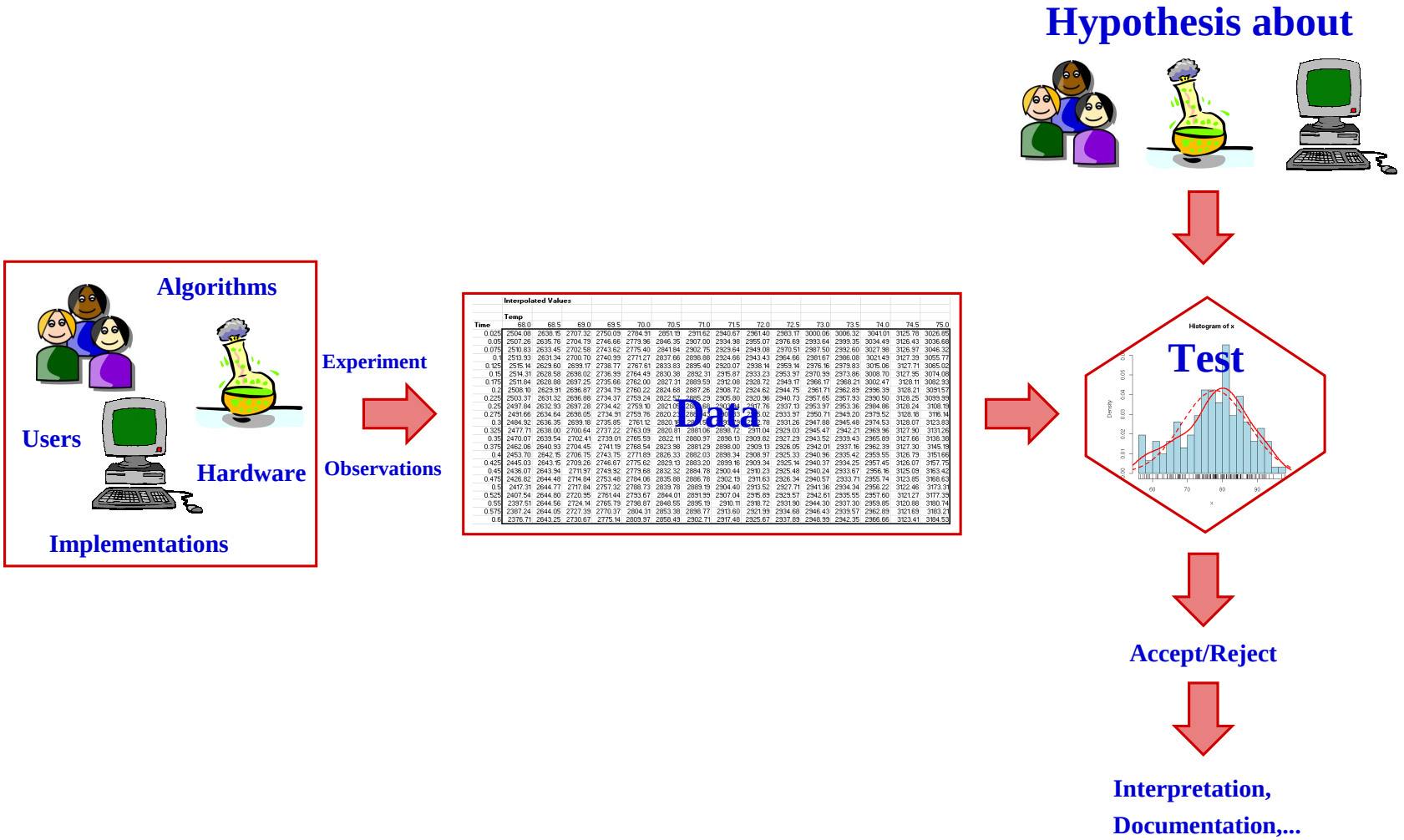
Experiments and Evaluation

Part 2: Statistical Testing II

Manfred Jaeger

Aalborg University

Review



A Hypothesis

- relates to the system/population which was partially observed/measured in controlled experiments or by passive data collection.
- is tested on the basis of the available data

