To analyse the data, we take the difference between the observations within each pair, and treat the differences as a single sample.

To estimate the population mean difference.

We test: $H_0: \mu = 0$

A $(1-\alpha)100\%$ confidence interval for μ:

$$\bar{x} \pm t_{\alpha/2} \cdot SE(\bar{x})$$

$\bar{x}$: sample mean difference

$SE(\bar{x}) = \frac{s}{\sqrt{n}}$

s: sample standard deviation of the differences

n: number of differences

$$t = \frac{\bar{x} - 0}{SE(\bar{x})} \rightarrow \text{find p-value and draw a conclusion}$$

Chi-Square Tests (for categorical data)

Chi-square tests for one-way tables

In a one-way table, observations are classified according to a single categorical variable.

example: famous genetics experiment in 1905 (investigation of inheritance in sweet peas)

One pure line of peas that had purple flowers and long pollen grains was crossed with another pure line that had red flowers and round pollen grains.
This first generation was then self-crossed

Results for 381 plants: phenotypes

purple/long	purple/round	red/long	red/round
284	21	21	55

If these two genes are inherited independently, a 9:3:3:1 ratio would be expected

H_0: True ratio of phenotypes is 9:3:3:1

H_a: True ratio of phenotype is not 9:3:3:1

→ in other words:

		Expected counts
P(Purple and long)	$= \frac{9}{16}$	$\cdot 381$	214,3
P(purple and round)	$= \frac{3}{16}$	$\cdot 381$	71,4
P(red and long)	$= \frac{3}{16}$	$\cdot 381$	71,4
P(red and round)	$= \frac{1}{16}$	$\cdot 381$	23,8

Are the observed counts significantly different from those expected under the null hypothesis?

	purple/long	purple/round	red/long	red/round
observed:	284	21	21	55
expected:	214,3	71,4	71,4	23,8

↓ differences seem quite large

Chi-Square test statistic: $X^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \Rightarrow X^2 = 134,7$

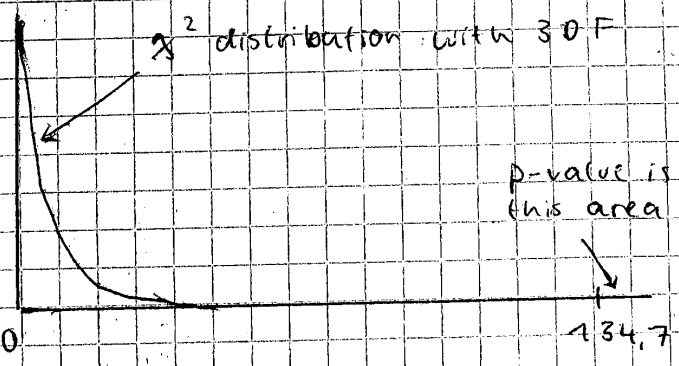
$DF = 4 - 1 = 3$

If the null hypothesis is true the chi-square test statistic is going to have approximately a <u>chi-square distribution</u>.

degrees of freedom = # of cells − 1 (i.e number of categories)

If the observed counts are close to the expected counts then x^2 is going to be small.
→ Larger values of x^2 give us evidence against "H_0"

When we get our p-value, we take the observed value of our test statistic x^2. The p-value is going to be the probability, under the null hypothesis of getting that value or sth even larger. (The area to the right of x^2 under a chi-square distribution)



There is extremely strong evidence that these phenotypes do not occur in a 9:3:3:1 ratio (aka not genes are not independent)

→ This study was significant in the discovery of genetic linkage, where genes close to each other on the same chromosome are more likely to be inherited together.

Chi-square Tests of Independence (two-way tables)

Example: A study of 11 160 alcohol drinkers on university campuses revealed:

	binge drinking: never	occasional	frequent	total
trouble with police	71	154	398	623
no trouble with police	4992	2808	2737	10537
total	5063	2962	3135	11160

H_a: They are not independent

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = 469.6$$

* Expected count $= \dfrac{\text{row total} \times \text{column total}}{\text{overall total}}$

$DF = (\#\text{ rows} - 1) \times (\#\text{ columns} - 1) = (2-1)\cdot(3-1) = 2$

$\longrightarrow$ p-value $< 2.2 \cdot 10^{-16}$

$\longrightarrow$ There is extremely strong evidence that binge drinking and trouble with police are not independent.

ANOVA

Analysisis of variance, or in short ANOVA, is a technique from statistical inference that allows us to deal with several populations.

