Inference about a Population Rate (λ)

JH notes on rates

Sahir Bhatnagar and James Hanley

EPIB 607
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/EPIB607/

November 3, 2018

Poisson Model for Sampling Variability of a Count in a Given Amount of “Experience”

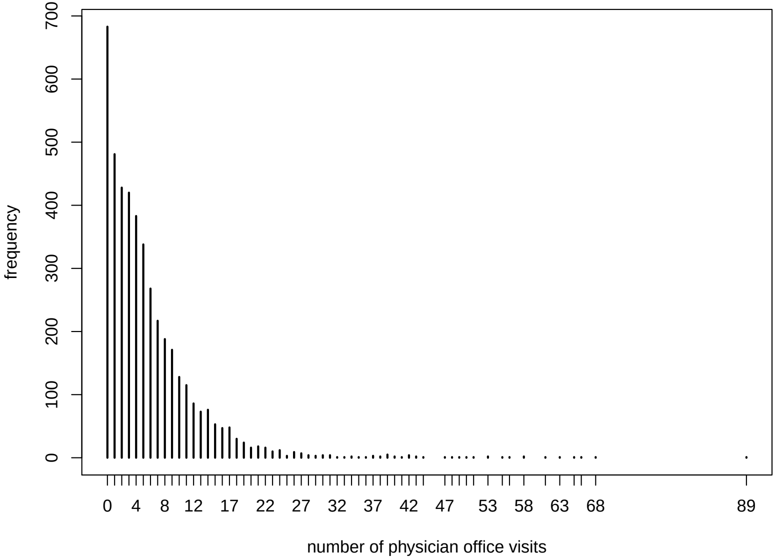
Motivating example: Demand for medical care

- Data from the US National Medical Expenditure Survey (NMES) for 1987/88

- 4406 individuals, aged 66 and over, who are covered by Medicare, a public insurance program

- The objective of the study was to model the demand for medical care - as captured by <u>the number</u> of physician/non-physician office and hospital outpatient visits - by the covariates available for the patients.

Motivating example: Demand for medical care



Some observations about the previous plot

- Discrete outcome $\rightarrow$ 0, 1, 2, 3, ... visits
- There are rare events, e.g. 1 individual with 89 visits
- The data are far from normally distributed
- Can theoretically go on forever

The Poisson Distribution

- The binomial distribution was derived by starting with an experiment consisting of trials or draws and applying the laws of probability to various outcomes of the experiment.

- There is no simple experiment on which the Poisson distribution is based, although we will shortly describe how it can be obtained by certain limiting operations.

The Poisson Distribution: what it is, and features

- The (infinite number of) probabilities $P_0, P_1, ..., P_y, ...,$ of observing $Y = 0, 1, 2, \ldots, y, \ldots$ events in a given amount of "experience."

- These probabilities, $P(Y = k) \rightarrow$ `dpois()`, are governed by a single parameter, the mean $E[Y] = \mu$ which represents the expected **number** of events in the amount of experience actually studied.

- We say that a random variable $Y \sim \mathrm{Poisson}(\mu)$ distribution if

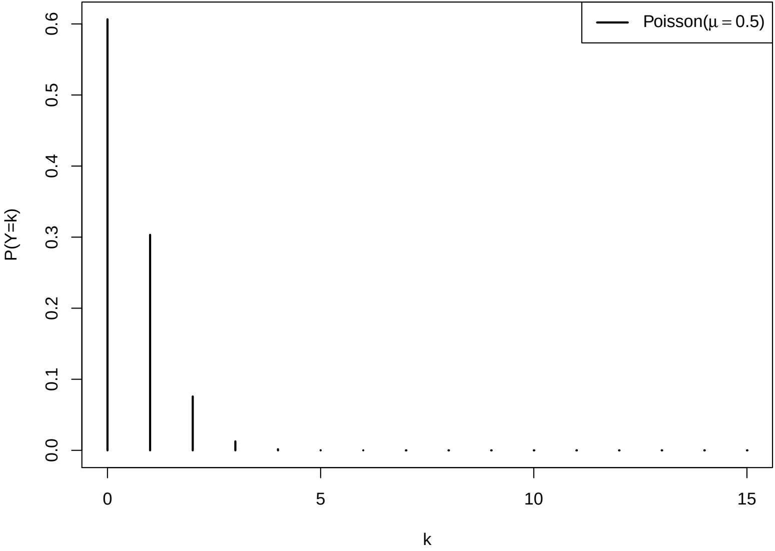
$$P(Y = k) = \frac{\mu^k}{k!} e^{-\mu}, \quad k = 0, 1, 2, \ldots$$

- Note: in `dpois()` μ is referred to as `lambda`

- Note the distinction between μ and λ
  - μ: expected **number** of events
  - λ: **rate** parameter

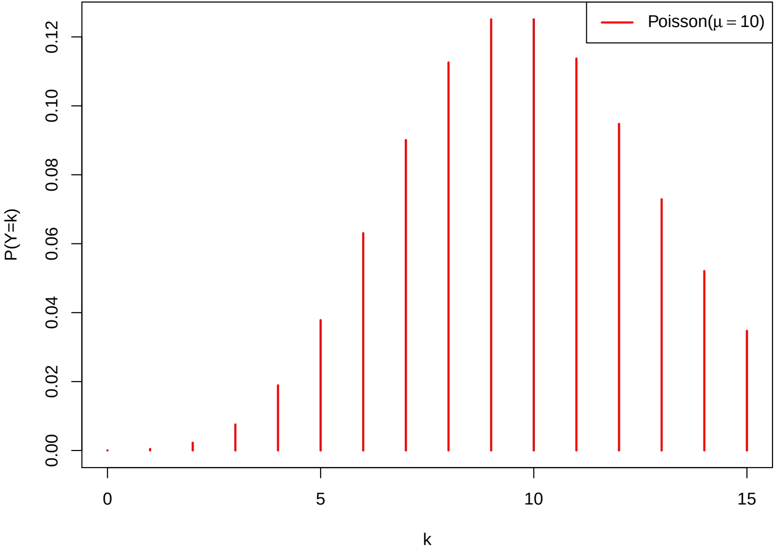
The probability mass function for $\mu = 0.5$

```
dpois(x = 0:15, lambda = 0.5)
```

