Computational Pattern Analysis and Statistical Learning

Lecture 0.3: Basics of Probability & Statistics

Konstantin Tretyakov, Tijl de Bie

November 7, 2006

What do we know

- Pattern analysis is important and interesting.
- The general idea:

  Search the space of patterns for the most interesting one with respect to the data.
- Space of patterns $\Leftrightarrow$ Problem of interest (regression/classification/clustering/outliers/...)
- Searching $\Leftrightarrow$ Optimization
  - $\nabla f(\mathbf{x}) = \mathbf{0}$
  - $\nabla L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$

What do we know

- Pattern analysis is important and interesting.
- The general idea:

  Search the space of patterns for the most interesting one with respect to the data.
- Space of patterns $\Leftrightarrow$ Problem of interest (regression/classification/clustering/outliers/...)
- Searching $\Leftrightarrow$ Optimization
  - $\nabla f(\mathbf{x}) = \mathbf{0}$
  - $\nabla L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{0}$
- "Interesting patterns" $\Leftrightarrow$ Statistics.

Today

1. Probability Theory
   - Random Variables
   - Probability Distributions
   - Conditional Distribution. Bayes Rule. Independence
2. Statistics
   - Hypothesis Testing
   - Confidence Intervals
   - Model estimation
   - Consistency

What is your height?

What is your height?

Is it a fixed number?

What is your height?

Is it a fixed number?

- Frequentist: Yes, it is. But we can't measure it exactly.
- Bayesian: No, it is not. It's a *distribution*.

In any case:

> It's inconvenient to model reality with fixed numbers because **randomness** is an inherent property of nature.

What is your height?

- Is it between 1m and 2m?

What is your height?

- Is it between 1m and 2m?
- Is it between 1.70 and 1.71?

What is your height?

- Consider all statements of the form *your height* $\underline{H}$ *is* $a \leq \underline{H} < b$.
- Suppose for each such statement S we can state our *confidence* $\mathbf{P}_H(S)$ that it holds.
- It's natural to require that
  - $\mathbf{P}_H(S) \geq 0$ for any S.
  - $\mathbf{P}_H(S) = 1$ if S is "100% true".
  - $\mathbf{P}_H(S_1 \text{ or } S_2) = \mathbf{P}_H(S_1) + \mathbf{P}_H(S_2)$

    if S_1 and S_2 are *mutually exclusive*.
- $\mathbf{P}_H$ thus describes *all* the information we have about H.
- We call $\underline{H}$ a *random variable* and $\mathbf{P}_H$ its *probability distribution*.

CDF & PDF

- It turns out we can always represent $\mathbf{P}_H(a \leq \underline{H} < b)$ as

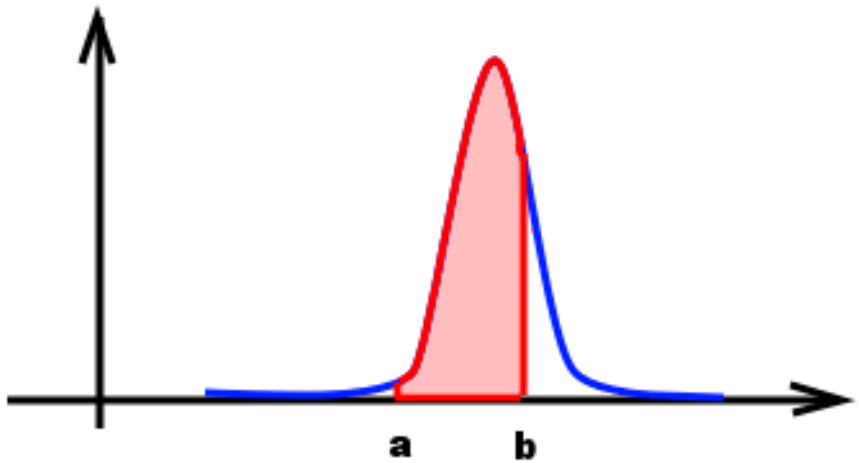
$$\mathbf{P}_H(a \leq \underline{H} < b) = F(b) - F(a)$$

  for some nondecreasing $F : \mathbf{R} \to \mathbf{R}$.
- F is called the *cumulative distribution function (CDF)* of H.
- $F(x) = \mathbf{P}_H(\underline{H} < x)$

CDF & PDF

- Moreover, in most cases F is differentiable and then:

$$\mathbf{P}_H(a \leq \underline{H} < b) = \int_a^b f(x)\,\mathrm{d}x$$



where $f(x) = F'(x)$.

