

Modelagem e Análise de Sistemas - COS767

Aula passada

Teste de hipótese duas
médias (caso 3)

Teste de hipótese:
variância

Aula de hoje

Teste de hipótese: variância

Goodness of fit: v.a. discreta

Goodness of fit: v.a. contínua

Kolmogorov-Smirnov

Fitting de função
distribuição

Teste a respeito de variância

- Suponha que a hipótese nula seja do tipo $H_0: \sigma = \sigma_0$, e assumindo amostras de uma população com distribuição Normal (μ, σ_0^2) . Podemos definir a seguinte estatística:

$$X_{n-1}^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

X_{n-1}^2 tem distribuição chi-squared com $(n-1)$ graus de liberdade

Teste a respeito de variância

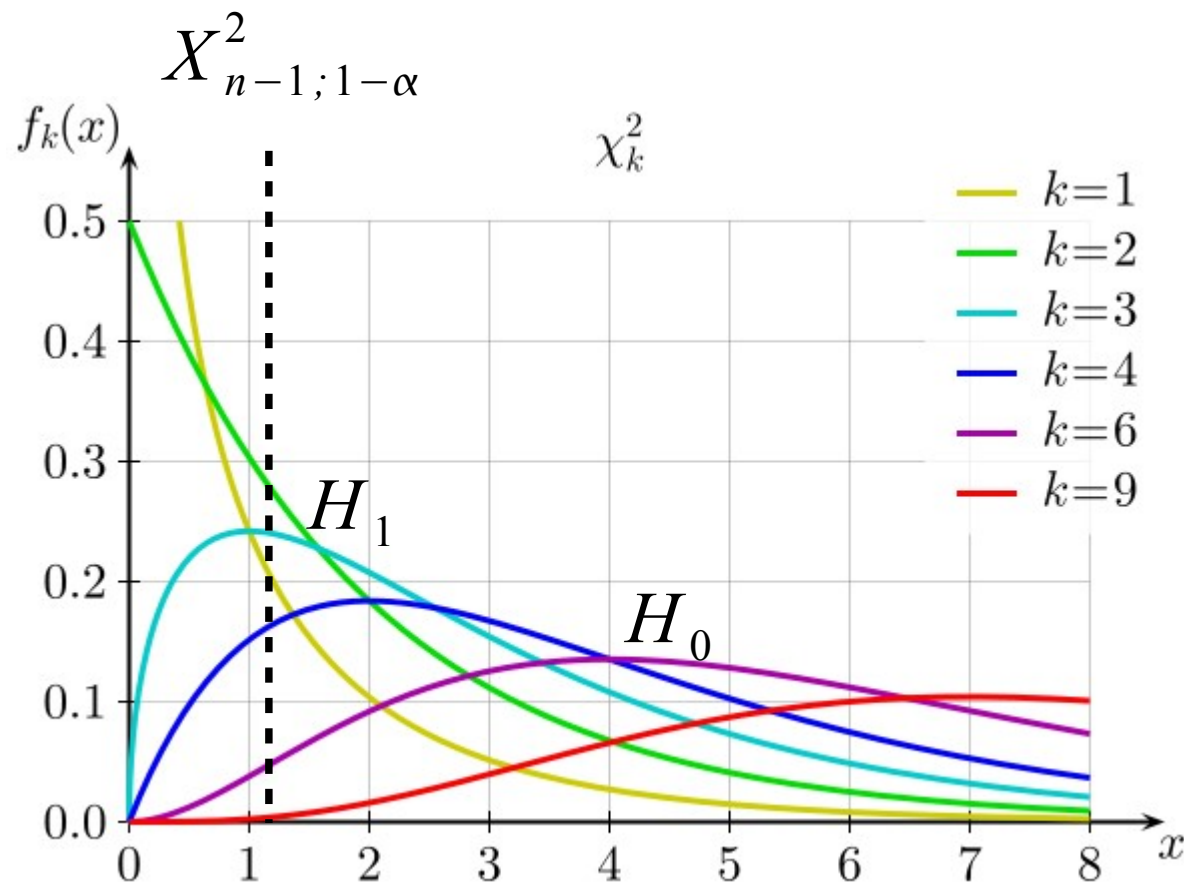
■ Variância da v.a. Chi-square = $2(n-1) = 2k$

■ Considere o teste:

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2$$

$$X_{n-1}^2 < X_{n-1;1-\alpha}^2$$



Teste a respeito de variância

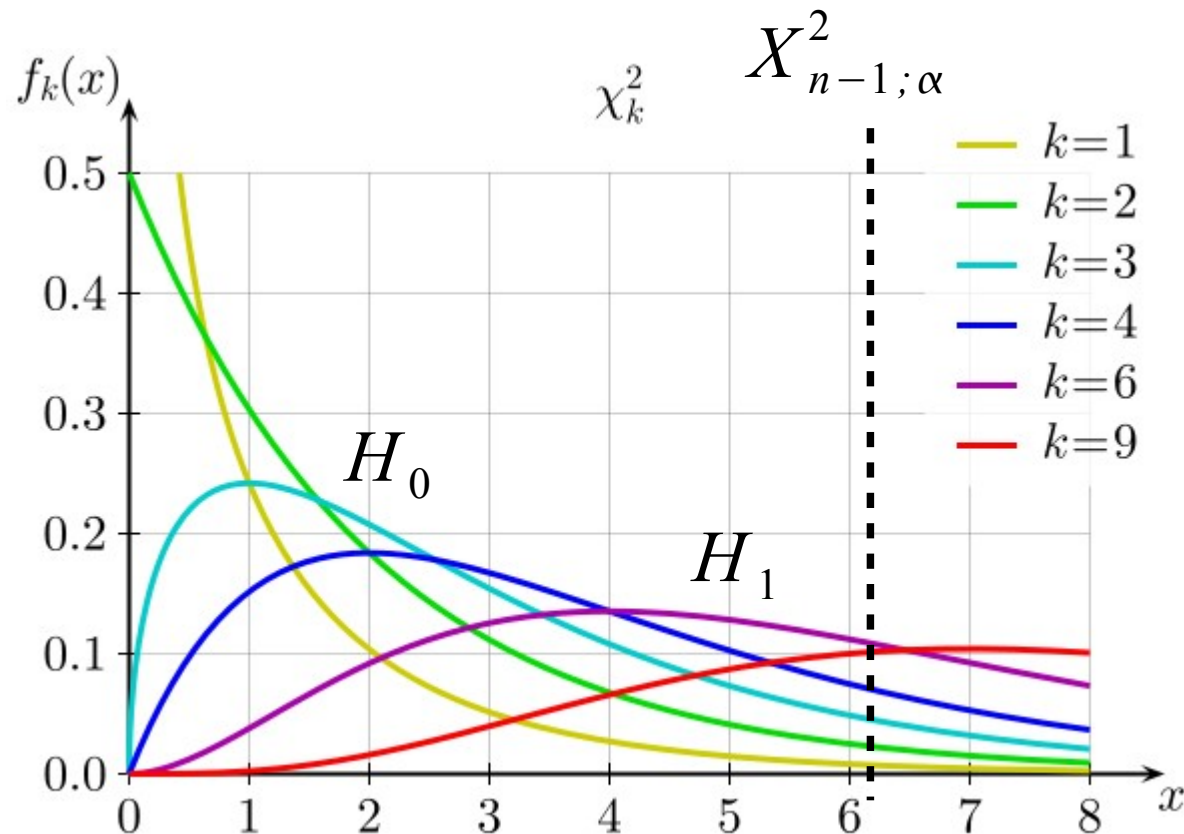
■ Variância da v.a. Chi-square = $2(n-1) = 2k$

■ Considere o teste:

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2$$

$$X_{n-1}^2 > X_{n-1;\alpha}^2$$



Teste a respeito de variância

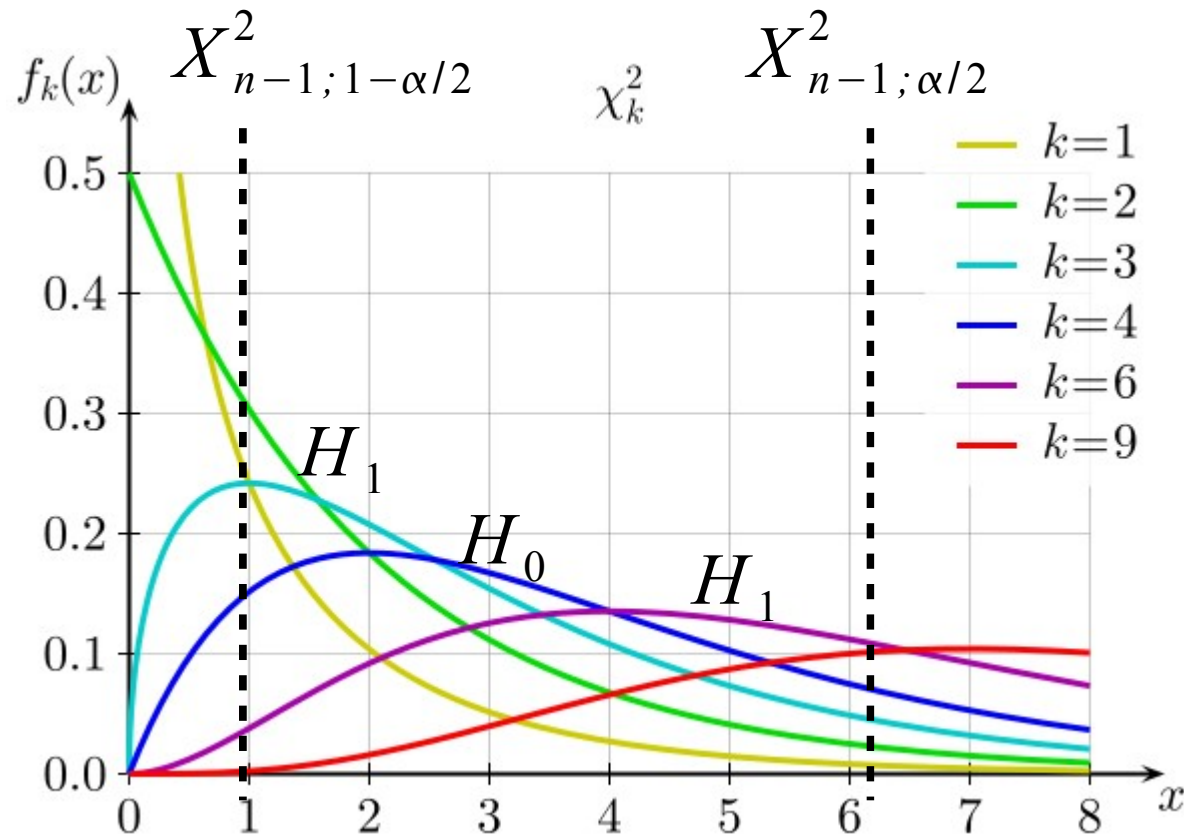
■ Variância da v.a. Chi-square = $2(n-1) = 2k$

■ Considere o teste:

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

$$X_{n-1}^2 > X_{n-1;\alpha/2}^2 \quad \text{ou} \\ X_{n-1}^2 < X_{n-1;1-\alpha/2}^2$$



Teste a respeito de variância

- Exemplo 10.38 do Livro do Trivedi (pg. 713)

Testes para distribuição de probabilidade

- Usado para testar se uma distribuição empírica corresponde a uma determinada distribuição da literatura
- Considere X uma v.a. discreta com pmf $p_X(i)=p_i$
- Hipóteses:

$$H_0: p_i = p_{i_0}, \quad i=0,1,\dots,k-1$$

$$H_1: \text{not } H_0$$

Teste para distribuição de probabilidade (v.a. discreta)

- Considere que são coletadas n observações e que a v.a. N_i representa o número de vezes que a v.a. X assumiu o valor i .
- N_i tem distribuição binomial com parâmetros n e p_i .
- $E[N_i] = np_i$ e $\text{Var}[N_i] = np_i(1-p_i)$
- É mostrado que Q possui distribuição chi-square com $(k-1)$ graus de liberdade

$$Q = \frac{\sum_{i=0}^{k-1} (N_i - np_i)^2}{np_i}$$

Teste para distribuição de probabilidade (v.a. discreta)

- Sob a hipótese $H_0 : p_i = p_{i0}$, a estatística Q pode ser calculada a partir de:

$$X_{k-1}^2 = \frac{\sum (\text{observed} - \text{expected})^2}{\text{expected}}$$

- A hipótese H_0 é rejeitada se: $X_{k-1}^2 > X_{k-1; 1-\alpha}^2$

Exemplo de teste v.a. discreta

We generated 1,000 random numbers for normal, double exponential, t with 3 degrees of freedom, and lognormal distributions. In all cases, a chi-square test with $k = 32$ bins was applied to test for normally distributed data. Because the normal distribution has two parameters, $c = 2 + 1 = 3$

The normal random numbers were stored in the variable Y1, the double exponential random numbers were stored in the variable Y2, the t random numbers were stored in the variable Y3, and the lognormal random numbers were stored in the variable Y4.

