Introduction to Statistics A

Maria Anastasiadi

2025-04-25

Statistical Foundations Part1

ILOs

- Critically assess the basic principles of different statistical techniques.

- Be able to select the correct statistical test depending on the experimental design and data type.

- Use R syntax and ecosystem to perform data analysis tasks.

Contents

1. Sample Statistics

▪ A **Sample or Descriptive Statistic** is a number that summarises data.

▪ Some of the most common sample statistics are the **mean**, the **standard deviation**, the **median**, the **maximum**, and the **minimum**.

1.1 Measures of Central Tendency

A) Mean

The most popular measure of central tendency is the **mean** also known as *the simple average*.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_1}{n}$$

Downside of using the mean as measure of central tendency?

B) Median

A better measure of Central Tendency is the **Median** which represents the *middle number in an ordered dataset* and is NOT affected by outliers.

1.2 Asymmetry

A measure of Asymmetry in a dataset is **Skewness**.

Skewness indicates if the observations in a dataset are concentrated (skewed) on one side.

Example:

The file `event_times.txt` contains the time (s) when consecutive cell divisions occur in a cell line culture.
We want to examine the distribution of waiting times between successive cell divisions.

1. Calculate waiting times

We are interested in calculating the waiting times between cell divisions

```
#1. Load the data in a vector:
div.time <- scan("Data/event_times.txt")

#2. Calculate waiting times
diff.time <- diff(div.time)
```

2. Get descriptive statistics

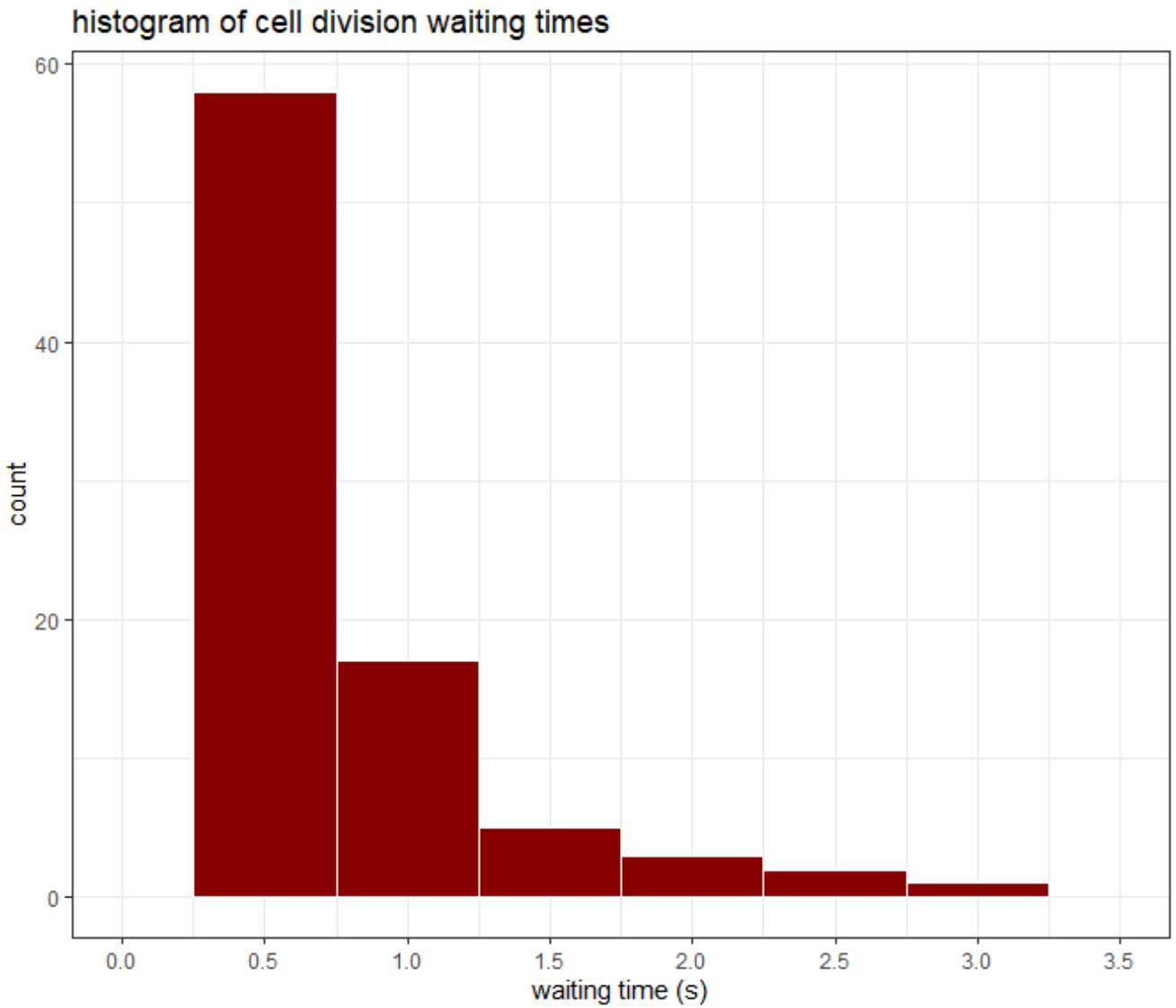
We can use the `summary()` function to get the main descriptive statistics for this dataset:

```
summary(diff.time)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00634 0.19084 0.33132 0.54319 0.69994 3.10818
```

