The Significance of HEP Observations

Pekka K. Sinervo

Department of Physics

University of Toronto

20 March 2002

1. What is Significance?
2. Frequentist Approach
3. A Few Case Studies
4. Some Observations
5. Summary

What Do We Mean by Significance?

- **Typical HEP approach**
  - **Have a set of observations**
  - **We say the data are “statistically significant” when**
    - We can use data to support a specific hypothesis, eg.
      - **“We see a phenomenom not predicted by the Standard Model”**
      - **“We report the discovery of X”**
    - The interpretation eliminates a number of competing hypotheses
    - The conclusion will not likely be altered with larger statistics or further analysis
- **Want a statistical framework that**
  - **Measures “degree of belief”**
  - **Ensures robust conclusions**

Some “Obvious” Discoveries

- **Observation of $B^0\overline{B}^0$ Mixing**
  - 24.8±7.6±3.8 like-sign events vs 25.2±5.0±3.8 opposite sign
  - “3σ” discovery

Albrecht et al., PLB 192, 245 (1987)

- **W Boson**
  - 6 eν events
  - No background!

Arnison et al., PLB 122, 103 (1983)

- **Upsilon**
  - 770 events on 350 background
  - Described as “significant” but no measure of it

Herb et al., PRL 39, 252 (1977)

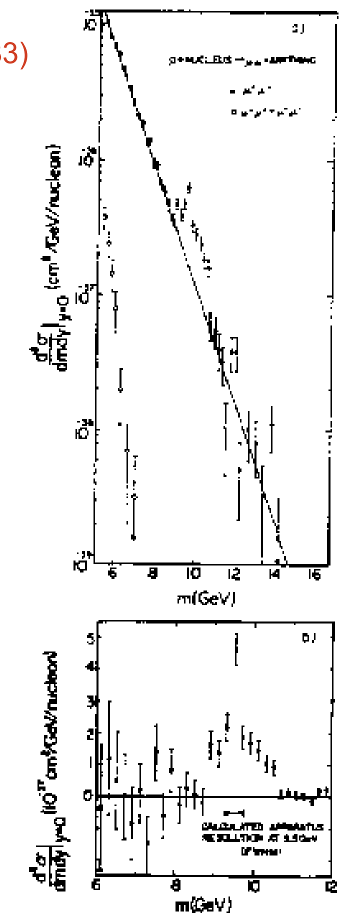


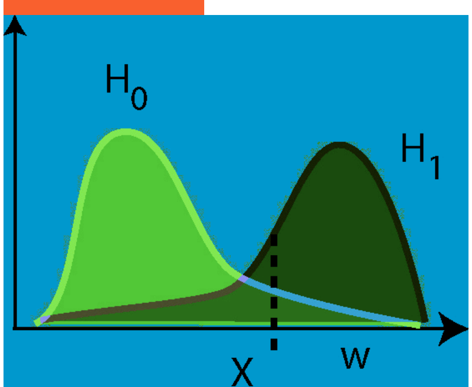
FIG. 3. (a) Measured dimuon production cross sections as a function of the invariant mass of the muon pair. The solid line is the continuum fit outlined in the text. The equal-sign—dimuon cross section is also shown. (b) The same cross sections as in (a) with the smooth exponential continuum fit subtracted in order to reveal the 9–10-GeV region in more detail.

- **B mesons**
  - 18 events on 4-7 background
  - No measure of significance

Behrends et al., PRL 50, 881 (1983)

A Frequentist Definition

- **Significance defined in context of "hypothesis testing"**
  - **Have two hypotheses, H_0 and H_1, and possible set of observations X**
    - Choose a "critical region", w, in the space of observations X
    - Define **significance**, □, as the probability of X □w when H_0 is true
    - Define the **power**, 1-□, to be the probability of X □w when H_1 is true



Typically, H_0 is "null" hypothesis

- **In this language, an observation is "significant" when**
  - **Significance □ is small & □ is small**
    - Typically □ < few 10^{-5}

Some Comments on Formal Definition

- **Definition depends on**
  - **Choice of statistic X**
    - Left up to the experimenter as part of design
    - More on that later
  - **Choice of “critical region” w**
    - Depends on hypotheses
    - Often chosen to minimize systematic uncertainties?
    - Not necessarily defined in advance!
  - **Definition of “probability”**
    - A frequentist definition
    - Raises issue of how systematic uncertainties are managed
  - **Choice of □ and □**
    - Matter of “taste” and precedent
    - A small □ is safe, but comes with less “discovery reach”

