Chapter 1. Introduction to Statistics

1.1 Fundamental Statistical Concepts

Defining the Problem

The purpose of the study is to determine whether or not the average combined Math and Verbal scores on the Scholastic Aptitude Test(SAT) at Carver County magnet high school is 1200 – the goal set by the school board.

Concept

1. *population*: a collection of all objects about which information is desired.
2. *sample*: a subset of the population. a representative of the population, meaning that the sample characteristics are similar to the population's characteristics.

Parameters and Statistics

	Population Parameters	Sample Statistics
Mean	μ	$\bar{x}$
variance	σ^2	s^2
Standard deviation	σ	s

Parameters are characteristics of populations. Because populations usually cannot be measured in their entirety, parameter values are generally unknown. *Statistics* are quantities calculated from the values in the sample.

Describing Your Data

The goals when you are describing data are to

- screen for unusual data values
- inspect the spread and shape of continuous variables
- characterize the central tendency
- draw preliminary conclusions about your data

Distributions

A *distribution* is a collection of data values that are arranged in order, along with the relative frequency. It is important that you describe the location, spread, and shape of your distribution using graphical techniques and descriptive statistics.

"Typical Values" in a Distribution

- mean(평균): the sum of all the values in the data set divided by the number of values

- median(중앙값): the middle value (also known as the 50thpercentile)
- mode(최빈값): the most common or frequent data value

The Spread of a Distribution : Dispersion

- range: the difference between the maximum and minimum data values
- interquartile range: the difference between the 25th and 75th percentiles
- variance: a measure of dispersion of the data around the mean
- standard deviation: a measure of dispersion expressed in the same units of measurement as your data (the square root of the variance)

➔ 평균, 표준편차를 알면 특정 데이터가 전체 데이터 중 어느 정도 위치에 있는지 파악할 수 있다.

➔ Difference between *variance* and *MSE*.

The *variance* just measures the dispersion of the values, whereas *MSE* indicates how different the values of the estimator and the actual values of the parameters are. The *MSE* is a comparison of the estimator and the true parameter, as it were. This is why *MSE* includes both the *variance* of the estimator and its *bias* ($\mathrm{E}(\hat{\theta} - \theta)$).

$$MSE = variance + bias^2$$

The MEANS Procedure

```
PROC MEANS DATA=SAS-data-set <options>;
    VAR variables;
RUN;
```

VAR specifies numeric variables for which you want to calculate descriptive statistics. If no **VAR** statement appears, all numeric variables in the data set are analyzed.

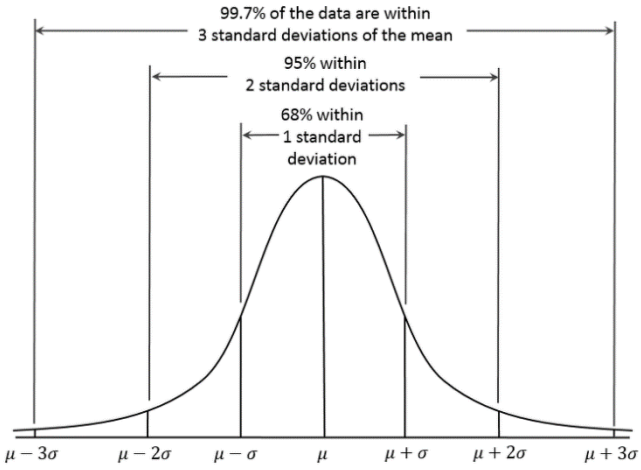
1.2 Picturing Distributions

Picturing Distributions: Histogram

It is a good idea to look at your data to see if the distribution of your sample data can reasonably be assumed to come from a population with that distribution. A histogram is a good way to get an idea of what the population distribution is shaped like.

Normal Distribution

- is symmetric. (If you draw a line down the center, you get the same shape on either side)
- is fully characterized by the mean and standard deviation. Given those two parameters, you know all there is to know about the distribution.
- is bell shaped.
- has mean = median = mode.