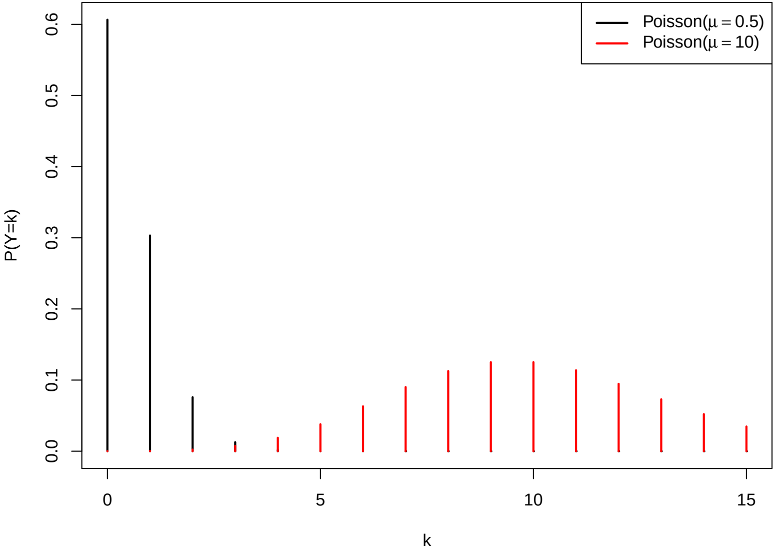


The probability mass function for $\mu = 10$

```
dpois(x = 0:15, lambda = 10)
```



The probability mass function



The Poisson Distribution: what it is, and features

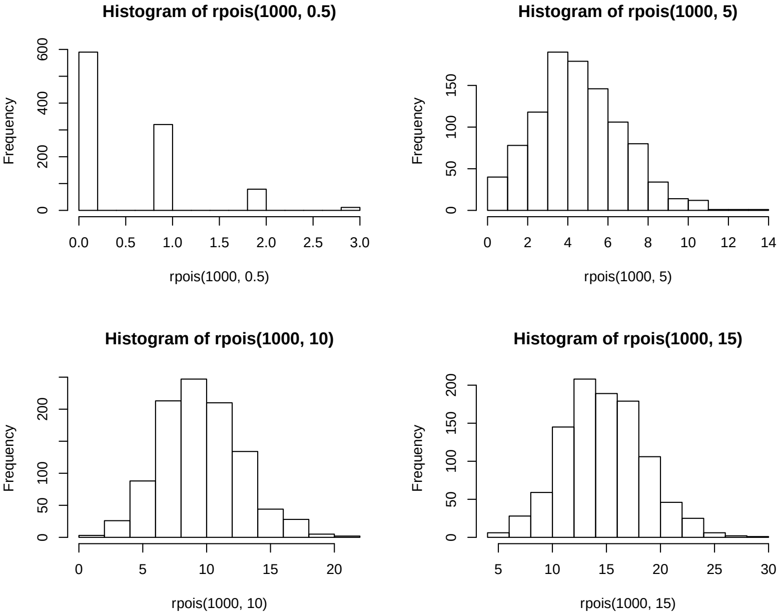
- $\sigma_Y^2 = \mu \ \rightarrow \ \ \sigma_Y = \sqrt{\mu}.$

- Approximated by $\mathcal{N}(\mu, \sqrt{\mu})$ when $\mu >> 10$

- Open-ended (unlike Binomial), but in practice, has finite range.

- Poisson data sometimes called ”numerator only”: (unlike Binomial) may not “see” or count “non-events”

Normal approximation to Poisson is the CLT in action



How it arises

- Count of events or items that occur randomly, with low homogeneous intensity, in time, space, or 'item'-time (e.g. person–time).

- Binomial(n, π) when $n \rightarrow \infty$ and $\pi \rightarrow 0$, but $n \times \pi = \mu$ is finite.

- $Y \sim \textit{Poisson}(\mu_Y)$ if time (T) between events follows an $T \sim \text{Exponential}(\mu_T = 1/\mu_Y)$. http://www.epi.mcgill.ca/hanley/bios601/Intensity-Rate/Randomness_poisson.pdf

- As sum of ≥ 2 *independent* Poisson random variables, with same **or different** μ's:
  $Y_1 \sim \text{Poisson}(\mu_1)\ \ Y_2 \sim \text{Poisson}(\mu_2) \Rightarrow Y = Y_1 + Y_2 \sim \text{Poisson}(\mu_1 + \mu_2)$.

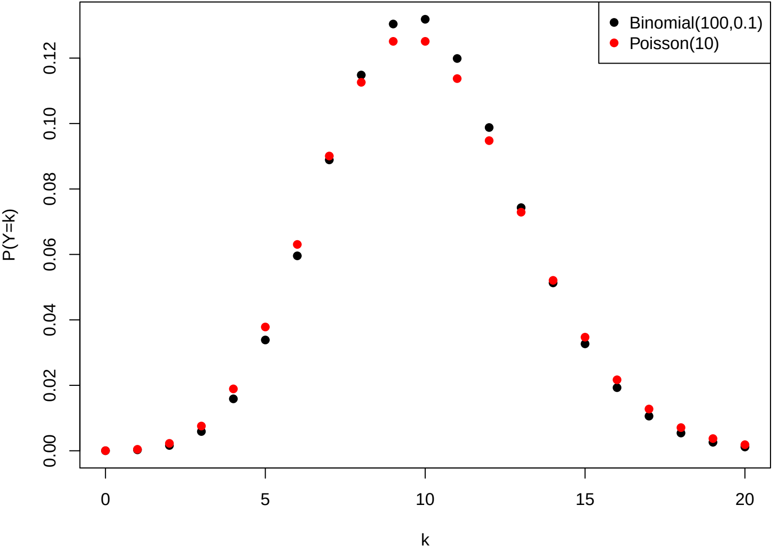
Poisson distribution as a limit

The rationale for using the Poisson distribution in many situations is provided by the following proposition.

Proposition 1 (Limit of a binomial is Poisson)

Suppose that $Y \sim$ *Binomial*(n, π). *If we let* $\pi = \mu/n$, *then as* $n \to \infty$, *Binomial*$(n, \pi) \to$ *Poisson*(μ). *Another way of saying this: for large* n *and small* π, *we can approximate the Binomial*(n, π) *probability by the Poisson*$(\mu = n\pi)$.

Poisson approximation to the Binomial



Examples

- numbers of asbestos fibres
- deaths from horse kicks*
- needle-stick or other percutaneous injuries
- bus-driver accidents*
- twin-pairs*
- radioactive disintegrations*
- flying-bomb hits*
- white blood cells
- typographical errors
- cell occupants – in a given volume, area, line-length, population-time, time, etc. [1]

[1] * included in http://www.epi.mcgill.ca/hanley/bios601/Intensity-Rate/

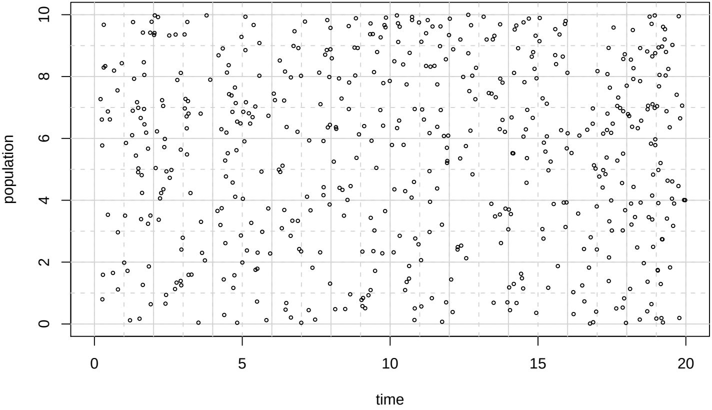


Fig.: Events in Population-Time randomly generated from intensities that are constant within (2 squares high by 2 squares wide) ‘panels’, but vary between such panels. In Epidemiology, each square might represent a number of units of population-time, and each dot an event.

