Benedict Kongyir Qualifying Exams

1. You have assumed in your demonstrations that the exact distribution your data comes from under the alternative (non-normal) is known. However, this will be unknown in reality. The user may use some pilot data to estimate this. So, in your framework, you do have the option to do your simulations from $\hat{F}$, the empirical cdf of the pilot data assuming it is a good estimator of F, the true cdf generating the data. This is still a limitation since, $\hat{F}$ is just an approximation of F. Conduct a small simulation study as below to assess the effect of this approximation.

   (a) Choose your favorite demonstration set up with a chosen set of test 1 and test 2. Choose an F that you like, e.g. an exponential distribution. Conduct N simulations to obtain the AUC-Power and AUC-Type-I-Error for a varying range of sample sizes. These can be thought of as the "true" parameters showing the efficacy*.

   (b) Now, conduct M repeats of this process where for each repeat, you simulate a pilot data from this distribution, and use the pilot data to generate data in the alternative case. You will get M AUC-Power and M AUC-Type-I-Error values. These can be considered estimates of the true AUC-Power and AUC-Type-I-Error. Create a boxplot for each of these estimates and also show the true parameters next to the boxplot. This should show how good the estimation is.

   Note that you will get AUC-Power and AUC-Type-I-Error for each of test 1, test 2 and adaptive. Therefore there will be six parameters in total.

   Choose N, M and range of sample sizes realistically depending on the computation time. N and M should be fairly large. If that makes it too computationally intensive, consider doing it for just one (or two) sample size(s) and focus on power and type-I error of the tests instead of looking at the AUC.

   *They are actually not the "true" parameters since we have used Monte Carlo simulations to estimate the power and type-I error, but this is a reasonable thing to assume in this context as long as N is large.

   For this question, feel free to ask for clarifications/hints.

Solution

In our user framework in the dissertation, we assumed that the exact family of distribution from which our data comes from is known, however, in practice, this is not always true. One possible, practical solution is to sample from a pilot data, if the user has one, and use that to approximate the true distribution from which the data comes from. But, how well this approximates the true distribution is uncertain. The goal of this simulation study is to assess how well this approximation is by sampling from a pilot data to approximate the true distribution. The empirical cdf $\hat{F}$ constructed from the pilot data is a plug-in estimate for the true F. However, for any finite pilot sample, $\hat{F}$, is only an approximate. We use the two-sample location as a case study to evaluate the effects of this approximation.

Simulation Study Design: We choose Test 1 as the two sample t-test, test 2 as the Mann-Whitney U test, and adaptive test uses test 1 if normality test is insignificant, and test 2, otherwise. For the true distributions (F), exponential(3) and N(0,1) are chosen. Samples from each distribution are standardized to have mean and variance 0, and 1 respectively. A pair of sample of sizes $n \in \{10, 20, 30, 40, 50\}$ are generated. Effect size, $\delta = 0.5$ is applied to get H_1.

For each test and sample size, we estimated the empirical Type I error rate and power at downstream significance level, $\alpha = 0.05$. To summarize performance across all sample sizes, we computed the Area Under the Curve (AUC) for both power and Type I error. where AUC-Power is the integrated area under the power function across the specified sample sizes and AUC-Type-I-Error is the integrated area under the Type I error function across the same sample sizes.

These AUC metrics provide a single value representing a test's overall performance across the experimental design.

Part (a): "True" performance under known F

Here, we evaluate the "true" performance of each test when the data is generated from the known theoretical distribution F. For each combination of distribution F, sample size n, and procedure (two-sample t–test, Mann-Whitney U, adaptive), we simulate $N_{\text{sim}} = 10{,}000$ Monte Carlo replications from both null and alternative hypothesis. For each procedure and n, we estimate the type I error probability

$$\hat{\alpha}(n) = \Pr(\text{reject } H_0 \mid H_0)$$

and the power

$$\hat{\pi}(n) = \Pr(\text{reject } H_0 \mid H_1),$$

We also calculated the AUC-Power and AUC-Type-I as the "true" operating characteristics of each method under each distribution.

