Data & Things

(Spring 25)

Monday February 10

Lecture 4: Statistics

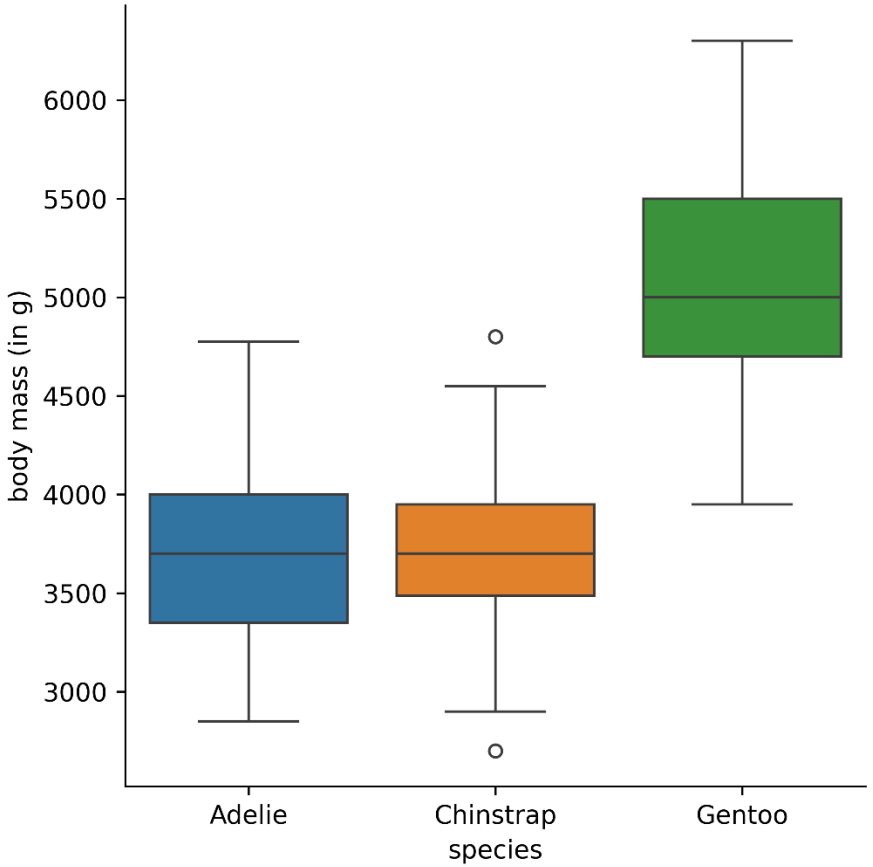
Jens Ulrik Hansen

Roskilde Universitet

Exploratory data analysis and descriptive statistics

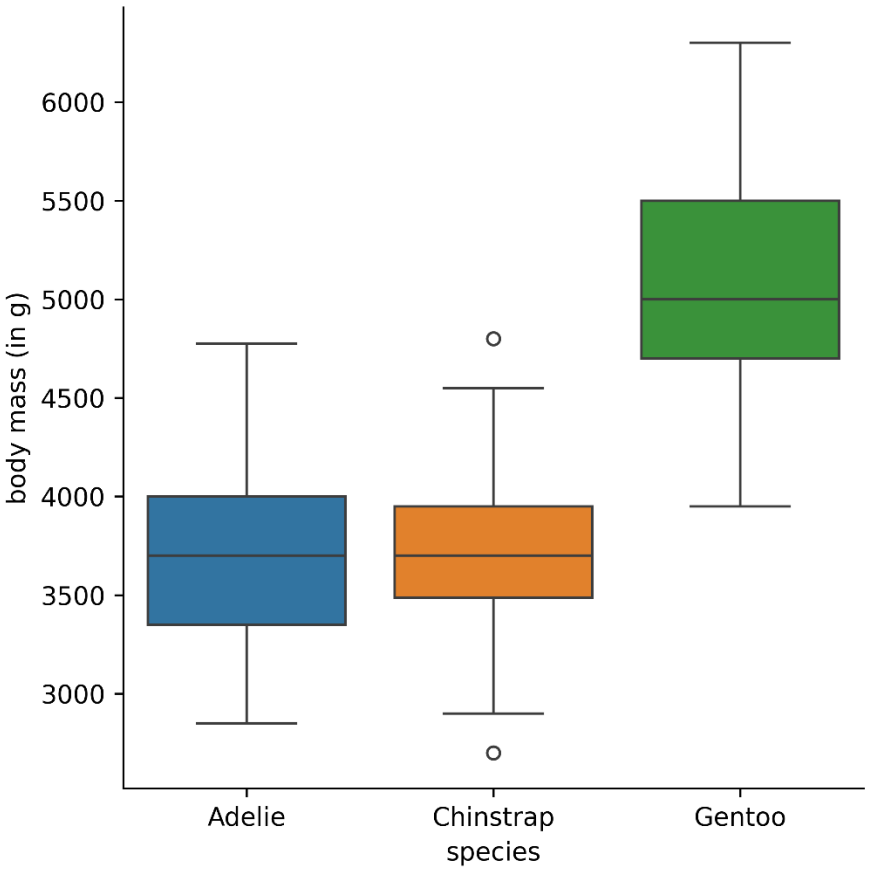
- Our exploratory data analysis of last time involved a lot:
  - Data visualizations such as bar plots, histograms, boxplots, scatterplots etc.
  - Calculations of descriptives such as mean, median, mode, variance, standard deviations, and correlations.
- Essentially, we were doing statistics… ***descriptive statistics.***

species	count	mean	std	min	25%	50%	75%	max
Adelie	151.0	3700.662252	458.566126	2850.0	3350.0	3700.0	4000.0	4775.0
Chinstrap	68.0	3733.088235	384.335081	2700.0	3487.5	3700.0	3950.0	4800.0
Gentoo	123.0	5076.016260	504.116237	3950.0	4700.0	5000.0	5500.0	6300.0



Beyond descriptive statistics

- We saw in the exercises last week that we can partly distinguish between different species of penguins based on their body mass…
  - It looks like Adelie and Gentoo penguins have different body mass on average, while Adelie and Chinstrap penguins on average have the same body mass…
  - How do express and decide this rigorously, and how can we make sure it is true about all penguins, when our dataset only contains data on 344 penguins?
    - More generally, to what extent can we be certain that the differences and relationships we find in exploratory data analysis actually holds true?
  - ***Inferential statistics*** can help us with exactly this…



Outline of this lecture

- What is statistics?
- A primer on probabilities, distributions, and a brief recap of descriptive statistics
- Central statistical concepts: Hypothesis testing, p-values, and significance level
- Comparison between groups
- Relationship between variables
- Exercises