System/Population	Data	Hypothesis
Coin	Result of N flips	Is the coin fair?
Implementation	Execution time on N random inputs	Does the implementation satisfy a given performance requirement?
Patients with disease X	Health status for N patients treated with novel drug Y	Is drug Y effective for treating patients with X?
Students in two study programs BLA and CRU	Exam results from N students in course X	Do BLA students perform as well as CRU students in X?

Basic principle:

- Model data by random variables that follow a certain probability distribution
  (Coin: H/T-valued "Bernoulli" variable with probability p for H)
- Express hypothesis about the population as a hypothesis over the distribution of the random variables
  (Coin: $p = 0.5$)
- *reject* the hypothesis if the observed data would be very unlikely when the hypothesis was true.
  (Coin: check probability of observed number of H if $p = 0.5$ was true)

One Sample t Test (Review)

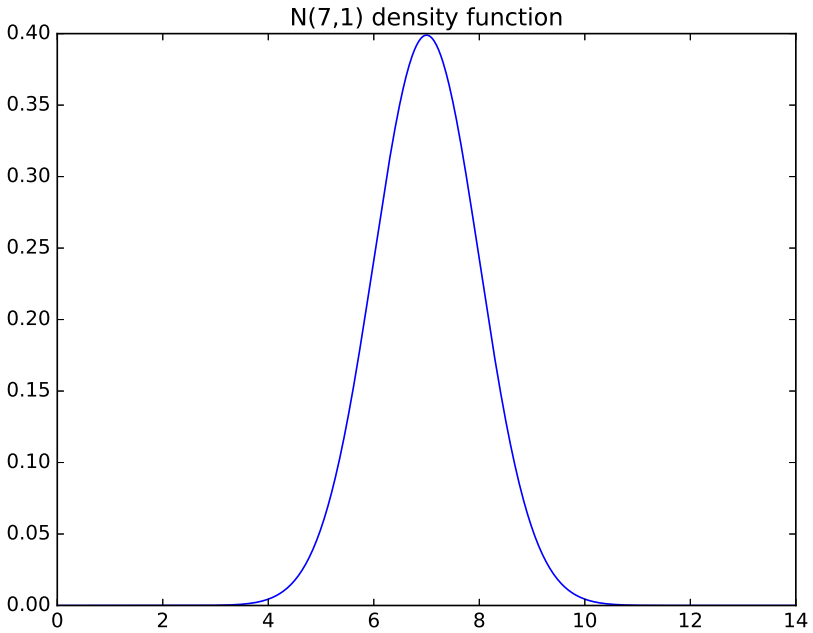
Observed runtimes of implementation on test cases:

Trials	Times (in ms)
1	4.4
5	6.41, 3.36, 4.71, 4.07, 6.41
100	4.17, 7.05, 4.38, 5.44, 5.69, 5.45, 5.76, ...
1000	4.76, 4.73, 6.42, 3.30, 6.44, 3.78, 7.62, 4.99, ...

Hypothesis: the average runtime is greater or equal 7ms (e.g.: customer defined performance requirement).

Binomial: each sample generated by a 0/1 valued Bernoulli variable characterized by parameter p

Assume that runtimes follow a **Gaussian distribution** with some *mean* μ and *standard deviation* σ



Plot for $\mu = 7.0, \sigma = 1.0$

Binomial: just count the number of heads in a sample

Compute from the raw data the relevant *test statistic*. Here from a sample of size N

$$x_1, x_2, \ldots, x_N$$

compute:

Sample Average : $\bar{x} := \frac{1}{N} \sum_{i=1}^{N} x_i$

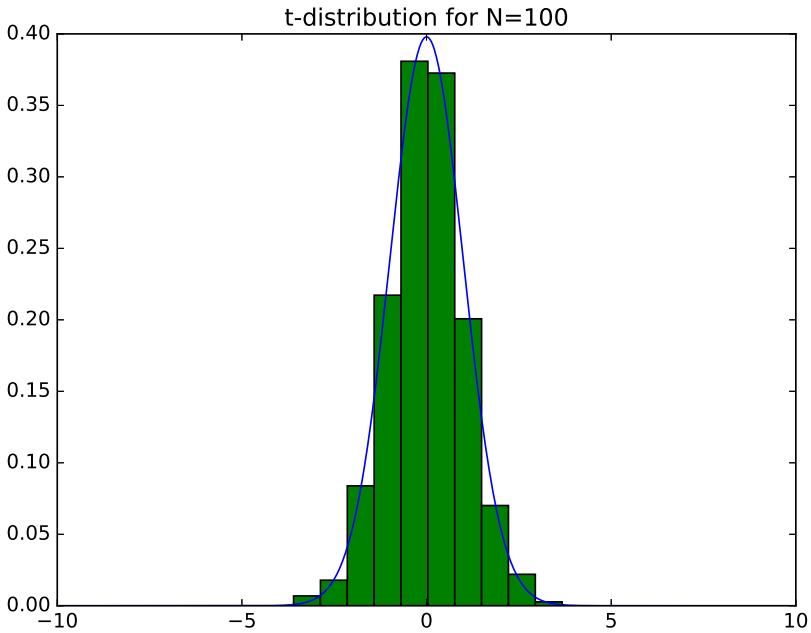
Sample Variance: $\bar{s} := \frac{1}{N} \sum_{i=1}^{N} (\bar{x} - x_i)^2$

t-Statistic: $t := \sqrt{\frac{N}{\bar{s}}}(\bar{x} - 7.0)$ (contains the hypothesized mean value of 7.0!)

Trials	$\bar{x}$	$\bar{s}$	t
1	4.4	0.0	$-\infty$
5	4.99	1.51	-3.644
100	5.44	0.75	-17.903
1000	5.54	1.0	-45.93

Binomial: if every sample point comes from a Bernoulli distribution with parameter p then #successes has a Binomial distribution with parameters N, p

If every sample point comes from a normal distribution with mean μ, then the t-statistic has a **t-distribution with N-1 degrees of freedom**



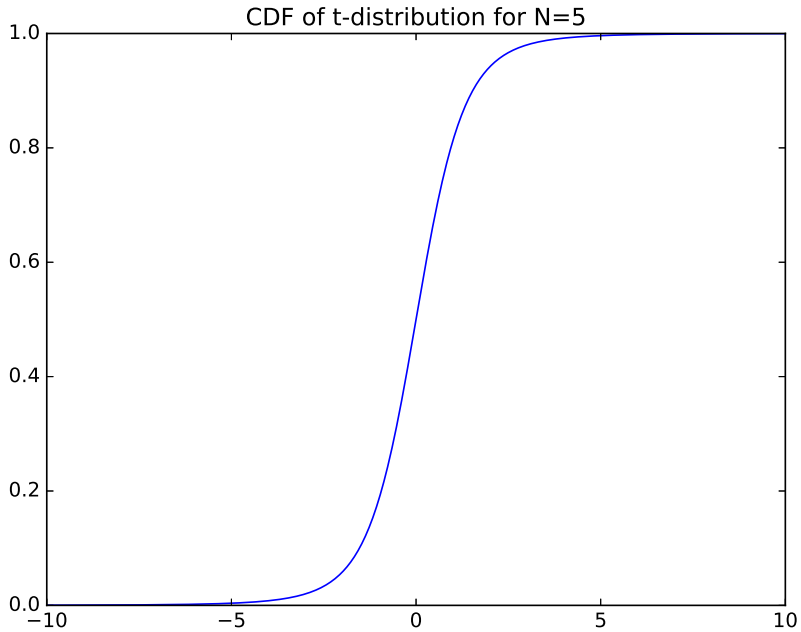
Blue line: t-distribution with 99 degrees of freedom

