

Basic Statistics and Probability Theory

Based on

“Foundations of Statistical NLP”

C. Manning & H. Schütze, ch. 2, MIT Press, 2002

*“Probability theory is nothing but common sense
reduced to calculation.”*

Pierre-Simon Laplace (1749-1827)

PLAN

1. Event Space, and Probability Function
2. Conditional Probabiliblity
3. Bayes' Theorem
4. Independence of Probabilistic Events

5. Random Variables: the Discrete and the Continuous Case
6. Mean, Variance and Standard Deviation
7. Standard Distributions
8. Joint, Marginal and and Conditional Distributions
9. Independence of Random Variables

10. Elementary Information Theory

Elementary Notions

- **sample space:** Ω (either discrete or continuous)
- **event:** $A \subseteq \Omega$
 - the certain event: Ω
 - the impossible event: \emptyset
- **event space:** $\mathcal{F} = 2^\Omega$ (or a subspace of 2^Ω closed under complement and countable union)
- **probability function/distribution:** $P : \mathcal{F} \rightarrow [0, 1]$ such that:
 - $P(\Omega) = 1$
 - $\forall A_1, \dots, A_k$ disjoint events, $P(\cup A_i) = \sum P(A_i)$

Consequence: for a uniform distribution in a finite space:

$$P(A) = \frac{\# \text{favorable events}}{\# \text{all events}}$$

Conditional Probability

- $$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Note: $P(A | B)$ is called the **a posteriori probability** of A, given B.

- **The “multiplication” rule:**

$$P(A \cap B) = P(A | B)P(B)$$

- **The “chain” rule:**

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2 | A_1)P(A_3 | A_1, A_2) \dots P(A_n | A_1, A_2, \dots, A_{n-1})$$

- **The “total probability” formula:**

$$P(A) = P(A | B)P(B) + P(A | \neg B)P(\neg B)$$

More generally:

if $A \subseteq \cup B_i$ and $\forall i \neq j B_i \cap B_j = \emptyset$, then

$$P(A) = \sum_i P(A | B_i)P(B_i)$$

- **Bayes' Theorem:**

$$P(B | A) = \frac{P(A | B) P(B)}{P(A)}$$

Independence of Probabilistic Events

- **Independent events:** $P(A \cap B) = P(A)P(B)$
- **Conditionally independent events:**
 $P(A \cap B | C) = P(A | C)P(B | C)$

Random Variables

Definitions

Let Ω be a sample space, and
 $P : 2^\Omega \rightarrow [0, 1]$ a probability function.

- A **random variable** of distribution P is a function

$$X : \Omega \rightarrow \mathbb{R}^n$$

- For now, let us consider $n = 1$.
- The **cumulative distribution function** of X is $F : \mathbb{R} \rightarrow [0, \infty)$ defined by

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega \mid X(\omega) \leq x\})$$

Discrete Random Variables

Definition:

Let $P : 2^\Omega \rightarrow [0, 1]$ be a probability function, and X be a random variable of distribution P .

- If $Image(X)$ is either finite or unfinite countable, then X is called a **discrete random variable**.
- For such a variable we define the **probability mass function** (pmf) $p : \mathbb{R} \rightarrow [0, 1]$ as $p(x) \stackrel{def}{=} p(X = x) = P(\{\omega \in \Omega \mid X(\omega) = x\})$.
(Obviously, it follows that $\sum_{x_i \in Image(X)} p(x_i) = 1$.)

