Recherche Zen
Séance 4 : Analyses

Carlos Ramisch and Manon Scholivet
Partly based on the course by Adeline Paiement

03 avril 2023

Expectation. . .

	dataset	metric1	metric2	metric3 [1]
SOTA model	DS1	82.3	75.9	48.0
Our model	DS1	**95.3**	**89.8**	**65.4**
SOTA model	DS2	67.7	65.2	56.8
Our model	DS2	**80.3**	**91.1**	**69.8**
SOTA model	DS3	77.6	74.1	92.8
Our model	DS3	**84.9**	**78.3**	**98.1**

1. Higher is better

Expectation. . .

	dataset	metric1	metric2	metric3 [1]
SOTA model	DS1	82.3	75.9	48.0
Our model	DS1	**95.3**	**89.8**	**65.4**
SOTA model	DS2	67.7	65.2	56.8
Our model	DS2	**80.3**	**91.1**	**69.8**
SOTA model	DS3	77.6	74.1	92.8
Our model	DS3	**84.9**	**78.3**	**98.1**

$\Longrightarrow$ Our model is better than state of the art ! 🎉

1. Higher is better

. . . Vs. reality !

	dataset	metric1	metric2	metric3
SOTA model	DS1	**82.3**	75.9	48.0
Our model	DS1	80.7	**76.2**	**50.4**
SOTA model	DS2	67.7	**65.2**	**56.8**
Our model	DS2	**67.9**	nan	49.6
SOTA model	DS3	77.6	**74.1**	92.8
Our model	DS3	**79.0**	**74.1**	**93.4**

. . . Vs. reality !

	dataset	metric1	metric2	metric3
SOTA model	DS1	**82.3**	75.9	48.0
Our model	DS1	80.7	**76.2**	**50.4**
SOTA model	DS2	67.7	**65.2**	**56.8**
Our model	DS2	**67.9**	nan	49.6
SOTA model	DS3	77.6	**74.1**	92.8
Our model	DS3	**79.0**	**74.1**	**93.4**

$\Longrightarrow$ Wake up and smell the coffee

Results analysis

- Identify overall trends
- Identify potential sources of problems (or bugs)
- Ensure conclusions are valid, claims are (statistically) sound

Experimental results

- Diversity of experiments $\implies$ diversity of results
  - $\rightarrow$ Task at hand
  - $\rightarrow$ Datasets
  - $\rightarrow$ Evaluation metrics
  - $\rightarrow$ ...
- This course : no silver bullet, rather a toolbox
- Focus on examples

Statistics

- A mathematical framework to analyse data
- Solid foundations : probability theory

  $\rightarrow$ Statistics = data + probability theory
- Statistical inference $\Longrightarrow$ data science, machine learning

  $\rightarrow$ Also : finances, health, biology, physics, social sciences, ...
- Identify trends, check hypotheses, measure correlations, ...

The problem with statistics

Finding good learning materials in statistics is hard

Too applied :

Too theoretical :

Weak Law of Large Numbers

The weak law of large numbers (cf. the strong law of large numbers) is a result in probability theory also known as Bernoulli's theorem. Let $X_1, ..., X_n$ be a sequence of independent and identically distributed random variables, each having a mean $\langle X_i \rangle = \mu$ and standard deviati σ. Define a new variable

$$X \equiv \frac{X_1 + \ldots + X_n}{n}.$$

Then, as $n \to \infty$, the sample mean $\langle x \rangle$ equals the population mean μ of each variable.

$$\begin{aligned}\langle X \rangle &= \left\langle \frac{X_1 + \ldots + X_n}{n} \right\rangle \\ &= \frac{1}{n}\left(\langle X_1 \rangle + \ldots + \langle X_n \rangle\right) \\ &= \frac{n\,\mu}{n} \\ &= \mu.\end{aligned}$$

In addition,

$$\begin{aligned}\operatorname{var}(X) &= \operatorname{var}\left(\frac{X_1 + \ldots + X_n}{n}\right) \\ &= \operatorname{var}\left(\frac{X_1}{n}\right) + \ldots + \operatorname{var}\left(\frac{X_n}{n}\right) \\ &= \frac{\sigma^2}{n^2} + \ldots + \frac{\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Therefore, by the Chebyshev inequality, for all $\epsilon > 0$,

$$P(|X - \mu| \geq \epsilon) \leq \frac{\operatorname{var}(X)}{\epsilon^2} = \frac{\sigma^2}{n\,\epsilon^2}.$$

