50.034 - Introduction to Probability and Statistics

Week 9 – Lecture 16

January–May Term, 2019

Lecture 16 Assigned Readings: 7.5, 7.6

Outline of Lecture

- Sampling from exponential distribution
- Sampling from normal distribution
- Maximum likelihood estimators (M.L.E.'s)
- Sample variance
- M.L.E.'s versus Bayes estimators

Recall: Conjugate priors

Consider a statistical model where $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

- If $X_1, \ldots, X_n$ are Bernoulli, and if we **start with any beta distribution** as our prior distribution for θ, then our posterior distribution will **remain as a beta distribution**, no matter how many observations we make, and no matter what the observed values are.
- Hence, we say that the family of beta distributions is a conjugate family of prior distributions, or more simply, a family of conjugate priors.

Key Idea: Let Ψ be a family of conjugate priors. If we choose a prior from Ψ, then the posterior will also be in Ψ.

- Remember, your initial guess (i.e. prior distribution) can be ANY distribution. If you choose a complicated prior, then the posterior would be complicated and very tedious to calculate.
- However, if you choose a prior from a family of conjugate priors, then the posterior becomes very easy to calculate!

Conjugate family of prior distributions

Suppose $X_1, \ldots, X_n$ are observable R.V.'s that are conditionally iid given the parameter θ.

- If $X_1, \ldots, X_n$ are R.V.'s in a specific class of distributions (e.g. Bernoulli distributions) whose corresponding family of conjugate priors (e.g. beta distributions) is denoted by Ψ, then we say that Ψ is closed under sampling from this given class of distributions.
  - e.g. sampling from the Bernoulli distribution for $X_1, \ldots, X_n$, starting with a beta prior distribution, would give a beta posterior distribution.

There are other families of conjugate priors:

Sampling from	**Family of conjugate priors**
Bernoulli distribution	beta distributions
binomial distribution	beta distributions
geometric distribution	beta distributions
Poisson distribution	gamma distributions
exponential distribution	gamma distributions
normal distribution	normal distributions

Sampling from exponential distribution

Theorem: Let $X_1, \ldots, X_n$ be observable **exponential** R.V.'s that are conditionally iid given the parameter θ. Suppose that θ is a gamma R.V. with prior hyperparameters α and β. Then, given the observed values $X_1 = x_1, \ldots, X_n = x_n$, the posterior hyperparameters of θ are $\alpha + n$ and $\beta + (x_1 + \cdots + x_n)$.

Interpretation:

- We started with the initial guess that θ follows the gamma distribution with parameters α and β.
- Let X_i represent the elapsed time between successive occurrences of an event of interest.
- Given observed values $X_1 = x_1, \ldots, X_n = x_n$, the total time elapsed is $(x_1 + \cdots + x_n)$ over n experiments conducted.
- We update our guess for the distribution of θ to a new gamma distribution with the following parameters:

$$\alpha' = \alpha + (\text{number of experiments}) = \alpha + n,$$

$$\beta' = \beta + (\text{total elapsed time}) = \beta + (x_1 + \cdots + x_n).$$

Example 1

A lighting company has designed a new light bulb model. They are interested in finding out how long each light bulb lasts.

Consider a statistical model consisting of observable exponential R.V.'s $X_1, \ldots, X_{10}$ that are conditionally iid given the parameter θ. Each X_i represents the lifespan (in hours) of the i-th light bulb. Suppose that θ is a gamma R.V. with prior hyperparameters 10 and 4500.

Suppose that the observed values are

$$X_1 = 500.2, X_2 = 500.1, X_3 = 500.2, X_4 = 500.1, X_5 = 500.2,$$
$$X_6 = 500.1, X_7 = 500.2, X_8 = 500.1, X_9 = 500.2, X_{10} = 500.1$$

Question: What is the posterior pdf of θ?

