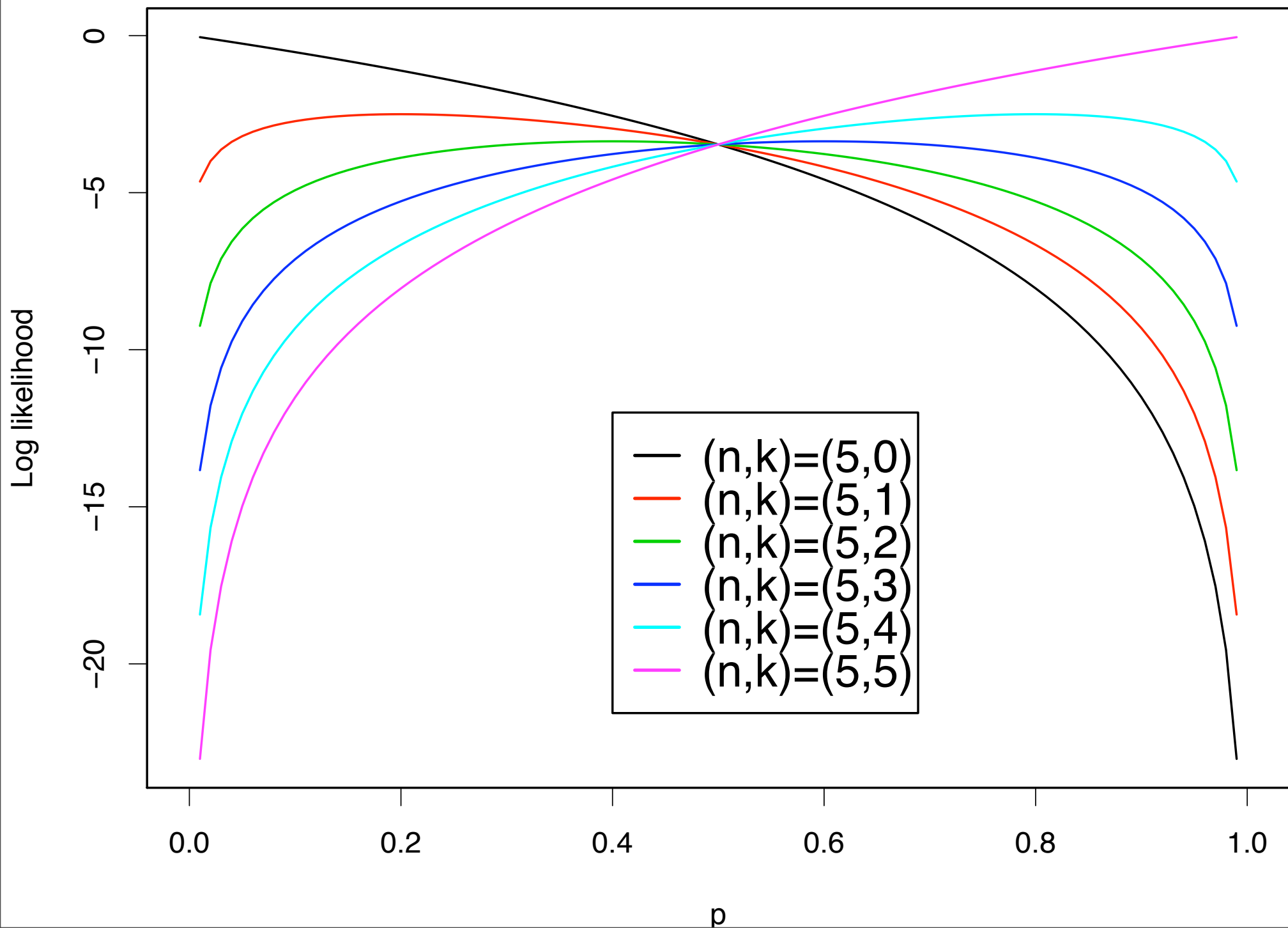


# Estimation

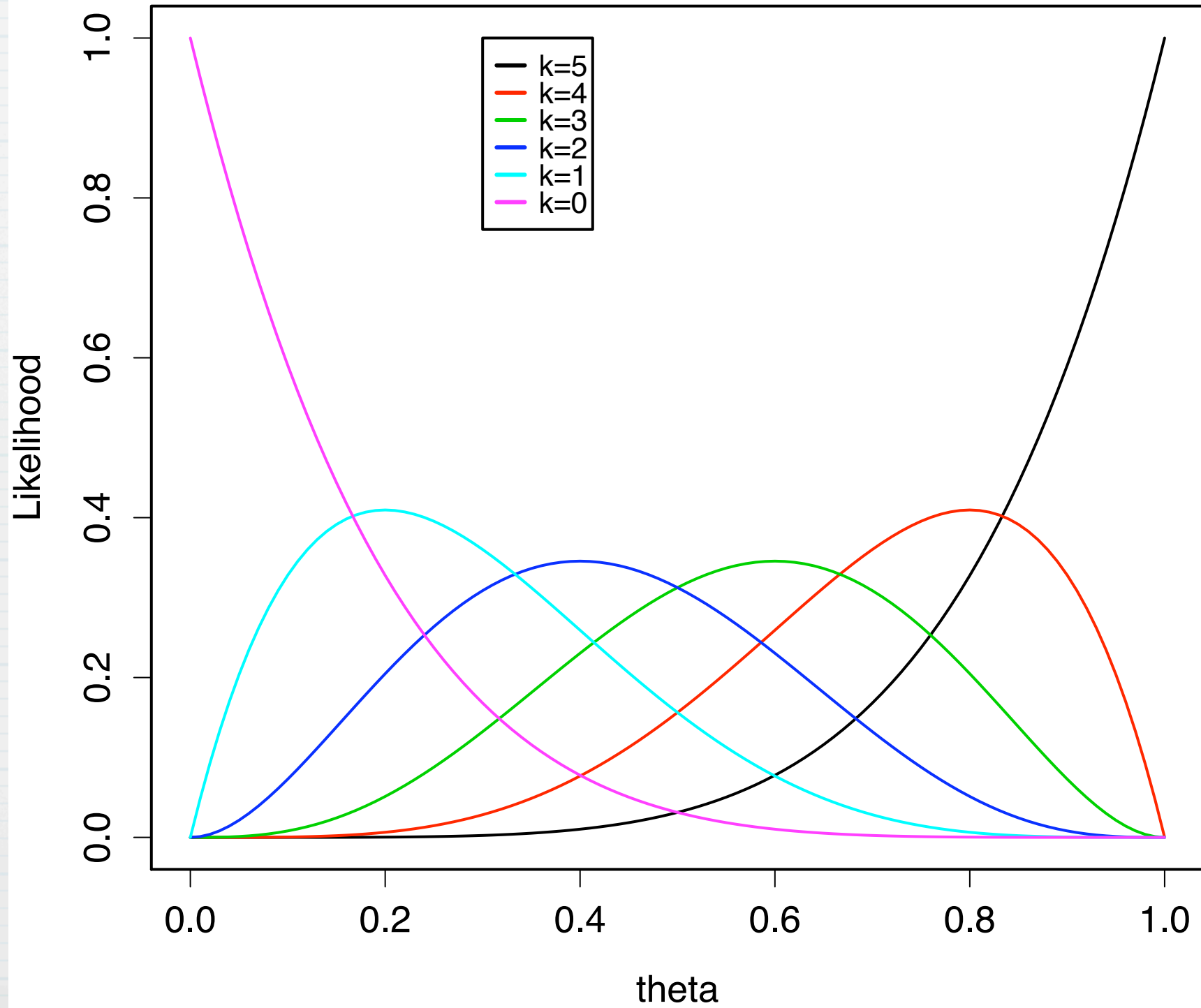
---

2/8/06

Log likelihood for binomial model



Likelihood for binomial model,  $n=5$



# Hardy-Weinberg Equilibrium

- \* Three possible genotypes:  $aa$ ,  $Aa$ ,  $AA$ .  
Each person has one genotype.
- \* Theoretically, should have proportions  $\theta^2$ ,  $2\theta(1-\theta)$ ,  $(1-\theta)^2$ .
- \* In a sample of 1029 Chinese in Hong Kong, measured 342, 500, 187. (Genotypes in this case named  $N$ ,  $MN$ ,  $M$ .)
- \* Goal: Estimate  $\theta$ .

