Analysis of variance

(Module 8)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019

Contents

Aims of this module

- Introduce the **analysis of variance** technique, which builds upon the variance decomposition ideas in previous modules.
- Revisit linear regression and apply the ideas of hypothesis testing and analysis of variance.
- Discuss ways to derive optimal hypothesis tests.

Overview

- **Analysis of variance (ANOVA).** Comparisons of more than two groups
- **Regression.** Hypothesis testing for simple linear regression
- **Likelihood ratio tests.** A method for deriving the best test for a given problem

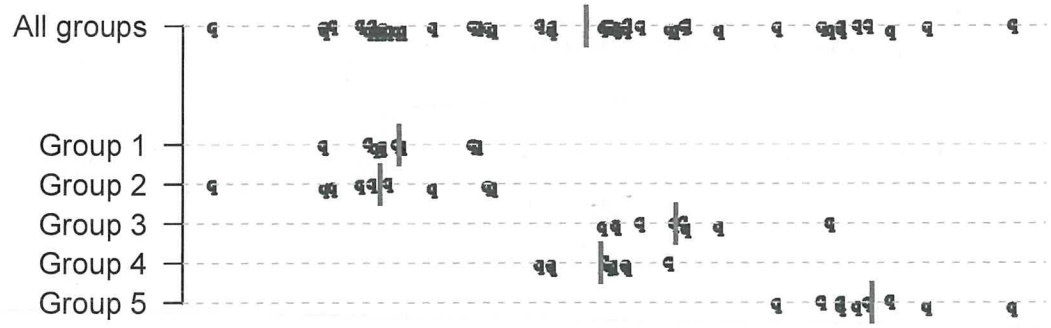
1 Analysis of variance (ANOVA)

1.1 Introduction

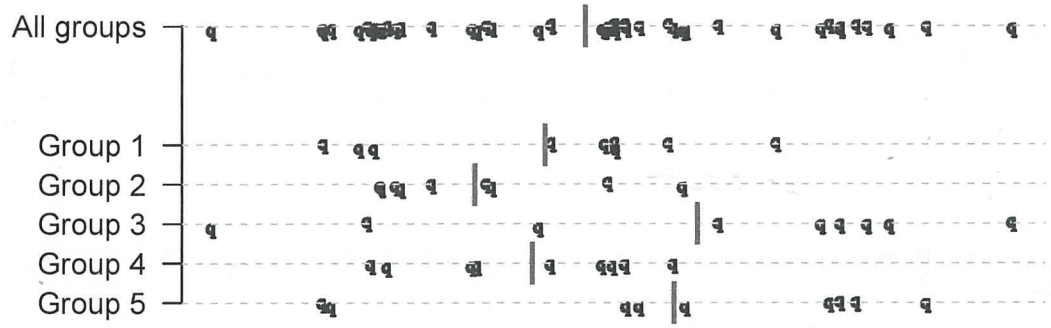
Analysis of variance: introduction

- Initial aim: compare the means of **more than two** populations
- Broader and more advanced aims:
  - Explore components of variation
  - Evaluate the fit a (general) linear model
- Formulated as hypothesis tests
- Referred to as *analysis of variance*, or *ANOVA* for short
- Involves comparing different summaries of variation
- Related to the 'analysis of variance' and 'variance decomposition' formulae we derived previously

Example: large variation between groups



Example: smaller variation between groups



1.2 One-way ANOVA

ANOVA: setting it up

- We have random samples from k populations, each having a normal distribution
- We sample n_i iid observations from the ith population, which has mean μ_i
- All populations assumed have the **same** variance, σ^2
- Question of interest: do the populations all have the same mean?
- Hypotheses:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu \quad \text{versus} \quad H_1 : \bar{H}_0$$

($\bar{H}_0$ means 'not H_0')

- This model is known as a *one-way ANOVA*, or *single-factor ANOVA*

Notation

Population	Sample	Statistics	
$\mathrm{N}(\mu_1, \sigma^2)$	$X_{11}, X_{12}, \ldots, X_{1n_1}$	$\bar{X}_{1\cdot}$	S_1^2
$\mathrm{N}(\mu_2, \sigma^2)$	$X_{21}, X_{22}, \ldots, X_{2n_2}$	$\bar{X}_{2\cdot}$	S_2^2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathrm{N}(\mu_k, \sigma^2)$	$X_{k1}, X_{k2}, \ldots, X_{kn_k}$	$\bar{X}_{k\cdot}$	S_k^2
	Overall	$\bar{X}_{\cdot\cdot}$	

$$n = n_1 + \cdots + n_k \quad \text{(total sample size)}$$

$$\bar{X}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij} \quad \text{(group means)}$$

$$\bar{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{X}_{i\cdot} \quad \text{(grand mean)}$$

Sum of squares (definitions)

- We now define statistics each called a *sum of squares* (*SS*)
- The *total SS* is:

$$SS(TO) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot\cdot})^2$$

- The *treatment SS*, or *between groups SS*, is:

$$SS(T) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2$$

- The *error SS*, or *within groups SS*, is:

$$SS(E) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2 = \sum_{i=1}^{k}(n_i - 1)S_i^2$$

Analysis of variance decomposition

- It turns out that:

$$SS(TO) = SS(T) + SS(E)$$

- This is similar to the analysis of variance formulae we derived earlier, in simpler scenarios (iid model, regression model)
- We will use this relationship as a basis to derive a hypothesis test
- Let's first prove the relationship...
- Start with the 'add and subtract' trick:

$$\begin{aligned}
SS(TO) &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot} + \bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 \\
&\quad + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) \\
&= SS(E) + SS(T) + CP
\end{aligned}$$

- The cross-product term is:

$$\begin{aligned} CP &= 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot})(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot}) \\ &= 2\sum_{i=1}^{k}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_{i\cdot}) \\ &= 2\sum_{i=1}^{k}(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})(n_i\bar{X}_{i\cdot}-n_i\bar{X}_{i\cdot}) \\ &= 0 \end{aligned}$$

- Thus, we have:

$$SS(TO) = SS(T) + SS(E)$$

Sampling distribution of $SS(E)$

- The sample variance from the ith group, S_i^2, is an unbiased estimator of σ^2 and we know that $(n_i-1)S_i^2/\sigma^2 \sim \chi^2_{n_i-1}$
- The samples from each group are independent, so we can usefully combine them,

$$\sum_{i=1}^{k}\frac{(n_i-1)S_i^2}{\sigma^2} = \frac{SS(E)}{\sigma^2} \sim \chi^2_{n-k}$$

- Note that: $(n_1-1)+(n_2-1)+\cdots+(n_k-1) = n-k$
- This also gives us an unbiased pooled estimator of σ^2,

$$\hat{\sigma}^2 = \frac{SS(E)}{n-k}$$

- These results are true irrespective of whether H_0 is true or not

Null sampling distribution of $SS(TO)$

- If we assume H_0, we can derive simple expressions for the sampling distributions of the other sums of squares
- The combined data would be a sample of size n from $\mathrm{N}(\mu, \sigma^2)$. Hence $SS(TO)/(n-1)$ is an unbiased estimator of σ^2 and

$$\frac{SS(TO)}{\sigma^2} \sim \chi^2_{n-1}$$

Null sampling distribution of $SS(T)$

- Under H_0, we have $\bar{X}_{i\cdot} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n_i})$
- $\bar{X}_{1\cdot}, \bar{X}_{2\cdot}, \ldots, \bar{X}_{k\cdot}$ are independent
- (Can think of this as a sample of sample means, and then think about what its variance estimator is)
- It is possible to show that (proof not shown):

$$\sum_{i=1}^{k}\frac{n_i(\bar{X}_{i\cdot}-\bar{X}_{\cdot\cdot})^2}{\sigma^2} = \frac{SS(T)}{\sigma^2} \sim \chi^2_{k-1}$$

and that this is independent of $SS(E)$

Null sampling distributions

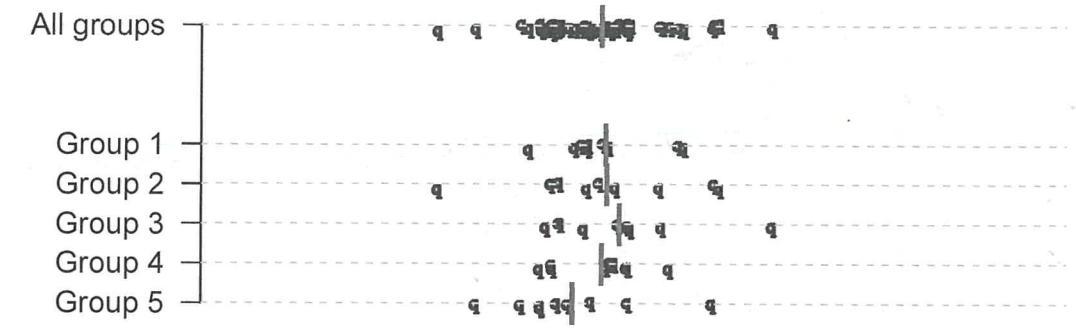
In summary, under H_0:

$$\frac{SS(TO)}{\sigma^2} = \frac{SS(E)}{\sigma^2} + \frac{SS(T)}{\sigma^2}$$

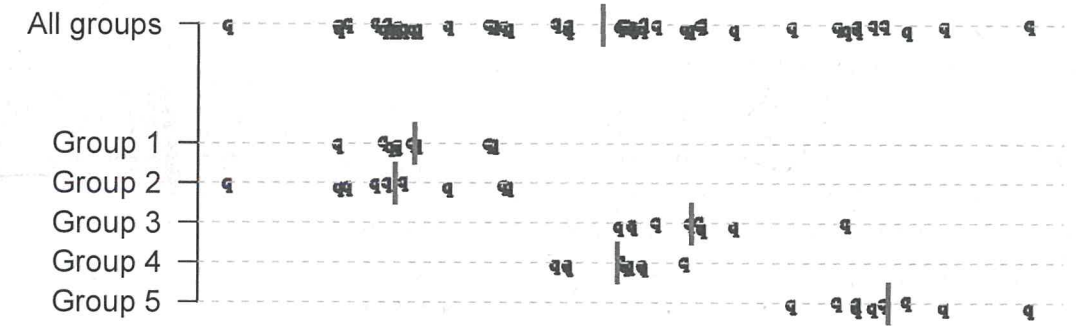
$$\frac{SS(TO)}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{SS(E)}{\sigma^2} \sim \chi^2_{n-k}, \quad \frac{SS(T)}{\sigma^2} \sim \chi^2_{k-1},$$

$SS(E)$ and $SS(T)$ are independent

H_0 is true



H_1 is true



$SS(T)$ under H_1

- What happens if H_1 is true?
- The population means differ, which will make $SS(T)$ larger
- Let's make this precise...

