5
aula

DISTRIBUIÇÃO NORMAL - PARTE II

META

Apresentar a segunda parte do conteúdo de distribuição normal.

OBJETIVOS

Ao final desta aula, o aluno deverá:
analisar as características da distribuição normal;
definir variável normal padronizada;
determinar a probabilidade de valores e intervalos de valores a partir da função de densidade de probabilidade.

PRÉ-REQUISITOS

O aluno deverá compreender a relação entre o número de ocorrências mostrado no histograma de probabilidade, obtida pela distribuição normal.

INTRODUÇÃO

No final da última aula, conhecemos a forma matemática de f(x)dx, que é a probabilidade de ocorrência de um valor da variável no intervalo que vai de x a x+dx. Precisamos ter em mente que quando estamos frente a dados reais, como por exemplo, a dosagem de um fármaco em um lote de comprimidos, há um valor ótimo para esta variável. Porém, em termos práticos, trabalha-se com faixas aceitáveis: a dosagem deve manter-se dentro de um intervalo definido por limites, pois se houver somente **um valor** absoluto aceitável, a produção seria inviável. Além disso, vimos que os valores medidos dependem do método escolhido e sempre estarão sujeitos a erros aleatórios. Portanto, o que vamos ver nesta aula é como prever, a partir da distribuição normal, a probabilidade de que uma medida aleatória realizada resulte em valores dentro de uma determinada faixa.

Caro aluno ou querida aluna, para reforçar a aprendizagem, vamos rever alguns conceitos, destacando inicialmente as características relevantes da distribuição normal que vimos na aula passada.

DISTRIBUIÇÃO NORMAL

Distribuição Normal.

É perfeitamente simétrica em torno do ponto central, que é a média m; se observarmos a forma da curva, veremos que a "queda" da curva após o máximo, de cada lado, é a imagem especular do outro lado.

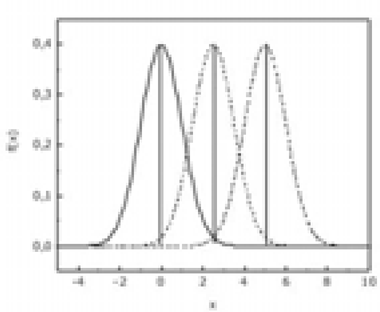
O valor de probabilidade é máximo sobre a média.

A três desvios-padrão da distância da média, a densidade de probabilidade reduz-se a zero, ou seja, a "queda" da curva de cada lado em torno da média vai até cerca de 3 desvios-padrão.

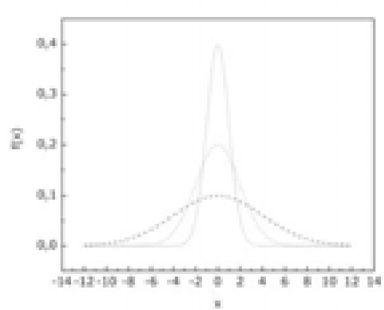
Agora, veja o que acontece com a curva de distribuição normal quando o valor da média e da variância aumentam.

Aumentando

A média:



A variância:



No caso do aumento somente da média (note que estamos estabelecendo que neste caso a variância não vai mudar), simplesmente a curva inteira se desloca para valores maiores no eixo x. A forma da curva não se altera. No caso do aumento da variância, vemos que o máximo da curva fica no mesmo lugar (no eixo x), mas a curva se alarga e se achata. A variância representa o "espalhamento" dos valores observados em relação à média: se a variância é baixa, indica que só foram observados valores muito próximos à média. Imaginemos um caso extremo: todos os valores iguais entre si, a curva seria um traço com altura igual ao número de observações. Vamos imaginar agora que estes valores começam a se modificar gradualmente: surge um valor ligeiramente acima da média e um ligeiramente abaixo dela. Isto faz com que nossa curva não seja mais um traço único, mas observamos um ligeiro alargamento dos dois lados. Juntamente com isso, a altura do traço inicial também abaixa um pouco, porque não corresponde mais a todas as observações. Conforme aumenta a variabilidade de valores, novos valores vão aparecendo, distanciando-se da média e fazendo sua altura baixar.

