How to Approach Statistical Analysis

Assignment #7: Reviewing Proposals (due 15th June)

- ✗ You have your review assignments on HotCRP: https://kaist-cs492d-i2r2020.hotcrp.com
- ✗ Finish your reviews by the deadline!
  - Remember, it may take more time than you thought, if you want to do a good job of it!

✗ There are lies, damned lies, and **statistics.**

Often attributed to Mark Twain, Benjamin Disraeli, and many others, but origin not clear (https://en.wikipedia.org/wiki/Lies,_damned_lies,_and_statistics).

Overview

- We cannot cover “statistics” as a topic: it is an entire branch of mathematics.
- Instead, I want to comment on the following topics
  - How to measure things correctly
  - Descriptive Statistics and EDA
  - How to interpret “significance” properly
  - Essential Skills

Definition

- statistic, noun, a fact or piece of data obtained from a study of a large quantity of numerical data: the statistics show that the crime rate has increased. (Oxford English Dictionary)
- The study of statistic is called statistics.
- The mean is a statistic.

Why bother with statistics at all?

- Many of CS studies are empirical and require investigation of data. Statistics is a discipline that concerns interpretation and representation of data.
- It allows you to write a more convincing story.
  - "No statistical analysis" can be a legitimate reason to reject.
- Being a good statistician never hurts!
  - But is is HARD.
  - Taking a course is highly recommended.
  - Awareness and basic knowledge is critical.

Measurement Theory

- Every quantitative study starts from measuring something.
- If you are not measuring something that actually is related to the "property" you are interested it, it becomes a threat to construct validity (i.e., the study is not designed to measure what you want to measure)

- "Measurement is the process of empirical objective assignment of numbers to entities, in order to characterise a specific attribute."
  – Prof. Norman Fenton, Queen Mary University, UK

Types of Scale

- ✗ Nominal: categorical
- ✗ Ordinal: order matters
- ✗ Interval: interval between numbers meaningful
- ✗ Ratio: ratio between numbers meaningful

Nominal Scale

- ✗ Simply a classification: no order, no size, no quantitative meaning.
  - Blood groups: O, A, B, AB
  - Which programming language is this program written in? A. Java, B. Python, ...
- ✗ Obviously cannot do any numerical analysis over nominal scale.

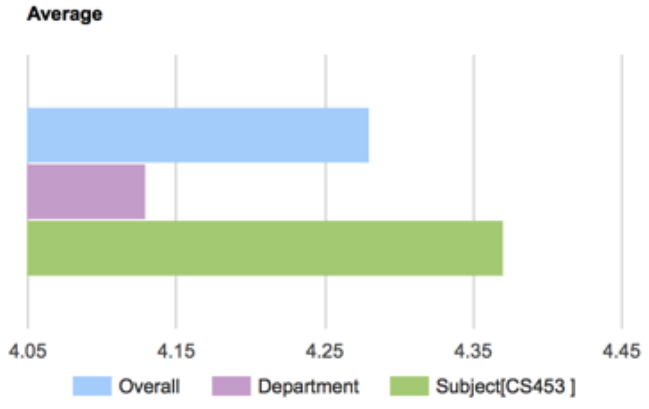
Ordinal Scale

- ✗ Order matters, but nothing else! You still cannot compare sizes, differences, etc. The most famous example is the Likert Scale (named after the psychologist Rensis Likert). Yet it is very frequently abused.
- ✗ Do you agree that you can report the average of a measurement taken in Likert scale?

1	2	3	4	5
Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly disagree

Classification		Lecture-Oriented Course
	1. Structure	1. Was the overall course constructed and carried out s ystematically? Strongly disagree(0-20%) Disagree(20-40%) Neutral(40-60%) Agree(60-80%) Strongly agree(80-100%)

Required (used in average calculation)			
No.1 [Structure]	**No.2 [Effectiveness]**	**No.3 [Participation]**	**No.4 [Understanding]**
4.35	4.47	4.47	4.18



