# A Test for Two Poisson Processes in the Presence of Background Events

Matthew Tom

Emmanuel College

June 8, 2006

# Overview

- Description of the original problem
- Features of the problem
- Methodology and formulas
- Test run A: non-conditioned inference
- Test run B: conditional inference
- Results and conclusions
- Final Thoughts

# The original problem

- Two arrays for detecting cosmic rays:
  - High-Resolution Fly’s Eye (HiRes), Utah
  - Akeno Giant Airshower Array (AGASA), Japan
- 1996: AGASA reports observing clusters of ultrahigh energy (over $4 \times 10^{19}$eV) particles.
- 1997-2003: HiRes detects similar clusters of rays.
- Are these particles coming from the same cosmic sources?

# The original problem (cont.)

- Some of the “events” detected by the arrays is just background noise.
- The hits on both detectors are an unknown mix of background events and “real” events.
- Simulations have yielded estimates for the expected number of background events at AGASA and HiRes.
- Are the sources of AGASA’s real events the same as the sources of HiRes’s?
- Do real events come into the HiRes and AGASA detectors at the same rate?

# Declaring variables

- Constants:
  - $T_1$: time that the HiRes detector is open
  - $T_2$: time that the AGASA detector is open
  - $\mu_{1f}$: expected number of background events at HiRes
  - $\mu_{2f}$: expected number of background events at AGASA
- Parameters:
  - $f_{1r}$: frequency of non-background events at HiRes
  - $f_{2r}$: frequency of non-background events at AGASA
  - $f_{0r}$: frequency of non-background events if $f_{1r} = f_{2r}$.
  - $T_1 f_{1r} + \mu_{1f}$: expected number of total events at HiRes
  - $T_2 f_{2r} + \mu_{2f}$: expected number of total events at AGASA
- Data:
  - $x_1$: number of events detected at HiRes
  - $x_2$: number of events detected at AGASA

# The model

- Model:
  - $x_1$ and $x_2$ are independent random variables
  - $x_1$ has a Poisson distribution with mean $T_1 f_{1r}+\mu_{1f}$.
  - $x_2$ has a Poisson distribution with mean $T_2 f_{2r}+\mu_{2f}$
- Hypotheses:
  - The null hypothesis is $H_0$: $f_{1r}=f_{2r}=f_{0r}$.
  - The alternative is two-sided.
- Under the alternative, $f_{1r}$ and $f_{2r}$ can be estimated separately.
  - If the number of events $x_1$ is less than the expected number of background events $\mu_{1f}$, then the MLE for $f_{1r}$ is 0.
  - If the number of events $x_2$ is less than the expected number of background events $\mu_{2f}$, then the MLE for $f_{2r}$ is 0.
  - The MLE for $f_{1r}$ is the maximum of 0 and $(x_1- \mu_{1f})/T_1$.
  - The MLE for $f_{2r}$ is the maximum of 0 and $(x_2- \mu_{2f})/T_2$.

# Point estimation

- To get the MLE for $f_{0r}$ under the null, we maximize the joint Poisson likelihood:

$$L_0(x_1, x_2, f_{0r}) = e^{-(T_1+T_2)f_{0r}-\mu_{1f}-\mu_{2f}} \frac{(T_1 f_{0r} + \mu_{1f})^{x_1}}{x_1!} \frac{(T_2 f_{0r} + \mu_{2f})^{x_2}}{x_2!} \quad \text{and}$$

$$\begin{aligned} \ell_0(x_1, x_2, f_{0r}) = &-(T_1 + T_2) f_{0r} - \mu_{1f} - \mu_{2f} \\ &+ x_1 \log(T_1 f_{0r} + \mu_{1f}) + x_2 \log(T_2 f_{0r} + \mu_{2f}) - \log(x_1! x_2!)\,. \end{aligned}$$

- The MLE for $f_{0r}$ is the maximum of 0 and

$$\begin{aligned} \hat{f}_{0r} = &\frac{1}{2}\frac{x_1 + x_2}{T_1 + T_2} - \frac{1}{2}\frac{\mu_{2f}T_1 + \mu_{1f}T_2}{T_1 T_2} \\ &+ \frac{\sqrt{\begin{array}{r}((T_1+T_2)(\mu_{1f}T_1 + \mu_{2f}T_2) - (x_1+x_2)T_1T_2)^2 \\ +4(T_1+T_2)T_1T_2\,(x_1T_1\mu_{2f} + x_2T_2\mu_{1f} - (T_1+T_2)\mu_{1f}\mu_{2f})\end{array}}}{2T_1T_2(T_1+T_2)} \end{aligned}$$

