



Detecting Language Contact by Lexical Skewing

James W. Minett

Language Engineering Laboratory

City University of Hong Kong

April 8, 2003



Lexicostatistical Classification

- Aim: to identify the **genetic relationships** among a group of related languages
- Method: languages with **high lexical similarity** are inferred to be closely genetically related
- ... but genetic transmission is **not the only mechanism** by which languages can appear lexically similar
 - **borrowing** due to **language contact**
e.g. Chinese → Japanese; French → English
- Can language contact be **detected**?



Language Contact Among 3 Languages — **concept**

- The Africanist **Tom Hinnebusch** quotes a comment by Heine (1974):
 - “The Nilotic languages Samburu and Nandi **share 9.9 percent** lexical resemblances on the basis of the 200-word list.”
 - “The percentage between Masai and Nandi, on the other hand, amounts to **15.7**. These two languages have been in **close contact** over the last few centuries.”
 - “It seems reasonable to assume that the difference of **5.8** percent between Samburu/Nandi and Masai/Nandi is a result of the process of **borrowing** which took place **between Masai and Nandi**.”

- Define the *lexical skewing* between 2 languages, L_i and L_j , with respect to a third language, L_k :

$$\delta S_{ij}^k = \frac{1}{N} \sum_{q=1}^N (s_{ik}^q - s_{jk}^q)$$

where N denotes the number of characters available and s_{ij}^q indicates when two languages, L_i and L_j , share the same state for character c^q :

$$s_{ij}^q = \begin{cases} 0 & : c_i^q \neq c_j^q \\ 1 & : c_i^q = c_j^q \end{cases}$$

e.g. $\delta S_{ij}^k = 15.7\% - 9.9\% = 5.8\%$ for Masai/Samburu w.r.t. Nandi



Lexical Skewing

- But it is **not enough** to know just the degree of lexical skewing, dS_{ij}^k
- An analogy:
 - if I toss a coin **20** times I might not be too surprised to observe **15 heads** and **5 tails**, i.e. skewing of **10** (Pr = 1.5%)
 - but if I toss a coin **10** times I would be very surprised to observe **10 heads** and **0 tails**, i.e. skewing of **10** (Pr = 0.1%)
 - and if I toss a coin only **5** times it is impossible to observe skewing of **10**
- The amount of skewing that is a significant indicator of contact depends on the **number of skewed characters**, N_{ij}^k

- The number of **skewed** characters, N_{ij}^k , for which:

$$s_{ik}^q \neq s_{jk}^q \quad \text{i.e. } c_i^q = c_k^q \quad \text{and} \quad c_j^q \neq c_k^q$$

e.g. **MOUNTAIN**: English “**mountain**,” German “**berg**,” French “**montagne**”

- The number of **positively skewed** characters, n_{ij}^k , for which:

$$s_{ik}^q = 1 \quad \text{and} \quad s_{jk}^q = 0$$

- The number of **negatively skewed** characters, n_{ji}^k , for which:

$$s_{ik}^q = 0 \quad \text{and} \quad s_{jk}^q = 1$$



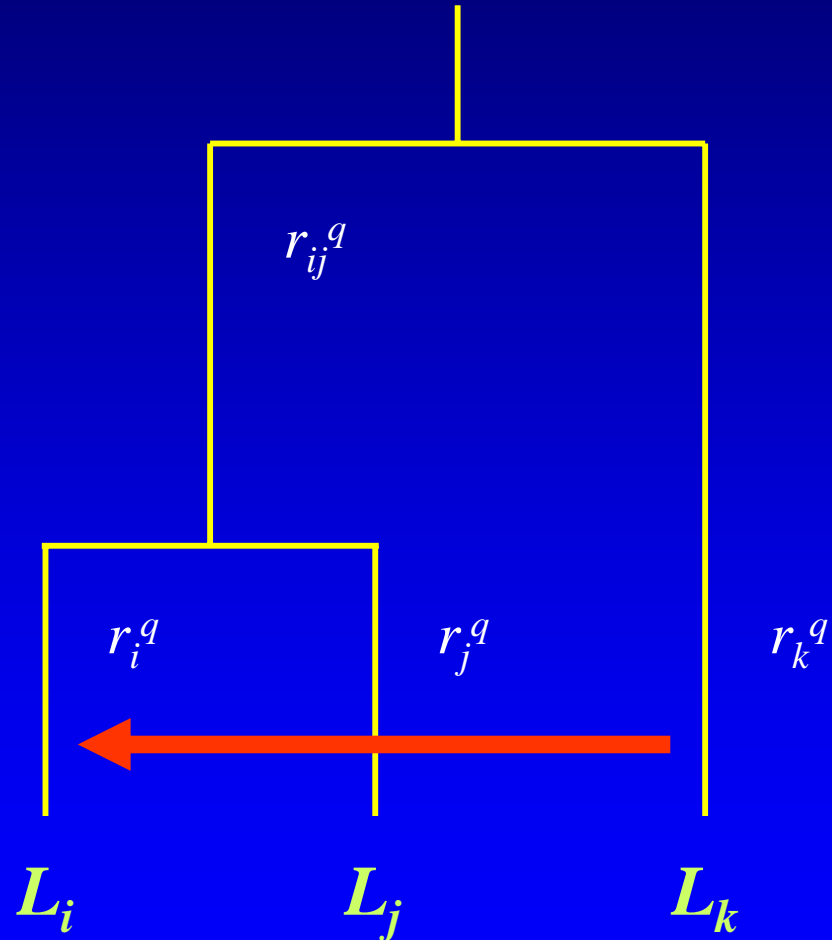
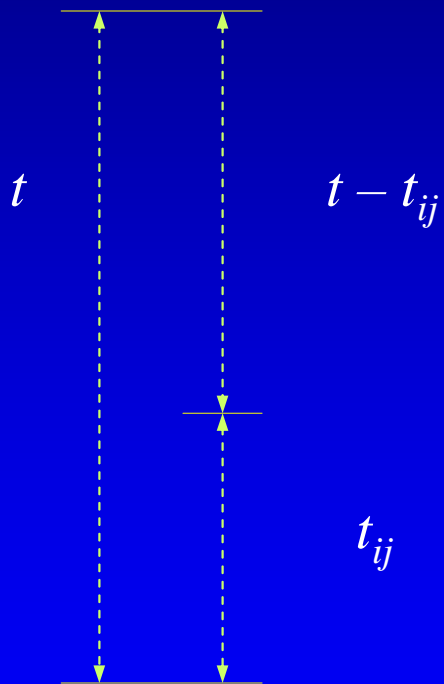
Calculating the Skewing

character states:	L_i	L_j	L_k	n_{ij}^k	n_{ji}^k	
c^1	1	1	2	—	—	
c^2	1	2	3	—	—	
c^3	1	2	2	—	—	} negatively skewed
c^4	1	2	2	—	—	
c^5	2	1	2	+	—	} positively skewed
c^6	2	1	2	+	—	
c^7	2	1	2	+	—	
c^8	2	1	2	+	—	
	—	—	—	4	2	$\delta S_{ij}^k = n_{ij}^k - n_{ji}^k$



Language Contact Among 3 Languages — **framework**

time depths:



- Aim to **detect contact** by implementing the decision rule:

$$\begin{array}{ccc}
 & \text{contact} & \\
 n_{ij}^k & \geq & \nu \\
 & < & \\
 & \text{no contact} &
 \end{array}
 \quad \text{i.e.} \quad
 \begin{array}{ccc}
 & \text{contact} & \\
 \delta S_{ij}^k & \geq & \frac{2\nu - N_{ij}^k}{N} \\
 & < & \\
 & \text{no contact} &
 \end{array}$$

