Machine Learning

Lecture 2: probability & statistics refresher

Jan Chorowski
Instytut Informatyki
Wydział Matematyki i Informatyki Uniwersytet Wrocławski

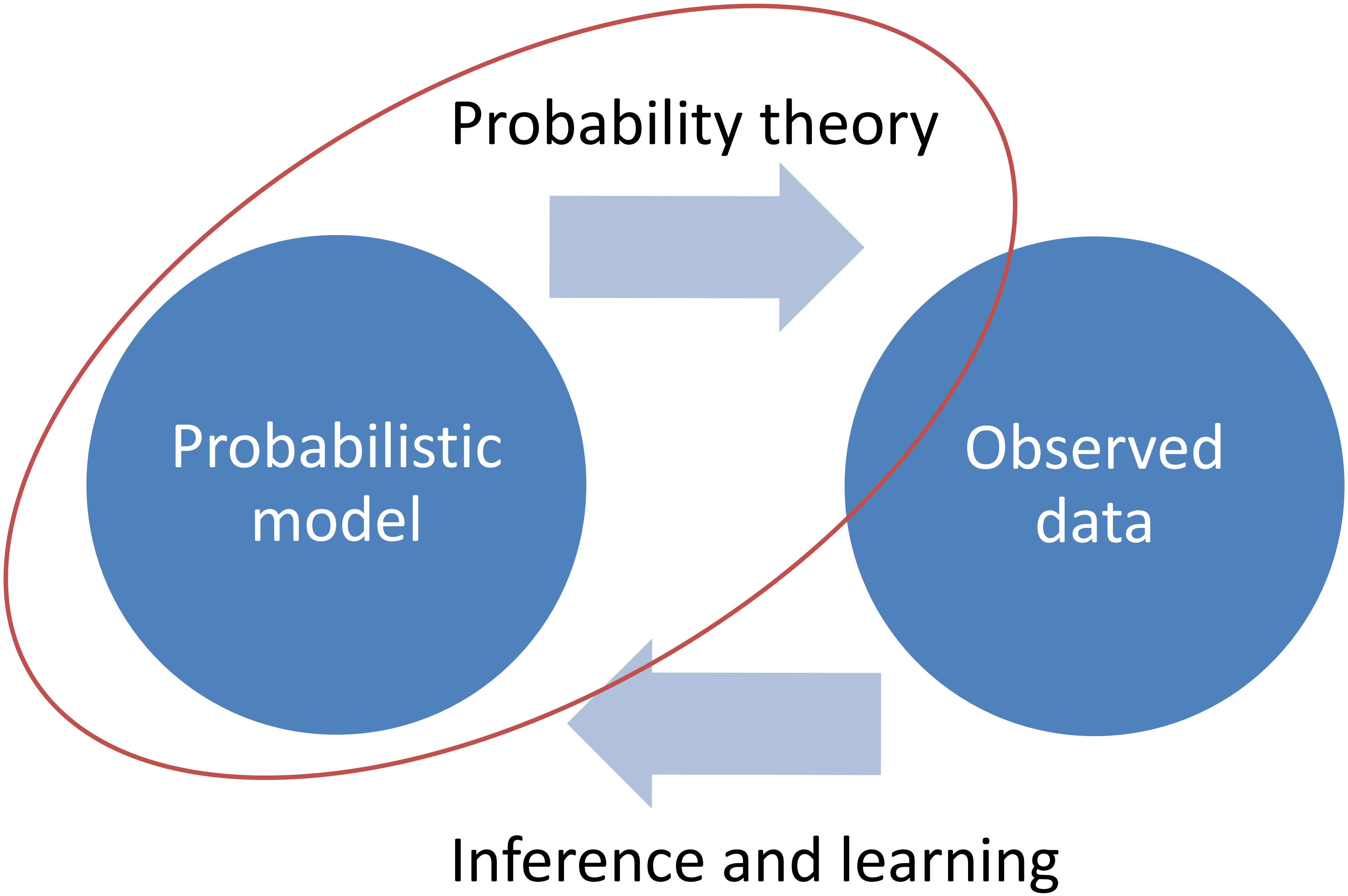
2019

Slides taken from presentations by: Y. LeCun, A. Ng

Additional materials

- http://cs229.stanford.edu/section/cs229-prob.pdf
- https://argmax.ai/docs/ml-course/01_lectureslides_ProbTheory.pdf
- Murphy, chapter 2
- Goodfellow et al. chapter 3 (the book webpage also hosts slides)
- Slides from LXMLS Summer School: http://lxmls.it.pt/2016/Lecture_0.pdf

