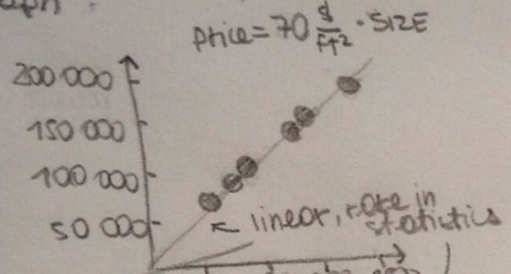
Intro to Statistics

DATA → STATISTICS → DECISIONS

Scatter plot

- Each data item is a dot in a graph.

SIZE	PRICE
1400	98000
2400	168000
1800	126000
1900	133000
1300	91000
1100	77000



A linear data set is really easy to predict

often we find deviations because of the many factors

Even without a fixed dollar price per squarefoot, the relationship can be linear

price = 30 $/ft² · SIZE + 2000 $

SIZE	PRICE
1700	53000
2100	65000
1900	59000
1300	41000
1600	50000
2200	68000

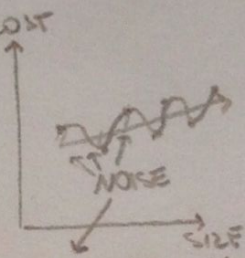
some prices may be outliers (more later...) and not fit in the relationship

Scatter plots aren't great when there's what's called "noise". That is, the data deviates from the expectation, in some random, noisy way

Barcharts address the issue of noise in data by pooling data points into a single cumulative bar

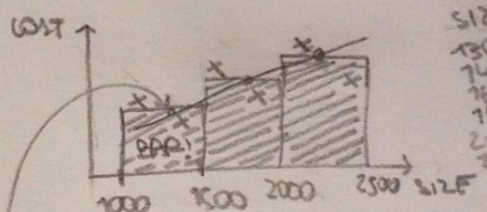
Bar charts

SIZE	COST
1300	88 000
1400	72 000
1600	94 000
1900	86 000
2100	112000
2300	98 000



There might be factors that really affect the price beyond the size.
If those factors aren't included, to a statistician that's called "random noise"

In a bar chart, we take our raw data and pull it together



SIZE	COST
1300	88 000
1400	
1600	
1900	
2100	
2300	

all the data that falls into that interval should be summarized by a single value...

such a value would lie halfway in between these two datapoints

A bar chart is a much finer representation of the data. There's a much better way to really understand the dependence of cost to data.

While the bar doesn't give you the linear relationship, it really gives you a sense, as you go up in house sizes, the cost increases

...so a bar chart really helps you to pool together groups of data into a single bar, and understand global trends

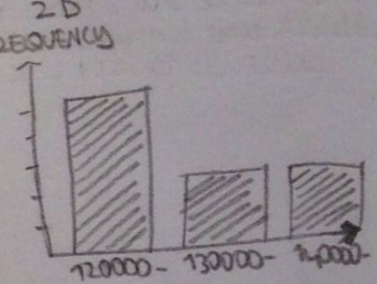
Useful with a lot of data points. A scatter plot with a lot of data points looks tells you very little

Histograms - special case of a bar chart

1D

INCOME
132,754
137,792
122,177
147,121
143,000
126,010
128,200
124,322
128,132

2D



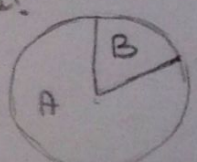
It shows the count for the salaries that fall into each category

Pie charts

In statistics you use pie charts to visualize relative data!

PARTY A: 724,000 votes
PARTY B: 181,000 votes

$\frac{181000}{905000} = 0.2$ $\frac{724000}{905000} = 0.8$



It is invariant to the total number of votes but it shows that A got many more votes than B

```
From plotting import *
data = [...]
histplot(data)
scatterplot(x, y)
```

Admissions Case Study

... From UC Berkeley to show whether admissions were gender biased

Simpson's Paradox

(Data is made up to illustrate the effect)

	MALE APPLIED	ADMITTED	RATE	FEMALE APPLIED	ADMITTED	RATE
MAJOR A	900	450	50%	100	80	80%
MAJOR B	100	10	10%	900	180	20%
BOTH	1000	460	46%	1000	260	26%

From that looks like Female students are being Favored

When you look at both majors together, males have a higher admissions rate than Females

Statistics is ambiguous. In choosing how to graph your data you can majorly impact what people believe to be the case

Statistics is deep and often manipulated

=

Probability

... is just the opposite of statistics

DATA —statistics→ CAUSES

DATA ←probability— CAUSES

Probability gives us a language, to describe the relationship between data and underlying causes

Flipping Coins

$P(HEADS) = 0.5$

$P(TAILS) = 0.5$

$P(HEADS) + P(TAILS) = 1$

$P(H,H) = 0.25$

TRUTH TABLE ← Draw out every possible outcome

FLIP-1	FLIP-2	
H	H	0.25
H	T	0.25
T	H	0.25
T	T	0.25

Since H and T are equally likely, each of these outcomes is equally likely

Since we know all add to 1, each outcome has a 0.25 chance

Another way to see this

$P(H,H) = P(H) * P(H)$ ← Independence: The second outcome doesn't depend on the first outcome

0.25 = 0.5 * 0.5

Say a loaded coin is flipped twice

$P(H) = 0.6$ $P(T) = 0.4$

FLIP-1	FLIP-2	
H	H	0.36
H	T	0.24
T	H	0.24
T	T	0.16

← Using the product rule above

=

$P(\text{Exactly one } H) = P(H,T) + P(T,H)$

All possibilities from truth table with exactly one H

Conditional Probability

We're going to study the more interesting case where the outcome of the first does impact the outcome of the second

$P(CANCER) = 0.1$

$P(\neg CANCER) = 0.9$

blood test for cancer

$P(POSITIVE \mid CANCER) = 0.1$ ← Test makes mistakes 10% of the cases

$P(NEGATIVE \mid CANCER) = 0.9$

Means the outcome of the test depends on whether the person has cancer or not. That is called a conditional probability

| says what the probability of the stuff on the left, given that we assume the stuff on the right is actually the case

$P(POSITIVE \mid \neg CANCER) = 0.2$

$P(NEGATIVE \mid \neg CANCER) = 0.8$

CANCER	TEST	P( )
Y	P	0.09
Y	N	0.01
N	P	
N	N	

$P(CANCER) * P(POSITIVE \mid CANCER)$

That notation is called Total Probability!

