# **Electoral Integrity Statistical Indicators**

## **Bihar Election Analysis**

## **CATEGORY 1: BASIC CLASSIFICATIONS**

### **L - Deletion Category**

**What it measures:** Groups booths into simple deletion ranges for easy analysis  
**Example:** Valmikinagar Booth 100 has 67 deletions = "Medium (51-100)", Booth 1 has 224 deletions = "Extreme (>200)"  
**Why important:** Helps quickly identify which booths had unusually high voter deletions  
**How calculated:** Creates four buckets: ≤50, 51-100, 101-200, >200  
**Results mean:**

* "Low (≤50)" = Normal administrative deletions
* "Medium (51-100)" = Moderate deletions requiring monitoring
* "High (101-200)" = High deletions requiring investigation
* "Extreme (>200)" = Extremely suspicious, needs immediate investigation  
   **Implications:** Booths in "Extreme" category should be audited first - they represent less than 1% of all booths but account for disproportionate deletions  
   **Formula:** ,=IF(E2<=50,"Low (≤50)",IF(E2<=100,"Medium (51-100)",IF(E2<=200,"High (101-200)","Extreme (>200)")))

### **N - Booth Number Category**

**What it measures:** Groups polling booths by their sequence numbers  
 **Example:** Booth 100 = "001-100", Booth 350 = "300+", showing administrative zones  
 **Why important:** Helps detect if deletions follow administrative patterns (early vs late booths)  
 **How calculated:** Creates ranges: 001-100, 101-200, 201-300, 300+  
 **Results mean:** Can reveal if deletions are concentrated in specific administrative zones  
 **Implications:** If "001-100" booths consistently have higher deletions, it suggests administrative rather than demographic causes  
 **Formula:** ,=IF(D2<=100,"001-100",IF(D2<=200,"101-200",IF(D2<=300,"201-300","300+")))

### **P - Competitive Index**

**What it measures:** How close the election was (margin of victory)  
 **Example:** Valmikinagar with 12% margin = "Moderate", while a constituency with 3% margin = "Very Close"  
 **Why important:** Tight races are more vulnerable to manipulation through voter deletions  
 **How calculated:** Groups margins: <5% Very Close, 5-10% Close, 10-20% Moderate, >20% Safe  
 **Results mean:** "Very Close" races are high-stakes where even small deletions could change outcomes  
 **Implications:** Focus investigation resources on "Very Close" and "Close" races where deletions could have altered results  
 **Formula:** ,=IF(H2<0.05,"Very Close",IF(H2<0.1,"Close",IF(H2<0.2,"Moderate","Safe")))

## **CATEGORY 2: STATISTICAL DISTRIBUTION ANALYSIS**

### **AX - Overall Deletion Mean**

**What it measures:** Average deletions across ALL booths in the dataset  
 **Example:** If overall mean is 84 deletions, then Valmikinagar Booth 1 (224 deletions) is 2.7 times higher than normal  
 **Why important:** Baseline for comparing individual booth performance  
 **How calculated:** Sum all deletions ÷ total number of booths  
 **Results mean:** Any booth significantly above this suggests unusual activity  
 **Implications:** Booths with deletions >3x the mean deserve immediate investigation regardless of local conditions  
 **Formula:** ,=AVERAGE($E$2:$E$232433)

### **BC - Constituency Deletion Skewness**

**What it measures:** Whether most booths have low deletions with few extreme cases, or vice versa  
 **Example:** Skewness of +2.5 means most Valmikinagar booths have low deletions but a few have extremely high ones  
 **Why important:** Normal administrative processes should follow predictable bell-curve patterns  
 **How calculated:** Statistical measure where 0 = symmetric, positive = right-skewed, negative = left-skewed  
 **Results mean:** Values above +1 or below -1 indicate unnatural patterns  
 **Implications:** High positive skew suggests targeted manipulation at specific booths rather than systematic administrative issues  
 **Formula:** ,=SKEW(INDIRECT("$E$"&MATCH(B2,$B$2:$B$232433,0)+1&":$E$"&MATCH(B2,$B$2:$B$232433,1)+1))

## **CATEGORY 3: ANOMALY DETECTION**

### **S - Deletion Z-Score**

**What it measures:** How many standard deviations a booth's deletions are from normal  
 **Example:** Booth with 350 deletions in constituency where mean=80, SD=25 gets Z-score of 10.8 (extremely abnormal)  
 **Why important:** Scientific standard for identifying true outliers vs normal variation  
 **How calculated:** (Individual value - Overall average) ÷ Overall standard deviation  
 **Results mean:** Z-score >3 = 99.7% confidence it's abnormal, >2 = 95% confidence  
 **Implications:** Any booth with Z-score >3 has less than 0.3% chance of being normal - investigate immediately  
 **Formula:** ,=(E2-AVERAGE($E$2:$E$232433))/STDEV($E$2:$E$232433)