What usually happens

- A given statistical tool is used without (full) understanding
- Statistical tools applied because supervisor/reviewer asked
- Give up trying to understand, just use it as a blackbox

From scratch : random variables i

- Experiment : flip 3 different coins, note head (H) or tail (T)
- The sample space S contains all possible experiment outcomes

  $\rightarrow$ The subsets of S are called events E_i
- The random variable X denots the number of heads (H)
  - A variable whose exact value is unknown or irrelevant
  - We know (or estimate) its probability distribution $P\{X = x_i\}$

E_i	$\{HHH\}$	$\{THH, HTH, HHT\}$	$\{TTH, THT, HTT\}$	$\{TTT\}$
$P(E_i)$	1/8	1/8 + 1/8 + 1/8	1/8 + 1/8 + 1/8	1/8
X	0	1	2	3
$P\{X = x_i\}$	1/8	3/8	3/8	1/8

From scratch : random variables ii

Formalisation

A random variable is a function $X : S \to \mathbb{R}$ such that :

1. Discrete random variable :
   - $\rightarrow$ Its set of possible values $X(S) = \{x_i, i \in \mathbb{N}^*\}$ is countable
   - $\rightarrow$ For all $x_i \in X(S)$: $\{X = x_i\} \Leftrightarrow \{e_i \in S | X(e_i) = x_i\} \in \mathcal{F}$
   - $\rightarrow$ $\mathcal{F}$ is the set of all possible events (subsets) of S
   - $\rightarrow$ $p(x_i) = P\{X = x_i\}$ is the probability mass function of X
2. Continuous random variable :
   - $\rightarrow$ $\forall$ value $x \in (-\infty, +\infty)$, $\forall$ interval $B \in \mathbb{R}$
   - $\rightarrow$ A non-negative function $P\{X \in B\} = \int_B f(x)\,dx$ exists
   - $\rightarrow$ $f(x)$ is the probability density function of X

Independence assumptions

- Data items $X_1 \dots X_n$ can be seen as n random variables
- We assume that all items come from the same distribution
- We assume that all items are independent, that is :

  $\rightarrow \forall X_i \neq X_j, \forall a, b \in X_i(S) \;\; P\{X_i = a | X_j = b\} = P\{X_i = a\}$
- This is often stated as independent and identically distributed

  $\rightarrow$ The acronym i.i.d. is usually employed

Expected value, mean, law of large numbers

- The expected value of a discrete random variable :

$$E[X] = p(x_1)x_1 + p(x_2)x_2 + \ldots = \sum_{x_i \in X(S)} p(x_i)x_i$$

- The arithmetic mean of a collection of i.i.d. items $x_1 \ldots x_n$:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- The law of large numbers states that $\overline{x} \to E[X]$ for large n

  $\rightarrow$ The (sample) mean $\overline{x}$ is an estimator of the expected value $E[X]$

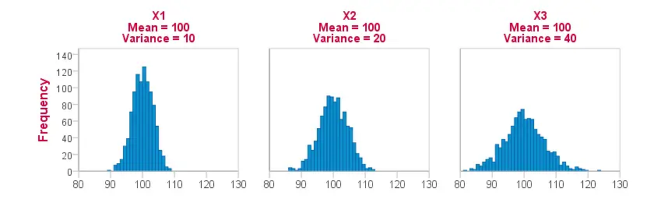
  $\rightarrow$ The mean summarise the distribution in a single value

Variance, standard deviation i

- Variance characterises the dispersion/spread of a distribution
  - → Intuition : average distance from the expected value
  - → $x_i - \overline{x}$ can be positive or negative $\Longrightarrow$ square it !