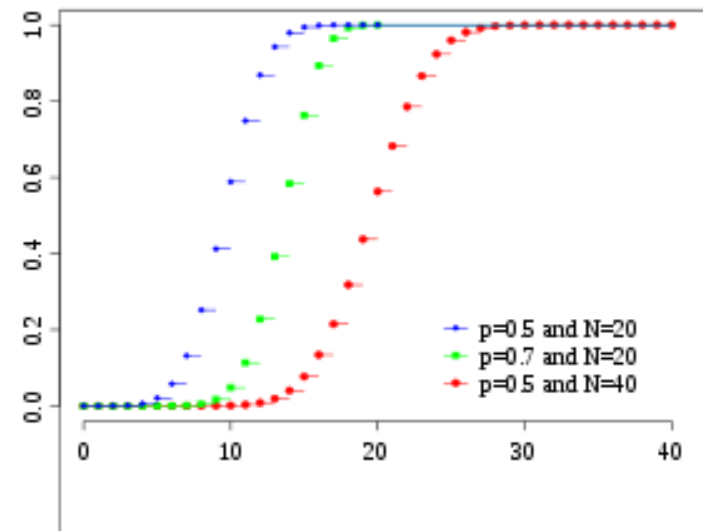
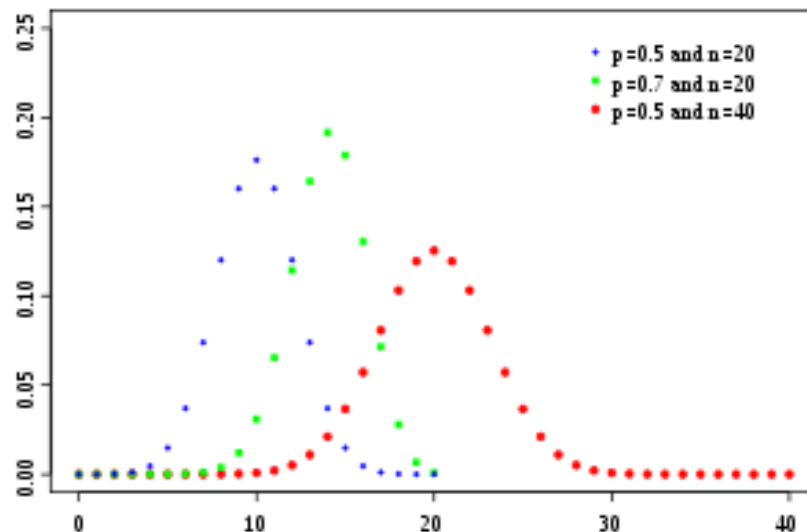
Mean, Variance, and Standard Deviation:

- **Expectation / mean** of X :
 $E(X) = \sum_x xp(x)$ if X .
- **Variance** of X : $Var(X) = E((X - E(X))^2)$.
- **Standard deviation**: $\sigma = \sqrt{Var(X)}$.

Exaeplication:

- **the Binomial distribution:** $b(r; n, p) = C_n^r p^r (1 - p)^{n-r}$ ($0 \leq r \leq n$)
 mean: np , variance: $np(1 - p)$
- **the Bernoulli distribution:** $b(r; 1, p)$

The probability mass function and the cumulative density function of the **Binomial distribution**:



Continuous Random Variables

9.

Definitions:

Let $P : 2^\Omega \rightarrow [0, 1]$ be a probability function, and $X : \Omega \rightarrow \mathbb{R}$ be a random variable of distribution P .

- If $\text{Image}(X)$ is unfinite non-countable set, and F , the cumulative distribution function of X is continuous, then X is called a **continuous random variable**.

(It follows, naturally, that $P(X = x) = 0$, for all $x \in \mathbb{R}$.)

- If there exists $p : \mathbb{R} \rightarrow [0, \infty)$ such that $F(x) = \int_{-\infty}^x p(t)dt$, then X is called **absolutely continuous**.

In such a case, p is called the **probability density function** (pdf) of X .

- For $B \subseteq \mathbb{R}$ for which $\int_B p(x)dx$ exists,
 $Pr(B) \stackrel{\text{def}}{=} P(\{\omega \in \Omega \mid X(\omega) \in B\}) = \int_B p(x)dx$.

- In particular, $\int_{-\infty}^{+\infty} p(x)dx = 1$.

- **Expectation / mean** of X : $E(X) = \int xp(x)dx$.

