ms-1 Distribution of $y_1 \mid y_1+y_2$, when $y_1 \sim Poisson(\mu_1)$ and (independently) $y_2 \sim Poisson(\mu_2)$

Exercise 4.15, Casella & Berger, p194.

ms-2 Two ways of deriving the Negative Binomial Distribution

• As a sum of k i.i.d. Geometric r.v.'s

The '**zero**-origin' version of the Geometric (G) probability distribution is the probability distribution of the number Y of failures **before** the first success, supported on the set {0, 1, 2, 3, ... }, when the probability of success on each trial is π.

$$Prob_G[Y = y] = (1 - \pi)^y \times \pi.$$

With this zero-origin version, the Negative Binomial (NB) probability distribution with parameters k and π is the probability distribution of the number Y of failures **before** the k-th success, again supported on the set $\{0, 1, 2, 3, ...\}$.

Exercise: Show that, so defined,

$$Prob_{NB}[Y = y] \;=\; {}^{(y+k-1)}C_{k-1} \times (1 - \pi)^y \times \pi^k.$$

and (from the pmf, or directly from the definition as a sum) find its expectation and variance.

• The negative binomial distribution can also be thought of as a continuous mixture of Poisson (P) distributions where the mixing distribution of the Poisson rate [or mean–JH] is a gamma distribution. Formally, this means that the mass function of the negative binomial distribution can also be written as

$$\begin{aligned} Prob_{NB}[Y = y] &= \int_0^\infty pmf_P[y \mid \mu] \times pdf_\Gamma[\mu \,|k, (1 - \pi)/\pi \,]d\mu \\ &= \ ... \\ &= \frac{\Gamma[k + y]}{y!\ \Gamma[k]} \times \pi^k \times (1 - \pi)^y. \end{aligned}$$

Because of this, the negative binomial distribution is also known as the gamma-Poisson (mixture) distribution.

Exercise: fill in the omitted steps.

(above text from `http://en.wikipedia.org/wiki/Negative_binomial_distribution`)

From `R` documentation... [see `nbinom` in `R`]

A *negative binomia*l distribution can arise as a mixture of Poisson distributions with mean distributed as a Γ (gamma) distribution with scale parameter $(1 - \pi)/\pi$ and shape parameter k. (This definition allows non-integer values of k.) In this model $\pi = scale/(1 + scale)$, and the mean is $k \times (1 - \pi)/\pi$.

The alternative parametrization (often used in ecology) is by the mean μ_{NB}, and k, the dispersion parameter, where $\pi = k/(k + \mu)$. The variance is $\mu_{NB} + \mu_{NB}^2/k$ in this parametrization or $k \times (1 - \pi)/\pi^2$ in the first one.

Overdispersed Poisson: The negative binomial distribution, especially in its alternative parametrization described above, can be used as an alternative to the Poisson distribution. It is especially useful for discrete data over an unbounded positive range whose sample variance exceeds the sample mean. If a Poisson distribution is used to model such data, the model mean and variance are equal. In that case, the observations are overdispersed with respect to the Poisson model. Since the negative binomial distribution has one more parameter than the Poisson, the second parameter can be used to adjust the variance independently of the mean...

ms-3 Sample size formulae for comparison of rates

See section 3.1 of the Notes. Derive the "Expected numbers of events required ... specified power" from 'scratch,' using normal approximations to the Poisson distributions of the observed numbers of events in the two (equal) amounts of population-time.

-1- Extended Work Duration and the Risk of Self-reported Percutaneous Injuries in Interns

Refer to rows 2 and 3 of Table 3. in this article, by Ayas et al. in JAMA on Sept 6 of 2006. [Resources - Intensity]

i. Manually calculate ORs and 95% CIs, and repeat by computer software.

ii. Explain why your answers do not match those reported (hint: see the paragraph beginning "To assess the relationships..." in the last column of page 1057 of the article.

iii. exactly what (and how many) numbers would you need to carry out their analysis for row 3 (injuries in ICU). Answer in the form of a 1-paragraph request to the authors asking for these specific numbers (but do not e-mail the authors! JH has in fact obtained these numbers from Dr Ayas, and they will form the basis for some of a future homework).

iv. Is OR the correct term for the ratio being estimated here?

-2- John Snow's "Grand Experiment"

"According to a return which was made to Parliament, the Southwark and Vauxhall Company supplied 40,046 houses from January I to December 31, 1853, and the Lambeth Company supplied 26,107 houses during the same period; " So, the *denominators* were...