- Let $\bar{\mu} = n^{-1}\sum_{i=1}^{k} n_i\mu_i$, and then,

$$\begin{aligned}
\mathbb{E}[SS(T)] &= \mathbb{E}\left[\sum_{i=1}^{k} n_i(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2\right] = \mathbb{E}\left[\sum_{i=1}^{k} n_i\bar{X}_{i\cdot}^2 - n\bar{X}_{\cdot\cdot}^2\right] \\
&= \sum_{i=1}^{k} n_i\,\mathbb{E}(\bar{X}_{i\cdot}^2) - n\,\mathbb{E}(\bar{X}_{\cdot\cdot}^2) \\
&= \sum_{i=1}^{k} n_i\left[\mathrm{var}(\bar{X}_{i\cdot}) + \mathbb{E}(\bar{X}_{i\cdot})^2\right] - n\left[\mathrm{var}(\bar{X}_{\cdot\cdot}) + \mathbb{E}(\bar{X}_{\cdot\cdot})^2\right] \\
&= \sum_{i=1}^{k} n_i\left[\frac{\sigma^2}{n_i} + \mu_i^2\right] - n\left[\frac{\sigma^2}{n} + \bar{\mu}^2\right] \\
&= (k-1)\sigma^2 + \sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2
\end{aligned}$$

- Under H_0 the second term is zero and we have,

$$\frac{\mathbb{E}(SS(T))}{k-1} = \sigma^2$$

- Otherwise (under H_1), the second term is positive and gives,

$$\frac{\mathbb{E}(SS(T))}{k-1} > \sigma^2$$

- In contrast, we always have,

$$\frac{E(SS(E))}{n-k} = \sigma^2$$

F-test statistic

- This movitates using the following as our test statistic:

$$F = \frac{SS(T)/(k-1)}{SS(E)/(n-k)}$$

- Under H_0, we have $F \sim \mathrm{F}_{k-1,n-k}$, since it is the ratio of independent χ^2 random variables
- Under H_1, the numerator will tend to be larger
- Therefore, reject H_0 if $F > c$
- This is known as an *F-test*

ANOVA table

The test quantities are often summarised using an *ANOVA table*:

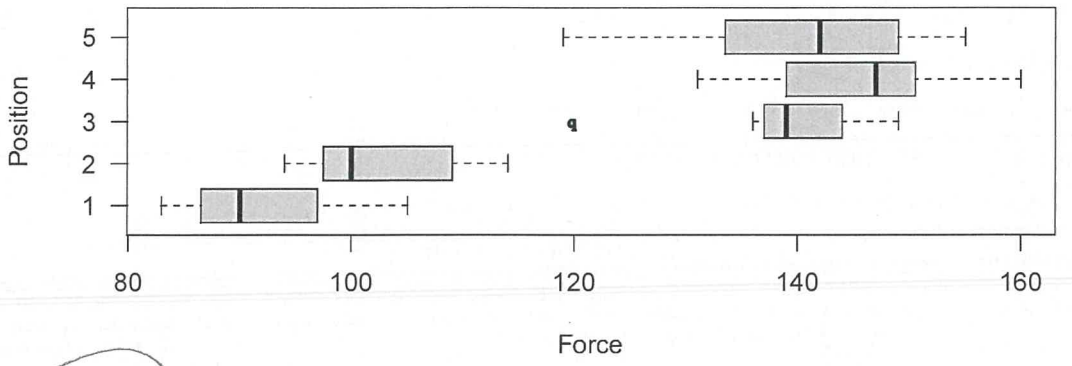
Source	df	SS	MS	F
Treatment	$k-1$	$SS(T)$	$MS(T) = \frac{SS(T)}{k-1}$	$\frac{MS(T)}{MS(E)}$
Error	$n-k$	$SS(E)$	$MS(E) = \frac{SS(E)}{n-k}$	
Total	$n-1$	$SS(TO)$		

Notes:

- MS = 'Mean square'
- $\hat{\sigma}^2 = MS(E)$ is an unbiased estimator

Example (one-way ANOVA)

Force required to pull out window studs in 5 positions on a car window.




```
> head(data1)
  Position Force
1        1    92
2        1    90
3        1    87
4        1   105
5        1    86
6        1    83

> table(data1$Position)

1 2 3 4 5
7 7 7 7 7

> model1 <- lm(Force ~ factor(Position), data = data1)
> anova(model1)
Analysis of Variance Table

Response: Force
                Df  Sum Sq Mean Sq F value    Pr(>F)
factor(Position)  4 16672.1  4168.0  44.202 3.664e-12 ***
Residuals        30  2828.9    94.3
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Notes:

- Need to use `factor()` to denote categorical variables
- R doesn't provide a 'Total' row, but we don't need it
- `Residuals` is the 'Error' row
- `Pr(>F)` is the p-value for the F-test

We conclude that the mean force required to pull out the window studs varies between the 5 positions on the car window (e.g. p-value < 0.01)

This was obvious from the boxplots: positions 1 & 2 are quite different from 3, 4 & 5

1.3 Two-way ANOVA

Two factors

- In one-way ANOVA, the observations were partitioned into k groups
- In other words, they were defined by a single categorical variable ('factor')
- What if we had two such variables?

- We can extend the procedure to give *two-way ANOVA*, or *two-factor ANOVA*
- For example, the fuel consumption of a car may depend on type of petrol and the brand of tyres

Two-way ANOVA: setting it up

- Factor 1 has a levels, Factor 2 has b levels
- Suppose we have exactly one observation per factor combination
- Observe X_{ij} with factor 1 at level i and factor 2 at level j
- Gives a total of $n = ab$ observations
- Assume $X_{ij} \sim \mathrm{N}(\mu_{ij}, \sigma^2)$, $i = 1, \ldots, a$, $j = 1, \ldots, b$, and that these are independent
- Consider the model:

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$
$$\text{with } \sum_{i=1}^{a} \alpha_i = 0, \ \sum_{j=1}^{b} \beta_j = 0$$

- μ is an overall effect, α_i is the effect of the ith row and β_j the effect of the jth column.
- For example, $a = 4$ and $b = 4$,

	1	2	3	4
1	$\mu+\alpha_1+\beta_1$	$\mu+\alpha_1+\beta_2$	$\mu+\alpha_1+\beta_3$	$\mu+\alpha_1+\beta_4$
2	$\mu+\alpha_2+\beta_1$	$\mu+\alpha_2+\beta_2$	$\mu+\alpha_2+\beta_3$	$\mu+\alpha_2+\beta_4$
3	$\mu+\alpha_3+\beta_1$	$\mu+\alpha_3+\beta_2$	$\mu+\alpha_3+\beta_3$	$\mu+\alpha_3+\beta_4$
4	$\mu+\alpha_4+\beta_1$	$\mu+\alpha_4+\beta_2$	$\mu+\alpha_4+\beta_3$	$\mu+\alpha_4+\beta_4$

- We are usually interested in $H_{0A}: \alpha_1 = \alpha_2 = \cdots = \alpha_a = 0$ or $H_{0B}: \beta_1 = \beta_2 = \cdots = \beta_b = 0$
- Let

$$\bar{X}_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} X_{ij}, \quad \bar{X}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b} X_{ij}, \quad \bar{X}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a} X_{ij}$$

- Arguing as before,

$$\begin{aligned} SS(TO) &= \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{\cdot\cdot})^2 \\ &= \sum_{i=1}^{a}\sum_{j=1}^{b}\big[(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot}) + (\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot}) \\ &\qquad + (X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})\big]^2 \\ &= b\sum_{i=1}^{a}(\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2 + a\sum_{j=1}^{b}(\bar{X}_{\cdot j} - \bar{X}_{\cdot\cdot})^2 \\ &\qquad + \sum_{i=1}^{a}\sum_{j=1}^{b}(X_{ij} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j} + \bar{X}_{\cdot\cdot})^2 \\ &= SS(A) + SS(B) + SS(E) \end{aligned}$$

- If both $\alpha_1 = \cdots = \alpha_a = 0$ and $\beta_1 = \cdots = \beta_b = 0$, then we have $SS(A)/\sigma^2 \sim \chi^2_{a-1}$, $SS(B)/\sigma^2 \sim \chi^2_{b-1}$ and $SS(E)/\sigma^2 \sim \chi^2_{(a-1)(b-1)}$ and these variables are independent (proof not shown)
- Reject $H_{0A}: \alpha_1 = \cdots = \alpha_a = 0$ at significance level α if:

$$F_A = \frac{SS(A)/(a-1)}{SS(E)/((a-1)(b-1))} > c$$

where c is the $1-\alpha$ quantile of $\mathrm{F}_{a-1,(a-1)(b-1)}$

- Reject H_{0B}: $\beta_1 = \cdots = \beta_b = 0$ at significance level α if:

$$F_B = \frac{SS(B)/(b-1)}{SS(E)/((a-1)(b-1))} > c$$

where c is the $1-\alpha$ quantile of $F_{b-1,(a-1)(b-1)}$

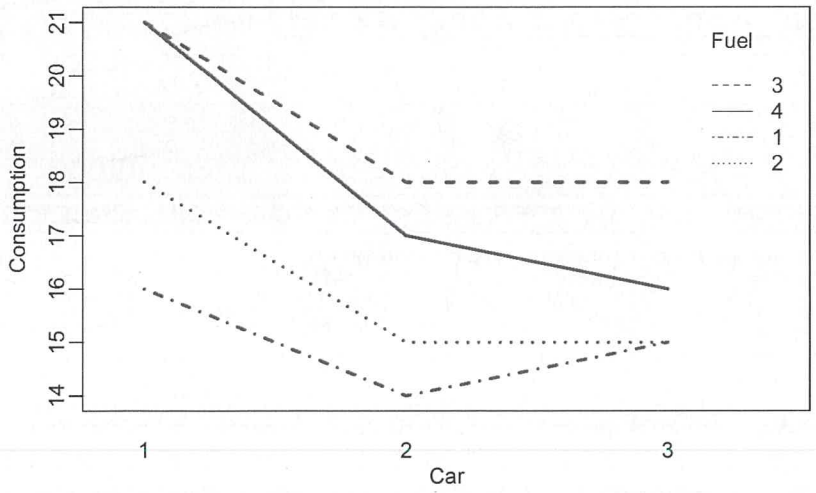
ANOVA table

Source	df	SS	MS	F
Factor A	$a-1$	$SS(A)$	$MS(A) = \frac{SS(A)}{a-1}$	$\frac{MS(A)}{MS(E)}$
Factor B	$b-1$	$SS(B)$	$MS(B) = \frac{SS(B)}{b-1}$	$\frac{MS(B)}{MS(E)}$
Error	$(a-1)(b-1)$	$SS(E)$	$MS(E) = \frac{SS(E)}{(a-1)(b-1)}$	
Total	$ab-1$	$SS(TO)$		

Example (two-way ANOVA)

Data on fuel consumption for three types of car (A) and four types of fuel (B).