To see why we need ANOVA, we will consider an example. Suppose we are trying to determine if the mean weights of green, red, blue and orange M&M candies are different from each other. We will state the mean weights for each of these populations μ_1, μ_2, μ_3 and μ_4 respectively. We may use the appropriate hypothesis test several times:

- $H_0: \mu_1 = \mu_2$
- $H_0: \mu_2 = \mu_3$
- $H_0: \mu_3 = \mu_4$
- $H_0: \mu_4 = \mu_1$
- $H_0: \mu_1 = \mu_3$
- $H_0: \mu_2 = \mu_4$

There are many problems with this kind of analysis. We will have six p-values. Our confidence in the overall process will be $0.95 \cdot 0.95 \cdot 0.95 \cdot 0.95 \approx 0.74$. Thus the probability of a type I error has increased. And we cannot compare these four parameters as a whole by comparing them two at a time.

One-way analysis

One-way ANOVA is a statistical method that tests:

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$ (It is a test on means!)

by comparing the variability between groups to the variability within groups

The assumptions are the same as those of the pooled-variance two-sample t-test (population variances are equal)

Let:

- k represent the number of groups
- X_{ij} represent the j^{th} observation in the i^{th} group
- $\bar{X}_i$ represent the mean of the ith group (k of them)
- $\bar{X}$ represent the overall mean
- s_i represent the standard deviation of the ith group
- n_i represent the number of observations in the ith group
- $n = n_1 + n_2 + \dots + n_k$ → total number of observations

The total sum of squares:

$$SS(Total) = \sum_{\text{all obs}} (X_{ij} - \bar{X})^2 \quad \left(\rightarrow s^2 = \frac{SS(Total)}{n-1}\right)$$

The total sum of squares is partitioned into two components:

Sum of square treatment (between groups) $SST = \sum_{\text{groups}} n_i (\bar{X}_i - \bar{X})^2$

sum of squares error (within groups) $SSE = \sum (n_i - 1) s_i^2$ — if there is a lot of variability within the groups SSE tends to be large

$$\Rightarrow SS(Total) = SST + SSE$$

$$DF(Total) = DFT + DFE$$

source	DF	Sum of squares	Mean square	F
treatments	$k-1$	SST	$SST/(k-1)$	MST/MSE
error	$n-k$	SSE	$SSE/(n-k)$	–
total	$n-1$	SS(Total)	–	–

$MSE = s_p^2$
(mean square error)

The test statistic is $F = \frac{MST}{MSE}$

If H_0 is true, this test statistic will have an F distribution and MST and MSE have to be the same quantity. (F close to 1 where its' mean is)

numerator df: $k-1$
denominator df: $n-k$

If H_0 is false MST will tend to be bigger than MSE and the test statistic will tend to be large.

<u>example</u>: can self-control be restored during intoxication?

An experiment randomly assigned 11 males to each of four groups

- Group A received 0,62 $\frac{mg}{kg}$ of alcohol
- Group AC received alcohol plus caffeine
- Group AR received alcohol and a monetary reward for performance
- Group P received a placebo

Scores on a word stem completion task involving "controlled memory processes" were recorded.

Higher scores were indicative of greater self-control

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: The means are not all equal

Group A	Group B	Group AR	Group P
$\bar{x}_1 = 0,08$	$\bar{x}_2 = 0,266$	$\bar{x}_3 = 0,441$	$\bar{x}_4 = 0,4$
$s_1 = 0,194$	$s_2 = 0,170$	$s_3 = 0,182$	$s_4 = 0,167$

<u>resulting ANOVA table is:</u>

source	DF	SS	MS	F	p-value
treatments	3	0,972	0,324	10,18	0,00004
error	40	1,272	0,0318		
total	43	2,243	–		

There is very strong evidence (p-value = 0,00004) that the groups do not all have the same population mean. (the different treatments do not all have the same effect)

but now you could investigate further and calculate confidence intervals for all the pairwise differences ($\mu_1 - \mu_2$, $\mu_3 - \mu_4$, etc.)

Two-way ANOVA

We use a two-way ANOVA when we want to know how two independent variables, in combination, affect a dependent variable.
(quantative)

Both of your independent variables should be categorical.

Again, we use the F-test which compares the variance in each group to the overall variance in the dependent variable.

If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-value, and therefore a higher likelihood that the difference observed is real and not due to chance.

assumption: the dependent variable should be normally distributed

example: customer satisfaction at friendly bank

Factor 1: time of day	Factor 2: Type Contact: automated	Factor 2: Type Contact: bank repr.	Row means
Morning	6,5,8,4 $\bar{x} = 5,75$	8,7,9,9 $\bar{x} = 8,25$	$\bar{x} = 7$
afternoon	3,5,6,5 $\bar{x} = 4,75$	9,10,6,8 $\bar{x} = 8,25$	$\bar{x} = 6,5$
evening	5,5,7,5 $\bar{x} = 5,5$	9,10,10,9 $\bar{x} = 9,5$	$\bar{x} = 7,5$
column means	$\bar{x} = 5,33$	$\bar{x} = 8,75$	total $\bar{x} = 7$