# Reparametrize

aa	Aa	AA	total
342	500	187	1029

Approach 1:  $X_i$  is 1 if person  $i$  has genotype aa, 0 otherwise.

$$\mu_1 = E X_i = \theta^2$$

$$\hat{\theta} = \sqrt{\hat{\mu}_1} = \sqrt{342/1029} = 0.5765$$

# Reparametrize

aa	Aa	AA	total
342	500	187	1029

Approach 1:  $X_i$  is 1 if person  $i$  has genotype aa, 0 otherwise.

$$\mu_1 = E X_i = \theta^2$$

$$\hat{\theta} = \sqrt{\hat{\mu}_1} = \sqrt{342/1029} = 0.5765$$

Approach 2:  $X_i$  is 1 if person  $i$  has genotype Aa, 0 otherwise.

$$\mu_1 = E X_i = (1 - \theta)^2$$

$$\hat{\theta} = 1 - \sqrt{\hat{\mu}_1} = 1 - \sqrt{187/1029} = 0.5737$$

# Reparametrize

aa	Aa	AA	total
342	500	187	1029

Approach 1:  $X_i$  is 1 if person  $i$  has genotype aa, 0 otherwise.

$$\mu_1 = E X_i = \theta^2$$

$$\hat{\theta} = \sqrt{\hat{\mu}_1} = \sqrt{342/1029} = 0.5765$$

Approach 2:  $X_i$  is 1 if person  $i$  has genotype Aa, 0 otherwise.

$$\mu_1 = E X_i = (1 - \theta)^2$$

$$\hat{\theta} = 1 - \sqrt{\hat{\mu}_1} = 1 - \sqrt{187/1029} = 0.5737$$

Approach 3:  $X_i$  is 1 if person  $i$  has genotype AA, 0 otherwise.

$$\mu_1 = E X_i = 2\theta(1 - \theta)$$

$$\hat{\theta} = \frac{1}{2}(1 + \sqrt{1 - \hat{\mu}_1}) = \frac{1}{2}(1 + \sqrt{1 - 500/1029}) = 0.5839.$$

We have 4 observations from a normal model

0.1532655 2.3245976 1.6162693 1.1856458

Estimate the parameters for

1. Location normal model with  $\sigma=1$ .
2. Location-Scale Normal model.

# Location Normal Model

One parameter:  $\theta = \mu$ . Data are independent, Normally distributed with expectation  $\mu$  and variance  $\sigma^2 (= 1$  in this case).

$$\begin{aligned} L(\theta | X_1, \dots, X_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i - \mu)^2 / 2\sigma^2} \\ &= (2\pi)^{-n/2} \sigma^{-n} e^{-\sum_{i=1}^n (X_i - \mu)^2 / 2\sigma^2} \end{aligned}$$

$$\begin{aligned} \ell(\theta | X_1, \dots, X_n) &= -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum (X_i - \mu)^2 \\ &= -\frac{n}{2} \log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum (X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= -\frac{n}{2} \log(2\pi\sigma) - n \log \sigma - \frac{1}{2\sigma^2} \sum (X_i - \bar{X})^2 + \frac{n}{\sigma^2} (\bar{X} - \mu)^2 \end{aligned}$$

$$\begin{aligned} \ell(\mu, \sigma | X_1, \dots, X_n) &= \frac{n}{\sigma^2} (\bar{X} - \mu)^2 \\ &\quad - \frac{1}{2\sigma^2} \sum (X_i - \bar{X})^2 - \frac{n}{2} \log(2\pi\sigma^2) \end{aligned}$$

Maximum when  $\mu = \bar{X}$ . MLE  $\hat{\mu} = \bar{X}$

The pair (Mean( $X_i$ ), SD( $X_i$ )) is a sufficient statistic.

What is the sampling distribution for the estimator?

Is the estimator any good?

What does that tell us?

# Sampling distribution

$\hat{\mu} = \bar{X}$  has  $N(\mu, \sigma^2/n)$  distribution.

MSE (Mean-squared error):

$$\mathbb{E}[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}.$$

# What is a good estimator?

\* Unbiased:  $E[\hat{\mu}] = \mu$ .

\* Minimum MSE: We'd like it to be true that  $E[(\hat{\mu} - \mu)^2] \leq E[(\hat{\mu}' - \mu)^2]$  for any other estimator  $\hat{\mu}'$

\* Weaker: An estimator for  $\mu$  is admissible if for any alternative estimator there is some  $\mu$  for which

$$E_{\mu}[(\hat{\mu} - \mu)^2] \leq E_{\mu}[(\hat{\mu}' - \mu)^2]$$

That is, there's no other estimator which "dominates"  $\hat{\mu}$ .  
i.e., that is always better, no matter what the reality.

