**Handout 1: Basic Notations**

**Population:** refers to all humans beings or a large part of humanity, and that we are trying to find out something that holds for these people. “Population” could also refer to a species of animals or plants. We refer to its members as **individuals.** We call the members of the population, **individual** **measurements.**

**Sample:** A small subset of population

**Sampling:** Selecting the sample with an *unbiased* and *random* *fashion* is important, only then we are able to draw mathematically valid conclusions from the sample.

**Inference:** Process of transferring findings from the sample to the population. *This is by necessity an imperfect process but probability theory will tell us the chance of making a mistake.*

**Variable: “**aspect” is called a variable, example:age, daily internet usage, gender.

**Scales:** A measurement requires a **scale** for the results. Right scale helps statisticians to select the correct procedure.

*Scales can be distinguished in two main kinds:*

**Categorical:** Results can take one of a small finite set of possible values. Examples: male/female, home/EU/overseas students; parties in an election.

**Quantitative:** Results are numbers, examples: reaction time, weight; memory usage.

Sometimes the categories under consideration are *ordered,* for example: strongly agree/agree/neutral/disagree/strongly disagree. Such a scale is called **ordinal.**

**Five Inference Tasks:** to classify statistical inferences tasks into five different groups:

**C: Categorical-**Distribution of a categorical variable in a population done by measuring its distribution in a sample.

**Q: Quantitative-**mean of a quantitative variable in a population and obtain an approx. value by calculating the mean in a sample.

**“⇒”** considering the hypothesis that the first variable influences the second.

**C⇒C** Value in one categorical variable influences on another categorical variable.

**C⇒Q** Value in one categorical variable has an influence on a quantitative variable.

**Q⇒Q** Relationship between two quantitative variables.

**Handout 2: Descriptive Statistics**

**C- one categorical variable:** we report the *number* of individuals in each category.  If there are more than two categories, then the **mode** is of interest. (**Mode: value that appears the most)**

Representation of categorical data as a **Pie Chart** or even **Bar Graph**.

**C⇒C - two categorical variables:** list the numbers as a table, pie charts, or bar graphs.

**Q - one quantitative variable: percentiles**

**BOX PLOT (box & whisker plot)**

1. Order your data points
2. Determine your quartiles
3. Compute the interquartile range as the difference between the 3rd & 1st quartile
4. Compute u as the 3rd quartile plus 1.5 times the interquartile range; compute l as the 1st quartile minus 1.5 times the interquartile range; u and l are called “fences”
5. Find the largest data point below u, call it U;
6. Find the smallest data point above l, call it L;
7. Draw a box from the first to third quartile, indicating the median as an additional line inside the box; draw whiskers from 3rd quartile to U & from 1st quartile to L;
8. Indicate any values that do not lie between L & U as separate points (“outliers”)

**Mean/Average = (Sum of Values)/Number of Values (N)**

**Standard Deviation = Sqrt((Xi - u)^2)/N)** (If finding for a sample, then denominator → N -1)

**Median ⇒** If there n results, the MEDIAN will be (n+1)/2 example: 6 no.s, the median will be (6+1)/2 = 3.5 *(The 3.5th entry or between 3rd & 4rd is the median)*

*The median is much more accurate “midpoint” that the mean, as it’s not affected by any outliers.*

*For unevenly* distributed data: the data is **right skewed** if data points above the median are further away from the median than data points below. With right skewed data, the mean will be above the median.

(the **measure of dispersion** that goes with the mean is the **standard deviation**)

**Handout 3: The Normal Distribution; the Central Limit Theorem**

**Bell Shape-** when we have a large population & we make the subdivision for the individual bars very fine, then we can observe that the “histogram” takes a characteristic.

**Normal** or **Gaussian** Characteristics(computing the standard deviation of a population & computing the proportion of data points we find certain intervals): **68-95-99.7% Rule**

https://lh4.googleusercontent.com/gzcHWj7AFtCnVd3pwfC6KBsrFJMbgScPRk8z5PP-N6XLnvXjOlrIt5WHd39sxO4XcHKl7_1_yysnQUDv5yFEbAzrdt1xorjD7oTgFixmYwKXhYffv0Y5rJ_nav5Wsg81ej_vsnI_https://lh3.googleusercontent.com/vdi14RSmEmJztscRfqepK5AQo2JW0bg0qQhWXfVDhxtdBbxHyyobvpI9WG1XfD8oXzRdaD4kcJxyEzkkXjcCVPHnw_2tGVlGfX7Tc2VW2dl1AMwUEJkeeq8KHJI3Gf0XaaonrkTdhttps://lh6.googleusercontent.com/s4bCyvTPa7B53JRb3qHUTVdQMWv20VHWUNqtb94P9S29iY3IWwv7UEphrLqXnuQhan4nfLVe8dGRjy86YU--2JG5RCGijEbB_rUMSEwt0U6zoGXNk2vTfpUswmyntB9n2crU6fsB

