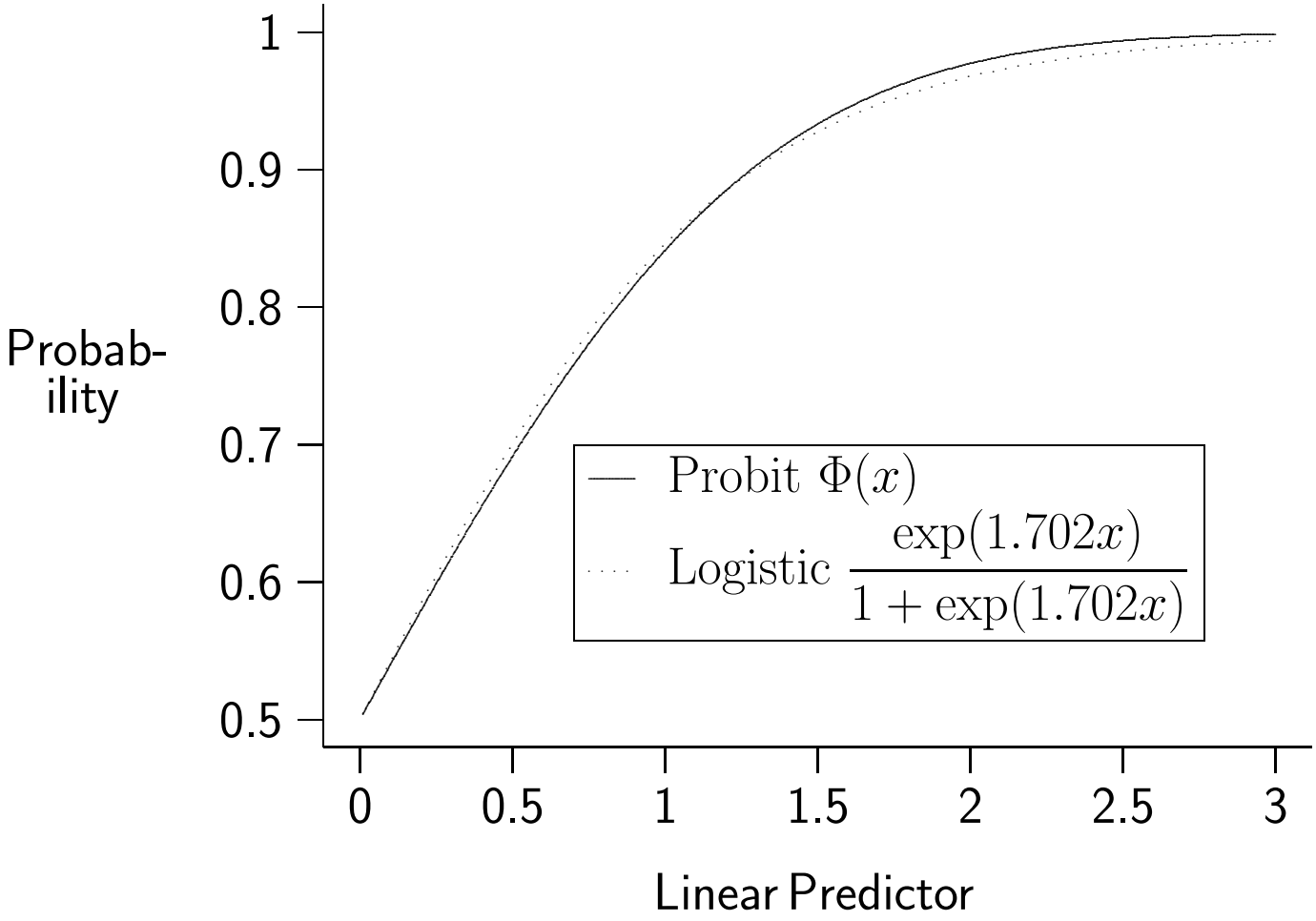
7. Continuous covariates may also be used
   a. As with simpler regression models, one should consider the proper scale for continuous covariates
   b. Consider adding polynomial terms
   c. Estimates could be seriously impacted by other variables in model
      i. mild effect of collinearity
      ii. Impact can be minimized by subtracting out mean.

