Probability and Inference

September 13, 2013

Simple example number 1

Suppose we have a sample of N observations on some interesting variable, say daily returns on a stock or weekly expenditures on food by those families in a survey with a single parent and three or more children under 4.

Simple example number 1

Suppose we have a sample of N observations on some interesting variable, say daily returns on a stock or weekly expenditures on food by those families in a survey with a single parent and three or more children under 4.

We are interested in the population mean. For the stock prices, we think of the period for which we have observations as representative of other time periods in which we think the stock prices behave in the same way. For the food expenditures, we are interested in the mean over the whole population and recognize that we have only a random sample of the population.

Simple example number 1

Suppose we have a sample of N observations on some interesting variable, say daily returns on a stock or weekly expenditures on food by those families in a survey with a single parent and three or more children under 4.

We are interested in the population mean. For the stock prices, we think of the period for which we have observations as representative of other time periods in which we think the stock prices behave in the same way. For the food expenditures, we are interested in the mean over the whole population and recognize that we have only a random sample of the population.

We could postulate that the observations x_j, $j = 1, \ldots, N$ are independently and identically distributed (i.i.d.) across j, and further that they have a $N(\mu, \sigma^2)$ distribution, where μ is the population mean and σ^2 the population variance.

Distribution of the observations, likelihood

The **probability density function**, or **pdf** of the data is

$$p(x_1, \ldots, x_N \mid \mu, \sigma^2) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x_j-\mu)^2}{2\sigma^2}}$$

Notice: The "$p(\cdots \mid \ldots)$" notation stands for the *conditional* probability density of what is before the "|" given what comes after it. Usually what is conditioned on is itself a random variable. So here, as elswhere in this course, we are treating the "parameters" μ and σ as random variables.

When we evaluate the pdf at the observed sample values of x_j and then treat it as a function of μ and σ^2, we call it the **likelihood function.** The likelihood function is large around values of μ, σ^2 that that make the observed data highly "likely".

Log likelihood

The log of the likelihood simplifies to

$$-\frac{N}{2}\log(2\pi) - N\log\sigma - \tfrac{1}{2}\sum_{j=1}^{N}\frac{(x_j-\mu)^2}{\sigma^2}$$

$$= -N\log(2\pi) - N\log\sigma - \tfrac{1}{2}\frac{(x-\mu\,\mathbf{1})'(x-\mu\,\mathbf{1})}{\sigma^2}\,,$$

where x is the vector $(x_1,\ldots,x_N)'$ and $\mathbf{1}$ is a column vector of 1's.

The obvious estimator

As you already know, a good estimator of μ here is

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j \, .$$

The obvious estimator

As you already know, a good estimator of μ here is

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j \ .$$

Why "obvious"? Here $\mu = E[x_j]$, the expectation of the x's under the assumed normal population distribution. $\bar{x}$ is the "sample average" of x. When we are trying estimate something that can be described as $E[f(x_j)]$ (where f is some arbitrary function), it often is reasonable to estimate it as the sample average value of $f(x_j)$.

Arguments for $\bar{x}$ based on behavior of the estimator

$\bar{x}$ is an **unbiased** estimator of μ, meaning that under the population distribution, $E[\bar{x}] = \mu$, regardless of the true value of μ or of σ^2.

Arguments for $\bar{x}$ based on behavior of the estimator

$\bar{x}$ is an **unbiased** estimator of μ, meaning that under the population distribution, $E[\bar{x}] = \mu$, regardless of the true value of μ or of σ^2.

Note that this is a claim about the behavior of the estimator across hypothetical additional samples that are not the one we have observed.

Arguments for $\bar{x}$ based on behavior of the estimator

$\bar{x}$ is an **unbiased** estimator of μ, meaning that under the population distribution, $E[\bar{x}] = \mu$, regardless of the true value of μ or of σ^2.

Note that this is a claim about the behavior of the estimator across hypothetical additional samples that are not the one we have observed.