We have a bag with two coins

$P_1(H) = 0.5$ $P_2(H) = 0.9$

We randomly pick a coin and flip it probabilities comes out head:

PICK	FLIP		
1	H	0.25 ←	$P(H) = 0.7$
1	T	0.25	
2	H	0.45 ←	
2	T	0.05	

... In flipping it twice, I want first heads and then tail:

PICK	FLIP-1	FLIP-2		
1	H	T	0.125	$P(H,T) = 0.17$
1	T	H		
2	H	T	0.045	
2	T	H		

Bayes rule! ← Holy grail of probabilistic inference

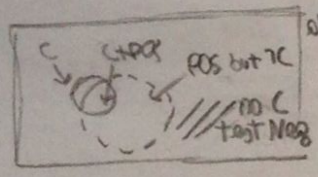
REV THOMAS BAYES who used this principle to infer the existence of God

..But he created a new family of methods that has vastly influenced AI and statistics

$P(C) = 0.01$

TEST: 90%. it is positive if you have C
90%. it is negative if you don't have C

Says test is positive, what is P(C)



Bayes rule

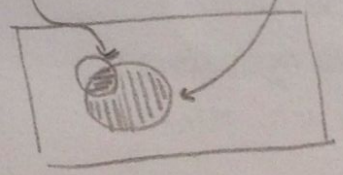
prior probability · test evidence → posterior probability

semantically, it incorporate some evidence from the test into your prior probability to arrive at a posterior probability

prior $P(C) = 0.01$ $P(Pos|C) = 0.9$
$P(Neg|\neg C) = 0.9$ $P(Pos|\neg C) = 0.1$

$P(C,Pos) = P(C) \cdot P(Pos|C) \rightarrow 0.009$
$P(\neg C,Pos) = P(\neg C) \cdot P(Pos|\neg C) \rightarrow 0.099$

posterior often called "joint"

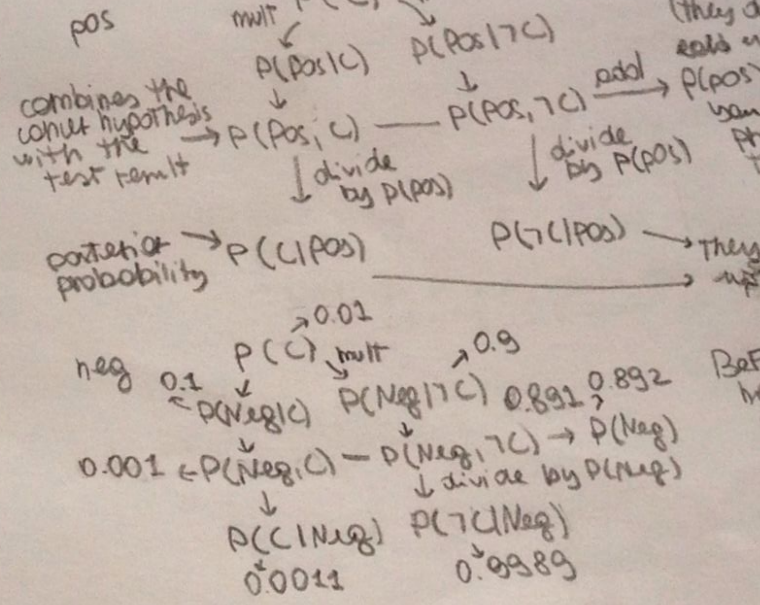


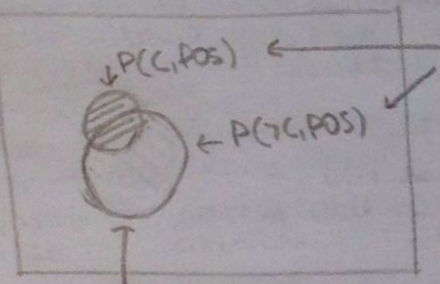
Normalizer $P(Pos) = P(C,Pos) + P(\neg C,Pos) = 0.108$

posterior $P(C|Pos) = P(Pos|C) / P(Pos) = 0.0833$ ↘ 1
$P(\neg C|Pos) = P(Pos|\neg C) / P(Pos) = 0.9267$ ↗

=

P(C) prior
P(Pos|C) sensitivity
P(Neg|¬C) specificity





... and then we normalized P(C,Pos) and P(¬C,Pos) by the total area to get the relative area assigned to P(C,Pos) versus P(¬C,Pos)

In Bayes rule there will be more than two causes of C, ¬C, and we can apply exactly the same math but we have to keep track of more values.

There might also be more than two test outcomes, and this means that our measurement probability will be more elaborate. More information is needed, but the math remains exactly the same

Before the test we had 0.01 probability of having cancer, After a negative test we have 0.0011 chance of having cancer

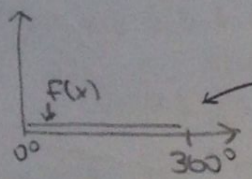
Continuous probability distributions

In continuous distribution, every outcome has probability 0!

$$P(a < x \leq b) = \frac{|b-a|}{360}$$

↑ probability that a spinning bottle stops at a given angle.

Density



outcome of turning the bottle lies between 0° and 360°

We want to assign 0 to anything outside that range

There's a function that makes every function in the range equally likely.

