Index of 12-L

Practical Information

Home assignments:

- no. III returned today: solution posted and brief review planned,
- no. IV at webpage, and due in one week (April 4);
  - don't wait too long to start looking for articles...

Today's lecture:

- regression (last parts):[1]
  - residual analysis, with Minitab demonstrations,
  - "warnings" and extensions (suppl. notes: Moodle),
- correlation (full review):[1]
  - correlation coefficent and statistical inference for correlation,
- links between models/procedures for correlation and regression,
- Bayesian statistics:
  - for your information and not part of the syllabus,
  - not in textbook: use as reference the lecture notes + supplementary notes (at webpage).

Lab session: tomorrow 1-4pm, due to Easter holidays.

[1] See lecture 11 for textbook references.

Tools for Model Checking: Residuals

Residuals:

- = "estimates" of random variables ε_i in model,
- calculated as "observed − expected", e.g.,
  - linear regression: $\hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$,
  - 1-way ANOVA: $\hat{\varepsilon}_{ij} = X_{ij} - \bar{X}_i$,
- always: SSE = sum of squared residuals,
- properties of residuals if the model is correct:
  - normally distributed with mean 0 and a computable standard error (may be the same for all residuals, depending on design),
  - residuals are *not independent*.

Other versions/variables:

- *standardized residuals* (i.e., divided by their standard error): $r_i = \hat{\varepsilon}_i / \mathrm{SE}(\hat{\varepsilon}_i)$, approximately distributed as $\mathrm{N}(0, 1)$,
- *(advanced) deletion residuals*: predicted value from model *without* current observation, also standardized,[2]
- *(advanced) influence statistics*: special statistics to assess the impact of a single observation on the fitted regression line,[3]
  - idea: it may be "problematic" if estimates or conclusion depends strongly on a single or a few observation(s).[4]

[2] Deletion residuals can be used for formal outlier tests (VHM 802/812).

[3] Several different statistics (leverage, Cook's distance, DF(F)ITS) with their specific interpretations, but beyond this course to go into details with them.

[4] The best way to assess if a particular observation is influential, is to analyze the data with and without this observation, and compare the results.

Model Checking in Regression/ANOVA

Proposed use of residuals for model checking:

- variance homogeneity:
  plot residuals ($\hat{\varepsilon}_i$ or r_i) against model's fitted values ($\hat{y}_i$):
  — should get a noisy pattern with no "fan" shapes,
- linear relation:
  plot residuals ($\hat{\varepsilon}_i$ or r_i) against explanatory variable x_i:[5]
  — should get a noisy pattern with no "parabolic" shapes,
- outliers:
  check very large or small values of stand. residuals (r_i):
  — extreme r_i-values can be assessed (approximately) in $N(0,1)$:
  - values outside (-2,2) "suspect" in a small dataset,
  - values outside (-3.5,3.5) "suspect" in moderate-sized dataset,
- normal distribution:
  normal probability plot of stand. residuals (r_i),[6]
- data errors:
  plot residuals ($\hat{\varepsilon}_i$ or r_i) against data order (if applicable).

"Unusual observations" (in Minitab listing):

- stand. residuals beyond (-2,2) (indicated with R),
- high *leverage* values (indicated with X): extreme among the (x_i) values, so the observation is *potentially influential*.

[5] In simple linear regression, plots of the residuals against x_i and $\hat{y}_i$ are practically the same (so one of them will suffice).

[6] Note that P-values for normality tests only apply approximately to residuals, of any type, because of their dependence.

Correlation

Correlation, usually denoted by ρ ("rho"),
= a *parameter/property* of a two-dimensional, continuous distribution (simultaneous distribution of two quantitative variables), expressing the strength and direction of *linear* association between them in their population.

Sample correlation coefficient: $r = \frac{1}{n-1} \sum_i \left(\frac{X_i - \bar{X}}{s_x} \right) \left(\frac{Y_i - \bar{Y}}{s_y} \right)$
= a descriptive statistic for a sample of pairs of variables (quantitative, response variables), and an estimate of the population correlation ρ: $\hat{\rho} = r$. (*Pearson* correlation coef.)

