}^{2} \binom{25}{x} \cdot (0.05)^x \cdot (0.95)^{20-x}$$





Find the values of $E(X)$, $Var(X)$ and $\sqrt{Var(X)}$

$$E(X) = np = 25 \times 0.05 = 1.25$$

$$Var(X) = npq = 25 \times 0.05 \times 0.95 = 1.1875$$

$$\sqrt{Var(X)} = \sqrt{1.1875} = 1.09$$

Example:

Suppose that only 20% of all drivers come to a complete stop at an intersection having flashing red lights in all directions when no other cars are visible.

What is the probability that, of 20 randomly chosen drivers coming to an intersection under these conditions:

1. At most five will come to a complete stop?
2. Exactly five will come to a complete stop?
3. At least five will come to a complete stop?
4. How many of the next 20 drivers do you expect to come to a complete stop?

Sol: $n = 20, p = 0.2$

X = number of cars come to a complete stop among 20 chosen cars

$X \sim Binomial(n = 20, p = 0.2)$

1. Pr(at most 5) = $\Pr(X \le 5)$

2. Pr(exactly 5) = $\Pr(X = 5)$

3. Pr(at least 5) = $\Pr(X \ge 5)$

4. $E(X) = np = 20 \times 0.2 = 4$

II. Poisson random variable (波耳松分配)

Consider a <u>stochastic process</u> which records the occurrence of random events

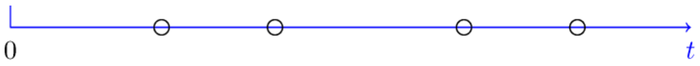
Examples of random events:

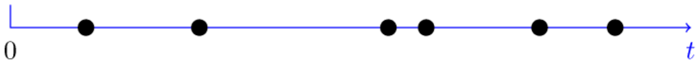
car accidents, earthquakes, receiving a phone call, disease occurrence

Poisson process: satisfies the following conditions

- The probability that *exactly one event* occurs in a given interval of length h is proportional to the length, say λh, where λ = the average number of events in a unit of time (space).
- When h is small, the probability that <u>two or more</u> events occur = 0.
- The numbers of events in two *disjoint* intervals are independent.

Plots of Poisson process (time)





Plots of Poisson process (space)


在零膨脹資料結構下進行卜瓦松迴歸模型與遞延分配模型之模...
ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/login?o=dnclcdr&s=id...
... 藉由1998-2008年高雄地區登革熱病例案件數及每日溫度資料進行模擬，在對立假設下將登革熱病例資料與溫度的關係由零膨脹卜瓦松分布(Zero-inflated Poisson ...

南台灣登革熱族群傳輸動態及感染風險評估 - 臺灣博碩士論文...
ndltd.ncl.edu.tw/handle/44847339654105523541
2012年6月21日 - 本研究亦發展以統計指標為基礎之卜瓦松迴歸模式(Poisson regression model)評估影響登革熱每月發生率之潛在因子，最後並以曲棍球桿 ...

國立臺灣大學生工系生物模擬與控制研究室: 研究主題-環境毒...
homepage.ntu.edu.tw/~cmliao/research_7.htm
本研究以台灣2001 – 2008年登革熱盛行地區之流行病學調查資料，利用波以松迴歸分析(Poisson regression analysis)探討登革熱發生率與氣候因子(溫度、相對濕度 ...

[PDF] 2051KB - 行政院環境保護署
www.epa.gov.tw/cpDownloadCtl.asp?id=17104
1994年12月31日 - 的一致性則沒有氣象因子對於登革熱流行來的明顯與明確，後續進一步大量 ...... 進一步利用Poisson regression with Generalized Additive Model ...

Q: How do we judge whether the Poisson process is appropriate?

Example 1: the number of car accidents in Kuan-Fu Road <u>within a month</u>

Yes → if probabilities in two disjoint time intervals are independent

No → if the situation changes (more policemen, change of regulations)

after a big accident within the same time interval

Example 2: the number of phone calls received by an operator <u>within 4 working hours</u>

Yes → if the patterns of phone calls are roughly the same

And also the chance of getting two calls at the same time is rare

Example3: the number of flaws in a book of <u>30 pages</u>

Usually yes, unless some pages require special printing techniques.

