USING STATISTICS IN RESEARCH

David A. Stephens

Department of Mathematics, Imperial College

d.stephens@imperial.ac.uk
stats.ma.ic.ac.uk/~das01/StatsShortCourse/

3rd March 2004

Module 2 : 3rd March

Extensions to Statistical Hypothesis Testing

- Analysis of Variance
- Non-normal/integer valued data
- Non-parametric Tests
- Simulation-based methods

HYPOTHESIS TESTING: THE GENERAL PROCEDURE

For a data sample $x_1, ..., x_n$, random variables $X_1, ..., X_n$, we

1. consider a pair of competing **hypotheses**, H_0 and H_1
2. define a suitable test statistic random variable $T = T(X_1, ..., X_n)$ (that is, some function of the original random variables; this will ultimately be computed to give the **observed test statistic**)
3. **assume that H_0 is true**, and compute the sampling distribution of T, f_T or F_T; this is the **null distribution**
4. compute the **observed** value of T, $t = T(x_1, ..., x_n)$; this is the **test statistic**
5. assess whether t is a surprising observation from the distribution f_T. If it **is** surprising, we have evidence to **reject** H_0; if it is not surprising, we **cannot reject** H_0

KEY POINT:
This is a **generic** approach that we have seen applied in the normal, one and two-sample case. Effectively, for the p-value, we have computed

P [Data at least **as extreme as** the data we did observe | Null model is **TRUE**]

or,

$$P\left[T(X) \geq t \mid H_0\textbf{TRUE}\right]$$

This strategy can be applied to more complicated normal examples, and also non-normal and non-parametric testing situations. It is a general strategy for assessing the statistical evidence for or against a hypothesis.

Note that we only have to compute the probability conditional on the "null" hypothesis being **true**.

SECTION 1.

ANALYSIS OF VARIANCE

The first extension we consider still presumes a normality assumption for the data, but extends the ideas from Z and T tests, which compare at most two samples, to allow for the analysis of any number of samples.

Analysis of variance or **ANOVA** is used to display the sources of variability in a collection of data groups.

The ANOVA F-test compares variability **between** groups with the variability **within** groups.

1.1 ONE-WAY ANOVA

The T-test can be extended to allow a test for differences between more than two data samples. Suppose there are K groups of sizes $n_1, ..., n_K$ (let $n = n_1 + ... + n_K$) from different populations. Let y_{kj} be the jth observation in the kth group, then

$$y_{kj} = \mu_k + \varepsilon_{kj}$$

for $k = 1, ..., K$, and $\varepsilon_{kj} \sim N\left(0, \sigma^2\right)$. This model assumes that

$$Y_{kj} \sim N\left(\mu_k, \sigma^2\right)$$

and that the expectations for the different groups are different. We can view the data as a table comprising K columns, with each column corresponding to a sample.

The groups are commonly referred to as **FACTORS.**

EXAMPLE: ANTIBIOTIC/SERUM PROTEIN BINDING

	Penicillin G	Tetra-cyclin	Strepto-mycin	Erythro-mycin	Chloram-phenicol
	29.6	27.3	5.8	21.6	29.2
	24.3	32.6	6.2	17.4	32.8
	28.5	30.8	11.0	18.3	25.0
	32.0	34.8	8.3	19.0	24.2
Mean	28.6	31.4	7.8	19.1	27.8

Is there any evidence that the amount of serum-binding differs across antibioitics ?

To test the hypothesis that each column (or "population") has the same mean, that is, the hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_K \\ H_1 &: \text{not } H_0 \end{aligned}$$

an **Analysis of Variance (ANOVA)** F-test may be carried out.

The alternative hypothesis H_1 corresponds to the model where **at least one** of the μ parameters, the mean levels for the factors, is different from the others.

To carry out a test of the hypothesis, the following **ANOVA table** should be completed;

Source	D.F.	Sum of squares	Mean square	$ANOVA-F$ F_a
Between Samples	$K-1$	FSS	$FSS/(K-1)$	$\frac{FSS/(K-1)}{RSS/(n-K)}$
Within Samples	$n-K$	RSS	$RSS/(n-K)$	
Total	$n-1$	TSS		

The test is completed by evaluating a p-value using the **observed** $ANOVA-F$ statistic, f_a, that is, the probability

$$P\left[F_a \geq f_a | F \text{ has a } Fisher-F\left(K-1, n-K\right) \text{ distribution }\right]$$

where

$$TSS = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (y_{kj} - \overline{y}_{..})^2 \quad RSS = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (y_{kj} - \overline{y}_{k})^2$$

$$FSS = \sum_{k=1}^{K} n_k (\overline{y}_{k} - \overline{y}_{..})^2$$

where

- TSS is the **total** sum-of-squares (i.e. total deviation from the overall data mean $\overline{y}_{\cdot}$)
- RSS is the **residual** sum-of-squares (i.e. sum of deviations from individual group means $\overline{y}_k$, $k = 1, ..., K$) and
- FSS is the **fitted** sum-of-squares (i.e. weighted sum of deviations of group means from the overall data mean, with weights equal to number of data points in the individual samples)

