Investigation of the Ability of Normality Tests to Detect Issues in Downstream Tests

Ella Li

Oklahoma State University

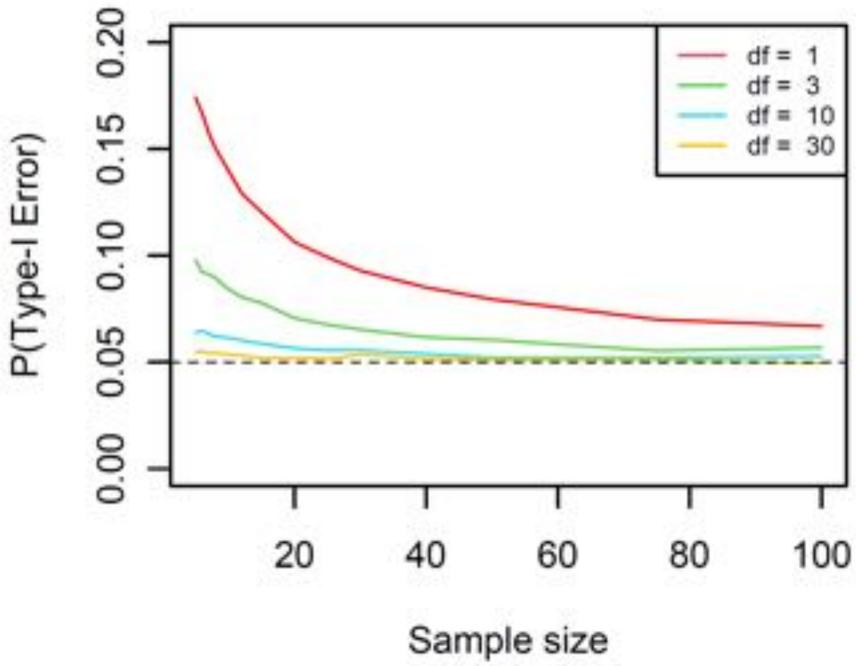
Joint work with Dr. Pratyaydipta Rudra

Problems of Violation of Normality

- Potentially Inflated type-I error rate
- Potential power loss

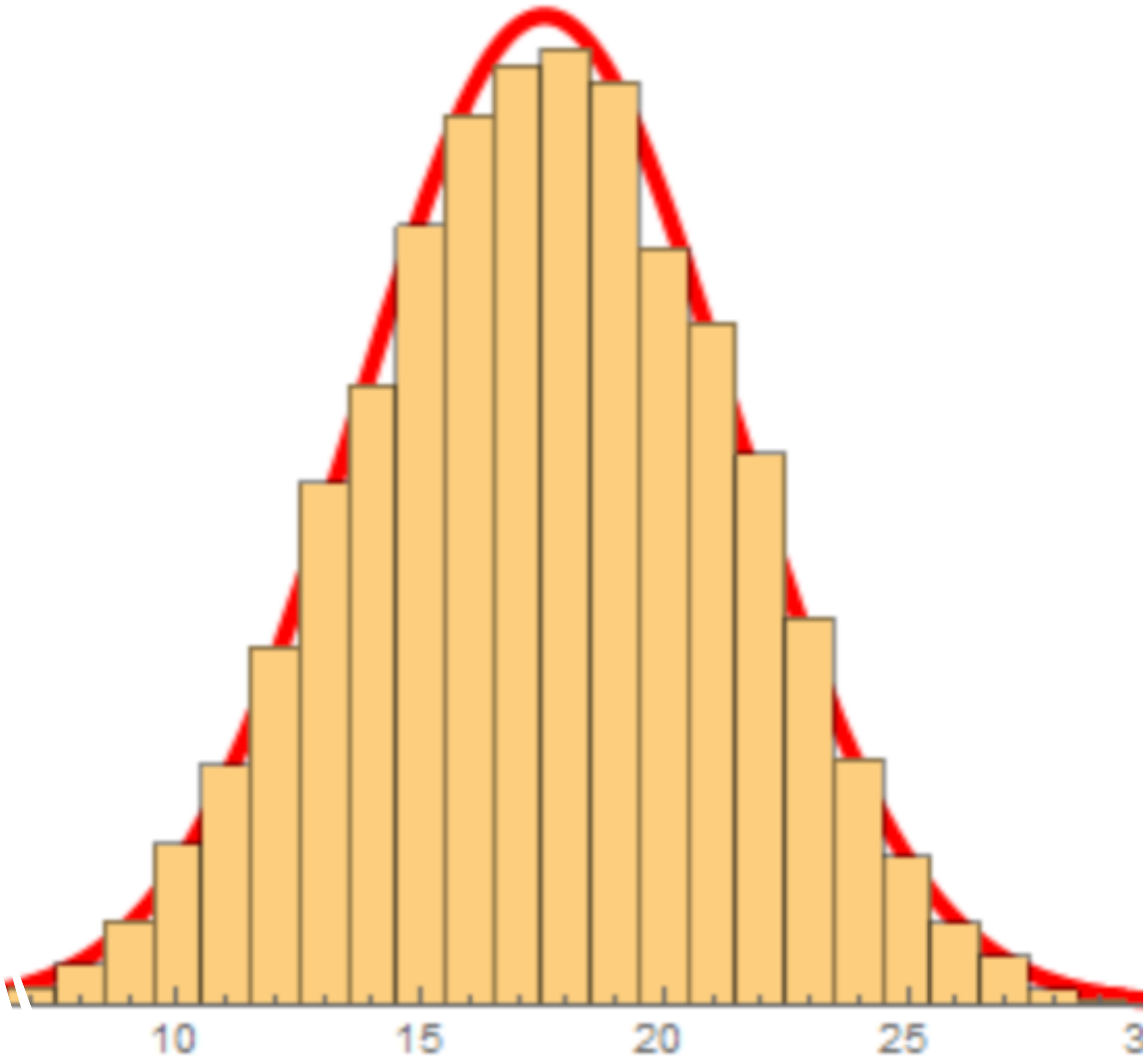
Example:

Type-I error inflation for one-sample t-test when data comes from χ^2 distribution



Normality Tests

- **Shapiro-Wilk Test**
- **D'Agostino-Pearson Test**
- **Kolmogorov-Smirnov Test**
- **Jarque-Bera Test**
- **...**



Shapiro-Wilk Test

- The Shapiro–Wilk test tests the null hypothesis that a sample $x_1, ..., x_n$ came from a normally distributed population

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

D’Agostino-Pearson Test

- a goodness-of-fit measure of departure from normality

- Based on transformations of the sample kurtosis and skewness
- Calculates how far each of these values differs from the value expected with a normal distribution, and computes a single P value from the sum of the squares of these discrepancies
- Has power only against the alternatives that the distribution is skewed and/or kurtic.

Kolmogorov-Smirnov Test

- The KS test can be applied to test whether the data follow any specified distribution, not just the normal distribution

- One-sample Kolmogorov–Smirnov test statistic

$$F_n(x) = \frac{\text{number of (elements in the sample} \leq x)}{n} = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(X_i),$$

where $1_{(-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and equal to 0 otherwise.

Jarque-Bera Test

- The Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

Where S is the sample skewness,

K is the sample kurtosis

Power of Normality tests
for Normal and non-normal distributions

Tests/ Sample Size	T				Beta				Chi-square				Uniform			
	Test Power				Test Power				Test Power				Test Power			
	KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB
5	3.6	3.4	-	-	4.7	5	-	-	4.5	5.2	-	-	3.5	5.2	-	-
6	4.8	4.4	-	-	6.9	6.8	-	-	6.1	6.2	-	-	4.1	6.2	-	-
7	5.1	4.6	-	-	5.6	5.6	-	-	5.6	6.9	-	-	6.1	6.7	-	-
8	4.7	4.7	5.7	0.3	6.7	9.2	8.2	0.2	5.9	6.7	8.7	0.6	5.2	7.3	2.4	0
9	6	6.7	8.1	0.8	7.8	9.2	8.8	1.9	6.4	6.5	7.4	0.8	5.8	7.7	3.3	0.4
10	5	5.7	6.9	1.2	7.2	9.5	9.1	1.4	5.6	7	8.9	1.5	6.6	7.7	2.8	0.2
11	5.8	7	8	1.9	8	9.3	6.7	1.8	6.7	7.5	8.2	1.5	5.9	9.9	1.8	0
12	6.7	5.3	7.2	1.7	7.6	8.7	8	1.8	6.7	7.5	8.4	3.3	6.1	9.9	3	0.2
13	5.5	6.4	7.5	2.3	8.1	9.9	8.5	2.2	6.5	10.6	10.1	4	5.9	11.6	5.4	0.1
14	4.4	5.1	6.1	1.7	9.2	12.1	8	2.8	6.8	9.2	8.3	3.1	8.5	12.8	7.3	0
15	5.1	6.7	7.4	3.6	8.6	10.3	7.9	2.3	6.5	9	9	4.5	8.1	13.8	7.3	0
20	5.4	4.9	5.8	2.5	11.5	16.6	11.7	5.1	7.8	11.6	12.1	6.7	8.9	19.6	14.3	0