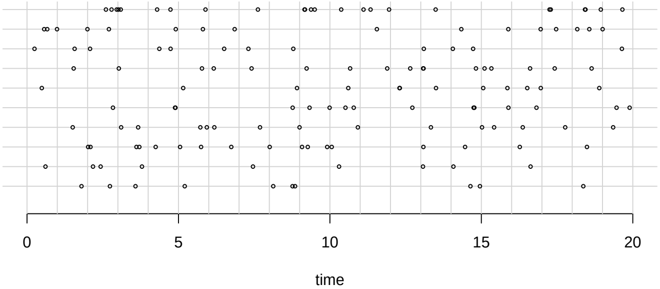


Fig.: Events in Time: 10 examples, randomly generated from constant over time intensities. Simulated with 1000 Bernoulli($_{\text{small}}\ \pi$)'s per time unit.

Does the Poisson Distribution apply to.. ?

1. Yearly variations in numbers of persons killed in plane crashes
2. Daily variations in numbers of births
3. Weekly variations in numbers of births
4. Daily variations in numbers of deaths
5. Daily variations in numbers of traffic accidents
6. Variations across cookies/pizzas in numbers of chocolate chips/olives

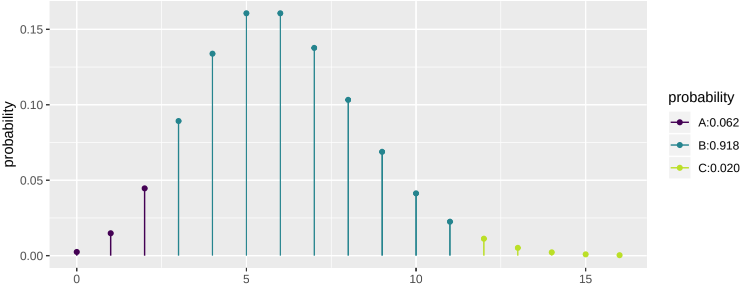
Inference regarding μ, based on observed count y

Confidence interval for μ

- If the CLT hasn't kicked in, then the usual CI might not be appropriate:

$$\text{point-estimate} \pm z^{\star} \times \text{standard error}$$

```r
mosaic::xqpois(c(0.025, 0.975), lambda = 6)
```



```
## [1]  2 11
```

Confidence interval for μ

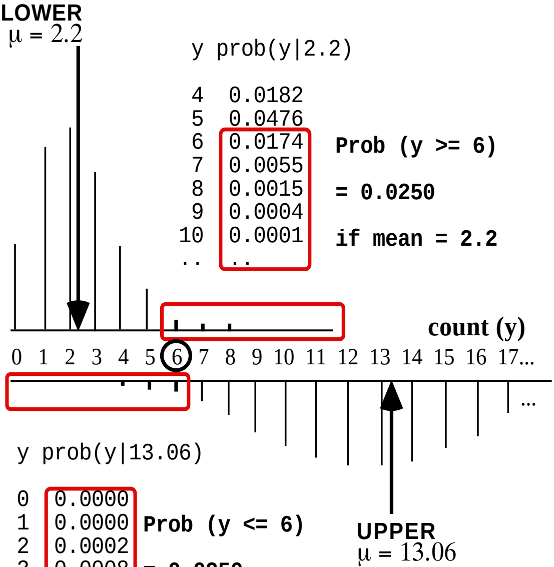
```r
manipulate::manipulate(
mosaic::xqpois(c(0.025, 0.975), lambda = LAMBDA),
LAMBDA = manipulate::slider(1, 200, step = 1))
```

Confidence interval for μ

- Similar to the binomial (Clopper-Pearson CI), we consider a *first-principles* $100(1-\alpha)\%$ CI $[\mu_{LOWER},\ \mu_{UPPER}]$ such that

$$P(Y \geq y \mid \mu_{LOWER}) = \alpha/2 \quad \text{and} \quad P(Y \leq y \mid \mu_{UPPER}) = \alpha/2.$$

- For example, the 95% CI for μ, based on $y = 6$, is $[\underline{2.20}, \underline{13.06}]$.



y	prob(y|2.2)
4	0.0182
5	0.0476
6	0.0174
7	0.0055
8	0.0015
9	0.0004
10	0.0001
..	..

Prob (y >= 6)

= 0.0250

if mean = 2.2

y	prob(y|13.06)
0	0.0000
1	0.0000
2	0.0002
3	0.0008
4	0.0026
5	0.0067
6	0.0147
7	0.0274
..	..

Prob (y <= 6)

= 0.0250

if mean = 13.06

6 **observed count**

Poisson 95% CI for μ when $y = 6$

```r
# upper limit --> lower tail needs 2.5%
manipulate::manipulate(
mosaic::xppois(6, lambda = LAMBDA),
LAMBDA = manipulate::slider(0.01, 20, step = 0.01))
```

```r
# lower limit --> upper tail needs 2.5%
# when lower.tail=FALSE, ppois doesnt include k, i.e., P(Y > k)
manipulate::manipulate(
mosaic::xppois(5, lambda = LAMBDA, lower.tail = FALSE),
LAMBDA = manipulate::slider(0.01, 20, step = 0.01))
```

Confidence interval for μ

- 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 **Gamma** 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 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*.