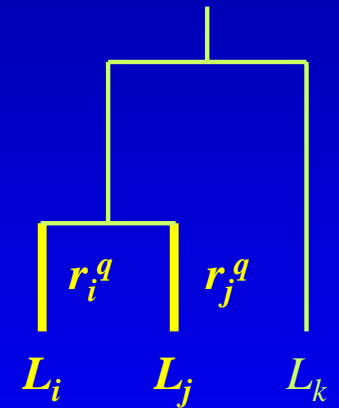
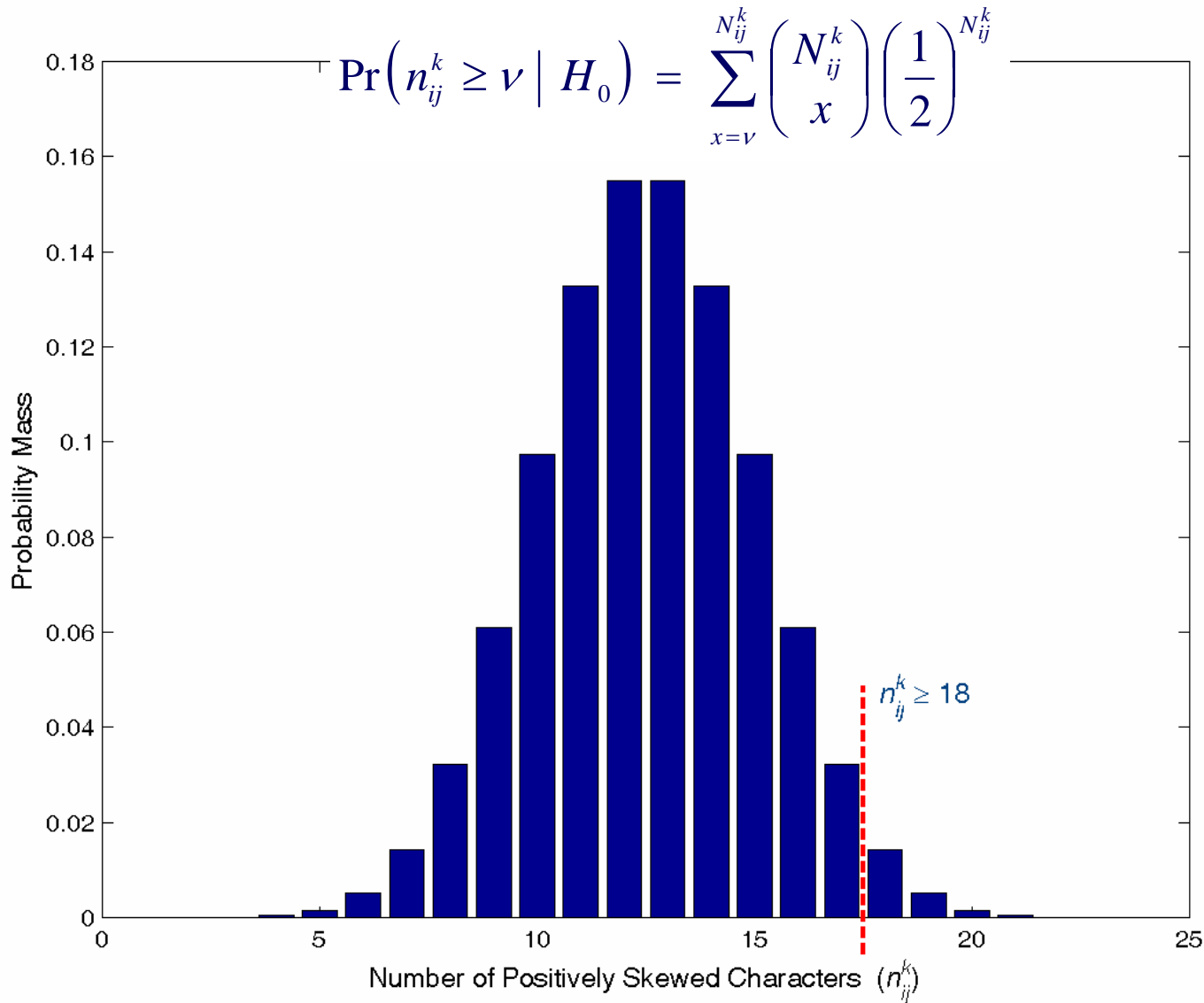
where ν is some threshold such that

