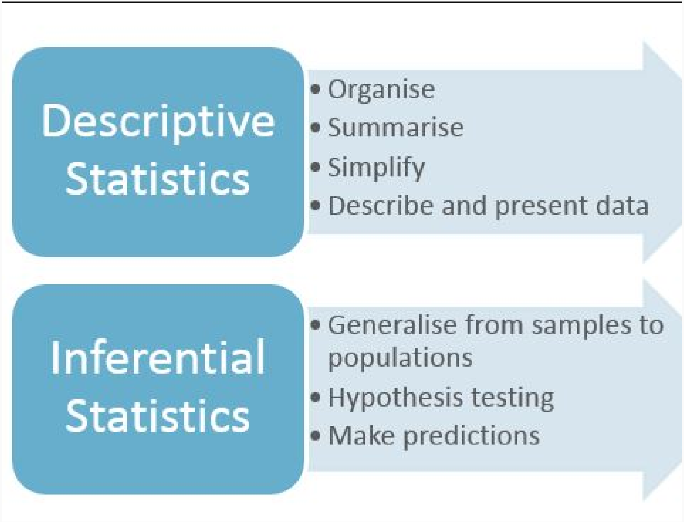
Neuroschool 2022

Descriptive Statistics and Hypothesis Testing

Stefania Sarno

June 17, 2022

Descriptive VS inferential statistics

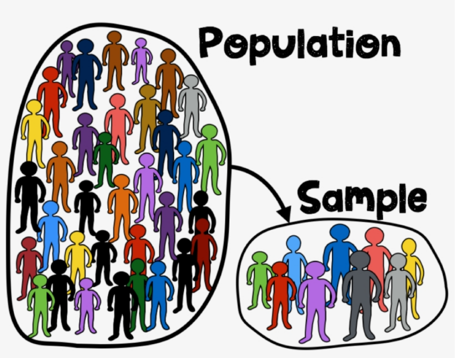


Descriptive Statistics

Descriptive statistics: outline

- Population and Sample
- Types of data (numerical, categorical)
- Graphical presentation (tables and plots)
- Measures of the centre of a set of observations
- Measures of variability

Population VS Sample

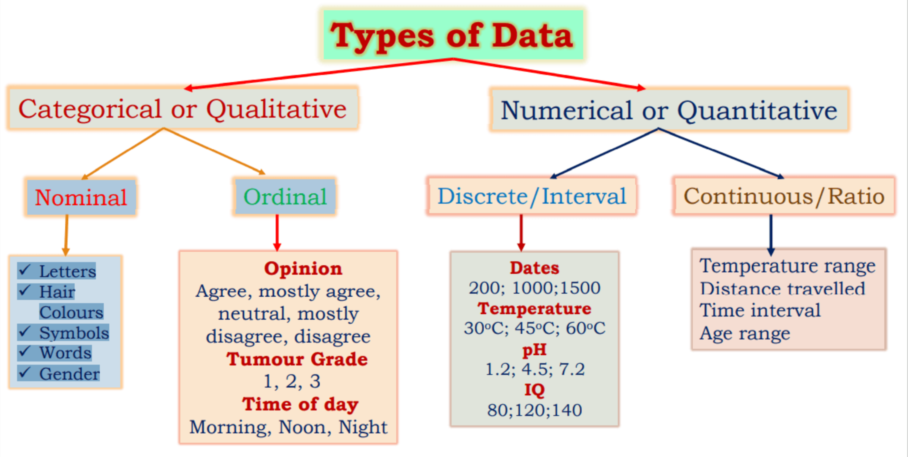


- Population: set of all individuals of interest in a particular study
- Sample: a relatively small number of individuals taken from the population.

