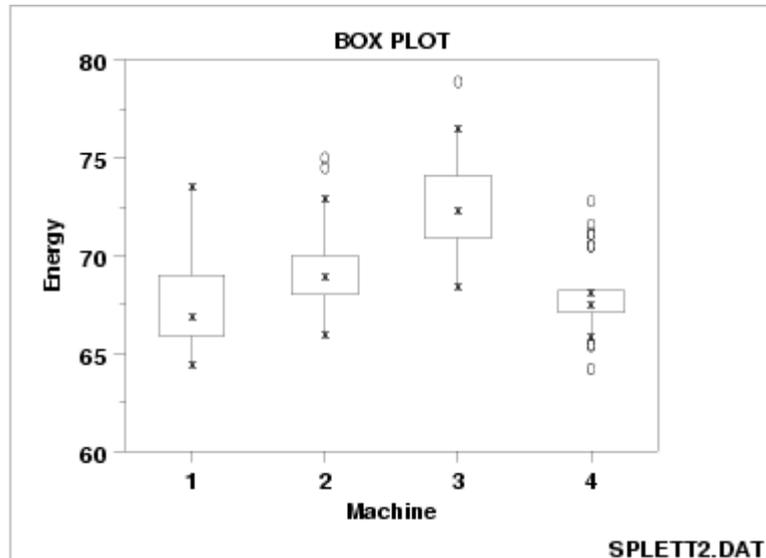


O QUE É BOX PLOT ?

*Finalidade:
Verifique
deslocamentos
da posição e
da variação*

Os lotes da caixa (câmaras 1983) são uma ferramenta excelente para fazer saber à informação da posição e da variação em séries de dados de, particularmente para detectar e ilustrar mudanças da posição e da variação entre grupos diferentes dos dados.

*Lote Da
Amostra:
Este lote da
caixa revela
que a
máquina tem
um efeito
significativo
na energia
com respeito à
posição e
possivelmente
à variação*



Este lote da caixa, comparando quatro máquinas para a saída da energia, mostra que a máquina tem um efeito significativo na energia com respeito à posição e à variação. A máquina 3 tem a resposta da energia a mais elevada (aproximadamente 72,5); a máquina 4 tem menos resposta variável da energia com as aproximadamente 50% de suas leituras que estão dentro de 1 unidade da energia.

Definição

Os lotes da caixa são dados forma perto
Linha central vertical: Variável da resposta
Linha central horizontal: O fator do interesse
Mais especificamente, nós

1. Calcule o número médio e os quartiles (o quartile mais baixo é o 25 percentile e o quartile superior é o 75th percentile).
2. Trace um símbolo no número médio (ou extraia uma linha) e extraia uma caixa (daqui o nome -- encaixote o lote) entre os quartiles mais baixos e superiores; esta caixa representa os 50% médio dos dados -- o "corpo" dos dados.
3. Extraia uma linha do quartile mais baixo ao ponto mínimo e uma outra linha do quartile superior ao ponto máximo. Um símbolo é

extraído tipicamente nestes pontos mínimos e máximos, embora este seja opcional.

Assim o lote da caixa identifica os 50% médio dos dados, do número médio, e dos pontos extremos.

Os únicos ou lotes múltiplos da caixa podem ser extraídos

Um único lote da caixa pode ser extraído para um grupo dos dados com nenhuns grupos distintos. Alternativamente, os lotes múltiplos da caixa podem ser extraídos junto para comparar séries de dados de múltiplas ou para comparar grupos em uma única série de dados de. Para um único lote da caixa, a largura da caixa é arbitrária. Para lotes múltiplos da caixa, a largura do lote da caixa pode ser ajustada proporcional ao número dos pontos no grupo ou na amostra dada (algumas execuções do software do lote da caixa ajustam simplesmente todas as caixas à mesma largura).

Lotes da caixa com cercas

Há uma variação útil do lote da caixa que identifica mais especificamente outliers. Para criar esta variação:

1. Calcule o número médio e os quartiles mais baixos e superiores.
2. Trace um símbolo no número médio e extraia uma caixa entre os quartiles mais baixos e superiores.
3. Calcule a escala interquartile (a diferença entre o quartile superior e mais baixo) e chame-a IQ.
4. Calcule os seguintes pontos:

$$L1 = \text{quartile mais baixo} - 1.5 * IQ$$

$$L2 = \text{quartile mais baixo} - 3.0 * IQ$$

$$U1 = \text{quartile superior} + 1.5 * IQ$$

$$U2 = \text{quartile superior} + 3.0 * IQ$$

5. A linha do quartile mais baixo ao mínimo é extraída agora do quartile mais baixo ao ponto o menor que é mais grande do que L1. Do mesmo modo, a linha do quartile superior ao máximo é extraída agora ao ponto o maior menor do que U1.
6. Os pontos entre L1 e L2 ou entre U1 e U2 são extraídos como círculos pequenos. Os pontos menos do que L2 ou mais grande do que U2 são extraídos como círculos grandes.

