Statistical Thinking using Randomisation and Simulation (ETC2420)

Contents (click the links to go directly to the page):

Key:

- Red underlined means new lecture commenced
- Purple highlight heading means content relevant to a certain slide package
- Green highlight heading means new main topic
- Blue highlight heading means new sub topic

Estimation

- **Estimate parameters of a distribution from the sample data**
  - e.g. given a sample of heights but haven't been told the distribution they have come from, and haven't told us what the population parameters are either
- Common approach is **maximum likelihood estimation**
  - Requires assuming we know the basic functional form

Want to do → assume a distribution and estimate the population parameters

Also want to → find out what distribution matches the sample the best

Common approach to doing that is MLE

Maximum likelihood estimate (MLE)

- Estimate the unknown parameter θ using the value that maximises the probability (i.e. likelihood) of getting the observed sample
  - Theta can have more than one element, depending on the distribution of the function we may have more than one population parameter
- Likelihood function
  - Likelihood function is made from multiplying the PDF of the distribution together over and over again, after the sample values drawn have been subbed into the function
  - The probability of observing the first draw from the distribution up to the nth draw from the distribution assuming that we know theta
  - Density function evaluated at each sample value and then the product of those

$$
\begin{aligned}
L(\theta) &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \theta) \\
&= f(x_1|\theta) f(x_2|\theta) \cdots f(x_n|\theta) \\
&= \prod_{i=1}^{n} f(x_i; \theta)
\end{aligned}
$$

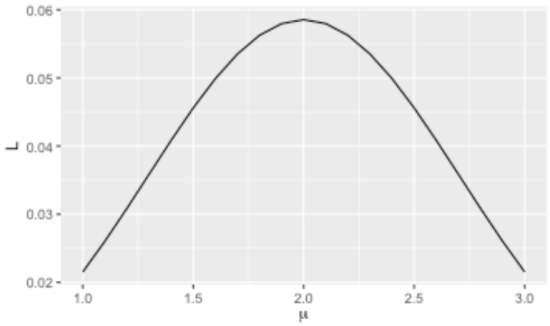
- This is now a function of θ.
- Use function maximisation to solve.

Example - Mean of normal distribution, assume variance is 1

- MLE estimate of the population mean for a normal model is the sample mean
- Run this numerically
- Suppose we have a sample of two: x1 = 1.0, x2 = 3.0
- Likelihood

$$L(\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(1.0-\mu)^2}{2}} \times \frac{1}{\sqrt{2\pi}} e^{-\frac{(3.0-\mu)^2}{2}} = \frac{1}{2\pi} e^{-\frac{(1-\mu)^2+(3-\mu)^2}{2}}$$

Plot it (i.e. Plot the likelihood function, and the peak of it corresponds to the maximum likelihood estimate for the distribution)



The maximum is at 2.0. This is the sample mean, which we can prove algebraically is the MLE.

Estimate mean and variance (i.e. now we don't know both of the parameters)

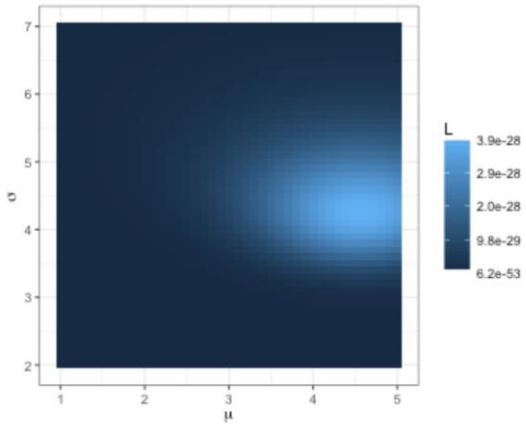
Sample

```
#>  [1]  7.31  3.96  2.34  0.55  5.12 10.33  3.74 -6.30 -1.14  6.55  9.13
#> [12] 10.34  5.93  1.72  3.71  3.68 -1.36 11.71  1.99  5.13  8.35  7.50
```

We know it comes from a normal distribution. What are the best guesses for the μ, σ? → need to compute the likelihood function by subbing in each of the 22 sample values into the distribution and then finding the product of those and finding the MLE of that overall function

Compute the likelihood for a range of values of both parameters.

