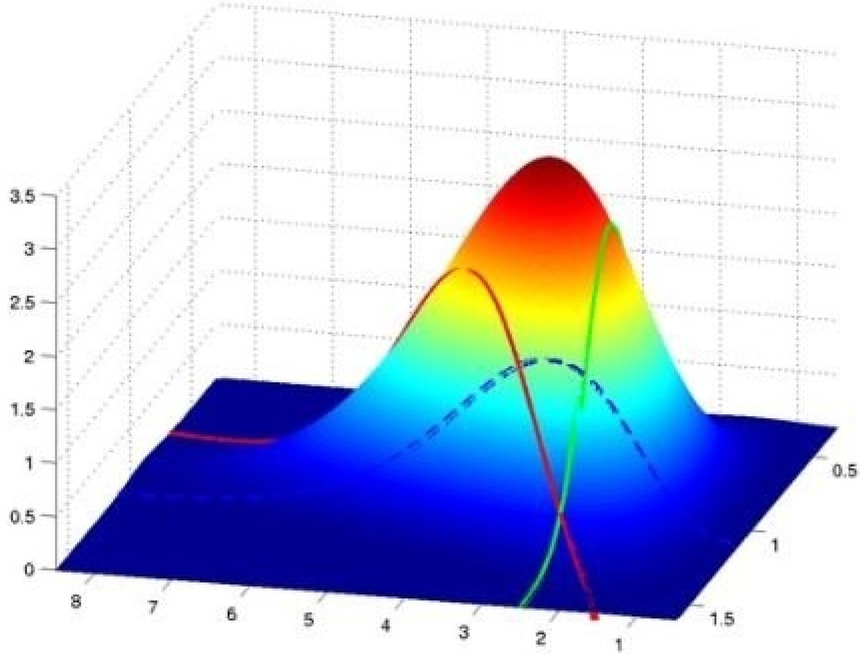
Tópicos em Gestão da Informação II

Aula 04 – **Medidas de posição relativa**



Prof. Dalton Martins
dmartins@gmail.com

Gestão da Informação
Faculdade de Informação e Comunicação
Universidade Federal de Goiás

Determinando sua posição: percentil

- Em geral, é interessante sabermos como um valor está em relação aos outros → é muito, pouco, etc...
- A maneira mais comum de informar a posição relativa é por meio do uso do percentil;
- O percentil é a porcentagem de indivíduos dentro de um conjunto de dados que estão abaixo da posição que você quer investigar;
- Se a posição desejada estiver no percentil 90º, por exemplo, isso significa que 90% dos dados estão abaixo de sua posição.

Calculando o percentil

- Para calcular o Kº percentil (onde K é um número entre 0 e 100), siga os passos abaixo:
  - Ordene todos os números do conjunto de dados do menor para o maior;
  - Multiplique a porcentagem, k, pela quantidade total de números, n;
  - Arredonde o resultado para o número inteiro mais próximo;
  - Conte os números do menor para o maior até encontrar o valor calculado no terceiro passo.

Open Office

- PERCENTIL
- Retorna o alfa-percentil dos valores dos dados em uma matriz. Um percentil retorna o valor de escala para uma série de dados que vai do menor (Alfa=0) ao maior valor (Alfa=1) de uma série de dados. Para Alfa = 25%, o percentil significa o quartil; Alfa = 50% é a MED.
- Sintaxe
- PERCENTIL(Dados; Alfa)
- Dados representa a matriz dos dados.
- Alfa representa a porcentagem da escala entre 0 e 1.
- Exemplo
- =PERCENTIL(A1:A50;0,1) representa o valor no conjunto de dados, que iguala 10% da escala total dos dados em A1:A50.

Exercício

- Para a planilha de exercício que estamos utilizando, construa uma matriz de percentil para as seguintes posições:
  - 10%
  - 20%
  - 50%
  - 80%
  - 90%
  - 95%
- Reflita: o que isso mostra da distribuição dos dados?

