STAT 3503/8109 Lecture 3 Notes

Edoardo Airoldi

*Scribe: Srikar Katta**

Fall 2020: September 14, 2020

1 Introduction

In the last set of notes, we had discussed the translating the language of modeling across disciplines into statistics. However, there is a fair bit of heterogeneity within statistical modelers as well. For example, "probability density function" refers to the likelihood of a certain outcome/set of outcomes in our *continuous* sample space occurring. However, people *use* probability density functions to refer to both continuous and discrete spaces, so being comfortable with the different uses and references of variables will be very beneficial as a researcher. There are many non-card-carrying statisticians in the statistics world, so we should be prepared to collaborate with any of them without being confused by their references to certain topics.

1. Notation/Transformation Theorem/Sufficient Statistics
2. Data Generating Process vs Probabilistic Graphical Models

2 Unpacking

Imagine, we are writing a model for customer expenditure. Although we are tracking expenditures to 4,000,000 customers for two years monthly, the best approach is to find the

*Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

minimum set of quantities that are necessary to write the likelihood. In all reality, this is one statistic - the sufficient statistic - allowing you to write the likelihood for one normally distributed quantity rather than a series of quantities since the sufficient statistic is a combination of the parameters for all individuals. The sufficient statistic is a way for you to simplify the problem into one quantity to work with rather than a set of random variables to have to wrangle through.

A lot of statistics modeling is an exercise in *unpacking*: condensing your complicated problem statement into a much more palatable one.

Example 2.1. *What is the* $\mathbb{E}[y]$ *given*

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1), i = 1 \dots n$$
$$y \equiv \frac{1}{n} \sum_{i=1}^{n} x_i,$$
$$\mathbb{P}(y|\theta) \textit{ is a function of } \mathbb{P}(x_1|\theta) \dots \mathbb{P}(x_n|\theta) \textit{ for } y \in Y?$$

Notice that there are many ways we can formulate or represent $\mathbb{E}[y]$:

- $\mathbb{E}[y] = \mathbb{E}[y|\theta]$
- $\mathbb{E}[y] = \mathbb{E}[\frac{1}{n} \sum x_i|\theta]$
- $\mathbb{E}[y] = \mathbb{E}_{\mathbb{P}(y|\theta)}[y] = \mathbb{E}_{\mathbb{P}(y|\theta)}[y|\theta]$

We can also write our expectation as multivariate integrals (because x_i is IID):

$$\begin{aligned}
\mathbb{E}[y] &= \mathbb{E}[\frac{1}{n} \sum x_i|\theta] \\
&= \mathbb{E}_{\mathbb{P}(x_1 \dots x_n|\theta)}[\frac{1}{n} \sum x_i|\theta] \\
&= \mathbb{E}_{\prod \mathbb{P}(x_i|\theta)}[\frac{1}{n} \sum x_i|\theta] \\
&= \int_{x_1} \dots \int_{x_n} \frac{1}{n} \sum x_i \prod \mathbb{P}(x_i|\theta) dx_1 \dots dx_n
\end{aligned}$$

However, because $\mathbb{P}(y|\theta)$ is a function of $\mathbb{P}(x|\theta)$, we can also rewrite our expectation as *one integral* as follows instead:

$$\mathbb{E}[y|\theta] = \int_y y\mathbb{P}(y|\theta) dy \text{ note: } y \text{ is the support of this function.}$$

This resulting integral is a much simpler version of the multivariate integral we had before, which will allow you to save time and effort. The transformation theorem allows you to simplify your original problem into a much more manageable one. However, we do run into a slight hiccup: we don't explicitly know what $\mathbb{P}(y|\theta)$ is.

If $x_i \overset{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$, then $\frac{1}{n}\sum x_i$ is $\mathcal{N}(\theta, \frac{\sigma^2}{n}$, which is exactly what you would know if you understood the transformation theorem, so there is no need to compute the multivariate integral. So,

$$\begin{aligned}\mathbb{E}[y|\theta] &= \int_y y\mathbb{P}(y|\theta)dy \\ &= \int_y y\mathcal{N}(\theta, \frac{\sigma^2}{n})dy \\ &= \theta.\end{aligned}$$

In this particular case, there is an even *simpler* solution, so we can use the properties of the expectation operator to find the optimal answer. Since the expectation is linear,

$$\begin{aligned}\mathbb{E}[\frac{1}{n}\sum x_i|\theta] &= \frac{1}{n}\sum \mathbb{E}[x_i\theta] \\ &= \sum_i \frac{\theta}{n} \\ &= \frac{n\theta}{n} \\ &= \theta.\end{aligned}$$

