# DATA MINING
# DATA EXPLORATION AND STATISTICS

Exploratory analysis of data – basic statistics

# Exploratory analysis of data – What does my data look like?

- Summary statistics: numbers that summarize properties of the data

- Summarized properties include frequency, location and spread
  - Examples: location - mean
              spread - standard deviation

- Most summary statistics can be calculated in a single pass through the data

- Computing data statistics is one of the first steps in understanding our data

# Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
    - For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

- We can visualize the data frequencies using a value histogram
- Frequency and frequency histogram are the empirical analogue of probability and probability distribution

# Example

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | NULL | 60K | No |
| 7 | Yes | Divorced | 220K | NULL |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

Marital Status

| Single | Married | Divorced | NULL |
|--------|---------|----------|------|
| 4 | 3 | 2 | 1 |

Mode: Single

# Example

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|---------------|---------------|-------|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 10000K | **Yes** |
| 6 | No | NULL | 60K | **No** |
| 7 | Yes | Divorced | 220K | **NULL** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 90K | **No** |
| 10 | No | Single | 90K | **No** |

Marital Status

| Single | Married | Divorced | NULL |
|--------|---------|----------|------|
| 40% | 30% | 20% | 10% |

# Example

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 10000K | **Yes** |
| 6 | No | NULL | 60K | **No** |
| 7 | Yes | Divorced | 220K | **NULL** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 90K | **No** |
| 10 | No | Single | 90K | **No** |

We can choose to ignore NULL values

Marital Status

| Single | Married | Divorced |
|---|---|---|
| 44% | 33% | 22% |
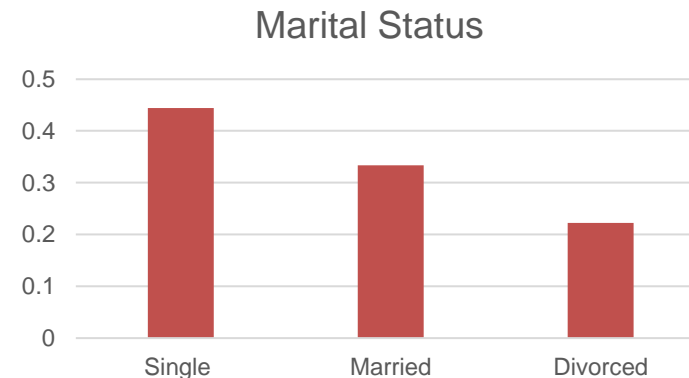


Marital Status

# Data histograms

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 10000K | **Yes** |
| 6 | No | NULL | 60K | **No** |
| 7 | Yes | Divorced | 220K | **NULL** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 90K | **No** |
| 10 | No | Single | 90K | **No** |

Use binning for numerical values



Refund

Marital Status

Income

REFUND

Yes

No

Marital Status

45%

33%

22%

Single   Married   Divorced

INCOME

<100K   [100K,200K]   >200K

50%

30%

20%
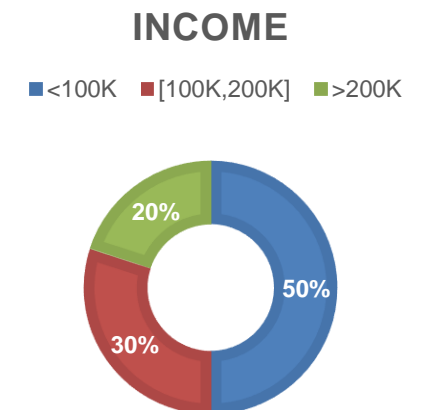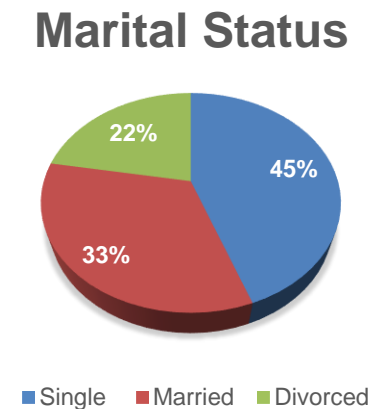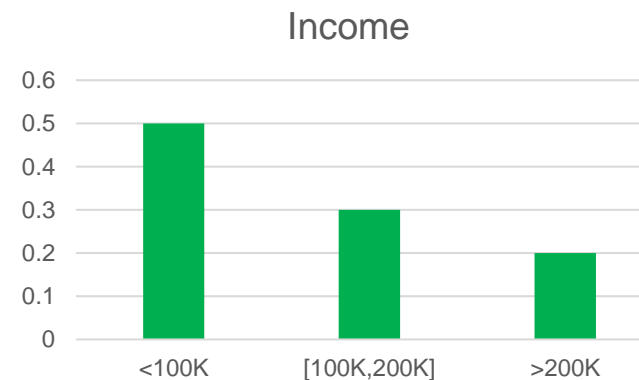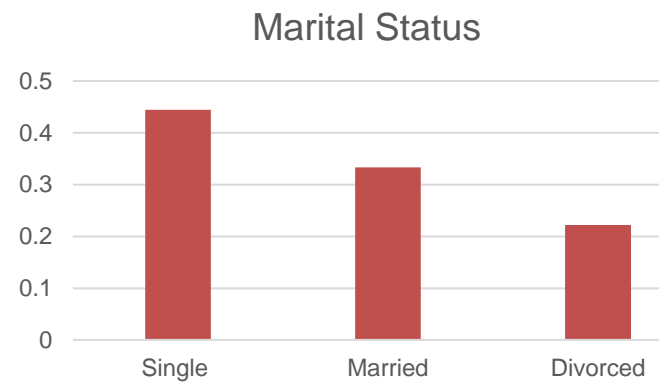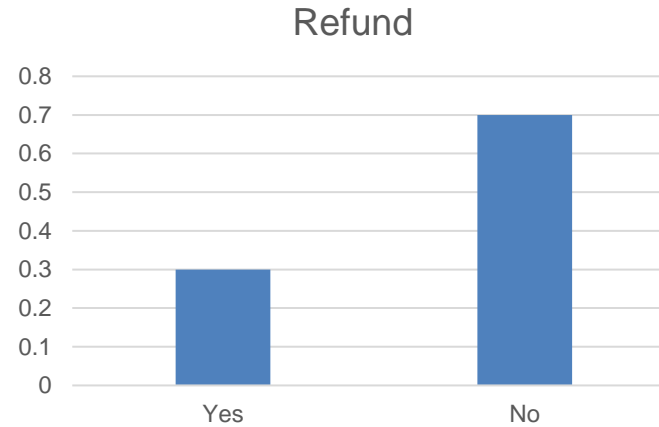
# Percentiles

- For continuous data, the notion of a percentile is more useful.

Given an ordinal or continuous attribute $x$ and a number $p$ between $0$ and $100$, the $p^{\text{th}}$ percentile is a value $x_p$ of $x$ such that $p\%$ of the observed values of x are less or equal than $x_p$.

- For instance, the 80th percentile is the value $x_{80\%}$ that is greater or equal than 80% of all the values of x we have in our data.

- The percentiles are the empirical analogue of the cumulative probability function

# Example

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | NULL | 60K | No |
| 7 | Yes | Divorced | 220K | NULL |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

| Taxable Income |
|----------------|
| 10000K |
| 220K |
| 125K |
| 120K |
| 100K |
| 90K |
| 90K |
| 85K |
| 70K |
| 60K |

$x_{80\%} = 125K$

# Measures of Location: Mean and Median

- The mean is the most common measure of the location of a set of points.

$$\text{mean}(x) = \overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

- However, the mean is very sensitive to outliers.
- Thus, the median is also commonly used.

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

- Or the trimmed mean: the mean after removing min and max values

# Example

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | NULL | 60K | No |
| 7 | Yes | Divorced | 220K | NULL |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

Mean: 1090K

Trimmed mean (remove min, max): 105K

Median: (90+100)/2 = 95K

# Measures of Spread: Range and Variance

- Range is the difference between the max and min

- The variance or standard deviation is the most common measure of the spread of a set of points.

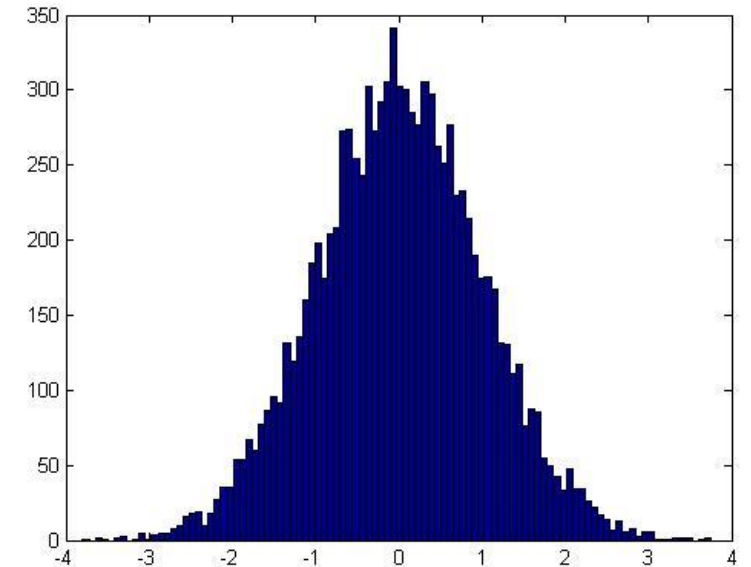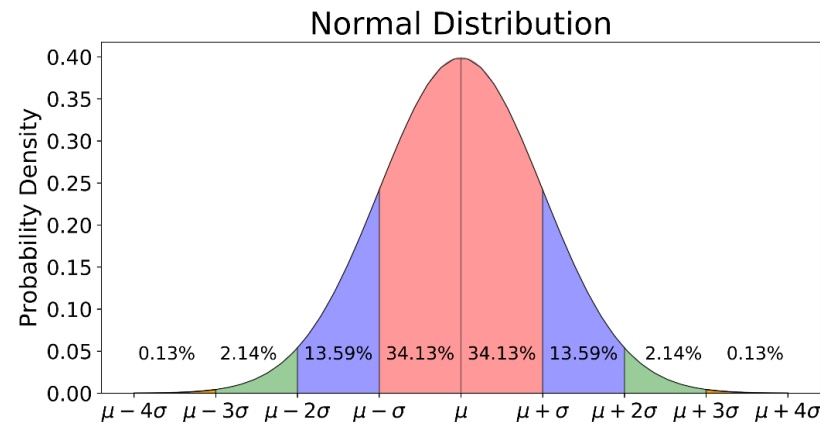$$var(x) = \frac{1}{m-1}\sum_{i=1}^{m}(x - \bar{x})^2$$

$$\sigma(x) = \sqrt{var(x)}$$

$m$ or $m - 1$?