Take a note of the Mean and Median values

3. Create a histogram of waiting times

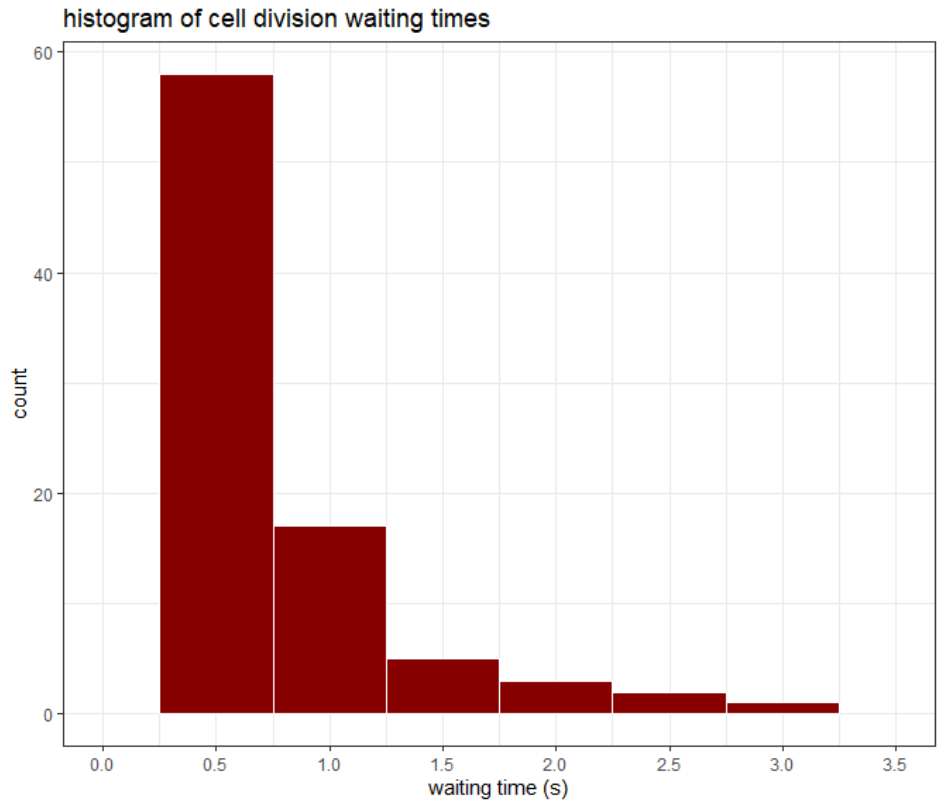


4. Summary

- If **Mean > Median** the data have a *positive or right skew*.
- If **Mean < Median** the data have a *negative or left skew*.
- If **Mean = Median** the data are completely *symmetrical*.

Quiz

What type of distribution/skew do we have in this case?



1.3 Measures of variability

Univariate Measures of Dispersion:

- **Variance**
- **Standard Deviation**
- **Coefficient of Variation**

A) Variance

	Population	Sample
Variance	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$

ADJUSTS FOR LARGER VARIABILITY

Sample Variance

- Sample variance s^2 measures the *dispersion* of a set of data points around their mean value.

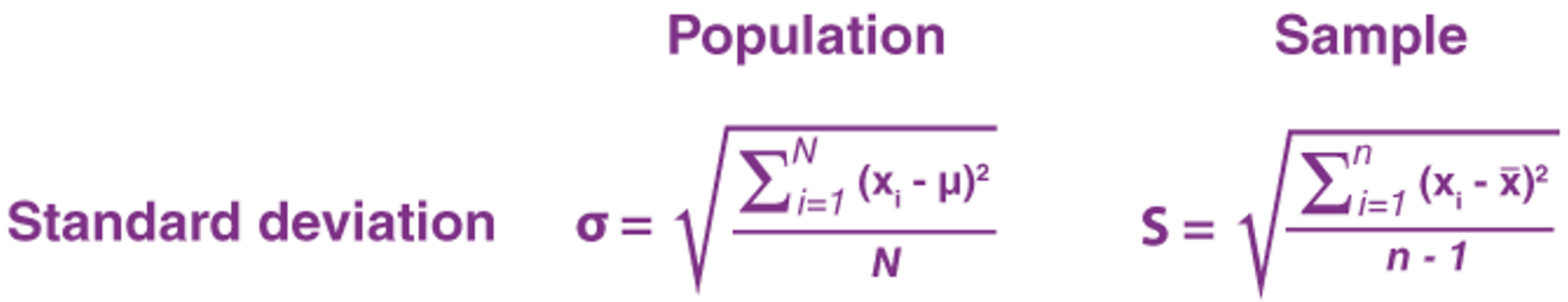
- The variance formula for the sample is more **conservative**.

- The **(n-1)** term in the formula accounts for the possibility that the variance captured by the sample is more than the variance of the population.

B) Standard Deviation

- Variance calculations can often result in large values as the term $(x_i - \bar{x})^2$ is squared.

- Solution: use the square root instead.



C) Coefficient of Variation

▪ The **Coefficient of Variation (CV)** or **Relative Standard Deviation (RSD)** is used to compare the standard deviations of data recorded in different units, e.g. Kg vs g.

$$\boldsymbol{RSD = \frac{s}{\bar{x}} \quad (for\ sample\ standard\ deviation)}$$

$$\boldsymbol{RSD = \frac{\sigma}{\mu} \quad (for\ population\ standard\ deviation)}$$

▪ The RSD or CV can also be expressed as a percentage.

2. Inferencial Statistics

Inferential Statistics refer to the branch of Statistics that rely on *Probability Theory* and *Distributions* to predict population values based on sample data.

2.1 Sampling Error

- Sample statistics can be used for making inferences for the whole population.
- But how can we be sure that these statistics are reliable and close to the true population parameters?

Let’s go back to the example of trying to calculate the mean diastolic blood pressure (MDBP) for the adult population of Massachusetts. In an effort to standardise our experiment, we have collected three samples of 20 volunteers each and the mean values are: **75.2, 79.5, 80.1 mm Hg**.

This difference between sample means is called **Sampling Error** or **Sampling Variability**

Manage Sampling Error

The best way to reduce the sampling error is by increasing the sample size.

2.2 Sampling distribution

We can use a simple simulation example to see what happens when we draw repeated samples of equal size from the same population.

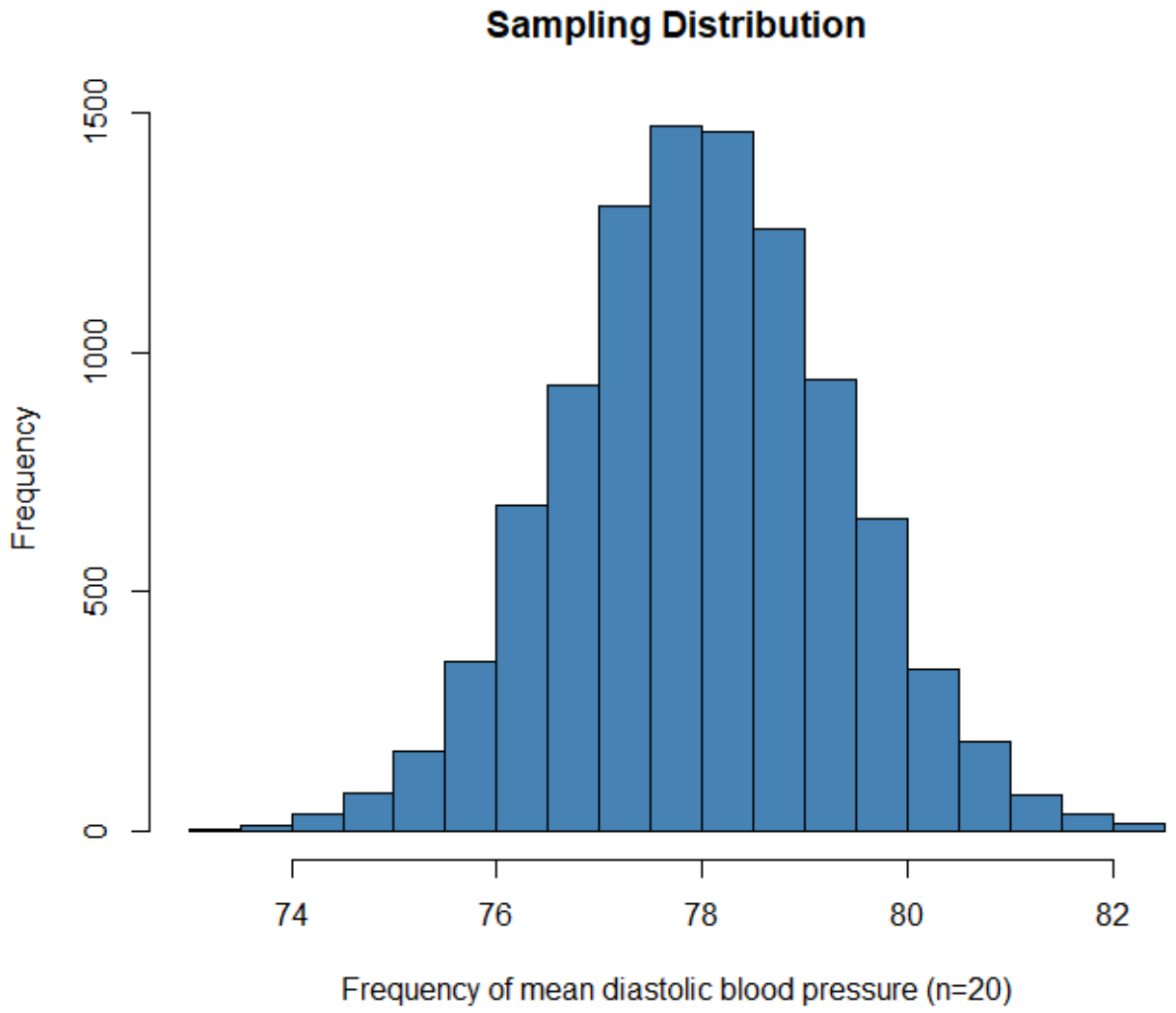
This means any variation observed will be due to sampling error.

