# Chapter 5: Sampling Distributions

### Introduction

Statistical inference is used to draw conclusions about a population or proces based on data. The data is received by tests, such as means, proportions and slopes of regression lines. For making future predictions based on data derived from samples and using probability, there are several steps and possibilities.

### 5.1: Statistical Inference

*Statistical inference* is using facts about a sample to draw conclusions or make predictions about a population.

- A *parameter* is a number that describes the population. A parameter is a set number, but we do not actually know its value.  For example, it is impossible to know exactly how many Dutch people are against abortion.
- A *statistic* is a number that describes a sample. The value of a statistic known after we have selected a sample, but this value can differ per sample as well. We often use a statistic to estimate an unknown parameter.

### Sampling Variability

*Sampling variability* means that the value of a statistic per sample will be different. Random samples remove bias by selecting a sample based on randomness; however, it appears that the selection of many random samples (of the same size and from the same population) the variation between samples follow a predictable pattern. Statistical inferences are based on the idea that the reliability of sampling depends on the repetition of procedures.

In practice, it is too expensive to run an unlimited number of trials; but we are able to imitate taking many random samples by using *simulation.*

### Sampling Distribution

The *sampling distribution* of a statistic is the distribution of all the values that adopt the statistic in all possible samples of the same size and from the same population. If this distribution is plotted on a histogram, it appears that:

- The histogram has a normal distribution
- The histogram also shows that the means and medians are roughly the same.

- In practice, it appears that values from samples of considerable size (e.g.,> 2500) have much less spread than the values of smaller samples (e.g., having a size of 100). This is because larger samples are more representative of the population than smaller samples.

These three facts are true whenever we use random sampling.

**Bias and Variability**

- A statistic that describes a parameter is *unbiased* if the average of the corresponding samples of distribution is equal to the true value of the estimated parameter. Bias is reduced by using random sampling techniques.
- The *variability* of a statistic is described by the spread of its sampling distribution. This spread is determined by the design and the sample size of the sample (n). Variability is reduced by having larger sample sizes.
- The *margin of error* is a measure of the spread of a sampling distribution
- Low bias can coexist with great variability and low variability may be associated with a lot of bias. In a good research, there is little spread and minimal bias.

So far, we have learned the following concepts:

- A test of a random sample or a randomized experiment is a random variable.
- The *sampling distribution* of a statistic shows how the statistic (such as a mean) will vary if it were repeatedly tested.
- The *population distribution* of a variable is a distribution that contains all the values that a variable can take for each on member of the population. The population distribution is also the probability distribution of a random variable when we choose one random individual from the population.

**5.2 The Sampling Distribution of a Sample Mean**

When data is derived from randomly dedacted samples, then a test is a random variable that can be understood using the rules of chance:

- A test of a random sample or a randomised experiment is a random variable.
- The *sampling distribution* shows how a test (for instance a mean) will vary when repeated samples are made.
- The *population distribution* of a variable is a distribution that contains all scores a variable can get for members of the population.

**Sample Means**

*Sample means* are averages of observations. There are two important things to remember about them:

- Sample means are less variable than individual observations.
- Sample means are more normally distributed than individual observations.

**The Mean and Standard Deviation of $\bar{x}$**

The sample mean $\bar{x}$ is an estimate of the underlying mean, μ, a population. The distribution of samples is determined by (1), the design that is used to collect data, (2) the sample size n, and (3), the population distribution.

The sample mean for an SRS is:

$\bar{x}$ = 1/n ($X_1$ + $X_2$ + $X_3$ + … + $X_n$), where n is the sample size.

The standard deviation of the sample mean is:

$\bar{x}$ = σ/$\sqrt{n}$ .

In short, the sample mean is the same as the population mean, because $\bar{x}$ is an unbiased predictor of μ. The standard deviation of a sample is the population standard deviation divided by the square root of the number of participants.

**The Central Limit Theorem**

If the distribution of a population is normal, the distribution for the sample mean is also normal:

- If a population has a distribution of N (μ, σ), the sample mean $\bar{x}$, of n observations has a distribution of N (μ., Σ / $\sqrt{n}$ ).

In practice, however, many populations are not normally distributed, however the *central limit theorem* states that for any sample with a large enough n,

- $\bar{x}$ is approximately N (μ, σ/$\sqrt{n}$ ).

In short, if your sample size is large enough, the distribution of your sample will be approximately normal. There are three other important things to keep in mind with regard tot he central limit theorem:

- Any linear combination of independent normal random variables also has a normal distribution

- The distribution of a sum or average of many small random quantities is also normally distributed.
- The central limit theorem also applies to discrete variables

**5.3 Sampling Distribution for Counts and Proportions**

A random variable X is a count if we count the number of occurrences of a certain outcome. For example you can count how many people answer "yes" to the question ofwhether prostitution should be legal.

