Mathematics/Statistics Bootcamp

Part IV: Basics of Statistical Inference

Brian Cozzi[1] Michael Valancius[1] Graham Tierney[1]
Becky Tang[1]

Department of Statistical Science
Duke University

Graduate Orientation, August 2019

Overview

Limiting Theorems

Data Reduction
- Sufficiency
- Likelihood

Point Estimation
- Bayesian Estimation
- Evaluating Estimators

Hypothesis Testing

Confidence Intervals

Limiting Theorems

Probability Inequalities

Theorem

Markov's Inequality: *Let X be a non-negative random variable and suppose that $E[X]$ exists. Then for any $t > 0$,*

$$Pr(X > t) \leq \frac{E[X]}{t}$$

Theorem

Chebyshev's Inequality: *Let $\mu = E[X]$ and $\sigma^2 = Var(X)$. Then,*

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$$

The Law of Large Numbers (LLN)

Suppose $\{X_1, X_2, \ldots\}$ is a sequence of independently and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$. Let $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ be the sample average. Then:

- The **Weak Law**: $\bar{X}_n \xrightarrow{p} \mu$ when $n \to \infty$, that is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$

- This is a straightforward application of Chebychev's Inequality.
- The **Strong Law**: $\bar{X}_n \xrightarrow{a.s.} \mu$ when $n \to \infty$, that is,

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

LLN Application: Monte Carlo Methods

In Monte Carlo simulations, the LLN is invoked to calculate expectations of functions. The applications are diverse, including calculating expectations, probabilities and integrals.

Let X have pdf $f_X(x)$, and let $h_n(X) = \sum_{i=1}^{2} h(X)$. By definition of expectation, $E[h(X)] = \int h(x) f_X(x) dx$. From the WLLN, if $E[h(X)]$ exists, then $\lim_{n\to\infty} P(|h_n(X) - E[h(X)]| > \epsilon) = 0$.

The idea: in Monte Carlo sampling, n samples are drawn from $f_X(x)$ to give $h_n(X)$, allowing for the approximation of $E[h(X)]$.

Monte Carlo Examples: Probability

If $X \sim N(\mu, \sigma^2)$, then:

$$\begin{aligned} Pr(X < c) &= \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} I_{(X<c)} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right) dx \\ &\simeq \frac{1}{n}\sum_{i=1}^{n} I(x_i < c) \end{aligned}$$

More generally, $P(X \in A) = E[I_{(A)}(X)] \simeq \dfrac{1}{n} \sum_{i=1}^{n} I(x_i \in A)$

Monte Carlo Examples: Integration

Another not so obvious application of Monte Carlo sampling is the ability to calculate integrals. Consider the integral $\int_0^2 x^2 dx$. From calculus, the solution is known to be $8/3$. An approximation to this can be found by multiplying the integral by 0.5 / 0.5 = 1 and identifying this integral as $0.5 * E[x^2]$ where X is uniformly distributed on the interval of [0,2].

$$\int_0^2 x^2 dx = 2\int_0^2 \frac{1}{2} x^2 dx = 2 * E[X^2] \simeq \frac{2}{n}\sum_{i=2}^{n} x_i^2$$

```
> n <- 10000
>
> x <- runif(n, 0, 2)
>
> 2*sum(x^2)/n
[1] 2.661915
>
```

The Central Limit Theorem (CLT)

Suppose $\{X_1, X_2, \ldots\}$ is a sequence of independently and identically distributed (i.i.d.) random variables with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ be the sample average, then as $n \to \infty$, the random variable $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to $N(0, \sigma^2)$:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