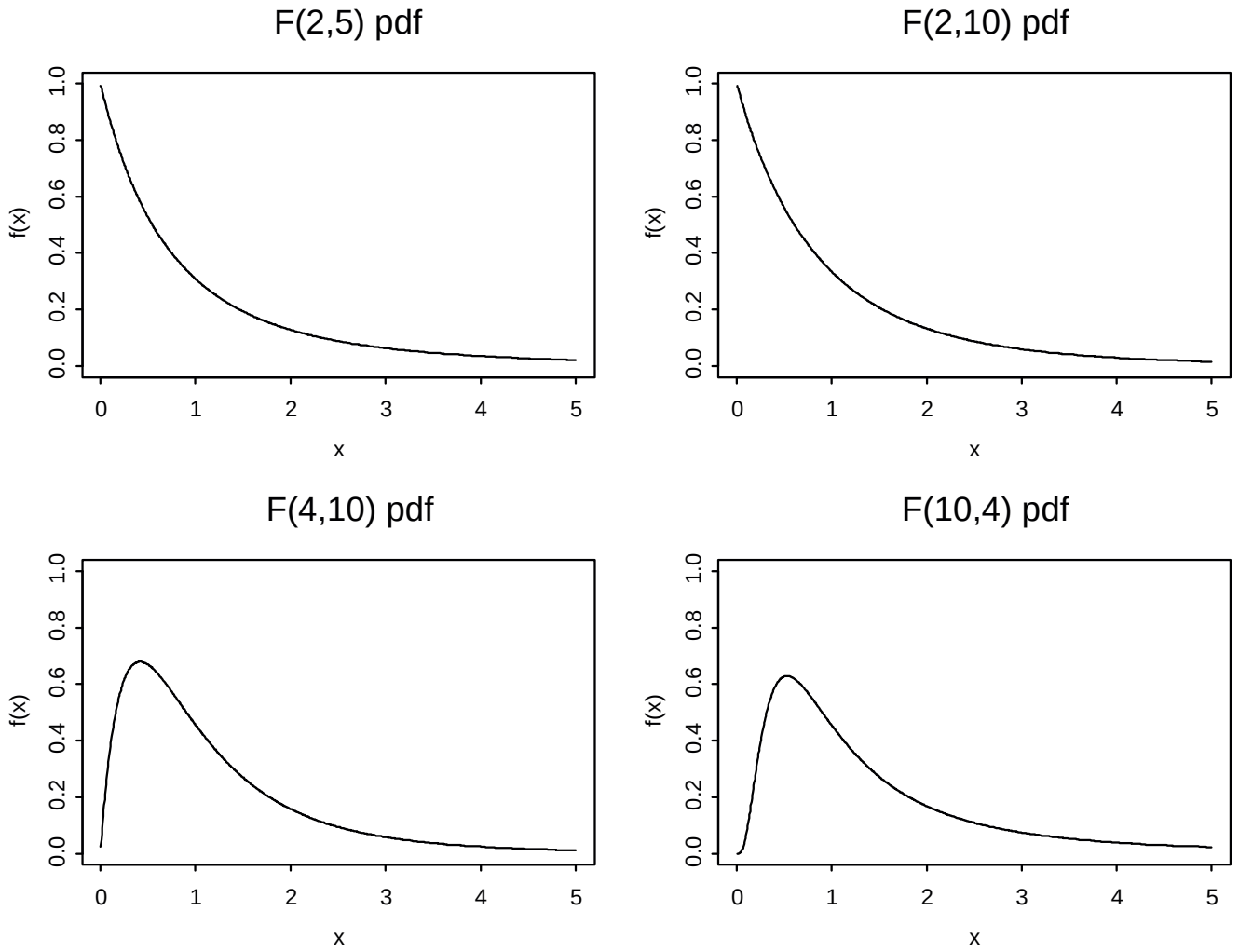
Note:that

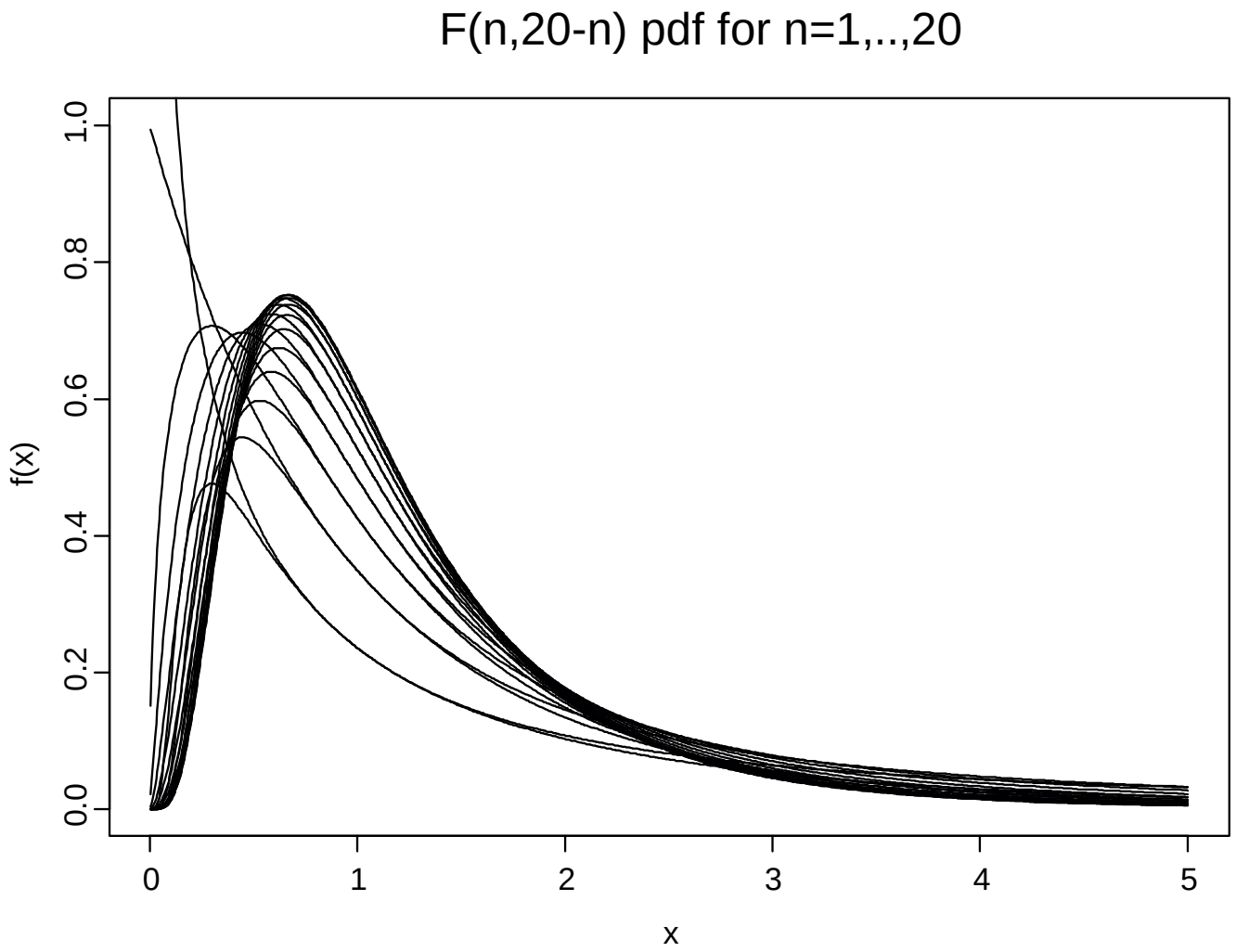
$$TSS = FSS + RSS$$

The definitions of these three sums of squares quantities gives insight into how ANOVA works by decomposing the total variation in the observed data