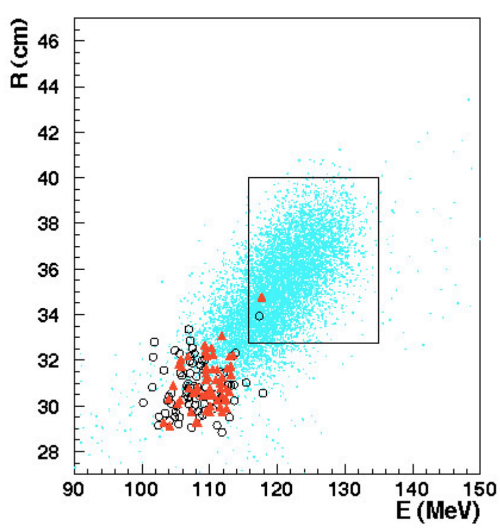
- **More fundamentally:**
  - **Is this an adequate definition of “significance?”**

The Choice of Statistic & Critical Region

- **Choice of statistic motivated by specific experimental design**
  - **Informed by the measurement to be made**
  - **Critical region is chosen at the same time**
  - **Good example: E787/E949 search**

$K^+ \to \pi^+ \nu\bar{\nu}$

  - Look for $\pi^+ \to \mu^+ \to e$ decay
  - Define a “box” a priori
    - **Expected 0.15±0.05 event bkgd**



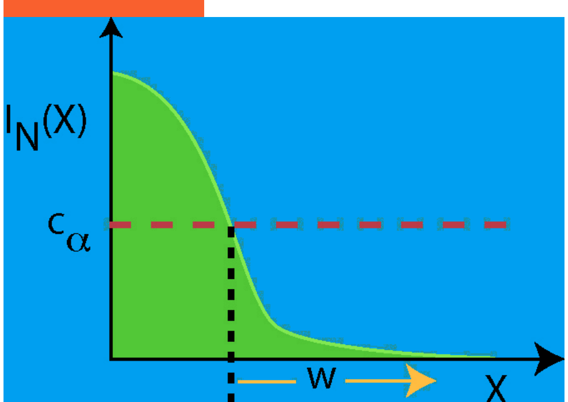
Only two events
Observed

Significance 0.02%

Have used the “box”
Since 1988

Optimal Tests: Neyman-Pearson

- **In some cases, possible to identify the "most powerful" test**
  - **Must involve only "simple" hypotheses (no free parameters)**
    - PDF's given by $f_i(X)$
    - Must have two hypotheses
  - **For given α, can identify region to minimize β for alternative H_1**
    - Order observations by
      $$l_N(X) \equiv f_0(X) / f_1(X)$$
    - Can minimize β by choosing critical region as all X s.t. $l_N(X) \geq c_\alpha$
      - **Chose c_α so that**



$$\int_w \mathbf{f}_0(\mathbf{X})dX = \alpha$$

Caveats to Neyman-Pearson

- **Neyman-Pearson limited**
  - **Only true for simple hypotheses**
    - Not for composite hypotheses (where unknown parameter)
  - **Compares two hypotheses**
    - Depends on alternative hypothesis
    - Makes results model-dependent

- **But does give some insight**
  - **The ratio $I_N(X)$ is proportional to ratio of likelihoods**

$$f_0(X) / f_1(X) \propto \mathcal{L}_0(X) / \mathcal{L}_1(X)$$

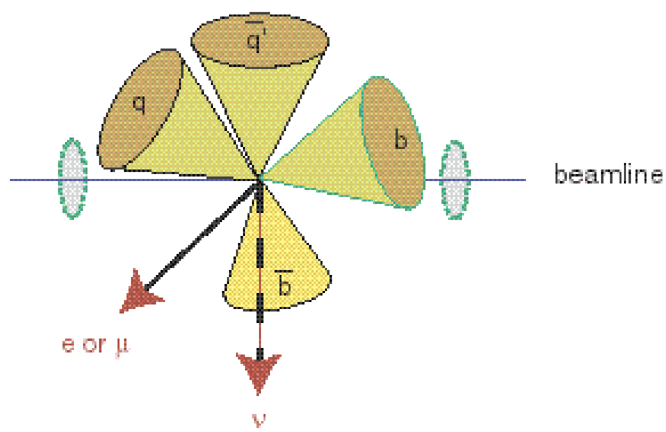
  - **Provides guidance for definition of effective tests**