When computing the sample variance m-1 is used

For large data it does not make much difference

# Normal Distribution



$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

This is a value histogram

- An important distribution that characterizes many quantities and has a central role in probabilities and statistics.
- Appears also in the central limit theorem: the distribution of the sum of IID random variables.
- Fully characterized by the mean $\mu$ and standard deviation $\sigma$

# Not everything is normally distributed

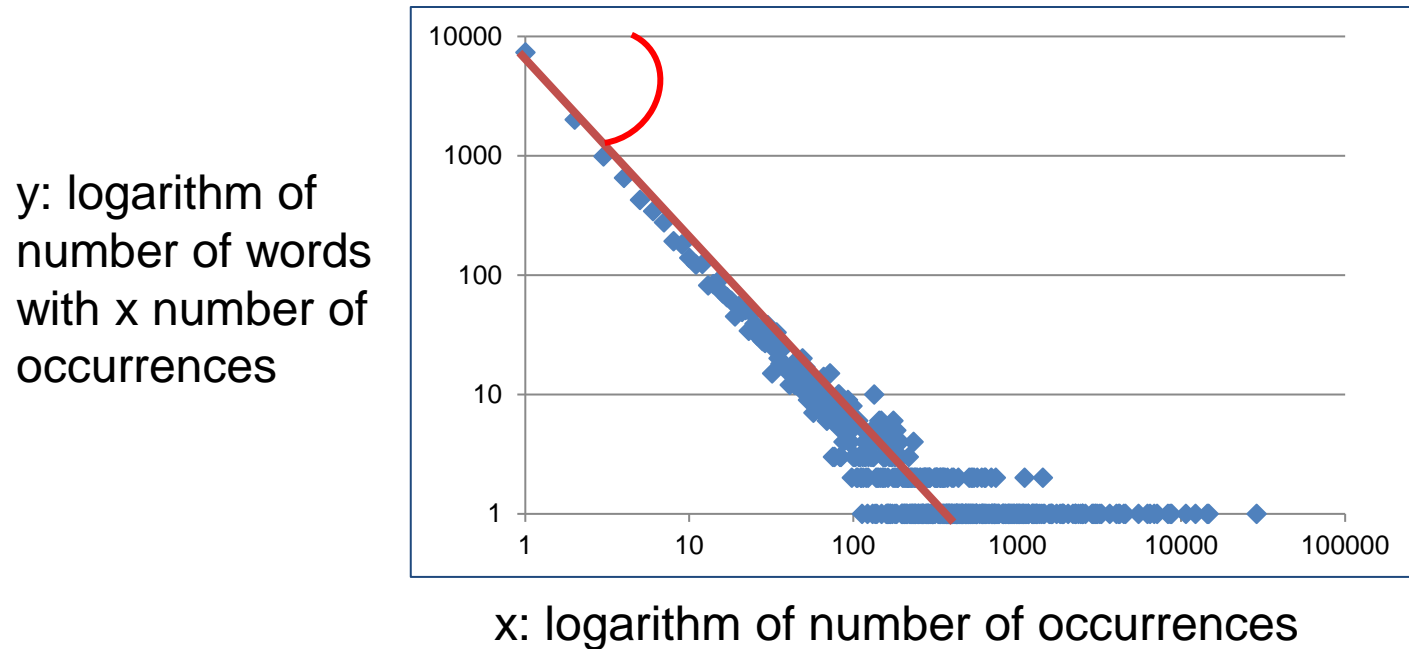- Plot of number of words with x number of occurrences

y: number of words with x number of occurrences



x: number of occurrences
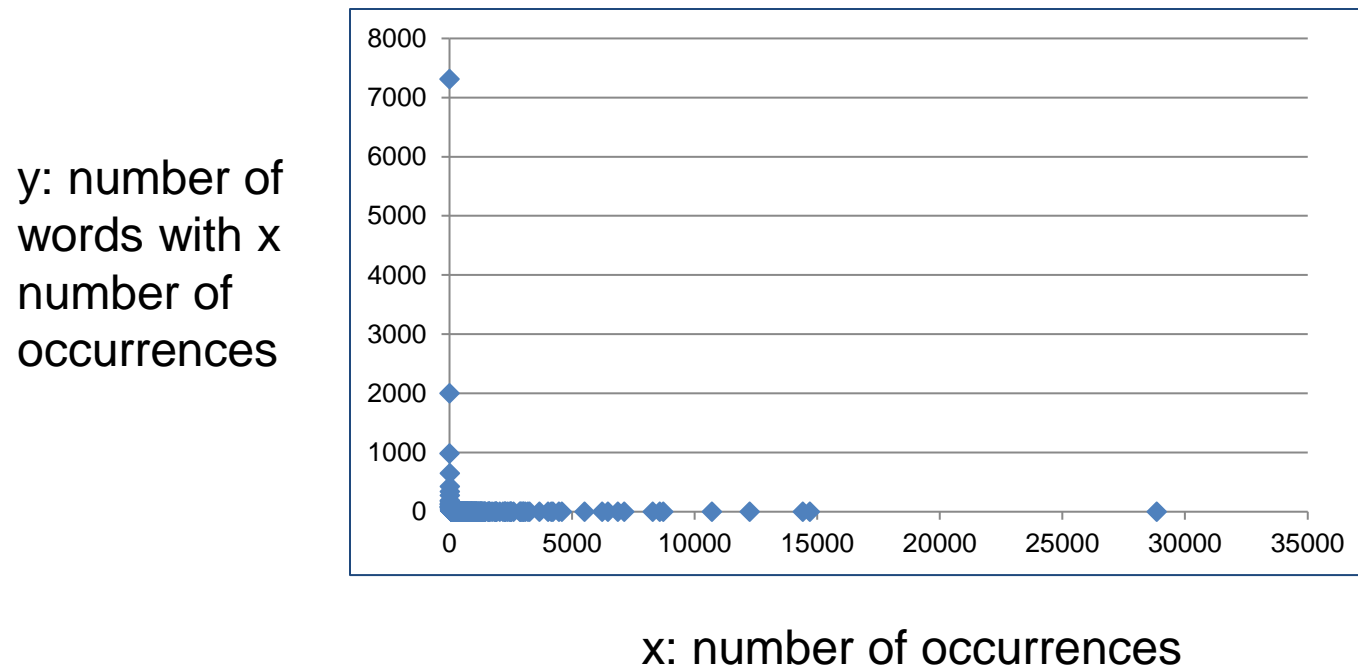
- If this was a normal distribution we would not have number of occurrences as large as 28K

# Power-law distribution

- We can understand the distribution of words if we take the log-log plot

y: logarithm of number of words with x number of occurrences



x: logarithm of number of occurrences

Power-law distribution:
$$p(k) = k^{-a}$$

The slope of the line gives us the exponent α

Linear relationship in the log-log space
$$\log p(x = k) = -a \log k$$

# Power-laws are everywhere

- Incoming and outgoing links of web pages, number of friends in social networks, number of occurrences of words, file sizes, city sizes, income distribution, popularity of products and movies
  - Signature of human activity?
  - A mechanism that explains everything?
  - Rich get richer process

- Related distribution: log-normal
  - Taking the log of the values gives a normal distribution
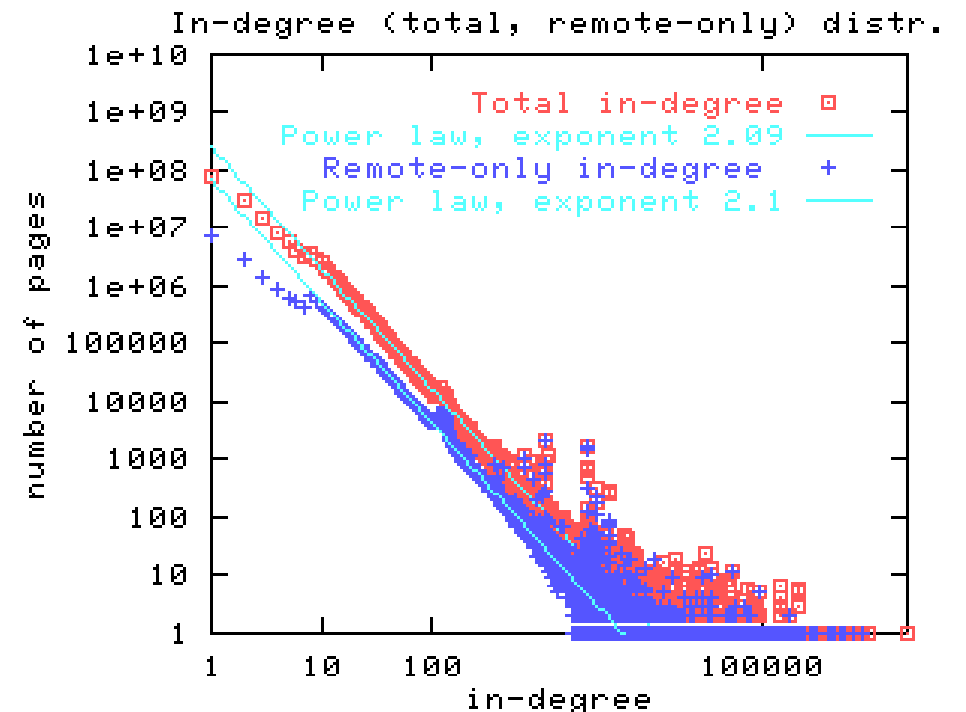
# Zipf's law

