Baseado no Cap. 15 do Mary L. Boas e Cap3 do Harvey Gould

Para introduzir os conceitos de probabilidade e algumas técnicas iremos usar o problema do caminho aleatório em 1-D. Apesar de ser um dos modelos mais simples, tem aplicações no mov. de partículas em um viscoso

Antes, alguns conceitos básicos de prob.

→ Probabilidades surgem quando não conseguimos prever com absoluta certeza o resultado de uma observação em um dado sistema.

Preparamos o sistema sempre da mesma maneira, mas os resultados das observações são diferentes

→ Fica claro que para se falar de prob. precisamos observar (ou fazer medidas) mais que uma vez sobre o sistema preparado exatamente da mesma maneira.

De modo equivalente podemos ter um conjunto de sistemas preparados da mesma maneira e observar cada membro do conjunto (os sistemas devem ser idênticos e independentes)

A esse conjunto de sistemas idênticos damos o nome de ensemble (grupo em francês)

→ Assim não faz sentido falar de prob. para um único sistema. Temos sempre que pensar no sistema como parte de um ensemble e as prob. irão depender de que ensemble específico estamos considerando

Mais formalmente: considere um ensemble Σ de sistemas idênticos S.
A prob. de ocorrer um evento x (um certo valor p/ alguma medida) é dado por

$$\mathcal{P}(x) = \lim_{\Omega(\Sigma)\to\infty} \frac{\Omega(x)}{\Omega(\Sigma)}$$

→ # de total de sistemas no ensemble Σ

$\Omega(x)$: # de sistemas que exibem o evento x

Fica claro que $0 \leq \mathcal{P}(x) \leq 1$

se $\mathcal{P}(x) = 0$ → nenhum sistema exibe o evento x, mesmo para os sistemas

se $\mathcal{P}(x) = 1$ → Todos os sistemas exibem o evento x.

Por exemplo p/ uma moeda: devemos ter N moedas idênticas e jogar cada uma de modo idêntico. então contar o número de vezes que temos o evento x (cara).

Essa é uma visão bem objetivo e experimental de probabilidade.

Uma outra possibilidade é considerar todos os possíveis resultados de um experimento, que devem ser mutuamente exclusivos.

Este conjunto é chamado <u>espaço amostral</u>

Para 1 moeda : { cara, coroa }

1 dado : { 1, 2, 3, 4, 5, 6 }

2 dados : { cara cara, cara coroa, coroa cara, coroa coroa }

Note que há mais de uma possibilidade de espaço amostral. Para o caso de 2 dados por exemplo podemos considerar: { 2 caras, 1 cara, 0 caras }

{ 0 caras, pelo menos 1 cara }

Só temos que tomar cuidado p/ que os resultados sejam mutuamente excludentes.

Cada resultado (ou evento) no espaço amostral é considerado um ponto do espaço amostral. E a cada evento associamos uma probabilidade.

ponto	cara	coroa	2 caras	1 cara	0 caras
prob	1/2	1/2	1/4	1/2	1/4
	espaço amostral uniforme		espaço amostral não uniforme		

Com isso definimos a prob. de 1 evento somando as prob. de cada ponto "favorável" ao evento

prob. de pelo menos 1 cara = prob. de 1 cara + prob. de 2 caras

$\frac{1}{2} + \frac{1}{4} = \frac{3}{4}$

Note que para espaços amostrais uniformes temos

$$\text{prob.} = \frac{\text{\# de eventos (pontos) favoráveis ao evento}}{\text{\# total de pontos do espaço amostral}}$$

Nessa definição temos que assumir a priori o conceito de prob.

No caso do espaço amostral uniforme pode-se argumentar em termos de simetria. Não temos nenhuma informação que leve a previlegiar um dos eventos, assim devem ser equiprováveis

Essas interpretações de prob. são chamados de frequentistas: associam prob. com a freq. que um evento ocorre. Existem outras definições ou interpretações que são mais subjetivas. A principal delas é a Bayesiana, baseado em um trabalho do clérigo e matemático amador Thomas Bayes de 1763

Na interpretação frequentista (ou observacional) a a distr. de prob. é uma propriedade do sistema sendo observado e não depende do observador. Já na interpretação de Bayes ela depende do observador.