$$\Pr(n_{ij}^k \geq \nu \mid \text{no contact}) \leq \alpha$$

- Put simply, infer **contact** when sufficiently **large skewing** is observed

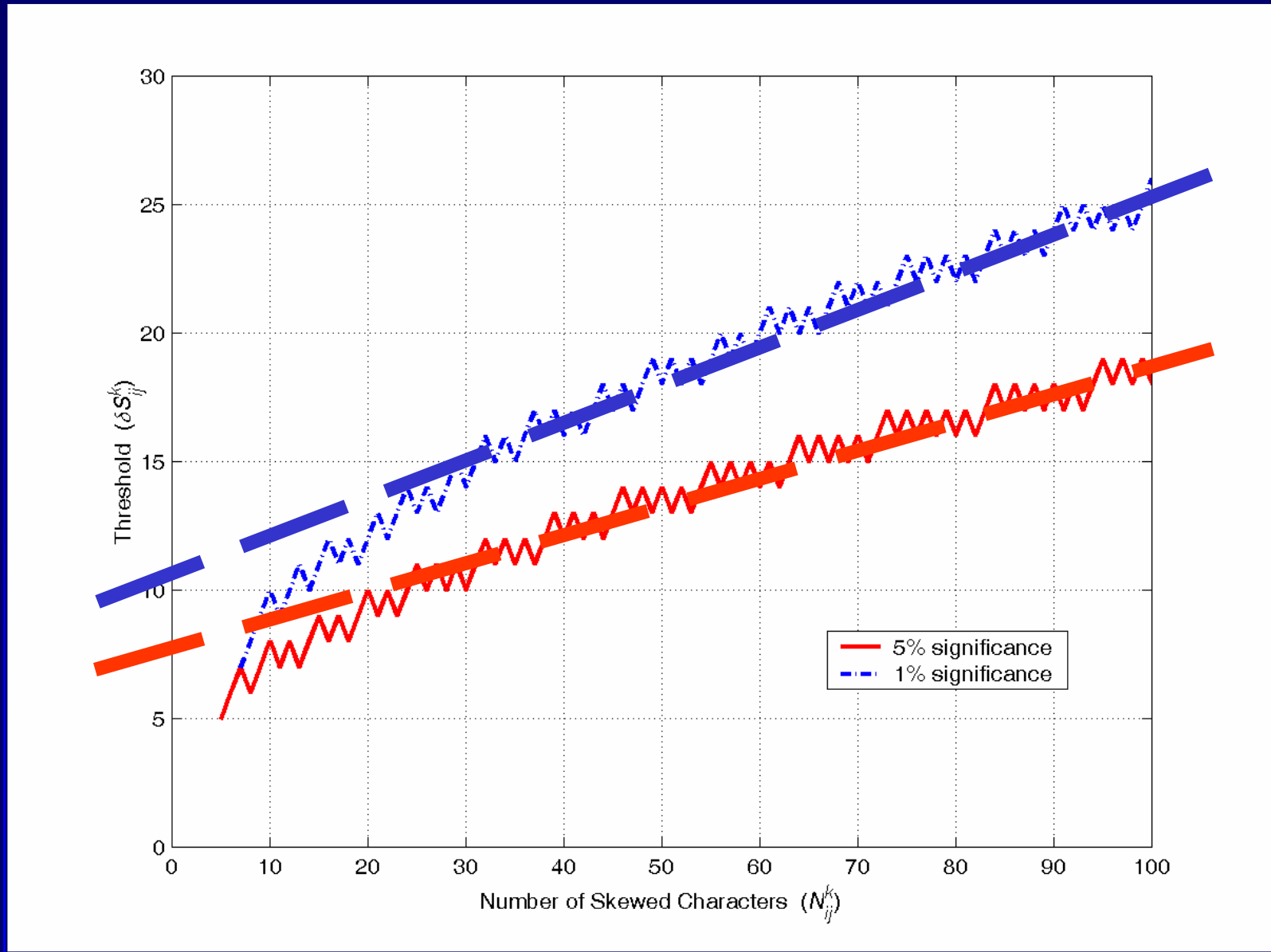


Distribution of Skewing — no contact



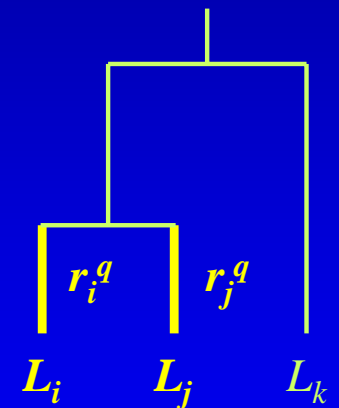
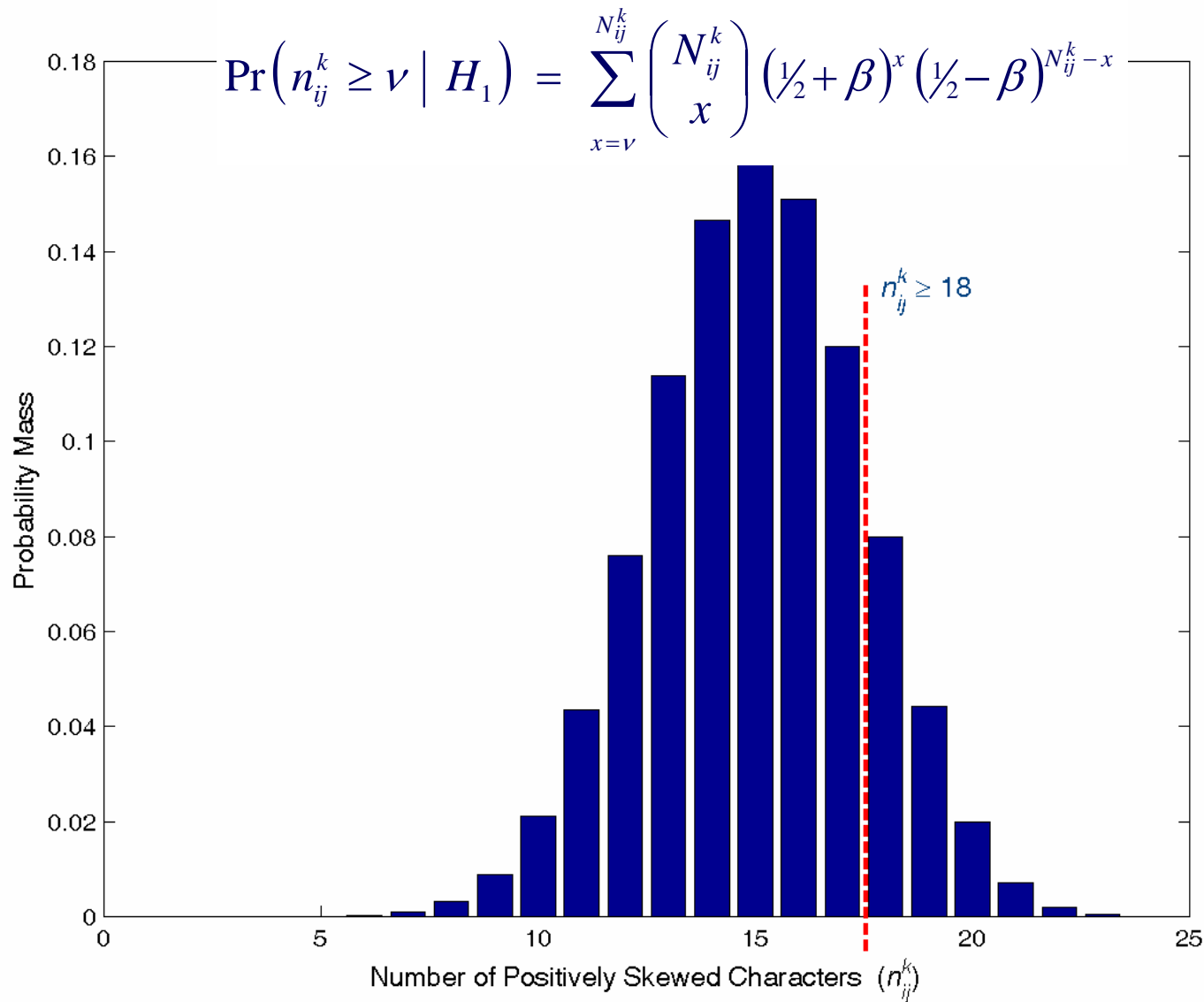


Skewing Required to Infer Contact



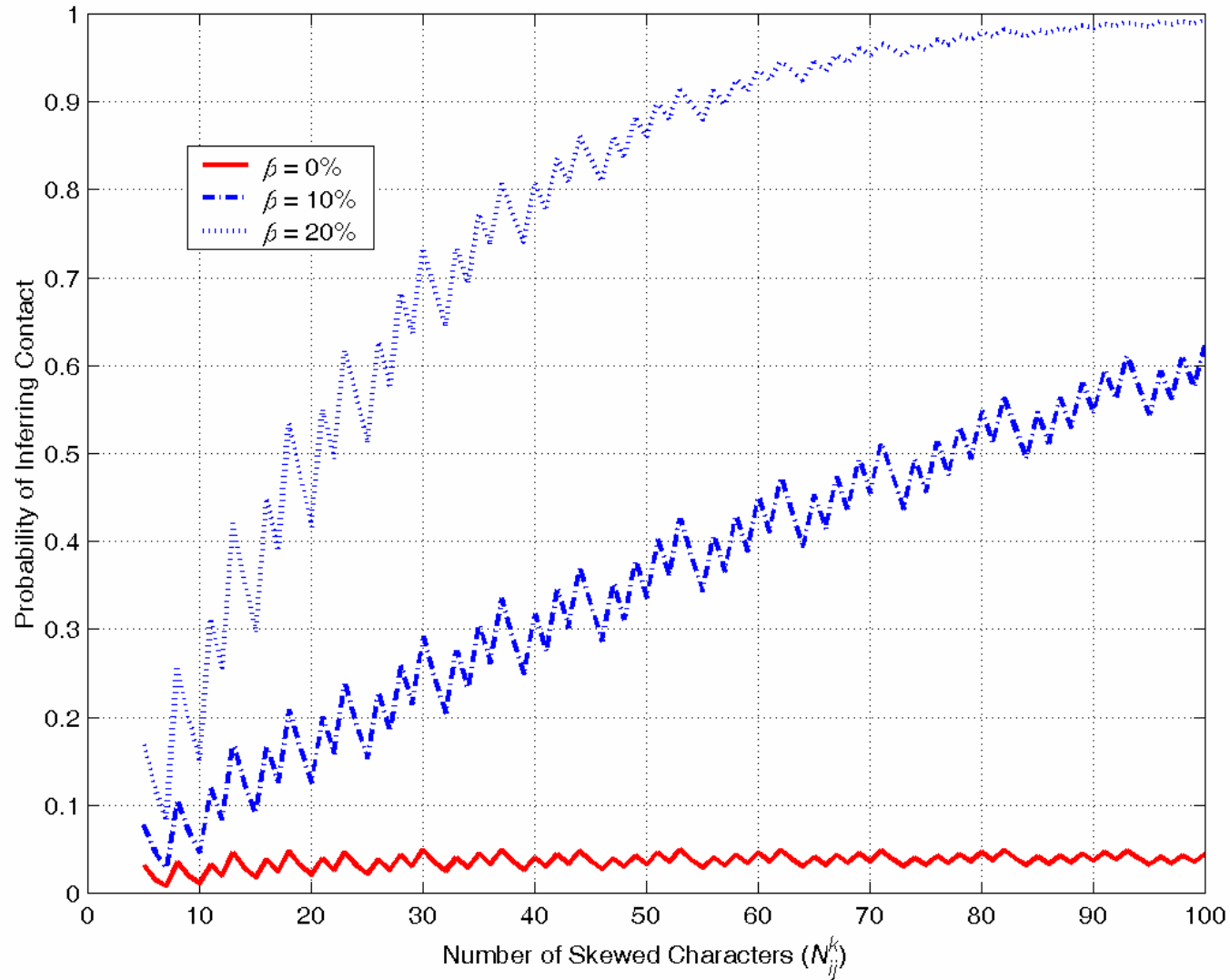


Distribution of Skewing — 10% contact





Probability of Correctly Detecting Contact



- Suppose the retention rate of is **90%** except for L_i with retention rate **85%**:

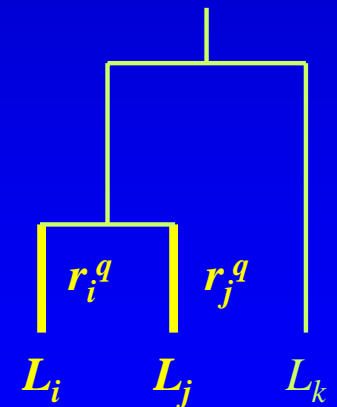
$$\Pr(\text{positive skewing}) = (0.85) \times (1 - 0.90) \times (0.90) \times (0.90)^2 = 6.19\%$$

$$\Pr(\text{negative skewing}) = (1 - 0.85) \times (0.90) \times (0.90) \times (0.90)^2 = 9.84\%$$

