Contents...

1 (Poisson) Model for (Sampling)Variability of Count in a given amount of "experience"

The Poisson Distribution: what it is, and some of its features

- The (infinite number of) probabilities $P_0, P_1, ..., P_y, ...$, of observing $Y = 0, 1, 2, \ldots, y, \ldots$ "events" in a given amount of "experience."
- These probabilities, $Prob[Y = y]$, or $P_Y[y]$'s, or P_y's for short, are governed by a single parameter, the mean $E[Y] = \mu$.
- $P[y] = \exp[-\mu]\, \mu^y / y!$ {note recurrence relation: $P_y = P_{y-1} \times (\mu/y)$.}
- Shorthand: $Y \sim$ Poisson(μ).
- $Var[Y] = \mu$; i.e., $\sigma^2_Y = \mu_Y$.
- Approximated by $N(\mu,\, \sigma_Y = \mu^{1/2})$ when $\mu >> 10$.
- Open-ended (unlike Binomial), but in practice, has finite range.
- Poisson data sometimes called "numerator only": (unlike Binomial) may not "see" or count "non-events": but there is (an invisible) denominator 'behind' the no. of incoming "wrong number" phone calls you receive.

How it arises / derivations

- Count of events (items) that occur randomly, with low homogeneous intensity, in time, space, or 'item'-time (e.g. person–time).
- Binomial(n, π) when $n \to \infty$ and $\pi \to 0$, but $n \times \pi = \mu$ is finite.[1]
- $Y \sim Poisson(\mu_Y) \Leftrightarrow T$ time b/w events $\sim$ Exponential($\mu_T = 1/\mu_Y$).[2]
- As sum of ≥ 2 *independent* Poisson rv's, with same **or different** μ's: $Y_1 \sim$ Poisson(μ_1) $Y_2 \sim$ Poisson(μ_2) $\Rightarrow Y = Y_1 + Y_2 \sim$ Poisson($\mu_1 + \mu_2$).
- **Examples**: numbers of asbestos fibres, deaths from horse kicks*, needlestick or other percutaneous injuries, bus-driver accidents*, twin-pairs*, radioactive disintegrations*, flying-bomb hits*, white blood cells, typographical errors, "wrong numbers", cancers; chocolate chips, radioactive emissions in nuclear medicine, cell occupants – in a given volume, area, line-length, population-time, time, etc. [*e.g. on website]

[1] See also: derivation & applications (counting yeast cells in beer) in Student's 1907 paper "On the Error of Counting with a Haemacytometer"; and Ch. from Armitage et al.
[2] cf. ***** "Randomness at the root of things: Poisson sequences" – Physics Education.

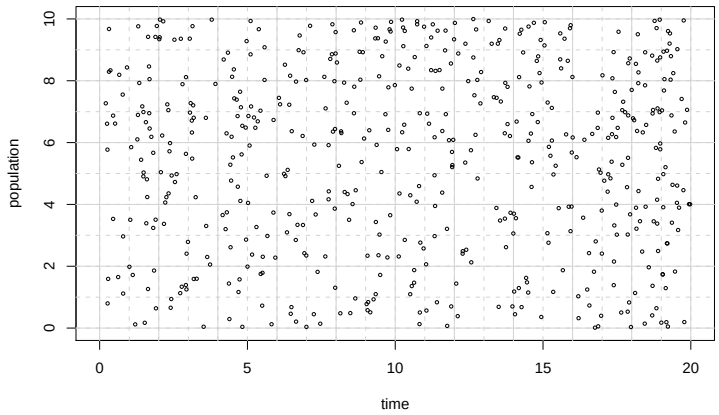


Figure 1: Events in Population-Time.. randomly generated from intensities that are constant within (2 squares high by 2 squares wide) 'panels', but vary between such panels. In Epidemiology, each square might represent a number of units of Population-time, and each dot an event.

1.1 Does the Poisson Distribution apply to.. ?

- Yearly variations in no.s of persons killed in plane crashes? [3]
- Daily variations in numbers of births?[4]
- Daily variations in no.s of deaths [variation over the seasons]
- Daily variations in numbers of traffic accidents [variation over the seasons, and days of week, and with weather etc.]
- Daily variations in numbers of deaths in France in summer 2002 & 2003[5]

[3] Yearly variations in no.s of plane *crashes* may be closer to Poisson [apart from variation due to improvements in safety, fluctuations in no.s of flights etc].

[4] See e.g. Number of weekday and weekend births in New York in August 1966 on web page: the variations are closer to Poisson if use *weekly* count.

[5] c.f. Impact sanitaire de la vague de chaleur en France survenue en août 2003. Rapport d'étape 29 aot 2003 [on course webpage] and Vanhems P et al. Number of in-hospital deaths at Edouard Herriot Hospital ,and Daily Maximal Temperatures during summers of 2002 and 2003, Lyon, France. New Eng J Med Nov 20, 2003, pp2077-2078. ibid. see Resources.

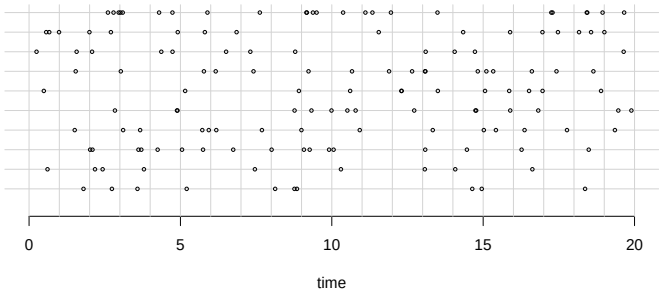


Figure 2: Events in Time: 10 examples, randomly generated from constant over time intensities. Simulated with 1000 Bernoulli($_{\text{small}}\ \pi$)'s per time unit.

- Variations across cookies/pizzas in no. of chocolate chips/olives.
- Variations across days of year in no. of deaths from sudden infant death syndrome.

1.2 Calculating Poisson probabilities:

1.2.1 Exactly

- pdf: formula for P_y (can use recursion).
- cdf:
  - Summation of terms in pdf
  - Using link between this sum and the cdf of χ^2 Distribution[6].
- Spreadsheet — Excel function `POISSON`(y, μ, cumulative)
- Statistical Packages: `SAS` function `POISSON`; see `www.ats.ucla.edu/stat/stata/faq/pprob.htm` for 'how to' in `Stata`; `R` functions `dpois(), ppois(), qpois()` probability, distribution, and quantile functions.

[6] Fisher 1935: see Resources

1.2.2 Using Gaussian Approximations to distribution of y or transforms of it

Described below, under Inference.

