Stats / Probability

"E" a random event.

S: set of all possible outcomes associated with E.

Sample space

E.g. E = coin flip → S = {H, T}.

Random Variable: X (or R.V.) is assigned a number according to outcome of a random event.

RVs: Discrete or Continuous

e.g. Yes/No, coinflip

e.g. height, weight, Time, etc.

Probability Distributions

→ called a PMF (point-mass) for Discrete RVs

→ called a Density for Continuous RVs.

① $0 \leq P(X = i) \leq 1$ for all $i \in S$

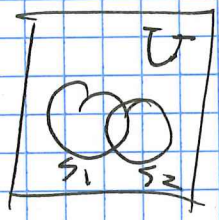
② $\sum_{i \in S} P(X = i) = 1$ $\left(\int_{-\infty}^{\infty} P(x)\,dx = 1 \right)$

(Notation: $P(x_i) = P(X = x_i)$, etc.)

Events in S are **disjoint** if: $S_1 \cap S_2 = \emptyset$ $(S_1, S_2 \in S)$

If S_1, S_2 disjoint:

Then: $P(S_1 \text{ OR } S_2) = P(S_1) + P(S_2)$



More Generally, **Additive Rule of Prob:**

$$P(S_1 \text{ OR } S_2) = P(S_1) + P(S_2) - P(S_1 \text{ AND } S_2)$$

Conditional Probability

$$P(A|B) \overset{\text{def.}}{=} \frac{P(A \& B)}{P(B)}$$

Prob. of A, given B

$$\left(P(A \& B) = P(A|B) \cdot P(B)\right)$$

Multiplication Rule:

$$P(A \& B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Independence

We say events A & B are **independent** if outcome of A has no bearing on B & vice versa.

If A & B are independent, more formally:

$$P(A \& B) = P(A)P(B)$$

(i.e. the joint probability factors)

Also, equivalently:

If A, B independent, then:

$$P(A|B) = P(A), \quad P(B|A) = P(B)$$

Thus, if A, B independent:

$$P(A \& B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

By Mult. Rule; By independence

equivalence

(2) Major Theorems in Elementary STATS:

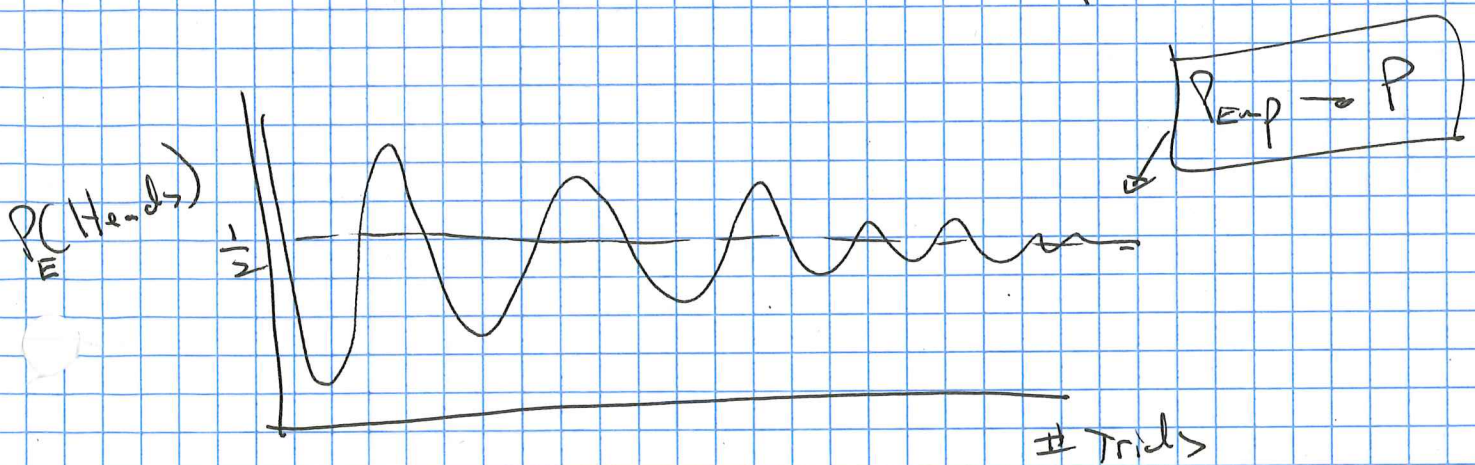
(1) Law of Large Numbers (LLN), (2) Central Limit Theorem (CLT)

LLN: (Paraphrasing) Experimental (i.e. empirical probabilities) $\overset{\text{Probs}}{\rightarrow}$ converge To Their associated Theoretical probability density as the number of Trials Tends To infinity:

$$\lim_{n \to \infty} P_{Emp}(x) = P(x)$$

e.g. Consider a single flip of a fair coin, i.e.

