Machine Learning
CS342

Lecture 9: Probability Theory Refresher

Dr. Theo Damoulas
T.Damoulas@warwick.ac.uk

Office hours: Mon & Fri 10-11am @ CS 307

Assignment is out!

Should be very straightforward to get 11/15 marks (73%)

Builds on Lab material directly, you have most of the functions needed

Remaining 4 marks will be for given for improved CV performances

Competitive element

Better understand your data and explore transformations of your data (feature expansions and feature engineering). Use CV to guide you

If struggling with Python speak to your Tutors for help

Questions?

selection/sampling bias

Feedback

Table 1

		Yes/About Right	Too Fast	No/Too Slow		
	Can you hear the lecturer?	40				
	Can you read the lecturer's handwriting/ Presentation Slides?	40				
	Is the rate of delivery…	32				
	Do the lectures seem well organised?	40				
	Has the lecturer made the academic objectives of the module clear?	39				
	Do you feel that your understanding is a sufficient grounding for this module?	37				
	Have you been informed of any relevant textbooks?	40				
	Have you been made aware of the assessment methods for this module?	40				
	Are you attending the support classes for this module (if provided)?	38 (95%)		2 (5%)		

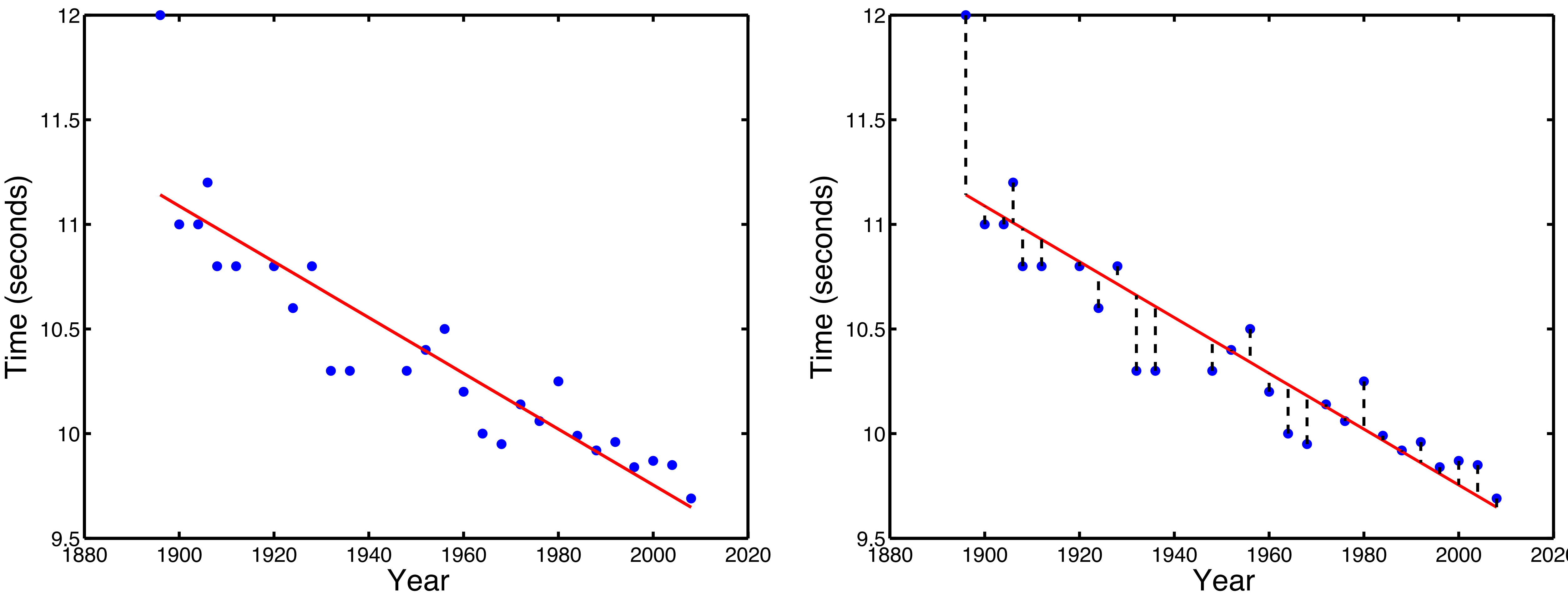
aha…

Department of COMPUTER SCIENCE

From error to noise and random variables

$$\hat{t} = \hat{w}_0 + \hat{w}_1 x$$ Equation of the fitted line

for n^{th} observation $\hat{t}_n = \hat{w}_0 + \hat{w}_1 x_n$ Point on the fitted line



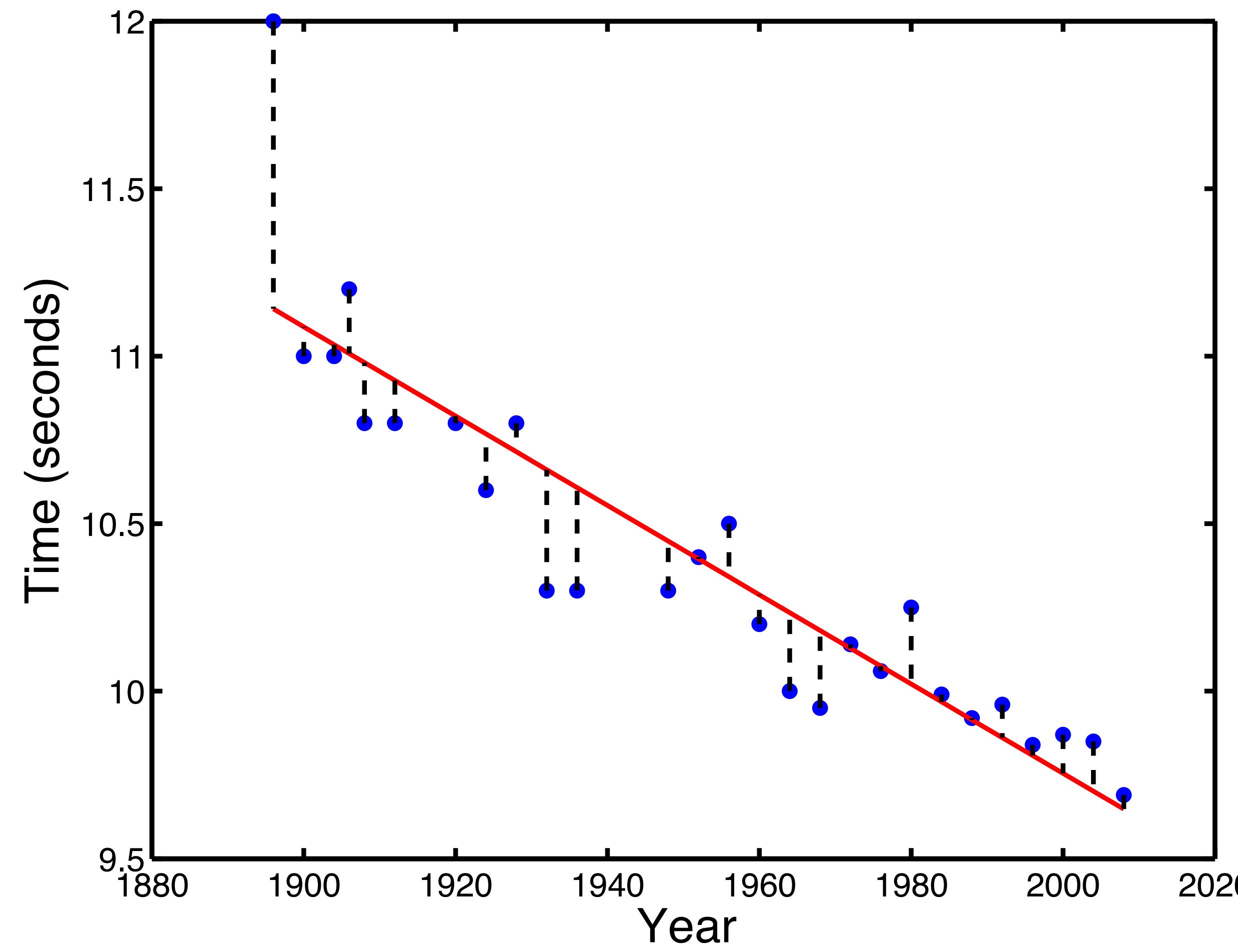
t_n output/target/response of n^{th} observation x_n

We will model the errors - think generatively

$$t_n = \hat{w_0} + \hat{w_1} x_n + \epsilon_n$$
$$= \hat{t}_n + \epsilon_n$$

Overfitting = fitting the noise

Imagine increasing model complexity



Noise appears both negative and positive
Seems different for each n
Does not seem to be a relationship between noise at different n
Looks very hard to model exactly (random…)

Random variables

Random variables example:

If I toss a coin and assign the variable X the value 1 if the coin lands heads and 0 if it lands tails, X is a random variable.

