SESA6085 – Advanced Aerospace Engineering Management

Lecture 3

2023-2024

Dr David Toal

Parameter Estimation

Previously…

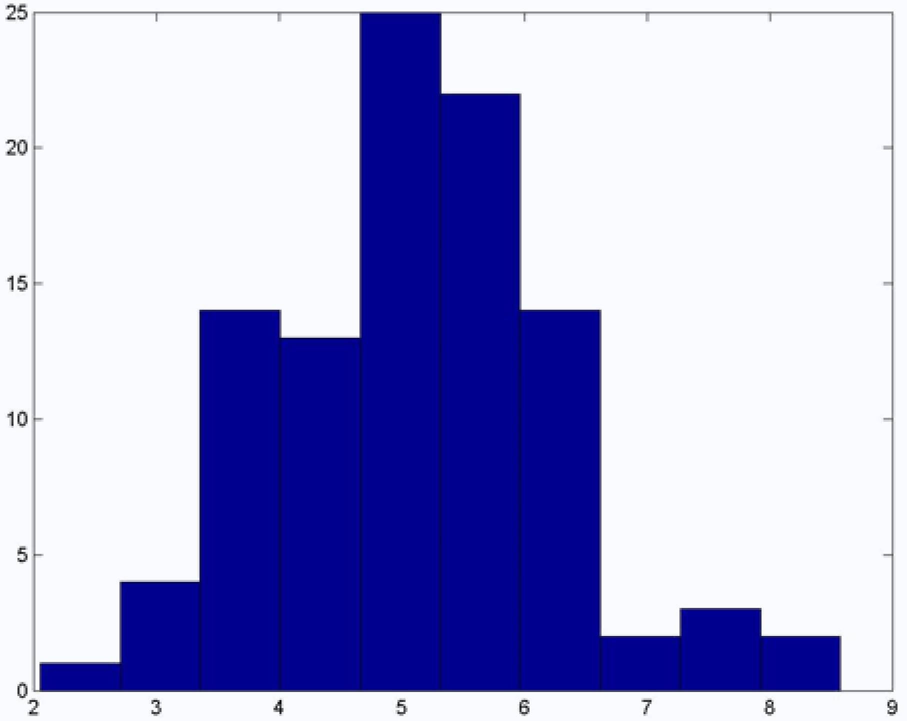
- Introduced a number of different continuous distribution functions e.g. normal, Weibull, exponential etc.
- The formulas for their PDF, CDF, reliability and hazard functions were also derived
- The first tutorial will illustrate how we can use these functions to calculate various probabilities and reliabilities
- But in each case the defining parameters were provided

Parameter Estimation

- How do we define such distributions in the first place?
- How do we calculate the parameters defining each model?
- This process is generally referred to as parameter estimation
- There are a number of different parameter estimation techniques but all require some data from which the parameters can be estimated

Parameter Estimation

- Given that we have some data, this could be in the form of measurements, failure times etc., how do we fit the PDF to it?



Parameter Estimation

- For all parameter estimation techniques the following should be true:
  - Unbiased – the estimator should not consistently under or overestimate the true value of the parameter
  - Consistent – the estimator should converge to the true value as the sample size increases
  - Efficient – the estimator should be consistent with a standard deviation in that estimate smaller than any other estimator for the same population
  - Sufficient – the estimator should use all of the information about the parameter that the data sample possesses

Possible Techniques

- There are three widely used techniques for estimating parameters:
  - The method of moments < Not examined
  - The maximum likelihood method
  - The least squares method
- All of these techniques depend on the quality of the data e.g. the presence of outliers which should be removed beforehand
- The following lecture will concentrate on maximum likelihood estimation but in some instances the methods are identical

  Often all these methods will produce the same estimation (because they are sufficient)

  - For those interested, information on the method of moments and least squares fitting can be found at the end of these slides

Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation

- Suppose that each observation is drawn from a common PDF of form:

V we use time, but could be pressure, force ect

$$f(t;\theta_1,\theta_2,\dots,\theta_m)$$

- Where the form of f is known i.e. it's Gaussian, exponential etc. but its parameters, θ_1, θ_2,… θ_m, are unknown
- For example for a Gaussian PDF then θ_1 & θ_2 might represent the mean and standard deviation
- Let's shorten the notation to:

$$f(t;\theta)$$

Maximum Likelihood Estimation

- If the density function is of the form:

$$f(t;\theta)$$

- Then the likelihood of the observations is defined by

$$L(\theta) = \prod_{i=1}^{n} f(t_i;\theta)$$

By multiplying the probability of each individual event (in the context of the tested prob function) we can see the overall likelihood of those observations being related to the function in question

Effectively the joint probability ^

- Generally working with the log-likelihood is more convenient

$$l(\theta) = \log L(\theta)$$

We'll generally get REALLY small numbers for L, hence use of log's for numerical stability and convenience

- Why is this the case?

Maximum Likelihood Estimation

- Let's define the maximum likelihood estimates of ϑ as

$$\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$$

- These are the values which maximise the log-likelihood which can be achieved using an optimisation algorithm
- Alternatively the MLEs can be found by solving the likelihood equations

$$\frac{\partial l}{\partial \theta_j} = 0 \quad (j = 1,2, \ldots, m)$$

MLE of a Normal Distribution

- We have a set of *n* data observations, how do we use MLE to fit a normal distribution to these observations?
- Recall our PDF for a normal distribution:

$$f(t) = \frac{1}{\sigma(2\pi)^{1/2}} \exp\left[-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right]$$

- So we need to find?

$$\hat{\mu} \quad \& \quad \hat{\sigma}$$

MLE of a Normal Distribution

- Taking natural logs of our pdf we get

$$\ln(f(t_i)) = -\ln(\sigma) - \frac{1}{2}\ln(2\pi) - \frac{1}{2}\left(\frac{t_i - \mu}{\sigma}\right)^2$$

- Transforming the likelihood function into the log-likelihood

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta)$$

$$l(\mu, \sigma) = \ln(L(\mu, \sigma)) = \ln(\prod_{i=1}^{n} f(t_i; \mu, \sigma))$$

$$l(\mu, \sigma) = \sum_{i=1}^{n} \ln(f(t_i; \mu, \sigma))$$

MLE of a Normal Distribution

- So for our *n* data points the log-likelihood becomes

$$l(\mu,\sigma) = -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n}\frac{1}{2}\left(\frac{t_i-\mu}{\sigma}\right)^2$$

- Our maximum likelihood estimates are therefore when the derivatives of this function with respect to μ and σ are zero

Turning points ^

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu - t_i)$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(t_i - \mu)^2$$

MLE of a Normal Distribution

- Our MLEs become

(as you'd expect, it's just the mean)

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

(shockingly, we find the standard deviation, lol)

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - \hat{\mu})^{2}}$$

- Do these look familiar?

MLE of an Exponential Distribution

- Derive the MLE for an exponential distribution
- What should we do?
  1. Define the formulation of the PDF
  2. Define the parameter to estimate
  3. Define the natural log of the PDF
  4. Define the log likelihood function
  5. Define it’s derivative(s)
  6. Equate the derivative(s) to zero and solve for the parameter(s)

MLE of an Exponential Distribution

- What is the PDF?

$$f(t) = \lambda \exp(-\lambda t)$$

- What is the parameter we wish to estimate?

$$\hat{\lambda}$$

- What is the natural log of the PDF?

$$\ln(f) = \ln(\lambda) - \lambda t$$

MLE of an Exponential Distribution

- What is the log-likelihood function for a sample of *n* data points?

$$l(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} t_i$$

- What is the derivative of the log-likelihood function with respect to λ?

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i$$

MLE of an Exponential Distribution

- Finally, what is the maximum likelihood estimator of λ?

$$\frac{\partial l}{\partial \lambda} = 0 = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i \qquad \therefore \frac{n}{\lambda} = \sum_{i=1}^{n} t_i$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}$$

(equivalent to 1/mean time to failure)

Maximum Likelihood Estimation

- The exact same process can be applied to most functions
- However, some distributions do not have a closed-form solution to the MLE. What do we do in this case?
- Typically some form of optimisation algorithm is employed to maximise the log-likelihood function

(Turn it into a optimization problem)

MLE in Matlab & Python

- Matlab & Python are able to perform MLE for a wide range of continuous distribution functions
- Where possible they employ the analytical solution to the MLE
- Otherwise an optimisation algorithm is used to find the parameters
- Type "`help mle`" in Matlab for more information
- See the documentation for the `.fit` operator in `scipy.stats`

If Matlab & Python Can Do It….