Interval & Ratio

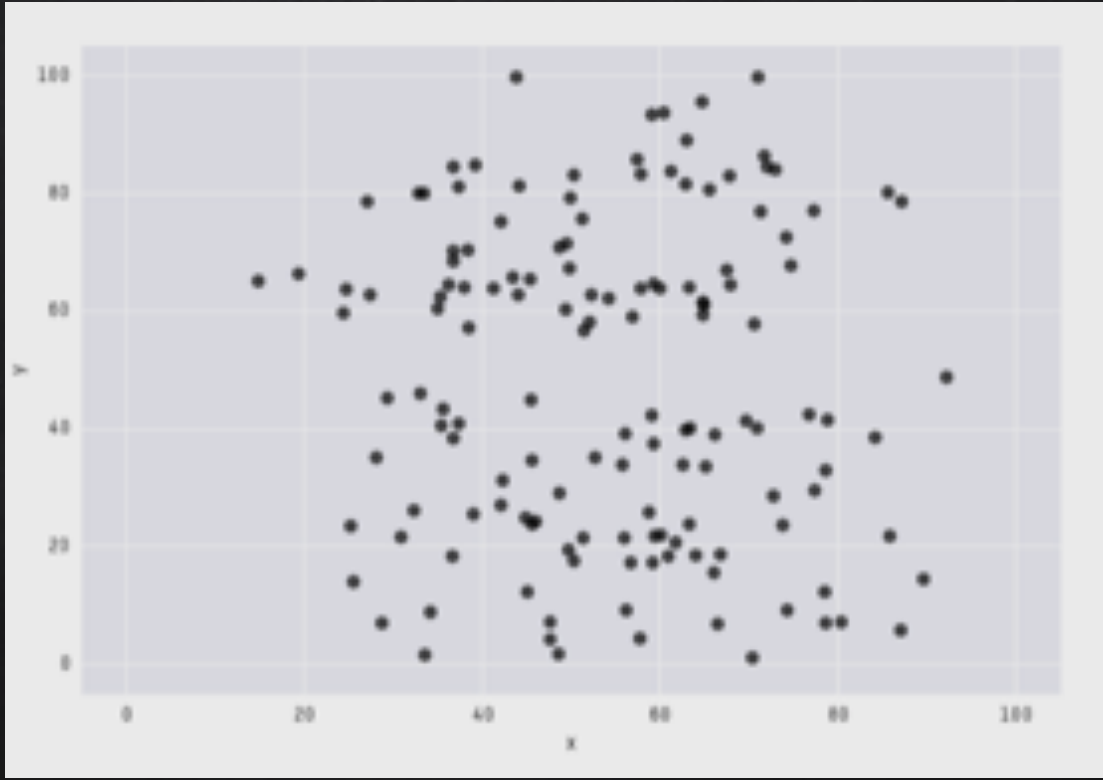
- Interval scale: order and intervals between numbers are meaningful, but not the absolute values and ratios. Example: temperature in C/F.
  - "The difference in temperature between Seoul and Cairo is half of the difference in temperature between Seoul and Moscow." → meaningful statement.
  - Note that only intervals in temperature are meaningful. "Today, Seoul is 30C whereas Moscow is 15C, therefore Seoul is twice hotter." → not meaningful.
- Ratio: can do everything (ex. Temperature in K)

Best practices for presenting data

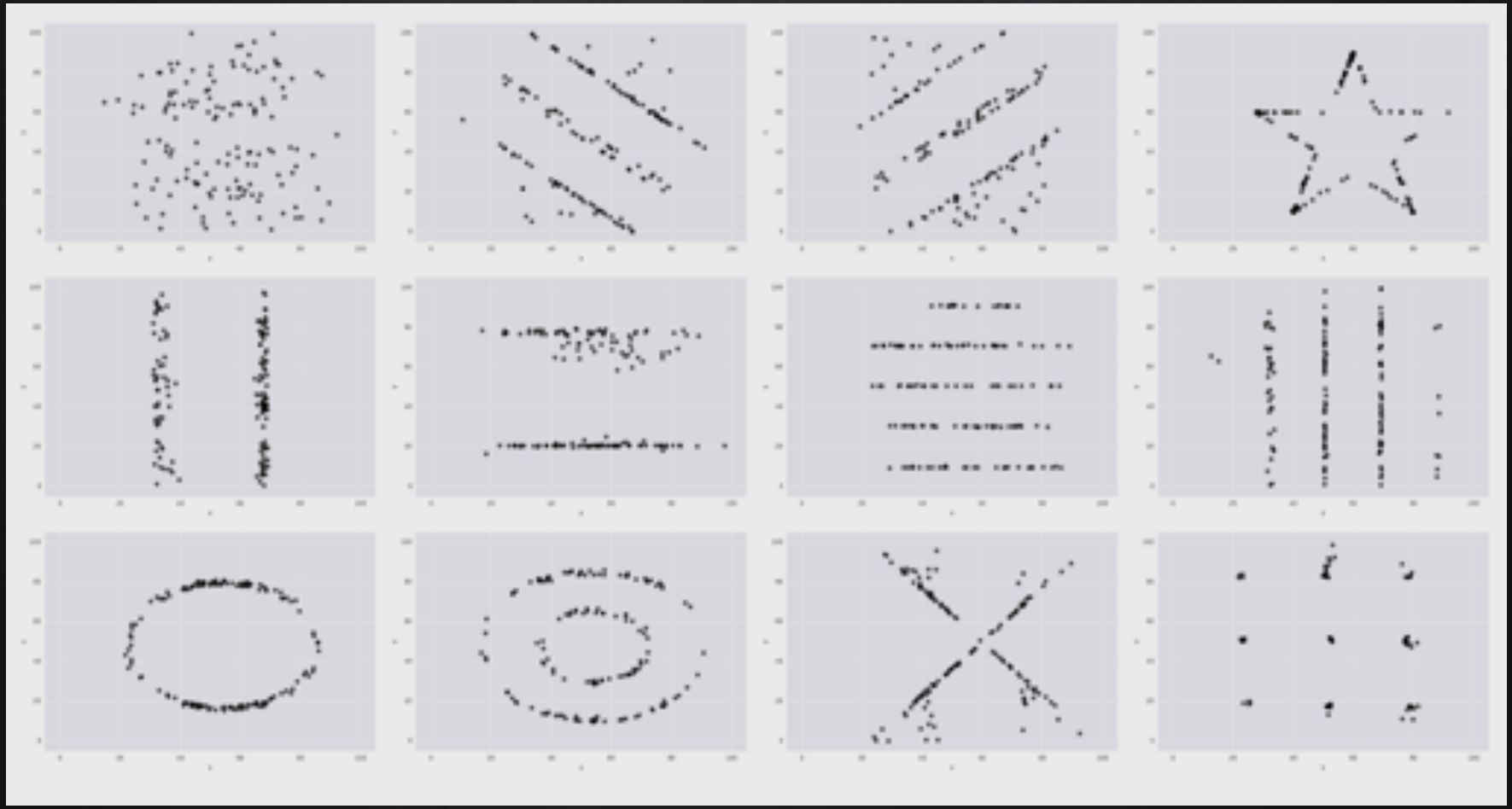
- Roughly speaking, there are two major branches in statistics
  - Descriptive statistics: how to summarise large data
  - Inferential statistics: how to draw conclusion under the influence of random variations
- You need both to understand your data
  - But it is very easy to make statistical analysis a mechanical "ritual" where you apply an arbitrary set of steps without thinking