Example: We want to examine the blood pressure of all adult males with a schizophrenia diagnosis in Ireland:

- Population is all adult males with a schizophrenia diagnosis in Ireland
- Sample is a random sample of 100 adult males with a schizophrenia diagnosis

Types of data

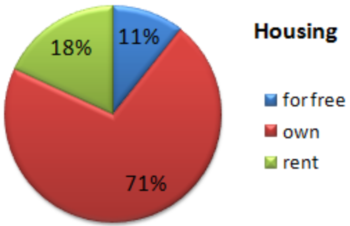


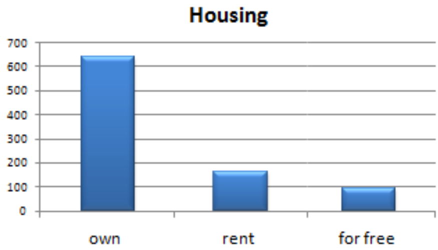
Visualization: categorical/discrete numerical data

Tables, Pie Charts, Bar Charts

Frequency Table

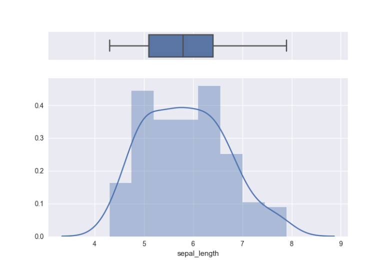
Housing	Count	Count%
for free	96	10.67%
own	641	71.22%
rent	163	18.11%

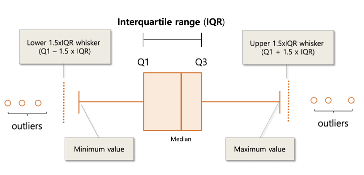




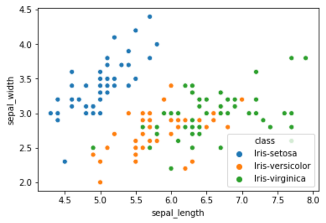
Visualization: continuous numerical data

Histograms and box plots





Scatter plots: to investigate correlations or relationships between two sets of measurements



Measuring the centre of the observations

- Suppose we have a set of numerical observations and we want to choose a single value that will represent this set of observations
- How do we choose such a value? What is meant by the average of a set of observations?
- We will look at 3 measures of the centre of the observations:
  - Median
  - Mean
  - Mode

Median

Individual #	IQ Score
1	75
2	81
3	79
4	69
5	85
6	98
7	100
8	102
9	76
10	84

- Rank the observations, i.e., write them down in ascending order:
  69, 75, 76, 79, 81, 84, 85, 98, 100, 102
- When n (total number of observations) is odd:
  median = $((n+1)/2)^{th}$ observation
- When n is even:
  median = half way between the $(n/2)^{th}$ observation and the $((n/2)+1)^{th}$ observation
- Here n is even:
  69, 75, 76, 79, 81, 84, 85, 98, 100, 102
  median = (81 + 84)/2 = 82.5

Mean

- The arithmetic mean is defined to be the sum of all the observations divided by the number of observations:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where x_i refers to each of the individual observations

- $\bar{x}$ is used as an estimate of , the mean of the population μ

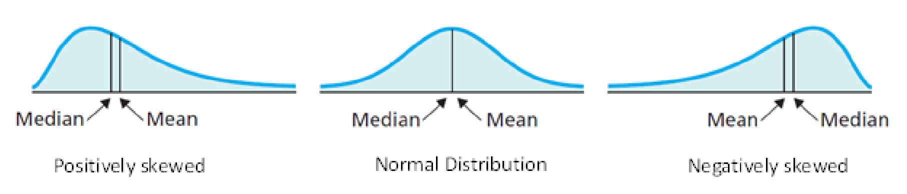
- For the IQ data above, the mean IQ is 84.9
mean = (75 + 81 + 79 + 69 + 85 + 98 + 100 + 102 + 76 + 84)/10 = 84.9

Median or mean

Individual #	IQ Score
1	75
2	81
3	79
4	69
5	85
6	98
7	100
8	~~102~~ 500
9	76
10	84

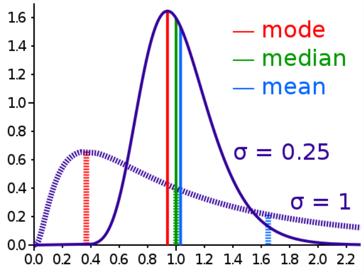
- Median is unaffected by outliers
  Here median = 82.5
- The mean, because it takes all values into account, is affected
  Here mean = 124.7
- Mean has better mathematical properties as it takes all data into account
- Median is usually used for descriptive statistics

Median or mean



- For symmetric data, the median and the mean are the same
- The median can be a better measure than the mean when the data are skewed

Mode



- The mode is that value of the variable which occurs most frequently
- Less commonly used than either the mean or median
- The mode can be used for categorical measurements
- Some sets of observations may have no mode and some may have more than one mode (unimodal = 1 peak, bimodal = 2 peaks)

Measures of Variability

In statistice it is crucial to measure how our data vary, not just their central tendecy.