25	6.1	6.3	8.8	5.5	12.4	21.9	12.1	5.9	9.8	14	14.3	8.7	10.4	26.9	27.9	0.1
30	5.3	6.3	8.2	5.4	16	29.2	17.4	9.9	9.8	14.8	13.5	9.6	13.1	36	39.2	0
35	5.5	6.1	8.2	5.7	16.9	31.8	17.2	10	11.7	17.1	16.7	12.1	13.8	46.2	53	0.2
40	6	7.7	8.6	7	19.9	41	20.7	12.5	12.1	19.8	16.7	12.9	20	58.3	64.2	0
45	5	5.6	7.7	5.1	21.5	43.9	21.9	12.7	13.9	21.6	17.5	14.3	22.6	65.5	70.4	0
50	4.8	7.1	9.5	7.6	26.8	48.6	22.9	15.2	4	11	12.3	11.5	27.9	75.7	82.4	0
75	5.1	8.1	8.2	8.2	39.2	76.1	40.2	30.1	18.4	35.9	30.8	26.5	41.7	95.1	97.3	7.5
100	5.8	7.8	9.2	8.5	50.4	89	59.2	48.5	22.7	43.7	37.9	34.5	57.3	99.7	99.5	55
150	6	7.9	10.6	12	72.6	99	90.4	86.3	34.4	60.5	54.4	51.6	84.4	100	100	98
175	6.1	8.8	10.7	12	77.2	99.6	96	94.2	39.1	68.9	62.2	61	89.9	100	100	100
200	4.9	8.5	10.4	12	83.1	100	98.7	97.7	44.8	74.6	68.7	66.1	93.7	100	100	100

T				Beta				Chi-square				Uniform			
Test Power				Test Power				Test Power				Test Power			
KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB	KS	SW	DAP	JB
3.6	3.4	-	-	4.7	5	-	-	4.5	5.2	-	-	3.5	5.2	-	-
4.8	4.4	-	-	6.9	6.8	-	-	6.1	6.2	-	-	4.1	6.2	-	-
5.1	4.6	-	-	5.6	5.6	-	-	5.6	6.9	-	-	6.1	6.7	-	-
4.7	4.7	5.7	0.3	6.7	9.2	8.2	0.2	5.9	6.7	8.7	0.6	5.2	7.3	2.4	0
6	6.7	8.1	0.8	7.8	9.2	8.8	1.9	6.4	6.5	7.4	0.8	5.8	7.7	3.3	0.4
5	5.7	6.9	1.2	7.2	9.5	9.1	1.4	5.6	7	8.9	1.5	6.6	7.7	2.8	0.2
5.8	7	8	1.9	8	9.3	6.7	1.8	6.7	7.5	8.2	1.5	5.9	9.9	1.8	0
6.7	5.3	7.2	1.7	7.6	8.7	8	1.8	6.7	7.5	8.4	3.3	6.1	9.9	3	0.2
5.5	6.4	7.5	2.3	8.1	9.9	8.5	2.2	6.5	10.6	10.1	4	5.9	11.6	5.4	0.1
4.4	5.1	6.1	1.7	9.2	12.1	8	2.8	6.8	9.2	8.3	3.1	8.5	12.8	7.3	0
5.1	6.7	7.4	3.6	8.6	10.3	7.9	2.3	6.5	9	9	4.5	8.1	13.8	7.3	0
5.4	4.9	5.8	2.5	11.5	16.6	11.7	5.1	7.8	11.6	12.1	6.7	8.9	19.6	14.3	0
6.1	6.3	8.8	5.5	12.4	21.9	12.1	5.9	9.8	14	14.3	8.7	10.4	26.9	27.9	0.1
5.3	6.3	8.2	5.4	16	29.2	17.4	9.9	9.8	14.8	13.5	9.6	13.1	36	39.2	0
5.5	6.1	8.2	5.7	16.9	31.8	17.2	10	11.7	17.1	16.7	12.1	13.8	46.2	53	0.2
6	7.7	8.6	7	19.9	41	20.7	12.5	12.1	19.8	16.7	12.9	20	58.3	64.2	0
5	5.6	7.7	5.1	21.5	43.9	21.9	12.7	13.9	21.6	17.5	14.3	22.6	65.5	70.4	0
4.8	7.1	9.5	7.6	26.8	48.6	22.9	15.2	4	11	12.3	11.5	27.9	75.7	82.4	0
5.1	8.1	8.2	8.2	39.2	76.1	40.2	30.1	18.4	35.9	30.8	26.5	41.7	95.1	97.3	7.5
5.8	7.8	9.2	8.5	50.4	89	59.2	48.5	22.7	43.7	37.9	34.5	57.3	99.7	99.5	55
6	7.9	10.6	12	72.6	99	90.4	86.3	34.4	60.5	54.4	51.6	84.4	100	100	98
6.1	8.8	10.7	12	77.2	99.6	96	94.2	39.1	68.9	62.2	61	89.9	100	100	100
4.9	8.5	10.4	12	83.1	100	98.7	97.7	44.8	74.6	68.7	66.1	93.7	100	100	100

