%%%%%%%%%% cv4\_NV

L0: # Exercise 4 - Random vector

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Example from a collection

L3: Random vector $ Z = (Y; X) ^ T $ has a probability function specified by the table! [Image.png] (attachment: image.png)

L4: ## a) Determine the missing value of the combined probability function,

L5: possibly byrow = ...

L6: do not run this cell twice, otherwise you will set the value back to 0,

L7: Do you know why?

L8: ## b) Specify the distribution function

L9: \*\* Attention! The vector Z is $ (Y, X) ^ T $ so the first parameter is the value Y and the second value X. \*\*

L10: F (2.8; 7.1)

L11: = P (Y <2.8, X <7.1)

L12: we go through the rows and columns, we always take one value

L13: from the relevant row or column

L14: ## c) Determine the marginal distribution

L15: ## d) Conditional probabilities and conditional probability functions $ P (x | y), P (y | x) $

L16: P (Y> 2.1 | X <5.3)

L17: = P (Y> 2.1 ∧ X <5.3) / P (X <5.3)

L18: P (X = 5 | Y = 1)

L19: = P (X = 5 ∧ Y = 1) / P (Y = 1)

L20: \*\* $ P (x | y) = \ frac {P (X = x, Y = y)} {P\_Y (y)} $ \*\*

L21: it's the same size, so we'll steal the formatting

L22: \*\* $ P (y | x) $ \*\*

L23: it's the same size, so we'll steal the formatting

L24: ## e) basic characteristics of random variables X and Y

L25: ## f) conditional mean E (X | Y = 2)

L26: P (x | Y = 2)

L27: ## g) covariance and correlation

L28: matrix where in each column is the value x \* y

L29: mean value of E (X \* Y)

L30: covariance

L31: correlation

%%%%%%%%%% cv12-extra

L0: # Exercise 12. Multiselective tests - Extra for those interested ## Michal Béreš

L1: ## We load test data and produce post-hoc + effects for ANOVA and KW test

L2: data is in standard data format

L3: POST-HOC ANOVA

L4: ANOVA effects calculation

L5: overall average

L6: averages in groups

L7: effects

L8: list sorted

L9: POST-HOC KW

L10: post hoc - another function with an output that suits us better

L11: numerically corresponds to that used for the exercise

L12: install.packages ("FSA")

L13: FSA library

L14: KW effect calculation

L15: overall average

L16: group averages

L17: effects

L18: list sorted

L19: # For those interested (optional) - creation of a sorted table of p-values / letter scheme automatically

L20: install.packages ("strings")

L21: this is a text search library

L22: we'll look for smurf names in paired post-hoc tests

L23: we initialize the matrix (for a nice result as a text)

L24: 7x7 because we have 7 smurfs

L25: we name its columns and rows according to the sorted smudges

L26: loop through all tests in post-hoc (over column names)

L27: Which dwarves are present in this paired test?

L28: what are the indices of these dwarves

L29: indexes for writing to the matrix - always 2 values

L30: I write to the matrix (the first index is smaller -> automatically to

L31: upper triangle)

L32: we write text (if the matrix is text, the numbers are automatically

L33: convert to text), values to thousands

L34: ### Functions for automated sign scheme (handwritten and from package)

L35: #### Handwritten functions (what we would do on paper)

L36: table of p-values

L37: number of groups

L38: we name its columns and rows according to sorted types

L39: loop through all tests in post-hoc (over column names)

L40: Which dwarves are present in this paired test?

L41: what are the indices of these dwarves

L42: indexes for writing to the matrix - always 2 values

L43: I write to the matrix (the first index is smaller -> automatically to

L44: upper triangle)

L45: letter diagram from the table

L46: how big is the matrix

L47: matrix initialization

L48: line names - copy from input

L49: set the diagonal to 1 - is in the given group

L50: cycle through all columns where we can fill something

L51: cycle through all the rows in the column where we follow the gallop

L52: if pval> alpha then we add to hom. groups

L53: #### How to use handwritten functions for ANOVA and KW test?

L54: How to do it from POST-HOC ANOVA:

L55: we produce input data

L56: we produce a sorted phodnot table

L57: draw rounded to thousands

L58: we make a letter scheme from the phodnot table

L59: How to do it from POST-HOC KW:

L60: we produce input data

L61: we produce a sorted phodnot table

L62: draw rounded to thousands

L63: we make a letter diagram from the phodnot table

L64: ## Letter scheme using the built-in Rk function

L65: rcompanion package, cldList function

L66: in case of ANOVA

L67: first we make a dataframe with columns of pairs and phodnot

L68: letter scheme, library rcompanion

L69: install.packages ("rcompanion")

L70: in the case of KW

L71: first we make a dataframe with columns of pairs and phodnot

L72: letter scheme, library rcompanion

L73: install.packages ("rcompanion")

%%%%%%%%%% cv9

L0: # Exercise 9. Interval estimates (one selection) ## Michal Béreš, Martina Litschmannová, Veronika Kubíčková

L1: # Demonstration at the beginning - what is interval estimation?

L2: Consider a random variable from the normal distribution with a mean value of $ \ mu $ and a standard deviation of $ \ sigma $. We will work with selections from this random variable and help them try to estimate the mean value of the distribution (here we know its true value, but in practice its value is unknown).

L3: selection size

L4: mean value

L5: direction. deviation

L6: simulation of random selection from a given random variable

L7: sampling average as a point estimate

L8: selection direction. departure

L9: For clarity, we can visualize the selection.

L10: width of graphs in Jupyter

L11: graph matrix 1x2

L12: ### The construction of the interval estimation itself using a selection characteristic

L13: We will use this selection characteristic: (we assume that we do not know any real parameters of the distribution, only that it is normal) <br> $ Y = \ frac {\ bar X - \ mu} {S} \ sqrt {n} \ sim t\_ {n-1} $ <br> Since we know the distribution of Y, we are able to compute $ a $ a $ b $ in the following expression: <br> $ P (a <Y <b) \ geq 1 - \ alpha $ <br > - $ \ alpha $ is called the significance level (the probability that the searched value is outside our range) - $ 1- \ alpha $ is called the reliability of the interval estimate

L14: $ a $ a $ b $ are chosen so that they are symmetric in probability, ie: - $ P (Y <a) \ leq \ alpha / 2 \ rightarrow a = t \_ {\ alpha / 2; n-1} $ - $ P (b <Y) \ leq \ alpha / 2 \ rightarrow P (Y \ leq b) \ geq 1 - \ alpha / 2 \ rightarrow b = t\_ {1- \ alpha / 2; n-1} $

L15: maximum probability with which we allow

L16: real st. hours lay outside the constructed interval

L17: relevant quantiles of the student's distribution

L18: Next we just add to the expression and modify: <br> $ P (t \_ {\ alpha / 2; n-1} <\ frac {\ bar X - \ mu} {S} \ sqrt {n} <t\_ {1 - \ alpha / 2; n-1}) \ geq 1 - \ alpha $ <br> $ P (\ bar X - t\_ {1- \ alpha / 2; n-1} \ frac {S} {\ sqrt { n}} <\ mu <\ bar X - t \_ {\ alpha / 2; n-1} \ frac {S} {\ sqrt {n}}) \ geq 1 - \ alpha $ <br>

L19: This particular estimate can also be obtained using the Rkov function t.test:

L20: ### Testing interval estimation on multiple selections

L21: number of selections

L22: selection size

L23: mean value

L24: guide. deviation.

L25: significance level

L26: relevant quantiles of the student's distribution

L27: plot of the actual mean value

L28: cycle through individual selections

L29: select the plot color, depending on whether the IO contains the middle hour.

L30: draw the IC as a vertical line

L31: I will return the width to the standard size

L32: width of graphs in Jupyter

L33: # Types of interval estimates

L34: (Examples of estimating the mean value of data from a normal distribution.)

L35: ## Bottom / Left IC

L36: - $ P (M\_D ^ \* <\ mu) = 1- \ alpha $ - v Rku \*\* alternative = "greater" \*\*

L37: ## Top / Right IO

L38: - $ P (\ mu <M\_H ^ \*) = 1- \ alpha $ - v Rku \*\* alternative = "less" \*\*

L39: ## Double-sided IC

L40: - $ P (M\_D <\ mu <M\_H) = 1- \ alpha $ - v Rku \*\* alternative = "two.sided" \*\*

L41: # Overview of selection parameters and their point / interval estimates

L42: We usually have more IO constructions (functions in Rk that will do this for us), but each construction has different data requirements and produces different "quality" (in terms of IO size) estimates. We will always select the "best quality" IC that \*\* has met \*\* the prerequisites for use. <br> The order of the various ICs below will always be from "best" to most robust.

L43: ## Position measures of one selection

L44: By position measures we mean the data that determines the position of the data, no matter how scattered. For data from the normal distribution we can estimate the mean value, for others the median.

L45: #### a) student's IC t-test

L46: - we estimate the mean value - the point estimate is the sample average - the data must come from a normal distribution - exploratory: skewness and sharpness lie in (-2,2) - exploratory: the QQ graph has points approximately on the line - exactly: using a statistical test, eg Shapiro-Wilk test (shapiro.test (data))

L47: exploratory normality test

L48: library (moments) - we can avoid this by calling moments ::

L49: it's safer - we're sure we're calling a function from this package

L50: exact data normality test

L51: the resulting p-value must be greater than hl. challenge (eg 0.05)

L52: point estimate

L53: IO

L54: #### b) Wilcoxn IC test

L55: - we estimate the median - the point estimate is the sample median - the data must come from a symmetric distribution - exploratory: the skewn lies in (-2,2) - exploratory: the histogram looks approximately symmetrical - exact: using a statistical test, eg the "lawstat" package , function "symmetry.test (data, boot = FALSE)" - function in Rk requires additional parameter (conf.int = TRUE)

L56: exploratory

L57: exact: symmetry test

L58: install.packages ("lawstat")

L59: the resulting p-value must be greater than hl. challenge (eg 0.05)

L60: point estimate

L61: IO

L62: #### c) sign test IO test

L63: - we estimate the median - the point estimate is the selection median - the selection of a larger range (> 10) - the function in Rku requires an additional parameter (conf.int = TRUE) - it requires the "BSDA" library - as the most robust test, it can be used for discontinuous data - eg order in a list

L64: true median

L65: point estimate

L66: quantile (select, probs = 0.5)

L67: IO

L68: install.packages ("BSDA")

L69: ## Measures of variability of one selection

L70: By measures of variability we mean the data determining the scatter / variability of the data, regardless of the total values. For data from the normal distribution, we can estimate the standard deviation.

L71: #### IO standard deviations

L72: - we estimate the standard deviation - the point estimate is the sample standard deviation - the data must come from a normal distribution - exploitative: skewness and sharpness lie in (-2,2) - exploitative: the QQ graph has points approximately on the line - exactly: using a statistical test , eg Shapiro-Wilk test (shapiro.test (data)) - requires package "EnvStats" - function in Rku, gives calculation of variance - necessary square root of the result

L73: exploratory normality test

L74: exact data normality test

L75: the resulting p-value must be greater than hl. challenge (eg 0.05)

L76: point estimate

L77: IO

L78: install.packages ("EnvStats")

L79: We add a manual calculation: - we start from the statistics: $ \ frac {S ^ 2} {\ sigma ^ 2} (n-1) \ sim \ chi ^ 2\_ {n-1} $ - Upper limit: - $ P (\ frac {S ^ 2} {\ sigma ^ 2} (n-1) <\ chi ^ 2 \_ {\ alpha / 2, n-1}) = \ alpha / 2 $ - $ P (\ frac {S ^ 2} {\ chi ^ 2 \_ {\ alpha / 2, n-1}} (n-1) <\ sigma ^ 2) = \ alpha / 2 $ - Lower limit: - $ P (\ frac {S ^ 2} {\ sigma ^ 2} (n-1)> \ chi ^ 2\_ {1- \ alpha / 2, n-1}) = \ alpha / 2 $ - $ P (\ frac {S ^ 2} {\ chi ^ 2\_ {1- \ alpha / 2, n-1}} (n-1)> \ sigma ^ 2) = \ alpha / 2 $ - Together: $ P (\ frac {S ^ 2} {\ chi ^ 2\_ { 1- \ alpha / 2, n-1}} (n-1) <\ sigma ^ 2 <\ frac {S ^ 2} {\ chi ^ 2 \_ {\ alpha / 2, n-1}} (n-1 )) = 1 - \ alpha $

L80: manual calculation

L81: ## Probability of occurrence with one selection

L82: #### IO probabilities

L83: - we estimate the probability - the point estimate is the relative frequency - we require a sufficient number of data: $ n> \ frac {9} {p (1-p)} $ - Clopper - Pearson estimate (binom.test) - does not take data as a parameter, but the number of successes and the number of observations - Wald's - from the selection characteristics

L84: verification of assumptions

L85: point estimate

L86: Clopper-Pearson interval estimation

L87: Wald's interval estimation

L88: lower limit of IO

L89: upper limit of IO

L90: Calculation of the 11 most frequently used confidence intervals param. bin. distribution

L91: using the binom package

L92: install.packages ("binom")

L93: # Examples

L94: ## Example 1.

L95: During control tests of 16 light bulbs, an estimate of the mean value of $ \ bar x $ = 3,000 hours and the standard deviation s = 20 hours of their life were determined. Assuming that the lamp life has a normal distribution, determine a 90% interval estimate for the µ and σ parameters

L96: We estimate the mean value and standard deviation of the lamp life

L97: Data normality information is included in the entry

L98: file range

L99: hours .... average (point estimate of mean value)

L100: hours .... sample standard deviation (point estimate of standard deviation)

L101: significance level (reliability 1-alpha = 0.9)

L102: Bilateral interval estimate of the mean

L103: lower IO limit

L104: upper limit of IC

L105: Bilateral interval estimate of the standard deviation

L106: lower IO limit

L107: upper limit of IC

L108: ## Example 2.

L109: The depth of the sea is measured with an instrument whose systematic error is zero and the random errors have a normal distribution with a standard deviation of 20 m.

L110: We determine the estimate of the required range of selection (number of required measurements)

L111: We assume data normality, with known variance (according to assignment)

L112: meters .... known standard deviation

L113: significance level (reliability 1-alpha = 0.95)

L114: meters ... permissible measurement error

L115: Estimate the range of the selection

L116: Y = delta / sigma \* sqrt (n) ~ N (0,1), delta = X-mu

L117: P (Y> Z\_ (1-alpha / 2)) = alpha / 2

L118: ## Example 3.

L119: The task is to determine the average serum cholesterol level in a certain population of men. In a random sample (derived from the normal distribution) of 25 men, the sample mean is 6.3 mmol / l and the sample standard deviation is 1.3 mmol / l.

L120: We estimate mean serum cholesterol levels

L121: We assume data normality (according to assignment)

L122: file range

L123: mmol / l .... average (point estimate of mean value)

L124: mmol / l .... selection direction. deviation (point estimate of deviation)

L125: significance level (reliability 1-alpha = 0.95)

L126: Bilateral interval estimate of the mean

L127: lower IO limit

L128: upper limit of IC

L129: ## Example 4.

L130: Suppose that in a random selection of 200 young men, 120 of them have higher than recommended serum cholesterol levels. Determine a 95% confidence interval for the percentage of young men with higher cholesterol levels in the population.

L131: We estimate the proportion of men with higher cholesterol levels in the entire population,

L132: ie the probability that a randomly selected man will have a higher cholesterol level

L133: file range

L134: number of "successes"

L135: relative frequency (probability point estimate)

L136: significance level (reliability 1-alpha = 0.95)

L137: Verification of assumptions

L138: Bilateral Clopper - Pearson (exact) int. binom. distribution

L139: # Wald's (asymptotic) estimate (z-statistic) - approx. normal distribution according to CLV

L140: lower IO limit

L141: upper limit of IO

L142: ## Example 5.

L143: In a research study, we work with a random selection of 70 women from the Czech population. Hemoglobin was measured in each of the women with an accuracy of 0.1 g / 100 ml. The measured values are listed in the Hemoglobin.xls file. Find 95% interval estimates of standard deviation and mean hemoglobin in the population of Czech women. (Check the normality based on the exploration graphs.)

L144: # We estimate the mean and standard deviation of hemoglobin in serum

L145: # Read data from xlsx file (using readxl package)

L146: # Exploratory analysis

L147: Data does not contain remote observations.

L148: Verification of normality - exploratory

L149: Slanting and pointing corresponds to standards. distribution.

L150: normality verification: exact - normality test.

L151: If we know hypothesis testing, we verify Shapir. Wilk's test.

L152: In hl. significance 0.05

L153: 95% bilateral interval estimate of the mean

L154: # 95% two-way interval standard deviation estimate

L155: ## Example 6.

L156: What must be the number of observations if we want to determine the average hemoglobin value in neonates with an error of at most 1.0 $ g / l $ with a probability of 0.95. Population variance is estimated at $ 46.0 g ^ 2 / l ^ 2 $.

L157: We determine the estimated required range (number of newborns we have to test)

L158: We assume data normality, without this assumption the example is unsolvable

L159: g / l .... known standard deviation

L160: significance level (reliability 1-alpha = 0.95)

L161: g / l ... permissible measurement error

L162: Estimate selection range

L163: Y = delta / sigma \* sqrt (n) ~ N (0,1), delta = X-mu

L164: P (Y> Z\_ (1-alpha / 2)) = alpha / 2

L165: ## Example 7.