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

→ Variance is always positive, expected value not necessarily

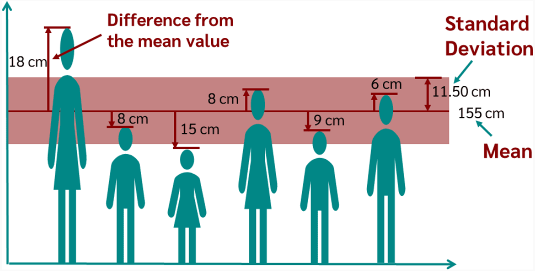


https://www.spss-tutorials.com/descriptive-statistics-one-metric-variable/

Variance, standard deviation ii

- Variance averages *squared* differences
  - $\rightarrow$ Its absolute value is hard to interpret
  - $\rightarrow$ Bring back to original value range $\rightarrow$ squared root
- The squared root of variance is called standard deviation

$$\sigma = \sqrt{Var(X)}$$



https://datatab.net/tutorial/dispersion-parameter

Variance, standard deviation iii

- Variance for known probability distribution :

$$Var(X) = E[(X - E[X])^2] = \sum_{x_i \in X(S)} (x_i - \overline{x})^2 p(x_i)$$

- Population variance estimator :

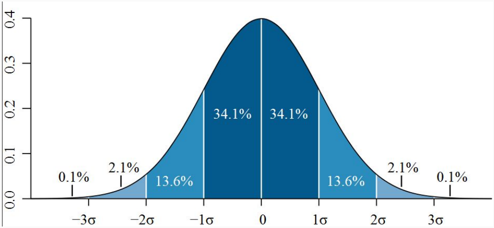
$$Var(X) = E[(X - E[X])^2] = \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{n} \qquad \sigma_X = \sqrt{Var(X)}$$

- Sample variance, unbiased estimator :

$$Var(X) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} \qquad s_X = \sqrt{Var(X)}$$

Normal distribution

- Well known distribution for continuous random variables
- Probability density function is a Gaussian bell-shaped curve
- Characterised by $E[X] = \mu$ and σ parameters
- Can be used to approximate binomial distribution for large n



Central limit theorem

- A properly normalised sum of i.i.d. random variables is normally distributed

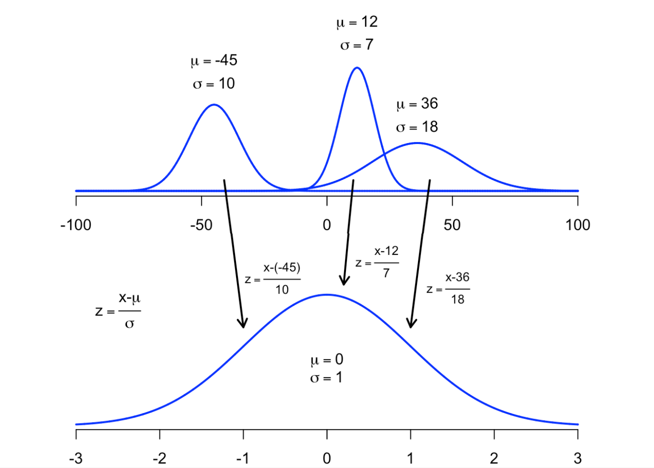
  $\longrightarrow$ Even if the variables are not normally distributed !

- The mean of i.i.d. random variables is normally distributed

  $\longrightarrow$ Comes in handy to analyse metrics when they are means

Standardization

- Normal is hard to integrate analytically
  - $\rightarrow$ Standardize $z = \frac{x-\mu}{\sigma}$
  - $\rightarrow$ Use cumulative function table $\Phi(a)$



Plan

Example : compositionality

- *Is a dry run litteraly a run which is dry ?*

  → not at all ←0 - 1 - 2 - 3 - 4 - 5 → absolutely yes
- Compositionality : average over 10-15 annotators
- Datasets : 180 compounds for English, French, Portuguese

  → `https://aclanthology.org/J19-1001/`

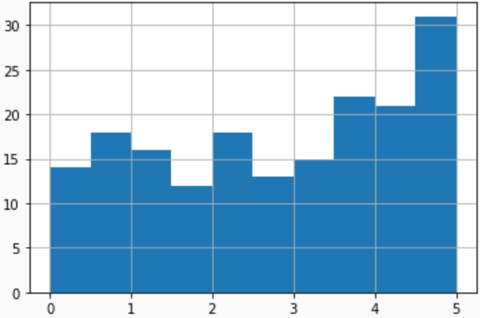
Compositionality of compounds

	compound_lemma	compositionality
134	poule_mouillé	0.0000
127	pied_noir	0.1333
19	carte_blanc	0.2000
151	septième_ciel	0.2143
15	bouc_émissaire	0.2308
...	...	...
0	activité_physique	4.9333
55	eau_potable	5.0000
170	téléphone_portable	5.0000
96	matière_gras	5.0000
52	eau_chaud	5.0000

