CPEQS

AULA 02
Distribuição de probabilidade normal

Ernesto F. L. Amaral

02 de outubro de 2013

Centro de Pesquisas Quantitativas em Ciências Sociais (CPEQS)
Faculdade de Filosofia e Ciências Humanas (FAFICH)
Universidade Federal de Minas Gerais (UFMG)

Fonte:
Triola, Mario F. 2008. "Introdução à estatística". 10 ª ed. Rio de Janeiro: LTC. Capítulo 6 (pp.192-249).

ESQUEMA DA AULA

– A distribuição normal padrão

– Aplicações da distribuição normal

– O Teorema Central do Limite

– Determinação de normalidade

– Utilização de pesos amostrais

A DISTRIBUIÇÃO NORMAL PADRÃO

VARIÁVEL ALEATÓRIA

– **Variável aleatória** é uma variável que tem um único valor numérico, determinado pelo acaso, para cada resultado de um experimento.

– **Distribuição de probabilidade** descreve a probabilidade de cada valor da variável aleatória.

– **Variável aleatória discreta** tem uma quantidade finita de valores ou uma quantidade enumerável de valores.

– **Variável aleatória contínua** tem infinitos valores, sem saltos ou interrupções.

GRÁFICOS DAS DISTRIBUIÇÕES

– O **histograma de probabilidade** é um gráfico de uma distribuição de probabilidade discreta.

– A **curva de densidade** é um gráfico de uma distribuição de probabilidade contínua, em que:

  – A área total sob a curva tem que ser igual a 1.

  – Cada ponto na curva tem que ter uma altura vertical que é 0 ou maior, não estando abaixo do eixo *x*.

DISTRIBUIÇÕES DE PROBABILIDADE

– Como a **área total** sob o gráfico de uma distribuição de probabilidade é igual a 1, há correspondência entre área e probabilidade (ou frequência relativa).

– Isto possibilita **calcular probabilidades** com utilização das áreas.

– **É importante:**
  – Desenvolver a habilidade para determinar áreas correspondentes a várias regiões sob o gráfico da distribuição.
  – Encontrar valores da variável *z* que correspondem a áreas sob o gráfico.

DISTRIBUIÇÕES UNIFORMES

– Na **distribuição uniforme**, uma variável aleatória contínua apresenta valores de probabilidade que se espalham uniformemente sobre a faixa de valores possíveis.

– Em geral, a área de um retângulo se torna 1 quando fazemos sua altura igual ao valor de 1/amplitude.

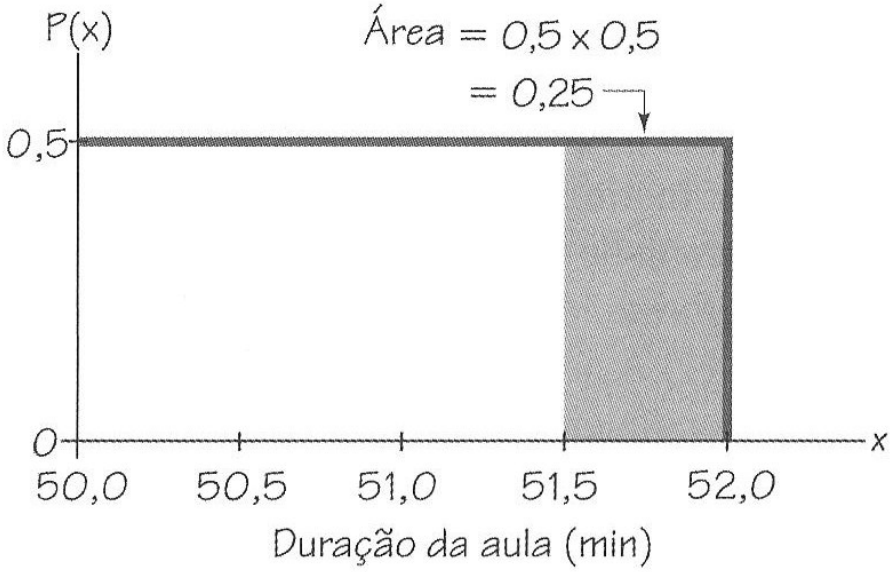


FIGURA 6-3 Usando a Área para Achar a Probabilidade

DISTRIBUIÇÃO NORMAL

– As **distribuições normais** são importantes, porque elas ocorrem frequentemente em situações reais e desempenham papel importante nos métodos de inferência estatística.

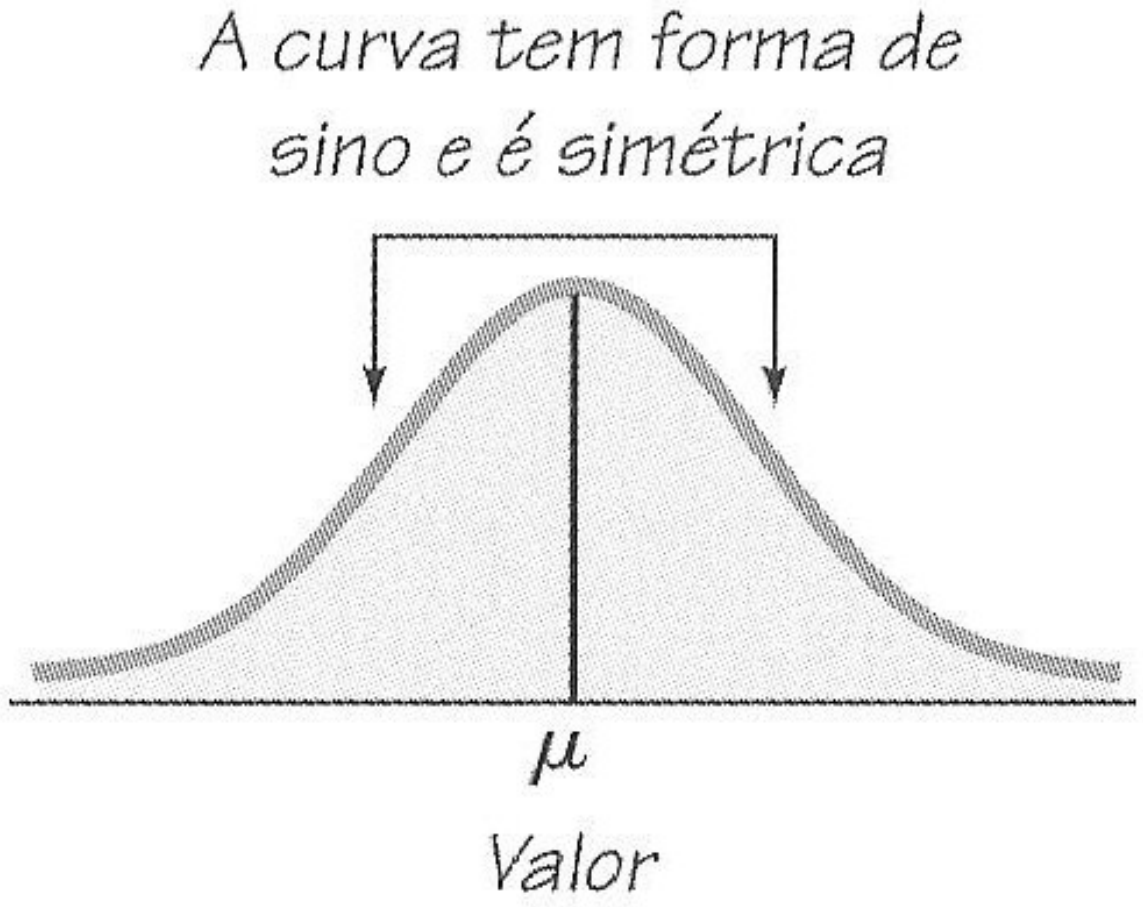
– A distribuição é normal se uma variável aleatória contínua tem uma distribuição com um gráfico simétrico em forma de sino.

