Wrap up statistics

DataResults by web2technology

What is statistics?

- Statistics is the science of collecting, organizing, analyzing and interpreting numerical facts which we call data.
- Statistics is needed in machine learning to estimate the quality of results and to check and improve the data quality

- Statistics is the science of collecting, organizing, analyzing and interpreting data in order to make decision in the presence of uncertainty.

Do we need statistics at all?

Quora

However, when you realize you don't know how to interpret the results of a logistic regression, are baffled by how bad your models perform due to non normalized predictors, and are using the wrong splitting criterion for your tree based models (hint, if the predictors don't all have the same levels, mutual information is probably going to be a bad idea), it's a pretty clear sign that you need some more background knowledge.

How good do you need to be in mathematics and statistics to become a data scientist?

Answer | Request | Follow 160 | Comment 1 | Share 1 | Downvote

Jobs and Careers in Data Science | Data Mining | Data Analysis | +3

How do data scientists use statistics?

1) Parameter Estimation
2) Hypothesis testing
3) Bayesian Analysis
4) Identifying the best estimator
5) Other Statistical Theory

You should have basic knowledge of statistics, such as random variable, probability, common distributions such as Gaussian, etc. Having such basic knowledge is sufficient for you to understand the derivation of many classic machine learning algorithms. And you should do so - otherwise you're just blindly applying models and know nothing about it.

How do I learn statistics for data science?

What statistics book do you recommend to a wannabe data scientist who is familiar with basic statistics and mathematics?

Scope

- Wrap up statistics for machine learning experts
- Focus on practice rather than on exact mathematical proofs – go further in detail if you need
- Introduction to the practical aspects of exploratory data analysis.
- Focus on computational approach, which has several advantages over mathematical approaches for engineering.
- Accompanying notebook on Apache Spark with Scala code
- Start practicing

Base categories: Population

- The **population** is the set representing all measurements of interest to the investigator.
- A population is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.
- In order to make any generalizations about a population, a **sample**, that is meant to be representative of the population, is often studied.

Base categories: Sample

- **A sample** is a subset of measurements selected from the population of interest.
- The sample should be representative of the population.
- In many cases the population is conceptual.
- In most cases measuring the whole population is too costly, unnecessary or impossible.
- There are many possible samples when sample is a subset of the whole population.

Descriptive and inference statistics

Descriptive statistics

- Methods for organizing, displaying and describing data by using tables, graphs and summary statistics.
- Descriptive statistics describe patterns and general trends in a data set. It allows better understanding of the data and access the quality of the data.

Inference statistics

- Statistical Inference makes use of information from a **sample** to draw conclusions about the **population** from which the sample was taken.
- Methods that making decisions or predictions about a population based on sampled data.

Base categories: Types of Variables

Qualitative

- Qualitative variables measure a quality or characteristic on each experimental unit.

Quantitative

- Quantitative variables measure a numerical quantity on each experimental unit.
  - Discrete if it can assume only a finite or countable number of values.
  - Continuous if it can assume the infinitely many values corresponding to the points on a line interval.

Descriptive statistics

- Descriptive statistics are numbers that are used to summarize and describe data
- Descriptive statistics do not involve generalizing beyond the data
- Gives insights of the data at hand, helps to find the meaning of variables

Frequency

- Use a data distribution to describe:
  - What values of the variable have been measured
  - How often each value has occurred

- Absolute frequency h_j - how often a value occured
- Relative frequency $f_j = \frac{h_j}{n}$ (if missing values it must be decided – what is n)
- In Percent (100 x relative frequency)

Cumulated frequency

- only applicable to ordinal scaled values
- How many values are less than a_j

grade	absolut	relative	cumulative
1	3	11.1	11.1
2	7	25.9	37.0
3	9	33.3	70.4
4	8	29.6	100

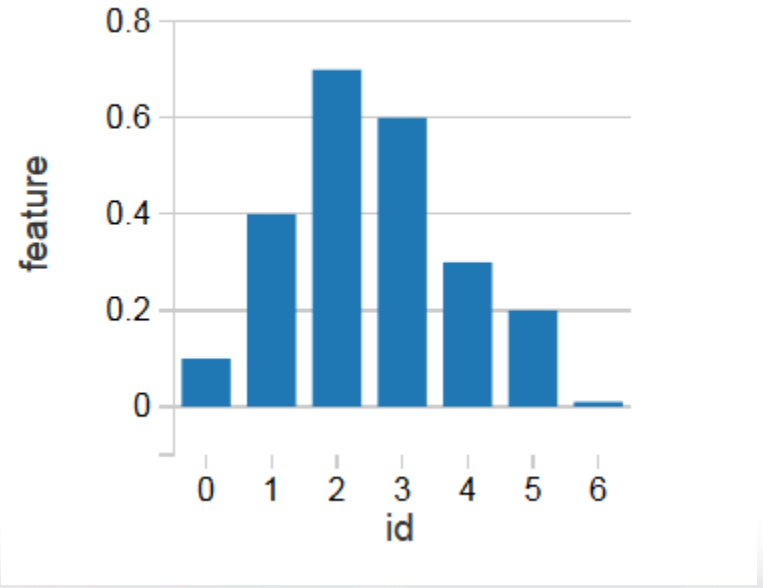
Classified frequencies

age	percent
20-24	4
25-29	12
30-34	16
35-39	22
40-44	25
45-49	12
50-	9

- Classified frequencies are obtained from continous numerical values by grouping. Thats also calle bucketize.
- Groups can have equal range (intervals) or different ones.