**Sampling a Population:** When we sample *randomly* from a (large) population, then we expect the distribution of values in the population to determine the *probability* of obtaining a certain value in the sample. *If we repeat a random experiment many times then the relative frequencies tend to get close & close to the theoretical probabilities. If they don’t, then we assume that the assumed probabilities were wrong or that our experiments are not properly random or independent of each other.*

**Sampling a Normal Population:** If we know that a population is normally distributed with mean and standard deviation then we can translate the 68-95-99.7 rule directly into statements about probabilities.

**Testing for Normality:** We have no way of measuring a variable on the whole of the population but only have the data from a sample. To infer whether the population is normally distributed or not, we look into the sample that is distribute similarly to the population - we can draw the histogram of the sample data & try to judge whether it looks normal or not. **Q-Q Plot:**

1. If we have N data points in our sample - list in ascending order: d1<d2<...<dN
2. Compute the quantiles of the Standard Normal Distribution (where =0 & =1) for the probabilities https://lh3.googleusercontent.com/pnyvE9faYQAsWEQNDS_9TbULoSGzNLk5VM5_ZsX07y-HlEW29GTXGzC8hrMff1j8M_mhyh1A7pHSN90V7mnZ3BEEBhzMorPp-O-JnNsouZ0sdDzaEGRJF0_MgfDgIg7r84OvEIfZto obtain q1, q2, ….qN-1
3. Draw Scatter Plot for the pairs https://lh6.googleusercontent.com/uyAfyxjxDltVkYx69pvhZ0pdcl-GMRCVMyLe8zXD_UzQgzeA3Vs-KBXe85_IilNtjf5HMhAttZFQvRDKcsFdPcCbP6ZcmS8D9pHIxLc0xIXlrxBWaAN4SV_hctxE76CxiseId8Jb
4. For Normal Distribution: Points will lie On or Near a Straight Line.

**Standard Normal Distribution:** If we know that a population is normally distributed & we know its mean & standard deviation then we can make a judgement as to whether a given data point is unusual or not. Translating Standard Normal Distribution to a General One (Normal Distribution): If Y is a variable that is normally distributed with **mean** & **standard deviation** (we say, N(, )-distributed) then https://lh3.googleusercontent.com/Ds1fVxU3UNo3SrMKWbu9WFF0OPU_dAUCoqs--4XqYnjWTk9QjSAXnaPgILKKl3WRXRZ0WKvtJV_Bsunwvc7xFNCtXRpDhmKFmWIN3fHQXMQE8loCgFORpZE1XNjsXe83CwctaJp8is Standard Normally Distributed.

**Example**: Assume Y is normally distributed with= 3 & = 5 & would like to know how likely it is to get a result between -1 & 2.

https://lh3.googleusercontent.com/Ds1fVxU3UNo3SrMKWbu9WFF0OPU_dAUCoqs--4XqYnjWTk9QjSAXnaPgILKKl3WRXRZ0WKvtJV_Bsunwvc7xFNCtXRpDhmKFmWIN3fHQXMQE8loCgFORpZE1XNjsXe83CwctaJp8X = (Y-3) / 5 N(0,1)-distributed ⇒ 5X = Y - 3 ⇒ 5X + 3 = Y

With these translations between Y & X we transform the bounds on Y to bounds for X:

-1Y2 ⇒-1 5X+32⇒ -4/5 X -1/5 ⇒ -0.8 X-0.2

**Random Variables:** When we pick an individual at random & take the measurement, then the outcome is dependent on the randomness of the sampling process. Instead of **“mean”**, we used **expected value** and write **E(x)** instead of , and **SD(X)** instead of .

**Density Function:** function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random **variable**) can be interpreted as providing a relative likelihood that the value of

For the population we measure frequencies & for the random variable we measure the probability of the obtaining a value in a certain interval.

