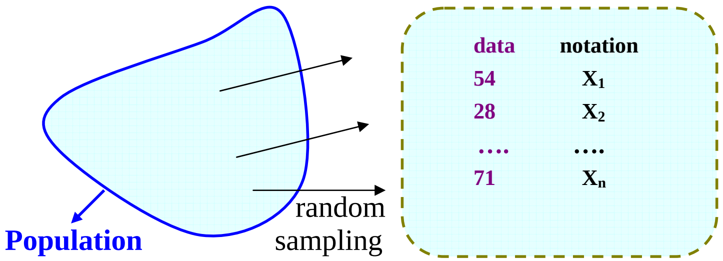
Lecture 02. (Part II) Descriptive Statistics



- **Terminology**

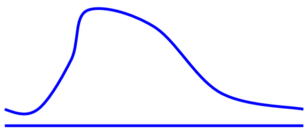
○ **random sample** ○ **random variable**

○ **sample size**

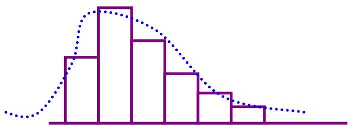
○ **probability density function (and cumulative distribution function) for continuous random variable**

○ **probability mass function for discrete random variable**

Distribution



Histogram



Population level

- mean: $\int_{-\infty}^{\infty} x f(x)dx \equiv \mathrm{E}x \equiv \mu$
- variance: $\int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \equiv \mathrm{Var}x \equiv \sigma^2$
- standard deviation: $\sqrt{\mathrm{Var}x} \equiv \sigma$
- skewness: $(1/\sigma^3)\int_{-\infty}^{\infty} (x-\mu)^3 f(x)dx$
- kurtosis: $(1/\sigma^4)\int_{-\infty}^{\infty} (x-\mu)^4 f(x)dx$
- p-th percentile: $\inf\{x : F_n(x) \le p\}$

  $p = 0.25 \rightarrow Q_1$

  $p = 0.50 \rightarrow Q_2(\mathrm{median})$

  $p = 0.75 \rightarrow Q_3$
- mode: $argmax f(x)$

Sample level

- mean: $\sum_{i=1}^{n} x_i/n \equiv \bar{x}$
- variance: $\sum_{i=1}^{n}(x_i-\bar{x})^2/n-1 \equiv S^2$
- standard deviation: $\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2/n-1} \equiv S$
- skewness: $\{(1/S^3)\sum_{i=1}^{n}(x-\bar{x})^3/n-1\} \times C_{3n}$
- kurtosis: $\{(1/S^4)\sum_{i=1}^{n}(x-\bar{x})^4/n-1\} \times C_{4n}$

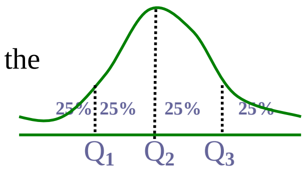
......

* The sample version of skewness and kurtosis involve 'correction' terms, C_{3n} and C_{4n}, which have values close to 1 when n is large. (Ref.: SAS8.2 Help System)

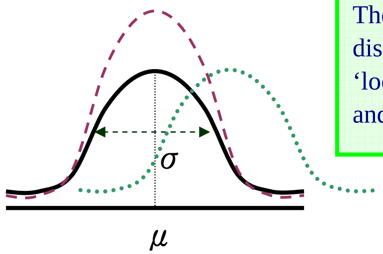
* At the so-called sample level, the variance has a denominator of 'n-1'. It is due to the requirement of 'unbiasedness'. Where, n-1 is termed as the degrees of freedom (df). However, note that 'unbiasedness' is not a necessary property that we ask for.

Some remarks

- Q_1,Q_2,Q_3 are the three most commonly used **percentiles**, and are referred to as the **quartiles**. They represent the 25-, 50- and 75-percentage points, respectively.



- The measures of "**central tendency**" include: mean, median, and mode. Measures of "**dispersion**" include: variance (std. dev.), Q_3-Q_1 (inter-quartile range; **IQR)**, the (full) range, kurtosis, etc.



