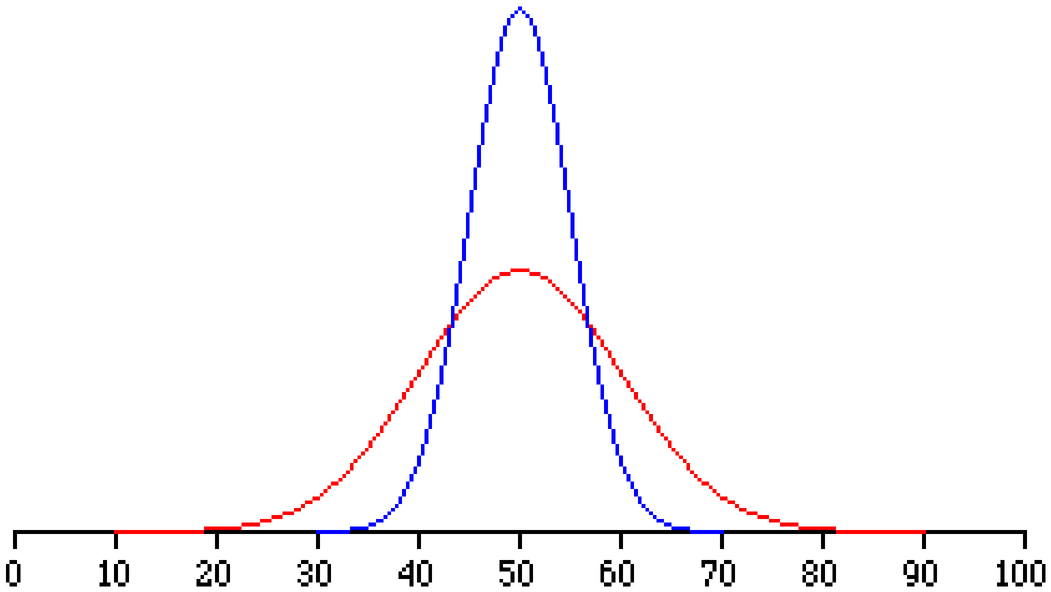
Tópicos em Gestão da Informação II

Aula 05 – **Variabilidade estatística**



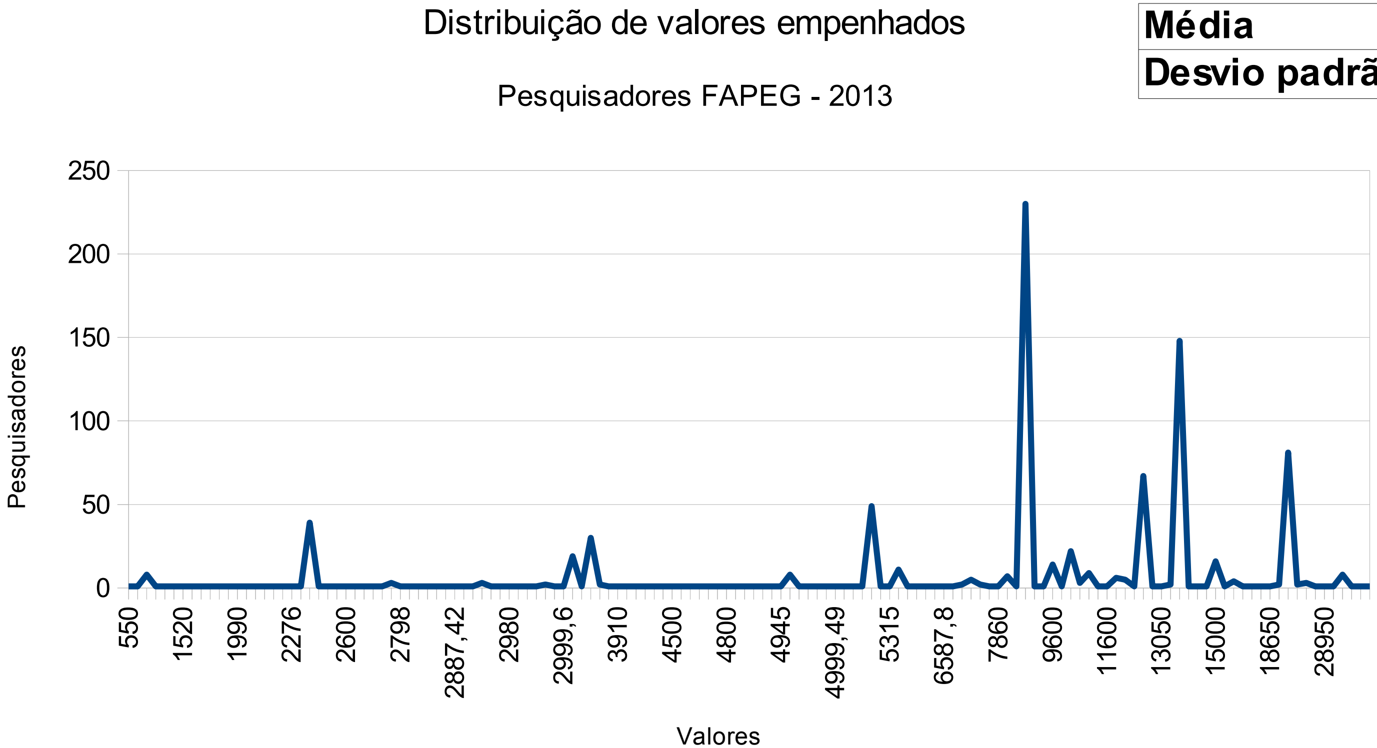
Prof. Dalton Martins
dmartins@gmail.com

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

Exercício da aula passada

- Distribuição dos valores sem CNPJ e sem 0s

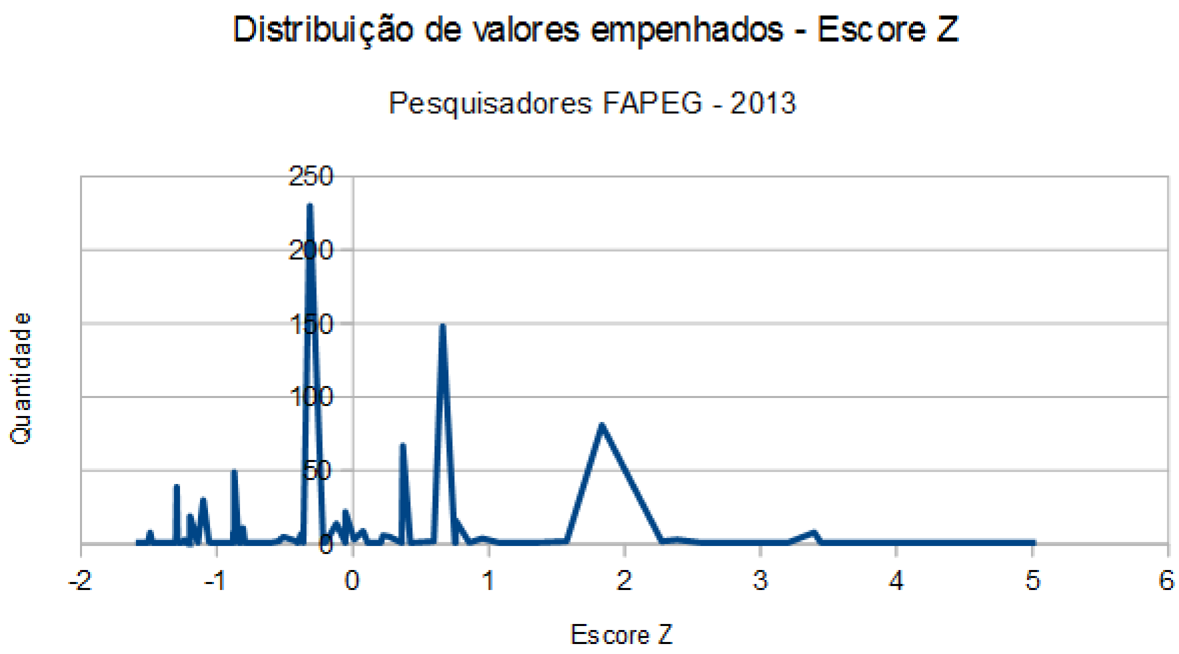
Média	10352,03
Desvio padrão	6138,82



O resultado se parece com uma distribuição normal?

Exercício da aula passada

- Distribuição dos valores sem CNPJ e sem 0s com escore padrão Z e gráfico de dispersão



O resultado se parece com uma distribuição normal?

Exercício da aula passada

- Vejamos algumas propriedades dessa distribuição com escore Z para avaliar se pode ser dita normal ou não:
  - Média é ~0;
  - Desvio padrão é 1;

Desvio padrão	Dados	%
1 desvpad	657	71,57%
2 desvpad	899	97,93%
3 desvpad	905	98,58%

A regra empírica diz que se a distribuição possui a forma normal, então:
68% dos valores até 1 desvio padrão
95% dos valores até 2 desvios padrão
99,7% dos valores até 3 desvios padrão

Valores muito próximos, com média e desvio padrão seguindo o esperado → distribuição é normal

Exercício da aula passada

- Conclusão do exercício:
  - Transformar qualquer conjunto de dados em escore padrão Z nos dá condições melhores de avaliar se a distribuição está próxima do esperado para uma distribuição normal ou não;
  - Visualmente podemos nos confundir apenas observando o formato da distribuição, apesar dele já dar indícios que precisam ser avaliados com mais cuidado;
