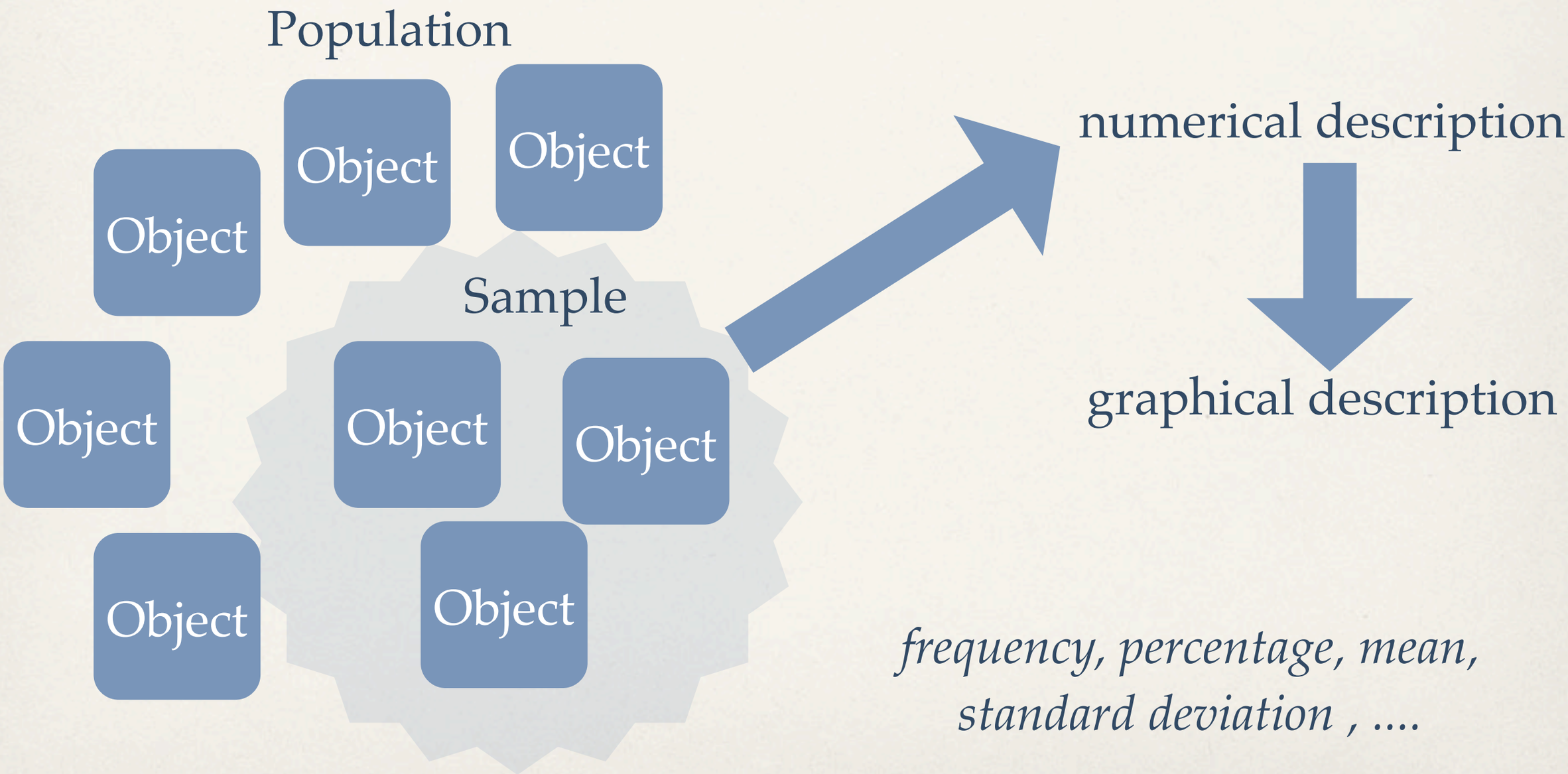
Dr. Peter Tröger
Hasso Plattner Institute, University of Potsdam

Software Profiling Seminar, 2013

Statistics 101

Descriptive Statistics



frequency, percentage, mean, standard deviation , ...