For instance, this likelihood function has two variables, so when we plot it the third dimension here is showing the probability (i.e. the colour is most concentrated at the far right of the x-axis (mu) and part way up the y-axis (sigma)) if we were to visualise it, it would be like a mountain with the peak being the MLE (maximum likelihood estimate)



Quantiles

- quantiles are cut-points dividing the range of a probability distribution into contiguous (sharing a common border) intervals with equal probabilities
    - 2-quantile is the median (divides the population into two equal halves)
    - 4-quantile are quartiles, Q1, Q2, Q3, dividing the population into four equal chunks
- quantiles are values of the random variable X (actual values we might observe)
- useful for comparing distributions
    - **e.g. very useful for checking if our sample is consistent with the shape of the distribution (confirms the full distribution)**
- Also, can be useful for things such as test scores, i.e. knowing what value corresponded to 90% of people achieving above

Example:

- 12-quantiles for a N(0,1)

```
qnorm(seq(1/12,11/12,1/12))
#>  [1] -1.4e+00 -9.7e-01 -6.7e-01 -4.3e-01 -2.1e-01 -1.4e-16  2.1e-01
#>  [8]  4.3e-01  6.7e-01  9.7e-01  1.4e+00
```

- 23-quantiles from a Gamma(2,1)

```
qgamma(seq(1/23,22/23,1/23), 2)
#>  [1] 0.33 0.49 0.63 0.75 0.87 0.99 1.11 1.23 1.35 1.48 1.61 1.75 1.90 2.06
#> [15] 2.23 2.42 2.63 2.88 3.18 3.55 4.06 4.91
```

Percentiles

- indicate the value of X below which a given percentage of observations fall, e.g. 20th percentile is the value that has 20% of values below it
- 17th percentile from N(0,1)

```
qnorm(0.17)
#> [1] -0.95
```

- 78th percentile from Gamma(2,1)

```
qgamma(0.78, 2)
#> [1] 2.9
```

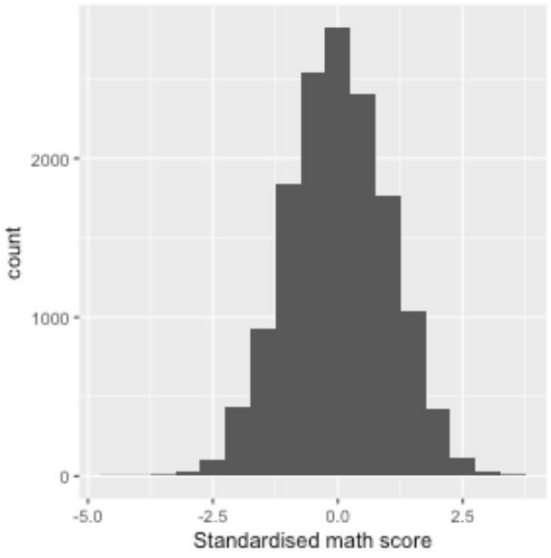
Goodness of fit → a reason for generating quantiles

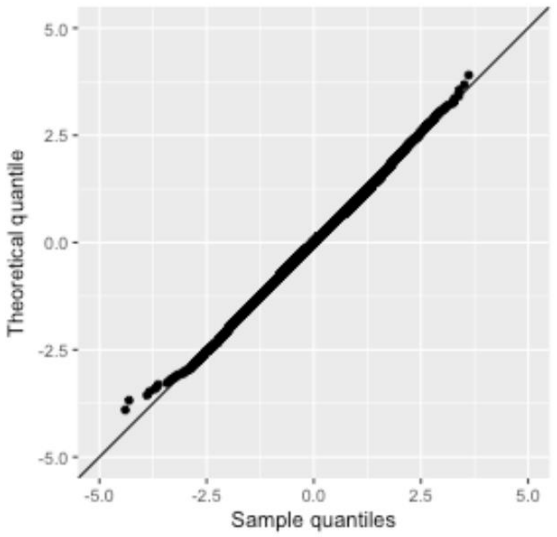
- "Have a sample of data and want to check if it is consistent with a specific probability distribution"
    - If data is consistent with distribution → can use that model to estimate probabilities
- **Quantile-quantile plot (QQplot) plots theoretical vs sample quantiles** (could even do sample vs sample if wanted to compare insurance products for instance)
    - "QQ" comparing quantiles from any distribution to quantiles from the sample
- Lets check the distribution of PISA math scores

<u>Method for checking if our data below is similar to a normal distribution in shape or not:</u>

Generate a sample from an actual normal distribution then calculate quantiles from those, and then **plot those theoretical quantiles against our sample quantiles** (we obtain our sample quantiles from sorting our data we are investigating from lowest to highest → the sample values are paired with the theoretical values)

i.e. the qqplot below says that our sample very closely matches what we would expect if we drew a sample from a normal distribution





Standardised values (i.e. mean 0, st 1 for normal dist) make it easier to understand QQ plot

Real data is not always perfect, in maths scores they are adjusted using some model (hence why the distribution is so perfect), privacy can also be protected through this