Continuous Probability Distributions and Moments

9/25/18
A discrete probability distribution $P(x)$ is one in which $x$ takes values from a countable set of possibilities.

Example: For one roll of a fair die, $x \in \{1,2,3,4,5,6\}$, with $P(x) = 1/6$ for all $x$.

A continuous probability distribution $P(x)$ is one in which $x$ takes values from a continuous (uncountable) set of possibilities.

Example: A computer-generated random number $x \in [1,6]$ has, for practical purposes, a continuous $P(x)$. 
Example: Uniform on $[a,b]$

If the probability distribution is uniform (constant) on $x \in [a, b]$ and zero elsewhere, it has the illustrated graph.

Tricky question: If $a = 1$ and $b = 6$, what is the probability that $x$ is exactly 2.5?

Example: Uniform on \([a,b]\)

Tricky question: If \(a = 1\) and \(b = 6\), what is the probability that \(x\) is exactly 2.5?
Answer: 0, because there is an infinite number of exact values of \(x\) in \([1,6]\).

Thus, \(P(x)\) is not a probability, as it would be in the discrete case. It is a probability density.

$P(x)$ as a Probability Density

That $P(x)$ is a probability density means that $P(x) \, dx$ is the probability that the observed result will be in the interval $[x, x+dx]$.  

Example: In quantum mechanics, the probability that a particle will be found in the interval $[x, x+dx]$ is $P(x) \, dx$, where $P(x) = \psi(x) \, \psi^*(x)$.

What are other examples?

Consider

\[ \text{Prob}(x \leq X < x + \delta x | I) \]

in the limit that \( \delta x \to 0 \)

This gives us a density rather than a function. We call this the **probability density function**

\[ \text{pdf}(X = x | I) = \lim_{\delta x \to 0} \frac{\text{Prob}(x \leq X < x + \delta x | I)}{\delta x} \]
We can then integrate the pdf to obtain the probability that $X$ takes a value within a finite range.

\[
\text{Prob}(x_1 \leq X < x_2 | I) = \int_{x_1}^{x_2} \text{pdf}(x | I) \, dx
\]

**NOTE THAT:** \[ \int_{h_1}^{h_2} \text{pdf}(H | I) \, dH = 1 \]

\[ \text{probability} \]
\[ \text{pdf} \times \text{Volume} \]
\[ \text{density} \times \text{Volume} = \text{mass} \]
More Abuses of Notation

I am horribly lazy, and will write

\[ p(n \mid I) \equiv \text{Prob}(n \mid I) = \text{Prob}(N = n \mid I) \]

for \( N \) being a discrete parameter

\[ P(x \mid I) \equiv p(x \mid I) = p(x \mid I) \]

for \( X \) being a continuous parameter
$P(n | I)$ is a probability

$p(x | I)$ is a probability density

To get a probability from a probability density you must multiply times the volume:

$$p(x | I) \, dx \text{ is a probability}$$
Integrating a PDF

\[
\text{Pdf}(X=x | I) = \text{Prob}(x \leq X < x + dx | I)
\]

Consider

\[
\text{Prob}( (x \leq X < x + dx) \cup (x + dx \leq X < x + 2dx) | I)
\]

\[
= \text{Prob}(x \leq X < x + 2dx | I)
\]

\[
= \text{Prob}(x \leq X < x + dx | I) + \text{Prob}(x + dx \leq X < x + 2dx | I)
\]

\[
- \text{Prob}(x \leq X < x + dx) \wedge (x + dx \leq X < x + 2dx)
\]

\[
= \text{Pdf}(X=x | I) + \text{Pdf}(X=x + dx | I)
\]

Extending this ...

\[
\text{Prob}(a \leq X < b | I) = \int_a^b dx \text{Pdf}(X=x | I)
\]
Notation

We will take some liberties with notation to simplify our equations.

