9.0 Estimation of a Random Variable's possible value

When we have collected data from an experiment, we can now use statistical estimation to determine the likely value of such variable when a future observation is to be performed.

Statistical inference consists of using methods by which one makes inferences or generalizations about a population.

Point Estimator = a rule or formula that tells us how to calculate an estimate based on the measurements contained in a sample. The number that results from the calculation is called a **point estimate**.

Interval Estimator = a formula that tells us how to use sample data to calculate an interval that estimates a population parameter θ.

Estimation Formulas and applied problems

A. Estimating the mean

A.1 Case 1: σ^2_x known

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma_x}{\sqrt{n}}$$

(1) The quality control manager at a light bulb factory needs to estimate the average life of a large shipment of light bulbs. The process standard deviation is known to be 100 hours. A random sample of 50 light bulbs indicated a sample average life of 350 hours.

a. Set up a 90% confidence interval estimate of the true average life of light bulbs in this shipment.

b. Set up a 95% confidence interval estimate of the true average life of light bulbs in this shipment.

c. Tell why an observed value of 320 hours would not be unusual even though it is outside the confidence interval you calculated.

A.2 Case 2: σ^2_x not known

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \qquad (v = n -1)$$

(2) The Director of Quality of a large health maintenance organization wants to evaluate patient waiting time at a local facility. A random sample of 25 patients selected from the appointment book. The waiting time was defined as the time

from when a patient signed in to when he or she was seen by the doctor. The following data represent the waiting time (in minutes)

19.5	30.5	45.6	39.8	29.6
25.4	21.8	28.6	52.0	25.4
26.1	31.1	43.1	4.9	12.7
10.7	12.1	1.9	45.9	42.5
41.3	13.8	17.4	39.0	36.6

Set up a 95 % confidence interval estimate of the population average waiting time.

B. Estimating the difference between two means $(\mu_1-\mu_2)$

B.1 Case 1: Large independent samples $(\bar{x}_1-\bar{x}_2)\pm Z_{\alpha/2}\sqrt{\dfrac{\sigma_1^{\ 2}}{n_1}+\dfrac{\sigma_2^{\ 2}}{n_2}}$

(3) It is desired to estimate the difference between the mean starting salaries for all bachelor's degrees graduates of DLSU in the Colleges of Engineering and Computer Science during the past year. The following information is available from the Career Development Office:

- A random sample of 40 starting salaries for Engineering graduates produced a sample mean of P230,500 and a standard deviation of P38,515.
- A random sample of 30 starting salaries for CompSci graduates produced a sample mean of P192,000 and a standard deviation of P35,452.

Construct a 95% confidence interval for the difference between mean starting salaries for graduates of the two colleges. Interpret the interval.

B.2 Case 2: Small independent samples with equal variances

$$(\bar{x}_1-\bar{x}_2)\pm t_{\alpha/2}S_p\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad \text{where} \quad S_p=\sqrt{\frac{(n_1-1)S_1^{\ 2}+(n_2-1)S_2^{\ 2}}{n_1+n_2-2}}$$

value of $t_{\alpha/2}$ is based on (n_1+n_2-2) degrees of freedom

(4) The Farm has received claims from its customers that the weight of point-of-sale adult pigs from its two farms are different by at least 5 kgs. The resident veterinarian decided to check the validity of this claim. He took 10 pigs from each farm's available pigs for sale, and the following data was found.

Farm 1	Farm 2

sample size	10 pigs	10 pigs
Mean weight	90.5	85.0
standard deviation of wts	3.2	3.4

a. Is the claim valid at 5% level of significance ?
b. Estimate a 95 % confidence interval for mean difference in pig weights between those of farm 1 and those of farm 2.

B.3 Case 3: Small independent samples with unequal variances

$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}$ where the distribution of $t_{\alpha/2}$ has degrees of freedom ν

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

The value of ν should be rounded *down* to nearest integer.

