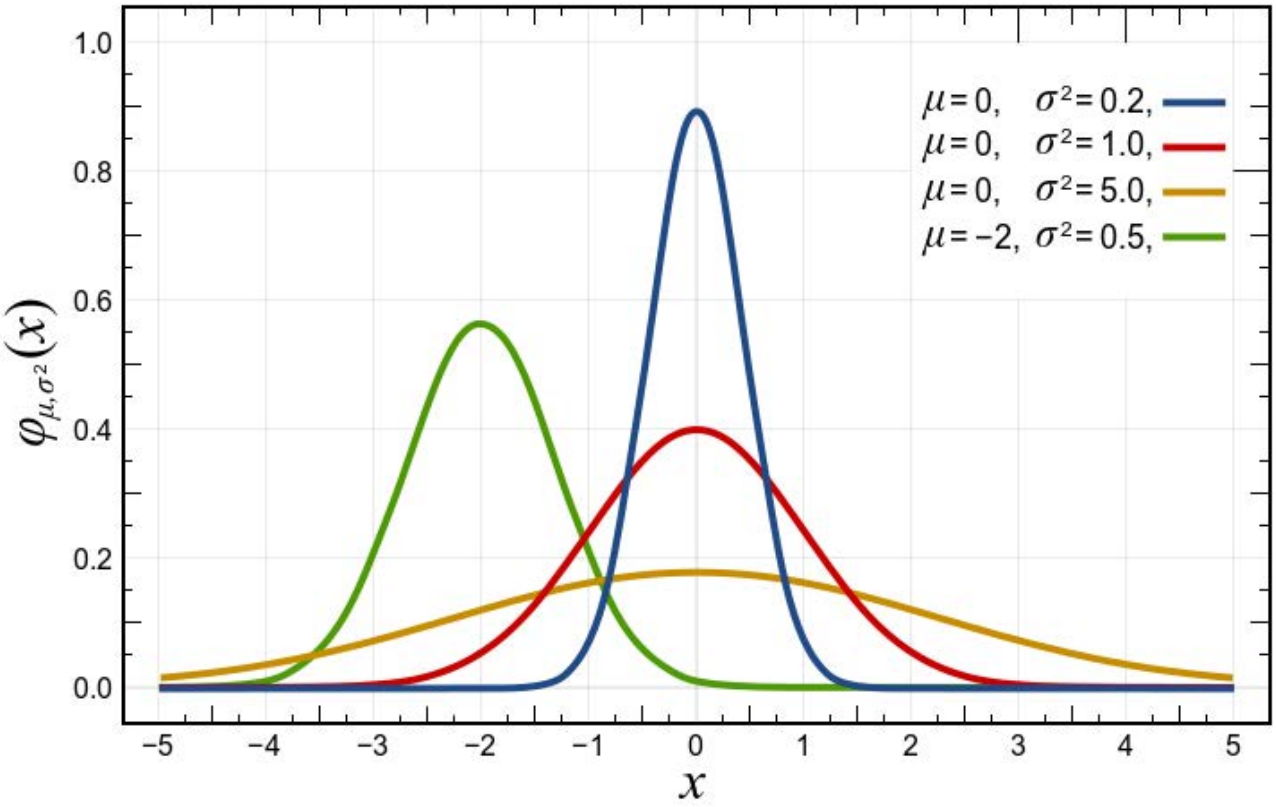
Applied Statistics using R

FACTSHEET

UNIT 2



Medical Statistics Team

University of Aberdeen

Copyright Notice

Contents

Chapter 1

Probability Definitions

1.1 Concept

An experiment is an activity whose outcome is uncertain.

An event consists of one or more possible outcomes.

The probability of any event, A, is denoted P(A).

- The probability of any outcome is always between 0 and 1.
- Some outcomes may be equally likely, e.g. heads or tails on a fair coin.
- The sum of the probabilities of all possible outcomes is always 1.

1.2 Example

Throwing a fair die once

The possible outcomes are (1, 2, 3, 4, 5 or 6) and all six outcomes are equally likely to occur with probability 1/6.