- Power laws can be detected also by a linear relationship in the log-log space for the rank-frequency plot

y: number of occurrences of the r-th most frequent word



Zipf distribution:
$$f(r) = r^{-\beta}$$
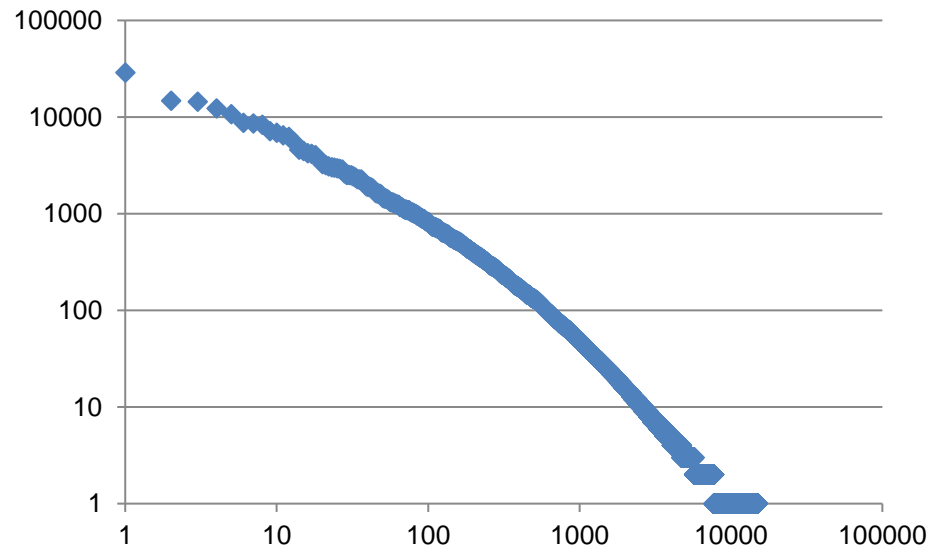
r: rank of word according to frequency (1st, 2nd …)

- $f(r)$: Frequency of the r-th most frequent word

$$\log f(r) = -\beta \log r$$

# The importance of correct representation

- Consider the following three plots which are histograms of values. What do you observe? What can you tell of the underlying function?
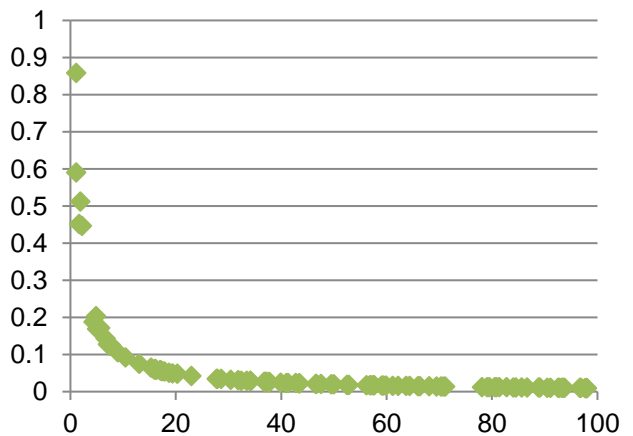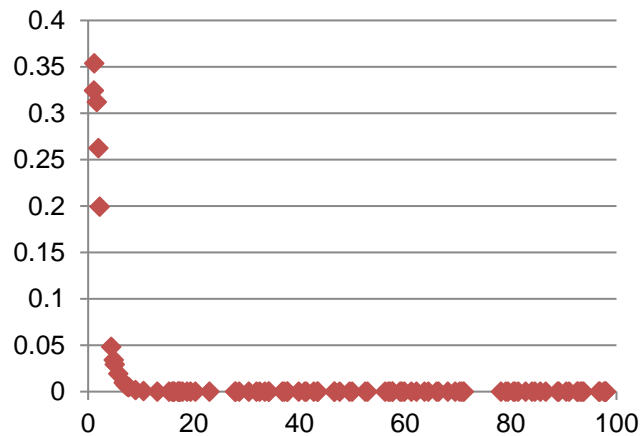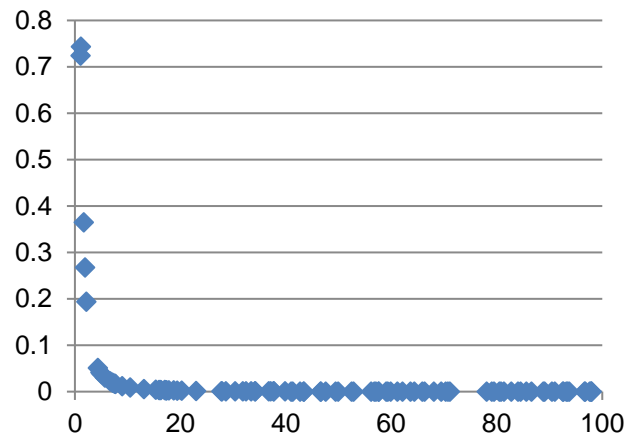
# The importance of correct representation

- Putting all three plots together makes it clearer to see the differences



- Green falls more slowly. Blue and Red seem more or less the same

# The importance of correct representation

- Making the plot in log-log space makes the differences more clear
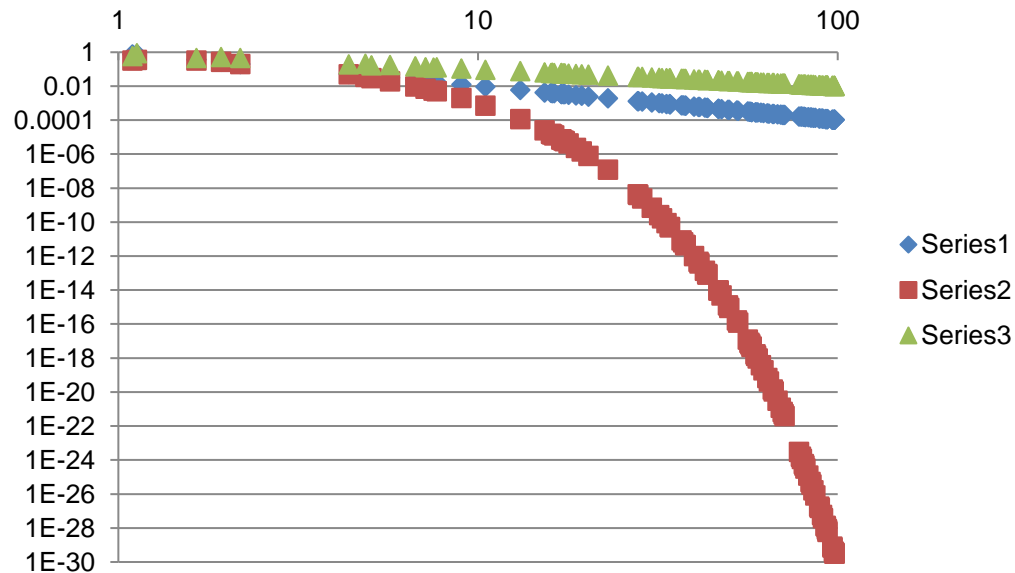


Linear relationship in log-log means polynomial in linear-linear
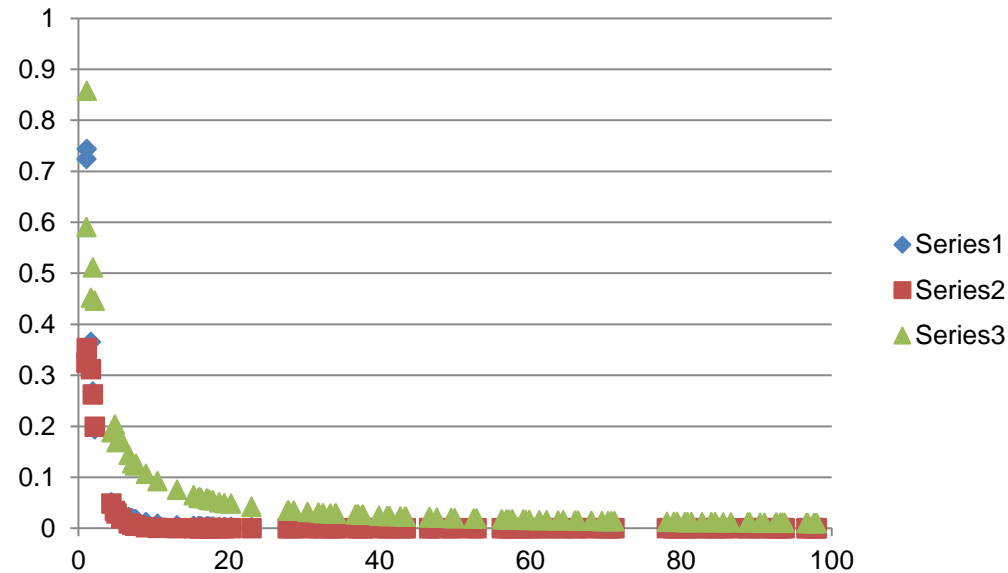The slope in the log-log is the exponent of the polynomial

Exponential relationship remains exponential in log-log

- Green and Blue form straight lines. Red drops exponentially.

  - $y = \frac{1}{2x+\epsilon}$  $\qquad\qquad \log y \approx -\log x + c$

  - $y = \frac{1}{x^2+\epsilon}$  $\qquad\qquad \log y \approx -2\log x + c$

  - $y = 2^{-x} + \epsilon$  $\qquad\qquad \log y \approx -x + c = -10^{\log x} + c$

# Attribute relationships

- In many cases it is interesting to look at two attributes together to understand if they are correlated
  - E.g., how does your marital status relate with tax cheating?
  - E.g., Does refund correlate with average income?
  - Is there a relationship between years of study and income?
- How do we visualize these relationships?

