

***Birds of Feather Flock Together:  
Implications of intra-cluster  
correlation on Analysis of Size  
Composition Data***

Catch Comparison Workshop  
University of Rhode Island  
Dec 8-9, 2004

*Paul Rago*  
*National Marine Fisheries Service*  
*Woods Hole, MA 02543*

# Objectives

- Describe intra-cluster correlation
- Implications for Kolmogorov Smirnov Tests
- Implications for precision of selectivity model estimates in SELECT (Share Each Lengths Contribution to Total)

# Intra-Cluster Correlation (1)

Defined in Cochran 1977, p 209, but earlier references include Hansen et al. 1953. Sample survey methods and theory. Vol 1 & 2. Wiley

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left( \frac{N-1}{N} \right) [1 + (n-1)\rho_w]$$

Intra-cluster correlation **INCREASES** the variance of estimate compared to that which would be obtained from a simple random sample.

At the extreme, if  $\rho_w=1$  then

$$V(\bar{y}_{sy}) = \frac{S^2}{n} \left( \frac{N-1}{N} \right) [1 + (n-1)]$$

$$V(\bar{y}_{sy}) = S^2 \left( \frac{N-1}{N} \right)$$

$\therefore$  effective sample size = 1.

# Intra-Cluster Correlation (2)

- Introduced to fisheries applications by Pennington and Volstad 1994. Biometrics 50:725-732
- Examination of Tow Duration
- Assume  $n$  stations and  $m_i$  fish caught at each station and length  $x_{i,j}$  is measured at each station
- The mean  $x$  is estimated as

$$\bar{x}_r = \frac{\sum_{i=1}^n \sum_{j=1}^{m_i} x_{i,j}}{\sum_{i=1}^n m_i}$$

- Variance of  $\bar{x}_r$  is

$$V(\bar{x}_r) \approx \sigma_x^2 \left( \frac{1 + (\bar{M} - 1 + \frac{\sigma_m^2}{\bar{M}}) \rho}{\bar{M}n} \right)$$

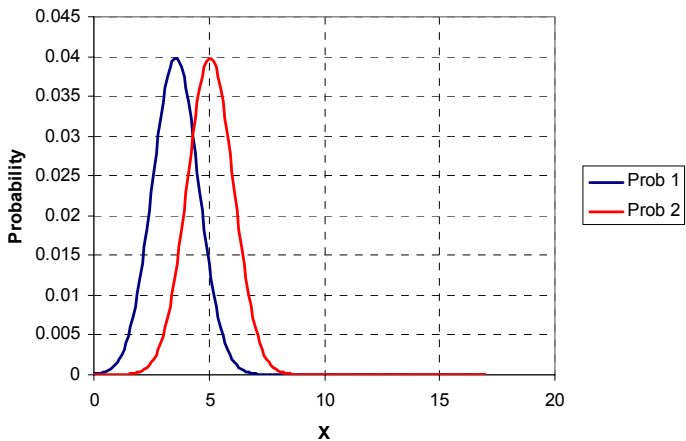
# **Intra-Cluster Correlation (3)**

- *Pennington, Burmeister, Hjellvik 2002 Fish. Bull. 100:74-80* examined effects on size composition.
- Formulas are complicated but essentially boil down to equating the variance of mean length taken over all tows and measurements within tows to the variance of the composite length frequency distribution (p 75)
- Effective sampling size is simply the ratio of latter to the former.
- Applied method to Northeast Arctic cod (1995-1999), Northeast Arctic haddock (1995-1999), deepwater hake (Namibia, 1985-99)
- **Range of effective sample size per station (0.1, 6.3)**