Exemplification:

- **Normal (Gaussean) distribution:** $N(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x - \mu)^2}{2\sigma^2}}$

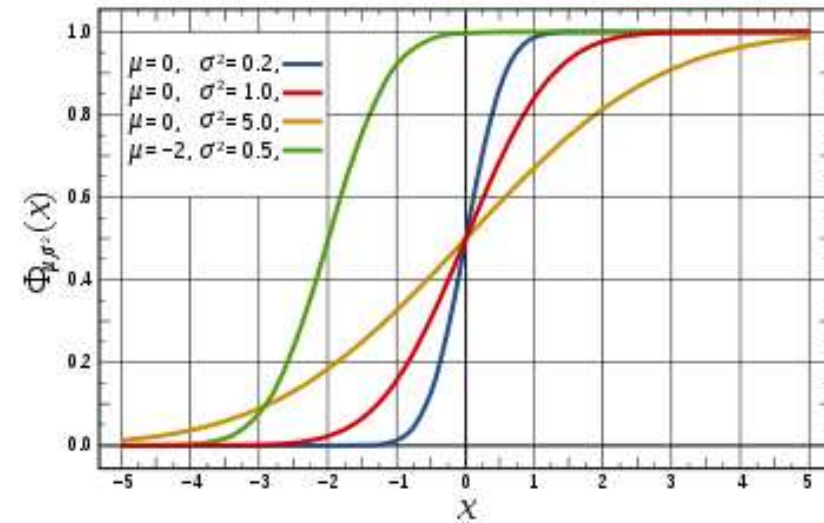
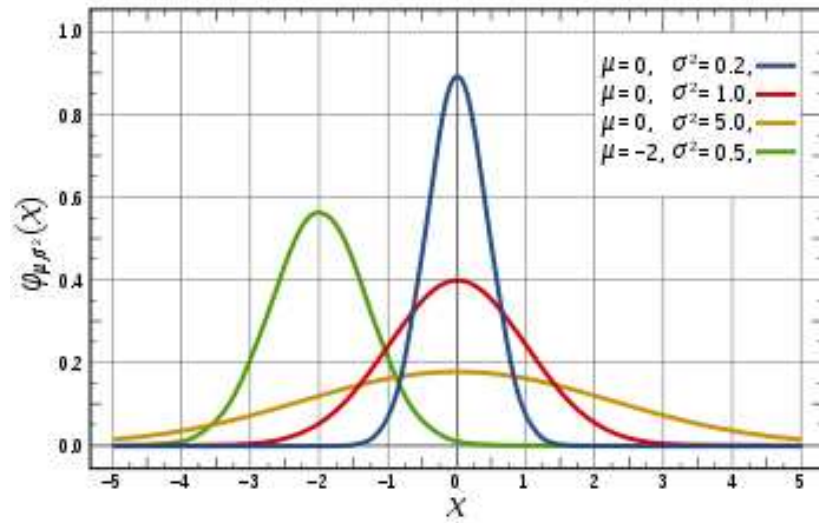
mean: μ , variance: σ^2

- **Standard Normal distribution:** $N(x; 0, 1)$

- **Remark:**

For n, p such that $np(1 - p) > 5$, the Binomial distributions can be approximated by Normal distributions.

The Normal distribution:
the probability density function and the cumulative density
function



Basic properties

Let $P : 2^\Omega \rightarrow [0, 1]$ be a probability function,

$X : \Omega \rightarrow \mathbb{R}^n$ be a random discrete/continuous variable of distribution P .

- If $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a function, then $g(X)$ is a random variable.

If $E(g(X))$ is discrete, then $E(g(X)) = \sum_x g(x)p(x)$.

If $E(g(X))$ is continuous, then $E(g(X)) = \int g(x)p(x)dx$.

- $E(aX + b) = aE(X) + b$.
- If g is non-linear $\not\Rightarrow E(g(X)) = g(E(X))$.
- $\text{Var}(X) = E(X^2) - E^2(X)$.
- $\text{Var}(aX) = a^2 \text{Var}(X)$.

Joint, marginal and conditional distributions

Exemplification for **the bi-variate case**:

Let Ω be a sample space, $P : 2^\Omega \rightarrow [0, 1]$ a probability function, and $V : \Omega \rightarrow \mathbb{R}^2$ be random variable of distribution P .