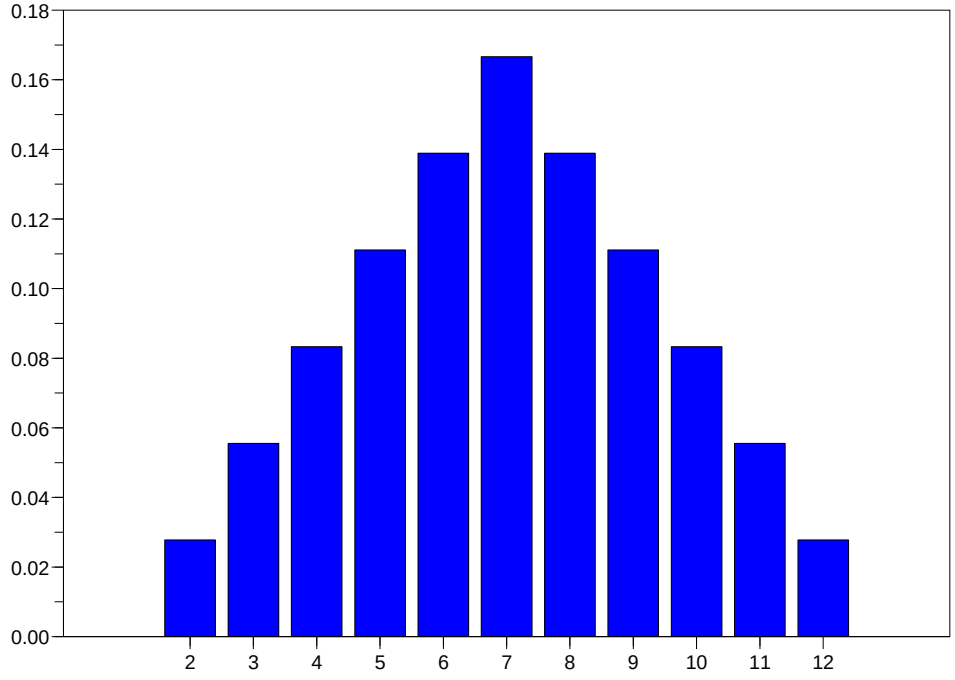
- f is called the *probability density function (PDF)* of $\underline{H}$.

The Discrete Case

- I’ve thrown a pair of dice. What number N did I get?

The Discrete Case

- I’ve thrown a pair of dice. What number N did I get?



The Discrete Case

- Similarly we can speak about *CDF*:

$$F(x) = \mathbf{P}(\underline{N} < x) \approx \frac{\text{number of throws with } \underline{N} < x}{\text{number of throws}}$$

- and *PDF*:

$$f(x) = \mathbf{P}(\underline{N} = x) \approx \frac{\text{number of throws with } \underline{N} = x}{\text{number of throws}}$$

Describing Distributions

Let $\underline{X}$ be a random variable with a density function $f(x)$. Then

- Mean of $\underline{X}$:

$$\mu = \mathbf{E}(\underline{X}) = \int_x x f(x)\,\mathrm{d}x = \sum_x x\mathbf{P}(\underline{X} = x) \approx \frac{\sum_i x_i}{\text{number of trials}}$$

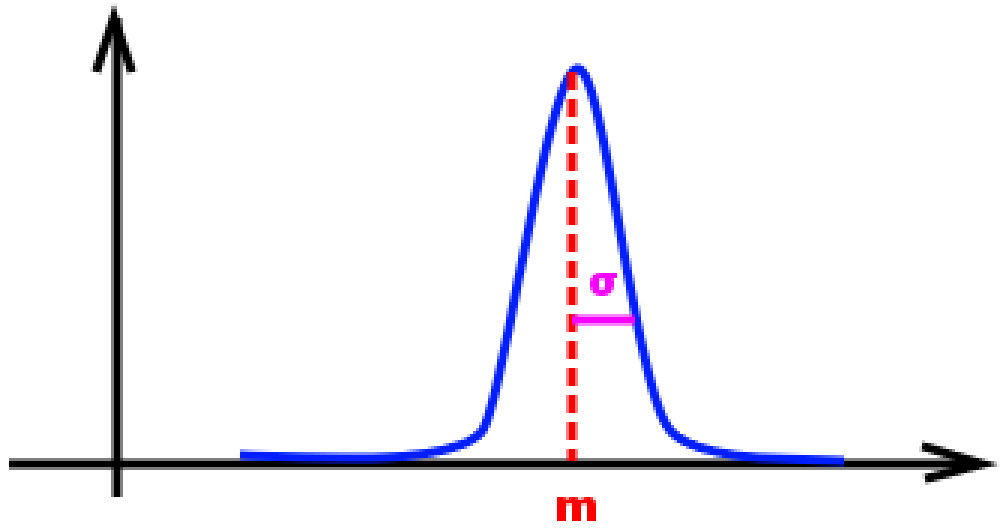
- Variance of $\underline{X}$:

$$\sigma^2 = \mathbf{D}(\underline{X}) = \int_x (x-\mu)^2\,\mathrm{d}x = \sum_x (x-\mu)^2\mathbf{P}(\underline{X} = x)$$