L166: In the data file pr7.xlsx you will find the measurement of noise caused by the computer fan [dB]. Calculate the 95% interval estimate of the average noise and the 95% interval estimate of the noise variability.

L167: read data

L168: visualization

L169: OP removal

L170: data normality test exploratory

L171: normality test exactly

L172: point and interval estimation of the mean

L173: point and interval estimate of standard deviation

L174: ## Example 8.

L175: In the data file pr8.xlsx you will find the measurement of the time to failure of the electrical component [h]. Calculate the 99% interval estimate of the average life of a given component type.

L176: read data

L177: visualization and verification of OP

L178: data normality test exploratory

L179: normality test exactly

L180: symmetry test exploratory

L181: exact: symmetry test

L182: install.packages ("lawstat")

L183: the resulting p-value must be greater than hl. challenge (eg 0.05)

L184: median point and interval estimation

L185: IO

L186: install.packages ("BSDA")

%%%%%%%%%% cv6

L0: # Exercise 6 - Selected distributions of a continuous random variable

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Overview of divisions and their functions

L3: ## Introduction: Probability Density, Distribution Function and Quantile Function

L4: ### Probability density

L5: - starts with the letter \*\* d \*\*: p = d ... (x, ...)

L6: ### Distribution function

L7: - starts with the letter \*\* p \*\*: $ p = P (X <x) $: p = p ... (x, ...)

L8: ### Quantile function

L9: - starts with the letter \*\* q \*\*: find x for the given p: $ p = F (x) \ rightarrow x = F ^ {- 1} (p) $: x = q ... (p, .. .)

L10: ## Even distribution: $ X \ sim Ro (a, b) $

L11: - the random variable acquires only values greater than a and less than b - all values have the same density -> the probability density is constant between a and b, elsewhere zero

L12: Probability density f (x)

L13: from where

L14: where

L15: we plot the probability density

L16: cex is the size of the markers

L17: Distribution function F (x) = P (X <x)

L18: from where

L19: where

L20: let's draw the Distribution function

L21: quantile function F ^ (- 1) (p) = x: P (X <x) = p

L22: from where

L23: where

L24: plot - quantile function F ^ (- 1) (p) = x

L25: ## Exponential distribution: $ X \ sim Exp (\ lambda) $

L26: - time to 1st event, time between events (only in the period of stable life - Poisson process) - parameter $ \ lambda $ is the same as in Poisson distribution - mean value is: $ E (X) = 1 / \ lambda $

L27: Probability density f (x)

L28: we plot the probability density

L29: Distribution function F (x) = P (X <x)

L30: let's draw the Distribution function

L31: quantile function F ^ (- 1) (p) = x: P (X <x) = p

L32: plot - quantile function F ^ (- 1) (p) = x

L33: ## Weibull distribution: $ X \ sim W (\ theta, \ beta) $

L34: - time to 1st event (faults) (suitable choice of β allows use in any period of fault intensity) - extension of exponential distribution Exp (λ) = W (Θ = 1 / λ, β = 1)

L35: Probability density f (x)

L36: equivalent 1 / lambda in exp. distribution

L37: beta = 1 -> exponential distribution

L38: we plot the probability density

L39: Distribution function F (x) = P (X <x)

L40: equivalent of 1 / lambda in exp. distribution

L41: beta = 1 -> exponential distribution

L42: we draw the Distribution function

L43: quantile function F ^ (- 1) (p) = x: P (X <x) = p

L44: equivalent of 1 / lambda in exp. distribution

L45: beta = 1 -> exponential distribution

L46: plot - quantile function F ^ (- 1) (p) = x

L47: ## Normal distribution: $ X \ sim N (\ mu, \ sigma ^ 2) $

L48: - distribution modeling eg measurement errors, sum / average behavior of many other random variables - see. The central limit theorem - $ \ mu $ is directly the mean value of the distribution: $ E (X) = \ mu $ - $ \ sigma $ is the directly standard deviation of the distribution: $ D (X) = \ sigma ^ 2 $ - with parameters $ \ mu = 0, \ sigma = 1 $ is called the normalized normal distribution

L49: Probability density f (x)

L50: we plot the probability density

L51: Distribution function F (x) = P (X <x)

L52: we draw the Distribution function

L53: quantile function F ^ (- 1) (p) = x: P (X <x) = p

L54: plot - quantile function F ^ (- 1) (p) = x

L55: # Examples

L56: ## Example 1.

L57: Height in the population of boys aged 3.5-4 years has a normal distribution with a mean value of 102 cm and a standard deviation of 4.5 cm. Determine what percentage of boys at that age are less than or equal to 93 cm in height.

L58: X ... height of boys aged 3.5 to 4 years (cm)

L59: X ~ N (mu = 102, sd = 4.5)

L60: P (X <= 93) = F (93)

L61: ## Example 2.

L62: The average life of a machine part is 30,000 hours. Assume that the component is in a period of stable life. Specify:

L63: X ... component life (h)

L64: X ~ Exp (lambda), where E (X) = 1 / lambda

L65: ### a)

L66: probability that the component will not last more than 2,000 hours,

L67: a) P (X <2000) = F (2000)

L68: ### b)

L69: probability that the component will last more than 35,000 hours,

L70: b) P (X> 35000) = 1-F (35000)

L71: ### c)

L72: time until 95% of the components fail.

L73: c) P (X <t) = 0.95 -> F (t) = 0.95 -> t… 95% quantile

L74: ## Example 3.

L75: The production facility fails on average once every 2000 hours. The quantity Y representing the fault waiting time has an exponential distribution. Determine the time T0 so that the probability that the device will run longer than T0 is 0.99.

L76: X ... fault waiting time (h)

L77: X ~ Exp (lambda), where E (X) = 1 / lambda

L78: P (X> t) = 0.99 -> 1-F (t) = 0.99 -> F (t) = 0.01 -> t… 1% quant.

L79: ## Example 4.

L80: The measurement results are loaded only with a normally distributed error with zero mean and a standard deviation of 3 mm. What is the probability that there will be an error in the interval (0 mm; 2.4 mm) at least once in 3 measurements?

L81: Y měření measurement error size (mm)

L82: Y ~ N (mu = 0, sigma = 3)

L83: pp… true that the measurement error will be in int. 0.0-2.4mm

L84: X… number of measurement errors in int. 0 mm -2.4 mm in 3 measures.

L85: X ~ Bi (n = 3, p = pp)

L86: P (X> = 1) = 1-P (X = 0)

L87: ## Example 5.

L88: An average of 25 users per hour log on to a large computer network. Determine the probability that:

L89: ### a)

L90: no one logs in during 14:30 - 14:36,

L91: X… number of users logged in in 6 minutes

L92: X ~ Po (lt = 2.5)

L93: P (X = 0)

L94: ### b)

L95: 2-3 minutes will elapse before the next login.

L96: Y… time until next login

L97: Y ~ Exp (lambda = 25/60), where E (X) = 1 / lambda

L98: P (2 <Y <3) = F (3) -F (2)

L99: ### c)

L100: Specify the maximum length of the time interval so that the probability that no one will log in is at least 0.90.

L101: P (Y> t) = 0.90 -> 1-F (t) = 0.90 -> F (t) = 0.10 -> t… 10% kv.

L102: ## Example 6.

L103: The random variable X has a normal distribution N (µ; σ). Specify:

L104: ### a)

L105: P (µ - 2σ <X <µ + 2σ),

L106: P (µ - 2σ <X <µ + 2σ) = F (µ + 2σ) - F (µ - 2σ)

L107: X ~ N (µ, σ)

L108: it does not matter what values we choose

L109: ### b)

L110: smallest k ∈ Z, so that P (µ - kσ <X <µ + kσ)> 0.99.

L111: normal distribution is symmetric

L112: P (µ - kσ <X <µ + kσ) =

L113: = 1 - (P (X <µ - kσ) + P (X> µ + kσ)) =

L114: = 1 - 2 \* P (X> µ + kσ) = 0.99 -> P (X> µ + kσ) = 0.005

L115: -> P (X <µ + kσ) = 0.995

L116: x = µ + kσ

L117: ## Example 7.

L118: An accompanying film about the life of the author of the exhibited works is screened on a tour of the exhibition. His screening begins every 20 minutes. Determine the probability that if you accidentally come to the screening room,

L119: Y… time until the start of the next projection

L120: Y ~ Ro (a = 0, b = 20)

L121: ### a)

L122: you won't have to wait more than 5 minutes for the movie to start,

L123: P (X <5)

L124: ### b)

L125: you will wait between 5 and 10 minutes,

L126: P (5 <X <10)

L127: ### c)

L128: mean and standard deviation of the waiting time for the beginning of the movie.

L129: ## Example 8.

L130: During quality control, we only take over the part if its dimension is in the range of 26-27 mm. The dimensions of the components have a normal distribution with a mean value of 26.4 mm and a standard deviation of 0.2 mm. What is the probability that the size of the part randomly selected for inspection will be within the required limits?

L131: X ... part size (mm)

L132: X ~ N (mu = 26.4, sigma = 0.2)

L133: P (26 <X <27) = F (27) -F (26)

L134: ## Example 9.

L135: The length of jumps of the athlete Jakub measured in cm has a normal distribution N (µ1; σ1), where µ1 = 690 and σ1 = 10. The length of jumps of the athlete Aleš measured in cm also has a normal distribution N (µ2; σ2), where µ2 = 705 and σ2 = 15. The one who jumps more than 700 cm from two jumps at least once qualifies for the races.

L136: SJ ... length of Jacob's jump

L137: SJ ~ N (mu = 690, sigma = 10)

L138: SA Ale Aleš's jump length

L139: SA ~ N (mu = 705, sigma = 15)

L140: J ... Jakub's jump is successful (longer than 700 cm)

L141: And ... Aleš's jump is successful (longer than 700 cm)

L142: P (J) = P (SJ> 700) = 1-F (700)

L143: P (A) = P (SA> 700) = 1-F (700)

L144: KJ… Jakub qualifies for races,

L145: P (KJ) = 1- (1-P (J)) (1-P (J))

L146: KA… Aleš qualifies for races,

L147: P (KA) = 1- (1-P (A)) (1-P (A))

L148: ### a)

L149: How likely are they both to qualify for the race?

L150: ada)

L151: ### b)

L152: With what probability does Aleš qualify, but Jakub does not?

L153: adb)

%%%%%%%%%% cv1

L0: # Exercise 1 - Brief introduction to R, Combinatorics

L1: ## Adéla Vrtková, Michal Béreš, Martina Litschmannová

L2: First we will go through some basics of Rk and implement some combinatorial functions.

L3: # Brief introduction to R

L4: simple arithmetic operations

L5: BEWARE of brackets! Only round ones are used for counting!

L6: Square and compound have a different function in R.!

L7: combination number, factorials

L8: data types -> numeric, character, logical, (complex)

L9: the class function determines the type of the object

L10: ## data structures in R

L11: - vector (we mean column vector) - factor (special case of vector) - matrix (matrix with dimensions nxm) - data.frame (data frame with dimensions nxp) - sheet (list)

L12: vector definition (column = column)

L13: other options

L14: creates a vector with four ones

L15: sequence from 1 to 10 with step 2

L16: sequence from 1 to 10 with step 1

L17: redefining an object to another type - eg as.vector, as.matrix, as.factor, ...

L18: working with vectors - merging by columns / rows

L19: # matrix definition

L20: diagonal matrix

L21: # matrix operations - pay attention to matrix multiplication ->% \*%

L22: <hr>

L23: # Combinatorics

L24: ## Variations

L25: V (n, k) - variation without repetition, the first argument will be the total number of entities, the second argument the size of the selection

L26: the function is created by the fucntion command, it is an object whose name is given by a variable

L27: to which I will assign this object

L28: here I enter the number of parameters and their names

L29: the whole body of the function is enclosed in parentheses {...}

L30: the factorial in the original Rku exists so we will use it

L31: what the function returns is given in the return (...) statement

L32: V \* (n, k) - variation with repetition

L33: ## Permutation

L34: P (n) = V (n, n) - permutation

L35: P \* (n1, n2, n3, ...., nk) - permutation with repetition, input will be a vector with individual numbers of unique entities

L36: vec\_n is the vector of values eg: vec\_n = c (2,2,2,4,3)

L37: we calculate how many values we have in total

L38: their factorial = value in the numerator

L39: a simple loop starts with the for statement, then the iterator name az follows in parentheses

L40: what list will be taken

L41: count is an iterator and will gradually take values from the vector vec\_n

L42: we gradually divide by the factorial of each number of unique entities

L43: ## Combination

L44: C (n, k) - combination

L45: the for combination function already exists in Rku and is called choose

L46: C \* (n, k) - combination with repetition

L47: we use a known formula

L48: # Exercise tasks

L49: ## Example 1.

L50: Which password is more secure? \* Eight-character password consisting of numbers only. \* A five-character password composed only of letters of the English alphabet.

L51: password 1

L52: password 2

L53: ## Example 2.

L54: How long would it take to solve a business traveler's problem for n = 10 cities by brute force if the evaluation of the length of each of the possible journeys takes 1 µs?

L55: ## Example 3.

L56: How to divide the loot between 2 robbers to get both items in the same value (or as close as possible). I.e. can I divide the N numbers entered into two groups so that the sum of the numbers in both groups is the same? \*\* How many options would be needed if we solved the task with brute force? \*\*

L57: ## Example 4.

L58: How many anagrams of the word "AUTO" can we create? How many anagrams of the word "CAR" can we create? How many of them start with "K"?

L59: ## Example 5.

L60: They have 6 types of colored cups in the store. - How many different ways can we buy 4 different-colored mugs? - How many different options can we buy 5 cups (if we don't mind more from the same color)? - How will the situation change if they have only 4 pieces from each (and we don't mind more of the same color)?

L61: ## Example 6. (collection chap. 1, ex. 7.8)

L62: From the urn with three balls, two red and one white, two balls will be selected at the same time. The student and the teacher place a bet. If both balls are the same color, the student wins. If the balls have different colors, the teacher wins. - Is the game fair? What are the probabilities of a teacher and a student winning? - What ball should be added to make the game fair?

L63: ## Example 7.

L64: There are 5 different pairs of socks in the package (the left and right socks are always the same). - How many different pairs of socks can be chosen? - How many different ways can I get dressed? (ie it depends on what is on which leg)

L65: ## Example 8.

L66: I have 12 weights weighing 1.2, ..., 12 kg. - How many ways can I divide them into 2 piles? - How many ways can I divide them into 3 piles? - How many ways can I divide them into 3 piles if they all have the same number of weights? - How many ways can I divide them into 3 piles of the same number of weights if none of them can exceed 40 kg?

L67: ## Example 9.

L68: I have 20 seeds from each of the three vegetables (carrots, radishes, celery). Unfortunately, she interfered. - I'll plant 5 random seeds in the box. What is the probability that there will be at least three radishes among them? - I'll plant 5 random seeds in the box. What is the probability that there will be more carrots than celery among them?

%%%%%%%%%% cv3

L0: # Exercise 3 - Discrete random variable

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Examples

L3: ## Example 1.

L4: The owner of the service center offered a car dealership that set up a car rental company. For each car rented through it, they will receive CZK 500 from the car rental company. At the same time, however, he undertook to invest CZK 800 in the maintenance of rented cars every day. The number of cars rented through the service center in 1 day is described by the following probability function:! [Image-3.png] (attachment: image-3.png)

L5: ### 1. a)

L6: The value of the probability function for 5 cars was difficult to read. Specify it:

L7: Computer arithmetic can be annoying here

L8: round to the hundredth

L9: the notation for x = 5 is the 6th position

L10: We start with graphics -> we further modify the graph - we use other parameters

L11: parameter for the y-axis range

L12: parameter for y-axis description

L13: parameter for graph name

L14: determines what type of graph it is (p -> points, point)

L15: parameter for the appearance of displayed points

L16: colors

L17: parameter for adjusting the size of the entire graph (in the sense of magnifying 2x)

L18: parameter separately for the size of the axis names

L19: parameter separately for the size of the values on the axes

L20: parameter for chart name size

L21: Probability function

L22: solid wheels - in actual values

L23: horizontal grid

L24: vertical grid

L25: that we want to draw in one graph

L26: empty circles - where a non-zero value is defined

L27: set values to X

L28: and Y

L29: Notes on the introduction to graphics are based on the so-called high-level functions that create a graph (ie open a graphics window and draw according to the specified parameters) followed by so-called low-level functions that add something to the active graphics window, themselves the "text" function used above does not open itself the low-level function - it adds text to the existing active graphics window other low-level functions - eg abline, points, lines, legend, title, axis ... which add a line, points , legend ... ie. before using the "low-level" function it is necessary to call the "high-level" function (eg plot, boxplot, hist, barplot, pie, ...) you can find other graphic parameters in the help or eg here http: // www .statmethods.net / advgraphs / parameters.html or here https://flowingdata.com/2015/03/17/r-cheat-sheet-for-graphical-parameters/ or http://bcb.dfci.harvard.edu /~aedin/courses/BiocDec2011/2.Plotting.pdf

L30: ### 1. b)

L31: Determine and plot the distribution function of the random variable X, which is defined as the number of rented cars.

