# STA 6133: HOMEWORK 4, Spring 2021

**(Posted April 5; due April 16)**

1. The one–sample $t$–test is a statistical procedure to test the hypotheses

$$\mathrm{H}_0 : \mu = \mu_0 \quad \text{against} \quad \mathrm{H}_1 : \mu \neq \mu_0, \tag{1}$$

when the data are a random sample $X_1, \ldots, X_n$ from the $\mathrm{N}(\mu, \sigma^2)$ distribution, with $\mu, \sigma^2$ unknown and $\mu_0$ a fixed number. It has been found that this test is robust to mild departures from normality. The goal of this problem is to assess the robustness of the test to strong departures from normality, by assuming the data are a random sample from one of the following distributions

$$\mathrm{N}(1,2), \qquad \chi^2_{(1)}, \qquad \mathrm{unif}(0,2), \qquad \exp(1). \tag{2}$$

(a) Assume the nominal significance level is $\alpha = 0.1$ or $0.05$. Use Monte Carlo simulation to investigate whether the empirical type I error of the $t$–test for (1) is approximately equal to the nominal significance level when the data are a random sample of size $n = 30$ or $300$ from a distribution in (2).

(b) For the $t$–test with nominal significance level $\alpha = 0.05$ and $n = 300$, use Monte Carlo simulation to estimate the power functions when the data are from a family of distributions like those in (2), but with mean $\mu \in (0, 4)$.

2. Let $X_1, \ldots, X_n$ be a random sample from a distribution that is symmetric about $\theta$, meaning that $X - \theta \stackrel{d}{=} \theta - X$. For continuous random variables this means that for every $x \in \mathbb{R}$ the cdf of $X$ satisfies $F(\theta - x) = 1 - F(\theta + x)$ and the pdf satisfies $f(\theta - x) = f(\theta + x)$. For any of these distributions we have $E_\theta(X) = \mathrm{med}_\theta(X) = \theta$ (provided $E_\theta(X)$ exists), so two natural estimators of $\theta$ are $\bar{X}$ and $M =$ sample median of $\{X_1, \ldots, X_n\}$. A third estimator of $\theta$ is the $\alpha$–trimmed mean, defined as

$$\bar{X}_\alpha = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} X_{(i)},$$

where $\alpha \in (0, 1/2)$, $k = \lfloor n\alpha \rfloor =$ (integer part of $n\alpha$) and $X_{(1)}, \ldots, X_{(n)}$ are the order statistics. For the questions below use as sample size $n = 30$ and as simulation size $M = 100000$.

(a) Consider the following families of distributions:

$$\{\mathrm{N}(\theta, 3) : \theta \in \mathbb{R}\}, \qquad \{\mathrm{Dexp}(\theta, \sqrt{3/2}) : \theta \in \mathbb{R}\}, \qquad \{\mathrm{t}_3(\theta, 1) : \theta \in \mathbb{R}\}.$$

For each of these families compute the mean square error of the estimators $\bar{X}$, $M$ and $\bar{X}_{0.1}$, and comment on how these estimators compare.

(b) Consider now the family of distributions:

$$F_\theta(x) = (1 - \epsilon)\Phi\Big(\frac{x - \theta}{\sqrt{3}}\Big) + \epsilon G_\theta(x), \qquad \theta \in \mathbb{R},$$

where $G_\theta$ is the cdf of the $\mathrm{t}_3(\theta, 1)$ distribution and $\epsilon \in (0, 1)$. This mixture model, usually called a contamination model, indicates that on average $100(1 - \epsilon)\%$ of the observation come from the $\mathrm{N}(\theta, 3)$ distribution, while the rest come from the $\mathrm{t}_3(\theta, 1)$ distribution. In this case we also have $E_\theta(X) = \mathrm{med}_\theta(X) = \theta$. Compute the mean square error of the estimators $\bar{X}$, $M$ and $\bar{X}_{0.1}$, when $\epsilon$ is 0.1 and 0.3, and comment on how these estimators compare.

