References:

*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

Outline [# *can omit on 1st reading*]

jh. course 626 Dec 2003

updated December, 2003: *Comments/corrections welcome*

The Poisson Distribution

How it arises

(1) When counting events (items) that occur randomly, and with low homogeneous intensity, in space or time

asbestos fibres	*deaths from horse kicks*
white cells	*typographical errors*
'"wrong numbers"	*cancers*
chocolate chips	*radioactive emissions*
nuclear medicine	*cell occupancy*

Distribution depends on single parameter μ = Expected (count)
(For "Expected ", can read "Average")

$$\text{Prob}(\text{count} = y) = \exp(-\mu)\,\frac{\mu^y}{y!} \quad \text{or} \quad e^{-\mu}\,\frac{\mu^y}{y!} \qquad (\ y = 0, 1, 2, \ldots\ )$$

y :	**0**	**1**	**2**	**3**	**..**
prob:	$e^{-\mu}$	$e^{-\mu}\ \mu$	$e^{-\mu}\ \mu^2/\ (1\times2)$	$e^{-\mu}\ \mu^3/\ (1\times2\times3)$	..

[*computing tip use the recurrence relation ...*

$e = \exp(1.0) = 2.718..$ $\quad \text{prob}(Y = y) = \text{prob}(Y = y\text{-}1) \times \frac{\mu}{y}$]

(2) As the limiting case of the Binomial (another "counting" distrn.)

For a Binomial with large number of "trials" (n) and small probability (π) of a "positive" on any one trial, the effective range of variation in the **count** Y of positive individuals is confined to small part of the lower end of the 0-n range. Used for "per mille" ($^{o}/_{oo}$) rather than percent (%) situations.

(3) As the sum of 2 or more independent Poisson random variables, with (same or) different expected values

Eg. Y_1 and Y_2 : counts of no. of 'events' from 2 indep. sources:

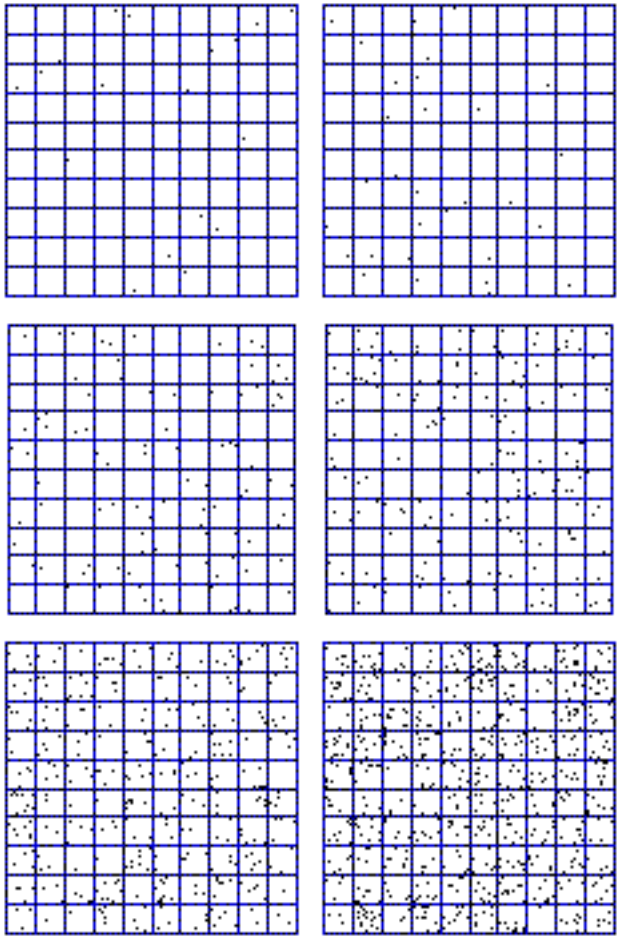
$Y_1 \quad \sim \quad \text{Poisson}[\ \mu_1\ ]$

$Y_2 \quad \sim \quad \text{Poisson}[\ \mu_2\ ]$

$Y_2 + Y_1 \ \sim \text{Poisson}[\ \mu_1 + \mu_2\ ]$ *see pages 8/9*

Random Distributions of Dots (events) in Space/Time

(different dot "intensity" in each panel)



(In epidemiology, each "cell" is a unit of person-time; each dot an event)

Excerpt from RA Fisher's
"Statistical Methods for Research Workers"

Whereas the normal curve has two unknown parameters, μ and σ, the Poisson series has only one. This value may be estimated from a series of observations, by taking their mean*, the mean being a statistic as appropriate to the Poisson series as it is to the normal curve. It may be shown theoretically that if the probability of an event is exceedingly small, but a sufficiently large number of independent cases** are taken to obtain a number of occurrences, then this number will be distributed in the Poisson series. For example, 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. The following data (Bortkewitch's data) were obtained from the records of ten army corps for twenty years, supplying 200 such observations, 1 per "corps-year". [See elsewhere on 626 website for even more detailed data]

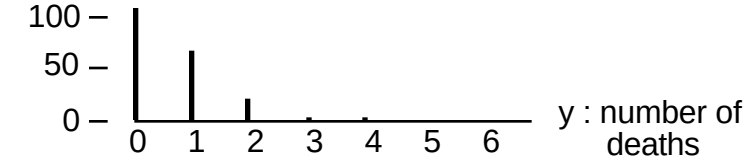
* [jh] We will see later (p 25) that -- if the Poisson moel is indeed appropriate) one can estimate the mean (and more inportantly, the variability of the estimate) from a ***single*** observation

** [jh] choice of word "case" is unfortunate here; he means ***a large number of independent "bits of experience" that make up a substantial denominator***

Bortkewitch's data, cited by Fisher

# Deaths in corps-year (y)	Frequency (# of "corps-years" with this many deaths) Observed	Expected †	y × freq.
0	109	108.67	0
1	65	66.29	65
2	22	20.22	44
3	3	4.11	9
4	1	0.63	4
5	-	0.08	0
6	-	0.01	0
(Total) Σ	# CorpsYears: 200	200	# deaths: 122

freq (no. of corps-years with y deaths)



†Expected number = prob(# Deaths=y)†† × 200

$$\hat{\mu} = \frac{\text{Total No. of deaths}}{\text{Total No. of corps-years}} = \frac{0\times109 + 1\times65 + ..}{200}$$

$$= \frac{122}{200} = 0.61 \ \frac{\text{deaths}}{\text{corps-year}}$$

var(y) = 0.61 = mean(y) in line with a Poisson distribution !!!

SD(y) = $\sqrt{0.61}$ = 0.78. Poisson distrn. ==> SD = $\sqrt{\text{mean}}$

(reverse not necessarily true!)

†† *p(# Deaths=y) = expected proportion of "corps-years" with y deaths*

* prob(# Deaths=0)	= exp{–0.61}	= 0.54
prob(# Deaths=1)	= p(0) x 0.61 / 1	= 0.33
prob(# Deaths=2)	= p(1) x 0.61 / 2	= 0.10
etc.		

Colton (p 78,79) defines the Poisson distribution as

> "that probability distribution on the integers 0 to ∞ with the probability of observing y events given by the formula
>
> $$prob(y) = \exp(-\mu)\ \mu^y / y! \qquad "$$

He gives one example

> y = bacteria counts in samples of one unit volume each from a thoroughly mixed large volume of bacterial suspension with an average of λ bacteria per unit volume [Colton's λ = our μ]

and one "non-example" (see discussion below)

> how accidents are distributed among (708) bus drivers; y_i = number of accidents for driver i ; $\Sigma\, y_i = 1623$

He doesn't derive the formula but says that "it can be deduced mathematically from the assumption that the bacteria are randomly distributed in the suspension". The derivation is not difficult and in fact it is instructive to derive the formula to show what the assumptions are and to help one recognize when the Poisson distribution might apply.

Derivation of formula for Poisson Probabilities [*]

Rather than tackle the derivation in the context of Colton's specific examples, it is easier to derive the Poisson distribution in the abstract as the limiting case of the Binomial distribution. and to then see how these and other examples fit the requirements.

---

[*] Armitage et al. [section 3.7, in 4th edition] give a similar derivation

The binomial distribution is characterized by a sample size n and the probability π that any one sampled member will be 'positive'. There are also the requirements that the n sample members be chosen randomly and independently of each other and that the probability π be the same from one "trial" to the next. Thus the average number of positives in n should be nπ ; we can call this quantity the 'expected' number and refer to it by the single parameter μ (note that for the binomial the definition of μ requires that one specify n and π first). Then it is easy to show that

$$\text{Binomial Prob(y "positives" | n, } \pi) = {}^nC_y\ \pi^y (1 - \pi)^{n-y}$$

Now, if n is large and π is small, then n is almost equal to n-1 or n-2 or n-3. Also, in this case, the binomial probability of observing **0** (zero) positives in n, i.e.

$$\text{Binomial prob}(\mathbf{0}) = \pi^0 (1 - \pi)^n$$

is very well approximated by exp(-nπ) or, in our notation, exp(-μ) ...

Indeed, a definition of exp[-x] is the limit, as n->infinity, of $(1 - x/n)^n$ ... try it out on your calculator or spreadsheet for say n=100 and π = 0.01, where you will see that the approximation is good to 2 decimal places.

Thus

$$\text{Binomial Prob}(\mathbf{0} \mid n = 100, \pi = 0.01)$$

$$= 0.01^0\ 0.99^{100}$$

$$= (1)(0.366) = 0.366.$$

$$\text{Poisson Prob}(\mathbf{0} \mid \mu)$$

$$= \exp(-1) = 0.368.$$

So we can write

$$\text{Binomial prob}(\mathbf{0}) \approx \exp(-\mu)$$

The binomial probability of observing **1** positive in n is

$$\text{Binomial prob}(1) = [\, n / 1! \,]\ \pi\ (1 - \pi)^{n-1}$$

and since we assume that n-1 is practically the same as n, so that $(1 - \pi)^{n-1}$ is approximately equal to $(1 - \pi)^{n}$, which is approximately equal to $\exp(-n\pi)$, the probability can be approximated by

$$\text{Binomial prob}(1) \approx [\, n\pi / 1! \,]\ \exp(-n\pi)$$

$$= [\, \mu^{1} / 1! \,]\ \exp(-\mu)$$

Likewise, the probability of **2** positives in n is

$$\text{Binomial prob}(\mathbf{2}) = [\, n(n-1) / 2! \,]\ \pi^{2}\ (1 - \pi)^{n-2}$$

and by the same arguments about n-1 and n-2 being approximately the same as n, the approximation is

$$\text{Binomial prob}(\mathbf{2}) \approx [\, (n\pi)^{2} / 2! \,]\ \exp(-n\pi)$$

$$= [\, \mu^{2} / 2! \,] \bullet \exp(-\mu)$$

By now the pattern is clear i.e. **when n is large and π small, the Binomial probabilities can be approximated by a simpler Poisson formula**

$$\text{prob}(y) = [\, \mu^{y} / y! \,]\ \exp(-\mu)$$

involving the single parameter $\mu = n\pi$. Also, because π is small, the binomial probabilities go to practically zero long before the largest possible outcome i.e. there is no need to calculate Binomial prob(n). This is what gives the Poisson the open-ended look, unlike the usual binomial which, with a smaller n and larger π, could stretch for quite a way over the 0/n to n/n range of possibilities. Note also that if we use the fact that the variance of a binomial count is $n\pi(1-\pi)$, and that $(1-\pi)$ is practically unity, we can infer that the variance of the Poisson is equal to $n\pi$, i.e. its variance is equal to its mean.

