Statistical Inference

Hyungsuk (Tak) Tak

University of Notre Dame

Summer School in Statistics for Astronomers XV
June 3-4, 2019

Inspired by the lecture notes prepared by Bing Li, Kwame Kankam, and James Rosenberger (Penn State), and Chapter 3 of Feigenson and Babu (2012), “Modern Statistical Methods for Astronomy”. 
Probability and inference are two sides of the same coin.

Probability explains how likely various outcomes (observations) are, given the model parameter $\theta$, while inference quantifies the uncertainty about $\theta$, given observed data $x$.

Statistical inference allows quantitative evaluation of parameters within the context of astronomical and astrophysical models.

Astronomers measure the properties of a limited sample of objects (often chosen to be brighter or closer than others) to learn about the properties of the vast underlying population of similar objects in the Universe.
Statistical inference is so pervasive throughout astronomical and astrophysical investigations. For example,

- smoothing over discrete observations to understand the underlying continuous phenomenon,
- seeking to quantify relationships between observed properties,
- testing whether an observation agrees with an assumed astrophysical theory,
- trying to compensate for flux limits and non-detections,
- investigating the temporal behavior of variable sources,
- inferring the evolution of cosmic bodies from studies of objects at different stages, and
- characterizing and modeling patterns in wavelength, images or space.
**Terminology in statistical inference**

**Population**: Entire group of objects about which we make inferences.

**Sample**: A part of the population on which we actually collect data.

**Statistic**: A numerical summary (a function) of observed (sample) data, $g(X_1, X_2, \ldots, X_n)$, e.g., mean, median, maximum, variance, etc. Note that $X_1, X_2, \ldots, X_n$ denote the data that we have not yet observed, i.e., random variables, and $x_1, x_2, \ldots, x_n$ are the observed (realized) data.

**Parameter** $\theta$: A numerical summary of a statistical model or population.

**Estimator**: A statistic meant to guess the parameter $\theta$ (e.g. $\overline{X} \rightarrow \mu$).

**Estimate**: A value of estimator computed by the observed data (e.g. $\overline{x}$).
Two different thoughts on $\theta$

**Frequentist vs Bayesian:** Is $\theta$ an unknown ‘constant’ or ‘random variable’?

- Frequentists consider $\theta$ as an unknown constant, not a random variable. Since $\theta$ is a fixed constant (just unknown), the realization (sampling) of the data is the only source of randomness. Thus, the sampling distribution of the estimator for $\theta$ is the key to the inference. “If we can repeat the experiment (or sampling) infinitely many times, e.g., bootstrapping in practice, how does the estimator for $\theta$ vary around the unknown constant $\theta$ on average?”

- Bayesians consider $\theta$ as a random variable and the data as fully known constants, which is the opposite to frequentist’s thought. Because $\theta$ is a random variable, it must have a probability distribution, call a prior distribution. Bayesian inference on $\theta$ is based on the posterior distribution of $\theta$ that updates prior belief (prior distribution of $\theta$) by the observed data (likelihood).
Overview of statistical inference

Statistical inference can be

- parametric (which requires that scientists make some assumptions regarding the mathematical structure of the population, and this structure has parameters to be estimated from the data at hand)
- non-parametric (which makes no assumption about the model structure or the distribution of the population)
- semi-parametric

Two main aspects of statistical inference are

- Estimation
  - Point estimation: Estimating a single best value of $\theta$.
  - Interval estimation: Estimating a range of plausible values of $\theta$.
- Hypothesis testing
  - Whether the given data are consistent with a stated hypothesis.
Two decisions must be made:

1. The functional model and its parameters must be specified.
   - If the model is not well-matched to the astronomical population or astrophysical process under the study, then the best fit obtained by the inferential process may be meaningless (model mis-specification).
   - Statistical procedures for model validation (or goodness-of-fit) and model selection are available.

2. The method by which best-fit parameters are estimated must be chosen.
   - Method of moments
   - Least squares (often called a minimum $\chi^2$ in astronomy)
   - Maximum likelihood estimation
   - Bayesian maximum a posteriori estimate

For many situations, we can find a single best-fit parameter closest to the true value, with the greatest accuracy (smallest uncertainty) or with the highest likelihood.
Principles of point estimation (cont.)

Settings:

- The observed data \(x_1, x_2, \ldots, x_n\) are assumed to be realizations of independent random variables \(X_1, X_2, \ldots, X_n\) with a common probability distribution function (PDF) \(f\).
- The PDF \(f\) is characterized by a small number of parameters \(\theta = (\theta_1, \theta_2, \ldots, \theta_p)\).

A point estimator of \(\theta\), denoted by \(\hat{\theta}\) is a function of the random variables (unrealized data) of the underlying population under study, i.e.,

\[
\hat{\theta} = g(X_1, X_2, \ldots, X_n),
\]

and is computed from a realization of the population (observed data).

If one validates that the data are consistent with the model, the result of parameter estimation can be a great simplification of a collection of data into a few easily interpretable parameters.
“Best” or “best-fit”? Important criteria to evaluate point estimators.

- **Unbiasedness**: An estimator \( \hat{\theta} \) of a parameter \( \theta \) is unbiased if
  \[
  \text{Bias} = E(\hat{\theta}) - \theta = 0 \quad \text{or} \quad E(\hat{\theta}) = \theta.
  \]
  Heuristically, \( \hat{\theta} \) is unbiased if its long-term average value is \( \theta \).

- **Minimum variance unbiased estimator**: If there are several unbiased estimators, one with the smallest variance is preferred.

- **Mean squared error (MSE)**: A composite measure of bias and variance. \( \text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2 \). Smaller MSE is preferred.

