5. Sampling and sampling distributions

5.1 Random sampling

A random sample on a random variable $X$, is a sequence of independent random variables $X_1, X_2, \ldots, X_n$ each having the same distribution as $X$.

Realisations of the random variables $X_1, X_2, \ldots, X_n$ are denoted by $x_1, x_2, \ldots, x_n$.

Example A random sample of 100 on $X$, where $X \overset{d}{=} N(\mu=10, \sigma^2=4)$ [the population is specified by $X$] consists of the independent random variables $X_1 \overset{d}{=} N(\mu=10, \sigma^2=4), X_2 \overset{d}{=} N(\mu=10, \sigma^2=4), \ldots, X_{100} \overset{d}{=} N(\mu=10, \sigma^2=4)$.

The observed sample consists of realisations of these random variables:

$x_1 = 11.43, \ x_2 = 8.27, \ldots, \ x_{100} = 9.19$.

From a random sample on $X$, we wish to make inferences about the distribution of $X$. We wish to estimate the pmf (or the pdf) and the cdf; we wish to estimate measures of location and spread, or other characteristics of the distribution of $X$.

To do this, we use statistics:

a statistic is a function of the sample variates:

$$W = \psi(X_1, X_2, \ldots, X_n)$$

For example, the sample mean:

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

Also, the sample median, the sample standard deviation, the sample interquartile range, etc., are statistics which are used as estimators of their population counterparts.

$W$ is a random variable; its realisation is given by

$$w = \psi(x_1, x_2, \ldots, x_n).$$

For example, the sample mean:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_{100}) = \frac{1}{100}(11.43 + 8.27 + \cdots + 9.19) = 10.13.$$

A statistic has a dual role: a measure of a sample characteristic and an estimator of the corresponding population characteristic.
Each of the statistics we have met can be regarded as an estimator of population characteristic, usually referred to as a parameter. Thus the statistic \( \bar{X} \) is an estimator of the parameter \( \mu \); the statistic \( \hat{c}_q \) is an estimator of \( c_q \), and so on. We start with the mean.

If the statistic \( W \) is an estimator of the parameter \( \theta \), then in order to make inferences about \( \theta \) based on \( W \) we need to know something about the probability distribution of \( W \).

Thus, we turn to consideration of the probability distributions of some of the statistics defined above.

### 5.2 The distribution of \( \bar{X} \)

\[ \bar{X} \text{ gives an estimate of the value of } \mu. \]

But what other values of \( \mu \) are likely? plausible? possible?

The answer is provided by the distribution of \( \bar{X} \).

**Remember:**
- mean of a sum = sum of the means
- variance of a sum = sum of the variances  
  (for independent random variables)

\[ E(cX) = cE(X) \text{ and } \text{var}(cX) = c^2\text{var}(X) \]

Hence:

\[ E(\bar{X}) = E\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{1}{n}(n\mu) = \mu \]

\[ \text{var}(\bar{X}) = \text{var}\left(\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right) = \left(\frac{1}{n}\right)^2(\sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \]

**E(\bar{X}) = \mu \text{ and var}(\bar{X}) = \frac{\sigma^2}{n}**

Further, from the Central Limit Theorem (which says that the sum of a lot of independent variables is approximately normal) we have:

\[ X \overset{d}{\approx} N(\mu, \frac{\sigma^2}{n}) \]

and this approximation applies (for large \( n \)) no matter what the population distribution.

**Example** sample of \( n = 100 \) on \( X \overset{d}{=} N(\mu=10, \sigma^2=4) \): \( \bar{X} \overset{d}{\approx} N(10, \frac{4}{100}) \).

**Example** sample of \( n = 100 \) on \( X \overset{d}{=} \text{Poi}(10) \): \( \bar{X} \overset{d}{\approx} N(10, \frac{100}{100}) \).

**Example** sample of \( n = 100 \) on \( X \overset{d}{=} \text{exp}(0.5) \): \( \bar{X} \overset{d}{\approx} N(2, \frac{4}{100}) \).
5.3 Inference on the population mean, $\mu$

| population mean $\mu$ | $\rightarrow$ | sample mean, $\bar{X}$ |

Example A random sample of $n = 40$ observations is obtained on $X$, where $X$ has a Poisson distribution. $\lambda$ is supposed to be 10, and so the population mean is supposed to be 10. We observe $\bar{x} = 12.4$. What do you think about the idea that $\lambda = 10$?

