统计与数据科学学院

《大数据的统计学基础》第三次作业

姓　　名：蒋贵豪

年　　级：2021 级

专　　业：应用统计学

学　　号：B+X9bo

完成日期：2021 年 12 月 27 日

目录

1 假设检验

题目 1. 对于从一个均值为 μ 和已知方差为 σ^2 的正态总体中抽取的样本量为 1，4，16，64，100 的样本，画出以下假设的 LRT 的功效函数图形，取 $\alpha = 0.05$。

(a) $H_0 : \mu \leq 0$ 对 $H_1 : \mu > 0$。

(b) $H_0 : \mu = 0$ 对 $H_1 : \mu \neq 0$。

解答. 由于方差 σ^2 是已经给定的，我们不妨设：$\sigma^2 = 1$。

(a) 计算该**单侧检验下的功效函数为：**

$$\beta(\mu) = \mathbb{P}\left(\frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} > u_{1-\alpha}\right) = \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} > u_{1-\alpha} - \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}\right) = \mathbb{P}\left(Z > u_{1-\alpha} - \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}}\right) \tag{1.1}$$

其中 $Z \sim N(0,1)$ 为标准正态分布，$\mu_0 = 0$。通过编写 **Matlab** 代码，我们得到了**单侧检验下的 LRT 功效函数如图1.1所示。**

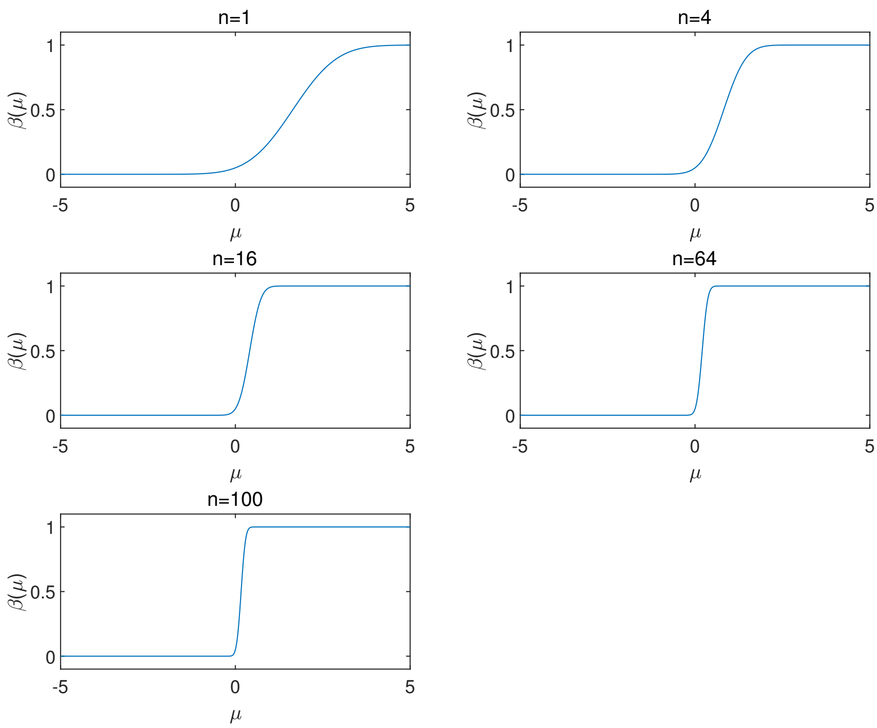

图 1.1: 单侧检验在不同样本量下的功效函数

(b) **双侧检验的功效函数为：**

$$
\begin{aligned}
\beta(\mu) &= \mathbb{P}\left(\left|\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}}\right| > u_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}} > u_{1-\frac{\alpha}{2}}\right) + \mathbb{P}\left(\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}} < u_{\frac{\alpha}{2}}\right) \\
&= \mathbb{P}\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} > u_{1-\frac{\alpha}{2}} - \frac{\mu-\mu_0}{\sqrt{\sigma^2/n}}\right) + \mathbb{P}\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} < u_{\frac{\alpha}{2}} - \frac{\mu-\mu_0}{\sqrt{\sigma^2/n}}\right) \\
&= \mathbb{P}\left(Z > u_{1-\frac{\alpha}{2}} - \frac{\mu-\mu_0}{\sqrt{\sigma^2/n}}\right) + \mathbb{P}\left(Z < u_{\frac{\alpha}{2}} - \frac{\mu-\mu_0}{\sqrt{\sigma^2/n}}\right)
\end{aligned} \tag{1.2}
$$

其中 $Z \sim N(0,1)$ 为标准正态分布，$\mu_0 = 0$。通过编写 **Matlab** 代码，我们得到了**双侧检验下的 LRT 功效函数如图1.2所示**。

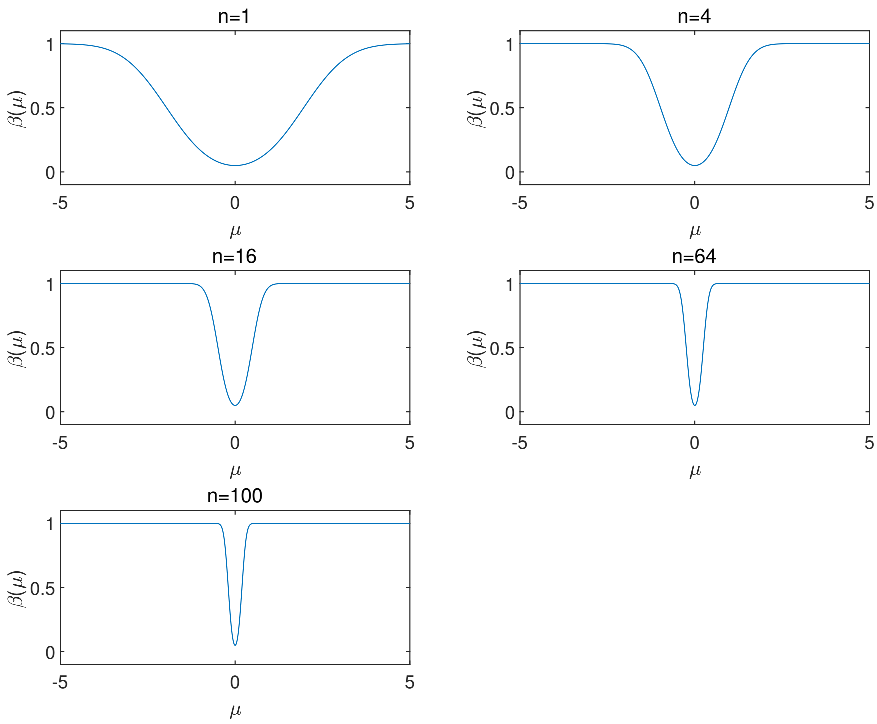

图 1.2: 双侧检验在不同样本量下的功效函数

用到的 **Matlab** 代码如下：

绘制单侧检验功效函数：

```
n = 1;
x = linspace(-5,5,1000);
y =  1-normcdf((norminv(0.95,0,1)-x*sqrt(n)/1),0,1);
subplot(3,2,1)
```

```
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=1')
n = 4;
y =  1-normcdf((norminv(0.95,0,1)-x*sqrt(n)/1),0,1);
subplot(3,2,2)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=4')
n = 16;
y =  1-normcdf((norminv(0.95,0,1)-x*sqrt(n)/1),0,1);
subplot(3,2,3)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=16')
n = 64;
y =  1-normcdf((norminv(0.95,0,1)-x*sqrt(n)/1),0,1);
subplot(3,2,4)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=64')
n = 100;
```

```
y =  1-normcdf((norminv(0.95,0,1)-x*sqrt(n)/1),0,1);
subplot(3,2,5)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=100')
```