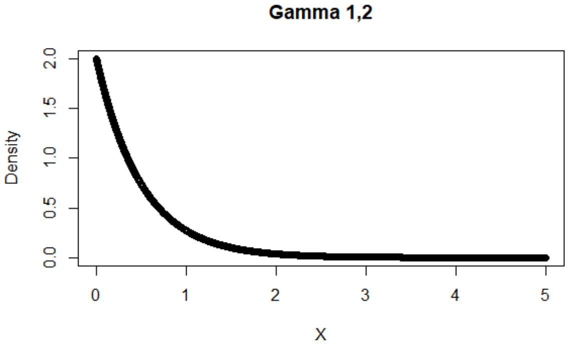
Notes on CLT

1. The central limit theorem applies regardless of the underlying distribution of the data, so long as the variance of the data is finite and the samples are i.i.d.
2. This is a statement about the sample average, not individual data points.
3. That the distribution of the sample mean is normal is an asymptotic result ($n \to \infty$).

Simulated Example

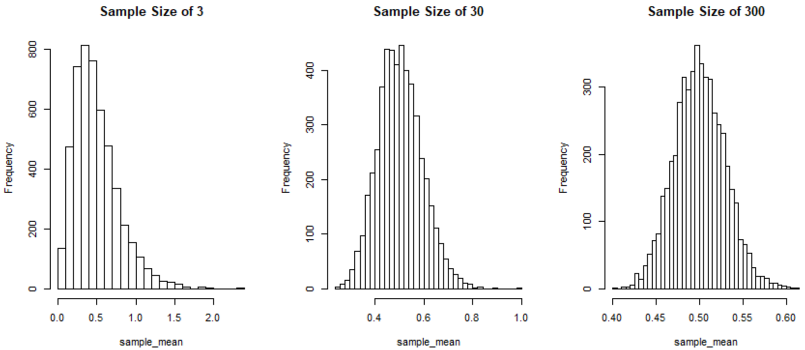
```r
### Generating a sequence of points to evaluate density at
xs <- seq(from = 0, to = 5, length.out = 5000)

### Plotting the density
plot(y = dgamma(xs, shape = 1, rate = 2), x = xs,
     xlab = "X", ylab = "Density", main = "Gamma 1,2")
```



Simulated Example Cont.

```r
samp_size <- c(3, 30, 300)
iterations <- 5000
for(x in samp_size){
  sample_mean <- vector(length = iterations)
  for(i in 1:iterations){
    sample_mean[i] <- mean(rgamma(x, shape = 1, rate = 2))
  }
  hist(sample_mean, breaks = 30, main = paste0("Sample Size of ", as.character(x)))
}
```



Data Reduction

Sufficient Statistics

- A statistic $T = T(\boldsymbol{X})$ that is a function of the data $\boldsymbol{X}$ is said to be **sufficient** for θ if the conditional distribution of $\boldsymbol{X}|T = t$ does not depend on θ.
- Intuition behind the terminology: T captures all the information the data $\boldsymbol{X}$ tell us about θ. So after assuming the distribution of $\boldsymbol{X}$ and being given the value $T = t$, there is nothing more to learn about θ.
- Sufficient statistics are not unique!
  - For example, all the data $\boldsymbol{X}$ are sufficient for θ, but depending on the model we may find simpler $T(\boldsymbol{X})$, such as $\sum X_i$

Sufficient Statistic Example

- Let two random variables X_1, X_2 be i.i.d. Poisson(λ). So $P_\lambda(X_i = j) = e^{-\lambda}\lambda^j/j!$ for $j = 0, 1, 2, \ldots$
- Let $T = X_1 + X_2$. Claim: T is a sufficient statistic.
- Consider the conditional distribution:

$$
\begin{aligned}
P(X_1 = x, X_2 = t - x | X_1 + X_2 = t) &= \frac{P(X_1 = x, X_2 = t - x)}{P(X_1 + X_2 = t)} \\
&= \frac{e^{-\lambda}\lambda^x/x! \times e^{-\lambda}\lambda^{t-x}/(t-x)}{e^{-2\lambda}(2\lambda)^t/t!} \\
&= \left(\frac{1}{2}\right)^t \times \frac{t!}{x!(t-x)!} \\
&= \left(\frac{1}{2}\right)^t \binom{t}{x} \\
&= \text{Binomial}(t, 1/2)
\end{aligned}
$$

Sufficient Statistic Example Cont.

