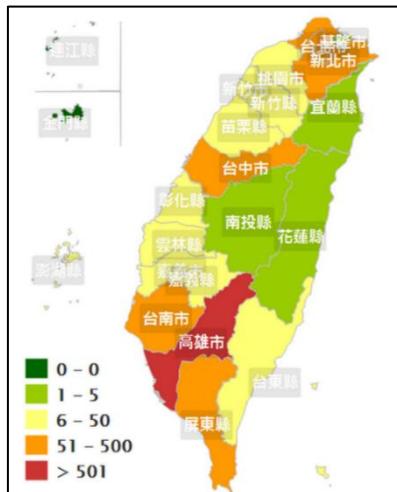


## 資料來源與彙整方式

研究對象：資料的區間定為 2023 年全年度登革熱確診資料，這裡以高雄地區為例，從登革熱病例地理分布圖可以看出高雄市為台灣嚴重的縣市。且高雄地區的氣候特徵為高溫濕潤，符合蚊蟲的生存條件。



圖片取自衛福部疾管署登革熱統計資料 記者陳俊廷翻攝

(<https://www.peoplemedia.tw/news/70b63bcd-2977-469b-be59-e5b9338d9a65>)

## 壹、登革熱確診數

(<https://data.cdc.gov.tw/dataset/dengue-daily-determined-cases-1998>)

條件：1. 高雄地區 2. 2023 年 3. 排除境外移入

資料粒度：日

A	B	C	D	E	F	G	H
發病日	通報日	性別	年齡層	居住縣市	居住鄉鎮	是否境外移入	確定病例數
1998/1/2	1998/1/7 M		40-44	屏東縣	屏東市	否	1
1998/1/3	1998/1/14 M		30-34	屏東縣	東港鎮	是	1
1998/1/13	1998/2/18 M		55-59	宜蘭縣	宜蘭市	是	1
1998/1/15	1998/1/23 M		35-39	高雄市	苓雅區	否	1
1998/1/20	1998/2/4 M		55-59	宜蘭縣	五結鄉	否	1
1998/1/22	1998/2/19 M		20-24	桃園市	蘆竹區	是	1
1998/1/23	1998/2/2 M		40-44	新北市	新店區	否	1
1998/1/26	1998/2/19 F		65-69	台北市	北投區	否	1
1998/2/11	1998/2/13 F		25-29	臺南市	南區	是	1
1998/2/16	1998/2/24 M		20-24	高雄市	楠梓區	是	1
1998/2/17	1998/2/23 F		30-34	高雄市	鳳山區	否	1

## 貳、高雄地區各區人口數

(<https://cabu.kcg.gov.tw/Stat/StatRpts/StatRpt1.aspx?yq=112&mq=1&dq=>)

資料粒度：月 => 將每月的資料分配到每日的資料

新增類別數據：切分南北高雄 以高雄火車站為界

A	B	C	D	E	F	G	H	I	J	K	L	M
1 行政區	1	2	3	4	5	6	7	8	9	10	11	12
2 左營區	196003	196192	196479	196620	196769	196824	197023	197009	196953	197026	197078	197276
3 橡梓區	191521	191672	191853	191929	192105	192242	192485	192667	192777	192888	192944	193139
4 前金區	26728	26808	26856	26897	26929	26924	26972	27011	27033	27053	27064	27098
5 小港區	154852	154817	154878	154878	154931	154977	155026	155066	155042	155034	154970	155009
6 凤山區	356359	356633	356693	356712	356813	356713	356762	356594	356397	356463	356476	356536
7 仁武區	95309	95522	95745	95934	96131	96353	96582	96808	97008	97155	97261	97337
8 大寮區	111575	111578	111574	111663	111770	111735	111845	111906	111910	111923	111916	111986
9 林園區	68459	68374	68340	68371	68303	68292	68325	68354	68328	68299	68254	68216
10 橋頭區	40551	40689	40819	40878	40971	41036	41138	41346	41488	41605	41685	41712

## 參、各空氣品質資料

( [https://airtw.moenv.gov.tw/CHT/Query/His\\_Data.aspx](https://airtw.moenv.gov.tw/CHT/Query/His_Data.aspx) )

