Data, Inference & Applied Machine Learning

Course: 18-785

Patrick McSharry

patrick@mcsharry.net

www.mcsharry.net

Twitter: @patrickmcsharry

Fall 2022

ICT Center of Excellence
Carnegie Mellon University

Week 4

Course outline

Week	Description
1	Measurement, data types, data collection, data cleaning
2	Data manipulation, data exploration, visualization techniques
3	Probability, statistical distributions, descriptive statistics
4	Statistical hypothesis testing, quantifying confidence
5	Time series analysis, autoregression, moving averages
6	Linear regression, parameter estimation, model selection, evaluation

Data & Inference

WEEK 4A

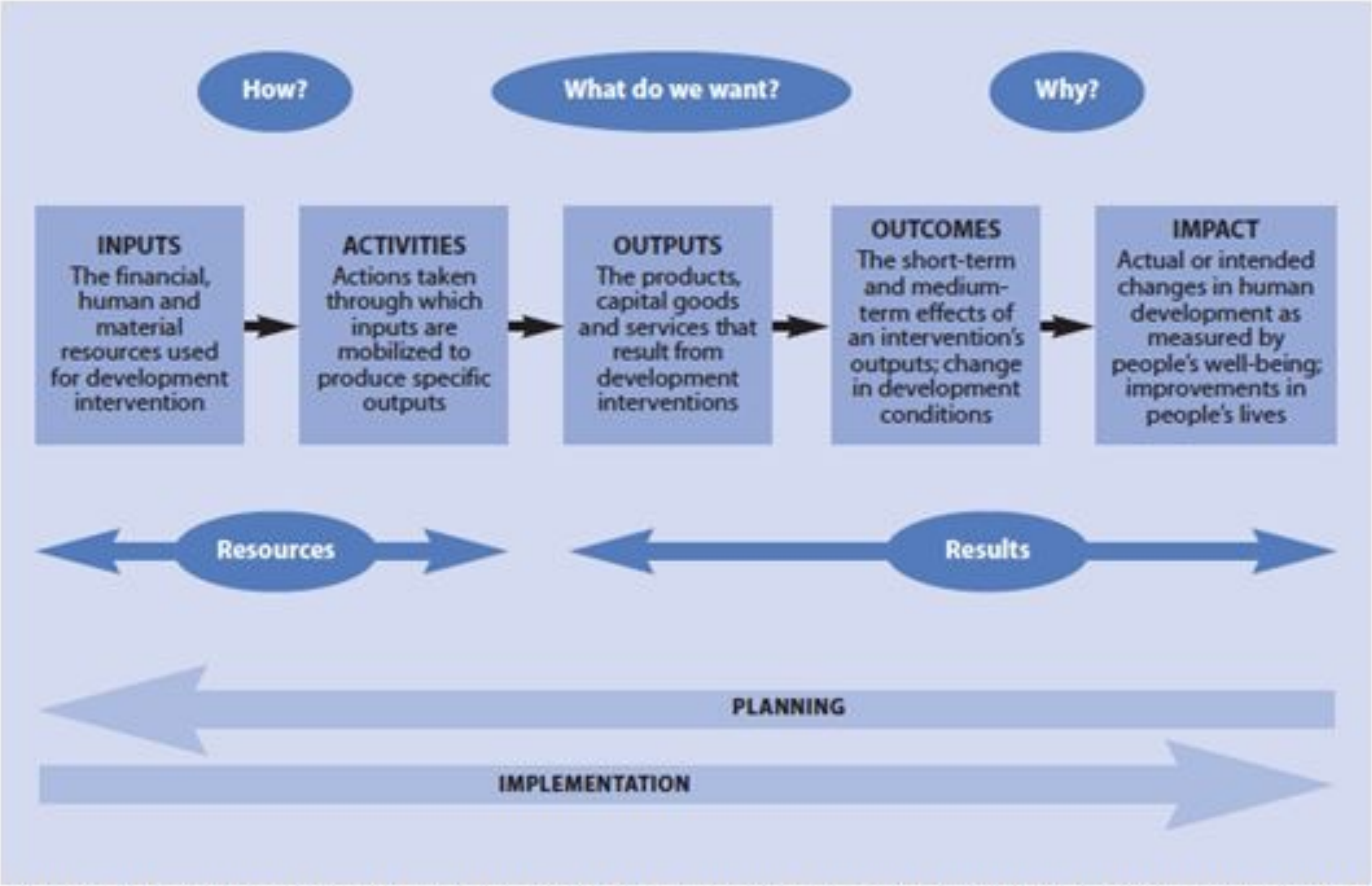
Today’s Lecture

No.	Activity	Description	Time
1	Challenge	Monitoring the quality of a product	10
2	Discussion	Making decisions with statistics	10
3	Case study	Guinness	10
4	Analysis	Statistical hypothesis testing	20
5	Demo	Techniques for testing	20
6	Q&A	Questions and feedback	10

Monitoring and Evaluation (M&E)

- Monitoring and evaluation (M&E) is a process that helps improve performance and achieve results.
- Its goal is to improve current and future management of outputs, outcomes and impact.
- Inputs are used in order to carry out activities.
- Activities lead to services or products delivered (**outputs**).
- The outputs start to bring about change (**outcomes**) and eventually this will hopefully contribute to the **impact**.

M&E in development



How?

What do we want?

Why?

INPUTS The financial, human and material resources used for development intervention

ACTIVITIES Actions taken through which inputs are mobilized to produce specific outputs

OUTPUTS The products, capital goods and services that result from development interventions

OUTCOMES The short-term and medium-term effects of an intervention's outputs; change in development conditions

IMPACT Actual or intended changes in human development as measured by people's well-being; improvements in people's lives

Resources

Results

PLANNING

IMPLEMENTATION

Source: UNDP Handbook on Planning, Monitoring and Evaluation for Development Results (2009, p55)

M&E

M&E terminology	Description	Example
Inputs	Resources (financial, human and material)	Costs of seeds, transport & staff
Outputs	List of activities	• Travelling to field to deliver seeds • Conduct training with farmers • Amount of seeds distributed • Number of farmers trained
Outcomes	Observed effects of the outcomes	• Farmers plant the seeds • Seeds grow into crops • Crops are harvested, and then eaten or sold.
Impact	Degree to which the outcomes observed are attributable to the activities	Better standard of living in the long-term for farmers and their families.

Source: www.intrac.org

Imagine if …

- A university randomly selected some students and pays for transport and lunch
- Offers some classes where the content was of interest to a few students
- Counts the number of students and the number of classes
- Employs a survey to determine if the students liked the classes, transport and lunch
- Provides certificates to confirm that the student attended the classes

M&E

M&E terminology	Description	Example for CMU Africa
Inputs	Resources (financial, human and material)	Costs of building, equipment & staff
Outputs	List of activities	• Attracting students • Delivery of MSc courses • Number of students enrolled • Number of students graduated
Outcomes	Observed effects of the outcomes	• MSc courses • Students doing practicums • Students doing internships • Students graduating
Impact	Degree to which the outcomes observed are attributable to the activities	Better standard of living in the long-term as students follow careers in ICT sector. • Jobs, positions and salaries

With No Boys Born in Nearly 10 Years, a Polish Village Finds Fame in Its Missing Males

MIEJSCE ODRZANSKIE, Poland — television crews have come searching for answers about a small Polish village's strange population anomaly.

The detail first attracted the attention of the Polish news media when the village sent an all-girl team to a regional competition for young volunteer firefighters.

It has now been almost a decade since the last boy was born in this place, with the most recent 12 babies all having been girls.

Source: New York Times, 16-Aug -2019

Class Poll 1

- Is “No Boys Born in Nearly 10 Years”?
- Significant
- Not significant

Slido.com
#41843

Is this significant?

- The probability of 12 consecutive girls being born in Miejsce Odrzanskie is $(½)^{12}$ =1/4096.
- What is the probability of the last 12 children born in some town somewhere in the world all being the same sex?
- The GeoNames database is an online database containing details of every town in the world with a population of over 500, and it suggests there are just under 200,000 such towns across the planet.
- Based on this, we'd actually expect roughly 50 towns in the world with 12 consecutive girls (1/4096 x 200,000), and another 50 with 12 consecutive boys.
- So, although this run of girls seems like a strange and unique event to the people of Miejsce Odrzanskie, there are in fact probably about 99 other places in the world where something similar is happening right now.
- For a small village of just 272 people with a birth rate of not much more than one per year. That means that this run of 12 girls is extended over almost a decade, which is what has attracted so much attention.

Craig Anderson, University of Glasgow
https://theconversation.com/polish-village-hasnt-seen-a-boy-born-in-nearly-10-years-heres-how-that-computes-122176

Drug testing

- A new drug is tested to see if it is effective in curing a certain disease.
- The Thalidomide tragedy led to strict rules about testing before a drug can be put on the market.
- We therefore assume the drug is not effective, or it is actually harmful, until the tests indicate otherwise.

Case Study

Guinness

- Guinness is one of the most successful beer brands worldwide.
- Brewed in almost 60 countries and is available in over 120.
- Annual sales total 850 million litres (1.5 billion Imperial or 1.8 billion US pints).
- Guinness generates almost €2 billion annually.
- Guinness ran an advertising campaign in the 1920s which stemmed from market research – when people told the company that they felt good after their pint, the slogan was born – "Guinness is Good for You".