3. Suppose you have $k \geq 2$ independent random samples $X_{11}, X_{12}, \ldots, X_{1n_1}$; $X_{21}, X_{22}, \ldots, X_{2n_2}$; $\ldots\ldots$; $X_{k1}, X_{k2}, \ldots, X_{2n_k}$, with $n_i \geq 2$ and $\sigma_i^2 = \text{var}(X_{ij})$, $i = 1, \ldots, k$; $j = 1, \ldots, n_i$. Consider testing the hypothesis of equality of variances:

$$H_0 : \sigma_1^2 = \cdots = \sigma_k^2,$$

against the alternative $H_1 : \sigma_{i_1}^2 \neq \sigma_{i_2}^2$ for some $i_1 \neq i_2$. A test to do this with (approximate) significance level $\alpha$ is *Bartlett's* test, which has test statistic

$$T = \frac{(N-k)\log(S_p^2) - \sum_{i=1}^k (n_i - 1)\log(S_i^2)}{1 + \frac{1}{3(k-1)}\Big(\sum_{i=1}^k (\frac{1}{n_i - 1}) - \frac{1}{N-k}\Big)},$$

and rejection region $\{T > \chi^2_{k-1,\alpha}\}$, where $N = \sum_{i=1}^k n_i$,

$$S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2, \qquad \bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \qquad i = 1, \ldots, k,$$

and $S_p^2 = \frac{1}{N-k}\sum_{i=1}^k (n_i - 1)S_i^2$ (the pooled estimate of variance); see `bartlett.test` in `R`. The theory supporting this test assumes all the samples are normally distributed, since in this case the asymptotic distribution of $T$ is $\chi^2_{k-1}$. It has been found that this test is non-robust to deviations form normality, and rejection of $H_0$ is often due to either differences in variance or non-normality. A more robust test is *Levene's* test, which has test statistic

$$W = \frac{(N-k)\sum_{i=1}^k n_i(\bar{Z}_i - \bar{Z})^2}{(k-1)\sum_{i=1}^k \sum_{j=1}^{n_i}(Z_{ij} - \bar{Z}_i)^2},$$

and rejection region $\{W > F_{k-1,N-k,\alpha}\}$, where

$$Z_{ij} = |X_{ij} - \bar{X}_i|, \qquad \bar{Z}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{ij}, \qquad \bar{Z} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij};$$

see `levene.test` in the `R` package `lawstat`.
Suppose $k = 3$, $(n_1, n_2, n_3) = (10, 10, 20)$, $\alpha = 0.05$, and all the random samples are from one of three families: Normal, $t_3$ or exponential. Without lost of generality, for the Normal and $t_3$ families set the all the means to zero. For the questions below use as simulation size $m = 10000$.

(a) For the null models $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$ and $(10, 10, 10)$ estimate the significance level of Bartlett's and Levene's tests for each of the three families of distributions.

(b) For the alternative models $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 3)$ and $(1, 2, 6)$ estimate the power of Bartlett's and Levene's tests for each of the three families of distributions.

(c) Summarize in a concise way your findings in (a) and (b).

4. The following are observations on the time (in hours) between failures of an air conditioning equipment:

```
3   5   7   18   43   85   91   98   100   130   230   487
```

(a) Let $\mu$ be the expected value of the failure time. Assuming these data are a random sample from the exp($\mu$) distribution, use bootstrap to estimate the bias and standard deviation of the MLE of $g(\mu) = 1/\mu$.

(b) Under the same assumption as in (a), compute exact, asymptotic and bootstrap 90% confidence intervals for $\mathrm{P}_\mu(X > 100)$. Comment on the differences.

(c) Assume now the data are a random sample from an unknown cdf $F$. Compute the bootstrap–t, percentile and $\mathrm{BC}_a$ 90% confidence intervals for $\mathrm{P}_F(X > 100)$. Comment on the differences.