2018 AAPM: Normal and non-normal distributions: Why understanding distributions are important when designing experiments and analyzing data

Statistical Failings that Keep Us All in the Dark

Normal and non-normal distributions: Why understanding distributions are important when designing experiments and analyzing data

Jenghwa Chang, Ph.D.[1,2]
[1]Department of Radiation Medicine, Northwell Health
[2]Hofstra Northwell School of Medicine at Hofstra University

Northwell Health

Conflict of Interest Disclosure

- I have no conflict of interest to disclosure.

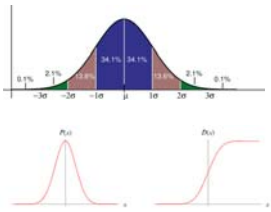
Outline

1. Why is the normal distribution so important?
2. How to tell if your data is normally distributed?
3. What to do if your data is NOT normally distributed?
4. Conclusions

Why is normal distribution so important?

Normal distribution plays a very predominant role in statistics

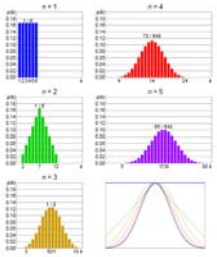
- Nearly normal distribution are encountered quite frequently in nature
- Sampling distributions based on a parent normal distribution are fairly manageable analytically.
- Distributions of functions of sample observations (i.e., statistics) is easier



- $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2},\ \mu_X = \mu,\ \sigma_X^2 = \sigma^2$
- $F_X(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(\tau-\mu)^2/2\sigma^2} d\tau = \frac{1}{2} + \mathrm{erf}\left(\frac{x-\mu}{\sigma}\right)$

Central limit theorem:

- Liapounov central limit theory: when many small independent, non-identically distributed random variables are added, their sum tends toward a normal distribution regardless of the form of the original density functions of the individual random variables.
- Therefore, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ approximate a normal distribution for large n (e.g., $n > 30$) even if individual observation X_i is not normally distributed.



Sample means of *n* fair 6-sided dice show their convergence to a normal distribution with increasing *n*.
https://en.wikipedia.org/wiki/Central_limit_theorem

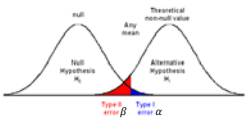
Statistics for normal distribution

- Sample mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample variance: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- χ^2statistic: $U = \frac{(n-1)S^2}{\sigma^2}$ a statistical distance (e.g., random organ motion); can be used for hypothesis testing of similarity
- Z statistic: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$
- t statistic: $T = \frac{Z}{U} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ For hypothesis testing of the mean
- F statistic: $F = \frac{U_1/n_1}{U_2/n_2}$

Hypothesis testing of the mean

- Purpose: confirm if average different between a new procedure and the current procedure is μ_0
- $\mu_0 = 0$ simply tests if the new is different from the current
- Null hypothesis (H_0): the average different is μ_0, and
- Alternative hypothesis (H_1): the average different is not μ_0

	Decision	
	Accept H_0	Reject H_0
H_0 (true)	Correct decision	Type I error (α error)
H_0 (false)	Type II error (β error)	Correct decision



Statistics (z, t, F) for testing mean are like signal to noise ratio

- $X_i, i = 1,2,\ldots n$: random selection of size n from a normal population
- Signal: $\bar{X} - \mu_0$, magnitude of difference
- Noise: $\sigma/\sqrt{n}$, variation of smapled data
- H_1 will be accepted if $|\bar{X} - \mu_0| \gg \sigma/\sqrt{n}$, which requires
  - $|\bar{X} - \mu_0| \gg 0$ or the new procedure is really different
  - $\sigma/\sqrt{n}$ is small or a large sample size (i.e., statistical power).
- Need to define when the difference is statistically significant.

Magnitude of difference

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Variation of sampled data

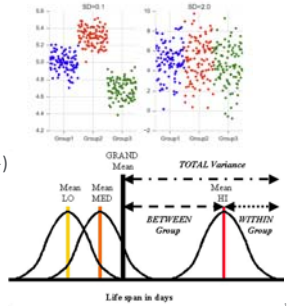
This is similar to SNR!

t statistics is similar to Z statistics except σ is unknown

- $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ standardizes $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ with known μ, σ^2
- $T = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{S^2}/\sigma^2} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ statistics achieve the similar purpose using sample variance S^2 to estimate σ^2.
- A statistics cannot have unknowns:
  - σ^2 is unknown but is cancelled out in Tstatistics
  - μ is assumed a hypothetic value μ_0 although its really value is unknown.