# Looking at the formulas

- Three lines partition the outcome space into 6 zones.
  - $f_{1r}=0$ if $x_1 < \mu_{1f}$.
  - $f_{2r}=0$ if $x_2 < \mu_{2f}$.
  - $f_{0r}=0$ if $T_1\mu_{2f}x_1 + T_2\mu_{1f}x_2 < (T_1+T_2)\mu_{1f}\mu_{2f}$.
  - The three lines meet at a single point.
- The estimate for $f_{0r}$ is $(x_1+x_2)/(T_2+T_2)$ minus a correction term for the interference of $\mu_{1f}$ and $\mu_{2f}$.
- If $\mu_{1f}=0$ and $\mu_{2f}=0$, if there is no background noise, then we get exactly $(x_1+x_2)/(T_2+T_2)$.

# The likelihood ratio statistic

- Let $l_0(x_1,x_2,f_{0r})$ be the log-likelihood under the alternative. Let $l_A(x_1,x_2,f_{1r},f_{2r})$ be the log-likelihood under the alternative. The likelihood ratio statistic is

$$
\begin{aligned}
LR &= 2\left(\ell_A\left(x_1,x_2,\hat{f}_{1\mathrm{r}},\hat{f}_{2\mathrm{r}}\right)-\ell_0\left(x_1,x_2,\hat{f}_{0\mathrm{r}}\right)\right)\\
&= 2\begin{pmatrix}
-T_1\hat{f}_{1\mathrm{r}}-T_2\hat{f}_{2\mathrm{r}}-\mu_{1\mathrm{f}}-\mu_{2\mathrm{f}} \\
\qquad +x_1\log\left(T_1\hat{f}_{1\mathrm{r}}+\mu_{1\mathrm{f}}\right)+x_2\log\left(T_2\hat{f}_{2\mathrm{r}}+\mu_{2\mathrm{f}}\right)-\log\left(x_1!x_2!\right) \\
+(T_1+T_2)\hat{f}_{0\mathrm{r}}+\mu_{1\mathrm{f}}+\mu_{2\mathrm{f}} \\
\qquad -x_1\log\left(T_1\hat{f}_{0\mathrm{r}}+\mu_{1\mathrm{f}}\right)-x_2\log\left(T_2\hat{f}_{0\mathrm{r}}+\mu_{2\mathrm{f}}\right)+\log\left(x_1!x_2!\right)
\end{pmatrix}\\
&= 2\begin{pmatrix}
x_1\log\left(\frac{T_1\hat{f}_{1\mathrm{r}}+\mu_{1\mathrm{f}}}{T_1\hat{f}_{0\mathrm{r}}+\mu_{1\mathrm{f}}}\right)-T_1\hat{f}_{1\mathrm{r}}-T_2\hat{f}_{2\mathrm{r}} \\
+x_2\log\left(\frac{T_2\hat{f}_{2\mathrm{r}}+\mu_{2\mathrm{f}}}{T_2\hat{f}_{0\mathrm{r}}+\mu_{2\mathrm{f}}}\right)+(T_1+T_2)\hat{f}_{0\mathrm{r}}
\end{pmatrix}
\end{aligned}
$$

