Inference I & II

Hyungsuk (Tak) Tak
Pennsylvania State University

Short review of probability

A random variable X (r.v. hereafter) always accompanies a probability distribution that governs the randomness of its realization. That is, an *unrealized* r.v. X is *randomly realized* into a specific value x according to X's probability distribution.

For example, if $X \sim \text{Normal}(0, 1)$, its realized value will be between $(-1, 1)$ with 68%, between $(-2, 2)$ with 95%, and between $(-3, 3)$ with 99%. For example,

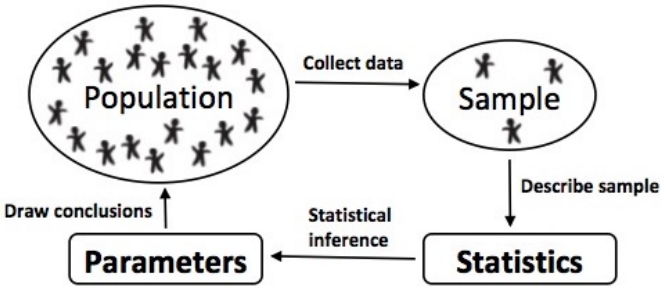
If $X \sim \text{Bernoulli}(0.1)$, only one out of ten realizations is expected to be 1. For example,

A function of r.v.(s) $g(X)$ is also an r.v., and so the function of r.v.(s) $g(X)$ has its own probability distribution that can be different from the original distribution of X.

A set of parameters in a probability distribution completely determines the shape of the distribution, and is typically denoted by a vector θ.

Terminology in Statistical Inference

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



- Population: The entire group of objects about which we make inferences.
- Sample: A fraction of the population on which we actually collect data.

  Let's assume that the sample size is n. Then, we can denote the data (to be collected) by $X_1, X_2, \ldots, X_n$. For example,

Statisticians use these n *random variables* to model the uncertainty of the data realizations by the probability distributions of the n random variables.

In statistical modeling, population characteristics of interest are often modeled by unknown parameter(s) of a probability distribution. For example,

Thus, estimating the unknown parameter(s) is the main theme of statistical inference. For reference, is θ an unknown fixed constant or a value that may vary (r.v.)?

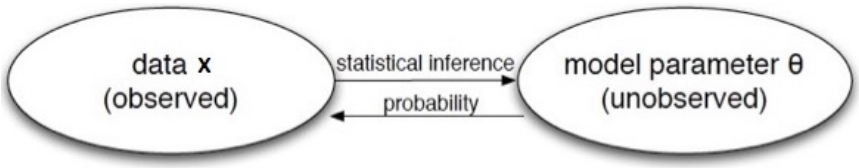
- Statistics: A numerical summary (i.e., a function) of the data, such as mean, median, maximum, etc. This is an r.v. because it is a function of the data (r.v.s).

- Estimator: A statistic (a function of the data) designed to infer a specific parameter. For 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 have learned 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 data generating process.

Example (statistical modeling): We are interested in measuring the brightness of some galaxies. The brightness of galaxy i ($i = 1, 2, \ldots, n$) 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

$$X_i = \mu_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0,\ \sigma_i^2)$$

or equivalently,

$$X_i \sim N(\mu_i,\ \sigma_i^2).$$

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

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

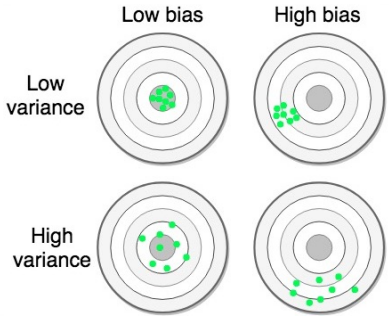
In statistical inference, our main interest is about how to obtain the most likely values of the model parameters, $\theta = (\mu_1, \mu_2, \ldots, \mu_n)$, given the model and observed data of size n, $(X_1 = x_1, \ldots, X_n = x_n)$. In this lecture, we will learn three main topics in statistical inference, i.e., point estimation, interval estimation, and hypothesis testing.