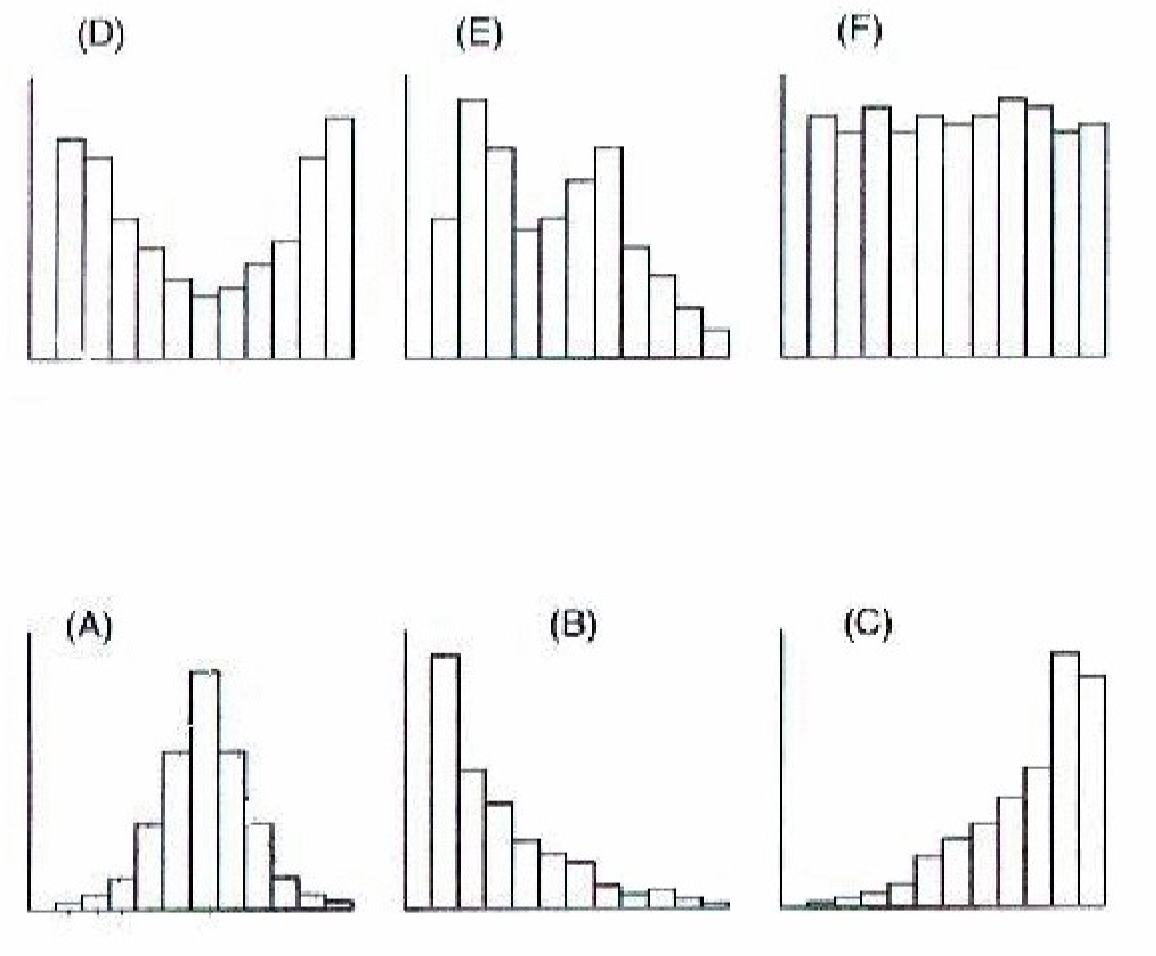
Posição relativa

- A única maneira de realmente conseguir interpretar os resultados estatísticos é ter algo com o que os comparar → precisamos então colocá-los em algum tipo de perspectiva;
- O primeiro passo para determinar onde um resultado em particular se encontra é conseguir uma lista ou uma imagem de todos os possíveis resultados que a variável pode assumir dentro de uma população → isso se chama distribuição!!!

