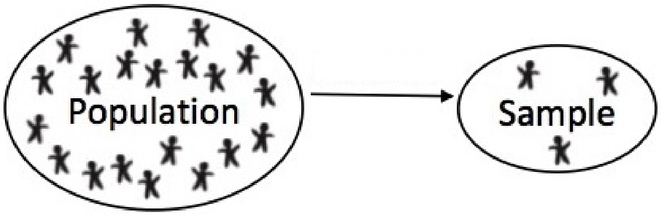
Statistical Inference

Hyungsuk Tak
Pennsylvania State University

Short review of probability

A random variable (r.v. hereafter) X is a mapping rule (deterministic function) to assign a number to each object in a sample. For example, suppose we take a random sample of size 100 from a population of interest, and define a function X that returns the magnitude of a randomly selected astronomical object.



Then, we get 100 values of X. A histogram of these 100 values shows a distribution of X, showing frequencies of possible values of X.

Let us set up a probabilistic model by assuming that these 100 values are 100 random realizations from a Normal(μ, σ^2) distribution. This assumption (model) may or may not be the case. But it appears useful for explaining possible variations (uncertainties) of the data as it captures the shape of X's distribution well.

Note that statistical modeling starts with a distributional assumption on the data.

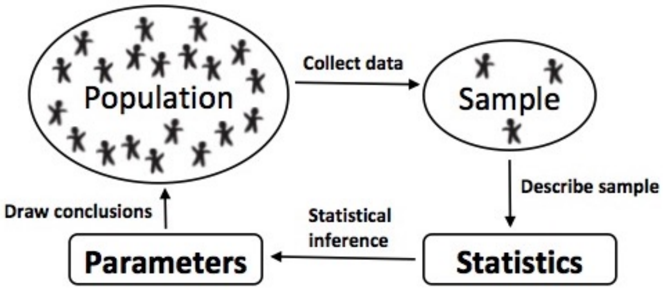
If $X \sim$ Normal$(0, 1)$, its realized value will be between $(-1, 1)$ with 68.3% (that is, within 1σ range), between $(-2, 2)$ with 95.4% (within 2σ range), and between $(-3, 3)$ with 99.7%.

If $X \sim$ Bernoulli(0.5), then about half of the random realizations will be 1. One possible sequence of these realizations is

A set of parameters in a probability distribution completely determines the shape of the distribution, and is typically denoted by a vector θ. For example, a Normal distribution has two parameters, mean (μ) and variance (σ^2). When $X \sim$ Normal(μ, σ^2), the shape of this Normal distribution will be completely determined by the values of $\theta = (\mu, \sigma^2)$.

Terminology in Statistical Inference

Interested in the proportion of Earth-like exoplanets orbiting stars within 10 pc of the Sun.



- ______________: The entire group of objects that we want to know about.

  Example: All of the exoplanets orbiting stars within 10 pc of the Sun.

- ______________: A fraction of the population from which we actually collect data.

  Example: Those observed by telescopes. Let's assume that the sample size is n.

Object	Features				
number	mag.	z	$\cdots$	M	I
1	m_1	z_1	$\cdots$	M_1	I_1
2	m_2	z_2	$\cdots$	M_2	I_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	m_n	z_n	$\cdots$	M_n	I_n

- __________________________ is an effort to explain a potential data generation process. It starts with an assumption that the data are r.v.s, and the observed data are random realizations of these r.v.s. Then, we assume a specific probability distribution on the data to better understand the randomness (uncertainty) involved in the data.

  Example: In the data, we already have numeric values of features, for example, $I_1 = 1, I_2 = 0, I_3 = 0, \ldots, I_n = 1$. Statistical inference starts with a question: What might have generated these binary values?

  In statistical modeling, population characteristics of interest are represented by unknown parameter(s) of a probability distribution.

  Example: The parameter θ is the probability of the i-th exoplanet being Earth-like.

Statistical inference is about estimating the unknown parameter(s) θ from the data.

For your reference, statistics has two big schools; frequentist and Bayesian. Is the unknown quantity of interest θ a fixed constant or an r.v. (that is, a value that may vary with some uncertainty)?

