Sampling and Sample Size Issues II

Michael Mohr

National Marine Fisheries Service
Southwest Fisheries Science Center
Santa Cruz, California

PSC GSI Workshop 2, 2007

Outline

Outline

1. Assumptions
2. Metrics
   - Probability of detecting ≥ 1 individuals from stock
   - Coefficient of variation of $\hat{p}$
   - Confidence interval width for p
   - Minimum detectable difference between p_1 and p_2

Outline

1. Assumptions

2. Metrics
   - Probability of detecting ≥ 1 individuals from stock
   - Coefficient of variation of $\hat{p}$
   - Confidence interval width for p
   - Minimum detectable difference between p_1 and p_2

3. Example application

Outline

1. Assumptions
2. Metrics
   - Probability of detecting ≥ 1 individuals from stock
   - Coefficient of variation of $\hat{p}$
   - Confidence interval width for p
   - Minimum detectable difference between p_1 and p_2
3. Example application
4. Additional comments

Assumptions

- All collected tissues processed and classified

Assumptions

- All collected tissues processed and classified
- All classifications are 100% accurate

Assumptions

- All collected tissues processed and classified
- All classifications are 100% accurate
- Type I and Type II error rates apply to p_i not $\{p_i\}$

Assumptions

- All collected tissues processed and classified
- All classifications are 100% accurate
- Type I and Type II error rates apply to p_i not $\{p_i\}$
- Simple random sampling within each stratum

Assumptions

- All collected tissues processed and classified
- All classifications are 100% accurate
- Type I and Type II error rates apply to p_i not $\{p_i\}$
- Simple random sampling within each stratum
- Sampling fraction is negligible (n relative to C, or N)

Probability of detecting ≥ 1 individuals from stock

Binomial model: $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$ implies

Probability of detecting ≥ 1 individuals from stock

Binomial model: $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$ implies

$$Q = P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1-p)^n$$

Probability of detecting ≥ 1 individuals from stock

Binomial model: $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$ implies

$$Q = P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1-p)^n$$

Probability of detecting ≥ 1 individuals from stock

Binomial model: $P(Y=y)=\binom{n}{y}p^y(1-p)^{n-y}$ implies

$$Q = P(Y \geq 1) = 1 - P(Y=0) = 1-(1-p)^n$$

Probability of detecting ≥ 1 individuals from stock

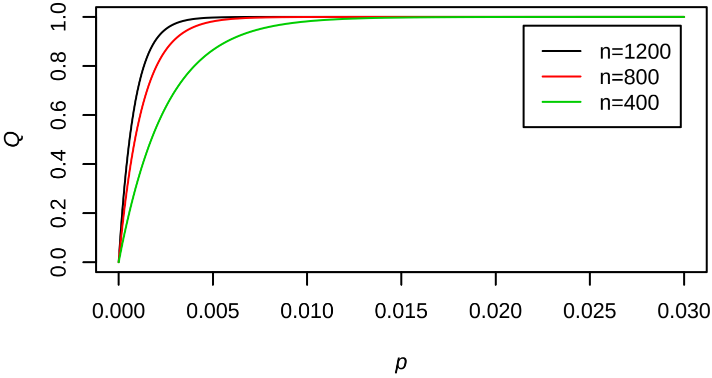
Binomial model: $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$ implies

$$Q = P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1-p)^n$$

Probability of detecting ≥ 1 individuals from stock

Binomial model: $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$ implies

$$Q = P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1-p)^n$$



Probability of detecting ≥ 1 individuals from stock

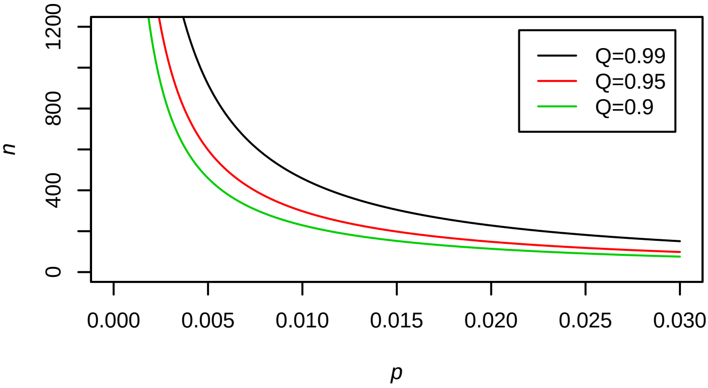
Binomial model: $P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}$ implies

