Data Analitiği

Keşfedici Veri Analizi

(Exploratory Data Analysis)

“Dr. Cahit Karakuş”

Outline

- Exploratory Data Analysis
  - Grafik türleri
  - Bazı önemli dağıtımlar
  - Hipotez testi

Tanımlayıcı ve Çıkarımsal İstatistikler
(Descriptive vs. Inferential Statistics)

- **Descriptive:** e.g., Median; describes data you have but can't be generalized
- **Inferential:** e.g., t-test, that enable inferences about the population beyond our data these are the techniques we’ll leverage for Machine Learning and Predictionç
- Çıkarımsal (Inferential): T-testi örneğinde olduğu gibi verilerimizin ötesinde popülasyon hakkında çıkarımlar sağlayan Makine Öğrenimi ve Tahmin için yararlanılan tekniklerdir.

Examples of Business Questions

- **Basit (açıklayıcı) İstatistikler:** "En karlı müşteriler kimlerdir?"
- **Hipotez testi:** "Bu müşterilerin şirket için bir değer farkı var mı?"
- **Segmentasyon/Sınıflandırma:** Bu müşterilerin ortak özellikleri nelerdir?
- **Tahmin**: Yeni müşteri karlı bir müşteri olacak mı? Eğer öyleyse, ne kadar karlı?

- **Simple (descriptive) Stats**
  - "Who are the most profitable customers?"
- **Hypothesis Testing**
  - "Is there a difference in value to the company of these customers?"
- **Segmentation/Classification**
  - What are the common characteristics of these customers?
- **Prediction**
  - Will this new customer become a profitable customer?   If so, how profitable?

Applying techniques

- İşle ilgili soruların çoğu nedenseldir: Olsaydı ne olurdu? (ör. bir reklam gösteriliyor)
- Ama ilişkisel sorular sormak daha kolay (geçmişte bu reklam gösterildiğinde ne oldu).
- Denetimli Öğrenme: Sınıflandırma ve Regresyon
- Denetimsiz Öğrenme: Kümeleme ve Boyut azaltma
- Not: Denetimsiz Öğrenme genellikle daha büyük denetimli öğrenme probleminde kullanılır. Örneğin, görüntü tanıma sinir ağları için otomatik kodlayıcılar.

- Most business questions are causal: what would happen if? (e.g. I show this ad)
- But its easier to ask correlational questions, (what happened in this past when I showed this ad).
- **Supervised Learning:**
  - Classification and Regression
- **Unsupervised Learning:**
  - Clustering and Dimension reduction
- Note: Unsupervised Learning is often used inside a larger Supervised learning problem.
  - E.g. auto-encoders for image recognition neural nets.

Applying techniques

- **Supervised Learning:**
  - kNN (k Nearest Neighbors)
  - Naïve Bayes
  - Logistic Regression
  - Support Vector Machines
  - Random Forests
- **Unsupervised Learning:**
  - Clustering
  - Factor analysis
  - Latent Dirichlet Allocation

Data Presentation

- Data Art

facebook
December 2010

The “R” Language

- EDA için Bell laboratuvarlarında geliştirilen “S” dilinin bir evrimi.
- Fikir, verilerin etkileşimli olarak keşfedilmesine ve görselleştirilmesine izin vermekti.
- Diğer birçok veri bilimci tarafından kullanılan istatistikçiler için tercih edilen dil.
- Özellikleri:Muhtemelen en kapsamlı istatistiksel modeller ve dağılımlar koleksiyonu.
- CRAN: çok büyük bir açık kaynak istatistiksel model kaynağı.

Chart types

- Single variable
  - Dot plot
  - Jitter plot
  - Error bar plot
  - Box-and-whisker plot
  - Histogram
  - Kernel density estimate
  - Cumulative distribution function

(note: examples using qplot library from R)

Chart types

- **Dot plot**

```
> f500.ca <- subset(f500, state_location == "CA")
> f500.ca$state_location <- factor(f500.ca$state_location)
> qplot(revenues, state_location, data=f500.ca)
```

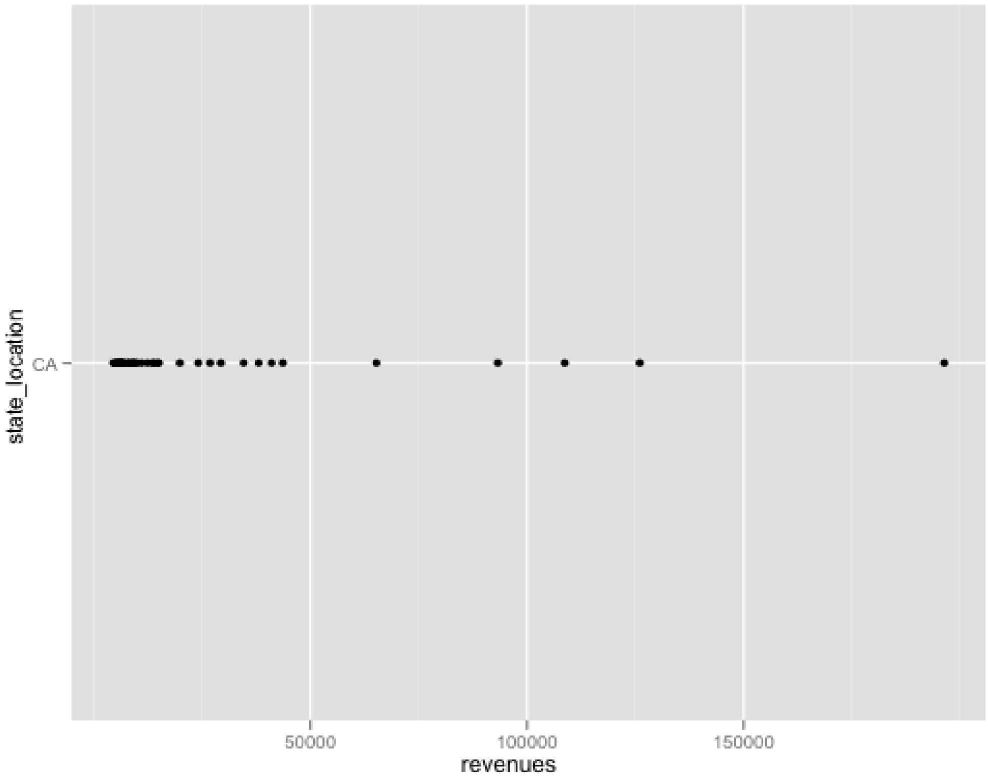
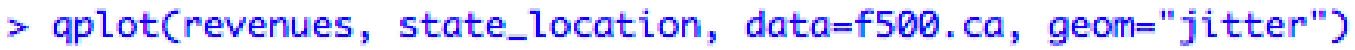


Chart types

- **Jitter plot**



```
> qplot(revenues, state_location, data=f500.ca, geom="jitter")
```

- Noktaları yaymak için y eksenine eklenen gürültü

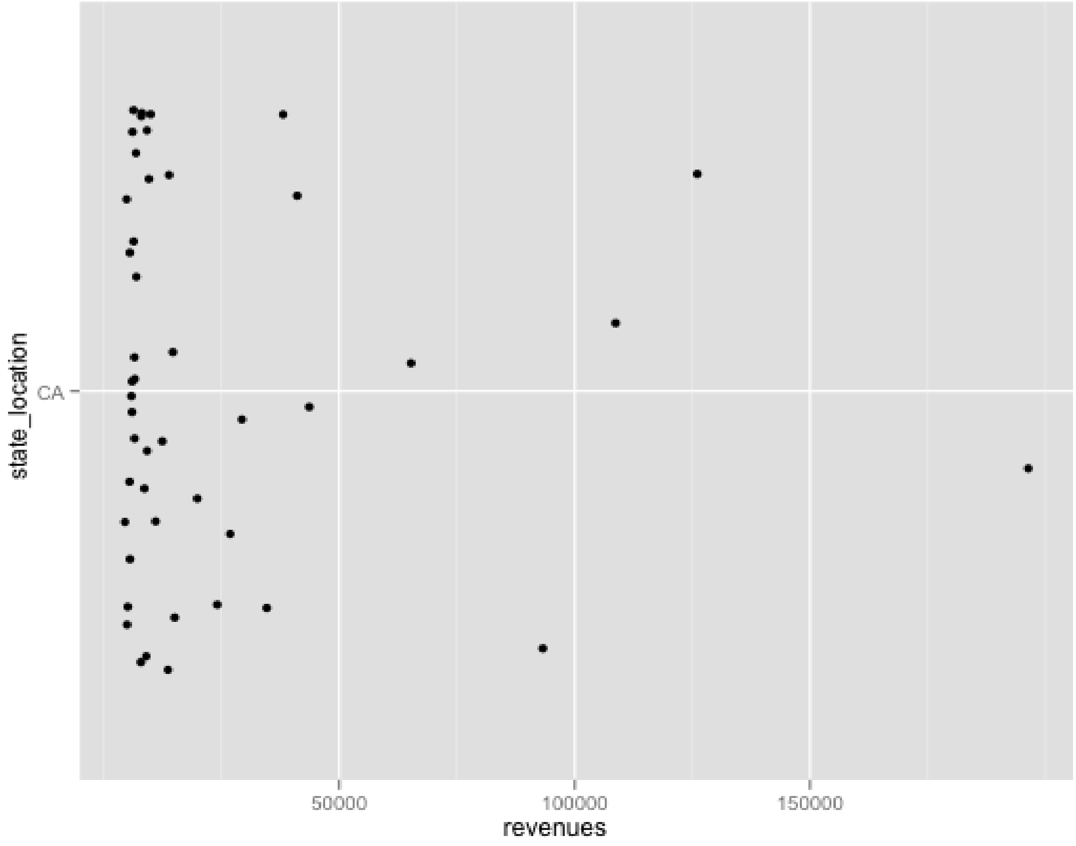


Chart types

- **Error bars:** usually based on confidence intervals (CI). 95% CI means 95% of points are in the range, so 2.5% of points are above or below the bar.
- Not necessarily symmetric:

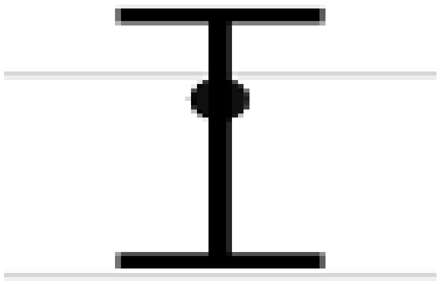


Chart types

- **Box-and-whisker plot** : a graphical form of 5-number summary (Tukey)

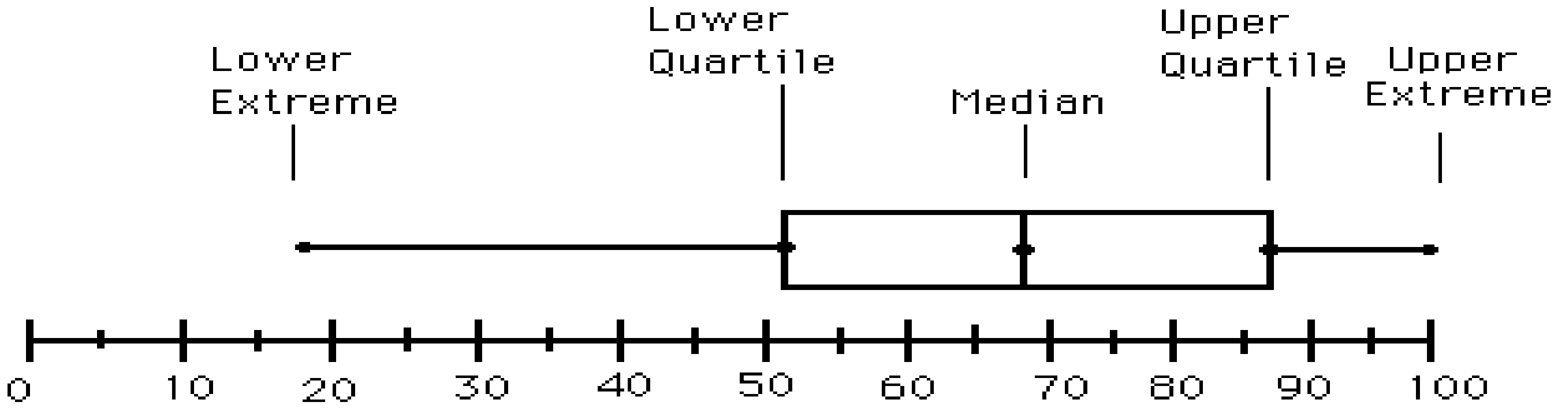
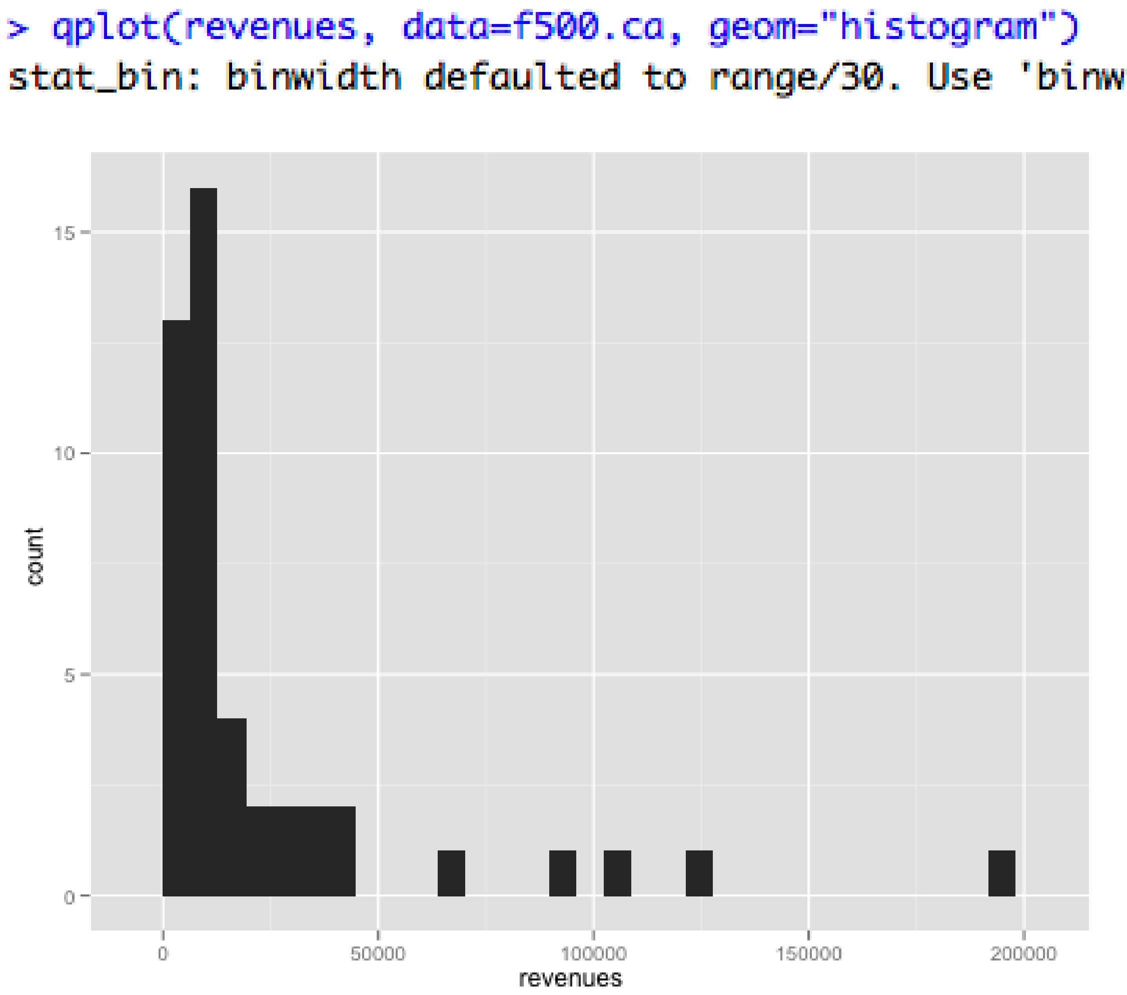


Chart types

- **Histogram**



```
> qplot(revenues, data=f500.ca, geom="histogram")
stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

Chart types

- **Kernel density estimate**

```
> qplot(revenues, data=f500.ca, geom="density")
```

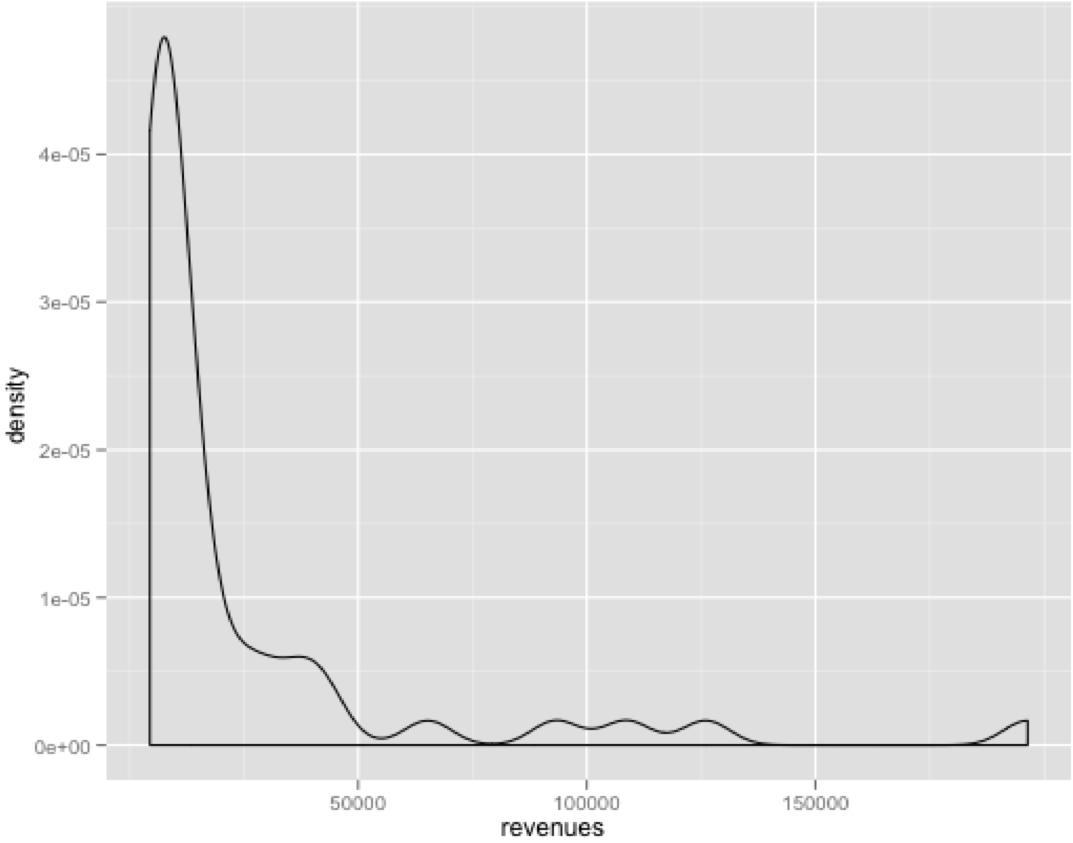


Chart types

- Histogram and Kernel Density Estimates
  - Histogram
    - Proper selection of bin width is important
    - Outliers should be discarded
  - KDE (like a smooth histogram)
    - Kernel function
      - Box, Epanechnikov, Gaussian
    - Kernel bandwidth

Chart types

- **Cumulative distribution function**
- Integral of the histogram – simpler to build than KDE (don’t need smoothing)

