

# Statistical Inference

Kwame Kankam  
Department of Statistics  
Penn State University

Adapted from notes prepared by James L Rosenberger,  
Penn State University

Summer School in Statistics for Astronomers

# Overview of Statistical Inference

- Given a sample from a population, how do we make inferences about the population?
  - Estimation
    - Point estimation
    - Interval estimation
  - Hypothesis testing

# Overview of Statistical Inference

- Some classical problems of statistical inference:
  - Tests and confidence intervals for an unknown population mean (one sample problem).
  - Tests and confidence intervals for the difference of two population means (two sample problem).
  - Tests for equality of several means (analysis of variance).
  - Tests for equality of several variances.
  - Chi-square goodness-of-fit test.

# A Motivating Example

- Van den Bergh (1985, ApJ 297, p. 361) considered the luminosity function (LF) for globular clusters in various galaxies.
- V-d-B's conclusion: The LF for clusters in the Milky Way is adequately described by a normal distribution (with pdf:)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

- $M_0 \equiv \mu$  : Mean visual absolute magnitude
- $\sigma$  : Std. deviation of visual absolute magnitude

# A Motivating Example

- Here is a diagram from van den Bergh (1985), providing complete data for the M31 Globular Clusters in the Milky Way. (Notice that the data *appear* to be non-Gaussian)

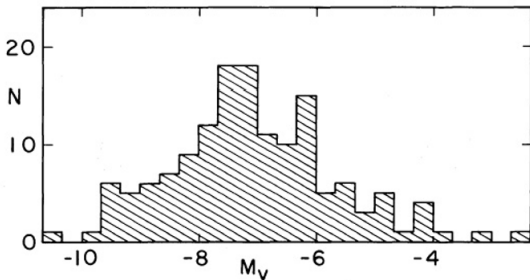


FIG. 2.—Luminosity function of Galactic globular clusters (one object at  $M_V = -1.7$  is not plotted). Note that the luminosity function is asymmetrical with a long tail extending to faint magnitudes.

# A Motivating Example:

- From a statistical viewpoint, the main questions of interest are:
  - ① On the basis of collected data, estimate the parameters  $\mu$  and  $\sigma$ . Also, derive a plausible range of values for each parameter; etc.
  - ② V-d-B, etc., conclude that the LF is “adequately described” by a normal distribution.  
How can we quantify the plausibility of their conclusion?

## Some Definitions

- $X$ : A random variable.
- *Population*: The collection of all values of  $X$ .
- $f(x; \theta)$ : The prob. density function (p.d.f.) of  $X$ . For this lecture, we take  $\theta = (\theta_1, \dots, \theta_k)$ . Thus we have  $\theta = (\theta_1, \theta_2) = (\mu, \sigma)$  for the p.d.f. of the LF for Globular Clusters.
- *Statistical model*: A choice of p.d.f. for  $X$ . We choose a model which “adequately describes” data collected on  $X$ .
- *Parameter*: A number which describes a property of the population.  $\theta_1, \dots, \theta_k$  are parameters.
- *Parameter space*: The set of permissible values of the parameters. The parameter space for the p.d.f. of the LF example is

$$\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$$

# Population Parameter vs Sample Statistic

- *Random sample*: Mutually independent random variables  $X_1, \dots, X_n$  which all have the same distribution as  $X$

## Difference between parameter and statistic

*Parameter*: A number computable only from the entire population

*Statistic*: A number computed from the random sample  $X_1, \dots, X_n$

- Some examples of statistics are

the sample mean:  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ , and

the sample variance:  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

- In general, a *statistic* is a function  $Y = t(X_1, \dots, X_n)$  of the observations.
- *Sampling distribution*: The probability distribution of a statistic



# Estimation

- Consider a random variable  $X$  with density  $f(x, \theta)$ .
- Let  $\theta$  be an unknown parameter.
- Once  $\theta$  is known, the density function is completely specified.

## The problem of estimation

Given a random sample  $x_1, \dots, x_n$  from a population with above density, how to estimate the value of the unknown parameter  $\theta$ .

- A statistic  $t(X_1, \dots, X_n)$  whose value is used to represent or estimate the unknown parameter  $\theta$  is called a point estimator or estimator for short.
- $\hat{\theta}$  will represent an estimator.
- We distinguish between an estimator (a random variable) and an estimate (a specific value of the estimator).

## Return to LF Example

- $X$ : LF for globular clusters
- Model:  $N(\mu, \sigma^2)$ , the normal distribution with mean  $\mu$  and variance  $\sigma^2$
- Problem: Given a random sample  $x_1, \dots, x_n$ , estimate  $\mu$
- $\bar{x}$  is a very good estimate of  $\mu$
- $\hat{x}$ , the sample median, is a good plausible estimate of  $\mu$
- $x_{(n)}$ , the largest observed value in the LF, is obviously a poor estimate of  $\mu$ , *since it almost certainly is much larger than  $\mu$ .*
- Statistics like  $\bar{X}$ ,  $\hat{X}$  and  $X_{(n)}$  are examples of *point estimators* of  $\mu$ .

