Introduction to Bayesian methods in macromolecular crystallography

Tom Terwilliger
Los Alamos National Laboratory

Why use a Bayesian approach?

- **We often know how are measurements are related to our model...**
- **The Bayesian approach gives us the probability of our model once we have made a measurement**
- **It is useful for dealing with cases where there are errors (uncertainties) in the model specification (missing parts of model)**

Introduction to Bayesian methods in macromolecular crystallography

Basics of the Bayesian approach

- Working with probability distributions
- Prior probability distributions
- How do we go from distributions to the value of “x”?
- Bayesian view of making measurements
- Example: from “400 counts” to a probability distribution for the rate
- Bayes' rule
- Applying Bayes' rule
- Visualizing Bayes' rule

Marginalization: Nuisance variables and models for errors

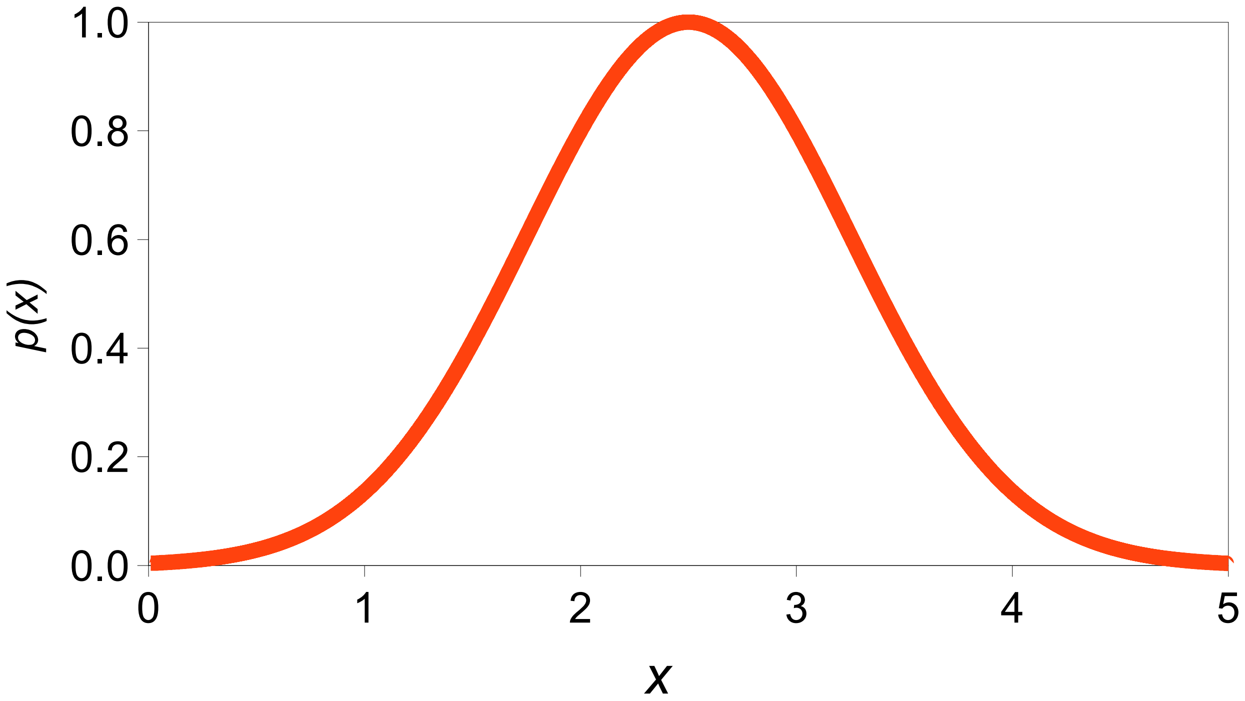
- How marginalization works
- Repeated measurements with systematic error

Applying the Bayesian approach to any measurement problem

Basics of the Bayesian approach
Working with probability distributions

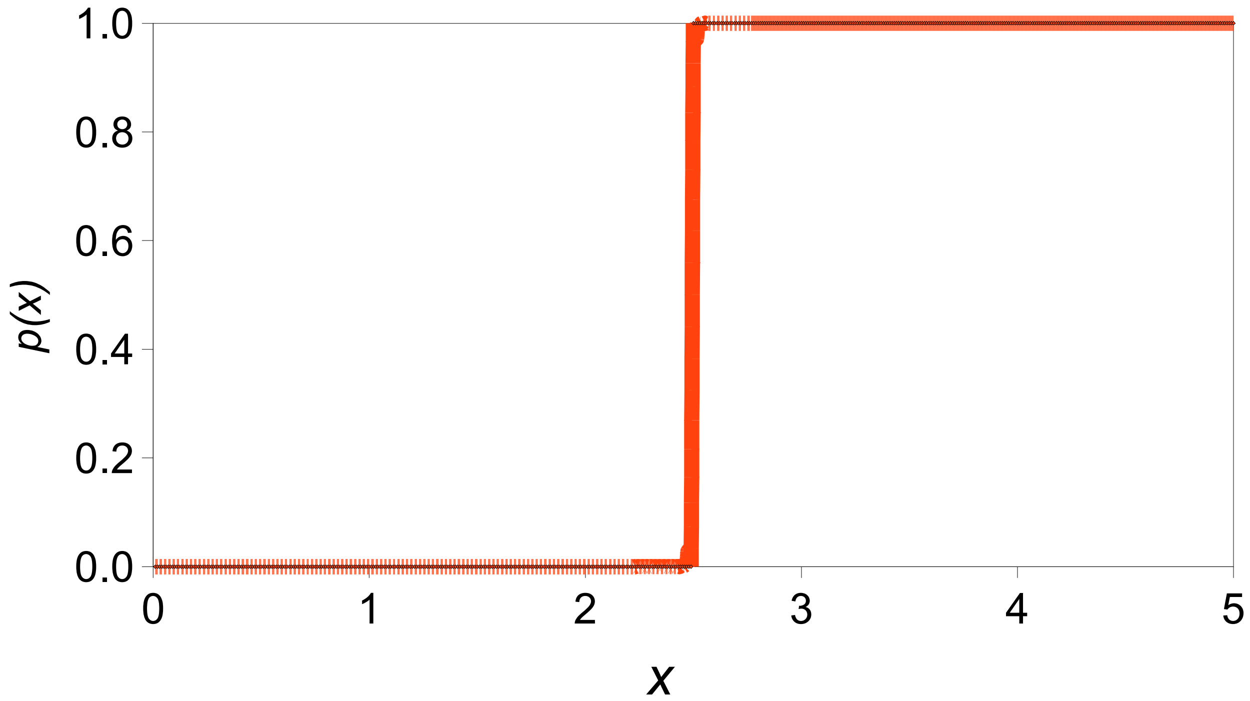
Representing what we know about x as a probability distribution

p(x) tells us the relative probability of different values of x



p(x) does not tell us what x is...
...just the relative probability of each value of x

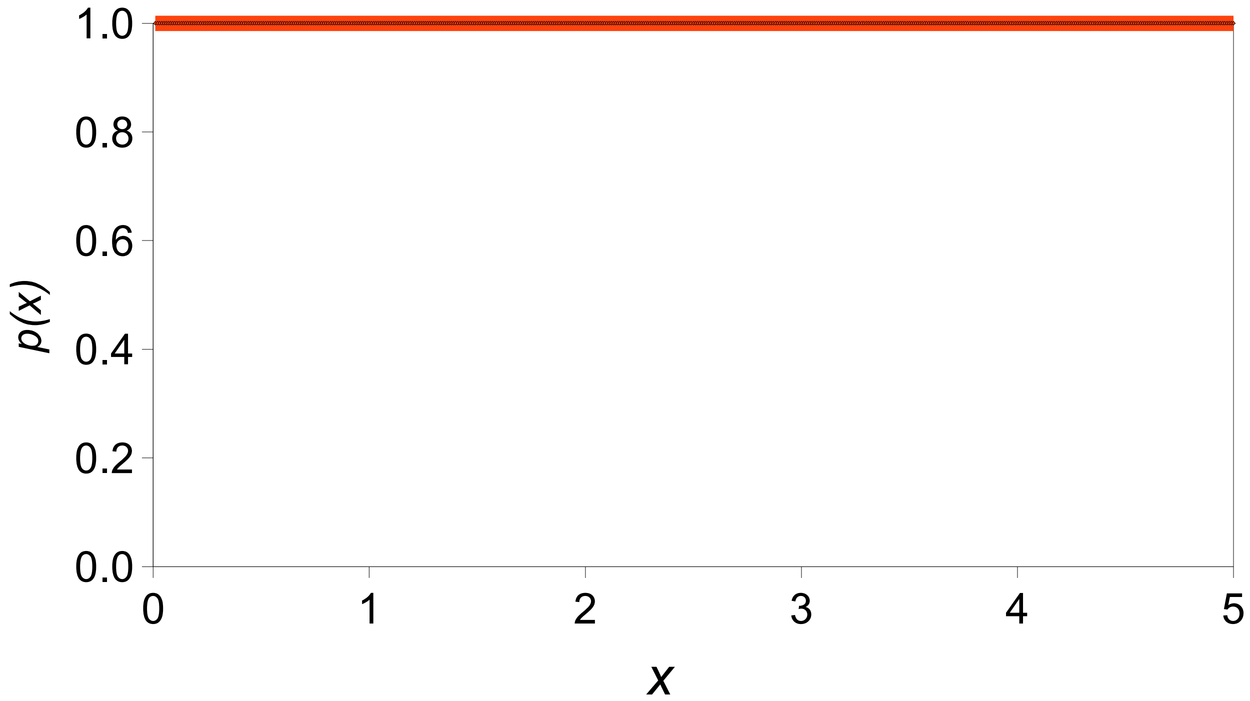
Prior probability distributions
What we know before making measurements



