NUS National University of Singapore | School of Computing

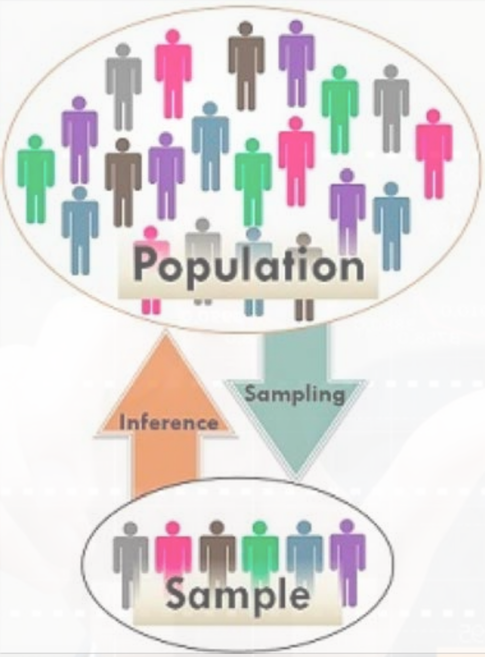
Leading The World With Asia's Best

# Hypothesis Testing

# Learning Objectives

- Be able to explain the purpose of hypothesis testing
- Understand what is the null and alternative hypotheses, and be able to formulate hypotheses correctly
- Understand the possible outcome that results from hypotheses test and be able to explain what is Type 1 and Type 2 errors
- Know how to choose the right test statistic for hypotheses tests involving means and proportions
- Be able to draw conclusions using p-values for one- and two-tailed hypothesis tests
- Be able to explain the purpose of and to conduct the analysis of variance (ANOVA) test

# Statistical Inference




- focuses on drawing conclusions about populations from samples.
- includes estimation of population parameters and hypothesis testing, which involves drawing conclusions about the value of the parameters of one or more populations.

# Hypothesis Testing

- Involves drawing inferences about <u>two</u> contrasting propositions (each called a <u>hypothesis</u>) relating to the value of one or more population parameters.
- $H_0$ Null hypothesis: describes an existing theory
- $H_1$ Alternative hypothesis: the complement of $H_0$
- Using sample data, we either:
  - *reject* $H_0$ and conclude the sample data provides sufficient evidence to support $H_1$, or
  - *fail to reject* $H_0$ and conclude the sample data does not support $H_1$.

***No proof that H1 is true or false***

# Hypothesis Testing

### Analogy for Hypothesis Testing

- In the US legal system, a defendant is innocent until proven guilty.
  - $H_0$: Innocent
  - $H_1$: Guilty
- If evidence (sample data) strongly indicates the defendant is guilty, then we reject $H_0$.
- If the evidence is not sufficient to indicate guilt, then we cannot reject the "innocent" hypothesis ($H_0$).
  - However note that we have not *proven* guilt or innocence!

# Hypothesis Testing Procedure

Steps in conducting a hypothesis test:

1. Identify the population parameter and formulate the hypotheses to test.
2. Select a level of significance (the risk of drawing an incorrect conclusion).
3. Determine the decision rule on which to base a conclusion.
4. Collect data and calculate a test statistic.
5. Apply the decision rule and draw a conclusion.

# One-Sample Hypothesis Tests

- Three forms:
  1. $H_0$: parameter $\leq$ constant
     $H_1$: parameter $>$ constant
  2. $H_0$: parameter $\geq$ constant
     $H_1$: parameter $<$ constant
  3. $H_0$: parameter $=$ constant
     $H_1$: parameter $\neq$ constant

- The equality part of the hypotheses is always in the null hypothesis.

# One-Sample Hypothesis Tests

## Example of Formulating a One-Sample Hypothesis Test

- CadSoft: producer of computer-aided design software for aerospace industry, receives numerous calls for technical support
- Average response time has been at least 25 mins
- So the company decide to upgrade its systems to help reduce response time.
- Sample data for 44 customers was collected to see if the new system is effective

How can you set up the hypothesis testing to see if the new system is effective?

- H0: population mean response time ≥ 25
- H1: population mean response time is < 25

- Using the proper symbol for population parameter,

  H0: $\mu \geq 25$

  H1: $\mu < 25$

# One-Sample Hypothesis Tests

## Potential Errors in Hypothesis Testing

- Sampling error - Risk in drawing an incorrect conclusion
- Four outcomes are possible:

$H_0$ is actually

| | FALSE | TRUE |
|---|---|---|
| Reject $H_0$ | Correct | TYPE I error ($p=\alpha$) |
| Accept $H_0$ | TYPE II error ($p=\beta$) | Correct |

charge a person guilty by mistake

person is guilty but lack evidence to proof

$H_0$: Innocent
$H_1$: Guilty

# One-Sample Hypothesis Tests

<u>Understanding Potential Errors in Hypothesis Testing</u>

- $\alpha = P($ rejecting $H_0 \mid H_0$ is true)

  $\beta = P$(not rejecting $H_0 \mid H_0$ is false)

- $\alpha$ is typically set to 0.01 (strong), 0.10 (weak) or 0.05 (commonly accepted).