So far, you might say "**so what**!" since you already have a perfectly good formula for the binomial. One benefit is convenience: some calculators cannot handle the large ${}^{n}C_{y}$'s that appear in the binomial formula, whereas the only difficult part in the Poisson formula is obtaining $\exp(-\mu)$ needed in Poisson prob(0); the successive probabilities for prob (1), (2), (3)... can be obtained [manually* via a calculator, or by a spreadsheet] by multiplying each immediately preceding probability by factors of μ, $\mu/2$, $\mu/3$, etc. The fact that n is large doesn't come into the formulae explicitly, but only implicitly through the use of the "composite" parameter $\mu = n\pi$.

---

[*] The Poisson probabilities can also be obtained directly in Excel using the inbuilt `POISSON(count, μ, Cumulative?)` function.

- Use `POISSON(count, μ, FALSE)` to obtain the probability of obtaining the indicated count [probability mass function] .
- `Use POISSON(count, μ, TRUE)` to **sum** the probabilities for 0, 1, up to the indicated count. [cumulative probability function] See worked e.g.'s later.

But how does this derivation relate to Colton's examples, or to count data in epidemiology?

[Note that the fact that an observed distribution is NOT Poisson may be an important finding .. indeed one simple definition of epidemiology is "disease is not distributed at random"]

(i) Bacterial suspension [same e.g. as ABM p73] **:** Imagine that samples have an average of μ bacteria per unit volume, and that each unit volume is divided up into a very large number (n) of very small equal-sized sub volumes (each large enough for a single bacterium to fit in); then the small chance of finding a bacterium within any one selected sub volume is μ/n ; for good reasons, which we will see later, we will denote this small probability by π. If the suspension is well mixed, the chances of finding y bacteria in a randomly selected unit volume (or equivalently in the n sub volumes) should be given by the binomial probability prob(y) i.e.

$$\text{Binomial prob}(y) = {}^{n}C_{y} \; \pi^{y} \; (1 - \pi)^{n-y}$$

But as we have seen, for large n and small π, this probability can be approximated by

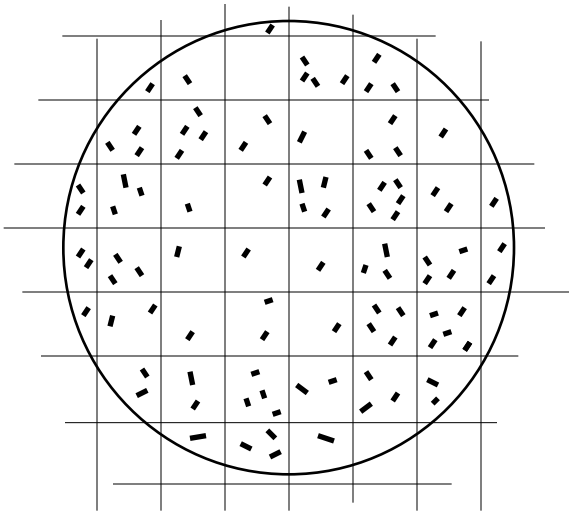
$$\text{Poisson prob}(y) = [\, (n\pi)^{y} / y! \,] \bullet \exp(-n\pi).$$

Since in our case $\pi = \mu / n$, then $n\pi = \mu$, and so

$$\text{Poisson prob}(y) = [\, \mu^{y} / y! \,] \; \exp(-\mu).$$

The same formulation can be used for example when counting asbestos fibres or white blood cells (randomly distributed in space) or emissions from a radioactive source (randomly distributed in time)

(from Feller p 163) Figure I reproduces a photograph of a Petri plate with bacterial colonies, which are visible under the microscope as dark spots. The plate is divided into small squares. Table 7 reproduces the observed numbers of squares with exactly k dark spots in five experiments with as many different kinds of bacteria.[I7] We have here a representative of an important practical application of the Poisson distribution to spatial distributions of random points.



y:	0	1	2	3	4	5	6
#cells	5	19	26	26	21	13	8
pred. #	*6.1*	*18.0*	*26.7*	*26.4*	*19.6*	*11.7*	*9.5*
o	26	40	38	17	7		
e	*27.5*	*42.2*	*32.5*	*16.7*	*9.1*		
o	59	86	49	30	20		
e	*55.6*	*82.2*	*60.8*	*30.0*	*15.4*		
o	83	134	135	101	40	16	7
e	*75.0*	*144.5*	*139.4*	*89.7*	*43.3*	*16.7*	*7.4*
o	8	16	18	15	9	7	
e	*6.8*	*16.2*	*19.2*	*15.1*	*9.0*	*6.7*	

The last entry in each row includes the figures for higher classes and should be labeled "y or more".

(ii) "Accidents": To test if accidents are truly distributed "randomly" over drivers, consider one specific (but generic) bus driver. If accidents are really that (i.e. if accidents shouldn't be more likely to happen to any driver rather than any other), then each time one occurs, there is a small $\pi = 1/708$ chance that it happens to the one driver we are considering. There are n=1623 such accidents to be 'distributed' so our driver has n "opportunities" to be dealt an accident [we ignore practical details such as whether the driver is still off work from the last one!].

We could therefore work out the binomial probability that the driver has 0, 1, 2, .. accidents. Now n is large enough and π small enough that using the Poisson formula with $\mu = n\pi = 1623 / 708$ will be quite accurate and save a lot of calculation. We then use the probability that any one driver will have y accidents as a synonym for the proportion of all drivers that will have y accidents

We can now compare the observed distribution of accidents and see how well it agrees with this theoretical Poisson distribution. Colton (pages 16-17) examines the fit of the Poisson model to the actual data...

Colton Table 2.5 Observed and "expected" numbers of accidents during a 3-year period among 708 Ulster (Northern Ireland Transport Authority) bus drivers.

# accidents in 3-year period	Number of drivers with this many accidents		
(y)	Observed	Expected †	O × y
0	117	71.5	0
1	157	164.0	157
2	158	187.9	316
3	115	143.6	345
4	78	82.3	312
5	44	37.7	220
6	21	14.4	126
7	7		49
8	6		48
9	1	(7-11 combined) 6.6	9
10	3		30
11	1		11
Σ	708	708	1623

$$\hat{\mu} = \frac{\#\text{accidents}}{\#\text{drivers}} = \frac{0\times117 + 1\times157 + ..}{708} = \frac{1623}{708} = 2.29.$$

var(y) = 3.45 >> mean(y).

†Expected number = Poisson Prob(# accidents=y $|\hat{\mu}$) × 708

(e.g. prob(0) = exp[-2.29]= 0.101 ; expected # of 0's = 708 × 0.101 = 71.5)

"Comparison of observed and expected tabulations reveals more than the expected number of drivers with no accidents and 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. From this example, what conclusions are justified concerning the random or nonrandom distribution of bus accidents?"

(iii) clustering in general: This second use of the Poisson probability distribution allows epidemiologists to calculate the probability of observing a certain size (or bigger) "cluster" of events if the only force operating were the play of chance. For example, the probability can be used to calculate the expected number, out of say 10000 communities of a certain size, that would, even with a homogeneous intensity of the event of interest, would have "clusters" of various sizes.

(iv) [more complex] Non-random choices of lottery combinations: The Lotto 6/49 draw does not produce a winning number in as many of the draws as one would predict. There are n = 14 million possible combinations. If N combinations are purchased, the average number of holders per combination is N/n [= μ say] . For many of the draws, N >> n, so that μ >> 1. If combinations were selected at random by their purchasers, then the chance that nobody holds the combination drawn is the Poisson probability P(0) calculated with parameter μ_Y. [Y = # of winning combinations purchased] If N = n, so that there are as many tickets sold as there are combinations, then μ_Y = 1.0 and P(0) = exp(-1) = 0.368 or 37%; if N = 4n or 55 million tickets sold, as it was in the mid 1980's for the draw with an accumulated prize of approx. $10million, then μ = 4 and P(0) = exp(-4) = 0.018 or only about 2% -- yet there was no winner! We can infer that either (i) the combinations were purchased at random and one of those "unexpected" 2% events did in fact occur or (ii) -- more plausibly -- that the combinations were not purchased at random -- they were not spread out at random over the 14 million possibilities. some numbers, and thus some ticket combinations, are oversubscribed and many are under subscribed).

(v) EVENT DATA IN EPIDEMIOLOGY: The distribution of bacteria in a large volume has a close analogy with the rate of disease events in a certain amount of person time.

One key assumption for the Poisson distribution to hold is that the events be "**well mixed**" and that there be no "lumpiness" or dependency between neighbouring counts. For example, if we were looking at variations in the numbers of lightning deaths (or worse still persons killed in airline crashes) from year to year, we would see "extra-Poisson variation. This can result from a number of factors, such as multiple fatalities to a related group or from a common source.

What, for example, if the person time is made up of contributions from different say age categories and what if the event intensities differ by age? Can we expect the TOTAL number of events in the amalgamated person time to be Poisson?

Provided again there is no other "lumpiness", the total number of events can still have a Poisson distribution -- , albeit governed by a more complex parameter. *Think of the contribution from each sub-category separately, so that the number of events y_j in person-time category j can be regarded as a realization of a Poisson variable with its own number n_j of person-time and its parameter μ_j .*

Then the sum (TOTAL) Σy_j of Poisson counts is a Poisson count with parameter $\Sigma \mu_j$.

An example of our (**triple**) use of this law is in dealing with the total numbers of cancers -- male and female -- in the Alberta study. We treat

$$O_{male} \sim \text{Poisson}(\mu_{male})$$
$$O_{female} \sim \text{Poisson}(\mu_{female})$$
$$O_{total} \sim \text{Poisson}(\mu_{male} + \mu_{female})$$

[we use the rule "The Sum of Poisson Random Variables is itself a Poisson Random Variable" **3** times -- first to sum *across ages within males*, second to sum *across ages within females* and third to sum *the overall count for males* and *the overall count for females*.