- **Consistency**: \( \hat{\theta} \to \theta \) (in probability) as the sample size \( n \) increases.
Techniques of point estimation: MM

**Moments:** The $k$-th moment of a random variable $X$ is $\mu_k = E(X^k)$.

**Sample moments:** Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The $k$-th sample moment is defined as $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$.

**Method of moments:** A way to obtain estimators by matching moments of a random variable with sample moments.

1. Calculate low-order moments ($k \leq p$) and relate them to unknown parameters, e.g., if $p = 1$, computing $\mu_1$ is enough. Note that moments ($\mu_k$'s) are functions of parameters ($\theta$), e.g., $\mu_1 = g_1(\theta)$.

2. Solve the equation so the parameter is in terms of moments: $\theta = g_1^{-1}(\mu_1)$.

3. Plug in the sample moment and obtain an estimator for the parameter: $\hat{\theta} = g_1^{-1}(\hat{\mu}_1)$

**Pros:** It is easy to derive and yields consistent estimators ($\hat{\theta} \xrightarrow{p} \theta$).

**Cons:** It often results in biased estimators and is particularly problematic in small samples. Also, sometimes estimates are outside valid parameter spaces (e.g. a negative variance estimate).
Example: van den Bergh [1985] considers the luminosity function for globular clusters in various galaxies.

vdB’s conclusion: The luminosity function for clusters in the Milky Way is adequately described by a Gaussian (Normal) distribution. Its PDF is

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

- \(\mu (= M_0)\): Mean visual absolute magnitude.
- \(\sigma\): Standard deviation of visual absolute magnitude.

Find MoM estimators for \(\mu\) and \(\sigma^2\) with a sample of globular clusters.

1. Two moments: \(E(X) = \mu, \ E(X^2) = \text{Var}(X) + E(X)^2 = \sigma^2 + \mu^2\).
2. Solve equations w.r.t. \(\mu\) & \(\sigma^2\): \(\mu = E(X), \ \sigma^2 = E(X^2) - \mu^2\).
3. Plug in sample moments, replacing \(E(X^k)\) with \(\frac{1}{n} \sum_{i=1}^{n} X_i^k\):
   \[
   \hat{\mu}_{\text{MM}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X},
   \]
   \[
   \hat{\sigma}^2_{\text{MM}} = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.
   \]
Techniques of point estimation: LS

Method of least squares (LS): Originally developed in the early 19th century to solve problems in celestial mechanics (Gauss, Legendre, and Laplace). This method, often called minimum $\chi^2$ method in astronomy, finds values of parameters that minimize the sum of squared errors (or the sum of weighted squared errors if errors are heteroscedastic).

Least square estimation procedure

1. Set up a model for an observation $x_i$, parametrized by $\theta$, i.e., $\hat{x}_i(\theta)$.
2. Find a value of $\theta$ that minimizes

$$\sum_{i=1}^{n} \frac{(x_i - \hat{x}_i(\theta))^2}{\sigma^2}$$

if measurement errors are homoscedastic, or

$$\sum_{i=1}^{n} \frac{(x_i - \hat{x}_i(\theta))^2}{\sigma_i^2}$$

if heteroscedastic with different measurement error variances $\sigma_i^2$. 

Techniques of point estimation: LS (cont.)

**Pros:** The LS method is heavily used in a classical regression setting, e.g., \( Y_i = f(X_i) + \epsilon_i \) \( \equiv \) \( \alpha + \beta X_i + \epsilon_i \) to find a line that minimizes the sum of squared vertical distances to observations (see the figure below).

It is intuitive and easy-to-implement. In a linear regression setting with some conditions, the LS estimator is the best linear unbiased estimator (Gauss-Markov theorem).

**Cons:** It is quite sensitive to outliers. Usually, sub-optimal to the maximum likelihood estimator (that will appear next).
Example: A simple linear regression model with the observed data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ is specified as
\[ y_i = \alpha + \beta x_i + \epsilon_i \]

The LS estimates for $\alpha$ and $\beta$ that minimize the sum of the squared lengths of the blue and red vertical lines, i.e., $\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$ are
\[
\hat{\beta}_{LS} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\alpha}_{LS} = \bar{y} - \hat{\beta}_{LS} \bar{x}.
\]
Techniques of point estimation: MLE

**Likelihood function** (R. A. Fisher, 1922) is the joint probability density (or mass) function of the observed data as a function of $\theta$. Thus, it indicates how likely the observed data are as a function of $\theta$, and maximizing the likelihood function determines the parameters that are most likely to produce the observed data. If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with PDF $f(x; \theta)$, the likelihood function of $\theta$ is

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) \overset{\text{ind.}}{=} \prod_{i=1}^{n} f(x_i; \theta).$$

**Maximum likelihood estimation procedure**

1. Write down the log-likelihood function of $\theta$, i.e., $\ell(\theta) = \ln(L(\theta))$.
   Why log? Easier. It is monotonic, preserving the maximizing value.

2. Maximize it in terms of the desired parameter: $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \ell(\theta)$
Techniques of point estimation: MLE (cont.)

Fisher's Information, denoted by $I_n(\theta)$, measures the amount of information about $\theta$ that the $n$ observed data contain. Intuitively, if we have more observations, the amount of information about $\theta$ increases.

$$I_n(\theta) = E \left[ \left( \frac{\partial}{\partial \theta} \ell(\theta) \right)^2 \right] = -E \left( \frac{\partial^2}{\partial \theta^2} \ell(\theta) \right)$$

Cramér-Rao inequality states that the inverse of Fisher's information is the smallest variance that an unbiased estimator can have, i.e.,

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)}.$$

Asymptotic optimality of MLE: As $n \uparrow$, MLE is unbiased (i.e., consistent), the most efficient (smallest variance, achieving the Cramér-Rao lower bound), and Normally distributed estimator:

$$\hat{\theta}_{\text{MLE}} \sim N \left( \theta, \frac{1}{I_n(\theta)} \right).$$
Techniques of point estimation: MLE (cont.)