绘制双侧检验功效函数:

```
n = 1;
x = linspace(-5,5,1000);
y = 1- (normcdf((norminv(0.975,0,1)-x*sqrt(n)/1),0,1)-...
normcdf((-norminv(0.975,0,1)-x*sqrt(n)/1),0,1));
subplot(3,2,1)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=1')
n = 4;
y = 1- (normcdf((norminv(0.975,0,1)-x*sqrt(n)/1),0,1)-...
normcdf((-norminv(0.975,0,1)-x*sqrt(n)/1),0,1));
subplot(3,2,2)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=4')
n = 16;
y = 1- (normcdf((norminv(0.975,0,1)-x*sqrt(n)/1),0,1)-...
normcdf((-norminv(0.975,0,1)-x*sqrt(n)/1),0,1));
```

```
subplot(3,2,3)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=16')
n = 64;
y = 1- (normcdf((norminv(0.975,0,1)-x*sqrt(n)/1),0,1)-...
normcdf((-norminv(0.975,0,1)-x*sqrt(n)/1),0,1));
subplot(3,2,4)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=64')
n = 100;
y = 1- (normcdf((norminv(0.975,0,1)-x*sqrt(n)/1),0,1)-...
normcdf((-norminv(0.975,0,1)-x*sqrt(n)/1),0,1));
subplot(3,2,5)
plot(x,y)
xlabel('\mu')
ylabel('\beta(\mu)')
ylim([-0.1,1.1])
title('n=100')
```

题目 2. 设一个随机变量 X 在 H_0 和 H_1 之下的概率质量函数由表1.1给出：

表 1.1: 概率质量函数表

x	1	2	3	4	5	6	7
$f(x\|H_0)$	0.01	0.01	0.01	0.01	0.01	0.01	0.94
$f(x\|H_1)$	0.06	0.05	0.04	0.03	0.02	0.01	0.79

利用 Neyman-Pearson 引理求 H_0 对 H_1 的的最大功效检验，取 $\alpha = 0.04$。并计算这个检验犯第二类错误的概率。

解答. 首先我们计算概率质量函数之比，如表1.2所示：

表 1.2: 概率密度之比

x	1	2	3	4	5	6	7
$\frac{f(x\|H_1)}{f(x\|H_0)}$	6	5	4	3	2	1	0.84

从表1.2中我们知道，当 x 较小时，我们倾向于拒绝 H_0；当 x 较大时，我们倾向于接受 H_0。因为 $\alpha = 0.04$，且当 H_0 成立时，$\sum_{i=1}^{4} \mathbb{P}(x = i|H_0) = 0.04 < \sum_{i=1}^{5} \mathbb{P}(x = i|H_0)$。于是，我们**取 $4 < k < 5$，此时的最大功效检验为：当 $x < k$ 时，拒绝 H_0；当 $x > k$ 时，接受 H_0**。

犯第二类错误概率为 H_1 成立下却接受了 H_0，那么**第二类错误概率 β 为：**

$$\beta = \sum_{i=5}^{7} \mathbb{P}(x = i|H_1) = 0.02 + 0.01 + 0.79 = 0.82 \tag{1.3}$$

题目 3. 证明 Poisson 族 Poisson(λ) 具有 MLR。

解答. 我们知道，Poisson(λ) 的概率质量函数为：

$$p_k = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1.4}$$

对于 $\forall\lambda_1 > \lambda_2$，我们有：

$$\frac{g(k|\lambda_1)}{g(k|\lambda_2)} = \frac{\frac{\lambda_1^k}{k!}e^{-\lambda_1}}{\frac{\lambda_2^k}{k!}e^{-\lambda_2}} = \frac{\lambda_1^k}{\lambda_2^k}e^{\lambda_2-\lambda_1} = \left(\frac{\lambda_1}{\lambda_2}\right)^k e^{\lambda_2-\lambda_1} \tag{1.5}$$

由于 $\frac{\lambda_1}{\lambda_2} > 1$，于是**式1.5是关于 k 单调递增的，因此 Poisson 分布族具有 MLR**。

题目 4. 设 $X_1, ..., X_n$ 是 $i.i.d$ 于 Poisson 分布 Poisson(λ) 的。

(a) 求：关于 $H_0 : \lambda \leq \lambda_0$ 对 $H_1 : \lambda > \lambda_0$ 的一个 UMP 检验。

(b) 考虑特殊情况 $H_0 : \lambda \leq 1$ 对 $H_1 : \lambda > 1$。利用中心极限定理确定确定样本量 n 以使得一个 UMP 检验满足 $\mathbb{P}(\text{拒绝}H_0|\lambda = 1) = 0.05$ 和 $\mathbb{P}(\text{拒绝}H_0|\lambda = 2) = 0.9$。

解答. (a) 首先证明：$T = \sum_{i=1}^{n} X_i$ 是 λ 的充分统计量。由 Poisson(λ) 的质量函数如式1.4所示，我们首先写出联合密度函数：

$$p(\mathbf{X}) = \frac{\lambda^{\sum_{i=1}^{n} X_i}}{X_1!X_2!...X_n!}e^{-n\lambda} \tag{1.6}$$

我们取：$g(t;\lambda) = \lambda^t e^{-n\lambda}, h(\mathbf{X}) = (X_1!X_2!...X_n!)^{-1}$ 则 $p(\mathbf{X}) = g(t;\lambda)h(\mathbf{X})$，由因子分解定理，我们知道 $T = \sum_{i=1}^{n} X_i$ 是 λ 的充分统计量。

我们知道，Poisson 分布具有可加性，于是 $T \sim$ Poisson$(n\lambda)$。而在**题目 3** 中，我们已经证明了 Poisson 分布族具有 MLR。于是，**通过 Karlin-Rubin 定理，取**：$\alpha = \mathbb{P}(T > k|\lambda = \lambda_0)$。**当 $T > k$ 时拒绝 H_0；当 $T < k$ 时接受 H_0 是一个水平为 α 的 UMP 检验。**

(b) 我们知道 Poisson 分布 Poisson(λ) 的均值和方差都是 λ。由中心极限定理，当 n 较大时，$T \sim AN(n\lambda, n\lambda)$。于是：

$$\mathbb{P}(T > k) = \alpha \Leftrightarrow \mathbb{P}(Z > \frac{k - n\lambda}{\sqrt{n\lambda}}) = \alpha \tag{1.7}$$

其中 Z 为标准正态分布。而为了满足题目中的条件，我们有：

$$\begin{cases} \mathbb{P}(Z > \frac{k-n}{\sqrt{n}}) = 0.05 \\ \mathbb{P}(Z > \frac{k-2n}{\sqrt{2n}}) = 0.9 \end{cases} \Rightarrow \begin{cases} \frac{k-n}{\sqrt{n}} = 1.6449 \\ \frac{k-2n}{\sqrt{2n}} = -1.2816 \end{cases} \tag{1.8}$$

求解式1.8中的右式，我们可以得到：$n = 11.9533 \approx 12, k = 17.6981$。

从而**我们确定的样本量** $n = 12$，**我们的 UMP 检验为：当** $T = \sum_{i=1}^{12} X_i > 17.70$ **时，我们拒绝** H_0**，当** $T = \sum_{i=1}^{12} X_i \leq 17.70$ **时，我们接受** H_0。

题目 5. 设 $X_1, ..., X_n$ 是来自正态总体 $N(\theta, \sigma^2)$ 的一组随机样本。考虑检验 $H_0 : \theta \leq \theta_0$ 对 $H_1 : \theta_0 > \theta_0$。

(a) 若 σ^2 已知，证明：当 $\bar{X} > \theta_0 + z_\alpha \sqrt{\frac{\sigma^2}{n}}$ 就拒绝 H_0 的检验是一个真实水平为 α 的检验。证明这个检验可以从一个 LRT 导出。

(b) 证明 (a) 中的检验是一个 UMP 检验。

(c) 若 σ^2 未知，证明：当 $\bar{X} > \theta_0 + t_{n-1,\alpha} \sqrt{\frac{S^2}{n}}$ 就拒绝 H_0 的检验是一个真实水平为 α 的检验。证明这个检验可以从一个 LRT 导出。

解答. (a) 我们知道，对于正态分布 $N(\theta, \sigma^2)$，$\bar{X}$ 是全空间 Θ 中均值的 MLE。因此，我们有：

$$\begin{aligned}\sup_{\theta\in\Theta} L(\theta|\mathbf{X}) &= \sup_{\theta\in\Theta} \left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp(-\frac{\sum_{i=1}^{n}(X_i-\theta)^2}{2\sigma^2}) \\ &= \left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp(-\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{2\sigma^2})\end{aligned} \tag{1.9}$$

而对于我们的原假设空间 Θ_0，当 $\bar{X} < \theta_0$ 时：

$$\begin{aligned}\sup_{\theta\in\Theta_0} L(\theta|\mathbf{X}) &= \sup_{\theta\in\Theta_0} \left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp(-\frac{\sum_{i=1}^{n}(X_i-\theta)^2}{2\sigma^2}) \\ &= \left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp(-\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{2\sigma^2})\end{aligned} \tag{1.10}$$