2 Inference re μ, based on observed count y

2.1 "First Principles" Confidence Interval

By first-principles $100(1-\alpha)\%$ CI, we mean "not usual point-estimate $\pm$ some multiple of standard error," but rather the pair $(\mu_{LOWER}, \mu_{UPPER})$ such that

$$P(Y \geq y \,|\, \mu_{LOWER}) = \alpha/2 \;\; \& \;\; P(Y \leq y \,|\, \mu_{UPPER}) = \alpha/2.$$

2.1.1 Exact – see Figure 3 for example, based on $y = 6$

Tables: For a given α, there is just one CI for each value of y; these exact CI's have been extensively tabulated and made available in several texts and Tables, e.g., the Documenta Geigy, and Biometrika Tables for Statisticians.[7]

If don't have tables.. Can find exact lower and upper limits $\mu_{LOWER/UPPER}$ that yield the target $\alpha/2$'s either by **trial & error** (rapidly with software that evaluates Poisson tail areas) or **directly** using the Link between the tail areas of the Poisson and tail areas of Chi- Square distributions (full details in article by Fisher, 1935, under Resources on webpage), $\mu_{LOWER} = \frac{1}{2}\chi^2_{2y,\alpha/2}$, $\mu_{UPPER} = \frac{1}{2}\chi^2_{2y+2,1-\alpha/2}$. Some specialized software packages (eg in `R`) also have functions that provide them directly.[8]

2.1.2 Quite accurate approximation to exact tail area

Using Wilson-Hilferty approximation to Chi-square quantiles[9] This has high accuracy for $y > 10$; it uses z, the normal standardized variate

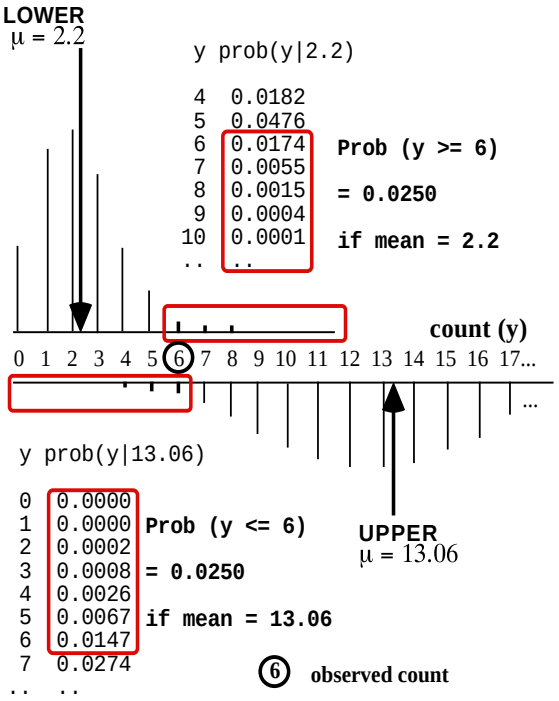


Figure 3: Example of Exact 95% CI of $\{2.2, 13.06\}$ for μ, based on $y = 6$.

corresponding to $\alpha/2$, e.g., $\alpha = 0.10 \rightarrow z = 1.645; \alpha = 0.05 \rightarrow z = 1.96,$ etc.

$$\mu_{LOWER} = y \times \{1 - [9y]^{-1} - z \times [9y]^{-1/2}\}^3$$

$$\mu_{UPPER} = (y+1) \times \{1 - [9(y+1)]^{-1} + z \times [9(y+1)]^{-1/2}\}^3$$

2.1.3 Not quite as accurate an approximation, but 1st principles

Using $Y \simeq N(\mu,\, \sigma = \mu^{1/2})$:

[7] (See (homemade) "Confidence limits for the expectation [i.e. the 'mean' parameter] of a Poisson random variable" on last page and (more fully) under Resources.

[8] Note that the above "First Principle" is a *general* and important one; it "just so happens" that in this particular discrete distribution, if one has access to the percentiles of the Chi-Square distribution, the link helps avoid the trial and error process involved.

[9] Rothman[2002], page 134, provides an adaptation from "D. Byar, unpublished" in which he makes a further approximation, using the average $(y + 0.5)$ for both the lower an upper limits, rather than the more accurate y for the lower and $y + 1$ for the upper limit. JH is surprised at Rothman's eagerness to save a few keystrokes on his calculator, and at his reference to an unpublished source, rather than the 1931 publication of Wilson and Hilferty. The Full Wilson and Hilferty citation, and evaluation of the above equation, can be found in Liddell's article "Simple exact analysis of the standardized mortality ratio" in J Epi and Comm. Health 37 85-88, 1984, available on website.

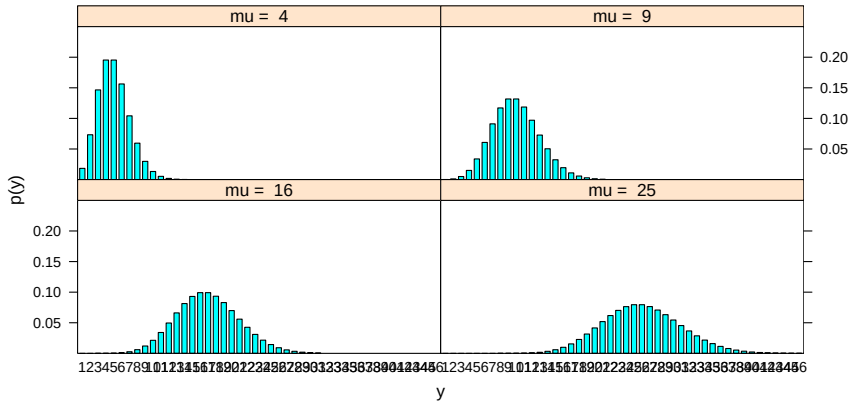


Figure 4: Poisson Distributions, over range $y = 0$ to 40, corresponding to $\mu = 4,\ 9,\ 16,$ and 25.

Obtained by solving the two equations:

$$y = \mu_{LOWER} + z \times \{\mu_{LOWER}\}^{1/2} \ ; \ y = \mu_{UPPER} - z \times \{\mu_{UPPER}\}^{1/2}$$

to give

$$\mu_{LOWER,UPPER} = ([y + z^2/4]^{1/2} \mp z/2)^2.$$

2.1.4 Variance-stabilizing transformation, so first principles

With μ large enough, $Y^{1/2} \sim (approx) N(\mu^{1/2},\ \sigma = 1/2)$, i.e., the SD is independent of $\mu^{1/2}$, thus providing for a first-principles interval:

$$\mu_{LOWER,UPPER} = y \mp z \times c^{1/2} + (z/2)^2.$$

2.2 "Not First Principles" Confidence Intervals, based on SE calculated at point estimate

2.2.1 Based on $Y \simeq N(\mu,\ \sigma = y^{1/2})$.

If lazy, or don't care about principles /accuracy, or if y is large, can solve

$$y = \mu_{LOWER} + z \times y^{1/2} \ ; \ y = \mu_{UPPER} - z \times y^{1/2}$$

to give

$$\mu_{LOWER,UPPER} = y \mp z \times y^{1/2}.$$

"Large-n": How Large is large?: The same rule of thumb: when expected no. of events, $\mu = E[Y] > 5$. or the same JH rule: when the tables don't go as high as your value of y. It works well if the distribution is not 'crowded' into the left corner (cf. Figure 3), i.e., if, with the symmetric Gaussian approximation, the lower tail of the distribution does not spill over the 0 boundary.