- Because the conditional distribution (Binomial($t, 1/2$)) is independent of the unknown parameter λ, by definition, T is sufficient for λ

Likelihood Function

- If $X_1, \dots, X_n$ are an i.i.d. sample from a population with pdf or pmf $f(\mathbf{x}|\theta_1, \dots, \theta_k)$, the **likelihood function** is

$$L(\theta|\mathbf{x}) = L(\theta_1, \dots, \theta_k|x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i|\theta_1, \dots, \theta_k).$$

- Density function vs Likelihood function. The density function $f(\mathbf{x}|\theta_1, \dots, \theta_k)$ is a non-negative function that integrates to 1. The likelihood function is a function of the parameter(s) θ and typically will not integrate to 1.
- For computational purposes, typically we worked with the log of the likelihood function.

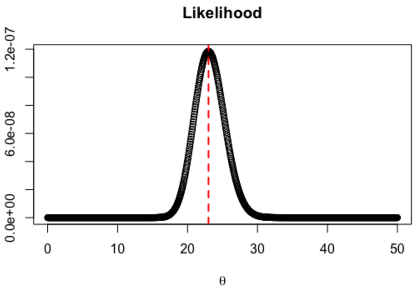
Likelihood Example

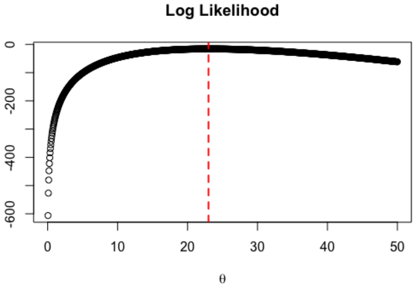
Suppose we observe (iid) values $\mathbf{x} = (15, 30, 21, 29, 20)$ and are modeling the distribution as Poisson.

$$L(\theta|\mathbf{x}) = e^{-5\theta} \frac{\theta^{115}}{15!\ 30!\ 21!\ 29!\ 20!}$$

$$Log\ \ L(\theta|\mathbf{x}) = -5\theta - \sum ln(x_i!) + ln(\theta) * 115$$

Likelihood Example





Point Estimation

Point Estimation

- A **point estimator** is any function of the sample.
- **Estimator** vs. **Estimate**: The former is a function, while the latter is the realized value of the function (a number) that is obtained when a sample is actually taken.
- Examples include the arithmetic mean ($\bar{\mathbf{X}}$ and $\bar{\mathbf{x}}$) and linear regression coefficients (β and $\hat{\beta}$).

Maximum Likelihood Estimators

- For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.
- If the likelihood function is differentiable (in θ_i), **possible candidates** for the MLE are the values of $(\theta_1, \ldots, \theta_k)$ that satisfy
$$\frac{\partial}{\partial \theta_i} L(\theta|\mathbf{x}) = 0, \quad i = 1, \ldots, k.$$
- Since $log(\theta)$ is a monotonically increasing function of θ, for any positive valued function f, $arg\ max_\theta f(x) = argmax_\theta$ $log\ f(x)$. That is, maximizing the log likelihood results in the same MLE estimates as maximizing the likelihood.

MLE: Normal Example

Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, 1)$, and let $L(\theta|\mathbf{x})$ denote the likelihood function. Since maximizing

$$L(\theta|\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{(-1/2)\sum_{i=1}^{n}(x_i-\theta)^2},$$

is equivalent to maximizing $\log L(\theta|\mathbf{x})$, we reduce the problem to maximizing

$$h(\theta) = \log((2\pi)^{-n/2}) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2,$$

a quadratic function of θ.

Since $\hat{\theta} = \bar{x} = (\sum_{i=1}^{n} x_i)/n$ is the global maximum point of $h(\theta)$, it is also the global maximum point of $L(\theta|\mathbf{x})$. Therefore $\hat{\theta}$ is the MLE.