資料粒度：小時 => 針對環境變數取每日最大值/最小值/平均數

A	B	C	D	E	F	G	H	I	J	K	L	M
1 測站	日期	測項	00	01	02	03	04	05	06	07	08	09
2 大寮	2023/1/1 00:00	AMB_TEMP	19.2	19.1	18.9	18.6	18.6	18.5	18.5	18.8	19.2	20.0
3 大寮	2023/1/1 00:00	CH4	2.19	2.18	2.18	2.14	2.18	2.14	2.18	2.2	2.15	2.15
4 大寮	2023/1/1 00:00	CO	0.5	0.5	0.47	0.45	0.45	0.44	0.46	0.51	0.45	0.47
5 大寮	2023/1/1 00:00	NMHC	0.16	0.16	0.16	0.1	0.13	0.11	0.11	0.17	0.13	0.11
6 大寮	2023/1/1 00:00	NO	0.8	0.7	0.7	0.7	0.7	0.9	0.9	3.4	3.3	3.3
7 大寮	2023/1/1 00:00	NO2	21.8	21.3	21.7	15.6	17.1	16	15.2	19.8	17	17
8 大寮	2023/1/1 00:00	NOx	22.6	22.1	22.5	16.4	17.8	16.9	16.1	23.2	20.4	17.3
9 大寮	2023/1/1 00:00	O3	7	8.3	8.4	10.6	7.7	8.6	7.5	4.6	10.8	17.3
10 大寮	2023/1/1 00:00	PM10	73	68	59	56	50	46	46	50	51	56
11 大寮	2023/1/1 00:00	PM2.5	43	28	34	35	33	30	32	28	33	30
12 大寮	2023/1/1 00:00	RAINFALL	0	0	0	0	0	0	0	0	0	0

原始資料 ( 單位：小時 / 測站 )

合併資料&針對每一個測項計算 min / max / mean ( 單位：天 )

station	date	min_A	min_C	min_C	min_N	min_N	min_N	min_O	min_PI	min_PI	
大寮區	2023/1/1	18.5	2.05	0.35	0.06	0.6	12.8	14.6	4.6	46	28
大寮區	2023/1/2	18.9	1.97	0.24	0.02	0.3	8.3	9.1	7.8	18	14
大寮區	2023/1/3	17.5	2.02	0.31	0.03	0.3	9.7	10.3	3.4	38	18
大寮區	2023/1/4	16.5	1.98	0.27	0.07	0	10.8	13.4	3	24	12
大寮區	2023/1/5	16.6	2.01	0.33	0.08	0	14.4	15.2	0.5	46	27
大寮區	2023/1/6	17.8	1.98	0.29	0.07	0.8	11.6	12.8	3.7	47	28
大寮區	2023/1/7	17.9	2	0.31	0.03	0.3	9	9.4	0.6	48	27
大寮區	2023/1/8	19.9	2.03	0.38	0.05	0.4	8.4	9.7	1.7	44	33
大寮區	2023/1/9	18.6	2.01	0.35	0.08	0.5	15	16.2	2.4	69	26
大寮區	2023/1/10	19.3	1.98	0.29	0.06	0.2	17.6	18.3	1.1	44	22
大寮區	2023/1/11	19.8	1.99	0.34	0.08	0.3	11	13.4	2.8	49	32
大寮區	2023/1/12	20.3	1.98	0.33	0.1	0.6	14.7	16.3	0	46	26
大寮區	2023/1/13	18.9	1.87	0.11	0	0	4	4.2	0.1	16	9

轉置後資料 ( 單位：天 )

# 資料前處理

## 壹、 篩選變數

我這邊是以  $\pm 0.3$  為基準，因為在統計上大於 0.3 就算有一定程度的相關性，可以排除掉那種幾乎沒有線性關係的變數，只留下比較有解釋力的。

	變數名稱	相關係數	相關方向
0	max_AMB_TEMP	0.5069827340326898	正
1	max_RH	0.49377868515655604	正
2	mean_AMB_TEMP	0.47962792961377676	正
3	mean_RH	0.4220574537322701	正
4	max_PM2_5	0.31952839442923947	正
5	mean_PM2_5	0.31379396853865893	正

