Lecture on Analysis of Data from Designed Experiments

Wim Krijnen Measurement Models for a Chemical Quantity

May 1, 2023

Outline

Steps of Statistical tests

1. Specify the research question pertaining to the aim of the research. Give the parameter of interest.
2. Give the null and the alternative hypothesis with respect to the engineering parameter.
3. Specify how the test is conducted: give test statistic, significance level, rejection region.
4. Obtain data from engineering experiment, perform the computations, and give decision on null-hypothesis.
5. Formulate general conclusion in non-statistical terms pertaining to aim of research.

Independent two-sample t-test

- Two groups of normally distributed measurements $y_{11}, \cdots, y_{1n_1}$; $y_{21}, \cdots, y_{2n_2}$
- Test the hypothesis

$$H_0 : \mu_1 - \mu_2 = 0, \quad \text{against} \quad H_A : \mu_1 - \mu_2 \neq 0$$

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(\text{df}), \quad df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}.$$

- Decision: Reject null hypothesis if

$$|t| > t_{1-\alpha/2}(\text{df}) \Leftrightarrow p - value < 0.05 = \alpha.$$

Arsenic concentration in public drinking water

- Metropolitan Phoenix $3, 7, 25, 10, 15, 6, 12, 25, 15, 7$ ppb
- Rural Arizona $48, 44, 40, 38, 33, 21, 20, 12, 1, 18$ ppb.
- $H_0 : \mu_1 - \mu_2 = 0$ against $H_1 : \mu_1 - \mu_2 \neq 0$.
- $n_1 = n_2 = 10$, $\overline{y}_1 = 12.5, \overline{y}_2 = 27.5$, $s_1 = 7.63, s_2 = 15.3$ (ppb).

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} = \frac{\left(\frac{7.63^2}{10} + \frac{15.3^2}{10}\right)^2}{\frac{(7.63^2/10)^2}{9} + \frac{(15.3^2/10)^2}{9}} = 13.196$$

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{12.5 - 27.5}{\sqrt{\frac{7.63^2}{10} + \frac{15.3^2}{10}}} = -2.77$$

- $t_{0.975,13.196} = 2.160 < 2.77 = |t|$.
- Reject: null hypothesis of equal means
- Conclusion: Arsenic concentration in public drinking water in Rural Arizona is larger than in Metropolitan Phoenix

Example: Example Arsenic concentration

```
> x <- c(3,7,25,10,15,6,12,25,15,7)    #Metropolitan P
> y <- c(48,44,40,38,33,21,20,12,1,18) #Rural Arizona
> t.test(x,y)
        Welch Two Sample t-test
data:  x and y
t = -2.7669, df = 13.196, p-value = 0.01583
alternative hypothesis: true difference in means is nc
95 percent confidence interval:
 -26.694067  -3.305933
sample estimates:
mean of x mean of y
     12.5      27.5
```

testing for equal variances

Two sets of measurements are equally precise if their variances are equal; $H_0 : \sigma_1^2 = \sigma_2^2$

```
> y1 <- c(3,7,25,10,15,6,12,25,15,7)
> y2 <- c(48,44,40,38,33,21,20,12,1,18)
> var.test(y1,y2)
        F test to compare two variances
data:  y1 and y2
F = 0.24735,num df=9, denom df=9, p-value=0.04936
alternative hypothesis: true ratio of variances is not
95 percent confidence interval:
 0.06143758 0.99581888
ratio of sample variances
         0.2473473
```

Decision: Reject $H_0 : \sigma_1^2 = \sigma_2^2$; Conclusion: Variances are not equal

Steps of Statistical tests

1. Specify the research question pertaining to the aim of the research. Give the parameter of interest.
2. Give the null and the alternative hypothesis with respect to the engineering parameter.
3. Specify how the test is conducted: give test statistic, significance level, rejection region.
4. Obtain data from engineering experiment, perform the computations, and give decision on null-hypothesis.
5. Formulate general conclusion in non-statistical terms pertaining to aim of research.

Outline

Model formulation for One-way Analysis of Variance

- The model for one-way analysis of variance can be written as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \ \{\varepsilon_{11}, \cdots, \varepsilon_{IJ}\} \sim IIND(0, \sigma^2), \tag{1}$$

  where μ is the grand mean, α_i the i-th effect of factor α, and ε_{ij} te error terms.
- To estimate this we need model constraints identifying parameters:
  Constraint type 1: $\sum_{i=1}^{I} \alpha_i = 0$
  Constraint type 2: $\alpha_1 = 0$
- Test H_0 equals means; $H_0 : \mu_i = \mu$ for all i against
  H_A : one of means differs from others
- Testing: the larger model against the smaller
  Model 2: $E[Y_{ij}] = \mu + \alpha_i = \mu_i$,
  Model 1: $E[Y_{ij}] = \mu$