No. of Houses with...	
Water Source	
Dirtier	Cleaner
40046	26107

"286 fatal attacks of cholera took place, in the first four weeks of the epidemic, in houses supplied by the former company, and only 14 in houses supplied by the latter."

i. Calculate a 95% CI to accompany the rate ratio.

ii. But what if the sizes of the two denominators were not readily available – but the numerators were? It would be a lot of 'leg work' (also known as 'shoe-leather epidemiology' to determine the water source of each of 40046 + 26107 = 66153 houses!

Use `R (or SAS or SPSS)` to form a sample of 100 houses that could form a '*denominator series.*' and thus to provide a point and interval estimate of the relative sizes of the two sources of water. Use the *case series* assembled by Snow, and the (armchair / virtual / desktop) *denominator series* obtained by you and `R`.

iii. Repeat this virtual epidemiology, but using denominator series of 300, 600, and 2000. Comment on the estimates.

iv. Explain to a journalist why are the CI's based on the virtual denominator series are wider than the one based on the actual "return which was made to Parliament"?

-3- A population-based study of measles, mumps, and rubella vaccination and autism[1]

Background: It has been suggested that vaccination against measles, mumps, and rubella (MMR) is a cause of autism.

Methods: We conducted a retrospective cohort study of all children born in Denmark from January 1991 through December 1998. The cohort was selected on the basis of data from the Danish Civil Registration System, which assigns a unique identification number to every live-born infant and new resident in Denmark. MMR-vaccination status was obtained from the Danish National Board of Health. Information on the childrens autism status was obtained from the Danish Psychiatric Central Register, which contains information on all diagnoses received by patients in psychiatric hospitals and outpatient clinics in Denmark. We obtained information on potential confounders from the Danish Medical Birth Registry, the National Hospital Registry, and Statistics Denmark.

Results: Of the 537,303 children in the cohort (representing 2,129,864 person-years), 440,655 (82.0 percent) had received the MMR vaccine. We identified 316 children with a diagnosis of autistic disorder and 422 with a diagnosis of other autistic-spectrum disorders. After adjustment for potential confounders, the relative risk of autistic disorder in the group of vaccinated children, as compared with the unvaccinated group, was 0.92 (95 percent confidence interval, 0.68 to 1.24), and the relative risk of another autistic-spectrum disorder was 0.83 (95 percent confidence interval, 0.65 to 1.07). There was no association between the age at the time of vaccination, the time since vaccination, or the date of vaccination and the development of autistic disorder.

Conclusions: This study provides strong evidence against the hypothesis that MMR vaccination causes autism.

• No. of Cases of autism (numerators) among children who did / did not receive MMR vaccination ...

Vaccinated		
Yes (1)	No (0)	
263	53	316 (**Cases**)

• No. of children-years (cy) of follow-up [contributed by 0.54 m children]

Vaccinated		
Yes (1)	No (0)	
1.65m cy	0.48m cy	2.13m cy (**Denominators**)

• Crude* Rate Ratio ...

Vaccinated				
Yes (1)	No (0)	**Rate Ratio (RR)**	ME	**95% CI for RR**
$\frac{263}{1.65m\ cy}$	$\frac{53}{0.48m\ cy}$	1.44*	1.34	1.07 to 1.93

$ME = \exp[1.96 \times \{1/263 + \ 1/53\}^{1/2}]$;
$RR_{lower} = 1.44 \div 1.34$; $RR_{upper} = 1.44 \times 1.34$.

* The reason for the large difference between the crude (1.44) and adjusted (0.92, 95% CI 0.68 to 1.24) rate ratios will be discussed when we come to confounding. The crude ratio in this example is simply for didactic purposes.

i. Hand-calculate the CI's for the 1.44 ratio by your usual ± way, and

[1] Madsen KM et al. N Engl J Med 2002;347:1477-82.

compare the 'work' involved with that in the "multiplied-by/divided-by" method – ie be a 'hand-calculator consultant' to Rothman.

ii. Which method do you prefer? (if you have software that does it for you, this is merely a heuristic issue!)

-4- A Controlled Trial of a Human Papillomavirus Type 16 Vaccine

Background: Approximately 20 percent of adults become infected with human papillomavirus type 16 (HPV-16).Although most infections are benign, some progress to anogenital cancer. A vaccine that reduces the incidence of HPV-16 infection may provide important public health benefits.