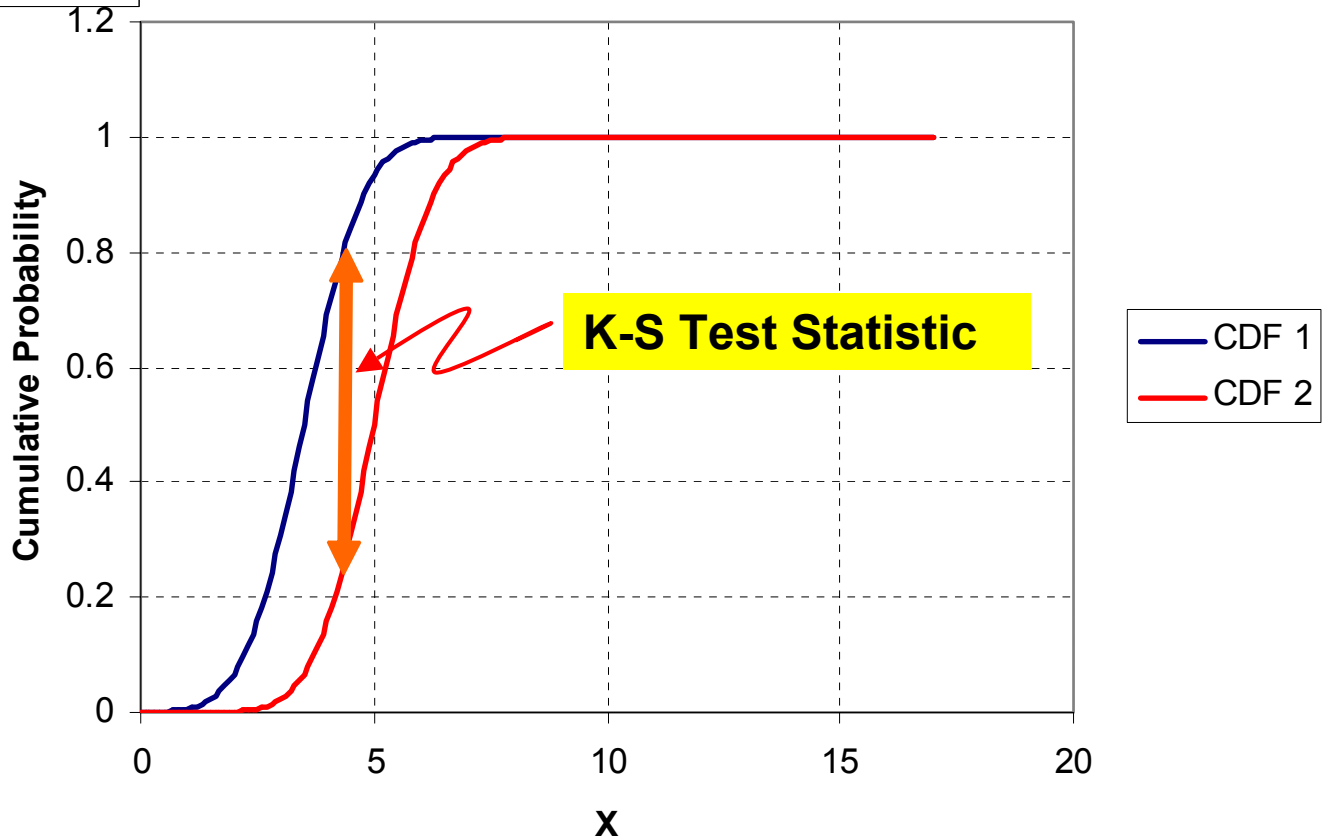
## **Implications of Effective Sample Size on the Kolmogorov Smirnov Test**

- K-S test is based on the maximum difference between the cumulative distribution functions of the control and treatment populations.
- The range of test statistics is between 0 and 1.
- The large sample approximation of the critical value at  $\alpha=0.05$  is  $1.92/\sqrt{n}$