Properties of correlations (both types of correlations):

- $-1 \leq r \leq 1$,
- $\left. \begin{array}{c} r > 0 \\ r = 0 \\ r < 0 \end{array} \right\} \sim \left\{ \begin{array}{c} \text{positive} \\ \text{no} \\ \text{negative} \end{array} \right\}$ linear association,
- $r = -1$ and $r = 1$ correspond to perfect linear association (all points on a straight non-horizontal/-vertical line),
- correlation between X and Y is same as betw. Y and X,
- correlation defined from standardized variables $\Rightarrow$ unaffected by changes in mean or standard deviation,
- X and Y independent variables $\Rightarrow \rho = 0$ (and $r \approx 0$),
- extended variance addition formulae (IPS Section 4.4):

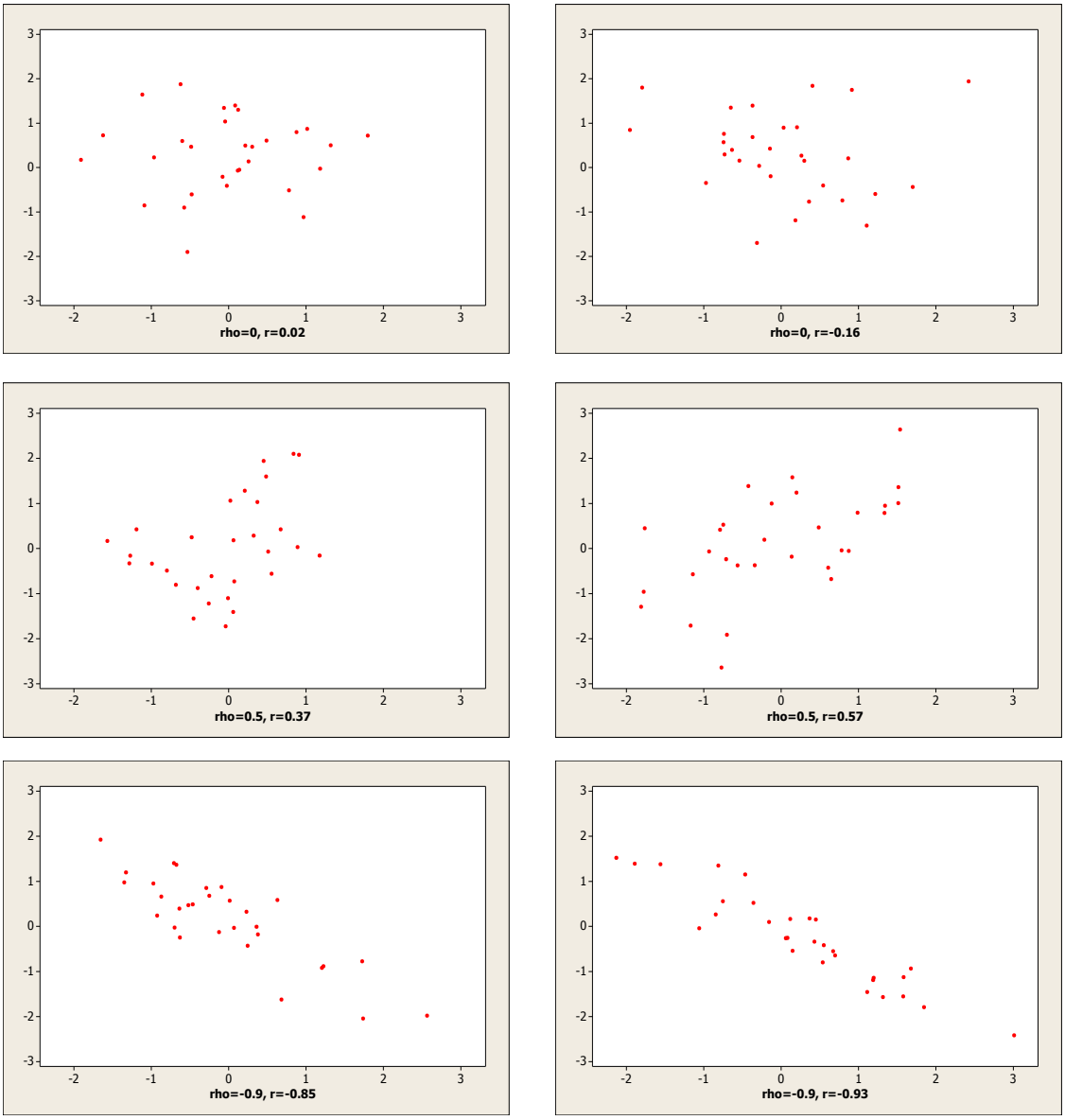
$$\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \pm 2\rho\, \mathrm{sd}(X)\mathrm{sd}(Y).$$