One could naturally see V as a pair of two random variables $X : \Omega \rightarrow \mathbb{R}$ and $Y : \Omega \rightarrow \mathbb{R}$. (More precisely, $V(\omega) = (x, y) = (X(\omega), Y(\omega))$.)

- the **joint pmf/pdf** of X and Y is defined by

$$p(x, y) = p_{X,Y}(x, y) = P(X = x, Y = y) = P(\omega \in \Omega \mid X(\omega) = x, Y(\omega) = y).$$

- the **marginal pmf/pdf** functions of X and Y are:

for the discrete case:

$$p_X(x) = \sum_y p(x, y), \quad p_Y(y) = \sum_x p(x, y)$$

for the continuous case:

$$p_X(x) = \int_y p(x, y) dy, \quad p_Y(y) = \int_x p(x, y) dx$$

- the **conditional pmf/pdf** of X given Y is: $p(x \mid y) = p_{X|Y}(x \mid y) = \frac{p(x, y)}{p_Y(y)}$

Independence of Random Variables

Definitions:

- Let X, Y be random variables of the same type (i.e. either discrete or continuous), and p their joint pmf/pdf.

X and Y are said to be **independent** if

$$p(x, y) = p_X(x) \cdot p_Y(y)$$

for all possible values x and y of X and Y respectively.

- Similarly, let X, Y and Z be random variables of the same type, and p their joint pmf/pdf.

X and Y are **conditionally independent** given Z if

$$p_{X,Y}(x, y) = p_{X|Z}(x | z) \cdot p_{Y|Z}(y | z)$$

for all possible values x, y and z of X, Y and Z respectively.

Basic properties

- $E(X + Y) = E(X) + E(Y)$.
- If X, Y are independent, then
 $E(XY) = E(X)E(Y)$.
- If $E(XY) = E(X)E(Y) \not\Rightarrow X, Y$ are independent.
- If X, Y are independent, then
 $Var(X + Y) = Var(X) + Var(Y)$.

A first proof: $E[X + Y] = E[X] + E[Y]$

The discrete case:

$$\begin{aligned} E[X + Y] &= \sum_{\omega \in \Omega} (X(\omega) + Y(\omega)) \cdot P(\omega) \\ &= \sum_{\omega} X(\omega) \cdot P(\omega) + \sum_{\omega} Y(\omega) \cdot P(\omega) = E[X] + E[Y] \end{aligned}$$

The continuous case:

$$\begin{aligned} E[X + Y] &= \int_x \int_y (x + y) p_{XY}(x, y) dy dx \\ &= \int_x \int_y x p_{XY}(x, y) dy dx + \int_x \int_y y p_{XY}(x, y) dy dx \\ &= \int_x x \int_y p_{XY}(x, y) dy dx + \int_y y \int_x p_{XY}(x, y) dx dy \\ &= \int_x x p_X(x) dx + \int_y y p_Y(y) dy = E[X] + E[Y] \end{aligned}$$

A second proof:

Let X and Y be random variables of the same type (i.e. either discrete or continuous).
If X and Y are independent, then

$$E[XY] = E[X] \cdot E[Y].$$

The discrete case:

$$\begin{aligned} E[XY] &= \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} xy P(X = x, Y = y) = \sum_{x \in \text{Val}(X)} \sum_{y \in \text{Val}(Y)} xy P(X = x) \cdot P(Y = y) \\ &= \sum_{x \in \text{Val}(X)} \left(x P(X = x) \sum_{y \in \text{Val}(Y)} y P(Y = y) \right) = \sum_{x \in \text{Val}(X)} x P(X = x) E[Y] = E[X] \cdot E[Y] \end{aligned}$$

The continuous case:

$$\begin{aligned} E[XY] &= \int_x \int_y xy p(X = x, Y = y) dy dx = \int_x \int_y xy p(X = x) \cdot p(Y = y) dy dx \\ &= \int_x x p(X = x) \int_y y p(Y = y) dy dx = \int_x x p(X = x) E[Y] dx \\ &= E[Y] \cdot \int_x x p(X = x) dx = E[X] \cdot E[Y] \end{aligned}$$

Elementary Information Theory

Definitions:

Let X and Y be discrete random variables.

