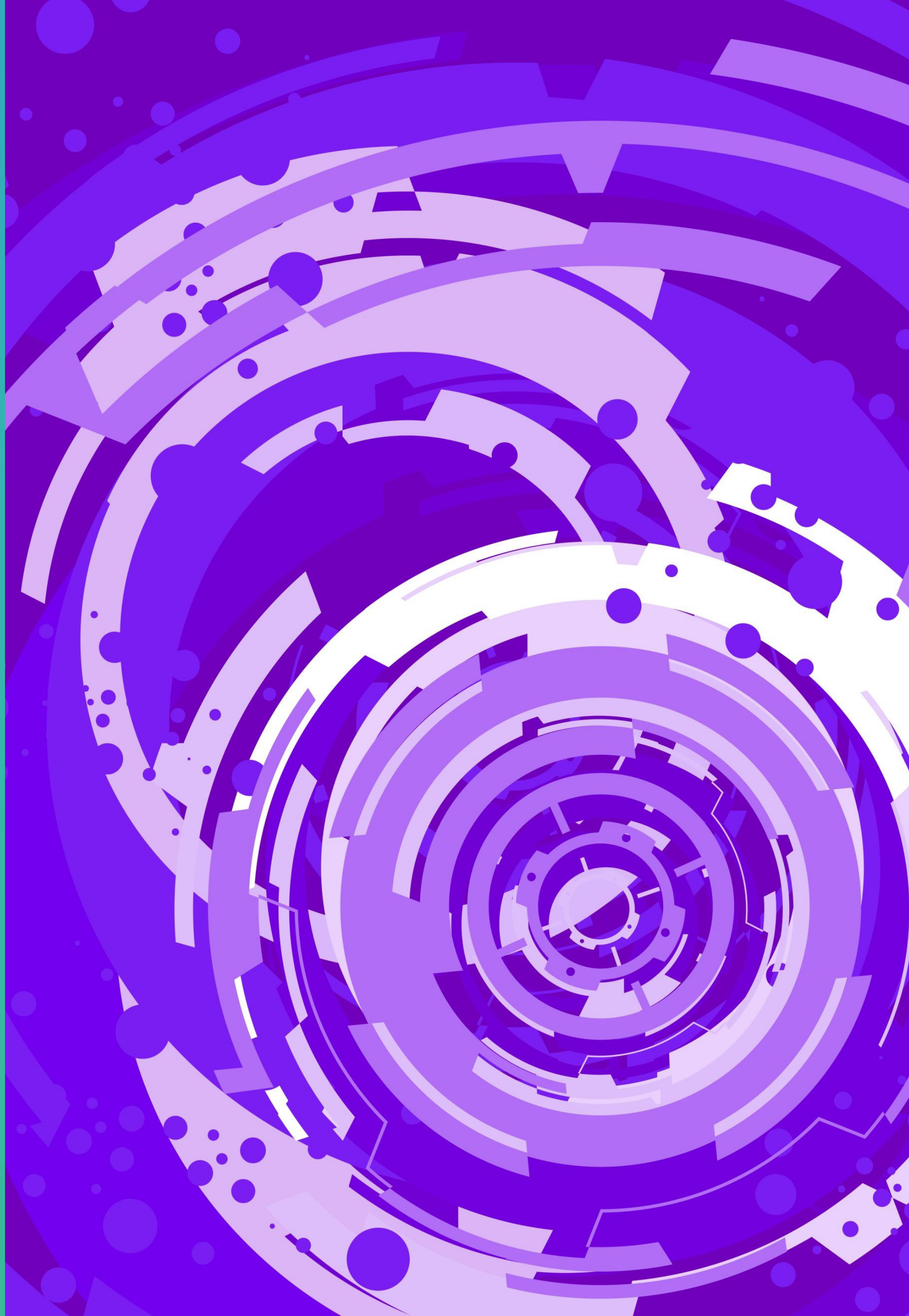
URBAN GEOGRAPHIC INFORMATION SYSTEM

Python Statistics & Regression

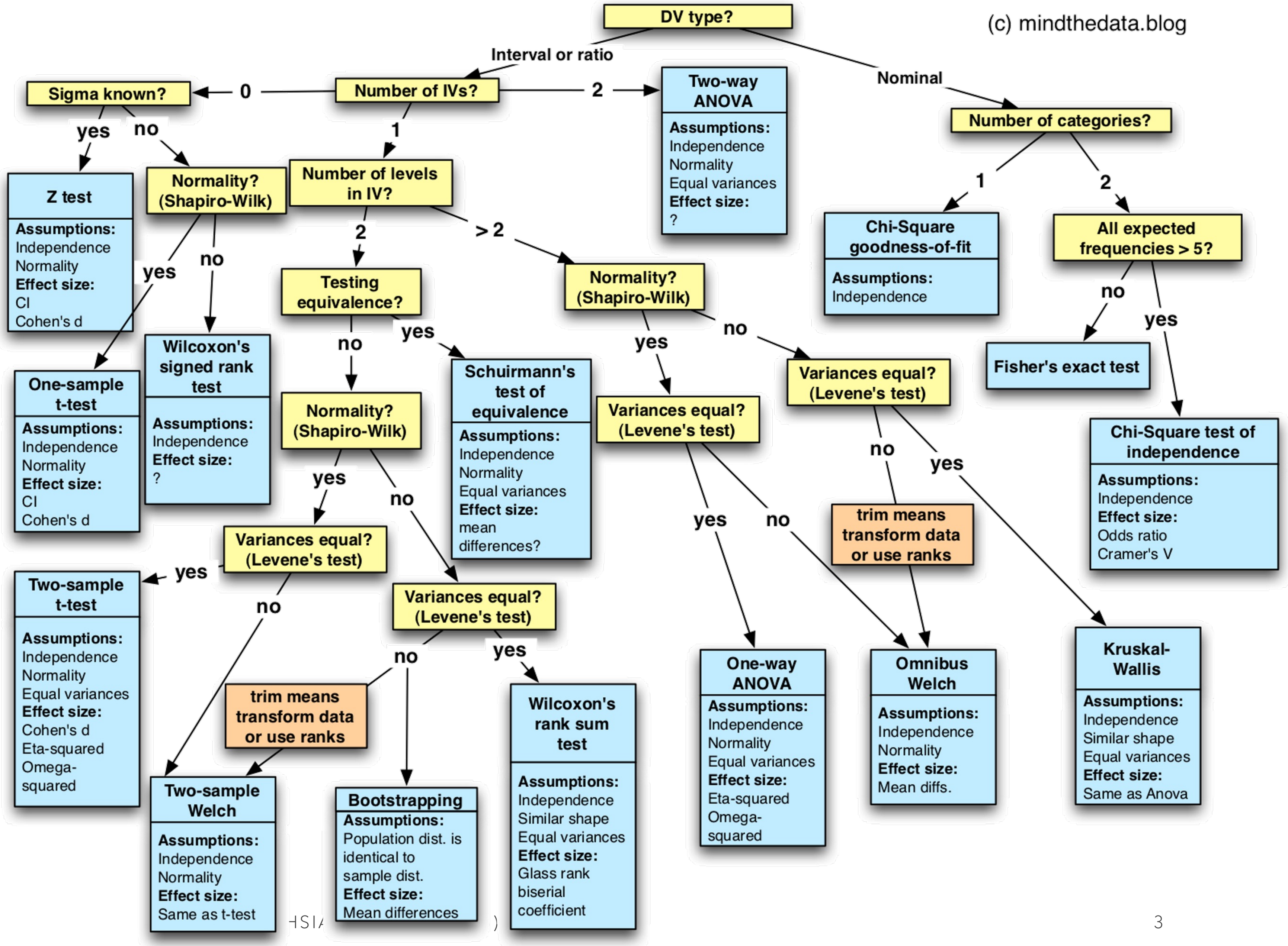
Chun-Hsiang Chan
Department of Geography,
National Taiwan Normal University



Outline

- Statistical Analysis
- Normality Test
- Inferential Statistics
- Correlation Analysis
- Scale Problem
- Regression

Statistical Analysis Road Map



Statistical Analysis Road Map

Table 1 Choice of statistical test from paired or matched observation

Variable	Test
Nominal	McNemar's Test
Ordinal (ordered categories)	Wilcoxon
Quantitative (discrete or non-normal)	Wilcoxon
Quantitative (normal)	Paired t-test

Table 2 Parametric and nonparametric tests for comparing two or more groups

Parametric Test	Situation	Nonparametric Test
t-test	Two independent population	Wilconxon rank sum test
t-test		Mann-Whitney U test
One way analysis of variance	Three or more populations	Kruskal Wallis test
Paired t-test	Paired population	Sign test
		Wilconxon rank sign test
Pearson correlation	Correlation	Spearman correlation

Source: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests

Statistical Analysis Road Map

Table 3 Choice of statistical test for independent observations

		Outcome variable					
Input variable		**Nominal**	**Categorical (>2)**	**Ordinal**	**Quantitative Discrete**	**Quantitative Non-normal**	**Quantitative Normal**
	Nominal	χ^2 or Fisher's	χ^2	χ^2-trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank[a]	T-test
	Categorical (>2)	χ^2	χ^2	Kruskal-Wallis[b]	Kruskal-Wallis[b]	Kruskal-Wallis[b]	ANOVA[c]
	Ordinal	χ^2-trend or Mann-Whitney	[e]	Spearman rank	Spearman rank	Spearman rank	Spearman rank or Linear regression[d]
	Quantitative Discrete	Logistic regression	[e]	[e]	Spearman rank	Spearman rank	Spearman rank or Linear regression[d]
	Quantitative Non-normal	Logistic regression	[e]	[e]	[e]	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and Linear regression
	Quantitative Normal	Logistic regression	[e]	[e]	[e]	Linear regression[d]	Pearson or Linear regression

