Inference re Epidemiologic Parameter: **Prevalence or Risk (proportion)** [1]

Theoretical: P or π or p [2]
Empirical: $\hat{P}$ or $\hat{\pi}$ or $\hat{p} = n_+/n = y/n$; $\hat{q} = 1 - \hat{p}$ [3]
Model: $y \sim Binomial(n, p)$ [4]

P-value: Exact [5]
- $P[\ y \geq y_{obs} \mid p_{null}\ ]$ (1-tail) or [6]
  $(1/2)P[\ y = y_{obs}] + P[\ y > y_{obs}]$ ("mid-P" version) [7]

Approx. [8]
- Normal Approx. to distr'n of $\hat{p}$, or transform $t(\hat{p})$ of $\hat{p}$ [9]

CI: Exact [10]
- $p_L : P[y \geq y_{obs} | p_L] = \alpha/2;\ p_U : P[y \leq y_{obs} | p_U] = \alpha/2$ [11]

Approx. [12]
- reverse transform of $\{t(\hat{p}) \pm z_{a/2} \times SE[t(\hat{p})]\}$ [13]

transform	$t(\hat{p})$	$SE[t(\hat{p})] = Var^{1/2}$	reverse	[14]
identity	$\hat{p}$	$\{\hat{p}\hat{q}/n\}^{1/2}$	n/a	[15]
log	$\log[\hat{p}]$	$\{[\hat{q}/\hat{p}]/n\}^{1/2}$	$e^{\{\}}$ or $\exp\{\}$	[16]
logit	$\log[\hat{p}/\hat{q}]$	$\{[1/\hat{p} + 1/\hat{q}]/n\}^{1/2}$	$e^{\{\}}/(1+e^{\{\}})$	[17]
arcsin	$\arcsin[\sqrt{\hat{p}}]$	$\{0.25/n\}^{1/2}$		[18]

Bayes: Posterior distribution $f_{post}(p) \mid y/n$ and $f_{prior}(p)$. [19]

Notes:
[1] *Different concepts* (prop'n *in*, prop'n who *change* state), but *same statistical structure.*
[2] One au. avoids π and $\hat{P}/\hat{\pi}/\hat{p}$ by using upper & lower case for theoretical & empirical.
[4] Assuming n *independent* observations.
[6] cf. *jh* notes, c607/ch6, or Armitage&Berry, re 2-tailed P-value if non-symmetric distr'n.
[7] cf. *jh* notes, c607/ch6, or Armitage&Berry, re *"mid-P"* P-value.
[9] Not listed here: augmenting of numerator and denominator (e.g. "Wilson").
[11] Known as *Klopper-Pearson* CI: see *jh* site, c607/ch6 and ch8, and footnote to Table.
[13] Compute CI in new scale, then back-transform to original 0-1 scale. cf jh c607 ch4/8.
[14] Transform known as links in generalized linear models. cf ch8 for `SAS`, `Stata` & `R` code.
[15] Untransformed: note familiar unit variance $\hat{p}(1-\hat{p}) = \hat{p}\hat{q}$ in orginal 0 to 1 p scale.
[16] Natural log (ln). ie $\log 1 = 0; \log 2 = 0.69; \log 3 = 1.10; \log 0.5 = -0.69$, etc.
[17] $\hat{p}/\hat{q}$ = *odds*. So, logit$[\hat{p}]$ = *log odds*. cf. c607/ch4 for prop'n, odds & logits scales.
[18] Known as *variance-stabilizing* transformation, since variance is independent of p.
[19] cf Bayesian inf. for p - ch 8. $Beta(\alpha, \beta)$ prior $\rightarrow Beta(\alpha + y, \beta + n - y)$ posterior.