* we could separately test the differences of these means with one-way ANOVA *

* same with these

→ but instead with two-way ANOVA we can perform the test at the same time

→ two-way ANOVA studies: average customer satisfaction regarding:

1. the interaction of contact type and time of day
2. type of contact
3. time of day

Variables (factors)	Levels
1. time of day	morning, afternoon, evening
2. type of contact	automated, bank rep.

n = number of items in cell
p = number of rows
i = number of columns

Analysis of variance response

source	DF	SS	MS
contact type	1	66,67	66,67
contact time	2	4,00	2,00
interaction	2	2,33	1,17
Error	18	29,00	1,61
Total	23	102,00	

degrees of freedom
row factor = $p-1$
column factor = $i-1$
interaction = $(p-1)(i-1)$
error = $pi(n-1)$
total = $n \cdot pi - 1$

$$MS = \frac{SS}{d.f.}$$

test statistic: $F = \frac{MS}{MS_{error}}$

critical value: $\alpha = 0,05$, $d.f._N$, $d.f._D$ (numerator, denominator)

1. H_0: There is no interaction

H_1: There is interaction

test statistic: $F = \frac{MS_{interaction}}{MS_{error}} = \frac{1,17}{1,61} = 0,73$

critical value: $\alpha = 0,05$ $d.f._N = 2$ $d.f._D = 18$

$\rightarrow F = 3,55$

→ The test statistic is so small that we don't get out far enough to get into the rejection region, so there is no interaction between contact type and contact time

2. H_0: No difference in customer satisfaction between the two types of contact

H_1: There is a difference

test statistic: $F = \frac{MS_{contact\ type}}{MS_{error}} = \frac{66,67}{1,61} = 41,41$

critical value: $d.f._N = 1$ $d.f._D = 18$ → $F = 4,41$

→ Certainly the 41,41 puts us far enough out to where we "know" we're in the rejection region so we reject H_0 and there is a difference between contact type

3. H_0: There is no difference in customer satisfaction among the times of contact

H_1: There is a difference

Test statistic: $F = \frac{MS_{time}}{MS_{error}} = \frac{2}{1,61} = 1,24$

critical value: $\alpha = 0,05$ $d.f._N = 2$ $d.f._0 = 18$

$\rightarrow F_0 = 3,55$

$\rightarrow$ 1,24 is not in the critical region, therefore we accept ~~the~~ H_0, meaning times of contact doesn't affect customer satisfaction

Correlation & Regression

Objectives of regression:

- find out whether there is a relationship between two variables
- forecast observations

Regression analysis explores the relationship between a quantitative response variable and one or more explanatory variables

If there is only one explanatory variable, we call it simple regression and if there are more than one explanatory variable, we call it multiple regression.

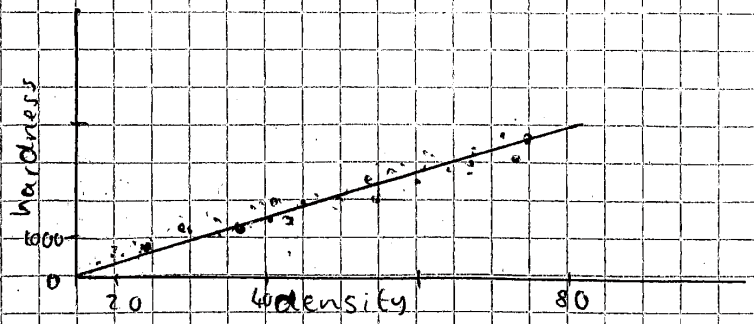
Simple linear regression:

Example: Can the density of Australian timber be used to predict its Janka hardness?

The first 4 observations (of 36):

x (explanatory variable (independent))	density	24,7	39,4	53,4	24,8
y (response variable (dependent))	hardness	484	1210	1880	427

We call the variable we are predicting y and the variable using to predict it x variable



We attempt to fit a line through those points and then we use that line for prediction

We will assume a linear relationship between Y and X:

$$E(Y|X) = \beta_0 + \beta_1 x + \epsilon$$

β_0: Y intercept; β_1: slope; ϵ: irreducible error

→ unknown

The observed values of Y vary along the line. We account for that variability with a random error component ϵ.

