
Oleksandr Romanko, Ph.D.

Associate Director, Financial Risk Quantitative Research, SS&C Algorithmics

Adjunct Professor, University of Toronto

MIE1624H – Introduction to Data Science and Analytics

Lecture 3 – Basic Statistics

University of Toronto

September 26, 2023

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

Nov 9, 2016 6:33 pm ET

 9 COMMENTS

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



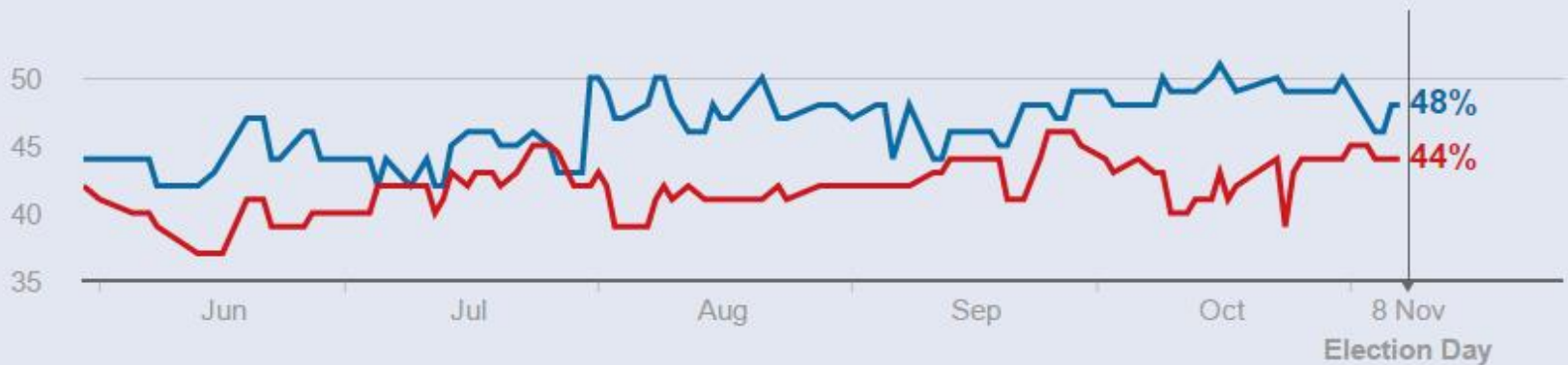
48%

Hillary Clinton



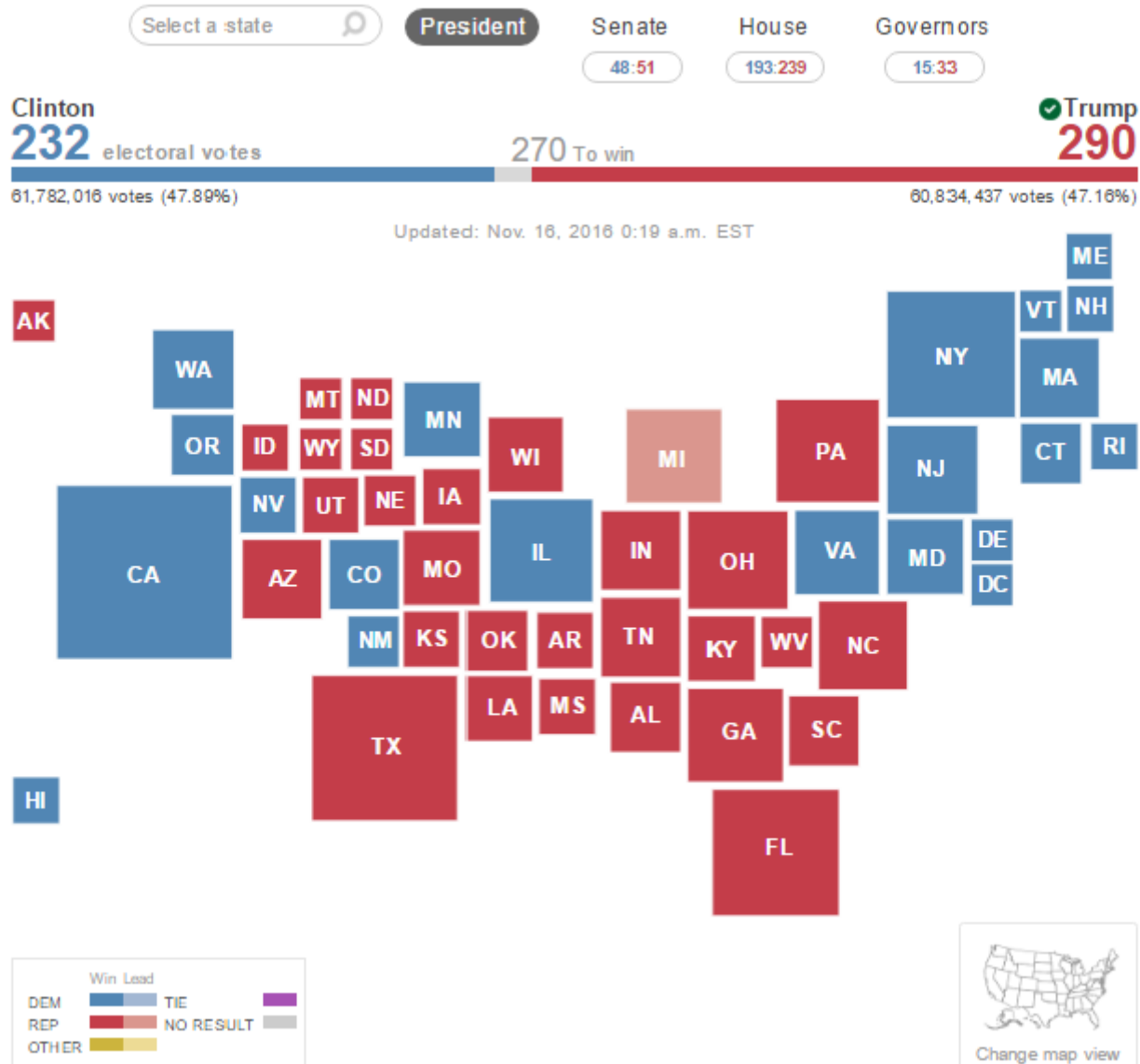
44%

Donald Trump



Source: BBC poll of polls

US presidential elections 2016





US presidential elections 2020



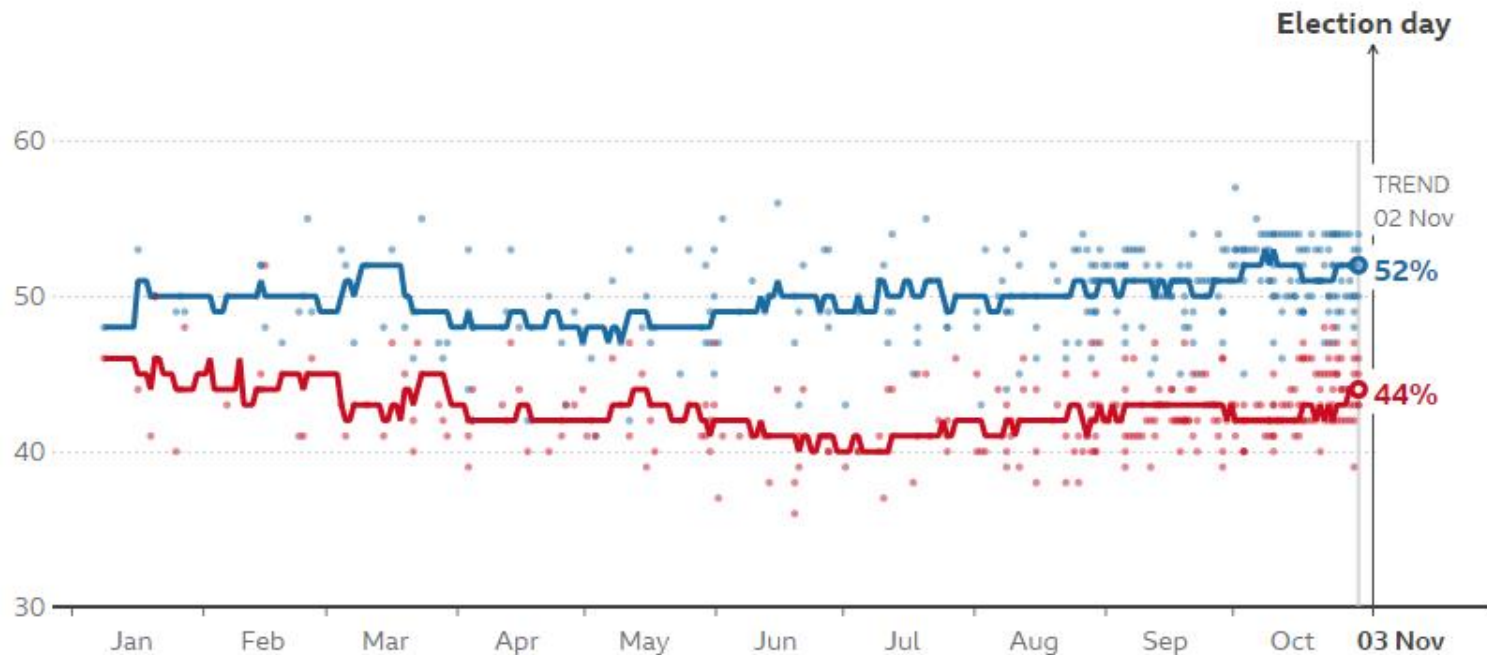
DEMOCRAT
BIDEN
52%

REPUBLICAN
TRUMP
44%



 Trend line showing average voting intention based on  individual polls

☒ Show individual polls



Source: BBC poll of polls <https://www.bbc.com/news/election-us-2020-53657174>

US presidential elections 2020

▼ Date	▼ Poll	▼ BIDEN	▼ TRUMP	▼ MoE	▼ Sample
Nov 02	Economist/YouGov	53	43	3.3	1,363 LV
Nov 02	IBD/TIPP*	50	46	3.2	1,212 LV
Nov 02	Reuters/Ipsos	52	45	3.7	914 LV
Nov 02	USC Dornsife	54	43	-	5,423 LV
Nov 01	CNBC/Change Research (D)	52	42	2.3	1,880 LV
Nov 01	IBD/TIPP	50	45	3.2	1,080 LV
Nov 01	IBD/TIPP*	49	46	3.2	1,080 LV
Nov 01	Quinnipiac	50	39	2.5	1,516 LV
Nov 01	Rasmussen Reports	48	47	2.5	1,500 LV
Nov 01	USC Dornsife	53	43	-	5,382 LV
Oct 31	IBD/TIPP	50	44	3.2	1,072 LV
Oct 31	JTN/RMG Research*	51	44	2.8	1,200 LV
Oct 31	NBC News/Wall St. Jnl	52	42	3.4	1,000 RV
Oct 31	SurveyUSA	52	44	3.2	1,265 LV