Of course, if there is wide variation in the event rates across categories, it may not make a lot of sense to speak of a single estimated rate of events (i.e. estimated as $\Sigma y_j / \Sigma n_j$) without specifying the category composition of the Σn_j . An overall count or rate may only make sense, if it is clear what mix of person-time "strata" it applies to.

The overall count is often compared with the overall number one would expect in a similar person-time structure which experience the event rates of a comparison population.

Examples of "lumpy" counts-- that show "extra-Poisson" variation

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 variations in numbers of births
{see e.g. *Number of weekday and weekend births in New York in August 1966* on 626 web page}
[closer to Poisson if use weekly count]

Daily variations in numbers of deaths
[variation over the seasons]

Daily variations in numbers of traffic accidents
[variation over the seasons, and days of week, and with weather etc.]

Daily variations in numbers of deaths in France in summers of 2002 and 2003

Impact sanitaire de la vague de chaleur en France survenue en août 2003. Rapport d'étape 29 août 2003 [on course 626 webpage]

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]

A summary of Features of Poisson Distribution

- Like Binomial, describes variation in counts
- Open-ended (unlike Binomial) : 0, 1, 2, 3, ...
- Never goes to full range (because π so small)
- If one thinks of it as limiting case of Binomial(n; π), then n and π appear only through their product $\mu = n\pi$

  (so same distribution for n=1000, π=0.0021 as for n=10000, π=0.00021; mean = $n \bullet \pi = 2.1$ in both)
- All the details of distribution (e.g. variance, 95% limits of variation, ...) are derived from its 1 parameter : μ
- "Denominator " can be person-years or other measure of "amount of experience"
- Poisson data often referred to as "numerator only" data, in sense that (unlike Binomial) one does not "see" or count "non-events"; instead, the denominator is a measure of amount of experience (what IS the "denominator "behind" the number of incoming "wrong numbers" on the phone? see e.g. on www page)
- To make inferences about ratios of event-rates, need only know the relative sizes of the "denominators"
- If counts follow a Poisson distribution with mean μ,

  Variance(counts) = μ

  SD(counts) = $\sqrt{\text{average count}} = \sqrt{\mu}$

  [via Binomial: If $\mu = n\pi$ and π is small, then $1-\pi$ is close to 1, and so Variance(count) = $n\ \pi(1-\pi) \approx n\pi = \mu$]
- CI / Test for μ : use Poisson tails if distrn. is not approximately Gaussian at the limit; Gaussian approxn. otherwise (see later)

Examples (some may not be Poisson; , counts may show "extra-" or "larger-than- " Poisson variation):

(1) "Cluster" (6 twin pairs in a school of 375 children)

Is it unusual to have 6 sets of twins in a school of 375 students? If expect 1 set of twins per 270 births, then μ = 375/270 = average of 1.3 twins pairs per school of size 375.

See Poisson Probability table for μ =1.3 [Table on p 16 doesn't have μ = 1.3, but can look at μ =1.0 or μ =1.5 and interpolate]. A count of 6 (or more) not that common, but if screen enough schools of that same size, will find a few schools per 1000 that will have 6 or more, even when the mean is 1.3. [See Hanley JA "Jumping to coincidences: defying odds in the realm of the preposterous". American Statistician, 1992, 46(3) 197-202.]

Can use the "Poisson counts as Cell-Occupancy counts" using 1300 visits(twin pairs) to 1000 cells (schools) since μ = 1.3 -- these were generated by the Excel macro. [see page 14]

(2) Radioactive disintegrations. (more details on separate file -- material from Feller -- on www page) A radioactive substance emits alpha-particles; the number of particles reaching a given portion of space during time *t is* the best-known example of random events obeying

the Poisson law. Of course, the substance continues to decay, and in the long run the density of alpha-particles will decline. However, with radium it takes years before a decrease of matter can be detected; for relatively short periods the conditions may be considered constant, and we have an ideal realization of the hypotheses which led to the Poisson distribution.

In a famous experiment (Rutherford, Chadwick, and Ellis *Radiations from radioactive substances,* Cambridge 1920, p.172.) a radioactive substance was observed during different time intervals of 7.5 seconds each; the number of particles reaching a counter was obtained for each period. Table 3 [Feller: see web page] records the number of periods with exactly *y* particles. The average number of particles per period is 3.87. The theoretical values for the number of periods with exactly *y* particles are seen to be rather close to the observed numbers.

(**3) Flying-bomb hits on London.** As an example of a spatial distribution of random points consider the statistics of flying-bomb hits in the south of London during World War II. The entire area is divided into 576 small areas of *1/4* square kilometers each, and table 4 records the number of areas with exactly *y* hits. The total number of hits is 537, an average 0.93 per small area. The fit of the Poisson distribution is surprisingly good; as judged by the X^2-criterion, under ideal conditions some 88 per cent of comparable observations should show a worse agreement. It is interesting to note that most people believed in a tendency of the points of impact to cluster. If this were true, there would be a higher frequency of areas with either many hits or no hit and a deficiency in the intermediate classes. Table 4 [again, see Feller] indicates perfect randomness and homogeneity of the area; we have here an instructive illustration of the established fact that to the untrained eye randomness appears as regularity or tendency to cluster.

(4) Connections to wrong telephone number. Table 6 (see Feller on www page) shows statistics -- from a 1926 publication! -- of telephone connections to a wrong number. A total of 267 numbers was observed; N_k indicates how many numbers had exactly *k* wrong connections. The Poisson distribution *Poisson(*$\mu = 8.74$) shows again an excellent fit. (As judged by the X^2-criterion the deviations are near the median value.). Sometimes (as with party lines, calls from groups of coin boxes, etc.) there is an obvious interdependence among the events, and the Poisson distribution no longer fits.

(5) Chromosome interchanges in cells. Irradiation by X-rays produces certain processes in organic cells which we call chromosome interchanges. As long as radiation continues, the probability of such interchanges remains constant, and, according to theory, the numbers *Nk* of cells with exactly k interchanges should follow a Poisson distribution. The theory is also able to predict the dependence of the parameter A on the intensity of radiation, the temperature, etc., but we shall not enter into these details. Table 5 [Feller] records the result of eleven different series of experiments.l5 These are arranged according to goodness of fit. The last column indicates the approximate percentage of ideal cases in which chance fluctuations would produce a worse agreement (as judged by the x2- standard). The agreement between theory and observation is striking.

(6) cancers in area in Alberta exposed to sour gas: based on provincial rates, areas with the same population had an average of 85.9 cancers in the same time period. How much variation should there be from area to area?

See separate "alberta.pdf" file on 626 www page

(7) An **estimate of WBC concentration** can be made by manually **counting enough fields (n) until say y=200 have been observed.** This is <u>not quite a Poisson distribution</u> 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 a margin of error (ME) of up to <u>13% high</u> to <u>15% low</u> -- since a single count of 200 (marked by * below) could be a low reading from a concentration which should produce an average of 230 for the same n, or a high reading from a concentration which should produce an average of 170 in the same n, i.e.

it could be that y=200(*) is an <u>over</u>estimate

```
μ per n  160 170 180 190 200 210 220 230 240
          |   |   |   |   |   |   |   |   |
     ----------μ----------*
```

200 ≈ **173** (Lower limit) + 2√**173**

(ignoring asymmetry; & using z = +/- 2 SD for 95%)

or it could be that y is an <u>under</u>estimate

```
μ per n  160 170 180 190 200 210 220 230 240
          |   |   |   |   |   |   |   |   |
                              *----------μ----------
```

200 ≈ **230** (Upper limit) - 2√**230**

General Note on Correct Method of Constructing a CI
Note: A number of students, used to thinking of a CI as
(point estimate - ME, point estimate + ME),
asked me if I made a mistake in the previous example. No!. It is only when the sampling distribution is symmetric and it's shape does not depend on the level that the CI is symmetric; otherwise, one starts at say the lower limit and works out, by trial and error if necessary, where to put this limit so that the observed y is at the (1-a/2)%ile with respect to that limit. When one moves to establish the upper limit, so that the observed y is now at the a/2(100)%ile with respect to this upper limit, the variation in y applies at this upper limit may be different than it is at the lower limit. See "Method of Constructing CI's (in general) in my notes on Chapter 6 in course 607.

Examples that may not fit the Poisson distribution (or, at least not without some further aggregation)

(8) **Daily numbers of births in New York city 9 months after the blackout** of 1965.

see separate "blackout.pdf" file on 626 www page

More than Poisson variation in daily numbers of birth (check the weekends!!)
Might be closer to Poisson if combine births for 7 days;
Over the year, one would still expect some non-homogeneity in the number per week

(9) **deaths before & after religious & ethnic holidays.**

see separate "holidays.pdf" file on 626 www page

More than Poisson variation in weekly numbers of deaths (seasonal cycles!!)

(10) **deaths in U.S.A. after Chernobyl accident.**

see separate "accident.pdf" file on 626 www page

(11) **Chocolate chips in cookies, olives on pizzas**

A useful exercise to illustrate how to think about whether a Poisson distribution might apply is to imagine that the task was to estimate, on average, **how many μ a manufacturer puts in a chocolate chip cookie**, or on average, **how many olives μ a pizza maker puts on a pizza**, by observing a **single** randomly selected cookie or pizza.

A count obtained in one randomly sampled unit is a valid but uncertain estimate of μ Can one use the Poisson distribution to estimate its uncertainty?

In the chocolate chips case, if the cookie dough was mixed and divided by a machine, we should be able to think of the count as a Poisson variable with expected value μ . If μ is large, we can use $\hat{\mu}$. with $\sqrt{\mu}$ -based margins of error to form a "large μ" confidence interval for μ . If μ is small, we can use Tabulated CI's (see below)

In the olives example, it is not evident how much variability to expect. If one knew that a pizza maker is obliged by his company to always use the same number of olives per pizza, then a single pizza provides an error-free estimate. This would result in considerable "less-than-Poisson" variation from slice to slice or pizza to pizza.