```
> head(data2)
  Car Fuel Consumption
1   1    1          16
2   1    2          18
3   1    3          21
4   1    4          21
5   2    1          14
6   2    2          15
```




```
> model2 <- lm(Consumption ~ factor(Car) + factor(Fuel),
+              data = data2)
> anova(model2)
Analysis of Variance Table

Response: Consumption
             Df Sum Sq Mean Sq F value   Pr(>F)
factor(Car)   2     24 12.0000      18 0.002915 **
factor(Fuel)  3     30 10.0000      15 0.003401 **
Residuals     6      4  0.6667
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

From this we conclude there is a clear difference in fuel consumption between cars (we reject H_{0A}: $\alpha_1 = \alpha_2 = \alpha_3$) and also between fuels (we reject H_{0B}: $\beta_1 = \beta_2 = \beta_3 = \beta_4$).

1.4 Two-way ANOVA with interaction

Interaction terms

- In the previous example we assumed an additive model:

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

- This assumes, for example, that the relative effect of petrol 1 is the same for all cars.
- If it is not true, then there is a *statistical interaction* (or simply an *interaction*) between the factors
- A more general model, which includes interactions, is:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

where γ_{ij} is the *interaction term* associated with combination (i, j).

- In addition to our previous assumptions, we also impose:

$$\sum_{i=1}^{a} \gamma_{ij} = 0, \quad \text{and} \quad \sum_{j=1}^{b} \gamma_{ij} = 0$$

- The terms α_i and β_j are called *main effects*
- When written out as a table they are also often referred to as the *row effects* and *column effects* respectively
- Writing this out as a table:

	1	2	3	4
1	$\mu+\alpha_1+\beta_1+\gamma_{11}$	$\mu+\alpha_1+\beta_2+\gamma_{12}$	$\mu+\alpha_1+\beta_3+\gamma_{13}$	$\mu+\alpha_1+\beta_4+\gamma_{14}$
2	$\mu+\alpha_2+\beta_1+\gamma_{21}$	$\mu+\alpha_2+\beta_2+\gamma_{22}$	$\mu+\alpha_2+\beta_3+\gamma_{23}$	$\mu+\alpha_2+\beta_4+\gamma_{24}$
3	$\mu+\alpha_3+\beta_1+\gamma_{31}$	$\mu+\alpha_3+\beta_2+\gamma_{32}$	$\mu+\alpha_3+\beta_3+\gamma_{33}$	$\mu+\alpha_3+\beta_4+\gamma_{34}$
4	$\mu+\alpha_4+\beta_1+\gamma_{41}$	$\mu+\alpha_4+\beta_2+\gamma_{42}$	$\mu+\alpha_4+\beta_3+\gamma_{43}$	$\mu+\alpha_4+\beta_4+\gamma_{44}$

- We are now interested in testing whether:
  - the row effects are zero
  - the column effects are zero
  - the interactions are zero (do this first!)
- To make inferences about the interactions we need more than one observation per cell
- Let X_{ijk}, $i = 1, \ldots, a$, $j = 1, \ldots, b$, $k = 1, \ldots, c$ be the kth observation for combination (i, j)
- Let

$$\bar{X}_{ij\cdot} = \frac{1}{c} \sum_{k=1}^{c} X_{ijk}$$

$$\bar{X}_{i\cdot\cdot} = \frac{1}{bc} \sum_{j=1}^{b} \sum_{k=1}^{c} X_{ijk}$$

$$\bar{X}_{\cdot j\cdot} = \frac{1}{ac} \sum_{i=1}^{a} \sum_{k=1}^{c} X_{ijk}$$

$$\bar{X}_{\cdots} = \frac{1}{abc} \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} X_{ijk}$$

- and as before

$$\begin{aligned} SS(TO) &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk}-\bar{X}_{...})^2 \\ &= bc\sum_{i=1}^{a}(\bar{X}_{i..}-\bar{X}_{...})^2 + ac\sum_{j=1}^{b}(\bar{X}_{.j.}-\bar{X}_{...})^2 \\ &\quad + c\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{X}_{ij.}-\bar{X}_{i..}-\bar{X}_{.j.}+\bar{X}_{...})^2 \\ &\quad + \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk}-\bar{X}_{ij.})^2 \\ &= SS(A)+SS(B)+SS(AB)+SS(E) \end{aligned}$$

Test statistics

- Familiar arguments show that to test

$$H_{0AB}: \gamma_{ij}=0, \quad i=1,\ldots,a, \quad j=1,\ldots,b$$

we may use the statistic

$$F=\frac{SS(AB)/[(a-1)(b-1)]}{SS(E)/[ab(c-1)]}$$

which has a F distribution with $(a-1)(b-1)$ and $ab(c-1)$ degrees of freedom.

- To test

$$H_{0A}: \alpha_i=0, \quad i=1,\ldots,a$$

we may use the statistic

$$F=\frac{SS(A)/[(a-1)]}{SS(E)/[ab(c-1)]}$$

which has a F distribution with $(a-1)$ and $ab(c-1)$ degrees of freedom.

- To test

$$H_{0B}: \beta_j=0, \quad j=1,\ldots,b$$

we may use the statistic

$$F=\frac{SS(B)/[(b-1)]}{SS(E)/[ab(c-1)]}$$

which has a F distribution with $(b-1)$ and $ab(c-1)$ degrees of freedom.

