数学实验第九次实验报告

计 76　张翔　2017011568

2020 年 5 月 23 日

1　实验目的

1. 掌握数据的参数估计、假设检验的基本原理、算法，及用 MATLAB 实现的方法;

2. 练习用这些方法解决实际问题。

2　Ch12-P5 供货问题

2.1　问题分析与模型建立

对于题中所给的产品，只有合格品与不合格品之分，可用 $X = 1$ 表示合格，$X = 0$ 表示不合格，那么总体 X 服从 0-1 分布。若合格率为 p，则 $\mu = EX = p, \sigma^2 = DX = p(1-p)$。虽然 X 不服从正态分布，但根据中心极限定理，当样本容量 n 充分大时，对样本均值 $\overline{x}$ 有，$z = \frac{\overline{x}-\mu}{\sigma/\sqrt{n}}$ 近似服从 $N(0,1)$。由于甲方承诺合格率为 $\mu_0 = 90\%$，可以对总体合格率 p 作如下的假设检验:

$$H_0 : \mu \geq \mu_0 = 0.9;\ H_1 : \mu < \mu_0 = 0.9$$

由于这里是合格率问题，而合格率高于约定的标准均是可接受的，因此在这里采用假设检验的单侧检验形式。

这里可以利用正态总体均值的 z 检验，取 $N(0,1)$ 的 α 分位数 u_α，设样本的合格率为 $\overline{x}$，记 $z = \frac{\overline{x}-\mu_0}{\sqrt{\mu_0(1-\mu_0)/n}}$ 满足 $P(z \geq u_\alpha) = 1-\alpha$，那么假设检验的规则为

$$\text{当} z \geq u_\alpha \text{时接受} H_0\text{；否则拒绝} H_0\ \text{（接受} H_1\text{）}$$

如果在某一条件下乙方接受货物，改变显著性水平 α，总体均值 μ_0 以及抽样数目 n 均能对是否接受该批货物造成影响。

2.2　算法设计

本题需要 z 检验,但没有提供原始数据,无法直接使用 `ztest`,这里需要自己编写检验函数 `exp_ztest`。对于第 (2) 问，可以改变显著性水平、总体均值以及抽样数目，讨论三者对于结果的影响。由于题目没有给出更多的数据，可以假设抽样数目 n 增加后，样本的合格率仍维持原来的水平。

编写的 `exp_ztest` 返回值为 `[h, sig]`，在题给的情况下，若 sig $> \alpha$，此时 H_0 会被接受，反之则被拒绝。

2.3　Matlab 程序

编写的检验函数如下

```
function [h, sig] = exp_ztest(mu, xbar, n, alpha, tail)
%% normalize
z = (xbar - mu) / sqrt(mu * (1 - mu) / n);
%% test procedure
switch tail
    case 0
        % x == mu
        u = norminv(1-alpha/2);
        sig = 2 * (1 - normcdf(abs(z)));
        if abs(z) <= u
            h = 0;
        else
            h = 1;
        end
    case 1
        % x <= mu
        u = norminv(1-alpha);
        sig = 1 - normcdf(z);
        if z <= u
            h = 0;
        else
            h = 1;
        end
    case -1
        % x >= mu
        u = norminv(alpha);
        sig = normcdf(z);
        if z >= u
            h = 0;
        else
            h = 1;
        end
end
end
```

第 (1) 问代码如下

```
%% input data
mu = 0.9;
xbar = 43 / 50;
n = 50;
alpha = 0.05;
tail = -1;

%% solve
```

```
[h, sig] = exp_ztest(mu, xbar, n, alpha, tail);
```

第 (2) 问代码在 (1) 的基础上，增加如下部分，对不同参数进行迭代

```
%% change params and solve
alpha_list = 0.05:0.0001:0.50;
for i=1:length(alpha_list)
    [h, sig] = exp_ztest(mu, xbar, n, alpha_list(i), tail);
    if h == 1
        fprintf("Found alpha: %.4f\n", alpha_list(i));
        break;
    end
end

mu_list = 0.9:0.0001:1;
for i=1:length(mu_list)
    [h, sig] = exp_ztest(mu_list(i), xbar, n, alpha, tail);
    if h == 1
        fprintf("Found mu: %.4f\n", mu_list(i));
        break;
    end
end

n_list = 50:1:10000;
for i=1:length(n_list)
    [h, sig] = exp_ztest(mu, xbar, n_list(i), alpha, tail);
    if h == 1
        fprintf("Found n: %d\n", n_list(i));
        break;
    end
end
```