The Density Function for a random variable that is perfectly normally distributed with expected value 0 & standard deviation 1 is computed as:

https://lh4.googleusercontent.com/TxIcRH3BzQDXP6C1q5WS1paR8R2_2G9xfJ0bXODtZnAat_Ya2QT_YRgW4gmxDwDdAqxxQlELIu2mk4NVzsSCpRv29khc0_4v_gj5rMObf-sx94xX0shr_9Dc6VmC2XFVbXH4PlUv

**Finding the mean of a population:** in order to find out about a population we should sample it randomly many times (N > 30, say) and study the data so obtained. For example, if we want to get an estimate for the mean µ of the population we can just compute the mean m of our sample. How good an estimate are we likely to get? **Central Limit Theorem** which makes a statement about the sample mean as a random variable:

**m is an unbiased estimator:** The expected value of the sample mean is the population mean: E(m) = µ.

**“Law of small numbers”:** As a random variable, the sample mean is distributed approximately like a Gaussian. This is true even for relatively small values of N (say, N ≥ 10).

**“Law of big numbers”:** The standard deviation of the sample mean is SD(m) = √σ N (where σ is the population standard deviation). In particular, as N increases, the standard deviation decreases.

**Central Limit Theorem:** The sample mean is a random variable that is approximately distributed as an N(µ, √σ N ) Gaussian.

Central Limit Theorem makes no assumption about the population’s distribution. The population does not have to be normally distributed. The mean is normally distributed even if the population is not!

In one experiment, 166 children from a certain background (parents with depression) were given this test and the test average was m = 55.71. This does not seem like a big variation from the established mean of 50 especially since we know that the standard deviation is 10. However, this is the standard deviation for testing a single person whereas 55.71 is the averaged result from 166 people. By the Central Limit Theorem we know that the standard deviation for a sample size of 166 is not 10, but 10/166= 0.776. So now the observed deviation of 5.71 looks very big. More precisely, we know that the sample average m behaves approximately like a normal distribution with expected value E(m) = µ = 50 and standard deviation SE(m) = 10/166= 0.776.

Normalisation applied to this gives us the N(0,1) distributed variable

https://lh6.googleusercontent.com/wJe8vHy2Cy6Y6P3Usr-TJ43csAmDRSZrKf2pVpbm1WiMsIh2m2XB_9NCPAb6LbdpjFf7Z1pCLfdyY5BwRvaYdPhhQh7EI2T_Ph1c9g0A8IplNflV9k0COpuxtx3F9v5WUYdqvOnpz = (55.71-50)/10/166    ⇒ z = 55.71/0.776 = 7.36

We get a value of 7.36 for a random variable that is N(0,1) distributed. This is hugely far away from the mean 0 and a value this high could only be expected with probability p = 0.00000000000018 (there are no table values for this anymore). Thus we may confidently state that these children are significantly different from their peers at p < 0.01 (we could go lower). Note that we did not use the sample standard deviation because we (believe we) know the population standard deviation.

**Sampling Categorical Variables:**  if the probability of the first outcome is p and that of the second q = 1− p, then the random variable of picking one sample has expected value p and standard deviation pq. This means that the average from N samples is a random variable with expected value p and standard deviation pq/N.

**NOTE: Mean of Population:** µ (usually not known to us)

**Mean of a Sample: m** & computed as X/N

**Expected Value of a Random Variable** X: **E(X)** (but if random variable is normally distributed, then µ - same applied for Standard Deviation)

**Standard Deviation of a Population:** σ (usually not known to us)

**Standard Deviation of a Sample: s https://lh6.googleusercontent.com/T0sSvi3MOMw8ZVUplgSNYeR4b5Q1nEE_fpXL5Mk5z-7o5efGCwmGVfA-KnNuKrJV1O31RawNFfKTN2bd_NfWnIadlVOSYAx8pcY5FmG1yNdBn2E2XMd80Ne-acRxXKAlHWEeoZ6R**

**Right Skewed:** Data to the right of the median tends to be further away than data on the left.

**PROBLEM WITH BOXPLOTS + OUTLIERS:** if there are many outliers (e.g. shoe prices, some shoes cost millions) then it could make our boxplot unreadable. We can overcome this by removing outliers and listing them separately.