We usually use the method of least squares to estimate β_0 and β_1

The Least Squares Regression Line

We use sample data to find the estimated regression line:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\text{Residual} = \text{Observed} - \text{Predicted}$$

$$e_i = Y_i - \hat{Y}_i$$

The residuals are the vertical distances between the points and the line points. Points below the line have negative residuals and points above the line have positive residuals

We minimize the sum of the absolute value of the residuals

$\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimize:

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$

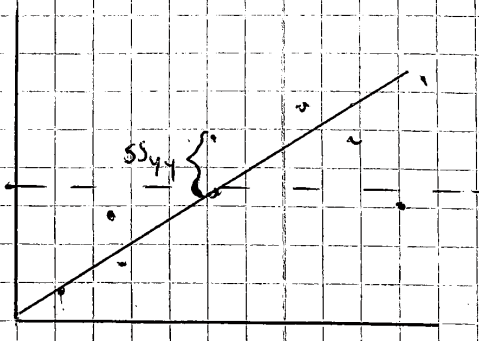
$$= \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

$$SS_{xx} = \sum (X_i - \bar{X})^2 \qquad s_x^2 = \frac{SS_{xx}}{n-1}$$

$$SS_{yy} = \sum (Y_i - \bar{y})^2 \qquad s_y^2 = \frac{SS_{yy}}{n-1}$$

$$SP_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{y})$$

$$cov(x, y) = \frac{SP_{xy}}{n-1}$$



$$\rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{SP_{xy}}{SS_{xx}} = \frac{cov(x,y)}{var(x)}$$

Now if we took all distances and added them up that would add to zero.

Exercise

x	-1	1	2	4	6	7
y	-1	2	3	3	5	8

$n = 6$

Does there exist a linear relationship?

x	y	x^2	y^2	xy
-1	-1	1	1	1
1	2	1	4	2
2	3	4	9	6
4	3	16	9	12
6	5	36	25	30
7	8	49	64	56
$\sum$	19	20	107	112	107

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_1 = \frac{(\sum x)(\sum y) - n \sum xy}{(\sum x)^2 - n \sum x^2} = \frac{19 \cdot 20 - 6(107)}{(19)^2 - 6(107)} = \frac{-262}{-281} \approx 0,93$$

$$\hat{\beta}_0 = \frac{(\sum x)(\sum xy) - (\sum y)(\sum x^2)}{(\sum x)^2 - n \sum x^2} = \frac{19 \cdot 107 - 20 \cdot 107}{19^2 - 6 \cdot 107} = \frac{-107}{-281} \approx 0,3808$$

$$\rightarrow \hat{y} \approx 0,9324 x + 0,3808$$

Correlation calculation:

$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1}} = \sqrt{\frac{107 - \frac{1}{6} 19^2}{5}} \approx 3,0605$$

$$s_y = \sqrt{\frac{112 - \frac{1}{6}(20)^2}{5}} \approx 3,0111$$

Correlation r: $r = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{(n-1) s_x s_y} \approx 0,9477$

↓

strong correlation

Multiple Regression

General parametric equation:

$$y = f(X) + \epsilon$$

($f(X)$ depends on statistical method)

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

This Regression type is more relevant in real life since it can take multiple factors (e.g. height, age to determine marital status) into account.

Bayesian Statistics

We have covered the Bayes' theorem already but as a quick review:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(A \wedge B)}{P(B)}$$

where A and B are events, $P(A|B)$ is the conditional probability that event A occurs given that event B has already occurred.

example: picking a card from a pack of traditional playing cards

There are 52 cards in the pack, 26 of them are red and 26 of them are black.

What is the probability of the card being a 4 given that we know the card is red?

A: card is a 4 B: card is red

$P(B|A) = P(\text{red}|4) = \frac{2}{2} = 1$ (2 red cards left)

$P(A) = P(4) = \frac{4}{52} = \frac{1}{13}$

$P(B) = P(\text{red}) = \frac{1}{2}$ (half the cards)

$$P(A|B) = P(4|\text{red}) = \frac{\frac{1}{2} \cdot \frac{1}{13}}{\frac{1}{2}} = \frac{1}{13}$$

Prior probability and posterior probability

example: diagnostic testing on HIV

early HIV testing in the military

- first screen with ELISA
- if positive, two more rounds of ELISA
- if either positive test with two western blot assays
- only if both positive, determine HIV infection

data:

ELISA
- sensitivity (true positive): 93%
- specificity (true negative): 99%

Western blot
- sensitivity: 99,9%
- specificity: 99,1%

prevalence: $\frac{1,48}{1000}$ (adult Americans were HIV positive by 1980)

What is the probability that a recruit who tested positive in the first ELISA has HIV?

$$P(\text{has HIV} \mid \text{ELISA} +) = ?$$

$$P(\text{HIV}) = 0,00148$$

$$P(+ \mid \text{HIV}) = 0,93 \qquad P(- \mid \text{no HIV}) = 0,99$$

prior probability (≙ probability that an observation will fall into a group before you collect your data – best rational assessment of the probability of an outcome based on the current knowledge before an experiment is performed)

→ prior to any testing what probability should be assigned to a recruit having HIV?