# Plotting attributes against each other

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|---------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

## Confusion or Contingency Matrix

| | No | Yes |
|---|---|---|
| **Single** | 2 | 1 |
| **Married** | 4 | 0 |
| **Divorced** | 2 | 1 |

# Plotting attributes against each other

| | No | Yes |
|---|---|---|
| **Single** | 2 | 1 |
| **Married** | 4 | 0 |
| **Divorced** | 2 | 1 |

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 10000K | **Yes** |
| 6 | No | Married | 60K | **No** |
| 7 | Yes | Divorced | 220K | **No** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 90K | **No** |
| 10 | No | Single | 90K | **No** |

## Joint Distribution Matrix

| | No | Yes |
|---|---|---|
| **Single** | 0.2 | 0.1 |
| **Married** | 0.4 | 0.0 |
| **Divorced** | 0.2 | 0.1 |

| | No | Yes |
|---|---|---|
| **Single** | 0.2 | 0.1 |
| **Married** | 0.4 | 0.0 |
| **Divorced** | 0.2 | 0.1 |

It can also be represented as a heatmap

# Plotting attributes against each other

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

## Joint Distribution Matrix

|  | No | Yes |  |
|---|---|---|---|
| **Single** | 0.2 | 0.1 | 0.3 |
| **Married** | 0.4 | 0.0 | 0.4 |
| **Divorced** | 0.2 | 0.1 | 0.2 |
|  | 0.8 | 0.2 | 1 |

Marginal distribution for Marital Status

Marginal distribution for Cheat

# Plotting attributes against each other

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

How do we know if there are interesting correlations?

## Joint Distribution Matrix P

| | No | Yes | |
|---|-----|-----|-----|
| **Single** | 0.2 | 0.1 | 0.3 |
| **Married** | 0.4 | 0.0 | 0.4 |
| **Divorced** | 0.2 | 0.1 | 0.2 |
| | 0.8 | 0.2 | 1 |

## Independence Matrix E

| | No | Yes | |
|---|------|------|-----|
| **Single** | 0.24 | 0.06 | 0.3 |
| **Married** | 0.32 | 0.08 | 0.4 |
| **Divorced** | 0.16 | 0.04 | 0.2 |
| | 0.8 | 0.2 | 1 |

Compare the values $P_{xy}$ with $E_{xy}$

The product of the two marginal values 0.2*0.8

# Plotting attributes against each other

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

## Joint Distribution Matrix P

| | No | Yes | |
|---|---|---|---|
| Single | 0.2 | 0.1 | 0.3 |
| Married | 0.4 | 0.0 | 0.4 |
| Divorced | 0.2 | 0.1 | 0.2 |
| | 0.8 | 0.2 | 1 |

## Independence Matrix E

| | No | Yes | |
|---|---|---|---|
| Single | 0.24 | 0.06 | 0.3 |
| Married | 0.32 | 0.08 | 0.4 |
| Divorced | 0.16 | 0.04 | 0.2 |
| | 0.8 | 0.2 | 1 |

We can compare specific pairs of values:
- If $P(x,y) > E(x,y)$ there is positive correlation (e.g, Married, No)
- If $P(x,y) < E(x,y)$ there is negative correlation (e.g., Single, No)
- Otherwise there is no correlation

The quantity $\frac{P(x,y)}{E(x,y)} = \frac{P(x,y)}{P(x)P(y)}$ is called Lift, or Pointwise Mutual Information

# Plotting attributes against each other

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |

## Joint Distribution Matrix P

| | No | Yes | |
|---|---|---|---|
| Single | 0.2 | 0.1 | 0.3 |
| Married | 0.4 | 0.0 | 0.4 |
| Divorced | 0.2 | 0.1 | 0.2 |
| | 0.8 | 0.2 | 1 |

## Independence Matrix E

| | No | Yes | |
|---|---|---|---|
| Single | 0.24 | 0.06 | 0.3 |
| Married | 0.32 | 0.08 | 0.4 |
| Divorced | 0.16 | 0.04 | 0.2 |
| | 0.8 | 0.2 | 1 |

Or compare the two attributes:

Pearson $x^2$ Independence Test Statistic:

$$U = N \sum_x \sum_y \frac{(P_{xy} - E_{xy})^2}{E_{xy}}$$

# Hypothesis testing

- How important is the statistic value $U$ that we computed?
- Formulate a null hypothesis $H_0$:
  - $H_0$ = the two attributes are independent
- Compute the distribution of the statistic $U$ in the case that $H_0$ is true
  - In this case we can show that the statistic $U$ follows a $\chi^2$ distribution
- For the statistic value $U = \theta$ we observe in our data, compute the probability $P(U \geq \theta)$ under the null hypothesis
  - For most distributions there are tables that give these numbers for our data
- This is the p-value of our experiment:

> The p-value is the probability under $H_0$ (independence) of observing a value of the test statistic the same as, or more extreme than the one that was actually observed

- We want it to be small (ideally $\leq 0.01$ , $\leq 0.05$ is good , $\leq 0.1$ is ok)
  - This means that the observed value is interesting

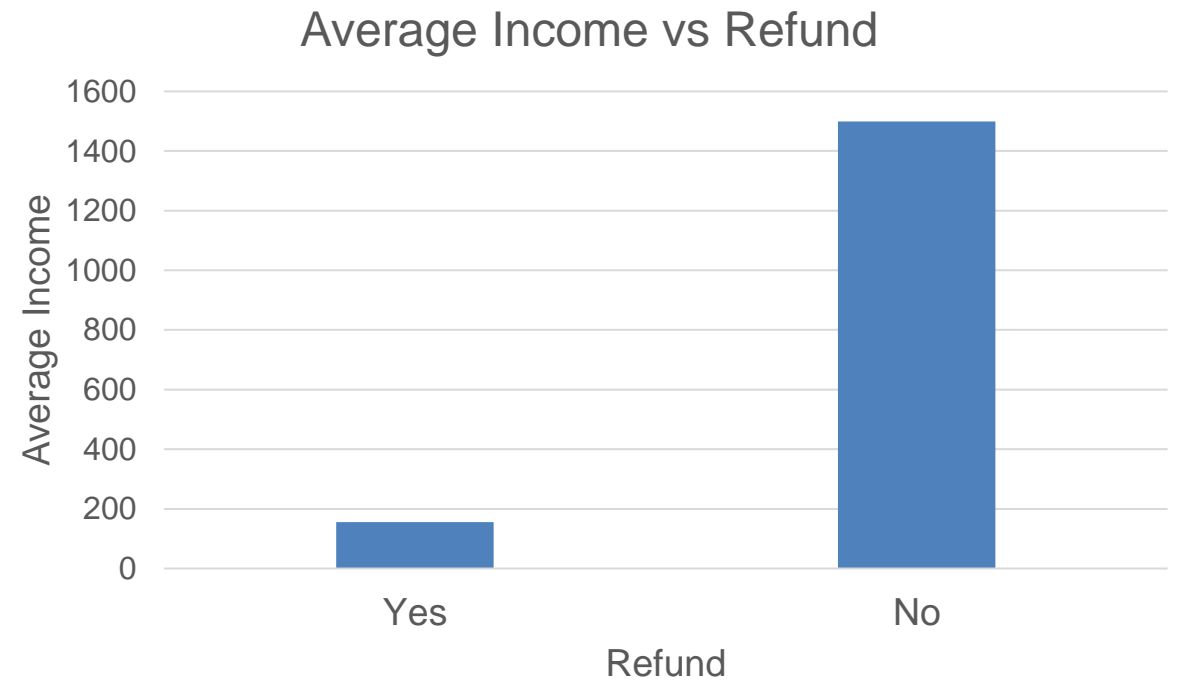# Hypothesis Testing and P-values – A simple example

- A coin is tossed 20 times, and we get 16 heads.

- Hypothesis $H_1$ ="The coin is not fair"

- Null Hypothesis $H_0$ = "The coin is fair" (probability 50% for head)

- p-value: What is the probability of getting a number of heads that is the same or more extreme than 16?

  - One-sided p-value: $\Pr(H \geq 16) = 0.0059$

  - Two-sided p-value: $\Pr(H \geq 16) + \Pr(H \leq 4) = 0.0118$

- With significance level $\alpha = 0.05$ we can conclude that we can reject the null hypothesis

# P-values

- The p-value tells us the probability that the value we observe could appear in data generated under the null hypothesis.
  - The null hypothesis proposes a (random) model for the data generation
  - The p-value answers the question: "If the null hypothesis model was correct how unlikely would it be to observe the value we observe"?
- Be careful!
  - A p-value $\phi$ does not mean that the null hypothesis is correct with probability $\phi$
    - A high p-value (e.g., 90%) does not mean that the null hypothesis is true, it only means that the data is consistent with the model of the null hypothesis
  - A p-value $\phi$ does not mean that our hypothesis is correct with probability $1 - \phi$
    - A p-value of 3% does not mean that our hypothesis is correct with probability 97%
    - It only means that the data is not consistent with the null hypothesis random model

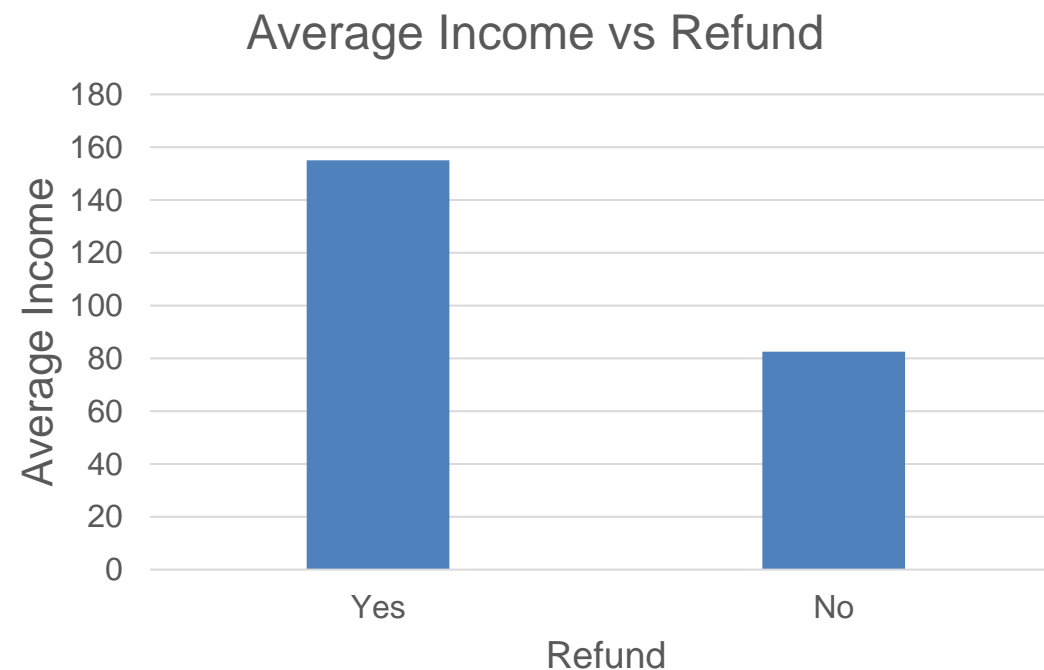# Categorical and numerical attributes

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 10000K | Yes |
| 6 | No | NULL | 60K | No |
| 7 | Yes | Divorced | 220K | NULL |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 90K | No |
| 10 | No | Single | 90K | No |



Average Income vs Refund

# Categorical and numerical attributes

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|-----|--------|----------------|----------------|-------|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 10000K | **Yes** |
| 6 | No | NULL | 60K | **No** |
| 7 | Yes | Divorced | 220K | **NULL** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 90K | **No** |
| 10 | No | Single | 90K | **No** |

After removing the outlier value



Average Income vs Refund

How informative are the means?

