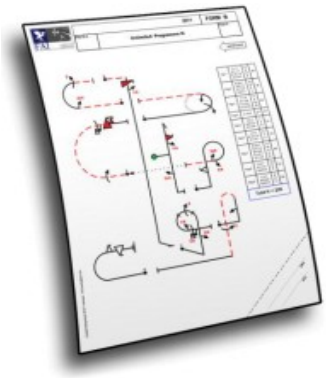
Processamento de notas dos Juizes e o Sistema Fair Play da CIVA (FPS)

Uma revisão completa de por que um “sistema” é necessário no julgamento da competição de acrobacia aérea, e o que o FPS faz para nós



Resultados na competição e Sistemas de Julgamento

Em muitos esportes competitivos, selecionar o ganhador é fácil... vai ser o primeiro carro a passar a linha de chegada, ou o time de futebol que marcar mais gols, e assim por diante. Entretanto, alguns esportes requerem juízes experientes para avaliar as habilidades técnicas e artísticas de uma demonstração, e a competição de acrobacia aérea é uma das muitas atividades que requer um especialista treinado para dizer o quão foi atingido o padrão técnico desejável. Onde tantos julgamentos complicados são necessários, o correto é assumir que inicialmente a performance pode ser teoricamente perfeita, e partindo daí, devemos simplesmente descontar o números de erros cometidos. O vencedor é claro, é aquele que tiver a maior pontuação remanescente após a subtração dos erros cometidos.

Um infeliz aspecto nesse processo de subtração de notas, é que a variação de conhecimento dos juízes pode trazer um efeito reverso. Um juíz menos experiente ou mais tímido tende a reconhecer menos erros e consequentemente pode conceder maiores notas com uma variação bem pequena, e isto vai influenciar mais o resultado do que um juíz mais experiente que deverá ver mais downgrades, dando mais notas baixas e com uma variação mais ampla. É também muito difícil para um juíz prevenir que algum entrosamento pessoal ou inimizade possa afetar suas decisões, seja isso aplicado conscientemente ou não. Em eventos internacionais, as influências de nacionalidade poem ser intrusivas e difíceis de se evitar.

Prática de julgamento de acrobacia aérea

Em eventos de acrobacia aérea, os juízes usam seus conhecimentos para acumular downgrades em cada figura para um valor múltiplo de 0,5, e depois subtrair este total de um “perfeito” dez, para dar a nota final que vai variar de 10 até 0,0, ou zero numérico.

Existe também, casos específicos de alguns erros técnicos, como quando um snap-roll, tail-slide, ou parafuso não apresentar as características de apresentação necessárias, então é assinalado um Perception Zero (PZ), e também, se a figura voada não é a especificada no documento dos juízes, então um Hard Zero (HZ) deve ser assinalado. O PZ é uma visão pessoal de cada juíz e precisa ser considerada como uma nota numérica, porém se algum juíz assinalou um HZ, então o juíz Chefe deve conferir com o painel de juízes e decidir se este HZ deve ser assinalado por todos, se possível, utilizando recurso de video , ou se este HZ deve ser rejeitado e alguma nota numérica atribuída. Para possíveis momentos de desconcentração, o juíz também pode dizer “Ops! Perdi aquela figura.” e assinalar um

“Avarage” (AV), no qual o sistema calcula e assinala uma média para aquela figura.

Estabelecendo diferenças de opiniões

Para humanos, a maneira mais comum de lidar com uma coleção de opiniões potencialmente não confiáveis, é provendo de o máximo de observações possíveis e fazer uma média delas para minimizar alguma influência de um elemento. Esta é uma estratégia válida se considerarmos uma ocasional discrepância que alguns juízes questionáveis vão causar. Entretanto as notas finais dos pilotos do topo da tabela podem apresentar uma diferença muito pequena, e aceitar todas as notas sem questionar pode levar a publicação do resultado errado. Deveria ter uma maneira melhor de identificar notas que simplesmente não “se encaixam” para que elas tenham a atenção que merecem, e com o FPS isto certamente é possível.

Combinando isto em um plano

Todas as notas dadas pelos juízes são transferidas para o computador. O que precisamos agora é:
. Um sistema de preparação que revele os efeitos e diferenças no estilo e capacidade de julgamento.

. Uma maneira de detectar notas “incomuns” quando comparadas com as notas de outros juízes sobre a mesma figura.
. Um teste prático no qual consiga-se avaliar notas incomuns como “OK” ou “Não OK”, e...
. Um método justo de substituição por uma nota mais adequada onde uma decisão “Não OK” for encontrada.
. Tudo isso deve ser feito de uma maneira “aberta” para que todos os juízes e pilotos saibam o que está realmente sendo feito, e com informações de suporte suficientes para que todos entendam porque as alterações foram feitas.