### **O - Deletion Severity Flag**

**What it measures:** Automatically flags booths with statistically extreme deletions  
 **Example:** Booth with 450 deletions gets "EXTREME" flag, 180 deletions gets "HIGH" flag  
 **Why important:** Computer-generated alerts prevent human bias in identifying problems  
 **How calculated:** Compares each booth to 2 and 3 standard deviation thresholds  
 **Results mean:** "EXTREME" = investigate immediately, "HIGH" = monitor closely, "NORMAL" = OK  
 **Implications:** "EXTREME" booths represent statistical impossibilities under normal operations and require forensic audit  
 **Formula:** ,=IF(E2>AVERAGE($E$2:$E$232433)+3*STDEV($E$2:$E$232433),"EXTREME",IF(E2>AVERAGE($E$2:$E$232433)+2*STDEV($E$2:$E$232433),"HIGH","NORMAL"))

## **CATEGORY 4: ELECTORAL COMPETITION ANALYSIS**

### **M - Deletion-to-Margin Ratio**

**What it measures:** Relationship between voter deletions and victory margins  
 **Example:** Booth with 200 deletions in race won by 1,000 votes = ratio of 200 (high concern)  
 **Why important:** High deletions in close races could indicate targeted manipulation  
 **How calculated:** Deletions ÷ victory margin × 1000 for readability  
 **Results mean:** High ratios = lots of deletions relative to how close the race was  
 **Implications:** Ratios >100 in competitive races suggest deletions may have been strategically targeted to affect outcomes  
 **Formula:** ,=IF(G2=0,0,E2/G2\*1000)

### **AE - Party Dominance Index**

**What it measures:** How much stronger the winning party is than the second-place party at each booth  
 **Example:** BJP gets 60% at one booth, Congress gets 15% = dominance index of 45% (very high)  
 **Why important:** Extreme dominance in some booths might indicate vote manipulation  
 **How calculated:** Winner's vote share minus highest non-winner vote share in same booth  
 **Results mean:** Very high values = suspicious dominance, very low = competitive  
 **Implications:** Booths showing >70% dominance in otherwise competitive constituencies warrant investigation for irregularities  
 **Formula:** ,=K2-MAXIFS($K$2:$K$232433,$A$2:$A$232433,A2,$I$2:$I$232433,"<>"&I2)

## **CATEGORY 5: COMPREHENSIVE SCORING SYSTEMS**

### **R - Red Flag Score (0-3)**

**What it measures:** Overall suspicion level combining multiple risk factors  
 **Example:** Booth with 250 deletions in race decided by 2% gets score of 3 (highest concern)  
 **Why important:** Single number that prioritizes which booths need investigation first  
 **How calculated:** 1 point for >100 deletions, 2 for >300 deletions, 3 for >200 deletions + close race  
 **Results mean:** 0=no concerns, 1=minor monitoring, 2=investigate, 3=immediate priority  
 **Implications:** Score 3 booths should receive forensic audit within 48 hours as they represent highest manipulation risk  
 **Formula:** ,=IF(AND(E2>200,H2<0.05),3,IF(OR(E2>300,AND(E2>150,H2<0.1)),2,IF(E2>100,1,0)))

### **AA - Statistical Alert**

**What it measures:** Three-tier classification system for prioritizing action  
 **Example:** Booth with extreme deletions + close race = "INVESTIGATE", moderate issues = "MONITOR"  
 **Why important:** Clear guidance for investigators on where to focus limited resources  
 **How calculated:** Combines Red Flag Score, Z-Score, and competitiveness into action categories  
 **Results mean:** "OK"=no concerns, "MONITOR"=watch for patterns, "INVESTIGATE"=urgent attention  
 **Implications:** "INVESTIGATE" booths should be audited immediately while "MONITOR" booths need increased oversight in future elections  
 **Formula:** ,=IF(OR(R2>=3,S2>3,AND(E2>100,H2<0.05)),"INVESTIGATE",IF(OR(R2>=2,S2>2),"MONITOR","OK"))

## **CATEGORY 6: ADVANCED STATISTICAL VALIDATION**

### **CN - Normality Score (0-10)**

**What it measures:** Comprehensive assessment of how "natural" a constituency's deletion pattern looks  
 **Example:** Constituency with score 9 follows expected patterns, score 2 shows highly artificial patterns  
 **Why important:** Normal administrative processes create predictable statistical patterns  
 **How calculated:** Points awarded for meeting skewness, kurtosis, and distribution criteria  
 **Results mean:** 8-10=natural patterns, 6-7=mostly normal, 4-5=questionable, 0-3=highly artificial  
 **Implications:** Constituencies scoring below 4 show strong evidence of systematic manipulation rather than administrative variation  
 **Formula:** ,=(IF(ABS(BC2)<0.5,2,IF(ABS(BC2)<1,1,0))+(IF(ABS(BV2)<1,2,IF(ABS(BV2)<2,1,0))+(IF(BZ2>0.65,2,IF(BZ2>0.6,1,0))+(IF(CA2>0.92,2,IF(CA2>0.9,1,0))+(IF(CB2>0.95,2,IF(CB2>0.9,1,0)))))