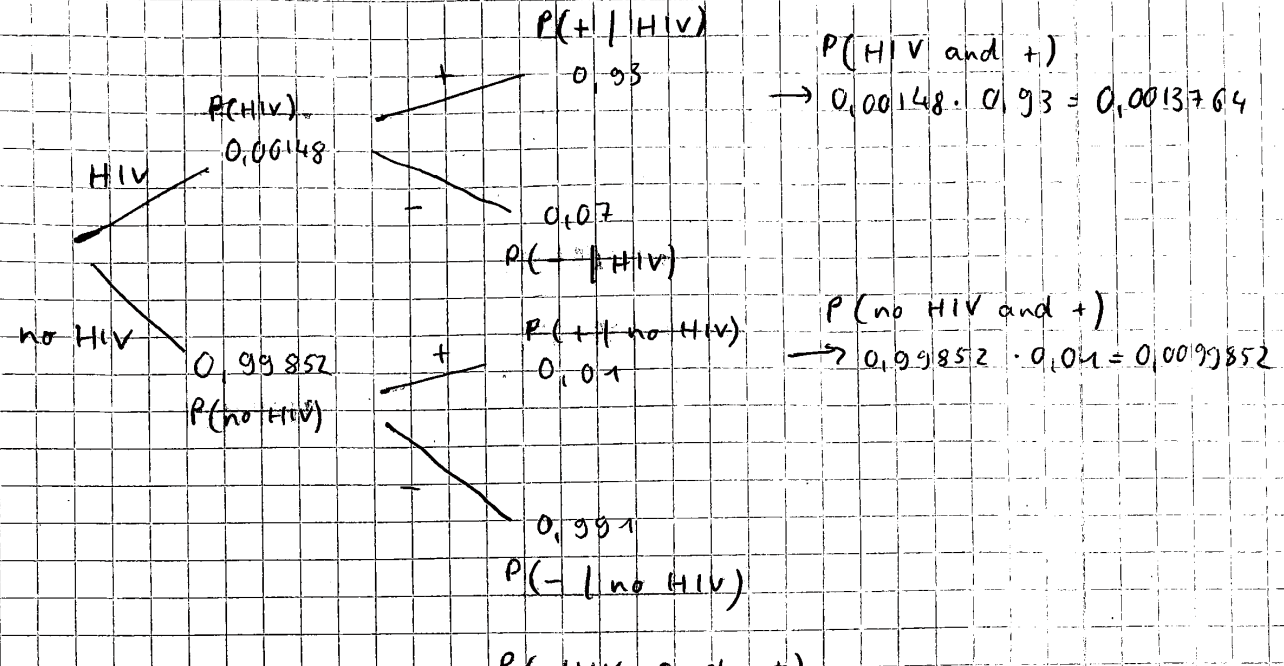
→ prevalence of the disease in the population

$$P(\text{HIV}) = 0,00148$$

posterior probability (≙ the updated probability of an event occurring after taking into consideration new information)

When a recruit goes through HIV screening there are two competing claims: recruit has HIV and recruit doesn't have HIV.

→ If the ELISA yields a positive result, what is the probability this recruit has HIV?



$P(\text{HIV and } +) \rightarrow 0{,}00148 \cdot 0{,}93 = 0{,}0013764$

$P(\text{no HIV and } +) \rightarrow 0{,}99852 \cdot 0{,}01 = 0{,}0099852$

$$\rightarrow P(\text{HIV} \mid +) = \frac{P(\text{HIV and } +)}{P(+)}$$

$$= \frac{0{,}0013764}{0{,}0013764 + 0{,}0099852} \approx \underline{\underline{0{,}12}}$$