Independente da interpretação probabilidades devem obedecer a algumas regras (podem ser vistas como axiomas)

1) Se existem N possíveis eventos x_i então

$$\sum_{i=1}^{N} P(x_i) = 1$$ condição de normalização

↳ cada sistema do ensemble ter que exibir um dos possíveis eventos

2) Se temos dois eventos distintos x_1 e x_2 (ou seja são mutuamente excludentes)

$$P(x_1 \text{ ou } x_2) = P(x_1) + P(x_2)$$

↳ prob. de obter 1 ou 3 quando jogo 1 único dado é $1/6 + 1/6 = 2/6 = 1/3$

3) Para dois eventos independentes (um não influencia o outro)

$$P(x_1 \text{ e } x_2) = P(x_1)\,P(x_2)$$

↳ prob. que ambos ocorrão. Pode ser um após o outro, ou ao mesmo tempo, já que são independentes

x_1 e x_2 é um evento composto

Exemplo: prob. de obter 1 num dado e 2 em outro

$$P(1 \text{ e } 2) = 1/6 \; 1/6 = 1/36$$

Note que a ordem é importante, já que não consideramos a possibilidade de obter 2 no primeiro e 1 no segundo. Estamos identificando os dados!

Se usarmos a def. de prob. como

$$\frac{\#\text{ de eventos } E}{\#\text{ total de eventos}}$$ → assumimos que todos os eventos são equiprováveis

Podemos facilmente deduzir as regras acima

de eventos onde x_1 ou x_2 ocorrem = # de eventos x_1 + # de eventos x_2

No caso do dado

espaço amostral é $\{1,2,3,4,5,6\}$ → 6 eventos

de eventos onde ocorre 1 ou 2 → 2 eventos

Logo $P(1 \text{ ou } 2) = 2/6 = P(1) + P(2) = 1/6 + 1/6$

No caso de dois eventos x_1 e x_2 ocorrem tenho que considerar o espaço amostral de um experimento em 2 sistemas. Logo o # de de eventos o quadrado do caso de 1 único sistema

E o número de eventos onde ocorre x_1 e x_2 é o produto dos # de eventos individuais.

No caso de 2 dados

espaço amostral: 11, 12, 13, 14, 15, 16
↓ 36 elementos
21, 22, ... 26
31, 32, ... 36
⋮
61, 62, ..., 66

de eventos com 1 e depois 2 → 1 → $P(1 \text{ e } 2) = 1/36$

Um caso mais interessante é prob de obter

$x_1 \leq 3$ e $x_2 \leq 2$

$P(x_1 \leq 3) = \frac{3}{6} = \frac{1}{2}$ há três possib, $x_1 = 1, 2$ ou 3

$P(x_2 \leq 2) = \frac{2}{6} = \frac{1}{3}$ há duas possib $x_2 = 1$ ou 2

$P(x_1 \leq 3 \text{ e } x_2 \leq 2) = \frac{6}{36} = \frac{1}{6} = \frac{1}{2}\,\frac{1}{3}$

↳ há 6 possibibilidades e fica claro que é o produto do # de possib. de x_1 vezes o # de possib. de x_2: posso combinar cada possib de x_1 com a outra de x_2

Nesse caso fica claro pq a prob. do evento conjunto fatora nas prob. dos eventos individuais

Como já mencionado, independente da interpretação probabilidades devem obedecer a algumas regras básicas

1) $P(i) \geq 0$ 2) $\sum P(i) = 1$

2) $P(i \text{ ou } j) = P(i) + P(j)$ p/ eventos i e j mutuamente excludentes

2) $P(i \text{ e } j) = P(i)\,P(j)$ p/ eventos independentes

Em 1933 o matemático russo A.N. Kolmogorov formulou um conjunto completo de axiomas para a definição matemática de probabilidade (são as três primeiras regras)

Formalmente pode-se definir um espaço de probabilidades usando o espaço amostral, os eventos como pontos, e as probabilidades associadas aos eventos como "medidas" neste espaço

→ A distribuição de prob. $\mathcal{P}(x)$ de uma dada variável aleatória x, é a descrição mais completa do sistema (do ponto de vista estatístico)

É interessante definir o valor esperado (ou valor médio ou primeiro momento) da variável aleatória. É a soma ponderada de todos os valores possíveis

Definição:

→ tb se usa a notação $E[X]$ ou $\langle x \rangle$

$$\bar{x} \equiv \frac{\sum_i P(x_i)\,x_i}{\sum_i P(x_i)}$$

⇝ caso $\mathcal{P}(x_i)$ não esteja normalizado.