Methods: In this double-blind study, we randomly assigned 2392 young women (defined as females 16 to 23 years of age) to receive three doses of placebo or HPV-16 virus-like particle vaccine (40 ?g per dose), at day 0, month 2, and month 6. Genital samples to test for HPV-16 DNA were obtained at enrollment, one month after the third vaccination, and every six months thereafter. Women were referred for colposcopy according to a protocol. Biopsy tissue was evaluated for cervical intraepithelial neoplasia and analyzed for HPV-16 DNA with use of the polymerase chain reaction. The primary end point was persistent 16 infection, defined as the detection of HPV-16 DNA in samples obtained at two or more visits. The primary analysis was limited to women who were negative for HPV-16 DNA and HPV-16 antibodies at enrollment and HPV-16 DNA at month 7.

Results: The women were followed for a median of 17.4 months after completing the vaccination regimen. The incidence of persistent HPV-16 infection was 3.8 per 100 woman-years at risk in the placebo group and 0 per 100 woman-years at risk in the vaccine group (100 percent efficacy; 95 percent confidence interval, 90 to 100; P¡0.001).All nine cases of HPV-16related cervical intraepithelial neoplasia occurred among the placebo recipients.

Conclusions: Administration of this HPV-16 vaccine the incidence of both HPV-16 infection and HPV- 16related cervical intraepithelial neoplasia. Immunizing HPV-16negative women may eventually reduce the incidence of cervical cancer.

(N Engl J Med 2002;347:1645-51.). See full article on Resources-Applications webpage.

i. Why this design rather than a "fixed number of woman-years-of-follow-up" design?

ii. With I denoting incidence, v denoting the vaccinated and u the unvaccinated, Efficacy (E) is defined here as a percentage

$$E = 100 \times (I_u - I_v)/I_u = 100 \times (1 - I_v/I_u).$$

Consider a very large R.C.T. (so random variation is not an issue), with 1/2 receiving the vaccine and 1/2 the placebo, and concentrate on the total number of cases (of persistent infection). What is the relation between the (theoretical) proportion (Π) of these cases that would be in the vaccinated group (i.e. what fraction of cases would be 'vaccinated' cases) and the vaccine efficacy E?

To answer, calculate for every 1 case in the unvaccinated, how many cases (c_v) there would be in the vaccinated; then express the c_v as a proportion of $(c_u + c_v)$.

$E(\%)$	0	25	50	75	80	90	99
c_u	1	1	1	1	1	1	1
c_v	--	--	--	--	--	--	--
Π	--	--	--	--	--	--	--

Π =proportion of cases that received $v = c_v/(1 + c_v)$.

iii. Suppose that in the actual (finite) study, subject as it was to random variations, the authors had analyzed the data when the *total* number of cases was $c_u + c_v = 31$, i.e. when the observed proportion of cases that had been vaccinated was p=0/31 i.e., when the point estimate for Π was p=0.0. This point estimate translates into an 'exact' 95% 2-sided [binomial-based] CI for Π of 0.0 to 0.11. From this CI, and interpolation in the table you just constructed, find 'exact' 95% limits for E.

-5- Women are Safer Pilots

LONDON- Initial results of a study by Britain's Civil Aviation Authority shows that women behind the controls of a plane might be safer than men. The study shows that male pilots in general aviation arc more likely to have accidents than female pilots. Only 6 per cent of Britain's general aviation pilots are women. According to the aviation magazine Flight International, there have been 138 fatal accidents in general aviation in the last 10 years, and only two involved women - less than 1.5 per cent of the total.

Woman News, page F1 The Montreal Gazette, August 21st, 1995

The large-sample methods for obtaining a CI for a rate ratio are accurate when there are enough events in each of the compared categories. But in-4-above, and in the "Women are Safer Pilots" example, the small number of events in one of the categories renders large-sample methods inaccurate or even impossible. In such situations, the conditional approach, in which one bases the inference on the distribution of the number of events in one category, conditional on the sum of the numbers of events in the two categories, is a way around this problem (we use a similar conditioning strategy when dealing with Fisher's exact test).

i. Compare the rate of accidents in women relative to men pilots (i.e. the rate ratio)

   (a) Assuming that on average, the women pilots fly just as many hours as the men pilots, and that all other relevant factors are equal [although they probably are not!]. Based on the information given, use software to calculate an exact CI for the rate ratio

   (b) Assuming that on average the women pilots fly half as many hours as the men.

ii. In your own words, describe the basis for the exact method. [JH will use your answers to judge how clear or muddled *his* description is!]