[a] If data are censored. [b] The Kruskal-Wallis test is used for comparing ordinal or non-Normal variables for more than two groups, and is a generalisation of the Mann-Whitney U test. [c] Analysis of variance is a general technique, and one version (one way analysis of variance) is used to compare Normally distributed variables for more than two groups, and is the parametric equivalent of the Kruskal-Wallistest. [d] If the outcome variable is the dependent variable, then provided the residuals (the differences between the observed values and the predicted responses from regression) are plausibly Normally distributed, then the distribution of the independent variable is not important. [e] There are a number of more advanced techniques, such as Poisson regression, for dealing with these situations. However, they require certain assumptions and it is often easier to either dichotomise the outcome variable or treat it as continuous.

Source: https://www.healthknowledge.org.uk/public-health-textbook/research-methods/1b-statistical-methods/parametric-nonparametric-tests

Normality Test :: Shapiro–Wilk test

- The Shapiro–Wilk test is a test of normality in frequentist statistics. **The Shapiro–Wilk test tests the null hypothesis that a sample $x_1, \dots, x_n$ came from a normally distributed population.** The test statistic is

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

(with parentheses enclosing the subscript index i; not to be confused with x_i) is the ith order statistic, i.e., the ith-smallest number in the sample; $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n), 8 \le n \le 50$ is the sample mean.

Shapiro–Wilk test

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

$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ is the sample mean.

- The coefficient a_i are given by: $(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{C}$, where C is a vector norm: $C = \|V^T m\| = \sqrt{m^T V^{-1} V^{-1} m}$ and the vector m, $m = (m_1, \ldots, m_n)^T$ is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution; finally, V is the covariance matrix of those normal order statistics.

Source: https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
Source: http://blog.excelmasterseries.com/2015/05/how-to-create-completely-automated_4.html

Shapiro–Wilk test

$\alpha = 0.05$	$n = 15$							Data Pairs					
Raw Data		Sorted Data		a value		Upper value		Lower value		Difference	a*Difference		
1	20	1	18	a1	0.5150	#15	22	#01	18	4	2.0600		
2	19	2	18	a2	0.3306	#14	21	#02	18	3	0.9918		
3	18	3	18	a3	0.2495	#13	21	#03	18	3	0.7485		
4	19	4	18	a4	0.1878	#12	21	#04	18	3	0.5634		
5	22	5	19	a5	0.1353	#11	20	#05	19	1	0.1353		
6	18	6	19	a6	0.0880	#10	20	#06	19	1	0.0880		
7	21	7	19	a7	0.0433	#09	19	#07	19	0	0.0000		
8	19	8	19										
9	21	9	19										
10	18	10	20										
11	18	11	20										
12	19	12	21										
13	20	13	21										
14	21	14	21										
15	19	15	22										

$\sum_{i=1}^{n} a_i x_{(i)}$	4.59	$\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2$	21.0681
$\sum_{i=1}^{n} (x_i - \bar{x})^2$	23.733	W	0.886541
		W critical	0.881

Kolmogorov-Smirnov Test

- The Kolmogorov–Smirnov test (K-S test or KS test) is a **nonparametric test** of the equality of continuous (or discontinuous), one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test), or to compare two samples (two-sample K–S test), where n is larger than 50.
- **The null distribution of this statistic is calculated under the null hypothesis that the sample is drawn from the reference distribution (in the one-sample case) or that the samples are drawn from the same distribution (in the two-sample case).**

Kolmogorov-Smirnov Test

- The two-sample K–S test is one of the most useful and general **nonparametric** methods for comparing two samples, as it is **sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples**.

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

- The Kolmogorov-Smirnov statistic for a given cumulative distribution function $F(x)$ is

$$D_n = \sup_x |F_n(x) - F(x)|,$$

$$where \sup_x is\ the\ supermum\ of\ the\ set\ of\ distances.$$

- Intuitively, the statistic takes the largest absolute difference between the two distribution functions across all x values.