```
> plot(ecdf(f500.ca$revenues))
```

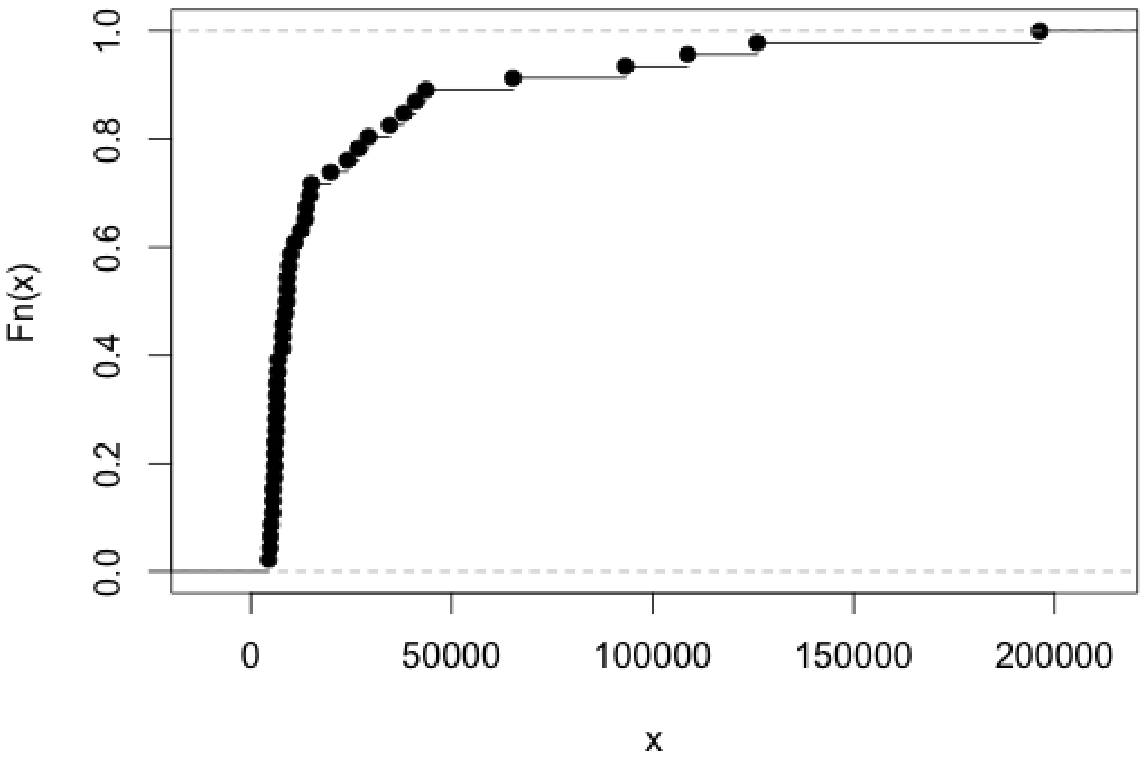


Chart types

- Two variables
    - Bar chart
    - Scatter plot
    - Line plot
    - Log-log plot

Chart types

- **Bar plot:** one variable is discrete

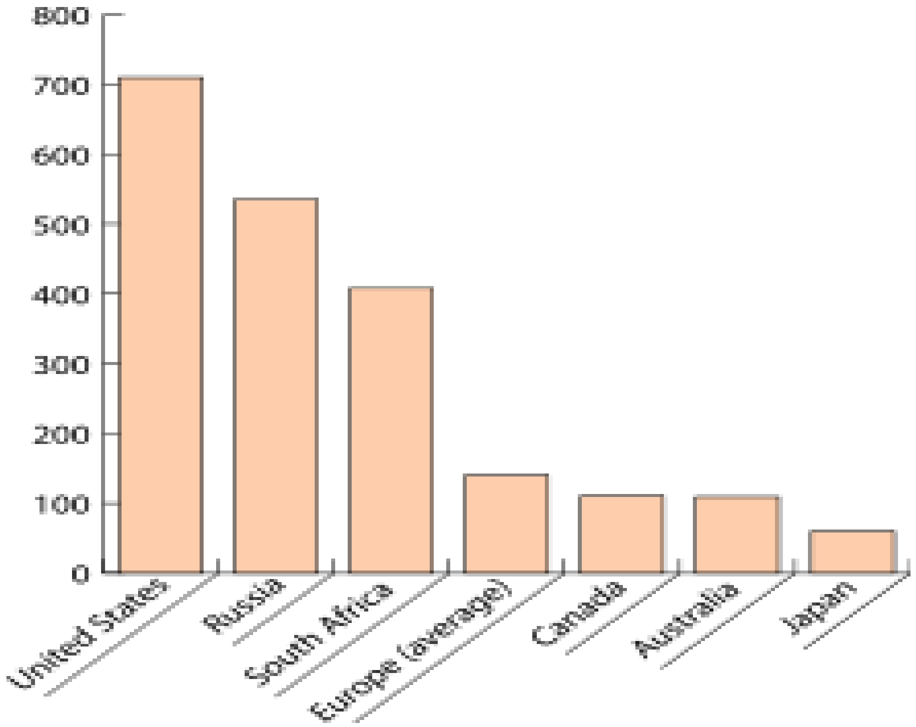


Chart types

- **Scatter plot**