H_0 : the data are normally distributed

H_a : the data are not normally distributed

Y1 Test statistic: $\chi^2 = 32.256$

Y2 Test statistic: $\chi^2 = 91.776$

Y3 Test statistic: $\chi^2 = 101.488$

Y4 Test statistic: $\chi^2 = 1085.104$

Significance level: $\alpha = 0.05$

Degrees of freedom: $k - c = 32 - 3 = 29$

Critical value: $\chi^2_{1-\alpha, k-c} = 42.557$

Critical region: Reject H_0 if $\chi^2 > 42.557$

Teste confirma hipótese que amostras foram geradas a partir de uma v.a. Normal

Teste rejeita hipótese de que amostras foram geradas a partir de uma v.a. Normal

Exemplo de teste v.a. discreta

- Exercício 10.41 do livro do Trivedi, pag. 716.

Foram coletados dados do número de erros que ocorre em um sistema. O número total de erros observados em 50 semanas foi 95. Queremos saber se podemos representar o número de erros por uma v.a. Poisson.

Exemplo de teste v.a. discreta

$$p_{i0} = P[Y=i],$$

onde Y é v.a.

Poisson

$$np_{i0}$$

i	N_i		
<i>Number of errors (i) in one-week period</i>	<i>Number of one-week periods with i errors</i>	<i>Poisson probabilities</i> $\hat{p}_{i0}$	<i>Expected frequencies</i> $n\hat{p}_{i0}$
0	14	0.150	7.50
1	11	0.284	14.20
2	9	0.270	13.50
3	6	0.171	8.55
4	5	0.081	4.05
5+	5	0.044	2.20

Exemplo de teste v.a. discreta

- An estimate of the rate parameter is

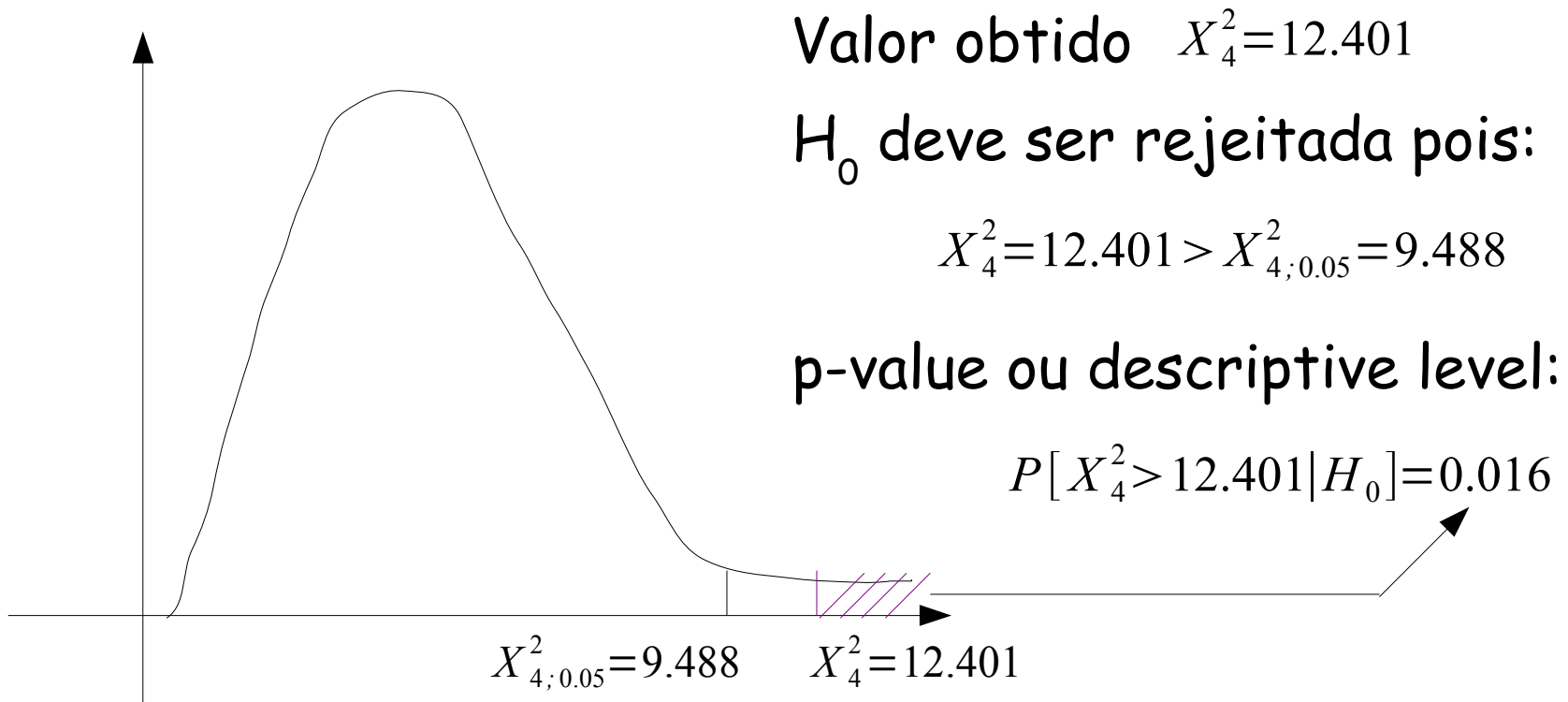
$$\hat{\lambda} = \frac{\text{total number of observed errors}}{\text{total number of weeks observed}} = \frac{95}{50} = 1.9 \text{ per week}$$