The Invariance Property of MLEs

Theorem

If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$ of θ, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

Further properties of MLEs

- Under certain conditions, the MLE is CAN (**C**onsistent and **A**symptotically **N**ormal). Consistency means that it converges in probability to the true value. The asymptotic normality means that as $n \to \infty$, $\hat{\theta}_{MLE} \sim Normal\left(\theta, \frac{1}{nI(\theta)}\right)$.
- $I(\theta)$ is the Fisher's Information and is defined to be $-\mathrm{E}\left[\frac{\partial^2 logf_\theta(X)}{\partial\theta^2}\right]$. If θ is a scalar, $I(\theta)$ is a scalar, and if θ is a vector, then $I(\theta)$ is a matrix.

Bayesian Estimates: More than Point Estimation

- In maximum likelihood estimation, a random sample $X_1 ... X_n$ is drawn from a population with a probability distribution that is indexed by an unknown, fixed θ.
- The maximum likelihood estimate is a best guess at the true value of θ based on the sample. It is a point estimate because the value obtained is a single point.
- In Bayesian approaches, uncertainty about θ is itself described by a probability distribution. This distribution, $p(\theta)$, is called the prior distribution.
- After a sample is obtained from the distribution $p(y|\theta)$, we **update** our beliefs about θ through Bayes rule. We now have an updated probability distribution: $p(\theta|x_1 ... x_n)$, the posterior distribution.

Comparison via Simple Example

Consider an example involving coin tosses. Let $X_1 ... X_{10}$ be 10 coin tosses where $X_i = 1$ if the coin lands heads up and zero if it lands tails up.

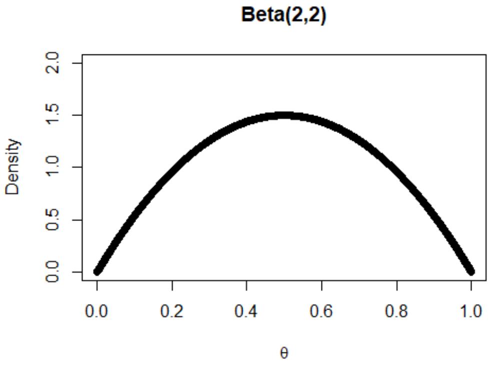
In both cases, we are assuming that the data are independent and identically distributed, with $p(X|\theta) \sim \text{Bernoulli}(\theta)$. We are interested in estimating θ.

In maximum likelihood estimation, the maximum of the likelihood of the data produces our point estimate for the probability of heads (θ). The MLE estimate is $\bar{X}$.

In Bayesian estimation, a prior distribution for θ is chosen. This prior distribution is a probability distribution that represents our belief in what value θ might be **before** observing the data. We might think that, given no other information, since we are dealing with a coin, there is a better chance that the probability a given flip is a heads is somewhere around 0.5 as opposed to 0.1 or 0.9.

Example Continued

A prior distribution should reflect this belief and be consistent with the structure of the problem (that is, $p(\theta) > 0$ only in $[0, 1]$). The calculations are simplified if the prior is conjugate, as discussed before. For this example, a beta(2,2) satisfies these beliefs.



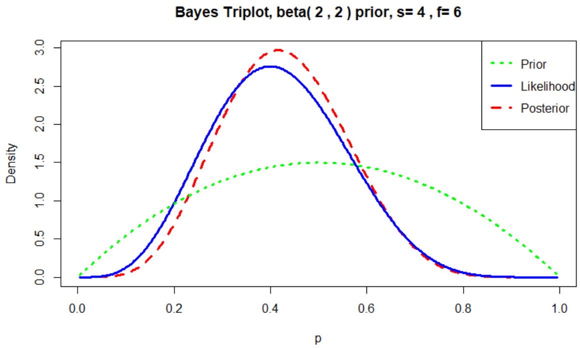
Example Continued

Tossing the coin 10 times gives the sequence H,H,T,T,T,T,T.H,H,T. The MLE estimate is $\hat{\theta} = 0.40$.