We don't know which value X will take but we do know the possible values x and how likely they are

Random events with outcomes that we can count: Discrete random variables

$$0 \leq P(X = x) \leq 1 \qquad \sum_{x} P(X = x) = 1$$

e.g. coin toss, rolling a dice, draw a card, number of emails per day

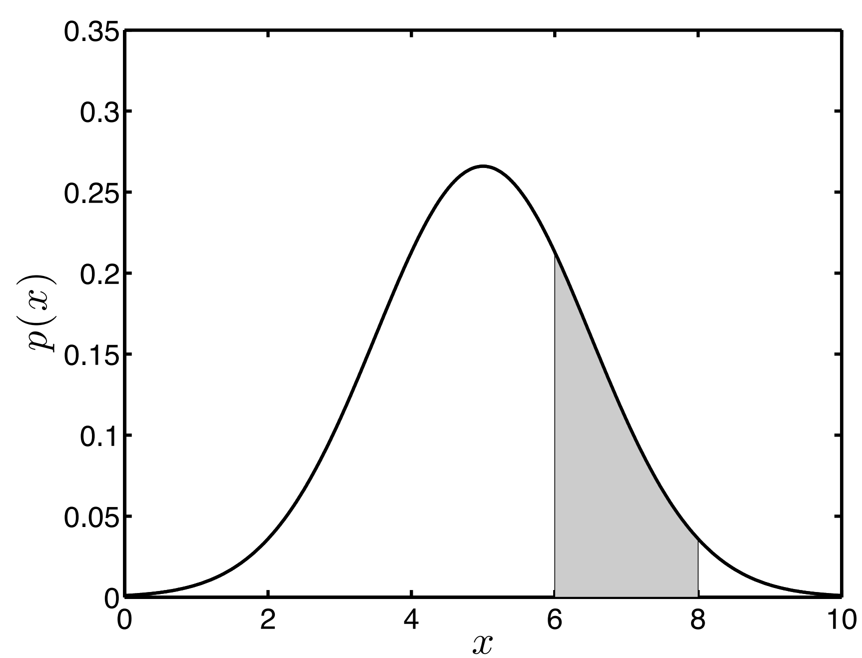
Random variables

Random events with outcomes that we cannot count:
Continuous random variables (RVs)

e.g. Winning time in Olympic dataset, noise in our data

Can't write down a probability for an event, since we cannot count them

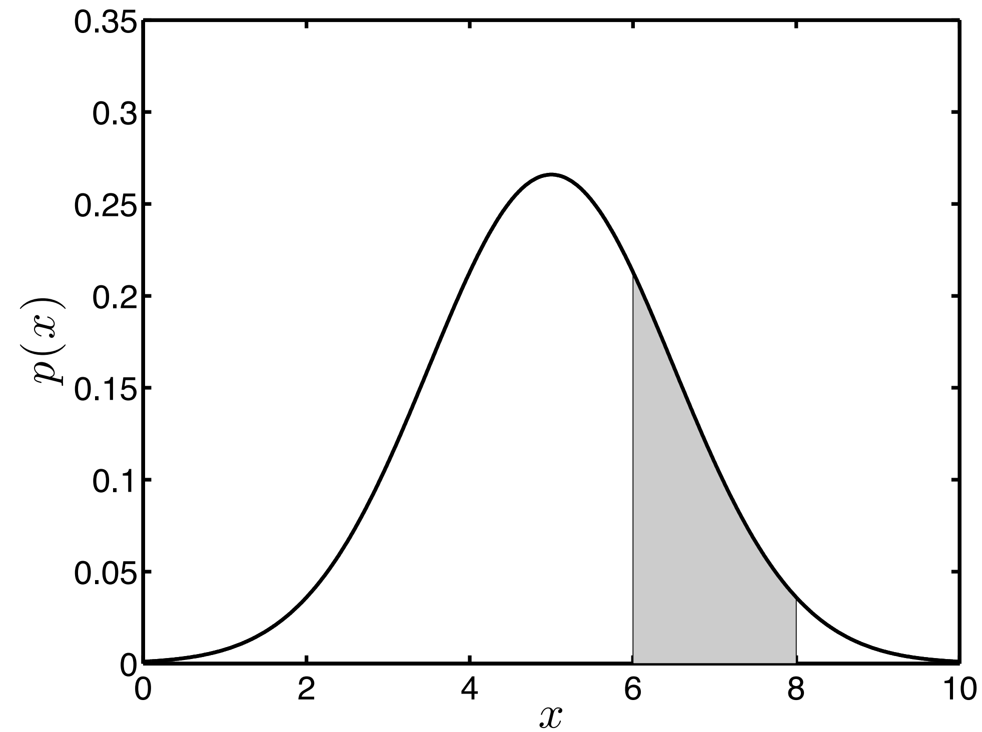
Instead we define a density function p(x)



Note: Lowercase p(x) for continuous RVs vs P(X=x) for discrete RVs

Department of COMPUTER SCIENCE

Random variables



p(x) is not the probability of x
(infinite x values
so would not make sense)

It is a **density** function

If you want the probability of a range of x values (e.g. 6 to 8), it is the area under the density function

$$P(6 \leq X \leq 8) = \int_{x=6}^{x=8} p(x)dx$$

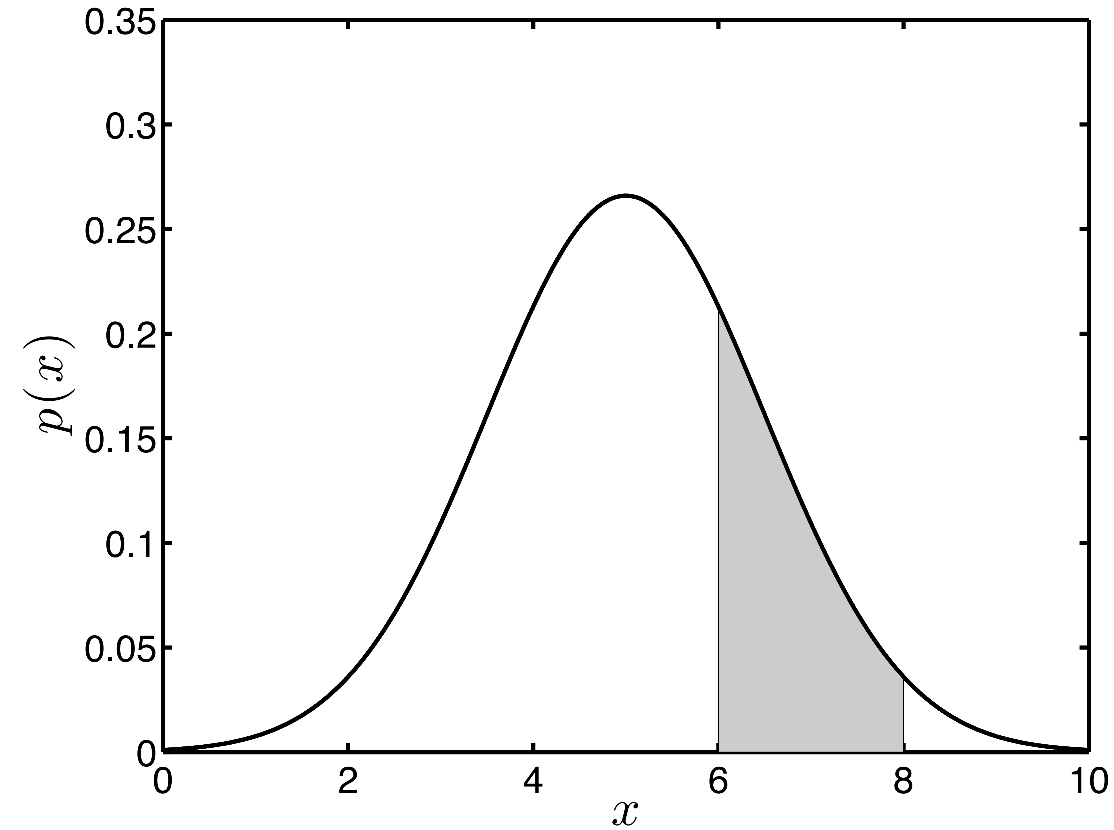
So if I consider the whole range of possible values of X, what should be the probability that X takes a value x in there? It should be 1