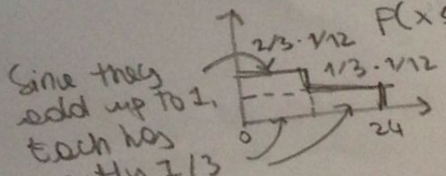
- Each outcome has equal value
- The area (under the function) integrates to 1

$$f(x) = \frac{1}{360}$$

=

Time of day when people are born.
Twice as likely to be born before noon

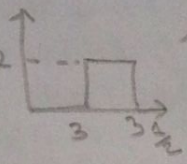
$$f(x \leq \text{noon}) = 2 \cdot f(x > \text{noon})$$



Since they add up to 1, each has exactly 1/3

You're dropping a package out of an airplane and it takes between 3 and 3½ min to reach the ground

In between the prob. of density is uniform



a = 2

½ · 2 gives us the area of 1

densities can be larger than 1!

Correlation and causation!

SICK	IN HOSPITAL	DIED
	40	4
AT HOME	8 000	DIED 40

Chances of dying in hospital are 40 times larger than at home

correlation

It doesn't mean by staying you'll reduce your chances of dying

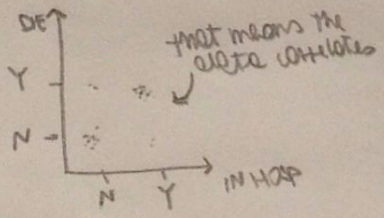
Being in a hospital increases your probability of dying by a factor of 40 ← causation!

The correlation does not imply causation

	IN HOSPITAL	DIED
SICK	36	4
HEALTHY	4	0

	AT HOME	DIED
SICK	40	20
HEALTHY	7960	20

since you're sick, your probability of dying is higher at home



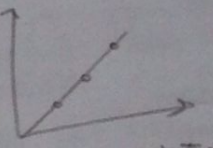
that means the data correlates

The example omitted the sickness variable, that in fact is cause you to die, and also affected your decision on whether you go into a hospital or not.

Once you knew you were sick being in a hospital negatively correlated with you dying

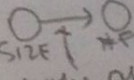
In statistics we call it a confounding variable!

# FIREFIGHTERS	# SIZE FIRE
10	100
40	400
200	2000
70	700



The # of firefighters is correlated with size of fire

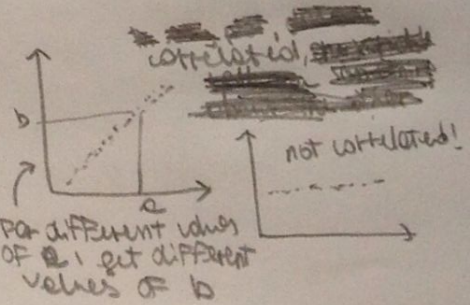
Reverse causation. Firefighters cause fire



But the graph is obvious of the direction of this

causation, size causes # firefighters

We use what know to deduce the direction of the causation



For different values of a, get different values of b

Estimators!

coin flips: 1 = head, 0 = tail

100101 P(HEADS) = 0.5

11011 P(HEADS) = 0.8

Empirical Frequency

DATA $x_1 x_2 x_3 \ldots x_N$ → $\frac{1}{N}\sum x_i$ → always between 0 and 1

maximum likelihood estimator

1 6 6 3 2 6 5 4 6 2 — using MLE

P(1)=0.1 P(2)=0.2 P(3)=0.1

P(4)=0.1 P(5)=0.1 P(6)=0.4

=

Maximum Likelihood Estimator (MLE)

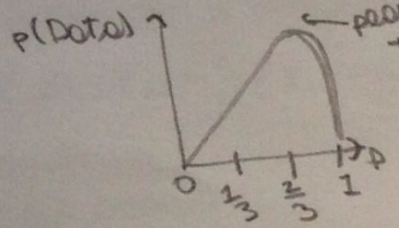
Estimation problem: Data ⇝ P

P ⇝ P(Data)

101 $P=\frac{1}{2}$ $p(Data)=0.125$ $(0.5 \cdot 0.5 \cdot 0.5)$ multiple independent events!

$P=\frac{1}{3}$ $p(Data)=0.074$ $(0.33 \cdot 0.66 \cdot 0.33)$

$P=1$ $p(Data)=0$ $(1 \cdot 0 \cdot 1)$



peaks at 2/3

that point is the one that maximizes the likelihood of the data

Therefore it's called MLE!

=

MLE always assumes the coin is loaded when we have few data

Add FAKE data! DATA ⇝ DATA + HEADS + TAILS ← LAPLACIAN ESTIMATOR!

1	1	0.667
001	0.33	0.4
1010	0.5	0.5
11	1	0.75

FAKE data pulls everything towards 0.5

The last has more data therefore is further away from 0.5

FLIP

1 P(HEADS) = 1

0 P(HEADS) = 0

...

... with plenty of data, the laplacian estimator gives about the same result as the maximum estimator, but when data is scarce, works much better than the maximum est. the laplacian est.

=

Roll a die, you get... 1232 123456 ← Laplacian estimator: one data point for each possible outcome

	MLE	LAPLACE
P(1)	0.25	0.2
P(2)	0.5	0.3

$\frac{1}{N+K}(1+\sum x_i)$ K = datapoints

Averages!

MMM

mean median mode

Mean $\frac{1}{N}\sum x_i = 174$K

↑ sum all prices and divide by # of prices

HOUSE PRICES

190 K

170 K

165 K

180 K

165 K

The mean isn't useful when there are strong outliers

HOUSE PRICES

2 M

125 K

148 K

162 K

170 K

325 K

180 K

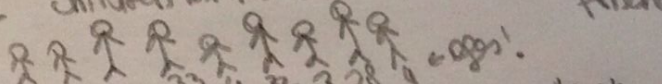
Median

House prices → SORT → pick the one in the middle

... → 148K → 160K

...