Making beer

- Raw materials like hops, barley, and malt were extremely sensitive to how they were grown, processed, and stored.
- The fermentation process, which produced the alcohol in the beer, was extremely sensitive to minor changes.
- Too much yeast could completely ruin a batch. Yet live yeast cultures were constantly growing and very difficult for workers to measure under a microscope.
- If the degree of saccharine in the malt extract was too low, the beer was weak.
- If it was too high, the beer was too strong and its stability and shelf life were compromised.

http://blog.minitab.com/blog/statistics-and-quality-data-analysis/beer-statistics-and-quality

William Gosset (1873-1937)

Quality of barley

- In 1908, Guinness was testing new barley varieties that had been grown in trials.
- Many of the trials were small and could not easily be repeated.
- It would take another year to grow more grain, and the weather would be different.
- William Gosset devised a technique for evaluating the results produced from small samples.

Student’s t-test

- Guinness staff could not publish under their own name then, because of a previous case when a staff member published trade secrets.
- Gosset published his technique in an international journal, *Biometrica*, under the pseudonym “Student”.
- His paper was entitled “The probable error of a mean” and his technique, “Student’s t-test”, is widely used by statisticians.

Obama enjoys Guinness in Ireland

Visiting his ancestral home in Moneygall, Ireland, this week Barack Obama announced that the Guinness tastes better in Ireland than anywhere else in the world.
"The first time I had Guinness," Obama said, "is when I came to the Shannon airport. We were flying into Afghanistan and so stopped in Shannon. It was the middle of the night. And I tried one of these and I realised it tastes so much better here than it does in the States ... You're keeping all the best stuff here!"

Trustworthy AI (Class Poll 2)

- Trustworthy AI should be:
- (1) lawful - respecting all applicable laws and regulations
- (2) ethical - respecting ethical principles and values
- (3) robust - both from a technical perspective while taking into account its social environment

Slido.com
#41843

Statistical significance (Class Poll 3)

- What does statistical significance mean to you?

Slido.com
#41843

Hypothesis testing

- Hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true.
- A test result is called statistically significant if it has been predicted as unlikely to have occurred by chance alone, according to a threshold probability known as the *significance level*.
- Hypothesis tests are used in determining what outcomes of a study would lead to a rejection of the null hypothesis for a pre-specified level of significance.

Hypothesis testing in four steps

- 1. Construct the null hypothesis H_0 and alternative hypothesis H_1:
  - H_0: the observations are the result of pure chance; and
  - H_1: the observations show a real effect combined with a component of chance variation.
- 2. Identify a test statistic that can be used to assess the truth of the null hypothesis.
- 3. Compute the p-value, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis were true.
  - The smaller the p-value, the stronger the evidence against the null hypothesis.
- 4. Compare the p-value to an acceptable significance value α.
  - If $p<=\alpha$, that the observed effect is statistically significant, the null hypothesis is ruled out, and the alternative hypothesis is valid.

Interpretation

- The difference between a particular measurement and the population mean can be thought of as the signal.
- The noise relates to the amount of variability that one might expect by chance, assuming that the null is true.
- Hypothesis testing aims to quantify how likely such an outcome could occur by chance.
- Test statistic studies signal divided by the noise.

Weight gain?

- The Centers for Disease Control (CDC) reported on trends in weight, height and body mass index from the 1960's through 2002.
- The general trend was that Americans were much heavier in 2002 as compared to 1960; both men and women gained approximately 24 pounds, on average, between 1960 and 2002.
- In 2002, the mean weight for men was reported at 191 pounds.
- Suppose that an investigator hypothesizes that weights are even higher in 2006 (i.e., that the trend continued over the subsequent 4 years).
- The **null hypothesis** is that there is no change in weight, and therefore the mean weight is still 191 pounds in 2006.
- The **alternative hypothesis** is that the mean weight in men in 2006 is more than 191 pounds.

H_0: $\mu = 191$

H_1: $\mu > 191$

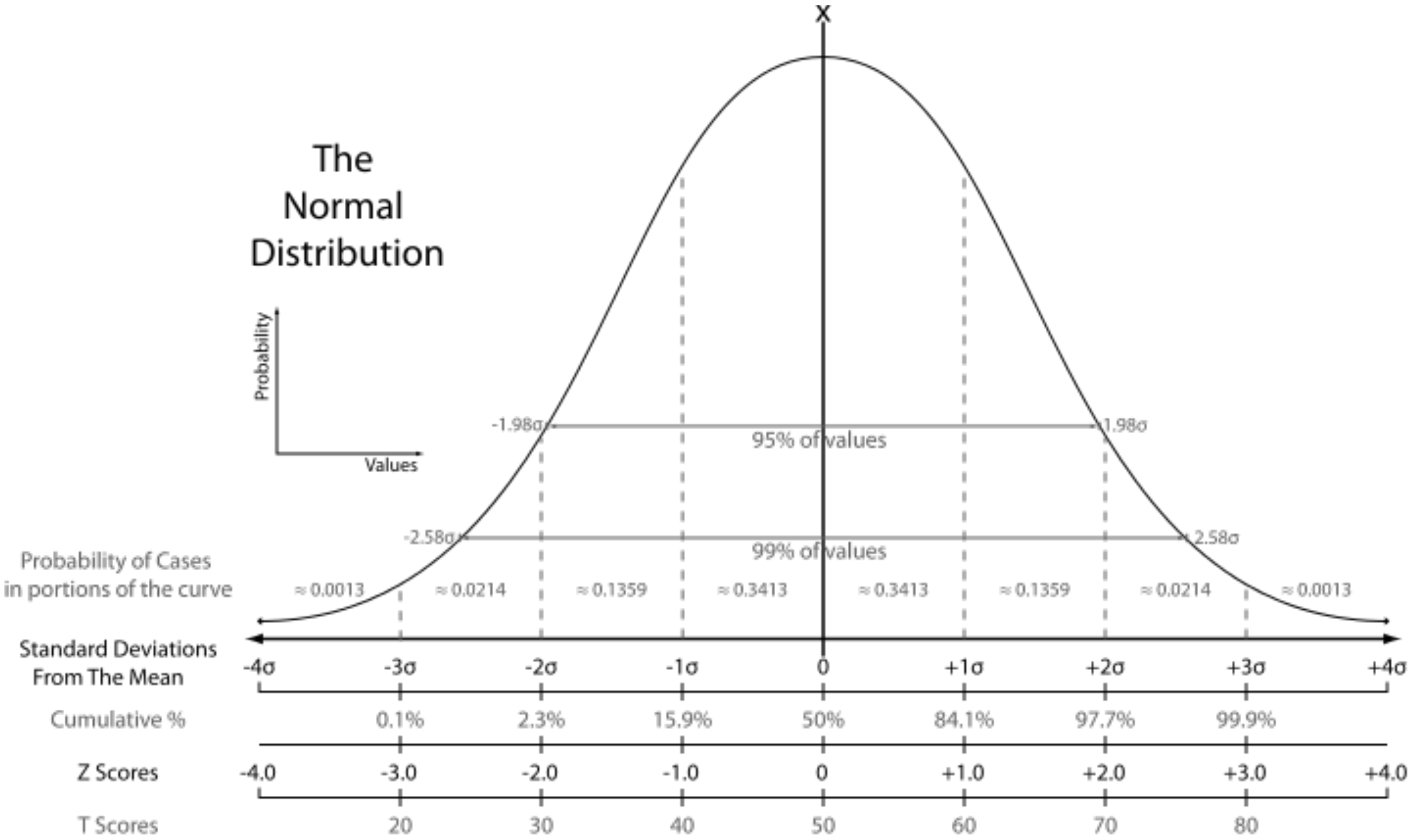
Testing the weight gain hypothesis

- Select random sample of American males in 2006 and measure weights.
- After recruiting n=100 men, the mean weight is 197.1 and the standard deviation is 25.6.
- Do the sample data support the null or research hypothesis?
- The sample mean of 197.1 is numerically higher than 191.
- However, is this difference more than would be expected by chance?

Z-score

- In statistics, the standard score is the (signed) number of standard deviations an observation or datum is above the mean.
- Thus, a positive standard score indicates a datum above the mean, while a negative standard score indicates a datum below the mean.
- It is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation.

Normal distribution and scores



Standard error of the mean

- Standard error of the mean (SEM) describes the variation of the mean of n observations with mean μ and standard deviation σ:

$$SEM = \sigma / \sqrt{n}$$

- The distribution of the averages has a smaller spread than that of the original distribution.
- It is possible to calculate an empirical estimate using $SEM = s / \sqrt{n}$

Test statistic

- Assume that the null hypothesis holds until proven otherwise.
- Determine the likelihood of observing a sample mean of 197.1 or higher when the true population mean is $\mu_0 = 191$.
- We can compute this probability using the z-score:

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

The mathematics

- Given: n=100, $\overline{x}$ = 197.1, μ = 191 and s = 25.6
- Test statistic: z = (197.1-191)/(25.6/10) = 2.38
- P($\overline{X}$> 197.1) = P(Z>2.38) = 1 – 0.9913 = 0.0087
- Hence there is less than a 1% probability of observing a sample mean as large as 197.1 when the true population mean is 191.
- Based on the data and our analysis, the null hypothesis is probably not true.
- P<0.01: null hypothesis rejected

Upper, lower and two-tailed tests

- The alternative hypothesis can take one of three forms.
- An investigator might believe that the parameter has increased, decreased or changed.
- 1. H_1: $\mu > \mu_0$; upper-tailed test
- 2. H_1: $\mu < \mu_0$; lower-tailed test
- 3. H_1: $\mu \neq \mu_0$; two-tailed test

The t distribution

- What if you only have a small number of observations (n < 30)?
- Then you should use the t-test instead.
- The t distribution has heavier tails than the normal distribution.
- It is important to correct for small samples when using the empirical estimate of the SEM.
- We expect more large deviations due to the fact that we divide by a SEM that is too small.