Part (b): Effect of Estimating the Distribution via a Pilot Sample

Here, we investigate the impact of replacing the true distribution, F by an empirical approximation $\hat{F}$ based on pilot data, focusing on the non-normal (exponential) case. We investigate how well one can approximate the "true" AUCs from Part (a) when the underlying distribution F is unknown and is replaced by an empirical estimate $\hat{F}_n$ obtained from a finite pilot sample.

For each sample size $n \in \{10, 20, 30, 40, 50\}$, we mimick the practical situation where only a sample of size n is available to estimate F'. In each replication, $m = 1, \ldots, M = 100$, and for each n, we draw a pilot sample of size n from F' and treat it as empirical distribution, $\hat{F}_n$.

Conditional on a given pilot sample and sample size n, we approximate the operating characteristics of the three methods by resampling from $\hat{F}_n$. For each $(n, \hat{F}_n)$, we perform $N_{\text{sim}} = 1{,}000$ Monte Carlo replications. Under the null hypothesis, two groups X and Y (each of size n) were generated by sampling with replacement from the same empirical distribution $\hat{F}_n$. Under the alternative, group X was again sampled from $\hat{F}_n$, while group Y was sampled from $\hat{F}_n$ and then shifted by $\delta = 0.5$.

For each M replicate, we compute p-values from each test procedure (two-sample t-test, Mann-Whitney U test, adaptive test). Empirical Type I error and power are estimated for each method at each n and AUC-Power and AUC–Type I Error are computed across the sample sizes using the same trapezoid rule as in Part (a).

Repeating this procedure over M independent sets of pilot samples produced a sampling distribution of AUC-Power and AUC–Type I Error estimates for each method when the true distribution F' is replaced by $\hat{F}_n$. These sampling distributions are then compared to the corresponding "true" AUCs obtained in Part (a) under the known exponential distribution using a box-plot as shown in Figure 2.

Results and Discussions

We present the results for both part a and the comparison of part a and b below.

