Analyticbridge Theorem: *If observations are assigned a random bin ID (labeled $1 \cdots k$), then the estimator $\hat{p}$ of any proportion computed on these k random bins satisfies*

$$P(\hat{p} \leq p_{(1)}) = \frac{1}{k+1} = P(\hat{p} \geq p_{(k)})$$

Also, for $m = 1, \cdots, k$, we have:

$$P(\hat{p} \leq p_{(m)}) = \frac{m}{k+1} = P(\hat{p} \geq p_{(k-m+1)})$$

Note that $p_{(1)} = \min p_j$ and $p_{(k)} = \max p_j, j = 1 \cdots k$. The $p_{(j)}$'s represent the order statistics, and p_j is the observed proportion in bin j. A proof of the result for $m = 1$ and $k = 2$ is as follows:

Proof : We must prove that $\hat{p} < p_{(1)}$ with probability $1/(k+1)$ and $\hat{p} > p_{(k)}$ with probability $1/(k+1)$. We will only prove the first assertion, because of the symmetry of the problem. Let us assume that the underlying distribution for $\hat{p}$ is $F(p)$, with $f(p) = dF(p)/dp$ denoting the density. So, $P(\hat{p} < p) = F(p)$ by notation.

$$\begin{aligned}
P(\hat{p} < p_{(1)}) &= P(\hat{p} < \min(p_1, p_2)) \\
&= \int\int P(\hat{p} < \min(p_1, p_2)|p_1, p_2) f(p_1) f(p_2) dp_1 dp_2 \\
&= \int\int_{p_1<p_2} P(\hat{p} < \min(p_1, p_2)|p_1, p_2) f(p_1) f(p_2) dp_1 dp_2 \\
&\quad + \int\int_{p_1>p_2} P(\hat{p} < \min(p_1, p_2)|p_1, p_2) f(p_1) f(p_2) dp_1 dp_2 \\
&= \int\int_{p_1<p_2} F(p_1) f(p_1) f(p_2) dp_1 dp_2 \\
&\quad + \int\int_{p_1>p_2} F(p_2) f(p_1) f(p_2) dp_1 dp_2 \\
&= \int_{p_2} f(p_2) \Big\{ \int_{p_1<p_2} F(p_1) f(p_1) dp_1 \Big\} dp_2 \\
&\quad + \int_{p_1} f(p_1) \Big\{ \int_{p_2<p_1} F(p_1) f(p_1) dp_2 \Big\} dp_1 \\
&= 2 \cdot \int_{p_2} f(p_2) \Big\{ \int_{p_1<p_2} \frac{1}{2} \frac{dF^2(p_1)}{dp_1} dp_1 \Big\} dp_2 \\
&= \int_{p_2} f(p_2) F^2(p_2) dp_2 = \int_{p_2} \frac{1}{3} \cdot \frac{F^3(p2)}{dp_2} dp_2 = \frac{1}{3} F^3(p_2) \Big]_{p_2=-\infty}^{p_2=+\infty} = \frac{1}{3}.
\end{aligned}$$

The general proof is similar but involves more complicated combinatorial arguments and the use of the Beta function. Note that the final result does not depend on the distribution F.

You can find more details in the book *Statistics of Extremes* by E.J. Gumbel, pages 58-59 (Dover edition, 2004) ∎

Values of k and m can be chosen to achieve the desired level of precision. This theorem is a fundamental result to compute simple, per-segment, data-driven, model-free confidence intervals in many contexts, in particular when generating predictive scores produced via logistic / ridge regression or decision trees / hidden decision trees (e.g. for fraud detection, consumer or credit scoring).

Application:

A scoring system designed to detect customers likely to fail on a loan, is based on a rule set. On average, for an individual customer, the probability to fail is 5%.

In a data set with 1 million observations (customers) and several metrics such as credit score, amount of debt, salary, etc. if we randomly select 99 bins each containing 1,000 customers, the 98% confidence interval (per bin of 1,000 customers) for the failure rate is (say) $[4.41\%, 5.53\%]$, based on the Analyticbridge Theorem with $k = 99$ and $m = 1$. Now, looking at a non-random bin with 1,000 observations, consisting of customers with credit score < 650 and less than 26 years old, we see that the failure rate is 6.73%. We can thus conclude that the rule *credit score* < 650 *and less than 26 years older* is actually a good rule to detect failure rate.

Indeed, we could test hundreds of rules, and easily identify rules with high predictive power, by looking at how far the observed failure rate (for a given rule) is from a standard confidence interval. This allows us to rule out effect of noise, and process and rank numerous rules at once.