Difference in Prevalence or Risk in *index* ($_1$) *vs. reference* ($_0$) category

Theoretical: $p_1 \& p_0 \rightarrow p_1 - p_0 (RD);\ p_1 \div p_0 (RR);\ p_1/q_1 \div q_0/q_0 (OR, \psi)$
Empirical: $\hat{p}_1 \& \hat{p}_0 \rightarrow \hat{p}_1 - \hat{p}_0;\ \hat{p}_1 \div \hat{p}_0;\ \hat{p}_1/\hat{q}_1 \div \hat{p}_0/\hat{q}_0 = \hat{OR} = \hat{\psi}$
Model: $y_i \sim Binomial(n_i, p_i), i = 0, 1;\ y_1$ independent of y_0.

P-value: $H_0 :\ RD = 0;\ RR = 1;\ OR = 1;$ H_0 doesn't have to be 0 or 1

Exact
- Fisher's exact test, condn'l on fixed $y_1 + y_0$

Approx. [using the $\hat{p}_i$'s, or transforms, $t(\hat{p}_i)$, of them]
- $z = \sqrt{X^2} = \{t(\hat{p}_1) - t(\hat{p}_0)\}/\{Var[t(\hat{p}_1)] + Var[t(\hat{p}_0)]\}^{1/2}$

CI: Exact – for OR only – conditional
- $\psi_L : P[y_1 \geq y_{1obs} | \psi_L] = \alpha/2;\ \psi_U : P[y_1 \leq y_{obs} | \psi_U] = \alpha/2$

Approx. – all 3 measures
- ci: $\{t(\hat{p_1}) - t(\hat{p_0}) \pm z_{a/2}(Var[t(\hat{p_1})] + Var[t(\hat{p_0})])^{1/2}\} \rightarrow CI$

measure	transform	$\hat{t(p)}$	$ci \rightarrow CI$
RD Risk Diff.	identity	$\hat{p}$	n/a
RR Risk Ratio	log	$\log[\hat{p}]$	e^{ci}, *i.e.*, $\exp(ci)$
OR Odds Ratio	logit	$\log[\hat{p}/\hat{q}]$	e^{ci}, *i.e.*, $\exp(ci)$
NNT	reciprocal of CI for Risk Difference		

Notes:
[1] Index and reference categories of "Exposure" (Determinant)
[2] Fisher, and Breslow & Day, use Greek ψ for OR.
[4] Assuming n_1 & n_0 *independent* observations.
[7] So, *all 4* marginal frequencies are *fixed.* Central (null) hypergeometric distr'n.
[9] Several *equivalent* versions of X^2 for 2×2 table .. See jh c607/ch9.
... $X^2 = \sum\{n_{ij} - E[n_{ij}]\}^2/E[y_{ij}],\ \ i = 0, 1; j = +/-$; $\sum$ over 4 cells.
... $X^2 = n(ad - bc)^2/\{r_1 r_2 c_1 c_0\}$ $\quad r_1, r_2, c_1, c_2, n$: row/col/overall totals
... $X^2 = (\hat{p_1} - \hat{p_0})^2/\{Var_{null}[\hat{p_1}] + Var_{null}[\hat{p_0}]\}$ square of z-statistic
... $X^2 = \{a - E_{null}[a]\}^2/Var_{null}[a]; Var_{null}[a] = r_1 r_2 c_1 c_2/n^3;\ n^2(n-1)$ in M-H version.
[11] 1st-principles CI, using non-central (non-null) hypergeometric distr'n. cf Fisher 1935.
... Can use same Excel spreadsheet (jh c607 ch 8 resources) for exact *test* and exact *CI.*
[14] For more on *test-based* CI's, see jh 607/ch 8.2.
[14] To fit measures using generalized linear models, cf c607, ch8.2 `SAS`, `Stata` & `R` code.
[17] $Var[\log \hat{OR}] = 1/y_1 + 1/(n_1 - y_1) + 1/y_0 + 1/(n_0 - y_0) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$. Woolf 1955
[15-17] Rothman2002Ch7 emphasizes ease of manual calculation over heuristics.