(5) You're trying to determine if a new route from your house to school would save you at least 10 minutes of travelling time. You recorded 4 weeks' travelling time using the two different routes and your data showed:

		Mean travel time	Std deviation
Old Route	(13 times)	55.2 minutes	5.2 minutes
New Route	(7 times)	42.7 minutes	10.3 minutes

Estimate a 90 % confidence interval of the difference in travelling times if you took the new route instead of the old one.

B.4 Case 4: Matched pairs (Dependent samples)

Let d_1, d_2 .. d_n represent the differences between pairwise observations in a random sample of n matched pairs. Then the small sample confidence interval for $\mu_d=(\mu_1-\mu_2)$ is

$\bar{d} \pm t_{\alpha/2}\left(\frac{S_d}{\sqrt{n}}\right)$ where $\bar{d}$ and S_d are the mean and std.dev. of the n sample differences.

(6) A new diet program called *!Give It Up!* claims to be effective in taking out the unwanted pounds off of obese people. As a benchmark to compare against, you used the Pritikin program. You randomly selected 30 people's and acquired their body weights before and after each program. The following table shows the data. Construct a 95% confidence interval for the mean weight loss under each program.

!Give It Up! person	Before	After	Pritikin person	Before	After
1	100 kgs	70 kgs	1	124 kgs	70 kgs
2	124	85	2	115	81
3	115	80	3	125	68
4	125	84	4	115	80
5	115	89	5	85	84
6	112	75	6	84	89
7	105	85	7	75	75
8	112	85	8	125	85
9	108	92	9	115	92
10	95	75	10	105	75
11	85	70	11	112	70
12	84	81	12	108	81
13	75	68	13	95	68
14	80	64	14	85	64
15	96	75	15	96	75

C. Estimating the variance $\dfrac{(n-1)s^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le \dfrac{(n-1)s^2}{\chi^2_{\left(1-\alpha/2\right)}}$

(7) Construct a 95 % CI on the variance based on the following set of data (the amount of Krypton gas (in milliliters) that leaked out each time its container was dropped from 4 ft above ground):

15.5	16.8	16.7	15.4	16.4	17.5
17.8	17.5	18.3	14.5	18.1	15.7
19.5	18.6	19.7	15.8	18.2	16.8

D. Estimating the ratio of two variances $\dfrac{s_1^2}{s_2^2}\cdot\dfrac{1}{F_{\alpha/2}\left(\nu_{1,}\nu_2\right)} \le \left[\dfrac{\sigma_1^2}{\sigma_2^2}\right] \le \dfrac{s_1^2}{s_2^2}\cdot\dfrac{1}{F_{\left(1-\alpha/2\right)}\left(\nu_{1,}\nu_2\right)} \equiv$

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2}\left(\nu_{1,}\nu_2\right)} \le \left[\frac{\sigma_1^2}{\sigma_2^2}\right] \le \frac{s_1^2}{s_2^2}\cdot F_{\alpha/2}\left(\nu_{2,}\nu_1\right)$$

(8) An investor wants to compare the risks associated with two different computer stocks, IBM and PhilCom, where the risk of a given stock is measured by the variation of daily prices. The investor obtains random samples of daily price changes for IBM and PhilCom. The sample results are summarized in the accompanying table. Compare the risks associated with both stocks by forming a 95 % c.i. for the ratio of the true population variances.

IBM	PhilCom
n_1 =21	n_2 = 21
X_1 = 0.585	X_2 = 0.572
S_1 = 0.023	S_2 = 0.014

E. Estimating a proportion $\bar{p} \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}\cdot\bar{q}}{n}}$

(9) In a recent employee satisfaction survey made by J.Rizal Industries, 356 out of 400 employees stataed that they were very satisfied or moderately satisfied with their jobs. Create a 95% c.i. estimate of the population proportion of the employees who were satisfied with their jobs.

