EpiNotes

Lewis Campbell

2019-04-12

Introduction

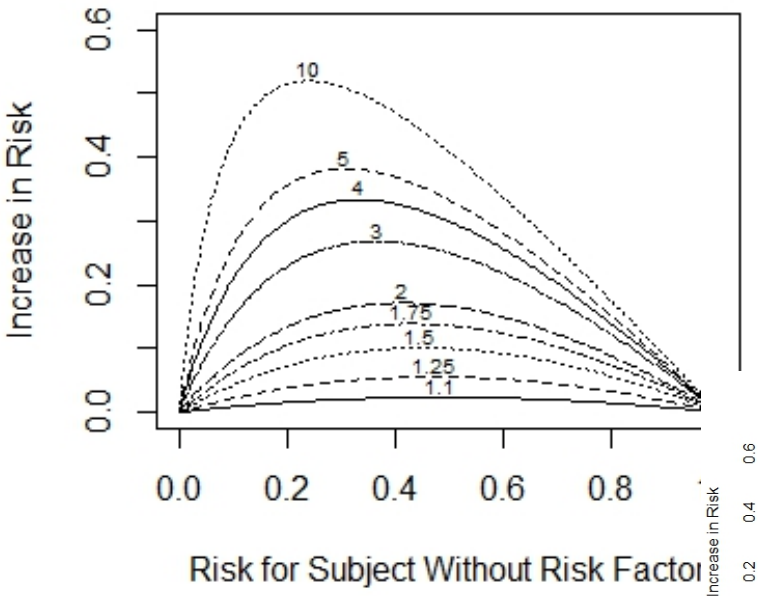
This is the notes from my *MSc Epidemiology* from LSHTM, one of the two most respected Epidemiology institutions in the world. It's a dog.

I wish I had done this twenty years ago, and spent the intervening time studying real statistics and computer science, instead of listening to people vomit party lines in my forties.

Headings

THE BASELINE ODDS IS KEY TO UNDERSTANDING[1] the effect of a given odds ratio on the absolute risk difference. While the odds ratio is the best way to state a constant effect of an exposure on the likelihood of an outcome, it needs to be applied to a baseline absolute risk to display itself.

[1] This is just a test



#The contingency table

A table allows things to be seen, and calculations referred to. "Expected values" for the cells in the table can be calculated under vari-

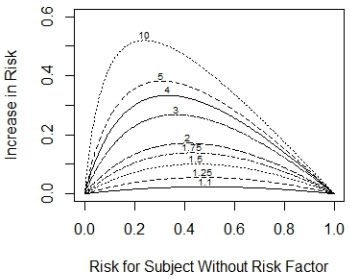


Figure 1: Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.

ous hypotheses using the marginal totals: these E don't have to be all possible at once, so "expected" for rows and columns will differ.

	+	-	
D	a	b	N_D
D^c	c	d	N_{D^c}
	N_+	N_-	N

Table 1: Marginal calculations

sensitivity	$= P(+\|D)$	= a/(a+b)
specificity	$= P(-\|DC)$	= d/(c+d)
ppv	$= P(D\|+)$	= a/(a+c)
npv	$= P(DC\|-)$	= d/(d+b)
accuracy	= P(correct)	= (a+d)/(a+b+c+d)
variance	$= \frac{(a+b).(a+c).(b+d).(c+d)}{n^2(n-1)}$	

DLR+ $= \frac{P(+|D)}{P(+|D^C)} = \frac{\frac{a}{(a+b)}}{\frac{c}{(c+d)}} = \frac{(ac+ad)}{(ac+ab)}$ = sensitivity/(1-specificity)

```
= (diagnostic likelihood ratio of a positive test) = Bayes Factor

        because PostOdds = Prior*DLR,

a DLR of N means the H(D) is N times more supported by data than is $H(D^C)$
```

SE for log odds ratio $= \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$ so the Error Factor is $e^{z_\alpha \times \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$ log 95% CI is OR +/- log EF , so 95% CI is OR/EF to OR*EF or exp(log Estimate +/- log EF)

Measures of association: Relative risk is re/ru Attributable risk is the risk difference and obviously unique to pairs of risks re-ru

Attributable fraction AF is the proportion of the higher risk that is unique to the higher risk or $AF = \frac{r_e - r_u}{r_e}$, whence, multiplying all by ru and substituting $RR = \frac{r_e}{r_u}$: $AF = \frac{RR-1}{RR}$ Population attributable risk PAR is the extra population risk due to exposure

$$PAR = r_t - r_u$$

or the attributable risk times the prevalence of exposure

$$PAR = p.AR = p(r_e - r_u); PAF = p' AF_E = \frac{p'(RR-1)}{RR}$$

,

```
where p is proportion of the population exposed, and p' is proportion of cases are exposed.
```

because total cases if a fraction p of the population is exposed, 0<p<1, it's like a binomial

$$r_t = p.r_e + (1-p)r_u$$

So anyway PAF is like attributable fraction,

$$\frac{r_t - r_u}{r_t} = \frac{PAR}{r_t} = \frac{p.(r_e - ru)}{r_t} = \frac{p.(RR-1)}{p.(RR-1)+1}$$

Call the parameter of interest theta; whether it's risk ratio, rate ratio, odds ratio (I know, sshh…) A cohort estimates theta directly but only estimates p and p' if it's a random sample, otherwise external estimates are used. Then the above equations are used. In cross sectional studies prevalence is estimated, incidence can't be. In case control, "theta is estimated by the odds ratio"; as $\theta = ad/bc$ and $PAF = p'(theta–1)/theta$ and $p' = a/n$ so $PAF = \frac{1-(b/n_1)}{(d/n_0)}$, the ratio of cases unexposed to controls unexposed. Remember, kids: the assumption that PAF describes a causal relationship might not be correct: it might describe a confounder, or have an effect modification. A confidence interval for PAF is given using 1-PAF:

$$1 - (1 - \text{PAF} \overset{\times}{\div} e^{1.96\sqrt{\frac{a}{bn_1} + \frac{c}{dn_0}}})$$

PAF for several levels of an exposure is given by $\sum \frac{p'_k(\theta_k - 1)}{\theta_k}$ or the equation $\frac{p'1-\theta}{\theta}$, using as theta a regression estimate with whatever covariates are desired to be included. The joint PAF for independent exposure is

$$1 - PAF_T = 1 - PAF_1 + 1 - PAF_2 ... 1 - PAF_i$$

AF_E is often used to assign blame by transposing the conditional: as it is the proportion of cases that are associated with the exposure it is called the probability of causation or the assigned share of causation. This is clearly an error unless rare conditions are met.