Point Estimation

What is a *good* estimator (a function of the data) for estimating θ? Can we say anything about the closeness of an estimate to an unknown parameter value?

Mean Squared Error

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.

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 at constants), viewed as a function of θ. For example, if $X = \{X_1, X_2, \ldots, X_n\}$ are modeled by continuous r.v.s whose probability distribution is parameterized by θ,

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

If the data are modeled by discrete r.v.s,

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

In this discrete case, the likelihood function is the "probability" that we observe the data $\{X = x\}$ under a model with θ. For example, let's say $L(0.8) \gg L(0.2)$. It means that the probability of observing the current data $P(X = x \mid \theta)$ is much higher when $\theta = 0.8$. So, we can say that the data are supporting $\theta = 0.8$ much more than $\theta = 0.2$. 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: Driggers et al. (2019) and Pratten et al. (2020).

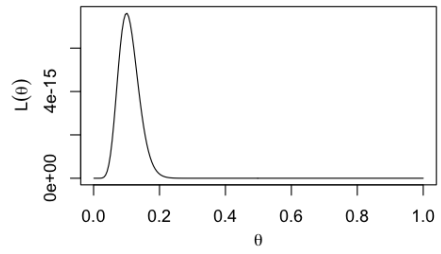
Example: We are interested in the proportion of Earth-like exoplanets orbiting stars within 10 parsecs of the Sun. We have a random sample of 100 exoplanets orbiting stars within 10 parsecs of the Sun. Our statistical model (assumption) is to assume that there is a Bernoulli r.v. (a binary indicator) for each exoplanet taking on 1 if the exoplanet is Earth-like and 0 otherwise, i.e.,

$$X_1, X_2, \ldots, X_{100} \sim \text{Bernoulli}(\theta),$$

where $\theta \in [0, 1]$ is the true population proportion we want to know about. Just for an illustration, let's say we have observed 10 Earth-like exoplanets out of 100.

The probability mass function of the Bernoulli(θ) distribution is $P(X_i = x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$. Then we can derive the likelihood function of the unknown proportion θ given the observed data (10 ones out of 100 binary values) as follows.

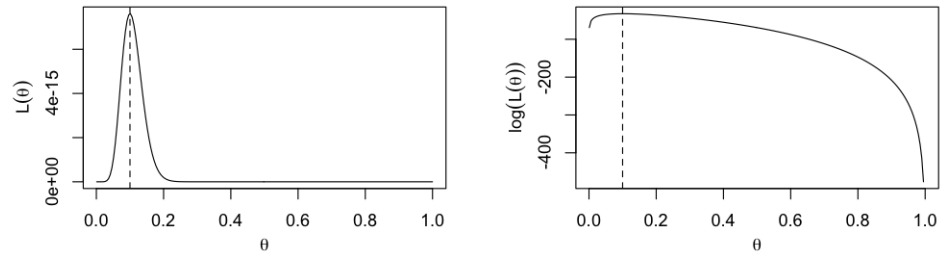
$$\begin{aligned} L(\theta) = P(X = x \mid \theta) = P(X_1 = x_1, \ldots, X_{100} = x_{100} \mid \theta) &\overset{\text{indep.}}{=} \prod_{i=1}^{100} P(X_i = x_i \mid \theta) \\ = \prod_{i=1}^{100} \theta^{x_i}(1-\theta)^{1-x_i} &= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} = \theta^{10}(1-\theta)^{90}. \end{aligned}$$



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

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 product 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 $\ell(\theta)$.

Example: In the previous example, the maximum likelihood estimate was 0.1. The same maximum likelihood estimate can be obtained by a value maximizing the log-likelihood function that is mathematically easier to handle and computationally more stable.



Example: The length of the cosmic ray path is modeled by an Exponential distribution with scale θ (Protheroe et al., 1981). Let X be the length of the path. Given $\theta > 0$, its probability density function is

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

To obtain the MLE of θ given a random sample of n cosmic ray path lengths, $x_1, \ldots, x_n$, we first derive log-likelihood function:

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

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

Next we 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 estimateof θ is $\hat{\theta}_{\text{MLE}} = \bar{x}$.

