Resources:
• Course webpage
• section 13.2.3 (Poisson Distribution) in under-construction online book
• Some history • Gosset • Rutherford
• Clayton & Hill's Chapter 5 on (Likelihood) Inference for Rates.

CONTENTS

[1] See also: derivation & applications (counting yeast cells in beer) in Student's 1907 paper "Counting with a Haemacytometer"; Ch. from Armitage et al.; earlier versions of 'Poisson' distribution and "Randomness at the root of things: Poisson sequences".

1 (Poisson) Model for (Sampling)Variability of a 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"/"instances" 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 $E[Y] = \mu$.
- $P[y] = \exp[-\mu] \frac{\mu^y}{y!}$ `dpois():` too bad 'μ' is referred to as '`lambda`.' cf §3.
- Shorthand: $Y \sim$ Poisson(μ).
- $\sigma_Y^2 = \text{Variance}[Y] = \mu$; $\sigma_Y = \sqrt{\mu_Y}$.
- Fairly well approximated by $N(\mu, \sigma_Y = \mu^{1/2})$ when $\mu >> 10$ (see p.4).
- 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) person-time denominator "behind' the no. of "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.
- $Y \sim Poisson(\mu_Y) \Leftrightarrow T$ time b/w events $\sim$ Exponential($\mu_T = 1/\mu_Y$). See articles by physicists: here, and (Marsden & Barratt) here.
- 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. [* included in these Resources.]

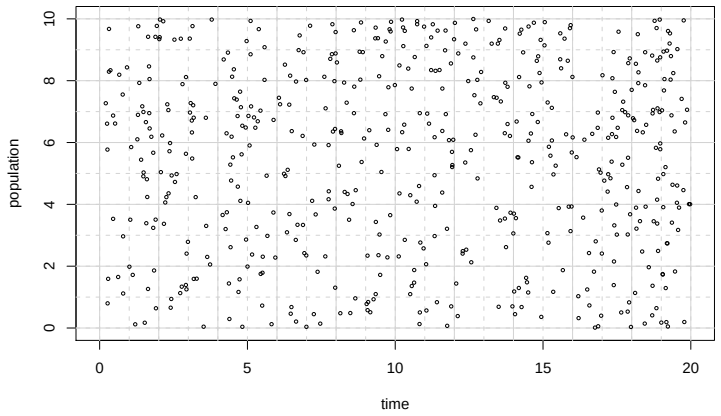


Figure 1: Events in Population-Time randomly generated from intensities that are constant within (2 squares high by 2 squares wide) 'panels', but vary between such panels. In Epidemiology, each square might represent a number of units of population-time, and each dot an event.

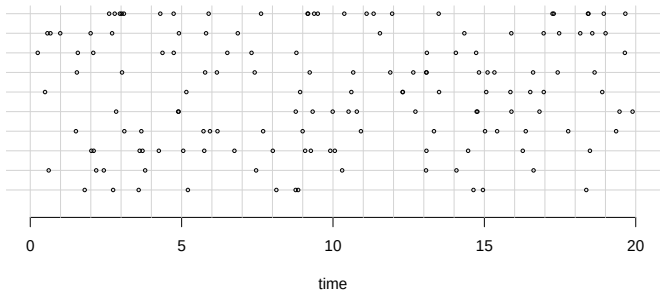


Figure 2: Events in Time: 10 examples [1/row], randomly generated from constant over time intensities. Simulated with 1000 Bernoulli($_{\text{small}}\,\pi$)'s per time unit. See earlier exercise on *Tire Ruptures*.

1.1 Does the Poisson Distribution apply to. . . ?

- Yearly variations in numbers of *persons* killed in plane crashes? [2]
- Daily variations in numbers of births?[3]
- Daily variations in numbers of deaths [extra 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[4]
- Variations across cookies/pizzas in numbers of chocolate chips/olives.
- Variations across days of year in numbers of deaths from sudden infant death syndrome. See link.

1.2 Calculating Poisson probabilities:

- Exactly
  - R: `dpois, ppois, qpois rpois` mass/cum./quant./rand
  - SAS: `POISSON`;
  - Stata: `www.ats.ucla.edu/stat/stata/faq/pprob.htm`
  - Spreadsheet — Excel function `POISSON`$(y, \mu,$ cumulative)
- Using Gaussian Approximations to distribution of y or transforms of it: described below, under Inference.
- In 'the old days' i.e. in the years BC ('Before Computers'), Poisson *tail probabilities* were often calculated using links to the tail areas of other better-tabulated continuous distributions, such as the Chi-Square distribution. See a more modern way.

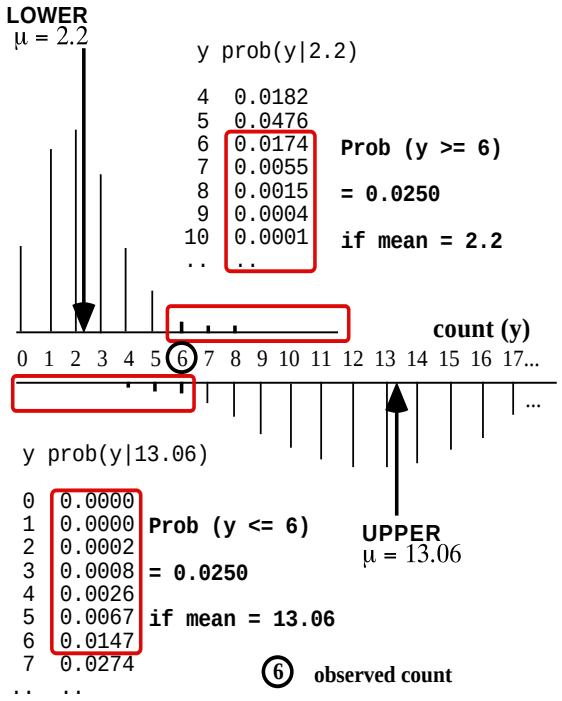
[2] Yearly variations in no.s of plane *crashes* may be a bit closer to Poisson [apart from variation due to improvements in safety, fluctuations in numbers of flights etc]. See here.

[3] See e.g. Number of weekday and weekend births in New York in August 1966; daily, in England and hourly, in USA; Variations are closer to Poisson if use *weekly* counts.

[4] c.f. Impact sanitaire de la vague de chaleur en France survenue en août 2003. Rapport d'étape 29 août2003 [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.

2 Inference re μ, based on observed count y

Instead of the usual "point-estimate ± some (z or t) multiple of standard error," a *first-principles* $100(1-\alpha)\%$ CI is the pair $(\mu_{LOWER}, \mu_{UPPER})$ such that $P(Y \geq y|\mu_{LOWER}) = \alpha/2$ and $P(Y \leq y|\mu_{UPPER}) = \alpha/2$. For example, as is shown below, the 95% CI for μ, based on $y = 6$, is $\{2.20, 13.06\}$.



For a given confidence level, there is one CI for each value of y. Each one can be worked out by trial and error, or – as has been done for the **last 80 years** – directly from the (exact) link between the tail areas of the Poisson and χ^2 distributions. These CI's – for y up to at least 30 – were found in special books of statistical tables or in textbooks. As you can check, z-based intervals (sub-section 2, next page) are more than adequate beyond this y. **Today**, if you have access to `R` (or `Stata` or `SAS`) you can obtain the first principles CIs directly **for *any* value of** y. [See next column, and Hanley, Statistics in Medicine, 2019.

However, *for those occasions where you do not have access to them or to the internet*, you can use the z-based ones – provided you can remember (or reconstruct) the formulae.

1st-principles CIs for the expectation [i.e. the μ parameter] of a Poisson random variable

Example: if observe 6 events in a certain amount of experience, then 95% CI for the mean count μ *for this same amount of experience* is 2.20 to 13.06.

y	95%		90%		80%	
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

To obtain these, we leave behind the well-worn link between the Poisson and χ^2 distributions[5], and use the natural link between the Poisson and the *gamma* distributions.[6] In `R`, e.g., the 95% limits for μ based on $y = 6$ are obtained as $\{\mu_L, \mu_U\}$ = `qgamma( c(0.025,0.975), c(6, 7) )`, or, generically, for *any* y, as $\{\mu_L, \mu_U\}$ = `qgamma( c(0.025,0.975), c(y, y+1) )`.

These limits can *also* be found using `stats::poisson.test` or (the less verbose) `survival::cipoisson` [both `R` functions use the *gamma* quantiles].

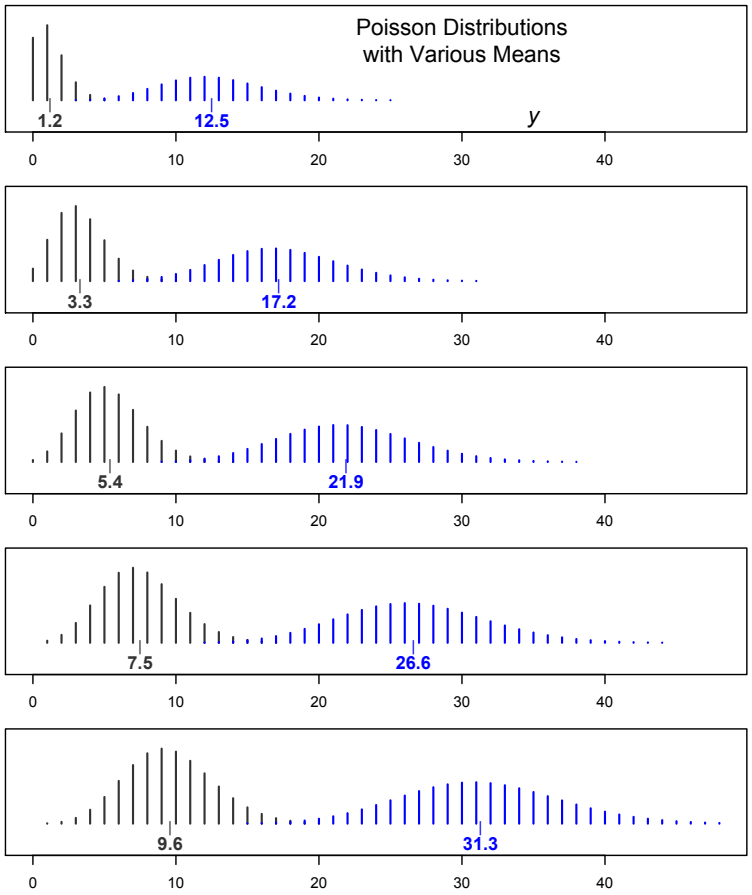
```
stats::poisson.test(6) [poisson.test(x, T=1,r = 1,..,conf.level=0.95)]
--> Exact Poisson test number of events = 6, time base = 1, p-value = 0.0006
    alternative hypothesis: true event rate is not equal to 1
    95% confidence interval: 2.20 13.06 sample estimates: event rate    6

survival::cipoisson(6) [cipoisson(k, time = 1, p = 0.95, method = c("exact", "anscombe"))]
--> lower: 2.20   upper: 13.06
```

[5] The real link is with the gamma distributions. The χ^2 ones were used because textbooks tabulated their key %-iles, whereas none tabulated the %-iles of the gamma distributions.

[6] See Links between the discrete & continuous and '*A more intuitive and modern way to compute a small-sample confidence interval for the mean of a Poisson distribution*.'

z-based confidence intervals

The Figure below shows that once μ is in the upper teens, the Poisson distributions can be reasonably well approximated by Gaussian distributions.



Thus, a 'point-estimate $\mp\, z \times SE$' CI based on SE $= \hat{\sigma} = \sqrt{\hat{\mu}} = \sqrt{y}$, is simply

$$\{\mu_L,\ \mu_U\} = y\ \mp\ z \times \sqrt{y}. \qquad [1]$$

From a single realization y of a $N(\mu, \sigma_Y)$ random variable, we can't estimate **both** μ and σ_Y: for a SE, we would have to use *outside* information on σ_Y. In the Poisson(μ) distrn., $\sigma_Y = \mu^{1/2}$, so we calculate a "model-based" SE.

How large a y? Above, when $\mu > 5$, the distr^{n} isn't 'crowded' into the corner; the lower tail of the Gaussian approx^{n} doesn't spill over the 0 boundary.

A more principled z-based way, à la Wilson.: Use the $\mu_L \to y \leftarrow \mu_U$ reasoning rather than the $\mu_L \leftarrow y \to \mu_U$ one: i.e., $Y \simeq N(\mu,\ \sigma = \mu^{1/2})$.

Thus, if the the *lower* limit ('y is an over-estimate' scenario) is μ_L, then the observed y is at the upper (say 97.5) percentile,

$$y = \mu_L + z_{975} \times \sqrt{\mu_L}.$$

Solving this (quadratic in $\sqrt{\mu_L}$) equation and back-transforming to μ_L yields

$$\mu_L = \left\{ \frac{-z_{975} + \sqrt{z_{975}^2 + 4y}}{2} \right\}^2.$$

Likewise, if the *upper* limit ('y is an under-estimate' scenario) is μ_U, then the observed y is at the lower (say 2.5) percentile,

$$y = \mu_U + z_{025} \times \sqrt{\mu_U}.$$

Solving this (quadratic in $\sqrt{\mu_L}$) equation for μ_L yields

$$\mu_U = \left\{ \frac{-z_{025} + \sqrt{z_{025}^2 + 4y}}{2} \right\}^2.$$

Together, we have, in the `R` language,

```
((-qnorm(c(0.975, 0.025)) + sqrt(qnorm(c(0.975, 0.025))^2 + 4*y))/ 2)^2
```

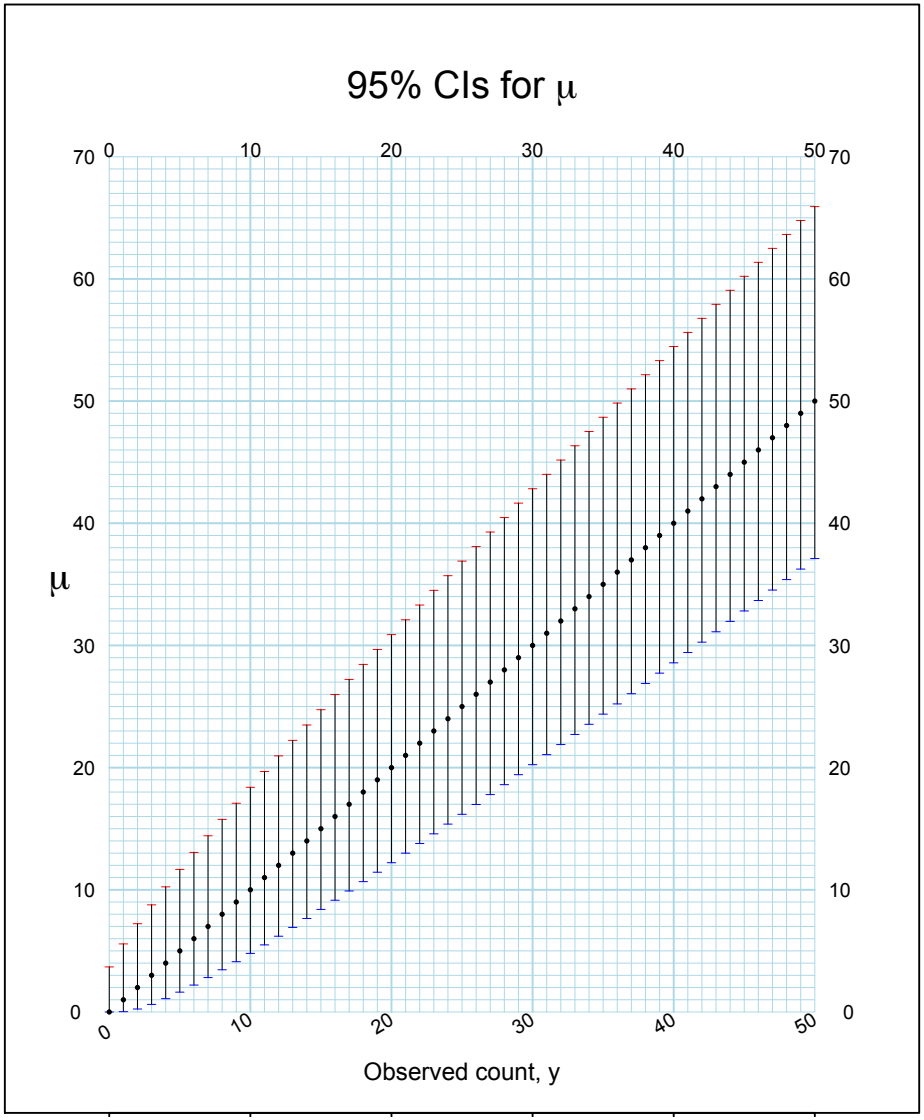
At $y = 30$, this yields the still-slightly-non-symmetric[7] 95% interval 21.0 to 42.8, just slightly narrower than the (already somewhat conservative) 20.2 to 42.8 obtained earlier.[8] Note that differences tend to be greater at the μ_L end, where the Poisson distribution is not as close to Gaussian as it is at the μ_U end. This also emphasizes that any 'rules of thumb' about when the Gaussian distribution provides a good approximation to the Poisson distribution should refer to the purpose/focus. If the focus is on a (null) hypothesis test of $\mu = \mu_0$ say, then the concern is with the shape of the Poisson distribution with mean μ_0. If the focus is on a CI, then the concern is more the shape at the lower limit μ_L, not with the shape when $\mu = y$.