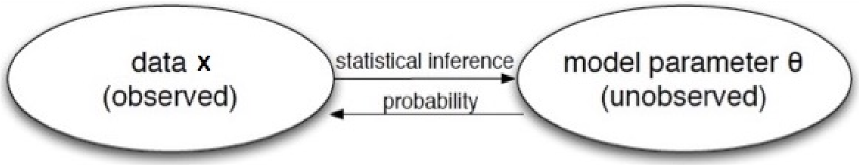
- ______________: A numerical summary (i.e., a function) of the data, such as mean, median, maximum, etc., to infer θ. This is an r.v. as a function of r.v.s (data) is an r.v.

  Example:

  A specific value of an estimator computed from the data is called an estimate.

Probability and Statistical Inference

Probability and statistical inference are two sides of the same coin.



In probability, we learn several families of probability distributions. For example, Bernoulli, Binomial, Poisson, Normal, Exponential, Gamma, Beta, etc. These are building blocks to construct a statistical model that is a set of assumptions about *probability distributions of the data* to account for the randomness of the data realizations or of data generating process.

Example: We are interested in estimating the true brightness of some galaxies from measurements. The brightness of galaxy i is measured by a certain telescope, and we assume that it is measured around the unknown true brightness with a Gaussian measurement error (Eddington, 1913). We can express this statistical model as

$$m_i = \mu_i + \epsilon_i, \quad \text{where} \;\; \epsilon_i \sim N(0,\ \sigma^2)$$

or equivalently,

$$m_i \sim N(\mu_i,\ \sigma^2),$$

meaning that the measured brightness in the data is assumed to be one realization of this Gaussian distribution.

Does the model represent the true data generation process? Absolutely not. But this model may be useful for understanding the uncertainty in the measurement process.

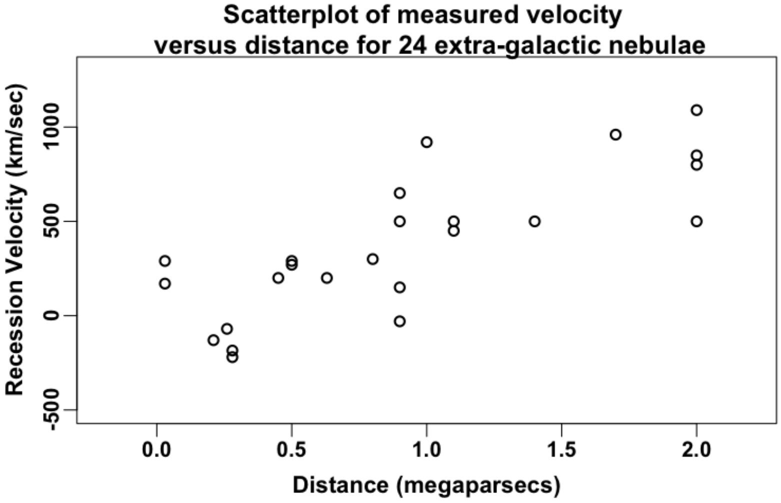
"All models are wrong, but some are useful." — George Box.

In statistical inference, our main interest is about how to obtain the most likely values of the model parameters, $\theta = (\mu_1, \mu_2, \ldots, \mu_n, \sigma^2)$, given the model and observed data $(m_1, \ldots, m_n)$.

In this lecture, we will learn four main topics in statistical inference, i.e., point estimation, interval estimation, hypothesis testing, and model comparison in light of the following example.

Hubble Constant Estimation (Hubble, 1929)

Edwin Hubble used the power of the Mount Wilson Observatory telescopes to measure features of nebulae outside the Milky Way. He was surprised to find a relationship between a nebula's distance from earth and the velocity with which it was going away from the earth. Hubble's initial data on 24 nebulae are displayed below (Hubble, 1929, Proceedings of the National Academy of Science, 15, 168–173).



The vertical axis measures the recession velocity, in kilometers per second, which was determined by the redshift in the spectrum of light from a nebula with considerable accuracy. The horizontal axis measures distance from the Earth, in megaparsecs. Distances were measured by comparing mean luminosities of the nebulae to those of certain star types, a method that is not particularly accurate.