Pros:

• In most probability structures in astronomy, MLE exists and unique.
• In linear regression settings, MLE=LSE if error terms are Gaussian.
• Asymptotic optimality:

\[ \hat{\theta}_{\text{MLE}} \sim N\left( \theta, \frac{1}{I_n(\theta)} \right). \]

• MLE is invariant, i.e., for any one-to-one function \( g \), the MLE of \( g(\theta) \) is simply \( g(\hat{\theta}_{\text{MLE}}) \). Its asymptotic distribution with \( g'(\theta) \neq 0 \) is

\[ g(\hat{\theta}_{\text{MLE}}) \sim N\left( g(\theta), \frac{g'(\theta)^2}{I_n(\theta)} \right). \]

Cons:

• The MLE is sometimes biased with finite sample size.
• It does not always provide a closed-form solution. Algorithms for optimization and root-finding, such as Newton-Raphson algorithm and EM (Expectation/Maximization) algorithm, can be used.
Example 1: The shape of the cosmic ray path length distribution is modeled by an Exponential distribution with scale $\theta$ (Protheroe et al., 1981). Let $X$ be the length of a path. Given $\theta > 0$, its PDF is

$$f(x; \theta) = \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right)$$

Find the MLE for $\theta$ with a sample of cosmic ray path lengths, $x_1, \ldots, x_n$.

1. Derive log-likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \frac{1}{\theta^n} \exp\left(-\frac{\sum_{i=1}^{n} x_i}{\theta}\right).$$

$$\ell(\theta) = \ln(L(\theta)) = -n \ln(\theta) - \frac{\sum_{i=1}^{n} x_i}{\theta}$$

2. Find a value (estimate) that maximizes $\ell(\theta)$:

$$\frac{d}{d\theta} \ell(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0 \quad \Rightarrow \quad \theta = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \quad \Rightarrow \quad \hat{\theta}_{\text{MLE}} = \bar{x}.$$

Thus, the maximum likelihood estimator of $\theta$ is $\hat{\theta}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$. 


Example 2: Luminosity function for clusters in the Milky Way. Its PDF:

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

Find the MLEs for \( \mu \) & \( \sigma^2 \) with a sample of globular clusters, \( x_1, \ldots, x_n \).

1. Derive log-likelihood function:

\[
L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right),
\]

\[
\ell(\mu, \sigma^2) = \ln(L(\mu, \sigma^2)) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}.
\]

2. Find values (estimates) that maximize \( \ell(\mu, \sigma^2) \):

\[
\frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\sigma^2} = 0 \quad \rightarrow \quad \hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}
\]

\[
\frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^4} = 0 \quad \rightarrow \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n}(x_i - \bar{x})^2
\]
Comparing estimators

There are two competing estimators for $\sigma^2$:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 \quad \text{vs} \quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1} \hat{\sigma}^2_{\text{MLE}}.$$  

The sample variance $S^2$ is unbiased, but the maximum likelihood estimator $\hat{\sigma}^2_{\text{MLE}}$ is biased:

$$E(S^2) = \sigma^2, \quad E(\hat{\sigma}^2_{\text{MLE}}) = \frac{n-1}{n} \sigma^2.$$  

However, the variance of $S^2$ is larger:

$$\text{Var}(S^2) = \text{Var}\left(\frac{n}{n-1} \hat{\sigma}^2_{\text{MLE}}\right) = \left(\frac{n}{n-1}\right)^2 \text{Var}(\hat{\sigma}^2_{\text{MLE}}) > \text{Var}(\hat{\sigma}^2_{\text{MLE}}).$$

We compare their mean squared errors ($\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$):

$$\frac{\text{MSE}(S^2)}{\text{MSE}(\hat{\sigma}^2_{\text{MLE}})} = \frac{2n^2}{2n^2 - 3n + 1} > 1.$$
Confidence interval

In addition to a point estimate, we also want a margin of error around this estimate to give a sense of uncertainty around the point estimate.

For a given value of $\alpha \in (0, 1)$ (typically $\alpha = 0.05$), the estimated $100(1 - \alpha)\%$ confidence interval of an estimator for some parameter $\theta$ is an interval $(l, u)$ such that

$$P(l < \theta < u) = 1 - \alpha.$$ 

This means that if the experiment of interest (or random sampling) is repeated 100 times and we compute 100 confidence intervals from the resulting 100 datasets, then 95 confidence intervals are expected to contain the unknown true parameter $\theta$. Thus, a single confidence interval may or may not contain the true parameter $\theta$. 
Confidence interval for $\mu$ (\(\sigma\) known)

**Example:** Luminosity function for clusters in the Milky Way. Assume that \(\sigma\) is completely known (\(\sigma = 1.2\) mag in vdB (1985)). Given a random sample of globular clusters, \(X_1, \ldots, X_n \sim N(\mu, \sigma^2)\), what is a 95% confidence interval for \(\mu\)?

We first need to derive the probability distribution of the sample mean \(\bar{X}\). \(\bar{X}\) is Normally distributed because any linear combination of Normal random variables is still a Normal random variable. To uniquely determine its Normal distribution, we need compute its mean and variance:

\[
E(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu, \quad \text{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n}.
\]

Thus,

\[
\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \quad \text{and thus} \quad Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).
\]
Confidence interval for $\mu$ ($\sigma$ known) (cont.)

From the property of the standard Normal distribution, we know that

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

If we re-arrange the inequalities inside the above probability in terms of $\mu$,

$$P(-1.96 < Z < 1.96) = P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95.$$
Confidence interval for $\mu$ (\(\sigma\) known) (cont.)

The resulting interval

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

is a 95\% confidence interval for $\mu$ when we estimate $\mu$ by $\bar{X}$.