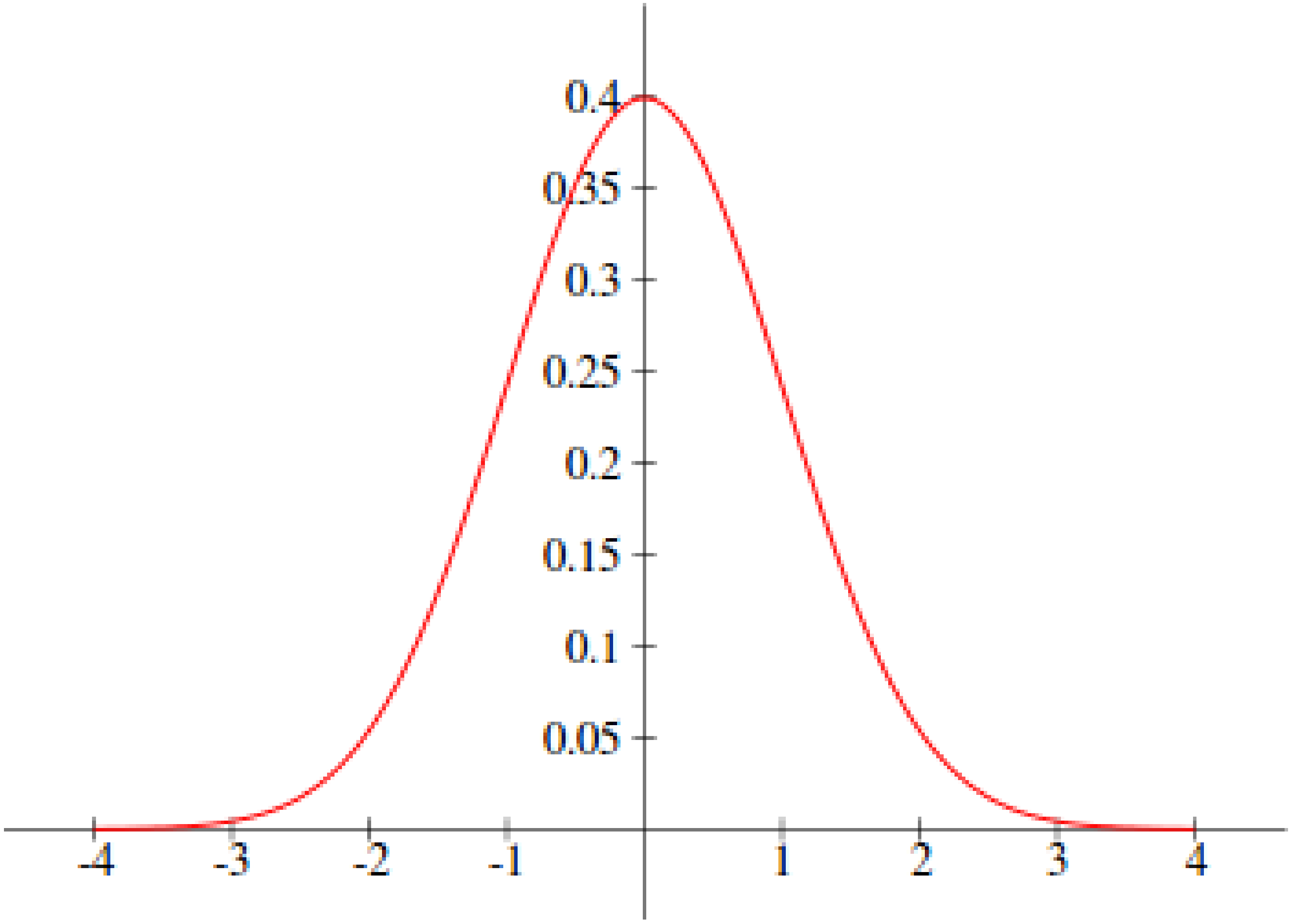
Distribuições

- Muitos tipos diferentes de distribuições são possíveis...

a) Distribuição Normal
b) Distribuição J
c) Distribuição J
d) Distribuição U
e) Distribuição Aleatória
f) Distribuição Uniforme



Curva em forma de sino



Também conhecida como Distribuição Normal, Gaussiana ou curva de Gauss!

História

- A distribuição normal foi introduzida pela primeira vez por Abraham de Moivre em um artigo no ano 1733, que foi reproduzido na segunda edição de seu The Doctrine of Chances (1738) no contexto da aproximação de distribuições binomiais para grandes valores de n. Seu resultado foi estendido por Laplace, em seu livro Analytical Theory of Probabilities (1812), e agora é chamado o teorema de Moivre-Laplace.
- Laplace usou a distribuição normal na análise de erros de experimentos. O importante método dos quadrados mínimos foi introduzido por Legendre, em 1805. Gauss, que alegou ter usado o método desde 1794, demonstrou-o rigorosamente em 1809 supondo uma distribuição normal para os erros.
- O nome "curva em forma de sino" ou "curva de sino" remonta a Esprit Jouffret que primeiro utilizou o termo "superfície de sino" em 1872 para um normal bivariada com componentes independentes (atentar que nem toda curva de sino é uma gaussiana). O nome "distribuição normal", foi inventado independentemente por Charles S. Peirce, Francis Galton e Wilhelm Lexis, por volta de 1875.

Caracterizando a distribuição normal

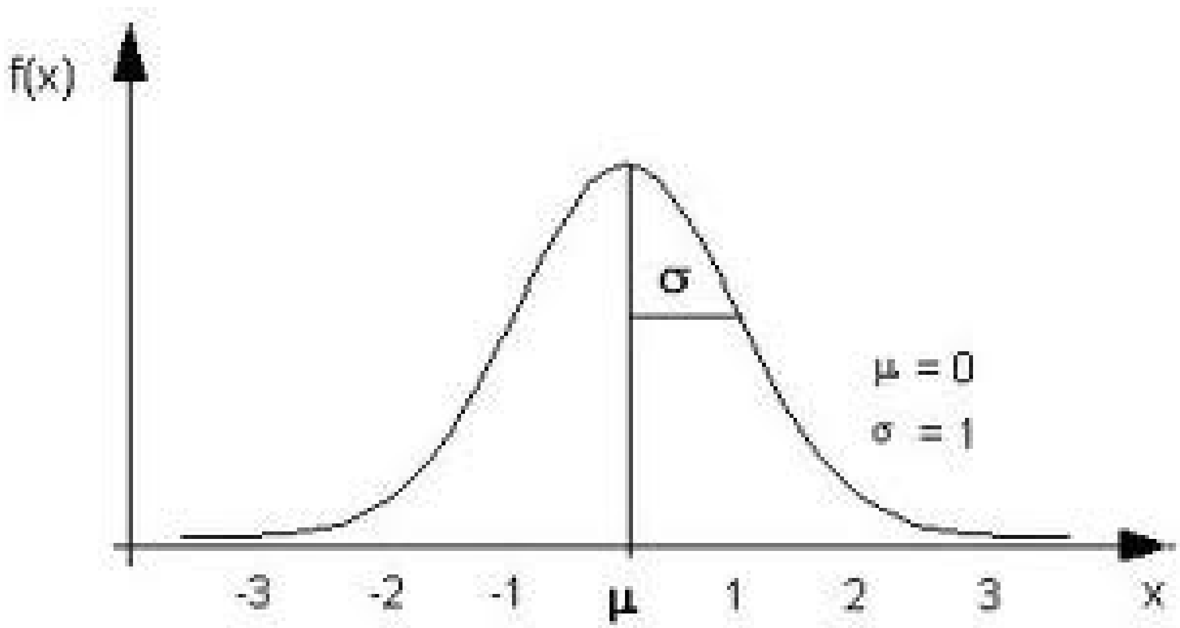
- Toda curva em forma de sino tem certas propriedades;
- Você pode utilizar essas propriedades para ajudar-lhe a determinar a posição relativa de qualquer resultado individual dentro da distribuição;
- Vejamos as propriedades:
  - A forma da curva é simétrica;
  - Possui uma elevação em seu centro e pontas que vão tanto para a direita quanto para a esquerda;
  - A média encontra-se exatamente no meio da distribuição;
  - A média e a mediana (em situações ideais) têm o mesmo valor graças a simetria da curva;
  - O desvio padrão representa o distância mais comum entre a média e todos os demais dados;
  - Cerca de 95% dos valores se encontram dentro de dois desvios padrões da média!