$$n = \log(1-Q)/\log(1-p)$$

Probability of detecting ≥ 1 individuals from stock

Binomial model: $P(Y=y) = \binom{n}{y} p^y (1-p)^{n-y}$ implies

$$n = \log(1-Q)/\log(1-p)$$



Coefficient of variation of $\hat{p}$

Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

Coefficient of variation of $\hat{p}$

Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

$$CV = \sqrt{V(\hat{p})}/p = \sqrt{(1-p)/(np)}$$

Coefficient of variation of $\hat{p}$

Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

$$CV = \sqrt{V(\hat{p})}/p = \sqrt{(1-p)/(np)}$$

Coefficient of variation of $\hat{p}$

Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

$$CV = \sqrt{V(\hat{p})}/p = \sqrt{(1-p)/(np)}$$

Coefficient of variation of $\hat{p}$

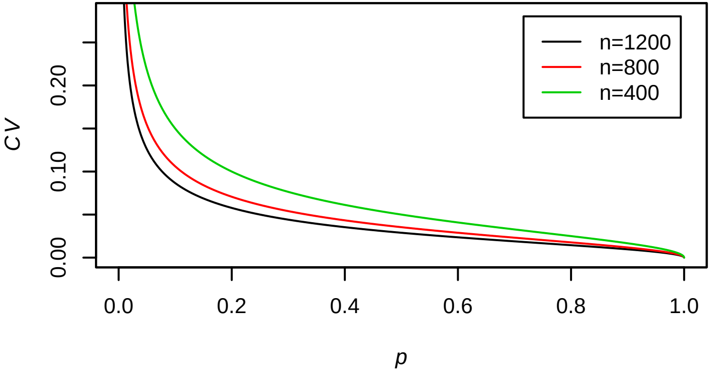
Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

$$CV = \sqrt{V(\hat{p})}/p = \sqrt{(1-p)/(np)}$$

Coefficient of variation of $\hat{p}$

Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

$$CV = \sqrt{V(\hat{p})}/p = \sqrt{(1-p)/(np)}$$



Coefficient of variation of $\hat{p}$

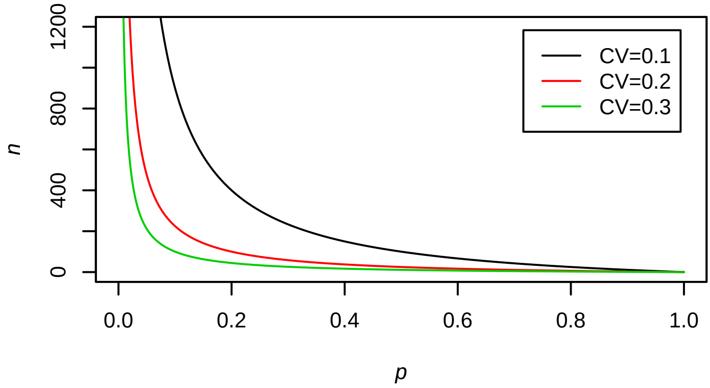
Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