$\bar{x}$ fornece informação sobre a "localização" de $\mathcal{P}(x)$ ou em torno de quais valores x_i está distribuído

O nome valor esperado sugere que $\bar{x}$ é o valor que esperamos obter num evento. No entanto devemos ser cuidadosos com o que queremos dizer com isso

No caso de um dado de 6 faces com iguais prop. para cada uma ($\rho = 1/6$). O valor médio será

$$\bar{x} = 1/6\,(1) + 1/6\,(2) + \dots + 1/6\,(6) = 3,5$$

. Porém o valor 3,5 não é um evento possível; nunca iremos observá-lo em um lançamento. Neste sentido não é o valor que esperamos obter. A priori tampouco podemos esperar obter mais valores em torno de 3,5

O significado de $\bar{x} = 3,5$ é que se lançarmos o dado N vezes e calcularmos o valor médio obtido (a média aritmética dos N resultados) ele irá convergir para $\bar{x}$ conforme $N \to \infty$ (isso é garantido pela lei dos grandes números que discutiremos depois). Assim num lançamento de N dados (ou após lançar N vezes o mesmo dado) esperamos obter $\bar{x}$ como o valor esperado calculado e nesse sentido dizemos valor esperado. Outro exemplo é o caso de moedas, onde esperamos obter metade caras e metade coroas (valor esperado é "nulo") Nesse caso quanto maior o número de lançamentos mais confidente podemos estar de que o número de caras será igual ao de coroas

Note que estamos comparando o valor esperado teórico $\bar{x}$ com o valor médio obtido em um número FINITO N de observações (por isso usar o nome valor esperado para $\bar{x}$ teórico, e média para o obtido em N eventos pode causar menos confusão. No entanto não garanto manter este padrão)

Logo $\bar{x}$ não é o valor esperado no sentido habitual.

Num lançamento (ou evento aleatório) a melhor previsão que você pode fazer é usar o valor mais provável ($\rho(x)$ é máximo)

Como veremos em muitos casos o valor médio é igual ao mais provável (o que ajuda a confusão)

Mencionamos que $\bar{x}$ dá a localização de $\rho(x)$ já que os valores de x_i estão distribuídos ao redor de $\bar{x}$.

Podemos perguntar quão dispersos os pontos estão em relação a $\bar{x}$. Um modo de quantificar isso seria obter

$\overline{\Delta x} \equiv \overline{(x - \bar{x})}$: desvio médio

Mas $\overline{(x - \bar{x})} = \bar{x} - \bar{x} = 0$ (a média da soma é a soma das médias)

↳ da própria definição da média temos desvios positivos e negativos que se cancelam.

Como estamos interessados somente na magnitude do desvio devemos obter

$$\overline{|x_i - \bar{x}|}$$

Porém esta grandeza é matematicamente um pouco inconveniente por usar a função módulo.

Uma segunda opção e melhor para

$\overline{\Delta x^2} \equiv \overline{(x - \bar{x})^2}$ chamado de desvio quadrático médio ou dispersão ou variância ou segundo momento (em relação a $\bar{x}$)

(+) se usa a letra $\sigma_x^2 \equiv \overline{\Delta x^2}$

$\overline{\Delta x^2}$ da uma ideia da dispersão

Note que

- $\overline{\Delta x^2} \geq 0$ já que é uma soma de termos ≥ 0
- $\overline{\Delta x^2} = 0 \underset{\text{se e somente se}}{\Longleftrightarrow} x_i = \bar{x}$ p/ $\forall$ i
- $\overline{\Delta x^2} = \overline{(x - \bar{x})^2} = \overline{(x^2 - 2x\bar{x} + \bar{x}^2)}$
$= \overline{x^2} - 2\bar{x}\,\bar{x} + \bar{x}^2$
$= \overline{x^2} - \bar{x}^2$

$\overline{\Delta x^2}$ quantifica quão dispersos os pontos estão em relação a $\bar{x}$. Se ele for pequeno a maior parte dos pontos está perto da média. Assim podemos definir intervalos de confidência: um intervalo onde a maior parte dos pontos está. Para a distribuição gaussiana (ou normal) por exemplo sabemos que $\simeq 67\%$ dos "pontos" estão no intervalo $\bar{x} \pm \sqrt{\overline{\Delta x^2}}$.

Assim $\overline{\Delta x^2}$ é um valor esperado em 2 sentidos

1) Se obtivermos a média aritmética dos valores de $(x-\bar{x})^2$ numa amostra de N eventos, esperamos que tal valor convirja p/ $\overline{\Delta x^2}$ no limite de $N \to \infty$.

