R：Statistical Programming Methods
R：程式、機率與統計

Significance Testing (1)

A/B Testing

- An A/B test is an experiment with two groups to establish which of two treatments, products, procedures is better or superior
- Example:
  - Testing two therapies to determine which suppresses cancer more effectively
  - Testing two prices to determine which yields more net profit
- Treatment group v.s. control group

Hypothesis Tests/ Significance Tests

- To learn whether random chance might be responsible for an observed effect.
- 區分觀測到的差異是來自於抽樣誤差還是母體本身的差異 (Misinterpreting Randomness)

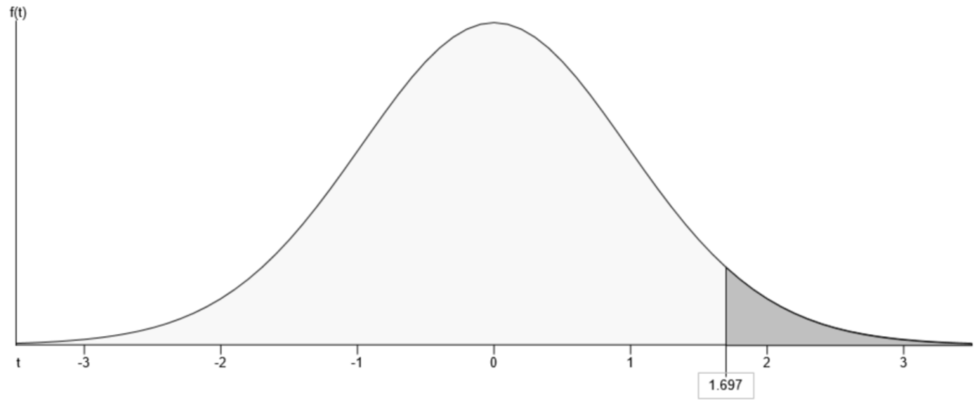
- The Null Hypothesis (虛無假說)
  - Baseline assumption that the treatments are equivalent, and any difference between the groups is due to chance
- Alternative Hypothesis (對立假設)
  - Counterpoint to the null (what you hope to prove)

Two-tailed and One-tailed Tests -1

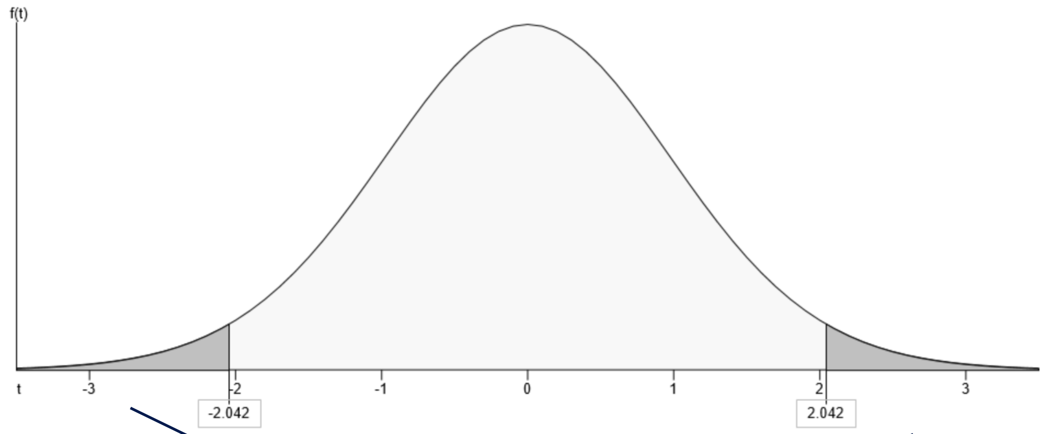
- When there is a direction, the alternative will use '>' or '<'
- When there is no direction, use ≠ (no equal)

- For example, if unsure as to whether Cambridge or Oxford graduates get paid more, following hypothesis will be used:
- H0: $\mu 1 = \mu 2$
- H1: $\mu 1 \neq \mu 2$