Example: van den Bergh (1985) considers the luminosity function for globular clusters, and concludes that the luminosity function for clusters in the Milky Way is adequately described

by a Gaussian (Normal) distribution. Its probability density function is

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

where μ is the mean visual absolute magnitude and σ is the standard deviation of visual absolute magnitude.

To find the MLEs for μ & σ^2 given a sample of n globular clusters in the Milky Way, $x_1, \ldots, x_n$, we first derive the log-likelihood function:

$$L(\mu, \sigma^2) \overset{\text{indep.}}{=} \prod_{i=1}^{n} f(x_i \mid \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right),$$

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

Then, we find values (estimates) that maximize $\ell(\mu, \sigma^2)$:

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

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

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

$$\hat{\theta}_{\text{MLE}} \dot{\sim} N\left(\theta,\ \hat{\sigma}^2_n\right).$$

As n becomes large, no other estimator can have the mean squared error smaller than the MSE of $\hat{\theta}_{\text{MLE}}$.

So for many (not all!) problems involving a large number of observations, the MLE becomes the Normally-distributed minimum variance unbiased estimator.

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.

Interval Estimation

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.

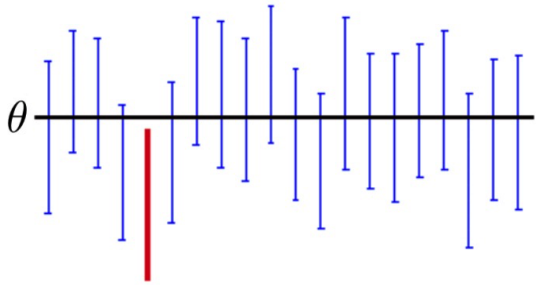
Confidence Interval

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 of an estimator for some parameter θ 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 single fixed and unknown value of θ. That means, it is the interval that is random, not the parameter θ. The randomness comes from which data we observe. This is why 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.

After the data are observed $(X = x)$, we end up with a single realization of this interval $(l(x),\ u(x))$, which is also called a 95% confidence interval of θ. Note that once the interval is crystalized, it either covers θ or does not. Thus the most common mistake is to interpret this interval as the probability of θ being in this specific interval $(l(x),\ u(x))$ with 95% chance.

Instead, its interpretation must be made under a hypothetical scenario; if the experiment of interest (or random sampling of the data) were repeated 100 times under the same condition, computing 100 confidence intervals from the resulting 100 data sets (in the same way), then 95 confidence intervals out of 100 are expected to contain the unknown true parameter θ. That is, we would expect 95% of the intervals obtained from the repeated experiments to cover the true parameter we are estimating.



In reality, just one out of these confidence intervals is what we actually get from the data.

Example: In the example of the luminosity function for globular clusters in the Milky Way (vdB, 1985), σ is completely known as $\sigma = 1.2$ mag. Given a random sample of globular clusters, $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, what is a 95% confidence interval for μ?

From the previous example, the maximum likelihood estimator was $\bar{X}$. One of the important results in probability is that

when $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$. Once we standardize this sample mean, we can get the

standard Normal r.v. Z:

From the property of the standard Normal distribution, we can say that

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

Next, we derive lower and upper bounds for the parameter of interest, re-arranging the inequalities inside the above probability as follows.

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

The resulting interval is a 95% confidence interval for μ when σ is a known constant.

Example: In the vdB example, the observed data are magnitudes of 148 ($= n$) globular clusters, and the average brightness is $\bar{x} = -7.1$ mag. Assuming that $\sigma = 1.2$ mag, find a 95% confidence interval for the population average brightness μ.

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

Hypothetically, if we were able to repeat the data collection process, computing a 95% confidence interval each time, then 95% of these intervals would contain μ. The observed interval $(-7.29, \ -6.91)$ is just one of these possible intervals.