What can you read from these numbers?

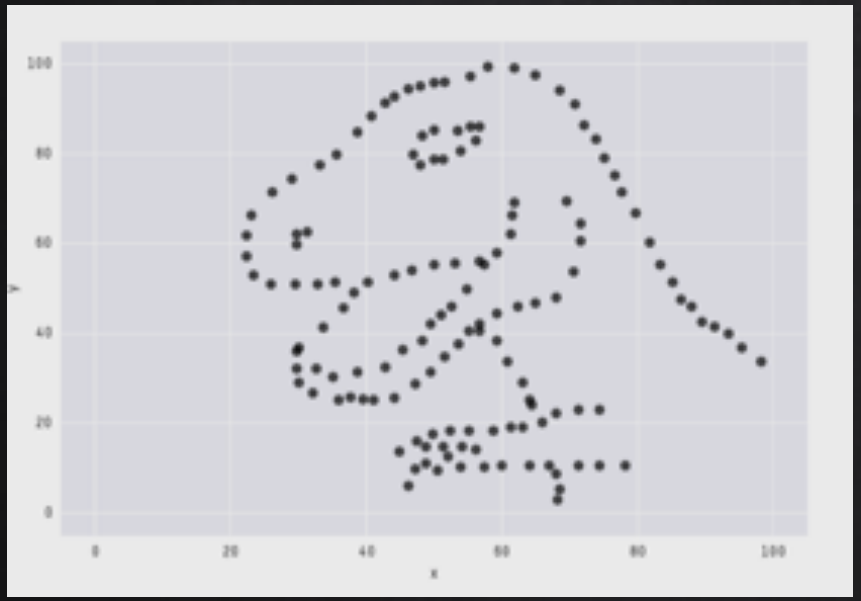
- We measured two variables, X and Y, from each subject.
  - X Mean: 54.26
  - Y Mean: 47.83
  - X standard deviation: 16.76
  - Y standard deviation: 26.93
  - Correlation between X and Y:– 0.06



Sounds about right?



They all share the same statistics. In fact…



This one too!

- “The Datasaurus Dozen”, from:

  Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing

  Justin Matejka, George Fitzmaurice
  CHI 2017 (Honourable Mention)

  https://www.autodeskresearch.com/publications/samestats

Descriptive Statistics & EDA

- EDA stands for Exploratory Data Analysis, which means the analysis of data to extract their characteristics, often aided by visualisation
- Do not rely on numerical descriptive statistics alone
  - You should have some expectation of how the data would "look"
  - Then always "see" the data to see if they match your expectation
  - Scatterplots and histograms should often be the first and the most important plot you produce, and not more sophisticated analysis

Descriptive Statistics

- A summary statistic is a summary statistic that quantitatively describes of summarises features of a sample.
  - Central tendency: what represents the sample?
  - Dispersion: who much variance is in the sample?

Scale	Central Tendency	Dispersion
Nominal	Mode (the most frequent choice)	Frequency of each item
Ordinal	Median (the midpoint when sorted)	Percentile (e.g., values at top and bottom 25%)
Interval	Arithmetic Mean $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$	Arithmetic Standard Deviation $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N-1}}$
Ratio	Geometric Mean $\mu_g = (\prod_{i=1}^{N} x_i)^{\frac{1}{N}}$	Geometric Standard Deviation $\sigma_g = \exp(\sqrt{\frac{\sum_{i=1}^{N}(\ln \frac{x_i}{\mu_g})^2}{N}})$

Statistical Significance

- You have compared technique X and Y . You think X is better than Y , based on N samples. But N is so smaller than the size of the population.
  - How can you be sure that what you observed is not just by chance (i.e. statistically significant)?

Hypothesis Test

- Hypotheses ask whether samples are from the same distribution. That is:
- Are X and Y different? (Variants: Is X greater/less than Y ?)
- Hypothesis tests tell you whether the observed results are sufficient to reject the null hypothesis. There are many different tests, depending on the assumption about the population distribution as well as the internal mechanism.
- But what does it really mean to be statistically significant? Some of you may have heard of "$p < 0.05$ therefore significant". But what does the p-value represent?

What does it mean to be statistically significant?