- The probability that a skewed character is **positively skewed** is

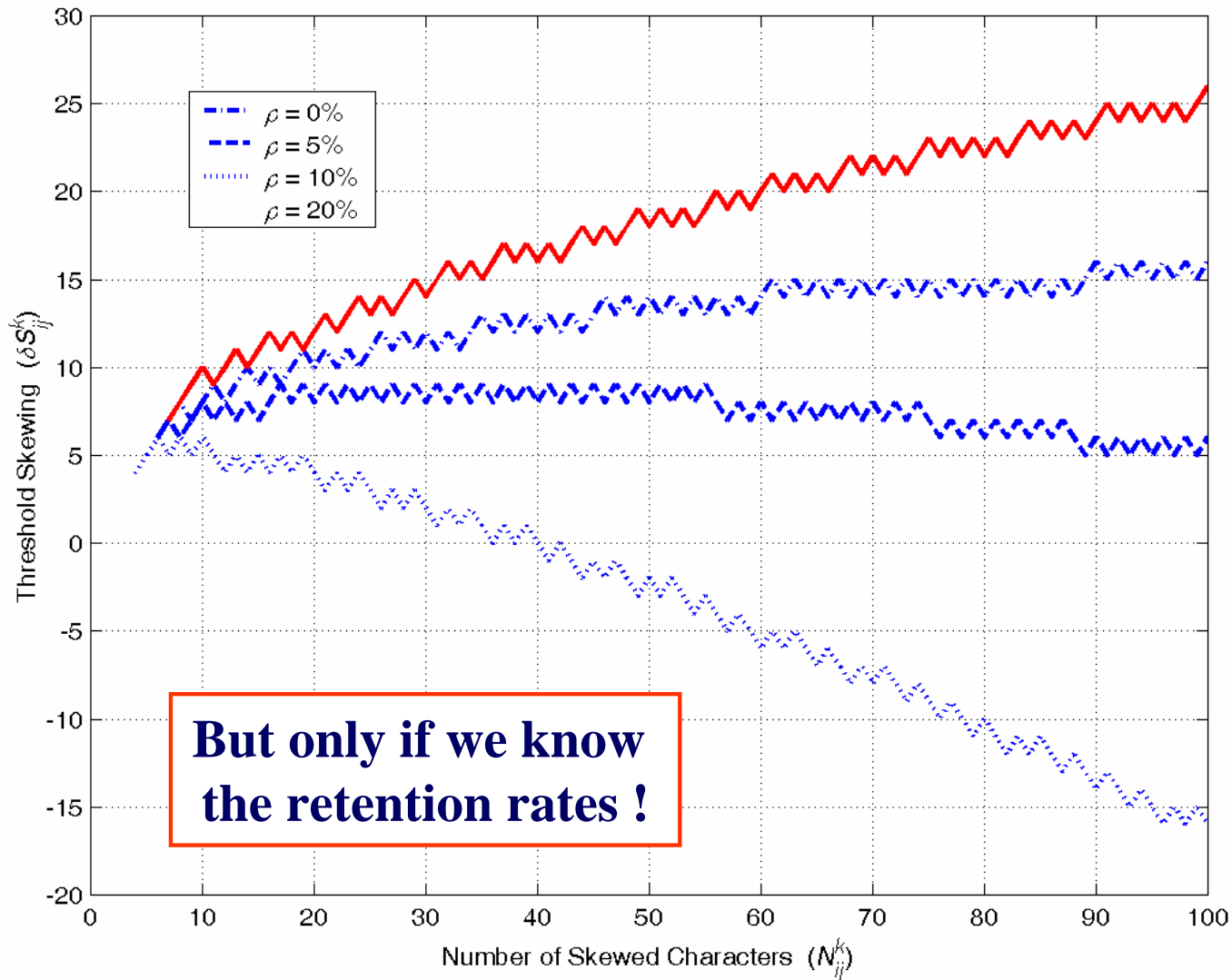
$$p_{ij}^k = \frac{6.19}{6.19 + 9.84} = 38.6\%$$

rather than 50% when the retention rate is homogeneous





Skewing Required to Infer Contact — non-homogeneous





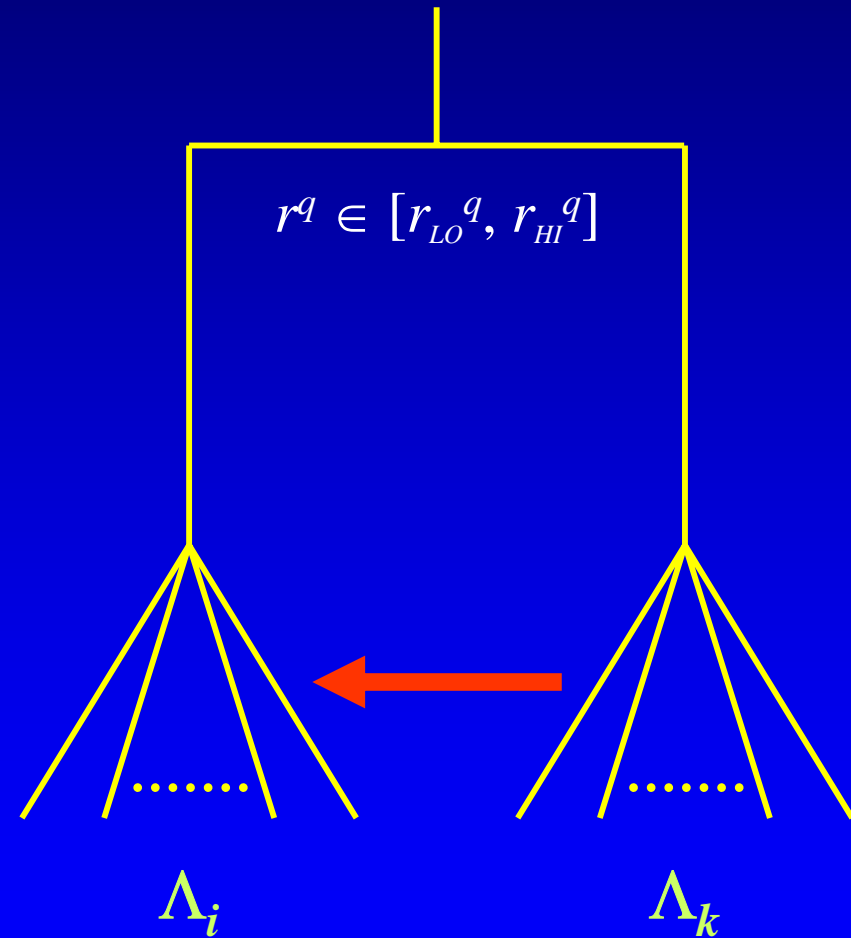
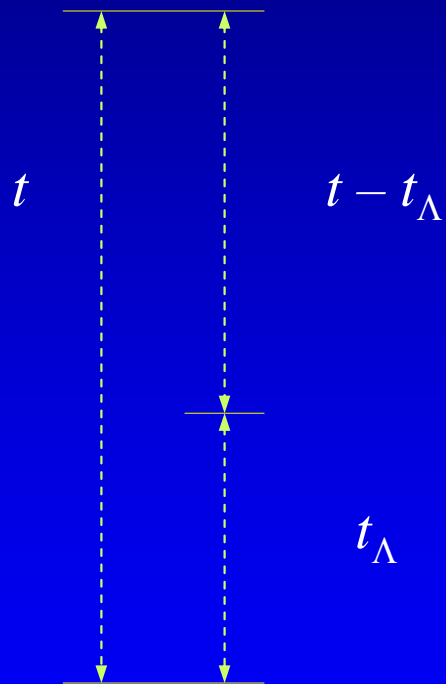
So where does this leave us?

- Hinnebusch's **concept** is fine:
 - borrowing **does** induce skewing in the lexical similarities
- Hinnebusch's **implementation** cannot be analyzed statistically:
 - must consider the numbers of **positively** and **negatively** skewed characters
- Skewing due to contact-induced **borrowing** is **indistinguishable** from skewing due to **non-homogeneous retention rate** ...
- ...unless we have an accurate estimate of the **retention rates**
- Can this approach be extended to sets of languages of **arbitrary number**?



Language Contact Among Multiple Languages — framework

time depths:



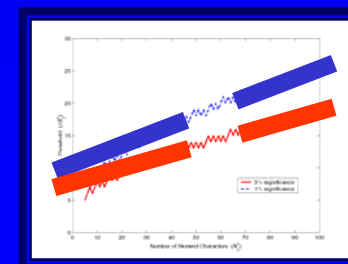
- Define the **aggregate skewing** of language $L_i \in \Lambda_i$ with respect to $L_k \in \Lambda_k$ as the average lexical skewing between L_i and **all** its siblings with respect to L_k :

$$\delta S_i^k = \frac{1}{l} \sum_{L_j \in \Lambda_i} (\delta S_{ij}^k)$$

- Aim to **detect contact** by implementing a decision rule of the form:

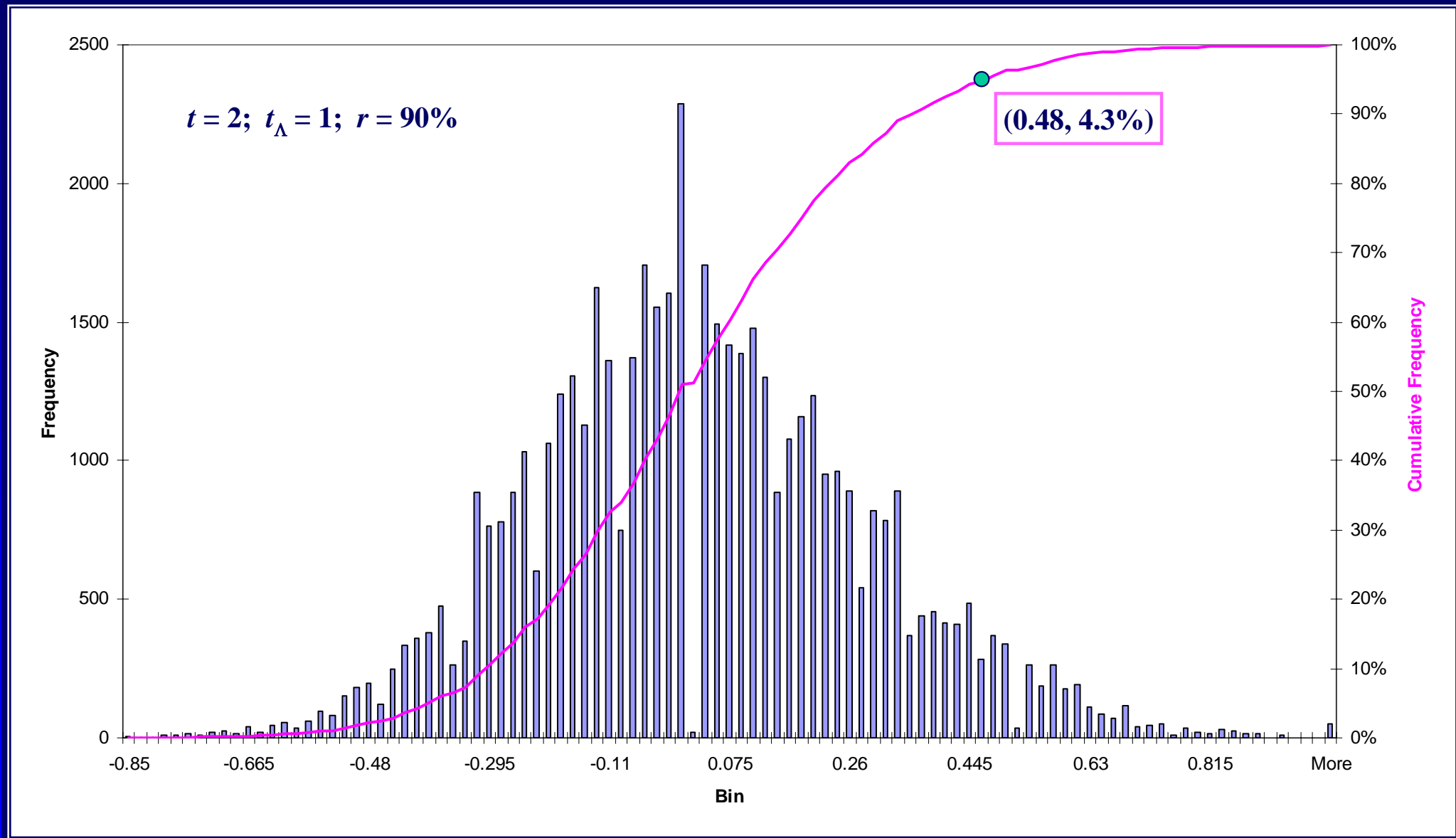
$$\delta S_i^k \begin{matrix} \text{contact} \\ \geq \\ < \\ \text{no contact} \end{matrix} \nu$$

- We will work with the ratio of the aggregate skewing, δS_i^k , and the **total number of skewed characters**, N_i^k .