Statistical modeling and inference



Definitions

- Ω is a **sample space**, e.g. two coin tosses $\Omega = \{HH, HT, TH, TT\}$
- $A \in 2^\Omega$ is an **event**, e.g. "first head" $\{HH, HT\}$

- $P: 2^\Omega \to \mathbb{R}$ is a **probability distributions** if:
  - $P(A) \geq 0$ for every A
  - $P(\Omega) = 1$
  - If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$

Discrete probability properties

- If $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cap B) \leq \min\big(P(A), P(B)\big)$
- (Union bound) $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega \backslash A) = 1 - P(A)$
- (Law of Total Probability)
  If $A_1 \dots A_k$ are disjoint and $\bigcup_{i=1}^{k} A_i = \Omega$, then
  $\sum_{i=1}^{k} P(A_i) = 1$.

Random Variables

A RV is a mapping $X\colon\ \Omega \to \mathbb{R}$.

- Discrete RV has countable values: $\{0,1\}$, $\mathbb{N}$
- RV X takes value x with a probability $\mathrm{P}_{\mathrm{X}}(x = X)$
- E.g. Binomial distribution
  X is the number of heads in n tosses. Tosses are independent, each with head probability Θ.

$$P_X(X = k) = P_X(k) = \binom{n}{k} \Theta^k (1 - \Theta)^{n-k}$$

Continuous RV

- Continuous RV has uncountable values: $[0,1], \mathbb{R}$
- A continuous RV X has an associated **Probability Density Function** $f_X(x)$:
  - $\forall x f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x) dx = 1$
  - $P(a < X \leq b) = \int_a^b f_X(x) dx$
  - For a continuous RV it is possible that $f_X(x) > 1$!
- Note: in the later lectures we will drop the distinction between probability $P()$ and probability density $f()$, using $P()$ in both contexts.

Cumulative distribution function (CDF)

- $F_X(x) = P_X(X \leq x)$
- $F_X(x) = \sum_{t \leq x} P_X(T)$ $\qquad$ $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$

Transformation of RVs

$$Y = g(X)$$

$$P_Y(y) = \sum_{x:y=g(x)} P_X(x)$$
$$= \sum_{x \in g^{-1}(y)} P_X(x)$$

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{\partial g^{-1}(y)}{\partial y} \right|$$
$$= f_X(x) \left| \frac{\partial x}{\partial y} \right|$$

Assumption:
g is a bijection

Intuition:

$$f_Y(y)dy \approx f_X(x)dx$$

Expected values

- The expected value of a function r of a RV X is:

$$\mathbb{E}[r(X)]_{X\sim\, P(x)} = \sum_x r(x)P\,(x)$$

$$\mathbb{E}[r(X)]_{X\sim\, f_X} = \int r(x)f_X(x)dx$$

- Example: the mean value of X is $\mu = \sum_x xP(x)$
- The expectation is linear:
  - $\mathbb{E}[X + c] = \mathbb{E}[X] + c \qquad \mathbb{E}[cX] = c\mathbb{E}[X]$
  - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ for all RV X and Y.

