Material Teórico - Módulo de INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

Normalização

Segundo Ano do Ensino Médio

Autor: Prof. Francisco Bruno Holanda
Revisor: Prof. Antonio Caminha Muniz Neto

8 de Agosto de 2020

1 Introdução

O objetivo dessa aula é apresentar as razões pelas quais as distribuições normais são tão importantes. Antes de chegarmos a tal ponto, desenvolveremos algumas nomenclaturas e simbologias fundamentais no estudo da Estatística Inferencial.

Definição 1. *Uma variável aleatória em um espaço amostral Ω é uma função $X : \Omega \to \mathbb{R}$ que associa cada evento de um experimento aleatório a um número real.*

Alguns exemplos de variáveis aleatórias:

- O tempo que um cliente gasta em uma loja.
- A expectativa de vida de uma pessoa nascida no Brasil.
- O peso[1] (em gramas) de um pacote de salgadinhos.

Uma vez que a cada experimento aleatório há uma probabilidade P associada, podemos definir a probabilidade da variável X estar entre dois valores reais distintos a e b como sendo $P(a \leq X \leq b)$.

Definição 2. *Uma variável aleatória X terá distribuição normal com média μ e desvio-padrão σ se $P(a \leq X \leq b)$ for igual à medida da área da região sob o gráfico da função*

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

acima do eixo das abscissas e situada entre as retas $x = a$ e $x = b$, para quaisquer a e b reais. Nesse caso, denotamos $X \sim \mathcal{N}(\mu, \sigma)$.

2 Normalização

O próximo resultado nos informa como as distribuições normais relacionam-se umas com as outras.

Teorema 3. *Se uma população X segue uma distribuição normal com média μ e desvio-padrão σ (em relação a um certo experimento), então*

$$\frac{X - \mu}{\sigma}$$

seguirá a distribuição normal padrão.

A demonstração desse teorema está fora do escopo deste material. Para saber mais, consulte os livros listados na seção de referências.

Este resultado nos permite utilizar a tabela normal padrão para descobrir as probabilidades de intervalo de qualquer outra distribuição normal. Por exemplo, se estivermos lidando com uma normal de média igual a 3 e desvio-padrão igual a 2, então a probabilidade de obtermos um valor menor ou igual a cinco será a mesmo de obter um resultado menor ou igual a $\frac{5-3}{2} = 1$ em uma distribuição normal padrão. Por sua vez, este valor, segundo a tabela, é igual a 0.8413.

[1] Aqui, estamos adotando o uso costumeiro, de chamar de *peso* o que, rigorosamente falando, é a *massa* do pacote.

3 O teorema do limite central

Teorema 4 (Teorema do limite central)**.** *Sejam $X_1, X_2, \ldots$ variáveis aleatórias independentes e identicamente distribuídas. Então, as variáveis dadas pelas médias amostrais até n,*

$$\overline{X}_n := \frac{X_1 + X_2 + \cdots + X_n}{n},$$

têm distribuições de probabilidade que se aproximam cada vez mais de uma curva normal com média μ e desvio-padrão $\frac{\sigma}{\sqrt{n}}$, à medida que n cresce.

A demonstração do Teorema do Limite Central (TLC) utiliza técnicas e conhecimentos que estão bem além do escopo deste material. Por outro lado, este resultado é fundamental para um estudo mais sofisticado de métodos estatísticos. A partir desse teorema, podemos explicar porque a média amostral é uma boa aproximação da média populacional de uma distribuição. Mais ainda, o resultado também nos informa qual é a curva de distruição da média amostral. Dessa forma, podemos calcular a probabilidade (*ex ante*) de que a média amostral pertença ao interior de um determinado intervalo em torno da média populacional.

Para explicar melhor estas ideias, apresentaremos os dois exemplos a seguir, que, apesar de pouco realísticos, são bem didáticos.

Exemplo 5. *A população de uma cidade tem* 10.000 *indivíduos. Suponha que foi realizado um censo completo nessa cidade (com todos os moradores) e foi constatado que a idade média das pessoas era igual a* 37 *anos, com desvio-padrão populacional $\sigma = 5$. Agora, suponha que Vanessa deseja fazer uma pesquisa e coletar (de forma aleatória e independente) a idade de um grupo de* 100 *pessoas dessa cidade. Qual será a probabilidade do valor encontrado por Vanessa estar no intervalo* $[36.5, 37.5]$*?*

Antes de resolvermos o exemplo anterior, é importante fazer o seguinte comentário: No enunciado, a frase "de forma aleatória e independente" garante que *não há viés de seleção* e que a hipótese de independência do TLC está garantida.