Agora vamo-nos preocupar com a questão das probabilidades de ocorrência de valores. Como podemos calcular tais probabilidades? A probabilidade de que o valor da variável aleatória seja observado no intervalo [a,b] será dada pela integral da função no intervalo que vai de a até b:

$$P(a < x < b) = \int_b^a f(x)dx$$

Será que para cada intervalo e cada problema específico vamos ter que calcular integrais? Na verdade, não! Mas, mesmo que fosse necessário, vários programas de computador do tipo planilha fazem estes cálculos facilmente. Mas existe outra alternativa ainda mais prática, que é a padronização da distribuição. Neste caso, o que se faz é construir a partir da variável x original, com sua média m e variância ó^2, uma nova variável chamada de z, obtida pela seguinte expressão:

Variável Normal Padronizada

$$z = \frac{x - \mu}{\sigma}$$

em que
x = variável aleatória com distribuição N(m,ó^2)
z = variável aleatória com distribuição N(0,1)

Para entender a vantagem, basta considerar (olhando para a equação) que subtraímos, inicialmente, de cada valor observado o valor da média. Isso faz com que todos os valores menores fiquem negativos, que a própria média fique igual a zero e que todos os valores maiores do que a média fiquem positivos. Notem também que quando fazemos a diferença entre cada valor observado e a média, estamos obtendo o desvio. Este entra no próprio cálculo da variância (elevado ao quadrado e depois dividido por N-1, em que N é o número de observações), que depois é dividido pelo desvio padrão ó. A vantagem do uso da variável padronizada e da distribuição normal padronizada é a curva toda ficar distribuída simetricamente em torno do zero e a variância desta distribuição ser igual a 1.

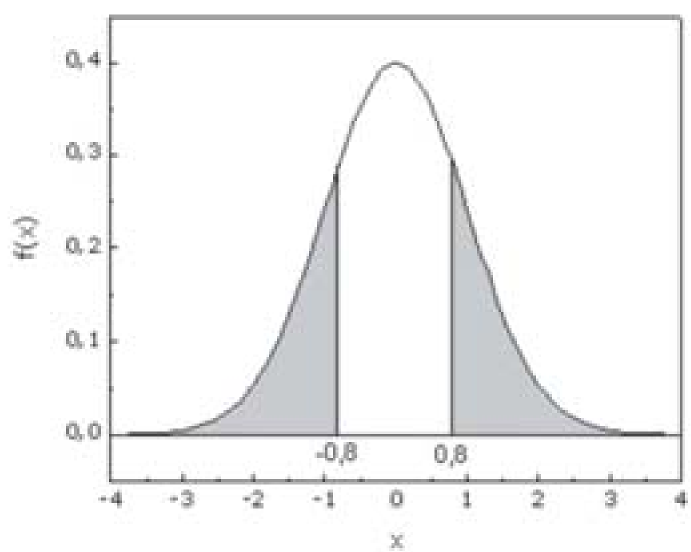
A forma matemática da função de distribuição normal padronizada fica:

Distribuição Normal Padronizada

$$f(z)dz = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dx$$

Você deve estar se perguntando como isso pode ser útil no cálculo das probabilidades. Muito bem! Qualquer distribuição normal pode ser padronizada. Para a distribuição padronizada, foi construída uma tabela que relaciona os valores de z (variável padronizada) com a área da cauda da curva (que é justamente a probabilidade, pois a integral nos permite calcular áreas). A seguir, mostramos uma porção da tabela completa das áreas da cauda da distribuição normal padronizada.

Como ela deve ser usada? Na primeira coluna (vertical), temos o valor de z com uma casa decimal; e, na primeira linha (horizontal), a segunda casa decimal correspondente ao valor de z. Por exemplo, o primeiro valor que mostramos na coluna é 0,8. Se usarmos a primeira posição da primeira linha (0,00), teremos 0,80; se usarmos a segunda posição da primeira linha (0,01), teremos 0,81. No primeiro exemplo (0,80), a célula que corresponde a este valor de z tem um valor de área igual a 0,2119. Esse valor significa o seguinte: da área total da curva, esse valor é a fração localizada à direita de z = 0,80. Como a curva é simétrica, uma área igual a esta está situada à esquerda de -0,8, como mostrado na figura abaixo:



Somando-se as duas áreas, temos: 0,4238, que significa que uma observação aleatória qualquer tem 42,38 % de probabilidade de situar-se à esquerda de -0,8 ou à direita de 0,8. Por conseqüência, a probabilidade de que a observação se situe no restante da curva é 57,62 %. Então, também podemos colocar a resposta da seguinte maneira: uma observação aleatória tem 57,62 % de probabilidade de situar-se entre -0,8 e 0,8.