Example 1 - Solution

We are given the following information:

- Prior distribution is gamma with hyperparameters 10 and 4500.
- $X_1 = 500.2, X_2 = 500.1, X_3 = 500.2, X_4 = 500.1, X_5 = 500.2,$ $X_6 = 500.1, X_7 = 500.2, X_8 = 500.1, X_9 = 500.2, X_{10} = 500.1.$
- Number of experiments: 10.
- Total sum of the X_i's: 5003

Thus, the posterior distribution of θ is gamma with parameters

$$\alpha' = 10 + 10 = 20, \qquad \beta' = 4500 + 5003 = 9503.$$

Therefore, by letting $\mathbf{x}$ denote the given vector of observed values, and using the fact that $\Gamma(20) = 19!$, we conclude that the posterior pdf of θ is

$$\xi(\theta|\mathbf{x}) = \begin{cases} \frac{9503^{20}}{19!}\theta^{19}e^{-9503\theta}, & \text{if } \theta \geq 0; \\ 0, & \text{if } \theta < 0. \end{cases}$$

Sampling from normal distribution

Theorem: Let $\sigma > 0$ be a known real number. Let $X_1, \ldots, X_n$ be observable **normal** R.V.'s each with unknown mean θ and known variance σ^2. Suppose that $X_1, \ldots, X_n$ are conditionally iid given the parameter θ. If the prior distribution of θ is normal with mean μ_0 and variance v_0^2, then given the observed values $X_1 = x_1, \ldots, X_n = x_n$, the posterior distribution of θ is normal with mean μ_1 and variance v_1^2 given as follows:

$$\mu_1 = \frac{\sigma^2 \mu_0 + v_0^2 (x_1 + \cdots + x_n)}{\sigma^2 + n v_0^2},$$

$$v_1^2 = \frac{\sigma^2 v_0^2}{\sigma^2 + n v_0^2}.$$

- We started with the initial guess that θ follows the normal distribution with mean μ_0 and variance v_0^2.
- Given observed values $X_1 = x_1, \ldots, X_n = x_n$, we update our guess for the distribution of θ to a new normal distribution with mean μ_1 and variance v_1^2.

Recall: Estimators and Bayes estimators (Lecture 15)

Definition: Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by a parameter θ.

- An estimator of θ is a real-valued function $\delta(X_1, \ldots, X_n)$.
- Given δ and a vector $\mathbf{x} = (x_1, \ldots, x_n)$ of observed values, the real number $\delta(\mathbf{x})$ is called an estimate of θ.
- For every estimate a of θ obtained from $\mathbf{x}$, the Bayes risk of θ given a, $\mathbf{x}$, and some loss function $L(x, y)$, is defined by

$$\mathbf{E}[L(\theta, a)|\mathbf{x}] = \sum_{\theta \in \Omega} L(\theta, a)\xi(\theta|\mathbf{x}) \quad \text{or} \quad \int_{\Omega} L(\theta, a)\xi(\theta|\mathbf{x})\, d\theta.$$

Definition: A Bayes estimator of θ is a real-valued function $\delta^*(X_1, \ldots, X_n)$ such that for every possible observed vector $\mathbf{x}$, we set $\delta^*(\mathbf{x})$ to be a value such that the Bayes risk is minimized over all possible estimates, i.e.

$$\mathbf{E}[L(\theta, \delta^*(\mathbf{x}))|\mathbf{x}] = \min_{a \in \mathbb{R}} \mathbf{E}[L(\theta, a)|\mathbf{x}].$$

Once the vector $\mathbf{x}$ of observed values is actually observed, we say that the real number $\delta^*(\mathbf{x})$ is a Bayes estimate of θ.

Consistent estimators

Let $X_1, X_2, X_3, \ldots$ be a sequence of observable R.V.'s that are conditionally iid given the parameter θ.

- For every integer $n \geq 1$, let $\delta_n = \delta_n(X_1, \ldots, X_n)$ be an estimator of θ for the random sample $\{X_1, \ldots, X_n\}$.
- Remember, each estimator δ_n is a function of R.V.'s and hence a R.V.
- If the sequence $\delta_1, \delta_2, \delta_3, \ldots$ converges in probability to the true value of θ, then we say that $\delta_1, \delta_2, \delta_3, \ldots$ is a consistent sequence of estimators.

Nice property of many Bayes estimators:
For many classes of loss functions, and under quite general conditions, the Bayes estimators of many parameters will form a **consistent sequence of estimators** as the sample size $n \to \infty$.

- **Interpretation:** Such Bayes estimators would give estimates that are very close to the true value of the parameter when the sample size n is sufficiently large.

Maximum Likelihood Estimation

Key Idea: Choose the estimate of a parameter θ that **maximizes the likelihood function**.

- **Note:** This particular kind of estimation is very widely used, partly because it is the "easiest" kind of estimation.

Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by a parameter θ. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a given vector of observed values for $(X_1, \ldots, X_n)$.