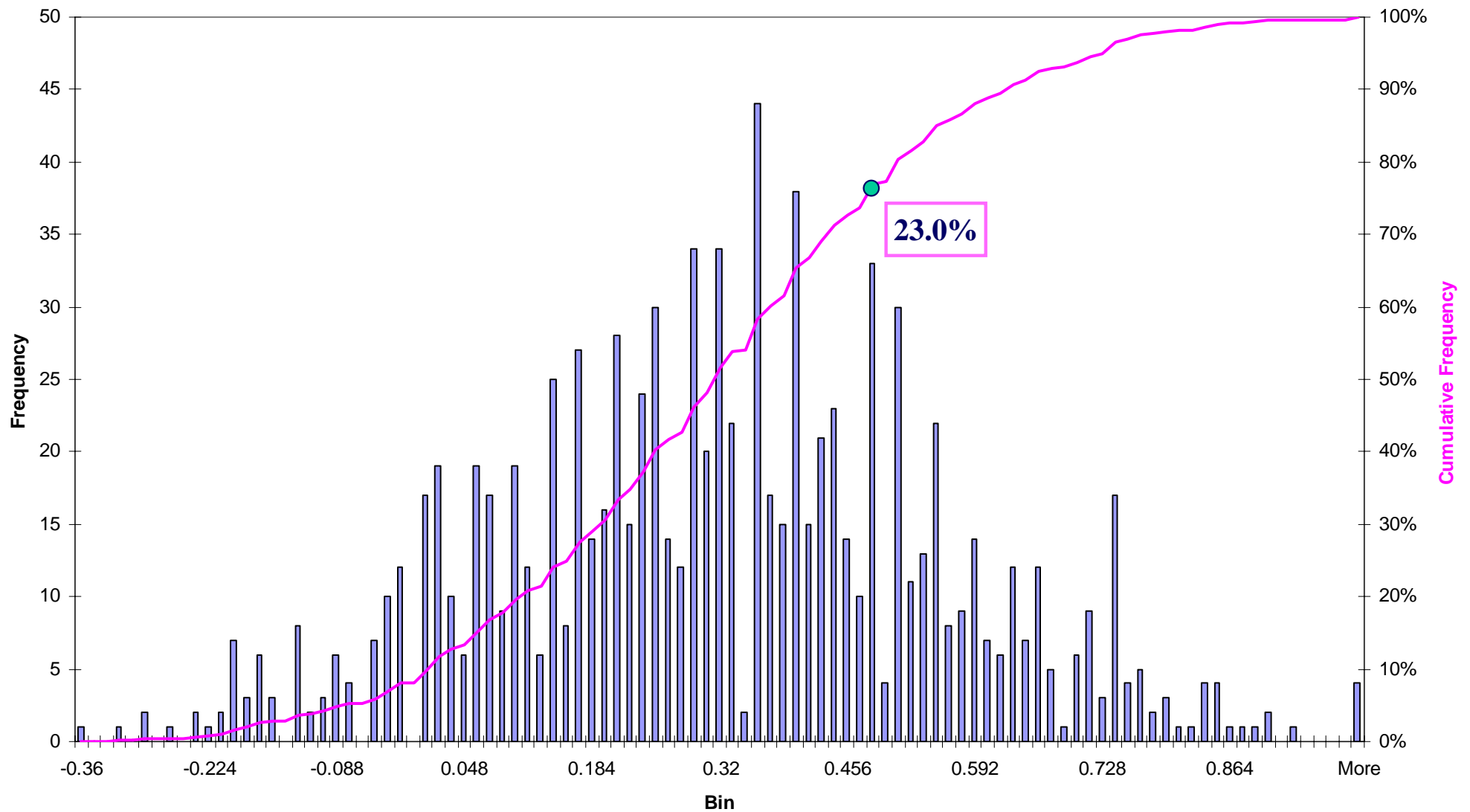


Distribution of Aggregate Skewing — no contact



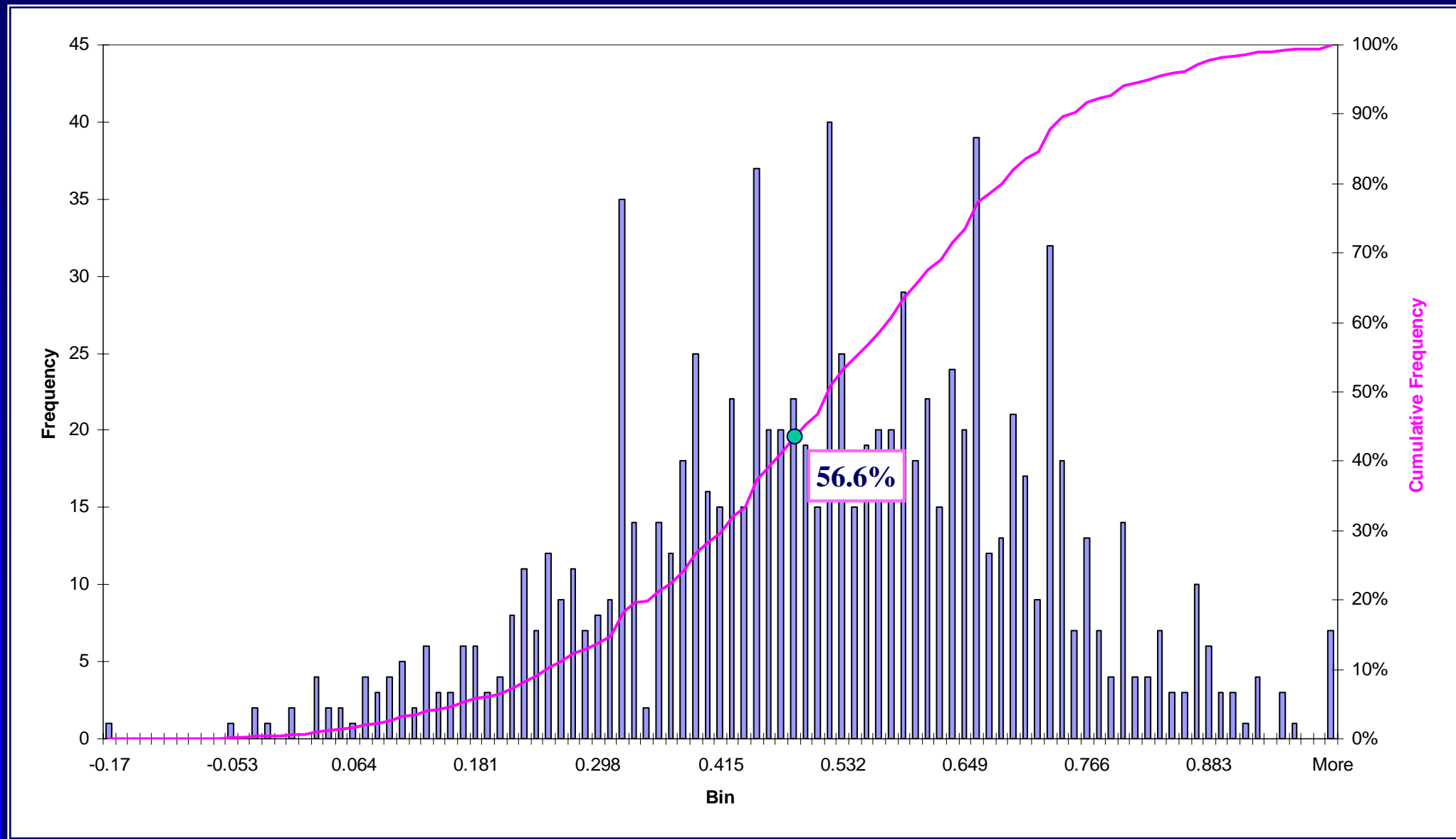


Distribution of Aggregate Skewing — 10% contact



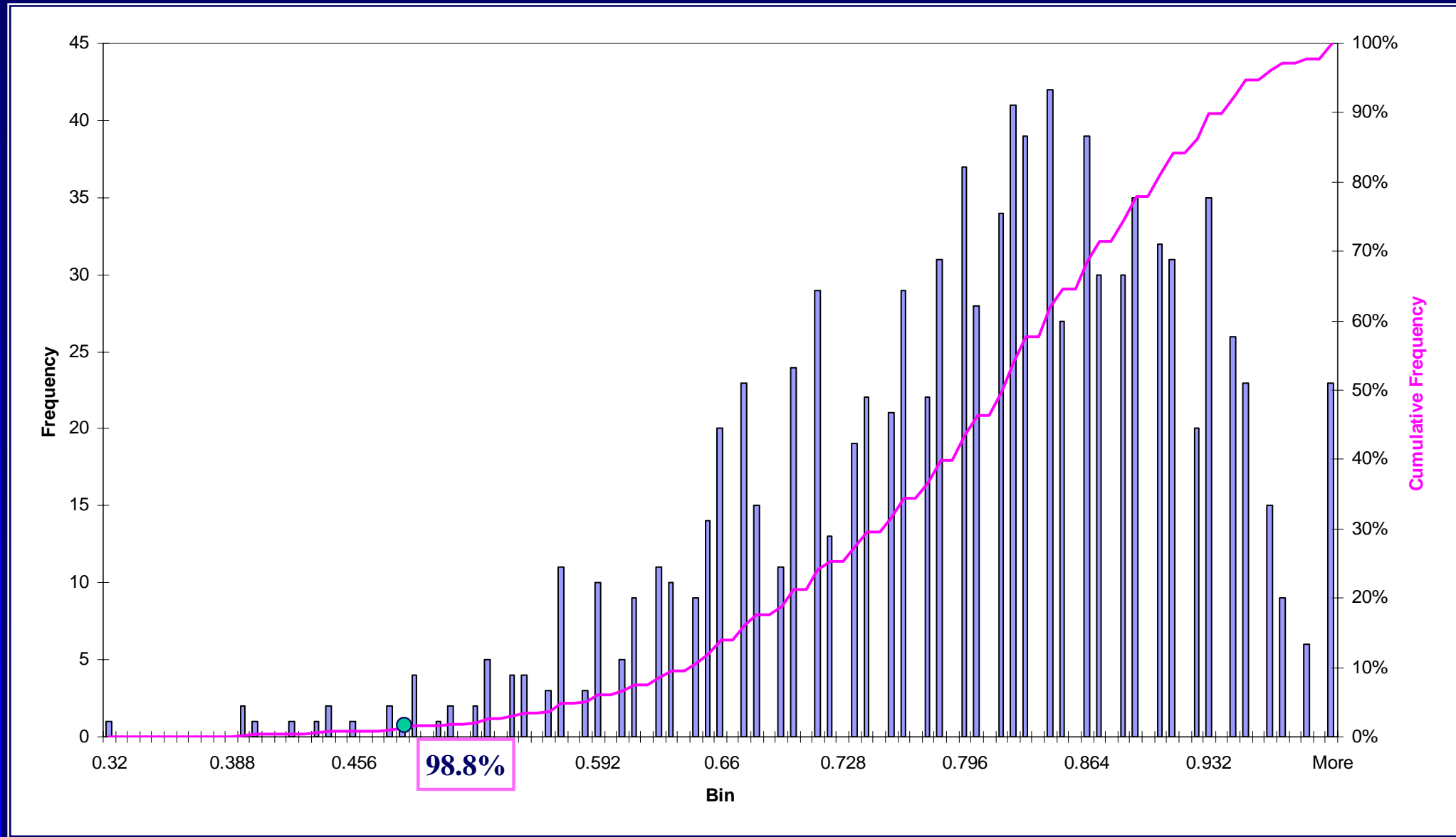


Distribution of Aggregate Skewing — 20% contact



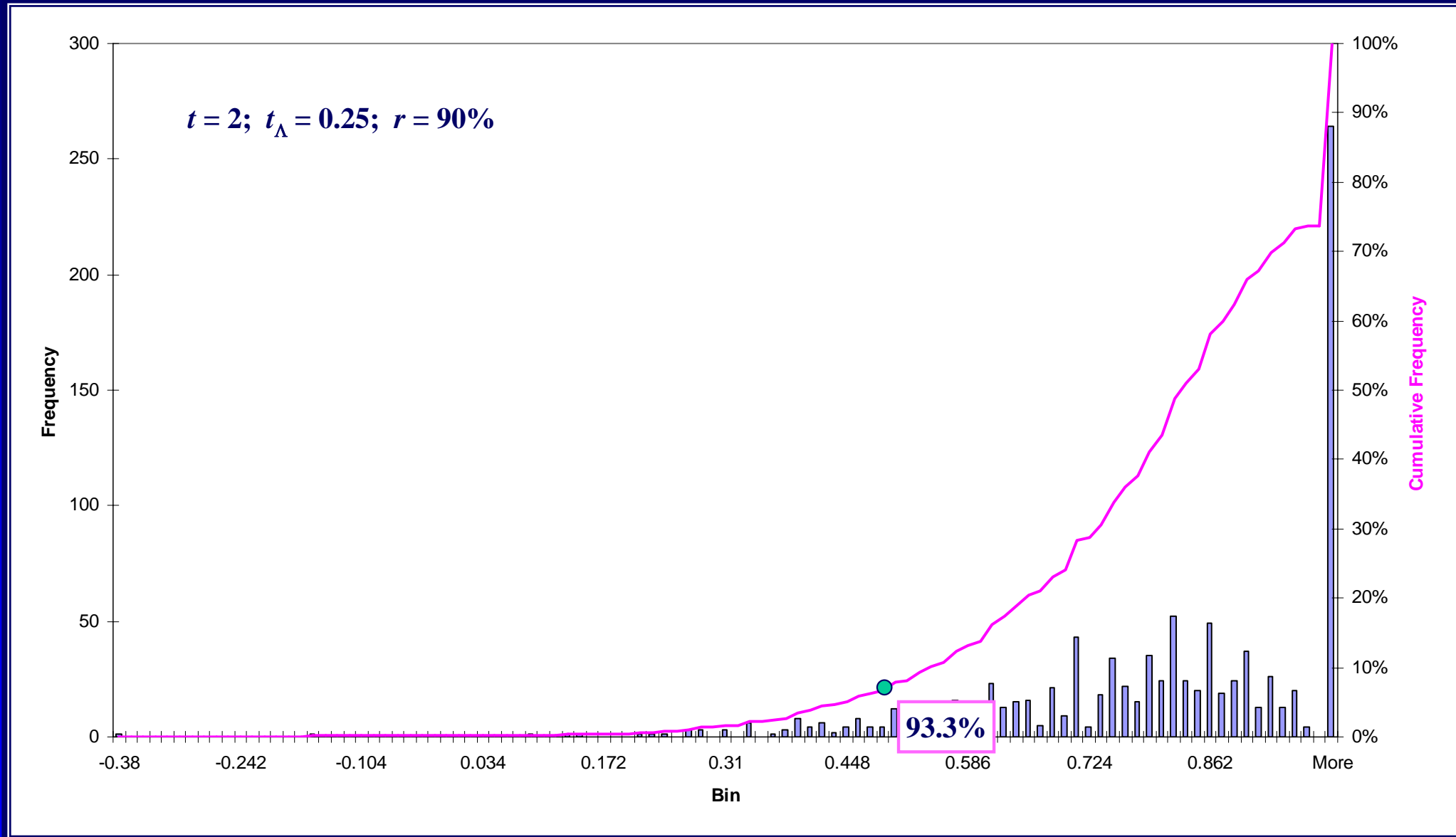
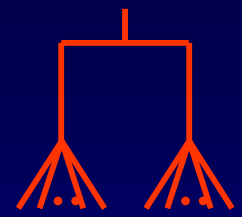