	Outcome	No Outcome	
Exposure	30	20	50
No Exposure	10	40	50
	40	60	100

AR = 30/50 – 10/50 = .4 AF = ((30/30+20) – (10/10+40))/(10/10+40) = 2/3 RR E+:E- = (30/30+20) / (10/10+40) = 3 Odds of O+ if E+ = 30/20 = 1.5 Odds of O+ if E- = 10/40 = .25 Odds ratio of E+ versus E- = (30*40)/(10*20) = 6 If the RR doubles to 6 by clockwise rotation to rbind(c(36, 14), c(6, 44)) the OR is 18.857 If the RR doubles to 6 because right shift to rbind(c(24, 26), c(4, 46)) OR is 10.615 PAR = 40/40+60 – 10/10+40 = .2 PAF = 40/40+60 – 10/10+40 / 40/40+60 = 0.5 = using RR (0.5).(3-1) / (0.5).(3-1)+1 = 0.5 = using OR (0.5).(6-1) / (0.5).(6-1)+1 = 0.71

##Probability and sampling Probability is always a feature of populations, not of data; the population is connected to the data by a probability model using assumptions so a sample quantity is an estimator of the estimand, population value 1. Probability of something

happening is 1 2. Probability of nothing happening is 0 3. Probability of a thing is 1-its opposite 4. Probability of at least one of some mutually exclusive things is the sum of their probabilities 5. If event A implies event B then probability of A < probability of B 6. The joint probability of any 2 events is the sum minus their intersection:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Probabilities in the same probability space are independent if $P(A \cap B) = P(A).P(B)$

The conditional probability is obtained by dividing a joint probability by a marginal probability, eg

$$P(x_3|y_2) = \frac{p(x_3, y_2)}{p(y_2)}$$

##Bayes rule and diagnosis Diagnostic likelihood ratio of a positive test (DLR+ve)

$$\frac{P(+|D)}{P(+|D^c)} = \frac{sensitivity}{(1 - specificity)}$$

and DLR-ve

$$\frac{P(-|D)}{P(-|D^c)} = \frac{(1 - sensitivity)}{specificity}$$

Because $P(D) = (1 - P(D^c))$ the complementary Bayes rules for positive predictive value is

$$P(D|+) = \frac{P(+|D).P(D)}{P(+|D).P(D) + P(+|D^c).P(D^c)}$$

```
$$P(D^{c}|+) = \frac {P(+|D^{c}).P(D^{c})}
```

{P(+|D).P(D) + P(+|D^{c}).P(D{c})}$$

```
and dividing one into the other gives the posterior odds being equal to the diagnostic likelihood ratio
```

$$\frac{P(D|+)}{P(D^c|+)} = \frac{P(+|D)}{P(+|D^c)} \times \frac{P(D)}{P(D^c)}$$

Presneill

$$P(B_k|A) = \frac{P(A|B_k)}{\sum_{i=1}^{n}[P(A|B_i).P(B_i)]}.P(B)$$

Chebyshev's inequality

The probability that a variable takes the value of k standard deviations from the mean is a maximum of 1/k squared, for any distribution where the mean has a single value:

$$P(absX - \mu > k\sigma) = \frac{1}{k^2}$$

Beyond subjectivity and objectivity

Gelman and Hennig at the RSS 2017 presented on the false dichotomy that hides bad behaviour by both sides. They propose that Objectivity be replaced by transparency, consensus, impartiality and correspondence to observable reality, and subjectivity replaced by awareness of multiple perspectives and context dependence. Together with stability, these make up a collection of virtues that they think is helpful in discussions of statistical foundations and practice.

Bias is any systematic effect (London:"error") in the design or conduct of a study which results in an incorrect estimate of the association between exposure and outcome. Information bias is present when there is a difference in the accuracy of information collected for exposure or outcome. Reporting bias is due to subjects reporting one with different accuracy, depending on their status in the other (for example reporting exposure differently if they have the outcome, or reporting outcomes differently if they were exposed). Observer bias is due to measurements giving different answers on one depending on the status of the other (for example, the outcome is reported with greater deviation or variation in the exposed, or vice versa) Selection bias is present when there is a systematic difference in the exposure or outcome between those who are observed and those who are not. AKA Collider Stratification bias. The sampling frame may not be representative of the target population The groups may not be comparable (their sampling frames may differ)

Exposure measurement

The true exposure is like True Grail, only seen by indirect evidence of its presence by an Instrument: this is the Measured Exposure. Exposure measurement errors arises from the Instrument's design, its protocols, training or attention or malice or prejudice of those who use it; from the subjects in their memory, cyclic or random or circumstantial variability or recall; and from entering or coding or transforming data. Bias can be caused by differential measurement exposure error. If a parameter X is a proxy measure of the true parameter T, with a fixed bias across a sample b and an additional random error E then $t_i = x_i + b + e_i$, where E(E)=0, E(X) = E(T)-b and so on. If the bias is not fixed, then the regression is not simple but multiple. This can be hard to detect if not suspected on the basis of mechanistic understanding, which couples with unpredictable "test coverage" of regression models to reflect the true situation. Non fixed bias is introduced by differential measurement error, among other things. Precision is given by the correlation of T with X, the validity coeffi-

cient. This all remains a bit theoretical unless the values of T can be known. $\rho^2_{TX} = 1 - \frac{\sigma^2_F}{\sigma^2_X} = \frac{\sigma^2_T}{\sigma^2_X}$, and assuming linear relationships for Y with both T and X, if both $Y = \alpha_T + \beta_T T$ and $Y = \alpha_O + \beta_O T$ then $\beta_O = \rho^2_{TX}\beta_T$ but if X is a function of T rather than an addend, this doesn't hold.

This assumes a linear model, or rather two simultaneous models for T and X. Logistic models for log(OR) can also be built, so that the observed odds ratio depends on the correlation of T and X, tending to 1 as the correlation tends to 0.

$$OR_O = OR_T^{\rho^2_{TX}}$$

The correlation can be expressed separately as sensitivity and specificity of X for T. The odds ratio or sign of the regression coefficient don't cross the null as long as (sens + spec) >= 1, or as long as an exposed case is classified as exposed with greater probability than an unexposed control is classified as exposed.

Scatter plots of data allow clues about linearity or remote or influential points.

Meta Analysis

```
- Provide a data display and objective review
- Give a summary interpretation
- Test an overall hypothesis
- Estimate an average exposure effect
- Assess whether data compatible with the exposure effect being the same in all studies
```

Fixed Effect assumes the effect really is the same and any difference is due to sampling variation: provides a summary using a weighted average of the individual study estimates. Weights are eg 1/variance of the log odds ratio, so the summary OR is

$$\psi_F = \frac{\sum_{i=1}^k w_i\psi_i}{\sum_{i=1}^k w_i}, \pm\frac{1}{\sum_{i=1}^k w_i},$$

which is the variance regenerated from the averaged weights. A Forest Plot represents each study (box size = weight, bars for CI, diamond width = CI of summary estimate) and the heterogeneity between them uses a Chi Squared test $Q = \sum_{i=1}^k w_i(\psi_i - \psi_F)$. If heterogenous then use Random Effect.

Random Effect assumes the true effects vary around an average: the between study variance of the estimates (or alternatively, the residuals around the average of the estimates) is used to modify weights of each study. A summary OR is the same shape as the fixed model summary OR but with between-study-variance-adjusted weights, w*i. These

weights are far less dispersed than the Fixed Effect weights so smaller studies are weighted more heavily and the confidence intervals tend to be wider. So Random Effects models are more conservative.

Likelihood

```
A probability conditioned for a model; or a probability of a model given some data; the likelihood is nu
```

Other points are divided by the MLE to yield the likelihood ratio, which rejects data but adds clarity and comparability. If the likelihood and hence the likelihood ratio curves are normal, so log(likelihood ratio) is quadratic and so easily solved using the pattern $x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$. If it's not standard normal then it's not easy to do and iteration is probably as good as algebra. The quadratic is chosen so that the curvature of the actual log likelihood and the quadratic fitted approximation are identical at the fit point, whether the MLE or the null, depending on whether the aim is to test or derive a confidence interval, or to determine the MLE. Wald tests fit at the MLE.