ANOVA table

Source	df	SS	MS	F
Factor A	$a-1$	$SS(A)$	$MS(A)=\frac{SS(A)}{a-1}$	$\frac{MS(A)}{MS(E)}$
Factor B	$b-1$	$SS(B)$	$MS(B)=\frac{SS(B)}{b-1}$	$\frac{MS(B)}{MS(E)}$
Factor AB	$(a-1)(b-1)$	$SS(AB)$	$MS(AB)=\frac{SS(AB)}{(a-1)(b-1)}$	$\frac{MS(AB)}{MS(E)}$
Error	$ab(c-1)$	$SS(E)$	$MS(E)=\frac{SS(E)}{ab(c-1)}$	
Total	$abc-1$	$SS(TO)$		

Example (two-way ANOVA with interaction)

- Six groups of 18 people
- Each person takes an arithmetic test: the task is to add three numbers together
- The numbers are presented either in a down array or an across array; this defines 2 levels of factor A
- The numbers have either one, two or three digits; this defines 3 levels of factor B

- The response variable, X, is the average number of problems completed correctly over two 90-second sessions
- Example of adding **one-digit** numbers in an **across** array:

$$2 + 5 + 1 = ?$$

- Example of adding **two-digit** numbers in an **down** array:

$$\begin{array}{r} 13 \\ 87 \\ +\ 51 \\ \hline ? \end{array}$$

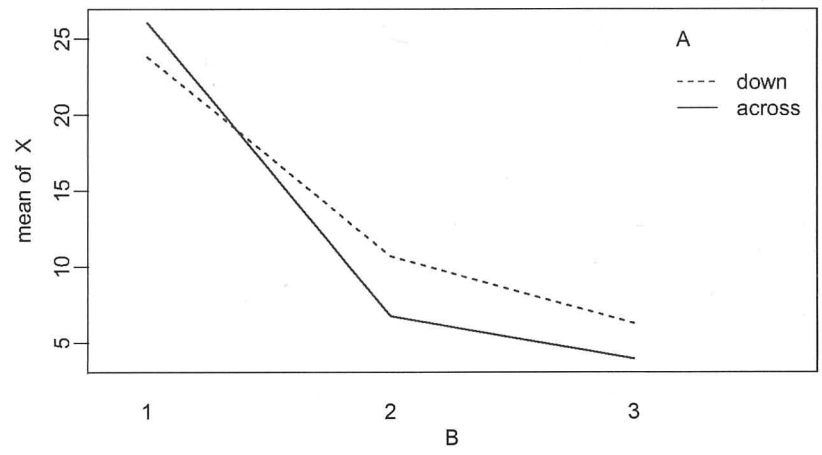
```
> head(data3)
     A B    X
1 down 1 19.5
2 down 1 18.5
3 down 1 32.0
4 down 1 21.5
5 down 1 28.5
6 down 1 33.0

> table(data3[, 1:2])
        B
A          1  2  3
  down    18 18 18
  across 18 18 18
> model3 <- lm(X ~ factor(A) * factor(B), data = data3)
> anova(model3)
Analysis of Variance Table

Response: X
                    Df Sum Sq Mean Sq  F value  Pr(>F)
factor(A)            1   48.7    48.7   2.8849 0.09246 .
factor(B)            2 8022.7  4011.4 237.7776 < 2e-16 ***
factor(A):factor(B)  2  185.9    93.0   5.5103 0.00534 **
Residuals          102 1720.8    16.9
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Note the use of '*' in the model formula.

The interaction is significant at a 5% level (or even at 1%).

Interaction plot

```
with(data3, interaction.plot(B, A, X, col = "blue"))
```



Beyond the F-test

- We have rejected the null... now what?
- This is often only the beginning of a statistical analysis of this type of data
- Will be interested in more detailed inferences, e.g. CIs/tests about individual parameters
- You know enough to be able to work some of this out...
- ...and later subjects will go into this in more detail (e.g. MAST30025)

2 Hypothesis testing in regression

Recap of simple linear regression

- Y a response variable, e.g. student's grade in first-year calculus
- x a predictor variable, e.g. student's high school mathematics mark
- Data: pairs $(x_1, y_1), \ldots, (x_n, y_n)$
- Linear regression model:

$$Y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ is a random error

- **Note:** α here plays the same role as α_0 from Module 5. We have dropped the '0' subscript for convenience, and also to avoid confusion with its use to denote null hypotheses.
- The MLE (and OLS) estimators are:

$$\hat{\alpha} = \bar{Y}, \quad \hat{\beta} = \frac{\sum_{i=1}^n Y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- and

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n [Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})]^2$$

- We also derived:

$$\hat{\alpha} \sim \mathrm{N}\left(\alpha, \frac{\sigma^2}{n}\right)$$

$$\hat{\beta} \sim \mathrm{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

- and

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n \left[Y_i - \hat{\alpha} - \hat{\beta}(x_i - \bar{x})\right]^2}{\sigma^2} \sim \chi^2_{n-2}$$

- From these we obtain,

$$T_\alpha = \frac{\hat{\alpha} - \alpha}{\hat{\sigma}/\sqrt{n}} \sim t_{n-2}$$

$$T_\beta = \frac{\hat{\beta} - \beta}{\hat{\sigma}/\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

- We used these previously to construct confidence intervals
- We can also use them to construct hypothesis tests
- For example, to test $H_0 \colon \beta = \beta_0$ versus $H_1 \colon \beta \neq \beta_0$ (or $\beta > \beta_0$ or $\beta < \beta_0$), we use T_β as the test statistic

Example: testing the slope parameter (β)

- Data: 10 pairs of scores on a preliminary test and a final exam
- Estimates: $\hat{\alpha} = 81.3$, $\hat{\beta} = 0.742$, $\hat{\sigma}^2 = 27.21$
- Test $H_0: \beta = 0$ versus $H_1: \beta \neq 0$ with a 1% significance level
- Reject H_0 if:

$$|T_\beta| \geqslant 3.36 \quad (0.995 \text{ quantile of } t_8)$$

- For the observed data,

$$t_\beta = \frac{0.742 - 0}{\sqrt{27.21/756.1}} = 3.91$$

so we reject H_0, concluding there is sufficient evidence that the slope differs from zero.

Note regarding the intercept parameter (α)

- Software packages (such as R) will typically fit the model:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

- This is equivalent to

$$Y_i = \alpha^* + \beta(x_i - \bar{x}) + \epsilon_i$$

where $\alpha = \alpha^* - \beta\bar{x}$

- The formulation $Y_i = \alpha^* + \beta(x - \bar{x}) + \epsilon$ is easier to examine theoretically.
- We saw that

$$\hat{\alpha}^* = \bar{Y}, \quad \text{and} \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

- $\hat{\alpha}$ or $\hat{\alpha}^*$ are rarely of direct interest

Using R

Use R to fit the regression model for the slope example:

```
> m1 <- lm(final_exam ~ prelim_test)
> summary(m1)

Call:
lm(formula = final_exam ~ prelim_test)

Residuals:
   Min     1Q Median     3Q    Max
-6.883 -3.264 -0.530  3.438  8.470

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.6147    13.0622   2.344  0.04714 *
prelim_test   0.7421     0.1897   3.912  0.00447 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.217 on 8 degrees of freedom
Multiple R-Squared: 0.6567,  Adjusted R-squared: 0.6137
F-statistic:  15.3 on 1 and 8 DF,  p-value: 0.004471
```

The t-value and the p-value are for testing $H_0: \alpha = 0$ and $H_0: \beta = 0$ respectively.

Interpreting the R output

- Usually most interested in testing H_0: $\beta = 0$ versus H_1: $\beta \neq 0$
- If we reject H_0 then we conclude there is sufficient evidence of (at least) a linear relationship between the mean response and x
- In the example,

$$t = \frac{0.7421}{0.1897} = 3.912$$

- This test statistic has a t-distribution with $10-2=8$ degrees of freedom, and the associated p-value is $0.00447 < 0.05$ so at the 5% level of significance we reject H_0
- It is also possible to represent this test using an ANOVA table

2.1 Analysis of variance approach

Deriving the variance decomposition formula

- Independent pairs $(x_1, Y_1), \ldots, (x_n, Y_n)$
- Parameter estimates,

$$\hat{\alpha} = \bar{Y}, \quad \hat{\beta} = \frac{\sum_{i=1}^n Y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Fitted value (estimated mean),

$$\hat{Y}_i = \bar{Y} + \hat{\beta}(x_i - \bar{x})$$

- Do the 'add and subtract' trick again:

$$\begin{aligned}\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ &\quad + 2\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})\end{aligned}$$

- Deal with the cross-product term,

$$\begin{aligned}\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) &= \sum_{i=1}^n \left[Y_i - \bar{Y} - \hat{\beta}(x_i - \bar{x})\right] \hat{\beta}(x_i - \bar{x}) \\ &= \hat{\beta} \sum_{i=1}^n \left[Y_i - \bar{Y} - \hat{\beta}(x_i - \bar{x})\right] (x_i - \bar{x}) \\ &= \hat{\beta} \left[\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^n (x_i - \bar{x})^2\right] \\ &= \hat{\beta} \left[\sum_{i=1}^n Y_i(x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^n (x_i - \bar{x})^2\right] \\ &= 0\end{aligned}$$

- That gives us,

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

- We can write this as follows,

$$SS(TO) = SS(E) + SS(R)$$

where $SS(R)$ is the *regression SS* or *model SS*

- The regression SS quantifies the variation **due to** the straight line

- The error SS quantifies the variation **around** the straight line
- To complete the specification,

$$MS(E) = \frac{SS(E)}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \hat{\sigma}^2$$

$$MS(R) = \frac{SS(R)}{1} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

- Then we have the test statistic,

$$F = \frac{MS(R)}{MS(E)} \sim F_{1,n-2}$$

ANOVA table

Source	df	SS	MS	F
Model	1	$SS(R)$	$MS(R) = \frac{SS(R)}{1}$	$\frac{MS(R)}{MS(E)}$
Error	$n-2$	$SS(E)$	$MS(E) = \frac{SS(E)}{n-2}$	
Total	$n-1$	$SS(TO)$		

Using R

```
> anova(m1)
Analysis of Variance Table

Response: final_exam
            Df Sum Sq Mean Sq F value   Pr(>F)
prelim_test  1 416.39  416.39  15.301 0.004471 **
Residuals    8 217.71   27.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notes:

- The F-statistic tests the 'significance of the regression'
- That is, $H_0: \beta = 0$ versus $H_1: \beta \neq 0$

3 Likelihood ratio tests

Is there a 'best' test?