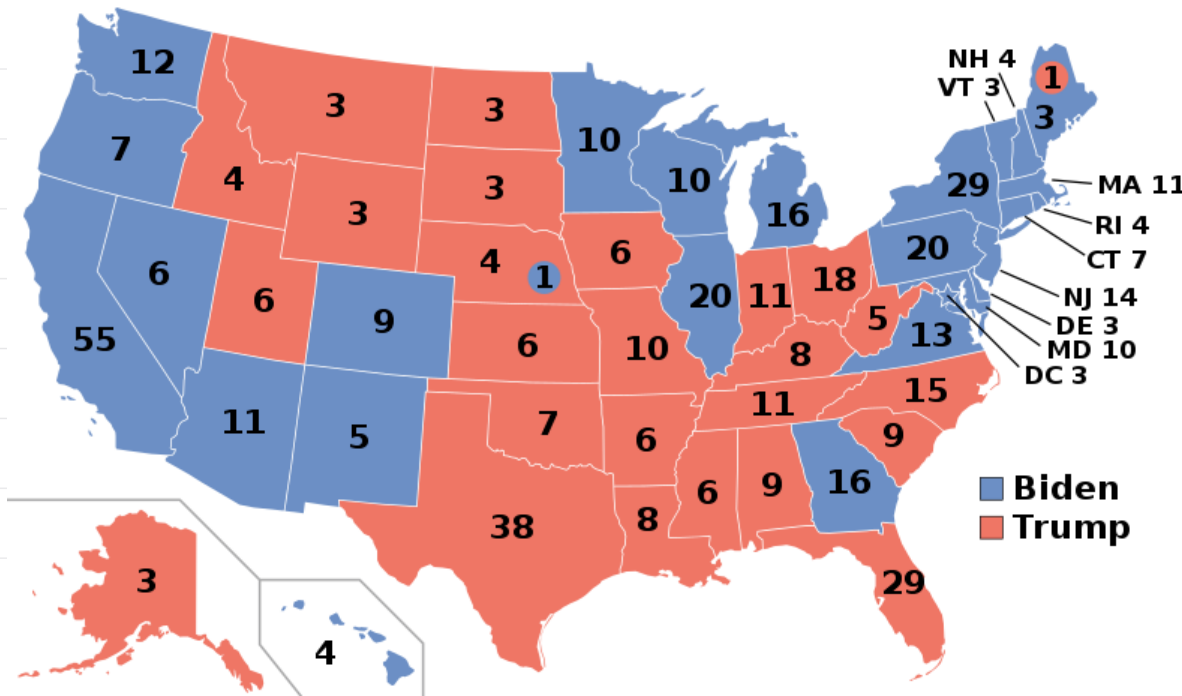
Source: <https://www.bbc.com/news/election-us-2020-54094119>

RV – registered voters
LV – likely voters

US presidential elections 2020

Latest polling averages in battleground states

▲ State	▼ BIDEN	▼ TRUMP	▼ Who won in 2016?
Arizona	47.9%	47.0%	Trump by 3.6%
Florida	47.9%	47.0%	Trump by 1.2%
Georgia	47.2%	48.2%	Trump by 5.2%
Iowa	45.6%	47.6%	Trump by 9.5%
Michigan	50.0%	45.8%	Trump by 0.2%
Minnesota	48.0%	43.7%	
Nevada	48.7%	46.3%	
New Hampshire	53.4%	42.4%	
North Carolina	47.6%	47.8%	
Ohio	46.3%	47.3%	
Pennsylvania	48.7%	47.5%	
Texas	46.5%	47.8%	
Virginia	51.7%	40.3%	
Wisconsin	51.0%	44.3%	



Source: Real Clear Politics, Associated Press. Last updated: 3 November

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)

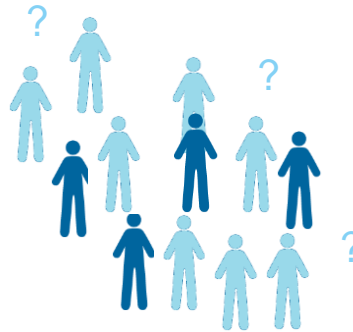
- Type of data drives the direction of your analysis
 - How to plot
 - How to summarize
 - How to draw inferences and conclusions
 - How to issue predictions

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)



Greater
uncertainty

Sparse data

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

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

Sources of uncertainty

Uncertainty from data collection

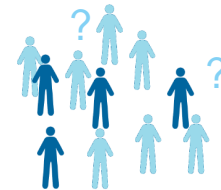
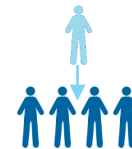


Uncertainty in model



Uncertainty in descriptive statistics, predictions and forecasts

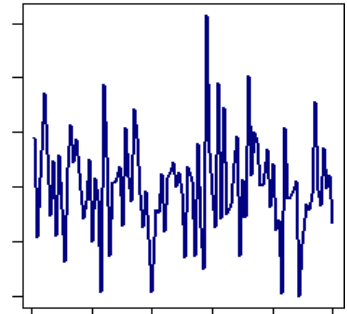
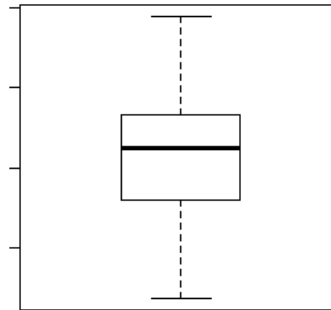
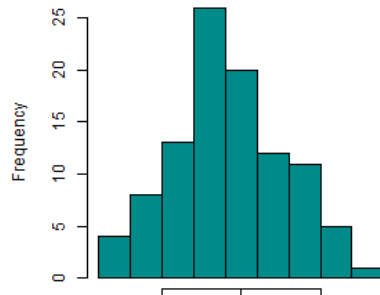
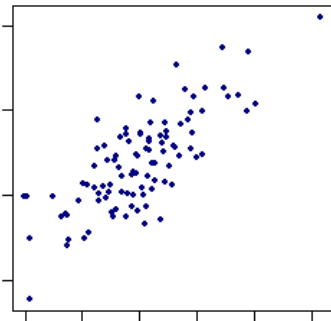
- Average vs. Individual (Standard Deviation)
- Data vs. Reality (Confidence Interval, Margin of Error)
- Prediction/Forecast (Prediction Intervals)



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

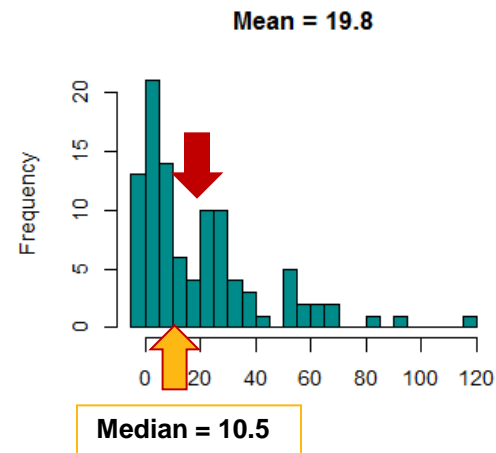
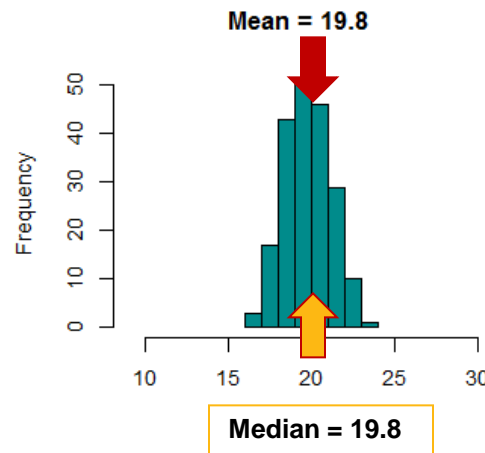
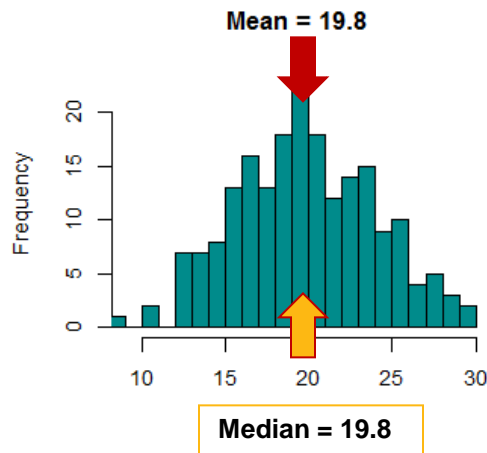
- Mean

- Average, obtained by summing all observations and dividing by the number of obs.

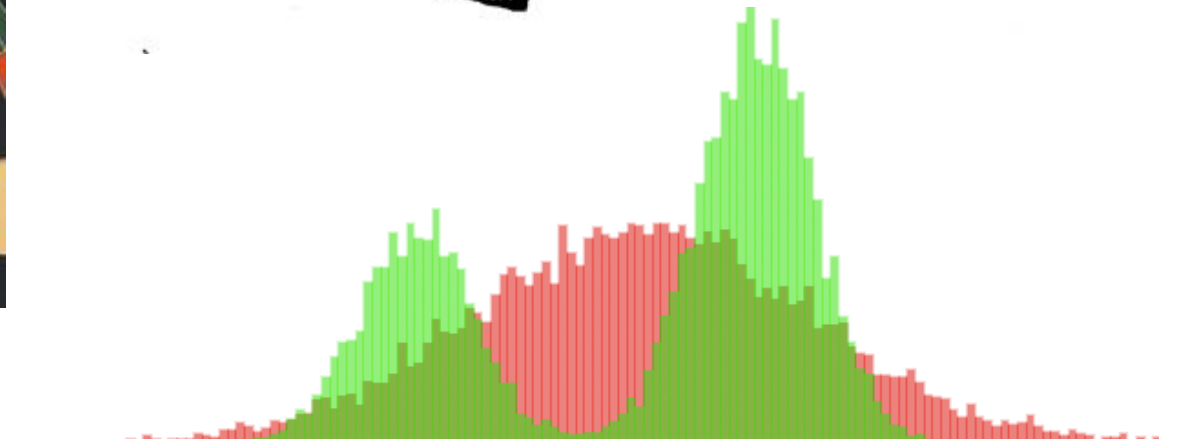
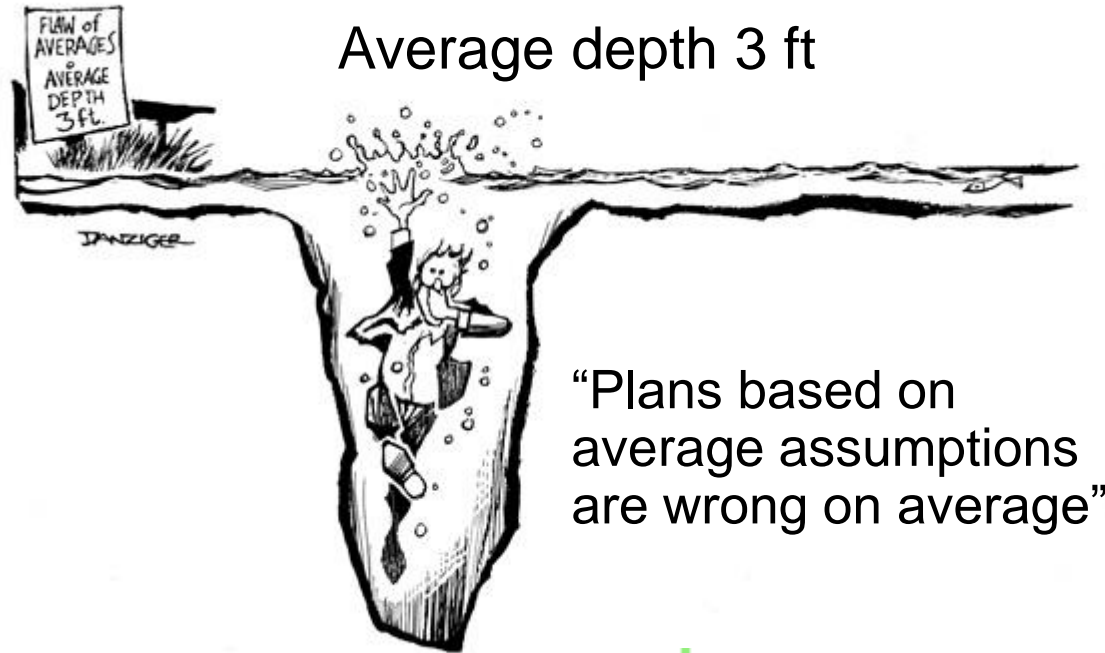
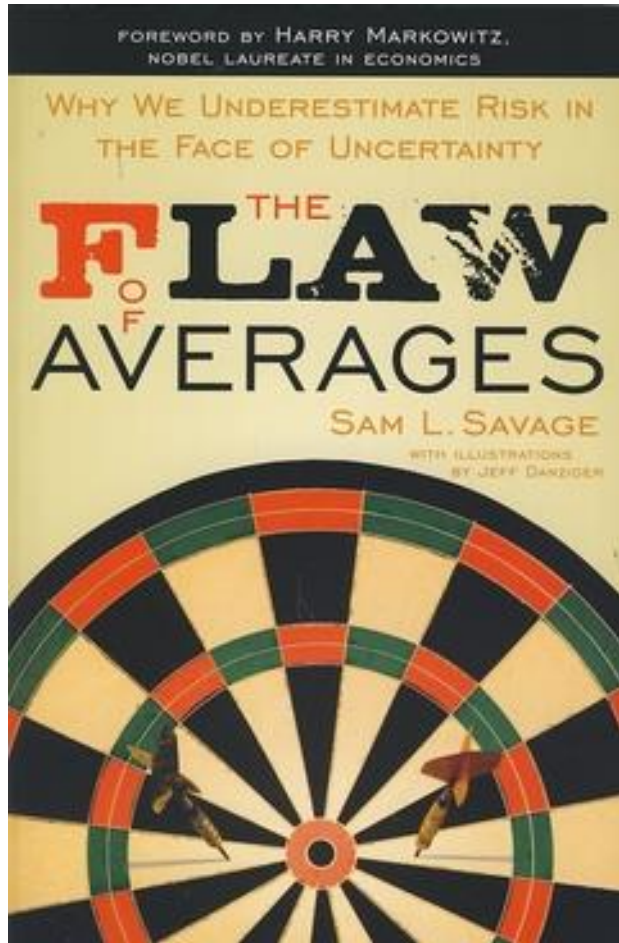
- Median

