

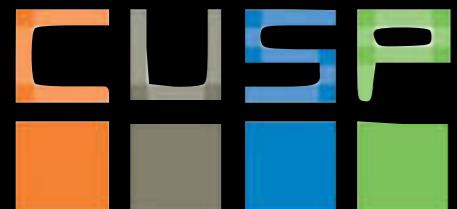
# Urban Informatics

Fall 2018

dr. federica bianco [fbianco@nyu.edu](mailto:fbianco@nyu.edu)

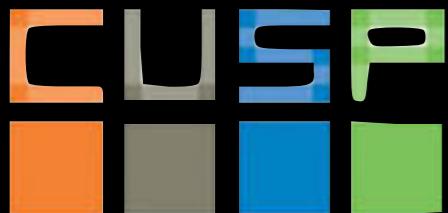


@fedhere



## Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing:  $p$ -value, statistical significance

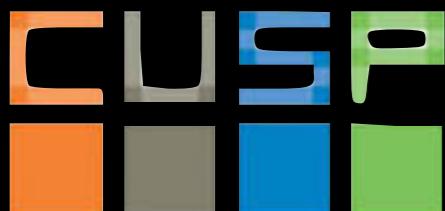


## Recap:

- Good practices with data: falsifiability, reproducibility
- Basic data retrieving and munging: APIs, Data formats
- Basic statistics: distributions and their moments
- Hypothesis testing:  $p$ -value, statistical significance

## This class:

- How to choose the right statistical test
- Z, t, F test and tests for correlation
- Correlation vs Causation

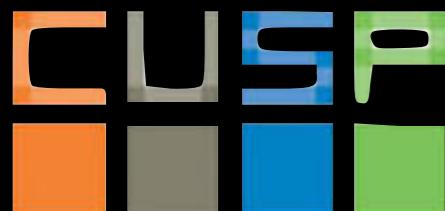


# hypothesis testing

did we detect a phenomenon  
(e.g. as a result of an implemented policy)?

*null hypothesis:* no relationship between (two) measured phenomena  
*or* no difference between (two) groups  
if you have a test control sample: test sample and  
control sample are the same - no effect

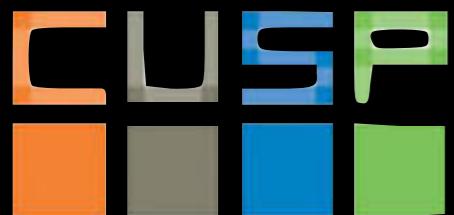
*falsify the null hypothesis:* do you see an effect?  
do you see a difference b/w samples?



A simple (too simple?) answer

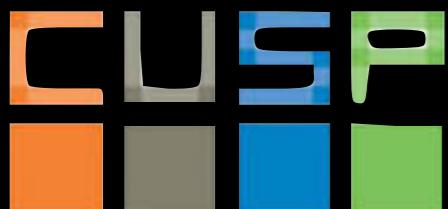
did we detect a phenomenon  
(e.g. as a result of an implemented policy)?

*p-value* a measure of the probability that the result you observed could have been observed by chance under the *Null hypothesis*



# Steps in Null-Rejection Hypothesis Testing

1. Formulate Null (and alternative) Hypothesis
2. Choose a significance level  $\alpha$
3. Measure a *statistic* for a *sample* to be compared to the *parameter of a population*  
OR  
Measure a *statistic* for *two or more samples* to be compared to *each other*
4. Assess if your statistics is significant or not. In practice: compare the statistics (Z, t, F, chisq) with a distribution table



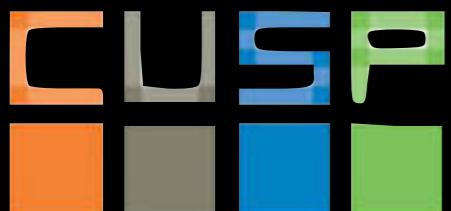
# Steps in Null-Rejection Hypothesis Testing

1. Formulate Null (and alternative) Hypothesis
2. Choose a significance level  $\alpha$
3. Measure a *statistic* for a *sample* to be compared to the *parameter of a population*

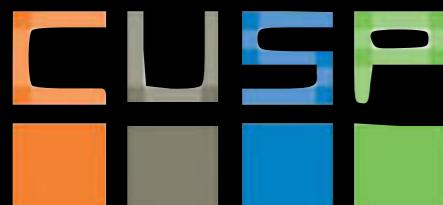
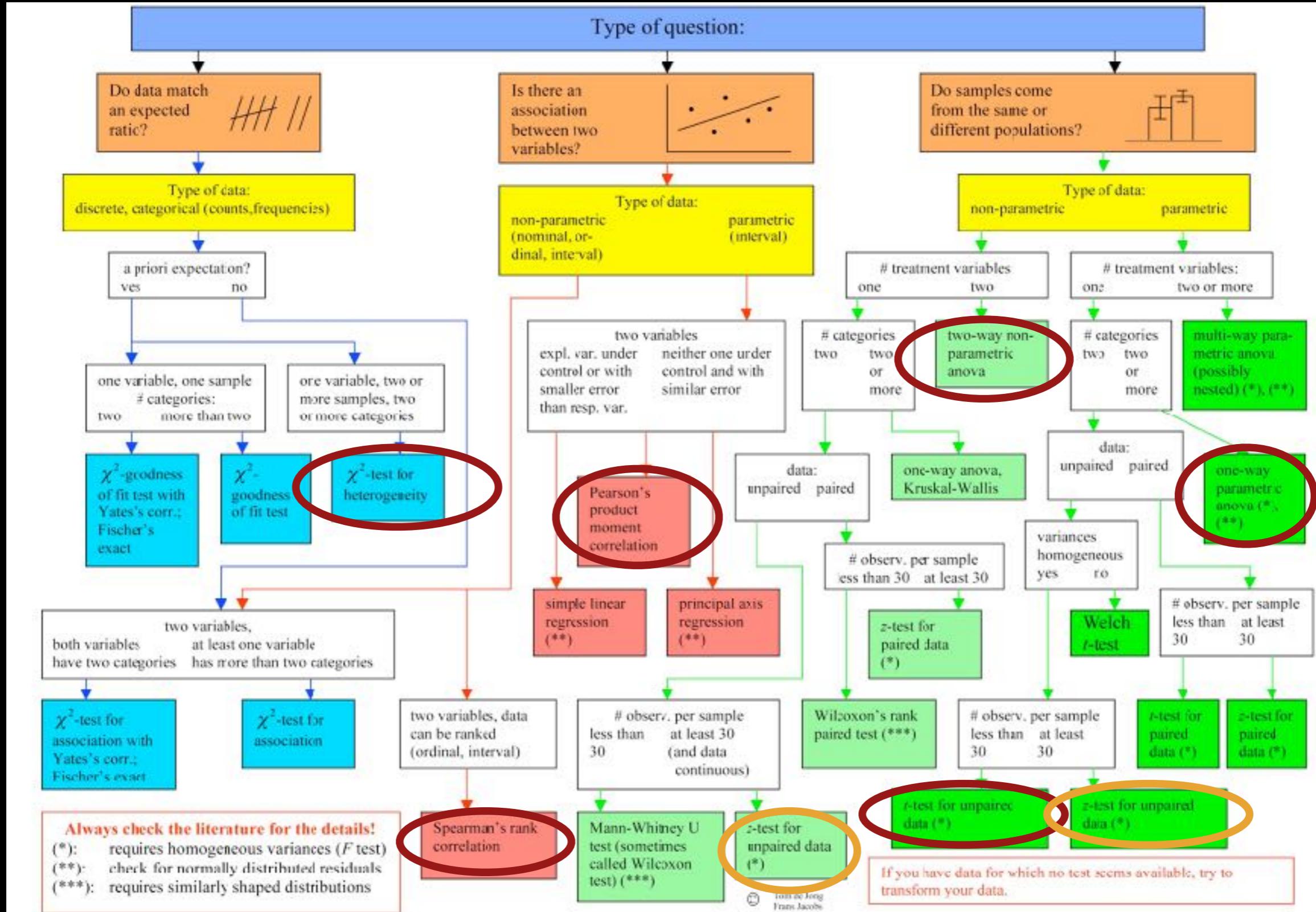
OR

Measure a *statistic* for *two or more samples* to be compared to *each other*

- 4. Assess if your statistics is significant or not. In practice: compare the statistics (Z, t, F, chisq) with a distribution table



# Steps in Null-Rejection Hypothesis Testing



which statistics?  
how to compare it?

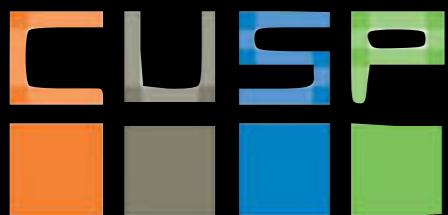
IV: Statistical analysis

<https://stats.idre.ucla.edu/other/mult-pkg/whatstat/>

# how to choose the right statistical test

[http://www.csun.edu/~amarenco/Fcs%20682/  
When%20to%20use%20what%20test.pdf](http://www.csun.edu/~amarenco/Fcs%20682/When%20to%20use%20what%20test.pdf)

Statistical Analyses	Independent Variables		Dependent Variables		Control Variables	Question Answered by the Statistic
	# of IVs	Data Type	# of DVs	Type of Data		
<b>Chi square</b>	1	categorical	1	categorical	0	Do differences exist between groups?
<b>t-Test</b>	1	dichotomous	1	continuous	0	Do differences exist between 2 groups on one DV?
<b>ANOVA</b>	1 +	categorical	1	continuous	0	Do differences exist between 2 or more groups on one DV?
<b>ANCOVA</b>	1 +	categorical	1	continuous	1 +	Do differences exist between 2 or more groups after controlling for CVs on one DV?
<b>MANOVA</b>	1 +	categorical	2 +	continuous	0	Do differences exist between 2 or more groups on multiple DVs?
<b>MANCOVA</b>	1 +	categorical	2 +	continuous	1 +	Do differences exist between 2 or more groups after controlling for CVs on multiple DVs?
<b>Correlation</b>	1	dichotomous or continuous	1	continuous	0	How strongly and in what direction (i.e., +, -) are the IV and DV related?
<b>Multiple regression</b>	2 +	dichotomous or continuous	1	continuous	0	How much variance in the DV is accounted for by linear combination of the IVs? Also, how strongly related to the DV is the beta coefficient for each IV?
<b>Path analysis</b>	2 +	continuous	1 +	continuous	0	What are the direct and indirect effects of predictor variables on the DV?
<b>Logistic Regression</b>	1 +	categorical or continuous	1	dichotomous	0	What is the odds probability of the DV occurring as the values of the IVs change?



# how to choose the right statistical test

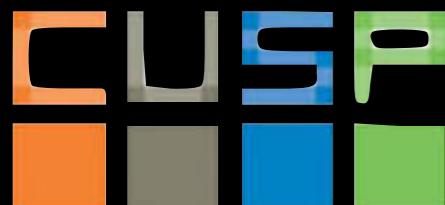
## Assignment 1:

Browse PlosOne's abstracts to find 1 paper for each of the 3 tests:

[http://www.csun.edu/~amarenco/Fcs%20682/  
When%20to%20use%20what%20test.pdf](http://www.csun.edu/~amarenco/Fcs%20682/When%20to%20use%20what%20test.pdf)

Statistical Analyses	Independent Variables		Dependent Variables		Control Variables	Question Answered by the Statistic
	# of IVs	Data Type	# of DVs	Type of Data		
<b>Chi square</b>	1	categorical	1	categorical	0	Do differences exist between groups?
<b>t-Test</b>	1	dichotomous	1	continuous	0	Do differences exist between 2 groups on one DV?
<b>ANOVA</b>	1 +	categorical	1	continuous	0	Do differences exist between 2 or more groups on one DV?
<b>ANCOVA</b>	1 +	categorical	1	continuous	1 +	Do differences exist between 2 or more groups after controlling for CVs on one DV?
<b>MANOVA</b>	1 +	categorical	2 +	continuous	0	Do differences exist between 2 or more groups on multiple DVs?
<b>MANCOVA</b>	1 +	categorical	2 +	continuous	1 +	Do differences exist between 2 or more groups after controlling for CVs on multiple DVs?
<b>Correlation</b>	1	dichotomous or continuous	1	continuous	0	How strongly and in what direction (i.e., +, -) are the IV and DV related?
<b>Multiple regression</b>	2 +	dichotomous or continuous	1	continuous	0	How much variance in the DV is accounted for by linear combination of the IVs? Also, how strongly related to the DV is the beta coefficient for each IV?
<b>Path analysis</b>	2 +	continuous	1 +	continuous	0	What are the direct and indirect effects of predictor variables on the DV?
<b>Logistic Regression</b>	1 +	categorical or continuous	1	dichotomous	0	What is the odds probability of the DV occurring as the values of the IVs change?

choose 1  
choose 1  
choose 1 ←

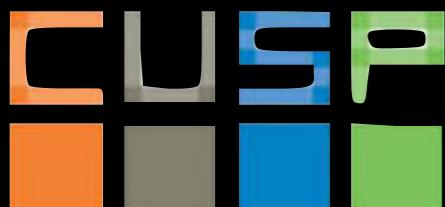


# how to choose the right statistical test

## Assignment 1:

### Example: ANCOVA

Statistical Analyses	Dependent Variables	Independent Variables	Control Variables	Question Answered
ANCOVA	Ratings about their values <b>type: ordinal</b>	did Self Affirmation or not <b>type: categorical</b>	age <b>type: continuous</b>	self-affirmation group rates the value significantly more than control group



<https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0062593.t001>

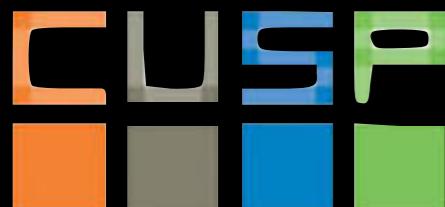
IV: Statistical analysis

## how to choose the right statistical test

Assignment 1: Prepare a markdown table describing the use of the test in each of the papers you chose (3 rows in your table)

Example: ANCOVA

Statistical Analyses	Dependent Variables	Independent Variables	Control Variables	Question Answered
ANCOVA	Ratings about their values <i>type: ordinal</i>	did Self Affirmation or not <i>type: categorical</i>	age <i>type: continuous</i>	self-affirmation group rates the value significantly more than control group



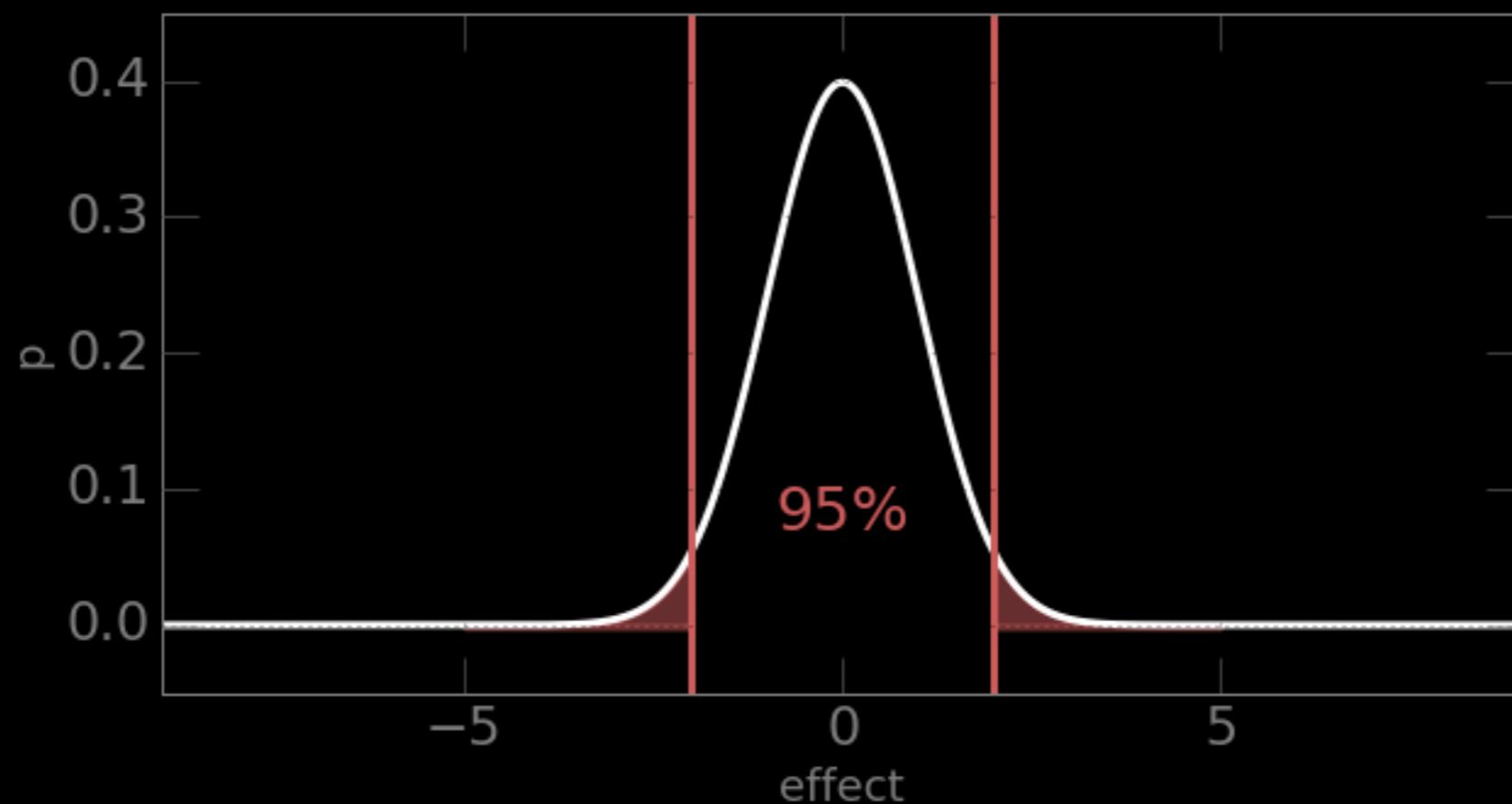
<https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0062593.t001>