### **BW - Jarque-Bera Test Statistic**

**What it measures:** Formal statistical test for whether deletion patterns follow natural administrative processes  
 **Example:** Score of 12.5 (above 5.99 threshold) means 95% confidence that patterns are not natural  
 **Why important:** Academic gold standard for detecting artificial vs natural patterns  
 **How calculated:** Combines skewness and kurtosis into single statistical test  
 **Results mean:** Values >5.99 indicate statistically significant deviation from normal administrative patterns  
 **Implications:** Constituencies failing this test (>5.99) require immediate administrative review and potential investigation for manipulation  
 **Formula:** ,=COUNTIF($B$2:$B$232433,B2)/6\*(BC2^2+(BV2^2)/4)

## **CATEGORY 7: BIAS DETECTION (CRITICAL FOR ACCURATE ANALYSIS)**

### **CT - Simpson's Paradox Flag**

**What it measures:** Whether conclusions reverse when data is aggregated vs analyzed separately  
 **Example:** Party X shows higher deletions in each constituency individually, but lower deletions overall  
 **Why important:** Prevents drawing completely wrong conclusions from misleading aggregated data  
 **How calculated:** Compares correlation direction at constituency level vs state level  
 **Results mean:** "SIMPSON'S PARADOX"=don't trust simple totals, "Consistent Trend"=aggregation OK  
 **Implications:** When present, all aggregate-level conclusions are invalid and constituency-by-constituency analysis is required  
 **Formula:** ,=IF(SIGN(CR2)<>SIGN(CS2),"SIMPSON'S PARADOX","Consistent Trend")

### **DD - Bonferroni Adjusted P-Value**

**What it measures:** Statistically corrected significance after testing 241 constituencies simultaneously  
 **Example:** Unadjusted p-value of 0.03 becomes 0.03×241=7.23 (not significant) after correction  
 **Why important:** Prevents false discoveries when analyzing hundreds of constituencies at once  
 **How calculated:** Multiplies regular p-value by number of constituencies tested (241)  
 **Results mean:** More stringent test - only truly significant patterns survive correction  
 **Implications:** Only patterns passing this corrected test should be reported to avoid misleading conclusions based on statistical flukes  
 **Formula:** ,=MIN(1,IF(ABS(S2)>1.96,2\*(1-NORM.S.DIST(ABS(S2),TRUE)),1)\*241)

## **CATEGORY 8: PRACTICAL FIELD CLASSIFICATION**

### **W - Booth Type Classification**

**What it measures:** Practical categories for field investigation teams  
 **Example:** Booth with 300 deletions in 4% margin race = "High Risk", 350 deletions in 25% margin = "High Deletion"  
 **Why important:** Helps allocate investigative resources efficiently based on risk and impact  
 **How calculated:** Combines deletion levels with race competitiveness for practical categories  
 **Results mean:** "High Risk"=immediate forensic audit, "High Deletion"=process review, "Competitive"=election security  
 **Implications:** "High Risk" booths get forensic teams, "High Deletion" gets administrative review, "Competitive" gets enhanced monitoring  
 **Formula:** ,=IF(AND(E2>150,H2<0.1),"High Risk",IF(E2>200,"High Deletion",IF(H2<0.05,"Competitive",IF(E2<25,"Low Activity","Normal"))))

### **X - Constituency Deletion Rank**

**What it measures:** How this booth ranks within its constituency for deletions  
 **Example:** In Valmikinagar, Booth 1 with 224 deletions might rank #1 (worst), Booth 100 with 67 deletions ranks #15  
 **Why important:** Identifies the most problematic booths within each local area  
 **How calculated:** Counts how many booths in same constituency have higher deletions, adds 1  
 **Results mean:** Rank 1 = highest deletions in constituency, higher ranks = relatively better  
 **Implications:** Rank 1-3 booths in each constituency should be prioritized for detailed audit as they represent local extremes  
 **Formula:** ,=COUNTIFS($B$2:$B$232433,B2,$E$2:$E$232433,">"&E2)+1

*Never rely on a single indicator. The power of this system comes from multiple statistical approaches confirming each other. A booth flagged by only one indicator needs verification, while a booth flagged by 3+ different methods requires immediate investigation.*