The figure displays how different distributions can have different 'locations' (central tendency) and/or 'scales' (dispersions)

- The IQR, different from the standard deviation, is also a very useful measure of dispersion, which reveals the 'local dispersion' of data within the range between Q_3 and Q_1. The reader can refer to the corresponding **box-plot** to see the local distribution in this range.
- CV (**coefficient of variation**) is also often used to describe the 'dispersion', while adjusted for the 'overall magnitude' of the data. This is done through dividing the standard deviation by the sample mean:

$$\mathrm{CV} = \frac{S}{\bar{\mathbf{x}}} \times 100\%$$

- As will be shown in the following artificial example, we illustrate how two distributions may possess close variances but have very distinct kurtosis.

```
data kurt1;
 input x1 @@;
cards;
20 20 20 30 30 30 30 30 30 40 40 40 40 40 40 40 50 50 50 50
50 50 50 50 60 60 60 60 60 60 60 70 70 70 70 70 70 80 80 80
;
proc univariate plot normal;
  var x1;
run;

data kurt2;
 input x2 @@;
cards;
0 10 20 30 40 40 40 40 40 50 50 50 50 50 50 50 50 50 50 50
50 50 50 50 50 50 50 50 50 50 50 60 60 60 60 60 70 80 90 100
;
proc univariate plot normal;
  var x2;
run;
```

N	40	Sum Weights	40
Mean	50	Sum Observations	2000
Std Deviation	17.24633	Variance	297.435897
Skewness	0	Kurtosis	-0.9059297

N	40	Sum Weights	40
Mean	50	Sum Observations	2000
Std Deviation	17.3943699	Variance	302.564103
Skewness	0	Kurtosis	3.17769272

```
Stem Leaf                    #
   8 000                     3
   7
   7 000000                  6
   6
   6 0000000                 7
   5
   5 00000000                8
   4
   4 0000000                 7
   3
   3 000000                  6
   2
   2 000                     3
     ----+----+----+----+
 Multiply Stem.Leaf by 10**+1
```

```
Stem Leaf                         #
  10 0                            1
   9 0                            1
   8 0                            1
   7 0                            1
   6 00000                        5
   5 0000000000000000000000      22
   4 00000                        5
   3 0                            1
   2 0                            1
   1 0                            1
   0 0                            1
     ----+----+----+----+--
 Multiply Stem.Leaf by 10**+1
```

Example 1 (with sas code)

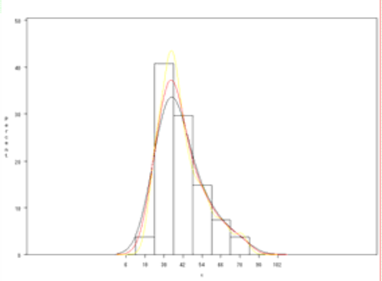
```
data lab01;
input scor @@;
cards;
22  24  24  27  28  29
32  33  33  33  34
35  37  37  38  39
40  43  44  47  48
53  55  57
64  69
79
;
proc univariate normal plot;
var scor;
run;
```

Note: (on the syntax of SAS)

SAS-output

```
                              Moments

N                          27    Sum Weights                27
Mean               40.8888889    Sum Observations         1104
Std Deviation      14.2810615    Variance           203.948718
Skewness           1.04935012    Kurtosis           0.75446305
Uncorrected SS          50444    Corrected SS       5302.66667
Coeff Variation    34.9265091    Std Error Mean     2.74839157


                   Basic Statistical Measures

         Location                    Variability

    Mean     40.88889     Std Deviation            14.28106
    Median   37.00000     Variance                203.94872
    Mode     33.00000     Range                    57.00000
                          Interquartile Range      16.00000
```

```
Quantile      Estimate

100% Max            79
99%                 79
95%                 69
90%                 64
75% Q3              48
50% Median          37
25% Q1              32
10%                 24
5%                  24
1%                  22
0% Min              22
```



```
 Stem Leaf                     #             Boxplot
    7 9                        1                0
    7
    6 9                        1                |
    6 4                        1                |
    5 57                       2                |
    5 3                        1                |
    4 78                       2             +-----+
    4 034                      3             |  +  |
    3 57789                    5             *-----*
    3 23334                    5             +-----+
    2 789                      3                |
    2 244                      3                |
      ----+----+----+----+
  Multiply Stem.Leaf by 10**+1
```

Example 2. (Length of stay of stroke patients at CMUH)