IV: Statistical analysis

*if you knew how your statistics should be distributed...*

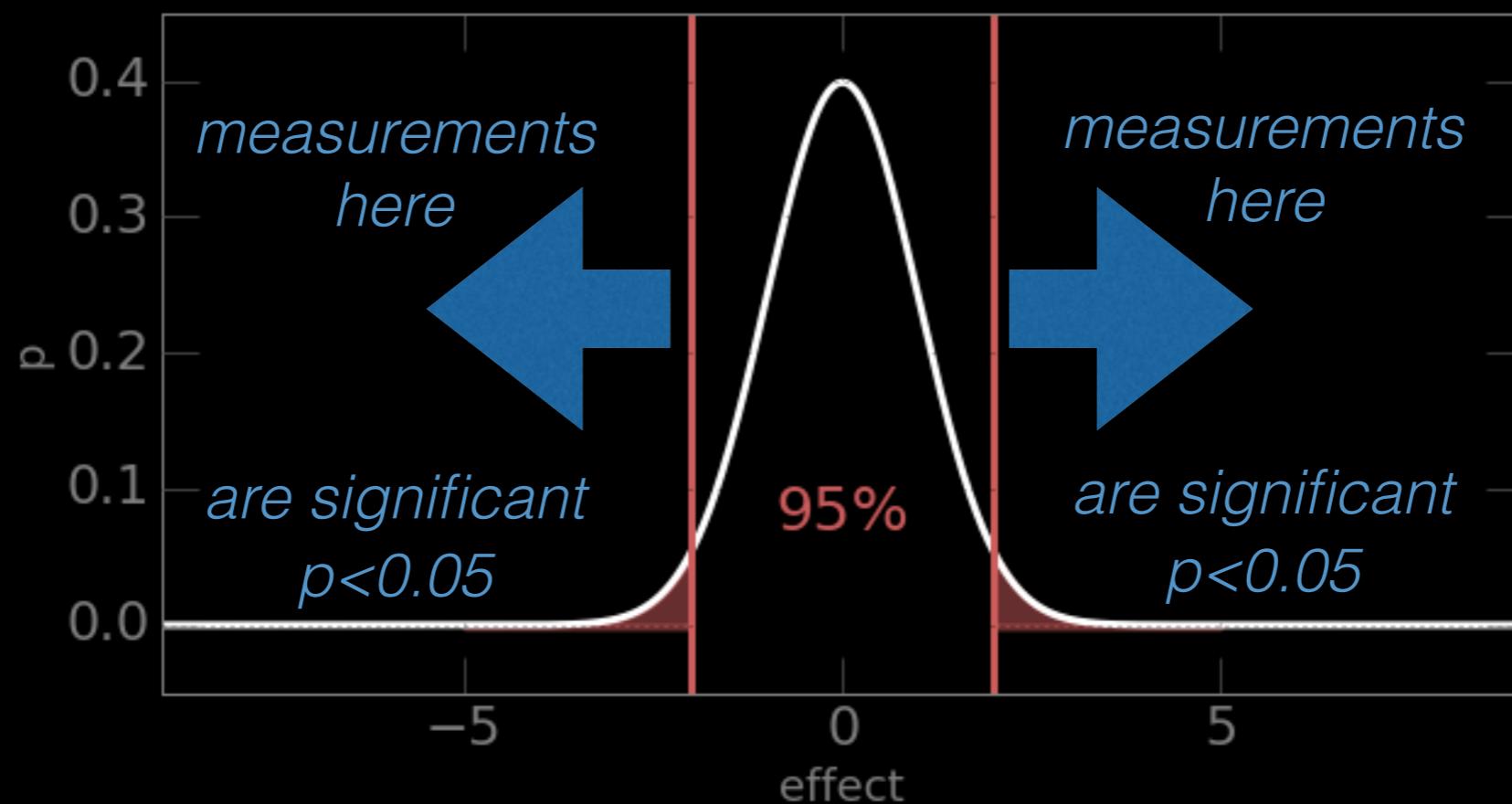
$$\alpha = 0.05$$

$$1 - 0.05 = 0.95 \Rightarrow 95\%$$



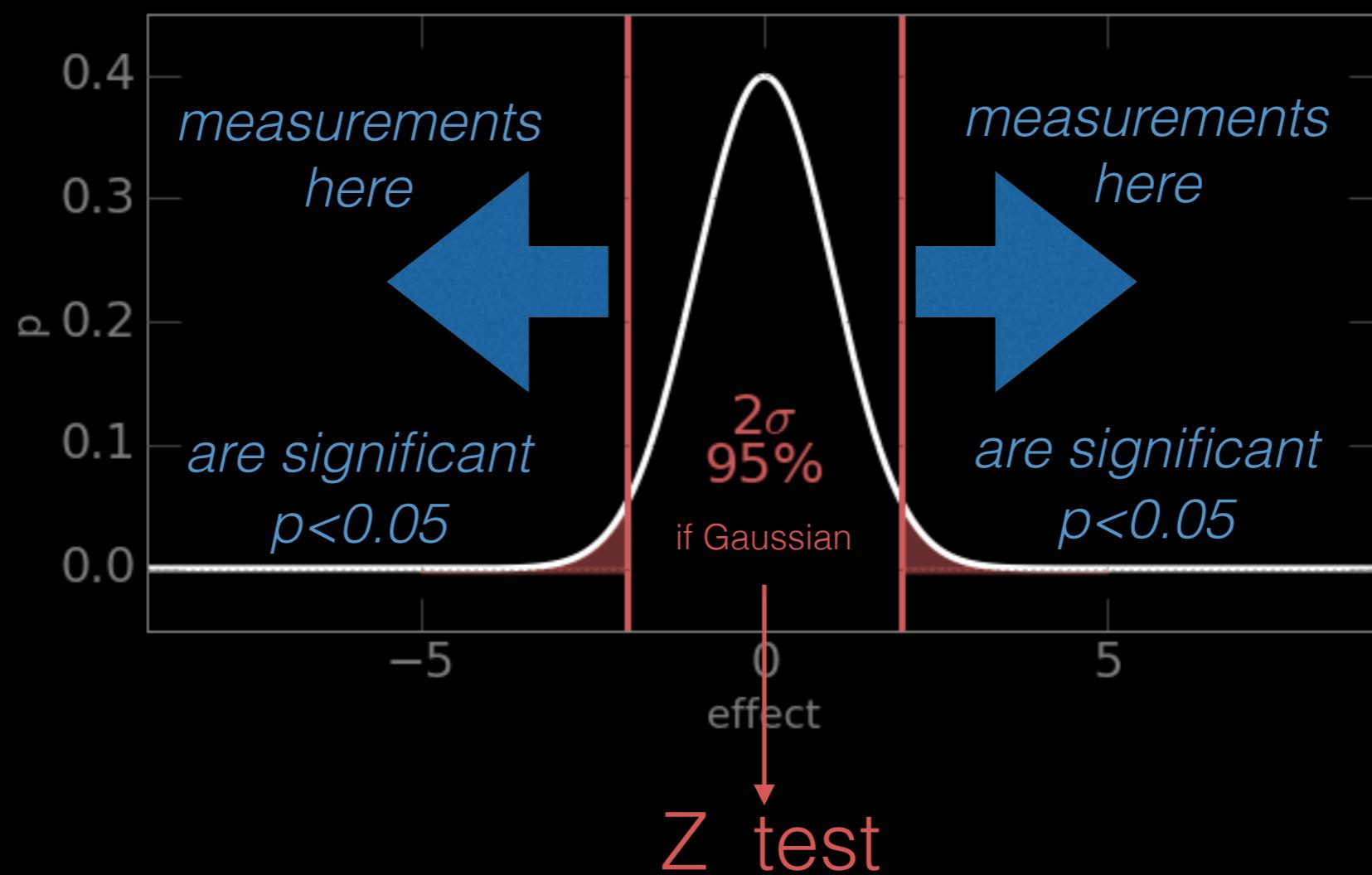
$$\alpha = 0.05$$

$$1 - 0.05 = 0.95 \Rightarrow 95\%$$

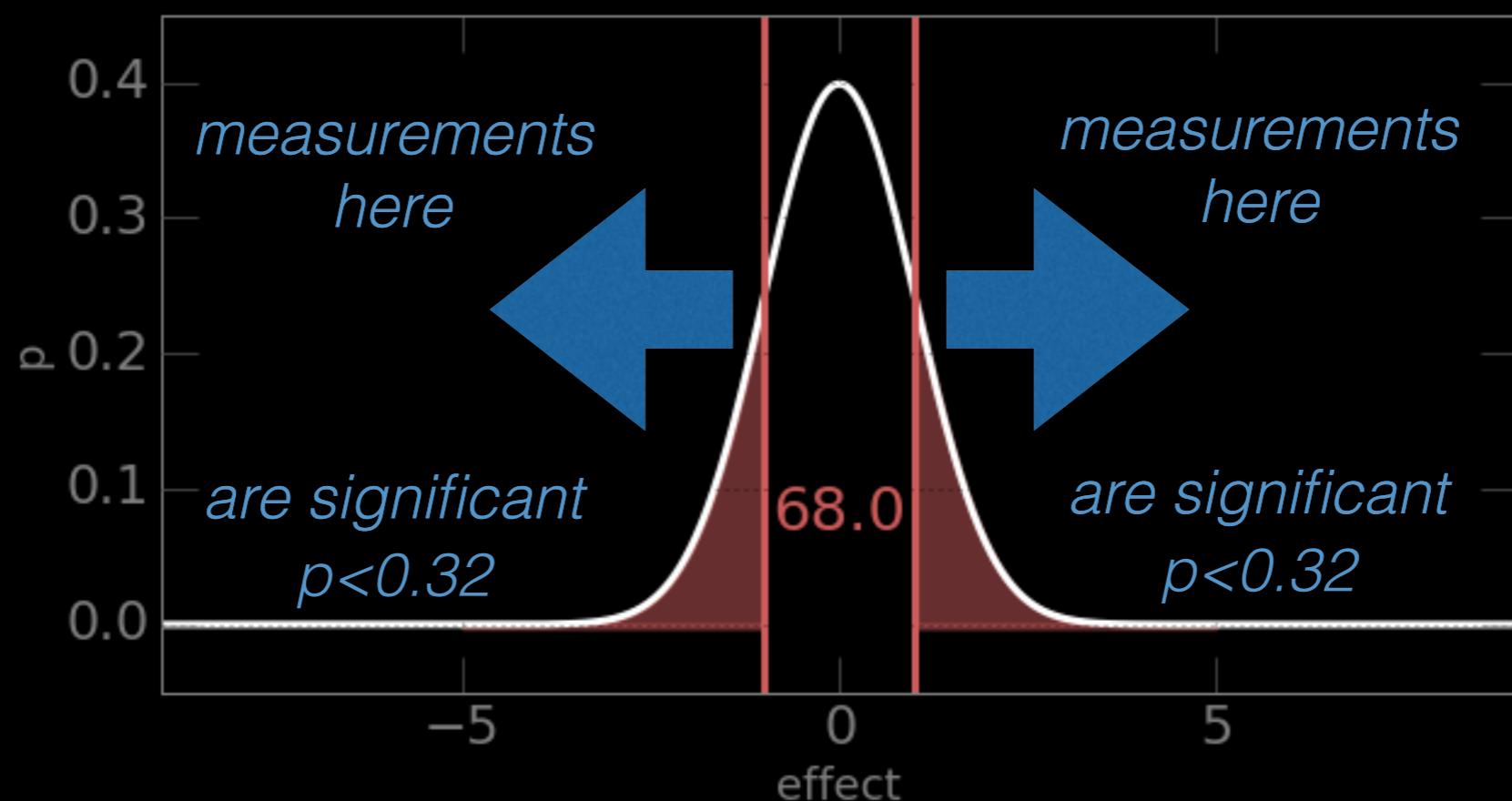


$$\alpha = 0.05$$

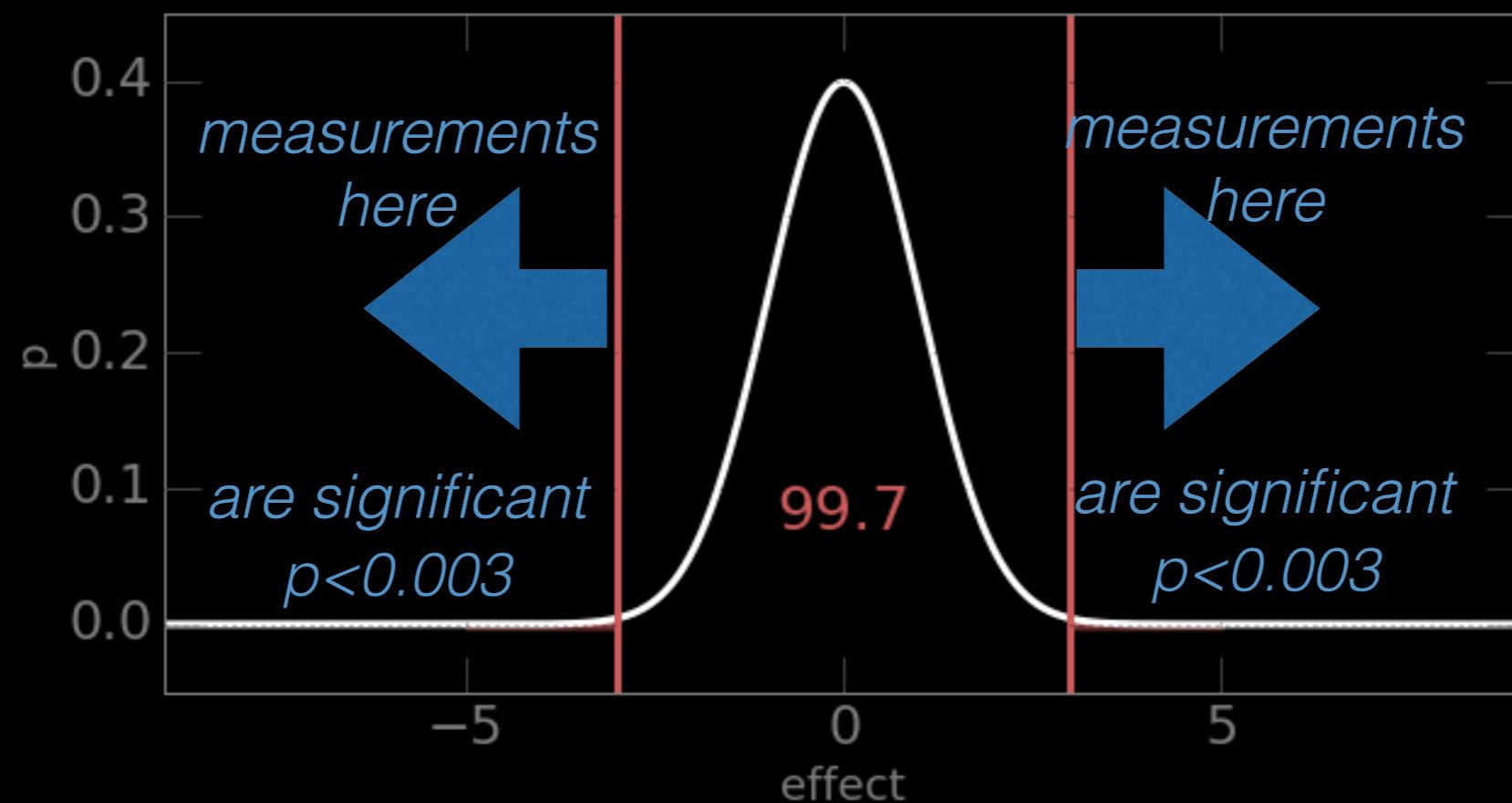
$$1 - \alpha = 0.95 \Rightarrow 95\%$$



$$a = 0.32$$



$$\alpha = 0.003$$



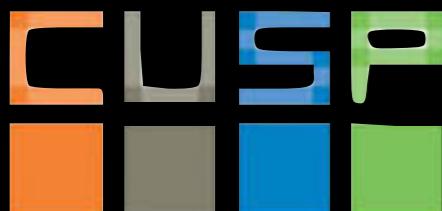
What is the *probability distribution* of a *statistics*?

To measure the probability of the value of a statistics that was measured after an experiment we need to know *how the statistics is distributed* under the null hypothesis.

Quantities that follow a specific distribution under the null hypothesis are called *pivotal*.

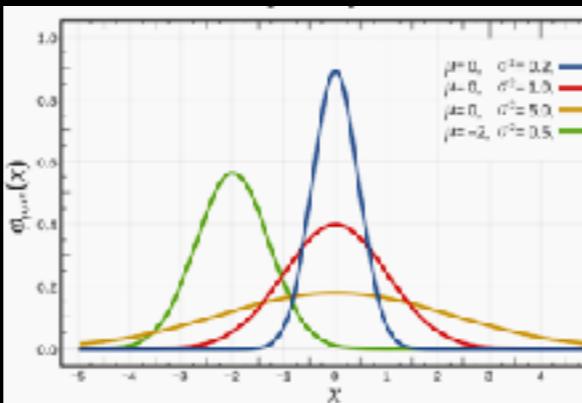
Each *statistics* follows some distribution.

Which one though?