```
> qplot(revenues, profits, data=f500)
```

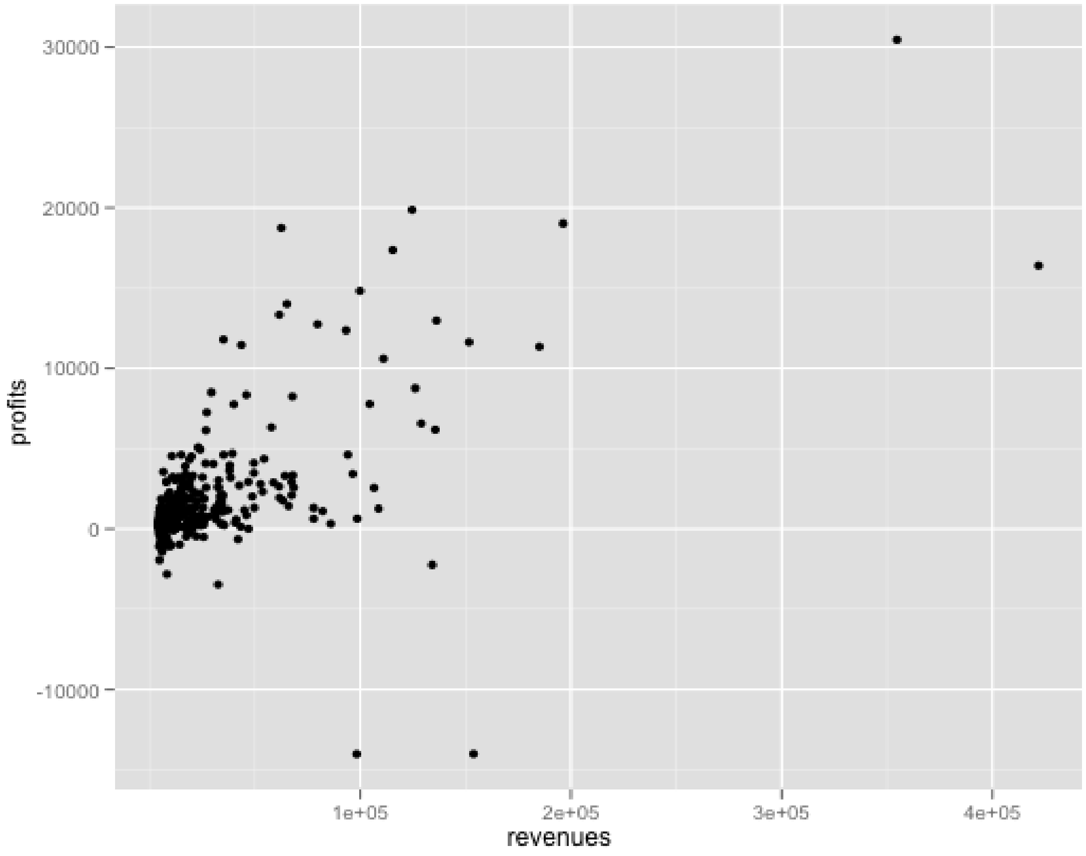


Chart types

- **Line plot**

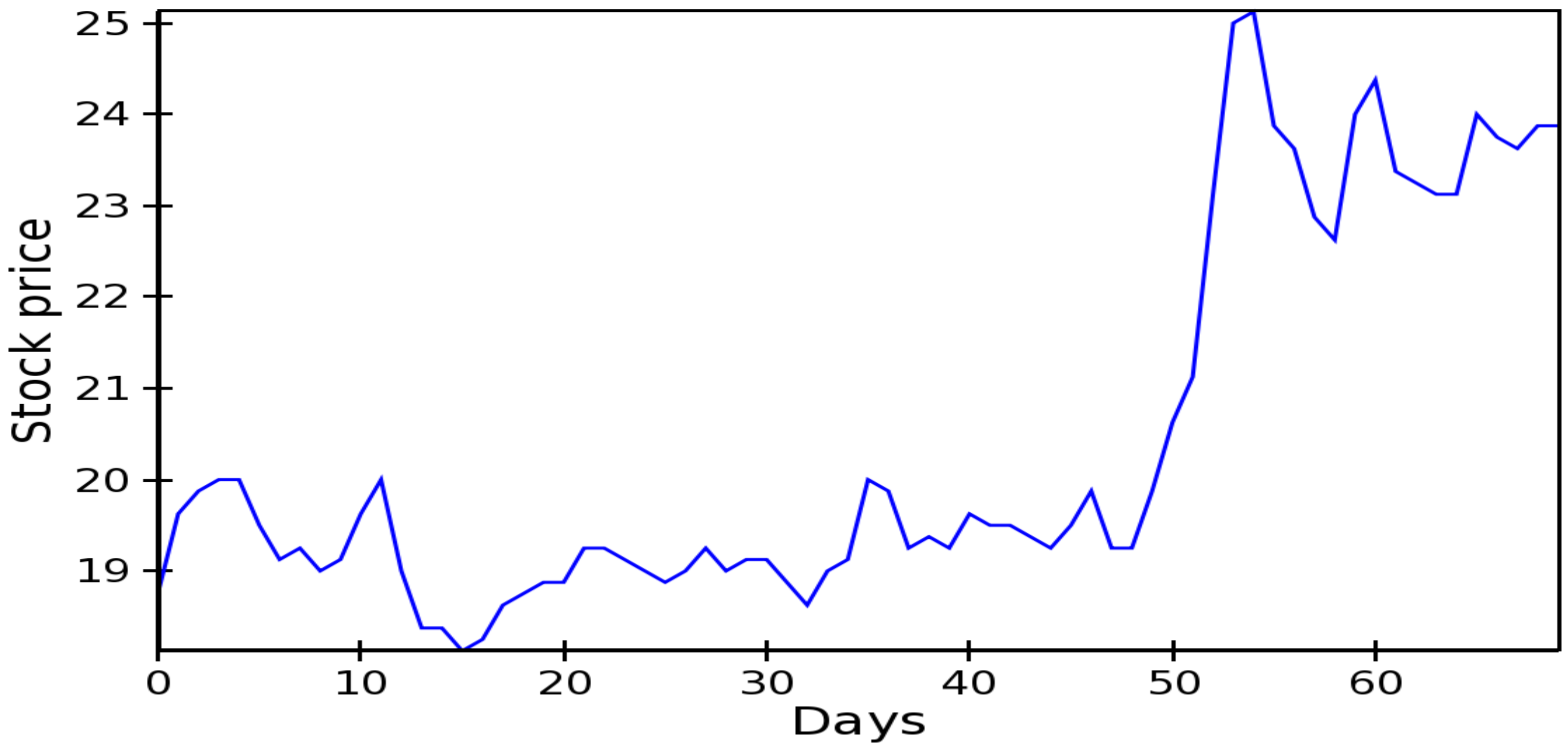
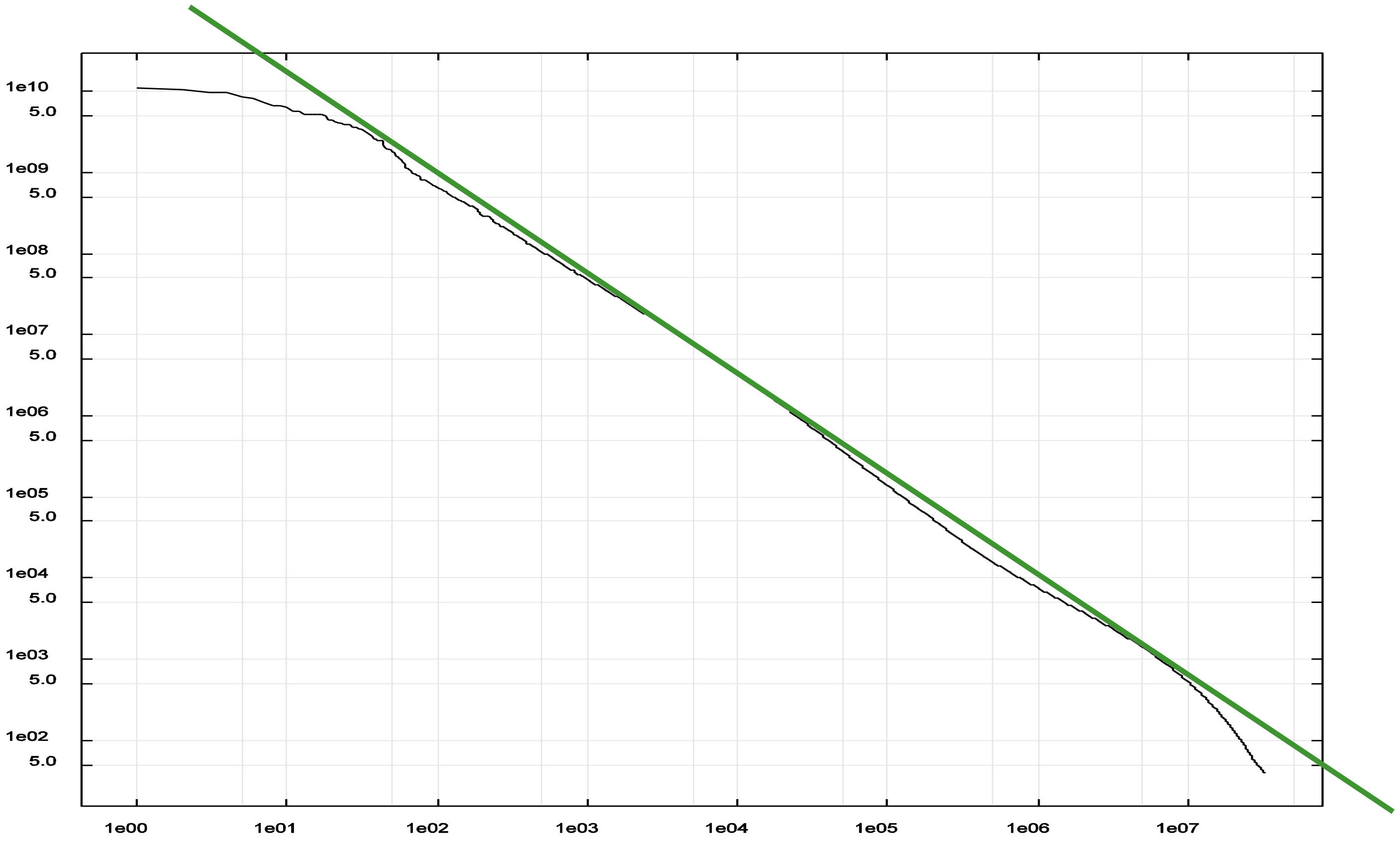


Chart types

- **Log-log plot:** Very useful for power law data

Frequency of
words in tweets



slope ~ -1

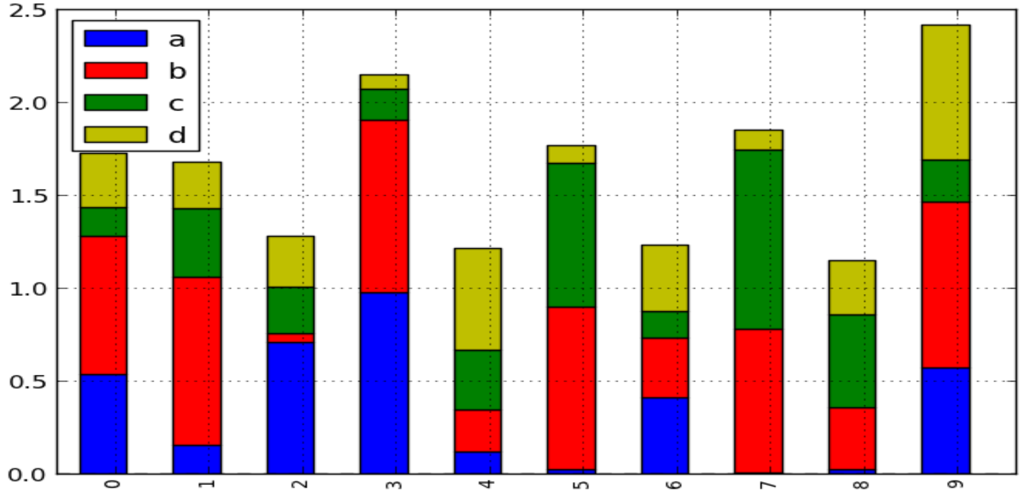
Rank of words in tweets, most frequent to least:
I, the, you,…

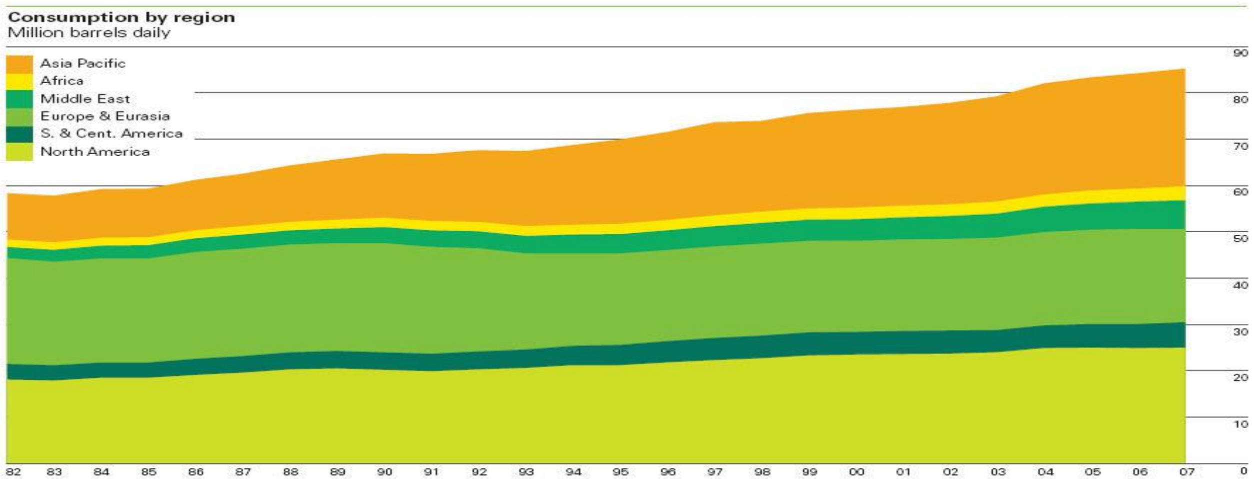
Chart types

- More than two variables
  - Stacked plots
  - Parallel coordinate plot

Chart types

- **Stacked plot:** stack variable is discrete:

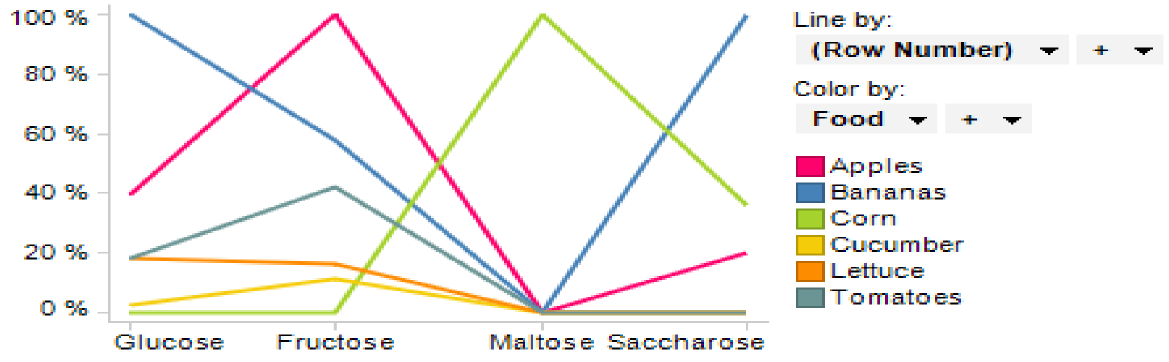




World oil consumption rose by about 1mmb/d in 2007, just below the 10-year average. OECD consumption declined nearly 400,000b/d. China accounted for the largest increment to consumption even though the growth rate was below average. Consumption in oil exporting regions was robust.

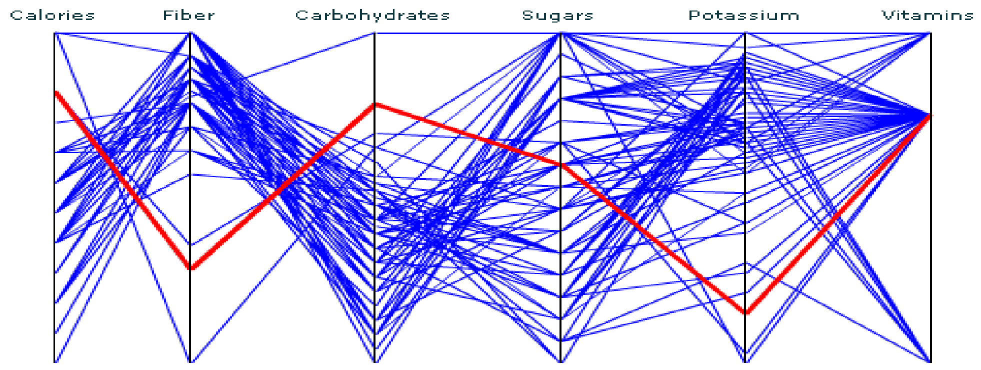
Chart types

- **Parallel coordinate plot:** one discrete variable, an arbitrary number of other variables:



Line by: (Row Number)

Color by: Food



Normal Distributions, Mean, Variance

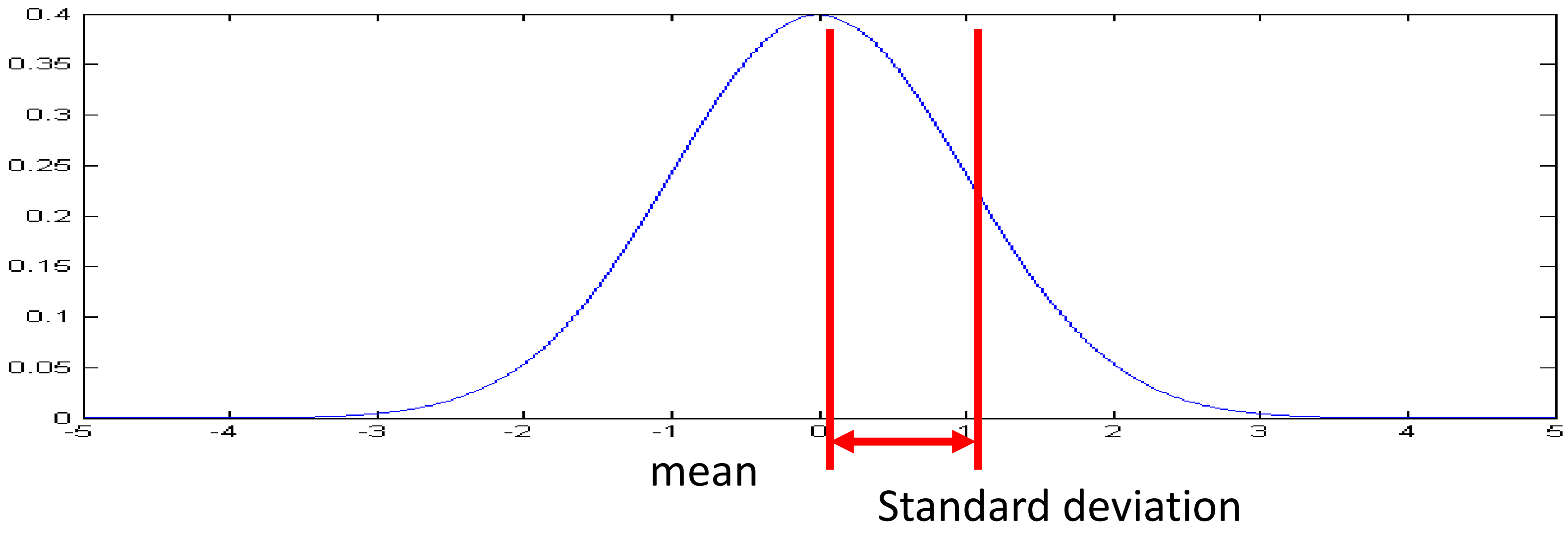
The mean of a set of values is just the average of the values.

Variance a measure of the width of a distribution. Specifically, the variance is the mean squared deviation of samples from the sample mean:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

The standard deviation is the square root of variance.

The normal distribution is completed characterized by mean and variance.

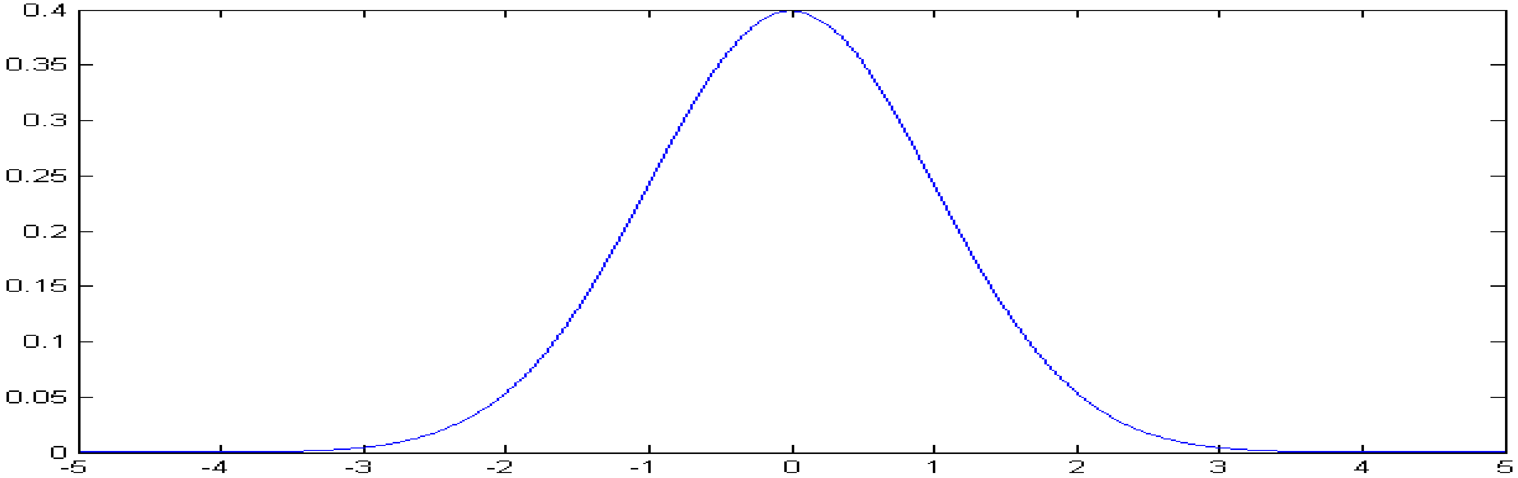


Central Limit Theorem

The distribution of the sum (or mean) of a set of n identically-distributed random variables Xi approaches a normal distribution as n → ∞.

The common parametric statistical tests, like t-test and ANOVA assume normally-distributed data, but depend on sample mean and variance measures of the data.

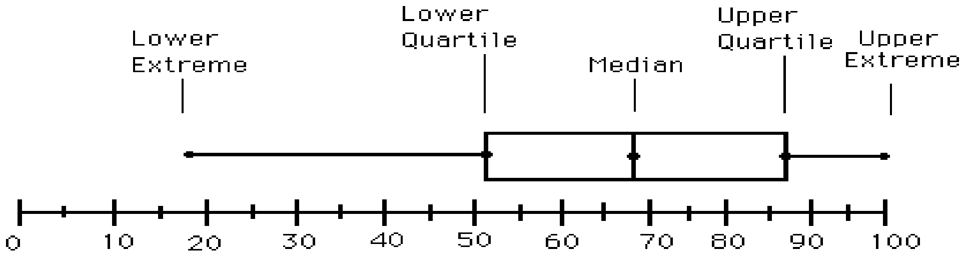
They typically work reasonably well for data that are not normally distributed as long as the samples are not too small.



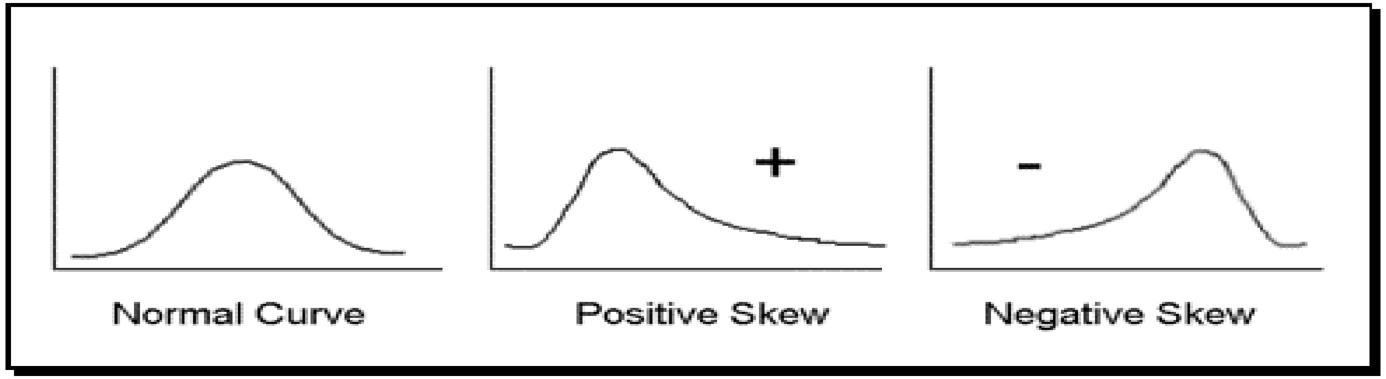
Correcting distributions

Many statistical tools, including mean and variance, t-test, ANOVA etc. **assume data are normally distributed**.

Very often this is not true. The box-and-whisker plot is a good clue



Whenever its asymmetric, the data cannot be normal. The histogram gives even more information

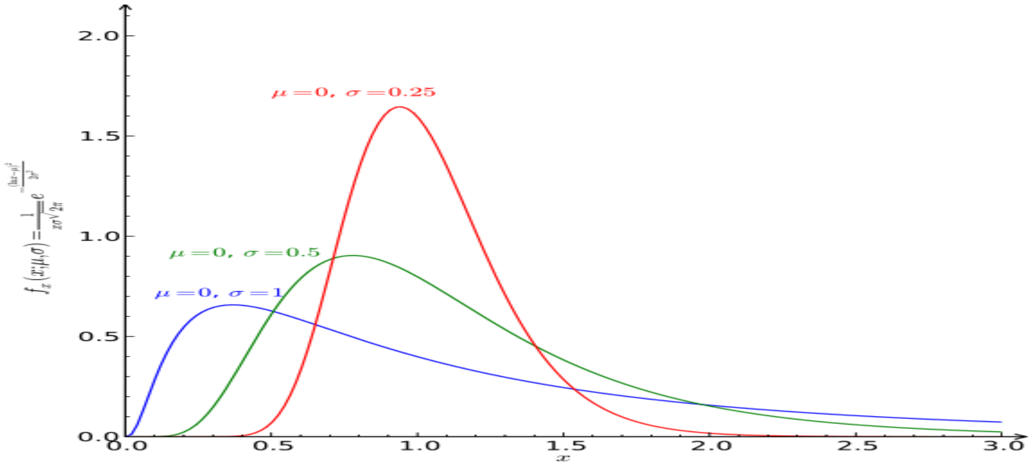


Correcting distributions

In many cases these distribution can be corrected before any other processing.

Examples:

- X satisfies a log-normal distribution, Y=log(X) has a normal dist.