- TSS is the **overall** variation
- FSS is the variation caused by the **systematic** component (that is, the differences in group means)
- RSS is the **random** variation

If the F statistic is calculated in this way, and compared with an F distribution with parameters $K-1$, $n-K$, the hypothesis that all the individual samples have the same mean can be tested. We write $F_{K-1,n-K}$ for this Fisher-F distribution.


F(2,5) pdf
F(2,10) pdf
F(4,10) pdf
F(10,4) pdf
f(x)
x
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5


F(n,20-n) pdf for n=1,..,20
f(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
x

EXAMPLE: ANTIBIOTIC/SERUM PROTEIN BINDING

Source	D.F.	Sum of squares	Mean square	F
SERUM	4	1480.82	370.21	40.88
Residual	15	135.82	9.05	
Total	19	1616.64		

which gives a p-value (of 6.74×10^{-8}) in comparison with a *Fisher* $F_{4,15}$ distribution)

This is a highly statistically significant result, and thus there is strong evidence to **reject** the null hypothesis that the mean serum protein binding is equal for all antibiotics (under the ANOVA assumptions).

EXAMPLE Three genomic segments were used to studied in order to discover whether the distances (in kB) between successive occurrences of a particular motif were substantially different. Several measurements were taken using for each segment;

	Method		
	SEGMENT A	SEGMENT B	SEGMENT C
	42.7	44.9	41.9
	45.6	48.3	44.2
	43.1	46.2	40.5
	41.6		43.7
			41.0
Mean	43.25	46.47	42.26
Variance	2.86	2.94	2.66

For these data, the ANOVA table is as follows;

Source	D.F.	Sum of squares	Mean square	F
SEGMENTS	2	34.1005	17.0503	6.11
Residual	9	25.1087	2.7899	
Total	11	59.2092		

and the F statistic must be compared with an $F_{2,9}$ distribution. For a significance test at the 0.05 level, F must be compared with the 95th percentile (in a **one-sided** test) of the $F_{2,9}$ distribution. This value is 4.26. Therefore, the F statistic **is** surprising, given the hypothesized model, and therefore there is evidence to reject the hypothesis that the segments are identical.

1.2 POST-HOC TESTS

The hypothesis of equal means across all groups is not necessarily the only hypothesis of interest. We may wish to test, for example

$$H_0 : \mu_r = \mu_s$$

against the general alternative for any possible pair of columns r and s, even if the null hypothesis of equal means in all columns is not rejected.

Pairwise tests for equality of column means that are carried out after an F-test has led to the rejection of the ANOVA null hypothesis are referred to as **post-hoc** tests. The key consideration for such tests is the appropriate correction for **multiple testing;** a number of methods have been proposed.

1.3 TWO-WAY ANOVA

One-way ANOVA can be used to test whether the underlying means of several groups of observations are equal Now consider the following data collection situation Suppose there are K treatments, and L groups of observations that are believed to have different responses, that all treatments are administered to all groups, and measurement samples of size n are made for each of the $K \times L$ combinations of treatments $\times$ groups. The experiment can be represented as follows: let y_{klj} be the jth observation in the kth treatment on the lth group, then

$$y_{klj} = \mu_k + \delta_l + \varepsilon_{klj}$$

for $k = 1, ..., K$, $l = 1, ..., L$, and again $\varepsilon_{klj} \sim N\left(0, \sigma^2\right)$.

This model assumes that $Y_{kj} \sim N\left(\mu_k + \delta_l, \sigma^2\right)$ and that the expectations for the different samples are different. We can view the data as a 3 dimensional-table comprising K columns and L rows, with n observations for each column $\times$ row combination, corresponding to a sample.

It is possible to test the hypothesis that each **treatment**, and/or that each **group** has the same mean, that is, the two null hypotheses

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = ... = \mu_K \\ H_0 &: \delta_1 = \delta_2 = ... = \delta_L \end{aligned}$$

against the alternative H_1 not H_0 in each case.

For these tests, a **Two-way Analysis of Variance (ANOVA)** F-test may be carried out.

The Two-Way ANOVA table is computed as follows

Source	D.F.	Sum of squares	Mean square	F
TREATMENTS	$K-1$	FSS_1	$FSS_1/(K-1)$	$\dfrac{FSS_1/(K-1)}{RSS/(R+1)}$
GROUPS	$L-1$	FSS_2	$FSS_2/(L-1)$	$\dfrac{FSS_2/(L-1)}{RSS/(R+1)}$
Residual	$R+1$	RSS	$RSS/(R+1)$	
Total	$N-1$	TSS		

where $N = K \times L \times n$, $R = N - L - K$. and again

$$TSS = FSS_1 + FSS_2 + RSS.$$

In the table below, there are $K = 6$ Treatments, and $L = 3$ Groups, and $n = 1$

	I	II	III	GROUP totals
1	0.96	0.94	0.98	2.88
2	0.96	0.98	1.01	2.95
3	0.85	0.87	0.86	2.58
4	0.86	0.84	0.90	2.60
5	0.86	0.87	0.89	2.62
6	0.89	0.93	0.92	2.74
TREATMENT totals	5.38	5.43	5.56	16.37

There are two natural hypotheses to test; first, do the TREATMENTS differ, and second, do the GROUPS differ ?