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

Probability Density functions

Also: $p(x) \geq 0$

If density negative then I would have negative probabilities



This is the pdf for the Gaussian distribution in 1-D (univariate)

$$p(x) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Integrate to get 1!

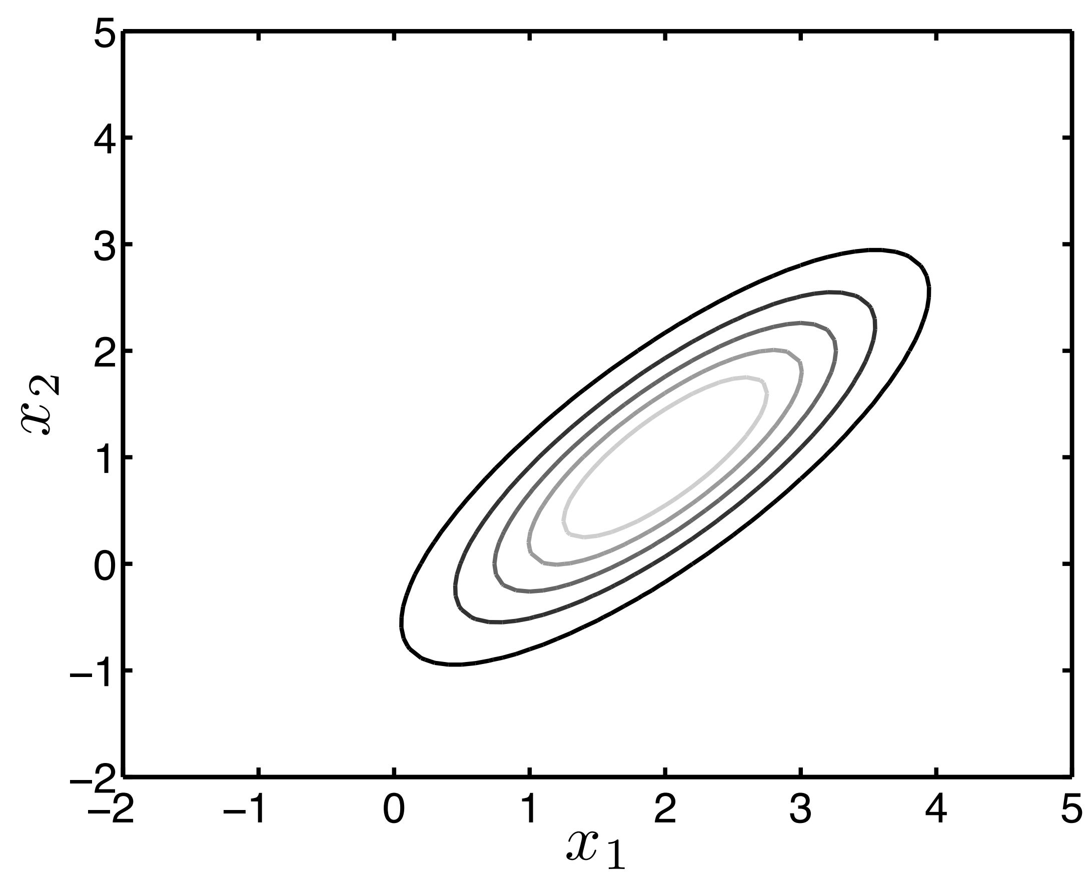
Normalising Constant

Bell-shape

Joint Probabilities and Densities

Joint probabilities: For two discrete RVs, X and Y, P(X=x, Y=y) is the probability that RV X has value x **and** RV Y has value y

Joint densities: For two continuous RVs, x_1 and x_2, the joint density is given by $p(x_1,x_2)$. If I integrate over ranges of values for x_1, x_2 I will get the probability that the RVs values both fall into those ranges



Independence

Let X be a discrete RV that describes a coin toss.
X=1 is heads and X=0 denotes tails.

Let Y be another discrete RV that describes rolling a fair dice.
Y=1 is rolling 1 and Y=6 is rolling a 6.

The joint distribution P(X=1, Y=5) describes the probability
that the coin will be heads **and** the dice will be 6

I could assume that tossing a coin and rolling a dice is independent (i.e. the outcome of the coin toss tells me nothing about the outcome of the dice roll)

Then **Independence** means: $P(X=1, Y=5) = P(X=1)*P(Y=5)$

Dependence and Conditional Probability

Dependence means: P(X=1, Y=5) does not equal P(X=1)*P(Y=5)

$$P(X = 1, Y = 5) \neq P(X = 1)P(Y = 5)$$

Because knowledge of one event tells me something about the other

Lets assume that Y=5 (rolling a 5) tells me something about flipping heads X=1. Dependent so use **Conditional Probability**

$$P(X = 1|Y = 5) \neq P(X = 1)$$

and the joint probability

$$P(X = 1, Y = 5) = P(X = 1|Y = 5)P(Y = 5)$$

What if X=1 is telling me something about Y=5?

Conditioning - continuous

Gaussian Likelihood for linear regression

$$p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \mathcal{N}(\mathbf{x}_n\mathbf{w}, \sigma^2)$$

The density of t_n conditioned on specific values for x and model parameters

$$P(0.3 \leq t_n \leq 1|\mathbf{x}_n, \mathbf{w}, \sigma^2)$$

This is the probability of t_n falling in that range given values for x and model parameters

Summary:

We should model the noise

We can model it as a random variable

It is continuous so we will choose a density function (pdf)