Visualize data - histograms

- A **relative frequency histogram** for a quantitative data set is a bar graph in which the height of the bar shows “how often” (measured as a proportion or relative frequency) measurements fall in a particular class or subinterval.



Distributions

Describing Data with Numerical Measures

- Numerical measures can be created for both populations and samples.
  - A parameter is a numerical descriptive measure calculated for a population.
  - A statistic is a numerical descriptive measure calculated for a sample.

- Distribution of frequencies,
- Central tendencies

Measure of central tendency
Mode

The **mode** is the measurement which value occurs most frequently.

value	hj	%
0-4999	6	22.2
5000-9999	6	22.2
10000-14999	2	7.4
15000-19999	**9**	**33.3**
20000-24999	1	3.7
25000-	3	11.1

Measure of central tendency
Median

- All measurements have to be ordered
- **Median** is a measure along the horizontal axis of the data distribution that locates the center of the distribution
- The median of a set of measurements is the middle measurement when the measurements are ranked from smallest to largest.

100	200	300	500	**900**	1000	1200	1500	1800

- $\tilde{x} = x_{(\frac{n+1}{2})}$ for odd n,
- $\tilde{x} = \frac{1}{2} x_{\frac{n}{2}} + x_{\frac{n}{2+1}}$

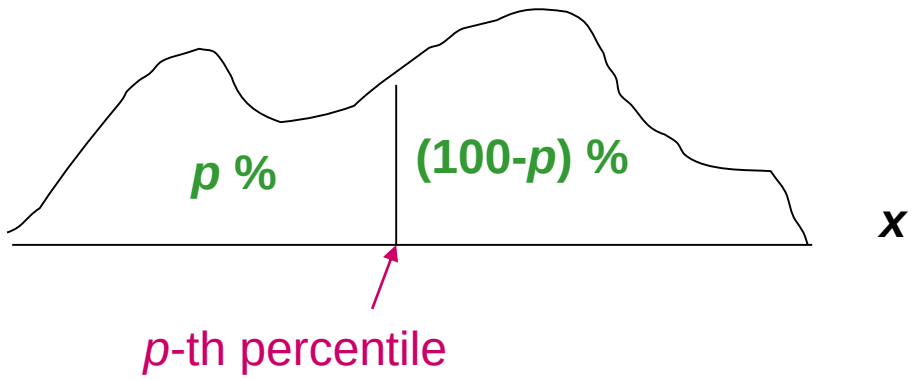
Measure of central tendency
Mean

- Only applicable for numerical data
- The **mean** of a set of measurements is the sum of the measurements divided by the total number of measurements
- $\bar{x} = \frac{\sum x_i}{n}$

i	xi
1	2000
2	5000
3	4000
4	1500
5	2500
$\bar{x}$	**3000**

Measures of dispersion
Quantiles

- Quantiles are analogue to the median - how many measurements lie below the measurement of interest? This is measured by the ***p*th percentile.**



- 50th Percentile is median
- 25th Percentile is lower quartile Q1
- 75th Percentile is upper quartile Q3

Measures of dispersion
Quartiles

- The **lower quartile (Q_1)** is the value of *x* which is larger than 25% and less than 75% of the ordered measurements.
- The **upper quartile (Q_3)** is the value of *x* which is larger than 75% and less than 25% of the ordered measurements.
- The range of the “middle 50%” of the measurements is the **interquartile range,**
  - $IQR = Q_3 - Q_1$

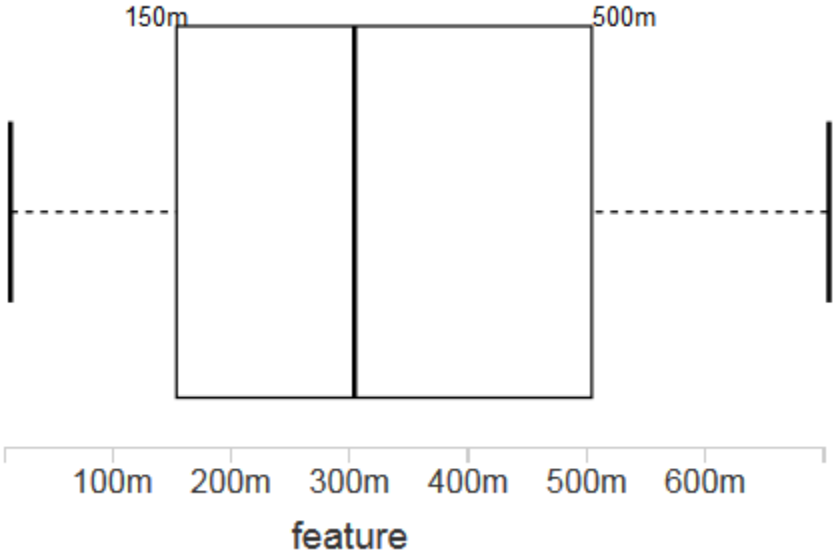
Using measures of center and spread the Box Plot