Also above a LR of 0.1465 lies 95% of the likelihood curve and twice the log of the likelihood ratio is algebraically equivalent to the chi squared distribution, giving a chi squared statistic for every value of the parameter, given the data, under those assumptions. The log(LR) at that point is -0.192, useful for the Gaussian assumption underlying the quadratic approximation: $log(LR) = -\frac{1}{2}\frac{x-\mu}{\sigma\sqrt{n}}$. The fit away from its fitted point is less good, so the choice is to fit at the MLE or at the null. The quadratic is chosen to meet the log LR curve at the MLE and to have the same curvature: log(LR) $= -\frac{1}{2}(\frac{M-\pi_0}{S})^2$ as for D failures in N trials D/N = M

Fuckedly, London uses M not only for D/N but also for log(D/N-D). $S^2 = M(1-M)$, $M = log(\frac{D}{N-D})$ and $S = \sqrt{\frac{1}{D} + \frac{1}{N-D}} = \sqrt{\frac{M(1-M)}{n}}$ $\frac{1}{S^2}$ is called the information of the data: the larger, the greater the curvature of log(LR), and the more precise the estimate and so on. A supported range of parameter values with attendant likelihood over a certain cutoff likelihood can be calculated and as S is the standard error so the LR at the 95% CI limits is $e^{\frac{-1.96^2}{2}} = 0.1465$ if the likelihood curve is normal across the parameter space.

Use likelihood ratio statistic of the value of the log(LR) at the null value of the parameter (or for any 2 models where one is a restricted form of the other)

LRS = -2 log(LR) = -2(lognull – logMLE) = 2(logMLE - lognull) ~ 2 df = (r-1).(c-1) or the Wald statistic of the fitted quadratic approximation to the log(LR) at the null $LRS_{Wald} = (\frac{MLE}{S})^2$, and its square root if H0:log(LR)=0 is $z = \frac{MLE}{S}$

or the Score Test: alternative quadratic matching value, gradient

and curvature at null, not MLE Score test $= -2\log(\mathrm{LR})_{null_{quadfit}} = \frac{U^2}{V}$ where U=gradient and -V = curvature(fitted) at null from which $\chi^2_{MH} = \frac{U^2}{V}$ is derived for the Mantel Haenszel test, $\frac{score^2}{score\ variance}$

Binomial likelihoods

are calculated so the most likely value of p is found for k successes in n trials Likelihood $= \binom{n}{k} p^k (1-p)^{n-k}$, approximated by $-\frac{1}{2}\frac{D/N-\pi}{S}$ with S set as $\sqrt{\frac{p(1-p)}{N}}$ by calculus to match the curvature of the actual likelihood at its maximum, being the sample standard error. the plot of which against p is identical to its ratio, whose log is approximated by the normal assumption: $\log(\mathrm{LR}) = -\frac{1}{2}(\frac{M-\pi_0}{S})^2$, making the Wald test $-2\log(LR) = (\frac{M-\pi_0}{S})^2 \approx \chi^2, \mathrm{df} = 1$

Poisson likelihoods

are calculated so for best guesses at baseline 0 and rate ratio for d0 and d1 events over times T0 and T1 log Likelihood $L = (d_0 + d_1)log(\lambda_0) + d_1 log(\theta) - \lambda_0 T_0 - \theta\lambda_0 T_1 + constant$ which generates a surface maximal at $\lambda_0 = \frac{d_0+d_1}{T_0+T_1}$. Substituting this into the above, L is hence maximal at $0 = d_1 log(\frac{\theta T_1}{T_0}) - (d_0 + d_1)log(1 + \frac{\theta T_1}{T_0}) + constant$ which is the "profile log likelihood for "; its plot against is the same shape as its ratio-to-maximum against ; its plot against log( ) is approximated by: $\log(\mathrm{LR}) = -\frac{1}{2}(\frac{MLE-\theta}{S})^2$ where S is the SE of More complex likelihoods are fitted by iteration on the MLE, eg taking nulls (L=MLE=0) for all n parameters, and working out, calculating gradient and curvature at each value to sketch the best approximation n-surface with its n-dimensional maximum being the improved estimate of MLE and repeating to convergence which doesn't work if the data are insufficient to estimate the number of parameters or if the profile log likelihoods are non-quadratic when eg Poisson, Logistic and Cox regression uses log transformations Or the similar Score Test of form $-2\log(LR) = \frac{U^2}{V}$ where U is the gradient (*aka* the Score) and V the negative of the curvature of a quadratic approximation to the likelihood fitted at the null (aka the Score Variance), not at the MLE.

The plausible values of all but the likelihood ratio test depend on the units of the fitted quadratic. Also, the fitted values don't have a definite integral for anything above a quadratic equation and more complex likelihoods don't have such approximations.

In regression the LRS tests the joint null that all the variables equal their null values and tests any two models where one is a restricted form of the other.

Robustness

Robust objects (estimators, statistics, models…) are ones which are changed little by perturbations in the data; good robust objects retain efficiency while being robust. Efficient objects are ones which need few observations to attain a given performance. Performance is the ability of an object to describe or predict reality, such as to describe variation reliably, reject a false claim or detect a true difference.

Sampling

The study population is the population available for study A sampling frame is the entire extant list of possible units which could be sampled "Equal probability" selections (of each final stage unit): population value is estimated by sample values (self-weighting) eg Simple Random Sampling (and K&S alledges that Systematic Sampling is an EPS…?) Probability proportional to size followed by numerical SRS (eg 4 at each stage 1) SRS at stage 1 followed by proportional SRS (eg 20% of each stage 1) Others need weighting at the analysis stage proportionate to their oversampling Stratified – sample within strata and add together Hierarchical – Primary or First Stage (or Tier) or 1°, within which Second Stage (2°, etc) "Probability Proportional to Size" then SRS with constant Tier 2 sample size if different sized 1° samples, gives equal probability of sampling each 2° unit SRS then SRS with 2° being a constant sampling fraction of 1° if equally sized 1° samples

Balance covariates by stratification (usually logical), blocking (see Fisher and Student's fight), minimisation (re-read Senn 2004), matching (by pairs, read Eldridge 2012). Matching case-control by a variable allows effect modification by that variable to be estimated, and for ignoring the effect of the variable. Can increase precision or efficiency; may add an unknown amount of control for unacknowledged confounders if those are correlated with the matching variable. Matching eg by malaria rates last year, can cause problems. I intuit that if the matched variable is not the right one then no inferential efficiency is gained, eg if "rank of malaria incidence" is very different in the year of study then they're not matched by malaria incidence and random baseline variability may be considered to have been reduced when it actually hasn't. So that's a problem like HTE or undermatched confounders. As it happens they're not very correlated at all, r=0.2 or so. Matching doesn't lose much power unless the matching and exposure variables are strongly correlated; but it can do it. Also overmatching is logistically hard and increases concordant pairs. And if the matched variable is on the causal pathway it underestimates the effect of exposure on outcome. "Break-even values" for the correlation between

matching and outcome variables for matching or stratification to be beneficial. Matched Odds ratios are derived from MH summary odds ratios where only discordant pairs count: if individually matched then each pair is a stratum, if the tabulation is at pair level then two boxes contain zeroes for each concordant pair. Overall numbers of pairs are tabulated for each combination.

Randomisation is distinct from random sampling "In a properly randomised trial any difference between the groups outside that imparted by the intervention are due to chance." "Any difference between the trial groups should be due to the outcome." True randomisation depends on randomness of sequence and allocation concealment.