- $H_1: \text{"Her pet is a Dog"}$

$\text{Prob}(H_1 | I)$ is the degree to which everything I know implies that "Her pet is a Dog"
• $q = \text{"The number of quarters in my pocket is } q\text{"}

$\text{Prob}(q \mid I) \text{ is the degree to which everything I know}
\text{implies that I have } q \text{ quarters in my pocket, where } q \in \mathbb{I} \text{ st } q \geq 0.$

I might write:

$$\text{Prob}(q \mid I) = \frac{1}{Z} e^{-\frac{q}{\lambda}} \quad \text{where } Z \text{ is a normalization constant}$$

Since $\sum_{q=0}^{\infty} \text{Prob}(q \mid I) = 1$

• $X = \text{"The length of my pen"}$

$x_1 \leq X < x_2 = \text{"The length of my pen is between } x_1 \text{ and } x_2 \text{"}$

$\text{Prob}(x_1 \leq X < x_2 \mid I)$
Expressing PDFs

A probability density function is an infinite number of values, one for each value of \( x \). In addition, very often the function can’t be written in closed form or is not even known.

To express the PDF in useful ways, we often use a finite number of values defined for convenience. *Moments* of the PDF are examples of such values.
The angle brackets mean that you sum (discrete case) or integrate (continuous case) what’s inside them times the probability, $P(x)$ (discrete) or $P(x)dx$ (continuous).
Moments of a prob. dist

The ZEROth moment

\[ \mu_0' = \int p(x) \, dx = 1 \]

The FIRST moment is the MEAN

\[ \mu = \mu_1' = \int x \, p(x) \, dx = \langle x \rangle \]

The SECOND moment

\[ \mu_2' = \int x^2 \, p(x) \, dx = \langle x^2 \rangle \]

Moments of Inertia

The ZEROth moment

\[ \int m(x) \, dx = M \text{ total mass} \]

The FIRST moment

\[ \int x \, m(x) \, dx = x_{cm} \text{ center of mass} \]

The SECOND moment

\[ \int x^2 \, m(x) \, dx = I \text{ moment of inertia} \]
Example of Zeroth Moment

\[ \int_{0}^{\infty} P(x) \, dx = \int_{0}^{a} (0) \, dx + \int_{a}^{b} \frac{1}{b-a} \, dx + \int_{b}^{\infty} (0) \, dx \]

\[ = 0 + \frac{1}{b-a} x \big|_{a}^{b} + 0 \]

\[ = \frac{b-a}{b-a} \]

\[ = 1 \]

Example of Zeroth Moment

The value of $P(x)$ on $[a,b]$ is determined by the requirement that the total probability, also called the *probability mass*, be 1.

\[
\int_0^\infty P(x)dx = \int_0^a (0)dx + \int_a^b \frac{1}{b-a} dx + \int_b^\infty (0)dx = 1
\]
Moments taken about the mean

\[ \langle f(x-\mu) \rangle \]

The second moment about the mean is the variance

\[ \sigma^2 = \langle (x-\mu)^2 \rangle = \int (x-\mu)^2 p(x) \, dx \]

\[ = \int (x^2 - 2\mu x + \mu^2) p(x) \, dx \]

\[ = \int x^2 p(x) \, dx - 2\mu \int x p(x) \, dx + \mu^2 \int p(x) \, dx \]

\[ = \langle x^2 \rangle - 2\mu^2 + \mu^2 \]

\[ = \langle x^2 \rangle - \mu^2 \]

\[ \sigma^2 = \langle x^2 \rangle - \langle x \rangle^2 \]
Precision and Accuracy

*Accuracy* is a measure of how close the measurement is to the (usually unknown) true value of the quantity measured.

*Precision* is a measure of how close repeated measurements are to each other. The standard deviation $\sigma$, or square root of the variance, is related to precision: large $\sigma$ means low precision.
Quantifying Uncertainty

We will revisit accuracy later.

Now we will consider our uncertainty.