180 rows × 2 columns

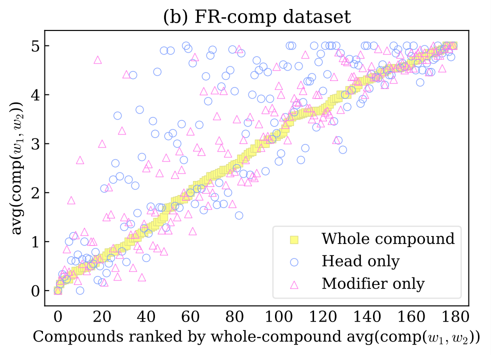
Simple descriptive statistics

count	180.000000
mean	2.770321
std	1.505560
min	0.000000
max	5.000000



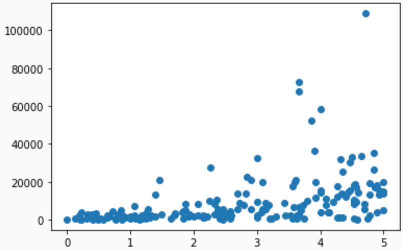
Two variables : scatter plot

- Variable X on x-axis, variable Y on y-axis
- `plt.scatter(x,y)`
- Linear regression can help visualise association



Example : compositionality and frequency

- Hypothesis : frequent compounds are judged less compositional
- How much variation in compositionality can be "accounted for" by variation in frequency ?
- Relation between two real-valued random variables



Covariance

- Covariance is the normalized product of centered values [2]

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

  $\rightarrow$ Both differences are positive or negative : product is positive

  $\rightarrow$ Both vary in opposite directions : product is negative

- Expected value of the product of (centered) variables

  $\rightarrow Cov(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$

- What if X and Y have very different ranges ?

  $\rightarrow$ Covariance is unbounded - ranges from $-\infty$ to $+\infty$

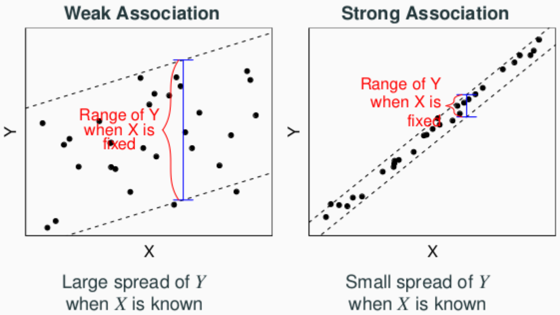
- Indicates whether a linear relation exists, but not its strength

2. Use n in denominator for population covariance

Pearson's linear correlation (r)

- Covariance normalised by individual variances

$$r_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Cov(X,Y)}{s_X s_Y}$$

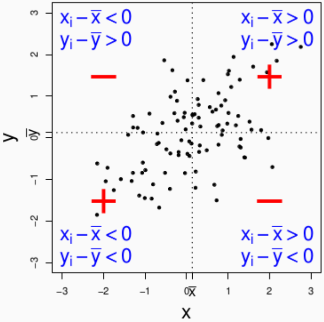


https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation and standarisation

$$r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X s_Y}$$

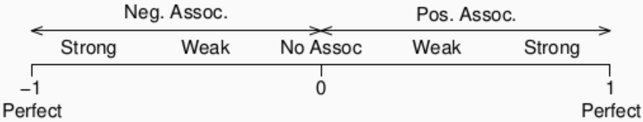
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right)$$



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation interpretation

- Ranges from -1 to $+1$
  - $\rightarrow$ $r \approx +1$: strong positive association
  - $\rightarrow$ $r \approx -1$: strong negative association
  - $\rightarrow$ $r \approx 0$: weak/no linear relationship

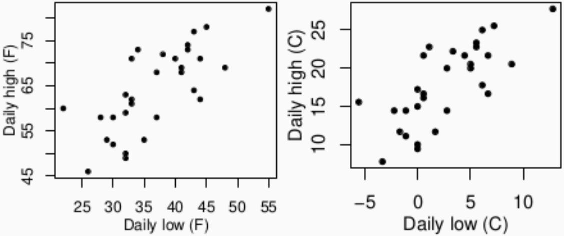


https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation is unit-less

- Covariance is unbounded, depends on variable ranges
- Correlation : compare metrics with different ranges