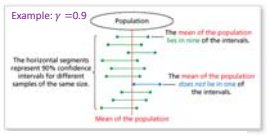
F statistic is similar to t statistic but for > two samples

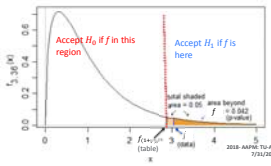
- k independent samples $X_1, X_2, \ldots, X_k$ with $n_1, n_2, \ldots, n_k$ observations
- Want to test at least one $\mu_j \neq \mu$
- Cannot use T-test as the type I error accumulates quickly.
- Can compare variance between group (SS_B) and variance within groups (SS_W)
  - SS: square sum
  - Which distribution is for SS_B/SS_W?
  - $f = \frac{SS_B/(k-1)}{SS_W/(n-k)}$, $n = n_1 + \cdots + n_k$

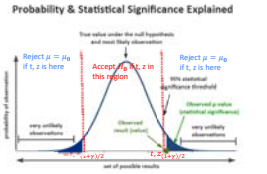


Hypothesis test of the mean using Z, t or F test

- Calculate z, t, f: ratio of difference (mean of diff, diff of mean, SS of diff) to dispersion of data
- Ratio needs to be large enough to reject H_0
- For a γ, identify critical value $z, t, f_{(1+\gamma)/2}$
- Do not reject H_0 if $|z, t, f| < z, t, f_{(1+\gamma)/2}$, μ_0 in (T, Z_1, T, Z_2), or $p < \alpha = 1 - \gamma$





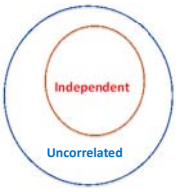


Checking Assumptions

- Independence
- Normality
- Homogeneity of variance
- Robustness

Assumption of independence is an unforgiving assumption

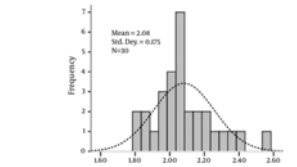
- Assumption of independence means that the data are not connected
- Two independent assumptions:
  - The **observations between groups**.
  - The observations **within each group** .
- Difficult use the study's sample data to test the validity of this prerequisite condition.
- Make sure your data is independent *while you are collecting it*.

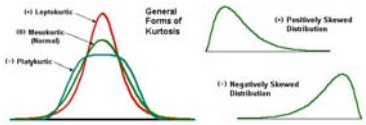


Correlated⇒
Dependent

Test of Normality

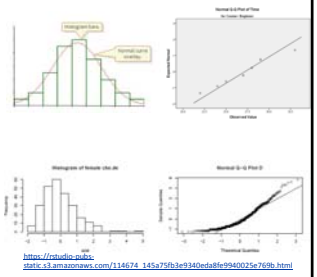
- Histogram: from sampled data
- Normal Curve with
  - Mean is sample mean
  - S.D. is squared root of sample variance
- Ways to test normality:
  - Eyeballing
  - Descriptive statistics
  - Chi-square goodness-of-fit test
  - Software package using more advanced goodness-of-fit tests





Eyeballing normality

- The simplest way to check normality.
- Visually check if the normal curve fit the histogram
- Normal Q-Q (quantile-quantile) plot:
  - The points closely follow the line for normal distribution.
  - Any deviations from normality leads to deviations of these points from the line.
- If undetermined, use descriptive statistics or normality test.



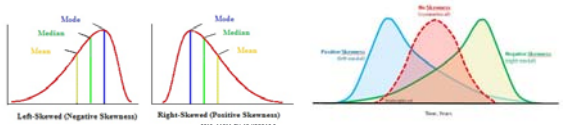
https://rstudio-pubs-static.s3.amazonaws.com/114674_145a75fb3e9340eda8fe9940025e769b.html

Descriptive Statistics

- Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ shows the central tendency
- Sample variance $\mathcal{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ measures the dispersion
- Sample central moment $m_r = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^r$
  - $m_1 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}) = 0$
  - $m_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ Note: $m_2 \neq \mathcal{S}^2$
  - $m_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3$ is the skewness
  - $m_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4$ is the Kurtosis

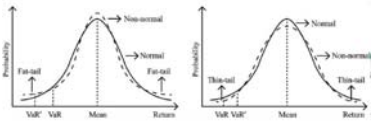
Test of normality using coefficients ofskewness

- Coefficient of skewness $= \frac{m_3}{\mathcal{S}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{3/2}}$ is a measure of symmetry
- Ideally, the skewness is 0 for normal distribution.
- Considered deviated from the normal distribution if the value is greater than ± 1.96.