The sum of the probabilities of all possible outcomes:

$P(all\ events) = P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 6 \times 1/6 = 1$

If event (A) is getting a number 3 or less on the die, then the probability of event A occurring:

$P(A) = P(1\ or\ 2\ or\ 3\ showing\ on\ the\ die) = 3/6 = 0.5$

1.3 Addition rule

The addition rule states that:

$P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$

1.4 Mutually exclusive

Two events are mutually exclusive if one happening means that the other cannot happen (cannot have head and tail from one toss of one coin).

If events A and B are mutually exclusive then:

$P(A \text{ or } B) = P(A) + P(B)$

Note: this is the same as the addition rule, but the last part P(A and B) is zero.

1.5 Independent events

Two events are independent if one happening has no bearing on whether the other happens or not. If A and B are independent then the probability of both happening is the multiplication of $P(A)$ and $P(B)$ then:

$P(A \text{ and } B) = P(A) \times P(B)$

1.6 Conditional probability

Independence is a strong assumption. Consider instead the probability of one event happening given that other has occurred. This is covered by Bayes' theorem (or rule or law).

The probability of A given B has occurred is:

$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$

Conditional probability is very useful for evaluating diagnostic tests through measures discussed later in the course.

Chapter 2

Probability Distributions

2.1 Normal distribution

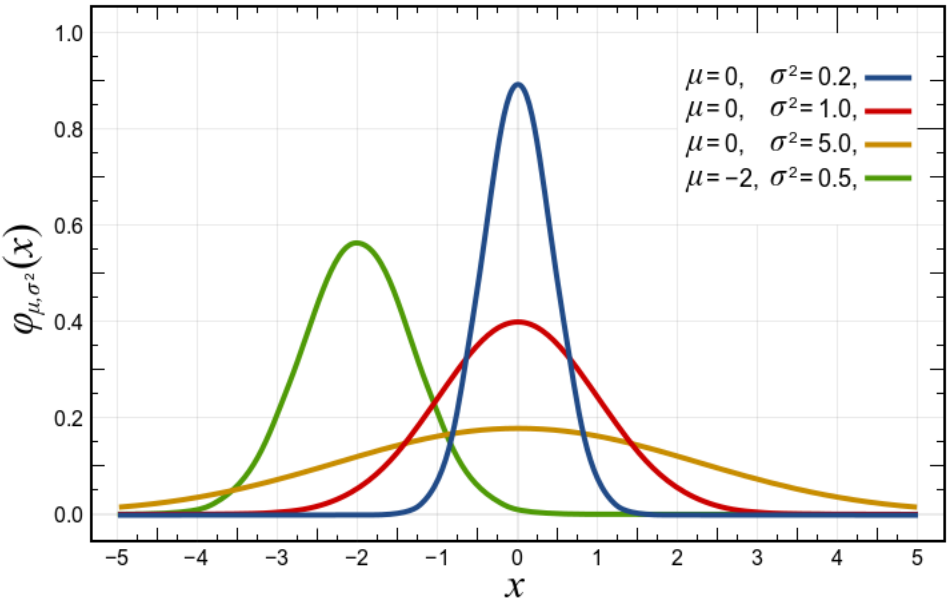


Image credit: Wikipedia

- A smooth bell-shaped curve $(-\infty \; to + \infty)$
- Symmetrical about μ with spread of σ
- The probability is the area under the curve
- X is the value of interest
- μ is the population mean
- σ is the population standard deviation
- $\pm 1.96\sigma$ relates to approx. 95% of the population

2.2 Standard Normal distribution

The Z-score is defined as $Z = \frac{X-\mu}{\sigma}$

- Z is value of interest
- $Mean = 0$
- $SD = 1$
- ± 2 relates to approx. 95% of the population
- More accurately this is at $Z = \pm 1.96$