- 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



Standard deviation

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

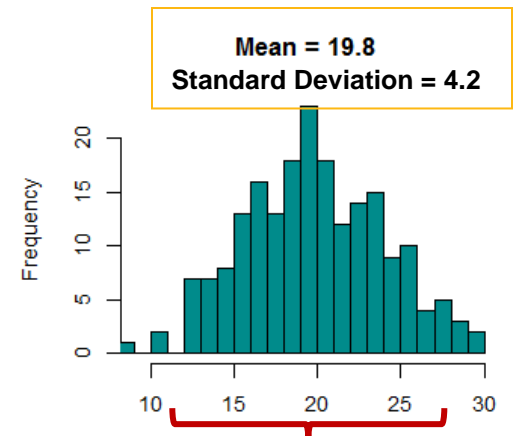
$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

- Rule of thumb for interpreting standard deviation values:

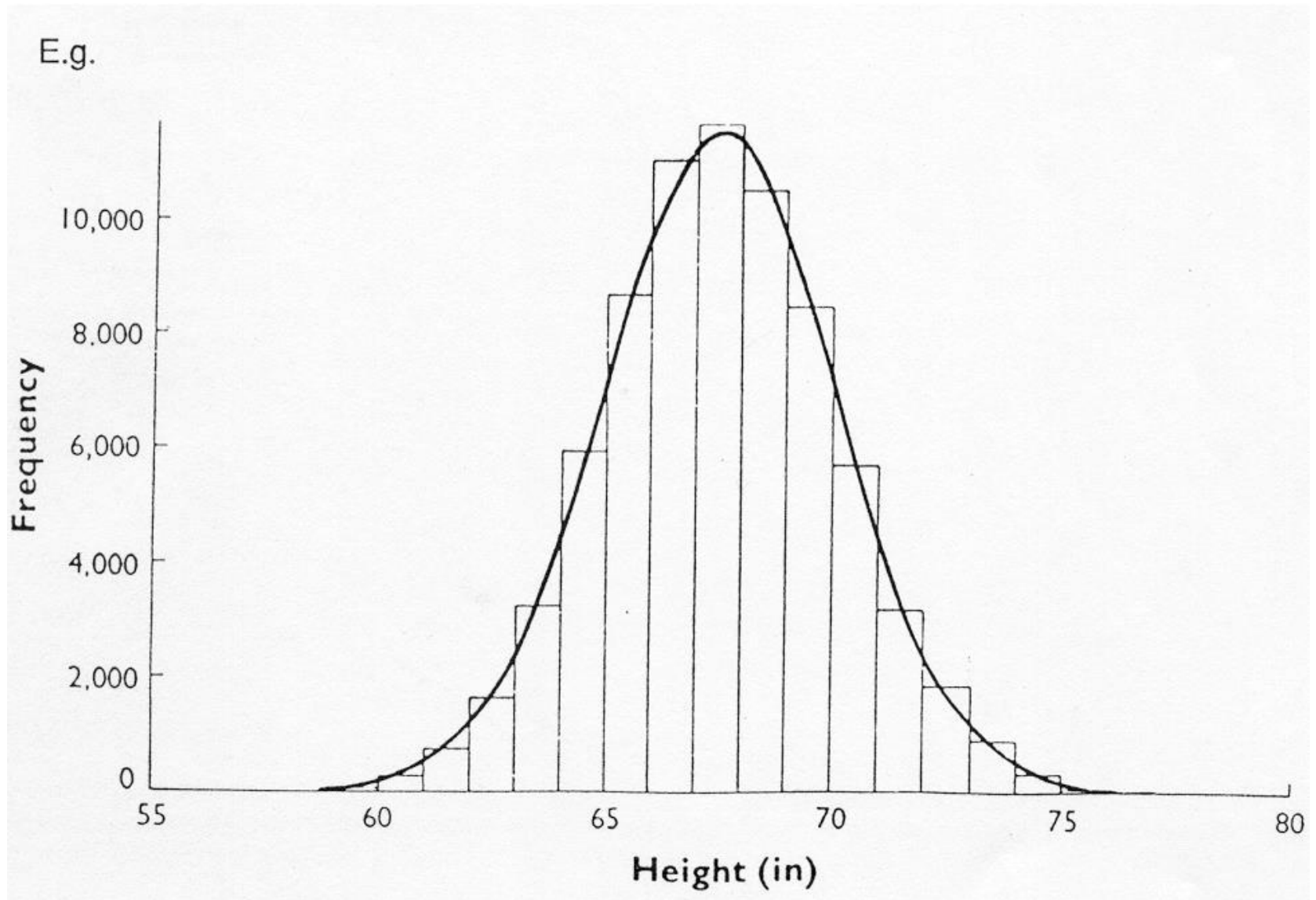
If the data is normally distributed

“Most observations fall within ± 2 standard deviations of the mean.”