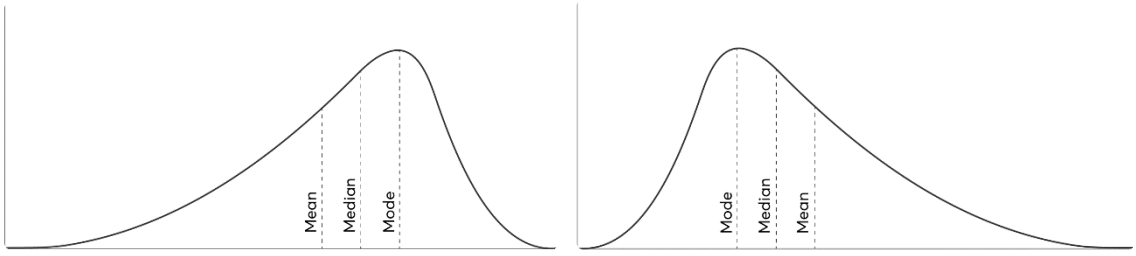


The center of the distribution is the population mean (μ). The standard deviation (σ) describes how variable the distribution is about μ. A larger standard deviation implies a wider normal distribution.

- μ (location parameter)
- σ (scale parameter)

Skewness (왜도)

The *skewness* statistic measures the tendency of your distribution to be more spread out on one side than the other. A distribution that is approximately symmetric has a skewness statistic close to 0.



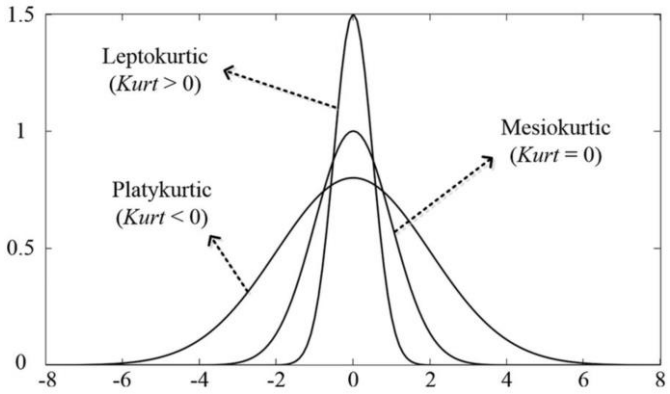
If your distribution is more spread out on the

- left side, then the statistic is negative, and the mean is less than the median. This is sometimes referred to as a *left-skewed* or *negatively skewed* distribution.

- right side, then the statistic is positive, and the mean is greater than the median. This is sometimes referred to as a *right-skewed* or *positively skewed* distribution.

➔ 왜도는 분포의 비대칭성을 측정하는 척도이며, 분포의 형태가 대칭인 경우에는 0에 가깝게 나타나고, 오른쪽으로 치우치면 양수, 왼쪽으로 치우치면 음수를 나타낸다.

Kurtosis (첨도)



The kurtosis statistic measures the tendency of your data to be distributed toward the center or toward the tails of the distribution. A distribution that is approximately normal has a kurtosis statistic close to 0 in SAS.

If your kurtosis statistic

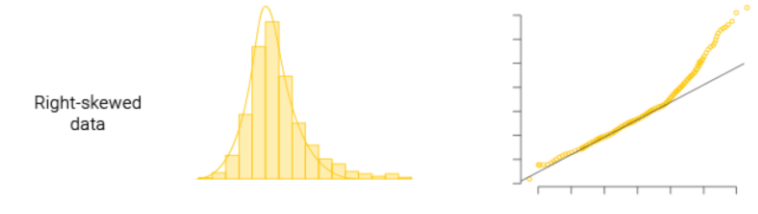
- is negative, the distribution is said to be *platykurtic*(평성) compared to the normal. A platykurtic distribution is often referred to as *light-tailed*.
- is positive, the distribution is said to be *leptokurtic*(급첨) compared to the normal. A leptokurtic distribution is often referred to as *heavy-tailed*. Leptokurtic distributions are also sometimes referred to as *outlier-prone* distribution.