L32: simplified distribution function graph

L33: Function for calculating and plotting the distribution function

L34: we stretch F by 0 at the beginning

L35: axz both sides

L36: empty wheels

L37: that we want to draw in one graph

L38: full wheels

L39: horizontal grid

L40: vertical grid

L41: graph - lines

L42: set values to X

L43: a Y

L44: ### 1. c)

L45: Determine the mean, variance, standard deviation, and mode of the number of cars rented per day.

L46: Mean value

L47: Scattering

L48: second general moment

L49: Standard deviation

L50: Functions for calculating basic numerical characteristics

L51: write the results to the table

L52: ### 1. d)

L53: Determine the probability function and the distribution function of the random variable Y, which is defined as the daily income of the service owner.

L54: Distribution function

L55: ### 1. e)

L56: Determine the mean, standard deviation, and mode of receipt of the service owner from rented cars within one day.

L57: ### 1. f)

L58: Determine the probability that the service owner's income (random variable Y) from car rental will exceed his expenses.

L59: profit

L60: income exceeds expenses when profit is positive

L61: ### 1. g)

L62: Determine the mean, standard deviation, and mode of the random variable Z, which is defined as the service owner's profit from rented cars in one day.

L63: ## Example # 2

L64: For the distribution function of the random variable X: $ F (x) = \ begin {cases} 0 & x \ leq -1 \\ 0.3 & -1 <x \ leq 0 \\ 0.7 & 0 <x \ leq 1 \ \ 1 & -1 <x \ end {cases} $

L65: ### 2. a)

L66: Determine the probability function of a random variable X, its mean and standard deviation.

L67: ### 2. b)

L68: Random variable Y = 1 - 3X, determine P (y), F (y), E (Y), D (Y).

L69: Nonsensical output - what is the cause?

L70: The order function returns the sorted order indexes

L71: ### 2. c)

L72: Random variable W = $ 3X ^ 2 $, determine P (w), F (w), E (W), D (W).

L73: initialize an array of the same size

L74: ## Example 3.

L75: There are two machines working independently in the workshop. The probability of failure of the first machine is 0.2, the probability of failure of the second machine is 0.3. The random variable X is defined as the number of machines that have failed at the same time. Specify:

L76: ### 3. a)

L77: probability function of a random variable X,

L78: we calculate the individual probabilities of the number of broken machines

L79: 0 broken, so both in operation

L80: 2 so broken both

L81: just one - either the first or the second

L82: ### 3. b)

L83: distribution function of random variable X,

L84: ### 3. c)

L85: mean and variance of a random variable X.

%%%%%%%%%% cv8

L0: # Exercise 8. Selection characteristics ## Michal Béreš

L1: # Other selected continuous distributions

L2: ## $ \ chi ^ 2 $ - Chi-square distribution (Pearns distribution)

L3: - Usage: when estimating the standard deviation (using selection) - Has a single parameter - Number of degrees of freedom - $ \ frac {S ^ 2} {\ sigma ^ 2} (n-1) \ sim \ chi ^ 2\_ {n -1} $ - $ S $ is the sample standard deviation

L4: number of degrees of freedom

L5: x-axis

L6: chi-quad probability density. distribution

L7: distrib. fce. chi-kvad. distribution

L8: ## $ t $ - Student's distribution

L9: - Usage: when estimating the mean without exact variance knowledge (sample variance only) - $ \ frac {\ bar X - \ mu} {S} \ sqrt {n} \ sim t\_ {n-1} $ - $ \ bar X $ is the sample mean - $ S $ is the sample standard deviation - with increasing number of degrees of freedom converges to the normalized normal distribution

L10: number of degrees of freedom

L11: x-axis

L12: probability density of the student's distribution

L13: norm values. normal birth

L14: to the last graph

L15: probability density of the student's distribution

L16: norm values. normal birth

L17: to the last graph

L18: ## $ F $ - Fisher-Snedecor distribution

L19: - Used to test stairs of variance - $ \ frac {S\_1 ^ 2 / \ sigma\_1 ^ 2} {S\_2 ^ 2 / \ sigma\_2 ^ 2} \ sim F\_ {n\_1 - 1, n\_2 - 1} $

L20: number of degrees of freedom selection. 1

L21: number of degrees of freedom selection. 2

L22: x-axis

L23: chi-quad probability density. distribution

L24: chi-quad probability density. distribution

L25: # How does the average of the values from the normal distribution behave?

L26: The function \*\* rnorm (n, mean, sd) \*\* generates \*\* n \*\* values from the normal distribution with the mean value \*\* mean \*\* and the standard deviation \*\* sd \*\*.

L27: ### Random variable: average of values

L28: numeric produces vector 0

L29: # How does the average of values from the uniform distribution behave?

L30: The function \*\* runif (n, min, max) \*\* generates \*\* n \*\* values from the uniform distribution U (\*\* min, max \*\*).

L31: nah\_vyber

L32: ### Random variable: average of values

L33: # Examples

L34: ## Example 1.

L35: The load on an aircraft with 64 seats shall not exceed 6,000 kg. What is the probability that this value will be exceeded at full occupancy if the passenger mass has a mean value of 90 kg and a standard deviation of 10 kg?

L36: X ... weight 64 passengers

L37: X ~ N (64 \* 90; 64 \* 100)

L38: P (X> 6000) = 1 - F (6000)

L39: ## Example 2.

L40: The consignment contains 300 products of a certain type. It is known that the probability of making a defective product of this type is 0.04.

L41: ### a)

L42: Estimate the probability that the absolute deviation of the proportion of defective products in the consignment from the probability of producing a defective product will be less than 1%.

L43: X = (p - π) / sqrt (π \* (1 - π)) \* sqrt (n) ∼ N (0, 1)

L44: P (-0.01 / sqrt (π \* (1 - π)) \* sqrt (n) <X <0.01 / sqrt (π \* (1 - π)) \* sqrt (n))

L45: ### b)

L46: How will the result change if the shipment contains 3,000 products?

L47: ## Example 3.

L48: Passengers regularly travel to and from public transport. It is known that the waiting time for the arrival of public transport ranges from 0 to 3 minutes. What is the probability that the total waiting time for an employee to arrive by public transport in 23 working days will be less than 80 minutes?

L49: Y ... waiting time for public transport

L50: y ~ R (0; 3)

L51: X ... total waiting time in 23 days (round trip ⇒ 46 waiting)

L52: X ~ N (46 \* EY; 46 \* DY)

L53: P (X <80)

L54: ## Example 4.

L55: Assume that the average electricity consumption of households in a given city in January is 120 kWh and the standard deviation of consumption is 100 kWh. Determine the probability that the average consumption of 100 randomly selected households will be greater than 140 kWh.

L56: Xi ... consumption of the i-th household

L57: X ... average consumption of 100 households

L58: X ~ N (EXi; Dxi / n)

L59: P (X> 140)

L60: ## Example 5.

L61: Acme Battery Company has developed a new type of mobile phone battery. On average, batteries last 60 minutes on a single charge. The standard deviation of this time is 4 minutes. Assume that the production department runs a quality control test after 6 months. They performed two random selections with a range of 10 batteries and in both found a standard deviation of battery life greater than 6 minutes. How likely were they to expect such a result?

L62: X = (n - 1) S ^ 2 / σ ^ 2

L63: X ∼ χ\_n-1

L64: P (S> 6) = P (X> ...)

L65: ## Example 6.

L66: Mortality tables show a probability of 0.99 that a 35-year-old man will live another year. The annual premium for this age group is CZK 2,000, in the event of death the insurance company will pay CZK 100,000. What is the probability that the profit of 500 insured men aged 35 will be at least CZK 500,000? (Solve in two ways - using the binomial distribution and using the binomial distribution approximation by the normal distribution.)

L67: X ... number of men out of 500 who won't live to see another year

L68: X ~ Bi (500; 0.01)

L69: Z = 500 · 2,000 - X · 100,000

L70: P (Z ≥ 500,000) = P (X ≤ 5)

L71: X ~ Bi (500; 0.01) ~ N (500 \* 0.01; 500 \* 0.01 \* (1-0.01))

L72: P (X ≤ 5) ~ P (X <5.5) (continuity correction)

L73: ## Example 7.

L74: Assume that approximately 60% of young men in the population have higher than recommended serum cholesterol levels. In a random selection of 200 young men, how likely will more than 120 of them have higher than recommended serum cholesterol levels?

L75: X ... number of young men out of 200 with higher than recommended serum cholesterol levels

L76: X ∼ Bi (200; 0.6)

L77: P (X> 120) = 1 - P (X ≤ 120)

L78: X ~ N (200 \* 0.6; 200 \* 0.6 (1-0.6)), ie X ≈ N (120; 48)

L79: 1 - P (X ≤ 120) ~ 1 - P (X <120.5) (continuity correction)

%%%%%%%%%% cv5

L0: # Exercise 5 - Selected distributions of a discrete random variable

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: # Overview of divisions and their functions

L3: ## Introduction: Probability, Cumulative Probability (Distribution) and Quantile functions

L4: ### Probability function

L5: - starts with the letter \*\* d \*\*: $ p = P (X = x) $: p = d ... (x, ...)

L6: ### Cumulative Probability (Distribution Function)

L7: - starts with the letter \*\* p \*\*: $ p = P (X \ leq x) $: p = p ... (x, ...) - note Cumulative probability is with the alternative definition $ P (X \ leq t) $ - for our distribution function $ F (t) = P (X <t) $: F (t) = p ... (t - 1, ...)

L8: ### Quantile function

L9: - starts with the letter \*\* q \*\*: $ p \ geq P (X \ leq x) $: x = q ... (p, ...) - searches for the smallest $ x $ for which is $ P (X \ leq x) $ greater than $ p $

L10: ## Binomial (Alternative): $ X \ sim Bi (n, π), X \ sim A (π) = Bi (1, π) $

L11: - number of successes in $ n $ Bernoulli attempts (or for one attempt in the case of Alternative) - each attempt has a chance of success $ π $

L12: Probability function P (X = x)

L13: value for which we are looking for a p-st function

L14: selection range

L15: probability of success

L16: this can be used to turn off warnings

L17: turn it on again

L18: draw a probability function

L19: minimum 0, maximum n has a positive probability

L20: Cumulative probability function P (X <= x)

L21: value for which we are looking for a cumulative p-st function

L22: selection range

L23: probability of success

L24: Distribution function F (x) = P (X <x)

L25: value for which we are looking for a cumulative p-st function

L26: selection range

L27: probability of success

L28: or

L29: we draw the distribution function

L30: minimum 0, maximum n has a positive probability

L31: or

L32: minimum 0, maximum n

L33: check correctness at 10

L34: minimum 0, maximum n

L35: find x for given q: q = P (X <= x)

L36: h

L37: selection range

L38: probability of success

L39: Quantile function (inversion of dist. Function): q = F (x) = P (X <x)

L40: probability for which we are looking for a quantile

L41: selection range

L42: probability of success

L43: ## Hypergeometric: $ X \ sim H (N, M, n) $

L44: - number of successes in $ n $ dependent attempts - dependency type:

L45: - $ N $ objects,

L46: - of which $ M $ objects with specified property,

L47: - size selection $ n $

L48: - \*\* we do not return when selecting - the probability of selecting an object with a given property changes with each additional selected object \*\*

L49: - \*\* R function takes as parameters \* hyper (k, M, N - M, n) \*\*

L50: - k is the number of successes for which we calculate the probability,

L51: - M is the number of objects with the specified property,

L52: - NM is the number of objects without the specified property,

L53: - n is the target size of the selection.

L54: Probability function P (X = x)

L55: value for which we are looking for a p-st function

L56: total number of objects

L57: of which with specified property

L58: selection size

L59: draw a probability function

L60: minimum 0, maximum n or M has a positive truth.

L61: Distribution function F (x) = P (X <x)

L62: value for which we are looking for dist. function

L63: total number of objects

L64: of which with specified property

L65: selection size

L66: let's draw the Distribution function

L67: minimum 0, maximum n or M has a positive truth.

L68: Quantile function (inversion of dist. Function): q = P (X <x)

L69: probability for which we are looking for a quantile

L70: total number of objects

L71: of which with specified property

L72: selection size

L73: ## Negative binomial (Geometric): $ X \ sim NB (k, π), X \ sim Ge (π) = NB (1, π) $

L74: - number of attempts up to $ k $. success (inclusive) - each attempt has a chance of success $ π $ - \*\* Negatively binomial NV is defined in Rku as the number of failures before the success \*\*

L75: - therefore we will send x - k as the first parameter

L76: Probability function P (X = x)

L77: number of attempts for which we are looking for truth. fci

L78: required number of successes

L79: truth. individual trials

L80: Note that the first argument must be the number of failures

L81: draw a probability function

L82: minimum k, maximum unlimited

L83: values 0,1,2,3,4 have P (x) = 0

L84: Distribution function F (x) = P (X <x)

L85: number of attempts for which we are looking for truth. fci

L86: required number of successes

L87: truth. individual trials

L88: Note that the first argument must be the number of failures

L89: let's draw the Distribution function

L90: minimum 0, maximum n or M has a positive truth.

L91: Quantile function (inversion of dist. Function): q = P (X <x)

L92: truth. for quantile

L93: required number of successes

L94: true individual trials

L95: ## Poisson: $ X \ sim Po (λt) $

L96: - number of events in the Poisson process in a closed area (in time, on area, in volume) - with occurrence density $ λ $ - in time / area / volume of size $ t $

L97: Probability function P (X = x)

L98: number of attempts for which we are looking for truth. fci

L99: occurrence density

L100: true individual trials

L101: draw a probability function

L102: minimum 0, maximum unlimited

L103: Distribution function F (x) = P (X <x)

L104: number of attempts for which we are looking for truth. fci

L105: density of occurrence

L106: true individual trials

L107: let's draw the Distribution function

L108: minimum 0, maximum n or M has a positive truth.

L109: Quantile function (inversion of dist. Function): q = P (X <x)

L110: true for quantile

L111: density of occurrence

L112: true individual trials

L113: # Examples

L114: ## Example 1.

L115: Bridge is played with 52 bridge cards, which are dealt among 4 players. There are always 2 players playing together. When dealing (13 cards) you received 2 aces. What is the probability that your partner will have the remaining two aces?

L116: X ... number of aces among 13 cards

L117: X ~ H (N = 39, M = 2, n = 13)

L118: P (X = 2)

L119: 52-13

L120: calculation

L121: which is dhyper (2,2,37,13)

L122: graph of probability function

L123: all possible implementations of NV X.

L124: values of the probability function for x

L125: ## Example 2.

L126: Experiments have shown that a radioactive substance emits within 7.5 s an average of 3.87 α-particles. Determine the probability that this substance will emit at least one α-particle in 1 second.

L127: X ... number of radiated alpha particles during 1 s

L128: X ~ Po (lt = 3.87 / 7.5)

L129: frequency of occurrence

L130: in 1 second

L131: Poisson distribution parameter

L132: P (X> = 1) = P (X> 0) = 1 - P (X <= 0)

L133: graph of probability function

L134: theoretically up to an infinite number of particles can be emitted,

L135: from a certain value the probability is negligible

L136: values of the probability function for x

L137: ## Example 3.

L138: A friend sends you to the cellar to bring 4 bottled beers - two dozen and two twelve. You don't know where to light it, so you take 4 bottles blindly from the bass. How likely were you to comply if you knew that there were a total of 10 tens and 6 twelve in the base?

L139: X ... number of 10 ° beers among 4 selected

L140: X ~ H (N = 16, M = 10, n = 4)

L141: P (X = 2)

L142: probability function graph

L143: all possible implementations of NV X.

L144: values of the probability function for x

L145: ## Example 4.

L146: On average, there are 15 certain microorganisms in one milliliter of a perfectly mixed solution. Determine the probability that there will be less than 5 of these micro-organisms in a test tube if a sample of 1/2 milliliter is randomly selected.

L147: X ... number of microorganisms in 0.5 ml of solution

L148: X ~ Po (lt = 15/2)

L149: Poisson distribution parameter

L150: P (X <5) = P (X <= 4)

L151: or

L152: graph of probability function

L153: theoretically there can be up to an infinite number of microorganisms in solution,

L154: from a certain value the probability is negligible

L155: values of the probability function for x

L156: ## Example 5.

L157: Pour 15 coins on the table. What is the probability that the number of coins lying face up is from 8 to 15?

L158: X ... number of coins that fall face up out of a total of 15 coins

L159: X ~ Bi (n = 15, p = 0.5)

L160: P (8 <= X <= 15) = P (X <= 15) - P (X <8) = P (X <= 15) - P (X <= 7)

L161: otherwise: P (8 <= X <= 15) = P (X> 7) = 1-P (X <= 7)

L162: graph of probability function

L163: all possible implementations of NV X

L164: values of the probability function for x

L165: ## Example 6.

L166: The probability that we will call the studio of the radio station that has just announced a telephone competition is 0.08. What is the probability that we will appeal on the 4th attempt at the most?

L167: X ... number of attempts before we call the radio studio

L168: X ~ NB (k = 1, p = 0.08) or G (0.08)

L169: P (X <= 4)

L170: probability function graph

L171: theoretically we can make infinitely many attempts,

L172: from a certain value the probability is negligible

L173: values of the probability function for x

L174: ## Example 7.