The above model is used if one fits a generalized linear model, with Poisson error but **IDENTITY link**. Example with $y = 4$:

e.g. In SAS:

```
PROC GENMOD; model y = /  dist = POISSON link = IDENTITY WALFCI;
```

e.g. In Stata

```
input y
1 * glm doesn't like file with 1 'observation'
3 *so ........... split across 2 'observations'
end
glm y , family (poisson) link (identity)
```

e.g. In R: `y=4; summary(glm(y ~ 1,family=poisson (link=identity) ))`

2.2.2 Based on $\log Y \simeq N(\log[\mu],\ \hat{\sigma} = \{1/\hat{\mu}\}^{1/2})$

- **Derivation:**

  Use the "Delta Method" to derive the approximate variance for the random variable $\log Y$, assuming that $\text{Prob}[Y = 0]$ is negligible.

  $\text{Var}[\log Y] \approx \text{Var}[Y] \times \{(d \log Y/dY)|_{Y=\mu_Y}\}^2 = \mu \times (1/\mu_Y)^2 = 1/\mu_Y.$

  We will make a lot of use of this result, especially for the variance of the log of a rate, and for the variance of the **log of a rate ratio** (i.e., the variance of the difference of two log rates).

  (Empirical) *rate ratio(rr)* or *id ratio (idr)*: $\hat{\lambda}_2 \div \hat{\lambda}_2 = \frac{y_2}{PT_2} \div \frac{y_1}{PT_1}$

$$\begin{aligned} Var[\log(rr)] &= Var[\log y_2 - \log y_1] \\ &= Var[\log y_2] + Var[\log y_1] \\ &= 1/\mu_2 + 1/\mu_1. \\ &= 2 \div \{ \text{ Harmonic Mean of } \mu_2 \text{ and } \mu_1\} \end{aligned}$$

- **CI:**

  CI: $\exp[\log(y) \mp z \times (1/y)^{1/2}]$.

  e.g., ...

  In SAS: `MODEL y = / dist = POISSON link = LOG WALFCI;`

  In Stata: `glm y , family (poisson) link (log)`

  In R: `summary(glm(y ~ 1,family=poisson(link=log) ))`

3 Applications, and Notes

3.1 How large a count so that margin of error $< 15\%$?

An estimate of WBC concentration can be made by manually counting enough fields (n) until say a total of $y = 200$ cells have been observed. This is not quite a Poisson distribution since $y = 200$ is fixed ahead of time and n is the random variable – but the variability in the estimate $200/n$ is close to Poisson-based, so as a first approximation we will treat the y as the variable and the denominator n as fixed. The estimate has margins of error (ME) of 13% and 15% – since a total count of 200 (marked by ↑ below) could be a *high* reading from a concentration which produce a μ of 173 (for the same n), or a *low* reading from a concentration which produces an average of $\mu = 230$, i.e.

y `per` n: `160..170..180..190..`***200***`..210..220..230..240...`

$\ldots\ldots\ldots\ldots\ldots\mu_L\ldots\ldots\ldots\ldots\uparrow\ldots\ldots\ldots\ldots\mu_U\ldots\ldots$

$\ldots\ldots\ldots\ldots\ldots$ ***200*** $= 173 + 1.96 \times \{173\}^{1/2}\ldots\ldots\ldots\ldots$

$\ldots\ldots\ldots\ldots\ldots$ ***200*** $= 230 - 1.96 \times \{230\}^{1/2}\ldots\ldots\ldots\ldots$

3.2 Leukemia Rate Triples near Nuke Plant: Study

OTTAWA (CP)[10] - Children born near a nuclear power station on Lake Huron have 3.5 times the normal rate of leukemia, according to figures made public yesterday. The study conducted for the Atomic Energy Control Board, found the higher rate among children born near the Bruce generating station at Douglas Point. But the scientist who headed the research team cautioned that the sample size was so small that that actual result could be much lower - or nearly four times higher.

Dr. Aileen Clarke said that while the Douglas Point results showed 3.5 cases of leukemia where one would have been normal[11], a larger sample size could place the true figure somewhere in the range from 0.4 cases to 12.6 cases.[12]

Clarke will do a second study to look at leukemia rates among children aged five to 14. The first study was on children under age 5. Clarke was asked whether parents should worry about the possibility that childhood leukemia rates could be over 12 times higher than normal around Douglas point. "My personal opinion is, not at this time," she said. She suggested that parents worried by the results should put them in context with other causes of death in children.

"Accidents are by far and away the chief cause of death in children, and what we're talking about is a very much smaller risk than that of death due to accidents," she said.

The results were detailed in a report on a year-long study into leukemia rates among children born within a 25-kilometre radius of five Ontario nuclear facilities. The study was ordered after British scientists reported leukemia rates among children born near nuclear processing plants were nine times higher than was normal. The Ontario study was based on 795 children who died of leukemia between 1950 and 1986 and 951 children who were diagnosed with cancer between 1964 and 1985.

It showed a lower-than-normal rate among children born near the Chalk River research station and only slightly higher than expected rates at Elliot Lake and Port Hope, uranium mining and conversion facilities.

At the Pickering generating station, the ratio was slightly higher still, at 1.4 - meaning there were 1.4 cases for every expected case. But the confidence interval - the range of reliability - for that figure set the possible range between 0.8 cases and 2.2 cases.[13]

Comment: It is interesting that it is the more extreme, but much less precise, SIR of 3.5, based on $O = 2, E = 0.57$ that made the headline, while the less extreme, but much more precise, SIR of 1.4, based on $O = 18, E = 12.8$, was relegated to the last paragraph.

[10] Montreal Gazette, Friday May 12, 1989.

[11] $SIR = 3.5 = No.Observed/No.Expected$. It is not $O = 3.5, E = 1$, since one cannot observe a fractional number of cases): SIR = 3.5; she simply scaled the O and the E so that E (reference "rate") is 1

[12] CI = (CI derived from O)/Expected = 0.4 to 12.6 (a 31-fold range). O is an integer. By trial and error, starting with O=1, and "trying all the CI's on for size" until one gets a 31-fold range, one comes to $O = 2$. (CI 0.242 to 7.22, range 31 fold). Dividing 2 by 3.5 gives an E of 0.57. Check: 95% CI for SIR (0.242 to 7.22) / 0.57 = 0.4 to 12.6.