Variance

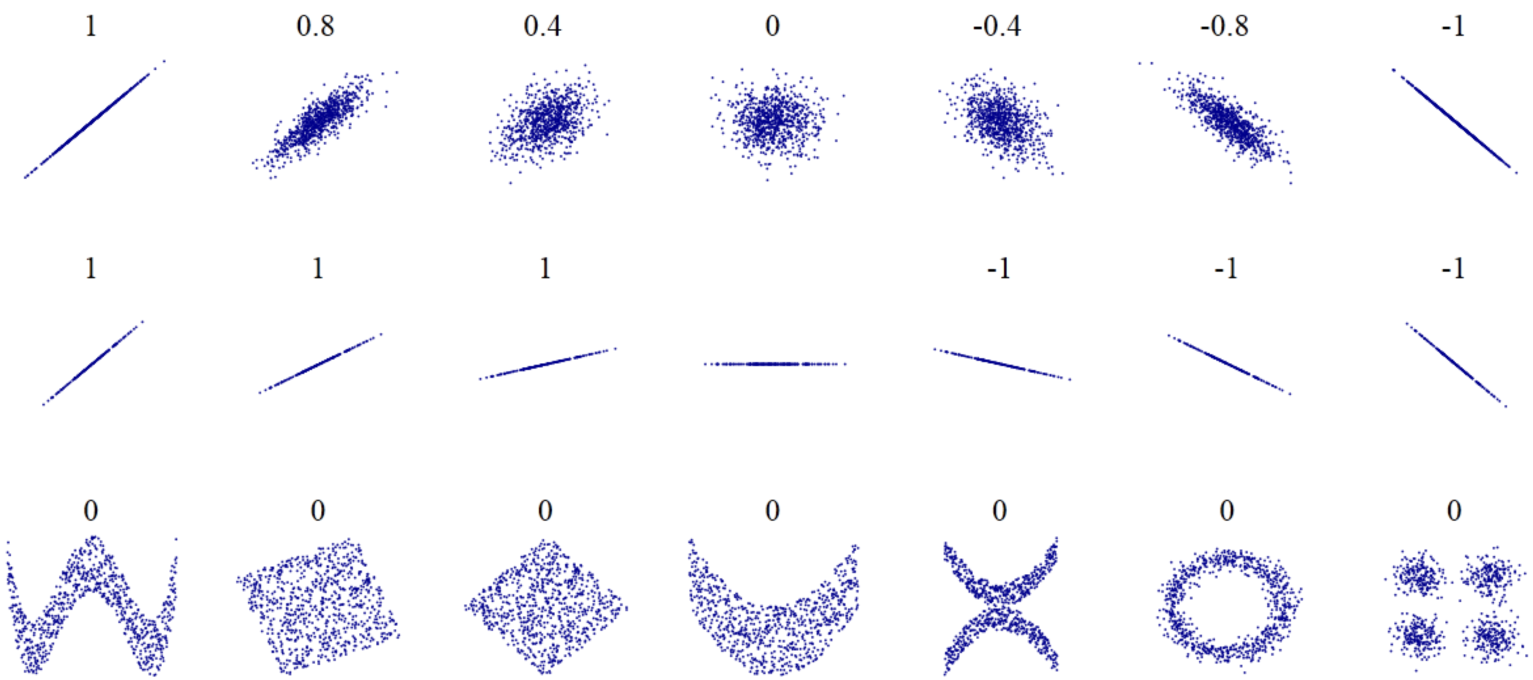
- Variance measures the spread of a RV X:
  $\sigma^2 = \mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_x (x - \mathbb{E}[X])^2$
- Standard deviation $\sigma_X = \sqrt{\mathrm{Var}[X]}$
- The Covariance between X and Y is:
  $\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- Properties of variance:
  - $\mathrm{Var}[X - c] = \mathrm{Var}[X]$
  - $\mathrm{Var[cX]} = \mathrm{c^2Var[X]}$
  - $\mathrm{Var[aX + bY]} = \mathrm{a^2Var[X]} + \mathrm{b^2Var[Y]} + 2ab\mathrm{Cov[X, Y]}$
  - When X and Y are independent:
    $\mathrm{Var[aX + bY]} = \mathrm{a^2Var[X]} + \mathrm{b^2Var[Y]}$

Correlation

- Correlation coefficient is normalized Covariance:

$$\rho_{X,Y} = \frac{\mathrm{Cov}[X,Y]}{\sigma_X \sigma_Y}$$

- $-1 \leq \rho_{X,Y} \leq 1$
- Independent $\Rightarrow$ uncorrelated



Joint probability

- Given two RVs X and Y $P(x, y)$ denotes the event that $X = x$ and $Y = y$.
- X and Y are independent iff $P(x, y) = P(x)P(y)$
- Marginal probability: $P(x) = \sum_y P(x, y)$
- Conditional probability (read probability of x given y):

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

Bayes theorem

$$P(x, y) = P(x|y)P(y) = P(y|x)P(x)$$

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)} = \frac{P(y|x)P(x)}{\sum_{x'} P(x', y)}$$

E.g. compute p(car crash | drunk driving)

Bayes theorem in action

We want: $P(\text{crash}|\text{drunk})$

Can’t get people drunk and send on the road…

$$P(\text{crash}|\text{drunk}) = \frac{P(\text{drunk}|\text{crash})P(\text{crash})}{P(\text{drunk})}$$

That’s ethical – we can estimate all need probabilities from police statistics!

Entropy

$$H(X) = \mathbb{E}_{x\sim P_X(x)}\left[\log\left(\frac{1}{P_X(x)}\right)\right]$$
$$= -\sum_x P_X(x)\log(P_X(x))$$

Interpretation:

$H(x)$: average number of nats (bits when $\log_2$) needed to transmit a message from X.

Conditional Entropy

$$H(Y|X) = \sum_{x} p(x) H(Y|X = x)$$

$$= -\sum_{x} p(x) \sum_{y} p(y|x) \log p(y|x)$$

The average entropy of Y when X is known.

Entropy for discrete RV

- $H(X) \geq 0$
- $H(Y|X) = 0$ iff Y is deterministic given X
- $H(Y|X) = H(Y)$ iff X and Y are independent
- $H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$
- $H(X, Y) \leq H(X) + H(Y)$

Entropy of continuous RV

- Entropy can be computed for continuous RV, giving the so-called **differential entropy**:

$$H(X) = -\int f_X(x) \log\left(\frac{1}{f_X(x)}\right) dx$$

- Unlike entropy, differemtial entropy can be negative!