Poisson counts as Cell-Occupancy counts (see Excel macro to the right of "W 8 Discrete Random Variables" on 323 webpage)

Poisson Distributions: μ = 1.0, 1.3 & 2.0

μ = 1.0 100 Random Visits to 100 Cells

For each visit, the target cell was chosen randomly from the numbers 1 to 100. The entries in the 100 cells are the number of times these cells were visited. For example, cell # 1 was visited 1 time, cell # 2 was visited 0 times, etc.

```
 1  -  -  2  -  -  1  3  -  5  1  2  -  2  2  1  1  -  -  2
 -  3  -  -  3  -  -  -  -  2  1  1  -  1  -  1  4  -  1  1
 -  1  1  3  1  -  2  -  3  2  -  4  2  -  2  -  2  2  -  1
 -  1  -  1  1  1  -  -  -  -  2  -  1  3  3  -  1  2  2  -
 -  1  -  -  -  1  2  1  1  1  -  -  2  -  1  1  -  2  1  1
 100 visits   to   100 cells      mean μ =  1.000/cell
```

FREQUENCIES f(y) = number of cells that were visited y times

y :	0	1	2	3	4	5	6	
observed f(y):	*42*	*30*	*18*	*7*	*2*	*1*	*0*	*sd(y)= 1.1*
expected f(y):	*37*	*37*	*18*	*6*	*2*	*0*	*0*	

μ = 1.3 130 Random Visits to 100 Cells

```
 -  2  1  1  1  -  3  2  1  1  2  -  2  1  1  1  1  1  -  1
 -  1  4  -  2  -  -  -  1  -  3  4  2  -  1  -  -  3  1  3
 2  -  2  -  2  1  1  -  2  3  2  1  1  2  1  -  -  1  2  1
 -  -  1  2  2  1  2  1  3  1  1  3  3  -  1  3  3  2  2  3
 2  1  4  -  -  -  1  1  3  -  2  1  1  -  1  1  2  1  4  -
 130 visits   to   100 cells      mean μ = 1.300/cell
```

y :	0	1	2	3	4	5	6	7	
observed f(y):	*27*	*36*	*21*	*12*	*4*	*0*	*0*	*0*	*sd(y)= 1.1*
expected f(y):	*27*	*35*	*23*	*10*	*3*	*1*	*0*	*0*	

μ = 2.0 200 Random Visits to 100 Cells

```
 1  2  3  3  2  1  1  3  3  -  4  1  3  3  2  1  5  3  3  -
 3  2  2  4  1  6  3  -  1  2  2  1  2  1  2  -  1  1  6  -
 1  3  4  2  2  -  5  2  1  2  2  2  2  4  1  3  3  1  -  3
 1  -  2  1  1  2  -  2  -  -  -  4  2  2  4  4  5  2  2  2
 3  -  -  2  3  1  2  3  3  5  -  -  1  1  3  2  2  2  3  1
 200 visits   to   100 cells      mean μ = 2.000/cell
```

y :	0	1	2	3	4	5	6	7	8	9	
observed f(y):	*16*	*22*	*29*	*20*	*7*	*4*	*2*	*0*	*0*	*0*	*sd(y)= 1.4*
expected f(y):	*14*	*27*	*27*	*18*	*9*	*4*	*1*	*0*	*0*	*0*	

Larger N, smaller π Does it make a difference?

1000 Random Visits to 1000 Cells

```
 1  -  1  -  1  -  1  1  1  -  -  -  1  -  2  4   ....
 -  -  3  -  -  -  1  1  -  -  -  1  1  -  2  2   ....
 1  -  3  1  3  1  1  1  -  1  -  -  -  1  1  -   ....
 1  -  -  -  1  -  1  1  -  1  -  -  1  2  2  2   ....
 -  -  3  1  2  1  -  -  -  1  1  -  3  1  -  -   ....
 .  .  .  .  .  .  .(remainder not shown)  .  .  .   ....
```

y :	0	1	2	3	4	5	6	7	
observed f(y):	*370*	*367*	*178*	*66*	*16*	*3*	*0*	*0*	*sd 1.0*
expected f(y):	*368*	*368*	*184*	*61*	*15*	*3*	*1*	*0*	*μ =1.0*

1300 Random Visits to 1000 Cells
EG 1300 TWIN PAIRS in 1000 SCHOOLS

```
 -  2  -  1  1  1  -  2  -  1  2  2  2  1  2  4   ....
 1  3  -  -  2  1  -  -  -  3  2  -  3  -  -  1   ....
 -  2  1  1  -  -  -  3  1  -  -  1  -  -  -  -   ....
 -  1  -  1  2  3  4  2  1  3  1  2  2  -  -  1   ....
 .  .  .  .  .  .  . (remainder not shown) .  .  .   ....
```

y :	0	1	2	3	4	5	6	7	8	
observed f(y):	*279*	*357*	*217*	*97*	*38*	*7*	*4*	*1*	*0*	*sd= 1.2*
expected f(y):	*273*	*354*	*230*	*100*	*32*	*8*	*2*	*0*	*0*	*μ = 1.3*

2000 Random Visits to 1000 Cells

```
 1  1  1  1  2  1  1  2  1  1  2  2  3  1  4  8   ....
 1  3  3  -  2  1  1  1  -  3  2  1  3  -  2  3   ....
 1  2  2  1  -  1  1  1  2  1  -  2  1  -  -  -   ....
 -  2  3  2  3  1  2  2  -  3  2  -  3  1  2  3   ....
 .  .  .  .  .  .  . (remainder not shown) .  .  .   ....
```

y :	0	1	2	3	4	5	6	7	8	9	10	
obs f(y):	*128*	*287*	*258*	*191*	*81*	*39*	*9*	*5*	*2*	*0*	*0*	*sd=1.4*
exp f(y):	*135*	*271*	*271*	*180*	*90*	*36*	*12*	*3*	*1*	*0*	*0*	*μ =2.0*

Distributions are characterized by μ , not by whether μ is a produced by N × π or 10N × (π/10).

(This material is more complex and can be skipped on 1st reading)

Want a tip on winning the 6/49 lottery jackpot? It's pure luck.

excerpt from article by JENNIFER JACKSON, OTTAWA CITIZEN
in Montreal Gazette March 1992

OTTAWA--Wouldn't you like to know which numbers win most often in Lotto 6/49?

OK, here they are: **31**, **43**, **44**, **47**, **7**, **49**, **27**.

From the time the lottery started June 12, 1982, to Dec. 28, 1991, No. 31 has come up **141** times, No. 43 was drawn 139 times, Nos. 44 and 47 both 134 times, No. 7 popped up 132 times, No. 49 rolled down the chute 131 times and No. 27 followed at 128 times.

Conversely, the numbers to stay away from over those years were No. **12** (pulled only **96** times), **15** (99 times), 6 and 48 (103 times each), No. **13** (104 times); No. **14** (105 times), **2** (106 times) and **24** (109 times).

Too bad that these figures are absolutely no indication of which numbers will prevail in the next draw-- never mind over the next 10 years. The only way to guarantee a win is to buy one of each possible combination, and in the 6/49 at least, that's a practical impossibility.

JH: What is the point of this Lottery example? It illustrates a common tendency to pick out extremes and ignore the fact that at least a part of the extremeness of these particular ones must be random variation ...

The Extremal Quotient (EQ)

In studies of geographical variation in rates, it is common to summarize the variation as the "Extremal Quotient" i.e.

$$\frac{\text{highest rate}}{\text{lowest rate}} .$$

One needs to be able to judge whether the observed variation is more than one would expect if all that was operating was chance, i.e. if the rates exhibit "extra-Poisson or extra-Binomial" variation [remember that the Binomial distribution assumes equal size denominators (n's) and equal probabilities of a "positive"].

The article on the lottery is a very good example of the Extremal Quotient ($\frac{\textbf{141}\text{ times}}{96\text{ times}}$.) and how, if the n's or the numerators are small, the variation can be substantial even when there is no underlying variation in the rates or probabilities. We are not given the total number of winning numbers drawn but one can deduce that they must be of the order of 5733 (=117 x 49: the average number of times numbers have been drawn seems to be about 117). Imagine recording them all on the 49-cell grid that players use to record their choices. **How would one expect these 5733 winning numbers to distribute over the 49 cells?**

They are drawn in blocks of 6 without replacement, but we will take some small poetic licence for now and assume that all 5733 are drawn with replacement -- this way we can use the Poisson distribution as an approximation to the cell occupancy distribution. (and even use the Excel macro)

We are told that #31 had 141 "hits", and that #12 had only 96, so the extremal quotient is

$$\frac{141}{96} = 1.46$$

This variation could be predicted from the Poisson distribution (close enough here) as follows:

Average number of 'hits' per #: = 117

SD[Poisson count] ~ √mean = √117 = 11 (approx.)

117 ± 2SD = 95 to 139

Observed range = 96 to 141 **(!!!)**

Variation in surgery rates: This can help us to quickly measure if variation in surgery rates is "more than chance alone". For example, the Conseil d'évaluation des technologies de la santé du Québec recently analysed the variation in surgical rates across the 32 Québec DSC's (the average population size per DSC is thus approximately 250,000).

For one particular procedure, there were more than 5600 operations in the time period studied, so the average μ per DSC [assuming all DSC's of equal size to make life simple] is approximately 5600/32=175. Thus, by chance alone, we would expect an EQ of approximately

$$\frac{175 + 2\sqrt{175}}{175 - 2\sqrt{175}} = 1.3$$

(it would in fact be somewhat more since the DSC's are not all of equal size -- smaller DSC's would have greater variability in their rates)

For procedures with 12,800 operations, that's 400 per (average sized) DSC on average. Going through the same calculation as above gives a range of 360<-->440 or an extremal quotient of 1.22.

One could use the Excel spread sheet with 5733 visits to 49 cells to look at the extremal quotient in the 6/49 case

[Using sampling with replacement, rather than sampling without replacement for each block of 6, in the Excel macro (and the Poisson model) slightly overestimates the variation:- the hyper-geometric distribution has less variation than the corresponding Binomial]

For further references on this topic, and an application to rates of admission of pediatric patients to Québec hospitals, see PhD thesis by M. Hodge.