Two-tailed and One-tailed Tests -2



5% of the data



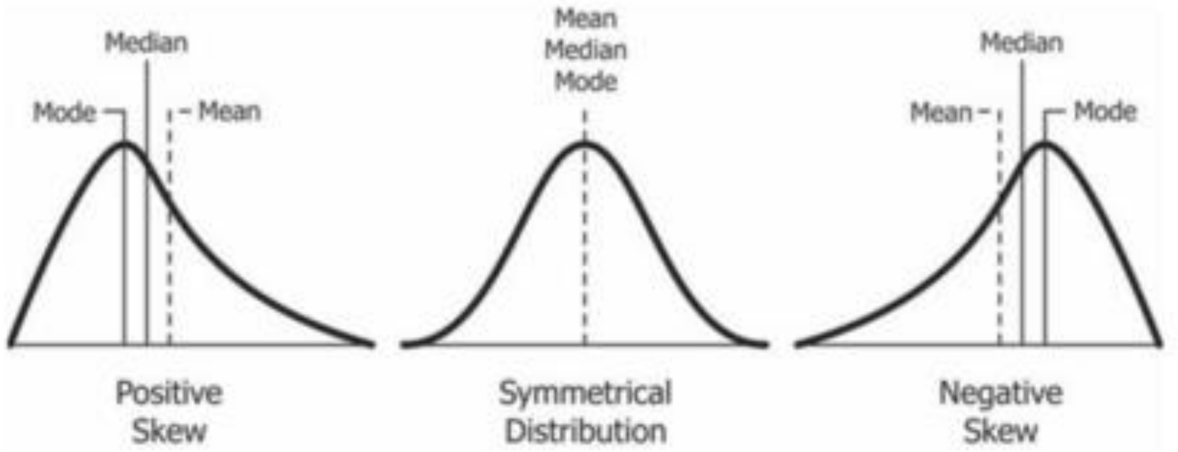
2.5% of the data

Shaded areas represents significance (reject null hypothesis).

Parametric Assumptions - Normality

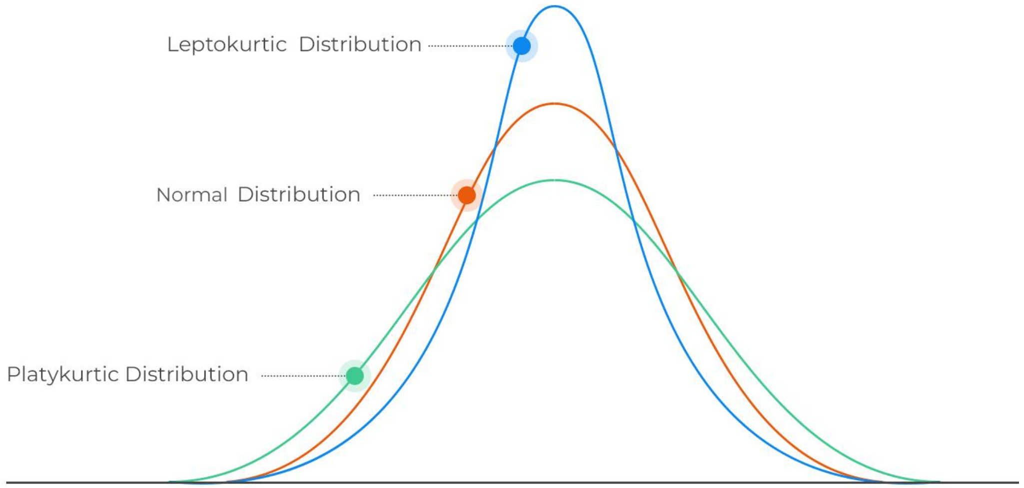
Skewness 偏度

- whether the data is skewed to larger or smaller values, and kurtosis indicates the propensity of the data to have extreme values



Kurtosis峰度

- the propensity of the data to have extreme values



Parametric Assumptions - Homogeneity of variance

- the variance of the two groups is equal
- Levene's test
  - H_0: $\sigma_i^2 = \sigma_j^2$
  - H_1: $\sigma_i^2 \neq \sigma_j^2$

T-test

One Sample T-test

- A manager claims that attendance is higher than 40 hours a week. He gathers data from 25 employees and finds that the average hours spent are 42 with a sample standard deviation of 5. Using a 95% confidence level, did the analyst find enough evidence to support her claim?

Hypothesis Testing

- Step 1 – Write down the hypotheses
  - H0: μ = 40
  - H1: μ > 40
- Step 2 – Calculate the standard error
  - $\mathrm{SE} = \frac{S}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$
- Step 3 – Find the t-value using the standard error as a replacement for σ in the formula:
  - t = $\frac{42 - 40}{1}$ = 2
- Step 4 – Find t_a for 95% confidence: t_a = 1.71 [note that the hypotheses have a direction. H1 focuses on if μ is **greater** than 40, not **different** to 40]
- If t > t_a, then we found enough evidence to support the analyst's claim (H1) and therefore reject the null hypothesis (H0). Otherwise we fail to reject the null hypothesis.

One-sample T-test

```
#one-sample t test
set.seed(1234)
df <- c(rnorm(500, mean=140, sd=0.01)) #randon draw from normal distribution
t.test(df, mu=140)

##

##  One Sample t-test

##

## data:  df

## t = 0.039734, df = 499, p-value = 0.9683
## alternative hypothesis: true mean is not equal to 140

## 95 percent confidence interval:

##  139.9991 140.0009

## sample estimates:

## mean of x

##       140
```

Can NOT reject null Hypothesis: H0=140

One-sample T-test

```
df.w <- women #data inside R
t.test(df.w$height, mu=63.7) #test if the average height is equal to 63.7 inches
##
##  One Sample t-test
##
## data:  df.w$height
## t = 1.1258, df = 14, p-value = 0.2792
## alternative hypothesis: true mean is not equal to 63.7
## 95 percent confidence interval:
##  62.52341 67.47659
## sample estimates:
## mean of x
##        65
```

What will be the result?

Two sample T-test -1

- Which group sleep longer?