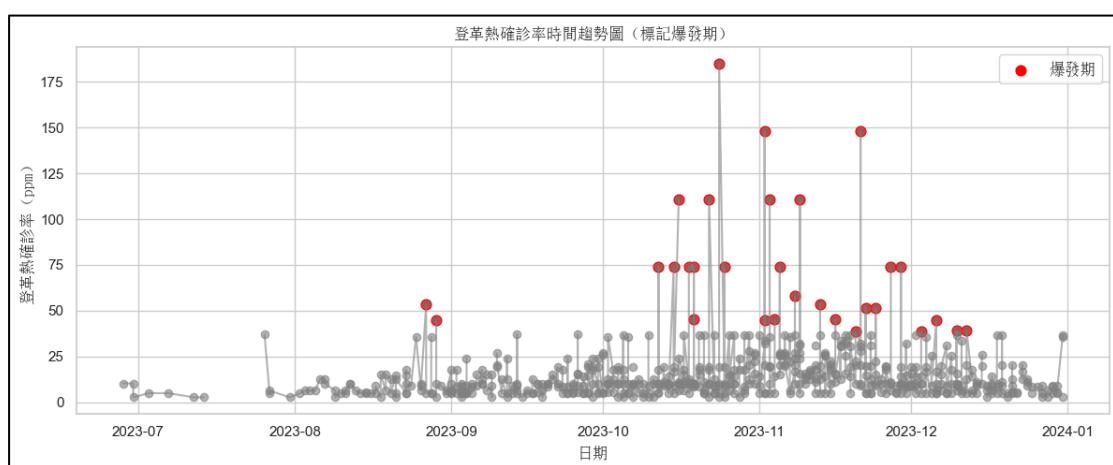
## 貳、 缺失值

因資料以每日最大、最小及平均值彙整，少量遺漏時段已於計算中自動忽略，最終資料無缺失值。

	資料完整性	欄位數量	樣本總數
0	<input checked="" type="checkbox"/> 全部欄位無缺失值	14	695

## 參、 離群值

以四分位數法(Q1、Q3、IQR)檢視離群值，確診率高值多屬疫情爆發期，具流行意義，故予以保留並標註，不刪除。



## 敘述性統計

研究變項：

Y（應變數）：每一百萬人中就有幾個人確診登革熱。

$$\text{登革熱確診 ppm} = \frac{\text{登革熱確診數}}{\text{人口數}} * 1000000$$

X（自變數）：空氣品質、人口密度及其他環境因素

資料：

變數名稱	單位	類型	中文說明
station		類別	測站代碼
date	YYYY-MM-DD	datetime	日期
max_AMB_TEMP	°C	連續	當日最高溫度
mean_AMB_TEMP	°C	連續	當日平均溫度
max_PM2_5	µg/m³	連續	當日最高 PM2.5
mean_PM2_5	µg/m³	連續	當日平均 PM2.5
max_RH	%	連續	當日最高相對濕度
mean_RH	%	連續	當日平均相對濕度
item_ppm	ppm	連續	登革熱確診率（應變數）

## 壹、反應變數 – 登革熱確診 ppm

這裡抓取登革熱確診數除以該地區總人口數即為登革熱確診 ppm，表示每一百萬人中就有幾個人確診登革熱，公式如下：

$$\text{登革熱確診 ppm} = \frac{\text{登革熱確診數}}{\text{人口數}} * 1000000$$

登革熱確診率 ppm	
樣本數 (N)	695.0
平均值	15.8019
標準差	17.0174
偏態 (Skewness)	4.4331
峰度 (Kurtosis)	29.239
Shapiro-Wilk 檢定 p 值	0.0

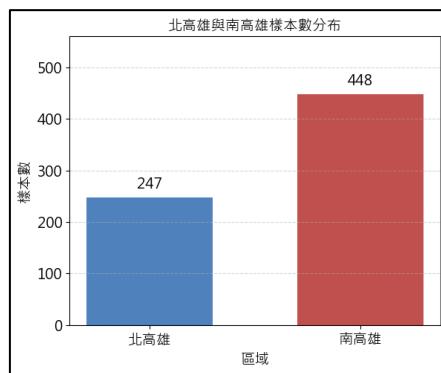
登革熱確診率平均約 15.8 ppm，變異程度大且右偏，p 值顯示不符常態分布。