Another way to approach this is to say,

$$\begin{aligned}\mathbb{E}_{\mathbb{P}(x_i|\theta)}[x_i|\theta] &= \int_{-\infty}^{\infty} x_i\mathcal{N}(\theta, 1)dx_i \\ &= \theta.\end{aligned}$$

2.1 Terminologies

In Statistics 101, we learn that if some random variable X is discrete, it has a probability mass function (the distribution of values in our sample). So, for example, the sample space of a fair die would be $\Omega = \{1, 2, 3, 4, 5, 6\}$, and each outcome has probability $\frac{1}{2}$. Now, if X is continuous, then it has a probability *density* function, written as a function $f : (a, b) \subseteq \mathbb{R} \to$

$[0, 1]$. However, in statistics notation, people do not differentiate between the continuity or discreteness of X by specifying $\mathbb{P}[X]$ rather than $f(X)$. We have to tolerate that. So, we have to know the **support** of X: the possible values that X can adopt. This will allow us to understand whether we are interested in the mass or density of X. This can also help us understand what *distribution* we should place on X.

For example, consider X is a random variable where $X \in \{1, 2, 3, 4, 5, 6\}$. The support are these discrete values. If X is normal, then the support of X is $\mathbb{R}$, the real numbers. If X has a Poisson distribution, then $X \in \mathbb{N}^{\geq 0}$, the natural numbers greater than or equal to 0. If we instead said that the support of X in $\mathbb{N}^{\geq 0}$, then we can figure out what distribution X might take on. It could be $X \sim exp(\theta)$, $X \sim Gamma(\alpha, \beta)$, etc.

Take the following situation: X is time. What is the support of X? But the distribution we assign to X depends on the support of X. Even if we specified that X is age measured in years, we have possibilities that could work:

- $X \sim Poisson(\mu) \implies x \in \{0, 1, 2, 3, \ldots, \infty\}$
- $X \sim Binomial(N, \theta) \implies x \in \{0, 1, 2, 3, \ldots, N\}$

Obviously, the second distribution is better because it is impossible to have an age of ∞. But even then, I have a problem: what should my N be? If I place a cap of 122, the age of the oldest person to live (wiki), there could still be a lot of missing entries for our older ages. If we set N to 80, then it is possible that we might have to throw some data points out. Another concern: what about differentiating between 18 years and 18.5 years?

In summary, the model statement carries a lot of information. We first need to unpack all of this information into something more palatable. And then we can figure out these three ideas to begin modeling our data:

- which quantities are relevant
- what possible values these quantities can be (support)
- what distributions we place on our quantities

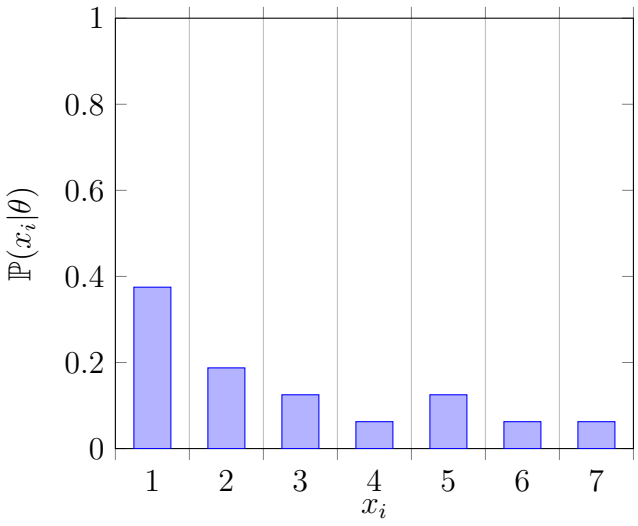
3 Data Generating Process (DGP) vs Graphical Model Representation

Examples will motivate most of the content for this section using the following problem statement: we want to create a model that sees, based on 43 students, the number of classes each student enrolls in.