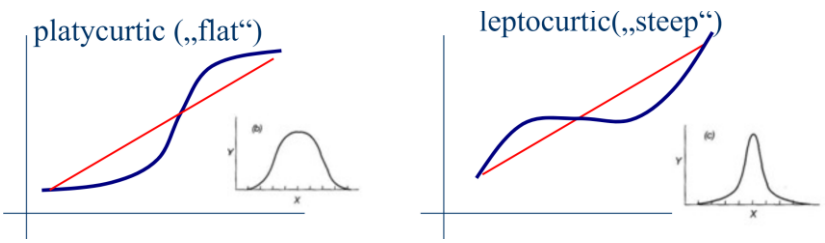
➔ 첨도는 분포의 첨예 정도를 나타내는 척도로, 분포가 뾰족한 형태인 경우에는 꼬리가 무겁다고 표현하며 이때 첨도는 양의 값을 갖는다. 반면, 분포가 퍼져 있는 경우에는 꼬리가 가볍다고 하며 이때 첨도는 음의 값을 갖는다.

Graphical Displays of Distributions

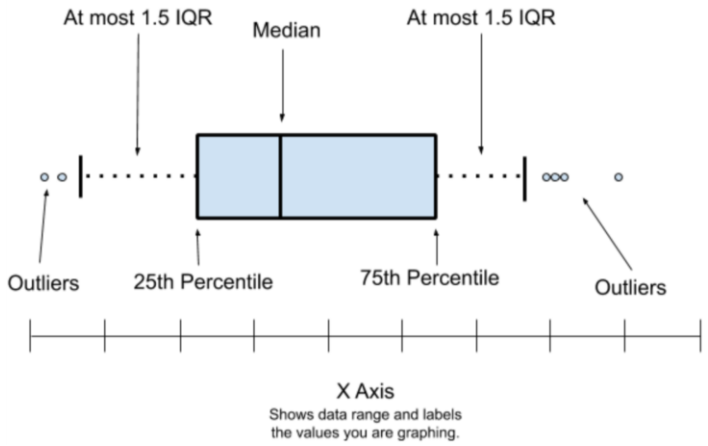
- histograms
- normal probability plots
  - visual method for determining whether or not your data comes from a distribution that is approximately normal.
  - x-axis: the expected percentiles from a standard normal distribution

    y-axis: the actual data values





- box-and-whisker plots
  - provide information about the variability of data and the extreme data values.



Statistical Graphics Procedures in SAS

- **PROC SGSCATTER** creates single-cell and multi-cell scatter plots and scatter plot matrices with optional fits and ellipses.
- **PROC SGPLOT** creates single-cell plots with a variety of plot and chart types.

- **PROC SGPANEL** creates single-page or multi-page panels of plots and charts conditional on classification variables.
- **PROC SGRENDER** provides a way to create plots from graph templates that you have modified or written yourself.

The UNIVARIATE Procedure

```
PROC UNIVARIATE DATA=SAS-data-set <options>;
    VAR variables;
    ID variable;
    HISTOGRAM variables </ options>;
    PROBPLOT variables </ options>;
    INSET keywords </ options>;
RUN;
```

VAR specifies numeric variables to analyze. If no **VAR** statement appears, then all numeric variables in the data set are analyzed.

ID specifies a variable used to label the five lowest and five highest values in the output.

HISTOGRAM creates high-resolution histograms.

PROBPLOT creates a high-resolution probability plot, which compares ordered variable values with the percentiles of a specified theoretical distribution.