The distance D_i between the Earth and the i-th nebula, and the velocity V_i at which they appear to be going away from each other satisfy the following physical relationship:

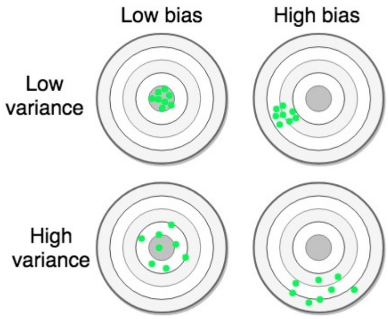
where T is the time elapsed since the Big Bang (the age of the Universe), and thus its inverse $1/T$ is the Hubble constant ($H_0 = 1/T$) in km s^{-1} Mpc^{-1}.

Let us adopt the following statistical model (assumption on the data) to check whether the observed data actually support the physical relationship between V_i and D_i.

Mean Squared Error

What is a *good* estimator (a function of the data) for estimating the parameter of interest in the model, that is, H_0? Can we say anything about the closeness of an estimate to the unknown true value of the parameter?

The most popular measure of closeness between an estimator and the unknown quantity of interest is the mean squared error (MSE), a composite measure of bias and variance



Since MSE strikes a balance between the bias and variance, an estimator with the smallest MSE is desirable. In practice, people also use the root MSE (RMSE) because its unit is the same as that of the parameter (or estimate).

Maximum Likelihood Estimation

Likelihood function (R. A. Fisher, 1922) of a model $f(x \mid \theta)$ is the joint probability density or mass function of the observed data $x = \{x_1, x_2, \ldots, x_n\}$ (fixed constants), viewed as a function of θ. For example, if the data $X = \{X_1, X_2, \ldots, X_n\}$ are modeled as continuous r.v.s whose probability density function f is parameterized by θ,

$$L(\theta) = f(x \mid \theta) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta), \text{ if } X_i\text{'s are independent.}$$

If the data are modeled by discrete r.v.s (mostly counts),

$$L(\theta) = P(X = x \mid \theta) = P(X_1 = x_1, \ldots, X_n = x_n \mid \theta) = \prod_{i=1}^{n} P(X_i = x_i \mid \theta), \text{ if independent.}$$

In this discrete case, the likelihood function is the "probability" of observing the current data $\{X = x\}$ as a function of θ. For example, let's say $L(0.8) \gg L(0.2)$. It means that the probability of observing the current data $P(X = x \mid \theta)$ $(= L(\theta))$ is much larger when $\theta = 0.8$. So, we can say that the data are supporting $\theta = 0.8$ much more than $\theta = 0.2$ (under the given model). In this sense, the likelihood function can be considered as a tool to let the data speak more about which parameter value they prefer!

Example: Pratten et al. (2020) and Driggers et al. (2019). Be careful about the notation.

Example: In the Hubble constant estimation problem, derive the likelihood function of the model parameter, $\theta = (H_0, \sigma^2)$. The adopted model is

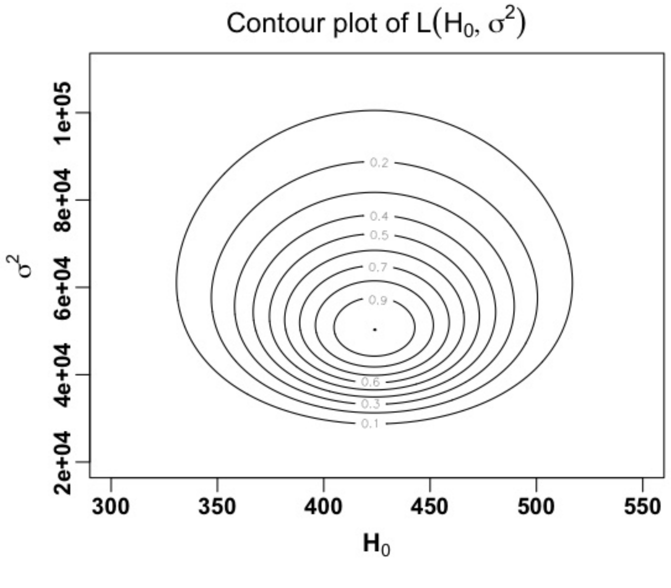
$$V_i \sim N(H_0 d_i,\ \sigma^2).$$

For your reference, the probability density function of a $N(\mu, \sigma^2)$ r.v. X is

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

The resulting likelihood function of $\theta = (H_0, \sigma^2)$ is defined as follows.