We will talk much more about this next time

Outline of this lecture

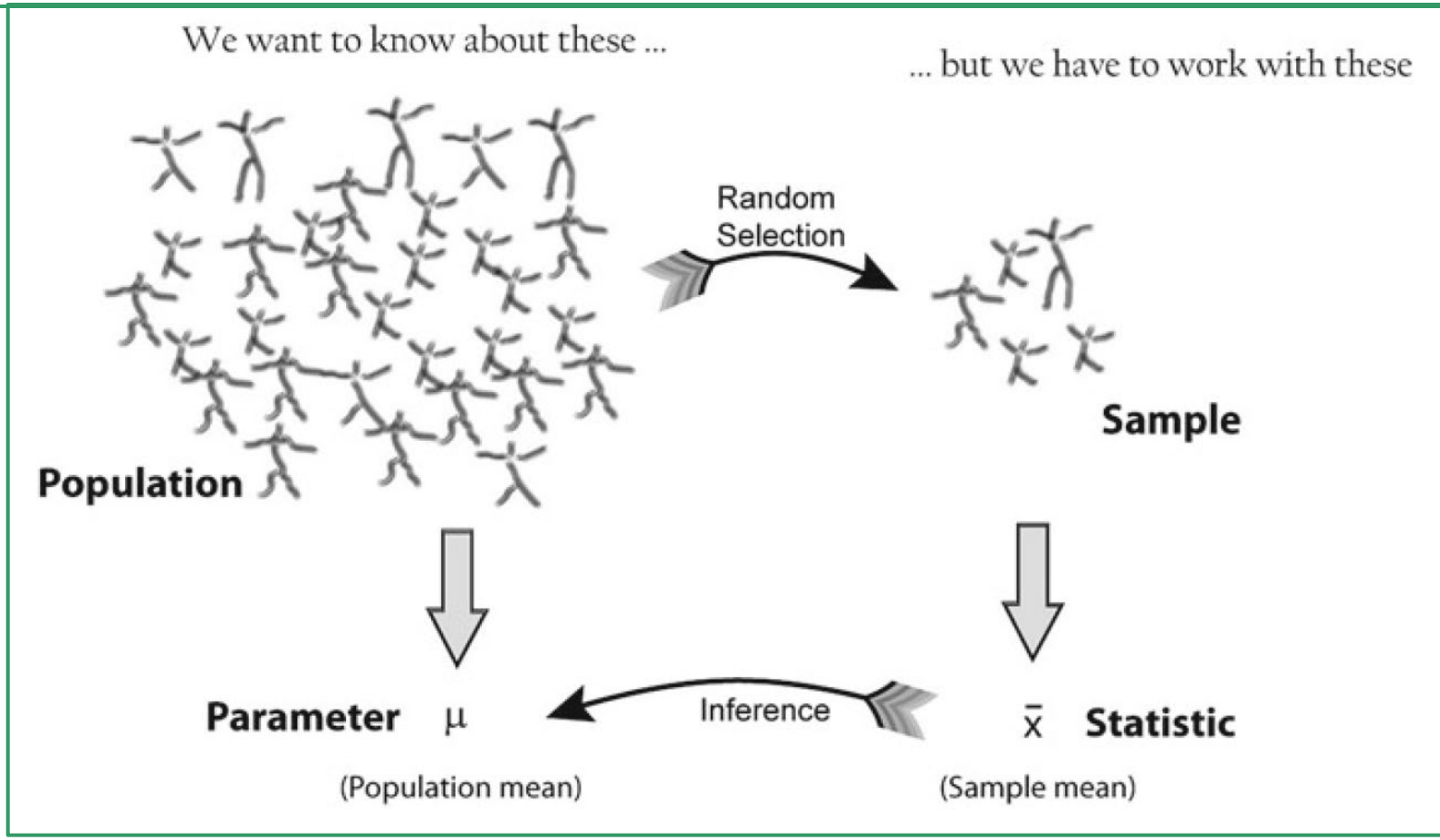
- What is statistics?
- A primer on probabilities, distributions, and a brief recap of descriptive statistics
- Central statistical concepts: Hypothesis testing, p-values, and significance level
- Comparison between groups
- Exercises

What is statistics?

- A central question when dealing with data: ***Is the fluctuation and differences observed in our data due to randomness or not?***
  - Statistics can also be viewed as the science of finding patterns in imperfect, noisy, or partial data, or as the science of quantifying uncertainty.
- Two types of statistical analysis:
  - **Descriptives statistics:** Identifying and describing patterns in a specific dataset
    - We did this last time in Exploratory Data Analysis when we calculated mean, median, … etc.
  - **Inferential statistics:** Using a dataset to answer questions about a larger population – by analyzing a given dataset what can we infer about the larger population it was sampled from?
    - In classical statistics we do this through ***hypothesis testing*** or ***parameter estimation***

What is statistics?

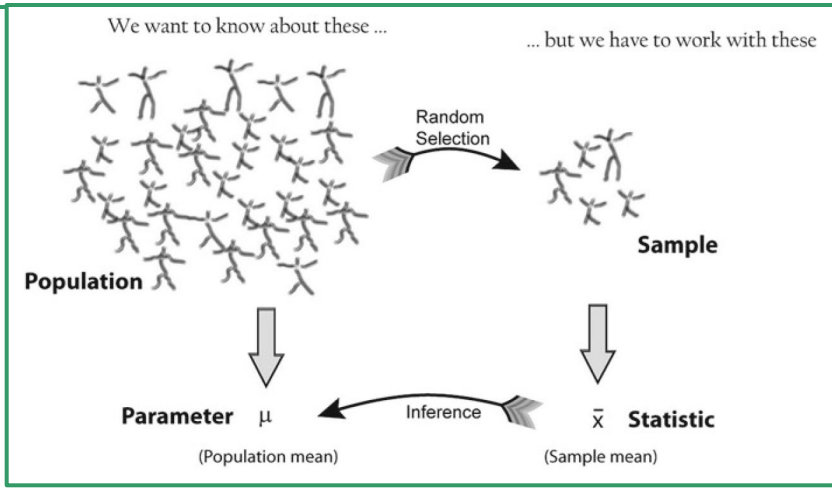
- **Population vs sample**
  - The population is all the individuals we are interested in
  - The sample is the subset of the population for which we have data
  - Inferential statistics example: From knowing the mean of the sample, what can we infer about the mean of the population?
    - How good is the sample mean as an estimate of the population mean (parameter)?



Haslwanter, T. (2022). *An Introduction to Statistics with Python - With Applications in the Life Sciences*. Springer, Cham.

What is statistics?

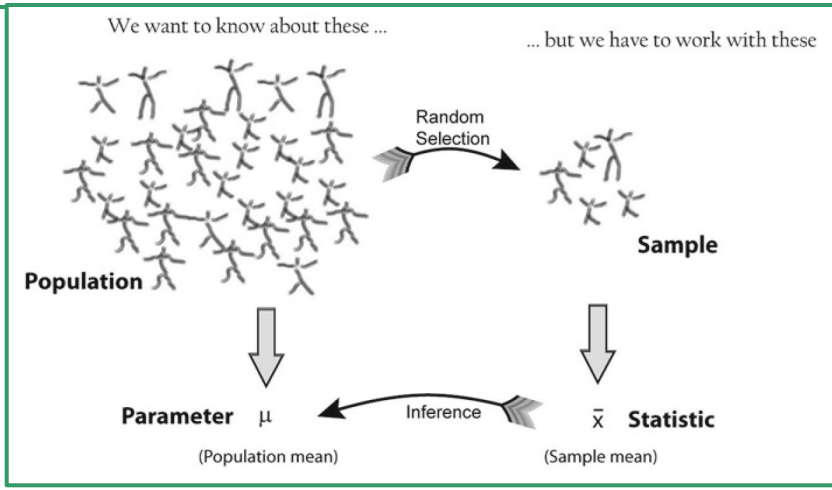
- **Further examples**
  - ***Election polling***
    - The population is all the voters (who will vote)
    - The sample is the approx. 1000 persons we asked
    - To what extent do the distribution of the votes in the sample resembles what we can expect of the entire population?
  - ***Randomized Controlled Trials of new drugs***
    - The population is all future people with a given disease
    - The sample is the people participating in the experiment
    - If the drug works for the people in the experiment how certain can we be that it will work for the entire population?