2.3 Students t-distribution

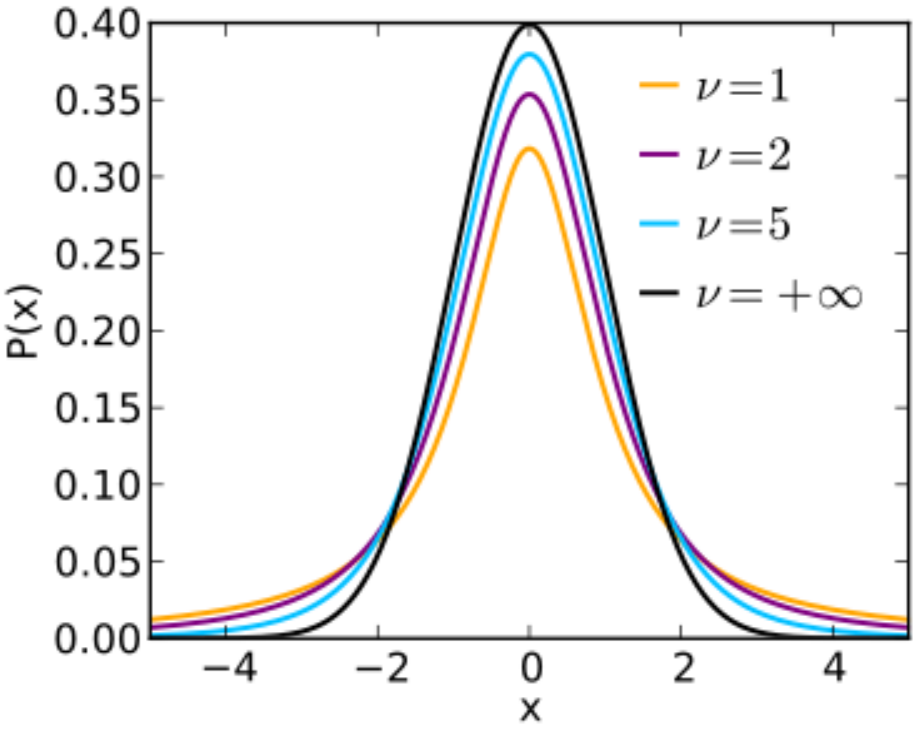


Image credit: Wikipedia

- A symmetrical curve ($-\infty$ *to* $+\infty$)
- One of the parameters are degrees of freedom (related to sample size)
- Again the probability is the area under the curve
- As degrees of freedom increase, t-value approaches the z-value

2.4 Chi-squared distribution

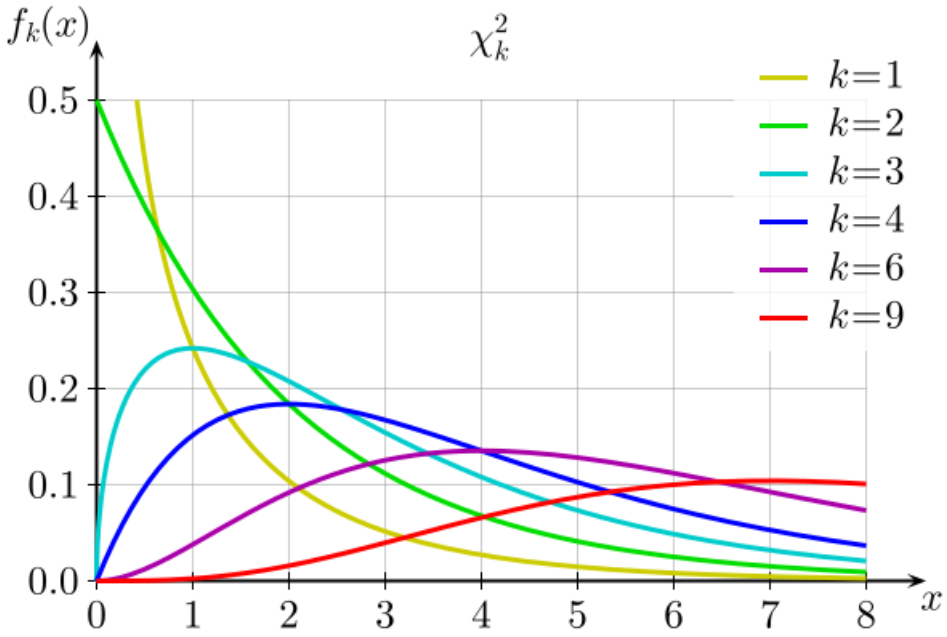


Image credit: Wikipedia

- A fat, chubby, skewed distribution ($0\ to + \infty$)
- One of the parameters are degrees of freedom