-6- The 1954 Field Trial of the Salk Poliomyelitis Vaccine[2]

Summary of Study Cases by Diagnostic Class and Vaccination Status (Rates per 100,000): **Placebo control areas: All Reported Cases***

Gp.	SP	A_c	A_r	T_c	T_r	PP_c	PP_r	NP_c	NP_r	FP_c	FP_r
V	200,745	82	41	57	28	**33**	***16***	24	12	-	-
Pl	201,229	162	81	142	71	**115**	***57***	27	13	4	2
NI	338,778	182	54	157	46	**121**	***36***	36	11	-	-
IV	8,484	2	24	2	24	1	12	1	12	-	-
All	749,236	428	57	358	48	270	36	88	12	4	1

V, Pl, NI, IV: Vaccinated, Placebo, Not Inoculated, and Incomplete Vaccinations groups.

SP: Study population (number of children);
A_c and A_r: *All* reported $c$ases and $r$ate;
T_c and T_r: Total poliomyelitis $c$ases and $r$ate;
PP_c and PP_c: Paralytic Poliomyelitis $c$ases and $r$ate;
NP_c and NP_r: Non-Paralytic poliomyelitis $c$ases and $r$ate;
FP_c and FP_r Fatal poliomyelitis $c$ases and $r$ate.
Some 70 reported cases were deemed to be "Not Polio" (25 in V, 20 in Pl, and 25 in the NI, are shown in Meier's table, but omitted here because of space constraints. Meier's Source: Adapted from Francis (1955), Tables 2 and 3.

Exercise: Compute point and interval estimates of the *difference in the rates* of paralytic polio with the Salk vaccine and Placebo, together with the percent *efficacy*. Use all of the approximate and exact approaches that you are aware of, and compare the results.

[2]Paul Meier. Chapter 2 The Biggest Public Health Experiment Ever: in Tanur JM et al. (Editors) *Statistics: A Guide to the Unknown.* Holden-Day San Francisco 1972.

-7- Effect of Raloxifene on Risk of BrCa in Postmenopausal Women

Context: Raloxifene hydrochloride is a selective estrogen receptor modulator that has antiestrogenic effects on breast and endometrial tissue and estrogenic effects on bone, lipid metabolism, and blood clotting. **Objective**: To determine whether women taking raloxifene have a lower risk of invasive breast cancer. **Design and Setting**: The Multiple Outcomes of Raloxifene Evaluation (MORE), a multicenter, randomized, double-blind trial, in which women taking raloxifene or placebo were followed up for a median of 40 months (SD, 3 years), from 1994 through 1998, at 180 clinical centers composed of community settings and medical practices in 25 countries, mainly in the United States and Europe. Participants A total of 7500 postmenopausal women, younger than 81 (mean age, 66.5) years, with osteoporosis. Women who had a history of breast cancer or who were taking estrogen were excluded. **Intervention**: Raloxifene, 60 mg, 2 tablets daily; or raloxifene, 60 mg, 1 tablet daily and 1 placebo tablet; or 2 placebo tablets. **Main Outcome**: Measures New cases of breast cancer, confirmed by histopathology. Deep vein thrombosis or pulmonary embolism were determined by chart review. **Results**: Thirteen cases of breast cancer were confirmed among the 5000 women assigned to raloxifene vs 26 among the 2500 women assigned to placebo (relative risk [RR], 0.25; 95% confidence interval [CI], 0.13-0.49; Chi-Square = 19.5, P¡.001). To prevent 1 case of breast cancer, 128 women would need to be treated. Raloxifene increased the risk of venous thromboembolic disease (RR, 3.0; 95% CI, 1.5-6.1), but did not increase the risk of endometrial cancer (RR, 0.8; 95% CI, 0.2-2.7). **Conclusion**: Among postmenopausal women with osteoporosis, the risk of invasive breast cancer was decreased by 75% during 3 years of treatment with raloxifene.
Notes: (1) The numbers in the abstract have been "rounded" to make calculations easier. (2) The data from the 2 regimens were combined in the abstract. Thus, twice as many women received raloxifene as placebo.

i. Show how the authors calculated that "128 would need to be treated"

ii. Reproduce the CI accompanying the RR of 0.25.

iii. Would a test-based CI to give close to the same CI? Compute it and see.

iv. If 3750 women each had been allocated to raloxifene & placebo, and the cancer rates been the same, would CI be narrower, wider or same?

v. Balance of benefits and risks: the abstract does not to report the numbers of thromboembolic disease, only the RR of 6 and the CI. From the information provided, determine – analytically or by trial and error – how many cases there were [Hint: $\log 1.5 = 0.4; \log 3 = 1.1; \log 6.1 = 1.8$; and use 2 as an approximation to 1.96]

vi. In Table 2 of the paper, the authors report 15,000 and 7,500 women years of follow-up in the two groups. Does using these rather than the "person" denominators in the abstract change the CI's? Why/why not?