Bayesian interval, with Jeffreys' prior

$$\{\mu_L, \mu_U\} = \texttt{qgamma(c(0.025,0.975), y+0.5)}.$$

[7] The fast but unprincipled 'point.est $\mp$ SE' gives $30 \mp 1.96 \times \sqrt{30} = 19.3$ to 40.7.

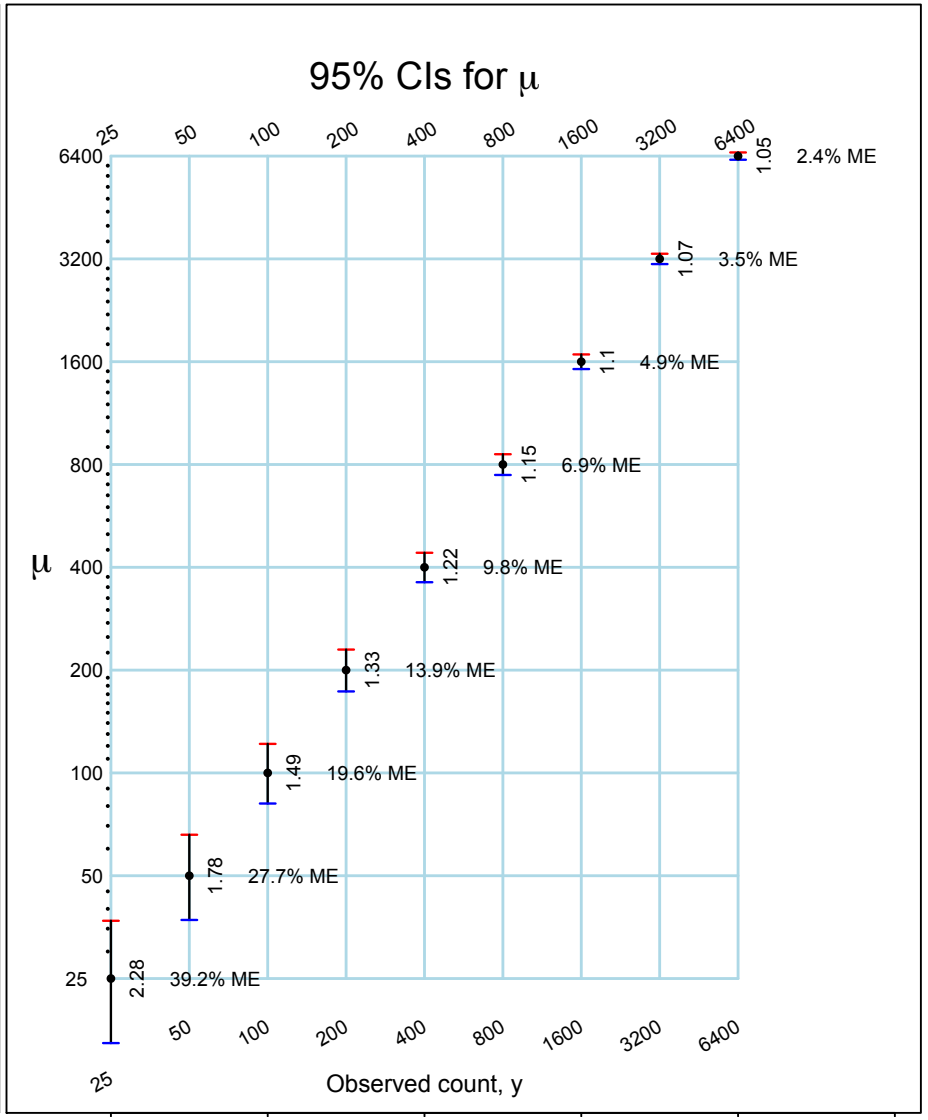
[8] As told by JH in Statistics in Medicine, 2019, that 'earlier' method that uses the Poisson tail probabilities themselves, rather than a Gaussian approximation to them, dates back to Garwood, *Fiducial Limits for the Poisson Distribution*, Biometrika, Volume 28, Issue 3-4, 1 December 1936. and to Fisher's 1935 paper.



The panel above shows 95% CI's for the mean, μ, of a Poisson distribution, based on an observed count y, ranging from 0 to 50.

Note the greater symmetry with larger y.

The nomogram can also be read horizontally, i.e., $\mu \rightarrow y$.



This panel shows 95% CI's for the mean, μ, of a Poisson distribution, based on a count y, ranging from 25×1, 2, 4, 8, 16, 32, 64, 128, and 256, i.e., from 25 to 6400. With these large y's, the percentage margins of error are 100 $\times 1.96 \times \frac{\sqrt{y}}{y} \approx \frac{200\%}{\sqrt{y}}$. Also shown is the upper/lower ratio. Thus, if say $y = 25$, the upper limit is more than double the lower limit (ratio 2.28). if $y = 400$, the uncertainty is 1.22 'fold' ('fois'). i.e., μ_U is approximately 1.22 times μ_L.

3 Inference re an event rate parameter λ, based on observed number of events y in a known amount of population-time, PT

Some writers refer to the parameter λ as Incidence Density ('ID'). For the origin of this term, see Section 3.1 of the groundbreaking 1976 article:

> 3.1. The parameters. *Incidence density ("force of morbidity" or "force of mortality") – perhaps the most fundamental measure of the occurrence of illness – is the number of new cases divided by the population-time (person-years of observation) in which they occur.*

So far, we have focused on inference regarding μ, the expected **number** of 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 or incidence density of $211/(232{,}978\text{WY}) = 0.00091/\text{WY}$ or **91** per 100,000WY. This makes it easier to compare the rate with the rate in the same age group in 1971, namely 33 lung cancer deaths in 131200WY, or $0.00025/\text{WY} =$ **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.

3.1 CI for the rate parameter λ

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

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

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

$$\boxed{\text{CI for } \lambda = \{\text{CI for } \mu\} \div \text{PT}.} \quad (1)$$

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

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

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

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 a non-symmetric first principles CI for μ. Our table of such CIs stops at $y = 24$, so we will use the

$$\{\mu_L, \mu_U\} = \texttt{qgamma(c(0.025,0.975), c(33,34))}$$

`R` function with a count of y=33. It yields lower and upper limits of 22.7 and 46.3 for μ. Thus, to accompany the point estimate of 25 deaths per 100,000WY, we have

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

We get the same limits with `stats::poisson.test(x=33,T=131200)` and `survival::cipoisson(k=33,t=131200)`

3.2 Test of $H_0: \mu = \mu_0$, i.e. $\lambda = \lambda_0$, & inference re. Standardized Incidence Ratios (SIR's), SMR's etc

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

Example: Cancers in area surrounding the Douglas Point and Pickering nuclear stations in Ontario [full story below]:

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

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

To get the CI for the SIR, divide the CI for Douglas Point μ_{DP} by the null $\mu_0 = 0.57$ (Ontario scaled down to the same size and age structure as Douglas Point.) We treat it as a constant because the Ontario rates used in the scaling are measured with much less sampling variability that the Douglas Point ones.

From page 3 of the notes, the $y = 2$ cases translates to a 95% CI for μ_{DP} of 0.24 to 7.22, so the CI for the SIR is 0.24/0.57 to 7.22/0.57 or 0.4 to 12.7.

We can *trick* `stats::poisson.test` or `survival::cIpoisson` to get the same CI by putting time as 0.57, i.e., `stats::poisson.test(x=2,T=0.57)`, `stats::poisson.test(k=2,t=0.57)`

```
data:  2 time base: 0.57
number of events = 2, time base = 0.57, p-value = 0.11
alternative hypothesis: true event rate is not equal to 1 [*]
95 percent confidence interval: 0.42 12.67
sample estimates:
event rate 3.5    [* Using T=0.57 means 3.5 is the point estimate of the rate RATIO]

The OTHER way is to use c(2,5700) for the counts, and c(1,10000) for the times:
 poisson.test(c(2,5700),c(1,10000))  Comparison of Poisson rates

data:  c(2, 5700) time base: c(1, 10000)
count1 = 2, expected count1 = 0.57014, p-value = 0.1122
alternative hypothesis: true rate ratio is not equal to 1
95 percent confidence interval: 0.42 12.68
sample estimates: rate ratio 3.5

**************
```

At the **Pickering** generating station, the ***O***bserved number was **18**, versus an ***E***xpected of **12.8**, for an *SIR* of 1.4.

Exercises: Calculate the *uppertail* P-value i..e. $P = \sum_{y \geq 18} \text{Prob}[y \,|\, \mu_0 = 12.8]$, and compute a 95% CI for the SIR.

Comment on the newspaper headline (see section 6) , and what re-wording or other emphasis Dr Clarke might have suggested, had the reporter given her a chance to see the completed story before it was published.

4 Generalized linear (regression) models for counts and event rates

4.1 For counts

• **First** with Poisson error but **identity** link.

Example with `y = 24`, using the `glm` function in `R`:

```
summary( glm(y  ~  1, family = poisson(link=identity) ) )

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   24.000      4.899   4.899 9.63e-07 ***
```

Note the ('Wald') `Std.  Error (SE)` = namely $\sqrt{24} = 4.899$, calculated at the point estimate y, namely $\sqrt{24} = 4.899$..

• **Second** with Poisson error but the default **log** link.

This default fits a linear model to $\log(\mu)$, so in this simple case, with no regressor variables, we are just fitting the constant $\log(\mu)$.

```
summary( glm(y  ~  1, family = poisson ) )

coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.1781     0.2041   15.57   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

3.1781 is the estimate of $\log(\mu)$, i.e., $\widehat{\log(\mu)} = 3.1781$, so $\widehat{\mu} = \exp(3.1781) = 24$.

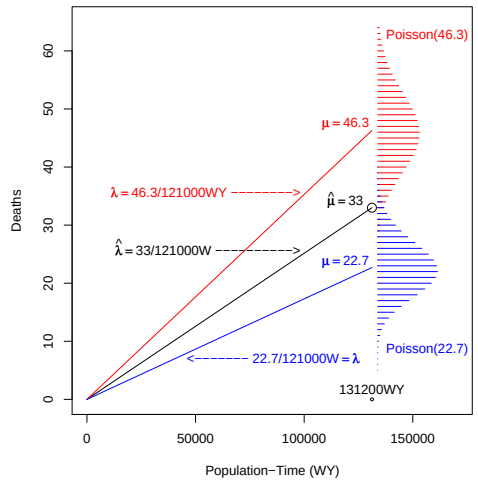
What statisticians call the "Delta Method" yields the approximate SE for a function of the random variable $\widehat{\log(\mu)}$, assuming the Prob$[Y = 0]$ is negligible.

$$SE[\widehat{\log(\mu)}] \approx \text{SE}[\widehat{\mu}] \times \{(d \log x/dx)|_{x=\widehat{\mu}}\} = \sqrt{\widehat{\mu}} \times (1/\widehat{\mu}) = 1/\sqrt{\widehat{\mu}} = \sqrt{1/y}.$$

Thus the ('Wald') `Std.  Error (SE)` for $\widehat{\log(\mu)}$ is $\sqrt{1/24} = 4.899$, calculated at the point estimate y, namely $\sqrt{1/24} = 0.2041$.

We will make a lot of use of this SE, especially for the variance (SE^2) 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 the logs of two rates).

4.2 For event rates



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, $PT = 0\ WY$) and the point ($Y = 33, PT = 121300\ WY$). 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$[Y \leq 33 \,|\, \mu = 46.3) = 0.025$.

This Figure is a simple mathematical reversal of the fundamental epidemiological 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.,

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

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

$$\mu = \lambda \times \text{amount of population-time.}$$

or, if we take logs, with the log of the product as the sum of the logs,

$$\log(\mu) = \log(\lambda) \times \underline{1} + \underline{1} \times \log(\text{amount of population-time}).$$

This re-statement has 2 important implications

(i) ***in epidemiology, we are students of rates*** and

(ii) Generalized ***Linear Models (GLMs) allow us to fit regression 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 rate models are the additive and multiplicative forms:-

$$\textbf{\textit{rate}}[x] = \beta_0 + \beta_1 x \quad \& \quad \textbf{\textit{rate}}[x] = \exp(\beta_0 + \beta_1 x).$$

• This code fits the simple rate model (just 1 fitted constant, λ) to the data above, (the '-1' removes the β_0, or 'intercept', term: if 0 time, 0 deaths!. '`link = identity`' means fit μ itself)

```
y  =     33; PT = 131200

fit.mu = summary(glm(y ~ -1 + PT, family = poisson(link=identity) ))
fit.mu

Coefficients:
    Estimate Std. Error z value Pr(>|z|)
PT 2.515e-04  4.378e-05   5.745 9.22e-09 ***

round(10^5*qnorm(c(0.025,0.50,0.975),     #
         mean = fit.mu$coefficients[1,1],
         sd   = fit.mu$coefficients[1,2]),1)

[1] 16.6 25.2 33.7
```

The fitted constant ($25.1/10^5$) is the fitted rate, with SE = $4.38/10^5$.

• $\log(\mu) = \log(\lambda) \times \underline{1} + \underline{1} \times \log(\text{PT}) = \beta_0 \times x_0 + \beta_1 \times x_1$, where $\beta_0 = \log(\lambda)$, $x_0 = 1$, $\beta_1 = 1$, and $x_1 = \log(PT)$. But β_1 is known to be 1, and should be fitted as such, and so $x_1 = \log(PT)$ has special status, namely an 'offset' variable.

This code fits the simple log rate model (just 1 fitted constant) to the data above, (the '`1`' says fit $\log(\lambda)$ as the 'intercept' in the $\log(\mu)$ model. The (default) '`link = log`' means fit the $\log(\mu)$ model. Specifying that `log(PT)` is an '`offset`' *sets* its accompanying regression coefficient to 1.)

