**Oleksandr Romanko**, Ph.D.
Senior Research Analyst, Risk Analytics, Watson Financial Services, IBM Canada
Adjunct Professor, University of Toronto

# MIE1624H – Introduction to Data Science and Analytics

## Lecture 3 – Basic Statistics

University of Toronto
October 2, 2018

# Lecture outline

**Basic statistics**

- Before you analyze your data
- Sources of uncertainty
- Summarizing and interpreting your data
  - Quantitative data
  - Categorical data
- Distributions
- Law of Large Numbers and Central Limit Theorem

# 🕮 Before You Analyze Your Data

# Where does your data come from?

- Do you have access to complete data, or only a sample?

Entire database of sales transactions

Sample of sales transactions
- How was the subset selected?
- Systematically, randomly?

HR data about all employees

Data for a subset of employees
- Randomly selected?
- Voluntary response?

- How the data was collected will drive what kind of conclusions we may be able to draw, and how confident we can be in those conclusions.

Complete demographic data of NYC users of web service

Conclusions about all NYC users of the service?

Conclusions about all NYC inhabitants?

# Election polling

- In many cases margins of error reported by pollsters substantially over-states the precision of poll-based forecasts
  - Usually reported margin of error is 3% (for a random and representative sample of around 1000 people)
  - Trump vs. Clinton election, why polls were wrong?
- Current polling practice
  - Low response rates (less than 10%)
  - Inadequate coverage
  - Hidden dependence (who tends to answer phone?)
  - Question design and the order in which questions are asked:
    - who would you vote for?
    - would you go and vote?
  - Pollster's methodology often produces results that lean to one side of politics or the other
  - Opinion polls tell us a historical fact on the date people were polled
- Sampling approach does not randomly select people from the entire population
- Segments of the population are excluded

# US presidential elections 2016

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business Tech Markets Opinion Arts Life Real Estate

## Bad Election Day Forecasts Deal Blow to Data Science

Prediction models suffered from narrow data, faulty algorithms and human foibles

By KIM S. NASH, STEVEN NORTON AND SARA CASTELLANOS

9 COMMENTS

Nov 9, 2016 6:33 pm ET

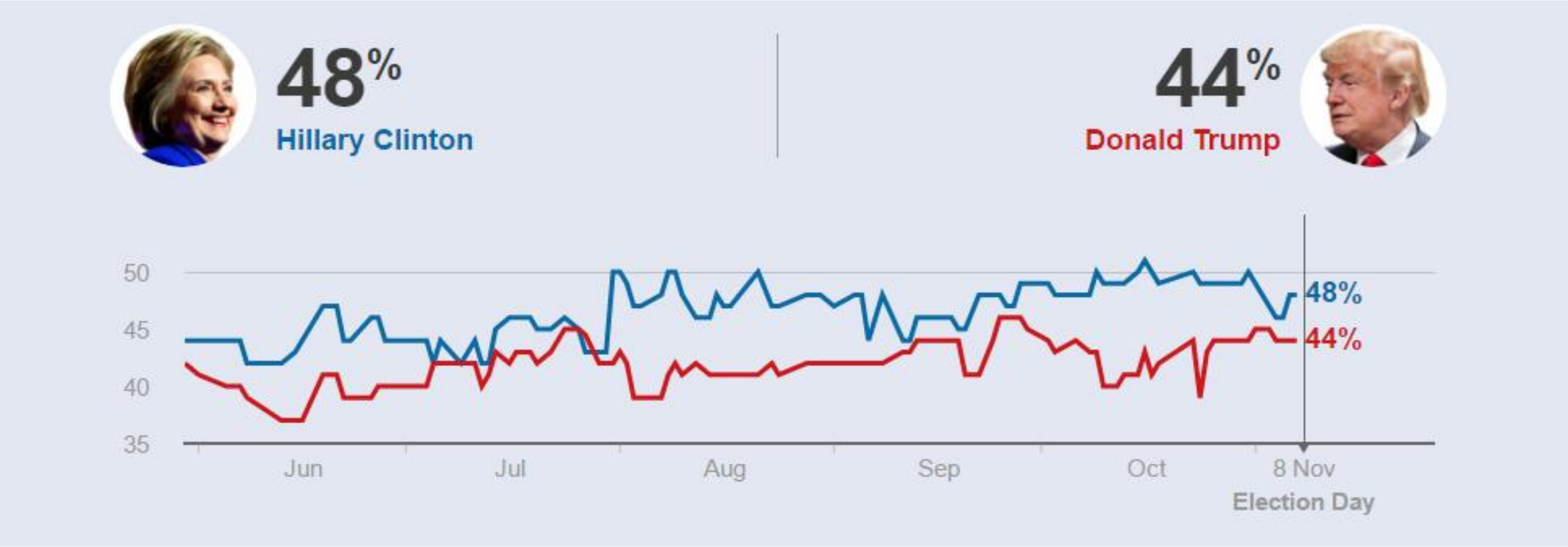
The New York Times

ELECTION 2016 Full Results | Exit Polls | Trump's Cabinet

## *How Data Failed Us in Calling an Election*

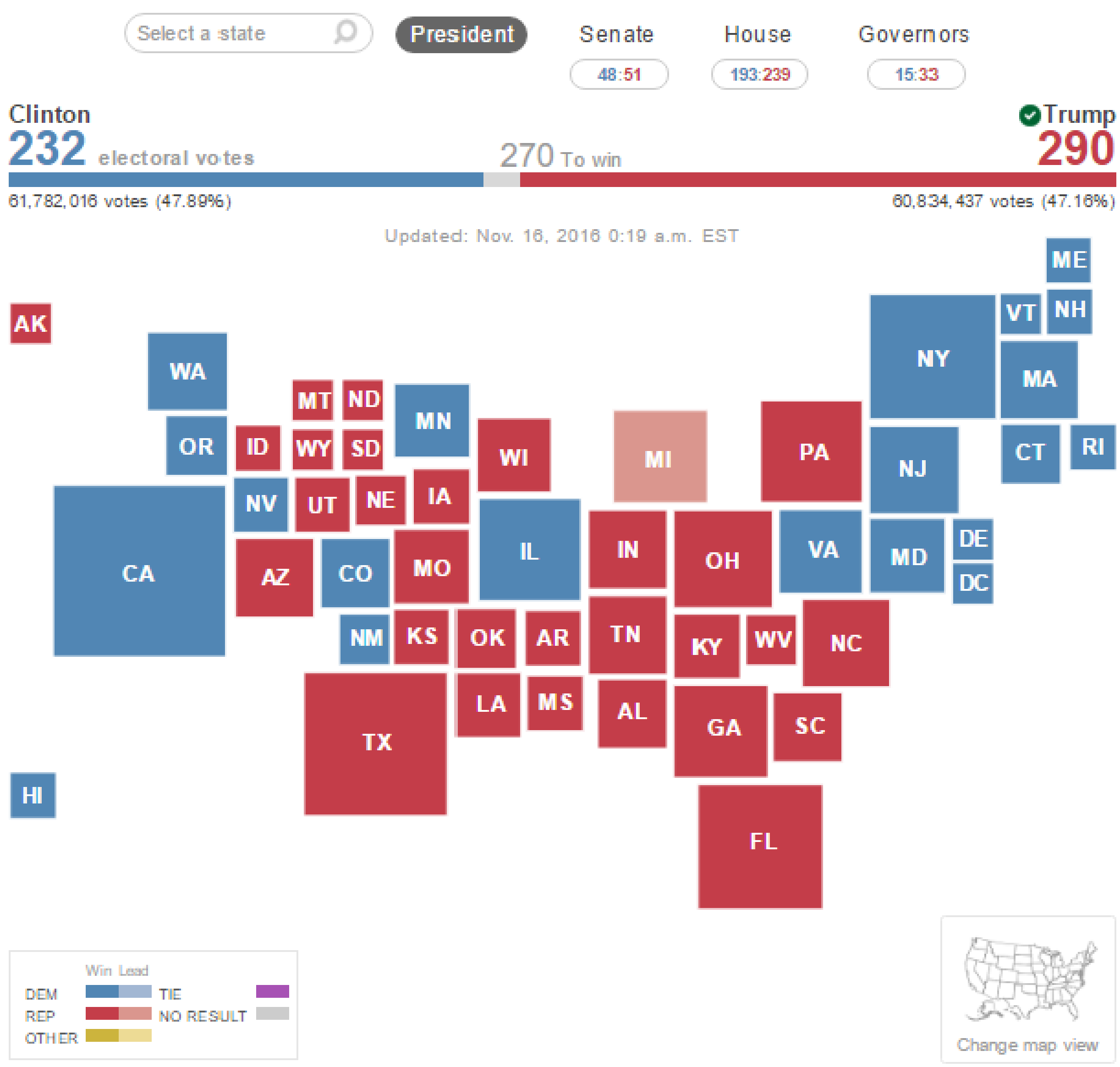
By STEVE LOHR and NATASHA SINGER NOV. 10, 2016

# US presidential elections 2016




Source: BBC poll of polls

# US presidential elections 2016



# What kind of data are we dealing with?

- Types of data
  - Quantitative
  - Categorical (ordered, unordered)