If $X \overset{d}{=} P(10)$, then $\bar{X} \overset{d}{=} N(10, \frac{10}{40})$. Therefore a 95% probability interval for $\bar{X}$ is $10 \pm 2 \times \frac{1}{2} = 9 < \bar{X} < 11$. So it is unlikely to observe $\bar{x} = 12.4$.

$Pr(\bar{X} \geq 12.4) = Pr(\bar{X} > \frac{12.4 - 10}{0.5}) = Pr(N > 4.8) = 0.00000079 \approx 0$.000

If everything else is right (random sample from a Poisson distribution) then the assumption that $\lambda = 10$ looks very dubious.

Example A random sample of $n = 400$ observations is to be obtained from a population with mean $\mu = 50$ and standard deviation $\sigma = 10$. Specify the approximate distribution of the sample mean and hence find approximately $Pr(49 < \bar{X} < 51)$.

$\bar{X} \overset{d}{=} N(50, \frac{100}{400})$; since $n = 400$, $\mu = 50$ and $\sigma^2 = 10^2$;

$Pr(49 < \bar{X} < 51) = Pr(-2 < \bar{X} < -2) = 0.9544$.

Given $\mu$ we can make a statement about $\bar{X}$:

$Pr\left(\mu - 2 \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 2 \frac{\sigma}{\sqrt{n}}\right) \approx 0.95$.

This means that, given $\bar{x}$ we can make a statement about $\mu$.

$Pr\left(\bar{X} - 2 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2 \frac{\sigma}{\sqrt{n}}\right) \approx 0.95$.

i.e. if $\bar{X}$ is within $\epsilon$ of $\mu$, then $\mu$ is within $\epsilon$ of $\bar{X}$.

Example A random sample of $n = 400$ observations is to be obtained from a population with standard deviation $\sigma = 10$. If we observed the sample mean, $\bar{x} = 50.8$, what are plausible values for the unknown population mean $\mu$?

We have $\bar{X} \overset{d}{=} N(\mu, \frac{100}{400})$, i.e. $X \overset{d}{=} N(\mu, 0.5^2)$.

$\therefore \quad Pr(\mu - 2 \times 0.5 < \bar{X} < \mu + 2 \times 0.5) \approx 0.95$

i.e. $Pr(\mu - 1 < \bar{X} < \mu + 1) \approx 0.95$.

Hence $Pr(\bar{X} - 1 < \mu < \bar{X} + 1) \approx 0.95$

So (“95%”) plausible values for $\mu$ are $50.8 \pm 1$, i.e. $(49.8 < \mu < 51.8)$.

If $\mu = 49.8$, then what we observed ($\bar{x} = 50.8$) would be just on the upper “plausible” limit for $\bar{X}$ values.

If $\mu = 51.8$, then what we observed would be just on the lower “plausible” limit.

This set of (95%) plausible values is called a (95%) confidence interval for $\mu$. 
There is one other problem if this approach is to be used to estimate an unknown population mean $\mu$:
If $\mu$ is unknown, then often $\sigma$ will be too. So, in most cases, we don’t know the value of $sd(\bar{X})$!

What to do? If $\sigma$ is unknown, then we estimate it. So:

$$sd(\bar{X}) = \frac{\sigma}{\sqrt{n}} \approx \frac{\hat{\sigma}}{\sqrt{n}} = se(\bar{X})$$

i.e. replace the unknown parameter $\sigma$ by an estimate.

**The estimate of the standard deviation of $\bar{X}$ is called the standard error of $\bar{X}$.**

Usually, but not always, $\hat{\sigma} = s$.

**Example** Suppose that nothing at all is known about the population!
If a random sample of $n$ observations is obtained from this population, then $\bar{x}$ is an estimate of the population mean $\mu$, with standard error $\frac{s}{\sqrt{n}}$.

**Example** Suppose that it is known that the population has a Poisson distribution; i.e. $X \overset{d}{=} \text{Pn}(\lambda)$ for some unknown value of $\lambda$.
Now the population has a structure and our aim is to estimate $\lambda$.
If a random sample of $n$ observations is obtained from this population, then $\bar{x}$ is an estimate of $\lambda$.
In this case $var(\bar{X}) = \frac{\lambda}{n}$, since $\sigma^2 = \lambda$ for $X \overset{d}{=} \text{Pn}(\lambda)$. So to estimate $var(\bar{X})$, we need to estimate $\lambda$, for which we use $\bar{x}$. Thus, $se(\bar{x}) = \sqrt{\frac{\bar{x}}{n}}$.