80%, 90% and 95% CI for mean count μ if we observe 0 to 30 events in a certain amount of experience

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

95% CI for mean count μ with **q** function

- To obtain these in R we use the natural link between the Poisson and the *gamma* distributions.[2]

- In R, e.g., the 95% limits for μ based on $y = 6$ are obtained as

  ```
  qgamma(p = c(0.025,0.975), shape = c(6, 7))
  ```

  ```
  ## [1]  2.201894 13.059474
  ```

- More generically, for *any y*, as

  ```
  qgamma(p = c(0.025,0.975), shape = c(y, y+1))
  ```

[2] details found here

95% CI for mean count μ with canned function

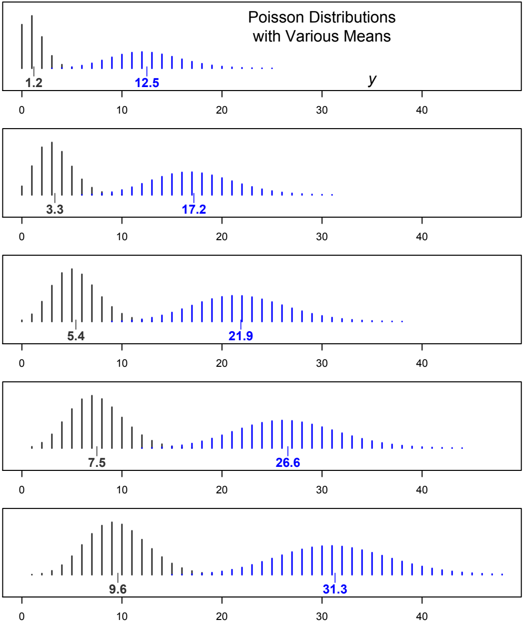
- These limits can also be found using the canned function in R

```r
stats::poisson.test(6)
```

```
##
## ^^IExact Poisson test
##
## data:  6 time base: 1
## number of events = 6, time base = 1, p-value = 0.0005942
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##   2.201894 13.059474
## sample estimates:
## event rate
##          6
```

z-based confidence intervals

once μ is in the upper teens, the Poisson → the Normal



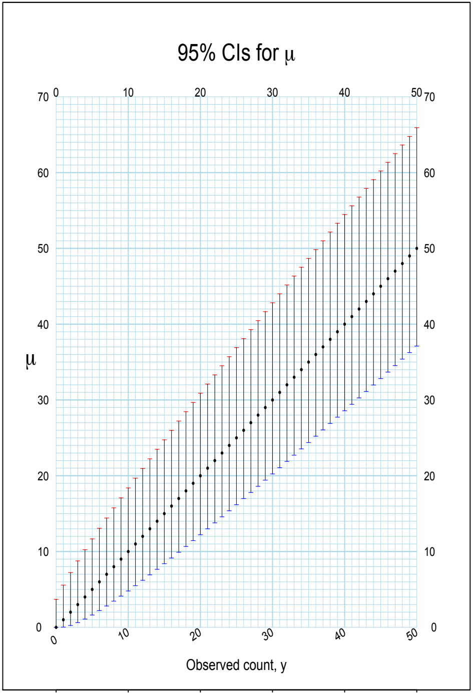
z-based confidence intervals

- Thus, a plus/minus CI based on SE = $\hat{\sigma} = \sqrt{\hat{\mu}} = \sqrt{y}$, is simply
$$[\mu_L,\ \mu_U] = y\ \pm\ z^{\star} \times \sqrt{y}.$$

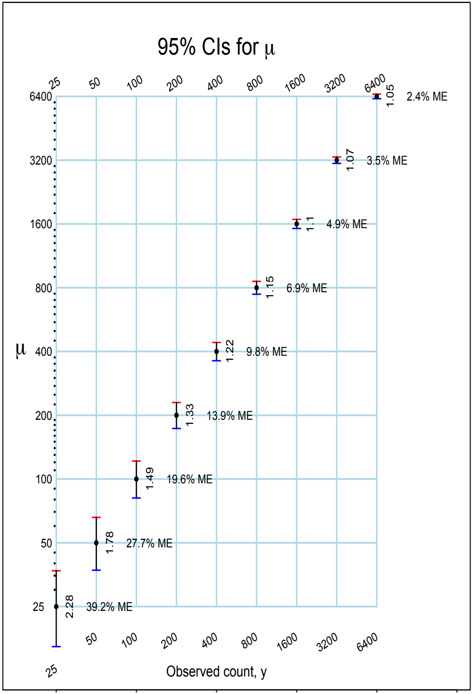
- Equivalently we can use the **q** function:
$$qnorm(p = c(0.025, 0.975), mean = y, sd = \sqrt{y})$$

- 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) distribution, $\sigma_Y = \sqrt{\mu}$, so we calculate a "model-based" SE.



95% CIs for μ



95% CIs for μ

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

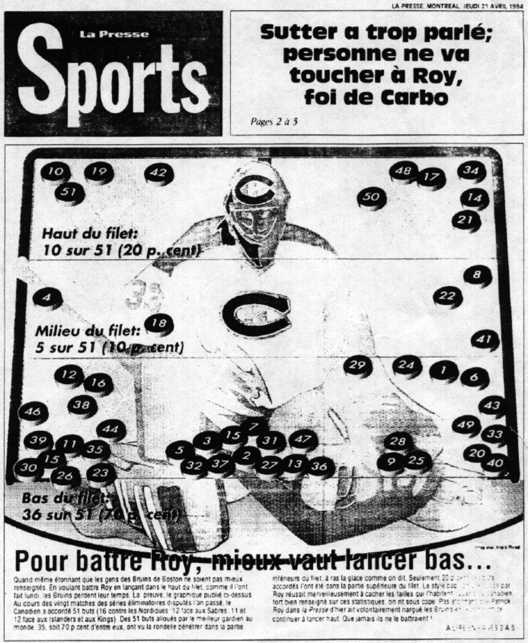
Rates are better for comparisons

year	deaths (y)
1971	33
2002	211

Table: Deaths from lung cancer in the age-group 55-60 in Quebec in 1971 and 2002

A researcher asks: Is the situation getting worse over time for lung cancer in this age group?

Your reply: **What's the denominator??**



LA PRESSE, MONTREAL, JEUDI 21 AVRIL 1994

La Presse

Sports

Sutter a trop parlé; personne ne va toucher à Roy, foi de Carbo

Pages 2 à 5

10
19
42
51
48
17
34
50
14
21
Haut du filet:
10 sur 51 (20 p. cent)
8
22
4
18
Milieu du filet:
5 sur 51 (10 p. cent)
41
29
24
1
6
12
16
38
46
43
44
7
15
3
31
47
28
49
33
39
11
35
5
2
20
15
30
26
23
32
37
27
13
36
9
25
40
Bas du filet:
36 sur 51 (70 p. cent)

Infographie André Rivest

Pour battre Roy, mieux vaut lancer bas...