I am sure *x* is at least 2.5

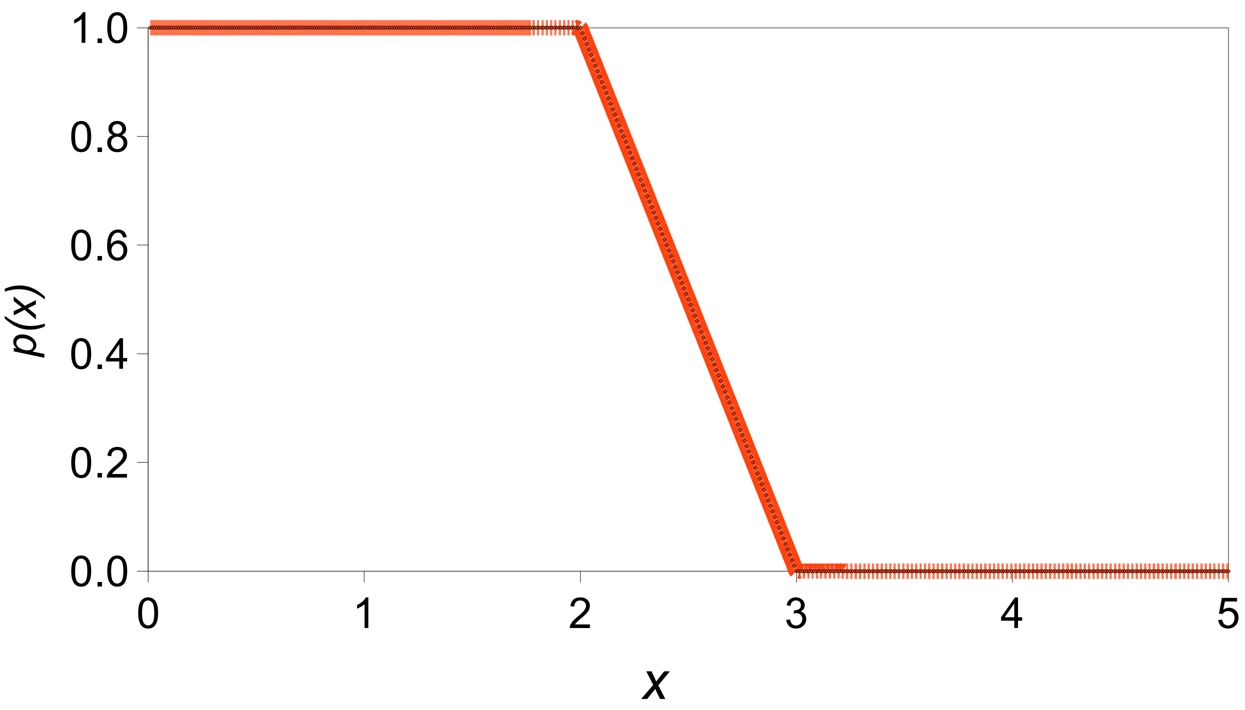
Prior probability distributions

What we know before making measurements



All values of *x* are equally probable

Prior probability distributions
What we know before making measurements

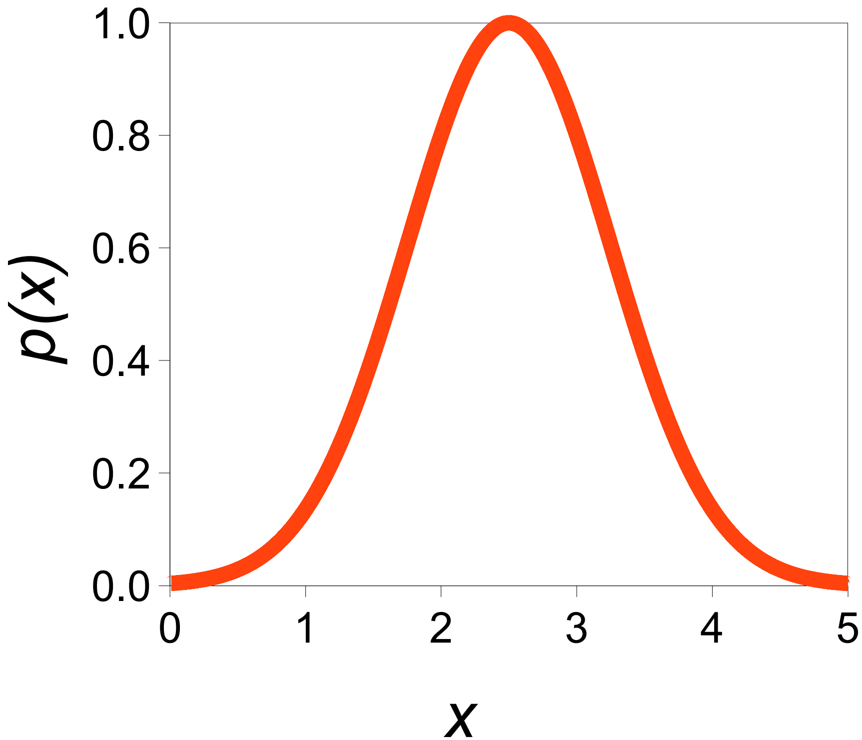


x is less than about 2 or 3

Working with probability distributions

What is the "value" of *x* ?

$\langle x \rangle$



We don't know exactly what "x" is...

but we can calculate a weighted estimate:

$$\langle x \rangle = A \int x \, p(x) \, dx$$

Weight each value of x

by its relative probability p(x)

$$A = 1 / \int p(x) \, dx$$

A is normalization factor

A Bayesian view of making measurements

A crystal is in diffracting position for a reflection
The beam and crystal are stable...

We measure 400 photons hitting the corresponding pixels in our detector in 1 second

What is the probability that the rate of photons hitting these pixels is actually less than 385 photons/sec?

A Bayesian view of making measurements

A crystal is in diffracting position for a reflection
The beam and crystal are stable...

We measure 400 photons hitting the corresponding pixels in our detector in 1 second : $N_{obs} = 400$

A good guess for the actual rate k of photons hitting these pixels is 400:

$k \sim 400$

What is the probability that k is actually < 385 photons/sec?

$$\text{What is } p(\, k<385 \mid N_{obs} = 400)$$

A Bayesian view of making measurements

Start with prior knowledge about which values of k are probable: $p_o(k)$

Make measurement N_{obs}

For each possible value of parameter k (385...400...)