- We have examined a variety of commonly used tests
- We used test statistics that:
  - Seemed useful
  - We were familiar with
- Did we use the 'best' one?
- Is there a general procedure for finding a good/best test statistic?
- We will introduce a general procedure now, and discuss why it is optimal later in the semester

Likelihood ratio test

- The *likelihood ratio test (LRT)* is a general procedure that can find the best test for a given problem

- Suppose we have H_0 and H_1 and both are composite and of the form:

$$H_0 : \theta \in A_0 \quad \text{versus} \quad H_1 : \theta \in A_1$$

where A_0 and A_1 are sets of possible parameter values consistent with each of the hypotheses.

- Note: we have mostly dealt with A_0 that has only one element (simple null hypothesis)
- The *likelihood ratio* is:

$$\lambda = \frac{L_0}{L_1} = \frac{\max_{\theta \in A_0} L(\theta)}{\max_{\theta \in A_1} L(\theta)}$$

- L is the likelihood function
- Clearly $\lambda \geqslant 0$
- Large $\lambda \Rightarrow$ more support for H_0 over H_1
- λ near zero $\Rightarrow$ more support for H_1 over H_0
- Therefore, we want a critical region of the form, $\lambda \leqslant k$
- Choose k to give the desired significance level

Example 1 (likelihood ratio test)

- $X_i \sim \mathrm{N}(\mu, \sigma^2 = 5)$, i.e. σ is known
- $H_0 : \mu = 162$ versus $H_1 : \mu \neq 162$
- When H_0 is true, $\mu = 162$ so $L_0 = L(162)$
- When H_1 is true, need to maximise the likelihood, $L_1 = L(\hat{\theta}) = L(\bar{x})$
- The likelihood ratio is,

$$\begin{aligned} \lambda = \frac{L_0}{L_1} &= \frac{L(162)}{L(\bar{x})} = \frac{(10\pi)^{-n/2} \exp\left[-\frac{1}{10}\sum_{i=1}^{n}(x_i - 162)^2\right]}{(10\pi)^{-n/2} \exp\left[-\frac{1}{10}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]} \\ &= \exp\left[-\frac{n}{10}(\bar{x} - 162)^2\right] \end{aligned}$$

- $\lambda \leqslant k$ same as

$$\frac{|\bar{x} - 162|}{\sigma/\sqrt{n}} \geqslant c$$

- A critical region for a size α test is

$$\frac{|\bar{x} - 162|}{\sigma/\sqrt{n}} \geqslant \Phi^{-1}(1 - \alpha/2)$$

- Note: this required knowledge of the distribution of $\bar{X}$!

Example 2 (likelihood ratio test)

- $X_i \sim \mathrm{N}(\mu, \sigma^2)$, i.e. σ is unknown
- $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$
- Under H_0 we have $\mu = \mu_0$, and under H_1 we need to use its MLE
- Under either hypothesis, σ^2 is unspecified, so in both cases we need its MLE (conditional on the specified value of μ).
- So, under H_0 we use:

$$\hat{\mu} = \mu_0, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

- And under H_1 we use:

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Some simplification yields

$$\lambda = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2}\right]^{n/2}$$

- and

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2$$

- Substitute and rearrange to get

$$\lambda = \left[\frac{1}{1 + \frac{n(\bar{x}-\mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right]^{n/2}$$

- Therefore, we have $\lambda \leqslant k$ when,

$$\frac{n(\bar{x} - \mu_0)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \geqslant c$$

- When H_0 is true, $\sqrt{n}(\bar{X} - \mu_0)/\sigma \sim \mathrm{N}(0,1)$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$, and is independent of $\bar{X}$.
- Therefore,

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)/\sigma}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2}} = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

- So we reject H_0 when $|T|$ is too large, with the following critical region for a test with significance level α,

$$|T| \geqslant d, \quad \text{where } d \text{ is the } 1 - \frac{\alpha}{2} \text{ quantile of } t_{n-1}$$

Remarks

- Usually easy to find the **form** of the test
- What is harder is to find the corresponding sampling distribution
- Manipulating λ until we have something whose distribution we know can be tricky!
- Many of the standard tests arise from the likelihood ratio

Asymptotic distribution & optimality

- The likelihood ratio itself is a statistic and therefore has a sampling distribution.
- For large sample sizes, this approaches a known distribution
- Also, the LRT gives the optimal test
- We will cover this theory later in the semester

Order statistics, quantiles & resampling

(Module 9)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2019

Contents

Aims of this module

- Go back to **order statistics** and **sample quantiles**
- More detailed definitions
- Derive **sampling distributions** and construct **confidence intervals**
- See examples of CIs that are **not** of the form $\hat{\theta} \pm \text{se}(\hat{\theta})$
- Learn some more distribution-free methods
- See how to use computation to avoid mathematical derivations

Unifying theme

- Use the data 'directly' rather than via assumed distributions
- Use the **sample cdf** and related summaries (such as order statistics)

1 Order statistics

1.1 Introduction

Definition (recap)

- Sample: $X_1, \ldots, X_n$
- Arrange them in increasing order:

$$
\begin{aligned}
X_{(1)} &= \text{Smallest of the } X_i \\
X_{(2)} &= \text{2nd smallest of the } X_i \\
&\vdots \\
X_{(n)} &= \text{Largest of the } X_i
\end{aligned}
$$

- These are called the *order statistics*

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$$

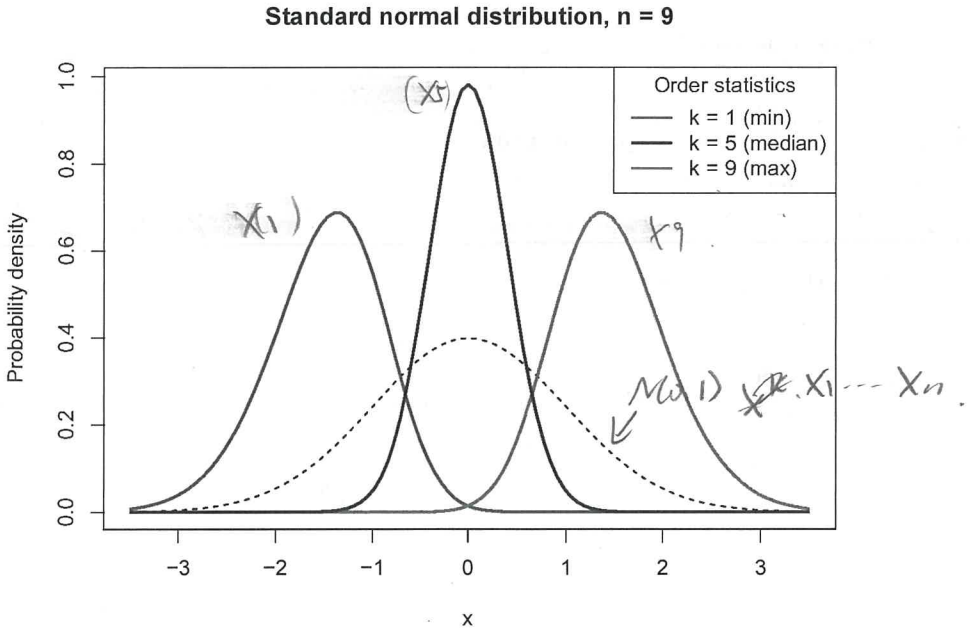
- $X_{(k)}$ is called the *kth order statistic* of the sample
- $X_{(1)}$ is the *minimum* or *sample minimum*
- $X_{(n)}$ is the *maximum* or *sample maximum*

aim ⇒ want to know distribution of $X_{(k)}$ ↳ cdf / pdf

Motivating example

- Take iid samples $X \sim \mathrm{N}(0,1)$ of size $n = 9$
- What can we say about the order statistics, $X_{(k)}$?
- Simulated values:

```
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,] -0.76 -1.94 -1.32 -0.85 -1.96  <-- Minimum
[2,] -0.32 -0.17 -0.53 -0.30 -0.98
[3,] -0.23  0.06 -0.44  0.14 -0.83
[4,]  0.05  0.18 -0.10  0.25 -0.63
[5,]  0.08  0.76  0.17  0.35 -0.47  <-- Median
[6,]  0.18  0.96  0.26  0.68  0.05
[7,]  0.27  1.07  0.60  0.69  0.34
[8,]  0.73  1.42  0.66  1.13  1.26
[9,]  0.91  1.77  1.93  1.98  1.26  <-- Maximum
```



1.2 Sampling distribution

Example (triangular distribution)

- Random sample: $X_1, \ldots, X_5$ with pdf $f(x) = 2x$, $0 < x < 1$
- Calculate $\Pr(X_{(4)} \leqslant 0.5)$
- Occurs if at least four of the X_i are less than 0.5,

$$\begin{aligned}\Pr(X_{(4)} \leqslant 0.5) &= \Pr(\textbf{at least 4 } X_i\text{'s less than } 0.5) \\ &= \Pr(\textbf{exactly 4 } X_i\text{'s less than } 0.5) \\ &\quad + \Pr(\textbf{exactly 5 } X_i\text{'s less than } 0.5)\end{aligned}$$

- This is a binomial with 5 trials and probability of success given by

$$p = \Pr(X_i \leqslant 0.5) = \int_0^{0.5} 2x\,dx = \left[x^2\right]_0^{0.5} = 0.5^2 = 0.25$$