**Probability from a density function graph:** work out the area under the graph

Density function of a sample: basically is a Gaussian graph. With the mean at the centre. Don’t forget: almost zero three s.d. away from the midpoint. Proof is CLT.

Numerical quantity that describes the population is called a **parameter**. Mean, median, standard deviation, minimum etc.

Numerical quantity that describes the sample is called a **statistic**. Sample mean is a statistic. A statistic is a random variable because we always assume that the sample is chosen randomly.

Sample standard deviation is a poor statistic, because it does not average out towards the population standard deviation. Biased estimator

Variance, however is unbiased.

**Case I: Estimating the pop mean: known population standard deviation**

z is N(0,1) distributed. Hence, for 95% we have the bounds of results in

p-value: 0.05 means 95% confidence level.

**Case II: Estimating the pop mean: unknown pop standard deviation**

The bounds are dependent on N since we have two estimates s and m, N=5:[-2.8, 2.8], N=10:[-2.3, 2.3], N=20:[-2.1, 2.1], N=30:[-2.05, 2]

It is only correct if m is normally distributed. By the CLT, if N is suff large, N > 4 for nice symm pop and N >29 for a skewed pop

**Case III: Estimating the pop mean: a categorical variable**

0 for Brexiteers and 1 for non-Brexiteers

1 with probability and 0 with , from this we can calculate .

1. Determine confidence interval, 95%: [-2, 2], 99%:[-2.6, 2.6], 99.7%:[-3, 3]
2. Calculate the interval, same as z-stat
3. Find minimum N by using:

***Haphazard (Sample) :*** relies on belief that we have ability ourselves to choose in a random fashion. BAD: humans exhibit all kinds of biases when confronted with free choice. A sample is haphazard if it is dependent on a choice by the experimenter. ***Convenience:*** “comes our way”. ADV: doesn’t require much work by experimenter. BAD: artificially restrict population from which we sample. ***Self-selected***: restricts to those willing 2 take part. BAD: self-selected so can be misleading if ppl chose themslvs, even with large sample size. We cant force people to take part in experiment, so self-selection may never be avoided in an experiment. ***Systematic:*** select from a list of individuals in a systematic fashion. Order individuals appear is not related to our question we are studying, hence list is *random from the point of view of our variables of interest*. ***Cluster***: randomly sample from particular subgroup. ADV certain diversity/coverage guaranteed, even if some subgroup members are hard 2 reach and would otherwise not be considered.***Stratified***: considering different groups in the population. Sampling = no of individual from each group. (men vs women, same number of ppl sampled)***Probability***: emphasis on randomness. Selection of individuals by probability alone. BAD: non-response bias:think real statisticians are spammers. USE IF WE HAVE ACCESS 2 WHOLE POPULATION.

***Survey***: we ask ppl to report variables of interest themselves, because there is no way to measure it. BAD: 1. People lie, especially if not anon (2) phrasing of question (leave EU and regain our country vs remain member & co-op with other nations.***Observation***: measure variables of interest BUT don’t influence them in any way. ***Experimental***: experimenter has complete control of explanatory/independent var hence experimenter measures response var.

***T-TEST***when we have population s.d. σ, we standardize ***m*** by working out ***z***, obtaining a random variable that is approx. N(0,1) distributed. Doing the same but with sample s.d. ***s*** instead of σ we get the rand.var ***t*** instead.

**Standard norm dist vs density function of t(graph of t vs norm)**: tails are heavier, means it is more likely for t to take values far away from 0. THIS DEPENDS ON sample size N. Bigger effect for smaller samples, for big samples (N >100), t is practically norm.dis (N(0,1)). *Distribution of t depends on N*. t-distributed with **df n-1**. **T & z** statistics are only **true** if (1) distribution of population is reasonably symmetric(NO HEAVY SKEW) (2) sample size sufficiently large

***Within-participants study***: measure response var y FOR EACH SINGLE participant under TWO CONDITIONS.

Within is interested whether the sample average of all differences (4 single participants) is significantly different from 0. We are in case 2,t.test. Compare with cv vals for t-distr under n-1 df. Remember under null m = 0 so t has 0 in it.

***Between-p study***: 2 populations. We are interested in testing whether the difference between 2 means is significantly different from 0.To calculate 95% cv 4 t-distr, there are fewer *df* ; we apply welsh procedure -> which selects the t-distr for determining cut-off points.