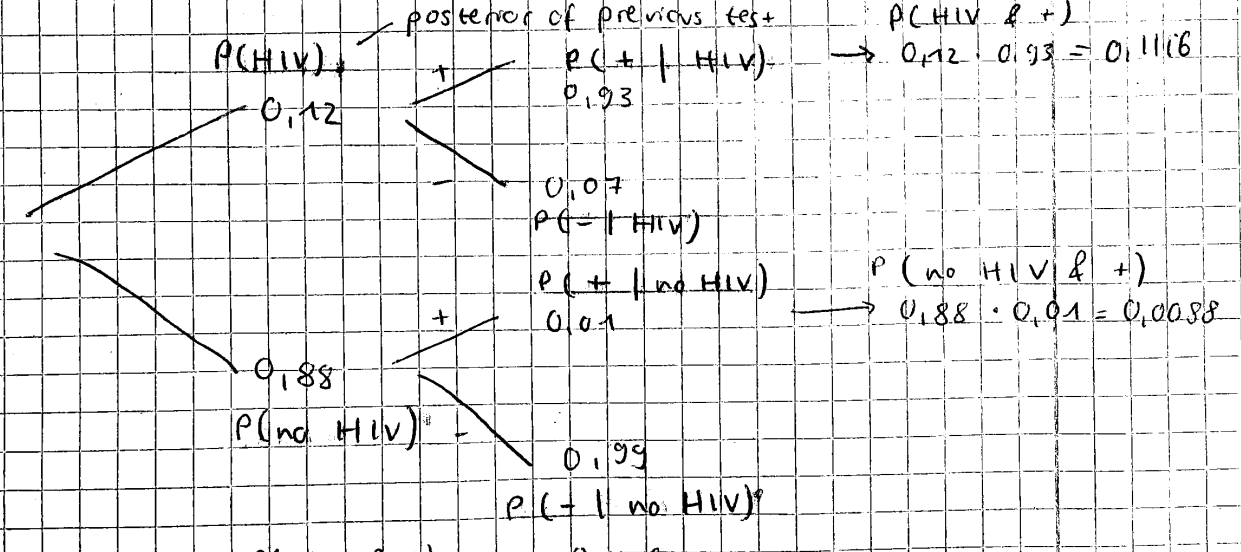
↑ posterior probability

The probability that a recruit who tested positive in the first ELISA has HIV is 0,12.

Bayes Updating

retesting: Since a positive outcome on the ELISA doesn't necessarily mean that the recruit actually has HIV, they are retested.

What is the probability of having HIV if this second ELISA also yields a positive result?



$P(\text{HIV and } +) \rightarrow 0{,}12 \cdot 0{,}93 = 0{,}1116$

$P(\text{no HIV and } +) \rightarrow 0{,}88 \cdot 0{,}01 = 0{,}0088$

$$P(\text{HIV} \mid +) = \frac{P(\text{HIV and } +)}{P(+)} = \frac{0{,}1116}{0{,}1116 + 0{,}0088} = \underline{\underline{0{,}93}}$$

→ Due to a second ELISA test we could increase the probability that a recruit who tested positive has HIV by 81 %.

! The difference between prior and posterior probabilities characterizes the information we have gotten from the experiment or measurements.

→ We learned that a posterior probability can subsequently become a prior for a new updated posterior probability as new information arises and is incorporated into the analysis.

Bayesian vs. frequentist

definitions of probability

- frequentist: relative frequency in large number of trials

$$P(E) = \lim_{n \to \infty} \frac{n_E}{n}$$ (proportion of times event occurs in n trials, when n goes to infinity)

- bayesian:

example: indifferent between winning

> $1 if event E occurs

or

> winning $1 if you draw a blue chip from a box with $1000 * p$ blue chips and $1000 * (1-p)$ white chips

This means you're equating the probability of events E, $P(E)$, to the probability of drawing a blue chip from this box, p.

$$\rightarrow P(E) = p$$

→ It is an interpretation or estimate of probability as a personal judgement or degree of belief about how likely a particular event is to occur, based on the state of knowledge and available evidence.

definitions of statistics:

- frequentist:

Frequentist statistics tests whether an event occurs or not. It calculates the probability of an event in the long run of the experiment. With a frequentist approach the result of an experiment is dependent on the number of times the experiment is repeated.

→ flaws in frequentist statistics.

- p-values measured against a sample (fixed size) statistic changes with change in the stopping intention and sample size. So if two people work on the same data and have different stopping intention, they may get two different p-values for the same data, which is undesirable.

  e.g. Person A chooses to stop tossing a coin after 100 tosses while B stops at 1000. For different sample sizes, we get different t-scores, thus different p-values

- Confidence intervals are not probability distributions therefore they do not provide the most probable value for a parameter.

  e.g. "We are 95% confident that 60% to 64% of Americans think the federal government does not do enough for middle class people" (poll on 1500 adults)

  means 95% of random samples of 1500 adults will produce confidence intervals that contain the true proportion of Americans who think the federal government does not do enough for middle class people.

  but not: The true population proportion is in this interval 95% of the time

  → in other words for a single given confidence interval of a sample the true parameter is either in it or not.

- Bayesian statistics

A mathematical procedure that applies probabilities to statistical problems, where probability expresses a degree of belief in an event.
The degree of belief may be based on prior knowledge about the event, for instance results of previous experiments or on personal beliefs about the event.

Bayesian statistical methods use Bayes' theorem to compute and revise probabilities after acquiring new data.

Bayesian inference

As we already know, statistical inference is the process of deducing properties about a population or probability distribution from data.

Bayesian inference specifically, uses the Bayes' theorem.

<u>example:</u> Probability of ice-cream sale with regards to the type of wheather.