2.4 计算结果与分析

第 (1) 问的计算结果为 $h = 0, sig = 0.1729$，即接受 H_0，可以认为该批货物的合格率不低于约定值 90%，说明乙方应该接受甲方的货物。注意到此时 sig 比显著性水平 α 大得多。

第 (2) 问通过对不同参数进行迭代（每次只改变一种参数，其余参数与 (1) 问相同），得到不接受 H_0 时的各个参数值，如下

提高显著性水平 α 计算得显著性水平上升到 $\alpha = 0.1729$ 时，H_0 不被接受，此时刚好不满足 $sig > \alpha$，相当于双方约定的置信概率降低到不超过 82.71%。

增大约定的总体合格率 μ_0 计算得到 $\mu_0 = 0.9223$ 时，H_0 开始不被接受，只要约定合格率不低于该值，就能拒绝 H_0。

增大抽样数量 n 当 $n = 153$ 时，H_0 开始不被接受。理论上只要乙方增加抽检数，使其不低于 153，即可拒绝该批货物，但由于抽样存在偶然性，实际操作时可能影响样本的合格率均值 $\overline{x}$，从而不一定能达成拒绝该批货物的目的。

上述三种拒绝货物的方法中，提高显著性水平 α（降低置信概率）使得样品均值落入 H_0 接受域的概率减小，但增大了原本成立 H_0 被拒绝（第一类错误）的概率，如果甲方不够可靠，适当提高 α 是合理的；提高总体均值 μ_0 的方法直观上容易理解，当均值增大后，$z \geq \mu_0$ 的概率减小，H_0 越不容易被接受；提高抽样数目 n 实际上减小了样本方差，使得样本更能代表总体，如果次数样本均值不变且低于总体均值，则该批货物不符合要求的概率将增大。

2.5 结论

1. 在题给条件下，接受 $H_0 : \mu \geq \mu_0 = 0.9$，即乙方应该接受这批货物；

2. 要使得乙方不接受该批货物，则可以采用下列三种方式之一（其余参数保持不变）

   - 使得双方约定的置信概率不超过 82.71%；
   - 甲方承诺的总体合格率不低于 92.23%；
   - 增加抽样数目，使得样本容量不低于 153。

3 Ch12-P6 身高体重估计

3.1 问题分析与模型建立

一般情况下，可以认为群体的身高体重总体上为正态分布。本题的样本容量为 $n = 100$，可以使用 Jarque-Bera 检验。检验为正态分布后，可以对样本进行参数估计，得到总体均值和区间估计。

对于第 (3) 问，10 年前的数据为普查结果，即为总体的均值，记为 μ_0，为回答学生的身高/体重是否发生变化，可对二者作假设检验

$$H_0 : \mu = \mu_0;\ H_1 : \mu \neq \mu_0$$

由于方差未知，上述问题可以使用 t 检验，由于 $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} \sim t(n-1)$，取 $t(n-1)$ 的分位数 $t_{1-\alpha/2}$，满足

$$P(|t| \leq t_{1-\alpha/2}) = 1 - \alpha$$

t 检验规则为

$$\text{当} |t| \leq t_{1-\alpha/2} \text{ 时接受} H_0\text{，否则拒绝}$$

3.2 算法设计

对于第 (1) 问，可以使用 MATLAB 的 `histfit` 函数对样本数据进行作图，它能作出样本直方图，并估计正态分布曲线。对于正态分布检验，由于样本容量不小，可以使用 Jarque-Bera 检验，使用 `jbtest` 函数即可。

