PSI:
Relationships Between Variables (2)

Technological University Dublin City Campus

School of Computer Science

t-test for comparing two groups

[a numeric variable and a dichotomous categorical variable]

- We have already encountered the t-test, in the section on hypothesis tests, where it was used to decide if a population had a hypothesised parameter value (one-sample t-test). Now we look at how the t-distribution is used to decide if the difference of means for two groups of observations is statistically significant (i.e. not likely to be due to statistical variation). This is called a two-sample t-test.
- Defined by William Sealy Gosset (1876-1937), who had to publish the t-test under the pseudonym Student, because of rules imposed by his employer, the Guinness Brewery

- There are two types of t-test:
  - **with paired samples**: applicable when the values in the two groups of observations are closely connected, instance by instance (e.g. one group of observations are blood pressure measurements for a group of patients before treatment and the other group of observations are blood pressure measurements for *the same group of patients* after treatment)
  - **with unpaired samples**: applicable when the grouped values come from different instances, grouped by some attribute (e.g. one group of observations are blood pressure measurements for the patients in hospital A and the other group are blood pressure measurements for patients in hospital B)

Unpaired two-sample t-test

- Assumptions
  - One variable is numeric (at least interval scale), the other binary.
  - Either the samples are large (> 30) or the data are normally distributed.
  - The number of values in the two groups does not have to be the same, although equal group sizes allow the use of Student's t-test (rather than Welch's test) regardless of the difference between standard deviations.

- As with all hypothesis tests, the t-test uses a statistic that is in fact a standardised residual* quantity:

$$\frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_{E(\text{difference of means})}} = \frac{\bar{x}_1 - \bar{x}_2}{S_{E(\text{difference of means})}}$$

  where x is the numeric variable, $\bar{x}_1$ and $\bar{x}_2$ are the sample means for the two groups defined by instance membership of the two categories, $\mu_1 - \mu_2$ is the hypothesised difference between the means (equal to 0) and $S_{E(\text{difference of means})}$ is the standard error of the difference of means.

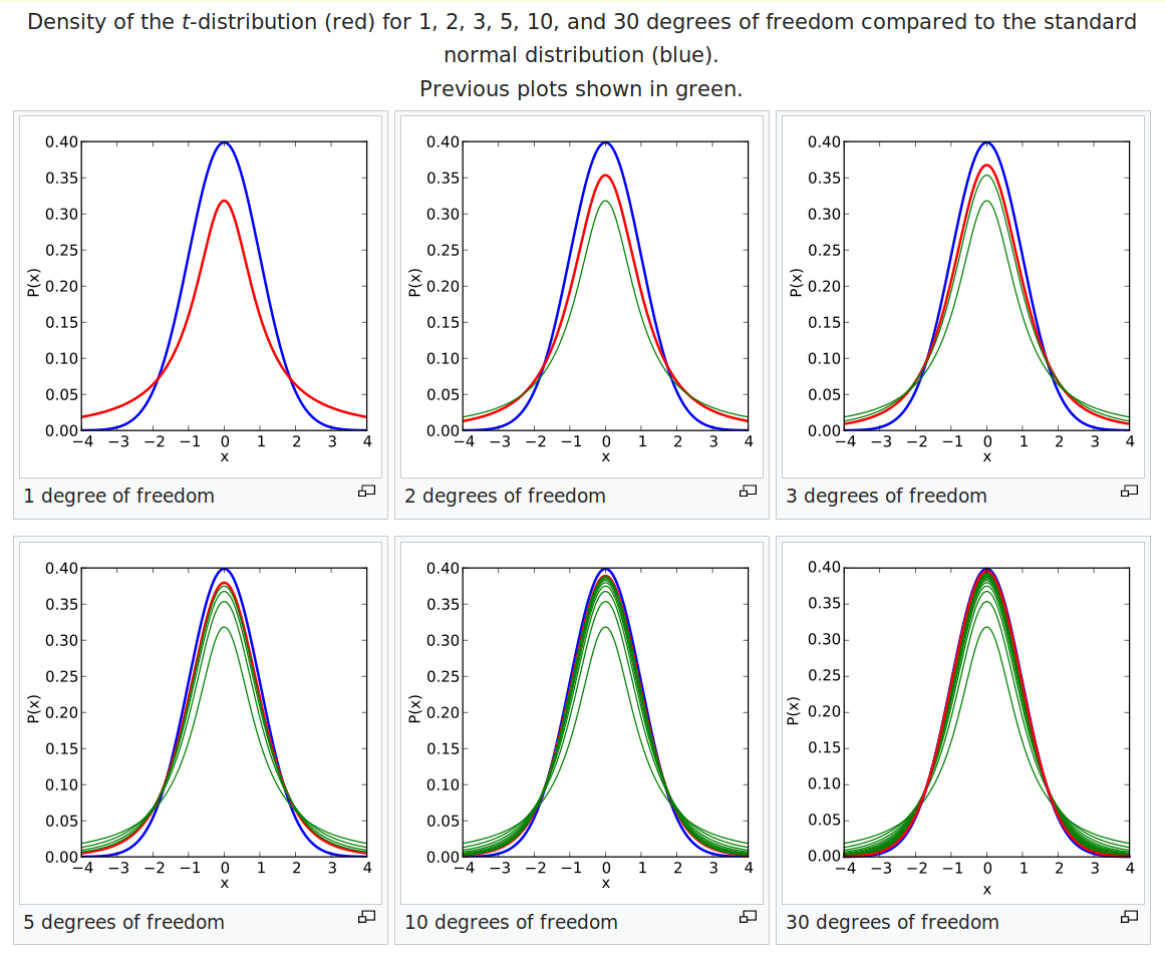
* Residuals represent the difference between experimentally obtained values (sample statistics) and hypothesised values (i.e. model values).