Perguntas

O lote da caixa pode dar respostas às seguintes perguntas:

1. É um fator significativo?
2. A posição difere entre subgrupos?
3. A variação difere entre subgrupos?

4. Há algum outliers?

<i>Importância: Verifique o significado de um fator</i>	O lote da caixa é uma ferramenta importante de EDA para determinar se um fator tiver um efeito significativo na resposta com respeito à posição ou à variação. O lote da caixa é também uma ferramenta eficaz para sumariar quantidades grandes da informação.	Com o extra ir um Boxp lot
<i>Técnicas Relacionadas</i>	<u>Lote Médio</u> <u>Análise de variação</u>	Há um méto do geral ment e aceit ado de
<i>Estudo De Caso</i>	O lote da caixa é demonstrado <u>no estudo de caso</u> cerâmico dos dados da força.	
<i>Software</i>	Os lotes da caixa estão disponíveis em a maioria de programas estatísticos do software da finalidade geral, including <u>Dataplot</u>	

extrair as suíças em um boxplot. Entretanto, há um plethora dos métodos para extrair a caixa. Alguns destes métodos foram desenvolvidos porque estendem agradável aos percentiles à exceção de 25% e de 75%, outros foram escolhidos por causa das considerações teóricas e alguns foram desenvolvidos para o simplicity. Bob Hayden escolheu seu método em parte para razões metaphysical! Neste papel eu opt para o simplicity, com um pouco de metaphysics jogado dentro para a medida boa.

Para uma discussão extensiva em extrair um boxplot recolheu dos email [à lista que enviar do Internet ap-ap-stats](#) eu sugiro que você lê as caixas [Ticky-ticky-Tacky do artigo](#).

Extraindo a caixa

Nós temos que começar com a caixa, como construir as suíças nos requer usa a caixa determinar se algum valor dos dados for outliers. Eu descreverei dois métodos, ambos os simples, de que você pode escolher. O primeiro está historicamente correto e tem a propriedade metaphysical agradável de Bob, quando o segundo for o método usado a calculadora por Ti-82 e por Ti-83 representando graficamente. Se seus estudantes usarem esta tecnologia e você quê-los começar as mesmas respostas que sua calculadora então você deve opt para este método.

Método De Tukey

O método recomendado por Tukey, que inventou o boxplot, é como segue:

Encontre o número médio. Então encontre o número médio dos valores dos dados cujos os Rank são MENOS DO QUE OU IGUALE-o ao Rank do número médio. Este será um valor dos dados ou será meia maneira entre dois valores dos dados.

Com uma série de dados com um número impar dos valores, inclua o número médio em cada uma das duas metades da série de dados e encontre então o número médio de cada metade. Isto dá os primeiros e terceiros quartiles. Se a

série de dados tiver um número uniforme dos valores, rache apenas os dados em duas metades, e encontre o número médio de cada metade.

Está aqui um exemplo usando uma série de dados pequena, que contenha *um número* ímpar dos valores:

35 47 48 50 51 53 54 70 75

Rache os dados em duas metades, *cada um including o número médio* :

35 47 48 50 51 e 51 53 54 70 75

Encontre o número médio de cada metade. Neste exemplo, o primeiro quartile é 48 e o terceiro quartile é 54. Daqui a escala interquartile é $54-48 = 6$.

Eu adicionarei um número à série de dados acima para ilustrar como encontrar os quartiles para um número uniforme dos valores (o que o heck, os dados é bogus de qualquer maneira):

35 47 48 50 51 53 54 60 70 75

Rache os dados em duas metades:

35 47 48 50 51 e 53 54 60 70 75

Encontre agora o número médio de cada metade. Neste exemplo, o primeiro quartile é 48 e o terceiro quartile é 60. Daqui o IQR é $60-48 = 12$.

Bob Hayden prefere este método porque o sumário de cinco números de cinco números dá os cinco se numera. Para o exemplo, faça exame a série de dados 1 de 4 78 81 345. O mínimo é 1, o máximo é 345, o número médio é 78. Rachar a série de dados em duas metades cada uma que contem o número médio dá Q1 como 4 e Q3 como 81. Muito puro. Esta é a propriedade metaphysical que anotou.