Applying the t-test

- The t distribution also require knowledge about the degrees of freedom, df.
- For the t-test, we use df = n-1.
- The decision of whether to accept or reject the null hypothesis is based on the quantile from the t distribution with df=n-1.
- This gives a p-value which can be compared with our specified acceptable significance value α.

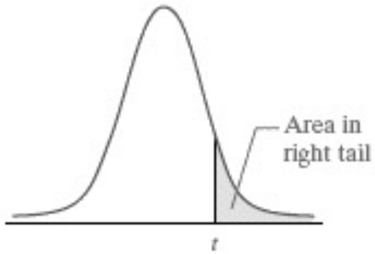


TABLE III

t-Distribution

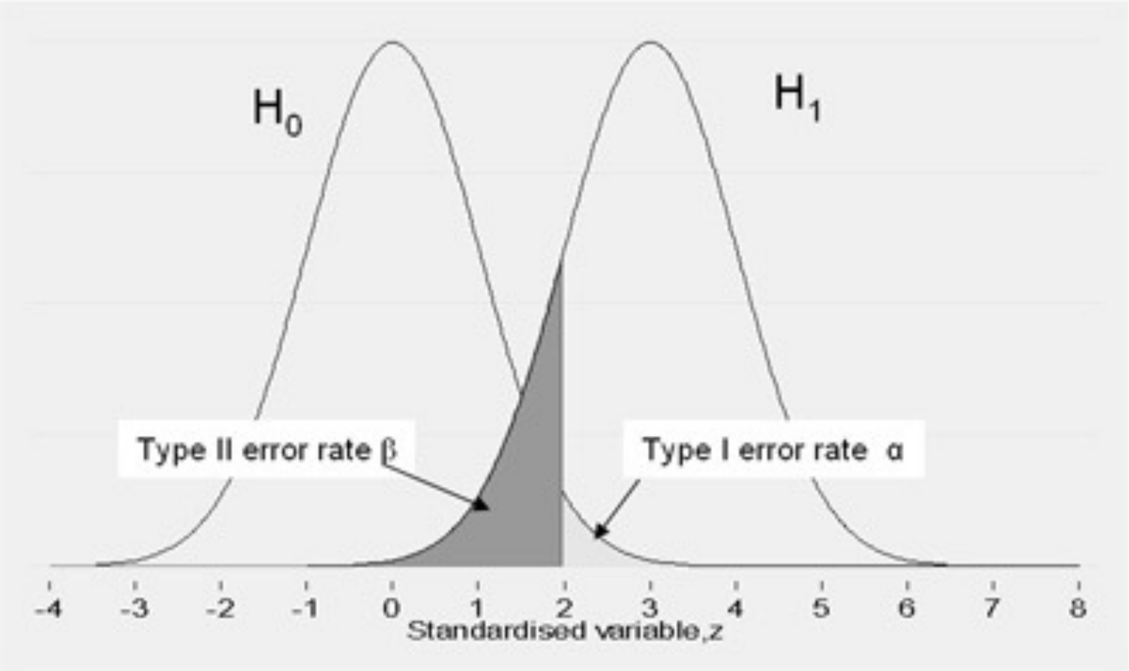
	Area in Right Tail								
df	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**	**0.005**
1	1.000	1.376	1.963	3.078	6.314	12.706	15.894	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787

Type I and II errors

Test \ H_o	True	False
Rejected	Type I error False positive Probability = α	Correct decision True positive Probability = 1-β
Not rejected	Correct decision True negative Probability = 1-α	Type II error False negative Probability = β

The power of a test is defined as 1 - β, and is the probability of rejecting the null hypothesis when it is false.

The most common reason for type II errors is that the study is too small.



www.healthknowledge.org.uk

Hypothesis Testing

- Null hypothesis
- Alternative hypothesis
- Exploring what happens by chance
- What kind of test?
- One-tailed (left and right tailed)
- Two-tailed test

Matlab functions

- mean
- std
- sqrt
- ttest

Q&A

Data & Inference

WEEK 4B

Assignments

- Class attendance and piazza participation are important for this course
- Check piazza using An:Qm format for Assignment n and Question m
- Please do not expect TAs to repeat answers already given on piazza as this is not an efficient use of the limited TA resource
- Justifications for assignment delays (medical, technical or otherwise) must be validated and supported by independent source (e.g. Faculty Advisor)

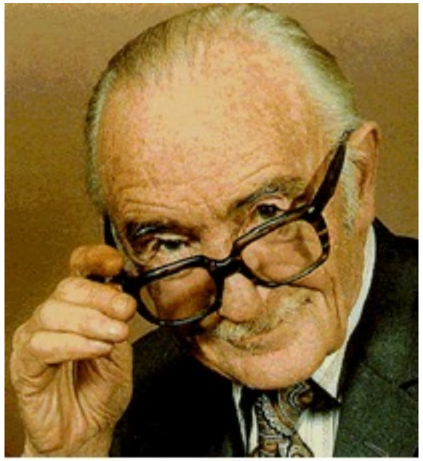
Today's Lecture

No.	Activity	Description	Time
1	Challenge	Measuring the effect of an intervention	10
2	Discussion	Quantifying impact using data	10
3	Case study	A/B testing	10
4	Analysis	Quantifying confidence	20
5	Demo	Techniques for measuring confidence	20
6	Q&A	Questions and feedback	10

Measuring impact

- In 1925, Ronald Fisher, the father of modern statistics, proposed using random assignments to test whether particular medical interventions had some effect.
- The first, so called randomized controlled trials (RCTs) for humans investigated the impact of antibiotics on tuberculosis in the 1940s.

Cochrane



- Archibald Leman Cochrane (1909–1988) was a Scottish doctor noted for his book Effectiveness and Efficiency: Random Reflections on Health Services.
- This book advocated for the use of randomized control trials to make medicine more effective and efficient.

Cochrane and WW2

- During World War II as a captured Medical Officer in prisoner of war camps, his experience led him to believe that much of medicine did not have sufficient evidence to justify its use.
- He said, "I knew that there was no real evidence that anything we had to offer had any effect on tuberculosis, and I was afraid that I shortened the lives of some of my friends by unnecessary intervention.”
- As a result, he spent his career urging the medical community to adopt the scientific method.

UK A-level Grading Algorithm

- Teachers will be asked to supply for each pupil for every subject:
  - An estimated grade (based on mock exam results)
  - A ranking compared with every other pupil at the school within that same estimated grade
- These will be put through an algorithm, which also factors in the school's performances in each subject over the previous three years.
- The idea is that the grades, even without exams due to COVID, will be consistent with how schools had done in the past.

Class Poll 1
Slido.com
#29181

https://www.bbc.co.uk/news/explainers-53807730

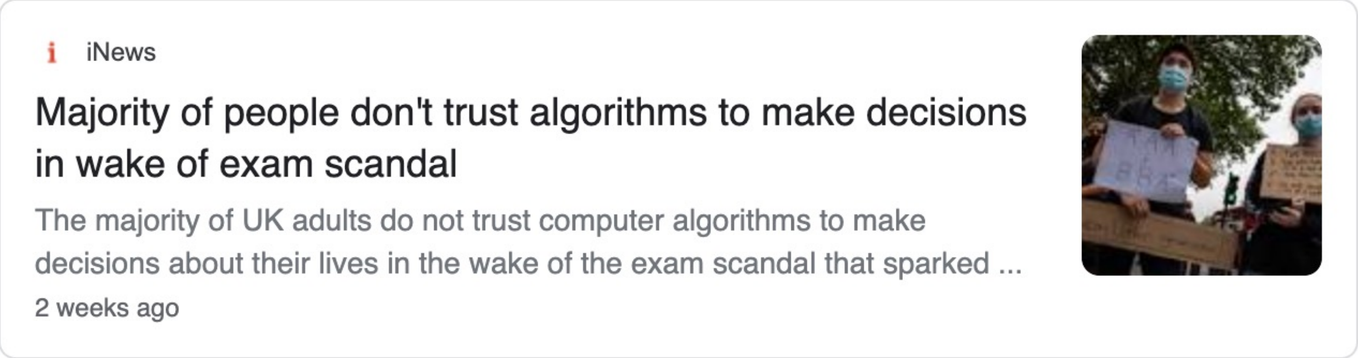
Grading Algorithms

- **How a computer algorithm caused a grading crisis in British schools – CNBC (21-Aug-2020)**
  - Approximately 39% of A-level results were downgraded by exam regulator Ofqual’s algorithm.
  - Disadvantaged students were the worst affected as the algorithm copied the inequalities that exist in the U.K.’s education system.
  - The U.K. government did a U-turn on the grading method as students went on protest.