Correlation II

Some cautions about correlation:

- only meaningful for *quantitative*, *response* variables,
- only meaningful for roughly *linear associations*,
- the sample correlation coefficient r is not resistant.

Simulated patterns ($n = 30$):[7]



[7] See also PSLS 3e: Figure 3.5; S: p. 173; IPS 7e: Figure 2.16.

Statistical Inference for Correlation

Definition: A pair of variables (X, Y) follows a *joint normal distribution* $\mathrm{N}(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$, if

- $X \sim \mathrm{N}(\mu_x, \sigma_x)$, and $Y \sim \mathrm{N}(\mu_y, \sigma_y)$,
- and the correlation between X and Y is ρ.[8]

Statistical inference for correlation:

- feasible only based on an i.i.d. sample (or SRS) $(X_1, Y_1), \ldots, (X_n, Y_n)$ from $\mathrm{N}(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$,
- in this model (only!): $\rho = 0 \Rightarrow X$ and Y independent, and regressions of Y on X (or reversely X on Y) have slope 0,
- Hypothesis H_0: $\rho = 0$ can be tested against a one- or two-sided alternative H_a by the t-statistic,

$$t = r\sqrt{\frac{n-2}{1-r^2}}\,, \qquad \text{and } t \sim t(n-2) \text{ under } \mathrm{H}_0,$$

  which is *exactly the same!!* as the t-test for slope $= 0$ in any of the two regressions,
- standard errors for r, and tests for $\rho =$ known value: difficult to calculate, and not easily accessible,
- nonparametric correlation: *Spearman's* rank correlation coefficient (i.e., r computed for ranks) $\rightarrow$ lab problem.

[8] and the conditional distributions of X given Y are normal distributions, and the same for Y given X.

Correlation vs. Regression

Correlation and least-squares regression very closely related:

- r = slope of least-squares regression line when both variables measured in standardized units ($\hat{\beta}_1 = r s_y / s_x$),
- test for $\rho = 0$ in jointly normal model is same as test for slope $= 0$ in the two conditional regressions,
- in the ANOVA table for linear regression: $r^2 = \text{SSM}/\text{SST}$ $\Rightarrow$ r^2 interpretable as *the proportion of variation explained by the regression, out of the total variation*:
  - r^2 large means good predictive power of the model,
  - r^2 large does not necessarily mean a good model,[9]
  - r^2 (usually denoted R^2) is widely misused to indicate the model's "quality".

How to choose between correlation and regression?
look at which model assumptions are most reasonable:

- normal distribution for pairs (X, Y) (correlation), or
- normal distribution for errors in linear regression,

for example,

- only one response variable: always regression,
- two response variables: interest in predicting one from the other (regression)? — or primarily measure/test their degree of linear association (correlation)?

[9] For an illustration, see Extra exercise 20 (x:20).

