Probability Review

Lyle Ungar

I'm comfortable working with probability density functions

A) yes

B) no

MLE/MAP

Bernoulli and beta distributions

Probability density functions

Gaussians

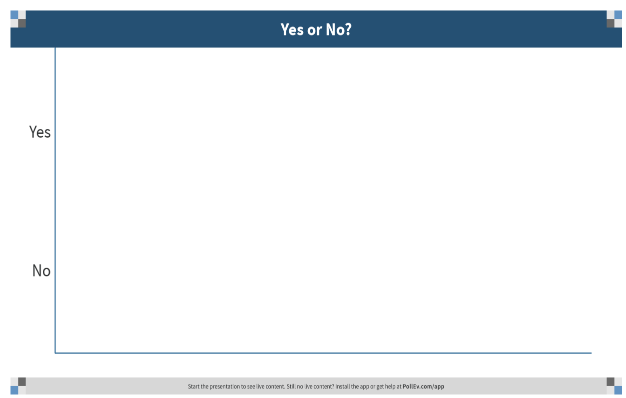
Expected value

- **Lectures are recorded**
  - but it’s much better to come to class
- **Piazza rocks!**
- **Questions? (chat window)**

I know the difference between MLE and MAP

A) yes

B) no



MLE/MAP

- MLE maximizes what?
- MAP maximizes what?
- When is MLE the same as MAP?
- We will almost always use MAP. Why?

Questions?

Top

Probability Densities (PDFs)

Originally by

Andrew W. Moore

Heavily edited by Lyle Ungar

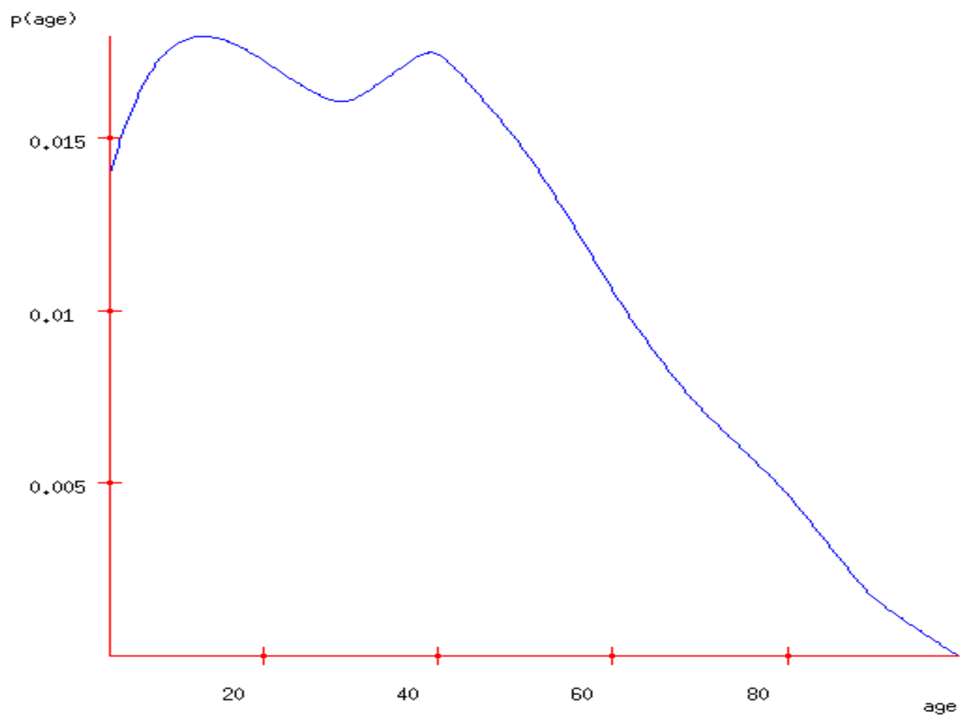
Probability Densities in ML

- Why we should care
- Notation and fundamentals of continuous PDFs
- Multivariate continuous PDFs
- Expected value, variance, covariance

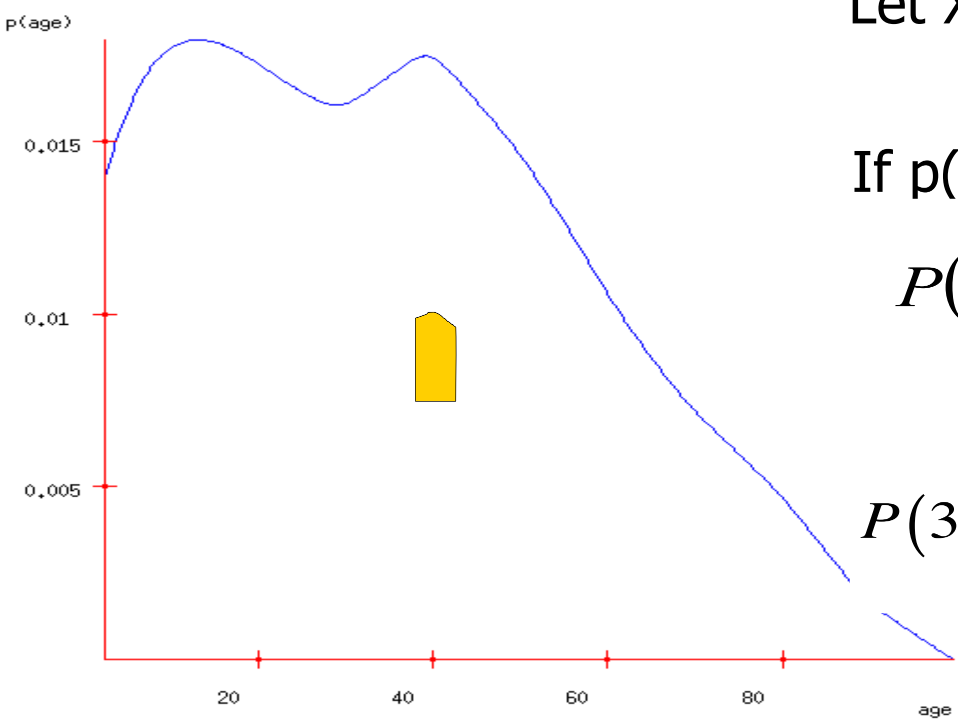
Why we should care

- **Real numbers occur most real data**
  - Can't always quantize them
- **Parameters in models are real valued**
- **You'll need to *intimately* understand PDFs for**
  - kernel methods,
  - clustering with mixture models
  - time series, HMMs
  - proofs about regression

A PDF of American Ages in 2000



A PDF of American Ages in 2000



Let X be a continuous random variable.

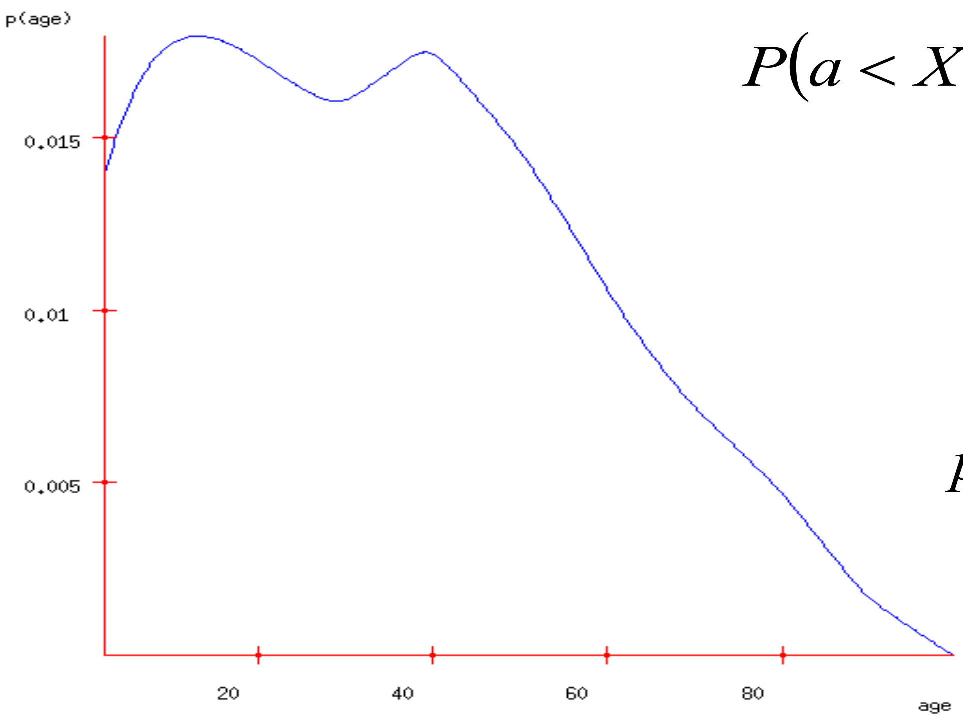
If p(x) is a Probability Density

$$P(a < X \leq b) = \int_{x=a}^{b} p(x)dx$$

$$P(30 < \text{Age} \leq 50) = \int_{\text{age}=30}^{50} p(\text{age})\,d\text{age}$$

$$= 0.36$$

Properties of PDFs



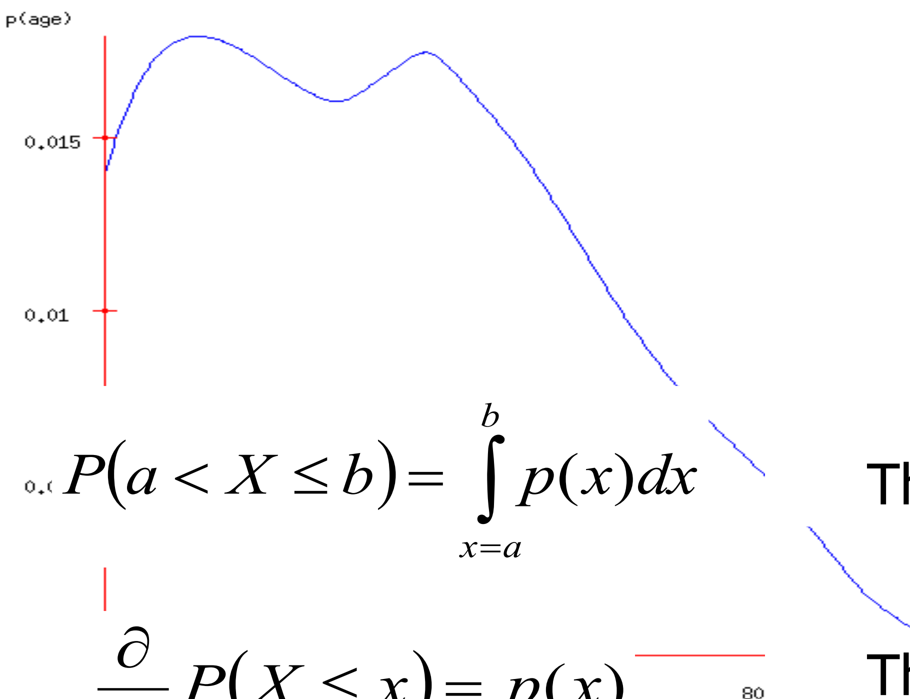
$$P(a < X \leq b) = \int_{x=a}^{b} p(x)\,dx$$

That means…

$$p(x) = \lim_{h \to 0} \frac{P\left(x - \frac{h}{2} < X \leq x + \frac{h}{2}\right)}{h}$$

$$\frac{\partial}{\partial x} P(X \leq x) = p(x)$$

Properties of PDFs



$$P(a < X \le b) = \int_{x=a}^{b} p(x)dx$$

Therefore…

$$\int_{x=-\infty}^{\infty} p(x)dx = 1$$

$$\frac{\partial}{\partial x} P(X \le x) = p(x)$$

Therefore…

$$\forall x : p(x) \ge 0$$

Talking to your stomach

- **What's the gut-feel meaning of p(x)?**

If

$p(5.31) = 0.06$ and $p(5.92) = 0.03$

then

when a value X is sampled from the distribution, you are 2 times as likely to find that X is "very close to" 5.31 than that X is "very close to" 5.92.