# Surprising digression

- \* Proving admissibility is difficult.
- \* Suppose we observe  $X_1, \dots, X_n, Y_1, \dots, Y_n, Z_1, \dots, Z_n$ , independent from normal distributions with variance 1 and expectations  $\mu_x, \mu_y, \mu_z$ .
- \* The estimator  $(\bar{X}, \bar{Y}, \bar{Z})$  is not admissible
- \* For more details, come to my Math Club talk on March 8.

# Confidence Interval

$Z := \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$  has standard normal distribution.

$$\mu = \bar{X} - \frac{\sigma}{n} Z$$

$$P\left\{\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right\} = \alpha,$$

where  $z_{\alpha}$  is chosen so that  $\Phi(-z_{\alpha}) = \alpha$ .

We have 4 observations from a normal model with known  $\sigma=1$ .

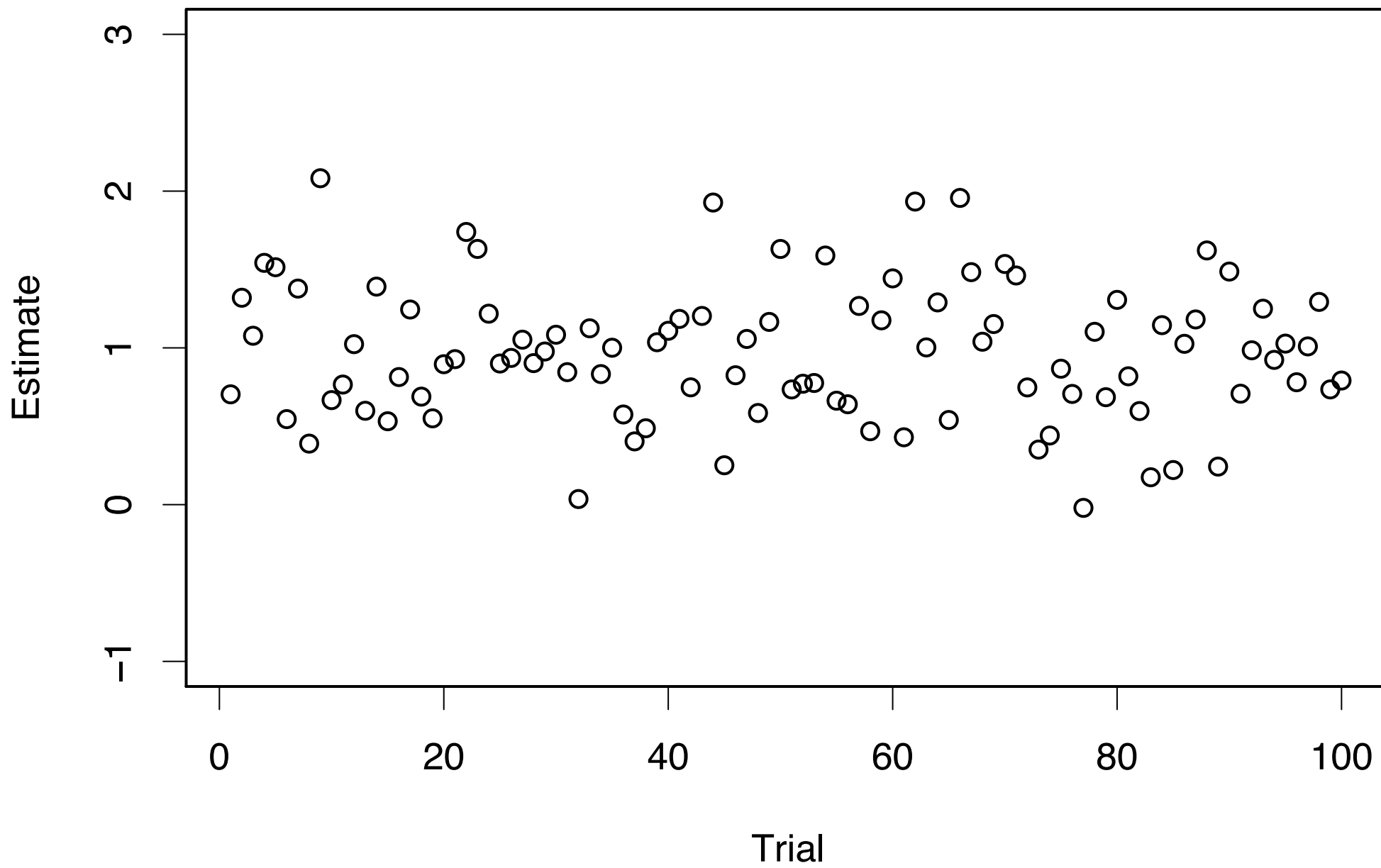
0.1532655 2.3245976 1.6162693 1.1856458

Estimate  $\hat{\mu} = \bar{x} = 1.32$ .