# Categorical and numerical attributes

| Tid | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | **No** |
| 2 | No | Married | 100K | **No** |
| 3 | No | Single | 70K | **No** |
| 4 | Yes | Married | 120K | **No** |
| 5 | No | Divorced | 10000K | **Yes** |
| 6 | No | NULL | 60K | **No** |
| 7 | Yes | Divorced | 220K | **NULL** |
| 8 | No | Single | 85K | **Yes** |
| 9 | No | Married | 90K | **No** |
| 10 | No | Single | 90K | **No** |

Compute error bars



Average Income vs Refund

Error bars give a measure of the variability of the mean

# Error bars

- Error bars may be:
  - The range
  - The standard deviation
  - The standard error
  - The 95% confidence interval

<span style="color:red">Descriptive</span> error bars: They tell us something about the underlying distribution of the data

<span style="color:red">Inferential</span> error bars: They tell us something about the quality of the estimation of the mean

- Inferential error bars get more informative the more data we collect.

- We should always specify what the error bars mean in a plot.

# Standard Error (of the Mean)

- The Standard Error (SE) is usually defined for the mean of a sample of values $X$ (it is also known as SEM – Standard Error of the Mean)
- It is defined as:

$$se = \frac{\hat{\sigma}(X)}{\sqrt{n}}$$

where $\hat{\sigma}(X)$ = empirical standard deviation.

- As the sample size grows the SE is reduced (we have a better estimation of the mean)

- Computation follows from the fact that

$$se = \hat{\sigma}(\hat{\mu}), \hat{\mu} = \frac{1}{n}\sum_i X_i$$

- We assume that $X_i$ are independent samples of the random variable $X$ that come from the same distribution. We use the fact that:

$$Var(\sum_i \alpha_i X_i) = \sum_i \alpha_i^2 \, Var(X_i)$$

# Confidence interval

- We want to estimate the average income $\mu$ which is a fixed value.
- We have a sample of the population and the measurements $\{X_i\}$ of incomes and we estimate the average income as:

$$\hat{\mu} = \frac{1}{n}\sum_i X_i$$

- The $p$-confidence interval of the value $\mu$ is an interval of values $C_n$ such that

$$P(\mu \in C_n) \geq p$$

  - We usually ask for the 95% confidence interval
- Important: The probability is taken over the many different samples of the population
  - Different samples will generate different confidence intervals
  - There is a 95% chance that each of these intervals contains the true mean $\mu$
  - It is incorrect to say that this is the probability that $\mu$ belongs to the interval
- The value $\hat{\mu}$ follows a normal distribution for large $n$. For normal distributions the 95% confidence interval (for large enough $n$) is:

$$(\hat{\mu} - 2se, \hat{\mu} + 2se)$$

# Example

- If we obtain an estimate of the mean for 20 different population samples, we will obtain 20 different 95%-confidence intervals.

- We expect that 1/20 of these intervals will not contain the true mean (the dotted line)
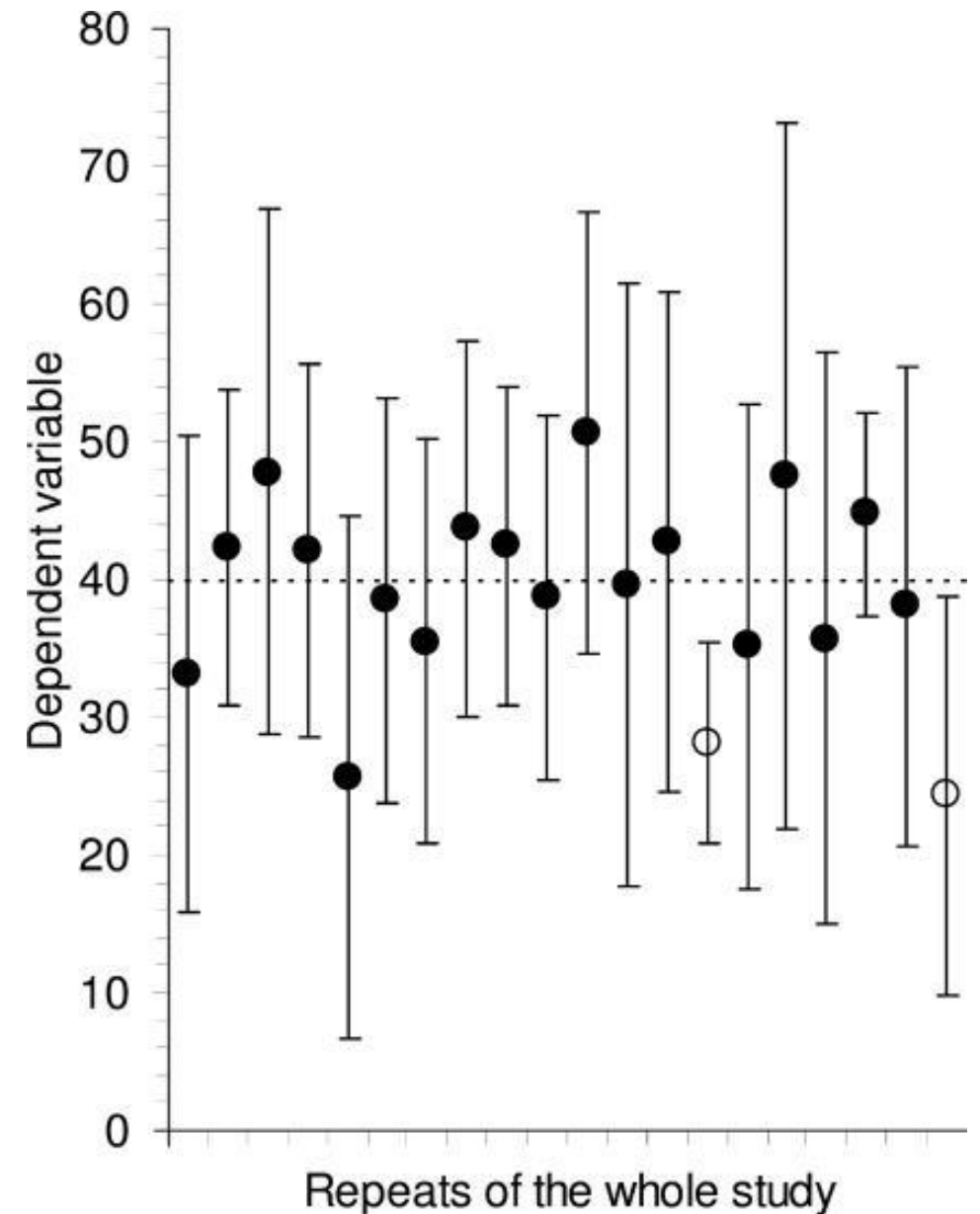
# Error bars example

- The different error bars and how they change as the sample size increases

- Out of the four different error bars, the confidence interval is probably the most informative.
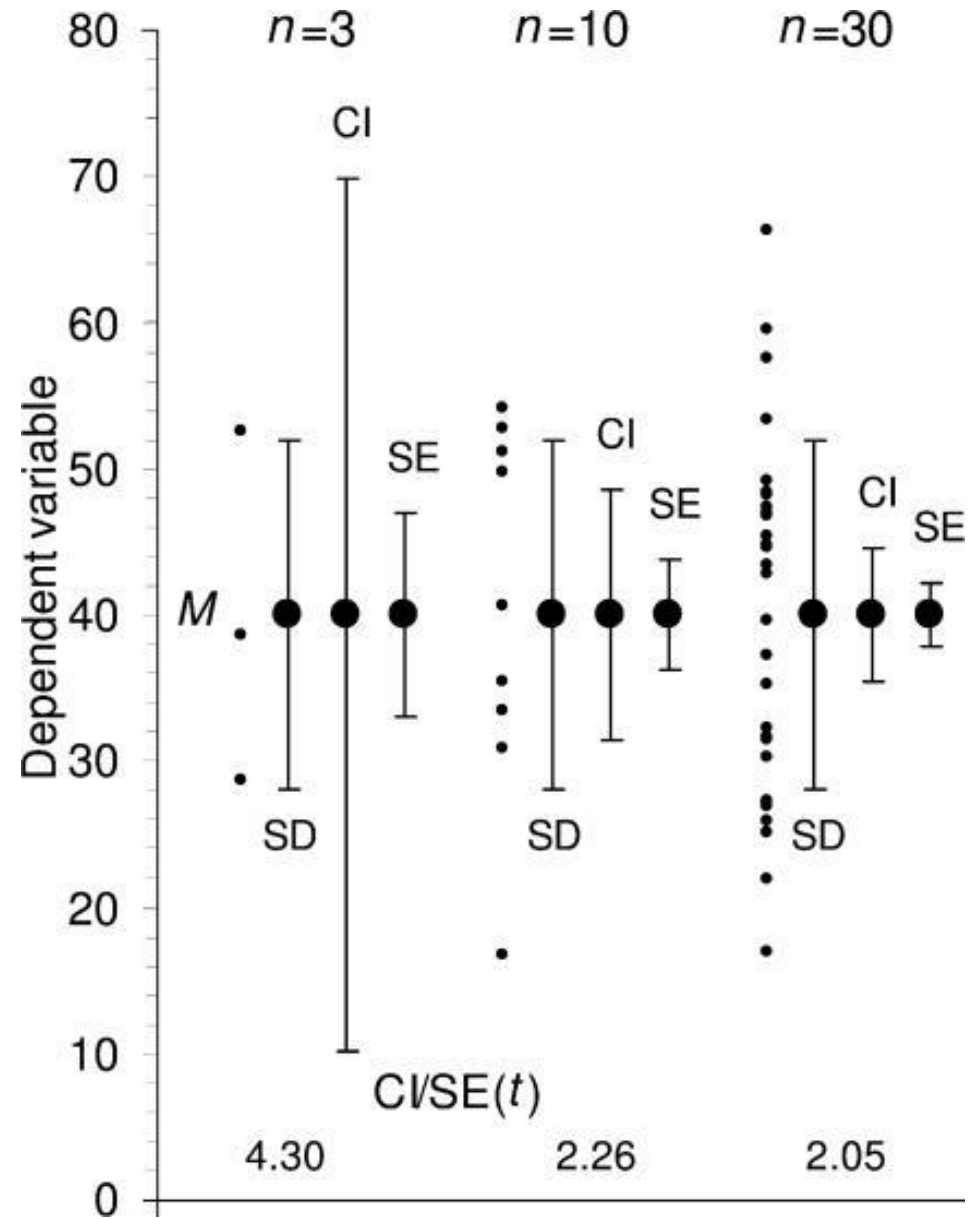
# Statistical significance