  - O cálculo da média, desvio padrão e da distribuição pela regra empírica facilitam termos uma avaliação mais precisa e concluir se é normal ou não.

Usando percentil para calcular resultados

- Os percentis são utilizados em uma variedade de formas visando sempre a comparação e determinação da posição relativa;
- Vejamos uma situação prática:
  - Uma estudante de fisioterapia fez uma prova para conseguir seu certificado e tirou 235.
  - Ela sabe que os resultados da prova possuem uma distribuição normal com média 250 e desvio padrão 15, o significa que seu escore padrão é -1, ou seja, um desvio padrão abaixo da média;
  - A pergunta é: sua nota ficou abaixo da média mas ainda é suficiente para passar? 60% dos estudantes que conseguiram as melhoras notas passam e 40% são reprovados;
  - Sabendo disso, é possível saber se a aluna passou? É possível saber qual a nota de corte? Os percentis vão nos ajudar nisso....

Usando percentil para calcular resultados

- A nota de corte para passar ou reprovar está no 40º percentil, conforme a informação que anterior!
  - Lembre-se que o percentil refere-se a porcentagem de resultados abaixo de um determinado ponto;
  - Sabemos que a pontuação da estudante está a 1 desvio padrão abaixo da média;
  - Usando a regra empírica, sabemos que 68% dos resultados se encontram entre 235 e 265 e o resto, 32%, se encontram fora dessa variação;
  - Metade dos que se encontram fora da variação de +- 2 desvios padrão pontuaram menos que 235 → 16%
    - A pontuação da estudante se encontra no 16º → como sua nota está exatamente a 1 desvio padrão pra baixo cai em cima do 16º percentil → ela não passou no teste.