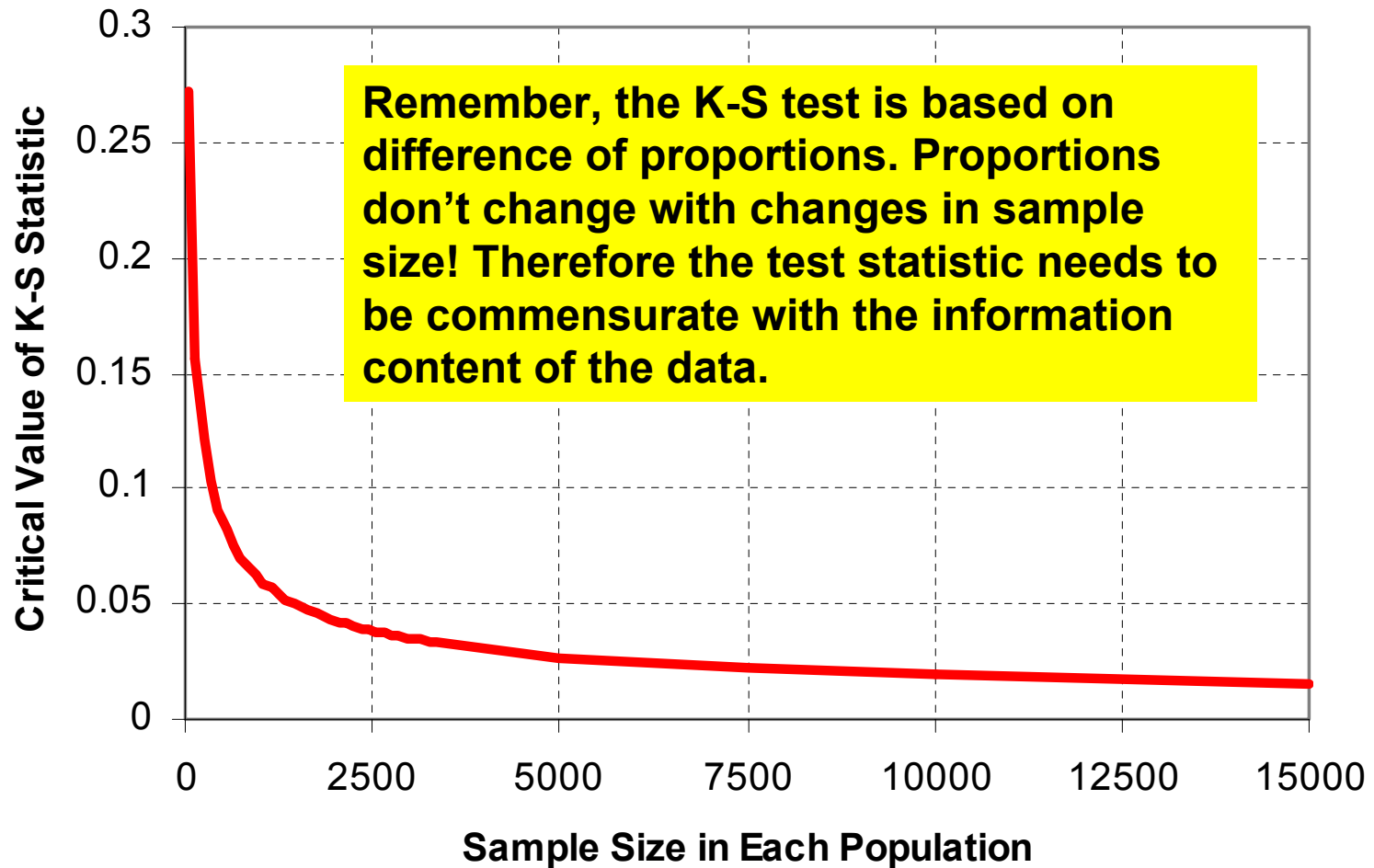
Comparison of Normal Distributions



Comparison of Cumulative Distribution Functions



***Critical Value for two Sample Kolmogorov-Smirnov Test with equal sample size (alpha=0.05)***



# *Implications of Intra-cluster correlation for Gear Selectivity*

- Millar's (1992) SELECT method has become the de facto standard for analysis of gear selectivity study.
- Uses a conditional likelihood approach that distinguishes between the relative fishing intensity of a type of gear ( $p$ ) and the parameters which define the selectivity of capture ( $a$  and  $b$  in the standard logistic curve).
- General approach--fit a function to the ratio of catches in gear 1 to the total catch in gear 1 and 2
- Millar's important contribution to selectivity was to recognize not only proper statistical properties of the conditioned ratio, but also to incorporate the difference in relative fishing intensity ( $p$ ).

# General Equations for SELECT

General Selection Curve to describe probability of capture after contact with the gear.

$$r(L) = \frac{e^{a+bL}}{1 + e^{a+bL}}$$

Relate the observed fraction caught at length to the modeled fraction

$$\frac{n_{L,treatment}}{n_{L,treatment} + n_{L,control}}$$



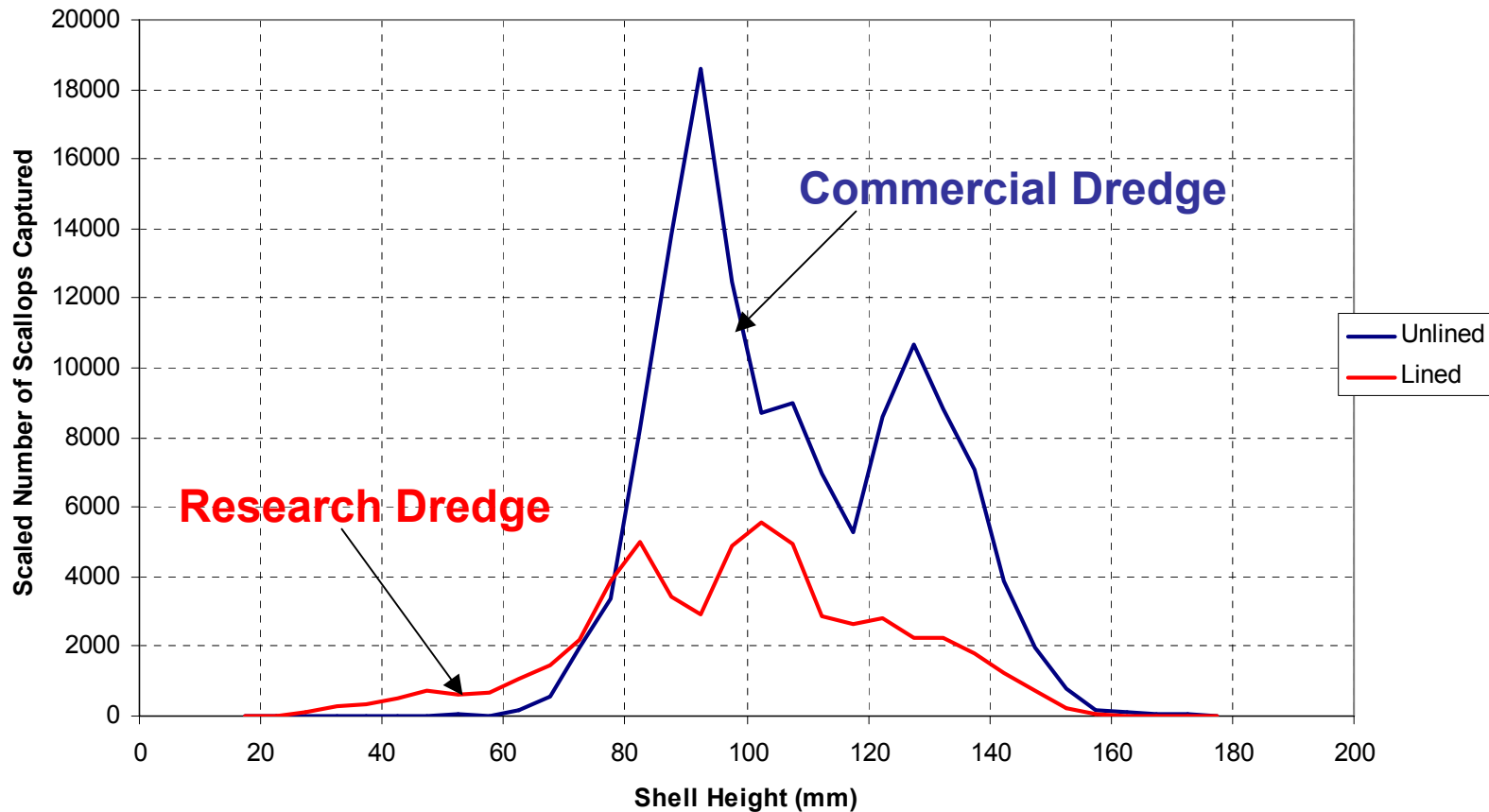
$$\phi(L) = \frac{p r(L)}{(1 - p) + p r(L)}$$