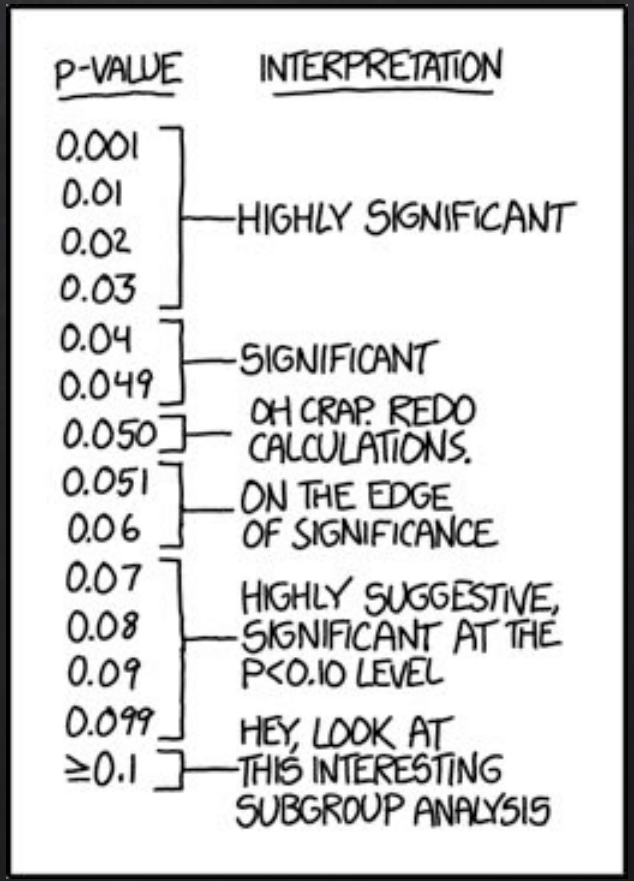
✗ The p-value is:

1. The probability that null hypothesis is true.
2. The probability that the alternative hypothesis is false.
3. The probability that the observation was produced by random chance alone.

✗ Answer 3 comes close but is not true. The p-value represents the probability that, when the null hypothesis is actually true, a statistical summary is equal to or greater than the observed summary.

P-value Fallacy

- ✗ There are numerous misunderstandings and fallacies surrounding p-value:
  - $p < 0.05$ does NOT mean that the null hypothesis is false. $p < 0.05$ does NOT mean that the alternative hypothesis is true.
  - A lower p-value does NOT denote the magnitude of difference.
  - The value 0.05 is not a magic number: it is merely a convention, inherited from long time ago.
- ✗ This should NOT be the final gate-keeper that says your result is meaningful, yet too many people use it that way.

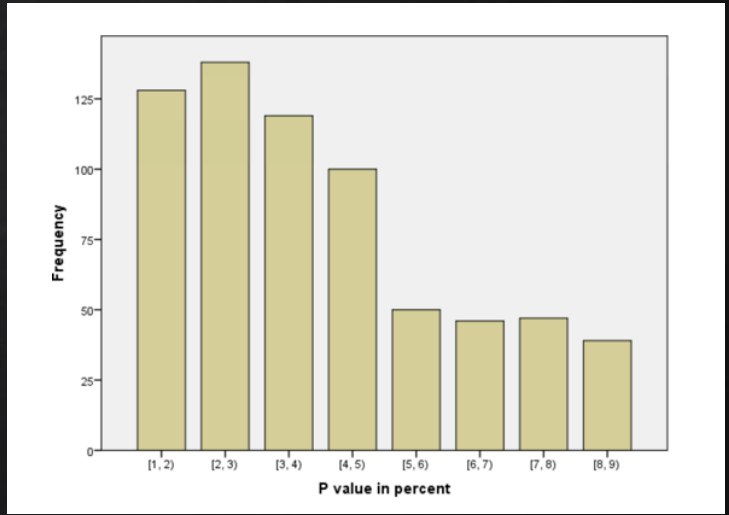


https://xkcd.com/1478/

Distribution of Reported P-values

- ✗ "The Distribution of P-values in Medical Research Articles Suggested Selective Reporting Associated With Statistical Significance"

  Perneger & Combescure
  Journal of Clinical Epidemiology
  2017
  (https://pubmed.ncbi.nlm.nih.gov/28400294/)



Practical Significance

- ✗ Statistical significance measures whether the difference you are observing may actually exist or not.
- ✗ It DOES NOT tell you anything about how much the difference is likely to be (the magnitude) or how often you are likely to observe the difference (the frequency).
- ✗ These are practical significance, also known as the effect size.

Effect Size Families

- ✗ d-family: measures magnitude of differences
- ✗ r-family: measures association between observations
- ✗ Common Language (CL) effect size: recent efforts to report effect sizes in intuitive and plain language

Effect Size: d-family

When comparing dichotomous variables (i.e., success vs. failure, detected vs. non-detected): assume that success rate in the treatment group was p, and the rate in the control group was q:

- Risk Difference: $p - q$
- Risk Ratio: $\frac{p}{q}$
- Odds Ratio: $\frac{p}{1-p} / \frac{q}{1-q}$

When comparing continuous variables, we want to argue about the difference in mean values:

- Cohen's $d = \frac{\mu_1 - \mu_2}{\sigma_{pooled}}$
- Glass's $d = \frac{\mu_1 - \mu_2}{\sigma_{control}}$
- Hedge's $d = \frac{\mu_1 - \mu_2}{\sigma^*_{pooled}}$ (σ weighted using sample sizes)