KL Divergence

$$D_{KL}(P||Q) = -\sum_x P(x)\log\frac{Q(x)}{P(x)}$$

$$= -\sum_x P(x)\log Q(x) + \sum_x P(x)\log P(x)$$

$$= -\sum_x P(x)\log Q(x) - H_P(X)$$

expected number of nats to encode message from P using code for Q

minus

expected number of nats to encode message from P using code for P

Mutual information

$$I(X;Y) = D_{KL}(P_{X,Y}||P_X P_Y)$$

Information that X and Y share.

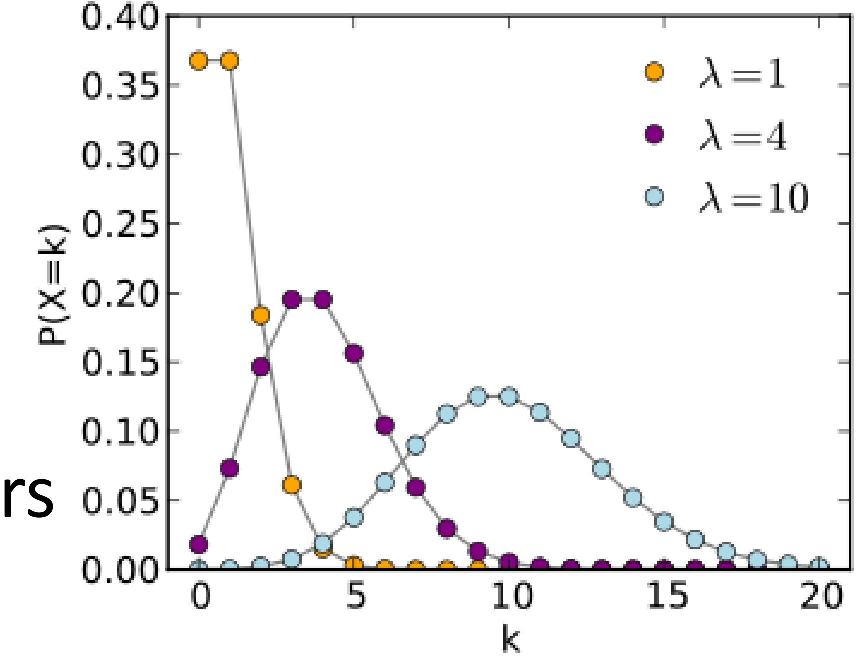
Difference of the joint from the product of marginals

$$\begin{aligned}
\mathrm{I}(X;Y) &\equiv \mathrm{H}(X) - \mathrm{H}(X|Y) \\
&\equiv \mathrm{H}(Y) - \mathrm{H}(Y|X) \\
&\equiv \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y) \\
&\equiv \mathrm{H}(X,Y) - \mathrm{H}(X|Y) - \mathrm{H}(Y|X)
\end{aligned}$$

Bernoulli and Binomial

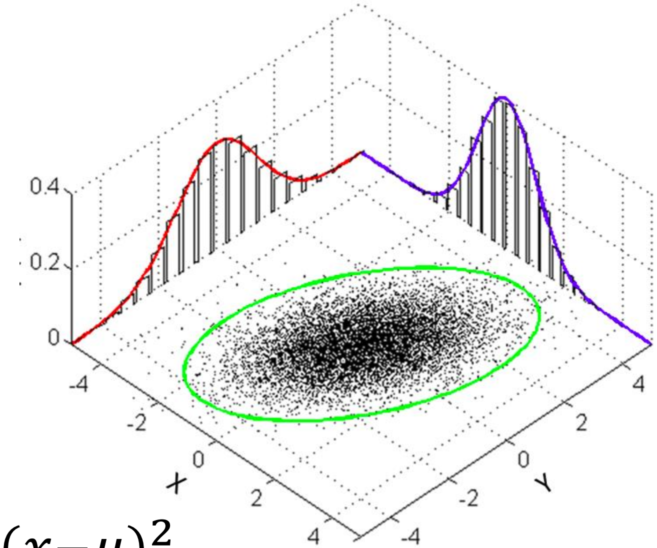
- Bernoulli:
  - X is binary
    $P(X = 1) = \phi, P(X = 0) = 1 - \phi$
  - $\mathbb{E}[X] = 0(1 - \phi) + 1\phi = \phi$
  - $\mathrm{Var}[X] = (0 - \phi)^2(1 - \phi) + (1 - \phi)^2\phi = \phi(1 - \phi)$
- Binomial:
  - RV $K =$ sum of n independent Bernoulli(ϕ) trials
  - $P(k; \phi, n) = \binom{n}{k} \phi^k (1 - \phi)^{n-k}$
  - $\mathbb{E}[K] = n\phi$
  - $\mathrm{Var}(K) = n\phi(1 - \phi)$

