Bar plot 和 Histogram 都是用于展示数据的可视化工具，但它们之间存在一些关键的区别。

1. 描述的数据类型：

* Bar plot（条形图）：通常用于展示分类数据。在条形图中，不同的条形代表不同的分类，条形的高度或长度则代表每个分类的数值。
* Histogram（直方图）：通常用于展示连续数据。在直方图中，每个条形的宽度代表一个数值范围（即 bin），高度或长度则代表落在这个范围内的数据点数量。

1. 数据分组和展示方式：

* Bar plot：不需要对数据进行分组，每一个条形代表一个分类的数值。条形之间通常是相互独立的，没有直接的数值联系。
* Histogram：需要对连续数据进行分组（binning）。每一个 bin 代表一个数值范围，多个数值落在这个范围内就会被归为一组，形成一个条形。因此，直方图中的条形之间存在直接的数值联系，可以比较不同 bin 之间的数据分布。

1. 展示的信息：

* Bar plot：除了可以展示分类数据的数量之外，还可以展示分类之间的相对大小关系以及它们的变化趋势。
* Histogram：主要展示数据的分布情况，例如数据的集中趋势、离散程度以及分布形状等。

1. 适用场景：

* Bar plot：适用于展示类别之间的比较，例如不同地区的销售数据、不同产品的市场份额等。
* Histogram：适用于展示数据的分布特征和变化规律，例如用户年龄分布、一段时间内用户的点击次数的分布等。

总的来说，Bar plot 和 Histogram 的主要区别在于它们的数据类型、分组方式、展示信息和适用场景。在实际应用中，应根据需要展示的数据特征和目标来选择合适的可视化工具。

数据清理的步骤主要包括以下几项：

1. **检查缺失值**：首先，我们需要检查数据集中是否存在缺失值。在R中，我们使用is.na()函数来识别包含缺失值的行。如果有缺失值，我们需要决定如何处理它们。常见的处理方式有删除含有缺失值的行、填充缺失值或使用某些统计方法处理。在这个例子中，我们选择删除含有缺失值的行，使用complete.cases()函数。
2. **处理异常值**：异常值是远离数据集主体的值，可能由于测量错误或数据错误等原因而存在。异常值可能会扭曲数据集的统计性质，因此通常需要处理。常见的处理方式包括删除异常值、将异常值替换为近似的正常值或使用统计方法来识别和修正异常值。
3. **数据类型转换**：有时候，数据集中可能存在不同类型的数据，如字符型和数值型。在进行统计分析之前，确保所有变量都是正确的数据类型是很重要的。在这个例子中，我们将“Feature”和“Genotype”列转换为因子类型（factor），因为它们是分类变量。
4. **清理重复行**：如果数据集中存在重复行，这些重复行可能会对统计分析造成干扰。我们需要检查并删除重复行。在R中，我们通常使用duplicated()函数来识别重复行，并使用na.omit()函数来删除它们。
5. **处理异常值和离群点**：有时候，数据集中可能存在极端值或离群点，这些值可能会对统计分析产生负面影响。我们需要识别并处理这些异常值。在R中，我们可以使用诸如IQR（四分位距）或Z分数等方法来识别离群点，并决定是否删除它们或用其他方式处理它们。
6. **填充缺失值**：对于含有缺失值的数据，除了删除包含缺失值的行外，还可以选择填充这些缺失值。填充缺失值的方法有很多种，包括使用均值、中位数、众数、插值或其他预测方法来填充缺失值。
7. **格式化和标准化**：有时候，数据集中的数值可能需要进行格式化或标准化以适应特定的分析方法。例如，将日期格式化为标准格式、将分类变量编码为数值变量等。

通过执行这些步骤，我们可以确保数据集的质量和准确性，从而为后续的数据分析和建模提供可靠的基础。

eval

eval控制了代码块是否执行。若设定为eval=TRUE，则markdown会执行代码块里的代码；若eval=FALSE，则markdown不会执行代码。

echo

echo参数控制了markdown是否显示代码块。若echo=TRUE，则表示代码块显示在markdown文档显示代码块；反之，代码块不出现在输出结果中。

include

Lesture 3

Sampling error is the error resulting from the process of selecting samples from a larger population. It typically manifests as the difference between the sample mean (or rate) and the population mean (or rate). Due to the inherent variations among individuals in the population, the sample mean only represents a portion of the total population, leading to a difference between the sample and population means. This difference is known as sampling error. Sampling error is inevitable but can be minimized by increasing the sample size or improving the sampling methods.