1. Simulate sampling distribution

- Define a normally disitributed polulation with mean value MDBP = 78 mm Hg and sd = 6.
- Draw 10,000 samples of size 20 from the above population.
- Examples of sample means

```
## [1] 77.98933 78.66990 77.30639 77.07685 79.63759 77.48844
```

2. Sampling Distribution Histogram

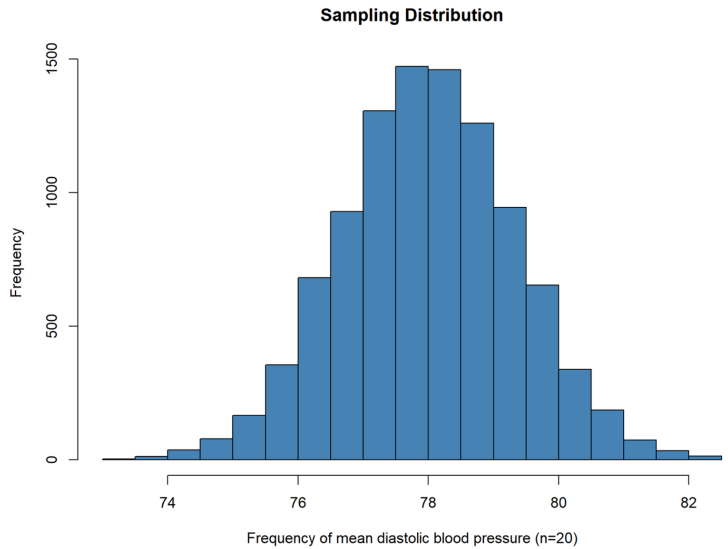


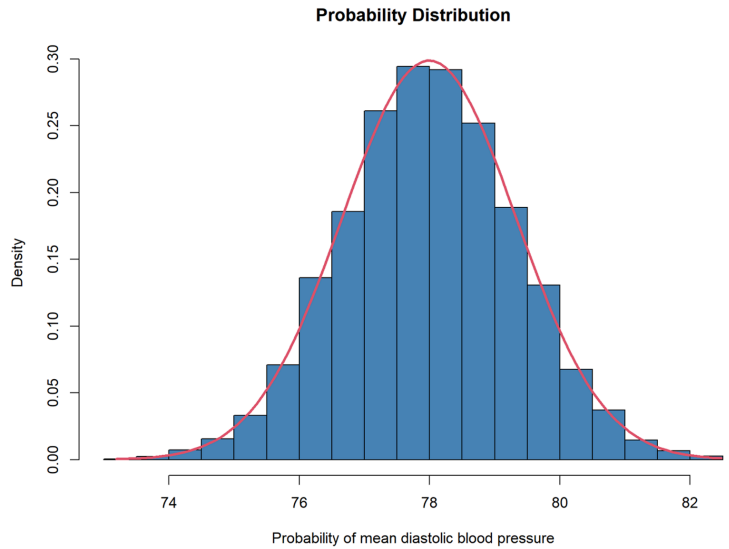
2.2.1 Probability Distributions

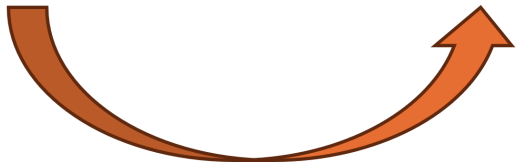
- A sampling distribution shows the expected range and frequency of outcomes when we repeat the same sampling process.
- An alternative way of thinking of distributions is in terms of how **likely** it is for an outcome to occur instead of how often it occurs.

Probability Distributions

They help convert the frequency of an outcome into a **probability of observing this outcome** by consulting the probability distribution.