Distribution of Aggregate Skewing — 50% contact



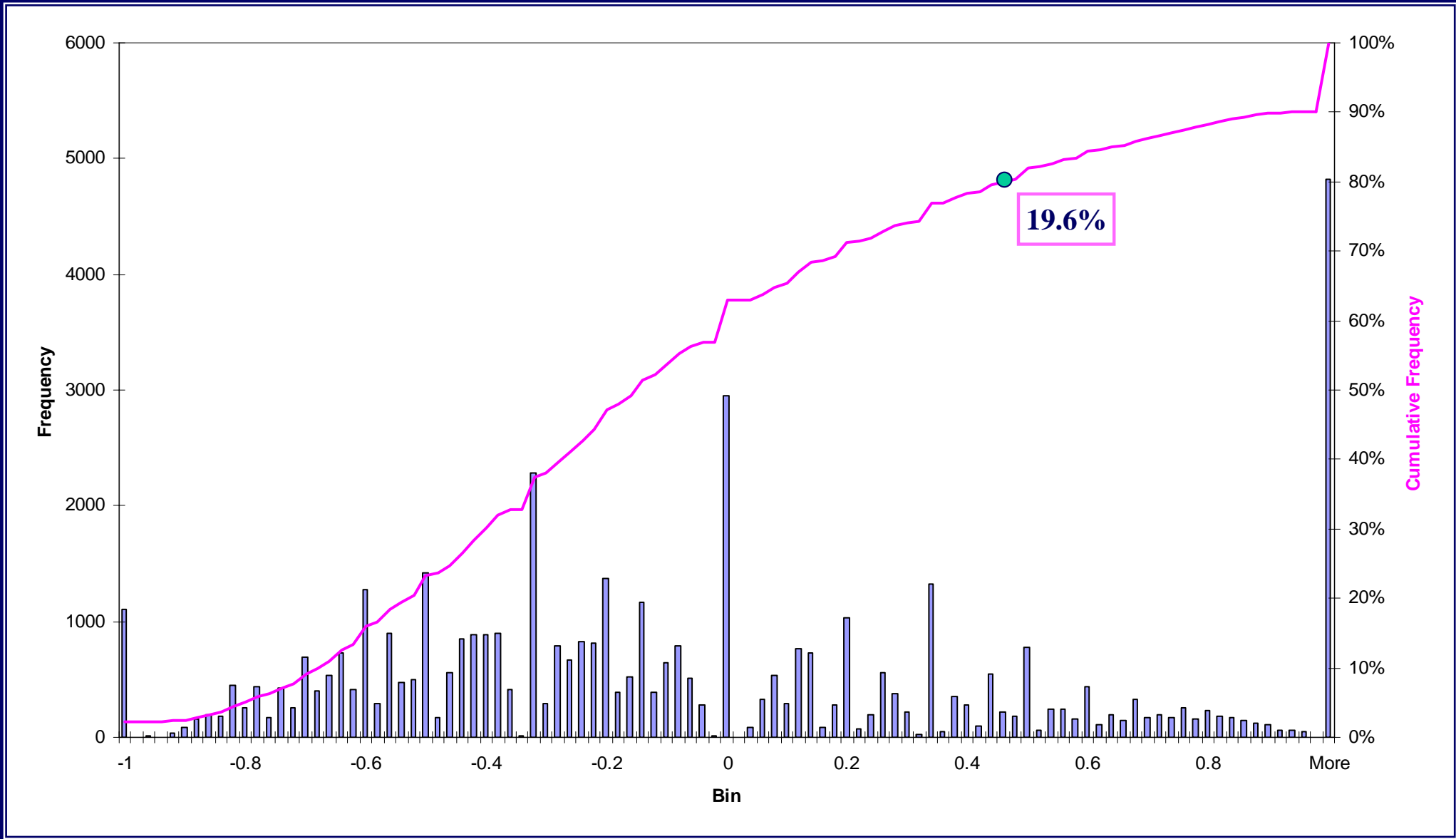


20% contact; 250 years of divergence



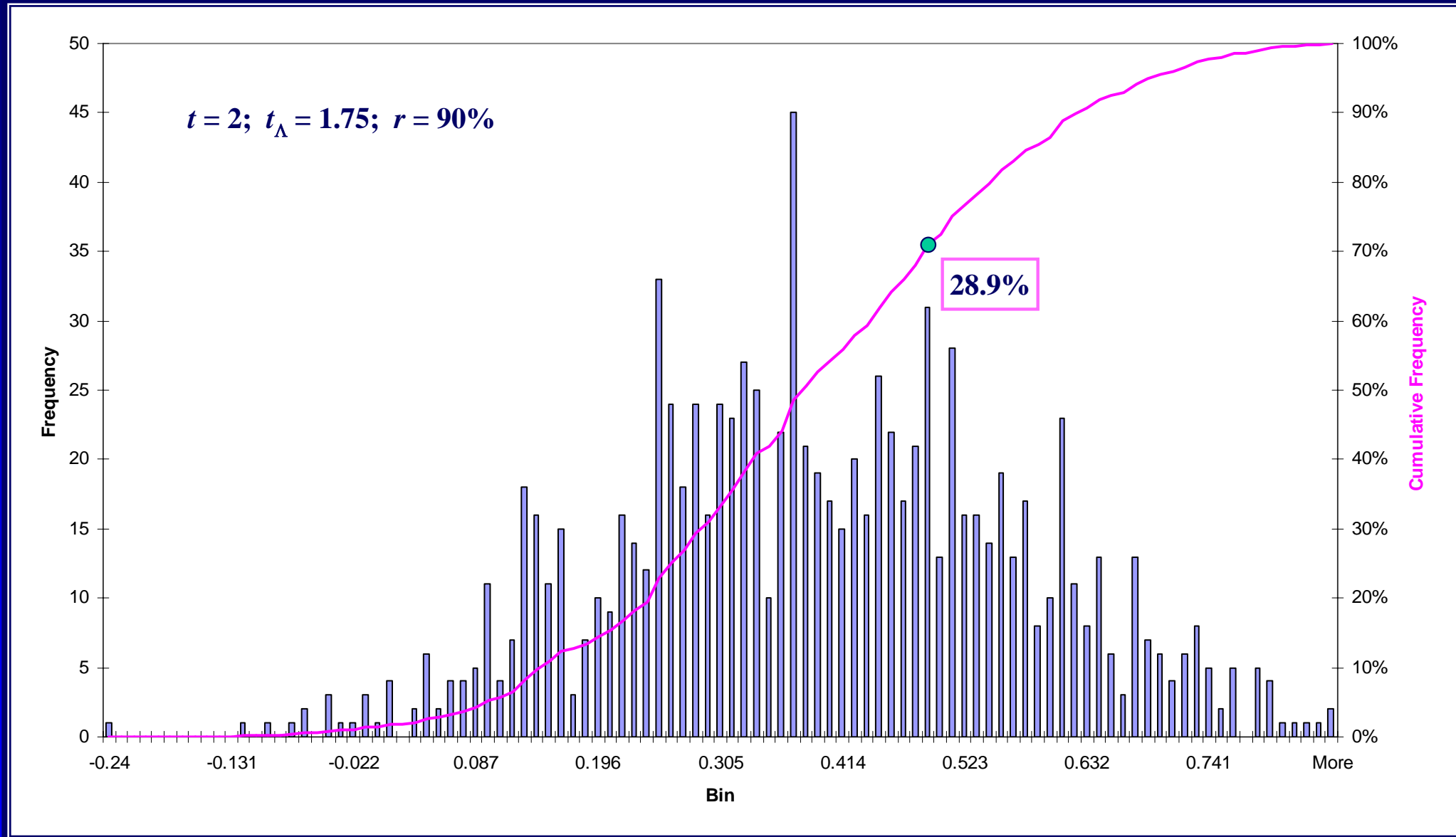
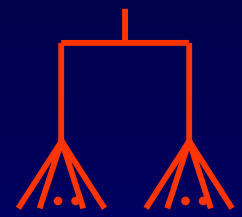