Poisson



- The count of rare events
- Defined for natural numbers
- $P(X = k; \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}$
- $\mathbb{E}[X] = \lambda$
- $\mathrm{Var}[X] = \lambda$
- Sum of independent Poissons is Poisson: if $X \sim \mathrm{Pois}(\lambda_X)$ and $Y \sim \mathrm{Pois}(\lambda_Y)$ then $X + Y \sim \mathrm{Pois}(\lambda_X + \lambda_Y)$

Normal distribution



- $X \sim \mathcal{N}(\mu, \sigma^2)$
- Univariate:

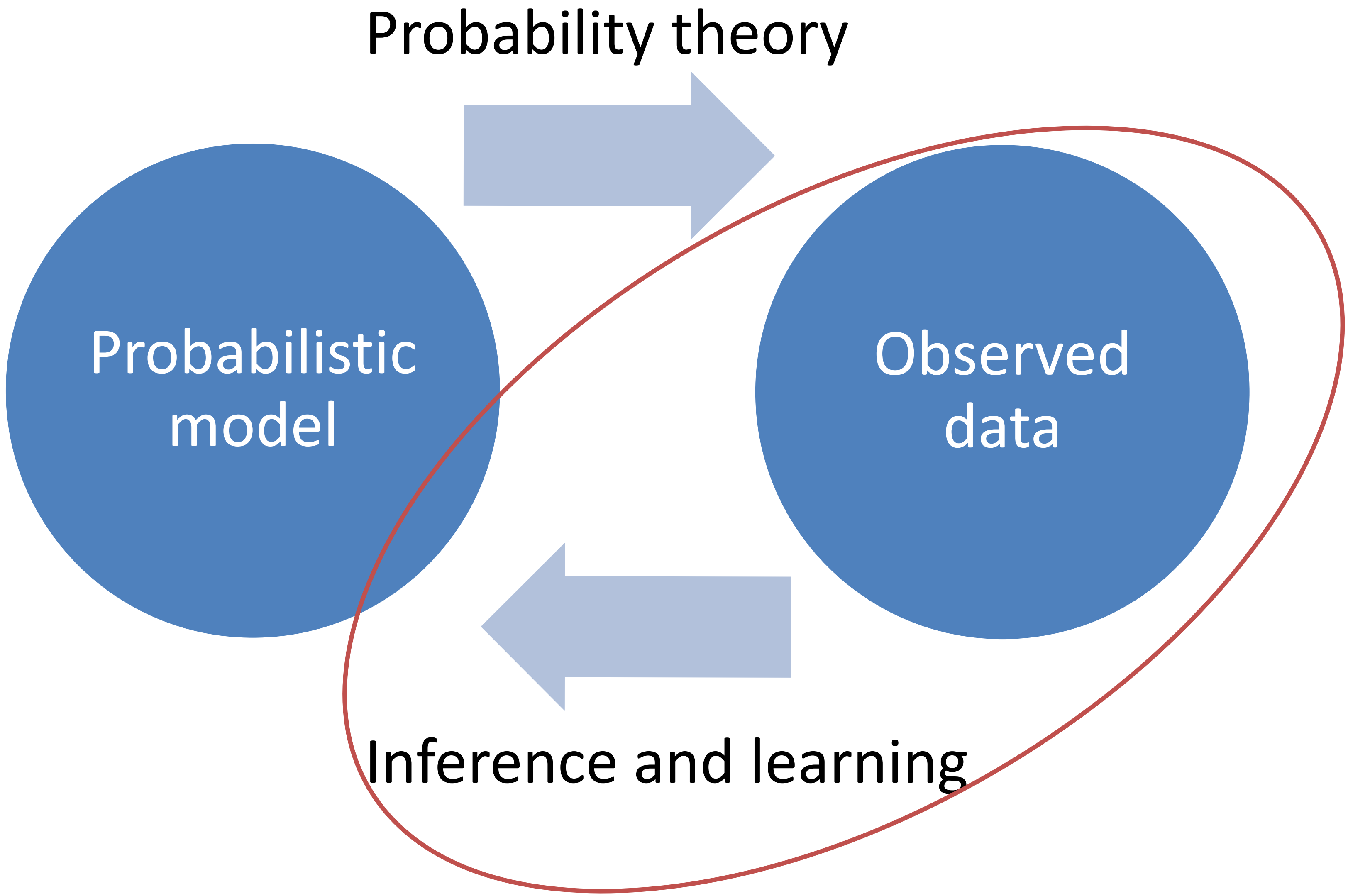
$$P(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Multivariate, k-dimensional:

$$P(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

- Mean: $\boldsymbol{\mu}$
- Variance: $\boldsymbol{\Sigma}$ (in 1D case σ)
- Conditionals, sums, and marginals of Gaussians are Gaussian