  → value of $\alpha$ can be controlled.

- The value of $\beta$ cannot be specified in advance and depends on the value of the (unknown) population parameter.

- Generally, as $\alpha$ decreases, $\beta$ increases.

# One-Sample Hypothesis Tests

## Illustrating how *β* depends on the True Population Mean

- In the CadSoft example:
  $H_0$: mean response time ≥ 25
  $H_1$: mean response time < 25
- If the true mean was 15, then the sample mean will most likely be less than 25. [less likely to NOT reject $H_0$]
- If the true mean is 24, then the sample mean may or may not be less than 25. [more likely to NOT reject $H_0$]
- The further away the true mean from the hypothesized value, the smaller the value of *β*.

# Improving the Power of the Test

- Power of test = 1- *b*
- Power of test
  - probability of not committing a type II error
  - should be high to make a valid conclusion
- How to ensure sufficient power?
  - Power of test is sensitive to sample size
    - small sample sizes → low power
    - power can be increased by taking larger samples
    - large sample required for small $\alpha$

# Selecting the Test Statistic

- Decision to reject or fail to reject a null hypothesis is based on computing a test statistic from sample data
- Test statistic used depends on type of hypothesis test and certain assumptions about the population
- Test statistics for one-sample hypothesis tests for means:

| Type of Test | Test Statistic |
|---|---|
| One-sample test for mean, $\sigma$ known | $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ |
| One-sample test for mean, $\sigma$ unknown | $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ |
| One-sample test on a proportion | $z = \dfrac{\hat{p} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$ |

# One Sample Hypothesis Tests

**Computing the Test Statistic: An example**

At CadSoft, sample data for 44 customers revealed a mean response time of 21.91 minutes and a sample standard deviation of 19.49 minutes.

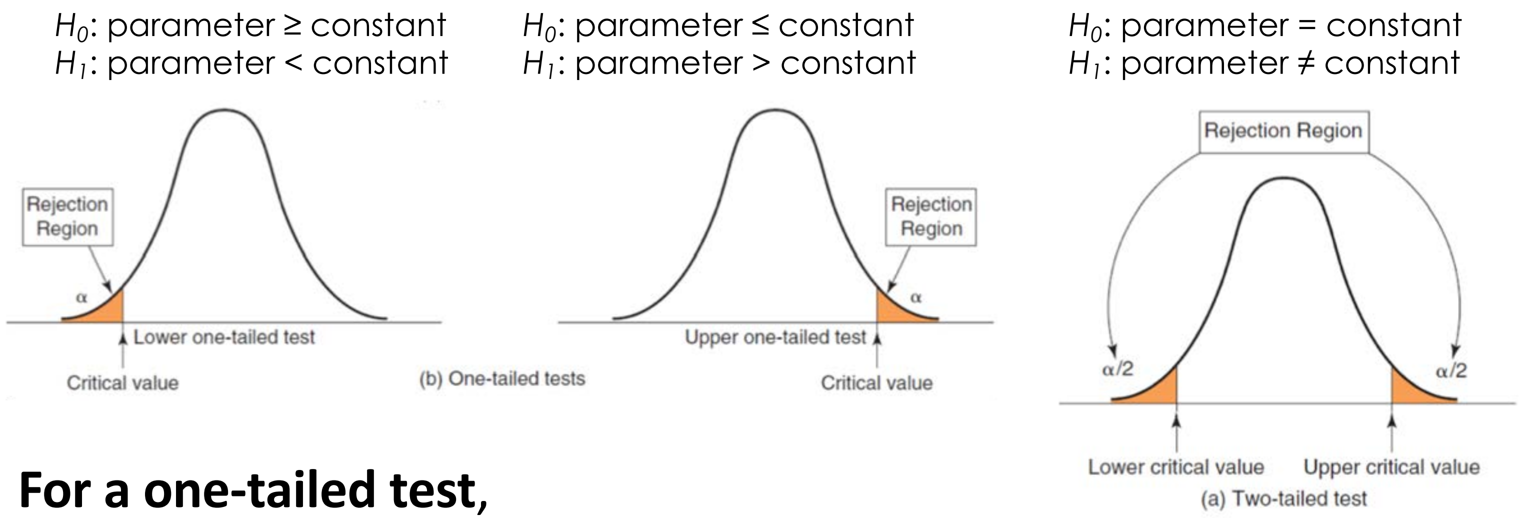
$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{21.91 - 25}{19.49/\sqrt{44}} = \frac{-3.09}{2.938} = -1.05$$

$t$ = -1.05 indicates that the sample mean of 21.91 is 1.05 standard errors below the hypothesized mean of 25 minutes

# Drawing a Conclusion

- reject or fail to reject $H_0$
  - compare value of test statistic to "critical value"
- critical value
  - value from sampling distribution of test statistic when null hypothesis is true at chosen level of significance, $\alpha$
  - divides sampling distribution into rejection region and non-rejection region
- sampling distribution of test statistic
  - usually normal distribution, t-distribution, or some other well-known distribution.
- conclusion
  - reject $H_0$ if test statistic falls into the rejection region; otherwise, we fail to reject it.

