Probabilidade e Estatística - Parte X

População e Amostra

Definição 1 (População): *A totalidade das observações que podem ser feitas em determinado experimento, sejam elas finitas ou infinitas, constitui o que se chama de população.*

Algumas populações são tão grandes que são consideradas, na prática, infinitamente grandes. Cada observação feita na população é uma variável aleatória X_i que é relacionada com a função probabilidade $f(x)$.

Em estatística, estamos interessados em chegar à conclusões acerca da população, partindo-se das variáveis aleatórias coletadas. Em muitos casos é impossível ou impraticável de se estudar toda a população, por isso que se tenta obter estas informações à partir dos dados observados, ou seja, da amostra.

Definição 2 (Amostra): *É um subconjunto da população.*

A amostra deve ser feita através da escolha de variáveis completamente aleatórias, a fim de evitar bias nas conclusões. Assim, consequentemente, os vários valores de X observados serão independentes entre si.

Classes de Problemas em Estatística

Os problemas em estatística podem ser separados em algumas classes principais:

Predição Nesta classe de problemas, tendo observado somente uma amostra de número relativamente pequeno de elementos da população, estamos interessados em saber qual é a probabilidade de que os próximos elementos observados tenham determinados valores. Algumas outras vezes, sabemos ou temos fortes indícios de quais são as funções densidades de probabiliades envolvidas com a variável aleatória observada, mas precisamos determinar qual é o valor provável dos parâmetros desta distribuição. Neste caso o problema é conhecido como problema de estimação.

de Decisão Neste tipo de problema, não estamos interessados em na estimação de determinado parâmetro, mas estamos interessados em saber se este parâmetro pode estar dentro de determinado limite. Por exemplo, queremos determinar se a taxa de falha na fabricação de determinado produto é menor do que 5% ou não. Para isso pegamos uma pequena amostra de todos os produtos fabricados mas precisamos inferir se a taxa observada nesta amostra condiz com a hipótese apresentada (taxa de falha menor do que 5%).

Desenho Experimental Dependendo do tipo de pergunta que se quer fazer sobre o observável X, podemos realizar o experimento de diversas maneiras diferentes. Isto pode ser obtido através de um estudo de desenho experimental, uma vez que a forma como o experimento é realizado pode facilitar ou até mesmo inviabilizar que determinadas perguntas possam ser feitas sobre a população.

Estatísticas

Definição 3 (Estatística): *Suponha que as variáveis observáveis de determinado experimento sejam $X_1, X_2, ..., X_n$. Seja r uma função real das n variáveis observáveis. Então a variável aleatória $r(X_1, X_2, ..., X_n)$ é chamada de uma* **estatística**.

Temos vários tipos de estatísticas, como por exemplo o valor médio $\bar{X}$, o menor valor X_{min}, a mediana $\tilde{X}$, etc...

Estatísticas de Localização

Definição 4 (Média da Amostra): *Sejam $X_1, X_2, ..., X_n$, n observações de uma população, a média da amostra $\bar{X}$ é dada por:*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Definição 5 (Mediana da Amostra): *Sejam $X_1, X_2, ..., X_n$, n observações de uma população, a mediana da amostra $\tilde{X}$ é dada por:*

$$\tilde{X} = \begin{cases} X_{(n+1)/2}, & \text{se } n \text{ for ímpar} \\ (1/2)[X_{n/2} + X_{(n+1)/2}], & \text{se } n \text{ for par} \end{cases}$$

Definição 6 (Moda): *Sejam $X_1, X_2, ..., X_n$, n observações de uma população, moda é o valor que ocorre com mais frequência entre todos os n valores.*

Medidas de Variabilidade

Definição 7 (Variância da Amostra): *Sejam $X_1, X_2, ..., X_n$, n observações de uma população, a variância da amostra S^2 é dada por:*

$$S^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left[X_i - \bar{X}\right]^2$$