Code :: Normality Test

```
# normality test
# Kolmogorov-Smirnov Test
from statsmodels.stats.diagnostic import kstest_normal
ks, pval = kstest_normal(df['avganncount'], dist='norm', pvalmethod='table')
print('Kolmogorov-Smirnov Test:',ks, pval)
# Shapiro-Wilk Test
from scipy.stats import shapiro
sw, pval = shapiro(df['avganncount'])
print('Shapiro-Wilk Test:',sw, pval)
# D'Agostino and Pearson's: test that combines skew and kurtosis to produce an omnibus test of normality
from scipy.stats import normaltest
nor, pval = normaltest(df['avganncount'])
print('D'Agostino and Pearson's:',nor, pval)
```

F Test

- The definitional equation of sample variance is

$$s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

- The fundamental technique is a partitioning of the total sum of squares SS into components related to the effects used in the model.

$$SS_{Total} = SS_{treatments} + SS_{Error}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\ldots\right)^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}\ldots)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y_{i\cdot}}\right)^2$$

F Test

- The F-test is used for comparing the factors of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F test statistic.

$$F = \frac{variance\ between\ treatments}{variance\ within\ treatments}$$

$$F = \frac{MS_{Treatments}}{MS_{Error}} = \frac{\frac{SS_{Treatments}}{I-1}}{\frac{SS_{Error}}{n_T-1}}$$

where MS is mean square, I is the number of treatments and n_T is the total number of cases.

T Test

- In addition to one-sample t-test, there are three types of t-test, including paired t-test, and two-sample independent t-test (assume that the variance of two samples or populations are [not] equal).
- In the following slides, we will give some examples to show their differences.

Question X

How do we determine whether the variances between two samples or populations are equal?

Paired T Test

- If the two samples or populations are from matched or paired sources or a replicated measurement, you must select a paired t-test.

$$t = \frac{\overline{X_D} - \mu_0}{\frac{s_D}{\sqrt{n}}}$$

$\overline{X_D}$ *and* s_D are the average and standard deviation of the differences between all pairs, the constant μ_0 is zero if we want to test whether the average of the difference is significantly different, and n is the number of pairs.

Source: https://en.wikipedia.org/wiki/Student%27s_t-test
Source: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/
Source: https://www.omnicalculator.com/statistics/t-test#p-value-from-t-test

Two-sample Independent T-test

- If the variance of two samples or populations are equal (or very similar).

$$t = \frac{\overline{x_1} - \overline{x_2} - \mu_0}{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

s_p is the pooled standard deviation, defined by

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Source: https://en.wikipedia.org/wiki/Student%27s_t-test
Source: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/
Source: https://www.omnicalculator.com/statistics/t-test#p-value-from-t-test

Two-sample Independent T-test

- If the variance of two samples or populations is **unequal** (or very similar), refer to Welch's t-test.

$$t = \frac{\overline{x_1} - \overline{x_2} - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Source: https://en.wikipedia.org/wiki/Student%27s_t-test
Source: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/
Source: https://www.omnicalculator.com/statistics/t-test#p-value-from-t-test

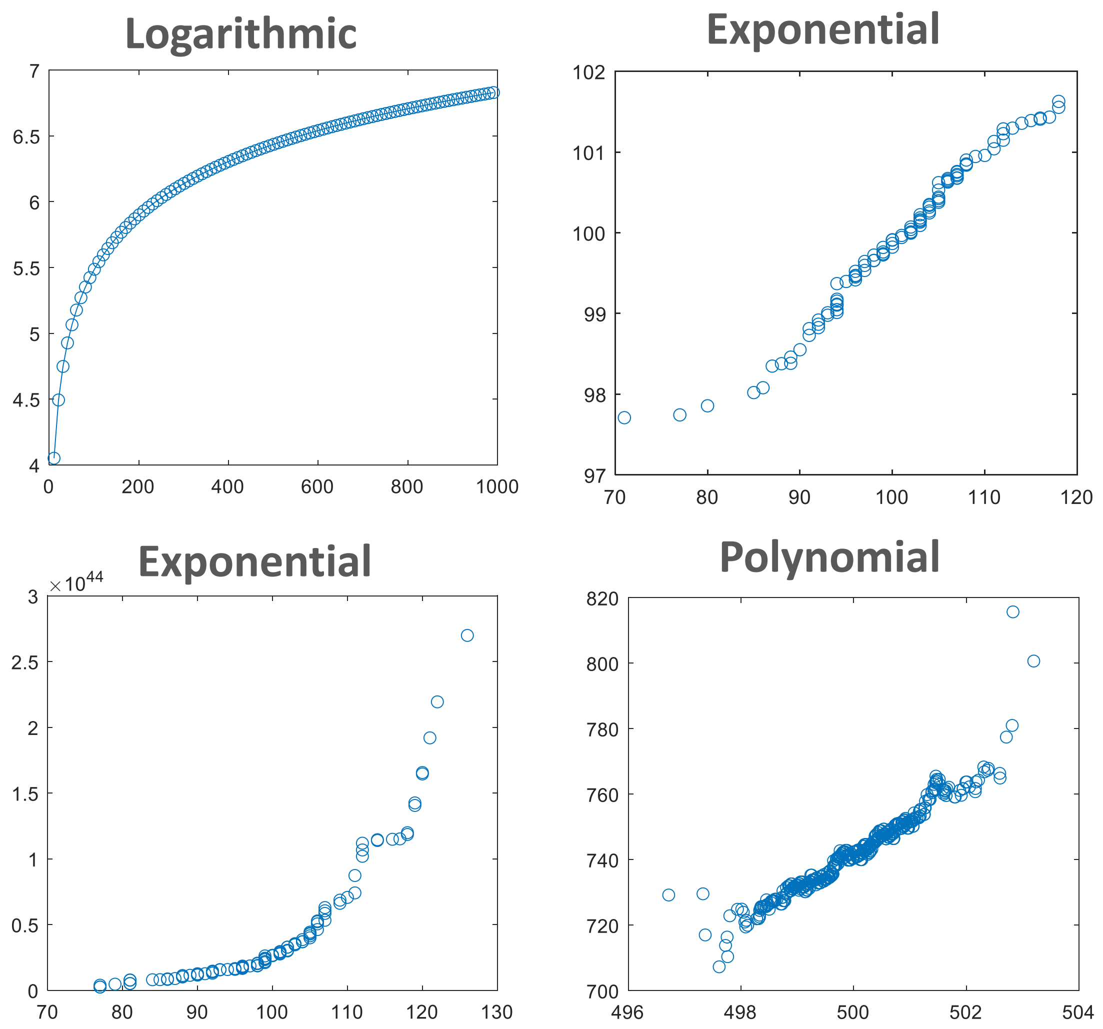
Code :: Two-sample Independent T-test

```
# levene Test
from scipy.stats import levene
levene(cal['medincome'], col['medincome'])
# two independent variables - t test
from scipy.stats import ttest_ind
ttest_ind(cal['medincome'], col['medincome'], axis=0, equal_var=True)
```