$P(\text{Heads}) = \frac{1}{2}$ ("True" or Theoretical probability)



CDF (Cumulative Density Function)

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} p(n)\,dn$$

$F_X(x) = .25$ ($x = Q_1$ Quantile)

$F_X(x) = .5$ ($x = Q_2$: Median)

$F_X(x) = .75$ ($x = Q_3$)

Note: $\frac{d}{dx} F(x) = p(x)$ — why?

(cdf) (pdf)

Some Essential Distributions

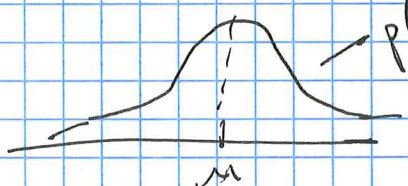
Normal / Gaussian (1-D) (continuous)

we write: $X \sim N(\mu, \sigma)$

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

Normalization constant

Standard Normal: $N(0,1)$



Empirical Rule

$x \sim N(\mu, \sigma)$

$x \pm \mu$: 68%

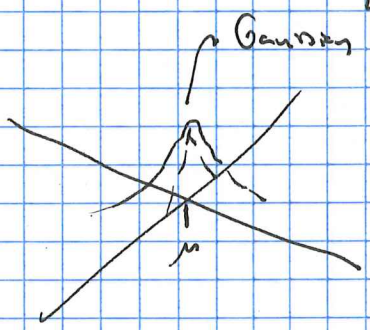
$x \pm 2\mu$: 95%

$x \pm 3\mu$: 99.7%

MVN (Multi-variate Normal)

$$P(\vec{x}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left[-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})\right]$$

$(x \in \mathbb{R}^d)$



Note: $\Lambda = \Sigma^{-1}$

referred to as Precision matrix

$$\underbrace{\Phi}_{\text{(Cumulative Normal)}}(x) = \int_{-\infty}^{x} P(z)\,dz = \frac{1}{2}\left(1 + \frac{\operatorname{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)}{\text{Gaussian}}\right)$$

where: $\operatorname{erf}(\cdot)$ is error function, has no closed-form expression.

Bernoulli (Discrete)

$$p(X=1) = \theta$$
$$p(X=0) = 1-\theta$$

$0 \le \theta \le 1$

e.g. flip a coin w/ Prob. Heads $= \theta$
Tails $= 1-\theta$

Binomial Distribution

(Discrete)

(·) Case of n Bernoulli trials

Binary Outcomes

$S = \{0, 1\}$

let: $p(X=1) = \theta$ "success"

$p(X=0) = 1-\theta$ "failure"

$$P(X=k) = \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

prob. "k successes in n trials"

ex. flip biased coin 10 Times $p(H) = .6$

$p(T) = .4$

$$P(\text{exactly 7 H in 10 flips}) = \binom{10}{7} (.6)^7 (.4)^3$$

Poisson Distribution

(Discrete)

λ: Mean # successes in given Time interval

$$P(X=k) = \frac{e^{-\lambda} \lambda^k}{k!}$$

$k = 0, 1, 2, \ldots$

Q: If, on average, 4 people visit a given webpage per minute, what is prob of Two or fewer visitors?

