

Performances of Tests for Two Sample Location Problem

Kun Chen
Yunlong Xie

December 10, 2007

Abstract

The Wilcoxon Rank-Sum test, the Kolmogorov-Smirnov test (K-S), and the Two-sample t test are commonly used for two sample location problem. However, their assumptions and hypothesis are not the same, and their performances may vary under different circumstances. For example, The Wilcoxon Rank-Sum test and K-S test are based on non-parametric method, and the former tests whether the distributions are the same but shifted by some amount, while the latter tests whether the two samples come from exactly the same distribution. The Two-sample t test assumes that the samples come from normal distributions with equal variance. It tests whether the means of the distributions are the same or whether the samples come from the same normal distribution if assumptions hold. In this project, we use simulation technique to compare the performances (type I and II error rates) of those three tests for datasets with different sample sizes from several different distributions.

1 Introduction

1.1 The Wilcoxon Rank-Sum test

1. Assumptions:

(x_1, x_2, \dots, x_n) are independent random samples from continuous population 1 having CDF $F(t)$, and (y_1, y_2, \dots, y_n) are independent random sample from continuous population 2 have CDF $G(t)$.

The X's and Y's are mutually independent.

2. Hypothesis:

$$H_0 : F(t) = G(t) \quad H_a : F(t) = G(t - \Delta)$$

1.2 The Kolmogorov-Smirnov (KS) test

1. Assumptions:

(x_1, x_2, \dots, x_n) are independent random samples from continuous population 1 having CDF $F(t)$, and (y_1, y_2, \dots, y_n) are independent random sample from continuous population 2 have CDF $G(t)$.

The X's and Y's are mutually independent.

2. Hypothesis:

$$H_0 : \forall t, F(t) = G(t) \quad H_a : \exists t, F(t) \neq G(t)$$

1.3 The Two-sample t test

1. Assumptions:

(x_1, x_2, \dots, x_n) are independent random samples from normally distributed population 1 and (y_1, y_2, \dots, y_n) are independent random sample from normally distributed population 2.

These two normal distributions have the same variance.

The X's and Y's are mutually independent.

2. Hypothesis:

$$H_0 : \mu_x = \mu_y \quad H_a : \mu_x \neq \mu_y$$

1.4 Comparison

The Wilcoxon Rank-Sum test and K-S test have the same assumption and the same null hypothesis; however, their alternative hypotheses are different. The former tests whether the distributions are the same but shifted by some amount, while the latter tests whether the two samples come from exactly the same distribution. The Two-sample t test assumes that the samples come from normal distributions with equal variance. It tests whether the means of the distributions are the same or whether the samples come from the same normal distribution if assumptions hold.

2 Method

- R was used to generate 1000 pairs of random samples for each Normal, Uniform, Logistic, Exponential, Chi-square, and Cauchy distribution (Table 1) with sample sizes of 10, 20, 30, 50, and 70. Random samples in each pair were independent to each other, and had equal sample size.
- Type I and II error rates of each test were conducted for the two samples of each size from each distribution. We also tested whether the simulated Type I error rate was significantly different from the nominal significance level.

Distribution	Parameters	Notation
Normal	$\mu = \text{mean}, \sigma^2 = \text{variance}$	$N(\mu, \sigma^2)$
Uniform	$l = \text{lower bound}, u = \text{upper bound}$	$U(l, u)$
Logistic	$\theta = \text{threshold parameter}, b = \text{scale parameter}$	$Logistic(\theta, b)$
Exponential	$\lambda = \text{rate parameter}$	$Exp(\lambda)$
Chi-square	$p = \text{degree of freedom}$	$Chisq(p)$
Cauchy	$l = \text{location parameter}, s = \text{scale parameter}$	$Cauchy(l, s)$

Table 1: Descriptions of distributions

Let p =Type I error rate; α =Nominal significance level; n =Simulation times,

$$H_0 : p = \alpha \quad H_a : p \neq \alpha$$

$$Z = \frac{p - \alpha}{\sqrt{\frac{p(1-p)}{n}}}$$

Under H_0 , Z follows $N(0, 1)$ distribution. Reject H_0 if $|Z| \geq Z_{0.025}$.

In this project, we use $\alpha=0.05$ and $n=1000$, thus we will conclude p is significantly different from α if $p \geq 0.0635$ or $p \leq 0.0365$.