- “If they can do it for me what's the point of this lecture?”
- There are a number of reasons for this:
  1. You might not always have access to Matlab or Python
  2. A deeper understanding of the mathematics in operation is always helpful

If Matlab & Python Can Do It….

- Not a good enough answer?

  1. MLE is applied in a wide range of subjects beyond reliability some of which require quite complex and costly optimisations therefore knowing how to do it is an advantage
  2. There are special cases, even for the basic PDFs mentioned previously, in which the MLE formulation may need to be considerably modified. For these cases you'll need to derive your own MLE function and knowing the basics of MLE helps with this

You may have incomplete samples, such as only having for a certain time period. Then having to have a deep enough understanding of stats to work around that limitation.

Other MLE

- If you are interested in the derivations of MLEs for any of the other distributions presented previously you may find the following useful:
  - Chapter 4 of “Reliability Engineering”
  - Section 3.5.2 of “Practical Reliability Engineering”
  - Chapter 3 of “Statistical Analysis of Reliability Data”

Correct Distribution?

(Not examinable)

What Distribution?

- For the purposes of this module you will be told the distribution to use
- However, in the "real world" this will not be the case
- We could fit multiple distributions to a dataset
  - The MLE maths, doesn't care if the distribution is correct
- To decide which distribution is more appropriate there are statistical tests that can be applied e.g.
  - χ^2 test (pronounced chi-squared) < A test done to test if a distribution fits well for the provided data

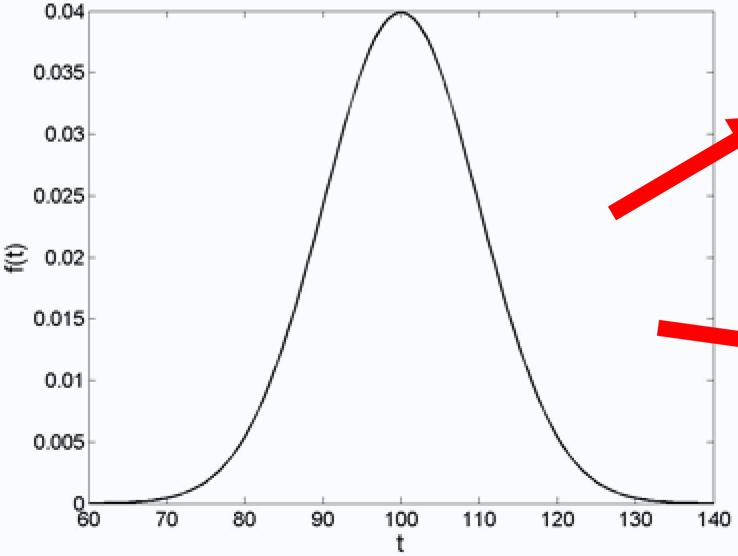
Remember…

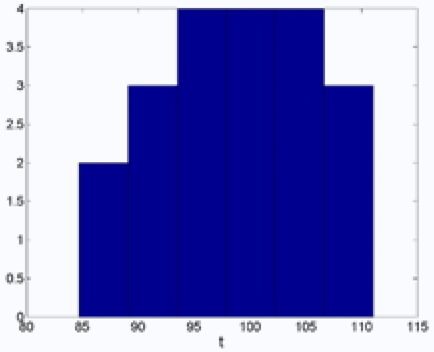
unknown distribution
data scientist
Gaussian

Parameter Confidence

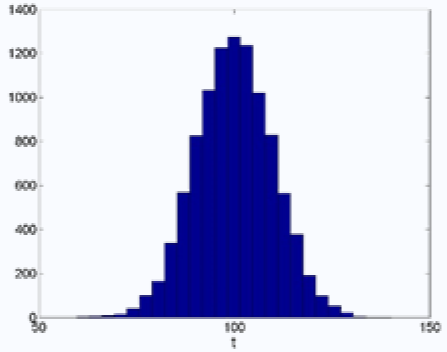
Is MLE Consistent?