# Rejection Regions




**For a one-tailed test,**

- if $H_1$ is stated as $<$, rejection region is in lower tail with critical value = $\propto$
- if $H_1$ is stated as $>$, rejection region is in upper tail with critical value = $\propto$

(think of the inequality as an arrow pointing to the proper tail direction)

**For a two-tailed test,**

- $H_1$ is stated as $\neq$, the rejection region is at both tails with $\propto/2$

as the critical value

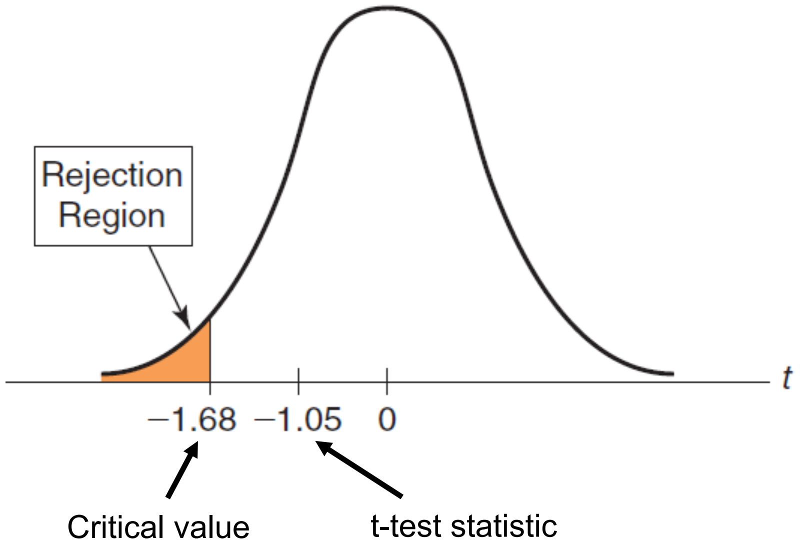
# One-Sample Hypothesis Tests

Cadsoft example (continued):

Finding the Critical Value and Drawing a Conclusion

$t$ = -1.05 does not fall in the rejection region.

Fail to reject $H_0$.




qt(∝ , n-1) = qt(0.05,43)

Even though the sample mean of 21.91 is less than 25, we are unable to conclude that the true population mean is less than 25 minutes as the sampling error is too large.

# One-Sample Hypothesis Tests

Alternative approach: Using p-values

- use *p*-value rather than the critical value.
- *p*-value is the observed significance level.
- *p*-value decision rule: Reject $H_0$ if the *p*-value < α

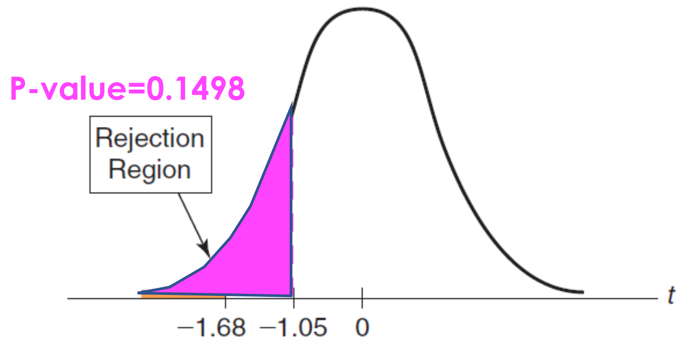
# One-Sample Hypothesis Tests

## Using *p*-Values: Cadsoft example

- *p*-value: left tail area of the observed test statistic, $t$ = -1.05.
- *p*-value = pt(-1.05, 43) in R {t=-1.05, df=43}

  = 0.1498
- Do not reject $H_0$ because the:

  *p*-value (=0.1498) is not less than α (=0.05)



# One-Sample Hypothesis Tests

## Example: Conducting a Two-Tailed Hypothesis Test for the Mean

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Vacation Survey | | | | |
| 2 | | | | | |
| 3 | **Age** | **Gender** | **Relationship Status** | **Number of Vacations per Year** | **Number of Children** |
| 4 | 24 | Male | Married | 2 | 0 |
| 5 | 26 | Female | Married | 4 | 0 |
| 6 | 28 | Male | Married | 2 | 2 |
| 7 | 33 | Male | Married | 4 | 0 |
| 8 | 45 | Male | Married | 2 | 0 |