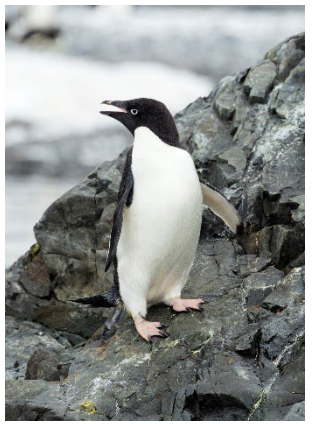
Haslwanter, T. (2022). *An Introduction to Statistics with Python - With Applications in the Life Sciences*. Springer, Cham.

What is statistics?

- **Assumptions of inferential statistics**
  - The sample must be ***representative*** of the population, i.e. share the same characteristics as the population and follow the same distribution (the proportion of different groups should be the same in the sample as the population etc.)
  - ***The right hypothesis test*** must be used that match both the research question, the data sample, and theoretical assumptions.
- **Example**
  - Assume that our penguin data is ***representative*** of all penguins
  - Assume that the body mass of a penguin is normally distributed (***theoretical assumption***)
  - Then we can estimate the average body mass of an Adelie penguin (***the research question***) by the sample mean of the body mass of Adelie penguins (***the right data sample***) in our data sample.
  - We can also test whether there is a statistically significant difference in body mass between Adelie and Gentoo penguins (the research question) by preforming a Student t-test (***the right hypothesis test*** given the assumption of body mass being normally distributed and the number of samples in our dataset).



Haslwanter, T. (2022). *An Introduction to Statistics with Python - With Applications in the Life Sciences*. Springer, Cham.



Wikipedia

Outline of this lecture

- What is statistics?
- A primer on probabilities, distributions, and a brief recap of descriptive statistics
- Central statistical concepts: Hypothesis testing, p-values, and significance level
- Comparison between groups
- Exercises

A primer on probabilities, distributions, and a brief recap of descriptive statistics

- **Random Events**
  - The role of dice or repeated flip of a coin, for instance. Or drawing a random reel number between 0 and 1.
  - Events are for instance: “the dice rolled 5”, “the dice rolled a number below 4”, “we got 4 heads on 10 flip of a coin”, “we drew the real number 6.4532”, “we drew a real number between 0.3 and 0.4”.
- **Probabilities**
  - Random events might be assigned probabilities (Notation: “P(A)”, if A is the event)
  - Probabilities are numbers between 0 and 1.
  - The probability of a (fair) dice rolling 5 is 1/6, the probability that a (fair) dice rolled a number below 4 is ½, etc.
  - The probability of randomly drawing the real number 6.4532 is meaningless as we have infinitely many reel number, but the probability of randomly drawing a real number between 0.3 and 0.4 make sense and is 0.1.

A primer on probabilities, distributions, and a brief recap of descriptive statistics

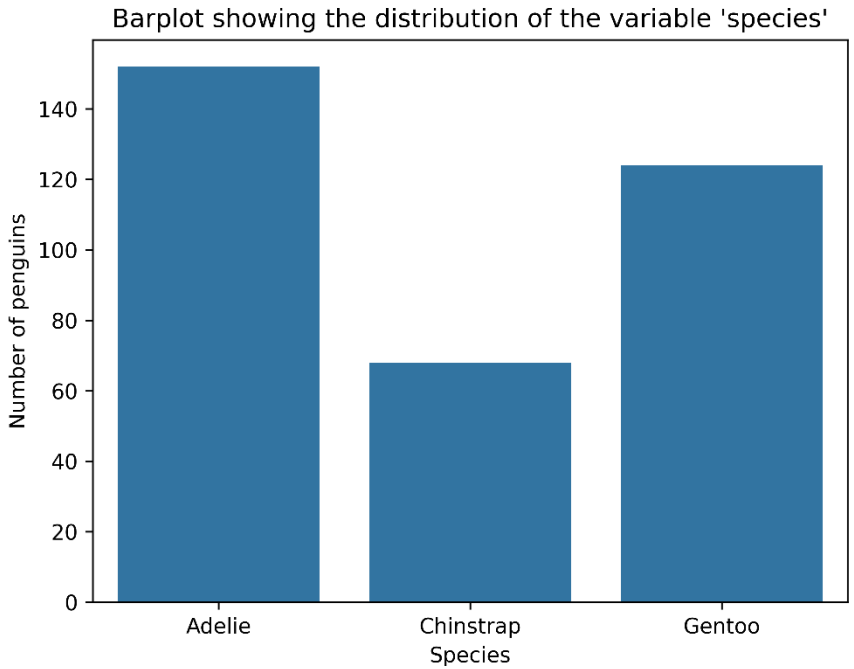
- **Laws of probabilities**
  - The probability of something certain is 1, the probability of something impossible is 0.
  - If the probability of an event A is P(A), the probability of "not A" is 1 – P(A).
  - The probability of event "A or B" is: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - If A and B are independent: $P(A \cup B) = P(A) + P(B)$
  - The probability of event "A and B", if A and B are independent: $P(A \cap B) = P(A) * P(B)$
  - The conditional probability of A given B: $P(A|B) = P(A \cap B) / P(B)$
  - Bayes theorem: $P(A|B) = P(B|A)*P(A) / P(B)$

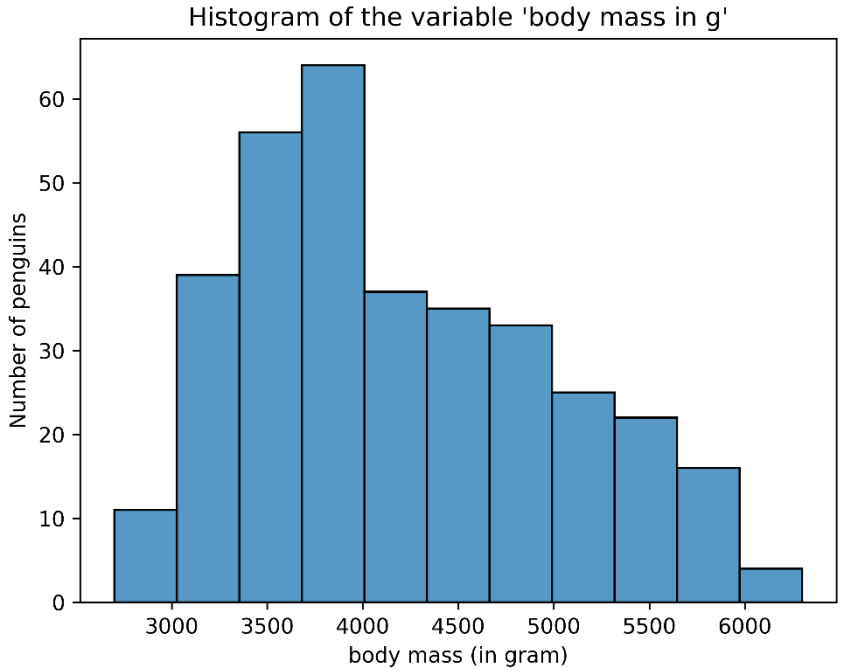
A primer on probabilities, distributions, and a brief recap of descriptive statistics