# Z statistics Gaussian

$$Z = \frac{\mu - \bar{x}}{\sigma / \sqrt{n}}$$

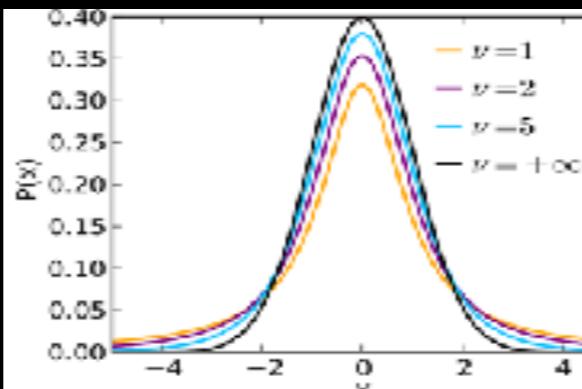


Notation	$N(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ — mean (location) $\sigma^2 > 0$ — variance (squared scale)
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2} \left[ 1 + \operatorname{erf}\left( \frac{x-\mu}{\sigma\sqrt{2}} \right) \right]$
Quantile	$\mu + \sigma\sqrt{2} \operatorname{erf}^{-1}(2F - 1)$
Mean	$\mu$
Median	$\mu$
Mode	$\mu$
Variance	$\sigma^2$

Quantile	$\mu + \sigma\sqrt{2} \operatorname{erf}^{-1}(2F - 1)$
Mean	$\mu$
Median	$\mu$
Mode	$\mu$
Variance	$\sigma^2$

# Student's t

$$t = \frac{\mu - \bar{x}}{s / \sqrt{n}}$$

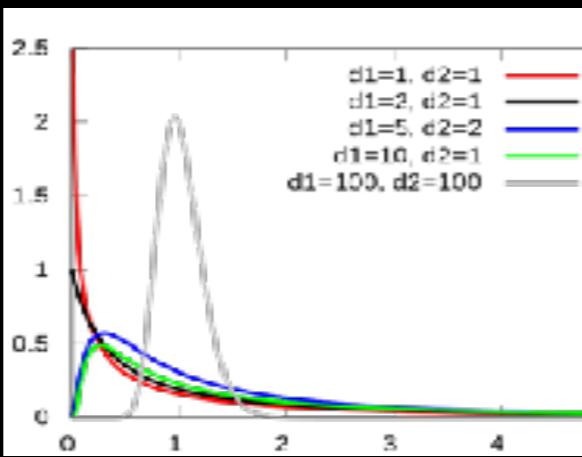


Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \\ {}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right) \\ \sqrt{\frac{\nu}{\nu+2}} \Gamma\left(\frac{\nu}{2}\right)$
where ${}_2F_1$ is the hypergeometric function	

Mean	0 for $\nu > 1$ , otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$ , = for $1 < \nu \leq 2$ , otherwise undefined

# F statistics

$$F = \frac{\sum_i n_i (\bar{x}_i - \bar{\bar{x}})^2 / (K-1)}{\sum_{ij} (x_{ij} - \bar{x}_i)^2 / (N-K)}$$

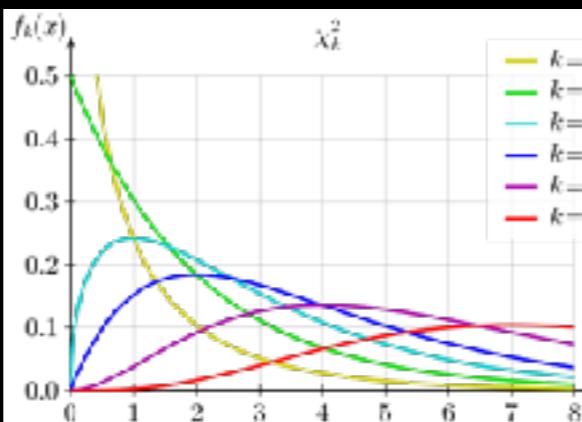


Parameters	$d_1, d_2 > 0$ deg. of freedom
Support	$x \in [0, +\infty)$
PDF	$\frac{\sqrt{(d_1 x)^{d_1} d_2^{d_2}}}{\sqrt{(d_1 + d_2)^{d_1+d_2}}} \\ x \operatorname{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$
CDF	$I \frac{d_1 x}{d_1 x + d_2} \left(\frac{d_1}{2}, \frac{d_2}{2}\right)$

Mean	$\frac{d_2}{d_2 - 2}$ for $d_2 > 2$
Mode	$\frac{d_1 - 2}{d_1} \frac{d_2}{d_2 + 2}$ for $d_1 > 2$
Variance	$\frac{2 d_2^2 (d_1 + d_2 - 2)}{d_1 (d_2 - 2)^2 (d_2 - 4)}$ for $d_2 > 4$
Skewness	$\frac{(2d_1 + d_2 - 2)\sqrt{8(d_2 - 4)}}{(d_2 - 6)\sqrt{d_1(d_1 + d_2 - 2)}}$ for $d_2 > 6$

# Pearson's $\chi^2$

$$\chi_P^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

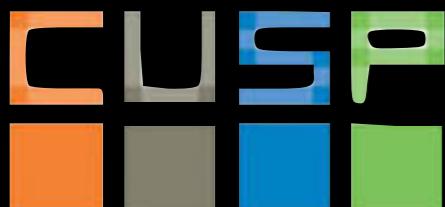


Notation	$\chi^2(k)$ or $\chi_k^2$
Parameters	$k \in \mathbb{N}_{>0}$ (known as "degrees of freedom")
Support	$x \in [0, +\infty)$
PDF	$\frac{1}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$
CDF	$\frac{1}{\Gamma\left(\frac{k}{2}\right)} \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$

Mean	$k$
Median	$\approx k \left(1 - \frac{2}{9k}\right)^3$
Mode	$\max\{k-2, 0\}$
Variance	$2k$
Skewness	$\sqrt{\frac{8}{k}}$

see

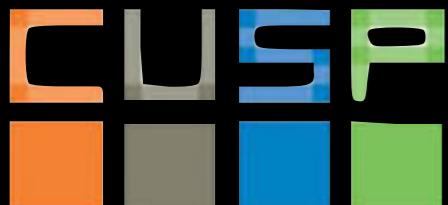
goodness of fit  $\chi^2$   $\chi_F^2 = \sum_i \frac{(m_i - x_i)^2}{e_i}$  - Statistics in a Nutshell  
IV: Statistical analysis



# Steps in Hypothesis Testing

1. Formulate Null (and alternative) Hypothesis
2. Choose a significance level  $\alpha$
2. Measure a *statistic* for a *sample* to be compared to the *parameter of a population*  
OR  
Measure a *statistic* for *two or more samples* to be compared to *each other*
4. Assess if your statistic is significant or not. In practice: compare the probability of your statistic (Z, t, F, chisq) value with a distribution table

<https://documents.software.dell.com/statistics/textbook/distribution-tables>



Is there a difference between means or population and sample,  
difference between proportion in 2 samples?

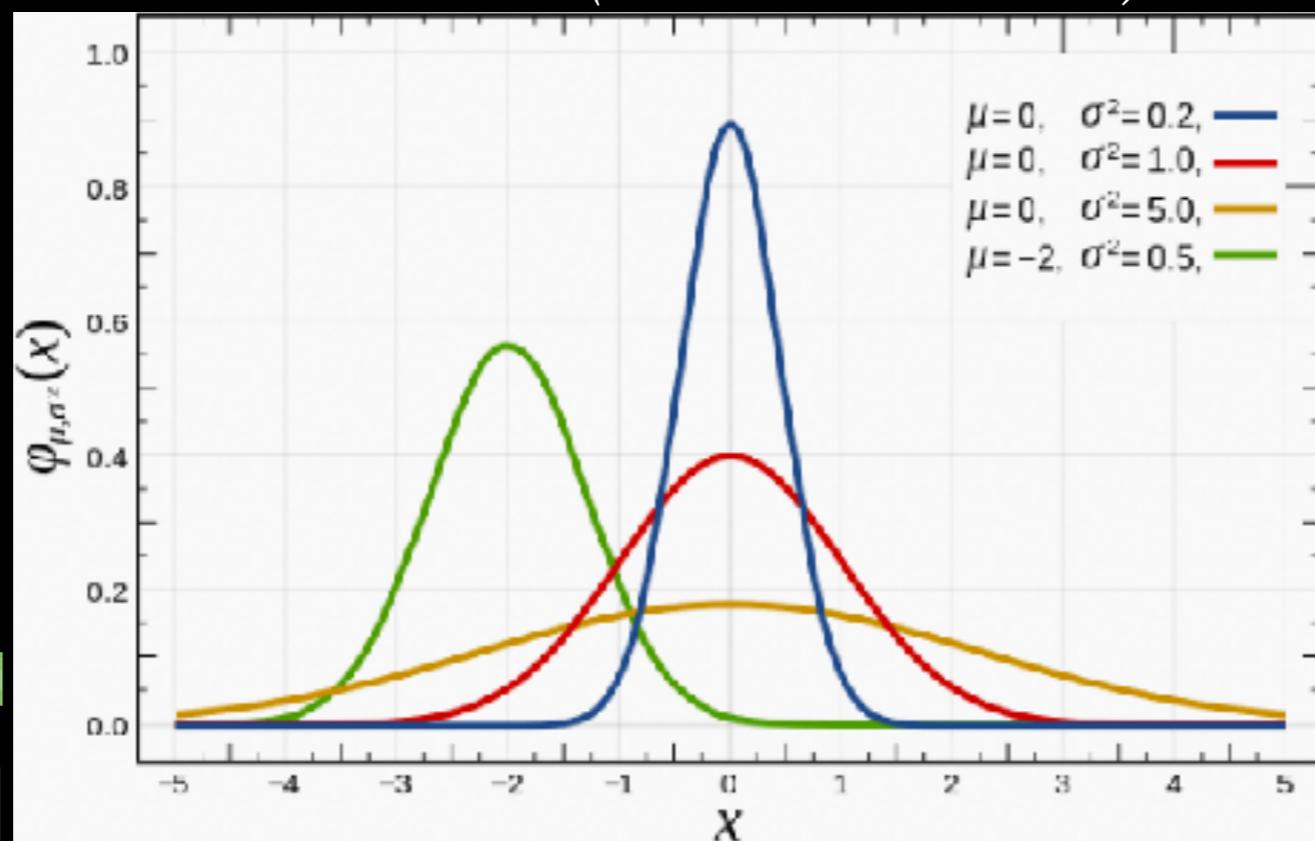
## Z statistics

$$Z = \frac{\mu - \bar{x}}{\sigma / \sqrt{n}}$$

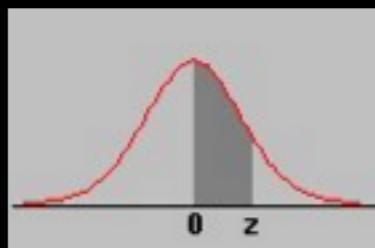
*In absence of effect (i.e. under the Null)*

**Z** is distributed according to a **Standard Normal**  $N(\mu=0, \sigma=1)$

(i.e. Gaussian)



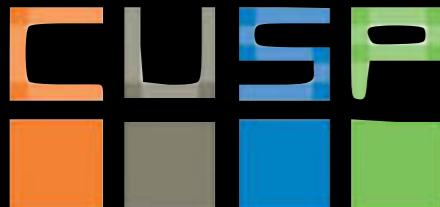
Notation	$\mathcal{N}(\mu, \sigma^2)$
Parameters	$\mu \in \mathbb{R}$ – mean ( <b>location</b> ) $\sigma^2 > 0$ – variance (squared <b>scale</b> )
Support	$x \in \mathbb{R}$
PDF	$\frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
CDF	$\frac{1}{2} \left[ 1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right) \right]$
Quantile	$\mu + \sigma\sqrt{2} \text{erf}^{-1}(2F - 1)$
Mean	$\mu$
Median	$\mu$
Mode	$\mu$
Variance	$\sigma^2$

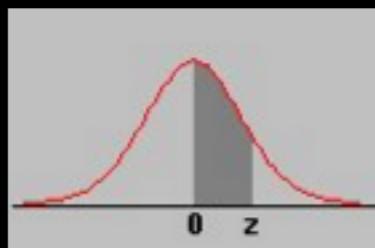


$$Z = \frac{\mu_{\text{pop}} - \mu_{\text{sample}}}{\sigma / \sqrt{N}} = 2.56$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359	1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753	1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141	1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517	1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879	1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224	2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549	2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852	2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133	2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389	2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621	2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830	2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015	2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177	2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319	2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
											3.0	0.4987	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

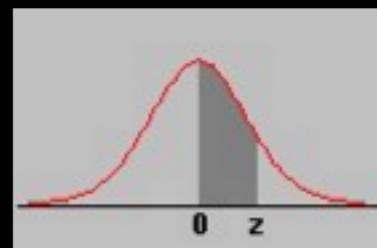
<https://github.com/fedhere/UInotebooks/blob/master/HowToReadZandChisqTables.md>





$$Z = \frac{\mu_{\text{pop}} - \mu_{\text{sample}}}{\sigma / \sqrt{N}} = 2.56$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359	1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753	1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141	1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517	1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879	1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224	2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549	2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852	2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133	2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389	2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621	2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830	2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015	2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177	2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319	2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
											3.0	0.4987	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

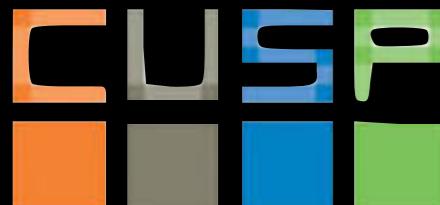


$$Z = \frac{\mu_{\text{pop}} - \mu_{\text{sample}}}{\sigma / \sqrt{N}} = 2.56$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359	1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753	1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141	1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517	1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879	1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224	2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549	2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852	2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133	2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389	2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621	2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830	2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4951	0.4962	0.4963	0.4964
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015	2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177	2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319	2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
											3.0	0.4987	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

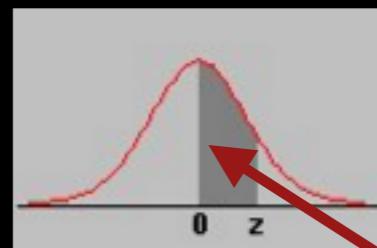
<https://github.com/fedhere/UInotebooks/blob/master/HowToReadZandChisqTables.md>

1 sided test     $1 - 0.4948 - 0.5 = 0.0052$   
                              $p < 0.05$



$H_0$  IS REJECTED ( $p < 0.05$ )

IV: Statistical analysis

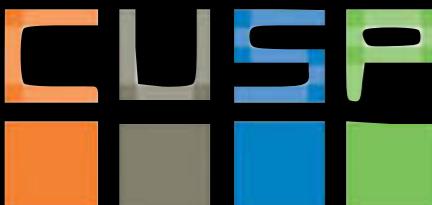


$$Z = \frac{\mu_{\text{pop}} - \mu_{\text{sample}}}{\sigma / \sqrt{N}} = 2.56$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359		1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753		1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141		1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517		1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879		1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224		2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549		2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852		2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133		2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389		2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621		2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830		2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4951	0.4962	0.4963	0.4964
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015		2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177		2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319		2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
												3.0	0.4987	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

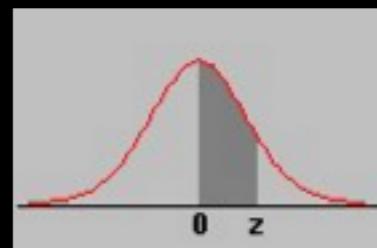
<https://github.com/fedhere/UInotebooks/blob/master/HowToReadZandChisqTables.md>

1 sided test     $1 - 0.4948 - 0.5 = 0.0052$   
 $p < 0.05$



$H_0$  IS REJECTED ( $p < 0.05$ )

IV: Statistical analysis

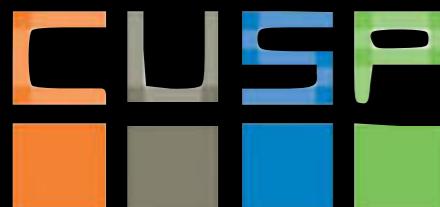


$$Z = \frac{\mu_{\text{pop}} - \mu_{\text{sample}}}{\sigma / \sqrt{N}} = 2.55$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359	1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753	1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141	1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517	1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879	1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224	2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549	2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852	2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133	2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389	2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621	2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830	2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015	2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177	2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319	2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
											3.0	0.4987	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4990	0.4990

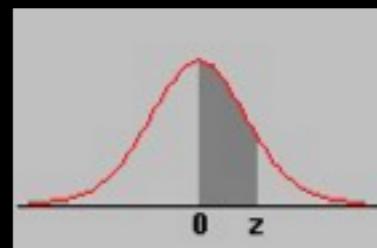
<https://github.com/fedhere/UInotebooks/blob/master/HowToReadZandChisqTables.md>

1 sided test     $1 - 0.4946 - 0.5 = 0.0054$   
                              $p < 0.05$



$H_0$  IS REJECTED ( $p < 0.05$ )

IV: Statistical analysis

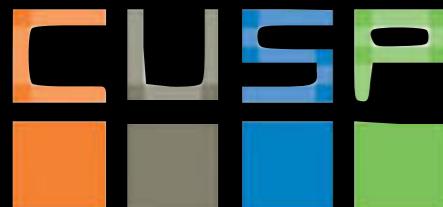


$$Z = \frac{\mu_{\text{pop}} - \mu_{\text{sample}}}{\sigma / \sqrt{N}} = 1.57$$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09		0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359	1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753	1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141	1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517	1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879	1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224	2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549	2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852	2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133	2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389	2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621	2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830	2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015	2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177	2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319	2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
											3.0	0.4987	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990

<https://github.com/fedhere/UInotebooks/blob/master/HowToReadZandChisqTables.md>

2 sided test       $1 - 0.4418^2 = 0.1164$   
 $p > 0.05$



$H_0$  CANNOT BE REJECTED

IV: Statistical analysis

# Is there a difference between means of 2 sample?

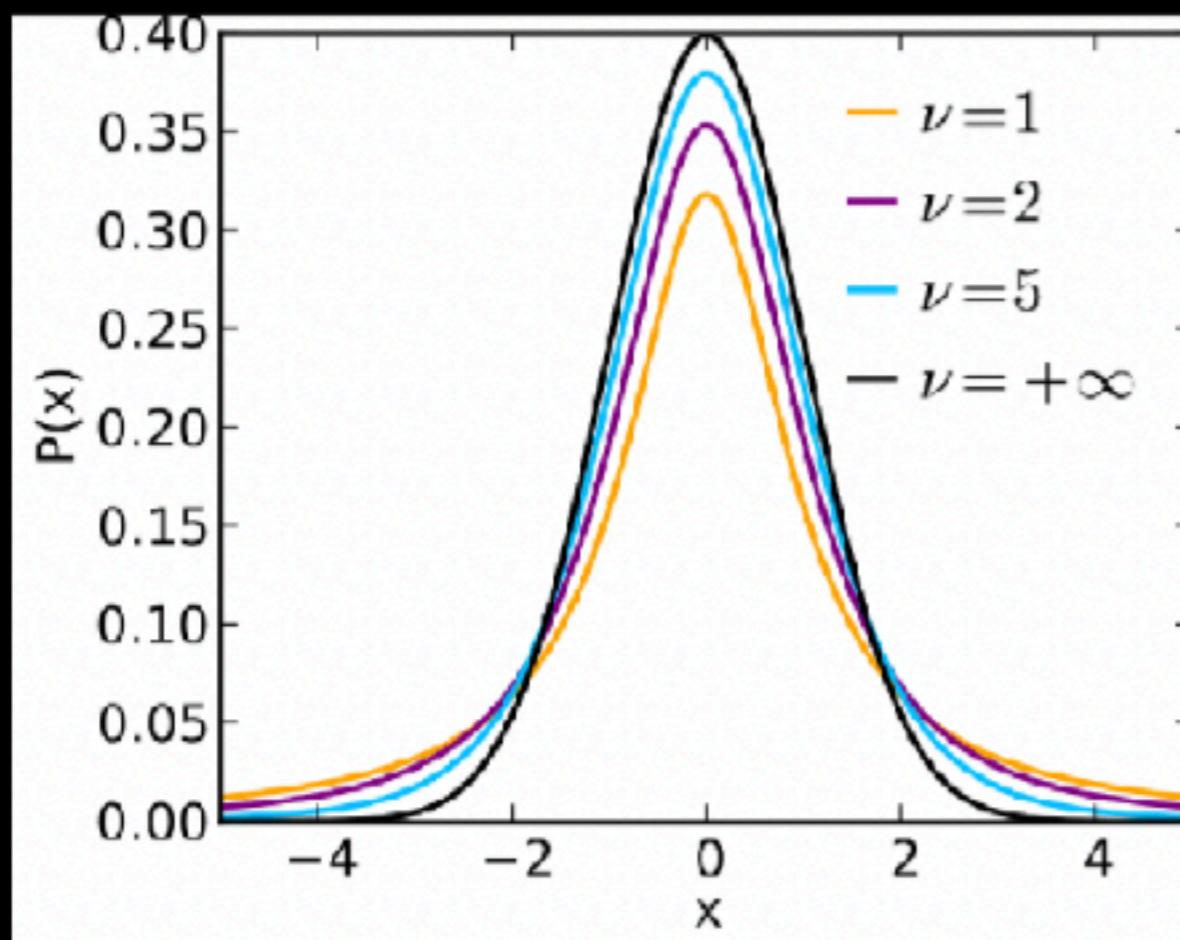
*In absence of effect (i.e. under the Null)*