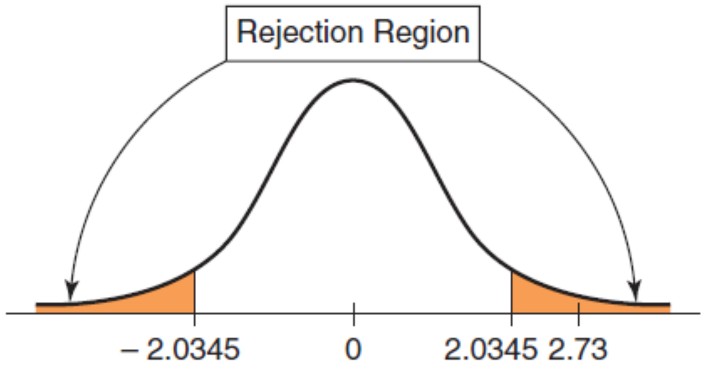
- Vacation Survey data set:
  - test whether average age of respondents is equal to 35
- Set up hypothesis:
  - H0: average age = 35; H1: average age ≠ 35
- Test at $\alpha$= 5% significance level
- For 34 respondents,
  - sample mean = 38.677
  - sample standard deviation = 7.858

# One-Sample Hypothesis Tests

Example (cont'd): Conducting a Two-Tailed Hypothesis Test for the Mean

$H_0$: mean age = 35
$H_1$: mean age ≠ 35

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{(38.677 - 35)}{(7.858/\sqrt{34})} = 2.73$$




Critical value = qt(0.975,33) in R ($\alpha/2$=0.025, df=33)
= 2.0345

We reject $H_0$ because t=2.73 falls in the rejection region.

*p*-value = 2*(pt(2.73,33,lower.tail=FALSE)) =2*(0.00504)
= 0.01008

We reject $H_0$ because p-value=0.01008 is < 0.05

# Confidence Intervals & Hypothesis Tests

~ Conduct hypothesis test using a confidence interval

Suppose you have the following hypothesis:

$H_0: \mu = \mu_0$
$H_1: \mu \neq \mu_0$

To conduct a hypothesis test at 5% level of significance, check:

- if $\mu_0$ does not lie within 95% confidence interval → reject H0
- if $\mu_0$ lies within the 95% confidence interval → cannot reject H0

For one-tailed test:

- lower-tailed test – reject H0 when CI falls entirely below $\mu_0$
- upper-tailed test – reject H0 when CI falls entirely above $\mu_0$

# Two-Sample Hypothesis Tests

~ involve comparing two populations for differences in means, proportions or other population parameters.

Suppose: Population parameter (1) and Population parameter (2) from Population (1) and (2) and $D_0$ is a constant.

*Three forms:*

1. $H_0$: population parameter (1) – population parameter (2) $\leq D_0$
   $H_1$: population parameter (1) – population parameter (2) $> D_0$
2. $H_0$: population parameter (1) – population parameter (2) $\geq D_0$
   $H_1$: population parameter (1) – population parameter (2) $< D_0$
3. $H_0$: population parameter (1) – population parameter (2) $= D_0$
   $H_1$: population parameter (1) – population parameter (2) $\neq D_0$

Usually $D_0 = 0$ if just comparing population parameters

eg in (1) $H_0$: population parameter (1) $\leq$ population parameter (2)
$H_1$: population parameter (1) $>$ population parameter (2)

# Two-Sample Tests for Differences in Means

*Example: Comparing Supplier Performance in Purchase Orders* database

Set up a hypothesis test for determining if the mean lead time for Alum Sheeting ($\mu_1$) is greater than the mean lead time for Durrable Products ($\mu_2$).

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **Purchase Orders** | | | | | | | | | | |
| 2 | | | | | | | | | | | |
| 3 | **Supplier** | **Order No** | **Item No.** | **Item Description** | **Item Cost** | **Quantity** | **Cost per order** | **A/P Terms (Month** | **Order Date** | **Arrival Date** | **Lead Time** |
| 4 | Alum Sheeting | A0443 | 1243 | Airframe fasteners | $ 4.25 | 10,000 | $ 42,500.00 | 30 | 08/08/11 | 08/14/11 | 6 |
| 5 | Alum Sheeting | B0247 | 1243 | Airframe fasteners | $ 4.25 | 9,000 | $ 38,250.00 | 30 | 09/05/11 | 09/12/11 | 7 |
| 6 | Alum Sheeting | B0567 | 1243 | Airframe fasteners | $ 4.25 | 10,500 | $ 44,625.00 | 30 | 10/10/11 | 10/17/11 | 7 |
| 7 | Alum Sheeting | A0223 | 4224 | Bolt-nut package | $ 3.95 | 4,500 | $ 17,775.00 | 30 | 10/15/11 | 10/20/11 | 5 |
| 8 | Alum Sheeting | A0433 | 5417 | Control Panel | $255.00 | 500 | $ 127,500.00 | 30 | 10/20/11 | 10/27/11 | 7 |
| 9 | Alum Sheeting | A0446 | 5417 | Control Panel | $255.00 | 406 | $ 103,530.00 | 30 | 09/01/11 | 09/10/11 | 9 |
| 10 | Alum Sheeting | B0447 | 5634 | Side Panel | $185.00 | 150 | $ 27,750.00 | 30 | 10/25/11 | 11/03/11 | 9 |
| 11 | Alum Sheeting | B0479 | 5634 | Side Panel | $185.00 | 140 | $ 25,900.00 | 30 | 10/29/11 | 11/04/11 | 6 |

$H_0$: ?