Statistical modeling and inference



Statistical Inference

Consider the polling problem:

- There exists a population of individuals (e.g. voters).
- The individuals have a voting preference (party A or B).
- We want the fraction of voters that prefer A.
- But we don't want to ask everyone (run an election)!

Polling

- Choose a **sample** of eligible voters
- Get the fraction $\bar{\phi}$ of A's supporters
- Questions:
  - How are ϕ and $\bar{\phi}$ related?
  - What is the error $(\phi - \bar{\phi})$
  - How many people to ask to have ± 3 perc. points accuracy with a high probability?

Polling model

If the population is very large, we can assume that our poll is a set of n independent Bernoulli(ϕ) trials.

The sample is IID – Independent Identically Distributed.

This corresponds to a binomial distribution:

$$\mathrm{P}(k; n, \phi) = \binom{n}{k} \phi^k (1 - \phi)^{n-k}$$

where k is the count of A’s supporters among n polled.

Likelihood

- The probability of seeing k supporters is:

$$\mathrm{P}(k; n, \phi) = \binom{n}{k} \phi^{k} (1 - \phi)^{n-k}$$

- Taken as a function $\mathcal{L}(\phi)$ we call it the likelihood.
- We will estimate the real, unknown ϕ by $\hat{\phi}$, the maximizer of the sample likelihood:

$$\begin{aligned}\hat{\phi} = \arg\max_{\phi} \mathcal{L}(\phi) &= \arg\max_{\phi} P(k; n, \phi) \\ &= \arg\max_{\phi} \log P(k; n, \phi) \\ &= \arg\max_{\phi} k \log(\phi) + (n - k) \log(1 - \phi)\end{aligned}$$

Maximum Likelihood

$$\begin{aligned}\hat{\phi} &= \arg\max_{\phi} ll(\phi) \\ &= \arg\max_{\phi} k \log \phi + (n-k) \log 1 - \phi\end{aligned}$$

At maximum the derivative wrt. ϕ is 0:

$$\frac{\partial ll(\phi)}{\partial \phi} = \frac{k}{\phi} - \frac{n-k}{1-\phi}$$

Solve for $\hat{\phi}$:

$$\frac{k}{\hat{\phi}} = \frac{n-k}{1-\hat{\phi}}$$

$$\hat{\phi} = \frac{k}{n}$$

The MLE (Maximum Likelihood Estimator) for $\hat{\phi}$ is just the sample mean $\bar{\phi} = \frac{k}{n}$!

Polling accuracy

$\frac{k}{n} = \bar{\phi}$, the fraction of A voters in the poll is an estimator for populations' fraction ϕ! How accurate is $\bar{\phi}$?

Observation: $\bar{\phi}$ is an RV!

It maps polls to results!

- $P\left(\bar{\phi} = \frac{k}{n}\right) = \mathrm{Binomial}(k; n, \phi)$
- $\mathbb{E}[\bar{\phi}] = \mathbb{E}\left[\frac{\sum_i \mathrm{trial_i}}{n}\right] = \frac{1}{n}\sum_i \mathbb{E}[\mathrm{trial_i}] = \phi$
- $\mathrm{Var}[\bar{\phi}] = \mathrm{Var}\left[\frac{1}{n}\sum_i \mathrm{trial_i}\right] = \frac{1}{n^2}\sum_i \mathrm{Var}[\mathrm{trial}_i] = \frac{\phi(1-\phi)}{n}$

Desired accuracy

Observation: the higher n, the less variable $\bar{\phi}$

We want to find n such that:

$$P(\phi - 0.03 \leq \bar{\phi} \leq \phi + 0.03) \geq 0.95$$

Then we will say that our 95% confidence interval is $\pm 3\%$ points.

That means, that if we did 100 polls, 95 would return an estimator within 3 perc. points from the true value.

