Atividade 09 - Teorema do Limite Central

Objetivos:

1) Gerar valores aleatórios em planilhas.
2) Introduzir o teorema do limite central e o conceito de desvio padrão

PARTE 1

Análise da distribuição da média dos lançamentos de um número fixo de dados

Quando lançamos dados, temos ideia da probabilidade teórica de cada lançamento. Sabemos, por exemplo, que a média esperada para o lançamento de um dado de 6 faces é:

$$\frac{1+2+3+4+5+6}{6} = 3{,}5$$

Se lançarmos ao acaso M dados, a média dos valores provavelmente não será 3,5: ora será maior, ora será menor e, às vezes, igual a 3,5. O que vamos fazer é repetir N vezes o lançamento dos M dados estudando o comportamento da média à medida que N cresce. Esse tipo de estudo é fundamental em situações em que a média não é conhecida *a priori* e precisa ser determinada por meio de amostras.

Vamos trabalhar com uma quantidade de lançamentos da ordem de milhares. Para isso, vamos substituir os dados físicos por dados virtuais. Utilizaremos a função `aleatórioentre` das planilhas do Google. Uma célula com entrada `=aleatórioentre(1;6)` sorteia, usando um gerador de números aleatórios, um número entre 1 e 6, simulando um dado. Experimente o uso dessa função.

Observe que cada vez que você mexe na planilha, um novo número é sorteado. Essa função é uma função *volátil*[1]: o valor dela muda quando qualquer célula é alterada. Para evitar que esse valor fique mudando, utilizaremos uma caixa de seleção para desativar os sorteios enquanto editamos a planilha.

PARTE 2

Geração da distribuição das médias de 3, 6 e 9 dados

2.1) Crie uma nova aba na planilha e a renomeie para “Distribuição das Médias”. Na célula A1, insira uma “caixa de seleção” (Menu *Inserir* -> *Caixa de seleção*). Vamos utilizá-la para controlar a geração dos valores dos dados.

Na célula B1 escreva “Jogadas”. Preencha as células de B2 até B151 com os números de 1 a 150 usando o preenchimento automático. Alternativamente, você pode utilizar a função `=SEQUENCE(150)` na célula B2 e ela preencherá o restante.

2.2) Na célula C1, escreva “DADO 1”. Em C2, digite:

```
=SE($A$1;ALEATÓRIOENTRE(1;6);"---")
```

Perceba que toda vez em que a caixa de seleção é selecionada, um novo valor é sorteado. Isso será útil em nossa análise.

Posicione o mouse sobre o canto inferior direito da célula C2 até aparecer o símbolo "+" e clique duas vezes para o preenchimento automático até a linha 150. Você terá gerado 150 lances aleatórios do DADO 1.

---

[1] O Google Planilhas ainda não tem implementadas versões não-voláteis dessas funções, mas o LibreOffice e o MS EXcel têm algumas funções não voláteis de geradores de números aleatórios: `=ALEATORIO.NV()` e `=ALEATORIOENTRE.NV()`. Para contornar esta limitação vamos utilizar a função `SE()` e mostrar um conteúdo temporário enquanto não marcamos a caixa de seleção,e assim evitar que a planilha fique recalculando centenas de números aleatórios toda vez que editarmos uma célula.

2.3) Repita o procedimento 2.2 até a coluna L de modo que você tenha 10 dados.

2.4) Na célula M1 escreva "Média 2". Na célula M2 calcule a média aritmética do resultado da jogada 1 dos DADOS 1 e 2 (média dos valores das colunas C e D), isto é, calcule:

```
=média($C2:D2)
```

Faça o preenchimento automático para calcular a média aritmética das 150 jogadas para os DADOS 1 e 2. **Note** que colocamos um $ antes de C2 (E não antes do 2). Isso significa que você poderá fazer o preenchimento automático para a direita e economizar alguma digitação no próximo passo.

2.5) Repita o procedimento 2.4 nas colunas N a T, mas calcule a "Média 3" até "Média 10", ou seja, a média aritmética do resultado do lançamento de 3 a 10 dados.

Com isto, temos os dados necessários para iniciar a análise.

PARTE 3

Análise da distribuição das médias de 2 a 10 dados

Vimos nas aulas anteriores que os valores de um dado honesto têm igual probabilidade, mas o valor médio do lançamento de *M* dados tem uma distribuição semelhante a um sino, sendo mais concentrada em torno do valor médio. Além disso, vimos que a largura dessa distribuição vai ficando menor à medida em que *M* aumenta. Nesta parte da aula vamos procurar quantificar esse comportamento.