Importância

- É um importante modelo probabilístico para populações estatísticas relativas a variáveis contínuas. É de grande utilidade para uma **ampla gama de fenômenos naturais**, sejam eles físicos, ambientais, comportamentais, psicométricos, etc., além de erros de medida.
- O **teorema central do limite** mostra que, em situações muito gerais, a soma de uma grande quantidade de variáveis aleatórias ou realizações de uma variável aleatória converge para a distribuição normal.
- É a distribuição amostral, exata ou aproximada, de estatísticas de grande importância, como a **média amostral**, o que a torna extremamente relevante para realização de **inferências estatísticas**.
- Seu tratamento matemático é elegante e relativamente simples.
- **Aspectos computacionais**, como cálculo de probabilidades, estimação de parâmetros, cálculo de quantis, geração de valores, etc., estão muito bem desenvolvidos e os recursos estão amplamente disponíveis.
- A distribuição de uma grande quantidade de variáveis não normais, como muitas variáveis aleatórias discretas, **converge para a distribuição normal.**
- Muitas **variáveis não normais podem ser tratadas como normais** após transformações simples.
- A pressuposição de normalidade é exigida em muitos **procedimentos de inferência estatística** mas esses procedimentos frequentemente são robustos em relação à essa pressuposição.
- Nos modelos lineares e modelos não lineares pressupõe-se usualmente que os **erros têm distribuição normal**, muitas vezes agregando outras pressuposições.

Ponto Sela

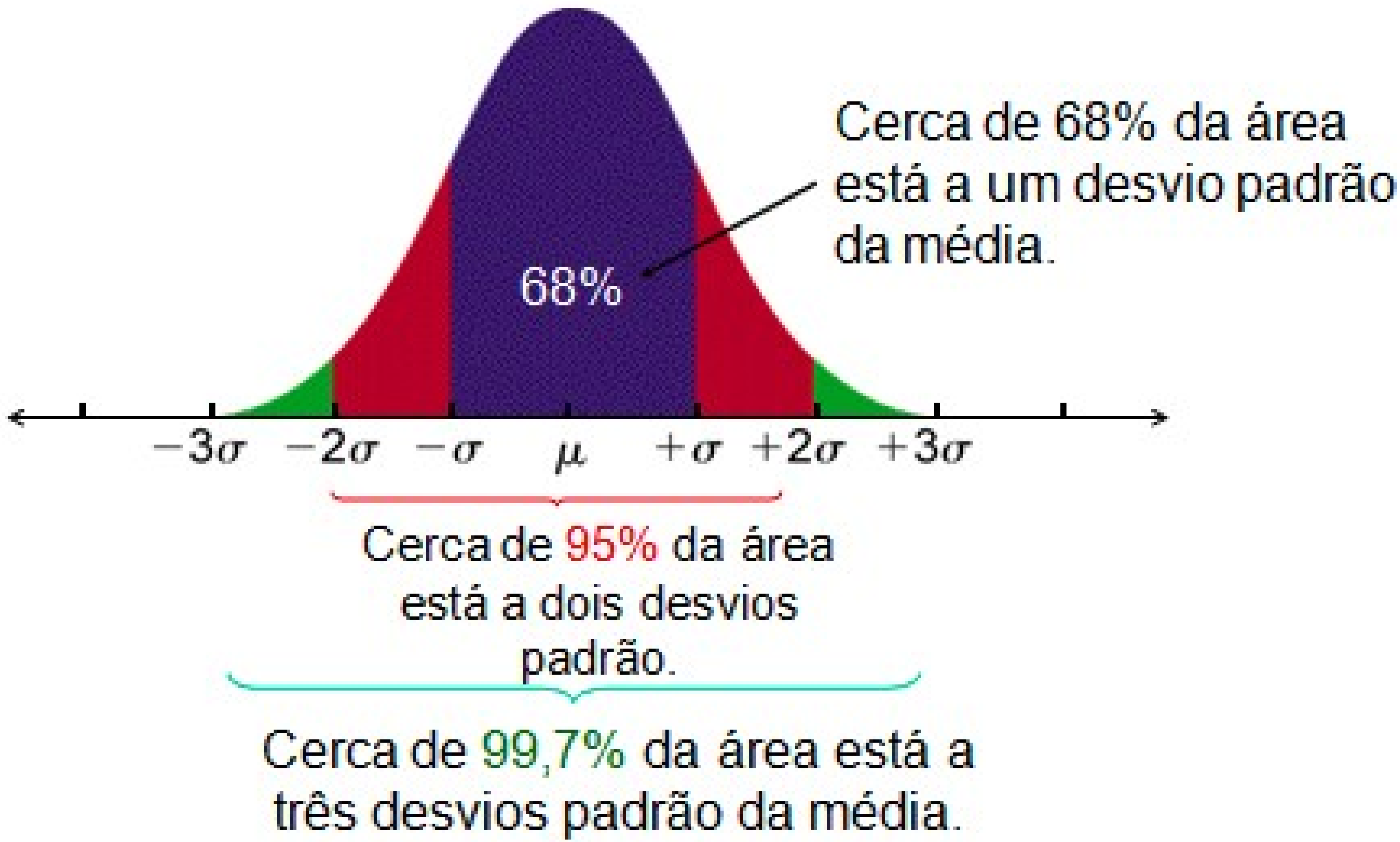
- Em uma distribuição normal, o desvio padrão tem um significado especial:
  - Determina a distância da média até um local dentro da distribuição conhecido como ponto sela.