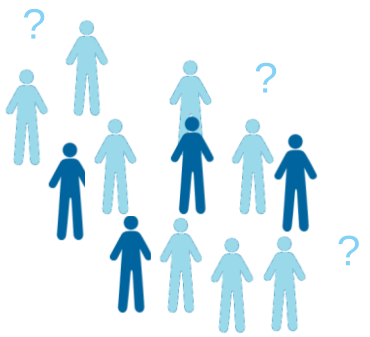
- Data collection
  - Independent observations (one observation per subject)
  - Dependent observations (repeated observation of the same subject, relationships within groups, relationships over time or space)

- Type of data drives the direction of your analysis
  - How to plot
  - How to summarize
  - How to draw inferences and conclusions
  - How to issue predictions

# Uncertainty stemming from the data collection process

No uncertainty

Complete data
e.g., census (in theory), database of all business transactions in the past, Big Data (in some cases)




Sparse data
e.g., survey data, sensor data, experiments

Greater uncertainty

Uncertainty due to data from only a sample, in addition to uncertainty in the measurement tool

# Sources of uncertainty

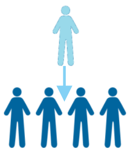
Uncertainty from data collection

+

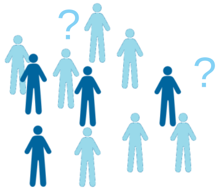
Uncertainty in model

Uncertainty in descriptive statistics, predictions and forecasts

- Average vs. Individual (Standard Deviation)



- Data vs. Reality (Confidence Interval, Margin of Error)



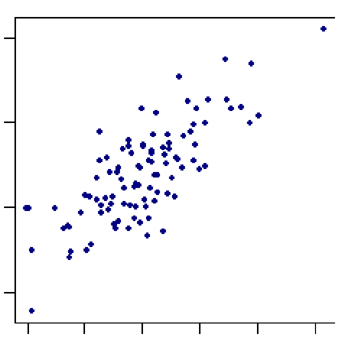
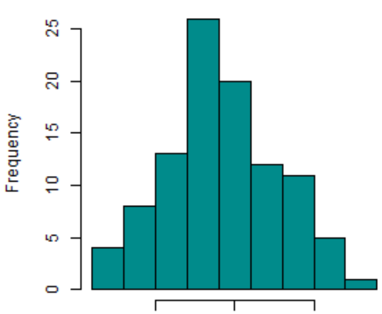
- Prediction/Forecast (Prediction Intervals)

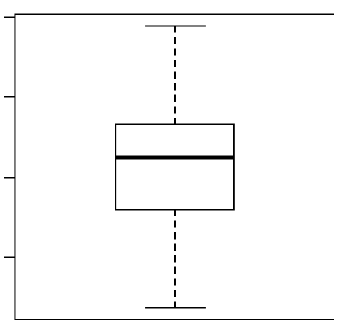
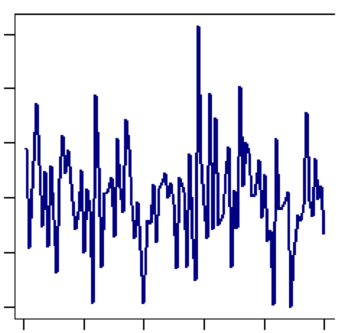
# 📖 Quantitative Data

# Quantitative data

- Examples: temperature, age, income
- Quick check: “Does it makes sense to calculate an average?”

- Appropriate summary statistics:
  - Mean and Median
  - Standard Deviation
  - Percentiles
- More advanced predictive methods: Regression, Time Series Analysis, …
- Plot your data!

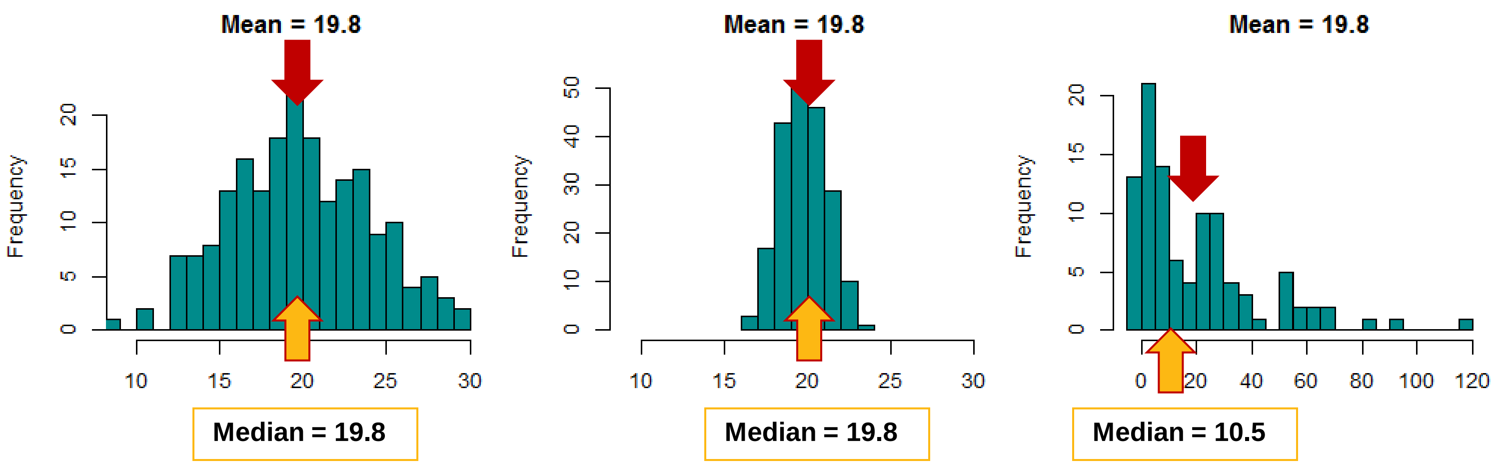



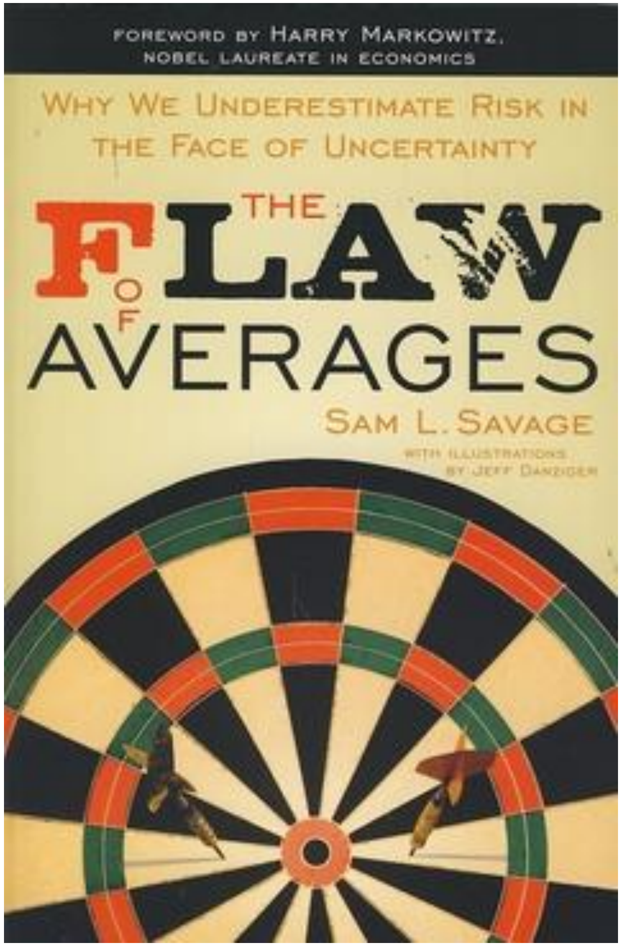
# Summarizing quantitative data

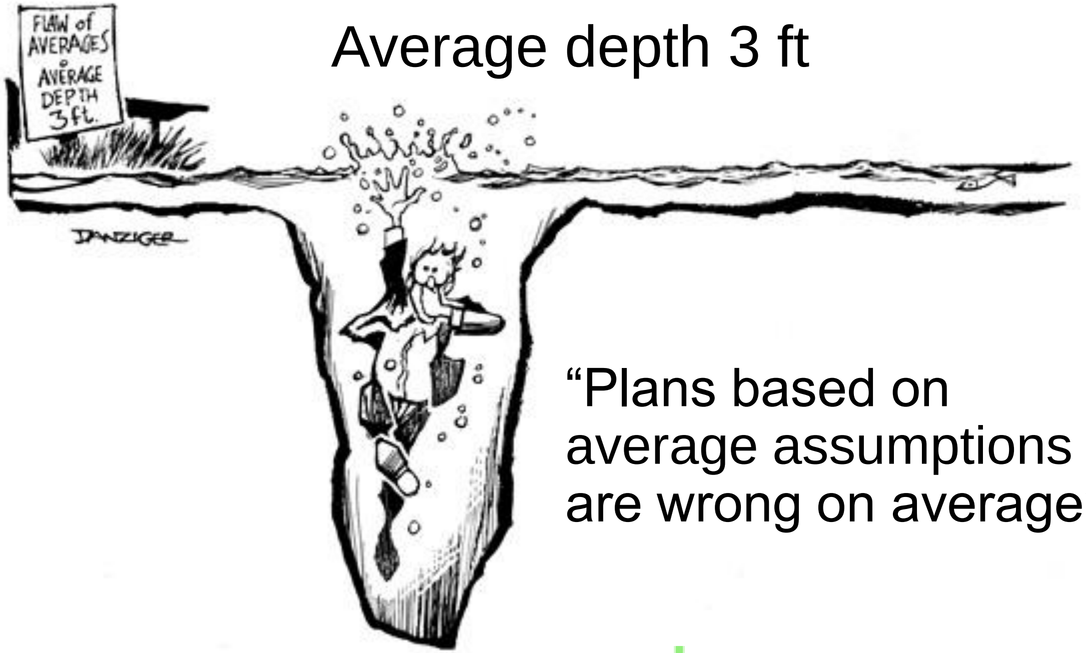
- One-number summaries
  - Mean
    Average, obtained by summing all observations and dividing by the number of obs.
  - Median
    The center value, below and above which you will find 50% of the observations.