With **df = n1 + n2 – 2** *if the s.d. of the populations are the same*.

***Confidence interval :*** if explanatory var **X** has a significant impact on **Y** (Response), we can confirm **c.i.** calculation to estimate how far apart 2 population means are. Bounds are for 95% c.i. the pop2 – pop1 for s. For large samples the cv is not far from 2. R function t.test() returns 96% confidence by default.

**One-tail:** impossible for μ1 smaller than μ2 or we don’t care about the situation, then we only need to test whether it is actually bigger/not. In one tail case null hypo is μ1 < μ2 . for 95% confidence compare with 95% cut-off point.

two tail: Sometimes μ1 IS smaller than u2, hence statistic can be -/+! This case we have significance at 95% confidence if below 2.5% cut-off or above 97.5% cut-off. [FOR R : alternative=”greater” / alternative = “less” to t.test for one-tail test to be computed. ]

***ANOVA*** if there r several levels, don’t pair t-tests. More likely to commit Type1 error, as probability of making an error is 95% taking this to the power of comparisons brings it down. 5 levels = 10 pairwise tests = prob of not making type 1 error 0.95^10=0.6. ADV 4 ANOVA. Anova avoids this.

Anova allows us to measure the dispersion of the individual outcomes simultaneiously. Sample var is unbiased estimator for pop var.

THIS IS THE ONLY REASON WHY ANOVA WORKS with var instead of s.d. Also aim of anova is not 2 find the population var, but 2 compare the two estimates for it.

***1st estimator of σ2 WITHIN GROUPS*** : C=> Q but K>2 levels. N measurements for each K level (table N x K). For each level work out the variance. Then take the mean. Swithin = mean(each levels variance). THIS DOESN’T DEPEND ON NULL HYPO.

**2nd estimator of *σ2* BETWEEN GROUPS**: UNDER THE NULL HYPO: mean of each level, is estimate of overall population mean. By CLT: the statistic is normally distributed with EX(pop mean) and var/N. sample var of all the means for each level is approx.

This is sample var(means). DEPENDS on null hypo. If the null is false, then s^2 IS MUCH LARGER THAN σ ^2/N.

***F-statistic:*** If null hypo is true, f=1, but due to randomness of sampling, get close to 1.

Sampling distribution F is known.as both swithin & sbetween (are ^2) = guaranteed +ve, hence F is +ve. F is one-sided.

DegreeofF; nwhere k – levels & N are participants

**Assumptions of Anova**

1. Population normally distributed: but F is robust here.
2. Variance of each population remains unaffected by treatment levels: sometimes not true as some people react +vely/some -vely. ANOVA NOT ROBUST AGAINST THIS VIOLATION. F stat will report more significant deviations from 1 than significance level indicates.   
   can invoke welchs procedure, if variances are not similar between populations, but welch can only compensate for small violations fo equal variance assumption.   
   can be solved by:  
   -draw boxplots and get visiual representation of how samples are distributed  
   -or can inspect vector of sample variances – will tell us if something is wrong. If there is a big difference between sample variances, we can decide that F test is not appropriate. Can do Resampling. We resample the sample
3. Population for each level was assumed to have same size N, different sizes can be accommodated with more complicated formulas but resulting F test becomes more susceptible to other violations.

Important: **result of F + df.** If result not likely under null (less than 5%), then we say: the treatment has a STATISTICALLY SIGNIFICANT EFFECT. Still remains to us to find out what the effect is exactly.

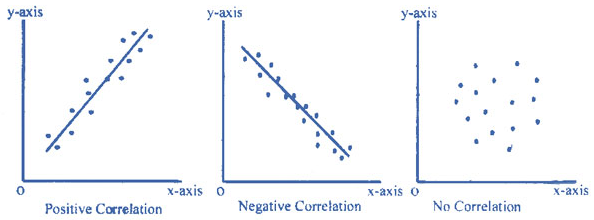
Pearsons correleation coefficient

Q->Q

i.e. are smoking and life expectancy related? Are highschool A-level results and uni exam perf correlated?

If X and Y are large simultaneously, both exceed their means and the product (X-m\_x)(Y-m\_y) will be positive. When they are both small, the product of two negative numbers will again be positive. If r is greater than zero then X and Y are positively correlated and if its negative they are negatively correlated. -1 < r < 1. Null Hypothesis says its close to 0.