Podemos re-escrever a expressão acima:

$$\begin{aligned}
S^2 &= \frac{1}{(n-1)} \sum_{i=1}^{n} \left[X_i - \bar{X}\right]^2 \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} \left[X_i^2 - 2X_i\bar{X} + \bar{X}^2\right] \\
&= \frac{1}{(n-1)} \left\{\sum_{i=1}^{n} \left[X_i^2\right] - 2\bar{X}\sum_{i=1}^{n}\left[X_i\right] + \bar{X}^2\sum_{i=1}^{n}\left[1\right]\right\} \\
&= \frac{1}{(n-1)} \left\{\sum_{i=1}^{n} \left[X_i^2\right] - 2n\bar{X}\bar{X} + n\bar{X}^2\right\} \\
&= \frac{1}{(n-1)} \left\{\sum_{i=1}^{n} \left[X_i^2\right] - n\bar{X}^2\right\}
\end{aligned}$$

Definição 8 (Desvio Padrão da Amostra): *O desvio padrão da amostra S é definido como:*

$$S = \sqrt{S^2}$$

A Lei dos Grandes Números

Um Exemplo Introdutório

Consideremos o caso de uma moeda justa. Neste caso, a probabilidade p de termos cara é dada por 1/2. Assim, o valor médio em 10 jogadas é 5. A probabiliade do número de caras em 10 jogadas ser exatamente 5 é dada por:

$$P(X = 5, n = 10) = \binom{10}{5} (0,5)^5 (0,5)^5 = 0,2461$$

Por outro lado a probabilidade do número de caras ser exatamente igual a 50 em 100 jogadas é dada por:

$$P(X = 50, n = 100) = \binom{100}{50} (0,5)^{50} (0,5)^{50} = 0,0796$$

De qualquer maneira, podemos calcular a probabilidade da soma estar entre 10% do valor médio em ambos os casos:

$$P(4 \leq X \leq 6, n = 10) = \sum_{i=4}^{6} \binom{10}{i} (0,5)^i (0,5)^{10-i} = 0,6563$$

e

$$P(40 \leq X \leq 60, n = 100) = \sum_{i=40}^{60} \binom{100}{i} (0,5)^i (0,5)^{100-i} = 0,9648.$$

Assim, podemos ver que quanto maior o número de jogadas, mais próximo do valor esperado vamos estar.

Desigualdades de Markov e Chebyshev

Teorema 1 (Desigualdade de Markov): *Suponha que X seja uma variável aleatória tal que $Pr(X \geq 0) = 1$. Então para cada número real $t > 0$,*

$$Pr(X \geq t) \leq \frac{E[X]}{t}$$

Prova: Temos, para uma distribuição discreta:

$$\begin{aligned}
\mathsf{E}[X] &= \sum_x x\, f(x) \\
&= \sum_{x<t} xf(x) + \sum_{x\geq t} xf(x) \\
&\geq \sum_{x\geq t} xf(x) \\
&\geq \sum_{x\geq t} tf(x) \\
&\geq t\sum_{x\geq t} f(x) \\
&\geq t\mathsf{Pr}(x \geq t)
\end{aligned}$$

Se usou acima o fato de que $x \geq 0$. Como $t \geq 0$, podemos isolar $\mathsf{Pr}(X \geq t)$ e temos a prova buscada.

Teorema 2 (Desigualdade de Chebyshev): *Suponha que X seja uma variável aleatória tal que $0 \leq Var(X) \leq \infty$. Então para cada número real $t > 0$,*

$$Pr(|X - \mu| \geq t) \leq \frac{Var(X)}{t^2}$$

Prova: Temos, para uma distribuição discreta:

$$\begin{aligned}
\mathsf{Var}(X) &= \sum_x (x-\mu)^2 f(x) \\
&= \sum_{|x-\mu|<t} (x-\mu)^2 f(x) + \sum_{|x-\mu|\geq t} (x-\mu)^2 f(x) \\
&\geq \sum_{|x-\mu|\geq t} (x-\mu)^2 f(x) \\
&\geq \sum_{|x-\mu|\geq t} t^2 f(x) \\
&\geq t^2 \sum_{|x-\mu|\geq t} f(x) \\
&\geq t^2\mathsf{Pr}(|x-\mu| \geq t)
\end{aligned}$$

Se usou acima o fato de que $x \geq 0$. Como $t^2 \geq 0$, podemos isolar $\mathsf{Pr}(|x-\mu| \geq t)$ e temos a prova buscada.