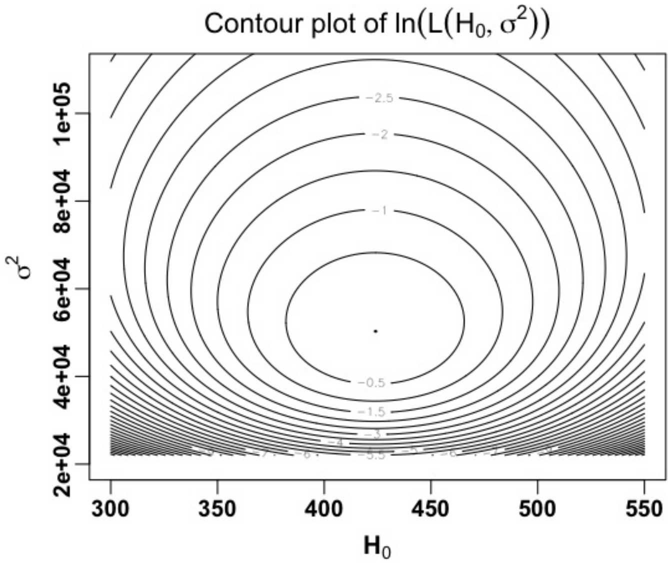
$$\begin{aligned} L(\theta) = f(v \mid \theta) = f(v_1, v_2, \ldots, v_{24} \mid \theta) &\overset{\text{ind.}}{=} \prod_{i=1}^{24} f(v_i \mid \theta) = \prod_{i=1}^{24} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(v_i - H_0 d_i)^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{24} \left(\sigma^2\right)^{-12} \exp\left(-\frac{\sum_{i=1}^{24}(v_i - H_0 d_i)^2}{2\sigma^2}\right) = c\left(\sigma^2\right)^{-12} \exp\left(-\frac{\sum_{i=1}^{24}(v_i - H_0 d_i)^2}{2\sigma^2}\right). \end{aligned}$$



Maximum likelihood estimator: A widely used method of obtaining a point estimate for a model parameter is to find the maximum likelihood estimate (MLE). As the name suggests, the MLE is defined as a value that jointly maximizes the likelihood function $L(\theta)$.

In practice, we obtain the MLE by maximizing $\ell(\theta) = \log_e(L(\theta))$ instead of maximizing $L(\theta)$ for a few reasons; (i) since $L(\theta)$ involves a multiplication when the data are independent, it is mathematically more convenient to work with the (natural) logarithm of the likelihood function. (Summation is easier to handle than multiplication mathematically!); (ii) Also, the logarithmic function is strictly increasing, preserving the maximizing value, i.e., the value of θ that maximizes $\ell(\theta)$ also maximizes $L(\theta)$; (iii) lastly, when an analytic solution is not available, we need to find a numerical solution. In this case it is computationally more stable to find the value of θ that maximizes $\ell(\theta)$.

$$\ell(\theta) = \ln(L(\theta)) = \ln(c) - 12\ln(\sigma^2) - \frac{\sum_{i=1}^{24}(v_i - H_0 d_i)^2}{2\sigma^2}.$$



Why MLE? Asymptotic optimality of MLE: As the data size increases ($n \uparrow$), MLE becomes an unbiased, most efficient (smallest variance), and approximately Normally distributed estimator:

$$\hat{\theta}_{\text{MLE}} \dot{\sim} N\Big(\theta,\ \hat{\tau}_n^2\Big)\,.$$

As n becomes large, no other estimator can have MSE smaller than the MSE of $\hat{\theta}_{\text{MLE}}$. So for many problems involving a large number of observations, the MLE becomes the Normally-distributed minimum-variance unbiased estimator.

The (asymptotic) uncertainty of the MLE $\hat{\theta}$, denoted by $\hat{\tau}_n^2$ above, can be computed numerically with the Hessian matrix that is typically produced as a byproduct of optimization.

Limitations: The MLE is sometimes biased with the sample size is small. Also, it does not always provide a closed-form solution. In this case, algorithms for optimization, such as Newton-Raphson and Expectation-Maximization algorithms, are used.