By pluggation

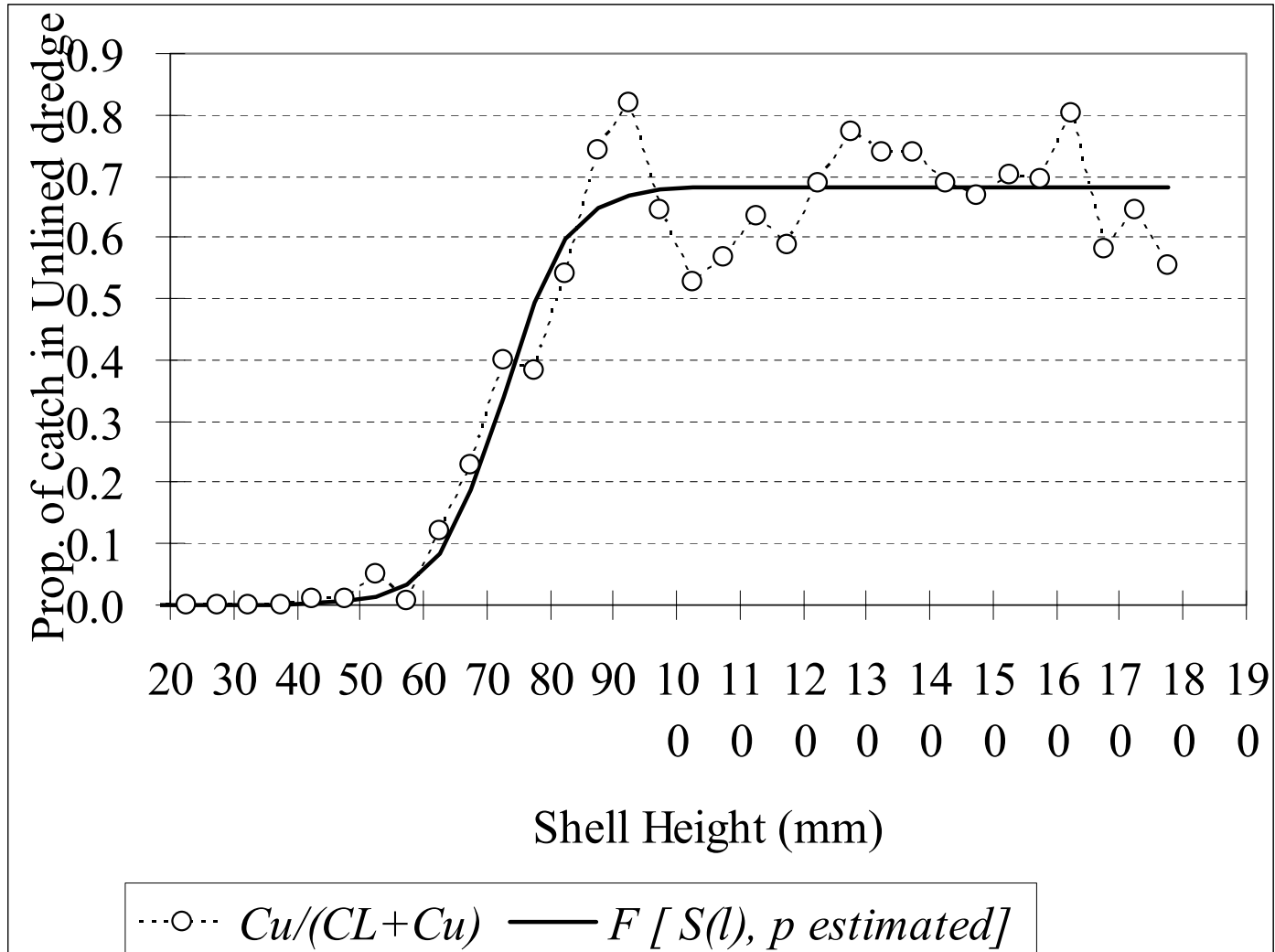
$$\phi(L) = \frac{p r(L)}{(1 - p) + p r(L)} = \frac{p \frac{e^{a+bL}}{1 + e^{a+bL}}}{(1 - p) + p \frac{e^{a+bL}}{1 + e^{a+bL}}} = \frac{p e^{a+bL}}{(1 - p) + e^{a+bL}}$$