Talking to your stomach

- **What's the gut-feel meaning of p(x)?**

If

p(a) = 0.06 and p(b) = 0.03

then

when a value X is sampled from the distribution, you are 2 times as likely to find that X is "very close to" a than that X is "very close to" b .

Talking to your stomach

- **What's the gut-feel meaning of p(x)?**

If

p(a) = $2z$ and p(b) = z

then

when a value X is sampled from the distribution, you are 2 times as likely to find that X is "very close to" a than that X is "very close to" b .

Talking to your stomach

- **What's the gut-feel meaning of p(x)?**

If

$p(a) = \alpha z$ and $p(b) = z$

then

when a value X is sampled from the distribution, you are α times as likely to find that X is "very close to" a than that X is "very close to" b

Talking to your stomach

- **What's the gut-feel meaning of p(x)?**

If

$$\frac{p(a)}{p(b)} = \alpha$$

then

when a value X is sampled from the distribution, you are α times as likely to find that X is "very close to" a than that X is "very close to" b .

Talking to your stomach

- **What's the gut-feel meaning of p(x)?**

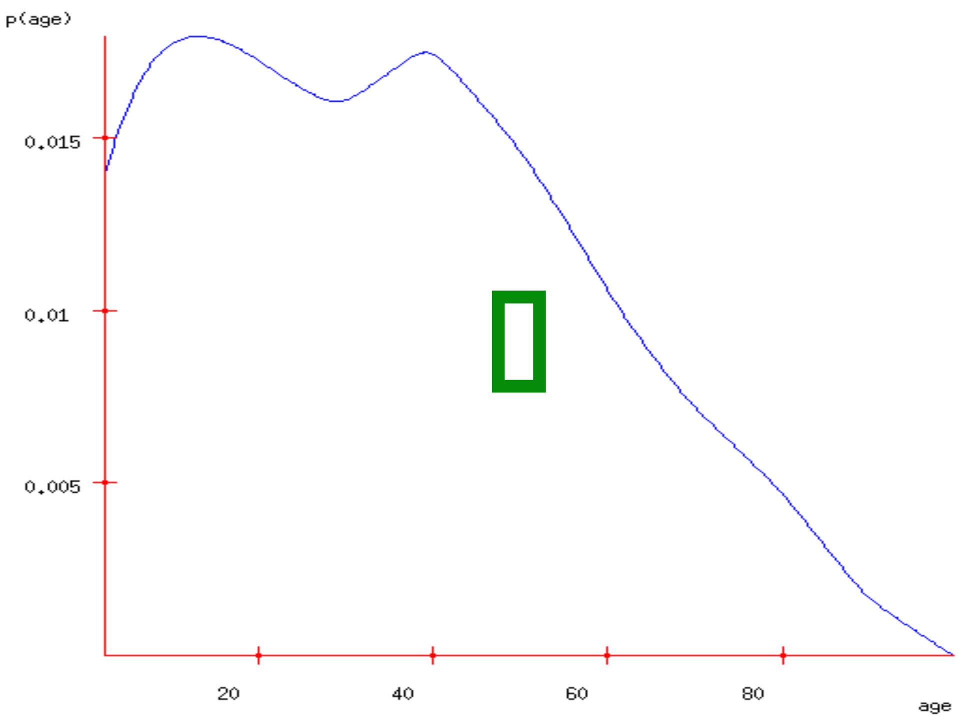
If

$$\frac{p(a)}{p(b)} = \alpha$$

then

$$\lim_{h \to 0} \frac{P(a-h < X < a+h)}{P(b-h < X < b+h)} = \alpha$$

Yet another way to view a PDF



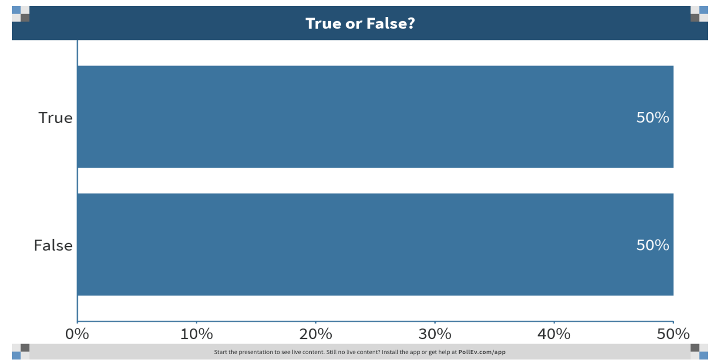
A recipe for sampling a random age.

1. Generate a random dot from the rectangle surrounding the PDF curve. Call the dot (age, d)
2. If d < p(age) stop and return age
3. Else try again: go to Step 1.

Test your understanding

- **True or False:**

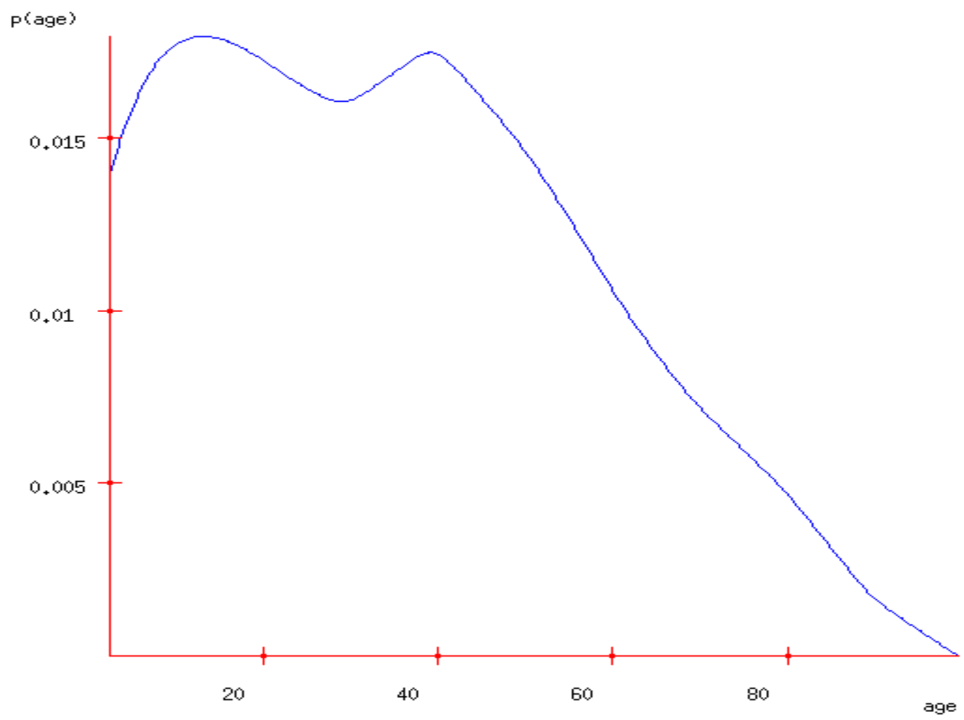
$$\forall x : p(x) \le 1$$



- **True or False:**

$$\forall x : P(X = x) = 0$$

Expectations

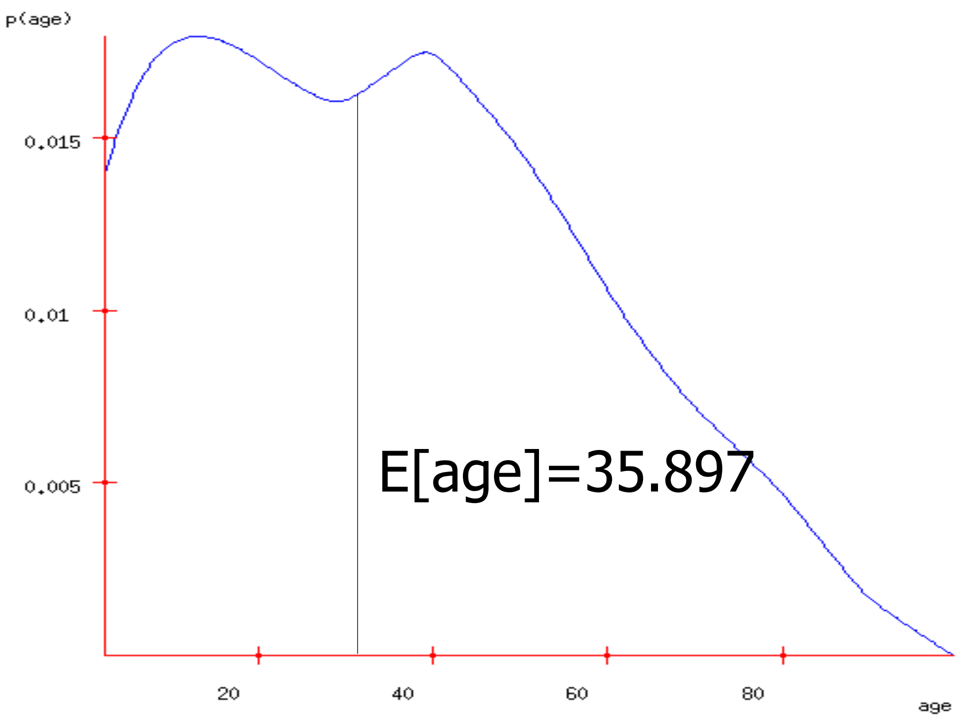


E[X] = the expected value of random variable X

= the average value we'd see if we took a very large number of random samples of X

$$= \int_{x=-\infty}^{\infty} x \, p(x) \, dx$$

Expectations



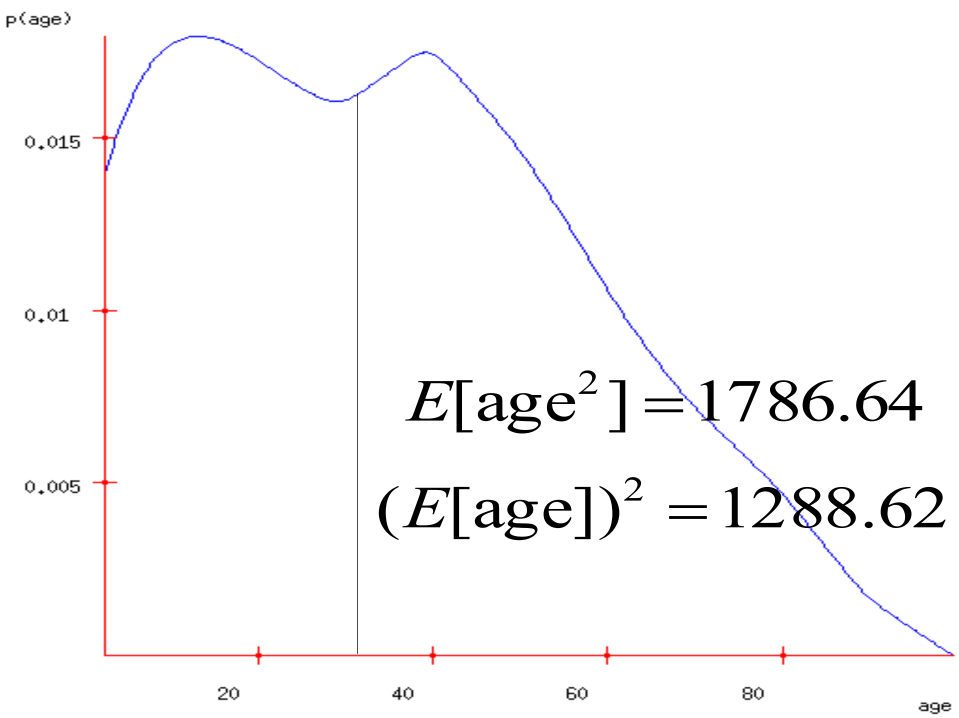
E[X] = the **expected value** of random variable X

= the **average value** we'd see if we took a very large number of random samples of X

= the **first moment** of the shape formed by the axes and the blue curve

= the **best value** to choose if you must guess an unknown person's age and you'll be fined the square of your error

Expectation of a function



$$E[\text{age}^2] = 1786.64$$

$$(E[\text{age}])^2 = 1288.62$$

μ=E[f(X)] = the expected value of f(x) where x is drawn from X's distribution.