Example 3.1.

$$x_i \overset{\text{iid}}{\sim} Binomial(7, \theta), i = 1 \dots 43$$

From this model statement, we know the *support* of x_i is $0, 1, 2, 3, 4, 5, 6, 7$. And since x_i is discrete, we can create a probability mass function. For example, it might look like this:



Notice, there is nothing greater than 7 and less than 1. Additionally, sum of the probability of outcome will be 1. Let us say that we observe $x_1, \dots, x_{43}$ but not θ. This information will allow us to build a 2x2 table:

	Observed	**Unobserved**
Variable	$x_1 \dots x_{43}$	NA
Constant	NA	θ

Using this, we can identify the likelihood (in this case, the complete and proper likelihood

are both the same):

$$\begin{aligned}
\textit{likelihood} &= \mathbb{P}(x_1 \dots x_{43}|\theta) \\
&= \prod_{i=1}^{43} \textit{Binomial}(x_i, \theta) \text{ (take product because } x_i \text{ is IID)} \\
&= \prod_{i=1}^{43} \mathbb{P}(x_i\theta) \\
&= \prod_{i=1}^{43} \binom{7}{x_i} \theta^{x_i}(1-\theta)^{1-x_i} \\
&= \left(\prod_{i=1}^{43} \binom{7}{x_i}\right) \left(\prod_{i=1}^{43} \theta^{x_i}(1-\theta)^{1-x_i}\right) \\
&= \left(\prod_{i=1}^{43} \binom{7}{x_i}\right) \left(\theta^{\sum_{i=1}^{43} x_i}(1-\theta)^{43-\sum_{i=1}^{43} x_i}\right).
\end{aligned}$$

Note: Usually, when we have a "loop" ($i = 1 \dots n$), that generally indicates the use of a product somehow. So, in general, we first create the 2x2 table, then we identify the probability distributions for our variables, and then we calculate the likelihood.

Example 3.2. *Now consider the same problem statement but with the following information:*

$$z_i = \begin{cases} 0 & \textit{if } i \textit{ is undergrad} \\ 1 & \textit{if } i \textit{ is graduate} \end{cases}$$

$$x_i|z_i \sim \begin{cases} \textit{Binomial}(7, \theta) & z_i = 0 \\ \textit{Binomial}(3, \theta) & z_i = 1 \end{cases} \textit{for } i = 1 \dots 43$$

Here, we observe both $x_1 \dots x_{43}$ and $z_1 \dots z_{43}$, but not θ_0 or θ_1.

Let's first create our 2x2 table. Since $x_1 \dots x_{43}$ are observed and have *theoretical distributions*, we classify it as an observed variable. Since $z_1 \dots x_{43}$ are observed but have no theoretical distribution, they are observed constants. Lastly, since θ_0, θ_1 are not observed and have no theoretical variation, we classify them as unknown constants:

	Observed	**Unobserved**
Variable	$x_1 \dots x_{43}$	NA
Constant	$z_1 \dots z_{43}$	θ_1, θ_2

Note that some may include 7 and 3 as known constants, but it is tacitly implied. The likelihood calculations would remain the same though. So, the likelihood would read as follows (proper and complete likelihood are still the same:

$$\begin{aligned} likelihood &= \mathbb{P}(x_1 \dots x_{43} | z_1 \dots z_{43}, \theta_0, \theta_1) \text{because of our IID assumption} \\ &= \mathbb{P}(x_1 | z_1, \theta_0, \theta_1) * \mathbb{P}(x_2 | z_2, \theta_0, \theta_1) * \dots * \mathbb{P}(x_{43} | z_{43}, \theta_0, \theta_1) \\ &= \prod_{i=1}^{43} \mathbb{P}(x_i | z_i, \theta_0, \theta_1) \\ &= \prod_{i=1}^{43} \left(Binomial(x_i | \theta_0)^{1-z_i} Binomial(x_i | \theta_1)^{z_i} \right). \end{aligned}$$

Let's break this down. We call z_i an indicator since it denotes what type something is. Note that when $z_i = 0$, our student is an undergraduate, so their distribution would be different than if $z_i = 1$, who would be a graduate. Taking advantage of z_i, we can mathematically indicate which probability distribution to use: When $z_i = 1$, our inner product reads as follows:

$$\begin{aligned} Binomial(x_i | \theta_0)^{1-z_i} Binomial(x_i | \theta_1)^{z_i} &= Binomial(x_i | \theta_0)^{1-1} Binomial(x_i | \theta_1)^1 \\ &= Binomial(x_i | \theta_0)^0 Binomial(x_i | \theta_1)^1 \\ &= 1 * Binomial(x_i | \theta_1)^1 \\ &= Binomial(x_i | \theta_1), \end{aligned}$$

which is identical to our distribution for x_i for graduate students. We can find the similar result when $z_i = 0$ for undergrads, just replace z_i with 0 and you see that

$$Binomial(x_i | \theta_0)^{1-z_i} Binomial(x_i | \theta_1)^{z_i} = Binomial(x_i | \theta_0)$$