- Summarizing your data with one number may not tell the whole story:



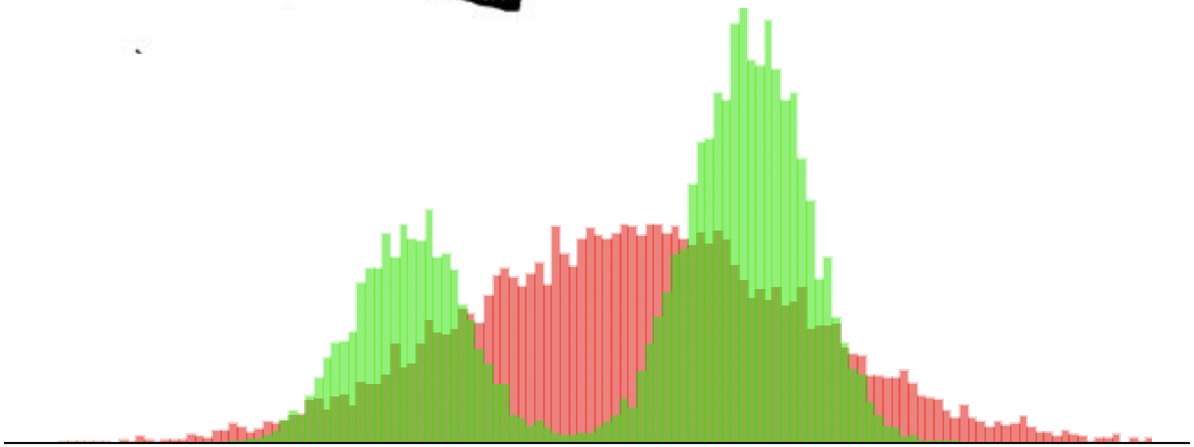
# Flaw of averages




Average depth 3 ft

“Plans based on average assumptions are wrong on average”



# Standard deviation

- The standard deviation $s$ is a measure of how spread out the $n$ observations $x_i$ are around the mean $\bar{x}$

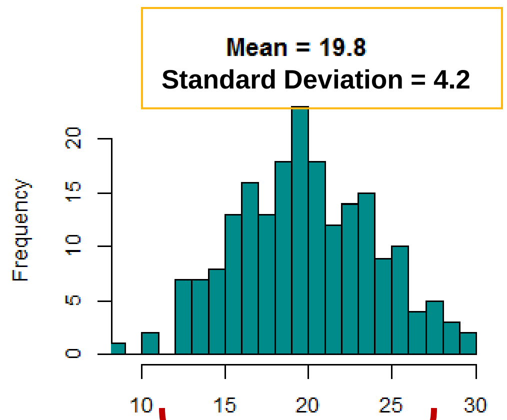
$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

- Rule of thumb for interpreting standard deviation values:

*If the data is normally distributed*

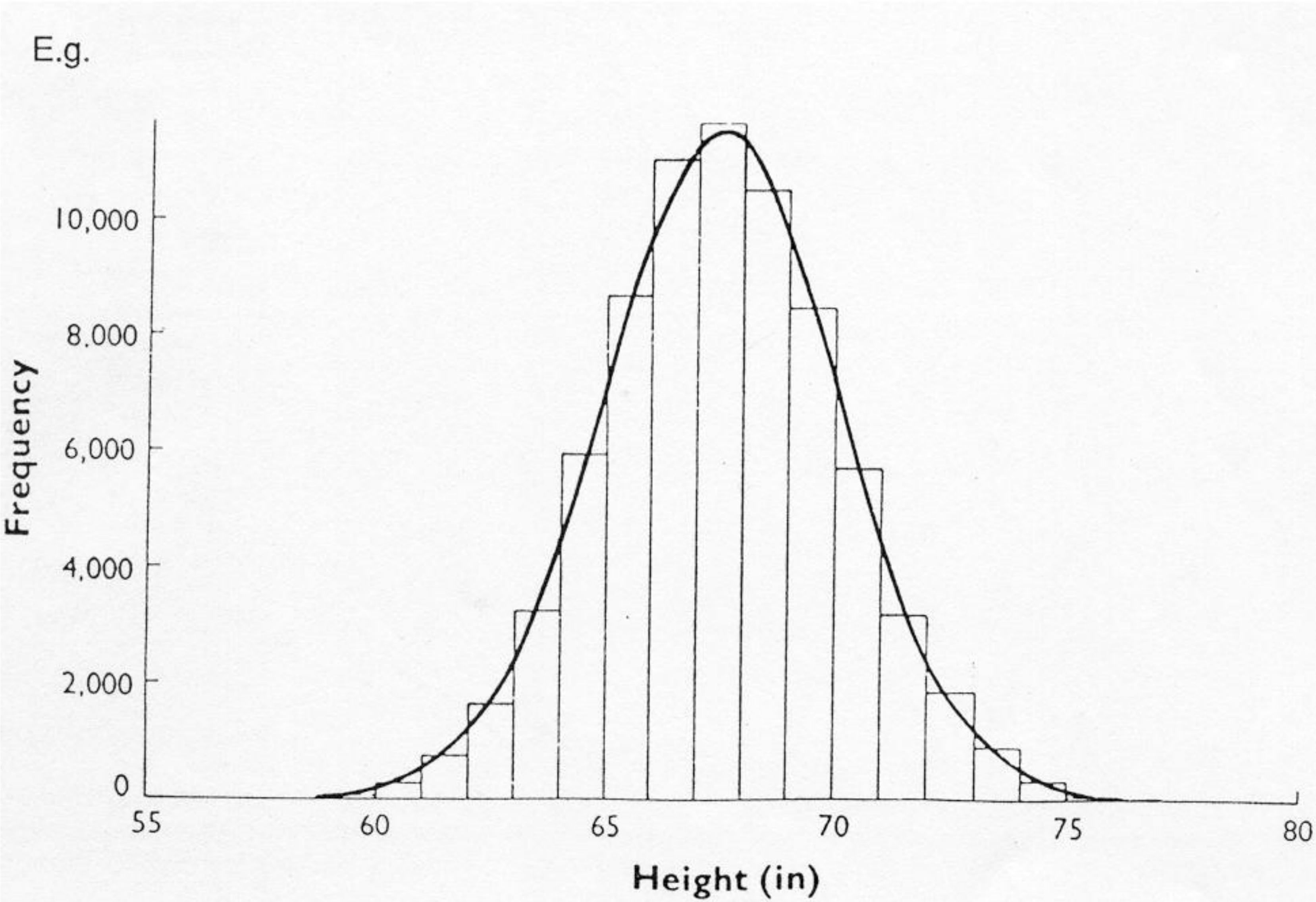
"Most observations fall within $\pm 2$ standard deviations of the mean."