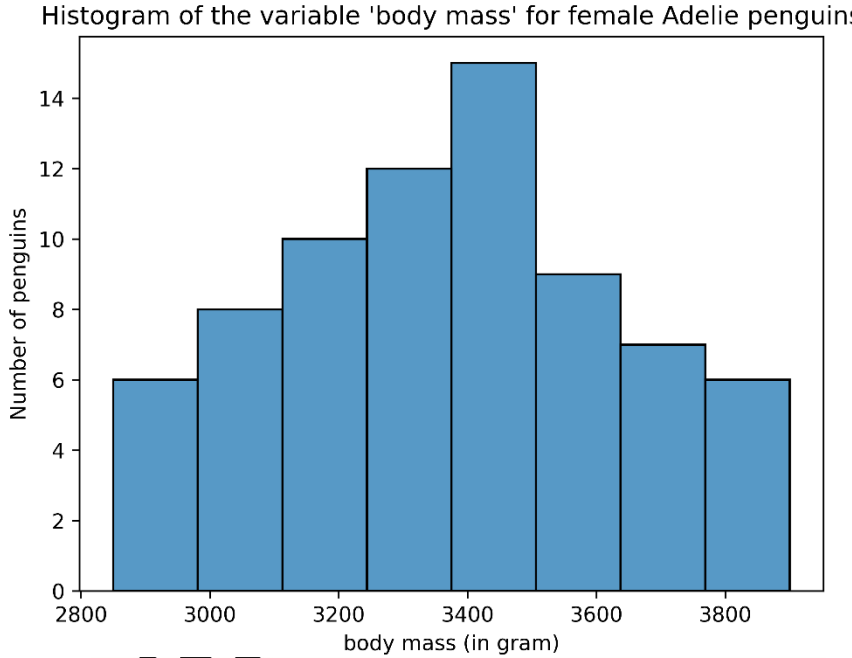
- **Random variables (/Stochastic variables)**
  - *A quantity which depends on random events*
  - *A variable that takes on numerical values dependent on the outcome of a chance event*
- Examples:
  - The eyes of a rolled dice
  - The sum of eyes for the roll of two dices
  - The number of heads in ten flip of a coin
  - The body mass of random penguin
  - The age of a random passenger on Titanic
  - The Passenger class of a random passenger on Titanic
  - …
- We can often view the features/attributes/columns of our datasets as random variables – it is a something that can be measured of our cases/observations/individuals
- We can distinguish between discrete/categorical and continuous/numeric random variables

A primer on probabilities, distributions, and a brief recap of descriptive statistics

- The ***distribution*** of a variable is how the individual values of the variable distribute them selves
  - If we have a sample of data, we can talk about the sample distribution and visualize it through, bar plots, histograms or boxplots

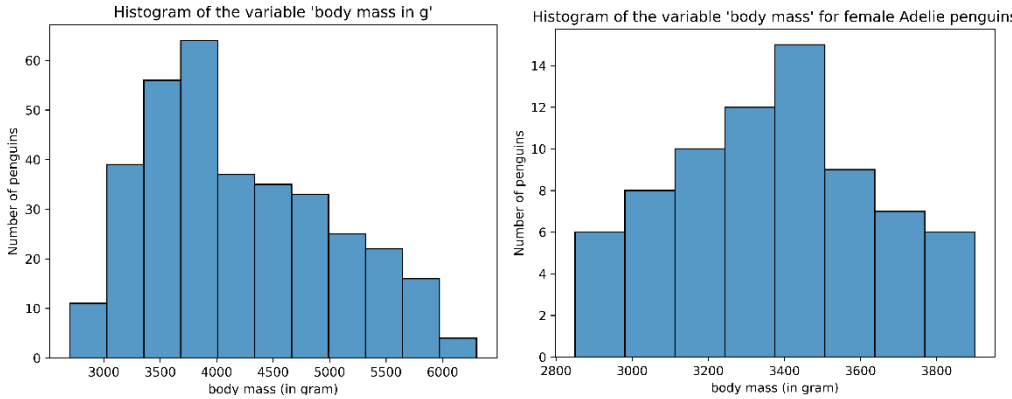


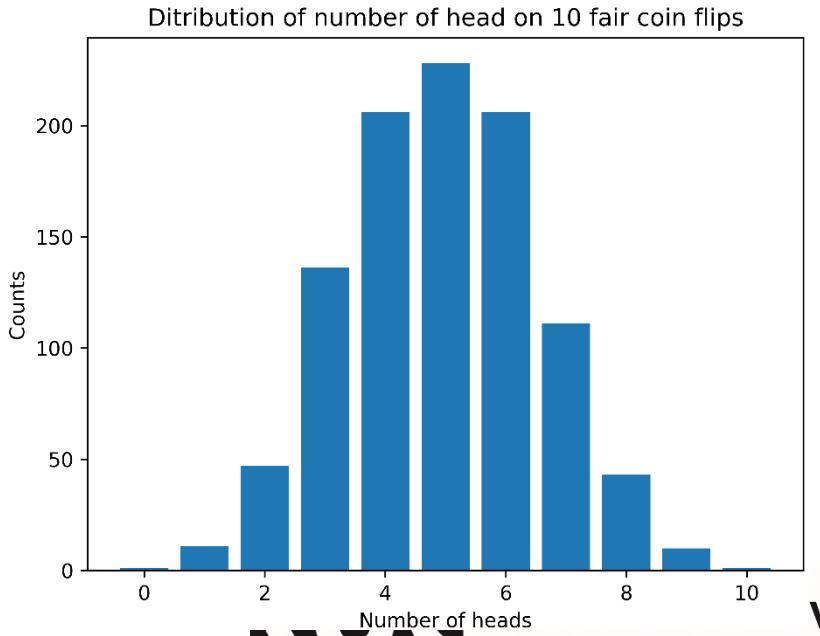




A primer on probabilities, distributions, and a brief recap of descriptive statistics

- The theoretical ***distribution*** of a random variable is probability that the random variable take on a particular value (or a value below a certain threshold).
  - If we have a population in mind (say penguins), we can talk about the theoretical distribution of a variable (say body mass in g) as the assumed distribution of the data – we may assume that body mass for pinguins are normally distributed
  - Similar for flipping a coin ten times, we theoretically assume that the number of heads is binomially distributed (with probability of heads 0.5)