Clusters Used when there's a direct effect plus indirect (ecological, herd, cooperative, etc) effect on participants, or when it's difficult to allocate individuals. Indirect include, for infections, that which reduces Quantity (herd, reduced carriage, vector death, epidemic periodicity) or Quality (less or more virulent strains, resistance, non-reproducing strains) of infection or the Immunity of victim (immune recency, cross reaction, polyvalency, multi-hit, immune mediated damage). "Direct" + "indirect" = "total"; if all in a cluster participate the effect is "total", otherwise the "overall" effect is the weighted average of direct + indirect on participants and indirect on non-participants. Individually allocated trials measure the Direct effect; cluster trials measure the Overall which depends on participation fraction and HTE combined with differential participation. Second stage randomisation of individuals within clusters clarifies direct v indirect. Spin the bottle: random direction to walk from centre to border of cluster Used by the Expanded Programme on Immunisation, EPI of the WHO as 30 clusters of 7 randomly select units along the line of walk to be cluster centres clusters are "next nearest until quota of 7 children reached", no callbacks Not probability based ("not self weighting") and In vaccine surveys mothers interviewed, Est is +- 10% at 95%CI assuming Design Effect 2 Probability of choosing a child is awfully roughly $P_i = m \times \frac{M_i}{M} \times \frac{n}{N_i}$ Where n is the very approximately equal number of children surveyed per cluster, Ni is total children in the ith cluster and others as below and $\hat{V}ar(\hat{R}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} (y_i/n - \hat{R})^2$ so CI is $\hat{R} \pm t_{1-\alpha,m-1}\sqrt{\hat{V}(\hat{R})}$ and the bias is $\bar{\rho} - R = -\rho_{p,g}\sqrt{\frac{V(p)V(g)}{\bar{g}}}$ where g is the proportional change in population since the census, p is the true vaccinated proportion, rho is the correlation between p and g, C is the number of clusters in the whole population and R is the true ratio $R = \sum_{i=1}^{C} \frac{g_i p_i}{C\bar{g}}$.

Design effect for clusters all of the same size is $1 + ((b-1) \times ICC)$ where b is the number per cluster and ICC is the correlation coefficient within clusters, as an average across all the clusters.

Compact segment sampling;

another cluster method. Areas are initially chosen with PPS, eg from last census The areas are divided into equal numbers of segments Within each segment of a given area are an equal number of houses, sketched and recorded A segment is randomly chosen from each area, and all houses in the segment are sampled Probability of choosing a house in the ith segment and therefore a child is very roughly $P_i = \frac{m}{S_i} \times \frac{M_i}{M_{tot}}$ where m is the number of selected clusters, Si is the number of segments in the ith cluster, Mtot the census population of all clusters in the sampling frame, Mi census population in the ith cluster IF population has changed uniformly and segments are exactly chosen eg vaccine coverage is estimated by a ratio of two random variables number vaccinated and number chosen, each of which is weighted for the probability of being chosen, which therefore has to be known to estimate the true values.

$$\hat{Var}(\hat{R}) = (\frac{m}{(m-1)\hat{N}^2}) \sum_{i=1}^{m} \frac{(y_i - n_i\hat{R})^2}{P_i^2}$$

$$\hat{R} = \frac{\sum_{i=1}^{m} y_i/P_i}{\sum_{i=1}^{m} n_i/P_i}$$

IntraClass Correlation Coefficient (ICC or) is Between Cluster Variance/Total Variance $\rho = \frac{\sigma_c^2}{\sigma_c^2+\sigma_e^2}$, also $\rho = \frac{\sigma_c^2}{\pi(1-\pi)}$, or $\rho = \frac{\sigma_B^2}{\sigma_B^2+\sigma_W^2}$ where c is the between cluster SD and e is everything else (here, within-cluster variation) ICC can be used for means, proportions and counts but is not defined for rates. k, the Between-Cluster Coefficient of Variation, is SD for cluster means/mean = $\frac{\sigma_c}{x} = \frac{\sigma_c}{\pi}$ so

$$\rho = \frac{k^2\pi}{1-\pi}$$

Design Factor is SE for design/SE under simple random, terminology analogous to Error Factor Design Effect DEFF = Sample size / size by simple random sampling or Variance/Variance by SRS = Design Factor squared if it's proportions (Also quoted as SE by design / SE by SRS!) = 1 + (Number in each cluster-1) * rate of homogeneity $= 1 + \rho(m-1)$ DEFF on total N is $1 + (\bar{m}-1)\rho$ where m is number in a cluster or, if cluster sizes vary, $1 + \rho(m(1 + (\frac{sd_m}{m})^2) - 1)$ where $\frac{sd_m}{\bar{m}}$ is coefficient of variation of cluster size rate of homogeneity **roh** (here) is the ICC for single stage clustering, and an equivalent ration of total variance for multi stage clustering k is estimated for power reasons from c : see below or $m = \frac{N(1-\rho)}{(c-\rho N)}$ for cluster size given total N under simple random and number of clusters c ICC can't be used when the outcome is a rate per time; so k is used here.

Standard cluster sample size calculations

For two rates:

$$n = \frac{(z_{1-\beta} + z_{1-\alpha/2})^2.(\lambda_1 + \lambda_2)}{(\lambda_1 - \lambda_2)^2}$$

$$\text{As } \mathrm{E}(s^2) = \lambda Av(1/y_j) + \sigma_c^2 = \lambda Av(1/y_j) + k^2\lambda^2 \text{ so}$$

$$\hat{\sigma}_c^2 = s^2 - rAv(1/y_j)$$

and

$$\hat{k} = \frac{\hat{\sigma}_c}{r}$$

For two means:

$$n = \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2}{d^2}, d = \frac{\delta}{\sigma}$$

or

$$\frac{(u+v)^2(\sigma_1^2 + \sigma_2^2)}{\delta^2}$$

or

```
As $E(s^2) = \sigma^2 Av(1/n_{j}) + \sigma_c^2$, so $\hat \sigma_c^2 = s^2 - \hat \sigma^2 Av(1/n_j
```

For two proportions,

$$n = \frac{2[z_{1-\alpha/2}\sqrt{2q(1-q)} + z_{1-\beta}\sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)}]^2}{(\pi_1 - \pi_2)^2}, q = \frac{\pi_1 + \pi_2}{2}$$

or

$$n = \frac{[u\sqrt{\pi_1(1-\pi_1) + \pi_2(1-\pi_2)} + v\sqrt{2\bar{\pi}(1-\bar{\pi})}]^2}{(\pi_1 - \pi_2)^2}$$

As $E(s^2) = \pi(1-\pi)Av(1/n_j) + \sigma_c^2$ so $\hat{\sigma}_c^2 = s^2 - p(1-p)Av(1/n_j)$ and $\hat{k} = \frac{\hat{\sigma}_c}{p}$ where p is the combined proportion across all clusters combined These look simpler but sacrifice transparency if written for the number of clusters c for rates

$$c = 1 + f\frac{[\frac{\lambda_0+\lambda_1}{y} + k^2(\lambda_0^2 + \lambda_1^2)]}{(\lambda_0 - \lambda_1)^2}$$

or for proportions

$$c = 1 + f\frac{[\frac{\pi_0(1-\pi_0)+\pi_1(1-\pi_1)}{m} + k^2(\pi_0^2 + \pi_1^2)]}{(\pi_0 - \pi_1)^2}$$

```
where f is a factor combining z  and z , eg f=7.84 for power 0.8 and alpha 0.05, 10.5 for power 0.9 and
```

If SRS but unequal randomisation ratios with size in each group n, then if the smaller group is n1, then alter the size of each group by $k = \frac{n_2}{n_1}$ and $n_1 = \frac{n(k+1)}{2k}$, which converges on $n_1 = \frac{n}{2}$ as the ratio k increases, at the cost of imprecision in estimates of the effect in group 1, and less information on rare events.