Quand même étonnant que les gens des Bruins de Boston ne soient pas mieux renseignés. En voulant battre Roy en lançant dans le haut du filet, comme il l'ont fait lundi, les Bruins perdent leur temps. La preuve, le graphique publié ci-dessus. Au cours des vingt matches des séries éliminatoires disputés l'an passé, le Canadien a accordé 51 buts (16 contre les Nordiques, 12 face aux Sabres, 11 et 12 face aux Islanders et aux Kings). Des 51 buts alloués par le meilleur gardien au monde, 35, soit 70 p.cent d'entre eux, ont vu la rondelle pénétrer dans la partie inférieure du filet, à ras la glace comme on dit. Seulement 20 p. [illegible] buts accordés l'ont été dans la partie supérieure du filet. Le style [illegible] par Roy réussit merveilleusement à cacher les failles qui l'habitent. [illegible] Canadien, fort bien renseigné sur ces statistiques, on rit sous cape. Pas étonnant que Patrick Roy dans *la Presse* d'hier ait volontairement nargué les Bruins en [illegible] continuer à lancer haut. Que jamais ils ne le battraient !

A LIRE EN PAGES 2 À 5

Rates are better for comparisons

- 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**, **intensity** or **incidence density (ID)**.

year	deaths (y)	person-time (PT)	rate ($\hat{\lambda}$)
1971	33	131,200 years	25 per 100,000 women-years
2002	211	232,978 years	91 per 100,000 women-years

Table: Deaths from lung cancer in the age-group 55-60 in Quebec in 1971 and 2002

Rates are better for comparisons

- The *statistic*, the empirical rate or empirical incidence density, is
$$\textit{rate} = \hat{ID} = \hat{\lambda} = y/\mathrm{PT}.$$

- 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 $\hat{ID}$ or $\hat{\lambda}$ as a point estimate of the (theoretical) Incidence Density *parameter*, ID or λ.

CI for the rate parameter λ

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

$$\begin{aligned}\lambda &= \mu \div \mathrm{PT}\\ \hat{\lambda} &= \hat{\mu} \div \mathrm{PT}\\ &= y \div \ \mathrm{PT}\end{aligned}$$

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

CI for the rate parameter λ

- $y = 211$ deaths from lung cancer in 2002 leads to a 95% CI for μ:

  ```
  qgamma(p = c(0.025, 0.975), shape = c(211, 212))
  ```

  ```
  ## [1] 183.4885 241.4725
  ```

- From this we can calculate the 95% CI **per 100,000 WY** for λ using a PT=232978 years:

  ```
  qgamma(p = c(0.025, 0.975), shape = c(211, 212)) / 232978 * 1e5
  ```

  ```
  ## [1]  78.75788 103.64607
  ```

- $y = 33$ deaths from lung cancer in 131200 women-years in 1971 leads to a 95% CI per 100,000 WY for λ of

  ```
  qgamma(c(0.025,0.975), c(33,34)) / 131200 * 1e5
  ```

  ```
  ## [1] 17.31378 35.32338
  ```

CI for the rate parameter λ using canned function

```r
stats::poisson.test(x = 33, T = 131200)
```

```
##
## ^^IExact Poisson test
##
## data:  33 time base: 131200
## number of events = 33, time base = 131200, p-value < 2.2e-16
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##  0.0001731378 0.0003532338
## sample estimates:
##   event rate
## 0.0002515244
```

Test of $H_0 : \mu = \mu_0 \quad \Leftrightarrow \quad \lambda = \lambda_0$

Statistical evidence and the *p*-value

Recall:

- P-Value = Prob[y or more extreme | 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 surrounding nuclear stations

- Cancers in area surrounding the Douglas Point nuclear station
- Denote by $\{CY_1, CY_2, \ldots\}$ the numbers of Douglas Point $\underline{c}$hild-$\underline{y}$ears of experience in the various age categories that were pooled over.
- Denote by $\{\lambda_1^{Ont}, \lambda_2^{Ont}, \ldots\}$ 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

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

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

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

Age *Standardized Incidence Ratio (SIR)* = $O/E = 2/0.57 = 3.5$.

Q: Is the $O = 2$ significantly higher than $E = 0.57$

Question:

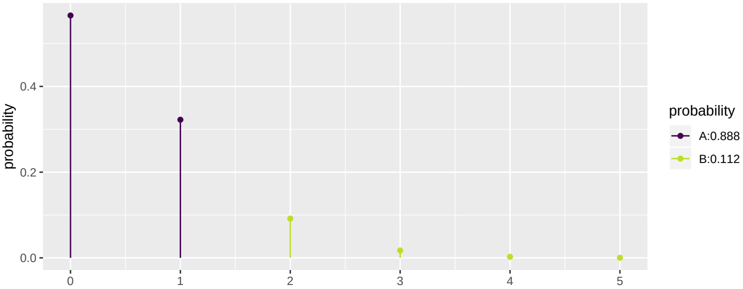
- Is the $y = 2$ cases of leukemia observed in the Douglas Point experience statistically significantly higher than the $E = 0.57$ cases "expected" for this many child-years of observation if in fact the rates in Douglas Point and the rest of Ontario were the same?

- Or, is the $y = 2$ observed in this community compatible with $H_0 : y \sim \mathrm{Poisson}(\mu = 0.57)$?

A: Is the $O = 2$ significantly higher than $E = 0.57$

- Answer: Under H_0, the age-specific numbers of leukemias $\{y_1 = O_1, y_2 = O_2, \ldots\}$ 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 $\mu = 0.57$.

```r
mosaic::xppois(1, lambda = 0.57, lower.tail = FALSE)
```



```
## [1] 0.1121251
```

95% CI for the SIR by hand

- 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.

- The $y = 2$ cases translates to
  - 95% CI for $\mu_{DP} \rightarrow$ [0.24, 7.22]
  - 95% CI for the SIR $\rightarrow$ [0.24/0.57, 7.22/0.57]=[0.4, 12.7].

95% CI for the SIR using canned function

- We can *trick* `stats::poisson.test` to get the same CI by putting time as 0.57:

```
stats::poisson.test(x=2,T=0.57)
```

```
##
## ^^IExact Poisson test
##
## data:  2 time base: 0.57
## number of events = 2, time base = 0.57, p-value = 0.1121
## alternative hypothesis: true event rate is not equal to 1
## 95 percent confidence interval:
##   0.4249286 12.6748906
## sample estimates:
## event rate
##   3.508772
```

Examples of Poisson and not-so Poisson variation

Examples of Poisson, ‘Extra-Poisson’ and ‘Less-than-Poisson’ Variation

- One definition of epidemiology:

  “Disease is not distributed at random”

- When might we expect (just) Poisson variation?

- When might we expect more than (i.e., extra-) Poisson variation?

  And what do we do about it?

- Can you think of a context where counts show less-than-Poisson variation?

Deaths from Horsekicks

“The chance of a man being killed by horsekick on any one day is exceedingly small, but if an army corps of men are exposed to this risk for a year, often one or more of them will be killed in this way" [R. A. Fisher, 1925, using just 10 of 14 corps used by Bortkiewicz in his 1898 work, Das Gesetz der kleinen Zahlen.]

