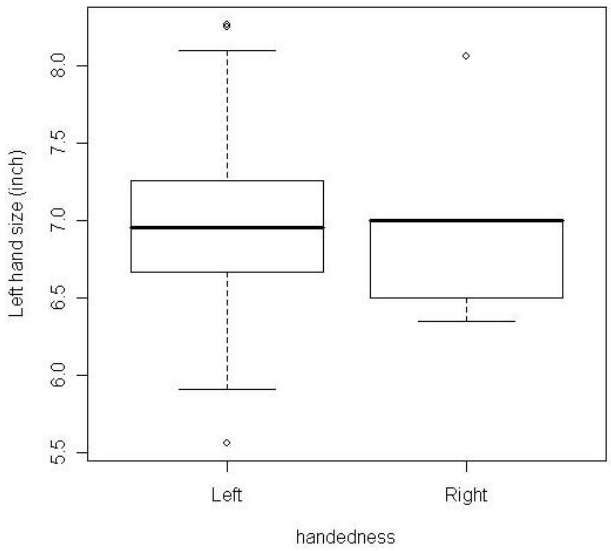
1. The following boxplots show left hand lengths. The boxplot on the left shows data from left handers. The boxplot on the right shows data from right handers.



Which of the following statements is (are) correct about the distributions of the left hand size data for the two handedness groups? Check all that apply. [3 marks]

____ **There are outliers in the distributions of the left hand size data for both handedness groups.**

____ For the distribution among the right handers, the first quartile is the same as the third quartile.

____ More than 25% of the left handers have left hand sizes that are shorter than the minimum left hand size of the right handers.

____ None of the above.

2. Your friend, Carl, has been doing a research project about the coffee drinking habit of UBC students. Last Wednesday during 7:30-8:30 am, he stood next to the Starbucks coffee counter in the SUB and interviewed 30 students who bought coffee there.

The average number of cups of coffee consumed daily by the interviewed students, 1.78, is (check only one): [2 marks]

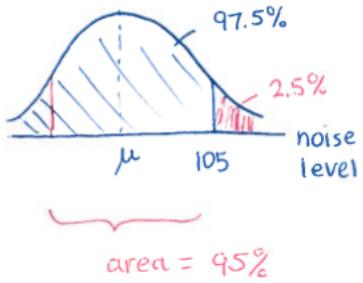
____ a population.

____ a sample.

____ a parameter.

____ **a statistic.**

3. At the present time the noise level per jet takeoff in one neighbourhood near the airport is believed to follow the normal distribution with mean 102 decibels and standard deviation 5 decibels. Suppose a regulation is passed that requires jet noise in this neighbourhood to be lower than 105 decibels 97.5% of the time. How much should the mean noise level be lowered to comply with the regulation? [4 marks]



By 68-95-99.7 rule, 105 is 2 SD's above the mean. So $\mu = 105 - 2(5) = 95$. The mean noise level should be lowered by $102 - 95 = 7$ dB.

4. Circle the most appropriate answer: [2 marks each]

   a) Form a data set that consists of four integer numbers from 1 to 10 (inclusive, without repeats).

   Among all the possible data sets that can be formed (as described in the above), which of the following statements is **NOT** correct?

   i. The set of numbers 1, 2, 3, 4 gives the smallest possible standard deviation.
   ii. The set of numbers 4, 5, 6, 7 gives the smallest possible standard deviation.
   iii. **The set of numbers 1, 5, 6, 10 gives the largest possible standard deviation.**
   iv. The set of numbers 1, 2, 9, 10 gives the largest possible standard deviation.

   b) In linear regression, which of the following features observed in the residual plot indicates that the linear fit is appropriate?

   i. There are the same number of positive residuals as negative residuals.
   ii. **The residuals are randomly scattered.**
   iii. The residuals show a linear pattern.
   iv. Both (i) and (ii).

   c) The length of a ball of yarn is a random variable with mean 150 ft and standard deviation 2 ft. The variance of the average length of sixteen randomly chosen balls of yarn will be

   i. **0.25 ft^2**
   ii. 0.5 ft^2
   iii. 2 ft^2
   iv. 4 ft^2

d) Sixty-four percent of the teenager population is nearsighted. Consider random samples of 4 teenagers drawn from this population. The following are three statements about the sampling distribution of the sample proportion (of 4 teenagers) who do not suffer from nearsightedness.