Bayes' theorem gives us the tools to use prior knowledge about the likelihood of selling ice cream on any other type of day (rainy, windy, snowy)

Let A represent the event that we sell ice cream and B the event of the wheather.

Then we might ask what is the probability of selling ice-cream on any given day given the type of wheather?

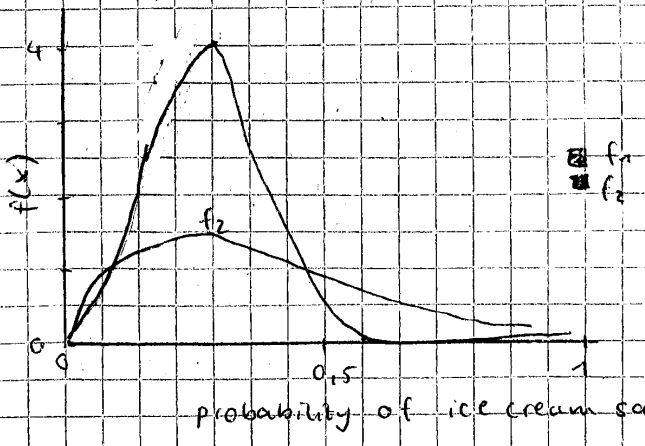
$$\rightarrow P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$ marginal

$P(A)$ is our prior, so in our case the probability of selling ice cream regardless of the type of wheather outside. We might know this by looking at data that said 30 people out of a potential 100 actually bought ice-cream at some shop somewhere.
So $P(A) = \frac{30}{100} = 0,3$, prior to knowing anything about the wheather.

(note: the prior belief can also be made up, though the resulting calculation will be affected by this choice)

If $P(A) = 0,3$ was just our best guess using a distribution of our prior belief is more appropriate, which is known as the <u>prior distribution</u>.

eg.



The peak of both curves occur around 0,3, which as we said, is our best guess. The fact that $f(x)$ is non-zero of other values of x shows that we're not completely certain that 0,3 is the true value. F_1 shows that it's likely to be between 0 and 0,5 whereas f_2 shows that's it's likely to be between 0 and 1, meaning the prior prob. expressed by f_2 is less certain.

Model form of Bayes' Theorem

Instead of event A, we usually use θ called theta. Theta is of our interest, it represents the set of parameters.
So if we're trying to estimate the parameter values of a Guassian distribution then θ represents the mean μ and the standard deviation σ

$\rightarrow \theta = \{\mu, \sigma\}$ (It usually represents a proposition (e.g. "a coin lands on heds 50% of the time"))

Instead of event B, we'll see <u>data</u> or $y = \{y^1, y^2, \dots y^n\}$
These represent the data, ie the set of observations of measurments we have, we can also see it as our evidence.

$$\rightarrow P(\theta \mid data) = \frac{P(data \mid \theta) \cdot P(\theta)}{P(data)}$$

$P(\theta \mid data)$ is our <u>posterior distribution</u>, representing our belief about the parameter values after we have calculated everything on the right hand side, taking the observed data into account.

$P(data \mid \theta)$ is our <u>likelihood distribution/function</u>, which can be interpreted as the prob. of the evidence give θ is true.
→ It measures the goodness of fit of a statistical model to a sample of data for given values of parameters
the "likelihood" is equal to the probability that a particular outcome is observed when the true value of parameter is θ

→ The Bayes' theorem implies that the posterior probability is proportional to the likelihood times the prior probability.

Models are the mathematical formulation of the observed events. Parameters are the factors in the models affecting the "observed" data.

<u>example:</u> fairness of a coin (θ).

Our question could be:

Given an outcome data what is the probability of the coin being fair ($\theta = 0,5$)

Thus $P(\theta)$ is the prior, ie the strength of our belief in the fairness of the coin <u>before</u> the toss.

$P(data \mid \theta)$ is the likelihood of observing our result given our distribution for θ.

$P(data)$ is the probability of data determined by summing across all possible values of θ, weighted by how strongly we believe in those particular values of θ

$P(\theta \mid data)$ is the posterior belief of our parameters after observing the evidence, i.e the number of heads.

Bernoulli likelihood function

to recap: our likelihood function is the probability of observing a particular number of heads in a particular number of flips for a given fairness of coin.

