Lecture 11

Fundamental Sampling Distributions and Data Distributions I

Lecture Information

Ceng272 *Statistical Computations* at May 10, 2010

Dr. Cem Özdoğan
Computer Engineering Department
Çankaya University

Contents

Random Sampling I

- This chapter connects (bridges) the previous knowledge and the understanding of statistical inference.
- Outcome of a statistical experiment:
  - **Numerical value**: total value of a pair of dice tossed.
  - **Descriptive representation**: blood types in blood test.
- We focus on
  - sampling from distributions or populations
  - study such important quantities as the sample mean and sample variance.
- We extend the concept of probability distribution to that of a **sample statistic**.
- For instance, the distribution of a sample mean $\bar{X}$, which is a random variable because the different samples may result in different values of sample mean $\bar{x}$.
- The use of high speed computer enhances the use of formal statistical inference with graphical techniques.

Random Sampling II

- **Definition 8.1**:

  A population consists of the totality of the observations with which we are concerned.

- The number of observations in the population is defined to be the size of the population.
  - **Finite size**: 600 students are classified according to blood type: a population of size 600.
  - **Infinite size**: measuring the atmospheric pressure; some finite populations are so large.

Random Sampling III

- Each observation in a population is a value of a random variable X having some probability distribution $f(x)$.
- If one is inspecting items coming off an assembly line for defects, then each observation in population might be a value 0 or 1 of the binomial random variable X with probability distribution

$$b(x; 1, p) = p^x q^{1-x}, \ x = 0, 1$$

where 0 indicates a non-defective item and 1 indicates a defective item.

- **Definition 8.2**:

A **sample** is a subset of a population.

- Sometimes, it is impossible or impractical to observe the entire set of observations that make up the population.

Random Sampling IV

- Obtain representative samples to have a valid inference.
- **Biased sampling** procedure produces inference that consistently overestimate/underestimate some characteristics of the population.
- **Random sample**: selected independently and at random,
- **Definition 8.3**:

  Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $f(x)$ (identically distributed).
  Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size n from the population $f(x)$ and write its joint probability distribution as

  $$f(x_1, x_2, \ldots, x_n) = f(x_1) f(x_2) \ldots f(x_n) = \Pi_{i=1}^{n} f(x_k)$$

- If we assume the population of battery lives to be normal, the possible values of any X_i, $i = 1, 2, \ldots, 8$, will be precisely the same as those in the original population, and hence X_i has the same identical normal distribution as X.

Some important statistics I

- Random samples are selected to elicit information about the unknown population parameters.
- Some important statistics:
  - sample mean
  - sample variance
- **Definition 8.4**:

  Any function of the random variables constituting a random sample is called a **statistic**.
- Say p is a function of the observed values in the random sample.
- We would expect p to vary somewhat from sample to sample.
- That is a value of a random variable P, called a **statistic.**

Some important statistics II

- **Definition 8.5**:

  If $X_1, X_2, \ldots, X_n$ represent a random sample of size n, then the **sample mean** is defined by the statistic

  $$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The mean, median, and mode are the most commonly used statistics for measuring the central tendency.
- The computed value of $\bar{X}$ for a given sample is denoted by $\bar{x}$.
- Sample mean is not the same thing as the mean of a random variable but they are very closely related.
- Sample mode is the observation value that occurs the most number of times in a sample.
- Sample median is the middle value of a sample after sorting.

Some important statistics III

- **Definition 8.6**:

  If $X_1, X_2, \ldots, X_n$ represent a random sample of size n, then the **sample variance** is defined by the statistic

  $$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- The computed value of S^2 for a given sample is denoted by s^2.
- Again this is very related to the standard deviation of a random variable but is not the same thing.

Some important statistics IV

- **Example 8.1**: A comparison of coffee prices at 4 randomly selected grocery stores in San Diego showed increases from the previous month of 12, 15, 17, and 20 cents for a 1-pound bag.
- Find the variance of this random sample of price increases.
- Solution:

$$\bar{x} = \frac{12+15+17+20}{4} = 16$$

$$s^2 = \frac{\sum_{i=1}^{4}(x_i-16)^2}{4-1} = \frac{(12-16)^2+(15-16)^2+(17-16)^2+(20-16)^2}{3} = \frac{34}{3}$$

- **Theorem 8.1**:

If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]$$

Some important statistics V

- **Definition 8.7**:
  The **sample standard deviation**, denoted by S, is the positive square root of the sample variance.
- **Example 8.2**: Find the variance of the data 3, 4, 5, 6, 6, and 7, representing the number of trout caught by a random sample of 6 fishermen.
- Solution:

$$\sum_{i=1}^{6} x_i^2 = 171 \qquad \sum_{i=1}^{6} x_i = 31 \qquad \sigma^2 = \frac{6 * 171 - 31^2}{6 * 5} = \frac{13}{6}$$