St: 11.2

8. You can use another function instead of logit
   a. Must still map $\Re$ into $[0, 1]$
   b. Logit has some mathematical properties we will discuss later
   c. Normal CDF is sometimes used
      i. Called the probit
      ii. Results from discretizing standard multiple regression.
         - Suppose $Y_j = \boldsymbol{x}_j\boldsymbol{\beta} + \sigma\epsilon_j$, $\epsilon_j \sim N(0, 1)$
         - $Z_j = \begin{cases} 1 & \text{if } Y_j > c \\ 0 & \text{otherwise} \end{cases}$.
         - $\mathsf{P}\left[Z_j = 1\right] = \mathsf{P}\left[Y_j > c\right] = \mathsf{P}\left[\epsilon_j > (c - \boldsymbol{x}_j\boldsymbol{\beta})/\sigma\right] = 1 - \Phi((c - \boldsymbol{x}_j\boldsymbol{\beta})/\sigma) = \Phi((\boldsymbol{x}_j\boldsymbol{\beta} - c)/\sigma)$



- After rescaling, probit and logit are very close. See Figure 6.

Se: 9 pp. 298–310

G. Regression Models for case–control studies

1. We want to model $\mathrm{P}\left[\boldsymbol{X}_j = \boldsymbol{x}_j | Y_j = y_j\right] =$
   $\mathrm{P}\left[\boldsymbol{X}_j = \boldsymbol{x}_j \& Y_j = y_j\right] / \mathrm{P}\left[Y_j = y_j\right] =$
   $\mathrm{P}\left[Y_j = y_j | \boldsymbol{X}_j = \boldsymbol{x}_j\right] \mathrm{P}\left[\boldsymbol{X}_j = \boldsymbol{x}_j\right] / \mathrm{P}\left[Y_j = y_j\right]$
2. All three probabilities might depend on $\boldsymbol{\beta}$
   a. The first is most direct $\mathrm{P}\left[Y_j = y_j \forall j | \boldsymbol{X}_j = \boldsymbol{x}_j \forall j\right] =$

$$\prod_j \frac{\exp(\boldsymbol{x}_j\boldsymbol{\beta}y_j)}{1+\exp(\boldsymbol{x}_j\boldsymbol{\beta})}$$

b. If a certain risk factor makes elements in a population more likely to die, P $\left[\boldsymbol{X}_j = \boldsymbol{x}_j\right]$ will be influenced by $\boldsymbol{\beta}$

i. Remove this dependence by conditioning

- on $\{\boldsymbol{X}_j\}$,
- not on the Y_j they are associated with.

ii. For example, in 2×2 table from case–control study, we condition on number of exposed and number unexposed.

iii. We say $\{\boldsymbol{X}_j\}$ is $ancillary$.

c. Marginal distribution of Y_j will certainly be influence by $\boldsymbol{\beta}$

i. Greater effect for risk factors $\Rightarrow$ more cases

ii. P $\left[Y_j = y_j \forall j | \{\boldsymbol{X}_j\}\right] = \sum \prod_j \frac{\exp(\boldsymbol{w}_j\boldsymbol{\beta}y_j)}{1+\exp(\boldsymbol{w}_j\boldsymbol{\beta})}$.

iii. sum is over all rearrangements $\boldsymbol{w}_j$ of the $\boldsymbol{x}_j$

3. Then P $\left[\boldsymbol{X}_j = \boldsymbol{x}_j \forall j | Y_j = y_j \forall j, \{\boldsymbol{X}_j\}\right] =$

$$\prod_j \frac{\exp(\boldsymbol{x}_j\boldsymbol{\beta}y_j)}{1+\exp(\boldsymbol{x}_j\boldsymbol{\beta})} / \sum \prod_j \frac{\exp(\boldsymbol{x}_j\boldsymbol{\beta}y_j)}{1+\exp(\boldsymbol{x}_j\boldsymbol{\beta})} = \frac{\prod_j \exp(\boldsymbol{x}_j\boldsymbol{\beta}y_j)}{\sum \prod_j \exp(\boldsymbol{w}_j\boldsymbol{\beta}y_j)}$$