[13] $SIR = 1.4 = O/E; CI = (CI$ derived from $O)/E$ has 0.8 to 2.2. This 2./0.8= 2.75-fold uncertainty comes from uncertainty generated by O. Examine range of 95% CI associated with each possible value of O, until come to 10.67 to 28.45 when $O = 18$. Divide 18 by 1.4 to get $E = 12.8$. Check 95% CI 10.67 to 28.45)/12.8 = 0.8 to 2.2.

3.3 Self-reported Percutaneous Injuries in Interns

Table 1. Rates of Percutaneous Injuries by Residency Program.[14]

Type of Residency	No. of Intern-Months	No. of Percutaneous Injuries	Rate (95% CI*) per Intern-Month
All	17003	498	0.0293 (0.0268-0.0318)
Internal medicine	3995	57	0.0143 (0.0106-0.0179)
Surgery	1730	124	0.0717 (0.0595-0.0838)
Family medicine	2008	51	0.0254 (0.0185-0.0323)
Emergency medicine	1007	40	0.0397 (0.0277-0.0518)
Pediatrics	2159	24	0.0111 (0.0067-0.0155)
Psychiatry	658	1	0.0015 (0.0000-0.0045)
Pathology	283	15	0.0530 (0.0269-0.0791)
Obstetrics/gynecology	964	94	0.0975 (0.0788-0.1160)
Other specialties	4199	92	0.0219 (0.0175-0.0263)

*Method not specified, but $\{498 \mp 498^{1/2}\} \div 17003 = \{0.0267, 0.0318\}$.

3.4 Cell Occupancy, Lotto 6/49, the Extremal Quotient, and Geographical Variation in Surgery Rates

What do these have in common? The answer may be easier to understand after seeing a few runs of the Excel Macro for visits to cells (in Resources).

3.5 Note: How is it that one can form a CI for μ from a *single* observation y?

If we had a single realization y of a $N(\mu, \sigma_Y)$ random variable, we could not, from this single y, estimate **both** μ and σ_Y: one would have to rely on *outside* information concerning σ_Y. However, the Poisson(μ) distribution is different in that $\sigma_Y = \mu^{1/2}$, so we can calculate a "model-based" SE (or SE's if use a first principles CI) from this relationship between the mean and the variance.

Another way to understand why a SE is possible without going "outside" is to take advantage of the "divisibility" of a Poisson denominator, and its corresponding numerator.

We can split up the overall sample or slice of experience into (n) small enough sub samples so that the subcount y_i in each sub sample will be either a 0 or a 1 The (unit) variance of the observed sub counts should be $p(1-p)$ where p is the proportion of sub counts that are 1. Thus the estimated variance of the total count $y = \sum_i y_i$ should be n times the unit variance, or $n \times p(1-p)$. But if p is small, so that $1-p$ is near unity, then the variance of the sub count is approximately $n \times p$, which is simply the observed overall count y. i.e. the variance of a Poisson variable is equal to its mean. see more under Resources.

The sum of independent Poisson r.v.'s with different expectations is still a random variable with a Poisson distribution. The same is not true of a sum of independent Bernoulli (or Binomial) r.v.'s with different expectations.

If you were told that $Y_1 \sim$ Bernoulli($\pi_1 = 0.1$) and $Y_2 \sim$ Bernoulli($\pi_2 = 0.7$), you would not argue that the distribution of $Y = Y_1 + Y_2$ is Binomial($n = 2, \pi = 0.4$). You can check that yes, E(Y) = 0.8, but that $P_0 = 0.27$, $P_1 = 0.66$, $P_2 = 0.07$, much more concentrated than the Binomial(2,0.4) probabilities $P_0 = 0.36$, $P_1 = 0.48$, $P_2 = 0.16$.

BUT, what if you were told that $Y_1 \sim$ Poisson($\mu_1 = 0.1$) and $Y_2 \sim$ Poisson($\mu_2 = 0.7$). Would you argue that the distribution of $Y = Y_1 + Y_2$ is Poisson($\mu = 0.8$)? You can check that in fact it is.

In epidemiology, prevalence and other proportion-type statistics have denominators made up of (indivisible) individuals; the *person* is the statistical atom. However, when dealing with incidence density statistics, the denominators are made up of an infinite number of *person-moments.*

3.6 CI for Rate or Incidence Density parameter, ID (λ)

So far, we have focused on inference regarding μ, the **expected number on events in the amount of experience actually studied**. However, for comparison purposes, the frequency is more often expressed as a **rate**, or **intensity**. e.g., we convert $y = 211$ deaths from lung cancer in 232978 women-years (WY) in the age-group 55-60 in Quebec in 2002 into a rate of incidence density of 211/(232,978WY) = 0.00091/WY or **91** per 100,000WY. This makes it easier to compare the frequency with the rate in the same age group in 1971, namely 33 lung cancer deaths in 131200WY, or 0.00025/WY = **25** per 100,000WY.

The *statistic*, the empirical rate or empirical incidence density, is

$$rate = id = \hat{ID} = \hat{\lambda} = y/\text{PT}.$$

where y is the observed number of events and PT is the amount of Population-Time in which these events were observed. We think of id or $\hat{ID}$ or $\hat{\lambda}$ as a point estimate of the (theoretical) Incidence Density *parameter*, ID or λ.

[14] Ayas NT, et al. JAMA. 2006;296:1055-1062

To calculate a CI for the ID parameter, we treat the PT denominator as a constant, and the numerator, y, as a Poisson random variable, with expectation $E[y] = \mu = \lambda \times PT$, so that

$$\lambda = \mu \div \text{PT},$$

$$\hat{\lambda} = \hat{\mu} \div \text{PT} = y \div \text{ PT},$$

$$\text{CI for } \lambda = \{\text{CI for } \mu\} \div \text{PT}.$$

The $y = 211$ leads to a (large-sample, SE-based)

$$95\% \text{ CI for } \mu: \ 211 \mp 1.96 \times 211^{1/2} \Rightarrow 211 \mp 28.5 \Rightarrow \{182.5, 239.5\}$$

$$95\% \text{ CI for } \lambda: \ \{182.5, 239.5\} \div 232,978\text{WY} \Rightarrow \{\mathbf{78.3}, \mathbf{102}\} \text{ per } 100,000\text{WY}$$

Whereas it matters little which method – exact or approximate – to use for the 95% CI from the 2002 data, the number of deaths in 1971 is a much smaller $y = 33$. Thus we will use the exact first principles CI for μ. The available tables stop at $y = 30$, so we will use the Excel spreadsheet, in the Resources, with a count of 33. It yields lower and upper limits of 22.7 and 46.3. Thus, to accompany the point estimate of 25 per 100,000WY, we have

$$95\% \text{ CI for } \lambda: \ \{22.7, 46.3\} \div 131,200\text{WY} \Rightarrow \{\mathbf{17.3}, \mathbf{35.3}\} \text{ per } 100,000\text{WY}$$

4 Test of $H_0 : \mu = \mu_0$, i.e. $\lambda = \lambda_0$.

Evidence: P-Value = Prob[y or more extreme $\mid H_0$], with 'more extreme' determined by whether H_{alt} is 1-sided or 2-sided.