Reference: Öztuna et al. (2006). *Investigation of four different normality tests in terms of type 1 error rate and power under different distributions*.

Why Normality tests may not be very useful?

- Central Limit Theorem says that as the sample size (n) increases, the sample mean converges to a normal distribution:

$$\bar{X} \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

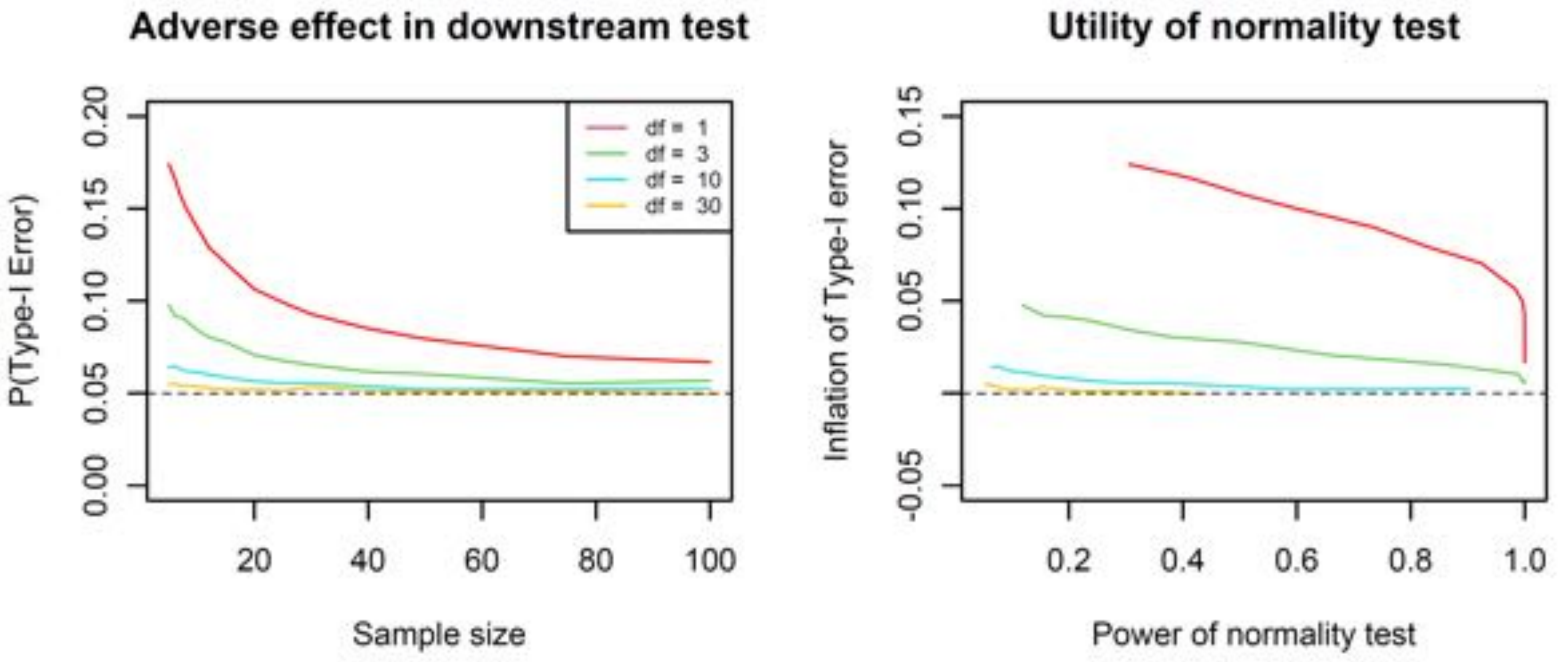
- Therefore, for many downstream methods based on the sample mean, the need to verify normality is less for larger sample sizes.
- However, one criticism of the normality tests is that they show good power only when the sample size is large, when we need them the least.
- We provide a simulation framework to explore the utility of the normality tests in detecting issues (inflated type-I error, loss of power) in downstream tests.

