60615A : Decision Analysis
Session 3 - Probabilistic Modeling II

Prof. Carolina Osorio
Department of Decision Sciences
HEC Montréal

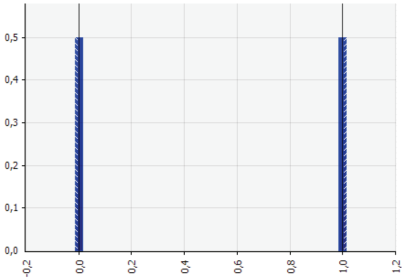
Outline

- Choosing a parametrized probability distribution function
- Parameter fitting
- Bayesian approach

Why use parametrized distribution functions

- Certain forms are natural choices to represent the uncertainty of certain types of physical processes.
- The selection of a small number of parameters allows for the definition of a density measure over a continuous space.
- The parameters can be estimated from available data.
- You are still responsible for the form that you choose to use.

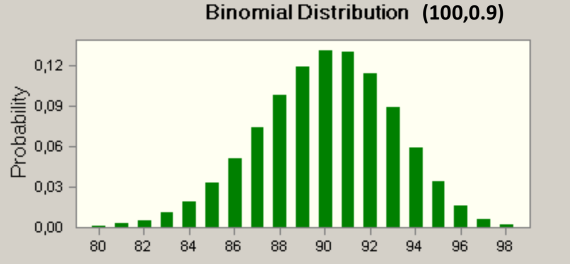
Bernoulli distribution



$$P(Z = z; p) = \begin{cases} p & \text{if } z = 1 \\ 1 - p & \text{if } z = 0 \end{cases}$$

- Represents the fact that an event happens or not.
- Example : the next client will purchase at least one product.

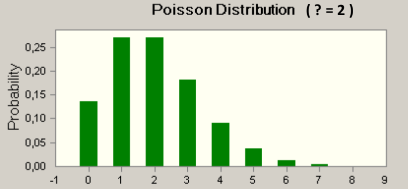
Binomial distribution



$$P(Z = k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Represents the number of times a random event (with prob. p) will take place in the context of n experiments.
- Example : how many clients will purchase a product in a sample of 100 customers.

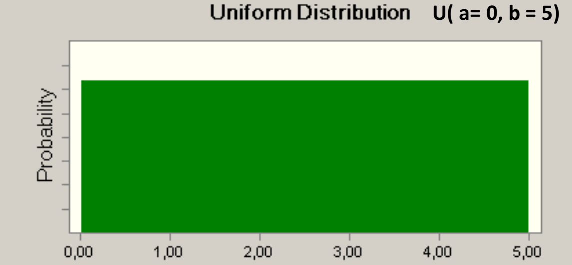
Poisson distribution



$$P(Z = k; \lambda) \quad = \quad \frac{\lambda^k}{k!} e^{-\lambda}$$

- Represents the prob. that k events take place within a period of time where the expected (i.e., average) quantity is λ and the time between two events follows an exponential distribution
- Example : the number of clients that present themselves to the store between 5-6pm.

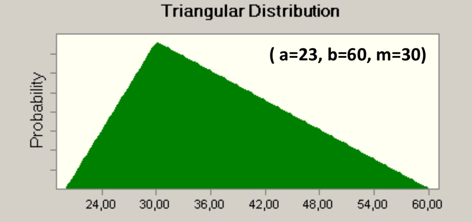
Uniform distribution



$$f(z; a, b) \quad = \quad \mathbb{1}\{z \in [a, b]\} \cdot 1/(b - a)$$

- Represents complete uncertainty with respect to the position of Z except for the fact that Z is between a & b
- Example : a delivery will take place between 10-11am

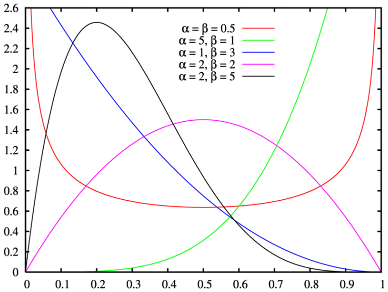
Triangular distribution



$$f(z; a, b, m) \;=\; \mathbb{1}\{z \in [a, b]\} \cdot \min\left(\frac{2(z-a)}{(m-a)(b-a)}, \frac{2(b-z)}{(b-m)(b-a)}\right)$$

- Frequently used when the minimum, the maximum and the most likely value of the Z variable are known (see also Beta distribution).
- Example : the project should be completed between 20 and 60 days from now, but expected in 30 days.