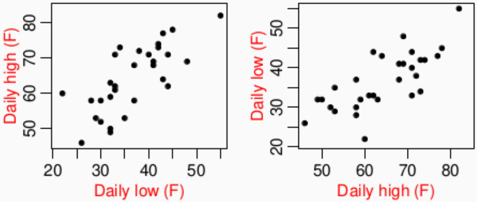
  $\longrightarrow$ Example : temperature in Celsius or Farehnheit – $r = 0.74$



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation is symmetric

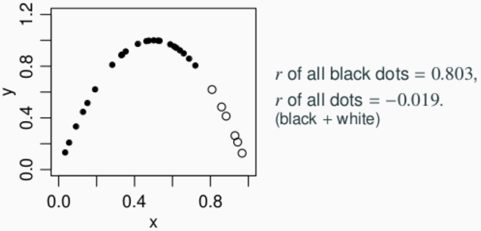
- Correlation is symmetric



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation shows linear association

- Correlation does not model non-linear association



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Spearman's rank correlation

- The actual compared X and Y values may be irrelevant
  $\rightarrow$ Does X rank itmes more or less in the same order as Y?
- Spearman's ρ : linear (Pearson) correlation between ranks
  $\rightarrow$ Models monotonic correlation
- In the presence of ties, correction is needed
  $\rightarrow$ Assign fractional ranks, for example

Spearman example

$$\rho = \frac{6\sum d_i^2}{n(n^2-1)}$$

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Source: `https://en.wikipedia.org/wiki/Spearman_correlation`

Kendall-tau correlation

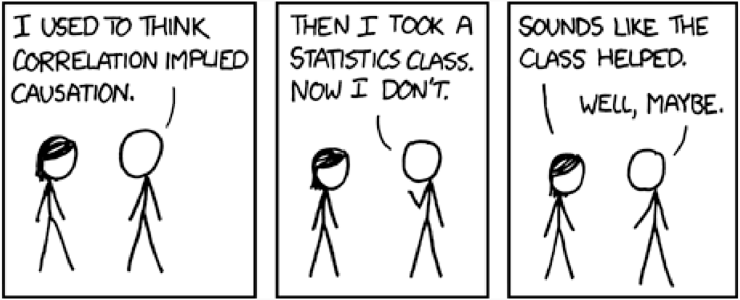
- Rank correlation, distinguishes local/distant mismatches
  - → Ranking an item 5 instead of 3 is not too bad
  - → Ranking an item 58 instead of 3 is really bad
- Consider all possible pairs (x_i, x_j) and (y_i, y_j) with $i < j$
  - → If $x_i < x_j$ and $y_i < y_j \implies$ concordant
  - → If $x_i > x_j$ and $y_i > y_j \implies$ concordant
  - → Else, discordant pairs

$$\tau = \frac{\#(concordant\ pairs) - \#(discordant\ pairs)}{\#(total\ pairs)}$$
$$= 1 - \frac{2 \times \#(discordant\ pairs)}{\binom{n}{2}}$$

Example : https://www.statisticshowto.com/kendalls-tau/

Confounders

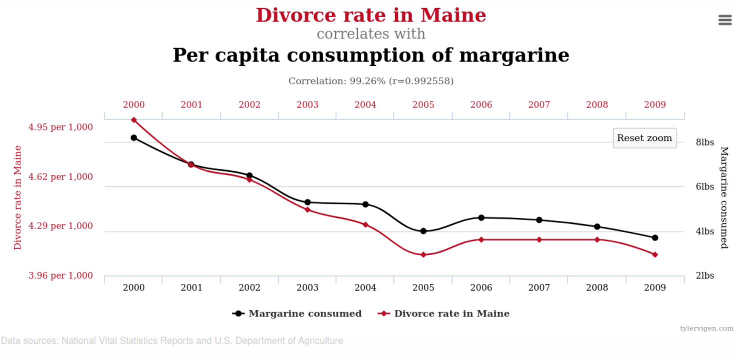
- Suppose X independent and Y dependent variables
- A confounder can influence both X and Y
- Correlation is not causation



Source: `https://xkcd.com/552/`

Spurious correlations

- Correlations can be found between unrelated variables
- Procrastinate : `https://www.tylervigen.com/spurious-correlations`
  - $\longrightarrow$ What possible confounders could explain these correlations ?