- X poisson with mean k and sdev. sqrt(k). Then sqrt(X) is approximately normally distributed with sdev 1.

Distributions

Some other important distributions:

- **Poisson:** the distribution of counts that occur at a certain “rate”.
  - Observed frequency of a given term in a corpus.
  - Number of visits to a web site in a fixed time interval.
  - Number of web site clicks in an hour.
- **Exponential:** the interval between two such events.
- **Zipf/Pareto/Yule distributions:** govern the frequencies of different terms in a document, or web site visits.
- **Binomial/Multinomial:** The number of counts of events (e.g. die tosses = 6) out of n trials.

- You should understand the distribution of your data before applying any model.

Rhine Paradox*

Joseph Rhine was a parapsychologist in the 1950’s (founder of the *Journal of Parapsychology* and the *Parapsychological Society, an affiliate of the AAAS*).

He ran an experiment where subjects had to guess whether 10 hidden cards were red or blue.

He found that about 1 person in 1000 had ESP, i.e. they could guess the color of all 10 cards.

Q: what’s wrong with his conclusion?

2^10 ~ 1000 so 1 in 1000 is just the rate by of guessing correctly by chance

Rhine Paradox

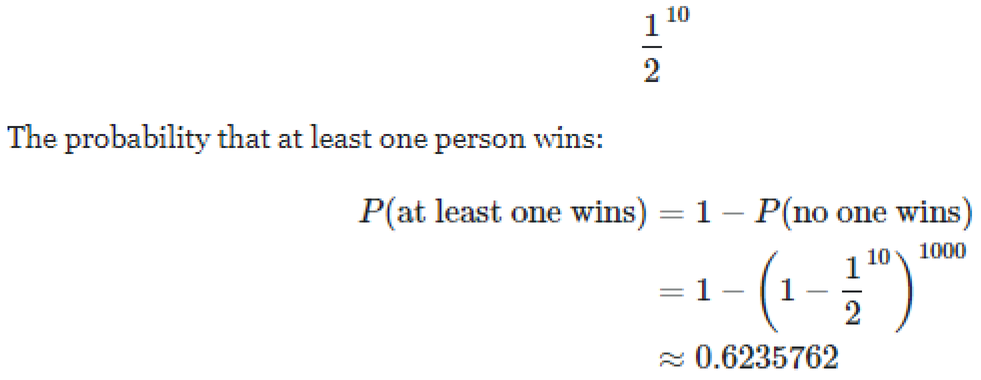
He called back the “psychic” subjects and had them do the same test again. They all failed.

He concluded that the act of telling psychics that they have psychic abilities causes them to lose it…(!)

- “Psişik” denekleri geri çağırdı ve aynı testi tekrar yapmalarını istedi. Hepsi başarısız oldu.
- Medyumlara psişik yeteneklere sahip olduklarını söyleme eyleminin onları kaybetmelerine neden olduğu sonucuna vardı…(!)

Rhine paradox

- In the Rhine paradox described here, the author arrives at a probability of 0.98 that at least 1 in 1,000 people would guess at ten cards right.
- Burada açıklanan Ren paradoksunda yazar, 0,98 olasılığa ulaşır ve 1000 kişiden en az 1'i on kartı doğru tahmin eder.
- 10 kartın rastgele seçildiğini varsayarsak, bir kişinin kazanma olasılığı



$$\frac{1}{2}^{10}$$

The probability that at least one person wins:

$$\begin{aligned} P(\text{at least one wins}) &= 1 - P(\text{no one wins}) \\ &= 1 - \left(1 - \frac{1}{2}^{10}\right)^{1000} \\ &\approx 0.6235762 \end{aligned}$$

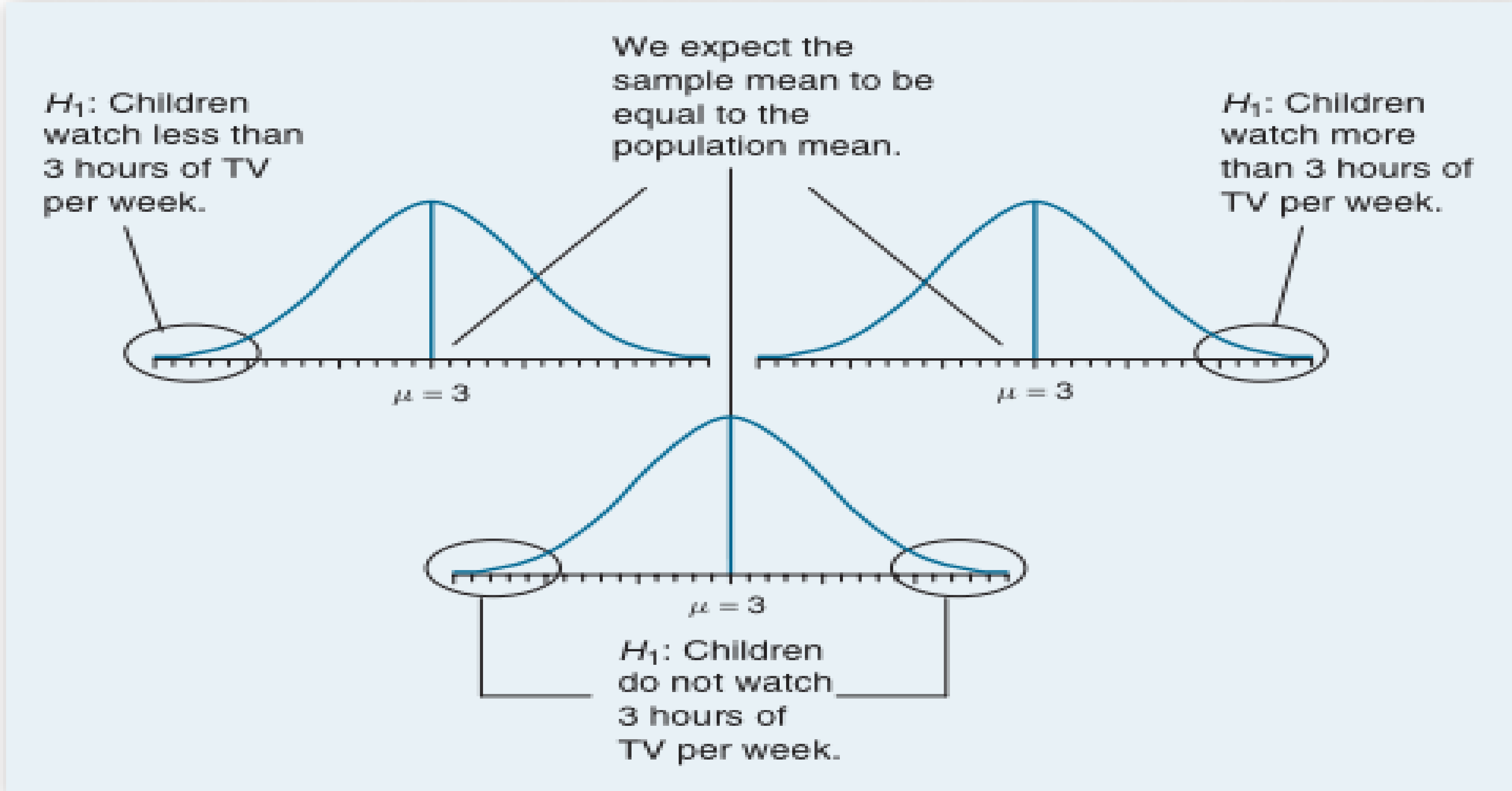
Hypothesis Testing

- We want to prove a hypothesis H_A, but its hard so we try to **disprove a null hypothesis H_0**.
- A **test statistic** is some measurement we can make on the data which is likely to be **big under H_A** but **small under H_0**.
- We chose a test statistic whose distribution we know if H_0 is true: e.g.
  - Two samples a and b, normally distributed, from A and B.