This gives a recipe for an approximate 95% confidence interval applicable in many situations:

**approx 95% CI:** $\text{est} \pm 2 \times se$

**Example**

$X \overset{d}{=} N(\mu, \sigma^2) \Rightarrow var(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow se(\bar{X}) = \frac{s}{\sqrt{n}}$,

$X \overset{d}{=} \text{Pn}(\lambda) \Rightarrow var(\bar{X}) = \frac{\lambda}{n} \Rightarrow se(\bar{X}) = \sqrt{\frac{\bar{x}}{n}}$.

**Example** A random sample of $n = 20$ observations on $X \overset{d}{=} \text{Pn}(\lambda)$ yields $\bar{x} = 56$. Find an approximate 95% confidence interval for $\lambda$.

$\hat{\lambda} = 56$, $se(\hat{\lambda}) = \sqrt{56/20} = 0.53$;

approx 95% CI: $5.6 \pm 2 \times 0.53 = (4.54, 6.66)$. 
5.4 Statistics as estimators

Sample characteristic = statistic
(1) describes the sample
(2) estimates the corresponding popln characteristic
Population characteristic = parameter

<table>
<thead>
<tr>
<th>Sample statistic</th>
<th>→</th>
<th>Population parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample mean $\bar{x}$</td>
<td>→</td>
<td>Population mean $\mu$</td>
</tr>
<tr>
<td>Sample variance $s^2$</td>
<td>→</td>
<td>Population variance $\sigma^2$</td>
</tr>
<tr>
<td>Relative frequency $\frac{\text{freq}(A)}{n}$</td>
<td>→</td>
<td>Probability $\Pr(A)$</td>
</tr>
<tr>
<td>Sample quantile $\hat{c}_q$</td>
<td>→</td>
<td>Population quantile $c_q$</td>
</tr>
</tbody>
</table>

Point estimation

An estimate of (a parameter) $\theta$ is the observed value of the estimator. This is a point estimate — it is a “guess” (hopefully an informed guess) at the unknown value of $\theta$.

Example  A sample from a $N(\mu, \sigma^2)$ population gives:
13.9, 17.4, 14.3, 15.7, 16.2, 14.7, 16.0, 14.8, 15.2
An estimate of $\mu$ is $\bar{x} = 15.36$.

Interval estimation

Generally, a point estimate alone is not sufficient indication of the unknown value of $\theta$. We also wish to know by how much the estimate might be in error, or better still to give an interval within which we are, say, 95% sure that the unknown value of $\theta$ will lie.

This leads to the concept of interval estimation, or confidence intervals.

Example  A sample from a $N(\mu, \sigma^2)$ population gives:
13.9, 17.4, 14.3, 15.7, 16.2, 14.7, 16.0, 14.8, 15.2
An estimate of $\mu$ is $\bar{x} = 15.36$, with standard error 0.36. 
(obtained using $\frac{s}{\sqrt{n}} = \frac{1.085}{\sqrt{9}}$)
An approximate 95% confidence interval for $\mu$ is $(14.64, 16.08)$. 
(obtained using $\text{est} \pm 2\text{se}$)

We have seen how an approximate confidence interval can be obtained from the standard error. In a number of cases we can do better. This is considered in more detail in Chapter 6, where we deal with sampling from a Normal population.
Appendix  (supplementary information: not required for examination)

Methods of estimation

Often the estimate is obvious (as when we use the sample characteristic to estimate the population characteristic) and we will stick with the “obvious”. However, in some cases it’s not so obvious, and something more is required.

There is a range of possible methods:

1. Method of moments — equating population moments and sample moments (mean, variance, …);

2. Method of quantiles — equating population quantiles and sample quantiles (median, interquartile range, …);

3. QQ-plots — fitting a straight line to a plot of sample quantiles against standard population quantiles [mentioned in Chapter 6];

4. Least squares — minimising a sum of squares \( \sum (x_i - \mu_i(\theta))^2 \) [see Chapter 10];

5. Maximum likelihood — choosing the parameter value which makes the data most probable: the most likely explanation of the data, as my mate Holmes would have it.