# Test data set

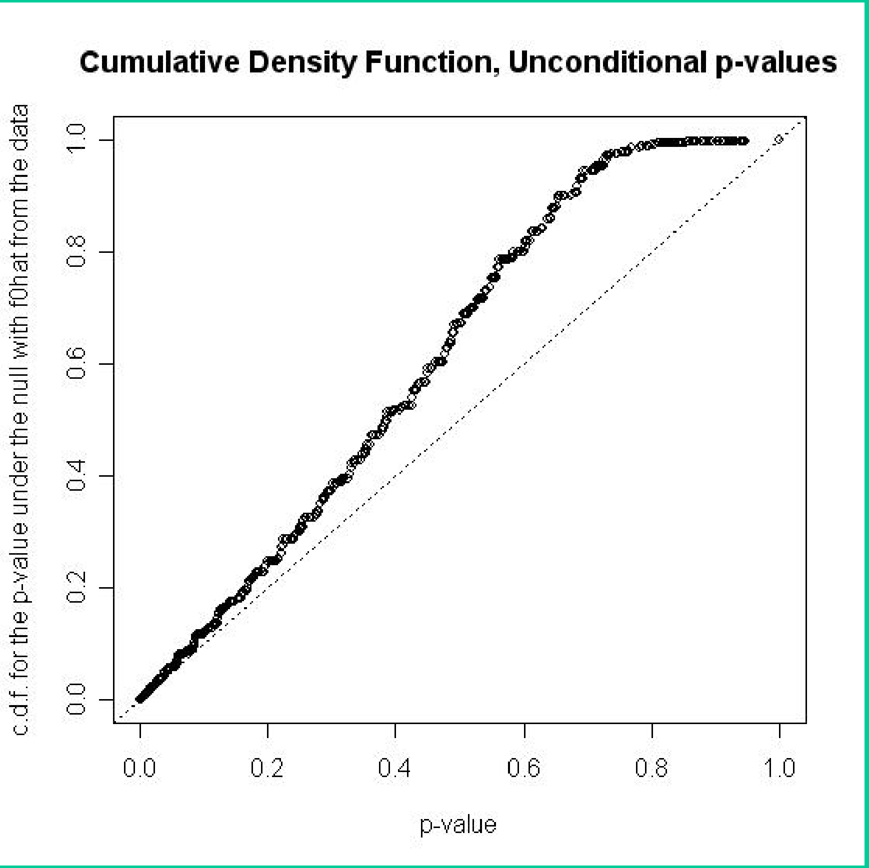
- These numbers were based on conversations with Professor Belz in the Department of Physics at the University of Montana.
- The difference in the time exposures cancel out the difference in the sizes of the arrays, so we can let $T_1=T_2=1$.
- According to simulations, $\mu_{1f}=3.6$ and $\mu_{2f}=6.4$.
- There were $x_1=6$ events at HiRes and $x_2=13$ events at AGASA.
- The MLEs are $f_{1r}=2.4$, $f_{2r}=6.6$, and $f_{0r}=6.923737$.
- The likelihood ratio statistic is LR=2.312918.

# Test Run A: unconditional inference

- Each possible data pair $(x_1,x_2)$ has an $f_0$ and a $LR(x_1,x_2)$.
- We'd like to use these $LR(x_1,x_2)$ to obtain a p-value.
- Let $f_{0rdata}$ be the data pair's estimate for $f_{0r}$. Let $LR_{data}$ be the data pair's likelihood ratio statistic.
- We use the distribution for $(x_1,x_2)$ over the whole quarter-plane.
- Our p-value is the sum of the probabilities of all the $(x_1,x_2)$ with higher LR's: $\Sigma_{LR(x1,x2)\geq LRdata}P(x_1,x_2|f_{0rdata})$.
- With the test data, the p-value is 0.2218632.
- Each possible data set has its own $f_0$, so each data set gets its own p-value based on its own distribution under the null.
- If the null is true, then the distribution for the p-value should be Uniform(0,1), plus or minus the discrete nature of the model.

