Comparsion of Normality Tests Using Monte Carlo Simulation

Yihong Gu
Department of Computer Science
Tsinghua University
yihong15.math@gmail.com

Abstract

In this paper, we carefully compare the performance of different methods of normality test. We found that Shapiro-Francia test outperform other methods in most case according to the evaluation of power function. Moreover, all the tests have large sample property, i.e., when significance level α is fixed, the power will approximate to 1 when n approxmiate ∞. We also plot the power function under some careful design of parametric data generation model.

1 Hypothesis Testing

We firstly review some basic ideas of hypothesis testing. Here we assume that $X_1, X_2, \cdots, x_n \sim$ i.i.d.$F(x)$, where $F(x)$ is the distribution function in the functional space $\mathcal{F}$, here we divide the whole functional space $\mathcal{F}$ into two subsets $\mathcal{F}_0, \mathcal{F}_1$, i.e., $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$ and $\mathcal{F}_0 \cap \mathcal{F}_1 =$.

We using the perspective of frequenist, we assume the true function $F^*(x)$ actually exists and is in the functional space $\mathcal{F}$.

Using $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$, we define our **null hypothesis** $H_0 : F^*(x) \in \mathcal{F}_0$ and the **alternative hypothesis** $H_1 : F^*(x) \in \mathcal{F}_1$.

After define our null hypothesis and alternative hypothesis, we then introduce the **reject area** $\mathscr{R}(c)$, which is the subset of the sample space $\mathscr{X}$. Note that the sample space is the product space of X's corresponding probability space $(\Omega, \mathscr{F}, \mathbb{P}, \mathrm{X})$. Moreover, c is the (vector) of parameter which will determine $\mathscr{R}$. Afterwards we determine the **significance level** α, and let $c = c^*$, which satisfy

$$\sup_{F \in \mathcal{F}_0} \mathbb{P}_F(x \in \mathscr{R}(c^*)) = \alpha$$

Moreover, in reality, we always get in a **test statistic** T, and covert the expression $x \in \mathscr{R}$ into $T \in E$. Pratically, E is often an interval.

we use **power function** to measure the performace of our test, which is define as followings

$$\beta(F) = \mathbb{P}_F(x \in \mathscr{R}_\alpha)$$

A good test method will make $\beta(F)$ approximates 0 when $F \in \mathcal{F}_0$ and approximates 1 when $F \in \mathcal{F}_1$.

In order to ensure whether we want to reject the hypothesis, we just need to calculate the value of test statistic according to the samples we get and then check if $T \in E$. Moreover, we can use another method to do the above steps, instead of concrete $\mathscr{R}$ according to α, we define **p-value** $p(x) = \inf_{\alpha(0,1)}\{x \in \mathscr{R}_\alpha\}$ and reject the null hypothesis when $p(x) < \alpha$.

2 Normality Tests

In this paper, we mainly consider the following five normality tests in 'nortest' package in R:

- ad test, short for Anderson-Darling test
- cvm test, short for Cramer-von Mises test
- lillie test, short for Lilliefors (Kolmogorov-Smirnov) test
- pearson test, short for Pearson chi-square test
- sf test, short for Shapiro-Francia test

Here we emphasize the principle of Shapiro-Francia test

RLab homework for course Linear Regression, Spring 17, Tsinghua University, Student ID: 2015011249
The code can be found in Github: https://github.com/wmyw96/LinearRegression-Spring17/

2.1 Shapiro-Francia test

The test statistic is

$$W = \frac{\sum_{i=1}^{n} a_i x_{(i)}}{\sum_{i=1}^{n} (x - \bar{x})^2} \tag{1}$$

where

- $x_{(i)}$ is the ith order statistic, i.e., the ith-smallest number in the sample;
- $x = (x_1 + \cdots + x_n)/n$ is the sample mean;
- the constants a_i are given by (2)

$$(a_1, \cdots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}} \tag{2}$$

where $m = (m_1, \cdots, m_n)^T$, and $m_1, \cdots, m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics.