- If the number of observations is n, then the *sample proportion* is X / n., where X stands for the number of people who support the legalization of prostitution.

**The Binomial Distribution**

The binomial setting has a specific set of characteristics:

- There are n number of observations, and the observations are independent
- Each observation falls into one of two categories. These categories are called for convenience 'success' and 'failure'.
- The probability of a success (p) is the same for each observation.

Binomial Distributions also have a specific set of characteristics:

- The distribution of X (the count of the number of successes in a binomial setting) is entirely determined by the number of observations (n) and the chance of success (p).
- The possible values of X are whole numbers between 0 and n.
- We denote the binomial distribution as: B (n, p)
- The binomial distribution is important if we want to draw conclusions about the population on the proportion of 'successes'. Choosing an SRS from a population, however, is not really a binomial situation.
- A population contains proportion p of successes. If the population is much larger than the sample, the count X (number of successes in a size of n SRS) has approximately the binomial distribution B (n, p). The accuracy of this approximation increases as the population size increases. The binomial distribution is used when the population size is at least 20 times as large as the sample size.

**Finding Binomial Probabilities**

Often, binomial probabilities are calculated by using software. It is also possible to calculate these manually by using table C in the back of the book. To use this table the probability of individual outcomes for the binomial random variable X

should be known.

**Binomial Mean and Standard Deviation**

If a count, X, has the binomial distribution B(n, p), the mean can be found by using the following formula:

$\mu_x = np$

The standard deviation can be found using:

$$\sigma_x = \sqrt{np(1-p)}$$

**Sample Proportions**

What percentage of adults are in favour of abortion? Using sampling distributions, we often want to estimate the proportion p of 'successes' in a population. The proportion of successes in a population can be estimated using:

- $\hat{p}$ = X/n

It is important to know that X here is not the same as a count, X. The count X assumes a whole number between 0 and n, but a proportion is always a number between 0 and 1. In a binomial setting, the count X has a more-or-less binomial distribution, while the population, $\hat{p}$ does not.

We can, however, perform probability calculations for $\hat{p}$ by writing them in terms of count X and using binomial calculations.

**Mean and Standard Deviation**

The mean and standard deviation of a proportion of successes of an SRS, $\hat{p}$, can be found by using the following formulas:

The mean of $\hat{p}$ = p

The standard deviation of $\hat{p}$ = $\sqrt{p(1-p)/n}$

where p is the proportion of successes.

**Normal Approximation for Counts and Proportions**

The sampling distribution of a sample proportion, $\hat{p}$, is approximately a normal distribution. Now we also know that the distribution of $\hat{p}$ is that of a binomial count divided by sample size. n:

- X is approximantely N (np, $\sqrt{np\,(1-p)}$)
- $\hat{p}$ is approximately N(p,$\sqrt{p\,(1-p)\,/n}$)

This applies when n is large. As a rule of thumb, this approach is used only for values n and p that satisfy the conditions np ≥ 10 and n (1-p) is ≥ 10.

**Binomial Formulas**

A *continuity correction factor* is use in cases where a continuous function is used to approximate a discrete function. One example of this is using a normal distribution to approximate a binomial. For a detailed description, see Fig. 5.15 in the textbook.

In order to predict the occurrence of a binomial random variable, we use the following formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

This formula gives us the *binomial coefficient*, where k is the number of successes and n is the number of observations.

- n! = n x (n-1) x (n-2) x ... x 3 x 2 x 1. 0! = 1.
- It is also important to note that the notation $\binom{n}{k}$ has nothing to do with the fraction n/k. This is, instead, read as "binomial coefficient n choose k"
- If X is the binomial distribution B (n, p) with n observations and p chance of success for each observation, the *binomial probability* is:

  $$P(X = k) = \binom{n}{k} p^k (1\text{-}p)^{n-k}$$

  where k is any of the possible values of X.

**Poisson Distributions**

A count X has a binomial distribution when it is produced in a binomial setting. If one or more facets of these settings do not hold true, the count X will have a different distribution. The Poisson Distribution is one of these cases and is used in open-ended situations. It can be used in the following situations:

- The number of successes that exists in two non-overlapping units of measure are independent
- The probability of a success occurs in a unit of measurement is the same for all units of the same size and is proportional to the size of the unit.
- The probability that more than one event occurs in a measurement unit is negligible for very small units. In other words: the events happen one by one.

The Poisson Distribution can be expressed using the following formula:

$$\mathrm{P}[x = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

where λ is the mean and k is the number of successes that occur in the experiment.

When the mean of the Poisson distribution is large, it may be difficult to calculate Poisson probabilities, even with a calculator or software. Fortunately, when μ is large, calculations can be made using the Normal Distribution.