(Poisson) probabilities of observing y events if "expected" or average number is μ

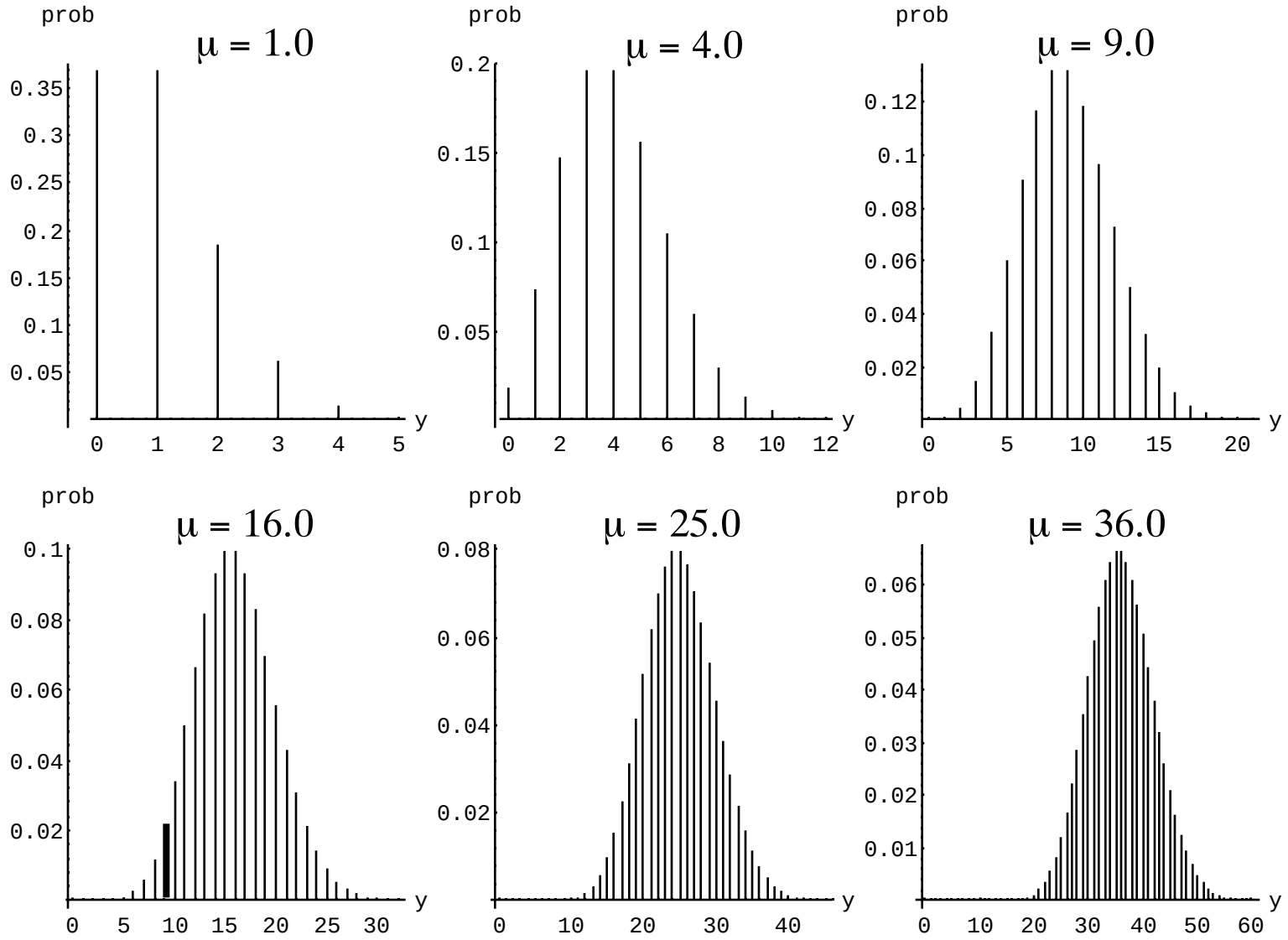
Probabilities are expressed "per 1000" i.e. 905 is 905/1000 or 0.905 or 90.5% The formula is Prob(y) = [exp(-μ)] • [μ to power y] / y!

E.g.: if μ = 1.5, then Prob(y=3) is exp(-1.5)•1.5 cubed / [1x2x3] = 126/1000 or .126 or 12.6% ... Reminder: 3! = 1x2x3=6]

μ =	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.0	1.5	2	3	4	5	7	10	16	20	25	30	36
y																						
0	905	819	741	670	607	549	497	449	407	368	223	135	50	18	7	1	0	0	0	0	0	0
1	90	164	222	268	303	329	348	359	366	368	335	271	149	73	34	6	0	0	0	0	0	0
2	5	16	33	54	76	99	122	144	165	184	251	271	224	147	84	22	2	0	0	0	0	0
3	0	1	3	7	13	20	28	38	49	61	**126**	180	224	195	140	52	8	0	0	0	0	0
4	0	0	0	1	2	3	5	8	11	15	47	90	168	195	175	91	19	0	0	0	0	0
5	0	0	0	0	0	0	1	1	2	3	14	36	101	156	175	128	38	1	0	0	0	0
6	0	0	0	0	0	0	0	0	0	1	4	12	50	104	146	149	63	3	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	1	3	22	60	104	149	90	6	1	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	1	8	30	65	130	113	12	1	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	3	13	36	101	125	21	3	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	1	5	18	71	125	34	6	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	2	8	45	114	50	11	1	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	26	95	66	18	2	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	14	73	81	27	3	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	52	93	39	6	1	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	35	99	52	10	1	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	22	99	65	15	2	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	93	76	23	3	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	83	84	32	6	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	70	89	42	9	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	56	89	52	13	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	43	85	62	19	2
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	31	77	70	26	4
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	22	67	76	34	6
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	56	80	43	8
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	45	80	51	12
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	34	76	59	17
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	25	71	66	22
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	18	63	70	29
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	13	54	73	36
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	45	73	43
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	37	70	50
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	29	66	56
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	22	60	61
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	16	53	64
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	45	66
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	38	66
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	31	65
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	24	61
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	19	56

(Poisson) probabilities of observing y events if "expected" or average number is μ= 1, 4, 9, 16, 25, 36

Using μ *and mean interchangeably.* μ *is not usually an integer: it can be any non-negative real number , such as 1.3)*



E.G. Prob(Y=9 | Mean = 16.0) = 0.021

<u>Message</u>: Gaussian approxn. to Poisson distrn. reasonably accurate when μ **is in the double digits'** [cf also Armitage & Berry]

[See frequency distributions on previous page]

The normal distribution is often useful as an approximation to the Poisson distribution. The Poisson distribution with mean μ approaches normality as μ increases indefinitely (see diagram of Poisson distributions as a function of μ). For sufficiently large μ, A Poisson variable y may, therefore, be regarded as "approximately normal" with mean μ and standard deviation $\sqrt{\mu}$

If tables of the normal distribution are to be used to provide approximations to the Poisson distribution, account must be taken of the fact that this distribution is discrete whereas the normal distribution is continuous. It is useful to introduce what is known as a *continuity correction,* whereby the exact probability for, say, the Poisson variable y (taking integral values) is approximated by the probability of a normal variable between y - 0.5 and y + 0.5 *[jh: round the continuous values between y-0.5 and y+0.5 to the nearest integer -- imagine the "spikes" in the distribution on the previous page converted to rectangles with no gaps]*. Thus, the probability that a Poisson variable took values greater than or equal to y when y > μ (or less than or equal to y when y < μ) would be approximated by the normal tail area beyond a standardized normal deviate $z = \frac{|\, y - \mu \,| - 0.5}{\sqrt{\mu}}$ (1)

Table 2.7:

Examples of approximation with continuity correction

Mean (μ)	SD ($\mu^{1/2}$)	Values of y	Exact Prob.	Approx Prob*	z
5	2.236	0	*0.0067*	*0.0221*	2.013
		≤2	*0.1246*	*0.1318*	1.118
		≥8	*0.1334*	*0.1318*	1.118
		≥10	*0.0318*	*0.0221*	2.013
20	4.472	≤10	*0.0108*	*0.0168*	2.124
		≤15	*0.1565*	*0.1572*	1.006
		≥25	*0.1568*	*0.1572*	1.006
		≥30	*0.0218*	*0.0168*	2.124
100	10.000	≤80	*0.0226*	*0.0256*	1.950
		≤90	*0.1714*	*0.1711*	0.950
		≥110	*0.1706*	*0.1711*	0.950
		≥120	*0.0282*	*0.0256*	1.950

* Normal approximation with continuity correction, with z as in (1)

(1- 2α) Confidence limits for the expectation [i.e. the 'mean' parameter] of a Poisson random variable

E.g. if observe 6 events in a certain amount of experience, then 95% CI for the μ count for this same amount of experience is (2.20, 13.06)

1−2α	*0.998*		*0.99*		*0.98*	
α	*0.001*		*0.005*		*0.01*	
count	Lower	Upper	Lower	Upper	Lower	Upper
0	0.00	6.91	0.00	5.30	0.00	4.61
1	0.00	9.23	0.01	7.43	0.01	6.64
2	0.05	11.23	0.10	9.27	0.15	8.41
3	0.19	13.06	0.34	10.98	0.44	10.05
4	0.43	14.79	0.67	12.59	0.82	11.60
5	0.74	16.45	1.08	14.15	1.28	13.11
6	1.11	18.06	1.54	15.66	1.79	14.57
7	1.52	19.63	2.04	17.13	2.33	16.00
8	1.97	21.16	2.57	18.58	2.91	17.40
9	2.45	22.66	3.13	20.00	3.51	18.78
10	2.96	24.13	3.72	21.40	4.13	20.14
11	3.49	25.59	4.32	22.78	4.77	21.49
12	4.04	27.03	4.94	24.14	5.43	22.82
13	4.61	28.45	5.58	25.50	6.10	24.14
14	5.20	29.85	6.23	26.84	6.78	25.45
15	5.79	31.24	6.89	28.16	7.48	26.74
16	6.41	32.62	7.57	29.48	8.18	28.03
17	7.03	33.99	8.25	30.79	8.89	29.31
18	7.66	35.35	8.94	32.09	9.62	30.58
19	8.31	36.70	9.64	33.38	10.35	31.85
20	8.96	38.04	10.35	34.67	11.08	33.10
21	9.62	39.37	11.07	35.95	11.83	34.35
22	10.29	40.70	11.79	37.22	12.57	35.60
23	10.96	42.02	12.52	38.48	13.33	36.84
24	11.65	43.33	13.26	39.74	14.09	38.08
25	12.34	44.64	14.00	41.00	14.85	39.31
26	13.03	45.94	14.74	42.25	15.62	40.53
27	13.73	47.23	15.49	43.50	16.40	41.76
28	14.44	48.52	16.25	44.74	17.17	42.98
29	15.15	49.80	17.00	45.98	17.96	44.19
30	15.87	51.08	17.77	47.21	18.74	45.40

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

• Computed from (exact) Poisson tail areas i.e. Prob(COUNT >= count | μ_{Lower}) = Prob(<= count | μ_{Upper}) = α. See also the spreadsheet "Exact confidence limits on a Poisson parameter" on 626 website • Limits in above Table computed using exact relationship b/w Poisson and Chi-square tail areas (later).

