Space Telescope Science Institute statistics mini-course

October 2011

Inference I: Estimation, Confidence Intervals, and Tests of Hypotheses

James L Rosenberger

Acknowledgements:
Donald Richards, William Harkness, and B.V Rao

Department of Statistics
Center for Astrostatistics
Penn State University

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

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

$M_0 \equiv \mu$: Mean visual absolute magnitude

σ: Std. deviation of visual absolute magnitude

Magnitudes are log variables (a log-normal distribution)

Statistical Problems:

1. On the basis of collected data, estimate the *parameters* μ and σ. 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?

Here is a diagram from van den Bergh (1985), providing complete data for the Milky Way (notice that the data *appear* to be non-Gaussian)

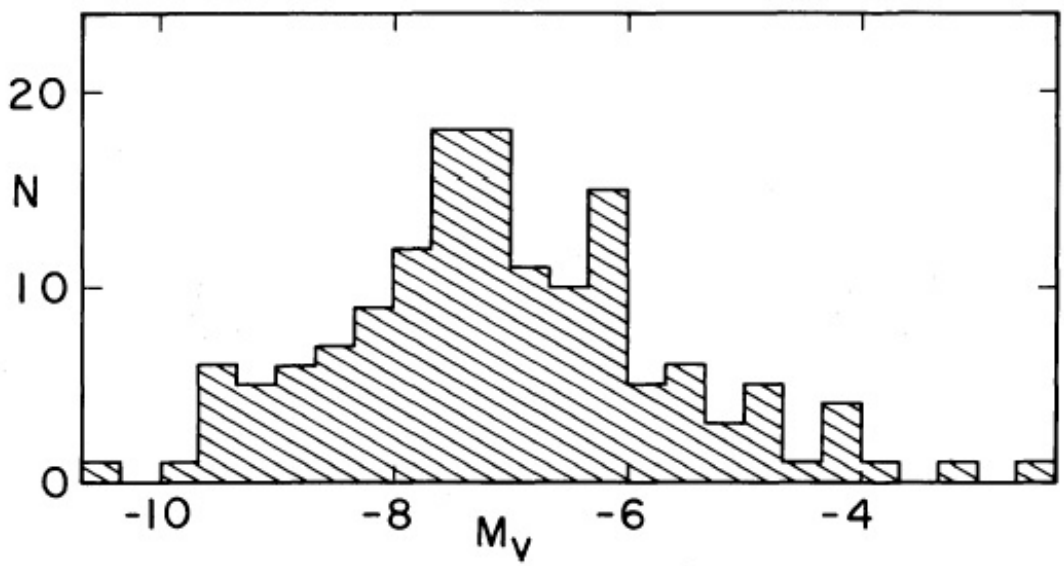

FIG. 2.—Luminosity function of Galactic globular clusters (one object at $M_V = -1.7$ is not plotted). Note that the luminosity function is asymetrical with a long tail extending to faint magnitudes.

A second diagram from van den Bergh (1985); truncated dataset for M31 (Andromeda galaxy)