Mode children birthday party



4 3 32 33 4 32 3 28 4 ages!

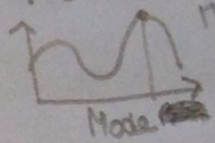
Mean = 17 There are no teenagers though

Median = 4 Not useful as well

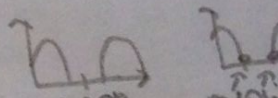
...

Mode, the age most frequently represented at the party

Mode = 4



Mode is useful if your distribution of data is multimodal



They are usually irrelevant here

Ties are broken at random (with Mode and Median)

Variance! close to the mean

Friends ages 17 19 18 17 19 mean=18

Family ages 7 38 4 23 18 mean=18

all over the place

Mode, median and mean don't really capture the spread of the data

what does capture the spread is called the variance

1 subtract mean from each data item

Friends -1 1 0 -1 1 mean=0

Family -11 20 -14 5 0 mean=0

2 Add up the squares of those values and normalize!

Friends 1 1 0 1 1

→ $\frac{4}{5}$ ← variance

Family

121 400 196

25 0

→ 148.4

...

Variance

$\frac{1}{N}\Sigma(x_i-\mu)^2$ ← this is the mean

↑ Variance is a measure of how far the data is spread

standard deviation!

The variance is the quadratic deviation from the mean

IF you don't want something quadratic

STANDARD DEVIATION = $\sqrt{\text{VARIANCE}}$

FRIENDS ... MEAN = 18, VARIANCE = 0.8

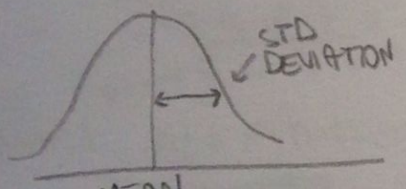
STD DEVIATION = 0.8944

FAMILY ... MEAN = 18, VARIANCE = 148.4

STD DEVIATION = 12.18 ↑ How much you would expect the age of an individual family member to deviate from the mean

... While here you expect the deviation to be just below 1 year

=



$\mu = \frac{1}{N}\Sigma x_i$ $\sigma^2 = \frac{1}{N}\Sigma(x_i-\mu)^2$ $\sigma = \sqrt{\sigma^2}$

σ^2 without $\mu = \frac{1}{N}\Sigma x_i^2 - \frac{1}{N^2}(\Sigma x_i)^2$

Velocity raise

Fixed amount 1000 ↑

$\mu' = \mu + 1000$ ← mean shifts of 1000

$\sigma' = \sigma$ ← std deviation and variance don't change

Relative raise 20%

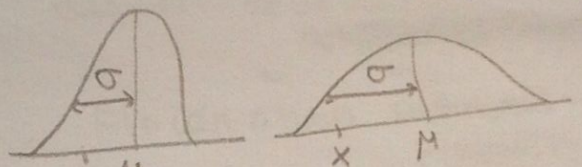
$\mu' = 1.2 \cdot \mu$

$\sigma' = 1.2 \cdot \sigma$

(For variance would be 1.2^2)

Standard score

For any gaussian, you can state how far in σ out a point x is



↑ relative to the mean and variance these points correspond

That's called the standard score

$z = \frac{x-\mu}{\sigma}$

Outliers!

NAME	AGE
JOSEPH	22
MARIA	21
SUSAN	24
MARC	20
TOM	211
JACK	23

probable typo just ignore it!

Quartiles

3 3 4 5 6 6 7 7 8 8 9 ← median!

3 3 4 5 6 | 6 7 7 8 8 9

↓ lower quartile ↑ upper quartile

interquartile = 8 − 4 range

This range is the data you use to calculate things such as the mean, data outside that range are the ignored ↑ to remove outliers

The correct # of elements is given by

$4 \cdot n + 3$

↑ # quartiles

↑ lower quartile = median = upper quartile

Everything else you just pick something in between

=

Ages 19 20 21 22 23 24 25

Frequency 2 1 1 3 2 1 1

lower quartile = 20

median = 22

upper quartile = 23

mean = 21.85 ↑ within interquartile range

=

-89 13 17 19 33 1489

interquartile range

mean = 290.6 ↓ mean = 22

Percentile

10th percentile

10% 90%

↑ k-th percentile is k% through the data

20 21 22 24 27

↑ remove the upper 20% percentile will remove 211

Binomial Distribution

2 COIN FLIPS

H H
H T
T H
T T

HEADS = # TAILS

5 COIN FLIPS
HEADS = # TAILS
→ 0

4 COIN FLIPS

HHTT
HTHT
HTTH
[redacted]
THHT
THTH
TTHH

HEADS = # TAILS

10 coins 4 heads

$\frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} =$ 210

10 coins 5 heads

$\frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 252$

$\frac{n!}{(n-k)!}$ → $k!$ →

$n! = n(n-1)(n-2)(n-3)\cdots 1$

[redacted] $\frac{n!}{k!(n-k)!}$ ← # coins, # heads

FLIP COIN 3 TIMES

HHH THH
HHT THT
HTH [redacted]
HTT TTT

P(HEADS) = 0.8

P(# HEADS = 1) is 0.096

$0.8 \cdot 0.2 \cdot 0.2 = 0.032$

Flip coin n times P(# HEADS = k)
P(HEADS) = p

# of outcomes with that property $\left[\frac{n!}{(n-k)!k!} p^k * (1-p)^{n-k} \right]$

Binomial distribution
cumulative outcome of many identical coin flips

FLIP COIN 5 TIMES P(HEADS) = 0.5

[redacted] SIZE OF TRUTH TABLE

OUTCOMES WITH ONE HEADS → 5 / 25

P(# HEADS = 1) is 0.15625

$\frac{5!}{2!3!} = 10$ $\frac{10}{32}$

P(# HEADS = 3) is 0.3125

Central limit theorem

"Distribution of the sum of many things"

coin: (0,1) $p(\sum x_i = k) = \frac{n!}{(n-k)!k!} 2^{-n}$ ← if coin is fair

	0	1	2	3	4
n=1	1	1			
n=2	1	2	1		
n=3	1	3	3	1	
n=4	1	4	6	6	1

← pascal's triangle

coin: (0,1,2)

	0	1	2	3	4	5	6	7	8
n=1	1	1	1						
n=2	1	2	3	2	1				
n=3	1	3	6	7	6	3	1		
n=4	1	4	10	16	19	16	10	4	1

$\sum_{i=1}^{n} x_i$ n large → n ≥ 30 → GAUSSIAN!

