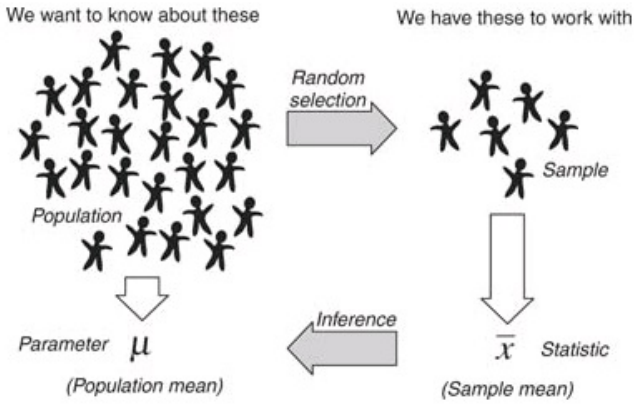
Chapter 4. Descriptive Statistics

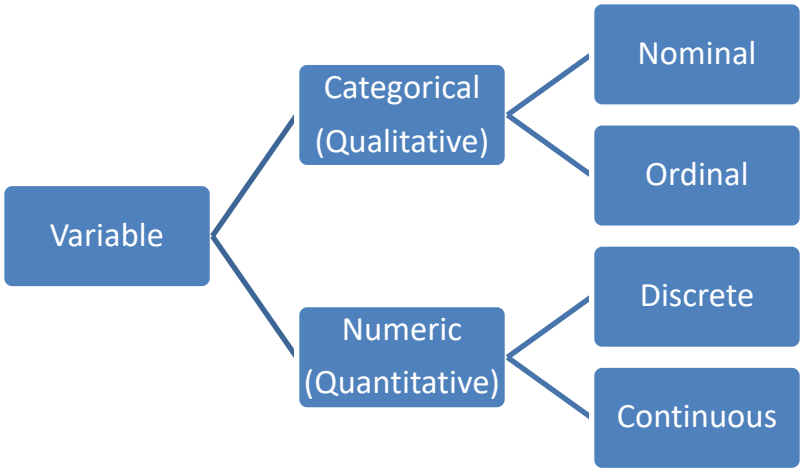
4.1. What is Statistics?

Cartoon by Jim Borgman, first published by the Cincinnati Inquirer and King Features Syndicate 1997 Apr 27; Forum section: 1
Reprinted in the New York Times, 27 April 1997, E4.

4.2. Population (Parameter) and Sample (Statistic)



4.3. Descriptive Statistics



- Distribution of a variable tells us what values it takes and how often it takes these values.
- Tabular description: Frequency table (one-way, two-way, …)
- Graphical description
  - Stem-and-leaf plot
  - Dot plot
  - Bar graph (Categorical)
  - Histogram (Continuous)
  - Box plot (cf. 5-number summary)
  - Scatterplot
- Measure of location
  - Mean
  - Median
  - Mode
- Measure of dispersion
  - Range
  - Quantile (Percentile)
  - 5-number summary (Min, Q1, Median, Q3, Maximum), Interquartile range (IQR)
  - Variance / Standard deviation
  - Coefficient of variation (CV)

4.4. Summarize Categorical Variables: PROC FREQ

- Count frequencies of both *character* and *numeric* variables in one-, two-, …, *n*-way tables.
- For n-way contingency table, separate each name with '*' in TABLES statement.
- Create output datasets containing counts and percentages.
- Compute various statistics such as chi-squared test, Fisher's exact test and odds ratio.
- The first listed variable forms the rows of the table, and the second forms the columns.
- The third variable creates multiple tables (stratification).

  e.g. var1 * var2 * var3: Create tables of var2 (row) and var3 (col) for each level of the var1

General Syntax

```
proc freq data=dataset;
     tables variable-combinations / <options>;
     * e.g. var1 var1*var2 var1*var2*var3 ...;
run;
```

- PROC FREQ options (Appear after a slash in the TABLES)

Option	Description
LIST	Print cross-tabulations in list format rather than grid
MISSPRINT	Include missing values in frequencies but not in percentages
MISSING	Include missing values in frequencies and percentages
NOCOL	Suppress printing of column percentage in cross-tabulations
NOROW	Suppress printing of row percentage in cross-tabulations
NOPERCENT	Suppress printing of global percentages
OUT = `out-dataset`	Write a dataset containing frequencies

4.5. Summarize Continuous Variables: PROC MEANS

- Primarily used for reporting various summary statistics of *numeric* variables.
- Without options, it will calculate the summary statistics for all numeric variables.