If there was no variability there would be no need for statistics.

Typical measure of variability are:

- Variance
- Standard deviation
- Range
- Kurtois
- Skew

Variance and standard deviation

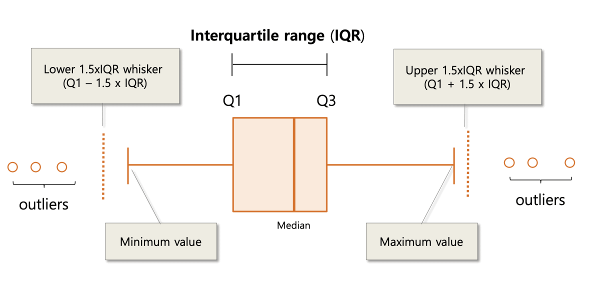
- The sample variance is defined as:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- s^2 is the sample estimate of the population variance σ^2
- The sample standard deviation is given by the square root of the variance
- Small standard deviation says the observations cluster closely around the mean, larger standard deviation says the observations are more scattered
- Standard deviation is often used as it has the same units as the mean

Range and inter-quartile range

- Range = difference between minimum an maximum values of the data
- Inter-quartile range (IQR) = difference between the upper (Q3) and lower (Q1) quartiles



- – Q1 = 25 th percentile = lower quartile = median of the lower half of the data
- – Q3 = 75 th percentile = upper quartile = median of the upper half of the data

Kurtois and skewness

Skewness

- The skewness is a measure of symmetry.
- The skewness can be defined as:

$$skewness = \frac{\sum_i (x_i - \bar{x})^3}{(n-1) \cdot s^3}$$

- The skewness for a normal distribution is zero, and any symmetric data should have a skewness near zero.

Kurtosis

- The kurtosis is a measure of peakedness or flatness of a distribution.
- The kurtosis can be defined as:

$$kurtosis = \frac{\sum_i (x_i - \bar{x})^4}{(n-1) \cdot s^4}$$

- The kurtosis for standard normal distribution is three.

Probability distributions

Probability distributions and random variable

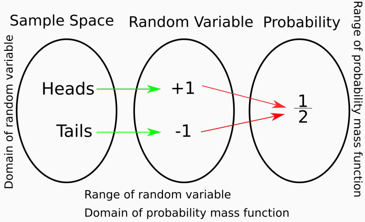
A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.

The concept of probability distribution is strictly related to that of random variable.

Random variable: its value depends on chance on different repetitions of the same experiment.

Examples:

- Toss a coin:
  - Coin comes up either a head or a tail
  - Probability of 0.5 for either a head or a tail
- Throw a dice:
  - Either a 1, 2, 3, 4, 5, 6 will come up
  - Probability of 1/6 for each of 1, 2, 3, 4, 5, 6



Common probability distributions

Commonly used probability distributions:

Discrete distributions:

- Bernoullidistribution
- Binomial distribution
- Poisson distribution
- Geometric distribution

Continuous distributions:

- Gaussian (Normal) distribution
- Chi-square distribution
- F distribution
- Student's t-distribution

Bernoulli distribution

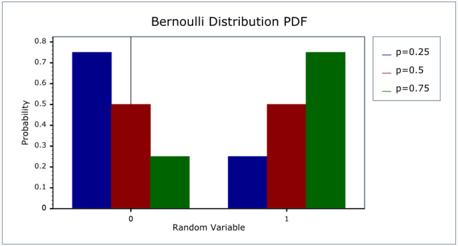
- The Bernoulli distribution is the probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability $q = 1 - p$

The probability mass function Ber of this distribution is:

$$Ber(k|p) = \begin{cases} p, & \text{if } k = 1. \\ 1 - p, & \text{otherwise} \end{cases}$$