Usando percentil para calcular resultados

- Quando a pontuação caiu exatamente a 1 desvio padrão, ficou fácil descobrir o percentil. E se fosse a 1,3 desvio padrão?
- Para isso, há uma tabela de Escores Padrões e Percentis Correspondentes da Distribuição normal Padrão que ajuda nessa análise.
- Vejamos essa tabela...

Tabela de Escores e Percentis

padrão igual a 0 ... Capítulo 5, para mais sobre ...

Tabela 8-1 Escores Padrões e Percentis Correspondentes da Distribuição Normal Padrão

Escore Padrão	Percentil	Escore Padrão	Percentil	Escore Padrão	Percentil
-3,4	0,03%	-1,1	13,57%	+1,2	88,49%
-3,3	0,05%	-1,0	15,87%	+1,3	90,32%
-3,2	0,07%	-0,9	18,41%	+1,4	91,92%
-3,1	0,10%	-0,8	21,19%	+1,5	93,32%
-3,0	0,13%	-0,7	24,20%	+1,6	94,52%
-2,9	0,19%	-0,6	27,42%	+1,7	95,54%
-2,8	0,26%	-0,5	30,85%	+1,8	96,41%
-2,7	0,35%	-0,4	34, 46%	+1,9	97,13%
-2,6	0,47%	-0,3	38,21%	+2,0	97,73%
-2,5	0,62%	-0,2	42,07%	+2,1	98,21%
-2,4	0,82%	-0,1	46,02%	+2,2	98,61%
-2,3	1,07%	0,0	50,00%	+2,3	98,93%
-2,2	1,39%	+0,1	53,98%	+2,4	99,18%
-2,1	1,79%	+0,2	57,93%	+2,5	99,38%
-2,0	2,27%	+0,3	61,79%	+2,6	99,53%
-1,9	2,87%	+0,4	65,54%	+2,7	99,65%
-1,8	3,59%	+0,5	69,15%	+2,8	99,74%
-1,7	4,46%	+0,6	72,58%	+2,9	99,81%
-1,6	5,48%	+0,7	75,80%	+3,0	99,87%
-1,5	6,68%	+0,8	78,81%	+3,1	99,90%
-1,4	8,08%	+0,9	81,59%	+3,2	99,93%
-1,3	9,68%	+1,0	84,13%	+3,3	99,95%
-1,2	11,51%	+1,1	86,43%	+3,4	99,97%

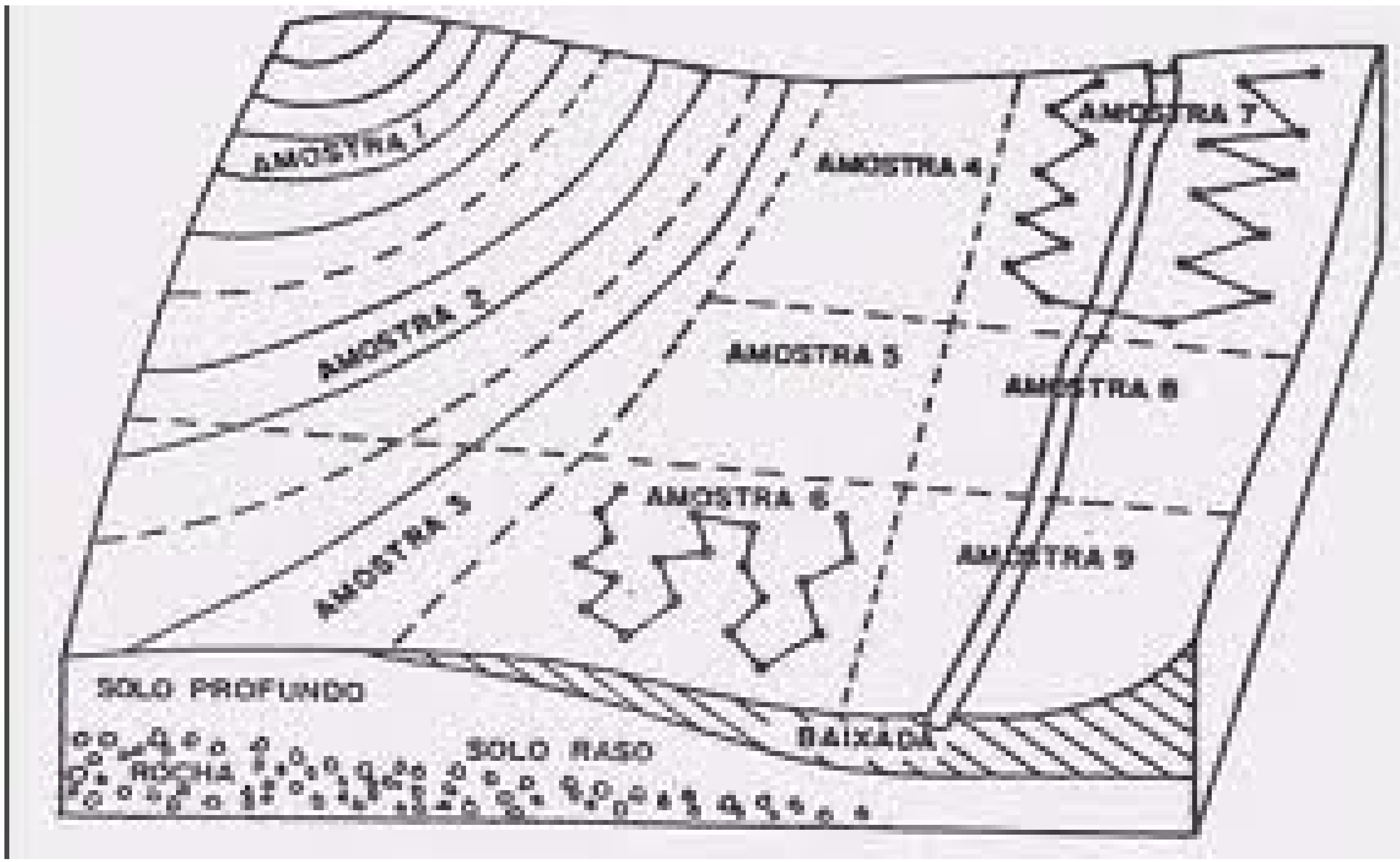
Usando percentil para calcular resultados

