Data Analysis II

Road Map

- Choosing the number of bins in a histogram
- Correcting Histograms for Systematic Uncertainty
  - Weighting Events
Recap

- Started analysis of SK-I two ring data
- Begin by looking at the basic variable distributions
- Found that there are differences between observation and expectations
- Systematic uncertainty

Pearson's $\chi^2 = 70.6$ for 40 d.o.f. (probability $\sim$0.2%)

This is a clear difference between data and MC. Need to find explanation
Other Distributions

Data

Simulation

Mean: 1.843 ± 0.06

Estimated uncertainty in mean is RMS/sqrt(entries)

Notice $\pi^0$ events don't have decay signals
Choosing Histogram Bin Sizes

- **Goldilocks Problem**
  - Fine bins
    - Can see the “fine-structure”
    - More statistical variation
  - Coarse bins
    - Loose the “fine-structure”
    - Minimizes statistical variation
  - Just right bins
    - Keep any real structure
      - i.e. approximate uncertainty of measurement
    - While maximizing number of events per bin
- Possible to vary bin size, but practicality → compromise

![Histogram of Two Ring Events](image)

- Not coarse enough
Small Bins, Small Statistics

- Too many bins hurts the statistical accuracy
- For high statistics, follows function better.
Medium Bins, Small Statistics

- Enough statistics per bin
- Moderately good at showing features of PDF
- At high statistics, PDF is washed out.
Big Bins

- Bins are too big to get information about PDF. This might be OK if you are looking for a simple excess of events.
- Not OK when you want information about the PDF shape.
Rules of Thumb for Bin Width

Lots of possible rules:
- Choose the number of bins
  \[ N_{\text{bins}} = \log_2 n_{\text{events}} + 1 \]
- Choose the bin width
  \[ W_{\text{bin}} = 3.5 \frac{\sigma}{\sqrt[3]{n}} \]

None of these rules work very well in real life.
- Choose the number of bins so that bin width is smaller than “smallest interesting feature”
- At the same time, choose the number of bins so that the “typical” signal bin and background bins are well separated
  \[ |N_{\text{signal bin}} - N_{\text{bkg bin}}| \gg \sqrt{N_{\text{signal bin}} + N_{\text{bkg bin}}} \]
True Direction vs Fit Direction

Very rough match between true and fit directions

Four Bins are about Right

Compromise
Event Direction

- Match between data and simulation isn't bad
  - It's suspicious
    - left side high
    - right side low

- From neutrino oscillations?
  - Left
    - Down-going (short neutrino path)
  - Right
    - Up-going (long neutrino path)

Pearson's $\chi^2 = 6.6$ for 3 d.o.f. (probability ~8.5%)
What we know about Neutrino Osc.

- The survival probability for muon neutrinos
  \[ P(\nu_\mu \rightarrow \nu_\mu) = 1 - \sin^2(2\theta) \sin^2\left(1.27 \Delta m^2 \frac{L}{E_\nu}\right) \]
  \(L\) in km, \(E\) in GeV

- Path length related to neutrino direction
  \[ L = R_{sk} \cos \theta_{\nu} + \sqrt{R_{\nu}^2 + R_{sk}^2 (\cos^2 \theta_{\nu} - 1)} \]

To keep things confusing, atmospheric neutrinos use the “zenith” angle, not “direction” angle (\(\cos \theta = - \cos \theta_z\))
Missing Pieces for Oscillations

Neutrino Energies

Muon Neutrino Oscillation Parameters

$\Delta m^2 \sim 0.0025$, $\theta \sim 45^\circ$
Event Directions w/ and w/o Decays

Pearson's $\chi^2 = 14$ for 3 d.o.f. (probability $\sim 0.2\%$)

Can't ignore neutrino oscillations
Event Weighting: Adding Effect of Osc.

- Include the oscillation affect assuming current best knowledge
  - $\Delta m^2$ is 0.0025
  - $\sin^2 2\theta$ is 1
- For each event in the MC, multiply by the survival probability
  - Weights must all be less than 1.
- Build histogram of the weighted events.

Pearson's $\chi^2 = 3.9$ for 3 d.o.f. (probability $\sim 27\%$)
Applying the Weights to a Histogram

```c
for (struct eventData* ev = gEventData; ev < &gEventData[gNEvents]; ++ev) {
    if (ev->decay<0.001) continue;
    if (ev->type == 0) {
        dataDecayDir->Fill(ev->direction);
    }
    else {
        mcDecayDir->Fill(ev->direction);
        double weight = oscWeight(ev->type,
                                    ev->nuType,
                                    ev->nuEnergy,
                                    ev->nuDirection);
        mcDecayDirOsc->Fill(ev->direction,weight);
    }
}
```

Fill histogram w/o muon neutrino osc.

Fill histogram w/ muon neutrino osc.
### The Oscillation Weight

```cpp
double path(double cosz) {
    double SKrad = 6368.38;
    double Erad = 6368.00;
    double NUHeight = 15.0;
    double NUrad = Erad + NUHeight;
    return sqrt(NUrad*NUrad + SKrad*SKrad*(cosz*cosz-1)) + SKrad*cosz;
}

double oscWeight(int type, int nuType, float nuEnergy, float nuDir) {
    if (std::abs(nuType) != 14) return 1.0;
    if (type != 1) return 1.0;
    double L = path(nuDir);
    double sstt = 1.0;
    double dms = 0.0025;
    double delta = 1.27*dms*L/nuEnergy;
    double weight = 1.0 - sstt*sin(delta)*sin(delta);
    if (delta>13*TMath::PiOver2()) weight = 0.5;
    return weight;
}
```
Weighting Comments

- Statistical uncertainty in a reweighted histogram
  \[ \sigma_{\text{bin}} = W_{\text{bin}} \sqrt{N_{\text{bin}}} \]

- You can use weighting to optimize the fit between MC and data
  - The chi-squared between the weighted MC and the data is just a function of the weight input parameters

\[ \chi^2 = 3.9 \]

\[ \chi^2 = 2.6 \]
Finally

- When plotting histograms:
  - Always think about the appropriate bin size before you make the plots
    - You should have some expectation about what the data will look like!
  - But, don't tune the bin size to based on making a particular significance
    - Trials Factor: More on this when we get to confidence limits
- Simulated data can be re-weighted to account for systematic uncertainties
  - Always reweight the simulation, and not the data
  - Make sure that the correct normalization is maintained when you reweight
  - Weights should always be less than one.

The End