*95 % of observations*

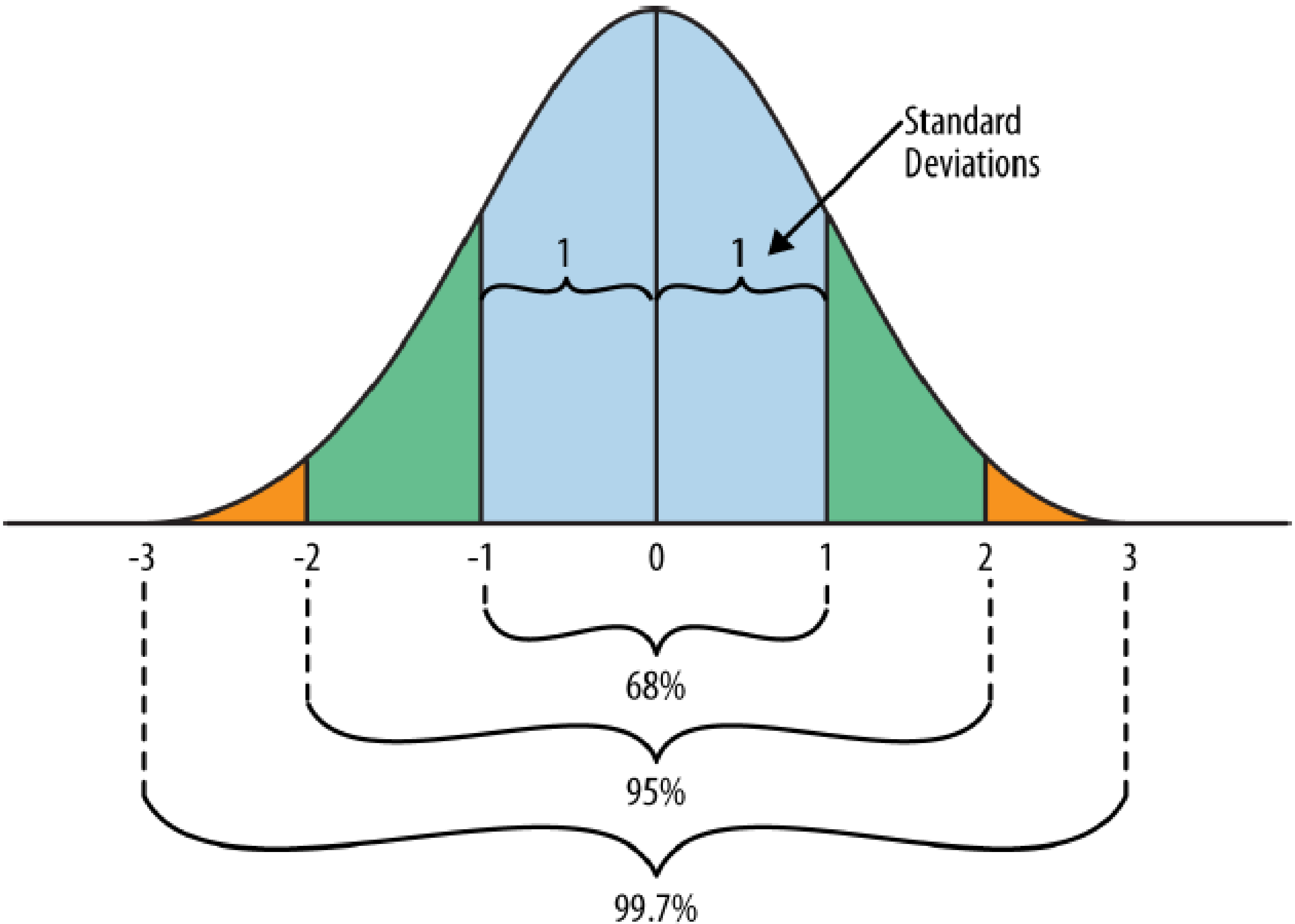



~95% of observations between 11.4 and 28.2

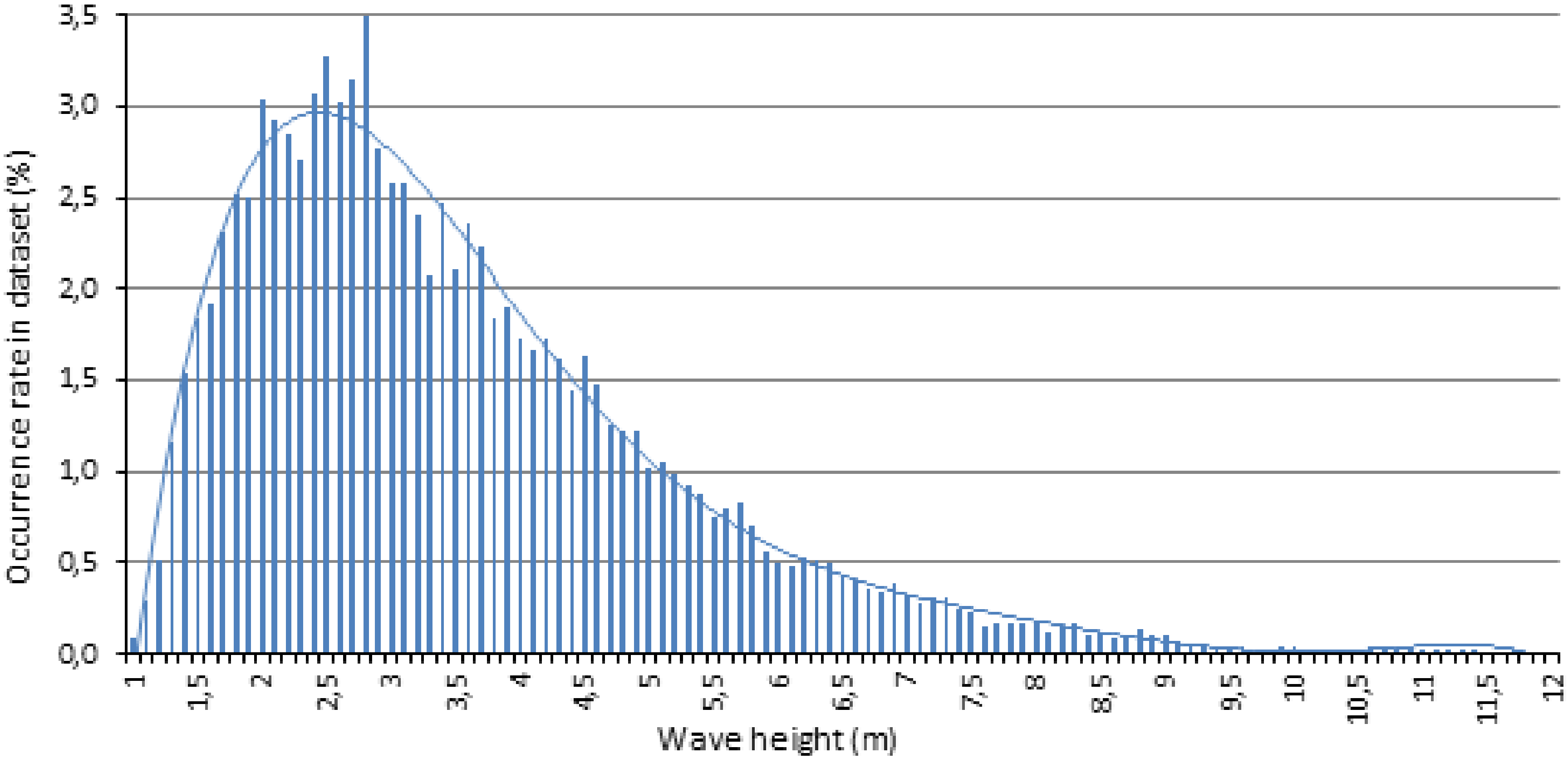
# Distributions: Normal distribution



# Distributions: Normal distribution



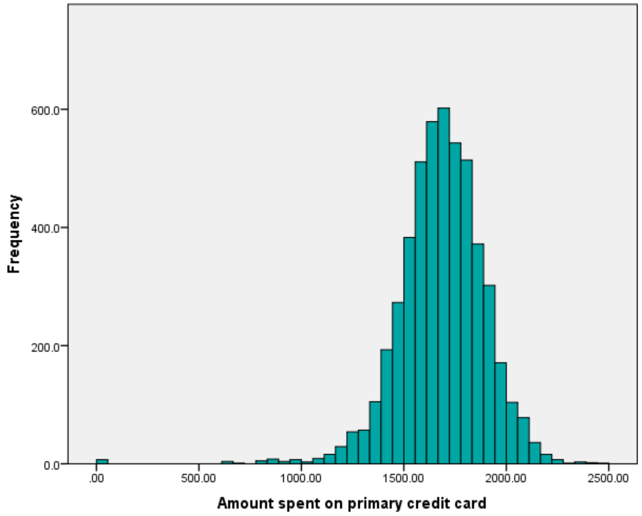
# Distributions: Non-Normal distribution

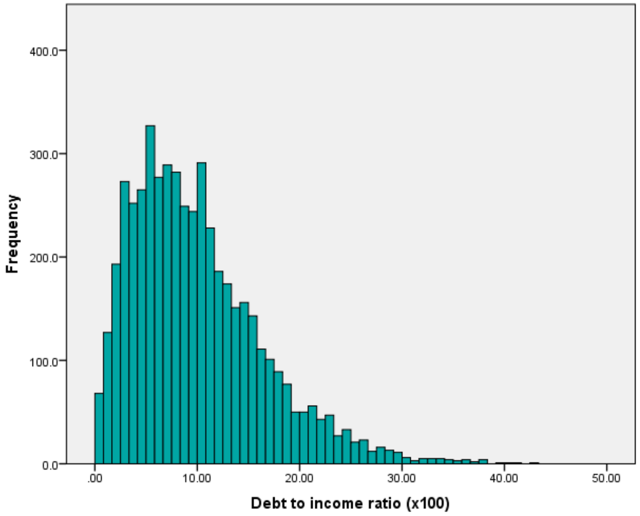


# Descriptive statistics - example

- Random sample of 5000 customers of a credit card company

| | | Amount spent on primary card last month | Debt to income ratio (x100) |
|---|---|---|---|
| N | Valid | 5000 | 5000 |
| | Missing | 0 | 0 |
| Mean | | 1683.7340 | 9.9578 |
| Median | | 1690.0670 | 8.8000 |
| Std. Deviation | | 210.26680 | 6.42317 |
| Minimum | | .00 | .00 |
| Maximum | | 2482.72 | 43.10 |



# Percentiles

- Generalizations of the median ($50^{th}$ percentile).
- The $p^{th}$ is the data point below which $p$ percent of the observations fall.

