

Rain Model for Events

Dorian Girod

October 22, 2026

1 Climatology

1.1 Estimation of the probability of rain

We start from the fact that rain exceeding a certain threshold t (for example 2 mm) on a given day D follows a Bernoulli distribution:

$$X_D \sim \text{Bernoulli}(p) \quad (1)$$

If we denote by $X_1, \dots, X_n$ the indicators of the n days of a given time window and consider that they are i.i.d., the number of rainy days in this window follows a binomial distribution:

$$K = \sum_{i=1}^n X_i, \quad K \sim \text{Bin}(n, p), \quad \Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (2)$$

The time window we consider contains all the days within ± 7 days of the event day, over the past 30 years. This means that $n = 15 \times 30 = 450$, and k , the observed value of K , is the number of days with rain exceeding our threshold t . In reality, consecutive days are not independent, since rainy days tend to come in series. The X_i are therefore positively correlated, which makes the true variance of $\hat{p}$ larger than the one we compute below, and our confidence intervals too narrow.

We know from intuition, and directly from the maximum likelihood estimator (MLE), that the best estimator of p is

$$\hat{p} = \frac{k}{n} \quad (3)$$

with k the number of rainy days exceeding t and n the total number of days in our sample.

Considering such a time window first allows us to use much more data than only the exact same day of the previous years. Second, it gives us a large enough n to use the Central Limit Theorem (CLT) to compute confidence intervals.

1.2 Moments

From the binomial distribution (2), by linearity of the expectation and independence of the X_i , we know that:

$$\mathbb{E}[K] = \sum_{i=1}^n \mathbb{E}[X_i] = np \quad \text{and} \quad \text{Var}(K) = \sum_{i=1}^n \text{Var}(X_i) = np(1-p) \quad (4)$$

Therefore, considering the estimator $\hat{p} = K/n$ (the random version of (3)), we compute the expected value and the variance of $\hat{p}$ using (4):

$$\mathbb{E}[\hat{p}] = \frac{1}{n} \mathbb{E}[K] = \frac{np}{n} = p \quad \text{and} \quad \text{Var}(\hat{p}) = \frac{1}{n^2} \text{Var}(K) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n} \quad (5)$$

1.3 De Moivre–Laplace theorem and CLT

We can approximate the binomial distribution by a normal distribution if these two conditions are met:

- $np \geq 10$
- $n(1 - p) \geq 10$

With $n = 450$, these two conditions are satisfied if $p \in [10/450, 1 - 10/450] \approx [2.2\%, 97.8\%]$. As p is unknown, we check these conditions in practice with $\hat{p}$.

Then, using the moments (4) and (5), we know that:

$$K \approx \mathcal{N}(np, np(1 - p)) \quad \text{and} \quad \hat{p} = \frac{K}{n} \approx \mathcal{N}\left(p, \frac{p(1 - p)}{n}\right) \quad (6)$$

This result gives the approximate distribution of our estimator $\hat{p}$, which we can standardize to build tests and confidence intervals more easily:

$$\frac{\hat{p} - p}{\sqrt{p(1 - p)/n}} \approx \mathcal{N}(0, 1) \quad (7)$$

1.4 Confidence interval and inference

From (7), for a confidence level $1 - \alpha$ we have:

$$\Pr\left(|\hat{p} - p| \leq z_{1-\alpha/2} \sqrt{\frac{p(1 - p)}{n}}\right) \approx 1 - \alpha \quad (8)$$

where $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution ($z_{1-\alpha/2} \approx 1.96$ for $\alpha = 0.05$). From now on, we write $z = z_{1-\alpha/2}$.

The Wald confidence interval replaces p by $\hat{p}$ in the square root of (8), so that p can be isolated directly: $\text{CI}_{\text{Wald}} = \hat{p} \pm z\sqrt{\hat{p}(1 - \hat{p})/n}$. The problem with this interval is that it comes from two approximations: the CLT, which only works under the two conditions we saw, and the use of $\hat{p}$ instead of p in the standard error. Because of this second approximation, the interval can go below 0 or above 1, and collapses to $[0, 0]$ when $k = 0$ (or to $[1, 1]$ when $k = n$). To compensate for this second problem, we preferred the Wilson interval.

1.5 Wilson interval

The Wilson interval still relies on the CLT, but it removes the second approximation: it keeps p in the standard error. The Wilson interval is therefore more accurate (even if asymptotically equivalent to the Wald interval), is asymmetric around $\hat{p}$, stays within $[0, 1]$ and does not degenerate for extreme cases ($k = 0$), but requires more computations to obtain the interval. For each candidate value p_0 , we test $H_0 : p = p_0$ using (8), and the interval contains all the values of p_0 that are not rejected:

$$\left\{ p_0 \in [0, 1] : |\hat{p} - p_0| \leq z\sqrt{\frac{p_0(1 - p_0)}{n}} \right\} \quad (9)$$

The inequality in (9) could be difficult to solve analytically, but luckily this is not the case for a binomial model. Squaring the inequality in (9) gives a second-degree polynomial inequality in p_0 . Since the leading coefficient is positive, the inequality holds between the two roots of the polynomial, which are the bounds of the interval.

The Wilson interval is therefore defined as

$$\text{CI}_{\text{Wilson}} = \left[\frac{\hat{p} + \frac{z^2}{2n} - z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \frac{\hat{p} + \frac{z^2}{2n} + z \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}} \right] \quad (10)$$

We denote by p_- and p_+ its lower and upper bounds, so that $\text{CI}_{\text{Wilson}} = [p_-, p_+]$.