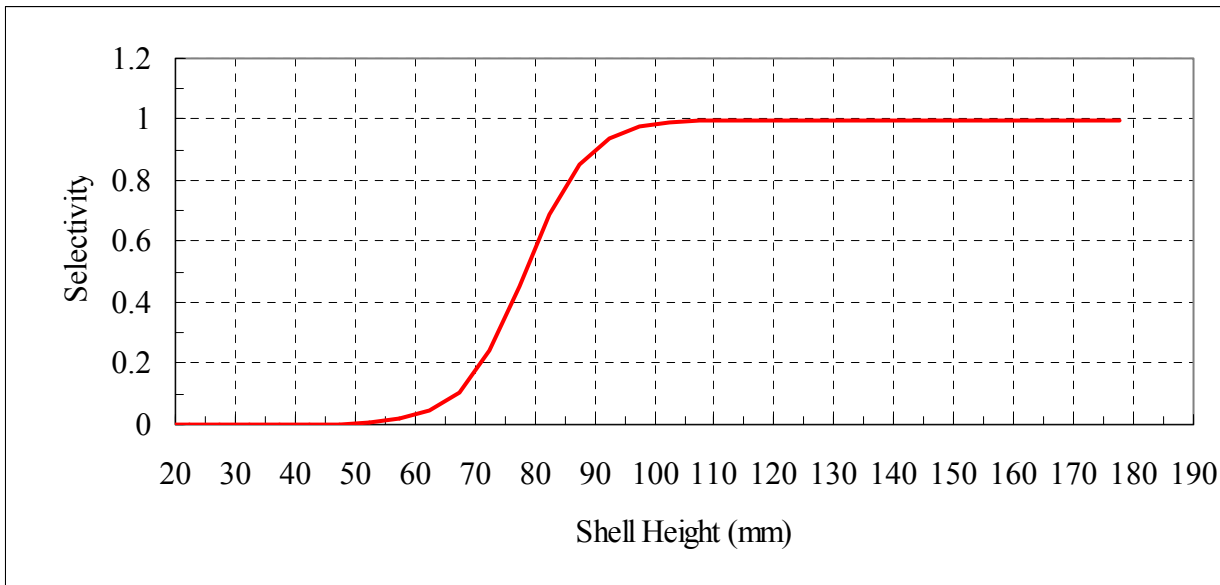
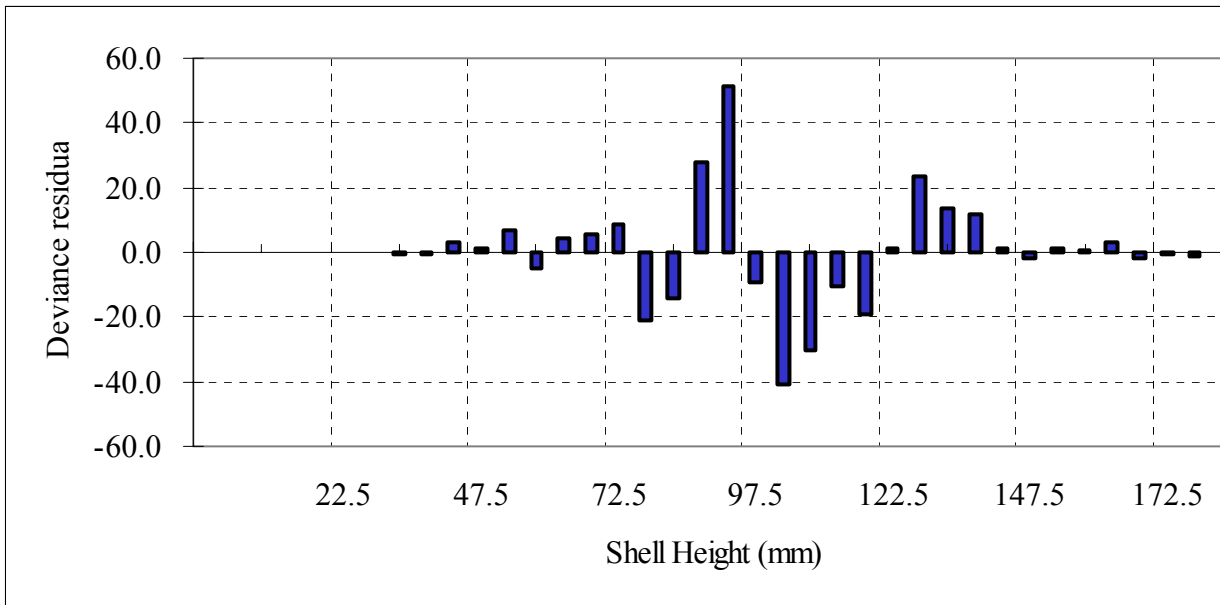
# Return to the F/V Tradition Experiment