Confidence interval for μ when $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ with σ unknown

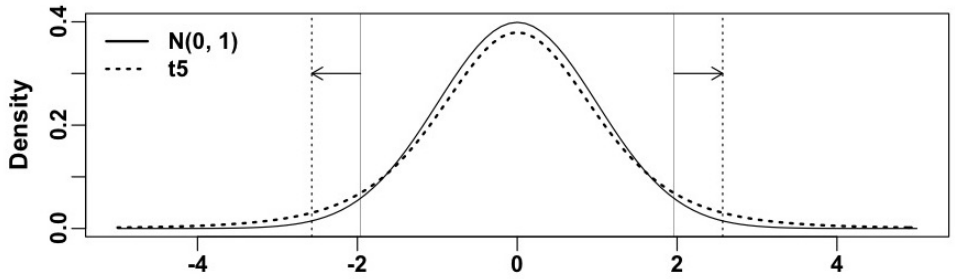
Previously, we use the standardized sample mean to derive a 95% confidence interval:

What if we do not know the value of σ above? *A principle in statistics is to replace any unknown quantity with its good estimator.*

When a random sample of size n (observed data) comes from a Normally distributed population distribution with mean μ and variance σ^2 like the vdB example, a good estimator for σ^2 is the sample variance:

If we replace σ with S, it is known that

The t_ν-distribution is heavy-tailed: For example, if $T \sim t_5$ with a sample size $n = 6$,



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

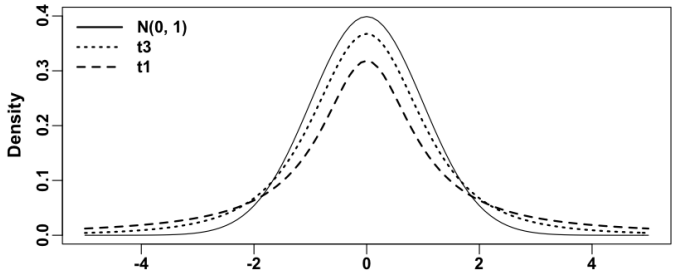
In this case, a 95% confidence interval for μ with unknown σ^2 is

because

$$P(-2.57 < T < 2.57) = P\left(-2.57 < \frac{\bar{X} - \mu}{S/\sqrt{6}} < 2.57\right) = P\left(\bar{X} - 2.57\frac{S}{\sqrt{6}} < \mu < \bar{X} + 2.57\frac{S}{\sqrt{6}}\right) = 0.95.$$

For the same confidence level, the t_ν-distribution results in an interval wider than a Normal-based interval to account for the additional uncertainty of not knowing σ^2 (i.e., extra uncertainty of using an estimator S^2 for the unknown σ^2 instead of using the true value of σ^2).

Note that the t_{n-1} distribution approaches the standard Normal distribution as the sample size n increases.



Thus, if n is large (typically $n > 30$), $t_{n-1} \dot{\sim} N(0, 1)$, and a 95% confidence interval for μ 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 μ

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

1. If σ^2 is known, a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{X} - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

   where $q_{1-\alpha/2}$ is a constant satisfying $P(Z < q_{1-\alpha/2}) = 1 - \alpha/2$ when $Z \sim N(0,1)$.

2. If σ^2 is unknown, a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{X} - q_{1-\alpha/2}\frac{S}{\sqrt{n}}, \quad \bar{X} + q_{1-\alpha/2}\frac{S}{\sqrt{n}}\right),$$

   where $q_{1-\alpha/2}$ is a constant such that $P(T < q_{1-\alpha/2}) = 1 - \alpha/2$ when $T \sim t_{n-1}$.

3. If σ^2 is unknown but n is large ($n > 30$), a $100(1-\alpha)\%$ confidence interval for μ is

$$\left(\bar{X} - q_{1-\alpha/2}\frac{S}{\sqrt{n}}, \quad \bar{X} + q_{1-\alpha/2}\frac{S}{\sqrt{n}}\right),$$

   where $q_{1-\alpha/2}$ is a constant satisfying $P(Z < q_{1-\alpha/2}) = 1 - \alpha/2$ when $Z \sim N(0,1)$.