For **formal test**, at level α, compare this P-value with α.

Examples:

1. **Cancers at Douglas Point**:

   Denote by $\{CY_1, CY_2, \dots\}$ the numbers of Douglas Point child-years of experience in the various age categories that were pooled over. Denote by $\{\lambda_1^{Ont}, \lambda_2^{Ont}, \dots\}$ the age-specific leukemia incidence rates during the period studied. If the underlying incidence rates in Douglas Point were the same as those in the rest of Ontario, the ***E***xpected total number of cases of leukemia for Douglas Point would be

   $$E = \mu_0 = \sum_{ages} CY_1 \times \lambda_i^{Ont} = 0.57.$$

   The actual total number of cases of leukemia ***O***bserved in Douglas Point was

   $$O = y = \sum_{ages} O_i = 2.$$

   So, (age-) *Standardized Incidence Ratio (SIR)* $= O/E = 2/0.57 = 3.5$.

   Q: Is the $y = 2$ cases of leukemia observed in the Douglas Point experience statistically significantly higher than the $E = 0.57$ cases "expected" for this many child-years of observation if in fact the rates in Douglas Point and the rest of Ontario were the same? Or, is the $y = 2$ observed in this community compatible with $H_0 : y \sim \text{Poisson}(\mu = 0.57)$?

   A: Since, under H_0, the age-specific numbers of leukemias $\{y_1 = O_1, y_2 = O_2, \dots\}$ in Douglas Point can be regared as independent Poisson random variables, their sum y can be regarded as a Poisson random variable with $\mu = 0.57$. Thus we can calculate $P = Prob[Y \geq y \mid \mu = 0.57] = P[2] + P[3] + \dots$, i.e.

   $$P_{uppertail} = 1 - \{P[0] + P[1]\} = 1 - \{\ exp[-0.57] \times (1 + 0.57/1!)\} = 0.11.$$

   At the **Pickering** generating station, the **O**bserved number was **18**, versus an **E**xpected of **12.8**, for an *SIR* of 1.4. These larger numbers give us a chance to compare the *uppertail* P-values obtained by the exact method, i..e. $P = \sum_{y \geq 18} \text{PoissonProb}[y \mid \mu_0 = 12.8]$ with those obtained from various **approximations** to the Poisson($\mu_0 = 12.8$) distribution:-

   - Exactly $P = \text{Poisson Prob}[18|12.8] + P[19|12.8] + \dots = 0.099$
   - No (dis)-continuity correction: $P = \text{Prob}[Z \geq \frac{18 - 12.8}{12.8^{1/2}}] = 0.073$
   - No (dis)-continuity correction: $P = \frac{1}{2}\text{Prob}[\chi^2 \geq \frac{(18-12.8)^2}{12.8}] = 0.073$
   - No correction: [15] $P = \text{Prob}[Z \geq \log(18/12.8)/\{1/12.8\}^{1/2}] = 0.111$
   - No correction: [16] $P = \text{Prob}[Z \geq (18^{1/2} - 12.8^{1/2})/0.5] = 0.092$
   - With the correction: $P = \text{Prob}[Z \geq \frac{|18-12.8|-0.5}{12.8^{1/2}}] = 0.094$
   - With the correction: $P = \frac{1}{2}\text{Prob}[Z \geq \frac{(|18-12.8|-0.5)^2}{12.8}] = 0.094$

[15] Using $\log y \simeq N(\log \mu,\ \sigma = (1/\mu)^{1/2})$.
[16] Using $y^{1/2} \simeq N(\mu^{1/2},\ \sigma = 1/2)$.

2. "**Cluster of Events**" Story in *Montreal Gazette* in May 1989[17]

> Double Trouble in Moose Jaw School
> *(caption to a photograph showing 6 sets of twins)*
> Every morning, teachers at Prince Arthur school in Moose Jaw, Saskatchewan see double – and its not because of what they were up to the night before. Six pairs of identical twins attend the school, which has an enrollment of 375. Identical births occur once in 270 births.

What is the probability P of having 6 or more sets of twins in a school of size $n = 375$, when the twinning probability is $\pi = 1/270$?

This can be obtained with the Binomial(n, π) distribution; because n is large and π is small, the distribution can be approximated by the Poisson(μ) distribution, where $\mu = n \times \pi = 1.3$.

$$P = P[Y \geq 6] = 1 - P[Y \leq 5],$$

i.e., as

$$1 - \exp[-1.3] \times \{1 + 1.3/1! + 1.3^2/2! + 1.3^3/3! + 1.3^4/4! + 1.3^5/5!\} = 0.0022.$$

Thus, the probability is low that **this particular school** would have six or more sets. BUT, on average, in 1000 schools of this size, there will be 2.2 with this many or more. Thus, if we scan over a large number of such schools, finding **some school somewhere** with this extreme a number is not didfficult. If the newswires scanned a large number of schhols in 2007, there is a good chance the Montreal Gazette could re-use the headline – but they would have to change "Moose Jaw" to "Town X", with "X" to be filled in.

Moral: The Law of Large Numbers at play here is the same as the one in the video display terminals and miscarriages" story. *Natural "clusters" do occur by chance alone*, and distinguishing ones caused simply by chance from ones caused by some environmental or other such factor is not an easy task.

3. (Large-sample) Example Where does the $O = 78$ cases of cancer in the "Sour Gas" community of Alberta fall relative to $E = 85.9$ "expected" for "non-sour-gas" communities with the same person years of experience and at Alberta cancer rates?

[17] See Hanley JA "Jumping to coincidences: defying odds in the realm of the preposterous". the American Statistician, 1992, 46(3) 197-202. – under Resources

5 Modelling Incidence Densities, or Rates, (λ's) via regression

Figure 5 is a simple mathematical reversal of the fundamental epidemiologic definition of an empirical rate or incidence density (id)

$$rate = id = \frac{\text{number of cases}}{\text{amount of population-time that generated these cases}},$$

i.e.,

number of cases = $rate \times$ amount of population-time.