Independent Samples

- Advantages:
  - Easier to obtain data with independent samples

- Disadvantages
  - Less control of random variability from each group.
  - Less statistical power (chance to find an effect when one exists). Therefore, larger sample sizes are needed in order to find an effect

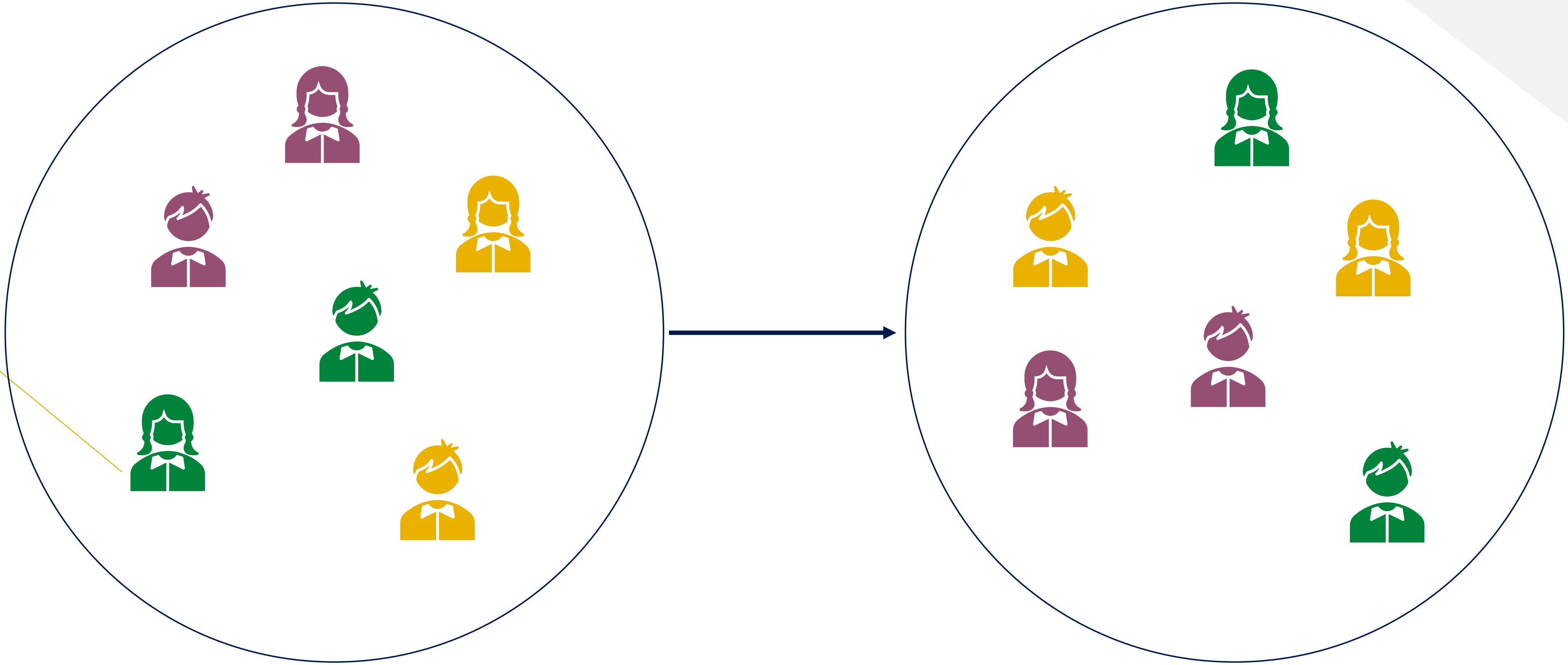
Two-sample T-test (Independent)

```
#two-sample t test (independent)
sample1 <- c(rnorm(100, mean=140, sd=6))
sample2 <- c(rnorm(100, mean=150, sd=6))
t.test(sample1, sample2, var.equal = TRUE)

##

##  Two Sample t-test

##

## data:  sample1 and sample2

## t = -13.129, df = 198, p-value < 2.2e-16

## alternative hypothesis: true difference in means is not equal to 0

## 95 percent confidence interval:

##  -11.840276  -8.747907

## sample estimates:

## mean of x mean of y

##  139.1787  149.4728
```

The means of two samples are statistically different.

Two sample T-test -2

- Do they sleep longer taking medicine?



Paired Samples

- Advantages:
  - More statistical power (increases chance of finding an effect when one exists)
  - Less sample size requirements
- Disadvantages
  - When working with participants it may mean less people, but also means twice the effort.
  - Parametric assumptions are harder to understand

Two-sample T-test (Paired)

```
#two-sample t test (paired)
sample3 <- c(rnorm(100, mean=50, sd=8))
sample4 <- c(rnorm(100, mean=40, sd=8))
t.test(sample3, sample4, paired=TRUE)

##

##  Paired t-test

##

## data:  sample3 and sample4

## t = 8.4359, df = 99, p-value = 2.77e-13

## alternative hypothesis: true mean difference is not equal to 0

## 95 percent confidence interval:

##   7.531866 12.164748

## sample estimates:

## mean difference

##        9.848307
```