对于第 (2) 问，在第 (1) 问检验了样本分布的正态性后，可以使用 MATLAB 的 `normfit` 函数对正态分布的参数和区间进行估计。

对于第 (3) 问，由于提供了样本的原始数据，使用 MATLAB 的 `ttest` 函数进行 t 检验即可。

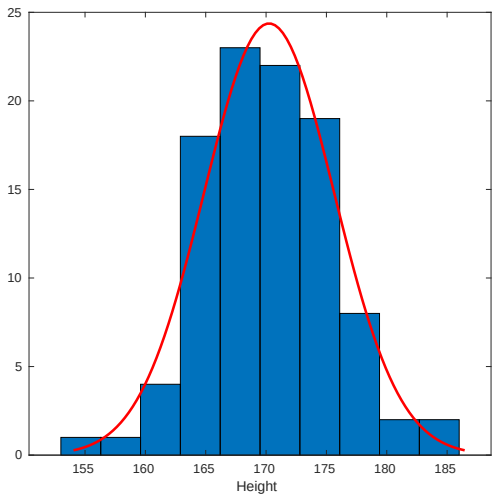
3.3 Matlab 程序

3 个小问的代码如下

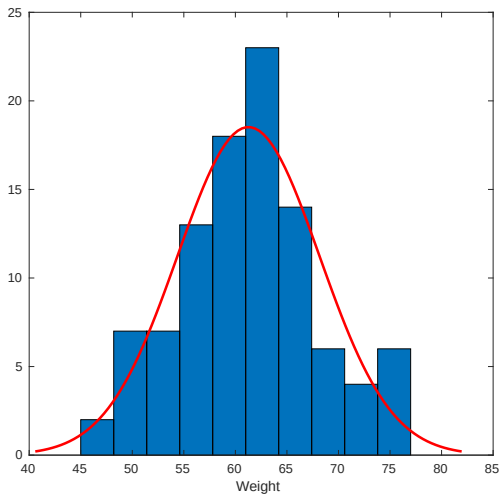
```
%% import data
data = readmatrix('data.txt');
height = reshape(data(:, 1:2:end), [], 1);
weight = reshape(data(:, 2:2:end), [], 1);

%% plot
figure(1);
histfit(height);
xlabel('Height');
figure(2);
histfit(weight);
xlabel('Weight');

%% norm test
h_height = jbtest(height);
h_weight = jbtest(weight);

%% norm fit
[mu_h, sigma_h, muci_h, sigmaci_h] = normfit(height, 0.05);
[mu_w, sigma_w, muci_w, sigmaci_w] = normfit(weight, 0.05);

%% test hypothesis
[h_h, p_h, ci_h, stats_h] = ttest(height, 167.5);
[h_w, p_w, ci_w, stats_w] = ttest(weight, 60.2);
```

3.4 计算结果与分析

第 (1) 问 身高体重样本的直方图如下



(a) 身高分布



(b) 体重分布

从上图可以看出，直方图与拟合的正态分布曲线（红色）比较接近，身高体重的分布均是近似正态的。使用 Jarque-Bera 检验，取显著性水平 $\alpha = 0.05$，均可以得到 $h = 0$，即接受 H_0（样本是从正态分布的总体中抽取的），与直方图相符。

第 (2) 问 可以取不同的显著性水平 α 对样本进行估计，得到的身高与体重的估计如表1与2所示。

显著性水平 α	均值点估计	标准差点估计	均值区间估计	标准差区间估计
0.05	170.25	5.4018	[169.18, 171.32]	[4.7428, 6.2751]
0.03	170.25	5.4018	[169.06, 171.44]	[4.6787, 6.3799]
0.01	170.25	5.4018	[168.83, 171.67]	[4.5590, 6.5904]

表 1: 不同显著性水平 α 下身高正态分布参数的估计

显著性水平 α	均值点估计	标准差点估计	均值区间估计	标准差区间估计
0.05	61.27	6.8929	[59.90, 62.64]	[6.0520, 8.0073]
0.03	61.27	6.8929	[59.75, 62.79]	[5.9702, 8.1410]
0.01	61.27	6.8929	[59.46, 63.08]	[5.8175, 8.4096]