To determine each student’s results, the U.K. decided to use an algorithm that looked at their mock exam results, as well as their school’s track record in the exams. Lawmakers said the software would give students a “fairer” result after concluding teachers could potentially try to inflate their pupil’s grades.
But the model ended up favoring students from private schools and affluent areas, leaving high-achievers from free, state-schools disproportionately affected.

https://www.cnbc.com/2020/08/21/computer-algorithm-caused-a-grading-crisis-in-british-schools.html

Reactions

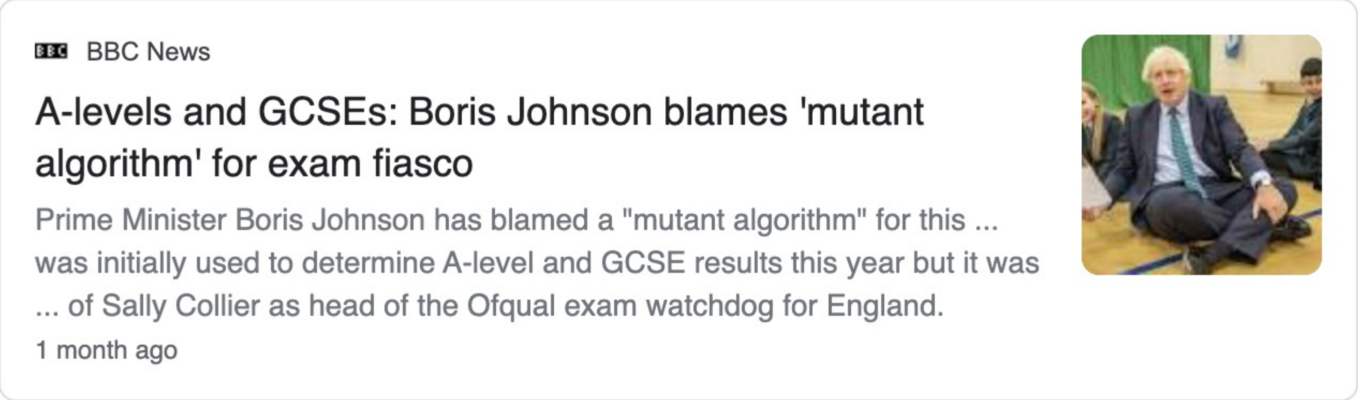


iNews

Majority of people don't trust algorithms to make decisions in wake of exam scandal

The majority of UK adults do not trust computer algorithms to make decisions about their lives in the wake of the exam scandal that sparked ...

2 weeks ago

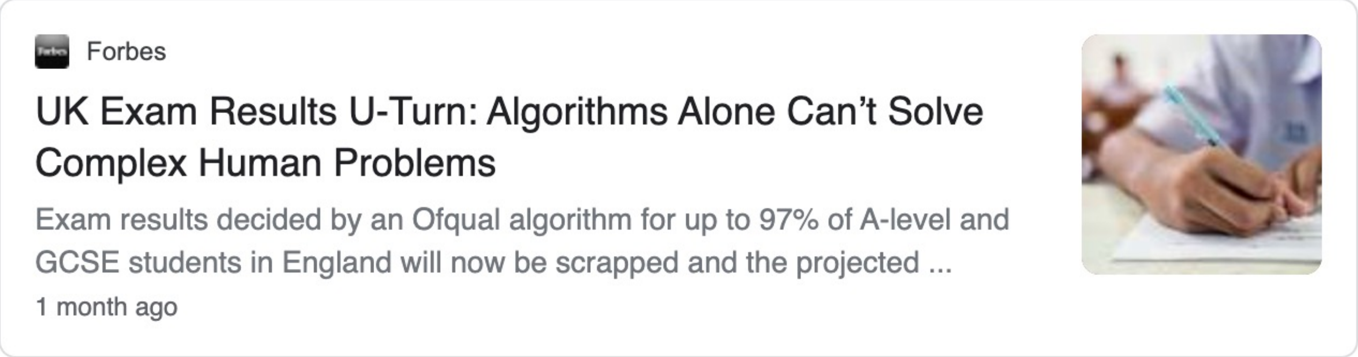


BBC News

A-levels and GCSEs: Boris Johnson blames 'mutant algorithm' for exam fiasco

Prime Minister Boris Johnson has blamed a "mutant algorithm" for this ... was initially used to determine A-level and GCSE results this year but it was ... of Sally Collier as head of the Ofqual exam watchdog for England.

1 month ago



Forbes

UK Exam Results U-Turn: Algorithms Alone Can’t Solve Complex Human Problems

Exam results decided by an Ofqual algorithm for up to 97% of A-level and GCSE students in England will now be scrapped and the projected ...

1 month ago

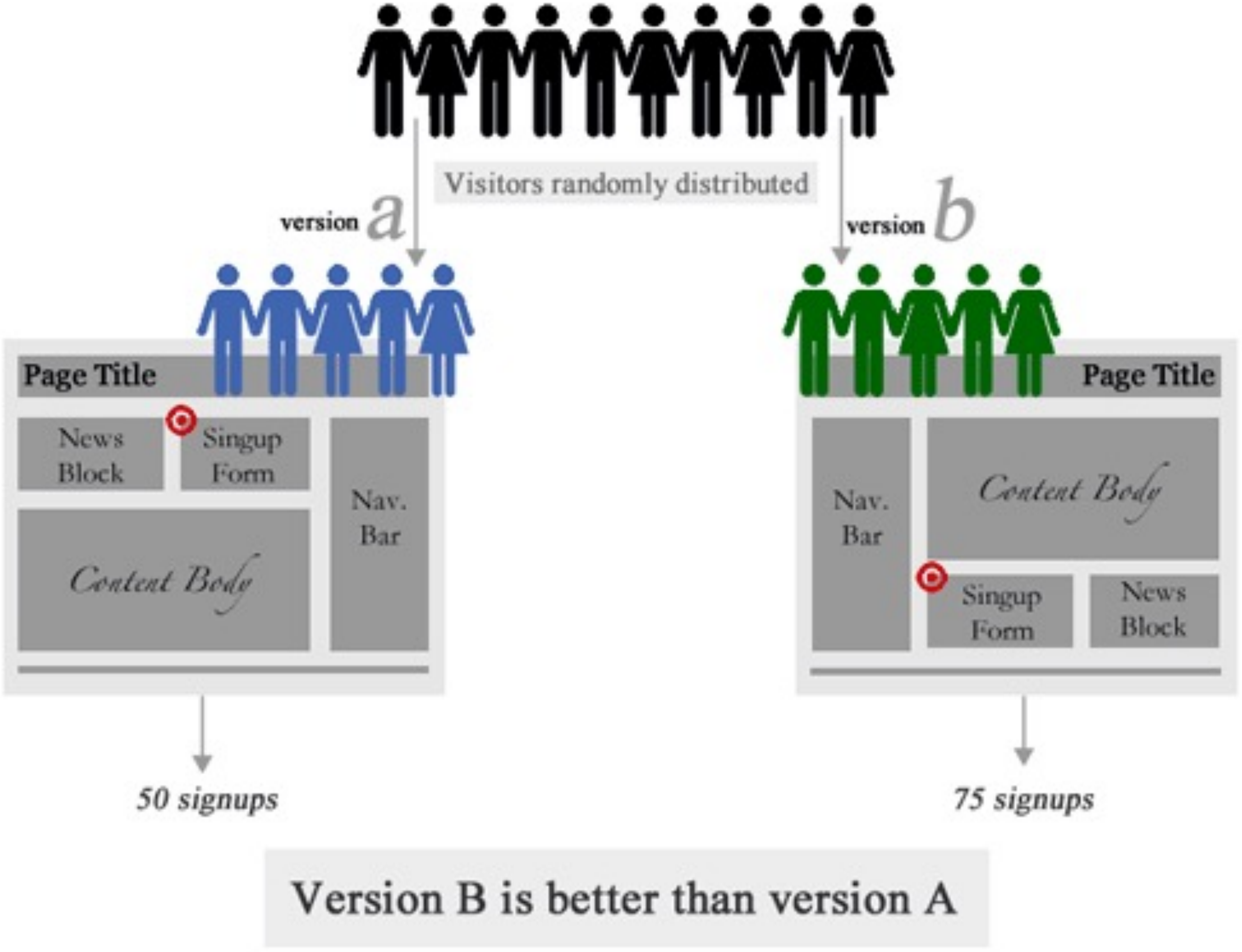
Ethics guidelines for trustworthy AI

- Human agency and oversight
- Technical Robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

Class Poll 2
Slido.com
#29181

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

A/B testing



Source: http://www.smashingmagazine.com

Importance of A/B testing

“ We should use the A/B testing methodology a lot more than we do today ”

Bill Gates, 2008

Capital One

- In 2006, CapOne ran more than 28,000 experiments (new products, advertising approaches and contract terms).
- Is it more effective to print on the outside envelope “Limited Time Offer” or “2.9% Introductory Rate”?
- In 1995, CapOne used a mailing list of 600,000 people and created six groups with 100,000 in each.

Capital One

- This large study allowed CapOne to test both the size and the duration of the teaser rate.
- The study showed that offering a teaser rate of 4.9% for six months was much more profitable than offering a 7.9% for twelve months.