– Qualquer distribuição normal é determinada pela média (μ) e desvio padrão (σ):

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

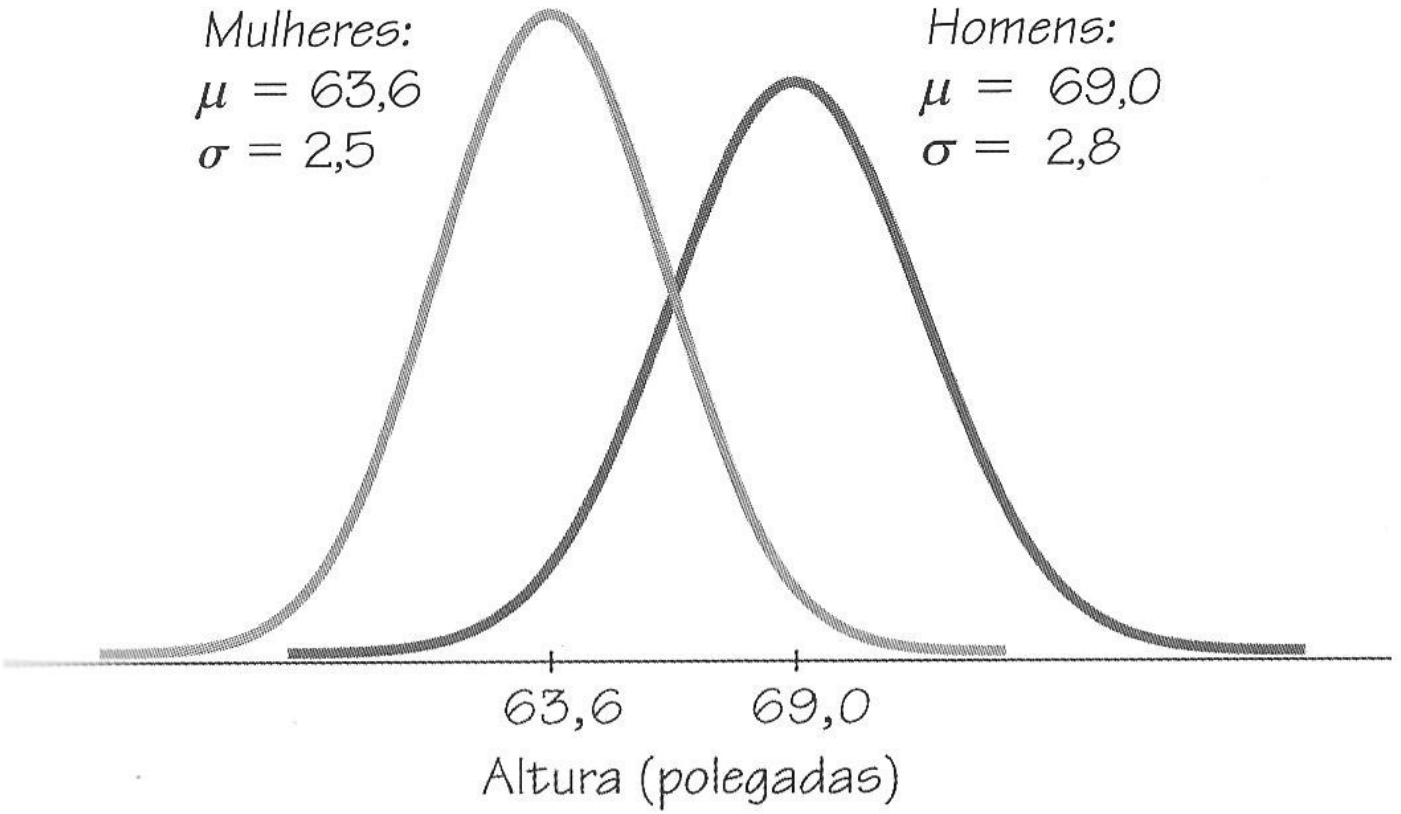
GRÁFICO DA DISTRIBUIÇÃO NORMAL

– De posse de valores específicos para μ e σ, podemos fazer o seguinte gráfico da distribuição normal.



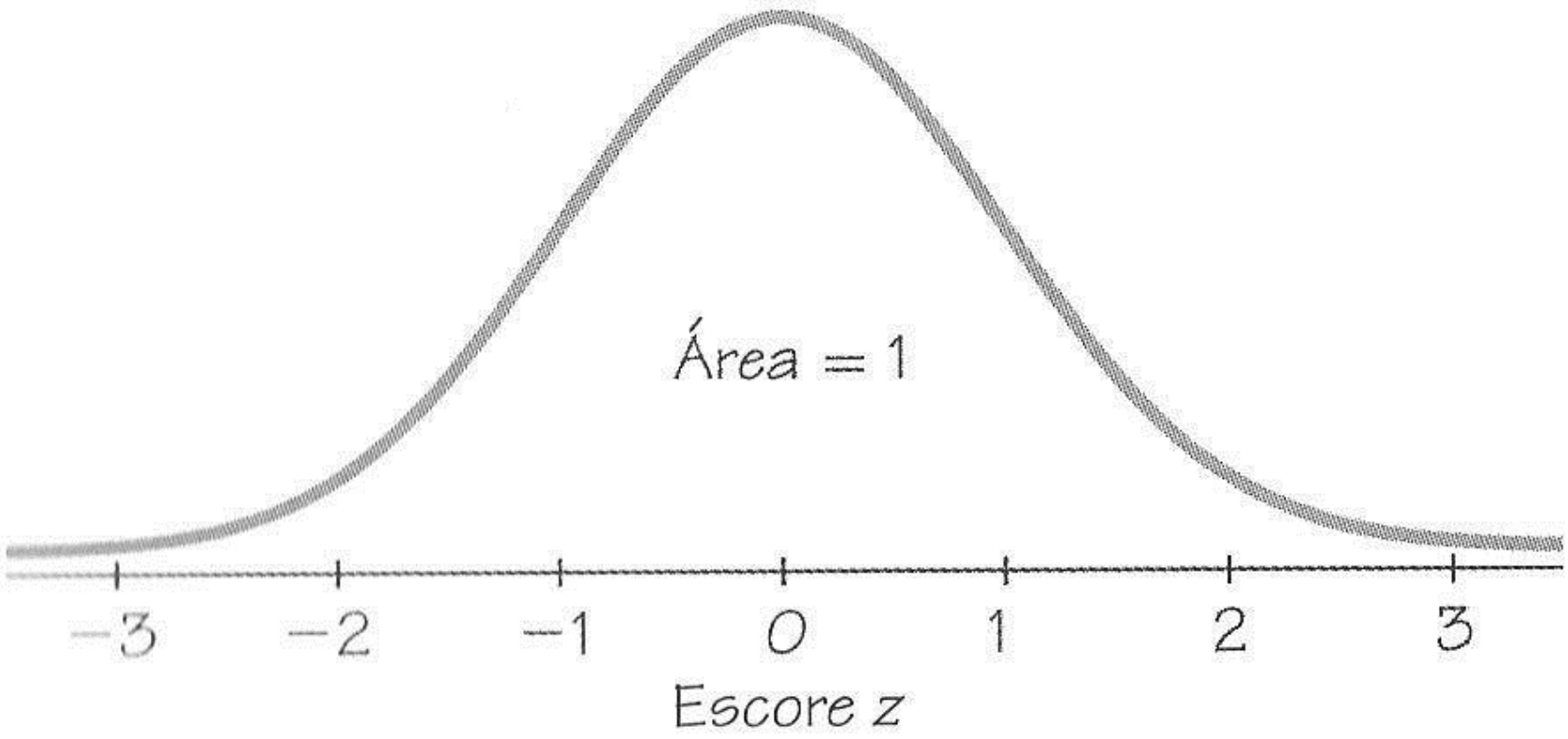
VARIAÇÃO NAS DISTRIBUIÇÕES NORMAIS

– Há muitas distribuições normais diferentes, dependendo de dois parâmetros: a média populacional (μ) e o desvio padrão populacional (σ).



DISTRIBUIÇÃO NORMAL PADRÃO

– A distribuição normal padrão é uma distribuição de probabilidade normal com média (μ) igual a 0 e desvio padrão (σ) igual a 1.



ENCONTRE PROBABILIDADES A PARTIR DE ESCORES *z*

– Usando a tabela das páginas 618-619, é possível achar áreas (ou probabilidades) para muitas regiões diferentes.
  – Se refere à distribuição normal padrão (μ=0 e σ=1).
  – Possui resultados para escores *z* negativos e positivos.

– **Escore *z*:** distância na escala horizontal da distribuição normal padrão:
  – Parte inteira e decimal: coluna à esquerda da tabela.
  – Parte do centésimo: linha no topo da tabela.

– **Área:** região sob a curva (valores no corpo da tabela).

Escores *z* POSITIVOS

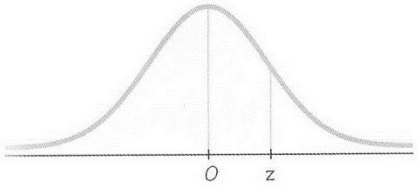


TABELA A.2 Área Acumulada à ESQUERDA *(continuação)*

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	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	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	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	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

– **Área acumulada à esquerda** de z=1,13 é igual a 0,8708.