A confidence interval is a random interval. The randomness comes from which data we observe (the realization of $\bar{X}$, i.e., $\bar{x}$, depends on which data we observe). This is why “a” 95\% confidence interval (out of infinitely many possible intervals), not “the” 95\% confidence interval.

In vdB (1985), 148 ($= n$) Galactic globular clusters are observed, and the observed sample mean is $\bar{x} = -7.1$ mag. Assuming that $\sigma = 1.2$ mag, a 95\% confidence interval for the population mean magnitude $\mu = M_0$ is

$$\left(-7.1 - 1.96 \frac{1.2}{\sqrt{148}}, \quad -7.1 + 1.96 \frac{1.2}{\sqrt{148}}\right) = (-7.29, \ -6.91).$$

If we had observed a different data set, then this confidence interval would have been different.
Confidence interval for $\mu$ (\(\sigma\) unknown)

Previously, we use the following fact to derive the confidence interval:

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

What if we do not know the value of $\sigma$? A principle in statistics is to replace any unknown parameter with a good estimator.

When a random sample of size $n$ (observed data) comes from a Normally distributed population with mean $\mu$ and variance $\sigma^2$, a good estimator for $\sigma^2$ is the sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \quad \left(= \frac{n}{n-1} \hat{\sigma}_{\text{MLE}}^2 \right)$$

If we replace $\sigma$ with $S$, what is the distribution of

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim ?$$
CONFIDENCE INTERVAL FOR $\mu$ ($\sigma$ UNKNOWN) (CONT.)

Representation of the $t_\nu$ distribution with $\nu$ degrees of freedom:
Let $Z \sim N(0, 1)$ and independently $X \sim \chi^2_\nu$, then

$$T = \frac{Z}{\sqrt{X/\nu}} \sim t_\nu \text{ distribution.}$$

Three important results in probability theory: If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$,

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

2. $X = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

3. $Z$ and $X$ are independent because $\overline{X}$ and $S^2$ are independent.

Then,

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{Z}{\sqrt{X/(n-1)}} \sim t_{n-1} \text{ distribution.}$$
Confidence interval for $\mu$ ($\sigma$ unknown) (cont.)

t distribution is heavy-tailed: For example, if $T \sim t_5$ ($n = 6$)

$P(-1.96 < Z < 1.96) = 0.95$ and $P(-2.57 < T < 2.57) = 0.95$

In this case, a 95% confidence interval for $\mu$ (when $\sigma^2$ is unknown) is

$$\left( \bar{X} - 2.57 \frac{S}{\sqrt{6}}, \bar{X} + 2.57 \frac{S}{\sqrt{6}} \right)$$

because

$$P\left(-2.57 < \frac{\bar{X} - \mu}{S/\sqrt{6}} < 2.57\right) = P\left(\bar{X} - 2.57 \frac{S}{\sqrt{6}} < \mu < \bar{X} + 2.57 \frac{S}{\sqrt{6}}\right) = 0.95.$$
Confidence interval for $\mu$ ($\sigma$ unknown) (cont.)

As the sample size $n$ increases, the $t_{n-1}$ distribution approaches the standard Normal distribution.

Thus, if $n$ is large (typically $n > 30$), $t_{n-1} \approx N(0, 1)$, and a 95% confidence interval for $\mu$ becomes close to

$$\left( \bar{X} - 1.96 \frac{S}{\sqrt{n}}, \quad \bar{X} + 1.96 \frac{S}{\sqrt{n}} \right).$$
Summary of confidence interval for \( \mu \)

Let \( X_1, X_n, \ldots, X_n \) be a random sample from \( N(\mu, \sigma^2) \).

1. If \( \sigma^2 \) is known, a 100(1 − \( \alpha \))% confidence interval for \( \mu \) is

\[
\left( \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right),
\]

where \( z_{\alpha/2} \) is a constant satisfying \( P(Z > z_{\alpha/2}) = \frac{\alpha}{2} \) if \( Z \sim N(0, 1) \).

2. If \( \sigma^2 \) is unknown, a 100(1 − \( \alpha \))% confidence interval for \( \mu \) is

\[
\left( \bar{X} - c_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + c_{\alpha/2} \frac{S}{\sqrt{n}} \right),
\]

where \( c_{\alpha/2} \) is a constant such that \( P(T > c_{\alpha/2}) = \frac{\alpha}{2} \) if \( T \sim t_{n-1} \).

3. If \( \sigma^2 \) is unknown and \( n \) is large (\( n > 30 \)), a 100(1 − \( \alpha \))% confidence interval for \( \mu \) is

\[
\left( \bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right).
\]
**Confidence interval for \( \sigma^2 \)**

Let \( X_1, X_2, \ldots, X_n \) be a random sample from \( N(\mu, \sigma^2) \), and we are interested in finding a 95% confidence interval for \( \sigma^2 \).

The sample variance \( S^2 \) is a good estimator for \( \sigma^2 \), and we know that

\[
\frac{(n - 1)S^2}{\sigma^2} \sim \chi^2_{n-1}.
\]

Then, using the \( \chi^2_{n-1} \) distribution, we can find \((a, b)\) such that

\[
P\left(\frac{(n - 1)S^2}{\sigma^2} < a\right) = P\left(\frac{(n - 1)S^2}{\sigma^2} > b\right) = \frac{\alpha}{2}.
\]

By re-arranging the inequalities inside the probability above, a 95% confidence interval for \( \sigma^2 \) becomes

\[
\left(\frac{(n - 1)S^2}{b}, \frac{(n - 1)S^2}{a}\right).
\]
Confidence interval for $\sigma^2$ (cont.)

For example, if $n = 148$, then $a = 115.33$ and $b = 182.46$, which makes each tail area of the $\chi^2_{147}$ distribution equal 0.025.