Continental Airlines

- Customer loyalty was examined in the aftermath of a “transportation event”.
- Group 1 received a letter of apology
- Group 2 received a letter of apology and a trial membership in Continental’s Presidential Club.
- Group 3 received nothing and acted as the control group.

Source: Ian Ayres, Super Crunchers

Continental Airlines

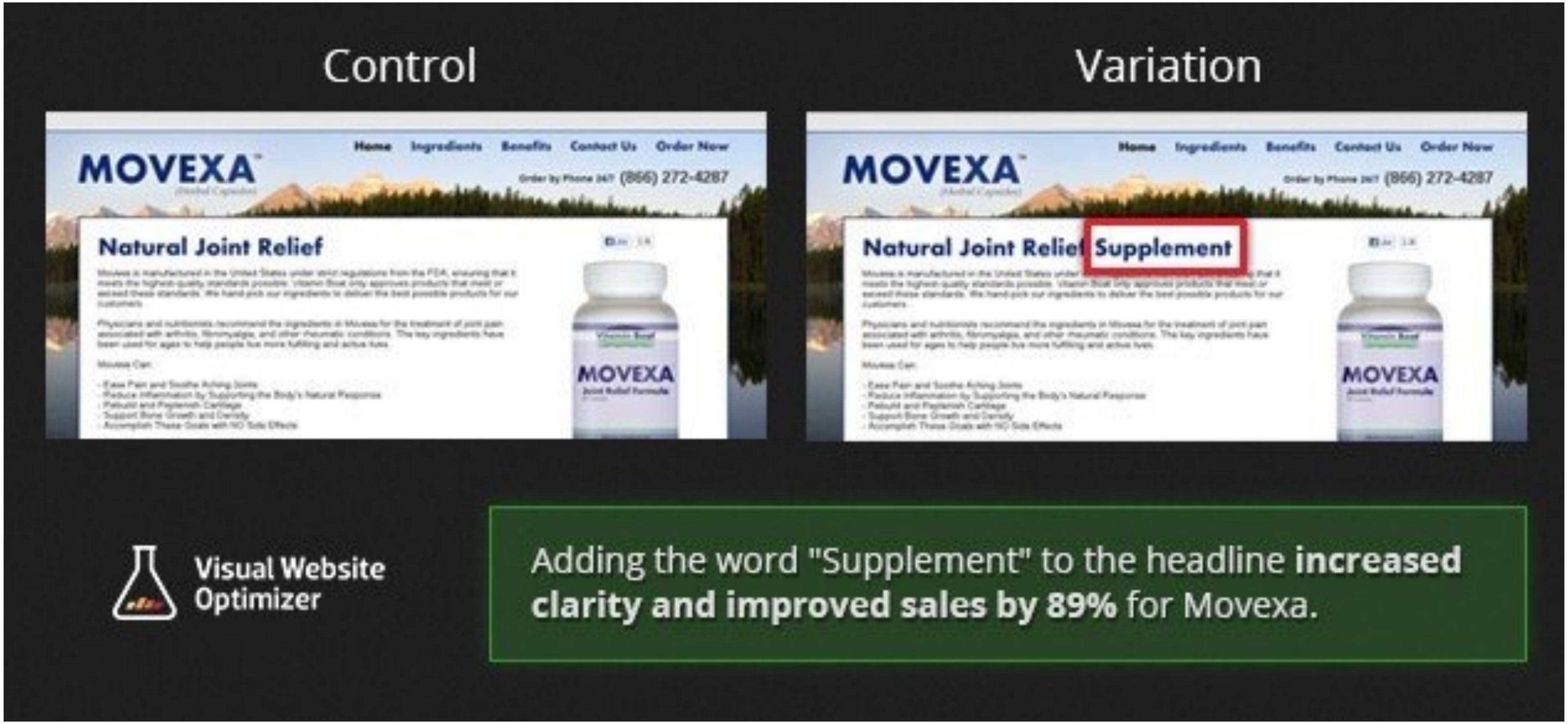
- The two groups receiving letters spent 8%more on Continental tickets in the next year.
- For just 4,000 customers, receiving letters, that translates to extra revenue of $6m.
- Expanding this program to the top 10% of Continental’s customers, the airline has seen $150m in additional revenues who might otherwise have gone elsewhere.

Advertising Survey

- Scenario: You are CEO of a company that sells a drug called Movexa that helps reduce joint pain and discomfort.
- The advertising team wants you to select between two slogans:
  - A: Natural Joint Relief
  - B: Natural Joint Reflief Supplement

Class Poll 3
Slido.com
#29181

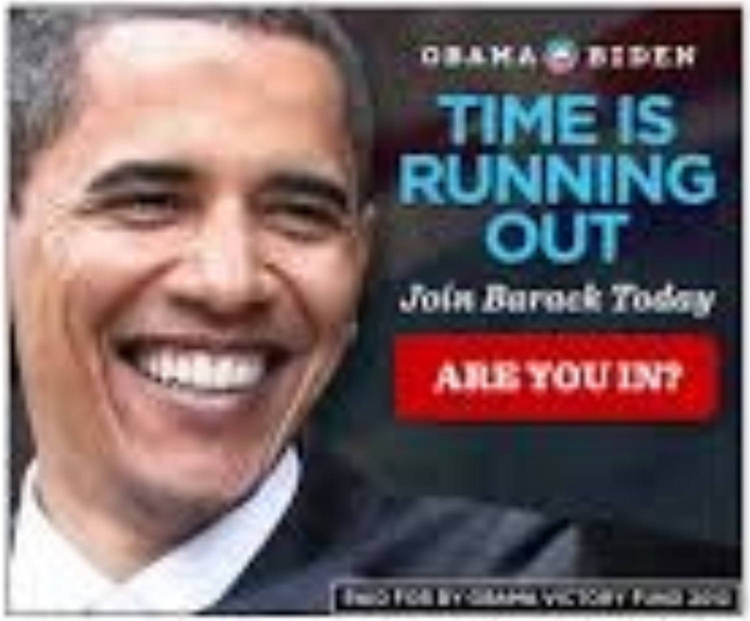
Advertising Result


Control
Variation
MOVEXA
Home Ingredients Benefits Contact Us Order Now
(866) 272-4287
Natural Joint Relief
MOVEXA
Home Ingredients Benefits Contact Us Order Now
(866) 272-4287
Natural Joint Relief Supplement
Visual Website Optimizer

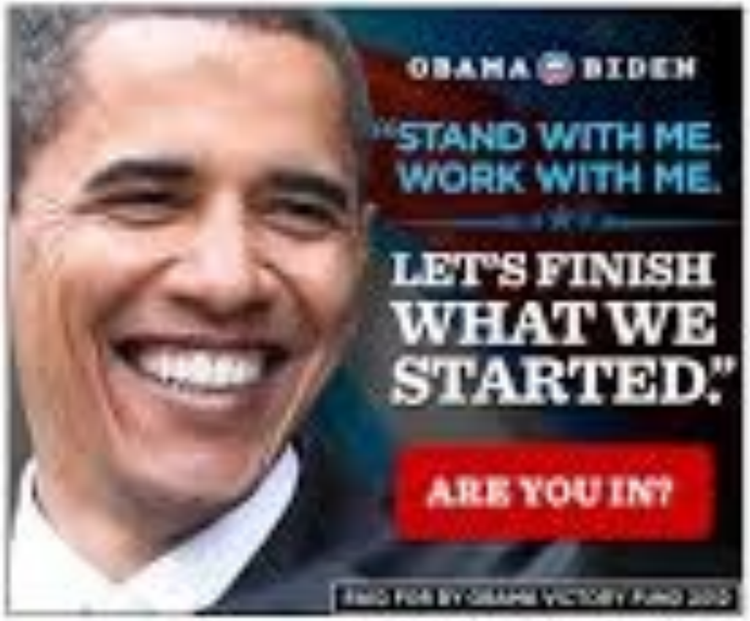
Adding the word "Supplement" to the headline **increased clarity and improved sales by 89%** for Movexa.

https://vwo.com/success-stories/movexa/

Which do you prefer?



PLAY ANIMATION


OBAMA BIDEN
"STAND WITH ME. WORK WITH ME.
LET'S FINISH WHAT WE STARTED."
ARE YOU IN?
PAID FOR BY OBAMA VICTORY FUND 2012

PLAY ANIMATION

A/B testing in politics

- Dan Siroker, CEO of Optimizely and ex-Google first applied A/B testing to politics.
- Obama’s campaign website was optimized.
- Challenge: turning the site’s visitors into subscribers and eventually into donors.
- an A/B test of three new word choices—”Learn More,” “Join Us Now,” and “Sign Up Now”—revealed that “Learn More” garnered 18.6 percent more signups per visitor than the default of “Sign Up.”

Sourcce: http://www.wired.com/2012/04/ff_abtesting/

Obama Campaign

- Similarly, a black-and-white photo of the Obama family outperformed the default turquoise image by 13.1%.
- Using both the family image and “Learn More,” signups increased by 40%.
- A video fared 30.3% worse than even the turquoise image.
- Siroker says that “Assumptions tend to be wrong”.

Overall impact

- Had the team listened to instinct—if it had kept “Sign Up” as the button text and swapped out the photo for the video—the sign-up rate would have slipped to 70% percent of the baseline.
- Instead the rate increased to 140% of baseline.
- By the end of the campaign, it was estimated that a full 4 million of the 13 million addresses in the campaign’s email list, and some $75 million in money raised resulted from Siroker’s A/B testing.