Compact form for the Bernoulli:

$$Ber(k|p) = p^k \cdot (1 - p)^{(1-k)}$$



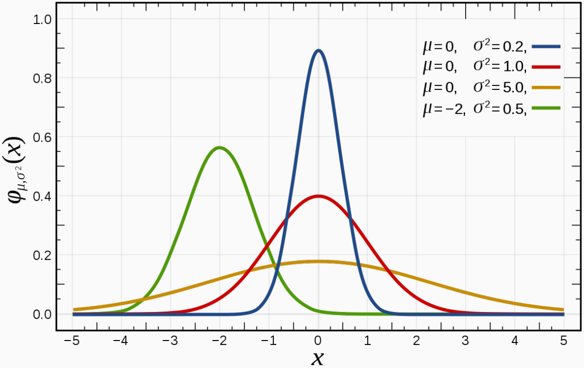
- Less formally, it can be thought of as a model for the set of possible outcomes of any single experiment that asks a yes–no question.

- It can be used to represent a coin toss where 1 and 0 would represent "head" and "tail" (or vice versa), respectively. In particular, unfair coins would have.

Normal distribution

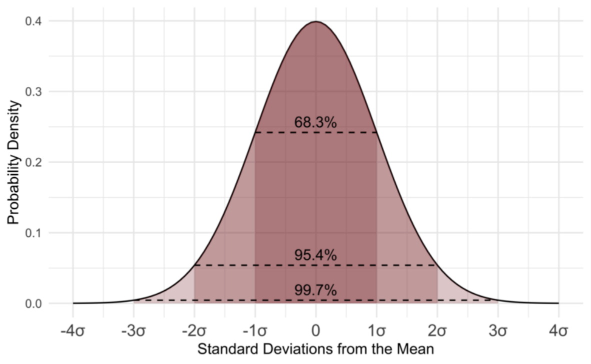
It is continuous probability distribution described/defined by two parameters, the mean μ and the standard deviation σ:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Standard normal distribution

- It is a Normal distribution with $\mu = 0$ and $\sigma = 1$
- %95 of the area is in the interval (-1.96, +1.96)



Hypothesis testing

Types of studies

Observational Studies

Methodical observation of a system without intervention

- Epidemiology: relationship between smoking and lung cancer
- Astronomy: relationship between the mass of a star and its brightness

Controlled experiment:

Manipulate one or more variables in order to determine the effect of the intervention

- Medicine: clinical drug trials
- Physics: relationship between electrical current, voltage and resistance

Types of studies

Example of an observational study:

- Background
  It is conjectured that patients with bipolar disorder tend to have a cognitive deficit (as measured by IQ) compared with unaffected people
- Objective
  To determine whether this is in fact the case

Hypothesis Testing Procedure

1. Define the research hypothesis, H_1:
   There is a difference in the mean IQ between affected and unaffected people
2. Define the null hypothesis, H_0:
   There is no difference in the mean IQ between affected and unaffected people
3. Define the significance threshold, α
   Tipycally $\alpha = 0.05$

Possible outcomes of the hypothesis test

	H_0 True	H_0 False
Reject	False Positive (Type I Error)	True Positive
Don't Reject	True Negative	False Negative (Type II Error)

Statistical Significance

- Statistical significance, p:

  The probability of rejecting the null hypothesis when it is in fact true (Type I error)

- Significance threshold, α:

  The critical value of p below which we reject the null hypothesis

Significance threshold

In theory this should depend on the experiment:

- How do we want to balance Type I and Type II errors?
- What prior evidence is there for our hypothesis?
- How important is it that we get the answer right?

In practice everyone chooses $\alpha = 0.05$

Outcome probabilities

- The probability of a false positive equals the significance threshold α
- The probability of a true negative equals $1-\alpha$ (specificity)
- The probability of a false negative is denoted β
- The probability of a true positive equals $1-\beta$ (sensitivity or power)

	H_0 True	H_0 False
Reject	α Type I Error Rate	$1-\beta$ Power
Don't Reject	$1-\alpha$ Specificity	β Type II Error Rate
	1	1

Hypothesis testing: methodology and types

Hypothesis testing typically involves four steps:

1. Formulation of the hypothesis
2. Select and collect sample data from the population of interest
3. Application of an appropriate test
4. Interpretation of the test results

We can distinguish two types of hypothesis testing

1. Parametric hypothesis test :
   it makes assumptions about the distribution of the population, typically a Normal distribution assumption
2. Nonparametric hypothesis test :
   it does not require a distribution to meet the required assumptions (we compare medians insted of means)