Variables		n	mean	(std)	Q1	Q2	Q3	t-test	K-W[2]
				Statistics[1]				Test	
Sex	0(female)	259	33.3	(21.1)	17.0	29.0	45.0	0.132	0.126
	1(male)	386	30.8	(20.2)	15.0	28.0	42.0		
Age	< 50	114	29.2	(19.3)	13.0	26.0	42.0	**0.032**	**0.019**
	50~64	210	34.9	(20.9)	19.0	32.0	48.0		
	65~79	275	31.1	(21.0)	16.0	28.0	42.0		
	>=80	46	27.8	(17.9)	12.0	26.0	35.0		
Comb	None	247	32.5	(22.3)	16.0	29.0	44.0	0.688	0.769
	DM	38	33.9	(21.4)	16.0	32.0	42.0		
	HYP	276	31.4	(19.1)	16.0	28.0	44.0		
	DM+HYP	84	29.9	(19.7)	16.0	27.5	40.0		
Phys	Yes	6	41.8	(25.7)	30.0	35.5	50.0	0.230	0.276
	No	639	31.7	(20.5)	16.0	28.0	43.0		
FIM	<29	161	38.6	(24.2)	21.0	34.0	49.0	**<0.001**	**<0.001**
	29~63	320	32.0	(19.1)	17.0	29.5	44.0		
	>=63	164	24.6	(16.8)	11.0	22.0	35.0		
MBI	0	172	35.5	(20.2)	20.5	33.0	47.5	**<0.001**	**<0.001**
	1~24	292	34.0	(21.7)	18.0	30.0	44.0		
	>=25	181	24.8	(17.2)	12.0	20.0	35.0		

1. n=Sample size; std=standard deviation; Q1, Q2, and Q3 are the 25-, 50- (median), and 75-percentage points.

2. Analysis of variance, reduces to t-test when K=2.

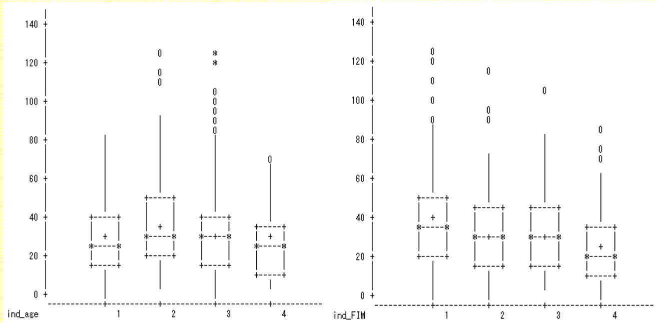
3. Kruskal-Wallis test, reduces to Wilcoxon's ranksum test when K=2

Source: "A model-based prediction on length of stay for rehabilitated stroke patients of mid-Taiwan" (by Chien-Lin Lin et al., CMUH; **preprint**)

Note: FIM (functional independence measure)的內容包括自我照顧能力、大小便控制、移位、走動、溝通、社會認知等因子，共分為18項，總分最高126，最低18。

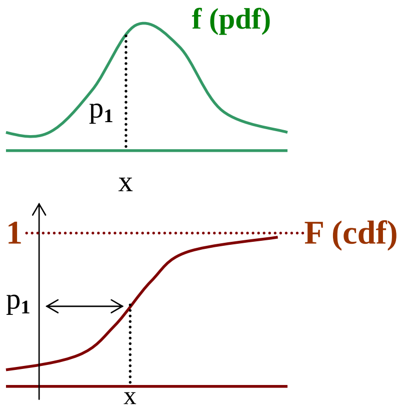
Example 2 (revisited with sas code) (★optional)

```
data strok;
infile 'd:\honda\lect03_example2.csv' dlm="," missover;
input id FIM age LOS;
/*proc print; run; */
proc univariate;   var FIM; run;
proc gplot;   plot (age FIM)*LOS;   run;
data strok1; set strok;
  if age<50 then do ind_age=1; end;
  if 50<=age<65 then do ind_age=2; end;
  if 65<=age<80 then do ind_age=3; end;
  if age>=80 then do ind_age=4; end;
  if FIM<30 then do ind_FIM=1; end;
  if 30<=FIM<45 then do ind_FIM=2; end;
  if 45<=FIM<65 then do ind_FIM=3; end;
  if FIM>=65 then do ind_FIM=4; end;
proc sort data=strok1; by ind_age;
proc univariate plot;   var LOS;   by ind_age;   run;
proc sort data=strok1; by ind_FIM;
proc univariate plot;   var LOS;   by ind_FIM; run;
```



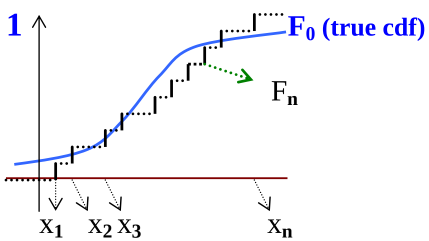
The Normal Probability Plot [a Q-Q plot] (★ optional)

• probability density function (**pdf**) and cumulative distribution function (**cdf**) of a continuous-type random variable (x):



$$\int_{-\infty}^{x} f(u)du = p_1 = F(x)$$
$$\int_{-\infty}^{\infty} f(u)du = 1 = F(\infty)$$
$$f(x) = \frac{dF(x)}{dx}$$