There is a corresponding equation for the **expected** number of cases, in terms of the **theoretical rate,** λ:

$$\mathbf{E}[\text{number of cases}] = \textbf{\textit{theoretical rate}} \times \text{amount of population-time.}$$

This simple re-statement has two important implications (i) ***in epidemiology, we are students of rates*** and (ii) Generalized ***Linear Models (GLMs) allow us to fit equations of this very type***. Even though we put the numbers of cases on the left hand side of the regression equation, these GLMs allow us to express the theoretical rates (the focus of our investigations) as functions of the determinants of interest (e.g. age, smoking, diet, calendar time, treatment, ... etc) while treating the amounts of population time as constants that are of no intrinsic interest. In the lung cancer mortality dataset, we could have a (no. deaths, PT) 'data point' for every 'covariate pattern' or x-vector.

The two most common theoretical rate models are the additive and multiplicative forms:-

$$\textbf{\textit{rate}}[x] = \beta x \quad \& \quad \textbf{\textit{rate}}[x] = \exp(\beta x).$$

More later...

6 Planning: Sample Size for CIs and Tests

6.1 Precision

Even though it is tempting to specify the 'sample size' in terms of the Amount of Experience that needs to be studied to achieve this precision, ultimately the precision is governed by the is safer to specify it it terms of the numbers of

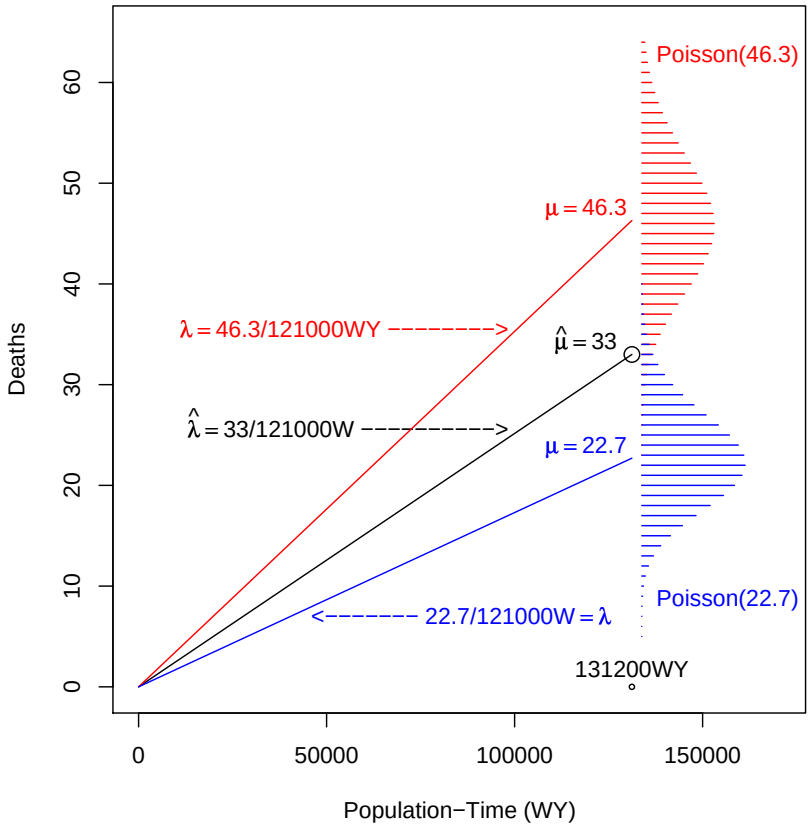
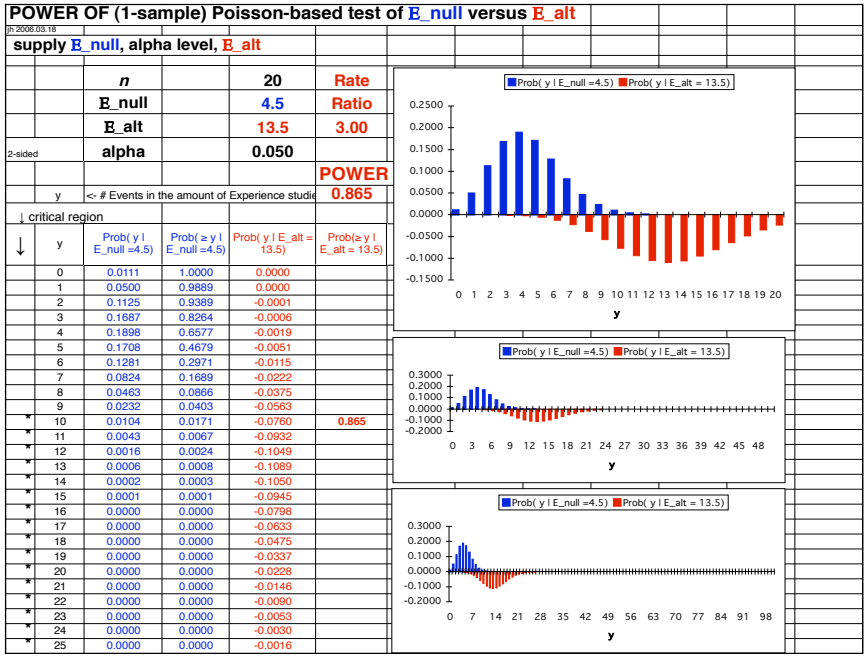


Figure 5: Depiction of empirical lung cancer mortality rate in age-group 55-60 in Quebec in 2002 as the slope of the line joining the point ($Y = 0$ cases, $PT = 0WY$) and the point ($Y = 33, PT = 121300WY$). Also shown are the Poisson Distributions, with $\mu_{UPPER} = 46.3$ and $\mu_{LOWER} = 22.7$ respectively, such that Prob$[Y \geq 33 \,|\, \mu = 22.7 =$ Prob$[Y \leq 33 \,|\, \mu = 46.3 = 0.025$.



↓	y	Prob(y | E_null =4.5)	Prob(≥ y | E_null =4.5)	Prob(y | E_alt = 13.5)	Prob(≥ y | E_alt = 13.5)
	0	0.0111	1.0000	0.0000	
	1	0.0500	0.9889	0.0000	
	2	0.1125	0.9389	-0.0001	
	3	0.1687	0.8264	-0.0006	
	4	0.1898	0.6577	-0.0019	
	5	0.1708	0.4679	-0.0051	
	6	0.1281	0.2971	-0.0115	
	7	0.0824	0.1689	-0.0222	
	8	0.0463	0.0866	-0.0375	
	9	0.0232	0.0403	-0.0563	
*	10	0.0104	0.0171	-0.0760	0.865
*	11	0.0043	0.0067	-0.0932	
*	12	0.0016	0.0024	-0.1049	
*	13	0.0006	0.0008	-0.1089	
*	14	0.0002	0.0003	-0.1050	
*	15	0.0001	0.0001	-0.0945	
*	16	0.0000	0.0000	-0.0798	
*	17	0.0000	0.0000	-0.0633	
*	18	0.0000	0.0000	-0.0475	
*	19	0.0000	0.0000	-0.0337	
*	20	0.0000	0.0000	-0.0228	
*	21	0.0000	0.0000	-0.0146	
*	22	0.0000	0.0000	-0.0090	
*	23	0.0000	0.0000	-0.0053	
*	24	0.0000	0.0000	-0.0030	
*	25	0.0000	0.0000	-0.0016	

Figure 6: Using exact Poisson Probabilities

Amount of experience In order to achieve a specified Coefficient of Variation (CV) for an estimated rate

See the example of the number of cells needed to count: approx. 200 so that have a margin or error of 15%.