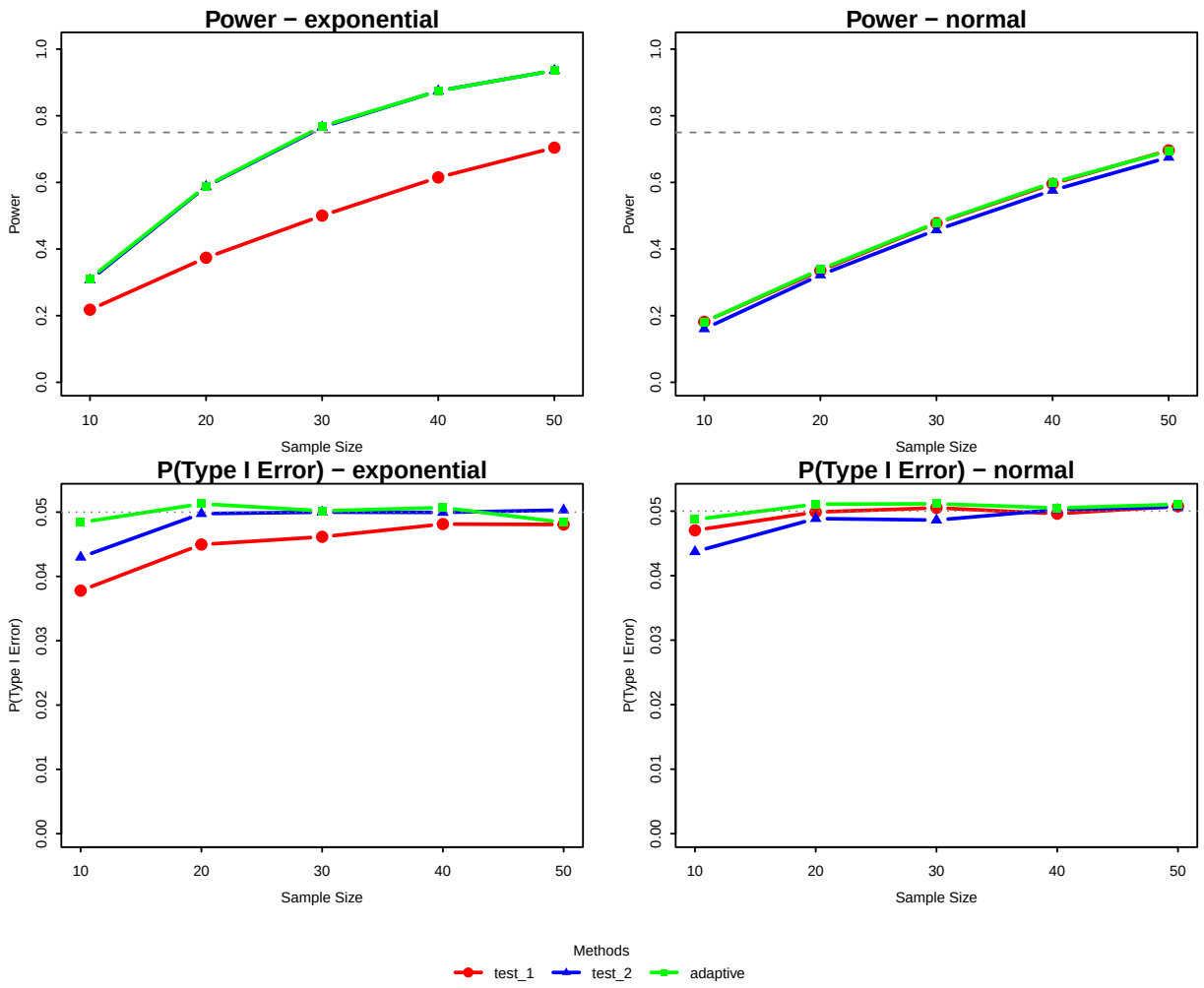

Figure 1: *Power and type I error curves for two-sample t-test, Mann-Whitney U test and adaptive test estimated from F.*

Figure 2 is the plot of power and type I error rates for each sample size generated from F for each of the three procedures (two-sample t–test, Mann-Whitney U, adaptive). Using the Trapezium rule, we approximate the area under each curve as the estimate of the overall performance of each test methods across varying sample sizes.

We compared the each procedure for $\hat{F}$ and F as shown in Table 1 below.

Table 1: Comparison of F and $\hat{F}$ through AUC

Method	Power		Type-I Error	
	F	$\hat{F}$	F	$\hat{F}$
Test 1	0.488	0.543	0.046	0.048
Test 2	0.712	0.732	0.049	0.048
Adaptive	0.714	0.733	0.050	0.048