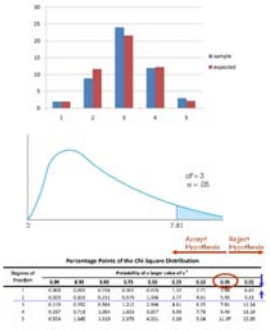
Test of normality using coefficients of excess kurtosis

- Coefficient of excess kurtosis= $\frac{m_4}{(m_2)^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)^2} - 3$, a measure of "tailedness" or "sharpness" of the shape of the probability distribution.
- Ideally, both are 0 for the normal distribution.
- Considered deviated from the normal distribution if the value is greater than ± 1.96



Chi-square test for goodness of fit

- $\chi^2 = \sum_{i=1}^{k}\frac{(f_i-e_i)^2}{e_i}$ where
  - f_i= observed frequency for bin i
  - e_i= expected frequency for bin i
  - k = number of categories
  - m = # of parameters, 2 for normal distribution
  - $DF = k - m - 1$
- H_0 : The distribution is normal
- H_1 : The distribution is not normal
- Critical value $P\left[X^2 < \chi^2_{(1+\gamma)/2}\right] = \gamma$
- Chi-square goodness-of-fit test:
  - If $\chi^2 < \chi^2_{(1+\gamma)/2}$, do not reject normal
  - Otherwise, reject normal



Available tests of normality:

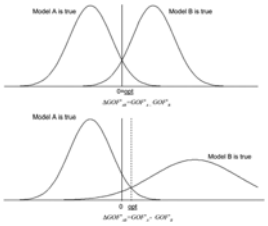
- Kolmogorov-Smirnov (K-S) test
- Lilliefors corrected K-S test
- Shapiro-Wilk test
- Anderson-Darling test
- Cramer-von Mises test
- D'Agostino skewness test
- Anscombe-Glynn kurtosis test
- D'Agostino-Pearson omnibus test
- Jarque-Bera test

The homogeneity of variance assumption for t and ANOVA test

- Assumption of homogeneity of variance: the populations from which the samples are obtained for testing have equal variances
- The standard deviation (σ) is unknown but cancelled out in T and F statistics

$$T = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{s^2/\sigma^2}} = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

- If violated, cannot confidently accept/reject the H_0.



http://goodsciencebadscience.nl/?p=101

Pooling of variance in unpaired t and F tests

- Two independent samples X_1 and X_2 with means μ_1, μ_2, unknown but equal σ, and n_1 and n_2 observations.
- $t = (\bar{x}_1-\bar{x}_2)/\sqrt{s^2/n_1 + s^2/n_2}$
- $\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1,i}$, $\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2,i}$, $s_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1,i}-\bar{x}_1)^2}{n_1-1}$, $s_2^2 = \frac{\sum_{i=1}^{n_1}(x_{2,i}-\bar{x}_2)^2}{n_2-1}$
- $s^2 = \frac{\sum_{i=1}^{n_1}(x_{1,i}-\bar{x}_1)^2+\sum_{i=1}^{n_2}(x_{2,i}-\bar{x}_2)^2}{n_1+n_2-2} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ which is the weighted sum of s_1^2, s_2^2
- For F test, s^2 is the weighted sum of s_1^2, s_2^2,..., s_k^2

$$s^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2+\cdots+(n_k-1)s_k^2}{n_1+n_2+\cdots+n_k-k}$$

Problem with pooled variance

- Estimating σ with s^2 might cause problems if
  - Unequal sample size: $n_1 \ll\gg n_2$
  - Unequal variance: $\sigma_1 \neq \sigma_2 \neq \sigma$
- Example:

			Pooled
	1	9	
Size	500	50	2.86
Size	50	500	8.59
Size	50	50	6.40

- $t = (\bar{x}_1-\bar{x}_2)/s\sqrt{1/n_1 + 1/n_2}$, $t \uparrow$ if $s \downarrow$ making t more likely $> t_{(1+\gamma)/2}$ and reject H_0.
- If the smallest sample size is the one with highest variance, the test will have increase the chance of rejecting H_0, or the type I error is inflated.