6.2 Power – to detect Rate Ratio $RR = E_{alt}/E_0$

Exactly, using a spreadsheet or R:

Approximately, using a Gaussian approximations to $Poisson(\mu = E_0)$ and $Poisson(\mu = E_{alt} = RR \times E_0)$: solve

$$Z_{\alpha/2} \times \{E_0\}^{1/2} + Z_\beta \times \{E_{alt}\}^{1/2} = E_{alt} - E_0.$$

7 References

*Clayton D and Hill M. Statistical Models for Epidemiology Chs 4, 5, 12.2

*Walker A Observation and Inference, p13,107,154.

Armitage P Berry G & Matthews JNS [4th edition, 2002] Statistical Methods in Medical Research sections 3.7 , 5.3, 6.3.

Colton T Statistics in Medicine, pp 16-17 and 77-78.

Kahn HA, Sempos CT Statistical Methods in Epidemiology pp 218-219.

Selvin S Statistical Analysis of Epidemiologic Data Ch 5 (clustering) and Appendix B (Binomial/Poisson).

Miettinen O Theoretical Epidemiology p 295.

Breslow N, Day N Statistical Methods in Cancer ResearchVol II: Analysis of Cohort Studies pp68-70 (SMR) pp131-135; sect. 7.2 (power/sample size) Statistical Methods in Cancer ResearchVol I: Analysis of Case-Control Studies p134 (test-based CI's).

Rothman K, Greenland S [1998] Modern Epidemiology pp 234- pp404-5 (overdispersion) 570-571 (Poisson regression).

Rothman K, Boice J Epidemiologic Analysis on a Programmable Calculator.

Rothman K [1986] Modern Epidemiology.

Rothman K [2002] Introduction to Epidemiology pp 133-4 & 137-139

1−2α	*0.95*		*0.9*		*0.8*	
α	*0.025*		*0.05*		*0.1*	
count	Lower	Upper	Lower	Upper	Lower	Upper
0	0.00	3.69	0.00	3.00	0.00	2.30
1	0.03	5.57	0.05	4.74	0.11	3.89
2	0.24	7.22	0.36	6.30	0.53	5.32
3	0.62	8.77	0.82	7.75	1.10	6.68
4	1.09	10.24	1.37	9.15	1.74	7.99
5	1.62	11.67	1.97	10.51	2.43	9.27
6	2.20	13.06	2.61	11.84	3.15	10.53
7	2.81	14.42	3.29	13.15	3.89	11.77
8	3.45	15.76	3.98	14.43	4.66	12.99
9	4.12	17.08	4.70	15.71	5.43	14.21
10	4.80	18.39	5.43	16.96	6.22	15.41
11	5.49	19.68	6.17	18.21	7.02	16.60
12	6.20	20.96	6.92	19.44	7.83	17.78
13	6.92	22.23	7.69	20.67	8.65	18.96
14	7.65	23.49	8.46	21.89	9.47	20.13
15	8.40	24.74	9.25	23.10	10.30	21.29
16	9.15	25.98	10.04	24.30	11.14	22.45
17	9.90	27.22	10.83	25.50	11.98	23.61
18	10.67	28.45	11.63	26.69	12.82	24.76
19	11.44	29.67	12.44	27.88	13.67	25.90
20	12.22	30.89	13.25	29.06	14.53	27.05
21	13.00	32.10	14.07	30.24	15.38	28.18
22	13.79	33.31	14.89	31.41	16.24	29.32
23	14.58	34.51	15.72	32.59	17.11	30.45
24	15.38	35.71	16.55	33.75	17.97	31.58
25	16.18	36.90	17.38	34.92	18.84	32.71
26	16.98	38.10	18.22	36.08	19.72	33.84
27	17.79	39.28	19.06	37.23	20.59	34.96
28	18.61	40.47	19.90	38.39	21.47	36.08
29	19.42	41.65	20.75	39.54	22.35	37.20
30	20.24	42.83	21.59	40.69	23.23	38.32

Figure 7: (Exact) Confidence limits for the expectation [i.e. the μ parameter] of a Poisson random variable E.g. if observe 6 events in a certain amount of experience, then 95% CI for the mean count for this same amount of experience is (2.20, 13.06)

Figure 8: POISSON, Siméon Denis 1781-1840

from...
`http://www.york.ac.uk/depts/maths/histstat/people/sources.htm`

See also...

`http://www.encyclopedia.com/topic/Simeon_Denis_Poisson.aspx`
`http://en.wikipedia.org/wiki/Poisson_distribution`

0 Exercises

0.1 (m-s) Working with logs of counts and logs of rates

In order to have a sampling distribution that is closer to Gaussian – sample counts, and ratios of them tend to have nasty sampling distributions –we often transform from the $(0, \infty)$ scale for a count y and its expectation, μ, to the $(-\infty, 0)$ $\log[y]$ and $\log[\mu]$ scale.

Thus, we do all our inference (SE calculations, CI's, tests) on the log scale, then transform back to the count or rate (or if comparative, rate ratio) scale.

1. Suppose $Y \sim$ Poisson(μ) with associated rate estimate $\hat{\lambda} = Y/PT$[18]. Derive the variances for the random variables $\log[Y]$ and $\log[\hat{\lambda}]$. Ignore the possibility of obtaining $\hat{\mu} = 0$ i.e., $\hat{\lambda} = 0/PT = 0$.
2. What is the variance for the log of a rate ratio, i.e., $\log[\hat{\lambda}_2 \div \hat{\lambda}_1]$?

0.2 (m-s) The Poisson Family as a 'Closed under Addition' Family

Show that if $Y_1 \sim Poisson(\mu_1)$ and $Y_2 \sim Poisson(\mu_2)$ are independent random variables, then $Y = Y_1 + Y_2 \sim Poisson(\mu_1 + \mu_2)$.

0.3 (m-s) Link between Poisson and Exponential Distributions

Show that if the random times $T_1, T_2, \ldots$ between successive events can be regarded as i.i.d observations from an exponential distribution with mean μ_T, then the number Y of events in a fixed time-window of length W has a Poisson Distribution with mean or expectation $\mu_Y = W \times \lambda = W \times (1/\mu_T)$.