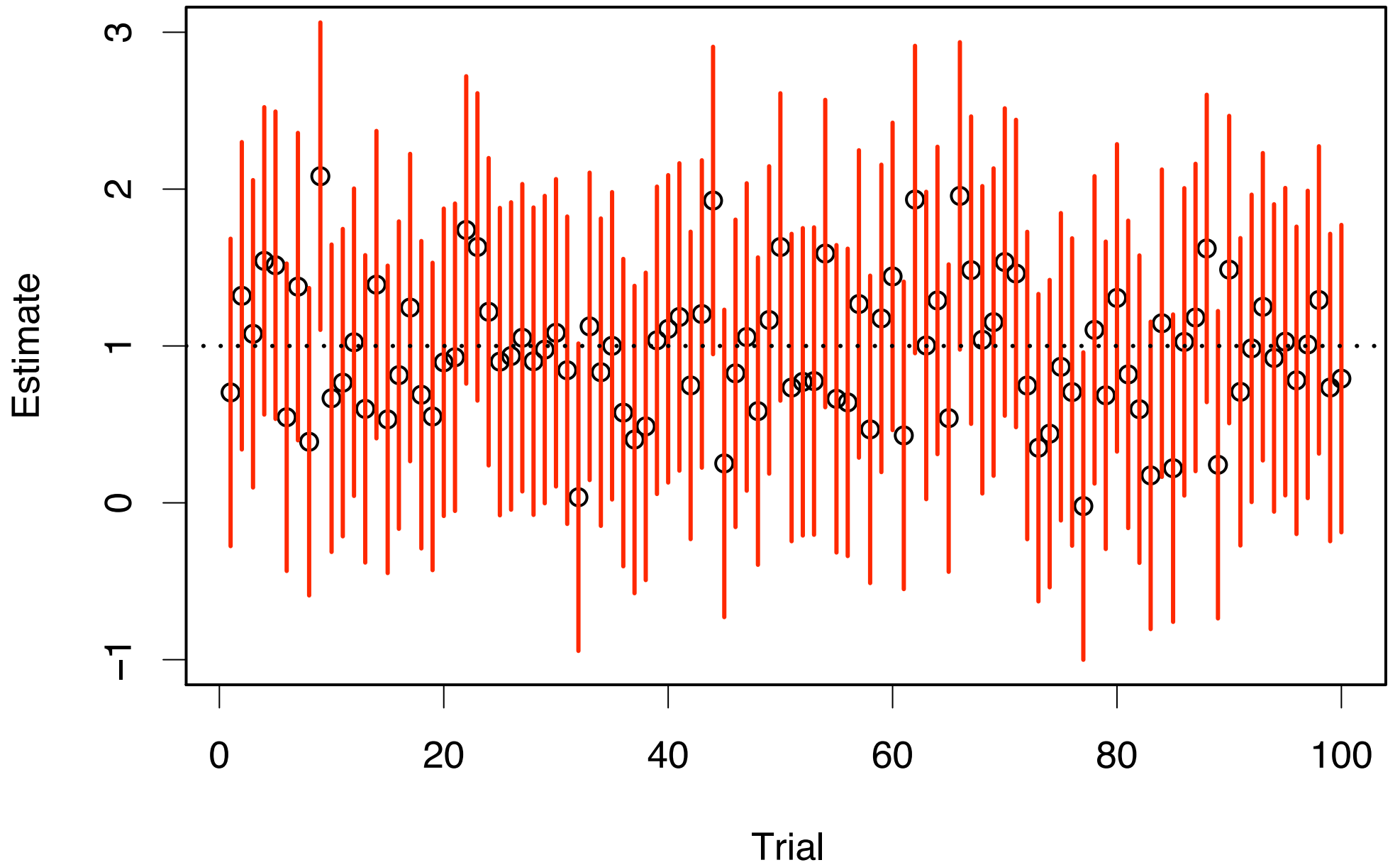
95% confidence interval:  $1.32 \pm (1/2)1.96 = (.34, 2.3)$

