STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 3

Last class

- Bayes' Theorem and the *ideal* approach to inference.
- Example: A political poll (a binomial experiment).

Political poll example revisited

An organization randomly selected 100 democrats and interviewed them about whether they support the incombent governer. Out of the 100 polled, 40 support the governer and 60 against. What can we say about the actual support rate of the governer θ?

Our model (or assumptions/hypotheses) are

(1) Given θ, X is Binomial(100,θ).

(2) θ is Uniform(0,1) *a priori*—representing our knowledge before data is observed.

Based on these two pieces, Bayes' Theorem allows us to use the observed data to update our knowledge about the parameter.

(3) θ is Beta(41,61) *a posteriori*—representing our updated knowledge after data is observed.

- Our analysis up to this point uses Uniform(0,1) to represent our prior knowledge about θ.
- More flexible choices for $\xi(\theta)$ can allow richer prior knowledge about the parameter to be incorporated into the analysis.

A richer class of priors for Binomial experiments

A flexible and convenient choice of the prior distribution for θ is the Beta(α, β) family of distributions.

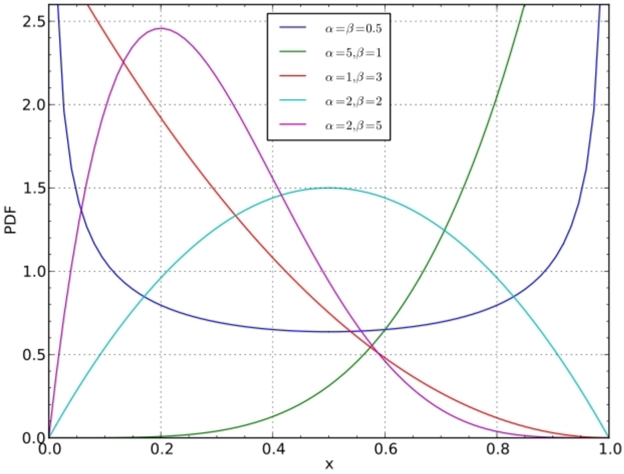
$$\begin{aligned} \xi(\theta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0 < \theta < 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

where $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$. (This is called the *Gamma* function.)

- Mean: $\mu_\theta = \frac{\alpha}{\alpha+\beta}$.
- Variance: $\sigma_\theta^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}$.

Parameters such as α and β that are used to characterize the prior distribution of θ are called *hyperparameters*.

Remark: When $\alpha = \beta = 1$, this is exactly the Uniform(0,1) distribution.



Source: Wikipedia.
`http://upload.wikimedia.org/wikipedia/commons/f/f3/Beta_distribution_pdf.svg`.

Bayesian inference using a Beta(α, β) prior

$$\begin{aligned}
\xi(\theta|x) &\propto p(x|\theta)\xi(\theta) \\
&= \binom{n}{x}\theta^x(1-\theta)^{n-x} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= C(x)\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \quad \text{for } 0<\theta<1 \\
&= 0 \quad \text{othewise.}
\end{aligned}$$

To make this a density, we must have

$$C(x) = \frac{1}{\int_0^1 u^{x+\alpha-1}(1-u)^{n-x+\beta-1}du}.$$

But we don't need to calculate the integral in the denominator due to the following trick.

A trick to get the posterior density

The part that depends on θ

$$\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \quad \text{for } 0<\theta<1$$

is the same as that in the p.d.f of a Beta$(x+\alpha, n-x+\beta)$ distribution. Thus the two density functions must be *identical*. So we must have

$$\begin{aligned} f(\theta|x) &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)}\theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \quad \text{for } 0<\theta<1 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

A quick sum up

To make inference on a Binomial experiment using a Beta prior, we can proceed as follows

1. We model the distribution of the number of successes X conditional on the success probability θ as Binomial(n,θ).
2. We can then model the prior distribution $\xi(\theta)$ as a Beta(α,β) density. The value of α and β are chosen to represent our prior knowledge about the parameter.
3. After observing $X=x$, Bayes' theorem tells us that the posterior distribution of θ is Beta$(x+\alpha, n-x+\beta)$.