- The test statistic is then

$$\begin{aligned}\chi^2 &= \sum \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i} \\ &= \frac{(14 - 7.50)^2}{7.50} + \frac{(11 - 14.20)^2}{14.20} + \frac{(9 - 13.5)^2}{13.5} + \frac{(6 - 8.55)^2}{8.55} \\ &\quad + \frac{(5 - 4.05)^2}{4.05} + \frac{(5 - 2.20)^2}{2.20} \\ &= 5.633 + 0.7211 + 1.5 + 0.7605 + 0.2228 + 3.564 \\ &= 12.401\end{aligned}$$

Exemplo de teste v.a. discreta

- A v.a. chi-square possui $k-2$ graus de liberdade. Um grau de liberdade é perdido pois as amostras são usadas para estimar o valor do parâmetro λ da distribuição Poisson.



Teste CDF v.a. contínua

- Suponha X v.a. contínua e queremos testar se X possui determinada função distribuição

$$H_0: \text{para todo } x, F_X(x) = F_0(x)$$

$$H_1: \text{existe um } x \text{ tal que } F_X(x) \neq F_0(x)$$

- Usaremos o teste de Kolmogorov-Smirnov (o teste chi-square poderia ser usado mas perderíamos precisão)

Teste CDF v.a. contínua

- Ordenar em ordem crescente os valores das amostras:

$$x_1 \leq x_2 \leq x_3 \dots \leq x_n$$

- Calcular a função distribuição empírica da v.a. X:

$$\hat{F}_n(x) = \begin{cases} 0, & x < x_1 \\ \frac{i}{n}, & x_i \leq x < x_{i+1} \\ 1, & x_n \leq x \end{cases}$$

Teste CDF v.a. contínua

- A estatística Komolgorov-Smirnov é definida como:

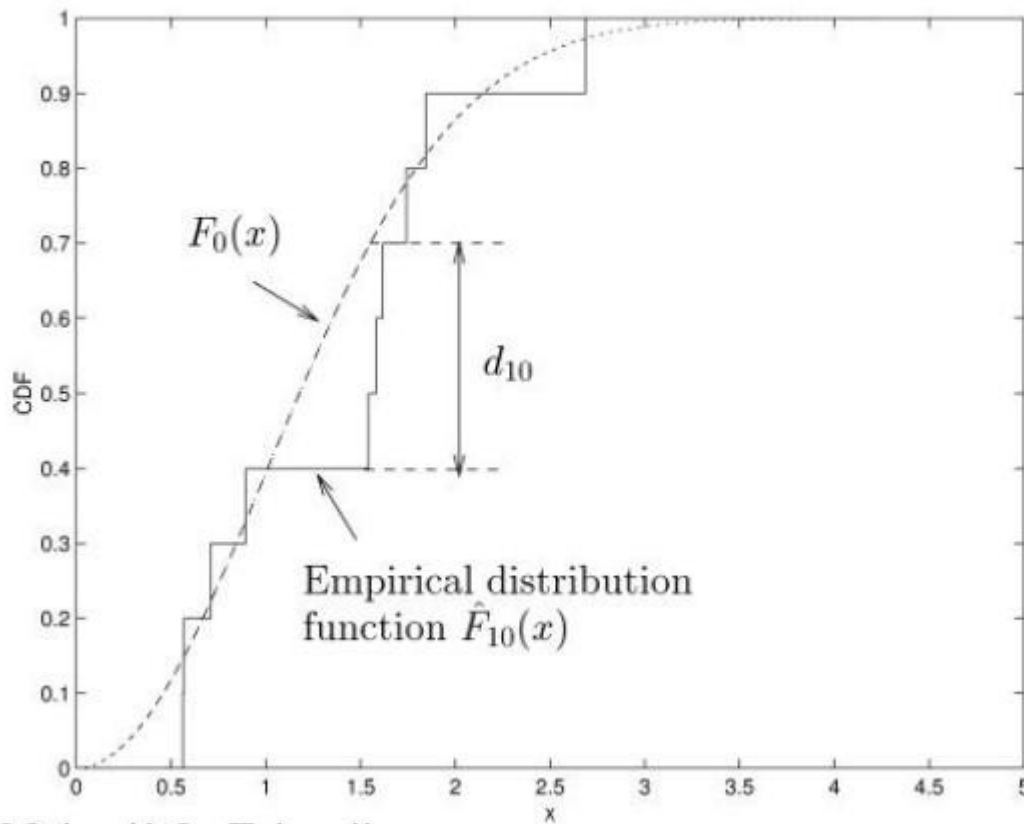
$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$$