Statistics = Uncertainty

- **Estimate** characteristics of a population, based on samples
- **Characterize natural variation** that is out of statistical control
- **Control and minimize**
  - **Sampling variation** - statistics depend on chosen subset
  - **Measurement variation** - same population object, different results
  - **Environmental variation** - influence of population environment

Measurement Variables

- **Categorical:** No reasonable way for numerical measurement
  - **Nominal scale**: Order has no meaning, values are mutually exclusive, examples: gender, names, yes or no, pass or fail, process states
  - **Ordinal scale**: Order has meaning, distance not,
    examples: school grades, satisfaction level, movie ratings
- **Quantitative:** Discrete or continuous numerical values
  - **Interval scale**: Order and distance have meaning, zero is just a value, examples: temperature in Celsius, dates
  - **Ratio scale**: Order and distance have meaning, has a data-independent zero as natural origin, example: height, weight, temperature in Kelvin

Measurement Variables

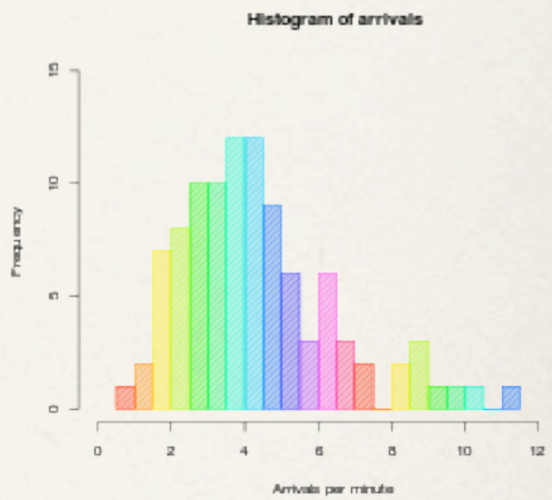
- Applying statistical tests on ordinal data is meanwhile common
  - Example: Average grade
- Helps to think about the feasibility of claims
  - Example: Interval and ratio claims about ordinal data
    - „Average grade in group 1 is B, average grade in group 2 is D.“
    - Is group 1 twice as good as group 2?

What Is Allowed

	Nominal	Ordinal	Interval	Ratio
has	no order	order	order, distance	order, distance, natural zero
Coefficient of variation	-	-	-	ok
Mean, standard deviation	-	-	ok	ok
Add or subtract	-	-	ok	ok
Median, percentiles	-	ok ?	ok	ok
Frequency distribution	ok	ok	ok	ok

Frequency Distribution

- Count of occurrences per value
- Visualized with histograms, bar chars, pie charts, ...