Example 4: the number of customers coming to a store <u>within 4 working hours</u>

Usually no: since more than one person may come to the shop together.

The patterns may not be the same within the 4-hour interval.

Remarks:

- We may restrict the time or space which provides a *homogeneous* condition.
- More examples:
  - the location of users in a wireless network;
  - the requests for individual documents on a web server;
  - photons landing on a photodiode;
  - The number of deaths attributed to typhoid fever (傷寒).
- The Poisson process is connected in various interesting ways to a number of special distributions, including the *Poisson, exponential, Gamma, Beta, uniform, binomial, and the multinomial*. These embracing connections and wide applications make the Poisson process a very special topic in probability.

Poisson Random Variables

From a Poisson process, pick an interval or space.

Define the number of events occurring in the interval or space.

Definition 1: Poisson random variable

Let X be the number of events in the time interval $[0,t]$ with length t.

Let λ = the average number of events in a unit of time.

Then $E(X) = \lambda t = \mu$.

Definition 1: Poisson random variable

X ~ Poisson($\mu = \lambda t$) with the probability function:

$$\Pr(X = x) = \frac{\exp(-\lambda t)(\lambda t)^x}{x!} = \frac{\exp(-\mu)(\mu)^x}{x!} \quad (x = 0, 1, 2, \ldots)$$ → no upper bound

$E(X) = \lambda t$ & $Var(X) = \lambda t$ (Poisson: the mean and variance are the same)

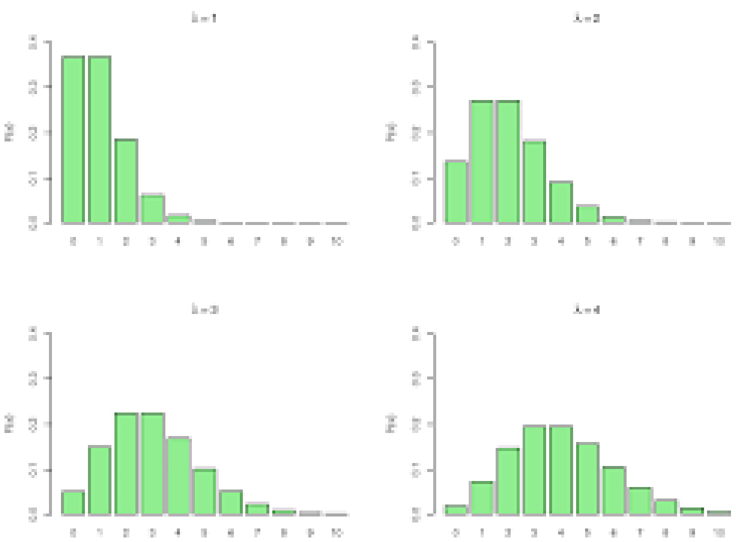
Definition 2: Poisson random variable

$$\Pr(X = x) = \frac{\exp(-\mu)\mu^x}{x!}, \quad x = 0,1,2,\ldots$$

$$E(X) = \sum_{x=0}^{\infty} x \frac{\exp(-\mu)\mu^x}{x!} = \mu.$$ (the definition of μ is the mean)

$$Var(X) = \sum_{x=0}^{\infty} (x-\mu)^2 \frac{\exp(-\mu)\mu^x}{x!} = \mu.$$

Poisson: Probability functions: x axis: x, y axis: $\Pr(X = x)$

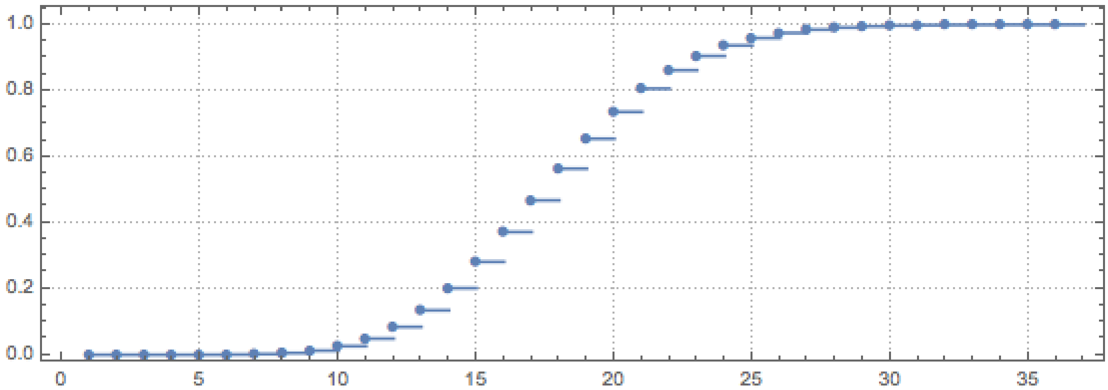


Cumulative distribution function:

$$\Pr(X \le a) = \sum_{x=0}^{x=a} \frac{\exp(-\lambda t)(\lambda t)^x}{x!} = \sum_{x=0}^{x=a} \frac{\exp(-\mu)(\mu)^x}{x!} \rightarrow \text{sum of finite series}$$

$\rightarrow$ no explicit formula

Poisson: cumulative distribution function: $\Pr(X \le x)$ $\leftarrow$ right continuous



Example: Births in a hospital occur randomly at an average rate of 1.8 births per hour.

Let X = No. of births in a given hour ~ Poisson(1.8)

1. What is the probability of observing 4 births in a given hour at the hospital?

$$\Pr(X = 4) = \frac{e^{-1.8}(1.8)^4}{4!} = \frac{0.1653 \times 10.5}{24} = 0.072$$

2. What about the probability of observing more than or equal to 2 births in a given hour at the hospital?

$$\Pr(X \ge 2) = 1 - \Pr(X = 0) - \Pr(X = 1) \approx 0.5347$$

$$\Pr(X = 0) = \frac{e^{-1.8}(1.8)^0}{0!} = 0.1653$$

$$\Pr(X = 1) = \frac{e^{-1.8}(1.8)^1}{1!} \approx 0.30$$

3. What about the probability of **observing no birth** within two hours at the hospital?

Hint: Let Y = No. of births in a two-hour period

$$E(Y) = 1.8 \times 2 = 3.6$$

$$\Pr(Y = 0) = \frac{e^{-3.6}(3.6)^0}{0!} = 0.$$

Derivations of Poisson probability: from Binomial（補充，不考）

Cut $[0,t]$ into n small non-overlapping intervals, each with length $\Delta t = \frac{t}{n}$ 。

Imagine that in each interval Δt, perform a Bernoulli random experiment, each with success probability $p = \frac{E(X)}{n} = \frac{\mu}{n}$.

$X =$ # of successes in the n trials $\Pr(X = x) = \binom{n}{x}(\frac{\mu}{n})^x(1-\frac{\mu}{n})^{n-x}$

$$= \frac{n(n-1)...(n-x+1)(n-x)...(1)}{[(x)(x-1)...1][(n-x)(n-x-1)...1]} \cdot (\frac{\mu}{n})^x \cdot (1-\frac{\mu}{n})^{n-x}$$

$$= \frac{1}{x!} \cdot \left(\frac{n}{n}\frac{n-1}{n}...\frac{n-x+1}{n}\right) \cdot \mu^x \cdot (1-\frac{\mu}{n})^{n-x}$$

$$= (1)\cdot(2)\cdot(3)\cdot(4)$$

Note: when $n \to \infty$

(1): $\frac{1}{x!}$

(2): $\frac{n}{n}\frac{n-1}{n}...\frac{n-x+1}{n} \underset{n\to\infty}{\to} 1$

(3): μ^x

(4): $(1-\frac{\mu}{n})^{n-x} = (1-\frac{\mu}{n})^n(1-\frac{\mu}{n})^{-x} \underset{n\to\infty}{\to} e^{-\mu}$

(4.1): $(1-\frac{\mu}{n})^n \underset{n\to\infty}{\to} e^{-\mu}$

(4.2): $(1-\frac{\mu}{n})^{-x} \underset{n\to\infty}{\to} 1$

Probability of Poisson random variable (taking $n \to \infty$)

$$\Pr(X = x) = \frac{1}{x!}e^{-\mu}\mu^x$$

Comparison between Binomial and Poisson distributions

	Binomial	Poisson
fixed	total # of experiment (n)	length or space
definition	count the number of successes	count the number of events
starting value	0	0
upper bound	n	no upper bound

