Statistics I
Distribution of sample proportions

Let X be the number of items (or people) with certain attribute.
Examples:
The number of people with IQ ≥ 95
The number of colour-blind people
The number of successes

Population:
X is not a variable in a population (size N).
It has a constant value c.
Population proportion is given by $\frac{c}{N}$ and it is also a constant value denoted by p.

Take random samples A, B, C, … of the same size n from the population.

A random sample → X is a random variable of value x. Sample proportion is given by $\frac{x}{n}$ and denoted by $\hat{p}$ → of size n.

In random sample A, the sample proportion is $\hat{p}_A$.
In random sample B, the sample proportion is $\hat{p}_B$ etc.

Sample proportion $\hat{p}$ varies from sample to sample because x varies from sample to sample, .: we can consider $\hat{p}_A$, $\hat{p}_B$, … as values of a random variable denoted by $\hat{P}$.

Notations: $\hat{P}$ represents sample proportion random variable and $\hat{p}$ is the value of $\hat{P}$ for a sample.

Taking random samples from a *small* population of size N (sampling without replacement)

X and $\hat{P}$ have the same sampling distribution given by the hypergeometric distribution $\Pr(\hat{P}=\hat{p})=\Pr(X=x)=\dfrac{\binom{c}{x}\binom{N-c}{n-x}}{\binom{N}{n}}$.

The mean and standard deviation of the sample proportion $\hat{P}$ are given by $\mathrm{E}(\hat{P})=\sum \hat{p}\times\Pr(\hat{P}=\hat{p})=p$

and $\mathrm{sd}(\hat{P})=\sqrt{\mathrm{E}(\hat{P}^2)-p^2}$.

Taking random samples from a *large* population (can be approximated as sampling with replacement)

X and $\hat{P}$ have the same sampling distribution and can be approximated by the binomial distribution

$$\Pr(\hat{P}=\hat{p})=\Pr(X=x)=\binom{n}{x}p^x(1-p)^{n-x}$$

The mean and standard deviation of the sample proportion $\hat{P}$ are given by:

$$\mathrm{E}(\hat{P})=p \text{ and } \mathrm{sd}(\hat{P})=\sqrt{\frac{p(1-p)}{n}}$$

If the random sample is large enough ($np>5$ and $n(1-p)>5$), the binomial distribution can be further approximated by a normal distribution with $\mathrm{E}(\hat{P})=p$ and $\mathrm{sd}(\hat{P})=\sqrt{\frac{p(1-p)}{n}}$.

In the discussion above, the population proportion p is a constant and its value is known/given. We use it to predict the distribution of $\hat{P}$ in the random samples.

Inference about the population from a sample

Usually we don't know what the population proportion p is.
To learn about the population we take a random sample of size n from it. $\hat{p}$ of the sample can be used as an estimate of the population proportion p. Here we infer about the population from a random sample.

A random sample → $X=x$ → of size n.

Calculate the sample proportion, one of the statistics of the random sample, by $\hat{p}=\frac{x}{n}$, if it is not given.

The value of $\hat{p}$ is a reasonable estimate of the population proportion p. The larger the sample size n, the better is the estimation.

If the random sample is large enough so that the binomial distribution can be approximated by a normal distribution, a better alternative to $\hat{p}$ as an estimator of p is to give an interval of p values that we are 95% sure contains the actual population proportion p, meaning about 95 out of 100 confidence intervals calculated from the random samples contain the actual population proportion p.

This interval is called a 95% confidence interval for p.

It is $\left(\hat{p}-1.96\sqrt{\frac{p(1-p)}{n}},\ \hat{p}+1.96\sqrt{\frac{p(1-p)}{n}}\right)$ approximately.

However, the population proportion p is generally unknown.

In the absence of any other information, the sample proportion $\hat{p}$ can be used instead of p in the calculation of the confidence interval, e.g. $\left(\hat{p}-1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ for the 95% confidence interval.

Note 1:
The larger the sample size *n*, the distribution is closer to normal and .: the confidence interval is more precise.

Note 2:

$\Pr(-1.96<Z<1.96)\approx\frac{95}{100}$, $\Pr(-2.85<Z<2.85)\approx\frac{99}{100}$

where random variable Z has a standard normal distribution.

Note 3:
In general, an A % confidence interval is approximately $\left(\hat{p}-z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p}+z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ where $\Pr(-z<Z<z)\approx\frac{A}{100}$, $z>0$.

