Maximization of likelihood

- We wish to model the distribution $P(Z)$ using data $\{z_1, z_2, ..., z_M\}$
- Hypothesis : $P(Z)$ takes a parametric form $f(z; \theta_1, \theta_2, \theta_3)$
- Identify the parameters that maximize the likelihood of the observed data :

$$\underset{\theta}{\text{maximize}} \prod_{i=1}^{M} f(z_i; \theta_1, \theta_2, \theta_3)$$

«Chi-square Goodness of Fit Test» (by bootstrap)

1. We wish to validate the hypothesis that $P(Z)$ is really $f(z;\theta_0^*)$ with the data $\{z_1^0, z_2^0, ..., z_M^0\}$
2. Calibrate new parameters $\theta_1^*, \theta_2^*, \ldots, \theta_J^*$ from J sets of M data points drawn according to $f(z;\theta_0^*)$
3. Pick a set of intervals $]a_k,\ a_{k+1}]$ with $k = 1, ..., K$ covering the support of $f(z;\theta^*)$
4. Calculate the χ^2 statistic of each set of data according to $f(z;\theta_j^*)$:

$$X_j^2 := \sum_{k=1}^{K} \frac{(P_{\theta_j^*}(a_k \leq Z \leq a_{k+1}) - \hat{P}_{\mathbf{z}_{1:M}^j}(a_k \leq Z \leq a_{k+1}))^2}{P_{\theta_j^*}(a_k \leq Z \leq a_{k+1})}$$

5. Compute the frequency «P-value» of the event $X_j^2 \geq X_0^2$
6. If the frequency is too small ($< 5\%$), reject the hypothesis that $f(z;\theta_0^*)$ is the underlying model

Exemple @Risk : Return on investment I

- Model the probability distribution characterizing the future return on investment using the software @Risk.
- Historical data are presented in the Financial Asset Modeling Excel file
- Justify the type of model proposed by @Risk

Insurance against major disasters

- Consider an insurance company that wishes to compute the annual premium for an insurance based on the total value of a property in case of a major disaster.
- The insurance contract in question states that the reimbursed value would be equal to the value of the property as evaluated by an expert the day before the disaster.
- The company considers that three criteria influence the value of a property : distance from downtown area (km), size of the land (square foot) and the presence of a swimming pool.
- Recent data is used to model the houses' values.
- The probability of a major disaster (in the following year) is estimated using data from the last 31 years for similar regions.

Question 1

4. Starting from the (incomplete) Home Insurance Excel file, apply a linear regression to identify a distribution for the property's value that is conditioned on the knowledge of the three influential variables.
5. Check how well calibrated the distribution is for a confidence interval of 80% on the test data.
6. Under the hypothesis that the probability of a disaster occuring in the next year is initially considered distributed according to the Beta(1,1) distribution, use the data from the Excel file to establish the a posteriori belief of this probability.

Question 2

- A client calls to learn about the annual premium that he would need to pay for this insurance. He resides 10km from the downtown area, his land is 5000 square foot and he doesn't have a pool.
- What is the lowest annual premium that the company can offer (without incurring a loss) ?

Relation to other variables : linear regression

- We wish to model the conditional prob. $P(Z|X,Y)$ with the data $\{(x_1,y_1,z_1),...,(x_M,y_M,z_M)\}$
- Hypothesis :
  1. $Z = \theta_0 + \theta_1 X + \theta_2 Y + \epsilon$
  2. The random variable ϵ is independent from X & Y
  3. ϵ concentrates around 0, e.g., $E[\|\epsilon\|^2]$ is small
- Identify the parameters that minimizes the empirical concentration as measured with the data :

$$\underset{\theta}{\text{minimize}} \frac{1}{M}\sum_{i=1}^{M}\|\epsilon_i\|^2 \quad \equiv \quad \underset{\theta}{\text{minimize}} \frac{1}{M}\sum_{i=1}^{M}\|z_i - \theta_0 - \theta_1 x_i - \theta_2 y_i\|^2$$

- Identify a model for $P(\epsilon)$ using $\{\epsilon_1,\epsilon_2,...,\epsilon_M\}$ where $\epsilon_i = z_i - \theta_0 - \theta_1 x_i - \theta_2 y_i$