The dierence between sample and population is called sampling error (“error” here means “difference from the true population”, not “mistake”)

Ideally, a sample is random and each individual in the population has an equal probability of being sampled. If this is not the case (i.e. if some parts of the population have a higher chance of being sampled than others), there is a sampling bias.

To design data collection procedures that avoid sampling bias, it is important to follow several guidelines:

1. Define the target population clearly: Before starting the data collection, it is essential to have a clear understanding of the target population. This helps in ensuring that the sampling frame represents the entire population and avoids any potential biases.
2. Use random sampling techniques: Random sampling techniques ensure that each member of the population has an equal chance of being selected for the study. This helps in reducing the risk of sampling bias and provides more representative data.
3. Ensure representativeness: It is important to ensure that the sample selected is representative of the target population. This can be achieved by ensuring that the sample is stratified based on key characteristics of the population and ensuring proportional representation from each stratum.
4. Collect sufficient sample size: A large sample size provides a more precise estimate of the population characteristics, reduces the sampling error, and allows for more reliable conclusions. It is crucial to determine an appropriate sample size before starting the data collection.
5. Use blinding techniques: Blinding techniques are used to reduce the risk of observer bias during data collection. This can be achieved by masking the identities of the individuals or groups being studied and ensuring that the data collectors are unaware of their identities.
6. Standardize procedures: Standardized procedures help to ensure that data are collected in a consistent manner across different samples or studies. This reduces the risk of introducing biases due to variations in data collection methods.
7. Train data collectors: It is essential to train data collectors on the procedures and methods used for data collection. This helps to ensure that they understand and follow the standardized procedures, reducing the risk of introducing biases during data collection.
8. Validate data collection tools: The reliability and validity of data collection tools should be established through appropriate validation procedures. This helps to ensure that the tools used for data collection are accurate and reliable, reducing the risk of introducing biases due to faulty tools or methods.
9. Monitor and adjust sampling plan: It is important to monitor the progress of the data collection and adjust the sampling plan if necessary. For example, if it becomes apparent that a particular group is underrepresented in the sample, additional efforts should be made to collect data from that group to ensure representativeness.

By following these guidelines, it is possible to design data collection procedures that minimize the risk of sampling bias and provide more representative and reliable data.

Lecture 5 sampling distribution

Sampling distribution refers to the distribution of sample results in a sampling survey. When conducting a sampling survey, samples are randomly selected from the population, and the statistical measures of these samples, such as the mean, variance, and other measures, form a distribution known as the sampling distribution. The sampling distribution is an important concept in probability theory and statistics that describes the possible range of values and probabilities for sample statistics. Through the sampling distribution, one can perform probability analysis and inference on sample results to understand the nature and characteristics of the population.

The Standard Error of the Mean (SEM) is a measure of how well your sample mean estimates the

true population mean.

Standard deviation measures variability in a dataset

Standard error of the mean measures how good your estimate of the population mean is

The central limit theorem

The central limit theorem (CLT) is a fundamental theorem in probability theory and statistics that states that the mean of a sufficiently large number of independent random variables, regardless of their distribution, will approximately follow a normal distribution.

For sample means

Even if a population is not normally distributed, the sampling distribution (for large enough

samples) will tend to be normal

More in general

If we take n independent random variables from any distribution, and take their (normalised)

sum, then that sum will tend towards a normal distribution with increasing n.

Even when things are not normally distributed, a normal

distribution often “comes out” of parameter combinations, such as taking the mean.

Long format

The "long format" refers to a data recording method that collects information about each individual observation (e.g., each participant) in a study, typically including multiple variables or measurements for each person. This format contrasts with the "wide format," where observations are recorded in separate columns for each variable. The long format is often used in data analysis software and statistical programming languages to organize and analyze data effectively. It facilitates data manipulation and merging, and is well-suited for statistical models that require observations to be related through individuals, such as repeated measures designs or longitudinal studies.

Wide format

The "wide format" refers to a data recording method where observations are recorded in separate columns for each variable. Each column typically represents a different variable or measurement, and the rows correspond to individual observations or cases. This format contrasts with the "long format," where information about each individual observation is collected in a single row, with multiple variables or measurements for each person.

The wide format is commonly used in spreadsheets and databases to organize and store data. It allows for easy visual representation and comparison of different variables across cases. However, it can become unwieldy when dealing with large datasets or when there are many variables to track. In such cases, the long format can provide a more efficient and manageable way to organize data.

