m-s-1. Working with logs of counts and logs of rates
In order to have a sampling distribution that is closer to Gaussian – sample counts, and ratios of them tend to have nasty sampling distributions –we often transform from the $(0, \infty)$ scale for a count y and its expectation, μ, to the $(-\infty, 0)$ $\log[y]$ and $\log[\mu]$ scale.

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

i. Suppose $Y \sim$ Poisson(μ) with associated rate estimate $\hat{\lambda} = Y/PT$[1]. Derive the variances for the random variables $\log[Y]$ and $\log[\hat{\lambda}]$. Ignore the possibility of obtaining $\hat{\mu} = 0$ i.e., $\hat{\lambda} = 0/PT = 0$.

ii. What is the variance for the log of a rate ratio, i.e., $\log[\hat{\lambda}_2 \div \hat{\lambda}_1]$?

m-s-2. The Poisson Family as a 'Closed under Addition' Family

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

m-s-3. Link between Poisson and Exponential Distributions

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

m-s-4. Link between tail areas of Poisson and χ^2 Distributions

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

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

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

[1] PT = amount of Population Time

m-s-5. The Fisher information that a Poisson random variable carries about its expectation and about the log of this expectation

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

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

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

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

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

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

2. Comparison of various CI's for the expectation, μ of a Poisson random variable, on the basis of a single count y
Fill in the blanks in the table below, and compare the accuracy of different approximations to the exact 95% CI for μ, based on a count of y.

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

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

3. Power Calculations

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

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

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

4. Perfect Results ?

The following excerpt is from the Vaccine Arm of Table 3 of an Article in the NEJM in 2002[2]. We will look at the comparison with the Placebo arm when we get to comparative studies.

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

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

(1) Primary per-protocol

(2) Including women with general protocol violations

(3) Secondary per-protocol

P = Persistent; T=transient

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

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

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

Questions

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

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

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