Comparison of Scallop Catch Rates in R/V Lined vs F/V Unlined Dredges on F/V Tradition 1999. Original Data



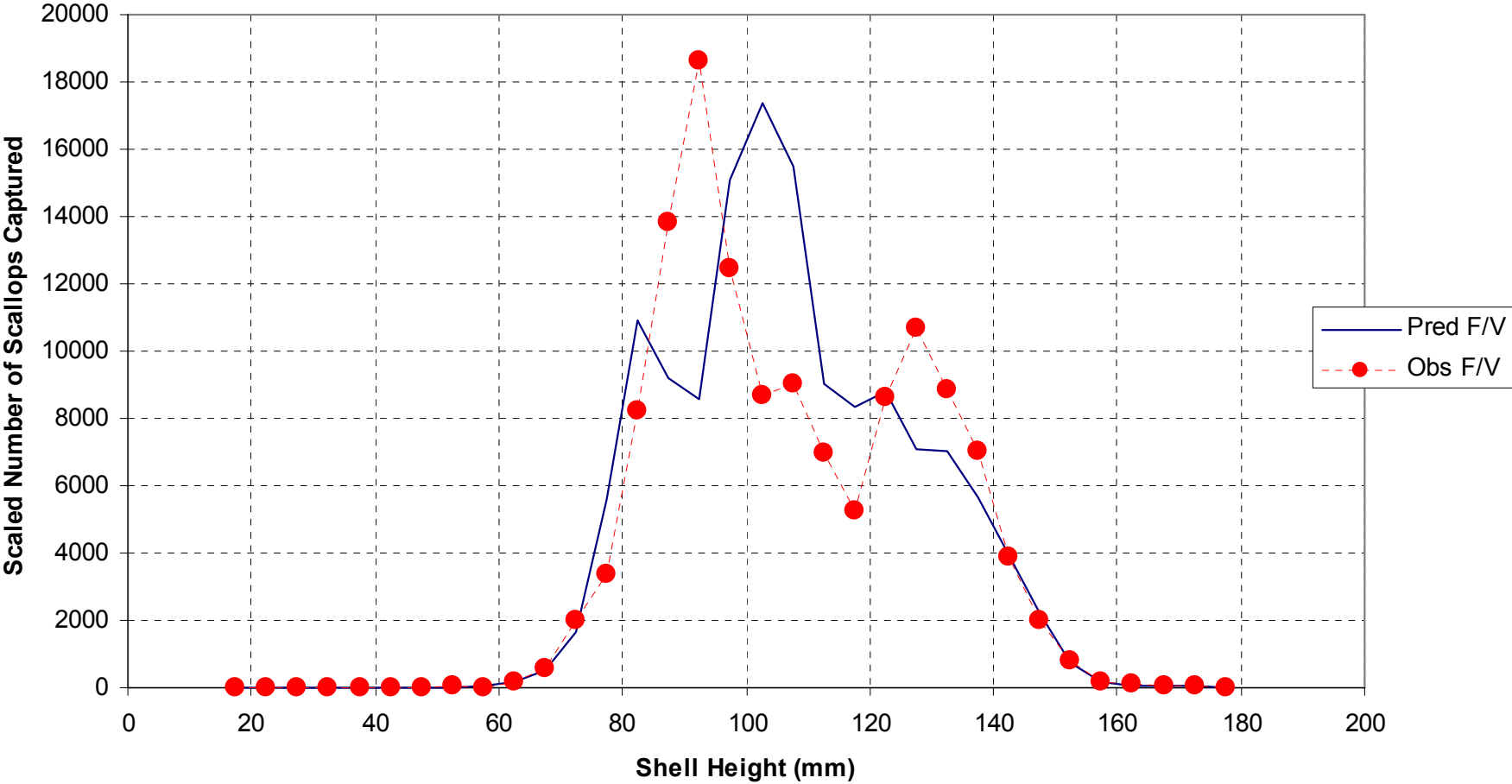
**Model Fit results for Scallop catch data in R/V and F/V Dredges. Data are taken from F/V Tradition Experimental Tows in 1999 and are fit using the SELECT model of Millar (1992).**





# Comparison of Obs vs Pred

**Comparison of Scallop Observed and Predicted Catch Rates in R/V Lined vs F/V Unlined Dredges (0n F/V Tradition 1999). Catches reduced by a multiplier =1.000**