Confidence Interval

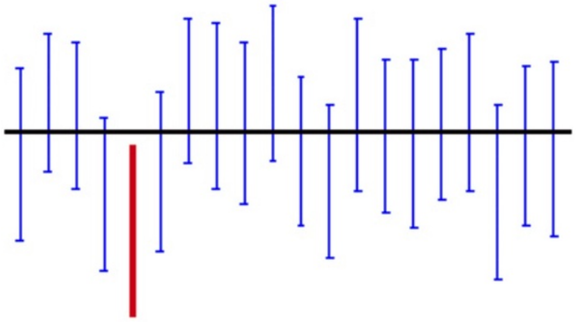
In addition to a point estimate for H_0, that is, $\hat{H}_{0,MLE} = 423.816$, we also want a margin of error around this estimate to give a sense of uncertainty around the point estimate. Astronomers typically report an interval based on 1σ error bar. For example,

What is the meaning of this (confidence, not Bayesian) interval? The most common mistake is to think of some probability of H_0 being in this interval.

Rigorously speaking, for a given value of $\alpha \in (0, 1)$ (typically $\alpha = 0.05$ in statistics and $\alpha = 0.32$ in astronomy), an $100(1-\alpha)\%$ confidence interval for H_0 is defined as an interval $(l(X),\ u(X))$ such that

This probability is taken over all possible (random) realizations of the data X for a fixed and unknown value of H_0. That means, it is the interval that is random, not the parameter H_0. The randomness comes from which data we observe (via random sapling). Thus, when $\alpha = 0.05$, we call it "a" 95% confidence interval (out of infinitely many possible intervals according to which random sample we get), not "the" 95% confidence interval.

Its interpretation must be made under a hypothetical scenario; if the experiment of interest (or sampling procedure of the data) were repeated 1,000 times under the same condition, computing 1,000 confidence intervals from the resulting 1,000 data sets, then 95% of these (hypothetical) intervals are expected to contain the unknown true parameter H_0.



In reality, we obtain just one of these hypothetical confidence intervals because we only obtain one data set and cannot repeat the data generating process under the same condition over and over again. This interval we obtain may be one of the blue intervals above, containing the true value of H_0, or a red one not containing H_0; no one knows which one is what we obtain from the data. Note that it is 'the way we construct confidence intervals' that guarantees the correct coverage (95%), not a single interval.

Approximate Confidence Interval based on MLE

A confidence interval for an unknown parameter is often analytically tractable if the data are Gaussian. However, the distribution of the data is not always Gaussian in astronomy, such as (Poisson) photon counts. How can we derive a confidence interval regardless of the data distribution?

We can construct an approximate confidence interval using the asymptotic optimality of an MLE. That is, we know that as $n \to \infty$,

$$\hat{\theta}_{MLE} \dot{\sim} N\Big(\theta,\ \hat{\tau}_n^2\Big) \quad \text{or equivalently} \quad \frac{\Big(\hat{\theta}_{MLE} - \theta\Big)}{\hat{\tau}_n} \dot{\sim} N(0,\ 1)\,.$$

Then, we can construct approximate 95% and 68% confidence intervals as follows.

$$\begin{aligned}&\text{An approximate 95\% confidence interval:} \quad \Big(\hat{\theta}_{MLE} - 1.96\hat{\tau}_n,\quad \hat{\theta}_{MLE} + 1.96\hat{\tau}_n\Big)\,.\\ &\text{An approximate 68\% confidence interval:} \quad \Big(\hat{\theta}_{MLE} - \hat{\tau}_n,\quad \hat{\theta}_{MLE} + \hat{\tau}_n\Big)\,.\end{aligned}$$

Example: Given the optimization result, find an approximate 68% confidence interval for H_0.

We interpret this 68% confidence interval in the following way. Hypothetically, if we were able to repeat the data collection process, computing a 68% confidence interval each time, then 68% of these intervals would contain H_0. The computed interval $(382.501,\ 465.131)$ is just one of these possible 68% intervals that may or may not contain the true value of H_0. It is the way we construct the interval that guarantees the correct coverage property.

Hypothesis Testing

Estimation is about pinning down the underlying values of unknown parameters from a potentially large number of possibilities with plausible ranges of parameter values. On the other hand, hypothesis testing asks the following: Given two hypotheses about the parameter value, which one do the data support more?