Unbiasedness does not require the normality assumption or the i.i.d. assumption. It holds so long as $E[x_j] = \mu$ for each j.

There are arguments, which we will not go into for now, that under our normality assumptions, or even somewhat weaker assumptions, $\bar{x}$ will be as close to μ on average (across hypothetical samples) as is possible — i.e. that it is in some sense **efficient**.

Arguments for $\bar{x}$ based on likelihood

Under the normality assumptions, $\bar{x}$ is the most likely value of μ no matter what the value of σ^2 might be. That is, $\bar{x}$ is the **maximum likelihood** estimator of μ.

If we are going to make probability statements about the unknown μ after seeing the data, it seems plausible that we might treat the likelihood function as tracing out a pdf for the unknown μ conditional on the observed data.

Arguments for $\bar{x}$ based on likelihood

Under the normality assumptions, $\bar{x}$ is the most likely value of μ no matter what the value of σ^2 might be. That is, $\bar{x}$ is the **maximum likelihood** estimator of μ.

If we are going to make probability statements about the unknown μ after seeing the data, it seems plausible that we might treat the likelihood function as tracing out a pdf for the unknown μ conditional on the observed data.

For now, consider the simple case where σ^2 is known. Then the likelihood function, as a function of μ alone, has the shape of a $N(\bar{x}, \sigma^2/N)$ pdf. So it is centered at $\bar{x}$, and looks "unbiased", in that it implies $E[\mu \mid x] = \bar{x}$.

Arguments for $\bar{x}$ based on likelihood

Under the normality assumptions, $\bar{x}$ is the most likely value of μ no matter what the value of σ^2 might be. That is, $\bar{x}$ is the **maximum likelihood** estimator of μ.

If we are going to make probability statements about the unknown μ after seeing the data, it seems plausible that we might treat the likelihood function as tracing out a pdf for the unknown μ conditional on the observed data.

For now, consider the simple case where σ^2 is known. Then the likelihood function, as a function of μ alone, has the shape of a $N(\bar{x}, \sigma^2/N)$ pdf. So it is centered at $\bar{x}$, and looks "unbiased", in that it implies $E[\mu \mid x] = \bar{x}$.

This way of thinking gives us both a natural estimator of μ, and a characterization of our uncertainty, given the sample x values, about μ.

Characterizing uncertainty about μ: after we've seen the sample

With σ^2 known, the likelihood, as a function of μ, is proportional to a normal density with mean $\bar{x}$ and variance σ^2/N.

Characterizing uncertainty about μ: after we've seen the sample

With σ^2 known, the likelihood, as a function of μ, is proportional to a normal density with mean $\bar{x}$ and variance σ^2/N.

So it seems natural to say, after seeing the sample, that we are 95% certain that μ is in the interval $(\bar{x} - 1.96\sigma, \bar{x} + 1.96\sigma)$ — that is, just cut of 2.5% of the probability in each tail of the likelihood, after scaling the likelihood to integrate to one.

Characterizing uncertainty about $\bar{x}$: before we've seen the sample

Before we see any data, we know that, conditional on μ and σ^2, $\bar{x} \sim N(\mu, \sigma^2)$. Thus, before we see the sample, we can say that, conditional on μ and σ^2,

$$P[\mu - 1.96\sigma < \bar{x} < \mu + 1.96\sigma] = .95\,.$$

Characterizing uncertainty about $\bar{x}$: before we've seen the sample

Before we see any data, we know that, conditional on μ and σ^2, $\bar{x} \sim N(\mu, \sigma^2)$. Thus, before we see the sample, we can say that, conditional on μ and σ^2,