= the average value we'd see if we took a very large number of random samples of f(X)

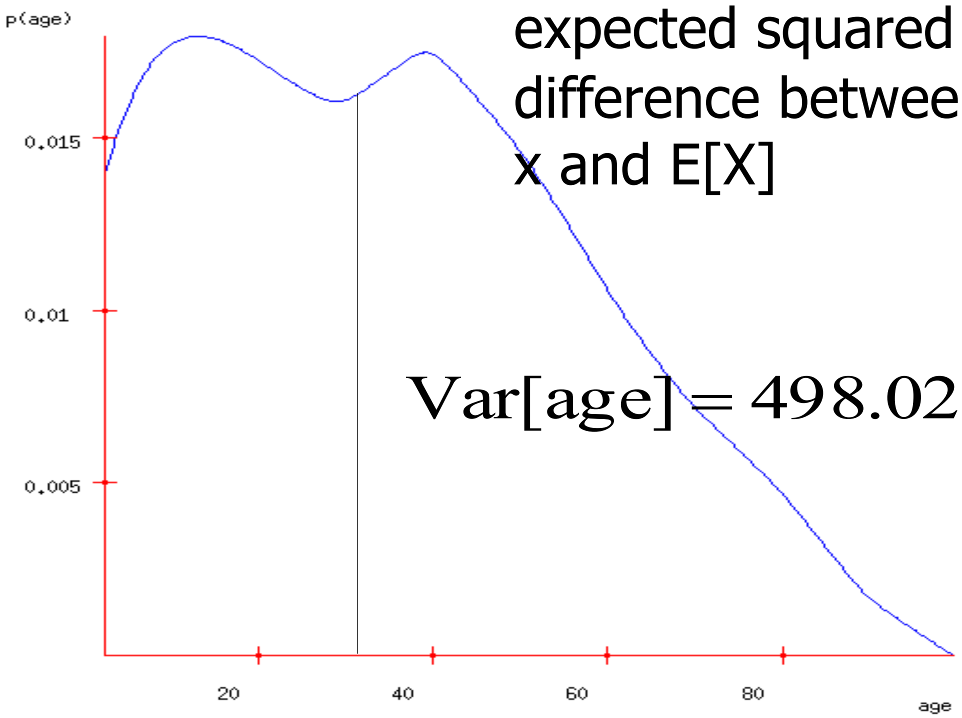
$$\mu = \int_{x=-\infty}^{\infty} f(x)\, p(x)\, dx$$

Note that in general:

$$E[f(x)] \neq f(E[X])$$

Variance

σ^2 = Var[X] = the expected squared difference between x and E[X]

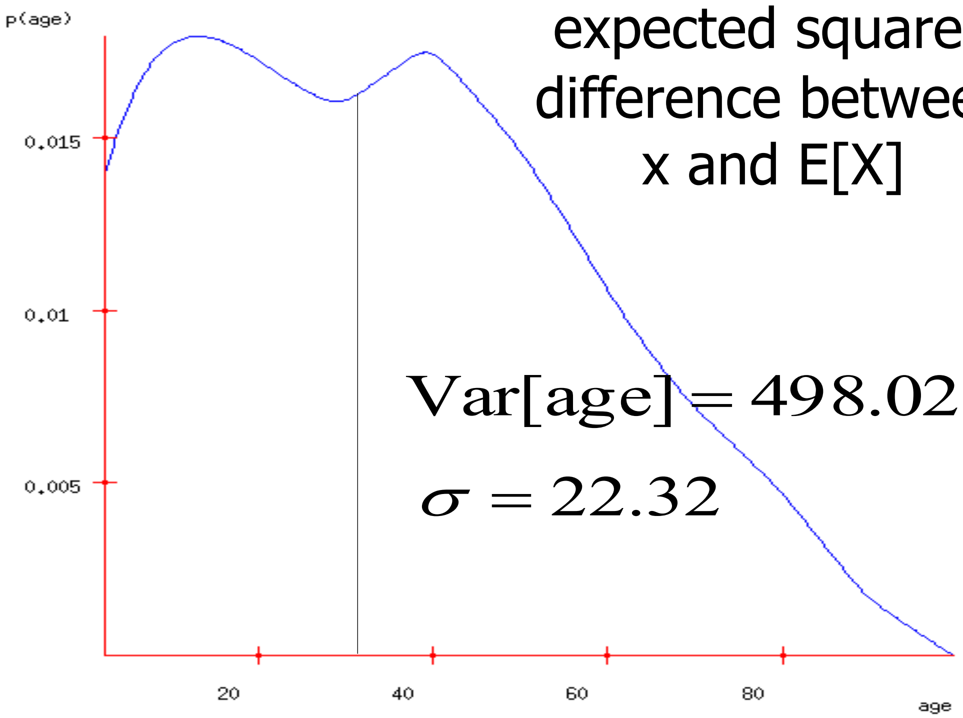


$$\mathrm{Var[age]} = 498.02$$

$$\sigma^2 = \int_{x=-\infty}^{\infty} (x-\mu)^2 \, p(x) \, dx$$

= amount you'd expect to lose if you must guess an unknown person's age and you'll be fined the square of your error, and assuming you play optimally

Standard Deviation



σ^2 = Var[X] = the expected squared difference between x and E[X]

$$\mathrm{Var[age]} = 498.02$$

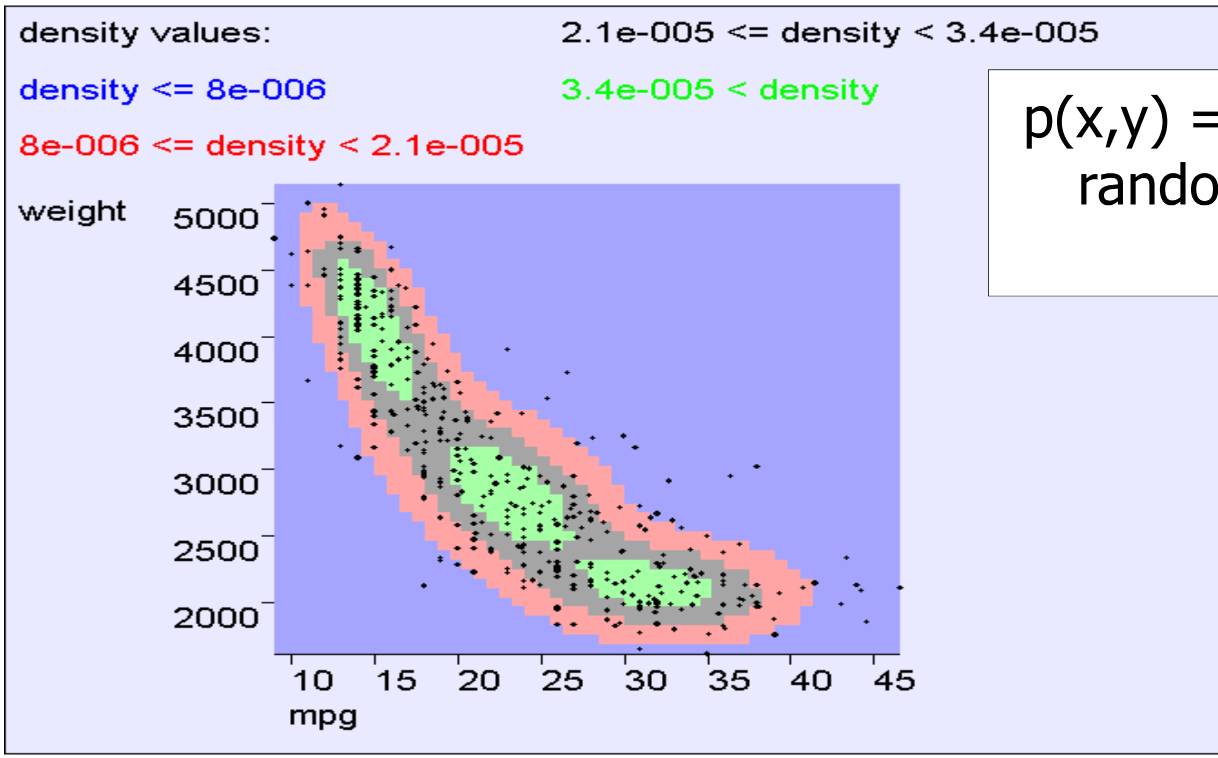
$$\sigma = 22.32$$

$$\sigma^2 = \int_{x=-\infty}^{\infty} (x-\mu)^2 \, p(x)\, dx$$

= amount you'd expect to lose if you must guess an unknown person's age and you'll be fined the square of your error, and assuming you play optimally

σ = Standard Deviation = "typical" deviation of X from its mean $\sigma = \sqrt{\mathrm{Var}[X]}$

In two dimensions

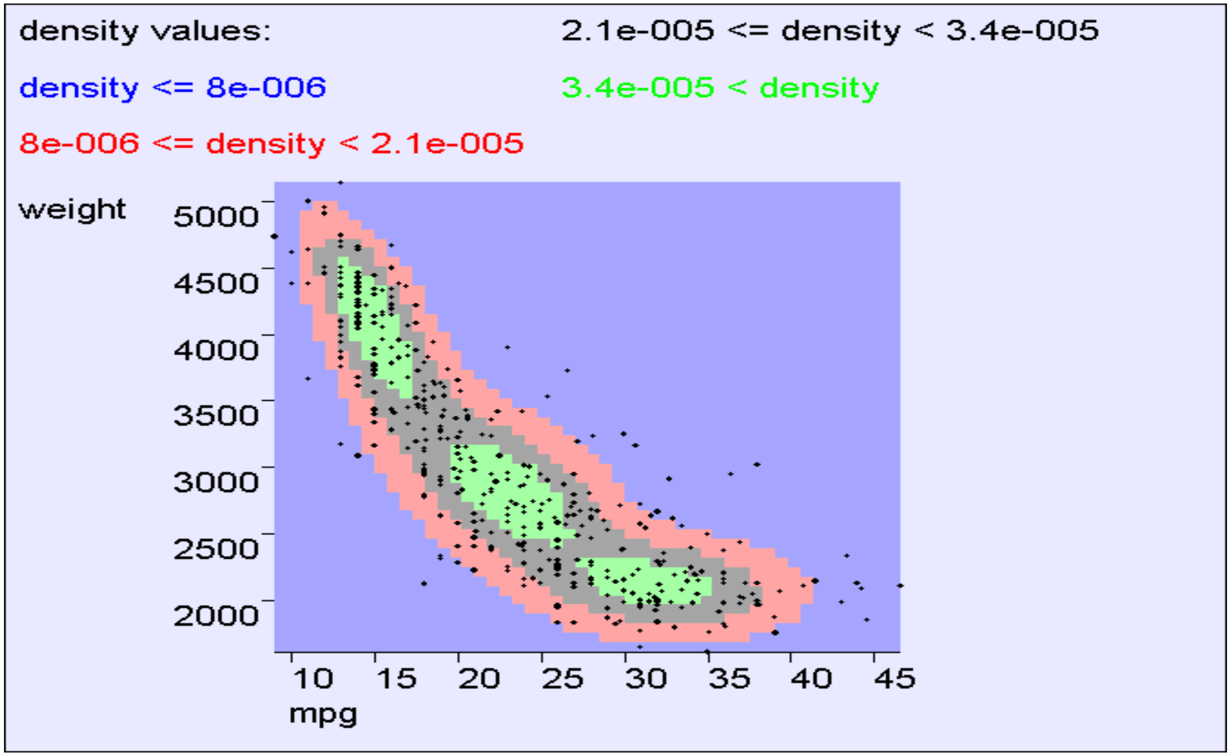


p(x,y) = probability density of random variables (X,Y) at location (x,y)