- Type II error rates of each test were conducted for the two samples of different amount of location shift from each distribution.
- We use samples from Non-equivariant Normal distributions to check the robustness of the Two-sample t test. We also discussed the Two-sample t test on the samples renormalized by the sample variances. (Instead of test x and y , we test $\frac{X}{S_x^2}$ and $\frac{Y}{S_y^2}$).
- Several kinds of charts are used in order to compare the performances of these three tests:
 - plot of Type I error rate on sample size.
 - plot of Type II error rate on sample size.
 - plot of Type II error rate on shift in location.

3 Result

3.1 Simulation of Type I error rate

Table 2 and Figure 1 summarize the results of type I error rates simulation.

Distributions	Tests	Sample Size				
		10	20	30	50	70
N(0,1)	Wilcoxon	0.05	0.052	0.056	0.047	0.04
	K-S	0.008*	0.036*	0.042	0.04	0.028*
	T	0.063	0.053	0.053	0.048	0.037
U(0,5)	Wilcoxon	0.043	0.059	0.054	0.044	0.053
	K-S	0.01*	0.036*	0.038	0.036*	0.029*
	T	0.048	0.054	0.053	0.044	0.056
L(0,1)	Wilcoxon	0.044	0.05	0.052	0.045	0.044
	K-S	0.012*	0.033*	0.038	0.046	0.027*
	T	0.05	0.052	0.056	0.04	0.046
Exp(1)	Wilcoxon	0.052	0.052	0.048	0.045	0.048
	K-S	0.01*	0.029*	0.035*	0.033*	0.025*
	T	0.041	0.047	0.045	0.04	0.042
Chisq(5)	Wilcoxon	0.054	0.044	0.046	0.037	0.042
	K-S	0.018*	0.031*	0.037	0.027*	0.032*
	T	0.061	0.046	0.049	0.044	0.046
Cau(0,1)	Wilcoxon	0.034*	0.047	0.057	0.055	0.054
	K-S	0.013*	0.03*	0.027*	0.041	0.041
	T	0.015*	0.018*	0.018*	0.025*	0.009*

Note: * means significantly different from α at 95% level.

Table 2: Type I error rate($\alpha=0.05$)

