NAKED STATISTICS

H1 - WHAT IS THE POINT

The point of statistics is to:

- summarize large data sets
- improve decision making
- answer important social questions
- recognize patterns
- detect anomalies
- assess the quality of policies, medical procedures, etc.

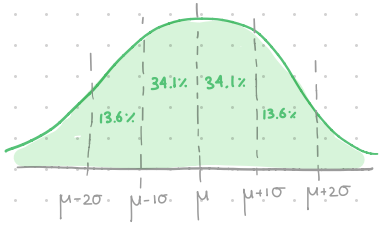
H2 - DESCRIPTIVE STATISTICS

Descriptive statistics organize complex information into a single number.
→ meaningfull summary of underlying phenomenon
→ can be very misleading due to over simplification

Absolute numbers can usually be interpreted without any context. → 3 goals
Relative numbers have meaning only in comparison to something else → 10th percentile

Standard deviation (σ) is a measure of how dispersed the data are from their mean
Variance is similar (σ^2) and puts more emphasizes on the outliers.
→ knowing the mean and std. provides much insight into the data!

NORMAL DISTRIBUTION



$1\sigma = 68.2\%$
$2\sigma = 95.4\%$
$3\sigma = 99.7\%$

PERCENTAGE DIFFERENCE = change

$$\frac{(\text{new} - \text{old})}{\text{old}}$$

Percentage change ≠ change in percentage points
↳ relative change ↳ absolute change

H3 - DECEPTIVE DESCRIPTION

Precision vs. Accuracy

- Precision = the exactitude whit which we can express something
  → 41.6 km → around 40 km → a long way
- Accuracy = whether what you measure is what you're trying to measure
  → no amount of precision can make up for inaccuracy

NAKED STATISTICS

H4 - CORRELATION

Correlation: the degree to which two phenomena are related to one another.

correlation coefficient is a number ranging from -1 to 1

→ correlation does not imply causation!

1 or -1 : every change in one variable is associated with an equivalent change in the other variable in in the same direction (1) or opposite (-1)

0 : variables have no meaningful association with one another

H5 - BASIC PROBABILITY

Binomial experiment (Bernoulli trial)

- fixed number of trials (e.g. 100 taste tests)
- each trial has 2 possible outcomes (classification)
- probability of success is the same in each trial
- assumption: trials are independent

Probability is the study of events and outcomes involving an element of uncertainty.
If you flip a coin 4 times in a row, you can never know the outcome in advance with certainty. Yet, you can determine in advance that some outcomes (two heads - two tails) are more likely than others.

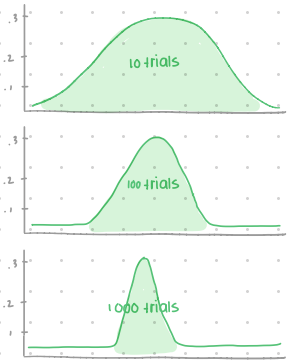
The probability of event A AND event B happening is prob. A * prob. B → no dependence
The probability of event A OR event B happening is prob. A + prob. B

Probability allows you to calculate the **expected value**. That is the sum of all different outcomes, each weighted by its probability and pay-off.

Example:
- game in which you roll a single dice
- pay-off 1 - $1, 2 - $2, 3 - $3, etc.
- probability: 1/6 for each possible outcome
- expected value: $1/6(\$1) + 1/6(\$2) + \ldots + 1/6(\$6) = \$3.50$

→ suppose you have to pay $3 to play the game, then the expected value is higher than the cost of playing. Thus, it's worth taking the risk.