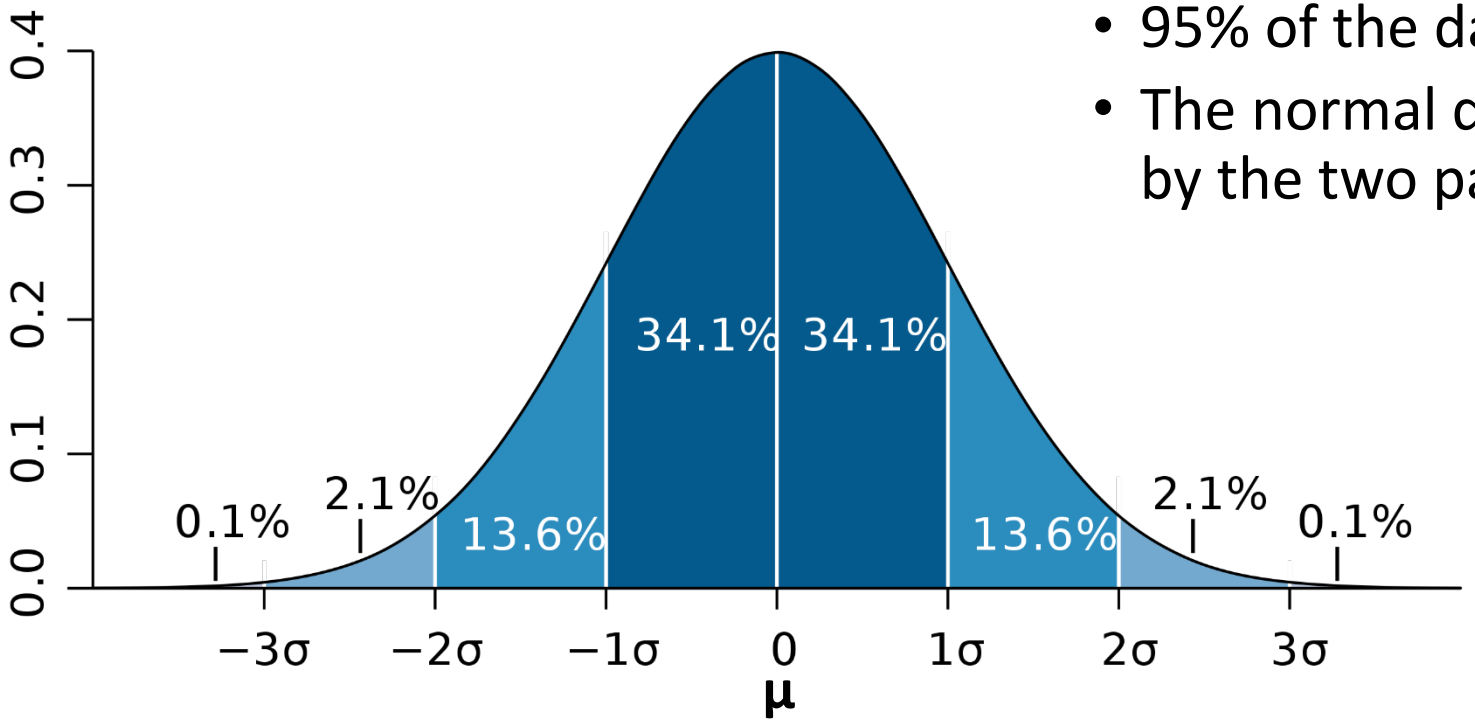
A primer on probabilities, distributions, and a brief recap of descriptive statistics

- **Descriptive statistics of a distribution**
  - For a distribution of a numeric variable, we can calculate:
    - Central tendencies: mean, median, mode, quantiles
    - Spread tendencies: variance, standard deviation
  - For a distribution of a categorical variable, we can calculate counts

	body_mass_g
count	73.000000
mean	3368.835616
std	269.380102
min	2850.000000
25%	3175.000000
50%	3400.000000
75%	3550.000000
max	3900.000000

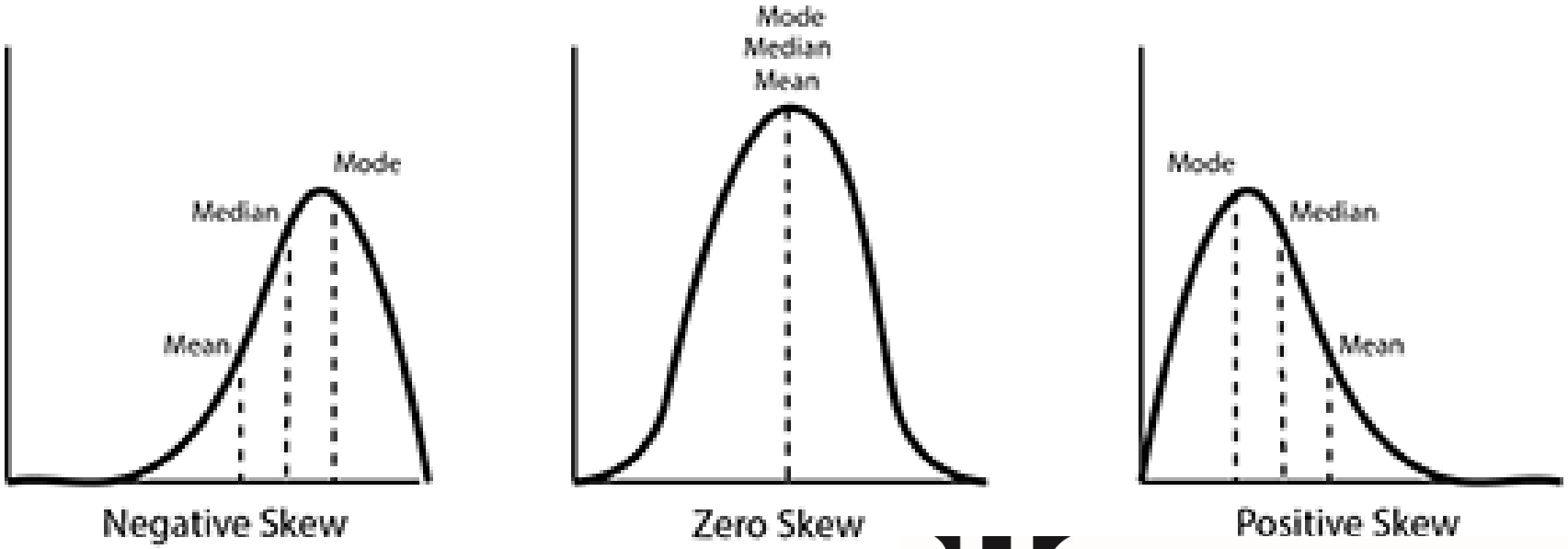
A primer on probabilities, distributions, and a brief recap of descriptive statistics

- **The normal distribution**
  - **μ** (my) denotes the mean
  - **σ** (sigma) denotes the standard deviation
  - 95% of the data falls within μ±1.96*σ
  - The normal distribution is completely specified by the two parameters μ and σ



A primer on probabilities, distributions, and a brief recap of descriptive statistics

- **Skewed distributions**
  - Left skewed = negative skewed
  - Right skewed = positive skewed
  - For skewed distributions, the median is a better measure of central tendency than the mean



Outline of this lecture

- What is statistics?
- A primer on probabilities, distributions, and a brief recap of descriptive statistics
- Central statistical concepts: Hypothesis testing, p-values, and significance level
- Comparison between groups
- Exercises

Central statistical concepts: Hypothesis testing

- Hypothesis testing, generally: A hypothesis about the populations is derived from the research question and data is collected to investigate the plausibility of the hypothesis
  - Like in the scientific method, where we might propose a hypothesis that we try to falsify (Universal laws can never be proved conclusive by data, but we can reject universal laws with a single data counter example – A black swan is a counter example to the universal laws "all swans are white")
- Similarly in the hypothesis testing of classical statistics, we propose a hypothesis, called the ***null hypothesis*** ($\boldsymbol{H_0}$) that we try to reject
  - If we can reject the null hypothesis it is considered as support of the ***alternative hypothesis*** ($\boldsymbol{H_1}$), which is the "opposite" of the null hypothesis
  - This type of hypothesis testing is also sometimes referred to as ***statistical significance testing***

Central statistical concepts: Hypothesis testing

- **Examples:**
  - Testing a new drug, the null hypothesis will be that it has no effect on the status of the disease in the patients. We can then conduct a randomized controlled experiment and collect data on which we can perform a hypothesis test. If the hypothesis test reject the null hypothesis that the drug has no effect, we conclude the alternative hypothesis that the drug did have an effect.
  - In advertisement we might want to investigate whether there is a correlation between the number of people who sees an add and how many product we sell. The null hypothesis will be there is no correlation. If, after doing a hypothesis test on the company's historical marketing and sales data, we can reject the null hypothesis, we are left with the alternative hypothesis that there is a correlation.