Approximation Method based on MLE

Sometimes, the distribution of the data is not Gaussian in astronomy. When the data are not Normally distributed, e.g., Poisson, Bernoulli, etc., the distribution of the MLE may not be Gaussian. For example, when $X_i \sim \text{Poisson}(\lambda)$, the MLE for λ is $\bar{X}$, but the distribution of $\bar{X}$ is not Gaussian. How can we derive a confidence interval with the non-Gaussian data?

Even in such a non-Gaussian case, we can still construct an approximate confidence interval using the asymptotic optimality of an MLE. That is, we know that as $n \to \infty$,

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

So, we can construct an approximate 95% confidence interval from this result:

$$\left(\hat{\theta}_{MLE} - 1.96\hat{\sigma}_n, \quad \hat{\theta}_{MLE} + 1.96\hat{\sigma}_n\right).$$

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: Abbott et al. (2016a).

Motivation and Terminology

The average brightness of the M31 globular clusters, denoted by μ, is known to be -7.7 magnitude. Tak feels that the brightness may not be -7.7 magnitude, and so he collects a random sample from M31 to check whether the data are consistent to his feeling. The question is

"Are the data strongly in support of Tak's claim that $\mu \neq -7.7$ magnitude?"

Statistical hypothesis (or hypothesis) is a statement about parameter(s) θ of a population. Under a statistical model, what we want to argue can be parameterized by θ.

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

Simple vs composite: Simple hypotheses are hypotheses where θ can only take on a single value, e.g., $H_0 : \theta = 0$ or $H_a : \theta = 3$. Composite hypotheses are hypotheses where θ can take on multiple or a range of values, e.g., $H_0 : \theta \leq 0$ or $H_a : \theta \in \{1, 2, 3\}$. In a particular problem, null and alternative hypotheses can have any combination of simple and composite hypotheses.

One-sided vs two-sided: If an alternative hypothesis is composite and represent one-side of the parameter space around some value w, we call this a one-sided test. For example,

$$\begin{aligned} H_0 : \theta \leq w \quad &\text{vs} \quad H_a : \theta > w \\ H_0 : \theta = w \quad &\text{vs} \quad H_a : \theta > w \\ H_0 : \theta \geq w \quad &\text{vs} \quad H_a : \theta < w \\ H_0 : \theta = w \quad &\text{vs} \quad H_a : \theta < w \end{aligned}$$

If H_0 is a simple hypothesis and H_a represents the rest of the parameter space of θ, we call this a two-sided test. For instance,

$$H_0 : \mu = -7.7 \quad \text{vs} \quad H_a : \mu \neq -7.7$$

Test statistic and rejection region: Let $X_1, \ldots, X_n$ be the data (a random sample) obtained from a population distribution with parameter θ. Let $T = h(X_1, \ldots, X_n)$ be a statistic (a function of the data) and let R be a subset (a particular range) 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.

Example: A search detection statistic in Abbott et al. (2016a) is the test statistic T.

Significance level α: 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, i.e., $P(T \in R \mid H_0 \text{ is the case})$, to be smaller than or equal to 0.05.

Likelihood Ratio Test

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

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

(Each of θ_0 and θ_1 is a value in the parameter space of θ). We want to make a choice between these two possibilities based on the data we observe. Let us consider this as a binary decision problem—reject or do not reject H_0.

A test (or equivalently a decision rule) for choosing one of the two hypotheses based on the data can be specified by the rejection region R:

$$\begin{aligned} &\text{If the data is in the rejection region, } X \in R, \text{ then we reject } H_0 \\ &\text{If the data is not in the rejection region, } X \notin R, \text{ then we do not reject } H_0 \end{aligned}$$

So constructing a test boils down to choosing an appropriate rejection region R.

The question is "How do we choose the rejection region R?" It turns out that there is a general procedure with which one can construct a good rejection region R. The idea is related to the maximum likelihood principle we have already learned. Let's start with the test of simple vs simple hypothesis above.

Comparing likelihoods: What are the likelihood values under the two hypotheses? Based on these likelihood values, it makes sense to choose a hypothesis with the higher likelihood (i.e., a hypothesis that the data support more). That means, it makes sense to reject H_0 if

The "best" (and optimal) hypothesis testing is very much along this line!