In the Bayesian context, the posterior distribution is derived via Bayes rule:

$$
\begin{aligned}
p(\theta | x_1 \ldots x_{10}) &= \frac{p(x_1 \ldots x_{10} | \theta) p(\theta)}{p(x_1 \ldots x_{10})} \\
&= \frac{\left(\prod_{i=1}^{10} (\theta)^{x_i} (1-\theta)^{1-x_i} \right) \left(\theta(1-\theta) \frac{1}{B(2,2)} \right)}{p(x_1 \ldots x_{10})} \\
&\propto c \ \theta^5 (1-\theta)^7 \\
&= Beta(6,8)
\end{aligned}
$$

Example Continued



Asymptotic Distribution of MLE

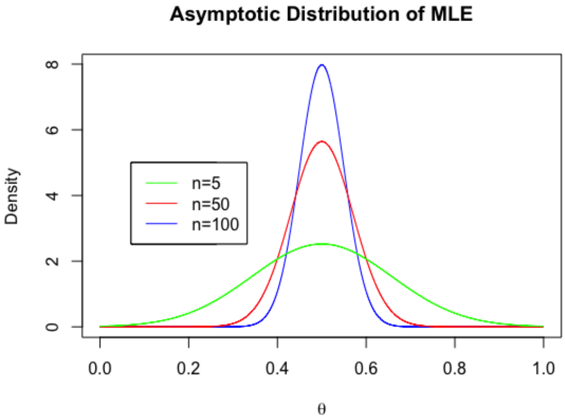
As noted earlier, the MLE is **CAN**: **C**onsistent and **A**symptoticaly **N**ormal. Let us suppose that these data were generated from a fair coin, i.e. one that has $\theta = 0.5$.

Then from statistical theory, $\hat{\theta}_{MLE} \sim N\left(\theta, \frac{1}{n\mathbf{I}(\theta)}\right)$

$$\begin{aligned}
I(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2} logf(x;\theta)|\theta\right] \\
&= -E\left[\frac{\partial^2}{\partial\theta^2} X\ log(\theta) + (1-X)log(\theta)|\theta\right] \\
&= E\left[\frac{-X}{\theta^2} - \frac{1-X}{(1-\theta)^2}|\theta\right] \\
&= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\
&= \frac{1}{\theta(1-\theta)}
\end{aligned}$$

MLE Distribution Continued

Thus, $\hat{\theta}_{MLE} \sim N\left(0.5, \frac{0.25}{n}\right)$



Summary of Differences

- In the example, we decided to model the data as Bernoulli.
- In Maximum Likelihood Estimation, we find the value $\hat{\theta}$ that maximizes the likelihood of the data. This single point has the property that it is consistent (will be defined shortly) and asymptotically normal.
- In Bayesian estimation, the likelihood of the data is combined with our prior knowledge of θ to produce a posterior distribution for θ. This probability distribution can be used to describe many features of $\theta | x_1 ... x_n$.

Monte Carlo Sampling: Posterior Distribution

- When the posterior distribution can be described by a common probability distribution (as is frequently the case with choosing a conjugate prior), Monte Carlo sampling provides a powerful framework for generating countless inferential summaries.
- Recall that in Monte Carlo sampling of a posterior distribution, S samples are drawn from the posterior distribution to approximate $E[g(\theta)|y_1 ... y_n]$. As $S \to \infty$, this approximation becomes more accurate.
- Examples:
  - $\frac{1}{S}\sum_{s=1}^{S} \theta^{(s)} \to E[\theta|y_1 ... y_n]$
  - $\sum_{s=1}^{S} I_{\theta^{(s)} \in A} \to Pr(\theta \in A|y_1 ... y_n)$
  - The median of $\theta^{(1)} ... \theta^{(S)} \to \theta_{0.5}$

Evaluating Estimators

The general question: given an estimator W of some parameter θ, how do we somehow assess it's quality? Ideally, the estimator exhibits two fundamental traits: low bias and low variance.