- O uso do supremo ao invés do máximo é porque a função empírica é descontínua.
- Pela definição acima, a diferença entre as funções $|\hat{F}_n(x) - F_0(x)|$ só será avaliada no extremo final do intervalo $[x_i, x_{i+1}]$ pois o valor no intervalo é constante.

Teste CDF v.a. contínua

■ A estatística Komolgorov-Smirnov é definida como:

$$D_n = \sup_x |\hat{F}_n(x) - F_0(x)|$$



■ A hipótese H_0 é rejeitada para um dado nível de significância α se:

$$D_n > d_{n,\alpha}$$

Teste CDF v.a. contínua

- A vantagem da estatística D_n é que não é preciso fazer nenhuma aproximação, ela é exata e só depende de n , não depende de $F_0(x)$.
- Já a estatística Q usada no teste para v.a. discretas (Chi-square test) é uma aproximação de uma v.a. Chi-square e é preciso um conjunto grande de amostras para que a aproximação seja boa.
- Exemplo 10.42, livro do Trivedi, pag., 718

Como fazer fitting de função distribuição ?

- Passo 1: Coletar amostras e parametrizar um conjunto de funções distribuição usando métodos dos momentos ou MLE.
- Passo 2: Usar teste Chi-square ou Kolmogorov-Smirnov
- Passo 3: Traçar QQ plot
- Passo 4: Traçar gráfico x versus $1-F_x(x)$ com eixo das ordenadas em escala log para analisar a cauda da distribuição.

Classificação de função distribuição

- Uma função distribuição pode ter até três parâmetros: location, scale, shape.
- Location: Desloca o gráfico da função densidade para direita ou esquerda com relação ao eixo das abcissas (ex: média da v.a. Normal)
- Scale: Aumenta ou diminui a largura da distribuição (ex: variância da v.a. Normal)
- Shape: Permite mudar a forma da distribuição (ex: parâmetro alpha da v.a. Weibull)

Parâmetros location e scale de uma função distribuição

- As distribuições podem pertencer a uma mesma família com relação aos parâmetros location, scale e location-scale.
- Se X é uma variável aleatória que pertence a uma certa família, então $Y =_d (a + bX)$ é uma outra distribuição que pertence a mesma família. ($=_d$ significa "possui a mesma distribuição que")

Parâmetro location

Location Model

- Say Z is some real random variable, then we can create a family of random variables by adding μ for any real μ .
- Possibly: measure a quantity μ but the physical measurement causes a random error Z . Then we observe $X = Z + \mu$. Or else we independently take repeated measures of the same quantity to get iid $X_i = Z_i + \mu$.
- We want to have a general distribution (known) depending on the circumstances Z .
- The density of X is

$$f_X(x) = f_Z(x - \mu)$$

Examples

Uniform $Z \sim \text{Unif}[-1, 1] \implies X \sim \text{Unif}[\mu - 1, \mu + 1]$.

Normal $Z \sim \mathcal{N}(0, 1) \implies X \sim \mathcal{N}(\mu, 1)$.

