Introduction to Artificial Intelligence

Week 7

What is Statistics?

- Statistics is the study of collecting, organizing, analyzing, and interpreting data in order to make decisions (Larson and Farber, 2006)
- Statistics is a measurement system for understanding randomness of variables and methods to compare randomness of systems

Applications

- Hard Sciences require statistics to understand the role of random events in the experimental process. There are often factors that are not controlled for or cannot be definitively measured and experiments control for this variation. Quite often these experiments use hypothesis testing, where two groups, one with treatment and one without (control group) are compared against each other for an observable difference.
- Statistics measures and quantifies the difference, and gives us an expectation of the likelihood that the results were due to the treatment and not due to chance

Applications

- Engineers work to control variance in processes and will work to identify and remove sources of variance to ensure a process puts out a quality product, without defect. Root cause analysis techniques and quality control, especially SixSigma, use statistics to guide process improvements
- The study of statistics is not only useful for the hard sciences and engineering. Studies of the soft sciences (economics, politics) often use polling data to understand the beliefs and preferences of consumers and voters

Descriptive v. Inferential Statistics

- Descriptive - organization, summarization, and display of data
- The mean temperature of Kazan over the year (Wikipedia)

Period	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Daily mean °C (°F)	−10.4 (13.3)	−10.2 (13.6)	−4.0 (24.8)	5.5 (41.9)	13.3 (55.9)	18.1 (64.6)	20.2 (68.4)	17.6 (63.7)	11.7 (53.1)	4.8 (40.6)	−3.4 (25.9)	−8.5 (16.7)	4.6 (40.3)

- Inferential - using a sample to make conclusions about a population
- The mean temperature in Kazan has been increasing over the last 30 years and it is likely we will see an increase every year from now on

Population v. Sample

- Often in statistical tests, we refer to the population and sample statistic
- The sample is any subset of the results seen in the full population, it is an approximation of the population
- The population is the measurement on the entire group that is of interest
- Imagine we are building car bumpers and wish to expect the number of defects which will occur. If one day we test one bumper of every 10 made (with 1000 produced), our sample is these 100 bumpers
- Is our population 1000 bumpers made that day?
- Our population is ALL the bumpers that was made and will be made on this production line

Data Collection Methods

- Observational Study – data is gathered by observing a part of the population in their normal actions
- Experimental Study – treatment is applied to a group (experimental group) and a second group is not given any treatment and observed (control). These groups are of a similar disposition to allow for a comparison between groups
- Simulation – use of computer, mathematical, or physical models to represent a situation. Usually done when data collection is dangerous or costly (e.g. crash testing cars)
- Survey – an investigation into the characteristics of a population by asking questions. This may be done by the census, asking everyone in the population, but more commonly done by taking samples, a count or measure of part of the population

Measures of Variation

- Mean – the average
  - Sample Mean – the average of a set of numbers x of cardinality N: $\overline{x} = \frac{\sum_i^N x_i}{N}$
  - Population mean μ
- Mode – the most frequent element
- Median – the middle of the ordered set
- Unbiased Sample Variance of a set of numbers x of cardinality N:
  - $s^2 = \frac{\sum_i^N (\bar{x} - x_i)^2}{N-1}$
- Corrected Sample Standard Deviation $s = \sqrt{s^2} = \sqrt{\frac{\sum_i^N (\bar{x} - x_i)^2}{N-1}}$

Hypothesis Testing

- Two actions, the null hypothesis is that things are equal or that there is no discernable difference and the alternative is that there is a difference
- H_o: $a = b; a \leq b; a \geq b.$
- H_a: $a \neq b; a > b; a < b.$
- We can either reject or fail to reject the null hypothesis
  - We can accept the alternative
  - WE NEVER ACCEPT THE NULL! YOU CAN NEVER PROVE THE NULL HYPOTHESIS! YOU CANNOT PROVE THAT TWO DISTRIBUTIONS ARE THE SAME WITH A SAMPLE BASED TEST
  - If there is one difference they are not the same – meaning you would have to look at the entire population

Example Hypothesis

- Two evolutionary algorithms (X, Y) were run on the same problem 30 times. We wish to test if there is a difference in the mean fitness value of the final generation.

Formulate the Null Hypothesis

- $H_o: \mu_x = \mu_y$
- There is no significant difference in the mean fitness value between the two algorithms

Formulate the Alternative Hypothesis

- $H_a: \mu_x \neq \mu_y$
- There is a significant difference in the mean fitness value between algorithm X and Y

Confidence

- A measure of the acceptance of Type I errors:
  - Rejecting the Null hypothesis when in reality we should not have rejected
  - Saying two things are different when they are in fact the same
  - Accusing an innocent man!
  - Saying someone has HIV when in reality they do not!
  - False Positive
- Confidence is measured by an $\alpha - value$ with 95% being the most common value in scientific literature. Most statistical tests aim to control confidence by a variable in the test

Power

- A measure of the acceptance of Type II errors:
  - Failing to reject the Null hypothesis when in reality we should have rejected
  - Saying two things are the same when they are in fact they are different
  - Acquitting a guilty man!
  - Saying someone does not have HIV when in reality they do!
  - False Negative
- Power is measured by the $\beta - value$ with most scientific measures ignoring it. The majority of statistical tests do not control for this, though increasing the sample size increases power

Student-T test

- Historically, was developed by a chemist, W. S. Gosset, for the Guinness Brewery, Dublin, Ireland which would not allow it to be published. The Guinness plant hired some of what were then the best graduates of Oxford and Cambridge. The Company was well known also for their policy of "study-leaves" for technical staff
- Was published under a pen name "Student"
- Can easily see why in quality control we want to use a small number of samples and not a z-test. We would have some very 'happy' quality control officers in the middle of our brewery

Test Against a Known Mean

- t-test for a target mean (μ):
- t value = $\frac{\bar{x}-\mu}{s/\sqrt{N}}$.
- Check against t-Table, with degrees of freedom = N-1.