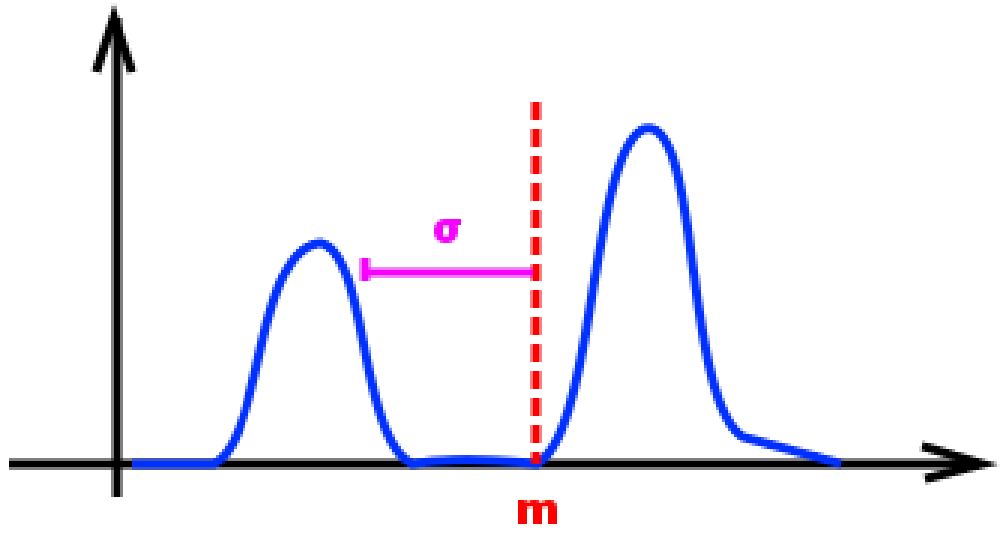
- Standard deviation of $\underline{X}$:

$$\sigma = \sqrt{\mathbf{D}(\underline{X})}$$

Describing Distributions



Describing Distributions



Common Distributions: Bernoulli

- A *random variable* $\underline{X}$ is said to have *Bernoulli distribution* if

$$\underline{X} = \begin{cases} 1 & \text{with probability } \lambda \\ 0 & \text{with probability } 1-\lambda \end{cases}$$

- To indicate this we write $\underline{X} \sim B(\lambda)$.
- Examples?

Common Distributions: Bernoulli

- A *random variable* $\underline{X}$ is said to have *Bernoulli distribution* if

$$\underline{X} = \begin{cases} 1 & \text{with probability } \lambda \\ 0 & \text{with probability } 1-\lambda \end{cases}$$

- To indicate this we write $\underline{X} \sim B(\lambda)$.
- Examples?
- $\mathbf{E}(\underline{X}) = 1 \cdot \lambda + 0 \cdot (1-\lambda) = \lambda$
- $\mathbf{D}(\underline{X}) = (1-\lambda)^2 \cdot \lambda + (0-\lambda)^2 \cdot (1-\lambda) = (1-\lambda)\lambda$

Common Distributions: Binomial

- A *random variable* $\underline{X}$ is said to have *Binomial distribution* if

$$\underline{X} = \underline{X_1} + \underline{X_2} + \cdots + \underline{X_n}$$

  where all $\underline{X_i} \sim B(\lambda)$ are *independent* Bernoulli-distributed variables.
- We denote $\underline{X} \sim B(n, \lambda)$.
- $\underline{X}$ is the number of successes in a series of n Bernoulli trials.
- $\mathbf{P}(\underline{X} = m) = \binom{n}{m} \lambda^m (1 - \lambda)^{n-m}$
- Examples?

Common Distributions: Binomial

- A *random variable* $\underline{X}$ is said to have *Binomial distribution* if

$$\underline{X} = \underline{X_1} + \underline{X_2} + \cdots + \underline{X_n}$$

  where all $\underline{X_i} \sim B(\lambda)$ are *independent* Bernoulli-distributed variables.
- We denote $\underline{X} \sim B(n, \lambda)$.
- $\underline{X}$ is the number of successes in a series of n Bernoulli trials.
- $\mathbf{P}(\underline{X} = m) = \binom{n}{m} \lambda^m (1-\lambda)^{n-m}$
- Examples?
- $\mathbf{E}(\underline{X}) = n\lambda$
- $\mathbf{D}(\underline{X}) = n(1-\lambda)\lambda$

Common Distributions: Uniform

- A *random variable* $\underline{X}$ is said to have *Uniform distribution* if

$$\underline{X} \in [0, 1], \qquad \mathbf{P}(\underline{X} < a) = a$$

- We denote $\underline{X} \sim U(0, 1)$.
- Examples?

Common Distributions: Uniform

- A *random variable* $\underline{X}$ is said to have *Uniform distribution* if

$$\underline{X} \in [0,1], \qquad \mathbf{P}(\underline{X} < a) = a$$

- We denote $\underline{X} \sim U(0,1)$.
- Examples?
- $\mathbf{E}(\underline{X}) = \frac{1}{2}$
- $\mathbf{D}(\underline{X}) = \frac{1}{12}$

Common Distributions: Normal

- A *random variable* $\underline{X}$ is said to have *Normal distribution* if

$$\underline{X} \in \mathbb{R}, \qquad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- We denote $\underline{X} \sim N(\mu, \sigma)$.
- Examples?