Plan

Model/system comparison

- Incremental research
  - State of the art or baseline **system B**
  - My own proposal **system A**
- How can I check if A is better than B ?
- What's the probability of drawing a wrong conclusion ?

Methodological framework

Take inspiration from health, biology, social siences

Randomised double-blind trial

- Randomly assign people to 2 groups :
  - Group A - treatment/vaccin
  - Group B - placebo
- Define a relevant metric, apply on A and B :
  - e.g. proportion P of healed people
- If $P_A > P_B$ the treatment/vaccin works
- Groups A and B - population sample
  - Is this sample large/representative enough ?
  - Is the observed difference $P_A - P_B$ significant ?

NLP system/model comparison

- We develop a system A
  - Is it better than baseline/SOTA system B ?
- Idea :
  - new/unseen data - test set
  - apply A and B on test set
  - compare their performances

Evaluation on held-out test set

- Test set
  - $x = x^{(1)} \dots x^{(m)}$ composed of m input examples
  - $y = y^{(1)} \dots y^{(m)}$ reference outputs (gold/correct/ground truth)
- Method :
  1. Apply A to x to obtain $\hat{y}_A$, compare to y
  2. Calculate the evaluation metric $M(A, x, y)$ - Example : accuracy
     $$M(A, x, y) = \frac{1}{m} \sum_{i=1}^{m} \delta(\hat{y}_A^{(i)}, y^{(i)})$$
  3. Do the same for B, obtain $M(B, x, y)$
  4. Calculate the difference (effect)
     $$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)$$
- $\delta_{A-B}(x, y) > 0 \implies$ system A better than B

Evaluation on held-out test set

- Test set
  - $x = x^{(1)} \dots x^{(m)}$ composed of m input examples
  - $y = y^{(1)} \dots y^{(m)}$ reference outputs (gold/correct/ground truth)
- Method :
  1. Apply A to x to obtain $\hat{y}_A$, compare to y
  2. Calculate the evaluation metric $M(A, x, y)$ - Example : accuracy
  $$M(A, x, y) = \frac{1}{m} \sum_{i=1}^{m} \delta(\hat{y}_A^{(i)}, y^{(i)})$$
  3. Do the same for B, obtain $M(B, x, y)$
  4. Calculate the difference (effect)
  $$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)$$
- $\delta_{A-B}(x, y) > 0 \implies$ system A better than B
  - Really ?

P-value

- Could the observed $\delta_{A-B}(x, y) > 0$ be due to chance ?
  - x, y is a sample of a joint random variable X, Y
  - What effect/difference would be observed for sample x', y' ?
  - What is the probability that A is actually no better than B ?

p-value

- Probability of drawing wrong conclusion
  - When stating A better than B
  - Given the observed effect $\delta_{A-B}(x, y)$
- We want to minimise this probability
- Usual threshold : $p < 0.05 \implies$ significant difference

Hypothesis testing

- $H_0 : \delta(X, Y) \leq 0 \implies$ if true, then A not better than B
- $H_1 : \delta(X, Y) > 0$

- $X, Y \rightarrow$ random variables, all possible test sets
  - Of which x, y is an m-sized sample
- Reject $H_0 \implies$ significant difference between the systems
- **P-value** : probability of observing $\delta_{A-B}(x, y)$ while H_0 is true :
  - $p - value = P[\delta(X, Y) \geq \delta_{A-B}(x, y) | H_0]$
  - probability to reject H_0 when it is true

Type I and type II error i

- **Type I error : false positives**
  - Rejecting H_0 when it is actually true, OR
  - Concluding that the observed difference greater than 0 $(A >> B)$ but it actually isn't $(A \leq\leq B)$
  - If p-value is below the significance level (usually $\alpha = 0.05$), we say that the difference is statistically significant
  - In other words, if probability of making type I errors (p-value) is sufficiently low, we can reject H_0

Type I and type II error ii

- **Type II error : false negatives**
  - Not rejecting H_0 when it is actually false
  - Concluding that the observed difference is no greater than 0 ($A \leq\leq B$) but it actually is ($A >> B$)
  - A **test's power** is its probability of avoiding type II errors

Goal :

- Guarantee that probability of type-I errors upper bounded by α
- Achieve as high power as possible