L175: The factory produces 10% of defective parts per day. What is the probability that if we remove thirty components from the daily production, at least two will be defective?

L176: X ... number of defective parts out of 30 selected

L177: X ~ Bi (n = 30, p = 0.1)

L178: P (X> = 2) = 1 - P (X <2) = 1 - P (X <= 1)

L179: or P (X> = 2) all except 0 and 1

L180: graph of probability function

L181: all possible implementations of NV X

L182: probability function values for x

L183: ## Example 8.

L184: There are 200 parts in stock. 10% of them are defective. What is the probability that if we remove thirty parts from the warehouse, at least two will be defective?

L185: X ... number of defective parts out of 30 selected from 200

L186: X ~ H (N = 200, M = 20, n = 30)

L187: P (X> = 2) = 1 - P (X <2) = 1 - P (X <= 1)

L188: graph of probability function

L189: all possible implementations of NV X

L190: probability function values for x

L191: ## Example 9.

L192: A company found that some illegal software was installed on 33% of computers. Determine the probability and distribution function of the number of computers with illegal software among the three computers inspected.

L193: X ... number of computers with illegal software out of 3 checked

L194: X ~ Bi (n = 3, p = 0.33)

L195: probabilistic function

L196: all possible implementations of NV X.

L197: values of the probability function for x

L198: rounding probabilities to 3 des. places

L199: completion of the last value by 1

L200: Create a table of probability functions

L201: graph of probability function

L202: distribution function

L203: simplified distribution function listing

L204: ## Example 10.

L205: Sports is a lottery game in which the bettor bets six numbers out of forty-nine, which he expects to fall in a future draw. To participate in the game, it is necessary to choose at least one combination of 6 numbers (always 6 numbers per column) and use these crosses to mark these numbers in the columns on Sazka as in the columns, starting with the first column. The bettor wins if he guesses at least three numbers from the drawn six numbers. What is the probability that in order for the bettor to win, he will have to fill in:

L206: First the probability that we get in one column

L207: Y ... number of guessed numbers in 6 drawn from 49

L208: Y ~ H (N = 49, M = 6, n = 6)

L209: P-st guess at least 3 numbers in one column

L210: P (Y> = 3) = 1 - P (Y <3) = 1 - P (Y <= 2)

L211: ### a)

L212: just three columns,

L213: X… the number of columns the bettor will have to fill in order to win

L214: X ~ NB (k = 1, p = pp)

L215: a) P (X = 3)

L216: ### b)

L217: at least 5 columns,

L218: b) P (X> = 5) = 1 - P (X <5) = 1 - P (X <= 4)

L219: ### c)

L220: less than 10 columns,

L221: c) P (X <10) = P (X <= 9)

L222:

L223: more than 5 and at most 10 columns?

L224: P (5 <X <= 10) = P (X <= 10) - P (X <= 5)

L225: or P (X <11) - P (X <6)

L226: ## Example 11.

L227: The probability of throwing 6 on a 6-wall cube is 1/6. We roll until we roll six times 10 times.

L228: ### a)

L229: What is the mean value of the number of throws.

L230: X… rolls the dice before we roll 10 sixes

L231: X ~ NB (k = 10, p = 1/6)

L232: ### b)

L233: How many throws do we have to count on if we want the probability of throwing 10 sixes to be at least 70%.

L234: P (X <= k)> = 0.7

%%%%%%%%%% cv4\_SNV

L0: # Exercise 4 - Continuous random variable

L1: ## Martina Litschmannová, Adéla Vrtková, Michal Béreš

L2: \*\* The content of this script is only as a supplementary illustration to the exercise, it is not necessary to know the exam. It is important to be able to calculate manually. \*\*

L3: ## Numerical integration in Rku

L4: Rkov function \*\* integrate \*\* <br> integrate (f, a, b) = $ \ int\_ {a} ^ {b} f (x) dx $ - \*\* f \*\* is Rková function (defined by us) which has one input argument - a vector of values in which to return its values - \*\* a \*\* lower integration limit - \*\* b \*\* upper integration limit

L5: x ^ 2

L6: # Examples

L7: ## Example 1.

L8: Random variable X has distribution function <br> $ F (x) = \ begin {cases} 0 & x \ leq 0 \\ cx ^ 2 & 0 <x \ leq 1 \\ 1 & 1 <x \ end { cases} $ <br> What values can the constant c take?

L9: deriving F (x) we get the density of truths. f (x)

L10: corresponding probability density on the interval <0.1>

L11: f (x) = 2x

L12: c = 1, so the distribution function looks like this:

L13: x ^ 2

L14: 0 for x <= 0

L15: 1 for x> 1

L16: points on the x-axis

L17: values of F (x)

L18: Draw as line

L19: ## Example 2.

L20: The distribution of the random variable X is given by the density <br> $ f (x) = \ begin {cases} 2x + 2 & x \ in <-1; 0> \\ 0 & x \ notin <-1; 0> \ end {cases} $ <br> Specify:

L21: ### 2. a)

L22: $ F (x) $,

L23: watch out for x <-1 because '<-' is in the assignment line

L24: 0 for x <= 0

L25: 1 for x> 1

L26: points on the x-axis

L27: values of f (x)

L28: draw dots (cex is the size)

L29: x ^ 2 + 2x + 1

L30: 0 for x <= 0

L31: 1 for x> 1

L32: points on the x-axis

L33: values of f (x)

L34: draw dots (cex is the size)

L35: ### 2. b)

L36: P (−2 ≤ X ≤ .50.5), P (−2 ≤ X ≤ −1), P (X> 0.5), P (X = 0.3)

L37: P (−2 ≤ X ≤ .50.5)

L38: P (−2 ≤ X − −1)

L39: P (X> 0.5)

L40: This will not always work

L41: P (X = 0.3)

L42: it is clear that this probability is 0

L43: corresponds to the integral sa = b, ie with zero size on the x-axis

L44: ### 2. c)

L45: mean, variance and standard deviation of the random variable X.

L46: E (X)

L47: we integrate only where we know that f (x) is nonzero

L48: E (X ^ 2)

L49: we integrate only where we know that f (x) is nonzero

L50: D (X)

L51: sigma (x)

L52: ### 2. d)

L53: mode $ \ hat {x} $

L54: mode = 0

L55: ### 2. e)

L56: median $ x\_ {0.5} $

L57: points on the x-axis

L58: first element zx for which F (x)> = 0.5

L59: ## Example 3.

L60: The random variable Y is defined as: Y = 3X + 1, where X is the random variable from the previous example. Specify:

L61: ### 3. a)

L62: $ F\_Y (y) $

L63: calculated from the relation FY (y) = P (Y <y) = P (3X + 1 <y) = ...

L64: points on the x-axis

L65: ### 3. b)

L66: $ f\_Y (y) $

L67: derivation of F\_Y

L68: 0 for x <-2

L69: 1 for x> 1

L70: total integral check

L71: points on the x-axis

L72: ### 3. c)

L73: E (Y), D (Y), σ (Y)

L74: E (Y)

L75: we integrate only where we know that f (y) is nonzero

L76: alternatively

L77: E (Y ^ 2)

L78: we integrate only where we know that f (y) is nonzero

L79: D (Y)

L80: alternatively

L81: sigma (Y)

L82: ## Example 4. (not from collection)

L83: Calculate $ \ omega $ such that the random variable X with probability density: <br> $ f (x) = \ begin {cases} 0 & x <0 \\ 3e ^ {- 3x} & x \ geq 0 \ end {cases} $ <br> was 0.3 greater than $ \ omega $

L84: 0 for x <= 0

L85: points on the x-axis

%%%%%%%%%% cv10

L0: # Exercise 10. Introduction to hypothesis testing, one-sample tests ## Michal Béreš, Martina Litschmannová

L1: # From interval estimates to hypothesis tests

L2: ## What is a statistical hypothesis test?

L3: Consider the following: - random variable X (for example, height of men) - selection from a random variable (measuring the height of 30 men) - alternative hypotheses For example: <br> $ H\_0 $: $ \ mu\_X = 175 $ <br> $ H\_A $: $ \ mu\_X> 175 $ <br> Since this is a statistical decision, it will always be tied to some level of significance $ \ alpha $. We can always reach only 2 different decisions: - I reject $ H\_0 $ in favor of $ H\_A $ - that is, I say that $ H\_0 $ does not apply - this decision is with the maximum error $ \ alpha $ (level of significance, type I error ) - this means that we are able to influence the size of this error - I do not reject $ H\_0 $ - this means that I claim that due to the obtained data (selection) $ H\_0 $ cannot be refuted - this decision is with error $ \ beta $ (error II This type is not directly controllable and depends on the type of test used. How hypothesis tests are related to interval estimates and how the level of significance enters them will be shown in the next section.

L4: ## Interval estimation and significance level

L5: width of graphs in Jupyter

L6: graph matrix 1x2

L7: oblique

L8: sharpness

L9: normality test

L10: We make a 95% interval estimate of the mean using a t-test:

L11: Now imagine that we want to test the hypothesis: <br> $ H\_0 $: $ \ mu = 100 $ <br> $ H\_A $: $ \ mu \ neq 100 $ <br> What would be the decision with respect to the calculated IO and so the significance level $ \ alpha = 0.05 $?

L12: Let's further imagine that we want to test the hypothesis: <br> $ H\_0 $: $ \ mu = 105 $ <br> $ H\_A $: $ \ mu \ neq 105 $ <br> What would be the decision with respect to the calculated IO and so the significance level $ \ alpha = 0.05 $?

L13: \*\* What we just did is called the classic test. \*\* <br> We'll show you the classic tests for one-sided alternatives. <br> $ H\_0 $: $ \ mu = 105 $ <br> $ H\_A $: $ \ mu> 105 $ <br>

L14: $ H\_0 $: $ \ mu = 105 $ <br> $ H\_A $: $ \ mu <105 $ <br>

L15: Note that the first of these one-sided alternatives led to a "rejection" of $ H\_0 $. This is because of the comparison of the unlikely $ H\_0 $ with the even less likely $ H\_A $.

L16: #### Net significance test and connection with IC

L17: An alternative to the classical test (where we create IO - in the terminology of classical tests the so-called field of admission and its addition to the R critical field) is the so-called pure significance test:

L18: H\_0: mu = 105

L19: H\_A: mu <> 105

L20: The net significance test results in a p-value. Based on it, we decide whether or not to reject $ H\_0 $. <br> p-value can be understood as the highest possible level of record, such that our decision is - I do not reject. Thus, the IO / field of acceptance would contain the examined value:

L21: H\_0: mu = 105

L22: H\_A: mu <> 105

L23: H\_0: mu = 105

L24: H\_A: mu> 105

L25: H\_0: mu = 105

L26: H\_A: mu <105

L27: ## Test overview

L28: ### Position measures

L29: By position measures we mean the data that determines the position of the data, no matter how scattered. For data from the normal distribution we can estimate the mean value, for others the median.

L30: #### a) student's t-test

L31: - we test the mean value - the data must come from a normal distribution - exploratory: skewness and sharpness lie in (-2,2) - exploratory: QQ graph has points approximately on the line - exactly: using a statistical test, eg Shapiro-Wilk test (shapiro.test (data))

L32: H\_0: mu = 100

L33: H\_A: mu <> 100

L34: H\_0: mu = 100

L35: H\_A: mu> 100

L36: H\_0: mu = 100

L37: H\_A: mu <100

L38: #### b) Wilcoxn test

L39: - we test the median - the data must come from a symmetric distribution - exploratory: the skewn lies in (-2,2) - exploratory: the histogram looks approximately symmetrical - exact: using a statistical test, eg the "lawstat" package, the "symmetry.test function" (data, boot = FALSE) "

L40: H\_0: X\_0.5 = 100

L41: H\_A: X\_0.5 <> 100

L42: H\_0: X\_0.5 = 100

L43: H\_A: X\_0.5> 100

L44: H\_0: X\_0.5 = 100

L45: H\_A: X\_0.5 <100

L46: #### c) sign test test

L47: - we test the median - select a larger range (> 10) - requires the "BSDA" library - as the most robust test, it can also be used for discontinuous data - eg order in a list

L48: H\_0: X\_0.5 = 100

L49: H\_A: X\_0.5 <> 100

L50: H\_0: X\_0.5 = 100

L51: H\_A: X\_0.5> 100

L52: H\_0: X\_0.5 = 100

L53: H\_A: X\_0.5 <100

L54: ### Variability measures

L55: By measures of variability we mean the data determining the scatter / variability of the data, regardless of the total values. For data from the normal distribution, we can estimate the standard deviation.

L56: #### standard deviation test

L57: - we test the standard deviation - the data must come from a normal distribution - exploratory: skewness and sharpness lie in (-2,2) - exploratory: QQ graph has points approximately on the line - exactly: using a statistical test, eg Shapiro-Wilk test (shapiro.test (data)) - requires package "EnvStats" - function in Rku, compares variance !!!

L58: H\_0: sigma = 10

L59: H\_A: sigma <> 10

L60: H\_0: sigma = 10

L61: H\_A: sigma> 10

L62: H\_0: sigma = 10

L63: H\_A: sigma <10

L64: ## Probability of occurrence with one selection

L65: #### IO probability

L66: - we test the probability - we require a sufficient number of data: $ n> \ frac {9} {p (1-p)} $ - Clopper - Pearson's estimate (binom.test) - does not take data as a parameter, but the number of successes and the number of observations

L67: H\_0: pi = 0.2

L68: H\_A: pi <> 0.2

L69: H\_0: pi = 0.2

L70: H\_A: pi> 0.2

L71: H\_0: pi = 0.2

L72: H\_A: pi <0.2

L73: # Examples

L74: ## Example 1.

L75: We have a selection of 216 patients and we measured their protein serum (file testy\_jednovyberove.xlsx list bilk\_serum). Verify that the average protein serum (Albumin) of all patients of this type (population average µ) differs statistically significantly from 35 g / l.

L76: Read data from xlsx file (using readxl package)

L77: Exploratory analysis

L78: sd is rounded to 3 valid digits

L79: sd and position rates are rounded to the nearest thousandth

L80: \*\* Position test \*\*

L81: Verification of normality - exploratory

L82: oblique

L83: sharpness

L84: width of graphs in Jupyter

L85: graph matrix 1x2

L86: We will use the normality test for the final decision on data normality.

L87: The presumption of normality is verified by the Shapirov-Wilkov test.

L88: H0: Data is a selection from a normal distribution.

L89: Ha: Data is not a selection from the normal distribution.

L90: p-value> 0.05 -> Na hl. significance of 0.05, the assumption of normality cannot be rejected.

L91: normal OK -> t.test

L92: H0: mu = 35 g / l

L93: Ha: mu <> 35 g / l

L94: p-value <0.05 -> Na hl. significance of 0.05 we reject the null hypothesis

L95: in favor of the alternative hypothesis

L96: The mean albumin value differs statistically significantly from 35 g / l.

L97: ## Example 2.

L98: Survival times for 100 lung cancer patients treated with the new drug are listed in the test\_jednovyberove.xlsx list previti file. It is known from previous studies that the average survival of such patients without the administration of a new drug is 22.2 months. Can these data suggest that the new drug prolongs survival?

L99: Reading data from an xlsx file (using the readxl package)

L100: # Exploratory analysis

L101: matrix of 1x2 graphs

L102: \*\* Data contains OP -> we can delete it. Or note that this is probably an exponential distribution and the OPs are not actually there (the division simply behaves this way.) \*\*

L103: Data contains remote observations. We can list them with the help of f-ce boxplot.

L104: if we decided to remove outliers, then

L105: We recommend that you do not overwrite the original data

L106: # Exploratory analysis for data without remote observations

L107: sd is rounded to 3 valid digits

L108: sd and position measures round. to tenths

L109: \*\* Position measure (mean / median) test \*\*

L110: Verification of normality - exploratory

L111: graph matrix 1x2

L112: QQ - graph and history show that the choice of truth. is not a choice of standards. distribution.

L113: Both skew and sharpness comply with standards. distribution.

L114: we will use the normality test.

L115: We verify the assumption of normality by the Shapirs. Wilkov's test.

L116: p-value <0.05 -> Na hl. significance 0.05, we reject the assumption of normality

L117: exploratory assessment of symmetry - historical height and skewness

L118: Assumption of symmetry - verification by test

L119: H0: data comes from symmetric distribution

L120: HA: ~ H0

L121: p-value <0.05 -> Na hl. significance 0.05 we reject the assumption of symmetry

L122: normality rejected -> symmetry rejected -> Sign. test

L123: H0: median = 22.2 months

L124: Ha: median> 22.2 months

L125: p-value> 0.05 -> Na hl. significance of 0.05, the null hypothesis cannot be rejected

L126: Median survival time is not statistically significantly greater than 22.2 months.

L127: H0: median = 22.2 months

L128: Ha: median <22.2 months

L129: ## Example 3.

L130: The machine produces piston rings of a given diameter. The manufacturer states that the standard deviation of the ring diameter is 0.05 mm. To verify this information, 80 rings were randomly selected and a standard deviation of 0.04 mm in diameter was calculated. Can this difference be considered statistically significant in terms of improving the quality of production? Verify with a clean significance test. Assume that the diameter of the piston rings has a normal distribution.