Common Distributions: Normal

- A *random variable* $\underline{X}$ is said to have *Normal distribution* if

$$\underline{X} \in \mathbb{R}, \qquad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- We denote $\underline{X} \sim N(\mu, \sigma)$.
- Examples?
- $\mathbf{E}(\underline{X}) = \mu$
- $\mathbf{D}(\underline{X}) = \sigma^2$

Multivariate Distributions

- So far we've been dealing with single real-valued random variables.
- In fact it's easy to generalize the idea to any other objects.
- Random vectors $(\underline{X_1}, \underline{X_2}, \ldots, \underline{X_n})$ are of particular interest.
- CDF, PDF, mean and variance are defined in a similar manner:

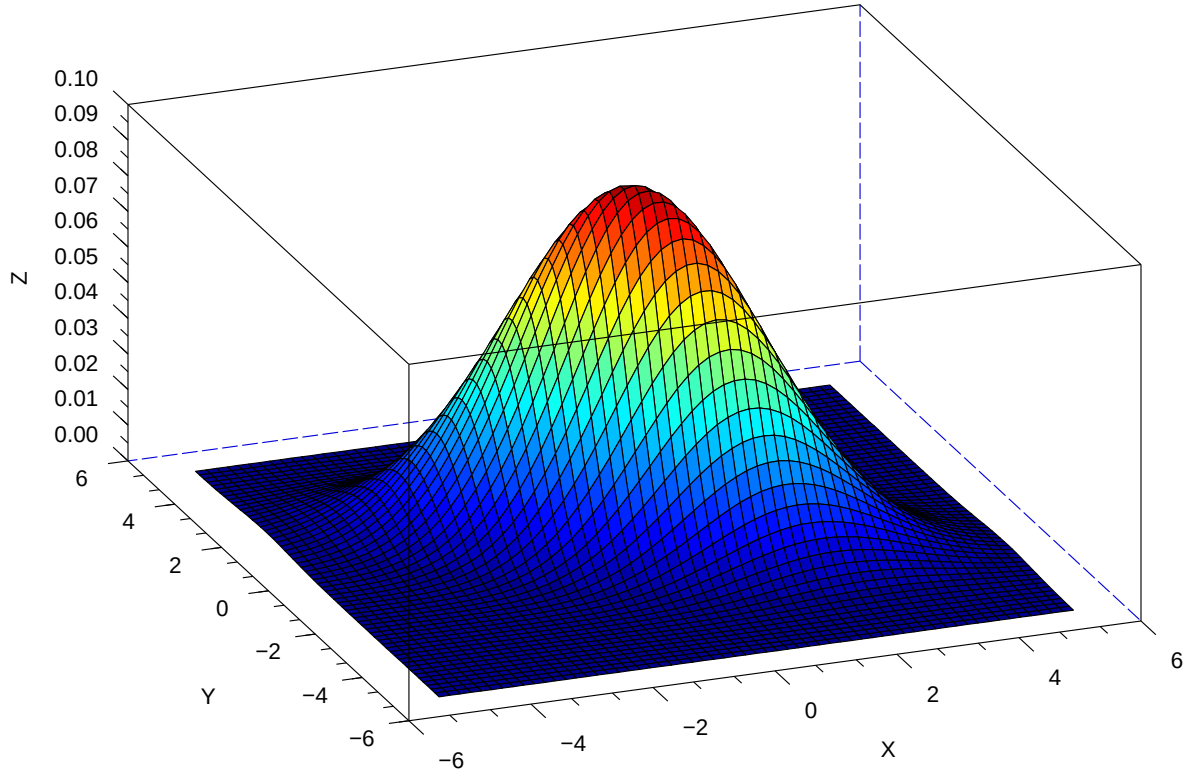
$$F(\mathbf{x}) = F(x_1, x_2, \ldots, x_n) = \mathbf{P}(\underline{X_1} < x_1, \underline{X_2} < x_2, \ldots, \underline{X_n} < x_n)$$