- Consider a normal distribution with μ= 100 & σ = 10
- Let’s use MLE to determine the parameters as the size of a sample from this distribution increases





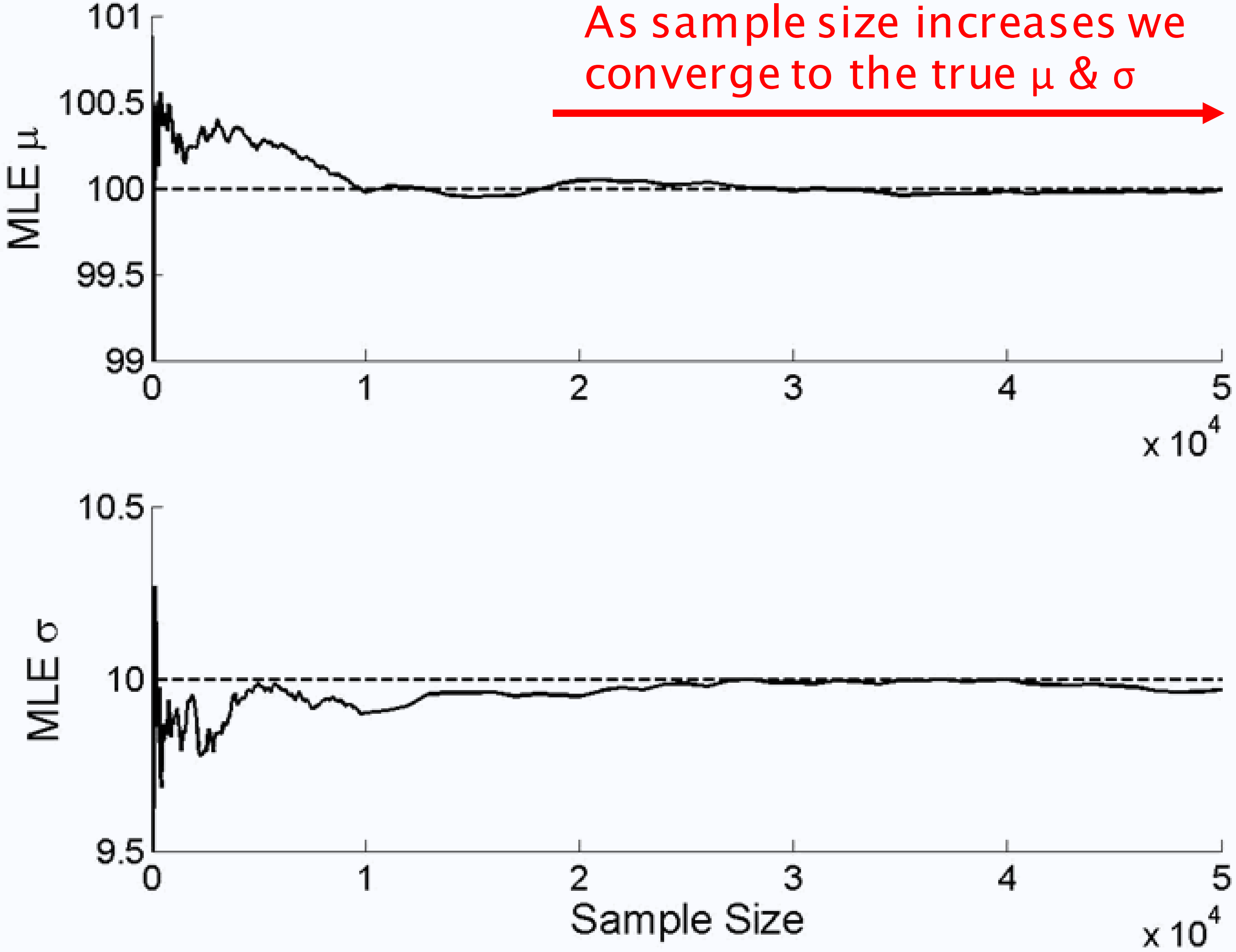
20 sample points



10,000 sample points

Is MLE consistent?