Basis for "First Principles" Poisson Confidence Interval : not your usual " point estimate +/- some multiple of the standarderror"

To form a (1-2α) CI for μ, based on # events c, need to find:

μ_{LOWER} such that Prob (c or more | μ_{LOWER}) = α

μ_{UPPER} such that Prob (c or fewer | μ_{UPPER}) = α

Example: 95% CI based on **c** = 6.

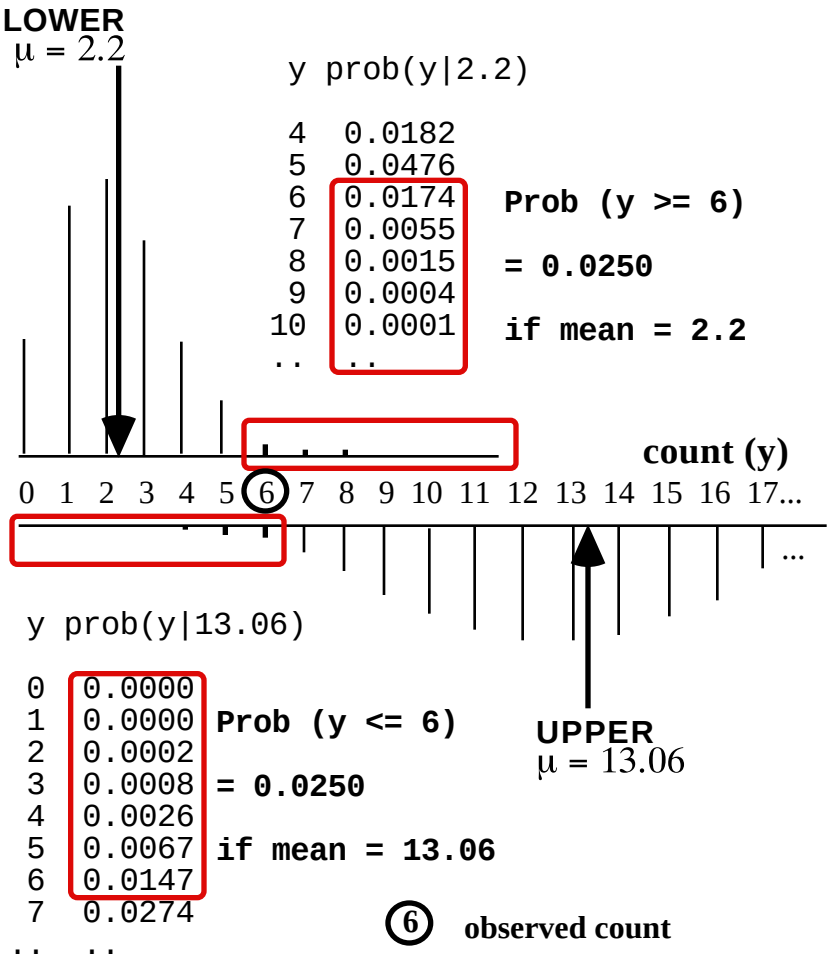
Need to find the μ_{LOWER} that makes the probability of 6 or more equal to α = 0.025.

Need to find the μ_{UPPER} that makes the probability of 6 or fewer equal to α = 0.025.

Finding lower and upper limits involves "trial and error" to find the appropriate μ_{LOWER} and μ_{UPPER} that yield the target α 's.

See below for a way to get there directly using Link between the tail areas of the Poisson and tail areas of Chi-Square distributions. Note that the above "First Principle" is a general and important one; it "just so happens" that in this particular discrete distribution, if one has access to the percentiles of the Chi-Square distribution, the link helps avoid the "trial and error" process involved in the actual calculation.

Note that unless c is large, the Poisson distribution corresponding to the lower limit will be skewed and not always amenable to a Gaussian approximation; the Poisson distribution at upper limit is less troublesome.



"Exact" CI for mean, μ , of a Poisson distribution using Link between Poisson and Chi-Square tail areas.

This is a surprising link, considering that the Poisson is a distribution on the discrete integers 0, 1, 2, ... and the Chi-square distribution is on the non-negative real numbers.

It has been known since early in 20th century, but has remained largely hidden. (the proof requires integration by parts!)

It used to be an important exact alternative to summing tail areas until the Poisson (and several other) tail areas became available from built-in functions in the statistical and spreadsheet packages. Even now, with a spreadsheet formula for Poisson tail areas, one has to find the limits by trial and error.

To form a (1-2α) CI for μ, based on a count c, we need to find

μ_{LOWER} such that Prob (c or more | μ_{LOWER}) = α

μ_{UPPER} such that Prob (c or fewer | μ_{UPPER}) = α

Example: Based on **c** = 6, find 95% CI for μ

Need to find the μ_{LOWER} that makes the probability of 6 or more equal to α = 0.025. and the μ_{LOWER} that makes the probability of 6 or fewer equal to α = 0.975.

(answer: μ_{LOWER} =2.20 and μ_{UPPER} =13.06)

The exact limits on μ**,** for any specified count **c** and confidence coefficient (1-2α), can -- without trial and error -- be found directly from the χ^2 distribution.

$\mu_{LOWER} = (1/2)\ \chi^2_{\alpha,\ df = 2\mathbf{c}}$.

$\mu_{UPPER} = (1/2)\ \chi^2_{1-\alpha,\ df = 2(\mathbf{c+1})}$.

Values of χ^2 for any α and df are readily available from many statistical packages or spreadsheets, or can be found from an adequately extensive tabulation of the χ^2 distribution.

In our example... **c** = 6 ; (1-2α) = 0.95,

so α = 0.025, 1- α = 0.975.

$\mu_{LOWER} = (1/2)\ \chi^2_{0.025,\ 12df} = (1/2)4.40 = 2.20$,

$\mu_{UPPER} = (1/2)\ \chi^2_{0.975,\ 14df} = (1/2)26.12 = 13.06$.

If you use Excel, reverse α and, 1- α .

Clever way to obtain exact limits, using Stata

use 2-sample comparison, with 'infinite' comparison group:-

epitab syntax is `iri #a #b #N1 #N2 [, level(#)]`

so set #b and N2 to be very large (a (our c)=6 events in 1 person-year, versus b=1000000 events in 1000000 person years:

```
      c very_large#  PT very_large_PT
iri 6    1000000    1    1000000
```

Approximate CI's for mean, μ , of a Poisson distribution, based on 5 different approximations to Poisson tail areas

(1) Wilson/Hilferty approxn. to Chi-square quantiles.

[helpful when appropriate Chi-square quantiles not readily available]

This approximation , which has high accuracy for **c** > 10, uses z , the normal standardized variate corresponding to α , e.g.. z = 1.645 for α = 0.05,1.96 for α = 0.025, etc.

$$\mu_{LOWER} = (\mathbf{c})\ \{\ 1 - (9\mathbf{c})^{-1} - z\ (9\mathbf{c})^{-1/2}\ \}^3$$

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

Note1: Rothman[2002], page 134, provides an adaptation from "D. Byar, unpublished" in which he makes a further approximation, using the average (c+0.5) for both the lower an upper limits, rather than the more accurate c for the lower and c+1 for the upper limit. This is called method 1' below. JH is surprised at Rothman's eagerness to save a few keystrokes on his calculator, and at his reference to an unpublished source, rather than the 1931 publication of Wilson & Hilferty. Full W-H citation, and evaluation of the above equation, in Liddell's "Simple exact analysis of the standardized mortality ratio" in J Epi and Comm. Health 37 85-88, 1984 available on 626 website.

Note2: Rothman uses the CI for the expected numerator of a Rate. {e.g.s below focus on number in same sized study, not rate *per se*.

(2) Square-root transformation of Poisson variable.

With μ large enough, √**c** is approximately Gaussian with mean √μ and variance 1/4 or SD 1/2 (the variance and SD are thus independent of √μ) .

This leads to (see ref. (3)):

$$\mu_{LOWER,UPPER} = \mathbf{c}\ \mathbf{-/+}\ z\ (\mathbf{c})^{1/2} + {}^1/_4(z)^2$$

This simpler formula is accurate when c > 100 or so.

(3) 1st Principles CI from c ~ Gaussian(μ , SD = √μ)

Obtained by solving the two equations:

$$c = \mu_{LOWER} + z\sqrt{\mu_{LOWER}}\ ;\ c = \mu_{UPPER} - z\sqrt{\mu_{UPPER}}$$

to give

$$\mu_{LOWER,UPPER} = (\sqrt{c + z^2/4}\ -/+\ z/2\ )^2$$

"First Principles" : it recognizes that Poisson variance is different (smaller) at μ = μ_{LOWER} than at μ = μ_{UPPER}.

(4) (Naive) CI based on c ~ Gaussian(μ , $\hat{SD} = \sqrt{c}$).

If really lazy, or don't care about principles or accuracy, or if c is large (3 digits) might solve

$$c = \mu_{LOWER} + z\sqrt{c}\ ;\ c = \mu_{UPPER} - z\sqrt{c}$$

to give

$$\mu_{LOWER,UPPER} = c\ -/+\ z\sqrt{\mathbf{c}}$$

Accuracy of 5 approximations (95% CI's) in 5 eg's

Method	c = 3*	c = 6	c = 33**	c=78***	c=100
Exact	(2.48,35.1)	(2.20,13.1)	(22.7,46.3)	(61.7,97.3)	(81,122)
(1)	(2.41,35.1)	(2.19,13.1)	(22.7,46.3)	(61.7,97.3)	(81,121)
(1')	(3.32,32.0)	(2.49,13.4)	(23.1,45.8)	(62.1,96.8)	(82,121)
(2)	(2.26,29.4)	(2.16,11.8)	(22.7,45.2)	(61.7,96.3)	(81,121)
(3)	(4.08,35.3)	(2.75,13.1)	(23.5,46.3)	(62.5,97.3)	(82,122)
(4)	(–1.6,25.6)	(1.20,10.8)	(21.7,44.3)	(60.7,95.3)	(80,120)

* Rothman2002 p134 "3 cases in 2500 PY; pt. est. of Rate:12 per 10 000PY
Focus: No. per 10000PY (Rate) *rather than* on ave. No. in 2500PY
Focus for c=6, 33, 78 & 100: ave. No in same-size study (no den. given)

** No. of cancers among females and ***overall in Alberta SourGas study

Montreal Gazette, Friday May 12, 1989.

OTTAWA (CP) - 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* **[jh - footnote 1]** *, a larger sample size could place the true figure somewhere in the range from 0.4 cases to 12.6 cases.* **[jh - footnote 2]**

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.* **jh - footnote 3]**

--foot notes by JH ----------

[1] SIR = **3.5** = $\frac{\text{Observed}}{\text{Expected}}$ 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.