$P(\text{Two or fewer visitors in 4 min}) =$

$P(X=0) + P(X=1) + P(X=2)$

$$= \frac{e^{-4} \cdot 4^0}{0!} + \frac{e^{-4} \cdot 4^1}{1!} + \frac{e^{-4} \cdot 4^2}{2!} \approx \boxed{.238}$$

CLT (Central Limit Theorem)
(Classical, non-Technical version)

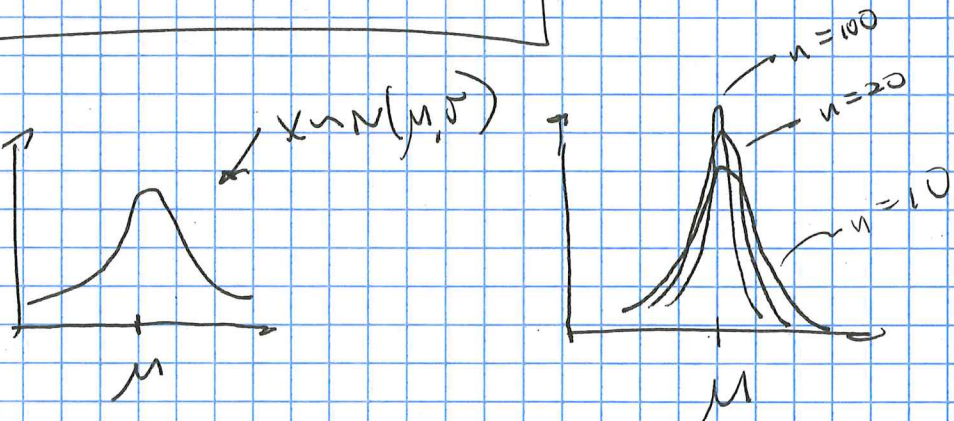
Given: $X_1, X_2, \ldots, X_n$ — **IID** → Independent, Identically Distributed

Random sample

where: $X_i \sim N(\mu, \sigma)$.

Then: $\boxed{\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$

In a picture:



Q: Given IQ scores have dist: $N(100, 15)$

what is prob: $P(85 \le X \le 115) = 68\%$

For 10 individual, what is: $P(85 \le \bar{X} \le 115) > 99\%$

Expectation (of a RV)

$$E[X] = \sum_i x_i P(X = x_i) \qquad \text{(Discrete case)}$$

$$E[X] = \int_{-\infty}^{\infty} x P(x)\, dx \qquad \text{(Continuous)}$$

Q: How many Heads are expected in 10 flips of a fair coin?

X	$P(X)$
0 (T)	$\frac{1}{2}$
1 (H)	$\frac{1}{2}$

$\Rightarrow$ PMF (Bernoulli Trials) $(n = 10)$

$$E[X] = \sum_{k=0}^{10} \binom{10}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} \cdot k$$

$$= \sum_{k=0}^{10} \binom{10}{k} \left(\frac{1}{2}\right)^{10} \cdot k = \left(\frac{1}{2}\right)^{10} \sum_{k=0}^{10} \binom{10}{k} \cdot k$$

$$= \boxed{5}$$

Variance & Standard Deviation (for RVs)

$$\mathrm{Var}[X] = E\left[(X-\mu)^2\right] = \sum_i \left(x_i - E[X]\right)^2 \left(P(X=x_i)\right)$$

where $\mu = E[X]$

$$SD[X] = \sqrt{\mathrm{Var}[X]}$$

Corollary: $\mathrm{Var}[X] = E[X^2] - \mu^2$

Proof:

$$\mathrm{Var}[X] = E\left[(X-\mu)^2\right] = E\left[X^2 - 2\mu X + \mu^2\right]$$

$$= E[X^2] - E[2\mu X] + E[\mu^2]$$

By "linearity" of Expectation; μ^2: a constant

$$= E[X^2] - 2\mu \cdot E[X] + \mu^2$$

By linearity; $E[\text{constant}] = \text{constant}$

$$= E[X^2] - 2\mu\cdot\mu + \mu^2 = E[X^2] - \mu^2. \quad \square$$

Covariance

X, Y RV's:

$$\boxed{Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]}$$

Lemma: If X, Y are independent, Then $Cov(X,Y) = 0$.