Definition of Critical Region

- **Challenge is not to bias choice of critical region with data**
  - **However, observer required to understand data**
    - Identify instrumental pathologies
    - Identify unexpected backgrounds
    - Estimate systematic uncertainties
    - Verify stable run conditions
  - **Studies may lead to unconscious bias (see, eg. RPP plots!)**
- **"Blind" analyses are popular**
    - Study data complementary to signal
    - However, implementation varies
      - **SNO's pure D_2O results set aside about 40% of data**
      - **Not clear that this really helps!**
    - Even E787/E949 reserve right to examine background rejection

Significance in Counting Experiments

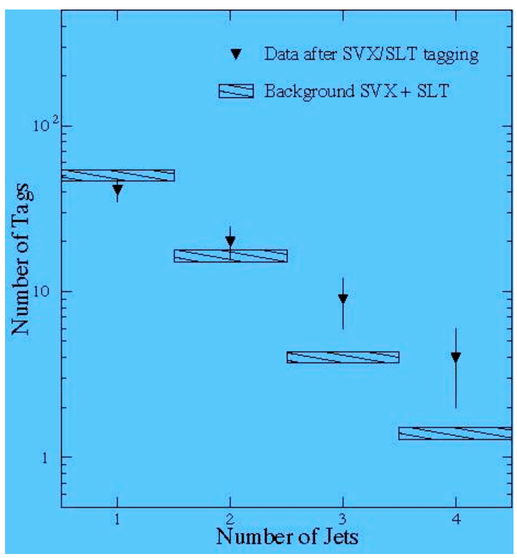
- **Top quark search is textbook example**
  - **By 1991, CDF had ruled out top quark with mass < 91 GeV/c^2**
  - **Searching for top quark pair production and decay into**
    - Lepton + + jets (20%)
    - Dilepton + + jets (8%)



- **In a sample of 20 pb^{-1}, expected handful of events**
  - **Large background from W + jets**
  - **"Fake" b-quark tags**

Definition of the Measurement

- **Defined clear strategy in 1990**
  - **Identify lepton+jets and dilepton candidates**
  - **Count "b" tags in lepton+jet events**
    - Use two b-tagging algorithms
      - **Use events with 1-2 jets as control**
      - **Signal sample events with ≥3 jets**
      - **Expected 3.5 evts (M_{top}=160 GeV/c^2)**



Observed **13** tagged "b jets" in 10 evts

7 SVX tags
6 lepton tags

Expect **5.4±0.4** tags from background

  - **For dileptons:**
    - Require 2 or more jets
    - Expected 1.3 evts (M_{top}=160 GeV/c^2)
    - Observed 2 evts, bkd of 0.6±0.3 evts

Significance Calculation

- **Calculated probability of background hypothesis**
  - **Dilepton significance $\square_{dil}$ = 0.12**
  - **Used MC calculation**
    - Treated background uncertainty as a normally distributed uncertainty on acceptance
  - **For lepton+jets, MC gives**
    - SVX b tags: $\square_{SVX}$ = 0.032
    - Soft lepton b tags: $\square_{SLT}$ = 0.038
- **To combine, take correlations in tags in background into account**
  - **Gives $\square_{tot}$ = 0.0026**
  - **If assume independent, then**

    $\square_{tot} = \square_{dil}\,\square_{ljets}\,[1 - \ln(\square_{dil}\,\square_{ljets})]$
    - Gives $\square_{tot}$ = 0.0088
  - **Collaboration reported only "evidence for top quark…."**
    - Factor 2 more data -- $\square_{tot}$ = few 10^{-5}

Power of the Top Quark Statistic

- **Choice of statistic driven by need to reduce background**
  - **Note $\square_{jets}$ = 0.074 before b-tagging**
    - Predict 12 events signal and 60 events background
    - Tagging efficiency 0.40
      - **Background "efficiency" 0.09**
  - **Definition of "power" problematic**
    - Arbitrary
      - **Power of lepton+jets selection?**
      - **Power of b-tagging?**
      - ***A posteriori* choice of $X = N_{tags} + N_{dil}$**
    - Experimenter chooses "critical region" based on hypothesis
      - **Lepton+jets Higgs search uses different selection**