The three steps in all Bayesian statistical analysis

This example illustrates the general process of *all* Bayesian statistical analysis.

1. Model the distribution of data X conditional on a set of parameter θ.
2. Choose a prior distribution $\xi(\theta)$ for the parameters, and
3. Apply Bayes' Theorem to find the posterior distribution $\xi(\theta|x)$.

This posterior distribution is a *probabilistic* summary of our knowledge about the parameters given the data. It allows us to make probabilistic statements about the parameters such as follows.

- *Given the data*, the probability for θ to be in (0.7,0.8) is ....
- *Given the data*, the expected value of θ is ....

Choosing prior distribution to represent prior knowledge

Suppose before the experiment is carried out, from historical background, we think that the actual proportion of supportors is about 0.5, “give or take” 0.1.

- Here by “give or take” I mean that we are willing to assume that prior standard deviation of θ is about 0.1. (Note that this is a probabilistic statement about θ!)

Which α and β values should we choose so that our prior Beta(α, β) will represent this knowledge?

We can choose the values for α and β so that

$$\mu_\theta = \frac{\alpha}{\alpha+\beta} = .5$$

and

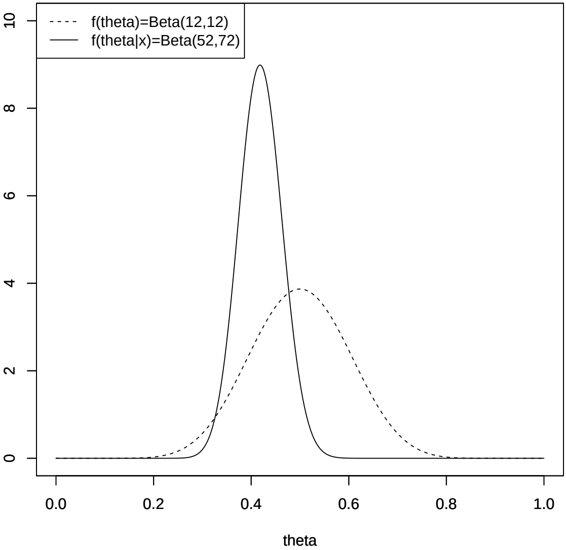
$$\sigma_\theta^2 = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1} = .1^2.$$

Solving these two equations we get

$$\alpha = \beta = 12.$$

Therefore we choose Beta(12,12) distribution as our prior distribution for θ.

From prior to posterior

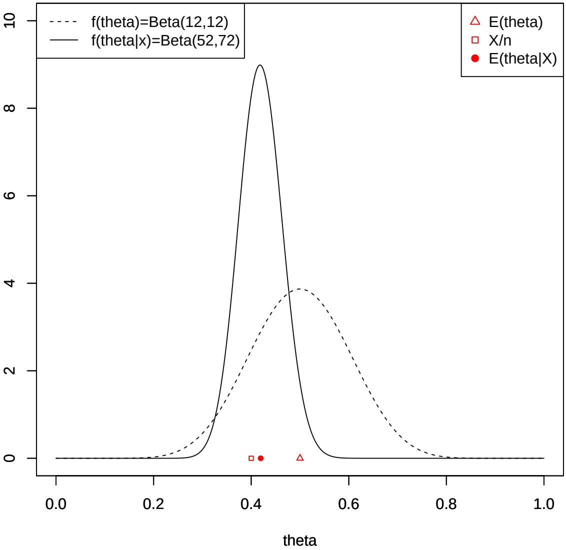


Summarizing the effect of data on the distribution of θ

- The center of the distribution is shifted.
- The spread of the distribution is also changed.

Let us quantify these two changes.