V hence is consistant



As sample size increases we converge to the true μ & σ

Data is Key

- Hopefully this illustrates that to create truly accurate PDFs the amount of data you use is very important
- This holds true for similar processes e.g. Monte Carlo analysis, design optimisation, surrogate model construction etc.
- Generally the more data the better
- Of course you still need to select the correct PDF in the first place
- If we have a small sample size we will not have the correct MLE but can we estimate how incorrect our value is?
  - Yes we can calculate variances and confidence values
  - What might these be a function of? (number of points)

University of Southampton

The Information & Covariance Matrices

- The ij^th element of the Fisher information matrix is:

$$I_{ij} = E\left[-\frac{\partial^2 l(t;\theta)}{\partial\theta_i \partial\theta_j}\right]$$

The expected negative of the second derivative of the log-likelihood w.r.t the parameters

- The inverse of this matrix equals the covariance matrix

$$I^{-1} = \begin{bmatrix} \mathrm{Var}(\theta_1) & \mathrm{Cov}(\theta_1,\theta_2) & \cdots & \mathrm{Cov}(\theta_1,\theta_k) \\ \mathrm{Cov}(\theta_2,\theta_1) & \mathrm{Var}(\theta_2) & \cdots & \mathrm{Cov}(\theta_2,\theta_k) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(\theta_k,\theta_1) & \mathrm{Cov}(\theta_k,\theta_2) & \cdots & \mathrm{Var}(\theta_k) \end{bmatrix}$$

- Where
  - $\mathrm{Var}(\theta_i)$ – the variance in the parameter θ_i
  - $\mathrm{Cov}(\theta_i,\theta_j)$ - the covariance of θ_i and θ_j

Normal Distribution Example

- Whilst deriving the MLE of μ & σ we found that:

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(\mu - t_i)$$

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(t_i - \mu)^2$$

- The second derivatives are therefore

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} \qquad \frac{\partial^2 l}{\partial \sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(t_i - \mu)^2 \qquad \frac{\partial^2 l}{\partial \mu \partial \sigma} = \frac{2}{\sigma^3}\sum_{i=1}^{n}(\mu - t_i)$$

Normal Distribution Example

- With the second derivatives we now calculate the negative of their expected values
- We do so by inputting the MLE values of the μ & σ and simplifying where possible
- For the second derivative w.r.t μ this is:

$$E\left[-\frac{\partial^2 l}{\partial \mu^2}\right] = \frac{n}{\hat{\sigma}^2}$$

- Using our equation for $\hat{\sigma}$ we find that:

$$E\left[-\frac{\partial^2 l}{\partial \sigma^2}\right] = -\frac{n}{\hat{\sigma}^2} + \frac{3n}{\hat{\sigma}^2}\frac{\sum_{i=1}^{n}(t_i - \mu)^2}{\sum_{i=1}^{n}(t_i - \mu)^2} = \frac{2n}{\hat{\sigma}^2}$$

Normal Distribution Example

- Similarly by recalling our MLE for μ:

$$E\left[-\frac{\partial^2 l}{\partial\mu\partial\sigma}\right] = -\frac{2}{\sigma^3}\sum_{i=1}^{n}(\hat{\mu} - t_i) = -\frac{2}{\hat{\sigma}^3}\left[n\hat{\mu} - \sum_{i=1}^{n} t_i\right] = 0$$

- Our information & covariance matrices are therefore:

$$I = \begin{bmatrix} {}^{n}/_{\sigma^2} & 0 \\ 0 & {}^{2n}/_{\sigma^2} \end{bmatrix} \qquad I^{-1} = \begin{bmatrix} {}^{\sigma^2}/_{n} & 0 \\ 0 & {}^{\sigma^2}/_{2n} \end{bmatrix}$$