Procurando pela maioria dos valores: a regra empírica

- A regra empírica diz que se a distribuição possui a forma de uma montanha, então:
  - 68% dos valores encontram-se dentro de um desvio padrão;
  - 95% dos valores encontram-se dentro de dois desvios padrões;
  - 99,7% dos valores encontram-se dentro de três desvios padrões.

Procurando pela maioria dos valores: a regra empírica



Exercício

- Em nossa planilha de teste, calcule:
  - A média e o desvio padrão dos dados (já calculados na aula anterior);
  - Quantas ocorrências estão dentro das faixas vistas anteriormente:
    - Quantas estão entre a média +- 1*desvio padrão;
    - Quantas estão entre a média +- 2*desvio padrão;
    - Quantas estão entre a média +- 3*desvio padrão.
  - O que esses resultados lhe permitem concluir de nossa distribuição?

Transformando em Escore Padrão

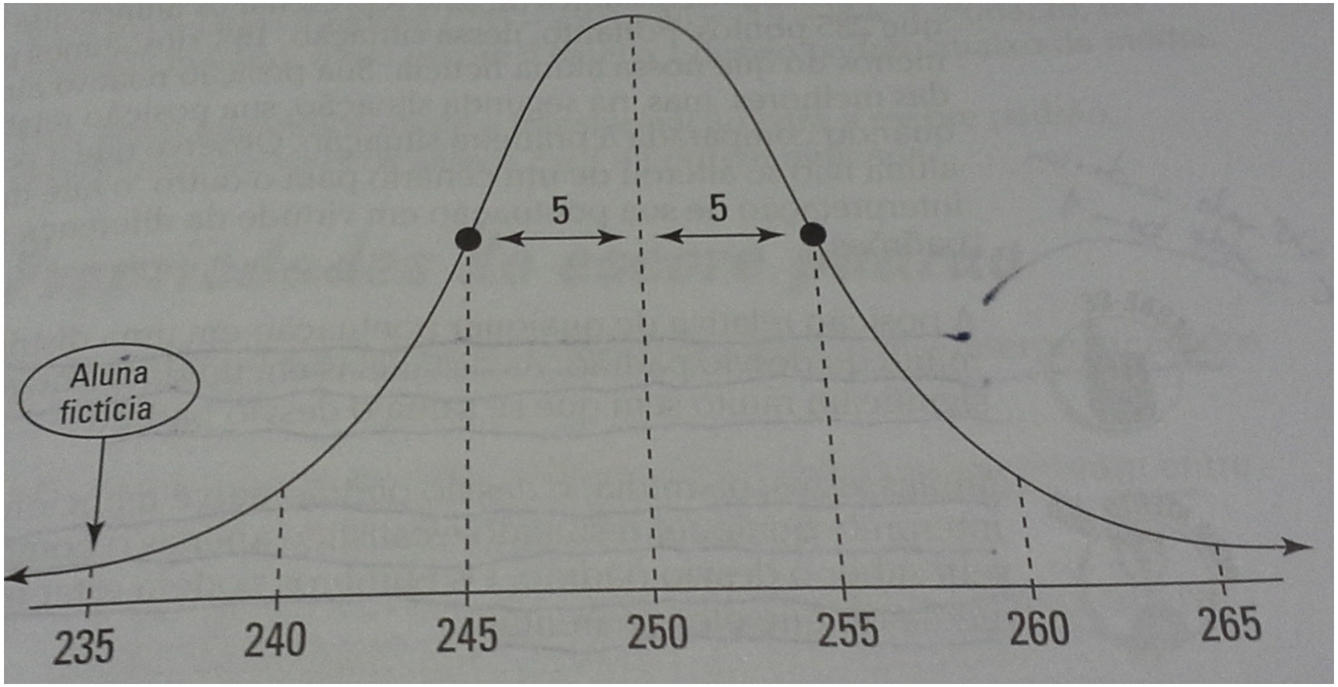
- Suponha que uma estudante de fisioterapia tenha feito um teste;
- Seus resultados indicam que sua pontuação foi de 235;
- Sabemos que a distribuição dos resultados neste teste foi uma distribuição normal;
- Será que essa pontuação da aluna foi boa, ruim ou ficou no meio do caminho?
- É impossível responder essa pergunta sem saber a posição desta estudante em relação aos outros que também fizeram a prova → é preciso determinar sua posição relativa!!!