In two dimensions

Let *X,Y* be a pair of continuous random variables, and let R be some region of (X,Y) space…

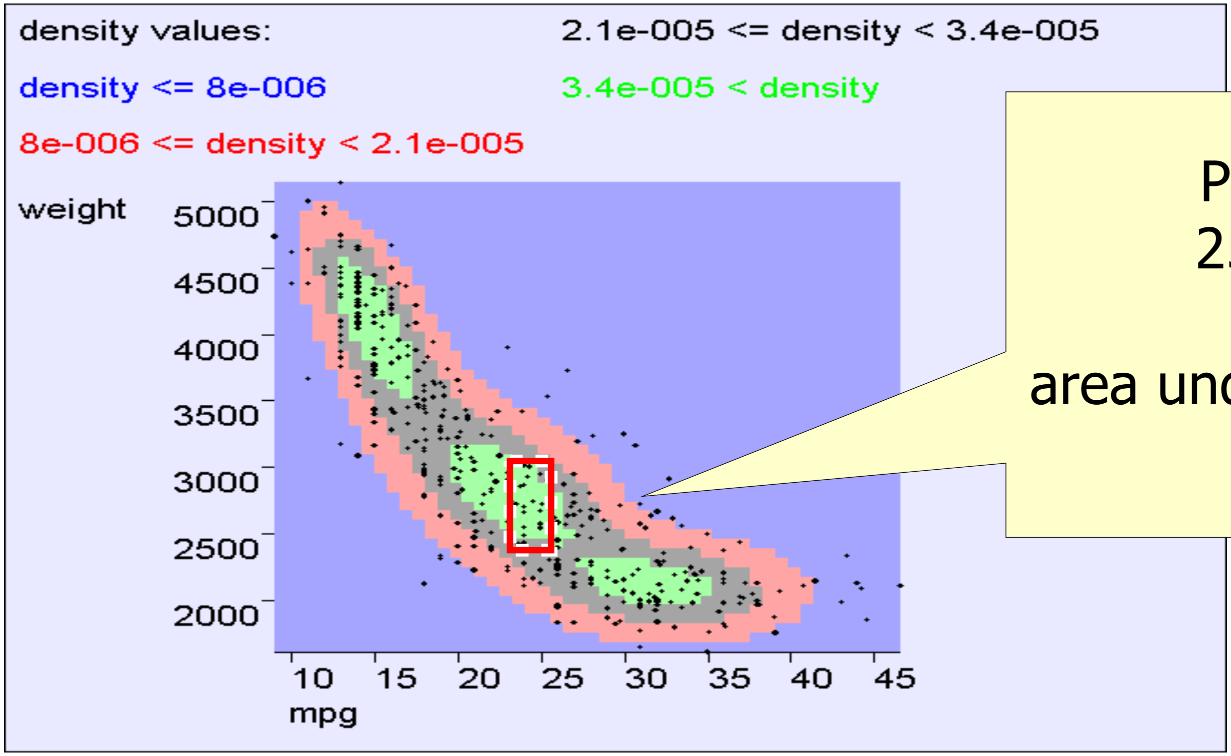
$$P((X,Y)\in R)=\iint_{(x,y)\in R} p(x,y)\,dy\,dx$$



In two dimensions

Let X,Y be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y)\in R) = \iint_{(x,y)\in R} p(x,y)\,dy\,dx$$



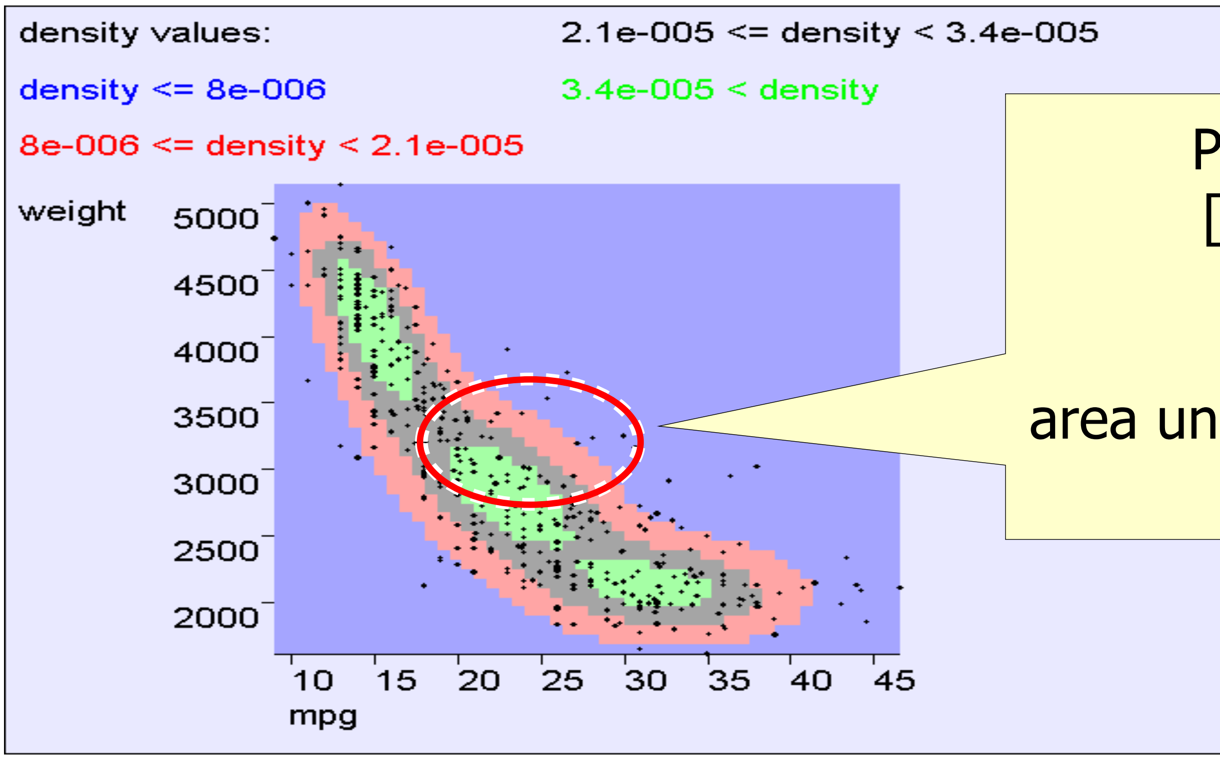
P(20<mpg<30 and 2500<weight<3000) =

area under the 2-d surface within the red rectangle

In two dimensions

Let X, Y be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y) \in R) = \iint_{(x,y)\in R} p(x,y)\,dy\,dx$$



P($[(\text{mpg}-25)/10]^2$ + $[(\text{weight}-3300)/1500]^2$ < 1) =

area under the 2-d surface within the red oval

In two dimensions

Let X,Y be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y)\in R) = \iint_{(x,y)\in R} p(x,y)\,dy\,dx$$

Take the special case of region R = “everywhere”.

Remember that with probability 1, (X,Y) will be drawn from “somewhere”.

$$\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} p(x,y)\,dy\,dx = 1$$

In two dimensions

Let *X,Y* be a pair of continuous random variables, and let R be some region of (X,Y) space…

$$P((X,Y)\in R)=\iint_{(x,y)\in R} p(x,y)\,dy\,dx$$

$$p(x,y)=\lim_{h\to 0}\frac{P\left(x-\frac{h}{2}<X\le x+\frac{h}{2} \quad\wedge\quad y-\frac{h}{2}<Y\le y+\frac{h}{2}\right)}{h^2}$$

In m dimensions

Let $(X_1, X_2, \ldots X_m)$ be an n-tuple of continuous random variables, and let R be some region of $\mathbf{R}^m$ …

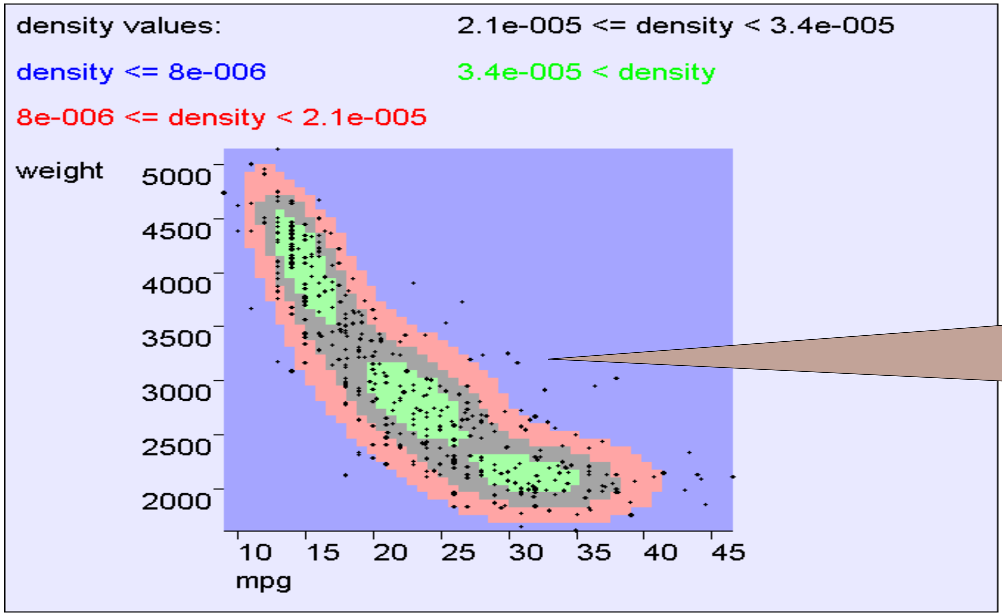
$$P((X_1, X_2, \ldots, X_m) \in R) =$$

$$\iint_{(x_1, x_2, \ldots, x_m) \in R} \cdots \int p(x_1, x_2, \ldots, x_m)\, dx_m, \ldots dx_2, dx_1$$

Independence

$$X \perp Y \text{ iff } \forall \mathrm{x}, \mathrm{y} : p(x, y) = p(x)\, p(y)$$