## 貳、解釋變數 - 類別變數

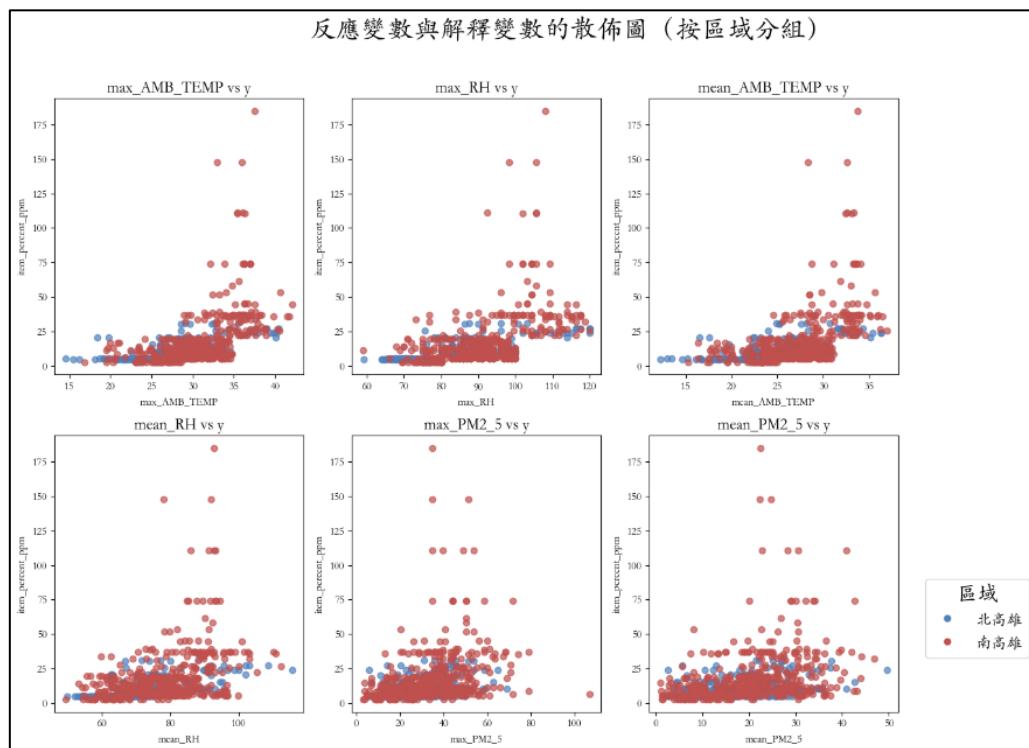
加入兩個有意義的類別變數，分別為地區、季節。

### 甲、 地區

這裡的地區以北高雄以及南高雄作為區分，高雄火車站以北為北高雄，以南為南高雄。



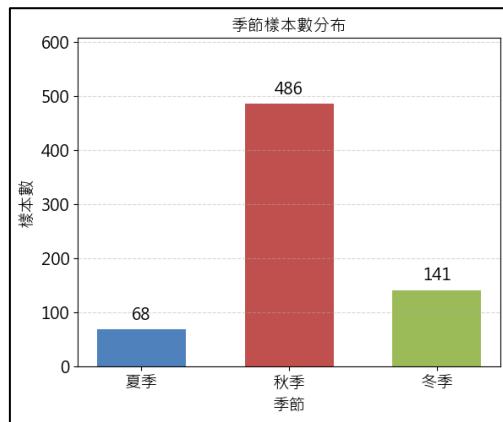
從此圖中可以看出南高雄比北高雄登革熱確診人數更多。



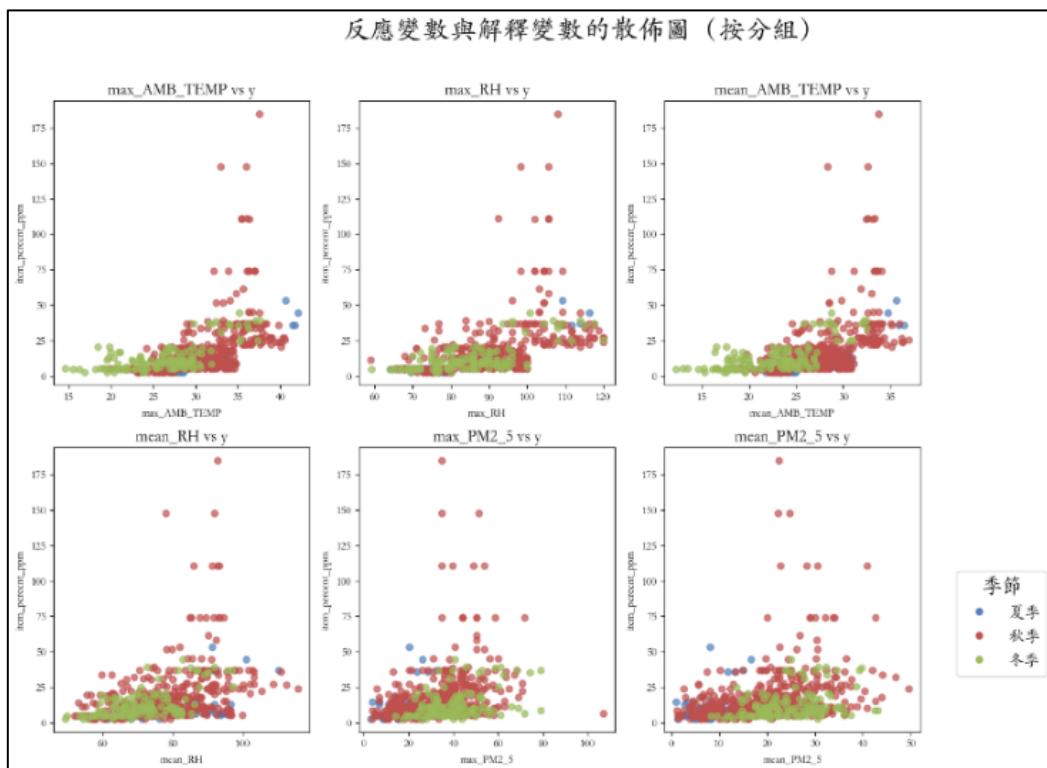