## Student's $t$

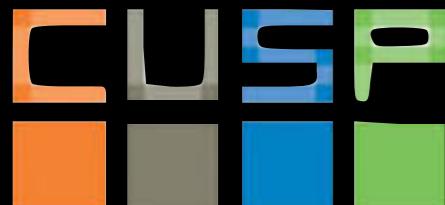
$$t = \frac{\mu - \bar{x}}{s/\sqrt{n}}$$

$t$  test for 1 small samples and a population

$t$  is distributed according to a  **$t$ -distribution** with  $v =$  degrees of freedom of the problem



Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}$ <p>where <math>{}_2F_1</math> is the hypergeometric function</p>
Mean	0 for $\nu > 1$ , otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$ , $\infty$ for $1 < \nu \leq 2$ , otherwise undefined



$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

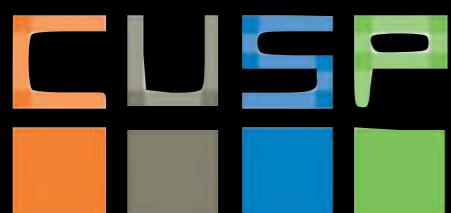
$t$  test between 2 samples (*unpaired t test*)

IV: Statistical analysis

# Comparing to population

## Standard DEVIATION of Sample Estimates

Sample mean, $\bar{x}$	$Z = \frac{\mu_{\text{pop}} - \bar{x}}{\sigma / \sqrt{n}}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Sample proportion, $p$	$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$\sigma_p = \sqrt{\frac{p_0(1-p_0)}{n}}$
Difference between proportions, $p_1 - p_2$	$z = \frac{(p_2 - p_1)}{\sqrt{p(1-p)(\frac{1}{n_2} + \frac{1}{n_1})}}, p = \frac{p_2 n_2 + p_1 n_1}{n_2 + n_1}$	$\sigma_{p_1 - p_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$

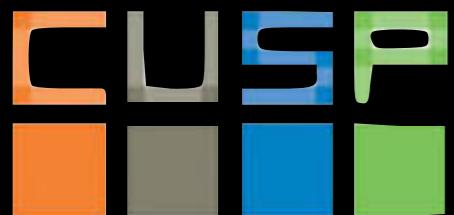


Use if you know the population *parameters*  
e.g.  $Z$ -test

# Comparing samples

## Standard ERROR of Sample Estimates

Sample mean, $\bar{x}$	$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$
Sample proportion, p	$SE_p = \frac{p(1-p)}{n}$
Difference between means, $x_1 - x_2$	$SE_{x_1 - x_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Difference between proportions, $p_1 - p_2$	$SE_{p_1 - p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$



Use if you DO NOT know the population *parameters*  
e.g.  $t$  - test

# Is there a difference between means of 2 sample?

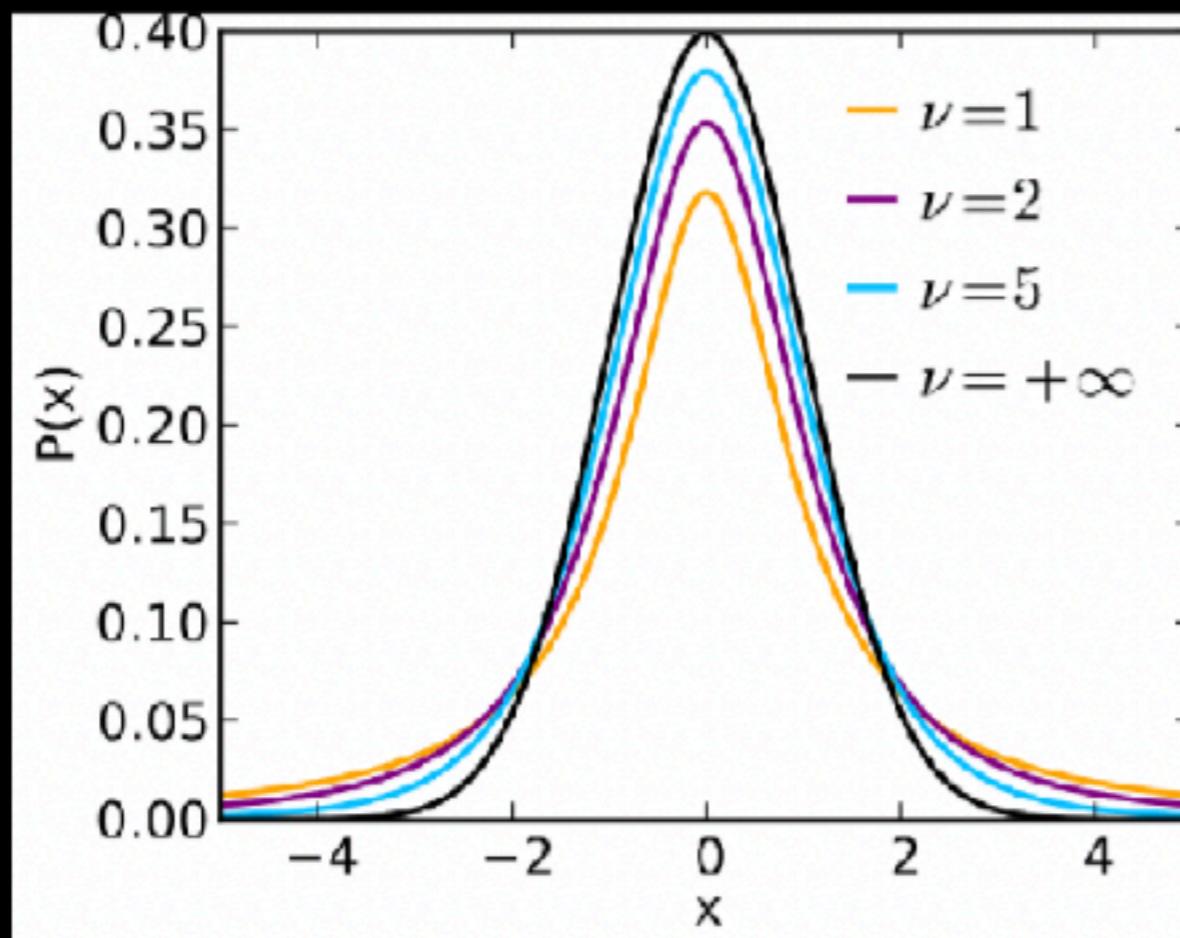
*In absence of effect (i.e. under the Null)*

## Student's $t$

$$t = \frac{\mu - \bar{x}}{s/\sqrt{n}}$$

$t$  test for 1 small samples and a population

$t$  is distributed according to a  **$t$ -distribution** with  $v =$  degrees of freedom of the problem

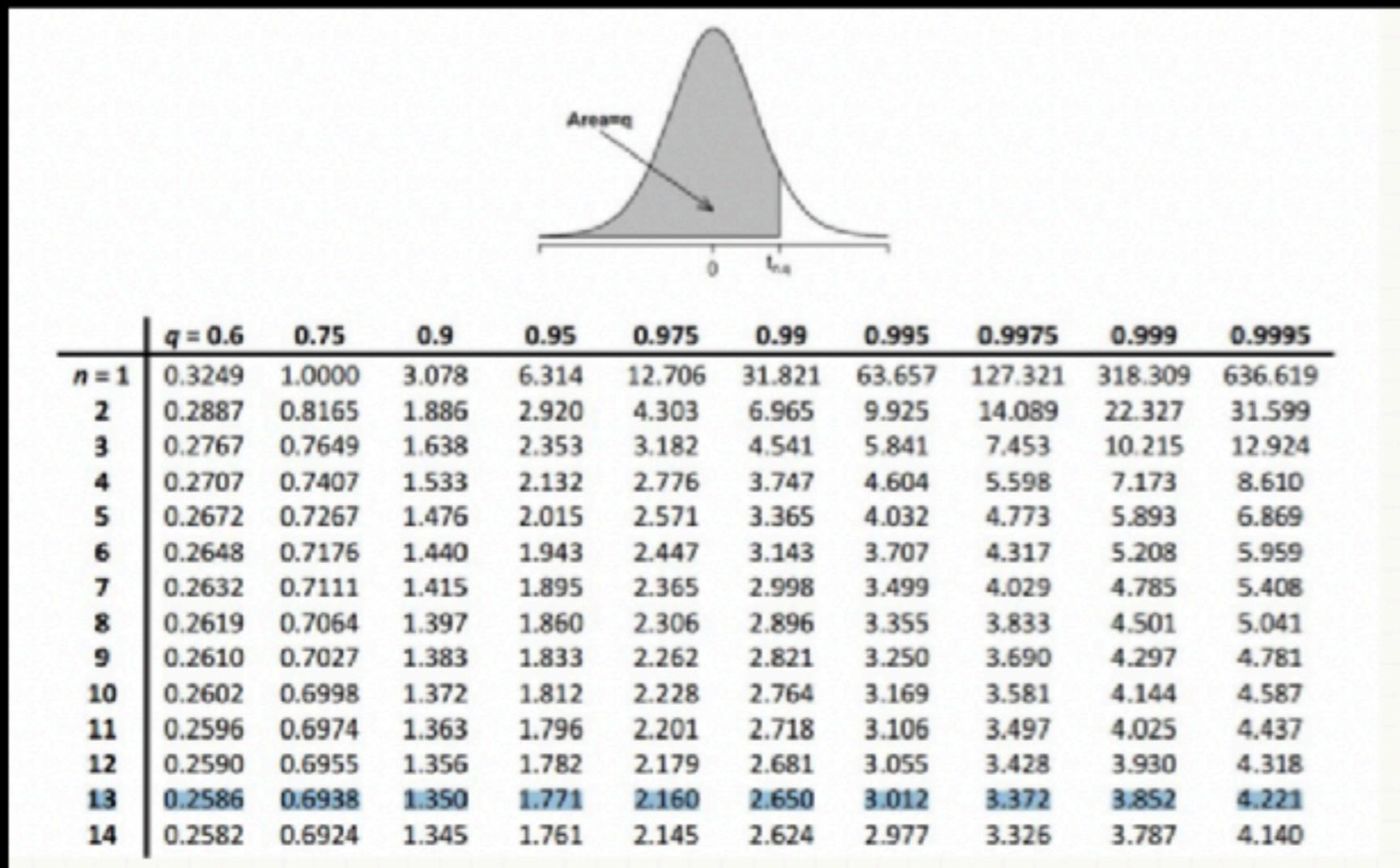


Parameters	$\nu > 0$ degrees of freedom (real)
Support	$x \in (-\infty; +\infty)$
PDF	$\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$
CDF	$\frac{1}{2} + x\Gamma\left(\frac{\nu+1}{2}\right) \times \frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; -\frac{x^2}{\nu}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}$ <p>where <math>{}_2F_1</math> is the hypergeometric function</p>
Mean	0 for $\nu > 1$ , otherwise undefined
Median	0
Mode	0
Variance	$\frac{\nu}{\nu-2}$ for $\nu > 2$ , $\infty$ for $1 < \nu \leq 2$ , otherwise undefined

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

# Is there a difference between means of 2 sample?

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} .75$$



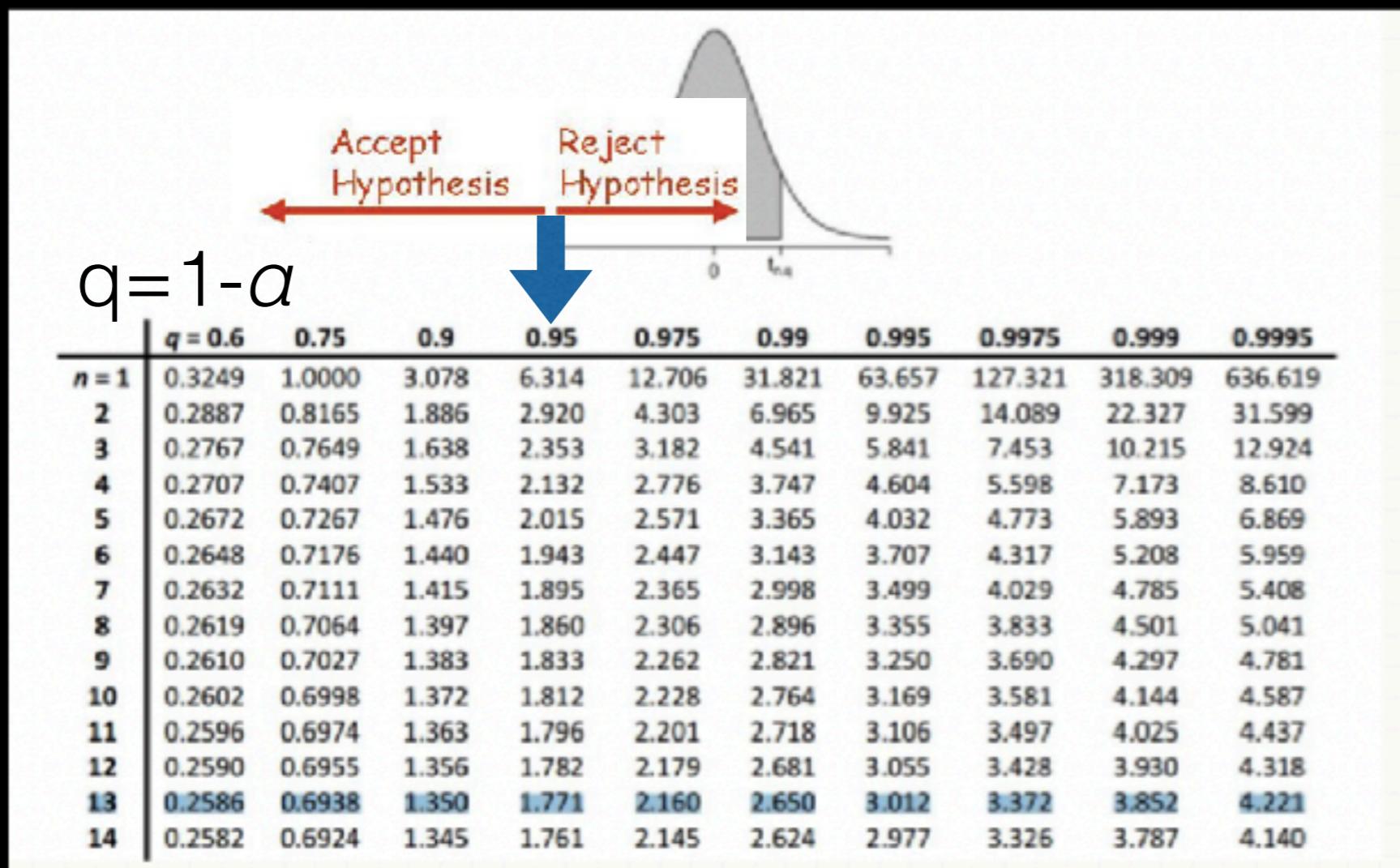
$\leftarrow 1 - \alpha$

↑  
d.o.f :  $n - 2$  (1 sample) or  $n_1 + n_2 - 2$  (2 samples)

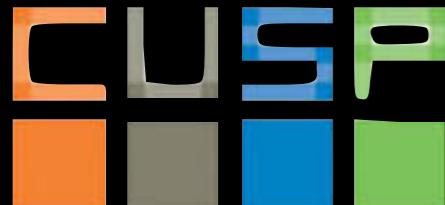
# Is there a difference between means of 2 sample?

$$t = \frac{\mu - x}{s/\sqrt{n}} = 1.75$$

$n=13$ ,  $q=.95$



d.o.f



Generally d.o.f =  $n_{observations} - 1$

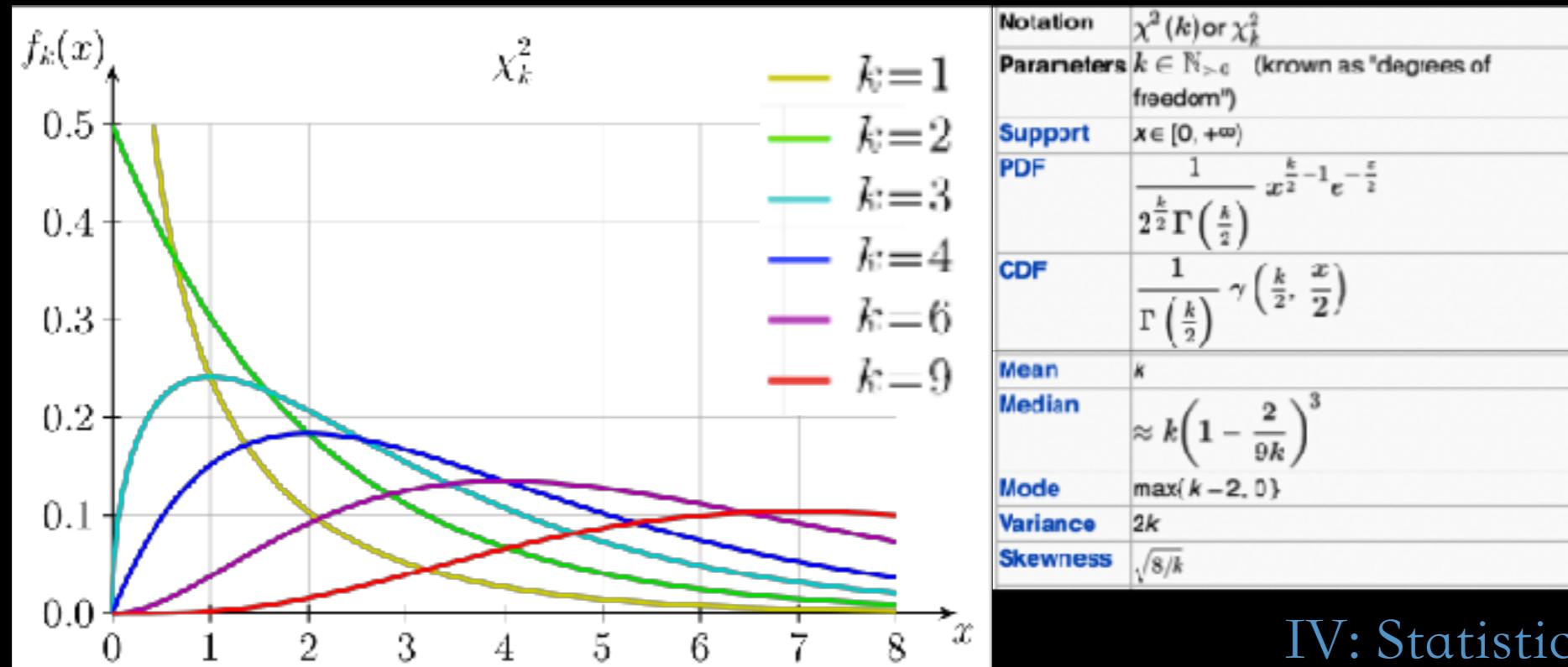
IV: Statistical analysis

Is there a difference between means or population and sample,  
difference between proportion in 2 samples?