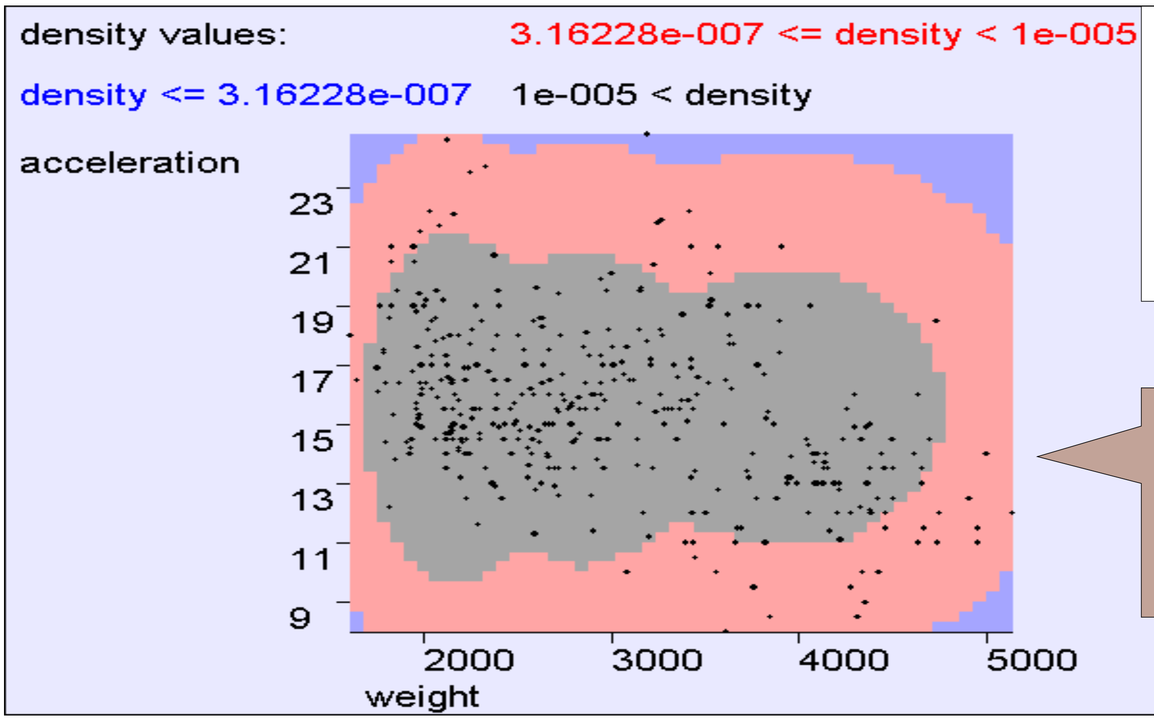
If X and Y are independent then knowing the value of X does not help predict the value of Y



mpg,weight NOT independent

Independence

$$X \perp Y \text{ iff } \forall \mathrm{x}, \mathrm{y} : p(x, y) = p(x)p(y)$$

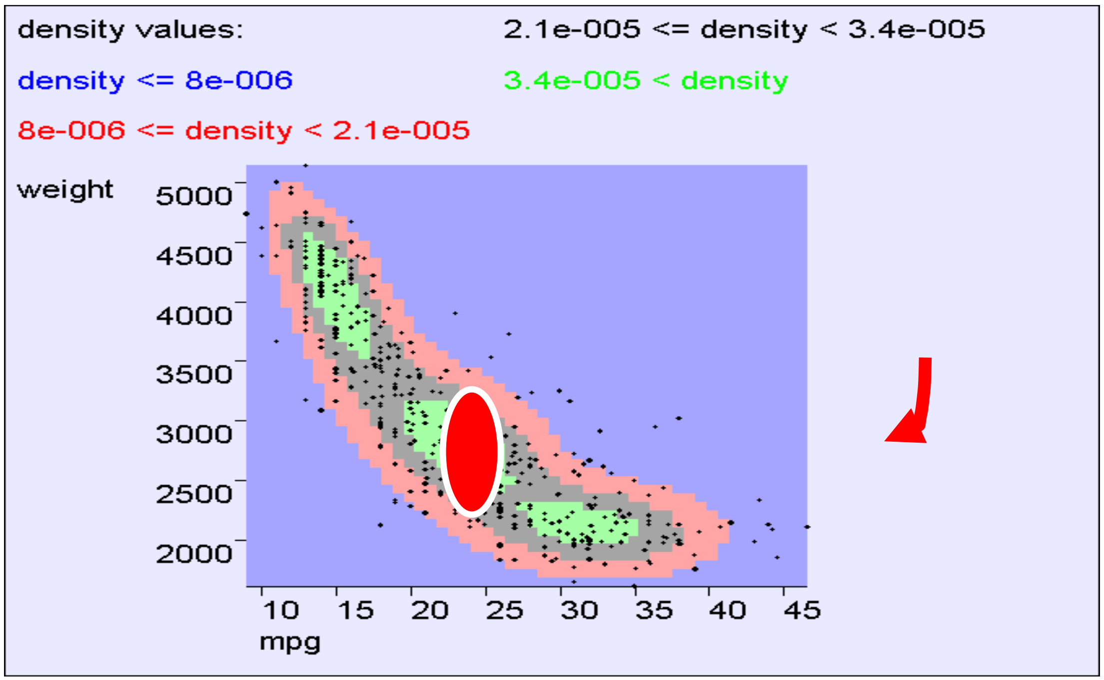


If X and Y are independent then knowing the value of X does not help predict the value of Y

the contours say that acceleration and weight are independent

Multivariate Expectation

$$\boldsymbol{\mu}_{\mathbf{X}} = E[\mathbf{X}] = \int \mathbf{x}\, p(\mathbf{x}) d\mathbf{x}$$



E[mpg,weight] = (24.5,2600)

The centroid of the cloud

Multivariate Expectation

$$E[f(\mathbf{X})] = \int f(\mathbf{x})\ p(\mathbf{x}) d\mathbf{x}$$

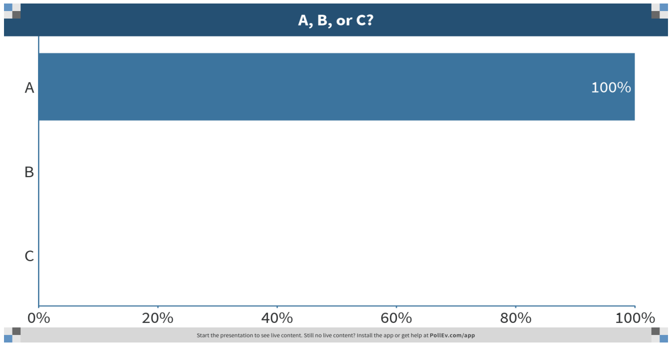
Test your understanding

Question : When (if ever) does $E[X+Y]=E[X]+E[Y]$?

A) All the time?

B) Only when X and Y are independent?

C) It can fail even if X and Y are independent?



Bivariate Expectation

$$E[f(x,y)] = \int f(x,y)\ p(x,y)dydx$$

$$\text{if } f(x,y) = x \text{ then } E[f(X,Y)] = \int x\ p(x,y)dydx$$

$$\text{if } f(x,y) = y \text{ then } E[f(X,Y)] = \int y\ p(x,y)dydx$$

$$\text{if } f(x,y) = x+y \text{ then } E[f(X,Y)] = \int (x+y)\ p(x,y)dydx$$

$$E[X+Y] = E[X] + E[Y]$$

Bivariate Covariance

$$\sigma_{xy} = \mathrm{Cov}[X, Y] = E[(X - \mu_x)(Y - \mu_y)]$$

$$\sigma_{xx} = \sigma^2{}_x = \mathrm{Cov}[X, X] = Var[X] = E[(X - \mu_x)^2]$$

$$\sigma_{yy} = \sigma^2{}_y = \mathrm{Cov}[Y, Y] = Var[Y] = E[(Y - \mu_y)^2]$$

Bivariate Covariance

$$\sigma_{xy} = \mathrm{Cov}[X, Y] = E[(X - \mu_x)(Y - \mu_y)]$$

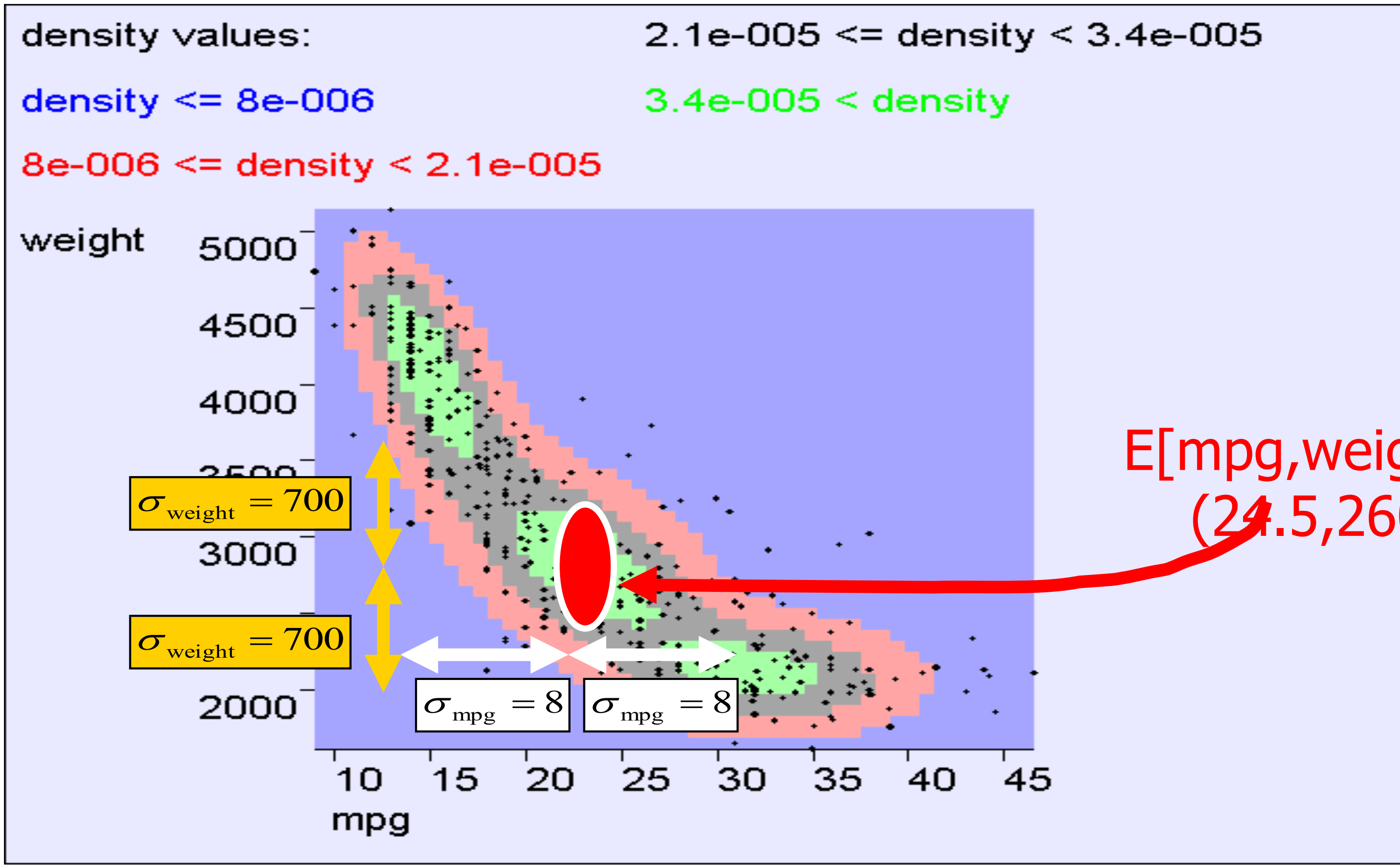
$$\sigma_{xx} = \sigma^2{}_x = \mathrm{Cov}[X, X] = Var[X] = E[(X - \mu_x)^2]$$

$$\sigma_{yy} = \sigma^2{}_y = \mathrm{Cov}[Y, Y] = Var[Y] = E[(Y - \mu_y)^2]$$

$$\text{Write } \mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}, \text{ then}$$