É claro – o computador não pode julgar! Mas pode fazer comparações muito inteligentes entre o que cada juíz diz e, assumindo que a visão geral do painel de juízes é a “correta”, este pode analisar meticulosamente todo elemento e aplicar técnicas matemáticas para alcançar um resultado que trata cada nota de cada juíz de uma maneira justa e balanceada, e que necessariamente assegura que este sempre erra a favor do piloto.

Como computar os resultados

Por muitos anos foram utilizadas a média geral das notas e enfrentado seus problemas, e depois por alguns anos a CIVA usou uma solução estatística chamada TBLP na qual uma simples tabela de todos os pilotos/figuras/juízes foi usada para comparar todas as notas juntas, substituindo por médias as notas que não passavam no chamado teste SD. Com o TBLP, entretanto, todas as notas de todos os pilotos afetavam todas as outras notas, e enquanto

isto oferecia algumas vantagens, foi visto que juízes poderiam adaptar seu estilo de julgamento para ter um resultado artificial melhorado... e eventualmente a confiança dos pilotos e organizadores dos campeonatos foi diminuindo. Em vez de se arriscar a ter que voltar as antigas notas médias, a CIVA se propôs a criar uma nova solução.

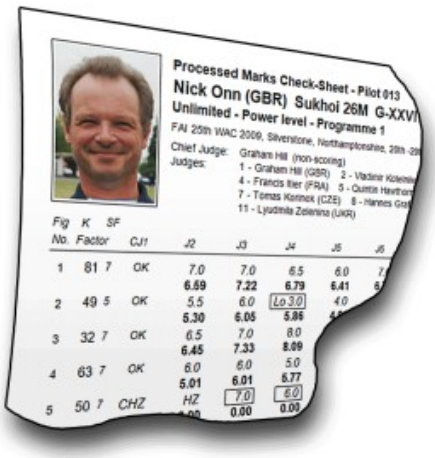
Sistema Fair Play da CIVA

O processo foi desenvolvido em 2005 com um contesto inteiramente novo que combina nossas experiências em julgamento de competição com um número de testes estatísticos robustos para encontrar um elevado padrão analítico desejado. O resultado provou ser um sistema de notas confiável que construiu um bom nível de credibilidade dentre os juízes e competidores.
O sistema funciona da seguinte maneira:

1. **Separar as notas dos juízes em grupos de figuras**
   Primeiro, o sistema monta as notas dadas pelos juízes em grupos figura-por-figura, para que que as comparações sejam feitas de diferentes opiniões sobre a mesma coisa. Para as sequencias Free e Free Unknown, nas quais as composições das figuras são mais flexíveis, um sistema de “Super Famílias” é usado para agrupar figuras de tipos similares junto, para garantir uma base sólida de comparação de julgamento.

2. **Balancear os juízes dentro de cada grupo de figuras.**
   Um essencial passo inicial para cada grupo é de rebalancear as notas dos juízes para que nenhum dos juízes tenha maior influência do que outro. A palavra estatística para este balanceamento é “normalização”, e sem isto, as comparações entre os juízes simplesmente não poderiam ser válidas. Em nossa normalização cada conjunto de notas “não-zero” do juiz é movido para cima ou para baixo e o desvio padrão das notas é estreitado ou alargado para que ambos sejam os mesmos da média do painel todo. Isto soluciona completamente o dilema do juiz experiente/inexperiente. A influência de cada juiz agora é a mesma. Esta é a alteração que muda as notas dos pilotos de um número múltiplo de 0,5 para um número com mais casas decimais.

3. **Identificar e solucionar notas incomuns**



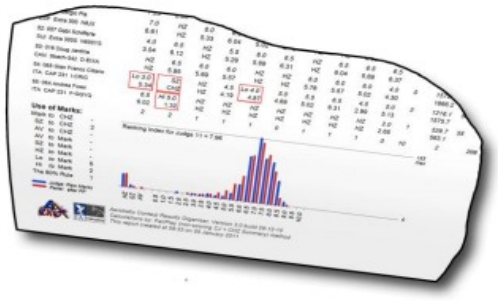
Para cada grupo de notas o FPS calcula uma tabela de notas de “Valor Ideal” de acordo com o estilo de julgamento de cada juiz. Um “teste estatístico de confiança” é executado agora, para avaliar a validade de cada nota normalizada com seu valor ideal correspondente. Se o teste encontra a confiança do FPS requerida então a nota é aceita e passada para o próximo estágio, mas se o teste falhar, então a nota original é assinalada “Rejeitada”. Assim toda nota normalizada vai ser assinalada “Aceita” ou