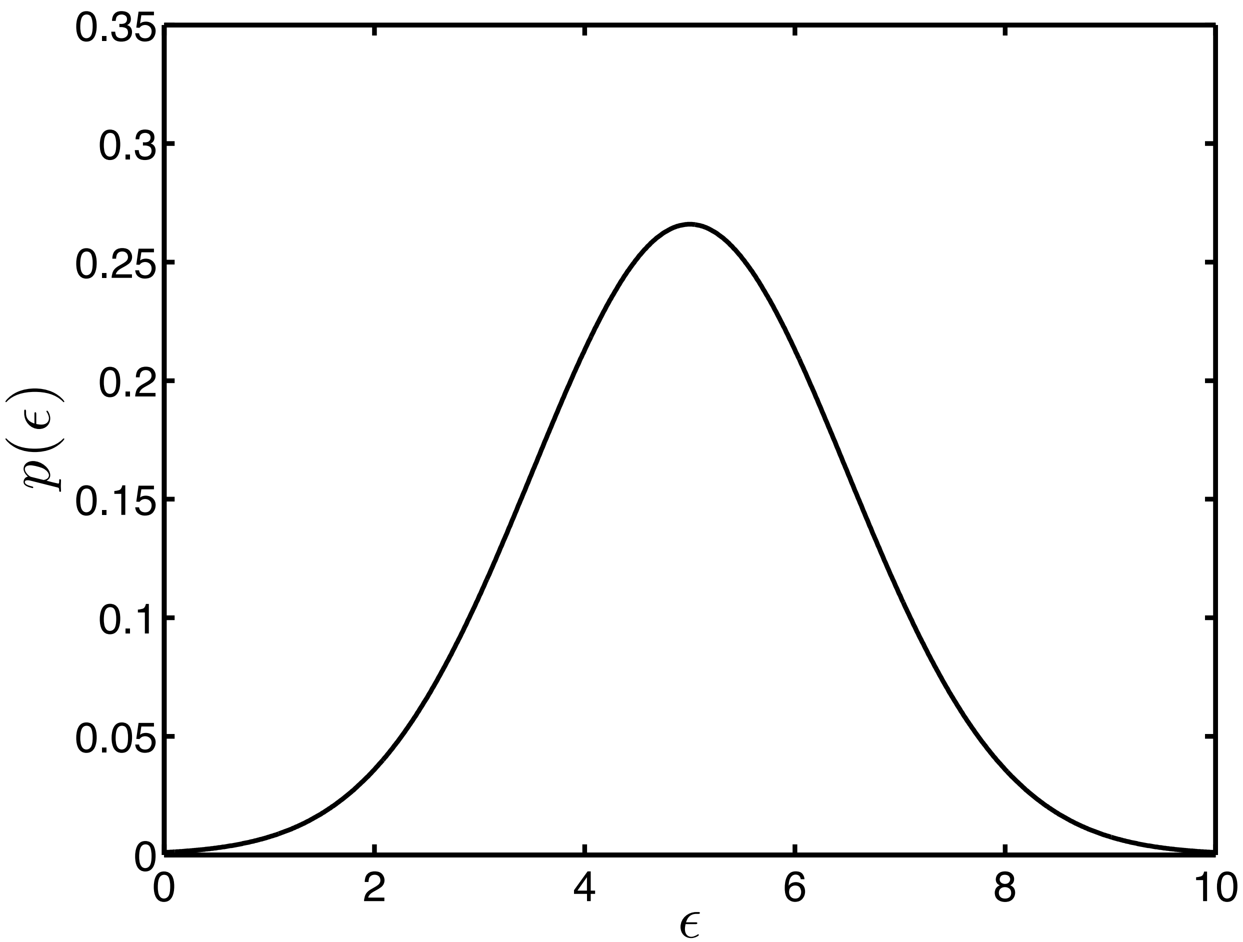
The Gaussian pdf seems reasonable due to our noise observations