$H_1$: ?

# Two-Sample Hypothesis Tests

Given $\mu_1$ is the population mean for Alum Sheeting and $\mu_2$ is the population mean for Durrable Products, and we are testing if $\mu_1$ is greater than $\mu_2$

Hypotheses:

$$H_0: \mu_1 - \mu_2 \leq 0 \qquad or \; H_0: \mu_1 \leq \mu_2$$
$$H_1: \mu_1 - \mu_2 > 0 \qquad H_1: \mu_1 > \mu_2$$

Using R to compute sample means:

```
> aggdata <-Purchase_Orders %>% group_by(Supplier) %>%
summarize(averLT = mean(`Lead Time`))
> print(aggdata)
```

| Supplier | averLT |
|---|---|
| <chr> | <dbl> |
| Alum Sheeting | 7 |
| Durrable Products | 4.92 |
| Fast-Tie Aerospace | 8.47 |
| Hulkey Fasteners | 6.47 |
| Manley Valve | 6.45 |
| Pylon Accessories | 8 |
| Spacetime Technologies | 15.2 |
| Steelpin Inc. | 10.2 |

$\overline{x}_1 = 7.00$
$\overline{x}_2 = 4.92$

**Significant Difference? Sampling Error**

# Two-Sample Hypothesis Tests

## Testing the Hypothesis for Supplier Lead-Time Performance

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

**Population variances are known**

*Conduct z-test*

*Formula:* $z = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\dfrac{\sigma_1^{\,2}}{n_1} + \dfrac{\sigma_2^{\,2}}{n_2}}}$

**Population variances are unknown**

*Conduct t-test*

*Formula:* $t = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\dfrac{s_1^{\,2}}{n_1} + \dfrac{s_2^{\,2}}{n_2}}}$

* no need to memorize the equations

*where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two samples, $\Delta$ is the hypothesized difference between the population means (0 if comparing the two means), $\sigma_1$ and $\sigma_2$ are standard deviations of the two populations, $s_1$ and $s_2$ are standard deviations of the two samples, and $n_1$ and $n_2$ are sizes of the two samples. The number of degrees of freedom is the smaller of $n_1$-1 and $n_2$-1.*

# Two-Sample Hypothesis Tests

## Testing the Hypothesis for Supplier Lead-Time Performance

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_1: \mu_1 - \mu_2 > 0$$

*Using R: Conduct t-test since population variances unknown*

```
> View(Purchase_Orders)
> PO<-Purchase_Orders
> PO_sub<-subset(PO, Supplier=='Alum Sheeting' | Supplier=='Durrable Products')
> View(PO_sub)
> View(PO_sub)
> PO_sub1<-PO[which(PO$Supplier=='Alum Sheeting' | PO$Supplier=='Durrable Products'), ]
> t.test(`Lead Time`~Supplier, alternative= "greater", data=PO_sub)

        Welch Two Sample t-test

data:  Lead Time by Supplier
t = 3.828, df = 9.5306, p-value = 0.001818
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.088604      Inf
sample estimates:
    mean in group Alum Sheeting mean in group Durrable Products
                       7.000000                        4.923077
```

t.test(… var.equal=FALSE…)

**Reject H0; Accept H1**

# Two-Sample Hypothesis Tests

What if we would like to test if mean lead time for Alum Sheeting ($\mu_1$) is less than the mean lead time for Durrable Products ($\mu_2$).

That is,

$H_0: \mu_1 - \mu_2 \geq 0; \quad H_1: \mu_1 - \mu_2 < 0$

```
> ttestres <- t.test(newdata$`Lead Time`~ newdata$Supplier, alternative="less")
> print(ttestres)

        Welch Two Sample t-test

data:  newdata$`Lead Time` by newdata$Supplier
t = 3.828, df = 9.5306, p-value = 0.9982
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 3.065242
sample estimates:
   mean in group Alum Sheeting mean in group Durrable Products 
                      7.000000                        4.923077 
```

**For α=0.05, critical value = qt(0.05,10) = -1.812461**

**Cannot Reject $H_0$ (p-value =0.998 > 0.05; t=3.83>-1.81)**

# Two-Sample Hypothesis Tests

~ if the two population distributions can be assumed to have equal variance, then the t-statistics will be computed differently.

s1 and s2 are pooled together, each weighted by the number of cases in each sample.