2 Forecast

The problem with the climatology inference is that it only considers past events and **not** the current state of the atmosphere and the current weather. We therefore need to include the forecast, which is much more precise if the event is soon but loses its value after roughly 2 weeks. The main idea here is to trust climatology more if the event is far away, and to put more emphasis on the forecast if the event is close to the current date.

2.1 Weight function

We therefore consider a weight function $w(h)$, with h the number of days between today and the event, and define our final estimate $\tilde{p}$ as follows:

$$\tilde{p} = w(h) p_f + (1 - w(h)) \hat{p}, \quad w(h) = \max\left(0, 1 - \frac{h}{16}\right) \quad (11)$$

with p_f the probability given by the forecast and $\hat{p}$ the probability given by the climatology (3). The number 16 comes from the fact that the API we use stops making forecasts after 16 days; it is not a statistical choice.

As we can see, we defined a linear weight for now; improving it is a possible improvement for the future. A good method could be to use the Brier score, choosing for each horizon h the weight that minimizes the squared error between the given probability and what really happened.

Another limitation of the model is that p_f represents the probability of any rain (even 0.1 mm of rain), whereas $\hat{p}$ represents the probability of rain exceeding the threshold t . As we want to interpret $\tilde{p}$ in (11) the same way as $\hat{p}$, we tend to overestimate $\tilde{p}$, because p_f drives $\tilde{p}$ higher than it should be. A correction using conditional probability could be a good idea:

$$\Pr(\text{rain} > t) = \Pr(\text{rain} > t \mid \text{rain} > 0) \Pr(\text{rain} > 0) \approx \hat{q} p_f, \quad \hat{q} = \frac{k}{k_0},$$

with k_0 the number of days with any rain in the time window. We could then replace p_f by $\hat{q} p_f$ in (11).

2.2 Confidence interval

We decide to assume that p_f is deterministic and has no variance (meaning that the probability given by the API is the true probability). This is a strong assumption and is not expected to be true, but we make it for the sake of simplicity. Applying the weights of (11) to the Wilson bounds (10), we compute the confidence interval as:

$$\text{CI}_{\text{final}} = [p_{\text{lo}}, p_{\text{hi}}] = w p_f + (1 - w) [p_-, p_+] \quad (12)$$

where $w = w(h)$ is the weight for the given horizon h . Note that for $h = 0$, we have $w = 1$, so CI_{final} reduces to the single point p_f : because of our assumption, the verdict (17) can then never be a close call.

3 Decision

3.1 Loss without protection

On the event day D , the rain is $X_D \sim \text{Bernoulli}(p)$, with p estimated by $\tilde{p}$ from (11), and the organizer's loss L is:

$$L = \lambda R \cdot X_D = \begin{cases} \lambda R & \text{with probability } p, \\ 0 & \text{with probability } 1 - p, \end{cases} \quad (13)$$

where R is the total revenue of the event and λ the share of the revenue lost in case of rain above the threshold t .

It means that the expected value of the loss is:

$$\mathbb{E}[L] = \lambda R p \quad (14)$$

3.2 Loss with protection

We suppose here that the protection is perfect, meaning that for a certain cost C , paid with certainty, the loss is brought back to 0.

3.3 Decision threshold

If the organizer compares the expected losses, they protect themselves when the protection is cheaper than the expected loss (14). In other words, when:

$$C < p \lambda R \quad \Longleftrightarrow \quad p > \frac{C}{\lambda R} \quad (15)$$

The threshold is then defined by:

$$\boxed{p^* = \frac{C}{\lambda R}} \quad (16)$$

which is a classic cost-loss model, with p^* being the cost-loss ratio.

3.4 Interpreting p^*

p^* is the price of the protection as a percentage of what we lose in case of a rainy day. The rule (15) is therefore pretty easy to understand: “protect yourself if the probability of rain exceeds the price of the protection divided by the potential loss.”

Here an important assumption is made regarding the organizer: **risk neutrality**. This assumption is often not realistic, in particular for small organizers. Indeed, a potential loss of 12 000 CHF with a probability of 25 % may not be judged the same way as a certain loss of 3 000 CHF, because the 12 000 CHF loss could be too damaging. This could be improved by modeling the risk aversion of the organizer, or by adding a constraint on the maximum acceptable loss.

3.5 Dealing with the uncertainty of p

If p were known exactly, there would be only two cases: $p > p^*$ or $p < p^*$. But p is estimated by $\tilde{p}$, with the interval $\text{CI}_{\text{final}} = [p_{\text{lo}}, p_{\text{hi}}]$ from (12). Depending on where the threshold p^* (16) lies relative to this interval, the verdict is different:

$$\text{verdict} = \begin{cases} \textbf{Protect} & \text{if } p_{\text{lo}} > p^* \quad (\text{the whole interval is above the threshold}) \\ \textbf{Don't protect} & \text{if } p_{\text{hi}} < p^* \quad (\text{the whole interval is below the threshold}) \\ \textbf{Close call} & \text{if } p_{\text{lo}} \leq p^* \leq p_{\text{hi}} \quad (\text{the threshold is inside the interval}) \end{cases} \quad (17)$$

3.6 Other assumptions

Other assumptions have been made for the sake of simplicity, and they are all possible ways to improve our model. Statistics is not about finding the truth, but about getting as close as possible to it :)

Assumption	Reality
Independent days	Consecutive days are correlated (rainy periods), so the effective sample size is smaller than n and the confidence interval is too narrow.
Binary rain (above or below t)	The loss depends on the intensity and duration of the rain.
Fixed λ	The loss also depends on when the rain falls (in the morning or during the concert).
Perfect protection	A tent does not save everything: some residual loss usually remains.
Single day	A three-day festival requires combining several days, which are correlated with each other.

Table 1: Assumptions of the model