  - H_0 hypothesis that mean(A) = mean(B), test statistic is:
    s = mean(a) – mean(b).
  - s has mean zero and is normally distributed under H_0.
  - But its “large” if the two means are different.

Hypothesis Testing – contd.

- s = mean(a) – mean(b) is our test statistic,
  H_0 the hypothesis that mean(A)=mean(B)
  - We reject if $\Pr(x > s \mid H_0) < p$
  - p is a suitable “small” probability, say 0.05.

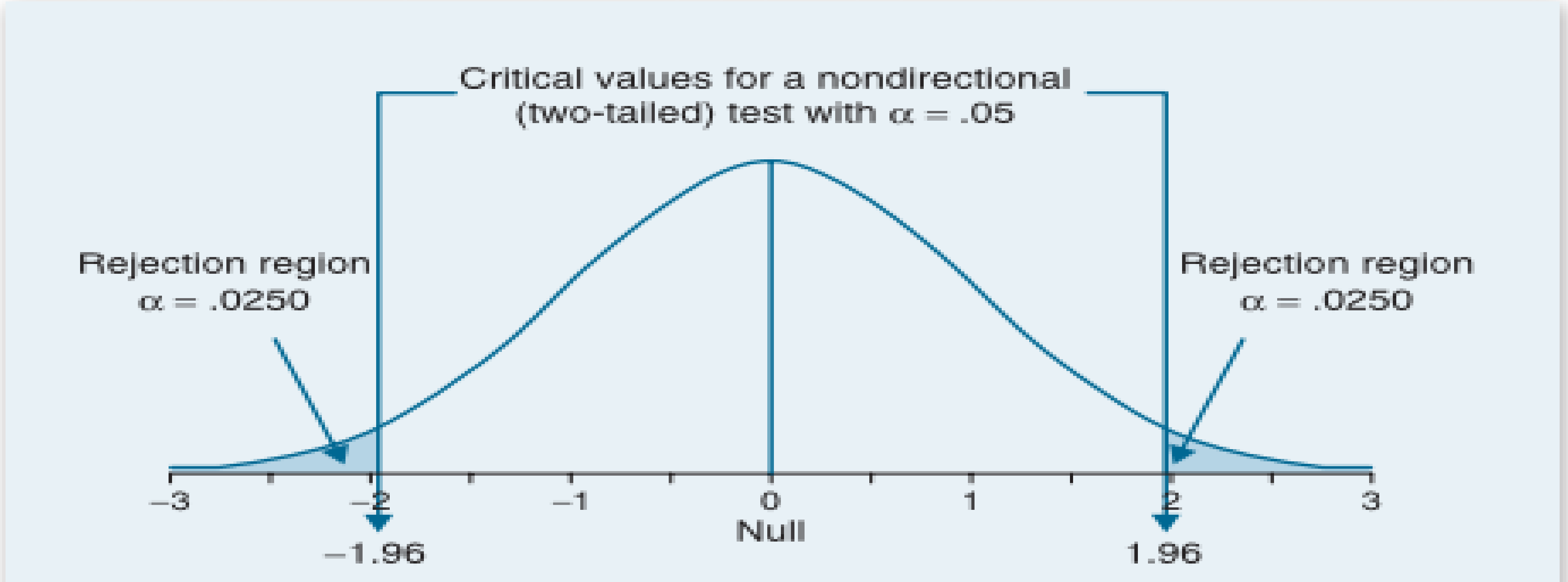
- This threshold probability is called a p-value.
  - P directly controls the false positive rate (rate at which we expect to observe large s even if is H_0 true).
  - As we make p smaller, the false negative rate increase – situations where mean(A), mean(B) differ but the test fails.
  - Common values 0.05, 0.02, 0.01, 0.005, 0.001



H_1: Children watch less than 3 hours of TV per week.
We expect the sample mean to be equal to the population mean.
H_1: Children watch more than 3 hours of TV per week.
$\mu = 3$
$\mu = 3$
$\mu = 3$
H_1: Children do not watch 3 hours of TV per week.

From G.J. Primavera, “Statistics for the Behavioral Sciences”

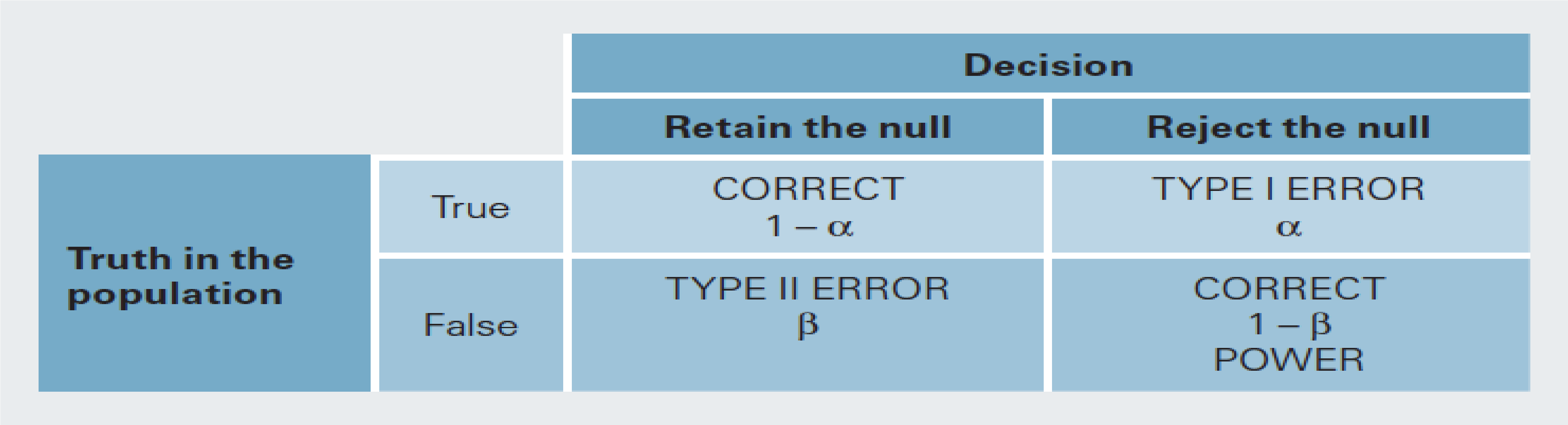
Two-tailed Significance



From G.J. Primavera, “Statistics for the Behavioral Sciences”

When the p value is less than 5% (p < .05), we reject the null hypothesis

Hypothesis Testing



		Decision	
		Retain the null	Reject the null
Truth in the population	True	CORRECT $1 - \alpha$	TYPE I ERROR α
	False	TYPE II ERROR β	CORRECT $1 - \beta$ POWER

Three important tests

- **T-test:** compare two groups, or two interventions on one group.

- **CHI-squared and Fisher’s test.** Compare the counts in a “contingency table”.

- **ANOVA:** compare outcomes under several discrete interventions.

T-test

Single-sample: Compute the test statistic:

$$\mathrm{t} = \frac{\bar{X}}{\bar{\sigma}}$$

where $\bar{X}$ is the sample mean and $\bar{\sigma}$ is the sample standard deviation, which is the square root of the sample variance Var(X).

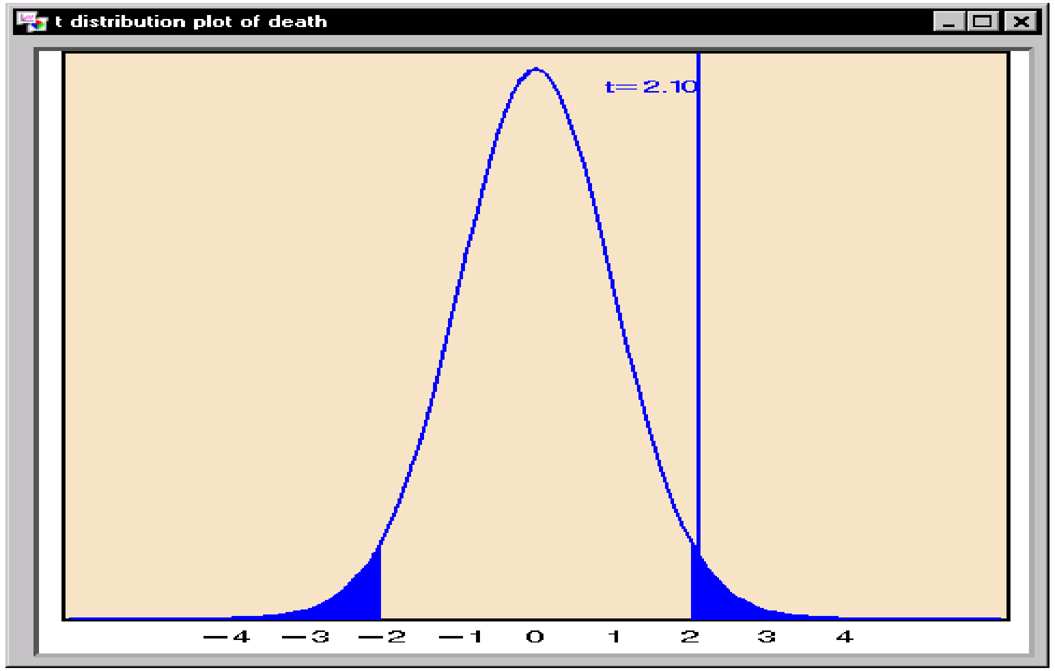
If X is normally distributed, t is **almost** normally distributed, but not quite because of the presence of $\bar{\sigma}$.

You use the single-sample test for **one group** of individuals in **two conditions**. Just subtract the two measurements for each person, and use the difference for the single sample t-test.

This is called a **within-subjects** design.

T-statistic and T-distribution

- We use the t-statistic from the last slide to test whether the mean of our sample could be zero.
- If the underlying population has mean zero, the t-distribution should be distributed like this:



- The area of the tail beyond our measurement tells us how likely it is under the null hypothesis.