3 Evaluation

3.1 Simulation and Calculation of Power Function

In Section 1, we found that the central task here is the calculation of the power function, i.e., given the F and α, and calculate

$$\beta(F) = \mathbb{P}_F(x \in \mathscr{R}_\alpha) \tag{3}$$

actually, due to the complexity of the test statistics, it is hard for us to calculate the power, so instead of calcuate if analytically, we use monte carlo simluation to calculate it. The process including the following steps

1. Fix the significance level α, and sample $X_1, X_2, \cdots, X_n$ from population $F(x)$.
2. Calculate the p-value of the test, and we will reject the hypothesis when $p(x) < \alpha$, here $x = (X_1, \cdots, X_n) \in \mathcal{X}$.

We will repeat the process above T times, and use frequency to estimate the probabilily, i.e., use

$$\hat{\mathbb{P}}_F(x \in \mathscr{R}_\alpha) = \frac{\#\{\text{Rejected Tests}\}}{T} \tag{4}$$

To estimate $\beta(F)$, and use the value monte carlo simulation provided to do the following analysis.

3.2 Data Generation and Experiment Design

We will perform our exprimenst in the following steps.

1. We fix the siginificance level $\alpha = 0.05$.
2. We set the data size $n = 30, 100, 1000$, and see how the estimators performed under diffenent scales of data.
3. We determine the true distribution $F(x)$:
   - Normal distribution: $\mathcal{N}(0,1)$, $\mathcal{N}(5,1)$, $\mathcal{N}(0,9)$
   - Other trivial continuous distribution: $U(0,1)$, Cauchy$(0,1)$, Gamma$(5,3)$, Exp(1), $\mathcal{X}_10^2$, T_10.
   - Mixture normal distribution: it will generate X from $\mathcal{N}(0,1)$ with probability π, and generate X from $\mathcal{N}(\mu, \sigma^2)$ with probability $1-\pi$. Here π, μ, σ^2 are both parameters.
   - Student t's approxmiation: it will generate X from T_ν, here ν is parameter.
   - Intrisic Error: we generate Y from $\mathcal{N}(0,1)$ and give X in the from $X = Y + \epsilon \sin(Y)$, here ϵ is parameter.
4. Calculate the estimate of the power $\hat{\beta}(\theta)$ according to eqation 4 using $T = 10000$ simluations, where θ refer to the parameter of the distribution (if it has parameters).

Our evaluation is divided into two main parts, for non-parametric distribution(Normal, Other trivial), we simply report the results and compare their performance according to the value. For paramteric distribution (Mixture, Student t's, Intrisic), we plot the power function curve according to each parameter and see how they behave according the change of the parameter.

Moreover, for the parametric distribution, our general hypothesis test will become parametric hypothesis test. For the case of Intrisic Error, our null hypothesis will become $\epsilon = 0$ and the alternative hypothesis will then become $\epsilon \neq 0$. For the case of Student t's approximation, since we can get that T_ν will converge to $\mathcal{N}(0,1)$ in distribution when $\nu \to \infty$, so the null hypothesis is $\nu = \infty$ and the alternative hypothesis is $\nu < \infty$. For the case of Mixture normal distribution, when we regard μ and σ^2 as constant, then the null hypothesis will become $\pi = 1$ and the alternative hypothesis will then become $\pi < 1$.

Because $\alpha = 0.05$ so we repeat 10000 times instead of 1000 times to get a stable result.

3.3 Implemention Details

We use the implementation in package 'nortest' to test the performance of different methods.

We fixed the random seed to be 123469 and use R-package 'ggplot2' to generate plots and 'xtable' to directly convert data.frame in R to table format in Latex.