Recall: The likelihood function of θ is either the joint conditional pmf $p_n(\mathbf{x}|\theta)$ (when all X_i's are discrete), or the joint conditional pdf $f_n(\mathbf{x}|\theta)$ (when all X_i's are continuous).

- In either case, the likelihood function is treated as a univariate function only in terms of the variable θ.
- **Note:** The definition of a likelihood function does not require that we choose a prior distribution for θ. Likelihood functions still make sense even when θ is an unknown constant.

Maximum Likelihood Estimator

Let $X_1, \ldots, X_n$ be observable R.V.'s whose joint distribution is parametrized by a parameter θ with parameter space Ω.

Definition: For every possible vector $\mathbf{x} = (x_1, \ldots, x_n)$ of observed values for $(X_1, \ldots, X_n)$, let $\delta(\mathbf{x})$ denote a value θ_0 in Ω such that the likelihood function $p_n(\mathbf{x}|\theta)$ or $f_n(\mathbf{x}|\theta)$ is maximized (over all possible $\theta \in \Omega$) at $\theta = \theta_0$. Different $\mathbf{x}$'s yield different θ_0's.

- Here, we assume that the max value θ_0 exists (for each $\mathbf{x}$).
- The function $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ that maps each $\mathbf{x}$ to $\delta(\mathbf{x})$ is called a maximum likelihood estimator (or M.L.E.) of θ.
- Once $\mathbf{x}$ is actually observed, we say that the real number $\hat{\theta}(\mathbf{x})$ is a maximum likelihood estimate of θ.
- The abbreviation M.L.E. is used to denote either maximum likelihood estimator or maximum likelihood estimate, assuming that it is clear from the context.
- **Note:** By definition, the maximum likelihood estimator is required to be an element of Ω.

Example 2

Suppose we have a coin that gives us heads with probability θ. We do not know what θ is, and we want to find an estimate for θ.

We shall toss this coin n times and record the outcomes. For our statistical model, let $X_1, \ldots, X_n$ be observable iid Bernoulli R.V.'s with unknown parameter θ, where $X_i = 1$ if the i-th toss is heads, and $X_i = 0$ otherwise.

Questions:

1. What is the likelihood function of θ?
2. What is the maximum likelihood estimator of θ?
3. Given that exactly 47% of the tosses are heads, what is the maximum likelihood estimate of θ?

Ask yourself this question: If you toss the coin 100 *times, and exactly* 47 *of the tosses are heads, what is the* ***most likely*** *value for* θ*? Remember, you only have the experimental data from these* 100 *tosses.*

Example 2 - Solution

1. [likelihood function of θ]
Since each X_i is Bernoulli, its conditional pmf given θ is

$$p_{X_i}(x_i|\theta) = \begin{cases} \theta^{x_i}(1-\theta)^{1-x_i}, & \text{if } x_i = 0 \text{ or } 1; \\ 0, & \text{otherwise.} \end{cases}$$

Consider any possible vector $\mathbf{x} = (x_1, \ldots, x_n)$ of observed values, i.e. each of $x_1, \ldots, x_n$ must be either 0 or 1.

Then the likelihood function of θ given $\mathbf{x}$ is

$$\begin{aligned} p_n(x_1, \ldots, x_n|\theta) &= p_{X_1}(x_1|\theta) \cdots p_{X_n}(x_n|\theta) \\ &= (\theta^{x_1}(1-\theta)^{1-x_1}) \cdots (\theta^{x_n}(1-\theta)^{1-x_n}) \\ &= \theta^{(x_1+\cdots+x_n)}(1-\theta)^{n-(x_1+\cdots+x_n)} \end{aligned}$$

- $X_1, \ldots, X_n$ are independent, so the joint conditional pmf equals the product of the marginal conditional pmf's.

Example 2 - Solution (continued)

2. [maximum likelihood estimator of θ]

Useful Trick: The value θ that maximizes $p_n(x_1, \ldots, x_n|\theta)$ will be the same value as the value θ that maximizes

$$\begin{aligned} g(\theta) &= \log p_n(x_1, \ldots, x_n|\theta) \\ &= (x_1 + \cdots + x_n) \log \theta + (n - (x_1 + \cdots + x_n) \log(1 - \theta), \end{aligned}$$

since log is an increasing function.

The derivative of $g(\theta)$ is

$$g'(\theta) = \frac{x_1 + \cdots + x_n}{\theta} - \frac{n - (x_1 + \cdots + x_n)}{1 - \theta}.$$

Solving for $g'(\theta) = 0$, we get $\theta = \frac{x_1 + \cdots + x_n}{n}$.