Test Between Two Sample Means

- t-test between two sample means:
- t value = $\frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{s_{\bar{x}_1-\bar{x}_2}}$, where $s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}$
- Check against t-Table with degrees of freedom as the smaller of $N_1 - 1$ and $N_2 - 1$

P-value

- Often we will want to figure the probability of making a Type I error exactly! The test is only ensuring it is meeting with set confidence. For this, we compute a P-value of the test
- We say that tests which have a:
  - P-value < 0.05 are "statistically significant" -> $\alpha > .95$ or 95% confidence
  - P-value < 0.01 are "extremely (very) statistically significant" -> $\alpha > .99$ or 99% confidence
- Note again: Statistical significance only measures that their difference is not caused by chance. It might be a very tiny effect, e.g. statistical significance vs. clinical significance of being exposed to a weak carcinogen

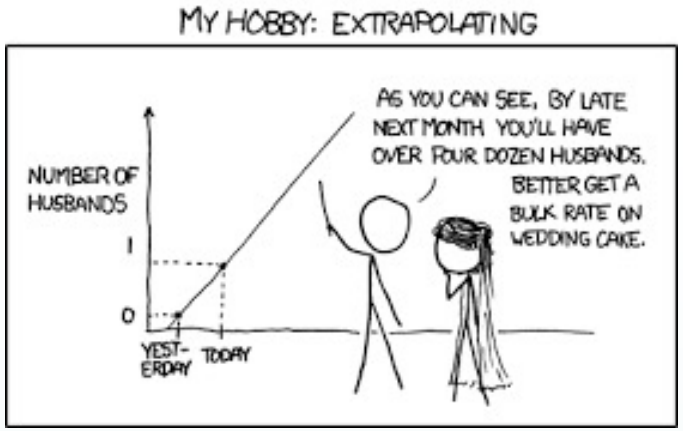
Confidence Interval

- Visualization of where the mean will be, upper gate and lower gate
- $\bar{x} - t_c \frac{s}{\sqrt{N}} < \mu < \bar{x} + t_c \frac{s}{\sqrt{N}}$
- 95% CI on a mean with n > 30
- $\bar{x} - 1.96 \frac{s}{\sqrt{N}} < \mu < \bar{x} + 1.96 \frac{s}{\sqrt{N}}$
- If Confidence Intervals do not overlap, we can be sure that a t-test would find it is significant at the X-confidence level
- If they do overlap, we do not necessarily know if we cannot reject the null hypothesis – for that we need to examine the rejection gates which are smaller than the CI

Look-Elsewhere Effect

- P-value of .05 means that 1/20 times we have demonstrated a difference
- If we do multiple tests with the same data against a group of other, non-independent trials such as checking multiple, we have a higher likelihood of seeing a Type I error (demonstrating a difference when no difference exists) across the group. This is the family error rate of the group of hypothesises, which we would like to reduce
- When comparing groups of hypothesis, this issue is controlled by increasing the sample sizes and applying a corrective measure – such as Bonferroni correction. Note, this reduces the POWER of the test, which is increased by razing the number of samples

Lies, Damn Lies, Statistics



- Learning statistics protects you from abuses of data. Politicians, activists, advertisers, sales people, and anyone wanting to convince you to do something will use data in a way that benefits their arguments – they are not aiming to be objective to the data.
- "One half of all students in our country are below the average math score, we need more funding for Education" – 'Expert' on a Canadian News Magazine show
- Learning statistics makes you more defensible to such tactics – it is a brain anti-virus for bad data
- "Our brand of cereal has 1/3 the fat of the next leading brand"
  - How much fat was there before? 1/3 of a small amount is not that much of an improvement
  - How much sugar did you add?
- "In the last year our consumer base has quadrupled – you can't pass up such an investment"
  - So you went from having one customer to four?

"My teen watches porn – Parentchannel.tv"

- What message is being given:
- "My teen watches porn" "57% of teens have watched porn online"
- Is 9-19 a good sample group? How many were sampled of each age?
  - 100% of their sampled 18-19 year olds could be porn watchers
  - 0% of 9-17 year olds have seen any
- Are parents going to be worried about an ADULT(18-19) seeing pornography?
- What is the definition of SEEN?
  - Actively watched or was seeking for it
  - Have had an advertisement pop-up
- What is the definition of PORN?
  - Does a love seen in a Hollywood movie count?
  - Does classical paintings, or nudity in art count?
- How was the data collected?
  - Was there a reporting bias?

Have to ask how a study was conducted

- Did it collect sensitive details anonymously?
  - People will LIE. Some to reduce (don't want to get in to trouble), some to inflate (maybe some 18-19 year olds trying to look cool to peers)
- Have to collect answers in a way which does not lead to a conclusion, and that adequately accounts for the possible beliefs or actions. So, we could correct the phrasing of our poll by asking:
  - In the last month how many times have you intentionally viewed an image or video in which the plot revolved about sexual actions including but not limited to sexual intercourse (i.e. pornographic materials)
- Have to ensure a sample which is REPRESENTATIVE of your population, and preferably one which is randomly selected
  - In a poll of 9-17 year olds, with 30 members in each age group, asked the previous question. 57% of them responded having at least one instance or more. The following is a breakdown by age group…
  - Making a poll of Farmers on how best to utilize land does not tell you how someone in a city will wish it used