B&D1: 7.0, 7.2

4. In a stratified study

a. Suppose there are K strata

b. $\mathsf{P}\left[Y_{jk}=1\right]=\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta}+\alpha_k)/(1+\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta}+\alpha_k))$

c. $\prod_{jk}\mathsf{P}\left[Y_{jk}=y_{jk}\forall j,k|\boldsymbol{X}_{jk}=\boldsymbol{x}_{jk}\right]=\prod_{jk}\exp([\boldsymbol{x}_{jk}\boldsymbol{\beta}+\alpha_k]y_{jk})/(1+\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta}+\alpha_k))$

d. Condition on the number of cases and control in each strata, and on the collection of covariate patterns $\{\boldsymbol{x}_{jk}\}$:

$$\prod_{jk}\mathsf{P}\left[Y_{jk}=y_{jk}\forall j,k|\{\boldsymbol{X}_{jk}\}=\{\boldsymbol{x}_{jk}\},\sum_j Y_{jk}=\sum_j y_{jk}\forall j\right]$$
$$=\sum\prod_{jk}\exp([\boldsymbol{x}_{jk}\boldsymbol{\beta}+\alpha_k]y_{jk})/(1+\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta}+\alpha_k))$$

i. Summation is over rearrangements of covariates within strata.

e. $\mathsf{P}\left[\boldsymbol{X}_j=\boldsymbol{x}_j\forall j|Y_j=y_j\forall j\right]=$

$$\prod_{jk}\frac{\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta}y_{jk})}{1+\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta})}/\sum\prod_{jk}\frac{\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta}y_{jk})}{1+\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta})}=\frac{\exp(\sum_{jk}\boldsymbol{x}_{jk}\boldsymbol{\beta}y_{jk})}{\sum\prod_{jk}\exp(\boldsymbol{x}_{jk}\boldsymbol{\beta}y_j)}$$

i. α_k s canceled because of linearity in exponent of probability

ii. Probability only depends on data through $\sum_{jk}\boldsymbol{x}_{jk}y_{jk}$

?: ?

iii. Models with this structure are called *canonical exponential families.*

B&D1: 7.1, 7.5

f. Suppose we fit a logistic regn. model to the stratified tables

i. ${O_{1k}}^i \sim \text{Bin}(\pi_{ki}, {O_{1+}}^i)$

ii. $\text{logit}(\pi_{ki}) = \alpha_i + \gamma_k$ with $\gamma_1 = 0$

- Probability only depends on data through $\sum_{jk} \boldsymbol{x}_{jk} y_{jk}$
- Best fit will make $\sum_{jk} \boldsymbol{x}_{jk} y_{jk} = \sum_{jk} \boldsymbol{x}_{jk} \pi_{jk}$

iii. concordant pairs match y_{jk} exactly

- Both exposed: $\hat{\alpha}_i = -\hat{\gamma}_2$
- Both unexposed: $\hat{\alpha}_i = -\hat{\gamma}_1$

iv. discordant pairs:

- P [diseased|exposed] + P [diseased|unexposed] $= \pi^i_{+k} = 1 \forall$ stratum
  - ▷ Hence $\exp(\hat{\alpha}_j)/(1 + \exp(\hat{\alpha}_j)) + \exp(\hat{\gamma}_2 + \hat{\alpha}_j)/(1 + \exp(\hat{\gamma}_2 + \hat{\alpha}_j)) = 1$
  - ▷ Hence $\exp(\hat{\alpha}_j) + \exp(2\hat{\alpha}_j + \hat{\gamma}_2) \exp(\hat{\gamma}_2 + \hat{\alpha}_j) + \exp(\hat{\alpha}_j + \hat{\gamma}_2) = \exp(\hat{\alpha}_j) + \exp(2\hat{\alpha}_j + \hat{\gamma}_2) \exp(\hat{\gamma}_2 + \hat{\alpha}_j) + \exp(\hat{\alpha}_j + \hat{\gamma}_2)$
  - ▷ Hence $\hat{\alpha}_j = -\hat{\gamma}_2/2$
- Sum of fitted values among exposed = number of exposed cases: $(n_{10} + n_{01}) \exp(\hat{\gamma}_2/2)/(1 + \exp(\hat{\gamma}_2/2)) = n_{10}$
  - ▷ Hence $(n_{10} + n_{01}) \exp(\hat{\gamma}_2/2) = (1 + \exp(\hat{\gamma}_2/2)) n_{10}$