To check if the function $g(\theta)$ is maximized at $\theta = \frac{x_1 + \cdots + x_n}{n}$ in the range $0 \leq \theta \leq 1$, we need to use some calculus.

Example 2 - Solution (continued)

2. (continued)

Let $\overline{x}_n$ denote $\frac{x_1+\cdots+x_n}{n}$. After some careful checking, we get:

- If $0 < \overline{x}_n < 1$, then $g''(\overline{x}_n) = \frac{-n}{\overline{x}_n(1-\overline{x}_n)} < 0$ i.e. $g(\theta)$ is maximized at $\theta = \overline{x}_n$.
- If $\overline{x}_n = 0$ (i.e. $x_1 = \cdots = x_n = 0$), then $g(\theta) = n\log(1-\theta)$ is maximized at $\theta = 0$ in the range $0 \leq \theta \leq 1$.
- If $\overline{x}_n = 1$ (i.e. $x_1 = \cdots = x_n = 1$), then $g(\theta) = n\log\theta$ is maximized at $\theta = 1$ in the range $0 \leq \theta \leq 1$.

Thus, in all cases, $g(\theta)$ is maximized at $\theta = \overline{x}_n$.

Therefore, the **maximum likelihood estimator** of θ is

$$\hat{\theta}(X_1, \ldots, X_n) = \frac{X_1 + \cdots + X_n}{n} = \overline{X}_n.$$

In other words, the maximum likelihood estimator of a random sample of **Bernoulli** random variables is exactly the **sample mean**!

Example 2 - Solution (continued)

3. [maximum likelihood estimate of θ]
From the previous part, we have computed that the maximum likelihood estimator of θ is

$$\hat{\theta}(X_1, \ldots, X_n) = \overline{X}_n.$$

We are given that exactly 47% of the tosses are heads, which means that $\overline{X}_n = 0.47$.
Therefore, the **maximum likelihood estimate** of θ is 0.47.

Intuition:

- If exactly 47% of the tosses are heads, then the "most likely" value for the unknown parameter θ is 0.47.
- Hence, maximum likelihood estimation matches our intuition.

Example 3

A lighting company has designed a new light bulb model. They are interested in finding out how long each light bulb lasts.

Consider a statistical model consisting of observable iid exponential R.V.'s $X_1, \ldots, X_n$ with unknown parameter θ, where each X_i represents the lifespan (in hours) of the i-th light bulb. Assume that $X_1, \ldots, X_n$ are all non-zero.

Questions:

1. What is the likelihood function of θ?
2. What is the maximum likelihood estimator of θ?
3. Suppose we are given that the sample mean of $\{X_1, \ldots, X_n\}$ is $\overline{X}_n = 504$. What is the maximum likelihood estimate of θ?

Example 3 - Solution

1. [likelihood function of θ]

Since each X_i is exponential and non-zero, its conditional pdf given θ is

$$f_{X_i}(x_i|\theta) = \begin{cases} \theta e^{-\theta x_i}, & \text{if } x_i > 0; \\ 0, & \text{if } x_i \leq 0; \end{cases}$$

Consider any possible vector $\mathbf{x} = (x_1, \ldots, x_n)$ of observed values, i.e. $x_1, \ldots, x_n$ are some unspecified positive real numbers.

Then the likelihood function of θ given $\mathbf{x}$ is

$$\begin{aligned} f_n(\mathbf{x}|\theta) &= f_{X_1}(x_1|\theta) \cdots f_{X_n}(x_n|\theta) \\ &= (\theta e^{-x_1\theta}) \cdots (\theta e^{-x_n\theta}) \\ &= \theta^n e^{-(x_1+x_2+x_n)\theta}. \end{aligned}$$

[Note: We used the fact here that $X_1, \ldots, X_n$ are all non-zero.]

- $X_1, \ldots, X_n$ are independent, so the joint conditional pmf equals the product of the marginal conditional pmf's.

Example 3 - Solution (continued)

2. [maximum likelihood estimator of θ]

Useful Trick: The value θ that maximizes $f_n(\mathbf{x}|\theta)$ will be the same value as the value θ that maximizes

$$g(\theta) = \log f_n(\mathbf{x}|\theta) = n \log \theta - (x_1 + \cdots + x_n)\theta,$$

since log is an increasing function.