Common hypothesis tests: overview

- Parametric hypothesis test
  - One sample z-test
  - One/Two sample t-test
  - Chi-squared test
  - ANOVA
- Nonparametric hypothesis test
  - Mann-Whitney U test
  - Wilcoxon signed-rank test
  - Kruskal-Wallis test

Example

- The average height of males in the population is believed to be approximately 175cm
- We want to know if male patients attending particular out- patient clinics are also this tall on average.
- Null hypothesis: the mean height of male patients is the same as the average height of males:

$$H_0 :: \mu = 175\mathrm{cm}$$

- Alternative hypothesis: the mean height of male patients is not the same as the average height of males:

$$H_0 :: \mu \neq 175\mathrm{cm}$$

We use a one-sample z-test

One-Sample z-Test: assumptions

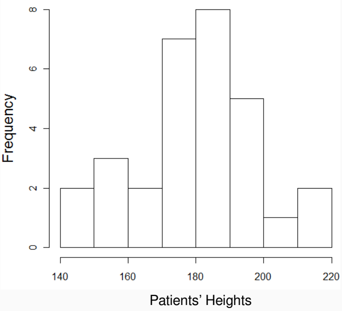
Assumptions of the one-sample z-test:

- Independent random sampling
- Large sample size (rough guide at least $n = 30$)
- Normally distributed population
- Standard deviation of the population known (or $\sigma \approx s$ for $n \geq 30$)

One-Sample z-Test: example

We collect data on the heights of 30 male patients from out-patient clinics

Patient ID	Height (cm)
01	148
02	197
03	173
04	192
05	174
⋮	⋮

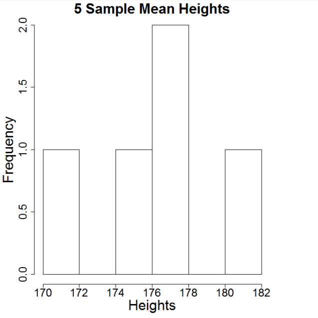


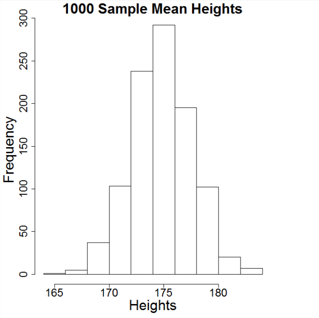
For our example data set, the sample mean is: $\bar{x} = 180.1$

Is this just by chance?

Sampling Distribution of the Mean

If we take repeated samples of size $n = 30$ from a population, we would expect the means of each of these samples to vary





- Mean of 5 sample means = 176
- SD of 5 sample means = 3.9

- Mean of 1000 sample means = 174.8
- SD of 1000 sample means = 2.7

Sampling Distribution of the Mean

If the true population mean and standard deviation are μ and σ respectively, then the sample means will have mean μ and standard deviation $\sigma/\sqrt{n}$

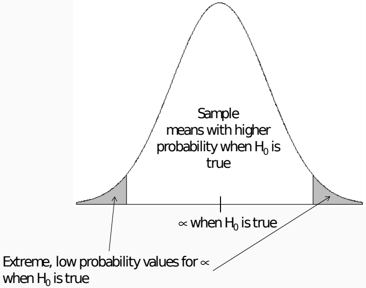
For large samples the distribution of the sample means will be Normal

Sampling Distribution of the Mean

If the true population mean and standard deviation are μ and σ respectively, then the sample means will have mean μ and standard deviation $\sigma/\sqrt{n}$

For large samples the distribution of the sample means will be Normal

When H_0 is true the sample means will not take unlikely values



One-sample z-test: test statistics

General procedure:

- Start with a normal variable that has a given mean and standard deviation
- Transform this normal variable so that it has a mean 0 and standard deviation 1
- The transformed variable has a standard normal distribution: Normal(0, 1)
- Check whether the transformed variable (test statistic) takes an acceptable value

In our example we have:

Sample Mean Hypothesized Mean

$$\frac{180.1 - 175}{2.7} = 1.9$$

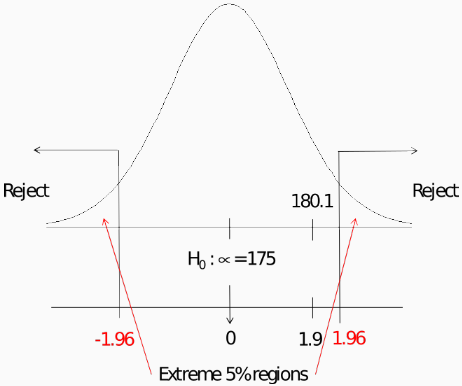
Test Statistic

Standard Error of the Mean

> The test statistic expresses the distance between the observed value and the hypothesized value as a number of standard

One-sample z-test: test statistics

The test statistics is used to decide whether the null hypothesis should be rejected



Thus at a significance level $\alpha = 0.05$ we fail to reject the null hypothesis that the mean height of the male patients attending the out-patient clinics is equal to 175 cm

Failing to Reject the Null Hypothesis

- The null hypothesis is never accepted
- We either reject or fail to reject the null hypothesis
- Failing to reject means that no difference is one of the possible explanations but we haven't shown that there is no difference
- The data may still be consistent with differences of practical importance

Hypothesis Testing: Errors

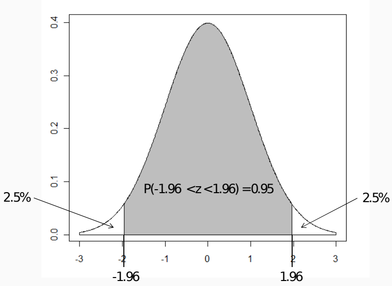
	H_0 True	H_0 False
Reject	False Positive (Type I Error)	True Positive
Don't Reject	True Negative	False Negative (Type II Error)

- The probability of a Type I Error is predetermined by the significance level α
- The probability of a Type II Error is denoted β
- The power of a statistical test is defined as $1 - \beta$ and is the probability of rejecting H_0 when H_0 is false
- A good test is one which minimises α and β

Confidence intervals

A confidence interval for a population characteristic (doesn't have to be the mean) is an interval of plausible values for that characteristic of interest

For z-test associated with 95 % confidence interval (CI):



Population mean

Sample mean

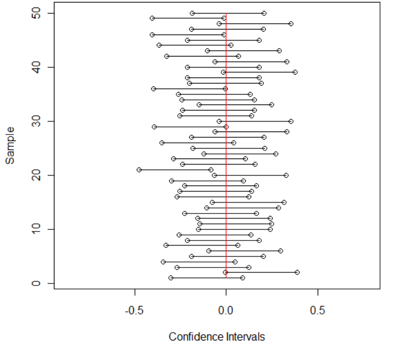
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Standard deviation of the sample mean, also known as the standard error of the mean

$$CI = \left(\bar{x} - 1.96 \cdot \sigma/\sqrt{n}, \bar{x} + 1.96 \cdot \sigma/\sqrt{n}\right)$$

Interpreting confidence intervals

- Repeatedly take samples of size *n* from the population of interest
- Calculate confidence intervals for each sample
- 95% of the time, these intervals would contain the true population value of the parameter of interest



Two Sample Hypothesis Test

Example:

- Group 1: Students received extra tuition before a test
- Group 2: Students did not receive extra tuition before a test
- Research Question: Does extra tuition help students to achieve better test scores or do they perform similarly to those who don't receive extra tuition?

Hypothesis generation:

- *Null hypothesis*: the population mean test score is the same in both groups

$$H_0 : \mu_1 = \mu_2 \qquad \text{or equivalently} \qquad H_0 : \mu_1 - \mu_2 = 0$$

- *Alternative hypothesis*: the population mean test score is not the same in both groups

$$H_0 : \mu_1 \neq \mu_2 \qquad \text{or equivalently} \qquad H_0 : \mu_1 - \mu_2 \neq 0$$

Independent two sample t-test

Assumptions:

- Two samples must be independent and random
- For partially paired data, the classical independent t-tests may give invalid results as the test statistic might not follow a t-distribution
- The underlying populations must not be skewed
- The means of the two populations being compared should follow normal distributions (often true for large samples, even when the distribution of observations in each group is non-normal)
- The two populations being compared should have the same variance