– Há uma probabilidade de 0,8708 de **selecionarmos aleatoriamente** um escore *z* menor que 1,13.

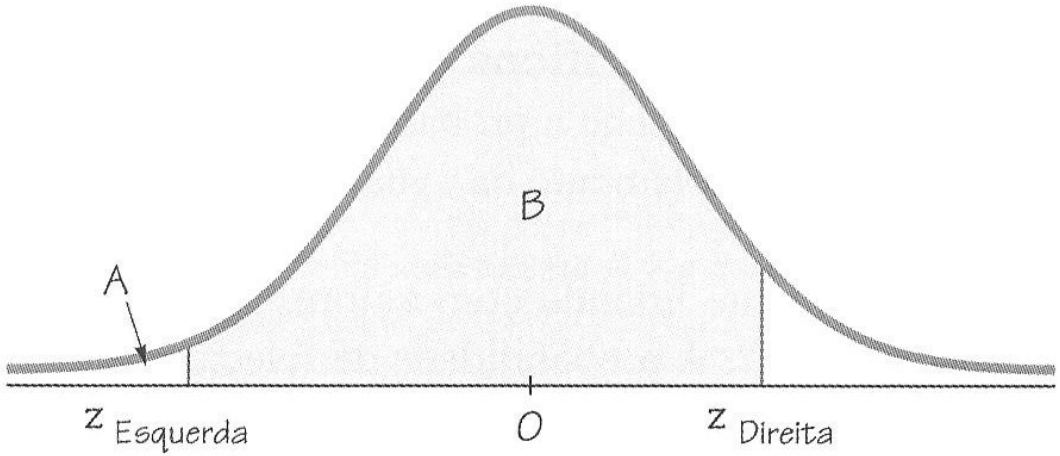
ENCONTRANDO PROBABILIDADES

– Para encontrar o valor da probabilidade, primeiro desenhe um gráfico, sombreie a região desejada e pense em uma maneira de achar a área correspondente.

– *P(a<z<b)*: probabilidade do escore *z* estar entre *a* e *b*.

– *P(z>a)*: probabilidade do escore *z* ser maior que *a*.

– *P(z<a)*: probabilidade do escore *z* ser menor que *a*.



Área sombreada B = (áreas A e B combinadas) − (área A)
= (área da Tabela A-2 usando $z_{Direita}$) − área da Tabela A-2 usando $z_{Esquerda}$)

PROBABILIDADE DE VALOR EXATO É IGUAL A ZERO

– Com uma distribuição de probabilidade contínua, a probabilidade de se obter qualquer valor único exato é zero:

$$P(z = a) = 0$$

– Por exemplo, há uma probabilidade 0 de selecionarmos aleatoriamente uma pessoa com altura exatamente igual a 1,763947 metros.

– Um ponto isolado na escala horizontal é representado por uma linha vertical, e não uma área sob a curva:

$$P(a \leq z \leq b) = P(a < z < b).$$

– A probabilidade de se obter um valor **no máximo igual *b*** é igual à probabilidade de se obter um valor **menor que *b***.

– É importante saber interpretar frases-chave: no máximo, pelo menos, mais do que, não mais do que...

APLICAÇÕES DA DISTRIBUIÇÃO NORMAL

APLICAÇÕES DA DISTRIBUIÇÃO NORMAL

– Vamos tratar de métodos para trabalhar com distribuições normais que não são padrões (ou $\mu \neq 0$, ou $\sigma \neq 1$, ou ambos).

– Podemos fazer conversão para transformar qualquer distribuição normal em distribuição normal padrão.

– Se convertermos valores para escores padronizados, os procedimentos para trabalhar com distribuições normais serão os mesmos daqueles usados para distribuição normal padrão:

$$z = (x - \mu) / \sigma$$

EQUIVALÊNCIA ENTRE NORMAL E NORMAL PADRÃO

– A área em qualquer distribuição normal limitada por um escore *x* é igual à área limitada pelo escore *z* equivalente na distribuição normal padrão.

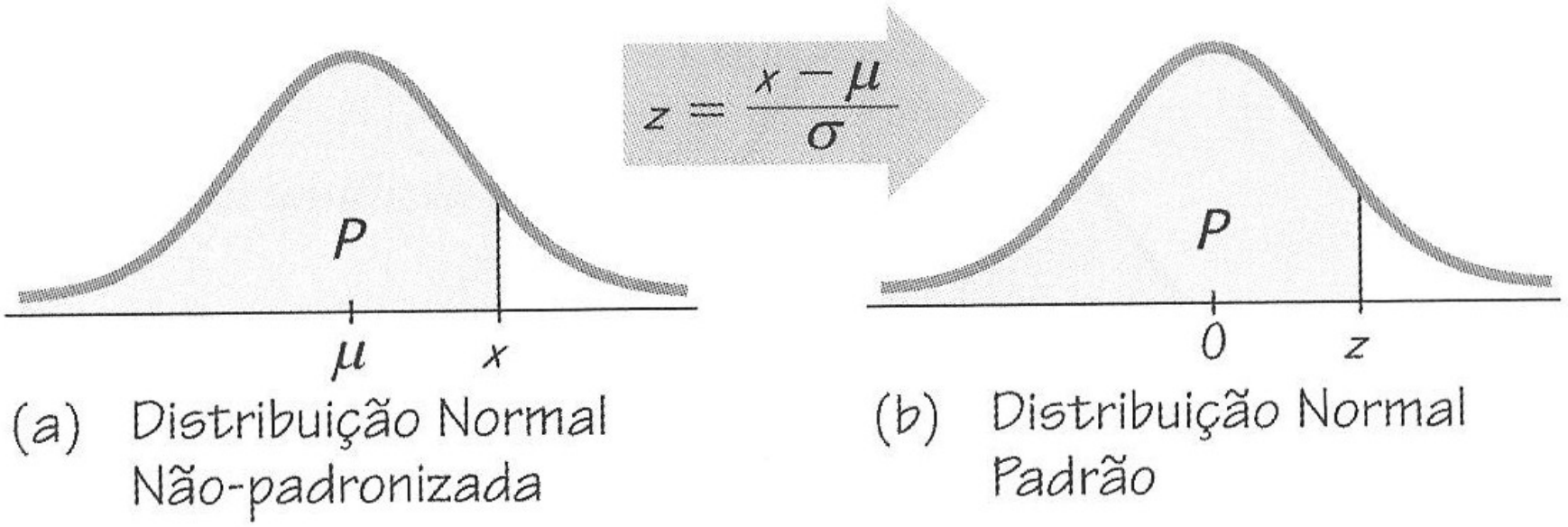


FIGURA 6-12 Conversão de uma Distribuição Normal Não-padronizada para a Distribuição Normal Padrão

O TEOREMA CENTRAL DO LIMITE

DISTRIBUIÇÕES AMOSTRAIS E ESTIMADORES

– **Distribuição amostral de uma estatística** (média amostral) é a distribuição de todos valores da estatística, quando todas amostras possíveis de mesmo tamanho *n* tiverem sido extraídas da mesma população.

– A distribuição amostral de uma estatística é geralmente representada por uma tabela, histograma de probabilidade ou fórmula.

– Estatísticas que atingem parâmetro (**estimadores não-viesados**): proporção, média, variância.

– Estatísticas que não atingem parâmetro (**estimadores viesados**): mediana, amplitude, desvio padrão.

ALGUNS PRINCÍPIOS

– Ao selecionar uma **amostra aleatória** de uma população com média (μ) e desvio padrão (σ):

  – **Se $n>30$**, então as médias amostrais têm uma distribuição que pode ser aproximada por uma distribuição normal com média (μ) e desvio padrão ($\sigma/\sqrt{n}$), independente da distribuição da população original.

  – **Se $n\leq30$** e a população original tem uma **distribuição normal**, então as médias amostrais têm uma distribuição normal com média (μ) e desvio padrão ($\sigma/\sqrt{n}$).

  – **Se $n\leq30$**, mas a população original **não tem uma distribuição normal**, então os métodos a seguir não se aplicam.

TEOREMA CENTRAL DO LIMITE (TCL)

– O **teorema central do limite** diz que...

  – se tamanho amostral é grande o suficiente...

  – a distribuição das médias amostrais pode ser aproximada por uma distribuição normal...

  – mesmo que a população original não seja normalmente distribuída.

PRESSUPOSTOS DO TCL

– A **variável aleatória *x*** tem uma distribuição (que pode ou não ser normal) com média μ e desvio padrão σ.

– **Amostras aleatórias simples (AAS)**, com mesmo tamanho amostral *n*, são selecionadas da população.