Data Display and Graphical Methods I

- Motivation: Use creative displays to extract information about properties of a set.
  - The stem and leaf plots provide the viewer a look at symmetry of the data.
  - Normal probability plots and quantile plots are used to check normal distribution.
- Characterize statistical analysis as the process of drawing conclusion about system variability.
- Statistics provide single measures, whereas a graphical display adds additional information in terms of a picture.
- **Box-and-whisker plot** encloses the interquartile range of the data in a box that has median displayed within.
- A graphical tool to get an idea about the center, variability and degree of asymmetry of a sample.
- **Interquartile range**: between the 75^{th} percentile (**upper quartile**) and the 25^{th} percentile (**lower quartile**).
- **Box plot** provides the viewer information about **outliers** which represent rare event.

Data Display and Graphical Methods II

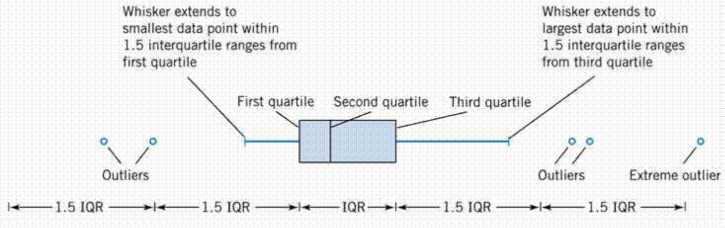

Figure: Box-and-Whisker plot.

Data Display and Graphical Methods III

- Nicotine content was measured in a random sample of 40 cigarettes. The data is displayed in the table.
- Mild outliers: 0.72, 0.85, and 2.55

Table: Nicotine Data for Example 8.3.

1.09	**0.85**	1.86	1.82	1.40	1.92	1.24	1.90
1.79	1.64	2.31	1.58	1.68	2.46	2.09	1.79
2.03	1.51	1.88	1.75	2.28	1.70	1.64	2.08
1.63	1.74	2.17	**0.72**	1.67	2.37	1.47	**2.55**
1.69	1.37	*1.75*	*1.97*	2.11	1.85	1.93	1.69

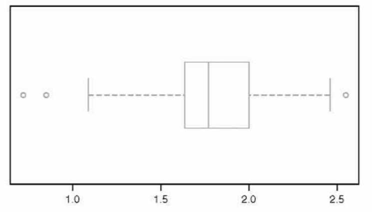

Figure: Box-and-Whisker plot for nicotine data.

- Sample size $n = 40$.
- Sort the sample.
- 25^{th} percentile: $\left(\frac{25*n}{100}\right)^{th}$ element in the sorted list.
- $q(0.25) = X_{sorted}(10) = 1.63$
- $q(0.50) = X_{sorted}(20) = 1.75$
- $q(0.75) = X_{sorted}(30) = 1.97$

Data Display and Graphical Methods IV

- Interquartile range:
  $q(0.75) - q(0.25) = 1.97 - 1.63 = 0.34$
- The whiskers are drawn at a distance of 1.5 times the interquartile range from the 25^{th} and 75^{th} percentiles.
- 1.63-1.5*0.34 & 1.977+1.5*0.34
- Anything outside that range is shown as an outlier.
- Another graphical tool: **Stem-and-leaf plot**.
  1. Split each observation into 2 parts: stem and leaf.
     - Stem can be the digit preceding the decimal,
     - Leaf can be the digit after the decimal.
  2. Make a table: List the stem values as rows. Add each leaf value with a specific stem value to that row.
- Gives an idea about what stem values occur more frequently.

Data Display and Graphical Methods V

```
Stem Leaf                        #          Boxplot
  25 5                           1             0
  24 6                           1             |
  23 17                          2             |
  22 8                           1             |
  21 17                          2             |
  20 389                         3          +-----+
  19 0237                        4          |     |
  18 2568                        4          |     |
  17 045599                      6          *--+--*
  16 3447899                     7          +-----+
  15 18                          2             |
  14 07                          2             |
  13 7                           1             |
  12 4                           1             |
  11                                           |
  10 9                           1             |
   9
   8 5                           1             0
   7 2                           1             0
     ----+----+----+----+
Multiply Stem.Leaf by 10**-1
```

Figure: Stem-and-leaf plot for the nicotine data.

Data Display and Graphical Methods VI

- **Example 8.4**: Consider the following data, consisting of 30 samples measuring the thickness of paint can ears.