Uncertainty is inversely related to the precision of our learning procedure.

\[
\text{LARGE UNCERTAINTY} \leftrightarrow \text{LOW PRECISION}
\]

\[
\text{WIDE DISTRIBUTION}
\]

\[
\text{SMALL UNCERTAINTY} \leftrightarrow \text{HIGH PRECISION}
\]

\[
\text{NARROW DISTRIBUTION}
\]
Example: Gaussian Distribution

is the Gaussian or normal distribution. It is specified by its average $\mu$ and std. dev. $\sigma$.

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

99.7% of the data are within 3 standard deviations of the mean
95% within 2 standard deviations
68% within 1 standard deviation

https://en.wikipedia.org/wiki/Normal_distribution
NEED TO QUANTIFY THE SPREAD OF A PDF

Consider a Gaussian PDF

\[ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \]

\[ f_{\max} = \frac{1}{\sqrt{2\pi \sigma^2}} \]

\[ x_{1/2} = \sigma \sqrt{2 \ln 2} = 1.17\sigma \]

FULL WIDTH at HALF MAXIMUM (FWHM): \( 2x_{1/2} = 2.35\sigma \)

STANDARD DEVIATION (Square Root of Variance): \( \sigma \)

1\( \sigma \) encloses \( \sim 68\% \) of probability mass

2\( \sigma \) encloses \( \sim 95\% \)

3\( \sigma \) encloses \( \sim 99\% \)

ENTROPY \( (-\sum p(x) \log p(x) \text{ or } -\int p(x) \log \frac{p(x)}{m_c(x)} \, dx) \) also measures uncertainty
Quantifying Uncertainty in General

The pdfs we obtain will not, in general, be Gaussian. However, if they are smooth and unimodal, we can make some approximations.

Let \( P = \text{pdf}(x | \text{data}) \).

At the mode \( x_0 \), the first derivative will be zero.

\[
\frac{dP}{dx}\bigg|_{x_0} = 0 \quad \text{and} \quad \frac{d^2P}{dx^2}\bigg|_{x_0} < 0 \quad \text{as it is a peak}
\]
It is often easier to work with the logarithm of the probability since the logarithm has a tendency to smooth out small bumps.

\[ L = \log_e \left[ \text{Pdf} \left( x \mid \text{data}, \theta \right) \right] \]

\[ \text{CAUTION: Sometimes } L \text{ stands for } \log P \text{ and sometimes } L \text{ stands for the likelihood.} \]

At the mode, we have a TAYLOR SERIES EXPANSION

\[ L(x) = L(x_0) + \left. \frac{dL}{dx} \right|_{x_0} (x - x_0) + \frac{1}{2} \left. \frac{d^2L}{dx^2} \right|_{x_0} (x - x_0)^2 + \ldots \]

\[ \frac{dL}{dx} = \frac{d \log P}{dx} = \frac{1}{P} \frac{dP}{dx} \Rightarrow \left. \frac{dL}{dx} \right|_{x_0} = 0 \text{ Log P still has a peak at } x_0! \]
\[ L(x) = L(x_0) + \left. \frac{dL}{dx} \right|_{x_0} (x-x_0) + \frac{1}{2} \left. \frac{d^2L}{dx^2} \right|_{x_0} (x-x_0)^2 + \ldots \]

\[ \frac{dL}{dx} = \frac{d \log P}{dx} = \frac{1}{P} \frac{dP}{dx} \Rightarrow \left. \frac{dL}{dx} \right|_{x_0} = 0 \text{ a peak at } x_0 \]

Exponentiating, we obtain an approximant to \( P \)

\[ \text{Pdf}(x|\text{data},\theta) \approx e^{L(x_0)} \exp \left[ \frac{1}{2} \left. \frac{d^2L}{dx^2} \right|_{x_0} (x-x_0)^2 \right] \]

\[ \approx A \exp \left[ -\frac{1}{2\sigma^2} (x-x_0)^2 \right] \text{ where } -\frac{1}{\sigma^2} = \left. \frac{d^2L}{dx^2} \right|_{x_0} \]