0.4 (m-s) Link between tail areas of Poisson and χ^2 Distributions

In section 5 of Fisher 1935, he states that 'it will be noticed' (from section 4) that, when its number of degrees n is even, the probability of the variate $\frac{1}{2}\chi^2$

[18] PT = amount of Population Time

exceeding any specified value μ is

$$e^{-\mu}\{1+\mu+\mu^2/2!+\cdots+\mu^{(n-2)/2}/[(n-2)/2]!\}.$$

From this, with $Y \sim Poisson[\mu]$, and $n = 2y$, derive a way to obtain $Prob[Y \geq y \mid \mu]$ from the cdf function of the χ^2 Distribution. From this, derive a way to obtain, from a single Poisson count y, the exact lower $\alpha/2$ and upper $\alpha/2$ limits for the mean of the Poisson Distribution it arose from.

0.5 (m-s) The Fisher information that a Poisson random variable carries about its expectation and about the log of this expectation

(Wikipedia) "The Fisher information is the amount of information that an observable random variable Y carries about an unknown parameter θ upon which the likelihood function of θ, $L(\mu) = f(Y;\theta)$, depends." The Fisher Information is defined as

$$I(\theta) = E\left\{ \left[\frac{d}{d\theta} \ln f(Y;\theta) \right]^2 \middle| \theta \right\}$$

1. Calculate the Fisher Information about the parameter μ in the case of the random variable $Y \sim Poisson(\mu)$, with

$$L(\mu) = f(Y;\mu) = \exp[-\mu] \times \mu^Y / Y!$$

2. Calculate the Fisher Information about the parameter $\theta = \log(\mu)$.

0.6 CI's for the incidence of percutaneous injuries in the various types of residencies

The authors of the NEJM paper did not say how they got the CIs for the Rates per Intern-Month, shown in their Table 1. The CI for the overall rate closely matches the large-sample one that JH has in his Notes. Apply the exact method to obtain CI's for the '3-P's', Pediatrics, Psychiatry and Pathology, where the observed numerators are all under 30.

0.7 Comparison of various CI's for the expectation, μ of a Poisson random variable, on the basis of a single count y

Fill in the blanks in the table below, and compare the accuracy of different approximations to the exact 95% CI for μ, based on a count of y.

Observe $y =$	3*	6	15	33**	78***	100
Exact CI:	?	?	?	?	?	?
Approximation						
Wilson-Hilferty	?	?	?	?	?	?
1st principles, y	?	?	?	?	?	?
1st principles, $y^{1/2}$	?	?	?	?	?	?
SE-based, y	?	?	?	?	?	?
SE-based, $\log(y)$	?	?	?	?	?	?

* Rothman 2002 p134: 3 cases in 2500 PY. ** No. of lung cancer deaths in women aged 55-60 in Quebec in 1971. ***Total number of cancers in concerned area in Alberta SourGas study.

0.8 Power Calculations

A researcher wishes to compare the numbers of new cases of a particular disease in the 'PT' Population-Time units exposed to a potentially noxious agent with the $E_0 = \mu_0 = 15.6$ that would be expected in this amount of Population Time if rates (already observed) in a LARGE unexposed experience prevailed. The researcher will use a 1-sided test with $\alpha = 0.05$ to test H_0 : $\mu_{\text{in this amount of exposed PT}} = 15.6$ vs. H_{alt} : $\mu_{\text{in this amount of exposed PT}} > 15.6$.

The amount of PT is fixed. Thus there is no point in the researcher calculating *what amount of PT would be required* so that, if $\mu_{exposed} =$ (say) $2 \times \mu_{un-exposed}$, there would be an 80% chance of obtaining a statistically significant elevation (i.e., an experience large enough to have 80% power to 'detect' a doubling of the incidence rate). Instead, the researcher decided to calculate the *power*, with the given fixed amount of PT that can be studied, to 'detect' a doubling or a tripling of the incidence rate.

Perform this power calculation. You may find it easier (and more transparent) to work with the exact Poisson probabilities (e.g. in a spreadsheet or in `R`).

0.9 Perfect Results ?

The following excerpt is from the Vaccine Arm of Table 3 of an Article in the NEJM in 2002[19]. We will look at the comparison with the Placebo arm when we get to comparative studies.

Efficacy Analyses of a Human Papillomavirus Type 16 L1 Virus-like-particle Vaccine.

	End point	HPV-16 VACCINE GROUP			
Efficacy Analysis	Type of HPV-16 Infection	No. of Women	Cases Of Infection	Woman-Yr At Risk	Rate per 100 Woman-Yr At Risk
(1)*	P.	768	0	1084.0	0
(2)**	P.	800	0	1128.0	0
(3)*	P. or T.	768	6	1084.0	0.6

(1) Primary per-protocol

(2) Including women with general protocol violations

(3) Secondary per-protocol

P = Persistent; T=transient

*The per-protocol population included women who received the full regimen of study vaccine and who were seronegative for HPV-16 and negative for HPV-16 DNA on day 0 and negative for HPV-16 DNA at month 7 and in any biopsy specimens obtained between day 0 and month 7; who did not engage in sexual intercourse within 48 hours before the day 0 or month 7 visit; who did not receive any nonstudy vaccine within specified time limits relative to vaccination; who did not receive courses of certain oral or parenteral immunosuppressive agents, immune globulin, or blood products; who were not enrolled in another study of an investigational agent; and who had a month 7 visit within the range considered acceptable for determining the month 7 HPV-16 status.

**The population includes women who received the full regimen of study vaccine and who were seronegative for HPV-16 and negative for HPV-16 DNA on day 0 and negative for HPV-16 DNA at month 7 and in any biopsy specimens obtained between day 0 and month 7.

[19]The New England Journal of Medicine Vol 347 Nov 21, 2002, p1645 A Controlled trail of a Human Papillomavirus Type 16 Vaccine. Laura A. Koutsky et al., for The Proof of Principle Study Investigators.

Questions

1. In their Statistical Methods, the authors state: "The study employed a fixed-number-of-events design. At least 31 cases of persistent HPV-16 infection were required for the study to show a statistically significant reduction in the primary end point (assuming that the true vaccine efficacy was at least 75 percent with a power of at least 90 percent). Accounting for dropouts and women who were HPV-16-positive at enrollment and assuming an event rate of approximately 2 percent per year, we estimated that approximately 2350 women had to be enrolled to yield at least 31 cases of HPV-16 infection. Although the study will continue until all women complete four years of follow-up, the primary analysis was initiated on August 31, 2001, as soon as at least 31 cases were known to have occurred. Thus, the primary analysis includes all safety and efficacy data from visits that occurred on or before that date."

   Why did the authors use a 'fixed-number-of-events' rather than 'fixed number of subjects for a fixed amount of time' design?

2. Calculate 95% 2-sided CIs to accompany the 3 point estimates of infection rate.