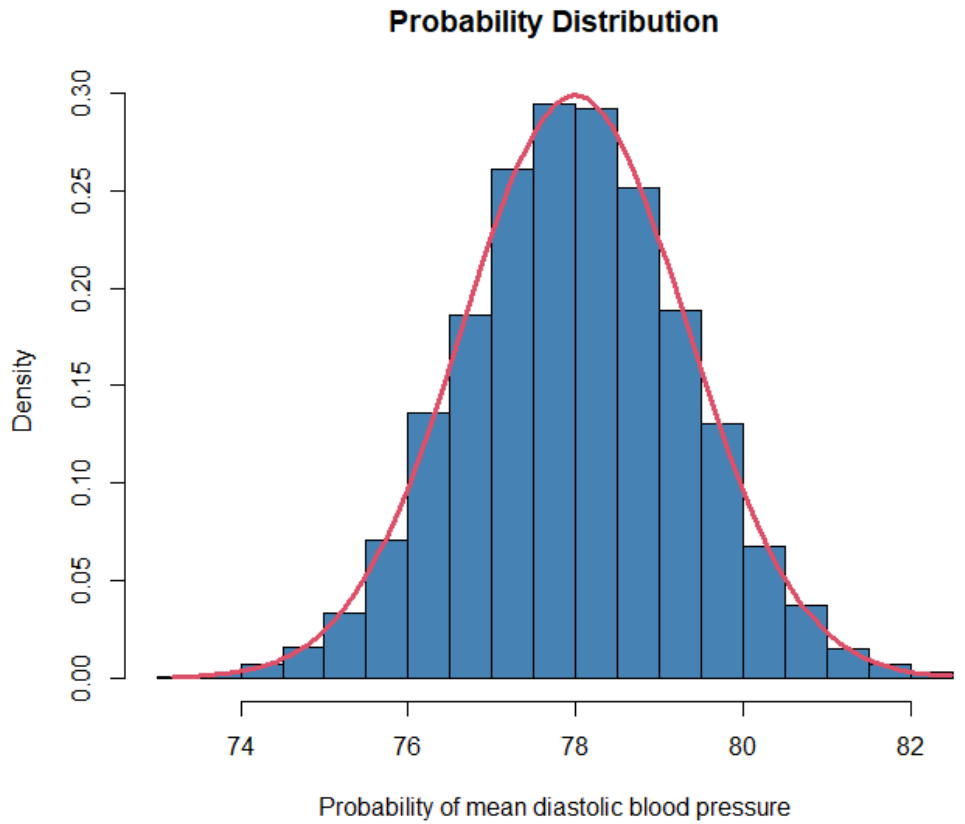




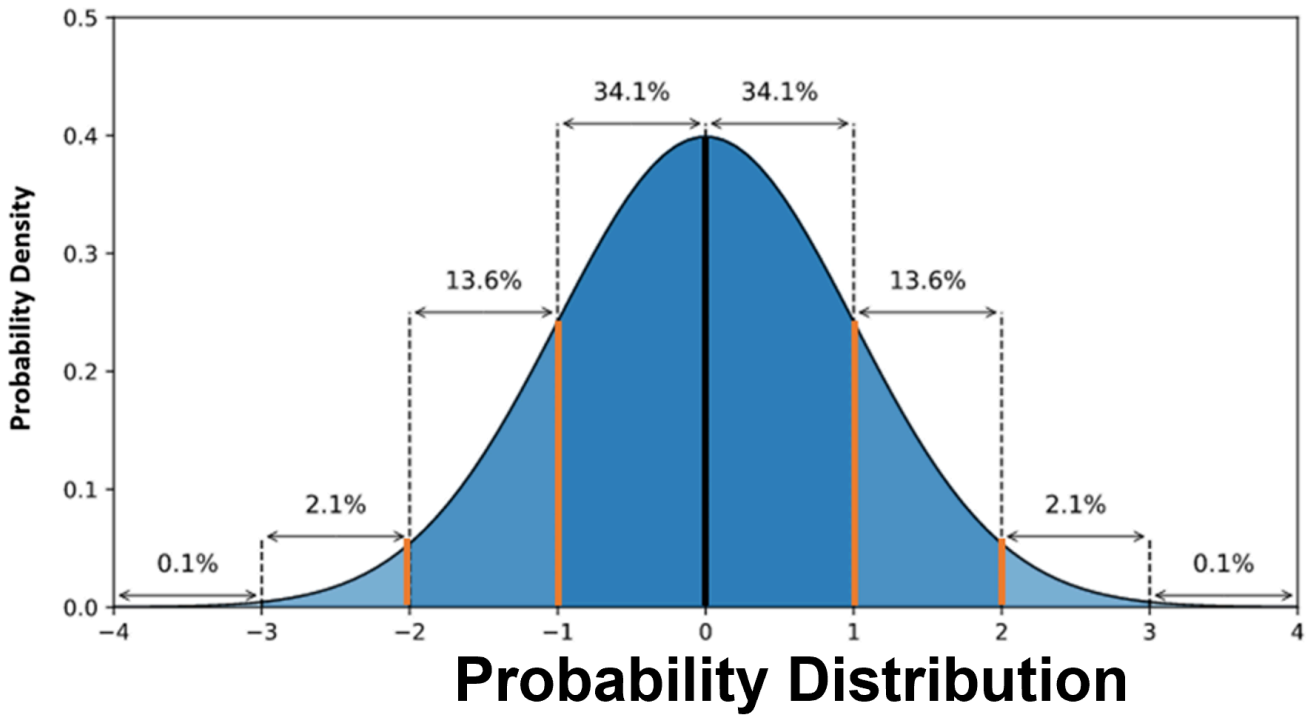


Normal Probability Distributions

Completely symmetrical with the most probable values centred around the mean.



Standard Normal Distribution



A special normal distibution with a mean = 0, and sd = 1 [$N(0,1)$].

z-score standardisation

If we have approximately normally distributed data, we can apply **z-score standardisation** to transform the dataset into one with a standard normal distribution.

$$\text{Z-scores} = \frac{\text{Original variable} - \text{Mean}}{\text{Standard deviation}}$$

z-scores

Once we have acquired the z-scores we can compare them against probability tables for the probability of getting this score.

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938

z-score = -2.53

Normal Distribution Empirical Rule

A distribution is normal if:

- Around 68% of scores fall within 1 standard deviation above and below the mean.
- Around 95% of scores fall within 2 standard deviations above and below the mean.
- Around 99.7% of scores fall within 3 standard deviations above and below the mean.

2.3 The Central Limit Theorem

The **Central Limit Theorem** shows the following:

When we increase the sample size (or the number of samples), then the sample mean will be closer to the population mean (**Law of Large Numbers**).

Central Limit Theorem

"If we have a sample with more than 30 observations, we can accept that it is coming from a sampling distribution with a mean equal to the population mean".

Central Limit Theorem Implications

▪ With multiple large samples, the **sampling distribution** of the mean is normally distributed, even if the original variable is not.

▪ We can use parametric tests for large samples from populations with any kind of distribution as long as other important assumptions are met.

▪ For small samples, the assumption of normality is important because the sampling distribution of the mean isn't known.

2.4 Standard Error

- The **standard error** statistic tells us how variable the *sampling distribution* is.

$$\sigma / \sqrt{n}$$

Standard Error Definition

"The standard error of an estimate is the standard deviation of the estimate's sampling distribution".

❗ The key point to remember is that the standard error (**SE**," or **se**) is a measure of the **spread**, or dispersion, of the sampling distribution.

Summary

Standard Deviation tells us how far each value lies from the mean within a single dataset (A **descriptive** statistic).

Standard Error tells us how accurately our sample data represents the whole population (An **inferential** statistic).

2.5 Confidence Intervals

- Another way of estimating how well the sample describes the population is by calculating **confidence intervals**.

- For given α the margin of error *m* for a CI is:

$m = mean \pm Z_{a/2} * SE.$

- Confidence Intervals are a range of values where the population mean is likely to fall.

Common confidence intervals and corresponding Z scores

Desired CI	Z Score
90%	1.645
95%	1.96
99%	2.576

NOTE 

- If se and CI are small, we can be fairly confident the sample mean is a good estimate of the population mean.