4 Experiments and Results

4.1 Normal Distribution: Type I Error

We found the choice of μ and σ don't affect the results, so without loss of generality, we only report the results when $\mu = 0$ and $\sigma^2 = 1$

	n=30	n=100	n=1000
ad	0.0483	0.0485	0.0448
cvm	0.0491	**0.0472**	0.0461
lillie	**0.0467**	**0.0472**	**0.0446**
pearson	0.0521	0.0492	0.0507
sf	0.0553	0.0504	0.0500

Table 1: Simluation Results of Type I Error, $X \sim \mathcal{N}(0,1)$, simluating $T = 10,000$ times

The table below reports the results when $T = 100,000$, we found that the result both approximate 0.05 but don't have large sample property($T \to \infty$).

	n=30	n=100	n=1000
ad	0.0497	**0.0496**	0.0486
cvm	0.0499	0.0510	0.0492
lillie	**0.0479**	0.0509	**0.0466**
pearson	0.0513	0.0524	0.0516
sf	0.0530	0.0517	0.0516

Table 2: Simluation Results of Type I Error, $X \sim \mathcal{N}(0,1)$, simluating $T = 100,000$ times

4.2 Other Trivial Distribution: Type II Error

Firstly, we plot the pdf of each distribution in Figure 1, note the black curve refer to $\mathcal{N}(0,1)$'s pdf, and red, orange, green, blue, purple, brown refers to the pdf curve of $U(0,1)$, Cauchy$(0,1)$, Gamma$(5,3)$, Exp(1), $\mathcal{X}^2_{10}$, T_{10} repspectively. All the pdf function are standardized except for Cauchy distribution in order to get a zero mean and unit variance. Also it should be noted that the green curve is overlapped by purple curve.

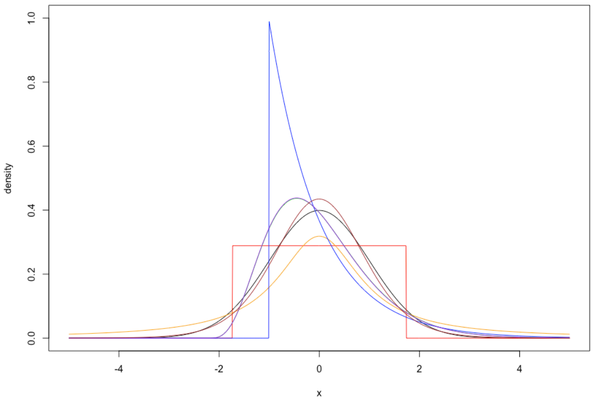

Figure 1: PDF of the distributions

We quickly report the results of Type II Error (both simulate $T = 10,000$ times) in Table 3 $\sim$ 8:

	n=30	n=100	n=1000
ad	**0.7028**	0.0486	**0.0000**
cvm	0.7697	0.1574	**0.0000**
lillie	0.8589	0.4087	**0.0000**
pearson	0.8926	0.5409	**0.0000**
sf	0.8285	**0.0311**	**0.0000**

Table 3: Simluation Results of Type II Error, $X \sim U(0,1)$

	n=30	n=100	n=1000
ad	0.0362	**0.0000**	**0.0000**
cvm	0.0365	**0.0000**	**0.0000**
lillie	0.0583	**0.0000**	**0.0000**
pearson	0.0963	0.0001	**0.0000**
sf	**0.0307**	**0.0000**	**0.0000**

Table 4: Simluation Results of Type II Error, $X \sim$ Cauchy$(0,1)$

	n=30	n=100	n=1000
ad	0.6874	0.1919	**0.0000**
cvm	0.7216	0.2625	**0.0000**
lillie	0.7870	0.3985	**0.0000**
pearson	0.8590	0.6090	**0.0000**
sf	**0.6419**	**0.1255**	**0.0000**