Two-way analysis of variance can be used to analyze such data. Given two sources of variation the data can be thought of as a table with the rows and columns representing these two sources . Two-way analysis of variance studies the variability due to

- the GROUP effect (here, variability between the columns),
- and the variability due to the TREATMENT effect (variability between the rows)

and calibrates them against the average level of variability in the data overall. Having performed the appropriate calculations, the results are displayed in an ANOVA table.

For example, for the data above

Source	D.F.	Sum of squares	Mean square	F
TREATMENT	5	0.040828	0.0081656	31.54
GROUP	2	0.002878	0.001439	5.57
Residual	10	0.002589	0.0002589	
Total	17	0.046295		

The two F statistics can be interpreted as follows;

- the first ($F = 31.54$) is the test statistic for the test of equal means in the **rows**, that is, that there is no difference between **TREATMENTS**. This statistic must be compared with an

  $$F_{5,10}$$

  distribution (the two degrees of freedom being the entries in the degrees of freedom column in the specimens and residual rows of the ANOVA table). The 95th percentile of the $F_{5,10}$ distribution is 3.33, and thus the test statistic is **more extreme** than this critical value, and thus the hypothesis that each specimen has the same mean can be **rejected**.

- The second F statistic, ($F = 5.57$), is the test statistic for the test of equal means in the **columns**, that is, that there is no difference between **GROUPS**. This statistic must be compared with an

$$F_{2,10}$$

distribution (the two degrees of freedom being the entries in the degrees of freedom column in the methods and residual rows of the ANOVA table). The 95th percentile of the $F_{2,10}$ distribution is 4.10, and thus the test statistic is **more extreme** than this critical value, and thus the hypothesis that each method has the same mean can be **rejected**.

Note: In this example, we do **not** have replicate data; this limits the complexity of the model that we can fit. Ideally we would like to be able to fit an **interaction** between the two-factors.

Mean parameters in two-way cross classification (full models):

No Interaction model

	I	II	III	IV	V
1	$\mu_1 + \delta_1$	$\mu_1 + \delta_2$	$\mu_1 + \delta_3$	$\mu_1 + \delta_4$	$\mu_1 + \delta_5$
2	$\mu_2 + \delta_1$	$\mu_2 + \delta_2$	$\mu_2 + \delta_3$	$\mu_2 + \delta_4$	$\mu_2 + \delta_5$
3	$\mu_3 + \delta_1$	$\mu_3 + \delta_2$	$\mu_3 + \delta_3$	$\mu_3 + \delta_4$	$\mu_3 + \delta_5$

Interaction model

	I	II	III	IV	V
1	γ_{11}	γ_{12}	γ_{13}	γ_{14}	γ_{15}
2	γ_{21}	γ_{22}	γ_{23}	γ_{24}	γ_{25}
3	γ_{31}	γ_{32}	γ_{33}	γ_{34}	γ_{35}

No Interaction Model: $8 = 3 + 5$ parameters
Interaction Model: $15 = 3 \times 5$ parameters; can be fit if replicate data are available.

1.4 ANOVA: KEY ASSUMPTIONS

In ANOVA, there are three key assumptions

(i) all data are **independent**

(ii) the data are **normally** distributed

(iii) the data subgroups (defined by the cross classification by factors) have **equal variances.**

Of these three points, (i) can be assessed by consideration of the study design, (ii) can be be tested formally using methods that will be described in later sections, and (iii) can be tested using statistical hypothesis testing in the following way using **Levene's Test**

LEVENE'S TEST FOR HOMOGENEITY OF VARIANCE

The Levene test is defined for the two hypotheses as follows: suppose that the data Y of size n is partitioned into K subgroups of sizes $n_1, ..., n_K$ where $n = n_1 + ... + n_K$. It is of interest to test whether the subgroups have the same variance, that is the hypothesis

$$\begin{aligned} H_0 &: \sigma_1 = \sigma_2 = ... = \sigma_K \\ H_1 &: \sigma_i \neq \sigma_j \text{ for at least one pair } (i, j). \end{aligned}$$

Test Statistic

$$W = \frac{(n-K)\sum_{i=1}^{K} n_i \left(\overline{Z}_i - \overline{Z}\right)^2}{(K-1)\sum_{i=1}^{K}\sum_{j=1}^{n_k} n_i \left(Z_{ij} - \overline{Z}_i\right)^2}$$

where Z_{ij} can have one of the following three definitions:

1. $Z_{ij} = |Y_{ij} - \overline{Y_i}|$, where $\overline{Y_i}$ is the mean of the ith subgroup.
2. $Z_{ij} = |Y_{ij} - Y_i^{(MEDIAN)}|$ where $Y_i^{(MEDIAN)}$ is the median of the ith subgroup.
3. $Z_{ij} = |Y_{ij} - Y_i^{(TRIMMED)}|$ where $Y_i^{(TRIMMED)}$ is the 10% trimmed mean of the ith subgroup.

The three choices for defining Z_{ij} determine the **robustness** (to not falsely detect unequal variances when the underlying data are not normally distributed) and **power** (to detect accurately unequal variances) of Levene's test. The Levene test rejects the hypothesis that the variances are equal at significance level α (typically, $\alpha = 0.05$) if

$$W > F_{K-1,n-K}(1-\alpha)$$

where $F_{K-1,n-K}(1-\alpha)$ is the $(1-\alpha)\,\%$ quantile of the Fisher F distribution with $K-1$ and $n-K$ degrees of freedom.

SECTION 2.

NON-NORMAL DATA

2.1 COUNTS AND PROPORTIONS

The one and two sample tests described in earlier sections can also be applied to non-normal data. A common form of non-normal data arise when the counts of numbers of "successes" or "failures" that arise in a fixed number of trials.

In this case, the Binomial distribution model is appropriate; in a one sample testing, we model the number of successes, X, by assuming

$$X \sim Binomial(n, \theta)$$

and test hypotheses about θ.