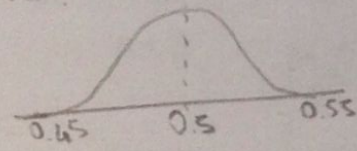
Normal distribution

empirical Frequency

FLIP COIN 1000 TIMES → $\frac{\#\text{HEADS}}{1000} \approx 0.5$

REPEAT EXPERIMENT (↑) 1000 TIMES

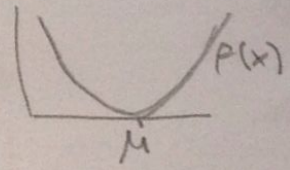
HISTOGRAMS OF THE MEANS I GOT



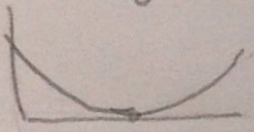
Normal distribution

For any outcome x

$F(x) = (x-\mu)^2$

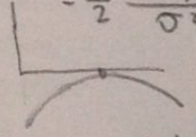


$F(x) = \frac{(x-\mu)^2}{\sigma^2}$ ← makes it wider. the bigger it is



IF smaller than 1 becomes tighter

$F(x) = -\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}$



normal distribution function

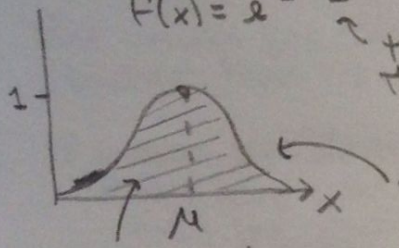
$F(x) = e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

$f(x)$ is maximum when $x = \mu$ and $F(\mu) = 1$

$F(x)$ is minimized when $x = \pm\infty$ and $F(\infty) = 0$

ooo

$f(x) = e^{\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]}$ ← this describes the limit of making infinitely many coin flips

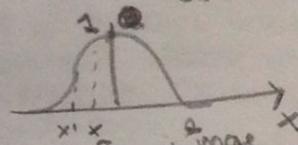


No matter what you do, when n is really large, you get a

That area however doesn't add up to 1, Bell curve like this but to $\sqrt{2\pi\sigma^2}$

We want all areas to add up to 1 as much as we wanted a coin flip and its complement to add up to 1. We normalize it

Normal distribution $\frac{1}{\sqrt{2\pi\sigma^2}} e^{\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]}$



x is more probable outcome than x'

- For a single outcome use p
- For a few outcomes use $\frac{N!}{(N-k)!k!}p^k(1-p)^{N-k}$ Binomial distribution
- For infinitely many i use outcomes $\frac{1}{\sqrt{2\pi\sigma}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$

[redacted]

Manipulating normal distributions

Suppose salaries in industry are normally distributed

Given raise of 10 000 $\mu = 60\,000$ $\sigma = 10\,000$

$\mu' = 70\,000$

$\sigma' = 10\,000$

$\exp\left\{\frac{1}{2}\frac{(x-60000)^2}{10\,000^2}\right\}$

$\exp\left\{-\frac{1}{2}\frac{(x_1-10000-60000)}{10\,000^2}\right\}$ ← μ becomes 70 000

$\exp\left\{-\frac{1}{2}\frac{(x_1-70000)}{10000^2}\right\}$ ← σ is not affected

Double salary

$\mu = 70\,000$ $\sigma = 10\,000$

$\mu' = 140000$ $\sigma = 20000$

$\exp\left\{-\frac{1}{2}\frac{(x-70000)^2}{10000^2}\right\} = \exp\left\{-\frac{1}{2}\frac{(\frac{1}{2}x'-70000)^2}{10\,000^2}\right\}$

$= \exp\left\{-\frac{1}{2}\frac{(\frac{1}{2}x'-\frac{1}{2}140000)^2}{10\,000^2}\right\} = \exp\left\{-\frac{1}{2}\frac{\frac{1}{4}(x'-140000)^2}{10000^2}\right\}$

$= \exp\left\{-\frac{1}{2}\frac{(x'-140000)^2}{20\,000^2}\right\}$

[redacted]

You're a golf player, and you hit the ball really hard, and it lands with some uncertainty (that's german) where the distance is $\mu = 100$ m and your variance is $\sigma^2 = 30\ m^2$. And then you do the something again

$\mu' = 200$ m (combined strokes, expected distance)

$\sigma^2 = 30\ m^2$ (combined variance)

Says we express things in terms of std deviation

$\mu = 100$ $\sigma = 10m$

$\sigma' = 14.14$ m

$N(\mu, \sigma^2) \to A$

$eA + b$

$\mu' = e\mu + b$

$\sigma^2 = e^2\sigma^2$

$A \sim N(\mu, \sigma^2)$ $B \sim N(\mu, \sigma^2)$

A+B $\mu' = 2\mu$ $\sigma^2 = 2\sigma^2$

A−B $\mu' = 0$ $\sigma^2 = 2\sigma^2$

Most better than average

Most drivers believe they drive better than the average driver

It's possible that even 99% of all people are smarter than the average person

4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5

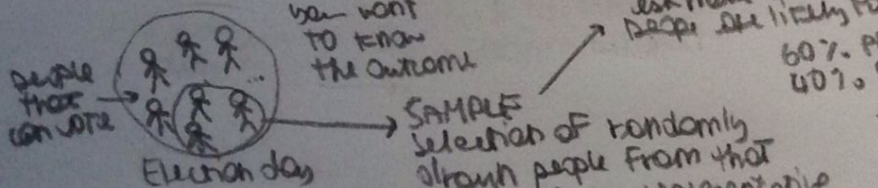
avg = 4.95 95% exceed the average

Sebastian's weight and Arafi

Most are good guesses, but some extreme guesses skew up the actual statistics

Scatterplot of today's weight vs last year's weight we get an approximately linear graph, with some outliers