兩組資料差異不大，暫時不需要分組分析，先以整體資料進行後續分析。

## 乙、季節

這裡的季節分為春季、夏季、秋季、冬季，其中春季為「三月」、「四月」、「五月」，夏季為「六月」、「七月」、「八月」，秋季為「九月」、「十月」、「十一月」，冬季為「十二月」、「一月」、「二月」。



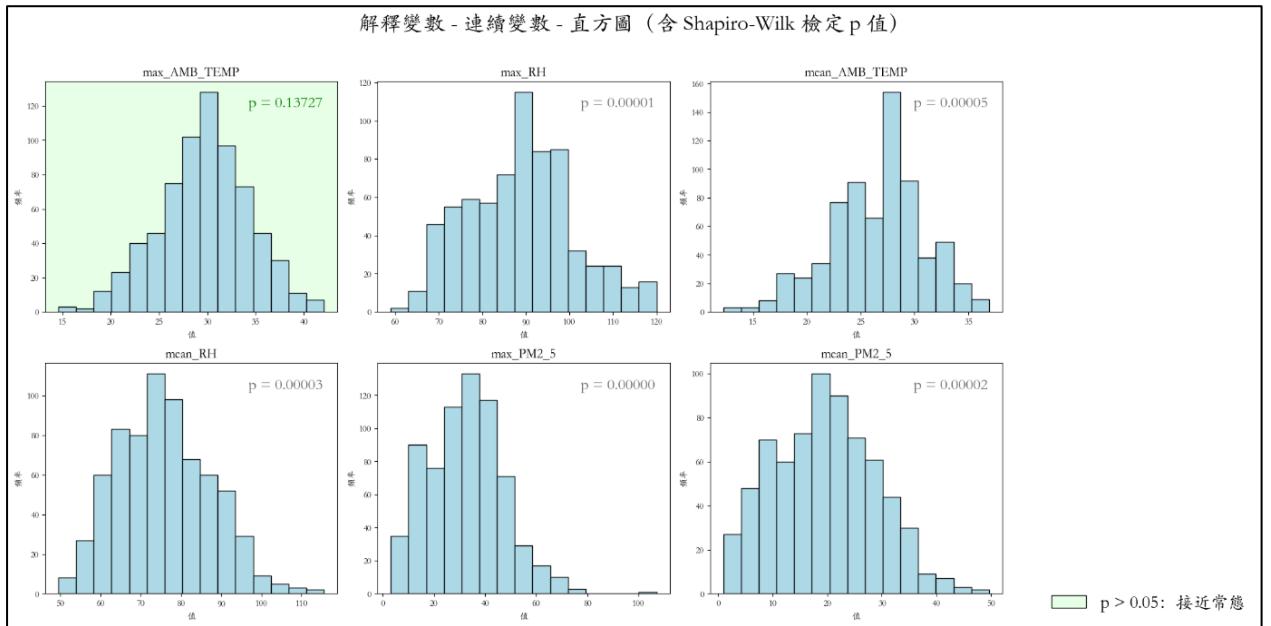
大多集中在秋季，猜測可能是因為夏季開始確診登革熱並且經過時間傳染，導致秋季病例數達到高峰，而冬季病例數則逐漸趨緩。