$$f(\mathbf{x}) = \nabla F(\mathbf{x})$$

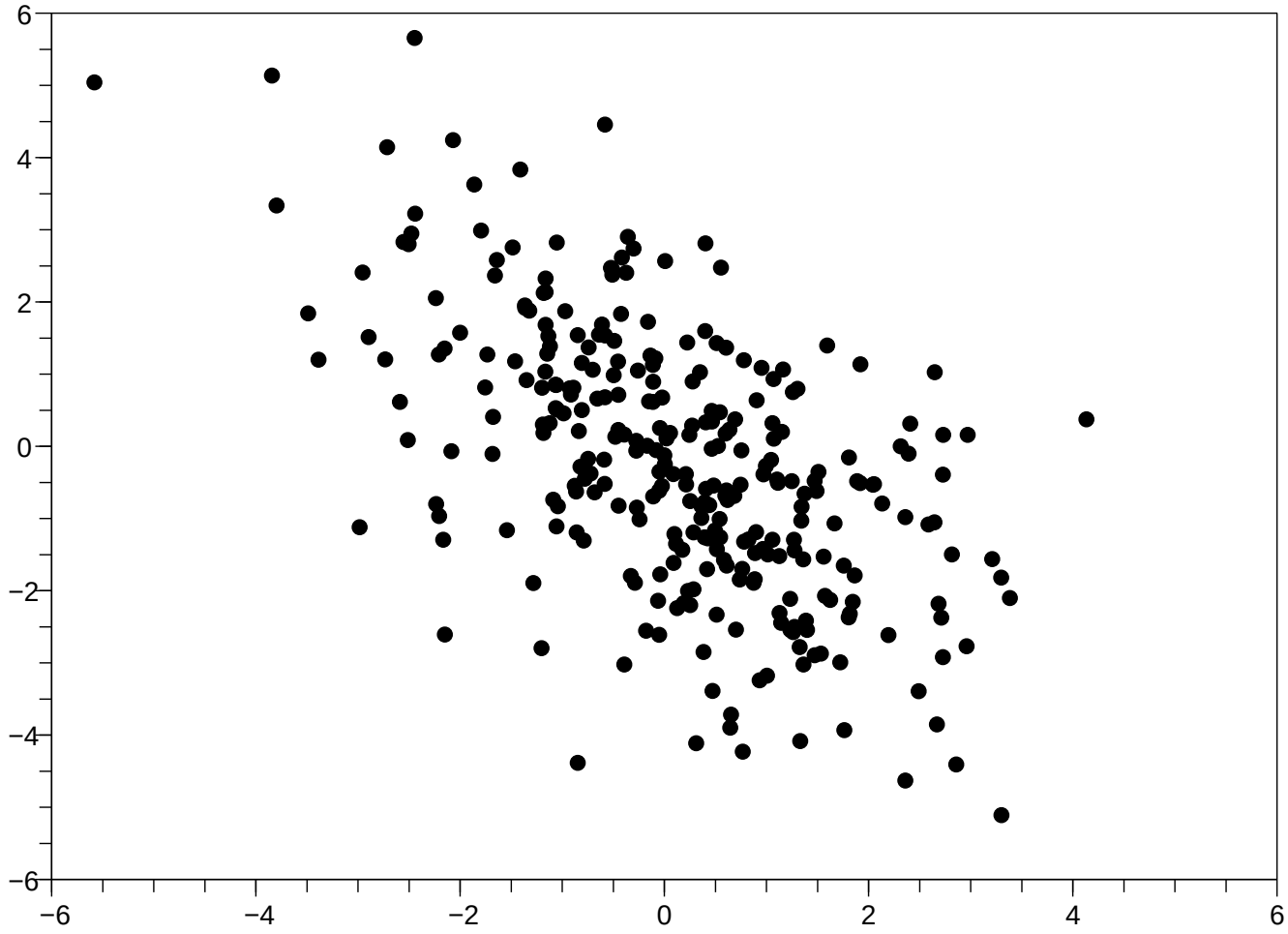
$$\mathbf{E}(\underline{X}) = \int_{\mathbf{x}} \mathbf{x} f(\mathbf{x})\, \mathrm{d}\mathbf{x} = \sum_{\mathbf{x}} \mathbf{x}\mathbf{P}(\underline{\mathbf{X}} = \mathbf{x}) = \frac{\sum_i \mathbf{x}_i}{\text{number of trials}}$$

$$\mathbf{D}(\underline{X}) = \int_{\mathbf{x}} (\mathbf{x} - \mu)^2 f(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