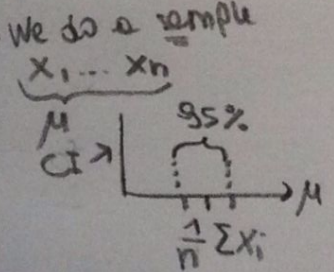
Confidence intervals

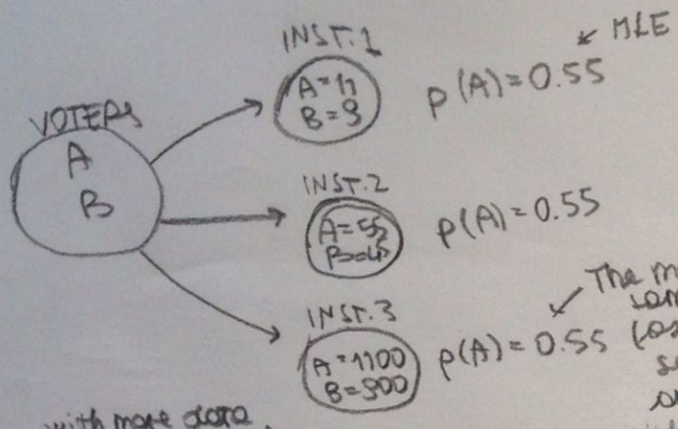


The interval 60% ± 3 is called Confidence Interval!
(we're confident the outcome will be within this range)

TRUE CHANCE OF RECEIVING A VOTE } p

As a statistician we can't ever know that





*** $p=0.1$ $Var(p)=0.09$

$1.96\frac{\sqrt{p(1-p)}}{\sqrt{N}}$

$N=1$	CI =	0.588
$N=10$		0.186
$N=100$		0.0588

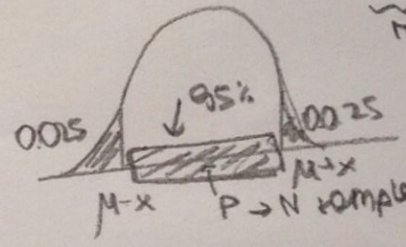
Normal Quantiles

$MEAN \pm 1.96\frac{\sigma}{\sqrt{N}}$

Central limit theorem

FOR LARGE N: $\frac{1}{N}\sum x_i \rightarrow$ NORMAL

($\frac{1}{N}\sum x_i$ = MEAN)



+/-	
90%	1.64
95%	1.96
99%	2.58

← quantiles!

Only correct for large samples (≥ 30)

The std deviation is not dependent on the sample, whether you have 100 or 1000 sample size when you get a pic the outcomes, the std deviation is not affected, but the CI is (it can become more accurate, but doesn't change that much)

=

SIZE OF CI ↔ STD DEVIATION

$\propto \frac{\sigma}{\sqrt{N}}$ σ

$Var(aX) = a^2 Var(X)$

$p=0.5$	$\mu=0.5$	$\sigma^2=0.25$				
	MEAN($\sum x_i$)	VAR($\sum x_i$)	VAR($\frac{1}{N}\sum x_i$)	STA DEV	CI ← STP. DEV * 1.96 for $N=30$	
$N=1$	0.5	0.25	0.25	0.5	0.98	
$N=2$	1	0.5	0.125	0.35	0.69	
$N=10$	5	2.5	0.025	0.16	0.31	

the spread of the mean, goes down as N increases!

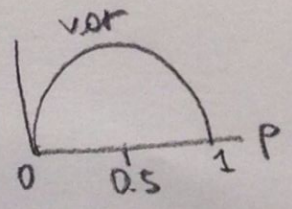
25	0.0025	0.05	0.098
$N=100$	50		

=

$p=0.1$ $\mu=p$ $\sigma^2=p(1-p)$

$p=0$ $\sigma^2=0$

$p=1$ $\sigma^2=0$

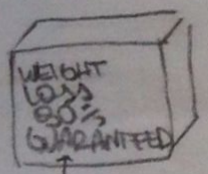


***

Hypothesis testing

SAMPLE —STATISTIC→ INFORMATION —STATISTIC→ DECISION → YES/NO

=



Null hypothesis (or H0)

$p = 0.9$

Lost weight?
YES 11 73.3%
NO 4 binomial!

Alternate hypothesis (or h1)

$p < 0.9$

Medication doesn't work as well as advertised

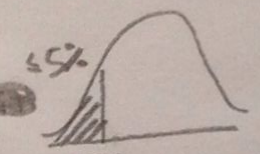
NULL HYPOTHESIS $H_0: p = 0.3$
ALT. HYPOTHESIS $H_1: p > 0.3$

THHTTTHTTHH → 0.45

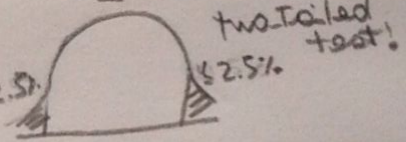
.02 .03 .12 .25 .22 .13 .06 .02 0 0 0 0
0 1 2 3 4 (5) 6 7 8 9 10 11

5% confidence boundary?

there's 5 heads in the outcome, to see if it in the left region (below 7 heads)