> ***Calculate probability of observing N_{obs} if k were correct: $p(N_{obs} \mid k)$***

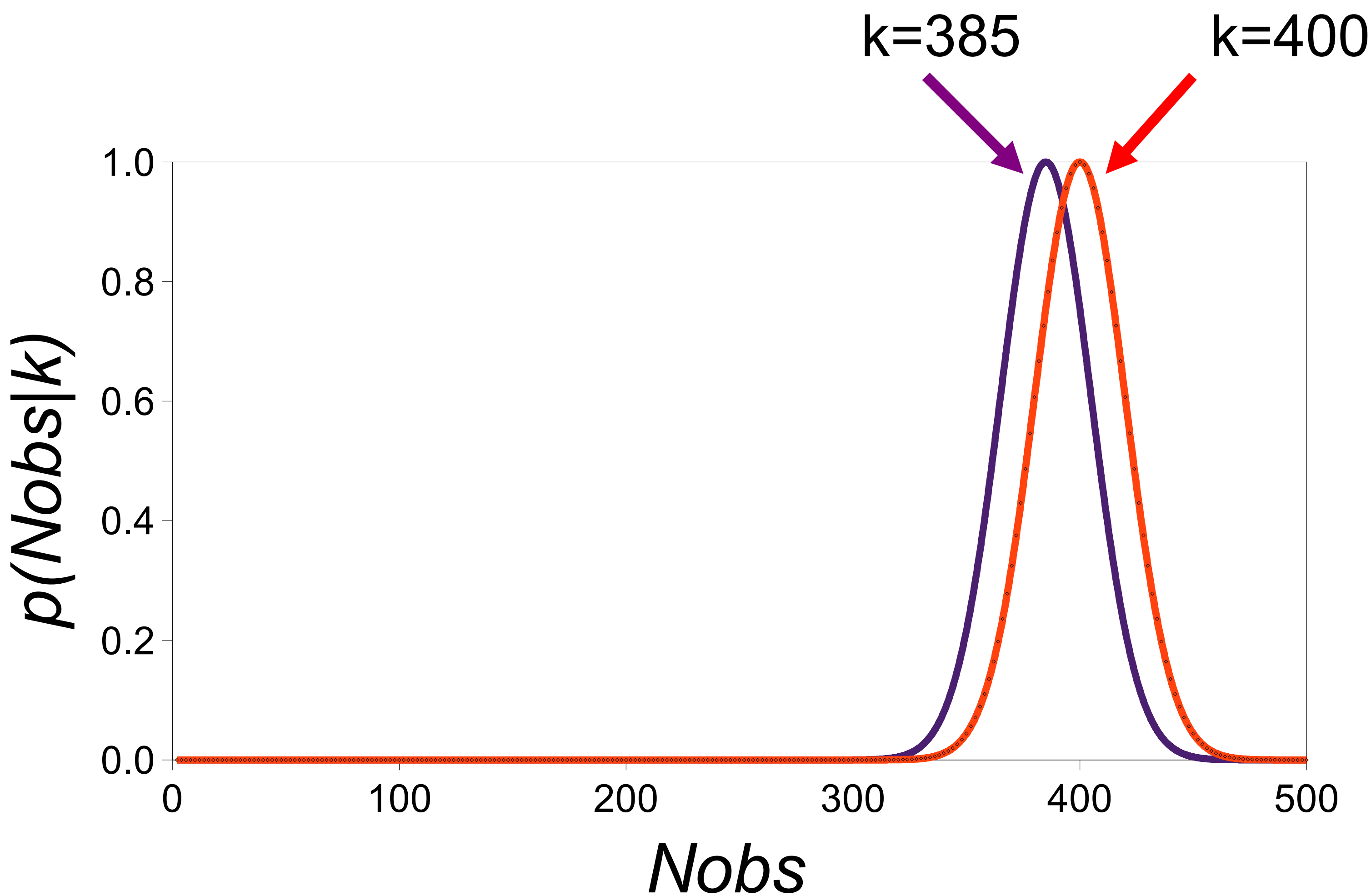
Use Bayes' rule to get p(k) from $p_o(k)$, N_{obs} and $p(N_{obs}|k)$:

$$p(k) \propto p_o(k)\, p(N_{obs}|k)$$

A Bayesian view of making measurements

What is the probability that we would measure N_{obs} counts if the true rate were k?

$$p\left(N_{obs} \middle| k\right)$$



Bayes' rule

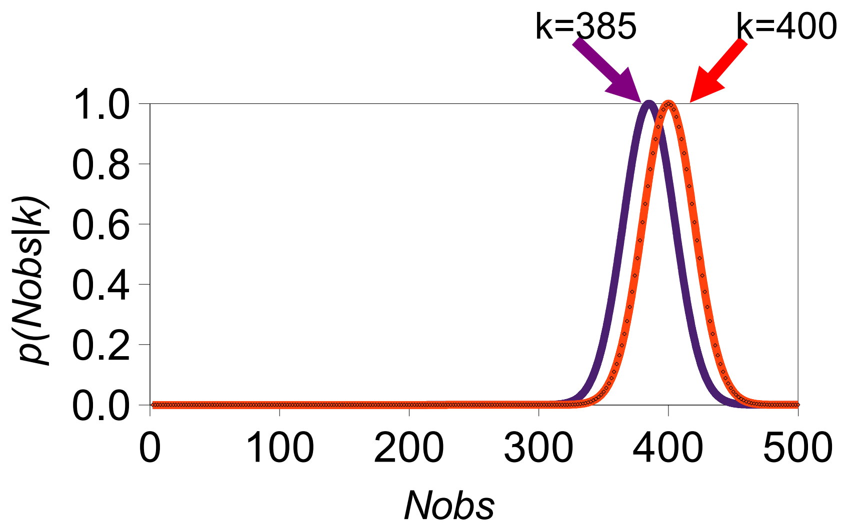
$$p(k) \propto p_o(k)\, p(N_{obs}|k)$$

The probability that k is correct is proportional to...

the probability of k from our prior knowledge

multiplied by...

the probability that we would measure N_{obs} counts if the true rate were k



Bayes' rule

$$p(k) \propto p_o(k)\, p(N_{obs}|k)$$

Prior

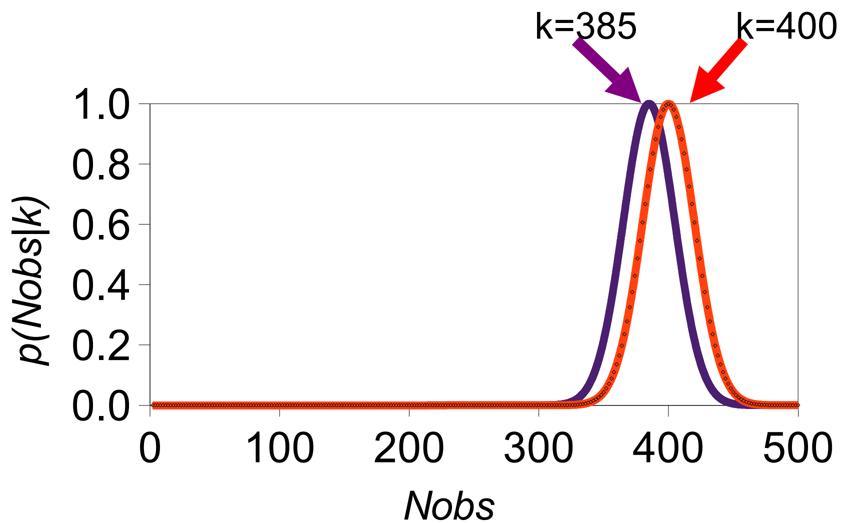
Likelihood

The probability that k is correct is proportional to...

the probability of k from our prior knowledge (prior)

multiplied by...

***the probability that we would measure N_{obs} counts if the true rate were k* (likelihood)**



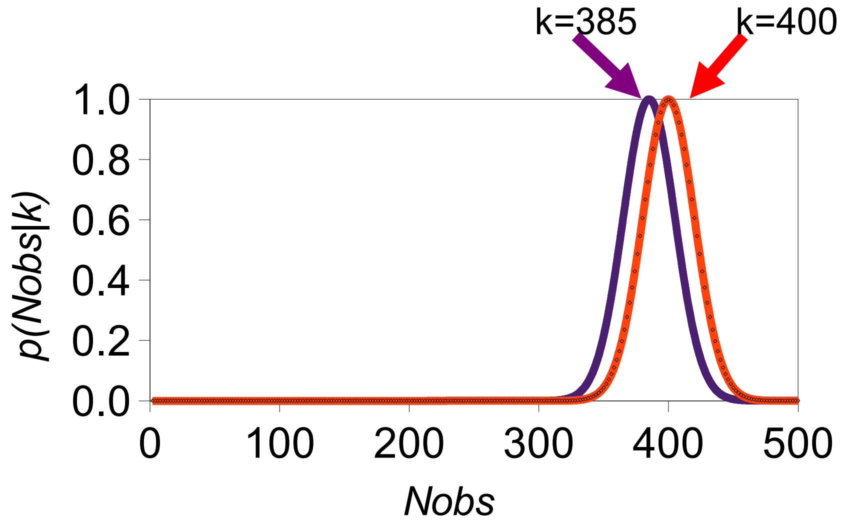
Application of Bayes' rule

$$p(k) \propto p_o(k)\, p(N_{obs}|k)$$

No prior knowledge: $p_o(k) = 1$

Poisson dist. for N_{obs} (large k)

$$p(N_{obs}|k) \propto e^{-[N_{obs}-k]^2/(2\mathrm{k})}$$



Application of Bayes' rule

Probability distribution for k given our measurement N_{obs} =400:

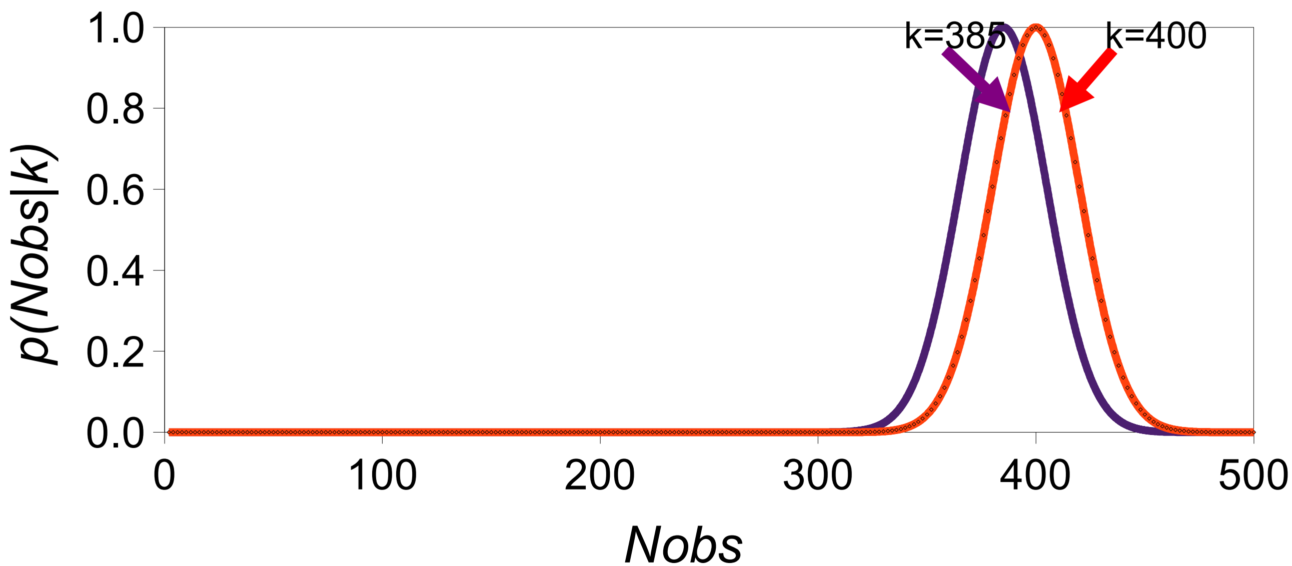
$$p(k) \propto e^{-[N_{obs}-k]^2/(2\mathrm{k})}$$

Probability that k < 385:

$$p(k<385) = A\int_{-\infty}^{385} p(k)\,dk$$

$$A = 1/\int_{-\infty}^{\infty} p(k)\,dk$$

p = 22%



Visualizing Bayes' rule

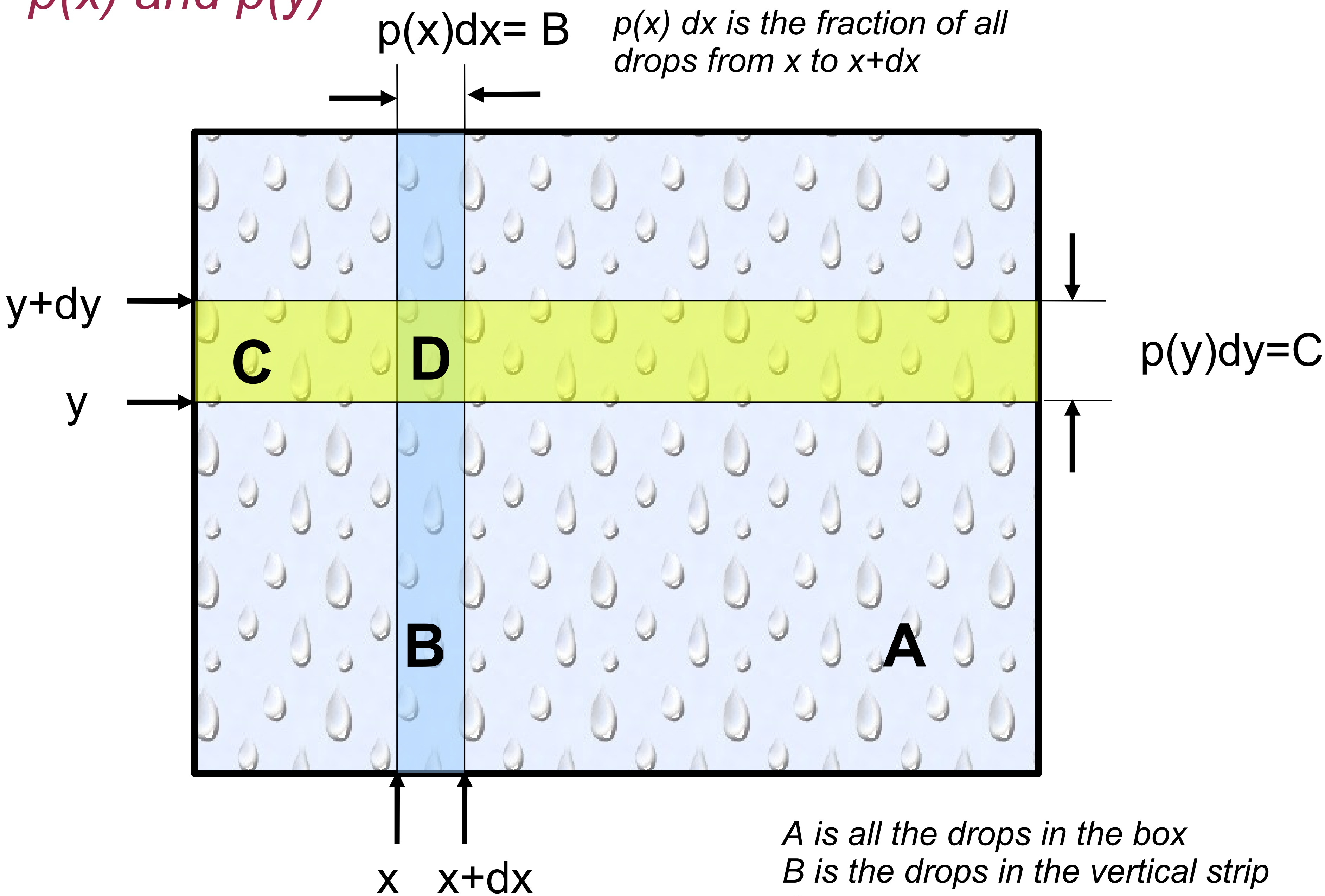
$$p(x|y_{obs}) \propto p_o(x)\, p(y_{obs}|x)$$

Where does Bayes' rule come from?

Using a graphical view to show how p(x|y) is related to p(y|x)

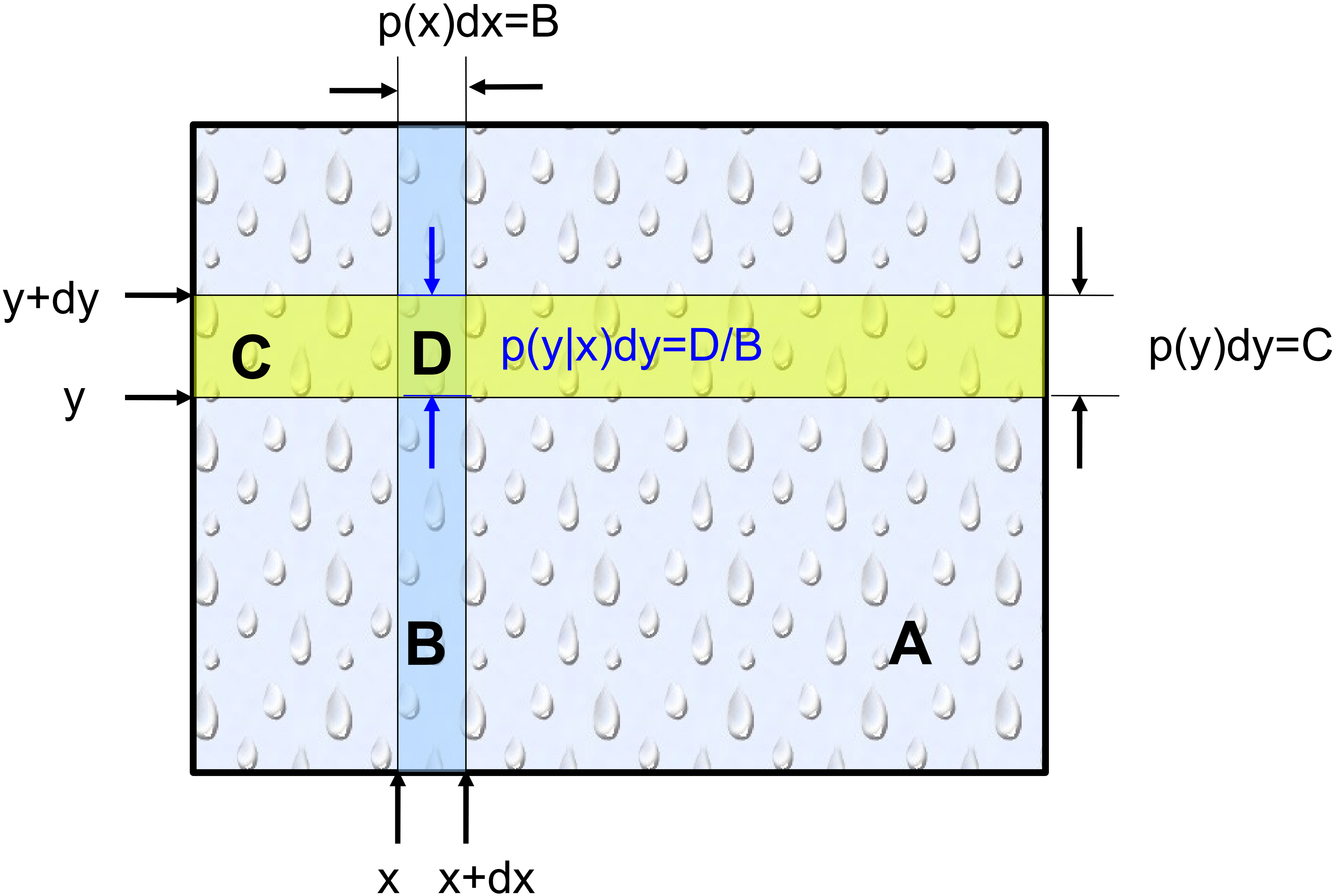
Visualizing Bayes' rule: *p(x) and p(y)*

