Statistical Inference

• Data generated from unknown probability distribution and statement on the unknown distribution are warranted. Determine parameters (e.g. \( \beta \) for exponential distribution, \( \mu \) and \( \sigma \) for normal distribution)

• Prediction of new experiments
Estimation of parameters

- Notation: \( f(x|\theta) \) is the probability density of sampling \( x \) given (conditioned on) parameters \( \theta \).

- For a set of \( n \) independent and identically distributed samples the probability density is:

\[
f(x_1,\ldots,x_n \mid \theta) = \prod_{i=1,\ldots,n} f(x_i \mid \theta) \equiv f(x \mid \theta)
\]

- However, what we want to determine now are the parameters... For example assuming the distribution is normal, we seek the mean \( \mu \) and the variance \( \sigma^2 \)

\[
f(x \mid \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma}} \exp\left( -\frac{(x - \mu)^2}{2\sigma^2} \right)
\]
Bayesian arguments

• What we want is the function $f(\theta \mid x)$ given a set of observations $x$, what is the probability that the set of parameters is $\theta$?

• Bayesian statistics: Think of the parameters like other random variables with probability $\xi(\theta)$.

The joint probability $f(x, \theta) \equiv f(x \mid \theta) \xi(\theta)$ is also $f(x, \theta) \equiv f(\theta \mid x) g(x)$.
The likelihood function

- We can formally write \( L(\theta | x) = \frac{f(x | \theta) \xi(\theta)}{g(x)} \)

which is the probability of having a particular set of parameter for the p.d.f provided a set of observation (what we wanted). Note that our prime interest here is in the parameter set \( \theta \) and the samples of \( x \) is given. Since \( g(x) \) is independent of \( \theta \) we can write the likelihood function

\[ L(\theta | x) \propto f(x | \theta) \xi(\theta) \]
Example: Likelihood function I

• Consider the exponential distribution

\[
f(x | \beta) = \begin{cases} 
  \beta \exp[-\beta x] & \text{for } x > 0 \\
  0 & \text{otherwise}
\end{cases}
\]

\[
f(x | \beta) = \begin{cases} 
  \beta^n \exp\left[-\beta \sum_{i=1}^{n} x_i \right] & \text{for } x > 0 \\
  0 & \text{otherwise}
\end{cases}
\]

• And assume the p.d.f. of the parameter \( \beta \) is a Gaussian with a mean and variance of 1.

\[
\xi(\beta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\beta^2}{2}\right)
\]
Example: Likelihood function II

\[ \xi(\beta \mid x) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{\beta^2}{2}\right) \beta^n \exp\left[-\beta \sum_{i=1,\ldots,n} x_i \right] \]
Maximum Likelihood

We look for a maximum of the function \( L(\theta) = \log(f_n(x|\theta)) \)
as a function of the parameters \( \theta \)

As a concrete example we consider the normal distribution

\[
L(\theta) = \log \left[ f_n(x|\mu,\theta) \right] \\
= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1,...,n} (x_i - \mu)^2
\]

To find the most likely set of parameters we determine the maximum of \( L(\theta) \)
Maximum of $L(\theta)$ for normal distribution

$$\frac{dL}{d\mu} = 0 = -\frac{1}{2\sigma^2} \sum_{i=1,...,n} 2(x_i - \mu) = -\frac{1}{\sigma^2} \left( \sum_{i=1,...,n} x_i - n\mu \right)$$

$\Rightarrow \mu = \frac{1}{n} \sum_{i=1,...,n} x_i$

$$\frac{dL}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1,...,n} (x_i - \mu)^2 = 0$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1,...,n} (x_i - \mu)^2$$
Determine a most likely parameter for the uniform distribution

\[
\begin{align*}
    f(x \mid \theta) &= \begin{cases} 
        \frac{1}{\theta} & \text{for } 0 \leq x \leq \theta \\
        0 & \text{otherwise}
    \end{cases} \\
    f(x \mid \theta) &= \begin{cases} 
        \frac{1}{\theta^n} & \text{for } 0 \leq x_i \leq \theta \ (i = 1, \ldots, n) \\
        0 & \text{otherwise}
    \end{cases}
\end{align*}
\]

It is clear that \( \theta \) must be larger than all the \( x_i \) and at the same time maximizes the monotonically decreasing function \( 1/\theta^n \), hence

\[
\theta = \max \left[ x_1, \ldots, x_n \right]
\]
Potential problems in maximum likelihood procedure

• Value of $\theta$ is underestimated (note that $\theta$ should be larger than all $x$, not only the ones we sample so far)

• No guarantee that a solution exists for the distribution below $\theta$ must be large than any $x$ but at the same time equal to the maximal $x$. This is not possible and hence, no solution

$$f(x | \theta) = \begin{cases} 
\frac{1}{\theta} & \text{for } 0 < x < \theta \\
0 & \text{otherwise}
\end{cases}$$

• The solution is not necessarily unique

$$f_n(x | \theta) = \begin{cases} 
1 & \text{for } \theta \leq x_i \leq \theta + 1 \ (i=1,...,n) \\
0 & \text{otherwise}
\end{cases}$$

$$\Rightarrow f_n(x | \theta) = \begin{cases} 
1 & \text{for } \max(x_1,...,x_n) - 1 \leq \theta \leq \min(x_1,...,x_n) \\
0 & \text{otherwise}
\end{cases}$$

$$\Rightarrow \max(x_1,...,x_n) - 1 \leq \theta \leq \min(x_1,...,x_n)$$
The $\chi^2$ distribution with $n$ degrees of freedom

$$f(x) = \frac{1}{2^{n/2} \Gamma(n/2)} x^{(n/2)-1} \exp\left(-\frac{x}{2}\right) \quad x > 0$$

$$E(x) = n \quad \text{var}(x) = 2n$$

There is a useful relation between the $\chi^2$ and the normal distributions
Theorem connecting $\chi^2$ and normal distributions