Parâmetro Scale

Scale Family

- If we take a random variable Z and multiply by a positive scaling quantity σ then we get $X = \sigma Z$

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x}{\sigma}\right)$$

Examples

Normal $Z \sim \mathcal{N}(0, 1) \implies X \sim (0, \sigma^2)$

Gamma $Z \sim \Gamma(\alpha, 1) \implies X \sim \Gamma(\alpha, \sigma)$

Parâmetro Location-Scale

Location-Scale Family

- The two things together.
- There exists a random variable Z that is

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

which has the same distribution under all values of the parameter.

- The density of the distributions in this model are

$$f_X(x) = \frac{1}{\sigma} f_Z\left(\frac{x - \mu}{\sigma}\right)$$

Examples

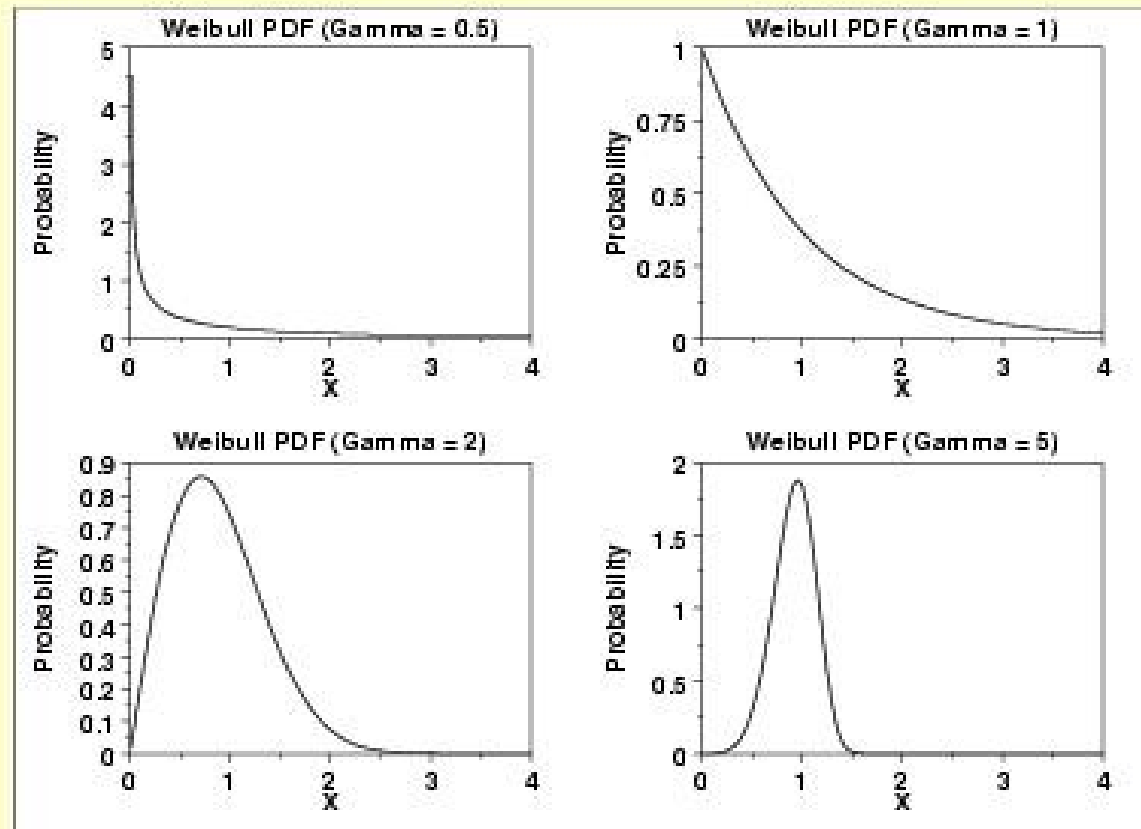
Normal $Z \sim \mathcal{N}(0, 1) \implies X \sim \mathcal{N}(\mu, \sigma^2)$.