$$p(x|y_{obs}) \propto p_o(x)\, p(y_{obs}|x)$$



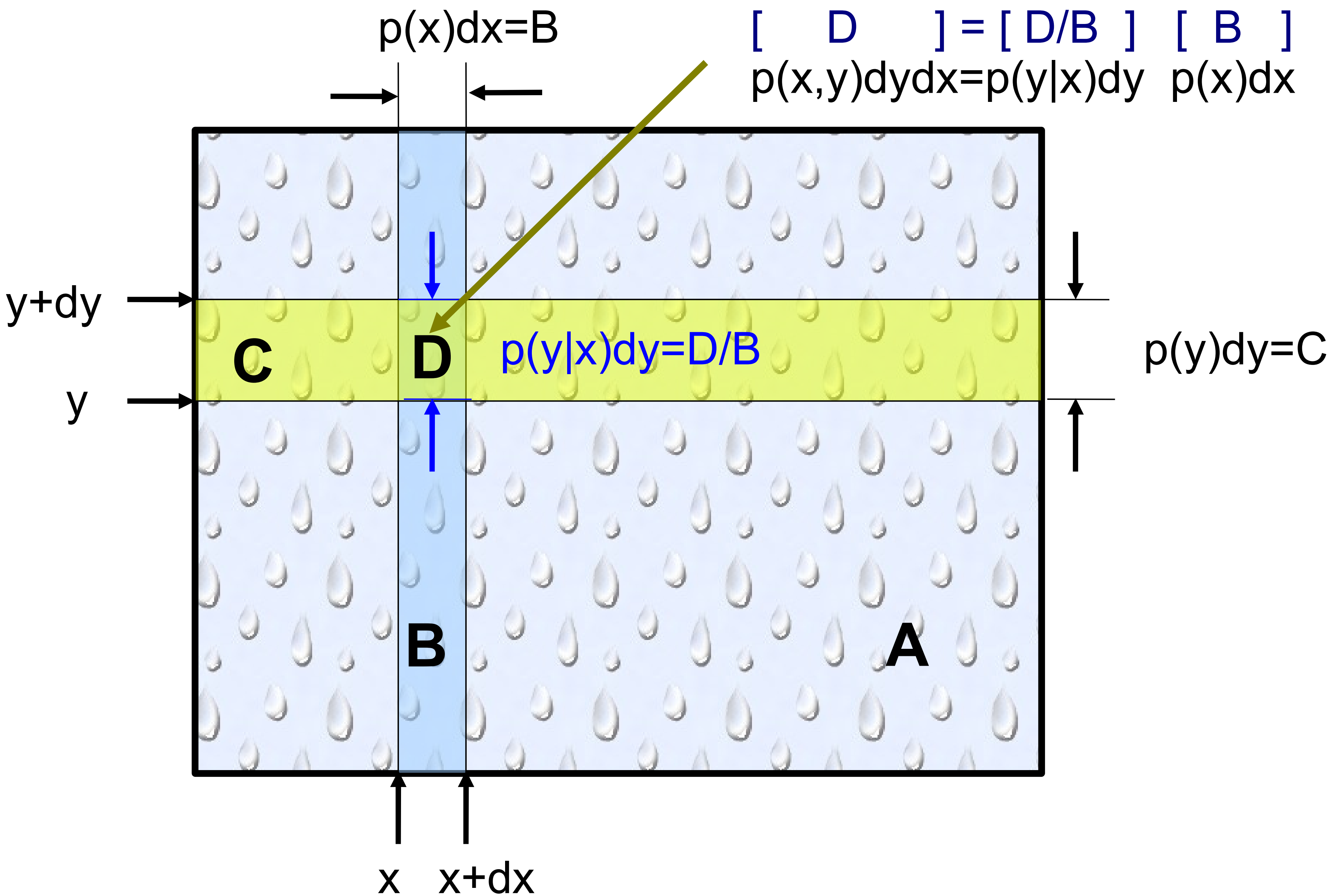
A is all the drops in the box
B is the drops in the vertical strip
C is drops in horizontal strip
D is the intersection of B and C

Visualizing Bayes' rule: *p(y|x) and p(x|y)*



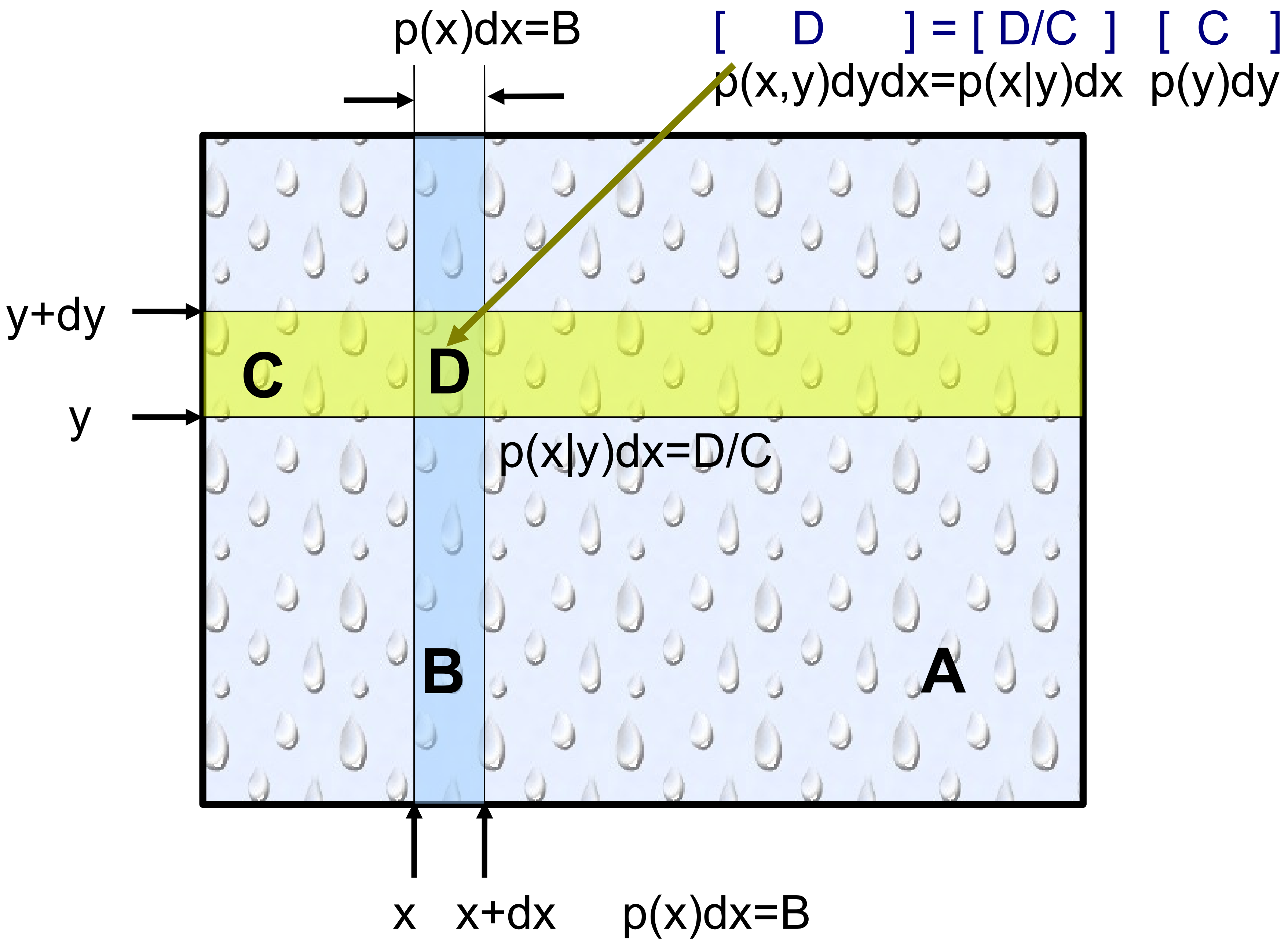
Considering only drops from x to x+dx, p(y|x)dy is the fraction of drops from y to y+dy

