**Universidade de São Paulo**

**Faculdade de Filosofia, Letras e Ciências Humanas**

**Departamento de Ciência Política**

FLS 5028 – Métodos Quantitativos e Técnicas de Pesquisa em Ciência Política

FLP 0406 – Métodos e Técnicas de Pesquisa em Ciência Política

1º Semestre de 2024

Prof. Glauco Peres da Silva

Laboratório 4 – Famílias de Distribuição

O exercício de determinar a probabilidade de um evento qualquer depende da variável a que se refere e há um sem-número de possibilidades de utilização devido às diferentes especificações matemáticas para o cálculo destas probabilidades. O laboratório de hoje tem como objetivo explorar os usos deste contexto para aquilo que nos interessa enquanto pesquisadores em ciências sociais.

Partiremos de algumas definições importantes.

De acordo com um importante axioma, temos a seguinte situação:

Suponha que tenhamos para um evento qualquer um espaço amostral dado por

$$S=\left\{s\_{1},…,s\_{n}\right\}.$$

Podemos definir uma função de probabilidade *P* com domínio em S que satisfaz:

1. $P(A)\geq 0$, para todo $A\in S$;
2. $P\left(S\right)=1$;
3. Se subconjuntos de $A\in S$ forem disjuntos dois a dois, a probabilidade da união destes subconjuntos é igual à soma das probabilidades individuais. (*Ou seja, se subconjuntos de A tiverem sua intersecção igual a Ф (vazio), a probabilidade de A é igual à soma das probabilidades individuais*).
* Assim, temos uma primeira informação importante: qualquer função que satisfaça às condições acima pode ser uma função de probabilidade. Isto nos permite supor que existam um sem-número delas, já que as condições acima são muito gerais.

Uma outra importante situação para nós se relaciona ao tipo de variável que construímos diante de cada experimento. Lembre-se do caso da aula passada em que escrevemos para o experimento do lançamento de dois dados, variáveis aleatórias diferentes.

Considere a situação em que a nossa variável aleatória é a soma destes dois dados. A partir de um espaço amostral *S* de um experimento (o lançamento de dois dados), construímos um conjunto de valores possíveis χ da variável aleatória, ou seja, um novo espaço amostral com os valores $χ=\left\{x\_{1},…,x\_{n}\right\}$ – no caso, o resultado da soma de dois dados = $\left\{2,…,12\right\}$. Com isso, podemos calcular a probabilidade de $X=x\_{i}$ a partir da associação entre as possibilidades de ocorrência em *S* que satisfazem $X=x\_{i}$.

A partir desta relação, podemos calcular a $P\left(X=x\_{i}\right)$.

* Então, temos duas situações cuja conjugação nos importa. Há diferentes funções que podem ser chamadas de funções de probabilidade e, para um mesmo evento, podemos criar diferentes tipos de variáveis. Estas constatações nos levam a considerar que cada tipo de variável pode estar associado a uma função diferente.

Por isto, os estatísticos falam da existência de diferentes famílias de funções de probabilidade.

Essa função de probabilidade depende primordialmente de a variável considerada ser discreta ou ser contínua. No primeiro caso, ela será chamada de *função massa de probabilidade (fmp)* ou só *função de probabilidade (fp)* e no segundo, de *função densidade de probabilidade (fdp)*.

Vamos considerar uma variável discreta. O exemplo mais trivial é o da distribuição uniforme, cujo exemplo básico é o do lançamento de um dado equilibrado. Vamos definir uma variável aleatória *X* como sendo aquela que computa o resultado do lançamento de um dado equilibrado de seis faces. Neste caso, a função que indica a probabilidade é expressa por:

$$P\left(X=x|N\right)=\frac{1}{N}, x=1,2,…,N,$$

onde *N* é um número inteiro especificado. Essa distribuição coloca massa igual em cada um dos resultados $1,2,…,N$.

1. Qual seria o histograma para uma variável como essa?

Uma outra distribuição de interesse é a chamada binomial. Neste caso, a variável é computada a partir de um evento com dois resultados possíveis (chamado usualmente de *sucesso* e *fracasso*) ao qual se associam valores respectivamente 1 e 0. Um problema que exemplifica o uso desta função é o experimento de lançar uma moeda equilibrada *n* vezes. Neste caso, definimos a variável aleatória X como o número de vezes que saiu o resultado *cara*, nos três lançamentos.