Effect Size: r-family

When showing the strength of relationship between two continuous variables:

- Pearson correlation ρ: when both variables are continuous (i.e., measured on interval or ratio scales), this measures the strength of linear correlation.

$$\rho_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Spearman correlation r_s: this is Pearson correlation applied to ranks (therefore it does not assume anything about linearity between raw data points).

$$r_s = \rho_{rg_x, rg_y} \text{ where } rg_x, rg_y \text{ are rankings of } x \text{ and } y.$$

Effect Size: Common Language family

Cohen's d is not very intuitive for nonspecialists. McGraw and Wong (1992) proposed Common Language effect size, based on the notion of probability of difference. Consider the height difference between men and women: men tend to be taller on average, and Cohen's d quantifies the difference. But if you pick a random pair of a man and a women, what is the probability of man being taller than the women? CL effect size is designed to answer that.

Vargha and Delaney (2000) extended McGraw and Wong to cater for ordinal scales. This statistic is called A_{12}.

$$A_{12} = (R_1/m - (m+1)/2)/n$$

where R_1 is the sum of X's rank, $m = |X|$, $n = |Y|$.

Best Practices

- ✗ Increasing sample size/feature count is not always a good thing to do.
  - With hypothesis testing: increasing sample size will decrease the p-value arbitrarily even when underlying distributions are the same
  - With machine learning: more features will cause the curse of dimensionality, which tends to make over-fitting much easier

Are larger sample sizes always better?

```
> x1 = rnorm(100, mean=5.11)
> x2 = rnorm(100, mean=5.21)
> t.test(x1 - x2)

        One Sample t-test

data:  x1 - x2
t = -0.025265, df = 99, p-value = 0.9799
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.3066499 0.2989388
sample estimates:
   mean of x
-0.003855518
```

Are larger sample sizes always better?

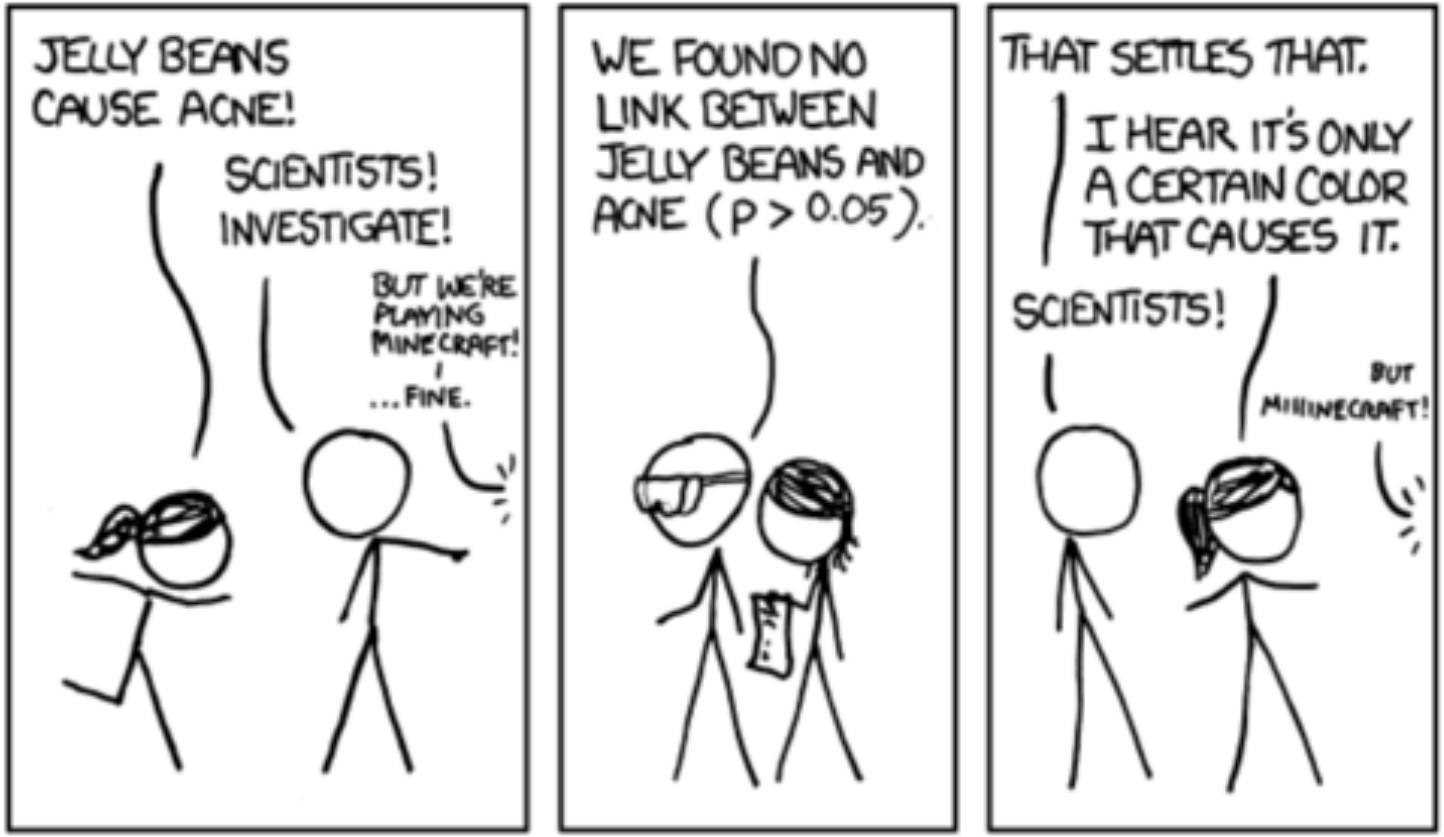
```
> x11 = rnorm(1000000, mean=5.11)
> x22 = rnorm(1000000, mean=5.21)
> t.test(x11-x22)

        One Sample t-test

data:  x11 - x22
t = -70.199, df = 1e+06, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.10210742 -0.09656057
sample estimates:
mean of x
-0.099334
```