Solução. Pelo TLC, a distribuição das idades das 100 pessoas segue uma distribuição aproximadamente normal com média 37 e desvio-padrão igual a $\frac{5}{\sqrt{100}} = \frac{1}{2}$. Portanto, a representamos da seguinte forma:

$$\overline{X} \sim \mathcal{N}(37, 0.5).$$

Assim, normalizando esta variável, temos

$$\frac{\overline{X} - 37}{0.5} = Z \sim \mathcal{N}(0, 1).$$

Daí,

$$\begin{aligned}
&\text{Prob}(36.5 \leq \overline{X} \leq 37.5) = \\
&= \text{Prob}\left(\frac{36.5 - 37}{0.5} \leq Z \leq \frac{37.5 - 37}{0.5}\right) \\
&= \text{Prob}(-1 \leq Z \leq 1).
\end{aligned}$$

Usando a tabela normal-padrão, temos que $P(Z \leq 1) = 0.8413$. Assim, $P(Z > 1) = 1 - 0.8413 = 0.1587$. Por simetria, $P(Z < -1) = 0.1587$. Logo,

$$\text{Prob}(-1 \leq Z \leq 1) = 1 - 2 \times 0.1587 = 0.6826.$$

Ou seja, antes mesmo de fazer a pesquisa, Vanessa sabe que a probabilidade que a sua média amostral esteja no intervalo $[36.5, 37.5]$ é de $68,26\%$.

Agora, analisaremos um exercício similar ao anterior. A única diferença será o número de dados coletados na amostra.

Exemplo 6. *Considere uma população de uma cidade com* 10.000 *indivíduos. Suponha que foi realizado um censo completo nessa cidade (com todos os moradores) e foi constatado que a idade média das pessoas era igual a* 37 *anos, com desvio-padrão populacional* $\sigma = 5$*. Agora, suponha que Carlos deseja fazer uma pesquisa e coletar (de forma aleatória e independente) a idade de um grupo de* 400 *pessoas dessa cidade. Qual será a probabilidade do valor encontrado por Carlos estar no intervalo* $[36.5, 37.5]$*?*

Solução. Pelo TLC, a distribuição da idade média das 400 pessoas segue uma distribuição aproximadamente normal com média 37 e desvio-padrão igual a $\frac{5}{\sqrt{400}} = \frac{1}{4}$. Representamos da seguinte forma:

$$\overline{X} \sim \mathcal{N}(37, 0.25).$$

Assim, normalizando esta variável,

$$\frac{\overline{X} - 37}{0.25} = Z \sim \mathcal{N}(0, 1).$$

Daí,

$$\begin{aligned}
&\text{Prob}(36.5 \leq \overline{X} \leq 37.5) = \\
&= \text{Prob}\left(\frac{36.5 - 37}{0.25} \leq Z \leq \frac{37.5 - 37}{0.25}\right) \\
&= Prob(-2 \leq Z \leq 2).
\end{aligned}$$

Usando a tabela normal-padrão, temos que $P(Z \leq 2) = 0.9772$. Assim, $P(Z > 2) = 1 - 0.9772 = 0.0228$. Por simetria, $P(Z < -2) = 0.0228$. Logo,

$$\text{Prob}(-2 \leq Z \leq 2) = 1 - 2 \times 0.0228 = 0.9544.$$

Ou seja, antes mesmo de fazer a pesquisa, Carlos sabe que a probabilidade de achar um valor para sua média amostral que esteja no intervalo $[36.5, 37.5]$ é de $95,44\%$.

Veja que, ao aumentarmos o tamanho da amostra, aumentamos drasticamente a probabilidade de obter uma média amostral "próxima" da média populacional. Essa constatação é garantida pelo TLC, que afirma que a média amostral tem desvio-padrão igual a $\frac{\sigma}{\sqrt{n}}$, onde n é o tamanho da amostra. Assim, quanto maior for o valor de n, menor será a variância da média amostral.

Além disso, observe que o TLC não impõe nenhuma condição sobre a distribuição da população. Mesmo assim, para valores "suficientemente grandes" de n, a distribuição da média amostral será aproximadamente normal. Agora, surge a questão natural:

A partir de que valores, n é considerado suficientemente grande para podermos utilizar o TLC?.

Em geral, ficou-se convencionado o valor de $n \geq 30$ para realizarmos experimentos empíricos como os apresentados nos exercícios anteriores. A explicação dos motivos que levaram os pesquisadores a escolherem esse valor de n também estão fora do escopo desse material, mas podem ser encontrados em *www.nrcse.washington.edu/research/struts/chapter2.pdf*

4 Sugestões aos Professores

Separe pelo menos dois encontros de 50 minutos para ensinar o conteúdo deste material. Se possível, ministre a aula em um único encontro de 140 minutos. O objetivo dessa aula é ensinar seus alunos a utilizarem a tabela da distribuição normal padrão. Se possível, utilize outras disposições de tabelas (diferentes da apresentada na página 3) para que os alunos aprendam a identificar as probabilidades corretas. Uma possibilidade é apresentar a tabela de $P(-k \leq z \leq k)$, ao invés da tabela de $P(z \leq k)$. Nesse caso, os alunos terão que usar fortemente a propriedade de simetria da curva normal.

Referências

[1] João Ismael Pinheiro et al. *Estatística Básica: a arte de trabalhar com dados.* Campus, 2009.

[2] Pedro A. Morettin and Wilton de O. Bussab. *Estatística Básica.* Saraiva, 2010.

Tabela 1: Tabela correspondente à distribuição acumulada da curva normal padrão.

$$F(z) = P(z \leq k)$$

	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990