Bias

The bias of an estimator W of θ is defined to be $E_\theta[W] - \theta$. An unbiased estimator is one for which the bias is zero.

Example: The MLE estimates for μ, σ^2 of a normal distribution are $\hat{\mu} = \bar{Y}$ and $\hat{\sigma^2} = \frac{1}{n}\sum(Y_i - \hat{\mu})^2$.

$$Bias(\hat{\mu}) = E[\hat{\mu}] - \mu = \frac{1}{n}\sum E[Y_i] - \mu = \frac{n\mu}{n} - \mu = 0$$

$$Bias(\hat{\sigma^2}) = E[\hat{\sigma^2}] - \sigma^2 \neq 0$$

Observation: MLE estimates are not guaranteed to be unbiased.

Consistency

A sequence of estimators $W_n = W_n(X_1 ... X_n)$ is a consistent sequence of estimators for θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$, $\lim_{n \to \infty} P_\theta(|W_n - \theta| \geq \epsilon) = 0$

- This is an asymptotic result: interest is in the behavior of a sequence of estimators.
- The general idea is that a consistent estimator gets closer to the parameter it is estimating as the amount of observations grow. Note, this does not say that, in general, W_i is closer to θ than W_j is if $i \geq j$ since that depends on the exact random sample.
- Earlier we noted that the MLE of σ^2 for normally distributed random variables is not unbiased. However, it is consistent (the difference between (n-1) and n becomes negligible as $n \to \infty$).

Mean Squared Error

Definition: The Mean Squared Error (MSE) of an estimator W of parameter θ is $E_\theta[(W-\theta)^2]$

- Simple algebraic manipulation provides the alternative definition: $E_\theta[(W-\theta)^2] = Var_\theta(W) + Bias(W)^2$
- Thus, MSE captures both the precision of the estimator (how much can we expect it to vary with different samples?) as well as the accuracy (is it biased?).
- An estimator that is biased (many Bayesian estimators) but more precise might be preferable to one that is unbiased but fluctuates wildly.

MSE Example

Recall that if $Y \sim Binom(\theta, n)$, then $\theta | y \sim Beta(y + a, n + b - y)$.

A reasonable estimator under of θ might be the posterior mean $E[\theta|y]$: $\hat{\theta}_B = \dfrac{a + y}{a + b + n}$.

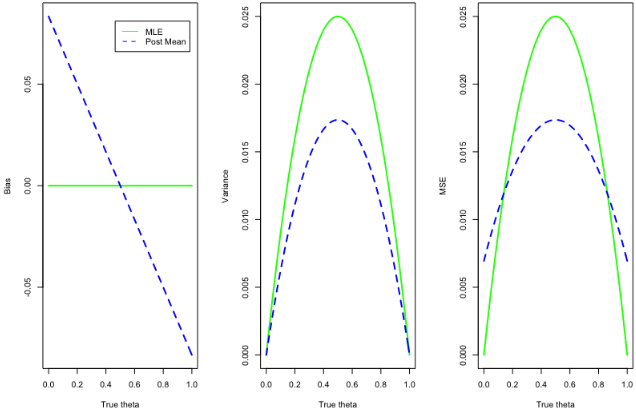
Another reasonable estimator of θ might be the MLE: $\hat{\theta}_{MLE} = \dfrac{y}{n}$.