Downstream Tests --- One sample t-test

- We conducted simulations (N = 10,000) to generate data from each of the following distributions with varying sample sizes.
  - Chi-Square Distribution --- Positively Skewed
  - Beta Distribution --- Chose parameters such that negatively skewed
  - Uniform Distribution --- Flat tails
- We estimated the power of Shapiro-Wilk test and the power and P(Type-I error) for a downstream one-sample t-test.

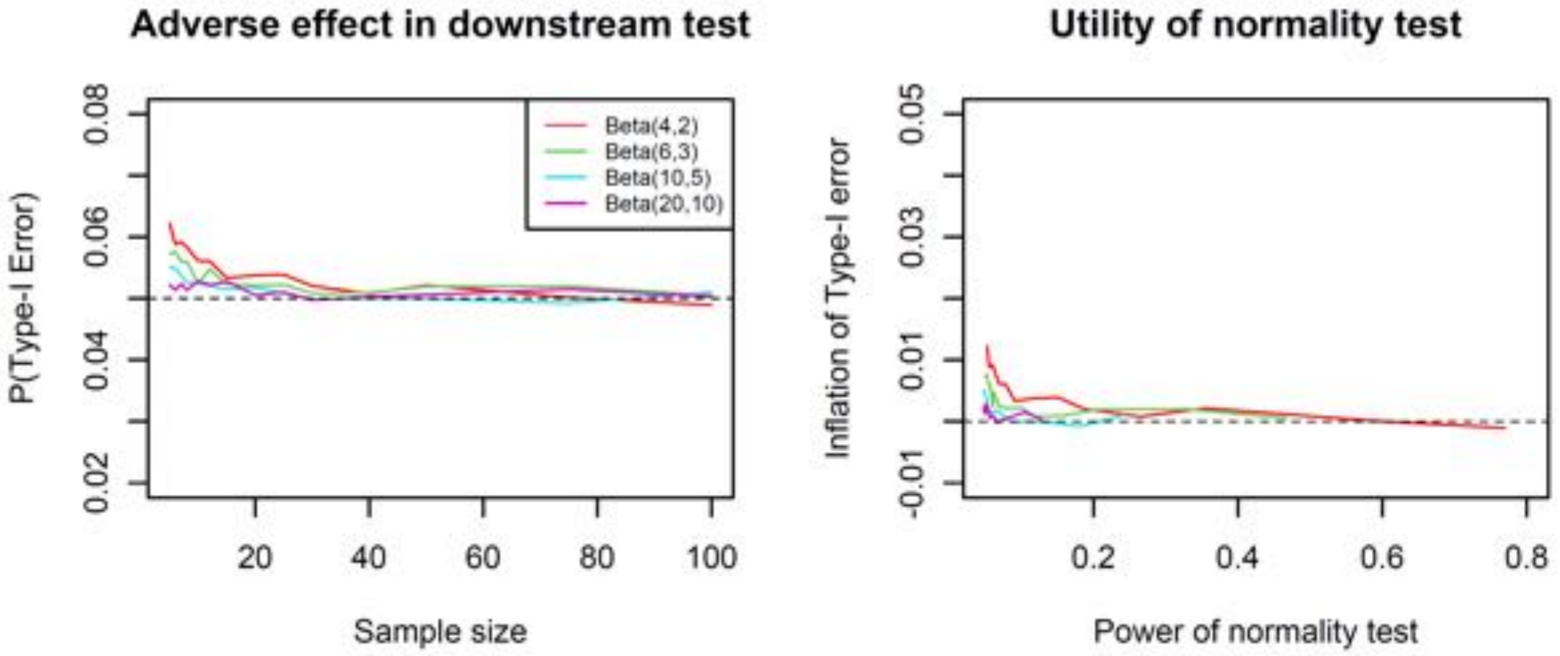
Chi-Square Distribution

- Inflated type-I error rate
- Less problematic for larger samples
- Normality tests somewhat useful for moderate sample sizes



