上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

上海交通大学

《概率论与数理统计》

大作业

姓名： 姚迪熙

学号： 518021910367

年级： 2018 级

所在院系： 电子信息与电气工程学院

任课教师： 仇璘

2019 年 12 月

概率统计小论文

对于患有心血管疾病的人群的BMI分布的研究

电子信息与电气工程学院
姚迪熙
518021910367

2019 年 12 月 18 日

目录

1 Introduction

1.1 研究内容

研究患有心血管疾病的人群，他们身体的BMI指数是如何分布的，并利用本学期课程中所学习到的方法，来对于分布情况进行一个简单的研究。

1.2 研究背景

当今世界上患有心血管疾病的人口越来越多，心血管疾病已经成为了全球疾病的头号死因，甚至超过癌症，因此如果可以能够早期的发现问题，及时治疗是非常有必要的。采样数据来源于世界知名网站kaggle上面所提供的Crdiovascular Disease dataset[1]该数据集记录了70000个来医院参与心血管疾病检查的病人的采样结果。

[1]https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

1.3 研究意义

如今人们对与身体情况越来越来约关心，通过统计方法建立一个BMI与患病情况的关系，使得人们在知道自己身高体重的时候可以能依照一定的统计学结果，做一些简单初步的判断比如，是否有概率患有心血管疾病以及是否有必要去医院做相关的检查。

2 Solution

2.1 原始数据处理与初始假设

因为本研究对象是患有心血管疾病人群，故采集样本中患有疾病的人群，经过筛选后共34979名病人，并且计算他们的BMI作为需要用的数据。

$$BMI = \frac{Weight}{Height^2} \tag{1}$$

由于实际人口远远大于一家医院的病人数量，因此依照*Lindeberg-Levy Central Limit Theorem*可以将总体的BMI分布近似为正态分布。

基础假设: 患有心血管疾病的病人BMI X服从正态分布$N(\mu, \sigma^2)$

部分的计算工作由python完成，代码地址[2]