# Point Estimation

- We focus on two concepts:
  - ① Methods for constructing an estimator
    - The method of maximum likelihood
    - The method of moments
    - Bayesian methods
    - Decision-theoretic methods
  - ② Evaluating and choosing estimators from a class

# The Method of Moments

- $X$ : Random variable with p.d.f.  $f(x; \theta_1, \theta_2)$
- Parameters to be estimated:  $\theta_1, \theta_2$
- Random sample:  $X_1, \dots, X_n$ 
  - 1 Calculate the first two sample moments:

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i, \quad m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2$$

- 2 Calculate  $E(X)$  and  $E(X^2)$ , the first two population moments:

$$E(X^k) = \int_{-\infty}^{\infty} x^k f(x; \theta_1, \theta_2) dx$$

The results are in terms of  $\theta_1$  and  $\theta_2$

- 3 Solve for  $\theta_1, \theta_2$  in the simultaneous equations:

$$E(X) = m_1, \quad E(X^2) = m_2$$

The solutions are the *method-of-moments estimators* of  $\theta_1, \theta_2$

# Method of Moments Example: LF for Globular Clusters

- We are given a random sample:  $X_1, \dots, X_n$  from  $N(\mu, \sigma^2)$ 
  - ① Obtain the first two sample moments:

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 = \frac{n-1}{n} S^2 + \bar{X}^2$$

- ② Obtain the first two population moments:

$$E(X) = \int_{-\infty}^{\infty} xf(x; \mu, \sigma^2) dx = \mu$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x; \mu, \sigma^2) dx = \mu^2 + \sigma^2$$

- ③ Solve:  $\hat{\mu} = m_1, \hat{\mu}^2 + \hat{\sigma}^2 = m_2$   
Solution:  $\hat{\mu} = \bar{X}, \hat{\sigma}^2 = m_2 - m_1^2 = \frac{n-1}{n} S^2$

- $X$ : a random variable
- $\theta$  is a parameter
- $f(x; \theta)$ : A statistical model for  $X$
- $X_1, \dots, X_n$ : A random sample from  $X$
- We want to construct good estimators for  $\theta$

# The Method of Maximum Likelihood

- R. A. Fisher (1912), "On an absolute criterion for fitting frequency curves," *Messenger of Math.* 41, 155-160
- Fisher's first mathematical paper, written while a final-year undergraduate in mathematics and mathematical physics at Cambridge University
- It's not clear what motivated Fisher to study this subject; perhaps it was the influence of his tutor, the astronomer F. J. M. Stratton.
- Fisher's paper started with a criticism of two methods of curve fitting, least-squares and the method of moments.

# Maximum likelihood Estimation

- Recall: LF for globular clusters in the Milky Way; van den Bergh's normal model,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

- $\mu$ : Mean visual absolute magnitude
- $\sigma$ : Std. deviation of visual absolute magnitude
- $\bar{X}$  is a good estimator for  $\mu$ ;  $S^2$  is a good estimator for  $\sigma^2$ . (More on this later!)
- We seek a method which produces good estimators automatically
- Fisher's brilliant idea: The method of maximum likelihood



# Maximum Likelihood Estimation

- Choose a globular cluster at random; what is the chance that the LF will be exactly -7.1 mag? Exactly -7.2 mag?
- For any continuous random variable  $X$ ,

$$P(X = c) = 0$$

- Suppose  $X \sim N(\mu = -6.9, \sigma^2 = 1.21)$
- $X$  has probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

- $P(X = -7.1) = 0$ , but

$$\begin{aligned} f(-7.1) &= \frac{1}{1.1\sqrt{2\pi}} \exp \left[ -\frac{(-7.1 + 6.9)^2}{2(1.1)^2} \right] \\ &= 0.37 \end{aligned}$$

# Maximum Likelihood Estimation

- Interpretation: In one simulation of the random variable  $X$ , the “likelihood” of observing the number  $-7.1$  is  $0.37$
- $f(-7.2) = 0.28$
- In one simulation of  $X$ , the value  $x = -7.1$  is 32% more likely to be observed than the value  $x = -7.2$
- $x = -6.9$  is the value which has the greatest (or maximum) likelihood, for it is where the probability density function is at its maximum

# Maximum Likelihood Estimation

- Return to a general model  $f(x; \theta)$
- Random sample:  $X_1, \dots, X_n$
- Recall that the  $X_i$  are independent random variables
- The *joint* probability density function of the sample is

$$f(x_1; \theta)f(x_2; \theta) \cdots f(x_n; \theta)$$

- Here the variables are the  $X$ 's, while  $\theta$  is fixed
- Fishers brilliant idea: Reverse the roles of the  $X$ 's and  $\theta$
- Regard the  $X$ 's as fixed and  $\theta$  as the variable

# Maximum Likelihood Estimation

- The *likelihood function* is

$$L(\boldsymbol{\theta}; X_1, \dots, X_n) = f(x_1; \boldsymbol{\theta})f(x_2; \boldsymbol{\theta}) \cdots f(x_n; \boldsymbol{\theta})$$

- Simpler notation:  $L(\boldsymbol{\theta})$
- $\hat{\boldsymbol{\theta}}$ , the maximum likelihood estimator of  $\boldsymbol{\theta}$ , is the value of  $\boldsymbol{\theta}$  where  $L$  is maximized
- $\hat{\boldsymbol{\theta}}$  is a function of the  $X$ 's
- Usually, it is easier to find the maximum of the logarithm of the likelihood.
- Caution: The MLE is not always unique.

- Example: “ ... cosmic ray composition - The path length distribution ... ”
- $X$ : Length of paths
- Parameter:  $\theta > 0$
- Model: The exponential distribution,

$$f(x; \theta) = \theta^{-1} \exp(-x/\theta), \quad x > 0$$

- Random sample:  $X_1, \dots, X_n$
- Likelihood function:

$$\begin{aligned} L(\theta) &= f(X_1; \theta) f(X_2; \theta), \dots, f(X_n; \theta) \\ &= \theta^{-n} \exp(-(X_1 + \dots + X_n) / \theta) \\ &= \theta^{-n} \exp(-n\bar{X} / \theta) \end{aligned}$$

- To maximize  $L$ , we use calculus
- It is also equivalent to maximize  $\ln L$ :

$$\ln L(\theta) = -n \ln(\theta) - n\bar{X}\theta^{-1}$$

$$\frac{d}{d\theta} \ln L(\theta) = -n\theta^{-1} + n\bar{X}\theta^{-2}$$

$$\frac{d^2}{d\theta^2} \ln L(\theta) = n\theta^{-2} - 2n\bar{X}\theta^{-3}$$

- Solve the equation  $\frac{d}{d\theta} \ln L(\theta) = 0$

$$\theta = \bar{X}$$

- Check that  $\frac{d^2}{d\theta^2} \ln L(\theta) < 0$  at  $\theta = \bar{X}$
- $\ln L(\theta)$  is maximized at  $\theta = \bar{X}$
- Conclusion: The MLE of  $\theta$  is  $\hat{\theta} = \bar{X}$