Pf: $E\left[(X-\mu_X)(Y-\mu_Y)\right] =$

$$E\left[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y\right] \overset{\text{By Linearity of } E[\cdot]}{=}$$

$$= E[XY] - \mu_Y E[X] - \mu_X E[Y] + E[\mu_X\mu_Y]$$

$$= E[XY] - \mu_Y\mu_X - \mu_X\mu_Y + \mu_X\mu_Y$$

$$\left(E[XY] = \sum_{x}\sum_{y} p(x,y)\cdot xy \overset{\text{since } X,Y \text{ independent}}{=} \sum_{x}\sum_{y} p(x)\cdot x\cdot p(y)\cdot y = E[X]\cdot E[Y]\right)$$

$$= E[X]\cdot E[Y] - \mu_X\mu_Y$$

$$= \mu_X\cdot\mu_Y - \mu_X\cdot\mu_Y = 0. \quad \square$$

Covariance Matrix

Let $\vec{X} = \langle X_1, \ldots, X_n \rangle$ (a vector of RV's)

$$\Sigma_{ij} = \mathrm{Cov}(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

Σ: Matrix of Covariances

Note:
$$\Sigma_{ij} = \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \vdots & \ddots & & \vdots \\ \mathrm{Cov}(X_n, X_1) & \cdots & & \mathrm{Var}[X_n] \end{bmatrix}$$

Σ is symmetric, positive, semi-definite.

Bayes' Theorem

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

"Re'Py Thayrem Theorem" of AI/ML.

Pf:
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)} \quad \square$$

Medical Diagnosis Example

Q: Physician knows $P(\text{Disease}) = 1\%$

$P(\text{No Disease}) = 99\%$

$P(\oplus \text{ Test} \mid D) = .792$ (True Positive)

$P(\oplus \text{ Test} \mid \text{No } D) = .096$ (False Positive)

Given That Patient $\underline{\text{receives } \oplus \text{ test result}}$,

what is probability they have disease?

Want: $P(D \mid \oplus) = \dfrac{P(\oplus \mid D)\,P(D)}{P(\oplus)}$

By Bayes'

$$= \frac{P(\oplus \mid D)P(D)}{P(\oplus \mid D)P(D) + P(\oplus \mid \text{No } D)P(\text{No } D)} = \frac{(.792)(.01)}{(.792)(.01) + (.096)(.99)}$$

$$\approx \boxed{.103}$$

Frequentists vs. Bayesians Statistics

Frequentists: Model parameters (θ) are fixed (i.e. Platonic); Data are drawn from "God's distribution", defined by θ.

Bayesians: Data are fixed (the observed data); data are observed from realized sample; we encode prior beliefs; parameters are described probabilistically.

Frequentists: Use MLE (Maximum likelihood estimate) for point estimate of θ:

("likelihood" of data)

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(D|\theta)$$

Bayesians: Compute MAP (Maximum a Posterior) estimate:

("prior" on θ)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta)$$

where posterior: $p(\theta|D) \propto p(D|\theta)p(\theta)$

Various "Pathologies" of Frequentist Prob. Exist:

Re Problem of Induction (Hume), Black Swan Paradox, limited exact solutions, reliance on "long-term" frequencies.

Bayesian Statistics have been called "statistics of Re 21st cent." (Efron)

Q: Do I need any Calculus for AI/ML?

A: Honestly, not That much!

At minimum, Know (2) Basic Things:

(1) Calculus gives us a framework to coherently/mathematically describe Re flow/dynamism of Re real-world.
(engineers, physicists, etc. like this)

(2) Principles in Calculus are helpful in [optimization]!

If $f(x)$ is [convex] / [concave]

it has a unique, global minimum/maximum.

We can use differential calculus to solve convex/concave problems (or even approximate sols. for non-convex sols)

Hill-Climbing

Suppose $f(x)$ is concave – we devise an iterative "hill-climbing" algorithm to approximate: $x^* = \arg\max_x f(x)$

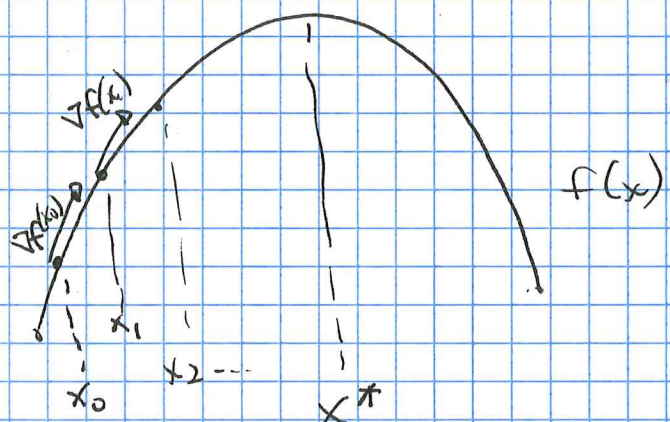
(1) Initial guess: x_0

(2) Compute gradient @ x_0:

$\nabla f(x_0)$

Recall: $\nabla f = \langle f_{x_1}, f_{x_2}, \ldots, f_{x_n} \rangle$

(vector of partial derivatives)



(3) Set $x_1 = x_0 + \gamma \nabla f(x_0)$

γ: learning parameter (controls convergence speed)

loop: (2)–(3) for $x_2, x_3, \ldots$

until stopping condition.

(Very Brief) Information Theory

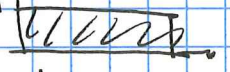
Entropy of a RV

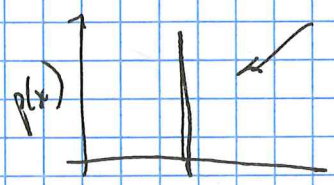
$$H(X) = -\sum_i p(X=i) \log_2 p(X=i)$$

(Define: $0 \log 0 \triangleq 0$)

quantifies disorder/uncertainty

e.g. By some accounts, entropy for English language (per letter) is ≈ 2.62 bits. (using N-gram model), i.e. on average we need ≈ 2.62 binary questions to guess N-th letter of a string. (Also a measure of redundancy)

(*) The distribution of Max entropy is the uniform distribution. → 

(*) The distribution of Minimum entropy is the (Dirac) delta function. → 

← deterministic, i.e. zero uncertainty

Ex. Consider Bernoulli RV: X

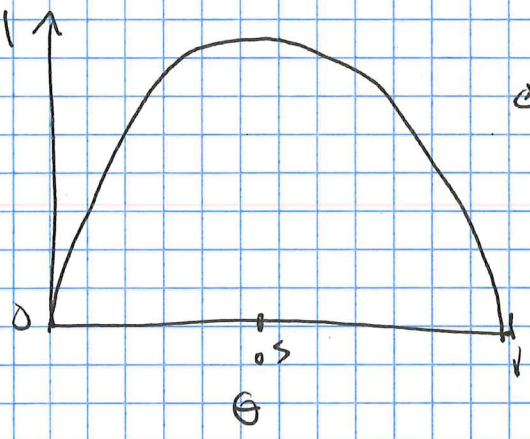
$$P(X=1) = \theta, \quad P(X=0) = 1-\theta$$

$$H(X) = -\sum_i p(x) \log_2 p(x)$$

$$= -\left[p(X=1)\log_2 p(X=1) + p(X=0)\log_2 p(X=0)\right]$$

$$= -\left[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)\right]$$

graph →



Max occurs when $\theta = .5$ (i.e. uniform);

Min occurs when $\theta = 0$ or 1 (i.e. deterministic)

KL Divergence (Kullback-Leibler)

(p, q are prob. distributions)

$$KL(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$$

Measure of dissimilarity b/w distributions.

$$= \sum_i p_i \log p_i - \sum_i p_i \log q_i$$

(entropy; "cross entropy")

$$= -H(p) + H(p,q)$$

Note: $KL(p\|q) \geq 0$ & $KL(p\|q) = 0$ iff $p = q$.

"Information Inequality"

Recall: Covariance (& Correlation) measure the linear dependence b/w RVs.

Using **KL-Divergence** we can develop a **more general** notion of **dependence**: **Mutual Information (MI)**

$$I(X;Y) = KL(p(x,y) \| p(x)p(y))$$
$$= \sum_x \sum_y p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

From above: $I(X;Y) \geq 0$ & $I(X;Y) = 0$ iff $p(x,y) = p(x)p(y)$

Thus: $I(X,Y)$ measures "similarity" between $p(x,y)$ (joint) & $p(x) \cdot p(y)$ (factored joint).

Shannon Source Coding Theorem: (for binary alphabet)

$$H(X) \leq E[L] \leq H(X) + 1$$

$E[L]$: expected word length