Q1. Was there an effect of treatment in this trial? Q2. What was the average effect of treatment in this trial? Q3. Was the treatment effect identical for all patients in the trial? Q4. What was the effect of treatment for different subgroups of patients? Q5. What will be the effect of treatment when used more generally (outside of the trial)?

Because few clusters tend to be randomised and the power calculations nastily fudged, baseline imbalance is common, with instability of the effect size estimates and hence increased false claims in both directions. Balancing of baseline covariates is often thought to limit this. Balance is achieved by matching (add 2 instead of 1 to the RHS of the simplifed equation, k is the average coefficient of variation only between matched pairs), stratifiction or restriction / constraint (reject all random allocations until one is generated with eg 10% difference in any covariate per arm). Among these only matching is random, depends on the matching variables being well chosen and doesn't permit analysis of the main effects of any matching variable.

The estimates are either calculated over the whole study, or combined from cluster-specific estimates. This is not straighforward : if equal weighting of cluster estimates is desired and the clusters were not randomly selected from a well defined target population then cluster estimates are better if they were randomly selected with PPS then the overall estimate is a consistent estimator of the true population value, and is of course easier but it doesn't allow simple t tests.

To derive the standard error the variance of the risk or, interchangeably, the rate ratio, for cluster specific estimates is estimated roughly by working in the log RR scale $Var(\log RR) = Var(\log R_1) + Var(\log R_0) \approx \frac{Var(R_1)}{R_1^2} + \frac{Var(R_0)}{R_0^2}$, and $Var(R_1) \approx \frac{s_1^2}{c_1}$, where s1 is the observed SD of the cluster specific estimates across the number c of clusters in arm 1 (the intervention, for example). So the estimate accounting for clustering is log RR $\pm$ $1.96\sqrt{Var(\log \text{RR})}$. A simple T or Wilcoxon rank sum test can be used for the summary estimates using these summary values.

Cluster level analysis

Adjusted estimates of the effect of treatment allocation rely on the differences (residuals) between a model that accounts for all important-seeming covariates (that is, all except for cluster effects and the effect of treatment allocation). These residuals will be randomly distributed across clusters under the null hypothesis, which is then tested by analysing them in place of the cluster level raw estimates. For example, a poisson regression provides estimates; residuals are calculated as the ratios of rate in each cluster to the predicted rate for those individuals under the model; the mean and SD of residuals are calculated

for each allocation group; these provide a T test (or equivalent) and the ratio of mean residuals is the estimate of effect size, with Var(RR) calculated analogously to the above, this time for the ratio of residuals.

Individual level analysis With linear models, assume the data are all iid but that there is a cluster effect (data are still independent within a cluster but are all drawn from the same independent distribution). Use poisson for rates and a gamma error, giving a negative binomial likelihood. Normal for risk and a normal error, giving a normal likelihood. Binomial for logit and normal error, giving a non analytic likelihood which is converted to a log which is approximated by a log normal using the minimum of a quadratic equation to pass through a certain number of points representing the data. This last might be unstable on changing the number of points to fit, if the approximation likelihood is not a good reflection of the whole data. This is revealed by a quadrature check with >1% difference in likelihood on changing the number of quadrature points; this is actually probably not necessary now with adaptive quadrature being used at least in Stata. If that's the case:

GEE assume that any two points from separate clusters are uncorrelated, but within a cluster all points are identically correlated with each other, represented by an exchangeable correlation matrix derived from the data. This estimate of rho is the primary output; then the regression coefficients and their standard errrors are estimated from it in turn. The regression coefficient is a population average odds ratio, assuming no cluster level random effect but only correlation… I thiink. Overall the buzz words are exchangeable correlation matrices and robust standard errors, with a Wald test $z = \frac{\beta}{\mathrm{SE}(\beta)}$.

Cluster level analysis with a t test is robust and has good coverage even with low cluster numbers (15 or fewer) but can't adjust well. Linear models are efficient, GEE robust. With binary data they differ: random effects models estimate cluster specific - averaged odds ratios and GEE estimate population averaged odds ratios. With 15 clusters in total you can estimate continuous or rate coefficients, with 30 their standard error; or 30 and 50 respectively for binary outcomes.

GEE inflate Type 1 error if few clusters; and vague hand waving about linear modelling's distributional assumptions not being easy to support if there are few clusters.

Moving field: Bayesian, restricted maximum likelihood.

Distributions

Probability Mass Function is >0 everywhere and the sum of the individual probabilities of possible values adds up to 1

Probability Density Function is >0 everywhere and has area =1 under it

```
The area under the pdf corresponds to the probability for those values of that random variable
The probability of a single specific number is 0
Integrating the CDF yields the CDF
```

Cumulative Distribution Function is the probability that a random variable takes the value x

$$F(x) = P(X \leqslant x)$$

Integrating the CDF at its limits yields an area of probability Survivor function is S(x) = 1-F(x) Quantiles: the th quantile of a distribution function F(x) is the point such that $F(x_\alpha) = \alpha$

Bernoulli mass function

$$p(X = x) = p^x.(1-p)^{1-x}$$

$$p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

(remembering that $\binom{n}{0} = \binom{n}{n} = 1$)

#Normal distribution Any normal distribution is a multiple of and an addend of from Z, the Standard Normal Distribution:

$$Z = \frac{x-\mu}{\sigma}, N \sim (0, 1)$$

So if X is a normally distributed variable with mean and sd then $X = \mu + \sigma Z, N \sim (\mu, \sigma^2)$

```
65%, 95% and 99% of the SND lie within 1, 2 and 3 SD of
```

Multinomial distribution is

$$P(x_1, \ldots x_n; k) = \frac{k!}{x_1! x_2! \ldots x_n!} . p_1^{x_1} \ldots p_n^{x_n}$$

```
of which a special case is the binomial distribution
    $$P(x, k) = \frac{k!}{x!(k-x)!} \times p^{x}.q^{k-x}$$
```

T distribution is descended from the standard normal, and assumed to be symmetric and centred on 0 so with only 1 parameter, the degrees of freedom. Skew can be exponentially transformed or another distribution used.