*Formula for pooled estimator of* $\sigma^2$: $s^2{}_p = \dfrac{(n_1 - 1)s_1{}^2 + (n_2 - 1)s_2{}^2}{n_1 + n_2 - 2}$

*Formula:* $t = \dfrac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{s^2{}_p\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

*The number of degrees of freedom is* $n_1 + n_2 - 2$.

* no need to memorize these equations

In R, we conduct t.test with var.equal=TRUE

# Paired Two-Sample Hypothesis Tests

Paired Sample t-test: used when comparing means from two samples that are naturally paired or matched.

Examples:

- a measurement taken at two different times (e.g. performance before and after attending a programme)
- a measurement taken under two different conditions (e.g. body temperature taken using two different types of thermometers)
- measurements taken from two halves/sides of a subject or experimental unit (e.g. writing speed with left and right hand)

Hypothesis:

$H_0$: $\mu_D \{\geq, \ \leq, \ =\} D_0$

$H_1$: $\mu_D \{<, \ >, \ \neq\} D_0$

where $\mu_D$ is the difference between the paired population means; and $D_0$ is usually 0 for comparing the two means

Test Statistic:

$$T = \frac{\bar{d} - D_0}{s_d / \sqrt{n}}$$

where there are *n* pairs, $\bar{d}$ is the mean and $s_d$ is the standard deviation of the paired sample differences, df = n-1

# Paired Two-Sample Hypothesis Tests

Example:

- Use the *Pile Foundation* data to test for a difference in the means of the estimated and actual pile lengths.

Use a 5% significance level.

$H_0$: $\mu_D = 0$

$H_1$: $\mu_D \neq 0$

where $\mu_D$ is the difference between the 2 population means.

| | A | B | C |
|---|---|---|---|
| 1 | Pile Foundation Data | | |
| 2 | | | |
| 3 | Pile | Estimated | Actual |
| 4 | Number | Pile Length (ft.) | Pile Length (ft.) |
| 5 | 1 | 10.58 | 18.58 |
| 6 | 2 | 10.58 | 18.58 |
| 7 | 3 | 10.58 | 18.58 |
| 8 | 4 | 10.58 | 18.58 |
| 9 | 5 | 10.58 | 28.58 |
| 10 | 6 | 10.58 | 26.58 |
| 11 | 7 | 10.58 | 17.58 |
| 12 | 8 | 10.58 | 27.58 |
| 13 | 9 | 10.58 | 27.58 |
| 14 | 10 | 10.58 | 37.58 |

# Two-Sample Hypothesis Tests

## Example ( cont'd): Using the Paired Two-Sample Test for Means

*t-Test: Paired Two-Sample for Means*

```
> ttestPF<-t.test(Pile_Foundation$PileLengthEst,Pile_Foundation$PileLengthAct, paired = TRUE)
> print(ttestPF)

	Paired t-test

data:  Pile_Foundation$PileLengthEst and Pile_Foundation$PileLengthAct
t = -10.912, df = 310, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.528856 -5.228508
sample estimates:
mean of the differences
              -6.378682
```

**p-value < 0,05; sufficient evidence to reject H0 and accept H1**

# Test for Equality of Variances

- Test for equality of variances between two samples using a new type of test, the *F-test*.
  - To use this test, we must assume that both samples are drawn from normal populations.
- Hypotheses:

$$H_0: \sigma_1^2 - \sigma_2^2 = 0 \qquad or\ H_0: \frac{\sigma_1{}^2}{\sigma_2{}^2} = 1$$

$$H_1: \sigma_1^2 - \sigma_2^2 \neq 0 \qquad or\ H_1: \frac{\sigma_1{}^2}{\sigma_2{}^2} \neq 1$$

- F-test statistic:

$$F = \frac{s_1^2}{s_2^2}$$

# F-Distribution

- The *F*-distribution has two degrees of freedom, one associated with the numerator of the *F*-statistic, $n_1 - 1$, and one associated with the denominator of the *F*-statistic, $n_2 - 1$.
- Population with larger variance will be assigned the numerator
- F-table below provides only upper-tail critical values, and the distribution is not symmetric.