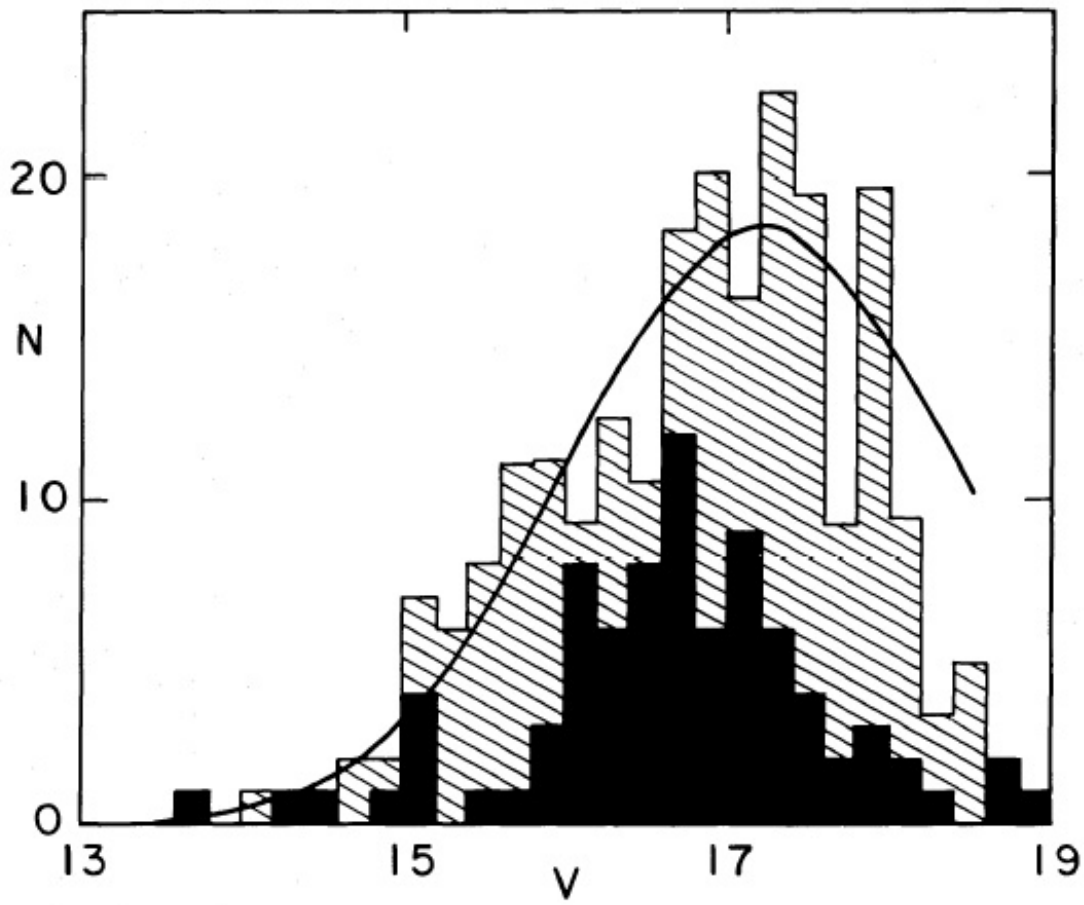

FIG. 1.—Luminosity function of clusters with $0.70 \leq B-V < 1.00$ in M31. Smooth curve is a Gaussian with $V(\max) = 17.2$ and $\sigma = 1.2$ mag. Lower histogram shows the luminosity function of the halo of M31 derived by Racine and Shara (1979).

X: A random variable

Population: The collection of all values of X

$f(x)$: The prob. density function (p.d.f.) of X

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

μ and σ are parameters for the p.d.f. of the LF for Galactic globulars

Values of the chosen p.d.f. depend on X and on the parameters: $f(x; \mu, \sigma)$

Parameter space: The set of permissible values of the parameters

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

Random sample

In practice: Data values $x_1, \ldots, x_n$ which are *fully representative* of the population

In theory: Mutually independent random variables $X_1, \ldots, X_n$ which all have the same distribution as X

Parameter: A number computable only from the entire population

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

Sample mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

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

In general, a statistic is a fucntion $Y = U(X_1, \ldots, X_n)$ of the observations.

Sampling distribution: The probability distribution of a statistic

X: LF for globular clusters

Model: $N(\mu, \sigma^2)$, the normal distribution with mean μ and variance σ^2

Problem: Given a random sample $x_1, \ldots, x_n$, estimate μ

$\bar{x}$ is a very good estimate of μ

$\widehat{x}$, the sample median, is a good plausible estimate of μ

$x_{(n)}$, the largest observed value in the LF, is obviously a poor estimate of μ, *since it almost certainly is much larger than* μ.

Statistics, like the sample mean $\bar{x}$ and the sample median $\widehat{x}$ are called *point estimators* of μ

Roman letters are used to denote Data and Greek letters to denote parameters.