We re-arrange the inequalities inside the probability below in terms of $\sigma^2$:

$$P\left( 115.33 < \frac{(147)S^2}{\sigma^2} < 182.46 \right) = P\left( \frac{(147)S^2}{182.46} < \sigma^2 < \frac{(147)S^2}{115.33} \right) = 0.95.$$ 

Thus, with the observed sample variance $s^2 = 1.1^2$, a 95% confidence interval for $\sigma^2$ is

$$\left( \frac{(147)1.1^2}{182.46}, \frac{(147)1.1^2}{115.33} \right) = (0.97, 1.54).$$
Hypothesis testing

It is known that $M_0 = -7.7$ mag for the M31 globular clusters. A researcher feels that $M_0 \neq -7.7$ and collects a random sample of data from M31. A natural question is

“Are the data strongly in support of the claim that $M_0 \neq -7.7$ mag?”

Hypothesis: A statement about parameters of a population.

- Null hypothesis $H_0$: Status quo.
- Alternative hypothesis $H_a$: What a researcher wants to argue.

For example, $H_0$: $M_0 = -7.7$ mag, $H_a$: $M_0 \neq -7.7$ mag.

An alternative hypothesis with ‘$\neq$’ is called a two-sided hypothesis because it means either $M_0 > -7.7$ mag or $M_0 < -7.7$ mag.
Hypothesis testing: A procedure for comparing observed data with a hypothesis whose plausibility is to be assessed. Three ingredients:

- Significance level $\alpha$ (typically $\alpha = 0.05$) is the probability of rejecting $H_0$ when $H_0$ is correct.

  Note that the common standard in astronomy of ‘3$\sigma$’ corresponds to $\alpha = 0.003$ for a model with $N(\mu, \sigma^2)$.

- A test statistic is calculated from the observed data, and will measure the compatibility of $H_0$ with the observed data.

- A rejection rule which specifies the values of the test statistic for which we reject $H_0$ at the significance level $\alpha$. 
**Terminology**

**Test statistic and critical value:** Let $X_1, \ldots, X_n$ be a random sample from a distribution with parameter $\theta$. Let $T = g(X_1, \ldots, X_n)$ be a statistic and let $R$ be a subset of the real line. For a specific set of hypotheses ($H_0$ vs. $H_A$) in a testing problem, suppose we choose to “reject $H_0$ if $T \in R$”. Then $T$ is called a test statistic and $R$ is called the critical region or rejection region of the test.

A rejection region is usually defined in terms of a critical value $c$. If it makes sense to reject $H_0$ when $T \geq c$, then find $c$ such that $P(T \geq c \mid H_0 \text{ is true}) = \alpha$. 


**Terminology (cont.)**

**Type I error rate**  
\[ \alpha = P(H_0 \text{ is rejected} \mid H_0 \text{ is true}) \]  
\[ = P(\text{Incorrectly rejecting } H_0) \]  
\[ = P(\text{Making a false-positive decision}) \]

E.g., when searching for a faint signal in noise with following hypotheses

\[ H_0 : \text{Signal does not exist} \quad \text{vs} \quad H_a : \text{Signal exists}, \]

Type I error occurs when we incorrectly infer that a signal is present although it truly is not.

**Testing power**  
\[ = P(H_0 \text{ is rejected} \mid H_a \text{ is true}) \]  
\[ = P(\text{Correctly rejecting } H_0) \]

**Type II error rate**  
\[ \beta = P(H_0 \text{ is NOT rejected} \mid H_a \text{ is true}) \]  
\[ = P(\text{Making a false-negative decision}) \]  
\[ = 1 - (\text{Testing power}) \]

E.g., incorrectly infer that a signal is absent when it truly is present.
**Uniformly most powerful test**: Ideally, we want to minimize both Type I and Type II error rates. But it is impossible to minimize two rates simultaneously, and thus we fix the Type I error rate at $\alpha$ and maximize the testing power (or minimize Type II error rate).

Hypothesis testing based on a test statistic that gives the highest test power for all possible parameter values under $H_0$ at a chosen significance level $\alpha$ is called the uniformly most powerful test.
**Example: One Sample t-test**

Luminosity function for clusters in the Milky Way. A random sample of 148 measurements has sample mean $\bar{x} = -7.5$ and sample variance $s^2 = 1.1^2$. From the data, a researcher feels that $\mu (= M_0) \neq -7.7$ mag.

1. Specify the null and alternative hypotheses:
   $$H_0 : \mu = -7.7 \text{ mag} \quad \text{vs} \quad H_a : \mu \neq -7.7 \text{ mag}.$$  

2. Set up a test statistic for $\mu$:
   $$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

3. Find the distribution of the test statistic under the null: As $n$ is large,
   $$T_0 = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} + 7.7}{S/\sqrt{n}} \sim N(0, 1)$$

4. Set up the rejection rule: Reject $H_0$ if $|T_0| > 1.96$. Otherwise, we fail to reject $H_0$. This choice of cut-off point (critical value) results in a 5% level of significance of the test of hypotheses.
Example: One Sample t-test (cont.)

5. Calculate the value of the test statistic using the observed data.

\[
\frac{\bar{x} + 7.7}{s/\sqrt{n}} = \frac{-7.5 + 7.7}{1.1/\sqrt{148}} = 2.21.
\]

6. We reject the null hypothesis \(H_0\) because the calculated (absolute) value of the test statistic exceeds the critical value, 1.96. We report that there is a statistically significant difference between the population mean and the hypothesized value \(-7.7\) mag.
Two other popular ways to conduct the same hypothesis testing:

- Derive a 95% confidence interval for $\mu$. Since $n$ is large enough,

$$
\left( \bar{x} - 1.96 \frac{s}{\sqrt{n}}, \; \bar{x} + 1.96 \frac{s}{\sqrt{n}} \right) = (-7.68, \; -7.32).
$$

This 95% confidence interval for $\mu$ does not contain the hypothesized value under the null hypothesis, i.e., $\mu = -7.7$, and thus we reject the null hypothesis at the 5% significant level.