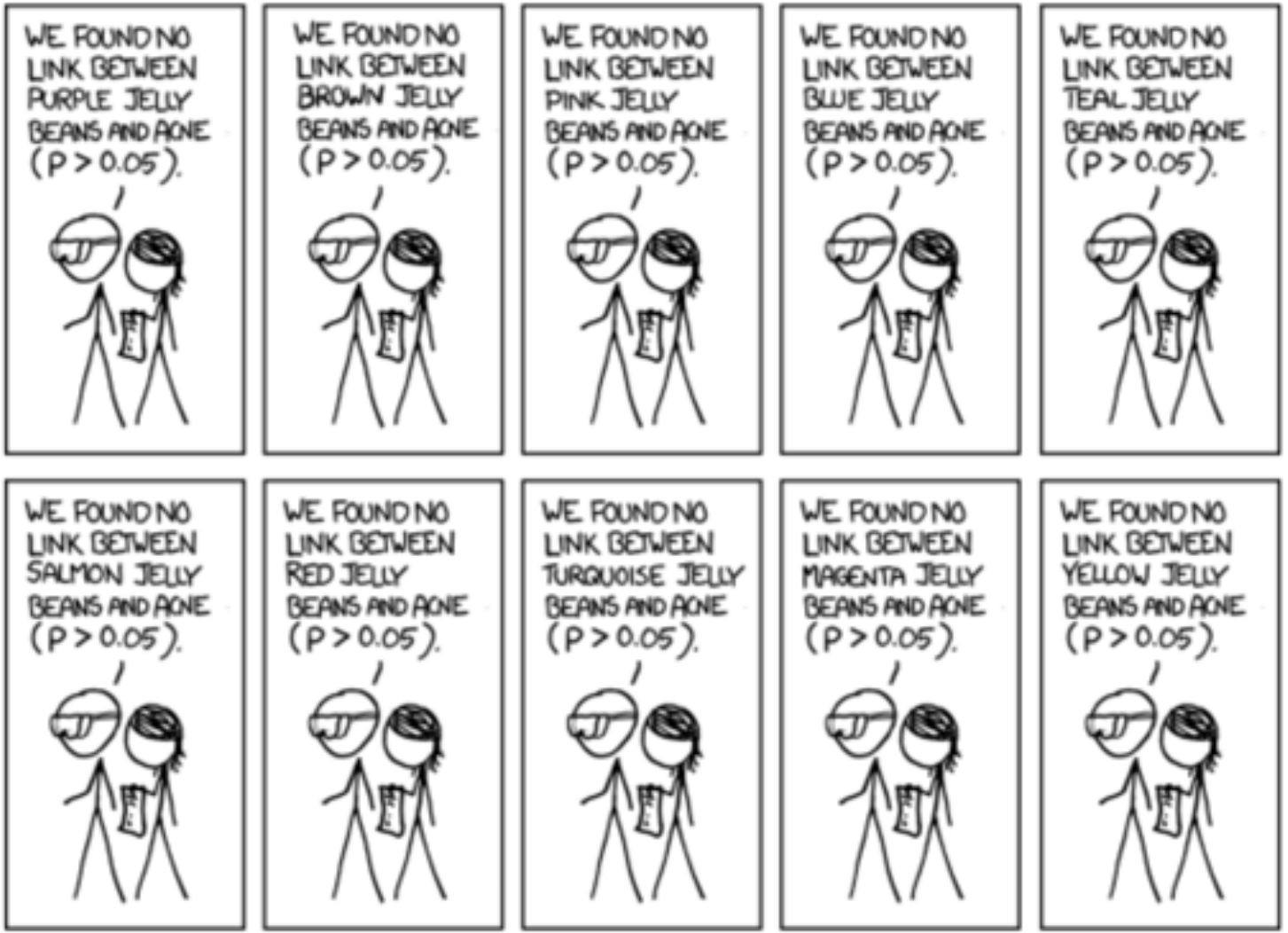
Best Practice

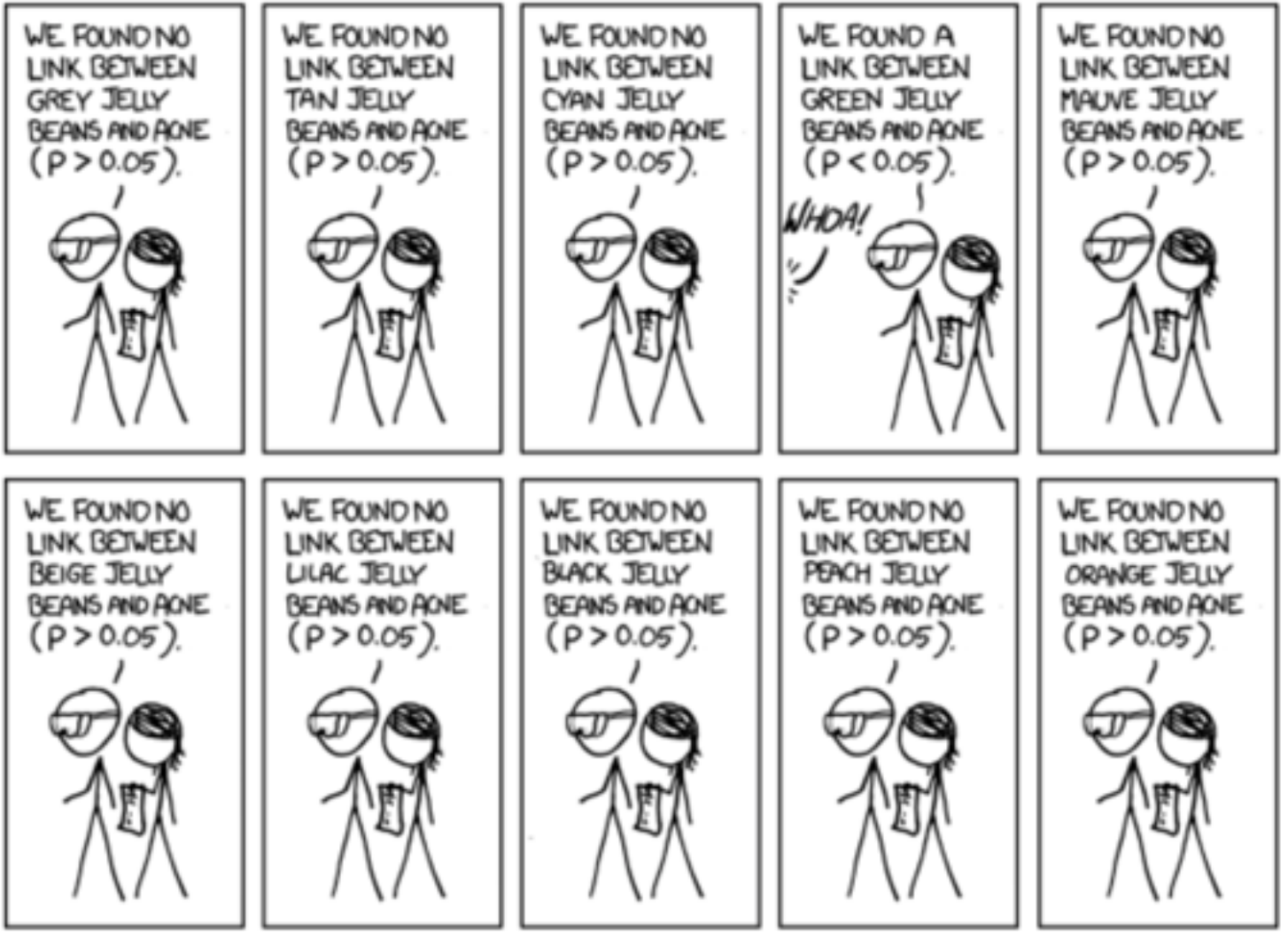
- ✗ Simply increasing the number of runs is also not always a good idea.
- ✗ "WHAT??"
- ✗ Remember what p-value was, once again. Then let's read one of my favourite XKCD comics together.