LAW OF LARGE NUMBERS



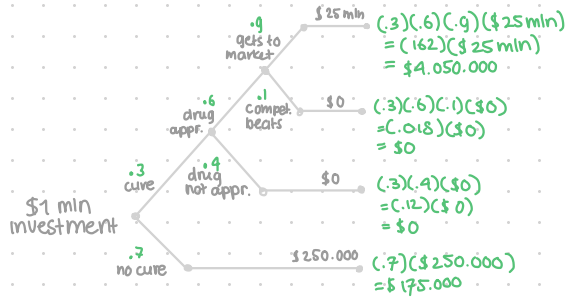
→ as the number of trials increases, the average outcome will get closer and closer to the expected value

→ this is how casino's and insurance companies earn money

NAKED STATISTICS

The expected value can help untangle complex decisions with many contingencies at different points of time. → decision trees

Expected pay-off:
$4.050.000 + $0 + $0 + $175.000
= $4.225.000



H5 ½ - MONTHY HALL PROBLEM

- 3 doors, 1 prize, 2 goats
- chose one door (e.g. door B)
- one of the other doors is opened, always revealing a goat (e.g. door C)
- do you want to change to door A?
  → yes, you should always change because:
  - stick with choice: 1/3 chance of winning
  - if you change, you technically choose two doors, and from those two you will always get the correct door, this increases chance of winning to 2/3

H6 - PROBLEMS WITH PROBABILITY

1. Assuming events are independent when they are not
   → e.g. engine failures of an airplane

2. Not understanding when events are independent
   → gambler's fallacy: mistaken belief that a certain random event is (more) likely to occur given previous event(s)
   → hot hand fallacy: mistaken belief that a person who experiences a successful outcome has a greater success in further attempts

3. Clusters happen: given enough repitition, unlikely events become more and more likely

4. Prosecutor's fallacy: when the context around statistical evidence is neglected. The chances of finding a coindidental one in a million match are relatively high if you run a sample through a database of a million records.

5. Reversion to the mean (or regression to the mean): any outlier is likely to be followed by outcomes that are more consistent with the long-term average given enough repeated samples (law of large numbers)

6. Statistical discrimination: statistics often contain social complications, it is wrong to unreasonably discriminate against someone on the basis of irrelevant factors using statistics.

NAKED STATISTICS

H7 - IMPORTANCE OF DATA

1. No amount of fancy analysis can make up for fundamentally flawed data
   → garbage in, garbage out

2. Data should be representative of some larger group or population.
   → bigger is better (smoothing away errors) BUT bigger bad sample is worse than bad small sample as it gives a false sense of confidence.

3. Data must have a source of comparison → control vs. treatmen group
   → it can be very difficult to isolate the impact of one specific attribute, therefore, randomization should be applied. Doing so, you assume that it will evenly divide all relevant characteristics between the two groups (both the things you can observe, as well as those you cannot measure.

Common examples of bad data / bias:

1. Selection bias
   → if each member of the population does not have an equal chance of ending up in the sample, e.g. those who are willing to answer a survey in a public place are different from those who prefer not to be bothered.
   → self-selection bias: whenever individuals volunteer to be tested.

2. Publication bias
   → positive findings are more likely to be published than negative findings.
   → unusual things happen once in a while (chance). If you conduct 100 studies, one is likely to turn up results that are nonsense. However, this one study that does find a link is published.

3. Recall bias
   → data that contains information that people had to recall (cross-sectional) are inherently less reliable than when this information was collected at that point in time (longitudinal)

4. Survivorship bias
   → when some or many of the observations fall out of the sample, changing the composition of the observations

5. Healthy user bias
   → people who faithfully engage in activities that are good for them - taking a drug as prescribed, or eating what they believe is a healthy diet - are fundamentally different from those who don't.

NAKED STATISTICS

H8 - CENTRAL LIMIT THEOREM

The Central Limit Theorem (CLT) states that a large properly drawn sample will resemble population from which it is drawn. The probability that any sample will deviate massively from the underlying population is low.

CLT allows us to make inferences:

- → make inferences about a sample from a population
  e.g. when only a sample of students is tested, the average test score will not deviate much from the average test score of the whole school
- → make inferences about a population from a sample
  e.g. the person testing only the sample of students can be reasonably certain that he gets a grasp of the performance of all students
- → the sample means will be distributed roughly as a normal distribution around the population mean, no matter what the distribution of the underlying population looks like.
  - → the larger the number of samples, the more closely it will approximate a normal distribution
  - → the larger the sample sizes, the tighter the distribution will be.