Hypothesis testing at $\alpha$ via a $100(1 - \alpha)$% confidence interval is valid for two-sided hypothesis testing procedures.
Example: One Sample \( t \)-Test (Cont.)

- Compute the \( p \)-value, i.e., the probability of observing the current test statistic or a more extreme one in the direction of the alternative, given that \( H_0 \) is correct. If \( H_a \) is two-sided,

\[
P \left( T_0 > \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \text{ or } T_0 < -\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \right) = 2P \left( T_0 > \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \right).
\]

The \( p \)-value in the example is \( 2P(T_0 > 2.21) = 0.03 \) that is smaller than the significance level 0.05. Thus, at the 5% significant level, there is enough evidence in data to reject \( H_0 \) (or to support \( H_a \)).

Note that the \( p \)-value is not the probability of \( H_0 \) being correct (the most common misunderstanding!).
**Example: One Sample Test on $\sigma^2$**

**Example:** Luminosity function for clusters in the Milky Way. Given a random sample of globular clusters, $X_1, \ldots, X_{148} \sim N(\mu, \sigma^2)$, vdB (1985) assumes that $\sigma = 1.2$ mag. Is this assumption consistent to the data?

1. Specify the null and alternative hypotheses:
   
   \[ H_0 : \sigma^2 = 1.2^2 \text{ (mag}^2) \text{ vs } H_a : \sigma^2 \neq 1.2^2 \text{ (mag}^2) \text{).} \]

2. Set up a test statistic for $\sigma^2$:
   
   \[ T = \frac{(147)S^2}{\sigma^2} = \frac{\sum_{i=1}^{148}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{147} \]

3. Find the distribution of the test statistic under the null:
   
   \[ T_0 = \frac{(147)S^2}{1.2^2} = \frac{\sum_{i=1}^{148}(X_i - \bar{X})^2}{1.2^2} \sim \chi^2_{147} \]
**Example: One Sample Test on $\sigma^2$ (Cont.)**

4. Set up the rejection rule at significance level $\alpha = 0.05$: Reject $H_0$ if

$$T_0 > 182.46 \text{ or } T_0 < 115.33,$$

where 123.52 and 182.46 are chosen to satisfy that when $X \sim \chi^2_{147}$

$$P(X > 182.46) = P(X < 115.33) = 0.025 \ (= \alpha/2)$$

5. Compute the test statistic and make a decision: With $s^2 = 1.1^2$,

$$T_0 = \frac{147 \times 1.1^2}{1.2^2} = 123.52.$$

At $\alpha = 0.05$, there is not enough evidence in the data to reject $H_0$. 

![Graph showing critical values and observed test statistic](image-url)
Another popular way to conduct the same hypothesis testing:

- Derive a 95% confidence interval for $\sigma^2$:
  \[
  \left( \frac{(n - 1)S^2}{182.46}, \frac{(n - 1)S^2}{115.33} \right) = (0.98, 1.54).
  \]

  This 95% confidence interval for $\sigma^2$ contains the hypothesized value under the null hypothesis, i.e., $\sigma^2 = 1.2^2 = 1.44$, and thus there is not enough evidence in the observed data to reject the null hypothesis at the 5% significant level.

- For the hypothesis testing for $\sigma^2$, the $p$-value approach is available only in a one-sided test. E.g., if $H_0 : \sigma^2 = 1.2$ vs $H_a : \sigma^2 < 1.2$,
  \[
p\text{-value} = P(T_0 < 123.52) = 0.079.
  \]
  Since the $p$-value is greater than the significance level $\alpha = 0.05$, we fail to reject $H_0$ at $\alpha = 0.05$. 

**Example: Two Sample t-Test**

We have two different populations of interest and obtain a sample of sizes $n_1$ from population 1 and independently a sample of size $n_2$ from population 2:

$$X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2).$$

Our primary interest is to see whether two population means are different.

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_a : \mu_1 \neq \mu_2.$$
**Example: Two Sample \( t \)-Test (cont.)**

1. Set up a test statistic for \( \mu_1 - \mu_2 \) under \( H_0 \): A good choice is \( \overline{X} - \overline{Y} \).
   
   Case 1: If \( \sigma_1^2 \) and \( \sigma_2^2 \) are known,
   
   \[
   T_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)
   \]
   
   Case 2: If \( \sigma_1^2 \) and \( \sigma_2^2 \) are unknown with large sample sizes, \( n_1, n_2 \),
   
   \[
   T_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}} \sim N(0, 1)
   \]
   
   Case 3: If \( \sigma_1^2 \) and \( \sigma_2^2 \) are unknown with an assumption that \( \sigma_1^2 = \sigma_2^2 \),
   
   \[
   T_0 = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}
   \]
   
   where \( S_p^2 = \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2 + \sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_1+n_2-2} \).

2. Set up the rejection rule: Reject \( H_0 \) if \( |T_0| > C_{\alpha/2} \), where \( C_{\alpha/2} \) is a constant such that \( P(|T_0| > C_{\alpha/2} \mid H_0 \text{ is correct}) = \alpha/2 \).
EXAMPLE: TWO SAMPLE $t$-TEST (CONT.)