$$P[\mu - 1.96\sigma < \bar{x} < \mu + 1.96\sigma] = .95\,.$$

Semantic trick

We can rewrite the expression as

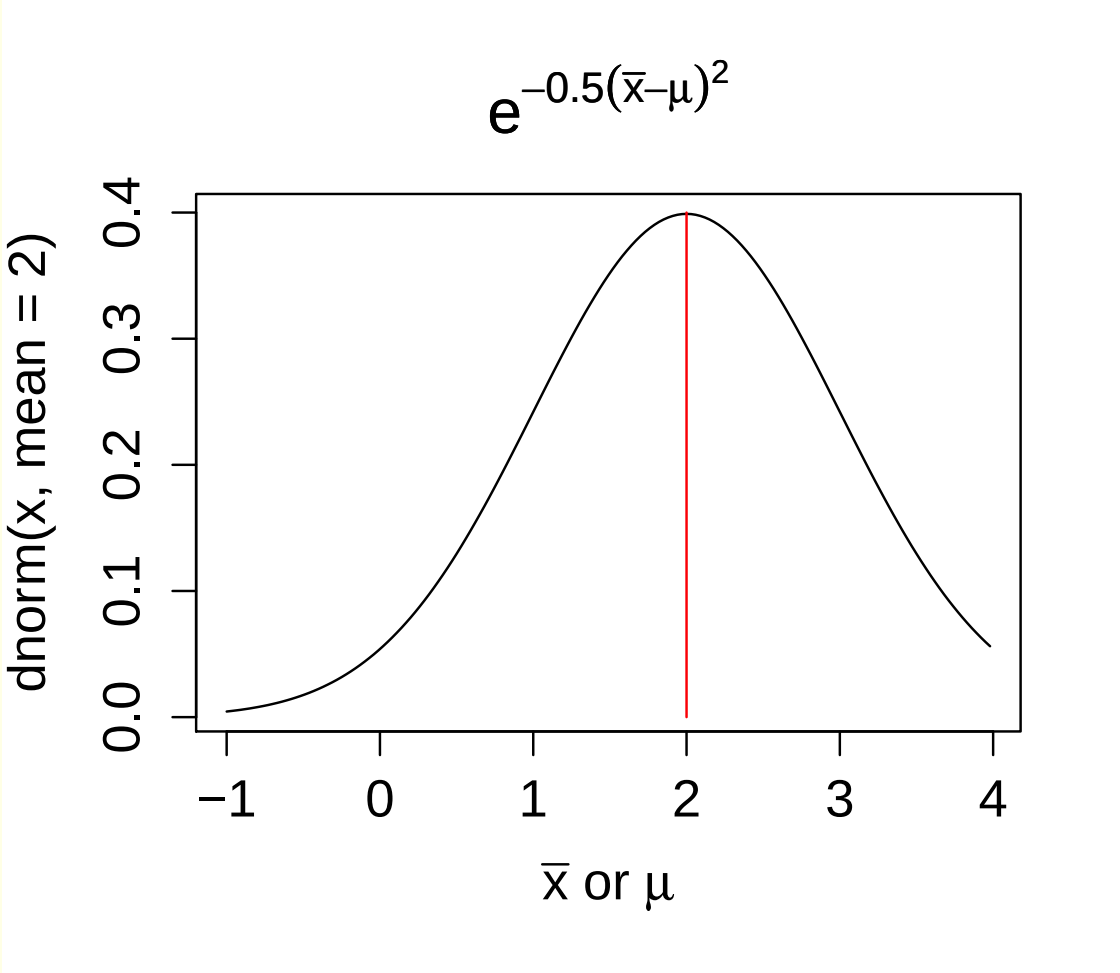
$$P[\bar{x} - 1.96 < \mu < \bar{x} + 1.96] = .95$$

and read it as "the probability that μ is in the interval $\bar{x} \pm 1.96$ is .95". But that does not change the fact that this is a probability statement about the random interval endpoints $\bar{x} \pm 1.96$, *before* we see the sample. If we are not willing to put probability distributions on unknown parameters, then after we have seen the sample, so that $\bar{x}$ is no longer uncertain, the interval either contains μ or it doesn't: the probability is zero or one.