- Boxplots give a quick summary of the data distribution.
- Create a boxplot

  - ✓ Calculate Q_1, the median, Q_3 and IQR.
  - ✓ Draw a horizontal line to represent the scale of measurement.
  - ✓ Draw a box using Q_1, the median, Q_3.
  - ✓ Isolate outliers by calculating
    - ✓ Lower fence: Q_1-1.5 IQR
    - ✓ Upper fence: Q_3+1.5 IQR
  - ✓ Measurements beyond the upper or lower fence is are outliers and are marked (*).
  - ✓ Draw “whiskers” connecting the largest and smallest measurements that are NOT outliers to the box.

Using measures of center and spread the Box Plot

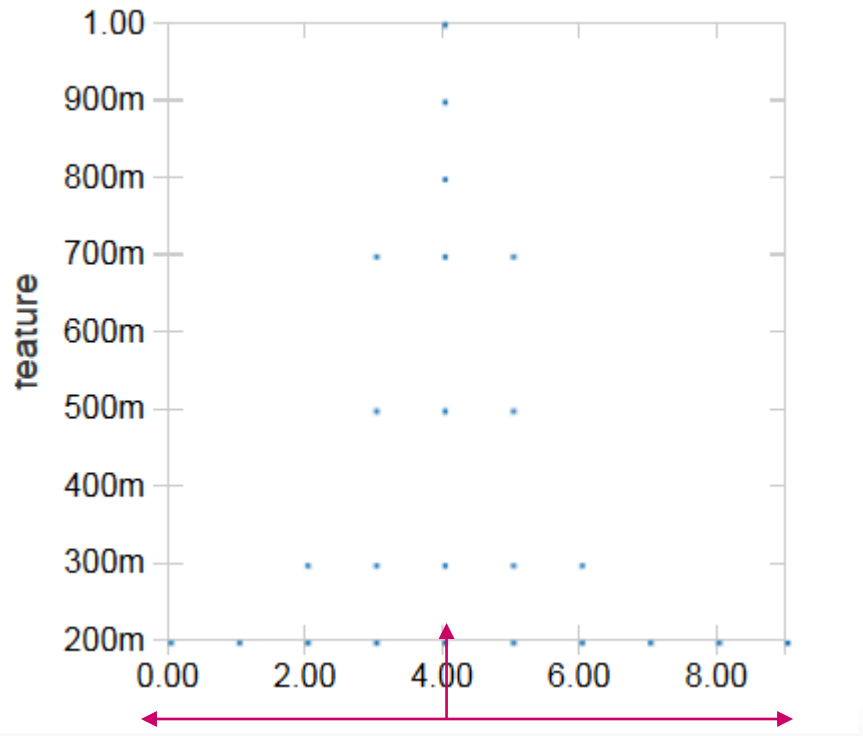
The five-number summary:



min Q1 median Q3 max

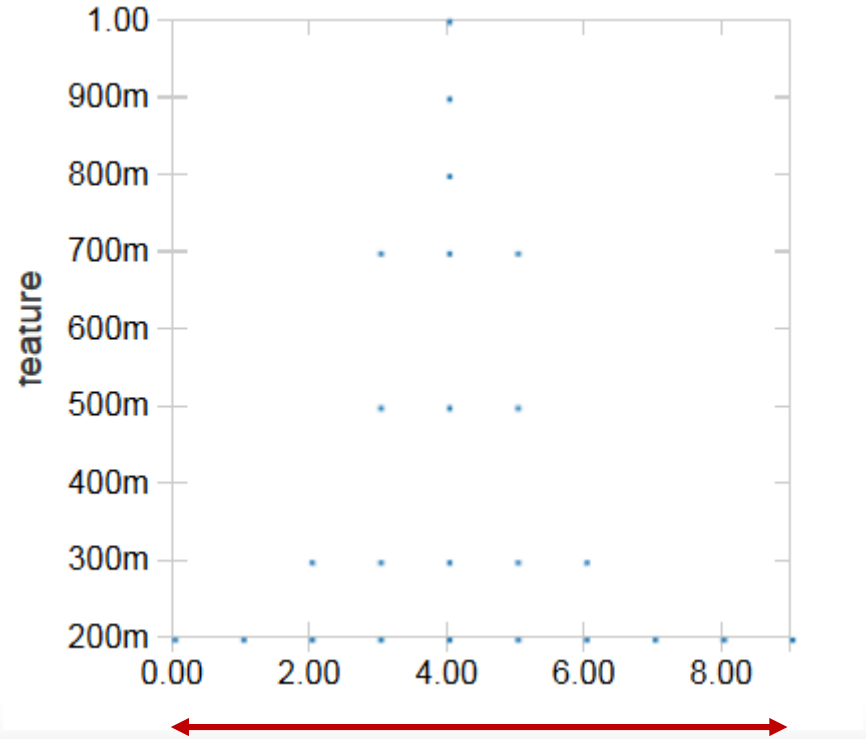
Measures of Variability

- A measure along the horizontal axis of the data distribution that describes the **spread** of the distribution from the center.
- is only applicable to metric data
- **Variance**
- **Deviation**
- different formulas for
  - population and
  - sample



Measures of Variability
Range

- The **range, R,** of a set of *n* measurements is the difference between the largest and smallest measurements (min-max).