  – AAS são amostras selecionadas de uma população de modo que todas possíveis amostras de tamanho *n* têm a mesma chance de ser escolhidas.

CONCLUSÕES DO TCL

– **Distribuição** das médias amostrais irá se aproximar de uma distribuição normal à medida que *n* aumentar.

– A **média** de todas médias amostrais é a média μ da população.

– O **desvio padrão** de todas médias amostrais é $\sigma/\sqrt{n}$.

DETERMINAÇÃO DE NORMALIDADE

DETERMINAÇÃO DE NORMALIDADE

– Alguns métodos estatísticos exigem que os dados amostrais tenham sido selecionados aleatoriamente de uma população que tenha distribuição normal.

– Podemos analisar histogramas, valores extremos (*outliers*) e gráficos de quantis normais para determinar se as exigências para uma distribuição normal são satisfeitas.

GRÁFICOS QUANTIL-NORMAL

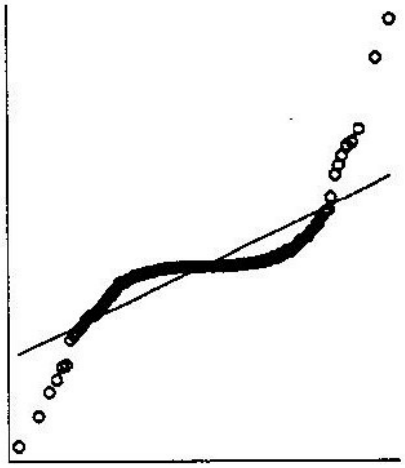
– Um gráfico dos quantis normais (ou gráfico de probabilidades normais) é um gráfico de pontos (x, y) em que um eixo possui o conjunto original de dados amostrais e o outro eixo apresenta o escore *z*, correspondente ao valor esperado do quantil da distribuição normal padrão.

– Se os pontos não se aproximam de uma reta ou se os pontos exibem um padrão simétrico que não seja um padrão linear, então os dados parecem provir de uma população que não tem distribuição normal.

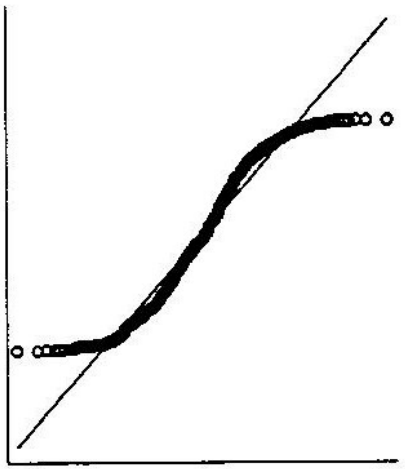
– Se o padrão dos pontos é razoavelmente próximo de uma reta, então os dados parecem provir de uma população com distribuição normal.

– Se a variável seguisse uma distribuição normal, os pontos se encontrariam exatamente sobre a linha diagonal.

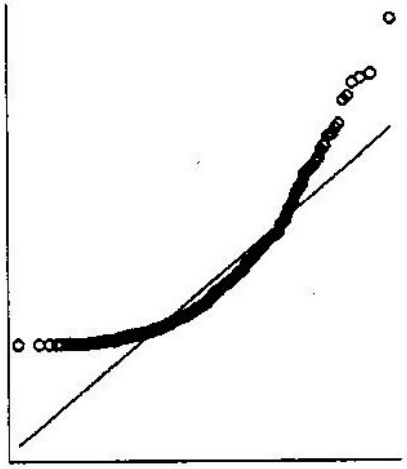
EXEMPLOS DE GRÁFICOS QUANTIL-NORMAL



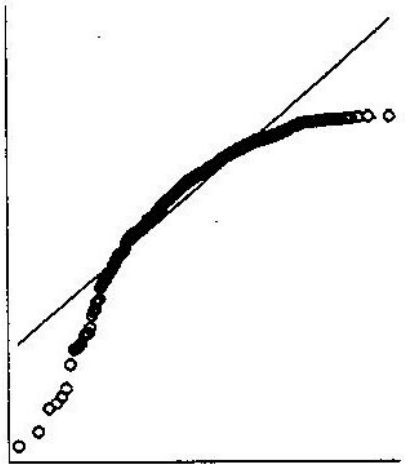
Heavy Tails, High and Low Outliers



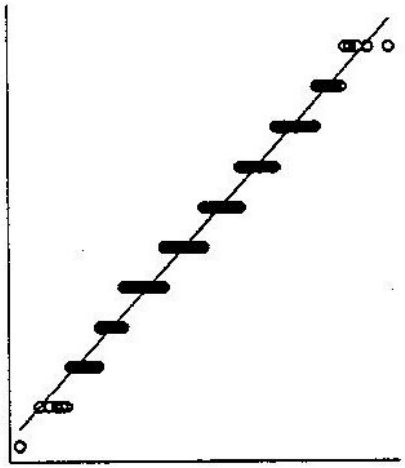
Light Tails, No Outliers



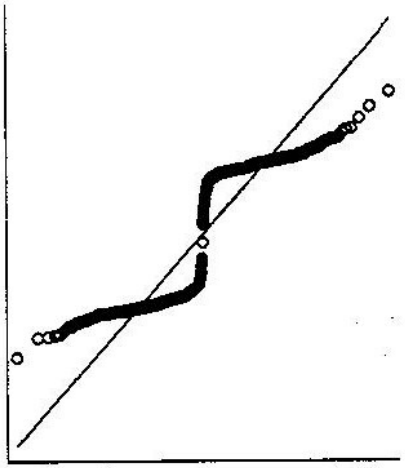
Positive Skew, High Outliers



Negative Skew, Low Outliers



Granularity
(discrete values)



Two Peaks, Central Gap
(bimodal)

Figure 1.10 Quantile-normal plots reflect distribution shape.

Hamilton (1992: 16).

TRANSFORMAÇÃO DE DADOS

– Lawrence Hamilton (“Regression with graphics”) pág.18-19:

Y^3 >>> q=3

Y^2 >>> q=2

Y^1 >>> q=1

$Y^{0,5}$ >>> q=0,5

log(Y) >>> q=0

$-(Y^{-0,5})$ >>> q=–0,5

$-(Y^{-1})$ >>> q=–1

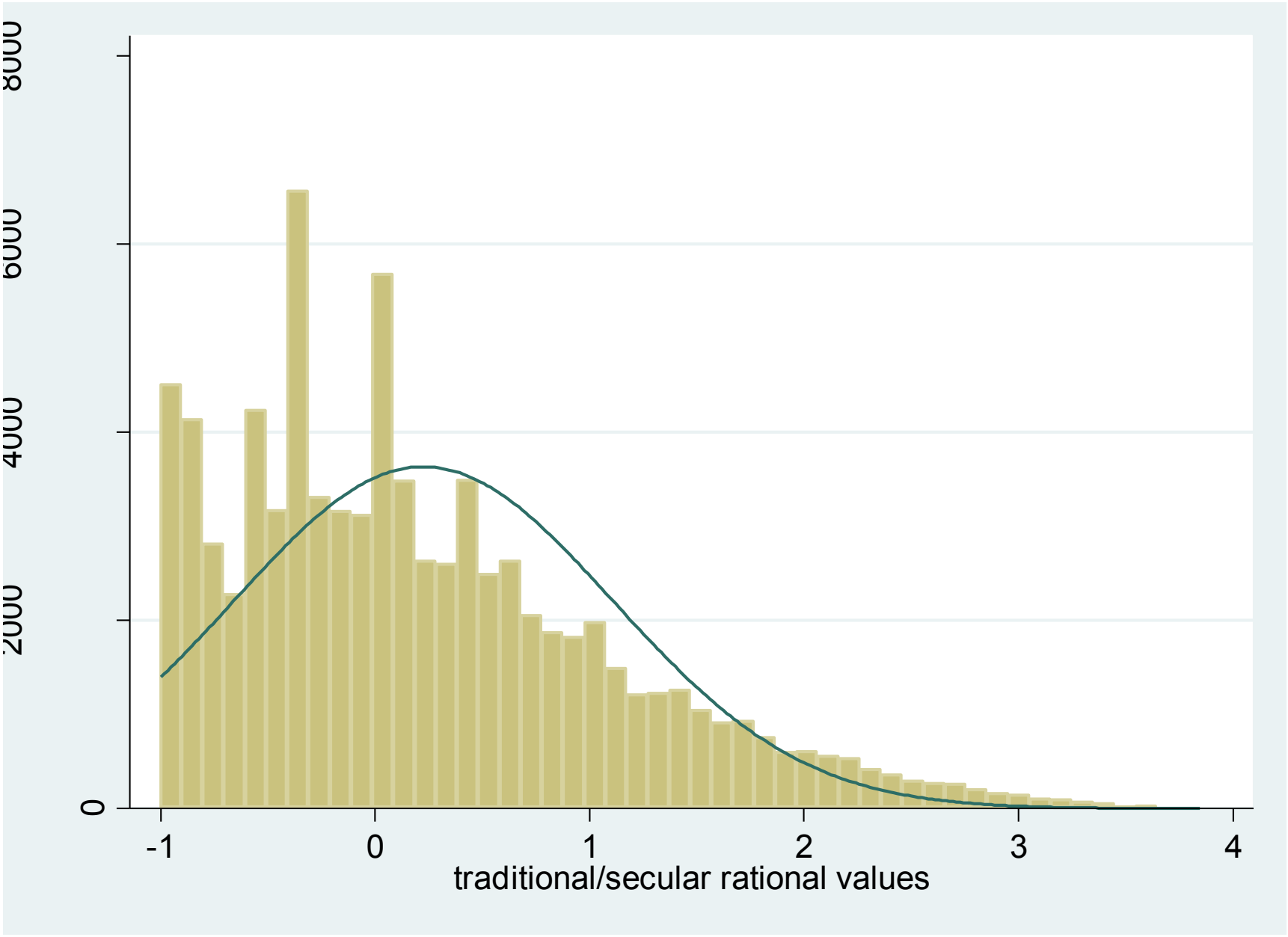
– q>1: reduz concentração à direita.