2.5 F-distribution

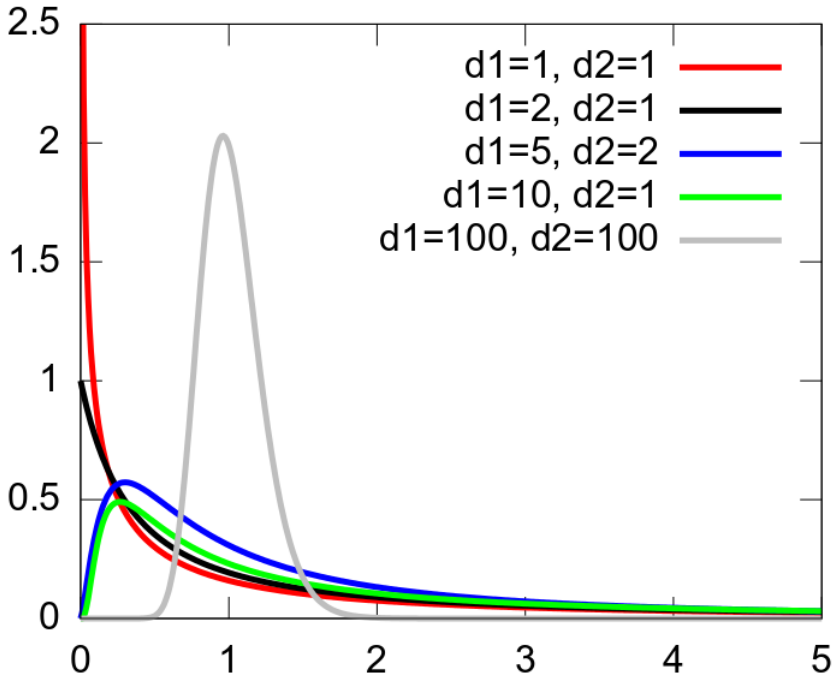


Image credit: Wikipedia

- This is also a skewed distribution ($0\ to + \infty$)
- Two important parameters are: numerator and denominator degrees of freedom

2.6 Using R

The following information is not required for the course.

If you are interested about distribution functions in R, it has four types of functions for each distribution with appropriate parameters as its arguments. Check the help file for further details.

- Density function (`d`)
  - Normal distribution: `dnorm`
  - t distribution: `dt`
  - Chi-squared distribution: `dchisq`
  - F distribution: `df`
- Distribution function (`p`)
  - Normal distribution: `pnorm`
  - t distribution: `pt`
  - Chi-squared distribution: `pchisq`
  - F distribution: `pf`
- Quantile function (`q`)
  - Normal distribution: `qnorm`
  - t distribution: `qt`
  - Chi-squared distribution: `qchisq`
  - F distribution: `qf`
- Generate random number from the distribution (`r`)
  - Normal distribution: `rnorm`
  - t distribution: `rt`
  - Chi-squared distribution: `rchisq`
  - F distribution: `rf`

Chapter 3

Sampling

3.1 What is it?

A *census* is the inclusion of every member in the population. *Sampling* is the selection of a small, but representative, group of individuals from the population for study so that we can generalise the findings from the sample to the population.

3.2 When is it appropriate?

Sampling is useful in order to reduce the cost of the research, to speed up the data collection and analysis process. It is also useful when we have limited resources such as trained personnel or specialised equipment.

3.3 Types of samples:

Non-probability samples such as convenience sample, judgment sample, referral sample and quota samples involve systematic error in the selection. Because the probability of member inclusion in the sample is not known, the representativeness of the sample to the population is not known in non-probability samples.

On the other hand, every element in population has a known, non-zero probability of being included in the **probability samples**.

3.4 Some definitions:

- *Population*: Complete set of individuals to which we want to generalise the study findings
- *Sampling frame*: List of every unit in the population
- *Sample*: Subset of a population used to study the population as a whole

- *Sample size*: Number of individuals or observations in the sample
- *Sampling method*: Procedures to select a sample that primarily determine the generalisability of research findings

3.5 Some frequently used probability sampling methods

- **Simple random sampling**

A random sample of required sample size is selected using random numbers from a sampling frame. Each member of the population has an equal chance of being included in the sample.

- **Systematic sampling**

A sampling interval is calculated by dividing the population size with the sample size required. A number between 1 and the sampling interval is randomly selected it becomes the first item in the sample. The sampling interval is added to the position of previously selected item to identify the next item to be included in the sample. This process is continued until the required sample size is achieved. This method is easy to apply even when the sampling frame is unavailable.