$$n = (1-p)/(p \cdot CV^2)$$

Coefficient of variation of $\hat{p}$

Binomial model: $V(\hat{p}) = p(1-p)/n$ implies

$$n = (1-p)/(p \cdot CV^2)$$



Confidence interval width for p

Binomial model + normal approx. for $\hat{p}$ implies

Confidence interval width for p

Binomial model + normal approx. for $\hat{p}$ implies

$$w = 2 \cdot Z_{1-\alpha/2} \cdot \sqrt{p(1-p)/n}$$

Confidence interval width for p

Binomial model + normal approx. for $\hat{p}$ implies

$$w = 2 \cdot Z_{1-\alpha/2} \cdot \sqrt{p(1-p)/n}$$

Confidence interval width for p

Binomial model + normal approx. for $\hat{p}$ implies

$$w = 2 \cdot Z_{1-\alpha/2} \cdot \sqrt{p(1-p)/n}$$

Confidence interval width for p

Binomial model + normal approx. for $\hat{p}$ implies

$$w = 2 \cdot Z_{1-\alpha/2} \cdot \sqrt{p(1-p)/n}$$

Confidence interval width for p

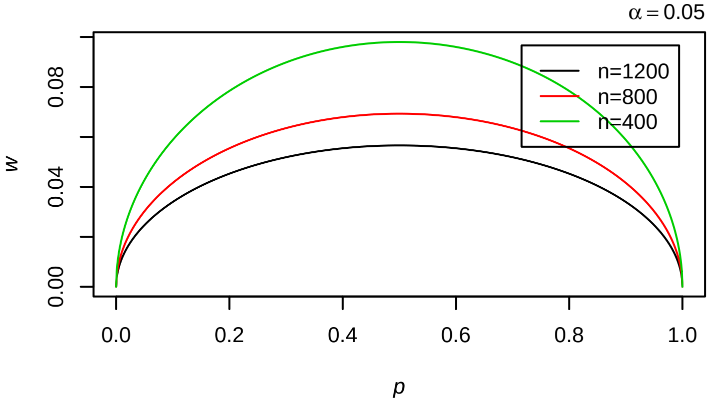
Binomial model + normal approx. for $\hat{p}$ implies

$$w = 2 \cdot Z_{1-\alpha/2} \cdot \sqrt{p(1-p)/n}$$

Confidence interval width for p

Binomial model + normal approx. for $\hat{p}$ implies

$$w = 2 \cdot Z_{1-\alpha/2} \cdot \sqrt{p(1-p)/n}$$



Confidence interval width for p

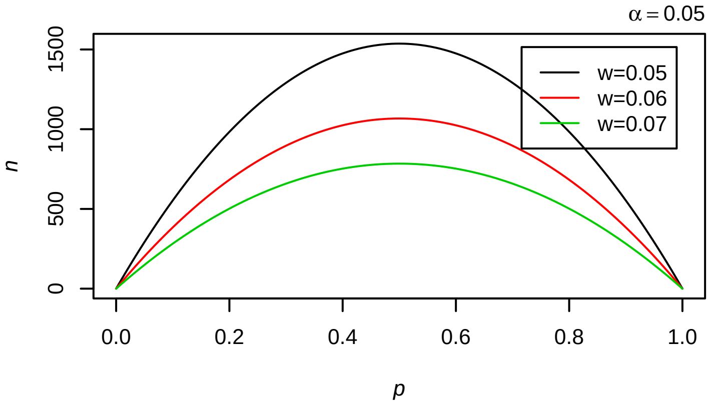
Binomial model + normal approx. for $\hat{p}$ implies

$$n = 4 \cdot Z^2_{1-\alpha/2} \cdot p(1-p)/w^2$$

Confidence interval width for p

Binomial model + normal approx. for $\hat{p}$ implies

$$n = 4 \cdot Z^2_{1-\alpha/2} \cdot p(1-p)/w^2$$



Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.

Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1+\sqrt{1+(4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1)+p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.

Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.

Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.

Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.

Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.

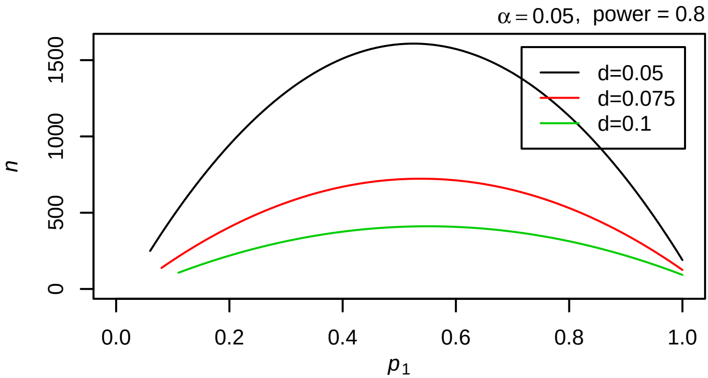
Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.



Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.

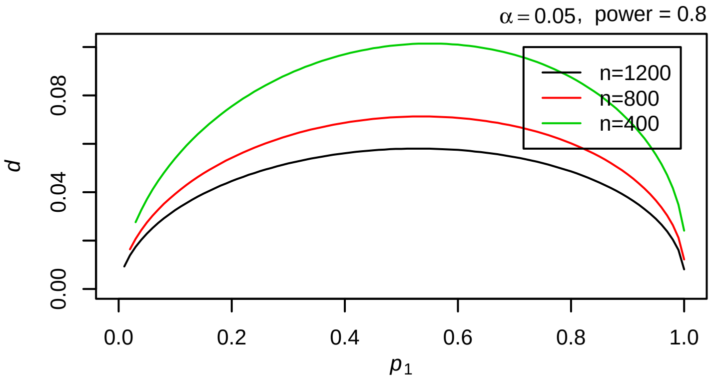
Minimum detectable difference between p_1 and p_2

Binomial model + normal approx. for $\hat{p}_1$ and $\hat{p}_2$ implies

$$n = (n'/4)\left[1 + \sqrt{1 + (4/(n'd))}\right]^2,$$

$$n' = \left[Z_{1-\alpha}\sqrt{2\bar{p}(1-\bar{p})} + Z_{\text{power}}\sqrt{p_1(1-p_1) + p_2(1-p_2)}\right]^2 / d^2.$$

with $p_1 > p_2$, $d = p_1 - p_2$, $\bar{p} = (p_1 + p_2)/2$.



Example Application: I

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery

Example Application: I

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery
- Statistical objectives (each stratum)
  - $Q \geq 0.999$ for $p \geq 0.01$
  - $CV \leq 0.20$ for $p \geq 0.03$

Example Application: I

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery
- Statistical objectives (each stratum)
  - $Q \geq 0.999$ for $p \geq 0.01$
  - $CV \leq 0.20$ for $p \geq 0.03$
- Minimum n per stratum
  - ≈ 800

Example Application: II

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery
- Statistical objectives (each stratum)
  - $Q \geq 0.999$ for $p \geq 0.01$
  - $CV \leq 0.20$ for $p \geq 0.03$
  - $w \leq 0.06$ for $p \geq 0.01$ with $\alpha = 0.05$

Example Application: II

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery

- Statistical objectives (each stratum)
  - $Q \geq 0.999$ for $p \geq 0.01$
  - $CV \leq 0.20$ for $p \geq 0.03$
  - $w \leq 0.06$ for $p \geq 0.01$ with $\alpha = 0.05$

- Minimum n per stratum
  - ≈ 1070

Example Application: III

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery
  - compare p inshore/offshore
- Statistical objectives (each stratum)
  - $Q \geq 0.999$ for $p \geq 0.01$
  - $CV \leq 0.20$ for $p \geq 0.03$
  - $w \leq 0.06$ for $p \geq 0.01$ with $\alpha = 0.05$
  - power ≥ 0.80 to detect $d = p_{off} - p_{in}$ with $\alpha = 0.05$

Example Application: III

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery
  - compare p inshore/offshore

- Statistical objectives (each stratum)
  - $Q \geq 0.999$ for $p \geq 0.01$
  - $CV \leq 0.20$ for $p \geq 0.03$
  - $w \leq 0.06$ for $p \geq 0.01$ with $\alpha = 0.05$
  - power ≥ 0.80 to detect $d = p_{off} - p_{in}$ with $\alpha = 0.05$
    - assume $p_{off} \geq 0.03$, and $p_{in} \leq p_{off}/2$