Likelihood ratio test statistic $L(\theta_1)/L(\theta_0)$ measures the relative evidence from the data under the two hypotheses. A rejection region based on this statistic is

A test with a rejection region of this form is called a *likelihood ratio test*. Although the rejection region above appears mysterious at first glance, it is always simplified to an inequality with respect to a test statistic (a function of an MLE in the end). For example, $R = \{x : T(x) > c'\}$, $R = \{x : T(x) < c'\}$, etc.

It turns out that likelihood ratio tests are the "best" (most powerful) tests under certain criteria, 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, given a specific significance level α.

Example: Another Abbott et al. (2016b).

Likelihood Ratio Test based on MLE

In the likelihood ratio test, the MLE for the parameter being tested or a function of the MLE (typically a standardized form under H_0) results in a test statistic T. For example, $\bar{X}$ is used for testing some value of μ in a Normal case, $\hat{p}$ $(= \bar{X})$ for testing some value of p in a Bernoulli or Binomial case, and $\hat{\lambda}$ $(= \bar{X})$ for testing some value of λ in a Poisson case. A rejection region is usually in the form of $T \geq c$, $T \leq c$, or $|T| \geq c$. The direction of the inequality is set by the direction of H_a.

Example: Find the rejection region (in particular, direction) when the null hypothesis is $H_0 : \theta = \theta_0$, and the alternative is $H_a : \theta > \theta_0, H_a : \theta < \theta_0$ or $H_a : \theta \neq \theta_0$

Example (t-test): One of the most popular likelihood ratio tests is a t-test. In the example of luminosity function for clusters in the Milky Way (vdB), a random sample of 148 measurements has sample mean $\bar{x} = -7.5$ and sample variance $s^2 = 1.1^2$. From the data, Tak feels that μ $(= M_0) \neq -7.7$ mag. We want to test it at the significant level $\alpha = 0.05$.

1. Specify the null and alternative hypotheses:

2. Set up a likelihood ratio test statistic for μ under H_0:

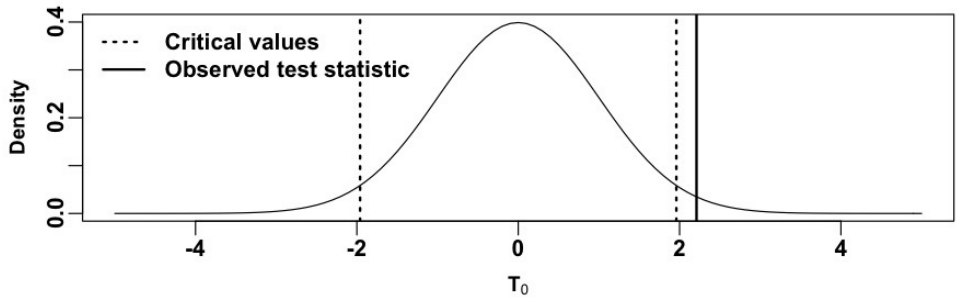
3. Find the distribution of the test statistic under the null:

4. Set up the rejection rule: Since H_a is two-sided, the likelihood ratio test has a rejection region in the form of $|T_0| > c$, where c is determined to make $P(|T_0| > c \mid H_0) = \alpha = 0.05$.

   Thus, the likelihood ratio test is to reject H_0 if $|T_0| > 1.96$. We consider a value of the test statistic T_0 falling in this rejection region as strong evidence that the null hypothesis is not the case because such a value of T_0 is very unlikely to happen (<5%) under the assumption that H_0 is true.

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

$$\frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{-7.5 + 7.7}{1.1/\sqrt{148}} = 2.21.$$



6. We reject the null hypothesis H_0 because the calculated value of the test statistic lies in the rejection region. We report that there is a statistically significant difference between the population mean μ and the hypothesized value $\mu_0 = -7.7$ mag.

Example: Going back to Figure 4 of Abbott et al. (2016a).

There are two other ways to conduct the same hypothesis testing.