Conversion rate

- The conversion rate, p, is calculated by observing a binomial random variable [conversion or non-conversion].
- We need to estimate p empirically using n visits to the website.
- After observing those n visits, we calculate how many visits resulted in a conversion.
- If we observe m conversions, then our estimate for p is m/n.

Uncertainty

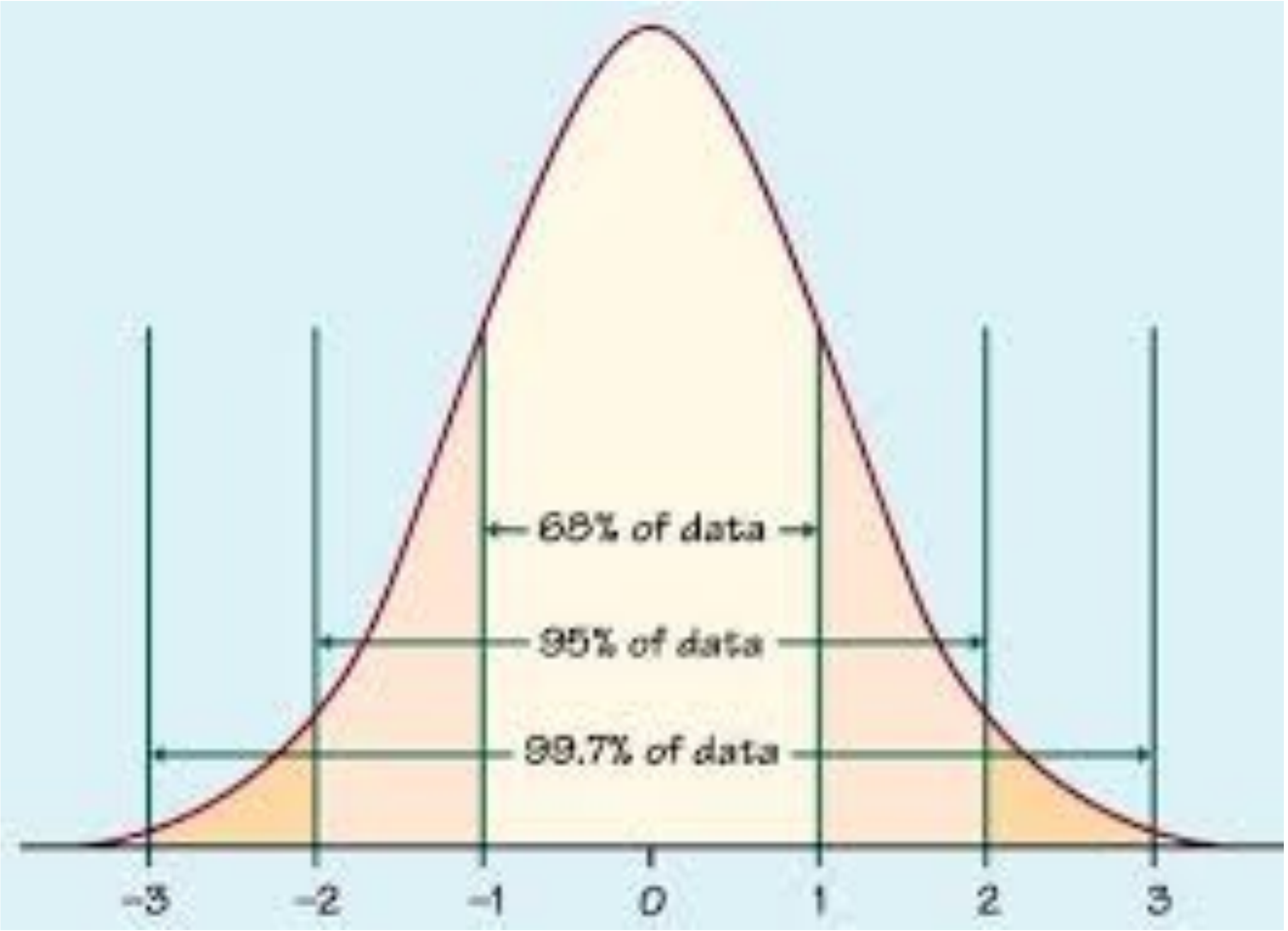
- Repeating this experiment multiple times, it is very likely that, due to chance, each experiment will yield a different value of *p*.
- Collecting different values of p provides a range for the conversion rate.
- To avoid doing repeated experiments, we use the standard error to quantify how much deviation from average conversion rate, p, can be expected if this experiment is repeated multiple times.

Confidence Intervals

- A statistical method for calculating a confidence interval around the conversion rate is used for each variation in A/B testing.
- For a given conversion rate, p, and n trials, the standard error is given by:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Standard Normal



Confidence in rates

- This allows us to determine the confidence interval for the conversion rate.
- To get the 95% range, we multiply the standard error value by 1.96 (based on the 95th percentile of a standard normal distribution).
- This results in a 95% confidence that the conversion rate, p, is in the range of

$$p \pm (1.96 \times SE)$$

Implementing A/B testing

- The A/B test uses conversion and view events to calculate the conversion rates, confidence intervals and statistical significance.
- The percentage change of the conversion rate between the test variation A and the control variation B:
- Change Percent = $(p_A - p_B)/p_A$

Significance

- We wish to determine whether the results are significant are not.
- This approach provides confidence that the observed conversion rates for each variation are not simply different because of random fluctuations.
- When making critical decisions about expenditure, it is important to be able to measure the likely impact in terms of improved profitability.
- A cost-benefit analysis should be performed before taking action.

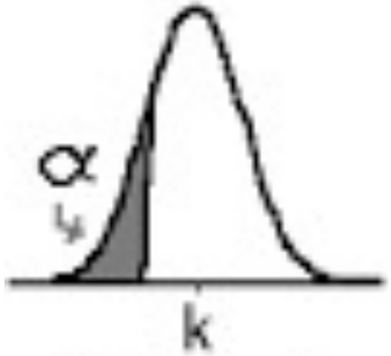
Z-score

- The z-score is calculated as follows:

$$z = \frac{p_A - p_B}{\sqrt{SE_A^2 + SE_B^2}}$$

- The z-score measures the number of standard deviations between the control and test variation mean values.
- Based on a standard normal distribution, the result is significant when the z-score is far into either the left or right tail of the distribution.

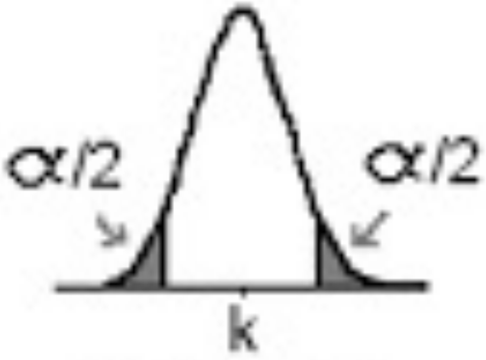
Interpreting the z-score



$H_0: \mu = k$

$H_1: \mu < k$

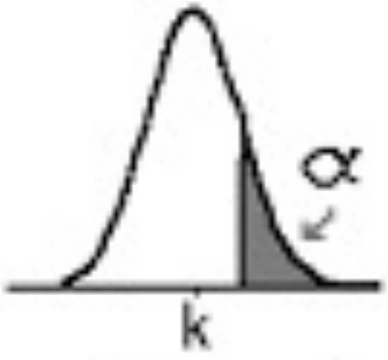
α	z critical
0.10	-1.28
0.05	-1.65
0.01	-2.33



$H_0: \mu = k$

$H_1: \mu \neq k$

α	z critical
0.10	± 1.65
0.05	± 1.96
0.01	± 2.58



$H_0: \mu = k$

$H_1: \mu > k$

α	z critical
0.10	1.28
0.05	1.65
0.01	2.33

Quantifying Chance

- Suppose the estimated z-score is z*.
- If z* is such that P[z<z*] <=0.5, then the chance of being different is 1-P[z<z*].
- If z* is such that P[z<z*] >0.5, then the chance of being different is P[z<z*].

A/B testing

- Statistical significance has been reached because the confidence is above 95%.
- There are more than 1000 views for each variation.