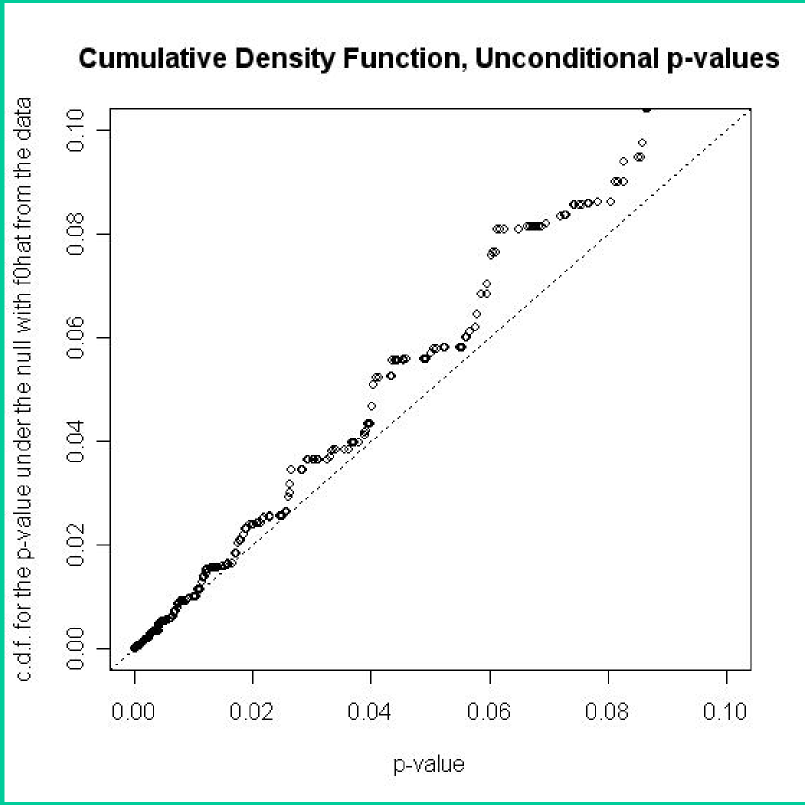
# The cumulative density function

- This is the distribution of the p-value if the null is true and $f_{0r}$ is the data's $f_{0r}$.
- For any $\alpha$, the probability of rejecting is always greater than $\alpha$.
- We expect some of this, since P(LR=0)>0, but this is bad.



# Focusing on practical α

- We can see the discretization.
- Data pairs that are CLEARLY not significant are still not significant.

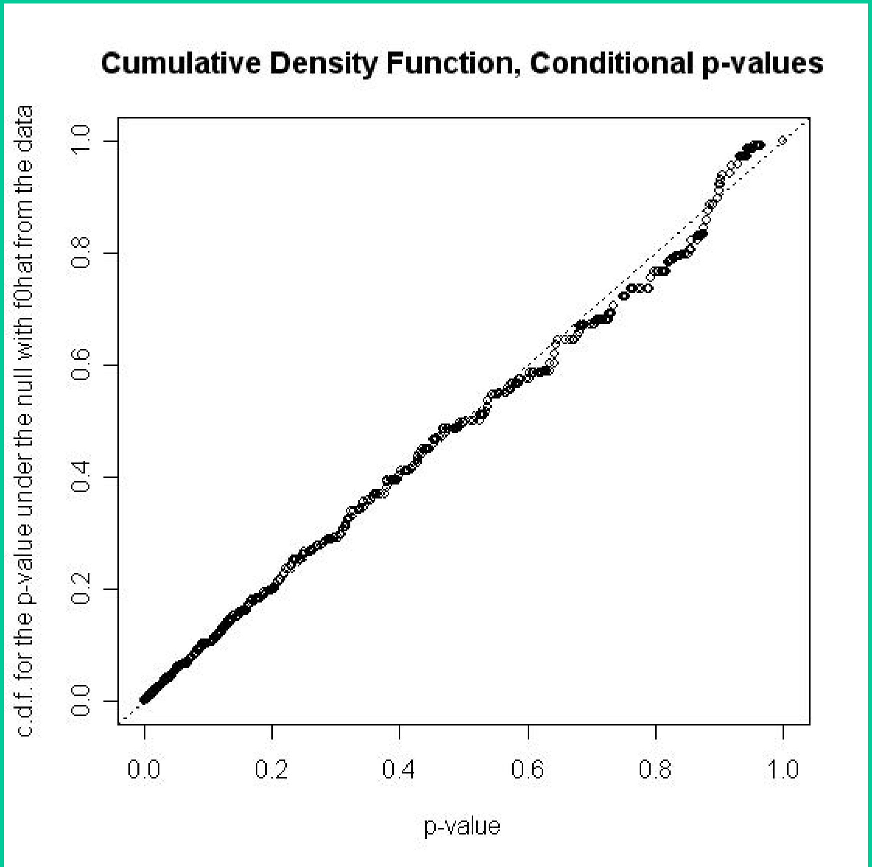


# Test Runs B: conditional inference

- This time, we're going to try conditioning on $f_{0r}$.
- We can't limit ourselves to data pairs with the exact same $f_{0r}$ as our data's. There won't be enough other data pairs. There might not be any.
- We can try choosing a percentage w, and then limit ourselves to $(x_1,x_2)$ with $f_{0r}$'s within $wf_{0rdata}$ of $f_{0rdata}$.
- For example, if we use w=10%, then we're conditioning on $f_0$ being between 90% and 110% of $f_{0rdata}$.
- Our p-value is: $\Sigma_{LR(x1,x2)\geq LRdata,|f0r(x1,x2)-f0data|<wf0data}P(x_1,x_2|f_{0rdata})$ / $\Sigma_{|f0r(x1,x2)-f0rdata|<wf0rdata}P(x_1,x_2|f_{0rdata})$.
- With the test data, the p-value is 0.2338813.
- Again, each possible data set has its own $f_{0r}$, so each data set gets its own p-value based on its own distribution under the null.