- Different formulae are used for the T-statistic, depending on whether the standard deviations differ a lot.

	Student's t-test: Standard deviations differ by a factor of 2 or less $\left(\frac{1}{2} \leq \frac{S_1}{S_2} \leq 2\right)$	**Welch's test:** Standard deviations differ by more than a factor of 2 $\left(\frac{S_1}{S_2} < \frac{1}{2} \quad \text{OR} \quad \frac{S_1}{S_2} > 2\right)$
SD ratio range		
T-statistic	$T = \frac{\bar{x}_1 - \bar{x}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}$	$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Degrees of freedom (df)	$df = n_1 + n_2 - 2$	$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$

In the table S_1^2 and S_2^2 are group variances, $\bar{x}_1$ and $\bar{x}_2$ are the group means, n_1 and n_2 are group observation counts.

- The t-distribution is not a single probability density function (PDF) but a series of distributions corresponding to different *degrees of freedom*.



Source: Wikipedia

- Once the correct t-distribution is identified based on the degrees of freedom, the T statistic value is tested in the same way as any other statistic, with respect to a certain level of significance (e.g. 5% or 1%).
- Significance: t-statistic critical value table for lookup can be found here.
- Effect size (see magnitude ranges):

Cohen's d

$$d = \frac{2t}{\sqrt{df}}$$

η^2 (eta squared)

$$\eta^2 = \frac{t^2}{t^2 + df}$$

- T-TEST AS A TEST OF RELATEDNESS

  The t-test can be viewed as a **test for whether there is a relationship between a numeric variable (that for which the values are grouped) and a categorical, grouping variable**.

  For example, if we are testing the difference of means between blood pressure measured for patients of hospital A on the one hand and the blood pressure measured for patients of hospital B, the hospital (A or B) is the categorical variable that groups the numeric blood pressure values. If it is found that the difference between the means is statistically different, then which hospital the measurement is taken in can tell us something about the value of blood pressure that we can expect for new patients and vice versa - this equates to a correlation.

Paired two-sample t-test

- Assumptions:

  These are the same as for the unpaired test, with the additional assumption that each value in the first group has a *pair* value in the second group, in that they are both associated with the same subject or instance, but based on measurements taken at different times, typically after a *treatment*. This implies that the two groups are of the same size.

- Test statistic:

$$t = \frac{\bar{D}}{\frac{S_D}{\sqrt{n}}}$$

  where $\bar{D}$ is the mean of the sample value pair differences, S_D is the sample value pair difference standard deviation and n is the number of value pairs in the sample

  Being calculated on the standard error of the mean of the value pair difference as variable (rather than the standard error of the difference of means as is the case with the unpaired test), this test has better *power* than the unpaired test.

- significance and effect size are derived in the same way as for the unpaired t-test

Non-parametric alternatives to the t-test

[a numeric variable and a dichotomous categorical variable]

Man-Whitney (or Wilcoxon's rank sum) test

- These are used as a non-parametric alternative to the unpaired two sample t-test when the assumption of normality is not met.
- As we have seen, heteroskedacity can be handled by Welch's variant of the t-test
- The two named tests differ in details but their methods are equivalent.
- Idea:
  - data belonging to both groups are ranked in a single sequence
  - the ranks for each group are summed up
  - if the data are centred similarly, the sums would be similar (provided an adjustment is made for group size)
  - if there is a difference between how the data are centred, the sums are different

  - distributions for these sum statistics in the case of similarity ('uninteresting' case) are available (having been compiled using Monte Carlo methods i.e. with the use of random numbers for values)
- significance is derived from these approximately normal distributions
- effect size can be expressed using Rosenthal's r:

$$r = \frac{z}{\sqrt{N}}$$

For magnitude interpretation see here.