From prior to posterior



Before observing data, the prior distribution is Beta(α, β). So

$$E(\theta) = \frac{\alpha}{\alpha+\beta} = \mu_\theta$$
$$\text{Var}(\theta) = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}.$$

After observing the data, $X = x$, the posterior distribution is Beta($\alpha + x$,$\beta + n - x$), and so

$$\begin{aligned} E(\theta|X=x) &= \frac{\alpha+x}{\alpha+\beta+n} \\ &= \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\mu_\theta + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n} = \mu_{\theta|x} \end{aligned}$$

and

$$\text{Var}(\theta|X=x) = \frac{\mu_{\theta|x}(1-\mu_{\theta|x})}{\alpha+\beta+n+1}.$$

An interpretation of the posterior mean

$$E(\theta|X=x) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\mu_\theta + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n}$$

- This is a weighted average of the prior mean μ_θ, and the observed average x/n.
- This is a compromise between our *prior expectation* (when there is no data) and the *observed average*.
- When we have a lot of data, i.e. n is very large compared to $\alpha+\beta$, the observed average x/n dominates $E(\theta|X=x) \approx x/n$.

An interpretation of the posterior mean

$$E(\theta|X=x) = \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\mu_\theta + \left(\frac{n}{\alpha+\beta+n}\right)\frac{x}{n}$$

- When $\alpha+\beta$ is much larger than n, then the prior expectation μ_θ donminates $E(\theta|X=x) \approx \mu_\theta$.
- $\alpha+\beta$ is the prior "sample size"—it measures the amount of information we have about θ.
- n is the observed sample size—it measures the amount of information we have in the data.

As for the variance

Before the data are observed,

$$\mathrm{Var}(\theta) = \frac{\mu_\theta(1-\mu_\theta)}{\alpha+\beta+1}$$

- For the same μ_θ, the larger $\alpha+\beta$ is, the smaller $\mathrm{Var}(\theta)$ is—we are more certain about the value of θ.

After the data are observed,

$$\mathrm{Var}(\theta|X=x) = \frac{\mu_{\theta|x}(1-\mu_{\theta|x})}{\alpha+\beta+n+1}.$$

- For the same $\mu_{\theta|x}$, the larger $\alpha+\beta+n$—total amount of information, the smaller $\mathrm{Var}(\theta|X=x)$.

The general applicability of Bayes' Theorem

Recall the two pieces we need to apply Bayes' Theorem

1. The distribution of the data *conditional* on the state of nature $f(x|\theta)$.
2. The *prior* distribution $\xi(\theta)$ on the state of nature θ.

Bayes' Theorem provide a probabilistic recipe for inference through learning the *posterior* distribution of the state of nature given the data, $\xi(\theta|x)$.

- We have seen a particular example where $\xi(\theta)$ is Beta(α, β) and $f(x|\theta)$ is Binomial.
- This scheme for inference applies generally to *all* choices of $f(x|\theta)$ and $\xi(\theta)$.
- Let us now look at another example.

Example: Measuring air quality with an imperfect device

Suppose we want to measure the amount (in terms of density) of a certain pollutant in the air.

- We have a device that is accurate “on average”.
- However it may in any single reading be off by an unpredictable amount, due to factors such as temperature, humidity, and the way the device is held, etc.

In formal terms, let

- θ be the actual amount of the pollutant in the air.
- X be a reading of the device, which given θ is a random quantity that is close but not exactly θ.

Goal of inference: What is θ?

The Bayesian procedure

Let us carry out a Bayesian inference just as we did for the political poll example. Again, we need to specify the two pieces:

1. The conditional distribution of the data given the parameter (or state of nature).
2. A *prior* distribution $\xi(\theta)$ for the state of nature that summarizes our knowledge about θ before observing data.

What may be a good model for the reading of the device given θ?