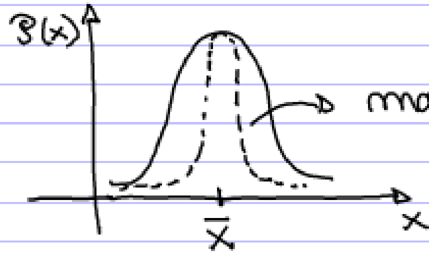
2) Esperamos encontrar uma fração dos pontos no intervalo $\bar{x} \pm \sqrt{\overline{\Delta x^2}}$ (a desigualdade de Chebyshev quantifica isso: $P(|x - \bar{x}| \geq k\sigma_x) \leq 1/k^2$)

Logo pode-se usar $\overline{\Delta x^2}$ como uma estimativa do desvio médio em x. Na verdade devemos usar

$\sqrt{\overline{\Delta x^2}}$ → tem a mesma unidade de x

↳ chamado de desvio padrão (tb se usa a notação σ_x. Tb conhecido com RMS, root mean square, de Δx)

$\sqrt{\overline{\Delta x^2}}$ mede a dispersão da distribuição



→ mais concentrado, menor σ_x

É importante notar que tem mais significado físico olhar para o valor relativo do desvio

$$\frac{\sqrt{\overline{\Delta x^2}}}{\overline{x}} \rightarrow \text{desvio padrão relativo}$$

↳ comparar $\sqrt{\overline{\Delta x^2}}$ com $\overline{x}$

→ Isso porque o valor de $\sqrt{\overline{\Delta x^2}}$ pode aumentar simplesmente porque o valor absoluto de $\overline{x}$ aumentou

↳ um desvio de 10 reais em 100 é "muito maior" que um desvio de 1000 em 100.000

Por completeza, deve-se mencionar que tb se define os chamados n-ésimos momentos de $S(x)$

$$\overline{x^n} \equiv \sum_i S(x_i)\, x_i^n$$

↳ as vezes se usa $(x_i - \overline{x})^n$

Como dito anteriormente a média ou valor esperado é o primeiro momento.

Dados todos os $\overline{x^n}$ podemos reconstruir (obter) a distribuição de prob. original

Por último, pode se definir valores esperados para uma função de x

$$\overline{f(x)} = \sum_i f(x_i)\, S(x_i)$$

E pode-se mostrar que (propriedades que já usamos acima)

i) $\overline{f(x) + g(x)} = \overline{f(x)} + \overline{g(x)}$

ii) $\overline{c\, f(x)} = c\, \overline{f(x)}$

Interpretações e Teoria de Informação

Essa definição que demos de probabilidade é empírica. Mas como já comentado há outros modos de obter a prob de um evento.

A interpretação frequentista tem alguns "problemas".

- assume que os eventos tem a mesma prob., ou seja, assume prob. p/ definir a prob.
- assume que as freq. serão as mesmas num novo experimento no futuro
- não temos como saber quantas medidas são suficientes para uma boa estimativa

Note que podemos usar o argumento de simetria para dizer que a prob. dos eventos é igual: não temos, a priori, nenhuma razão para dizer que as prob. sejam diferentes.

Usamos o conceito de probabilidade o tempo todo no nosso dia a dia, quando tomamos decisões. Temos assim uma intuição sobre probabilidade. Probabilidade está relacionado com o nosso grau de confiança de que um evento irá ocorrer: a prob. é uma medida do grau de confiança que temos de que o evento irá ocorrer.

Nesse caso a prob. de um dado evento irá depender do no conhecimento prévio que temos. Por exemplo, se jogarmos uma moeda 100 vezes e obtermos 100 caras em seguida, podemos usar essa informação como evidência de que a moeda não é perfeita e portanto a prob. de obter outra cara é alta.

Outro exemplo. Eu e você jogamos cada um uma moeda. Qual a prob. de ter 2 caras? Se eu vejo que minha moeda é coroa, vou dizer que a prob. é zero. Se não conheço os resultados de nenhuma das duas moedas irei dizer que a prob. é 1/4. Se sei que minha moeda é cara, então direi que a prob. é 1/2.

Assim sempre "calculamos" uma probabilidade condicional: prob. do evento E dado a informação I
↳ $P(E|I)$

Pessoas racionais com a mesma informação deveriam chegar a mesma conclusão sobre a prob. de um evento. Essa ideia é chamada de um "aposta coerente".

A aposta coerente deve ser considerada virtual. uma pessoa pode considerar um evento 99,999% provável e ainda se recusar a apostar R$1.000.000,00 contra R$1,00 se R$1.000.000,00 for muito mais dinheiro do que ele tem. Mas poderíamos convencer essa pessoa que essa aposta seria justa se ela tivesse uma quantidade infinita de dinheiro

Ou seja, decisões dependem não só na prob. como um grau de confiança de que um evento irá ocorrer, mas tb na importância subjetiva das consequências (nesse caso a quantidade de dinheiro)

↳ estimativas de prob devem ser separadas de outras questões subjetivas envolvidas na conseq. do evento.