Propriedades da Média da Amostra

Imagine que peguemos n variáveis aleatórias de uma distribuição qualquer. O valor médio das amostras $\bar{X}_n$ é dado por:

$$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + ... + X_n).$$

Teorema 3: *Sejam X_1, X_2, ... , X_n, n variáveis aleatórias pegas de uma distribuição com média μ e variânvia σ^2 e seja então $\bar{X}_n$ a média destas variáveis, temos que $E[\bar{X}_n] = \mu$ e $Var(\bar{X}_n) = \sigma^2/n$.*

Prova: Para o caso da média, podemos facilmente ver que:

$$\begin{aligned}
\mathsf{E}[\bar{X}_n] &= \mathsf{E}\left[\frac{1}{n}(X_1 + X_2 + ... + X_n)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}[X_i] \\
&= \frac{n\mu}{n} \\
&= \mu
\end{aligned}$$

Por outro lado, temos:

$$\begin{aligned}\text{Var}\left(\bar{X}_n\right) &= \text{Var}\left(\frac{1}{n}(X_1 + X_2 + ... + X_n)\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}\left(X_i\right) \\ &= \frac{n\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n}\end{aligned}$$

Acima se utilizou o fato de que as variáveis X_i são todas independentes.

Distribuição das Médias

Este tipo de experimento ocorre quando uma amostra de tamanho n é coletada de uma população e um valor de $\bar{X}$ é computado. A seguir, são feitas mais séries de amostragens de tamanho n e os demais valores de $\bar{X}$ são computados. Isto faz com que surja uma distribuição de valores médios. Estes valores devem orbitar ao redor do valor μ que é a média da população, conforme visto acima. Por outro lado, a média tende a ficar mais próxima de μ do que um valor X_i isoladamente, porque a variância da média ($\text{Var}\left(\bar{X}\right) = \sigma^2/n$) é n vezes menor do que a variância (σ^2) de X.

Atualizando a Desigualdade de Chebyshev

Agora podemos atualizar a equação para a desigualdade de Chebyshev para a variável aleatória $\bar{X}$:

$$\text{Pr}(|\bar{X} - \mu| \geq t) \leq \frac{\sigma^2}{n\,t^2}$$

Exemplo 1

Considere que jogamos uma moeda justa por um número n de vezes. Qual deve ser o número mínimo de jogadas que deve ser feito a fim de que a fração do número de caras em relação à coroas esteja entre 40% e 60% com probabilidade maior do que 0,7? Faça as contas considerando a distribuição binomial e a desigualdade de Chebyshev

Vamos primeiramente utilizar a desigualdade. Podemos olhar este problema sob duas ópticas diferentes. Num primeiro momento, podemos analisar somente a variável $\bar{X}$. O valor esperado desta variável é $\mu = 0,5$ e sua variância é dada por $0,25/n$ (podemos ver isso facilmente utilizando-se a distribuição binomial). Assim:

$$\begin{aligned}\text{Pr}(0,4 \geq \bar{X} \leq 0,6) &= 1 - \text{Pr}(|\bar{X} - 0,5| \geq 0,1) \\ \text{Pr}(|\bar{X} - \mu| \geq t) &\leq \frac{\text{Var}\left(\bar{X}_n\right)}{t^2} \\ \text{Pr}(|\bar{X} - \mu| \leq t) &\geq 1 - \frac{0,25}{n\,t^2} \\ \text{Pr}(|\bar{X} - \mu| \leq 0,1) &\geq 1 - \frac{0,25}{n\,(0,1)^2} \\ &\geq 1 - \frac{25}{n} \\ &\geq 1 - \frac{25}{n}\end{aligned}$$

Por outro lado, podemos usar como estatística, o fato de que $T = X_1 + X_2 + ... + X_n$. Como a moeda é justa, temos $p = 0,5$. Como o número de caras é uma distribuição binomial, sabemos que o valor médio é dado por $\text{E}[T] = np = 0,5\,n$ e a variância é dada por $\text{Var}(T) = n\,p\,(1-p) = 0,25n$. Assim, por Chebyshev:

$$\begin{aligned}\text{Pr}(0,4n \geq T \leq 0,6n) &= \text{Pr}(|T - \mu| \geq 0,1n) \\ \text{Pr}(|T_n - \mu| \geq t) &\leq \frac{\text{Var}\left(T_n\right)}{t^2} \\ \text{Pr}(|T_n - \mu| \leq 0,1\,n) &\leq 1 - \frac{0,25\,n}{(0,1n)^2} \\ &\geq 1 - \frac{25}{n}\end{aligned}$$