Visualizing Bayes' rule: *p(x,y)*



p(x,y)dxdy is the fraction of all drops inside the box from x to x+dx and y to y+dy

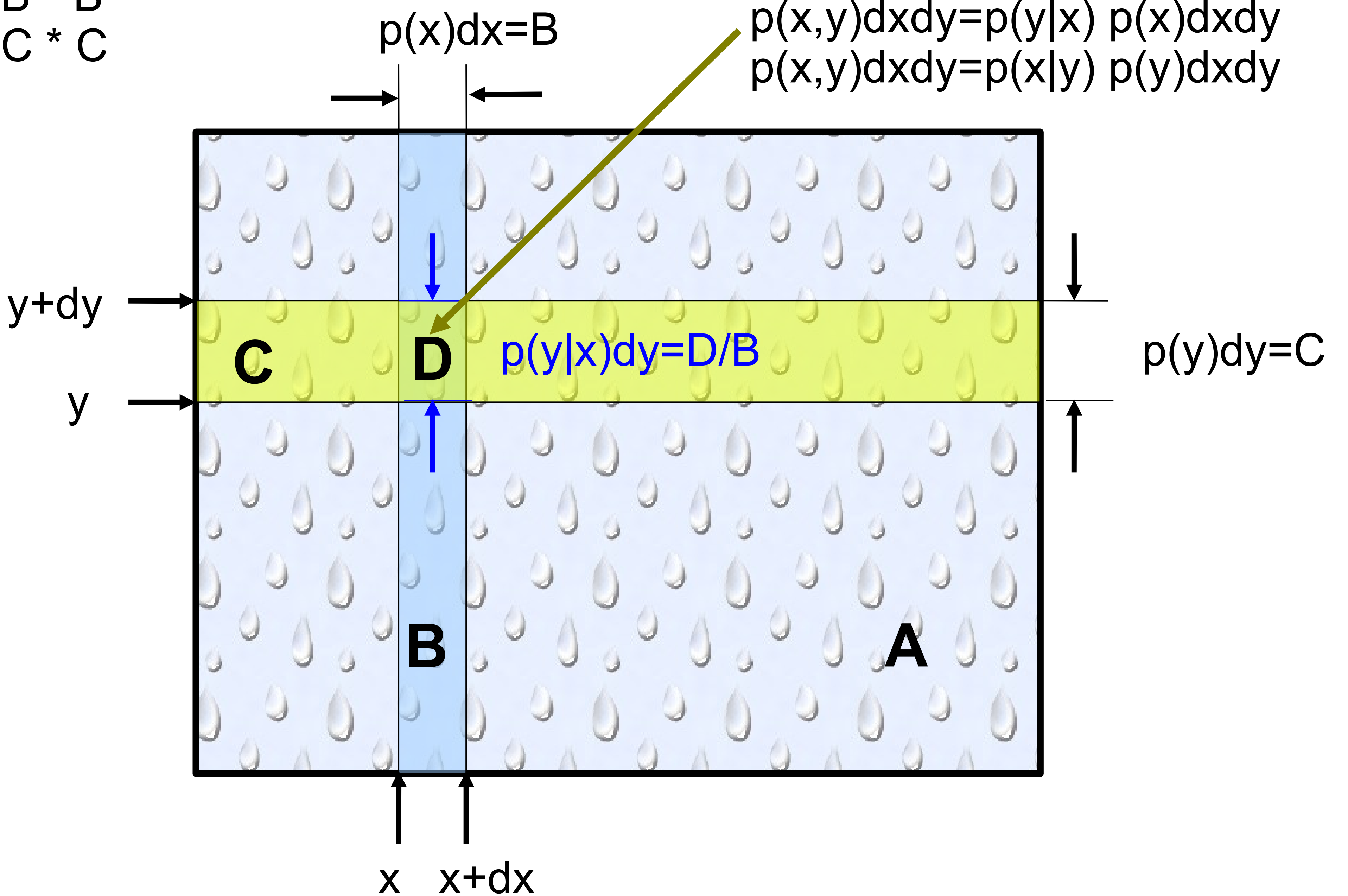
Visualizing Bayes' rule: *p(x,y)*



p(x,y)dxdy is the fraction of all drops inside the box from x to x+dx and y to y+dy

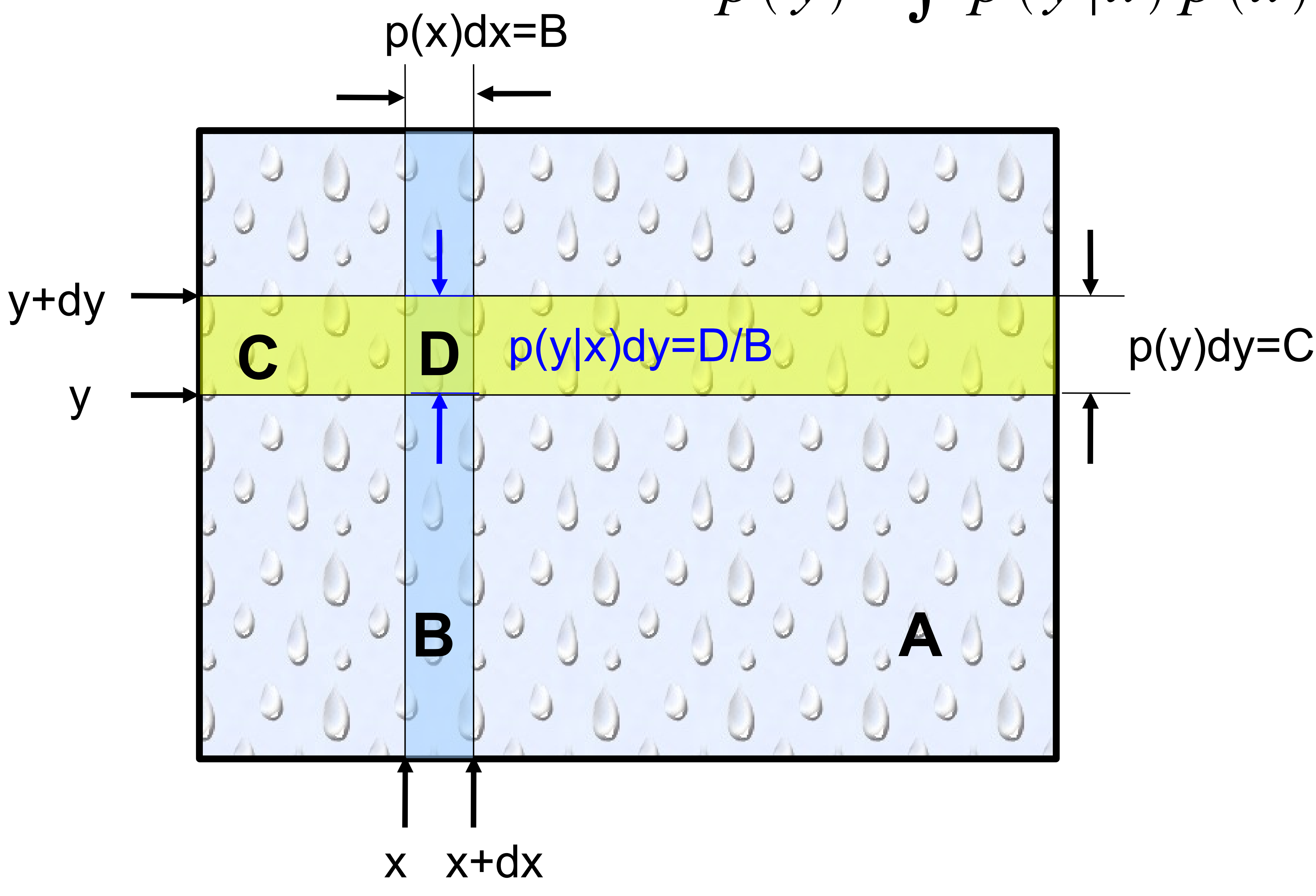
Visualizing Bayes' rule

D = D/B * B
D = D/C * C



An identity we will need now and later....