1 represents heads, 0 represents tails

$P(y=1 \mid \theta) = \theta^y$ (if coin is fair, prob of heads is 0,5)

$P(y=0 \mid \theta) = (1-\theta)^{1-y}$

$\longrightarrow P(y \mid \theta) = \theta^y \cdot (1-\theta)^{1-y}$ $\quad y = \{0,1\}$ $\theta = (0,1)$

when we want to see a series of flips or heads:

$$P(y_1, y_2, \dots y_n \mid \theta) = \prod_i P(y_i \mid \theta)$$

$$= \prod_i^n \theta^{y_i} \cdot (1-\theta)^{1-y_i}$$

Furthermore, if we are interested in the probability of number of heads z turning up in N number of flips:

$$P(z, N \mid \theta) = \theta^z \cdot (1-\theta)^{N-z}$$

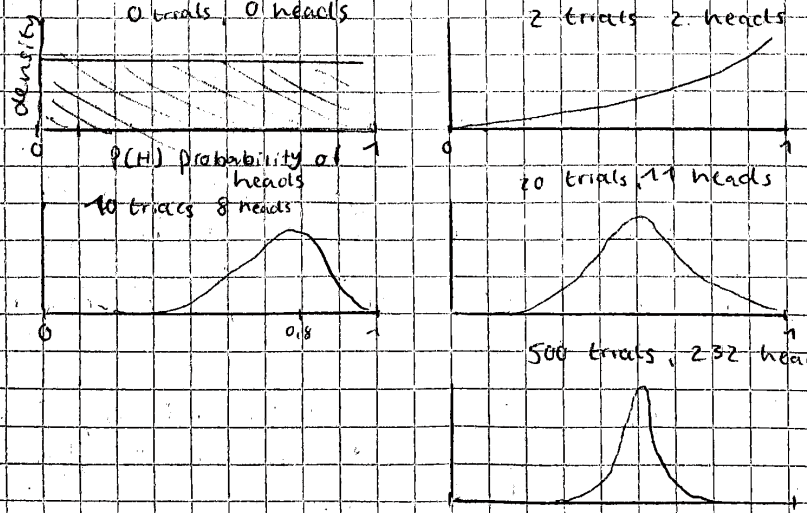
Prior belief distribution

The mathematical function to represent prior beliefs is known as beta distribution and has the following form:

$$x^{\alpha-1} \cdot (1-x)^{\beta-1} / B(\alpha, \beta) \qquad \mu = \frac{\alpha}{\alpha+\beta} \qquad \sigma = \sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}}$$

α and β are called the shape decoding parameters for the density function.

Here α is equivalent to the number of heads in the trials and β corresponds to the number of tails. here a few visualizations of beta distributions for different α and β



→ As more tosses are done and heads continue to come in larger proportion the peak narrows which increases our confidence in the fairness of the coin.

Posterior belief distribution

So let's put it all together

$$P(\theta \mid z, N) = \frac{P(z, N \mid \theta)\, P(\theta)}{P(z, N)}$$

$$= \frac{\theta^z (1-\theta)^{N-z} \cdot \theta^{\alpha-1} (1-\theta)^{\beta-1}}{B(\alpha, \beta) \cdot P(z, N)}$$

$$= \frac{\theta^{z+\alpha-1} (1-\theta)^{N-z+\beta-1}}{B(z+\alpha, N-z+\beta)}$$

$$\Rightarrow P(\theta \mid z+\alpha, N-z+\beta)$$

So this means by knowing the mean and standard deviation (as we can derive α and β from them) of our belief about ~~of~~ the parameter θ and by observing the number of heads in N flips, we can update our belief about θ.

For example you could suppose the coin is biased and has a mean of $0,6$ with $\sigma\ 0,1$ then $\alpha = 13,8$ and $\beta = 9,2$

which means our prior distribution will be biased on the right. And suppose you then observed 80 heads ($z = 80$) in 100 tips ($N = 100$)

So our posterior probability would look like this

$$P(\theta \mid z+\alpha, N-z+\beta) = P(\theta \mid 93,8, 29,2)$$

Bayes' factor

The Bayes' factor is the equivalent of the p-value in ~~b~~ the bayesian framework.

It is defined as the ratio of the likelihood of one particular hypothesis to the likelihood of another hypothesis.

Typically, it is used to find the ratio of the likelihood of an alternative hypothesis to a null hypothesis.

For example, if the Bayes Factor is 5 then it means the alternative hypothesis is 5 times as likely as the null hypothesis given the data.

Similar to p-values we can use thresholds to decide when we should reject a null hypothesis (e.g. 10)

A Bayes' factor of 1 means that there is no evidence.

Some statisticians believe that the Bayes factor offers an advantage over p-values because it allows you to quantify the evidence for and against two competing hypotheses.

Credibility Intervals (analogous to confidence interval in frequentist statistics)

They allow us to describe the unknown true parameter not as a fixed value but with a probability distribution.

Therefore we can make probabilistic statements about the parameter falling within that range

As the Bayesian inference returns a distribution of possible outcomes, the credibility interval is the range containing a particular percentage of probable values.

For instance, the 95% credible interval is simply the central portion of the posterior distribution that contains 95% of the values.

It allows us to say something like "given the observed data, the outcome has 95% probability of falling within this range"