$\chi^2$  statistics

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

In absence of effect (i.e. under the Null)  
 $\chi^2$  is distributed according to a  $\chi^2$   
distribution with  $k$ =degrees of freedom  
 $O$ : observed  
 $E$ : expected (also model prediction)  
 $N$ : observations



*observed* → *expected*

$$\chi_P^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

For test of proportion

*expected*

4 observations - 1 independent variable =  
3 degreeed of freedom

Accept Hypothesis      Reject Hypothesis

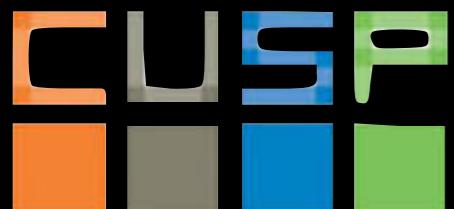
← →

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09

$$\alpha = 0.05$$

$$p < 0.05$$



$H_0$  CAN BE REJECTED

IV: Statistical analysis

For goodness of fit

$$model \quad \chi^2_F = \sum_{i=1}^4 \frac{(m_i - x_i)^2}{e_i} \quad data \quad error$$

4 observations - 1 independent variable =  
3 degrees of freedom

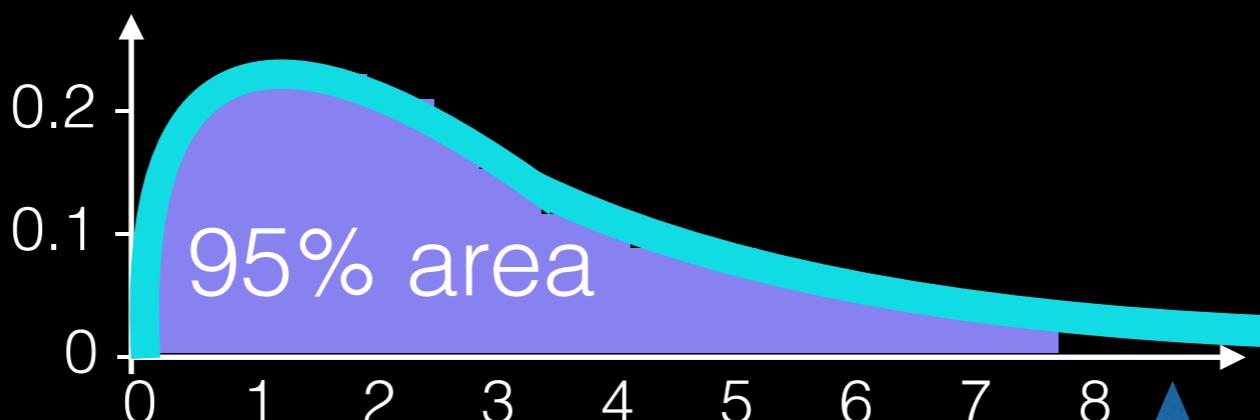
The table shows the percentage points of the Chi-Square distribution. The columns represent the probability of a larger value of  $\chi^2$ , ranging from 0.99 down to 0.01. The rows represent degrees of freedom (DF) from 1 to 5. A red circle highlights the value at DF=3, P=0.05, which is 7.81. A horizontal double-headed arrow above the table indicates the range of values where the null hypothesis is accepted (to the left) or rejected (to the right).

Degrees of Freedom	Probability of a larger value of $\chi^2$								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09

$$\alpha = 0.05$$

$$p < 0.05$$

$H_0$  CAN BE REJECTED



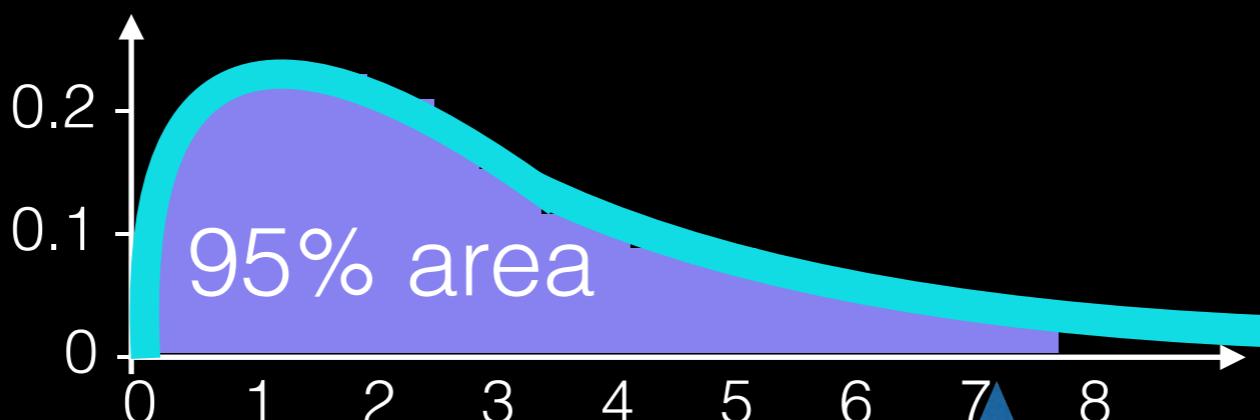
4 observations - 1 independent variable =  
3 degrees of freedom

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of $\chi^2$								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09

$$\alpha = 0.05$$

$$\chi_P^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 8.57 \quad 8.57 > 7.81 \\ p < 0.05$$



4 observations - 1 independent variable =  
3 degrees of freedom

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of $\chi^2$								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09

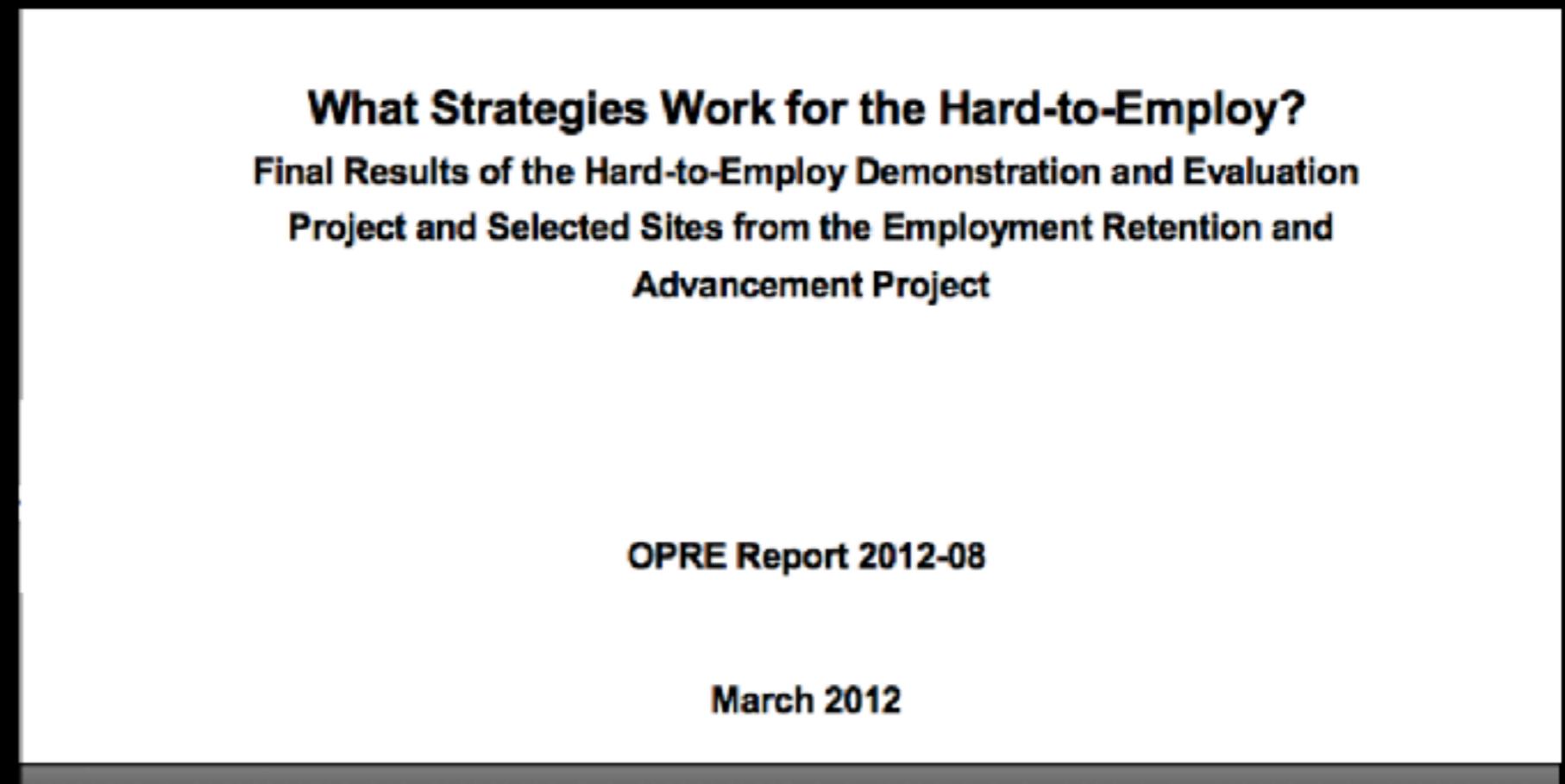
$$\alpha = 0.05$$

$$\chi_P^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = 7.11 \quad 7.11 < 7.81$$

$p \geq 0.05$

$H_0$  CANNOT BE REJECTED

**Example:** **NULL HYPOTHESIS:** the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,  
*significance level p=0.05*



<https://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf>

**NULL HYPOTHESIS:** the % of former prisoners employed 3 years after release is *the same or lower* for candidates who participated in the program as for the control group,  
*significance level p=0.05*

The Enhanced Services for the Hard-to-Employ Demonstration and Evaluation Project

Table 2.1

Summary of Impacts, New York City Center for Employment Opportunities

Outcome	Program Group	Control Group	Difference (Impact)	P-Value
<u>Employment (Years 1-3) (%)</u>	<b>P<sub>1</sub></b>	<b>P<sub>0</sub></b>		
Ever employed	83.8	70.4	13.4 ***	0.000
Ever employed in a CEO transitional job <sup>a</sup>	70.1	3.5	66.6 ***	0.000
Ever employed in an unsubsidized job	63.7	69.0	-5.3 *	0.078
<u>Postprogram unsubsidized employment (Years 2-3)</u>				
Ever employed in an unsubsidized job (%)	53.3	52.1	1.2	0.713
Employed in an unsubsidized job, average per quarter (%)	28.2	27.2	1.1	0.618
Employed for six or more consecutive quarters (%)	14.7	11.9	2.8	0.195
Total UI-covered earnings <sup>b</sup> (\$)	10,435	9,846	589	0.658
Sample size (total = 973) <sup>c</sup>	564	409		

$$H_0: P_0 - P_1 \geq 0$$

$$H_a: P_0 - P_1 < 0$$

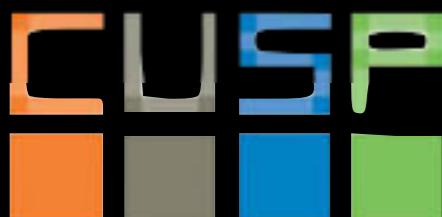
$$\alpha = 0.05$$

SOURCES: MDRC earnings calculations from the National Directory of New Hires (NDNH) database and employment calculations from the unemployment insurance (UI) wage records from New York State, MDRC calculations using data from the New York State Division of Criminal Justice Services (DCJS) and the New York City Department of Correction (DOC).

NOTES: Statistical significance levels are indicated as: \*\*\* = 1 percent; \*\* = 5 percent; \* = 10 percent.

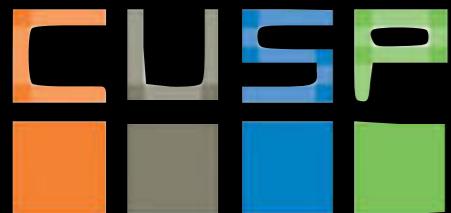
The p-value indicates the likelihood that the difference between the program and control groups arose by chance.

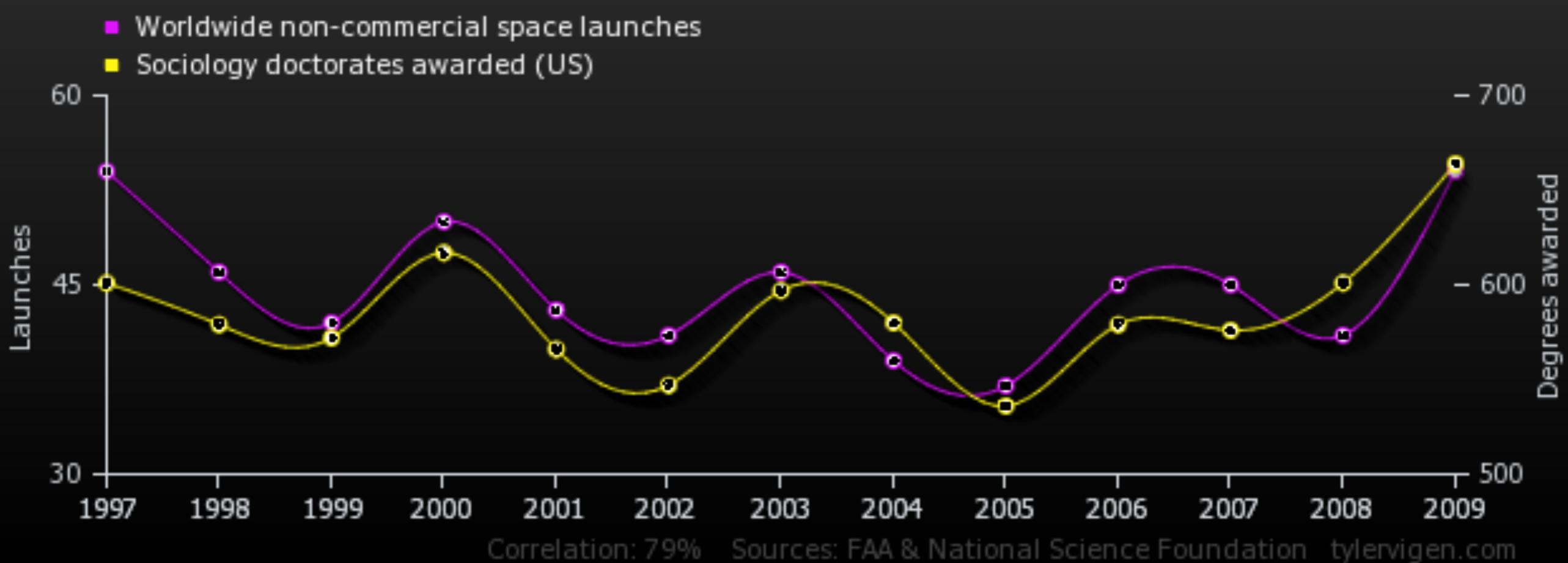
<http://www.mdrc.org/sites/default/files/What%20Strategies%20Work%20for%20the%20Hard%20FR.pdf>





[https://github.com/fedhere/PUI2018\\_fb55/blob/master/  
Lab5\\_fb55/effectiveness%20of%20NYC%20Post-  
Prison%20Employment%20Programs.ipynb](https://github.com/fedhere/PUI2018_fb55/blob/master/Lab5_fb55/effectiveness%20of%20NYC%20Post-Prison%20Employment%20Programs.ipynb)



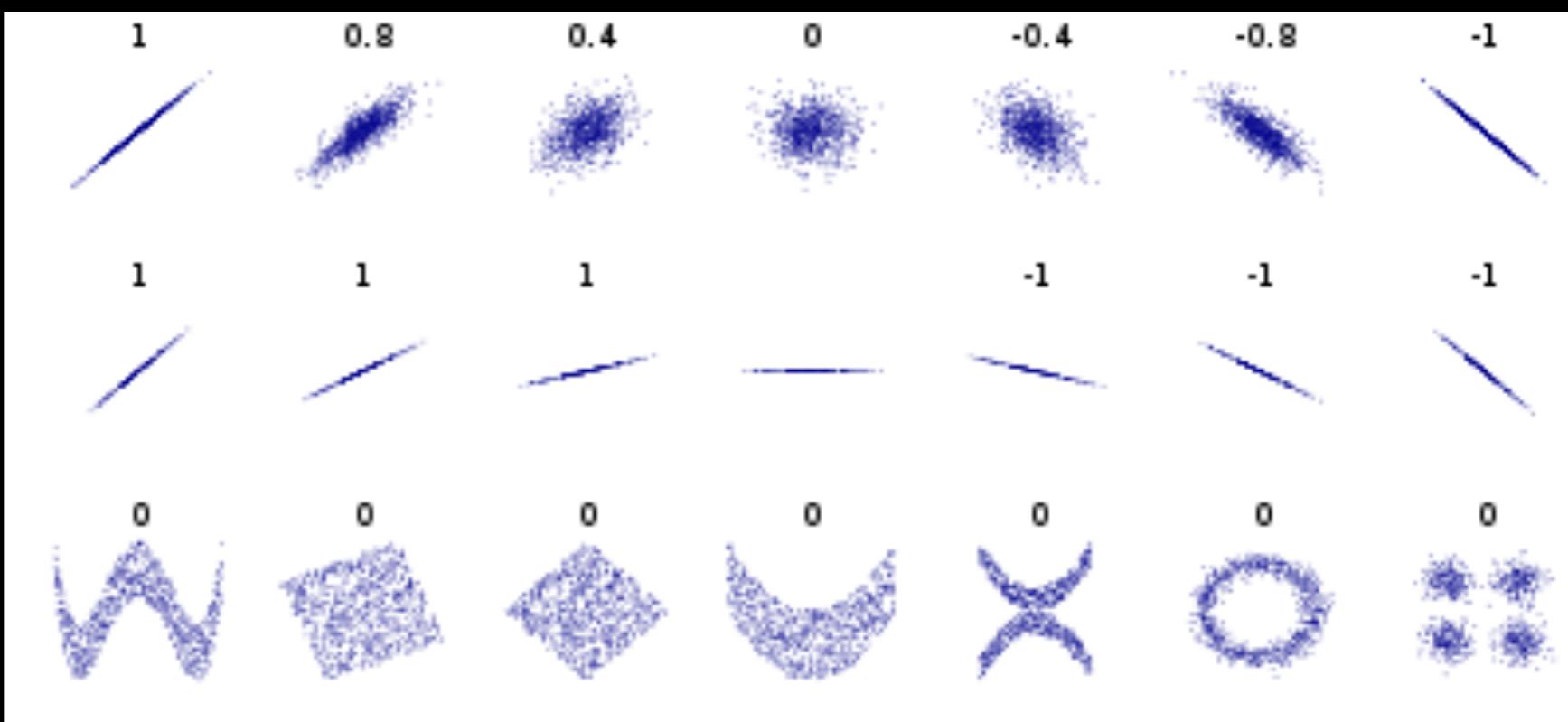


## Correlation

Pearson's test:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

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



Pearson's test:

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

Spearman's test:  
(Pearson's for ranks)

$$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}$$

### Choosing the test

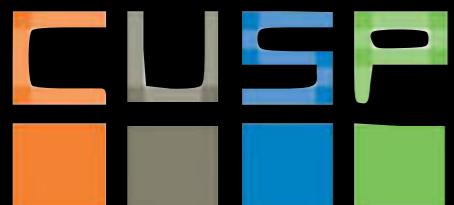
Use the table below to choose the test. See below for further details.