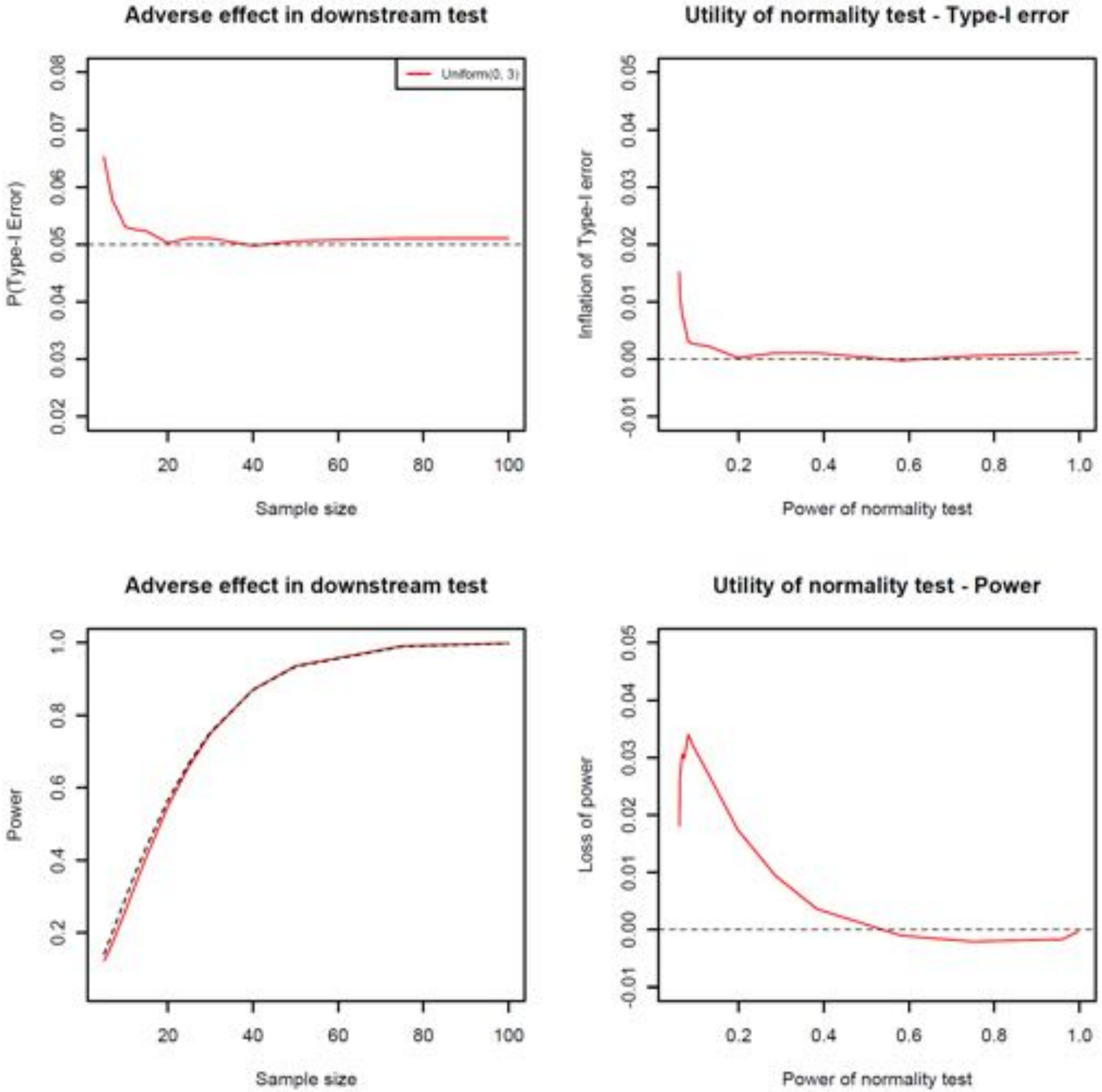
Beta Distribution

- Inflated type-I error rate
- Less problematic for larger samples
- Very little utility of the normality test



Uniform Distribution

- Inflated type-I error rate
- Also loss of power
- Both are less problematic for larger samples
- Very little utility of the normality test



Possible Solutions

- Use robust statistical methods that do not use parametric assumptions
- Use normality tests that are less problematic

Future Work

- Explore other kinds of departure from normality.
- Explore other downstream tests.
- Explore the performance of other normality tests.
- Find what normality tests are less problematic and more useful for a given downstream test.

THANK YOU FOR LISTENING!