Two other popular ways to conduct the same hypothesis testing:

- Derive a 95% confidence interval for $\mu_1 - \mu_2$.

\[
\text{Case 1: } \left( (\bar{X} - \bar{Y}) - 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{X} - \bar{Y}) + 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)
\]

\[
\text{Case 2: } \left( (\bar{X} - \bar{Y}) - 1.96 \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}}, (\bar{X} - \bar{Y}) + 1.96 \sqrt{\frac{S_X^2}{n_1} + \frac{S_Y^2}{n_2}} \right)
\]

\[
\text{Case 3: } \left( (\bar{X} - \bar{Y}) - 1.96S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\bar{X} - \bar{Y}) + 1.96S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)
\]

We reject the null hypothesis at the 5% significant level, if this 95% confidence interval for $\mu_1 - \mu_2$ does not contain zero (i.e., $\mu_1 = \mu_2$).
Example: Two Sample t-Test (cont.)

- Compute the $p$-value, i.e., the probability of observing the current test statistic or a more extreme one in the direction of the alternative, given that $H_0$ is correct. Since $H_a$ is two-sided,

$$\text{Case 1: } p\text{-value} = 2P\left( T_0 > \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right)$$

$$\text{Case 2: } p\text{-value} = 2P\left( T_0 > \frac{|\bar{x} - \bar{y}|}{\sqrt{s_x^2 \frac{1}{n_1} + s_y^2 \frac{1}{n_2}}} \right)$$

$$\text{Case 3: } p\text{-value} = 2P\left( T_0 > \frac{|\bar{x} - \bar{y}|}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)$$

We conclude that there is enough evidence in data to reject $H_0$ (or to support $H_a$) if this $p$-value is smaller than $\alpha = 0.05$. 
**Example: Two Sample $F$-Test**

The same two-sample setting as before:

$$X_1, X_2, \ldots, X_{n_1} \sim N(\mu_1, \sigma_1^2) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_{n_2} \sim N(\mu_2, \sigma_2^2).$$

Our primary interest is to see whether two population variances are different.

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad H_a : \sigma_1^2 \neq \sigma_2^2.$$
Example: Two Sample $F$-test (cont.)

This hypothesis testing is the same as whether the variance ratio is one:

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \ vs \ H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1.$$

A good estimator for the ratio is $\frac{S_X^2}{S_Y^2}$. What is its distribution?

Representation of $F_{a,b}$ distribution: If $X \sim \chi^2_a$ and independently $Y \sim \chi^2_b$,

$$F = \frac{X/a}{Y/b} \sim F_{a,b} \text{ distribution}$$

Since $\frac{(n_1-1)S_X^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ and independently $\frac{(n_2-1)S_Y^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$,

$$T = \frac{\frac{(n_1-1)S_X^2}{\sigma_1^2}/(n_1 - 1)}{\frac{(n_2-1)S_Y^2}{\sigma_2^2}/(n_2 - 1)} = \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}.$$

Under the null hypothesis,

$$T_0 = \frac{S_X^2}{S_Y^2} \sim F_{n_1-1,n_2-1}.$$
**Example: Two Sample $F$-Test (cont.)**

It makes sense to reject $H_0$ if this ratio of two sample variances is much larger or smaller than one. Thus, the decision rule at $\alpha$ is to reject $H_0$ if

$$T_0 > C_{\alpha/2} \text{ or } T_0 < C_{1-\alpha/2},$$

where $C_{\alpha/2}$ and $C_{1-\alpha/2}$ are constants such that when $F \sim F_{n_1-1,n_2-1}$

$$P(F > C_{\alpha/2}) = P(F < C_{1-\alpha/2}) = \alpha/2.$$

The same hypothesis test can be done by checking whether a 95% confidence interval for $\sigma_1^2/\sigma_2^2$ contains one ($\sigma_1^2/\sigma_2^2 = 1$): The interval is

$$\left(\frac{S_X^2}{S_Y^2 C_{0.025}}, \frac{S_X^2}{S_Y^2 C_{0.975}}\right)$$

because

$$P\left(C_{0.975} < \frac{S_X^2/\sigma_1^2}{S_Y^2/\sigma_2^2} < C_{0.025}\right) = P\left(\frac{S_X^2}{S_Y^2 C_{0.025}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_X^2}{S_Y^2 C_{0.975}}\right) = 0.95.$$
**Likelihood ratio test**

**Likelihood ratio test statistic**: One way to form a test statistic is to compare the value of the likelihood functions under the two hypotheses: \( H_0 : \theta \in \Omega_0 \) vs \( H_a : \theta \in \Omega_1 \). Let the likelihood ratio statistic be

\[
LR = \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)},
\]

where \( \Omega = \Omega_0 \cup \Omega_1 \) is the entire parameter space of \( \theta \). It makes sense to reject \( H_0 \) if \( LR \leq c \) for some critical value \( c \).

**Asymptotic distribution of the likelihood ratio test statistic**: When \( n \) is large and you have a two-sided alternative hypothesis, \( H_0 : \theta = \theta_0 \) vs \( H_A : \theta \neq \theta_0 \), then there is a standard asymptotic likelihood ratio test based on the log likelihood ratio

\[
-2 \log(LR) = -2 \log \left( \frac{\max_{\theta \in \Omega_0} L(\theta)}{\max_{\theta \in \Omega} L(\theta)} \right) \sim \chi^2_p,
\]

where \( p \) is the number of (free) parameters in \( \Omega \) minus that in \( \Omega_0 \). This test is valid if the model in \( H_0 \) is a special case of the one in \( H_a \) (nested).
**Likelihood ratio test (cont.)**

**Example:** A simple linear regression model with the observed data $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ with independent Gaussian measurement error is specified as $y_i = \alpha + \beta x_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, i.e.,

$$y_i \sim N(\alpha + \beta x_i, \sigma^2)$$

A relation between distance and radial velocity among extra-galactic nebulae (Hubble, 1929)

We are interested in whether the slope $\beta$ is non-zero (statistically significant), i.e., the corresponding hypotheses are

$$H_0 : \beta = 0 \; \text{vs} \; H_a : \beta \neq 0.$$
The likelihood and log-likelihood functions for \( \theta = (\alpha, \beta, \sigma^2) \) are

\[
L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left( -\frac{\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2}{2\sigma^2} \right),
\]

\[
\ell(\theta) = \ln(L(\theta)) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2}{2\sigma^2}.
\]

The maximum likelihood estimates for \( \theta \) are

\[
\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2
\]

Thus, under the entire parameter space, \( \Omega = \Omega_0 \cup \Omega_1 \),

\[
\max_{\theta \in \Omega} L(\theta) = L(\hat{\theta}) = 6.793551e-72.
\]
Likelihood ratio test (cont.)