How many dichotomous* (binary) variables?				
Both variables interval or ratio?				
0	Y	Measures are linear? (No = monotonic*)		
		Y Pearson correlation		
0	N	N Spearman correlation		
		Both variables are ordinal?		
1	Y	Kendall correlation		
		Both variables can be ranked?		
1	N	Y Kendall correlation		
		N Convert to frequency data and use Chi-square test for independence		
1 serial Correlation Coefficient				
2 x 2 table?				
2	Y	Phi		
		N Cramer's V		
Data has frequency values for each category?				
Y	Chi-square test for independence			

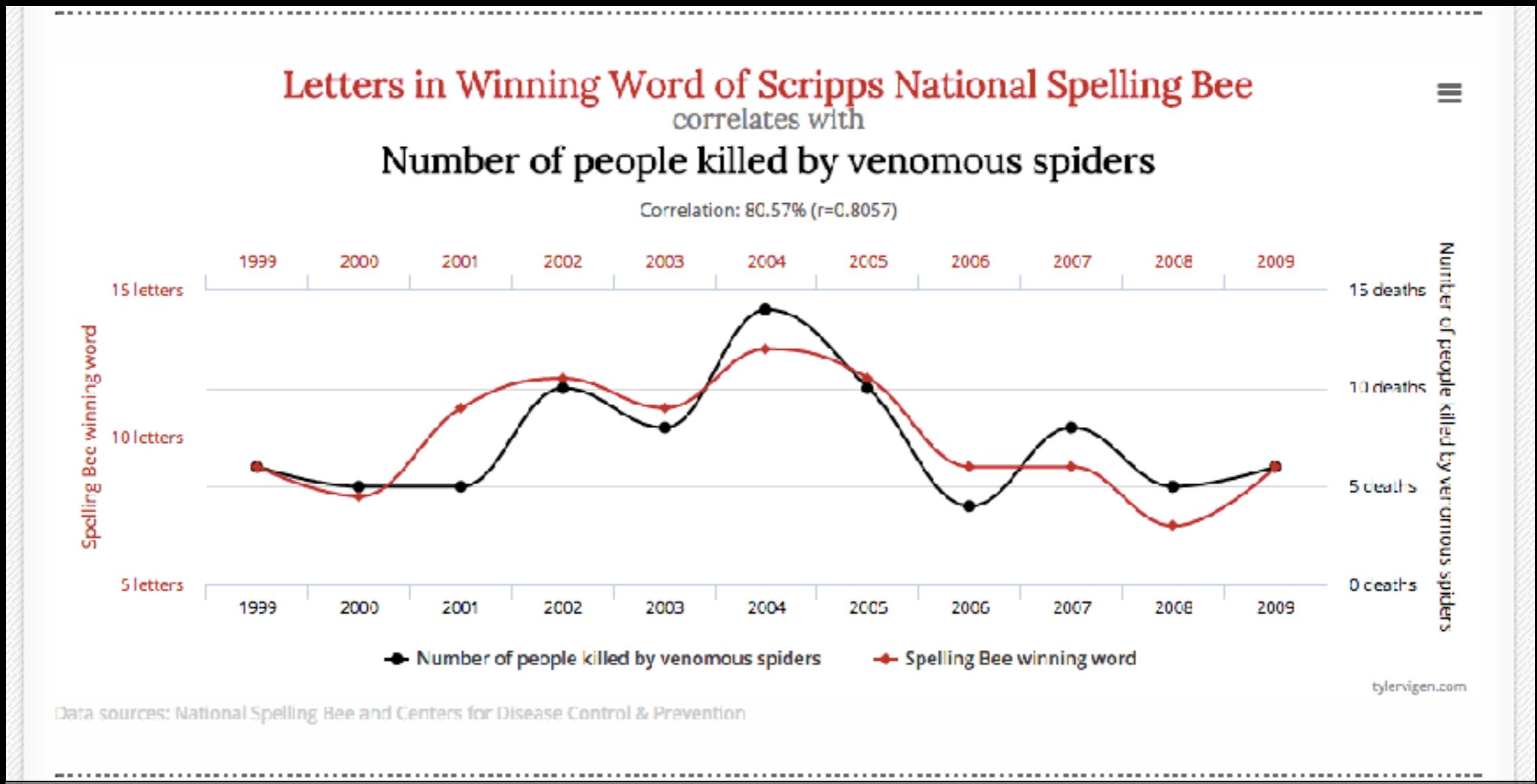
\*dichotomous = 'can have only two values' (eg. yes/no or 0/1).

†monotonic = constantly increasing or decreasing.

[http://changingminds.org/explanations/research/analysis/choose\\_correlation.htm](http://changingminds.org/explanations/research/analysis/choose_correlation.htm)



# WARNING: Correlation is not causation!



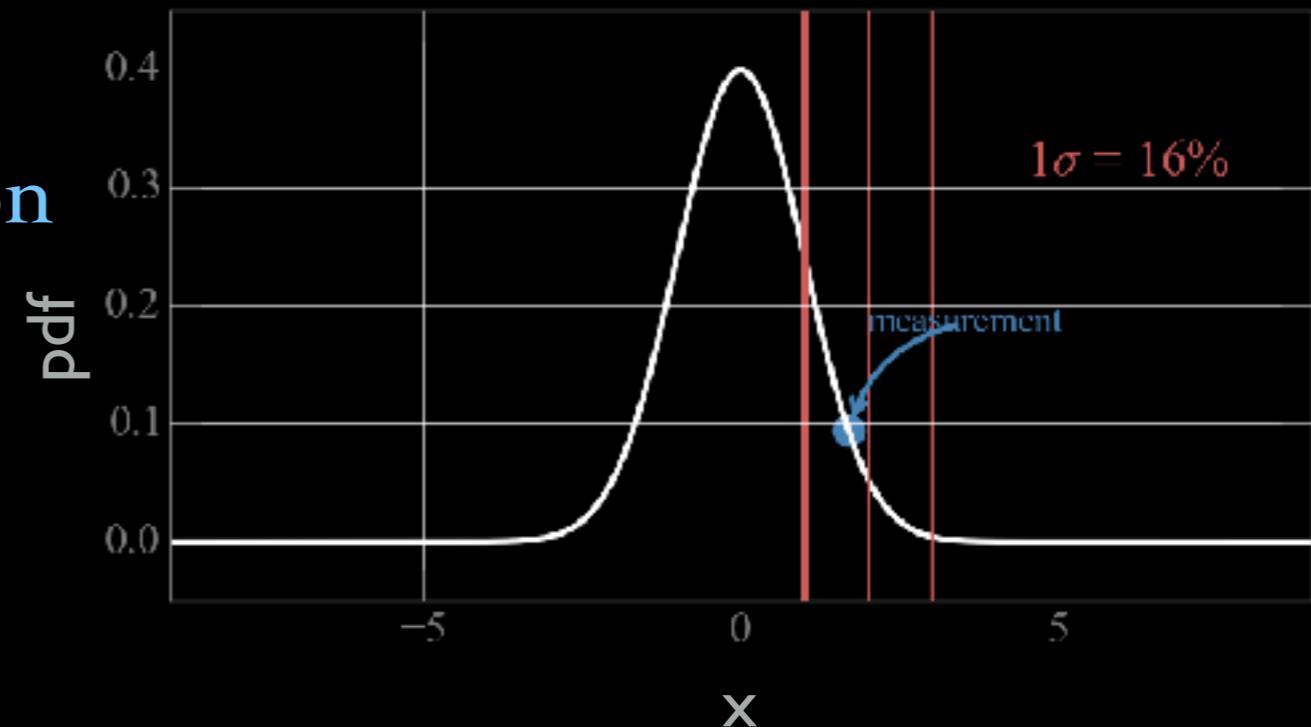
<http://www.tylervigen.com/spurious-correlations>