当 $\bar{X} \geq \theta_0$ 时：

$$\begin{aligned}\sup_{\theta\in\Theta_0} L(\theta|\mathbf{X}) &= \sup_{\theta\in\Theta_0} \left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp(-\frac{\sum_{i=1}^{n}(X_i-\theta)^2}{2\sigma^2}) \\ &= \left(\sqrt{2\pi\sigma^2}\right)^{-n} \exp(-\frac{\sum_{i=1}^{n}(X_i-\theta_0)^2}{2\sigma^2})\end{aligned} \tag{1.11}$$

因此，结合式1.9, 式1.10, 式1.11, 应用 LRT：

$$\lambda(\mathbf{X})=\frac{\sup\limits_{\theta\in\Theta_0}L(\theta|\mathbf{X})}{\sup\limits_{\theta\in\Theta}L(\theta|\mathbf{X})}=\begin{cases}\exp(\frac{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2-\sum\limits_{i=1}^{n}\left(X_i-\theta_0\right)^2}{2\sigma^2}) & \bar{X}\geq\theta_0\\ 1 & \bar{X}<\theta_0\end{cases}\tag{1.12}$$

我们计算式1.12在 $\bar{X}\geq\theta_0$ 情况下的指数部分的分子，有：

$$\begin{aligned}&\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2-\sum_{i=1}^{n}\left(X_i-\theta_0\right)^2\\&=\sum_{i=1}^{n}X_i^2-2\sum_{i=1}^{n}X_i\bar{X}+\sum_{i=1}^{n}\bar{X}^2-\sum_{i=1}^{n}X_i^2+2\sum_{i=1}^{n}X_i\theta_0-\sum_{i=1}^{n}\theta_0^2\\&=-2n\bar{X}+n\bar{X}^2+2n\bar{X}\theta_0-n\theta_0^2=-n\theta_0^2+2n\bar{X}\theta_0-n\bar{X}^2\\&=-n(\bar{X}-\theta_0)^2\end{aligned}\tag{1.13}$$

因此，最后**我们得到的 LRT 统计量为**：

$$\lambda(\mathbf{X})=\begin{cases}\exp(\frac{-n\left(\bar{X}-\theta_0\right)^2}{2\sigma^2}) & \bar{X}\geq\theta_0\\ 1 & \bar{X}<\theta_0\end{cases}\tag{1.14}$$

而题目中的检验为：

$$\mathbb{P}(\bar{X}>\theta_0+z_\alpha\sqrt{\frac{\sigma^2}{n}}|\theta_0)=\mathbb{P}(\frac{\bar{X}-\theta_0}{\sqrt{\sigma^2/n}}>z_\alpha|\theta_0)=\mathbb{P}(Z>z_\alpha)=\alpha\tag{1.15}$$

其中，Z 服从一个标准正态分布。对于式1.14, 我们的拒绝域为 $\lambda<c$, 也就是：

$$\exp(\frac{-n\left(\bar{X}-\theta_0\right)^2}{2\sigma^2})<c\Rightarrow -n(\bar{X}-\theta_0)^2<2\sigma^2\log c\Rightarrow\frac{\bar{X}-\theta_0}{\sqrt{\sigma^2/n}}>\sqrt{-2\log c}\tag{1.16}$$

因此，我们取 α 使 $z_\alpha=\sqrt{-2\log c}$, 就由 LRT 中导出了题中的检验。

(b) 我们知道 $T(\mathbf{X})=\bar{X}$ 为正态分布在方差已知情况下均值 θ 的充分统计量。下面证明 T 的分布族具有 MLR。对于 $\forall\theta_1>\theta_2$:

$$\begin{aligned}\frac{g(t|\theta_1)}{g(t|\theta_2)}&=\frac{\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\frac{\left(t-\theta_1\right)^2}{2\sigma^2/n}\right)}{\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left(-\frac{\left(t-\theta_2\right)^2}{2\sigma^2/n}\right)}\\&=\exp\left(-\frac{\left(t-\theta_1\right)^2-\left(t-\theta_2\right)^2}{2\sigma^2/n}\right)\\&=\exp\left(\frac{2t(\theta_1-\theta_2)+\theta_2^2-\theta_1^2}{2\sigma^2/n}\right)\end{aligned}\tag{1.17}$$

于是，我们知道式1.17是关于 t 单调递增的，从而正态分布的均值的充分统计量 T 具有 MLR。**结合 Karlin-Rubin 定理，(a) 是一个 UMP 检验。**

(c) 题目中的检验为：

$$\mathbb{P}(\bar{X} > \theta_0 + t_{n-1,\alpha}\sqrt{\frac{S^2}{n}}|\theta_0) = \mathbb{P}(\frac{\bar{X} - \theta_0}{\sqrt{S^2/n}} > t_{n-1,\alpha}|\theta_0) = \mathbb{P}(T > t_{n-1,\alpha}) = \alpha \tag{1.18}$$

其中，T 服从参数为 $n-1$ 的 t 分布。而另一方面，我们知道，正态分布 $N(\theta, \sigma^2)$ 的 MLE 为 $(\hat{\theta}, \hat{\sigma}^2) = (\bar{X}, \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n})$, 因此：

$$\begin{aligned}&\sup_{\theta,\sigma^2\in\Theta} L(\theta,\sigma^2|\mathbf{X}) = \sup_{\theta,\sigma^2\in\Theta}\left(\sqrt{2\pi\sigma^2}\right)^{-n}\exp(-\frac{\sum_{i=1}^{n}\left(X_i-\theta\right)^2}{2\sigma^2})\\&=\left(\sqrt{2\pi\hat{\sigma}^2}\right)^{-n}\exp(-\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{2\hat{\sigma}^2})\end{aligned} \tag{1.19}$$

对于我们的原假设空间 Θ_0，当 $\bar{X} < \theta_0$ 时：

$$\begin{aligned}&\sup_{\theta,\sigma^2\in\Theta_0} L(\theta,\sigma^2|\mathbf{X}) = \sup_{\theta,\sigma^2\in\Theta_0}\left(\sqrt{2\pi\sigma^2}\right)^{-n}\exp(-\frac{\sum_{i=1}^{n}\left(X_i-\theta\right)^2}{2\sigma^2})\\&=\left(\sqrt{2\pi\hat{\sigma}^2}\right)^{-n}\exp(-\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{2\hat{\sigma}^2}) = \left(\sqrt{2\pi\hat{\sigma}^2}\right)^{-n}\exp(-\frac{n}{2})\end{aligned} \tag{1.20}$$

当 $\bar{X} \geq \theta_0$ 时：

$$\begin{aligned}&\sup_{\theta,\sigma^2\in\Theta_0} L(\theta,\sigma^2|\mathbf{X}) = \sup_{\theta,\sigma^2\in\Theta_0}\left(\sqrt{2\pi\sigma^2}\right)^{-n}\exp(-\frac{\sum_{i=1}^{n}\left(X_i-\theta\right)^2}{2\sigma^2})\\&=\left(\sqrt{2\pi\hat{\sigma}_0^2}\right)^{-n}\exp(-\frac{\sum_{i=1}^{n}\left(X_i-\theta_0\right)^2}{2\hat{\sigma}_0^2}) = \left(\sqrt{2\pi\hat{\sigma}_0^2}\right)^{-n}\exp(-\frac{n}{2})\end{aligned} \tag{1.21}$$

其中，$\hat{\sigma}_0^2 = \frac{\sum_{i=1}^{n}\left(X_i-\theta_0\right)^2}{n}$。于是，结合式1.19, 式1.20, 式1.21。我们得到的 **LRT 统计量为：**

$$\lambda(\mathbf{X}) = \begin{cases}\left(\frac{\hat{\sigma}}{\hat{\sigma}_0}\right)^n & \bar{X} \geq \theta_0\\ 1 & \bar{X} < \theta_0\end{cases} \tag{1.22}$$