Note 4:
For constant sample size *n*, the higher the required confidence level, the wider is the required interval.

For a required confidence level, the larger the sample size *n*, the narrower is the required interval.

Example 1 (2016 VCAA MM Sample Exam 2 SECTION A Q13)

It is known that 26% of the 19-year-olds in a region do not have a driver's licence.
If a random sample of ten 19-year-olds from the region is taken, the probability, correct to four decimal places, that more than half of them will not have a driver's licence is

A. 0.0239
B. 0.0904
C. 0.2600
D. 0.9096
E. 0.9761

Binomial $n=10$, $p=0.5$:

$\Pr(X>5)=1-\Pr(X\le 5)\approx 0.0239$ A

Example 2 (2016 VCAA MM Sample Exam 2 SECTION A Q14)

An opinion pollster reported that for a random sample of 574 voters in a town, 76% indicated a preference for retaining the current council.
An approximate 90% confidence interval for the proportion of the total voting population with a preference for retaining the current council can be found by evaluating

A. $\left(0.76-\sqrt{\frac{0.76\times0.24}{574}},\ 0.76+\sqrt{\frac{0.76\times0.24}{574}}\right)$

B. $\left(0.76-1.65\sqrt{\frac{0.76\times0.24}{574}},\ 0.76+1.65\sqrt{\frac{0.76\times0.24}{574}}\right)$

C. $\left(0.76-2.58\sqrt{\frac{0.76\times0.24}{574}},\ 0.76+2.58\sqrt{\frac{0.76\times0.24}{574}}\right)$

D. $\left(436-1.96\sqrt{0.76\times0.24\times574},\ 436+1.96\sqrt{0.76\times0.24\times574}\right)$

E. $\left(0.76-2\sqrt{0.76\times0.24\times574},\ 0.76+2\sqrt{0.76\times0.24\times574}\right)$

$\Pr(-1.65<Z<1.65)\approx 0.90$ B

Example 3 (2016 VCAA MM Sample Exam 1 Q7)

A student performs an experiment in which a computer is used to simulate drawing a random sample of size *n* from a large population. The proportion of the population with the characteristic of interest to the student is *p*.
a. Let the random variable $\hat{P}$ represent the sample proportion observed in the experiment. If $p = 1/5$, find the smallest integer value of the sample size such that the standard deviation of $\hat{P}$ is less than or equal to 1/100. 2 marks

Each of 23 students in a class independently performs the experiment described above and each student calculates an approximate 95% confidence interval for *p* using the sample proportions for their sample. It is subsequently found that exactly one of the 23 confidence intervals calculated by the class does not contain the value of *p*.

b. Two of the confidence intervals calculated by the class are selected at random without replacement. Find the probability that exactly one of the selected confidence intervals does not contain the value of *p*. 2 marks

a $\text{sd}(\hat{P})\le\frac{1}{100}$, $\sqrt{\frac{p(1-p)}{n}}\le\frac{1}{100}$ where $p=\frac{1}{5}$

.: $\frac{2}{5\sqrt{n}}\le\frac{1}{100}$, $\sqrt{n}\ge 40$, $n\ge 1600$, .: smallest $n=1600$

b $\Pr(\text{exactly one success})=\Pr(SF)+\Pr(FS)$

$=\Pr(S)\Pr(F\mid S)+\Pr(F)\Pr(S\mid F)=\frac{1}{23}\times\frac{22}{22}+\frac{22}{23}\times\frac{1}{22}=\frac{2}{23}$

Example 4 (2016 VCAA MM Sample Exam 2 SECTION B Q3)

FullyFit is an international company that owns and operates many fitness centres (gyms) in several countries. It has more than 100000 members worldwide. At every one of FullyFit's gyms, each member agrees to have their fitness assessed every month by undertaking a set of exercises called **S**. If someone completes **S** in less than three minutes, they are considered fit.

a. It has been found that the probability that any member of FullyFit will complete **S** in less than three minutes is 5/8 . This is independent of any other member. A random sample of 20 FullyFit members is taken. For a sample of 20 members, let X be the random variable that represents the number of members who complete **S** in less than three minutes.

i. Find $\Pr(X \geq 10)$ correct to four decimal places. 2 marks

ii. Find $\Pr(X \geq 15 \mid X \geq 10)$ correct to three decimal places. 3 marks

For samples of 20 members, $\hat{P}$ is the random variable of the distribution of sample proportions of people who complete **S** in less than three minutes.