Last Comments about Correlation and Regression

Some cautionary points from textbooks[10]:

- Linear regression and correlation are based on *linear relationships* — always check if that is reasonable,
  note: some non-linear relations[11] can be transformed to linear ones and analyzed by a linear regression for the transformed variables, e.g.,

$$\begin{aligned} y = a \times x^b &\longrightarrow \log(y) = \log(a) + b \times \log(x), \\ y = a \times b^x &\longrightarrow \log(y) = \log(a) + \log(b) \times x, \\ y = \tfrac{a}{1+b\times x} &\longrightarrow 1/y = 1/a + (b/a) \times x. \end{aligned}$$

- Watch out for *outliers and influential observations*,
- Regression or correlation $\not\Rightarrow$ *causation*,
- *Lurking variables* can distort any relationship between variables (not new, but particularly important here),
- Beware of *extrapolation* (too far outside the data range),
- *Averaging and/or restricted range* affect estimation of correlation and linear regression.

Variants (extensions) of linear regression:

- multiple linear regression: more than one x-variable in model; textbooks[12] (and VHM 802/812), generalization of simple linear regression, and not impossible for you to do on your own,
- measurement error models: if x is a *response variable* and measured with (considerable) *error/variation/noise*, and interest is in linear regression on *true* value of x (without error), not prediction.

[10] PSLS 3e: Chapter 4; IPS 7e: Section 2.4.

[11] Transformation to achieve a linear relation is discussed in supplemental material for IPS 6e (from IPS6e website (discontinued), available at Moodle site).

[12] PSLS 3e: Chapter 28 (Suppl.); S: Section 10.4; IPS 7e: Chapter 11.

Introduction to Bayesian Approach

Two rivaling schools of statistics exist:

- Classical (likelihood-based, frequentist)
- Bayesian[13]

Statisticians often are strongly in favour of one of them, or have a more pragmatic view (using whatever works for the data at hand).

Fundamentally different concepts:

- probability (frequency interpretation vs. Bayesian subjective probabilities[14]),
- parameters in statistical models (next slide),

$\Rightarrow$ different statistical inferences.

Applications of Bayesian methods:

- getting more common in veterinary science/epidemiology,
- vast majority of statistics courses taughts are on classical methods,
- nowadays usually based on MCMC estimation,
  - simulation-based method using specialised software,
  - involves new issues related to model checking.

[13] After Reverend Thomas Bayes, 1702−1761, who "invented" Bayes' theorem.

[14] Unique to each person, also valid for events with no frequency interpretation, e.g. "there is at least one typo in these notes"; see also 3L−3.

Bayesian Approach

One-line characterisation of Bayesian modelling:

> "parameters are random quantities associated with probability distributions".[15]

Rough outline of Bayesian scientific approach:

1) Pose a question in terms of a parameter θ,
2) Based on existing knowledge and/or own belief, set up a *prior distribution* for θ,
3) Carry out an experiment to collect (additional) information about θ,
4) Combine the prior distribution and the data into a *posterior distribution* reflecting the knowledge about θ after the experiment.

Classical versus Bayesian approach:

Concept	Classical approach	Bayesian approach
parameter	constant	distribution
prior information on parameters	none	prior distribution
base of inference	likelihood function	posterior distribution
parameter value	(ML) estimate	statistic of posterior, e.g.: median, mode, mean
parameter range	confidence interval	prob. range of posterior
hypothesis statement	test	(Bayesian factors/P-values)

[15] In classical statistical models, parameters are unknown constants.

How it Works: Bayes' Formula in Action

Bayes' formula (see 3L−10) for events A and B:

$$P(A|B) = P(B|A) \cdot P(A)/P(B).$$

Also for probability functions and densities f:

$$f(X|Y) = f(Y|X) \cdot f(X)/f(Y).$$

Bayesian framework:

- $f(Y|\theta)$ — the likelihood function (often denoted $L(Y|\theta)$), giving the probability or density of the observed data Y when the parameter takes value θ,
- $f(\theta)$ — the prior distribution for θ,
- $f(\theta|Y)$ — the posterior distribution for θ,

With these definitions, Bayes' formula gives:

$$f(\theta|Y) = f(Y|\theta) \cdot f(\theta)/f(Y) = const(Y)\, f(Y|\theta)\, f(\theta),$$

where $const(Y)$ means a variable depending on Y but not on θ.