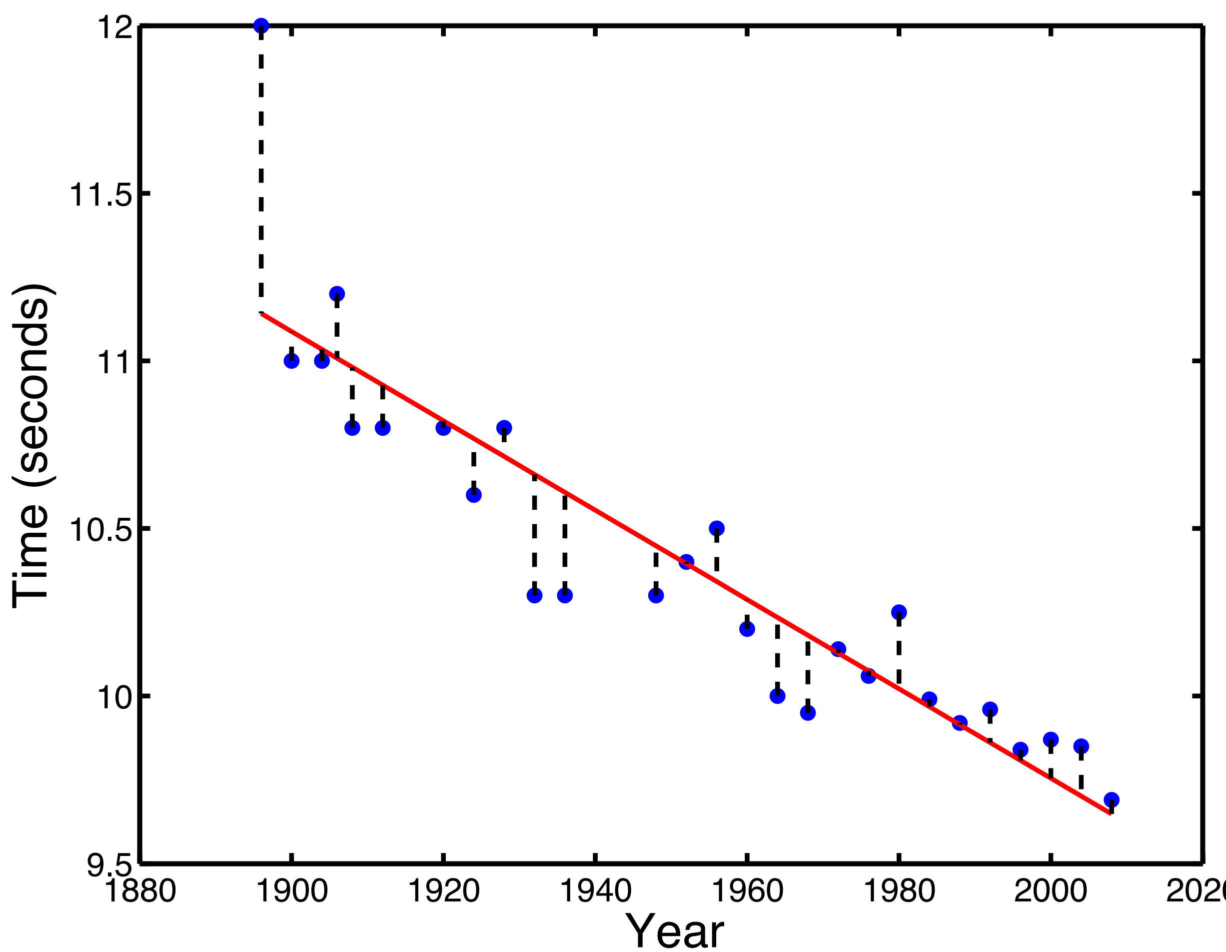
Hence t_n is a random variable too

Department of COMPUTER SCIENCE

Back to our model and noise term

$$t_n = \mathbf{x}_n \mathbf{w} + \epsilon_n$$



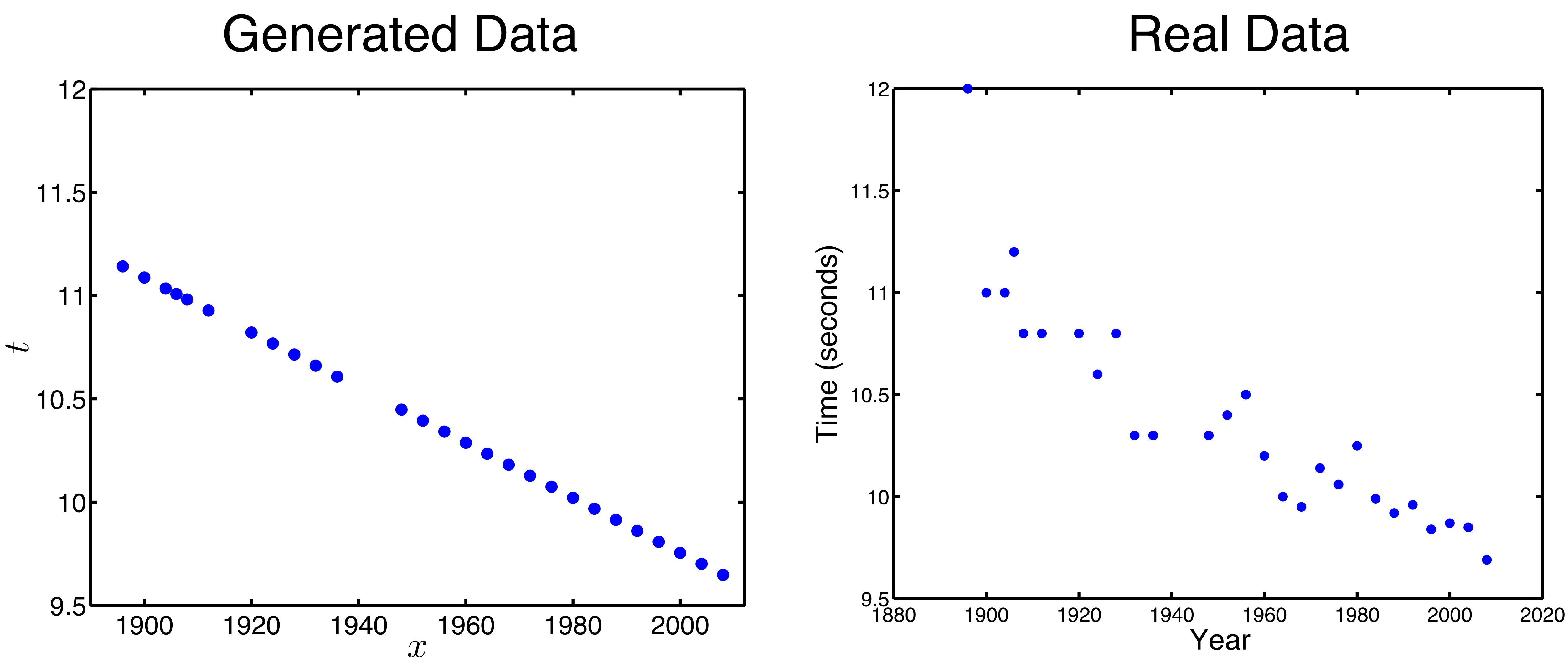


Noise appears both negative and positive

Seems different for each n

Does not seem to be a relationship between noise at different n