item	mean	std.dev	f test		t test	
			f value	p value	t value	p value
California	56102.91	15722.81	0.0135	0.9077	1.1682	0.2451
Colorado	52769.08	15144.20				

Correlation Analysis

- Correlation analysis is an inferential statistic to describe the relationship or association between one variable and another.
- Most formulae for correlation analyses are developed for linear relationships; therefore, other relationships (e.g., logistic, exponential, and cubic) are unsuitable. A nonlinear relationship could adopt the performance of curve-fitting results.



Correlation Analysis

Variable Y/X	Quantitative X	Ordinal X	Nominal X
Quantitative Y	Pearson r	Biserial r_b	Point Biserial r_{pb}
Ordinal Y	Biserial r_b	Spearman ρ/ Tetrachoric r_{tet}	Rank Biserial r_{rb}
Nominal Y	Point Biserial r_{pb}	Rank Biserial r_{rb}	Phi, L, C, V, Lambda

- There are two important outcomes from correlation analyses: significance and coefficient.
- **Significant correlation**: the consistency of the association between one variable and the other.
- **Coefficient of correlation**: the direction (i.e., positive or negative) and magnitude (i.e., value) of correlation between variables.

Pearson Correlation Coefficient r

- **Pearson correlation coefficient**, also known as Pearson product-moment correlation coefficient (PPMCC), is to measure the linear correlation between two variables or data.
- The definition of Pearson correlation coefficient is calculated by the covariance of the two variables divided by the product of their standard deviations. Its value ranges from -1 to +1.

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, when\ it\ is\ applied\ for\ population$$

$$r = \frac{cov(X,Y)}{s_X s_Y}, when\ it\ is\ applied\ for\ sample$$

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pearson Correlation Coefficient r

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y}, where\ cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)],$$

$$then\ \rho = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, and\ \ldots$$

$$\mu_X = \mathbb{E}[X]; \mu_Y = \mathbb{E}[Y];$$

$$\sigma_X^2 = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2;$$

$$\sigma_Y^2 = \mathbb{E}[(Y-\mathbb{E}[Y])^2] = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2;$$

$$\mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

$$\rho_{XY} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}\sqrt{\mathbb{E}[Y^2] - (\mathbb{E}[Y])^2}}$$

Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pearson Correlation Coefficient r

- Testing using t-distribution with degrees of freedom $n-2$, where standard error is denoted as,

$$\sigma_r = \sqrt{\frac{1-r^2}{n-2}}$$

- Therefore, t value is …

$$t = \frac{r}{\sigma_r} = r\sqrt{\frac{n-2}{1-r^2}}$$

The inverse function for determining the critical values for r is …

$$r = \frac{t}{\sqrt{n-2+t^2}}$$

Pearson Correlation Coefficient r

Sleeping/Day, X_i	Relax/Day, Y_i
7.5	1
8	12
9.1	2
6	10
10	5
8.4	6.1
9.1	7
2.4	8.2
6.7	7
6.8	6
9	4.5

$$\rho = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{-2.2370}{2.0011 \times 3.0509} = -0.3664$$