Conclusion: The posterior distribution (for θ) is essentially obtained by multiplying the prior distribution with the likelihood function.

Example: Inference about a Proportion

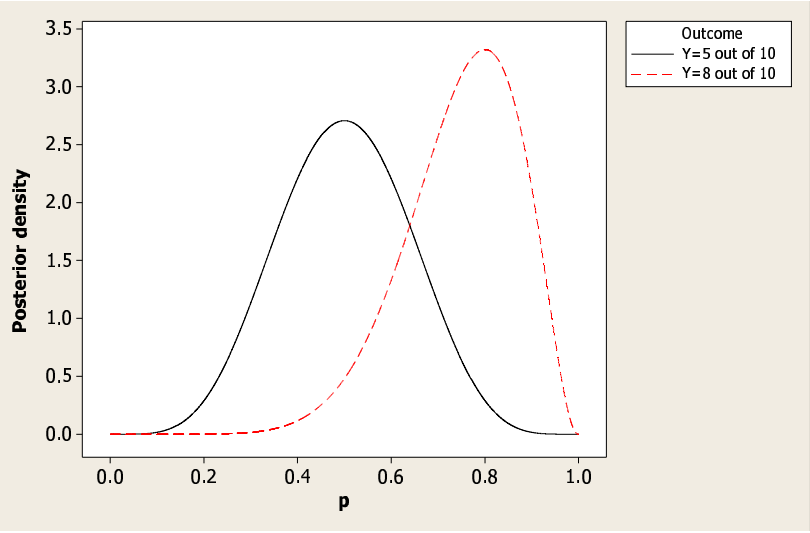
Consider a population of animals and a disease either present or absent in the animals; assume a binomial setting. If $Y=5$ or $Y=8$ animals out of 10 tested positive:

Classical answer:

- 5 out of 10: $\hat{p}=0.5$, 95% CI (plus 4) for p: (0.238,0.762),
- 8 out of 10: $\hat{p}=0.8$, 95% CI (plus 4) for p: (0.478,0.951),

Bayesian approach:

- choose a prior: uniform distribution on (0,1), (just one possibility — a "non-informative" prior),
- combine prior and data (Y out of 10 positives) into posterior: beta-distribution[16] $(Y+1, 10-Y+1)$, see figure:



- posterior "estimates" and intervals (2.5%–97.5% range):
  - 5 out of 10: median=0.5, interval: (0.234,0.766),
  - 8 out of 10: median=0.764, interval: (0.482,0.940).

[16] Beta-distributions (a, b) are distributions on (0,1) with parameters a and b.

Choice of Prior Distribution

— the cornerstone as well as the strength and weakness of the Bayesian approach; some possibilities:

- subjective prior: the researcher's personal belief,
- non-informative (vague, flat) prior: giving no (or only very weak) preference to a particular θ-value; commonly used despite some philosophical problems, will often lead to results similar to classical statistics,
- convenient prior: to allow simple formulae for posterior distributions in specific models (less used after the invention of MCMC methods; e.g. beta-distribution for p),
- posterior from previous experiment: allows for successive updates of information for every new experiment,
- structural prior: to specify certain regularities in complex data structures (image analysis, genetics),
- expert opinion: quantification of expert opinions obtained in questionnaires or by round-table discussion.

Personal views:

- In some situations (e.g., clinical trials, diagnostic testing), it is very appealing to include prior information; the question is how to quantify it in an "objective" way.
- In many simple situations, the prior distribution is rather unnatural and only complicates matters.