H0 HA

Week 6 hypothesis testing

Null hypothesis: nothing is happening

´ In our example: Swedish students are able to distinguish child mortality between countries where it differs with ratio 2:1 same as chimpanzees

The effect of the medication is the same as that of a placebo.

´ Lack of sleep does not affect cognitive function.

´ Wildtype mice and FMRP knockout mice show the same level of synaptic activity.

´ Rates of TB infection have not changed between 1920 and 2010.

´ Alternative hypothesis: something is happening

´ In our example: Swedish students are able to distinguish child mortality between countries where it differs with ratio 2:1 worse than chimpanzees

The medication has a different effect than a placebo.

´ Lack of sleep affects cognitive function.

´ Wildtype mice and FMRP knockout mice show different levels of synaptic activity.

´ Rates of TB infection changed between 1920 and 2010.

What is P value?

´ Probability of observing a value as or more extreme as the one you observed if the null hypothesis were true.

What do we do with it?

Level of significance – convention (no magic!)

´ p < 0.05: reject the null hypothesis

´ p > 0.05: cannot reject the null hypothesis (not the same as confirm it!)

NULL HYPOTHESIS:

There is no difference in the amount of food eaten in response to light and between trials

ALTERNATIVE HYPOTHESIS:

The amount of food eaten in response to light is significantly different from the amount eaten

during the inter-trial interval.

Week 7 data cleaning

脏：

• Special characters where not needed;

• Use of the wrong data structures;

• Duplicated rows;

• Misspelling;

• White spaces;

• Missing data;

• Zeroes instead of NULL or NA values;

• Other inaccuracies;

• Poor structure.

改进：

Free of duplicate rows/values

• Error-free (e.g. free of misspellings)

• Relevant (e.g. free of special characters)

• The appropriate data type for analysis

• Free of outliers (or only contain outliers that have

been identified/understood)

• Follows a “tidy data” structure.

The key steps of the data cleaning:

• Screen: Check the data set systematically

• Diagnose: Find out the nature of the problem.

• Treat: Delete, edit, or leave the data as it is.

• Document: Comment on each step to make sure you will not forget the reason of the action and the original state.

Possible signs of the bad data:

• Lack of data: Some columns/rows have fewer values. Why?

• Excess of data: Some rows are duplicated or include more values than it should be. Were some rows duplicated? Were there some additional values added?

• Outliers: Some values are far beyond the limits of the data. An error is possible.

• Strange patterns of data or results: The results look quite weird. Was there an error or falsification?

Missing data: Some answers were not recorded or test subjects were lost for the follow-up (dropped out from the study).

• Errors or typos: Errors can always take place.

• No error: The value is strange, but is actually valid.

The researcher has to accept it.

Week 10 the t-test

Assumptions required for using the t-statistic

1. Data is continuous and randomly-selected.

→ See lecture on sampling

2. The sample is normally distributed.

→ See lecture on sampling distributions

3. The mean and standard error are independent.

→ Nearly always true, but could test by simulation.

Question from W10 Practical

Q2. Is the t-distribution an appropriate test to answer this question? Do these data meet the assumptions required?

a. The first assumption is that the volumes recorded are continuous and randomly selected. A quick look at the data will reveal that the values are continuous.

We don’t know whether the bottles whose volumes were recorded were selected at random. In the real world you would likely know much more about where the data came from and could determine whether it was randomly-sampled. For the purposes of the rest of the problem sheet, we are going to assume that this the data is a random sample from the population of volumes.

b. The second assumption is that the sample is normally distributed. This can be done by plotting a histogram to determine whether the data looks normal enough to me - note here you have to make your own judgement.

不抄了 W10 problem set Note非常重要！！！

1.

one sample/paired T test: 正态分布+t/wilcox test

t.test(bottles$Volume, mu = 500)