Inference re Epidemiologic Parameter: **Rate or Incidence Density** [1]

Theoretical:	*Rate* or ID or λ	[2]
Empirical:	c cases in PT population-time units : $\widehat{ID} = \hat{\lambda} = c/PT$;	[3]
Model:	$c \sim Poisson(\mu = \lambda \times PT)$	[4]
P-value:	$\mathrm{H}_0:\ ID = ID_0,\ \lambda = \lambda_0;\ \rightarrow \mu = \lambda_0 \times PT = \mu_0$;	[5]
	Exact	[6]
	• $P[\ c \geq c_{obs}\ \mid\ \mu_{null}\ ]$ (1-tail)	[7]
	Approx.	[8]
	• N'l Approx. to distr'n of c or $\hat{\lambda}$, or transform $t(c)$ of c	[9]
CI:	Exact	[10]
	• $\mu_L : P[c \geq c_{obs} \vert \mu_L] = \alpha/2;\ \mu_U : P[c \leq c_{obs} \vert \mu_U] = \alpha/2$	[11]
	Approx.	[12]
	• reverse transform of $\{t(c) \pm z_{a/2} \times SE[t(c)]\}/PT$	[13]

transform	$t(c)$	$SE[t(\hat{p})] = Var^{1/2}$	reverse	[14]
identity	c	$c^{1/2}$	n/a	[15]
log	$\log[c]$	$\{1/c\}^{1/2}$	$e^{\{\ \}}$ or $\exp\{\ \}$	[16]
sqrt	$\sqrt{c}$	$\{0.25\}^{1/2} = 0.5$	$\{\sqrt{c} \pm 0.5 z_{\alpha/2}\}^2$	[17]

Notes:

[1] "Rate" in the incidence density sense.

[2-4] helps to separate obs'd & exp'd *numerator*, μ & c, from *rate* $(ID, \lambda, \hat{\lambda})$. $\mu = \lambda \times PT$

[3] We could use the usual 'y' as the numerator (i.e., the count) but 'c' is more meaningful.

[5] Interested in λ, not μ, but it is c which has the Poisson distribution!

[7,11] Can use tables and Excel tool in c634 Resources for Rates

[13] Note that CI shown is of the form $\{CI$ for $\mu\} \div PT$.

[15] The variance of a Poisson random variable is a function *only* of the mean μ.

[17] Variance-stabilizing transformation.

Difference in ID's or Rates in *index* $(_1)$ *vs. reference* $(_0)$ category

Theoretical:	λ_1 & $\lambda_0 \rightarrow \lambda_1 - \lambda_0\ (IDD);\ \ \lambda_1 \div \lambda_0\ \ (IDR)$
Empirical:	c_1/PT_1 & $c_0/PT_0 \rightarrow \hat{\lambda}_1 - \hat{\lambda}_0;\ \hat{\lambda}_1 \div \hat{\lambda}_0$
Model:	$c_i \sim Poisson(\mu_i = \lambda_i \times PT), i = 0, 1$; c_1 indep't of c_0.
P-value:	$\mathrm{H}_0:\ IDD = 0;\ IDR = 1$;
	Exact
	• $c_1 \vert (c_1 + c_0) \sim Binomial(''n'' = c_1 + c_0, \pi = \mu_1/\{\mu_1 + \mu_0\})$
	Approx. [using the $\hat{\lambda}_i$'s, or transforms, $t(\hat{\lambda}_i)$, of them]
	• $z = \sqrt{X^2} = \{t(\hat{\lambda}_1) - t(\hat{\lambda}_0)\}/\{Var[t(\hat{\lambda}_1)] + Var[t(\hat{\lambda}_0)]\}^{1/2}$
CI:	Exact – IDR only
	• $IDR_L : P[c_1 \geq c_{1obs} \vert IDR_L] = \alpha/2;\ IDR_U : -\ similarly$
	Approx. – both IDD and IDR
	• ci: $\{t(\hat{\lambda_1}) - t(\hat{\lambda_0}) \pm z_{a/2}(Var[t(\hat{\lambda_1})] + Var[t(\hat{\lambda_0})])^{1/2}\} \rightarrow CI$