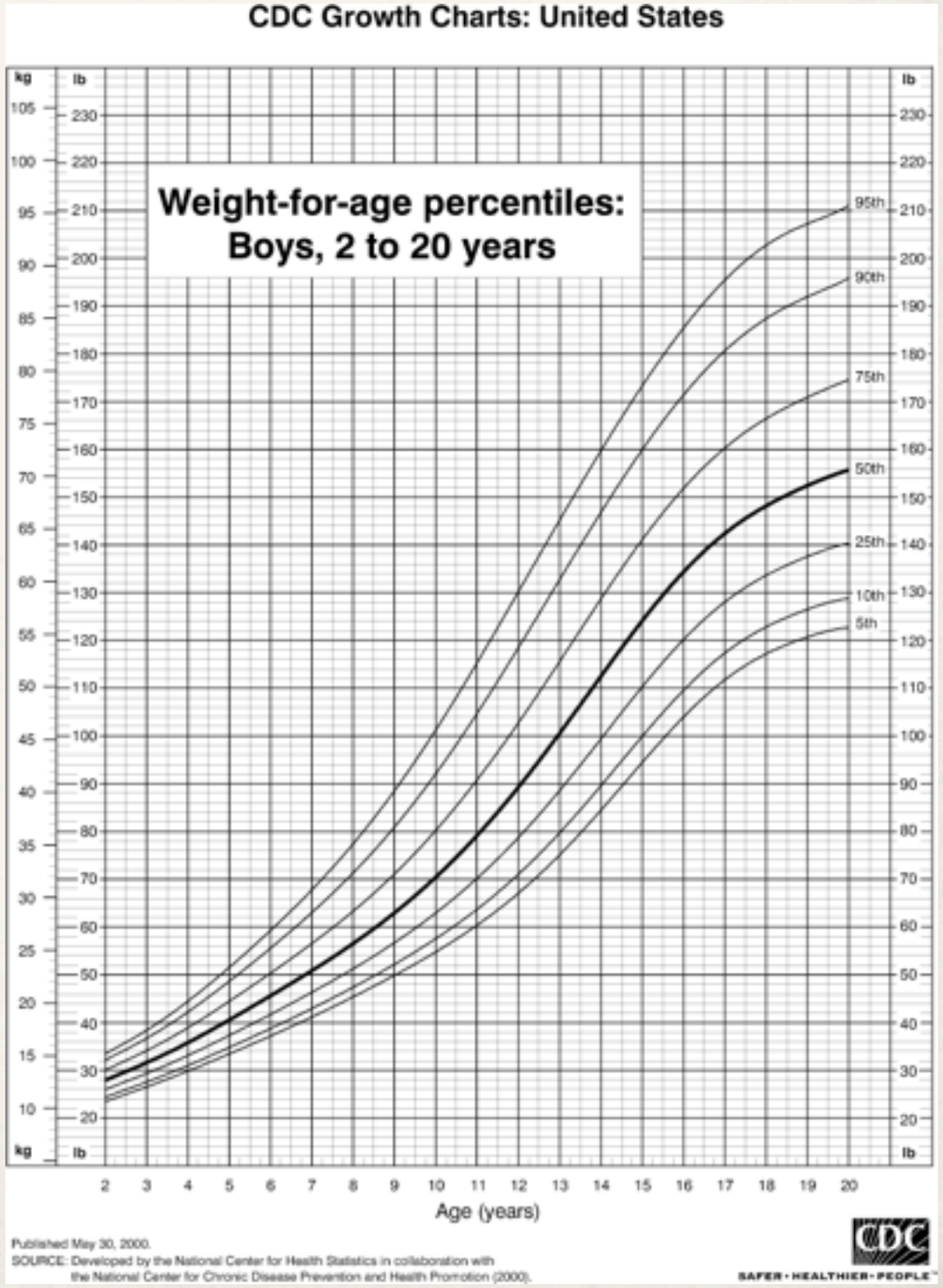
- Feasible for both categorical and numerical data
- Median, mean, standard deviation and others can be easily computed from such an overview
- **Relative frequency distribution**: Information given in percentage

Mean, Median, Mode

- **Mode**: Value that appears most often
  - May be not unique (bimodal, multimodal)
- **Arithmetic mean / average**: Sum divided by number of values
- **Median**: Value separating the higher from the lower half of data
  - For even number of samples, the **mean** of the middle values
- Median is more robust than mean with extreme outliers
  - `mean(1,2,4,6,7,9,10)=5,57` vs. `mean(1,2,4,6,7,9,44)=10,43`

Percentiles

- 75th percentile: The value that has 75% of the values below it
- The 50th percentile is the median
- Examples:
  - Assess growth of children in comparison to national standards
  - Taking rare bursts away from some average calculation
- Four equal quadrants: **quartiles**