$z_{0.025} = 1.96$

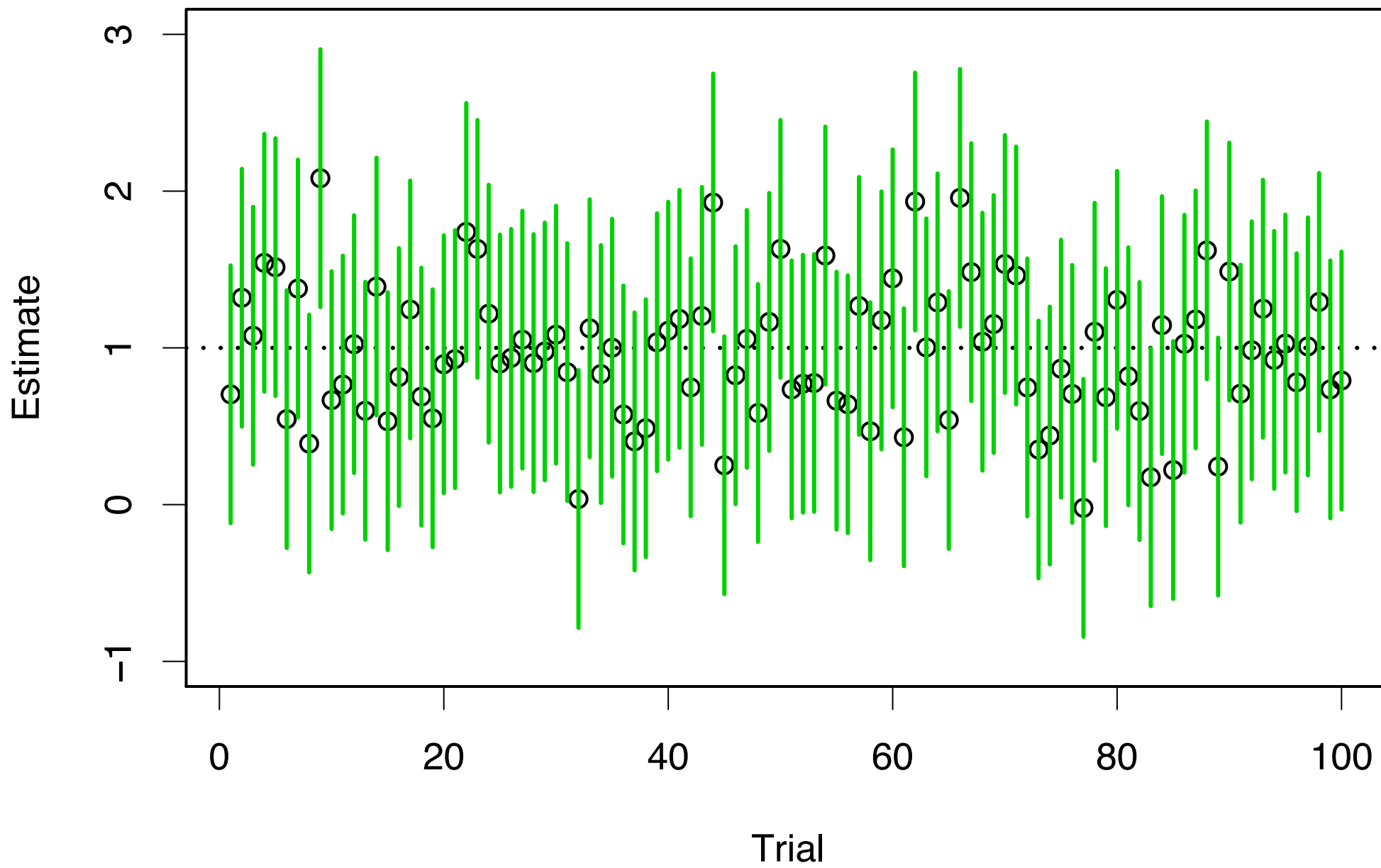
### 100 Trials of averaging 4 N(1,1)



# 95% confidence intervals



# 90% confidence intervals



# 68% confidence intervals