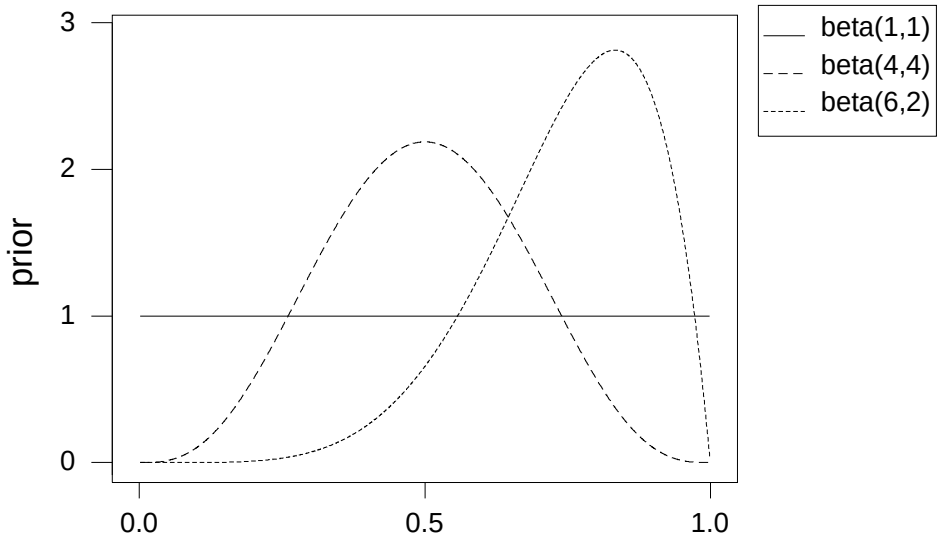
Proportion Example: Informative Priors

Consider again testing for presence of disease, and observing $Y = 5$ positives out of $n = 10$. Some possible priors for p:

1) uniform distribution on (0,1) — same as: beta-distr.(1,1),

2) beta-distr.(4,4) — symmetrical around 0.5,

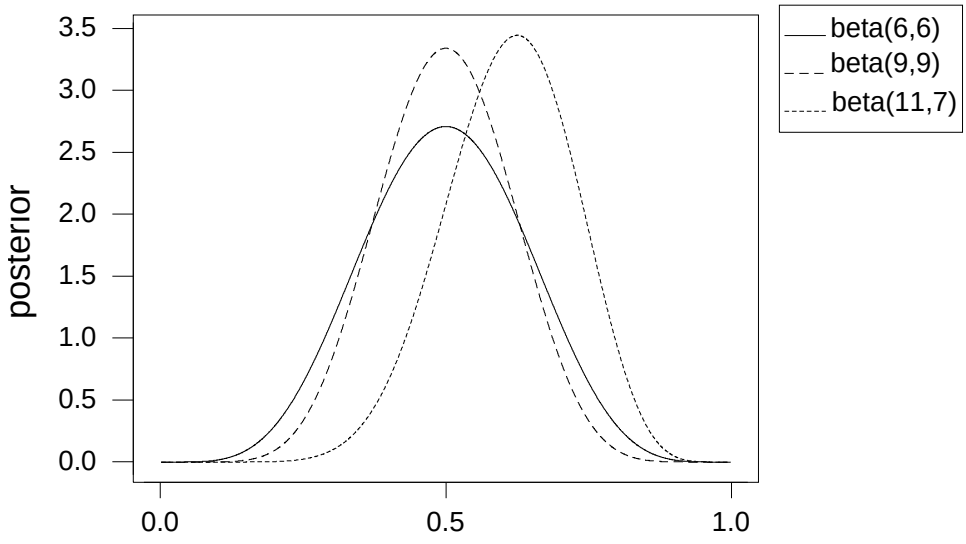
3) beta-distr.(6,2) — left skewed with mean 0.75.