In the two sample case, we assume that the number of successes in the two samples are random variables X_1 and X_2, where

$$X_1 \sim Binomial(n_1, \theta_1)$$

$$X_2 \sim Binomial(n_2, \theta_2),$$

and perhaps test the null hypothesis

$$H_0 : \theta_1 = \theta_2$$

against some alternative hypothesis ($\theta_1 \neq \theta_2, \theta_1 > \theta_2$ or $\theta_1 < \theta_2$)

2.2 ONE-SAMPLE TESTING

In the one sample case, two alternative approaches can be adopted:

- an **exact** test, where the distribution of the chosen test statistic under $H_0 : \theta = c$ is computed exactly, giving exact critical values and p-values
- an approximate test based on a Normal approximation to the binomial distribution.

For the exact test, we note that, **if H_0 is true**, and $\theta = c$, then $X \sim Binomial(n, c)$ so the critical values in a two-sided test can be computed directly by inspection of the $Binomial(n, c)$ c.d.f; that is

$$F_{BIN}\left(C_{R_1}; n, \theta = c\right) = 0.025 \qquad C_{R_2} = F_{BIN}\left(0.975; n, \theta = c\right)$$

where $F_{BIN}\left(-; n, \theta\right)$ is the c.d.f. of the $Binomial(n, \theta)$ distribution

$$F_{BIN}\left(x; n, \theta\right) = \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} \theta^i \left(1 - \theta\right)^{n-i}$$

where

$$\lfloor x \rfloor \text{ is largest whole number} \leq x.$$

For the approximate test, we use the fact that

$$X \sim Binomial(n, \theta) \approx Normal\left(n\theta, n\theta\left(1-\theta\right)\right)$$

and hence random variable Z

$$Z = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$$

is approximately distributed as $Normal(0, 1)$. For the approximate test of $H_0 : \theta = c$, we therefore use the test statistic

$$z = \frac{x - nc}{\sqrt{nc(1-c)}}$$

(x is the actual, observed count) and compare this with the standard normal c.d.f.. This test is virtually equivalent to the one-sample t-test.

2.3 TWO SAMPLE TESTING

For a two sample test of $H_0 : \theta_1 = \theta_2$, we use a similar normal approximation to the one-sample case. If H_0 is true, then there is a common probability θ determining the success frequency in both samples, and the maximum likelihood estimate of θ is

$$\widehat{\theta} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{x}{n}, \text{ say}$$

and thus it can be shown that the test statistic.

$$z = \frac{\dfrac{x_1}{n_1} - \dfrac{x_2}{n_2}}{\sqrt{\dfrac{(n_1 + n_2)}{n_1 n_2}\left(\dfrac{x_1 + x_2}{n_1 + n_2}\right)\left(1 - \dfrac{x_1 + x_2}{n_1 + n_2}\right)}}$$

has an approximate standard Normal distribution.

2.4 CONTINGENCY TABLES

Contingency tables are constructed when a sample of data of size n are classified according to D factors, with each factor having k_d levels or categories, for $d = 1, ..., D$. When the classification is complete, the result can be represented by a D-way table of $k_1 \times k_2 \times ... \times k_D$ "cells", with each cell containing a fraction of the original data. For example, if $D = 2$, the table consists of k_1 rows and k_2 columns, and the number data in cell (i, j) is denoted n_{ij} for $i = 1, ..., k_1$ and $j = 1, ..., k_2$, where

$$\sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n_{ij} = n$$

Such a table when $D = 2$, $k_1 = 4$ and $k_2 = 6$ is illustrated below

		COLUMN						
		1	2	3	4	5	6	Total
ROW	1	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}	$n_{1.}$
	2	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{26}	$n_{2.}$
	3	n_{31}	n_{32}	n_{33}	n_{34}	n_{35}	n_{36}	$n_{3.}$
	4	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{46}	$n_{4.}$
	Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	$n_{.6}$	n

This is a **cross-classification** table; it says that n_{ij} out of a total of n individuals had

- row classification i
- column classification j

2.5 CHI-SQUARED GOODNESS-OF-FIT TEST

It is often of interest to test whether row classification is **independent** of column classification, as this would indicate **independence** between row and column factors. An approximate test can be carried out using a **Chi-Squared Goodness-of-Fit** statistic; if the independence model is correct, the expected cell frequencies $\hat{n}_{ij}$ can be calculated as

$$\hat{n}_{ij} = \frac{n_{i.} n_{.j}}{n} \qquad i = 1, ..., k_1, \ j = 1, ..., k_2$$

where $n_{i.}$ is the *total* of cell counts in row i and $n_{.j}$ is the *total* of cell counts in column j, and that, under independence, the χ^2 test statistic

$$\chi^2 = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

has an approximate chi-squared distribution with $(k_1 - 1)(k_2 - 1)$ degrees of freedom.

2.6 LIKELIHOOD RATIO TEST

Another approximate test is based on a **Likelihood Ratio** (**LR**) statistic

$$LR = 2\sum_{i=1}^{k_1}\sum_{j=1}^{k_2} n_{ij} \log \frac{n_{ij}}{\hat{n}_{ij}}$$

This statistic also has an approximate Chi-squared distribution

$$\chi^2_{(k_1-1)(k_2-1)}$$

again given that H_0 **is true.**

It compares the "likelihood" under the independence model with the likelihood of the "saturated" model that fits a parameter for each cell in the table;

- the independence model has

$$1 + (k_1 - 1) + (k_2 - 1) = k_1 + k_2 - 1$$

  parameters, and hence

$$(k_1 \times k_2) - (k_1 + k_2 - 1) = (k_1 - 1)(k_2 - 1)$$

  degrees of freedom

- the saturated model has $(k_1 \times k_2)$ parameters, and hence 0 degrees of freedom

- the difference in degrees of freedom is hence

$$(k_1 - 1)(k_2 - 1)$$

EXAMPLE: 4×4 TABLE PHENOTYPIC RELATIONSHIP

		Hair colour				
		Black	Brunette	Red	Blonde	Total
	Brown	68	119	26	7	220
	Blue	20	84	17	94	215
Eye color	Hazel	15	54	14	10	93
	Green	5	29	14	16	64
	Total	108	286	71	127	592

Number of tables: 1,225,914,276,276,768,514

Any evidence of **dependence/association** between traits ?