Analisando o Desvio Padrão

- Existem muitas maneiras de determinar a posição relativa dessa estudante, algumas melhores do que outras:
  - Podemos observar essa pontuação em relação a pontuação total, que foi de 300 para a prova.
    - Porém, isso não compara ela com os outros, apenas com o total.
    - Continuamos sem saber sua posição relativa;
  - Outra maneira é comparar com a média. Suponha que a média tenha sido 250.
    - Sabemos que a pontuação da aluna está 15 pontos abaixo da média.
    - Mas, o que uma diferença de 15 pontos representa nessa situação específica?
- Para entender a posição relativa de qualquer valor em uma distribuição você tem que saber qual é o desvio padrão!

Analisando o Desvio Padrão

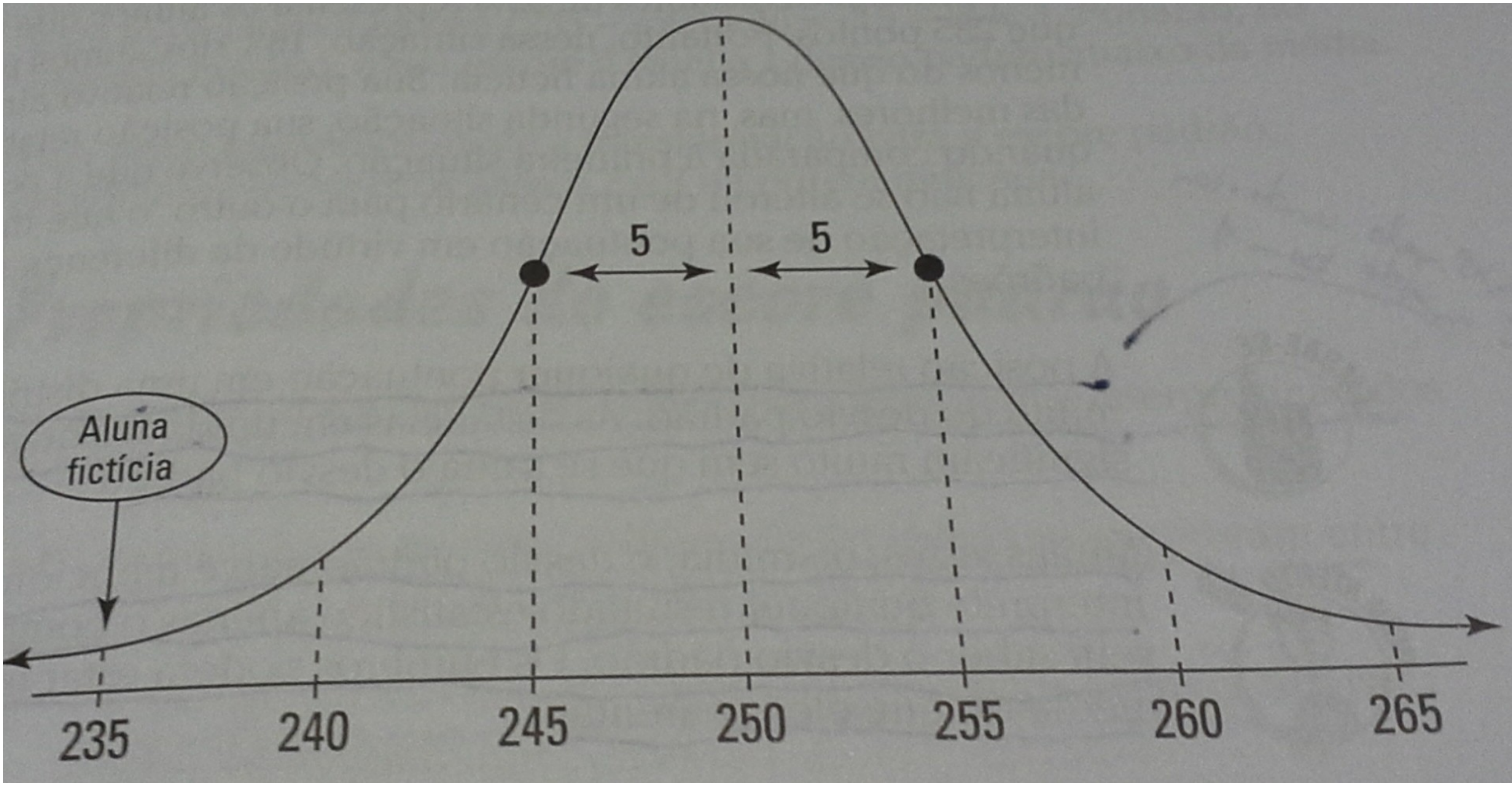
- Suponha que o Desvio Padrão para essa distribuição seja igual a 5 → significa que as pontuações foram bastante próximas umas das outras!
- 15 pontos aqui fazem muita diferença



Analisando o Desvio Padrão

- Suponha que o Desvio Padrão para essa distribuição seja igual a 5 → significa que as pontuações foram bastante próximas umas das outras!
- 15 pontos aqui fazem muita diferença;
- A pontuação da aluna é considerada bem abaixo da média pois está a 3 desvios padrões da média
  - Apenas uma minúscula fração de alunos pontuou menos do que ela → sabemos 99,7% dos alunos teve uma nota dentro da faixa de 3 desvios padrões (para cima ou para baixo)
  - Logo, apenas 0,3%/2 dos alunos teve pontuação igual ou menor, ou seja, apenas 0,15% deles pontuaram menos que 235.