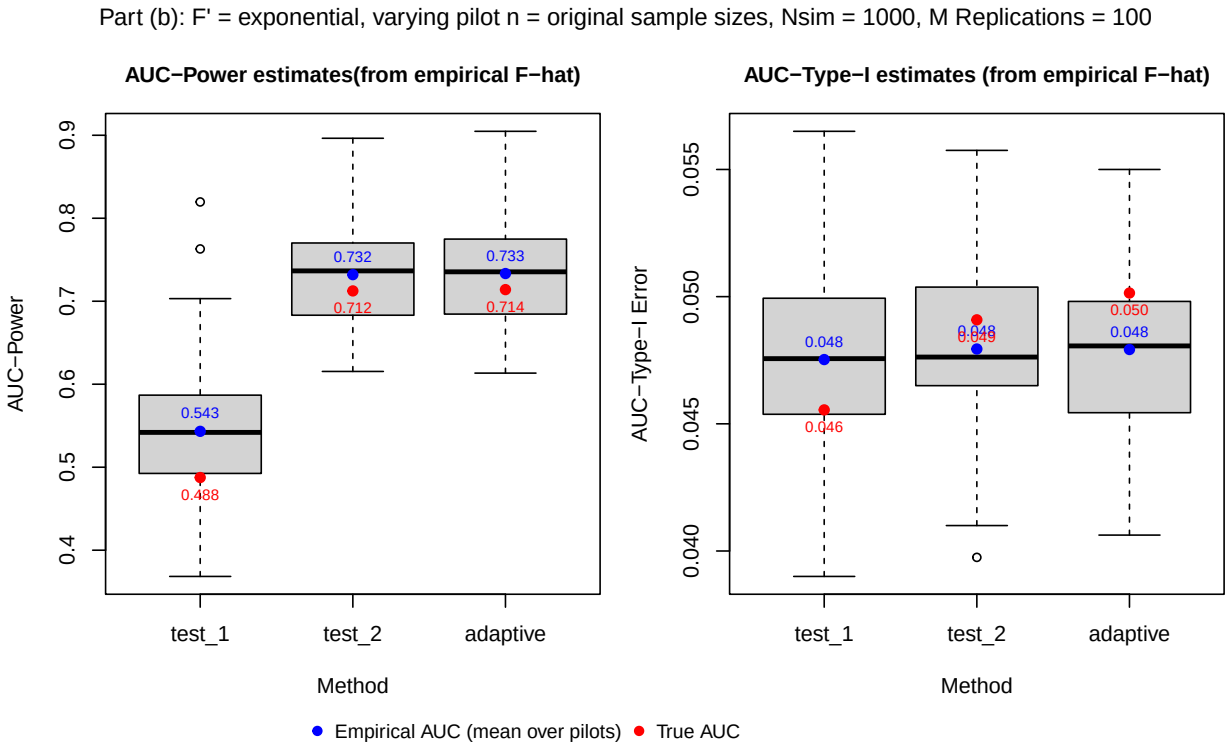

Figure 2: *Box plots of AUC estimates from $\hat{F}$ and F for power and type I error rates.*

The results above indicates that the empirical cdf, $\hat{F}$ is a good estimator of the true cdf, F as the empirical distribution contains the true parameters, AUCs, for all test procedures. Generally, we can say that approximation works well.

2. Show that the problem of selective inference is likely to be less severe if the normality test is really good. Below are some ideas for you do so.

   Consider that the data may come from either of the two cdf: F_0 and F_1 with probabilities π and $1-\pi$. Consider F_0 is cdf of a standard normal distribution. Suppose, a normality test has sensitivity $1-\beta$ and specificity $1-\alpha$.

   (a) Show that when α and β are both 0, the selective inference "issue" is not there. I intentionally kept it vague what I mean by "issue" above. You will have to clearly define it.

   (b) Show that, perhaps with some additional assumptions (about the conditional distributions), the "issue" will be less severe as α and β decrease. You will need to come up with these additional assumptions. This part is more open-ended. Just give it your best try.

   I am not asking you to do simulations here. Please answer using analytical arguments.

   Solution

   We begin by referring to important definitions, and prepositions in the dissertation.

   Definition 0.1 (Normality Pre-test)**.** *Let $\mathcal{D} = \{X_1, \ldots, X_n\}$ denote a random sample from an unknown distribution F with finite mean and variance. Let N_T denote any normality test procedure. Define the* normality pre-test *as a function $\phi : \mathbb{R}^n \to \{0, 1\}$ where $\phi(\mathbf{X}) = 1$ indicates rejection of the null hypothesis $H_0 : F \in \mathcal{N}$ at significance level α_{pre}, with $\mathcal{N}$ denoting the family of normal distributions.*

   Definition 0.2 (Adaptive Test Procedure)**.** *Let **Test 1** be a downstream test procedure whose validity relies on the normality assumption (e.g., a two-sample t-test) and **Test 2** be a downstream test procedure that is valid without distributional assumptions (e.g., a permutation test). The adaptive test procedure based upon a given normality pre-test N_T is defined as:*

$$T_{adaptive} = \begin{cases} T_2 & \text{if } N_T \text{ is significant at } \alpha_{pre} \\ T_1 & \text{otherwise} \end{cases}$$

   where T_1 and T_2 represent Test 1 and Test 2 respectively, applied to the same dataset $\mathcal{D}$.