1. Apresente a enumeração completa do valor de X para cada item do espaço amostral, se *n = 3*;
2. Determine o conjunto de valores para a variável aleatória;
3. Qual a probabilidade de ocorrência de cada um dos valores de X?

A função que representa esta situação pode ser expressa pela expressão:

$$P\left(X=x|n,p\right)=\left(\begin{matrix}n\\x\end{matrix}\right)p^{x}\left(1-p\right)^{n-x}, x=0,1,2,…,n,$$

em que *X* é chamada de uma *variável aleatória binomial (n, p)*.

Para variáveis contínuas, a situação é semelhante. Mas há uma consideração importante: ao invés de especificarmos $P\left(X=x\_{i}\right)$ usualmente criamos uma desigualdade como $P\left(X\leq x\_{i}\right)$, pois para variáveis contínuas, a probabilidade de a variável assumir um determinado valor específico é sempre 0; ou seja, $P\left(X=x\_{i}\right)=0$ .

A fim de evitar esta dificuldade, especifica-se apenas a função como *f(x)*.

Então, uma variável contínua uniforme teria como *fdp* a seguinte função:

$$f\left(x|a,b\right)=\left\{\begin{matrix}\frac{1}{b-a}, se x\in \left[a,b\right];\\ 0 caso contrário \end{matrix}\right.$$

1. Qual seria o histograma desta função? Esboce algumas possibilidades.

Para todas as funções acima, mesmo nas *fmp*, você pode notar que elas foram escritas a partir da seguinte notação: $f\left(.|.\right)$. Na parte à esquerda da barra, indicamos para a *fmp* a expressão $X=x\_{i}$ e na *fdp* apenas a variável *x*. É onde se indica qual a variável utilizada na função correspondente (que poderia ser indicada por outra letra, como *y*, p.ex.). Após a “|” indicou-se o parâmetro da função. Algumas funções não têm parâmetro, enquanto outras têm até 3; em alguns casos os parâmetros possuem significados, enquanto outros não. Por exemplo, para a distribuição normal, a *fdp* é dada por:

$$f\left(y | μ,σ^{2}\right)=\frac{1}{\sqrt{2π}σ}e^{\frac{-\left(y-μ\right)^{2}}{\left(2σ^{2}\right)}}, -\infty <y<\infty $$

A função, embora seja bastante “feia”, possui uma variável (*y*) e dois parâmetros que possuem interpretação: o primeiro, μ, é a média e o segundo, $σ^{2}$ , é a variância de *y*.

A família de uma distribuição normal recebe este nome (*família*) porque se os parâmetros mudam, as distribuições também mudam, embora a função original se mantenha.

1. O que acontece com o gráfico de uma função normal se cada um dos parâmetros se altera?

Por fim, um terceiro componente que precisa ser observado em relação às notações das funções de probabilidade é o domínio da variável a que a *fmp* ou a *fdp* se referem. Por exemplo, no caso da distribuição uniforme contínua, a variável precisa estar especificada entre *a* e *b* para que a probabilidade seja diferente de 0; no caso da normal, está entre $-\infty $ e $+\infty $. No caso das variáveis discretas, por vezes, algumas distribuições permitem que a variável aleatória assuma o valor 0; outras não.

Isto nos leva a estabelecer alguns critérios que permitem que julguemos a pertinência de utilizar uma determinada distribuição para criar um modelo:

1º. O fenômeno que se quer modelar:

Qual seu fato gerador? Qual a variável aleatória será definida?

2º. A variável gerada a partir do fenômeno de interesse é discreta ou contínua?

3º. Qual o intervalo de valores que a variável pode assumir?

Diante destas avaliações, temos a possibilidade de escolher uma distribuição que atenda aos critérios de interesse.

Vamos treinar para alguns casos. Considere o gráfico no qual há informações sobre diversas distribuições de probabilidade (no moodle, está no link Famílias de Distribuição (Casella e Berger)). Baseado nele, quais distribuições você poderia utilizar para modelar as seguintes variáveis:

A. Um lançamento de um dado?

B. Duração de uma guerra?

C. Número de chamadas necessárias até completar 20 questionários?

D. Número de deputados reeleitos no estado s?

E. Variação no peso do indivíduo i no ano t?

F. Duração da vida?

G. Proporção da renda de uma pessoa i gasta em queijo?