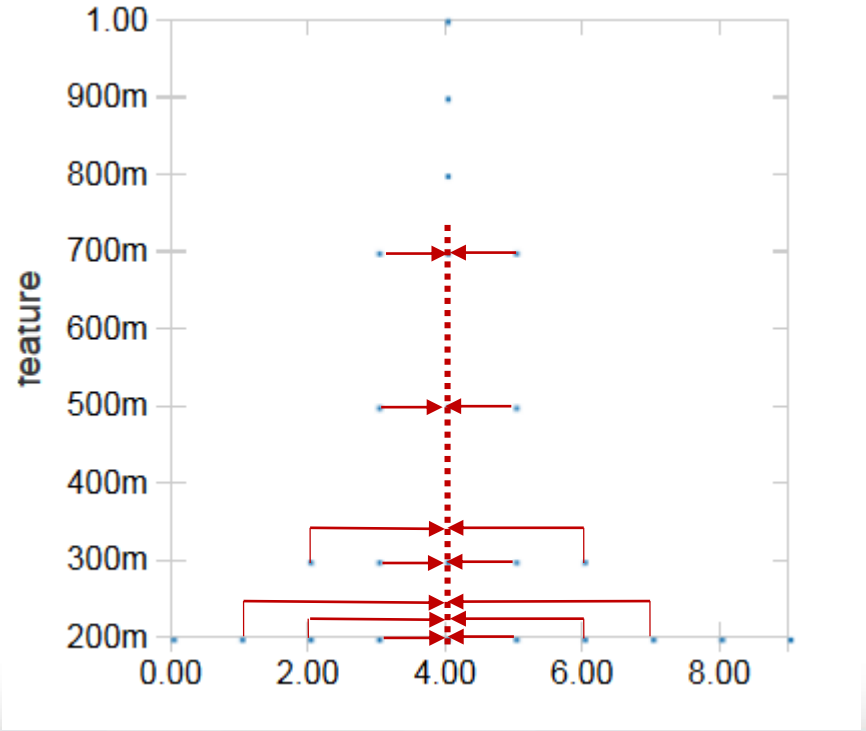
Measures of Variability
Variance

- The **variance** is measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.

Calculation:

- For populations $var_p = \frac{\sum(x_i - \mu)^2}{n}$

- For samples $var_s = \frac{\sum(x_i - \bar{x})^2}{n-1}$



Measures of Variability
Standard deviation

- In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.
- To return this measure of variability to the original units of measure, we calculate the **standard deviation**, the positive square root of the variance.

- Population: $\sigma = \sqrt{var_p}$
- sample: $s = \sqrt{var_s}$

Cumulative distribution function (CDF).

- The **cumulative distribution function** (**CDF**) of a real-valued random variable *X*, evaluated at *x*, is the probability that *X* will take a value less than or equal to *x*.
- In the case of a continuous distribution, it gives the area under the probability density function,
- The CDF is the function that maps from a value to its percentile rank.

$$cdf = count_{(value<x)}/count_{total}$$

Relations between variables Covariance

- Covariance – measure of joint variance of two variables

- Covariance for populations

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Covariance of samples

$$\hat{\sigma}_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Effect size – variance explained

- This effect size estimate the amount of the variance within an experiment that is "explained" or "accounted for" by the experiment's model.
- **Correlation coefficient**: a numerical measure of the **strength** and **direction** of the linear relationship between *x* and *y*
- The strength and direction of the relationship between *x* and *y* are measured using the **correlation coefficient (Pearson product moment coefficient of correlation), *r*.**
- Pearson's correlation only measures linear relationships. If there's a nonlinear relationship, ***r*** understates its strength.

Effect size
Pearson product moment coefficient

- Correlation formula:

$$r = \frac{S_{xy}}{S_x S_y}$$

- with

$$S_{xy} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{n-1}$$

Spearman's Rank Correlation

- Spearman's rank correlation is an alternative that mitigates the effect of outliers and skewed distributions.
- To compute Spearman's correlation, compute the rank of each value, which is its index in a sorted sample.
- Next compute Pearson's correlation for the ranks.

$$r_s = \frac{cov(r_g X, r_g Y)}{\sigma_{r_g X} \sigma_{r_g Y}}$$,where $cov(r_g X, r_g Y)$ is the covariance of the rank variables and $\sigma_{r_g X} \sigma_{r_g Y}$ is the standard deviation of the rank variables.