... but the **false alarm rate is increased**



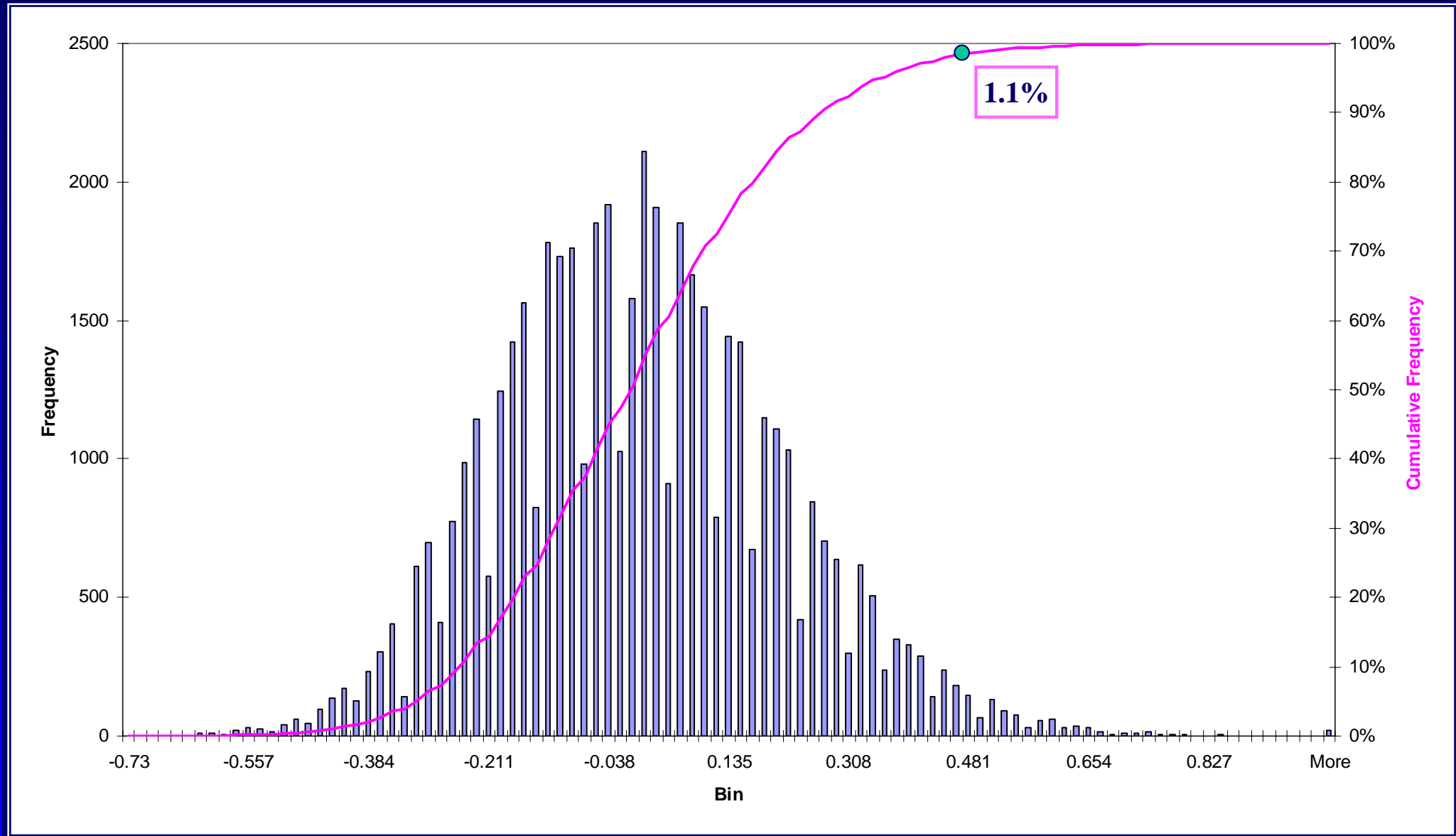


20% contact; 1,750 years of divergence



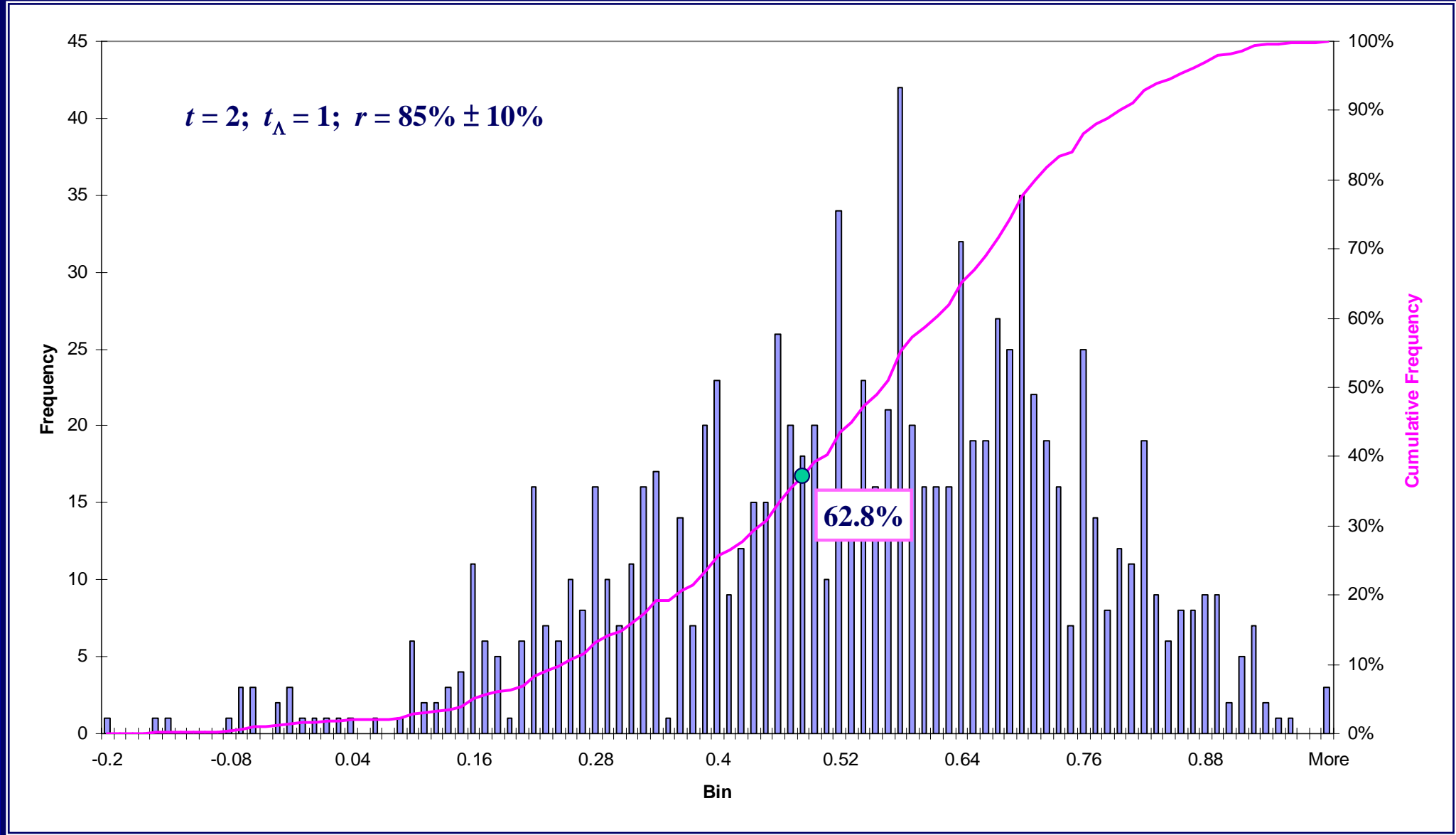


... but at least the **false alarm rate is decreased**





Non-homogeneous retention rate





Conclusion

- The Hinnebusch method can be extended to sets of languages of **arbitrary number**
- ... but the probability of error is **extremely sensitive to time depth**, but apparently **not very sensitive to retention rate**
- ... and the response to **multiple layers** of contact is unknown
- In summary, the method has only been demonstrated to be effective under **simple scenarios of contact** or when certain parameter values are **well known**
- We can still try **other detection methods**, e.g. ordered statistics