兩組資料差異不大，暫時不需要分組分析，先以整體資料進行後續分析。

## 參、 解釋變數 - 連續變數

從圖中可見，大部分變數（如相對濕度與 PM2.5）呈現右偏或長尾分布，僅有「當日最高溫度（max\_AMB\_TEMP）」未顯著偏離常態 ( $p = 0.137$ )，顯示其分布較為平滑。此結果屬於環境監測資料的常見特性，因此後續分析將以原始數據為主，不進行變數轉換。



## 研究目的：

目的	使用變數	統計/圖形
判斷地區與季節是否具關聯性	地區、季節	卡方
判斷不同季節登革熱確診率是否存在差異	季節、確診率 y	ANOVA
判斷哪一季確診情形最嚴重	季節、確診率 y	箱型圖
判斷確診率與多項環境因子是否顯著相關	環境因子、確診率 y	Pearson
確診率與環境因子之間的關係分布	環境因子、確診率 y	散佈圖

## 壹、判斷地區與季節是否具關聯性

為避免後續模型忽略地區與季節間的交互作用，這裡我們針對地區以及季節去做卡方檢定看他彼此之間是否獨立。假設檢定如下：

$H_0$ ：地區、季節兩變數是獨立的

$H_1$ ：地區、季節兩變數不是獨立的

FREQ 程序					area * season 之表格的統計值			
次數 百分比 列百分比 欄百分比	area * season 的表格				統計值	DF	值	機率
	season(季節)							
area(地區)	冬季	秋季	夏季	總計				
北高雄	43 6.19 17.41 30.50	176 25.32 71.26 36.21	28 4.03 11.34 41.18	247 35.54				
南高雄	98 14.10 21.88 69.50	310 44.60 69.20 63.79	40 5.76 8.93 58.82	448 64.46				
總計	141 20.29	486 69.93	68 9.78	695 100.00				

這裡可以看到佔比最高的是秋季的南高雄，佔了比較大的部分。另外可以看到卡方檢定 p value 大於 0.05 不拒絕虛無假設，本研究沒有顯著證據證明地區與季節兩者具有關聯性，後續模型（不）需加入  $area \times season$  交互作用或進行分層分析。。

參考：SAS Institute Inc. (2024). *The FREQ Procedure (Getting Started)*.

[https://documentation.sas.com/doc/en/statug/15.2/statug\\_freq\\_gettingstarted01.htm](https://documentation.sas.com/doc/en/statug/15.2/statug_freq_gettingstarted01.htm)

## 貳、判斷不同季節登革熱確診率是否存在差異

為了解登革熱在不同季節是否有明顯差異，這裡我們針對季節做 anova 變異數檢定。

$H_0$ ：每組(秋季、冬季、夏季) $\mu$  無顯著差異

$H_1$ ：每組(秋季、冬季、夏季)至少有一組 $\mu$  存在顯著差異

應變數: y 確診ppm					
來源	DF	平方和	均方	F 值	Pr > F
模型	2	5418.1742	2709.0871	9.59	<.0001
誤差	692	195559.4868	282.6004		
已校正的總計	694	200977.6610			

結果顯示 P value <0.0001 小於 0.05，拒絕虛無假設，表示在統計上組別之間至少有一組的 $\mu$  不相等，後續也可以進一步探討每兩組之間的 $\mu$  是否存在顯著差異。

參考：SAS Institute Inc. (2024). *The ANOVA Procedure (Getting Started)*.

[https://documentation.sas.com/doc/en/statug/15.2/statug\\_anova\\_gettingstarted01.htm](https://documentation.sas.com/doc/en/statug/15.2/statug_anova_gettingstarted01.htm)