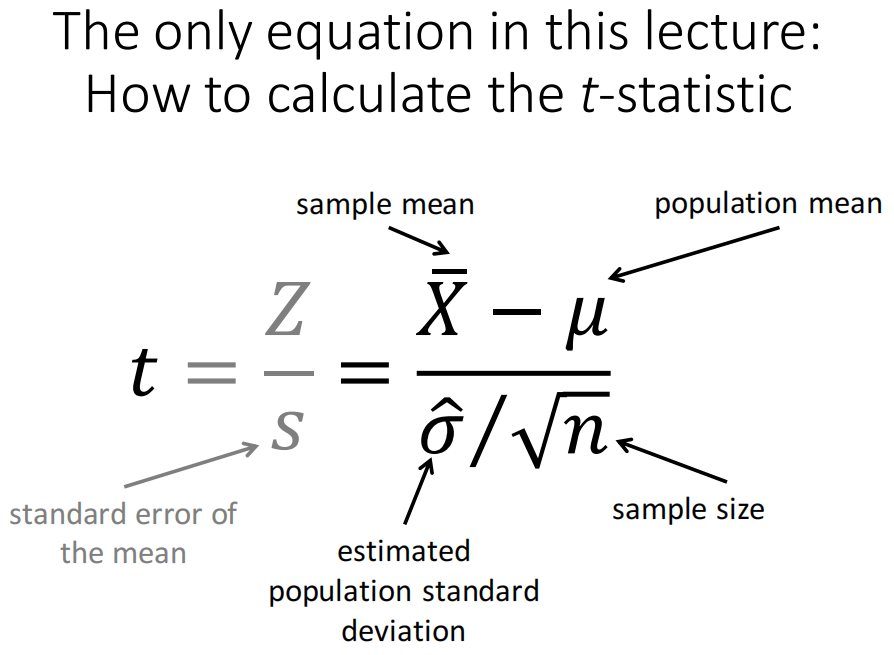
也就是说2 sample：正态分布+方差测试+t/wilcox test

还要检测independence

Appropriate for small sample sizes (n < 100).

• Generally, we assume a normal distribution of the underlying population.

What do we mean when we say two values are ‘different’?



• The t-distribution is a generalization of the normal distribution.

• Symmetric and bell-shaped.

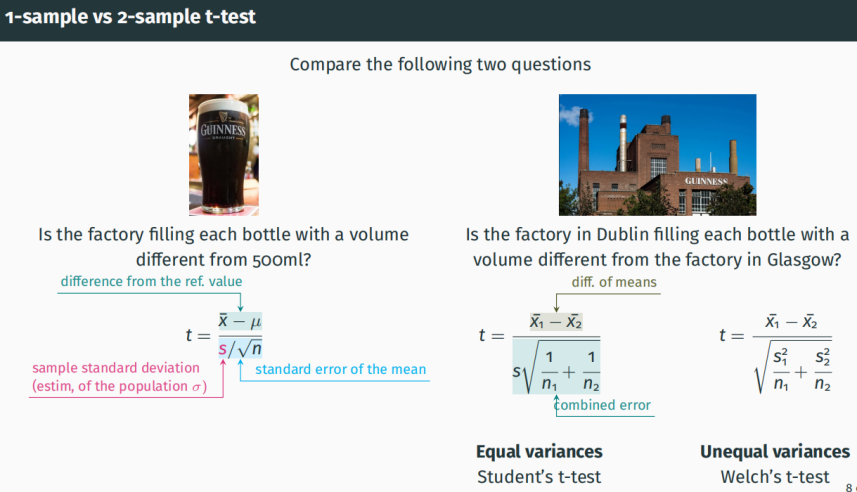
• Its distribution depends on a single parameter, the degrees of freedom (d.o.f, see next slide)

• It is heavier tailed compared to the normal distribution at small d.o.f.

• As d.o.f. increase, it converges to the normal distribution

The degrees of freedom (d.o.f.) in a statistical calculation represent the number of independent pieces of information that are used to calculate that statistic.

Week 11 T-test application



2.

Data cleaning中所有错误

单独分开讨论吧 这样比较清晰一些

不是的 long和wide是根据最后analysis的需求来用的 tidy是指数据里没有missing data（NA），没有重复数据，没有outlier，然后factor是factor，numerical variable是numerical

可以有NA NA删不删实际上很复杂 需要判断 但是在这个考试里直接删掉吧

3.

学长早上好！想问一下如果做paired T test的话，是不是不要求两个distribution是normal的，只要求两个sample的差是normal的就可以呀

是的 差符合normal就行 因为paired t test相当于one sample t test

3.

ggplot是需要long format的 还是转换吧

4.

Outlier实际depends 这个考试直接删掉

可以用boxplot看有没有异常值

或者大于mean +/- 三倍标准差的也属于异常值

好的，先画boxplot，然后根据outlier的x,y轴值的特征找到他然后删除

但是以防万一是其他老师改的 还是用boxplot吧

5.