- Para calcular um percentil quando os dados possuem uma distribuição normal:
  - Converta o valor original para o seu escore-z;
  - Use a tabela anterior e encontre o percentil correspondente.
- Podemos também fazer o contrário, transformar um percentil de volta ao valor original:
  - Encontre a média e o desvio padrão para a população com a que estamos lidando;
  - Multiplique o escore padrão pelo desvio padrão;
  - Some a esse valor a média.
    - É assim que podemos encontrar a nota de corte de uma prova sabendo em que percentil ela está!

Variabilidade estatística

- No mundo real, a menos que consigamos fazer uma pesquisa que possa abranger todos os elementos da população pesquisada (censo), os resultados obtidos sempre vão variar de amostra para amostra;
- A variabilidade pode ser maior do que gostaríamos ou imaginamos;
- Sempre espere que os resultados amostrais variem de amostra para amostra → não acredite em uma estatística sem alguma indicação de variação explícita.
- Vejamos um exemplo...

Variabilidade estatística



As mudanças do terreno criam características diferentes.

Cada pedaço terá efeitos diferentes sobre um plantio dependendo do que se queira plantar. Se avaliarmos apenas uma amostra, não teremos uma visão do todo.

Medindo a variabilidade

- Graças a alguns resultados estatísticos importantes (especialmente o teorema do limite central) é possível encontrar o quanto se deve esperar que as médias ou proporções amostrais variem sem ter que coletar todas as amostras possíveis;
- O teorema do limite central diz que a distribuição de todas as médias amostrais são normais, desde que o tamanho das amostras seja grande o bastante;
- O que é mais impressionante deste é que para esse resultado não importa a aparência da distribuição da população original.

Erros padrões

- A variabilidade em médias amostrais é medida em termos de erros padrões;
- O erro padrão tem o mesmo conceito básico de um desvio padrão; ambos representam uma distância típica da média;
- Os **valores da população original** desviam-se um dos outros graças a um fenômeno natural (as pessoas têm diferentes alturas, pesos, etc.) → daí temos o nome desvio padrão para medir sua variabilidade;
- As **médias amostrais** variam por causa do erro que ocorre por não sermos capazes de realizar um senso e temos que coletar amostras.

Erros padrões

Média de gastos domésticos anuais de residência norte-americana em 2001.

Despesa	**Média**	**Erro padrão**
Alimentação (comendo em casa)	$3.085,52	$42,30
Alimentação (comendo fora de casa)	$2.235,37	$38,35
Telefone	$914,41	$9,69
Combustível (para veículos)	$1.279,37	$12,88
Materiais de leitura	$141,00	$2,99

Distribuições Amostrais

- Uma lista de todos os valores que uma média amostral pode ter e a frequência com que tais valores ocorrem é chamada **distribuição amostral** da média amostral;
- Uma distribuição amostral, como qualquer outra distribuição, tem um formato, um centro e uma medida de variabilidade → erro padrão;
- Segundo o teorema do limite central, se as amostras forem grande o suficiente, a distribuição de todas as médias amostrais possíveis terá uma distribuição normal, com a mesma média da população original.

Usando a regra empírica para interpretar os erros padrões

- A regra empírica diz que:
  - 68% das médias amostrais encontram-se dentro de 1 erro padrão da média da população;
  - 95% das médias amostrais encontram-se dentro de 2 erros padrão da média da população;
  - 99,7% das médias amostrais encontram-se dentro de 3 erros padrão da média da população.
- Vejamos um exemplo de interpretação baseado na tabela anterior...