Example : Student's t-test

- Difference of means
  - Accuracy is a mean (Bernoulli trial averaged over m instances)
  - $M(A, x, y) = \frac{1}{m} \sum_{i=1}^{m} \delta(\hat{y}_A^{(i)}, y^{(i)})$
- $m = 25, M(A, x, y) = 0.88, M(B, x, y) = 0.79, SE = 0.08$ [3]

$$\text{t-stat} = \frac{M(A, x, y) - M(B, x, y)}{SE/\sqrt{m}} = 5,625$$

- P-value : check Student's t table, $m - 1$ degrees of freedom
- In practice : scipy `stats.ttest_rel`

3. SE = standard error, standard deviation of the difference $\hat{y}_A^{(i)} - y^{(i)}$.

Non parametric tests

- Problem of t-test : assumes $M(A, x, y) \sim$ normally distributed
- Other metrics :
  - Recall $R = tp/t$ linear wrt. tp, t constant

    $\rightarrow$ t-test OK ✓
  - Precision $P = tp/p$ depends on p, unknown distribution

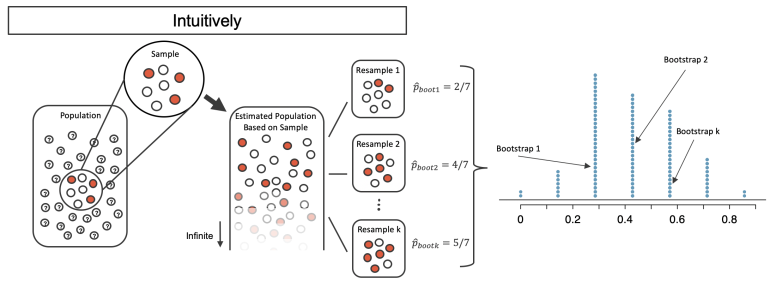
    $\rightarrow$ t-test not OK ✗
  - F-score $2PR/(P + R)$ depends on P, unknown distribution

    $\rightarrow$ t-test not OK ✗
- Alternative : non parametric tests
  - no sampling
    - Fast
    - Conservative, will not state $A > B$ for small δ (not powerful)
  - with sampling (slow, powerful)
    - E.g. randomised approximaiton, bootstrap test

Source : Yeh (2000) `https://aclanthology.org/C00-2137/`

Bootstrap

Idea : estimate M distribution by random re-sampling in x, y



https://bookdown.org/gregcox7/ims_psych/foundations-bootstrapping.html

Bootstrap for significance (Efron & Tibshirani 1993)

Input

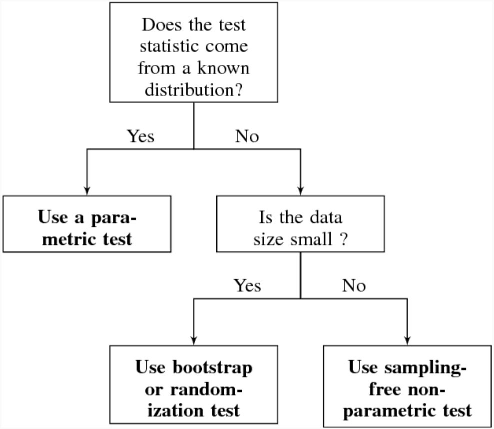
- test set $x = x^{(1)} \dots x^{(m)}, y = y^{(1)} \dots y^{(m)}$,
- predictions $\hat{y}_A^{(i)}$ et $\hat{y}_B^{(i)}$ of systems A and B for each item $x^{(i)}$
- evaluation metric $M(\cdot)$

```
deltaobs = M(A,x,y) - M(B,x,y)   # observed difference
for i in range(R) :              # R constant 10k - 100k
  xprim, yprim = sample(x,y,m)   # sample m with repetition
  deltasample = M(A,xprim,yprim) - M(B,xprim,yprim)
  if deltasample > 2 * deltaobs :
        r = r + 1
pvalue = r/R                     # % of surprising results
return pvalue
```

Evaluation metric M distribution vs. test

- Parametric test ($M(A, x, y)$ from known distribution)
  - Paired Student's t-test
- Non-parametric tests ($M(A, x, y)$ from unknown distribution)
  - No sampling (less powerful)
    - Sign test
    - McNemar's test
    - Wilcoxon signed rank test
  - Sampling (computationally expensive)
    - Permutation (randomized approximation) test
    - Bootstrap test

Which test to apply ?