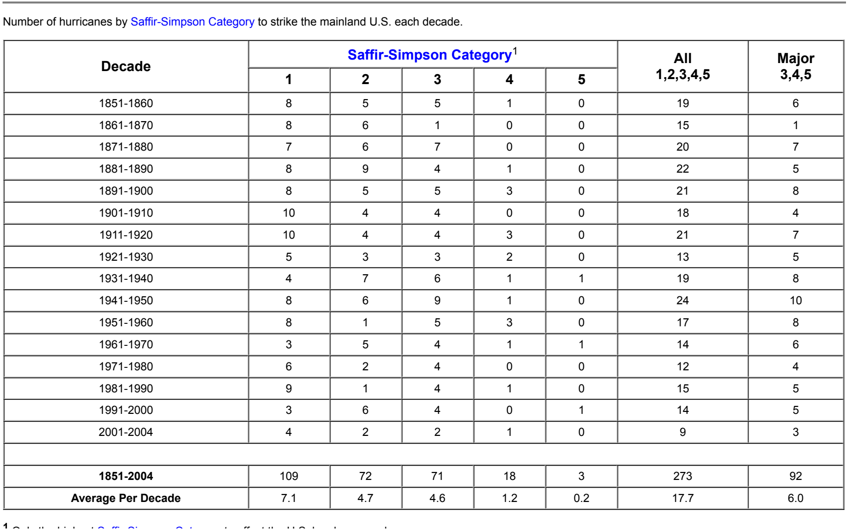
10 Prussian army-corps for 20 years, i.e., 200 observations, 1 per “corps-year” (CY).

No. Deaths in corps-year	Frequencies) (No. “corps-years” with *y* deaths)		
y	Observed	`Expected`$^{\$}$	$y \times$ *Obs. Freq.*
0	109	`108.7`	0
1	65	`66.3`	65
2	22	`20.2`	44
3	3	`4.1`	9
4	1	`0.6`	4
5	-	`0.1`	-
Sum:	200	`200`	No. Deaths: 122

$\hat{\lambda} = \bar{y} = \frac{122 \ deaths}{200CY} = 0.61/CY$; $^{\$}$`200 * dpois(0:6,lambda=0.61)`

- Poisson Distribution $\Rightarrow$ SD = $\sqrt{mean}$.
- In this series, the SD of the 200 *y*’s is 0.78, which happens to equal $\sqrt{0.61}$.
- “ SD = $\sqrt{mean}$ ” $\not\Rightarrow$ Poisson Distribution, but there is close agreement (’fit’) between the observed and expected/theoretical distributions.

`http://www.epi.mcgill.ca//hanley/bios601/Intensity-Rate/horsekicks_anthrax_poisson.pdf`
`http://www.epi.mcgill.ca//hanley/bios601/Intensity-Rate/horsekicks.txt`



Number of hurricanes by Saffir-Simpson Category to strike the mainland U.S. each decade.

Decade	Saffir-Simpson Category[1]					All 1,2,3,4,5	Major 3,4,5
	1	2	3	4	5		
1851-1860	8	5	5	1	0	19	6
1861-1870	8	6	1	0	0	15	1
1871-1880	7	6	7	0	0	20	7
1881-1890	8	9	4	1	0	22	5
1891-1900	8	5	5	3	0	21	8
1901-1910	10	4	4	0	0	18	4
1911-1920	10	4	4	3	0	21	7
1921-1930	5	3	3	2	0	13	5
1931-1940	4	7	6	1	1	19	8
1941-1950	8	6	9	1	0	24	10
1951-1960	8	1	5	3	0	17	8
1961-1970	3	5	4	1	1	14	6
1971-1980	6	2	4	0	0	12	4
1981-1990	9	1	4	1	0	15	5
1991-2000	3	6	4	0	1	14	5
2001-2004	4	2	2	1	0	9	3
1851-2004	109	72	71	18	3	273	92
Average Per Decade	7.1	4.7	4.6	1.2	0.2	17.7	6.0

15 FULL DECADES

mean(sd)	7.0(2.3)	4.7(2.2)	4.6(1.8)	1.1(1.1)	0.2(0.4)	17.6(3.2)	5.9(2.3)
$\sqrt{mean}$:	2.6	2.2	2.1	1.1	0.4	4.2	2.4 !!

Can use regression models to fit temporal trends to these (‘noisy’) data.

Simulate 15 decades ± a time pattern: `summary(rpois(15, lambda= ..  ))`

Source: `https://www.nhc.noaa.gov/pastdec.shtml`

Rate of *de novo* mutations and the importance of father's age to disease risk

http://www.epi.mcgill.ca/hanley/bios601/FathersAgeMutations.pdf

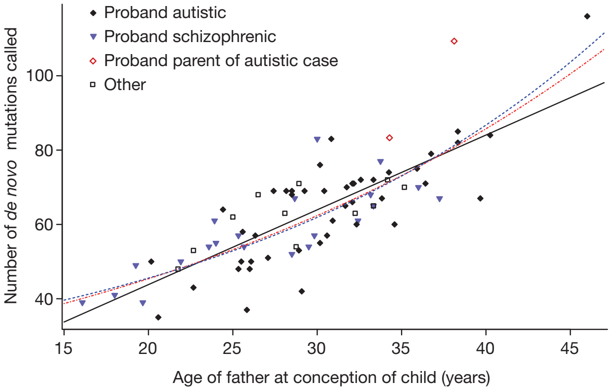


Figure 2 | Father's age and number of *de novo* mutations. The number of *de novo* mutations called is plotted against father's age at conception of child for the 78 trios. The solid black line denotes the linear fit. The dashed red curve is based on an exponential model fitted to the combined mutation counts. The dashed blue curve corresponds to a model in which maternal mutations are assumed to have a constant rate of 14.2 and paternal mutations are assumed to increase exponentially with father's age.

$\mu[y|age] \approx 2 \times age$; $\approx$ Poisson variation of y's around each $\mu[y|age]$; i.e. $\sigma[y|x] \approx \sqrt{\mu[y|age]}$

What events are these?

http://www.epi.mcgill.ca/hanley/mysteryData/

What about Daily No.s of BIRTHS 9 MONTHS AFTER BLACKOUT ?

On November 9, 1965, the power went out in New York City, and it stayed out for a day – the **Great Blackout**. Nine months later, the newspapers suggested that New York was experiencing a baby boom. The columns below show the numbers of babies born every day during a 25-day period, Aug1-Aug25, centered nine months and ten days after the Great Blackout. These numbers average out to 436. This turns out not to be unusually high for New York.