Independent two sample t-test

Test statistics:

Hypothesized value

Sample difference in means

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{se(\bar{x}_1 - \bar{x}_2)}$$

Standard error of the sample difference in means

- For the t-distribution the 95% confidence interval is (0.24, 8.6)
- We compute the value *t* and, using this CI, we decide whether to reject the null hypothesis

Chi squared test

Suppose we have two groups of individuals and we want to compare proportions

Example:

- Group 1: individuals have received a treatment
- Group 2: individuals have received a placebo
- After a period of time the individuals will either have responded to the treatment/placebo or not
- Research Question: Does the proportion of individuals that respond is the same in Group 1 and Group 2? tuition?

Chi squared test assumptions:

- Random sample
- Independent observations
- Expected counts need to be ≥ 5

Chi squared test

We use the contingency table to compute the test statistic

	Respond	Don't Respond	
Group 1	20	40	60
Group 2	35	35	70
	55	75	130

$$\chi^2 = \sum_{k=1}^{4} \frac{(observed_k - expected_k)^2}{expected_k}$$

Running the test:

- Compute the test statstic (for the example data we get $\chi^2 = 3.67$
- Comuting the degree of freedom: $\nu = (n_{rows} - 1) \cdot (n_{cols} - 1)$
- Compare the test statistic with the χ^2 of ν degrees of freedom (in the example $\nu = 1$)

In the given example we would fail to reject the null hypothesis

Parametric tests

- So far we have used probability distributions, and assumed that if the sample size is large enough, then the data will match some underlying distribution, e.g. the t-distribution, the chi-squared distribution, etc.
- If possible we want to use parametric tests as they are often the most powerful tests for a given data set
- But sometimes the test assumptions will not be satisfied, particularly the assumption of Normality....

Alternative to parametric tests

Transforming your data:

A number of different transformations are possible:

- Take the square root of each of the data points
- Take the square of each of the data points
- Take the logarithm of each of the data points

Nonparametric tests:

- Instead of comparing means we compare medians
- They use ranks, i.e order of the data (loos information about data spread)
- They will have less power
- They are not completely assumption-free

Mann-Whitney U test

The Mann-Whitney U test is the non-parametric alternative to the two sample t-test:

- H_0 : median of group 1 = median of group 2
- H_1 : median of group 1 $\neq$ median of group 2

Procedure

1. Pool the two groups, and rank all the data points
2. In each of the two groups, sum the ranks
3. Compute the test statistic U from these sums and the sample sizes in the groups

Nonparametric tests

Parametric Test	Non-Parametric Version
t test: Two sample	Mann-Whitney U test
	Wilcoxon rank sum test
t test: Paired	Sign test
	Wilcoxon signed rank test
ANOVA	Kruskal-Wallis test
Pearson correlation	Spearman correlation

One-way ANOVA

Analysis of variance (ANOVA) is a method for testing the hypothesis that there is no difference between three or more population means (let us say K means)

Basic ideas:

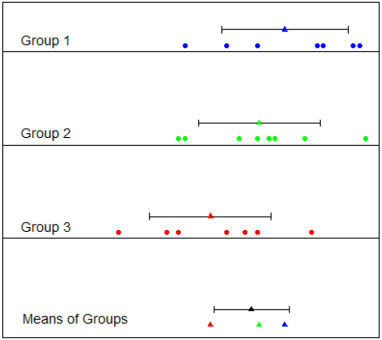
- Calculate the mean of the observations within each group
- Compare the variance of these means to the average variance within each group
- As the means become more different, the variance among the means increases

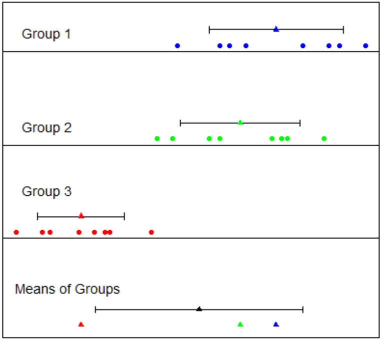
Assumptions:

- Each of the populations is Normally distributed with the same variance (homogeneity of variance)
- The observations are sampled independently, the groups under consideration are independent

One-way ANOVA

Why Look at Variance When Interested in Means?