- So we have,

$$\Pr(X_{(4)} \leqslant 0.5) = \binom{5}{4} 0.25^4\, 0.75 + 0.25^5 = 0.0156$$

- More generally we have,

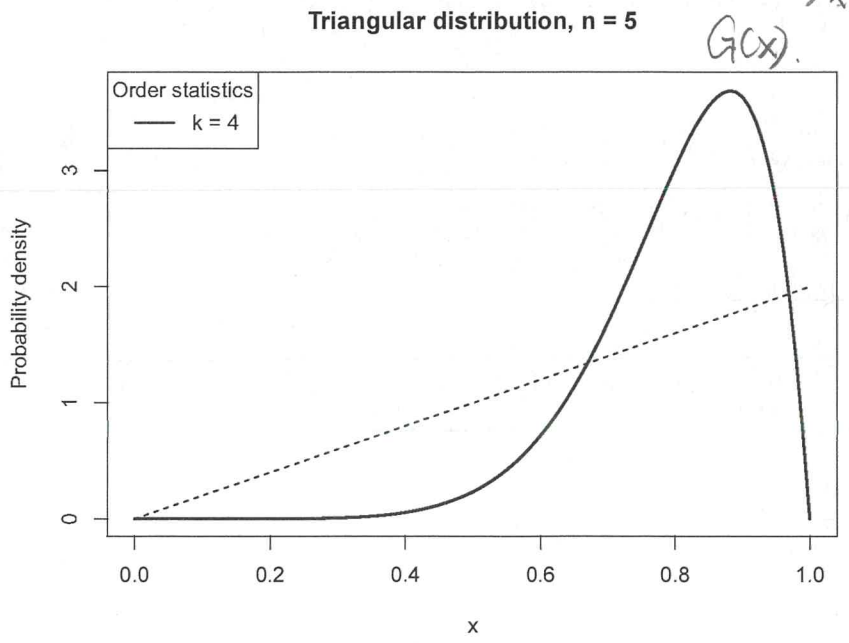
$$F(x) = \Pr(X_i \leqslant x) = \int_0^x 2t\,dt = \left[t^2\right]_0^x = x^2$$

$$G(x) = \Pr(X_{(4)} \leqslant x) = \binom{5}{4}(x^2)^4(1-x^2) + (x^2)^5$$

- Taking derivatives gives the pdf,

$$\begin{aligned} g(x) = G'(x) &= \binom{5}{4} 4(x^2)^3(1-x^2)(2x) \\ &= 4\binom{5}{4} F(x)^3\,(1-F(x))\,f(x)\end{aligned}$$

since we know that $F(x) = x^2$.



Distribution of $X_{(k)}$

- Sample from a continuous distribution with cdf $F(x)$ and pdf $f(x) = F'(x)$.
- The cdf of $X_{(k)}$ is,

$$\begin{aligned} G_k(x) &= \Pr(X_{(k)} \leqslant x) \\ &= \sum_{i=k}^{n} \binom{n}{i} F(x)^i \left(1 - F(x)\right)^{n-i} \end{aligned}$$

- Thus the pdf of $X_{(k)}$ is,

$$\begin{aligned} g_k(x) = G_k'(x) &= \sum_{i=k}^{n} i\binom{n}{i} F(x)^{i-1} \left(1 - F(x)\right)^{n-i} f(x) \\ &\quad + \sum_{i=k}^{n-1} (n-i)\binom{n}{i} F(x)^i \left(1 - F(x)\right)^{n-i-1} \left(-f(x)\right) \\ &= k\binom{n}{k} F(x)^{k-1} \left(1 - F(x)\right)^{n-k} f(x) \\ &\quad + \sum_{i=k+1}^{n} i\binom{n}{i} F(x)^{i-1} \left(1 - F(x)\right)^{n-i} f(x) \\ &\quad - \sum_{i=k}^{n-1} (n-i)\binom{n}{i} F(x)^i \left(1 - F(x)\right)^{n-i-1} f(x) \end{aligned}$$

- But

$$i\binom{n}{i} = \frac{n!}{(i-1)!(n-i)!} = n\binom{n-1}{i-1}$$

and similarly

$$(n-i)\binom{n}{i} = \frac{n!}{i!(n-i-1)!} = n\binom{n-1}{i}$$

which allows some cancelling of terms.

- For example, the first term of the first summation is,

$$\begin{aligned} &(k+1)\binom{n}{k+1} F(x)^k \left(1 - F(x)\right)^{n-k-1} f(x) \\ &= n\binom{n-1}{k} F(x)^k \left(1 - F(x)\right)^{n-k-1} f(x) \end{aligned}$$

- The first term of the second summation is,

$$\begin{aligned} &(n-k)\binom{n}{k} F(x)^k \left(1 - F(x)\right)^{n-k-1} f(x) \\ &= n\binom{n-1}{k} F(x)^k \left(1 - F(x)\right)^{n-k-1} f(x) \end{aligned}$$

- These cancel, and similarly the other terms do as well.
- Hence, the pdf simplifies to,

$$g_k(x) = k\binom{n}{k} F(x)^{k-1} \left(1 - F(x)\right)^{n-k} f(x)$$

- Special cases: minimum and maximum,

$$g_1(x) = n\left(1 - F(x)\right)^{n-1} f(x)$$
$$g_n(x) = n\,F(x)^{n-1} f(x)$$

- Also:

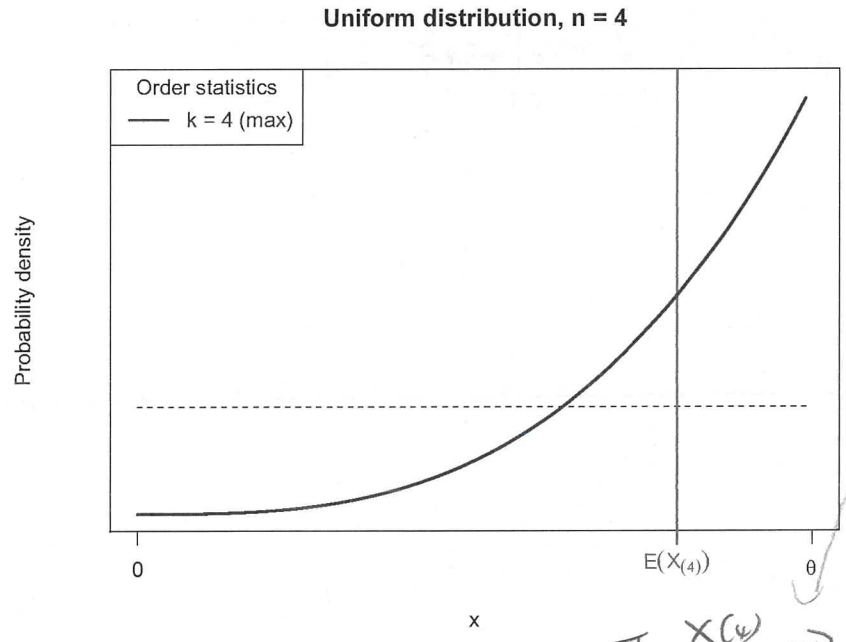
$$\Pr(X_{(1)} > x) = \left(1 - F(x)\right)^n$$
$$\Pr(X_{(n)} \leqslant x) = F(x)^n$$

Alternative derivation of the pdf of $X_{(k)}$

- Heuristically,
$$\Pr(X_{(k)} \approx x) = \Pr(x - \tfrac{1}{2}dy < X_{(k)} \leqslant x + \tfrac{1}{2}dy) \approx g_k(x)\,dy$$
- Need to observe X_i such that:
  - $k-1$ are in $(-\infty,\ x - \frac{1}{2}dy]$
  - One is in $(x - \frac{1}{2}dy,\ x + \frac{1}{2}dy]$
  - $n-k$ are in $(x + \frac{1}{2}dy,\ \infty)$
- Trinomial distribution (3 outcomes), event probabilities:
$$\begin{aligned} \Pr(X_i \leqslant x - \frac{1}{2}dy) &\approx F(x) \\ \Pr(x - \frac{1}{2}dy < X_i \leqslant x + \frac{1}{2}dy) &\approx f(x)\,dy \\ \Pr(X_i > x + \frac{1}{2}dy) &\approx 1 - F(x) \end{aligned}$$
- Putting these together,
$$g_k(x)\,dy \approx \frac{n!}{(k-1)!\,1!\,(n-k)!} F(x)^{k-1} \left(1 - F(x)\right)^{n-k} f(x)\,dy$$
- Dividing both sides by dy gives the pdf of $X_{(k)}$

Example (boundary estimate)

- $X_1, \ldots, X_4 \sim \text{Unif}(0, \theta)$
- Likelihood is
$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^4 & 0 \leqslant x_i \leqslant \theta, \quad i = 1, \ldots, 4 \\ 0 & \text{otherwise (i.e. if } \theta < x_i \text{ for some } i) \end{cases}$$
- Maximised when θ is as small as possible, so $\hat{\theta} = \max(X_i) = X_{(4)}$
- Now,
$$g_4(x) = 4\left(\frac{x}{\theta}\right)^3 \left(\frac{1}{\theta}\right) = \frac{4x^3}{\theta^4}, \quad 0 \leqslant x \leqslant \theta$$
- Then,
$$\mathbb{E}(X_{(4)}) = \int_0^\theta x \frac{4x^3}{\theta^4}\,dx = \left[\frac{4x^5}{5\theta^4}\right]_0^\theta = \frac{4}{5}\theta \neq \theta$$
- So the MLE $X_{(4)}$ is biased
- (But $\frac{5}{4}X_{(4)}$ is unbiased)



- Deriving a one-sided CI for θ based on $X_{(4)}$:
  1. For a given $0 < c < 1$, show that,
  $$1 - c^4 = \Pr(c\theta < X_{(4)} < \theta) = \Pr(X_{(4)} < \theta < X_{(4)}/c)$$
  2. Thus, a $100 \cdot (1 - c^4)\%$ confidence interval for θ is $(x_{(4)}, x_{(4)}/c)$
  3. Letting $c = \sqrt[4]{0.05} = 0.47$, we have a 95% confidence interval from $x_{(4)}$ to $2.11x_{(4)}$

2 Quantiles

2.1 Definitions

Population quantiles

- Informally, a **quantile** is a number that divides the range of a random variable based on the probabilities on either side.
- The *p-quantile*, π_p, of a continuous probability distribution with cdf F has the property:
$$p = F(\pi_p) = \Pr(X \leqslant \pi_p)$$
So, we can define it by the inverse cdf:
$$\pi_p = F^{-1}(p)$$
- More general definition (also works for discrete variables): the p-quantile is the smallest value π_p such that $p \leqslant F(\pi_p)$
- The most commonly used quantile is the *median*, $\pi_{0.5}$, often referred to simply as m
- Also the first and third *quartiles*, $\pi_{0.25}$ and $\pi_{0.75}$

Sample quantiles

- Want a statistic which estimates π_p
- There are many ways to do this
- R implements 9 different definitions!