Example 3.3. *Now take this new model statement (slight twist):*

$$\begin{aligned} z_i &\sim Bernoulli(p) \text{ so } z_i = 0 \text{ or } 1 \\ x_i | z_i &\sim \begin{cases} Binomial(7, \theta_0) & z_i = 0 \\ Binomial(3, \theta_1) & z_i = 1, \end{cases} \end{aligned}$$

where we observe $x_1 \dots x_{43}, z_1 \dots z_{43}$ *and don't observe* θ_0, θ_1.

Let's first create our 2x2 table:

	Observed	**Unobserved**
Variable	$x_1 \dots x_{43}, z_1 \dots x_{43}$	NA
Constant	7,3,43	p, θ_1, θ_0

Now, we can find the likelihood:

$$\begin{aligned} likelihood &= \prod_{i=1}^{43} \mathbb{P}(x_i, z_i | p, \theta_0, \theta_1) \\ &= \prod_{i=1}^{43} (\mathbb{P}(z_i|p)\mathbb{P}(x_i|z_i, \theta_0, \theta_1)) \\ &= \prod_{i=1}^{43} \left[\mathbb{P}(z_i = 0|p) Binomial(x_i|7, \theta_0)\right]^{1-z_i} \left[\mathbb{P}(z_i = 1|p) Binomial(x_i, 3, \theta_1)\right]^{z_i}. \end{aligned}$$

Let's do some cases to make sure these make sense. What is the situation when $Z = \{z_1 = 0, z_2 = 1, \dots\}$?

$$\begin{aligned} likelihood^{z_1} &= \left[\mathbb{P}(z_1 = 0|p) Binomial(x_1|7, \theta_0)\right]^{1-z_1} \left[\mathbb{P}(z_1 = 1|p) Binomial(x_1, 3, \theta_1)\right]^{z_1} \\ &= \left[\mathbb{P}(z_1 = 0|p) Binomial(x_1|7, \theta_0)\right]^{1-0} \left[\mathbb{P}(z_1 = 1|p) Binomial(x_1, 3, \theta_1)\right]^{0} \\ &= \left[\mathbb{P}(z_1 = 0|p) Binomial(x_1|7, \theta_0)\right]^{1} \\ &= (1-p)\binom{7}{x_1}\theta_0. \end{aligned}$$

Now, for $z_2 = 1$,

$$\begin{aligned} likelihood^{z_2} &= \left[\mathbb{P}(z_2 = 0|p) Binomial(x_2|7, \theta_0)\right]^{1-z_2} \left[\mathbb{P}(z_2 = 1|p) Binomial(x_2, 3, \theta_1)\right]^{z_i} \\ &= \left[\mathbb{P}(z_2 = 0|p) Binomial(x_2|7, \theta_0)\right]^{1-1} \left[\mathbb{P}(z_2 = 1|p) Binomial(x_2, 3, \theta_1)\right]^{1} \\ &= \left[\mathbb{P}(z_2 = 1|p) Binomial(x_2, 3, \theta_1)\right] \\ &= p\binom{3}{x_2}\theta_1^1(1-\theta_1)^0 \\ &= p\binom{3}{x_2}\theta_1. \end{aligned}$$

Then, the total likelihood would read as follows:

$$likelihood = \mathbb{P}(x_1 \dots x_{43}, z_1 \dots z_{43} | p, \theta_0, \theta_1)$$

$$
\begin{aligned}
&= \prod_{i=1}^{43} \left[p^{z_i}(1-p)^{1-z_i} \binom{7}{x_i} \theta_0^{x_i}(1-\theta_0)^{7-x_i} \right]^{1-z_i} \left[p^{z_i}(1-p)^{1-z_i} \binom{3}{x_i} \theta^{x_i}(1-\theta_1)^{3-x_i} \right]^{z_i} \\
&= \prod_{i=1}^{43} \left[(1-p) \binom{7}{x_i} \theta_0^{x_i}(1-\theta_0)^{7-x_i} \right]^{1-z_i} \left[p \binom{3}{x_i} \theta^{x_i}(1-\theta_1)^{3-x_i} \right]^{z_i} \\
&\implies f(p, \theta_0, \theta_1 | x_1 \dots x_{43}, z_1 \dots z_{43}),
\end{aligned}
$$

so now we have rewritten our likelihood so that it is a function of three unknown variables given observed data from a mathematical perspective. But from a statistical perspective, we use likelihood to see how probable the set of observations is for any set of parameters.