Método alternativo, como extraído pelo Ti-82 e pelo Ti-83

Se você desejar seus estudantes começar a mesma resposta que o Ti-82 e o Ti-83 calculadoras gráficas então esta é uma alternativa aceitável. Nota é não padronizada no mundo dos statistics.

Encontre o número médio. Encontre então o número médio dos valores dos dados cujos os Rank são MENOS DO QUE o Rank do número médio. Este será um valor dos dados ou será meia maneira entre dois valores dos dados.

Está aqui o mesmo exemplo usando os números acima, que contem *um número* ímpar dos valores:

35 47 48 50 51 53 54 70 75

Rache os dados em duas metades, *não including o número médio* :

35 47 48 50 e 53 54 70 75

Encontre o número médio de cada metade. Neste exemplo, o primeiro quartile é 47,5 e o terceiro quartile é 62. Daqui a escala interquartile é $62 - 47,5 = 14,5$.

Anote a diferença nas respostas entre os dois métodos! Não está surpreendendo realmente quando você considera que nós estamos fazendo um sumário de 5 números em um jogo de somente 9 números.

Extraindo as suíças

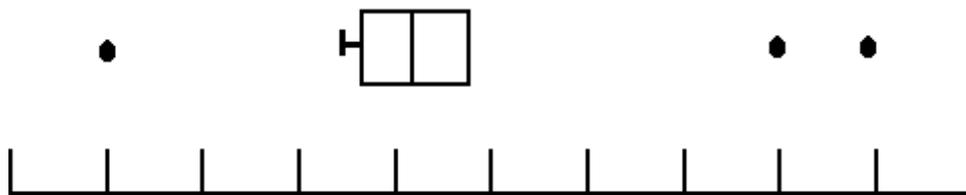
O método geralmente aceitado entre statisticians para extrair as suíças é um tanto mais complicado do que aquele descrito em a maioria de textos secundários sênior escritos aos syllabuses de Queensland. O comprimento máximo de cada suíça é 1,5 vezes a escala interquartile (IQR). Para extrair a suíça acima do quartile ^{de 3}rd ao, extraia-o o valor o maior dos dados que é menos do que ou igual ao valor que é 1,5 IQRs acima do quartile ^{de 3}rd. Todos os dados avaliam maior do que aquele deve ser marcado como um outlier. Alguns statisticians diferenciam-se entre outliers ' suaves ' e outliers ' severos '. Os outliers suaves encontram-se entre 1,5 e 3 IQRs acima do quartile ^{de 3}rd quando os outliers severos são mais de 3 IQRs acima do quartile ^{de 3}rd. Eu não acredito que nós necessitamos fazer a distinção entre outliers suaves e severos em nossos cursos.

Está aqui um exemplo, usando o primeiro jogo dos números acima de usar o método de Tukey de determinar Q1 e Q3:

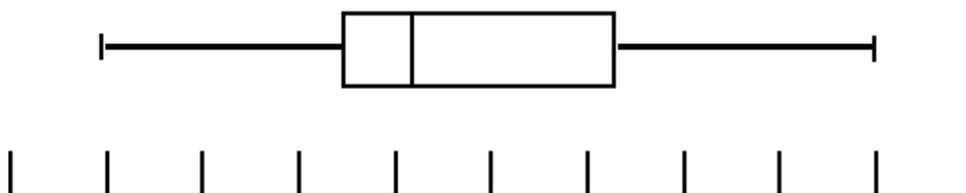
35 47 48 50 51 53 54 70 75

O IQR é 6. Agora 1,5 vezes 6 iguais 9. Este está a um comprimento máximo da suíça. Subtraia 9 do primeiro quartile: $48 - 9 = 39$. Anote que 35 são um outlier, e a suíça deve ser extraída a 47, que é o valor o menor que não é um outlier.

Adicione 9 ao terceiro quartile: $54 + 9 = 63$. Todo o valor maior de 63 é um outlier, assim que neste exemplo 70 e 75 estão os outliers. Extraia a suíça ao valor o maior na série de dados que não é um outlier, neste caso 54. Desde que este valor é o quartile ^{de 3}rd, nós não extraímos nenhuma suíça em tudo! Marque 70 e 75 como outliers. O boxplot é dado abaixo:



Se nós usarmos o método alternativo com o IQR de 14,5 então nós não temos nenhum outliers, e os olhares do boxplot como este:



A mensagem aqui pode ser: beware de construir boxplots em um jogo pequeno dos números!

Por que 1,5 IQRs?