Table: Data for Example 8.4.

Sample	Measurements					Sample	Measurements				
1	29	36	39	34	34	16	35	30	35	29	37
2	29	29	28	32	31	17	40	31	38	35	31
3	34	34	39	38	37	18	35	36	30	33	32
4	35	37	33	38	41	19	35	34	35	30	36
5	30	29	31	38	29	20	35	35	31	38	36
6	34	31	37	39	36	21	32	36	36	32	36
7	30	35	33	40	36	22	36	37	32	34	34
8	28	28	31	34	30	23	29	34	33	37	35
9	32	36	38	38	35	24	36	36	35	37	37
10	35	30	37	35	31	25	36	30	35	33	31
11	35	30	35	38	35	26	35	30	29	38	35
12	38	34	35	35	31	27	35	36	30	34	36
13	34	35	33	30	34	28	35	30	36	29	35
14	40	35	34	33	35	29	38	36	35	31	31
15	34	35	38	35	30	30	30	34	40	28	30

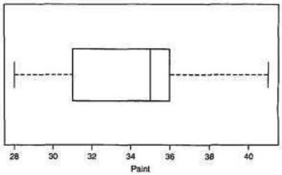

Figure: Box-and-whisker plot for thickness of paint can “ears”.

Data Display and Graphical Methods VII

- **Quantile plot**
  - Compare samples of data
  - Draw distinctions
  - Depict cumulative distribution function
- **Definition 8.8**:

  A **quantile** of a sample, $q(f)$, is a value for which a specified fraction f of the data values is less than or equal to $q(f)$.

- Sample median: q(0.5); 75^{th} percentile: q(0.75); 25^{th} percentile: q(0.25).
- A quantile plot simply plots the data values on the vertical axis against an empirical assessment of the fraction of observations exceeded by the data value.

Data Display and Graphical Methods VIII

- Let f_i be the i^{th} observation when they are sorted low to high.
- Then f_i is the $(i/n)^{th}$ quantile where n is the size of the sample.
- So we plot f_i vs (i/n). For theoretical purposes this fraction is computed as

Plotting position formula

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}} \qquad\qquad f_i = \frac{i - a}{n + 1 - 2a}$$

for some a

- where i is the order of the observations when they are ranked from low to high.
- In other words, if we denote the ranked observations as

$$y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n-1)} \leq y_{(n)}$$

then the quantile plot depicts a plot of $y_{(i)}$ against f_i.

Data Display and Graphical Methods IX

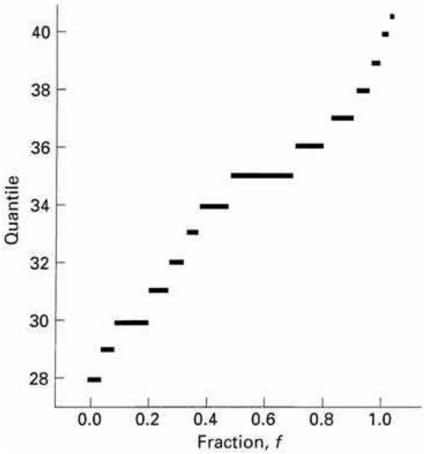

Figure: Quantile plot for paint data.

- In Fig. 5, quantile plot shows all observations.
- Large clusters: slopes near zero. e.g.: 36-38
- Sparse data: steeper slopes. e.g.: 28-30

Data Display and Graphical Methods X

- **Dedection of deviations from normality.**
- We often assumes that a data set are realizations of independently identically distributed normal random variables.
- Question: Did this sample come from a population with a normal distribution?
- Tool: We can take advantage of what is known about the quantiles of the normal distribution to answer this question.
- The diagnostic plot can often nicely augment a formal **goodness-of-fit test** on the data.

Data Display and Graphical Methods XI

- Approximation of quantile of normal distribution

$$q_{\mu,\sigma}(f) = \mu + \sigma \left\{ 4.91 \left[f^{0.14} - (1-f)^{0.14} \right] \right\}$$

$\mu = 0$ and $\sigma = 1$ for standard normal distribution

$$q_{0,1}(f) = 4.91 \left[f^{0.14} - (1-f)^{0.14} \right]$$

- **Definition 8.8**:

The **normal quantile-quantile plot** is a plot of $y_{(i)}$ ordered observations against $q_{0,1}(f_i)$, where

$$f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$$

- A nearly straight line relationship suggests that the data came from a normal distribution.
- The intercept on the vertical axis is an estimate of the population mean μ.
- The slope is an estimate of the standard deviation σ.