\( \Rightarrow \) near peak (mode), approximate the pdf as Gaussian.
GAUSSIAN INTEGRALS

\[ \int_{-\infty}^{\infty} e^{-ax^2} \, dx = \sqrt{\frac{\pi}{a}} \]

\[ \int_{-\infty}^{\infty} x \, e^{-ax^2} \, dx = 0 \quad \text{for} \quad a \text{ even} \]

\[ \int_{-\infty}^{\infty} x^2 \, e^{-ax^2} \, dx = 0 \quad \text{for} \quad a \text{ odd} \]

\[ \int_{-\infty}^{\infty} e^{-ax^2} \, dx = I \]

TRICK! SWITCH TO POLAR COORDINATES.
\[ \int_{-\infty}^{\infty} x^2 e^{-ax^2} \, dx = \int_{-\infty}^{\infty} (-1) \frac{3}{\sqrt{a}} e^{-ax^2} \, dx \quad \text{TRICK} \]

\[ = - \frac{3}{\sqrt{a}} \int_{-\infty}^{\infty} e^{-ax^2} \, dx \]

\[ = - \frac{3}{\sqrt{a}} \sqrt{\frac{\pi}{a}} \]

\[ = \frac{3}{2} \sqrt{\frac{\pi}{a^3/6}} \]

\[ \int_{-\infty}^{\infty} x^n e^{-ax^2} \, dx = - \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x^{n-1} e^{-ax^2} \, dx \quad \text{N EVEN} \]

\[ \int_{0}^{\infty} e^{-ax^2} \, dx = \frac{1}{\sqrt{\pi}} \]

\[ \int_{0}^{\infty} xe^{-ax^2} \, dx \quad \text{DO BY} \]

\[ \text{SUBSTITUTION} \]

\[ u = ax^2 \quad \Rightarrow \quad \frac{1}{2a} \]

\[ \int_{0}^{\infty} x^{n+2} e^{-ax^2} \, dx = - \frac{1}{2a} \int_{0}^{\infty} x^n e^{-ax^2} \, dx \]
GAUSSIAN IN STANDARD FORM  \(= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\)

-WRITTEN THIS WAY TO SHOW WHAT THE MOMENTS ARE.

0th MOMENT: \(\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \int_{-\infty}^{\infty} e^{-\frac{u^2}{2\sigma^2}} \, du = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2} u^2} \, du = 1\)

GAUSSIAN IS NORMALIZED.
1st moment: \[ \langle x \rangle = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma^2} x e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \]

\[ \langle x \rangle = \frac{1}{\sqrt{2\pi}\sigma^2} \left[ \int_{-\infty}^{0} (u+\mu) e^{-\frac{u^2}{2\sigma^2}} \, du + \int_{0}^{\infty} u e^{-\frac{u^2}{2\sigma^2}} \, du \right] \]

\[ = \frac{1}{\sqrt{2\pi}\sigma^2} \left[ 0 + \mu \sqrt{\frac{\pi}{\sigma^2}} \right] \]

\[ = \mu = \text{MEAN} \checkmark \]

2nd moment about \( \mu = \langle (x-\mu)^2 \rangle \): variance

\[ = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \]

\[ = \frac{1}{\sqrt{2\pi}\sigma^2} \int_{-\infty}^{\infty} u^2 e^{-au^2} \, du \]

\[ u = x-\mu \quad du = dx \quad a = \frac{1}{2\sigma^2} \]

\[ = \frac{1}{\sqrt{2\pi}\sigma^2} \left[ (-1)^{\frac{1}{2}} a \int_{-\infty}^{\infty} e^{-au^2} \, du \right] \]

\[ = \frac{1}{\sqrt{2\pi}\sigma^2} \frac{1}{\sqrt{a}} \sqrt{\pi} = \frac{1}{2} \sqrt{\frac{\pi}{2\pi\sigma^2}} \sigma^{-3/2} \]

\[ = \frac{1}{\sqrt{3\pi\sigma^2}} (2\sigma^2)^{3/2} = \sigma^2 \checkmark \]
Integrating a Gaussian

Consider a Gaussian Pdf

$$\text{Pdf}(x | \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Now

$$\text{Prob}(a \leq x < b | \mu) = \int_{a}^{b} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx$$

These integrals cannot be performed analytically, but can be expressed in terms of the Error function.
These integrals cannot be performed analytically, but can be expressed in terms of the Error function.

\[
\text{erf}^2(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt
\]

Not quite what we want.