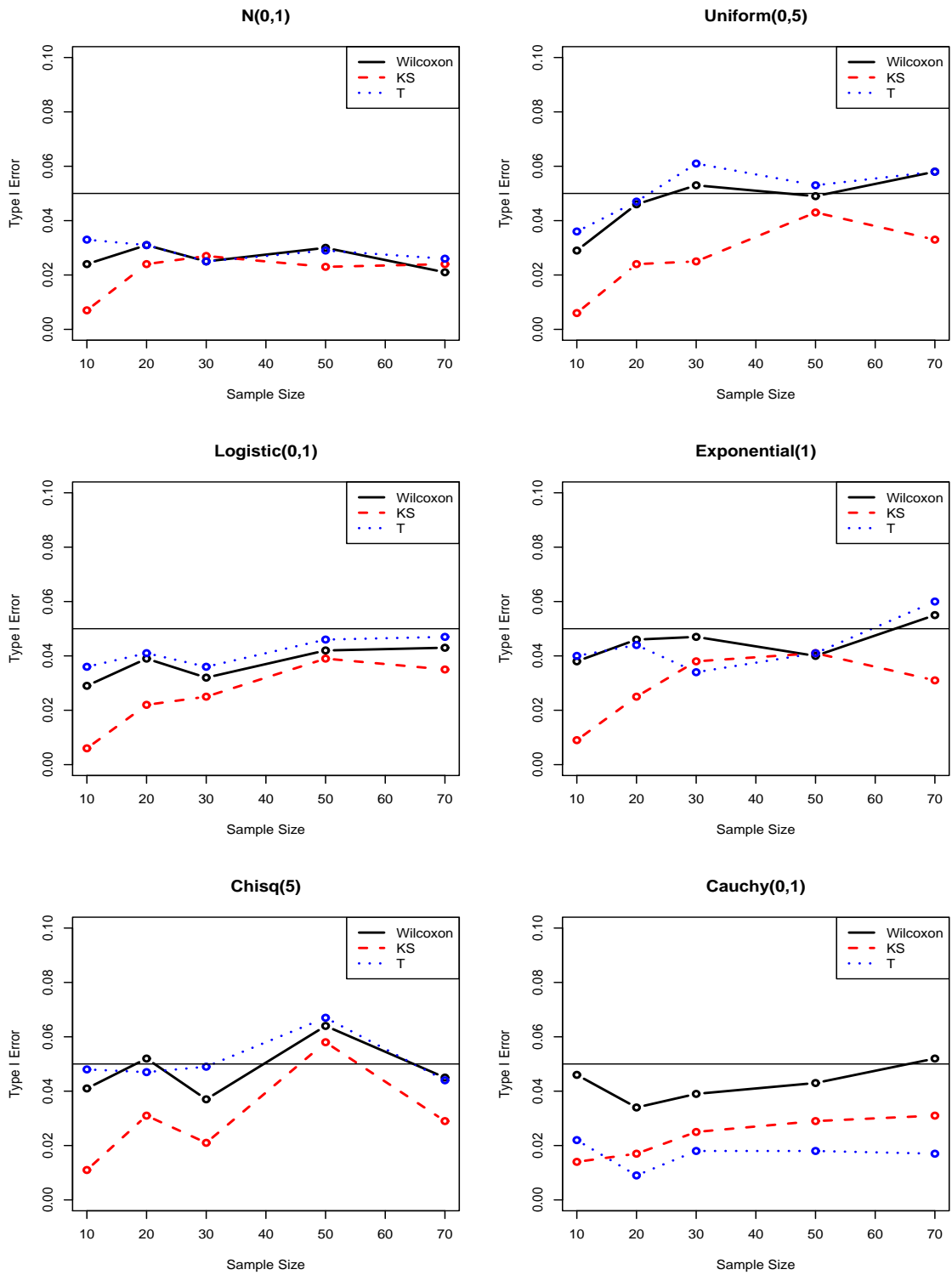


Figure 1: Simulation of Type I error rates

3.2 Type II error rate and sample size

Table 3 and Figure 2 summarize the relationship between type II error rate and sample size.

Distributions	Tests	Sample Size				
		10	20	30	50	70
N(0,1) vs N(0.5,1)	Wilcoxon	0.846	0.73	0.51	0.32	0.18
	K-S	0.95	0.799	0.68	0.466	0.356
	T	0.812	0.689	0.498	0.3	0.163
U(0,5) vs U(1,6)	Wilcoxon	0.761	0.49	0.31	0.105	0.041
	K-S	0.939	0.729	0.566	0.285	0.161
	T	0.713	0.452	0.258	0.066	0.018
L(0,1) vs L(0.5,1)	Wilcoxon	0.923	0.855	0.822	0.695	0.576
	K-S	0.982	0.906	0.877	0.772	0.703
	T	0.913	0.858	0.842	0.72	0.592
Exp(1) vs Exp(1)+0.5	Wilcoxon	0.68	0.394	0.206	0.068	0.015
	K-S	0.842	0.385	0.161	0.015	0
	T	0.765	0.637	0.491	0.299	0.154
Chisq(3) vs Chisq(3)+1	Wilcoxon	0.845	0.635	0.464	0.277	0.133
	K-S	0.95	0.725	0.566	0.314	0.153
	T	0.873	0.737	0.63	0.469	0.328
Cau(0,1) vs Cau(1,1)	Wilcoxon	0.802	0.657	0.481	0.267	0.13
	K-S	0.889	0.641	0.429	0.198	0.085
	T	0.93	0.922	0.933	0.908	0.915

Table 3: Type II Error Rate and Sample Size($\alpha=0.05$)