        W H □ l □ b b
  - **Usually characterized by sensitivity**
    - Size of expected signal

Significance using Data Distributions

- **Measurements often involve continuous observables**
  - **Can assess agreement with "null" hypothesis**
    - Generally "goodness-of-fit" tests
- **Number of tests in common use**
    - χ^2 Test
      - **Depends on choice of binning**
      - **Limited to "large" statistics samples**
        - **Bin contents > 5-10 (?)**
    - Smirnov-Cramer-Von Mises
      - **Define statistic based on cumulative distributions $S_N(x)$**

$$W^2 \equiv \int \left[S_N(X) - F(X)\right]^2 f(X)\, dX$$

      - **Probability distribution for W^2 independent of distribution**
        - **$E[W^2] = (6N)^{-1}$ and $V[W^2] = (4N-3)/180N^3$**
    - Kolmogorov-Smirnov
      - **Popular form of test based on $S_N(x)$**

$$D_N \equiv \max\left|S_N(X) - F(X)\right|$$

      - **Distribution for D_N proportional to χ^2**
        - **Can be converted into a significance**

Multivariate Significance

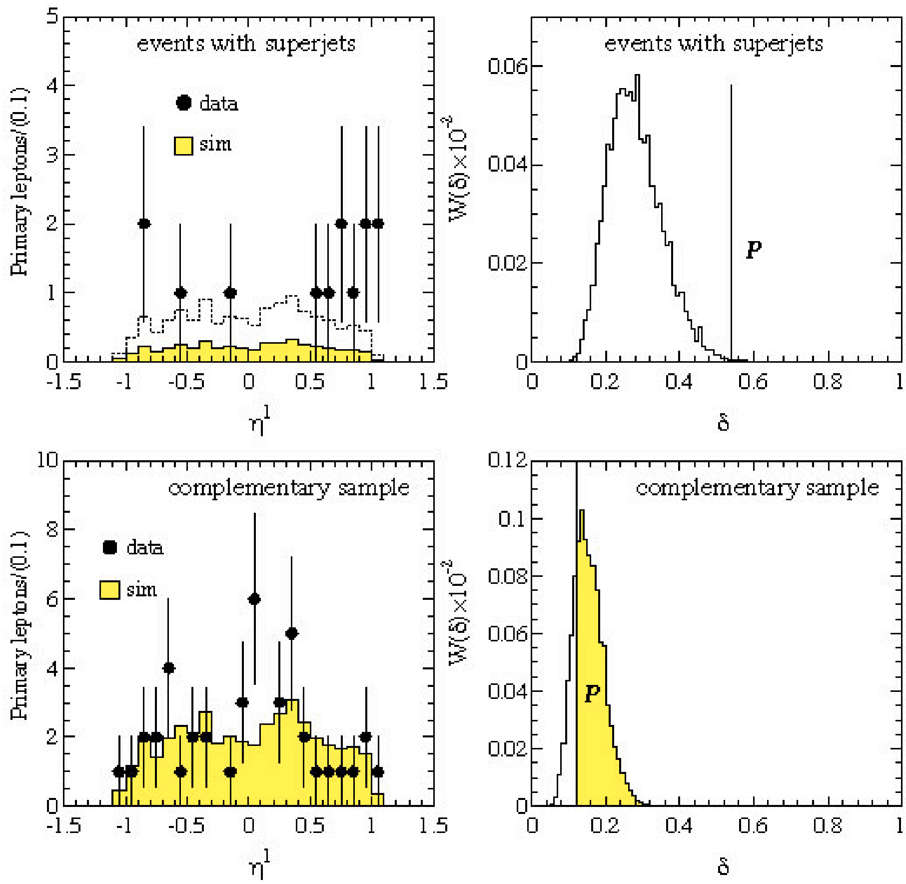
- **Often difficult to reduce data to one-dimensional statistic**
  - **Typical case has several variables**
    - Different correlations between signal and "null" hypothesis
    - Any straightforward transformation causes loss of information
  - **Several techniques used**
    - Characterize significance of each component and then combine into a single measure of significance
    - More sophisticated, e.g.
      - **Combine information using any one of the techniques discussed by Prosper, Towers, etc.**
- **In practice, two approaches:**
  1. **Assume independent statistics**
     - Check for any correlations
  2. **Model correlations using MC approaches or "bootstrapping"**
     - Computationally expensive
     - Relies on understanding correlations