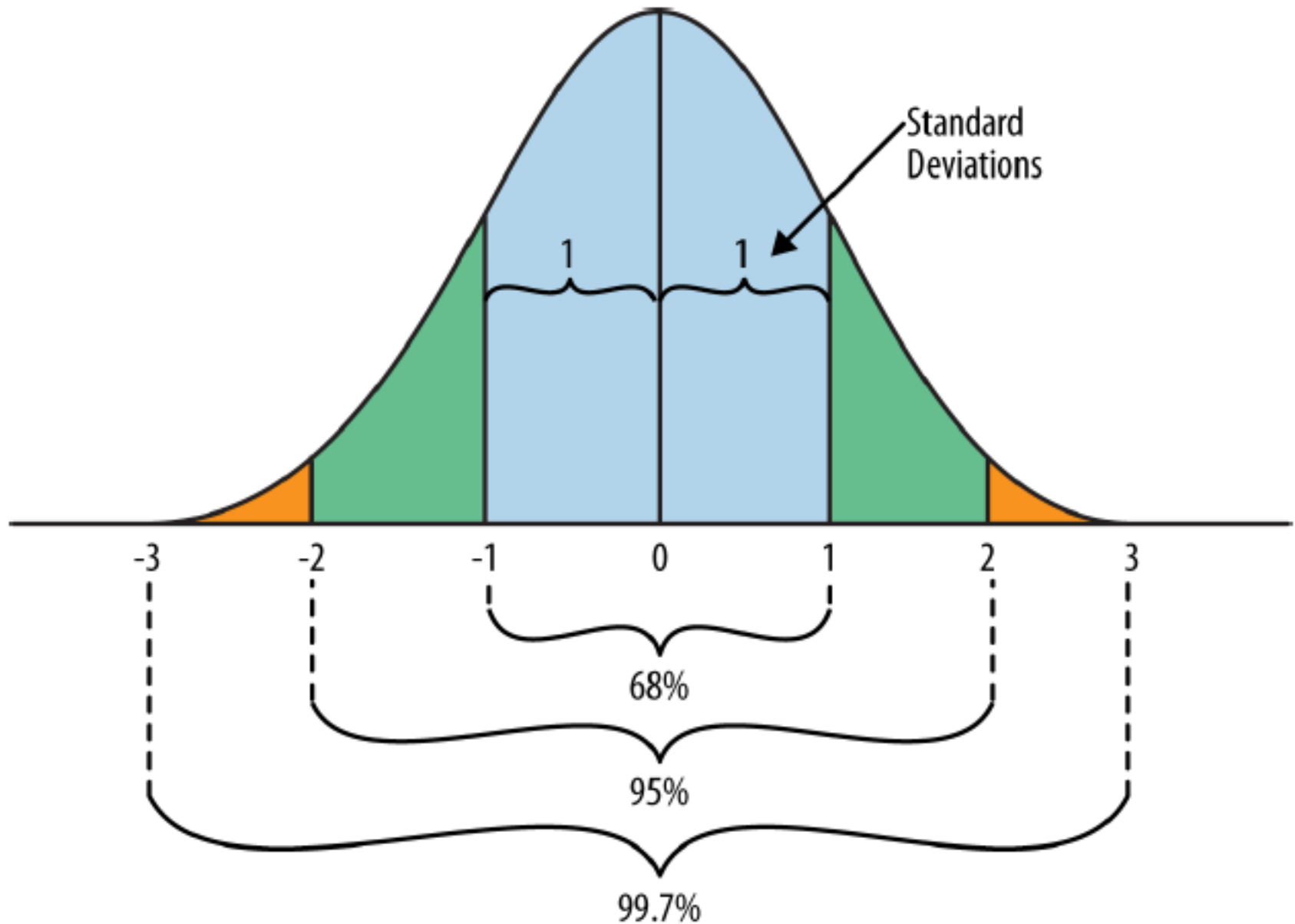
95 % of observations



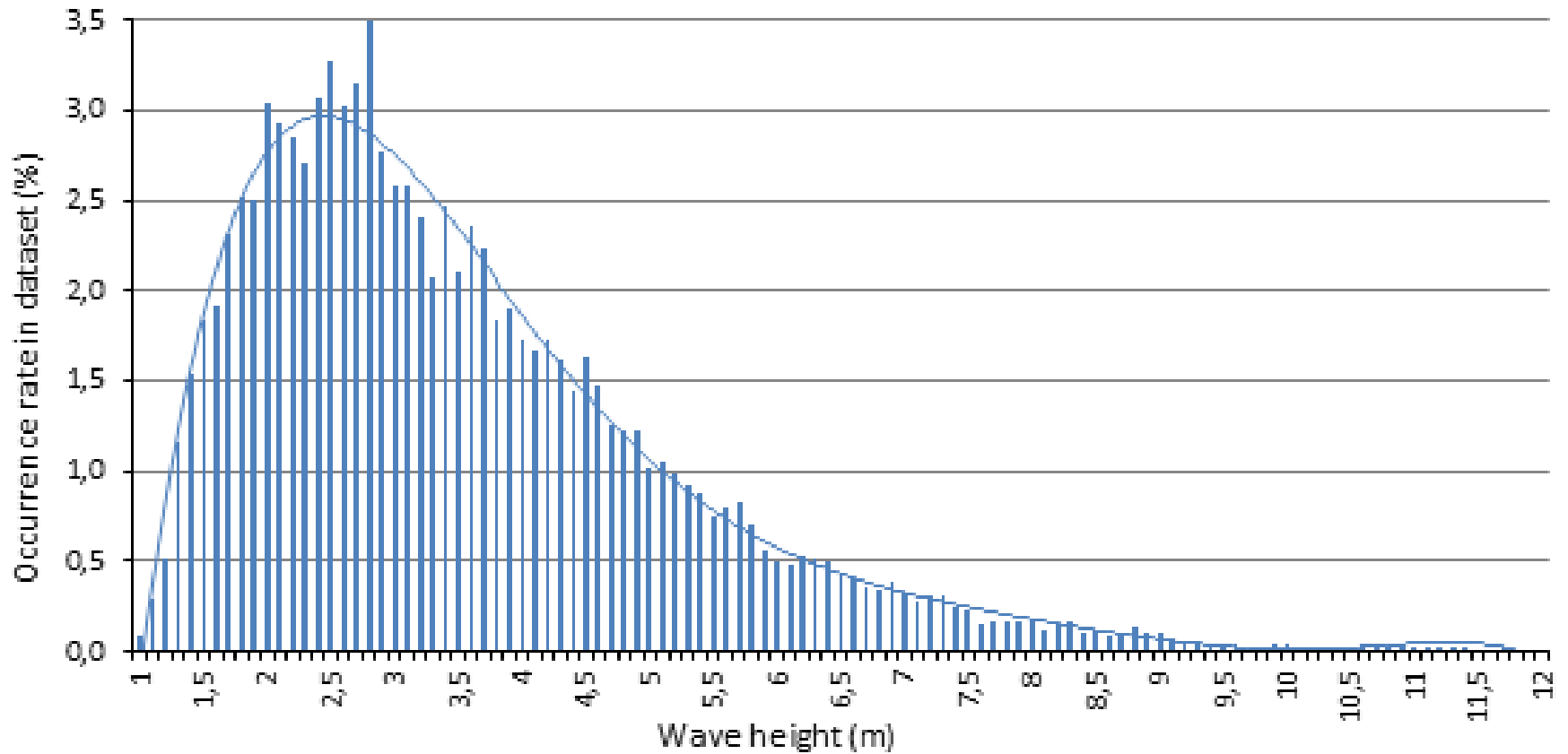
Distributions: Normal distribution



Distributions: Normal distribution



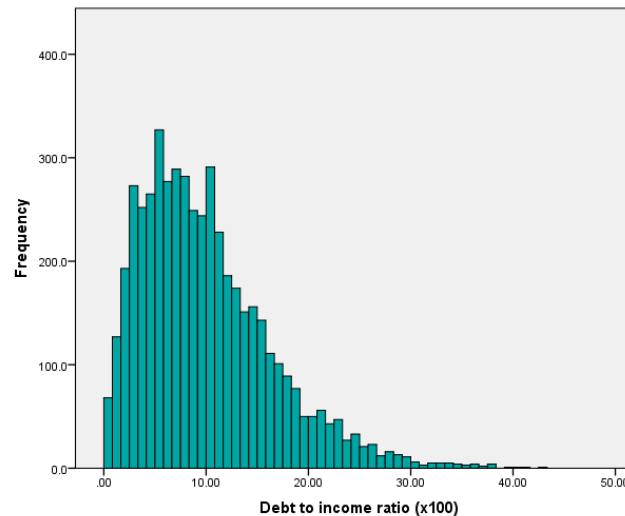
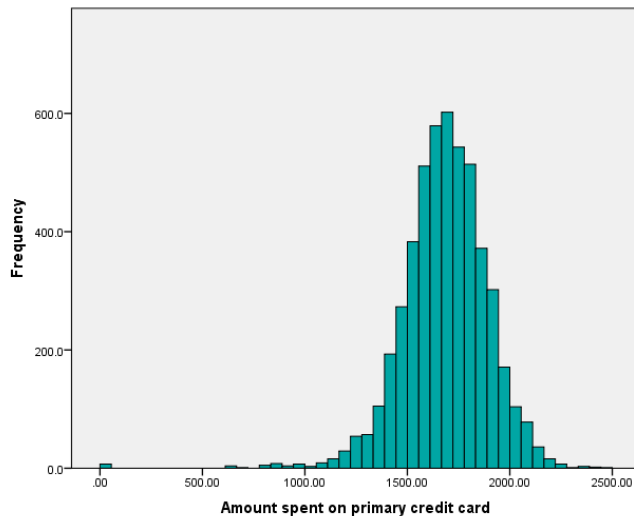
Distributions: Non-Normal distribution



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		339.6359	9.9578
Median		278.6550	8.8000
Std. Deviation		248.3830	6.42317
Minimum		.00	.00
Maximum		3926.41	43.10



Percentiles

- Generalizations of the median (50th 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 93th 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.”

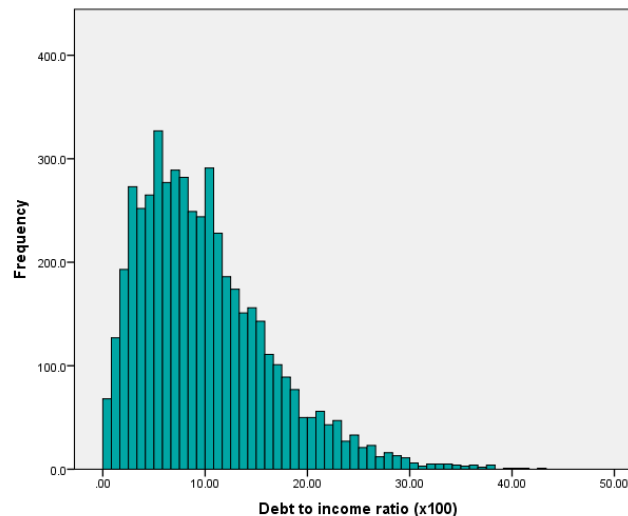
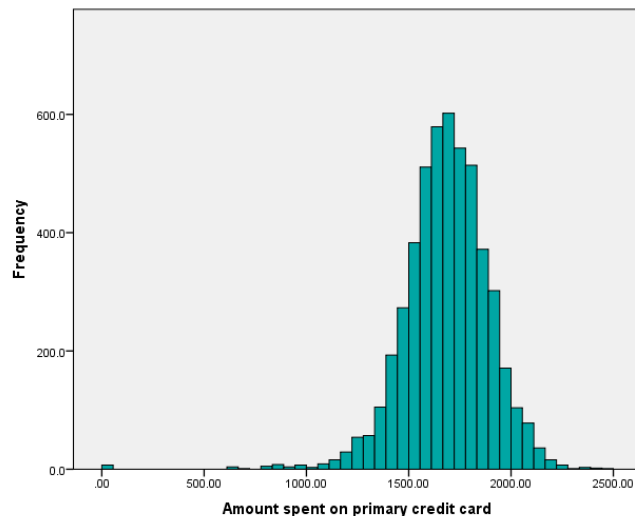
Percentiles - example

- 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
Percentiles	25	184.8600	5.1250
	50	278.6550	8.8000
	75	422.4025	13.5000
Maximum		3926.41	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).

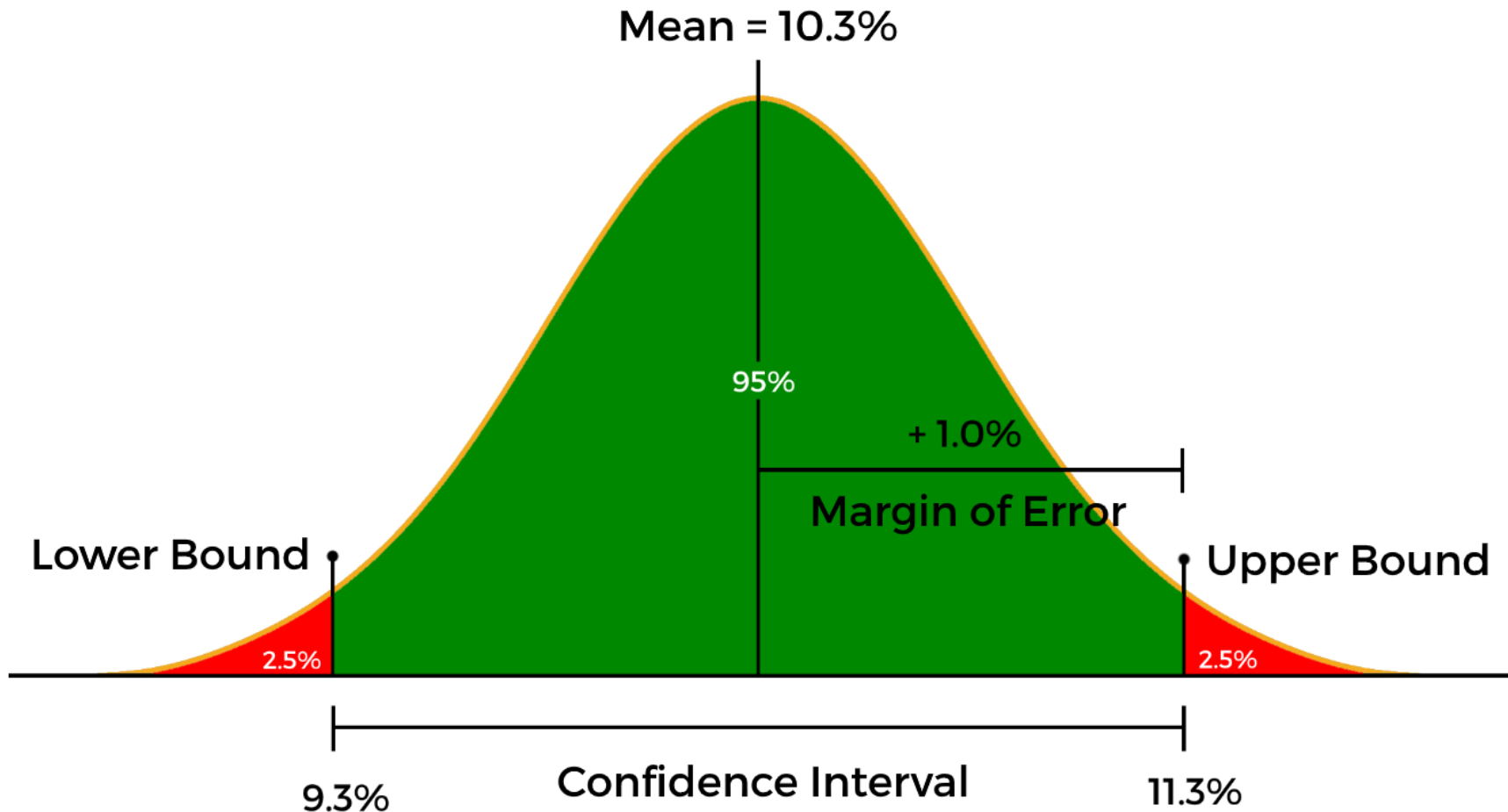
One-Sample Test		
	95% Confidence Interval of the Difference	
	Lower	Upper
Debt to income ratio (x100)	9.7797	10.1359
Amount spent on primary card last month	332.7519	346.5199

“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.

Quantifying uncertainty – confidence intervals



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	-.06599
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	-.03599
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.