3.1) O desvio-padrão como uma medida da largura da distribuição

Ainda na planilha anterior, vamos considerar os valores relativos às médias de M dados, onde M vai de 2 a 10. Vamos representar os valores das médias por x_{M1}, x_{M2}, ..., x_{MN}, onde *N* é o número de vezes em que o lançamento dos M dados foi realizado. No caso, N = 150. O valor médio dos lançamentos de M dados N vezes será representado por $\overline{x}$:

$$\overline{x}_M = \frac{1}{N}\sum_{i=1}^{N} x_{Mi}$$

Em uma nova coluna "Tamanho das amostras", coloque os valores de 2 até 10. Na coluna do lado, chamada "Médias", obtenha as médias correspondentes, use a função =MÉDIA(M2:M), onde M é a coluna para 2 dados, N para 3 etc. Note que o intervalo inicia na linha 2 e não apresenta uma linha para término. Isso é importante para podermos alterar o número de lançamentos com mais facilidade[2]. Expresse essas médias com duas casas decimais.

A largura da distribuição é medida pela quantidade:

$$\sigma_M = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{Mi} - \overline{x}_M)^2}$$

Em uma nova coluna, denominada "Desvio-padrão", use a função =DESVPAD(M2:M) para calcular o desvio-padrão das médias para cada tamanho das amostras. Perceba que o desvio-padrão vai caindo à medida que *M* aumenta.

Para avaliar isto, crie uma nova coluna, com o título "1/raiz(Ndados)", e preencha com a fórmula =1/RAIZ(n), onde n é o valor da coluna "Tamanho das amostras". Faça um gráfico de dispersão cujo eixo x é 1/raiz(n) e o eixo y é o desvio-padrão. Faça o ajuste linear (através da linha de

[2] É possível que este recurso não funcione no LibreOffice e/ou no MS Excel. Nesse caso, utilize um intervalo longo o suficiente (cerca de mil linhas)

tendência ou da função `=PROJ.LINEAR()`). Compare o coeficiente angular com o desvio-padrão da distribuição de *um* dado. O que você pode dizer a respeito desse resultado, considerando o conceito de *desvio padrão da média*?

3.2) O teorema central do limite

Nesta parte, vamos desenhar sobre os histogramas o gráfico de uma função especial, denominada distribuição normal ou gaussiana. Sua definição depende dos valores da média e do desvio padrão:

$$\text{Normal}(x) = \frac{A}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\overline{x}}{\sigma}\right)^2}$$

O número *A* é um número que depende da quantidade de lançamentos feitos. Essa função está disponível nas planilhas do Google (função `Norm.Dist()`, que será explicada a seguir).

Infelizmente, as planilhas do Google não permitem a sobreposição direta de curvas ao histograma. Vamos contornar isso fazendo manualmente o histograma, efetuando explicitamente a contagem dos valores e gerando um gráfico de colunas, para os casos da média de 2 dados e da média de 10 dados.

Vamos precisar das coordenadas *x* e y dos pontos a serem colocados nos gráficos: a primeira coluna serão as coordenadas *x*, que são os valores que podemos obter com a média dos dados. A segunda coluna será a contagem do número de vezes que cada valor foi sorteado para formar o histograma, e a terceira coluna será o valor previsto pela distribuição normal dada acima.

Para construir essa tabela, adicione uma nova coluna com os valores possíveis para a média de 2 dados, que chamamos de x_{2i}. Perceba que os valores possíveis de x_{2i} estão entre 1 e 6, em intervalos de 0,5.[3] Use a função `=CONT.SE(intervalo de valores de` x_{Mi}`;valor procurado)` para obter, na coluna ao lado, quantas vezes cada valor de x_{Mi} aparece nos lançamentos feitos. Em outra coluna, vamos calcular:

```
=150*0,5*NORM.DIST(valor; média; desvio-padrão; cumulativo)
```

onde `valor` é um valor possível para a medida; `média` é o valor da média da distribuição para 2 dados; `desvio-padrão` é o desvio-padrão da distribuição para 2 dados e `cumulativo` é uma variável booleana que deve ser FALSO. Note que estamos usando o número *A* = *N**0,5 (com *N*=1000).

Selecione as três colunas e faça um gráfico do tipo “Gráfico de combinação”. Esse tipo de gráfico permite a sobreposição de uma curva a um gráfico de colunas. E então? A distribuição normal representa bem a distribuição das médias?

Repita o procedimento para as amostras de tamanho 10. Desta vez, os valores possíveis serão de 1 a 6, de 0,1 em, 0,1 (por quê?). Ase A = N*0,1. Você entende por que antes era A = N*0,5 e agora A = N*0,1? Analise os dados!

3.3) O teorema central do limite: Mais pontos

Faça uma cópia desta planilha, e modifique a função `=SEQUENCE()` para gerar 1000 linhas. Na última linha com sorteio dos dados (linha 151), selecione as colunas `C` a `U` e clique duas vezes na bolinha do canto para preencher automaticamente o restante da planilha. Ajuste, claro, o coeficiente na função `NORM.DIST` para refletir o novo número de dados.

Compare os gráficos com os do caso anterior: quais as semelhanças e diferenças? A largura do gráfico mudou? O histograma parece mais compatível com a distribuição normal? A qualidade do ajuste linear dos desvios-padrão mudou?

---

[3] É possível preencher automaticamente esses valores com alguma prestidigitação: `=SEQUENCE(N*(6-1)+1;1;1;1/N)`, onde `N` é o número de dados. (por que esta fórmula funciona?)