- See `help(quantile)`
- Previously mentioned two of these...

'Type 6' quantiles

- Definition:

$$\hat{\pi}_p = x_{(k)}, \quad \text{where } p = \frac{k}{n+1}$$

- Linear interpolation otherwise
- Motivated by the following relationship (see later):

$$\mathbb{E}(F(X_{(k)})) = \frac{k}{n+1}$$

- We used this previously for QQ plots

'Type 7' quantiles

- Definition:

$$\hat{\pi}_p = x_{(k)}, \quad \text{where } p = \frac{k-1}{n-1}$$

- Linear interpolation otherwise
- Motivated by the following relationship (see later):

$$\text{mode}(F(X_{(k)})) = \frac{k-1}{n-1}$$

- This is the default in R (`quantile` function)

'Type 1' quantiles

- Can also apply the general quantile definition to the sample cdf:

$$\hat{\pi}_p = x_{(\lceil np \rceil)}$$

- The ceiling function, $\lceil b \rceil$, is the smallest integer not less than b
- In other words,

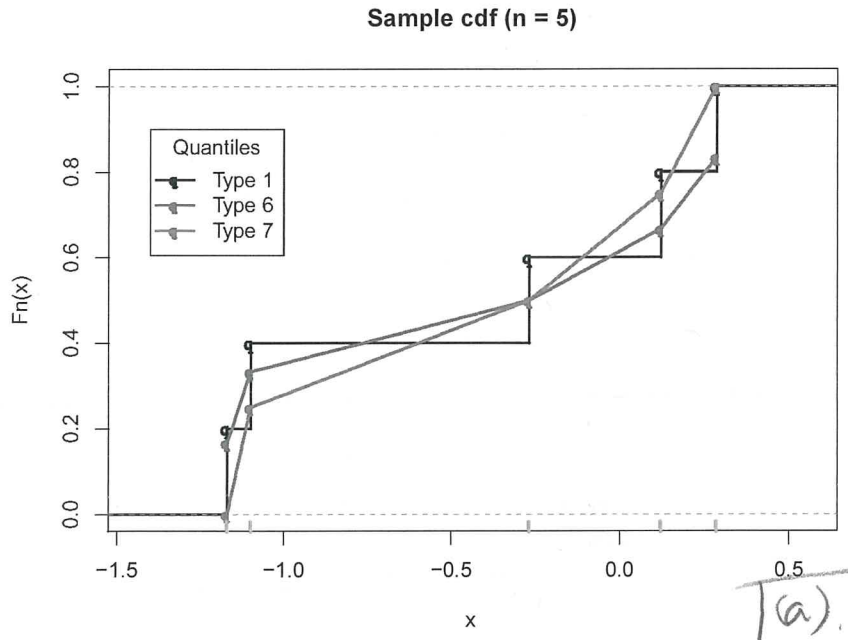
$$\hat{\pi}_p = x_{(k)}, \quad \text{if } \frac{k-1}{n} < p \leqslant \frac{k}{n}$$

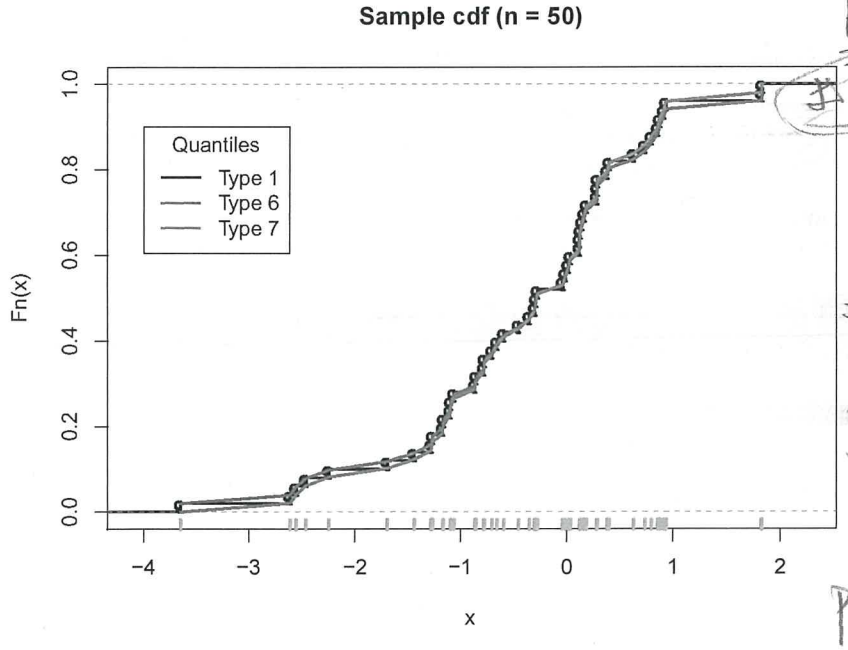
- Reminder: the *sample cdf* is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leqslant x)$$

Differences in definitions

- Different definitions imply different estimators for the cdf
- For large sample sizes, differences are negligible





2016 exam.

$X_1 \dots X_n$ on $[0,\theta]$ uniform. with pdf

$f(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \le x \le \theta \\ 0 & \theta, \text{otherwise} \end{cases}$

maximum likelihood estimator for θ is $Y = X_{(n)}$, and Y's pdf $g(y) = ny^{n-1}/\theta^n$ if $0 \le y \le \theta$, and 0 otherwise.

(a). derive an unbiased estimator of θ. using MLE

$E(Y) = \int_0^\theta (n y^{n-1}/\theta^n) y \, dy$

$= \int_0^\theta \frac{n y^n}{\theta^n} dy \left[\frac{n}{n+1} \frac{y^{n+1}}{\theta^n}\right]_0^\theta$

$= \frac{n}{n+1} \frac{\theta^{n+1}}{\theta^n} - 0 = \frac{n}{n+1}\theta$

biased

$\Rightarrow E\left(\frac{n+1}{n} Y\right) = \theta$

$\frac{n+1}{n} Y$ is an unbiased estimator

(b) verify that $\Pr(2^{\frac{1}{n}} \le \frac{Y}{\theta} \le 1) = 1-\alpha$ and use this to find $100*(1-\alpha)\%$ CI for θ

$\Pr(Y \le c) = \Pr(X_{(n)} \le c)$

$= [\Pr(X \le c)]^n = \left(\frac{c}{\theta}\right)^n$

$F(x)$ ← cdf of uniform

$\Pr(2^{\frac{1}{n}} \le \frac{Y}{\theta} \le 1) = \Pr(\theta 2^{\frac{1}{n}} \le Y \le \theta)$

$= \Pr(Y \le \theta) - \Pr(Y < \theta 2^{\frac{1}{n}})$

$= 1 - \left(\frac{\theta 2^{\frac{1}{n}}}{\theta}\right)^n$

$= 1 - 2$

$\Pr(Y \le \theta \le Y 2^{\frac{1}{n}}) \Rightarrow 100(1-\alpha)\%$ CI

$(y, y 2^{\frac{1}{n}})$ (2)

(c) 3.1 8.0 8.9 9.4 3.7

95% CI for θ

$Y = X_{(5)} = 9.4 \Rightarrow$ 95% CI is

$(9.4, 9.4 \times 0.05^{-\frac{1}{5}}) = (9.4, 17.1)$

Distribution on the cdf scale

- Reminder: for a continuous distribution, $F(X) \sim \text{Unif}(0,1)$
- Proof: for $0 \leqslant w \leqslant 1$,

$$G(w) = \Pr(F(X) \leqslant w) = \Pr(X \leqslant F^{-1}(w)) = F(F^{-1}(w)) = w$$

so the density is

$$g(w) = G'(w) = 1, \quad 0 \leqslant w \leqslant 1$$

so $F(X) \sim \text{Unif}(0,1)$.

