Insper

Ciência dos Dados

Modelos Probabilísticos Contínuos

Distribuição Normal
Probability plot

*Ver detalhes no livro

Magalhães e Lima, 7ª edição. Seção 6.2

Modelo Normal

Modelo fundamental em Probabilidade e Inferência Estatística.

“Em 12 de novembro de 1733, Abraham de Moivre publicou um artigo em latim contendo a dedução da distribuição normal como uma aproximação da distribuição binomial.”

Fonte: lista da ABE

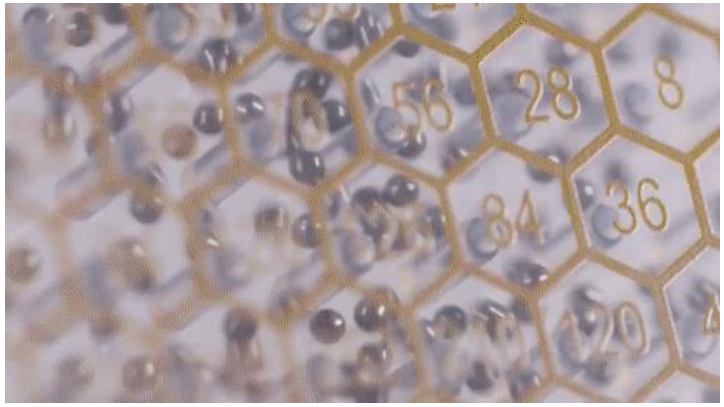
O uso dessa distribuição remonta a Gauss em seus trabalhos sobre erros de observações astronômicas, por volta de 1810, dando o nome de distribuição gaussiana para tal modelo.