- LF for globular clusters;  $X \sim N(\mu, \sigma^2)$

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

- Assume that  $\sigma$  is known (1.1 mag, say)
- Random sample:  $X_1, \dots, X_n$
- Likelihood function:

$$\begin{aligned} L(\mu) &= f(X_1; \mu)f(X_2; \mu) \cdots f(X_n; \mu) \\ &= (2\pi)^{-n/2} \sigma^{-n} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right] \end{aligned}$$

- Maximize  $\ln L$  using calculus:  $\hat{\mu} = \bar{X}$

- LF for globular clusters;  $X \sim N(\mu, \sigma^2)$

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$

- Both  $\mu$  and  $\sigma$  are unknown
- A likelihood function of two variables,

$$\begin{aligned} L(\mu, \sigma^2) &= f(X_1; \mu, \sigma^2) \cdots f(X_n; \mu, \sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right] \end{aligned}$$

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2$$

$$\frac{\partial}{\partial \mu} \ln L = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)$$

$$\frac{\partial}{\partial(\sigma^2)} \ln L = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^n (X_i - \mu)^2$$



- Solve for  $\mu$  and  $\sigma^2$  the simultaneous equations:

$$\frac{\partial}{\partial \mu} \ln L = 0, \quad \frac{\partial}{\partial (\sigma^2)} \ln L = 0$$

- We also verify that  $L$  is concave at the solutions of these equations (Hessian matrix)
- Conclusion: The MLEs are

$$\hat{\mu} = \bar{X}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

- $\hat{\mu}$  is unbiased:  $E(\hat{\mu}) = \mu$
- $\hat{\sigma}^2$  is not unbiased:  $E(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$
- For this reason, we use  $\frac{n-1}{n} \hat{\sigma}^2 \equiv S^2$
- We will define the concept of unbiasedness later

- Calculus cannot always be used to find MLEs
- Example: "...cosmic ray composition ..."
- Parameter:  $\theta > 0$
- Model:

$$f(x; \theta) = \begin{cases} \exp [-(x - \theta)], & x \geq \theta \\ 0, & x < \theta \end{cases}$$

- Random sample:  $X_1, \dots, X_n$

$$\begin{aligned} L(\theta) &= f(X_1; \theta) \cdots f(X_n; \theta) \\ &= \begin{cases} \exp (-\sum_{i=1}^n (X_i - \theta)), & \text{all } X_i \geq \theta \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

- $X_{(1)}$ : The smallest observation in the sample
- "all  $X_i \geq \theta$ " is equivalent to " $X_{(1)} \geq \theta$ "

$$L(\theta) = \begin{cases} \exp (-n(\bar{X} - \theta)), & \theta \leq X_{(1)} \\ 0, & \text{otherwise} \end{cases}$$

- Conclusion:  $\hat{\theta} = X_{(1)}$

# The ML Method for Linear Regression Analysis

- Scatterplot data:  $(x_1, y_1), \dots, (x_n, y_n)$
- Basic assumption: The  $x_i$ s are non-random measurements; the  $y_i$  are observations on  $Y$ , a random variable
- Statistical model:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \dots, n$$

- Errors  $\epsilon_1, \dots, \epsilon_n$ : a random sample from  $N(0, \sigma^2)$
- Parameters:  $\alpha, \beta, \sigma^2$
- $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$ : The  $Y_i$ 's are independent
- The  $Y_i$  are not identically distributed, because they have differing means

- The likelihood function is the joint density function of the observed data,  $Y_1, \dots, Y_n$

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right] \\ &= (2\pi\sigma^2)^{-n/2} \exp \left[ -\frac{\sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right] \end{aligned}$$

- Use partial derivatives to maximize  $L$  over all  $\alpha, \beta$  and  $\sigma^2 > 0$  (Wise advice: Maximize  $\ln L$ )
- The ML estimators are:

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

# Invariance Property of the MLE

## Theorem

Let  $\hat{\theta}$  be the MLE of  $\theta$  in the density  $f(x; \theta)$  where  $\theta$  is assumed to be one-dimensional. If  $h$  is a one-to-one function, then the MLE of  $h(\theta)$  is  $h(\hat{\theta})$ .

- Example: Consider the normal density with  $\mu = \mu_0$  known. It can be shown that the MLE of  $\sigma^2$  is

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2.$$

By the invariance property, the MLE of  $\sigma$  is

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2}.$$

# You Can Usually Bet on the MLE!

- The method of maximum likelihood works well when intuition fails and no obvious estimator can be found.
- When an obvious estimator exists the method of ML often will find it.
- The method can be applied to many statistical problems: regression analysis, analysis of variance, discriminant analysis, hypothesis testing, principal components, etc.
- We will return to describe some further properties of the MLE after we have developed some language for comparing estimators.

# Bias of an Estimator

- Let  $\hat{\theta}$  be an estimator for a parameter  $\theta$ . We define the **bias** of the estimator as:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- If we use the sample mean  $\bar{X}$  to estimate the population mean  $\mu$ , then the bias of  $\bar{X}$  is:

$$\text{Bias}(\bar{X}) = E(\bar{X}) - \mu$$

Similarly, if we use the sample variance  $S^2$  to estimate the population variance  $\sigma^2$ , then the bias of  $S^2$  is:

$$\text{Bias}(S^2) = E(S^2) - \sigma^2$$

- Whenever the bias of an estimator is equal to 0, we call the estimator an unbiased estimator for the parameter of interest.
- It is clear that the sample mean is *unbiased* for the population mean. It can be shown that the sample variance is also unbiased for the population variance.