表 2: 不同显著性水平 α 下体重正态分布参数的估计

可以看出，均值与标准差的点估计是固定的，而区间估计的区间长度则显著性水平的下降而上升，这与区间估计的理论是符合的。

第 (3) 问 选择显著性水平 $\alpha = 0.05$，可以得到，对于身高，$h = 1, p = 1.7 \times 10^{-6}$；对于体重，$h = 0, p = 0.1238$。说明对于身高，拒绝 H_0 假设，即学生的身高较 10 年前有了明显变化；对于体重，接受 H_0，即体重较 10 年前没有明显变化。

注意到对于身高来说，即使令 $\alpha = 0.01$，t 检验的 p 值 $p = 1.7 \times 10^{-6} \ll \alpha$，说明 H_0 大概率是不可靠的，即身高发生了变化；而对于体重，$p = 0.1238$，即使将显著性水平 α 增大，也仍有较多的余量使得 $p > \alpha$，使得最终接受 H_0，即体重没有发生明显变化。

这一点也可以从第 (2) 问中 $\alpha = 0.05$ 时得到的均值的置信区间得到，容易发现，167.5 不在身高的置信区间内，而 60.2 在体重的置信区间内，这与假设检验的结果是吻合的。

3.5 结论

1. 身高和体重的分布具有正态性，通过上述的直方图或 Jarque-Bera 检验均可以验证；
2. 在显著性水平为 $\alpha = 0.05$ 时，学生身高均值的点估计为 170.25，置信区间 [169.18, 171.32]（单位均为 cm）；标准差的点估计为 5.4018，置信区间为 [4.7428, 6.2751]。学生体重的的点估计为 61.27，置信区间 [59.90, 62.64]；标准差的点估计为 6.8929，置信区间为 [6.0520, 8.0073]（单位均为 kg）。
3. 根据 t 检验的结果，学生的平均身高有明显变化，体重无明显变化。

4 Ch12-P7 溶菌酶含量

4.1 问题分析与模型建立

假设患胃溃疡病人总体的溶菌酶含量与正常人总体的溶菌酶含量均服从正态分布，从题给的数据可以估计总体的均值，但各自的方差未知。记两类人群的溶菌酶含量均值、方差分别为 $\mu_1, \sigma_1, \mu_2, \sigma_2$，假

设 $\sigma_1^2 = \sigma_2^2$ 且未知，则问题转化为作假设检验

$$H_0 : \mu_1 = \mu_2;\ H_1 : \mu_1 \neq \mu_2$$

由于两组样本是从不同总体中抽取的，这是两总体均值的假设检验，此时 $t = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{s^2}{n_1}+\frac{s^2}{n_2}}} \sim t(n_1 + n_2 - 2)$，其中 $s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$。取 $t(n_1+n_2-2)$ 的 $1-\alpha/2$ 分位数 $t_{1-\alpha/2}$，则假设检验的规则为

当$|t| \leq t_{1-\alpha/2}$ 时接受H_0；否则拒绝H_0 （接受H_1）

4.2 算法设计

上述 t 检验的假设较强，它要求两组样本的分布均有正态性，且 $\sigma_1^2 = \sigma_2^2$。本题样本数较小 $(n = 30)$，不适合使用 Jarque-Bera 方法检验正态性，可以使用 Lilliefors 检验。如果符合正态分布，可以使用 MATLAB 自带的 F 检验函数 `vartest2` 检验方差是否相等。

在上述条件满足的情况下，使用 `ttest2` 可以对两组样本进行 t 检验，从而判断是否接受假设 H_0。

4.3 Matlab 程序

代码如下，第 (2) 问只需在此基础上，删去病人组的最后 5 个数据即可

```
%% import data
data = readmatrix('data_enzyme.txt');
patients = reshape(data(1:3, :), [], 1);
normals = reshape(data(4:6, :), [], 1);

%% print stats
fprintf('mu1=%.4f, mu2=%.4f\n', mean(patients), mean(normals));

%% t-test
[h, p, ci, stats] = ttest2(patients, normals, 0.05);
```