```
> fit.log.mu = summary( glm(y ~ 1+ offset(log(PT)), family = poisson))
> fit.log.mu
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.2880     0.1741  -47.61   <2e-16 ***

round(                          # *** CI in 'log rate' scale -> CI in rate scale    ***
 10^5 * exp(                    # *** The 'exp' function is the 'anti-log' function ***
  qnorm(c(0.025,0.50,0.975),
        mean = fit.log.mu$coefficients[1,1],
        sd   = fit.log.mu$coefficients[1,2])
   ),1)

[1] 17.9 25.2 35.4
```

The fitted constant (-8.28) is the (natural) log of the fitted rate, with SE = 0.1741 = $\sqrt{1/33}$.

5 Bootstrap CIs for the μ (or λ) that generated a event-count y in a given amount of experience, PT

Example, $y = 24$ 'events' (needle-stick injuries) in PT = 2159 Intern-Months (IM) of experience. The point-estimate for λ is 24 events per 2159 IM, or 0.0111 per IM, or 1.11 per 100 IM. (see data in §6.3.)

Even though it looks like we have only 1 'observation', we can split up the overall experience into a large number (N) of very small segments, each so small that it has at most 1 event (most will have 0 events).

In the diagram on next page, each rectangle represents a segment, and the total area represents 2159 IM.

So, we can make a long vector (`e` say) containing 24 1's and the rest 0s.

We can then bootstrap the sum of the entries:

```
bootstrap <- do(4000) * sum( resample(e) )
```

As you can see (in black at the bottom right of the Figure), the most likely sum is 24, but it was sometimes as few as 9 or 10 or as many as 40 or so. The 2.5%- and 97.5%-iles, obtained from `quantile(bootstrap$sum, probs=c(0.025, 0.975))` are 17 and 31.

So, can we (and should we) use 17/2159 and 31/2159 as a 95% CI for λ?

One limitation is that the bootstrap simply reproduces (to with Monte Carlo precision) the Poisson distribution with $\mu = 24$. **But** we know that the true rate (in an infinite amount of experience) is not limited to some integer divided by 2159 - the possible values for λ should be allowed to be on a continuum. λ should not be measured by a ruler with discrete values only.

We could do better, if instead of relying on the single sampling distribution based on the point estimate, we follow Wilson's 'more principled' approach. Remember that he would see the $y = 24$ as having been generated by (but at the 'edge of') a Poisson distribution with a mean, μ, to the right/left of the 24.

But if μ is above the 24 (or, as Wilson saw it, 24 is below μ), the sampling distribution will have a different shape than if μ is below the 24,

We need to put μ just far enough above that the 24 is at its 2.5%-ile. We can do this using the theoretical gamma-Poisson link, BUT one can also locate it by trial and error, using nothing more than re-sampling.

We could put say 33 (or 34, or 35, ..) events in the long vector representing 2019 IM's, and bootstrap each one, and pick the one for which the 2.5%-ile is 24. And so the same for μ below 24.

But how to use μ steps smaller than integers?

Instead of sampling 2157 IM from a vector representing 33 events in 2157 IM, why not sample 2157 IM (the same amount of IM that generated the observed data) from a (say) 100-times-longer vector (`E`, say) representing 3347 events in 215700 IM, i.e., from a vector that represents 33.47 events per 2157 IM?

It turns out that the `mosaic::resample` function allows you to sample 2157 from a larger universe, using the `size=` argument.

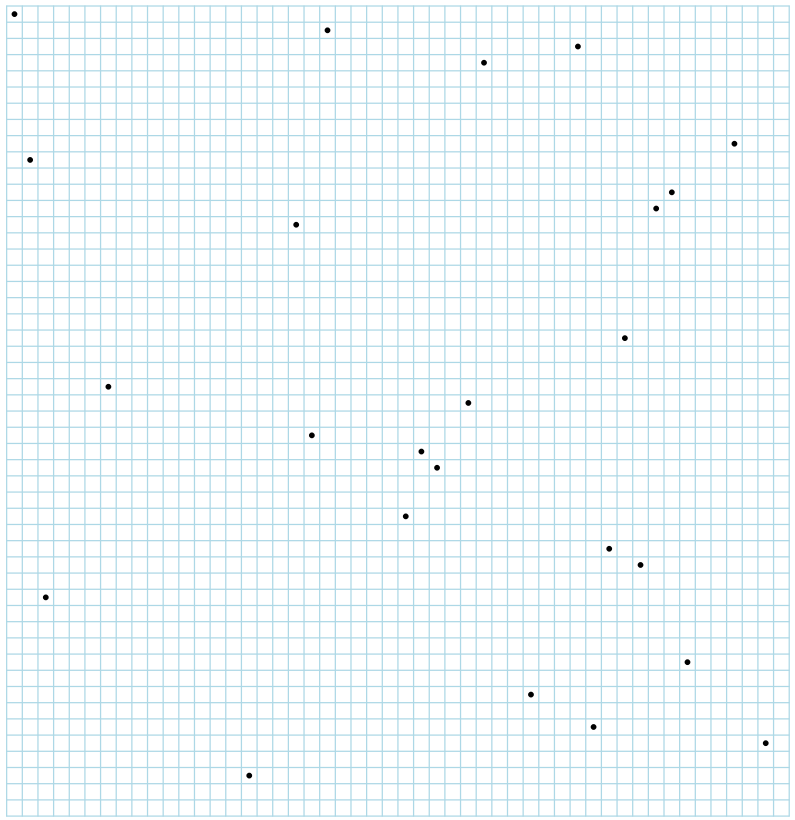
```
bootstrap = do(1000)*sum(resample(E,size=2157))
```

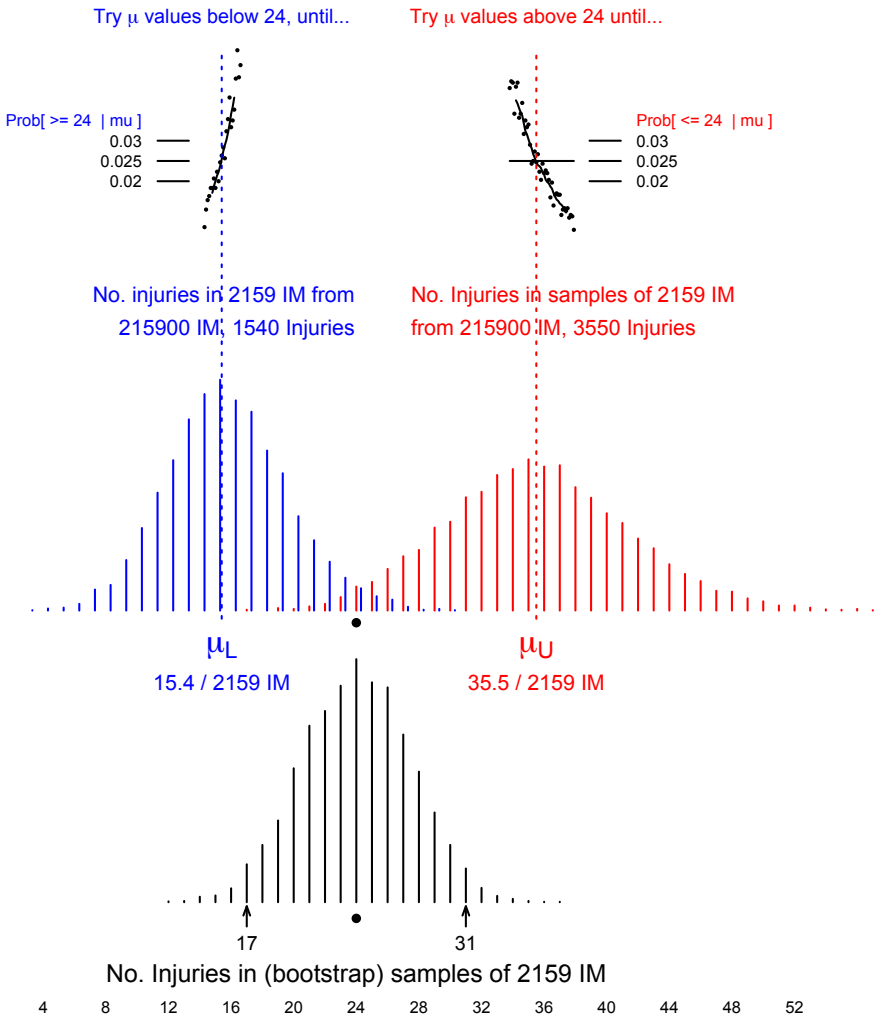
The μ_U value of 35.5 in the Figure was found by continuing to increase μ (to change the composition of `E`) until the 24 was at the 2.5%-ile. The μ_L value of 15.4 was found similarly by continuing to move μ to the left until 24 was at the 97.5%-ile.

If we use sufficient bootstrap samples and a fine enough search, this approach reproduces the limits based on the two Poisson distributions.

This example emphasizes why Wilson's $\mu_L \rightarrow y \leftarrow \mu_U$ approach is superior to the unprincipled 'point-estimate ± SE' or $\mu_L \leftarrow y \rightarrow \mu_U$ approach, especially if the sampling distribution is different at the lower and upper limits for the parameter being aimed at.

It also shows how, with a bit of extra computing time, we can compute limits for an event rate without explicitly invoking the Poisson distribution *per se*.





Computing limits for an event rate without explicitly invoking the Poisson distribution *per se*: The sample experience is split into a large number (N) of very small segments, each so small that it has at most 1 event (most will have 0 events): a long vector containing 24 1's and the rest 0s.

We can then bootstrap the sum of the entries, to get the (not quite satisfactory) sampling distribution shown in black. Thus, ...

we (again) follow Wilson's (trial-and-error) $\mu_L \rightarrow y \leftarrow \mu_U$ approach, but – to allow μ to vary on a continuum – we use small μ steps by sampling 2157 IM from (say) a 100-times-longer vector representing 3347 events in 215700 IM, i.e., 33.47 events per 2157 IM.

Note that the **shapes** and **widths** of the sampling distributions at the **lower** and **upper** limits for μ are **different**.

6 Applications, and Notes

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

An estimate of the White Blood Cell (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 [as one can derive from trial and error] a total count of 200 (marked by ↑ below) could be a *high* reading from a concentration which produces 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...`

`.....`μ_L`.....`↑`.....`μ_U`.....`

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

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

Can do this by trial and error in `R` using various counts (y): e.g. 120