Table 5: Simluation Results of Type II Error, $X \sim$ Gamma$(5,3)$

	n=30	n=100	n=1000
ad	0.0638	**0.0000**	**0.0000**
cvm	0.0988	**0.0000**	**0.0000**
lillie	0.2170	0.0001	**0.0000**
pearson	0.1419	**0.0000**	**0.0000**
sf	**0.0504**	0.0000	**0.0000**

Table 6: Simluation Results of Type II Error, $X \sim \text{Exp}(1)$

	n=30	n=100	n=1000
ad	0.6874	0.1919	**0.0000**
cvm	0.7216	0.2625	**0.0000**
lillie	0.7870	0.3985	**0.0000**
pearson	0.8590	0.6090	**0.0000**
sf	**0.6419**	**0.1255**	**0.0000**

Table 7: Simluation Results of Type II Error, $X \sim \mathcal{X}^2_{10}$

We can think that the results is consistent with our intuition after seeing the plot of pdf: (except Cauchy distribution), expoential distribution is the most simple one to be distinguished. Then follows the uniform distribution, the results of Gamma(5, 3) are same with ChiSquare(10) because they are identically distributed after standardized. It might be hard to detect student's t distribution and it's curve is most similar to the curve of normal distribution.

We can draw the following conclusions:

1. When the significance level is fixed, Shapiro-Francia test performed best out of the five methods.

2. Both test methods have large sample property: when $n \to \infty$, the type II error will converge to 0 (In probability or almost everywhere).

3. It might be difficult for test to distiguish T_ν and $\mathcal{N}(0, 1)$ when ν is very large.

Here we give some plots of the large sample property of each distribution in Figure 2, 3, 4, 5, 6, 7

	n=30	n=100	n=1000
ad	0.8975	0.8349	0.2235
cvm	0.9080	0.8574	0.3050
lillie	0.9188	0.8894	0.5021
pearson	0.9379	0.9241	0.8043
sf	**0.8486**	**0.7074**	**0.0757**