Change of variables

Let \( t = \frac{x - \mu}{\sqrt{2\sigma^2}} \) \( \Rightarrow \) \( x = \mu + \sqrt{2\sigma^2} t \)

\[
\text{erf}^2(z) = \frac{2}{\sqrt{\pi}} \frac{1}{\sqrt{2\sigma^2}} \int_{\mu}^{\mu+\sqrt{2\sigma^2}z} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx
\]

\[
= \frac{2}{\sqrt{2\pi\sigma^2}} \int_{\mu}^{\mu+\sqrt{2\sigma^2}z} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx
\]
Integrating a Gaussian cont.

\[\text{Prob} (a \leq x \leq b | \mathcal{I}) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx\]

\[= \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^\mu e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx - \frac{1}{\sqrt{2\pi\sigma^2}} \int_\mu^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx\]

To use our previous results, let \(b = \mu + \sqrt{2\sigma^2} \varepsilon\) \(\Rightarrow \varepsilon = \frac{b - \mu}{\sqrt{2\sigma^2}}\)

\[\frac{1}{2} \erf \left( \frac{b - \mu}{\sqrt{2\sigma^2}} \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_\mu^b e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx\]

Similarly with \(a\).
We get

\[
\frac{1}{\sqrt{2\pi \sigma^2}} \int_{a}^{b} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \frac{1}{2} \left[ \text{erf} \left( \frac{b-\mu}{\sqrt{2\sigma^2}} \right) - \text{erf} \left( \frac{a-\mu}{\sqrt{2\sigma^2}} \right) \right]
\]

To find probability mass within 1σ let \( b = \mu + \sigma \)

\[
\text{Prob} \left( \mu - \sigma \leq x < \mu + \sigma \mid \mathcal{I} \right) = \frac{1}{2} \text{erf} \left( \frac{\sigma}{\sqrt{2\sigma^2}} \right) - \frac{1}{2} \text{erf} \left( \frac{-\sigma}{\sqrt{2\sigma^2}} \right)
\]

\[
= \text{erf} \left( \frac{1}{\sqrt{2\sigma^2}} \right)
\]
Summarizing the Posterior Probability

The result of any Bayesian calculation is the **POSTERIOR PROBABILITY**.

The posterior is the **ANSWER** to the problem. Any other answer is a **SUMMARY** of the posterior.

The pdf is the solution to the problem. It tells us everything we can learn from the calculation.

A commonly encountered posterior pdf.

But how do we summarize our solution?
MODE - Most probable parameter value.

The mode is the highest peak. It is the most probable value. This is often called the maximum a posteriori estimate or MAP estimate for short.

MEAN - This is the weighted average value of the parameter. For a parameter $x$, this is $\bar{x} = \int x \rho(x|d,I) \, dx$

$$\bar{x} = \sum_{i} x_i \rho(x_i|d,I)$$

MEDIAN - This is the parameter value s.t. 50% of the probability mass is below this value and 50% is above this value.
Mean, Median, and Mode in Practice

For a unimodal symmetric distribution, the mode, mean and median have the same value.
DIFFICULTIES

For a skewed distribution, the three summary quantities no longer have the same values.

For multimodal distributions, these quantities can be meaningless.

The mode ignores the probability mass on the left. The mean and median point to relatively improbable values of the parameter.

LESSON: The posterior probability is the solution! All other quantities merely attempt to summarize the solution.