$$p(y) = \int p(y|x)\, p(x)\, dx$$



Visualizing Bayes' rule

p(x,y) written two ways $$p(x|y)\,p(y) = p(y|x)\,p(x)$$

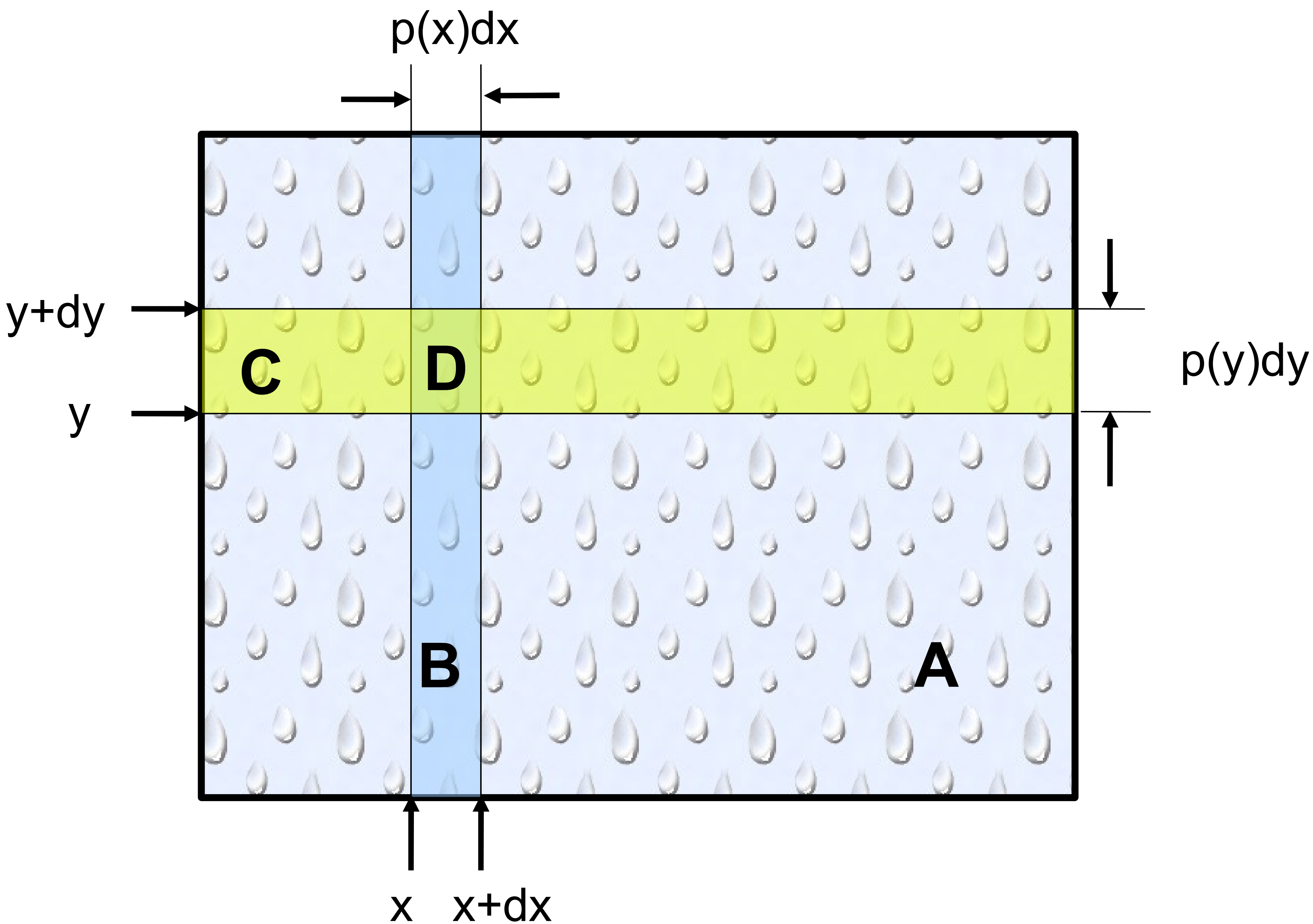
rearrangement... $$p(x|y) = p(y|x)\,p(x) / p(y)$$

An identity $$p(y) = \int p(y|x)\,p(x)\,dx$$

Substitution...Bayes' rule:

$$p(x|y) = p(y|x)\,p(x) / \int p(y|x)\,p(x)\,dx$$

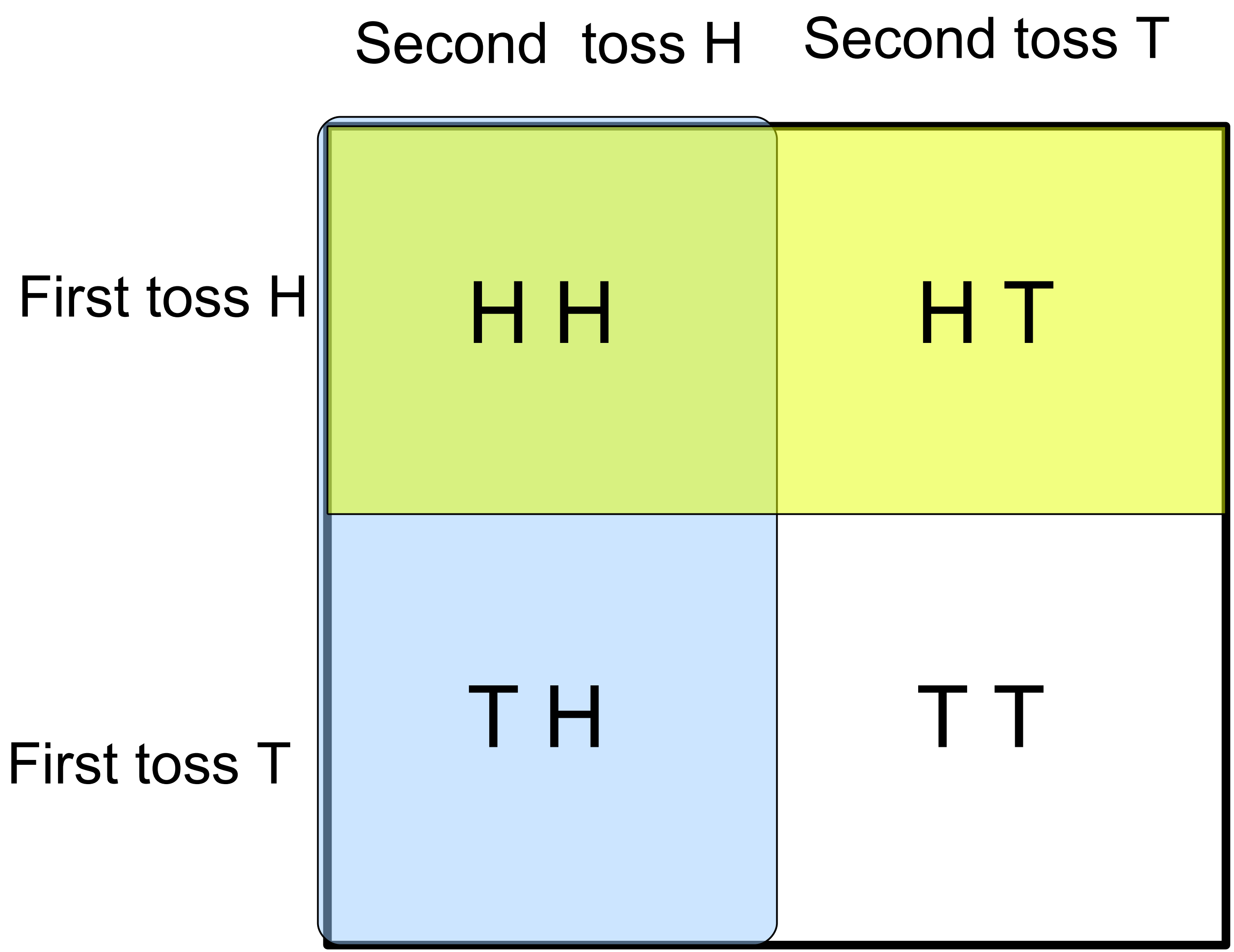
Bayes' rule as a systematic way to evaluate truth-tables



p(x) dx is the fraction of all drops from x to x+dx

Bayes' rule as a systematic way to evaluate truth-tables

We toss a coin twice and get at least one “heads”.
What is the probability that the first toss was a “head?”