EXAMPLE : DESCENDENTS OF QUEEN VICTORIA

Month of birth	Month of death												
	Jan	Feb	March	April	May	June	July	Aug	Sept	Oct	Nov	Dec	Total
Jan	1	0	0	0	1	2	0	0	1	0	1	0	6
Feb	1	0	0	1	0	0	0	0	0	1	0	2	5
March	1	0	0	0	2	1	0	0	0	0	0	1	5
April	3	0	2	0	0	0	1	0	1	3	1	1	12
May	2	1	1	1	1	1	1	1	1	1	1	0	12
June	2	0	0	0	1	0	0	0	0	0	0	0	3
July	2	0	2	1	0	0	0	0	1	1	1	2	10
Aug	0	0	0	3	0	0	1	0	0	1	0	2	7
Sept	0	0	0	1	1	0	0	0	0	0	1	0	3
Oct	1	1	0	2	0	0	1	0	0	1	1	0	7
Nov	0	1	1	1	2	0	0	2	0	1	1	0	9
Dec	0	1	1	0	0	0	1	0	0	0	0	0	3
Total	13	4	7	10	8	4	5	3	4	9	7	8	82

Any evidence of **association** between birth/death months ?
Note: A very "sparse" table.

EXAMPLE : CLASSIFICATION OF PURUM MARRIAGES

	Husband				
Wife	Marrim	Makan	Parpa	Thao	Kheyang
Marrim	-	5	17	-	6
Makan	5	-	0	16	2
Parpa	-	2	-	10	11
Thao	10	-	-	-	9
Kheyang	6	20	8	0	1

Structural zeros (marriages forbidden) "-" and real zeros 0.

Any evidence of symmetry/independence ?

EXAMPLE : SQUIRREL MONKEY DATA

Ploog [1967] observed the following distribution of genital display among the members of a colony of six squirrel monkeys (labeled as R, S, T, U, V, and W). For each display there is an active and passive participant, but a monkey never displays toward himself.

Active Participant	Passive Participant R	S	T	U	V	W	Totals
R		1	5	8	9	0	23
S	29		14	46	4	0	93
T	0	0	—	0	0	0	0
U	2	3	1		38	2	46
V	0	0	0	0		1	1
W	9	25	4	6	13		57
Totals	40	29	24	60	64	3	220

Wish to fit **symmetry** model ?

2.7 2×2 TABLES

When $k_1 = k_2 = 2$, the contingency table reduces to a two-way binary classification

		COLUMN		
		1	2	Total
ROW	1	n_{11}	n_{12}	$n_{1.}$
	2	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

In this case we can obtain some more explicit tests: one is again an **exact test**, the other is based on a normal approximation. The chi-squared test described above is feasible, but other tests may also be constructed:

- **FISHER'S EXACT TEST FOR INDEPENDENCE**

  Suppose we wish to test for **independence** between the row and column variables of a contingency table. When the data consist of two categorical variables, a contingency table can be constructed reflecting the number of occurrences of each factor combination. **Fisher's exact test** assesses whether the classification according to one factor is independent of the classification according to the other, that is the test is of the null hypothesis H_0 that the factors are independent, against the general alternative, **under the assumption that the row and column totals are fixed**.

- – The data for such a table comprises the row and column totals $(n_{1.}, n_{2.}, n_{.1}, n_{.2})$ and the cell entries

$$(n_{11}, n_{12}, n_{21}, n_{22})$$

The test statistic can be defined as the upper left cell entry n_{11}; for the null distribution, we compute the probability of the observing **all possible tables** with these row and column totals.. Under H_0 this distribution is **hypergeometric** and the probability of observing the table $(n_{11}, n_{12}, n_{21}, n_{22})$ is

$$p(n_{11}) = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n}{n_{.1}}} = \frac{n_{1.}!n_{.1}!n_{2.}!n_{.2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}$$

where $n! = 1 \times 2 \times 3 \times .. \times (n-1) \times n$.

- – For the p-value, we need to assess the whether or not the observed table is surprising under this null distribution; suppose we observe $n_{11} = x$, then we can compare $p(x)$ with all $p(y)$ for all feasible y, that is y in the range $\max\{0, n_{1.} - (n - n_{.1})\} \leq y \leq \min\{n, n_{.1}\}$. We are thus calculating the null distribution **exactly** given the null distribution assumptions and the row and column totals; if the observed test statistic lies in the tail of the distribution, we can reject the null hypothesis of independent factors.

- **MANTEL-HAENSZEL TEST FOR INDEPENDENCE**

  This test allows you to test for independence between two factors in the presence of a third, and possibly related variable. It extends the two-way Chi-squared test of independence described above; the test statistic is a chi-squared type statistic, and the null distribution under independence is a Chi-squared distribution.

- **McNEMAR'S TEST FOR SYMMETRY IN PAIRED SAMPLES**

  In a 2×2 table representing paired data (where observations are, for example, matched in terms of medical history or genotype, or phenotype) the usual chi-squared test is not appropriate, and **McNemar's test** can instead be used. Consider the following table for a total of n matched pairs of observations, in which each individual in the pair has been classified (or randomized to class) A or B, with one A one B in each pair, and then the outcome (disease status, survival status) recorded.

		A		
		YES	NO	Total
	YES	n_{11}	n_{12}	$n_{1.}$
B	NO	n_{21}	n_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

that is, n_{11} pairs were observed for which both A and B classified individuals had disease/survival status YES, whereas n_{12} pairs were observed for which the A individual had status NO, but the B individual had status YES, and so on.

An appropriate test statistic here for a test of symmetry or "discordancy" in these results (that is, whether the two classifications are significantly different in terms of outcome) is

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

which effectively measures how different the off-diagonal entries in the table are. This statistic is an adjusted Chi-squared statistic, and has a χ^2_1 distribution under the null hypothesis that there is no asymmetry. Again a one-tailed test is carried out: "surprising" values of the test statistic are large.

SECTION 3.

NON-PARAMETRIC TESTS

The standard test for the equality of expectations of two samples is the two-sample T-test. This test is predicated on the assumption of normality of the underlying distributions. In many cases, such an assumption is inappropriate, possible due to distributional asymmetry or the presence of outliers, and thus other tests of the hypothesis of equality of population locations must be developed.