Johann Carl Friedrich Gauss

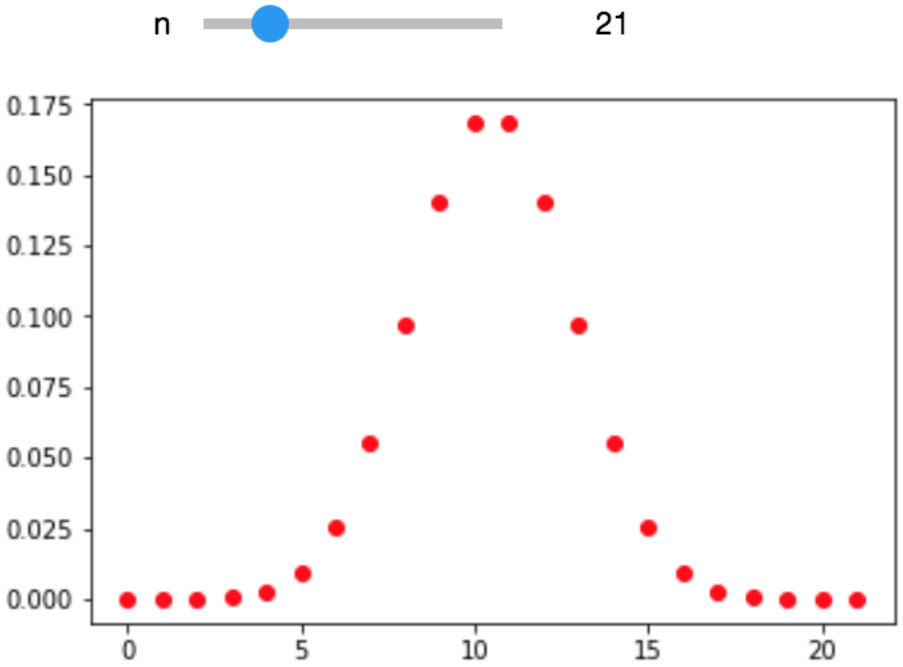
Alemanha 1777-1855

http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Gauss.html

Normal - origens



Binomial com muitos ensaios



Modelo Normal

a) Função Densidade de Probabilidade

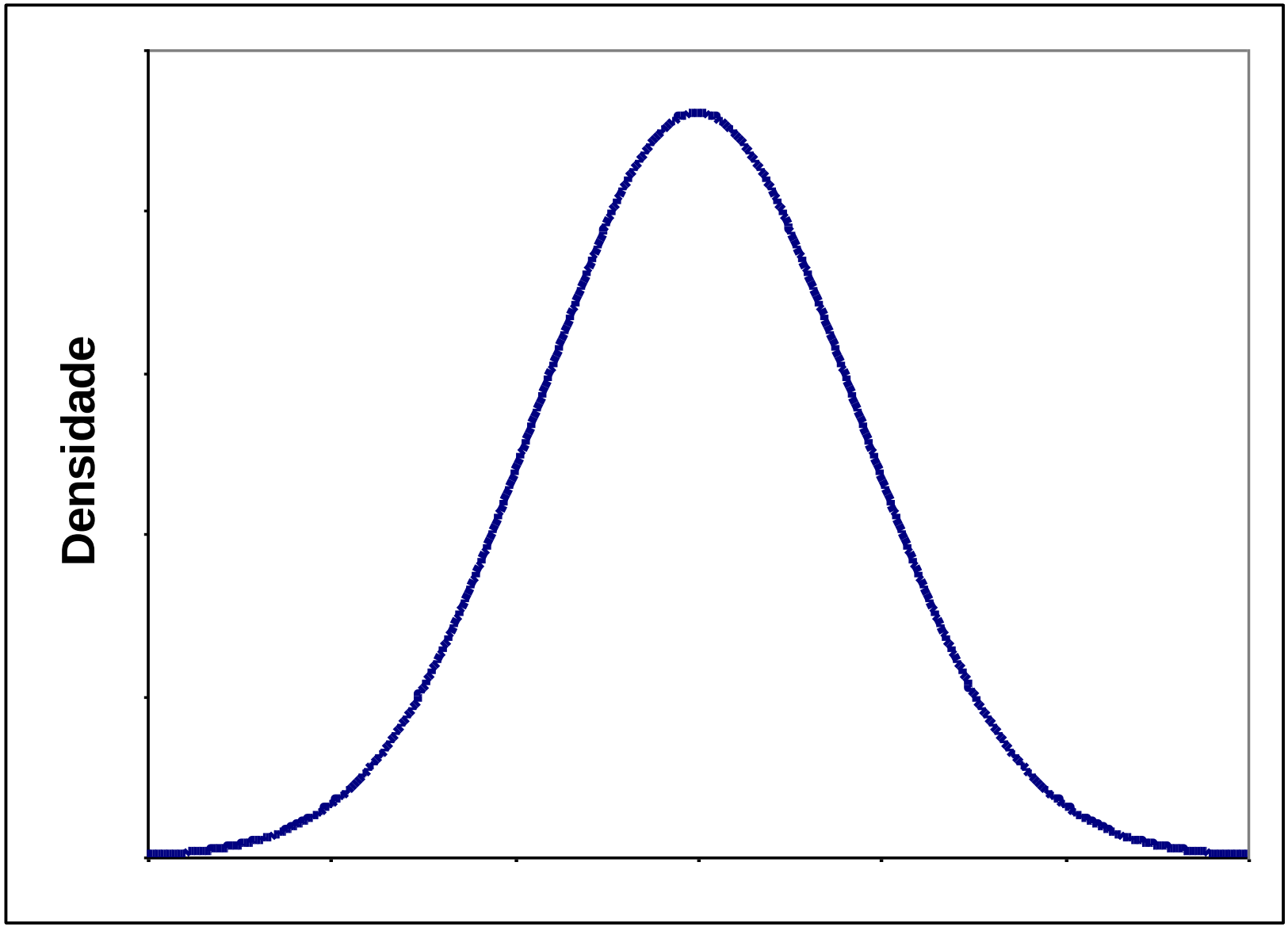
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

b) Notação: $X \sim N(\mu, \sigma^2)$

c) Parâmetros: $-\infty < \mu < \infty$ e $\sigma^2 > 0$.