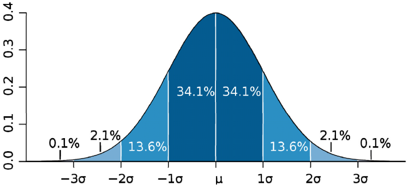
Beta distribution



$$f(z;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} z^{\alpha-1}(1-z)^{\beta-1}$$

- Used when the min $= 0$, the max $= 1$ and the most likely value $= \alpha/(\alpha+\beta)$
- Unlike with the triangular distribution, we can represent the level of concentration around the mode using $\alpha+\beta$.
- Example : the proportion of clients that will adopt a product after surveying a subset of them.

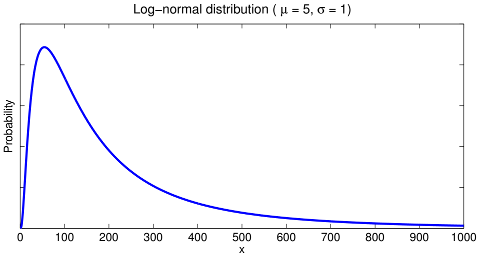
Normal distribution



$$f(z;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(z-\mu)^2}{2\sigma^2})$$

- According to the central limit theorem (CLT), the distribution of a sum of random variables (same mean and variance) converges to a normal distribution.

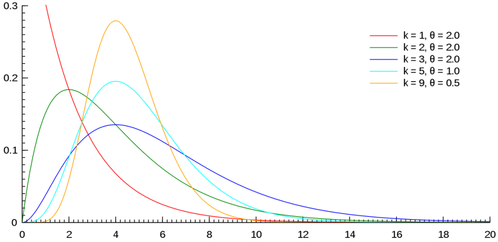
Log-normal distribution



$$f(z;\mu,\sigma) \quad = \quad \frac{1}{\sqrt{2\pi\sigma^2 z^2}} \exp\left(-(\ln(z)-\mu)^2/(2\sigma^2)\right)$$

- The logarithm of a log-normal random variable follows the normal distribution.
- Represents well the product of random values (CLT).
- Example : the cumulated returns of a stock.

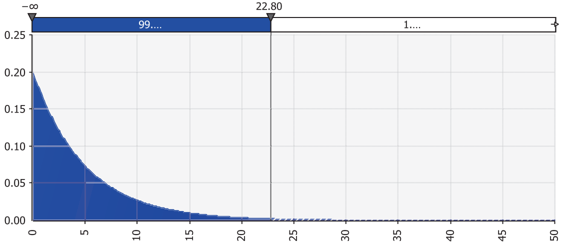
Gamma distribution



$$f(z; k, \theta) = \frac{z^{k-1} \exp(-z/\theta)}{\Gamma(k)\theta^k}$$

- The amount of time before the k^{th} event occurs, where the time between two events follows an exponential distribution ($\lambda = 1/\theta$).
- Ex. : the time to hire a team of 6 people.

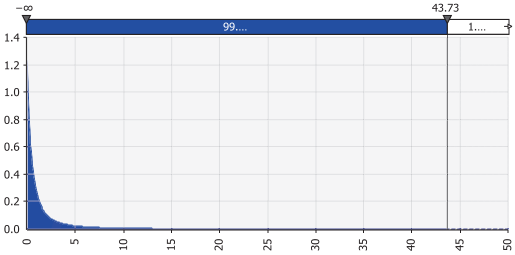
Exponential distribution



$$f(z;\lambda) \;=\; \lambda e^{-\lambda z}$$

- Represents the duration before the next event when this duration is independent of the time already spent waiting for this event (with expectation $1/\lambda$).
- Example : the waiting time before the next client.

Pareto distribution



$$f(z; z_0, \alpha) \;=\; \alpha z_0^{\alpha} / z^{\alpha+1}$$

- The severity of damages related to a (possibly disastrous) accident.
- Ex. : legal action, environmental risk, etc.
- Distribution characterized by «a heavy tail»

Outline

- Choosing a parametrized probability distribution function
- Parameter fitting

Why use data ?