Approximation: Use Poisson to approximate Binomial (very important)

Given $X \sim Binomial(n, p)$,

$Y \sim Poisson(\mu = \lambda t)$

There exists a situation that

when $E(X) = np = E(Y) = \mu = \lambda t$,

$\Pr(X = t) \approx \Pr(Y = t)$ for most $t = 0, 1, 2, \ldots$

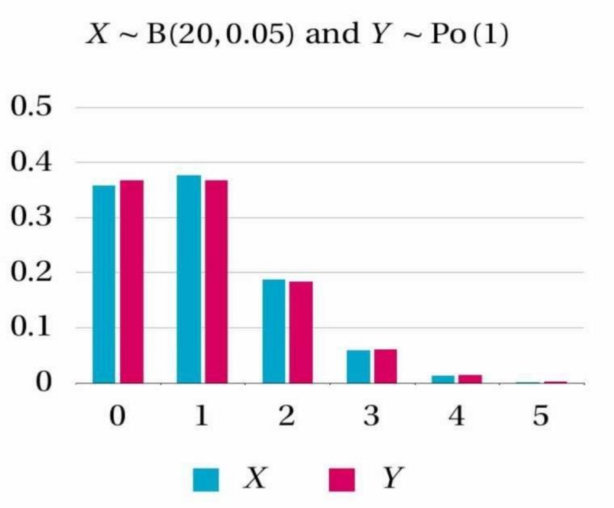
Question: When does the above situation happen?

Answer: **for large n and small p (for rare events)**

$$\Pr(X = t) = \binom{n}{t} p^t (1-p)^{n-t} \approx \Pr(Y = t) = \frac{e^{-np}(np)^t}{t!}$$

Specific criteria for the approximation:

- $n \geq 20, p \leq 0.05$ → good approximation
- $n \geq 100, np \leq 10$ → very good approximation



Remarks:

- Computing a Binomial probability is more complicated since it involves

$\frac{n!}{x!(n-x)!}$ and p^x (or $(1-p)^{n-x}$) if p is very small or large

- We usually use the Poisson distribution to approximate the corresponding Binomial distribution.

Example: bacteria mutate to become resistant to antibiotic.

$n = 2\times 10^9$ cells

$p = 10^{-9}$ (the probability of mutation of each cell)

X = # of mutation ~ Binomial($n = 2\times 10^9$, $p = 10^{-9}$)

$\Pr(X = x) = \binom{2\times 10^9}{x}(10^{-9})^x(1-10^{-9})^{2\times 10^9 - x}$ → based on Binomial

→ computationally impossible

Use Poisson for approximation

$E(X) = np = 2$ → approximated by X ~ Poisson(2)

$$\Pr(X = x) \approx \frac{e^{-2}(2)^x}{x!}.$$

$\Pr(X = 0) \approx e^{-2} = 0.1353$ = Pr(no mutation)

$\Pr(X \geq 1) = 1 - \Pr(X = 0)$ = 0.7247 = Pr(at least one mutation)

Example: Airline Overbook

- An airline knows that overall 3% of passengers do not turn up for flights .
- The airline decides to adopt a policy of selling more tickets than there are seats on a flight. → overbooked (超賣)
- For an aircraft with 196 seats, the airline sold 200 tickets for a particular flight. By using a suitable approximation, find
  - the probability that there is at least one empty seat on this flight

- the probability that at least one passenger will have to be bumped from the flight?"

Let X = number of passengers who buy the tickets but <u>will not show up</u>

= 有票但不會搭機的旅客人數

$X \sim Binomial(200, 0.03)$ → large n and small p

$X \sim Binomial(200, 0.03) \approx Y \sim Poisson(6)$

Pr(there is at least one empty seat on this flight)

= Pr(at most 195 passengers show up)

= Pr(at least 5 persons who buy the tickets but <u>will not show up</u>)

$= \Pr(X \geq 5) = 1 - \Pr(X \leq 4)$

$$= 1 - \sum_{x=0}^{x=4} \binom{200}{x} 0.03^x 0.97^{200-x}$$

= 0.7190 → sometimes very difficult to compute

$\approx \Pr(Y \geq 5)$

$$= 1 - \sum_{x=0}^{x=4} \frac{e^{-6} 6^x}{x!} = 1 - (0.002479) \sum_{x=0}^{x=4} \frac{6^x}{x!}$$