(I) The sampling distribution has a mean of 36%.

(II) The sampling distribution has a standard deviation of 24%.

(III) The sampling distribution follows the normal distribution approximately.

Which of the above statements is (are) correct?

i. Statement (I) only.

ii. **Statements (I) and (II) only.**

iii. All the three statements.

iv. None of the three statements.

e) There are two urns. Each urn contains 1 black marble and 2 white marbles. You randomly draw one marble from each urn. Let X be the number of white marbles out of the two being drawn. Then

i. **X is a Bin(2,$\frac{2}{3}$) random variable.**

ii. X is a Bin(3,$\frac{2}{3}$) random variable.

iii. X is a Bin(6,$\frac{1}{2}$) random variable.

iv. X is not a Binomial random variable.

5. The following contingency table shows the admission status and gender of applicants for the Master of Arts program in the economics and psychology departments of a national university. There are a total of 855 applicants across the two departments.

	Admitted	Not admitted	Total
Male	177	226	403
Female	177	275	452
Total	354	501	855

a) Write down the conditional distribution of gender among applicants who are admitted. [2 marks]

Male	**Female**	**Total**
177	**177**	**354**
(50%)	**(50%)**	**(100%)**

b) Now we look at the data for the two departments separately:

Economics department			
	# admitted	# applicants	Admission rate
Male	132	268	132/268 = 49.3%
Female	75	148	75/148 = 50.7%

Psychology department			
	# admitted	# applicants	Admission rate
Male	45	135	45/135 = 33.3 %
Female	102	304	102/304 = 33.6 %

Within each department, the admission rate is higher among females. However, when we combine the data from the two departments, males have a higher admission rate (177/403=43.9% for males vs. 177/452=39.2% for females). Briefly explain this phenomenon. [2 marks]

This is Simpson's Paradox. When we take into account the department, the direction of association (admission rates of the 2 gender groups) is reversed.

6. C-section has been one of the many delivery methods used among pregnant women. Let X represent the number of previous C-sections a randomly chosen pregnant woman has had before, and the probability distribution of X is given below (assume that no pregnant women underwent more than two previous C-sections):

x	0	1	2
$P(X = x)$	3a	a	b

for some unknown positive constants a and b.

a) Find the values of a and b if the mean of X is 0.39. [5 marks]

① sum of probabilities of non-overlapping events in sample space = 1

$P(X=0) + P(X=1) + P(X=2) = 1$

$3a + a + b = 1$

$4a + b = 1$

$b = 1 - 4a$

② $E(X) = 0.39$

$\Sigma x\, P(X=x) = 0(3a) + 1(a) + 2(b) = a + 2b = 0.39$

Substituting $b = 1 - 4a$ into $a + 2b = 0.39$ gives

$a + 2(1 - 4a) = 0.39$

$7a = 1.61$

$a = 0.23$

$b = 1 - 4(0.23) = 0.08$

b) Interpret the "mean of X is 0.39" in the context of this question. [2 marks]

It means if we are to repeatedly sample pregnant women with replacement from the population, the long run average number of previous C-sections these women have is 0.39.

7. Cardiovascular fitness is commonly measured by the amount of maximum oxygen uptake during a strenuous exercise. In a study of cardiovascular fitness among athletes, twenty middle-aged men had their maximum oxygen uptake (in milliliters per kilogram weight per minute) during a 2-mile run and run time (in seconds) measured. The summary statistics of the two variables are given below:

$$\begin{aligned} \text{Maximum oxygen uptake} &: \text{Mean} = 49.5,\ \text{SD} = 4.0 \\ \text{Run time} &: \text{Mean} = 842,\ \text{SD} = 50 \\ \text{Correlation coefficient } r &= -0.56 \end{aligned}$$

a) Suppose you and your friend took part in the study. If your friend took 50 more seconds to complete the run than you did, then you would predict his maximum oxygen uptake to be ____2.24____ mL/(kg×minutes) **above/<u>below</u>** (circle one) your maximum oxygen uptake during the run. (Fill in the blank and show your work below.) [3 marks]

The 2 run times differ by 50 seconds (run = 50, i.e. 1 SD_x) so the predicted max oxygen intake should differ by

$$r * SD_y = 0.56(4) = 2.24$$

b) If the run time was measured in minutes instead of seconds, what is the correlation coefficient between the maximum oxygen uptake and the run time? Check only one answer. [2 marks]

____ **-0.56.**
____ Some number between -1 and -0.56.
____ Some number between -0.56 and 0.
____ There is insufficient information to determine.