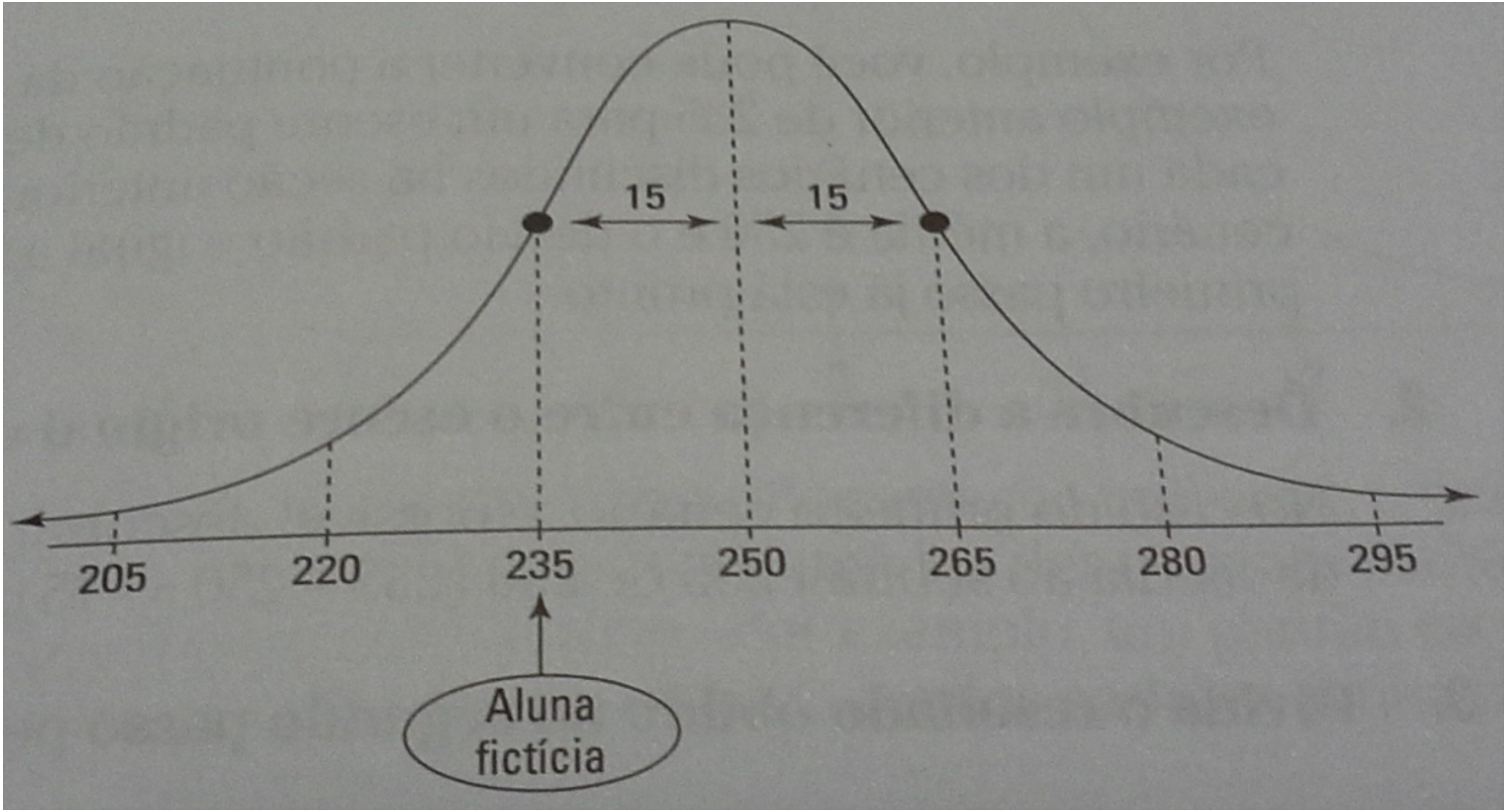
Analisando o Desvio Padrão



Analisando o Desvio Padrão

- Agora, suponha que o desvio padrão seja um valor diferente, vamos dizer, 15, sendo que a pontuação média continua 250.
- Um desvio padrão de 15 significa que as pontuações variaram muito mais (mais espalhadas) do que no exemplo anterior;
- Nessa caso, ter ficado 15 pontos abaixo da média não é tão ruim;
- Ela está a 1 desvio padrão da média → sabemos que 68% dos alunos teve nota dentro da faixa de um desvio padrão (para cima ou para baixo);
- Logo, 32% dos alunos está fora dessa faixa → 16% deles tirou nota menor que ela.
- Sua posição relativa melhorou substancialmente apenas analisando o desvio padrão!