"Rejeitada". Quando este grupo inicial for processado, caso alguma nota seja assinalada "Rejeitada", então a normalização é reexecutada e os valores ideais recalculados desde o começo, mas é claro agora sem as notas consideradas rejeitadas. Estes novos valores ideais, sendo livres de toda influência das notas rejeitadas e corretamente correspondendo ao estilo de julgamento de cada juiz, são agora usados para substituir cada nota considerada rejeitada e também qualquer "Average" que tenha sido solicitada. Estas substituições ficam circuladas nas folhas de notas dos pilotos para mostrar aonde a correção foi feita. Este grupo final de notas pode então ser multiplicado pelo fator K correspondente da figura para construir uma nova tabela de notas de cada piloto.

4. **Identificar e mostrar qualquer nota tendenciosa alta ou baixa**
   O Sistema Fair Play agora utiliza a tabela acima de notas como base para uma nova normalização, Valor Ideal e processo de rejeição de notas muito parecido com o anterior executado. Entretanto, desta vez o processo é usado para detectar qualquer nota incomum que tenha sobrevivido, e o nível de confiança é um pouco mais tranquilo (90%).
   As notas tendenciosas são possíveis porque mesmo que as notas incomuns tenham sido removidas, um juiz pode ainda dar notas mais altas ou baixas de forma geral para um competidor, e o resultado pode ser inaceitavelmente alto ou baixo em comparação aos outros juizes. Esta tendência pode por exemplo vir do resultado do entusiasmo quanto a um piloto da casa, ou simplesmente afinidades de nacionalidade. O FPS substitui qualquer nota que falhe seu teste de confiança com o o Valor Ideal de cada juiz, e novamente cada mudança é claramente mostrada na folha de notas dos pilotos.

5. **Remover qualquer possível influência dos pilotos de notas baixas nos líderes**
   Como um último passo, é necessário garantir que os pilotos com notas baixas "difíceis de julgar" não possam influenciar o ranking dos pilotos no topo da tabela.Os pilotos que fizerem menos de 60% da pontuação na Known e Free e menos que 50% na Unknown, são agora temporariamente excluídos, e todo processo do FPS é rodado novamente desde o começo. Uma tabela de resultados pode agora ser construída com este novo cálculo das notas maiores combinado com os resultados anteriores para os pilotos de menores notas. Finalmente as penalizações são descontadas e os resultados da sequencia está pronto para ser publicado.

6. **Criar um feedback detalhado para os Juizes**



   Agora o Sistema Fair Play pode apresentar sua outra grande força – uma revisão completa da performance dos juizes. Uma análise individual mostra para cada juiz como ele se compara a seus colegas, enquanto para o Juiz Chefe as estatísticas de todo painel são montadas para mostrar qual juiz se aproximou mais do painel e por quanto um determinado juiz se distanciou da performance dos outros. Assim, o FPS pode oferecer um feedback de fácil análise para todo time de juizes, algo não disponível até a introdução deste sistema.

Publicação dos resultados

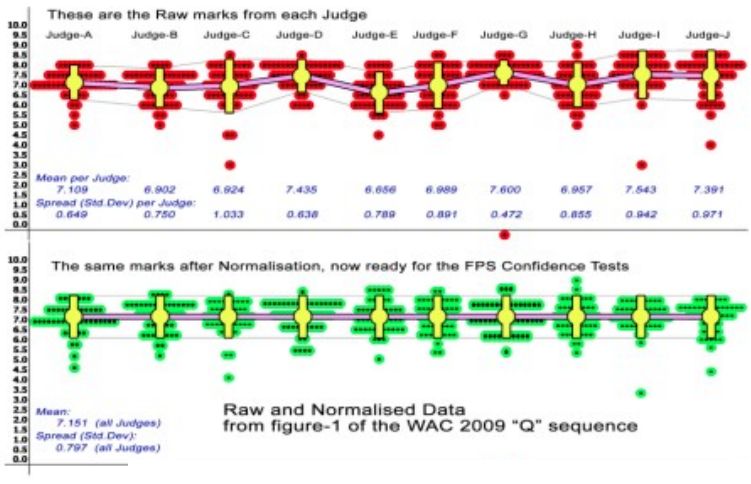
Após a aprovação do Juiz Chefe e do Juri, os resultados podem agora serem publicados no papel ou na web, e fazer com que a análise dos Juizes seja disponibilizada, e assim não só os pilotos mas também os juizes ficarão sabendo como foi sua própria performance na competição.