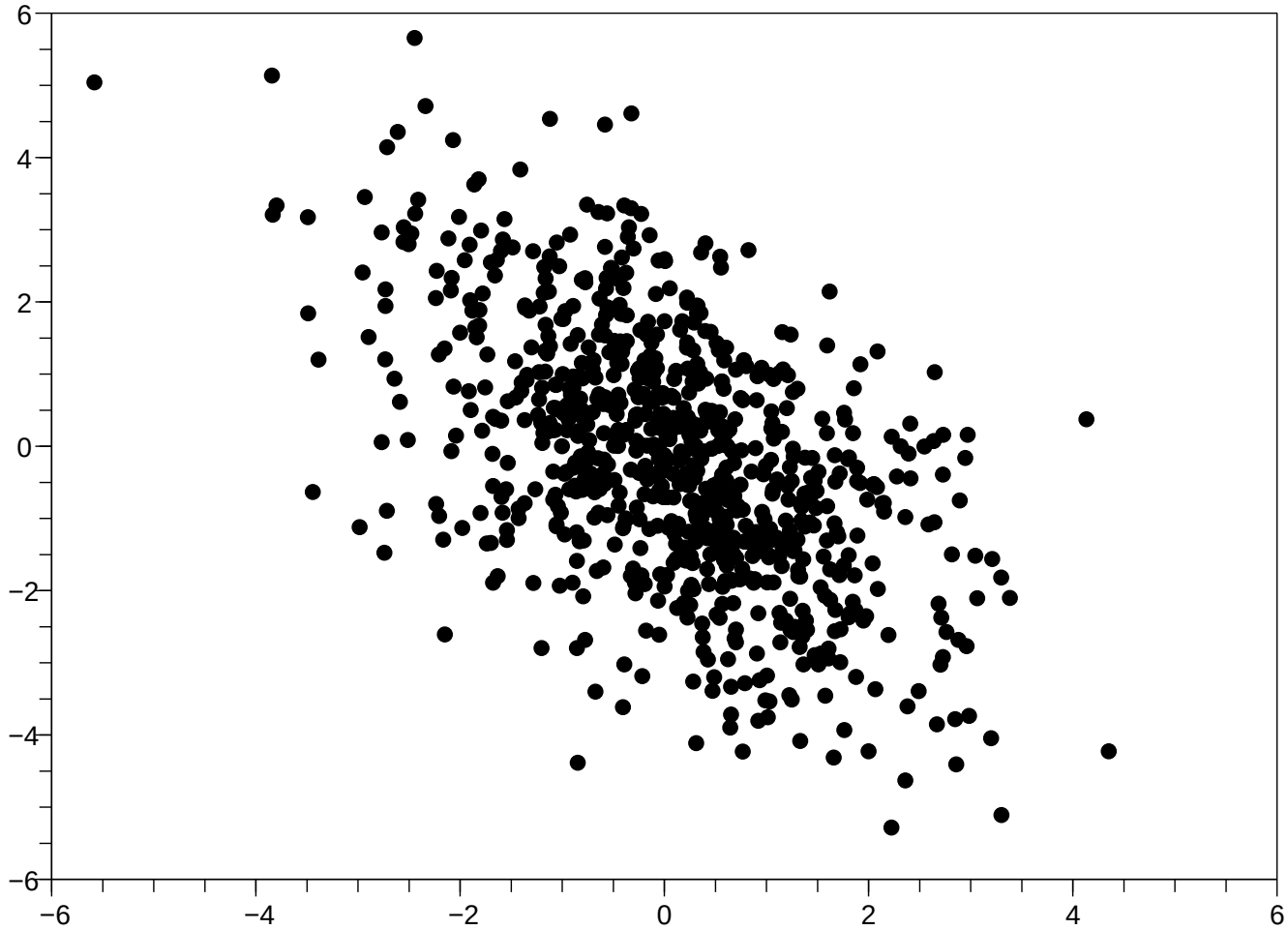
Multivariate Distributions



Multivariate Distributions



Multivariate Distributions



Conditional Distribution

- Consider a random vector $(\underline{X}, \underline{Y})$. Fix some value x for random variable $\underline{X}$. What are the possible values for $\underline{Y}$ when $\underline{X} = x$?
- They form a distribution known as the *conditional distribution of $\underline{Y}$ given $\underline{X}$*.
- Notation:
  - $\mathbf{P}(\underline{Y} = y \mid \underline{X} = x)$ — probability of $\underline{Y}$ *given* a value for $\underline{X}$.
  - $F(y \mid x)$ — corresponding CDF.
  - $f(y \mid x)$ — corresponding PDF.

Bayes Rule

- It is possible to show that:

$$\mathbf{P}(\underline{Y} = y \mid \underline{X} = x) = \frac{\mathbf{P}(\underline{X} = x, \underline{Y} = y)}{\mathbf{P}(\underline{X} = x)}$$

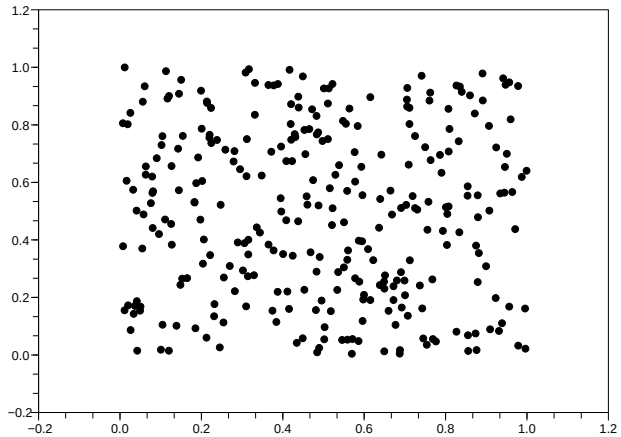
$$f(y \mid x) = \frac{f(x,y)}{f(x)}$$

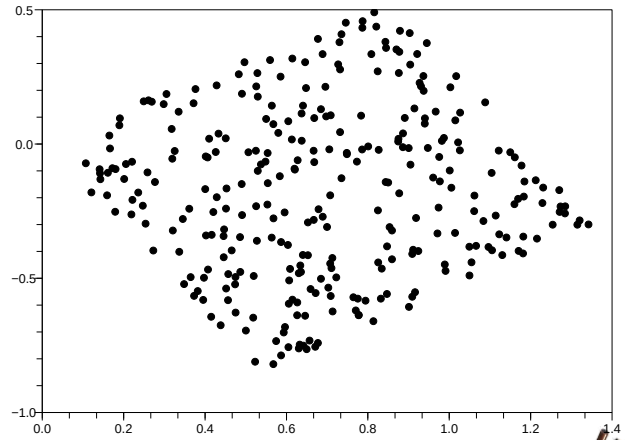
- This leads to the *Bayes rule*

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

Independence

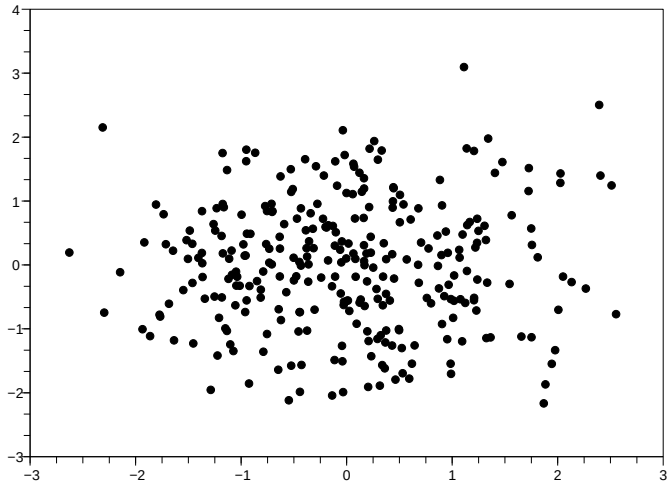
- We say that $\underline{X}$ and $\underline{Y}$ are *independent* if $f(y \mid x) = f(y)$ does not depend on x.