= 1- 0.2851 = 0.71494 → easier to obtain

Pr(at least one passenger will have to be bumped)

= Pr(at least 197 passengers show up)

= Pr(at most 3 persons who buy the tickets but <u>will not show up</u>)

$= \Pr(X \leq 3)$

$$\Pr(X \leq 3) = \sum_{x=0}^{x=3} \binom{200}{x} (0.03)^x (0.97)^{200-x} = .14715 \leftarrow \text{ more difficult}$$

$$\approx \sum_{x=0}^{x=3} \frac{e^{-6} 6^x}{x!} = 0.1512 \leftarrow \text{ easier}$$

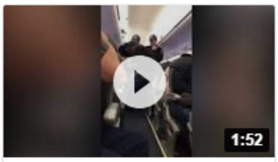
Officer who dragged bloodied doctor from overbooked plane sues ...
https://www.telegraph.co.uk › News ▾ 翻譯這個網頁
2018年4月12日 - A former aviation security officer who lost his job after assisting in dragging a man from an overbooked United Airlines flight in an incident that ...

Passenger dragged off overbook United flight - CNN - CNN.com
https://www.cnn.com/2017/04/10/travel/passenger-removed-united-flight.../index.html
2018年4月10日 - A man's refusal to give up his seat on an overbooked United Airlines flight ... He yelled that he was a doctor and that he was being profiled for ...

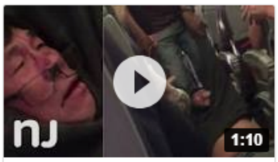
United Express Flight 3411 incident - Wikipedia
https://en.wikipedia.org/wiki/United_Express_Flight_3411_incident ▾ 翻譯這個網頁
On April 9, 2017, O'Hare International Airport Aviation Security Officers forcibly removed .... from the rude ticket agent who demanded that this man give up his seat on the flight United overbooked . ... with the supervisor not to remove him from the flight and explained that as a doctor he could not miss his return flight home.
Aftermath · United Airlines · Other airlines · United States government

影片

Doctor forcibly taken from overbooked United Airlines Flight
Tony's - 24/7 Eyes
YouTube - 2017年4月10日



Video shows a passenger forcibly dragged off a United Airlines plane
Business Insider
YouTube - 2017年4月11日



United Airlines passenger dragged off "overbooked" flight
NJ.com
YouTube - 2017年4月11日

A year after the infamous United dragging incident, has anything ...
https://www.washingtonpost.com/.../dr.../a-year-after-the-infamous-unit... - 翻譯這個網頁
2018年4月9日 - United Airlines settled with its passenger who was yanked from his seat to ... doctor was violently dragged off a United Airlines flight after refusing to give up ... has dropped significantly and airlines have reduced overbooking.

www.thenewslens.com › 國際 ▾ Translate this page
美國聯合航空超賣機位，亞裔男子竟遭航警強拖下機還被撞得 ...
Apr 11, 2017 — 因機位超賣、無法起飛，於是航空公司以800美元（約24000元台幣）希望徵求自願下機乘客，仍無人願意，只好用電腦隨機抽出四人。 標籤: 聯合 ...

www.cw.com.tw › ... › 經濟學人 ▾ Translate this page
聯航讓亞裔醫師被爆打一年後美國的航空變了｜天下雜誌
Apr 14, 2018 — 如果某架班機已超賣或可能超賣，有加入的人就會在飛機起飛5天前就收到通知，並選擇自己要 ... 延伸閱讀：消費者是健忘的聯合航空好輝煌）.