measure	transform	$t(\hat{\lambda})$	$ci \rightarrow CI$	test-based* CI
ID Diff.	identity	$\hat{\lambda}$	n/a	$\widehat{IDD} \times (1 \pm z_{\alpha/2}/X)$
ID Ratio	log	$\log[\hat{\lambda}]$	e^{ci}	$\widehat{IDR}^{(1 \pm z_{\alpha/2}/X)}$

Notes:

[1] Index and reference categories of "Exposure" (Determinant)

[2] "RR" has many interpretations. If I use the Rate Ratio, I prefer to spell it out.

[4] Assuming *independent* samples.

[7] Fixing the sum $c_1 + c_0$ makes it possible to eliminate 1 nuisance parameter.

[9] Again here, several *equivalent* versions of X^2 for 2 counts. See jh c607/ch9.

... Don't force c_1, c_0, PT_1, PT_0 into a 2×2 table. See depiction as a '2×1' table (jh Ch9).

[11] Use Binomial distr'n; π is determined by (function of) the IDR & ratio of the PT's.

... Use def'n. of μ's to show: $\mu_1/(\mu_1 + \mu_0) = IDR/(IDR + PT_0/PT_1)$

... Can use same Excel spreadsheet (jh c607 ch 8 resources) for exact *test* and exact *CI*.

[14] Can also use a test-based CIs for a *risk* difference/ratio or an odds ratio.

... Test-based CI's use the $Variance\ under\ the\ Null$

[14] To fit IDD & IDR using regression models, cf Resources for Rates.

... Think of $E[\ c \mid PT] = \lambda \times PT$, or $E[\ y \mid X] = \beta \times X$ as a *regression equation*!

[17] $Variance[I\hat{D}R]$ had just 2 terms, $1/c_1 + 1/c_0$, emphasizing *role of the no.of cases* in the reliability of an estimated ID *Ratio*. (not same principle for ID *Difference*!)

[15-17] Rothman2002Ch7 emphasizes ease of manual calculation over heuristics.

Epidemiologic Parameter: **Survival Proportion or Cumulative Incidence** .

Theoretical: $S(t)$, and its complement $1 - S(t) = R_{0-t} = CI_{0-t}$

Empirical(**1**): $\{\widehat{ID_1}, \widehat{ID_2}, \ldots, \widehat{ID_K}\}$ in K sub-intervals spanning $[0, t]$.

Model(**1**): Assume $Var[\widehat{ID_k}], \ldots, Var[\widehat{ID_K}]$, are available.

Let $w_1, \ldots, w_K$ be the widths of the intervals.

Point Est.: $\widehat{S(t)} = \exp\{-\sum \widehat{ID_k} \times w_k\} = \exp\{-integral\}$; $\widehat{CI_{0-t}} = 1 - \widehat{S(t)}$

Variance of the *integral*: $V = \sum Var[\widehat{ID_k}] \times w_k^2$

CI for $S(t)$ $\exp\{-[integral \pm z_{\alpha/2} \times V^{1/2}]\}$

======= ===

Empirical(**2**): J *narrow* death-containing intervals in $[0, t]$.
n_j at risk just before the death(s) in interval j.
s_j survive death-containing interval j. Remaining d_j do not.

