Uma ferramenta estatística para o futuro

Por Hindemburg Melão Jr.
http://www.sigmasociety.com

Quando eu tinha cerca de 15 anos, e tive meus primeiros contatos com as distribuições de QIs, ocorreram-me algumas idéias que na época eu não tinha como confirmar nem como me aprofundar.

Uma das idéias que tive é que algumas amostras poderiam acidentalmente conter elementos muito destoantes e com isso distorcer os parâmetros da distribuição. Por exemplo: numa classe com 50 alunos em que eu estivesse presente, a incidência de QIs acima de determinado valor seria diferente do esperado para uma amostra desse tamanho, e isso distorceria a posição da média e o tamanho do desvio-padrão, além de introduzir muita assimetria e curtose. Eu não conhecia a expressão “curtose”, mas a compreendia, bem como o conceito de assimetria, que é bastante intuitivo. Se um grupo de ganhadores de prêmio Nobel estivesse passeando pela praça da Sé, e um examinador escolhesse arbitrariamente 300 pessoas que estivessem passando pela praça para medir seus QIs, também haveria um efeito desse tipo. Para que o cálculo pudesse ser feito de modo a estimar corretamente a média e o desvio-padrão naquela população, seria necessário desconsiderar ou atenuar o efeito causado por aquelas presenças inusitadas. Em Xadrez havia um critério de desempate chamado “milésimos medianos”, que consistia basicamente numa média recortada (trimmed mean), eliminando algo como 10% dos resultados mais altos e 10% dos mais baixos, que era um jeito simples e meio porco de corrigir aquele tipo de efeito. Eu conhecia o sistema de desempate de milésimos medianos, mas não me parecia ser a solução mais apropriada. A exclusão pura e simples do(s) elemento(s) destoante(s), antes de fazer o cálculo dos parâmetros da distribuição, e a exclusão de um elemento aleatoriamente escolhido da outra metade da distribuição (ou tantos quantos equivalessem aos que foram excluídos do outro lado), poderia ser uma forma de contrabalançar esse efeito de maneira mais limpa, ou ainda balancear o peso de cada elemento em função da probabilidade de cada elemento fazer parte daquele grupo. Muitos anos depois eu soube da existência de Estatística Robusta, cujo propósito é exatamente esse.

Outra percepção que tive, e que de certo modo decorre da anterior, é que em vez de calcular os parâmetros da distribuição (média e desvio-padrão) com uso das fórmulas convencionais, seria mais acurado testar diferentes valores para aqueles parâmetros, até encontrar a curva que mais se assemelhasse visualmente à curva empírica. Muitos anos depois, eu descobri que isso existe e se chama “estimação por máxima verossimilhança”, e que a “comparação visual entre as formas das curvas” é feita objetivamente por testes de aderência.

Eu também tinha o entendimento de que se alguns elementos de uma amostra destoavam muito daquela amostra, e se havia inexatidões nas medidas, então se o tal elemento ou os tais elementos estivessem muito acima da média amostral, poderia indicar que na verdade todos os outros elementos da amostra deveriam ter valores um pouco maiores (o valor verdadeiro deveria ser um pouco maior que o valor aferido), para acompanhar o fato de existir o tal elemento, mas não tão maiores quanto se o cálculo dos parâmetros da distribuição fossem feitos à maneira convencional e incluindo o elemento como parte “normal” da amostra. A idéia de balancear a presença de outliers, em vez de os eliminar, me parecia o procedimento mais correto, e este balanceamento precisava ser bilateral para que fizesse sentido, isto é: se a amostra inteira tem sua média arrastada para cima devido a presença de um outlier muito alto, então o valor do outlier também precisa ser arrastado para baixo. Isso implicaria uma incerteza assimétrica em todas as medidas, especialmente nos elementos mais afastados da média e mais ainda nos outliers. Essa percepção de relação mútua entre como as probabilidades de