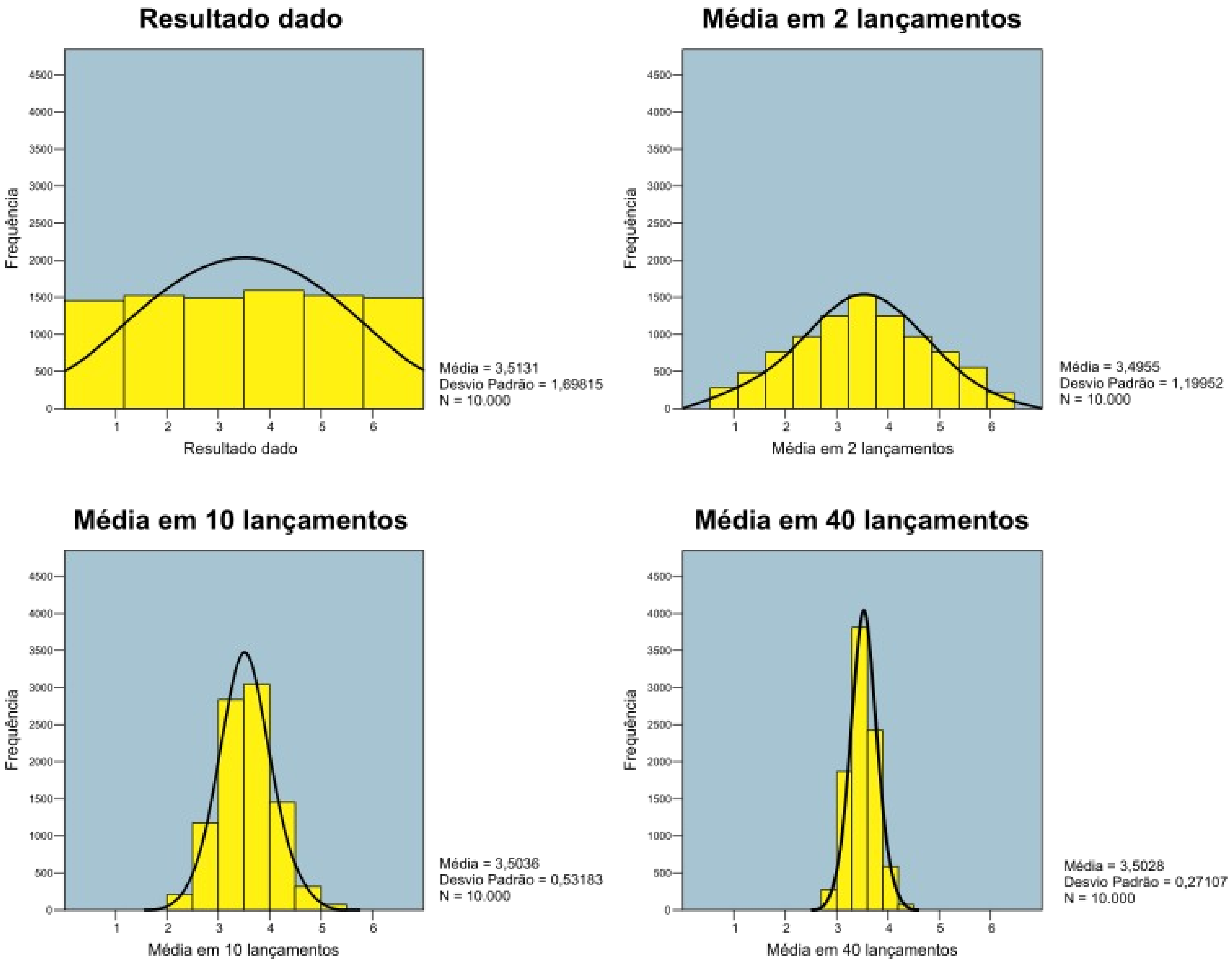
Usando a regra empírica para interpretar os erros padrões

- Utilizando a tabela anterior e a regra empírica, podemos afirmar o seguinte:
  - Em torno de 95% de **TODOS** os lares norte-americanos, terão como média de gasto de telefone para o ano de 2001 um valor entre \$895,03 e \$933,79;
- É a isso que chamamos de intervalo de confiança → o intervalo de variação dos valores;
- O valor somado ou subtraído (\$9,69, no caso) é chamado de margem de erro.

Especificidades do Teorema do limite central

- O teorema do limite central diz que para qualquer população com média ***m*** e desvio padrão ***d***:
  - A distribuição de todas as médias amostrais é, aproximadamente, normal para tamanhos amostrais grandes → isso significa que podemos utilizar a distribuição normal para responder perguntas ou tirar conclusões da média amostral;
  - Quanto maior o tamanho amostral (n), mais próxima de uma distribuição normal será a distribuição das médias amostrais → a maioria dos estatísticos concorda que se n for pelo menos **igual a 30**, fará um bom trabalho na maioria dos casos.
  - O erro padrão é calculado por $d / \sqrt{n}$
    - Conforme n aumenta, o erro diminui.
  - Se os dados normais possuírem uma distribuição normal, a média amostral sempre irá ter uma distribuição normal exata, independente do tamanho amostral;
  - Se o desvio padrão da população for desconhecido, podemos estimá-lo usando o desvio padrão da amostra para calcular o erro.