```
y = 120; L=qgamma(c(0.025,0.975),c(y,y+1));
paste(toString(round(100*L/y)),"%")
```

† As told in our history of what we now know as the 'Poisson' distribution, "To help users of a new Zeiss microscope that allowed one to count the number of blood cells in a sample, in 1879, Abbe determined the number of cells one needed to count in order to achieve a sufficiently small 'probable error' for an estimated blood cell concentration." In 1905, Schweidler provided error calculations for counting radioactive transformations. At the Guinness company in 1907, Gosset ('Student') was also concerned with the precision of 'counting' statistics.

6.2 Leukemia Rate Triples near Nuke Plant: Study

OTTAWA (CP)[9] - 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[10], a larger sample size could place the true figure somewhere in the range from 0.4 cases to 12.6 cases.[11]

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.[12]

Comment [JH]: 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.

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

[10] $SIR = 3.5 = No.Observed/No.Expected$. 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

[11] 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.

[12] $SIR = 1.4 = O/E; CI = (CI$ 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.

6.3 Risk of Self-reported Percutaneous ('Needle stick') Injuries in Interns

In Relation to Extended Work Duration (link)

Context In their first year of postgraduate training, interns commonly work shifts that are longer than 24 hours. Extended-duration work shifts are associated with in- creased risks of automobile crash, particularly during a commute from work. Interns may be at risk for other occupation-related injuries.

Objective To assess the relationship between extended work duration and rates of percutaneous injuries in a diverse population of interns in the United States.

Design, Setting, and Participants National prospective cohort study of 2737 of the estimated 18 447 interns in US postgraduate residency programs from July 2002 through May 2003. Each month, comprehensive Web-based surveys that asked about work schedules and the occurrence of percutaneous injuries in the previous month were sent to all participants. Case-crossover within-subjects analyses were performed.

Main Outcome Measures Comparisons of rates of percutaneous injuries during day work (6:30 AM to 5:30 PM) after working overnight (extended work) vs day work that was not preceded by working overnight (nonextended work). We also compared injuries during the nighttime (11:30 PM to 7:30 AM) vs the daytime (7:30 AM to 3:30 PM).

Results From a total of 17 003 monthly surveys, 498 percutaneous injuries were re- ported (0.029/intern-month). In 448 injuries, at least 1 contributing factor was reported. Lapse in concentration and fatigue were the 2 most commonly reported con- tributing factors (64% and 31% of injuries, respectively). Percutaneous injuries were more frequent during extended work compared with nonextended work (1.31/1000 opportunities vs 0.76/1000 opportunities, respectively; odds ratio [OR][13], 1.61; 95% confidence interval [CI], 1.46-1.78). Extended work injuries occurred after a mean of 29.1 consecutive work hours; nonextended work injuries occurred after a mean of 6.1 consecutive work hours. Injuries were more frequent during the nighttime than dur- ing the daytime (1.48/1000 opportunities vs 0.70/1000 opportunities, respectively; OR, 2.04; 95% CI, 1.98-2.11).

Conclusion Extended work duration and night work were associated with an increased risk of percutaneous injuries in this study population of physicians during their first year of clinical training. JAMA. 2006;296:1055-1062 www.jama.com

[13]This is **not** an odds ratio. It is a Rate ratio or an Incidence Rate ratio. Injuries occur in *person-time*. Odds are transforms of probabilities, and refer to people.

(part of) Table 1. Rates of Percutaneous Injuries by Residency Program.

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 $\{498 \mp 1.96 \times 498^{1/2}\} \div 17003 = \{0.0267, 0.0318\}$.

Exercise: Try to match the others to the methods described above.[14]

[We will come back, in the section of the course dealing with comparisons, to the issue of *event-rate differences* and *event-rate ratios*, and we will dispel the notion that one needs to invoke the **odds** ratios mentioned in the Results section.]

[14]If you don't succeed, try page134 in Chapter 7 in Rothman's 2002 edition of Epidemiology: An Introduction.

The one Rothman adapts from 'D. Byar (unpublished)' has at its core method No. 2 ('Wilson-Hilferty' [15]) among the 19 (!!) CI's described in this investigation: "Comparison of confidence intervals for the Poisson mean: some new aspects" REVSTAT – Statistical Journal Volume 10, Number 2, June 2012, 211?227, which can be found here. `https://www.ine.pt/revstat/pdf/rs120203.pdf`

Rothman's 2nd Edition, 2012, omits this CI method, and limits his presentation to the $y = 8$ example.

6.4 "Cluster of Events" Newspaper Stories

1. In the *Montreal Gazette* in April 25, 1989[16]

> 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 also 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

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

or as `1-ppois(5,lambda=1.3)`

Thus, the (computed before the fact) 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 difficult. If the newswires scanned a large number of schools 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. See also the 'Texas Sharpshooter' refererence in 2nd column p 441 of this article.

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.*

[16] See 'Births Case #3' in Hanley J, the American Statistician, 1992, "Jumping to coincidences: defying odds in the realm of the preposterous"

.

2. In the *Montreal Gazette*, week of May 8 , 1991

> Double trouble Down Under
> *(caption to a photograph showing five sets of twins)*
>
> It was a very busy week in the obstetrics department of Baulkham Hills Private Hospital in Sydney Australia, as five mothers gave birth to twins. Hospital officials offered no explanation of the sudden run of multiple births, but the proud mothers are happy to pose with their infants. Everyone is doing well.

!!

3. In the *Montreal Gazette*, week of May 15 , 1991...

> No Double trouble anywhere this week!
> no photograph, no twins!

4. In the *Montreal Gazette*, week of May 22 , 1991...

> No Double trouble anywhere this week!
> no photograph, no twins

!!!

5. But, ... stay posted..!

7 Planning: Sample Size for CIs and Tests

7.1 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 number of events**. So it is safer to specify sample size in these terms.

7.2 Amount of experience required 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%.

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

Exactly, using `qpois` and `ppois` in `R`:

(a) use `qpois` to establish the (smallest) number of cancers y^* would be (just be) 'statistically significantly higher' than the null $\mu = E_0$.

(b) use `qpois(y*,E0 × RR,lower.tail=FALSE)` to calculate the probability, if $\mu_{alt} = RR \times E_0$, of observing a statistically significant elevation.

Approximately, using Gaussian approximations to the sampling distributions $Poisson(\mu = E_0)$ and $Poisson(\mu = E_{alt} = RR \times E_0)$, by solving

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

(use $Z_{\alpha/2}$ if it is to be a 2-sided alternative). Note: if the power is more than 50%, $E_{alt} >$ critical value y^*, and Z_β will be a negative quantity, but the two absolute distances $Z_\alpha \times \{E_0\}^{1/2}$ and $|Z_\beta| \times \{E_{alt}\}^{1/2}$ will add to the positive quantity $E_{alt} - E_0$. We will switch the sign of Z_β later.

Substituting $RR \times E_0$ for E_{alt} , we get

$$Z_\alpha \times \{E_0\}^{1/2} + Z_\beta \times RR^{1/2} \times \{E_0\}^{1/2} = E_0[RR - 1],$$

or

$$Z_\beta \text{ (with sign corrected)} \; = -\frac{E_0^{1/2} \times [RR - 1] - Z_\alpha}{RR^{1/2}}.$$

(Working with Poisson distributions avoids issues of signs; if you *do* use Normal approximations, *draw a diagram* to get the signs for Z_α and Z_β correct).

POWER OF (1-sample) Poisson-based test of E_null versus E_alt

jh 2006.03.18

supply E_null, alpha level, E_alt

		n		**20**	**Rate**
		E_null		**4.5**	**Ratio**
		E_alt		**13.5**	**3.00**
2-sided		**alpha**		**0.050**	
					POWER
	y	<- # Events in the amount of Experience studie			**0.865**
↓ critical region					
↓	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	

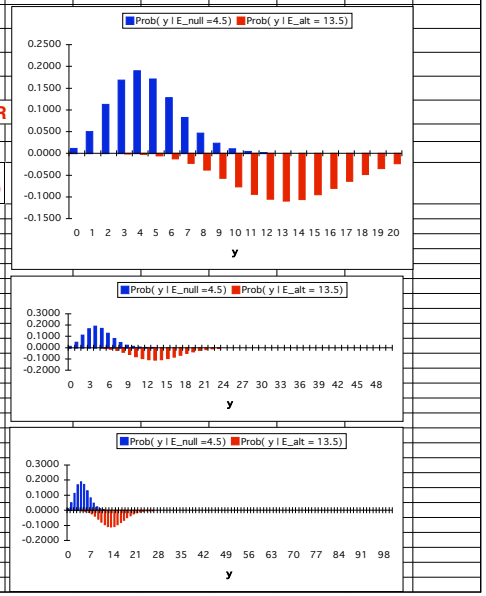


Figure 3: Using exact Poisson Probabilities [see 'Power for test of E = E_{null} vs E = E_{alt}: Excel worksheet' here. It could be made much nicer with `R` or with a `shiny` app.

8 From event-rates to risks

We can use the observed injury *rate* in Obstetrics/gynecology to derive the 1-year risk (probability) of suffering an injury.

Note that risk refers to the probability for an individual, estimated using a 12-month cumulative incidence - a proportion. It assumes the person is 'available' (at risk of being injured) for the full year – or at least until the injury.

See Rothman, 2nd Ed, 2012, page 38-39.

For simplicity, assume that individuals are subject to this constant[17] force of 0.0975 injuries per inter-month throughout the 12 months.

Hint: the answer is not 12 $\times 0.0975 = 1.17$, as some previous students calculated. It is 0.69. See Rothman's Epidemiology: An Introduction, 2ndE 2012, Chapter 4.

It may help to use Edmonds' concept of a 'person-year' as *1 person (not necessarily the same person) constantly at risk*, He used the term '*a given number of persons constantly living*' when originating the term 'force of mortality.'[18]

If you treat an intern-year as 3000 working hours (h), you could also write the incidence density (or λ) as 0.00039 percutaneous injuries h^{-1}, or so many per intern-week, or intern-decade, or intern-century!

As Edmonds did, take the 'given number of interns' to be one (1). Imagine a 'chain', starting at $t' = 0$ and extending for 12 months until $t'' = 12$. The chain is begun with a randomly selected never-injured intern, who continues until he/she either reaches 12 months or is injured before then. If the latter, and it occurs at age t, he/she is immediately replaced by another randomly selected never-injured intern. The chain proceeds, 'with further replacements as needed,' until $t'' = 12$. From t' to t'', the 1 constantly-serving candidate constitutes a *dynamic population* with a constant membership of 1.

The number of replacements required is a random variable, with possible val-

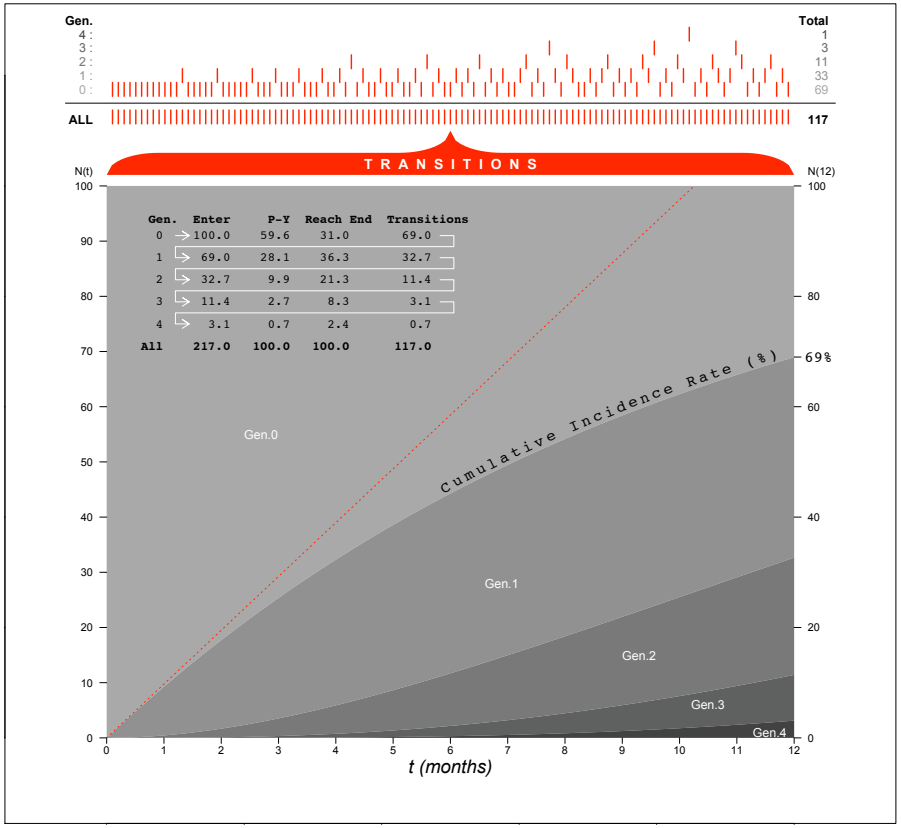


Figure 4: An average of 1.17 transitions (percutaneous injuries) in 1 intern-year (I-Y) of experience (117 in 100 I-Y), so that ID = 1.17 $year^{-1}$. 100 'chains' start at $t = 0$ (the 100 chains are represented by 100 horizontal lines, so close to each other that the total person time appears as a rectangle 100 interns high by 12 months wide); each chain continues for 12 months, each using as many replacements (Gen. 1, 2, ...) as necessary to complete the chain. The different shaded areas represent the population-time for generations 0, 1, ... The proportion of chains that are completed using the initial (Gen. 0) intern is $\exp[-1.17] = 0.31$, i.e., 31%, so the 1-year risk is 100% - 31% = 69%. The proportion of chains in which, by time t, the initial (Gen. 0) intern has been replaced, i.e., the cumulative incidence rate up to time t, is $1 - \exp[-ID \times t] = 1 - \exp[-(\text{integral up to time } t)]$ The straight line (the product of ID and time, scaled up by 100) involves a constant number of candidates at each time point, and thus overestimates the cumulative incidence rate – substantially so as generation 0 is replaced. The numbers of transitions do not sum exactly to 117 because of rounding.
The gen0, gen1, gen2, ... fractions 'still there' at 12 months are the Poisson probabilities of 0, 1, 2 ... 'replacements' when the replacement (injury) rate is $\lambda = 0.0975$ per intern-month, i.e., in an intern-year, $\mu = 1.17$. `round(dpois(0:5,1.17),2):` `0.31 0.36 0.21 0.08 0.02 0.01`

[17] See `http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/SurvivalAnalysis/IncidenceFunctionToRisk2018.pdf` for a more comprehensive treatment of the 'exponential' formula linking a rate function and risk, including cases where the event-rate function, $\lambda(t)$, varies considerably over the time/age span – one such rate-function is the force of human mortality over the lifespan).

[18] Edmonds, T. R. (1832) The Discovery of a Numerical Law regulating the Existence of Every Human Being illustrated by a New Theory of the Causes producing Health and Longevity. London: Duncan. Available as an on-line digital version at url-http://books.google.com.

ues 0, 1, 2, ... Its expected value (mean) is $\mu = 0.0975\ m^{-1} \times 12\ m = 0.00039\ h^{-1} \times 3000\ h = 1.17$ first injuries. Readers will recognize μ as integral of the $ID(t)$ function over the 12-month age-span. The probability that the chain is *completed by the same intern who initiated it* is the probability that *0 replacements are required*. The probability that it is not is the complement of this 'survival' probability. Since the number of replacements (transitions, first injuries) in the 12 months is a Poisson random variable, the probability that the chain *is* completed by the same intern who initiated it is the Poisson probability of observing 0 events when 1.17 are expected, i.e., as $\exp[-1.17] = \exp\left[-\int_{t'}^{t''} ID(t)dt\right] = 0.31$. The probability that this intern fails to complete the chain, i.e., *is injured before the 12 month period ends* is $1 - \exp\left[-\int_{t'}^{t''} ID(t)dt\right] = 1 - 0.31 = 0.69$. Thus the *12-month risk* of injury is 69%.

Fig 4, modeled after Fig 1 in Miettinen,[19] gives the expected values for a total of 100 separate such chains, and shows why *the product of ID and time (the 1.17, the integral) is not a risk, but rather an expected number of events (transitions, turnovers, injuries) in a dynamic population of size 1.* To provide 100 intern-years of service, an average of 217 interns is required. Of the 100 who began the chains (the average service of these 100 in 'generation 0' is 0.596 P-Y per intern) 31 complete them and 69 do not. Thus, the 12-month risk is 69%. On average, of their 69 replacements (generation 1), 36 complete the chains and 33 do not; and so on, so that in all – over the initial and replacement generations, totaling 100 P-Y – 117 do not and 100 do.

The proportion of chains in which, by time t, the 'Gen. 0' intern has been replaced, is $1 - \exp[-ID \times t] = 1 - \exp[-(\text{integral up to time } t)]$. The straight line (the product of ID and time, scaled up by 100) involves a constant number of candidates at each time point, and thus overestimates the cumulative incidence rate – substantially so as 'gen. 0' is replaced.

Table 4.2 & Fig. 4.3 in Rothman's 2nd Edition show a 20-year incidence proportion, but using an ID of 0.011 yr^{-1}, so the expected number of transitions in a dynamic population of 1 is $0.011\text{yr}^{-1} \times 1\ \text{yr} = 0.22$. His curve is identical to the first $0.22/0.0975 = 2.3$ months of the percutaneous injuries curve.

The expected numbers of 'cumulative deaths' column in Rothman's Table 4.2 can be (and probably were) arrived at using the 'exponential' formula

$$1000 \times \{\ 1 - \exp[-\ 0.011\text{yr}^{-1} \times\ (\text{number of years})]\ \}.$$

The 0.011 $\text{yr}^{-1} \times$ (number of years) is the integral of the ID function, i.e., the expected number of transitions, over the number of years in question.

[19] 5. Miettinen, O. S. (1976) Estimability and estimation in case-referent studies. Am. J. Epidem., 103, 226-235.

9 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 132-4

Rothman K [2012] Introduction to Epidemiology (2ndEd) pp 165-6

Baldi and Moore [3rdEdition] Ch 12, p.300-305.

Baldi and Moore [4thEdition] Ch 12, p.310-316.
(Neither B&M edition mentions Poisson in the Index!)

The "Poisson" Distribution: History, Reenactments, Adaptations",