- Often used to compare a single observation to a general population.

- Examples:
  - Standardized test scores
    If you scored in the $93^{th}$ percentile, your score was higher than that of 93% of test takers.
  - Child growth percentiles
  - Stock market/Options trading
    "The call/put volume ratio of 2.15 stands in the 82nd annual percentile, pointing to a heightened demand for long calls during the last two weeks."

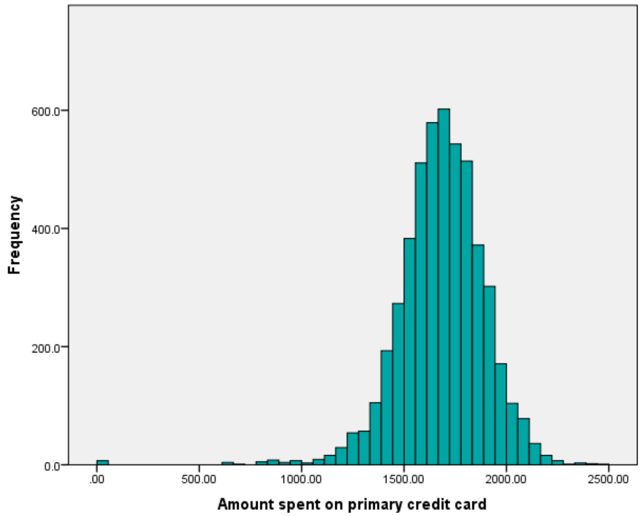
# Percentiles - example

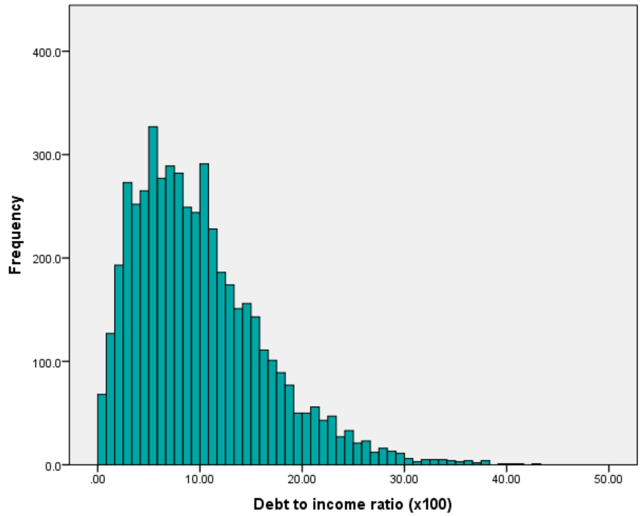
- Percentiles can be another way of describing how spread out data values are.

Example: 5-Number Summary

Minimum – 25th percentile – Median – 50th percentile - Maximum

| | | Amount spent on primary card last month | Debt to income ratio (x100) |
|---|---|---|---|
| Minimum | | .00 | .00 |
| | 25 | 1567.4658 | 5.1250 |
| Percentiles | 50 | 1690.0670 | 8.8000 |
| | 75 | 1814.5430 | 13.5000 |
| Maximum | | 2482.72 | 43.10 |



# Quantifying uncertainty – confidence intervals

- Unless we have complete data, we cannot be sure that the mean in the sample is equal to the true underlying mean (of the theoretically underlying complete data).

**One-Sample Test**

| | 95% Confidence Interval of the Difference | |
|---|---|---|
| | Lower | Upper |
| Debt to income ratio (x100) | 9.7797 | 10.1359 |
| Amount spent on primary card last month | 1677.9044 | 1689.5636 |

"We are 95% percent confident that the average Debt-to-Income ratio (x100) is between 9.78 and 10.14."

"The average Debt-to-Income ratio (x100) is 9.96 with a margin of error of .18"

- Confidence Intervals (CI) and Margins of Error (MoE) tell us how close we think the mean is to the true value, with a certain level of confidence.
- Generally, CIs and MoEs are calculated for 95% percent confidence.
  Other levels of confidence are labeled explicitly.

# Comparing means of two groups

- If two groups have different means in our data, can we conclude that the means would be different if we had complete information?
- In statistical terms, we want to test if the observed difference is ***statistically significant***.
- Once again, we consider the fact that there is uncertainty in our data.

- Example:
  In our sample of customers, women have higher Debt-to-Income ratio, but spent less on their primary credit card.
  Are these differences statistically significant?

**Group Statistics**

| | Gender | N | Mean | Std. Deviation |
|---|---|---|---|---|
| Debt to income ratio (x100) | Male | 2449 | 9.9292 | 6.37257 |
| | Female | 2551 | 9.9852 | 6.47251 |
| Amount spent on primary card last month | Male | 2449 | 356.6068 | 263.40686 |
| | Female | 2551 | 323.3435 | 231.93672 |

# Comparing means of two groups

- Example: **Independent samples t-test**

**Group Statistics**

| | Gender | N | Mean | Std. Deviation |
|---|---|---|---|---|
| Debt to income ratio (x100) | Male | 2449 | 9.9292 | 6.37257 |
| | Female | 2551 | 9.9852 | 6.47251 |
| Amount spent on primary card last month | Male | 2449 | 356.6068 | 263.40686 |
| | Female | 2551 | 323.3435 | 231.93672 |

**Independent Samples Test**

| | | t-test for Equality of Means | | | |
|---|---|---|---|---|---|
| | | t | df | Sig. (2-tailed) | Mean Difference |
| Debt to income ratio (x100) | Equal variances not assumed | -.308 | 4994.814 | .758 | -.05599 |
| Amount spent on primary card last month | Equal variances not assumed | 4.732 | 4862.365 | .000 | 33.26335 |

P-values

- A statistical test tells us whether an observed difference is statistically significant:

**P-value <.05:** The difference observed in the data is most likely not due to chance. We conclude the difference is also present in the unobserved population. ***The difference is statistically significant.***

**P-value >.05:** The difference observed could easily be simple due to chance. It is not safe to conclude that the difference is present in the underlying (unobserved) population.

# Comparing means of two groups

- Example: **Independent samples t-test**

**Independent Samples Test**

| | | t-test for Equality of Means | | | |
|---|---|---|---|---|---|
| | | t | df | Sig. (2-tailed) | Mean Difference |
| Debt to income ratio (x100) | Equal variances not assumed | -.308 | 4994.814 | .758 | -.05599 |
| Amount spent on primary card last month | Equal variances not assumed | 4.732 | 4862.365 | .000 | 33.26335 |

P-values

- In the case of Debt-to-Income ratio, we conclude that there is no significant difference between men and women (P-value = .758 >.05, not significant).
- In the case of Amount spent on primary card, we conclude that men tend to charge more on their primary card (P-value <.05, statistically significant).

- **Note**: The larger the sample, the more likely the difference of a given size will be significant.
- **Caveat**: Make sure all your observations are truly independent (repeated observations are cheating!)
- For any data scenario, there are different tests, that make their respective mathematical assumptions. When in doubt, consult your favorite statistician.

# 📖 Categorical Data

# Categorical data

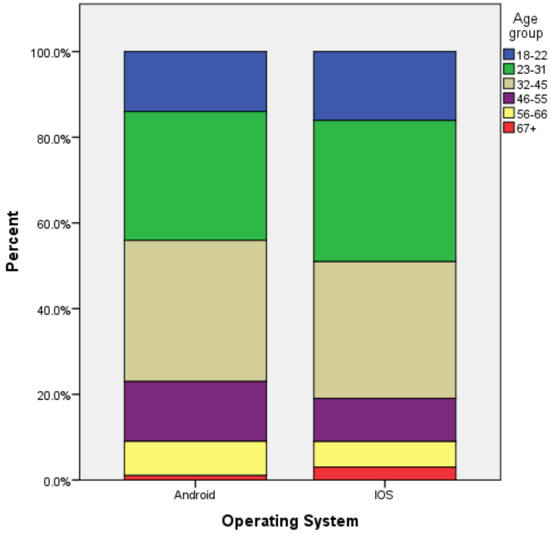
- Examples: gender, age groups, product category
- Summarize using frequencies and percentages in crosstabs

- More advanced predictive methods: Logistic Regression, Classification, …

- Example: IOS vs. Android users

| **Counts** | | Age group | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 18-22 | 23-31 | 32-45 | 46-55 | 56-66 | 67+ | Total |
| Operating System | Android | 93 | 200 | 219 | 93 | 5 | 7 | 665 |
| | IOS | 75 | 154 | 149 | 47 | 28 | 14 | 467 |
| Total | | 168 | 354 | 368 | 140 | 81 | 21 | 1132 |

| **% within Operating System** | | Age group | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 18-22 | 23-31 | 32-45 | 46-55 | 56-66 | 67+ | Total |
| Operating System | Android | 14.0% | 30.1% | 32.9% | 14.0% | 8.0% | 1.1% | 100% |
| | IOS | 16.1% | 33.0% | 31.9% | 10.1% | 6.0% | 3.0% | 100% |
| Total | | 14.8% | 31.3% | 32.5% | 12.4% | 7.2% | 1.9% | 100% |




Data source: www.forrester.com

# Margin of error for categorical data

- Confidence intervals and Margins of Error can be calculated for categorical data as well
- For this survey, the margin of error was 1.32% for 95% confidence.

| **% within Operating System** | | Age group | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 18-22 | 23-31 | 32-45 | 46-55 | 56-66 | 67+ | Total |
| Operating | Android | 14.0% | 30.1% | 32.9% | 14.0% | 8.0% | 1.1% | 100% |
| System | IOS | 16.1% | 33.0% | 31.9% | 10.1% | 6.0% | 3.0% | 100% |
| Total | | 14.8% | 31.3% | 32.5% | 12.4% | 7.2% | 1.9% | 100% |

- However, this data was based on a online survey, so the results might be biased!