- Given the means of two populations an important question is whether the difference we observe is <span style="color:red">statistically significant</span>

- Statistical significance is estimated by computing a <span style="color:red">p-value</span> with respect to a <span style="color:red">null hypothesis</span>

- The value is compared to a <span style="color:red">significance level $\alpha$</span> which is usually set to 0.05 (or 0.01)

Average Income vs Refund
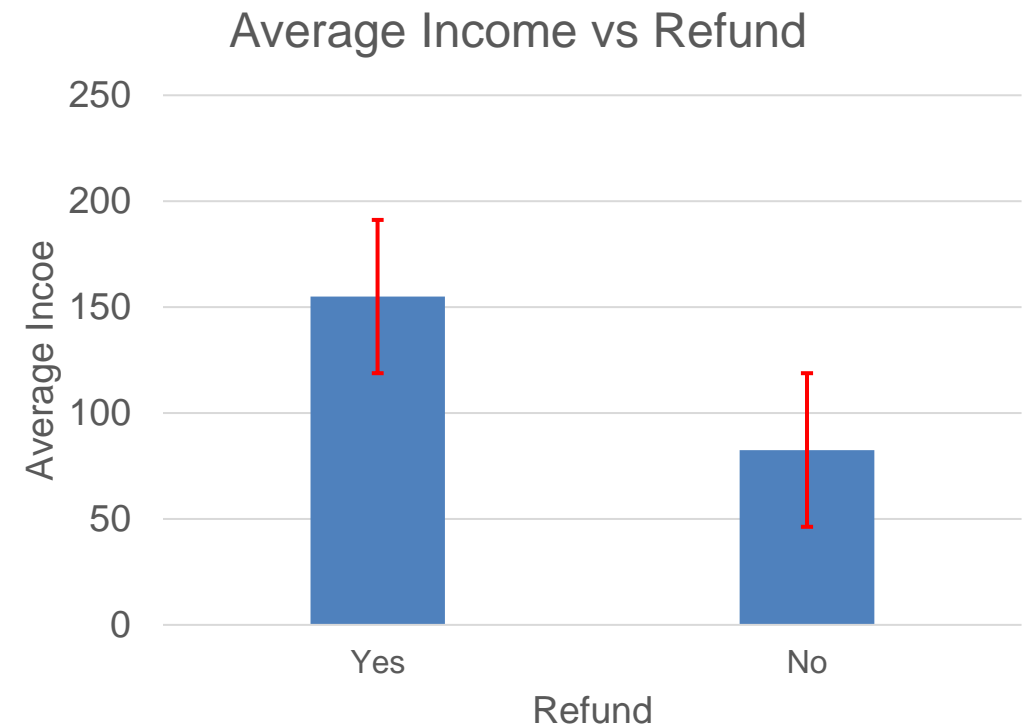
# Statistical significance via error bar overlap

- It is <span style="color:red">not</span> always safe to declare that there is statistical significance when error bars do not overlap
  - We may have statistically significant differences when there is overlap, or no statistical significance when there is no overlap
- We can say that there is statistically significant difference of means when sample sizes are comparable and the 95%-confidence intervals do not overlap
- There are a little more complex rules for the standard error.

# Statistical tests

- Statistical tests measure specific values and determine their statistical significance
  - For example measure the importance of the difference between the means (e.g., average grade) of two populations (e.g., students in cities vs students in rural areas).
- The magnitude of the value that is measured is also called the effect size
- The statistical significance of this value is measured with respect to a null hypothesis
  - For example: the difference of the means is zero
- The statistical test assumes a random model for the underlying data
  - For example, the data are generated by a Gaussian distribution
- The statistical test produces a p-value for the statistical significance of the values we observe

# Statistical tests – The Student t-test

- The Student t-test tests if the difference of the means of two samples is "big enough"

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\dfrac{\sigma_X^2}{N_X} + \dfrac{\sigma_Y^2}{N_Y}}}$$

- Large t-value (effect size):
  - Large difference between the means
  - Small variance in the samples (more accurate measurements)
  - Large sample sizes (more reliable)

# Statistical tests – The Student t-test

- The Student t-test produces a p-value: Measures the probability of the null hypothesis that the two distributions have zero difference in mean
  - This is what we care about, the t-value is usually not looked at
- Student t-test assumptions:
  - (near) Gaussian distribution of the data,
  - (near) same variance,
  - similar sample sizes.
- There is paired and unpaired Student t-test
  - Example of paired: behavior before and after a treatment.

# Statistical tests – The KS-test
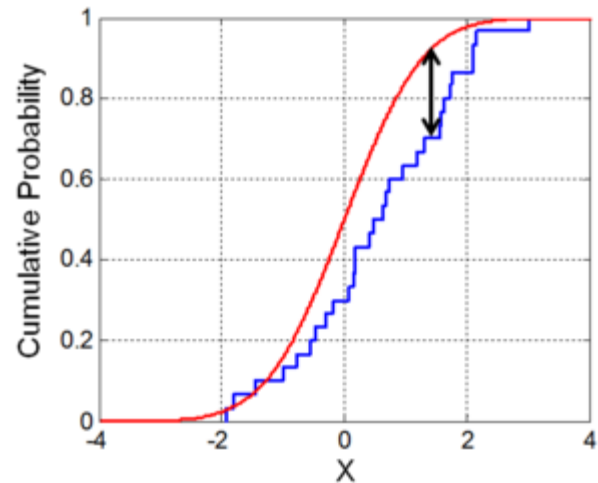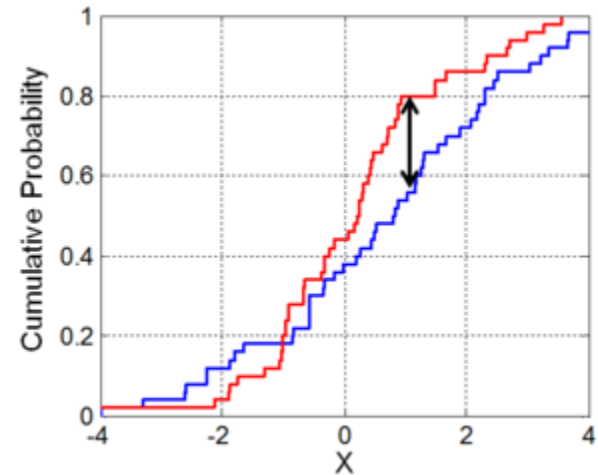
- The Kolomogorov-Smirnov (KS) test, tests if two samples come from the same distribution (or come from a specific distribution)
  - Take the cumulative distribution function (CDF) of the two distributions
  - Compute:

$$D(C_1, C_2) = \max_x |C_1(x) - C_2(x)|$$

- We can reject the null hypothesis if:

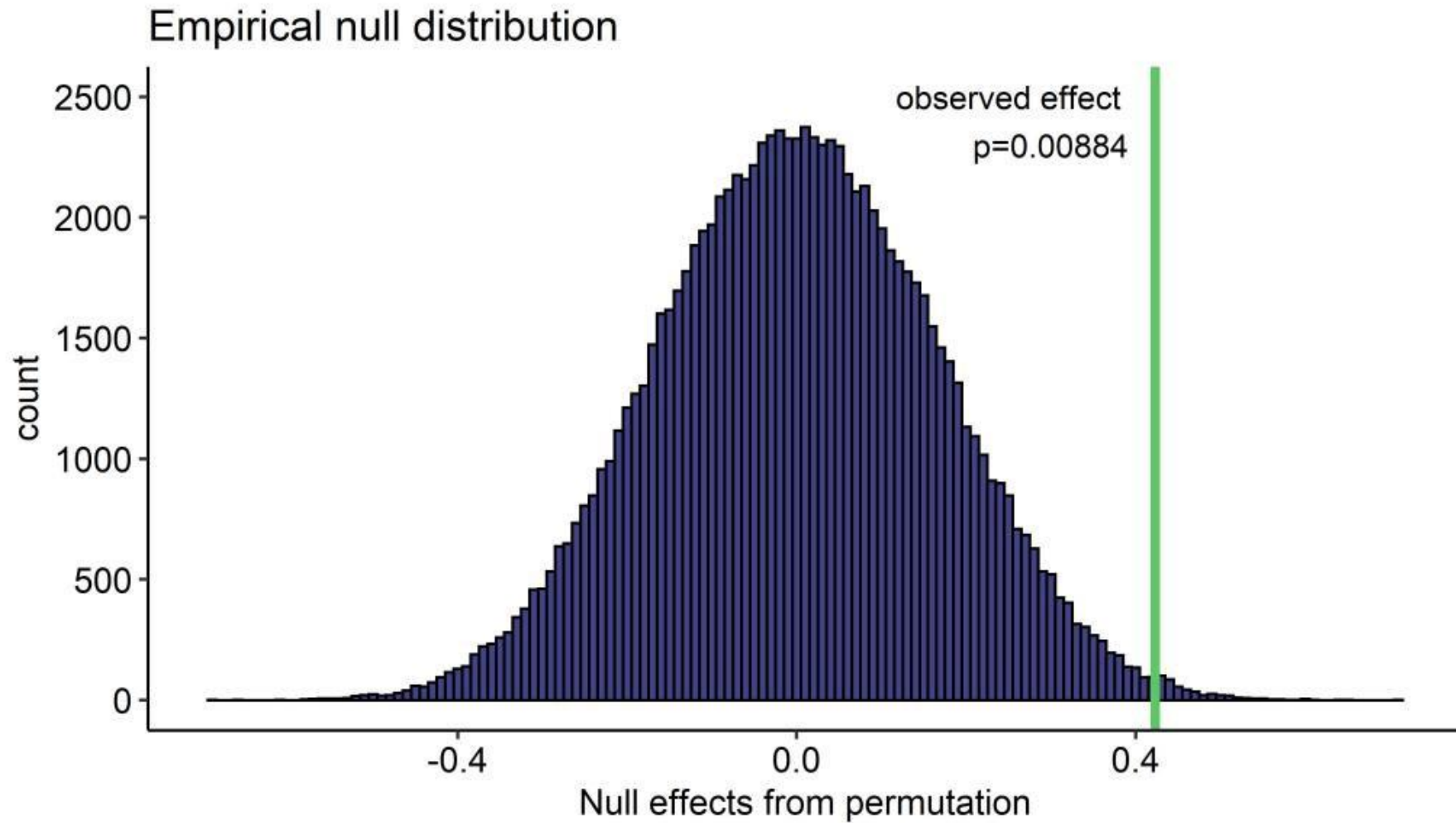$$D(C_1, C_2) > c(\alpha)\sqrt{\frac{N_1 + N_2}{N_1 N_2}}$$

  - $\alpha$ is the confidence level, $c(\alpha)$ is given by some tables

# Statistical tests – Permutation testing

- Most tests make some assumption about the underlying distribution of the data.

- A non-parametric statistical test is the permutation test

- Create random instances of the data by randomly permuting values
  - E.g., permute the Cheat labels randomly

- Compute a statistic of interest for the permuted data
  - E.g., the average income of the cheaters

- Repeat this several times (at least 1000)

- Compute the empirical p-value: the fraction of permutations where we have a value that is equal or more extreme than the one observed.
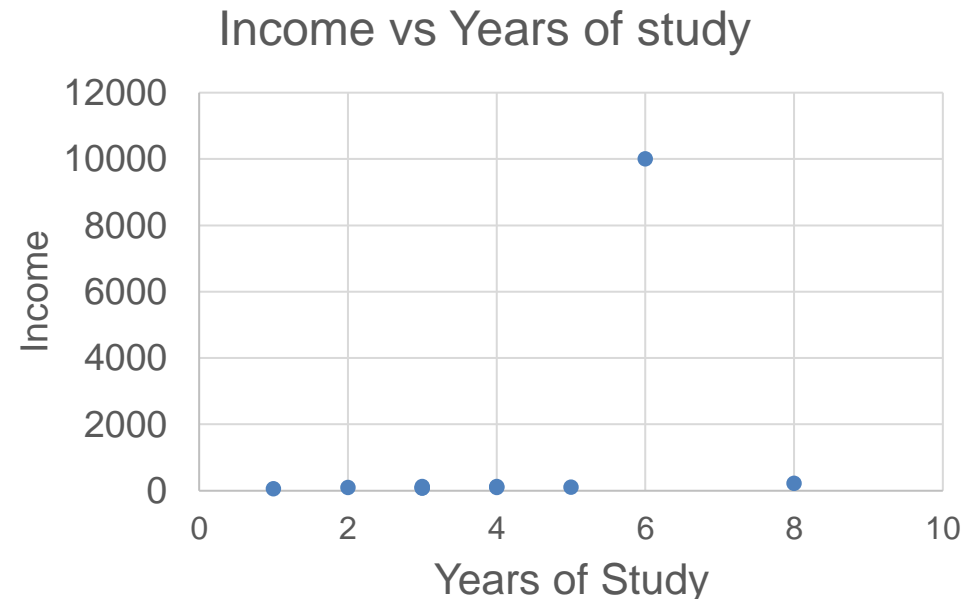
# Example



Empirical null distribution

observed effect
p=0.00884

count

2500
2000
1500
1000
500
0

-0.4          0.0          0.4

Null effects from permutation

# Correlating numerical attributes

| Tid | Refund | Marital Status | Taxable Income | Years of Study |
|-----|--------|----------------|----------------|----------------|
| 1 | Yes | Single | 125K | 4 |
| 2 | No | Married | 100K | 5 |
| 3 | No | Single | 70K | 3 |
| 4 | Yes | Married | 120K | 3 |
| 5 | No | Divorced | 10000K | 6 |
| 6 | No | NULL | 60K | 1 |
| 7 | Yes | Divorced | 220K | 8 |
| 8 | No | Single | 85K | 3 |
| 9 | No | Married | 90K | 2 |
| 10 | No | Single | 90K | 4 |

Scatter plot:

X axis is one attribute, Y axis is the other
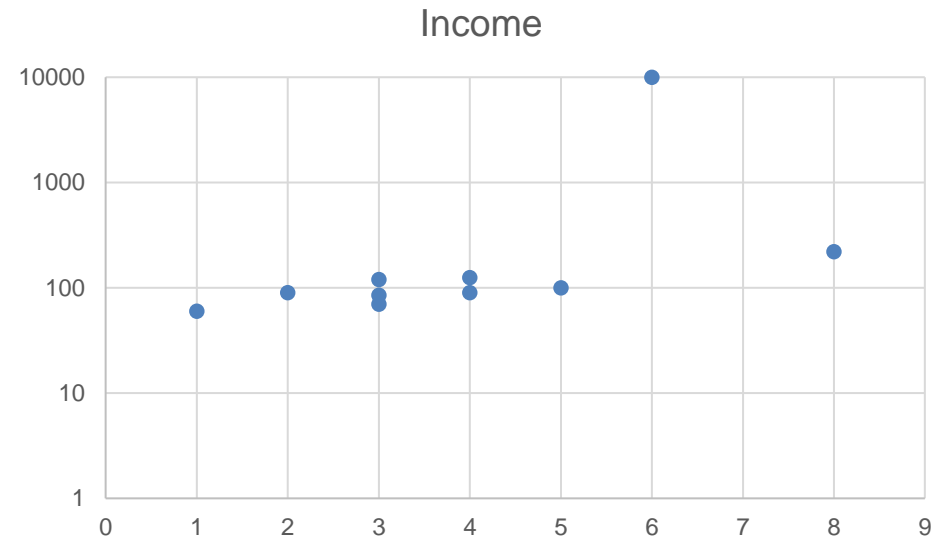For each entry we have two values
Plot the entries as two-dimensional points



Income vs Years of study

# Correlating numerical attributes

| Tid | Refund | Marital Status | Taxable Income | Years of Study |
|-----|--------|----------------|----------------|----------------|
| 1 | Yes | Single | 125K | 4 |
| 2 | No | Married | 100K | 5 |
| 3 | No | Single | 70K | 3 |
| 4 | Yes | Married | 120K | 3 |
| 5 | No | Divorced | 10000K | 6 |
| 6 | No | NULL | 60K | 1 |
| 7 | Yes | Divorced | 220K | 8 |
| 8 | No | Single | 85K | 3 |
| 9 | No | Married | 90K | 2 |
| 10 | No | Single | 90K | 4 |

Scatter plot:
X axis is one attribute, Y axis is the other
For each entry we have two values
Plot the entries as two-dimensional points

Log-scale in y-axis makes the plot look a little better



Income

# Plotting attributes against each other

| Tid | Refund | Marital Status | Taxable Income | Years of Study |
|-----|--------|----------------|----------------|----------------|
| 1 | Yes | Single | 125K | 4 |
| 2 | No | Married | 100K | 5 |
| 3 | No | Single | 70K | 3 |
| 4 | Yes | Married | 120K | 3 |
| 5 | No | Divorced | 10000K | 6 |
| 6 | No | NULL | 60K | 1 |
| 7 | Yes | Divorced | 220K | 8 |
| 8 | No | Single | 85K | 3 |
| 9 | No | Married | 90K | 2 |
| 10 | No | Single | 90K | 4 |

Scatter plot:

X axis is one attribute, Y axis is the other
For each entry we have two values
Plot the entries as two-dimensional points