– q=1: dados originais.

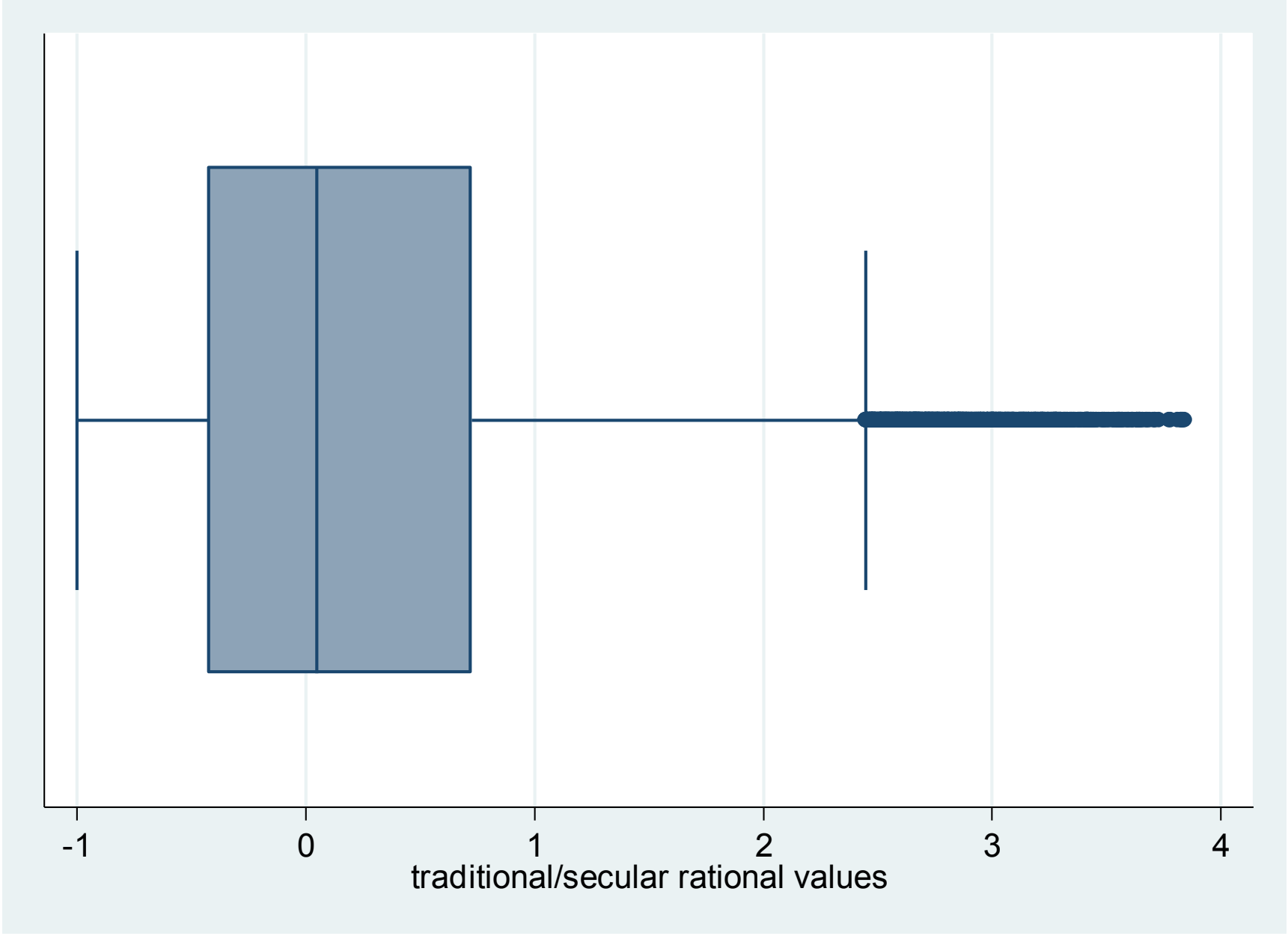
– q<1: reduz concentração à esquerda.

– *log*(*x*+1) viabiliza transformação quando *x*=0. Se distribuição de *log*(*x*+1) for normal, é chamada de distribuição lognormal.

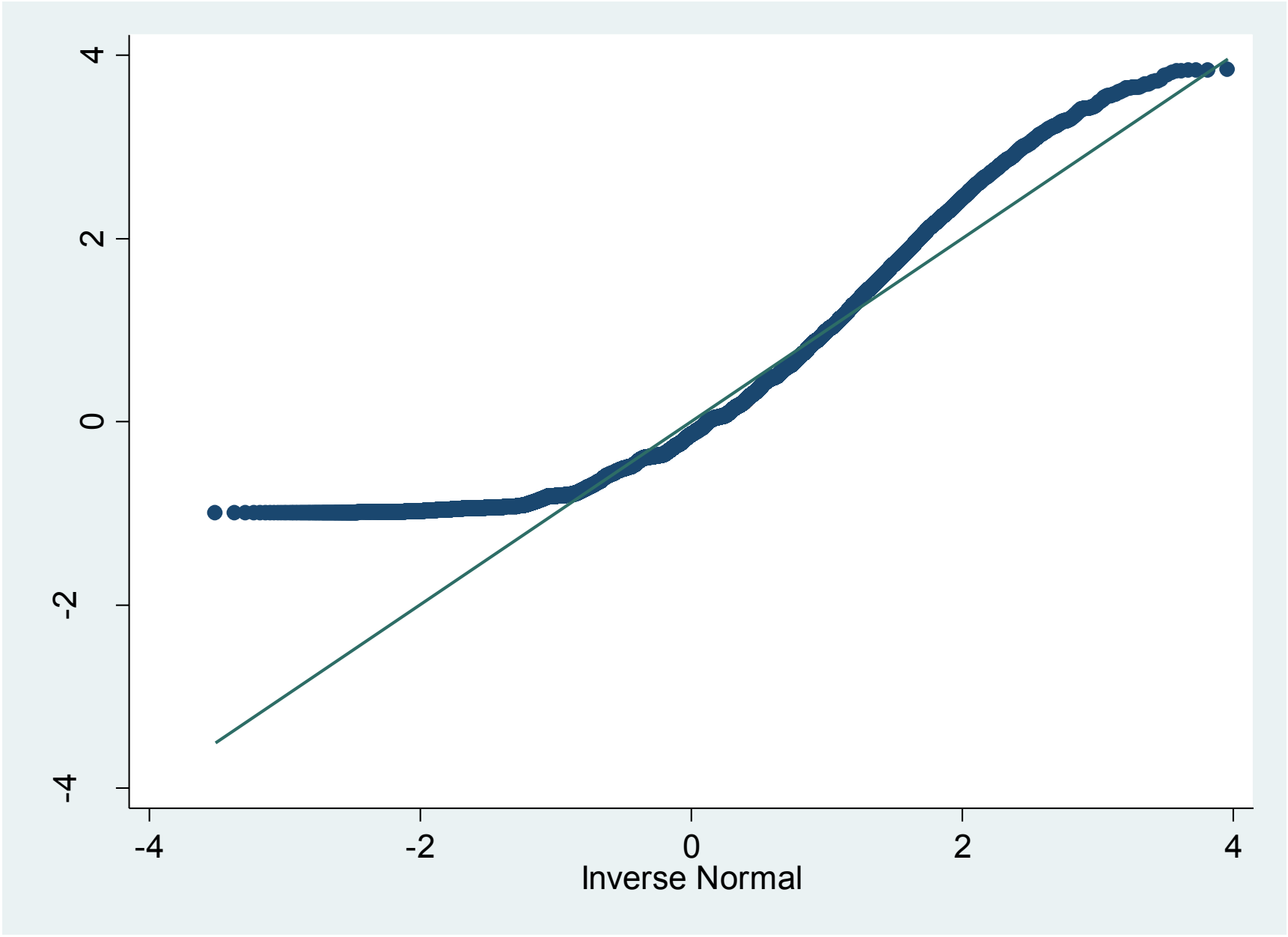
ÍNDICE VALORES RACIONAIS (TRADICIONAL/SECULAR)