**Upper critical values of the *F* distribution for numerator degrees of freedom $\nu_1$ and denominator degrees of freedom $\nu_2$, 5% significance level**

| $\nu_2$ \ $\nu_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 161.448 | 199.500 | 215.707 | 224.583 | 230.162 | 233.986 | 236.768 | 238.882 | 240.543 | 241.882 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 | 19.296 | 19.330 | 19.353 | 19.371 | 19.385 | 19.396 |
| 3 | 10.128 | 9.552 | 9.277 | 9.117 | 9.013 | 8.941 | 8.887 | 8.845 | 8.812 | 8.786 |
| 4 | 7.709 | 6.944 | 6.591 | 6.388 | 6.256 | 6.163 | 6.094 | 6.041 | 5.999 | 5.964 |
| 5 | 6.608 | 5.786 | 5.409 | 5.192 | 5.050 | 4.950 | 4.876 | 4.818 | 4.772 | 4.735 |
| 6 | 5.987 | 5.143 | 4.757 | 4.534 | 4.387 | 4.284 | 4.207 | 4.147 | 4.099 | 4.060 |
| 7 | 5.591 | 4.737 | 4.347 | 4.120 | 3.972 | 3.866 | 3.787 | 3.726 | 3.677 | 3.637 |
| 8 | 5.318 | 4.459 | 4.066 | 3.838 | 3.687 | 3.581 | 3.500 | 3.438 | 3.388 | 3.347 |
| 9 | 5.117 | 4.256 | 3.863 | 3.633 | 3.482 | 3.374 | 3.293 | 3.230 | 3.179 | 3.137 |
| 10 | 4.965 | 4.103 | 3.708 | 3.478 | 3.326 | 3.217 | 3.135 | 3.072 | 3.020 | 2.978 |

# Applying the F-Test for Equality of Variances

- Determine whether the variance of lead times is the same for Alum Sheeting and Durrable Products in the *Purchase Orders* data.
  - H0: $V_A - V_D = 0$ ;    H1: $V_A - V_D \neq 0$
  - The variance of the lead times for Alum Sheeting is larger than the variance for Durable Products, so was assigned to be numerator (if done manually).

```
> var.test(`Lead Time`~Supplier)

        F test to compare two variances

data:  Lead Time by Supplier
F = 3.4667, num df = 7, denom df = 12, p-value = 0.05719
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.9612235 16.1748764
sample estimates:
ratio of variances
          3.466667
```

**Critical F-value**
**=qf(.975, df1=7, df2=12)**
**=3.6065**
***F* = 3.47 < F-critical$H_0$ cannot be rejected**

**p>0.05**
**$H_0$ cannot be rejected**
**Hence, we conclude that there is no significant difference in variances at 5% level of significance.**

# Testing Equality/homogeneity of Variances – Other tests

- Bartlett’s Test
  - normally distributed data where there are 2 or more samples
  - R function bartlett.test()
- Levene’s test
  - an alternative to Bartlett’s test but less sensitive to departures from normality
  - R function leveneTest() in the car package
- Fligner-Killeen test
  - a non-parametric test which is very robust to departures from normality
  - R function fligner.test()

# Analysis of Variance (ANOVA)

- Used to compare the means of two or more population groups.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m$$

$$H_1: \text{at least one mean is different from the others}$$

- ANOVA derives its name from the fact that we are analyzing variances in the data.
- ANOVA measures variation between groups relative to variation within groups.
- Test statistic has an F-distribution so if the F-statistic is large enough based on the level of significance chosen and exceeds a critical value, we would reject H0.

# Analysis of Variance

## Example: Differences in *Insurance Survey* Data

Average health insurance satisfaction scores (1-5 scale) by education level

| | College Graduate | Graduate Degree | Some College |
|---|---|---|---|
| | 5 | 3 | 4 |
| | 3 | 4 | 1 |
| | 5 | 5 | 4 |
| | 3 | 5 | 2 |
| | 3 | 5 | 3 |
| | 3 | 4 | 4 |
| | 3 | 5 | 4 |
| | 4 | 5 | |
| | 2 | | |
| Average | 3.444 | 4.500 | 3.143 |
| Count | 9 | 8 | 7 |

**Means are quite different.**
**Are they significantly different?**

# Analysis of Variance

*Let* $\mu_1, \mu_2$, and $\mu_3$ be the mean satisfaction for individuals with College Degree, Graduate degree and some college education level respectively.

$H_0$: $\mu_1 = \mu_2 = \mu_3$

$H_1$: at least one mean is different from the others

| nder | Education | Marital Status | Years Employed | Satisfaction* | Prem |
|---|---|---|---|---|---|
| | Some college | Divorced | 4 | 4 | N |
| | Some college | Divorced | 2 | 1 | N |
| | Graduate degree | Widowed | 26 | 3 | N |
| | Some college | Married | 9 | 4 | N |
| | Graduate degree | Married | 6 | 4 | N |
| | Graduate degree | Married | 10 | 5 | N |
| | College graduate | Married | 4 | 5 | N |
| | College graduate | Divorced | 9 | 3 | N |
| | Graduate degree | Married | 6 | 5 | N |

Assign a variable to results of ANOVA (e.g. \`your.aov\`)