Generating data

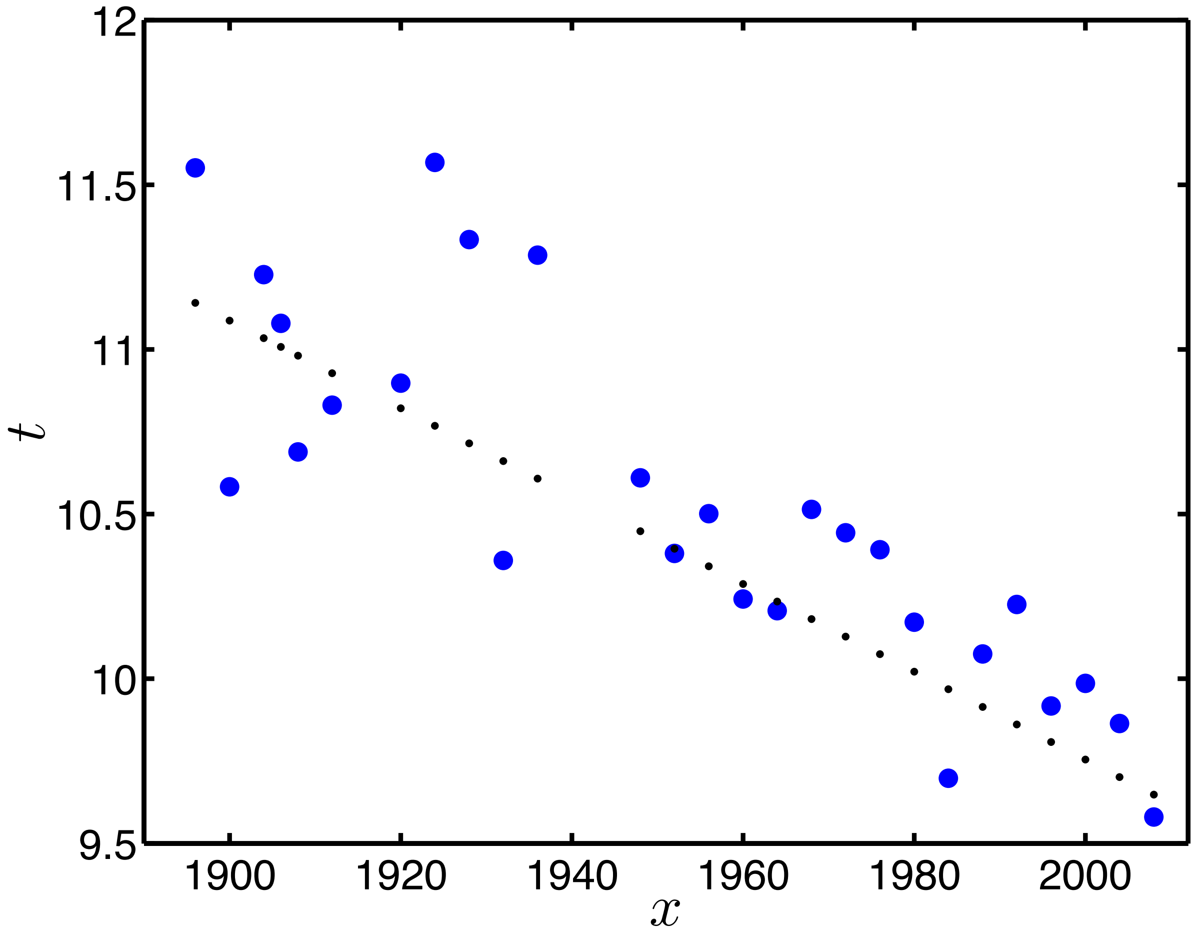


Fix some w

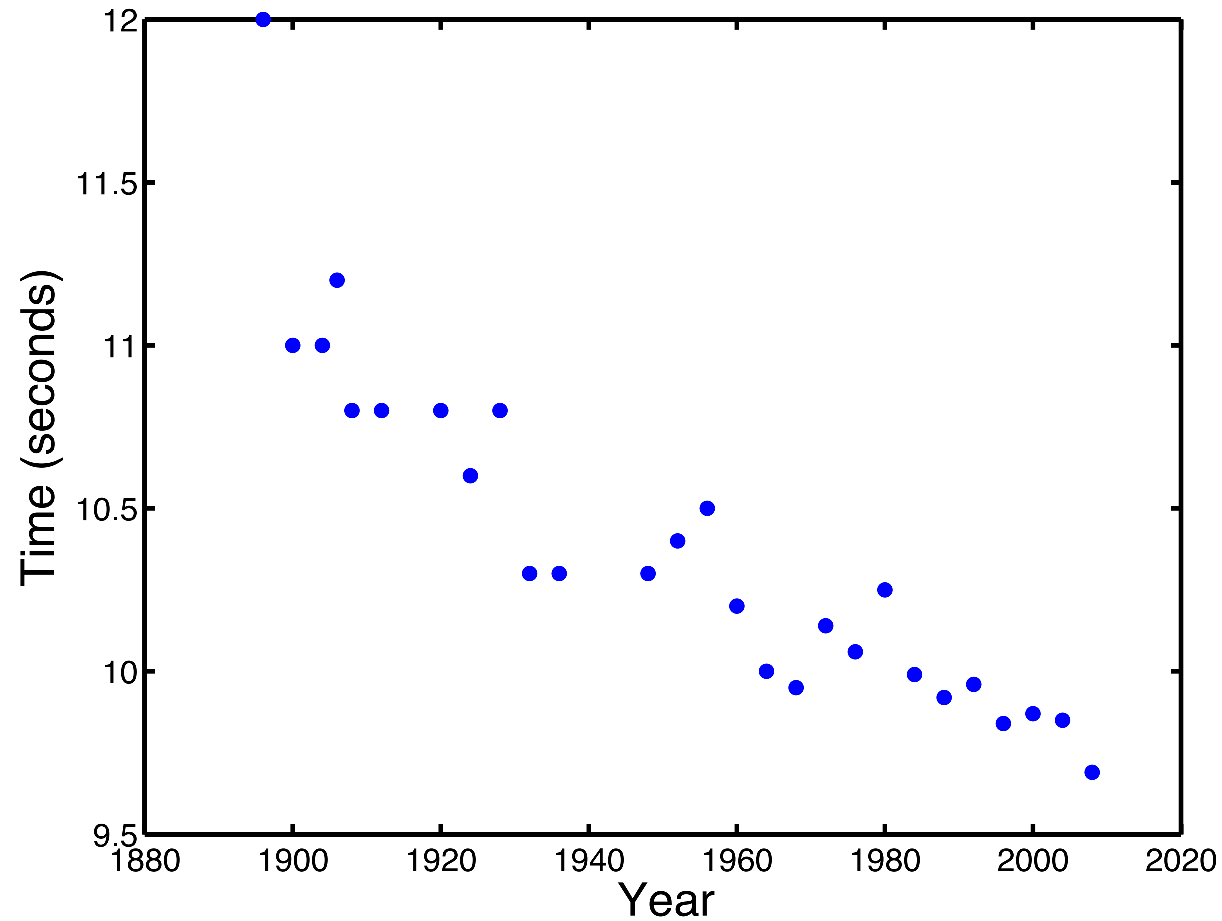
For a set of x values, compute wx

Generating data

Generated Data

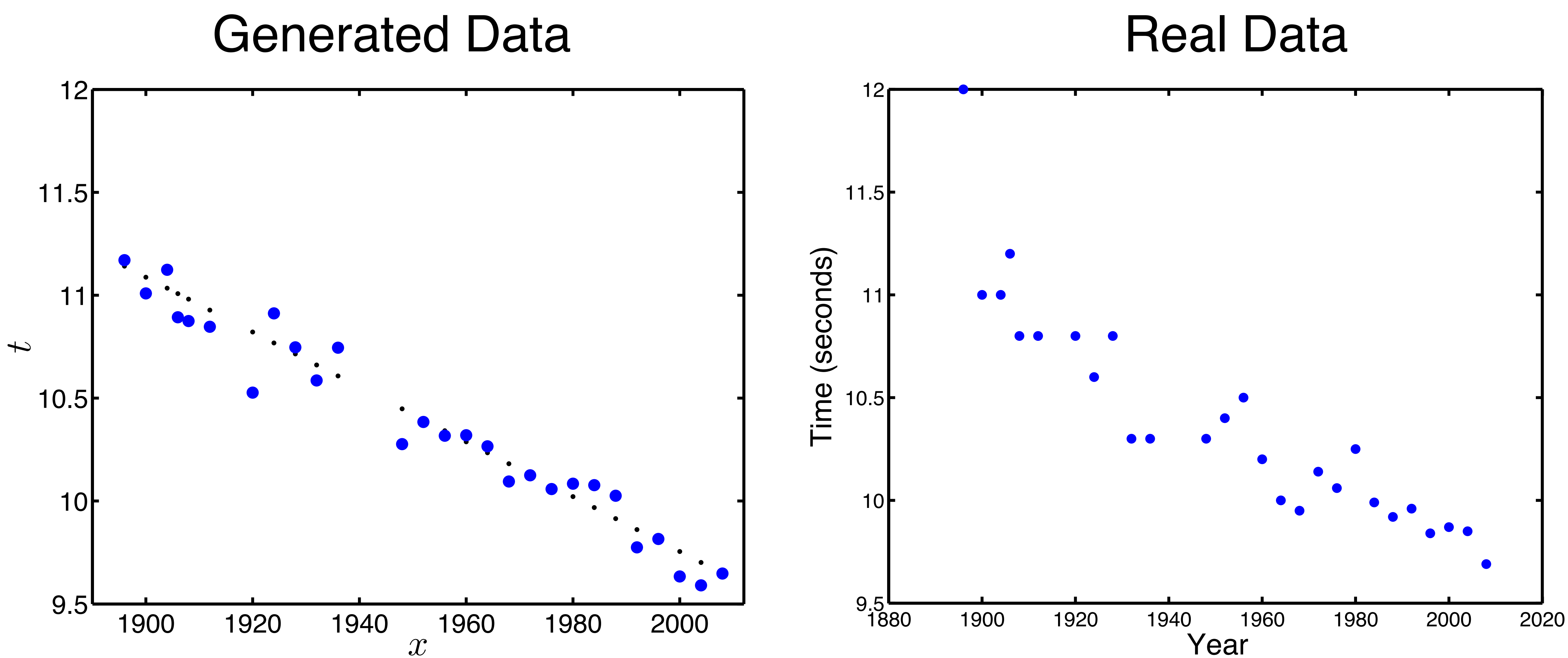


Real Data



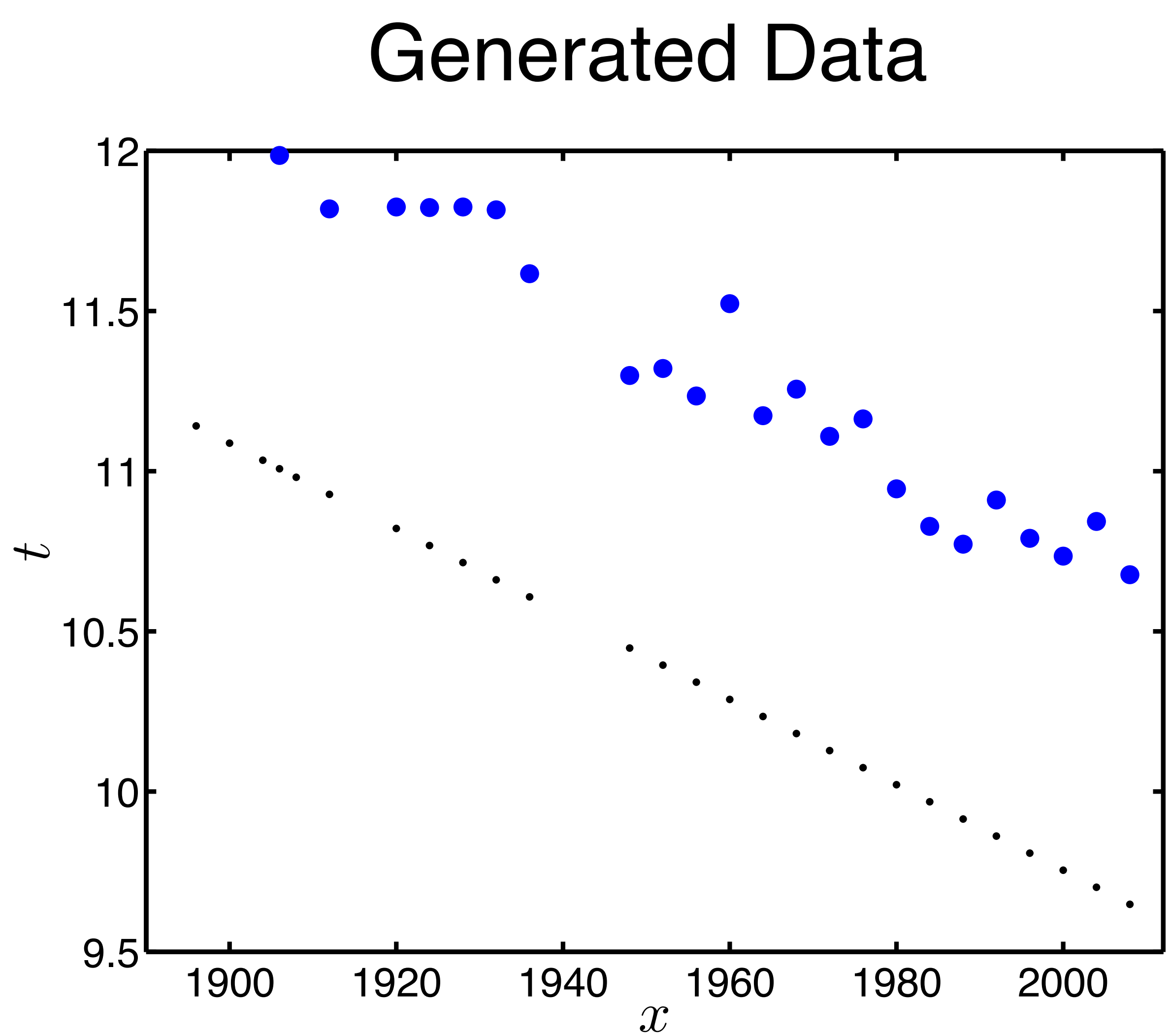
Sample noise RV from Gaussian distribution
with some mean and variance (0, 0.05)
Add to wx