Example: In mathematics, a function of a line is generally formulated as $y = a + bx$. The physically-motivated model in Hubble (1929) assumes that the intercept a is 0, that is, $V_i = 0 + H_0 D_i$, forcing the line to pass the origin. Do the data also support this physically-motivated assumption that $a = 0$? For reference, a represents the peculiar velocity in cosmology.

Terminology in Hypothesis Testing

Statistical hypothesis is a statement about parameter(s) θ of a probability distribution. Note that the quantity of interest is represented by the parameter θ under a statistical model.

Null vs alternative: The null hypothesis H_0 indicates status quo (default or baseline case), and the alternative hypothesis H_1 represents what a researcher wants to argue.

Example:

Test statistic: Let $T = h(X_1, \ldots, X_n)$ be a function of the data to be used to determine which hypothesis is more supported by the data. Then we call it a test statistic.

Significance level α (the same α used in the confidence interval): A test with significance level, e.g., $\alpha = 0.05$, means that the test controls the probability of rejecting H_0 when H_0 is the case, that is, $P(\text{reject } H_0 \mid H_0 \text{ is the case})$, to be smaller than or equal to 0.05.

Principle in hypothesis testing: We assume that H_0 is the case, find the distribution of the test statistic under H_0, and check whether the evidence in the data (a value of T computed from the observed data) is within the probable range of this distribution.

Example: Figure 4 of Abbott et al. (2016).

Likelihood Ratio Test

Now, we learn the most fundamental theoretical background of hypothesis testing. Let's consider the simplest situation where we test the following hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

We want to make a choice between these two possibilities based on the data we observe. It turns out that there is a general procedure with which one can construct a good testing procedure. The idea is related to the maximum likelihood principle we have already learned.

Comparing likelihoods: What are the values of the likelihood function under the two hypotheses? Comparing these values, it makes sense to choose a hypothesis with the higher likelihood, that is, a hypothesis that the data support more. This idea is to reject H_0 if

A testing procedure based on a likelihood comparison is called a *likelihood ratio test*.

It turns out that likelihood ratio tests are the "best" (most powerful) tests under mild conditions, which is theoretically proven in the Neyman-Pearson Lemma. Specifically, the Lemma says that the likelihood ratio test maximizes the probability of rejecting H_0 when H_0 is not the case, that is, P(correct scientific discovery), given a specific significance level α.

Likelihood Ratio Test based on MLE

The likelihood ratio test statistic T is derived from the MLE of the parameter being tested.

Example: In the Hubble constant estimation problem, a mathematically-motivated model including an intercept term a is

$$V_i = a + H_0 D_i + \epsilon_i, \quad \text{where } \epsilon_i \sim N(0, \sigma^2), \quad \text{or equivalently } \; V_i \sim N(a + H_0 D_i,\ \sigma^2).$$

Now the model contains three unknown parameters, $\theta = (a, H_0, \sigma^2)$. The resulting log-likelihood function is as follows.

$$l(\theta) = \ln(L(\theta)) = \ln(c) - 12\ln(\sigma^2) - \frac{\sum_{i=1}^{24}(v_i - a - H_0 d_i)^2}{2\sigma^2}.$$

We want to test the following two hypotheses:

$$H_0 : a = 0 \;\; \text{vs} \;\; H_1 : a \neq 0.$$

The simplest way to form a likelihood ratio test statistic is to obtain the MLE for the parameter being tested and its (asymptotic) uncertainty. These two values are

Next, we use the asymptotic property of the MLE:

$$\hat{a}_{MLE} \dot{\sim} N\Big(a,\ \hat{\tau}_n^2\Big) \quad \text{or equivalently} \quad \frac{(\hat{a}_{MLE} - a)}{\hat{\tau}_n} \dot{\sim} N(0,\ 1)\,.$$

The test statistic T can be obtained by replacing a with its hypothesized value under H_0, $a_0 = 0$.

We know that this test statistic is approximately distributed as $N(0, 1)$ if H_0 is true. In other words, if H_0 is the case, this test statistic is most likely to become a value close to 0 (near the highest density region). It turns out that the actual value of the test statistic is very close to 0, as expected under H_0. Our conclusion is that the data do not have statistically significant evidence against H_0.