由：$\hat{\sigma}^2=\dfrac{n-1}{n}S^2$, 以及：

$$\begin{aligned}
n\hat{\sigma}_0^2&=\sum_{i=1}^{n}\left(X_i-\theta_0\right)^2=\sum_{i=1}^{n}\left(X_i-\bar{X}+\bar{X}-\theta_0\right)^2\\
&=\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum_{i=1}^{n}\left(\bar{X}-\theta_0\right)^2+2\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(\bar{X}-\theta_0\right)\\
&=(n-1)S^2+n\left(\bar{X}-\theta_0\right)^2+2\left(\bar{X}-\theta_0\right)\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\\
&=(n-1)S^2+n\left(\bar{X}-\theta_0\right)^2
\end{aligned}\tag{1.23}$$

因此，我们可以得到，当 $\bar{X}\geq\theta_0$ 时：

$$\lambda(\mathbf{X})=\left(\frac{\hat{\sigma}}{\hat{\sigma}_0}\right)^n=\left(\frac{\frac{n-1}{n}S^2}{\frac{n-1}{n}S^2+\left(\bar{X}-\theta_0\right)^2}\right)^{\frac{n}{2}}=\left(\frac{1}{1+\frac{n}{n-1}\frac{\left(\bar{X}-\theta_0\right)^2}{S^2}}\right)^{\frac{n}{2}}\tag{1.24}$$

于是：

$$\begin{aligned}
&\lambda(\mathbf{X})<c\Rightarrow\left(\frac{1}{1+\frac{n}{n-1}\frac{\left(\bar{X}-\theta_0\right)^2}{S^2}}\right)^{\frac{n}{2}}<c\Rightarrow\frac{n}{n-1}\frac{\left(\bar{X}-\theta_0\right)^2}{S^2}>\left(\frac{1}{c}\right)^{\frac{2}{n}}-1\\
&\Rightarrow\frac{\left(\bar{X}-\theta_0\right)^2}{S^2}>\frac{(n-1)(c^{-\frac{2}{n}}-1)}{n}\approx\frac{(n-1)}{n}\frac{2}{n}\log\frac{1}{c}\approx\frac{2}{n}\log\frac{1}{c}\\
&\Rightarrow\frac{\bar{X}-\theta_0}{\sqrt{S^2/n}}>\sqrt{-2\log c}
\end{aligned}\tag{1.25}$$

因此在 n 足够大的情况下，我们取 α 使 $z_\alpha=\sqrt{-2\log c}$, 就由 LRT 中导出了题中的检验。

2 渐进评价

题目 6. 为检验 $H_0: p = p_0$ 对 $H_1: p \neq p_0$，假定我们观测到 $X_1, ..., X_n$ 是 $i.i.d$ 于 Bernoulli(p) 的。

(a) 推导出 $-2\log\lambda(\boldsymbol{x})$ 的表达式，这里 $\lambda(\boldsymbol{x})$ 为 LRT 统计量。

(b) 模拟 $-2\log\lambda(\boldsymbol{x})$ 的分布，并把结果与 $\mathcal{X}^2$ 近似进行比较。

(c) 通过模拟得出该检验的功效函数，探讨检验的功效以及影响功效的因素。

解答. (a) 由于：

$$\lambda(\mathbf{X}) = \frac{\sup\limits_{p\in\Theta_0} L(p|\mathbf{X})}{\sup\limits_{p\in\Theta} L(p|\mathbf{X})} \tag{2.1}$$

式2.1的分子部分为：

$$\sup_{p\in\Theta_0} L(p|\mathbf{X}) = p_0^{\sum\limits_{i=1}^{n} X_i}(1-p_0)^{n-\sum\limits_{i=1}^{n} X_i} \tag{2.2}$$

我们知道，Bernoulli(p) 中参数 p 的 MLE 为 $\hat{p} = \bar{X}$，于是式2.1的分母部分为：

$$\sup_{p\in\Theta_0} L(p|\mathbf{X}) = \bar{X}^{\sum\limits_{i=1}^{n} X_i}(1-\bar{X})^{n-\sum\limits_{i=1}^{n} X_i} \tag{2.3}$$

于是，$-2\log\lambda(\mathbf{X})$ **的表达式如式2.4所示：**

$$\begin{aligned}
-2\log\lambda(\mathbf{X}) &= -2\log\frac{p_0^{\sum\limits_{i=1}^{n} X_i}(1-p_0)^{n-\sum\limits_{i=1}^{n} X_i}}{\bar{X}^{\sum\limits_{i=1}^{n} X_i}(1-\bar{X})^{n-\sum\limits_{i=1}^{n} X_i}} \\
&= -2n\bar{X}\log p_0 - 2(n-n\bar{X})\log(1-p_0) + 2n\bar{X}\log\bar{X} + 2(n-n\bar{X})\log(1-\bar{X}) \\
&= 2n\bar{X}\log\left(\frac{\bar{X}}{p_0}\right) + 2(n-n\bar{X})\log(\frac{1-\bar{X}}{1-p_0})
\end{aligned} \tag{2.4}$$

(b) 我们编写模拟 $-2\log\lambda(\mathbf{X})$ 的 **Matlab** 代码，其中我们取 $p_0 = 0.5$，样本数 $n = 50$，共模拟了 10000 次 $-2\log\lambda(\mathbf{X})$，并将模拟的直方密度图和 $\mathcal{X}_1^2$ 分布进行比较，其结果如图所示。

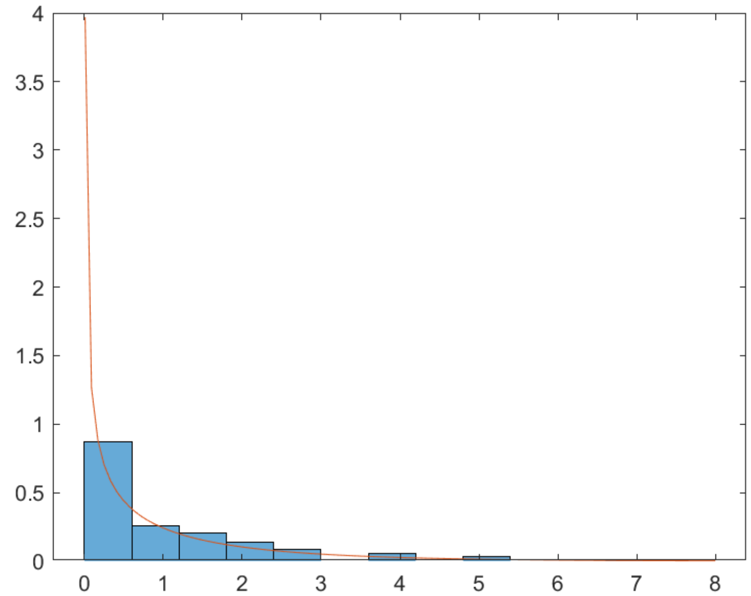

图 2.1: $-2\log\lambda(\mathbf{X})$ 分布和 $\mathcal{X}^2$ 近似的比较

从图2.1中我们可以看到，$-2\log\lambda(\mathbf{X})$ 的分布确实 $\mathcal{X}_1^2$ 近似相同，这也验证了本题注记中的定理。

(c) 我们和 (b) 中一样，我们研究抽取的样本数 n 和假设 p_0 与 p 的真实值的距离对功效函数的影响。在我们的模拟中，我们同样对每一个真实值 p 模拟 10000 次，得到该值下的拒绝原假设 H_0 的概率，得到的功效函数图像如图2.2和图2.3：