Table 8: Simluation Results of Type II Error, $X \sim T_{10}$

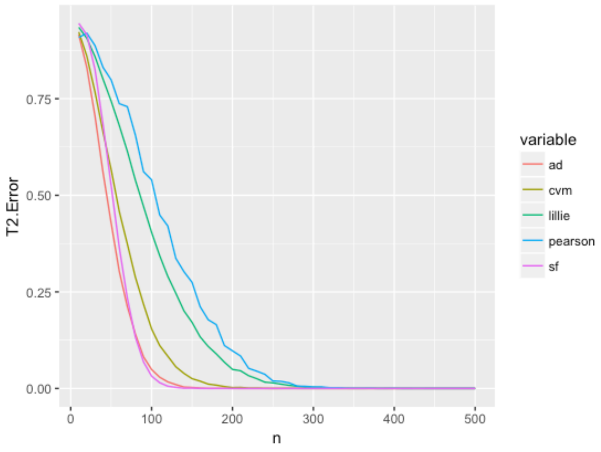

Figure 2: Type II Error versus n for $U(0, 1)$

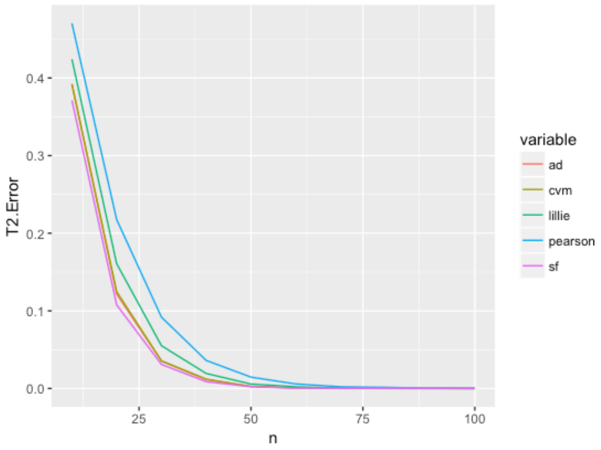

Figure 3: Type II Error versus n for Cauchy$(0, 1)$

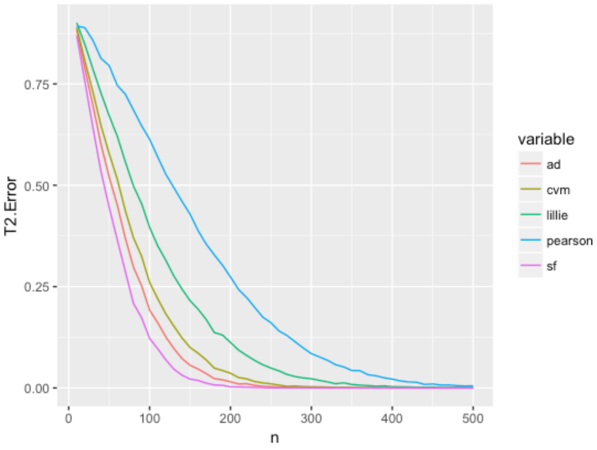

Figure 4: Type II Error versus n for Gamma$(5, 3)$

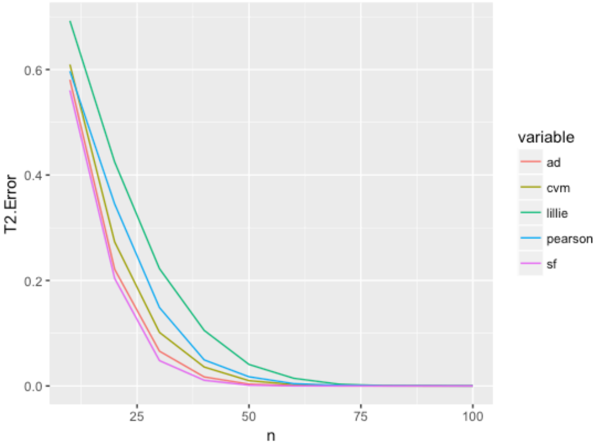

Figure 5: Type II Error versus n for Exp(1)

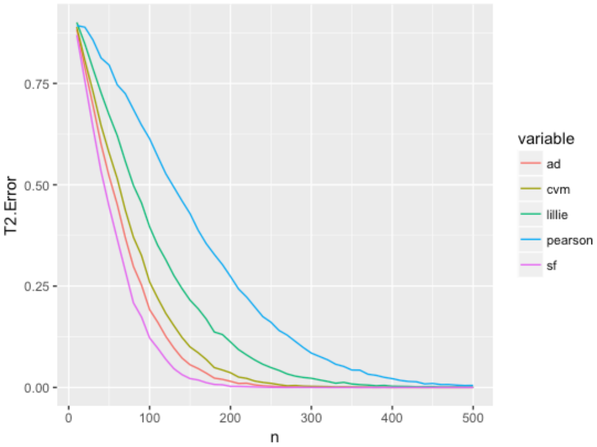

Figure 6: Type II Error versus n for $\mathcal{X}^2_{10}$

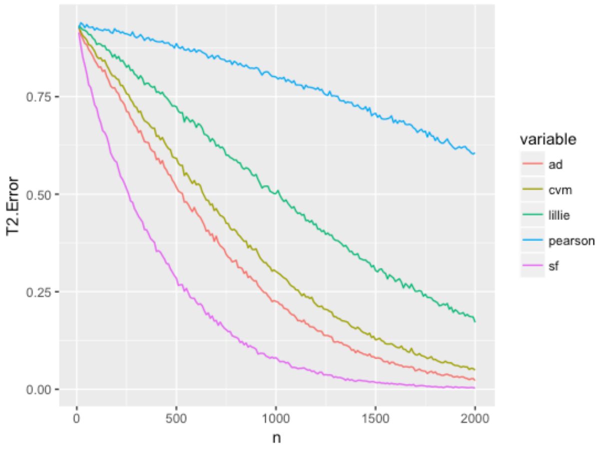

Figure 7: Type II Error versus n for $\sim T_{10}$

The plots give us some estimate information about number of samples we need to distinguish between other distributions and normal distribution with Type II Error is under a particular value, for example $\beta = 0.05$, for SF test, We need more than 100 samples for uniform distribution, more than 40 samples for cauchy distribution, more than 150 samples for Gamma$(\alpha = 5)$, more than 38 samples for exponential distribution, more than 1600 samples for T_{10}.