  (Default statistics: N (number of non-missing obs), Mean, Standard deviation, Min and Max)

General Syntax

```
proc means data=dataset;
     by list-of-variables;
     class list-of-variables;
     var list-of-variables;
     output out=out-dataset;
run;
```

- PROC MEANS options

Option	Description	Option	Description
MAX	Maximum value	N	Number of non-missings
MIN	Minimum value	NMISS	Number of missings
MEAN	Mean	RANGE	Range
MEDIAN	Median	STDDEV	Standard deviation
MODE	Mode	SUM	Sum
MAXDEC=*n*	Number of decimal places to be displayed	MISSING	Treat missing values as valid summary groups.
P20	20% quantile	NOPRINT	Do not print the means result.

- Optional statements

Option	Description
BY `list-of-variables`	Perform separate analyses for each level of the variables in the list. The dataset must first be *sorted* by these variables.
CLASS `list-of-variables`	Perform the same thing as BY statement, but the output is more compact. No sorting needed.
VAR `list-of-variables`	Specify which numeric variables to use in the analysis. If not specified, then SAS uses all numeric variables.

4.6. Examine Distribution of Continuous Variables: PROC UNIVARIATE

- Explore a dataset *before* conducting any statistical test.
- Produce statistics and graphs describing the distribution of a single variable.

  (e.g. mean, median, mode, standard deviation, skewness, kurtosis[1])
- Good for checking distributional assumptions (Normality).
- Without VAR statement, SAS will calculate statistics for all numeric variables in the dataset.

General Syntax

```
proc univariate data=dataset;
     var list-of-variables;
run;
```

[1] Skewness indicates how asymmetrical the distribution is; Kurtosis indicates how flat or peaked the distribution is.

Example

Raw Data

Obs	pregnant	blood	insulin	bmi	pedigree	age	test	BMIlevel
1	6	72	.	33.6	0.627	50	Positive	Obese
2	1	66	.	26.6	0.351	31	Negative	Overweight
3	8	64	.	23.3	0.672	32	Positive	Healthy
4	1	66	94	28.1	0.167	21	Negative	Overweight
5	0	40	168	43.1	2.288	33	Positive	Obese

SAS Code

```
proc freq data=pima;
     tables BMIlevel * test /
nocol missing out=freqout;
run;
```

```
proc means data=pima n nmiss mean std range;
     class test;
     var insulin blood bmi age;
     output out=meansout1;
run;

proc univariate data=pima normal;
     var insulin blood;
run;
```

Output

Frequency
Percent
Row Pct

Table of BMIlevel by test

BMIlevel	test(test) Negative	Positive	Total
	9 1.17 81.82	2 0.26 18.18	11 1.43
Healthy	95 12.37 93.14	7 0.91 6.86	102 13.28
Obese	253 32.94 53.60	219 28.52 46.40	472 61.46
Overweight	139 18.10 77.65	40 5.21 22.35	179 23.31
Underweight	4 0.52 100.00	0 0.00 0.00	4 0.52
Total	500 65.10	268 34.90	768 100.00

test	N Obs	Variable	Label	N	N Miss	Mean	Std Dev	Range
Negative	500	insulin	insulin	264	236	130.2878788	102.4822366	729.0000000
		blood	blood	481	19	70.8773389	12.1612228	98.0000000
		bmi	bmi	491	9	30.8596741	6.5607369	39.1000000
		age	age	500	0	31.1900000	11.6676548	60.0000000
Positive	268	insulin	insulin	130	138	206.8461538	132.6998982	832.0000000
		blood	blood	252	16	75.3214286	12.2998663	84.0000000
		bmi	bmi	266	2	35.4067669	6.6149824	44.2000000
		age	age	268	0	37.0671642	10.9682537	49.0000000