Comparing means of two groups

$$t = \frac{\mu - \mu_0}{\sqrt{s^2(\frac{1}{n} + \frac{1}{n_0})}}$$

Hypothesis testing

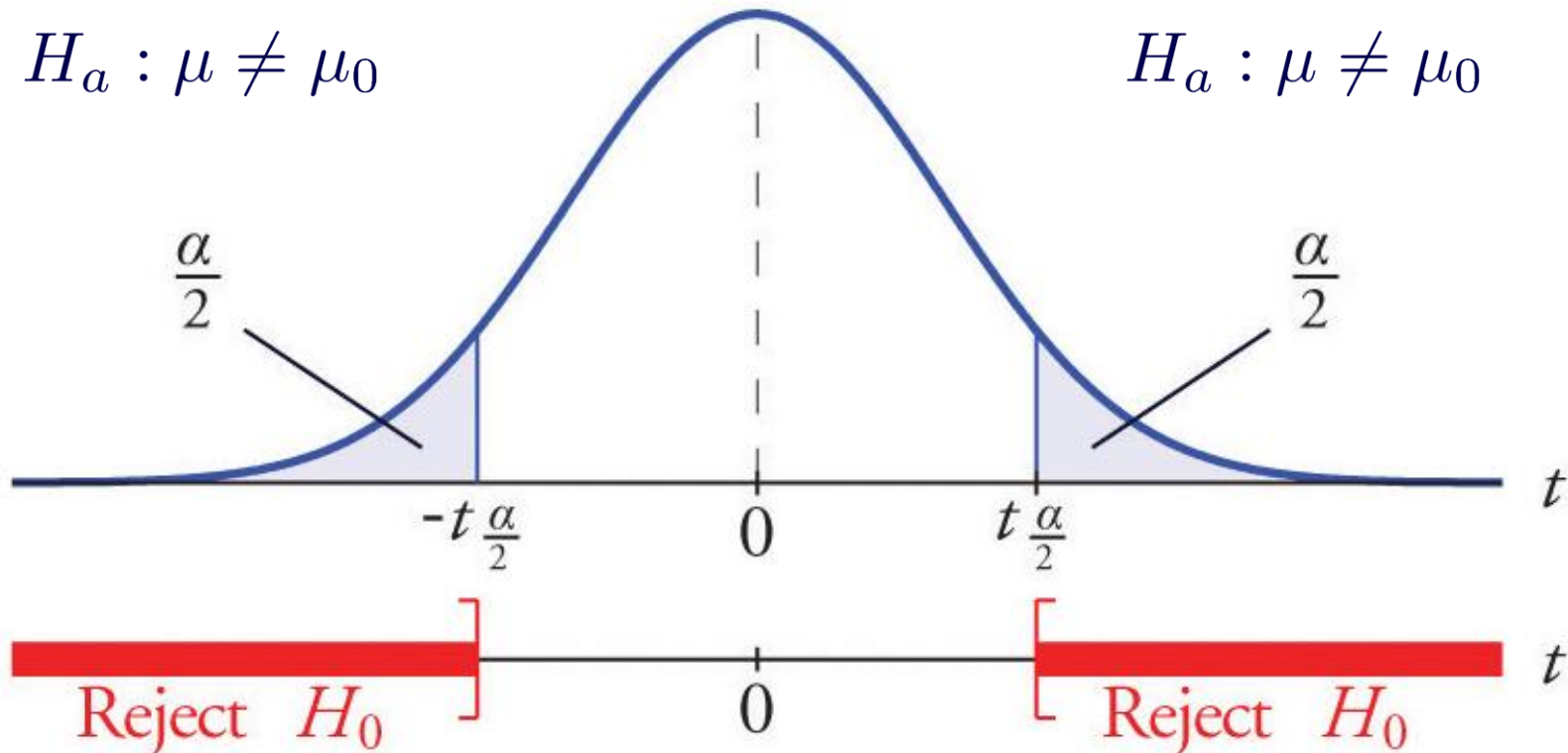
Null hypothesis: means of two groups are not different

Alternative hypothesis: means of two groups are different

$$H_0 : \mu = \mu_0$$

$$H_a : \mu \neq \mu_0$$

$$H_a : \mu \neq \mu_0$$



$$\alpha = 0.05 \Rightarrow t_{0.025} = 1.96$$

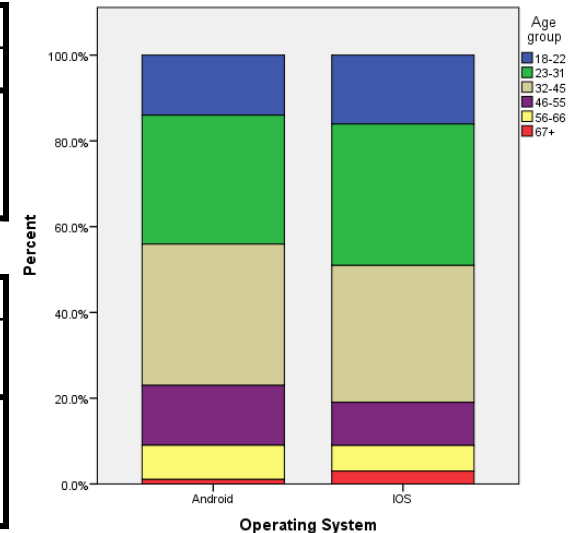
Categorical Data

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						Total
		18-22	23-31	32-45	46-55	56-66	67+	
Operating	Android	93	200	219	93	53	7	665
System	IOS	75	154	149	47	28	14	467
Total		168	354	368	140	81	21	1132

% within Operating System		Age group						Total
		18-22	23-31	32-45	46-55	56-66	67+	
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%



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						Total
		18-22	23-31	32-45	46-55	56-66	67+	
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%

- 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						Total
		18-22	23-31	32-45	46-55	56-66	67+	
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%

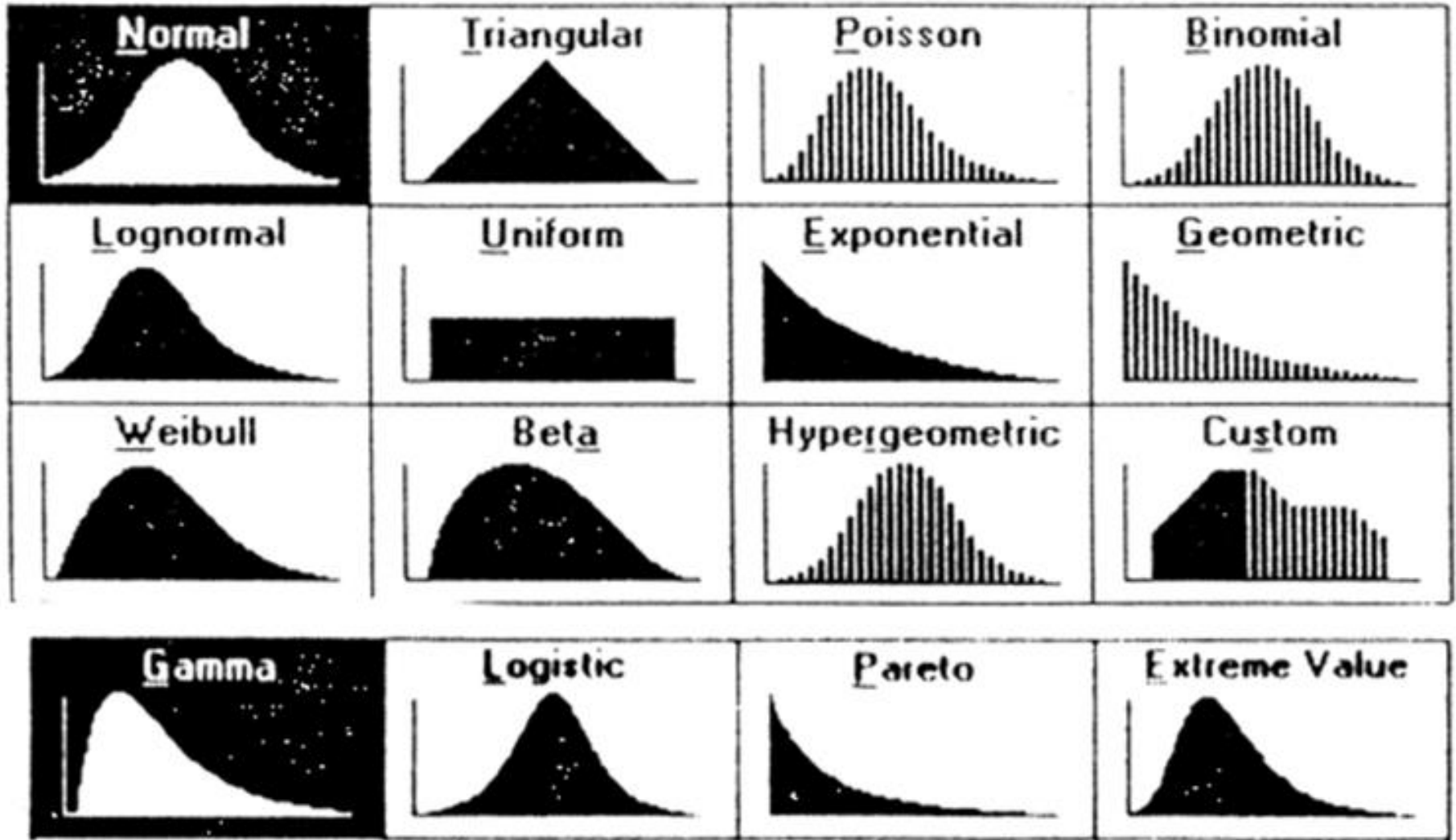
Chi-Square Test

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



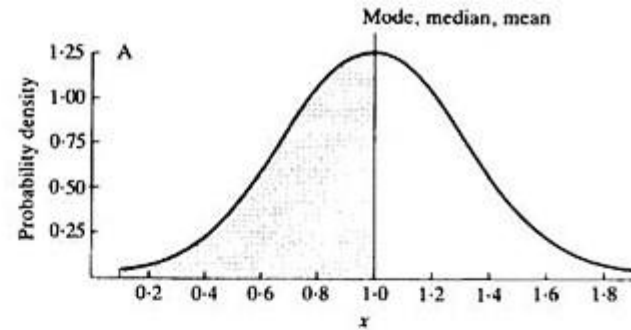
Distributions

Distributions

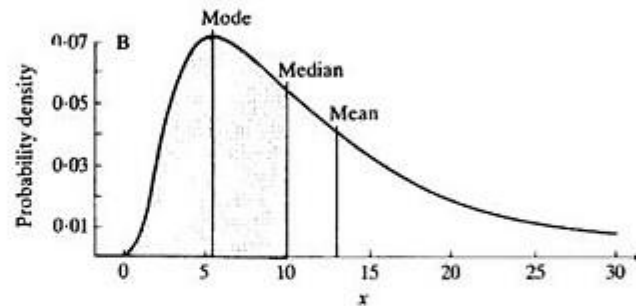


Distributions

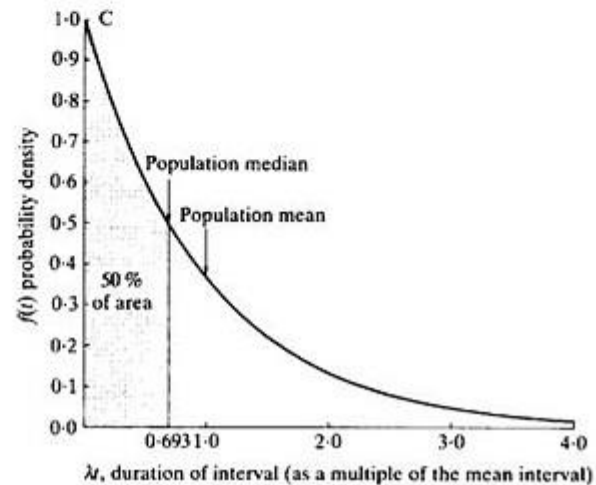
Gaussian p.d.f.



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



Exponential p.d.f.