along with the accompanying Gosset and Rutherford websites.

10 Credits

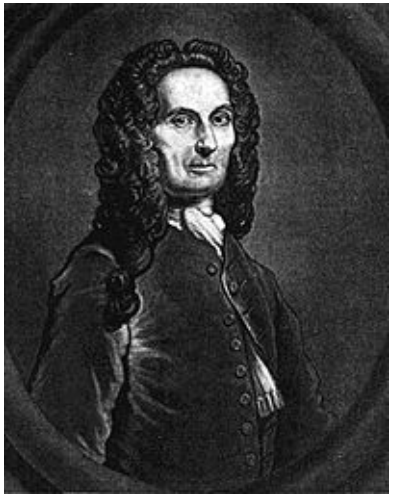
From 'Practical Applications of the Statistics of Repeated Events' Particularly to Industrial Accidents' by Ethel M. Newbold, Journal of the Royal Statistical Society, Vol. 90, No. 3 (1927), pp. 487-547 (https://www.jstor.org/stable/pdf/2341203.pdf)

The path to the (Poisson as a binomial) limit is very simple; as Bortkiewicz has remarked, it does not need a first-class mathematician like Poisson to obtain this limit, and it has been arrived at independently more than once by other people dealing with statistical or physical problems in time or space. "Student" (On the Error of Counting with a Haemacytometer," Biometrika, V, 1906-7, pp. 351-60) obtained it when searching for the probable error of the number of yeast cells counted in the squares of a hoemacytometer; Bateman,(The Probability Variations in the Distribution of a Particles, Phil. Mag., 6th series, 1910, vol. xx, p. 696) in an appendix to a paper by Rutherford and Geiger, by a different method of approach obtained it as an exact formula to describe the frequency of emission of α particles per unit of time in radioactive radiation. Von Bortkiewicz ("Das Gesetz der Kleinen Zahlen" 1898) and Mortara ("Sulle Variazioni di Frequenza di Alcuni Fenomeni Demografici Rari Annali di Statistica," serie V, vol. 4, 1912, pp. 5-61) have used it to describe infrequent events in vital and social statistics, and it has become familiar in many later applications.

Incidentally, P**oisson's name has stuck to the series**, because it has often been stated that he was the first to apply it to probability problems, but this is, I think, incorrect. Poisson's Recherches sur la Probabilite des Jugements was published in **1837**; but over a hundred years before, in **1718**, in the first edition of his Doctrine of Chances, **De Moivre** applied the exponential limit to the following problem :–Problem V (p. 14): "To find in how many Trials an Event will Probably Happen, or how many Trials will be requisite to make it indifferent to lay on its Happening or Failing; supposing that a is the number of Chances for its Happening in any one Trial, and b the number of Chances for its Failing." De Moivre make the chances of the event happening (i.e. happening at least once in the total number of trials) or failing equal, by equating the first term of the binomial (or, as he calls it, "Sir Isaac Newton's theorem") to the rest; his binomial is .. He then proceeds to two limits, first he makes the chance of an event =0.5, ... , then he makes x and q both infinite, the mean x/q remaining finite, and so the left hand of his equation becomes

$$1 + \frac{x}{q} + \frac{x^2}{2!q^2} + + \frac{x^3}{3!q^3} + ...,$$

and hence $\frac{x}{q} = \log_e 2$.
This problem of De Moivre's really contains the basis of Proposition LI of Whitworth's *Choice and Chance*: "If an event happens at random on an average once in time t, the chance of its not happening in a given period T is $e^{-T/t}$," i.e. the first term of a Poisson series. [Stigler disagrees with Newbold.]



Abraham de Moivre 1667 - 1754
https://en.wikipedia.org/wiki/Abraham_de_Moivre

Siméon Denis Poisson 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 and
http://en.wikipedia.org/wiki/Poisson_distribution and

Stephen M. Stigler. Poisson on the Poisson Distribution

Abstract. A translation of the totality of Poisson's own 1837 discussion of the Poisson distribution is presented, and its relation to earlier work of De Moivre is briefly noted.

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 $(-\infty, \infty)$ $\log[y]$ 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 $\hat{\lambda} = Y/PT$[20]. 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., $\log[\hat{\lambda}_2 \div \hat{\lambda}_1]$? Since, in practice, you do not know μ_1 and μ_2, substitute ('plug in') their empirical counterparts.

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 $Y = Y_1 + Y_2 \sim Poisson(\mu_1 + \mu_2)$. Then, look for and cite examples in earlier pages of these Notes where this property is (implicitly) used in practice.

0.3 Link between Poisson & Exponential (and Gamma) Distributions

1. 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 $\mu_Y = W \times \lambda = W \times (1/\mu_T)$.
2. Show the reverse.
3. Marsden and Barratt (see bottom of Rutherford website) gave empirical examples. How complete a two-way mathematical proof did they give?
4. Genest and Hanley (see here and here) gave a space-exploration-inspired example. How complete a two-way mathematical proof did they give?
5. Hanley's 'A more intuitive and modern way to compute a small-sample confidence interval for the mean of a Poisson distribution' (available here) uses this 2-way link. How complete a two-way mathematical proof does he give?
6. Do you consider the 'tire-ruptures' and 'space-journey' applications as 'proofs by example'? In other words, if you were grading the answers to parts 1 and 2, would you accept these applications as (algebra-less or calculus-less) 'proofs'?

[20] PT = amount of Population Time

0.4 (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

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

As per Casella and Berger, 2nd Ed. p338, in an exponential family we also have that

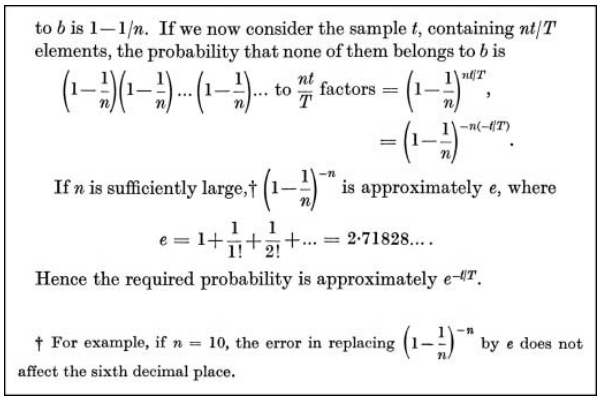
$$E\left\{ \left[\frac{d}{d\theta} \ln f(Y;\theta) \right]^2 \Big|_\theta \right\} = -E\left\{ \frac{d^2}{d\theta^2} \ln f(Y;\theta) \Big|_\theta \right\}.$$

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 $\theta = \log(\mu)$.
3. In both the above, you probably left the expressions in terms or the μ parameter. This '*expected*' information is helpful for planning purposes, i.e., for anticipating how much information you expect to have. But, the expression can also be used after the fact to compute a variance to be used in computing a standard error and margin of error. In this case you will probably plug in the point estimate of μ as a substitute for μ itself. This is called the (after the fact) '*observed*' information. Give a numerical example from the 'Needle-stick' (Percutaneous) Injuries in Interns study.

0.5 (m-s) The Poisson distribution as an approximation to the binomial distribution

Stigler, in The American Statistician, February 2013 (see Resources), writes

> "The Poisson distribution is often introduced as an approximation to the binomial distribution, an approximation that improves in accuracy as n, the number of binomial trials, increases, while np, the expected value, does not:
>
> $$\frac{e^{-np}(np)^k}{k!} \cong \binom{n}{k} p^k (1-p)^{n-k}$$
>
> The presentation is usually accompanied by a proof that invokes some version of the approximation $(1 - 1/n)^{-n} \cong e = 2.71828\ldots$. Poisson's own derivation proceeded in much the same manner (Poisson 1837, p. 206; Stigler 1982), as did a bestselling textbook published in 1936 by Hyman Levy and Leonard Roth. Those authors were, respectively, professor of Mathematics and assistant lecturer in Mathematics at Imperial College London. Figure 1 reproduces the relevant passage from Levy and Roth (1936).



to b is $1-1/n$. If we now consider the sample t, containing nt/T elements, the probability that none of them belongs to b is

$$\left(1-\frac{1}{n}\right)\left(1-\frac{1}{n}\right)\ldots\left(1-\frac{1}{n}\right)\ldots \text{ to } \frac{nt}{T} \text{ factors} = \left(1-\frac{1}{n}\right)^{nt/T},$$

$$= \left(1-\frac{1}{n}\right)^{-n(-t/T)}.$$

If n is sufficiently large,† $\left(1-\frac{1}{n}\right)^{-n}$ is approximately e, where

$$e = 1+\frac{1}{1!}+\frac{1}{2!}+\ldots = 2{\cdot}71828\ldots.$$

Hence the required probability is approximately $e^{-t/T}$.

† For example, if $n = 10$, the error in replacing $\left(1-\frac{1}{n}\right)^{-n}$ by e does not affect the sixth decimal place.

Figure 1. Part of page 80 of Levy and Roth (1936), showing the approximation and the footnote.

For many years, I have been presenting my class with a copy of this page from Levy and Roth and asking them, as a homework exercise, to answer a simple question: **Is the footnote correct?** ”

BIOS601 Exercise: Answer Stigler's question.

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

The NEJM authors did not say how they got the CIs for the Rates per Intern-Month, shown in Table 1 on page 12. 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. [Table on p. 3 may help]

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	8**	33**	78****	100
'Exact' (but conservative) CI:						
`stats::gamma( )`	?	?	?	?	?	?
`stats::poisson.test( )`	?	?	?	?	?	?
`survival::cipoisson(,method="exact")`	?	?	?	?	?	?
Approximation						
Wilson-Hilferty	?	?	?	?	?	?
1st principles, y	?	?	?	?	?	?
1st principles, $y^{1/2}$	?	?	?	?	?	?
SE-based, y	?	?	?	?	?	?
SE-based, $\log(y)$	?	?	?	?	?	?
`survival::cipoisson(,method="anscombe")`	?	?	?	?	?	?
Any others you wish to try	?	?	?	?	?	?

* Rothman (2002 p134) : 3 cases in 2500 PY.
** Rothman (2002 p133, 2012 p.165) : 8 cases in 85,000 PY.
** No. lung cancer deaths in 131,200 W-Y, women 55-60, Quebec,1971.
***Total no. cancers in concerned area in Alberta Sour Gas study. ($\mu_0 = 85.9$) (Table 5, `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1567589/pdf/envhper00423-0270.pdf`

“Cancer Downwind from Sour Gas Refineries”). “A total of 30,175 person-years of risk within Alberta were experienced by this cohort from 1970 to 1984.”

“The significance of the resulting standardized incidence ratios (SIR) was tested by computing 95% confidence intervals around them using methods described by Ederer, F., and Mantel, N. Confidence limits of the ratio of two Poisson variables. Am. J. Epidemiol. 100:165-167(1974).”

0.8 Power Calculations

From “Cancer Downwind from Sour Gas Refineries”:

`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1567589/pdf/envhper00423-0270.pdf`

> “*A priori* sample size calculations were based on an estimate of 2000 people in the IACS with an expected 30,000 person-years of observation. Approximately 102 incident cancers were expected (5), yielding greater than 90% power (6) to detect elevations in the incidence rate by [a factor of 1.2 or more [21] using a one sided test with a confidence level of 0.05.
>
> (5) Cancer Registration in Alberta. Edmonton: Provincial Cancer Hospitals Board, 1978.
>
> (6) Beaumont, J.J. and Breslow, N.E. Power considerations in epidemiologic studies of vinyl chloride workers. Am. J. Epidemiol. 114:725-734 (1981).”

Exercise: The amount of PT is fixed. Thus there is no point in the researcher calculating *what amount of PT would be required* for a desired power against a given alternative. Instead re-do the (pre-study) calculation by (as described in section 7.3)

[21] Their wording: 'detect elevations in relative risk of 1.2 or more.'

(a) establishing what (smallest) number of cancers y^* would be 'statistically significantly higher' than the null $\mu_0 = 102$ calculated on the basis of the sex-age-specific rates in the (much larger) reference areas.

(b) calculating the probability, if $\mu_{alt} = 1.2 \times \mu_0$, of observing a statistically significant elevation.

Compare your calculated power with that claimed by the authors. Do you think they made a 'Type III error' somewhere?

Comment: you may find it easier (and more transparent) to work with the exact Poisson probabilities in `R` rather than with the (quite good in this 3-digit zone) Normal distributions centered on μ_0 and μ_{alt}. However, working with the latter (as in section 7.3) makes it easier to develop generic power/sample size formula.

0.9 From Event-Rates (λ's) to Risks (π's)

Section 8 above introduced a special case of the *general* formula for converting a rate, λ, that may vary over time – and is thus written as the function $\lambda(t)$ – into a risk (a probability):

The risk, R, over the time span t' to t'' is

$$R = 1 - \exp\Big[- \int_{t'}^{t''} \lambda(t) dt \Big].$$

if, as we estimated from the obstetrics/gynecology experience, $\lambda(t) = 0.0975 m^{-1}$ (and constant-over-time) over the span $t' = 0$ to $t'' = 12m$ then the integral $\int_{t'}^{t''} \lambda(t) dt$ is simply $0.0975 m^{-1} \times 12m = 1.17$ injuries in 12-intern-months of continuous service. Thus, the 12-month risk is

$$R_{0 \to 12} = 1 - \exp[-1.17].$$

Exercise:

1. Find a 95% CI for the 1.17, and from it compute a 95% CI for the 12-month risk. *Hint*: what part of the 1.17 is random/subject to sampling variation?

2. Repeat the exercise using the $\hat{\lambda}$ for Pediatrics. Does Rothman's approximate risk equation, i.e., 4-1 in the 2012 Edition, do a good job approximating the 12-month risk in this specialty?

0.10 RCT of HPV vaccine

The following excerpt is from the Vaccine Arm of Table 3 of an Article in the NEJM in 2002 [22]. 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 violationsà

(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.

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."

[22] A Controlled trial of a Human Papillomavirus Type 16 Vaccine by Laura A. Koutsky et al., for The Proof of Principle Study Investigators.

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.

0.11 How well do Poisson models describe variation in. . . ?

1. Yearly Numbers of Dengue Fever Cases `https://www.nature.com/articles/d41586-018-05914-3` and here `https://gatesopenresearch.org/articles/2-36/v1`
2. Daily (and hourly!) Numbers of births? `https://www.significancemagazine.com/585` and `https://rss-onlinelibrary-wiley-com.proxy3.library.mcgill.ca/doi/full/10.1111/j.1740-9713.2017.01026.x` or `hanley/mysteryData/`
3. Daily numbers of Sudden Infant Deaths? `https://www.ncbi.nlm.nih.gov/pubmed/21059188`
4. Monthly numbers of earthquakes in a region? `https://earthquake.usgs.gov/earthquakes/search/`
5. Yearly numbers of major hurricanes? `https://www.nhc.noaa.gov/pastdec.shtml`
6. Yearly numbers/incidence of hospitalized injuries in a region? `http:www.medicine.mcgill.ca/epidemiology/hanley/c609/Material/LidkopingALL.pdf`
7. Yearly Accidents, Fatalities, and Rates, 1982 - 2000, U.S. Air Carriers Operating Under 14 CFR 121 `http:www.medicine.mcgill.ca/epidemiology/hanley/c626/airline-data-sas.txt`
8. Quarterly & Monthly (prevalence) rates of Spina Bifida and Anencephaly Among Births (in relation to fortification of Foods with Folic Acid) `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/folic_acid.pdf`. See more data on webpage `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/`.
9. (Yearly) fatal and nonfatal crash rates on a toll highway (following a 5-15 mph (8-24 kph) decrease in speed limits) `https://www.ncbi.nlm.nih.gov/pubmed/1251837`
10. Yearly numbers of accidents before/after change to Daylight Savings Time) `https://www.nejm.org/doi/full/10.1056/NEJM199604043341416`
11. Daily numbers of in-hospital deaths and Daily Maximal Temperatures during summers of 2002 and 2003 (France) `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/Heatwave_death_lyon.pdf`
12. The (daily) incidence of crimes reported to 3 police stations in different towns (one rural, one urban, one industrial) vis-a-vis the day of the lunar cycle `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/fullmoon.pdf`
13. Daily numbers of Deaths (Postponement of Death Until Symbolically Meaningful Occasions) `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/holidays.pdf`
14. Rates of audience fidget. (F Galton) `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/measure_of_fidget_galton.pdf`
15. Number of Deaths by Horsekicks in the Prussian Army from 1875-1894 for 14 Corps `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/horsekicks.txt` and `http://www.medicine.mcgill.ca/epidemiology/hanley/c626/horsekicks_anthrax_poisson.pdf`
16. Daily counts of individuals involved in a car crash with at least one fatality (here)
17. Weekly COVID-19 cases reported on McGill's campuses (McGill's Case Tracker)

0.12 Pure cultures - Gosset 1907 & Part A Exam 2019

'Student'/Gosset ended his 1907 Biometrika paper on counting yeast cells – where he derived the Poisson distribution from first principles – with an application to the creation of a pure culture (a population of cells that originates from a single cell, so that the cells are genetic clones of one another). He wrote:

> "To do so, it is customary to estimate the concentration of cells and then dilute so that each two drops of the liquid contain on an average one cell. Different flasks are then seeded with one drop of the liquid in each, and then most of those flasks which show growths are pure cultures."

He used the Poisson distribution to show that "*approximately three-quarters* of those flasks which show growth are pure cultures."

Exercise: Derive the analytic expression for this proportion.

0.13 With luck, will the Royal Mint have enough coins?

Refer to the story "Babies who share royal birthday will coin it" [23] on next page and to the 'average of 1,983 births a day.'

1. (From the information in the article) what is the probability that the Mint will have enough, if they mint 2,013 coins? State any assumptions made.
2. How many should they mint to be 99.99% sure of having enough?
3. The average number of births per day varies slightly by season, and substantially by day of the week – JH could not find day-of-week data for the UK[24], but did find 2010 data from the USA.[25] Rework questions 1 and 2 using a worst case scenario, and assuming the same day-of-week patterns seen in the USA apply in England and Wales [***scale row 1 of the CDC table for the USA down to match the size of UK***]
4. For shorthand purposes, refer to the probability of having enough coins as the '*non-exceedance*' probability.[26] How close is the mean of the 7 non-exceedance probabilities to the non-exceedance probability calculated at the mean no. of births per day? How close is the *median* non-exceedance probability? What if we switched focus to the *exceedance* probability rather than the non-exceedance probability?
5. (Again, under your worst case scenario) how many pink and blue pouches would you recommend they have ready?

[23] Seems that 'to coin it' means means 'to profit'

[24] http://www.statistics.gov.uk/hub/population/births-and-fertility/live-births-and-stillbirths

[25] http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm

[26] A New Zealand webpage entitled What does Annual Exceedance Probability or AEP mean? says 'This term is generally referred to in rules that regulate discharges of contaminants including stormwater, wastewater, greywater. It can also be referred to in rules that regulate the use of land that may result in a discharge including offal pits, storage facilities for animal effluent, stockpiling organic matter (including composting) and storage of hazardous substances. The Annual Exceedance Probability is the chance or probability of a natural hazard event (usually a rainfall or flooding event) occurring annually and is usually expressed as a percentage. Bigger rainfall events occur (are exceeded) less often and will therefore have a lesser annual probability. Example 1: 2% exceedance probability rainfall event: A 2% Annual Exceedance Probability rainfall event has a 2% chance of occurring in a year, so once in every 50 years. Example 2: 20% exceedance probability rainfall event: A 20% Annual Exceedance Probability rainfall event has a 20% chance of occurring in a year, so once in every 5 years.

No SIM 6:48 AM 100%

THE TIMES
Friday, 5 July 2013

News

Babies who share royal birthday will coin it

Valentine Low

Babies born on the same day as the Duke and Duchess of Cambridge's first child are to receive a commemorative silver penny, the Royal Mint announced yesterday.

It is giving away 2,013 of the pennies, which reflect the tradition of marking a new birth with a gift of silver for good luck. This, however, is a new tradition: it will be the first time the Royal Mint has marked a royal birth by giving away coins.

The Royal Mint will give away 2,013 of the silver pennies in special pouches
PA

The pennies are worth £28 each, making the overall cost to the state-owned company more than £50,000.

The silver penny, which will be presented in a pink or blue pouch, is marked with the year 2013 to commemorate the baby girl or boy's year of birth and features a shield of the Royal Arms designed by Matthew Dent.

Shane Bissett, Director of Commemorative Coin at the Royal Mint, said: "The birth of the royal baby will be a joyous occasion, not just for Their Royal Highnesses the Duke and Duchess of Cambridge, but also for the whole nation, as we prepare to celebrate another remarkable milestone in their life journey together.

"However, it will also be a special day for many mothers and fathers across the country as they, too, welcome the arrival of their new baby; hence why we wanted to extend this historical moment to them with a lucky silver penny."

Parents of babies born on the same day as the royal baby have 60 days to claim their silver penny by visiting Facebook.com/theroyalmint to register the birth of their child. With luck, the Royal Mint should have enough coins: an average of 1,983 babies are born in England and Wales each day. ■

CONTENTS EDITION LIVE NEWS

Beyond 20/20 WDS - Table view - ME_ROUT by DOB_WK (2010 Birth Data - State Detail)

CDC Home | NCHS Home | Contact NCHS | NVSS Home | VitalStats Home | Privacy Policy | Accessibility

CDC National Vital Statistics System
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Tables | Table | Chart

ME_ROUT by DOB_WK (2010 Birth Data - State Detail)

Other:

DOB_WK / ME_ROUT	Total	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Total	3,999,386	369,704	606,424	666,686	656,694	649,636	633,899	416,343
Vaginal-Spontaneous	1,931,624	203,437	280,188	310,516	309,225	306,652	295,023	226,583
Vaginal-Forceps	20,868	2,069	3,055	3,456	3,506	3,352	3,056	2,374
Vaginal-Vacuum	89,879	9,114	12,669	14,962	14,761	14,358	13,566	10,449
Cesarean	995,945	67,410	163,540	176,929	170,666	168,245	170,456	78,699
Not stated	17,568	1,574	2,682	3,022	2,964	2,793	2,730	1,803
Not on certificate	943,502	86,100	144,290	157,801	155,572	154,236	149,068	96,435

DayOfWeek	Total	January	February	March	April	May	June	July	August	September	October	November	December
Total	3999386	323249	301994	338613	325028	328273	334535	345199	349747	350745	336809	326220	338974
Sunday	369704	34516	27851	27331	27326	34675	28456	29104	36658	30460	36798	28536	27993
Monday	606424	45873	45899	57699	46839	54815	47351	44502	61073	45406	47795	60967	48205
Tuesday	666686	50373	49751	62979	50302	51071	63178	51325	65572	53189	51056	65209	52681
Wednesday	656694	49042	49277	62067	49237	50024	63541	52202	51530	67103	49698	49109	63864
Thursday	649636	49448	49230	48981	61017	49637	50787	64681	51511	67279	49963	45162	61940
Friday	633899	55630	48382	48155	59075	48472	48887	62232	49800	52728	61238	45280	54020
Saturday	416343	38367	31604	31401	31232	39579	32335	41153	33603	34580	40261	31957	30271

See 1. article, **by a student and teacher of** bios601, on this topic
https://www.significancemagazine.com/585 and

2. Mystery Data Quiz, in Significance Magazine, also on the timing of births.
https://rss-onlinelibrary-wiley-com.proxy3.library.mcgill.ca/doi/full/10.1111/j.1740-9713.2017.01026.x
or http://www.medicine.mcgill.ca/epidemiology/hanley/mysteryData/

0.14 2 (indep.) Poisson r.v.'s → 1 Binomial distribution

Suppose we wish to compare 2 event-rates, λ_1 in 'exposed' ($_1$) person time and λ_0 in 'unexposed' ($_0$) person time. Denote the (to-be-observed) numbers of events in Y_1 and Y_0 person-years by D_1 and D_0 respectively.[27]

Then

$$D_1 \sim \text{Poisson}(\mu_1) \text{ and } D_0 \sim \text{Poisson}(\mu_0),$$

where

$$\mu_1 = \lambda_1 \times Y_1 \text{ and } \mu_0 = \lambda_0 \times Y_0.$$

Show that by conditioning on (fixing) the sum $D = D_1 + D_0$, we obtain a binomial random variable:

$$(D_1 \mid D) \sim Binomial\Big(D,\ \pi = \frac{\mu_1}{\mu_1 + \mu_0} = \frac{\lambda_1 Y_1}{\lambda_0 Y_0 + \lambda_1 Y_1} = \frac{\theta Y_1}{Y_0 + \theta Y_1}\Big),$$

where θ is the Rate Ratio λ_1/λ_0,

and that

$$\Omega = \frac{\pi}{1-\pi} = \frac{E[D_1]}{E[D_0]} = \frac{Y_1}{Y_0} \times \frac{\lambda_1}{\lambda_0}.$$

0.15 Cancer screening trials: sample size/data-analysis

[new in 2017, and a prelude to the visit of Steven Skates (UK Ovarian Cancer Screening Trial) on Oct 3, 2017]

The following sections are taken from 'Biometric design of the Mayo Lung Project for early detection and localization of bronchogenic carcinoma.' by Taylor WF, Fontana RS. Cancer. 1972 Nov;30(5):1344-7. More material here, under 'Cancer Screening Trials']

ABSTRACT

Several important aspects of the Mayo Lung Project demand evaluation. These are: 1. Acceptance. Will people accept such a screening program? 2. Case finding. Does the screen pick out the people most likely to have or develop bronchogenic carcinoma? 3. Effectiveness. If an early case of bronchogenic carcinoma is found, will prompt treatment extend life beyond the time at which death from this disease would have occurred if treatment had been delayed? Direct measurement of effectiveness is not possible, and indirect methods must be used. A group of patients, all of whom are considered suitable for the screening program, are being divided randomly into two sub- groups, one to be screened and the other to be kept as an unscreened control. Mortality in the two groups is to be compared for 5 years, and hopefully for 10 years. We also consider here sample size requirements and reports on some of the characteristics of the first 500 patients.

DESIGN OF PROJECT

Subjects and methods: In the course of usual procedure at the Mayo Clinic, we identify each male patient who is 45 years of age or older and who smokes at least one pack of cigarettes a day. As part of the routine health examination of such patients, a standard 14 by 17-inch posterior-anterior chest roentgenogram is made and studied and a pooled 3-day "deep cough" sputum specimen is examined cytologically. We have the patients answer a Lung-Health Questionnaire as part of this project. All men found free from clinical evidence of lung cancer and free from other serious diseases (to the degree that life expectancy is estimated as at least 5 years) are included in this study. These patients are assigned at random to one of two groups.

1. One group, designated *controls*, receives care and advice of the standard which is current practice at Mayo Clinic. This includes the recommendation of the Clinic's Division of Thoracic Diseases that a chest roentgenogram and a sputum cytology test be obtained at least once a year and that the patient stop smoking. However, these men will be told nothing of the screening program. Rather, they will be examined and will receive care at their own request as if no screening program existed. A routine follow-up communication will be made with each man at least once a year for at least 10 years to determine survival status. If a man dies, his death certificate will be obtained and the circumstances of his death will be determined from his local doctor.

2. The other group, called *participants*, will be treated initially just as the first group, but these men will also be urged to participate in the intensive bronchogenic carcinoma screening project.[28] Men who refuse will not be dropped; they will be followed as closely as possible through correspondence and will be included when comparisons are made with the first group.

Analysis: If the work is carefully clone and if adequate time is allotted for the project, a moderate difference in observed lung cancer mortality can be deemed significant statistically and can be attributed to some aspect and effect of the screening procedure. (We may not know which aspect, but at least we will have established that screening and early treatment have some effect, and we will have incentive to pursue the matter further. Such aspects as how intensive the screening should be or how costs can be reduced are perhaps better delayed until the question of gross effectiveness is answered.)

Notice that we *will not merely compare survival time of early-discovered and late-discovered cases*. There is an unknown bias in favor of early-discovered cases, even if no treatment is employed. *Notice also that we do not rely on volunteers for one group and let the comparison group consist of nonvolunteers*. Instead, we divide the group of eligible people at random into two groups, offer the screening to one of them, and then compare the two groups in their entirety. Finally, it should be noted that we do not plan to make a comparison of the incidence of cancer or the survival of cancer patients among the unscreened controls with that of the participants, because to get such detailed information we would have to communicate with the control patients and thus lose part of the difference between control and screened patients. The screened group may have a higher observed incidence because we observe them more closely. We want the two groups to be observed with different intensity-within the bounds of currently acceptable medical practice-because this is what the study is all about.

A word about eligibility: An early benefit from this work results from the first screening. The cases of lung cancer found then will be interesting in themselves and will be worked up thoroughly. The initial screening should also eliminate from further study patients who for other reasons are considered to have an unusually short expectation of life. This, of course, will be somewhat subjective, but decisions will be made as consistently as possible,

[27] Clayton and Hills used the letter D, since it is short for numbers of 'deaths'; not all of the events in epidemiology are terminal, or unwanted.

[28] "We decided to use a 4-month screening interval because previous studies suggested that longer intervals were too long. We thought a 4-month interval would be acceptable to our patients and achievable by our technical personnel."

in accord with written guidelines.

Sample size and time required: We have considered sample size in relation to comparison of mortality from bronchogenic carcinoma in the two designated groups. Suppose we admit N men into each group. After 5 years there will have occurred T_1 and T_2 man-years of exposure in each group, and D_1 and D_2 deaths. If $T_1 \cong T_2$, as is likely, we merely must determine whether the control deaths D_1, exceed significantly the screened deaths D_2. **It is reasonable to consider the D_1 and D_2 deaths as independent binomial trials.** Let p denote the probability that, **given a death occurs**, it occurs in the controls. Let H_0 be the hypothesis $p = 1/2$, and let H_1 be the alternative of interest, $p = 2/3$. (*This corresponds to reducing the lung cancer death rate in the screened group to half that in the controls.*) We want the following two conditions to be met,

$$P(\text{reject } H_0 \text{ in favor of } H_1 | H_0 \text{ true}) = \alpha = 0.05$$

$$P(\text{reject } H_0 \text{ in favor of } H_1 | H_1 \text{ true}) = \beta = 0.95.$$

We reject H_0 in favor of H_1 whenever

$$\left[\left(\frac{D_1}{D_1 + D_2} - \frac{1}{2} \right) \Big/ \sqrt{\frac{1}{4} \, \frac{1}{D_1 + D_2}} \right] \geq 1.645.$$

The probability that this occurs under H_0 is about 0.05. The probability under H_1 is about 0.95 if $D_1 + D_2 = 90$.

Now the question is: how how many men must we examine for how long to get about 90 deaths from bronchogenic carcinoma? (The following information is in the nature of a first attempt to estimate this quantity.) Suppose we wish to get an answer in 5 years, and assume from published data and some educated guessing that 5 deaths per 1,000 man-years will occur among the controls and 2.5 deaths per 1,000 man-years among the participants in the close surveillance. We expect to have 60 deaths among the controls and 30 among the participants if we observe 12,000 man-years in each. These estimates, based on averages, do not take into account chance variation. If we wish to be 95% sure of obtaining 60 and 30 deaths, respectively, we need to observe 12,000 man-years in each group. We think we can obtain such numbers from our present case load but not without difficulty. Initial plans calling for a total of 6,000 men (3,000 in each group) may have to be modified and will be as soon as deemed essential. We anticipate some losses; there may well be men who refuse to continue under screening These are not to be entirely lost; their cases will be followed anyway by mail. But it does dilute the difference between the qroups and makes the true effects of screening more difficult to detect. The surveillance effort will have to be vigorous and encouraging.