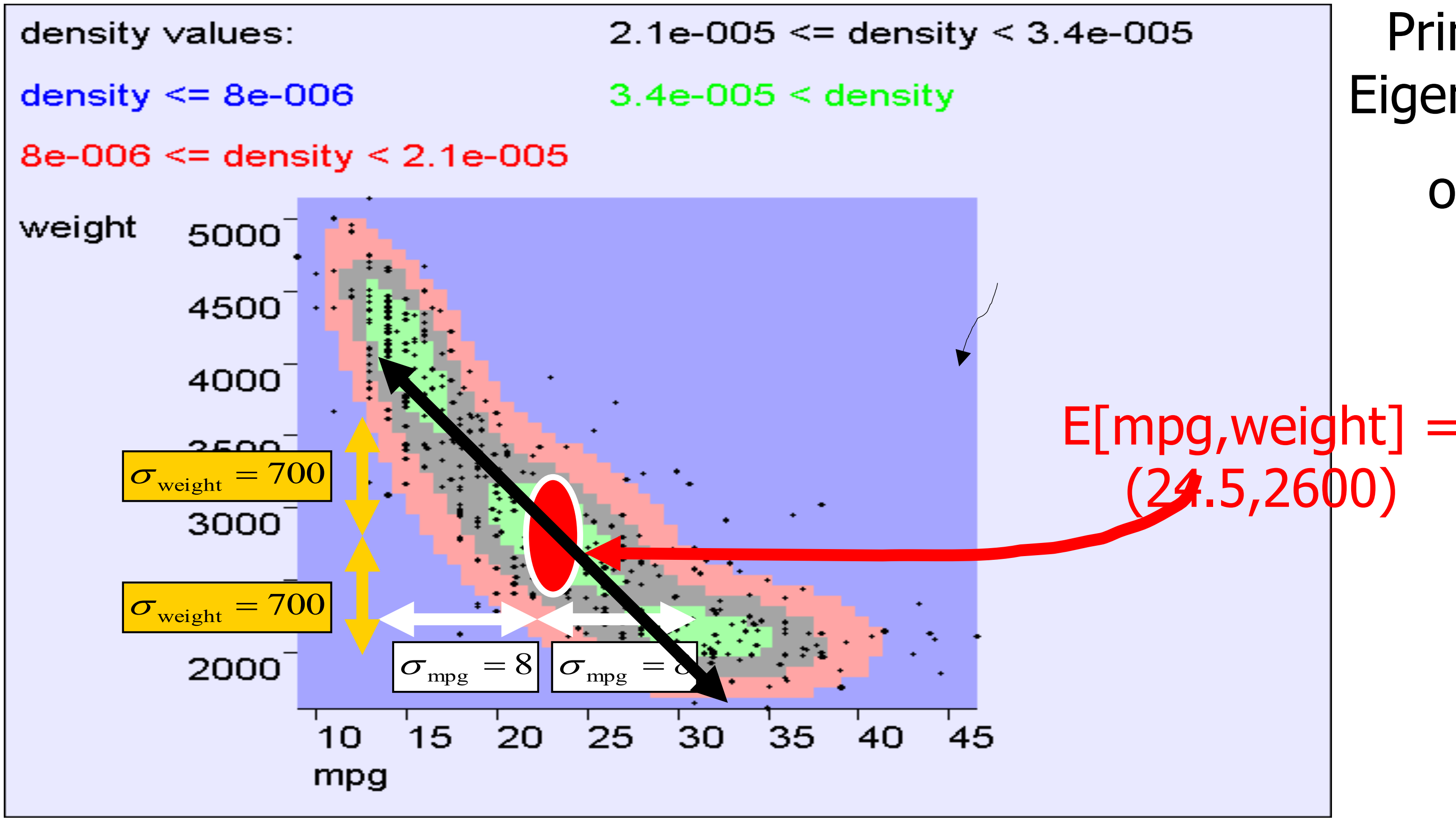
$$\mathbf{Cov}[\mathbf{X}] = E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

Covariance Intuition



E[mpg,weight] = (24.5,2600)

Covariance Intuition

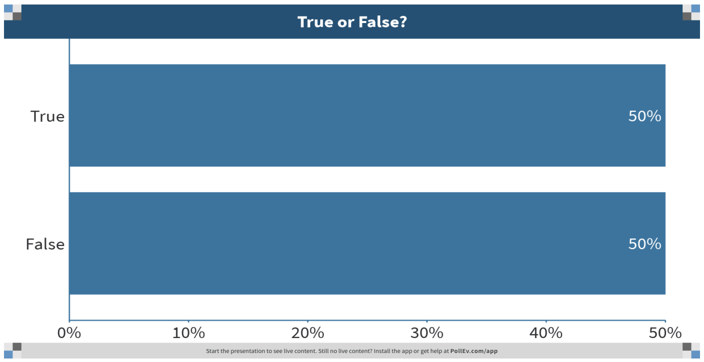


Principal Eigenvector of Σ

Covariance Fun Facts

$$\mathbf{Cov}[\mathbf{X}] = E[(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^T] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

- True or False: If $\sigma_{xy} = 0$ then X and Y are independent



How could you prove or disprove these?

Covariance Fun Facts

For example, let X be uniformly distributed in $[-1, 1]$ and let $Y = X^2$.
Clearly, X and Y are dependent, but

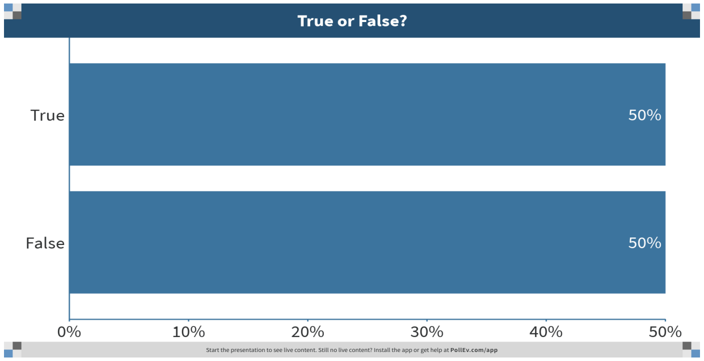
$$\begin{aligned}\operatorname{cov}(X, Y) &= \operatorname{cov}(X, X^2) \\ &= \mathrm{E}[X \cdot X^2] - \mathrm{E}[X] \cdot \mathrm{E}[X^2] \\ &= \mathrm{E}\left[X^3\right] - \mathrm{E}[X]\,\mathrm{E}[X^2] \\ &= 0 - 0 \cdot \mathrm{E}[X^2] \\ &= 0.\end{aligned}$$

https://en.wikipedia.org/wiki/Covariance#Uncorrelatedness_and_independence

Covariance Fun Facts

$$\mathbf{Cov}[\mathbf{X}] = E[(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^T] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

- True or False: If X and Y are independent then $\sigma_{xy} = 0$



https://en.wikipedia.org/wiki/Covariance#Uncorrelatedness_and_independence

Covariance Fun Facts

$$\mathbf{Cov}[\mathbf{X}] = E[(\mathbf{X}-\boldsymbol{\mu}_x)(\mathbf{X}-\boldsymbol{\mu}_x)^T] = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2{}_x & \sigma_{xy} \\ \sigma_{xy} & \sigma^2{}_y \end{pmatrix}$$

- True or False: If $\sigma_{xy} = \sigma_x \sigma_y$ then X and Y are deterministically related
- True or False: If X and Y are deterministically related then $\sigma_{xy} = \sigma_x \sigma_y$

How could you prove or disprove these?

General Covariance

Let $\boldsymbol{X} = (X_1, X_2, \ldots X_k)$ be a vector of k continuous random variables

$$\mathbf{Cov}[\,\mathbf{X}\,] = E[(\mathbf{X} - \boldsymbol{\mu}_x)(\mathbf{X} - \boldsymbol{\mu}_x)^T\,] = \boldsymbol{\Sigma}$$

$$\boldsymbol{\Sigma}_{ij} = Cov[X_i, X_j] = \sigma_{x_i x_j}$$

S is a k x k symmetric positive semi-definite (PSD) matrix

If all distributions are linearly independent it is positive definite

If the distributions are linearly dependent it has at least on zero eigenvalue

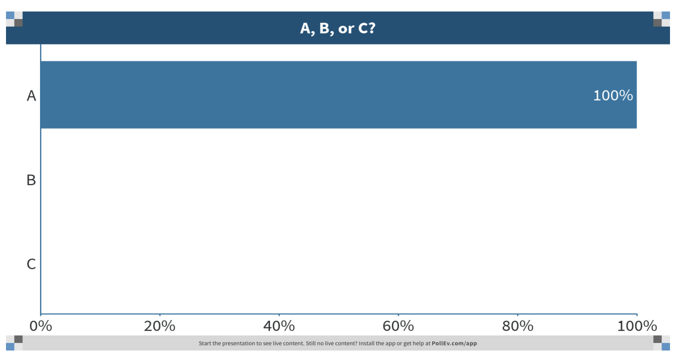
Test your understanding

Question : When (if ever) does $Var[X+Y] = Var[X] + Var[Y]$?

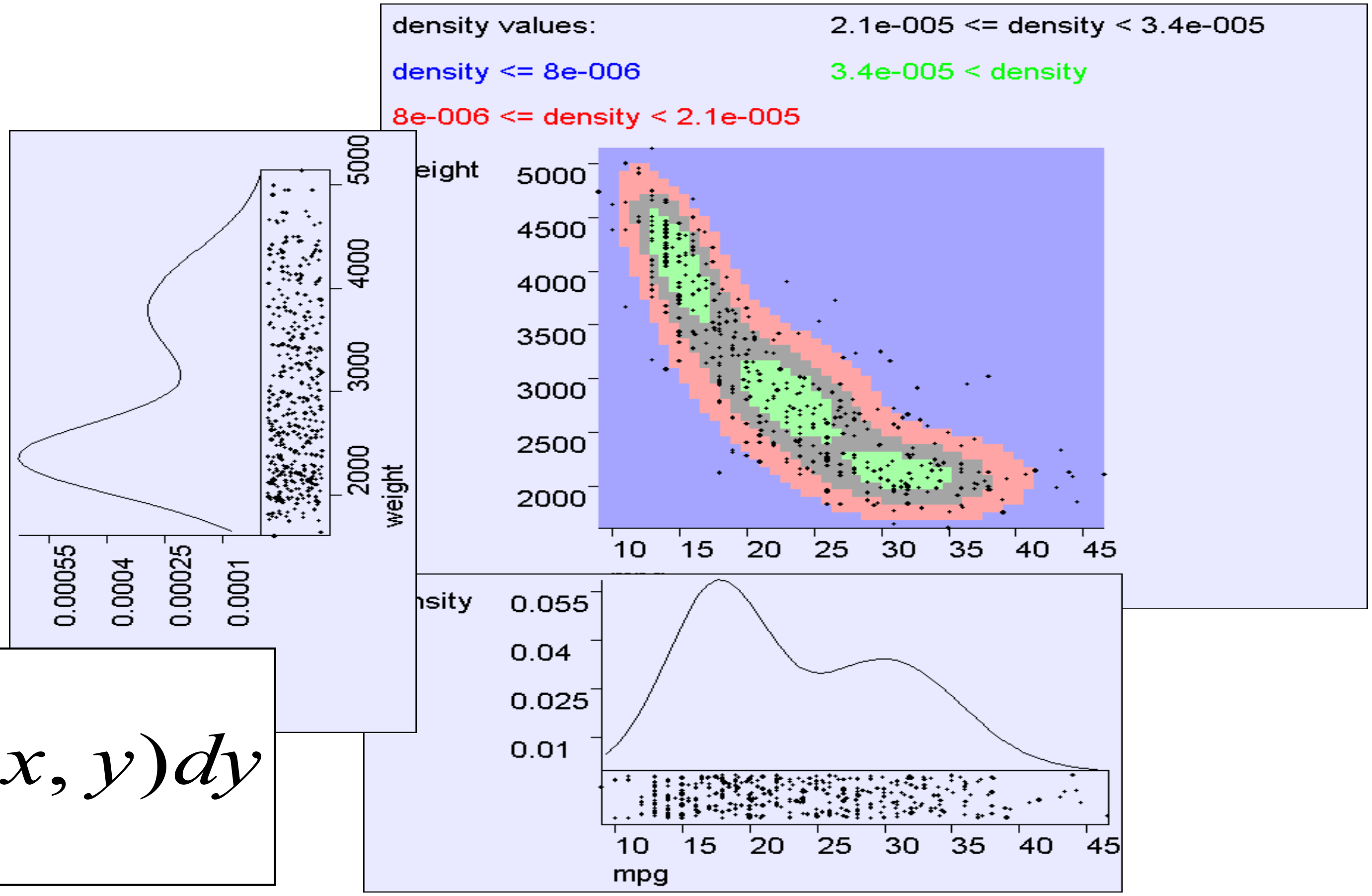
A) All the time?

B) Only when X and Y are independent?

C) It can fail even if X and Y are independent?