Central statistical concepts: p-values

- Preforming hypothesis testing
  - We calculate the probability of observing the data sample we actually have (or something more extreme), if the null hypothesis is true – this probability is called the ***p-value***
    - *This calculation requires assumptions about the what theoretical distribution generated the observed data (for parametric tests)*
  - ***If the p-value is large***, that means that is it fairly common to observe the data we observed if the null hypothesis is true and therefore, ***we cannot reject the null hypothesis***.
  - However, ***if the p-value is small enough***, we observed some data that was highly unlike to observe. Instead of accepting that we observed a really rare event, ***we instead reject the null hypothesis***.
  - Questions:
    1) When is p-value small enough?
    2) How do we calculate the p-value?

Essentially an empirical question

Different type of statistical tests do this differently based on assumption of the theoretical distribution

Central statistical concepts: Significance level

- When is a p-value small enough?
  - Whenever it is smaller than the ***significance level*** that was set when one proposed the research question…
  - Setting the significance level depends on the research question and the level of strength one wants associated with the hypothesis test
  - A commonly used significance level is ***0.05*** – which corresponds to rejecting the null hypothesis if there is less 5% probability of observing the data sample one observed
  - A smaller significance level will make it harder to reject the null hypothesis, but also make the evidence stronger, if one succeed in rejecting the null hypothesis

TABLE 9.1
p–value and Significance

p-value	Significance
>0.1	Little or no evidence of a difference or relationship.
0.05–0.1	Weak evidence of a difference or relationship.
0.01–0.05	Evidence of a difference or relationship.
0.001–0.01	Strong evidence of a difference or relationship.
<0.001	Very strong evidence of a difference or relationship.

Chapter 9: "Statistical Analysis with Python" by Han-I Wang, Christos Mandolas, and Dimitrios Xanthidis in the *Handbook of Computer Programming with Python*

Central statistical concepts: Choosing the type of statistical test

- How do we calculate the p-value?
  - Based on various factors, such as type of data (categorical or numerical), how it was collected, the sample size, and how it is distributed, we decide on a specific hypothesis test.
  - The specific hypothesis can calculate the p-value based on the assumption about the theoretical distribution that generated our sample data.
  - Each hypothesis test give us a p-value, we can compare to our significance level
  - We will look at various tests used for four particular cases:
    - Comparison between groups
    - Relationship between variables

Central statistical concepts: Choosing the type of statistical test

- We can also distinguish between ***parametric tests*** and ***non-parametric tests***
  - It is essentially a choice in how we assume our sample data to be distributed
  - **Parametric tests** make stricter assumptions about our sample data, such as it being ***normally distributed*** (parametric means that the distribution of the data is characterized by a finite set of parameters – the normal distribution is characterized by the mean and the standard deviation) and ***the sample size of each group is large enough***.
  - Parametric tests have more statistical power and can therefore detect significance more often than non-parametric tests
  - Parametric tests can test difference in means, while non-parametric tests may test difference in medians.
  - However, if we cannot be sure that the data is distributed in a nice way (as in the normal distribution), we cannot use parametric tests and must use non-parametric tests

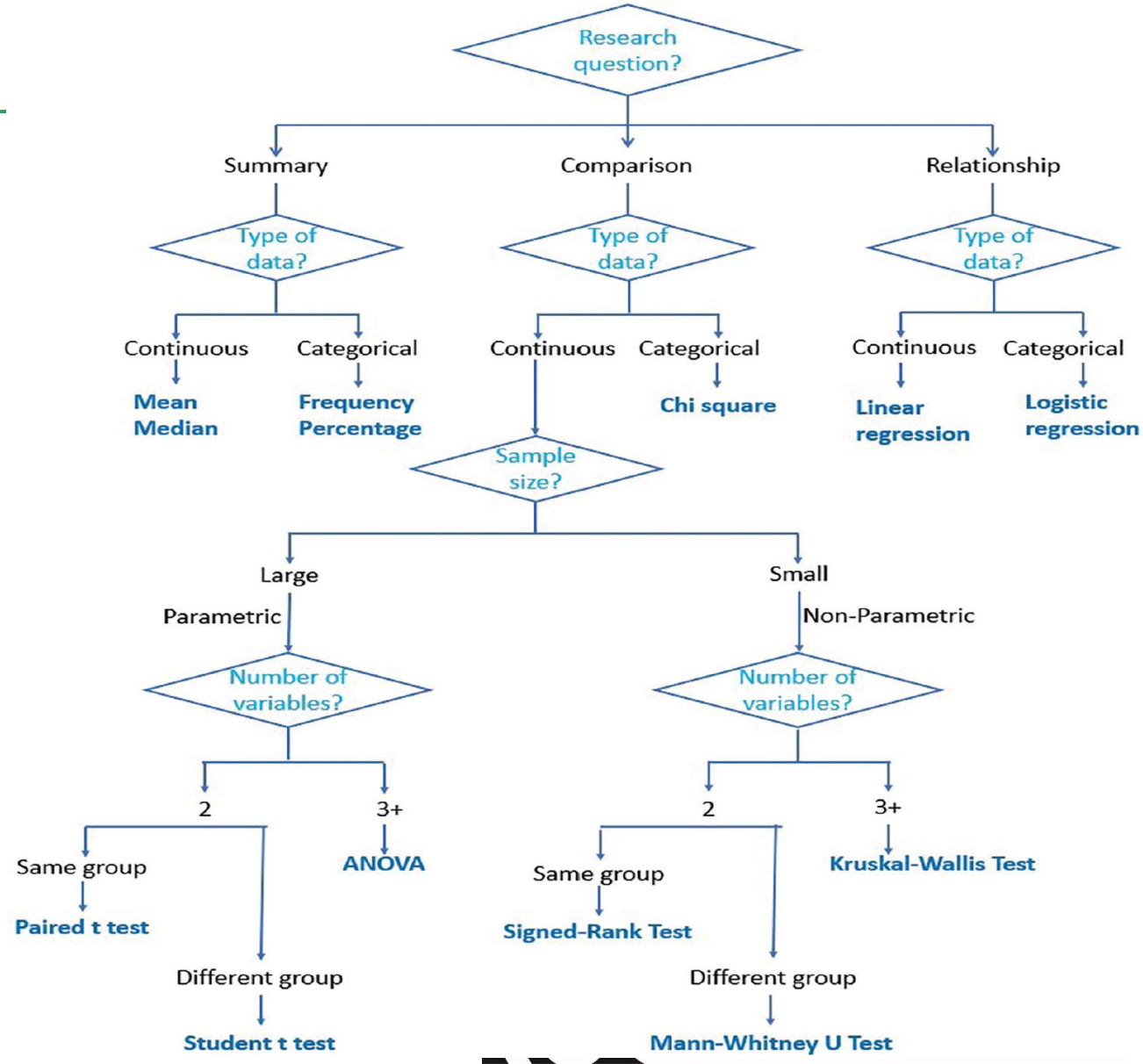
TABLE 9.4

Simple Guide for Choosing between Parametric and Non-Parametric Tests

Non-Parametric Tests	Sample Size	Parametric Tests
Mann-Whitney U test	$N=15$ in each group	Independent Student t-test
Wilcoxon Signed-Rank test	$N=30$	Dependent (Paired) Student t-test
Kruskal-Wallis, Mood's median test	Compare 2–9 groups, $n=15$ in each group	Analysis of Variance test (ANOVA)
	Compare 10–12 groups, $n=20$ in each group	