• **the empirical distribution**



$$F_n(x) = 0, \text{ if } x < x_1,$$
$$F_n(x) = \frac{1}{n}, \text{ if } x_1 \le x < x_2,$$
$$F_n(x) = \frac{2}{n}, \text{ if } x_2 \le x < x_3,$$
$$\ldots$$
$$F_n(x) = \frac{k}{n}, \text{ if } x_k \le x < x_{k+1}$$
$$\ldots$$
$$F_n(x) = 1, \text{ if } x \ge x_n,$$
or, in summary,
$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\{x_i \le x\}.$$

(For ease of exposition, let's assume $x_1<x_2<x_3<\ldots<x_n$)

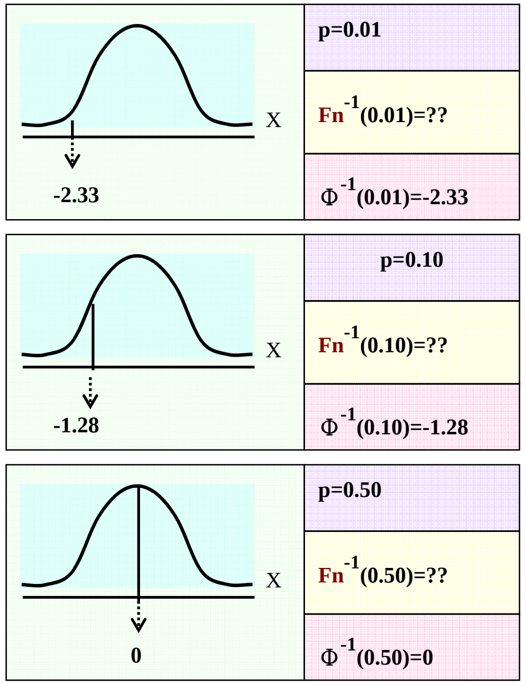
• The **indicator function** $\mathbf{1}\{A\}=1$, if the event A is true, and =0, otherwise.

• When n goes to infinity, the distance between $F_n(\bullet)$ and $F_0(\bullet)$ approaches 0 :

$$\sup_{\{x\}}|F_n(x)-F_0(x)| \rightarrow 0,$$

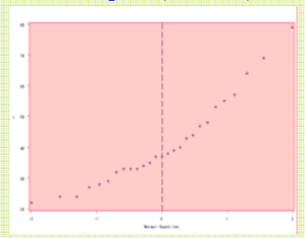
a basic property on which the famous **Kolmogorov-Smirnov (KS) test** is based. However, it's worth of noting that the KS test can be applied to test for goodness-of-fit problems with respect to any specific parametric distributions $F_0(\bullet)$.

• **The idea behind a 'Q-Q plot'** [If F_0 is a **Gaussian distribution**, cdf=Φ(x) and pdf=φ(x), then it is called a 'normal probability plot'.]



- Plotting $\mathbf{F_n^{-1}(p)}$ **vs.** $\mathbf{\Phi^{-1}(p)}$ for different values of p gives the so-called normal probability plot. [$F_n^{-1}(p)$ and $\Phi^{-1}(p)$ are the two **q**uantiles correspond to the same p for the empirical distribution F_n and cumulative (standard) normal distribution respectively. That's the reason why it is called a **Q-Q plot**.]

Example 1 (revisited)



- Ideally, if the data $X_1,\ldots X_n$ are drawn from the hypothetical F_0 (say the Gaussian distribution), then the points ($\Phi^{-1}(p)$, $F_n^{-1}(p)$) (for different p) will scatter around (very close to) the line x=y in the Cartesian X-Y plane (in case the data are suitably standardized.) This is a very useful **diagnostic plot** (or simply 'diagnostics'). In applications, seeking for powerful diagnostics is an ever-lasting effort for almost all sub-fields of statistics.
- **Q**: How to select p?

Ans.: For a sample of size n, it is convenient to consider the following p: p=1/n, 2/n,…, n-1/n, and n/n(=1) !!

Homework and exercise:

1. Please use the data offered in Example 1 to produce (using SAS or any other packages) the histogram, boxplot, empirical distribution (Fn), and the normal probability plot. Moreover, what is the 'quantile' $Fn^{-1}(0.3)$, and $Fn^{-1}(0.75)$? Please check the normal probability plot on your own calculation.

2. Please read pp.33~40, find the definition of sample correlation (r), and calculate 'r' of **Example 2.6a** in your textbook (page 38).

3. Do the following problems in your textbook (pp 41~54): 26, 29, 32