Let θ be a 'generic' parameter (for example, μ or σ)

$Y = u(X_1, \ldots, X_n)$, a function of the data;

Y is

(i) a point estimator of θ,

(ii) a random variable and therefore

(iii) has a probability distribution called *the sampling distribution of the statistic* Y.

Conceptually, we can calculate the moments of Y, including the mean $E(Y)$.

If $E(Y) = \theta$, then Y is said to be an unbiased estimator of θ, for example $\bar{x}$ is an unbiased estimator of the population mean μ.

Intuitively, Y is unbiased if its long-term average value is equal to θ

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 $\widehat{x}$, the sample median, is an unbiased estimate of μ also (based on the symmetry of $f(x)$).

$X_{(n)}$: the largest observed LF

$X_{(n)}$ is not unbiased: $E(X_{(n)}) > \mu$

We also want statistics which are "close" to θ

For all statistics Y, calculate $E[(Y-\theta)^2]$, the mean square error (MSE)

Choose as our point estimator the statistic for which the MSE is smallest

A statistic Y which minimizes $E[(Y-\theta)^2]$ is said to have *minimum mean square error*

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

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

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

Example: LF for globular clusters

$X \sim N(\mu, \sigma^2) = f(x : \mu, \sigma)$

Random sample of size $n = 3$: X_1, X_2, X_3

Two point estimators of μ:

Sample mean: $\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3)$

Place more weight on the last observation

A weighted average: $Y = \frac{1}{6}(X_1 + 2X_2 + 3X_3)$

Both estimators are unbiased: $E(\bar{X}) = \mu$, and

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

However,

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

while

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

$\bar{X}$ and Y are unbiased but $\text{Var}(\bar{X}) < \text{Var}(Y)$

The distribution of $\bar{X}$ is more concentrated around μ than the distribution of Y

$\bar{X}$ is a better estimator than Y

Note: For any sample size n, $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

Random sample: $X_1, \ldots, X_n$

$Y = u(X_1, \ldots, X_n)$: An estimator of θ

Bear in mind that Y depends on n

It would be good if Y "converges" to θ as $n \to \infty$

Y is *consistent* if, for any $t > 0$,

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

as $n \to \infty$

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

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

as $n \to \infty$

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

How do we construct good estimators?

Judicious guessing

The method of maximum likelihood

The method of moments

Bayesian methods

Decision-theoretic methods

Unbiased estimator

Consistent estimator

A consequence of Chebyshev's inequality: If Y is an unbiased estimator of θ and $\mathrm{Var}(Y) \to 0$ as $n \to \infty$ then Y is consistent.

The Method of Moments

X: Random variable with p.d.f. $f(x; \theta_1, \theta_2)$

Parameters to be estimated: θ_1, θ_2

Random sample: $X_1, \ldots, X_n$

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 θ_1 and θ_2

3. Solve for θ_1, θ_2 the simultaneous equations

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

The solutions are the *method-of-moments estimators* of θ_1, θ_2

Example: LF for globular clusters

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

Random sample: $X_1,\ldots,X_n$

1. 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. The first two population moments:

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

3. Solve: $\widehat{\mu} = m_1$, $\widehat{\mu}^2 + \widehat{\sigma}^2 = m_2$

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

$\widehat{\mu}$ is unbiased; $\widehat{\sigma}^2$ is not unbiased

Hanes-Whittaker (1987), "Globular clusters as extragalactic distance indicators ...," AJ 94, p. 906

M_l: The absolute magnitude limit of the study

T: A parameter identifying the size of a cluster

Truncated normal distribution:

$$f(x;\mu,\sigma^2,T) \propto \begin{cases} \frac{T}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], & x \le M_l \\ 0, & x > M_l \end{cases}$$

Method of moments: Calculate

1. The first three sample moments,

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1,2,3$$

2. The first three population moments

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

3. Solve the equations $m_k = E(X^k)$, $k = 1,2,3$

Garcia-Munoz, et al. "The relative abundances of the elements silicon through nickel in the low energy galactic cosmic rays," In: Proc. Int'l. Cosmic Ray Conference, 1978