Splitting hairs?

After all, we've already seen that by treating the likelihood shape as defining a pdf for μ given the observed sample, we can justify saying that the probability that $\bar{x} \pm 1.96\sigma$ contains μ is .95 *after* we've seen the sample, and treating μ as a random variable. It's the same interval! Why make such a fuss about pre-sample and post-sample probabilities?

Likelihood and pdf of $\bar{x}$, all in one plot



It matters

- This "pure location shift parameter" situation is the only one in which likelihood-based intervals and pre-sample confidence intervals are the same.

- It is often much easier and more intuitive to construct likelihood-based intervals and probability statements than to form the corresponding pre-sample confidence statements.

- Where confidence intervals do not correspond to post sample probability intervals, they can be a bad guide to decision making if misinterpreted as probability intervals.

It matters

- This "pure location shift parameter" situation is the only one in which likelihood-based intervals and pre-sample confidence intervals are the same.

- It is often much easier and more intuitive to construct likelihood-based intervals and probability statements than to form the corresponding pre-sample confidence statements.

- Where confidence intervals do not correspond to post sample probability intervals, they can be a bad guide to decision making if misinterpreted as probability intervals.

- Which leads us to:

Simple example number 2

A bank has made N mortgages of a certain new type, and all have been outstanding 5 years. $n << N$ of them have defaulted. It would like to estimate the probability p of default in the first five years for this type of mortgage, and get some idea of how much uncertainty there is about the probability, given the observed data.

The data x are a vector of 0's and 1's, the 0's standing for non-default and the 1's standing for default. We assume these are i.i.d., all with the same p. (More realistically, we might know something about the characteristics of the borrowers and the houses mortgaged, so we could model p as differing across mortgages, but here we keep things simpler by assuming a fixed p.)