Posterior distributions:

1) beta-distr.(6,6); 2) beta-distr.(9,9); 3) beta-distr.(11,7);

$*$ rule: beta-distr.(a, b) & B$(n, Y) \Rightarrow$ beta-distr.$(a + Y, b + n - Y)$.

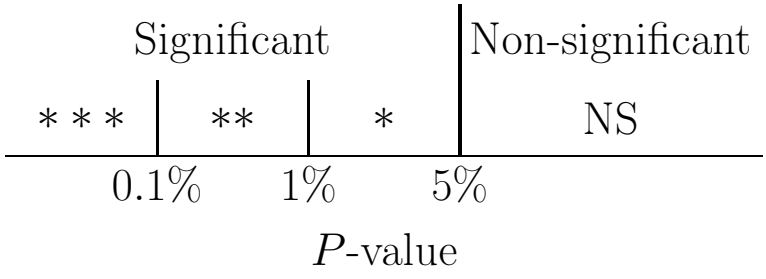


Home Assignment III

- generally: some deduction for omitting conclusions or models/ assumptions/conditions,
- Question 1: only use "plus four" calculation for CI, not for estimation (or test),
- Question 2 was answered incorrectly by everyone $\Rightarrow$ excluded from final score (except where close to correct!),
  - the difference to Q1 is its focus on persons instead of pairs,
  - 45 "single" and 8 "double" pairs $\sim$ 61 potential twins, and $2 \cdot 8 = 16$ of these are alcoholics themselves $\Rightarrow$ desired $\hat{p} = 16/61$ (so "simple" and still so difficult),
  - termed the *probandwise concordance rate* (see solution),
- Question 3: a significant X^2-test must be further explored, and looking at the X^2-contributions leads to the genetic interpretation: monozygotic pairs are more similar (either none or both twins are alcoholics) than dizygotic pairs,
- Question 4: biggest problem was using the and independence assumption and estimated probability to compute probabilities for $0, 1, 2$ alcoholics in a pair — binomial $\mathrm{B}(2, p)$ distribution!; similar genetic interpretation where the data (for dizygotic twins) shows too many similar twin pairs for independence,
- model and/or hypotheses needed for chi-square analysis!
  - Question 3: model I most natural (comparison of two populations),
  - Question 4: hypothesis tested is independence between twin 1 and 2 (but not a standard two-way table setup).

REVIEW: P-VALUES

Significance as determined from P-value[17]:



Interpretation and suggested wording (personal preferences):

- P-values above 0.05 are usually denoted *non-significant* and interpreted as indicating that the result (observed value of the test statistic) could have occurred by coincidence if H_0 is true, so that H_0 cannot be rejected,
- P-values just around 0.05 (no matter if just above or below) are *borderline significant*, and give *indication* that H_0 may not be true,
- formally, a P-value below 0.05 is *evidence against* H_0,
- P-values at 0.01 or below are *clearly significant* and give clear indication that H_0 is false, so that H_0 should probably be rejected,
- P-values at 0.001 or below are *strongly significant* and show that the data are incompatible with H_0, so that H_0 without doubt should be rejected.

[17] Warning ! — do not confuse P-values and values for p's:

- P-values are values (certain probabilities) computed to interpret test results,
- p's are (probability) parameters in particular models, e.g. $B(n, p)$.

Summary Notes

Key words and concepts for 2 quantitative (continuous) variables:

- scatterplot, response and explanatory variable, dependent (y) and independent (x) variable,
- linear relation: intercept, slope, prediction, extrapolation, transformation of y and/or x,
- linear regression model: normally distributed, vertical errors about line, least squares estimation, standard deviation about line, t-based inference, ANOVA table, confidence and prediction intervals,
- model checking: residuals, standardized residuals, residual plots, outliers, variance homogeneity,
- correlation: population parameter/estimate (Pearson's correlation coefficient), strength of linear association, range (-1,1), independence, addition formula for variances,
- correlation model: normal distributions, t-test for no association, links with linear regression, squared correlation (r^2 or R^2).