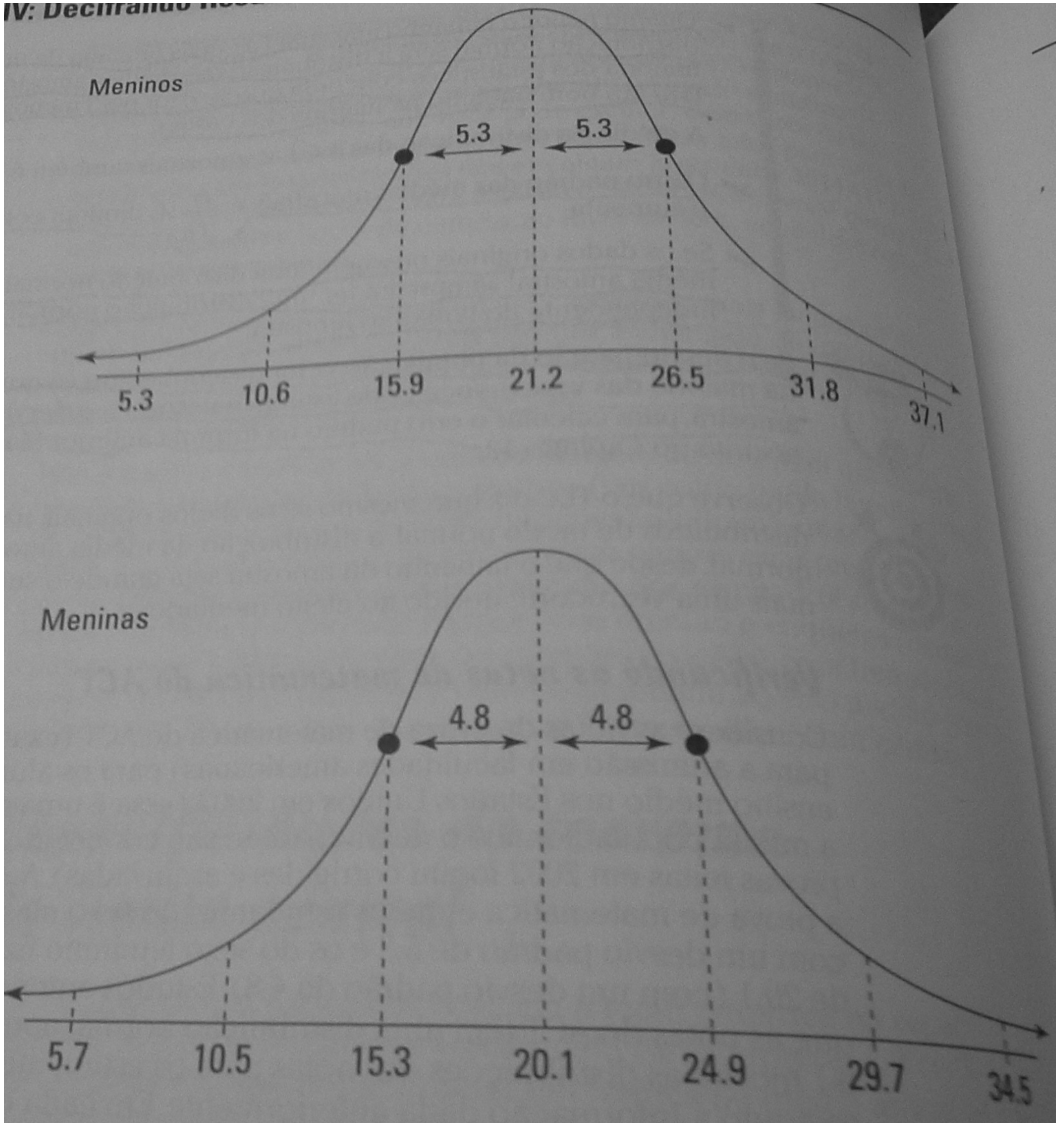
Exemplo


Resultado dado
Frequência
Resultado dado
Média = 3,5131
Desvio Padrão = 1,69815
N = 10.000
Média em 2 lançamentos
Frequência
Média em 2 lançamentos
Média = 3,4955
Desvio Padrão = 1,19952
N = 10.000
Média em 10 lançamentos
Frequência
Média em 10 lançamentos
Média = 3,5036
Desvio Padrão = 0,53183
N = 10.000
Média em 40 lançamentos
Frequência
Média em 40 lançamentos
Média = 3,5028
Desvio Padrão = 0,27107
N = 10.000

Vejamos um exemplo de como isso se aplica

- Considere as notas da prova de matemática do ENEM:
  - A nota média entre os estudantes do sexo masculino foi 21,2 e desvio padrão 5,3;
  - A nota média entre os estudantes do sexo feminino foi 20,1 e desvio padrão 4,8.
  - Usando a regra empírica:
    - 95% dos meninos pontuaram de 10,6 a 31,8
    - 95% das meninas pontuaram de 10,5 a 29,7.

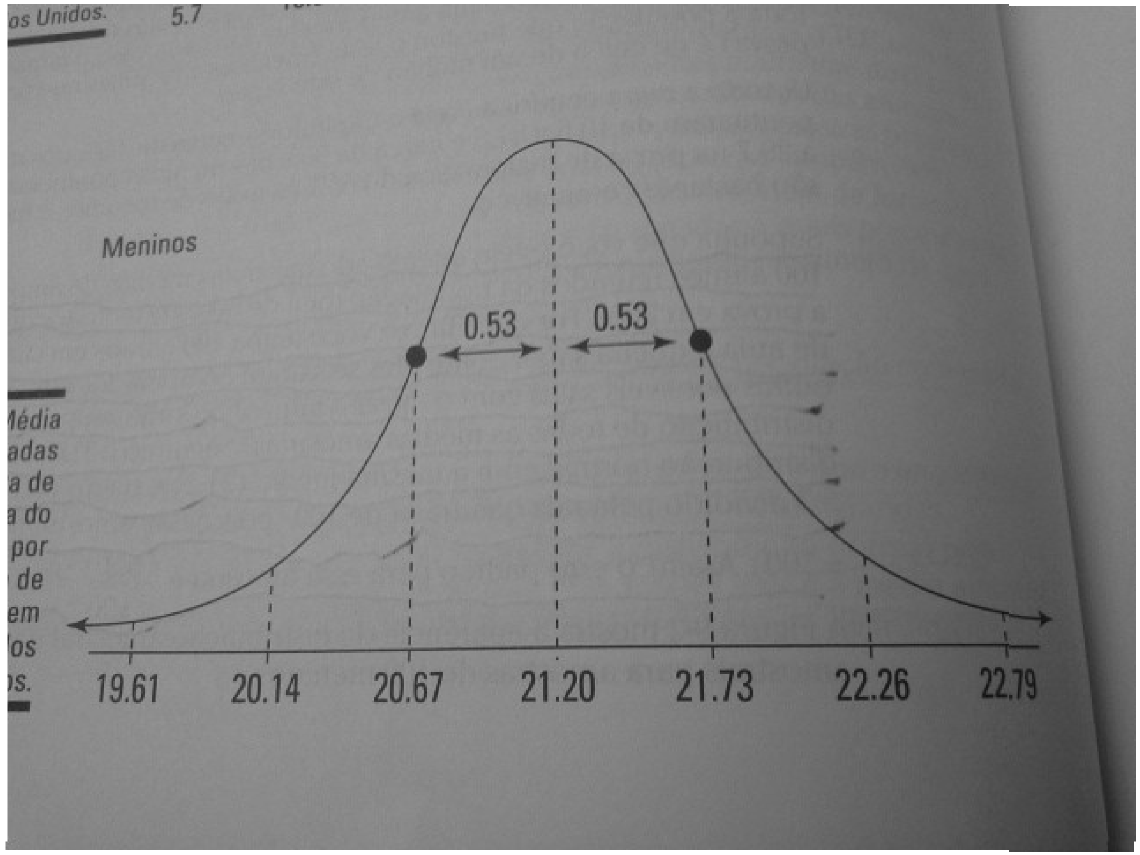
Vejamos um exemplo de como isso se aplica