4.3 Mixture normal distribution

We generate data from the population with pdf $f(x) = \pi\varphi(x) + (1 - \pi)\varphi((x - \mu)/\sigma)$, where $\varphi(x)$ is the pdf of $\mathcal{N}(0, 1)$. We can then regard the original hypothesis testing as a parametric hypothesis testing, with null hypothesis $\pi = 0$ and alternative hypothesis $\pi \in (0, 1/2]$.

Here we let $\mu = 5, \sigma = 1$ and change π, then plot the power function for all the methods for $n = 30$, $n = 100$ and $n = 1000$ in Figure 8, 9 and 10

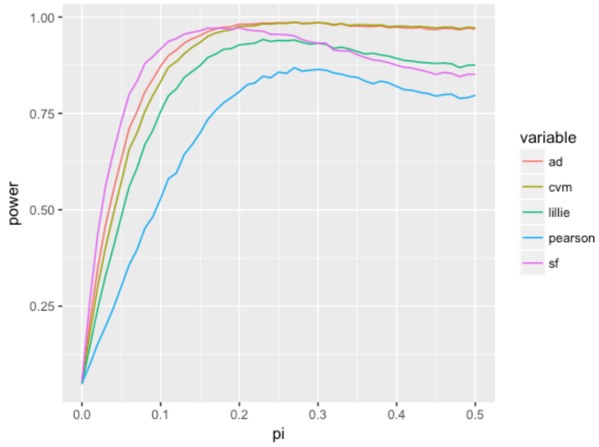

Figure 8: Power function with mixture normal popu-

lation, $n = 30$

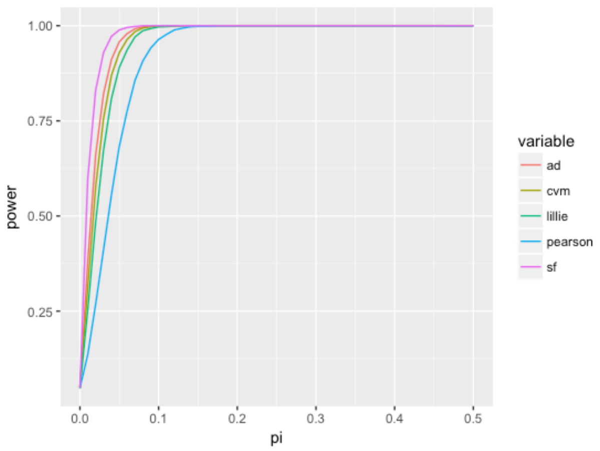

Figure 9: Power function with mixture normal population, $n = 100$

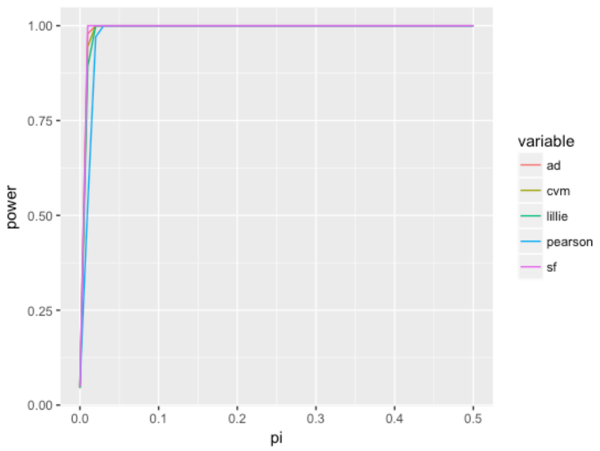

Figure 10: Power function with mixture normal population, $n = 1000$

We can see that SF test outperform other methods when $n = 100$ and $n = 1000$, which has the largest power. However, when $n = 30$ and $\pi \geq 0.2$ we can see the power decreases for lillie, pearson and sf method.