Independence

- We say that $\underline{X}$ and $\underline{Y}$ are *independent* if $f(y \mid x) = f(y)$ does not depend on x.



Independence

- We say that $\underline{X}$ and $\underline{Y}$ are *independent* if $f(y \mid x) = f(y)$ does not depend on x.
- $f(x, y) = f(y \mid x) f(x) = f(x) f(y)$
- We mostly need the notion of independence to describe *samples* — it's very common to assume that a certain dataset is a collection of *independent, identically distributed* random variables.
  In short: an *i.i.d. sample*.

Probabilities and Pattern Analysis

- We use a *probability distribution* to model our knowledge about things.
- In particular, we can use it to model *source of data*.
- E.g we can say "*the data comes from* $N(\mu, \sigma)$".
- What can you do with this model?

Probabilities and Pattern Analysis

- We use a *probability distribution* to model our knowledge about things.
- In particular, we can use it to model *source of data*.
- E.g we can say "*the data comes from* $N(\mu, \sigma)$".
- What can you do with this model?
  - We can use it to *predict* new data.
  - We can use just as a concise *description*.
  - In principle, the model for the data is nothing else than a *pattern* in the data.
  - We can use it to *detect* things coming from a wrong model.
- This leads us to the realm of *statistics*.

Statistics: The plan

1. Probability Theory
   - Random Variables
   - Probability Distributions
   - Conditional Distribution. Bayes Rule. Independence

2. Statistics
   - Hypothesis Testing
   - Confidence Intervals
   - Model estimation
   - Consistency

Probability vs Statistics

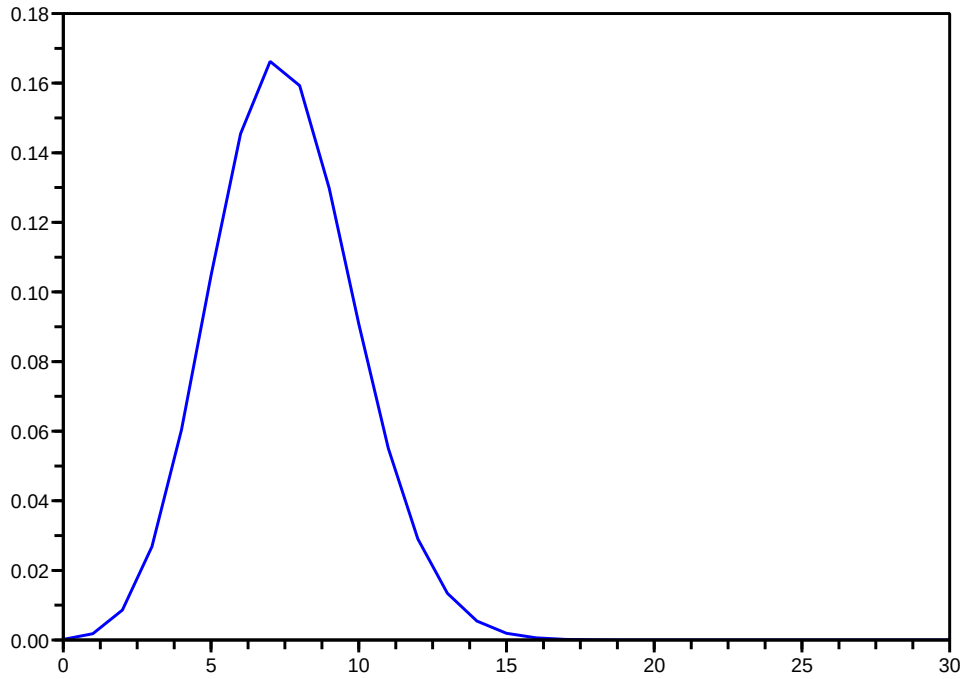
- Until now we were thinking about a “fixed” model that described the data.
- *Statistics* is about *doubting* whether the model we fixed is actually true, or even *selecting* a suitable model, by looking at the data.

Hypothesis Testing: Example

Would you please generate a random sequence of 0-s and 1-s by throwing a coin 60 times.
You don’t have to really throw a coin, you can cheat by making the sequence up “by hand”.