Data Display and Graphical Methods XII

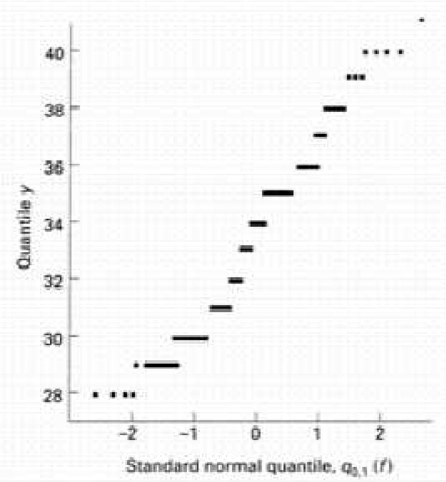

Figure: Normal quantile-quantile plot for paint data.

Data Display and Graphical Methods XIII

- **Normal probability plotting.**
- The vertical axis contains *f* plotted on special paper, known as probability paper.
- The scale used results in a straight line when plotted against the ordered values of a normal random variable.
- If the normal distribution adequately describes the data, the plotted points will fall approximately along a straight line.
- Construct a normal quantile-quantile plot and draw conclusions regarding whether or not it is reasonable to assume that the two samples are from the same $N(\mu, \sigma)$ distribution.

Data Display and Graphical Methods XIV

Table: Data for Example 8.5.

Number of Organisms per Square Meter			
Station 1		Station 2	
5,030	4,980	2,800	2,810
13,700	11,910	4,670	1,330
10,730	8,130	6,890	3,320
11,400	26,850	7,720	1,230
860	17,660	7,030	2,130
2,200	22,800	7,330	2,190
4,250	1,130		
15,040	1,690		

- Solution:
- Far from a straight line.
- Station 1 reflect a few values in the lower tail of the distribution and several in the upper tail.
- Unlikely!

Data Display and Graphical Methods XV

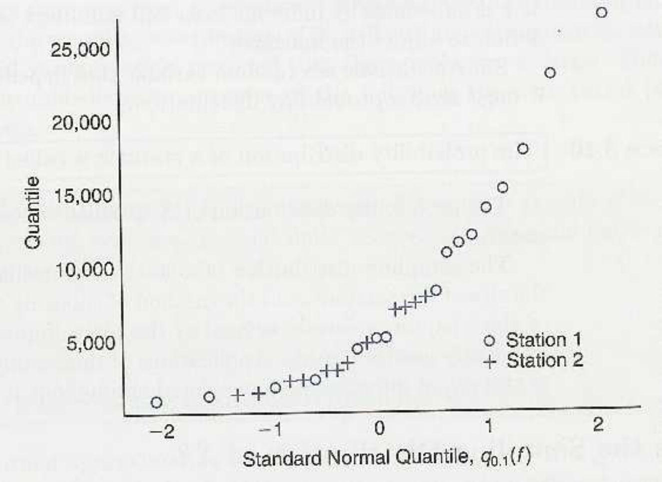

Figure: Standard Normal Quantile, $q_{0,1}(f)$.

Sampling Distribution I

- Statistical inference is concerned with **generalizations** and **predictions**.
- Based on the opinions of several people interviewed on the street, that in a forthcoming election 60% of the eligible voters in the city of Detroit favour a certain candidate.
- If we repeat the sampling, we would expect to obtain a different value for the sample mean.
- Therefore, like other random variables, the sample mean $\bar{X}$, possesses a probability distribution, which is more commonly called the **sampling distribution** of $\bar{X}$.
- Question: A company manufactures 100 Ohms resistors. A sample of 40 resistors from the assembly line is found to have a mean of 105 Ohms.
- How likely is the population mean (the mean of the probability density function) to be 100 Ohms?

Sampling Distribution II

- Answer: In questions like this, we need to make inferences about the population mean based on the sample mean.
- To do this, we need to know the probability distribution of the sample mean.
- **Definition 8.10**:

  The probability distribution of a statistic is called a **sampling distribution**.
- **Sampling Error**: The difference between the sample statistic and the value of the corresponding population parameter.
  - For the sample mean, the sampling error $= |\bar{x} - \mu|$. This is controllable by taking more *n*.
- **Nonsampling Error**: Human error. The error occurs while we collect, record or tabulate the data.
- The sampling distribution of a statistic depends on
  - the size of the population,
  - the size of the samples,
  - the method of choosing the samples.

Sampling Distribution of Means I

- Suppose that a random sample of n observations is taken from a normal population with mean μ and variance σ^2.
- By the reproductive property of the normal distribution (established in Theorem 7.11)

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$E(\bar{X}) = \mu_{\bar{X}} = \frac{\mu + \mu + \ldots + \mu}{n} = \mu$$

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2} = \frac{\sigma^2}{n} \; (or \; \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right))$$

The standard deviation of the sample mean, $\sigma_{\bar{X}}$ is called the standard error of $\bar{X}$.