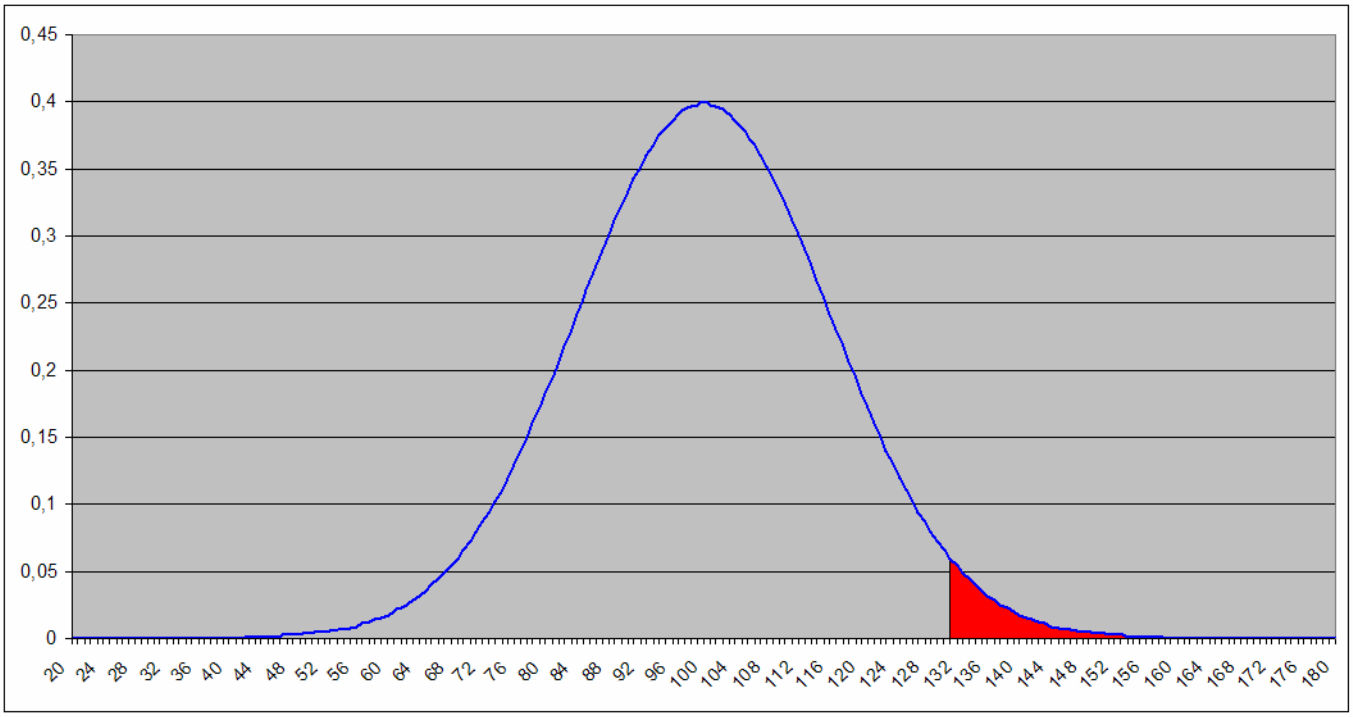
cada elemento ter assumido determinado valor influenciava nos valores de todos os demais elementos individualmente, bem como nos parâmetros da amostra como um todo, eu cheguei a aplicar em muitas ocasiões, inclusive em 2003, em meu artigo sobre como calcular paralaxes estelares, em que proponho que a probabilidade de uma estrela apresentar determinada luminosidade absoluta precisa ser considerada conjuntamente com a probabilidade de a distância dessa estrela estar situada em determinado limite fiduciário, caso contrário se pode fazer superestimativas ou subestimativas. Alguns anos depois, eu soube que isso se chama “Estatística Bayesiana”.

Em minhas tentativas de normatização do Sigma Test, concebi vários conceitos que são usados em Teoria de Resposta ao Item (TRI) e outros superiores aos usados em TRI, além de posteriormente ter ampliado e aprimorado alguns conceitos de TRI quando passei a conhecê-la.

Se aos 15 anos de idade eu dispusesse de conhecimentos matemáticos e instrumentos apropriados, ou mesmo que não conhecesse a matemática necessária, mas se eu tivesse computadores naquela época, e alguma motivação para dedicar tempo a isso, é provável que eu redescobrisse alguns fundamentos da Estatística Robusta, da Estatística Bayesiana, alguns métodos para estimação por máxima verossimilhança e recriasse a Teoria de Resposta ao Item.