- Example: The Luminosity Function LF for globular clusters
- The sample mean,  $\bar{X}$ , is unbiased:

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

- Similarly, if the LF has a normal distribution then  $\hat{X}$ , the sample median, is an unbiased estimate of  $\mu$  also (based on the symmetry of  $f(x)$ ).
- $X_{(n)}$ : the largest observed LF
- $X_{(n)}$  is not unbiased:  $E(X_{(n)}) > \mu$



# Variance and Standard Error of an Estimator

- Recall that an estimator is a random variable. When the variance of  $\hat{\theta}$  exists, its standard deviation (also called *standard error*) is given by  $\sigma_{\hat{\theta}} = \sqrt{\text{Var}(\hat{\theta})}$ .
- As an example, suppose that we have a random sample from a distribution with mean  $\mu$  and finite variance  $\sigma^2$ . The distribution of  $\bar{X}$  has mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .
- Thus we say that the standard error of the sample mean  $\bar{X}$  is:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Since  $\sigma$  is usually unknown, we can estimate it with the sample standard deviation,  $S$ . In that case, we obtain the estimated standard error of the sample mean  $\bar{X}$  as:

$$\hat{\sigma}_{\bar{X}} = \frac{S}{\sqrt{n}}$$

- We also want estimators which are “close” to  $\theta$
- For any estimator  $\hat{\theta}$ , we define the mean square error (*MSE*) as  $E \left[ \left( \hat{\theta} - \theta \right)^2 \right]$ .
- Choose as our point estimator the estimator for which the *MSE* is smallest
- An estimator  $\hat{\theta}$  which minimizes *MSE* is said to have *minimum mean square error*
- It can be shown that:

$$MSE(\hat{\theta}) = \text{Var}(\hat{\theta}) + (\text{bias}(\hat{\theta}))^2$$

- If  $\hat{\theta}$  is also unbiased then  $MSE = \text{Var}(\hat{\theta})$ , and  $\hat{\theta}$  is a *minimum variance unbiased estimator (MVUE)*

## Comparing Estimators: An Example

### Reminder

If  $R_1, R_2$  are random variables and  $a, b$  are constants then

$$E(aR_1 + bR_2) = aE(R_1) + bE(R_2) :$$

If  $R_1$  and  $R_2$  are also independent then

$$\text{Var}(aR_1 + bR_2) = a^2 \text{Var}(R_1) + b^2 \text{Var}(R_2) :$$

- Random sample of size  $n = 3$  :  $X_1, X_2, X_3$
- Two point estimators of  $\mu$
- Sample mean:  $\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3)$  versus
- A weighted average:  $Y = \frac{1}{6}(X_1 + 2X_2 + 3X_3)$
- Both estimators are unbiased:  $E(\bar{X}) = \mu$ , and

$$\begin{aligned} E(Y) &= \frac{1}{6} E(X_1 + 2X_2 + 3X_3) \\ &= \frac{1}{6} (\mu + 2\mu + 3\mu) = \mu \end{aligned}$$

# Comparing Estimators: An Example

- However,

$$\text{Var}(\bar{X}) = \frac{1}{3^2}(\sigma^2 + \sigma^2 + \sigma^2) = \frac{1}{3}\sigma^2$$

- while

$$\begin{aligned}\text{Var}(Y) &= \frac{1}{6^2} \text{Var}(X_1 + 2X_2 + 3X_3) \\ &= \frac{1}{36}(\sigma^2 + 2^2\sigma^2 + 3^2\sigma^2) = \frac{7}{18}\sigma^2\end{aligned}$$

- $\bar{X}$  and  $Y$  are unbiased but  $\text{Var}(\bar{X}) < \text{Var}(Y)$
- The distribution of  $\bar{X}$  is more concentrated around  $\mu$  than the distribution of  $Y$
- $\bar{X}$  is a better estimator than  $Y$
- Note: For any sample size  $n$ ,  $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

## Comparing Estimators: Another Example

- Suppose we want to estimate a population proportion,  $p$ . We obtain  $X$  “successes” in  $n$  trials.
- The usual (and natural) estimator of  $p$  is the sample proportion  $\hat{p} = \frac{X}{n}$ .  $\hat{p}$  is also the MLE and the MOM estimator of  $p$ .
- It is unbiased, and has  $\text{MSE}(\hat{p}) = \text{Var}(\hat{p}) = \frac{p(1-p)}{n}$
- An alternative estimator is

$$\tilde{p} = \frac{X + 2}{n + 4}$$

- This estimator is the sample proportion of successes computed after adding two successes and two failures to the sample.
- Can you think of an intuitive justification for this estimator?

# Comparing Estimators: Another Example

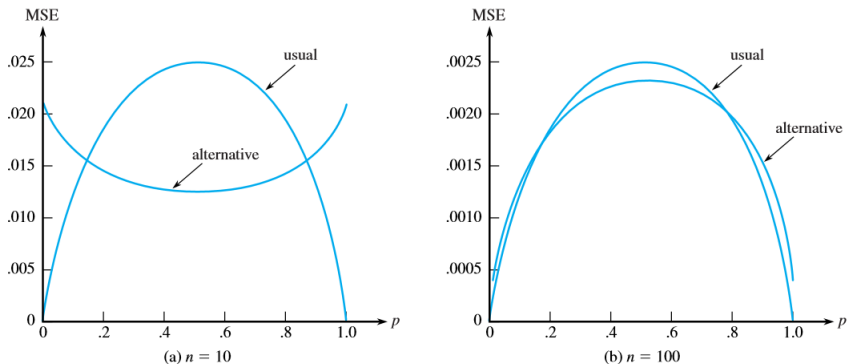


Figure 1: Comparison of MSE for the usual and alternative estimators of  $p$ . Taken from Devore and Berk [2007]