INSET places a box or table of summary statistics, called an inset, directly in a graph created with a **CDFPLOT**, **HISTOGRAM**, **PPPLOT**, **PROBPLOT**, or **QQPLOT** statement. The **INSET** statement must follow the plot statement that creates the plot that you want to augment.

Selected option for **HISTOGRAM** and **PROBPLOT**

NORMAL creates a normal probability plot. Options (MU= SIGMA=) determine the mean and standard deviation of the normal distribution used to create reference lines (normal curve overlay in **HISTOGRAM** and diagonal reference line in **PROBPLOT**).

The SGPLOT Procedure

```
PROC SGPLOT <option(s)>;
    DOT category-variable </ option(s)>;
    HBAR category-variable </ option(s)>;
    HBOX response-variable </ option(s)>;
    HISTOGRAM response-variable </ option(s)>;
    NEDDLE X=variable Y=numeric-variable </ option(s)>;
    REG X=numeric-variable Y=numeric-variable </ option(s)>;
    SCATTER X=variable Y=variable </ option(s)>;
    VBAR category-variable </ option(s)>;
    VBOX response-variable </ option(s)>;
RUN;
```

VBOX creates a vertical box plot that shows the distribution of you data.

REFLINE creates a horizontal or vertical reference line.

1.3 Confidence Intervals for the Mean

Point Estimates(점추정)

A point estimate is a sample statistic used to estimate a population parameter. Different samples yield different estimates of the mean for the same population. How close on average these sample means are to one another is the variability of the estimate of the population mean.

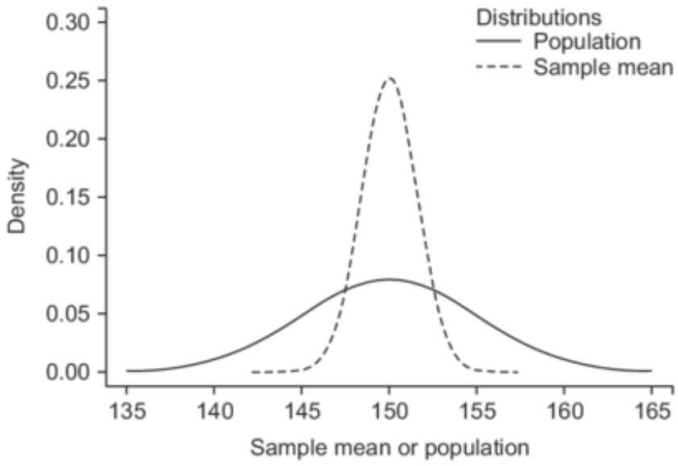
➔ 표본 데이터를 이용하여 하나의 수치로 모수값을 추정.

➔ 한계: 해당 수치가 얼마나 정학한지는 알 수 없다.

➔ 점추정 값은 표본에 따라 그 값이 매번 달라진다. (변동성 有)

➔ 표본 평균들이 매번 비슷하다면 모평균을 추정하기 위한 표본평균이 비교적 정확함을 의미한다.

Distribution of Sample Means



The variability of the distribution of the sample means is smaller than the variability of the distribution of the population.

Standard Error of the Mean(평균의 표준오차)

A statistic that measures the variability of your estimate is the standard error of the mean.

It differs from the sample standard deviation because

- the sample standard deviation deals with the variability of your data
- the standard error of the mean deals with the variability of your sample mean.

➔ standard deviation(표준편차): 개개의 데이터의 흩어진 정도

➔ standard error of the mean(평균의 표준오차): 표본 데이터로부터 계산된 통계량인 표본평균의 흩어진 정도 (=표본평균의 표준오차)

➔ 표본평균의 변동성 = 표준오차

➔ 표준오차가 작을수록 정밀한 값

➔ 표준오차가 작다 = 특정집단의 특성이 일관되다

= 모집단을 추정하는 데에 있어서 신뢰성을 가질 수 있다

Confidence Intervals

A 95% confidence interval states that you are 95% certain that the true population mean lies between two calculated values.