- If se and CI are large, this implies they are uninformative. The true population mean can fall anywhere in the range.

Confidence vs Significance

- The **se** and **CI** tells us how confident we are that we have captured the true population mean but does not tell us if the result is *statistically significant*.

- To be able to say this we need to look at the **probability statistics** calculated using hypothesis testing.

3. Hypothesis Testing

Steps in data driven decision making

3.1 Hypothesis Definition

So what is a hypothesis? 🤔

Intuitively **“A hypothesis is a statement that can be tested”**.

Example: The mean length of newborn babies in the UK is equal to 50cm.

A hypothesis can be **TRUE** or **FALSE**. The two scenarios are covered by the **Alternative** and **Null** Hypothesis respectively.

Null and Alternative hypothesis

- The Null hypothesis (H_0) says what our theory predicts will be FALSE.
- The Alternative hypothesis (H_1) says what our theory predicts will be TRUE.

When conducting hypothesis testing the alternative hypothesis can be two sided or one sided.

3.2 Two-sided Hypothesis Testing

Example of two-sided hypothesis test

Hypothesis	Notation	Outcome1	Outcome2
The mean length of newborn babies in the UK is 50 cm	$H_0: \mu_0 = 50$ cm	IF REJECTED ⌄	NOT REJECTED ⌄
The mean length of newborn babies in the UK is not 50 cm	$H_1: \mu_0 \neq 50$ cm	SUPPORTED	CAN NOT BE SUPPORTED

Important!

Remember that the alternative hypothesis H_1 cannot be proved.

What we are trying to do is reject the Null hypothesis H_0.

3.3 One-sided Hypothesis Testing

Example:

According to the National Institute of Health in the U.S. an estimated 31.9% of U.S. adolescents aged 13-18 had any anxiety disorder 😟 in the period 2001-2003.

Forming our Hypothesis

We hypothesise that in recent years the prevalence of anxiety disorders in adolescents in the U.S. has risen.

Example of one-sided hypothesis test

Hypothesis	Notation	Outcome1	Outcome2
The prevalence of anxiety disorders in US adolescents is 31.9%	H_0	IF REJECTED 	NOT REJECTED 
The prevalence of anxiety disorders in US adolescents > 31.9%	H_1	SUPPORTED	CAN NOT BE SUPPORTED

Hypothesis Testing Considerations

In hypothesis testing we have three things we need to define:

A) The Null Hypothesis H_0 we are trying to reject.

B) The rejection region.

C) The significance level.

3.4 Rejection Region

After defining the Null Hypothesis we need to define the Rejection Region.

How is the rejection region defined?

Example

Assume we are interested in testing the following statement: “The average mean birth weight of babies born 👶 in a large UK hospital🏥 is 3900 g”.

We don’t agree with this statement and we declare that: “The average birth weight of newborn babies in this hospital is different to 3900 g”.

H_0: birth weight = 3900 g

H_1: birth weight ≠ 3900 g

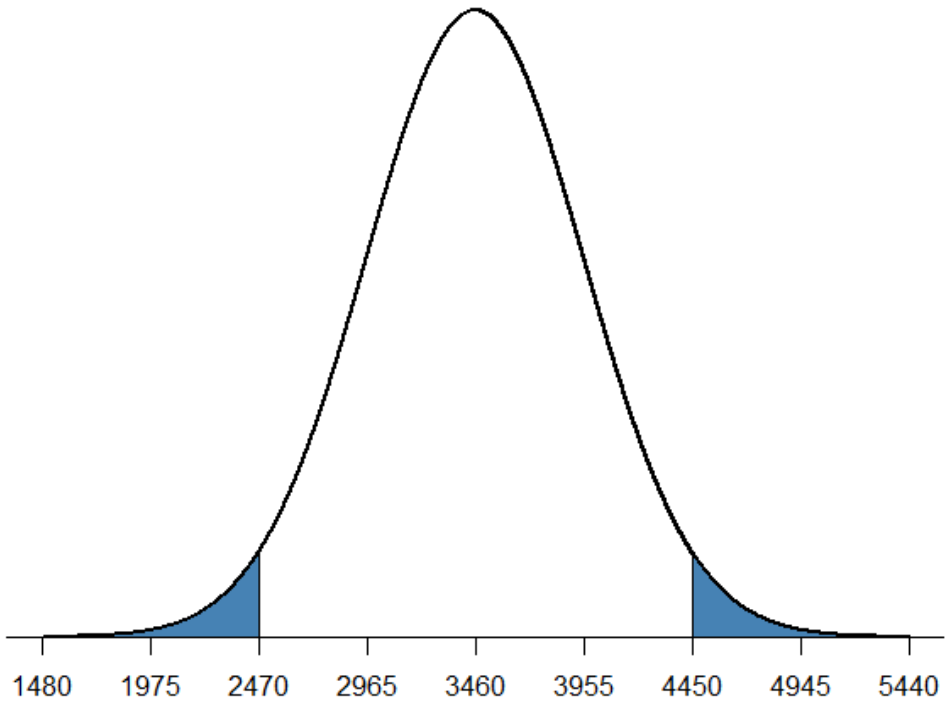
1. Define population

After obtaining the birth records for all babies in this hospital born in the last year the mean weight was **3460 g** with a sd of **495** and the data were normally distributed **N~(μ=3460, σ=495)**.

2. Draw the above distribution:

Two-sided rejection plot

Rejection region at *significance level* $\alpha = 0.05$.