- Besides the MSE, there are other ways of quantifying closeness to the true parameter value. A general framework is to use loss functions.
- A *loss function* is a real-valued function of two variables,  $L(\theta, a)$ , where  $\theta \in \Omega$  and  $a$  is a real number. We interpret this to mean that the statistician loses  $L(\theta, a)$  if the parameter equals  $\theta$  and the estimate equals  $a$ .
- Some examples of loss functions are:
  - Squared error loss

$$L(\theta, a) = (\theta - a)^2$$

- Absolute error loss

$$L(\theta, a) = |\theta - a|$$

# Loss and Risk

- For a given loss function  $L(\theta, a)$ , the risk function of an estimator  $\hat{\theta}$  is defined to be

$$R_{\hat{\theta}}(\theta) = E_{\theta} [L(\theta, a)]$$

- The risk function is an average loss.
- Consider the loss functions listed above. The corresponding risks are given by:

$$R_{\hat{\theta}}(\theta) = E \left[ (\hat{\theta} - \theta)^2 \right]$$

and

$$R_{\hat{\theta}}(\theta) = E \left[ \left| \hat{\theta} - \theta \right| \right]$$

- Note that the first is the familiar mean squared error. The second is called the mean absolute error.



# Consistency of an Estimator

- Random sample:  $X_1, \dots, X_n$
- $Y = u(X_1, \dots, X_n)$  : An estimator of  $\theta$
- Bear in mind that  $Y$  depends on  $n$
- It would be good if  $Y$  “converges” to  $\theta$  as  $n \rightarrow \infty$
- $Y$  is *consistent* if, for any  $t > 0$ ,

$$P(|Y - \theta| \geq t) \rightarrow 0$$

as  $n \rightarrow \infty$

- The Law of Large Numbers: If  $X_1, \dots, X_n$  is a random sample from  $X$  then for any  $t > 0$ ,

$$P(|\bar{X} - \mu| \geq t) \rightarrow 0$$

as  $n \rightarrow \infty$

- Very Important Conclusion: For any population,  $\bar{X}$  is a consistent estimator of  $\mu$ .

# The Cramér-Rao or Information Inequality

- Given two unbiased estimators, we prefer the one with smaller variance
- In our quest for unbiased estimators with minimum possible variance, we need to know how small their variances can be
- $X$ : Random variable with model  $f(x; \theta)$
- The “support” of  $f$  is the region where  $f > 0$ .
- We assume that the “support” of  $f$  does not depend on  $\theta$ .
- Given: a random sample:  $X_1, \dots, X_n$ . Let  $Y$  be an unbiased estimator of  $\theta$ .
- The Cramér-Rao Inequality: The smallest possible value that  $\text{Var}(Y)$  can attain is  $1/B$  where

$$B = nE \left[ \frac{\partial}{\partial \theta} \ln f(X; \theta) \right]^2 = -nE \left[ \frac{\partial^2}{\partial \theta^2} \ln f(X; \theta) \right]$$

- Example: "... cosmic ray composition - The path length distribution ..."
- $X$ : Length of paths
- Parameter:  $\theta > 0$
- Model:  $f(x; \theta) = \theta^{-1} \exp(-x/\theta)$ ,  $x > 0$

$$\ln f(X; \theta) = -\ln \theta - \theta^{-1} X$$

$$\frac{\partial^2}{\partial \theta^2} \ln f(X; \theta) = \theta^{-2} - 2\theta^{-3} X$$

$$\begin{aligned} E \left[ \frac{\partial^2}{\partial \theta^2} \ln f(X; \theta) \right] &= E(\theta^{-2} - 2\theta^{-3} X) \\ &= \theta^{-2} - 2\theta^{-3} E(X) \\ &= \theta^{-2} - 2\theta^{-3} \theta \\ &= -\theta^{-2} \end{aligned}$$

- The smallest possible value of  $\text{Var}(Y)$  is  $\theta^2/n$
- This is attained by  $\bar{X}$ . For this problem,  $\bar{X}$  is *the* best unbiased estimator of  $\theta$

# Efficiency of Estimators

- $Y$  : An unbiased estimator of a parameter  $\theta$
- We compare  $\text{Var}(Y)$  with  $1/B$ , the lower bound in the Cramér-Rao inequality:
- $\frac{1}{B} \div \text{Var}(Y)$
- This number is called the *efficiency* of  $Y$
- Obviously,  $0 \leq \text{efficiency} \leq 1$
- If  $Y$  has 50% efficiency then about  $1/0.5 = 2$  times as many sample observations are needed for  $Y$  to perform as well as the MVUE.
- The use of  $Y$  result in confidence intervals which generally are longer than those arising from the MVUE.
- If the MLE is unbiased then as  $n$  becomes large, its efficiency increases to 1.

# General Properties of the MLE $\hat{\theta}$

We summarize here some properties of the MLE using the language we have learned in the last few slides:

- 1  $\hat{\theta}$  may not be unbiased. We often can remove this bias by multiplying  $\hat{\theta}$  by a constant.
- 2 For many models,  $\hat{\theta}$  is consistent.
- 3 The Invariance Property: For many nice functions  $g$ , if  $\hat{\theta}$  is the MLE of  $\theta$  then  $g(\hat{\theta})$  is the MLE of  $g(\theta)$ .
- 4 The Asymptotic Property: For large  $n$ ,  $\hat{\theta}$  has an approximate normal distribution with mean  $\theta$  and variance  $1/B$  where

$$B = nE \left[ \frac{\partial}{\partial \theta} \ln f(X; \theta) \right]^2$$

The asymptotic property can be used to develop confidence intervals for  $\theta$

# Confidence Intervals

- Recall: LF for globular clusters in the Galaxy
- $X$  is  $N(\mu, \sigma^2)$
- Given a random sample:  $X_1, \dots, X_n$ , the sample mean  $\bar{X}$  is an unbiased estimator of  $\mu$ :  $E(\bar{X}) = \mu$ .
- What is the probability distribution of  $\bar{X}$ ?
- Let  $Y$  be a linear combination of independent normal random variables. Then  $Y$  also has a normal distribution.
- Conclusion:  $\bar{X}$  has a normal distribution

$$E(\bar{X}) = \mu, \text{ Var}(\bar{X}) = \frac{\sigma^2}{n}, \text{ so } \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- We want to use this fact to obtain an **interval estimate** for  $\mu$ .