L131: Standard deviation test

L132: We assume data normality (according to assignment)

L133: file range

L134: mm .... sample standard deviation (point estimate of standard deviation)

L135: H0: sigma = 0.05 mm

L136: Ha: sigma <0.05 mm

L137: p.value <0.05 -> At the significance level of 0.05 we reject the null hypothesis

L138: in favor of an alternative hypothesis

L139: Direction. the ring diameter deviation is statistically significantly less than 0.05 mm.

L140: ## Example 4.

L141: The machine produces piston rings of a given diameter. The manufacturer states that the standard deviation of the ring diameter is 0.05 mm. To verify this information, 80 rings were randomly selected and their diameter was measured (file testy\_jednovyberove.xlsx list krouzky). Can the results obtained be considered statistically significant in terms of improving the quality of production? Verify with a clean significance test.

L142: Reading data from xlsx file (using readxl package)

L143: # Exploratory analysis

L144: Data contains remote observations. We can list them with the help of f-ce boxplot.

L145: if we decided to remove outliers, then

L146: Exploratory analysis for data without remote observations

L147: sd is rounded to 3 valid digits

L148: sd and position measures round. per thousandths

L149: Verification of normality - exploratory

L150: graph matrix 1x2

L151: Both skew and sharpness comply with standards. distribution.

L152: We will use for the final decision on data normality

L153: normality test.

L154: We verify the assumption of normality by the Shapirs. Wilkov's test.

L155: p-value> 0.05 -> Na hl. significance of 0.05 cannot be assumed norms. reject

L156: variability test -> variance test

L157: H0: sigma = 0.05 mm

L158: Ha: sigma <0.05 mm

L159: p-value <0.05 -> At the significance level of 0.05, we reject H0 in favor of Ha

L160: How to find a 95% interval estimate of the standard deviation?

L161: ## Example 5.

L162: TT states that 1% of their resistors do not meet the required criteria. 15 unsuitable resistors were found in the tested delivery of 1000 pieces. Does this result confirm TT's assertion? Verify with a clean significance test.

L163: selection range

L164: number of "successes"

L165: relative frequency (probability point estimate)

L166: Verification of assumptions

L167: We further assume n / N <0.05, ie that the given population (resistors) has a range

L168: at least 1000 / 0.05 = 1000 \* 20 = 20,000 resistors

L169: # Clopper - Pearson (exact) test

L170: # H0: pi = 0.01

L171: # Ha: pi <> 0.01

L172: # Clopper - Pearson (exact) test

L173: # H0: pi = 0.01

L174: # Ha: pi> 0.01

L175: At the significance level of 0.05 we do not reject H0

L176: The share of defective resistors in production cannot be expected to be statistically significant

L177: exceeds 1%.

%%%%%%%%%% cv7

L0: # Exercise 7. Data preprocessing and exploratory analysis ## Adéla Vrtková, Martina Litschmannová, Michal Béreš

L1: # 1. Expansion feature packs - installation and loading

L2: You only need to install packages once (if you don't already have them)

L3: install.packages ("readxl")

L4: install.packages ("dplyr")

L5: install.packages ("openxlsx")

L6: Loading the package (must be repeated every time Rka is run, it is advisable to have it at the beginning of the script)

L7: contains notifications of overwritten functions or older versions of the package

L8: # 2. Working directory - where we load and where we store data

L9: - Attention, the current open folder in Rstudio, or the location of the Rskcript is not automatically a working directory

L10: Working directory listing

L11: Working directory settings -> in quotation marks, full path (relative or absolute)

L12: Where are we now?

L13: back again

L14: control

L15: # 3. Load data file

L16: ## From CSV file

L17: Basic functions - read.table, read.csv, read.csv2, ... It depends mainly on the file format (.txt, .csv), on the so-called separator of individual values, decimal point / dot

L18: Load and save a data file in csv2 format from the working directory

L19: Load and save a csv2 data file from the local disk to the data frame

L20: Load and save a csv2 data file from the Internet to the data frame

L21: ## From Excel (xlsx file)

L22: Loading and saving a data file in xlsx format from the local disk to the data frame We use the function from the readxl package, which we expanded in the introduction

L23: worksheet specification in xlsx file

L24: lines to be skipped

L25: ## Remove unnecessary rows / columns and name rows / columns for easier data addressing

L26: Indexing with negative indexes returns everything except the index value

L27: do not mix negative and positive indices!

L28: delete the first column with indexes

L29: Rename columns - if necessary

L30: #### Note (which is good to read until the end ....)

L31: (in Rstudio) it is possible to import using "Import Dataset" from the Environment window without having to write the code. In this case, however, there must be no special characters (hooks, commas) in the "path" to the file. Otherwise, an error will appear. The object imported this way will be in the new RSstudio as type "tibble". This is a more modern "data.frame" and in some features it can cause problems and throw errors! You can easily convert this object to type data.frame using \*\* as.data.frame () \*\* If you have a problem with a function that does not take a column from "tibble" as non-numeric output, you can fix it with the command pull: data [, 1] replace pull (data, 1)

L32: # 4. Pre-processing data + Dplyr library

L33: ### Overview of Dplyr library functions

L34: - \*\*%>% \*\* is a so-called pipe operator, a typical use is "res = data%>% operation", where the result is a data-driven operation - \*\* select (...) \*\* is one of the operations which we can insert into the "pipe" operator - used to select data

L35: - select (1) - selects the first column

L36: - select (A5) - selects the column named A5

L37: - select (1,3,5) - selects columns 1,3,5

L38: - \*\* mutate (new\_column = ...) \*\* is an operation that produces a new data column in the data frame using the specified calculation over the current columns

L39: - data%>% mutate (C = AB) produces a new column named "C" in the "data" data frame as the difference of the values in the existing columns "A" and "B"

L40: - \*\* filter (...) \*\* filters values from the data that meet the specified requirements

L41: - data%>% filter (manufacturer == "A" | manufacturer == "B") returns a data file that has only "A" or "B" values in the "manufacturer" column

L42: - data%>% filter (manufacturer == "A", values> 1000) if we write the requirements in a row (separated by a comma) we understand it as and at the same time

L43: - \*\* summarize (...) \*\* calculate the prescribed numerical characteristics within the specified columns (suitable for combination with group.by)

L44: - data%>% summarize (prum = mean (kap5), median = median (kap5))

L45: - \*\* arrange (...) \*\* ascending or descending row order

L46: - data%>% arrange ascending

L47: - data%>% arrange (desc) descending

L48: - \*\* group\_by (...) \*\* grouping of data according to unique values in the specified column

L49: - data%>% group\_by (manufacturer)

L50: Very useful Dgasr "cheat sheet" can be found here: https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf

L51: ### Column / row selections

L52: Data file listing

L53: Display of the first six lines

L54: Display of the last six lines

L55: Display of line 10

L56: Display of the 3rd column - several ways

L57: or (if we know the name of the variable written in the 3rd column)

L58: or using the dplyr package select function, which selects the selected columns

L59: <hr>

L60: Save the first and fifth columns of data. frames data to data. framework attempt

L61: or using the dplyr function

L62: or by name

L63: <hr> Exclude data from the file.

L64: Exclude the first and fifth columns from the data. data frames and data storage. framework attempt

L65: or using dplyr

L66: or by name

L67: <hr> Modification of data into several smaller logical units with different structure Note. when storing data, we think of clarity in names

L68: ### Basic conversion of a simple data matrix into a standard data format - stack (...)

L69: from the data we select those columns that correspond to measurements after 5 cycles

L70: Rename columns

L71: and transfer to st. data format

L72: and edit the column names once more

L73: Do the same for measurements performed after 100 cycles

L74: we select from the data those columns that correspond to measurements after 100 cycles

L75: Rename columns

L76: and transfer to st. data format

L77: and edit the column names once more

L78: Finally, we will create a data file in st. data format with all data

L79: merge "by columns"

L80: omit the extra second column

L81: omit rows with NA values

L82: \*\* !!! Handle the na.omit function extremely carefully so that you do not inadvertently lose data !!! \*\*

L83: <hr>

L84: ### Defining new columns in a data frame

L85: Defining a new drop variable

L86: or using a function from the dplyr package

L87: ### Select data from standard data format

L88: May be useful - create separate variables

L89: Class (type) numeric

L90: as follows with a data frame result

L91: filters rows corresponding to manufacturer A

L92: Selects only the values in column kap5,

L93: Other separate variables (only one method specified)

L94: ### More detailed window for Dplyr library functions - work on data in standard data format

L95: It is necessary to apply to data in st. data format !!! Pipe operator%>% - helps with chaining functions - in the new RSstudio shortcut key Ctrl + Shift + M

L96: #### filter - applies a filter to the given column

L97: filter - selects / filters rows based on given conditions

L98: Selection of products from the manufacturer

L99: Selection of products from manufacturer A or B

L100: separating conditions correspond to the logical "or"

L101: Selection of all products with a decrease of 200 mAh and more from the manufacturer C

L102: comma separating conditions corresponds to logical "and at the same time"

L103: #### mutate - produce a new column

L104: mutate - adds a new variable or transforms an existing one

L105: Creating a new column drop\_Ah, which indicates the drop in capacity in Ah (original data in mAh, 1 Ah = 1000 mAh)

L106: Attention! if we do not save the result with the new column, it will only be printed and disappear

L107: #### summarize - generates summary characteristics of various variables

L108: Calculation of the mean and median of all values of the variable kap5

L109: #### arrange - sorts rows according to the selected variable

L110: Ascending and descending order of rows according to the decrease value

L111: #### group\_by - groups values into groups according to the selected variable

L112: the table is "virtually" divided into groups for later processing, eg summarize

L113: Ideal for calculating summary characteristics for each manufacturer separately, eg average

L114: \*\* Final note on dplyr (which is good to finish until the end ...) Some operations may throw a "tibble" object. This is a more modern data.frame, however it can cause problems and cause error messages in some functions! You can easily convert this "tibble" object to data.frame using as.data.frame (). \*\*

L115: # 5. Data conversion to standard data format (for the two most common data formats)

L116: ## From data in Data Matrix format

L117: ### Reshape function

L118: Its parameters: - \*\* data \*\* - data to be converted must be fe format data.frame (as.data.frame (data)) - \*\* direction \*\* - which direction we want to do the transformation - "long" - to standard format - "wide" - back to the data matrix - \*\* varying \*\* - column names that indicate the same data for different categories - it is a sheet of vectors - each item of the sheet is one measurement - each vector is then a list of columns - \*\* v.names \*\* - column names in st. give. format - the number of names must match the number of vectors in varying - \*\* times \*\* - the names of the individual categories - ATTENTION !! must be in the same order as for the variable varying - \*\* timevar \*\* - column name with categories

L119: and if we want, we can convert the data back

L120: ## From a data file where the categories are in individual Excel sheets

L121: # 6. Exploratory analysis and visualization of a categorical variable

L122: ### Notes on graphics in R

L123: the basis is the so-called high-level functions, which create a graph (ie open the graphics window and draw according to the specified parameters) followed by the so-called low-level functions, which add something to the active graphics window, do not open new low-level functions - eg abline, points, lines, legend, title, axis ... which add a line, points, legend ... ie. before using the "low-level" function it is necessary to call the "high-level" function (eg plot, boxplot, hist, barplot, pie, ...) Further graphic parameters can be found in the help or eg here http: // www .statmethods.net / advgraphs / parameters.html or here https://flowingdata.com/2015/03/17/r-cheat-sheet-for-graphical-parameters/ or http://bcb.dfci.harvard.edu /~aedin/courses/BiocDec2011/2.Plotting.pdf Colors in R http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf https://www.nceas.ucsb.edu/~frazier /RSpatialGuides/colorPaletteCheatsheet.pdf Saving graphs is possible using the function dev.print, jpeg, pdf and others. More easily in the Plots -> Export window

L124: Table of absolute frequencies of the manufacturer's categorical variable ...

L125: listing - object of type "table" - usually more suitable, but more difficult conversion to type data.frame

L126: ... and using dplyr functions (more complex)

L127: number of products for each manufacturer

L128: listing - object type "tibble" - useful when we need to simply convert to data.frame

L129: ### Relative frequency table

L130: Direct calculation

L131: statement

L132: or using the prop.table function

L133: statement

L134: or using the dplyr functions, where absolute frequencies will also be included

L135: statement

L136: For all tables, rounding and the associated risk of rounding error must be observed.

L137: The procedure for frequency and frequency2 is the same.

L138: rounded to 1 decimal place

L139: rounding error monitoring

L140: The procedure for table\_abs\_rel is different due to a different format (tibble)

L141: #### Create table with absolute and rel. frequencies (without gas). We have:

L142: merge tables

L143: change column names

L144: #### Save table to csv file

L145: Where is the table stored? It is stored in the working directory without specifying the complete path in the previous command.

L146: <hr>

L147: ### Visualization using graphs

L148: Bar graph

L149: The basic (ie does not require any package) bar graph is based on the frequency table we have prepared

L150: simple division of the graphics window - 1 row, 1 column

L151: margins around each of the graphs in line numbers - - c (bottom, left, top, right)

L152: outer margins in number of lines - c (bottom, left, top, right)

L153: Change colors, add name

L154: alt. a vector of specific colors can be chosen, eg c ("blue", "yellow," red "," green ")

L155: or other scales (heat.colors, topo.colors, terrain.colors and many others)

L156: the space parameter creates a space between the columns

L157: Add additional captions and legends

L158: horizontal graph orientation

L159: does not draw a line around the bars

L160: The paste0 function allows you to merge text strings and variable values, the "\ n" symbol forms a new line in the text

L161: placing a legend next to a bar graph is very tricky

L162: it is much easier to work with ggplot2 in this case

L163: Add absolute and relative frequencies to the corresponding columns

L164: parameter pos specifies where the text will be given with respect to the given position (1 = below, 2 = left, 3 = above, 4 = right)

L165: Try to use the previous code and create a bar graph for the Manufacturer variable by yourself.

L166: # 7. Exploratory analysis and visualization of a quantitative variable

L167: Descriptive statistics

L168: Calculation of the average of one variable

L169: Beware of missing values

L170: Calculation of the median of one variable

L171: Range determination

L172: #### Other characteristics -> var (), sd (), min (), max (), ...

L173: Attention! The functions for calculating skewness and kurtosis are not part of the basic R, you will find them in the package moments. sharpness in the interval (1,5) To standardize the sharpness, it is necessary to subtract 3 from the calculated value. If you write the package name and "::" before the function name, you will ensure that the function from the given package will be used. packages have different functions under the same name

L174: install.packages ("moments")

L175: If we want to calculate the given characteristic for variable capacity after 5 cycles

L176: according to the manufacturers, we can use the tapply function

L177: or using dplyr - here pay attention to automatic (not always correct rounding)

L178: To simplify the work, we can use the dplyr function and put all the characteristics in one table

L179: without using group\_by for the whole variable kap5

L180: preventive at.rm = T

L181: coefficient of variation in percent

L182: precautionary package specification moments

L183: Don't forget the correct rounding!

L184: We use group\_by and get the characteristics for the capacity after 5 cycles according to the manufacturers

L185: Due to the incomplete statement, it is advisable to save the output and view it in a new window

L186: coefficient of variation in percent

L187: ### Box chart

L188: \*\* We draw for original data, we can add rendering for data without OP. \*\*

L189: Simple and fast rendering using the basic function only for manufacturer C

L190: Further modification of the graph, use of the points function to display the average

L191: adds a point showing the average to the existing graph

L192: Horizontal orientation, box width change

L193: When orienting horizontally, the opposite setting of the labels must be observed

L194: changes the box width to 1/2

L195: Use the previous code and create a box chart according to you.

L196: And draw a multiple box chart

L197: graphic parameters can be set similarly to the previous ones

L198: ### Histogram

L199: \*\* We always plot for data without remote observations !! \*\*

L200: Simple and fast rendering

L201: What do the different values of the breaks parameter do with the graph?

L202: Labels, colors and other parameters can be set traditionally

L203: fill color

L204: column border color

L205: adds the absolute frequencies of the given categories in the form of labels

L206: Scale the y-axis to plot the probability density estimate

L207: scaling on the y-axis -> f (x)

L208: Attaches a probability density estimation graph

L209: Generate normal distribution density and add to histogram

L210: Generate values for the x-axis

L211: Generate values for the y-axis

L212: Add a curve to the last graph based on the values generated above

L213: This combined graph can be used for visual assessment of normality.

L214: \*\* Use the previous code and create a histogram of yourself. \*\*

L215: ### QQ-graph

L216: \*\* We always plot for data without remote observations !! \*\*

L217: Simple and very fast rendering ...

L218: ... with adjustment of axis labels ...

L219: For advanced and interested - automation, use of for-cycle, multiple graphs in one image If we use basic functions (barplot, boxplot, histogram), then the function par () or layout () is used. In these functions we specify the structure - how we want to draw more pictures

L220: Eg. we want to draw a histogram and boxplot for capacity after 5 battery cycles from manufacturer A

L221: structure creation

L222: margin size adjustment

L223: Using for-cycle histograms and boxplots for all manufacturers

L224: Combination of histogram and QQ-fence

L225: # 8. Internal walls and identification of remote observations

L226: ## Manual removal by counting the inner walls

L227: data column separation for manufacturer

L228: interquartile margins

L229: calculation of the lower between the inner walls

L230: calculation of the upper between the inner walls

L231: set values that are out of limits to NA

L232: we can delete NA hondots

L233: ## Automatic removal according to the box fence

L234: ### How to do this for standard multi-category data?