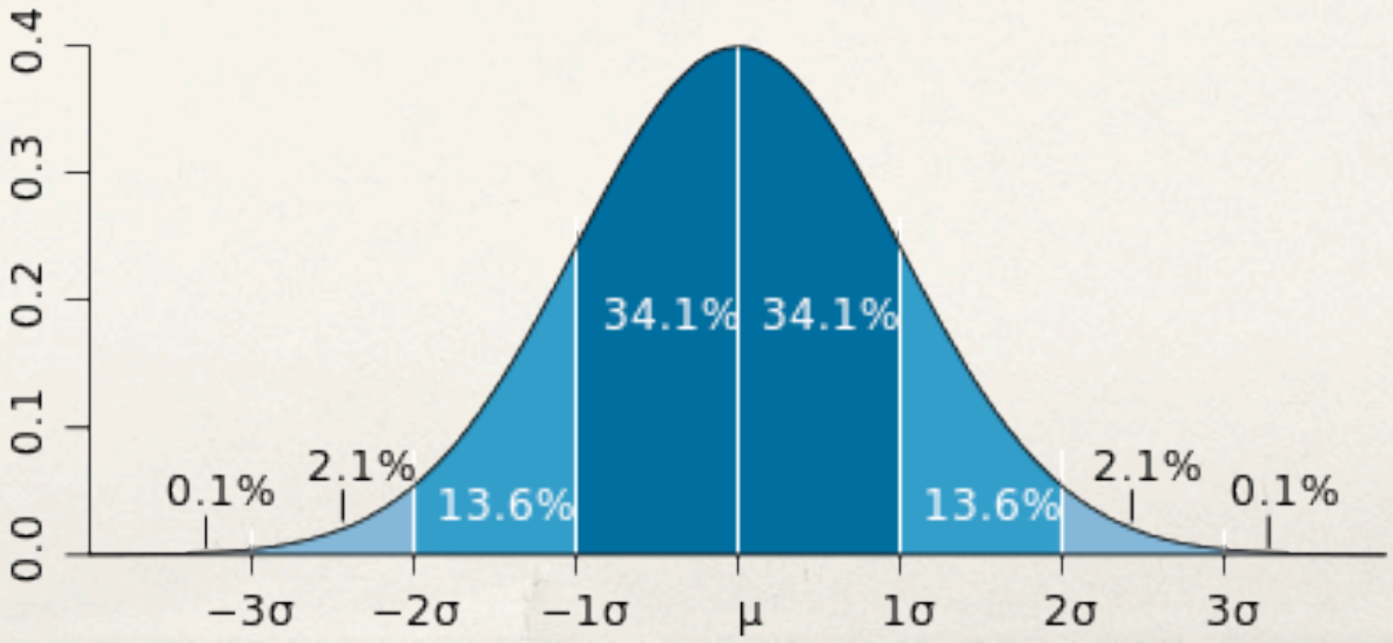


Range, Standard Deviation

- **Range**: Difference between minimum and maximum value
  - Looks at the most extreme values
- **Standard Deviation:** Better to have a ‚typical distance‘ from the mean
  - Describes **variation from the mean** or expected value
  - Expressed in the same units as the data, unlike **variance**
  - A sampled standard deviation has some **confidence interval**
    - With low number of samples, it may vary during multiple runs

Standard Deviation

- **Three sigma rule**: For normal distribution, nearly all values lie within 3 standard deviations
  - Identify impossible outliers
- Reported to be applicable in most cases of statistical analysis



Standard Deviation, CV

- If values are the same, the standard deviation is zero
- Mean and (!) standard deviation are heavily influenced by outliers
- If the data is spread out, you have a higher standard deviation
- **Coefficient of variation:** Ratio of standard deviation to the mean
  - Results in dimensionless number
  - Expresses relative variability, much easier to interpret
  - The secret weapon for cool statistics ...