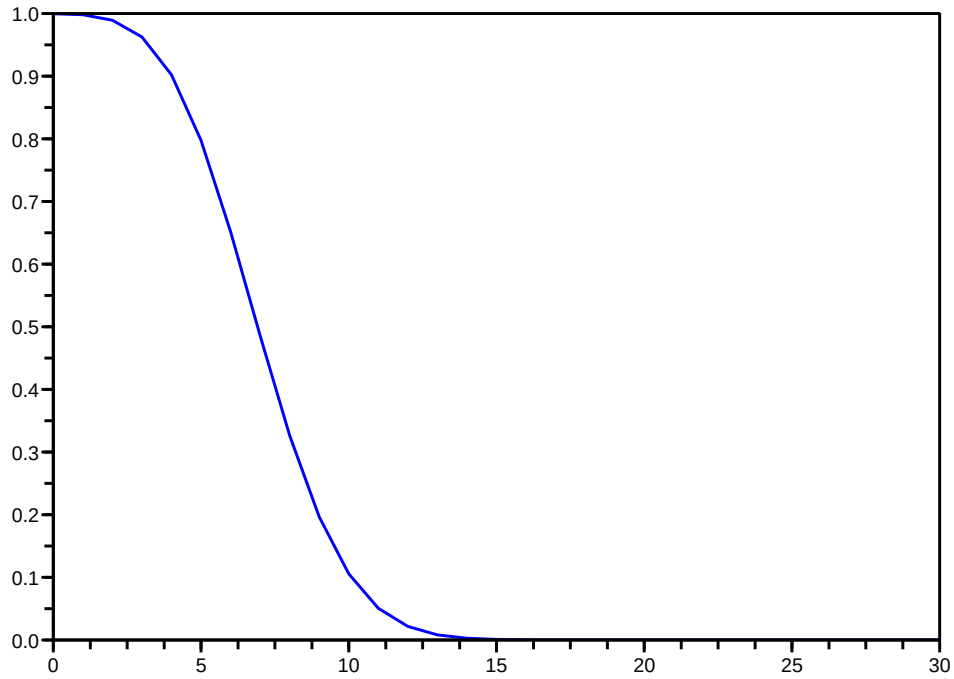
Hypothesis Testing: Example

Would you please generate a random sequence of 0-s and 1-s by throwing a coin 60 times.
You don’t have to really throw a coin, you can cheat by making the sequence up “by hand”. Right?



Hypothesis Testing: Example

Would you please generate a random sequence of 0-s and 1-s by throwing a coin 60 times.
You don’t have to really throw a coin, you can cheat by making the sequence up “by hand”. Right?



Hypothesis Testing: Summary

- Assume data comes from distribution **P**.
- Figure out some *statistic* $T(X)$.
- Select a *critical region* $[a, b]$ such that

$$\mathbf{P}(T(\underline{X}) \in [a, b]) = 0.05$$

- Now if, on our data it indeed holds that $T(X) \in [a, b]$ then conclude that **P** is not the correct distribution for the data.

Confidence Intervals: Example

“A recent poll indicated that 50% of dogs enjoy biting, but only 40% of cats.”

What’s wrong with that?

Confidence Intervals: Example

"A recent poll indicated that 50% of dogs enjoy biting, but only 40% of cats."
What's wrong with that?

- Not every dog or cat on Earth has been considered, only a *sample*.
- Suppose 50 dogs have been surveyed. What does it *really* tell us about the overall average?

Confidence Intervals: Example

"A recent poll indicated that 50% of dogs enjoy biting, but only 40% of cats."

What's wrong with that?

- Not every dog or cat on Earth has been considered, only a *sample*.
- Suppose 50 dogs have been surveyed. What does it *really* tell us about the overall average?
- $p \in (50\% - 1.96\sqrt{\frac{0.5(1-0.5)}{50}}, 50\% + 1.96\sqrt{\frac{0.5(1-0.5)}{50}})$ with 95% probability.
- $p \in (0.36, 0.64)$
- Similarly for cats.

Confidence Intervals: Summary

- Nothing is precise! Not even the parameters of your distributions.
- Ideally, you would describe knowledge about parameters by their own distributions ("metadistributions").
- In practice it's often simpler to just specify *confidence intervals*.
- A 95% *confidence interval* for a random variable $\underline{X}$ is an interval such that $\underline{X}$ belongs to it with probability 0.95.
- If $\underline{X} \sim N(0, 1)$ then $(-1.96, 1.96)$ is a 95% confidence interval for $\underline{X}$.

Estimating Models

- In the previous example we saw that we can use *data* to estimate a *confidence interval* for the parameter of the distribution.
- Sometimes we don't need a confidence interval, but an exact value (e.g. 50%).
- In short, given *data*, estimate *model*.
- Different approaches: Maximum likelihood, MAP.

Consistency

- The statistics up to now was all used to define *significant* patterns, ie patterns *present in a given dataset*.
- However, we *actually* want more than that.

Consistency

- The statistics up to now was all used to define *significant* patterns, ie patterns *present in a given dataset*.
- However, we *actually* want more than that.
- We want to be able to *generalize*.
- What is the probability that the patterns we've found in our dataset will also be present in another dataset like that?
- If we use a dataset to learn a classifier, will our classifier work for "new data"?
- Does adding data help?

Summary

- Probability theory is used to *model* phenomena.
- Statistics is used to
  - Find good models for data.
  - Detect whether data fits the model or not.
  - Analyze consistency of algorithms.

Questions?

?