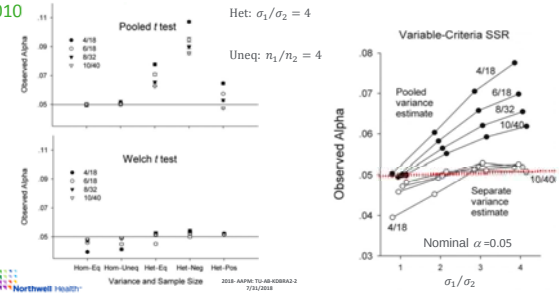
How to avoid inflated type I error when $n_1 \ll\gg n_2$

- Resample the sample with a larger sample size making $n_1 = n_2$
- User unpooled (e.g. Weltch's) t test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ where } s_1^2 = \frac{\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2}{n_1 - 1},\ s_2^2 = \frac{\sum_{i=1}^{n_1}(x_{2,i} - \bar{x}_2)^2}{n_2 - 1}$$

$$DF = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}$$

Inflation of type I error due to uneven sample size, unequal variance and correlated samples (Fitts DA. Behav Res Methods 2010



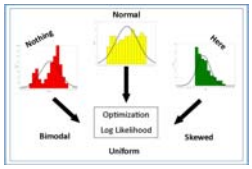
Robustness of t test

- When $n_1 \approx n_2$, violating the assumption of homogeneity of variance produces very small effects—the nominal value of $\alpha = 0.05$ is most likely within ± 0.02 of true α once $\sigma_1/\sigma_2 \leq 3$.
- Paired t-tests are generally robust as $n_1 = n_2$.
- Most unpaired t-tests are robust when $n_1 \approx n_2$.
- Not sensitive to normality one n is large (e.g., $n > 30$, larger if the distribution is skewed) due to central limit theorem.
- Unequal sample sizes do not affect α as long as $\sigma_1 = \sigma_2$
- Separate-variance estimates stabilize α when $\sigma_1 \neq \sigma_2$
- A combination of widely heterogeneous variances and unequal sample sizes should be avoided.

Non-parametric methods

Non-parametric methods

- Populations being sampled are not normally distributed.
- Sample sizes are small so assessing normality is not possible ($n_i \leq 20$).
- No assumption can be made about the population distribution
- When there are outliers
- Would like to test the difference in median instead of mean



https://www.jokejive.com/topic/statistics

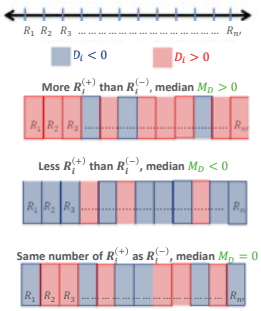
Order and rand order statistics

- Order statistics: Let $X_1, X_2, \ldots, X_n$ denotes a random sample of size n. Then $Y_1 \leq Y_2 \leq \cdots \leq Y_n$, which Y_i are the X_i arranged in order of increasing magnitudes and are defined to be the order statistics corresponding to the random sample $X_1, X_2, \ldots, X_n$.
- Rank Order statistics is the index of Y_i

Rank	1	2	3	4	5
Order	Y_1	Y_2	Y_3	Y_4	Y_5
R.S.	X_5	X_1	X_4	X_2	X_3

Wilcoxin W statistic

- Paird samples X_1, X_2 with sample size n
- Calculate difference $D_i = X_{1,i} - X_{2,i}$
- Exclude any differences which are zero, yielding n' nonzero D_i.
- Ignoring signs, rank the non-zero $|D_i|$.
- Assign ranks R_i from 1 to n' to each of $|D_i|$ with smallest $|D_i|$ getting rank 1.
- Reassign signs (+ or -) to n' ranks R_i.
- Wilcoxon W statistic $W = \sum_{i=1}^{n'} R_i^{(+)}$ is the sum of the positive ranks.



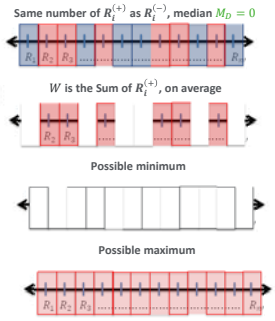
Mean and variance of W statistic if $M_D = 0$

- Wilcoxon W statistic $W = \sum_{i=1}^{n'} R_i^{(+)}$
- If $M_D = 0$, half of the ranks are for $R_i^{(+)}$ and the other half for $R_i^{(-)}$, so the expect value of W is:

$$\mu_W = \frac{1}{2}\sum_{i=1}^{n'} i = \frac{1}{2}\frac{n'(n'+1)}{2} = \frac{n'(n'+1)}{4}$$

$$\sigma_W = \sqrt{\frac{n'(n'+1)(2n'+1)}{24}}$$

- Possible min: $w_{\min} = 0$
- Possible max: $w_{\max} = \sum_{i=1}^{n'} i = \frac{n'(n'+1)}{2}$