▷ Hence $\hat{\gamma}_2 = 2\log(n_{10}/n_{01})$

B&D1: 7.6

g. Suppose we ignore stratification:

i. Recall $\psi = \frac{\pi_2(1-\pi_1)}{\pi_1(1-\pi_2)}$

ii. Use $\hat{\psi}^* = \frac{[\sum_i {O_{22}}^i][\sum_i {O_{11}}^i]}{[\sum_i {O_{21}}^i][\sum_i {O_{12}}^i]}$

iii. Estimate

$$\psi^* = \frac{[\sum_i(1-\pi_{1i})][\sum_i \pi_{2i}]}{[\sum_i(1-\pi_{2i})][\sum_i \pi_{1i}]} = \frac{[\sum_i(1-\pi_{1i})][\sum_i \psi\vartheta_i(1-\pi_{2i})]}{[\sum_i(1-\pi_{2i})][\sum_i \vartheta_i(1-\pi_{1i})]}$$
$$= \psi\frac{[\sum_i \vartheta_i(1-\pi_{2i})]/[\sum_i(1-\pi_{2i})]}{[\sum_i \vartheta_i(1-\pi_{1i})]/[\sum_i(1-\pi_{1i})]}$$

for $\vartheta_i = \pi_{1i}/(1-\pi_{1i})$.

h. Note

$$\frac{\sum_i \vartheta_i(1-\pi_{2i})}{\sum_j(1-\pi_{2j})} - \frac{\sum_i \vartheta_i(1-\pi_{1i})}{\sum_j(1-\pi_{1j})}$$
$$= \sum_i \vartheta_i \left[\frac{1-\pi_{2i}}{\sum_j(1-\pi_{2j})} - \frac{1-\pi_{1i}}{\sum_j(1-\pi_{1j})}\right]$$

i. Note $1-\pi_{1i} = \vartheta_i/(\psi+\vartheta_i)$ and $1-\pi_{2i} = 1/(1+\vartheta_i)$, and

$(1-\pi_{1i})/(1-\pi_{2i}) = (\psi+\vartheta_i)/(\psi(1+\vartheta_i))$

i. Big values are associated with big ϑ_i

H. Summary:

Number of Strata	Covar–iates?	Logist. Regr.	Cond. Logist. Regr.	Mantel–Haentzel
Cohort:				
Large	no	☹	☺	✓
Small	no	☺	🕒	✓
Large	yes	☹	☺	☹
Small	yes	☺	🕒	☹
Case–Control:				
Any	no	☹	☺	✓
Any	yes	☹	☺	☹

☺ Good

✓ Test OK; estmator suboptimal except for paired data

☹ Inappropriate

🕒 Slow

1. Mantel–Haentzel test and estimator reduces to McNemar's test and estimator when strata are pairs.

Se: 3 pp. 83–86

IV. Sample Size Calculations

A. Preliminaries

1. We'll do power for 1-sided tests

a. Conceptually easier (as we shall see)

b. Get power for 2-sided tests by doubling α

B. Exactly:

1. Select smallest C such that $\mathsf{P}_0[T \geq C] \leq \alpha$
2. Power is $\mathsf{P}_A[T \geq C]$.

C. Approximately,

1. Suppose
   a. $H_0 : T \sim \mathcal{N}(\mu_0, \sigma_0^2)$
   b. $H_A : T \sim \mathcal{N}(\mu_A, \sigma_A^2)$
2. Critical value: C
   a. Reject H_0 if $(T - \mu_0)/\sigma_0 \geq z_\alpha$
   b. $1 - \Phi(z_\alpha) = \alpha$
   c. Reject H_0 if $T \geq \mu_0 + \sigma_0 z_\alpha$
3. Power is $\mathsf{P}_A[T \geq C] = \Phi((\mu_A - \mu_0 - \sigma_0 z_\alpha)/\sigma_A)$
   a. Special Case: $\sigma_0 = \sigma_A$, power is $\Phi((\mu_A - \mu_0)/\sigma_0 - z_\alpha)$
4. Sample size:
   a. Assume that $\sigma_0 = \tau_0/\sqrt{n}$, $\sigma_A = \tau_A/\sqrt{n}$.
   b. Require power $1 - \beta$
      i. Typically, $.8 = 80\%$.