- Hence:

$$\mathrm{Var}(\hat{\mu}) = {}^{\hat{\sigma}^2}/_{n} \qquad \mathrm{Var}(\hat{\sigma}) = {}^{\hat{\sigma}^2}/_{2n}$$

Functions of n as expected

An Example

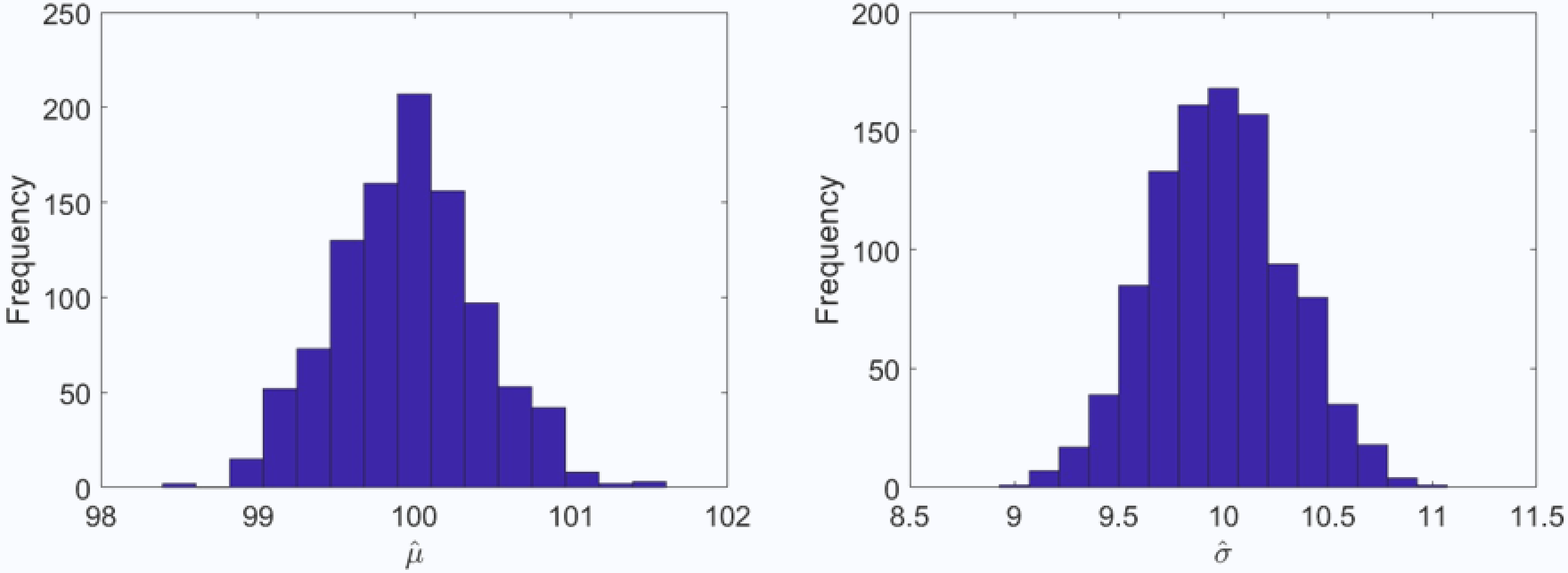
- Consider fitting a normal distribution to 500 data points drawn from our true distribution of $\mu = 100$ & $\sigma = 10$
- Via MLE we find $\hat{\mu} = 100.49$ & $\hat{\sigma} = 9.78$
- Using our equations for the variance we find

$$\mathrm{Var}(\hat{\mu}) = 0.191 \ \& \ \mathrm{Var}(\hat{\sigma}) = 0.096$$

- But what does this mean?
- Let’s consider a graphical interpretation

A Graphical Interpretation

- Imagine we repeatably draw sets of 500 numbers from our normal distribution and calculate $\hat{\mu}$ and $\hat{\sigma}$ for each
- We do this 1000 times and plot a histogram of resulting $\hat{\mu}$…



- What do we observe?