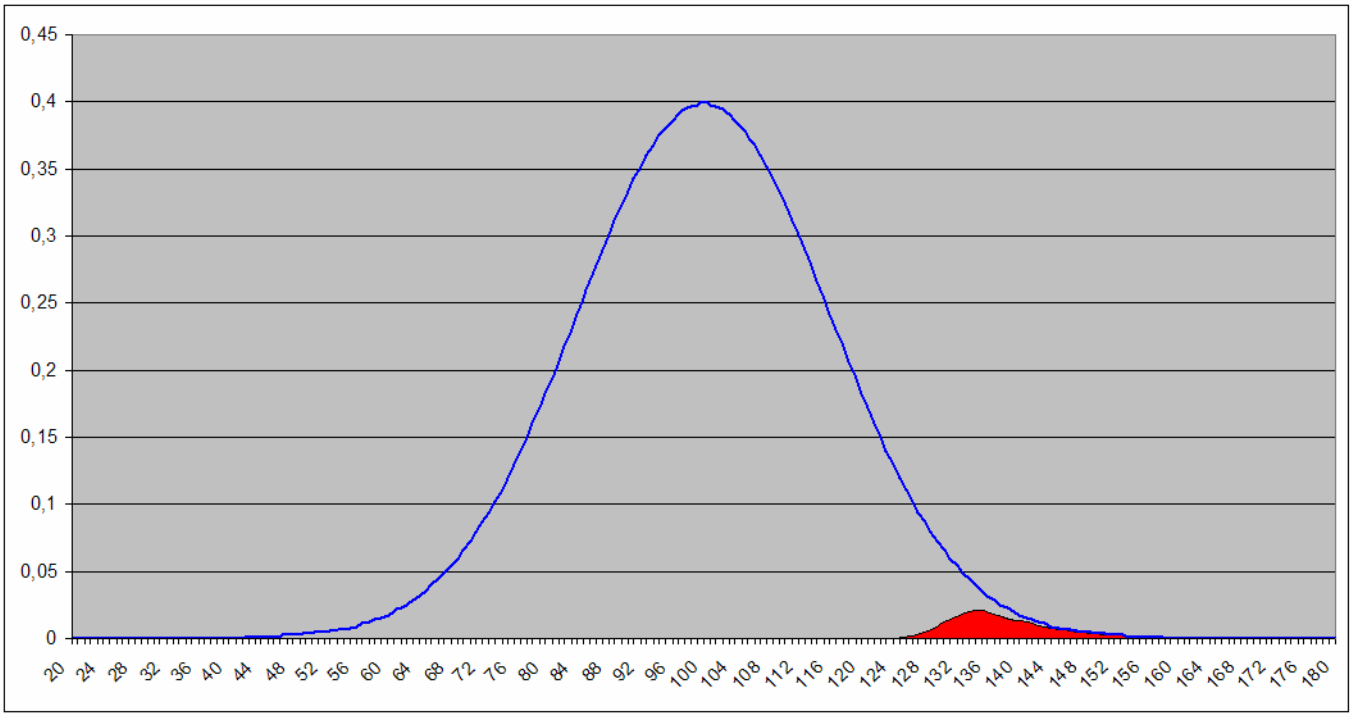
Uma das idéias que tive naquela época, mas que parece ainda não existir algo equivalente, e que pode ser uma ferramenta importante no futuro, talvez nos próximos séculos, é a determinação dos parâmetros de uma função que represente a distribuição de uma população com base numa amostra dessa população que esteja concentrada numa das caudas da distribuição populacional, portanto com parâmetros muito diferentes daqueles da própria população. Exemplo: Com base na medida dos QIs de todas as pessoas do MIT, calcular o QI médio da população dos EUA. O QI médio no MIT é cerca de 144 e o QI médio nos EUA é cerca de 98 (de acordo com Richard Lynn). Se não fosse conhecido, de antemão, o QI médio nos EUA, como calcular este valor com base apenas na medida dos QIs do MIT?