[2] CI = $\frac{\text{CI derived from O}}{\text{Expected}}$ = 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.**

[3] 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**.

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

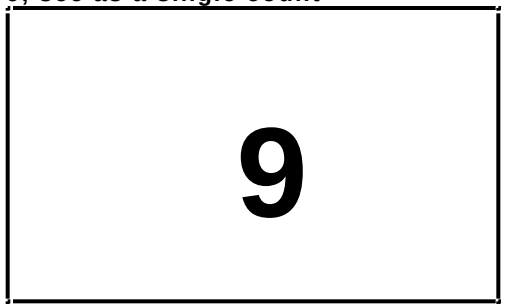
[More advanced] Uncertainty of an estimate based on a single Poisson count (see "CI for μ" earlier in this material)

It may seem strange that one can form a CI for μ from a single Poisson count c, whereas one would feel very uncomfortable doing so from a single measured (continuous) y. While it seems logical to use the single count c in a sample of one unit volume as the best point estimate of the **average** count per unit volume in the entire volume, that single number y does not seem to give an explicit measure of the uncertainty of the point estimate. In contrast, when "measuring" rather than counting, one uses an average $\bar{y}$ of the n sample quantities y_1 to y_n to estimate the μ per unit in the bigger universe;] one uses the empirical sd (s) of the n different y's to plug into the theoretical $\sigma/\sqrt{n}$ and thereby calculate a standard error $s/\sqrt{n}$ to accompany the point estimate; if n=1, so that one only has $\bar{y} = y_1$, one has to get an "outside" estimate of s.

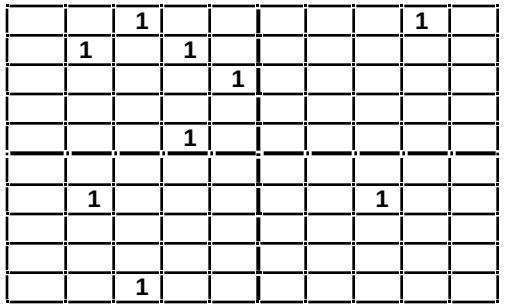
With some assumptions, a solution is possible without going "outside": split up the overall sample or slice of experience into (n) small enough sub samples so that the **subcount** y_i in each sub sample will be either a 0 or a 1 The variance of the observed sub counts should be p(1-p) where p is the proportion of sub counts that are 1. Thus the estimated variance of the total count c should be n times this, or np(1-p). But if p is small, so that 1-p is near unity, then the variance of the sub count is approximately np, which is simply the observed overall count c. i.e. the variance of a Poisson variable is equal to its mean.

Of course, when the count is small, it is not safe to use a Gaussian approximation, with a mean of μ and a sd of $\sqrt{\mu}$ to make inferences -- we should use the Poisson distribution itself to provide an idea of the uncertainty in a count or in any rate derived from it.

A count of c=9, see as a single count



The same 9, seen as a total of n=4 sub counts y_1 =5, .. y_4 =1 and as a sum of n=100 sub counts y_1 ... y_{100} (each 0/1)



$\hat{\mu} = \Sigma\, y_i = n\,\bar{y} = 9 = c.$ $\quad Var[\,\hat{\mu}\,] = Var[\,n\,\bar{y}\,] = n^2\, var[\,\bar{y}\,]$

Sub_ units n	size of each 1/n	sub_ counts $y_1 \ldots y_n$	$\hat{Var}(y)$ *	$\hat{Var}[\bar{y}]$	$\hat{Var}[n\,\bar{y}]$
1	entire	9	???		
4	1/4 ' s	5 1 2 1	4.6	1.16	18.3
100	1/100 's	1(9) 0(99)	$8.3\ 10^{-2}$	$8.3\ 10^{-4}$	8.3
1000	1/1000	1(9) 0(999)	$8.9\ 10^{-3}$	$8.9\ 10^{-6}$	8.9
10^m	$1/10^m$	$1(9)\ 0(10^m\text{-}9)$	$9.0\ 10^m$	$9.0\ 10^{2m}$	9.0

* the variance of the 4 sub counts 5, 1, 2, 1; the variance of the 91 zeroes and 9 ones, the 991 zeroes and 9 ones, etc.....

$\hat{Var}[\,\bar{y}\,] = var[\,y\,]\,/\,n$ *(familiar variance formula)* ; $\hat{Var}[n\,\bar{y}\,] = n^2\,\hat{Var}[\,\bar{y}\,]$.

Inferences regarding a single event rate parameter: i.e. rate of events per N [$=10^x$] units of experience

data: **c "events"** counted in sample of n units of "experience"; or Binomial(c,n) if c << n.

[can use c to calculate a rate i.e. empirical rate = $\frac{c}{n} \times$ N events per N units of experience; N usually 10^3 or 10^4 or the like]

See "Modern Epidemiology"(Rothman 1986) ; Observation & Inference (Walker) or Epidemiology: An introduction (Rothman, 2002, 133-134).

	Small no. of events	**Large no. of events**
CI for μ = E[c] E[c] is a parameter: the theoretical (unobservable) average number of events per n units; c refers to the realization in the observed sample ***Example***: If observe y=2 cases of leukemia in a certain amount of experience ('n'=P-Y) in a single "exposed" community , what is the 95% CI for the average number of cases (μ scaled to the same amount of experience) that (would) occur in (all such) exposed communities ?	• Use tabulated CI's e.g. p 20 in this material, the CRC handbook, Documenta Geigy scientific tables, Biometrika Tables for Statisticians, ... *(Most end at c=30 or c=50)* • If have to, can use (a) trial and error on spreadsheet, or .. (b) the link between the Poisson tail areas and the tail area of the chi-square distribution.	• Same as for small numbers, or... • One of 4 approximations on p 23 (1) Wilson/Hilferty approxn. to Chi-square quantiles (X^2<-->Poisson). (2) Square-root transformation of Poisson variable. (3) 1st Principles CI from c ~ Gaussian(μ , SD = $\sqrt{\mu}$) (4) (Naive) CI based on c ~ Gaussian(μ , $\hat{SD} = \sqrt{c}$). • X^2 and Likelihood Ratio (LR) methods (Miettinen Ch 10, pp 137-9)
CI for rate: $\frac{E[c]}{n} \times N$	$\frac{\textbf{CI for } \mu}{n} \times N$	$\frac{\textbf{CI for } \mu}{n} \times N$

See Liddell, FDK. Simple exact analysis of the standardized mortality ratio. Journal of Epidemiology and Community Health, March 1984, Vol 38, No. 1, pages 85-88.... on 626 website. This paper deals with SMR's but since the numerator of an SMR is treated as arising from a Poisson distribution, and the denominator as a constant, the results dealing with CI's for an SMR are also relevant just for the CI for a single Poisson parameter.

Inferences regarding a single event rate parameter: i.e. rate of events per N [$=10^x$] units of experience

data: **c "events"** counted in sample of n units of "experience"; or Binomial(c,n) if c << n. *(See again "Rothman and Walker).*

Small no. of events

Test $E[c] = E_0$

Example: Is the O=2 cases of leukemia at Douglas Point statistically significantly higher than the E=0.57 cases "expected" under the null for this many person years of observation?

Example What is the probability of getting 6 or more sets of twins in one school when the expected number, for schools of this size, is μ = 1.3?

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

P-Value obtained by adding the individual Poisson probabilities to obtain a tail area

(as done for Binomial and hypergeometric probabilities).

These individual probabilities are tabulated, for various 'round' values of E_0, on page 17 and in the sources listed above.

E or μ = 0.57 is not tabulated but μ=0.5 and μ=0.6 are.

P[2 or more events | μ=0.5] = (76+13+2)/1000 = **0.091**.
P[2 or more events | μ=0.6] = (99+20+3)/1000 = **0.122**. So,

P[2 or more events | μ=0.57] ≈ 0. 11 (upper tail p-value only)

Instead of interpolation for non-round values of E_0, use a calculator/ spreadsheet / statistical package. Excel and SAS have Poisson probability and cumulative probability functions built in.

E.g., the Excel Poisson(x, mean, cumulative) function returns a value of 0.89 when ones puts x=1, mean=0.57, cumulative = TRUE). This is the sum of the 2 tail probabilities P(0|E=0.57)= 0.57 andP(1|E=0.57) =0.32. The complement, 0.11, of the 0.89 is the upper tail p-value P(2) + P(3) + P(4) + ...

So the interpolation above is quite accurate.

Same procedure for c=6 vs. E=1.3 in twins data.

If one sets cumulative=FALSE, the Excel function calculates the probability at the integer x only, and does not sum all of the probabilities from 0 to x. For example, setting x=9, mean=16.0 and cumulative = FALSE (or 0) yields the P(9 | μ = 16.0) = 0.21 shown in the Figure on page 18 and in row 9 of the μ=16.0 column on p 17.

Large no. of events

- nomogram by Bailar & Ederer 1964*
- 2 Gaussian approximations (from page 23)

(2) **square root transformation of Poisson distribution** i.e.

$z = (\sqrt{c} - \sqrt{E_0})/(0.5)$.

$= (\sqrt{78} - \sqrt{85.9})/(0.5) =$ **- 0.87**

(4) **asymptotic normality of c** :

$z = (c - E_0) / \sqrt{E_0}$

$= (78 - 85.9) / \sqrt{85.9} =$ **- 0.85**

Squaring (4) gives X^2 form (1 df)

$X^2 = (c - E_0)^2 / E_0$

$= (78 - 85.9)^2 / 85.9 =$ **0.72**

- Miettinen Chapter 10

* Bailar, J.C. & Ederer, F. Significance factors for the ratio of a Poisson variable to its expectation. Biometrics, Vol 20, pages 639-643, 1964.