2.2 样本处理

采用全部数据作为样本的分布如下

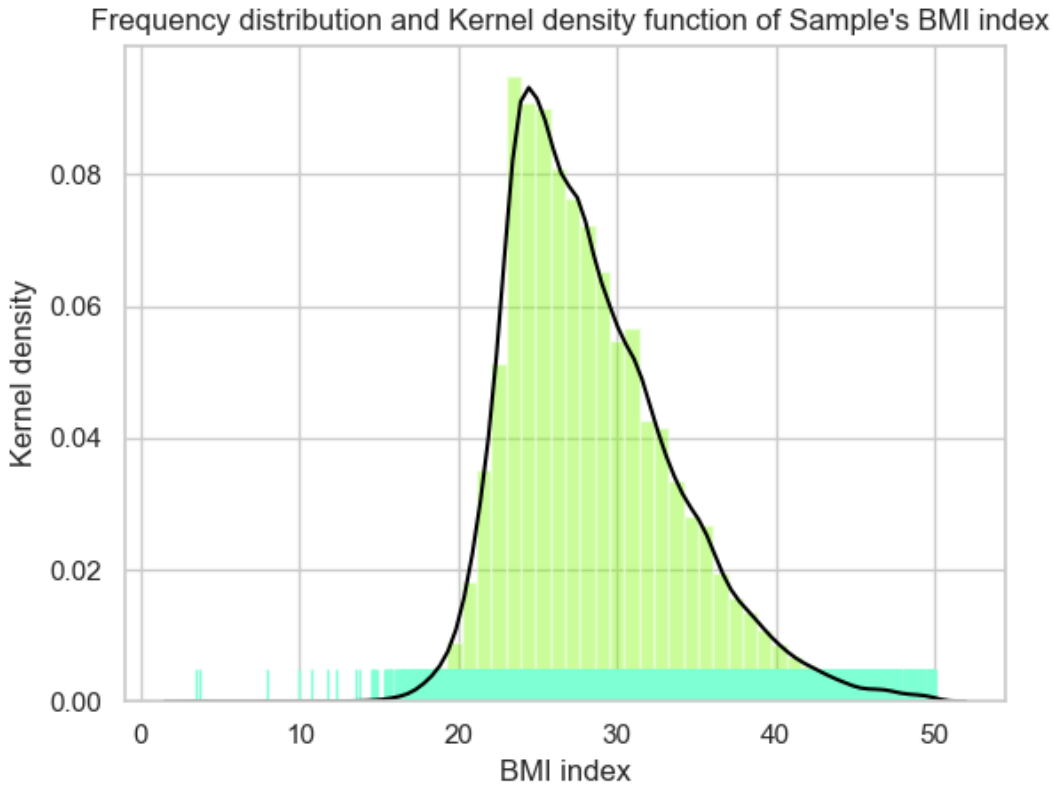

图 1: 全部数据样本的BMI指数频度分布与核密度函数

采用随机抽样观察的方法，样本容量 n=300 ，通过从34979个数据中随机抽取的方式，模拟从总体中进行随机抽样，共采集100组样本。

[2]https://jbox.sjtu.edu.cn/l/UHkD1m(密码：nksw)

随机抽样观察

样本空间: $(\Omega_1, \Omega_2 ... \Omega_{100})$

每个样本空间内样本: $(X_1, X_2 ... X_{300})$

关于各组样本空间的平均值所计算的结果见附录表1，接下来计算样本期望，方差

$$E(\bar{X}) = 28.582739434952963$$
$$D(\bar{X}) = 0.14169727784519617$$

又依据$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$可以计算获得$\mu = 28.58$,$\sigma = 6.51991$记为$\mu_a$,与$\sigma_a$。与之后通过参数估计所获得的结果进行比较。

2.3 参数估计

估计参数μ,σ

本小节所使用的样本空间包含全部34979个样本

2.3.1 点估计法

利用两种方法估计，矩估计法和最大似然估计

矩估计法: 已知

$$\begin{cases} \hat{\mu} = \frac{1}{n}\Sigma_{i=1}^{n} X_i = \bar{X} \\ \hat{\sigma^2} = \frac{1}{n}\Sigma_{i=1}^{n} X_i - \bar{X}^2 \end{cases} \tag{2}$$

求解得到$\hat{\mu}$=28.56606062687427,$\hat{\sigma}$=6.3835743883812315,记为μ_b与σ_b

最大似然估计法: X的密度为$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

从而似然函数$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$

$lnL(\mu, \sigma^2) = -\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(x_i - \mu)^2$

似然方程组为:

$$\begin{cases} \frac{\partial}{\partial \mu} lnL = \frac{1}{\sigma^2}\Sigma_{i=1}^{n}(x_i - \mu) = 0, \\ \frac{\partial}{\partial \sigma^2} lnL = \frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\Sigma_{i=1}^{n}(x_i - \mu)^2 = 0; \end{cases} \tag{3}$$

解的结果与矩估计法所获得的结果相同，故点估计法所估计获得的参数采用μ_b与σ_b

2.3.2 估计量的评价

无偏性:

因为$E(\mu_b) = \mu$,所以μ_b是μ的无偏估计，但是验证发现$E({\sigma_b}^2) \neq \sigma^2$,所以需要进行修正。$E({\sigma_b}^2) = \frac{n-1}{n}\sigma^2$，修正${\sigma_b'}^2 = \frac{n}{n-1}{\sigma_b}^2$,解得$\sigma_b' = 6.394240351358263$

有效性:

依据Rao-Cramer不等式,

$D(\mu_b) = G$,所以μ_b是μ最有效的估计量。

$D({\sigma_b'}^2) = \frac{2\sigma^4}{n-1}$, 下计算对应的G

$E([(\frac{\partial lnL}{\sigma^2})^2] = \frac{n^2}{4\sigma^4} + \frac{n}{2\sigma^6}E(\Sigma(x_i - \mu)^2) + \frac{1}{4\sigma^8}E(\Sigma(x_i - \mu)^4) = (\frac{n^2}{4} + n + \frac{1}{4})\frac{1}{\sigma^4}$