Wilcoxon's signed rank test

- This test is a non-parametric alternative to the paired two sample t-test.
- Similarly to the Man-Whitney test, it has an approximately normal distribution and its effect size can be expressed using Rosenthal's r (see previous page)

ANOVA

[a numeric variable and a categorical variable with $>$ 2 values]

- Tests the variance of *three or more groups of observations* for whether there is a significant difference between their means (i.e. probably not due to normal variation in the samples) - it is like the t-test but is performed on 3 or more groups of numeric values
- The name stands for *completely randomized one-way analysis of variance*
- A hypothesis test, where the null hypothesis is that the means of the groups are equal
- Central to ANOVA is a number called the F statistic, which is essentially the ratio between inter-group variation and intra-group variation
- Assumptions:
  - data in the groups are independent
  - if the group sizes are different then also
    - the group populations must have similar variances (homoskedasticity)
    - the distribution of each group must be normal (without significant skewness or kurtosis)

- Calculation of the F-statistic:

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1} = \frac{SSB}{k-1}$$

$$MSW = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{N-k} = \frac{SSW}{N-k}$$

$$F = \frac{MSB}{MSW}$$

  where k is the number of groups, N is the overall number of observations, n_i is the number of observations in group i and s_i is the standard deviation within group i, $\bar{x}_i$ is the mean within group i and $\bar{\bar{x}}$ is the overall mean; SSB, SSW, MSB, MSW are the sums of squares and mean squares between and within groups.

- The F-distribution depends on two **degrees of freedom** values, are used in looking up the F-table:

  $df_B = k-1, \quad df_W = N-k$

  Can you tell where these numbers come from? We have had similar examples earlier in the module.

- The F-statistic differs from the t and z statistics in that it is a ratio of variances rather than a ratio in the variable dimension (i.e. residual/standard error). For the case of two group comparison, $F = t^2$, and is in essence the same test, albeit using a different statistic.

- Significance: F-statistic sampling distribution for the correct degrees of freedom (w.r.t. sample size and number of groups) should be used. A file containing an F-table can be found here.

- Effect size (for more information see here):

r^2

$$r^2 = \eta^2 = \frac{SSB}{SST} = \frac{SSB}{SSB + SSW}$$

Somewhat biased.

ω^2

$$\omega^2 = \frac{SSB - df_W MSW}{SSB + SSW + MSW}$$

Less biased.

- **Comparison of groups two-by-two**
  - One of the reasons for performing ANOVA in the first place is to contain the type I errors, which would accummulate if set at a fixed value (e.g. 0.05) for individual t-tests perform between each pair of groups
  - However, this can be avoided **planning a sequence of t-tests** and by **post-hoc testing**
  - **Planned t-tests** involve progressive two-way splits of the data (along group boundaries) and using a t-test at each split. For example:
    * we are comparing groups G1, G2, G3 and G4
    * we first want to compare G1 against the other groups
    * we pick weights for the groups so that G1's weight is balanced out by the others and form a new variable V1 $=$ 3*G1 - G2 - G3 - G4, which is tested with a t-test with $H_0 : \bar{V1} = 0$
    * the remaining three groups are tested similarly, now leaving out G1: V2 $=$ 2*G2 - G3 - G4, with t-test for $H_0 : \bar{V2} = 0$

* and so on

– **Post-hoc testing** involves performing t-tests between all pairs of groups, finding the p-values and then deciding significance based on one of several patterns devised to contain the type I error

Table 10.7 Critical values for p based on variations on Bonferroni (* indicates that a comparison is significant)

		Bonferroni			Holm			Benjamini–Hochberg	
	p	$p_{crit}=\frac{\alpha}{k}$		j	$p_{crit}=\frac{\alpha}{j}$		j	$p_{crit}=\left(\frac{j}{k}\right)\alpha$	
NT–Super	.0000	.0083	*	6	.0083	*	1	.0083	*
Super–Hulk	.0014	.0083	*	5	.0100	*	2	.0167	*
Spider–Super	.0127	.0083		4	.0125		3	.0250	*
NT–Spider	.0252	.0083		3	.0167		4	.0333	*
NT–Hulk	.1704	.0083		2	.0250		5	.0417	
Spider–Hulk	.3431	.0083		1	.0500		6	.0500	