Nessa interpretação, prob. estão diretamente relacionados com o conceito de informação
Considere dois experimentos cada um com dois resultados E_1, E_2 com prob. P_1 e P_2: jogar uma moeda, por exemplo.
No primeiro experimento temos $P_1 = P_2 = 1/2$ (moeda perfeita)
No segundo experimento temos $P_1 = 1/5$ e $P_2 = 4/5$ (moeda viciada)

Nossa intuição diria que o primeiro experimento é mais incerto que o segundo (não?)

Consideremos outros 2 experimentos. No terceiro há quatro possíveis resultados com $P_1 = P_2 = P_3 = P_4 = 1/4$ e no quarto há seis possíveis resultados com $P_1 = P_2 = P_3 = P_4 = P_5 = P_6$. Apesar de que em ambos os casos os resultados são equiprováveis, diríamos que o quarto experimento é mais incerto que o terceiro, já que há mais possibilidades.

Podemos tornar essa intuição mais "precisa" introduzindo uma medida de incerteza que seja consistente com nossa intuição.
Queremos definir uma função da distribuição de prob. que meça quanta incerteza há: $S(P_1, P_2, \dots P_i, \dots)$.
No caso onde os eventos são equiprováveis, $P_1 = P_2 = \dots = P_i = 1/\Omega$ com Ω o # total de possíveis resultados, temos $S(1/\Omega, \dots, 1/\Omega)$ ou simplesmente $S(\Omega)$
Nesse caso $S(\Omega)$ tem que satisfazer algumas condições simples

1) $S(\Omega = 1) = 0$ se só há uma possibilidade ñ há incerteza

2) $S(\Omega_1) > S(\Omega_2)$ se $\Omega_1 > \Omega_2$ $\to$ $S(\Omega)$ aumenta monotonicamente com Ω

No caso de múltiplos eventos: jogar um dado com Ω_1 possíveis resultados, e uma moeda com Ω_2 possíveis resultados; todos equiprováveis. O número total de possíveis resultados é $\Omega = \Omega_1 \Omega_2$. Se o resultado do dado é conhecido a incerteza em relação a ele é zero mas ainda há incerteza associada a moeda. Assim devemos ter

3) $S(\Omega_1 \Omega_2) = S(\Omega_1) + S(\Omega_2)$ (Aditividade)

Pode-se mostrar que há uma única função que satisfaz a essas propriedades (ver Cap. 3 do Harvey Gould). Ela tem a forma

$$S(x) = A \ln x + B$$

B deve ser nulo p/ satisfazer a condição 1) e A é arbitrário e assim escolhemos $A = 1$.
Logo p/ prob. iguais temos

$$S(\Omega) = \ln \Omega$$

No caso onde as prob. não são iguais tb pode-se encontrar a função que obedece as propriedades acima. Nesse caso tem-se

$$S = -\sum_i p_i \ln P_i \quad \to \text{se reduz a } \ln \Omega \text{ para } P_i = 1/\Omega$$

Pode-se ver que se um dos resultados é certo $P_i = 1$ e $P_j = 0$ p/ $j \neq i$ então $S = -1 \ln 1 = 0$
Pode-se mostrar tb que o caso $P_i = 1/\Omega$, prob. iguais, maximiza a função S. Assim pode-se usar um "princípio da máxima incerteza ou ignorância" para justificar o uso das prob. iguais a priori

A expressão para S foi proposta Claude Shannon em 1949 nos seus trabalhos sobre teoria estatística de comunicação.
Shannon define a informação i_e trazida por um evento com prob. P_e como

$$i_e = -K \ln P_e$$

Aqui uma informação constitue um ou mais eventos escolhidos em relação a um conjunto finito de eventos possíveis. Uma informação é mais importante quanto menor for a prob. dela ocorrer (como no exemplo dos dois experimentos acima)

Se o evento ainda não ocorreu a possibilidade que o sistema traga informação correspondente ao evento e é o produto da informação que ele tem caso ocorra e a prob. de ocorrer efetivamente:

$$I_e = -K P_e \ln P_e$$

Considerando todos os eventos possíveis no sistema a possibilidade que o sistema leve informação ao exterior se pudéssemos distinguir cada um dos eventos possíveis é

$$-K \sum_e P_e \ln P_e$$

Antes de observar os eventos, podemos considerar que a expressão acima (S a menos de uma constante) representa a falta de informação, ou nosso grau de ignorância, sobre o sistema. Assim a expressão acima é quanto obtemos de informação, em média, quando observamos um evento. Ao observar um dado evento e, nosso grau de ignorância diminue, e devemos excluir este evento da soma acima.