- Since F is non-decreasing, we have

$$F(X_{(1)}) < F(X_{(2)}) < \cdots < F(X_{(n)})$$

- So $W_i = F(X_{(i)})$ are order statistics from a $\text{Unif}(0,1)$ distribution

- The cdf is $G(w) = w$, for $0 < w < 1$
- So the pdf of kth order statistic $W_k = F(X_{(k)})$ is

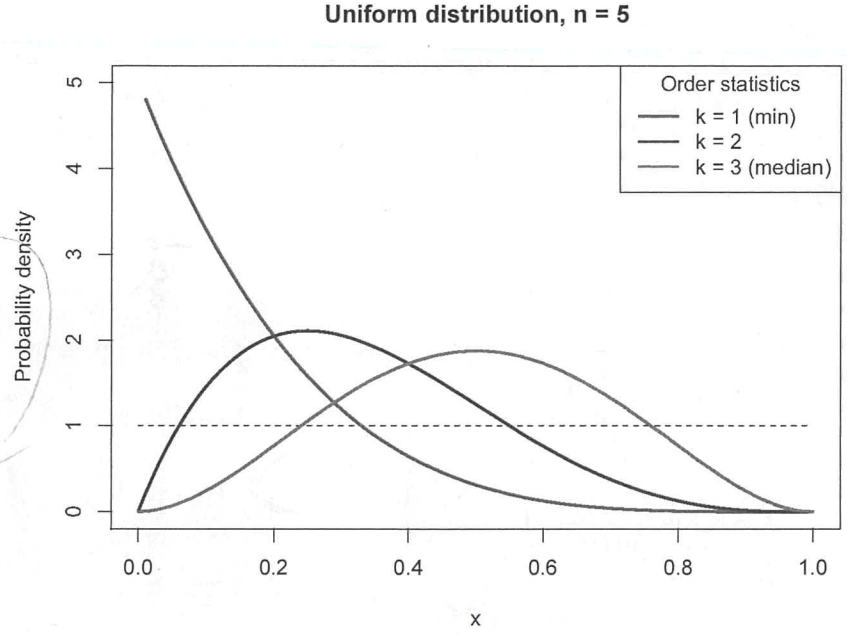
$$g_k(w) = k\binom{n}{k} w^{k-1} (1-w)^{n-k}$$

- This is a beta distribution,

$$F(X_k) \sim \text{Beta}(k, n-k+1)$$

- We can derive that:

$$\mathbb{E}(W_k) = \frac{k}{n+1}$$
$$\text{mode}(W_k) = \frac{k-1}{n-1}$$



Defining the estimators

- How does this relate to the definitions of the estimators?
- Consider:

$$\Pr(X \leqslant X_{(k)}) = F(X_{(k)})$$
$$\Pr(X \leqslant \pi_p) = F(\pi_p) = p$$

- Have $F(X_{(k)})$ probability to the left of $X_{(k)}$, need p probability to the left π_p
- Just need to relate them
- $F(X_{(k)})$ is the (random!) area to the left $X_{(k)}$
- We know its distribution, so can summarise it
- For example, $\mathbb{E}(F(X_{(k)})) = k/(n+1)$
- This suggests $X_{(k)}$ can be an estimator of π_p where $p = k/(n+1)$
- So, define $\hat{\pi}_p = X_{(k)}$ where $p = k/(n+1)$
- For other values of p, linearly interpolate

Sample median

- The *sample median* is

$$\hat{m} = \begin{cases} X_{((n+1)/2)} & \text{when } n \text{ is odd} \\ \frac{1}{2}\left(X_{(n/2)} + X_{((n/2)+1)}\right) & \text{when } n \text{ is even} \end{cases}$$

- Consistent with most definitions of the sample quantiles (not type 1!)

2.2 Asymptotic distribution

Asymptotic distribution

- For large sample sizes, it can be shown that

$$\hat{\pi}_p \approx \mathrm{N}\left(\pi_p, \frac{p(1-p)}{nf(\pi_p)^2}\right)$$

  where f is the pdf of the population distribution
- The median, $\hat{M} = \hat{\pi}_{0.5}$, is convenient special case,

$$\hat{M} \approx \mathrm{N}\left(m, \frac{1}{4nf(m)^2}\right)$$

Example (normal distribution)

- Random sample: $X \sim \mathrm{N}(\mu, \sigma^2)$ of size n
- Compare $\bar{X}$ and $\hat{M}$ as estimators of μ
- Already know,

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- Now we also know,

$$\hat{M} \approx \mathrm{N}\left(m, \frac{1}{4nf(m)^2}\right)$$

- Note that $m = \mu$ and,

$$f(m) = f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

- This gives,

$$\hat{M} \approx \mathrm{N}\left(\mu, \frac{\pi}{2}\frac{\sigma^2}{n}\right)$$

- Does the $\pi/2$ look familiar?
- ... problem 3, week 2!
- The sample mean, $\bar{X}$, is a more *efficient* estimator of μ than the sample median, $\hat{M}$
- In other scenarios, it can be the other way around

2.3 Confidence intervals for quantiles

Confidence intervals for quantiles

- Can we construct **distribution-free** CIs for quantiles?
- Can do so based on order statistics
- Procedure is the 'inverse' of the sign test

Example (CI for median)

- Take iid samples $X_1, \ldots, X_5$
- $X_{(3)}$ is an estimator of the median $m = \pi_{0.5}$
- For the median to be between $X_{(1)}$ and $X_{(5)}$ must have at least one $X_i < m$ but not five $X_i < m$
- If the distribution is continuous, $\Pr(X < m) = 0.5$
- Let W be the number of $X_i < m$, then $W \sim \text{Bi}(5, 0.5)$ and

$$\begin{aligned}\Pr(X_{(1)} < m < X_{(5)}) &= \Pr(1 \leqslant W \leqslant 4) \\ &= \sum_{k=1}^{4} \binom{5}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{5-k} \\ &= 1 - 0.5^5 - 0.5^5 = \frac{15}{16} \approx 0.94\end{aligned}$$

- So $(x_{(1)}, x_{(5)})$ is a 94% confidence interval for m

Confidence intervals for the median

- In general, want i and j so that, to the closest possible extent,

$$\begin{aligned}\Pr(X_{(i)} < m < X_{(j)}) &= \Pr(i \leqslant W \leqslant j-1) \\ &= \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} \approx 1 - \alpha\end{aligned}$$

- Need to use computed binomial probabilities (e.g. R) to determine i and j
- Or use the normal approximation to the binomial
- Note that these confidence intervals do not arise from pivots and cannot achieve 95% confidence exactly

Example (lengths of fish)

- Lengths of 9 fish (in cm), in ascending order:

  15.5, 19.0, 21.2, 21.7, 22.8, 27.6, 29.3, 30.1, 32.5

- Now,

$$\Pr(X_{(2)} < m < X_{(8)}) = \sum_{k=2}^{7} \binom{9}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{9-k} = 0.9610$$

- In R:

```
> pbinom(7, size = 9, prob = 0.5) -
+ pbinom(1, size = 9, prob = 0.5)
[1] 0.9609375
```

- So a 96.1% confidence interval for m is (19.0, 30.1)

Confidence intervals for arbitrary quantiles

- Argument can be extended to any quantile and any order statistics,
- For example, the ith and jth,

$$\begin{aligned}1 - \alpha &= \Pr(X_{(i)} < \pi_p < X_{(j)}) \\ &= \Pr(i \leqslant W \leqslant j-1) \\ &= \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}\end{aligned}$$

Example (income distribution)

- Incomes (in \$100's) for a sample of 27 people, in ascending order:

  161, 169, 171, 174, 179, 180, 183, 184, 186,
  187, 192, 193, 196, 200, 204, 205, 213, 221,
  222, 229, 241, 243, 256, 264, 291, 317, 376

- Want to estimate the first quartile, $\pi_{0.25}$
- W is the number of the X's below $\pi_{0.25}$
- $W \sim \text{Bi}(27, 0.25) \approx N(\mu = 27/4 = 6.75,\ \sigma^2 = 81/16)$
- This gives

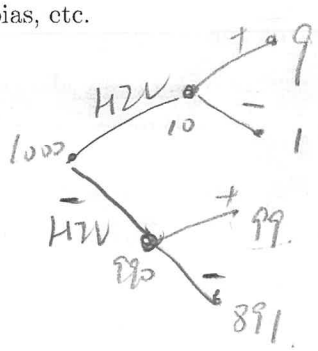
$$\begin{aligned}
&\Pr(X_{(4)} < \pi_{0.25} < X_{(10)}) \\
&= \Pr(4 \leqslant W \leqslant 9) \\
&= \Pr(3.5 < W < 9.5) \qquad \text{(continuity correction)} \\
&= \Phi\left(\frac{9.5 - 6.75}{9/4}\right) - \Phi\left(\frac{3.5 - 6.75}{9/4}\right) \\
&= 0.815
\end{aligned}$$

- So (\$17 400, \$18 700) is an 81.5% CI for the first quartile

3 Resampling methods

Resampling

- What if maths is too hard?
- Try a *resampling* method
- Replaces mathematical derivation with brute force computation
- Used for approximating sampling distributions, standard errors, bias, etc.
- Sometimes work brilliantly, sometimes not at all



Bootstrap

- Most popular resampling method: the *bootstrap*
- Basic idea:
  - Use the sample cdf as an approximation to the true cdf
  - Simulate new data from the sample cdf
  - Equivalent to sampling with replacement from the actual data
- Use these *bootstrap samples* to infer sampling distributions of statistics of interest
- This is an advanced topic
- Only a 'taster' is presented...
- ...in the lab (week 10)

$\Pr(+ \mid HIV) = 90\%$

$\Pr(- \mid \overline{HIV}) = 90\%$

$\Pr(HIV \mid +) = \frac{9}{9+99} = \frac{9}{108}$