# Comparative statistics for categorical data

- Is the distribution of one categorical variable independent of another categorical variable?

- Example:

  Is the distribution of age groups the same for IOS and Android users?
  It looks like IOS users tend to be younger than Android users.
  Is this difference ***statistically significant***?

| **% within Operating System** | | Age group | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 18-22 | 23-31 | 32-45 | 46-55 | 56-66 | 67+ | Total |
| Operating System | Android | 14.0% | 30.1% | 32.9% | 14.0% | 8.0% | 1.1% | 100% |
| | IOS | 16.1% | 33.0% | 31.9% | 10.1% | 6.0% | 3.0% | 100% |
| Total | | 14.8% | 31.3% | 32.5% | 12.4% | 7.2% | 1.9% | 100% |

# Comparative statistics for categorical data

- Example:

  Is the distribution of age groups the same for IOS and Android users?
  It looks like IOS users tend to be younger than Android users.
  Is this difference ***statistically significant***?

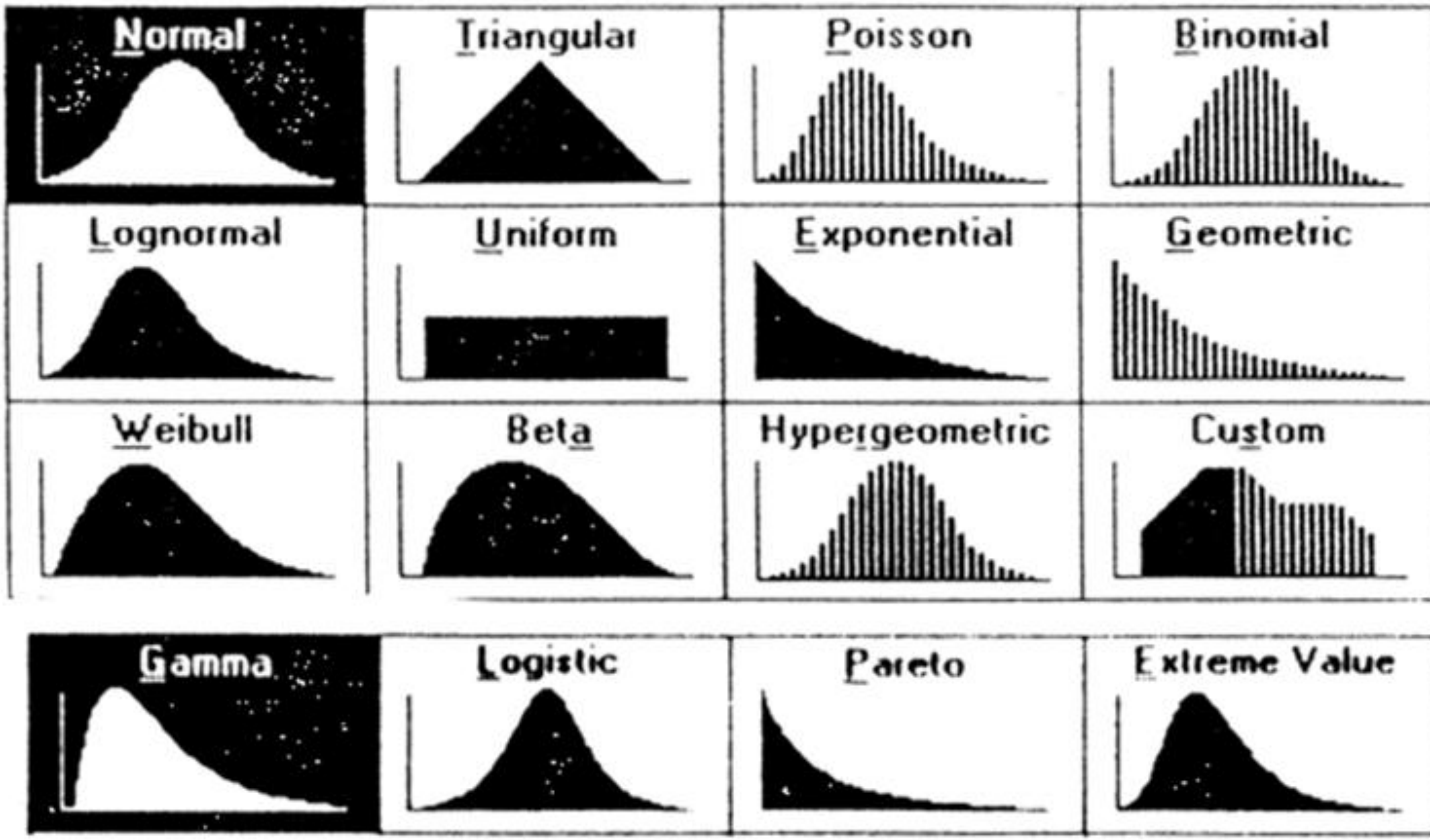
| **% within Operating System** | | Age group | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 18-22 | 23-31 | 32-45 | 46-55 | 56-66 | 67+ | Total |
| Operating System | Android | 14.0% | 30.1% | 32.9% | 14.0% | 8.0% | 1.1% | 100% |
| | IOS | 16.1% | 33.0% | 31.9% | 10.1% | 6.0% | 3.0% | 100% |
| Total | | 14.8% | 31.3% | 32.5% | 12.4% | 7.2% | 1.9% | 100% |

**Chi-Square Test**

| | Value | df | Asymp. Sig. (2-sided) |
|---|---|---|---|
| Pearson Chi-Square | 12.123[a] | 5 | .033 |
| N of Valid Cases | 1132 | | |