Analisando o Desvio Padrão



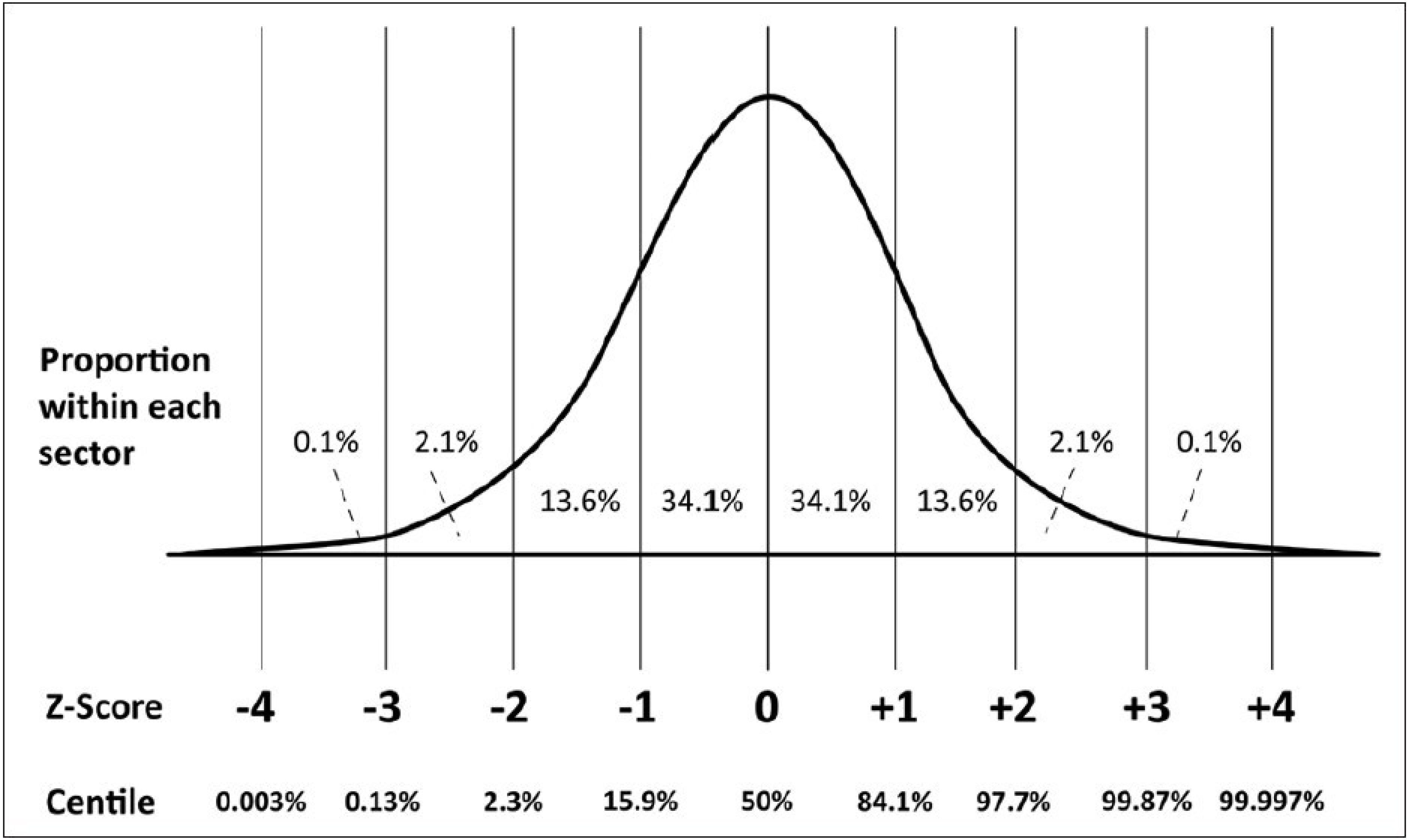
Calculando o Escore Padrão

- Para encontrar, mencionar e interpretar a posição relativa de qualquer valor em uma distribuição normal → transformamos a variável num escore padrão;
- O escore padrão é uma versão padronizada do escore original; ele representa o número de desvios padrões acima ou abaixo da média;
- Fórmula:
  - Escore padrão = (escore original – média) / desvio padrão

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

$\mu =$ Mean
$\sigma =$ Standard Deviation

Calculando o Escore Padrão



Propriedades do escore padrão

- As propriedades a seguir mostram-se muito úteis para a interpretação dos escores padrões:
  - Quase todos os escores padrões (99,7%) se encontram entre valores de -3 a +3, graças a regra empírica;
  - Um escore padrão negativo indica que o escore original estava abaixo da média;
  - Um escore padrão positivo indica que o escore original estava acima da média;
  - Um escore padrão igual a 0 indica que o escore original era a própria média;
  - Escores que vêm de uma distribuição normal, quando padronizados, têm uma distribuição normal especial com média 0 e desvio padrão 1 → distribuição normal padrão!

Propriedades do escore padrão

- Não compare resultados de diferentes distribuições sem antes transformar tudo em escores padrões!
- O escore padrão permite realizar uma comparação justa e imparcial dentro de uma mesma escala! → FUNDAMENTAL!!!

Exercício

- Utilizando nossa planilha de testes, faça as seguintes operações:
  - Crie uma nova planilha apenas com os Pesquisadores (exclua todas as empresas por seu CNPJ);
  - Apague todas as linhas que possuem valor 0 empenhado de pesquisadores → queremos apenas aqueles que receberam algum valor;
  - Ordene do maior para o menor e faça o gráfico de sua distribuição como vimos na aula passada;
  - Calcule a média e desvio padrão;
  - Crie uma nova coluna e calcule os escores padrões de todos os valores e faça um novo gráfico dessa distribuição;
  - Escolhe aleatoriamente 3 valores e calcule sua posição relativa indicando em que categoria de desvios padrões eles se encontram (1, 2 ou 3, apontando quantos valores estão acima e abaixo deles).