Taking the derivative of $g(\theta) = n \log \theta - (x_1 + \cdots + x_n)\theta$, we get

$$g'(\theta) = \frac{n}{\theta} - (x_1 + \cdots + x_n).$$

Solving for $g'(\theta) = 0$, we then get $\theta = \frac{n}{x_1 + \cdots + x_n}$.

We check that $g''(\theta) = -\frac{3}{\theta^2}$, which is negative at $\theta = \frac{n}{x_1 + \cdots + x_n}$.
Thus, $g(\theta)$ is maximized at $\theta = \frac{n}{x_1 + \cdots + x_n}$.

Therefore, the **maximum likelihood estimator** of θ is

$$\hat{\theta}(X_1, \ldots, X_n) = \frac{1}{\overline{X}_n}.$$

Example 3 - Solution (continued)

3. [maximum likelihood estimate of θ]
From the previous part, we have computed that the maximum likelihood estimator of θ is

$$\hat{\theta}(X_1, \ldots, X_n) = \frac{1}{\overline{X}_n}.$$

Therefore, given that $\overline{X}_n = 504$, we conclude that the **maximum likelihood estimate** of θ is $\frac{1}{504}$.

Intuition:

- The expectation of an exponential R.V. with parameter λ is $\frac{1}{\lambda}$.
- If we only have the observed values of $X_1, \ldots, X_n$, then the "most likely" value for the mean of each X_i is $\overline{X}_n$.
- Hence the "most likely" value for the parameter λ is $\frac{1}{\overline{X}_n}$.

M.L.E.'s of some distributions

Let $\{X_1, \ldots, X_n\}$ be a random sample, and let $\overline{X}_n$ denote its sample mean.

Random sample of	**Mean**	**M.L.E. of θ**
Bernoulli R.V.'s (parameter θ)	θ	$\overline{X}_n$
binomial R.V.'s (parameters N (known) and θ)	$N\theta$	$\frac{\overline{X}_n}{N}$
geometric R.V.'s (parameter θ)	$\frac{1-\theta}{\theta}$	$\frac{1}{1+\overline{X}_n}$
Poisson R.V.'s (parameter θ)	θ	$\overline{X}_n$
exponential R.V.'s (parameter θ)	$\frac{1}{\theta}$	$\frac{1}{\overline{X}_n}$
normal R.V.'s (mean θ and variance σ^2 (known))	θ	$\overline{X}_n$

In the computation of all these M.L.E.'s, the logarithm of the likelihood function is used (as a useful trick). Since this useful trick appears frequently in maximum likelihood estimation, the logarithm of the likelihood function is called the log-likelihood function.

Sample mean as a M.L.E.

Let $\{X_1, \ldots, X_n\}$ be any random sample with unknown mean μ.
Recall: The sample mean is $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$.

- By the law of large numbers, $\overline{X}_n \xrightarrow{\text{p}} \mu$.
- So for a large sample size n, we can use $\overline{X}_n$ to estimate μ.

Theorem: If $X_1, \ldots, X_n$ are **normal** R.V.'s, then the maximum likelihood estimator of the mean μ is

$$\hat{\mu}(X_1, \ldots, X_n) = \overline{X}_n = \frac{X_1 + \cdots + X_n}{n}.$$

Wait.. but don't we already know that μ can be estimated by $\overline{X}_n$?

- Yes, we know that $\overline{X}_n$ is an estimator of μ, and in fact a good estimator for large sample size n.
- The new insight is that the M.L.E. of μ (this specific estimator involving likelihood functions) is exactly the sample mean.
- i.e., the maximum likelihood estimation of the mean (as a parameter of normal distribution) coincides with our intuition that the sample mean is a "good" estimator of the "true" mean.

Sample Variance

Let $\{X_1, \ldots, X_n\}$ be a random sample with mean μ and variance σ^2. Suppose that μ and σ^2 are unknown, unspecified real numbers. Similarly, we can estimate the variance σ^2 by the sample variance, defined as

$$\hat{\sigma}_0^2(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Fact: $\hat{\sigma}_0^2(X_1, \ldots, X_n) \xrightarrow{p} \sigma^2$.

- So for a large sample size n, we can use $\hat{\sigma}_0^2$ to estimate σ^2.

Theorem: If $X_1, \ldots, X_n$ are **normal** R.V.'s, then the maximum likelihood estimator of the variance σ^2 is $\hat{\sigma}_0^2(X_1, \ldots, X_n)$.