Central statistical concepts: Choosing the type of statistical test

- A decision tree for choosing the right statistical test from Chapter 9: "Statistical Analysis with Python" by Han-I Wang, Christos Mandolas, and Dimitrios Xanthidis in the *Handbook of Computer Programming with Python.*
  - *Same group* means paired groups
  - *Different group* means non-paired/independent groups



Outline of this lecture

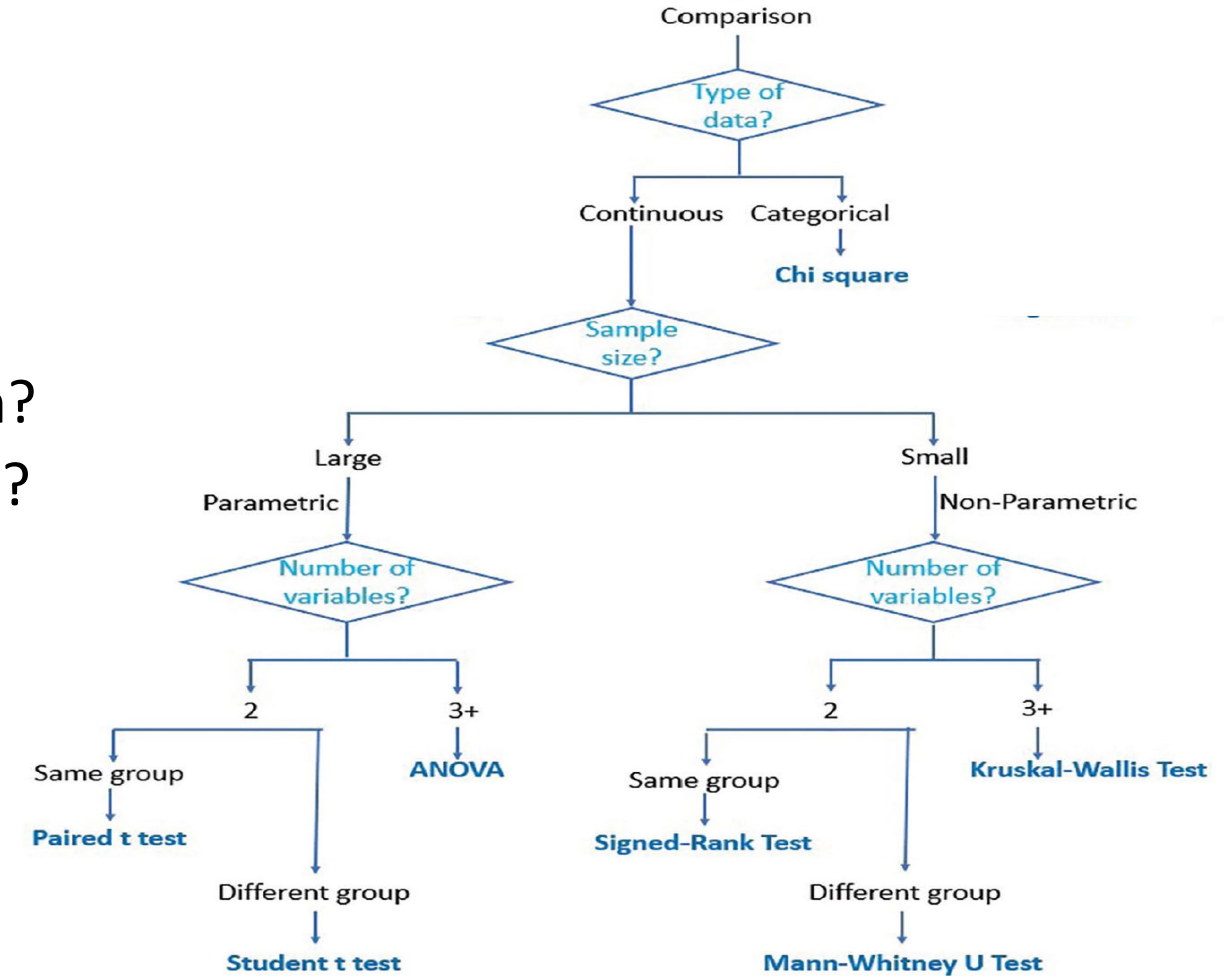
- What is statistics?
- A primer on probabilities, distributions, and a brief recap of descriptive statistics
- Central statistical concepts: Hypothesis testing, p-values, and significance level
- Comparison between groups
- Exercises

Comparison between groups

- ***Examples***
  - Is the average height among Danish and Dutch people different?
  - Is the blood pressure of patients different before and after taking a new drug
  - Is the time it take people to run a marathon different depending on whether they trained using method A or method B?
  - Are our video posts getting more reactions than our image posts on Instagram?
  - Is this implementation A of an algorithm using less resources than implementation B?
- ***Paired or non-paired tests***
  - Non-paired or independent tests: the two groups consists of different individuals
  - Paired or dependent tests: the two groups consists of the same individuals measured before and after an event

Comparison between groups

- **Choosing the right test**
  - Is it categorical or numeric data?
  - How big is your sample sizes?
  - Are your data normally distributed?
  - Do the groups have same distribution?
  - How many groups are you comparing?
  - (Note, we might still want to use a non-parametric test if we have a lot of data, but it is not normally distributed)



Comparison between groups

- **The tests** (for Python we use “from scipy import stats”)
  - ***Mann-Whitney U test***. Non-parametric, numeric variables, small sample size or skewed data.
    - Test whether the distribution of two independent samples are equal
    - With Scipy: stats.mannwhitneyu
  - ***Kruskal-Wallis test.*** Non-parametric, numeric or ordinal variables, small sample size and/or not normally distributed, but the distributions have same shape (and same skewness). (can also be used for more than two groups)
    - Test whether the distribution/median of two independent samples are equal
    - With Scipy: stats.Kruskal
  - ***Willcoxom Signed-rank test***. Non-parametric, numeric or ordinal variables, small sample size or skewed data
    - Test whether the distribution of two paired samples are equal
    - With Scipy: stats.Wilcoxon