# Confidence Intervals

- Consult the tables of the  $N(0, 1)$  distribution:

$$P(-1.96 < Z < 1.96) = 0.95$$

- For LF data,

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95$$

- Assume that  $\sigma$  is known,  $\sigma = 1.2$  mag for Galactic globulars (van den Bergh , 1985)
- Solve for  $\mu$  in the inequalities

$$-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

- The solution is

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$
$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- The probability that the interval

$$\left( \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

“captures”  $\mu$  is 0.95.

- This interval is called a 95% *confidence interval* for  $\mu$
- It is a plausible range of values for  $\mu$  together with a quantifiable measure of its plausibility
- Notes:
  - A confidence interval is a *random* interval; it changes as the collected data changes. This explains why we say “a 95% confidence interval” rather than “the 95% confidence interval”
  - We chose the “cutoff limits”  $\pm 1.96$  symmetrically around 0 to minimize the length of the confidence interval.



## Example (devised from van den Bergh, 1985):

- $n = 148$  Galactic globular clusters
- $\bar{x} = -7.1$  mag
- We assume that  $\sigma = 1.2$  mag
- $M_0$ : The population mean visual absolute magnitude
- A 95% confidence interval for  $M_0$  is

$$\begin{aligned} & \left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right) \\ &= \left( -7.1 - 1.96 \frac{1.2}{\sqrt{148}}, -7.1 + 1.96 \frac{1.2}{\sqrt{148}} \right) \\ &= (-7.1 \pm 0.193) \end{aligned}$$

- This is a plausible range of values for  $M_0$ .

- The Warning: Don't bet your life that your 95% confidence interval has captured  $\mu$ !
- Intervals with higher levels of confidence, 90%, 98%, 99%, 99.9%, can be obtained similarly.
- Intervals with confidence levels  $100(1 - \alpha)\%$  are obtained by replacing the multiplier 1.96 in a 95% confidence by  $Z_{\alpha/2}$ , where  $Z_{\alpha/2}$  is determined by

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1 - \alpha;$$

a 95% confidence has  $\alpha = 0.05$ .

- 90%, 98%, 99%, 99.9% confidence intervals correspond to  $\alpha = .10, .02, .01$ , and  $.001$ , respectively; the corresponding values of  $Z_{\alpha/2}$  are 1.645, 2.33, 2.58, and 3.09, respectively.

- If  $\sigma$  is unknown then the previous confidence intervals are not useful
- A basic principle in statistics: Replace any unknown parameter with a good estimator
- LF data problem; a random sample  $X_1, \dots, X_n$  drawn from  $N(\mu, \sigma^2)$
- We are tempted to construct confidence intervals for  $\mu$  using the statistic  $\frac{\bar{X} - \mu}{S/\sqrt{n}}$
- What is the sampling distribution of this statistic? It is not normally distributed.
- *The t-distribution:* If  $X_1, \dots, X_n$  is a random sample drawn from  $N(\mu, \sigma^2)$  then the statistic

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution on  $n - 1$  degrees of freedom

- We construct confidence intervals as before
- Suppose that  $n = 16$ , then see the tables of the t-distribution on 15 degrees of freedom:

$$P(-2.131 < T_{15} < 2.131) = 0.95$$

- Therefore

$$P\left(-2.131 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < 2.131\right) = 0.95$$

- Solve for  $\mu$  in the inequalities

$$-2.131 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < 2.131$$

- A 95% confidence interval for  $\mu$  is

$$\left(\bar{X} - 2.131 \frac{S}{\sqrt{n}}, \bar{X} + 2.131 \frac{S}{\sqrt{n}}\right)$$

- Example:  $n = 16$ ,  $\bar{x} = -7.1$  mag,  $s = 1.1$  mag.
- A 95% confidence interval for  $\mu$  is  $-7.1 \pm 0.586$

- Normal population  $N(\mu, \sigma^2)$
- We want to obtain confidence intervals for  $\sigma$
- Random sample:  $X_1, \dots, X_n$
- $S^2$  is an unbiased and consistent estimator of  $\sigma^2$
- What is the sampling distribution of  $S^2$ ?
- The *chi-squared* ( $\chi^2$ ) distribution:  $\frac{(n-1)S^2}{\sigma^2}$  has a chi-squared distribution on  $n - 1$  degrees of freedom.
- We now construct confidence intervals as before
- Consult the tables of the  $\chi^2$  distribution
- Find the percentage points, and solve the various inequalities for  $\sigma^2$

- Denote the percentage points by  $a$  and  $b$

$$P(a < \chi_{n-1}^2 < b) = 0.95$$

- We find  $a, b$  using tables of the  $\chi^2$  distribution
- Solve for  $\sigma^2$  in the inequalities:  $a < \frac{(n-1)S^2}{\sigma^2} < b$
- A 95% confidence interval for  $\sigma^2$  is  $\left(\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right)$
- Example:  $n = 16, s = 1.2$  mag
- Percentage points from the  $\chi^2$  tables (with 15 degrees of freedom): 6.262 and 27.49
- Note: The percentage points are not symmetric about 0
- A 95% confidence interval for  $\sigma^2$  is

$$\left(\frac{15 \times (1.2)^2}{27.49}, \frac{15 \times (1.2)^2}{6.262}\right) = (0.786, 3.449)$$

## The width of a confidence interval

All other things remaining constant:

The greater the level of confidence, the longer the confidence interval.

The larger the sample size, the shorter the confidence interval.