F-test from linear regression

- Model 1 with q parameters $\boldsymbol{\beta}_1$ is nested within the larger Model 2 with p parameters $\boldsymbol{\beta}_2$
- Then F-statistic can be used to test Model 1 against model 2.

$$
\begin{aligned}
F &= \frac{\left(\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_1) - \mathrm{RSS}(\widehat{\boldsymbol{\beta}}_2)\right)/(p-q)}{\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_2)/(n-p)} \\
&= \frac{\left(\|\mathbf{y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1\|^2 - \|\mathbf{y} - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\|^2\right)/(p-q)}{\|\mathbf{y} - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2\|^2/(n-p)} \\
&\sim F_{p-q,n-p}
\end{aligned}
$$

  the latter is the F distribution with $p-q, n-p$ degrees of freedom.
- Model 1 is rejected if $F > F^{(\alpha)}_{p-q,n-p}$; significance level $\alpha = 0.05$.
- Reduction in RSS by Model 2 in nominator is sufficiently large.
- The P-value is computed as $P(F_{p-q,n-p} > F)$ the probability of larger F-values than the F observed.

Design matrix

- The design matrix $\mathbf{X}$ is filled with zeros and ones according to whether the run is executed for level i; for 3 level we have

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix},$$

  length of **1** is number of observations n within level
- Measurements are in vector $\mathbf{y}$, so that

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} n & 0 & 0 \\ 0 & n & 0 \\ 0 & 0 & n \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^n y_{1j} \\ \sum_{j=1}^n y_{2j} \\ \sum_{j=1}^n y_{3j} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{bmatrix},$$

Contrast matrix

- For the estimator we have

$$\begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{bmatrix} = \hat{\beta} \sim N\left(\beta, \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1}\right)$$

$$= N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1/n & 0 & 0 \\ 0 & 1/n & 0 \\ 0 & 0 & 1/n \end{bmatrix} \right) = N\left(\boldsymbol{\mu}, \frac{\sigma^2}{n}\mathbf{I}\right)$$

- To find interpretable results we need a contrast matrix

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

Distribution of Effects from contrast matrix

- Since linear combinations of normally distributed random variables are normally distributed, we find

$$\mathbf{C}\hat{\beta} \sim N\left(\mathbf{C}\beta, \sigma^2\mathbf{C}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{C}^T\right),$$

- So that

$$\mathbf{C}\hat{\beta} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \bar{y}_3 - \bar{y}_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \end{bmatrix}, \sigma^2/n \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}\right).$$

- For practical purposes we replace σ^2 by its estimator $\hat{\sigma}^2$.
- All software systems give this output when a single factor indicates the level to which each run belongs.

Elementary formulation of One-way analysis of variance

- Model for testing of equality of three or more means

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i = 1, \cdots, p,\ j = 1, \cdots, n_i,$$

- μ is the grand mean,
- α_i the effect of treatment i, and
- error terms ε_{ij} are a random sample from $N(0, \sigma^2)$.
- mean for treatment i is $\mu_i = \mu + \alpha_i$
- null hypothesis of equality of means

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_p \Leftrightarrow H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p$$

- Mean for the i-th treatment $\bar{Y}_{i.} = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$,
- overall mean $\bar{Y}_{..} = \frac{1}{np} \sum_{j=1}^{n} \sum_{i=1}^{p} Y_{ij}$.

F-test for one-way analysis of variance

- sums of squares due to treatment and residual are

$$SS_{Tr} = \sum_{i=1}^{p} n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2 \qquad SS_{Res} = \sum_{i=1}^{p}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i.})^2,$$

- Our test-statistic

$$F = \frac{SS_{Tr}/(p-1)}{SS_{Res}/(p(n-1))} \sim F_{p-1,p(n-1)},$$

- Decision: Reject null-hypothesis, if

$$p - \text{value} = P(F_{p-1,p(n-1)} \geq F) \leq \alpha = 0.05$$

- Often first treatment is control type of condition (standard), then treatment effects are

$$\beta_0 = \mu_1;\ \beta_1 = \mu_2 - \mu_1;\ \beta_2 = \mu_3 - \mu_1, \cdots, \beta_p = \mu_p - \mu_1.$$

The beta coefficients are the differences in means (target values)

- These coefficients or sizes of effects are estimated by linear regression.

Example: Analysis of variance (ANOVA)

- Experiment on low-pressure vapor deposition of polysilicon was conducted in a large-capacity reactor at Semantech
- Three independent measurements at four different positions in table below
- Means indicate first position to have larger thickness.