- **Stratified sampling**

Population is separated into mutually exclusive, homogeneous groups called strata. There is little variation among members within each stratum but wide variation between strata. A random sample is selected within each stratum. Stratified sampling ensures that the population variability is adequately represented in the sample.

- **Cluster sampling**

Population is divided into natural groups called clusters. Clusters are heterogeneous and as such are mini populations. One or more clusters are selected randomly. Cluster sampling is administratively convenient.

3.6 Using R

The following information is not required for the course.

The R function `sample` takes a random sample of the specified size from the elements of `x` using either with or without replacement. Check the help function for further details.

Statisticians also use the standard random number generator function like `rnorm` to generate random samples from a normal distribution with given mean and standard deviation.

Chapter 4

Estimation, Standard Errors and Confidence Intervals

4.1 Inference

This is how we draw conclusions from sample data about larger population from which sample data selected.

4.2 Population parameters & sample statistics (estimates)

- A population parameter is *a measurable characteristic of the population.*
- Population parameters are *fixed* so long as the population itself does not change.
- Values based on a sample (sample statistics) are *estimates* of population parameters.
- Sample statistics vary from sample to sample even if samples are randomly selected.

4.3 Notation

- A *parameter* is a numerical descriptive measure of a *population.* It is calculated from the observations in the population
  - Population mean μ
  - Population standard deviation σ
- A *sample statistic* is a numerical descriptive measure of a *sample.* It is calculated from observations in the sample
  - Sample mean $\bar{x}$
  - Sample standard deviation s

4.4 Sampling distribution

In theory, we could select all possible random samples from a population and gain an *estimate* of the population *parameter* from each of the individual samples. A histogram of the sample estimates for each individual sample would form the sampling distribution of the population parameter (probability distribution). The sampling distribution of a sample statistic, calculated from a sample of n measurements, is the probability distribution of the statistic.

4.5 Sampling error and precision

Knowledge of the sampling distribution allows us to assess how close the estimate obtained from one individual sample is to the true population parameter. This is known as *precision*.

The *standard error* is defined as the standard deviation of the sampling distribution of the sample estimates. The standard error obtained from a sample is a measure of the *uncertainty* in the *sample statistic*. The smaller the standard error is, the greater the precision in the sample statistic.

4.6 Differences in the definitions of the standard deviation and standard error

The sample *standard deviation* is a measure of the variability of individuals in a sample.

The *standard error* obtained is a measure of the uncertainty in the sample statistic.

4.7 If we knew about the population mean and standard deviation

The mean of the sampling distribution = mean of the population distribution.

Standard deviation of the sampling distribution = $\sigma/\sqrt{n}$ = Standard error of the mean.

The sampling distribution of the mean is approximately Normal provided that there are at least 30 in the sample. This is true even if the original data are not Normally distributed.

4.8 Standard error of the mean

The standard error of the mean denoted SE(mean) or SEM is estimated from a single sample as:

$\sigma/\sqrt{n}$ = (standard deviation)/(square root of number in sample)

The standard error provides a measure of how far from the true population value the estimate is likely to be (the precision).

4.9 Practical definition of confidence intervals

Sample estimate and standard error are used together to produce a *confidence interval*. For practical purposes, a confidence interval is an interval around a sample estimate which is likely to contain population parameter. This interval gives suggestion of lower bound and upper bound for what might be value of population parameter

4.10 Textbook definition of 95% confidence interval

Consider an infinite number of random samples of size n from a population. For 95% of the samples from the population, the 95% CI around the sample mean will include the population mean.

4.11 Calculating the confidence interval

Confidence intervals can be calculated for most sample estimates using the following notation

Sample estimate $\pm$ critical value $\times$ standard error of sample estimate

Usually set critical value for 95% confidence interval. For large samples, >30, the critical value for a 95% confidence interval is 1.96.

A higher level of confidence (e.g. 99% rather than 95%) requires a wider interval and hence a higher critical value.

So the width of the confidence interval is controlled by: the size of the standard error which comes from the size of the sample and the level of confidence chosen (higher is wider).

4.12 Example: Confidence interval for a mean

If the relevant sample estimate is the sample mean then the 95% confidence interval is:

sample mean $\pm$ 1.96 $\times$ standard error of sample mean

The standard error of the mean is: $standard\ deviation/\sqrt{n}$.