"primeiramente de tudo, qualquer um sabe porque é habitual multiplicar $1.5 \cdot \text{IQR}$ a fim encontrar outliers? No detalhe, porque é 1,5 e não, para o exemplo, 2?"

De Paul Velleman

A resposta "oficial" de John Tukey (quando eu pedi) é: porque 1 é demasiado pequeno e 2 são demasiado grandes.

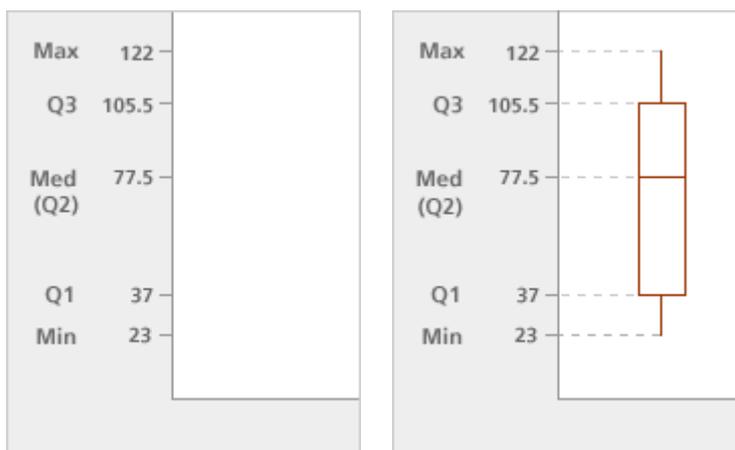
Há um papel Hoaglin e Ygelevich em que fizeram algumas simulações e mostraram por que a régua do outlier é realmente completamente boa através de uma disposição larga bonita das distribuições. Eu não sei a referência cheia.

Um artigo agradável do exercício ou da avaliação para estudantes da High School deve mostrar que para uma distribuição gaussian (IE um que seja distribuída normalmente), a régua $1.5 \cdot \text{IQR}$ iguala a aproximadamente 1 valor dos dados em 100 que estão sendo classificados como um outlier. Encontre uma solução trabalhada [Boxplots e a distribuição normal](#).

Lote Da Caixa

Nesta Parte: [Sumário do Cinco-Número com dados da medida](#) | Extraíndo um lote da caixa

Uma vez que nós temos o sumário do Cinco-Número, nós podemos indicá-lo que usa um tipo do gráfico sabido como um lote da caixa. Está aqui o lote da caixa para os dados que do noodle nós nos temos usado:



O lote da caixa é chamado também um lote das caixa-e-suiças. Embora olha muito diferente dos gráficos precedentes, é justo uma outra maneira representar a distribuição dos dados que nós temos trabalhado com tudo longitudinalmente:

- A suíça mais baixa estende do mínimo a Q1. O comprimento desta suíça indica a escala (ou, neste caso, no mais curto) do fourth o mais baixo dos dados requisitados.
- A suíça superior estende de Q3 ao máximo. O comprimento desta suíça indica a escala (ou, neste caso, no mais longo) do fourth o mais elevado dos dados requisitados.
- A caixa (a parcela retangular do gráfico) estende de Q1 a Q3, com um segmento de linha horizontal que indica Med.
- A parcela do retângulo entre Q1 e Med indica a escala do segundo fourth dos dados requisitados.
- A parcela do retângulo entre Med e Q3 indica a escala do terceiro fourth dos dados requisitados.
- O retângulo inteiro indica a escala da metade média (a escala interquartile) dos dados requisitados.

Anote que os lotes da caixa podem ser extraídos verticalmente ou horizontalmente, dependendo de se você indica o sumário do Cinco-Número ao longo de uma linha central vertical ou horizontal. Nota 2

Segmento Video

Neste segmento video, o professor Kader introduz o processo de construir um lote da caixa. Preste atenção a este segmento para rever o processo ou para ajudá-lo a extrair os lotes da caixa para o seguinte problema.

Nota: A série de dados de usada pelo onscreen participants é diferente de essa fornecida acima.

Se você estiver usando um VCR, você pode encontrar este segmento no vídeo da sessão aproximadamente 15 minutos e 10 segundos após o logo de Annenberg/CPB.

Vamos comparar nossos dados do noodle como representados pelo sumário do Cinco-five-Noodle, pelo sumário do Cinco-Número, e pelo lote da caixa:

Esta atividade requer o encaixe flash, que você pode download para livre do Web site de Macromedia. Para uma versão non-interactive desta ilustração que não requer o encaixe flash, reveja a seqüência das ilustrações na página precedente e nesta página, para seguir a progressão dos noodles através do lote da caixa.