- Experts aren't always available
- In certain contexts, data is abundant :
  - E-commerce, where millions of clients visit a website every day.
  - The information era grants us access to large amounts of historical data.
  - We can purchase certain data (stock price values, weather, surveys)
- Data provides a more objective argument.
- Data can complement an expert's opinion.

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 :

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

Maximization of likelihood

- We wish to model the distribution $P(Z)$ with the data $\{z_1, z_2, ..., z_M\}$
- Example 1 : Z is a Bernoulli trial, $p =$ prob. of success

$$\Rightarrow \quad p^* = \frac{\sum_i z_i}{M}$$

- Example 2 : $P(Z)$ is a normal distribution $\Rightarrow \mu^* = \frac{1}{M}\sum_i z_i$ & $\sigma^* = \sqrt{\frac{1}{M}\sum_i (z_i - \mu^*)^2}$

Chi-square Goodness of Fit Test (no bootstrapping)

1. We wish to validate the hypothesis that $P(Z)$ is really $f(z;\theta^*)$ with the data $\{z_1, z_2, ..., z_M\}$
2. Fit the parameter θ^* from our sample of M data points
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 according to $f(z;\theta^*)$:

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

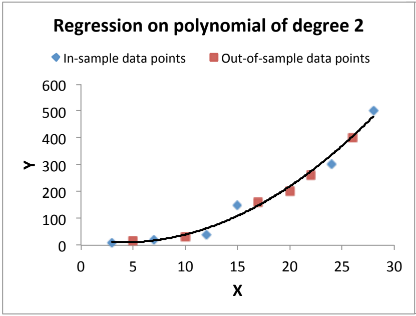
5. Compute the degrees of freedom $\nu = K - C - 1$ where C is the number of paramaters fitted (i.e., the dimension of θ^*)
6. Compute the «P-value» of the event $Y \geq X^2$, where Y follows a chi-square distribution with ν degrees of freedom
7. If the p-value is too small (e.g., below 0.05, which represents 5%), reject the hypothesis that $f(z;\theta^*)$ is the underlying model

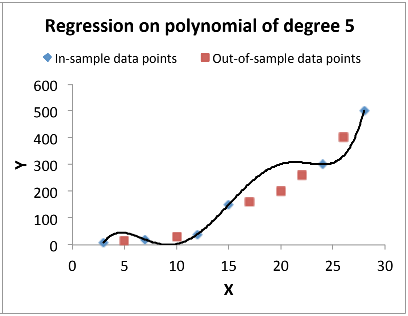
Beware of overfitting

- When the number of parameters grows, it takes more data to estimate them accurately (at least 10 data points per parameter)
- It is always better to test the performance of a set of parameters on a new dataset.
- The performance on this new dataset could improve when the model's complexity is reduced.

Example of overfitting

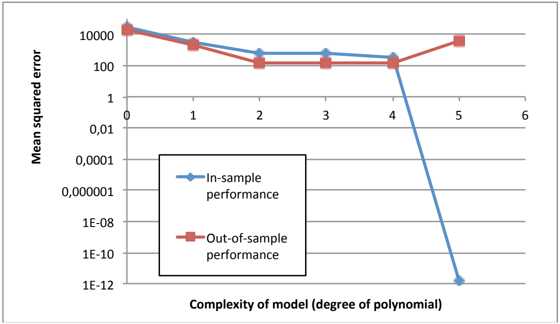
Which of the following two models overfits the data ?





Protecting oneself from overfitting

It is good practice to exclude part of the data during the fitting step so that it can be used to select the model's most appropriate level of complexity.



Outline

- Choosing a parametrized probability distribution function
- Parameter fitting
- **Bayesian approach**

Recall : Bayes' theorem

- Bayes' theorem tells us how to account for new information about random variables for which we already had information.
- In a case where we formulated $P(A)$ for $A = A_1, A_2, \cdots, A_n$ and receive the new information B
  - Characterize $P(B|A_1)$, $P(B|A_2), \cdots, P(B|A_n)$
  - Apply Bayes' theorem