- We can *model* the error made by the device on each reading as normally distributed with standard deviation τ. For simplicity, let us assume that this τ is known. The user's manual will typically give the technical specs of the device including its "precision". For example, $\tau = 2$. Under this model, the conditional density for $X = x$ given θ is

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(x-\theta)^2}{2\tau^2}} \quad \text{for } -\infty < x < \infty.$$

Take-home question: What if the value of τ is unknown? How can we carry out a Bayesian inference?

Choosing the prior distribution for θ

We can *model* our prior knowledge of the actual amount θ also as a Normal(μ, σ^2) distribution.

- The mean of the prior, μ, can be chosen as the historical average amount, e.g. $\mu = 100$.
- We can choose σ so that the historical records of the amount fall in the range of $\mu \pm \sigma$ about 2/3 of the times.

For example, if historically the amount of the pollutant falls in the range 100 ± 10 about 2/3 of the time, then we may choose

$$\mu = 100 \quad \text{and} \quad \sigma = 10.$$

Now that we have the two pieces, inference again is an application of Bayes' theorem.

$$\begin{aligned}
\xi(\theta|x) &\propto \xi(\theta) f(x|\theta) \\
&= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\theta-\mu)^2}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{(x-\theta)^2}{2\tau^2}} \\
&= \frac{1}{2\pi\sigma\tau} e^{-[\frac{(\theta-\mu)^2}{2\sigma^2}+\frac{(x-\theta)^2}{2\tau^2}]} \\
&\propto e^{-\frac{1}{2}[\frac{(\theta-\mu)^2}{\sigma^2}+\frac{(x-\theta)^2}{\tau^2}]}.
\end{aligned}$$

What is this distribution?

Note that

$$
\begin{aligned}
&\frac{(\theta-\mu)^2}{\sigma^2}+\frac{(x-\theta)^2}{\tau^2}\\
&=\theta^2(\frac{1}{\sigma^2}+\frac{1}{\tau^2})-2\theta(\frac{\mu}{\sigma^2}+\frac{x}{\tau^2})+\frac{\mu^2}{\sigma^2}+\frac{x^2}{\tau^2}\\
&=\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta^2-2\theta\left(\frac{\mu/\sigma^2+x/\tau^2}{1/\sigma^2+1/\tau^2}\right)+\left(\frac{\mu/\sigma^2+x/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2\right)+Const\\
&=\left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\theta-\frac{\mu/\sigma^2+x/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2+Const\\
&=\frac{(\theta-\tilde{\mu})^2}{\tilde{\sigma}^2}+Const
\end{aligned}
$$

where

$$\tilde{\mu}=\frac{\mu/\sigma^2+x/\tau^2}{1/\sigma^2+1/\tau^2}\quad\text{and}\quad\tilde{\sigma}^2=\frac{1}{1/\sigma^2+1/\tau^2}.$$

Thus

$$\xi(\theta|x) \propto e^{-\frac{(\theta-\tilde{\mu})^2}{2\tilde{\sigma}^2}} \quad \text{for } -\infty < \theta < \infty.$$

This is the same as the p.d.f of a Normal($\tilde{\mu}, \tilde{\sigma}^2$) distribution up to a normalizing constant. Therefore we must have

$$\xi(\theta|x) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} e^{-\frac{(\theta-\tilde{\mu})^2}{2\tilde{\sigma}^2}} \quad \text{for } -\infty < \theta < \infty.$$

and the posterior expectation and variance are

$$E(\theta|X=x) = \tilde{\mu} = \left(\frac{1/\sigma^2}{1/\sigma^2 + 1/\tau^2}\right)\mu + \left(\frac{1/\tau^2}{1/\sigma^2 + 1/\tau^2}\right)x$$

and

$$\text{Var}(\theta|X=x) = \tilde{\sigma}^2 = \frac{1}{1/\sigma^2 + 1/\tau^2}.$$

Note that the posterior expectation is again a weighted average of the prior expectation μ and the observed data x.

$$E(\theta|X=x) = \left(\frac{1/\sigma^2}{1/\sigma^2+1/\tau^2}\right)\mu + \left(\frac{1/\tau^2}{1/\sigma^2+1/\tau^2}\right)x.$$

The weights are determined by the relative sizes of $1/\sigma^2$ (which measures the amount of information about θ in the prior) and $1/\tau^2$ (which measures the amount of information about θ in the data).