Source: DSUR

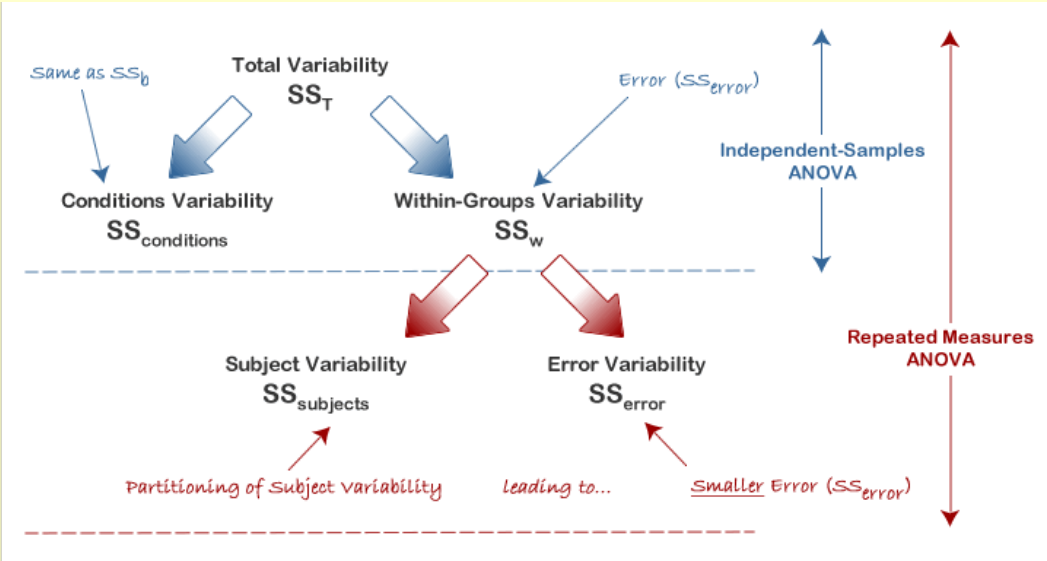
- ANOVA AS A TEST OF CORRELATION

  As with the t-test, ANOVA can be viewed as **a test for correlation between the numeric variable that contributes the values and a categorical variable that groups the values**.

  For example, the test could be performed on the heights of trees sampled from three different forests. In this case the forest (with values e.g. *North Forest*, *Big Forest*, *Old Forest*) is the categorical variable and the tree height is the numeric variable.

ANOVA with repeated measures

- Similarly to paired t-tests, ANOVA with repeated measures operates on groups of data with strictly related values by subject, having one value per subject in each group
- In general, this test has greater power than plain one-way ANOVA, resulting from the fact that the effect we are trying to model (the difference between groups) can be compared against a more focused variability that excludes differences between subjects (this could not be isolated in the general case ANOVA, where we did not know of any connections between the subjects from group to group)