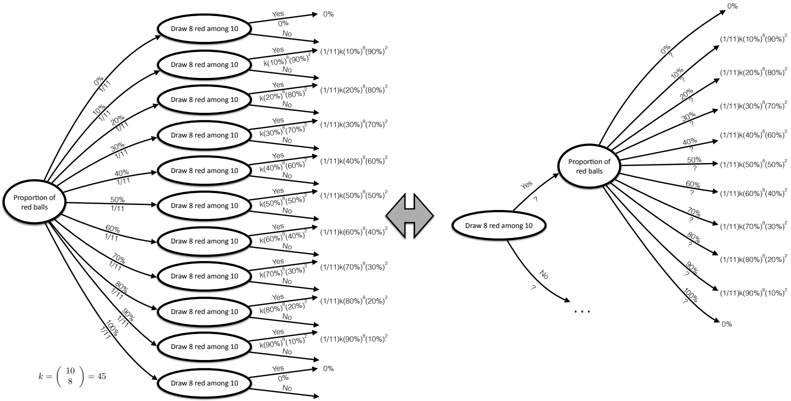
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
$$= \frac{P(B|A)P(A)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots P(B|A_n)P(A_n)}$$

Example of the urn containing 10 balls

- Consider an urn containing 10 balls. Out of 10 balls drawn randomly from the urn (with replacement), 8 were red. What is the probability of randomly drawing a red ball from the urn ?
- Subjective's answer :
  - Originally, I believed that all proportions were as likely (equal probability on each proportion).
  - After the experience, it is likely to be close to 0.8
  - I perceive that the probability that there is 8 red balls is approximately 0.33
  - However, if the urn was red this probability would be higher.

  Question : Confirm using Bayes' theorem that the probability of drawing 8 red balls (when drawing with replacement from the urn), given that 8 red balls were already observed in a set of 10 draws, is 0.33.

Example of the urn containing 10 balls - Solution



$$P(\text{Nb red} = 8|8 \text{ red among } 10) = \frac{(1/11)\cdot 45\cdot (0.8)^8\cdot (0.2)^2}{\sum_{i=1,2,\ldots,9}(1/11)\cdot 45\cdot (i/10)^8\cdot (1-i/10)^2} \approx 0.33325$$

(See Excel file for calculations.)

Bayesian inference

- We wish to model $P(Z)$ using $\{z_1, z_2, ..., z_M\}$
- Hypothesis followed by the Bayesian approach :
  1. Z follows a parametric form : $f(z;\theta)$
  2. Even before looking at our data, we have a subjective belief of the value of θ : i.e. $f(\theta)$
  3. We know that $f(\{z_1, z_2, ..., z_M\}|\theta) = \prod_{i=1}^{M} f(z_i;\theta)$
- Implications :
  1. Before seeing the data, $P(Z \in A) = \int P(Z \in A|\theta)f(\theta)d\theta = \int \int_A f(z;\theta)f(\theta)dzd\theta$
  2. After seeing the data, apply Bayes' theorem to determine $P(Z \in A|\mathcal{O})$, where $\mathcal{O} = \{z_1, z_2, ..., z_M\}$

$$f(\theta|\mathcal{O}) = \frac{f(\mathcal{O}|\theta)f(\theta)}{\int f(\mathcal{O}|\theta)f(\theta)d\theta} \qquad P(Z \in A|\mathcal{O}) = \int \int_A f(z;\theta)f(\theta|\mathcal{O})dzd\theta$$

where we exploit that Z is independent of $\mathcal{O}$ if θ is known

- Note : $f(\theta|\mathcal{O})$ is our belief of θ after studying $\mathcal{O}$

The conjugate prior of a distribution

- In general, it is difficult to compute $f(\theta|\mathcal{O})$ because of the integral

$$\int f(\mathcal{O}|\theta)f(\theta)d\theta = \int \left(\prod_{i=1}^{M} f(z_i;\theta)\right) f(\theta)d\theta$$

- If $f(\theta)$ is the conjugate prior for $f(z;\theta)$, then $f(\theta|\mathcal{O})$ takes the same parametric form as $f(\theta)$

Distribution	Conjugate prior	Parameters update
Bernoulli	Beta(α, β)	$\alpha' = \alpha + \sum_i z_i$, $\beta' = \beta + \sum_i (1 - z_i)$
Poisson	Gamma(k, θ)	$k' = k + \sum_i z_i$, $\theta' = \theta/(M\theta + 1)$
Exponential	Gamma(k, θ)	$k' = k + M$, $\theta' = \theta/(1 + \theta \sum_i z_i)$
...	...	...