O QI médio dos membros da Mensa é cerca de 138,4. O ponto de corte é 133. A distribuição dos QIs medidos no exame para admissão gera escores truncados em 133, porém se as pessoas aprovadas forem examinadas com outro teste equivalente, provavelmente será mantido o escore médio 138,4, com a diferença que muitas terão escores abaixo de 133. Os dois gráficos a seguir resumem a situação:



No gráfico acima, a linha azul representa a distribuição teórica de QIs na população em geral, e a região em vermelho representa a parte da população aprovada na Mensa nos próprios testes usados para seleção. Como o ponto de corte é 133, nenhum escore pode ser abaixo desse valor e a curva é quase perfeitamente aderente à parte correspondente da distribuição da população inteira.

Mas se for aplicado outro teste nas pessoas que foram aprovadas, o resultado provável é como no próximo gráfico:



Em que a média continua sendo 138,4, mas a forma da curva muda completamente e alguns escores podem ficar abaixo do ponto de corte. [Nota: a curva vermelha real tem forma quantitativamente diferente desta desenhada no gráfico, mas essencialmente conserva as mesmas propriedades desta representada acima]

A área vermelha é uma distribuição que apresenta maior aderência à curva azul à medida que se afasta da tendência central (forte correlação positiva entre aderência local e distância à tendência central) numa das metades da distribuição (na cauda direita, nesse caso) e o oposto na outra metade, ou seja, forte correlação negativa entre aderência local e distância à tendência central. Com base na variação da dissimilaridade entre a função vermelha e a azul em relação à distância da tendência central até o ponto considerado, pode-se definir um método para reconstruir a curva azul com base nos parâmetros da curva vermelha. Porém a curva azul, a priori, não é conhecida. Nesse caso há duas maneiras de lidar com o problema. A mais fácil é quando se conhece os escores de pelo menos alguns elementos da população geral, por meio dos quais se pode definir aproximadamente a média da população e assim posicionar a amostra conhecida em relação à população. Mas suponhamos que não se sabe nada, nem a média aproximada da população, então o cálculo é feito com base exclusivamente nos parâmetros da amostra e partindo de algumas suposições:

1) A distribuição da população é normal (mesocúrtica e simétrica).
2) Os efeitos de ego-seleção não são significativos na amostra.
3) A assimetria e a curtose na amostra são causadas pelo distanciamento da tendência central e são proporcionais à diferença entre a média da amostra e a média da população:
    a. Quanto maior a distância entre a média da amostra e a média da população, tanto maior é a assimetria na amostra.

b. Quanto maior a distância entre a média da amostra e a média da população, tanto menor é a curtose na amostra na direção da média da população e maior é a curtose na amostra na direção oposta à média da população. Isso gera um novo parâmetro a ser considerado: a “assimetria da curtose”.

A determinação quantitativa destas propriedades para diferentes posições da média amostral em relação à média populacional, possibilitam obter uma função (por regressão) que represente os parâmetros da população para quaisquer parâmetros de uma amostra, lembrando que a amostra deve, preferencialmente, contar com um quinto parâmetro (assimetria da curtose) para possibilitar o cálculo dos 2 parâmetros da população (média e desvio-padrão).

As informações necessárias para que se possa calcular os parâmetros da população são:

- Correlação entre assimetria da amostra e distância entre a média da amostra e média da população.
- Correlação entre aderência local da amostra à população e distância entre a média da amostra e média da população.
- Correlação entre curtose direita da amostra e distância entre a média da amostra e média da população.
- Correlação entre curtose esquerda da amostra e distância entre a média da amostra e média da população.
- Correlação entre assimetria da curtose da amostra e distância entre a média da amostra e média da população.
- Correlação entre desvio-padrão da amostra e distância entre a média da amostra e média da população.
- Correlação entre desvio-padrão direito da amostra e distância entre a média da amostra e média da população.
- Correlação entre desvio-padrão esquerdo da amostra e distância entre a média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação da assimetria da amostra em função da distância entre a média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação da aderência local da amostra em função da distância entre a média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação da curtose direita da amostra em função da média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação da curtose esquerda da amostra em função da média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação da assimetria na curtose da amostra em função da média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação do desvio-padrão direito da amostra em função da média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação do desvio-padrão esquerdo da amostra em função da média da amostra e média da população.
- Parâmetros para uma curva suave que represente a variação do desvio-padrão da amostra em função da média da amostra e média da população.