A Recent Example: "Superjets"

- CDF Run I data contained
  - Unusual lepton + □ + 2,3 jet events
    - 13 events with jets that are both SLT and SVX tagged
      - Expect 4.4±0.6 events from background sources
      - Significance is 0.001!
  - Led to examination of 9 kinematical distributions
      - P_T & □ for leptons & jets, and azimuthal angle between lepton, jet
      - P_T and □ for lepton+jet system
    - Perform independent K-S tests
      - Use control sample defined by events without a "supertag"
      - Combined significance of 1.6×10^{-6}
    - Also defined a new statistic
      - Sum of K-S distances
      - MC gives significance of 3.3×10^{-6}

K-S Tests on Superjet Data

- **Lepton distribution**



  - **Some approximations:**
    - Control sample events w/o superjet
    - Randomly pick 13 of 42 events
    - Also checked with MC calculation of background

Comments on Superjet Study

- Choice of statistic (number of superjets) problematic
  - Made ***a posteriori*** after anomaly noted
    - Significance difficult to assess
  - Ignored lepton + 1 jet data (where one observes a deficit of events)
    - Why?
- Choice of distributions also problematic
  - Justified ***a posteriori***
  - Correlations difficult to assess
- Aside:
  - Interpretation of excess requires unusual physics process
    - Not a problem in itself
    - But small statistics allow for many hypotheses

Some Practical Proxies for Significance

- **HEP suffers Gaussian tyranny**
  - **Many people will quote numbers of "□" as measures of significance**
    - Belief that this can be more readily interpreted by lay person
      - **Shorthand for the significance of an n□ measurement**
    - 5□ seems to have become conventional "discovery threshold"
      - **□ = 2.8×10^{-5}**
      - **Used for LHC discovery reach**
- **In situations where expected signal S and background B**
  - **Various figures of merit**
    - S/N -- signal versus noise
      - **Doesn't scale with N**
    - More natural definition is

$$\frac{S}{\sqrt{B}}$$

See talk by Bityukov & Krasnikov for more discussion

      - **Just normal Gaussian estimate of # of s.d.**
      - **Does scale with N**

The “Flip-Flopping” Physicist

- **Feldman & Cousins highlighted the problem of “flip-flopping”**
  - **A physicist who uses**
    - One set of criteria to set a limit in the absence of a signal
    - Different criteria to claim a significant signal
  - **Results in confidence intervals with ill-defined frequentist coverage**

- **This should be anticipated in any experiment that wishes to be sensitive to small signals**
  - **F-C propose their “unified approach”**

What About Reverend Bayes?

- **Bayesian approach to classifying hypotheses is**

$$\frac{P(H_1 \mid X)}{P(H_0 \mid X)} = \frac{P(X \mid H_1)}{P(X \mid H_0)} \bullet \frac{\pi(H_1)}{\pi(H_0)}$$

  - **Few comments:**
    - $P(X|H_i)$ is typically likelihood
    - Only meaningful in comparison of two hypotheses
    - Can handle composite hypotheses readily
      - **Just integrate over any “nuisance” variables**
- **Is it used? Not often…**
  - **Only relative “degree of belief”**
    - Requires at least two hypotheses
  - **“Prior” avoidance**
  - **Challenges where single points in parameter space are important**
    - Is $\sin 2\beta = 0$?

Some Recommendations

- **Define measurement strategy in advance of data analysis**
  - **Otherwise, significance estimates could and will be biased**
  - **"Blind" analyses can play a role**
    - However, this should not limit the ability to "explore" the data
- **Take consistent approach to CL setting & signal measurement**
  - **Avoid "flip-flopping" -- F-C offers one approach to this problem**
- **Describe clearly how you are determining "significance"**
  - **Things to remember:**
    - Definition of probability
    - Definition of critical region
    - What decisions were taken a posteriori?

Summary and Conclusions

- **Signal significance a well-established concept**
  - **Literature full of frequentist examples**
  - **Used to reject “null hypothesis”**
  - **Bayesian approaches haven’t entered mainstream**
- **Potential for abuse**
  - **Using *a posteriori* information makes any significance calculation suspect**
  - **Obligation to be explicit about assumptions**
- **HEP discovery “threshold”**
  - **Appears to be “5σ”**
    - Significance of 2.8×10^{-7}

Truly a conservative bunch!