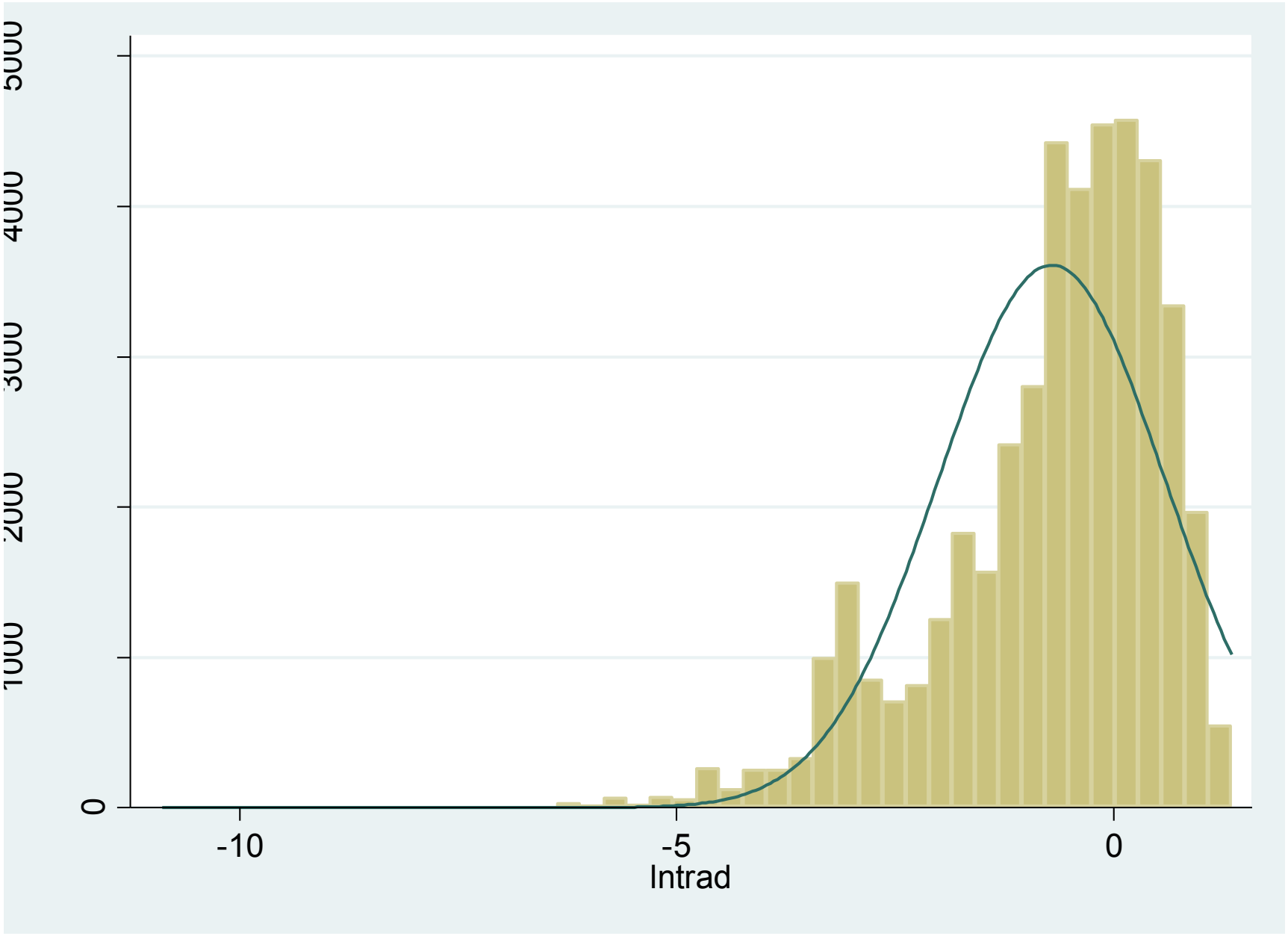
ÍNDICE VALORES RACIONAIS (TRADICIONAL/SECULAR)



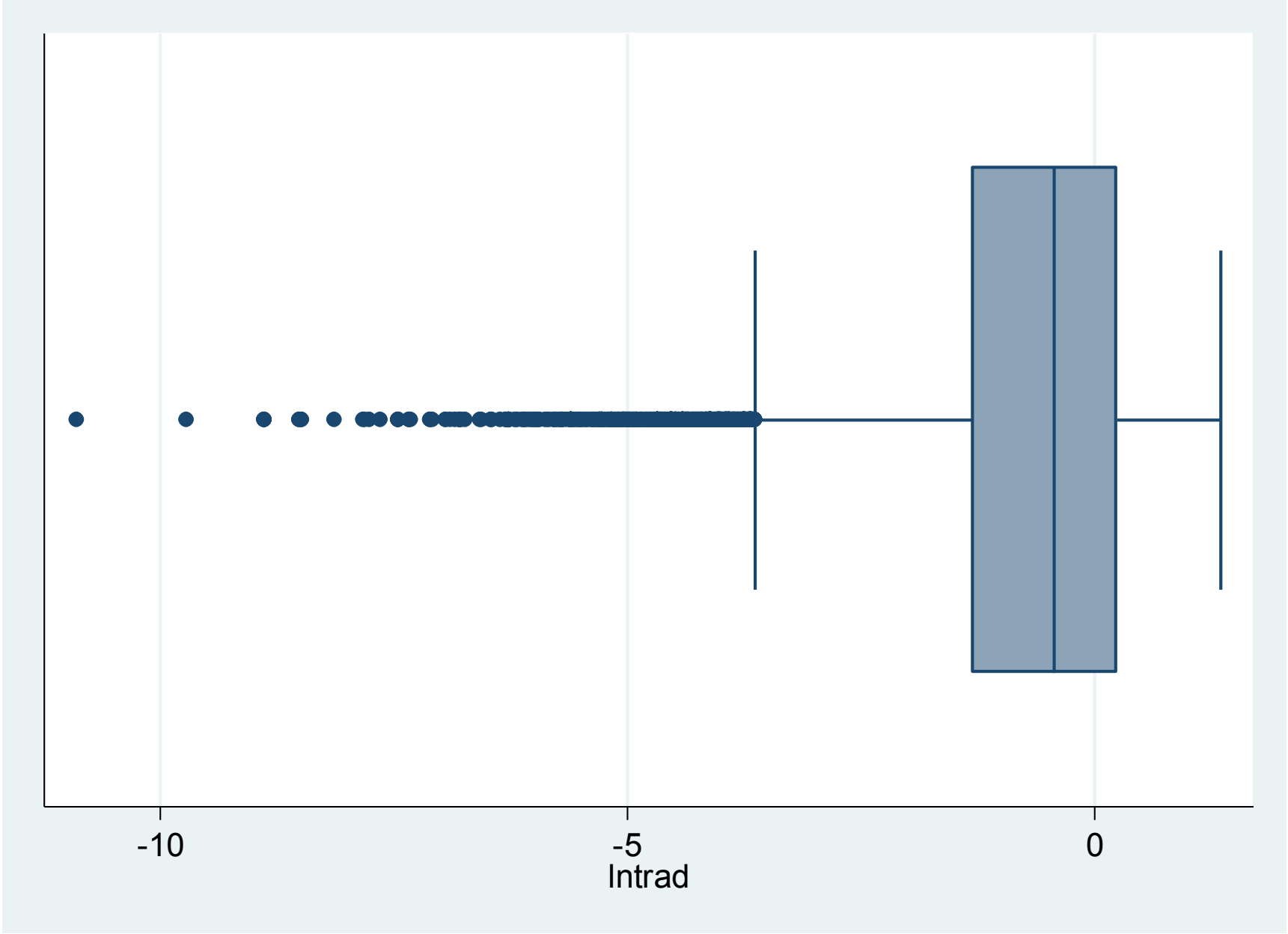
ÍNDICE VALORES RACIONAIS (TRADICIONAL/SECULAR)