# Estimated Parameters Assuming NO Intra-cluster correlation

Logistic equation

$$S(l) = \exp(a + bl) / [1 + \exp(a + bl)]$$

Standard errors

$$a = -15.24005826 \quad 0.251999234$$

$$b = 0.194261917 \quad 0.003408537$$

$$\text{Split parameter } p = 0.6828888 \quad 0.001253678$$

$$L50\% = 78.451$$

$$0.150703023$$

$$S.R. = 11.311$$

$$0.198457312$$

# Estimated Parameters Assuming Effective sample size of about 1 per Tow

Logistic equation

$$S(l) = \exp(a + bl) / [1 + \exp(a + bl)]$$

Standard errors

$$a = -15.24736251 \quad 7.973634723$$

$$b = 0.194358834 \quad 0.107851639$$

$$\text{Split parameter } p = 0.682881835 \quad 0.039639908$$

**Compare  
to 0.0012**

$$\text{L50\%} = 78.450$$

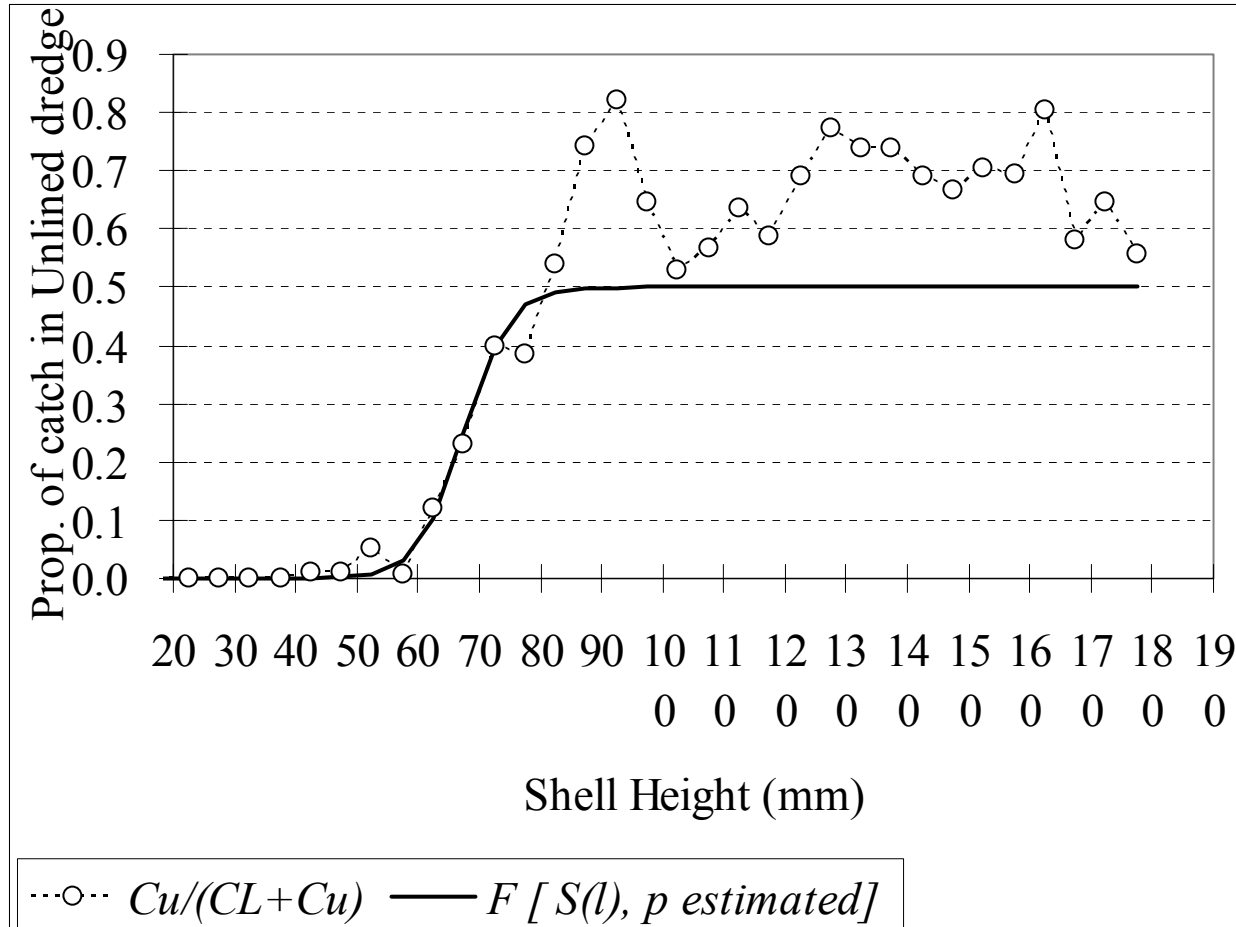
$$4.764089778$$

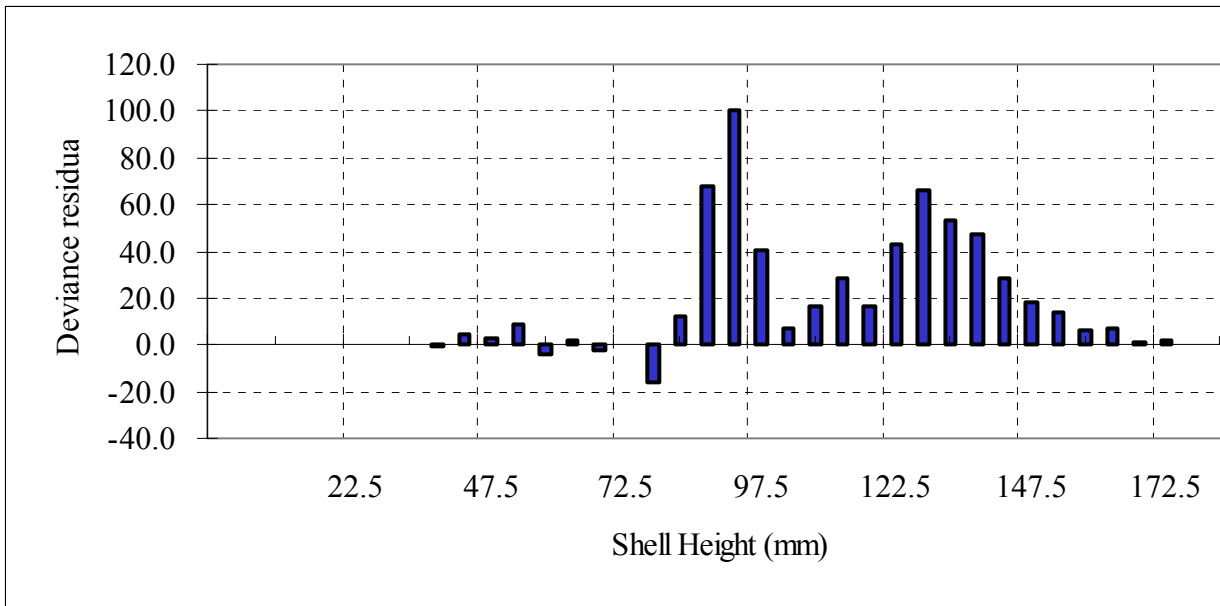
**Compare  
to 0.15**

$$\text{S.R.} = 11.305$$

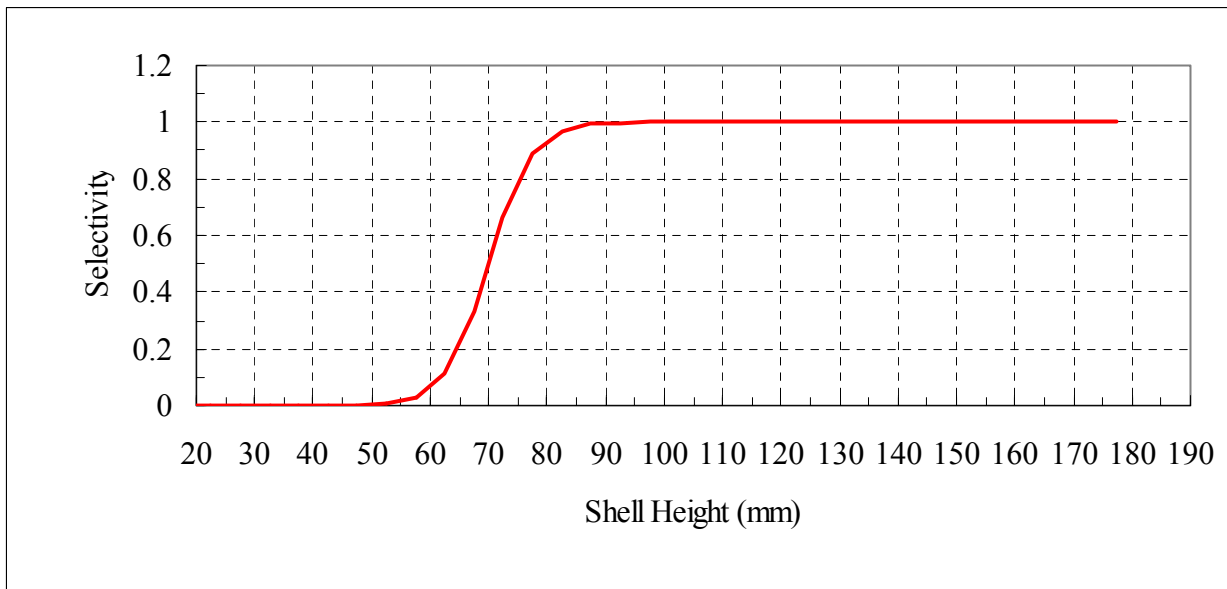
$$6.273250499$$

# SELECT model with fixed $p=0.5$





**The model fits rather poorly.**



**The estimated L50 is 70 mm as compared to 78 mm when the split is estimated.**

# Summary

- Intra-cluster correlation has major implications for evaluation of gear selectivity.

Over-estimating the information content of the data leads to:

- *Increased likelihood of rejecting the null hypothesis when it is true (Type 1 error)*
- *Overestimates of precision for selectivity parameters*