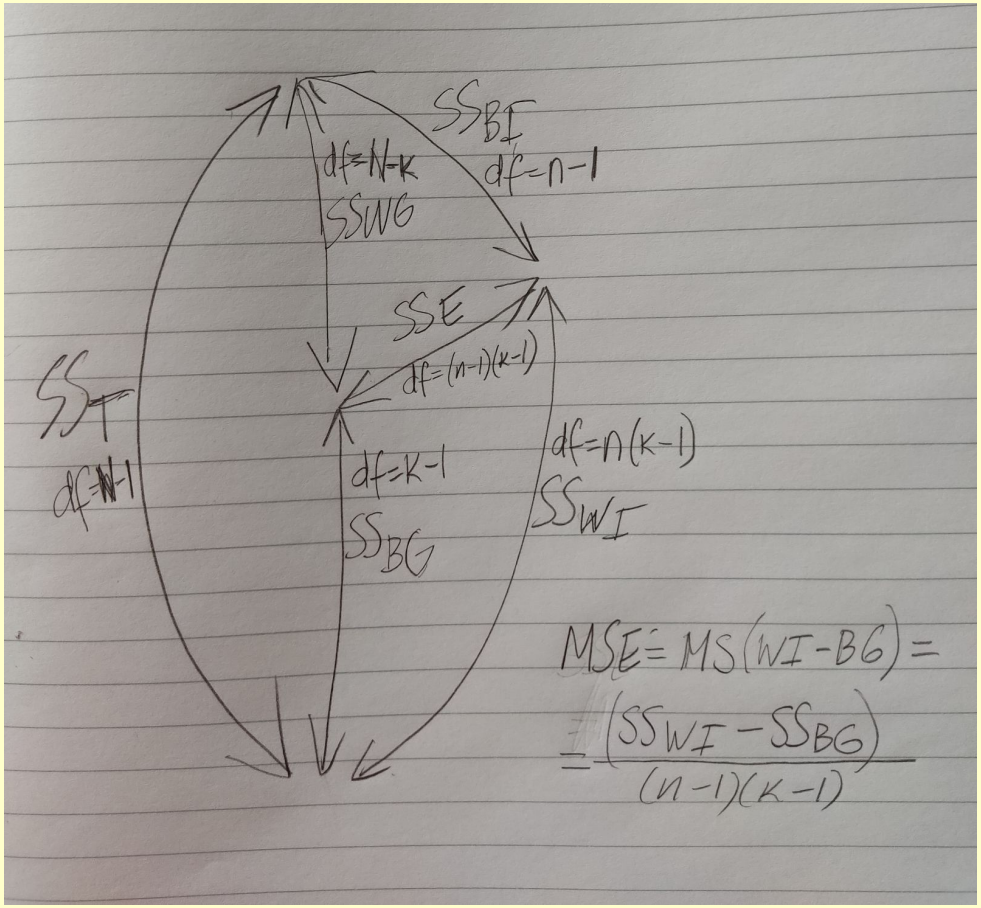


Source: laerd.com

- The statistic in this case is:

$$F = \frac{MSBG}{MSE} = \frac{MS(BG)}{MS(WI - BG)}$$

B - between; W - within; G - groups; I - individuals



Chi-squared

- Tests the independence of categorical variables (on the nominal or ordinal scale)
- Chi-squared can also be used to test goodnes of fit for a distribution (but we are not looking at that here)
- A hypothesis test where the hypothesis is that there is no relationship between the variables
- The statistic on which the hypothesis test is based is:

$$\chi^2 = \sum_{i=1,j=1}^{k_R,k_C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where k is the number of cells (categories) in the contingency table for the two variables, O_i is the observed frequency in cell i and E_i is the expected frequency for cell i. The expected cell frequency is calculated as:

$$E_{ij} = \frac{n_{Ri} \times n_{Cj}}{n}$$

where E_{ij} is the expected frequency for cell in row i and column j, n_{Ri} is the sum of frequencies in row i, n_{Cj} is the sum of frequencies in column j and n is the sum of frequencies across the entire table.

- The calculated χ^2 value is compared with the critical value for the required confidence level and number of degrees of freedom ($df = (r-1) \times (c-1)$, where r and c are the number of rows and columns, respectively, in the contingency table) from the standard chi-squared table. If the calculated χ^2 is greater than the critical value, the null hypothesis is rejected and it is taken that *there is a relationship* between the categorical variables.
- A Chi-squared table of critical values is shown on the next page.

Table of Chi-squared critical values

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

HOWTO: Calculating Chi-squared

Assumptions: We have the contingency table for two categorical variables.

Example scenario: We want to calculate Chi-squared for variables `car colour` and `accident in 10 years`, with contingency table shown on the right.

car colour → / accident ↓	silver/white	other
YES	22	27
NO	222	243

1) Calculate the row, column and overall counts

car colour→ / accident ↓	silver or white	other	ROW TOTALS
YES	22	27	$n_{r1} = 49$
NO	222	243	$n_{r2} = 465$
COLUMN TOTALS	$n_{c1} = 244$	$n_{c2} = 270$	$n = 514$

2) Using the formula for expected frequency ($E_{ij} = \frac{n_{Ri} \times n_{Cj}}{n}$), populate an expected frequencies table

For example: $E_{12} = \frac{49 \times 270}{514} = 25.74$

car colour → / accident ↓	silver or white	other
YES	23.26	25.74
NO	220.74	244.26

3) Calculate the term of the Chi-squared sum for each cell and place in a table; the formula is $T_{ij} = \frac{(E_{ij} - O_{ij})^2}{O_{ij}}$, where O_{ij} is the observed value for cell in row i and column j

For example: $T_{12} = \frac{(25.74 - 27)^2}{27} = 0.059$

car colour → / accident ↓	silver or white	other
YES	0.072	0.059
NO	0.007	0.007

4) Calculate Chi-squared by adding all the terms:

$$\chi^2 = \sum_{i=1,j=1}^{k_R,k_C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1,j=1}^{k_R,k_C} T_{ij}$$

$$\chi^2 = 0.072 + 0.059 + 0.07 + 0.07 = 0.145$$

This value can now be looked up in the Chi-squared table of critical values.

References Some pictures in this presentation were taken from the following books. The source for each picture is cited beside it.

[MSD] *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, by Glenn J. Myatt and Wayne P. Johnson, John Wiley & Sons, 2014.