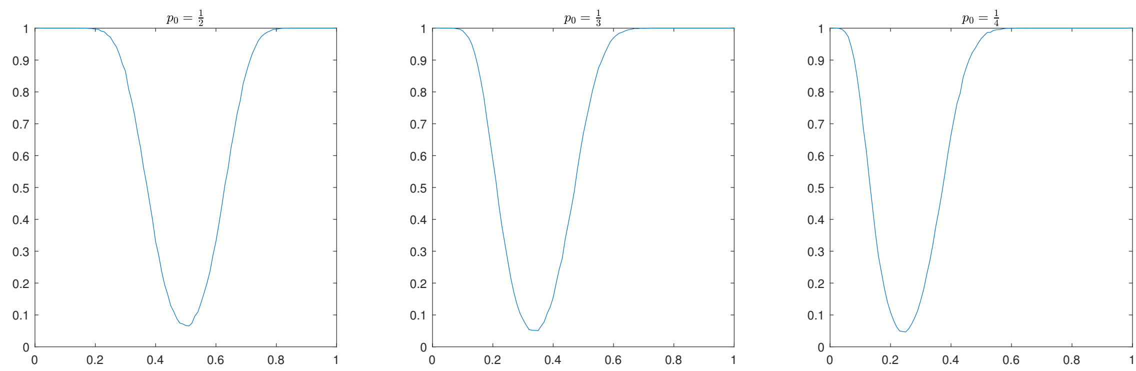

图 2.2: p_0 与 p 的真实值的距离不同时的功效函数

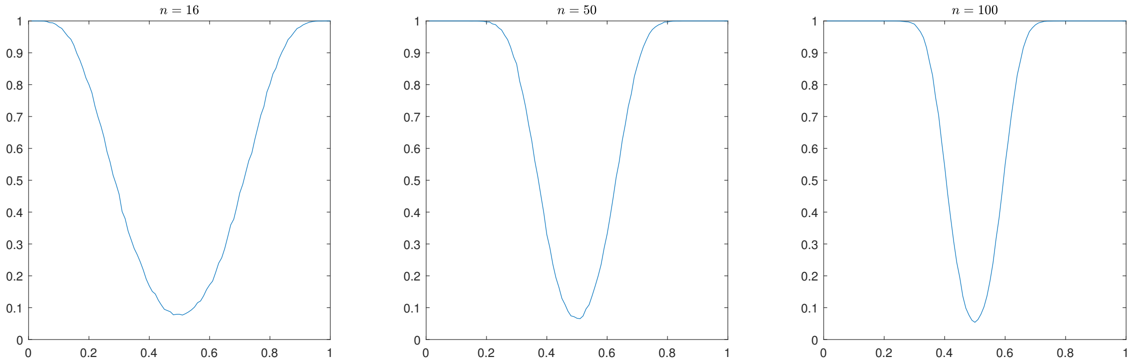

图 2.3: 样本量不同时的功效函数

从图2.2和图2.3中可以看出，离真实值越近时，功效函数越小。并且所取的样本量越大，功效函数在真实值附近越为陡峭，检验的准确率越高。

题目 6 的注记. 关于检验 $H_0 : \theta = \theta_0$ 对 $H_1 : \theta \neq \theta_0$，设 $X_1, ..., X_n$ 是 $i.i.d$ 于 $f(x|\theta)$ 的，$\hat{\theta}$ 是 θ 的 MLE，并且 $f(x|\theta)$ 满足 6 条正则性条件。则在 H_0 之下，当 $n \to \infty$, 有 $-2\log\lambda(\mathbf{X})$ 依分布收敛于 $\mathcal{X}_1^2$。

正则性条件：

(A1) 我们观察到 $X_1, ..., X_n$，$X_i \sim f(x|\theta)$ 是 $i.i.d$ 的。

(A2) 参数是可识别的，即如果 $\theta \neq \theta'$，则 $f(x|\theta) \neq f(x|\theta')$

(A3) 各个密度 $f(x|\theta)$ 有共同的支撑集，并且 $f(x|\theta)$ 关于 θ 可导。

(A4) 参数空间 Ω 包含一个开集 ω，以及真参数值 θ_0 为该开集的一个内点。

(A5) 对于每个 $x \in \mathcal{X}$，密度 $f(x|\theta)$ 关于 θ 是三阶可导的，其三阶导数是 θ 的连续函数，并且 $\displaystyle\int f(x|\theta)\mathrm{d}x$ 可以在积分号下微分三次。

(A6) 对于任何 $\theta_0 \in \Omega$，存在一个正数 c 和一个函数 $M(x)$(两者都可以依赖于 θ_0) 使得：

$$\left|\frac{\partial^3}{\partial\theta^3}\log f(x \mid \theta)\right| \leqslant M(x), \text{对于所有} x \in \mathcal{X}, \theta_0 - c < \theta < \theta_0 + c$$

以及 $\mathbb{E}_{\theta_0}|M(X)| < \infty$。

模拟 $-2\log\lambda(\mathbf{X})$ 的分布并绘制其与 $\mathcal{X}_1^2$ 分布比较图代码：

```
n = 50;
p_real = 0.5;
N = 10000;
p_0 = 0.5;
for j = 1:N
```

```
    for i = 1:n
        x(i) = binornd(1,p_real);
    end
    T(j) = 2*n*mean(x)*log(mean(x)/p_0)+2*(n-n*mean(x))*...
    log((1-mean(x))/(1-p_0));
end

histogram(T,'Normalization','pdf','BinWidth',0.6,'BinLimits',[0,8])
hold on
plot(linspace(0.01,8,100),chi2pdf(linspace(0.01,8,100),1))
```

功效函数绘制：

```
n = 100;%n = 50,16
p_real= linspace(0.01,1,100);
N = 10000;
p_0 = 1/2;%p_real = 1/4,1/3
for k = 1:100
    count = 0;
    for j = 1:N
        for i = 1:n
            x(i) = binornd(1,p_real(k));
        end
        mean_x = mean(x);
        T = 2*n*mean_x*log(mean_x/p_0)+2*(n-n*mean_x)...
        *log((1-mean_x)/(1-p_0));
        if T < chi2inv(0.95,1)
            count = count+1;
        end
    end
    rej(k) = 1-count/N;
end
```

题目 7. 设 $X_1, ..., X_n$ 是来自总体 $N(\mu, \sigma^2)$ 的随机样本。

(a) 如果 μ 未知而 σ 已知，证明 $Z = \sqrt{n}(\bar{X} - \mu_0)/\sigma$ 是检验 $H_0 : \mu = \mu_0$ 的 Wald 统计量。

(b) 如果 σ 未知而 μ 已知，求检验 $H_0 : \sigma = \sigma_0$ 的一个 Wald 统计量。

解答. (a) 由题意知：$\bar{X} \sim N(\mu, \dfrac{\sigma^2}{n})$。**因为** $\bar{X}$ **是均值** μ **的估计，** $\dfrac{\sigma^2}{n}$ **是** $\bar{X}$ **的方差**，因此：

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{2.5}$$

为检验 $H_0 : \mu = \mu_0$ 的 Wald 统计量。

(b) 我们知道，对于正态分布的方差 σ^2 的 MLE 为：$\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$。于是 σ 的估计 W_n 为：

$$W_n = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}} \tag{2.6}$$

下面我们先计算 $\hat{\sigma}^2$ 的方差，即 $\dfrac{\sum_{i=1}^{n} (X_i - \mu)^2}{n}$ 的方差。因为我们的 $X_i \sim N(\mu, \sigma^2)$，于是我们有：

$$\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \mathcal{X}_1^2 \tag{2.7}$$

由 $\mathcal{X}^2$ 分布的可加性，我们有：

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \mathcal{X}_n^2 \tag{2.8}$$

于是：

$$\mathrm{Var}\left(\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2\right) = 2n \tag{2.9}$$

那么，由式2.9我们可以得到：

$$\mathrm{Var}\left(\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sqrt{n}}\right)^2\right) = \mathrm{Var}\left(\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2\right) \times \sigma^4/n^2 \bigg|_{\sigma^2 = \hat{\sigma}^2} = \frac{2n\sigma^4}{n^2}\bigg|_{\sigma^2 = \hat{\sigma}^2} = \frac{2\hat{\sigma}^4}{n} \tag{2.10}$$

那么结合式2.10可以计算出 $\hat{\sigma}$ 的方差为：

$$\mathrm{Var}(\hat{\sigma}) = \left(\frac{\mathrm{d}}{\mathrm{d}\sigma^2}\sigma\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sqrt{n}}\right)^2\right)\bigg|_{\sigma^2 = \hat{\sigma}^2} = \frac{1}{4}\hat{\sigma}^{-2} \times \frac{2\hat{\sigma}^4}{n} = \frac{\hat{\sigma}^2}{2n} \tag{2.11}$$

从而，结合式2.6和式2.11，$H_0:\sigma=\sigma_0$ **的一个 Wald 统计量为：**

$$Z=\frac{\sqrt{\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{n}}-\sigma_0}{\sqrt{\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{2n^2}}} \tag{2.12}$$

题目 8. 设 $X_1,...,X_n$ 是来自总体 $N(\mu,\sigma^2)$ 的随机样本。

(a) 如果 μ 未知而 σ 已知，证明 $Z=\sqrt{n}(\bar{X}-\mu_0)/\sigma$ 是检验 $H_0:\mu=\mu_0$ 的记分统计量。