Table: Film thickness at four positions in a reactor.

measurement	Pos1	Pos2	Pos3	Pos4
1	2.76	1.43	2.34	0.94
2	5.67	1.70	1.97	1.36
3	4.49	2.19	1.47	1.65
mean $\overline{y}$	4.31	1.77	1.93	1.32
sd s	1.46	0.39	0.44	0.36

- Testing could suggest to take these differences serious

Inspection of Analysis of variance table

Table: ANOVA table on film thickness at four positions in a reactor.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Treatm (factor)	3	16.22	5.41	8.29	0.0077
Residuals	8	5.22	0.65		

- df treatment is $p - 1 = 4 - 1$
- df residuals is $p(n - 1) = 4(3 - 1) = 8$
- ANOVA table gives F-value 8.29
- F(3,8) distribution gives p-value 0.0077
- Reject the null hypothesis of equal means.
- Conclusion: There are differences in mean corresponding to the positions
- We do not know yet which differ!

Table: Effect estimation of film thickness at four positions in a reactor.

	Estimate	SE	t-value	Pr(>|t|)
$\hat{\beta}_0 = \hat{\mu}_1$ (Intercept)	4.31	0.46	9.23	0.0000
$\hat{\beta}_1 = \hat{\mu}_2 - \hat{\mu}_1$	-2.53	0.66	-3.84	0.0049
$\hat{\beta}_2 = \hat{\mu}_3 - \hat{\mu}_1$	-2.38	0.66	-3.61	0.0069
$\hat{\beta}_3 = \hat{\mu}_4 - \hat{\mu}_1$	-2.99	0.66	-4.53	0.0019

- The differences in mean wrt reference Position 1 are -2.53, -2.38, and -2.99. (Position effects)
- Decision: Marginal t-tests indicate rejection H_0 of equal means.
- Conclusion: Data provide strong evidence that mean Position 1 is larger than that of the other three positions.
- This standard output from any statistical analysis of variance software

Using R its apply functions

```
> y <- c(2.76,5.67,4.49,1.43,1.70,2.19,2.34,1.97,
  1.47,0.94,1.36,1.65)
> fac<-gl(4,3,labels=c("Pos1", "Pos2","Pos3","Pos4"))
  # construct factor indicating Positions of measureme
> dfa <- data.frame(y,fac)
> boxplot(y ~ fac, data=dfa, sd)
> points(y ~ fac, data=dfa,col=fac)
> with(dfa, tapply(y, fac, mean))
    Pos1     Pos2     Pos3     Pos4
4.306667 1.773333 1.926667 1.316667
> with(dfa, tapply(y, fac, sd))
     Pos1      Pos2      Pos3      Pos4
1.4636370 0.3852705 0.4366158 0.3569781
```

Using R its aggregate functions

Use model notation in aggregate to compute means and SD for measurements at each level of a factor

```
> aggregate(y ~ fac, data=dfa, mean)
   fac        y
1 Pos1 4.306667
2 Pos2 1.773333
3 Pos3 1.926667
4 Pos4 1.316667
> aggregate(y ~ fac, data=dfa, sd)
   fac         y
1 Pos1 1.4636370
2 Pos2 0.3852705
3 Pos3 0.4366158
4 Pos4 0.3569781
```

Analysis of variance by lm function

```
> summary(mod <- lm(y ~ fac, dfa))
Residuals:
     Min       1Q   Median       3Q      Max
-1.54667 -0.35167  0.04333  0.35333  1.36333
Coefficients:
            Estimate Std.Error t val Pr(>|t|)
(Intercept)   4.3067  0.4663   9.237 1.53e-05 ***
facPos2      -2.5333  0.6594  -3.842  0.00493 **
facPos3      -2.3800  0.6594  -3.609  0.00689 **
facPos4      -2.9900  0.6594  -4.535  0.00191 **
Residual standard error: 0.8076 on 8 degrees of freedo
Multiple R-squared:0.7566, Adjusted R-squared: 0.6653
F-statistic: 8.29 on 3 and 8 DF, p-value: 0.007747
```

12 measurements - 4 parameters = df 8

Analysis of variance table from R its aov function

```
> summary(fm1 <- aov(y ~ fac, dfa))
            Df Sum Sq Mean Sq F value  Pr(>F)
fac          3 16.220   5.407    8.29 0.00775 **
Residuals    8  5.217   0.652
```

- Same anova table as from lm
- We need the R object fm1 as output from aov for the Tukey procedure below

Comparing all possible differences adjusting for multiple testing

```
> TukeyHSD(fm1, "fac")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ fac, data = dfa)

$fac
            diff      lwr      upr     p adj
Pos2-Pos1 -2.533   -4.644   -0.421   0.0206120
Pos3-Pos1 -2.380   -4.491   -0.268   0.0283387
Pos4-Pos1 -2.990   -5.101   -0.878   0.0082686
Pos3-Pos2  0.153   -1.958    2.264   0.9952095
Pos4-Pos2 -0.456   -2.568    1.654   0.8971196
Pos4-Pos3 -0.610   -2.721    1.501   0.7927948
```