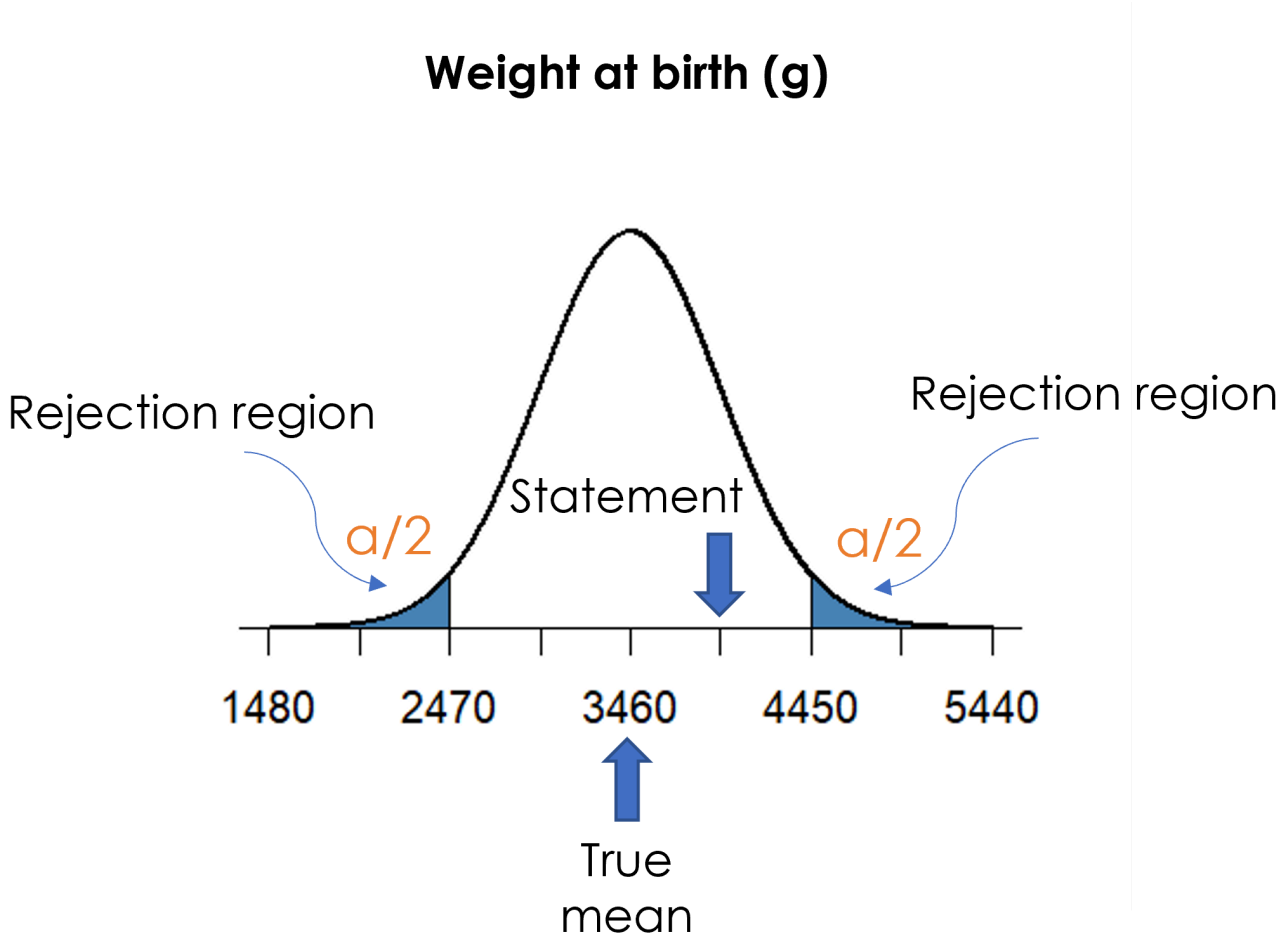
Conclusions

- The significance level **α** represents the probability of rejecting the null hypothesis **if it is true**.

- If the null hypothesis value falls inside the rejection region, then we can reject the Null hypothesis

Question:

From the Figure below, can we can reject the Null hypothesis?



3.5 Type I and Type II Errors

- **Type I** error is when we reject the Null hypothesis when it is in fact TRUE. **[FALSE POSITIVE]**.

- **Type II** error is when we fail to reject the Null hypothesis when it is in fact FALSE. **[FALSE NEGATIVE]**.

Type I Errors Facts

- The probability of making a Type I error is **α**.
- Type I errors are more serious and tests are usually designed to reduce the probability of type I errors (e.g. Post-Hoc tests).

Type II Error Facts

- The probability of a type II error is denoted as **β** and depends on the sample size and the population variance.
- **Power of a Test 1-β**: the probability of **TRUE POSITIVES**.
- **To increase the Power of a test (1-β) we can increase the sample size.**

Test Power Summary

Type I and Type II Error		
Null hypothesis is ...	**True**	**False**
Rejected	**Type I error** False positive Probability = α	**Correct decision** True positive Probability = 1 - β
Not rejected	**Correct decision** True negative Probability = 1 - α	**Type II error** False negative Probability = β

Power of the test

3.6 Test of Significance

A test of significance finds the probability of getting an outcome as extreme or more extreme than the actually observed outcome assuming the Null hypothesis is TRUE.

▪ we can use the **z scores** to assess how far away the estimate is from the population parameter.

▪ We call these scores a **test statistic** which has the purpose of measuring compatibility between the Null hypothesis and the data.

z-statistic

$$z = \frac{estimate - hypothesised\ value}{standard\ deviation\ of\ the\ estimate}$$

- estimate = the observed value for a statistic acquired from the sample.
- hypothesised value = the value we attribute to the parameter under the Null hypothesis.
- standard deviation of the estimate =the sd of the sampling distribution.

Example

Now assume we want to test whether there is a difference in birth weight between boys 👦 and girls 👧 in the country.

What does our hypothesis look like?

Sampling

To test the hypothesis we look at many different samples we find boys are on average 200 g heavier than girls with a sd=60 g.
Is this difference **statistically significant?**

1. Calculate the z-statistic

The z statistic in this case would be: $z = \frac{200-0}{60} = 3.33$

- This means that we have observed a sample estimate >3 SD away from the hypothesised value of the parameter (diff = 0).
- Since the sample sizes are sufficiently large the z statistic will have approximately the standard normal distribution ***N*(0,1)**.
- Based on the z statistic can we reject the Null hypothesis?

2. Conduct Significance Test

In our example this translates as:

$P(Z \leq -3.33 \text{ or } Z \geq 3.33) \rightarrow P(|Z| \geq 3.33) = 2P(Z \geq 3.33)$

From the table of z scores we find:

$2P(Z \geq 3.33) = 2(1-0.9996) = 0.0008.$

This is the **P-value** of the test.

P-value Definition

The **P-value** of a test is the probability that the test statistic would take a value as extreme or more extreme than that actually observed assuming that H_0 is true.

Statistical Significance

If the P-value is ≤ α, we say that the data are **statistically significant at level α**.

- Most commonly we choose **α=0.05** which means that if H_0 was indeed TRUE, we would not observe this test statistic value more than 5% of the time.

3.7 Estimating Population Mean

- When σ is **unknown**, we must first estimate σ before we can make any inference for μ!

- In this case, we use the sample standard deviation *s* to estimate the population standard deviation σ.

One-sample *t*-statistic

The new statistic is called: **one-sample *t*-statistic**:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The denominator is called the **standard error of the sample mean** and it is used to estimate the unknown standard deviation of the sample mean:

$$\sigma/\sqrt{n}$$

The *t*-statistic Distribution

- Unlike the **z** statistic the **t** statistic **does not** follow a normal distribution.

- It follows a new type of distribution called a ***t*-distribution** or **Student's *t*-distribution**.

