Final Project

He & Xu

10/25/2018

In our dataset, a number of smoke or non-smoke data is missing. So our goal is to find wether the data is come from smoke or non-smoke and their parameters μ and σ.Assume Smoker ~ $N(\mu_1, \sigma_1^2)$, Non-Smoke ~ $N(\mu_2, \sigma_2^2)$ $\theta_1 = [\mu_1, \sigma_1]^T$ $\theta_2 = [\mu_2, \sigma_2]^T$ $\theta = [\theta_1, \theta_2]$ –> the parameter need to be estimate. Let Z be the missing data, and set when $z = 1$,the data comes from smoke, else when $z = 0$ comes from non-smoke.

set initial parameters: $\mu_1 = 2000$, $\mu_2 = 1000$,$\sigma_1 = 100$, $\sigma_2 = 100$.

We need to design a cut-off point C, based on the normal gragh, to determine x_i belongs to which group, somker or non-smoker.

After grouping our initial data, we can calculate the MLE of μ and σ for both group.

Then we use $\hat{\mu}$ and $\hat{\sigma}$ to re-seclect group. We use Newtow iteration method until $\hat{\mu} = \mu$ and $\hat{\sigma} = \sigma$.

Finally, we will have mean cost and variance for smoker and non-smoker.