- How do we choose  $n$ ?
- In our 95% confidence intervals for  $\mu$ , the term  $1.96\sigma/\sqrt{n}$  is called the margin of error
- We choose  $n$  to have a desired margin of error
- To have a margin of error of 0.01 mag then we choose  $n$  so that

$$\frac{1.96\sigma}{\sqrt{n}} = 0.01$$

- Solve this equation for  $n$ :

$$n = \left( \frac{1.96\sigma}{0.01} \right)^2$$

# Confidence Intervals with Large Sample Sizes

- Confidence intervals with large sample sizes
- Papers on LF for globular clusters
- Sample sizes are large: 68, 148, 300, 1000, ...
- A modified Central Limit Theorem
- $X_1, \dots, X_n$ : a random sample
- $\mu$ : The population mean
- $\bar{X}$  and  $S$ : The sample mean and std. deviation
- The modified CLT: If  $n$  is large then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx N(0, 1)$$

- The conclusion does not depend on the population probability distribution
- The resulting confidence intervals for  $\mu$  also do not depend on the population probability distribution



# Tests of Hypotheses

- Alternatives to confidence intervals
- A LF researcher believes that  $M_0 = -7.7$  mag for the M31 globular clusters. The researcher collects a *random sample* of data from M31
- A natural question: “Are the data strongly in support of the claim that  $M_0 = -7.7$  mag?”
- *Statistical hypothesis*: A statement about the parameters of a population.
- *Statistical test of significance*: A procedure for comparing observed data with a hypothesis whose plausibility is to be assessed.
- *Null hypothesis*: The statement being tested.
- *Alternative hypothesis*: A competing statement.
- In general, the alternative hypothesis is chosen as the statement for which we hope to find supporting evidence.

- In the case of our M31 LF researcher, the null hypothesis is  $H_0 : M_0 = -7.7$
- An alternative hypothesis is  $H_a : M_0 \neq -7.7$
- Two-sided alternative hypotheses
- One-sided alternatives, e.g.,  $H_a : M_0 < -7.7$
- To test  $H_0$  vs.  $H_a$ , we need:
  - (a) A test statistic: This statistic will be calculated from the observed data, and will measure the compatibility of  $H_0$  with the observed data. It will have a sampling distribution free of unknown parameters.
  - (b) A rejection rule which specifies the values of the test statistic for which we reject  $H_0$ .

- Example: A random sample of 64 measurements has mean  $\bar{x} = 5.2$  and std. dev.  $s = 1.1$ . Test the null hypothesis  $H_0 : \mu = 4.9$  against the alternative hypothesis  $H_a : \mu \neq 4.9$

Step 1: The null and alternative hypotheses:

$$H_0 : \mu = 4.9, H_a : \mu \neq 4.9$$

Step 2: The test statistic:

$$T = \frac{\bar{X} - 4.9}{S/\sqrt{n}}$$

Step 3: The distribution of the test statistic under the assumption that  $H_0$  is valid:  $T \approx N(0, 1)$

Step 4: The rejection rule: Reject  $H_0$  if  $|T| > 1.96$ . Otherwise, we fail to reject  $H_0$ . This choice of cutoff point (also called critical value) results in a 5% level of significance of the test of hypotheses.

**Step 5:** Calculate the value of the test statistic: The calculated value of the test statistic is

$$\frac{\bar{X} - 4.9}{S/\sqrt{n}} = \frac{5.2 - 4.9}{1.1/\sqrt{64}} = 2.18$$

**Step 6:** Decision: We reject  $H_0$ ; the calculated value of the test statistic exceeds the critical value, 1.96.

We report that the data are *significant* and that there is a *statistically significant* difference between the population mean and the hypothesized value of 4.9

**Step 7:** The P-value of the test: The smallest significance level at which the data are significant.

$$P(t > 2.18) = 0.0165315$$

Thus the exact p-value is  $2 \times 0.01653 = 0.03306$

# Inference on the Means of Two Populations, Variances Known

- Premise: We have two different populations with unknown means  $\mu_1$  and  $\mu_2$ , respectively and known variances  $\sigma_1^2$  and  $\sigma_2^2$ , respectively. We have obtained two samples of sizes  $n_1$  and  $n_2$  from population 1 and population 2 respectively. It is notationally convenient to express the samples as

$$\mathbf{X}_1 = X_{11}, X_{12}, \dots, X_{1n_1}$$

$$\mathbf{X}_2 = X_{21}, X_{22}, \dots, X_{2n_2}$$

- We make two further assumptions. The first is that the samples  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are independent. The second is that both populations are normal, (or if not, then the conditions of the central limit theorem apply).

- At this point all we need to do are to establish procedures for making inference on the difference between the (unknown) means of the two populations,  $\mu_1 = \mu_2$ .
- The point estimator for the parameter of interest,  $\mu_1 - \mu_2$  is  $\bar{X}_1 - \bar{X}_2$ . The expected value of this point estimate is

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

and the variance is

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Now we may claim that

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution.

- Suppose someone makes a claim that the difference in means between populations 1 and 2 is  $\Delta_0$ . (We call  $\Delta_0$  the null or hypothesized difference.)
- To test this claim, we formulate the problem as a hypothesis test with null hypothesis  $H_0 : \mu_1 - \mu_2 = \Delta_0$ .
- Then an appropriate test statistic is given by

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- The remaining steps are similar to what we did in the case of a single mean.
- We usually have  $\Delta_0 = 0$ . In this case the null hypothesis  $H_0 : \mu_1 - \mu_2 = 0$  OR  $H_0 : \mu_1 = \mu_2$  can be put in words as:  $H_0$ : There is no difference in mean between population 1 and population 2.
- The two-tailed alternative is  $H_0 : \mu_1 - \mu_2 \neq 0$  OR  $H_0 : \mu_1 \neq \mu_2$ . In words, there is a difference in mean between population 1 and population 2.
- We can also have a one-tailed alternative in either direction.