Vejamos um exemplo de como isso se aplica

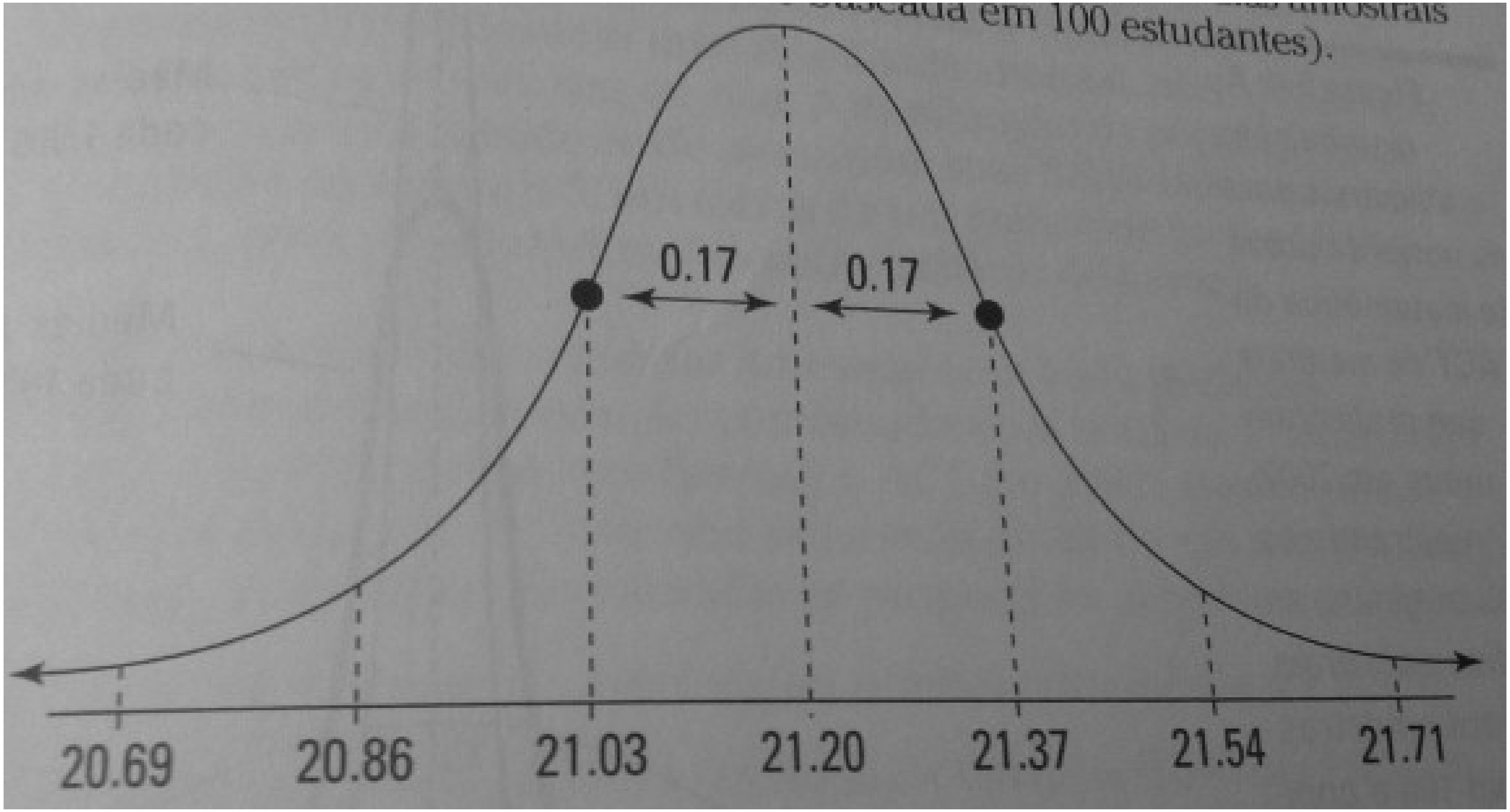
- Se estivemos uma escola qualquer com um grupo de 100 meninos que fizeram essa prova de um total de 500.000 (no país todo), como saber o intervalo de confiança para nosso grupo?
  - Sabemos que a média para meninos é 21,2 e o desvio padrão 5,3;
  - Sabemos que nossa amostra é de tamanho 100;
  - Calculamos o erro padrão: $5{,}3/\sqrt{100}=0{,}53$
  - Logo, a média dos alunos dessa sala é de 21,2 +-0,53