Insper

Modelo Normal



Distribuição Normal

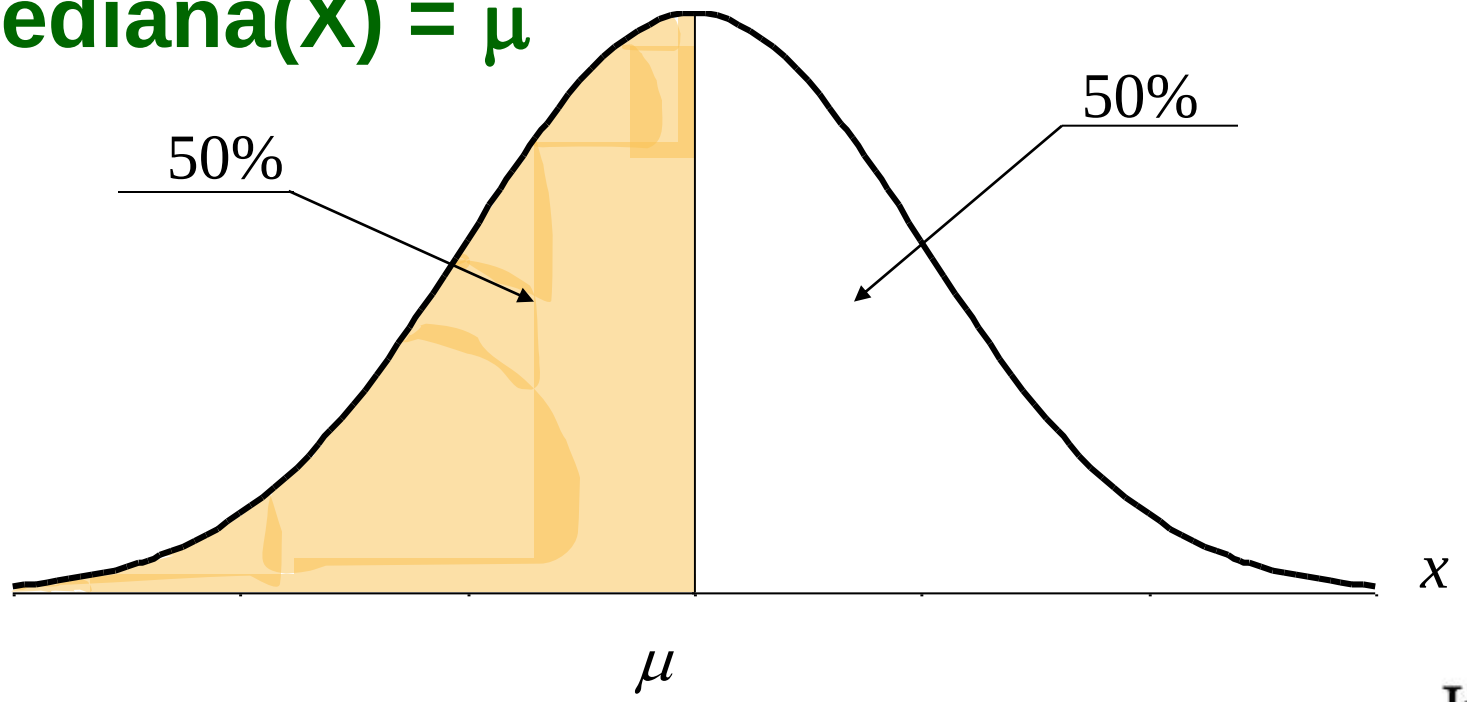
Med(X) = μ

Moda(X) = μ

Mediana(X) = μ

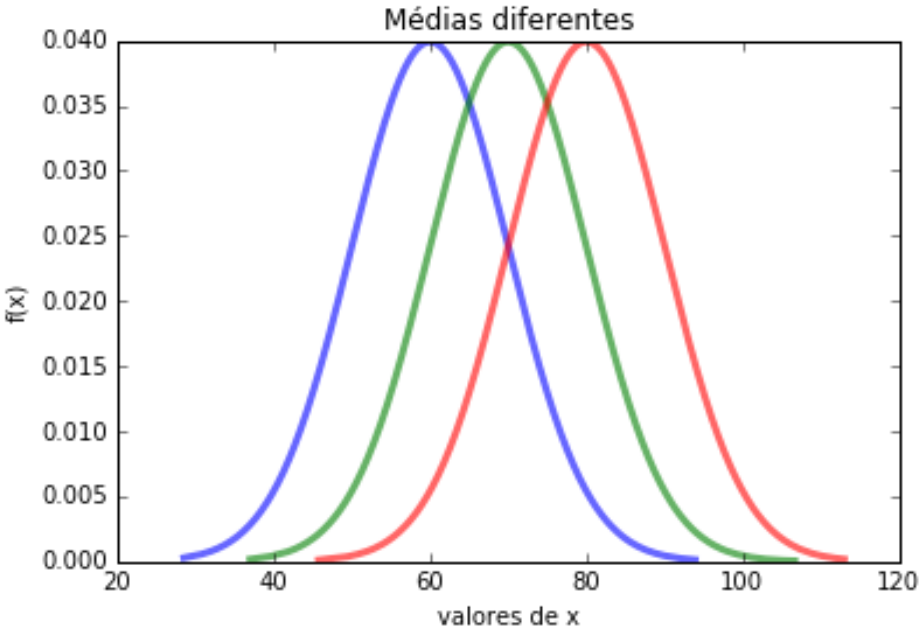
$$E(X) = \mu$$

$$Var(X) = \sigma^2$$



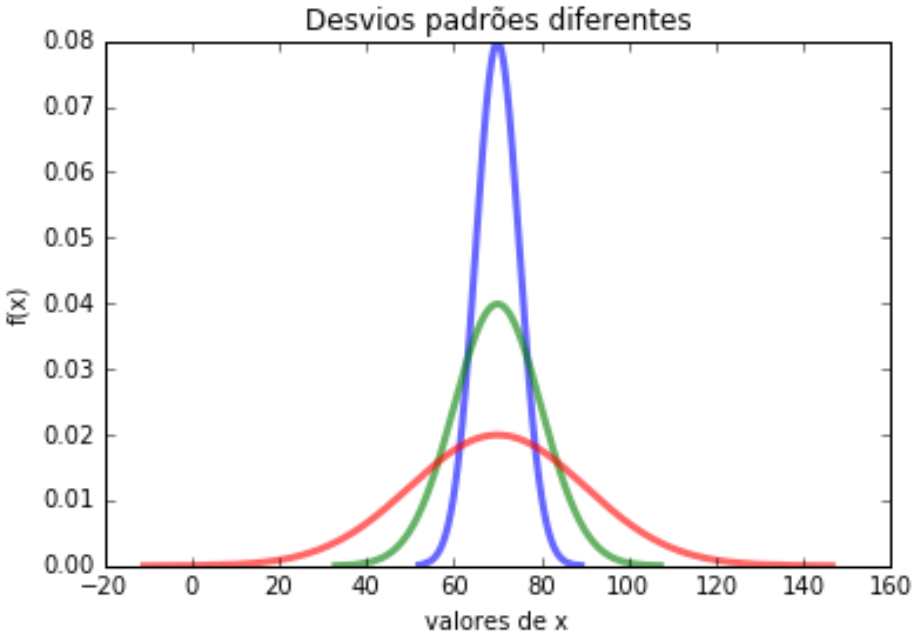
Insper

Distribuição Normal

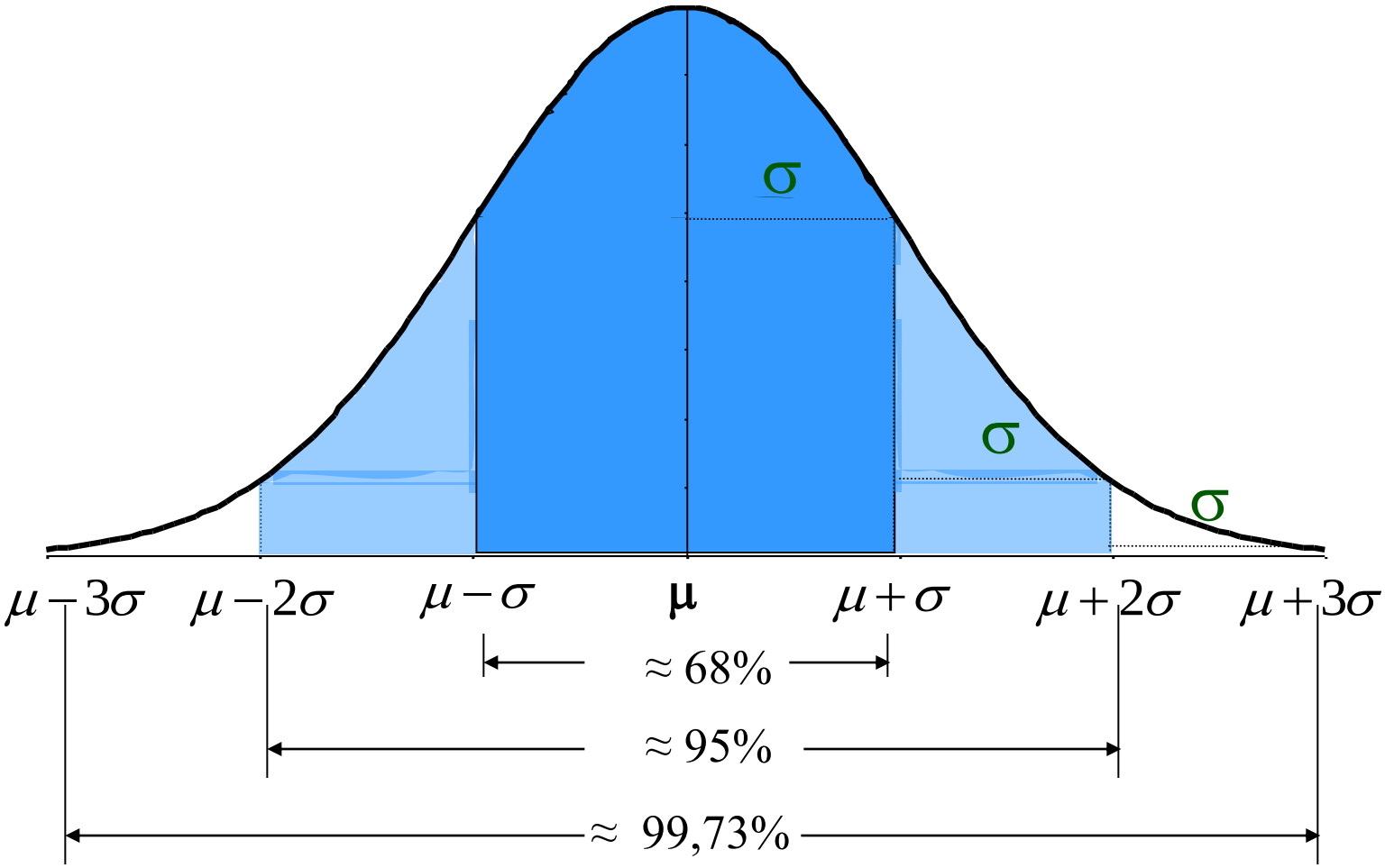


Médias diferentes e desvio padrões iguais

Médias iguais e desvio padrões diferentes



Distribuição Normal



Conhecido como Regra: 68-95-99

Insper

Distribuição Normal - Padronização

Muitas vezes estamos interessados em valores de probabilidade que a regra 68-95-99 não pode nos fornecer.

Como calcular a área abaixo da curva (probabilidade) nestes casos?

Cálculo da Integral

OU

Uso de algum software para obter probabilidade

OU