Source: Dror et al. (2018) `https://aclanthology.org/P18-1128/`

Multiple comparisons

- Multiple comparisons : probability of false claims increases
- Bonferroni's correction
  - Divide significance level α by the number of datasets N
- Replicability analysis

P-hacking

A significant *p*-value can always be obtained for large-enough samples

Community's practice

# papers that **do not** report significance	117	15
# papers that report significance	63	18
# papers that report significance but use the **wrong** statistical test	6	0
# papers that report significance but do not mention the test name	21	3
# papers that have to report replicability	110	19
# papers that report replicability	3	4
# papers that perform cross validation	23	5

Source: Dror et al. 2018

Plan

Advanced data analysis

- Correlation works well for 2 numerical variables
- What if the variables are categorical ?
- Waht if we have more than 2 variables ?

Advanced data analysis

- Correlation works well for 2 numerical variables
- What if the variables are categorical ?
- Waht if we have more than 2 variables ?

Further statistical tools

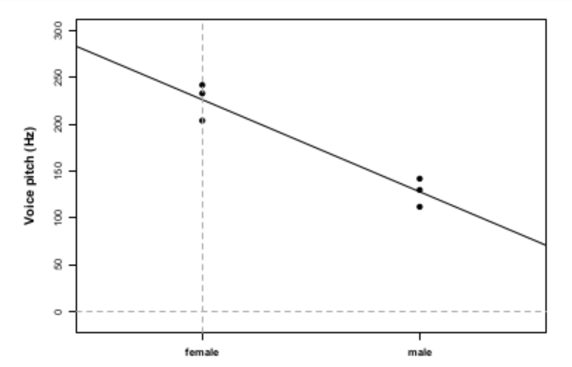
- Information theory
- ANOVA
- Linear models
- Mixed models
- ...

Information theory

- Entropy : alternative view of variability/skewness
  - $\rightarrow$ $H = -\sum p(x_i) \log p(x_i)$ $\rightarrow$ amount of uncertainty
  - $\rightarrow$ $H = \max$ for uniform distribution (unpredictable)
  - $\rightarrow$ $H = 0$ for highly skewed distribution (predictable)
- Other useful notions :
  - $\rightarrow$ Cross entropy
  - $\rightarrow$ Mutual information
  - $\rightarrow$ Kullbak-Leibler divergence (asymmetric)
  - $\rightarrow$ Jensen–Shannon divergence (symmetric)

Models for categorical variables

- ANOVA : Generalise t-test for more than 2 means
- Linear model : predict a linear regression slope
  - → Is the slope is significantly different from zero ?
  - → Notation : pitch $\approx$ sex $+\varepsilon$
- Mixed model : more sophisticated for multiple factors



Plan

Statistics libraries

- Visual : Excel, Libreoffice, ...
- Python : `matplotlib`, `numpy`, `scipy`, `sklearn`, ...
- R : multiple libraries including linear models
- Proprietary : Matlab, SPSS, ...

Error analysis

- Characterise the errors in our model
- Scripts to print characteristics of errors
  - $\longrightarrow$ Frequency, length, resolution, predicted/gold class, ...
  - $\longrightarrow$ Example : compounds predicted in wrongest positions
- Manual error annotation : taxonomies, guidelines
  - $\longrightarrow$ Gain insight on most promising improvements

Leaderboards, shared tasks

- Remember Goodhart's law (metric $\neq$ objective)
- Beating state of the art is good
- Learning something interesting about the problem is better
- From time to time : remember the research question

Negative results

- Well designed hypothesis have more interesting “negative” results
- Experiments require persistence and some faith
- Source of frustration : publish or perish

  $\longrightarrow$ Is it a problem with my results or with the system ?
- Negative results are publishable if sound experimental design

Confirmation bias

- Tendency to favour interpretations that confirm initial beliefs
- Well studied in psychology
- May lead to cognitive dissonance
- Tool : try to demonstrate the opposite of the initial hypothesis

  $\rightarrow$ If you fail for long enough, maybe the initial hypothesis is true

Sources

- Cours d'Adeline Paiement
- Statistical Significance Testing for NLP (Dror et al. 2020)
- `https://bodo-winter.net/tutorials.html` (thanks Leonardo Pinto Arata)
- Wikipedia
- Google images