Vejamos um exemplo de como isso se aplica



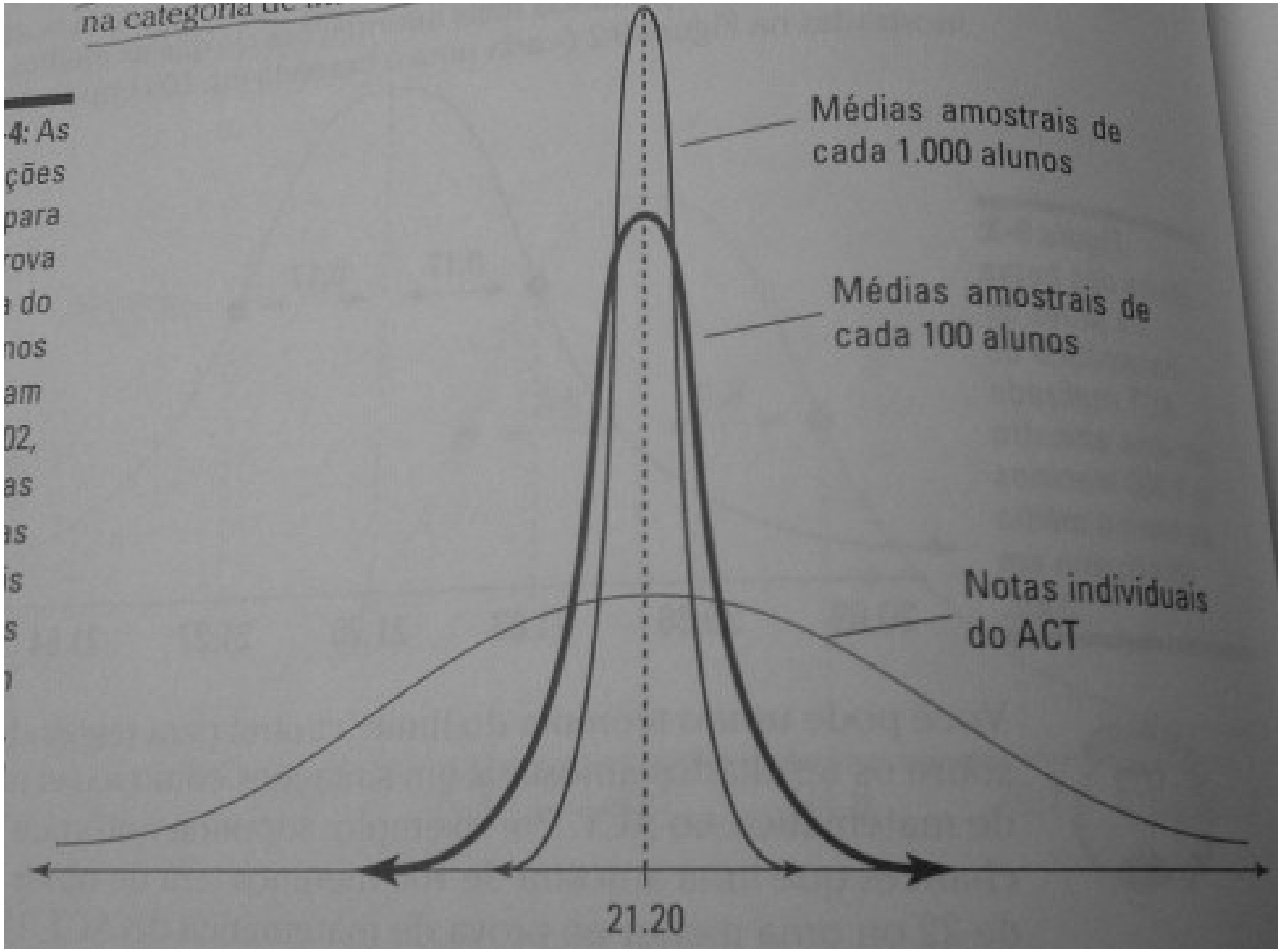
O erro padrão da média amostral é BEM menor do que o desvio padrão das notas originais, vistas no gráfico anterior. Isso ocorre pois temos uma média amostral de 100 alunos comparadas a cada nota individual para um único aluno no primeiro gráfico. Se tivemos uma amostra de 1000 alunos, o gráfico seria...

Vejamos um exemplo de como isso se aplica



O erro padrão ficou ainda menor....

Vejamos um exemplo de como isso se aplica



Quanto maior a amostra, mais o resultado se ajusta em torno da média.

Respondendo perguntas com o Teorema do limite Central

- Suponha que você queira saber as chances que uma mostra de 100 meninos terá de obter a nota média 22 ou menor na prova anterior;
- Podemos usar a técnica que vimos hoje do percentil para isso:
  - Transformamos a nota 22 em percentil com a fórmula **[(valor X – média)/erro padrão]**
  - Façamos como exercício!!! - Que porcentagem teremos como retorno?

Exercícios

- 1. A pontuação média obtida por 52 alunos em uma determinada avaliação foi de 70 pontos, com um desvio padrão igual a 5 pontos. Calcule a probabilidade dos alunos tirarem:
  - a) mais de 80 pontos;
  - b) menos de 80 pontos.
- 2. Em uma outra prova, os mesmos alunos citados acima obtiveram a mesma pontuação média, porém com um desvio padrão de 20 pontos. Calcule a probabilidade dos alunos tirarem:
  - a) mais de 80 pontos;
  - b) menos de 80 pontos.

Respostas

- 1.
  - a) mais de 80 pontos → 2,28 %
  - b) menos de 80 pontos → 97,72 %
- 2.
  - a) mais de 80 pontos → 30,85 %
  - b) menos de 80 pontos → 69,15 %