Will 5 years be long enough, even with the numbers of subjects proposed? Perhaps not; but regardless of the early outcome and regardless of whether the actual screening goes on beyond 5 years, these men should continue to be traced for at least a total of 10 years. In our opinion, important information about survival following early treatment will require more than 5 years' study. This opinion is based on possible recurrence of the initial cancer, as well as concern over development of an entirely new primary cancer, particularly in individuals with squamous cell carcinoma.

— — —

Questions - for `bios601` exercise

1. re-write the sentence "It is reasonable to consider the D_1 and D_2 deaths as independent binomial trials."
2. With $D = 90$ and (the null) $p = 0.5$, use the `pbinom` function to calculate $d_1^{critical}$, the smallest d_1 such that $\text{Prob}[D_1 \geq d_1] < \alpha$. [see Note[29]]
3. (Staying with $D = 90$) use the non-null $p = 2/3$ in the `pbinom` function to calculate $\text{Prob}[D_1 \geq d_1^{critical}]$ and check the value against the 'about 95%' [power] given by Taylor and Fontana.

   JH finds that rough diagrams are a big help in setting up power calculations like these.
4. Comment on their use of the letter β to denote this probability.
5. Taylor and Fontana did not have easy access to binomial calculations, so they used a Normal approximation to the binomial. i.e,

   $$(D_1 | D) \sim N[\ \mu = D \times p, \ \text{Var } = D \times p \times (1 - p)\ ].$$

   (Staying with $D = 90$) use this approximation to repeat the above calculations for $p = 1/2$ and $p = 2/3$.
6. In the above H_1 the alternative of interest, $p = 2/3$, corresponded to reducing the lung cancer death rate in the screened group to half that in the controls, i.e. (using their '1' to denote to denote the controls, and '2' to denote the participants) to a situation where $\lambda_2 = 0.5 \times \lambda_1$.

   But what if this alternative is **too optimistic**? Consider four more modest scenarios: $H_2 : \lambda_2 = 0.6 \times \lambda_1$; $H_3 : \lambda_2 = 0.7 \times \lambda_1$; $H_4 : \lambda_2 = 0.8 \times \lambda_1$; and $H_5 : \lambda_2 = 0.9 \times \lambda_1$, i.e., reductions of 40%, 30%, 20%, and 10% respectively. First, convert these 4 scenarios to the corresponding 4 non-null values of p and (staying with $D = 90$), calculate $\text{Prob}[D_1 \geq d_1^{critical}]$, i.e., the statistical power, for each of these.[30]
7. As you will have found, the power against H_4 (a 20% reduction) is low when D is just 90. By trial and error, or directly, calculate the D one would need to have 80% power (rather than their 95%) but against just a 20% reduction.

   Convert this required D to a required number of man-years, using a mortality rate of 3 per 1,000 man-years in the controls.[31]

[29] Note the values of `pbinom(3,4,.5)` and `pbinom(3,4,.5,lower.tail=FALSE)`

[30] Use exact binomials, or normal approximations, as you wish.

[31] This rate of 3/1,000 MY was calculated 'after-the-fact' in 1986, after 115 lung cancer deaths had been observed in 4,600 men followed for an average of just over 8 years. As you will have seen above, the rate used for planning purposes was 5 per 1,000 man-years.

AFTERMATH 1981, 1986, 2000 || CT screening: 2006, 2011

Some Results of Screening for Early Lung Cancer

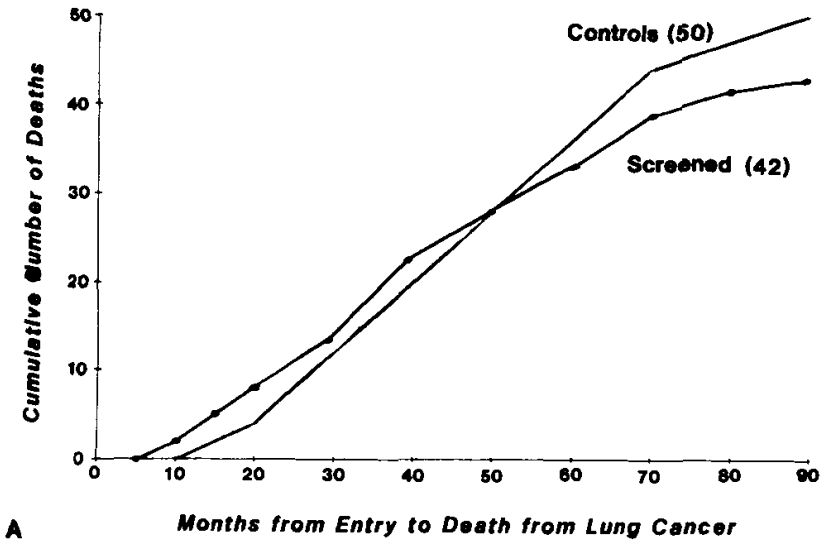
WILLIAM F. TAYLOR, PHD, ROBERT S. FONTANA, MD, MARY ANN UHLENHOPP, BA, AND CHARLES S. DAVIS, MS

Screening for lung cancer is somewhat controversial in that very few evaluations of the screening process have been made, and even fewer have involved the use of concomitant, unscreened controls. This report of the Mayo Lung Project provides evaluation of a randomly selected 4500 clinic patients, offered screening for lung cancer at four-month intervals for six years. Another 4500 randomly selected controls not offered screening were merely observed. Good screening is defined, the Mayo project is evaluated, and puzzling results are presented and discussed.

From the screened group, 98 new cases of lung cancer have been detected, 67 by study screening and 31 by spontaneous reporting of symptoms (15) or by x-ray examinations (16) done in other than study circumstances. From the controls, 64 new lung cancer cases have been detected, 43 by symptoms and 21 by other methods. Lung cancer mortality is 39 for study patients and 41 for controls. There is thus no evidence at this time that early case finding has decreased mortality from lung cancer.

Cancer 47:1114–1120, 1981.

JH is puzzled by the sentence 'Lung cancer mortality is 39 for study patients and 41 for controls.' in the above summary. The 39 and 41 do not agree with the numbers (42 and 50) given elsewhere in the text and in the various Figures.



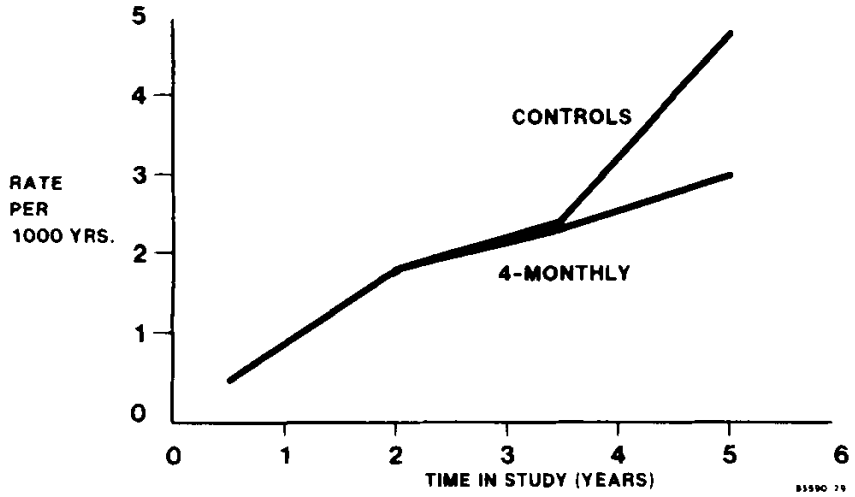


FIG. 7. Lung cancer death rates by time in study—control and screened patients.

[From Discussion] 'A second hopeful observation has to do with the actual lung cancer death rates for controls and screened patients, as shown in Figure 7. Attention is directed particularly to the rates for those patients who have been in the study four years or more. The death rate from lung cancer for the controls exceeds that of the screened group by a considerable amount, although this is not yet statistically significant either. However, this trend has been observed for the last three years, and the difference is growing.

We believe that lung cancer screening appears promising for squamous cancers and for adeno-carcinomas but not for small or large cell un-differentiated tumors. Our recommendation now is to continue observation well into the follow-up phase (at least three to five more years). We suggest that no lung cancer screening projects be established for the general population of older male smokers at this time. But, we also suggest that we do not now know enough about this matter to make definitive statements.

Lung Cancer Screening: The Mayo Program [1986]

Robert S. Fontana, MD; David R. Sanderson, MD; Lewis B. Woolner, MD; William F. Taylor, PhD; W. Eugene Miller, MD; and John R. Muhm, MD

Journal of Occupational Medicine/Volume 28 No. 8/August 1986

(Summary) The National Cancer Institute has sponsored three randomized controlled trials of screening for early lung cancer in large, high-risk populations to determine whether (1) lung cancer detection can be improved by adding sputum cytological screening every 4 months to chest roentgenography done either yearly or every 4 months; and (2) lung cancer mortality can be significantly reduced by this type of screening program, followed by appropriate treatment. Results of the three trials suggest that (1) sputum cytology alone detects 15% to 20% of lung cancers, almost all of which are squamous cancers with a favorable prognosis; and (2) chest roentgenography may be a more effective test for early-stage lung cancer than previous reports have suggested. Nevertheless, results of the randomized trial conducted at the Mayo Clinic showed that offering both procedures to high-risk outpatients every 4 months conferred no mortality advantage over standard medical practice that included recommended annual testing.

(From results section) In the MLP randomized trial, the death rates from all causes (per 1,000 person-years) were high: 24.8% in the screened every 4 months and 24.6% in the control group. The major competing death risk was ischemic cardiovascular disease.

There were 122 lung cancer deaths in the group screened every 4 months and 115 in the control group. Seven deaths in the group screened every 4 months and six deaths in the control group were attributed to surgery for lung cancer. These were treated as lung cancer deaths.

The death rate from lung cancer was 3.2/1,000 person-years in the group screened every 4 months and 3.0 among the control subjects. Like the cumulative numbers of unresectable cancers, the cumulative numbers of lung cancer deaths in the two groups were comparable, both during and after the period of active screening.

Comments

The results of the MLP randomized controlled trial do not justify recommending large-scale programs of radiological or cytological screening for lung cancer. Such programs are usually initiated by those who conduct them and should benefit the participants by reducing lung cancer mortality.' The MLP trial did not demonstrate this sort of benefit.

Neither do the results of the MLP mean that testing high-risk patients for lung cancer by chest x-ray film or sputum cytology is not useful, as some have claimed.' All who participated in the MLP trial received an initial (prevalence) radiological and cytological screening. The randomized trial simply shows that offering the two procedures every 4 months to high-risk Mayo outpatients who have had one negative screening confers no morality [*sic*][32] advantage over routine Mayo Clinic practice with a recommendation of annual testing. The randomized, controlled trials conducted at the Johns Hopkins Medical Institutions and at the Memorial Sloan-Kettering Cancer Center offered all participants annual chest roentgenograms. In addition, half of the men in each of these trials were randomly allocated to a group offered sputum cytology every 4 months. Results from both trials indicate that in the populations screened by x-ray film only, as well as in the populations screened by x-ray film and cytology, the proportion of early-stage, resectable lung cancers and the lung cancer survivorship have been substantially better than those observed in previously reported lung cancer screening programs. However, like the MLP, no significant difference in lung cancer mortality has been observed between the two populations in either the Hopkins or the Memorial trial.'

It should be emphasized that when the NCI randomized controlled trials commenced, it was generally accepted that yearly chest roentgenograms would not reduce lung cancer mortality. It was also believed that a large proportion of lung cancers would be detected cytologically, and the trials were designed with this in mind. Yet in all three screening programs, the great majority of lung cancers have been detected radiologically. Furthermore, sizable numbers were detected by nonstudy chest x-ray films in the control group of the MLP and by annual chest x-ray films in the control populations of the other two trials. It would be of interest to know what might have happened in these cases if chest roentgenograms had not been available to the control subjects.

The randomized controlled trial is ideal for assessing new procedures such as mammography, or new application of procedures such as screening populations at high risk of lung cancer by sputum cytology. Unfortunately, once a procedure has become an established part of medical practice, as the chest roentgenogram has (more than 80 million are taken year in the United States), it may become necessary to resort to other, less precise methods of evaluation, such as case-control studies.

Summary

Three large, long-term randomized controlled trials of screening for early-stage lung cancer by periodic chest x-ray film and sputum cytology have been conducted under the auspices of the National Cancer Institute. Cytological screening alone has detected only a small proportion of the lung cancers in these programs, although cytologically detected lung cancers tend to have a very favorable prognosis. Modern chest roentgenography appears to be a better method of detecting early-stage, resectable lung cancer than previous stud- ies have indicated.

Everyone who participated in the Mayo Clinic randomized trial had a satisfactory and negative initial (prevalence) radiological and cytological screening. The study group was then offered re-screening every 4 months, while the control group was offered standard medical care and advised to have annual chest radiography and sputum cytology.

The Mayo trial has shown significantly increased lung cancer detection, resectability, and survivorship in the study group compared with that of the control groups. Yet the death rates from lung cancer and from all causes have been almost identical in the two groups.

2000

Lung Cancer Mortality in the Mayo Lung Project: Impact of Extended Follow-up Pamela M. Marcus, Erik J. Bergstralh, Richard M. Fagerstrom, David E. Williams, Robert Fontana, William F. Taylor, Philip C. Prorok. [JNCI]

Background: The Mayo Lung Project (MLP) was a randomized, controlled clinical trial of lung cancer screening that was conducted in 9211 male smokers between 1971 and 1983. The intervention arm was offered chest x-ray and sputum cytology every 4 months for 6 years; the usual-care arm was advised at trial entry to receive the same tests annually. No lung cancer mortality benefit was evident at the end of the study. We have extended follow-up through 1996.

Methods: A National Death Index-PLUS search was used to assign vital status and date and cause of death for 6523 participants with unknown information. The median survival for lung cancer patients diagnosed before July 1, 1983, was calculated by use of Kaplan-Meier estimates. Survival curves were compared with the log-rank test.

Results: The median follow- up time was 20.5 years. Lung cancer mortality was 4.4 (95%

[32] `https://en.wikipedia.org/wiki/Sic`

confidence interval [CI] = 3.9-4.9) deaths per 1000 person-years in the intervention arm and 3.9 (95% CI = 3.5-4.4) in the usual-care arm (two-sided P for difference = .09). For participants diagnosed with lung cancer before July 1, 1983, survival was better in the intervention arm (two-sided P = .0039). The median survival for patients with resected early-stage disease was 16.0 years in the intervention arm versus 5.0 years in the usual-care arm.

Conclusions: Extended follow-up of MLP participants did not reveal a lung cancer mortality reduction for the intervention arm. Similar mortality but better survival for individuals in the intervention arm indicates that some lesions with limited clinical relevance may have been identified in the intervention arm. [J Natl Cancer Inst 2000;92:1308-16]

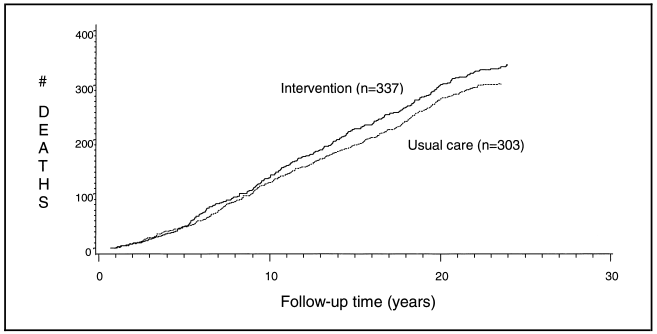


Fig. 1. Cumulative lung cancer deaths by study arm. Sample size was 4607 in the intervention arm **(solid line)** and 4585 in the usual-care arm **(dashed line). Numbers in parentheses** are the numbers of lung cancer deaths as of December 31, 1996. The National Death Index was used, as described in the text, to follow-up Mayo Lung Project participants for whom vital status on December 31, 1996, was unknown.

Table 2. Mortality in the Mayo Lung Project, as of December 31, 1996

Cause of death*	Deaths, No. (%)		Mortality rate (95% confidence interval) per 1000 person-years	
	Intervention arm (n = 4607)	Usual-care arm (n = 4585)	Intervention arm (76 760.7 person-years)	Usual-care arm (76 772.4 person-years)
Lung cancer	337 (7)	303 (7)	4.4 (3.9–4.9)	3.9 (3.5–4.4)
Causes other than lung cancer	2148 (47)	2133 (47)	28.0 (26.8–29.2)	27.8 (26.6–29.0)
Cancers other than lung cancer	403 (9)	391 (9)	5.3 (4.8–5.8)	5.1 (4.6–5.6)
Chronic obstructive pulmonary disease	156 (3)	149 (3)	2.0 (1.7–2.4)	1.9 (1.6–2.3)
Ischemic heart disease	816 (18)	816 (18)	10.6 (9.9–11.4)	10.6 (9.9–11.4)
Other respiratory causes	60 (1)	44 (1)	0.8 (0.6–1.0)	0.6 (0.4–0.8)
Other	712 (15)	733 (16)	9.3 (8.6–10.0)	9.5 (8.9–10.3)
All causes	2493 (54)	2445 (53)	32.5 (31.2–33.8)	31.8 (30.6–33.1)

*Seventeen participants (eight in the intervention arm and nine in the usual-care arm) had unknown causes of death.

LOW-DOSE CT SCREENING

The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 OCTOBER 26, 2006 VOL. 355 NO. 17

Survival of Patients with Stage I Lung Cancer Detected on CT Screening

The International Early Lung Cancer Action Program Investigators*