# Inference on the Means of Two Populations, Variances Unknown - An Example

- The diameter of steel rods manufactured on two different extrusion machines is being investigated. Two random samples of sizes  $n_1 = 15$  and  $n_2 = 17$  are selected, and the sample means and sample variances are  $\bar{x}_1 = 8.73$ ,  $s_1^2 = 0.35$ ,  $\bar{x}_2 = 8.68$ , and  $s_2^2 = 0.40$ . Assume that  $\sigma_1^2 = \sigma_2^2$  and that the data are drawn from a normal distribution.
  - (a) Is there evidence to support the claim that the two machines produce rods with different mean diameters? Use a P-value in arriving at the conclusion.
  - (b) Construct a 95% CI for the difference in mean rod diameter. Interpret this interval.

**Note that** for problems of this nature, there are two possible scenarios which will lead to different distributions of the test statistics:

- 1 We assume  $\sigma_1^2 = \sigma_2^2$
- 2 We assume  $\sigma_1^2 \neq \sigma_2^2$ .



$$(a) \boxed{H_0 : \mu_1 - \mu_2 = 0, H_1 : \mu_1 - \mu_2 \neq 0}$$

First we find the pooled sample variance,  $S_p^2$ .

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{14 \times 0.35 + 16 \times 0.40}{30} \\ &= 0.377 \end{aligned}$$

Therefore  $S_p = \sqrt{S_p^2} = 0.6137$ .

Now compute the test statistic:

$$\begin{aligned} T &= \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \\ &= \frac{8.73 - 8.68 - 0}{0.6137 \sqrt{\frac{1}{15} + \frac{1}{17}}} \\ &= 0.230 \end{aligned}$$

$$|t_0| = |0.230| = 0.230$$

- The test statistic follows a t distribution with degrees of freedom  $n_1 + n_2 - 2 = 30$ . Therefore the P-value is

$$2 \times P(t_{30} > |t_0|) = 2 \times P(t_{30} > 0.230) > 0.80$$

We fail to reject  $H_0$

- (b) From the t-table,  $t_{\alpha/2, n_1+n_2-2} = t_{0.025, 30} = 2.042$ . The 95% CI is given by

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq$$

$$\bar{X}_1 - \bar{X}_2 + t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$0.05 - 2.042(0.2174) \leq \mu_1 - \mu_2 \leq 0.05 + 2.042(0.2174)$$

$$-0.394 \leq \mu_1 - \mu_2 \leq 0.494$$

- Therefore, we can be 95% confident that the true mean difference in fill volume lies between  $-0.394$  and  $0.494$ .

# The ML Method for Testing Hypotheses

- $X \sim N(\mu, \sigma^2)$ ; parameters  $\mu$  and  $\sigma^2$
- Model:  $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$
- Random sample:  $X_1, \dots, X_n$
- We wish to test  $H_0 : \mu = 3$  vs.  $H_a : \mu \neq 3$
- Parameter space: The space of all permissible values of the parameters

$$\Omega = \{(\mu, \sigma) : -\infty < \mu < \infty, \sigma > 0\}$$

- $H_0$  and  $H_a$  represent restrictions on the parameters, so we are led to parameter subspaces

$$\omega_0 = \{(\mu, \sigma) : \mu = 3, \sigma > 0\}$$

$$\omega_a = \{(\mu, \sigma) : \mu \neq 3, \sigma > 0\}$$

$$\begin{aligned}
 L(\mu, \sigma^2) &= f(X_1; \mu, \sigma^2) \cdots f(X_n; \mu, \sigma^2) \\
 &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right]
 \end{aligned}$$

- Maximize  $L(\mu, \sigma^2)$  over  $\omega_0$  and  $\omega_a$
- The *likelihood ratio test statistic* is

$$\lambda = \frac{\max_{\omega_0} L(\mu, \sigma^2)}{\max_{\omega_0 \cup \omega_a} L(\mu, \sigma^2)} = \frac{\max_{\sigma > 0} L(3, \sigma^2)}{\max_{\sigma > 0, \mu} L(\mu, \sigma^2)}$$

- Fact:  $0 \leq \lambda \leq 1$
- $L(3, \sigma^2)$  is maximized over  $\omega_0$  at

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - 3)^2$$

$$\begin{aligned}\max_{\omega_0} L(3, \sigma^2) &= L\left(3, \frac{1}{n} \sum_{i=1}^n (X_i - 3)^2\right) \\ &= \left[ \frac{n}{2\pi e \sum_{i=1}^n (X_i - 3)^2} \right]^{n/2}\end{aligned}$$

$L(\mu, \sigma^2)$  is maximized over  $\omega_a$  at

$$\mu = \bar{X}, \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\begin{aligned}\max_{\omega_0 \cup \omega_a} L(\mu, \sigma^2) &= L\left(\bar{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right) \\ &= \left[ \frac{n}{2\pi e \sum_{i=1}^n (X_i - \bar{X})^2} \right]^{n/2}\end{aligned}$$

- The likelihood ratio test statistic:

$$\begin{aligned}\lambda &= \left[ \frac{n}{2\pi e \sum_{i=1}^n (X_i - 3)^2} \right]^{n/2} \div \left[ \frac{n}{2\pi e \sum_{i=1}^n (X_i - \bar{X})^2} \right]^{n/2} \\ &= \left[ \sum_{i=1}^n (X_i - \bar{X})^2 \div \sum_{i=1}^n (X_i - 3)^2 \right]^{n/2}\end{aligned}$$

- $\lambda$  is close to 1 iff  $\bar{X}$  is close to 3
- $\lambda$  is close to 0 iff  $\bar{X}$  is far from 3
- This particular LRT statistic  $\lambda$  is equivalent to the t-statistic seen earlier
- In this case, the ML method discovers the obvious test statistic

Devore, J. L. and Berk, K. N. (2007). *Modern Mathematical Statistics with Applications*. Thomson Brooks/Cole, Belmont, CA.