c. Then $-z_\beta = (\mu_0 + \sigma_0 z_\alpha - \mu_A)/\sigma_A$.

i. $(\sigma_A z_\beta + \sigma_0 z_\alpha) = \mu_A - \mu_0$

ii. $(\tau_A z_\beta + \tau_0 z_\alpha)/\sqrt{n} = \mu_A - \mu_0$

iii. $(\tau_A z_\beta + \tau_0 z_\alpha)/(\mu_A - \mu_0) = \sqrt{n}$

iv. $(\tau_A z_\beta + \tau_0 z_\alpha)^2/(\mu_A - \mu_0)^2 = n$

v. When $\tau_A = \tau_0$, $n = \tau_0^2(z_\beta + z_\alpha)^2/(\mu_A - \mu_0)^2 = n$

Se: 3 pp. 87–90

D. One-way analyses

1. External standards based on SMR

a. We could in principal perform calculations exactly

i. This is typically not considered necessary

b. Assume $O \sim \mathcal{P}(e)$

i. Approximately, $T = \sqrt{O} \sim \mathcal{N}(\sqrt{e}, \frac{1}{4})$.

ii. Recall $e = \sum_i R_i P_{1i} \Delta$

c. The power is $\approx 1 - \Phi(z_\alpha + 2(\sqrt{e} - \sqrt{e\varsigma}))$

d. Sample sizes are input and output via e

i. To get power $1 - \beta$ for level α test, need $-z_\beta = z_\alpha + 2(\sqrt{e} - \sqrt{e\varsigma})$ or $\frac{1}{4}(z_\beta + z_\alpha)^2/(1 - \sqrt{\varsigma})^2 = e$

e. Example:

i. In the first homework assignment, we saw that the maximal rate of nasal cancer in the unexposed population was 4.25 $\times 10^{-5}$ cases per year.

ii. Let λ be rate for nickel smelters.

iii. Test $H_0 : \lambda = 4.25 \times 10^{-5}$, vs. $H_A : \lambda > 4.25 \times 10^{-5}$ with $\alpha = 2.5\%$.

iv. Calculate power for alternative $\lambda = 5 \times 4.25 \times 10^{-5}$.

v. If we follow a cohort of 10000 exposed individuals for 5 years, we expect 2.125 cases under H_0

vi. Then the power is

$$1 - \Phi(1.96 + 2 \times (\sqrt{2.125} - \sqrt{2.125 \times 5})) = 95\%$$

vii. To get 80% power, we need $e = \frac{1}{4}(z_{.2} + z_{.025})^2 \times (1 - \sqrt{5})^{-2} = 1.284$, which if the null hypothesis rate is true corresponds to roughly 1010 individuals for 5 years, if our alternative is true.

Se: 3 p. 102

2. Dichotomous exposures

a. Without age stratification

b. H_0 : SMR=1 vs. H_A : SMR=ς

c. Inference is based on $O_1|O_+ \sim \text{Bin}(\pi_1, O_+)$

d. $T = O_1/O_+$.

e. $\mu_0 = Q_1/(Q_0 + Q_1)$, $\sigma_0 = \sqrt{\mu_0(1-\mu_0)/O_+}$.

f. $\mu_A = \varsigma Q_1/(Q_0 + \varsigma Q_1) = \frac{\pi_1^0 \varsigma}{1+(\varsigma-1)\pi_1^0}$ and $\sigma_A = \sqrt{\mu_A(1-\mu_A)/O_+}$.

g. No serious simplification to power and sample size formulae

h. Example:

i. For shipping example, suppose that we want to test H_0 : two ship classes have same accident rate

ii. Want 80% power to detect difference if one class has 50% more accidents

iii. Suppose that both kinds of ships have equal months at risk

iv. Need $O_+ = \frac{(\sqrt{.6\times.4}z_{.8} - z_{.025}.5)^2}{(.6-.5)^2} = 194$ accidents.

v. Median rate was .002 accidents per month

vi. Hence need to follow ships for $194/.002 = 95000$ total months.