Aug.	Births	**Day**
1	451	Mon
2	468	Tue
3	429	Wed
4	448	Thu
5	466	Fri
6	377	Sat
7	344	Sun
8	448	Mon
9	438	Tue
10	455	Wed
11	468	Thu
12	462	Fri
13	405	Sat
14	377	Sun
15	451	Mon
16	497	Tue
17	458	Wed
18	429	Thu
19	434	Fri
20	410	Sat
21	351	Sun
22	467	Mon
23	508	Tue
24	432	Wed
25	426	Thu

Statistics (2nd Ed) by D Freedman, R Pisani et al

- “Apparently, the *New York Times, NYT* sent a reporter around to a few hospitals on Monday, August 8, and Tuesday, August 9, nine months after the blackout.
- The hospitals reported that their obstetrics wards were busier than usual probably because of the general pattern that weekends are slow, Mondays and Tuesdays are busy.
- Headline, p.1, Wed, Aug10, ’66: “**Births Up 9 Months After the Blackout**’ http://www.epi.mcgill.ca//hanley/bios601/Intensity-Rate/BirthsUp9MonthsAfterBlackout.pdf
- This [overlooked source of variation] seems to be the origin of the baby-boom myth.”

Closer to Poisson variation if use week as unit, or if one ’conditions on’ day-of-week; Over the year, some variation in weekly numbers.

Freedman cites Izenman & Zabell, "Babies and the blackout: the genesis of a misconception," Soc. Sci. Res.,1981,282- 99. available here http://www.epi.mcgill.ca/hanley/bios601/Intensity-Rate/GenesisOfAMisconception.pdf

Daily (and hourly!) variations in 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 http://www.epi.mcgill.ca/hanley/mysteryData/

Are “accidents” distributed “randomly” over bus drivers?

Observed & “expected” numbers of accidents during a 3-year period among 708 Northern Ireland Transport Authority bus drivers. [Table 2.5]

No. accidents in 3-year period	Number of drivers with y accidents	
(y)	Observed	Expected[$]
0	117	71.5
1	157	164.0
2	158	187.9
3	115	143.6
4	78	82.3
5	44	37.7
6	21	14.4
7	7	4.7
8	6	1.4
9	1	0.3
10	3	0.1
11	1	0.0
	708	708

$\hat{\lambda} = \bar{y} = \frac{1623 \textit{ accidents}}{708 \textit{ drivers}} = \frac{2.29}{\textit{driver}}$.

SD of the 708 y's :1.86; $\sqrt{2.29} = 1.51$.

[$] `708 * dpois(0:11,lambda=2.29)`

Colton:

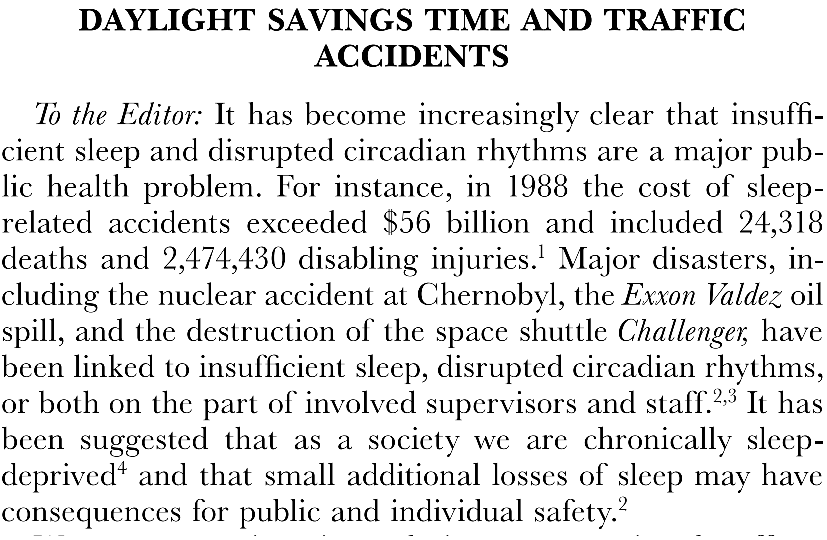
"Comparison of observed and expected frequencies:

- More than the expected number of drivers with no accidents
- More than the expected number of drivers with five or more accidents
- These data suggest that the accidents did not occur completely at random;

  in fact it appears that there is some indication of accident proneness.
- Ignoring this variation makes for (model-based) Standard Errors (SE's) and CI's that are too narrow, and that can lead to ‘false positive’ findings
- Encoding this (not identifiable) variation in ‘random effects’ or ‘random-intercept’ regression models makes for more realistic SE's and CI's.

http://www.epi.mcgill.ca/hanley/bios601/Intensity-Rate/AccidentsBusDrivers.pdf

DAYLIGHT SAVINGS TIME AND TRAFFIC ACCIDENTS, NEJM 1996.04,04



DAYLIGHT SAVINGS TIME AND TRAFFIC ACCIDENTS

To the Editor: It has become increasingly clear that insufficient sleep and disrupted circadian rhythms are a major public health problem. For instance, in 1988 the cost of sleep-related accidents exceeded $56 billion and included 24,318 deaths and 2,474,430 disabling injuries.[1] Major disasters, including the nuclear accident at Chernobyl, the *Exxon Valdez* oil spill, and the destruction of the space shuttle *Challenger,* have been linked to insufficient sleep, disrupted circadian rhythms, or both on the part of involved supervisors and staff.[2,3] It has been suggested that as a society we are chronically sleep-deprived[4] and that small additional losses of sleep may have consequences for public and individual safety.[2]

https://www.nejm.org/doi/full/10.1056/NEJM199604043341416

We can use noninvasive techniques to examine the effects of minor disruptions of circadian rhythms on normal activities if we take advantage of annual shifts in time keeping. More than 25 countries shift to daylight savings time each spring and return to standard time in the fall. The spring shift results in the loss of one hour of sleep time (the equivalent in terms of jet lag of traveling one time zone to the east), whereas the fall shift permits an additional hour of sleep (the equivalent of traveling one time zone to the west). Although one hour's change may seem like a minor disruption in the cycle of sleep and wakefulness, measurable changes in sleep pattern persist for up to five days after each time shift.[5] This leads to the prediction that the spring shift, involving a loss of an hour's sleep, might lead to an increased number of "microsleeps," or lapses of attention, during daily activities and thus might cause an increase in the probability of accidents, especially in traffic. The additional hour of sleep gained in the fall might then lead conversely to a reduction in accident rates.

We used data from a tabulation of all traffic accidents in Canada as they were reported to the Canadian Ministry of Transport for the years 1991 and 1992 by all 10 provinces. A total of 1,398,784 accidents were coded according to the date of occurrence. Data for analysis were restricted to the Monday preceding the week of the change due to daylight savings time, the Monday immediately after, and the Monday one week after the change, for both spring and fall time shifts. Data from the province of Saskatchewan were excluded because it does not observe daylight savings time. The analysis of the spring shift included 9593 accidents and that of the fall shift 12,010. The resulting data are shown in Figure 1.