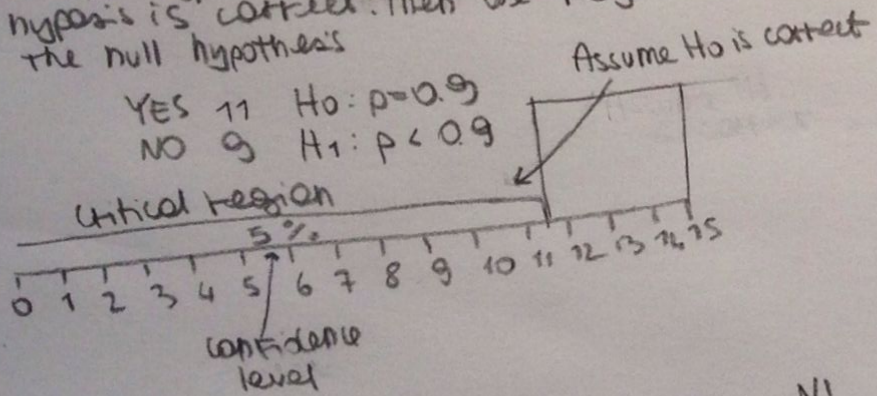
=



$H_0: p = 0.5$
$H_1: p \neq 0.5$

Unless proven wrong we believe null hypothesis is actually correct, or do we have sufficient evidence to be very likelihood the null hypothesis is wrong, the alternative hypothesis is correct. Then we reject the null hypothesis

YES 11 $H_0: p = 0.9$
NO 9 $H_1: p < 0.9$



=

Corresponding probability for every outcome $\frac{N!}{k!(N-k)!} p^k (1-p)^{(N-k)}$

0	0	6	0	12 0.12	YES 11
1	0	7	0	13 0.27	NO 4
2	0	8	0.0002	14 0.34	
3	0	9	0.001	15 0.20	
4	0	10	0.01		
5	0	11	0.04		

Outcomes up to 10 (<5%) is in the critical region

=

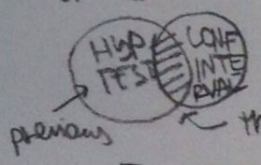
$N = 14$ # HEADS = 3

0 0 .005 .022 .06 .12 .18 .21 .18 .12 .06 .022 .005 0 0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14

≤2.5%

↑ critical region ≤2.5%

Hypothesis test - part 2



=

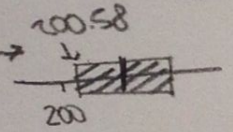
WIKIPEDIA CLAIM
NBA player is 200cm in average tall

take measurements yourself
199 200 201 202 203 204 205 206

confidence interval $\left\{ \frac{1}{N} \pm \alpha \sqrt{\frac{\sigma^2}{N}} \right.$ → 202.5 ± 1.916

↑ mean

$\alpha = 2.635$



we should reject, but depending on ~~with~~ town, we pick different players, and we didn't pick the town independently

=

Dance club average age 26

4:21
6:24
7:26
11:29
2:40

$N=30$
$\mu=26.97$
$\sigma=19.57$
$\alpha=1.96$ ↑ for 95%

$\pm = 26.97 \pm 1.91$

Accept claim based on the sample we got

But taking the sample of one night is not really independent. You should go at random nights and take a person each night

=

Given data $X_1 \dots X_n$

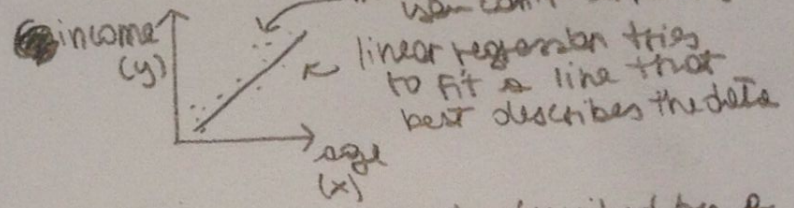
$\mu = \frac{1}{N} \sum X_i$

$\sigma^2 = \frac{1}{N} \sum (X_i - \mu)$

get $t(N-1, P) = \alpha$

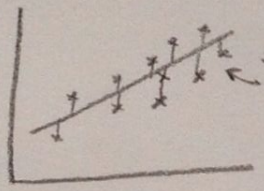
size CI = $\alpha \sqrt{\frac{\sigma^2}{N}}$

Regression



the line is commonly described by a functional relationship between x and y in the form $y = bx + a$

given data, get (a, b) that best fit data



"we assume that our data is the result of this ~~some~~ somewhat linear function: $y = bx + a + noise$ and if that noise happens to be gaussian, then minimizing the quadratic deviation between the data points and the line happens to be the right mathematical answer"

$\sum_{x_i} (bx_i + a - y_i)^2$

↑ distance we are minimizing

x	y
6	7
2	3
1	2
-1	0

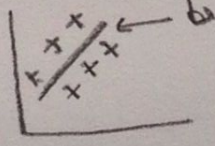
$\bar{x}$ 2 $\bar{y}$ 3

$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

$\rightarrow \frac{16 + 0 + 1 + 9}{16 + 0 + 1 + 9} \rightarrow b$ is 1!

$a = \bar{y} - b\bar{x}$

$\rightarrow 3 - (1 \cdot 2) = 1$



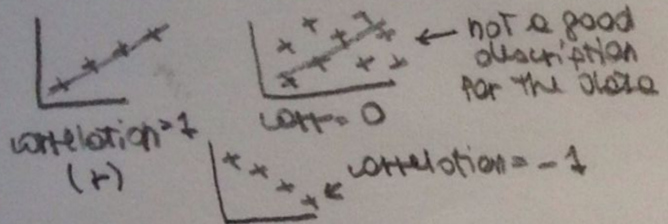
mean of the x_is / y_is

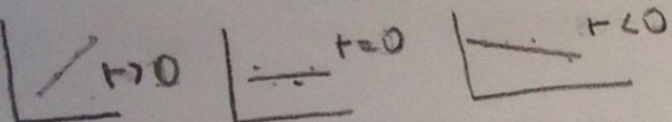
$b = \sum_i [(x_i - \bar{x})(y_i - \bar{y})$

$\sum_i (x - \bar{x})^2$ ← normalize (like the variance)

Correlation!