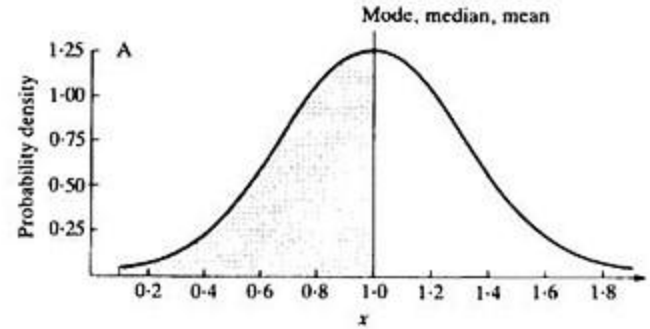
# 📖 Distributions

# Distributions

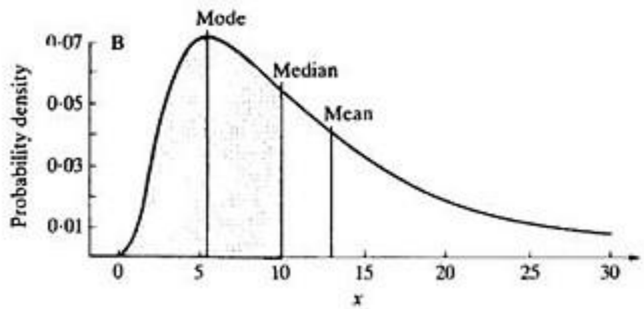


# Distributions

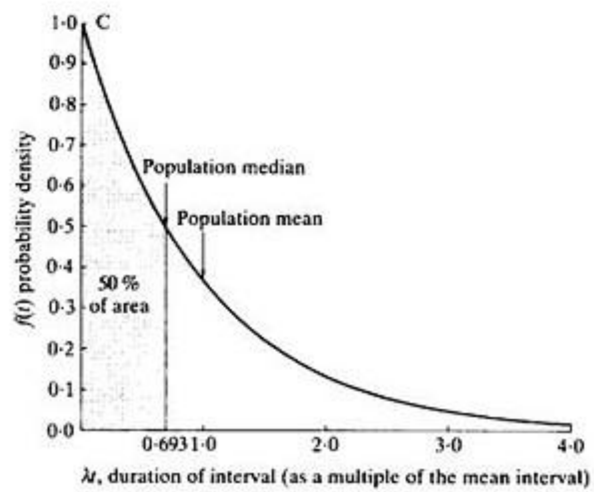
**Gaussian p.d.f.**




**Positively-skewed p.d.f. (e.g. lognormal')**




**Exponential p.d.f.**

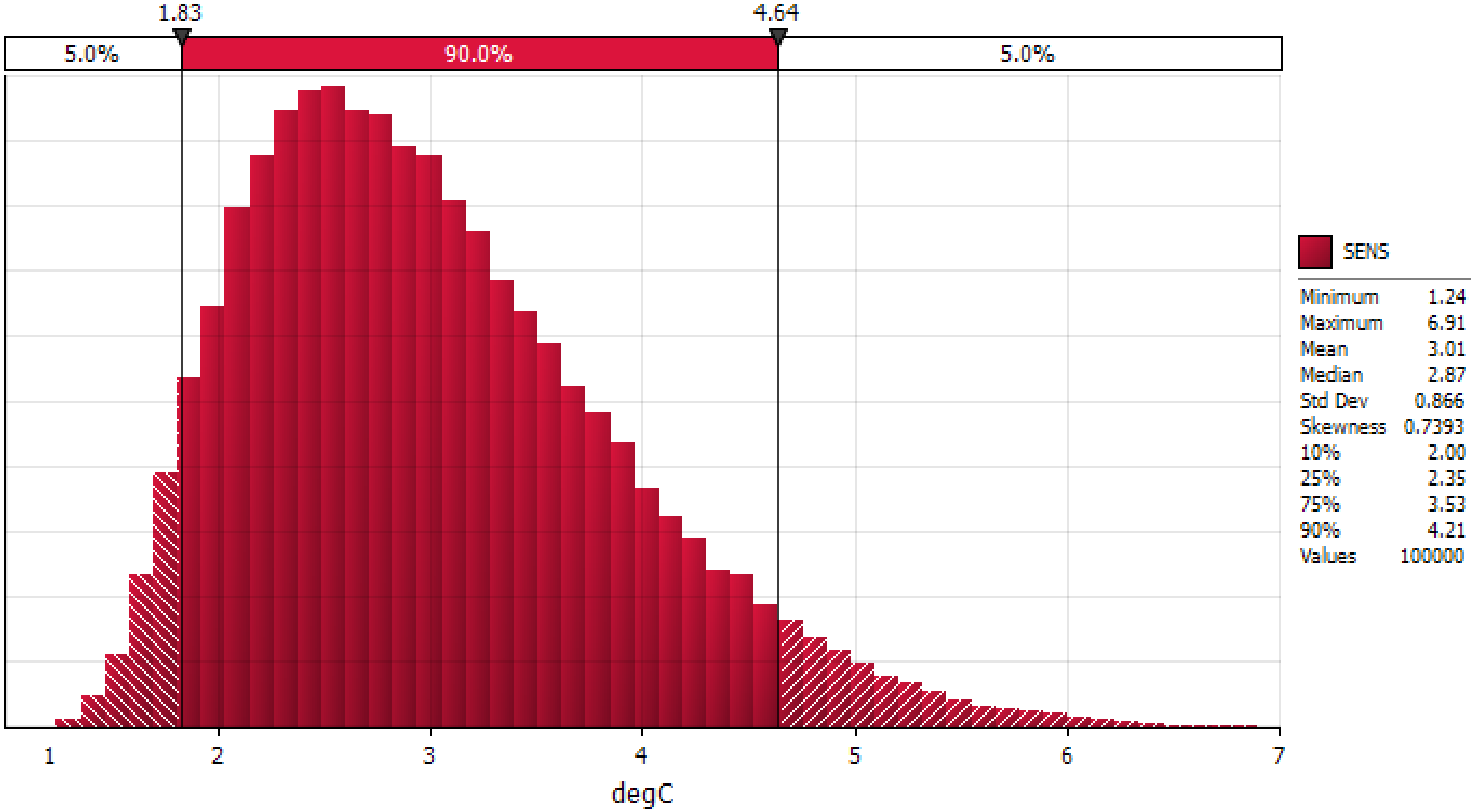


# Continuous distributions

| | Notation | $F_X(x)$ | $f_X(x)$ | $\mathbb{E}[X]$ | $\mathbb{V}[X]$ | $M_X(s)$ |
|---|---|---|---|---|---|---|
| Uniform | $\mathrm{Unif}(a,b)$ | $\begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases}$ | $\frac{I(a < x < b)}{b-a}$ | $\frac{a+b}{2}$ | $\frac{(b-a)^2}{12}$ | $\frac{e^{sb}-e^{sa}}{s(b-a)}$ |
| Normal | $\mathcal{N}(\mu,\sigma^2)$ | $\Phi(x) = \int_{-\infty}^{x} \phi(t)\,dt$ | $\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$ | $\mu$ | $\sigma^2$ | $\exp\left\{\mu s + \frac{\sigma^2 s^2}{2}\right\}$ |
| Log-Normal | $\ln\mathcal{N}(\mu,\sigma^2)$ | $\frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left[\frac{\ln x - \mu}{\sqrt{2\sigma^2}}\right]$ | $\frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(\ln x-\mu)^2}{2\sigma^2}\right\}$ | $e^{\mu+\sigma^2/2}$ | $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$ | |
| Multivariate Normal | $\mathrm{MVN}(\mu,\Sigma)$ | | $(2\pi)^{-k/2}\lvert\Sigma\rvert^{-1/2} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$ | $\mu$ | $\Sigma$ | $\exp\left\{\mu^T s + \frac{1}{2}s^T\Sigma s\right\}$ |
| Student's $t$ | $\mathrm{Student}(\nu)$ | $I_x\left(\frac{\nu}{2},\frac{\nu}{2}\right)$ | $\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}$ | $0$ | $0$ | |
| Chi-square | $\chi^2_k$ | $\frac{1}{\Gamma(k/2)}\gamma\left(\frac{k}{2},\frac{x}{2}\right)$ | $\frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2}e^{-x/2}$ | $k$ | $2k$ | $(1-2s)^{-k/2}\ s < 1/2$ |
| F | $\mathrm{F}(d_1,d_2)$ | $I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2},\frac{d_1}{2}\right)$ | $\frac{\sqrt{\frac{(d_1 x)^{d_1} d_2^{d_2}}{(d_1 x + d_2)^{d_1+d_2}}}}{x\mathrm{B}\left(\frac{d_1}{2},\frac{d_1}{2}\right)}$ | $\frac{d_2}{d_2-2}$ | $\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$ | |
| Exponential | $\mathrm{Exp}(\beta)$ | $1-e^{-x/\beta}$ | $\frac{1}{\beta}e^{-x/\beta}$ | $\beta$ | $\beta^2$ | $\frac{1}{1-\beta s}\ (s < 1/\beta)$ |
| Gamma | $\mathrm{Gamma}(\alpha,\beta)$ | $\frac{\gamma(\alpha, x/\beta)}{\Gamma(\alpha)}$ | $\frac{1}{\Gamma(\alpha)\,\beta^\alpha}x^{\alpha-1}e^{-x/\beta}$ | $\alpha\beta$ | $\alpha\beta^2$ | $\left(\frac{1}{1-\beta s}\right)^\alpha\ (s < 1/\beta)$ |
| Inverse Gamma | $\mathrm{InvGamma}(\alpha,\beta)$ | $\frac{\Gamma\left(\alpha,\frac{\beta}{x}\right)}{\Gamma(\alpha)}$ | $\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$ | $\frac{\beta}{\alpha-1}\ \alpha > 1$ | $\frac{\beta^2}{(\alpha-1)^2(\alpha-2)^2}\ \alpha > 2$ | $\frac{2(-\beta s)^{\alpha/2}}{\Gamma(\alpha)}K_\alpha\left(\sqrt{-4\beta s}\right)$ |
| Dirichlet | $\mathrm{Dir}(\alpha)$ | | $\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)}\prod_{i=1}^k x_i^{\alpha_i-1}$ | $\frac{\alpha_i}{\sum_{i=1}^k \alpha_i}$ | $\frac{\mathbb{E}[X_i](1-\mathbb{E}[X_i])}{\sum_{i=1}^k \alpha_i + 1}$ | |
| Beta | $\mathrm{Beta}(\alpha,\beta)$ | $I_x(\alpha,\beta)$ | $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ | $\frac{\alpha}{\alpha+\beta}$ | $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ | $1+\sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{s^k}{k!}$ |
| Weibull | $\mathrm{Weibull}(\lambda,k)$ | $1-e^{-(x/\lambda)^k}$ | $\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}$ | $\lambda\Gamma\left(1+\frac{1}{k}\right)$ | $\lambda^2\Gamma\left(1+\frac{2}{k}\right)-\mu^2$ | $\sum_{n=0}^{\infty}\frac{s^n\lambda^n}{n!}\Gamma\left(1+\frac{n}{k}\right)$ |
| Pareto | $\mathrm{Pareto}(x_m,\alpha)$ | $1-\left(\frac{x_m}{x}\right)^\alpha\ x \geq x_m$ | $\alpha\frac{x_m^\alpha}{x^{\alpha+1}}\ \ x \geq x_m$ | $\frac{\alpha x_m}{\alpha-1}\ \alpha > 1$ | $\frac{x_m^\alpha}{(\alpha-1)^2(\alpha-2)}\ \alpha > 2$ | $\alpha(-x_m s)^\alpha\Gamma(-\alpha,-x_m s)\ s < 0$ |