After removing the outlier value there is a clear correlation



Income vs Years of Study
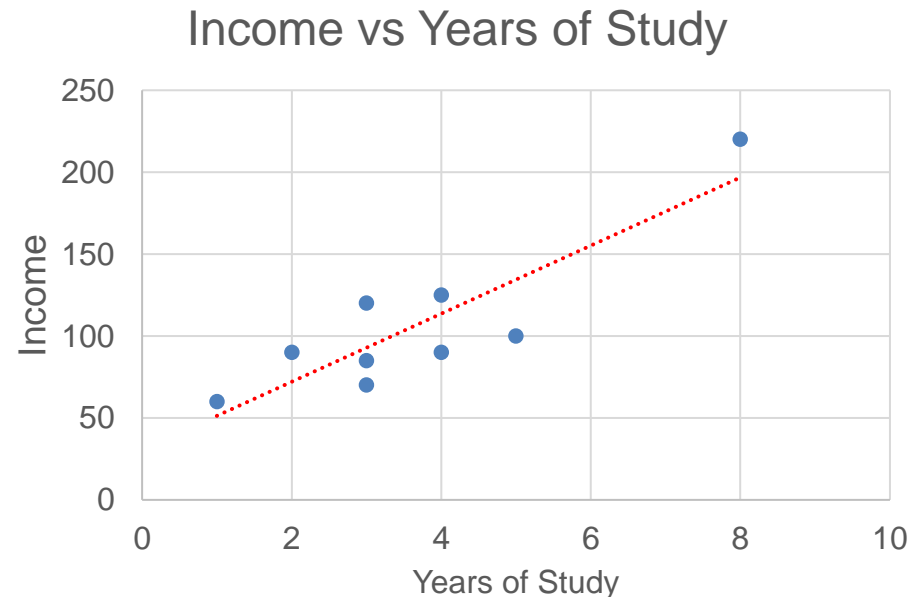
# Scatter Plot Array of Iris Attributes
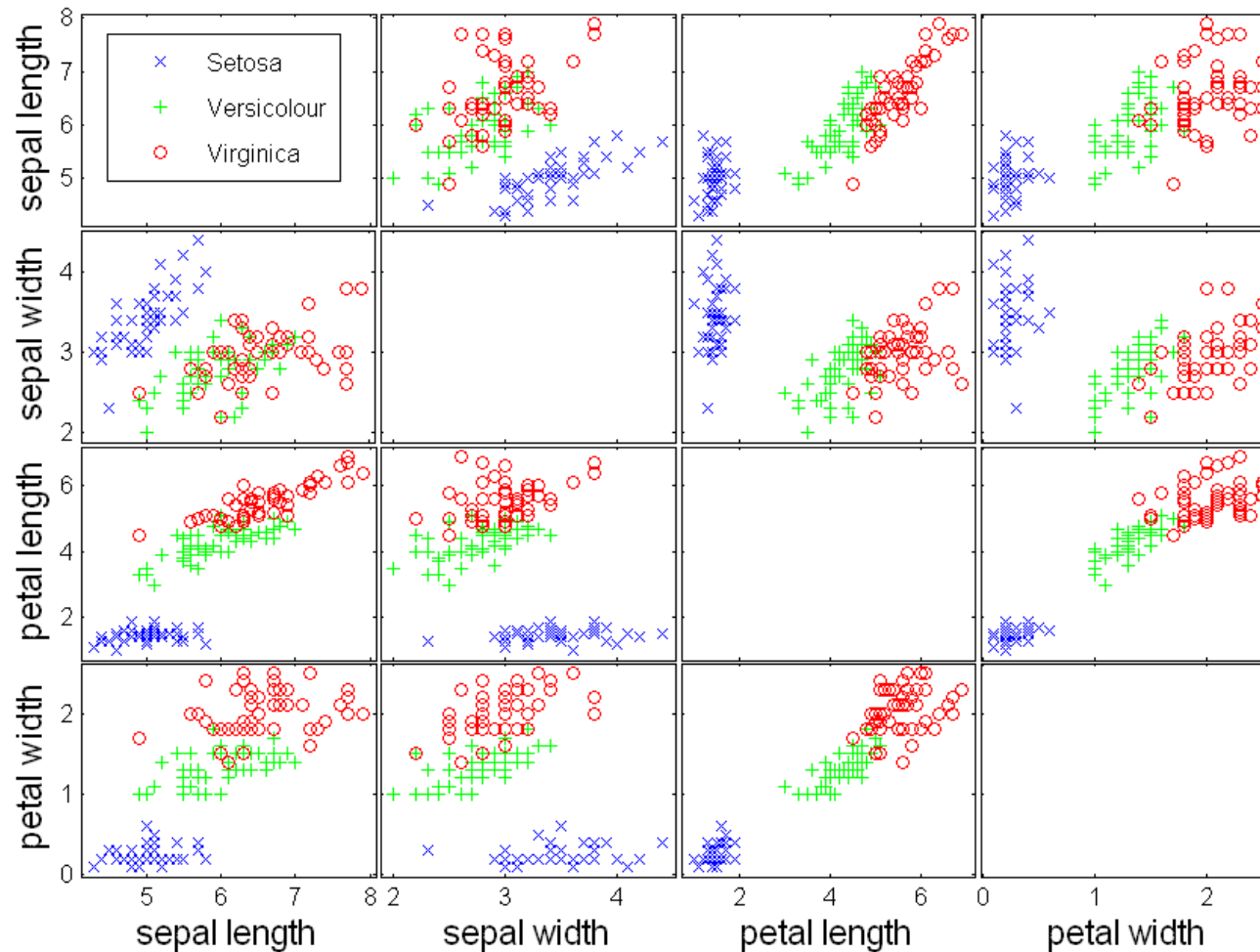
# Measuring correlation

- Pearson correlation coefficient: measures the extent to which two variables are linearly correlated
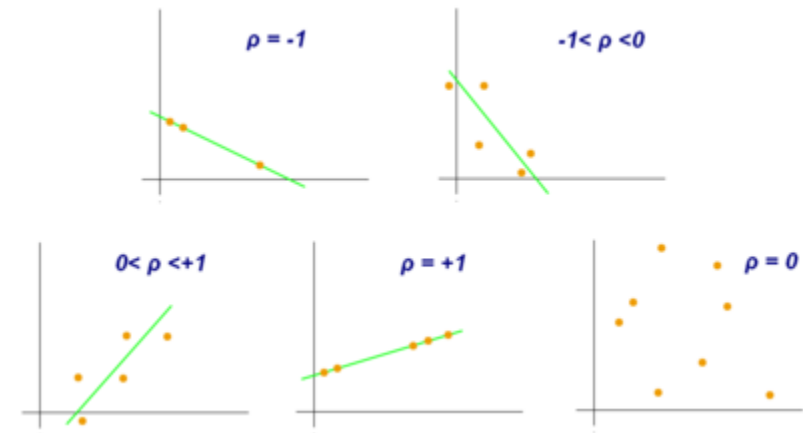  - $X = \{x_1, \ldots, x_n\}$
  - $Y = \{y_1, \ldots, y_n\}$

  Must have pairs of observations

  - $corr(X, Y) = \dfrac{\sum_i (x_i - \mu_X)(y_i - \mu_Y)}{\sqrt{\sum_i (x_i - \mu_X)^2} \sqrt{\sum_i (y_i - \mu_Y)^2}}$



- It comes with a p-value
  - The p-value is the probability that the correlation was by chance.

# Pearson correlation

- Assumptions:
  - Variables are normally distributed
  - No outliers
  - A linear relationship between the variables
- Caveats
  - For large samples p-values will always be small
  - Except for the p-value we need to also look at the effect size: the value of $r = corr(X, Y)$
- Interpretation
  - The value of $r^2$ measures the fraction of variance in one variable that is explained by the values of the other variable (shared variance)

# Rank correlation

- Spearman rank correlation coefficient: tells us if two variable are rank-correlated
  - They place items in the same order – Pearson correlation of the rank vectors
  - For ranking without ties it looks at the differences between the ranks of the same items
- Spearman coefficient also comes with a p-value
- Spearman coefficient does not assume a linear relationship, but a monotonic one

# Statistical significance vs Scientific significance

- Statistics place a lot of emphasis on the <span style="color:red">p-values</span> and the statistical significance

- However, p-values may be small but the finding to not be of <span style="color:red">scientific interest</span>

  - A difference or a correlation may be statistically significant, but too small to be of scientific interest

  - We need to evaluate the results beyond simply looking at the p-values.

  - We also need to look at the <span style="color:red">effect size</span>, or the impact of the computed difference.

# Plotting attributes together

| City | Product 1 | Product 2 |
|------|-----------|-----------|
| New York | 100 | 60 |
| Chicago | 70 | 150 |
| San Francisco | 30 | 80 |
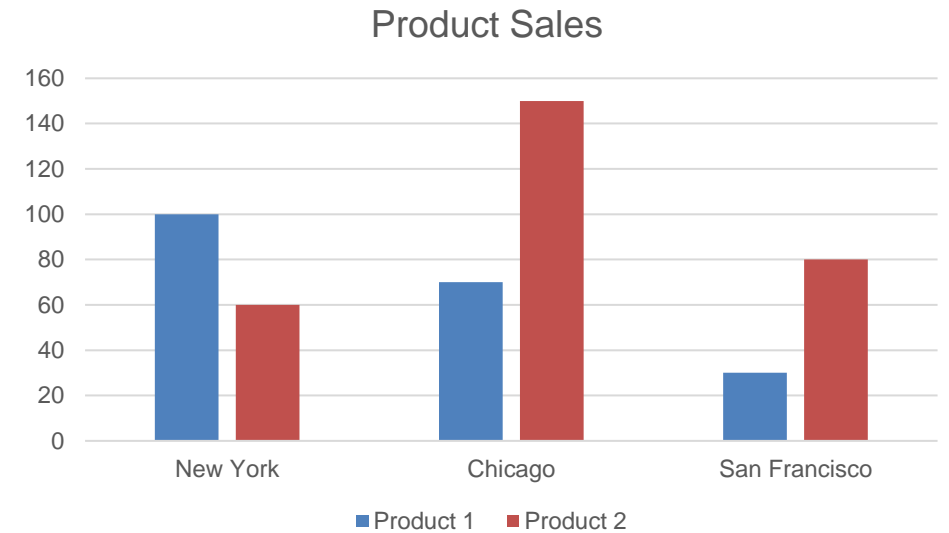


Product Sales

How would you visualize the differences between the product sales per city?

# Plotting attributes together

| Year | Product 1 | Product 2 |
|------|-----------|-----------|
| 2011 | 100 | 200 |
| 2012 | 200 | 250 |
| 2013 | 180 | 300 |
| 2014 | 300 | 350 |
| 2015 | 500 | 490 |
| 2016 | 600 | 500 |
| 2017 | 650 | 550 |
| 2018 | 640 | 540 |
| 2019 | 700 | 500 |
| 2020 | 200 | 100 |

How would you visualize the differences between the product sales over time?



Product Sales