Which estimator is "better"? We can compute their MSEs:

$$
\begin{aligned}
MSE(\hat{\theta}_B) &= [Bias \hat{\theta}_B]^2 + Var(\hat{\theta}_B) \\
&= \left(\frac{a - (a + b)\theta}{n + a + b} \right)^2 + \left(\frac{n}{n + a + b} \right)^2 \frac{\theta(1 - \theta)}{n} \\
MSE(\hat{\theta}_{MLE}) &= [Bias \hat{\theta}_{MLE}]^2 + Var(\hat{\theta}_{MLE}) \\
&= 0^2 + Var\left(\frac{y}{n} \right) \\
&= \frac{\theta(1 - \theta)}{n}
\end{aligned}
$$

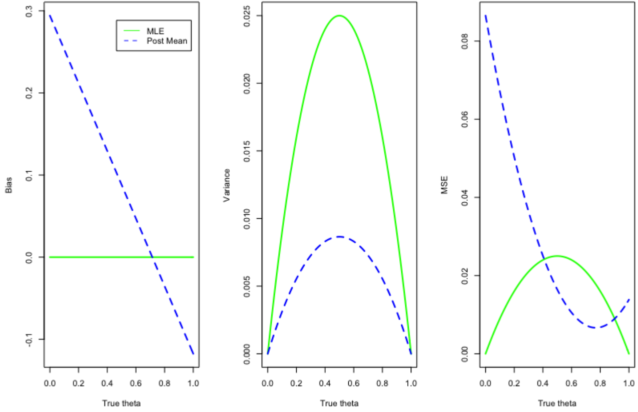
MSE Example Continued

In terms of MSE, does one estimator always outperform the other? Not necessarily! With these two estimators, let us first consider a Beta(1,1) prior on θ and $n = 10$.



MSE Example Continued

Here, with a Beta(5,2) prior on θ and still $n = 10$.



Hypothesis Testing

Hypothesis Testing

- In hypothesis testing, a default theory, the null hypothesis, is proposed, and we see if the data provides sufficient evidence to reject this hypothesis.
- If we do not reject the null hypothesis, we are said to retain the null hypothesis (sometimes referred to as accepting or failing to reject the null).

Hypothesis Testing: Likelihood Ratio Tests

Let $\theta = (\theta_1 ... \theta_q, \theta_{q+1}, ... \theta_r)$, while Θ_0 consists of all possible parameter values (θ) such that $(\theta_{q+1} ... \theta_r) = (\theta_{0,q+1} ... \theta_{0,r})$,

Remark about terminology

As a reminder, a parameter is a value that is passed in a probability model. For example, in the normal distribution, the parameter, θ, is a vector containing the mean μ and the variance σ^2. θ can take on values in Θ, which is referred to as the parameter space.

The **likelihood ratio test statistic** is defined as

$$\lambda = 2 * log\left(\frac{\sup_{\Theta} L(\theta|\mathbf{x})}{\sup_{\Theta_0} L(\theta|\mathbf{x})}\right) = 2 * log\left(\frac{L(\hat{\theta})|x}{L(\hat{\theta_0})|x}.\right)$$

The **likelihood ratio test** is to reject H_0 when $\lambda > \chi^2_{r-q,\alpha}$.

Test Errors and Power Function

- Type I Error and Type II Error:

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

- Suppose R denotes the rejection region for a test, then the probability of a Type I Error is $P(\mathbf{X} \in R|H_0)$, and the probability of a Type II Error is $P(\mathbf{X} \in R^c|H_1) = 1 - P_\theta(\mathbf{X} \in R|H_1)$.
- A level-α test is one such that $P(\mathbf{X} \in R|H_0) \leq \alpha$.

p-values

Definition:
A **p-value**, $p(\mathbf{X})$, is a statistics such that:

$$Pr(t(\mathbf{Y}^*) \geq t(\mathbf{y})|H_0)$$

Breaking down the definition:

- $t(\mathbf{Y}^*)$: A test statistic (such as $\frac{\sqrt{n}(\bar{Y} - \mu)}{\sigma^2}$) that is a function of random data ($\mathbf{Y}^*$) that you would get under the null hypothesis.
- $t(\mathbf{y})$: The same test statistic, but of your observed data.
- This is a conditional probability. It is conditioned on the null hypothesis being true.

p-values: An Example

A neurologist is testing the effect of a drug on response time by injecting 100 rats with a unit dose of the drug, subjecting each to neurological stimulus, and recording its response time. The neurologist knows that the response time for a rat not injected with the drug follows a normal distribution with a mean response time of 1.2 seconds. The mean of the 100 injected rats' response times is 1.05 seconds with a sample standard deviation of 0.5 seconds.