- For fixed τ^2, as $\sigma^2 \downarrow 0$, we have $E(\theta|X=x) \to \mu$ and $\mathrm{Var}(\theta|X=x) \downarrow 0$. We are *a priori* “certain” that $\theta = \mu$.
- For fixed σ^2, as $\tau^2 \downarrow 0$, we have $E(\theta|X=x) \to x$ and $\mathrm{Var}(\theta|X=x) \downarrow 0$. We have a “perfect” device.

The posterior variance

$$\mathrm{Var}(\theta|X=x)=\tilde{\sigma}^2=\frac{1}{1/\sigma^2+1/\tau^2}$$

is smaller than both σ^2 and τ^2. We can think of $1/\tilde{\sigma}^2$ as a measure of the information we have about θ. The total informaion is the sum of prior information and the information from data:

$$\frac{1}{\tilde{\sigma}^2}=\frac{1}{\sigma^2}+\frac{1}{\tau^2}.$$

For our ongoing example

Suppose we observe $X = 105$ and we know that $\tau = 2$, then

$$E(\theta|X = 105) = \frac{1/10^2}{1/10^2 + 1/2^2} \cdot 100 + \frac{1/2^2}{1/10^2 + 1/2^2} \cdot 105 = 104.8$$

$$\text{Var}(\theta|X = 105) = 1/(1/10^2 + 1/2^2) = 3.85.$$

If instead $\sigma = 1$, (so we are a lot more certain *a priori*),

$$E(\theta|X = 105) = \frac{1/1^2}{1/1^2 + 1/2^2} \cdot 100 + \frac{1/2^2}{1/1^2 + 1/2^2} \cdot 105 = 101$$

$$\text{Var}(\theta|X = 105) = 1/(1/1^2 + 1/2^2) = 0.89.$$

Bayes' Theorem for multiple independent observations

- Now suppose instead of taking one reading from the device, we take n *independent* readings $X_1, X_2, \ldots, X_n$. That is, conditional on θ, the X_i's are *independent identitically distributed* (i.i.d.) $N(\theta, \tau^2)$ random variables. That is

$$f(x_1, x_2, \ldots, x_n | \theta) = \prod_{i=1}^{n} f(x_i | \theta) = \frac{1}{(2\pi)^{n/2} \tau^n} e^{-\frac{\sum_{i=1}^{n} (x_i - \theta)^2}{2\tau^2}}.$$

- Then Bayes' Theorem applies just as before:

$$\begin{aligned} \xi(\theta | x_1, x_2, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n | \theta) \xi(\theta)}{\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n | u) \xi(u) du} \\ &\propto f(x_1, x_2, \ldots, x_n | \theta) \xi(\theta) \\ &= \xi(\theta) \prod_{i=1}^{n} f(x_i | \theta). \end{aligned}$$

Exercise

If we still use $N(\mu, \sigma^2)$ as the prior for the parameter θ, show that given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, the posterior distribution for θ is $N(\tilde{\mu}, \tilde{\sigma}^2)$ with

$$\tilde{\mu} = \left(\frac{1/\sigma^2}{1/\sigma^2 + n/\tau^2} \right) \mu + \left(\frac{n/\tau^2}{1/\sigma^2 + n/\tau^2} \right) \bar{x}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

is the sample average, and

$$\tilde{\sigma}^2 = \frac{1}{1/\sigma^2 + n/\tau^2}.$$

Next ...

- Conjugate models.
- Point estimation.