Como esta probabilidade tem que ser maior do que 0,7, temos:

$$\begin{aligned}1 - \frac{25}{n} &\geq 0,7 \\ -\frac{25}{n} &\geq -0,3 \\ \frac{25}{n} &\leq 0,3 \\ \frac{25}{0,3} &\leq n \\ n &\geq 83,333...\end{aligned}$$

Logo $n \geq 84$. Por outro lado, se resolvermos este problema utilizando a distribuição binomial, veremos que $n = 15$ é a solução do problema pois:

$$\text{Pr}(0,4 \leq \bar{X}_n \leq 0.6) = \text{Pr}(6 \leq T \leq 9) = 0,70.$$

Lei dos Grandes Números

Definição 9 (Convergência em Probabilidade): *Se $Z_1, ..., Z_n$ for uma sequência de números aleatórios que convergem em b em probabilidade se para qualquer número $\epsilon > 0$:*

$$\lim_{n\to\infty} Pr(|Z_n - b| < \epsilon) = 1$$

Denotamos a propriedade acima como:

$$Z_n \xrightarrow{p} b$$

Teorema 4 (Lei dos Grandes Números): *Seja X_1, X_2, ..., X_n números aleatórios pegos de uma distribuição com média μ e com variância finita, seja $\bar{X}_n$ a média desta amostra, então, $\bar{X}_n$ converge para μ em probabilidade, ou seja:*

$$\bar{X}_n \xrightarrow{p} \mu$$

Prova: Seja a variância de cada X_i igual a σ^2. Então segue da desigualdade de Chebyshev que para qualquer número $\epsilon > 0$:

$$\text{Pr}(|\bar{X}_n - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}.$$

E, portanto:

$$\lim_{n\to\infty} \text{Pr}(|\bar{X}_n - \mu| < \epsilon) = 1.$$

Assim, se um número grande de amostras aleatórias for pego de uma distribuição desconhecida, então a média dos números aleatórios escolhidos será uma boa estimativa para a média da distribuição desconhecida.

Teorema do Limite Central

Teorema 5 (Teorema do Limite Central): *Seja $\bar{X}$ a média de n variáveis aleatórias $X_1, X_2, ..., X_n$ pegas de uma distribuição com média μ e variância σ^2, então a forma limite de:*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

é a distribuição normal padrão.

A aproximação normal é boa para $n \geq 30$, desde que a população seja relativamente simétrica. Para $n < 30$ a aproximação é melhor quanto mais próxima de uma distribuição normal for $f(x)$.

Distribuição da Diferença de Duas Médias

Em determinadas situações temos duas amostras, a primeira delas com média μ_1 em n_1 amostras e a segunda com média μ_2 em n_2 amostras. Se σ_1^2 e σ_2^2 são as variâncias da população de cada uma das amostras, então podemos nos interessar pela variável $Y = \bar{X}_1 - \bar{X}_2$. Utilizando as ferramentas utilizadas anteriormente, vemos que:

$$\text{E}[Y] = \text{E}\left[\bar{X}_1\right] - \text{E}\left[\bar{X}_2\right] = \mu_1 - \mu_2$$

e

$$\text{Var}(Y) = \text{Var}\left(\bar{X}_1\right) + \text{Var}\left(\bar{X}_2\right) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

O Teorema do Limite Central pode ser extendido para este caso também.

Teorema 6: *Seja $\bar{X}_1$ a média de n_1 variáveis aleatórias pegas de uma distribuição com média μ_1 e variância σ_1^2. Seja $\bar{X}_2$ a média de n_2 variáveis aleatórias pegas de uma distribuição com média μ_2 e variância σ_2^2. Então a forma limite de:*

$$Z = \frac{\left[\bar{X}_1 - \bar{X}_2\right] - \left[\mu_1 - \mu_2\right]}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

é a distribuição normal padrão.

Novamente, se n_1 e n_2 forem maiores do que 30, então a aproximação acima é bastante razoável deste que as duas distribuições sejam simétricas. Para valores menores do que 30, a aproximação é melhor quanto mais próxima de distribuições normais estas forem.