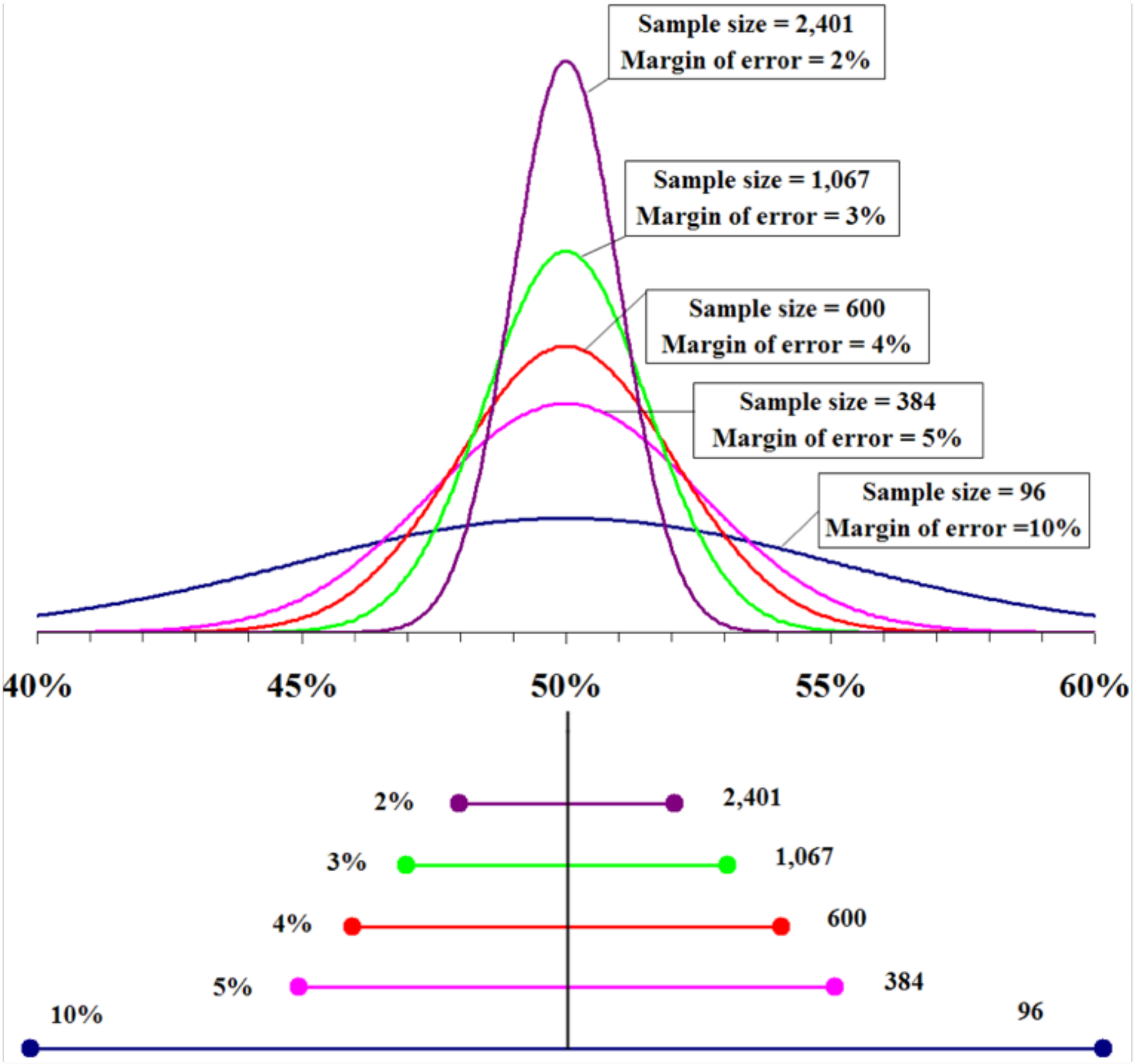
Coefficient of Variation

- Easier for comparison between data sets with widely different means
- When the mean is close to zero, the coefficient will approach infinity
  - Becomes sensitive to small changes
  - Typically indicator for a non-ratio scale
- Distributions with CV<1 are considered low variance
- Distributions with CV>1 are considered high variance
- Sensors need to deliver a CV close to zero (constant absolute error)

Confidence Interval

- If you would repeat the same experiment endlessly, what is the **confidence interval** that would include the real value of some **sample estimate** (e.g. mean) in X% of the cases?

- X is the **confidence level**, typical value is 95%
  - Decided by the researcher who wants to prove a hypothesis
- **Margin of error** to the left and right builds confidence interval
- Helps to deal with **sampling variation,** direct relation to **sample size**
- Can be used to check the quality of mean, median, or percentages



(from Wikipedia)

Hypothesis

- **Hypothesis**: Assumption being proved with empirical data
- **Null hypothesis**: Assuming no relation between cause and effect

- If the probability for it becomes low enough, then the **alternative hypothesis** becomes true ‚without any doubt‘
- **Statistical tests** only target the falsification aspect,
  so the null hypothesis is never intended to get verified by them
  - Decision about falsification based on some some **test result**
- Example: Innocence in jury trial

Statistical Significance

- You are very sure that your falsification of the null hypothesis is reliable

- **One-tailed test**: When the null hypothesis states a direction
  - Example: „Females will not score higher on IQ tests than males.“

- **Two-tailed test**: No direction is given
  - Example: „There is no significant gender-related difference in IQ test results.“

- Test typically relates to the acceptance of a maximum **error rate**

Chi-Square Test

- Mostly means **Pearson's chi-square test**
- Supports two types of comparison, based on frequency distribution
  - **Goodness of fit**: Does the observed frequency distribution differ from an assumed theoretical distribution?
  - **Test of independence**: Are paired frequency distribution investigations independent from each other?
- Many other statistical tests that help to verify / falsify a hypothesis

Summary

- Know your variable type
- Never claim averages without standard deviation
- Coefficient of variation is your friend
- A hypothesis is helpful
- Use some existing software for calculations (e.g. R)
- Graphs are ok

Sources

- http://www.graphpad.com/support/faqid/1089/
- http://statswithcats.wordpress.com/
- http://www.psychstat.missouristate.edu/introbook/sbk06m.htm
- http://www.usablestats.com
- http://www.statcan.gc.ca/edu/
- http://wikipedia.org
- http://www.statpac.com