To distinguish between the groups, the variability between the groups must be greater than the variability within the groups

One-way ANOVA

Procedure:

- Compute the within-groups Variance:

$$s_w^2 = \frac{1}{K}\sum_{i=1}^{K} s_i^2$$

  This is true because the population variances of the three groups is the same, and for K groups of the same size

- Compute the between-groups variance:

$$s_b^2 = n\sum_{i=1}^{K}\frac{(\bar{x}_i - \bar{x})^2}{(K-1)}$$

  This is true because:

  - The three means are observations from the same sampling distribution of the mean
  - The distribution of the mean is Normal with variance σ^2/n ($n =$ number of observations in each group)
  - The observed variance of the treatment means is an estimate of σ^2/n

One-way ANOVA

Procedure:

- Compute the F-statistics:

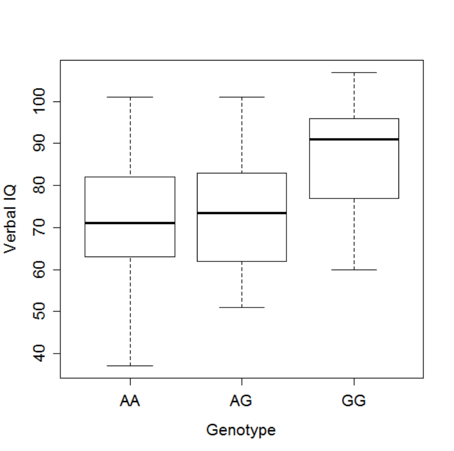
$$F = \frac{s_b^2}{s_w^2}$$

  with K-1 and K(n-1) degrees of freedom

- Another way of thinking about this ratio:

$$F = \frac{\text{Variability due to effect and chance}}{\text{Variability due chance}}$$

One-way ANOVA: Example



- 54 observations
- 18 AA observations
  mean IQ for AA = 71.6
- 18 AG observations
  mean IQ for AG = 72.7
- 18 GG observations
  mean IQ for GG = 87.1

One-way ANOVA: output table

For the genotype, verbal IQ data:

Source	DF	SS	MS	F	P
Genotype	2	2691	1346	6.25	0.004
Error	51	10979	215		
Total	53	13671			

Between-Groups
Between treatments

P-Value

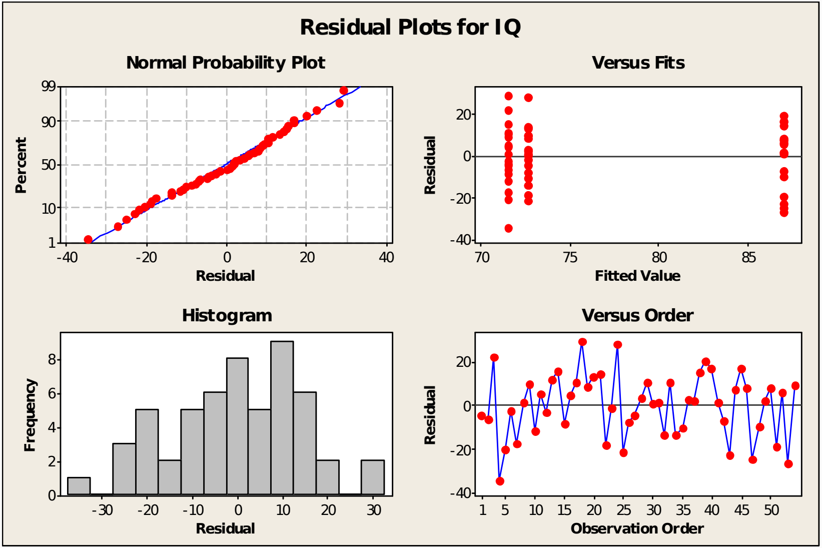
F Statistic

Within-Groups
Residual Variation

I-1 =3-1 =2, since
3 genotype groups,
AA, AG, GG

I(n-1)=3(18-1)
=51

One-way ANOVA: assumption checking



What to do with a Significant ANOVA Result

- If the ANOVA is significant and the null hypothesis is rejected, the only valid inference that can be made is that at least one population mean is different from at least one other population mean
- The ANOVA does not reveal which population means differ from which others
- To test which means are different we need to correct for multiple comparison

Thanks!