Example Application: III

- Goal
  - $\hat{p}$ for all stocks by time, area, fishery
  - compare p inshore/offshore
- Statistical objectives (each stratum)
  - $Q \geq 0.999$ for $p \geq 0.01$
  - $CV \leq 0.20$ for $p \geq 0.03$
  - $w \leq 0.06$ for $p \geq 0.01$ with $\alpha = 0.05$
  - power ≥ 0.80 to detect $d = p_{off} - p_{in}$ with $\alpha = 0.05$
    - assume $p_{off} \geq 0.03$, and $p_{in} \leq p_{off}/2$
- Minimum n_{in}, n_{off} per stratum
  - ≈ 1350

Age-specific stock composition

- Chinook
  - distribution is stock-age-specific
  - assessment requires stock-age-specific catch

Age-specific stock composition

- Chinook
  - distribution is stock-age-specific
  - assessment requires stock-age-specific catch

- Minimum n
  - treat stock-age components as “separate stocks”
  - apply sample size formulas

Age-specific stock composition

- Chinook
  - distribution is stock-age-specific
  - assessment requires stock-age-specific catch
- Minimum *n*
  - treat stock-age components as "separate stocks"
  - apply sample size formulas
- Additional requirements
  - GSI individual assignment (versus MSA)
  - age assignment (scale collection and reading)
  - both assignments assumed 100% accurate

Factors affecting stock proportion

- Size-selective fishery (for a stratum)

$$p_i = \frac{C_i}{\sum C_k} = \frac{N_i^* \alpha_i q_i f \ell_i}{\sum N_k^* \alpha_k q_k f \ell_k} = \frac{N_i^* \alpha_i q_i \ell_i}{\sum N_k^* \alpha_k q_k \ell_k}$$

Factors affecting stock proportion

- Size-selective fishery (for a stratum)

$$p_i = \frac{C_i}{\sum C_k} = \frac{N_i^* \alpha_i q_i f \ell_i}{\sum N_k^* \alpha_k q_k f \ell_k} = \frac{N_i^* \alpha_i q_i \ell_i}{\sum N_k^* \alpha_k q_k \ell_k}$$

Factors affecting stock proportion

- Size-selective fishery (for a stratum)

$$p_i = \frac{C_i}{\sum C_k} = \frac{N_i^* \alpha_i q_i f \ell_i}{\sum N_k^* \alpha_k q_k f \ell_k} = \frac{N_i^* \alpha_i q_i \ell_i}{\sum N_k^* \alpha_k q_k \ell_k}$$

Factors affecting stock proportion

- Size-selective fishery (for a stratum)

$$p_i = \frac{C_i}{\sum C_k} = \frac{N_i^* \alpha_i q_i f \ell_i}{\sum N_k^* \alpha_k q_k f \ell_k} = \frac{N_i^* \alpha_i q_i \ell_i}{\sum N_k^* \alpha_k q_k \ell_k}$$

- Assume $q_1 = q_2 = q_3 = \ldots$

$$p_i = \frac{N_i^* \alpha_i \ell_i}{\sum N_k^* \alpha_k \ell_k}$$

Factors affecting stock proportion

- Size-selective fishery (for a stratum)

$$p_i = \frac{C_i}{\sum C_k} = \frac{N_i^* \alpha_i q_i f \ell_i}{\sum N_k^* \alpha_k q_k f \ell_k} = \frac{N_i^* \alpha_i q_i \ell_i}{\sum N_k^* \alpha_k q_k \ell_k}$$

- Assume $q_1 = q_2 = q_3 = \dots$

$$p_i = \frac{N_i^* \alpha_i \ell_i}{\sum N_k^* \alpha_k \ell_k}$$

- Variation in p_i across strata (at time t)

  numerator: stock i

Factors affecting stock proportion

- Size-selective fishery (for a stratum)

$$p_i = \frac{C_i}{\sum C_k} = \frac{N_i^* \alpha_i q_i f \ell_i}{\sum N_k^* \alpha_k q_k f \ell_k} = \frac{N_i^* \alpha_i q_i \ell_i}{\sum N_k^* \alpha_k q_k \ell_k}$$