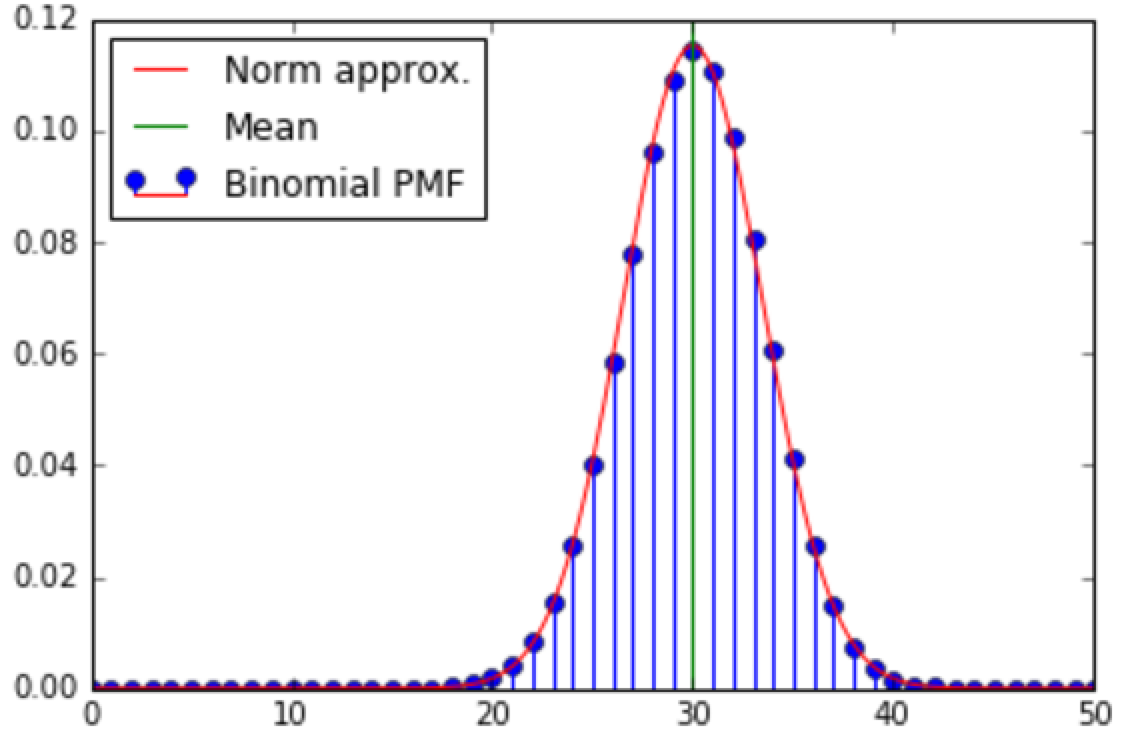
Gaussian approximation

We want to find n such that:

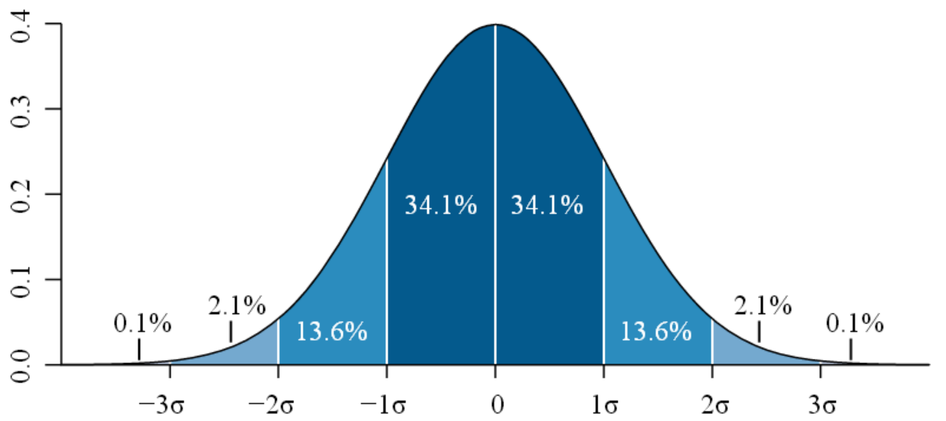
$$P(\phi - 0.03 \leq \bar{\phi} \leq \phi + 0.03) \geq 0.95$$

We know that $\mathbb{E}[\bar{\phi}] = \phi$ and $\mathrm{Var}[\bar{\phi}] = \frac{\phi(1-\phi)}{n}$.

Approximate with a Gaussian!



Gaussian confidence intervals



95% of the Gaussian's pdf lies in the range $\pm 1.96\sigma$

We want that

$$0.03 = 1.96\sigma = 1.96\sqrt{\mathrm{Var}[\bar{\phi}]} = 1.96\sqrt{\frac{\phi(1-\phi)}{n}}$$

Assume the worse case ($\phi = .5$) and solve for n!

$$n = \frac{\phi(1-\phi)}{(0.03/1.96\ )^2}$$

Bayesian Reasoning

Bayesian methods pose the problem in terms of our beliefs. This allows us to answer additional questions:

- How did my belief about the population change after seeing the poll?
- How to incorporate my prior knowledge?
- How to use small polls?