... is a measure that lies in the range -1, 1 that tells us how far the data is described by a line





Given the linear regression $y = \underline{4}x - 3$
r will be positive ← because of that

Correlation coefficient (r)
- between 1 and -1
- tells how how related two variables are
- 1 and -1 stand for perfectly linear data

$$r = \frac{\sum_i [(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (x_i - \bar{y})^2}}$$ ← covariance, variance of two co-occurring variables

x	3	4	5
y	7	8	9

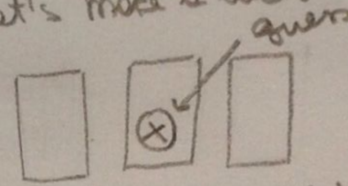
$\bar{x} = 4$, $\bar{y} = 8$

$x - \bar{x}$	-1	0	1
$y - \bar{y}$	-1	0	1

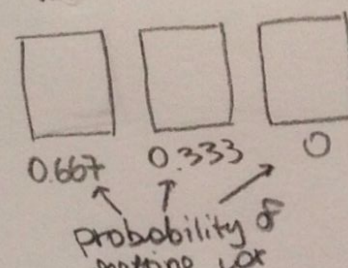
$$r = \frac{2}{\sqrt{2 \cdot 2}} = 1$$

Monty Hall Problem

Let's make a deal!



- 3 doors covered with a car behind one of those
- you get to place a bet,
- then one of the empty curtain is lifted and whether you want to switch



true location	Monty shows...	prob.
1	1	[illegible]
$p(1) = \frac{1}{3}$ 1	2	[illegible]
1	3	1/3
2	1 $\frac{1}{2}$	[illegible]
$p(2) = \frac{1}{3}$ 2	2	[illegible]
2	3 $\frac{1}{2}$	1/6
3	1	[illegible]
$p(3) = \frac{1}{3}$ 3	2	[illegible]
3	3	0

won't show the one I guessed

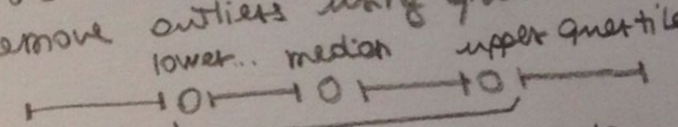
$\left.\right] \frac{1}{3}$

truth table provided that I picked door 2

so Monty picks door 3, we be in either three cases!
p(1) is 2/3 p(2) is 1/3 p(3) is 0

Weight case study

163 samples
Remove outliers using quartiles



picked! 83 samples

$n = 83$ $\sum x_i = 6618.47$ $\sum x_i^2 = 528679$

$\mu = 79.74$ VAR = 11.1
STD DEV = 3.33 $\pm = 0.72$ 95%
$\frac{VAR}{\sqrt{n}} \cdot 1.96$

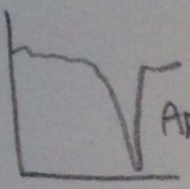
Sebastian's weight = 73 kg STD SCORE = 2

case study - regression
- Run scatterplot to LOOK AT DATA!
- Can't use quartiles as data is paired!
  ↑ Sebastian's weight now vs last year
- Remove [redacted] pair of datapoints where either component is outlier

$\bar{x} = 81.2$ $\bar{y} = 79.7$ stddev(x) = 10.6
stddev(y) = 9.34
r = 0.76 !
b = 0.67 ← [redacted] slope
a = 25.2 ← intercept

$y = 25.2 + 0.67x$

Flash crash example



typical trading day $x_1 \cdots x_{70\,000}$

$\Delta t = \frac{x_{t+1} - x_t}{x_t}$ MEAN(Δt) = -0.00074%.
5%. STDEV(Δt) = 0.07344%.

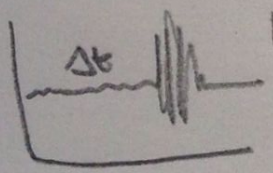
CI = -0.02708% - 0.0256%.

can get 3500 outliers in expectation, a day!

change to 6.5%.

-0.0881% - 0.08662%.

x_t	x_{t+1}	$\Delta t = \frac{x_{t+1} - x_t}{x_t}$	
286.85	286.83	0.00697	↑ Not abnormal!
247.6	247.55		
242.5	240.0		↖ Abnormal!
205.71	201.00		↙ Abnormal!



May 6, 2020

Challenger example

temperature Jan 28, 1986
36°

~~temperature~~ → temp

x_i	Y_i ← hoverings failed (affected by temperature)
70	1
57	1
63	1
70	1
53	2
75	2
58	1

Linear regression!

N=7 $\Sigma x_i = 446$

ΣY_i $\Sigma x_i^2 = 28816$

$\Sigma x_i Y_i = 574$

b is a positive slope, suggests that as temp. gets warmer the chance of failure go up

$$b = \frac{\Sigma x_i Y_i - \frac{1}{N}\Sigma x_i Y_i}{\Sigma x_i^2 - \frac{1}{N}(\Sigma x_i)^2} = 0.00243$$

$$a = \frac{1}{N}\Sigma Y_i - b\frac{1}{N}\Sigma x_i = 1.194$$

$36 \cdot b + a = 1.25$ ↑ actual expected number of hovering failures for the temperature of 36°

Full dataset (no failures included)

N = 23

$\Sigma x_i = 1600$

$\Sigma x_i^2 = 112{,}400$

$\Sigma Y_i = 9$

$\Sigma x_i Y_i = 574$

b = -0.0475

a = 3.698

$36 \cdot b + a = 1.99$