4.4 计算结果与分析

使用 `histfit` 函数作出两组数据的直方图如下

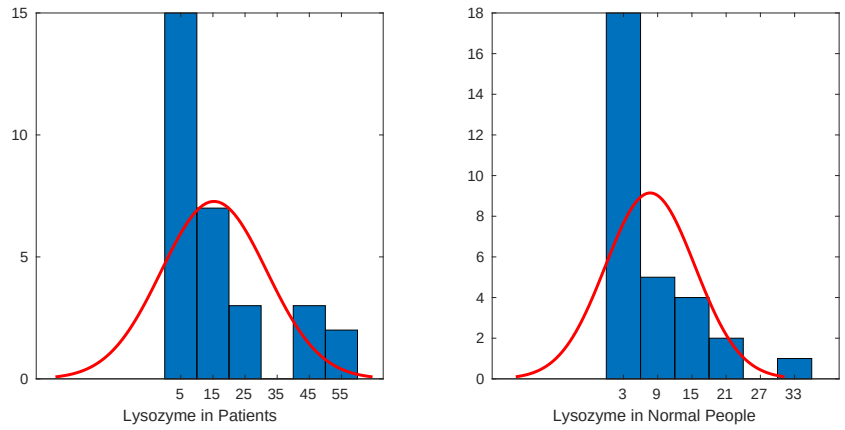


图 1: 胃溃疡病人/正常人溶菌酶含量的分布

使用 `lillietest` 均得到 $h=1$，结合上图可以判断样本的分布不符合正态分布。使用 `vartest2` 可以得到 $h=1$，说明拒绝 $\sigma_1^2=\sigma_2^2$ 的假设。但为了分析能够继续，仍然使用 `ttest2` 对样本进行检验。

第 (1) 问 在显著性水平为 $\alpha=0.05$ 时，t 检验得到 $h=1, p=0.0251, ci=[0.9886, 14.3114]$，说明此时拒绝 H_0。但由于 p 值不够小，如果 $\alpha=0.01$，此时将转为接受 H_0。计算两组样本的均值得到 $\mu_1=15.33, \mu_2=7.68$，相差较大。

第 (2) 问 删除胃溃疡病人组的最后 5 个数据后，在显著性水平为 $\alpha=0.05$ 时，t 检验得到 $h=0, p=0.1558, ci=[-1.5035, 9.1528]$，说明此时接受 H_1。此时两组样本的均值为 $\mu_1=11.51, \mu_2=7.68$，相比第 (1) 问更为接近一些，但仍然有较大差异。

讨论 两问的结果相差较大，第 (1) 问认为胃溃疡病人的溶菌酶含量与正常人有显著差别，从置信区间可以看出，胃溃疡病人的溶菌酶含量高于正常人；而第 (2) 问虽然二者均值仍然相差较大，但结论为胃溃疡病人的溶菌酶含量与正常人没有显著区别，且此时 p 值相比显著性水平大许多，说明接受 H_0 是相对可靠的。由此可以看出，只用样本均值来判断总体情况是不科学的。

第 (1) 问的 $p=0.0251$，在显著性水平 α 较小时，会产生相反的结论（接受 H_0），但考虑到本题所给数据的样本容量较小，且分布没有良好的正态性，α 不应取太小的值。

对比两问，虽然二者其他条件均相同，但第 (2) 问删去 5 个数据就得到了不同的结论，说明了原始数据的重要性。在统计推断前，不仅要通过科学的手段获得更全面的原始数据，还要通过分析去除不合理的数据。否则，相应统计推断结果的意义不大。本题由于样本容量较小，且可能包含了一些错误的数据，使得样本既无法通过正态检验，也无法通过 F 检验，导致最终 t 检验的结果不那么可靠。

4.5 结论

在题给条件下（显著性水平 $\alpha=0.05$），结论如下

1. 使用全部数据，认为胃溃疡病人的溶菌酶含量与正常人有显著差别；
2. 删除胃溃疡病人的最后 5 个数据，可认为胃溃疡病人的溶菌酶含量与正常人没有显著差别。

5 收获与建议

通过这次的实验，我掌握了使用 MATLAB 求解参数估计和假设检验问题的一般方法，并对概率论和数理统计的知识有了更深的理解。建议：新版 MATLAB 支持单个总体/两总体的方差检验（`vartest`/`vartest2`），希望老师能对课本上相应内容进行更新。