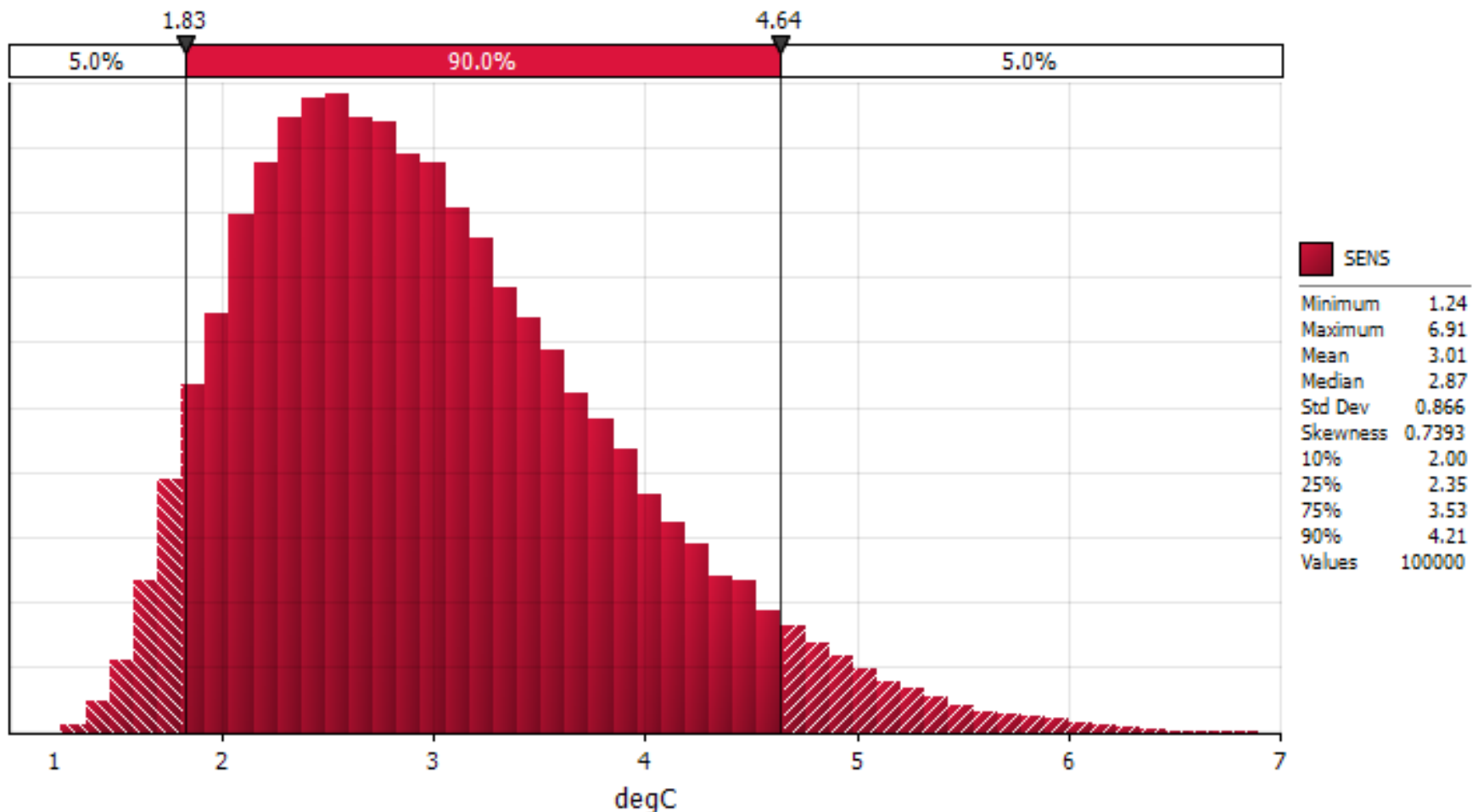


Continuous distributions

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

Distributions

Estimate of the probability distribution of global mean temperature resulting from a doubling of CO₂ 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)

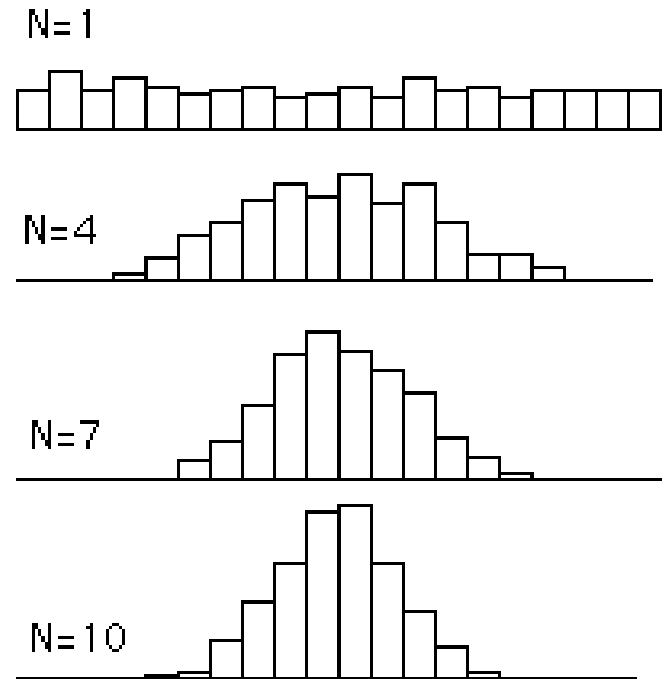
If $\mathbb{E}(x_i) = \mu$ and $\text{var}(x_i) = \sigma^2$ for all i (and independent) then:

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

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

Central Limit Theorem – example

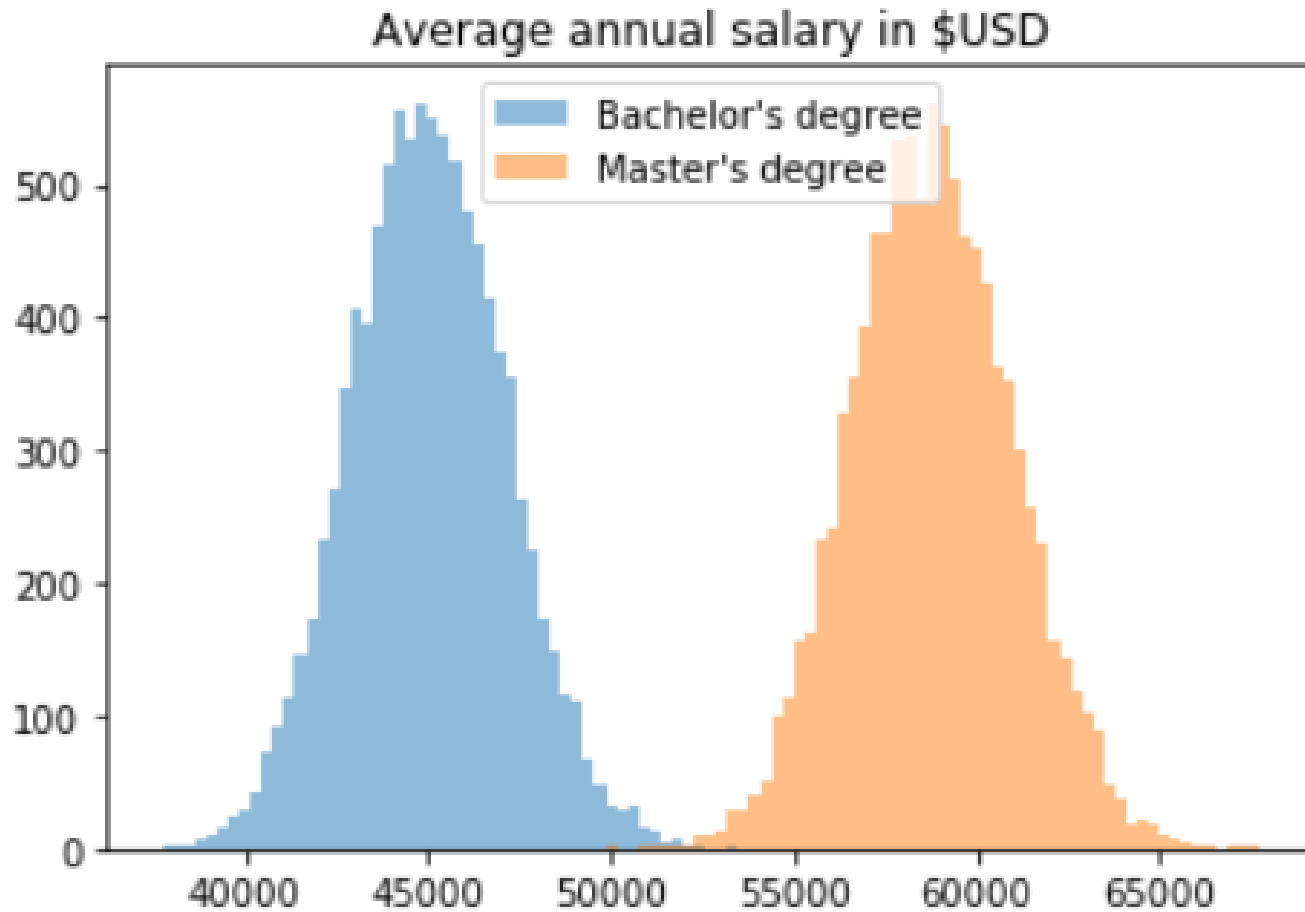
On the right are shown the resulting frequency distributions each based on 500 means. For $n = 4$, 4 scores were sampled from a uniform distribution 500 times and the mean computed each time. The same method was followed with means of 7 scores for $n = 7$ and 10 scores for $n = 10$.



When **n** increases:

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

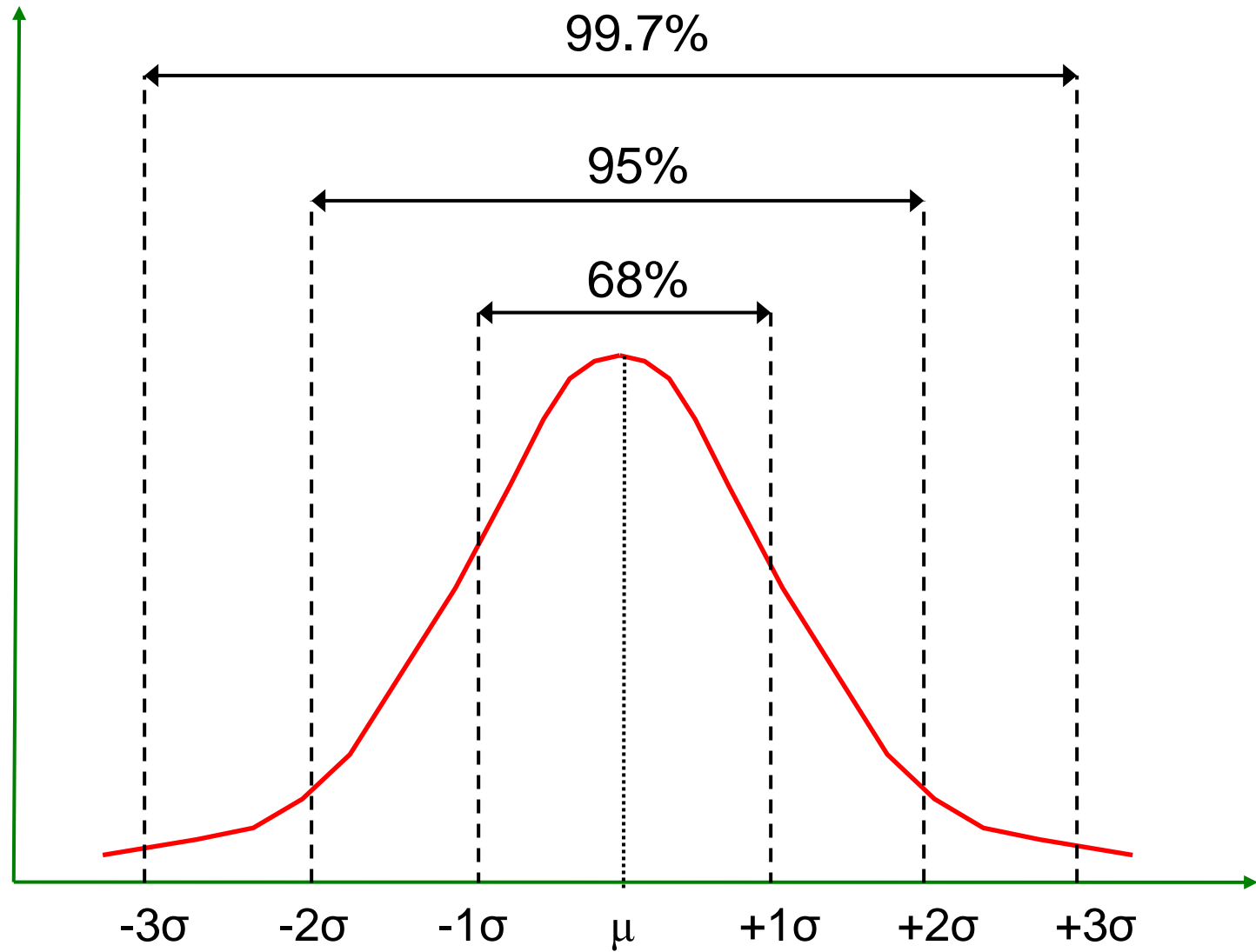
Central Limit Theorem – example (bootstrapping)