iii. Find the expected value and variance of $\hat{P}$. 3 marks

iv. Find the probability that a sample proportion lies within two standard deviations of 5/8. Give your answer correct to three decimal places. Do not use a normal approximation. 3 marks

v. Find $\Pr(\hat{P} \geq 3/4 \mid \hat{P} \geq 5/8)$. Give your answer correct to three decimal places. Do not use a normal approximation. 2 marks

b. Paula is a member of FullyFit's gym in San Francisco. She completes **S** every month as required, but otherwise does not attend regularly and so her fitness level varies over many months. Paula finds that if she is fit one month, the probability that she is fit the next month is 3/4, and if she is not fit one month, the probability that she is not fit the next month is 1/2. If Paula is not fit in one particular month, what is the probability that she is fit in exactly two of the next three months? 2 marks

c. When FullyFit surveyed all its gyms throughout the world, it was found that the time taken by members to complete another exercise routine, **T**, is a continuous random variable W with a probability density function g, as defined below.

$$g(w) = \begin{cases} \dfrac{(w-3)^3 + 64}{256} & 1 \leq w \leq 3 \\ \dfrac{w+29}{128} & 3 < w \leq 5 \\ 0 & \text{elsewhere} \end{cases}$$

i. Find $\mathrm{E}(W)$ correct to four decimal places. 2 marks

ii. In a random sample of 200 FullyFit members, how many members would be expected to take more than four minutes to complete **T**? Give your answer to the nearest integer. 2 marks

d. From a random sample of 100 members, it was found that the sample proportion of people who spent more than two hours per week in the gym was 0.6 Find an approximate 95% confidence interval for the population proportion corresponding to this sample proportion. Give values correct to three decimal places. 1 mark

a i Binomial: $n = 20$, $p = \dfrac{5}{8}$, $\Pr(X \geq 10) \approx 0.9153$

a ii $\Pr(X \geq 15 \mid X \geq 10) = \dfrac{\Pr(X \geq 15)}{\Pr(X \geq 10)} \approx \dfrac{0.1788}{0.9153} \approx 0.195$

a iii $\mathrm{E}(\hat{P}) = p = \dfrac{5}{8}$, $\mathrm{Var}(\hat{P}) = \dfrac{p(1-p)}{n} = \dfrac{\frac{5}{8} \times \frac{3}{8}}{20} = \dfrac{3}{256}$

a iv $\sigma = \sqrt{\dfrac{3}{256}} = \dfrac{\sqrt{3}}{16}$

$\dfrac{5}{8} - 2 \times \dfrac{\sqrt{3}}{16} \approx 0.4085$, $\dfrac{5}{8} + 2 \times \dfrac{\sqrt{3}}{16} \approx 0.8415$

$\Pr(0.4085 < \hat{P} < 0.8415) = \Pr(20 \times 0.4085 < X < 20 \times 0.8415)$

$= \Pr(9 \leq X \leq 16) \approx 0.939$

a v $\Pr\left(\hat{P} \geq \dfrac{3}{4} \mid \hat{P} \geq \dfrac{5}{8}\right) = \Pr(X \geq 15 \mid X \geq 12.5) = \dfrac{\Pr(X \geq 15)}{\Pr(X \geq 13)}$

$\approx \dfrac{0.1788}{0.5079} \approx 0.352$

b $\Pr(FFF') + \Pr(FF'F) + \Pr(F'FF)$

$= \dfrac{1}{2} \times \dfrac{3}{4} \times \dfrac{1}{4} + \dfrac{1}{2} \times \dfrac{1}{4} \times \dfrac{1}{2} + \dfrac{1}{2} \times \dfrac{1}{2} \times \dfrac{3}{4} = \dfrac{11}{32}$

c i $\mathrm{E}(W) = \displaystyle\int_1^3 \frac{w\left((w-3)^3 + 64\right)}{256}\, dw + \int_3^5 \frac{w(w+29)}{128}\, dw$

$\approx 0.978125 + 2.0677083 \approx 3.0458$ min

c ii $\Pr(W > 4) = \displaystyle\int_4^5 \frac{w+29}{128}\, dw \approx 0.261719$

Expected number of members $\approx 200 \times 0.261719 \approx 52$

d $\left(0.6 - 1.96\sqrt{\dfrac{0.6 \times 0.4}{100}}, 0.6 + 1.96\sqrt{\dfrac{0.6 \times 0.4}{100}}\right) \approx (0.504, 0.696)$