8. The following table shows the breakdown of the annual salaries of university graduates and the proportion of graduates falling in each income category.

Annual salary range	Proportion of university graduates
under \$20k	0.15
\$20k to \$40k	0.45
\$40k to \$60k	0.30
\$60k to \$100k	0.09
above \$100k	0.01

a) A university graduate is randomly chosen. Here are two events:

Event A : The chosen graduate has an annual salary under \$20k.

Event B : The chosen graduate has an annual salary under \$40k.

Are the two events independent? ____Yes ____ **No**
Explain your reasoning. [3 marks]
P(A) = 0.15
If we are given that B happens, P(A given B) increases to 0.25. The occurrence of B affects the probability of A. Hence, A and B are not independent.

b) Based on the information given in the above table, which of the following is the most plausible value of the median annual salary of university graduates? [2 marks]

____ \$19450
____ **\$37685**
____ \$41283
____ \$50400

c) A random sample of 400 university graduates is drawn from all university graduates. Approximate the probability that 180 or fewer of the sampled graduates earn more than \$40k annually. State any assumption(s) you have made in the calculation. [6 marks]

Let $\hat{p}$ be the sample proportion of graduates who earn more than \$40k annually. Hence we have

$$\hat{p} \sim N\left(p,\, \sqrt{\frac{p\,(1-p)}{n}}\right),$$

where p is the population proportion of graduates who earn more than \$40k annually, and n is sample size (i.e. 400). Based on the table given in the question, we have $p = 0.4$. In sum, we have

$$\hat{p} \sim N\left(0.4,\, \sqrt{\frac{0.4 \cdot (1-0.4)}{400}}\right).$$

Therefore, the probability that 180 or fewer of the sampled graduates earn more than \$40k annually is given by

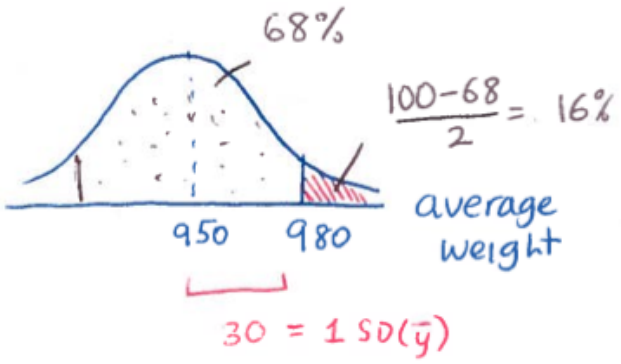
$$\Pr\left(\hat{p} \leq \frac{180}{400}\right) = \Pr\left(z \leq \frac{\frac{180}{400} - 0.4}{\sqrt{\frac{0.4 \cdot (1-0.4)}{400}}}\right) = \Pr\left(z \leq 2.04\right) \approx \Pr\left(z \leq 2\right) = 0.975.$$

9. According to a research study of fish in the Tennessee River, the weights of fish in the river follow the normal distribution with mean 950 grams and standard deviation 180 grams.

   a) True or false? The average weight of a random sample of 36 fish drawn from the Tennessee River follows the normal distribution. This is a result of the Central Limit Theorem. [2 marks]

   ____ True
   ____ **False**

   b) Find the probability that a random sample of 36 fish drawn from the Tennessee River has an average weight of at least 980 grams. [4 marks]



Let y = weight of a fish $\sim N(\mu = 950, \sigma = 180)$

$\bar{y}$ = average weight of 36 fish

$$= \frac{y_1 + y_2 + \ldots + y_{36}}{36}$$

$$\bar{y} \sim N\left(950, \frac{180}{\sqrt{36}} = 30\right)$$

$P(\bar{y} > 980)$ is the area for more than 1 SD($\bar{y}$) above the mean.
Answer = 16%