The MH odds ratio across strata

is the weighted mean sum of odds ratios across strata generally $OR_{MH} = \frac{\sum(w_i OR_i)}{w_i}$ weights $w_i = \frac{d_{0i}h_{1i}}{n_i}$ denominator of OR divided by the total or ORMH = Q/R where Q = $\sum \frac{d_{1i}h_{0i}}{n_i}$ and R = $\sum \frac{d_{0i}h_{1i}}{n_i}$ then se(log(ORMH)) = $\sqrt{\frac{V}{QR}}$, V = $\sum V_i = \sum \frac{d_i h_i n_{0i} n_{1i}}{n_i^2(n_i - 1)}$, V calculated from the marginal totals because each stratum has equal variance Chi squared is $\chi^2_{MH} = \frac{(\sum d_{1i} - \sum E_{1i})^2}{\sum V_i} = \frac{(O-E)^2}{V} = \frac{U^2}{V}$, df=1,

```
$E_{1i}= \frac{d_i n_{1i}}{n_iO} = \sum d_{1i}, \> E = \sum E_{1i} \textrm{, U = O-E}$
MH chi squared is (n-1)/n times the size of the non MH chi squared in the simple 2x2 table.
```

Rule of 5 to check validity: subtrahends of `c(sum(min(c(di, n1i))) , sum(max(c(0, (n1i - hi))))) - sum(Ei) > 5` Chi squared for heterogeneity tests H0: for all i, ORi = ORMH , so

$$\chi^2_{het} = \sum \frac{(d_{1i}h_{0i} - OR_{MH}d_{0i}h_{1i})^2}{OR_{MH}V_i n_i^2}$$

Poisson has MEAN AND VARIANCE EQUAL to

which is events / time, $\lambda = \frac{D}{t}$ and noting $1 - x \approx e^{-x}$ so $(1-x)^n \approx e^{-nx}$ its probability mass function is:

$$P(X = x; \lambda) = \frac{\lambda^x . e^{-\lambda}}{x!}$$

where x is a non negative integer and Risk = $1 - e^{-\lambda t}$

```
Standard Error of a rate is SE(number of events)/person-time at risk
$$SE_{events} = \sqrt{D}, SE_{risk} = \frac{\sqrt D}{t}$$
         and $\lambda = \frac{D}{t}$ so $SE = \sqrt{\frac{\lambda}{t}}$
SE for log rate is $\frac{1}{\sqrt{D}}$, CI is $\lambda^{\times}_{\div} EF$, EF=$e^{\frac{1.96}{\sqrt D

Used for example in modelling especially unbounded count data,
```

$X \sim Poisson(\lambda t)$ where $\lambda = E[X/t]$ approximating the binomial for large n and small p let =np modelling contingency tables

in censoring survival models (as an "unbounded" problem). NHST is the Wald against H0: 0 = 1, $z = \frac{\log(rate\ ratio)}{s.e.\log(rate\ ratio)}$

Mantel-Haenszel uses RR in stratum i weighted for total time in the stratum,

```
$$RR_{MH} = \frac{\sum{\frac{d_{1i} \times T_{0i}}{T_i}}}{\sum{\frac{d_{0i} \times T_{1i}}{T_i}}} = \fra
or if $w_i = \frac{d_{0i}T_{1i}}{T_i}$ then
$$RR_{MH} = \frac{\sum(w_i \times RR_i)}{\sum w_i}$$

 $se_{MH} = \sqrt{\frac{V}{QR}}$ where $V = \sum V_i$ and $V_i = \frac{d_i T_{1i} T_{0i}}{T_i^2}$ and NI
```

```
 $\chi^2_{MH} = \frac{(\sum d_{1i} - \sum E_{1i})^2}{\sum V_i}$ or $\frac{(O-E)^2}{V}$, df = 1

where Exi is overall stratumi rate times observation time for exposure x, $E_{1i} = \frac{d_i \times T_
```

To detect effect modification the Chi Square test for heterogeneity $H_0 : RR_i = RR_{MH}$

```
$$\chi^2_{het} = \sum \frac{(d_{1i} T_{0i} - RR_{MH} d_{01} T_{1i})^2}{RR_{MH} V_i T_i^2} \textrm{, df =
```

Cohort E(D1) = Y1 *(D/T) and U = D1 – E(D1)

Var(U) = $D\frac{t_1}{t}(1-\frac{t_1}{t})$ and test statistic $z = \frac{U}{SE(U)} = \frac{U}{\sqrt{Var(U)}} or Z^2 = \frac{U^2}{V}$, Chi square df=1

Poisson regression is very similarly shaped to other regressions: rate = constant x timeband x exposure. Under a poisson distribution the lambda is equal across strata; if it's not and especially if there's a linear trend, there is overdispersion. If overdispersion is suspected by qualitative examination of rates, or by the ratio of LRS:df being greater than 1, then ~~an adult must be asked for help~~ a value (constant or distributed) can be added to the regression model to account for the departure from poisson. The value is called the frailty and the resulting model is a negative binomial model.

Cox regression

limits each timeband to that containing one event (call it a timeclick). So rate = Changing baseline * exposure: the model assumes the exposure effect is constant (the proportional hazard assumption). Test this graphically or by taking logs of the cumulative hazard so that the "Nelson-Aalen plot" of $\sum \lambda_{ti} = \sum \lambda_{t0} \times \theta_i$ is parallel against time, and test by fitting interaction terms of time (by arbitrary epochs or as a superimposed linear trend) and using LRT or something.

Multivariate distributions

MV norm has mean as a vector, sigma is the variance / covariance matrix

These have parameters of location, scale and skew. Normal location is the mean, t location is the non centrality parameter which are a scalar for univariate and an n-element vector for n-dimensional distributions (eg an n-variate distribution of densities). Scale for a univariate is the variance for univariate normal and t distributions, but is a dispersion matrix sigma of n by n elements for for n-variate distributions (it's the variance / covariance matrix). The skew parameter is a vector of length n for n-variate distributions.

MV t has degrees of freedom usually presumed identical across al variates, a parameter for centrality delta, and a variance / covariance matrix sigma

density by `dmvt(x, delta=c(i,j)`, sigma= `as.matrix(r, t, y, u), df=df, log=T)`

Bivariate has delta=0 if standard and sigma=diag(2), ie rbind(c(2, 0, 0, 2))

Survival and logrank test

Kaplan Meier survival probability is the cumulative probability of an entity that is initially at risk surviving subsequent epochs until the one in question, each epoch being "data-defined" and usually as containing one event: as Pevent = nevents/nat-risk – (0.5*ncensored) if the censoring time is not defined for the interval P(Si) = (1-Pevent1).(1-Pevent2)…(1-Peventi) $P(S_x) = \prod_{i=1}^{x}(1-\frac{\text{events}}{\text{at risk}-0.5\times\text{censored}})_i$, or $\widehat{S}(t) = \prod_{i:\, t_i \le t}\left(1-\frac{d_i}{n_i}\right)$, A good KM curve has ticks on the line for censoring occurrences and nat-risk below the line Log Rank test is a Chi Squared, with Observed totals and Expected E = (row total x column total)/(grand total) summed across all intervals for each group, and added, giving the statistic: $\chi^2_{logrank} = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2}$, df=1 or $Z = \frac{\sum_{j=1}^{J}(O_{1j}-E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}}$ Not sure whether you could substitute P(Sx) times number at risk at t=0 for E.

The hazard function h(t) is the failure rate summed over ever smaller time intervals, which may be more than 1, hence is not a probability. This depends on a failure function, F(t), which is a probability, being a cumulative distribution function. The P(Si) above is equivalent to the reliability function R(t), also called the survival function S(t), the probability of "no failure" as a function of time. R(t) = 1 – F(t). Cumulative failure as a function of time, F(t) is the integral over time of the probability density of failure as a function of time, f(t).

$\tilde{H}(t) = \sum_{t_i \le t} \frac{d_i}{n_i}$

h(t) = f(t) / 1-F(t) = f(t) / R(t). $h(t) = \lim_{\Delta t \to 0} \frac{R(t)-R(t+\Delta t)}{\Delta t \cdot R(t)}$, which does not have to be parametric as long as there is a definite cumulative distribution.