4.13 Example: Confidence interval for a mean height

Sample estimate = 172.03 cm

Sample standard deviation = 6.03 cm

Sample size $n = 100$

Standard error = $6.03/\sqrt{n} = 0.603\ cm$

Lower limit of confidence interval = 172.03–$1.96 \times 0.0603 = 172.03 - 1.18188 = 170.85\ cm$

Upper limit of confidence interval $= 172.03 + 1.96 \times 0.0603 = 172.03 + 1.18188 = 173.21$ *cm*

The 95% confidence interval for mean height is 170.85 *to* 173.21 *cm*.

Note: Some books recommend rounding the lower limit down rather than up to ensure that the interval is definitely wide enough.

4.14 Using R

You can use the simple arithmetic calculator that we discussed in the Part 1 of the manual to do the calculation.

```
# SE & CI

xbar <- 172.03

xsd <- 6.03

n <- 100

se <- xsd / sqrt(n)

lCI <-  xbar - 1.96 * se

uCI <-  xbar + 1.96 * se

rst <-  c(xbar = xbar, xsd = xsd, n = n, se = se, lCI = lCI, uCI = uCI)

rst
```

```
    xbar      xsd        n       se      lCI      uCI
172.0300   6.0300 100.0000   0.6030 170.8481 173.2119
```

Chapter 5

Introduction to Hypothesis Testing

5.1 Overview of hypothesis testing

- Define the null and alternative hypotheses
- Collect data
- Calculate the probability (the p-value) of obtaining these results (or more extreme results) given the null hypothesis
- Use the p-value to decide whether to reject the null hypothesis

5.2 Null and alternative hypotheses

- The *null* hypothesis (H_0) is that there is *no difference* in the *population*
- The *alternative* hypothesis (H_1) is that there is a *difference* in the *population*

NB: the null and alternative hypotheses concern the whole population of interest, not the sample

5.3 One and two tailed tests

Usually a two-tailed (or two-sided) test is assumed. This means that the alternative hypothesis reflects that the results could go in either direction.

5.4 Example

- H_0: In the population mean blood pressure is the same before and after taking drug A
- Two-tailed H_1: In the population mean blood pressure is not the same before and after taking drug A
- One-tailed H_1: In the population mean blood pressure is lower after taking drug A

5.5 The p-value

The p-value is the probability of obtaining our results, or more extreme results, if the null hypothesis were true.

- In general, if the p-value is less than 5% ($p < 0.05$) we reject the null hypothesis in favour of the alternative hypothesis. The result is referred to as being statistically significant.
- If the p-value is greater than or equal to 5% ($p \geq 0.05$) we do not have enough evidence to reject the null hypothesis. The result is not statistically significant and we don't reject the null hypothesis.

5.6 Parametric versus non-parametric hypothesis tests

Parametric tests are based on an assumption that the data follow a known probability distribution, often the Normal distribution. Non-parametric tests do not assume any specific distribution for the data.

5.7 Hypothesis testing versus confidence intervals

Confidence intervals and hypothesis tests are closely related. If the hypothesised value (e.g., zero for a difference in means) for an effect lies outside the 95% confidence interval then the hypothesised value is implausible and the p-value will be less than 0.05.

A confidence interval gives more information than a p-value – as well as indicating statistical significance it also gives the possible range of plausible values for an effect.

5.8 Errors in hypothesis testing

- Type I error: rejecting the null hypothesis when it is true
- Type II error: not rejecting the null hypothesis when it is false

5.9 Multiple testing

When the null hypothesis is true and several hypothesis tests are conducted, the chance of making at least one Type I error is increased. Strategies to combat this include:

- Limiting the number of hypothesis tests conducted
- Using a reduce level of significance (e.g., $p < 0.001$ instead of $p < 0.05$)
- Bonferroni approach: divide the significance level by the number of tests conducted (e.g., if there are ten tests use $p < 0.005$ instead of $p < 0.05$)

5.10 Practical (or clinical) importance versus statistical significance

A test can be statistically significant ($p < 0.05$) but the observed effect may not be practically or clinically important. More commonly, especially for small samples, a study may show an effect of a clinically important magnitude but which is not statistically significant ($p > 0.05$). Interpretation of a hypothesis test should consider both practical and statistical significance.