Double Exponential Z has density $1/2e^{-|x|}$ then

$$f_X(x \mid \mu, \sigma) = \frac{1}{2\sigma} \exp\left[-\frac{|x - \mu|}{\sigma}\right].$$

Parâmetro shape

- O parâmetro shape permite a distribuição mudar de forma de acordo com o seu valor.
- Exemplo: v.a. Weibull



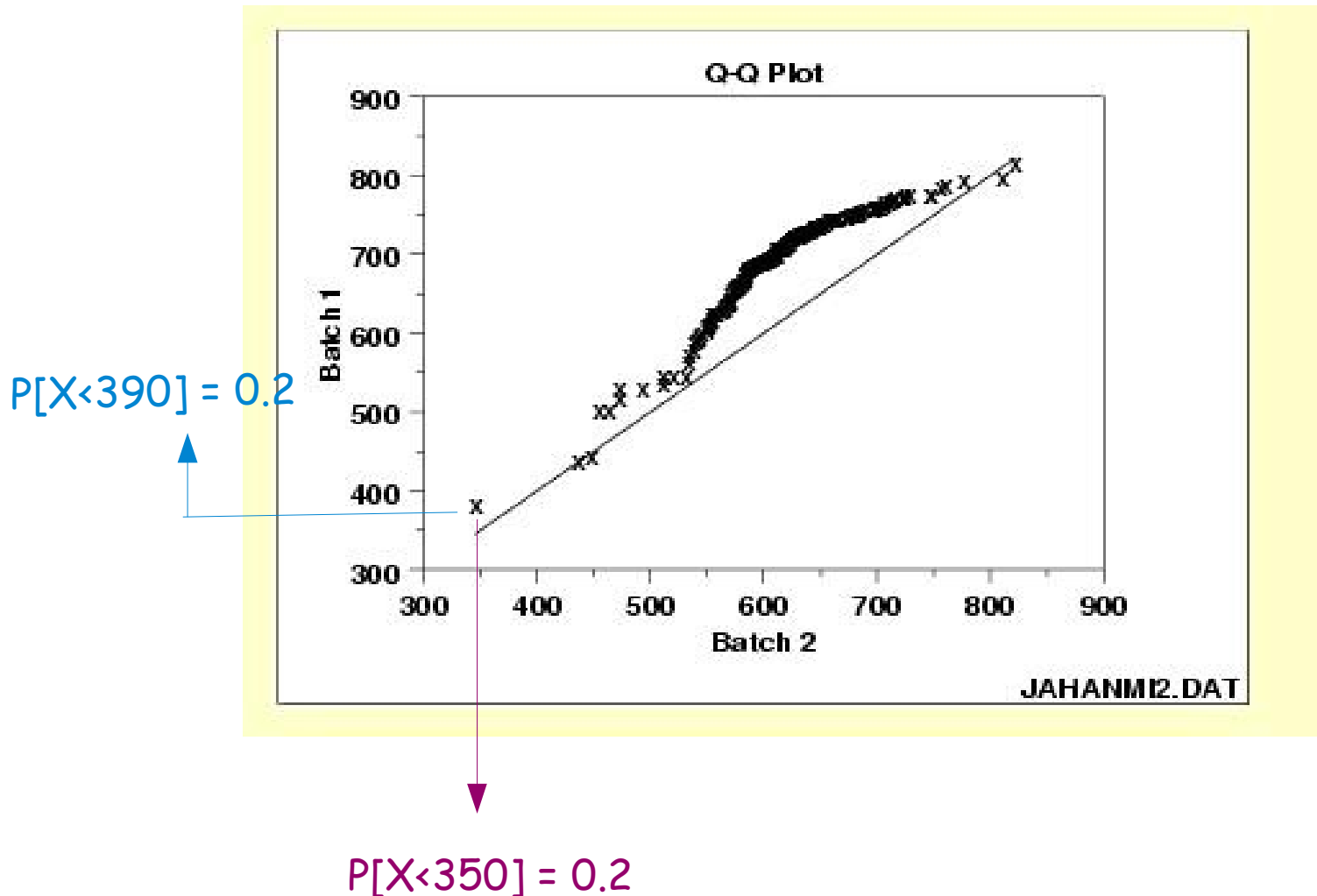
Quantile-Quantile Plot

- Método gráfico para indicar que dois conjuntos de amostras pertencem a populações com a mesma distribuição de probabilidade.
- By a quantile, we mean the point below which a given fraction (or percent) of points lies. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value.
- A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.

Quantile-Quantile Plot

- Vantagem: Os conjuntos de amostras podem ter tamanhos diferentes
- O que pode ser observado com relação a função de distribuição das amostras :
 - (1) presença de outliers
 - (2) deslocamentos em location
 - (3) deslocamentos em scale
 - (4) comportamento da cauda

Exemplo Q-Q Plot



1 -Pode-se notar que o valor das amostras do conjunto 2 é maior do que o valor das amostras do conjunto 1

2 - Os dois conjuntos tem alta probabilidade de não pertencerem a populações com a mesma distribuição de probabilidade