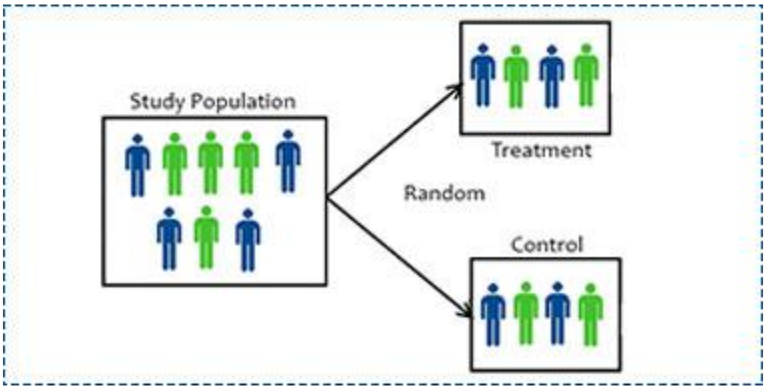
www.storm.mg › article ▾ Translate this page
聯航為何低頭道歉？業界律師：這不是超賣機位趕人下機要負民 ...
Apr 12, 2017 — 美國聯合航空公司日前一架準備從芝加哥飛往路易斯維爾的航班上，一名乘客因為拒絕改簽被安全人員強行拖下飛機，這個事件被同機乘客拍攝 ...

補充: Smoking and Lung Cancer

	Diseased (D)	not diseased (D^c)
Smoke (E)	a	b
Non-smoke (E^c)	c	d

Development of Clinical Trials（臨床試驗）

- RA Fisher: Experimental design **(1922-1926)**
  - Randomization
  - Replication
  - Blocking
- First randomized curative trial: **1946–1947**
  - test the efficacy of the chemical streptomycinfor curing pulmonary tuberculosis（肺結核）
  - Double-blind & placebo controlled
  - Compute Relative risk $\dfrac{\Pr(D\,|\,E)}{\Pr(D\,|\,E^C)}$



Problems:

- Sometimes we may not conduct clinical trials for ethical reasons.
- How to establish causal relationship from observational studies?

Fact: In 1950s, cigarette smoking was identified as an important risk factor in the development of lung cancer. But we can not conduct a clinical trial to examine whether smoking causes lung cancer.

History:

1952-1958: Many research found that

$$\frac{\Pr(\text{smoke|lung cancer})}{\Pr(\text{smoke|without lung cancer})} = \frac{\Pr(E \mid D)}{\Pr(E \mid D^C)} \approx 10$$

1958: Cornfield and other co-authors

Smoking and Lung Cancer: Recent Evidence and a Discussion of Some Questions

Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, Ernst L. Wynder

JNCI: Journal of the National Cancer Institute, Volume 22, Issue 1, January 1959, Pages 173–203, https://doi.org/10.1093/jnci/22.1.173

Published: 01 January 1959 **Article history ▼**

Cite Permissions Share ▼

Abstract

This report reviews some of the more recent epidemiologic and experimental findings on the relationship of tobacco smoking to lung cancer, and discusses some criticisms directed against the conclusion that tobacco smoking, especially cigarettes, has a causal role in the increase in bronchogenic carcinoma. The magnitude of the excess lung-cancer risk among cigarette smokers is so great that the results can not be interpreted as arising from an indirect association of cigarette smoking with some other agent or characteristic, since this hypothetical agent would have to be at least as strongly associated with lung cancer as cigarette use; no such agent has been found or suggested. The consistency of all the epidemiologic and experimental evidence also supports the conclusion of a causal relationship with cigarette smoking, while there are serious inconsistencies in reconciling the evidence with other hypotheses which have been advanced. Unquestionably there are areas where more research is necessary, and, of course, no single cause accounts for all lung cancer. The information already available, however, is sufficient for planning and activating public health measures.

Story: The debate on whether smoking caused lung cancer

Jerome Cornfield: The statistician who established risk factors for lung cancer and heart disease By Rick Wicklin March 18, 2013

https://blogs.sas.com/content/iml/2013/03/18/biography-of-jerome-cornfield.html

Jerome Cornfield: The early years

Jerome Cornfield was born the son of Russian Jewish immigrants in 1912. Like John Sall, co-founder of SAS, he got a degree in history.

He joined the US Bureau of Labor Statistics during the Depression (1935) and learned statistics by taking courses at the US Department of Agriculture from 1936–1938 (Greenhouse, p. 3)

In 1947 he joined a group that provided consultation to the fledgling National Institutes of Health (NIH).

The link between cigarettes and lung cancer

At the NIH, he became interested in a vexing problem: Why were so many people dying of lung cancer in 1950, when 50 years earlier there were so few cases reported? Was it better diagnosis? More factory pollution? A population that was living longer than ever?