- In other words, if 100 different samples were drawn from the same population and 100 intervals were calculated, approximately 95 of them would contain the population mean.

➔ 모평균의 참값이 해당 구간 사이에 있다는 것을 95% 확신한다는 의미이다.

Normality and the Central Limit Theorem(중심극한정리)

To satisfy the assumption of normality, you can either

- verify that the population distribution is approximately normal, or
- apply the central limit theorem.
  - The central limit theorem states that the distribution of sample means is approximately normal, regardless of the distribution's shape, if the sample size is large enough.
  - "Large enough" is usually about 30 observations: more if the data is heavily skewed, fewer if the data is symmetric.

➔ 무한 모집단에서 무작위로 추출된 확률변수 X가 독립적으로 동일한 분포에 따라 $\mathrm{E(X)} = \mu, \mathrm{V(X)} = \sigma^2$ 인 경우 표본의 크기가 커짐에 따라 표본 평균 $\bar{x} = \sum x_i/n$ 은 근사적으로 평균 μ, 분산 σ^2/n의 정규 분포를 따른다.

➔ 중심극한정리는 모집단의 변수의 분포형태에 좌우되지 않는다.

➔ $X_1, X_2, \cdots$ 는 독립이고 동일 분포를 가지고 이들의 공통 평균과 분산이 각각 μ와 σ^2 이라고 하자. $S_n = X_1 + X_2 + \cdots + X_n$이라고 할 때, $\mathrm{n} \to \infty$이면

$$\frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} \mathrm{Z}$$

를 만족한다. 여기서 Z는 표준 정규분포를 가지는 확률 변수이다.

1.4 Hypothesis Testing

In a criminal court, you put defendants on trial because you suspect they are guilty of a crime. But how does the trial proceed?

- alternative hypothesis(대립가설): your initial research hypothesis(연구자가 밝히고자 하는 가설) – the defendant is guilty
- null hypothesis(귀무가설): the logical opposite of the alternative hypothesis – the defendant is not guilty

Use a decision rule to make a judgement. If the evidence is

- sufficiently strong, reject the null hypothesis.
- not strong enough, fail to reject the null hypothesis. Note that failing to prove guilt does not prove that the defendant is innocent.

Types of Errors

	ACTUAL	
DECISION	**H0 Is True**	**H0 is False**
Fail to Reject Null	Correct	Type II Error
Reject Null	Type I Error	Correct

- Type I Error(제1종 오류)
  - the probability that you reject the null hypothesis when it is true.
  - also called the significance level of a test (alpha:α).
  - 귀무가설이 사실일 때 이를 기각하는 경우(유의수준)
- Type II Error(제2종 오류)
  - the probability that you fail to reject the null hypothesis when it is false.
  - beta:β
- the power of a statistical test is equal to 1 – β(검정력). This is the probability that you correctly reject the null hypothesis.(실제로 귀무가설이 사실이 아닌 경우, 귀무가설을 기각할 확률).
- p-value(유의확률)
  - measures the probability of observing a value as extreme or more extreme than the one observed. For example, if your null hypothesis is that the coin is fair and you observe 40 heads(60 tails), the p-value is the probability of observing a difference in the number of heads and tails of 20 or more from a fair coin tossed

100 times.

- If the p-value is large, you would often see a difference this large in experiments with a fair coin. If the p-value is small, however, you would rarely see differences this large from a fair coin. In the latter situation, you have evidence that the coin is not fair.
- 귀무가설이 사실이라는 전제 하에, 현재 관측된 데이터 또는 그 이상으로 대립가설을 입증할 만한 데이터가 관찰될 확률.

Comparing α and the p-value

In general, you

- reject the null hypothesis if $p-\text{value} < \alpha$.
- fail to reject the null hypothesis if $p-\text{value} > \alpha$.