In the parameter space under $H_0 : \beta = 0$, the MLEs for $\alpha$ and $\sigma^2$ are

$$\hat{\alpha}_0 = \bar{y} \quad \text{and} \quad \hat{\sigma}^2_0 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\alpha}_0)^2,$$

where the subscript zero denotes that these estimates are under $H_0$. Collectively, $\hat{\theta}_0 = (\hat{\alpha}_0, \beta = 0, \hat{\sigma}^2_0)$. Then, under the null space $\Omega_0$,

$$\max_{\theta \in \Omega_0} L(\theta) = L(\hat{\theta}_0) = 5.692142e-77.$$

The likelihood ratio test statistic is

$$LR = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} = \frac{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n} (y_i - \hat{\alpha}_0)^2}{2\hat{\sigma}_0^2}\right)}{\left(\frac{1}{\sqrt{2\pi\hat{\sigma}}}\right)^n \exp\left(-\frac{\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{2\hat{\sigma}^2}\right)},$$

and we can conduct the hypothesis test using its asymptotic distribution:

$$-2 \log(LR) \sim \chi^2_1.$$

At $\alpha = 0.05$, we reject $H_0$ if $-2 \log(LR) > 3.84$ where 3.84 is a constant such that $P(X > 3.84) = 0.05$ when $X \sim \chi^2_1$. The value of $-2 \log(LR)$ in this example is 23.38, and thus we reject $H_0$ at $\alpha = 0.05$. 
**Likelihood ratio test (cont.)**

**Example:** What if we want to test the following hypotheses?

\[
H_0 : \alpha = \beta = 0 \quad \text{vs} \quad H_a : H_0 \text{ is not true}.
\]

Under the entire parameter space, \( \Omega = \Omega_0 \cup \Omega_1 \), the MLEs are the same:

\[
\max_{\theta \in \Omega} L(\theta) = L(\hat{\theta}).
\]

Under the null space \((\alpha = \beta = 0)\), the MLE for \(\sigma^2\) is \(\hat{\sigma}^2_0 = \frac{1}{n} \sum_{i=1}^{n} y_i^2\).

The likelihood ratio test statistic is

\[
LR = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} = \frac{\left(\frac{1}{\sqrt{2\pi \hat{\sigma}_0}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{y_i^2}{2\hat{\sigma}_0^2}\right)}{\left(\frac{1}{\sqrt{2\pi \hat{\sigma}}}\right)^n \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}{2\hat{\sigma}^2}\right)} = 1.501873 \times 10^{-9},
\]

and we can conduct the hypothesis test using its asymptotic distribution:

\[-2 \log(LR) \sim \chi^2_2.\]

At \(\alpha = 0.05\), we reject \(H_0\) if \(-2 \log(LR) > 5.99\) where 5.99 is a constant such that \(P(X > 5.99) = 0.05\) when \(X \sim \chi^2_2\). The observed value of \(-2 \log(LR)\) is 40.63.
Penalized likelihood approaches have dominated model selection since the 1980s due to several limitations of the likelihood ratio test:

- LRT is only applicable to nested models.
- If a model $M_1$ is nested within another model $M_2$, the largest likelihood achievable by $M_2$ will always be larger than that achievable by $M_1$ simply because $M_2$ has more parameters (enabling $M_2$ to explain the data more elaborately).

If a penalty is applied to compensate for the obligatory difference in likelihoods due to the different number of parameters in $M_1$ and $M_2$, the desired balance between overfitting and underfitting might be found.
Model selection via information criteria (cont.)

- **Akaike information criterion** (AIC, 1973) is defined as
  \[
  \text{AIC} = -2\ell(\hat{\theta}) + 2p = (\text{goodness-of-fit}) + (\text{penalty}),
  \]
  where \(\ell(\hat{\theta})\) is the maximized log likelihood (evaluated at the MLE \(\hat{\theta}\)) and \(p\) is the number of parameters (dimension of \(\theta\)). The penalty term increases as the complexity of the model grows, and thus compensates for the necessary increase in the likelihood. A model with the ‘smallest’ AIC, i.e., a model that explains the data well with a small number of parameters, is preferred.

- **Bayesian information criterion** weights the penalty according to \(n\).
  \[
  \text{BIC} = -2\ell(\hat{\theta}) + p \log(n)
  \]
  As we collect more data, the penalty on an additional parameter becomes stronger than that of AIC. Thus, when \(n\) is large, BIC prefers even more parsimonious models than AIC does.
Model selection via information criteria (cont.)

Model 1: \( y_i = \alpha + \beta x_i + \epsilon_i \)
Model 2: \( y_i = \alpha + \epsilon_i \)
Model 3: \( y_i = \epsilon_i \)

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>333.65</td>
<td>355.10</td>
<td>370.37</td>
</tr>
<tr>
<td>BIC</td>
<td>337.19</td>
<td>357.45</td>
<td>371.55</td>
</tr>
</tbody>
</table>

For both information criteria, Model 1 is preferred.

A relation between distance and radial velocity among extra-galactic nebulae (Hubble, 1929)
References

Reference

Image credit
3. Pages 13, 14, 52: https://imagingsolution.net/math/least-square-method
5. Pages 23, 44, 48: The lecture notes prepared by Professor Bing Li.