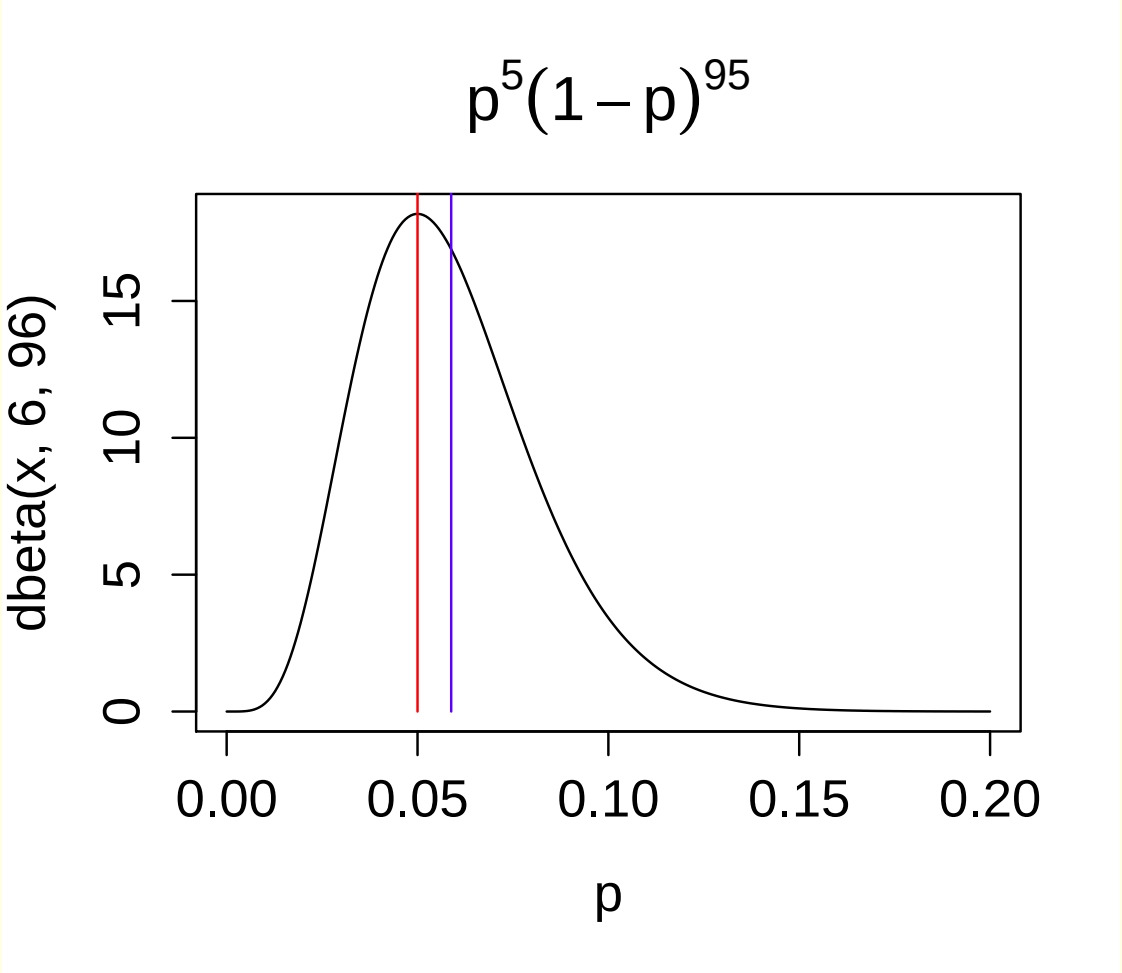
Likelihood

The pdf of x given p is $\prod_{j=1}^{N} p^{x_j}(1-p)^{1-x_j}$. Notice that for each j, the term in the product is either p (if $x_j = 1$) or $1-p$) (if $x_j = 0$). Collecting terms, the likelihood function is

$$p^n(1-p)^{N-n}\,.$$

Suppose $N = 100$ and $n = 5$.

The likelihood for $n = 5$, $N = 100$



95% interval

If we cut off the upper and lower 2.5% tails of this distribution, we get a 95% probability interval of (.0221, .1118). We can make the interval shorter by shifting it slightly to the left: (.0181, .1048).

95% interval

If we cut off the upper and lower 2.5% tails of this distribution, we get a 95% probability interval of (.0221, .1118). We can make the interval shorter by shifting it slightly to the left: (.0181, .1048).

(The shortest interval with 95% probability will have the likelihood the same height at each end. Proving this to yourself can exercise your calculus a bit.)

95% interval

If we cut off the upper and lower 2.5% tails of this distribution, we get a 95% probability interval of (.0221, .1118). We can make the interval shorter by shifting it slightly to the left: (.0181, .1048).

(The shortest interval with 95% probability will have the likelihood the same height at each end. Proving this to yourself can exercise your calculus a bit.)

Note that the interval is quite asymmetric about the maximum likelihood value of $p = .05$. The plot shows the MLE (maximum likelihood estimator) as a red vertical line and the expectation of p given the data, treating the likelihood as its pdf, as a blue vertical line. The expectation is $6/102 = .059$, slightly above the MLE because of the right-skewness of the distribution.

The skewness is because five defaults is much less likely with very low p than with p well above the MLE. With $p = 0$, for example, the probability of 5 defaults is obviously zero, whereas with $p = .1$ — the same distance above .05 as 0 is below .05 — the probability of five defaults is .034, about one fifth of the probability with $p = .05$, which is .180.

Pre-sample probability approach

There is an unbiased estimator available here, and it turns out to match the MLE:

$$\hat{p} = \frac{n}{N} \, .$$

This is again a sample average, and it estimates the population expectation $E[x_j \mid p] = p$. The pdf of this estimator, assuming we take N as fixed, takes values only on the discrete set of points $\{.01, .02, \ldots, 1\}$ — those are the only possible values for n/N.