4.4 Student's t Approximation

Since the pdf of the Students's t distribution is

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{5}$$

We can see that when $\nu \to \infty$, the term which contains t will coverge to $e^{-t^2/2}$, which approximate the form of normal distribution. When we regard ν as parameter, the original hypothesis testing could become a parametric hypothesis testing, therefore can easily plot the power function can evaluate its property. We assume that when ν is very large, the samples are sampling from normal distribution.

We plot the power function for all the methods for $n = 30$, $n = 100$ and $n = 1000$ in Figure 11, 12 and 13

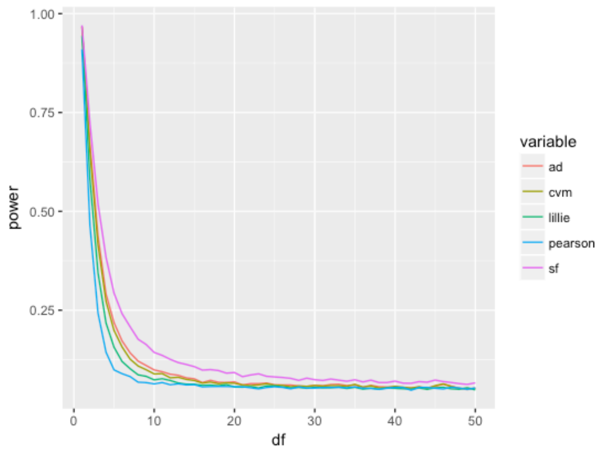

Figure 11: Power function with population T_ν, $n = 30$

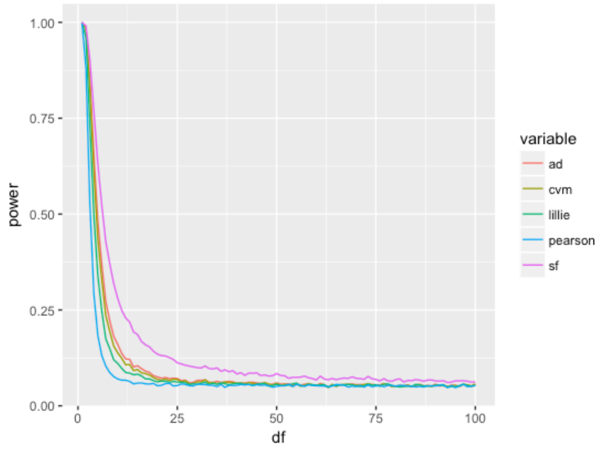

Figure 12: Power function with population T_ν, $n = 100$

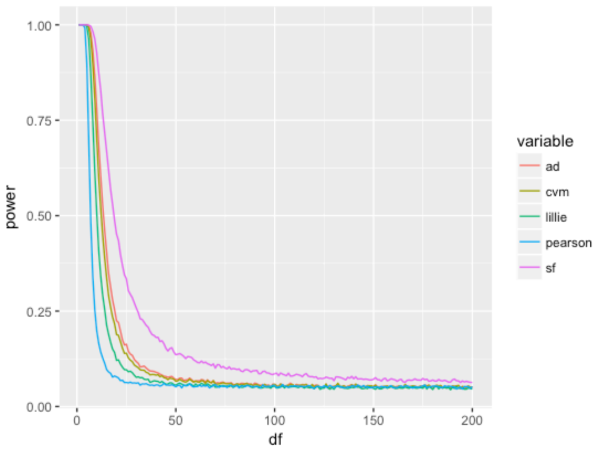

Figure 13: Power function with population T_ν, $n = 1000$

We can see that SF test outperform other methods, which has the largest power.