From every individual we obtain two vals, one for X and one for Y.

In R a dataframe could look like:

|  |  |  |
| --- | --- | --- |
| Individual | X | Y |
| 1 | X1 | Y1 |
| … | X2 | Y2 |
| N | X3 | Y3 |

Correlation is cheap, to gret zero exactly is nearly impossible. That does not mean that the correlation is significant. Correlation is not causation. Easy to tempted to postulate the causation when the correlation coefficient is high and when it agrees with our prejudice.

Find out whether 0 is in our conf interval. If it is we can’t claim a connection. r has no distribution like normal, t or F. Moreover r formula does not take the sample size into account. Sample size is crucial to the significance level. r can be transformed into a random variable that is normally distributed.

1. is N(, distributed (p is the true correlation in the population)

Standardisation to N(0,1) by

1. Conf. interval for 95% of N(0,1) are ±1.96 (two-tailed)
2. Inverse of g to the transformation f (as we don’t want the bounds of
3. Overall procedure:
   1. Compute r from sample, then
   2. Compute
   3. Compute
   4. The final conf. interval is []

Improve of our findings -> Increase N, to tighten the interval. We can not solve for N, in order to know how many measurements are need for a given confidence level as in t-statistic or z, because of the non-linear distortion by the transformation of f and g.

1. How big must N be to obtain after measuring sample corr of 0.5
2. From we get and and .
3. From , and from
4. Using , and
5. We conclude 200 measurements are need to get the desired bounds on p.

To get tight bounds on the true corr. coefficient, a high number of measurements is needed.

Regression

Regression model is able to make predictions for example of the lifespan of a given person, utilizing cigarettes smoked for instance.

We determine a and b such that we get the best prediction. Given a sample of value pairs (X,Y) and optimal a and b can found like:

where are averages of X and Y.

tells us how much variability of Y is explained by X. Since all practical situations r is smaller than 1, is even smaller than r.

r=0 explains nothing, r=1 explains everything, r must be 0.8 to get 40% of the variation explained.

Summary:

Strength: given by which is between 0 and 1

Explanatory value: , between 0 and 1, expressed as percentage

Slope:

This is linear regression, but non-linear can be dangerous because of over-fitting.

***Latex section:*** What is **accuracy**? 1.clearly state research hypo + null (2) describe sampling approach (3)describe experiment set-up in detail (4) describe how data is collected+recorded (5) include full raw data (6) include visualisation of data (7) state clearly which statistical test was applied

**Honest approach** 2 sampling: (1) don’t formulate research hypo after collecting data (2) include difficulties encountered during recruiting volunteer participants (ppl who declined) (3) argue why sample is representative, consider how sample is less perfectly random (4) consider biases (5) report steps taken to clean data – e.g. removing outliers.

Reporting outcome: (1) if significant effect found: give p-value (probability to obtain a result this extreme under the null) (2) if nothing is going on, say: we are unable to reject the null hypo. We encourage others to dig deeper

**Knitr and Latex:**

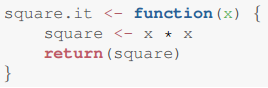
knitr is an instance of literature programming (text+code).

**<<>>=**  This tells knitr that an R program code chunk is entered. Insert inbetween <<*here*>>. Two important commands:

Echo = FALSE ; (default is true) by =F the R code chunk will appear in the final document and not just the output of the code. (As normally code in <<>> is hidden)

Eval = FALSE; (default = true); code will be shown but outcome wont. Will show whats in brackets, but not outcome like graph.

Why is knitr/literature programming important? (1) if u don’t use it then you will have 2 use some other txt software, copy + paste R code to your word doc, and you can MISS some important R info, leading to a report that is incomplete where readers cant verify your analysis. (2) if you forget something/make changes to R code/add more detail to R/ clean up code (outliers) -> then you are forced to keep copy+pasting. Hence, u can miss things which leads to inconsistent report where tables/graphs don’t correspond 2 data. Knits solves = problems with copying + updating. Knitr promises document includes all data, which ADV = supports reproducible research (can be verified by any reader/correct reproducible data).

 Vector:

Dataframe:



Get columns s:



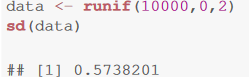














I

 QQ-Plot: qqnorm(), t-distribution: t(df, ncp=0)

pnorm(b, - pnorm(a,



ANOVA