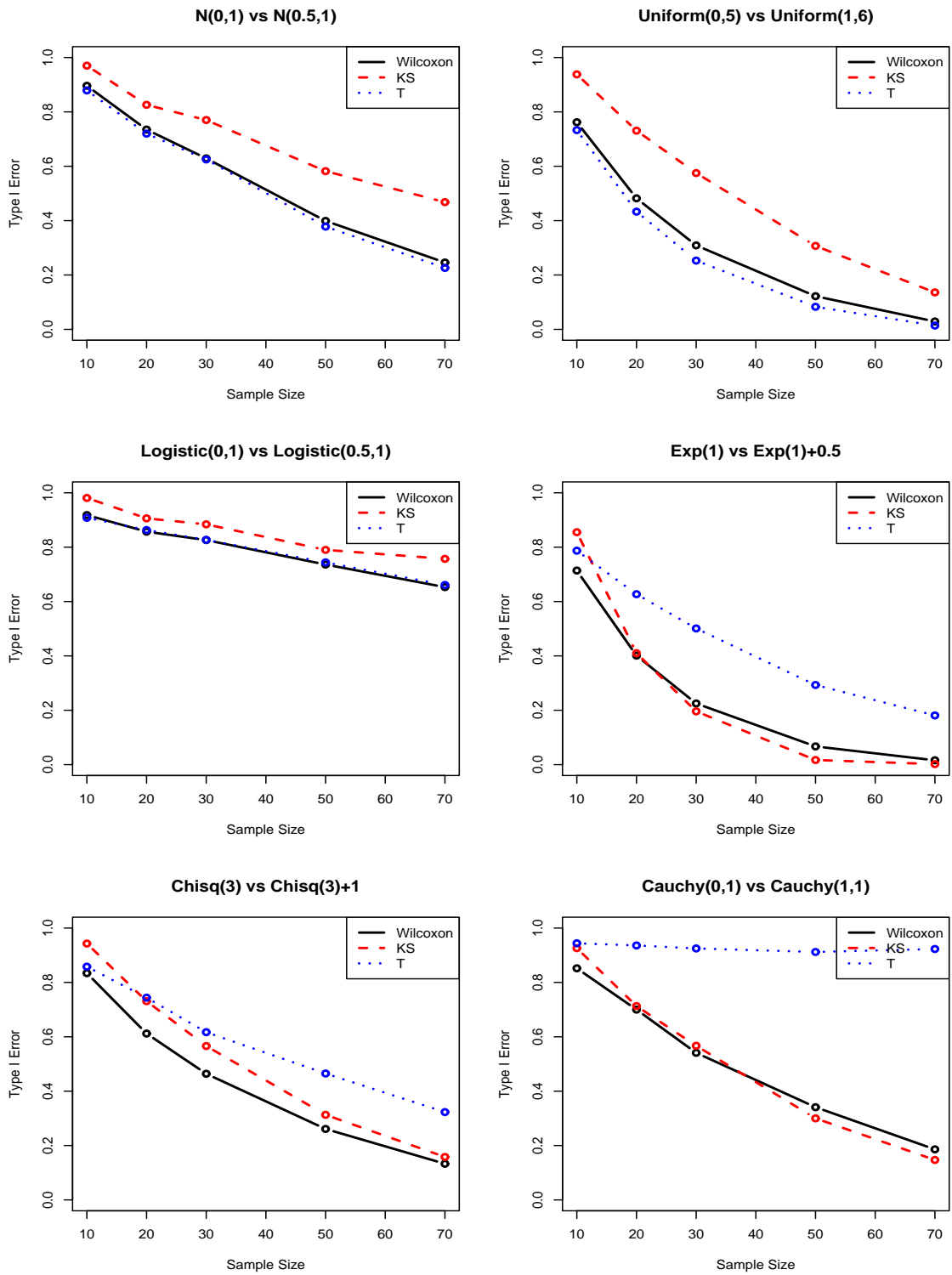


Figure 2: Type II error rate and sample size

3.3 Type II error rate and shift in location

Table 4 and Figure 3 summarize the relationship between type II error rate and the amount of shift in location. We use sample size 30 in this analysis.

Distributions	Tests	Shift in location					
		0.5	1	1.5	2	2.5	3
N(0,1) vs N(*,1)	Wilcoxon	0.93	0.828	0.675	0.49	0.303	0.161
	K-S	0.955	0.893	0.777	0.627	0.468	0.307
	T	0.938	0.817	0.649	0.471	0.283	0.132
U(0,5) vs U(*,*+5)	Wilcoxon	0.808	0.471	0.168	0.029	0.002	0
	K-S	0.908	0.702	0.38	0.117	0.013	0.002
	T	0.802	0.417	0.106	0.012	0	0
L(0,2) vs L(*,2)	Wilcoxon	0.94	0.845	0.734	0.572	0.424	0.247
	K-S	0.957	0.89	0.801	0.696	0.55	0.398
	T	0.943	0.85	0.739	0.58	0.449	0.271
Exp(0.5) vs Exp(0.5)+*	Wilcoxon	0.775	0.406	0.146	0.026	0.008	0.001
	K-S	0.845	0.437	0.101	0.008	0.001	0
	T	0.87	0.635	0.357	0.133	0.045	0.012
Chisq(3) vs Chisq(3)+*	Wilcoxon	0.861	0.649	0.362	0.15	0.052	0.015
	K-S	0.913	0.743	0.454	0.158	0.051	0.01
	T	0.896	0.734	0.506	0.263	0.128	0.041
Cau(0,2) vs Cau(*,2)	Wilcoxon	0.928	0.857	0.765	0.632	0.483	0.391
	K-S	0.942	0.878	0.771	0.62	0.444	0.338
	T	0.975	0.972	0.945	0.928	0.915	0.857