L235: \*\* Caution, this needs to be done by category, otherwise we risk deleting the data into the data. the file belongs to !!! \*\*

L236: c ("A", "B", ..)

L237: \*\* The analyst can always say that he will not delete remote observations, but he must include this information in the analysis report! \*\*

L238: # 9 rule 3 $ \ sigma $ and Chebyshev's inequality

L239: ## Empirical verification of normality

L240: \*\* Based on data after deleting remote observations: \*\*

L241: we will use the data from the removal example op

L242: We draw the QQ graph and calculate the skewness and sharpness:

L243: another definition shifted by 3

L244: - the dots in the QQ graph must lie approximately on the line - ie. the quantiles correspond approximately to the quantiles of the normal distribution - skewness must lie in the interval <-2, 2> - kurtosis must lie in the interval <-2,2>

L245: - be careful we have to reduce the result of the R function by 3

L246: \*\* If data normality is met -> rule 3σ \*\* <br> σ: P (µ - σ <X <µ + σ) = 0,6827 <br> 2σ: P (µ - 2σ <X < µ + 2σ) = 0.9545 <br> 3σ: P (µ - 3σ <X <µ + 3σ) = 0.9973 <br> <br> \*\* If data normality is not met -> Chebyshev inequality \*\* < br> σ: P (µ - σ <X <µ + σ) = 0 <br> 2σ: P (µ - 2σ <X <µ + 2σ) = 0.75 <br> 3σ: P (µ - 3σ < X <µ + 3σ) = 0.8889 <br>

L247: # 10. Rounding

L248: Everything needed for rounding can be found on the LMS in the rounding document. https://lms.vsb.cz/pluginfile.php/1298954/mod\_folder/content/0/Leg%C3%A1ln%C3%AD%20tah%C3%A1ky/zaokrouhlovani.pdf <br> The most important: - standard deviation rounded to the prescribed number of digits up (ceiling)

L249: - data file size = <2.10> -> 1 valid digit

L250: - data file size = (10.30> -> 2 valid digits

L251: - data file size = (30,2000> -> 3 valid digits

L252: - position measures (averages, quantiles, ...) are then rounded to the same valid digit as the standard deviation

%%%%%%%%%% cv2

L0: # Exercise 2 - Probability

L1: ## Adéla Vrtková, Michal Béreš, Martina Litschmannová

L2: In this exercise, we will go through an introduction to probability. We assume knowledge from the lecture, especially the terms: \*\* definition of probability, conditional probability, complete probability theorem, Bayes' theorem \*\*.

L3: # Auxiliary functions

L4: ## Total probability

L5: $ P (A) = \ sum\_ {i = 1} ^ {n} P (B\_i) P (A | B\_i) $

L6: probability calculation P (A) - complete probability theorem

L7: we consider P\_B as a vector of values P (B\_i) and P\_BA as a vector of values P (A | B\_i)

L8: ## Bayes' theorem

L9: $ P (B\_k | A) = \ frac {P (B\_k) P (A | B\_k)} {\ sum\_ {i = 1} ^ {n} P (B\_i) P (A | B\_i)} $

L10: calculation of conditional probability P (B\_k | A) - Bayes' theorem

L11: we consider P\_B as a vector of values P (B\_i), P\_BA as a vector of values P (A | B\_i)

L12: \*\* We will add functions from the ammunition exercise for counting combinatorial selections, they are in the combinatorics script.R \*\*

L13: # Examples

L14: ## Example 1.

L15: Determine the probability that a number greater than 14 will fall on a 20-wall fair dice roll.

L16: probability as a proportion favorable to all

L17: ## Example 2.

L18: Determine the probability that a number greater than 14 will fall on a 20-wall dice roll, if you know that even numbers fall twice as often as odd numbers.

L19: probability is

L20: ## Example 3.

L21: Determine the probability that you will guess 4 numbers in the athlete. (6 numbers out of 49 are drawn in sports)

L22: ## Example 4.

L23: From the alphabetical list of students enrolled in the exercise, the teacher selects the first 12 and offers them a bet: “If each of you was born in a different zodiac sign, I will give each of you CZK 100. However, if there are at least two students among you who were born in the same sign, each of you will give me CZK 100. ”Is it worthwhile for students to accept a bet? How likely are students to win?

L24: ## Example 5.

L25: Calculate the probability that an electric current will flow from point 1 to point 2 if part of the el. circuit, including the probability of failure of individual components is indicated in the following figure. (The failures of the individual components are independent of each other.)! [Image.png] (attachment: image.png)

L26: divided into blocks I = (A, B) and II = (C, D, E)

L27: result

L28: ## Example 6.

L29: The enclosure has a rectangular shape, the east and west walls are 40 m long, the south and north walls are 100 m long. A horse runs in this enclosure. What is the probability that it is closer to the south wall than to the other three?

L30: geometric probability

L31: close to the south

L32: probabilities

L33: ## Example 7.

L34: The patient is suspected of having one of four mutually exclusive diseases - N1, N2, N3, N4 with a probability of P (N1) = 0.1; P (N2) = 0.2; P (N3) = 0.4; P (N4) = 0.3. Laboratory test A is positive in the case of the first disease in 50% of cases, in the second disease in 75% of cases, in the third disease in 15% of cases and in the fourth in 20% of cases. What is the probability that the result of the laboratory test will be positive?

L35: complete probability theorem

L36: P (N1), P (N2), ...

L37: P (P | N1), P (P | N2), ...

L38: P (P)

L39: ## Example 8.

L40: Telegraphic characters consist of "dot", "comma" signals. It is statistically found that 25% of "dot" messages and 20% of "comma" signals are distorted. It is also known that signals are used in a 3: 2 ratio. Determine the probability that the signal was received correctly if a "dot" signal was received.

L41: Bayes' theorem

L42: P (O.), P (O-)

L43: P (P. | O.), P (P. | O-)

L44: k = 1 because correctly = O.

L45: ## Example 9.

L46: 85% of green taxis and 15% of blue taxis run in one city. The witness of the traffic accident testified that the accident was caused by the driver of the blue taxi, who then left. Tests carried out under similar lighting conditions showed that the witness identified the color of the taxi well in 80% of cases and was wrong in 20% of cases.

L47: - What is the probability that the culprit of the accident actually drove a blue taxi?

L48: - Subsequently, another independent witness was found who also claims that the taxi was blue. What is the probability now that the culprit of the accident actually drove a blue taxi?

L49: - Does the probability that the perpetrator of the accident actually drove a blue taxi affect whether the two witnesses mentioned above testified gradually or simultaneously?

L50: a) again Bayes' theorem

L51: P (Z), P (M)

L52: P (SM | Z), P (SM | M)

L53: blue is second

L54: b) first option - second pass through Bayes

L55: P (Z), P (M)

L56: P (S2M | Z), P (S2M | M)

L57: c) or answered at once

L58: P (Z), P (M)

L59: P (S1M & S2M | Z), P (S1M & S2M | M)

L60: ## Example 10.

L61: We need to find out the answer to a sensitive question. How to estimate what percentage of respondents will answer YES to the question and at the same time guarantee complete anonymity to all respondents? One of the solutions is the so-called double-anonymous survey: <br> We will let the respondents throw the crown and the double crown and those who fell on the crown will write the answer (YES / NO) to the sensitive question on the card. Other respondents will write if they fell on their double crown (YES / NO). How do we determine the proportion of students who answered YES to a sensitive question? <br> Assume that respondents were asked if they were cheating on an exam. From the questionnaires, it was found that 120 respondents answered "YES" and 200 respondents answered "NO". What percentage of students cheated at the exam?

L62: complete probability theorem

L63: P (A) = P (K\_lic) \* P (A | K\_lic) + P (K\_rub) \* P (D\_lic | K\_rub)

L64: equation 120/320 = 0.5 \* x + 0.5 \* 0.5

L65: ## Bonus - Monty Hall Problem

L66: Let's start with the generation of n instances of the competition - the price will be a random index of the door (1,2,3) behind which the price can be

L67: number of attempts

L68: random door selection

L69: head draws the first 6 elements / lines

L70: Same for our original choice - random door index.

L71: original selection

L72: In the first round, the moderator opens one empty door, this is how it can be simulated:

L73: vector initialization

L74: auxiliary variable - door identifiers

L75: initialization

L76: We must not open the door with the price

L77: not even our chosen door

L78: in the rest there are either 2 (if we hit) or 1 door (if not)

L79: if we open one

L80: if 2, we randomly select one and open it

L81: Our new choice if we decide - the sum of the indices is 1 + 2 + 3 = 6 so if we have an index selected, then some index will open, then the rest of the 6 are the third = our new choice.

L82: Success in the original selection:

L83: Replacement success:

%%%%%%%%%% cv12

L0: # Exercise 12. Multiselective tests ## Michal Béreš, Martina Litschmannová, Adéla Vrtková

L1: ## Test data for a function call example

L2: I will produce frame data with one column from give

L3: I will rename the column name

L4: I will add a type for all frame data

L5: I glue the lines together

L6: Convert type to type factor

L7: if there are any OPs, I will ignore them

L8: (I know the data is from a normal distribution!)

L9: (I also know they have the same variance)

L10: # Overview of functions

L11: ## Variability measures

L12: ### Bartlett test

L13: - verifies the agreement of variances - $ H\_0: \ sigma ^ 2\_1 = \ sigma ^ 2\_2 = \ sigma ^ 2\_3 = \ ldots $ - $ H\_A: \ neg H\_0 $ - assumption is normality of data (and of course independence and continuity)

L14: ### Levene's test

L15: - verifies the agreement of variances - $ H\_0: \ sigma ^ 2\_1 = \ sigma ^ 2\_2 = \ sigma ^ 2\_3 = \ ldots $ - $ H\_A: \ neg H\_0 $ - only independence and continuity are assumed

L16: ### Cochran's and Hartley's test

L17: - they also verify the agreement of variances - they require data normality and so-called balanced sorting - balanced sorting means that we have approximately the same amount of data in each group - we will not use them

L18: ## Position measures

L19: ### ANOVA (analysis of variance)

L20: - verifies the position match (mean values) - $ H\_0: \ mu\_1 = \ mu\_2 = \ mu\_3 = \ ldots $ - $ H\_A: \ neg H\_0 $ - assumptions: - data normality - homoskedasticity (identical variances) - (and of course independence and continuity) - if we reject $ H\_0 $ Post-Hoc analysis is required - using TukeyHSD test

L21: Basic ANOVA

L22: H0: mu1 = mu2 = mu3 = mu4

L23: HA: ~ H0 (H0 negation)

L24: Post-Hoc analysis

L25: effect counting

L26: overall average

L27: averages in groups

L28: effects

L29: list sorted

L30: ### Kruskal - Wallis test

L31: - verifies position match (medians) - $ H\_0: X\_ {0.5,1} = X\_ {0.5,2} = X\_ {0.5,3} = \ ldots $ - $ H\_A: \ neg H\_0 $ - assumptions: - symmetry data - (and of course independence and continuity) - if we reject $ H\_0 $ Post-Hoc analysis is required - using the Dunn test / method

L32: Basic KW test

L33: H0: X0.5,1 = X0.5,2 = X0.5,3 = X0.5,4

L34: HA: ~ H0 (H0 negation)

L35: Post-Hoc analysis

L36: altP = T sets the p-value so that when making a decision

L37: on statistical significance compared to alpha

L38: (default: altp = FALSE, then compare with alpha / 2)

L39: install.packages ("dunn.test")

L40: effect counting

L41: total median

L42: medians in groups

L43: effects

L44: list sorted

L45: # Examples

L46: ## Example 1.

L47: We test the null hypothesis µ1 = µ2 = µ3. It was found that the data we have available are selections from the normal distribution that satisfy the assumption of homoskedasticity (agreement of variances). Based on the data obtained by the exploratory analysis, complete the ANOVA table and the resulting conclusions. <br>! [Image.png] (attachment: 4429db71-6c58-4bf9-b61f-fd9ec4015d20.png)

L48: selection ranges

L49: averages in individual groups / classes

L50: direction. deviations in individual groups / classes

L51: total range of selections

L52: number of classes

L53: number of degrees of freedom - intergroup

L54: number of degrees of freedom - residual

L55: overall average (using weighted average)

L56: intergroup sum of squares

L57: residual sum of squares

L58: total squares

L59: variance between groups / classes

L60: variance within groups / classes

L61: F-ratio

L62: p-value

L63: At the significance level of 0.05, we reject the hypothesis of agreement of the mean values

L64: ie the mean values of at least one pair of groups become. significantly different.

L65: group effects estimates

L66: Group 2 shows the most below-average results compared to the overall average

L67: (about 10 units lower than the overall average). On the contrary, the average of group 3 is

L68: about 10 units higher than the overall average. Average results of group 1

L69: correspond to the overall average.

L70: ## Example 2.

L71: 122 patients who underwent heart surgery were randomly divided into three groups. <br> \*\* Group 1: \*\* Patients received 50% nitrous oxide and 50% oxygen mixed continuously for 24 hours. <br> \*\* Group 2: \*\* Patients received 50% nitrous oxide and 50% oxygen mixed only during surgery. <br> \*\* Group 3: \*\* Patients received no nitrous oxide but received 35-50% oxygen for 24 hours. <br> The data in the folic acid.xls file correspond to the folic acid salt concentrations in the red blood cells in all three groups after 24 hours of ventilation. Verify that the observed differences between the folic acid salt concentrations are statistically significant, ie that there is an effect of the composition of the mixture on the monitored parameter.

L72: rename columns

L73: conversion to standard data format

L74: Data do not contain remote observations.

L75: we test normality using S.-W. test

L76: Information needed to set up rounding

L77: sd is rounded to 3 valid digits

L78: sd and position measures are rounded to tenths

L79: Verification of variance agreement

L80: sampling variances

L81: According to the box chart and information on the ratio of the largest and smallest

L82: variances (<2) we do not assume that the variances differ statistically significantly

L83: The presumption of normality was not rejected -> Bartlett's test

L84: At the significance level of 0.05, the assumption of agreement of variances cannot be rejected

L85: (Bartlett test, x\_OBS = 0.878, df = 2, p-value = 0.645).

L86: We want to compare the mean values of independent samples from normal distributions

L87: with same variances -> ANOVA

L88: The aov () command requires data in the standard data format

L89: At the significance level of 0.05, we reject the hypothesis of agreement of the mean values

L90: (ANOVA, p-value << 0.001) -> multiple comparison

L91: post-hoc analysis

L92: effect counting

L93: overall average

L94: averages in groups

L95: effects

L96: list sorted

L97: If we consider a high folic acid content to be positive, then statistically

L98: Group 1 patients achieved significantly best results (average content

L99: folic acid about 27 units higher than av. folic acid content in the blood

L100: all patients tested) and statistically significantly worst results

L101: achieved by patients from group 2 (average folic acid content by about 26 units

L102: lower than average folic acid content in the blood of all tested patients).

L103: Folic acid content in the blood of group 3 patients

L104: corresponds to the overall average. All three groups of patients are according to each other

L105: folic acid content in the blood statistically significantly different.

L106: ## Example 3.

L107: Three breeds of rabbits are bred on the farm. An experiment was performed on kralici.xls, the aim of which was to find out whether, even if we keep and fatten all rabbits for the same time and under the same conditions, there is a statistically significant (conclusive) difference between breeds in rabbit weights. Verify.

L108: rename columns

L109: conversion to standard data format

L110: data contains OP

L111: Eliminate remote observation

L112: Box chart

L113: At the significance level of 0.05, we do not reject the assumption of normality.

L114: Information needed to set rounding

L115: sd is rounded to 2 valid digits

L116: sd and position measurements zaok. to hundredths (unification across species of rabbits)

L117: Verification of variance agreement

L118: According to the box chart and information on the ratio of the largest and smallest dispersion.

L119: (close to 2, but selection range <30) it is more difficult to estimate whether it is possible

L120: assume variance agreement. The test will help us decide.

L121: The presumption of normality was not rejected -> Bartlett's test

L122: At the significance level of 0.05, the assumption of agreement of variances cannot be rejected

L123: (Bartlett test, x\_OBS = 3.1, df = 2, p-value = 0.217).

L124: We want to compare the mean values of independent selections from normal ones

L125: distribution with same variances -> ANOVA

L126: The aov () command requires data in the standard data format

L127: At the significance level of 0.05, we reject the hypothesis of agreement of the mean values

L128: (p-value << 0.001, ANOVA) -> multiple comparison

L129: post-hoc analysis

L130: effect counting

L131: overall average

L132: averages in groups

L133: effects

L134: list sorted

L135: ## Example 4.

L136: The competition for the best product quality was sent to four manufacturers A, B, C, D by a total of 66 products. The jury compiled the order (only the order of the product from best to worst), which is listed in the file quality.xls. On the basis of the above data, assess whether the origin of the products affects its quality.

L137: rename columns

L138: data is already in standard format

L139: Verification of normality does not make sense - by nature it is a dis. data-order

L140: Information needed to set rounding

L141: sd is rounded to 2 valid digits

L142: sd and position measures are rounded to integers

L143: Variation compliance verification

L144: According to the box chart and information on the ratio of the largest and smallest

L145: Variance (<2) from assuming variance agreement.