(b) 如果 σ 未知而 μ 已知，求检验 $H_0:\sigma=\sigma_0$ 的一个记分统计量。

解答. (a) 我们先写出关于 μ 的似然函数：

$$L(\mu|\mathbf{X})=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\right) \tag{2.13}$$

于是：

$$\frac{\partial}{\partial u}\log L(\mu|\mathbf{X})=\frac{\sum\limits_{i=1}^{n}(X_i-\mu)}{\sigma^2},\qquad \frac{\partial^2}{\partial u^2}\log L(\mu|\mathbf{X})=\frac{-n}{\sigma^2} \tag{2.14}$$

由式2.14，$H_0:\mu=\mu_0$ **的记分统计量为：**

$$Z=\frac{\frac{\sum\limits_{i=1}^{n}(X_i-\mu_0)}{\sigma^2}}{\sqrt{\frac{n}{\sigma^2}}}=\frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} \tag{2.15}$$

(b) 对于 $H_0:\sigma=\sigma_0$，写出关于 σ 的似然函数：

$$L(\sigma|\mathbf{X})=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\right) \tag{2.16}$$

于是：

$$\frac{\partial}{\partial \sigma}\log L(\sigma|\mathbf{X})=\frac{-n}{\sigma}+\frac{\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^3},\qquad \frac{\partial^2}{\partial \sigma^2}\log L(\mu|\mathbf{X})=\frac{n}{\sigma^2}-\frac{3\sum\limits_{i=1}^{n}(X_i-\mu)^2}{\sigma^4} \tag{2.17}$$

由式2.17，我们有：

$$\mathbb{E}\left(-\frac{\partial^2}{\partial\sigma^2}\log L(\mu|\mathbf{X})\right)=-\frac{n\sigma^2-3n\sigma^2}{\sigma^4}=\frac{2n}{\sigma^2} \tag{2.18}$$

从而 $H_0:\sigma=\sigma_0$ 的一个记分统计量为：

$$Z=\frac{-\dfrac{n}{\sigma_0}+\dfrac{n\hat{\sigma}^2}{\sigma_0^3}}{\sqrt{\dfrac{2n}{\sigma_0^2}}}=\frac{\sqrt{n}}{\sqrt{2}}\frac{\hat{\sigma}^2-\sigma_0^2}{\sigma_0^2} \tag{2.19}$$

其中：$\hat{\sigma}^2=\dfrac{\sum\limits_{i=1}^{n}\left(X_i-\mu\right)^2}{n}$。

3 附加题

题目 9. 设 $X_1, ..., X_n$ 是来自 $N(\mu_X, \sigma_X^2)$ 的一组随机样本，设 $Y_1, ..., Y_m$ 是与之独立的来自 $N(\mu_Y, \sigma_Y^2)$ 的一组随机样本。我们的兴趣在于检验：

$$H_0 : \mu_X = \mu_Y \quad \text{对} \quad H_1 : \mu_X \neq \mu_Y \tag{3.1}$$

这里假定 $\sigma_X^2 = \sigma_Y^2 = \sigma^2$。

(a) 推导出关于这个假设的 LRT。证明这个 LRT 能够基于统计量：

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \tag{3.2}$$

其中：

$$S_p^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{m}\left(Y_i - \bar{Y}\right)^2\right) \tag{3.3}$$

量 S_p^2 有时也被称为一个合并方差估计值。

(b) 证明：在 H_0 下，$T \sim t_{n+m-2}$。(这个检验被称为两样本 t 检验)

(c) 木材样本取自某个拜占庭教堂的中心与外围。木材的年代经过鉴定，给出如下数据：

中心		外围	
1294	1251	1284	1274
1279	1248	1272	1264
1274	1240	1256	1256
1264	1232	1254	1250
1263	1220	1242	
1254	1218		
1251	1210		

(3.4)

利用两样本 t 检验检验中心的木材平均年龄与外围的木材平均年龄是否相同。

解答. (a) 似然比统计量为：

$$\lambda(\mathbf{X},\mathbf{Y})=\frac{\sup\limits_{(\mu_X,\mu_Y)\in\Theta_0}L(\mu_X,\mu_Y,\sigma^2|\mathbf{X},\mathbf{Y})}{\sup\limits_{(\mu_X,\mu_Y)\in\Theta}L(\mu_X,\mu_Y,\sigma^2|\mathbf{X},\mathbf{Y})}=\frac{L(\tilde{\mu}_X,\tilde{\mu}_Y,\tilde{\sigma}^2|\mathbf{X},\mathbf{Y})}{L(\hat{\mu}_X,\hat{\mu}_Y,\hat{\sigma}^2|\mathbf{X},\mathbf{Y})} \tag{3.5}$$

首先计算式3.5的分子部分。此时：$\tilde{\mu}_X=\tilde{\mu}_Y=\tilde{\mu}$。且 $\mathbf{X},\mathbf{Y}$ 来自同一分布 $N(\tilde{\mu},\sigma^2)$ 且独立，此时：

$$p(\mathbf{X},\mathbf{Y})=\left(2\pi\sigma^2\right)^{-\frac{m}{2}-\frac{n}{2}}\exp\left(-\frac{\sum\limits_{i=1}^{n}\left(X_i-\mu\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\mu\right)^2}{2\sigma^2}\right) \tag{3.6}$$

则：

$$\begin{aligned}&\frac{\partial}{\partial u}\log p(\mathbf{X},\mathbf{Y})\propto 2\sum_{i=1}^{n}X_i-2n\mu+2\sum_{i=1}^{m}Y_i-2m\mu\\&\frac{\partial}{\partial\sigma^2}\log p(\mathbf{X},\mathbf{Y})\propto-\frac{\frac{n+m}{2}}{\sigma^2}+\frac{\sum\limits_{i=1}^{n}\left(X_i-\mu\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\mu\right)^2}{2\left(\sigma^2\right)^2}\end{aligned} \tag{3.7}$$

于是，我们令式3.7两式为 0，即可得到关于正态分布两个参数的 MLE：

$$\tilde{u}=\frac{n\bar{X}+m\bar{Y}}{n+m},\qquad \tilde{\sigma}^2=\frac{\sum\limits_{i=1}^{n}\left(X_i-\tilde{\mu}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\tilde{\mu}\right)^2}{n+m} \tag{3.8}$$

因此，式3.5分子部分为：

$$L(\tilde{\mu}_X,\tilde{\mu}_Y,\tilde{\sigma}^2|\mathbf{X},\mathbf{Y})=\left(2\pi\tilde{\sigma}^2\right)^{-\frac{m}{2}-\frac{n}{2}}\exp\left(-\frac{\sum\limits_{i=1}^{n}\left(X_i-\tilde{\mu}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\tilde{\mu}\right)^2}{2\tilde{\sigma}^2}\right) \tag{3.9}$$

下面计算式3.5的分母部分，由于 μ_X,μ_Y 的 MLE 为：$\hat{\mu}_X=\bar{X},\hat{\mu}_Y=\bar{Y}$。于是，

$$\frac{\partial}{\partial\sigma^2}\log p(\mathbf{X},\mathbf{Y})\propto-\frac{\frac{n+m}{2}}{\sigma^2}+\frac{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2}{2\left(\sigma^2\right)^2} \tag{3.10}$$

令3.10为 0，得到：

$$\hat{\sigma}^2=\frac{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2}{n+m} \tag{3.11}$$

从而式3.5可写为：

$$
\begin{aligned}
\lambda(\mathbf{X},\mathbf{Y}) &= \frac{(2\pi\tilde{\sigma}^2)^{-\frac{m}{2}-\frac{n}{2}} \exp\left(-\frac{\sum\limits_{i=1}^{n}(X_i-\tilde{\mu})^2+\sum\limits_{i=1}^{m}(Y_i-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right)}{(2\pi\hat{\sigma}^2)^{-\frac{m}{2}-\frac{n}{2}} \exp\left(-\frac{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2}{2\hat{\sigma}^2}\right)} \\
&= \frac{(2\pi\tilde{\sigma}^2)^{-\frac{m}{2}-\frac{n}{2}} \exp\left(-\frac{n+m}{2}\right)}{(2\pi\hat{\sigma}^2)^{-\frac{m}{2}-\frac{n}{2}} \exp\left(-\frac{n+m}{2}\right)} = \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)^{-\frac{n+m}{2}}
\end{aligned}
\tag{3.12}
$$