Effect size – Cohen‘s d

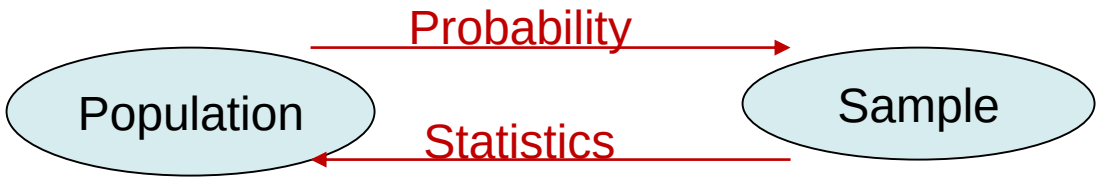
- The **effect size** is a quantitative measure of the strength of a phenomenon.
- Examples of effect sizes are the correlation between two variables or the mean difference.
- Another way to convey the size of an effect is to compare the difference between groups to the variability within groups.
- Cohen's d is a statistic do that
- Cohen’s d is: $d = \frac{\bar{x}_1 - \bar{x}_2}{s}$

Effect size - Interpretation of r

$-1 \leq r \leq 1$	Sign of *r* indicates direction of the linear relationship
$r \approx 0$	No relationship; random scatter of points
$r \approx 1$ or –1	Strong relationship; either positive or negative
$r = 1$ or –1	All points fall exactly on a straight line

Probability

- A probability provides a quantitative description of the chances or likelihoods associated with various outcomes
- It provides a bridge between descriptive and inferential statistics



- The probability **P(A)** of an event A measures “how often” A will occur.
- Suppose that an experiment is performed *n* times. If we let *n* get infinitely large, the relative frequency for an event A is

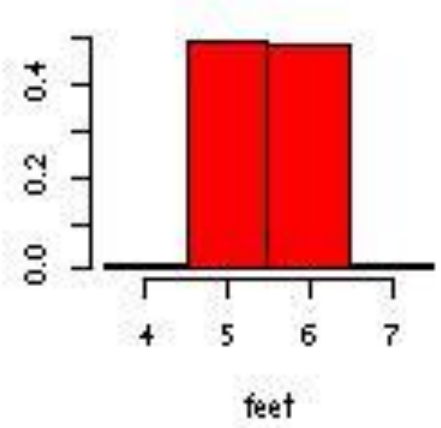
$$P(A) = \lim_{n \to \infty} \frac{f}{n}$$

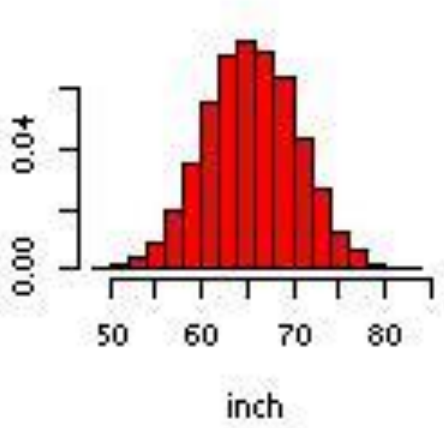
Random variables

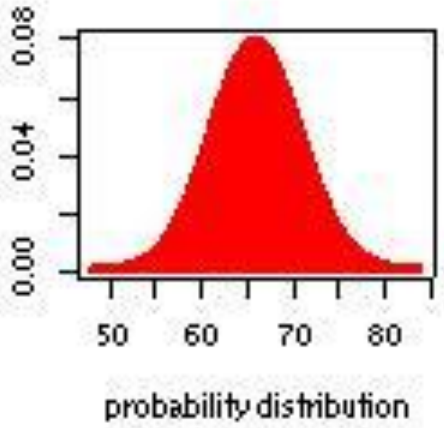
- A random variable is continuous if it can assume the infinitely many values corresponding to points on a line interval.
- A quantitative variable *x* is a **random variable** if the value that it assumes, corresponding to the outcome of an experiment is a chance or random event.
- Random variables can be **discrete** or **continuous.**

Continuous probability distribution

- If random variables are “discretized” by rounding to the nearest category, a continuous probability histogram can be obtained, if the categories intervals are very small.







The Normal Distribution

- The formula that generates the normal probability distribution is:

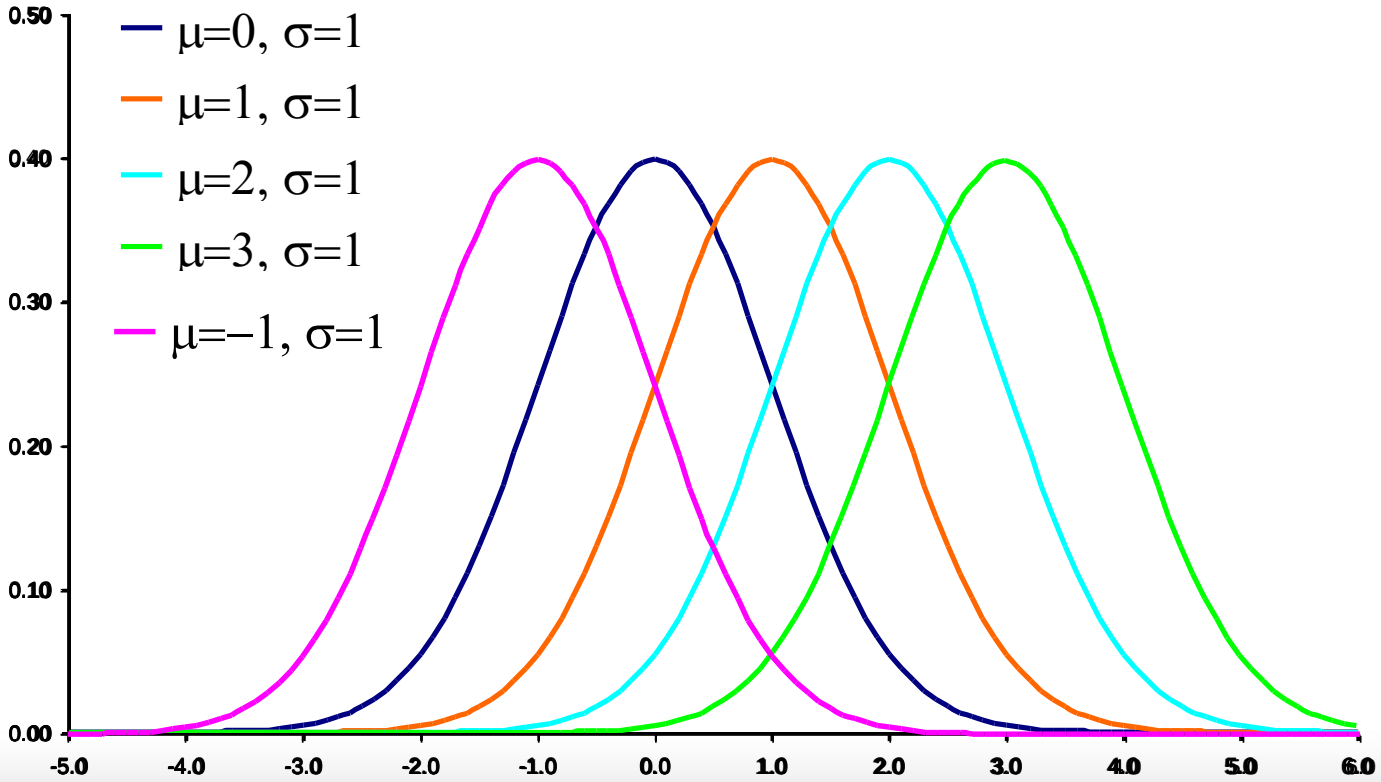
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

- Two parameters, mean and standard deviation, completely determine the Normal distribution. The shape and location of the normal curve changes as the mean and standard deviation change.

- The interval $\mu+\sigma$ contains approximately 68% of the measures,
- The interval $\mu+2\sigma$ contains approximately 95% of the measures,
- The interval $\mu+3\sigma$ contains approximately 99.7% of the measures.

The Normal Distributions: $\sigma = 1$

Normal Distributions: σ=1



The Standard Normal Distribution

- The Standard Normal Distribution is a Normal Distribution with **standardized** values.
- Each value of *x* is standardized by expressing it as a ***z*-score**
- The z-score calculation is a z-transformation of all values
- Z is a measure how far a value lies from the mean μ, divided by the standard deviation.
- $\boldsymbol{z = \frac{x-\mu}{\sigma}}$
- Has same distribution as the Normal Distribution

Probability mass functions

- A probability mass function (PMF),represents a distribution which maps from each value to its probability.
- A probability is a frequency expressed as a fraction of the sample size n.
- To get from frequencies to probabilities, we divide through by n, which is called normalization.

$$pmf = \frac{freq}{n}$$

Parameters

- **Parameters** are numerical descriptive measures for populations.
  - Two parameters for a normal distribution: mean μ and standard deviation σ.
  - One parameter for a binomial distribution: the success probability of each trial p.
- Often the values of parameters that specify the exact form of a distribution are **unknown**.
- You must rely on the **sample** to learn about these parameters.

Sampling distributions

- Any numerical descriptive measures calculated from the sample are called **statistics**
- Statistics vary from sample to sample and hence are **random variables**. This variability is called sampling variability.
- The probability distributions for statistics are called **sampling distributions.**
- In repeated sampling, they tell us what values of the statistics can occur and how often each value occurs

Central Limit Theorem

- If random samples of *n* observations are drawn from a nonnormal population with finite μ and standard deviation σ , then, when *n* is large, the sampling distribution of the sample mean is approximately normally distributed, with mean μ and standard deviation.
- The approximation becomes more accurate as *n* becomes large
- The **Central Limit Theorem** also implies that the sum of *n* measurements is approximately normal with mean $\boldsymbol{n\mu}$ and standard deviation
- This will allow us to describe their behavior and evaluate the **reliability** of our inferences.

Standard error

- The standard deviation of sampling measurements is called standard error.
- For large samplings the standard error has normal distribution.
- For means: $S.E._{mean} = \sqrt{\frac{\sigma_x^2}{n}}$
- For percents: $S.E._{perc} = \sqrt{\frac{\pi_1(1-\pi_1)}{n}}$

Inference statistics

- **Inference** from a set of observations to a reasonable hypothesis for the population
- Beginning with a set of observations, we ask what can be said about the system that generated them.
- Samples are Random Variables

Types of inference

- **Estimation:**
  - Estimating or predicting the value of parameters

    e.g**.** “What is (are) the most likely values of μ or *p*?”