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

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



$$P(Z \geq 2.0) = 0.0228$$

$$P(-2 \leq Z \leq +2) = 1 - 2 \cdot 0.0228 = 0.9544$$

$$P(Z \geq 1.96) = 0.025$$

$$P(-1.96 \leq Z \leq +1.96) = 1 - 2 \cdot 0.025 = 0.95$$

Central Limit Theorem

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

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

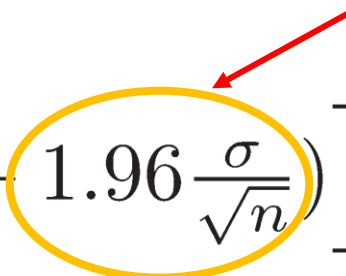
$$\text{prob} [(\mu - 1.96s) \leq \bar{x} \leq (\mu + 1.96s)] = 0.95$$

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

Rearranging

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

margin
of error



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**

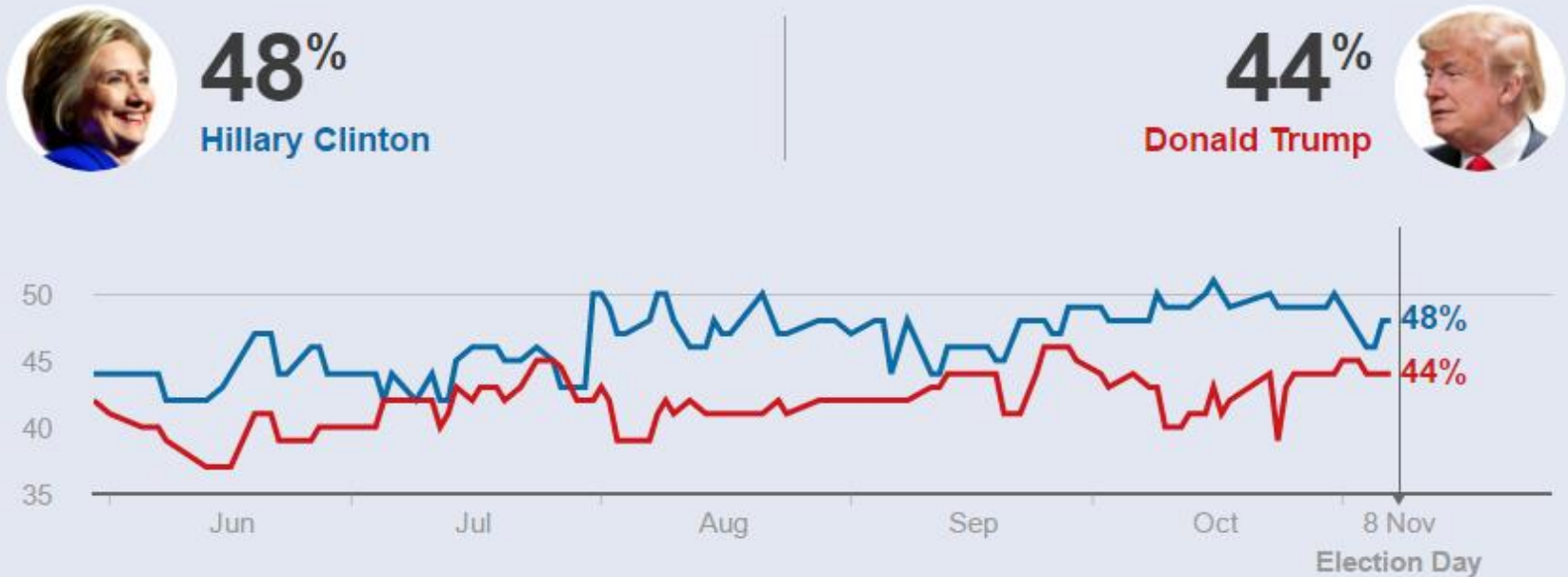
$$s = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p \cdot (1-p)}}{\sqrt{n}} = \frac{\sqrt{0.55 \cdot 0.45}}{\sqrt{1000}} = 0.0157$$

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



Election Statistics Revisited

US presidential elections 2016



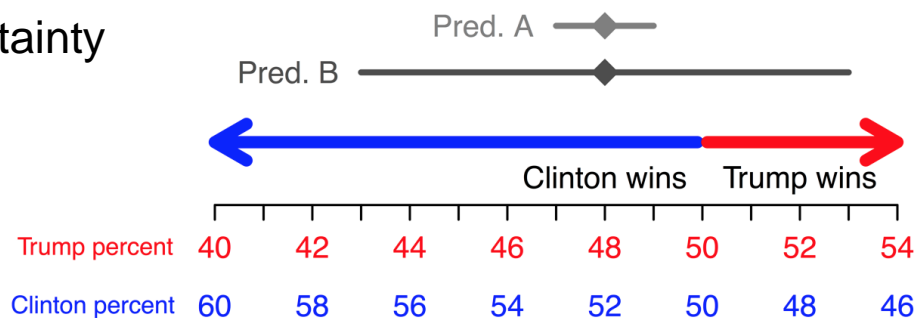
Source: BBC poll of polls

Election polling – quantitative issues relevant to the 2016 elections

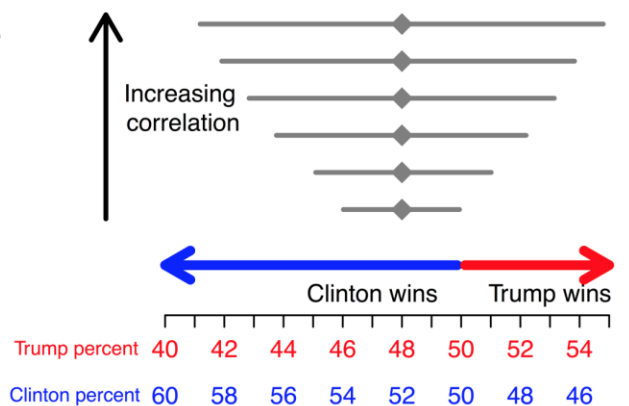
▪ Errors in sampling and polling

- ❑ Low response rates (less than 10%)
- ❑ Inadequate coverage
- ❑ ...

▪ Uncertainty



▪ Correlated errors



Random variables – mean and variance

■ Random variable

- x is a random variable, takes finite number of values, x_i for $i=1,2,\dots,n$
- a probability (associated with each event) represents the relative chance of an occurrence of x_i such that

$$\sum_{i=1}^n p_i = 1 \quad \text{and} \quad p_i \geq 0 \quad \text{for} \quad i = 1, 2, \dots, n \quad \text{relative frequency}$$

■ Expected value (mean value or mean)

- average value obtained by regarding probabilities as frequencies

$$\bar{x} = \mathbb{E}(x) = \sum_{i=1}^n p_i x_i = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{finite number of possibilities}$$

■ Variance - measure of possible deviation from the mean

$$\begin{aligned} \text{var}(x) &= \mathbb{E}[(x - \bar{x})^2] \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \end{aligned}$$

Expected value of squared variable
how much x tends to vary from its mean

Random variables – covariance and correlation

- **Covariance of two random variables x and y**

$$\begin{aligned}\text{cov}(x, y) &= \sigma_{x y} = \mathbb{E}[(x - \bar{x})(y - \bar{y})] \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}\end{aligned}$$

$$\sigma_{x y} = \sigma_{y x}$$

- **Correlation between two random variables x and y**

$$\text{cor}(x, y) = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}} \Rightarrow \rho_{x y} = \frac{\sigma_{x y}}{\sigma_x \sigma_y}$$

$$\rho_{x y} > 0 \Rightarrow \text{positively correlated}$$

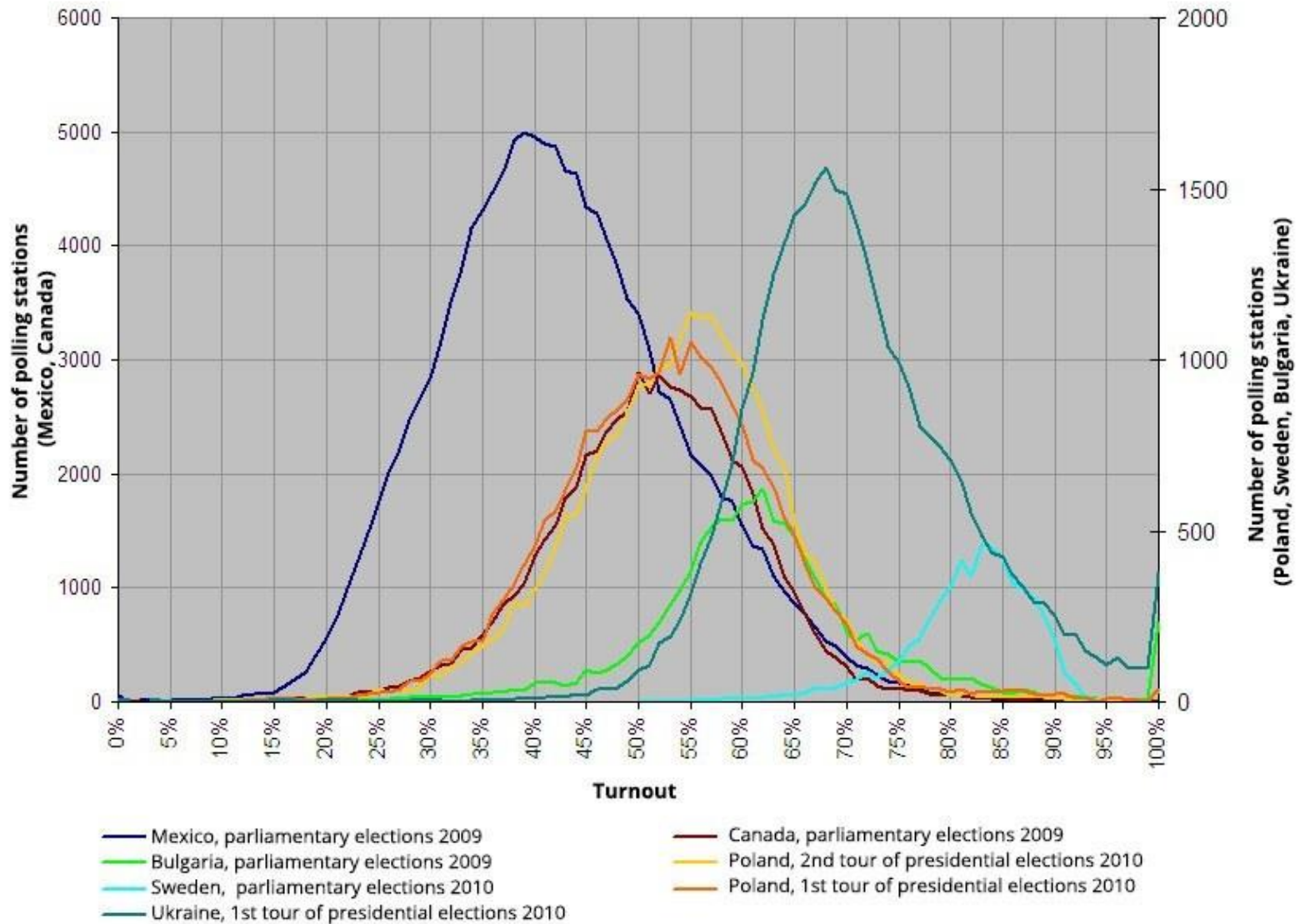
$$\rho_{x y} = 0 \Rightarrow \text{uncorrelated (maybe independent)}$$