(Guassian distrobutions, hence variance tells us the shape of the dist, from here confidence can be computed)

A Graphical Interpretation

- We could calculate the variance of the results in this histogram

	$\mathrm{Var}(\hat{\mu})$	$\mathrm{Var}(\hat{\sigma})$
Equations	0.191	0.0956
Graphical Example (1000)	0.211	0.1056

- The equations for the variance in the MLEs, therefore, defines the spread in the values of MLEs if we continually resampled the dataset used to calculate the MLE
- These variances can be used to calculate confidence intervals for our MLEs

MLE Confidence Intervals

$$P(\theta_l \leq \hat{\theta} \leq \theta_u) = \gamma$$

- θ_l and θ_u are called the confidence intervals and γ is the confidence level
- For example if we set γ equal to 0.95 then we can expect about 95% of the time the value of $\hat{\theta}$ is within the bounds θ_l and θ_u ^ Arbitrary of course, but reasonable
- Our confidence bounds are therefore…

$$\hat{\theta} \pm Z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})}$$

- Where $\alpha = 1 - \gamma$ and $Z_{\alpha/2}$ is the standard normal statistic
  - negative of the inverse CDF of a normal distribution with mean 0 and a standard deviation of 1 for the probability $\alpha/2$

Example Continued

- From our 500 data points we know…

$$\hat{\mu} = 100.49 \ \& \ \hat{\sigma} = 9.78$$

$$\mathrm{Var}(\hat{\mu}) = 0.191 \ \& \ \mathrm{Var}(\hat{\sigma}) = 0.096$$

- For 95% bounds, $\alpha/2$=0.025 and therefore $Z_{\alpha/2} = 1.96$
  - $Z_{\alpha/2}$ can be calculated or read from a look-up table
- Using $\hat{\theta} \pm Z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})}$ we obtain the following…

$$\hat{\mu} \pm 0.857 \qquad \hat{\sigma} \pm 0.607$$

$$99.63 \leq \hat{\mu} \leq 101.35 \qquad 9.17 \leq \hat{\sigma} \leq 10.39$$

- If we wanted 99% bounds, $Z_{\alpha/2} \approx 2.58$

University of Southampton

Method of Moments

Method of Moments

- The main aim here is to equate certain sample statistical characteristics (e.g. mean and variance etc.) to the expected values of the distribution and then solve the resulting equation
- Note that the expected value E[t] is given by the following general equation:

$$E[t] = \int_{q} t f(t) dt$$

- Where q here defines the range over which the pdf is applicable e.g. +-∞ for a normal distribution

Method of Moments - Example

- Let's take the exponential distribution as an example

$$f(t) = \lambda \exp(-\lambda t)$$

$$E[t] = \int_0^\infty t f(t) d\,t = \int_0^\infty t\lambda \exp(-\lambda t) d\,t = {}^{1}/_{\lambda}$$

- This corresponds to our first statistical moment. If we equate this to the same moment from our data we can solve for λ

Method of Moments - Example

$$\frac{\sum_{i=1}^{n} t_i}{n} = E[t] = {}^{1}/_{\lambda}$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}$$

- Where we distinguish our parameter estimate by a ^ symbol
- The same process can be applied to other PDFs using additional statistical moments if required

Least Squares

Least Squares

- This method attempts to minimise the sum of the squared errors between the observed data and the proposed distribution
- Can be used for a variety of models (linear, nonlinear etc.)
- Has uses beyond fitting pdfs e.g. fitting splines or polynomials to point cloud information
  - Geometry matching
  - Surrogate modelling etc.

Least Squares - Example

- Lets assume the data can be represented by a linear model

$$f(t_i) = a + bt_i + \varepsilon_i$$

- Where ε_i is some random noise
- If we propose to fit a model of the form

$$\hat{f}(t) = \hat{a} + \hat{b}t$$

- Then the error between the proposed and actual model is

$$e(t_i) = f(t_i) - \hat{f}(t_i)$$

Least Squares - Example

- The sum of the squares of the error is then

$$SS_E = \sum_{i=1}^{n} e^2(t_i) = \sum_{i=1}^{n} \left[f(t_i) - \hat{f}(t_i)\right]^2$$

- Which can be minimised to find the best value for *a* & *b*
- A search algorithm can be used, or alternatively, the problem may be solved analytically by taking partial derivatives and equating them to zero