In Bayesian reasoning we will treat the population's parameter ϕ as yet another RV!

Bayesian Reasoning

- The probability assigned to ϕ is subjective – it expresses *our* uncertainty about the real ϕ.
- We have seen poll results and …
  we will use the Bayes theorem:
  $$\mathrm{P}(\phi|\mathrm{poll}) = \frac{P(\mathrm{poll}|\phi)P(\phi)}{P(\mathrm{poll})}$$
- We know the likelihood term, $\mathrm{P}(\mathrm{poll}|\phi)$.
- We need the prior $\mathrm{P}(\phi)$!
- We don't need $\mathrm{P}(\mathrm{poll})$ – it's only a scaling constant!

Prior

For convenience we will choose a prior that has a similar formula to the likelihood.

- This is called a *conjugate prior*.

Recall that: $P(k|\phi; n) \propto \phi^k (1-\phi)^{n-k}$

Choose $P(\phi) \propto \phi^{\alpha-1}(1-\phi)^{\beta-1}$

- This is the $\mathrm{Beta}(\alpha, \beta)$ distribution

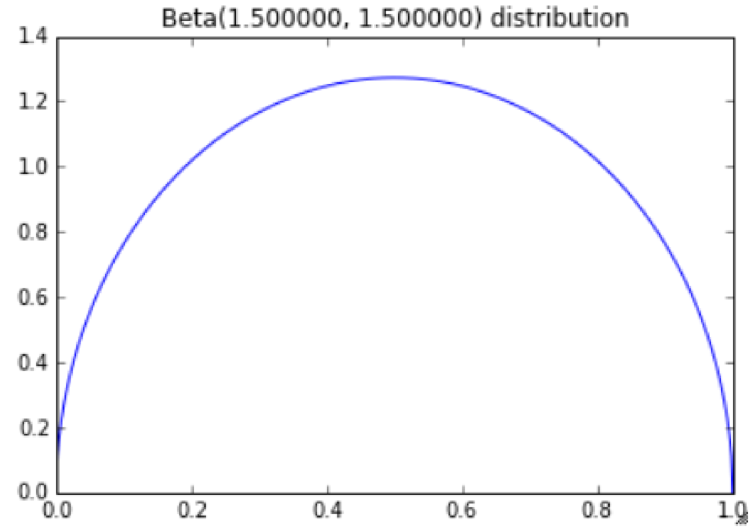
The posterior is then:

$$\begin{aligned} P(\phi|k) &\propto P(k|\phi)P(\phi) \\ &= \phi^{k+\alpha-1}(1-\phi)^{n-k+\beta-1} \end{aligned}$$

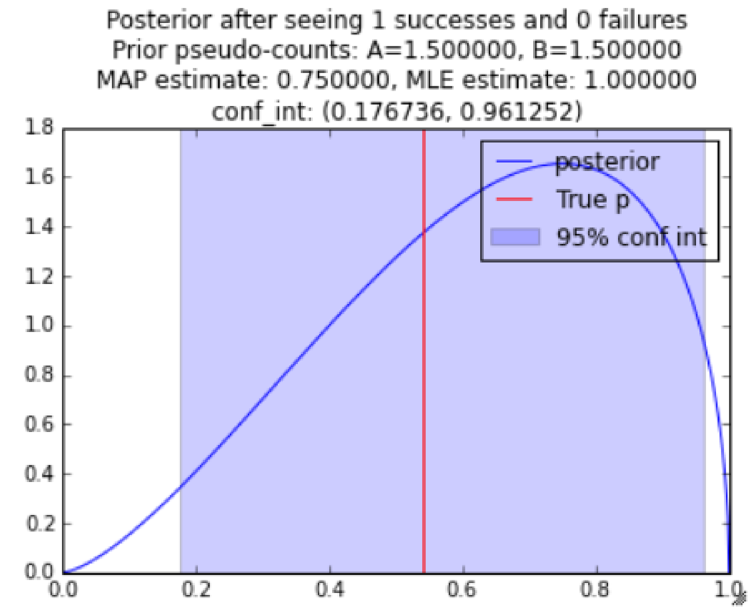
This is just $\mathrm{Beta}(k+\alpha, n-k+\beta)$.

Bayesian polling

This our prior ($\mathrm{Beta}(1.5, 1.5)$)



After seeing one success we update to $\mathrm{Beta}(2.5, 1.5)$.



In this case, the prior can be interpreted as *pseudo-counts*.