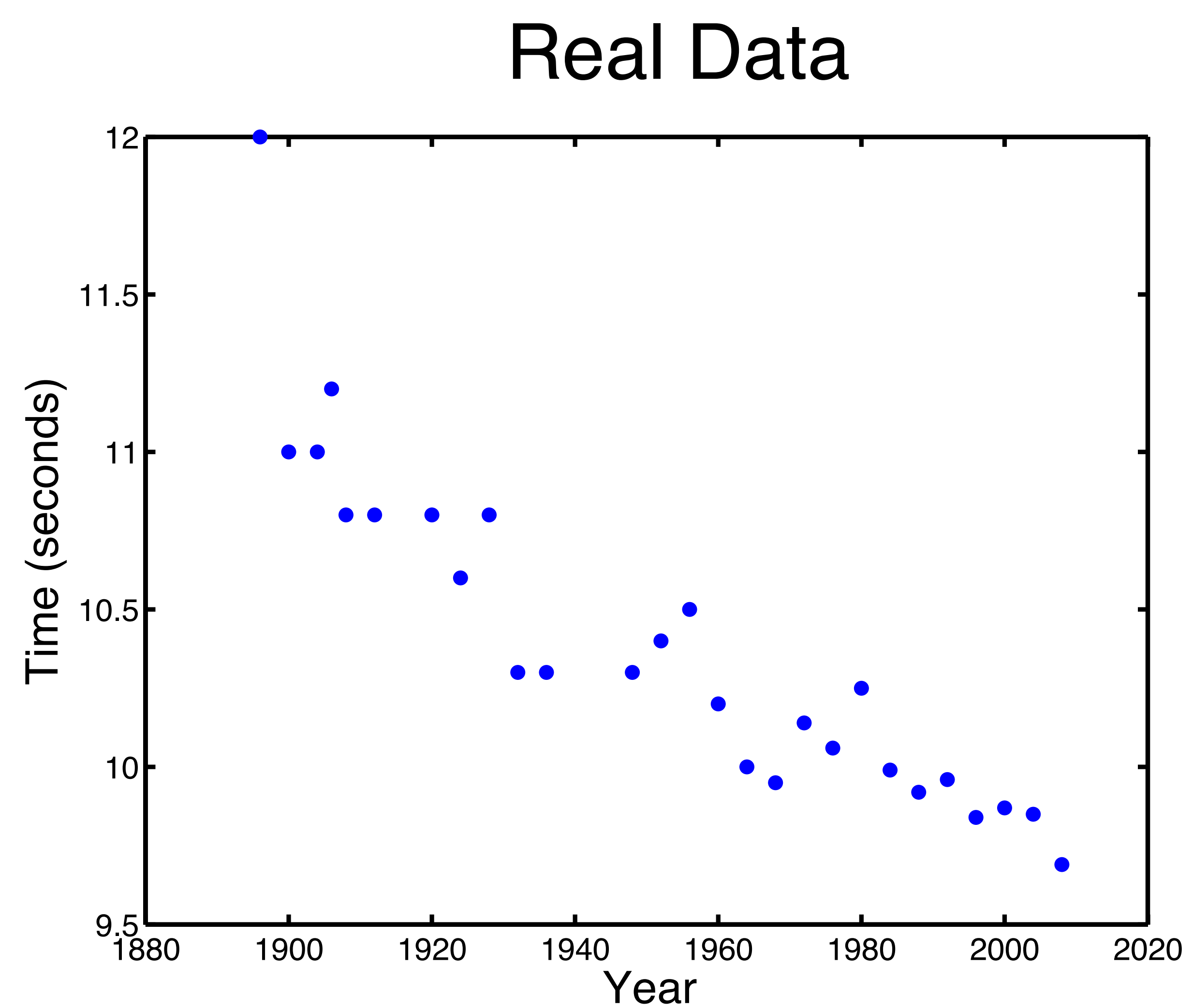
Generating data



Same but with mean, variance = (0, 0.01)

Generating data



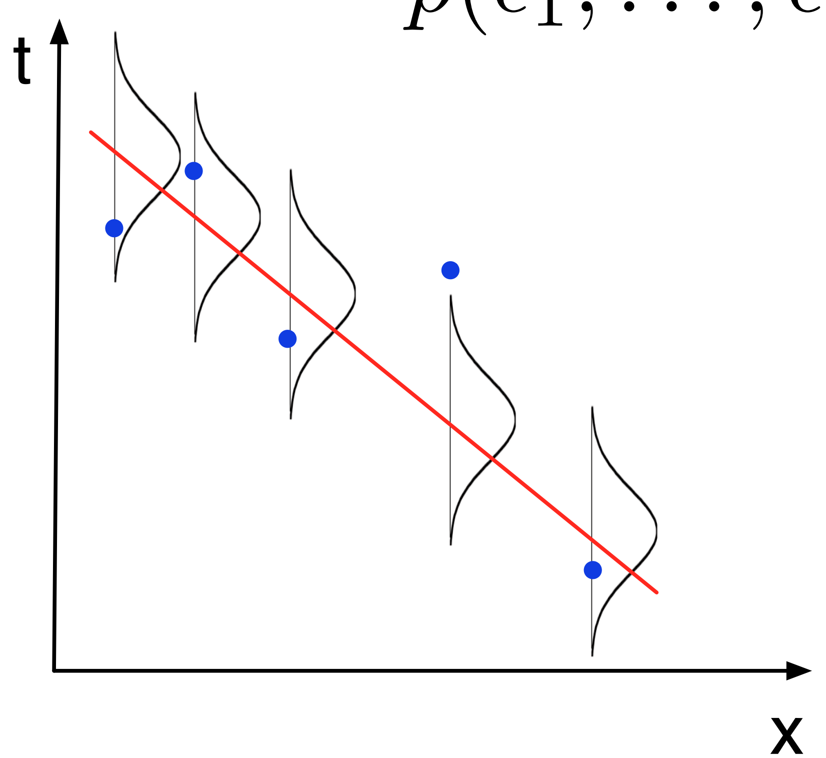


Try a mean not equal to 0, in this case (1, 0.01)
Nope.. mean to 0 seems like a good bet
We will learn the variance as a parameter

Recap: Maximum Likelihood

Assume that noise RVs are *independent* and *homoscedastic*:

$$p(\epsilon_1, \dots, \epsilon_N) = \prod_{n=1}^{N} p(\epsilon_n) = \prod_{n=1}^{N} \mathcal{N}(0, \sigma^2)$$



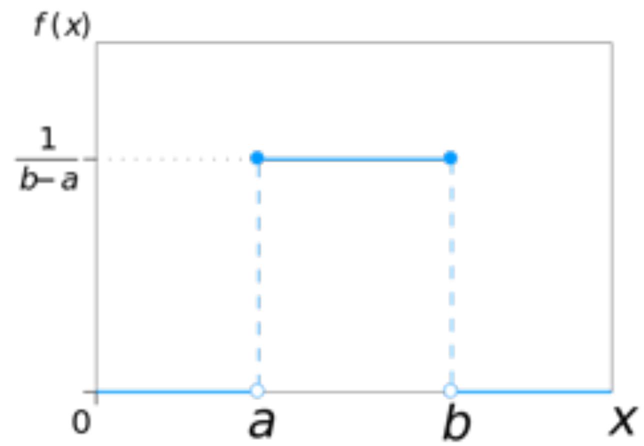
Likelihood

$$\boxed{p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n \mathbf{w}, \sigma^2)}$$