If the failure density is modelled as exponential (ie Poisson(1), that is the time until first failure is Poisson distributed and other failures are presumed to be independent in every case including those already failed) then the familiar result is that $f(t) = \lambda^1 e^{-\lambda t}$ so $F(t) = \int_0^t \lambda e^{-\lambda t} dt$ then the hazard function h(t) is $h(t) = \frac{f(t)}{R(t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$, which is time-blind / memoryless / constant with respect to time Assuming, then, that the hazard function is constant and the cumula-

tive hazard rises linearly: - Risk = events/number at risk; - hazard = (events/time at risk in the limit as t goes to 0) = lambda; - survival as a function of time S(t) = product of 1-risk across times; cumulative hazard = sum of risks for all times (Nelson-Aalen estimate of the cumulative hazard).

The log of the hazard ratios for two exposures is log(HR1) – log(HR2) = log(constant) H(t) = -log(S(T)) or S(t) = e^(-H(t)) S(t) = e^(-lambda t) H(t) = lambda t Risk up to time t = 1- e^(-lambda t) Average survival time = 1/lambda

ARIMA The error term ϵ in the usual linear model $Y_i = \alpha + \beta X_i + \epsilon_i$ is assumed to be composed of errors which are independent, identically distributed (usually normally distributed with common variance). This is the definition of white noise.

If the value of a variable at time t is either static or at least predicted mainly by its value at time t-1, then it's a time series. Some function of time ϕX_t, added to an error term, gives the Autoregression formula:

$$X_t = \phi X_{t-1} + \epsilon_t$$

These assume errors are white noise; but usually errors are correlated with the previous error, just as the underlying true value is autocorrelated. So now the model describes an autoreggressive process with autocorrelated errors:

$$\epsilon_t = W_t + \theta W_{t-1}$$

The autocorrelation magnitude can be estimated by comparing values that are a constant distance apart; lag 1 correlation is contiguous pairs, lag 2 is pairs separated by one intervening variable and so on. A series is stationary over a time, if the mean over that time is always constant. Some fiddling can make it stationary over a trend (so if there is an underlying spurious rise in price due to inflation). Examples of stationary processes are linear predictors where X_t is stationary or a random walk, where W is white noise added to the last value of X to produce the current value of X:

$$X_t = X_{t-1} + W_i$$

This is demonstrated by differencing in R, eg for the default lag 1

```
diff(x, lag = 1)
```

Correlation

categorical variance is by tabulation, eg mean pair agreement or Observed Agreement OA:

```
or Kappa statistic for excess agreement over max chance agreement $\kappa = \frac{OA-CA}{1-CA}$,

(where CA is $\frac {E[a]+E[d]}{a+b+c+d}$ and $E[a] = \frac{\textrm{Row total} \times \textrm{Column to
```

continuous is partitioned into systematic and random error eg by ANOVA; their ratio is the Intraclass Correlation Coefficient, aka reliability coefficient eg by Pearson's correlation of product-moment where for a population

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X.\sigma_Y}$$

, where

$$cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$

and using the identities

$$\sigma_X^2 = E[(X-E[X])^2] = E[X]^2 - E[X^2]$$

and

```
$$E[(X-\mu_{X}).(Y-\mu_{Y})] = E[(X-E[X]).(Y-E[Y])]$$, so
$$\rho_{(X, Y)} = \frac{E[XY]-E[X]E[Y]} {\sqrt{E[X]^2 - E[X^2]} \sqrt{E[Y]^2-E[Y^2]} }$$


then the sample correlation r, assuming
        a normal distribution of y for each xi
        a uniform variance for y across all x
        y is monotonic on x
    is the value that describes the closeness of all points to a linear relationship
    is useful for a description of the scatter about the least squares linear relationship
    is the number of standard deviations that y changes for a 1 SD change in x
        given by Pearson's Correlation Coefficient:

                $$r = r_{xy} = \frac{\sum_{i=1}^n (x_{i} - \bar{x})(y_{i} - \bar{y})}{\sqrt{\sum_{i=1}^
```

```
    can also be quoted as the mean of the standard scores of x and y:
```

$r = r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (\frac{x_i - x}{s_x})(\frac{y_i - y}{s_y})$, and $\frac{\sum x_i y_i - nxy}{n s_x s_y}$

=

$$\frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

, or

$$r_{xy} = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sqrt{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2} \sqrt{\sum y_i^2 - \frac{1}{n} (\sum y_i)^2}}$$

r squared explains the proportion of variance that is explained by the straight line, For that straight line the regression coefficient for

the straight line $y = \beta_0 x + \beta_1 x + \gamma x + \epsilon$ is $\beta_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ or $\hat{\beta}_1 = Cor(Y, X)\frac{sd_y}{sd_x}$

And Spearman's is Pearson's rho of the *ranks* of the values, rather than the values themselves.

Hypothesis Testing

`The confidence interval is the set of observations for which H0 is rejected, that is an inversion of the`
`Often overlooked is that evidence for a hypothesis is always relative to that for another hypothesis; o`

Correlation of parameters within a subgroup implies that, when viewed at the level of the population, there is some component of variation that is attributable to membership of that subgroup. A neater way to say this is that within cluster correlation measures the same phenomenon as between cluster variation. The information is therefore reduced in comparison to i.i.d observations. Robust standard errors, GEE and multilevel modelling are ways to address this.

Robust standard errors Robust variance is proportional to the sum of the squared residuals (residuals being the difference between the observed values and those predicted by the model). They can be calculated individually if the data are i.i.d or calculated for each cluster then summed, where there is correlation. This still assumes independence between clusters. Robust standard errors do not affect the calculation of the maximum likelihood and so a LRS and test are not valid and the MLE is not altered. A quadratic approximation is not used so the standard errors are not excessively narrow; the cluster level residuals are used to generate the variance instead. "Robust SE are correct provided the model is correct" and there are >30 clusters.

Generalised Estimating Equations GEE take account of correlation in the calculation of the effect estimate as well as the standard error. The correlation matrix structure is assumed: independent (iid), exchangeable (observations within a cluster are equally correlated and are not correlated with those outside the cluster) or autocorrelated (the same observation at different times, for example). Exchangeable correlation is the usual GEE setting, where the effect estimate is a weighted combination of the effect estimates in each cluster and the standard error is derived from the residuals outside the model. A population average effect is estimated: the average odds of an outcome among those exposed, divided by the average odds among those not exposed.

Multilevel Modelling (Random Effects models) The effect of cluster, whether innate, unmeasured confounder or a true mechanistic effect, is represented in the regression model by a separate term for each cluster: this is rendered tractable by assuming a stochastic distribu-

tion of the size of these terms rather than attempting to measure or specify the cluster effect for each cluster (hence "random effects"): the log odds of the jth person in the ith cluster is then given by , where for example ui ~ N(0, u2) and u2 is estimated as part of the model building. The same information is imparted by estimating rho, the within cluster (also called "intra-class") correlation coefficient. A cluster specific effect is estimated: the odds of an individual having the outcome if exposed, divided by the odds if not exposed. The intention is to derive a model that provides a fully specified likelihood: assuming that it has done so the LRT is appropriate. That likelihood becomes complicated easily: a bivariate normal distribution for the likelihood is produced for a normal error plus normally distributed cluster effect, which has an algebraic solution, as does the combination of poisson error and gamma cluster effects to produce a negative binomial; but the combination of binomial error and normal cluster effects, or poisson error and normal cluster effects, produce mixture distributions without roots. The reliability of estimates needs to be checked.