- **Entropy:** $H(X) = \sum_x p(x) \log \frac{1}{p(x)} = - \sum_x p(x) \log p(x).$

- **Conditional entropy:**

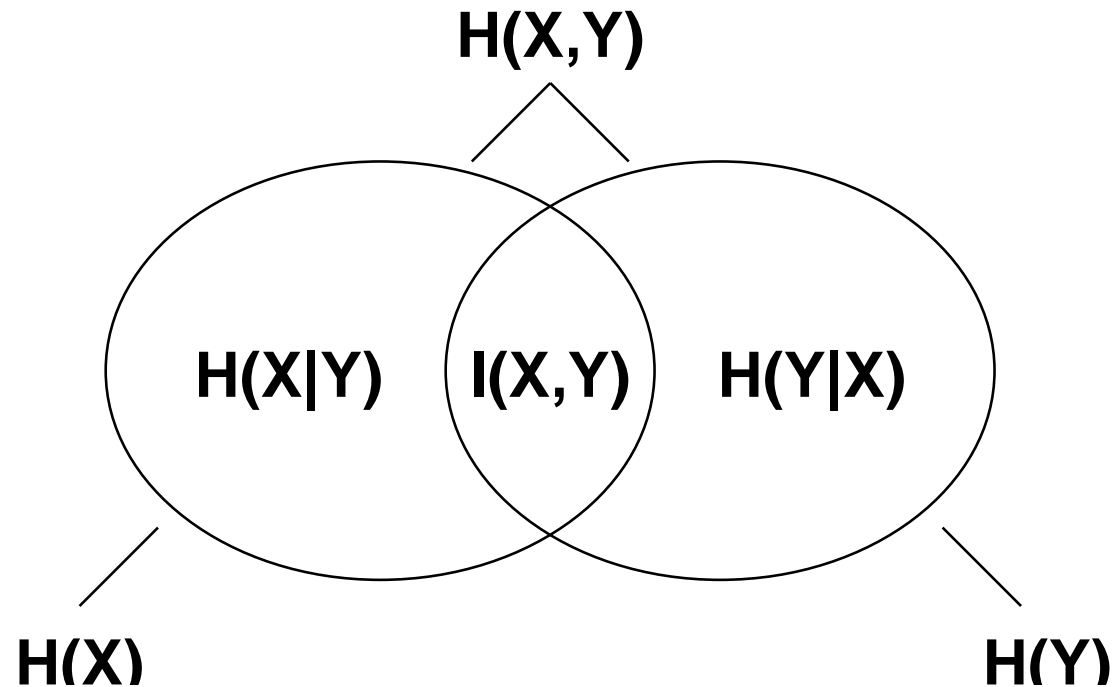
$$H(Y | X) = \sum_{x \in X} p(x) H(Y | X = x) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(y | x).$$

- **Joint entropy:** $H(X, Y) = H(X) + H(Y | X) = H(Y) + H(X | Y).$

- **Mutual information (or: Information gain):**

$$I(X, Y) = H(X) - H(X | Y) = H(Y) - H(Y | X).$$

The Relationship between
Entropy, Conditional Entropy, Joint Entropy and
Mutual Information



Other definitions:

- Let X and Y be discrete random variables, and p and q their respective pmf.

Relative entropy (or, Kullback-Leibler divergence):

$$D(p \parallel q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} = E_p(\log \frac{p(x)}{q(x)})$$

- Let X be a discrete random variable, p its pmf and q another pmf (usually a model of p).

Cross entropy: $H(X, q) = H(X) + D(p \parallel q)$

Properties:

$$I(X, Y) = D(p(x, y) \parallel p(x)p(y)).$$

$$H(X, q) = - \sum_x p(x) \log q(x) = E_p(\log \frac{1}{q(x)})$$

Recommended exercises

from [Manning & Schütze, 2002], ch. 2

Examples 1, 2, 4, 5, 7, 8, 9

Exercises 2.1, 2.3, 2.4, 2.5