# Tests for correlation and independence (continuous variables)

## Probability Distribution Function

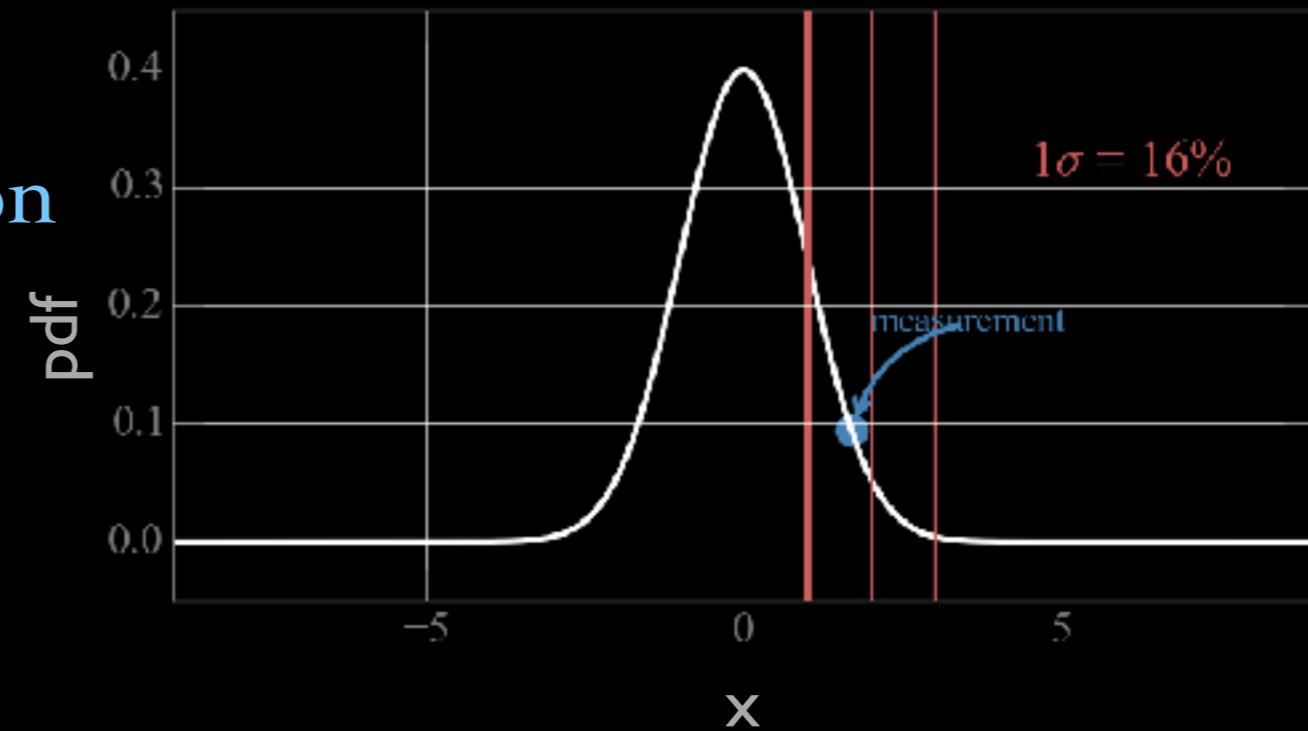
$$f_{x_0}(x) \sim p(x=x_0)$$

$$f_{x_0}(x) \sim p(x > x_0 - dx) \cap p(x < x_0 + dx)$$



## Probability Distribution Function

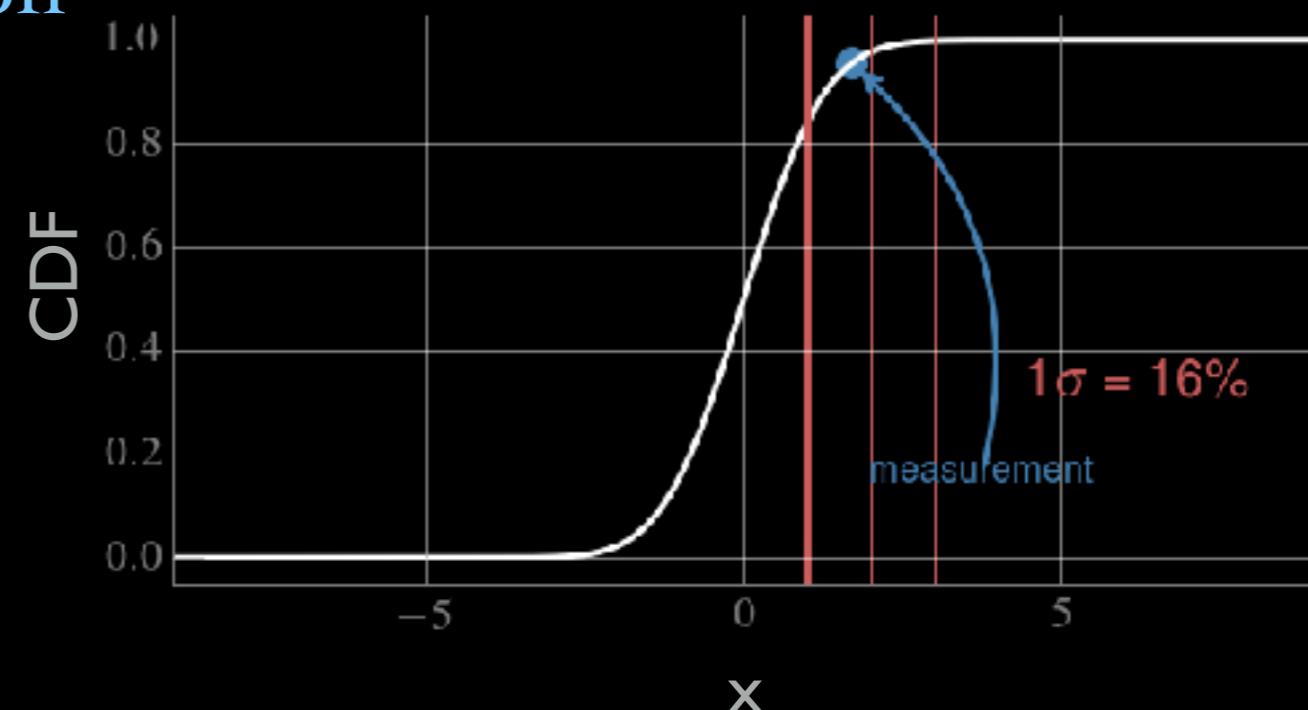
$$f_{x_0}(x) \sim p(x=x_0)$$

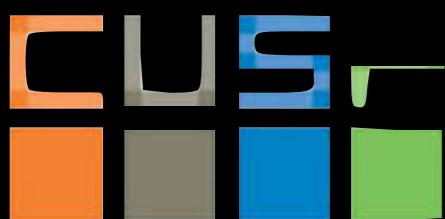
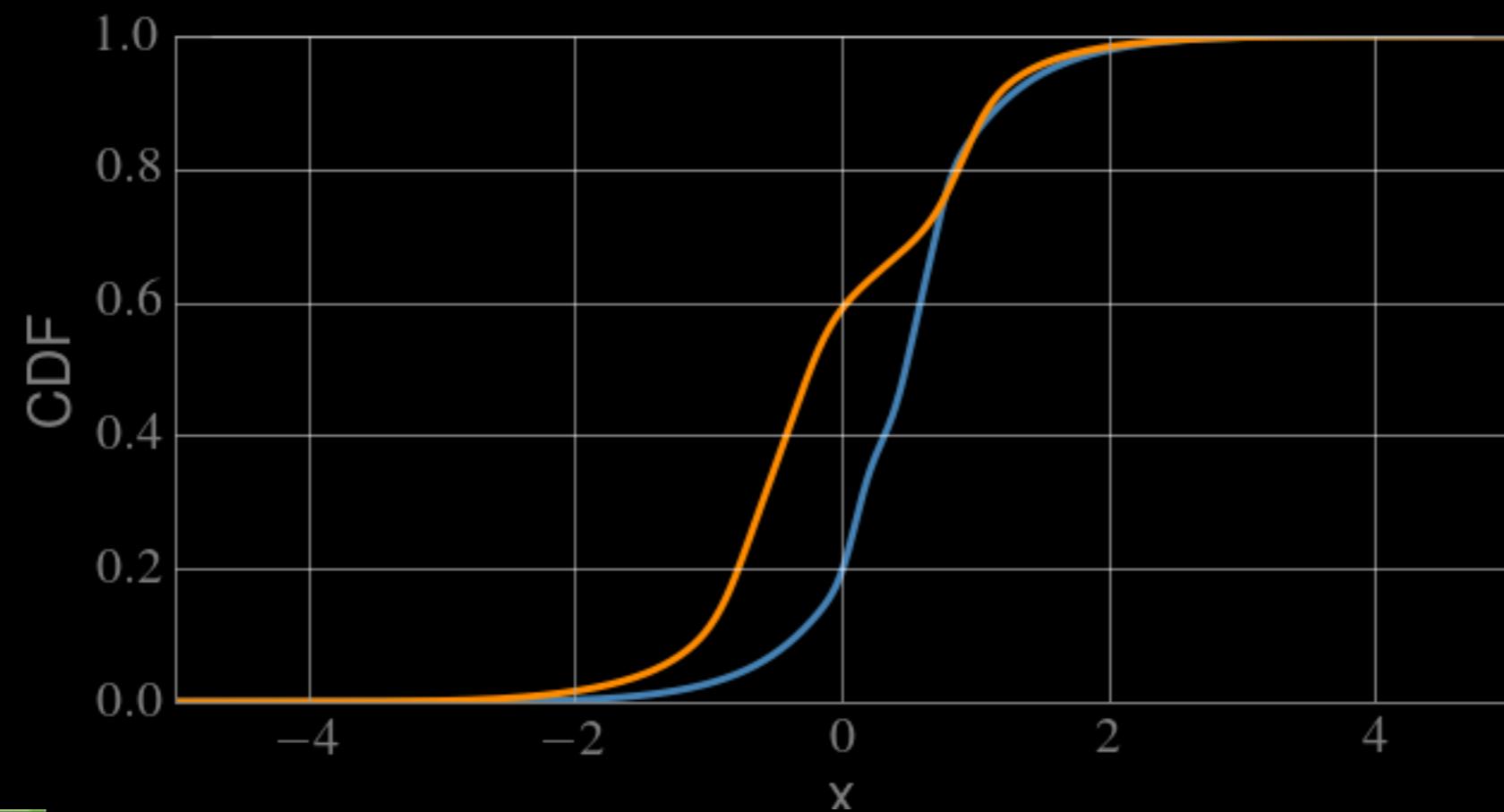
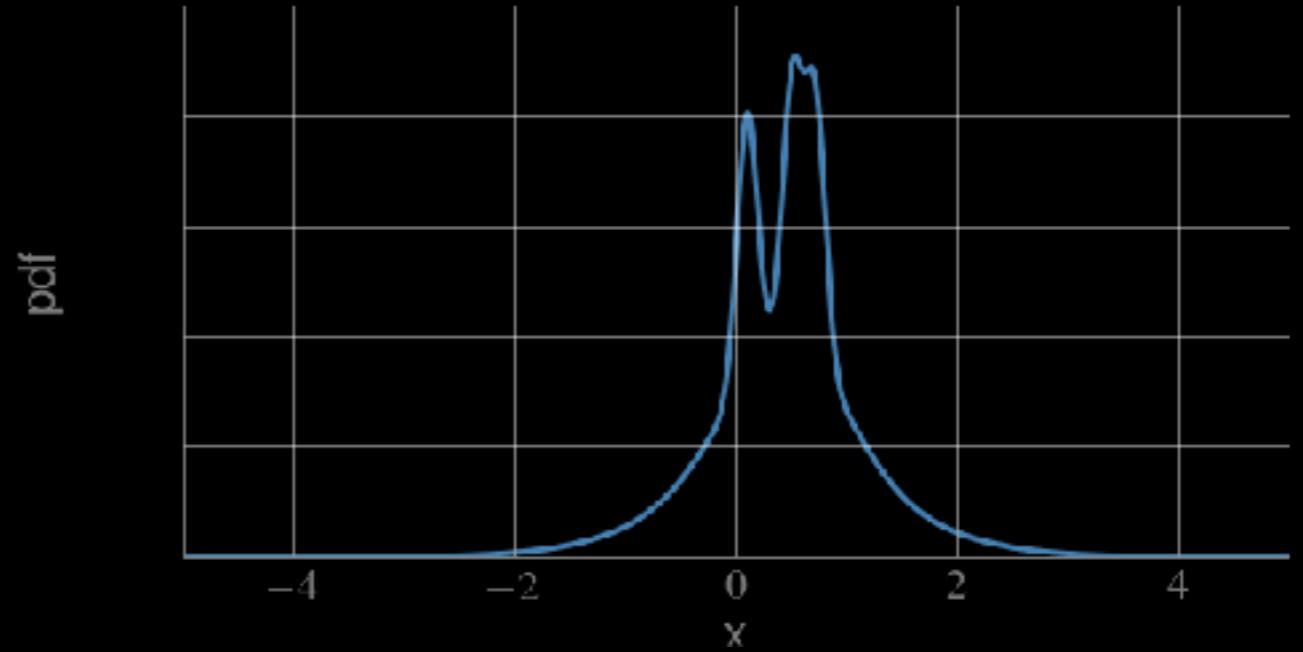
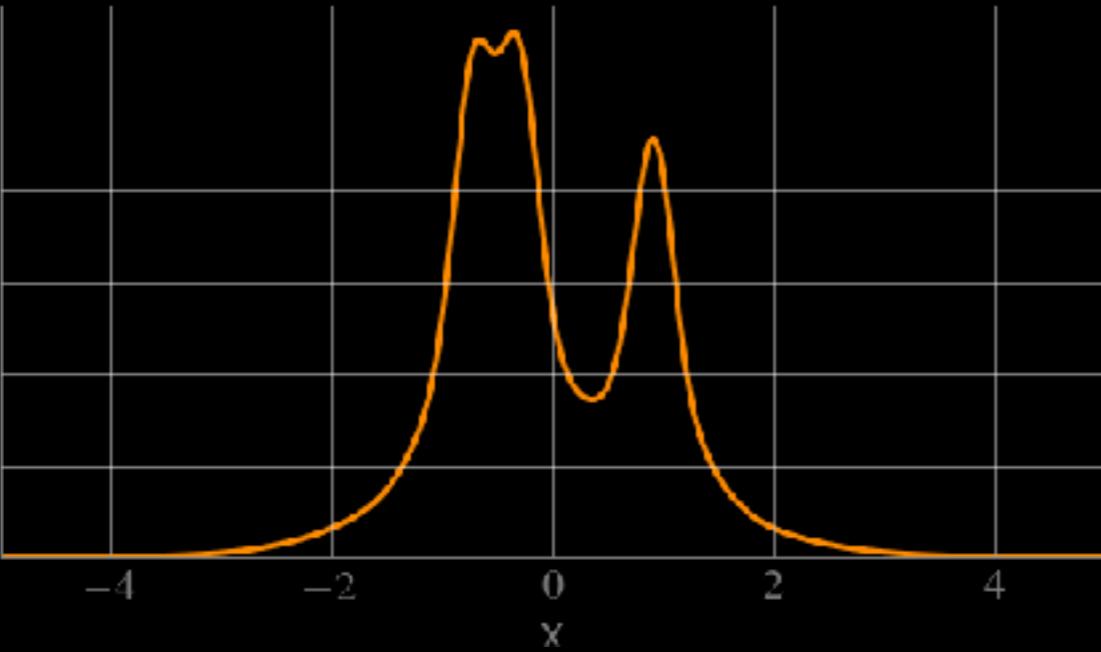


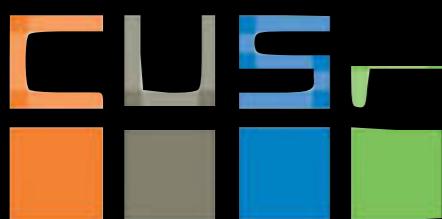
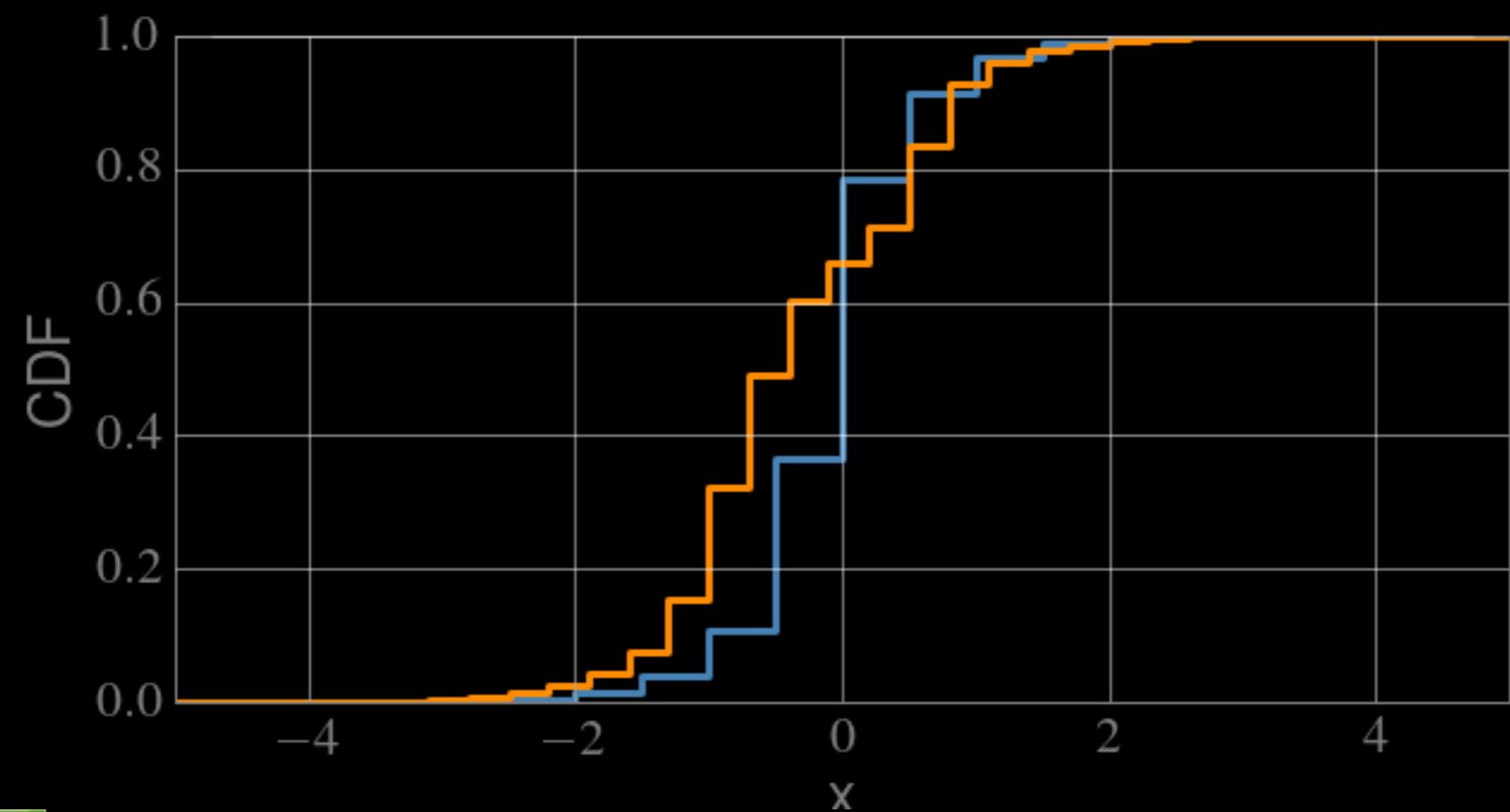
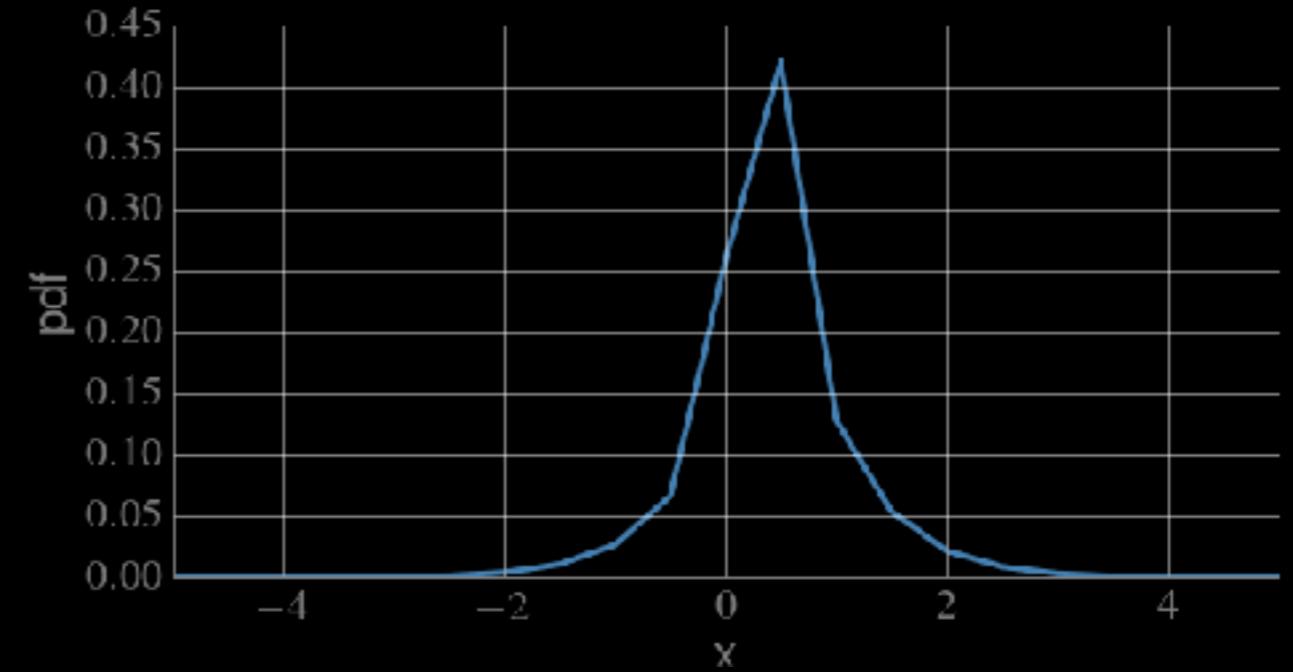
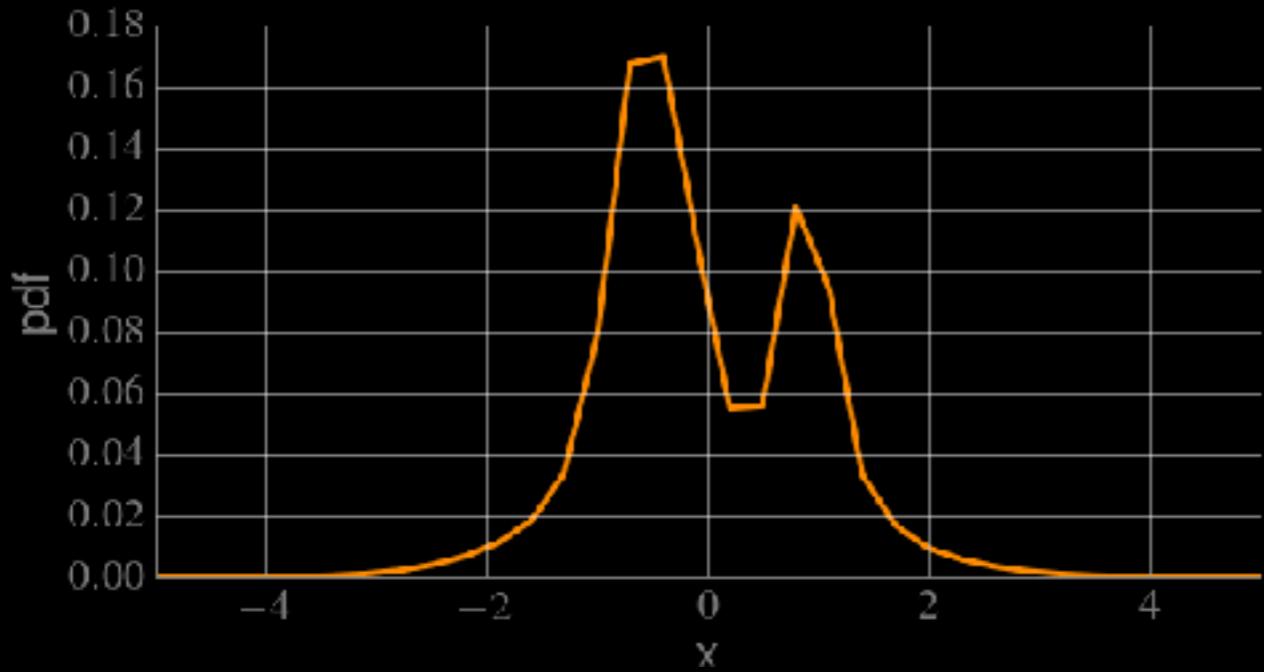
$$f_{x_0}(x) \sim p(x > x_0 - dx) \cap p(x < x_0 + dx)$$

## Cumulative Distribution Function

$$F_{x_0}(x) = P(x < x_0)$$







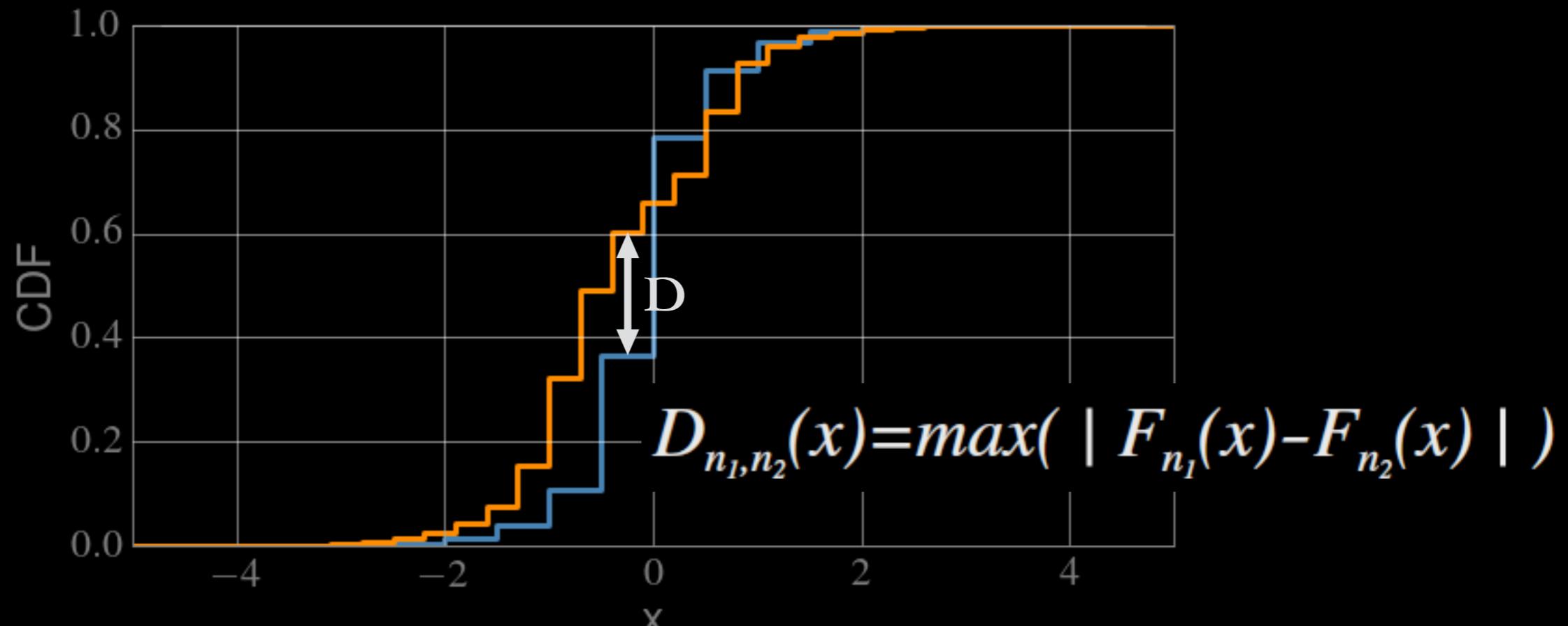
Two sample Kolmogorov Smirnoff test:

*null hypothesis*  $H_0$ : the samples come from the same parent distribution

$H_0$  is rejected at level  $\alpha$  if  $D(n_1, n_2) > c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$

with  $c(\alpha)$  given by a table

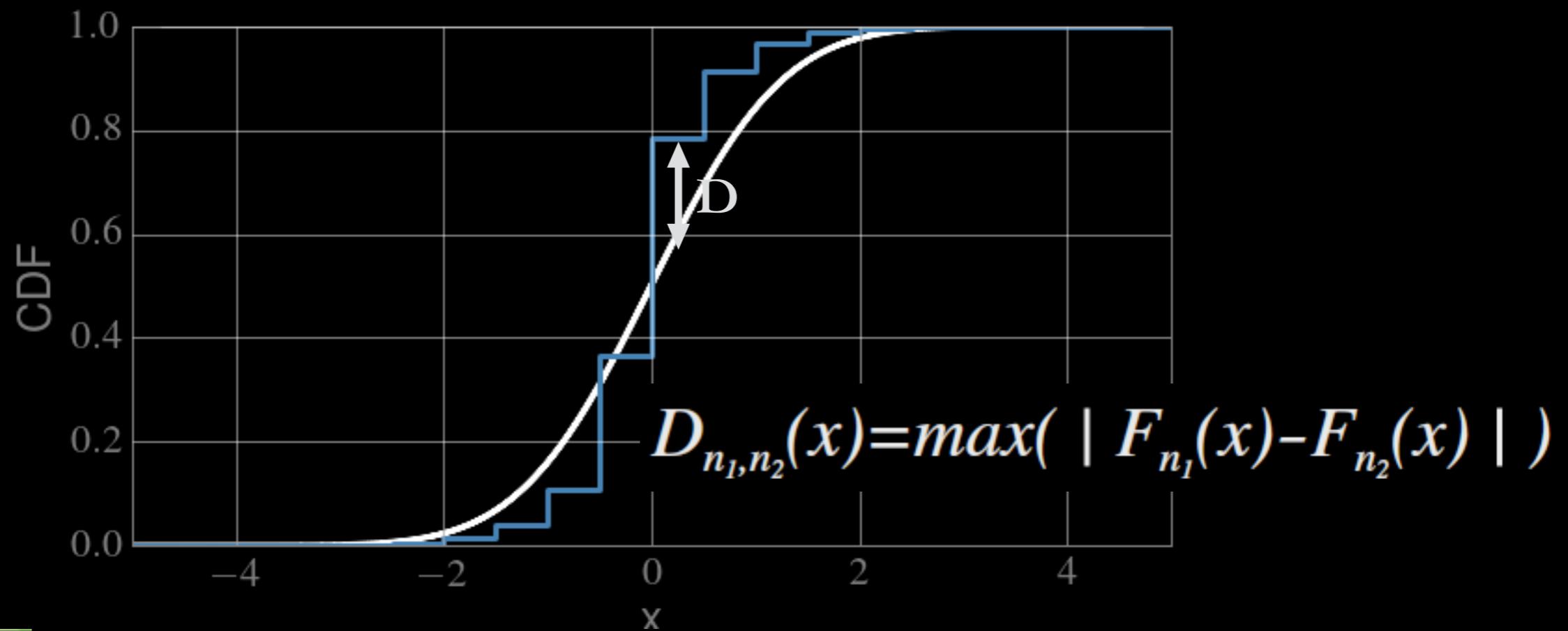
NOTE: it ONLY works in 2D where the Euclidian distance is uniquely defined!



Goodness-of-fit Kolmogorov Smirnoff test:

*null hypothesis*  $H_0$ : the sample does comes from the model distribution

$H_0$  is rejected at level  $\alpha$  if  $\sqrt{n} D_n > K_\alpha$  where  $P(K \leq K_\alpha) = 1 - \alpha$



# Tests Cheat Sheet:

## 2 (+) samples comparison

	metric (statistic)	compare to	
KS	$D_{n_1, n_2}(x) = \max( F_{n_1}(x) - F_{n_2}(x) )$	$c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$	Non parametric 2 samples only
K-sample Anderson-Darling	$ADK = \frac{n-1}{n^2(k-1)} \sum_{i=1}^k \frac{1}{n(i)} \left( \sum_{j=1}^L h_j \frac{(nF_{ij} - n_i H_j)^2}{H_j(n-H_j) - nh_j/4} \right)$	• AK table	Non parametric, N samples
Pearson's	$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$	The interpretation of a correlation coefficient depends on the context and purpose	-1 anticorrelated 0 uncorrelated 1 correlated .
Spearman's	$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}$	t test $t = r \sqrt{\frac{n-2}{1-r^2}}$	ranked data only p-value from t-test, Fisher's transformation +z score, permutation test



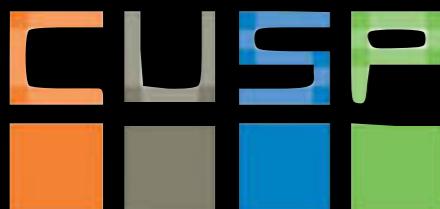
## LAB: Compare Tests for Goodness of fit (synthetic data)

The following are 5 tests that can be used to assess the goodness of fit of a model

- **K-S**
- **Pearson's Chi squared**
- **Anderson-Darling**
- **K-L Divergence**
- **(Likelihood ratio, you do not need to do this yet!)**

Use KS, K-L divergence, and one more test (AD or Chisq) to quantify the difference between a binomial & Gaussian distribution and a Poisson & Gaussian distribution as a function of the parameters of the first distribution (np for binomial,  $\lambda$  for poisson)

For each test plot the relevant parameter (the K-L parameter, Anderson-Darling statistics, p-value for KS, Chi-sq parameter), against the distribution parameter (np,  $\lambda$ )



## Compare Tests for Correlation and Goodness of fit:

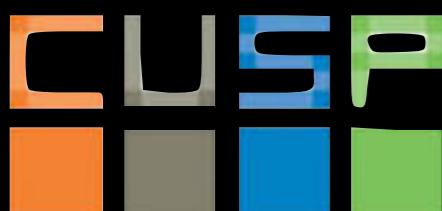
The following are 3 tests that assess correlation between 2 samples:

- Pearson's test e.g. Estimate number of MTA Bus passengers at different hours
- Spearman's test (morning, afternoon, or in time chunks as 7:30-10:30, 10:30-1:30, 1:30-3, 3:6, 6:9, you can do it per bus line, per origin or destination neighborhood...)
- K-S test

The following are 5 tests that can be used to assess the goodness of fit of a model

- K-S
- Pearson's Chi squared
- Anderson-Darling e.g. Estimate number of MTA Bus passengers per bus line within an interval of time: are the passengers randomly distributing on busses.
- K-L Divergence
- Likelihood ratio

In the lab/homework you will 2 out of these tests to assess if 2 samples are related (measure their correlation, or decide if they come from the same parent distributions) and 2 out of the goodness of fit tests to see if a dataset comes from a normal distribution, or from another distribution (where possible) of your choice.



## Compare Tests for Correlation and Goodness of fit:

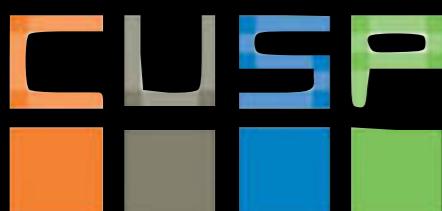
The following are 3 tests that assess correlation between 2 samples:

- Pearson's test e.g. Age distribution of male vs female Citibikes riders. Age
- Spearman's test distribution in different seasons. Age distribution for long/short
- K-S test trips

The following are 5 tests that can be used to assess the goodness of fit of a model

- K-S
  - Pearson's Chi squared
  - Anderson-Darling
  - K-L Divergence
  - Likelihood ratio
- e.g. Estimate Age of riders: could be Gaussian, could be lognormal, power law, bimodal

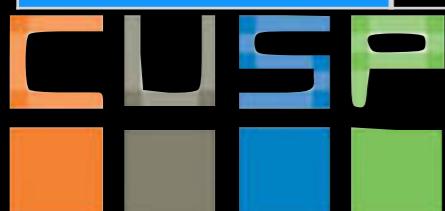
In the lab/homework you will 2 out of these tests to assess if 2 samples are related (measure their correlation, or decide if they come from the same parent distributions) and 2 out of the goodness of fit tests to see if a dataset comes from a normal distribution, or from another distribution (where possible) of your choice.



# Tests Cheat Sheet:

## 2 (+) samples comparison

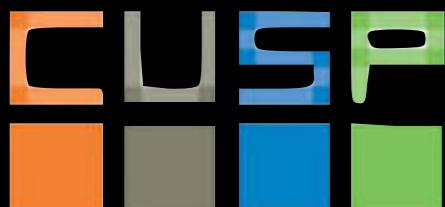
	metric (statistic)	compare to	
KS	$D_{n_1, n_2}(x) = \max( F_{n_1}(x) - F_{n_2}(x) )$	$c(\alpha) \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$	Non parametric 2 samples only
K-sample Anderson-Darling	$ADK = \frac{n-1}{n^2(k-1)} \sum_{i=1}^k \frac{1}{n(i)} \left( \sum_{j=1}^L h_j \frac{(nF_{ij} - n_i H_j)^2}{H_j(n-H_j) - nh_j/4} \right)$	• AK table	Non parametric, N samples
Pearson's	$r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$	The interpretation of a correlation coefficient depends on the context and purpose	-1 anticorrelated 0 uncorrelated 1 correlated .
Spearman's	$\rho = 1 - \frac{6 \sum (x_i - y_i)^2}{n(n^2 - 1)}$	t test $t = r \sqrt{\frac{n-2}{1-r^2}}$	ranked data only  p-value from t-test, Fisher's transformation +z score, permutation test



# Tests Cheat Sheet:

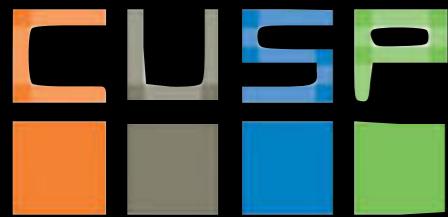
## goodness of fit

	metric (statistic)	compare to	
KS	$D_{n_1, n_2}(x) = \max( F_n(x) - F(x) )$	$\frac{K_\alpha}{\sqrt{n}}$	power in the core only
Pearson's chi square	$\chi^2_{red} = \frac{\chi^2}{df} = \frac{I}{df} \sum \frac{(O-E)^2}{\sigma^2}$	scipy.stats.chisquare(f_obs, f_exp=None, ddof=0, axis=0)[0]	
Anderson-Darling	$A = n \int_{-\infty}^{\infty} \frac{(F_n(x) - F(x))^2}{F(x)(1-F(x))} dF(x)$	scipy.stats.anderson(x, dist='norm')	power in the tails
K-L divergence	$D_{KL} = - \int_x p(x) \log(q(x)) + p(x) \log(p(x))$	scipy.stats.entropy(pk, qk=<not None>)	relates to information entropy
Likelihood ratio	$\frac{L(\text{model 1}   \text{data})}{L(\text{model 2}   \text{data})}$		suitable to bayesian analysis



## assignment: Z-test and chi sq test

- Reproduce the analysis of the Hard to Employ program. Reproduce the results in cell 2 and 10. Follow the notebook in the HW directory (turn in the python notebook in the HW GitHub folder)



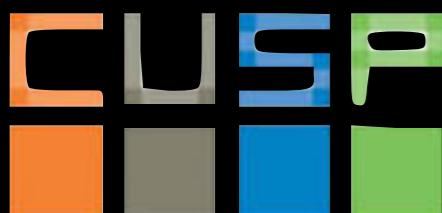
## assignment. Compare Tests for Correlation

The following are 3 tests that assess correlation between 2 samples of citibike data:

- **Pearson's test** (answer: are the 2 samples correlated?)
- **Spearman's test** (answer: are the 2 samples correlated?)
- **K-S test** (answer: are the 2 samples likely to come from the same parent distribution?)

Use:

-trip duration for day vs night. State your result in words in terms of the Null Hypothesis  
-Extra Credit : age of bikers in BK vs Man and assess the correlation/independence of the 2 samples in each case..



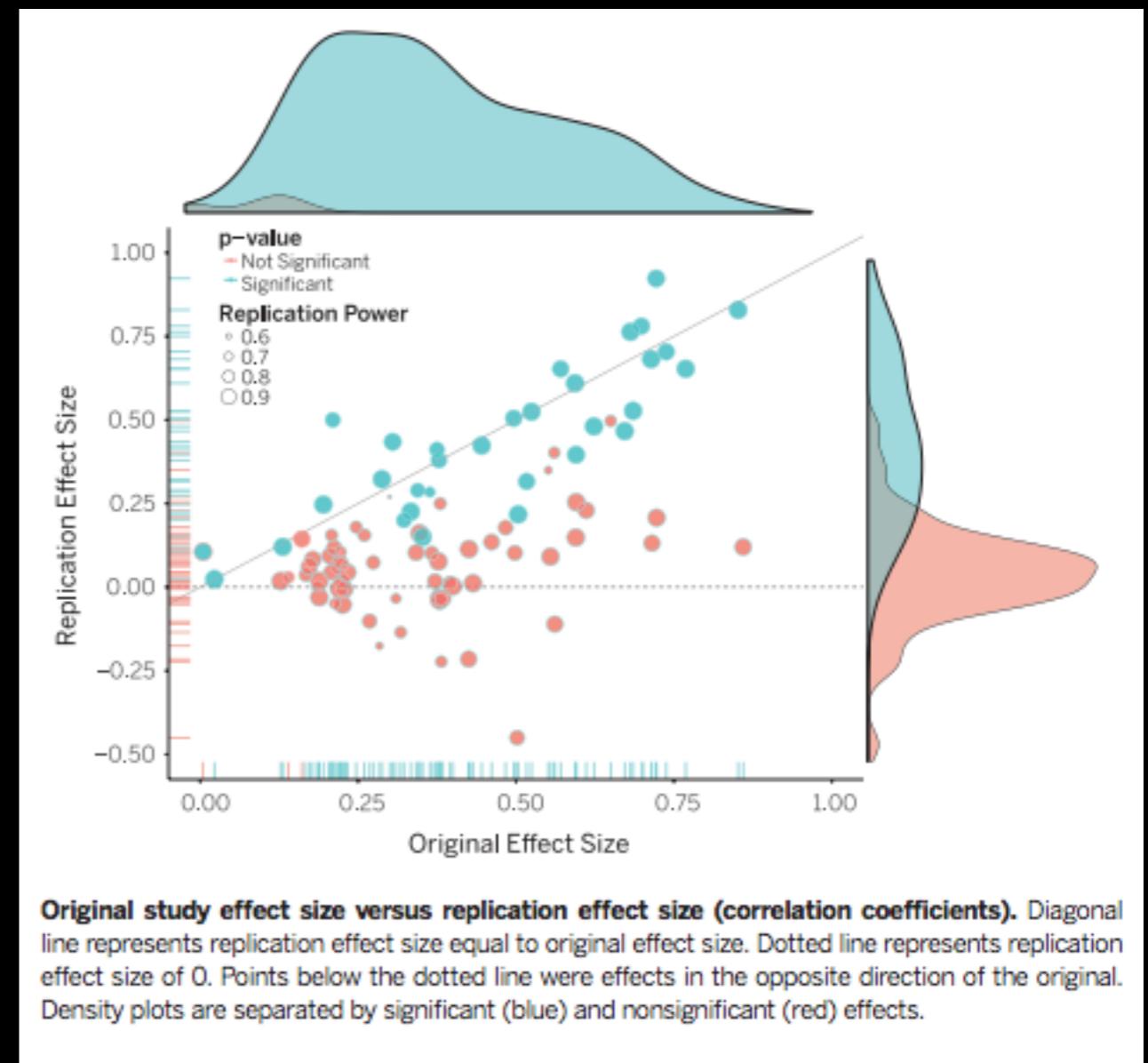
# Homework: READING

## RESEARCH ARTICLE SUMMARY

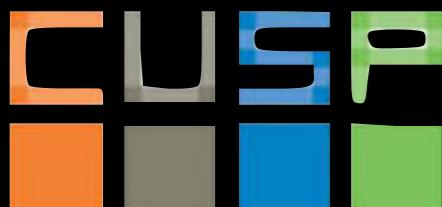
PSYCHOLOGY

### Estimating the reproducibility of psychological science

Open Science Collaboration\*



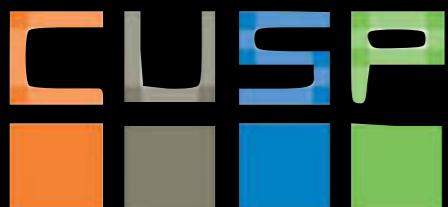
<http://www.sciencemag.org/content/349/6251/aac4716.full.pdf>



IV: Statistical analysis

## MUST KNOWS:

- How to choose (and perform) a statistical test
- Statistical errors
- how to perform Z and chisel test
- PDF vs CDF
- correlation vs causation
- KS test for 2 samples, Pearson's, Spareman



# Resources:

Sarah Boslaugh, Dr. Paul Andrew Watters, 2008

**Statistics in a Nutshell (Chapters 3,4,5)**

[https://books.google.com/books/about/Statistics\\_in\\_a\\_Nutshell.html?id=ZnhgO65Pyl4C](https://books.google.com/books/about/Statistics_in_a_Nutshell.html?id=ZnhgO65Pyl4C)

David M. Lane et al.

**Introduction to Statistics (XVIII)**

[http://onlinestatbook.com/Online\\_Statistics\\_Education.epub](http://onlinestatbook.com/Online_Statistics_Education.epub)

<http://onlinestatbook.com/2/index.html>

Reckova & Irsova

**Publication Bias in Measuring Climate Sensitivity**

IES Working Paper: 14/2015

[http://salserver.org.aalto.fi/vanhat\\_sivut/Opinnot/Mat-2.4108/pdf-files/emet03.pdf](http://salserver.org.aalto.fi/vanhat_sivut/Opinnot/Mat-2.4108/pdf-files/emet03.pdf)