$$t = \frac{r}{\sigma_r} = r\sqrt{\frac{n-2}{1-r^2}}$$

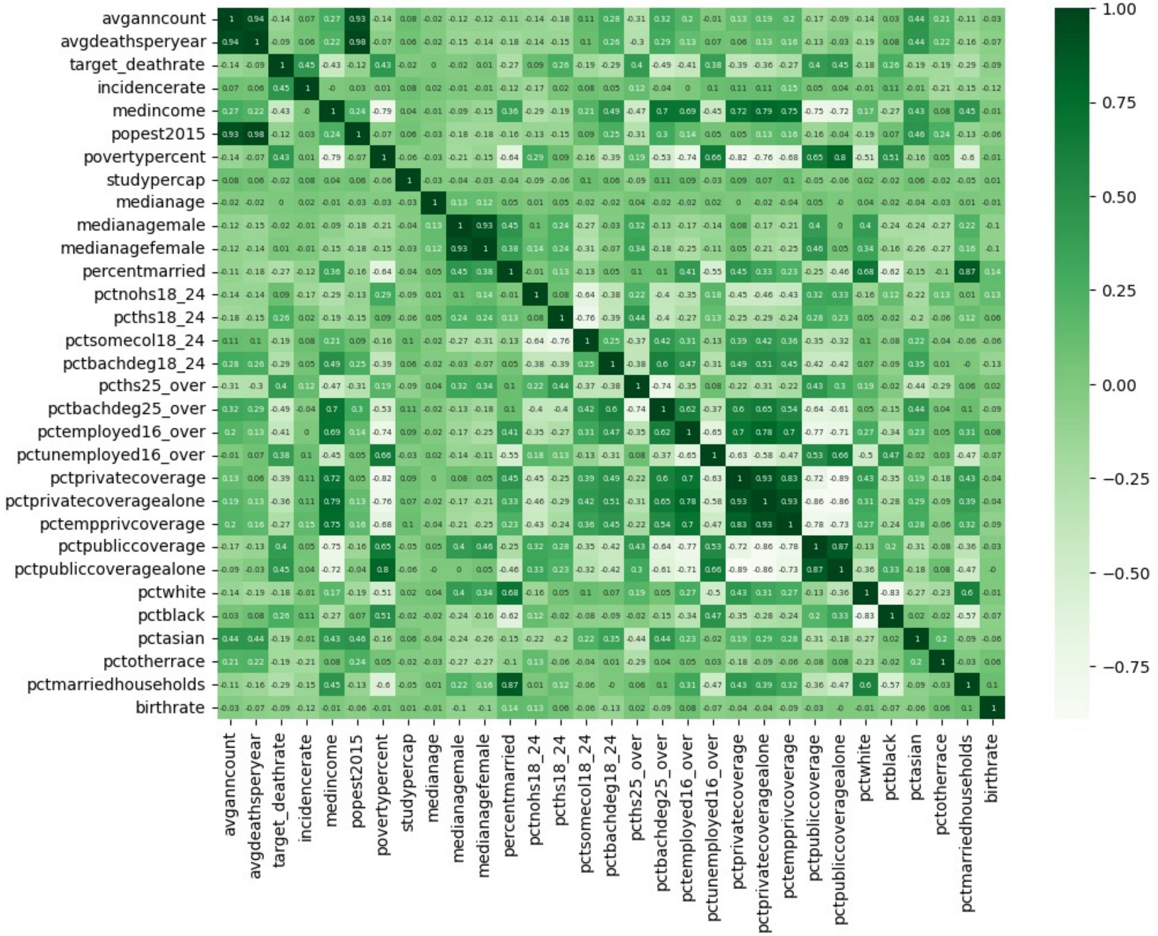
$$t = -0.3664 \times \sqrt{\frac{11-2}{1-(-0.3664)^2}}$$

$$= -1.18143$$

$$t_{-1.18143,9} = 0.133853$$

Code :: Pearson Correlation Coefficient r

```
# calculate Pearson correlation matrix
corr = df1.corr()
corr = np.round(corr,2)
# visualize correlation reuslts
plt.figure(figsize=(11,8))
sns.heatmap(corr, cmap="Greens",annot=True, annot_kws = {'size': 5})
plt.show()
```



Spearman Rank Correlation ρ

- **The Spearman correlation coefficient** is defined from the Pearson correlation coefficient between the rank variables.
- For a sample of size n, the n raw scores X_i, Y_i are converted to ranks $R(X_i), R(Y_i)$, and r_s is computed as

$$r_s = \rho_{R(X)R(Y)} = \frac{cov(R(X), R(Y))}{\sigma_{R(X)}\sigma_{R(Y)}}$$

where ρ denotes the Pearson correlation coefficient with rank variables, $cov(R(X), R(Y))$ is the covariance of the rank variables, $\sigma_{R(X)}\, and\, \sigma_{R(Y)}$ are the standard deviations of the rank variables.