F. Estimating the difference between two proportions $\left(\bar{p}_1 - \bar{p}_2\right) \pm Z_{\alpha/2}\sqrt{\frac{\bar{p}_1\cdot\bar{q}_1}{n_1} + \frac{\bar{p}_2\cdot\bar{q}_2}{n_2}}$

(10) In a controversial survey of dating preferences made at the University of the Philippines in 1992, 305 out of 500 students surveyed claimed "Good Conversationalist" as the one of the most preferred trait of a dating partner. Furthermore, 175 out of 500 claimed "sexual attractiveness" as the most important trait. Estimate the difference between the proportions of the UP studentry who preferred good talk over good looks at a 90% confidence level.

G. Test for Goodness of Fit between a hypothesized Distribution and actual data

Hypothesis: Data follows the hypothesized value.

Test statistic formula:
Where

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

where o_i = observed frequency of the ith interval cell (i=1 to k)
e_i = expected frequency according to the hypothesized distribution
= Total N x (Probability P_i for ith interval.)
= N P_i

Decision: Reject the hypothesis if $X^2 > X^2_{\alpha,k-1}$ value on chi-squared distribution table with probability of error α and degrees of freedom = k-1.

Example:

A die is to be tested if each side occurs as equally frequent as the others. To do this, 100 throws of the die was made and the number of occurrences per side were recorded.

X	1	2	3	4	5	6
Frequency	12	10	15	16	22	25

Can we say with a 5% probability of error (or 95% confidence) that the die results follow a uniform distribution?

PRACTICE PROBLEMS:

1. The production records for an automobile manufacturer show the following figures for production per shift.

688	700	691	679
656	701	694	703
711	702	664	708
677	688	630	688
625	688	688	699
703	667	547	697

a. Give a 95% confidence interval for the production output (cars, in this case)
b. Give an 95% confidence interval for the standard deviation of production output that should be known to occur.
c. What proportion of the shifts should you expect to produce 680 cars or more? Give a 95% confidence interval for this proportion.

2. A single leaf was taken from each of Luciano Tang's tobacco plants. Each was divided in half; one half was chosen at random and treated with preparation I and the other received preparation II. The object of the experiment was to compare the effects of the two preparations of mosaic virus on the number of lesions of half

leaves after a fixed period of time. For a 5% level of significance, examine the research hypothesis that the lesions that occurred from different preparations are significantly different. You can do this by making a confidence interval of the differences between each prep.

Plant	1	2	3	4	5	6	7	8	9	10	11
Prep I	18	20	9	14	38	26	15	10	25	7	13
Prep II	14	15	6	12	32	30	9	2	18	3	6

3. DJC Video Stores wants to know how long it takes for its customers to check-out a video rental. Suppose that you are to obtain a random sample of 20 video checkout times (in minutes) . The following table showed the ordered sequence of data collected: (read the data by row)

Day 1	1.12	2.76	3.81	4.91	1.28	5.06	5.67	6.00
Day 2	3.79	4.54	10.28	12.45	7.16	18.12		
Day 3	1.19	0.85	2.15	15.7	1.75	5.12		

Treat each problem below independently. Using a 0.05 level of significance :

a. Assuming normality, give an interval estimate of the time it takes for a customer to check out videos.
b. Assuming normality, give an interval estimate the proportion of customers who check out within 3 minutes.

4. Louie wore XXL sized clothes in June 2002. Today, he can consider himself normal Large sized (Size L). He shows you a month-by-month record of his body weight for the past year. He always weighed himself at the beginning of each month. He wants your opinion on certain statistical claims. He started on a diet program September 1.

Month	*June*	*July*	*Aug*	*Sept*	*Oct*	*Nov*	*Dec*	*Jan*	*Feb*	*Mar*	*Apr*	*May*
Weight	220	216	215	210	206	200	192	187	181	172	162	155

a. Give a 90% confidence interval for the month-to-month differences in his weight since he started on the diet program.
b. Has his average monthly weight significantly changed since he started with the program? Compare his average weight before the program and his average weight after the program. Use a=0.05 Is there a significant reduction?