Table 4: Type II Error Rate and Shift in location ($\alpha=0.05$)

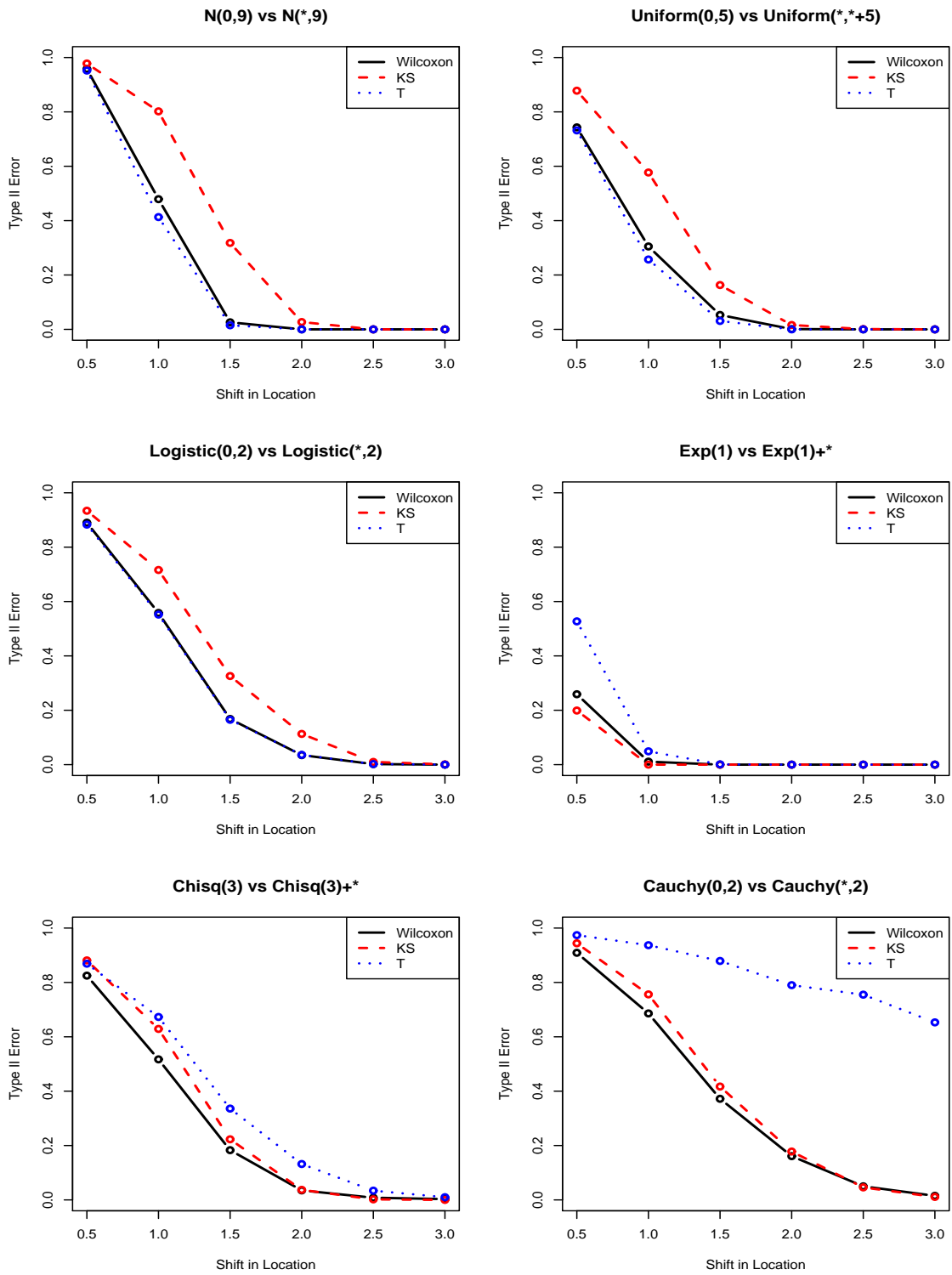


Figure 3: Type II error rate and shift in location

3.4 Non-equivariant variance Problem

Table 5 summarizes the Type I error and Type II error rates for samples from Non-equivariant variance Normal distributions.

Distributions	Error	Tests	Sample Size				
			10	20	30	50	70
N(0,1) vs N(0,9)	Type I	Wilcoxon	0.051	0.071*	0.057	0.071*	0.079*
		K-S	0.053	0.272*	0.498*	0.881*	0.974*
		T	0.049	0.045	0.046	0.049	0.058
		T#	0.051	0.045	0.046	0.049	0.058
N(2,1) vs N(0,64)	Type II	Wilcoxon	0.732	0.48	0.305	0.115	0.044
		K-S	0.802	0.3	0.103	0.004	0
		T	0.732	0.48	0.294	0.095	0.03
		T#	0.067	0.002	0	0	0
Note: * means significantly different from α at 95% level.							
Note: T# means Renormalized T.							

Table 5: Non-equivariant variance problem simulation($\alpha=0.05$)

4 Discussion

4.1 Type I error rates

- As the sample size increases, the type I error rates have no specific trend (not decreasing), and in fact type I error rates will be closer to the nominal significance level. For very large sample sizes, the type I error rates of these three tests are remain approximately at 0.05 (For Wilcoxon and K-S test, large sample approximation is used when sample size bigger than 50).
- Generally, the Type I error rates of Wilcoxon rank sum test and Two-sample t test are approximately at the nominal significance level (except Cauchy); however, for K-S test, many of the Type I error rates are significantly different from the nominal significance level.
- Among these three tests, except Cauchy data, the K-S test has the smallest type I error rates which are always below the nominal significance level (around 3% when $\alpha=0.05$). This means the K-S test is too conservative for two-sample location problem. If we assume rejecting the null hypothesis with a 5% risk of type I error, the risk is actually 3%.
- For Normal, Uniform, Logistic samples, there is no specific difference between the type I error rates of Wilcoxon Rank-Sum test and Two-sample t test. For Exponential and Chi-square samples, the type I error rates of Wilcoxon Rank-Sum test is a little smaller than the Two-sample t test.