- We call $\left(\frac{N-n}{N-1}\right)$ the finite population correction and it approaches 1 as $N \to \infty$.

Sampling Distribution of Means II

- **Example**: The following data gives the years of employment for all five employees (A, B, C, D, E) at the University Medical Center: 7, 8, 12, 7, 20.
- Let X denote the number of years of employment. The population distribution ($N = 5$) of X will be

X	7	8	12	20	$\sum p(x)$
$p(x)$	2/5	1/5	1/5	1/5	1.0

- Population mean; $\mu = \sum_{all\ x} x * p(x) = 10.8$ years
- Population variance; $\sigma^2 = \sum x^2 * p(x) - \mu^2 = 24.56$
- Now, we take a sample of size $n = 4$.
- There will be $\binom{5}{4} = 5$ ways of making combinations.

Sampling Distribution of Means III

- The following table shows the list all the possible samples (without replacement) that can be selected from this population.

Sample No	Sample	Sample Mean $\bar{x}$
1	(A,B,C,D) = 7,8,12,7	8.5
2	(A,B,C,E) = 7,8,12,20	11.75
3	(A,B,D,E) = 7,8,7,20	10.5
4	(A,C,D,E) = 7,12,7,20	11.5
5	(B,C,D,E) = 8,12,7,20	11.75

- Calculate the sample mean for each of these samples. Then, the sampling distribution of $\bar{X}$ is

$\bar{X}$	8.5	10.5	11.5	11.75	$\sum p(\bar{x})$
$p(\bar{x})$	1/5	1/5	1/5	2/5	1.0

Sampling Distribution of Means IV

- $E(\bar{X}) = \mu_{\bar{X}} = \sum_{all\ \bar{x}} \bar{x} * p(\bar{x}) = 10.8 = \mu$
- $\sigma^2_{\bar{X}} = \sum \bar{x}^2 * p(\bar{x}) - \mu^2_{\bar{X}} = 118.175 - (10.8)^2 = 1.535$
  - This can be verified by applying the finite population correction for the population variance

$$\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{24.56}{4}\left(\frac{5-4}{5-1}\right) = \frac{24.56}{4}\left(\frac{1}{4}\right) = 1.535$$

    which is exactly agreeable with sample variance of $\bar{x}$.

- If you chose sample number 3, then the sampling error$= |\bar{x} - \mu| = |10.5 - 10.8| = 0.3$ years.
- The sampling distribution of is normally distributed if the underlying population itself has a normal distribution.
- But what if the population distribution is not normally distributed or unknown?
- If a random sample of n observations is selected from a population (any population), then when n is sufficiently large, the sampling distribution of will be approximately a normal distribution.

Sampling Distribution of Means V

- **Theorem 8.2**:

  Central Limit Theorem. If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

  $$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

  as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

- The normal approximation for $\bar{X}$ will generally be good if $n \geq 30$.
- If $n < 30$, the approximation is good only if the population is not too different from a normal distribution.
- This is true no matter what the population distribution may be as long as the population has a finite variance σ^2.
- This marvellous and famous fact in probability theory is called the Central Limit Theorem.
- This is remarkable and an universal probability law.
- If the population is known to be normal, the sampling distribution of $\bar{X}$ will follow a normal distribution exactly, no matter how small the size of the samples.

Sampling Distribution of Means VI

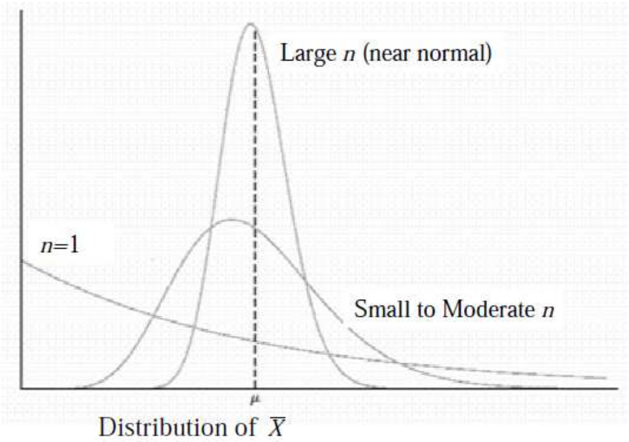

Figure: Illustration of the central limit theorem (distribution of $\bar{X}$ for $n = 1$, moderate n, and large n).