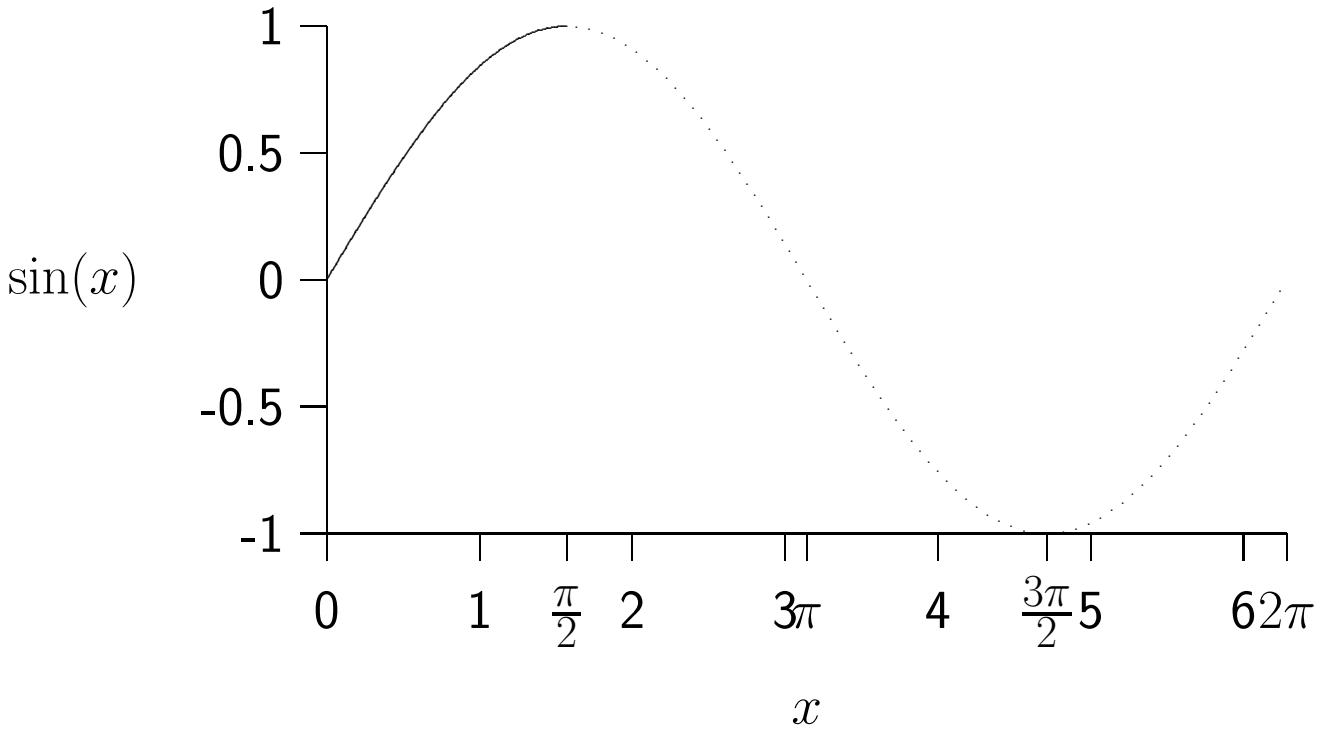
i. Power is approximate

i. Better approximation for Poisson uses fact that when $O \sim \mathcal{P}(\mu)$ then $\text{Var}\left[\sqrt{O}\right] \approx \mu \times (\frac{1}{2}\mu^{-1/2})^2 = \frac{1}{4}$.