Bayes' rule as a systematic way to evaluate truth-tables

We toss a coin twice and get at least one “heads”.
What is the probability that the first toss was a “head?”

	Second toss H	Second toss T
First toss H	H H	H T
First toss T	T H	T T

FS=head on first or second toss

H= heads first toss T= tails first toss

Bayes' rule:

$p(H)=A\ p_o(H)\ p(FS|H)$
$A = 1 /[\ p_o(H)\ p(FS|H) + p_o(T)\ p(FS|T)\ ]$

$p_o(H)=1/2$
$p(FS|H)= 1$
$p(FS|T)=1/2$

$A = 1/[1/2 + 1/2 * 1/2] = 4/3$
$p(H)= 4/3 * 1/2 = 2/3$

Quick Review of Bayes' rule

$$p(x \mid y_{obs}) \propto p_o(x)\, p(y_{obs} \mid x)$$

$p(x \mid y_{obs})$	Probability of x given our observations
$p_o(x)$	What we knew beforehand about x
$p(y_{obs} \mid x)$	Probability of measuring these observations if x were the correct value

Marginalization

What if the observations depend on *z* as well as *x* ?
(Maybe z is model error)

$$p(y_{obs}|x)$$

What we want to use in Bayes' rule

$$p(y_{obs}|x) = \int p(y_{obs}|x,z)\,p(z)\,dz$$

"Integrate over the nuisance variable z, weighting by p(z)"

Marginalization

y_{obs}=observations

$$p(y_{obs}) = \int p(y_{obs}|z)\,p(z)\,dz$$

Identity we saw earlier

$$p(y_{obs}|x) = \int p(y_{obs}|z,x)\,p(z|x)\,dz$$

The whole equation can be for a particular value of x

$$p(y_{obs}|x) = \int p(y_{obs}|z,x)\,p(z)\,dz$$

If z does not depend on *x*, *p(z)=p(z|x)*

"Integrate over the nuisance variable z, weighting by p(z)"

Marginalization with Bayes' rule

We want to get *p(x)* using $p(y_{obs}|x)$ in Bayes' rule...

y_{obs} is an experimental measurement of *y*

$$p(y_{obs}|y) \propto e^{-(y_{obs}-y)^2/2\sigma^2}$$

y depends on x and z (perhaps z is model error)

$$y = y(z, x)$$

...then we can integrate over *z* to get $p(y_{obs}|x)$:

$$p(y_{obs}|x) = \int p(y_{obs}|y(z,x))\, p(z)\, dz$$

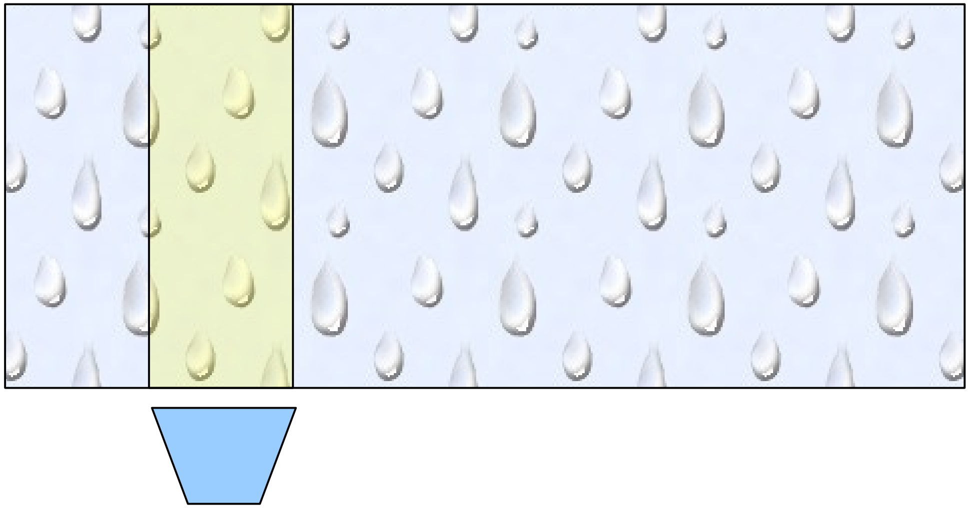
Repeated measurements with systematic error

We want to know on average how many drops D_{avg} of rain hit a surface per 100 cm^2 per minute.

The rain does not fall uniformly: $D(x)=D_{avg}+E(x)$ where the SD of $E(x)$ is e. However we only sample one place

We count the drops N falling in 1 minute into a fixed bucket with top area of 100 cm^2 m times (N_1, N_2...) with a mean of n.

What is the weighted mean estimate $<D_{avg}>$? What is the uncertainty in $<D_{avg}>$?



Repeated measurements with systematic error

We want to get $p(D_{avg})$ using $p(N_{obs}|D_{avg})$ in Bayes' rule...but the rate into our bucket *D depends on* D_{avg} *and E:*

$$D = D_{avg} + E$$

$$p(E) \propto e^{-E^2/2e^2}$$

N_{obs} is the number of drops we count with SD of $n^{1/2}$:

$$p(N_{obs}|D_{avg}, E) \propto e^{-(N_{obs} - (D_{avg}+E))^2/2s^2}$$

Including all *m* measurements N_1, N_2...

$$p(N_1, N_2 \ldots | D_{avg}, E) \propto e^{-\sum_i (N_i - (D_{avg} + E))^2/2s^2}$$

From previous slide

$$p(N_1, N_2 \ldots | D_{avg}, E) \propto e^{-\sum_i (N_i - (D_{avg} + E))^2 / 2s^2}$$

$$p(E) \propto e^{-E^2/2e^2}$$

We have $p(N_1,N_2\ldots|D_{avg},E)$. We want $p(N_1,N_2\ldots|D_{avg})$. Integrate over the nuisance variable E:

$$p(N_1, N_2 \ldots | D_{avg}) = \int p(N_1, N_2 \ldots | D_{avg}, E)\, p(E)\, dE$$

Yielding (where n is the mean value of N: $\langle N_1,N_2\ldots \rangle$)

$$p(N_1, N_2 \ldots | D_{avg}) \propto e^{-(D_{avg} - n)^2 / 2(e^2 + s^2/m)}$$

Now we have $p(N_1,N_2\ldots|D_{avg})$ and we are ready to apply Bayes' rule

We have the probability of the observations given D_{avg},

$$p(N_1, N_2 \ldots | D_{avg}) \propto e^{-(D_{avg}-n)^2/2(e^2+s^2/m)}$$

Bayes' rule gives us the probability of D_{avg} given the observations:

$$p(D_{avg} | N_1, N_2 \ldots) \propto p_o(D_{avg})\, e^{-(D_{avg}-n)^2/2(e^2+s^2/m)}$$

If the prior $p_o(D_{avg})$ is uniform:

$$p(D_{avg} | N_1, N_2 \ldots) \propto e^{-(D_{avg}-n)^2/2(e^2+s^2/m)}$$

$$\langle D_{avg} \rangle = n = \langle N \rangle \qquad \sigma^2 = e^2 + s^2/m$$

Summary: How to apply a Bayesian analysis to any measurement problem

1. Write down what you really want to know: $p(D_{avg})$

2. Write down prior knowledge: $p_o(D_{avg})=1$

3. Write down how the true value of the thing you are measuring depends on what you really want to know and any other variables: $D=D_{avg}+E$

4. Write down probability distributions for errors in measurement and for the variables you don't know:
$p(N_{obs}|D)$ ***and*** $p(E)$

How to apply a Bayesian analysis of any measurement problem

5. Use 3&4 to write probability distribution for measurements given values of what you want to know and of nuisance variables: $p(N_1, N_2 \ldots | D_{avg}, E)$

6. Integrate over the nuisance variables (E), weighted by their probability distributions p(E) to get probability of measurements given what you want to know:
$p(N_1, N_2 \ldots | D_{avg})$

7. Apply Bayes' rule to get the probability distribution for what you want to know, given the measurements:
$p(D_{avg} | N_1, N_2 \ldots) = p_o(D_{avg})\, p(N_1, N_2 \ldots | D_{avg})$

Applications of the Bayesian approach in macromolecular crystallography

- **Correlated MIR phasing (errors due to non-isomorphism are correlated among heavy-atom derivatives)**
- **Correlated MAD phasing (errors in heavy-atom model are correlated among wavelengths)**
- **Bayesian difference refinement (errors in model of macromolecular structure correlated between two structures)**
- **Macromolecular refinement (phase unknown and model errors present)**

Tutorials

- **Working through simple Bayesian exercises from handout in a group**
- **PHENIX demo and discussion**
- **Density modification and model-building theory and discussion**
- **Discussion of individual challenging examples and questions from students**