- For Cauchy data, Two-sample t test has the smallest type I error rate (around 0.02), and all are significantly different from the nominal significance level, while the Wilcoxon Rank-Sum test has the largest type I error rate (around 0.05). The type I error rate of K-S test is around 0.03.

In conclusion, we would say that the type I error rate of Wilcoxon Rank-Sum test is more accurate.

4.2 Type II error rates

- For samples from distributions other than Cauchy, as the sample size increases, the type II error rates of all three tests decrease. For the Cauchy samples, as the sample size increases, the type II error rates of Wilcoxon Rank-Sum test and K-S test decrease rapidly, but the type II error rates of Two-sample t test does not decrease as much.
- For samples from distributions other than Cauchy, as the shift in location increases, the type II error rates of all three tests decrease. For the Cauchy samples, as the shift in increases, the type II error rates of Wilcoxon Rank-Sum test and K-S test decrease rapidly, but the type II error rates of Two-sample t test does not decrease as much.

- For Normal, Uniform, Logistic samples, K-S test has the greatest type II error rates, which means the power of K-S test is smaller than the other two tests when sample sizes are equal. In order to achieve the same power, K-S test requires more samples. For Exponential, Chi-square and Cauchy samples, Two-sample t test has the greatest type II error rates, which mean the performance of Two-sample test is not good for long-tailed samples or samples with outliers.
- For Normal, Uniform, Logistic samples, the type II error rates of Wilcoxon Rank-Sum test and Two-sample t test appear close to each other. Even for Normal distributions, the type II error rate of Wilcoxon Rank-Sum test is slightly greater than Two-sample t test. For Exponential, Chi-square and Cauchy samples, the type II error rates of Wilcoxon Rank-Sum test is much smaller than Two-sample t test, which means Wilcoxon Rank-Sum test can do a good job for long-tailed samples or samples with outliers.
- For Cauchy data, the type II error rate of Two-sample t test are very large, and as sample size or shift of location increases, it decreases very slowly. This means for Cauchy, Two-sample t test lose almost all the power, while the type II error rate of Wilcoxon Rank-Sum test is smaller than K-S test, and they both decrease rapidly when sample size increases.