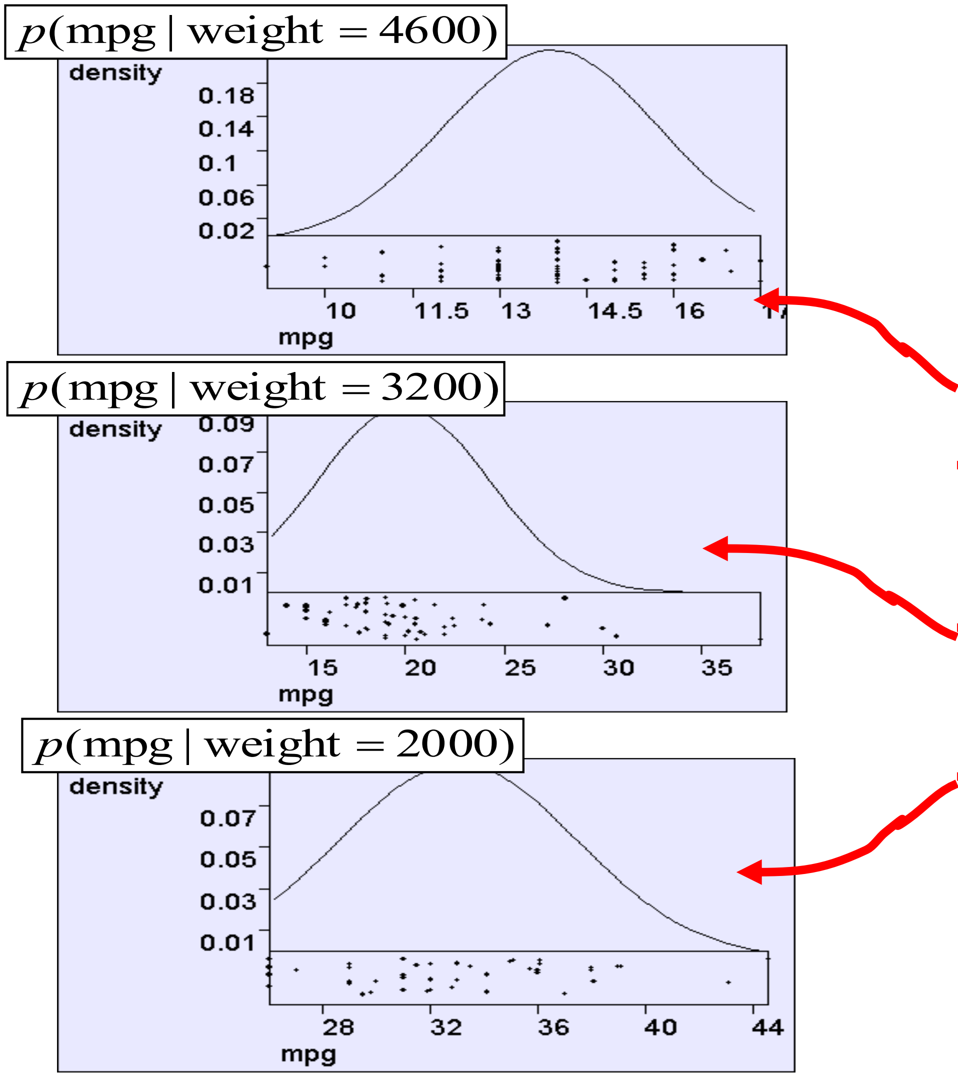


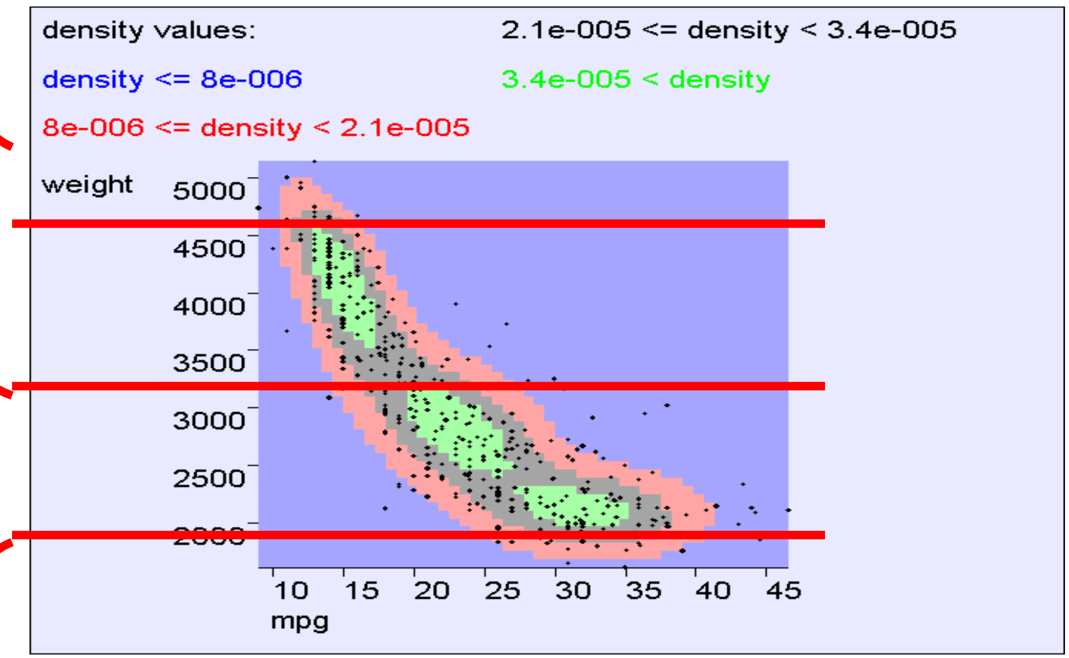
Marginal Distributions



$$p(x) = \int_{y=-\infty}^{\infty} p(x, y)\,dy$$

Conditional Distributions

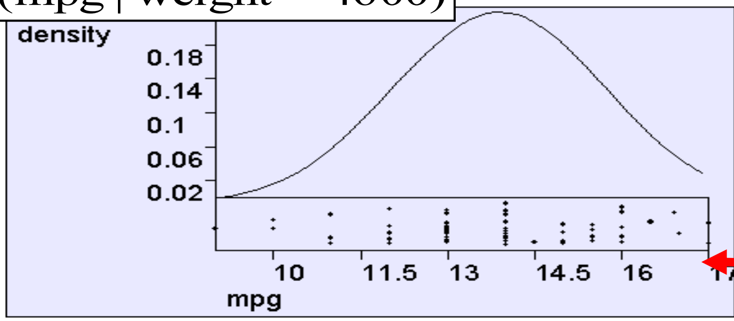


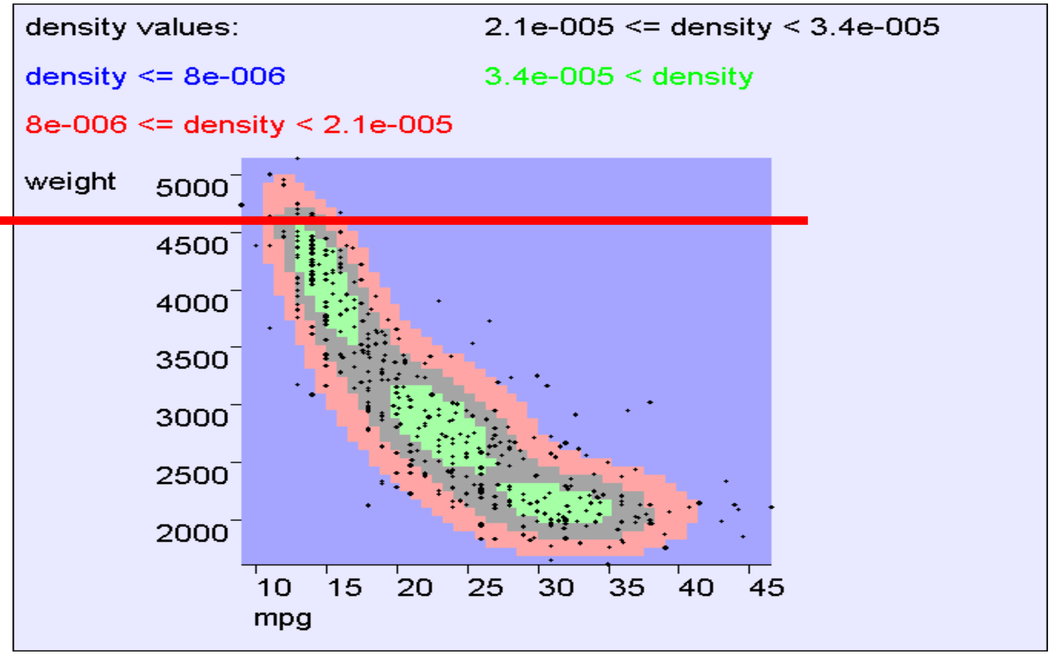


$$p(x \mid y) =$$
$$\text{p.d.f. of } X \text{ when } Y = y$$

Conditional Distributions

$p(\text{mpg} \mid \text{weight} = 4600)$





$$p(x \mid y) = \frac{p(x, y)}{p(y)}$$

Why?

$$p(x \mid y) =$$
$$\text{p.d.f. of } X \text{ when } Y = y$$

Independence Revisited

$$X \perp Y \text{ iff } \forall \mathrm{x}, \mathrm{y} : p(x, y) = p(x)p(y)$$

It's easy to prove that these statements are equivalent…

$$\forall \mathrm{x}, \mathrm{y} : p(x, y) = p(x)p(y)$$

$$\Leftrightarrow$$

$$\forall \mathrm{x}, \mathrm{y} : p(x \mid y) = p(x)$$

$$\Leftrightarrow$$

$$\forall \mathrm{x}, \mathrm{y} : p(y \mid x) = p(y)$$

More useful stuff

$$\int_{x=-\infty}^{\infty} p(x \mid y)\,dx = 1$$

(These can all be proved from definitions on previous slides)

$$p(x \mid y, z) = \frac{p(x, y \mid z)}{p(y \mid z)}$$

$$p(x \mid y) = \frac{p(y \mid x)\,p(x)}{p(y)}$$

Bayes Rule

Mixing discrete and continuous variables

$$p(x, A=v) = \lim_{h \to 0} \frac{P\left(x - \frac{h}{2} < X \leq x + \frac{h}{2} \wedge A = v\right)}{h}$$