Pre-sample probability approach

There is an unbiased estimator available here, and it turns out to match the MLE:

$$\hat{p} = \frac{n}{N} .$$

This is again a sample average, and it estimates the population expectation $E[x_j \mid p] = p$. The pdf of this estimator, assuming we take N as fixed, takes values only on the discrete set of points $\{.01, .02, \ldots, 1\}$ — those are the only possible values for n/N.

This is no longer pure-location-shift country. The distribution for low p values is tightly packed on low values of n/N, whereas for high p values it is more spread out and symmetric.

The unbiased estimator can look quite "biased", after we see the sample. For example, the sample might turn out to have $n = 0$. This is quite possible for low $p > 0$. In this case $\hat{p} = 0$, but it is obvious that $p < 0$ is impossible, while $p > 0$ is possible, so any reasonable expectation of p after we've seen $n = 0$ must be positive. 0 is surely downward "biased" in this sense, even though in the technical statistical sense it is an unbiased estimator.

A pre-sample confidence interval

In the normal example with known σ^2, we had what is known as a **pivot**, a function of the data and the unknown parameter that had a distribution that did not depend on the parameter: $\bar{x} - \mu$, which is $N(0, \sigma^2/N)$ no matter what μ is. This made it straightforward to find a confidence interval. But in our current example there is no pivot. It is still possible to construct a confidence interval. For each p, construct a test of the null hypothesis that p is the true parameter at the 5% level. For any given sample, collect all the p values for which the hypothesis is not rejected. This is a 95% pre-sample confidence interval. The calculations are messy, but the result, using two-sided, equal-tail tests, for our $n = 5$, $N = 100$ example is the interval (.0223, .1128) — slightly longer, but very similar to the equal-tailed likelihood-based 95% probability interval.

Moral of the story, so far

Treating the likelihood function as a pdf for the unknown parameter gives sensible results, is often easy to implement and display graphically, and often gives probability intervals that are similar to confidence intervals for the same problem.

Moral of the story, so far

Treating the likelihood function as a pdf for the unknown parameter gives sensible results, is often easy to implement and display graphically, and often gives probability intervals that are similar to confidence intervals for the same problem.

A really bad confidence interval

Confidence intervals can be unreasonable in some samples. For example, they can be empty or can include the entire parameter space. This happens with non-zero probability for some confidence intervals in complex models that are actually used in econometric practice.

A really bad confidence interval

Confidence intervals can be unreasonable in some samples. For example, they can be empty or can include the entire parameter space. This happens with non-zero probability for some confidence intervals in complex models that are actually used in econometric practice.

Here is a simple, though somewhat contrived, example. Suppose in our mortgage default probability example we knew from the start that the true default rate must be at least 4%. Possibly the new type of mortgage just loosens the criteria for borrower eligibility, and we know that the default rate was 4% under the old, tighter, eligibility requirements. So the only question is, how much over 4% is the default rate on this new type.

A really bad confidence interval

Confidence intervals can be unreasonable in some samples. For example, they can be empty or can include the entire parameter space. This happens with non-zero probability for some confidence intervals in complex models that are actually used in econometric practice.

Here is a simple, though somewhat contrived, example. Suppose in our mortgage default probability example we knew from the start that the true default rate must be at least 4%. Possibly the new type of mortgage just loosens the criteria for borrower eligibility, and we know that the default rate was 4% under the old, tighter, eligibility requirements. So the only question is, how much over 4% is the default rate on this new type.

How to construct the interval

The likelihood approach: Look at the likelihood function over (.04, 1), scale it to integrate to one, and proceed as before. Gives us (.04, .0681) as the 95% interval when $n = 0$.

Confidence interval approach: Use the same confidence intervals as before, just discarding the part of the interval that falls below .04. (Be sure you understand why, if we know $p \geq .04$, discarding the part of the confidence interval below .04 leaves it still a 95% confidence interval.) When $n = 0$, this gives us an empty interval. Every $p \geq .04$ is rejected by an equal-tail two-sided test.