In conclusion, we would say that for two sample location problem, Wilcoxon Rank-Sum test is very powerful. Even when the original assumption of t test is true, the Wilcoxon Rank-sum test is almost as good as t test. So , we suggest using Wilcoxon Rank-Sum test when one has non-Gaussian data or when the underlying distribution is unknown.

4.3 Non-equivariant Samples

- To compare means with Two-sample t test, we assume: the samples are Normal, independent, and have the same variance. When the equal variance assumption is violated, the Type I error rate of Two-sample t test remains correct (around 0.05), which means that the Two-sample t test is robust in this case.
- The Type I error of Wilcoxon Rank-Sum test is around 0.05, but the Type I error rate of K-S test is large, far from 0.05.
- In this case, the type II error rates of Wilcoxon Rank-Sum test and Two-sample t test are very high, which mean they lose some power. The K-S test has low Type II error rate, but Type I error rate of K-S test is far from the nominal significance level.

With non-equivariant samples, the p-value of the Two-sample test remains correct, but the power dramatically decreases. If the data looks normal but has different variances, we would rather normalize them in order to perform a Two-sample t test than perform a non-parametric test.

5 Appendix

R program:

1. Simulation of Type I error rates.

```
#initialization
sim_1<-function(rdist1,p1,rdist2, p2,name){
  #initialization
  N<-1000
  diff<-0
  terror<-vector()
  Ierror<-vector()
  kerror<-vector()
  size<-c(10,20,30,50,70)

  #simulation
  for(n in size){
    k <- vector()
    w <- vector()
    t <- vector()
    for (i in 1:N) {
      x <- rdist1(n,p1)
      y <-rdist2(n,p2)
      k <- c(k, (ks.test(x,y,exact=T)$p.value<0.05))
      w <- c(w, (wilcox.test(x,y,exact=T)$p.value<0.05))
      t <- c(t, (t.test(x,y)$p.value<0.05))
    }
    terror<-c(terror,sum(t)/N)
    Ierror<-c(Ierror,sum(w)/N)
    kerror<-c(kerror,sum(k)/N)
  }

  #output
  cat("Sample Size","\n",size,"\n")
  cat("Type I error of wilcoxon Test","\n",Ierror,"\n")
  cat("Type I error of K-S Test","\n",kerror,"\n")
  cat("Type I error of t Test","\n",terror,"\n")

  #figure
  plot(size,Ierror,ylim=c(0,0.10),col='black',lty=1,
        lwd=2,type='b',xlab="Sample Size",ylab="Type I Error")
  lines(size,kerror, col='red',type='b',lwd=2,lty=2)
  lines(size,terror,col='blue',type='b',lwd=2,lty=3)
  abline(h=0.05)
  legend("topright",
```

```

        c('Wilcoxon', 'KS', 'T'),
        lty=c(1,2,3),
        lwd=c(2,2,2),
        col=c('black', 'red', 'blue'))
        title(main=name)
    }

postscript("/mnt/nfs/fileserv/grad/kuchen/Desktop/figure1.eps",horizon=F)
par(mfrow=c(3,2))
sim_1(rnorm,c(0,1),rnorm,c(0,1),"N(0,1)")
sim_1(runif, c(0,5),runif, c(0,5),"Uniform(0,5)")
sim_1(rlogis,c(0,1),rlogis,c(0,1),"Logistic(0,1)")
sim_1(rexp,1,exp,1,"Exponential(1)")
sim_1(rchisq,5,rchisq,5,"Chisq(5)")
sim_1(rcauchy,c(0,1),rcauchy,c(0,1),"Cauchy(0,1)")
graphics.off()