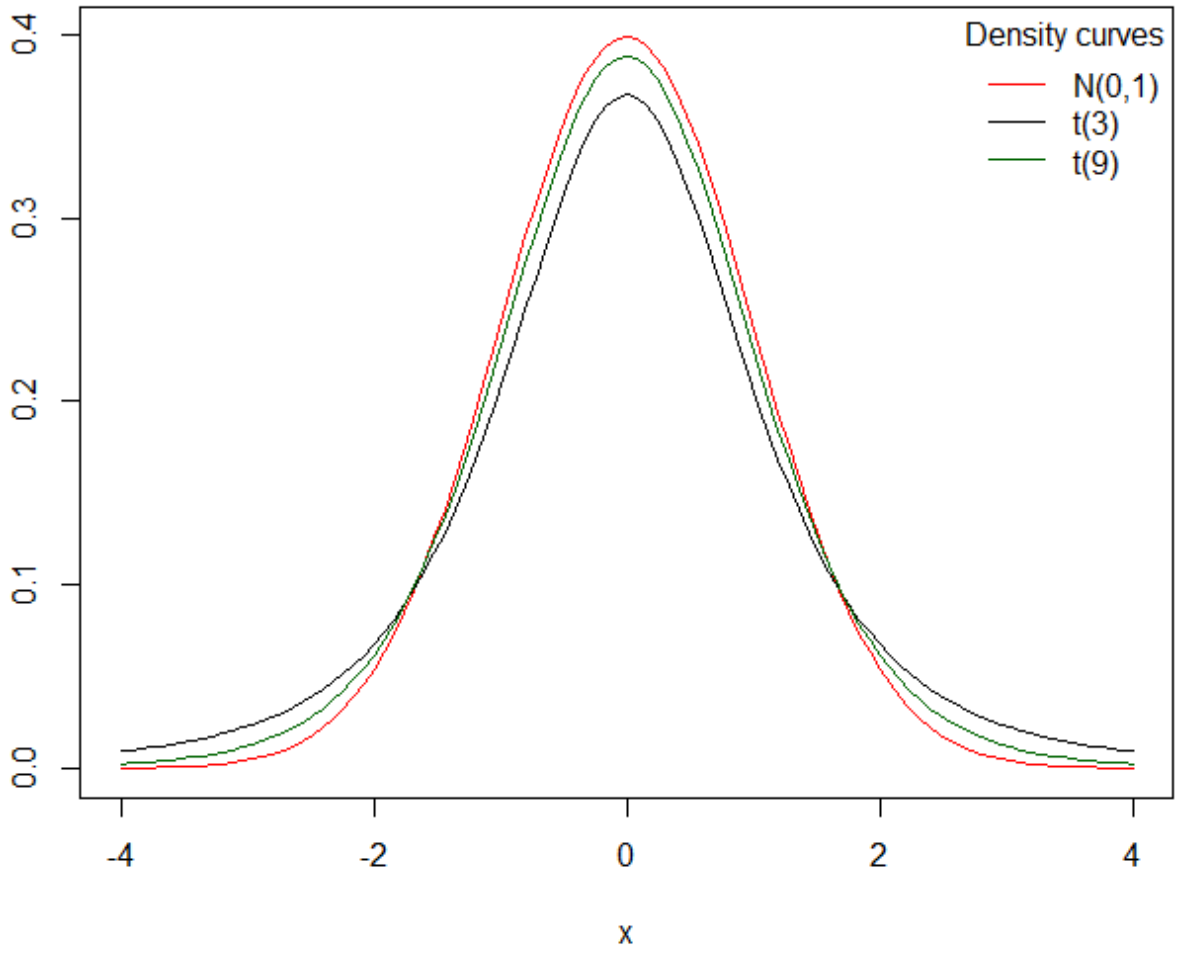
Important

- The type of *t*-distribution for a given sample is dependent on the sample size (n)!

- To know the type of *t*-distribution we need the **degrees of freedom k=n-1**.

- We use *t(k)* to define a *t* distribution with *k* degrees of freedom.

t-statistic Distribution Examples



t-Test P-Values

For a random standard variable T having the t**(n-1)** distribution, the **P-value** for a test of H_0 against all possible alternatives is calculated as:

A) For $H_1 : \mu > \mu_0$ the P-value is: P(T ≥ t)

B) For $H_1 : \mu < \mu_0$ the P-value is: P(T ≤ t)

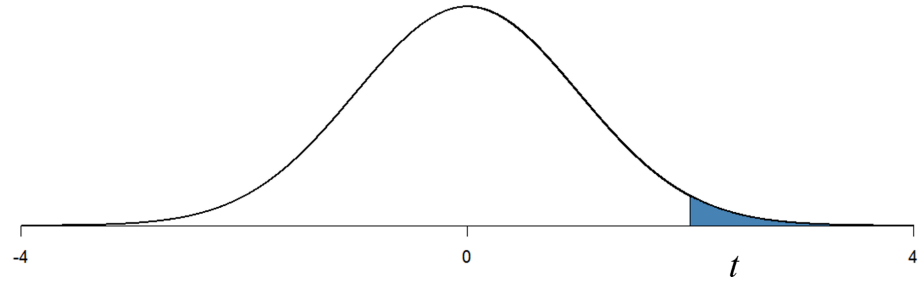
C) For $H_1 : \mu \neq \mu_0$ the P-value is: 2P(T≥ |t|)

t-Test for a Population Mean

t Test for Population Mean

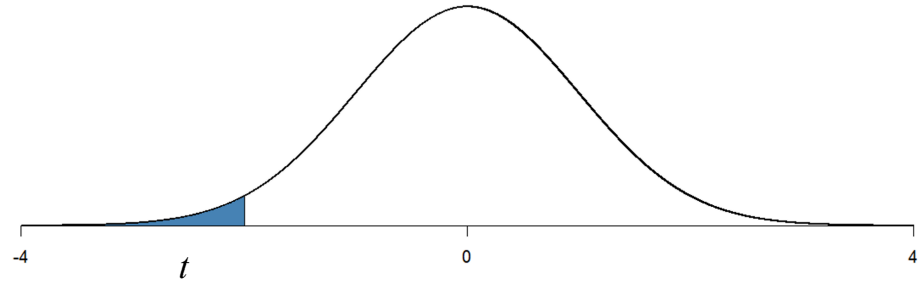
$H_1: \mu > \mu_0$

$P(T \geq t)$



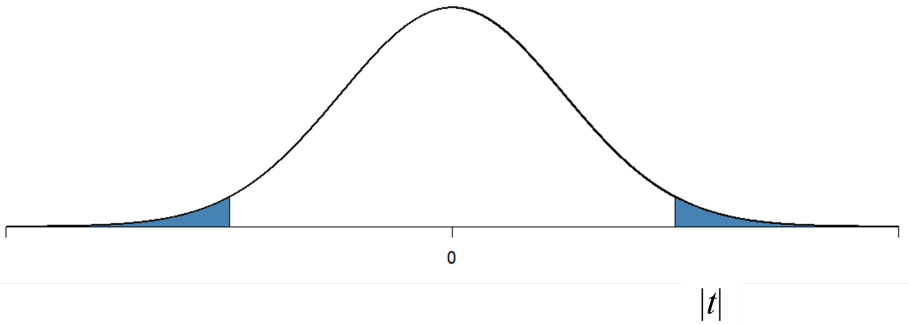
$H_1: \mu < \mu_0$

$P(T \leq t)$



$H_1: \mu \neq \mu_0$

$2P(T \geq |t|)$



3.8 Comparing Two Means

- In many research studies our purpose is to see if a treatment has an effect on a population.