Measured abundances compared with propagation calculations using distributions of path lengths; data suggest an exponential distribution truncated at short path lengths

Protheroe, et al. "Interpretation of cosmic ray composition - The path length distribution," ApJ., 247 1981

X: Length of paths

Parameters: $\theta_1, \theta_2 > 0$

Model:

$$f(x; \theta_1, \theta_2) = \begin{cases} \theta_1^{-1} \exp[-(x - \theta_2)/\theta_1], & x \geq \theta_2 \\ 0, & x < \theta_2 \end{cases}$$

LF for globular clusters in the Galaxy,

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

Random sample: $X_1, \ldots, X_n$

$\bar{X}$ is an unbiased estimator of μ

$\bar{X}$ has minimum variance among all estimators which are linear combinations of $X_1, \ldots, X_n$

S^2 is an unbiased estimator of σ^2

Given an actual data set, we calculate $\bar{x}$ and s^2 to obtain point estimates of μ and σ^2

Point estimates are not perfect

We wish to quantify their accuracy

Confidence Intervals

LF for globular clusters in the Galaxy

X is $N(\mu, \sigma^2)$

Random sample: $X_1, \ldots, X_n$

$\bar{X}$ is an unbiased estimator of μ: $E(\bar{X}) = \mu$

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

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

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

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

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

For LF data,

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

Assume that σ is known, $\sigma = 1.2$ mag for Galactic globulars (van den Bergh , 1985)

Solve for μ the inequalities

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

The solution is

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

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

The probability that the interval

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

"captures" μ is 0.95.

This interval is called a 95% *confidence interval for* μ

It is a plausible range of values for μ 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" ± 1.96 symmetrically around 0 to minimize the length of the confidence interval.

"Cutoff limits" are called "percentage points"

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

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

This is a plausible range of values for M_0.

The Warning: Don't bet your life that your 95% confidence interval has captured μ (but the odds are in your favor -19 to 1)

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 σ 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 $X_1, \ldots, X_n$ drawn from $N(\mu, \sigma^2)$

We are tempted to construct confidence intervals for μ 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 $X_1, \ldots, X_n$ is a random sample drawn from $N(\mu, \sigma^2)$ then the statistic

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

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

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

Solve for μ in the inequalities

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

A 95% confidence interval for μ 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 μ is -7.1 ∓ 0.586

Normal population $N(\mu, \sigma^2)$

We want to obtain confidence intervals for σ

Random sample: $X_1, \ldots, X_n$

S^2 is an unbiased and consistent estimator of σ^2

What is the sampling distribution of S^2?

The *chi-squared* (χ^2) distribution: $(n-1)S^2/\sigma^2$ has a chi-squared distribution on $n-1$ degrees of freedom.

We now construct confidence intervals as before

Consult the tables of the χ^2 distribution

Find the percentage points, and solve the various inequalities for σ^2

Denote the percentage points by a and b

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

We find a, b using tables of the χ^2 distribution

Solve for σ^2 the inequalities: $a < \frac{(n-1)S^2}{\sigma^2} < b$

A 95% confidence interval for σ^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 χ^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 σ^2 is

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

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 μ, 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

Papers on LF for globular clusters

Sample sizes are large: 68, 148, 300, 1000, ...

A modified Central Limit Theorem

$X_1, \ldots, X_n$: a random sample

μ: 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 μ 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 $H_a : \mu \neq 4.9$

1. The null and alternative hypotheses:
   $H_0 : \mu = 4.9, \quad H_a : \mu \neq 4.9$

2. The test statistic:

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

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

4. The rejection rule:
   Reject H_0 if $|T| > 1.96$, the upper 95% percentage point in the tables of the standard normal distribution. Otherwise, we *fail to reject* H_0.

This cutoff point is also called a *critical value*.

This choice of critical value results in a 5% *level of significance* of the test of hypotheses.

5. Calculate the value of the test statistic:

The calculated value of the test statistic is

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

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

7. The P-value of the test:

The smallest significance level at which the data are significant.