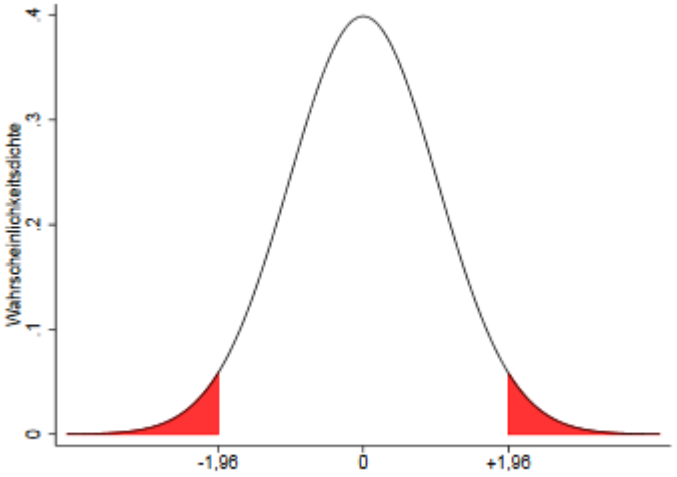
- **Hypothesis Testing:**
  - Deciding about the value of a parameter based on some preconceived ideas.

    E.g. “Did the sample come from a population with $\mu = 5$ or $p = .2$?”

95% Confidence Interval

- The 95% Confidence interval is the interval which includes a value with a probability of 95%
- Confidence intervals depend on sampling distributions.
- For large sample sizes, central limit theorem applies which allow us to use normal distributions
- Confidence interval for

a population mean μ: $\bar{x} \mp z\frac{\alpha}{2}\frac{s}{\sqrt{n}}$



Confidence Interval

- A confidence interval is calculated from **one** given sample. It either covers or misses the true parameter. Since the true parameter is unknown, you'll never know which one is true.
- If independent samples are taken **repeatedly** from the same population, and a confidence interval calculated for each sample, then a certain percentage (**confidence level**) of the intervals will include the unknown population parameter.
- The **confidence level** associated with a confidence interval is the success rate of the confidence interval.
- A parameter that is included in the 95% Confidence Interval is maximal 1,96 S.E. different from the real value.

Hypotheses testing

- Setting up and testing hypotheses is an essential part of statistical inference. In order to formulate such a test, usually some theory has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved.

- Hypothesis testing refers to the process of using statistical analysis to determine if the differences between observed and hypothesized values are due to random chance or to true differences in the samples.
  - Statistical tests separate significant effects from mere luck or random chance.
  - All hypothesis tests have unavoidable, but quantifiable, risks of making the wrong conclusion

Hypotheses testing

- Hypothesis testing is a tool in statistics to determine whether a result is statistically significant, whether this result occurred by chance or not.
- Hypothesis testing is essential for data-driven applications. A test result shows the statistical significance of an event unlikely to have occurred by chance.

Five Steps of a Statistical Test

A statistical test of hypothesis consist of five steps

1. Specify statistical hypothesis which include a **null hypothesis H_0** and a **alternative hypothesis H_a**
2. Identify and calculate **test statistic**
3. Identify distribution and find **p-value**
4. Make a **decision** to reject or not to reject the null hypothesis
5. State conclusion

Null and Alternative Hypothesis

The null hypothesis, H_0:

- The hypothesis we wish to falsify
- Assumed to be true until we can prove otherwise.

The alternative hypothesis, H_a:

- The hypothesis we wish to prove to be true

Two Types of Errors

- There are two types of errors which can occur in a statistical test:
- **Type I error**: reject the null hypothesis when it is true
- **Type II error**: fail to reject the null hypothesis when it is false

Actual Fact / Your Decision	H_0 true	H_0 false
Fail to reject H_0	Correct	Type II Error
Reject H_0	Type I Error	Correct

P-value

•The **p-value** is a measure of inconsistency between the hypothesized value under the null hypothesis and the observed sample.

• The p-value is the probability, assuming that H_0 is true, of obtaining a test statistic value at least as inconsistent with H_0 as what actually resulted.

• It measures whether the test statistic is **likely** or **unlikely**, assuming H_0 is true. Small p-values suggest that the null hypothesis is unlikely to be true. The smaller it is, the more convincing is the rejection of the null hypothesis. It indicates the strength of evidence for rejecting the null hypothesis H_0

Decision

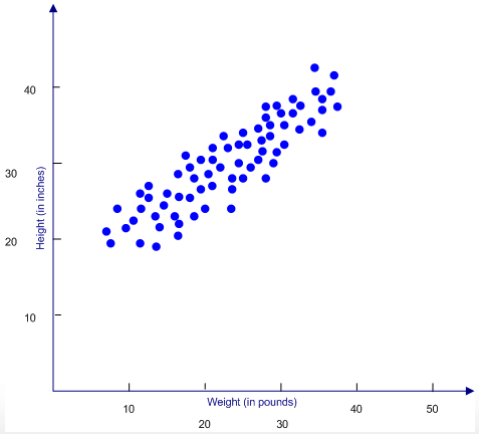
A decision as to whether H_0 should be rejected results from comparing the p-value to the chosen significance level α:

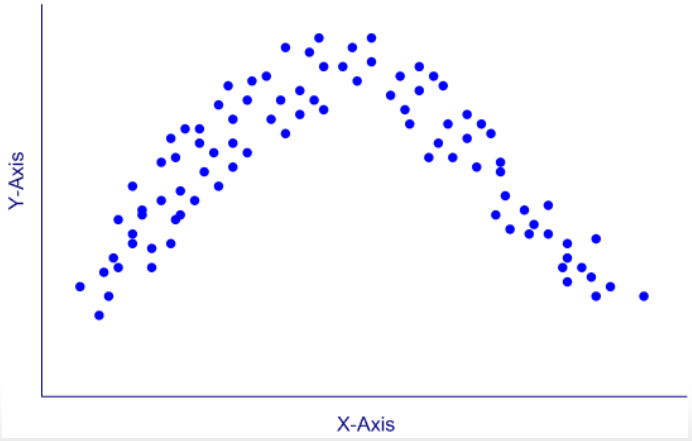
- **H_0 should be rejected if p-value $\leq \alpha$.**
- **H_0 should not be rejected if p-value $> \alpha$.**

1. The significance level α = P(type I error)
2. β = P(type II error)

Relationships, causalities

- Investigate **Relationships** between two variables and how to use one variable to **predict** another variable.
- Sample scatterplots show visually if there could be any relationship between the variables
- Relationships can be causal, but this can’t be measured by statistics





Cross tables

- A **contingency table** or **cross tabulation** is a type of table in a matrix format that displays the frequency distribution of the variables.
- They provide a basic picture of the interrelation between two variables and can help find interactions between them.
- Used to analyse relations between nominal or ordinal scaled parameters.
- Base for chi-squared tests

Cross tables

income| Male| Female	Male	Female
0-24999	104	862
25000-34999	945	2366
35000-44999	2880	0
45000-54999	1350	5110
55000-64999	3840	4710
65000-74999	8260	0
75000-	176	0

- values are grouped in intervals === categories

Chi-square χ^2 test by Pearson

- The chi-square distribution is used in tests of hypotheses concerning the independence of two random variables and concerning whether a discrete random variable follows a specified distribution.
- The chi-square test for independence is used to determine whether there is a significant association between two variables. Two random variables x and y are called independent if the probability distribution of one variable is not affected by the presence of another.

Chi-square χ^2 test by Pearson

- A chi-square random variable is a random variable that assumes only positive values and follows a chi-square distribution.

  Hypothesis:
  H0: variables are indpendent
  H1: variables are **not** independent.
- Calculation of deviations of expected absolute frequencies from measured frequencies.

Chi-square χ^2 test by Pearson

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(0_{ij} - e_{ij})^2}{e_{ij}}$$

Expected values are $e_{ij} = \frac{total_{row} * total_{col}}{total}$ are calcuated from the row, column and total sums,

Observations are o_{ij}

- The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level α

The Analysis of Variance

- The total variation in an experiment is measured by the **total sum of squares**: $SS_{total} = S_{yy} = \sum(y - \bar{y})^2$
- The **Total SS** is divided into two parts:
  - **SSR** (sum of squares for regression): measures the variation explained by including the independent variable *x* in the model: $SSR = \frac{(S_{xy})_2}{S_{xx}}$
  - **SSE** (sum of squares for error): measures the leftover variation not explained by *x:* $SSE = SS_{total} - SSR.$

Coefficient of Determination

- The **coefficient of determination** is defined as

$$r^2 = \frac{SSR}{SS_{total}} = 1 - \frac{SSE}{SS_{total}}$$

- **r^2** is the square of correlation coefficient
- **r^2** is a number between zero and one and a value close to zero suggests a poor model
- It gives the proportion of variation in y that can be attributed to an approximate linear relationship between x and y.

Coefficient of Determination

- A very high value of **r^2** can arise even though the relationship between the two variables is non-linear. The fit of a model should never simply be judged from the **r^2** value alone.

The F Test

- We can test the overall usefulness of the linear model using an F test. If the model is useful, MSR (mean squared regression) will be large compared to the unexplained variation, MSE (mean squared error).
- $H_0: \beta = 0 \; vs \; H_1: \beta \neq 0$
- Test statistic is $F = \frac{MSR}{MSE}$
- Reject $H_0 \; if \; F > F_\alpha$ with 1 and $(n-2)df$

Bootstrap

- Bootstrapping is a statistical technique of resampling (random sampling with replacement).
- Bootstrapping is used if the theoretical distribution of a statistic is not known. With bootstrapping repeatedly statistics are calculated based on only a single sample with resampling.
- Bootstrapping technique allows estimation of the sampling distribution of almost any statistic using random sampling methods

Bootstrap

- Bootstrapping is used if the theoretical distribution of a statistic is not known.
- With bootstrapping repeatedly statistics are calculated based on only a single sample with resampling.
- Bootstrapping technique allows estimation of the sampling distribution of almost any statistic using random sampling methods.

Bootstrap

- Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution.

- The basic idea of bootstrapping is that inference about a population from sample data, can be modeled by resampling the sample data and performing inference about a sample from resampled data.

- As the population is unknown, the true error in a sample statistic against its population value is unknowable.