$$\sum_{v=1}^{n_A} \int_{x=-\infty}^{\infty} p(x, A=v)\,dx = 1$$

$$p(x \mid A) = \frac{P(A \mid x)\,p(x)}{P(A)}$$

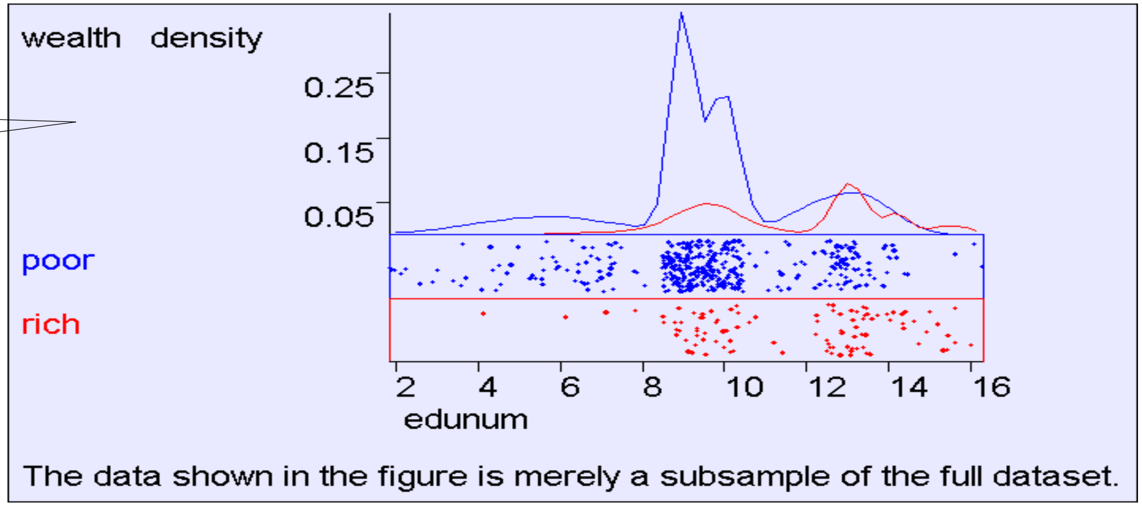
Bayes Rule

$$P(A \mid x) = \frac{p(x \mid A)\,P(A)}{p(x)}$$

Bayes Rule

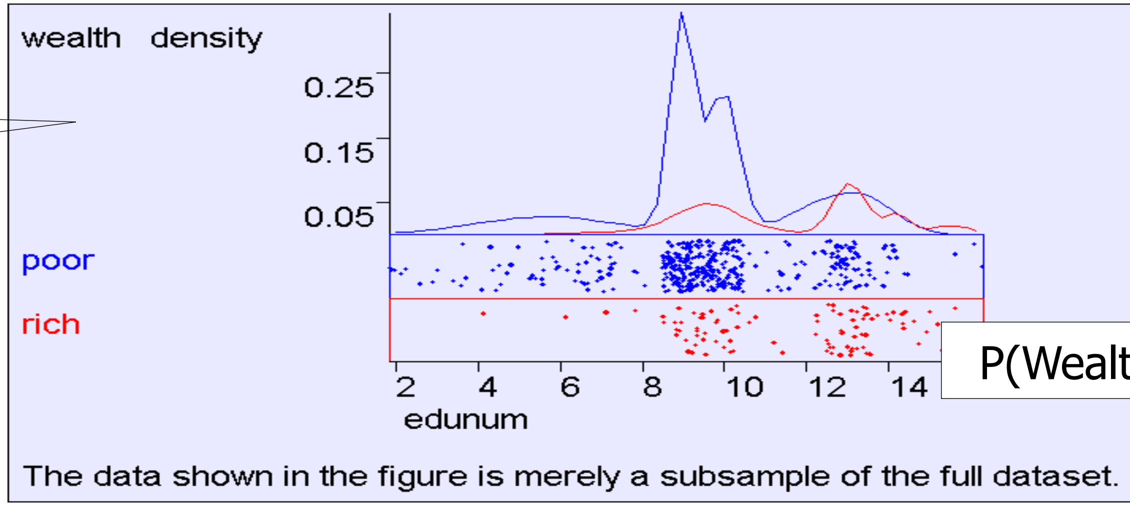
Mixing discrete and continuous variables

P(EduYears,Wealthy)

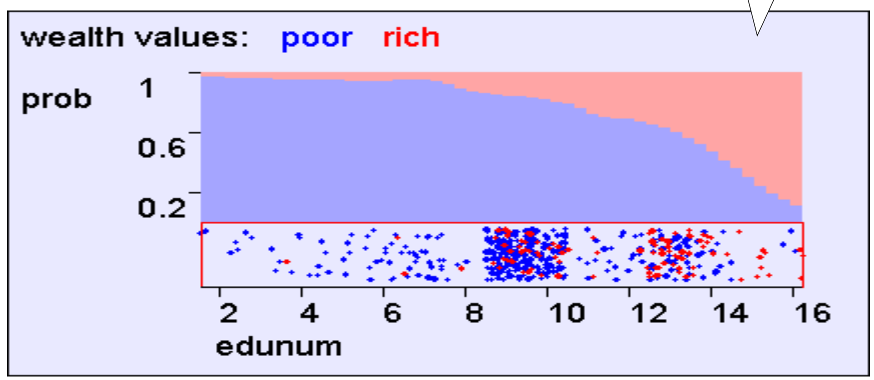


Mixing discrete and continuous variables

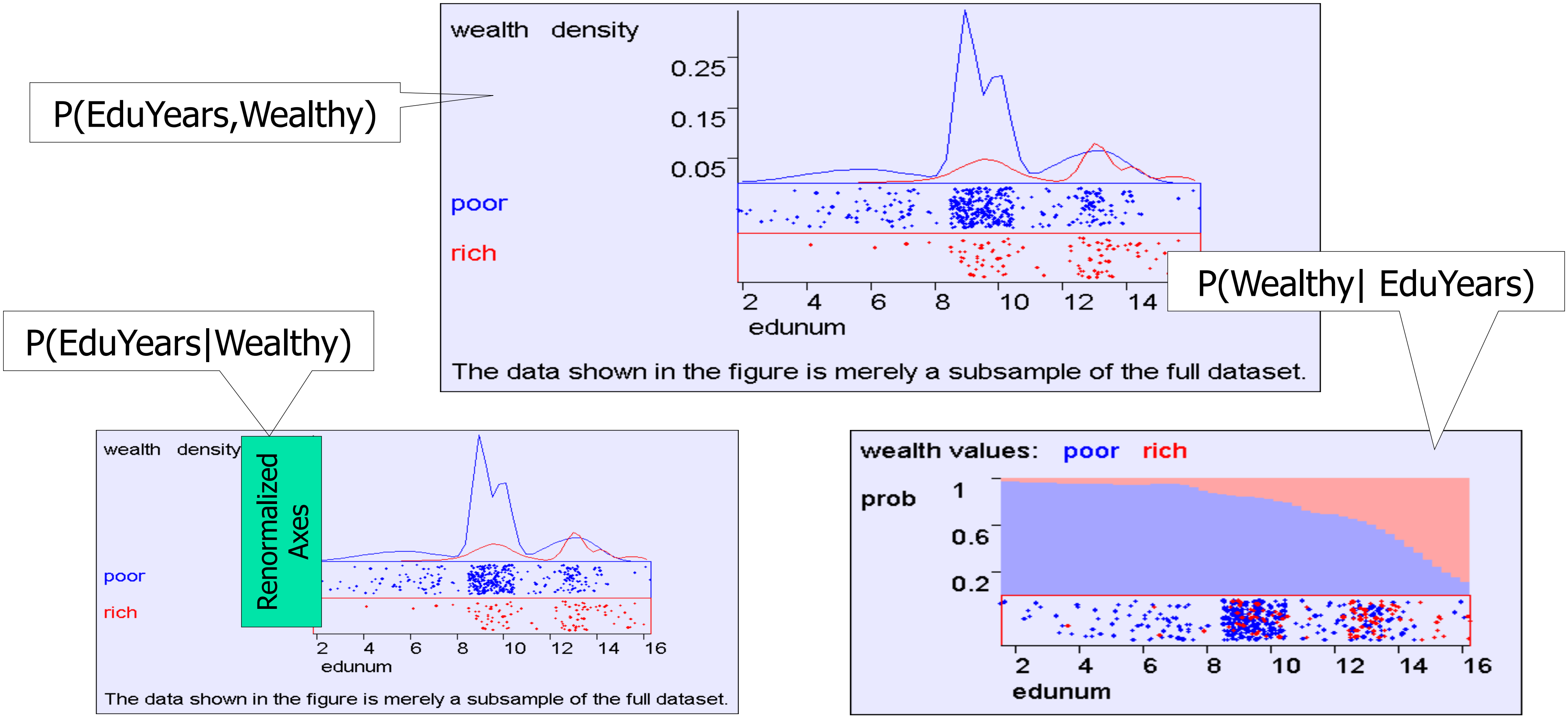
P(EduYears,Wealthy)



P(Wealthy| EduYears)



Mixing discrete and continuous variables



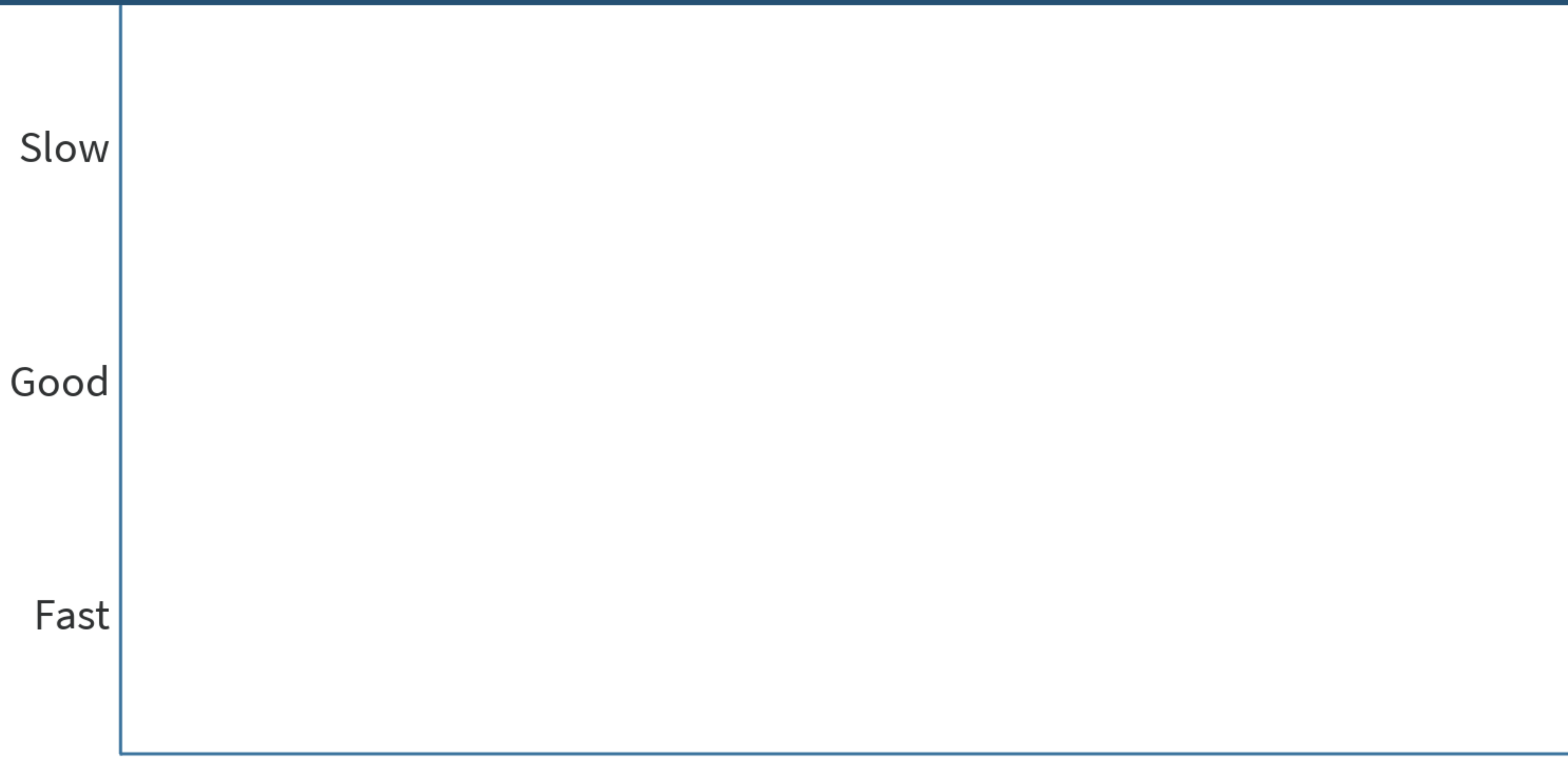
What you should know

- **You should**
  - be able to play with discrete, continuous and mixed joint distributions
  - be happy with the difference between p(x) and P(A)
  - be intimate with expectations, variance and covariance of continuous and discrete random variables
  - smile when you meet a covariance matrix
- **Independence and its consequences should be second nature**

What questions do you have on today's class?

Top

How is my speed?



Start the presentation to see live content. Still no live content? Install the app or get help at **PollEv.com/app**