- If that probability is low (say < 0.05) we reject the null hypothesis.

Two sample T-test

In this test, there are **two samples** X_1 and X_2. A t statistic is constructed from their sample means and sample standard deviations:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_{\overline{X}_1 - \overline{X}_2}}$$

where:

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

You should try to understand the formula, but you shouldn’t need to use it. most stat software exposes a function that takes the samples X_1 and X_2 as inputs directly.

This design is called a **between-subjects** test.

Chi-squared test

Often you will be faced with discrete (count) data. Given a table like this:

	Prob(X)	Count(X)
X=0	0.3	10
X=1	0.7	50

Where Prob(X) is part of a null hypothesis about the data (e.g. that a coin is fair).

The CHI-squared statistic lets you test whether an observation is consistent with the data:

O_i is an observed count, and E_i is the expected value of that count. It has a chi-squared distribution, whose p-values you compute to do the test.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Fisher’s exact test

In case we only have counts under different conditions

	Count1(X)	Count2(X)
X=0	a	b
X=1	c	d

We can use Fisher’s exact test (n = a+b+c+d):

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\ (c+d)!\ (a+c)!\ (b+d)!}{a!\ \ b!\ \ c!\ \ d!\ \ n!}$$

Which gives the probability directly (its not a statistic).

One-Way ANOVA

ANOVA (ANalysis Of VAriance) allows testing of multiple differences in a single test. Suppose our experiment design has an independent variable Y with four levels:

Y

Primary School	High School	College	Grad degree
4.1	4.5	4.2	3.8

The table shows the mean values of a response variable (e.g. avg number of Facebook posts per day) in each group.

We would like to know in a single test whether the response variable depends on Y, at some particular significance such as 0.05.

ANOVA

In ANOVA we compute a single statistic (an F-statistic) that compares variance between groups with variance within each group.

$$F = \frac{VAR_{between}}{VAR_{within}}$$

The higher the F-value is, the less probable is the null hypothesis that the samples all come from the same population.

We can look up the F-statistic value in a cumulative F-distribution (similar to the other statistics) to get the p-value.

ANOVA tests can be much more complicated, with multiple dependent variables, hierarchies of variables, correlated measurements etc.

Closing Words

All the tests so far are parametric tests that assume the data are **normally distributed**, and that the samples are **independent of each other and all have the same distribution** (IID).