ABSTRACT

BACKGROUND

The outcome among patients with clinical stage I cancer that is detected on annual screening using spiral computed tomography (CT) is unknown.

METHODS

In a large collaborative study, we screened 31,567 asymptomatic persons at risk for lung cancer using low-dose CT from 1993 through 2005, and from 1994 through 2005, 27,456 repeated screenings were performed 7 to 18 months after the previous screening. We estimated the 10-year lung-cancer–specific survival rate among participants with clinical stage I lung cancer that was detected on CT screening and diagnosed by biopsy, regardless of the type of treatment received, and among those who underwent surgical resection of clinical stage I cancer within 1 month. A pathology panel reviewed the surgical specimens obtained from participants who underwent resection.

RESULTS

Screening resulted in a diagnosis of lung cancer in 484 participants. Of these participants, 412 (85%) had clinical stage I lung cancer, and the estimated 10-year survival rate was 88% in this subgroup (95% confidence interval [CI], 84 to 91). Among the 302 participants with clinical stage I cancer who underwent surgical resection within 1 month after diagnosis, the survival rate was 92% (95% CI, 88 to 95). The 8 participants with clinical stage I cancer who did not receive treatment died within 5 years after diagnosis.

CONCLUSIONS

Annual spiral CT screening can detect lung cancer that is curable.

The members of the Writing Committee (Claudia I. Henschke, M.D., Ph.D., David F. Yankelevitz, M.D., Daniel M. Libby, M.D., Mark W. Pasmantier, M.D., and James P. Smith, M.D., New York Presbyterian Hospital–Weill Medical College of Cornell University, New York; and Olli S. Miettinen, M.D., Ph.D., McGill University, Montreal) of the International Early Lung Cancer Action Program assume responsibility for the overall content and integrity of the article. Address reprint requests to Dr. Henschke at New York Presbyterian Hospital–Weill Medical College of Cornell University, 525 E. 168th St., New York, NY 10021, or at chensch@med.cornell.edu.

*The International Early Lung Cancer Action Program investigators are listed in the Appendix.

N Engl J Med 2006;355:1763-71.

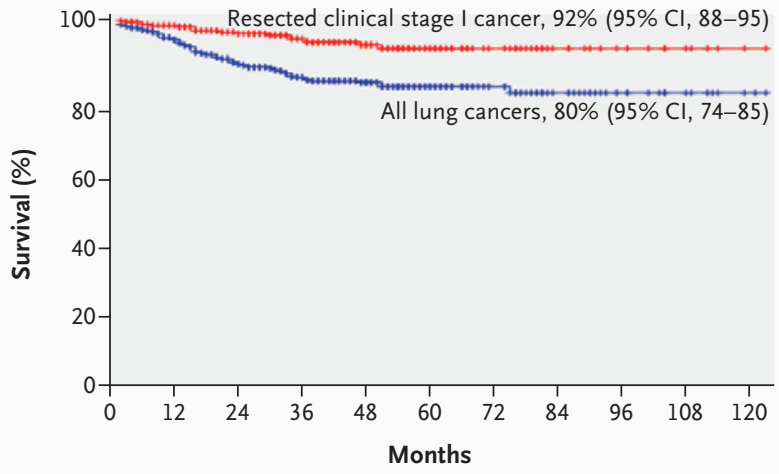


Figure 2. Kaplan–Meier Survival Curves for 484 Participants with Lung Cancer and 302 Participants with Clinical Stage I Cancer Resected within 1 Month after Diagnosis.

The diagnoses were made on the basis of CT screening at baseline combined with cycles of annual CT.

The National Lung Screening Trial:

Overview and Study Design [Gatsonis et al. Radiology: Volume 258: Number 1, January 2011]

The National Lung Screening Trial (NLST) is a randomized multicenter study comparing low-dose helical computed tomography (CT) with chest radiography in the screening of older current and former heavy smokers for early detection of lung cancer, which is the leading cause of cancer-related death in the United States. Five-year survival rates approach 70% with surgical resection of stage IA disease; however, more than 75% of individuals have incurable locally advanced or metastatic disease, the latter having a 5-year survival of less than 5%. It is plausible that treatment should be more effective and the likelihood of death decreased if asymptomatic lung cancer is detected through screening early enough in its preclinical phase. **For these reasons, there is intense interest and intuitive appeal in lung cancer screening with low-dose CT**. The use of survival as the determinant of screening effectiveness is, however, confounded by the well-described biases of lead time, length, and overdiagnosis. Despite previous attempts, no test has been shown to reduce lung cancer mortality, an endpoint that circumvents screening biases and provides a definitive measure of benefit when assessed in a randomized controlled trial that enables comparison of mortality rates between screened individuals and a control group that does not undergo the screening intervention of interest. The NLST is such a trial. The rationale for and design of the NLST are presented.

Sample Size Considerations

Preliminary computations of the required sample size for the NLST were made by using the approach of Taylor and Fontana, which is based on several simplifying assumptions and does not account for the number of screenings. The final computations were based on an elaboration of the approach of Hu and Zelen, modified to allow for staggered entry of participants and analyses based on calendar time instead of time on study. Parameters for the Hu-Zelen model are listed in Appendix E8 (online) and were estimated by using data from the Mayo Lung Project. **With 25 000 participants enrolled in each of years 1 and 2 of the trial, [i.e., 25,000 per arm, enrolled over 2 years]** statistical **power of 90% for detecting a 21% reduction in lung cancer mortality in the low-dose CT arm relative to the chest radiographic arm** may be achieved in an analysis conducted on events occurring through August 2008. Because of lags in data availability and entry, such an analysis would not occur until 2010. Therefore, we continued to collect information on lung cancer cases and deaths occurring through December 2009 so that information would not have to be obtained retroactively if needed.

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening

The National Lung Screening Trial Research Team*

ABSTRACT

BACKGROUND

The aggressive and heterogeneous nature of lung cancer has thwarted efforts to reduce mortality from this cancer through the use of screening. The advent of low-dose helical computed tomography (CT) altered the landscape of lung-cancer screening, with studies indicating that low-dose CT detects many tumors at early stages. The National Lung Screening Trial (NLST) was conducted to determine whether screening with low-dose CT could reduce mortality from lung cancer.

METHODS

From August 2002 through April 2004, we enrolled 53,454 persons at high risk for lung cancer at 33 U.S. medical centers. Participants were randomly assigned to undergo three annual screenings with either low-dose CT (26,722 participants) or single-view posteroanterior chest radiography (26,732). Data were collected on cases of lung cancer and deaths from lung cancer that occurred through December 31, 2009.

RESULTS

The rate of adherence to screening was more than 90%. The rate of positive screening tests was 24.2% with low-dose CT and 6.9% with radiography over all three rounds. A total of 96.4% of the positive screening results in the low-dose CT group and 94.5% in the radiography group were false positive results. The incidence of lung cancer was 645 cases per 100,000 person-years (1060 cancers) in the low-dose CT group, as compared with 572 cases per 100,000 person-years (941 cancers) in the radiography group (rate ratio, 1.13; 95% confidence interval [CI], 1.03 to 1.23). There were 247 deaths from lung cancer per 100,000 person-years in the low-dose CT group and 309 deaths per 100,000 person-years in the radiography group, representing a relative reduction in mortality from lung cancer with low-dose CT screening of 20.0% (95% CI, 6.8 to 26.7; P=0.004). The rate of death from any cause was reduced in the low-dose CT group, as compared with the radiography group, by 6.7% (95% CI, 1.2 to 13.6; P=0.02).

CONCLUSIONS

Screening with the use of low-dose CT reduces mortality from lung cancer. (Funded by the National Cancer Institute; National Lung Screening Trial ClinicalTrials.gov number, NCT00047385.)

The members of the writing team (who are listed in the Appendix) assume responsibility for the integrity of the article. Address reprint requests to Dr. Christine D. Berg at the Early Detection Research Group, Division of Cancer Prevention, National Cancer Institute, 6130 Executive Blvd., Suite 3112, Bethesda, MD 20892-7346, or at bergc@mail.nih.gov.

*A complete list of members of the National Lung Screening Trial research team is provided in the Supplementary Appendix, available at NEJM.org.

This article (10.1056/NEJMoa1102873) was published on June 29, 2011, at NEJM.org.

N Engl J Med 2011.

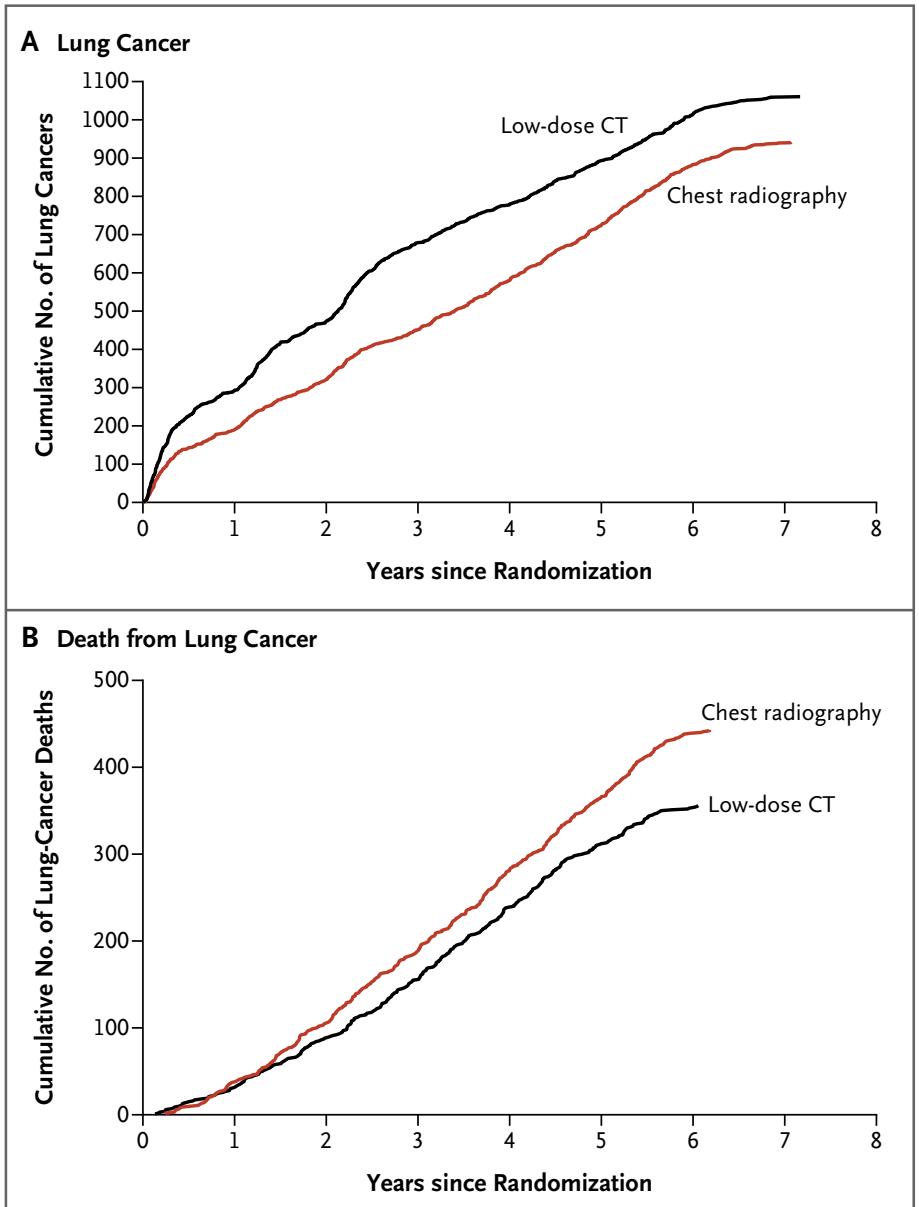


Figure 1. Cumulative Numbers of Lung Cancers and of Deaths from Lung Cancer.

The number of lung cancers (Panel A) includes lung cancers that were diagnosed from the date of randomization through December 31, 2009. The number of deaths from lung cancer (Panel B) includes deaths that occurred from the date of randomization through January 15, 2009.

0.16 Telephone calls; stars in the sky (bios700, 2020)

1. The Poisson distribution says that if, for example, during a given time the average number of calls is μ, then the probability (P_y) of y telephone calls being originated is $e^{-\mu}\mu^y/y!$ One way to derive this law is via a statistical equilibrium argument[33] which leads to the recurrence relation $P_y = P_{y-1} \times \frac{\mu}{y}$.

   Describe the remaining steps in the proof.
2. Mention one other way to derive the distribution, and the main steps involved.
3. An early application of this distribution involved an argument about whether the stars are scattered at random over the heavens, which Newcomb divided into 41,253 spaces of 1 square degree each. He supposed that there were approximately 1500 stars of the fifth and higher brightness level spread at random over these entire 41,253 square degrees of heavens; thus $1500/41{,}253 = 0.03636$ stars per square. The arguments focused on the six brightest stars in the Pleiades (a cluster of stars), and so he calculated the probability that *any square degree selected at random* contains six stars.
   (a) Write down the expression for this probability.
   (b) He went on to calculate the probability that *some one of the 41,253 square degrees would contain six stars.* [The original, written in 1860, had these exact words, and had them in italics; today he might have simply referred to the probability that *one of the 41,253* square degrees would ...]. Write down an expression for this, along with an approximation.
   (c) Explain to a 'p-hacker' why Newcomb's distinction between *any square degree selected at random* and *some one of the 41,253 square degrees* is important.
   (d) With today's computing power, one could go further, and try to find six stars so near together that 'a square degree could be fitted on so as to include them.' How does the probability of such a finding compare with those in (a) and (b)? How might you set about computing it?
4. Although it is an over-simplification, suppose that a business firm has certain busy days every week corresponding to a mean value μ_1, and certain less busy days corresponding to a mean value μ_2. Let the busy portion of the week be π_1, and the less busy portion $\pi_2 = 1 - \pi$. Suppose you wish to express the variations in the number of calls from day to day in terms of one single law of distribution, with a mean value and a variance. Write down expressions for these. [Example is from Erlang]
5. Suppose you wanted an infinite mix of μ's. What mixing distribution might you suggest, and why?
6. List 2 way to fit the parameters of your model from the previous question.

0.17 Vacancy rate in US Supreme Court

Refer to Updating a Classic: 'The Poisson Distribution and the Supreme Court' Revisited

1. Update Table 1 and calculate an updated vacancy rate (expressed as vacancies per year) for the period 1933-2020.
2. Based only on your updated point-estimate of the rate [i.e. without using any information on the health of the current court], what is the probability that the next US president (the one who takes office in 2021) will be able to appoint 0, 1, 2 ... new judges if (s)he stays in office for (a) four (b) 8 years?

[33] See Erlang's derivation in section 3.4 of the history article.

0.18 Mortality rates in the o l d e s t old

Refer to the '*cool Lexis diagram*' one of our graduates (now teaching at Berkeley) sent to JH.

1. Merging all ages and both sexes, calculate the overall incidence density (deaths per person year)
2. Ignoring the sexes, calculate age-band-specific incidence densities: use the age-bands 105-110, 110-115, and 115- .
3. Fit the incidence density (λ) as the following 2-parameter function of calendar time
$$\lambda[date] = \lambda[1955] \times \exp[\beta(date - 1955)].$$
4. Refine the model so that it includes age (and sex?).

0.19 Where Flying Bombs landed in London in 1944

The classic textbook *An introduction to probability theory and its applications* by Feller included the data from this 1946 article. The 'randomness' became a statistical legend and the story was included [along with Deaths from Horsekicks and the rate of audience fidget] in the examples (mainly from Feller's book) of Observations fitting the Poisson Distribution in JH's courses. It was finally debunked in 2018 and 2019. See the note from the Editor of Significance Magazine and the full Significance article, along with an even bigger-picture analysis by Canadian authors in 2018.

1. Write a short paragraph that updates the entry in Feller.[34]
2. Suggest how, if you had easier access to the counts in the 1km grid squares in Figure 5 of the 2018 article, you might model them – especially if you want to follow up on the 'revised target' hypothesis put forward in the 2018 and 2019 articles.

0.20 Re-analysis of data in Student 1907, and in Rutherford, Geiger and Bateman 1910

1. Repeat the calculation of the 2 moments and the GoF statistic for Student's counts from concentration IV (full spatial data in Table I).
2. Do likewise for the counts from concentration III. Note that (unlike those from IV) they show a *slight* extra-Poisson variation. Suggest a model that allows for extra-Poisson variation [Hint: see the 'infinitely compound' model adopted in Greenwood & Yule's 1920 article, section IV], and how you might fit it.
3. How many degrees of freedom did his reference χ^2 distribution have? In light of the work of Greenwood and Yule 1915, pp 117-119; Yule, 1922; and Fisher, 1922 how many should it have? [Pearson's 1900 paper, with may interesting worked examples, can be found here.]
4. Repeat the calculation of the 2 moments and calculate the GoF statistic for the counts in the 'Sum' row of Rutherford and Geiger's table (p 701). See Snow1911. Again, be careful with the number of degrees of freedom.

[34] If you would like to see the original, you can get temporary access via the McGill Library. The entry starts at page 160 in the 3rd Edition.

5. Summarize Bateman's and Erlang's (time-based) derivations of the Poisson distribution. Which 'time-based' argument do your prefer, Bateman's or [recounted in the draft ms.] Erlang's? Why?

0.21 Marsden and Barratt – see JH's 'Rutherford' website

These two physicists are (possibly) the first to point out the important statistical link between what you observe in the 'count' scale and the 'time between events' scale. It is the same one that we encountered in our more modern 'ruptured tires' example.

1. Summarize their argument as to why they thought using the (continuous) time between events scale provides a more stringent test of randomness than the count scale.
2. Instead of their data, collect and analyze your own (bin-count and 'time between events') data from the 2020 re-enactment of minute 1 [the frame-numbers can be regarded as from an 'effectively continuous' time-scale]. Naturally, you won't be able to say a lot from the 8 counts in the 7.5 sec. bins, but maybe there will more of a definite pattern to the distribution of the lengths of the (more numerous) inter-event intervals.
3. If Rutherford and Geiger had counted into 1-minute rather than 1/8 of a minute bins, would they have learned a lot about the frequency law that governs the 1/8 of a minute bins? Why/why not?

0.22 Underdispersion: A statistical anomaly in reported Covid data

1. In your own words, summarize the article by Dmitry Kobak in Significance in April 2022.
2. From the list in the 'Does the Poisson Distribution apply to...?' in section 1.1 of the Notes, can you identity any context where there might be *under*-dispersion? Add one of your own.