ii. Better approximation for binomial uses fact that

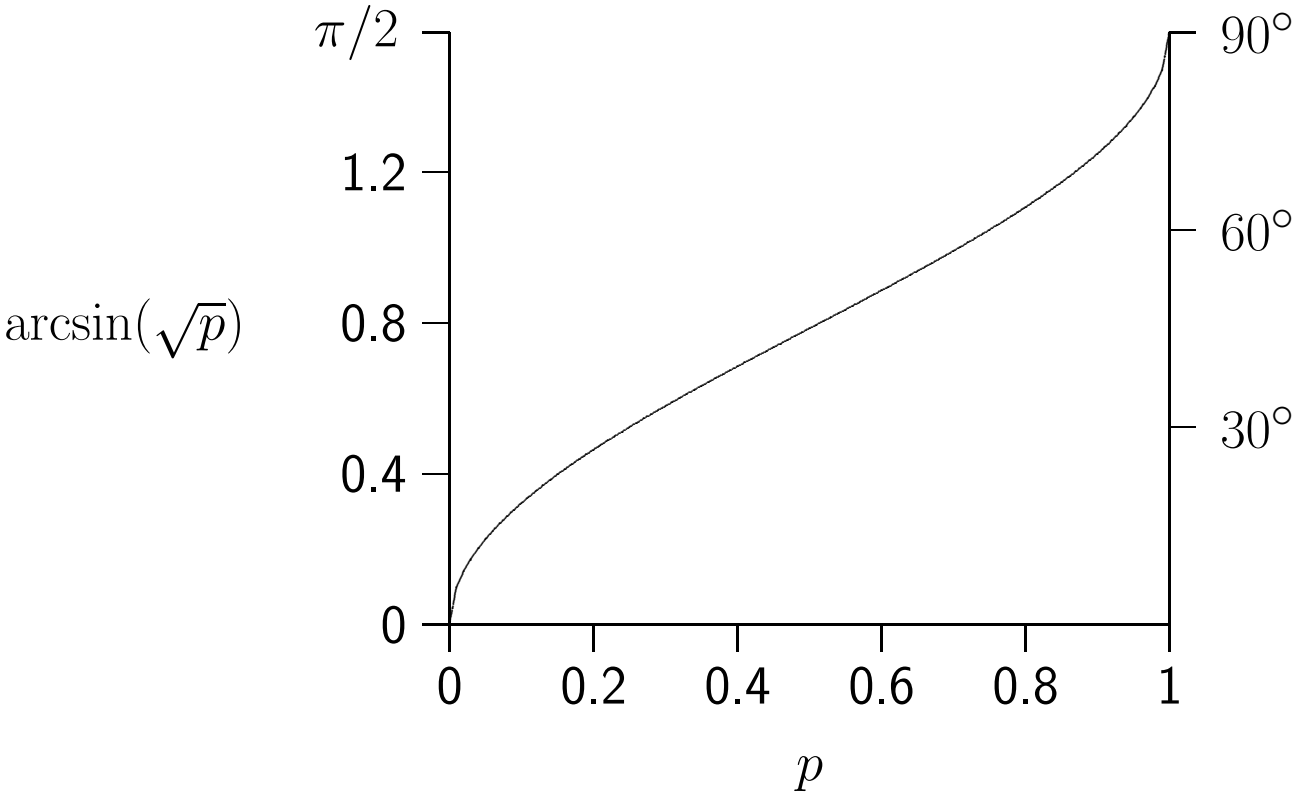
$$\arcsin(\sqrt{O_1/O_+}) \sim \mathcal{N}(\arcsin(\sqrt{\varsigma Q_1/(Q_0+\varsigma Q_1)}), 1/(4O_+))$$

- $\frac{d}{dx}\arcsin(x) = 1/\sqrt{1-x^2}$
- $\frac{d}{dx}\arcsin(\sqrt{x}) = 1/(2\sqrt{x}\sqrt{1-x})$ See Figures 7 and 8.



iii. $\mu_0 = \arcsin(\sqrt{Q_1/(Q_0+Q_1)})$, $\mu_A = \arcsin(\varsigma\sqrt{Q_1/(Q_0+\varsigma Q_1)})$, $\sigma_A = \sigma_0 = \sqrt{1/(4O_+)}$

iv. Power and sample size as before.



v. Shipping example:

$$\frac{(z_{.8} - z_{.025})^2}{\left(\arcsin\left(\sqrt{.5}\right) - \arcsin\left(\sqrt{\frac{1.5}{1+1.5}}\right)\right)^2 4} = 194$$

j. Can also calculate minimal sample size to give a desired CI width.