$G = \frac{\sigma^4}{\frac{n^3}{4}+n^2+\frac{n}{4}}$,由于n很大，所以$\sigma_b'$已经非常的接近最有效估计量了

一致性:

$$\lim_{x\to+\infty} D(\mu_b) = \lim_{x\to+\infty} \frac{\sigma^2}{n} = 0$$
$$\lim_{x\to+\infty} D({\sigma_b'}^2) = \lim_{x\to+\infty} \frac{2\sigma^4}{n-1} = 0$$

所以μ_b与σ_b'是一致估计量

2.3.3 区间估计

取置信度为0.95
方差σ^2未知，求均值μ的置信区间为:

$$(\bar{X} - t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}})$$

T分布在n=300的时候近似为标准正态分布,查表可知$u_{0.025} = 1.96$
解得μ的置信区间是$(27.8425, 29.2896)$
均值μ未知，求方差σ^2的置信区间为:

$$(\sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}})$$

卡方分布的极限是正态分布$\chi^2 \sim N(n, 2n)$,所以$\chi^2_{\frac{\alpha}{2}}(n-1) = 301.96, \chi^2_{1-\frac{\alpha}{2}}(n-1) = 298.04$
解得σ的置信区间是$(6.3628, 6.4045)$

2.3.4 综上所述

对于参数的估计如下(保留四位小数,置信度0.95)

$$\begin{cases} \hat{\mu} = 28.5661, \text{置信区间}(27.8425, 29.2896) \\ \hat{\sigma} = 6.3942, \text{置信区间}(6.3628, 6.4045) \end{cases} \tag{4}$$

下为拟合出的图像

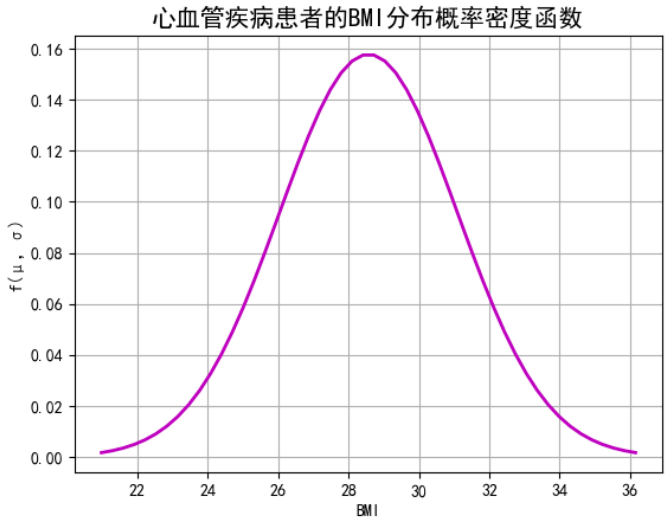

图 2: 心血管患者BMI分布

2.4 贝叶斯概率计算

已经获得对于患有心血管疾病的人群的BMI指数的正态分布$N(28.5661, 6.3942^2)$ 记事件A表示BMI指数为A，事件B表示是否患有心血管，则P(A)表示BMI指数是A的概率，$P(B_i)$表示患有心血管疾病的概率，B_0表示不患有心血管疾病，B_1表示患有心血管疾病，依据*Bernoulli's law of large numbers*，用世界人口中BMI指数为A的人所占的比例代替P(A),用患有心血管疾病人数的比例来代替$P(B_1)$,因而$P(A|B_i)$表示确认是否患有心血管疾病的情况下，BMI是A的概率，也就是本文主要做的工作。而$P(B_i|A)$则是知道BMI指数下，是否患有疾病的概率，也就是本节所要推导的量。查阅资料[3]可知目前全球患有心血管的人口比例大约为四分之一，由贝叶斯公式$P(B_i)|A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A|B_i)}{P(A)}$
可以获得如下的公式

$$\text{X表示BMI指数}$$
$$\begin{cases} P(A|B_i) = \Phi(\frac{X-28.5661}{6.3942}) \\ P(B_i) = 0.25 \\ P(A)\text{查资料获得} \end{cases} \tag{5}$$
$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(A)}$$