They may be arbitrarily inaccurate is those assumptions are not met. Always make sure your data satisfies the assumptions of the test you're using. e.g. watch out for:

- Outliers – will corrupt many tests that use variance estimates.
- Correlated values as samples, e.g. if you repeated measurements on the same subject.
- Skewed distributions – give invalid results.

Non-parametric tests

These tests make no assumption about the distribution of the input data, and can be used on very general datasets:

- K-S test
- Permutation tests
- Bootstrap confidence intervals

K-S test

The K-S (Kolmogorov-Smirnov) test is a very useful test for checking whether two (continuous or discrete) distributions are the same.

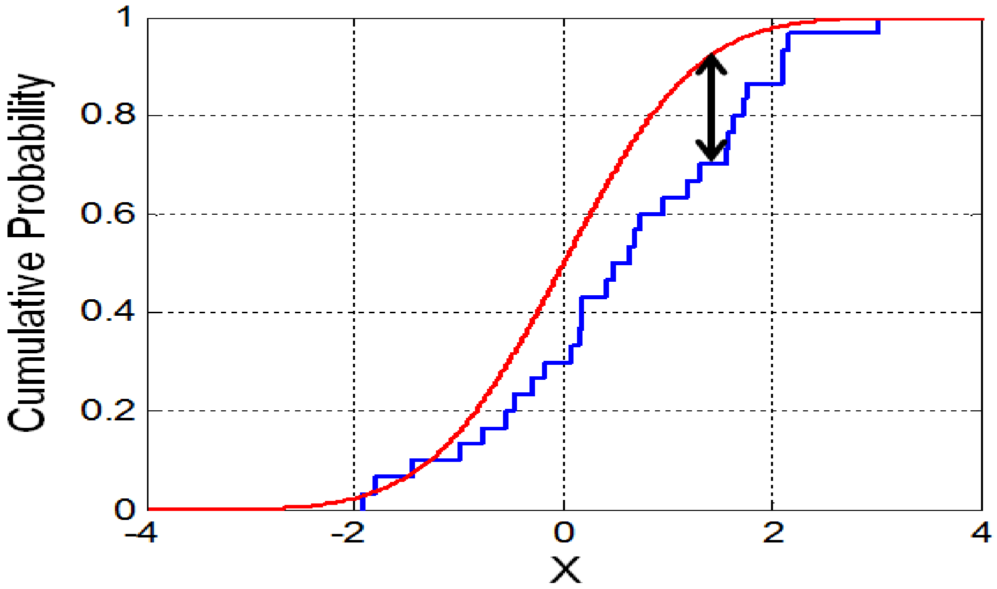
In the **one-sided test**, an observed distribution (e.g. some observed values or a histogram) is compared against a reference distribution.

In the **two-sided test**, two observed distributions are compared.

The K-S statistic is just the **max distance between the CDFs** of the two distributions.

While the statistic is simple, its distribution is not!

But it is available in most stat packages.



K-S test

The K-S test can be used to test **whether a data sample has a normal distribution** or not.

Thus it can be used as a sanity check for any common parametric test (which assumes normally-distributed data).

It can also be used to compare distributions of data values in a large data pipeline: **Most errors will distort the distribution of a data parameter and a K-S test can detect this**.

Non-parametric tests

Permutation tests

Bootstrap confidence intervals

- We wont discuss these in detail, but its important to know that non-parametric tests using one of the above methods exist for many forms of hypothesis.

- They make no assumptions about the distribution of the data, but in many cases are just as sensitive as parametric tests.

- They use computational cycles to simulate sample data, to derive p-value estimates approximately, and accuracy improves with the amount of computational work done.

Outline

- Exploratory Data Analysis
  - Chart types
  - Some important distributions
  - Hypothesis Testing

Usage Notes

- A lot of slides are adopted from the presentations and documents published on internet by experts who know the subject very well.
- I would like to thank who prepared slides and documents.
- Also, these slides are made publicly available on the web for anyone to use
- If you choose to use them, I ask that you alert me of any mistakes which were made and allow me the option of incorporating such changes (with an acknowledgment) in my set of slides.

Sincerely,
Dr. Cahit Karakuş
cahitkarakus@gmail.com