Some of the standard non-parametric tests used in statistical analysis are described below: we concentrate on two-sample tests for the most part. All of these tests can be found in good statistics packages.

References: Conover, *Practical Nonparametric Statistics*
Hollander and Wolfe, *Nonparametric Statistical Methods*

Non-parametric tests are usually based on the **ranks** of the data: typically, we

- sort the pooled data into ascending order (forming the **order statistics/empirical quantiles)**
- assign the ranks from 1 up to the total sample size to the data points
- examine statistics based on functions of the ranks (for example, the rank-sum) for data within the identified subgroups.
- base group comparison on differences in the rank statistics
- the rank statistics are used to construct a test statistic, whose distribution is typically approximated using a normal approximation.
- a "distribution-free" procedure.

3.1 THE MANN-WHITNEY-WILCOXON TEST

Consider two samples $x_1, ..., x_{n_1}$ and $y_1, ..., y_{n_2}$. The **Mann-Whitney-Wilcoxon** test proceeds as follows; first, sort the pooled sample into ascending order. Add up the ranks of the data from sample one to get u_1 say. Repeat for sample two to get u_2. Note that

$$u_1 + u_2 = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

The Mann-Whitney-Wilcoxon statistic is u_1. It can be shown that, under the hypothesis that the data are from populations with the equal medians, then u_1 has an approximate normal distribution with mean and variance

$$\frac{n_1(n_1 + n_2 + 1)}{2} \qquad \frac{n_1 n_2(n_1 + n_2 + 1)}{12}$$

This is the non-parametric alternative to the two sample t-test.

3.2 THE KOLMOGOROV-SMIRNOV TEST

The two-sample Kolmogorov-Smirnov test is a non-parametric test for comparing two samples via their **empirical cumulative distribution function**. For data $x_1, ..., x_n$, the empirical c.d.f. is the function $\widehat{F}$

$$\widehat{F}(x) = \frac{c(x)}{n} \qquad\qquad c(x) = \text{“Number of data } \leq x\text{”}$$

Thus, for two samples, we have two empirical c.d.f., and the (two-sided) **Kolmogorov-Smirnov** test that the two samples come from the same underlying distribution is based on the statistic

$$T = \max_x \left| \widehat{F}_1(x) - \widehat{F}_2(x) \right| .$$

It is easy to show that $0 \leq T \leq 1$, but the null distribution of T is not available in closed form. Fortunately, the p-value probability in the test for test statistic t, $p = \mathrm{P}[T > t]$ can be obtained for various different sample sizes using statistical tables or packages.

NOTE : There is a one-sample version of the Kolmogorov-Smirnov test for testing whether a sample are well represented by a specified probability model with cdf F_0. It is based on the test statistic

$$T = \max_x \left| \widehat{F}_1(x) - F_0(x) \right| .$$

It can be used as a **goodness-of-fit** test, to test against a specific distribution.

3.3 TESTING NORMALITY

- **THE CHI-SQUARED GOODNESS-OF-FIT TEST**
  The **chi-squared goodness-of-fit test** is a non-parametric test for which the null distribution of the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  can be well approximated by a Chi-squared distribution. In this formula, k is the number of "bins" into which the range of the data is broken down, and

  - O_i is the number of observations **observed** to fall into bin i
  - E_i is the number of observations **expected** to fall into bin i under the normal model

Alternative tests/Assessments:

- **The Shapiro-Wilk Test:** The **Shapiro-Wilk** test can be used to test this hypothesis; the test statistic is commonly denoted W, and critical and p- values from its null distribution are available from tables or statistics packages.

- The **Kolmogorov-Smirnov one-sample** test can be used in a one-sample problem to test any distributional assumptions, including normality.

- **Probability Plotting** or **Quantile-Quantile** (QQ) plotting involves plotting empirical quantiles versus theoretical quantiles; a straight line in the QQ plot indicates that the distributional assumption is valid.

3.4 THE KRUSKAL-WALLIS TEST

The **Kruskal-Wallis rank test** is a nonparametric alternative to a *one-way analysis of variance.*

- The null hypothesis is that the true location parameter is the same in each of the samples.
- The alternative hypothesis is that at least one of the samples has a different location.
- Unlike one-way ANOVA, this test does not require normality

3.5 THE FRIEDMAN RANK SUM TEST

The **Friedman rank sum test** is a nonparametric alternative to a specific *two-way analysis of variance*

- It is appropriate for data arising from an experiment in which exactly one observation was collected from each experimental unit, or group, under each treatment.
- The elements of the samples are assumed to consist of a treatment effect, plus a group effect, plus independent and identically distributed residual errors

SECTION 4.

EXACT TESTS AND SIMULATION-BASED METHODS

Chi-squared tests involved the construction of a chi-squared statistic of the form

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

The distribution of the test-statistic is approximated by a suitable Chi-squared distribution. This approximation is

- **good** when the sample size is large
- **poor** when the table is "sparse", with some low (expected) cell entries (under the null hypothesis)

We have also seen two examples of **Exact Tests:** the exact binomial test in section (2.1) and Fisher's Exact Test in section (2.7). For these tests, we proceeded as follows, mimicking the general hypothesis strategy outlined at the start of the section.

1. Write down a null hypothesis H_0 and a suitable alternative hypothesis H_1

2. Construct a test statistic T deemed appropriate for the hypothesis under study

3. Compute the null distribution of T, that is the sampling distribution of T if H_0 is true, f_T

4. Compare the observed value of T, $t = T(x)$ for sample data $x = (x_1, ..., x_n)$ with the null distribution and assess whether the observed test statistic is a surprising observation from f_T; if it is reject H_0

Step 3 is crucial: for some tests (for example, one and two sample tests based on the Normal distribution assumption), it is possible to find f_T analytically for appropriate choices of T in Step 2. For others, such as the chi-squared goodness of fit and related tests, f_T is only available approximately.

However, the null distribution (and hence the critical regions and p-value) can, in theory, **always** be found : it is the probability distribution of the statistic T under the model restriction imposed by the null hypothesis.