Pode-se ampliar essa lista, usando diferentes curvas suaves para representar as variações em função da distância entre as médias. Mas, preliminarmente, o método já pode atender a alguns propósitos básicos usando apenas um tipo de modelagem. Munido com estes resultados, pode-se calcular os 2 parâmetros (média e desvio-padrão) da população de 8 maneiras diferentes, bem como as incertezas. Ainda que cada um dos resultados seja uma estimativa grosseira, o conjunto global de resultados pode fornecer valores razoavelmente precisos para os parâmetros que se deseja conhecer, por isso é recomendável o uso de diferentes parâmetros da amostra para calcular repetidamente os mesmos parâmetros da população.

Depois de realizados estes cálculos com amostras e populações cujos parâmetros sejam conhecidos (isso é feito apenas uma vez), pode-se usar as relações encontradas sempre que for necessário para se calcular os parâmetros desconhecidos de outras populações, com base nos parâmetros de amostras extremas destas populações.

Quando se sabe, a priori, que a distribuição da população não é normal, e se sabe qual distribuição apresenta melhor qualidade de ajuste aos dados empíricos da população, pode-se fazer todo o processo equivalente usando outra distribuição em lugar da normal. Quando não se sabe qual é a distribuição que melhor representa a população, é recomendável adotar como hipótese que a distribuição seja normal.

Para que estes procedimentos produzam resultados aceitáveis, é importante que a distribuição da população seja muito aderente a uma gaussiana, inclusive e especialmente na cauda em que está a amostra considerada. Também é importante que a amostra seja suficientemente numerosa para que os 5 parâmetros possam ser calculados com boa precisão.

Desse modo, seria possível usar os escores de testes aplicados no MIT para se calcular a distribuição de QIs no país inteiro. A incerteza no resultado seria grande, porém possibilitaria encontrar uma solução para um problema que de outro modo permaneceria insolúvel.

Algumas aplicações que este método pode ter são:

Conhecendo a distribuição de freqüências de linfócitos numa amostra local de sangue, isto é, quantos linfócitos por mililitro se tem em 200 ml, divididos em 200 amostras de 1 ml cada, sendo toda a amostra retirada da mesma região do corpo e na mesma data, portanto uma amostra que provavelmente não é representativa da distribuição de toda a população de linfócitos em outras partes do corpo, por meio de nosso método é possível estimar a distribuição de linfócitos em todas as partes do corpo.

Conhecendo as propriedades de peças defeituosas, sem saber as propriedades de peças boas, pode-se estimar as propriedades das peças boas. Aplica-se à praticamente quaisquer peças de qualquer segmento industrial.

De modo geral: conhecendo as propriedades de outliers ou de apenas parte da cauda de uma amostra, é possível calcular as propriedades de toda a população da qual aquela amostra foi extraída, mesmo os parâmetros da amostra sendo muito diferentes dos parâmetros da população.

O uso de Bootstrap, Jackknife etc., podem melhorar os resultados.

O uso de correlações entre os logaritmos das medidas pode ser mais apropriado quando se está próximo às extremidades das caudas. Há outros detalhes que podem refinar os resultados.

No Mercado Financeiro, esta ferramenta é extremamente útil para calibrar parâmetros de sistemas automáticos, de modo a incorporar eventos raros, como o crash de 2008, e determinar propriedades “normais” para o Mercado, entre outras aplicações. Quando se conhece as propriedades tanto da população quanto de uma amostra concentrada numa das caudas, ou se conhece uma amostra representativa da população e outra concentrada numa das caudas, pode-se fazer modelagens muito mais úteis do que por métodos convencionais.

Até onde sei, não existe algo similar, e em caso afirmativo, parece-me que um nome apropriado a este processo seja: ***“Estimação porca de parâmetros populacionais com base em amostras descentralizadas”*** ou EPPPAD. A estimação é “porca” porque não segue um caminho elegante, quase todo o método é acoxambrado, porém é extremamente efetivo.

Perigos na aplicação indiscriminada desses procedimentos:

Se uma população apresenta distribuição não-normal, uma amostra representativa dessa população pode induzir a estimativas incorretas dos parâmetros populacionais. Para que estes procedimentos sejam utilizados com legitimidade é necessário que se tenha motivos para supor que a amostra coletada esteja situada num posição afastada da média populacional, ou com franca predominância numa das caudas da população.