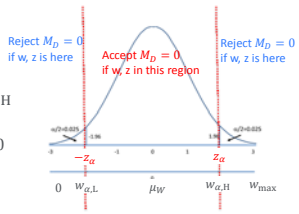


Wilcoxin Signed rank Test tests median difference (M_D) of dependent samples

- Null hypothesis H_0: $M_D = 0$
- Alternative hypothesis H_1: $M_D \neq 0$
- Calculate statistic $w = \sum_{i=1}^{n'} R_i^{(+)}$ is the sum of the positive ranks.
- For a α, identify critical values $w_{\alpha,\mathrm{L}}$, $w_{\alpha,\mathrm{H}}$
- Wilcoxin Signed rank test:
  - If $w_{\alpha,\mathrm{L}} < w < w_{\alpha,\mathrm{H}}$, do not reject $M_D = 0$
  - Otherwise, reject $M_D = 0$
- If $n' > 20$, can use Z test instead

$$\mathrm{Z} = \frac{W - \mu_W}{\sigma_W}$$

Example: $\alpha = 0.05$, $z_{0.95} = 1.960$



Parametric and Nonparametric Tests

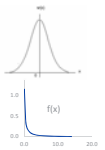
Sample Types	Parametric Tests	Nonparametric Tests
Dependent samples	Paired t-Test	Sign Test/ Wilcoxon Signed-Rank Test
2-independent samples	Two-Sample t-Test	Mann-Whitney Test/ Wilcoxon Rank Sum Test
k-independent samples	One way ANOVA	Kruskal-Wallis Test

Robustness of Non-parametric method
(Zimmerman *Psicológica (2004), 25,* 103-133.)

- For a wide variety of non-normal distributions, especially skewed distributions, the Type I error probabilities of both the t test and the Wilcoxon-Mann-Whitney test are substantially inflated by heterogeneous variances, even when sample sizes are equal.

$N_1 = N_2 = 25$

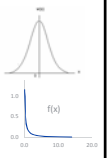
distribution	α	$\sigma_1/\sigma_2 = 1$: t	t on ranks	W	$\sigma_1/\sigma_2 = 2$: t	t on ranks	W	$\sigma_1/\sigma_2 = 3$: t	t on ranks	W
normal	.01	.011	.011	.009	.010	.014	.012	.012	.018	.015
	.05	.051	.050	.053	.052	.059	.058	.054	.068	.068
	.10	.100	.099	.102	.103	.113	.110	.105	.127	.126
Weibull shape parameter .5 scale parameter 1	.01	.003	.010	.009	.035	.515	.494	.067	.624	.609
	.05	.035	.049	.050	.094	.722	.723	.126	.803	.808
	.10	.090	.099	.098	.146	.812	.811	.176	.875	.876



Robustness of Non-parametric method
(Zimmerman *Psicológica (2004), 25,* 103-133.)

- The inflation of Type I error of the Wilcoxon-Mann-Whitney test increases with the sample size for skewed distribution like Weibull. Samples from symmetric distributions, are not affected in this way.

distribution	σ_1/σ_2	t: N_1,N_2 20	50	80	t on ranks: N_1,N_2 20	50	80	W: N_1,N_2 20	50	80
normal	1.25	.050	.050	.049	.050	.051	.051	.050	.050	.050
	2.00	.054	.050	.053	.058	.059	.059	.057	.057	.058
Weibull shape parameter .5 scale parameter 1	1.25	.043	.045	.045	.321	.651	.836	.319	.647	.835
	2.00	.097	.076	.069	.630	.946	.994	.628	.945	.994



Conclusions

- Normal (or Gaussian) distribution is a widely used probability distribution in natural and social science for describing random events.
  - The central limit theory: the sum of multiple independent random variables approximates the normal distribution if the sample size is sufficiently large
  - Well established (Z, χ^2, t, F) statistics for statistical inference
- Assumption check is critical
  - Normality: not a issue for large n (e.g., $n > 30$, larger is the distribution is skewed) due to central limit theory.
  - Homogeneity: up to three-time difference is ok if sample size is balanced.
- Balanced sample size is preferred: if the sample size is not balanced, use unpooled variance estimate to minimize inflation of type I error.
- When normality is in doubt and sample size is small, use non-parametric method instead.

Thank You!