为了计算方便，我们先计算式3.8中的 $\sum\limits_{i=1}^{n}(X_i-\tilde{\mu})^2$:

$$
\begin{aligned}
&\sum_{i=1}^{n}(X_i-\tilde{\mu})^2 = \sum_{i=1}^{n}\left(X_i-\frac{n\bar{X}+m\bar{Y}}{n+m}\right)^2 = \sum_{i=1}^{n}\left(X_i-\bar{X}+\bar{X}-\frac{n\bar{X}+m\bar{Y}}{n+m}\right)^2 \\
&= \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 + \sum_{i=1}^{n}\left(\bar{X}-\frac{n\bar{X}+m\bar{Y}}{n+m}\right)^2 + 2\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(\bar{X}-\frac{n\bar{X}+m\bar{Y}}{n+m}\right) \\
&= \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 + \sum_{i=1}^{n}\left(\bar{X}-\frac{n\bar{X}+m\bar{Y}}{n+m}\right)^2 = \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 + n\left(\bar{X}-\frac{n\bar{X}+m\bar{Y}}{n+m}\right)^2 \\
&= \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 + n\left(\frac{n\bar{X}+m\bar{X}-n\bar{X}-m\bar{Y}}{n+m}\right)^2 = \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 + nm^2\left(\frac{\bar{X}-\bar{Y}}{n+m}\right)^2
\end{aligned}
\tag{3.13}
$$

那么相应的，可以计算得：

$$
\sum_{i=1}^{m}(Y_i-\tilde{\mu})^2 = \sum_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2 + n^2m\left(\frac{\bar{X}-\bar{Y}}{n+m}\right)^2 \tag{3.14}
$$

于是，结合式3.12，似然比检验为，当 $\lambda(\mathbf{X},\mathbf{Y})<c$ 时，拒绝 H_0:

$$\begin{aligned}
&\lambda(\mathbf{X},\mathbf{Y})<c \Leftrightarrow \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right) > c^{-\frac{2}{n+m}} \\
&\Leftrightarrow \frac{\sum\limits_{i=1}^{n}\left(X_i-\tilde{\mu}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\tilde{\mu}\right)^2}{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2} > c^{-\frac{2}{n+m}} \\
&\Leftrightarrow \frac{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2+nm^2\left(\frac{\bar{X}-\bar{Y}}{n+m}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2+n^2m\left(\frac{\bar{X}-\bar{Y}}{n+m}\right)^2}{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2} > c^{-\frac{2}{n+m}} \\
&\Leftrightarrow 1+\frac{\frac{nm}{n+m}\left(\bar{X}-\bar{Y}\right)^2}{\sum\limits_{i=1}^{n}\left(X_i-\bar{X}\right)^2+\sum\limits_{i=1}^{m}\left(Y_i-\bar{Y}\right)^2} > c^{-\frac{2}{n+m}} \\
&\Rightarrow 1+\frac{\frac{nm}{n+m}\left(\bar{X}-\bar{Y}\right)^2}{(n+m-2)S_p^2} > c^{-\frac{2}{n+m}} \\
&\Rightarrow \frac{\left(\bar{X}-\bar{Y}\right)^2}{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)} > (n+m-2)(c^{-\frac{2}{n+m}}-1)
\end{aligned} \tag{3.15}$$

其中，S_p^2 如式3.3所示，**由式3.15那么这个假设的 LRT 就可以写为：**

$$T=\frac{\bar{X}-\bar{Y}}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}} \tag{3.16}$$

当

$$|T|>\sqrt{(n+m-2)(c^{-\frac{2}{n+m}}-1)} \tag{3.17}$$

拒绝 H_0。

(b) 在 H_0 下，$\mu_X=\mu_Y$，则：

$$\begin{aligned}
\bar{X} &\sim N\left(\mu_X,\frac{\sigma^2}{n}\right) \\
\bar{Y} &\sim N\left(\mu_Y,\frac{\sigma^2}{m}\right)
\end{aligned} \tag{3.18}$$

那么，由式3.18及 $\mathbf{X}$ 与 $\mathbf{Y}$ 独立，我们有：

$$\bar{X}-\bar{Y}\sim N\left(0,\frac{\sigma^2}{n}+\frac{\sigma^2}{m}\right) \tag{3.19}$$

X 与 **Y** 独立，那么：

$$\begin{aligned}\frac{(n-1)S_X^2}{\sigma^2} &\sim \mathcal{X}_{n-1}^2 \\ \frac{(m-1)S_Y^2}{\sigma^2} &\sim \mathcal{X}_{m-1}^2\end{aligned} \tag{3.20}$$

其中式3.20两式独立，那么由 $\mathcal{X}^2$ 分布的可加性：

$$\frac{(n-1)S_X^2}{\sigma^2}+\frac{(m-1)S_Y^2}{\sigma^2}=\frac{(n+m-2)S_p^2}{\sigma^2}\sim \mathcal{X}_{n+m-2}^2 \tag{3.21}$$

我们知道，$\bar{X}$ 和 S_X^2 是独立的，$\bar{Y}$ 和 S_Y^2 也是独立的。那么，$\bar{X}-\bar{Y}$ 和式3.21也是独立的，**因此，结合式3.19和式3.21，我们得到：**

$$T=\frac{\dfrac{\bar{X}-\bar{Y}}{\sqrt{\dfrac{\sigma^2}{n}+\dfrac{\sigma^2}{m}}}}{\sqrt{\dfrac{\dfrac{(n+m-2)S_p^2}{\sigma^2}}{n+m-2}}}=\frac{\bar{X}-\bar{Y}}{\sqrt{S_p^2\left(\dfrac{1}{n}+\dfrac{1}{m}\right)}}\sim t_{n+m-2} \tag{3.22}$$

因此，**在** H_0 **下，**$T\sim t_{n+m-2}$。

(c) 我们编写 **Matlab** 代码来进行双边的 t 检验，可以得到我们的统计量 $T=-1.3093$，我们取置信度 $\alpha=0.05$，那么 $t_{0.975}(n+m-2)=2.0796$，**从而:**$|T|<t_{0.975}(n+m-2)$**，我们接受** H_0**，即认为中心的木材平均年龄与外围的木材平均年龄相同。**

用到的 **Matlab** 代码如下：

双边 t 检验：

```
x = [1294 1251 1279 1248 1274 1240 1264 1232 1263 1220 1254 1212 ...
        1251 1210];
y = [1284 1274 1272 1264 1256 1256 1254 1250 1242];
n = length(x);
m = length(y);
s_p_2 = (var(x)*(n-1)+var(y)*(m-1))/(n+m-2);
T = (mean(x)-mean(y))/sqrt(s_p_2*(1/n+1/m))
t = tinv(0.975,n+m-2)
```

题目 10. 从多个总体收集到的二项数据常常用列联表表现出来。在两个总体的情形，列联表如：

表 3.1: 两个总体的列联表

	总体		
	1	2	总和
成功	S_1	S_2	$S = S_1 + S_2$
失败	F_1	F_2	$F = F_1 + F_2$
总和	n_1	n_2	$n = n_1 + n_2$

其中总体 1 是 binomial(n_1, p_1)，有 S_1 个成功，F_1 个失败；总体 2 是 binomial(n_2, p_2)，有 S_2 个成功，F_2 个失败。经常感兴趣的一个假设是：

$$H_0 : p_1 = p_2 \quad \text{对} \quad H_1 : p_1 \neq p_2 \tag{3.23}$$

(a) 说明可以基于统计量：

$$T = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)(\hat{p}(1-\hat{p}))} \tag{3.24}$$

进行检验，其中 $\hat{p}_1 = S_1/n_1, \hat{p}_2 = S_2/n_2, \hat{p} = (S_1+S_2)/(n_1+n_2)$。另外，证明当 $n_1, n_2 \to \infty$ 时，T 的分布趋于 $\mathcal{X}_1^2$。(这时所谓独立性 $\mathcal{X}^2$ 检验的一个特殊情形)

(b) 测量与 H_0 相违背的另一个方法时计算期望频数表。这个表的构造方法是，在给定边缘总和的条件下，根据 $H_0 : p_1 = p_2$ 填充表格，即：

表 3.2: 两个总体的期望频数表

	期望频数		
	1	2	总和
成功	$\dfrac{n_1 S}{n_1 + n_2}$	$\dfrac{n_2 S}{n_1 + n_2}$	$S = S_1 + S_2$
失败	$\dfrac{n_1 F}{n_1 + n_2}$	$\dfrac{n_2 F}{n_1 + n_2}$	$F = F_1 + F_2$
总和	n_1	n_2	$n = n_1 + n_2$