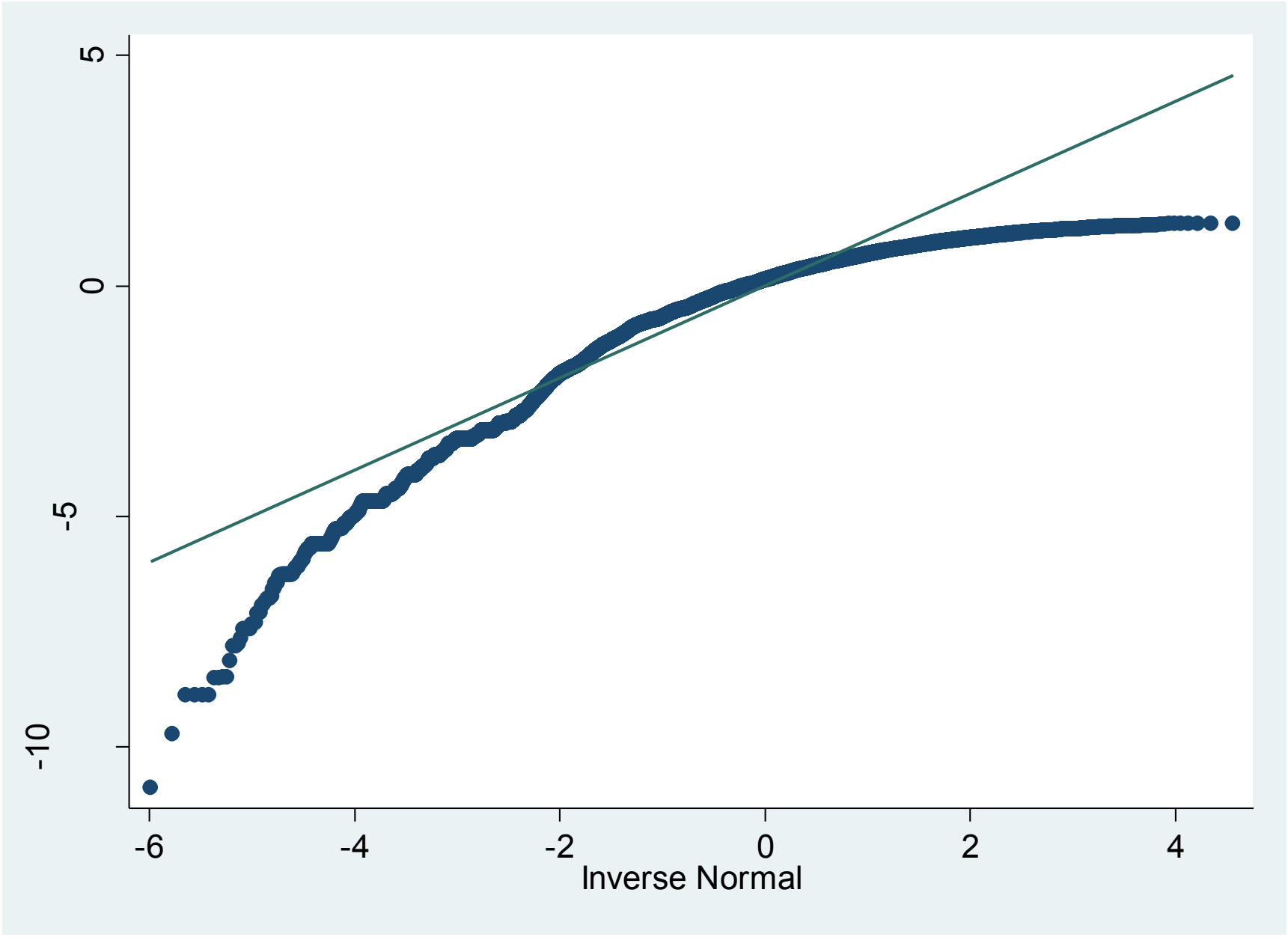
LOGARITMO DO ÍNDICE VALORES RACIONAIS



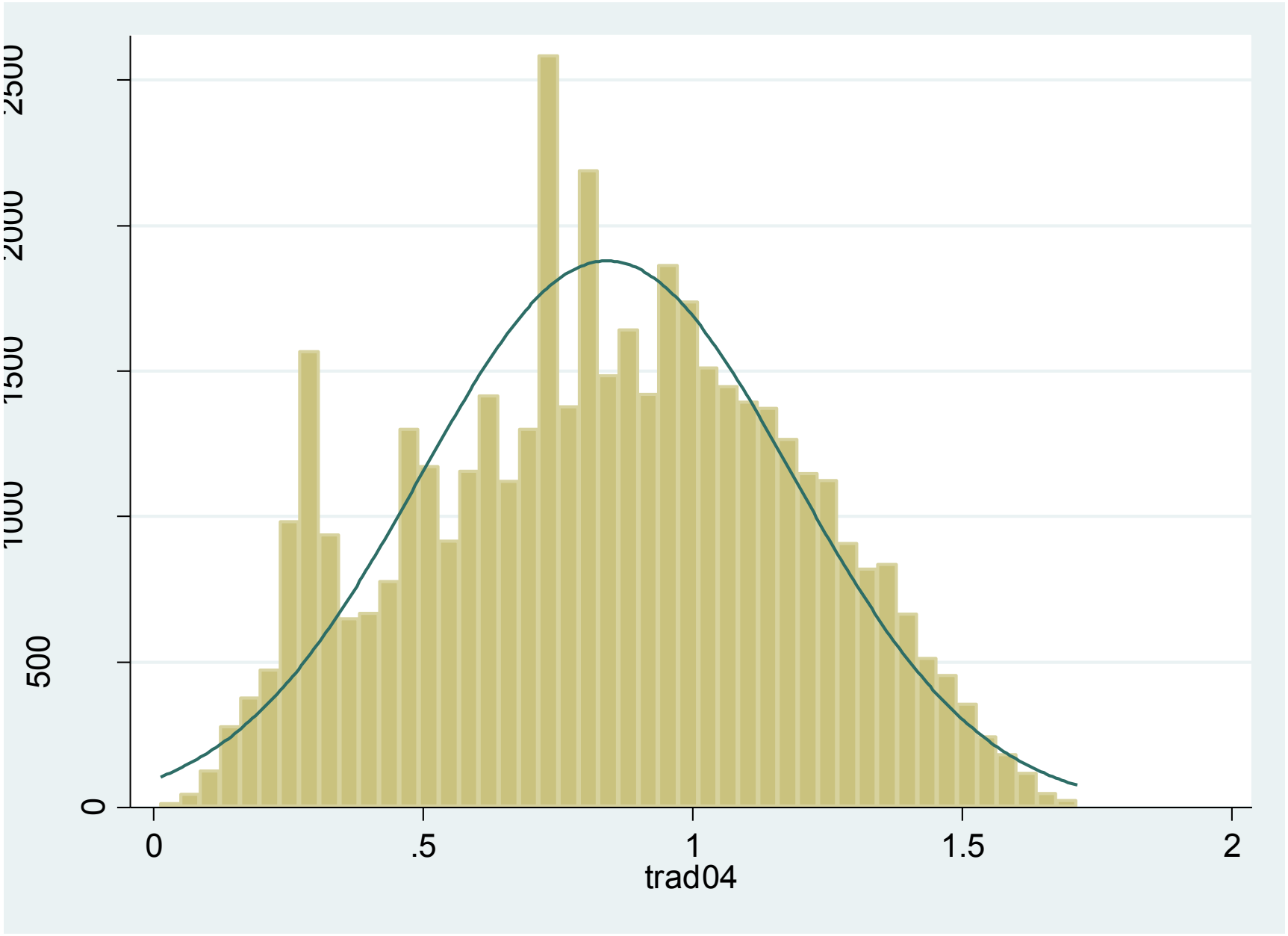
LOGARITMO DO ÍNDICE VALORES RACIONAIS



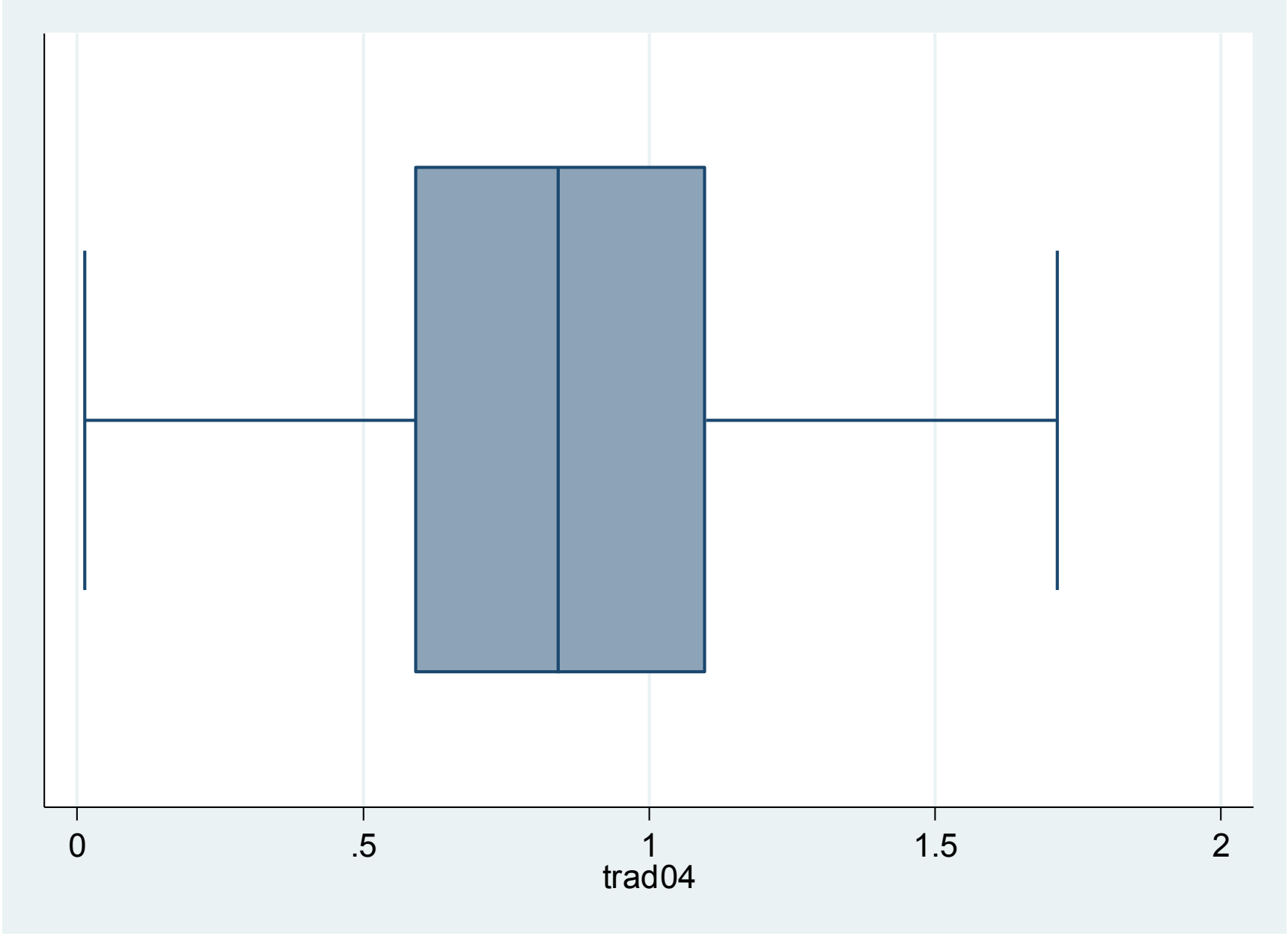
LOGARITMO DO ÍNDICE VALORES RACIONAIS



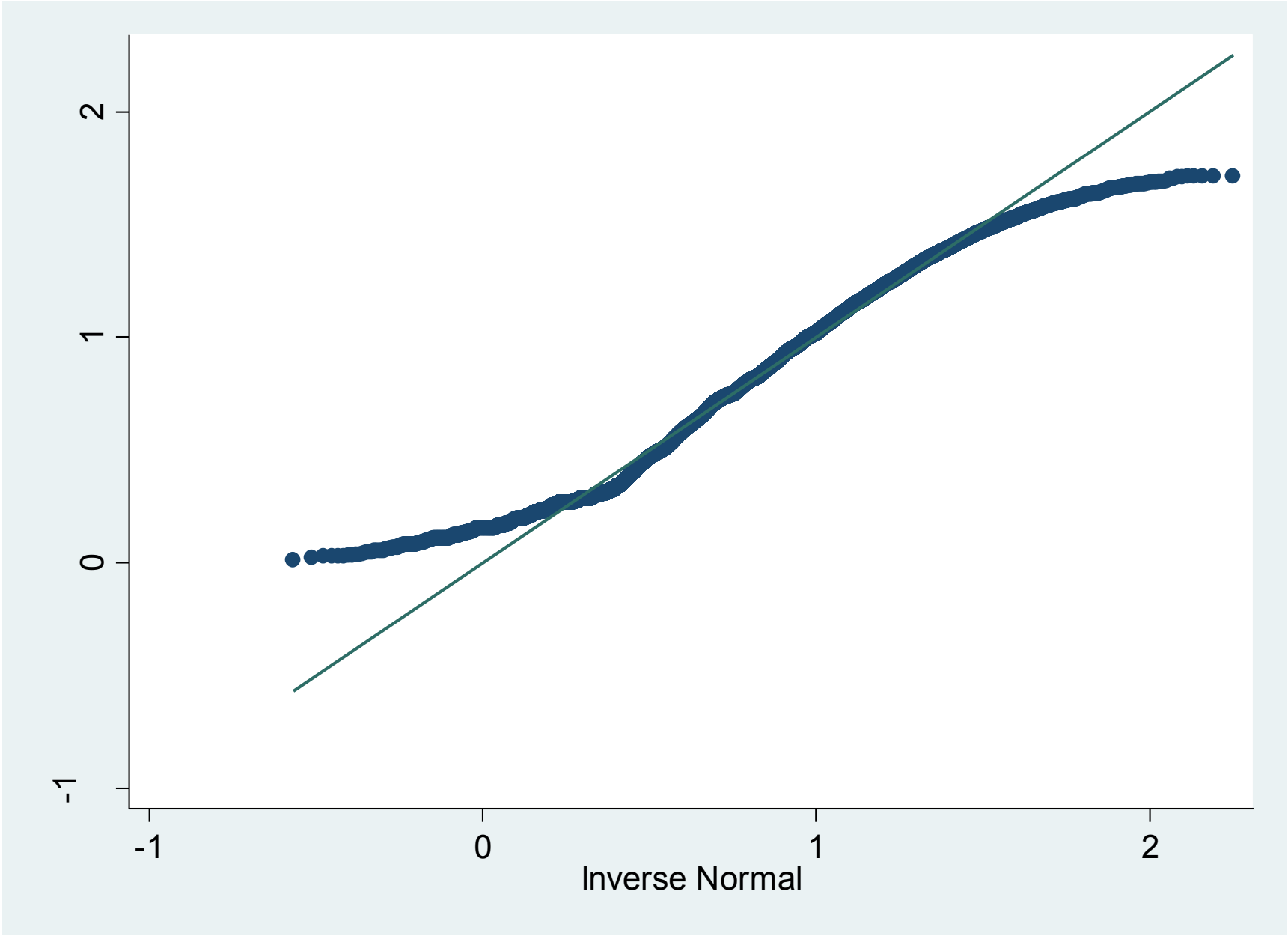
ÍNDICE VALORES RACIONAIS ELEVADO A 0,4



ÍNDICE VALORES RACIONAIS ELEVADO A 0,4



ÍNDICE VALORES RACIONAIS ELEVADO A 0,4



UTILIZAÇÃO DE PESOS AMOSTRAIS

PESOS EM BANCOS DE DADOS

– Expande os resultados da amostra para o tamanho populacional.

– Ao realizar inferência estatística, levamos em consideração o peso, o qual é o inverso da probabilidade da observação ser incluída no banco, devido ao desenho amostral.

– Por exemplo, o uso desse peso é importante na amostra do Censo Demográfico e na Pesquisa Nacional por Amostra de Domicílios (PNAD) do Instituto Brasileiro de Geografia e Estatística (IBGE) para expandir a amostra para o tamanho da população do país.

DIFERENTES PESOS

Indivíduo	Número de observações coletadas na amostra	Peso para expandir para o tamanho da população (N)	Peso para manter o tamanho da amostra (n)
João	1	4	0,8
Maria	1	6	1,2
Total	2	10	2

EXEMPLO:
Peso amostral do João =
Peso de frequência do João * (Peso amostral total / Peso de frequência total)