Spearman Rank Correlation ρ

- Only if all n ranks are distinct integers, it can be computed using the popular formula.

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where $d_i = R(X_i) - R(Y_i)$ is the difference between the two ranks of each observation, n is the number of observations.

- Significance measurement could be obtained from t distribution, where degree of freedom is $n - 2$.

$$t = r_s \sqrt{\frac{n - 2}{1 - r^2}}$$

Spearman Rank Correlation ρ

PR, X_i	Reading/Day, Y_i	x_i rank	y_i rank	d_i	d_i^2
60	1	1	2	-1	1
65	1	2	2	0	0
71	2	3	4	-1	1
75	1	4	2	2	4
78	5	5	6	-1	1
81	6	6	7.5	-1.5	2.25
85	7	7	9.5	-2.5	6.25
89	8	8	11	-3	9
91	7	9	9.5	-0.5	0.25
95	6	10	7.5	2.5	6.25
99	4	11	5	6	36

- **Spearman Rank Corr.**

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$r_s = 1 - \frac{6\times 67}{11(11^2-1)}$$

$$r_s = 0.695455$$

$$t = r_s\sqrt{\frac{n-2}{1-r^2}}$$

$$= 2.870262$$

$$t_{2.87,df=9,two} = 0.018469$$

Code :: Spearman Rank Correlation ρ

```
# correlation analysis
# Pearson Correlation: linear parametric
from scipy.stats import pearsonr
rho_p, pval_p = pearsonr(df['avganncount'], df['avgdeathsperyear'])
print('Pearson Correlation:', rho_p, pval_p)
# Spearman Correlation: nonlinear or nonparametric
from scipy.stats import spearmanr
rho_s, pval_s = spearmanr(df['avganncount'], df['avgdeathsperyear'])
print('Spearman Correlation:', rho_s, pval_s)
```

```
Pearson Correlation: 0.9394077833002424 0.0
Spearman Correlation: 0.8177036458376524 0.0
```

Scale Problem

- Typically, we have five numerical transformation methods.
    1. Normalization
    2. Standardization
    3. Binarization
    4. Centralization
    5. Feature scaling
- Among these five methods, standardization is the most common method in statistical analysis because of its range of transformed value.

Scale Problem

Let $p \geqq 1$ be a real number.

The $p - norm$ (= $\ell_p - norm$) of vector $x = (x_1, \ldots, x_n)$ is

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$

Given $p = 1$

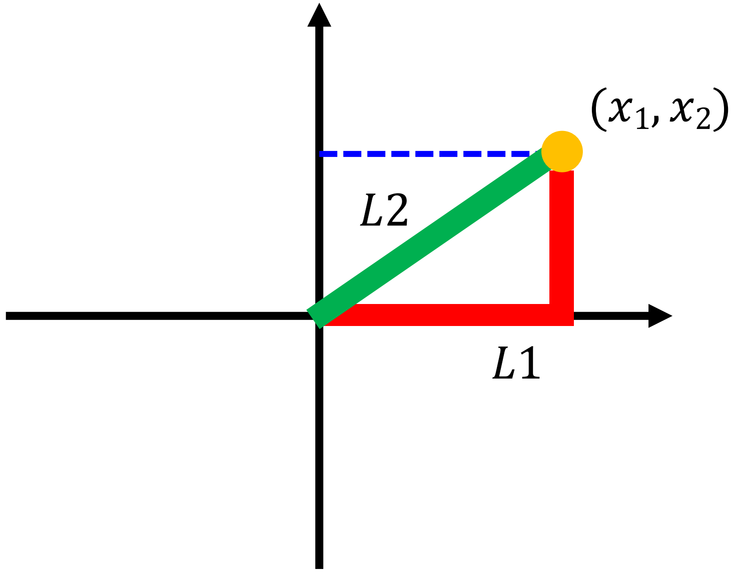
$$L1 = \|x\|_1 = |x_1| + |x_2|$$

Given $p = 2$

$$L2 = \|x\|_2 = \sqrt{|x_1|^2 + |x_2|^2}$$

Given $p = \infty$

$$L\infty = \|x\|_\infty = \max(x)$$



Scale Problem

Methods	Equations
Standardization	$x' = \frac{x-\mu}{\sigma}$
Binarization	$if\ x \geq threshold; x' = 1$ $else\ x' = 0$
Centralization	$x' = x - \frac{1}{n}\sum_{i=1}^{n} x_i$
Feature scaling	$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$

μ and σ denote mean and standard deviation, respectively.