Some retrospective case-control studies by Richard Doll and Bradford Hill in the UK had linked lung cancer to cigarette smoking. **But Cornfield wanted a stronger conclusion. He wanted to know whether the correlation was actually a cause-and-effect relationship.** He also wanted to use statistics to assess the risk that an individual who smokes would develop lung cancer. **By using Bayes' rule, Cornfield was able to combine Doll and Hill's results** (namely, the estimated probability that someone was a smoker, given that they had lung cancer) with NIH data to **answer the inverse question: the probability that someone would develop lung cancer, given that he was a smoker.** The result: smokers are many times more likely to develop lung cancer than nonsmokers.

Cornfield's work "stunned research epidemiologists" and his techniques "made much of modern epidemiology possible" (McGrayne, p. 111).

Although his work provoked the ire of the volatile and influential Sir Ronald Fisher (Fisher was a smoker, a paid consultant to the tobacco industry, and a fervent anti-Bayesian), Cornfield successfully defended his methods and results. In 1964, when the US Surgeon General warned "that cigarette smoking is causally related to lung cancer" (McGrayne, p. 113), Cornfield's work was among the evidence that was cited.

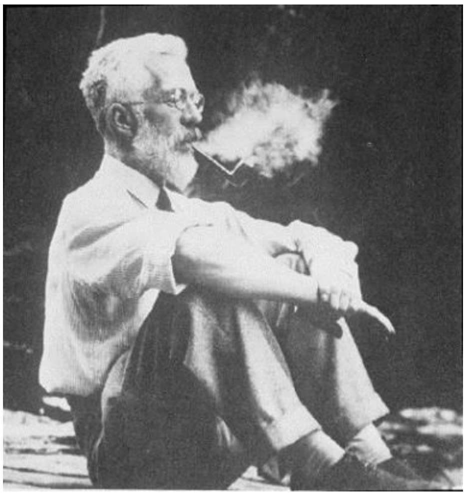
Cornfield: case-control study to study rare diseases.

Statistics in Epidemiology: The Case-Control Study N. E. Breslow

Journal of the American Statistical Association (1996), pp. 14-28

eters of primary interest—namely, the disease rates. This misconception was corrected by Jerome Cornfield (1951), who is widely credited with launching the modern era of case-control studies. Cornfield demonstrated that the exposure odds ratio for cases versus controls equals the disease odds ratio for exposed versus unexposed, and that the latter in turn approximates the ratio of disease rates provided that the disease is rare. Formally, if D denotes disease (1 for cases, 0 for controls) and X denotes exposure (1 for

RA Fisher: No! smoking did not cause lung cancer.



Cornfield's demonstration did not quiet all the critics, one of the most vociferous being R. A. Fisher (1957a,b; 1958a,b). Fisher raised the issue of association versus causation that clouds the interpretation of any observational study. In his famous constitutional hypothesis, he suggested that the smoking and lung cancer association could be explained by the confounding effects of a genotype that predisposed both to smoking and to lung cancer. Data on twins were used to substantiate his assertion that smoking behavior was influenced by genetics. Cornfield and colleagues (1959) responded to these charges and others in a lengthy review that is well worth reading for its insights regarding causal inference. A centerpiece of their argument was

Mathematics Genealogy Project 數學族譜

https://www.genealogy.math.ndsu.nodak.edu/id.php?id=73510

Mathematics Genealogy Project

Home
Search
Extrema
About MGP
Links
FAQs
Posters
Submit Data
Contact
Mirrors
Donate

A service of the NDSU Department of Mathematics, in association with the American Mathematical Society.

Weijing Wang

MathSciNet

Ph.D. Cornell University 1995

Dissertation: *Estimation and Modeling in Bivariate Survival Analysis*

Advisor: Martin Timothy Wells

Students:
Click here to see the students listed in chronological order.

Name	School	Year	Descendants
Chang, Wei-Hwa	National Chiao Tung University	2006	
Chen, Chien-Hao	National Chiao Tung University	2007	
Emura, Takeshi	National Chiao Tung University	2007	
Hsieh, Jin-Jian	National Chiao Tung University	2006	
Kor, Chew-Teng	National Chiao Tung University	2008	
Li, Bowen	National Chiao Tung University	2013	
Su, Chien-Lin	National Chiao Tung University	2015	