Padronização da curva Normal  **Tabela** *z*

Distribuição Normal - Padronização

Se X ~ *N*(μ, σ²), então a v.a. definida por

$$Z = \frac{X - \mu}{\sigma}$$

terá média zero e variância 1.

Ainda, prova-se que

$$Z = \frac{X-\mu}{\sigma} \sim N(0;1)$$

pois toda combinação linear de uma v.a. com distribuição normal também é uma normal.

Insper

Distribuição Normal - Padronização

Vimos que se X ~ *N*(μ, σ²), então

$$Z = \frac{X - \mu}{\sigma} \sim N(0;1)$$

Logo, para calcular áreas sob curvas normais que não a padrão, primeiro converta X em Z e depois procure o valor numa tabela apropriada ou no Excel ou no Python.

Distribuição Normal : Valores de P(Z ≤ z) = A(z)

Parte inteira e primeira decimal de z	Segunda decimal de z									
	0	1	2	3	4	5	6	7	8	9
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
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Exercício

Usando a tabela da normal padrão, calcule $P(Z > 0.32)$.

Resposta: Quebre o valor procurado em duas partes

1. Parte inteira + primeiro decimal:
   - Neste caso temos 0.3

2. Segundo decimal:
   - Neste caso temos 2

Agora basta consultar a tabela para obter $P(Z \leq 0.32) = 0.6255$

Contudo, estamos procurando o complemento disso! Logo a resposta é $P(Z > 0.32) = 1 - P(Z \leq 0.32) = 0.3745$

Distribuição Normal : Valores de P(Z ≤ z) = A(z)

Parte inteira e primeira decimal de z	Segunda decimal de z									
	0	1	2	3	4	5	6	7	8	9
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
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Funções úteis: cdf e ppf