Model(**2a**): *conditional* prob's: $\widehat{S_1} = \frac{s_1}{n_1}$, $\widehat{S_2} = \frac{s_2}{n_2}$, ..., $\widehat{S_J} = \frac{s_J}{n_J}$

$\widehat{S_j} \sim Binomial(n_j, S_j)$ – only for variance calculations below

Point Est.: $\widehat{S_{KM}(t)} = \widehat{S_1} \times \widehat{S_2} \cdots \times \widehat{S_J}$ ... **Kaplan-Meier Product Limit Estimator**

Variance $\log \widehat{S_{KM}(t)} = \sum \log \widehat{S_k} \rightarrow Var[\log \widehat{S_{KM}(t)}] = \sum \frac{d_j}{s_j \times n_j} = V$ (*say*)

CI for $S(t)$

- $\exp\{\log \widehat{S_{KM}(t)}] \pm z_{\alpha/2} \times V^{1/2}\}$ [ci in $\log S$ scale → CI in S scale]
- $\widehat{S_{KM}(t)} \pm z_{\alpha/2} \times \widehat{S_{KM}(t)} \times V^{1/2}$ [**Greenwood's formula**]
- others, based on other transformations; $t()$ = c-log-log recommended

Model(**2b**): "counting process"; $\widehat{d_j} \sim Poisson()$ – only for variance calculations below

Point Est.: $\widehat{S_{NA}(t)} = \exp\{-integral\} = \exp\{-\sum \frac{d_j}{n_j}\}$ ... **Nelson-Aalen Estimator**

Variance $Var[integral] = \sum \frac{d_j}{n_j^2} = V$ (*say*)

CI for $S(t)$ • $\exp\{-[integral \pm z_{\alpha/2} \times V^{1/2}]\}$ • $\widehat{S_{NA}(t)} \pm z_{\alpha/2} \times \widehat{S_{NA}(t)} \times V^{1/2}$

Object: **Comparison of 2 Survival or Cumulative Incidence curves**

======= ==

(**1**) Survival or Risk (i.e., Cum. Inc., CI) **Difference at a specific timepoint** t

Theoretical: $S_1(t) - S_2(t)$; or $CI(t)_1 - CI(t)_2$; or $Risk_{1[0 \to t]} - Risk_{2[0 \to t]}$

Empirical: $\widehat{S_1(t)} - \widehat{S_2(t)}$, along with SE_1 and SE_2 (Greenwood SE's)

Test Statistic: $ratio = \{\widehat{S_1(t)} - \widehat{S_2(t)}\}/\{SE_1^2 + SE_2^2\}^{1/2} \sim N(0,1)$ *under* H_0

Conf. Int: $\widehat{S_1(t)} - \widehat{S_2(t)} \mp z_\alpha \times \{SE_1^2 + SE_2^2\}^{1/2}$

======= ==

Object: **Test of equality (H_0) of 2 entire Survival or Cumulative Incidence curves**

Empirical: J *narrow* death-containing intervals in $[0, t_{max}]$.

n_j at risk just before the death(s) in interval j (the n_j persons make up '**riskset**' j)
s_j survive death-containing interval j. Remaining d_j do not.

2×2 table for j^{th} riskset, along with $E[d_{1j}|H_0]$ and $Var[d_{1j}|H_0]$

d_{1j} s_{1j} | n_{1j} $E[d_{1j}|H_0] = (n_{1j}/n_j) \times d_j$; $Var[d_{1j}|H_0] = n_{1j}n_{2j}d_js_j/\{n_j^2(n_j - 1)\}$
d_{2j} s_{2j} | n_{2j}
– – – – – ––
d_j s_j | n_j

Test Statistic: $X^2 = \frac{\{\sum_j d_{1j} - \sum_j E[d_{1j}|H_0]\}^2}{\sum_j Var[d_{1j}|H_0]} \sim \chi_1^2$

Terminology: This is called the "**Log-rank" test**

It has the same structure as Mantel & Haenszel's test of $H_0 : OR_1 = OR_2 = \cdots = OR_J = 1$.

In their application, the 2×2 tables were for different strata.

Here, they are for the different 'risksets', which happen to be nested one inside the one before.

Example: Armitage chapter; or p4. of JH Notes: Survival Analysis / Follow-up Studies .. Resources for survival analysis.