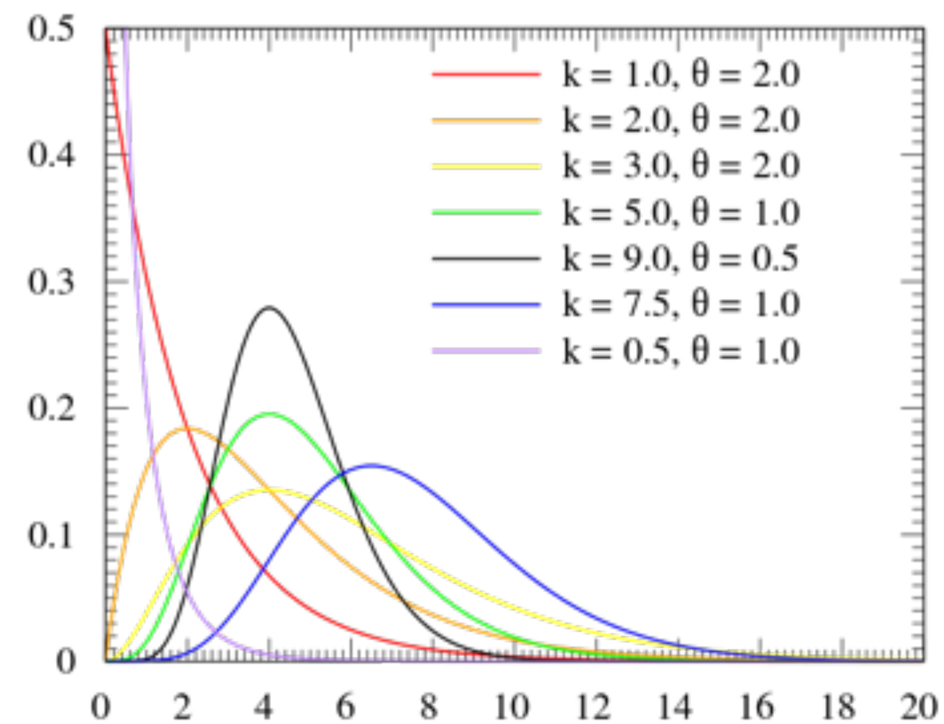
$$\mathbf{w}, \sigma \leftarrow \underset{\mathbf{w},\sigma}{\operatorname{argmax}} \; \log \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n \mathbf{w}, \sigma^2)$$

Other famous probability density functions

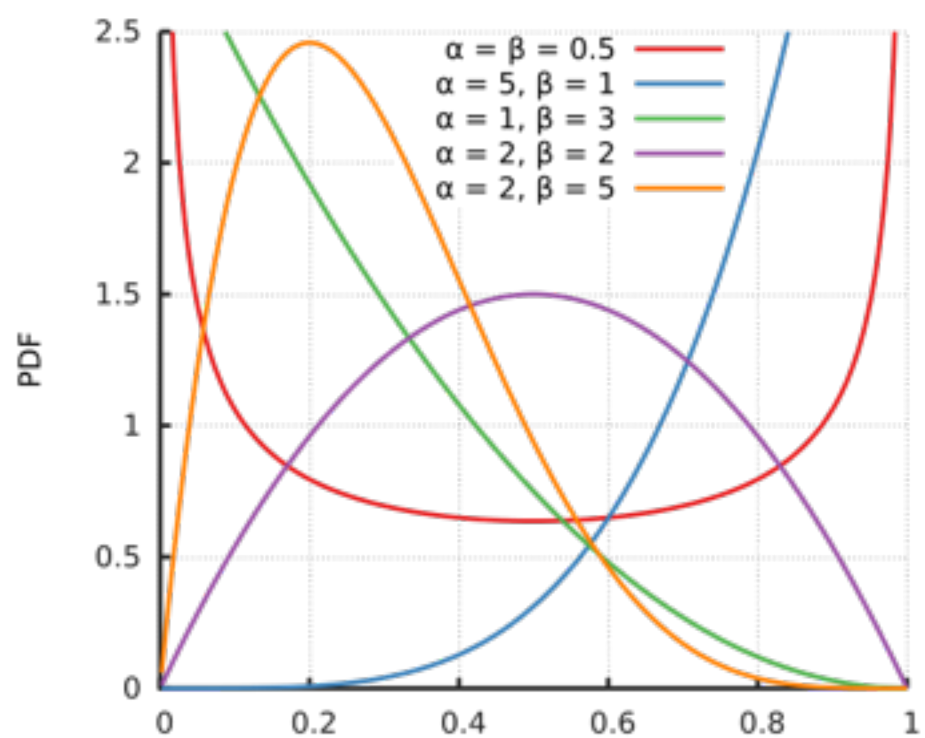
Uniform distribution



Gamma distribution

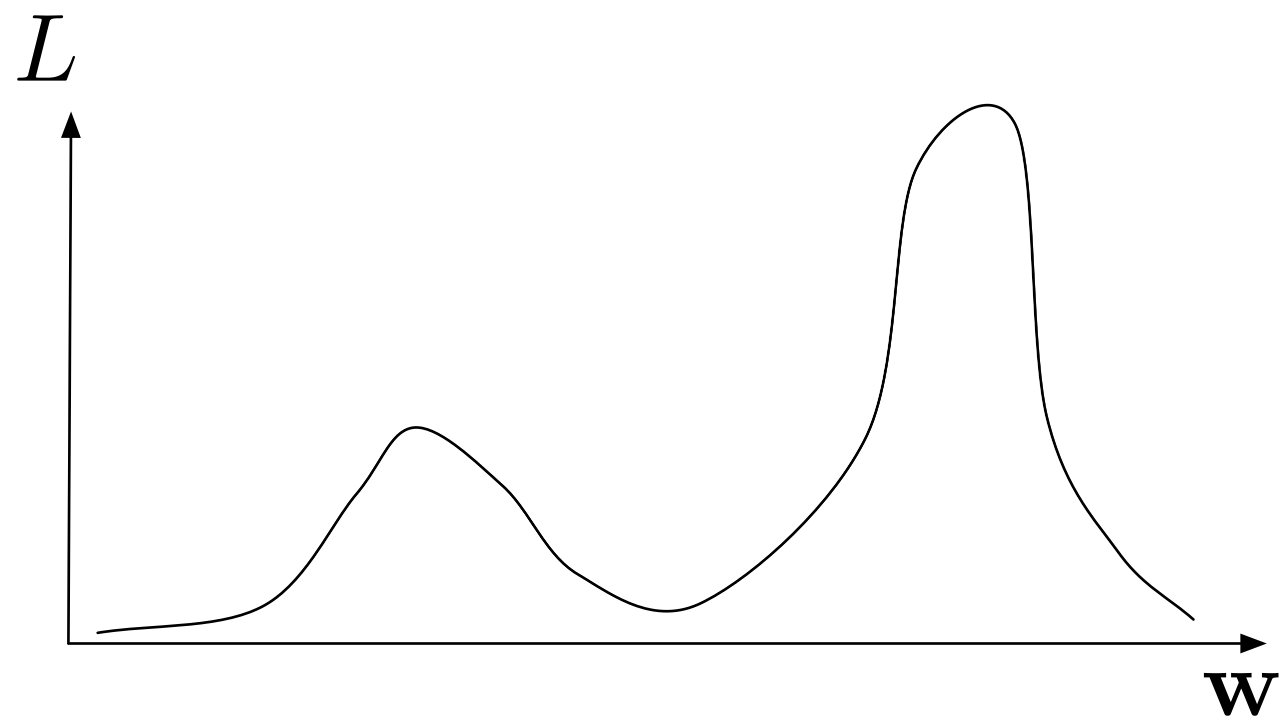


Beta distribution



The road to Bayes (full probabilistic inference)

So far we are finding the "best" parameters (Loss/Likelihood)
Is there one best parameter?



Might be more then one "best" parameter
Different values might give different predictions
How many values are "best" might be telling us something…
Uncertainty? Evidence?
Parameters as random variables! Place distributions…