```

2. Type II error rate of different sample size.

```

sim_2<-function(rdist1,p1,rdist2, p2,delta,name){
  #initialization
  N<-1000
  terror<-vector()
  Ierror<-vector()
  kerror<-vector()
  size<-c(10,20,30,50,70)

  #simulation
  for(n in size){
    k <- vector()
    w <- vector()
    t <- vector()
    for (i in 1:N) {
      x <- rdist1(n,p1)
      y <-rdist2(n,p2)+delta
      k <- c(k, (ks.test(x,y,exact=T)$p.value>0.05))
      w <- c(w, (wilcox.test(x,y,exact=T)$p.value>0.05))
      t <- c(t, (t.test(x,y)$p.value>0.05))
    }
    terror<-c(terror,sum(t)/N)
    Ierror<-c(Ierror,sum(w)/N)
    kerror<-c(kerror,sum(k)/N)
  }
}

```

```

#chart
plot(size,Error,ylim=c(0,1),col='black',lty=1,lwd=2,type='b',
xlab="Sample Size",ylab="Type I Error")
lines(size,kerror,col='red',type='b',lwd=2,lty=2)
lines(size,terror,col='blue',type='b',lwd=2,lty=3)
legend("topright",
      c('Wilcoxon', 'KS', 'T'),
      lty=c(1,2,3),
      lwd=c(2,2,2),
      col=c('black', 'red', 'blue'))
      title(main=name)
}

postscript("/mnt/nfs/fileserv/grad/kuchen/Desktop/figure2.eps",horizon=F)
par(mfrow=c(3,2))
sim_2(rnorm,c(0,1),rnorm,c(0,1),0.5,"N(0,1) vs N(0.5,1)")
sim_2(runif,c(0,5),runif,c(0,5),1,"Uniform(0,5) vs Uniform(1,6)")
sim_2(rlogis,c(0,1),rlogis,c(0,1),0.5,"Logistic(0,1) vs Logistic(0.5,1)")
sim_2(rexp,1,1,0.5,"Exp(1) vs Exp(1)+0.5")
sim_2(rchisq,3,rchisq,3,1,"Chisq(3) vs Chisq(3)+1")
sim_2(rcauchy,c(0,1),rcauchy,c(0,1),1,"Cauchy(0,1) vs Cauchy(1,1)")
graphics.off()

```

3. Type II Error with different location shifts.

```

sim_3<-function(rdist,p,name){
#initialization
N <- 1000
n<-30
terror<-vector()
Error<-vector()
kerror<-vector()
diff<-c(0.5,1,1.5,2,2.5,3)

#simulation
for(delta in diff){
k <- vector()
w <- vector()
t <- vector()
for (i in 1:N) {
x <- rdist(n,p)
y <-rdist(n,p)+delta
k <- c(k, (ks.test(x,y,exact=T)$p.value>0.05))
w <- c(w, (wilcox.test(x,y,exact=T)$p.value>0.05))
}
}
}

```

```

    t <- c(t, (t.test(x,y)$p.value>0.05))
  }
  terror<-c(terror,sum(t)/N)
  Ierror<-c(Ierror,sum(w)/N)
  kerror<-c(kerror,sum(k)/N)
}

#chart
plot(diff,Ierror,ylim=c(0,1),col='black',lty=1,lwd=2,type='b'
,xlab="Shift in Location",ylab="Type II Error")
lines(diff,kerror,col='red',type='b',lwd=2,lty=2)
lines(diff,terror,col='blue',type='b',lwd=2,lty=3)
legend("topright",
      c('Wilcoxon', 'KS', 'T'),
      lty=c(1,2,3),
      lwd=c(2,2,2),
      col=c('black', 'red', 'blue'))
  title(main=name)
}

postscript("/mnt/nfs/fileserv/grad/kuchen/Desktop/figure3.eps",horizon=F)
par(mfrow=c(3,2))
sim_3(rnorm,c(0,3),"N(0,9) vs N(*,9)")
sim_3(runif,c(0,5),"Uniform(0,5) vs Uniform(*,5)")
sim_3(rlogis,c(0,2),"Logistic(0,2) vs Logistic(*,2)")
sim_3(rexp,1,"Exp(1) vs Exp(1)+*")
sim_3(rchisq,3,"Chisq(3) vs Chisq(3)+*")
sim_3(rcauchy,c(0,2),"Cauchy(0,2) vs Cauchy(*,2)")
graphics.off()

```