Hypothesis testing via confidence interval: A test of significance level α for a parameter θ can be done by deriving a $100(1-\alpha)\%$ confidence interval for θ. We reject the null if the interval does not contain the hypothesized value of θ under the null (i.e., θ_0) because it is a rare event ($100\alpha\%$) if the null is true.

In the previous example, we have derived a 95% confidence interval for μ and we can use this to conduct the same hypothesis testing. Since n is large enough, a 95% confidence interval is

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

This 95% confidence interval for μ _____ contain the hypothesized value under the null hypothesis, i.e., $\mu_0 = -7.7$, and thus at the 5% significance level we _____ H_0.

Hypothesis testing via p-value: 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,

$$p\text{-value} = P\left(T_0 > \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \text{ or } T_0 < -\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \;\middle|\; H_0\right) = 2P\left(T_0 > \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \;\middle|\; H_0\right)$$

For instance, the p-value in the vdB example is $2P(T_0 > 2.21 \mid H_0) = 0.03$ that is _____ than the significance level 0.05. Thus, at the 5% significant level, there is enough evidence in data to _____ H_0.

Note that the p-value is not the probability of H_0 being correct (the most common mistake!).

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 one under H_a).
2. Let's say a model M_0 under H_0 is nested within another model M_a under H_a. The largest likelihood achievable by M_a will always be larger than that achievable by M_0 even when M_0 is the true model. This is because M_a has more parameters, enabling M_a to explain the data more elaborately.

If a penalty is applied to compensate for the difference in likelihoods due to the different number of parameters in M_0 and M_a, the desired balance between overfitting and underfitting can be found.

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 in a model.

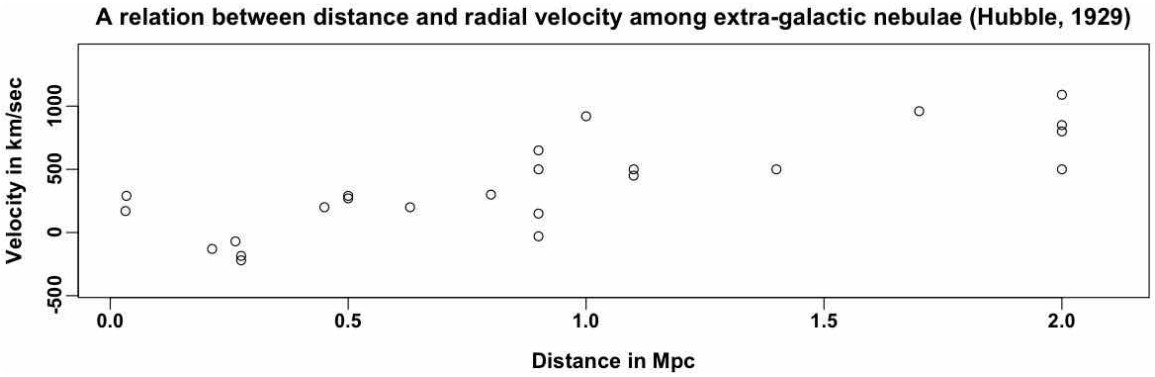
The penalty term, $2p$, increases as the complexity of the model grows, and thus compensates for the necessary increase in the likelihood. A model with the 'smallest' AIC, i.e., a model that explains the data well with a small number of parameters, is preferred.

Bayesian information criterion weights the penalty according to the data size n.

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

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.

Example: Let's compare the following four models on the original 24 data of Hubble (1929).



Let y denote the velocity and x denote the distance.

Model 1: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$.
Model 2: $y_i = \beta_0 + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$.
Model 3: $y_i = \beta_1 x_i + \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$.
Model 4: $y_i = \epsilon_i, \quad \epsilon \sim N(0, \sigma^2)$.

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

	Model 1	Model 2	Model 3	Model 4
AIC	333.65	355.10	**331.91**	370.37
BIC	337.19	357.45	**334.27**	371.55

The data prefer Model 3 (no intercept β_0) in a sense that both AIC and BIC are the smallest under Model 3. This is consistent to the Hubble's reasoning:

$$\text{Velocity} = \beta_1 \times \text{Distance} = H_0 \times \text{Distance}.$$