用这个期望频数表中的所有格子，计算统计量 T^*:

$$T^* = \sum \frac{(\text{ 观测频数} - \text{期望频数})^2}{\text{期望频数}}$$
$$= \frac{\left(S_1 - \dfrac{n_1 S}{n_1+n_2}\right)^2}{\dfrac{n_1 S}{n_1+n_2}} + \cdots + \frac{\left(F_2 - \dfrac{n_2 F}{n_1+n_2}\right)^2}{\dfrac{n_2 F}{n_1+n_2}} \tag{3.25}$$

用代数运算证明 $T^* = T$ 因此 T^* 是渐进 $\mathcal{X}^2$ 的。

(c) 检验 p_1 和 p_2 相等的另一个统计量是：

$$T^{**} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \tag{3.26}$$

证明,在 H_0 下,T^{**} 是渐进 $N(0,1)$ 的,因此,其平方渐进于 $\mathcal{X}_1^2$。进一步,证明 $(T^{**})^2 \neq T^*$。

(d) 在什么情况下一个统计量比另一个更好？

(e) 19 世纪后期 Joesph Lister 进行了一个著名的医学试验。当时手术的死亡率很高，而 Lister 猜测使用抗感染剂石炭酸可能有助于降低死亡率。在几年期间，Lister 做了 75 例截肢手术，有的用了石炭酸，有的没有用。数据如下：

		用了石炭酸？	
		用	没用
患者存活？	存活	34	19
	没有	6	16

用这些数据检验石炭酸的使用是否与患者死亡有关。

解答. (a) 我们知道，由中心极限定理：

$$\begin{aligned} \hat{p}_1 &\sim AN\left(p_1, \frac{p_1(1-p_1)}{n_1}\right) \\ \hat{p}_2 &\sim AN\left(p_2, \frac{p_2(1-p_2)}{n_2}\right) \end{aligned} \tag{3.27}$$

其中 $\hat{p}_1 = S_1/n_1, \hat{p}_2 = S_2/n_2$。如果，$\hat{p}_1, \hat{p}_2$ 是独立的，我们有：

$$\hat{p}_1 - \hat{p}_2 \sim AN\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right) \tag{3.28}$$

即：

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)(p(1-p))}}\sim AN(0,1) \tag{3.29}$$

我们先计算一下 p 的 MLE：

$$\begin{aligned} L(\mathbf{X}_1,\mathbf{X}_2)&=p^{S_1}(1-p)^{n_1-S_1}p^{S_2}(1-p)^{n_2-S_2}=p^{S_1+S_2}(1-p)^{n_1+n_2-S_1-S_2}\\ \log L(\mathbf{X}_1,\mathbf{X}_2)&=(S_1+S_2)\log p+(n_1+n_2-S_1-S_2)\log(1-p)\end{aligned} \tag{3.30}$$

得到：

$$\frac{\mathrm{d}}{\mathrm{d}p}L(\mathbf{X}_1,\mathbf{X}_2)=0\Rightarrow\frac{S_1+S_2}{p}=\frac{n_1+n_2-S_1-S_2}{1-p} \tag{3.31}$$

从而 p 的 MLE 为：

$$\hat{p}=\frac{S_1+S_2}{n_1+n_2} \tag{3.32}$$

于是，当 $n_1,n_2\to\infty$ 时。结合式3.29及 Slutsky 定理：

$$T=\frac{(\hat{p}_1-\hat{p}_2)^2}{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)(\hat{p}(1-\hat{p}))}\sim\mathcal{X}_1^2 \tag{3.33}$$

即可以基于 T 进行检验，且 T 的分布趋于 $\mathcal{X}_1^2$。

(b) 我们计算 T^*:

$$\begin{aligned} T^*&=\frac{n_1^2\left(\frac{S_1}{n_1}-\frac{S}{n_1+n_2}\right)^2}{n_1\frac{S}{n_1+n_2}}+\frac{n_2^2\left(\frac{S_2}{n_2}-\frac{S}{n_1+n_2}\right)^2}{n_2\frac{S}{n_1+n_2}}+\frac{n_1^2\left(\frac{F_1}{n_1}-\frac{F}{n_1+n_2}\right)^2}{n_1\frac{F}{n_1+n_2}}+\frac{n_2^2\left(\frac{F_2}{n_2}-\frac{F}{n_1+n_2}\right)^2}{n_2\frac{F}{n_1+n_2}}\\ &=\frac{n_1^2(\hat{p}_1-\hat{p})^2}{n_1\hat{p}}+\frac{n_2^2(\hat{p}_2-\hat{p})^2}{n_2\hat{p}}+\frac{n_1^2((1-\hat{p}_1)-(1-\hat{p}))^2}{n_1(1-\hat{p})}+\frac{n_2^2((1-\hat{p}_2)-(1-\hat{p}))^2}{n_2(1-\hat{p})}\\ &=\frac{n_1(\hat{p}_1-\hat{p})^2+n_2(\hat{p}_2-\hat{p})^2}{\hat{p}}+\frac{n_1(\hat{p}_1-\hat{p})^2+n_2(\hat{p}_2-\hat{p})^2}{1-\hat{p}}\\ &=\frac{n_1(\hat{p}_1-\hat{p})^2+n_2(\hat{p}_2-\hat{p})^2}{\hat{p}(1-\hat{p})}\\ &=\frac{n_1\left(\hat{p}_1-\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}\right)^2+n_2\left(\hat{p}_2-\frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}\right)^2}{\hat{p}(1-\hat{p})}\\ &=\frac{n_1n_2^2\left(\frac{\hat{p}_1-\hat{p}_2}{n_1+n_2}\right)^2+n_2n_1^2\left(\frac{\hat{p}_1-\hat{p}_2}{n_1+n_2}\right)^2}{\hat{p}(1-\hat{p})}\\ &=\frac{\frac{n_1n_2}{n_1+n_2}(\hat{p}_1-\hat{p}_2)^2}{\hat{p}(1-\hat{p})}=\frac{(\hat{p}_1-\hat{p}_2)^2}{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\hat{p}(1-\hat{p})}=T\end{aligned} \tag{3.34}$$

从而 $T^* = T$，T^* **是渐进** $\mathcal{X}^2$ **的。**

(c) 我们知道：

$$\hat{p}_1 (1-\hat{p}_1) \to p(1-p), \qquad \hat{p}_2 (1-\hat{p}_2) \to p(1-p) \tag{3.35}$$

结合式3.29及 Slutsky 定理：

$$T^{**} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim AN(0,1) \tag{3.36}$$

于是，在 H_0 **下，**T^{**} **是渐进** $N(0,1)$ **的，因此，**$(T^{**})^2$ **渐进于** $\mathcal{X}_1^2$。进一步，比较 $(T^{**})^2$ 和 T^* 的分母部分，当 $p_1 \neq p_2$ 时：

$$\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2} \neq \left(\frac{1}{n_1} + \frac{1}{n_2}\right)(\hat{p}(1-\hat{p})) \tag{3.37}$$

从而，在一般情况下，$(T^{**})^2 \neq T^*$。

(d) 我们在 (c) 中的讨论中可以看到，如果 $p_1 = p_2$，得到的 $(T^{**})^2 = T^*$。并行 T^* 是基于 $p_1 = p_2$，T^{**} 是基于 $p_1 \neq p_2$。因为我们的假设检验中，我们原来的目的是构建一个 H_0，而且是希望去拒绝 H_0，因此大多数情况下 $p_1 = p_2$ 不成立。于是，我们认为 T^{**} 更好。

(e) 由题意知：$\hat{p}_1 {=} \frac{34}{40}$, $\hat{p}_2 {=} \frac{19}{35}$, $\hat{p} {=} \frac{34+19}{40+35} = \frac{53}{75}$。我们计算式3.33中的统计量 $T = 8.4952$，而我们取 $\alpha = 0.05$, 有 $\mathcal{X}^2_{1,0.95} = 3.8415$。于是，$T > \mathcal{X}^2_{1,0.95}$，我们拒绝 H_0。也就是说，**我们认为石炭酸能降低死亡率。**