Variation	Conversion/ Views	Conversion Rate	Change	Confidence
A (control)	320/1064	30.08 ± 2.76%		
B (test)	250/1043	23.97 ± 2.59%	-20.30%	99.92%

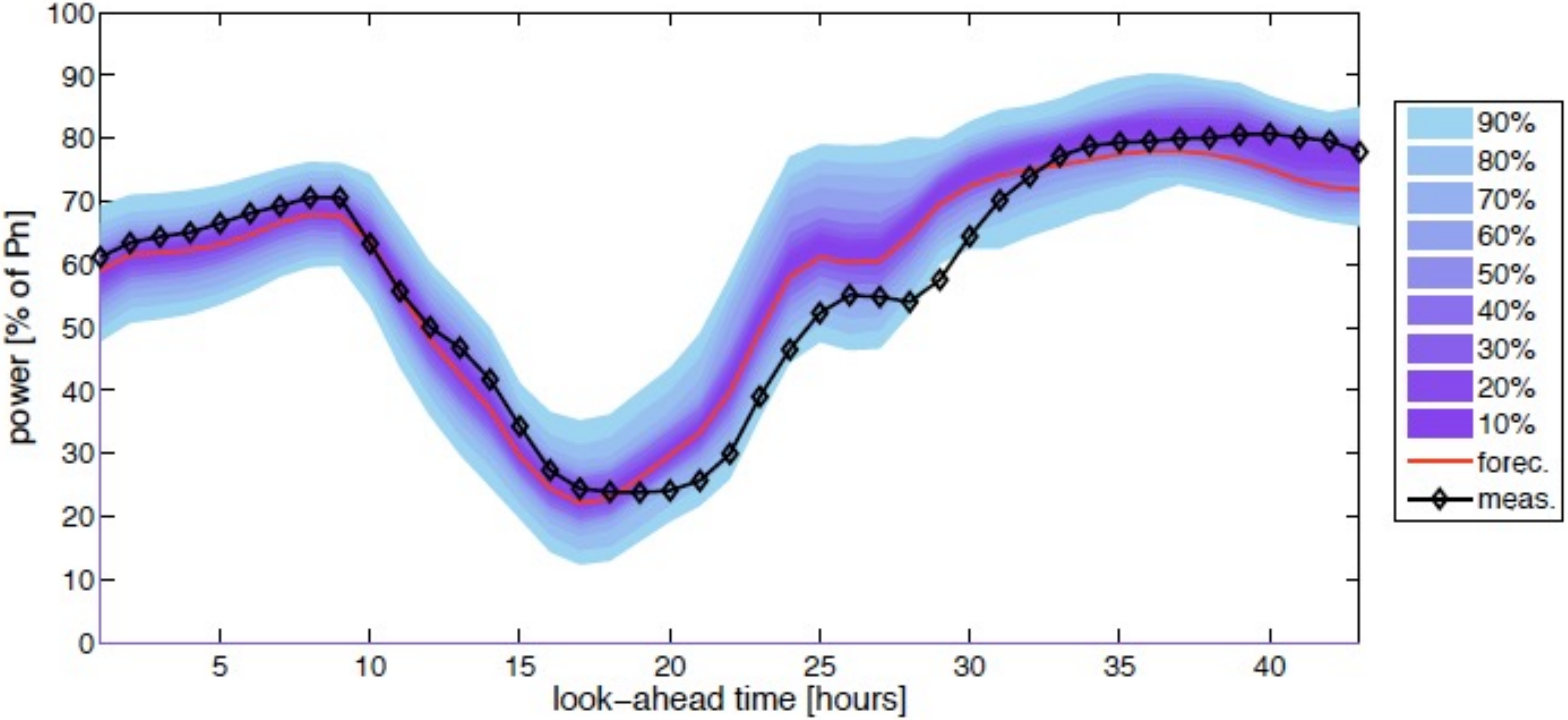
Communicating confidence

- Error bars provide a means of communicating the potential range of values that are likely to be associated with a particular estimate.
- Uncertainty is associated with a random variable and measured by a probability distribution.
- While centrality conveys the most likely outcome, we also need to explain the spread of the distribution.
- This helps to prevent surprises and to measure how often extremes are likely to occur.

Communicating uncertainty

- Wind power forecasting represents a substantial challenge due to the chaotic nature of the atmosphere.
- Power system operators depend on knowledge of wind power generation over the next three days.
- They know the forecasts are not always accurate.
- What is the best way to help them?

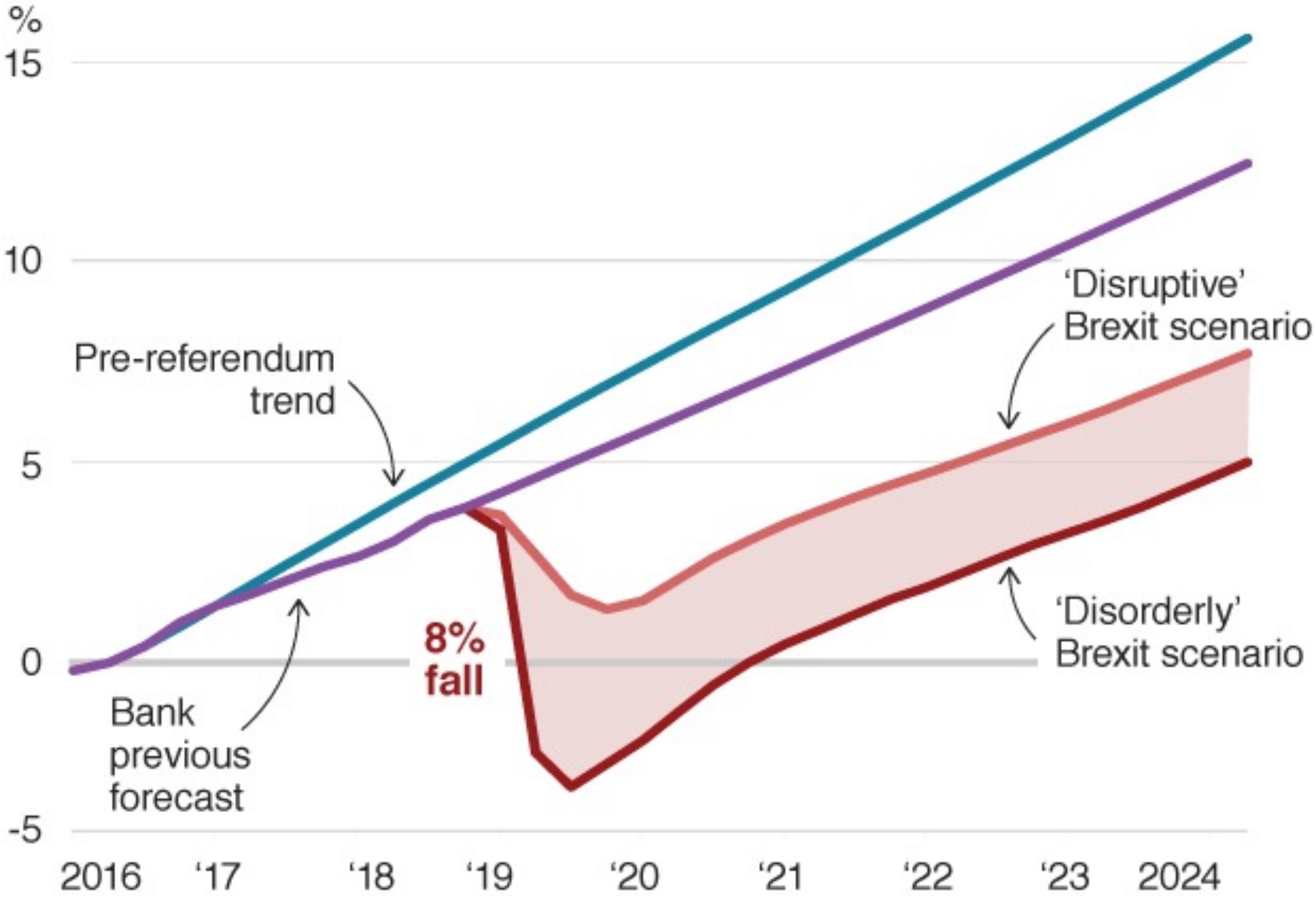
Probabilistic forecast of wind power in Denmark



Source: SafeWind, colaboration with Pierre Pinson

What does the Bank think could happen to UK growth?