L146: (Kruskal - Wallis test has more test power if the data is homosc.)

L147: This is "serial" data, there is no point in considering the assumption of norms.

L148: -> Levene's test

L149: At the significance level of 0.05, the assumption of agreement of variances cannot be rejected

L150: (Leven's test, x\_OBS = 0.4, df\_num = 3, df\_denom = 62, p-value = 0.750)

L151: Symmetry verification

L152: We want to compare the medians of independent samples -> Kruskal-Wallis test

L153: At the significance level of 0.05, the median agreement hypothesis cannot be rejected

L154: (Kruskal-Wallis test, x\_OBS = 3.7, df = 3, p-value = 0.295).

L155: Ie. statistically significant differences between producers (in terms of rank

L156: products in competition) do not exist.

L157: ## Example 5.

L158: The effect of three slides on blood clotting was studied. In addition to other indicators, the so-called thrombin time was determined. Data on 42 monitored persons are recorded in the file trombin.xls. Does the size of the thrombin time depend on which preparation was used?

L159: rename columns

L160: data is already in standard format

L161: exploratory analysis - verification of OP

L162: does not contain OP

L163: verification of normality

L164: At the significance level of 0.05 we reject the assumption of normality (for Ačka)

L165: Information needed to set rounding

L166: sd is rounded to 2 valid digits

L167: sd and position measures are rounded to hundredths (unification across groups)

L168: Verification of variance agreement (not necessary - we have to use KW anyway)

L169: According to the box chart and information on the ratio of the largest and smallest

L170: variances (>> 2) cannot match variances.

L171: The presumption of normality was rejected -> Levene's test

L172: thrombin.s $ group = as.factor (thrombin.s $ group)

L173: The assumption of homoskedasticity was rejected

L174: Symmetry verification

L175: we do not reject the assumption of data symmetry

L176: We want to compare medians than. selections that do not have standards. distribution

L177: -> Kruskal - Wallis test

L178: At the significance level of 0.05, we reject the hypothesis of a median agreement

L179: Ie. thrombin time is statistically significant

L180: affected by the preparation. -> multiple comparisons

L181: altP = T sets the p-value so that when making a decision

L182: on statistical significance compared to alpha

L183: (default: altp = FALSE, then compare with alpha / 2)

L184: effect counting

L185: overall average

L186: averages in groups

L187: effects

L188: list sorted

L189: ## Example 6. (multiple groups)

L190: When Snow White got to the seven dwarves, she sensed an opportunity to make a lot of money. The Dwarves basically beat the Snow White's hand and immediately handed over all the mined gold to it. However, even this is not enough for Snow White and she feels that she could benefit more from dwarves. Therefore, she began to record how many kilograms of gold a day she received from each of the dwarves (snehurka.xlsx). Verify that the dwarves differ in the amount of gold mined, if so, to form a homogeneous group in terms of gold mined.

L191: data is in standard data format

L192: data does not contain OP

L193: verification of normality

L194: At the significance level of 0.05, we do not reject the assumption of normality

L195: The presumption of normality was not rejected -> Bartlett's test

L196: At the significance level of 0.05, the assumption of agreement of variances cannot be rejected

L197: ANOVA

L198: We reject the presumption of conformity

L199: -> there are stat. significant differences in mean values

L200: POST-HOC

L201: effect counting

L202: overall average

L203: averages in groups

L204: effects

L205: list sorted

%%%%%%%%%% cv13

L0: # Exercise 13. Nonparametric tests, goodness-of-fit tests ## Michal Béreš, Martina Litschmannová, Adéla Vrtková

L1: # Conformance distribution probability testing of discrete NV (finite number of values) - good agreement test

L2: - we test whether the measured data (their relative frequencies) agree with some specific distribution (ie its probabilities) - we test using the $ \ chi ^ 2 $ good agreement test - test assumptions: (ATTENTION relate to the expected frequencies - ie those we would monitor if the measured data were 100% according to the distribution in the hypothesis) - Expected frequencies ≥ 2, - at least 80% of the expected frequencies> 5 - test statistics (the one that has a $ \ chi ^ 2 $ distribution) is $ G = \ sum\_ {i = 1} ^ k (O\_i - E\_i) ^ 2 / E\_i $ - distribution ma degree of freedom $ df = k - 1 - h $ - k is the number of possibilities - h is the number of estimated parameters (this applies to incompletely specified tests)

L3: example see example 1

L4: ## Good match test for a continuous random variable (or discrete with an infinite number of values)

L5: - we have to convert to a table with a finite number of values - for discrete (eg poison) we group from a certain number of columns eg 4,5,6, ... to "4 and more" - for continuous we produce a series of intervals and look at how many values will fall within the given interval - eg: (- $ \ infty $, 3), <3, 4), ..., <10, $ \ infty $) - then we have to calculate for each interval how many% data belongs to them, which gives a table of expected probabilities - we continue as before - to test the normality of the distribution there is a function pearson.test (data) from the package nortest

L6: example 2, 3, 4

L7: # PivotTables

L8: - tables containing data depending on two factors - one of the factors is usually an independent variable in which we monitor whether it affects the other factor (dependent variable) - the independent variable is usually in rows - it is usually dependent in columns - pay attention to the whole testing examines correlation, not causality! Causality can be assessed by "expert" evaluation - statistical conclusion: there is a statistically significant dependence between the independent and dependent variable (correlation) - expert assessment: the independent variable statistically significantly affects the dependent variable (causality)

L9: ## PivotTable Visualization

L10: - visualization eg using the barplot function - pay attention to what rows and columns are, we always want individual divided columns to be via independent variables (each column for one value of an independent variable) - beside = T determines whether we want to merge adjacent columns into one split column or not - preferred visualization using mosaicplot - the same as for barplot, linked columns must be via independent variables

L11: examples in Examples 5,6,7

L12: ## Dependency table dependencies

L13: - Correlation coefficient CC - Corrected correlation coefficient CCcor - Cramer's coefficient V - we will use this mainly - cramersV (cont.tab) function from the lsr package

L14: examples in Examples 5,6,7

L15: ## PivotTable Dependency Test

L16: - $ H\_0: $ there is no dependency between the independent (eg is a smoker) and dependent (eg suffering from illness) variable - $ H\_A: \ neg H\_0 $ - chisq.test function (cont.tab) - assumptions: Expected frequencies ≥ 2, at least 80% of expected frequencies> 5 - expected frequencies can be found from chisq.test (cont.tab) \ $ expected

L17: # Association tables

L18: - this is a special case of a PivotTable - it always has exactly 2 options for the dependent and exactly 2 options for the independent variable

L19: ## Mandatory association table format

L20: - lines indicate the possibilities of the independent variable - the first line is the so-called exposed part of the population (the exposed phenomenon we study - eg smokers try to influence the effects of smoking) - the second line is the unexposed part of the population - columns indicate the possibilities of the dependent variable phenomenon (eg occurrence of disease, product error, ...) - the second column indicates the rest - no occurrence of the investigated phenomenon

L21: ## Relative risk and odds ratio

L22: - Relative risk and odds ratio provide the same information, only in a different format - all point IOs are calculated using the epi.2by2 (associ.tab) function from the epiR package - the function takes the association table as input, which must be in the correct format!

L23: ### Relative risk

L24: - denotes $ RR $ - this is the risk ratio (probability of occurrence of the studied phenomenon) in exposed and unexposed populations - if it is equal to 1 it means the same probabilities of occurrence in both exposed and unexposed - if it is greater than 1 then the exposed population has greater probability of occurrence - if it is less than 1 then the exposed population has a lower probability of occurrence - the point estimate $ \ hat {RR} $ is calculated as the ratio rel. frequency of the studied phenomenon in exposed and unexposed populations - the function epi.2by2 provides interval estimates - if the IO does not contain the value 1 then there is a statistically significant dependence between the dependent and independent variable

L25: ### Odds ratio

L26: - denotes $ OR $ - this is the ratio of chances (chance of occurrence of the studied phenomenon) in exposed and unexposed population - if it is equal to 1 then it means the same chances of occurrence in both exposed and unexposed - if it is greater than 1 then the exposed population has greater chance of occurrence - if it is less than 1 then the exposed population has less chance of occurrence - the point estimate $ \ hat {OR} $ is calculated as the ratio of the chances of the exposed phenomenon in the exposed and unexposed population - epi.2by2 provides interval estimates - if IO does not contain the value 1 then there is a statistically significant dependence between the dependent and independent variable

L27: example in example 7

L28: # Examples (good match tests)

L29: ## Example 1.

L30: The dice were rolled 6,000 times and the number of falling stitches was recorded. [image.png] (attachment: 64f1169e-6bc1-470a-8afb-b282230c2c9f.png)

L31: H0: The cube is fair. (so all probabilities are 1/6)

L32: Ha: The cube is not fair. (H0 negation)

L33: test assumptions must be checked

L34: All expected frequencies are greater than 5.

L35: At the significance level of 0.05 we do not reject HO (p-value = 0.711,

L36: Chi-square test of independence, df = 5).

L37: ## Example 2.

L38: The manufacturing company estimates the number of failures of a particular device in 100 hours using a Poisson distribution with parameter 1.2. Employees recorded the actual number of failures at a total of 150 100-hour intervals for inspection (results are shown in the table). Verify with a clean significance test that the number of failures of a given device within 100 hours actually has a Poisson distribution with the parameter λt = 1.2.

L39: Fully specified test

L40: H0: The number of faults during 100 operating hours can be modeled

L41: Poisson distribution with parameter 1.2.

L42: Ha: The number of faults during 100 operating hours cannot be modeled

L43: Poisson distribution with parameter 1.2.

L44: test assumptions must be checked

L45: 4 of the 5 expected frequencies, ie 80%, are greater than 5.

L46: At the significance level of 0.05 we do not reject HO (p-value = 0.590,

L47: Chi-square test of independence, df = 4).

L48: ## Example 3.

L49: Employees recorded a total of 150 100-hour breakdowns at check (results are shown in the table). Use a clean significance test to see if the number of failures of a given device has a true Poisson distribution in 100 hours.! [Image.png] (attachment: 4da89057-87d1-4bcd-a488-3365237654f7.png)

L50: Incompletely specified test

L51: H0: The number of faults during 100 operating hours can be modeled

L52: Poisson distribution.

L53: Ha: The number of faults during 100 operating hours cannot be modeled

L54: Poisson distribution.

L55: Poisson distribution parameter estimate

L56: test assumptions must be checked

L57: 4 of the 5 expected frequencies, ie 80%, are greater than 5.

L58: At the significance level of 0.05 we do not reject HO (p-value = 0.491,

L59: Chi-square test of independence, df = 3).

L60: ## Example 4.

L61: Time intervals (s) between the passages of individual vehicles were measured on the motorway within a few minutes. The detected values of these distances are recorded in the file dalnice.xlsx. Verify that this is data from a normal distribution (use a good match test).

L62: automatic test of good agreement from continuous data

L63: generate values for the x-axis

L64: generate values for the y-axis

L65: Add a curve to the last graph based on the values generated above

L66: install.packages ("nortest")

L67: H0: Spacing between vehicles can be modeled by normal distribution.

L68: Ha: Spacing between vehicles cannot be modeled by normal distribution.

L69: Specify the number of degrees of freedom

L70: At the significance level of 0.05, HO can be rejected (p-value << 0.001,

L71: Chi-square test of good agreement, df = 12).

L72: a test you already know

L73: # Examples of PivotTable and Association Tables

L74: ## Example 5.

L75: Decide on the basis of the data file experimentovani-s-telem.xls (Dudová, J. - Experimentování s tělem (survey results), 2013. Available online at http://experimentovani-stelem.vyplnto.cz) whether there is a connection between gender of respondents and whether they have tattoos. Use Cramer V to assess the contingency rate.

L76: Preprocessing

L77: Variants of cat. Variables (factors) must be arranged and named so

L78: how they should be arranged and named in the accounts. table

L79: Exploratory analysis

L80: associated relative frequencies

L81: line relative frequencies

L82: columnar relative frequencies

L83: Visualization in standard R

L84: Cluster bar graph

L85: compare graphs, which of the graphs is more suitable for the presentation of the data

L86: width of graphs in Jupyter

L87: graph matrix 1x2

L88: Stacked bar graph

L89: width of graphs in Jupyter

L90: graph matrix 1x2

L91: Mosaic chart

L92: width of graphs in Jupyter

L93: Rotate y-axis labels

L94: compare which of the graphs is more suitable for the presentation of the given data

L95: install.packages ("lsr")

L96: Calculation of Cramer V ####

L97: PivotTable independence test

L98: H0: Data is independent -> whether the individual is male or female

L99: Does not affect his likelihood of having a tattoo

L100: HA: negation H0 (there is a dependency)

L101: Required to verify assumptions

L102: All expected frequencies are greater than 5.

L103: At the significance level of 0.05, HO can be rejected (p-value = 0.003,

L104: Chi-square test of good agreement, df = 1).

L105: The observed dependence can be assessed as weak (Cramer's V = 0.121).

L106: ## Example 6.

L107: For a differentiated approach in personnel policy, the company's management needs to know whether job satisfaction depends on whether it is a Prague plant or non-Prague plants. The results of the survey are in the following table. Display the data using a mosaic chart and, based on the independence test in the combination table, decide on the dependence of job satisfaction on the company's location. To assess the contingency rate, use Cramer V. <br>! [Image.png] (attachment: ebc6062c-1020-4fab-8855-6665d65d59a7.png)

L108: We do not have a data matrix, ie cont. we must enter the table "manually"

L109: Exploratory Analysis ####

L110: associated relative frequencies

L111: line relative frequencies

L112: columnar relative frequencies

L113: Visualization in standard R

L114: Mosaic chart

L115: Rotate yo axis labels 90

L116: Cramer's V.

L117: H0: There is no connection between job satisfaction and company location.

L118: Ha: There is a connection between job satisfaction and company location.

L119: Chi-square test of independence in PivotTable ####

L120: All expected frequencies are greater than 5.

L121: HO can be rejected at the significance level of 0.05 (p-value << 0.001,

L122: Chi-square test of good agreement, df = 3).

L123: The observed dependence can be assessed as moderate (Cramer's V = 0.296)

L124: ## Example 7. (Association table)

L125: Between 1965 and 1968, a cohort study of cardiovascular disease under the Honolulu Heart Program began monitoring 8,006 men, of whom 7,872 did not have a history of stroke at the start of the study. Of this number, there were 3,435 smokers and 4,437 non-smokers. When followed for 12 years, 171 men in the group of smokers and 117 men in the group of non-smokers suffered a stroke.

L126: #### a)

L127: Record the results in the association table.

L128: completion of the absolute frequency table

L129: addition of the table of relative frequencies

L130: #### b)

L131: Based on visual assessment, estimate the effect of smoking on the incidence of cardiovascular disease.

L132: Mosaic graph visualization in basic R

L133: Calculation of Cramer V ####

L134: According to the mosaic graph and Cramer's V (0.061) there is a connection between smoking

L135: and the occurrence of apoplexy evaluated as very weak.

L136: #### c)

L137: Determine the absolute risk of cardiovascular disease in smokers and non-smokers.

L138: risk = probability

L139: Smokers

L140: Assumptions check

L141: OK (3,435> 190.3)

L142: Calculation of point and 95% Clopper-Pearson interval estimation

L143: The smoker has an apoplexy risk of about 5.0%. 95% Clopper-Pearson

L144: the interval estimate of this risk is 4.2% to 5.8%.

L145: Non-smokers

L146: Assumptions check

L147: OK (4,437> 350.6)

L148: Calculation of point and 95% Clopper-Pearson interval estimation

L149: In non-smokers, the risk of developing apoplexy is about 2.6%. 95% Clopper-Pearson

L150: the interval estimate of this risk is 2.1% to 3.2%.

L151: #### d)

L152: Determine the relative risk (including 95% of the interval estimate) of cardiovascular disease in smokers and non-smokers. Explain the practical significance of the results obtained.

L153: install.packages ("epiR")

L154: Smokers have about 1.89 times higher risk of apoplexy than non-smokers. 95%

L155: the interval estimate of this relative risk is 1.50 to 2.38.

L156: According to the interval estimate of relative risk, it is clear that at the surface

L157: significance 0.05 is a statistically significantly higher risk for smokers

L158: apoplexy than in non-smokers.

L159: #### e)

L160: Determine the absolute chances of cardiovascular disease in smokers and non-smokers.

L161: In a smoker, the chance of apoplexy is about 52: 1,000. at 1,052 smokers

L162: Approximately 52 occurrences of apoplexy can be expected.

L163: In non-smokers, the chance of developing apoplexy is about 27: 1,000. for 1,027 non-smokers

L164: Approximately 27 occurrences of apoplexy can be expected.

L165: #### f) Determine the relative chances of cardiovascular disease in smokers.

L166: Smokers have about 1.93 (= 0.0524 / 0.0271) times higher chance of apoplexy

L167: than for non-smokers. 95% interval estimate of this ratio

L168: The odds are 1.52 to 2.46.

L169: According to the interval estimate of the odds ratio, it is clear that on the surface

L170: Significance 0.05 is for smokers

L171: statistically significantly higher chance of developing apoplexy than in non-smokers.