- Assume $q_1 = q_2 = q_3 = \dots$

$$p_i = \frac{N_i^* \alpha_i \ell_i}{\sum N_k^* \alpha_k \ell_k}$$

- Variation in p_i across strata (at time t)

  numerator: stock i

  denominator: other stocks, e.g. 50/100 vs 50/1000

Factors affecting stock proportion

- Size-selective fishery (for a stratum)

$$p_i = \frac{C_i}{\sum C_k} = \frac{N_i^* \alpha_i q_i f \ell_i}{\sum N_k^* \alpha_k q_k f \ell_k} = \frac{N_i^* \alpha_i q_i \ell_i}{\sum N_k^* \alpha_k q_k \ell_k}$$

- Assume $q_1 = q_2 = q_3 = \dots$

$$p_i = \frac{N_i^* \alpha_i \ell_i}{\sum N_k^* \alpha_k \ell_k}$$

- Variation in p_i across strata (at time t)

  numerator: stock i

  denominator: other stocks, e.g. 50/100 vs 50/1000

Some potential uses of stock proportion estimates

Quota management: $\hat{p}_i/\hat{\ell}_i$

Some potential uses of stock proportion estimates

Quota management: $\hat{p}_i/\hat{\ell}_i$

Cohort analysis: $[\widehat{C}\hat{p}_i/\hat{\ell}_i]$

Some potential uses of stock proportion estimates

Quota management: $\hat{p}_i / \hat{\ell}_i$

Cohort analysis: $[\widehat{C} \hat{p}_i / \hat{\ell}_i]$

Seasonal management: $[\widehat{C} \hat{p}_i / \hat{\ell}_i] \ / \ d$

Some potential uses of stock proportion estimates

Quota management: $\hat{p}_i/\hat{\ell}_i$

Cohort analysis: $[\widehat{C}\hat{p}_i/\hat{\ell}_i]$

Seasonal management: $[\widehat{C}\hat{p}_i/\hat{\ell}_i] \;/\; d$

Distribution studies: $[\widehat{C}\hat{p}_i/\hat{\ell}_i] \;/\; \hat{f}$

Some potential uses of stock proportion estimates

Quota management: $\hat{p}_i / \hat{\ell}_i$

Cohort analysis: $[\widehat{C}\hat{p}_i / \hat{\ell}_i]$

Seasonal management: $[\widehat{C}\hat{p}_i / \hat{\ell}_i] \ / \ d$

Distribution studies: $[\widehat{C}\hat{p}_i / \hat{\ell}_i] \ / \ \hat{f}$

- p itself not necessarily quantity of interest
- Compounded estimates compound error

Summary

- Binomial-based n formulas useful for planning

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling
- Addresses sampling errors only—not assignment errors

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling
- Addresses sampling errors only—not assignment errors
- Suite of metrics to consider

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling
- Addresses sampling errors only—not assignment errors
- Suite of metrics to consider
- Minimum n depends on statistical objectives

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling
- Addresses sampling errors only—not assignment errors
- Suite of metrics to consider
- Minimum n depends on statistical objectives
- Low $CV(\hat{p})$ for $p < 0.02$ probably not achievable

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling
- Addresses sampling errors only—not assignment errors
- Suite of metrics to consider
- Minimum n depends on statistical objectives
- Low $CV(\hat{p})$ for $p < 0.02$ probably not achievable
- For age-specific stock composition, treat stock-age components as separate stocks

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling
- Addresses sampling errors only—not assignment errors
- Suite of metrics to consider
- Minimum n depends on statistical objectives
- Low $CV(\hat{p})$ for $p < 0.02$ probably not achievable
- For age-specific stock composition, treat stock-age components as separate stocks
- Variation in p_i across strata can be due to variation in local abundance of stock i, or variation in local abundance of other stocks

Summary

- Binomial-based n formulas useful for planning
- Assumes random sampling
- Addresses sampling errors only—not assignment errors
- Suite of metrics to consider
- Minimum n depends on statistical objectives
- Low $CV(\hat{p})$ for $p < 0.02$ probably not achievable
- For age-specific stock composition, treat stock-age components as separate stocks
- Variation in p_i across strata can be due to variation in local abundance of stock i, or variation in local abundance of other stocks
- Compounded estimates compound error