UK GDP under different scenarios

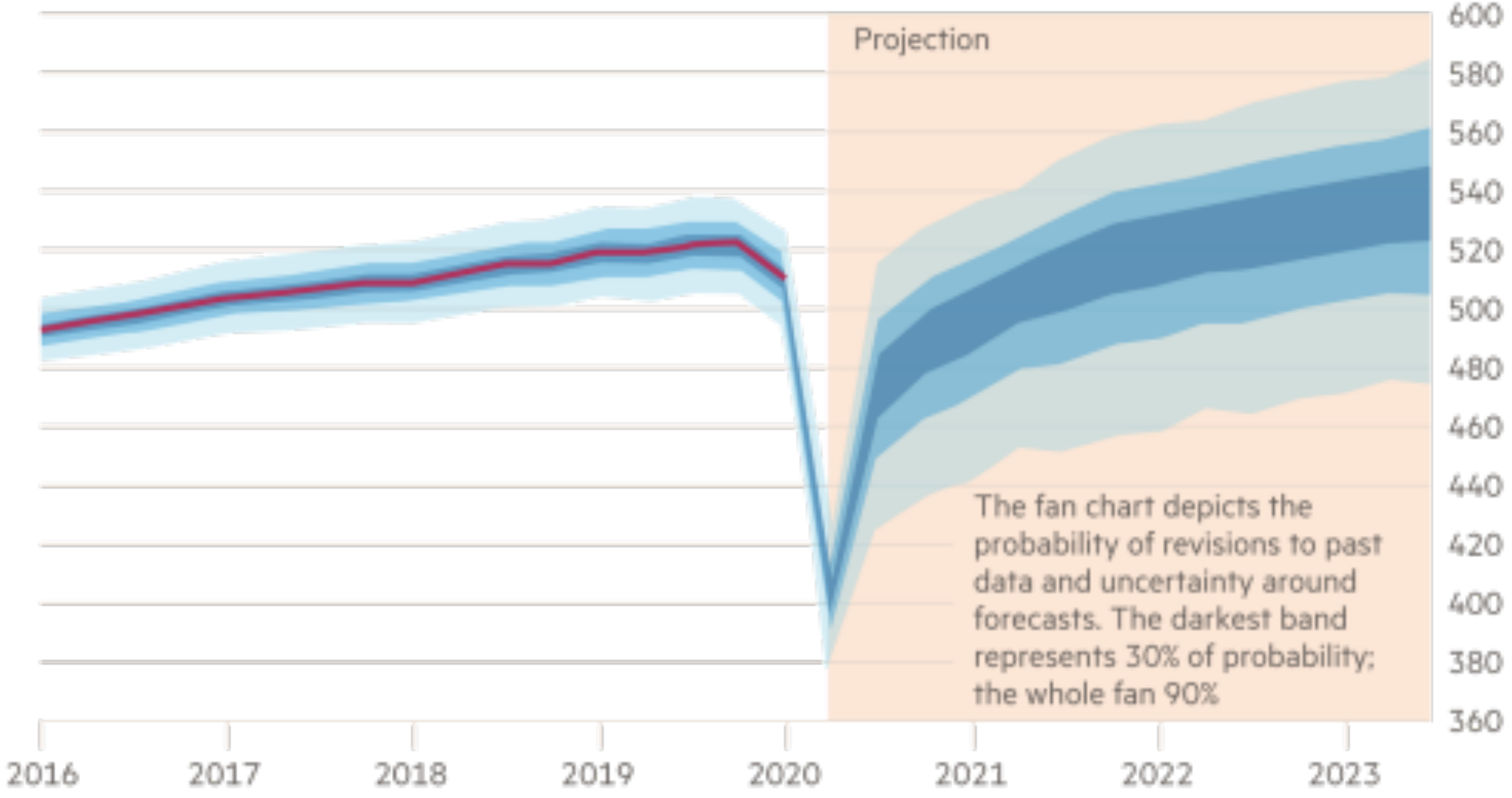


Source: Bank of England

BoE GDP fan chart

The MPC's unusually wide 'fan chart' reflects virus risks and committee splits

UK real GDP (quarterly values, £bn at 2016 constant prices)



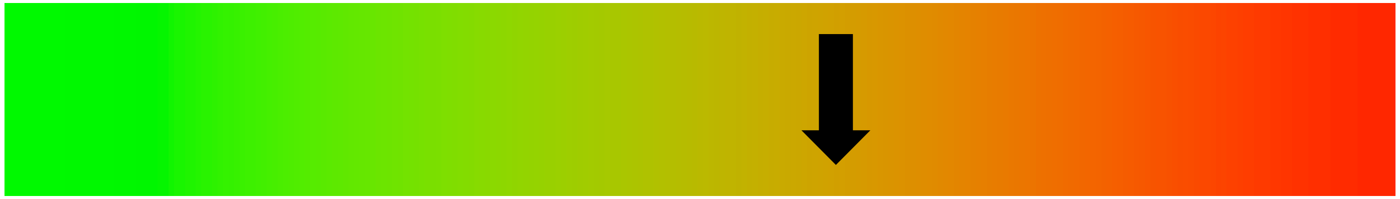
The fan chart depicts the probability of revisions to past data and uncertainty around forecasts. The darkest band represents 30% of probability; the whole fan 90%

Sources: ONS; Bank of England

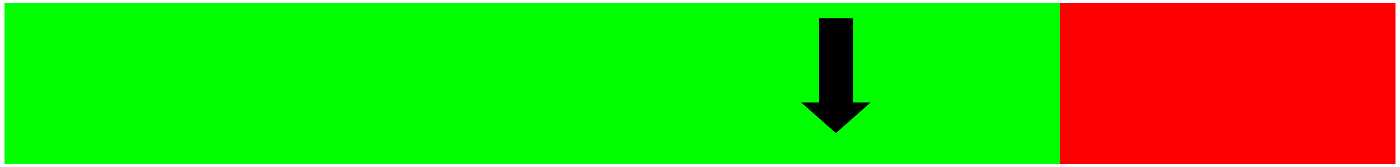
© FT

Risk communication

Continuous scale, with no pre-defined threshold:



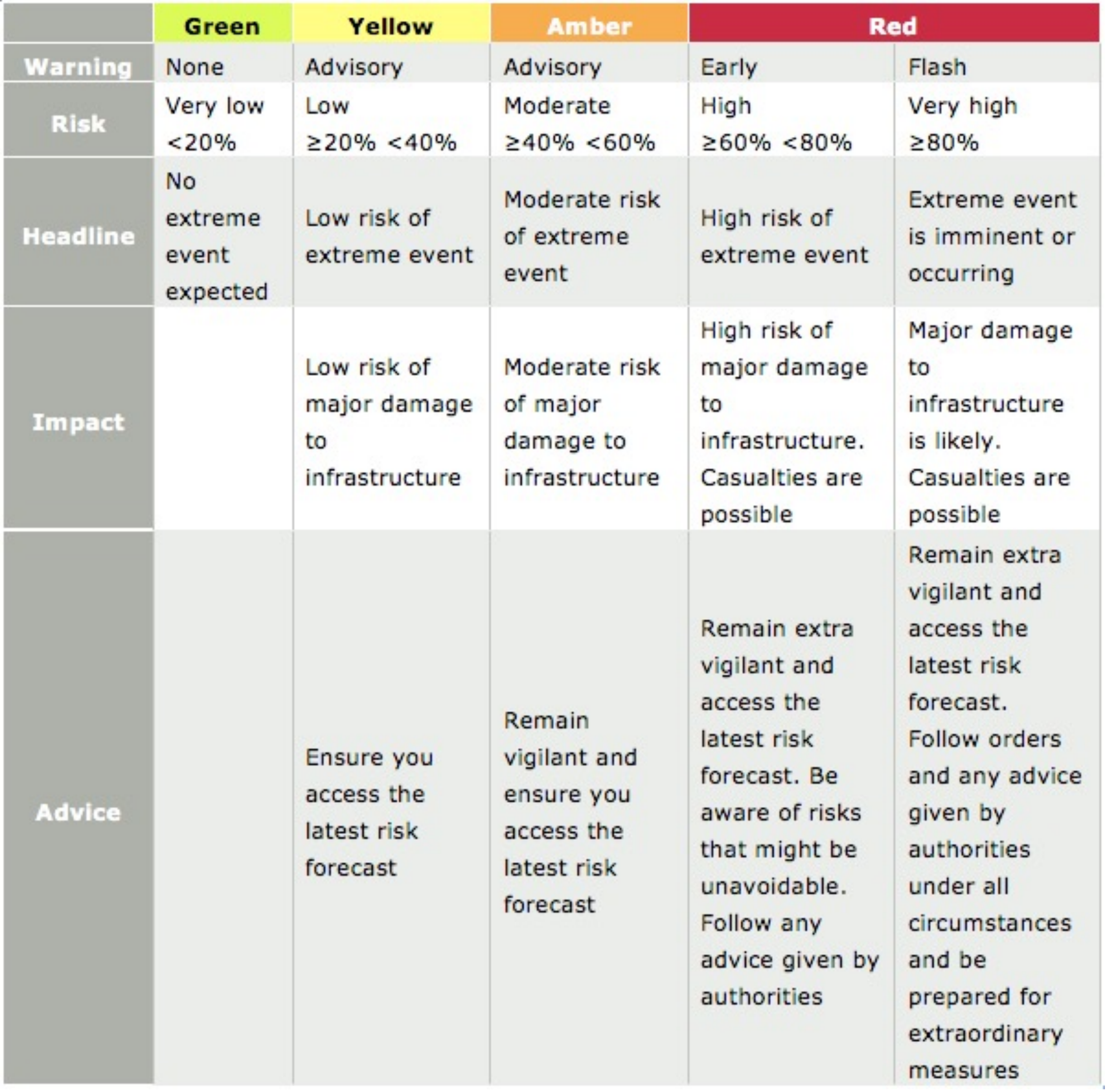
Quantised scale with pre-defined threshold:



Low risk

High risk

Probabilistic early warning



	Green	Yellow	Amber	Red	
Warning	None	Advisory	Advisory	Early	Flash
Risk	Very low <20%	Low ≥20% <40%	Moderate ≥40% <60%	High ≥60% <80%	Very high ≥80%
Headline	No extreme event expected	Low risk of extreme event	Moderate risk of extreme event	High risk of extreme event	Extreme event is imminent or occurring
Impact		Low risk of major damage to infrastructure	Moderate risk of major damage to infrastructure	High risk of major damage to infrastructure. Casualties are possible	Major damage to infrastructure is likely. Casualties are possible
Advice		Ensure you access the latest risk forecast	Remain vigilant and ensure you access the latest risk forecast	Remain extra vigilant and access the latest risk forecast. Be aware of risks that might be unavoidable. Follow any advice given by authorities	Remain extra vigilant and access the latest risk forecast. Follow orders and any advice given by authorities under all circumstances and be prepared for extraordinary measures

Following UK Met Office severe weather warnings.

Quantifying Confidence

- Simulating data
- Error bars
- Standard error
- Confidence intervals
- Prediction intervals

Matlab functions

- normrnd
- normcdf
- ttest2
- prctile
- fill

Q&A