In R:

```
> your.aov=aov(`Satisfaction*`~Education)
> your.aov
```

Classification factor should be character type or factor object

```
Call:
   aov(formula = `Satisfaction*` ~ Education)

Terms:
                Education Residuals
Sum of Squares   7.878968 21.079365
Deg. of Freedom         2        21

Residual standard error: 1.001888
Estimated effects may be unbalanced
```

```
> summary(your.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
Education    2  7.879   3.939   3.925 0.0356 *
Residuals   21 21.079   1.004
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P<0.05, hence we can reject $H_0$ and accept there is at least one mean satisfaction is different from the others.

# Analysis of Variance

- ANOVA does not compare mean across groups/samples directly

- Computes F statistic = Between group variance / within group variance,

where df (numerator) = k-1; df (denominator)=N-k for N observations and k samples/groups

```
> summary(your.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
Education    2  7.879   3.939   3.925 0.0356 *
Residuals   21 21.079   1.004
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

between group variance is the mean squares between groups

within group variance is the mean squares within groups (or residuals)

| Source of Variation | Sums of Squares (SS) | Degrees of Freedom (df) | Mean Squares (MS) | F |
|---|---|---|---|---|
| Between Treatments/Groups | $\text{SSB} = \Sigma n_j (\bar{X}_j - \bar{X})^2$ | k-1 | $\text{MSB} = \frac{SSB}{k-1}$ | $F = \frac{\text{MSB}}{\text{MSE}}$ |
| Error (or Residual) | $\text{SSE} = \Sigma\Sigma (X - \bar{X}_j)^2$ | N-k | $\text{MSE} = \frac{\text{MSE}}{N-k}$ | |
| Total | $\text{SST} = \Sigma\Sigma (X - \bar{X})^2$ | N-1 | | |

where

- X = individual observation,
- $\bar{X}_j$ = sample mean of the j[th] treatment (or group),
- $\bar{X}$ = overall sample mean (or called the grand mean),
- k = number of independent comparison groups,
- N = total number of observations or total sample size.

- Intuitively, between group variance tells us how far each sample mean is from the grand mean; within group variance tells us how far the observations are from the mean of their own group/sample. If group means are significantly different, we would expect the between group variance to be larger than the within group variance.
- Test if F statistic > F-critical at 5% level of significance

# ANOVA Assumptions

- Independence, Normality, and homogeneity of the variances
  1. Randomly and independently obtained
  2. Normally distributed
  3. Have equal variances

- 1 is easily validated if random samples were chosen
- ANOVA is fairly robust to departures from normality
- If sample sizes are equal, violation of the third assumption does not have serious effects, but with unequal sample sizes, it can. If variances are unequal, we can use Welch ANOVA test: *welch_anova_test(data,formula)*
- When comparing sample means of two populations that may have different variances, use t-test rather than ANOVA

# Post-Hoc Tests

- In ANOVA tests, when p-value is less than significance level

  → reject H0, accept not all group means are equal; but cannot determine which pair of means is significantly different

- Post-Hoc (a.k.a multiple comparisons) test explores differences between multiple group means while controlling experiment-wise(or family-wise) error rate
  - each hypothesis test has a type 1 error rate (false positive);
  - with one test, error rate = significance level
  - with more tests, chance of false positive increases and rapidly exceeds the significance level (see table)
  - post-hoc test calculates the significance level given the error level desired (e.g. 0.05)

| Groups (N) | # of Comparisons ( C = (N*(N-1))/2 ) | Experiment-wise error rate ( 1-(1-alpha)^C ) |
|---|---|---|
| 2 | 1 | 0.05 |
| 3 | 3 | 0.142625 |
| 4 | 6 | 0.264908109 |
| 5 | 10 | 0.401263061 |
| 6 | 15 | 0.53670877 |
| 7 | 21 | 0.659438374 |
| 8 | 28 | 0.762173115 |
| 9 | 36 | 0.842220785 |
| 10 | 45 | 0.900559743 |

# Post-Hoc Tests

- One-way ANOVA
  - use Tukey Honest Significant Differences (TukeyHSD function in R)

```
> TukeyHSD(your.aov)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = IS$`Satisfaction*` ~ IS$Education)

$`IS$Education`
                                  diff        lwr         upr     p adj
Graduate degree-College graduate  1.0555556 -0.1715336  2.28264475 0.1003252
Some college-College graduate    -0.3015873 -1.5742334  0.97105876 0.8230559
Some college-Graduate degree     -1.3571429 -2.6641246 -0.05016107 0.0409193
```

- Welch's ANOVA
  - use Games-Howell post hoc test (games_howell_test {rstatix} in R)

```
> IS %>% games_howell_test(`Satisfaction*`~Education)
# A tibble: 3 x 8
  .y.          group1       group2       estimate conf.low conf.high p.adj p.adj.signif
* <chr>        <chr>        <chr>           <dbl>    <dbl>     <dbl> <dbl> <chr>
1 Satisfacti… College gra… Graduate d…      1.06  -0.0668     2.18  0.067 ns
2 Satisfacti… College gra… Some colle…     -0.302 -1.83       1.23  0.859 ns
3 Satisfacti… Graduate de… Some colle…     -1.36  -2.82       0.105 0.069 ns
```