Sabendo p, qual o a? `norm.ppf(p, loc=mu, scale=sigma)`

Probabilidade *p* acumulada até a

$$p = P(X < a) = \int_{-\infty}^{a} f(x)dx$$

Valor a até o qual a probabilidade é p

$$a \mid P(X < a) = p$$

Sabendo a, qual o p? `norm.cdf(a, loc=mu, scale=sigma)`

Probplot – QQ Plot

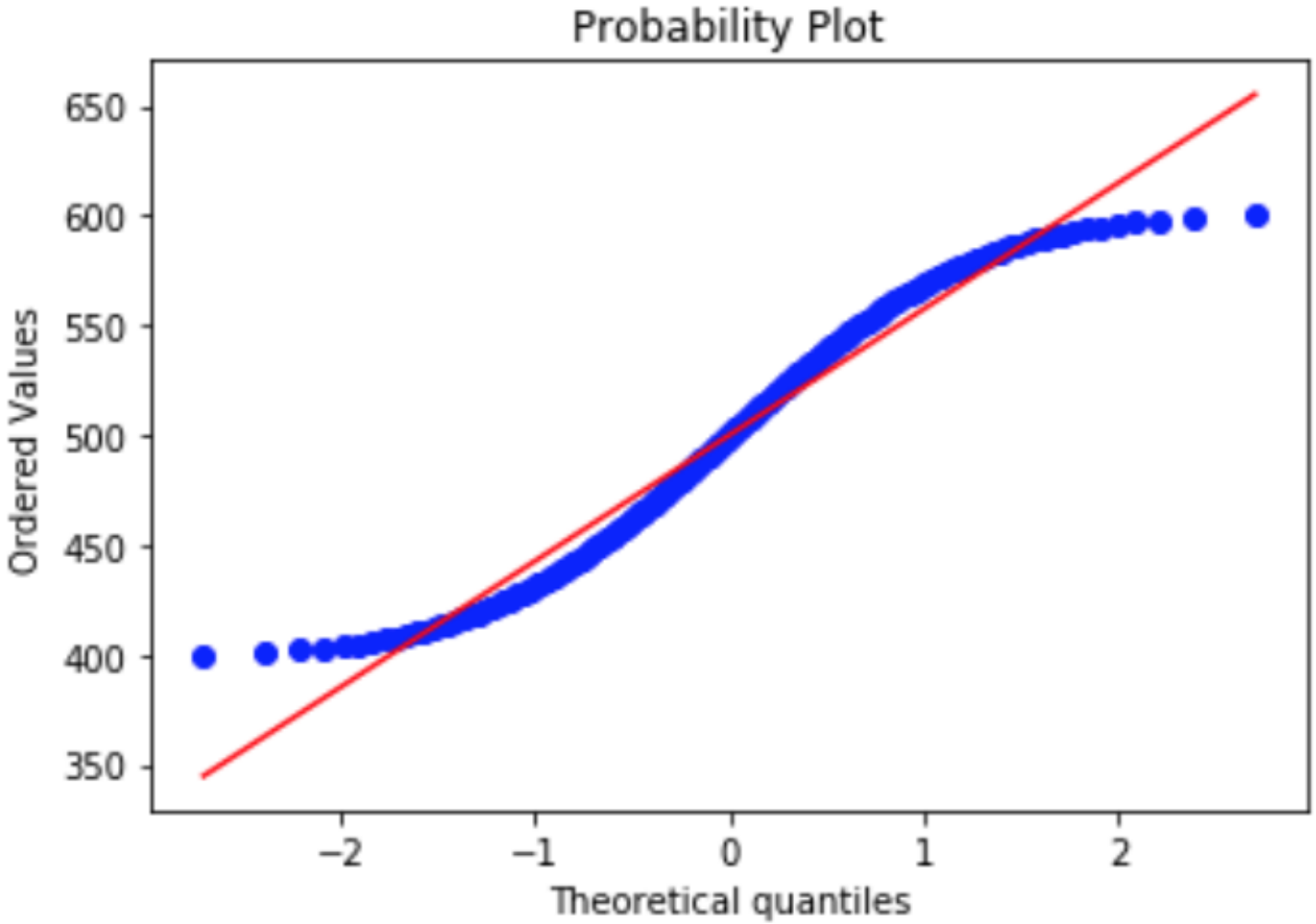
O probability plot (`scipy.stats.probplot`) é um tipo especial de plot quantil-quantil (QQ-Plot).