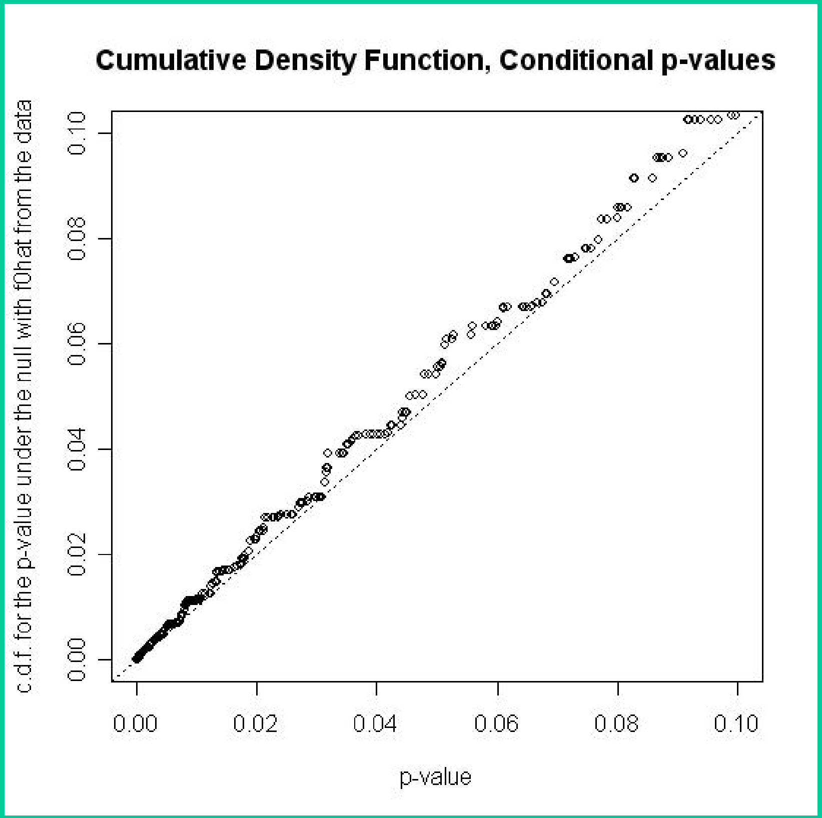
# The cumulative density function

- This is the distribution of the p-value if the null is true and $f_{0r}$ is the data's $f_{0r}$.
- The c.d.f. for this p-value is a lot closer to what it should be.



# Focusing on practical α

- We can still see the discretization, but it's not as bad.
- Even though $f_{0r}$ is not a sufficient statistic, conditioning still seems to work.



# Conclusions

- Maximum likelihood seemed despite the background events.
- Conditional inference looked more consistent than unconditional inference.
- When designing an ad hoc test statistic, check the distribution of the resulting p-value.

# Future work

- Here the conditioning window grew as $f_0$ grew. Do I do better or different with a fixed window width?
- The MLEs for $f_{0r}$, $f_{1r}$ and $f_{2r}$ are all biased. Does this bias create problems?
- As the arrays are open longer, $\mu_{1f}$ and/or $\mu_{2f}$ increase.
  - We really start with frequencies $f_{1f}$ and $f_{1f}$ for background events, and then get $\mu_{1f}=T_1f_{1f}$ and $\mu_{2f}=T_2f_{2f}$.
  - How large do $T_1$ and $T_2$ have to be to yield a test with reasonable power?
- At what point can we discard the exact test and switch to asymptotics?

# Acknowledgements

- Professor Russell Zaretski, University of Tennessee Department of Statistics, for inviting me
- Professor John Belz, University of Montana Department of Physics, for bringing the original problem to my attention
- B.J. Harshfield, for assisting with the software development
- The R Project for Statistical Computing, for providing the software environment