# Distributions

Estimate of the probability distribution of global mean temperature resulting from a doubling of $CO_2$ relative to its pre-industrial value, made from 100000 simulations



# 📖 Central Limit Theorem

# Central Limit Theorem

**Arithmetic means** from a **sufficiently large number of random samples** from the entire population will be **Normally distributed** around the population mean (regardless of the distribution in the population)

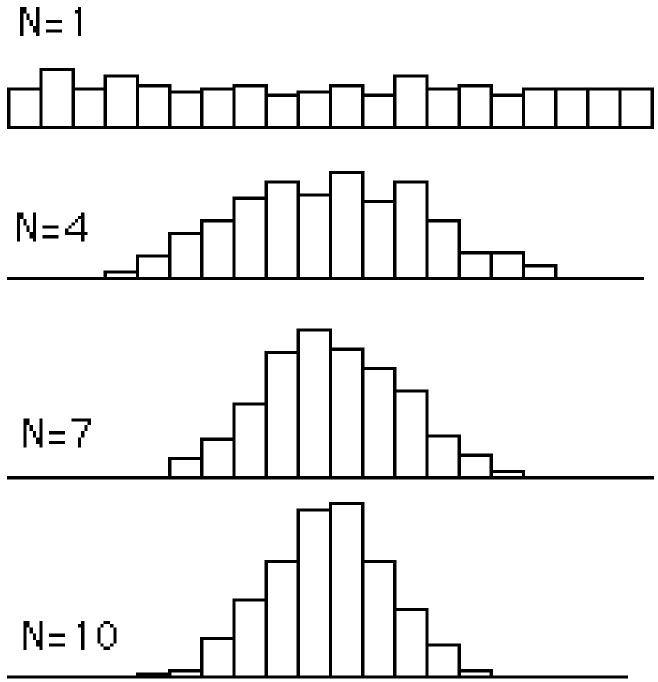
If $\mathbb{E}(x_i) = \mu$ and $\mathrm{var}(x_i) = \sigma^2$ for all *i* (and independent) then:

$$x_1 + \ldots + x_n \sim \mathcal{N}(n \cdot \mu,\ n \cdot \sigma^2)$$

$$\bar{x} = \tfrac{x_1 + \ldots + x_n}{n} \sim \mathcal{N}(\mu,\ \sigma^2/n)$$

## Central Limit Theorem – example

On the right are shown the resulting frequency distributions each based on 500 means. For *n* = 4, 4 scores were sampled from a uniform distribution 500 times and the mean computed each time. The same method was followed with means of 7 scores for *n* = 7 and 10 scores for *n* = 10.




When ***n*** increases:

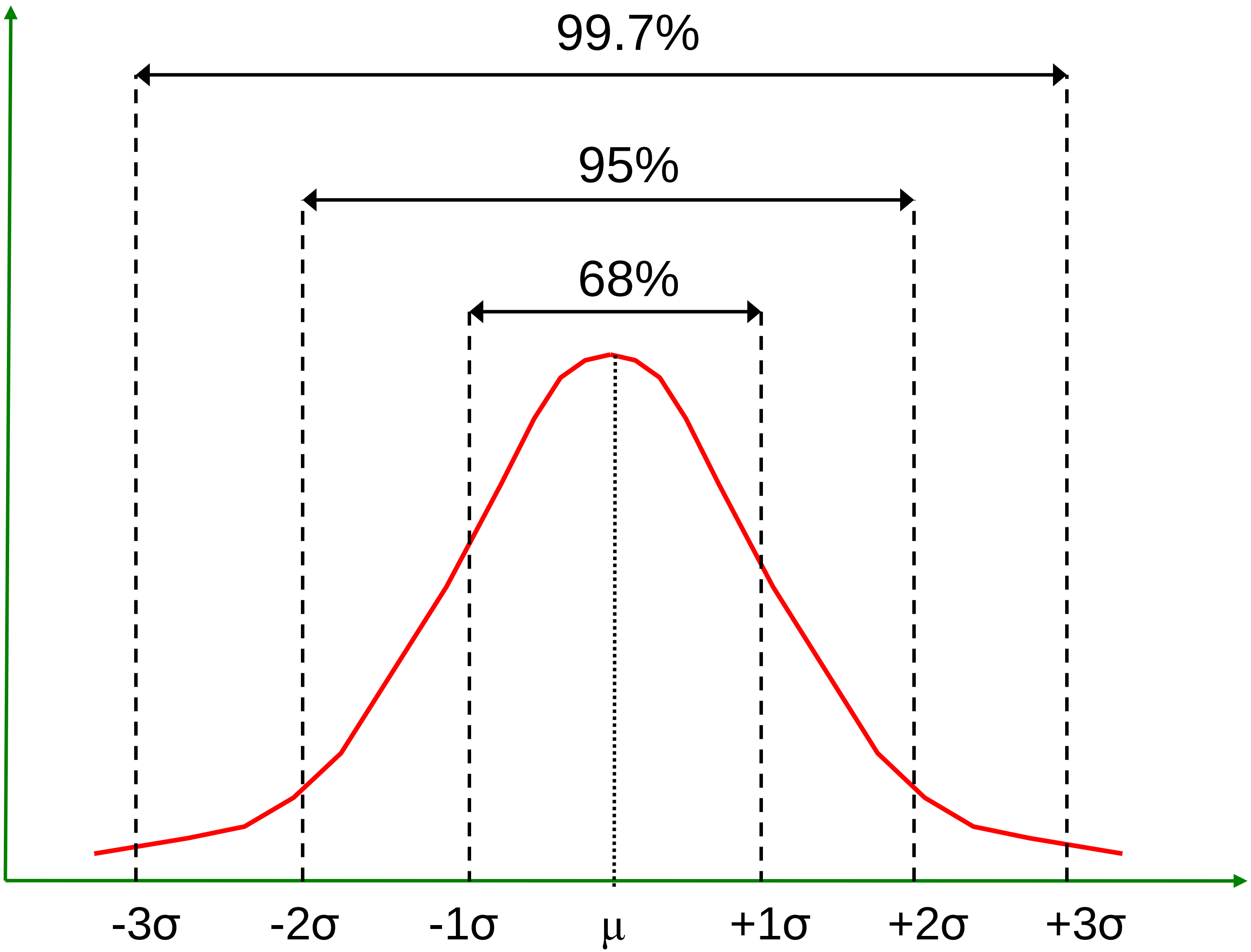
1. The distributions becomes more and more Normal
2. The spread of the distributions decreases

Source: http://davidmlane.com/hyperstat/A14461.html

# Central Limit Theorem

- The **sampling distribution of the mean** roughly follows a **Normal distribution**
- **95%** of the time, an individual sample mean should lie within 2 (actually **1.96**) standard deviations of the mean

$$\text{prob}\left[(\mu - 1.96s) \leq \bar{x} \leq (\mu + 1.96s)\right] = 0.95$$




P(Z>=2.0) = 0.0228 P(-2<=Z<=+2) = 1 – 2*0.0228 = 0.9544

P(Z>=1.96) = 0.025 P(-1.96<=Z<=+1.96) = 1 – 2*0.025 = 0.95

## Central Limit Theorem

- The **standard deviation** *s* of the sampling distribution of the mean of *x* is:

$$s^2 = \frac{\sigma^2}{n} \qquad s = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\text{prob}\,[(\mu - 1.96s) \leq \bar{x} \leq (\mu + 1.96s)] = 0.95$$

$$\text{prob}\left[(\mu - 1.96\frac{\sigma}{\sqrt{n}}) \leq \bar{x} \leq (\mu + 1.96\frac{\sigma}{\sqrt{n}})\right] = 0.95$$

Rearranging

margin of error

$$\text{prob}\left[(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}) \leq \mu \leq (\bar{x} + 1.96\frac{\sigma}{\sqrt{n}})\right] = 0.95$$

# Central Limit Theorem – election poll example

- Suppose we conduct a poll to try and get the outcome of an upcoming **election** with two candidates. We poll 1000 people, and 550 of them respond that they will vote for candidate A
- **How confident** can we be that a given person will cast their vote for candidate A?
- In this case we are working with a **binomial distribution** (i.e., a voter can choose Candidate A or B, which is a binomial function)
- We have a probability **estimator** from our sample, where the probability of an individual in our sample voting for candidate A was found to be 550/1000=0.55
- For the **binominal distribution**

$$s = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p \cdot (1-p)}}{\sqrt{n}} = \frac{\sqrt{0.55 \cdot 0.45}}{\sqrt{1000}} = 0.0157$$

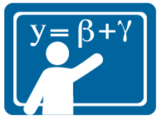
- **Margin of error** = 1.96 * 0.0157 = 0.031 = 3%

# 📖 Summary of Lecture 3

# Summary – good practices for data analysis

- Be aware of where your data comes from and how it was collected

- Plot your data
- Choose the appropriate summary statistics for your type of data




- Statistics generally have uncertainty associated with them
  - Keep standard deviation and confidence intervals in mind when interpreting results
  - Perform statistical tests to see if the difference in the data indicate a statistically significant difference
- Get familiar with distributions