If the random variables $X_1, \ldots, X_n$ are i.i.d. and if each of these variables has standard normal distribution, then the sum of the squares

$$Y^2 = X_1^2 + \ldots + X_n^k$$

Has a $\chi^2$ distribution with $n$ degrees of freedom

The distribution functions

$$F(y) = \Pr(Y \leq y) = \Pr(X^2 \leq y) = \Pr(-y^{1/2} \leq X \leq y^{1/2})$$

$$= \Phi\left(y^{1/2}\right) - \Phi\left(-y^{1/2}\right)$$

The p.d.f is obtained by differentiating both side $f'(y) = F'(y)$

$$\phi(y) = \Phi'(y). \text{ Note } \phi\left(y^{1/2}\right) = (2\pi)^{-1/2} \exp\left(-y/2\right). \text{ We have}$$

$$f(y) = \phi\left(y^{1/2}\right)\left(1/2y^{-1/2}\right) + \phi\left(-y^{1/2}\right)\left(1/2y^{-1/2}\right)$$

$$f(y) = (2\pi)^{-1/2} y^{-1/2} \exp\left(-y/2\right)$$

which is the $\chi^2$ distribution with one degree of freedom
Normal distribution: Parameters

Let $X_1, \ldots, X_n$ be a random sample from normal distribution having mean $\mu$ and variance $\sigma^2$. Then the sample mean (hat denotes M.L.E)

$$\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and the sample variance

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left( X_i - \bar{X}_n \right)^2$$

are independent random variables.

$\hat{\mu}$ has a normal distribution with a mean $\mu$ and variance $\sigma^2/n$.

$n\hat{\sigma}^2 / \sigma^2$ has a chi-square distribution of $n-1$ degrees of freedom

Why $n-1$? (next slide)
Parameters of the normal distribution: Note 1

- Let $x_1, \ldots, x_n$ be a vector of random numbers of length $n$ sampled from the normal distribution.
- Let $y_1, \ldots, y_n$ be another vector of $n$ random numbers, related to the previous vector by a linear transformation $A$ ($AA^t = I$).

\[ y = Ax \]

- Consider now the calculation of the variance (next slide).
Variance

• The formula we should use for the variance

\[
\text{var}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2
\]

• However, we do not know the exact mean, and therefore we use

\[
\text{var}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2
\]

• What are the consequences of using this approximation?
Variance is not changing upon linear transformations

Consider the expression

$$\sum_{i=1,\ldots,n} (Y_i - \bar{Y}_n)^2 = \sum_{i=1,\ldots,n} (AX_i - A\bar{X}_n)^t (AX_i - A\bar{X}_n)$$

$$\sum_{i=1,\ldots,n} (X_i^t - \bar{X}_n)^t A^t A (X_i^t - \bar{X}_n) = \sum_{i=1,\ldots,n} (X_i^t - \bar{X}_n)^2$$

The analysis is based on the unitarity of $A$. Hence, linear transformation does change the variance of the distribution. This makes it possible to exploit the difference between $\bar{X}_n$ and $\mu$. 
The $n-1$ (versus $n$) factor

- Since $A$ is arbitrary (as long as it is unitary). We can choose one of the transformation vectors $a$ to be $(1,\ldots,1)/n^{1/2}$
- The scalar product
  \[ X^t a - \overline{X}_n a = 0 \]
  - Is identically zero (remember how we compute the mean?)
  - Hence since we computed the average from the same sample we computed the variance, the variance lost one degree of freedom.
The $n-1$ factor II

- Note that the $n-1$ makes sense. Consider only a single sample point, which is of course very poor and leaves a high degree of uncertainty regarding the value of the parameters. If we use $n$ then the estimated variance becomes zero, while if we use $n-1$ we obtain infinite, which is more appropriate to the problem at hand, for which we have no information to determine the variance.
The $t$ distribution
(in preparation for confidence intervals)

- Consider two random variables $Y$ and $Z$, such that $Y$ has chi-2 distribution with $n$ degrees of freedom and $Z$ has a standard normal distribution the variable $X$ is defined by

$$X = Z \left/ \left( \frac{Y}{n^{1/2}} \right) \right.$$ 

Then the distribution of $X$ is the $t$ distribution with $n$ degrees of freedom.
The $t$ distribution

- The function is tabulated and can be written in terms of $\Gamma$ function

$$ t_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{(n\pi)^{1/2} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2} $$

$$ \Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x) \, dx $$

- The $t$ distribution is approaching the normal distribution as $n \to \infty$. It has the same mean but longer tails.
Confidence Interval

- Confidence interval provide an alternative to the use of estimator instead of the actual value of an unknown parameter. We can find an interval \((A, B)\) that we think has high probability of containing the desired parameter. The length of the interval gives us an idea how well we can estimate the parameter value.
Confidence interval for the mean of the normal distribution

• Let $X_1, \ldots, X_n$ for a random sample from a normal distribution with unknown mean and unknown variance. Let $t_{n-1}(x)$ denote the p.d.f of the $t$ distribution with $n-1$ degrees of freedom, and let $c$ be a constant such that

$$\int_{-c}^{c} t_{n-1}(x) \, dx = \gamma$$

• For every value of $n$, the value of $c$ can be found from the table of the $t$ distribution to fit the confidence (probability) $\gamma$