4.5 Intrisic Error

We firstly sample Y from $\mathcal{N}(0,1)$ and use transform $X = Y + \epsilon \sin(Y)$ to generate data X. We can also regard the original hypothesis testing as a parametric hypothesis testing, with null hypothesis $\epsilon = 0$ and alternative hypothesis $\epsilon \in (0,1]$.

We plot the power function for all the methods for $n = 30$, $n = 100$ and $n = 1000$ in Figure 14, 15 and 16

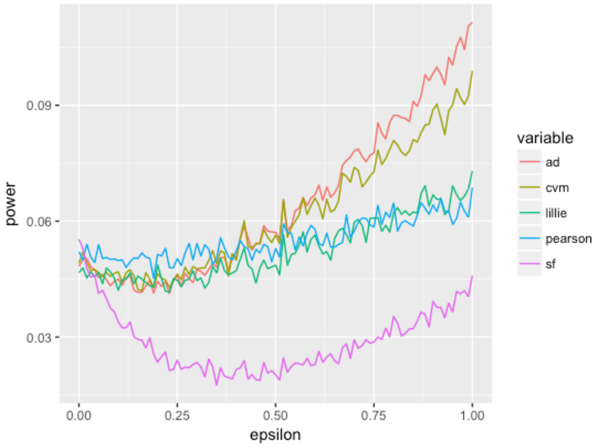

Figure 14: Power function with tranformed population $X = Y + \epsilon \sin(X)$, $n = 30$

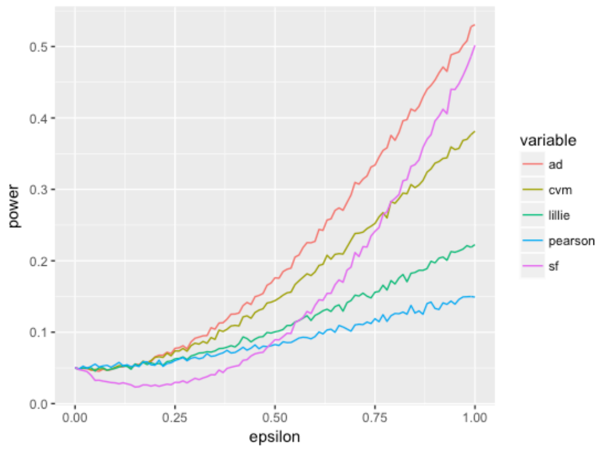

Figure 15: Power function with tranformed population $X = Y + \epsilon \sin(X)$, $n = 100$

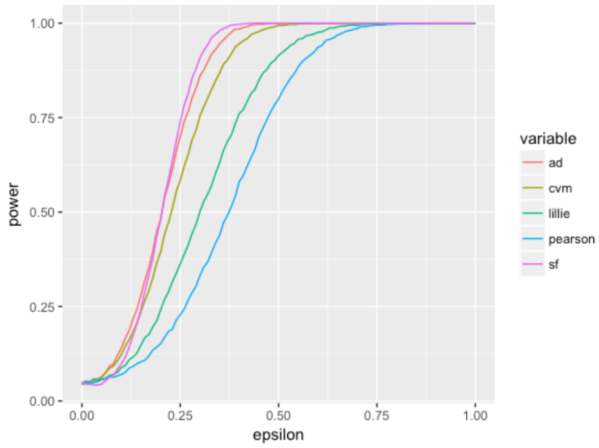

Figure 16: Power function with tranformed population $X = Y + \epsilon \sin(X)$, $n = 1000$

Also, when n becomes larger, all methods become better, when $n = 1000$, SF test becomes the best. When $n = 30$ and $n = 100$, SF test don't perform well and AD test might be better.