Inference concerning comparative parameters: Rate Difference (RD) and Rate Ratio (RR)

Rate Parameters $\mathbf{R_1}$ and $\mathbf{R_0}$; Rate Difference Parameter $\mathbf{RD = R_1 - R_0}$

data: c_1 and c_0 "events" (total $c = c_1 + c_0$) in n_1 and n_0 (total=n) units of experience"; empirical rates $r_1 = \frac{c_1}{n_1}$ and $r_0 = \frac{c_0}{n_0}$;

[e.g.Rothman & Boice compare c_1=41 in n_1 =28,010 person years (PY) with c_0=15 in n_0 =19,017 person years (PY)]

	Small no. of events	Large no. of events
CI for **RD**	"Exact" methods are difficult, since t he presence of a nuisance parameter complicates matters. See papers by Suissa and by Nurminen and Miettinen. Note however that even if numerators (c_1 and c_0) are small (or even zero!) one may still have considerable precision for a rate difference: if statistical uncertainty about each rate is small, the uncertainty concerning their difference must also be small. Contrast this with situation for RR, where small numerators make RR estimates unstable. (see report by J Caro on mortality following use of low and high osmolar contrast media in radiology)	$r_1 - r_0 \pm z\sqrt{\{SE[r_1]\}^2 + \{SE[r_0]\}^2}$ in our example... $\frac{41}{28010} - \frac{15}{19017}$ $\pm 1.96\sqrt{\frac{\frac{41}{28010}(1 - \frac{41}{28010})}{28010} + \frac{\frac{15}{19017}(1 - \frac{15}{19017})}{19017}}$ Can dispense with the "1 minus small rate" term in each (binomial) variance, so the standard error of the rd simplifies to $\sqrt{\frac{c_1}{n_1^2} + \frac{c_0}{n_0^2}}$

(see Walker ; or Rothman 2002, pp 137-138)

Inference concerning comparative parameters: Rate Difference (RD) and Rate Ratio (RR)

Rate Parameters $\mathbf{R_1}$ and $\mathbf{R_2}$ Rate Ratio Parameter $\mathbf{RR = R_1 / R_0}$ *See Rothman 2002, pp 137-138)*

data: c_1 and c_0 "events" (total $c = c_1 + c_0$) in n_1 and n_0 (total=n) units of experience"; empirical rates $r_1 = c_1/n_1$ & $r_0 = c_0/n_0$;

Small no. of events

CI for RR

Use distribution of c_1 conditional on $c = c_1 + c_0$ [56 in e.g. -- *not that small !*]

Conditioning on the total no. of cases, c, gets rid of one (nuisance) parameter, and lets us focus on the observed "proportion of exposed cases " (c_1/ c) and its theoretical (parameter) counterpart.

In e.g., proportion of "exposed" PY $= \frac{28010}{28010 + 19017} = 0.596 = 59.6\%$

There is a 1:1 correspondence between the expected proportion of exposed cases (call it π for short) and the RR parameter, and correspondingly between the observed proportion (p) of exposed cases and the point estimate, rr,of the rate ratio.

Under the null (RR=1), π clearly equals the proportion 0.596;

If RR > 1, this expected proportion π is higher; for example if RR=2, so that each exposed PY generates 2 times as many cases as an unexposed PY,

$$\pi = \frac{28010 \times 2}{28010 \times 2 + 19017} = 74.7\% = 0.747.$$

Thus, in our example... (and in general, $\pi = \frac{n_1 \times RR}{n_1 \times RR + n_0}$)

RR	0.25	0.50	1.00	2.00	4.00	8.00
π (proportion of exposed cases)	0.269	0.424	0.596	0.747	0.855	0.922

The observed proportion of exposed cases is p = 41/56 = 0.732; in our table, the 0.732 corresponds to an RR point estimate just below 2.

We can reverse the general formula to get RR = $\{\pi/(1-\pi)\} / \{n_1/n_0\} = \{\pi/(1-\pi)\}\{n_0/n_1\}$

So, in our e.g., the point estimate of RR is rr = (0.732/0.268) / (28010/19017) = 1.86 .

To obtain a CI, we treat the proportion of exposed cases, 0.732, as a binomial proportion, based on 41 "positives" out of a total of 56 cases (obviously, if the proportion were based on 8 exposed cases out of 11 cases, or 410 out of 560, the precision would be very different!)

From table/other source of CI's for proportions (see e.g. table on 607 web page), can determine that 95% CI for π is π_L=0.596 to π_U=0.842. Substitute these for the point estimate to get

RR_L = (0.596 / 0.404) / (28010/19017) = 1.00 RR_U = (0.842/0.158) / (28010/19017) = 3.61

Rothman & Walker emphasize formula $RR_{L,U} = \{\pi_{L/U} / (1-\pi_{L/U})\} / \{n_1 / n_0\}$ over basis for it.

SEE EXAMPLE IN 626 EXAM IN 2002 (0 and 41 seroconversions following vaccination vs HPV)

Large no. of events

• Use same conditional (binomial-based) formula as for small no. of events, but use Gaussian approxn. to get Binomial CI for π

• Test-based CI (Miettinen)

Uses fact that in vicinity of RR=1, can obtain SE for ln(rr) indirectly from null X^2 test statistic

X^2 statistic = square of Zstatistic = $4.33 = 2.08^2$ in e.g.

$$\text{Xstatistic} = \text{Zstatistic} = \frac{\ln(rr) - 0}{SE[\ln(rr)]}$$

$$\text{so } SE[\ln(rr)] = \frac{\ln(rr)}{\text{Xstatistic}}$$

$$\text{CI for } \ln(RR) = \ln(rr) \pm z \frac{\ln(rr)}{\text{Xstatistic}}$$

CI for RR: rr to power $[1 \pm \frac{z}{\text{Xstatistic}}]$

= 1.86 to power of [1 ± 1.96/2.08]
= 1.04 to 3.32 in e.g.

• $\text{Var}[\ln(rr)] = \frac{1}{c_1} + \frac{1}{c_0} + \frac{1}{\infty} + \frac{1}{\infty}$ (Woolf)

CI for **RR** $= rr \exp[\pm z \sqrt{\frac{1}{c_1} + \frac{1}{c_0}}]$

$1.96\ (1/41+1/15)^{1/2} = 0.59$ in e.g. ;

so exp[0.59]=1.81; So CI for RR

=.86 / 1.81 to 1.86*1.81 = (1.02,3.35)

Precision for ln(RR) estimate depends on numbers of events c_1 and c_0.

Inference concerning comparative parameters: Rate Difference (RD) and Rate Ratio (RR)

Rate Parameters $\mathbf{R_1}$ and $\mathbf{R_2}$ Rate Difference Parameter $\mathbf{RD = R_1 - R_0}$ Rate Ratio Parameter $\mathbf{RR = R_1 / R_0}$

data: c_1 and c_0 "events" (total $c = c_1 + c_0$) in n_1 and n_0 (total=n) units of experience"; empirical rates $r_1 = c_1/n_1$ & $r_0 = c_0/n_0$;

	Small no. of events	**Large no. of events**
test of **RD=0** **or** **RR=1**	• Null distribution of c_1 conditional on c $c_1 \mid c$ ~ Binomial, with c "trials", (see above) each with null probability $\pi = \frac{\mathbf{RR} \times n_1}{\mathbf{RR} \times n_1 + n_0}$. e.g. If RR =1 (RD=0) would expect the 56 cases to split into "exposed" and "unexposed" in the proportions 27010/(27010+19017) = 0.596 and 1-0.596=0.404 respectively. Can test if the observed proportion 41/56 = 0.732 is significantly different from this null expectation using a Binomial distribution with "n"=56 and π =0.596. Can use the Excel Binomial function with x=40,mean=0.596,cumulative=TRUE, to get the sum of all the probabilities up to and including 40. Subtract this quantity 0.976 from 1 to get the probability 0.024 of 41 or more (upper tail area). Double this for a 2-sided test. • Unconditional test for proportions / rates (Suissa)	• Use same "c_1 conditional on c" test but use Gaussian approx^n to Binomial (c, π) e.g. $z = \frac{[41/56 =]0.732 - 0.596}{\sqrt{0.596 \times 0.404/56}} = 2.08$ P(Z > z) = 0.019 (upper tail area). Double for 2-sided test. • $z = \frac{[\, r_1 - r_2\,] - RD_0}{\sqrt{\{SE[r_1 \mid H_0\}^2 + \{SE[r_2 \mid H_0\}^2}}$ {*SE's use r = Σc/Σn [pooled data]} • $X^2 = \frac{\{c_1 - E[\, c_1 \mid H_0]\}^2}{E[\, c_1 \mid H_0]} + \frac{\{c_0 - E[\, c_0 \mid H_0]\}^2}{E[\, c_0 \mid H_0]}$ $= \frac{\{c_1 - E[c_1 \mid H_0]\}^2}{Var[\, c_1 \mid H_0]}$ (Mantel-Haenszel version) *See my notes on Chi-square tests in on chapter 8 in 607 course*

Numbers in body of table are expected number of events required in Group 1 to give specified power if relative rate in Group 2 is R.

Relative Rate**	Expected events in Group 1 to give: * 80% Power	90% Power	95% Power
0.1	10.6	14.3	17.6
0.2	14.7	19.7	24.3
0.3	20.8	27.9	34.4
0.4	30.5	40.8	50.4
0.5	47.0	63.0	77.8
0.6	78.4	105.0	129.6
0.7	148.1	198.3	244.8
0.8	352.8	472.4	583.2
0.9	1489.6	1994.5	2462.4
1.1	1646.4	2204.5	2721.6
1.2	431.2	577.4	712.8
1.4	117.6	157.5	194.4
1.6	56.6	75.8	93.6
1.8	34.3	45.9	56.7
2.0	23.5	31.5	38.9
2.5	12.2	16.3	20.2
3.0	7.8	10.5	13.0
5.0	2.9	3.9	4.9
10.0	1.1	1.4	1.8

** Ratio of incidence rate in Group 2 to incidence rate in Group 1.

* Using a two-sided significance test with $p<0.05$.
The two groups are assumed to be of equal size *(B&D more general)*

Taken from Table 3.2 in Chapter 3 "Study Size" in "Methods for Field Trials of Interventions against Tropical Diseases: A Toolbox" Edited by P.G. Smith and Richard H. Morrow. Oxford University Press Oxford 1991. (on behalf of the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases)

Note that roles of Group 1 and 2 above are reversed from in Smith & Morrow text; *See also Breslow NE and Day NE Vol II, Section 7.3*

Formulae for calculating study size requirements for comparison of rates using two groups of equal size

from Table 3.4 of Morrow and Smith, with role of groups 1 & 2 reversed.

Formula — Section in text

Notation

• *Choosing study size to achieve adequate precision*

$$\mathbf{e_1 = (1.96/\log_e f)^2 (R + 1)/R}$$

e_1 = Expected no. of events in group 1 — 3.2
R = Rate in group 2/Rate in group 1
Gives 95 per cent CI from R/f to Rf

• *Choosing study size to achieve adequate power*

$$\mathbf{P\text{-}T = (z_\alpha + z_\beta)^2 (r_2 + r_1)/(r_2 — r_1)^2}$$

P-T = Person-time in each group — 4.2
r_i = Rate in group i

$\mathbf{z}_\alpha$ = 1.96 for significance at $p < 0.05$

Power	80%	90%	95%
$\mathbf{z}_\beta$	0.84	1.28	1.64