O Indice de Ranking dos Juizes

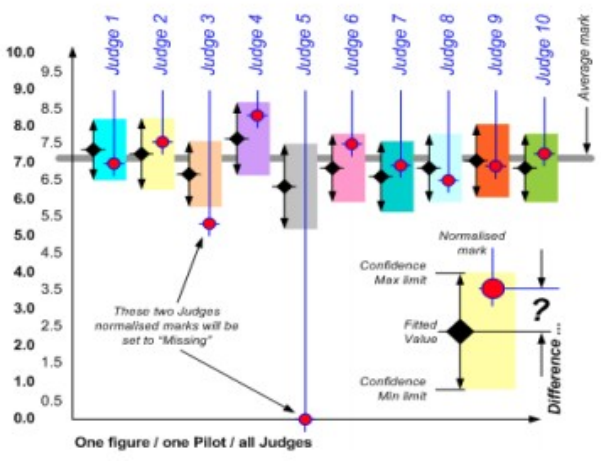
Em um mundo ideal cada juiz ranquearia os pilotos na mesma ordem do resultado final baseado nas visões de todo o painel. Embora pequenas diferenças, em geral, serem de pouca preocupação, alterações significativas no ranking do piloto comparado a conclusão final do painel, seria uma clara indicação de que esta visão do juiz não está sendo compartilhada e ela é menos provável de ser correta. Para medir este efeito o FPS determina o ranking de pilotos para cada juiz com um conjunto especialmente preparado de normalizações, levando em conta qualquer PZ rejeitado para o qual o juiz não é penalizado, e então cria um Indice de Ranking (RI) que será zero se o juiz estiver perfeitamente sintonia com o painel e que é aumentado caso alguma diferença de ranking e notas é combinado. Em uma competição, um valor de RI abaixo de 10 para cada sequencia indicaria um bom nível de concordância dos resultados publicados, e a medida que o número aumente disto, aumenta-se também o motivo para preocupação – uma revisão na análise das notas deste juiz se faz necessária para identificar onde as discrepâncias são encontradas.

Além da óbvia vantagem da facilidade com que qualquer juiz pode agora rever dua performance na competição em relação aos resultados publicados e ver aonde ele pode focar seu esforço para melhorar, a experiência mostra que este sistema pode agora ser usado com uma base confiável e sólida para a seleção de juizes para competições internacionais.

Um exemplo de normalização de notas



Primeiro diagrama
Cada ponto vermelho/preto representa uma nota dada por cada juiz com aquele valor. Os círculos amarelos mostram a média de cada juiz, as faixas verticais amarelas indicam o desvio padrão das notas dos juizes. As linhas rosas e cinzas enfatizam a diferença de estilo de cada juiz- alguns juizes dão notas maiores do que outros, e alguns juizes dão notas que variam mais do que os outros.
Segundo diagrama
Durante o processo de normalização, cada bloco de notas do juiz é movido para cima ou para baixo para que sua média de



notas seja igual a média geral de todo o painel, e o desvio padrão é comprimido ou expandido para que seja igual o desvio padrão médio do painel. Como agora os juizes tem um estilo de julgamento idêntico, é então possível iniciar a comparação de um juiz com o outro de uma maneira justa.

Como funciona o teste de confiança do FPS?

Considerando cada nota normalizada do grupo todo, o FPS executa um teste estatístico em cada uma para obter um valor de “Incerteza” para esta.
Isto é feito pegando a diferença entre a nota e o “Valor Ideal” que o FPS calculou e dividindo pelo desvio padrão do grupo. No diagrama acima, cada nota do juiz é mostrada como um círculo vermelho e o Valor Ideal como um diamante preto. A altura da flecha preta indica o intervalo de 97.5% de confiança dentro do qual nós devemos aceitar a nota. Qualquer nota que estiver acima ou abaixo deste intervalo é muito diferente do valor que esperávamos ser dado pelo juiz, e então não pode ser usado.

Se o resultado do teste de confiança exceder 2,24 então podemos dizer que a incerteza da nota é maior do que 97.5% e esta deve ser descartada. Para entender isto, olhe a distribuição normal das notas mostrado no diagrama abaixo. No FPS, as notas no centro da área de 97.5%, entre o desvio padrão de +/- 2,24, são aceitas, enquanto aquelas nos extremos esquerdo e direito (vermelhos) são os 2,5% que são mais diferentes dos outros e devem ser descartados.
Para as notas rejeitadas nas áreas vermelhas, as notas originais dos juizes são rejeitadas, e é substituída por um valor ideal calculado no próximo passo, o qual é agora livre de qualquer anomalia.