Regression Analysis

- **Objective function:**
  - **Linear Regression:**

$$\mathrm{Min}_w \frac{1}{m}\sum_i (y_i - w^T x_i)^2$$

  - **Lasso Regression:**

$$Min_w \frac{1}{m}\left(\sum_i (y_i - w^T x_i)^2 + \lambda \sum_j^n |w_j|\right)$$

  - **Ridge Regression:**

$$Min_w \frac{1}{m}\left(\sum_i (y_i - w^T x_i)^2 + \lambda \sum_j^n w_j^2\right)$$

- **Gradient Descent:**

$$w^{new} = w^{old} - \gamma dw, where\ \gamma\ is\ the\ learning\ rate$$

Regression Analysis :: OLS

- OLS is the simplest linear regression, which stands for ordinary least squares.
- If we use partial differential on w, then we get …

$$\frac{2}{m} X^T (Y - Xw)$$

```
x_scaled = stats.zscore(x_list, axis=0, ddof=1)
x_scaled = sm.add_constant(x_scaled)

OLS_reg = sm.OLS(y_list['incidencerate'], x_scaled).fit()
OLS_reg.summary()
```

Regression Analysis :: Lasso

- Using the **L1 norm** concept in the objective function
- If we use partial differential on w, then we get …

$$
\begin{cases}
\frac{2}{m} X_j^T (Y - Xw) + \frac{1}{m} \lambda, w_j \geq 0 \\
\frac{2}{m} X_j^T (Y - Xw) - \frac{1}{m} \lambda, w_j < 0
\end{cases}
$$

```
from sklearn.linear_model import Ridge, RidgeCV, Lasso
lassoReg = Lasso(alpha=1)
lassoReg.fit(x_scaled, y_list['incidencerate'])
```

Regression Analysis :: Ridge

- Using the **L2 norm** concept in the objective function
- If we use partial differential on w, then we get …

$$\frac{2}{m}(X^T(Y - Xw) + \lambda w)$$

```
from sklearn.linear_model import Ridge, RidgeCV, Lasso
ridgeReg = Ridge(alpha=1)
ridgeReg.fit(x_scaled, y_list['incidencerate’])
plt.figure(figsize=[12,6], dpi=300)
plt.bar(np.arange(x_scaled.shape[1]), ridgeReg.coef_, facecolor='#81D8D0')
plt.xticks(np.arange(x_scaled.shape[1]), x_scaled.columns.values, rotation=80)
plt.plot([0, x_scaled.shape[1]], [0, 0], 'k')
plt.show()
```

Regression Analysis :: GLM

When your data does not follow normal distribution, you must find a suitable distribution for regression analysis.

Moreover, if your data has many zeros, you must use the zero-inflated version of GLM to overcome the zero inflation problem.

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function, $\mathbf{X}\boldsymbol{\beta} = g(\mu)$	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Negative inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{n-\mu}\right)$	
Categorical	integer: $[0, K)$	outcome of single *K*-way occurrence		$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
Multinomial	*K*-vector of integer: $[0, N]$	count of occurrences of different types (1, ..., *K*) out of *N* total *K*-way occurrences			

statsmodels 0.14.0

statsmodels.discrete.count_model.ZeroInflatedPoiss

```
class statsmodels.discrete.count_model.ZeroInflatedPoisson(
    endog,
    exog,
    exog_infl=None,
    offset=None,
    exposure=None,
    inflation='logit',
    missing='none',
    **kwargs
)                                                          [source]
```

Poisson Zero Inflated Model

Parameters

endog : array_like

A 1-d endogenous response variable. The dependent variable.

exog : array_like

Regression Analysis :: GLMRegression Analysis :: GLM

```
# GLM: original data
model = sm.GLM(y_list['incidencerate'],
               x_scaled,
               family = sm.families.Gaussian())
result = model.fit()
result.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	incidencerate	**No. Observations:**	3047
Model:	GLM	**Df Residuals:**	3022
Model Family:	Gaussian	**Df Model:**	24
Link Function:	Identity	**Scale:**	2406.7
Method:	IRLS	**Log-Likelihood:**	-16173.
Date:	Mon, 20 Nov 2023	**Deviance:**	7.2729e+06
Time:	01:02:53	**Pearson chi2:**	7.27e+06
No. Iterations:	3	**Pseudo R-squ. (CS):**	0.2171
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	448.2686	0.889	504.391	0.000	446.527	450.010
medincome	7.4001	2.445	3.027	0.002	2.608	12.192
popest2015	2.5009	1.056	2.369	0.018	0.432	4.570
povertypercent	-0.1909	2.477	-0.077	0.939	-5.045	4.663
studypercap	2.5546	0.907	2.818	0.005	0.778	4.331
medianage	1.0829	0.899	1.205	0.228	-0.679	2.845

Question Time

- **Assignment:**
- **Download today's lab practice and upload to moodle.**
- **Thx**

The End

Thank you for your attention!

Email: chchan@ntnu.edu.tw
Web: toodou.github.io

SCAN ME

SCAN ME