$$\rho_{x y} < 0 \Rightarrow \text{negatively correlated}$$

$$\rho_{x y} = \pm 1 \Rightarrow \text{perfect } +(-) \text{ correlation}$$

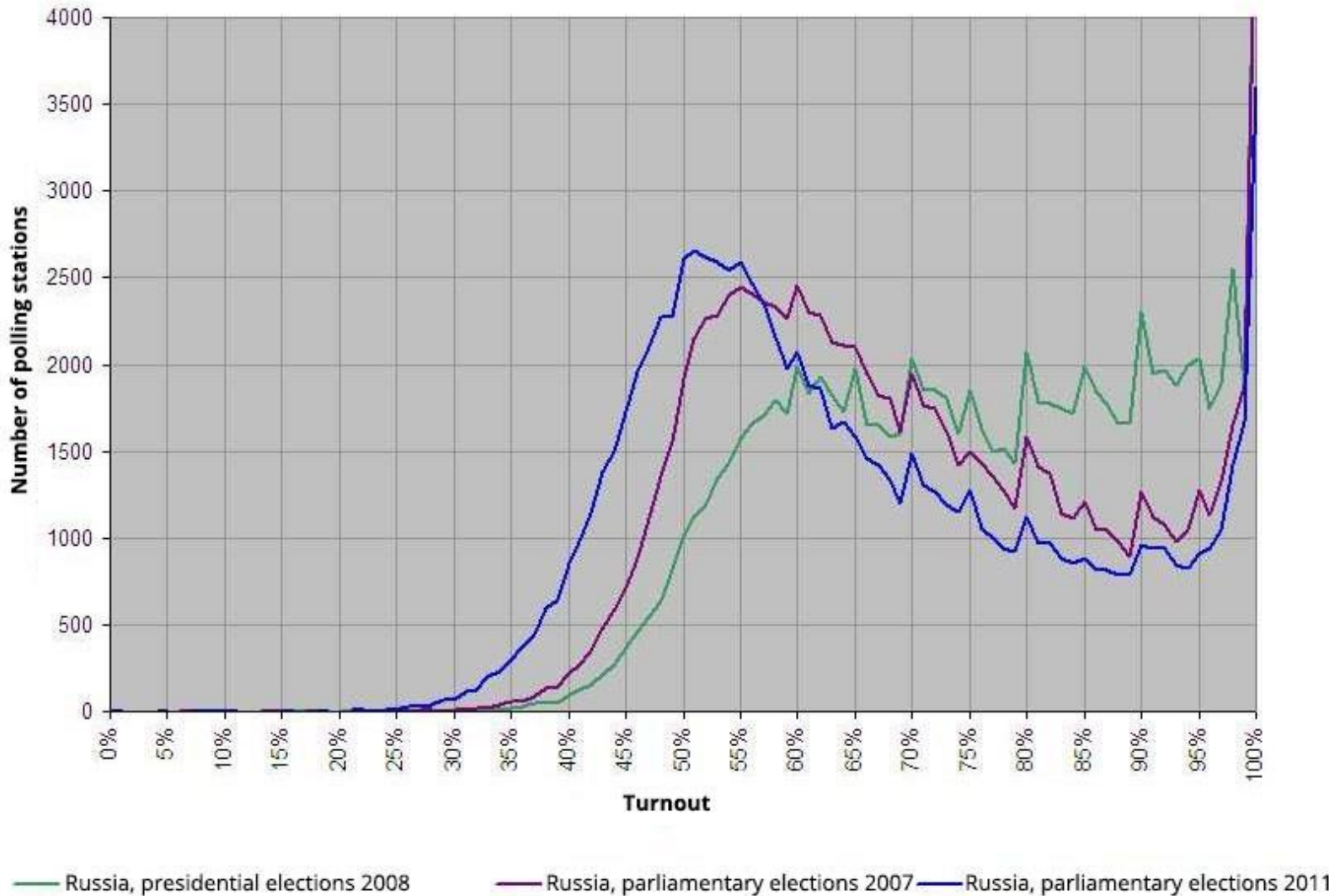
sign defines
direction of
the relationship

Voter turnout at elections – Canada, USA, Ukraine, Poland, Bulgaria



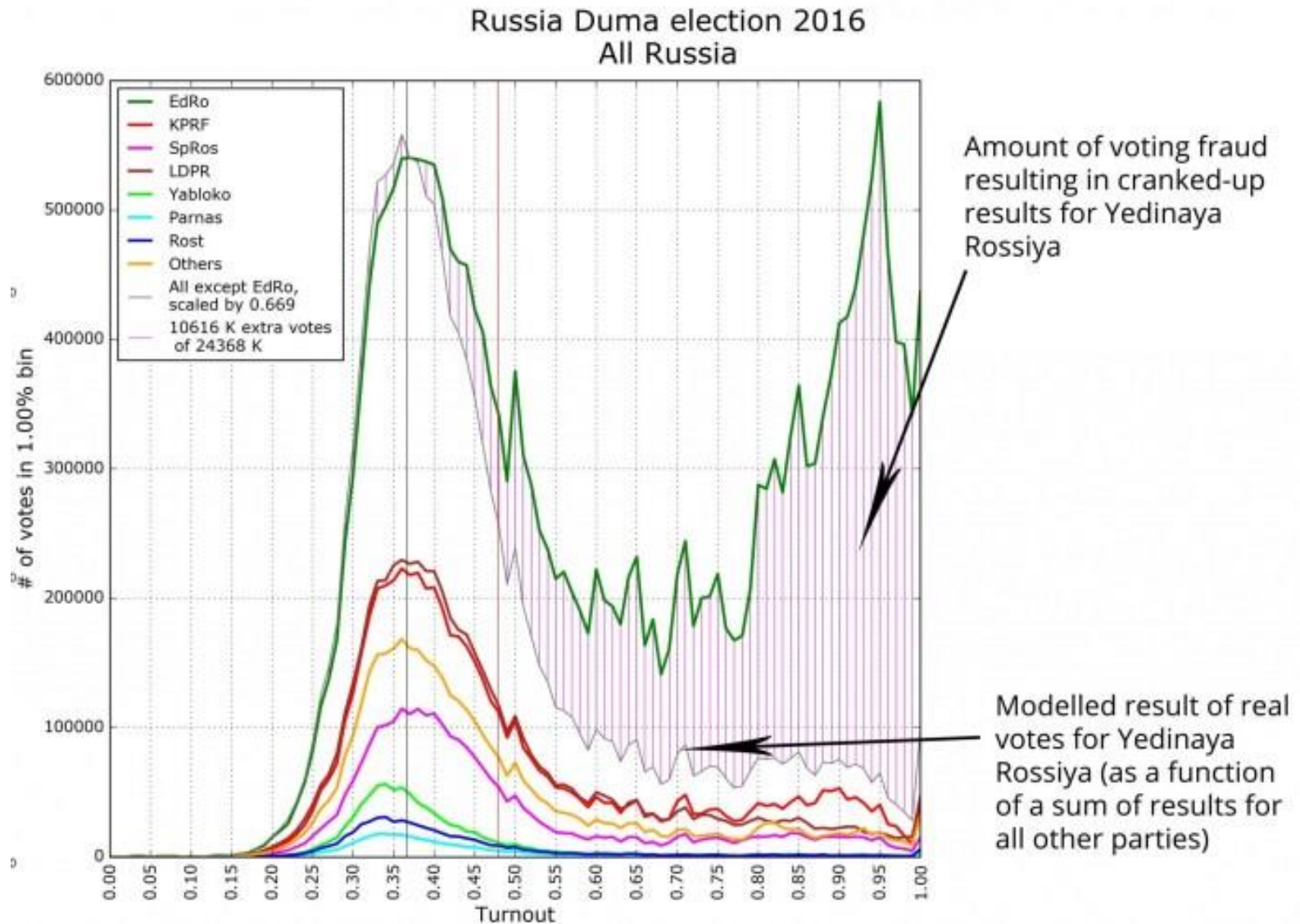
Source: <http://euromaidanpress.com/2016/09/21/statistical-method-measures-voting-fraud-of-russias-pro-putin-party/>

Voter turnout at elections – Russia election fraud

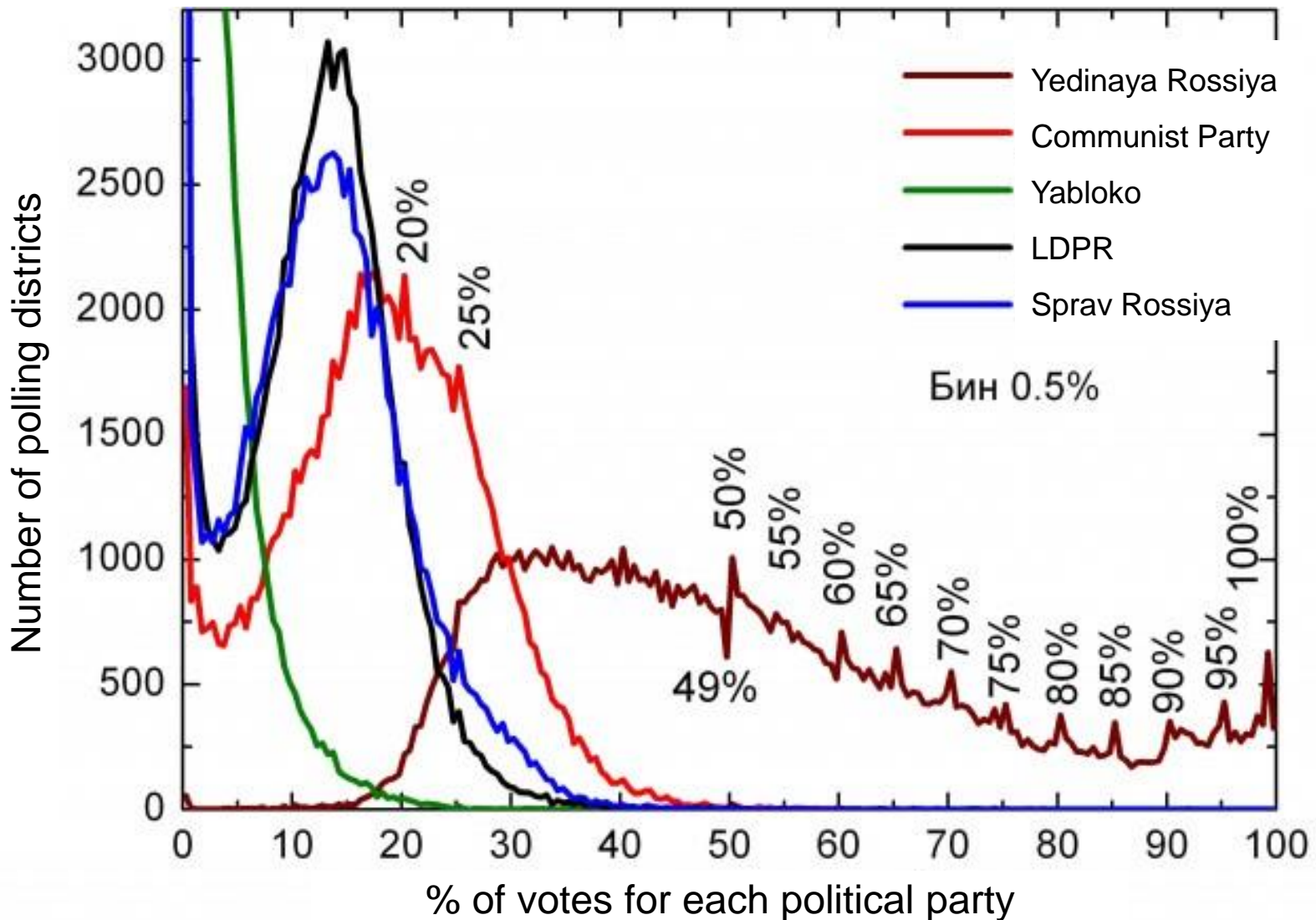


Source: <http://euromaidanpress.com/2016/09/21/statistical-method-measures-voting-fraud-of-russias-pro-putin-party/>

Voter turnout at elections – Russia election fraud



Number of votes for political parties – Russia election fraud





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
- Apply statistical analysis to solve business problems