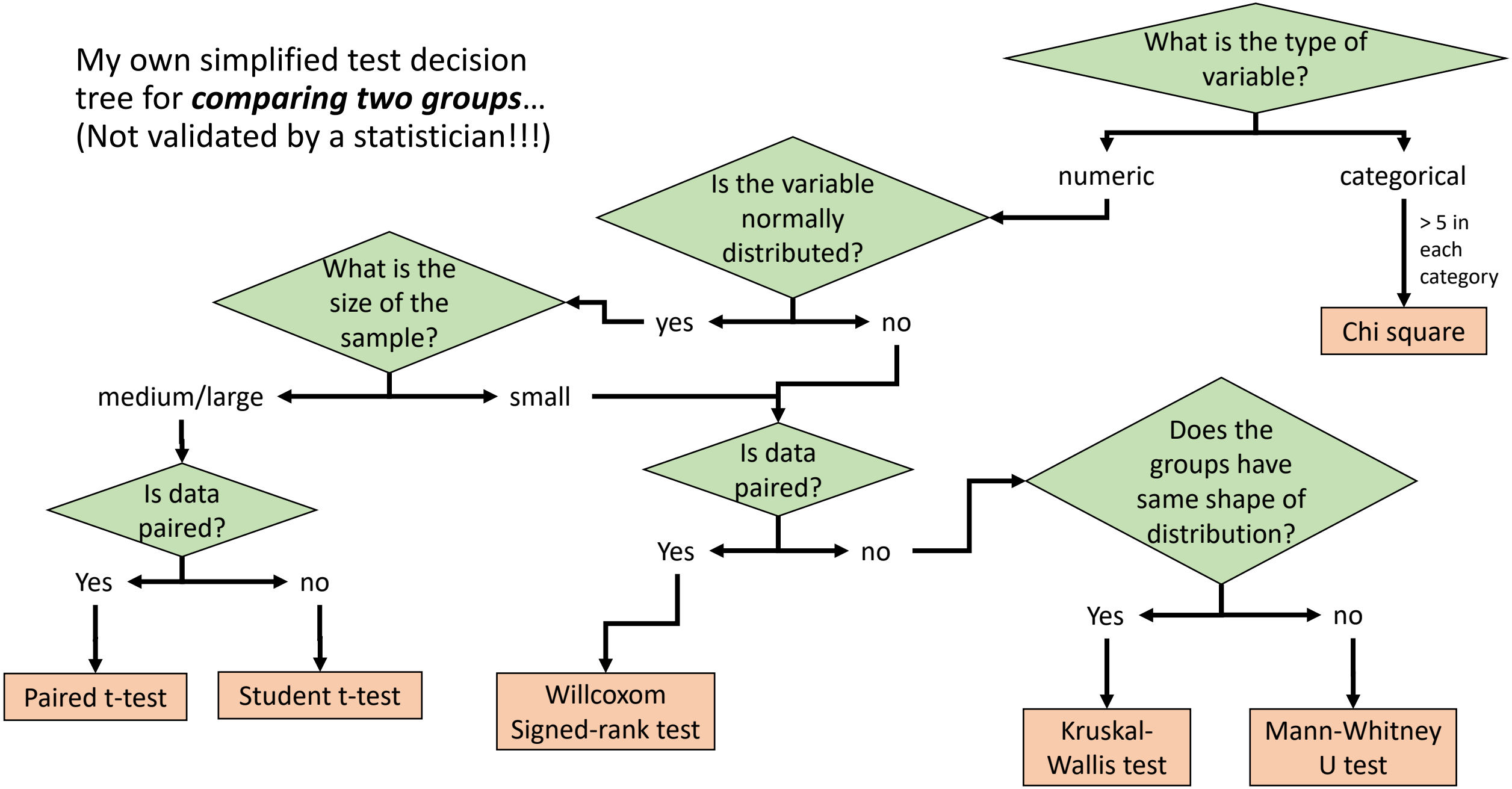
Comparison between groups

- **The tests** (for Python we use “from scipy import stats”)
  - ***Student t-test (Independent t-Test).*** Parametric, numeric variables, normally distributed data with no significant outliers.
    - Test whether the mean of two independent samples are equal
    - With Scipy: stats.ttest_ind
  - ***Paired t-test***. Parametric, numeric variables, normally distributed data with no significant outliers.
    - Test whether the mean of two paired samples are equal
    - With Scipy: stats.ttest_rel

Comparison between groups

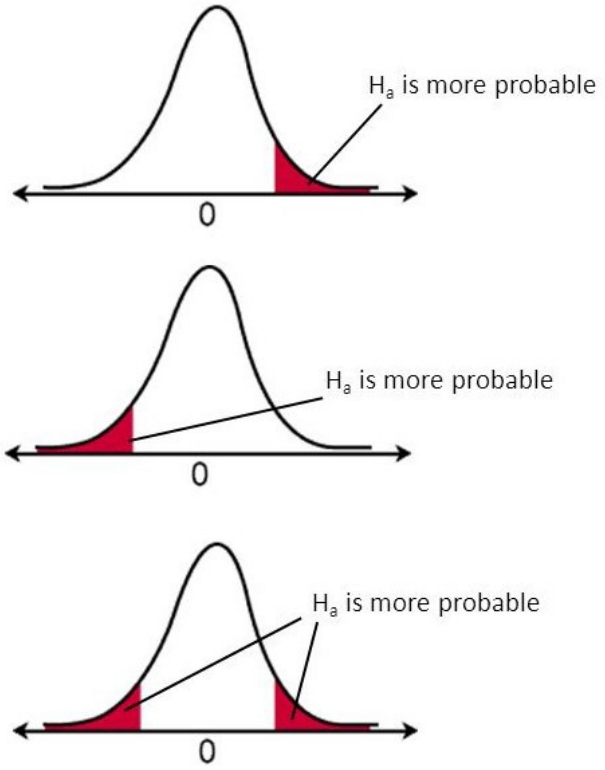
- **The tests** (for Python we use “from scipy import stats”)
  - ***ANOVA***. Parametric, numeric variables, normally distributed with homogeneity of variance.
    - Test whether the mean of several independent samples are equal
    - With Scipy: stats.f_oneway
    - It only test whether there is at least one difference in means. It does not tell between which particular groups, there is a difference!
  - ***CHI-Square***. Parametric, categorical variables, single sample, more than 5 observations in each group.
    - Tests whether categorical data from a single sample follow a specific distribution or is similar to another set of categorical data.
    - Like the t-test for categorical data
    - With Scipy:
      - stats.chisquare to compare a sample to expected values calculated from a known distribution
      - stats.chi2_contingency if you have a contingency table of two categorical variables.

My own simplified test decision tree for ***comparing two groups***…
(Not validated by a statistician!!!)



Comparison between groups

- **Important comments**
  - The alternative hypothesis matter!
    - Is the alternative to a null hypothesis of equal means just that the means are different or that the mean for one specific group is bigger?
    - If the alternative is merely that they are different, then we have a ***two sided*** test, otherwise one sided.
    - This can be passed to ttest_ind, for instance as an extra argument (called 'alternative')
    - There is also an argument to ttest_ind specifying whether we assume equal variance or not of the two groups
  - One needs to correct the significance level or the calculation of p-values if one performs **multiple tests**.
    - While there is a small chance of getting the wrong result in one test, the likelihood multiplies for the case of numerous test
    - There are many different techniques for doing this…



Right-tail test

$H_a: \mu > \text{value}$

Left-tail test

$H_a: \mu < \text{value}$

Two-tail test

$H_a: \mu \neq \text{value}$

Comparison between groups

- Let us look at the notebook “Comparison of groups.ipynb”

Outline of this lecture

- What is statistics?
- A primer on probabilities, distributions, and a brief recap of descriptive statistics
- Central statistical concepts: Hypothesis testing, p-values, and significance level
- Comparison between groups
- Exercises

Relationship between variables

- Do the exercises in the notebook “Exercises in statistics.ipynb”