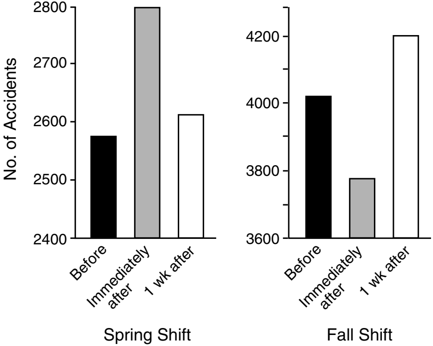
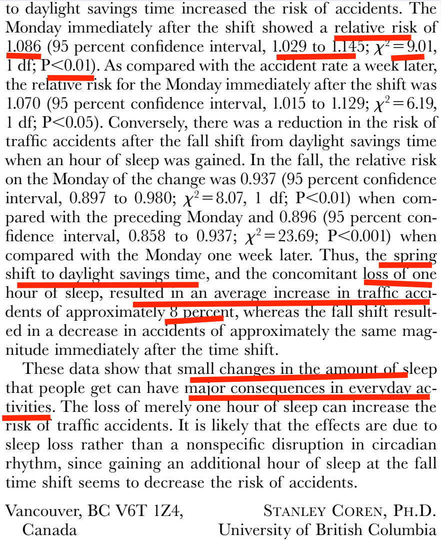


Figure 1. Numbers of Traffic Accidents on the Mondays before and after the Shifts to and from Daylight Savings Time for the Years 1991 and 1992.

There is an increase in accidents after the spring shift (when an hour of sleep is lost) and a decrease in the fall (when an hour of sleep is gained).



to daylight savings time increased the risk of accidents. The Monday immediately after the shift showed a relative risk of 1.086 (95 percent confidence interval, 1.029 to 1.145; $\chi^2 = 9.01$, 1 df; $P < 0.01$). As compared with the accident rate a week later, the relative risk for the Monday immediately after the shift was 1.070 (95 percent confidence interval, 1.015 to 1.129; $\chi^2 = 6.19$, 1 df; $P < 0.05$). Conversely, there was a reduction in the risk of traffic accidents after the fall shift from daylight savings time when an hour of sleep was gained. In the fall, the relative risk on the Monday of the change was 0.937 (95 percent confidence interval, 0.897 to 0.980; $\chi^2 = 8.07$, 1 df; $P < 0.01$) when compared with the preceding Monday and 0.896 (95 percent confidence interval, 0.858 to 0.937; $\chi^2 = 23.69$; $P < 0.001$) when compared with the Monday one week later. Thus, the spring shift to daylight savings time, and the concomitant loss of one hour of sleep, resulted in an average increase in traffic accidents of approximately 8 percent, whereas the fall shift resulted in a decrease in accidents of approximately the same magnitude immediately after the time shift.

These data show that small changes in the amount of sleep that people get can have major consequences in everyday activities. The loss of merely one hour of sleep can increase the risk of traffic accidents. It is likely that the effects are due to sleep loss rather than a nonspecific disruption in circadian rhythm, since gaining an additional hour of sleep at the fall time shift seems to decrease the risk of accidents.

Vancouver, BC V6T 1Z4,
Canada

STANLEY COREN, PH.D.
University of British Columbia

Vincent: `http://www.epi.mcgill.ca/hanley/bios601/Intensity-Rate/Sleep-Vincent.pdf` The results of both a graphical analysis and the variability estimation of 10 years of data failed, as had an earlier study, to support Coren's hypothesis. [JH: Coren's χ^2 statistic assumes that the 2 'before' counts are like 2 Geiger counts generated by a uniform source, and ignores all outside influences.]

Other examples of "lumpy" counts [“extra-Poisson” variation]

Examples that don’t fit Poisson distribution (or, not without further aggre-/segre-gation)

- Yearly variations in numbers of persons killed in plane crashes
  [Yearly variations in numbers of plane crashes may be closer to Poisson –apart from some extra variation over time due to improvements in safety, fluctuations in numbers of flights etc.]
- Daily/Yearly variations in numbers of deaths [variation over seasons, ’flu’ years]
- Daily numbers of Sudden Infant Deaths
  `https://www.ncbi.nlm.nih.gov/pubmed/21059188`
- Yearly numbers/incidence of hospitalized injuries in a region
  `http://www.epi.mcgill.ca/hanley/c609/Material/LidkopingALL.pdf`
- Lethality of Civilian Active Shooter Incidents With and Without Semiautomatic Rifles in the United States [The authors used a negative binomial model that allows extra-Poisson variation.] A bootstrap CI would also be appropriate.
  `https://jamanetwork.com/journals/jama/fullarticle/2702134`

Is Poisson Model appropriate for?

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 numbers of Sudden Infant Deaths?

   https://www.ncbi.nlm.nih.gov/pubmed/21059188

3. Yearly numbers/incidence of hospitalized injuries in a region?

   http://www.epi.mcgill.ca/hanley/c609/Material/LidkopingALL.pdf

4. Yearly Accidents, Fatalities, and Rates, 1982 - 2000, U.S. Air Carriers

   http://www.epi.mcgill.ca/hanley/c626/airline-data-sas.txt

5. Quarterly & Monthly (prevalence) rates of Spina Bifida and Anencephaly Among Births (in relation to fortification of Foods with Folic Acid)

   http://www.epi.mcgill.ca//hanley/c626/folic_acid.pdf. See more data on webpage http://www.epi.mcgill.ca//hanley/c626/.

6. (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

7. Daily numbers of in-hospital deaths and Daily Maximal Temperatures

   http://www.epi.mcgill.ca/hanley/c626/Heatwave_death_lyon.pdf

8. 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.epi.mcgill.ca/hanley/c626/fullmoon.pdf

9. Daily no.s (Postponement of Death Until Symbolically Meaningful Occasions)

   http://www.epi.mcgill.ca/hanley/c626/holidays.pdf

10. Rates of audience fidget. (F Galton)

    http://www.epi.mcgill.ca/hanley/c626/measure_of_fidget_galton.pdf

References [as of 2007]

– Walker A Observation and Inference, p13,107,154

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

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

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

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

– Miettinen O. Theoretical Epidemiology p 295

– Breslow N, Day N Statistical Methods in Cancer ResearchVol II: Analysis of Cohort Studies pp68-70 (SMR) pp131-135; sect. 7.2 (power/sample size)

– Statistical Methods in Cancer ResearchVol I: Analysis of Case-Control Studies p134 (test-based CI's)

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

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

– Rothman K [1986] Modern Epidemiology

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