Important Note: $\hat{\sigma}_0^2$ consistently underestimates σ^2 for finite samples, although its deviation from σ^2 approaches 0 as $n \to \infty$.

- For this reason, we refer to $\hat{\sigma}_0^2$ as the biased sample variance or the uncorrected sample variance
- There are several different estimators for variance!
  - We will see an "unbiased" sample variance later in this course, when we learn about **biased estimators** and **unbiased estimators** (subsequently in Lecture 18).

Some limitations of M.L.E.'s

Recall: A maximum likelihood estimate of a parameter θ is by definition a value in the parameter space of θ that maximizes the likelihood function of θ

Limitation 1: M.L.E.'s do not always exist.

- Not all functions have a maximum. Similarly, not every likelihood function has a maximum.
- See Example 7.5.8 in course textbook
  - Example involves a uniform distribution on an open interval.

Limitation 2: M.L.E.'s are not always uniquely determined.

- There could be multiple possible values in the parameter space that maximizes the likelihood function.
- Which do we choose as our maximum likelihood estimate??
- See Example 7.5.7 in course textbook.
  - Example involves a uniform distribution on the interval $[0, \theta]$.

Despite these limitations, there are many statistical models with parameters whose M.L.E.'s exist and are unique!

Nice properties of M.L.E.'s

Invariance Property: If $\hat{\theta}$ is a maximum likelihood estimator of a parameter θ, and if $g(t)$ is a one-to-one function, then $g(\hat{\theta})$ is a maximum likelihood estimator of $g(\theta)$.

- e.g. If $\hat{\theta}$ is a maximum likelihood estimator of the standard deviation of some R.V. X, then $\left(\hat{\theta}\right)^2$ is a maximum likelihood estimator of the variance of X.

Consistency Property: Let $X_1, X_2, X_3, \ldots$ be an infinite sequence of iid R.V.'s with parameter θ, where θ is an unknown constant. For every integer $n \geq 1$, let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be the M.L.E. of the random sample $\{X_1, \ldots, X_n\}$. (Remember, each M.L.E. is a function of R.V.'s and hence a R.V.) Then the sequence of R.V.'s $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \ldots$ is a consistent sequence of estimators, i.e. $\hat{\theta} \xrightarrow{p} \theta$.

- **Interpretation:** As the sample size increases, the M.L.E. $\hat{\theta}_n$ becomes closer to the true value θ.
  - The law of large numbers says that as the sampe size increases, the sample mean becomes closer to the true mean.
  - In contrast, θ could be any parameter, not necessarily mean.

M.L.E.'s versus Bayes estimators

Key difference: A Bayes estimator depends on a choice of some prior distribution, while a M.L.E. does not require any priors.

- In general, given some parameter θ, a Bayes estimator δ^* of θ would be very similar to a M.L.E. $\hat{\theta}$ of θ if the chosen prior distribution for δ^* is "close to" the uniform distribution on the parameter space of θ. (Precise details are out of syllabus.)
- However, we could choose our prior distribution to be ANY distribution on the parameter space.
  - For prior distributions "far from" the uniform distribution, the Bayes estimator would look very different.
  - In general, Bayes estimators give much more freedom and can provide "better" estimates that incorporate insights that you may already have before obtaining experimental data.

M.L.E.'s versus Bayes estimators (continued)

You could use Bayes estimators to get "quick" estimates, if you already have additional insights before obtaining experimental data and can choose the prior distribution accordingly.

Basketball Example:
Let θ be the 3-pointer success rate of a new basketball player. Based on data of 3-point shooting from **other** basketball players, you could choose a prior distribution that reflects the overall distribution of actual 3-pointer success rate.

- If you observed ten 3-point shot attempts, and all were successful, then the maximum likelihood estimate is $\theta = 1$, while the Bayes estimate could be $\theta = 0.42$.
- Hence, the Bayes estimate would be a much better estimate just from very limited data (observed values from ten 3-point shot attempts).

Summary

- Sampling from exponential distribution
- Sampling from normal distribution
- Maximum likelihood estimators (M.L.E.'s)
- Sample variance
- M.L.E.'s versus Bayes estimators

Reminder:

- There is **Mini-quiz 3** (15mins) this week during cohort class.
  - Tested on materials from Lectures 11–13 only.
  - As mentioned during Lecture 13, the focus for Mini-quiz 3 will be on materials covered in Lecture 13.