Power - is 1-beta and depends on the assumptions of the population structure - is a function that depends on the value of mu0 - the distributions of the observation under H0 versus Ha hypotheses are calculated - then a line is drawn across the null for alpha; - at that line the extreme tail under Ha is 1- = P

If testing Ha: > 0, $1 - \beta = P(X > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}; \mu = \mu_a)$, where $\bar{X} \sim N(\mu_a, \sigma^2/n)$

and 0, are known. If 3 of the 4 unknowns a, , n, are specified the last can be calculated

If assuming only the noncentral t distribution then power is $P = (\frac{X-\mu_0}{s/\sqrt{n}} > t_{1-\alpha,n-1}; \mu = \mu_a)$; note P= at a= 0 ; On the other hand power depends only on the difference in means divided by S.E.M $\frac{\sqrt{n}(\mu_a-\mu_0)}{\sigma}$, and the effect size is dimensionless and can be used across contexts a bit.

Multiple testing

	$\beta = 0$	$\beta \neq 0$	Hypotheses
Claim $\beta = 0$	U	T	m-R
Claim $\beta \neq 0$	V	S	R
Claims	m_0	$m - m_0$	m

Control of false positives involves an error measure and a correction for it. There are many such systems, their use depending on the needs of the end user of the data.

False positive rate $E[\frac{V}{m_0}]$ is the rate at which false results are called positive Family wise error rate $P(V \geq 1)$ which converges to 1 for

multiple tests even as alpha is fixed controlled by Bonferroni: $\alpha_{fwer} = \frac{\alpha}{m}$ (very conservative) or calculate p-values, order them as P1 … Pm with their null hypotheses H1 … Hm and either -(Holm, stable) let R be the smallest k such that $P_{(k)} > \frac{\alpha}{m+1-k}$, reject H1 to HR-1 or -(Hochberg, powerful, assumes positive dependence eg in reusing data) let R be the largest k such that $P_{(k)} \leq \frac{\alpha}{m+1-k}$, reject H1 to HR

False discovery rate is the rate at which claims of significance are false. controlled so that E($\frac{V}{R}$) is below a chosen threshold q: Benjamini-Hochberg order P1 … Pm with their null hypotheses H1 … Hm and for the largest k such that $P_k \leq \alpha.\frac{k}{m}$, reject Hi where i= 1 … k which is useful because $E[Q] \leq \frac{m_0}{m}\alpha \leq \alpha$ Or you could report "adjusted p-values" which are no longer p-values eg to control FWER for P1 … Pm take $P_i^{fwer} = \max m \times P_i, 1$ and call each Pi< significant

Expected values - are properties of distributions, as are variances of those distributions - the population value is estimated by the relevant sample value

E(x) is linear, so E(cx) = E.c(x) and E(aX+bY) = a.E(X)+b.E(Y) The centre of mass from a probability mass function of discrete data gives that expected value $E[X] = \sum_x x.p(x)$, extended to a sample mean where $\bar{X} = \sum_{i=0}^{n} x_i.p(x_i)$, where p is independent and identically distributed and is equal to $p = \frac{1}{n}$ for every x.

For continuous variables it's an area under the function t.f(t) where f(t) is the variable's PDF $E[X] = \int t.f(t)$

Given a normally distributed continuous variable X the population variance is

$$Var[X] = E([X - \mu]^2) = E[X^2] - E[X]^2$$

And because binary categorical random variables have expected value

$$E[\bar{X}] = (1 - p).x^c + p.x$$

then if x is coded as 0 if a variable is not present, eg tails, and 1 if present, eg heads, so $E[\bar{X}] = (1 - p) \times 0 + p \times 1 = p$ and $E[X^2] = (1 - p) \times 0^2 + p \times 1^2 = p$ then $E[Var(X)] = p - p^2 = p(1 - p)$ and also from the definition of Var[X] $Var[aX] = a^2.Var[X]$

The sample variance is sort of analogous as $s^2 = \frac{\sum_{i=1}(X_i - X)^2}{n-1}$ and $\frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1}$

THE SAMPLE VARIANCE ALSO HAS A DISTRIBUTION relating to the population of sample variances from which it is drawn: it is an unbiased estimator, meaning the mean of sample variances is an estimate of the population variance; and the variance of sample means is an estimate of the population variance, scaled for the sample size. I think this is incredibly important. The variance of the sample mean is the population variance divided by n $Var[\bar{X}] = \frac{\sigma^2}{n}$, so s.e.m. $= \sigma\sqrt{\frac{1}{n}}$, because $Var[\bar{X}] = Var[\frac{1}{n}.\sum X_i] = (\frac{1}{n^2}).Var[\sum X_i] = (\frac{1}{n^2}).\sum \sigma^2 = \frac{\sigma^2}{n}$

The SE for the mean of a log transformed variable, because $log(X) \simeq log(\mu) + (X - \mu)(log'(\mu))$ and $log'(\mu) = \frac{1}{\mu}$, is $SE(log(X)) \simeq SE(X)log'(\mu) = \frac{SE(X)}{\mu}$

For log proportions, using $\frac{SE(X)}{\mu}$, $SE(log(p)) \simeq \frac{\sqrt{\frac{(p(1-p)}{n}})}{d/n} = \sqrt{\frac{1}{d} - \frac{1}{n}}$ so for log risk ratio, $RR = \frac{p_1}{p_2}$ so $\log(\text{RR}) = \log(p_1) - \log(p_2)$ and $\text{SE}(\log(\text{RR})) = \sqrt{Var(\log(p_1)) + Var(\log(p_2))} = \sqrt{\frac{1}{d_1} - \frac{1}{n_1} + \frac{1}{d_2} - \frac{1}{n_2}}$

For log rates $\text{SE}(\log(\lambda)) = \frac{1}{\sqrt{D}}$ - The log of a product is the sum of the logs - The sum of the logs is the log of the products - The log of a quotient is the difference of the logs - The difference of the logs is the log of the quotient - The exponent on the argument is the coefficient of the log - The coefficient of the log is the exponent on the argument

Bootstraps Efron and Tibshinari. Permits confidence intervals without very complex maths. Approximate the sampling distribution of a statistic using the distribution implied by the data. Use the data you have and sample within it, with replacement, B times. Calculate the statistic you need to estimate for each new sample you have made. The resampling distribution non-parametrically approximates the population distribution, so the standard deviation of it implies the standard error of the median, confidence intervals can be generated etc. Bias Corrected and Accelerated (BCA) interval performs MUCH better than the raw CI; use bootstrap package to construct it

Permutation tests Powerful and manifold: eg Rank sum test permutes ranks for the rank sum, Fisher's Exact test permutes binary groups for a hypergeometric probability, an ordinary permutation test permutes the raw data. Randomisation tests eg if matched data have differences with signs randomly reassigned the signed rank test is the result; or a regressor of interest might be permuted for regression testing. When grouped / stratified data are compared the labels are irrelevant for forming H0. So take grouped data, calculate some comparative statistic. Reassort the group labels randomly across the data; calculate a statistic for each of these random bins. Repeat many times; this results in a distribution of statistics Observe number of times these differences are more extreme than the initially observed data.

#Asymptotics Really very useful topic: the behaviour of functions as the size nears infinity, in this case the behaviour of estimators as the sample size approaches infinity; eg Law of large numbers: sample averages converge on the population average as n increases. Central Limit Theorem: "the distribution of averages of i.i.d variables (properly normalised) becomes that of a standard normal as the sample size increases" So the average of samples, and their variance, converge on population values. The distribution also becomes more normal-like as