因而对于患有心血管疾病有个粗略的估计, 比如说我的BMI指数是22.64，大概在人口中可以占到三分之一的比例，计算得我患有心血管疾病的概率是13.35%，还是比较小的，也很符合我的BMI指数，因为我经常锻炼，身材不胖不瘦，因而确实患病几率不高，不需要太过担心。

3 Conclusion

在本次研究中研究了患有心血管疾病的人群的BMI指数是如何分布的，但是在选取样本容量，以及选取样本的时候还有很多的讲究，此外在进行参数估计的时候如何选取到最有效的估计量也是很重要的。总体上可以完成初定的目标，通过样本研究出总体的分布情况是什么情况。能够得出总体的分布规律，对参数有较为准确的估计。并且通过贝叶斯概率公式给出了计算公式，使得在知道BMI指数的情况下可以对于患心血管疾病的概率有一定规模的预测。

当然心血管疾病并不只是和一个人的BMI指数有关，还有很多因素包括生活作息，生活习惯，血压血常规，先天性因素等非常多的问题，因而在未来可以不管是研究单因素，而是对于多变量做一个统计的分析，能够给出更加准确的判断和预测。

除此之外还存在一些小问题，比如对于世界人口患有心脏病人数的准确统计，以及不同BMI指数人群在人口中所占的比例的数据不是绝对精确，因而会使得结果有偏差，本文仅仅是给出一个示例，具体在使用公式(5)的时候要结合具体情况进行判断。

另外，本人在本学期的另外一门课程CS241 问题求解与实践课程中使用机器学习的方法对于任意病人给出准确判断是否会患有心血管疾病，如有兴趣欢迎交流！[4]

4 Appendix

表 1: 各组样本均值计算

$\bar{X}_i$
28.3 29283765833328

[3] https://www.who.int/cardiovascular_diseases/about_cvd/zh/

[4] Jimmyyao18@sjtu.edu.cn

28.9784426416
27.82847335513333
28.11674664326667
28.50847043303333
28.280198193863338
28.167908558633336
28.4948710393
28.752732427699993
28.314091194899998
28.676760823733336
28.7225593215
28.381409129863332
28.654830374733333
29.05491875006667
28.533332222033334
28.944156838499996
28.758601053899994
28.489765134466666
28.3846108347
28.34932426463333
28.891147993266664
28.32229233076666
28.444907775833332
28.395341197766673
28.49847837403333
28.37656800856667
28.718232274233337
27.96345641383333
27.887611262866667
28.78980783503334
28.38589030906667
28.764212429466667
28.426702325266664
28.451903102333333
28.283119187833332
28.5344558002
28.41031383196667
28.80911396106666
28.987207236833335
28.456205628939994
28.719859732866666
28.616885945433328
28.780273067566668
28.16765646916667

28.443707421766668
29.07305433076667
29.0730562547
28.780260017233328
28.48600626653333
28.75796011806667
28.306060821533332
28.6795070163
28.965131014266667
28.132266635666667
28.91505180536667
28.38254818006667
28.32855779806667
28.8300897125
28.5407238514
28.571171509400003
29.10002587803334
28.327161792833333
28.93727640216667
28.4874741486
29.200191787366663
29.376269214500002
28.694063457333336
28.56132718946667
28.800287130466664
28.77983916173333
28.085781279373332
29.05167716343333
28.807717504533333
28.124188395833336
28.319898108733334
28.5038875393
28.926210005166666
28.811143027766665
28.884301787333335
28.31141044103334
28.88935234038666
28.723681641366664
28.5986866634
28.4600417023
28.881403291933335
28.017528915933333
27.882760941966662
28.515788417599996

28.533226413166666
29.079139849033332
28.460365138433335
28.681343536933333
28.972148809799993
28.4027801966
28.185999713066668
28.7322584932
28.66094745083333
28.054950319233335
28.770745023466667