The likelihood based probability interval will never be empty.

A rigorous argument for likelihood-based inference

If we're going to use likelihood to treat an unknown parameter as a random variable after we've seen the data, we should treat it as random before we see the data also. Its distribution before we see the data is called the **prior** distribution.

A rigorous argument for likelihood-based inference

If we're going to use likelihood to treat an unknown parameter as a random variable after we've seen the data, we should treat it as random before we see the data also. Its distribution before we see the data is called the **prior** distribution.

The model is in the form of a conditional probability distribution for the data x given the parameter β, $p(y \mid \beta)$. Applying the standard rule that the joint probability density of y and z is the marginal density for z, $q(z)$, times the conditional density $p(y \mid z)$, the joint pdf for the data and the parameter before we see the data is $p(y \mid \beta)\pi(\beta)$, where $\pi(\beta)$ is the prior pdf for the parameter.

A rigorous argument for likelihood-based inference

If we're going to use likelihood to treat an unknown parameter as a random variable after we've seen the data, we should treat it as random before we see the data also. Its distribution before we see the data is called the **prior** distribution.

The model is in the form of a conditional probability distribution for the data x given the parameter β, $p(y \mid \beta)$. Applying the standard rule that the joint probability density of y and z is the marginal density for z, $q(z)$, times the conditional density $p(y \mid z)$, the joint pdf for the data and the parameter before we see the data is $p(y \mid \beta)\pi(\beta)$, where $\pi(\beta)$ is the prior pdf for the parameter.

Bayes' rule

Now we apply two more rules about joint and conditional probability densities:

- The pdf of $y \mid x$ is their joint pdf divided by the marginal pdf of x.
- The marginal pdf of y is the integral over x of the joint pdf of y and x.

Putting these together gives us **Bayes' rule**:

$$r(\beta \mid x) = \frac{p(y \mid \beta)\pi(\beta)}{\int p(y \mid \beta)\pi(\beta)d\beta}.$$

$r(\beta \mid x)$ is called the **posterior** pdf.

But $r(\beta \mid x)$ is not the likelihood!

While this is true in general, in our mortgage-default example the parameter, p might have a uniform prior over $(0, 1)$. If it did, we would have $\pi(\beta) \equiv 1$, in which case $r(p \mid x)$ is indeed the likelihood, scaled to integrate to one.

But $r(\beta \mid x)$ is not the likelihood!

While this is true in general, in our mortgage-default example the parameter, p might have a uniform prior over $(0, 1)$. If it did, we would have $\pi(\beta) \equiv 1$, in which case $r(p \mid x)$ is indeed the likelihood, scaled to integrate to one.

This is generally true: the posterior pdf is the same as the normalized likelihood when $\pi(\beta)$ is constant. It might seem natural to say that a uniform prior over (0,1) in this example is reasonable if before we see the data we "have no idea" what p should be. (It's not so clear that this really is "natural", but let's not worry about that now.)

One way to put this is to say the likelihood is proportional to the posterior pdf when we have a "flat prior", which sometimes is taken to represent "ignorance".

A flat prior on the whole real line?

In our normal mean example, though, $\pi(\mu)$ can't be constant, as it then couldn't integrate to one over the entire real line. Still, it could be that, for example, $\pi(\beta)$ is a normal density with very large variance, centered not too far from $\bar{x}$. Then the sample information might make the likelihood $p(x \mid \mu)$ concentrate over a short interval relative to the standard deviation of the prior, so that over the relevant range the prior pdf is nearly constant. Then the posterior pdf would be nearly the same as the likelihood normalized to integrate to one.

Reporting the likelihood for a diverse audience

A second argument for reporting the likelihood as if it were the pdf is that if readers of your work might have different prior pdf's $\pi(\beta)$, they can all construct their own posterior pdf's $r(\beta \mid x)$ using your reported likelihood. r is just the likelihood times $\pi(\beta)$, normalized to integrate to one.