   Definition 0.3 (Conditional Type I Error)**.** *Let $A = \phi(\mathbf{X}) = 1$ denote the event that the normality test rejects H_0 of normality. Then, the conditional Type I error rates under the*

null hypothesis $H_0 : \theta \in \Theta_0$ are:

$$\alpha_1(F|A^c) = \mathbb{P}_F(\delta_1 = 1|\phi(\mathbf{X}) = 0, H_0) \quad \textit{and} \quad \alpha_2(F|A) = \mathbb{P}_F(\delta_2 = 1|\phi(\mathbf{X}) = 1, H_0)$$

where

$$\delta_j = \begin{cases} 1 & \textit{if } T_j \textit{ rejects } H_0 \\ 0 & \textit{otherwise} \end{cases}$$

for $j \in \{1, 2\}$

Proposition 0.4 (Unconditional Type I Error Rate of Adaptive Procedure)**.** *The overall Type I error rate of the adaptive procedure is:*

$$\alpha_{adapt}(F) = \mathbb{P}_F(A^c)\alpha_1(F|A^c) + \mathbb{P}_F(A)\alpha_2(F|A) \tag{1}$$

where $A = \{\phi(\mathbf{X}) = 1\}$.

Definition 0.5 (Expected Type I Error Inflation)**.** *The expected inflation in Type I error for using the adaptive procedure:*

$$\Delta_{size}(F) \equiv \alpha_{adapt}(F) - \alpha \tag{2}$$

where α is the chosen nominal level, commonly taken to be 0.05.

Definition 0.6 (Power Function of Adaptive Procedure)**.** *Let F_1 be a distribution under the alternative and $\vartheta \in H_1$. Define the power of test j as $\pi_j(F_1, \vartheta)$. Then, the power function of the adaptive procedure is:*

$$\pi_{adapt}(F, \vartheta) = \mathbb{P}_{F_1}(A^c)\pi_1(F, \vartheta \mid A^c) + \mathbb{P}_{F_1}(A)\pi_2(F, \vartheta \mid A) \tag{3}$$

Proposition 0.7 (Expected Power Loss)**.** *The expected power loss relative to always using T_1 is:*

$$\Delta_\pi(F_1, \vartheta) = \pi_{adapt}(F_1, \vartheta) - \pi_1(F_1, \vartheta) = \mathbb{P}_{F_1}(A)\left[\pi_2(F_1, \vartheta, |, A) - \pi_1(F_1, \vartheta, |, A)\right] \tag{4}$$

Part (a): Perfect Normality Test ($\alpha = 0$, $\beta = 0$)

The **selective inference issue** is defined as the inflation of the actual Type I error rate beyond the nominal level due to the adaptive selection process.

When $\alpha = 0$ and $\beta = 0$, we have perfect normality pre-test:

- Specificity $= \mathbb{P}_{F_0}(\phi(\mathbf{X}) = 0) = 1 - \alpha = 1$
- Sensitivity $= \mathbb{P}_{F_1}(\phi(\mathbf{X}) = 1) = 1 - \beta = 1$

Under F_0 (normal data):$\mathbb{P}_{F_0}(A) = \alpha = 0 \implies \mathbb{P}_{F_0}(A^c) = 1 - \alpha = 1$

Substituting into Equation (1) above, we get:

$$\begin{aligned}\alpha_{\text{adapt}}(F_0) &= \mathbb{P}_{F_0}(A^c)\alpha_1(F_0|A^c) + \mathbb{P}_{F_0}(A)\alpha_2(F_0|A) \\ &= 1 \cdot \alpha_1(F_0|A^c) + 0 \cdot \alpha_2(F_0|A) \\ &= \alpha_1(F_0|A^c)\end{aligned}$$

Since the test is perfect, conditioning on A^c under F_0 doesn't change the distribution, so:

$$\alpha_1(F_0|A^c) = \alpha_1(F_0) \leq \alpha$$

assuming T_1 is valid under normality.

Similarly, under F_1 (non-normal data): $\mathbb{P}_{F_1}(A) = 1 - \beta = 1 \implies \mathbb{P}_{F_1}(A^c) = \beta = 0$ Substituting into Equation (1) above we get:

$$\begin{aligned}\alpha_{\text{adapt}}(F_1) &= \mathbb{P}_{F_1}(A^c)\alpha_1(F_1|A^c) + \mathbb{P}_{F_1}(A)\alpha_2(F_1|A) \\ &= 0 \cdot \alpha_1(F_1|A^c) + 1 \cdot \alpha_2(F_1|A) \\ &= \alpha_2(F_1|A) \\ &= \alpha_2(F_1) \\ &\leq \alpha\end{aligned}$$

assuming T_2 is also valid under non-normality.

