Tutorial 4: Statistical Theory; Why is the Gaussian Distribution so popular?

Rob Nicholls – nicholls@mrc-lmb.cam.ac.uk

MRC LMB Statistics Course 2014

Task 1:

For a random variable $X \sim N(\mu, \sigma^2)$, R's `pnorm` function can be used to calculate $P(X \leq x)$.

1. Assuming $X \sim N(10, 4)$, Calculate $P(X \leq 9)$.
2. Assuming $X \sim N(10, 4)$, Calculate $P(|X - 8| \leq 1)$.

Task 2:

The *covariance* between two random variables X and Y is given by:

$$Cov(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right)$$

where μ_X and μ_Y are the means of X and Y, respectively.

Show that the covariance may also be expressed:

$$Cov(X, Y) = E(XY) - \mu_X \mu_Y$$

Task 3:

An artificial random sample following a Binomial distribution can be generated using the `rbinom` function.

1. Generate a random sample following a Binomial $Bin(100, 0.4)$ distribution, containing 10 observations.
2. Given that the probability density function corresponding to an empirical sample may be estimated using the `density` function, generate a plot of the probability density function of your $Bin(100, 0.4)$ sample created above (note that the output of the `density` function can be directly plotted using the `plot` function, i.e. a command of the form: `plot(density(SOME_DATA_VECTOR))` is acceptable).
3. As a consequence of the Central Limit Theorem, the Binomial distribution may be approximated by the Normal distribution with the same mean and variance as the Binomial distribution (for n large and p not close to 0 or 1). Identify the values of the mean and variance of a $Bin(100, 0.4)$ distribution (hint – google "binomial distribution" to find the mean and variance).

4. We will now visually investigate using the Normal distribution to approximate a Binomial. Begin by generating a vector containing all potential outcomes (i.e. integers from 0 to 100). The `dnorm` function is used to calculate the value of the Normal probability density function corresponding to a particular outcome/observation. Plot the Normal density function that would be used to approximate that of the $Bin(100, 0.4)$ distribution.

5. Using the `lines` function to allow overplotting, display both the Normal density function that would be used to approximate that of the $Bin(100, 0.4)$ distribution and the random sample of 10 $Bin(100, 0.4)$-distributed observations created earlier. Do the two density functions appear to correspond well?

6. Repeat the previous step, this time comparing the Normal density function with that of a random sample of 100 $Bin(100, 0.4)$-distributed observations. Does the Normal distribution appear to approximate the Binomial well?

7. Use a 'for loop' to iteratively recreate a new random $Bin(100, 0.4)$ sample of 100 observations and overplot the corresponding empirical density functions onto the plot of the Normal density function. The 'for loop' should iterate 50 times. Note – 'for loop' syntax:

   $$\texttt{for(i in 1:n)\{***INSERT\_COMMAND(S)***\}} \tag{1}$$

   Has your opinion changed regarding whether or not the Normal distribution appears to approximate the Binomial well?

8. Use the Normal distribution to estimate $P(X \leq 35)$ given that $X \sim Bin(100, 0.4)$.

9. Use the Binomial distribution to calculate $P(X \leq 35)$. How does this compare with the previous estimate?

10. Use a 'for loop' to iteratively estimate $P(X \leq 35)$ from random samples of 100 $Bin(100, 0.4)$-distributed variates, iterating 100 times, storing the resultant distribution of estimates in a vector. Plot the density function corresponding to this distribution of estimates. What is the mean of this distribution? What is the standard deviation of this distribution? Where do the values calculated in parts 8 and 9 (i.e. the actual value of $P(X \leq 35)$, and the estimate from the Normal approximation) lie in relation to this distribution? What conclusions can you draw?

Task 4:

The moment generating function (MGF) uniquely characterises a distribution – if the MGF of X is equal to the MGF of Y then both X and Y follow the same distribution. Note that the MGF of a standard Normal random variable X is:

$$m_X(t) = e^{t\mu + \frac{1}{2}t^2\sigma^2}$$

Use this information to show that if $X \sim N(\mu, \sigma^2)$ and $Y = \alpha X + \beta$ then:

$$Y \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$$

for some constants α, β, utilising the knowledge:

$$m_{f(X)}(t) = E(e^{f(X)t})$$

and

$$m_{\alpha X}(t) = m_X(\alpha t)$$