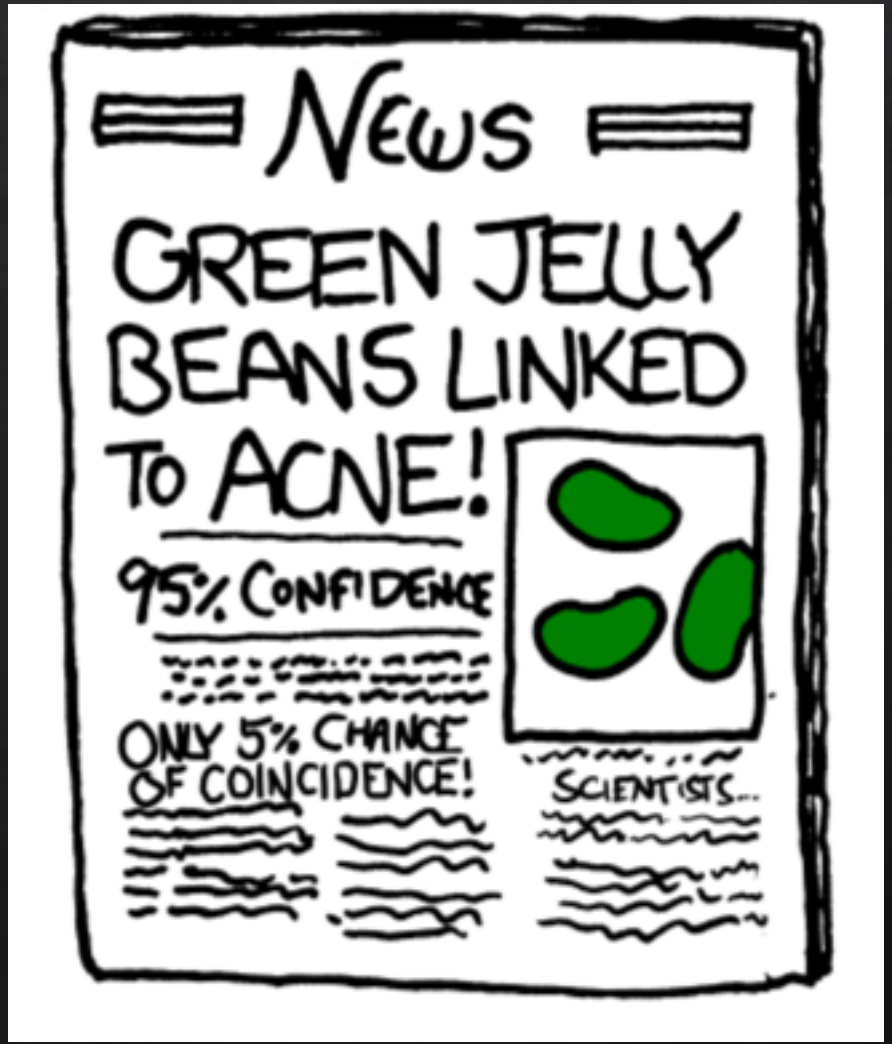
SIGNIFICANT

I< < PREV RANDOM NEXT > >I




WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).



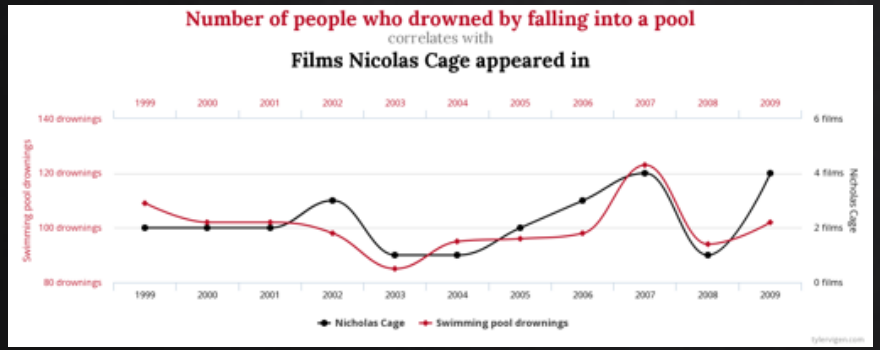
https://xkcd.com/882/

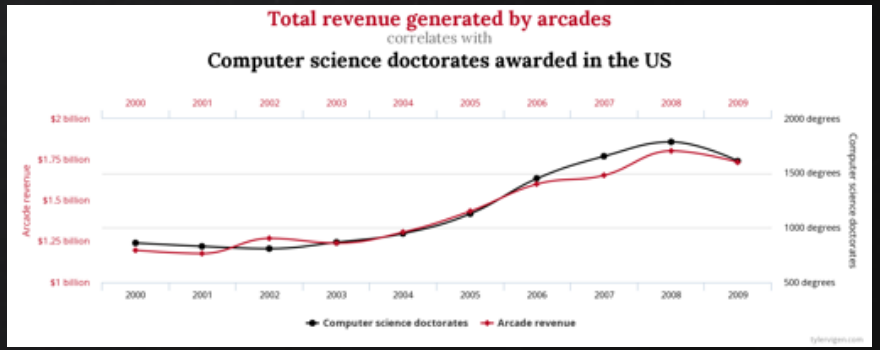
Bonferroni Correction

- When you run multiple tests (as in the green jellybeans/acne example), you increase the probability of Type 1 error (i.e., a high probability of rejecting at least one null hypothesis incorrectly).
- Bonferroni (1936) proposed that, when performing n tests, the significance level should be divided by n (or, equally, p-values multiplied by n), thereby reducing the Type 1 error risk conservatively. This is called Bonferroni Correction.
- Some people adopt Bonferroni correction, while other object. Objection is based on the fact that, faced with stricter significance level, researchers may choose to do fewer experiments.

Best Practice: Higher Level

- ✗ Do NOT hide data that you don't like!
- ✗ Do NOT do data-dredging (i.e., actively searching for data that fits your story)
  - "If you torture the data long enough, it will confess to anything" – Ronald H. Coase, Nobel Laureate (Economics)





https://www.tylervigen.com/spurious-correlations

Best Practices: Skillset

- ✗ Choose a tool from each of the following categories and GET REALLY GOOD at them – it will take time but it will PAY OFF
  - Data management: pandas, numpy, SQL, etc
  - Statistical Packages: scipy, GNU R, etc
  - Visualisation: matplotlib, ggplot, etc
- ✗ Learn how to automate these things together
  - Ideally you should have an automation pipeline that "builds" your paper from the raw data
- ✗ Do NOT rely on Microsoft Excel for everything.

Summaries

- ✗ Statistics is a story telling skill: get good at it if you need quantitative analysis for your research
- ✗ Do not treat statistics as a ritual
- ✗ Understand statistics enough to avoid all the anti-patterns
- ✗ Get familiar with essential tools – this will save a LOT of your time later