Do you suggest that the neurologist conclude that the drug has an effect on response time?

Solution to the Example

Suppose the mean response time for rats injected with the drug is μ, then we want to test

$$H_0 : \mu = 1.2s \text{ (the drug has no effect)}$$

against

$$H_1 : \mu \neq 1.2s \text{ (the drug has effect)} .$$

Construct the test statistic (here $\bar{X}$ is the sample mean, and S is the sample standard deviation)

$$Z = \frac{\bar{X} - 1.2}{S/\sqrt{100}}.$$

$Z \sim t_{99}$, which is approximately $N(0, 1)$. Plug in the observed data, $\bar{x} = 1.05, s = 0.5$, and $z = -3$, so the p-value is approximately $P(|W| \geq |z|) = P(|W| \geq 3) \approx 0.003$ (let $W \sim N(0, 1)$).

Confidence Intervals

Interval Estimation

- An **interval estimate** of a parameter θ is any pair of functions, $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$, of a sample that satisfy $L(\mathbf{x}) \leq U(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. The inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made once $\mathbf{X} = \mathbf{x}$ is observed. The **random interval** $[L(\mathbf{X}), U(\mathbf{X})]$ is called an **interval estimator**.
- We call $C_n = (L(X_1 ... X_n),\ U(X_1 ... X_n))$ a $1 - \alpha$ confidence interval if $P_\theta(X \in C_n) \geq 1 - \alpha$ for all $\theta \in \Theta$
- This is not a probability statement about θ: the interval is the random quantity, not the parameter. Such interpretation will be explored further in a Bayesian context.

A mini-exercise

Suppose that X is a random sample from a distribution with parameter θ, and $[L(X), U(X)]$ is a 95% confidence interval of θ. If we observe $X = x$, which of the following statements is correct?

A The probability that $\theta \in [L(x), U(x)]$ is 0.95;

B The probability that $\theta \in [L(x), U(x)]$ is either 1 or 0.

Example: Normal Confidence Interval

If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with σ^2 known, then $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is a standard normal variable ($Z \sim N(0,1)$). Then a confidence interval of μ can be

$$\{\mu : \bar{x} - a\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + a\frac{\sigma}{\sqrt{n}}\},$$

where a is a constant.

If σ^2 is unknown, then $T_{n-1} = (\bar{X} - \mu)/(S/\sqrt{n}) \sim t_{n-1}$ which is independent of μ. Thus, for any given $\alpha \in (0,1)$, a $1-\alpha$ confidence interval of μ is given by

$$\{\mu : \bar{x} - t_{n-1,(1-\alpha/2)}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,(1-\alpha/2)}\frac{s}{\sqrt{n}}\},$$

where $t_{df,p}$ is the $p \times 100\%$th quantile of a student-t distribution with df degrees of freedom.

Bayesian Confidence Region

- As discussed before, in the frequentist context, a **random interval** has 95% coverage for θ if, before data are gathered,

$$Pr(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})|\theta) \geq 0.95$$

- An **interval**, based on observed data, has 95% Bayesian coverage for the **random variable** θ if

$$Pr(L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})|\mathbf{x}) \geq 0.95$$

- The two main types of confidence regions in Bayesian Analysis are (1) quantile-based regions and (2) highest posterior density regions, both of which will be discussed in detail in STA 601.

Reference Guide

- *Statistical Inference* - Casella and Berger
- *A First Course in Bayesian Statistical Methods* - Hoff
- *All of Statistics* - Wasserman