Statistical Modeling

- Results of a statistical analysis have to be generalizable to be scientifically useful.
- A model proposes a general functional relation between the unknown parameter(s) and the observed data. It allows us to make predictions.
- The goal of a statistical analysis is to estimate the unknown parameter(s) in the proposed model.
- The likelihood function is central to the process of estimating the unknown parameters. Older and less sophisticated methods include the method of moments, and the method of minimum chi-square for count data. These estimators are not always efficient, and their sampling distributions are often mathematically intractable.

Example: Least squares vs. linear regression: One can fit a best straight line to explain the relation between two continuous variables X and Y using a least squares approach. The intercept ($\beta_0$) and slope ($\beta_1$) parameters can be estimated as the values that minimize the squared loss: $\sum (Y_i - \beta_0 - \beta_1 X_i)^2$. Using this approach, we cannot say anything about the precision of our estimates, or about future values of Y for a given X. In order to do so, we would need to make some assumption about the probability distribution of the error term, $\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$.

Example: Binned data: Likelihood provides a very flexible approach to combining data, provided one has a probability model for them. As a simple example, consider the challenge of estimating the mean $\mu$ from several independent observations for a $N(\mu, \sigma)$ process, but where each observation is recorded to a different degree of numerical ‘rounding’ or ‘binning.’ For example, imagine that because of the differences with which the data were recorded, the $n = 4$ observations are $y_1 \in [4, 6)$, $y_2 \in [3, 4)$, $y_3 \in [5, \infty)$, $y_4 \in (-\infty, 3.6)$. Even if we were told the true value of $\sigma$, the least squares method cannot handle this uni-parameter
estimation task. **Exercise:** Using the graphical methods described below determine the most plausible value of $\mu$.

### Likelihood

Let $X_1, X_2, ..., X_n$ have a joint density function $f(X_1, X_2, ..., X_n|\theta)$. Given $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ is observed, the function of $\theta$ defined by:

$$L(\theta) = L(\theta|x_1, x_2, ..., x_n) = f(x_1, x_2, ..., x_n|\theta)$$  \hspace{1cm} (1)

is the *likelihood function*.

- The likelihood function is not a probability density function.
- It is an important component of both frequentist and Bayesian analyses.
- It measures the support provided by the data for each possible value of the parameter.
  
  If we compare the likelihood function at two parameter points and find that $L(\theta_1|x) > L(\theta_2|x)$ then the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. This can be interpreted as $\theta_1$ is a more plausible value for $\theta$ than $\theta_2$.

**Example:** American or Canadian M&M’s? (Discrete parameter): M&M’s sold in the United States have 50% red candies compared to 30% in those sold in Canada. In an experimental study, a sample of 5 candies were drawn from an unlabelled bag and 2 red candies were observed. Is it more plausible that this bag was from the United States or from Canada? The likelihood function is: $L(p|x) \propto p^2(1-p)^3$, $p=0.3$ or $0.5$. $L(0.3|x) = 0.03087 < 0.03125 = L(0.5|x)$, suggesting that it is more plausible that the bag used in the experiment was from the United States.
Likelihood Principle

If x and y are two sample points such that \( L(\theta|x) \propto L(\theta|y) \) \( \forall \theta \) then the conclusions drawn from x and y should be identical.

Thus the likelihood principle implies that likelihood function can be used to compare the plausibility of various parameter values. For example, if \( L(\theta_2|x) = 2L(\theta_1|x) \) and \( L(\theta|x) \propto L(\theta|y) \) \( \forall \theta \), then \( L(\theta_2|y) = 2L(\theta_1|y) \). Therefore, whether we observed x or y we would come to the conclusion that \( \theta_2 \) is twice as plausible as \( \theta_1 \).

Example: Two samples, same distribution: Consider the distribution Multinomial\((n = 6, \theta, \theta, 1 - 2\theta)\). The following two samples drawn from this distribution have the same likelihood:

\[
X = (1, 3, 2) \Rightarrow \frac{6!}{1!3!2!} \theta^1 \theta^3 (1 - 2\theta)^2 \\
X = (2, 2, 2) \Rightarrow \frac{6!}{2!2!2!} \theta^2 \theta^2 (1 - 2\theta)^2
\]

This means both samples would lead us to the same conclusion regarding the relative plausibility of different values of \( \theta \).

Example: Same sample, two distributions: Two treatments A and B were compared in each of 6 patients. Treatment A was better in the first 5 patients, and treatment B in the sixth patient. Let us imagine this experiment was conducted by two investigators, each of whom, unbeknownst to the other, had a different study design in mind. The first investigator had originally planned to study 6 patients (a binomial experiment), while the second had planned to stop the study when the first patient who preferred treatment B was observed (a negative binomial experiment). Thus, the likelihood function according to the first investigator would be \( L(p|x) = \binom{5}{0} p^5(1 - p) \), where \( p = P(\text{Treatment A is preferred}) \). The likelihood function according to the second investigator would be \( L(p|y) = \binom{5}{5} p^5(1 - p) \).

(Note: The negative binomial density function for observing \( y \) failures before the \( r \)th success is \( P(Y = y) = \binom{y+r-1}{r-1} p^r(1-p)^k, k = 0, 1, 2, 3, \ldots \). The likelihood functions of both investigators are proportional implying that they should reach the same conclusion.
If were to use a frequentist approach for inference and calculate a p-value for the null hypothesis $H_0: p = 0.5$ against the alternative $H_1: p > 0.5$, then the first investigator would obtain a p-value of 0.11, while the second investigator would obtain a p-value of 0.03, potentially leading them to different conclusions. (See Goodman S, Towards evidence-based medical statistics. 1: The p-value fallacy. Annals of Internal Medicine, 1999)

**Plotting the likelihood function**

Example from Clayton and Hills (Continuous parameter): In section 3.1, Clayton and Hills describe a hypothetical study in which 10 patients are followed for a fixed time period at the end of which 4 of them die. The likelihood function is given by: $L(p|x) \propto p^4(1 - p)^6$. The likelihood of $p=0.5$ is $9.77 \times 10^{-4}$, whereas the likelihood of $p=0.1$ is $5.31 \times 10^{-5}$.

Likelihood function plot:

- Easy to see from the graph the most likely value of $p$ is 0.4 ($L(0.4|x) = 9.77 \times 10^{-4}$).
- Absolute values of likelihood are tiny not easy to interpret
- Relative values of likelihood for different values of $p$ are more interesting

Plotting the Likelihood ratio:
• Measures how likely different values of $p$ are relative to $p=0.4$.

• Can be used to define a group of values of $p$ that are supported by the data, i.e. a group of values whose likelihood ratio is above a critical value.

• Using a critical value of 0.258 we find that the range of supported values is 0.178 to 0.655. This critical value corresponds to a 90% confidence level.

• If we had in fact observed 20 deaths in a sample of 50 individuals, the most likely value of $p$ would still be 0.4 but the supported range would be narrower: 0.291, 0.516.

Plotting the log-Likelihood ratio: The (log-)likelihood is invariant to alternative monotonic transformations of the parameter, so one often chooses a parameter scale on which the function is more symmetric.
Exercise: Tumble Mortality data: Write down the log likelihood function for the data on annealed glasses. Assume the shape parameter, $\mu$, is known to be equal to 1.6. Plot the log likelihood function vs. possible values of the rate to determine the most plausible value of the rate for the observed data.