- For the study results to be valid we need to include a **control** group as well as the **treatment group**.

- This is often called the **Two-sample problem**.

Two-Sample Problem Summary

- The goal of inference is to compare the response in two groups.
- Each group is considered to be a sample from a distinct population with means μ_1 and μ_2, and sd σ_1 and σ_2 respectively.
- The responses of one group are independent of those of the other.

 In addition, there is no need the two groups to have the same size, as would be the case in matched-pair designs.

Example

We have a clinical trial where volunteers are randomly assigned to a group receiving a treatment and a control group receiving a placebo.

▪ The same variable is measured in both groups but we call the variable x_1 in the treatment group and x_2 in the placebo group as their distribution may be different.

Comparing Populations

- Our main aim is to compare the two population means by testing the hypothesis $H_0: \mu_1 = \mu_2$.

- Inference is based on the two samples comprised of the two groups of volunteers.

Population	Sample Size	Sample Mean	Sample standard deviation
1	n_1	$\bar{x_1}$	s_1
2	n_2	$\bar{x_2}$	s_2

3.9.1 The Two-Sample *t*-statistic

- When σ_1 and σ_2 is **unknown** we compute the **two-sample *t*-statistic**.

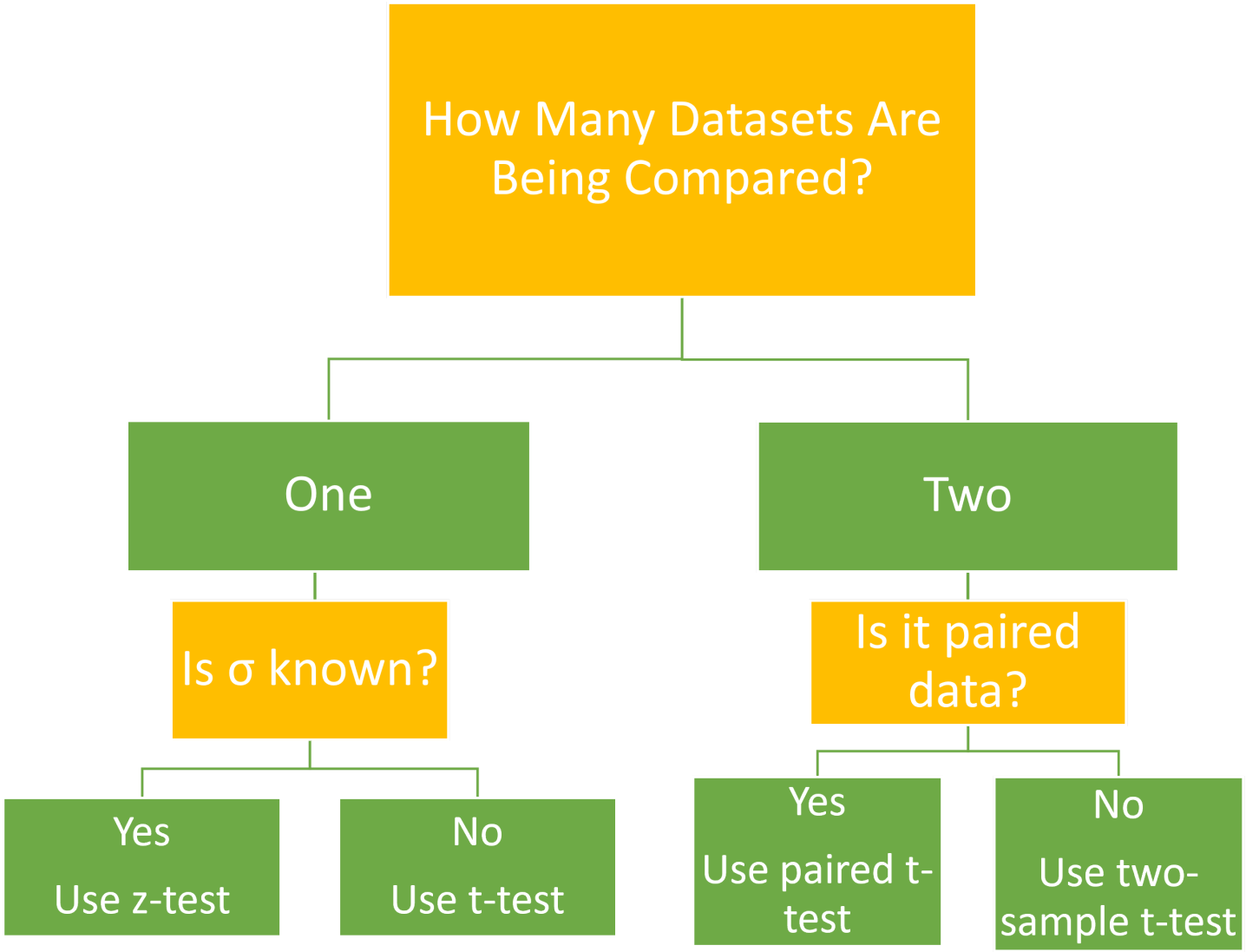
$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{(\frac{s_1^2}{n_1}) - (\frac{s_2^2}{n_2})}}$$

When to reject the Null Hypothesis?

 To decide whether we can reject the Null Hypothesis in favour of the Alternative $H_1: \mu_1 \neq \mu_2$, we look at the p-values for the *t(k)* distribution which is an approximation for the two-sample *t*-statistic distribution.

The degrees of freedom *k* are either approximated by software or are the smaller of $n_1 - 1$ vs $n_2 - 1$.

Summary of z & t-tests



Considerations on Statistical Significance tests

1. Exact some caution in putting too much weight on statistical significance.

2. Small effects can be highly significant (very small P-values) but the practical importance of this effect can be questionable.

3. On the other hand, if we fail to reject the Null hypothesis this doesn't necessarily mean H_0 is TRUE especially when the test has low power.

Effect size

- To know if an observed difference is not only statistically significant but also important or meaningful, we can calculate its **effect size.**

$$effect - size = \frac{mean_{treatm} - mean_{control}}{sd_{control}}$$

- Effect size is a standardized measure of the difference between groups.
- All effect sizes are calculated on a common scale.

Effect size interpretation

😕 < 0.1 = trivial effect
😐 0.1 - 0.3 = small effect
🙂 0.3 - 0.5 = moderate effect
🎉 > 0.5 = large difference effect