Example 3.4. *Now, consider the same exact problem statement before with one slight variation:*

$$
\begin{aligned}
z_i &\sim Bernoulli(p) \\
x_i | z_i &\sim \begin{cases} Binomial(7, \theta_0) & z_i = 0 \\ Binomial(3, \theta_1) & z_i = 1, \end{cases}
\end{aligned}
$$

where $x_1 \dots x_{43}$ is observed and $p, \theta_0, \theta_1, z_1 \dots z_{43}$ are not observed.

First, let's create the 2x2 table:

	Observed	**Unobserved**
Variable	$x_1 \dots x_{43}$	$z_1 \dots z_{43}$
Constant	43, 7, 3	p, θ_0, θ_1

So, our proper likelihood is now *finally* different from our complete likelihood.

$$
\begin{aligned}
likelihood(proper) &= \mathbb{P}(x_1 \dots x_{43} | p, \theta_0, \theta_1) \\
&= \int_{z_1 \dots z_{43}} \mathbb{P}(x_1 \dots x_{43}, z_1 \dots z_{43} | p, \theta_0, \theta_1) dz_1 \dots dz_{43}.
\end{aligned}
$$

So, because x_i and z_i are random variables, we can write these as marginal probability of the observed random variables expressed as the joint probability of observed random variables with respect to latent variables, while integrating latent variables out. In other words, our proper likelihood is the complete likelihood with the latent variables *integrated out*. Then, our likelihood would be

$$
likelihood = \mathbb{E}_{\mathbb{P}(z_1 \dots z_{43})} \left[\mathbb{P}(x_1 \dots x_{43}, z_1 \dots z_{43} | p, \theta_0, \theta_1) \right].
$$

We are simply marginalizing over one side. Let's return back to our integral though, breaking it down into its smaller components and then building back into our likelihood calculations. On the inside, we have

$$\begin{aligned}\mathbb{P}(x_1 \dots x_{43}, z_1 \dots z_{43}|p, \theta_0, \theta_1) &= \prod_{i=1}^{43} \mathbb{P}(x_i, z_i|p, \theta_0, \theta_1) \\ &= \prod_{i=1}^{43} \left[\mathbb{P}(z_i = 0|p) Binomial(x_i|7, \theta_0)\right]^{1-z_i} \left[p(z_i = 1|p) Binomial(x_i|3, \theta_1)\right]^{z_i}.\end{aligned}$$

Now, let's return this back to our integral:

$$\begin{aligned}likelihood(proper) &= \mathbb{P}(x_1 \dots x_{43}|p, \theta_0, \theta_1) \\ &= \int_{z_1 \dots z_{43}} \mathbb{P}(x_1 \dots x_{43}, z_1 \dots z_{43}|p, \theta_0, \theta_1) dz_1 \dots dz_{43} \\ &= \prod_{i=1}^{43} \int_{z_i} \mathbb{P}(x_1 \dots x_{43}, z_1 \dots z_{43}|p, \theta_0, \theta_1) dz_i \ ([1]) \\ &= \prod_{i=1}^{43} \sum_{z_i=0,1} \left[\mathbb{P}(z_i = 0|p) Binomial(x_i|7, \theta_0)\right]^{1-z_i} \left[p(z_i = 1|p) Binomial(x_i|3, \theta_1)\right]^{z_i} \ ([2]) \\ &= \prod_{i=1}^{43} \left[p * Binomial(x_i|\theta_1, 3)\right] + \left[(1-p) Binomial(x_i|\theta_0, 7)\right] ([3]).\end{aligned}$$

This is now the likelihood proper. We can be confident of this because we do not see any latent random variables, which in this case was $z_i \dots z_{43}$. We integrated them all out. Remember, the likelihood is *statistically* a function of all the x_i (observed variables) given p, θ_0, θ_1 (constants). But *mathematically* the likelihood is a function of the unknowns given our variable observations.