那是所有numeric的数据都要检查一下对吗

是的 要检查一下看是否为正确的data type

factor必须是factor numerical必须是numerical

outlier是针对数值型数据

如果是非数值型数据要检查重复和NA

Lecture 8 Comparing two means - simulation

1. Happiness ratings of UoE students based in Edinburgh and in Haining

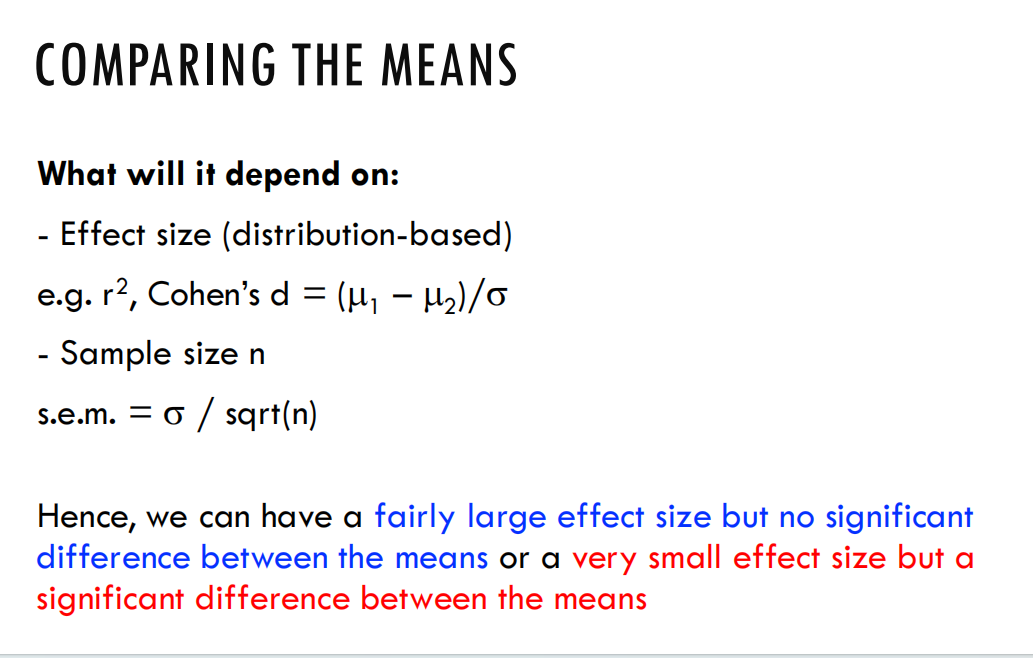
 2 independent samples

1. Satisfaction ratings of catering on ZJU international campus in 2018/19 and 2019/20 (assume the same people provide ratings)

 2 paired samples => take the difference, treat as 1 sample

1. Average activity of PFC neurons with a baseline (4 Hz)

 1 sample, compare with a number



好的，以下是一份英文的数据分析管道计划：

Data Analysis Pipeline Plan

Data Collection:

* 1. Identify the source of the data you need for your analysis. This could be internal databases, external APIs, or flat files.
  2. Determine the frequency at which the data will be collected. Is it real-time, daily, weekly, monthly, or annually?

Data Cleaning:

* 1. Eliminate or correct any missing or incorrect data.
  2. Handle duplicates and outliers.
  3. Ensure that the data is formatted correctly and is in the appropriate format for analysis.

Data Transformation:

* 1. Normalize or standardize the data if necessary.
  2. Perform any required calculations or transformations to make the data suitable for analysis.

Data Exploration:

* 1. Use exploratory data analysis techniques to gain insights into the data's structure, patterns, and relationships.
  2. Identify any potential issues or biases in the data.

Data Visualization:

* 1. Create visualizations to communicate insights and patterns in the data.
  2. Use charts, graphs, and maps to aid in understanding the data.

Statistical Analysis:

* 1. Apply appropriate statistical techniques to analyze the data and answer your research questions.
  2. Interpret the results and draw conclusions from the analysis.

Data Reporting:

* 1. Compile the results of your analysis into a report that summarizes your findings and provides insights.
  2. Ensure that the report is easy to understand and contains all relevant information.

Data Archiving and Documentation:

* 1. Archive the raw data and any intermediate files used in the analysis for future reference.
  2. Document your analysis process, including any decisions made during the pipeline, for reproducibility purposes.