L172: #### g)

L173: Decide at a significance level of 0.05 on the dependence of the incidence of cardiovascular disease on smoking.

L174: Attention! The epi.2by2 command does not output the expected frequency for

L175: Chi-square test of independence.

L176: It is not possible to verify the test assumptions!

L177: H0: There is no link between smoking and the occurrence of apoplexy.

L178: Ha: There is a link between smoking and the occurrence of apoplexy.

L179: All expected frequencies are greater than 5.

L180: HO can be rejected at the significance level of 0.05 (p-value << 0.001,

L181: Chi-square test of good agreement,

L182: df = 1). The observed dependence can be assessed as very weak

L183: (Cramer's V = 0.061).

%%%%%%%%%% cv1\_add

L0: # Exercise 1 - Combinatorics

L1: ## Adéla Vrtková, Michal Béreš, Martina Litschmannová

L2: ## Variations

L3: V (n, k) - variation without repetition, the first argument will be the total number of entities, the second argument the size of the selection

L4: the function is created by the fucntion command, it is an object whose name is given by a variable

L5: to which I will assign this object

L6: here I enter the number of parameters and their names

L7: the whole body of the function is enclosed in parentheses {...}

L8: the factorial in the original Rku exists so we will use it

L9: what the function returns is given in the return (...) statement

L10: V \* (n, k) - variation with repetition

L11: ## Permutation

L12: P (n) = V (n, n) - permutation

L13: P \* (n1, n2, n3, ...., nk) - permutation with repetition, input will be a vector with individual numbers of unique entities

L14: vec\_n is the vector of values eg: vec\_n = c (2,2,2,4,3)

L15: we calculate how many values we have in total

L16: their factorial = value in the numerator

L17: a simple loop starts with the for statement, then the iterator name az follows in parentheses

L18: what list will be taken

L19: count is an iterator and will gradually take values from the vector vec\_n

L20: we gradually divide by the factorial of each number of unique entities

L21: ## Combination

L22: C (n, k) - combination

L23: the function for combination already exists in Rku and is called choose

L24: C \* (n, k) - combination with repetition

L25: we use a known formula

L26: # Exercise tasks

L27: ## Example 1.

L28: There are three types of locks available in the store. To open the first lock, it is necessary to press four of the ten buttons marked with numbers 0 to 9. (The order does not matter - the buttons remain pressed.) The second lock will open if we press six of the ten buttons. To open the third lock, it is necessary to set the correct combination on the four discs. Which of these locks best protects against thieves?

L29: ## Example 2.

L30: The store offers two types of briefcase locking. The first briefcase is locked with a six-digit cipher. The second case is locked with two locks that open at the same time. The cipher of each of them consists of three digits. Determine for each briefcase the probability of opening by a thief on the first attempt. Which type of lock is safer?

L31: ## Example 3.

L32: There are 40 balls in the urn - 2 red and 38 white. We randomly pull 2 balls out of the urn. How likely are they both to be red?

L33: ## Example 4.

L34: The student had to prepare answers to 40 questions for the exam. He could not answer the two questions the examiner asked him, so he said, “I'm unlucky! These are the only two questions I can't answer. ”How likely is he to tell the truth?

L35: ## Example 5.

L36: A student passes a chemistry test if he underlines the only two aldehydes on the list of 40 chemical compounds. What is the probability that a student who passes the compounds at random will pass the test?

L37: ## Example 6.

L38: A group of 40 tourists returned from abroad, including 2 smugglers. At the border, customs officer 2 called for a personal search and it turned out that both were smugglers. The remaining tourists responded: "The customs officer was really lucky!", "Someone gave the smugglers!",. . .. How to deal with these statements? Is there a legitimate suspicion that someone smuggled smugglers?

L39: ## Example 7.

L40: From the urn with three balls, two red and one white, two balls will be selected at the same time. The student and the teacher place a bet. If both balls are the same color, the student wins. If the balls have different colors, the teacher wins. Is the game fair? What are the probabilities of a teacher and a student winning?

L41: the combn function produces combinations of the specified size - the first parameter is a vector of values, the second the size of the selection

L42: ## Example 8.

L43: The game described in Example 7 was not fair. What ball (red or white) do we need to add to the urn to make the game fair?

L44: ## Example 9.

L45: You want to play Man, don't be angry, but the dice are lost. How and how can a dice be replaced if you have playing cards (a deck of 32 cards) and 4 different colored balls?

L46: ## Example 10.

L47: You want to play Man, don't be angry, but the dice are lost. How can I replace a dice if you have 3 different colored balls?

L48: ## Example 11.

L49: They have a sales event in the Škoda car dealership in February. In addition to the standard equipment, they offer 3 items from the above-standard equipment free of charge. Extra equipment includes 7 items: - cruise control, seat heating, rear airbags, xenon headlights, roof window, transmission safety lock, special durable metallic paint. How many options does the customer have to choose 3 items from the above-standard equipment?

L50: ## Example 12.

L51: During the exam, 12 students sat in the 5th row. The examiner wants to determine for himself how to deploy these students in a row. - How many ways to deploy students? - Student Brahý asks to be able to sit on the edge and leave earlier to catch the train. How many options are there for deploying students if the examiner wants to meet Brahý's request? - How many possibilities are there for deploying students if Pažout and Horáček are not allowed to sit next to each other?

L52: a

L53: b

L54: c

L55: ## Example 13.

L56: How many anagrams can be created from the word STATISTICS?

L57: ## Example 14.

L58: They got new goods in Tesco - 6 kinds of boys' trick. They have at least 7 pieces of each species. The mother wants to buy her 4 T-shirts. How many options are there to choose from - should they all be different? - if he admits that they can all be the same?

L59: a

L60: b

L61: ## Example 15.

L62: How many passwords of length 5 can we create from alphabetic characters - if case insensitive? - if case sensitive?

L63: a

L64: b

%%%%%%%%%% cv11

L0: # Exercise 11. Two-sample tests / Interval estimates ## Michal Béreš, Martina Litschmannová

L1: # Overview of IO tests / constructions

L2: two-sample data

L3: - Paired data indicates data that are taken to two measurements of the same entities -> data columns are dependent. - If they are independent, it is a two-sample test. - For paired data, we calculate the difference between the columns (or another function according to the assignment) and use one-sample tests for this difference.

L4: ### Examples of paired data:

L5: - measuring bulbs at two different temperatures (if each piece is measured twice - at temperature 1 and temperature 2) - be careful here, it can happen that the tests are eg destructive and it is not possible to measure twice the same entity (product). Then we would consider two independent selections, each for one type of measurement -> independent data columns -> two-sample tests - measuring values in the patient's blood before and after drug administration - again pay attention to drug testing in two groups (placebo / real drug) -> two independent groups -> two-sample tests

L6: ## In general for two-sample tests / IO

L7: - the test is always tied to the relevant IO -> same conditions for use - if the test has conditions of use (eg: data normality, data symmetry) then this condition must meet \*\* both files \*\*, if at least one does not meet, we consider presumption as violated - one of the very important assumptions is the independence of the data - eg: measurement of the products of manufacturer A and the products of manufacturer B - here it is reasonable to assume that the products of manufacturer A are separate entities from the products of manufacturer B

L8: ## Two-sample tests / IO - difference of position measures

L9: we make test data - so it can be used everywhere

L10: ### Two-sample Student's t-test

L11: - Tests / estimates the difference of mean values: $ H\_0: \ mu\_ {1} - \ mu\_ {2} = a $ - requirements: - Data normality - Homoskedasticity (scatter matching) - selection independence - the function must have the parameter var.equal = TRUE

L12: H0: mu1 - mu2 = 2

L13: HA: mu1 - mu2! = 2

L14: H0: mu1 - mu2 = 2

L15: HA: mu1 - mu2> 2

L16: H0: mu1 - mu2 = 2

L17: HA: mu1 - mu2 <2

L18: ### Aspin-Welsh test

L19: - Tests / estimates the difference of mean values: $ H\_0: \ mu\_ {1} - \ mu\_ {2} = a $ - requirements: - Data normality - selection independence - function must have parameter var.equal = FALSE

L20: H0: mu1 - mu2 = 2

L21: HA: mu1 - mu2! = 2

L22: H0: mu1 - mu2 = 2

L23: HA: mu1 - mu2> 2

L24: H0: mu1 - mu2 = 2

L25: HA: mu1 - mu2 <2

L26: ### Mann-Whitney test

L27: - Tests / estimates median difference: $ H\_0: X\_ {0.5,1} - X\_ {0.5,2} = and $ - requirements: - independence of selections - (same type of distribution) - requires conf.int = TRUE, for calculation IO

L28: H0: X0.5,1 - X0.5,2 = 2

L29: HA: X0.5,1 - X0.5,2! = 2

L30: H0: X0.5,1 - X0.5,2 = 2

L31: HA: X0.5,1 - X0.5,2> 2

L32: H0: X0.5,1 - X0.5,2 = 2

L33: HA: X0.5,1 - X0.5,2 <2

L34: ## Two-sample tests / IO - variance ratio

L35: ### F-test

L36: - Tests / estimates the variance ratio: $ H\_0: \ sigma ^ 2\_ {1} / \ sigma ^ 2\_ {2} = a $ - requirements: - normality of data - independence of selections

L37: H0: sigma1 ^ 2 / sigma2 ^ 2 = 1

L38: H0: sigma1 ^ 2 / sigma2 ^ 2! = 1

L39: H0: sigma1 ^ 2 / sigma2 ^ 2 = 1

L40: H0: sigma1 ^ 2 / sigma2 ^ 2> 1

L41: H0: sigma1 ^ 2 / sigma2 ^ 2 = 1

L42: H0: sigma1 ^ 2 / sigma2 ^ 2 <1

L43: ### Levene's test

L44: - Tests equality of variances: $ H\_0: \ sigma ^ 2\_ {1} = \ sigma ^ 2\_ {2} $! - requirements: - independence of selections - requires data in standard data format - leveneTest function in the car package

L45: we produce data in a standard data format

L46: install.packages ("car")

L47: H0: sigma1 ^ 2 = sigma2 ^ 2

L48: HA: sigma1 ^ 2! = Sigma2 ^ 2

L49: ## Two-sample tests / IO - probability difference

L50: ### Homogeneity test of two binomial distributions

L51: - Tests the match / estimates the probability difference: $ H\_0: \ pi\_ {1} - \ pi\_ {2} = 0 $ - requirements: - sufficient size of selections: $ n\_i> \ frac {9} {p\_i (1-p\_i) } $ - independence of selections

L52: we make the appropriate data

L53: H0: pi1 - pi2 = 0

L54: HA: pi1 - pi2! = 0

L55: H0: pi1 - pi2 = 0

L56: HA: pi1 - pi2> 0

L57: H0: pi1 - pi2 = 0

L58: HA: pi1 - pi2 <0

L59: # Examples

L60: ## Příkald 1.

L61: Data in the cholesterol2.xls file indicate the blood cholesterol level of men of two different age groups (20-30 years and 40-50 years). Verify at the significance level 0.05 hypothesis that the cholesterol level in the blood of older men does not differ from the cholesterol level in the blood of younger men.

L62: Read data

L63: Conversion to standard data format

L64: Exploratory analysis

L65: Elimination of remote observations:

L66: be careful in the data we have NA and we have to reckon with it !!!

L67: (eg for length determination)

L68: rounding -> 3 valid digits -> according to sd to thousands

L69: \*\* Mean / median compliance test \*\*

L70: Verification of normality

L71: normality in hl. significance 0.05 OK

L72: Verification of variance agreement

L73: Exploratory

L74: exploratory assessment: the ratio of the largest to the smallest is> than 2

L75: -> I don't assume variance agreement

L76: Exactly by F-test

L77: H0: sigma.starsi = sigma.mladsi

L78: Ha: sigma.starsi <> sigma.mladsi

L79: I select the required data

L80: In hl. significance 0.05 we reject the assumption of agreement of variances

L81: The observed discrepancy between the variances can be at a significance level of 0.05

L82: Mark as statistically significant.

L83: Mean value verification (Aspin-Welch test)

L84: H0: mu.starsi - mu.mladsi = 0

L85: Ha: mu.starsi - mu.mladsi! = 0

L86: in hl. significance 0.05 we reject H0-> there is a stat. significant difference.

L87: H0: mu.starsi = mu.mladsi (mu.starsi - mu.mladsi = 0)

L88: Ha: mu.starsi> mu.mladsi (mu.starsi - mu.mladsi> 0)

L89: At the significance level of 0.05, we reject the assumption of medium agreement

L90: Cholesterol levels in groups of younger and older men in favor

L91: Alternatives that older men have higher mean cholesterol levels

L92: than men younger

L93: According to the results of the sample survey, we expect a medium content

L94: cholesterol in the blood of scared men will be about 0.524 mmol / l higher than

L95: medium chol. in younger men. According to 95% left intervention.

L96: to estimate the given difference, we expect the mean cholesterol content at

L97: older men at least 0.457 mmol / l higher than the mean value

L98: Cholesterol in younger men.

L99: ## Example 2.

L100: Data in the depression.xls file represent the length of remission in days from a simple random selection of two different groups of patients (patients with endogenous depression and patients with neurotic depression). Verify that the observed difference in mean remission length in these two groups of patients is statistically significant.

L101: Reading data from xlsx file (using readxl package)

L102: Conversion to standard data format

L103: Exploratory analysis

L104: Data do not contain remote observations.

L105: rounding -> 3 valid digits -> according to sd per unit

L106: \*\* Mean / median compliance test \*\*

L107: Verification of normality

L108: We assume the assumption of normality by the Shapir - Wilkov test.

L109: In hl. significance of 0.05, we reject the assumption of normality

L110: at least approximately, we will check the similarity of the distribution

L111: we choose data for easier processing

L112: Median compliance verification (Mann-Whitney test)

L113: According to the histograms, we assume that the data have the same type of distribution.

L114: H0: med.neuro = med.endo (med.neuro - med.endo = 0)

L115: Ha: med.neuro! = Med.endo (med.neuro - med.endo! = 0)

L116: in hl. significance 0.05 we reject H0-> there is a stat. significant difference

L117: H0: med.neuro = med.endo (med.neuro - med.endo = 0)

L118: Ha: med.neuro> med.endo (med.neuro - med.endo> 0)

L119: At the significance level of 0.05 we reject hyp. on the agreement of median times to

L120: disease remission for both groups of patients in favor of the alternative

L121: The median time to remission is statistically significant in patients with neurotic depression

L122: significantly greater than in patients with endogenous depression.

L123: The remission time of patients with neurotic depression is about 191 days longer

L124: than in patients with endogenous depression. According to 95% left

L125: interval estimation is expected for patients with neuro. depression have

L126: at least 168 days longer remission than endo patients. depression.

L127: ## Example 3.

L128: We monitor urine osmolality at the inpatient station at 08:00 and 11:00 at 16 men. Based on the results in the osmolality.xls file, verify that the osmolality has increased statistically significantly.

L129: Load data

L130: Osmolality increase calculation

L131: Exploratory analysis

L132: Data contains remote observations.

L133: Remove outliers

L134: Exploratory analysis for data without remote observations

L135: rounding -> 2 valid digits -> according to sd per unit

L136: Verification of normality

L137: The presumption of normality is verified by the Shapir - Wilk test.

L138: In hl. significance of 0.05, the assumption of normality cannot be rejected

L139: (Shapir-Wilk test, W = 0.949, p-value = 0.545).

L140: Paired t-test

L141: H0: mu.narust = 0 mm

L142: Ha: mu.narust> 0 mm

L143: According to a sample survey, urine osmolality can be expected to increase

L144: increases by approx. 24 mmol / kg between 8 and 11 o'clock. According to the 95% interval estimate

L145: osmolality can be expected to increase by at least 10 mmol / kg).

L146: At a significance level of 0.05, this increase can be described as statistically

L147: significant (paired t-test, t = 3.1, df = 13, p-value = 0.005).

L148: ## Example 4.

L149: Semiconductor components of two manufacturers - MM and PP - were tested. MM claims that its products have a lower percentage of defective pieces. To verify this claim, 200 components were randomly selected from MM production, of which 14 were defective. A similar experiment was performed at PP with the result of 10 defective out of 100 randomly selected components.

L150: ### a)

L151: Test MM's claim with a clean significance test.

L152: Verification of assumptions

L153: Furthermore, for both companies we assume that n / N <0.05, ie that the given population (components) has a range of at least 20 \* n, ie 20 \* 200 (4,000), resp. 20 \* 150 (3,000) components, which is probably a fairly realistic assumption.

L154: Pearson's X2 test

L155: H0: pi.PP = pi.MM

L156: Ha: pi.PP> pi.MM

L157: Due to the p-value> hl. significance 0.05 we do not reject H0 - ie assumption.

L158: identical error rates. Therefore, it cannot be said that MM has better production.

L159: Pearson's X2 test

L160: H0: pi.PP = pi.MM

L161: Ha: pi.PP! = Pi.MM

L162: ### b)

L163: Test MM's statement using an interval estimate of a significance level of 0.05.

L164: Based on 95% Clopper - Pearson right - hand interval estimation

L165: (-0.036; 1,000) the observed difference in production quality can be described as

L166: not statistically significant. We can reach the same conclusions on the basis of

L167: Pearson's right-hand test