We may not be able to compute the null distribution **analytically** (as for the tests for normal samples), but we can do it **numerically**, using **simulation**.

For most of the hypothesis tests above, we start with the assumptions and work forward to derive the sampling distribution of the test statistic under the null hypothesis.

- For **permutation tests,** we will reverse the procedure, since the sampling distribution involves the permutations which give the procedure its name and are the key theoretical issue in understanding the test.

- For **resampling** or **bootstrap** methods , we will resample the original data uniformly and randomly so as to explore the variability of a test statistic.

4.1 PERMUTATION TESTS

A permutation is a reordering of the numbers $1, ..., n$. For example, (1, 2, 3, 4, 5, 6), (1, 3, 2, 4, 5, 6), (4, 5, 2, 6, 1, 3) (3, 2, 1, 6, 4, 5) are all permutations of the numbers 1 through 6 (note that this includes the standard order in first line). There are $n! = 1 \times 2 \times 3 \times ... \times n$ permutations of n objects.

The central idea of permutation tests refers to rearrangements of the data. The null hypothesis of the test specifies that **the permutations are all equally likely**. The sampling distribution of the test statistic under the null hypothesis is computed by forming all (or many) of the permutations, calculating the test statistic for each and considering these values all equally likely.

Consider the following two group example, where we want to test for any significant difference between the groups.

$$\begin{aligned} \text{Group 1} &: 55, 58, 60 \\ \text{Group 2} &: 12, 22, 34 \end{aligned}$$

Here are the steps we will follow to use a permutation test to analyze the differences between the two groups. For the original order the sum for Group 1 is 173. In this example, if the groups were truly equal (**and the null hypothesis was true**) then randomly moving the observations among the groups would make no difference in the sum for Group 1. Some of the sums would be a little larger than the original sum and some would be a bit smaller. For the six observations there are 720 permutations of which there are 20 distinct combinations for which we can compute the sum of Group 1.

	GROUP 1	GROUP 2	SUM		GROUP 1	GROUP 2	SUM
1	55,58,60	12,22,34	173	11	12,22,60	55,58,34	94
2	55,58,12	60,22,34	125	12	12,58,22	55,60,34	92
3	55,58,22	12,60,34	135	13	55,12,22	12,55,58	89
4	55,58,34	12,22,34	148	14	12,34,60	55,58,34	106
5	55,12,60	58,22,34	127	15	12,58,34	55,22,60	104
6	55,22,60	12,58,34	137	16	55,12,34	12,58,60	101
7	55,34,60	12,22,58	149	17	22,34,60	55,58,34	116
8	12,58,60	55,22,34	130	18	22,58,34	55,22,60	114
9	22,58,60	12,55,34	140	19	55,22,34	12,58,60	111
10	34,58,60	12,22,55	152	20	12,22,34	55,58,60	68

Only **one** of the twenty orderings has a Group 1 sum that greater than that of the original ordering; thus the probability of a sum at least this large by chance alone is $1/20 = 0.05$; it can be considered statistically significant.

4.2 MONTE CARLO METHODS

Above, the permutation yielded an **exact test** because we were able to enumerate all of the possible combinations. In larger examples it will not be possible , so we will have to take a large number of random orderings, sampled uniformly from the permutation distribution.

Monte Carlo methods replace an **analytic** calculation of the probability function by a **numerical**, **simulation-based** one. The principal is that large **samples** from probability distributions can be used accurately to **approximate** the probability distribution itself.

A general **Monte Carlo** strategy for two sample testing is outlined below:

1. For two sample tests for samples of size n_1 and n_2, compute the value of the test statistic for the observed sample t^*

2. Randomly select one of the $(n_1 + n_2)!$ permutations, re-arrange the data according to this permutation, allocate the first n_1 to pseudo-sample 1 and the remaining n_2 to pseudo-sample 2, and then compute the test statistic t_1

3. Repeat 2. N times to obtain a random sample of $t_1, t_2, ..., t_N$ of test statistics from the TRUE null distribution.

4. Compute the p-value by reporting

$$\frac{\text{Number of } t_1, t_2, ..., t_N \text{ more extreme than } t^*}{N}$$

this value will be a good approximation to the true $p-$value if the Monte Carlo sample size N is large enough.

4.3 THE BOOTSTRAP AND JACKKNIFE

In statistical analysis, we usually interested in obtaining estimates of a parameter via some statistic, and also an estimate of the variability or uncertainty attached to this point estimate, and a confidence interval for the true value of the parameter.

Traditionally, researchers have relied on normal approximations to obtain standard errors and confidence intervals. These techniques are valid only if the statistic, or some known transformation of it, is asymptotically normally distributed. If the normality assumption does not hold, then the traditional methods should not be used to obtain confidence intervals. A major motivation for the traditional reliance on normal-theory methods has been computational tractability, computational methods remove the reliance on asymptotic theory to estimate the distribution of a statistic.

Resampling techniques such as the **bootstrap** and **jackknife** provide estimates of the standard error, confidence intervals, and distributions for any statistic. The fundamental assumption of bootstrapping is that the observed data are representative of the underlying population. By resampling observations from the observed data, the process of sampling observations from the population is mimicked. The key techniques are

- **THE BOOTSTRAP:** In bootstrap resampling, B new samples, each of the same size as the observed data, are drawn with replacement from the observed data. The statistic is first calculated using the observed data and then recalculated using each of the new samples, yielding a bootstrap distribution. The resulting replicates are used to calculate the bootstrap estimates of bias, mean, and standard error for the statistic.

- **THE JACKKNIFE:** In **jackknife** resampling, a statistic is calculated for the n possible samples of size $n-1$, each with one observation left out. The default sample size is $n-1$, but more than one observation may be removed. Jackknife estimates of bias, mean, and standard error are available and are calculated differently than the equivalent bootstrap statistics.

Using the bootstrap and jackknife procedures, all informative summaries (mean, variance, quantiles etc) for the sample-based estimates' sampling distribution can be approximated.

This is vitally important if we want to compute **measures of uncertainty** (standard errors, confidence intervals) for parameters in the model, or statistics.