RULE OF THUMB
For CLT to hold, the sample size must be at least 30

Standard error measures the dispersion of the sample means.

- → Standard deviation: dispersion in population (σ)
- → standard error: dispersion in sample means (SE)

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

Because sample means are distributed normally:

- 68% of all sample means lie within 1 SE of population mean
  - → 95% - 2 SE, 99.7% - 3 SE.

Example: population mean = 162, St.dev. = 36, sample mean = 194, sample size = 62.

- SE of sample: $5 / \sqrt{62} = 36./7.9 = 4.6$
- difference sample mean and population mean is 32.
- that is way more than 3 st.errors. (99.7%).
- it is extremely unlikely that the sample is drawn from the population

Most sample means will lie reasonably close to the population mean, the SE is what defines "reasonably close".

CLT tells us that the probability that a sample mean will lie within a certain distance of the population mean. It is relatively unlikely that a sample mean will lie more than two standard errors away from the population mean, etc.

NAKED STATISTICS

H9 - INFERENCE

Statistical inference cannot prove anything with certainty, it can observe some pattern or outcome and then use probability to determine the most likely explanation for that outcome.

Hypothesis testing:

→ Define null hypothesis → e.g. there is no difference.
& alt. hypothesis → e.g there is a difference.
→ alt.hyp. is a conclusion that must be true if null hyp. can be rejected
→ you are aiming to reject the null hypothesis → "success"
→ if you can't reject the null hypothesis, you typically say that you "failed to reject it"

RULE OF THUMB
Threshold for rejecting null hypothesis is 5% → signif. level (or 1% or 10%)

→ **Significance level:** the upper bound for the likelihood of observing some pattern of data if the null hypothesis were true.
→ **P-value:** the specific probability of getting a result at least as extreme as the one you've observed if the null hypothesis were true or likelyhood of observed difference in means.

Confidence interval
→ With 95% confidence the value is between two standard errors of the found value.

Type I vs. type II errors
→ I: wrongly rejecting a true null hypothesis (false positive)
→ II: failing to reject a false null hypothesis (false negative)

H10 - POLLING

The difference between a poll and other forms of sampling is that the interest is not in the mean, but rather the percentage or proportion.

Standard error for proportion: $\sqrt{p(1-p)/n}$
→ the standard error tends to be smaller when p and $(1-p)$ are far apart
→ when calculating with different proportions, usually the largest SE is used for all proportions
→ or assuming $p=.5$, generating the largest SE for that sample

The larger the sample size, the smaller the error.

H11 - REGRESSION ANALYSIS

Regression analysis allows you to quantify the relationship between a particular variable and an outcome that you care about while controlling for other variables.

NAKED STATISTICS

The regression algorithm seeks to find the "best fit" for a linear relationship between two variables
→ fitting a regression line using ordinary least squares (OLS) → $y = ax + b$

Dependent variable: variable that is being explained
Explanatory variables: variables used to explain dep. var.
↳ independent / control variables

RULE OF THUMB
significant if coefficient is at least twice the size of the standard error

R^2 is a measure for the total amount of variation explained by the regression equation.
→ $R^2 = 0$ means that the equation does no better than the mean
→ $R^2 = 1$ means that the equation perfectly predicts the value.

T-distribution if sample is $n < 30$ → looks like the normal distribution with fatter tails → when the number of degrees of freedom gets large, the t-distribution converges to the normal distribution.

HIZ - COMMON REGRESSION MISTAKES

1. Use regression to analyze non-linear relationship.
2. Correlation does not equal causation
3. **Reverse causality** → don't use expl.var. that migh be affected by the dep. variable
4. **Omitted variable bias** → when you leave out a crucial explanatory variable, another variable might "pick-up" the effect.
5. **Highly correlated explanatory variables** (multicollinearity) → use either one or create one combined variable.
6. **Extrapolating beyond the data** → results are valid only for a population that is similar to the sample on which the analysis is performed.
7. **Data mining** → too many variables → with too many variables, one is bound to meet the threshold for significance just by chance.