Área da cauda da distribuição normal padronizada.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011
3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008

Finalmente, vamos considerar a probabilidade de uma observação aleatória situar-se entre limites não-simétricos (diferente de -0,8 e 0,8), como por exemplo, entre valores como -0,95 e 1,95. Para -0,95, apesar de a tabela não trazer valores negativos, podemos usar a mesma área correspondente a 0,95, que é de 0,1711; e para 1,95, temos uma área de 0,0256. Significa que à esquerda de -0,95, temos 17,11 % de probabilidade e, à direita de 1,95, temos 2,56 % de probabilidade. Subtraindo da área total as áreas das duas caudas, temos a probabilidade de que a observação tenha valores dentro dos limites especificados: 100 -17,11 - 2,56 = 80,33 %.

ATIVIDADES

1. Determine a probabilidade de que uma observação aleatória tenha valores à direita e à esquerda de:
a)1,43 e -1,43
b)1,92 e -1,92
c) 2,95 e -2,95
d) 1,31 e -0,62
e)1,04 e -0,86
f) 0,90 e -2,21
2. Represente graficamente usando a curva de distribuição normal.

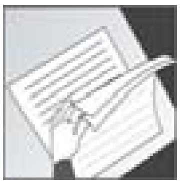
CONCLUSÃO

Podemos concluir, com base no conteúdo estudado nesta aula, que é possível, a partir de um método bastante simples, determinar as probabilidades de que observações aleatórias de variáveis que seguem uma distribuição normal ocorram dentro de certos intervalos. Isso é importante para que possam ser feitas previsões a respeito de populações, a partir de medidas de amostras.

RESUMO

Aprendemos que se a forma matemática para a distribuição normal for f(x), a expressão f(x)dx é a probabilidade de ocorrência de um valor da variável no intervalo que vai de x a x+dx. A importância disto é que quando calculamos estas probabilidades, estamos na realidade fazendo previsões sobre o restante da população, cujos valores nós não medimos diretamente em nossa amostra. Precisamos ter em mente que quando estamos frente a dados reais, como por exemplo, a dosagem de um fármaco em um lote de comprimidos, há um valor ótimo para esta variável. Porém, em termos práticos, trabalha-se com faixas aceitáveis: a dosagem deve manter dentro de um intervalo definido por limites. Então, como prever, a partir da distribuição normal, a probabilidade de que uma medida aleatória realizada resulte em valores dentro de uma determinada faixa? A probabilidade de que o valor da variável aleatória seja observado no intervalo [a,b] será dada pela integral da função no intervalo que vai de a até b. Será que para cada intervalo e cada problema específico vamos ter que calcular integrais? A saída que temos para isto vem do fato de que a distribuição normal pode ser padronizada. O que se faz é construir a partir da variável x original, com sua média m e variância ó^{2}, uma nova variável chamada de z. Para a distribuição padronizada foi construída uma tabela que relaciona os valores de z (variável padronizada) com a área da cauda da curva (que é justamente a probabilidade, pois a integral nos permite calcular áreas). Na primeira coluna (vertical), temos o valor de z com uma casa decimal e na primeira linha (horizontal), a segunda casa decimal correspondente ao valor de z. A grande vantagem disto é que apesar de cada amostra, de diferentes tipos de dados, ter seus valores específicos de médias e desvios, quando padronizamos as distribuições acabamos por deixar comparáveis diferentes conjuntos de dados que obedeçam a distribuição normal, com área total igual a 1, de modo que a probabilidade dependa somente da fração de área sob a curva.

PRÓXIMA AULA



Na próxima aula, conheceremos as razões pelas quais a distribuição normal é tão importante, no sentido de descrever o comportamento de muitos tipos de populações distintas. Além disso, vamos ver como podemos fazer estimativas de intervalos de confiança populacionais.

REFERÊNCIAS

BARROS NETO, B.; SCARMINIO, I. E.; BRUNS. R. E. **Planejamento e otimização de experimentos**. Campinas Editora da Unicamp, 1995.

BOX, G. E. P.; HUNTER, W. G.; HUNTER, J. S. **Statistics for experimenters. An introduction to design, data analysis and model building**. New York, Wiley: 1978.

BUSSAB, W. O.; MORETIN, P.A. **Estatística básica**. São Paulo: Atual, 1985.