Therefore:

$$\Delta_{\text{size}}(F_0) = \alpha_{\text{adapt}}(F_0) - \alpha \leq 0$$

When $\alpha = \beta = 0$ (perfect pretest procedure), the adaptive procedure will always use the correct tests (test 1 or test 2), thus $\Delta_{\text{size}}(F_0) = \alpha_{\text{adapt}}(F_0) - \alpha \leq 0$ provided both test is valid.

Therefore, the selective inference "issue" is not there.

Part (b): Assuming α and β decrease

We make the following assumptions:

(a) **Conditional Independence**: The conditional Type I error rates satisfy:

$$\alpha_1(F_0|A^c) = \alpha_1(F_0) = \alpha$$
$$\alpha_2(F_0|A) = \alpha_2(F_0)$$

That is, conditioning on the normality test outcome doesn't affect the type I error rate of T_1 under normality, but T_2 may have different unconditional size.

(b) **Test Performance Ordering**:

- Under F_0: $\alpha_2(F_0) \geq \alpha_1(F_0) = \alpha$ (T_2 may be conservative or have inflated size)
- Under F_1: $\pi_2(F_1, \vartheta|A) \geq \pi_1(F_1, \vartheta|A)$ (T_2 has better power for non-normal data)

(c) **Perfect Specificity under Normality**: The normality test has specificity $1 - \alpha$ and sensitivity $1 - \beta$, where:

$$\mathbb{P}_{F_0}(A) = \alpha \quad \text{(False positive rate)}$$
$$\mathbb{P}_{F_1}(A) = 1 - \beta \quad \text{(True positive rate)}$$

Under Assumption 1, the unconditional Type I error rate from Equation (1) becomes:

$$\begin{aligned}\alpha_{\text{adapt}}(F_0) &= \mathbb{P}_{F_0}(A^c)\alpha_1(F_0|A^c) + \mathbb{P}_{F_0}(A)\alpha_2(F_0|A) \\ &= (1-\alpha)\alpha + \alpha \cdot \alpha_2(F_0)\end{aligned}$$

The Expected Type I Error Inflation is then:

$$\begin{aligned}\Delta_{\text{size}}(F_0) &= \alpha_{\text{adapt}}(F_0) - \alpha \\ &= (1-\alpha)\alpha + \alpha \cdot \alpha_2(F_0) - \alpha \\ &= \alpha[\alpha_2(F_0) - \alpha]\end{aligned}$$

Taking the partial derivatives with respect to α and β we get:

$$\frac{\partial \Delta_{\text{size}}(F_0)}{\partial \alpha} = \alpha_2(F_0) - \alpha \tag{5}$$

$$\frac{\partial \Delta_{\text{size}}(F_0)}{\partial \beta} = 0 \tag{6}$$

Since $\alpha_2(F_0) \geq \alpha$ (Assumption 2), we have:

$$\frac{\partial \Delta_{\text{size}}(F_0)}{\partial \alpha} \geq 0$$

Hence, Type I error inflation decreases monotonically as α decreases.

Similarly, from Equation 4 above, the Expected Power Loss:

$$\begin{aligned}\Delta_{\pi}(F_1, \vartheta) &= \mathbb{P}_{F_1}(A)[\pi_2(F_1, \vartheta|A) - \pi_1(F_1, \vartheta|A)] \\ &= (1-\beta)[\pi_2(F_1, \vartheta|A) - \pi_1(F_1, \vartheta|A)]\end{aligned}$$

Assuming $\pi_2(F_1, \vartheta|A) \geq \pi_1(F_1, \vartheta|A)$ (Assumption 2), we have:

$$\frac{\partial \Delta_{\pi}(F_1, \vartheta)}{\partial \beta} = -[\pi_2(F_1, \vartheta|A) - \pi_1(F_1, \vartheta|A)] \leq 0$$

Thus power under non-normality increases as β decreases.

From the above results Type I error inflation under F_0, decreases to 0 as $\alpha \to 0$ and Power under non-normality increases as $\beta \to 0$.

Therefore, the selective inference issue becomes *less severe* as both the specificity $(1-\alpha)$ and sensitivity $(1-\beta)$ of the normality pre-test improve.