Permite comparar o quão bem um dataset ajusta uma distribuição

Exemplo de uso:

```
probplot(x, dist="norm", plot=plt)
```

Probability plot – QQ Plot



Insper

Fit – ajuste da distribuição aos dados

Todas as distribuições do pacote scipy.stats possuem um método `fit()`, que permite estimar como ajustar os dados pela distribuição

```
loc, scale  = norm.fit(dados_float)
```

Obs: No caso da normal, os parâmetros são `loc` e `scale`, mas poderiam ter outros significados em outras distribuições.

Exercícios

Exercício 1

Para $X \sim N(90, 100)$, calcular:

a) $P(X \leq 115)$.

b) $P(X \geq 80 \mid X < 100)$.

c) O valor de a tal que

$P(90 - a \leq X \leq 90 + a) = 0{,}99$.

d) O número d tal que $P(X < d) = 0{,}975$.

e) O número e tal que $P(X > e) = 0{,}95$.

Exercício 2

Seja $X \sim N(\mu, \sigma^2)$, encontre:

a) $P(X \geq \mu + 2\sigma)$.
b) $P(|X - \mu| \leq \sigma)$.
c) O número a tal que $P(\mu - a\sigma \leq X \leq \mu + a\sigma) = 0{,}99$.
d) O número d tal que $P(X > d) = 0{,}90$.

Exercício 3

As notas no quiz final de Ciência dos Dados distribuem-se segundo uma variável aleatória normal com média 6,5 e desvio padrão 1,6. O professor deseja dividir a classe em 3 categorias, da seguinte forma: os 30% que tiveram as melhores notas serão aprovados, os 50% com notas intermediárias ficarão de exame e os 20% que tiveram as piores notas serão reprovados.

a. Quais os limites de nota entre cada uma das categorias?

5,156 e 7,332

b. Caso a nota para aprovação (sem ir para exame) fosse igual a 7,0 e uma turma tivesse 50 alunos, quantos desses seriam aprovados sem ir para o exame? **37,83% => 19 alunos**

Insper

Exercício 4

Em um processo industrial, o diâmetro de um rolamento é uma parte importante do processo.

Sabe-se que a probabilidade de um rolamento ter diâmetro maior do 2,98 cm é de 80%.

Sabe-se, também que a probabilidade de que um rolamento tenha diâmetro abaixo de 2,97 cm é de 10%.

Admitindo que o diâmetro de um rolamento segue uma distribuição normal, determine a média e o desvio-padrão dos diâmetros dos rolamentos que saem da linha de produção. **2,9993 e 0,0227**

Determine a especificação que represente a maior distância da média, para mais ou para menos, contendo 95% dos rolamentos produzidos.

1,96*0,0227=0,04449

Insper

Exercício 5

Uma promotora de eventos está preparando um show. Este show será realizado no próximo mês e o local já foi escolhido, mas, como se trata de um espaço aberto, as condições do tempo devem ser consideradas. Ela sabe que em dias de chuva este tipo de show gera um lucro com distribuição Normal (média=100 mil Reais; desvio padrão = 20 mil Reais) e em dias ensolarados o lucro tem distribuição Normal (média = 110 mil Reais; desvio padrão = 30 mil Reais). A probabilidade de chuva no próximo mês é 0,60.

a) Em um dia de sol, qual é a probabilidade do lucro ser superior a 120 mil Reais?

b) Em um show que o lucro foi superior a 98 mil Reais, qual é a probabilidade de ter chovido?

c) Uma outra promotora também está interessada neste local e está oferecendo 105 mil Reais para poder utilizá-lo. Qual deveria ser a probabilidade de chuva para que o lucro esperado com a realização do show seja igual à quantia oferecida pela outra promotora?

a) 37,07% **b) 55,3%** **c) 50%**

Insper

Exercício 6

Um determinado calçado é vendido em lojas populares e em lojas sofisticadas. De todas as lojas, 70% são populares e 30% são sofisticadas. Nas lojas populares seu preço segue uma distribuição normal com média 8 e desvio-padrão 1,2. Já em lojas sofisticadas, o preço também segue uma distribuição normal de média 16 e desvio-padrão 3.

a) Determine o primeiro quartil da distribuição de preços de uma loja popular. **7,196**

b) Gastou-se mais de $10,00 para comprar o calçado. Qual é a probabilidade da compra ter sido feita numa loja popular? **10,2%**