4 Conclusion

We discussed notation and different methods of representing our model statement to make our lives the easiest they could be. We then extended our discussion of variable classification for our 2x2 table by also discussing likelihood. We will many times find ourselves in siutations

[1] The integral of product is equivalent to the product of the integral if our rvs are independent

[2] We changed our likelihood calculation from an integral to a sum because we are working with discrete variables

[3] We simply went from summation notation to evaluating at $z_i = 0$ and $z_i = 1$

with latent random variables. How do we go from here to the proper likelihood, which is a function of *observed* random variables, given constants? Well, we can integrate out these latent variables by taking advantage of joint probability distributions and marginalizing the observed random variables. In other words, we can integrate the complete likelihood, do some math and computation, and then get the proper likelihood, which is what we are looking for. In these remarks, we also included observations about the data generating process and how modeling that will help us make predictions and inference more succinctly and clearly. So in the rest of the first part of the course, we will discuss a series of models and DGP formulations with observed and latent variables, and then we will see how to compute the proper likelihood from the complete likelihood for several different models.

5 Appendix

5.1 Marginal Probabilities

Let us illustrate the working of marginal probabilities via an example (adapted from Bruce Hansen's Econometrics):

Let X and Y have the joint density $f(x,y) = \frac{3}{2}(x^2+y^2)$ on $0 \leq x \leq 1, 0 \leq y \leq 1$. Let's find the marginal probability, $f(x)$. We know that the marginal probability distribution can be rewritten as the following:

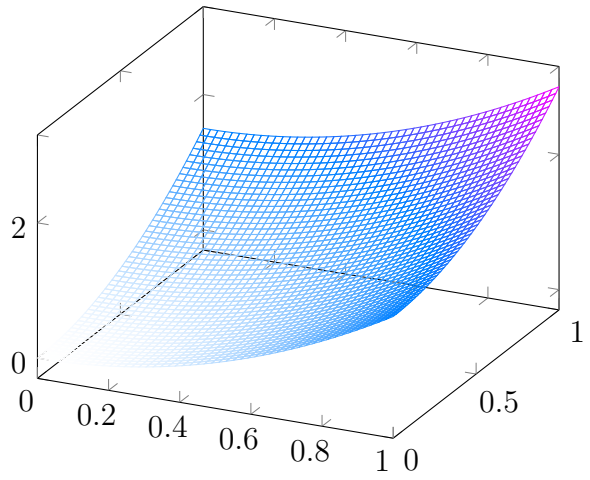
$$f(x) = \int_a^b f(x,y)dy,$$

where a, b are our lower and upper bounds for y. In other words, the continuous range (a,b) is the *support* of y. We can simply apply this to our formula and "do the math:"

$$\begin{aligned} f(x) &= \int_0^1 \frac{3}{2}(x^2+y^2)dy \text{ since } 0 \leq y \leq 1 \\ &= \frac{3}{2}(x^2y + \frac{1}{2}y^3) \mid_0^1 \\ &= \frac{3}{2}(x^2 + \frac{1}{2}) \\ &= \frac{3x^2+1}{2}. \end{aligned}$$

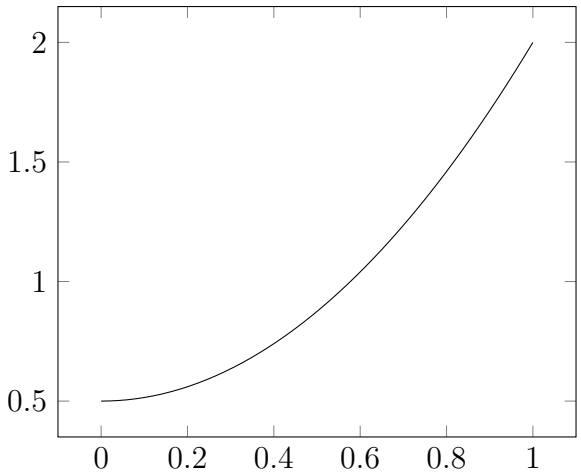
So, the marginal probability density function for x is $f(x) = \frac{3x^2+1}{2}$. Let's visualize this:

Joint Probability Distribution: $\frac{3}{2}(x^2 + y^2)$



Now, when we integrate out the latent random variable, we can "project" this joint probability onto one variable space into a marginal probability. So, if we integrate out y from the probability distribution, it would look something like this:

Marginalized Joint Probability Distribution: $\frac{3x^2+1}{2}$



By integrating over y, we have now "flattened" our distribution to one variable, the probability of x marginally.