Model Comparison via Information Criteria

Penalized likelihood approaches have dominated model selection since the 1980s due to several limitations of the likelihood ratio test.

1. The likelihood ratio test is only applicable to nested models (i.e., the model under H_0 is a special case of the model under H_1).

   Example:

2. Likelihood functions always prefer more complex models. Specifically, the maximized likelihood value always increases as more and more parameters are added to the model regardless of their usefulness.

   Example:

So, we have a motivation to adopt a more widely applicable criterion (enabling a comparison among non-nested models) that penalizes the model complexity to achieve a balance between overfitting and underfitting of the model to the data.

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 (i.e., the log-likelihood function evaluated at the MLE $\hat{\theta}$) and p is the number of parameters in a model.

Therefore, the penalty term, $2p$ compensates for the inevitable 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 the data size n.

$$\text{BIC} = -2\ell(\hat{\theta}) + \ln(n)p$$

As we collect more data ($n \geq 8$), 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.

Example: Let's compare the two models considered in the Hubble constant estimation problem via AIC and BIC.

Model 1: $V_i = a + H_0 D_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$.
Model 2: $V_i = H_0 D_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$.

The AIC and BIC are computed for each model as follows.

	Model 1	Model 2
AIC	333.65	**331.91**

	Model 1	Model 2
BIC	337.19	**334.27**

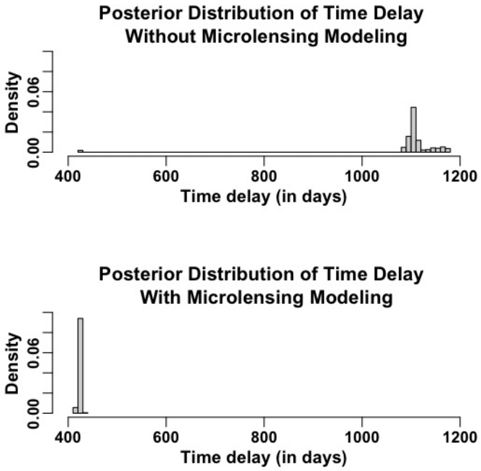
The data prefer Model 2 (no intercept a) because both AIC and BIC are uniformly smaller under Model 2. This empirical evidence in the data is consistent to the Hubble's theoretical reasoning:

$$\text{Velocity} = H_0 \times \text{Distance}.$$

Issues in model mis-specification

Statisticians often say “let the data speak” in estimating parameters of a model. Often neglected is that the data speak only through a given model. That is, the same data may speak something quite different depending on which model the data speak through.

Example: In estimating the time delay between gravitationally lensed light curves of Q0957+561 (Hainline et al., 2012), Tak et al. (2018) adopts a damped random walk model (also known as a continuous-time auto-regressive model of order one or Ornstein-Uhlenbeck process) as a data generating process. This model reveals multiple modes in the likelihood function for the time delay parameter as shown in the top panel. The height of the mode near 400 days is relatively much smaller than the modes near 1100 days. However, it turns out that the highest mode near 1100 days is spurious, caused by model misspecification. The modes near 1100 days disappear when the astronomical model additionally incorporates polynomial regression to account for the effect of microlensing (Tak et al., 2017) that is known to be present in the data (Hainline et al., 2012); see the bottom panel of the figure. Consequently, the mode near 400 days becomes prominent, in agreement with some previous analyses of this quasar (Schild 1990; Shalyapin et al. 2008).



Posterior Distribution of Time Delay
Without Microlensing Modeling

Posterior Distribution of Time Delay
With Microlensing Modeling

This example points out several important aspects in astronomical data analysis. First, different model fits on the same data can reveal completely different possibilities, e.g., for the time delay of Q0957+561. All of these possibilities are worth proper investigation in the context of available scientific knowledge, in an attempt to determine which are simply the result of model misspecification and which are new scientific discoveries. Second, blindly making inference based on the highest mode of the posterior distribution or likelihood function (or smallest loss function in machine learning methods) can be misleading, as illustrated in the top panel of the figure. Thus it is essential to check whether the model captures important characteristics of the data sufficiently well before drawing any conclusions. Lastly, it is the story embedded in the data that can provide insight for improved modeling of physical phenomena, such as microlensing. The better the statistical and astronomical models reflect the data, the better the quality of what the data reveal to us.