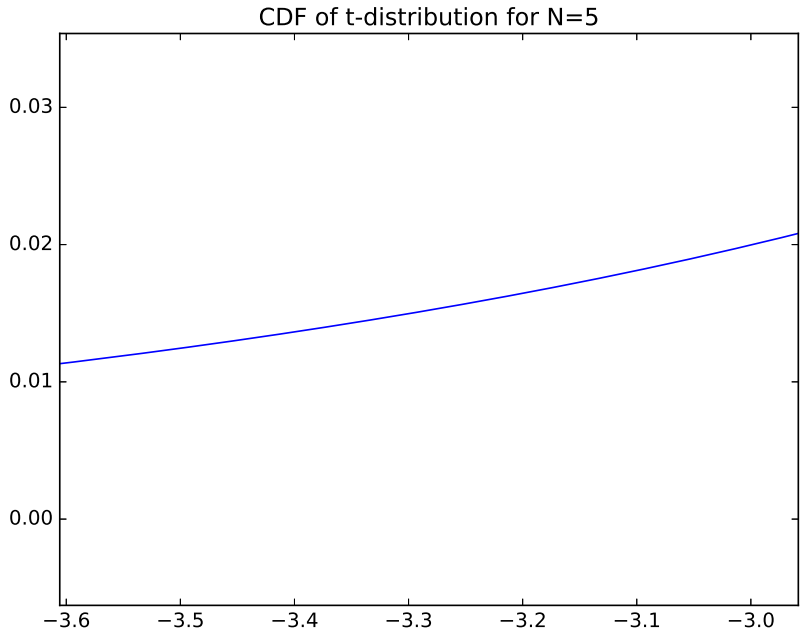
Green histogram: values of t-statistics computed for 1000 datasets of size 100 (datapoints sampled from normal distribution with mean 7.0).

Binomal: evaluating the CDF of the binomial distribution under the (null) hypothesis at the observed # successes

Evaluate the CDF of t-distribution at the observed value of the t-statistic (computed for the (null) hypothesis).

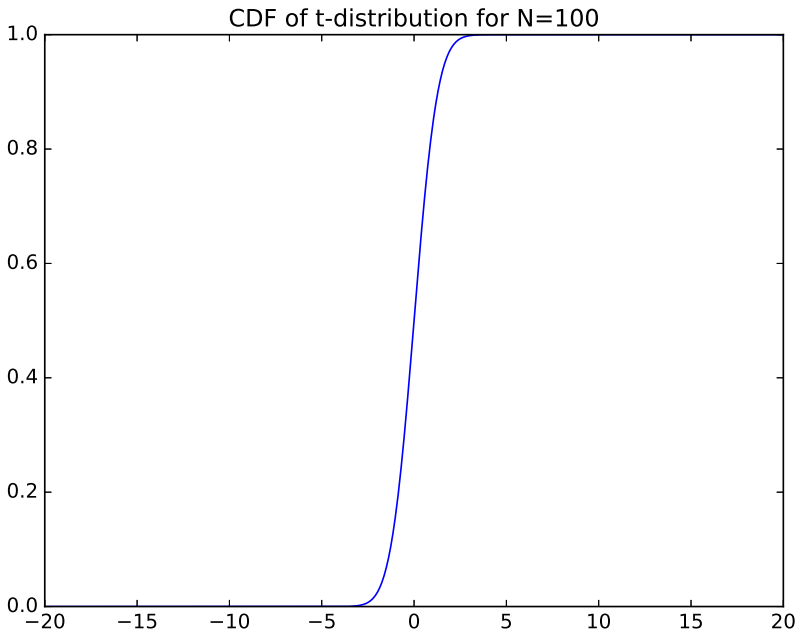
Our data at $N = 5$, $t = -3.644$

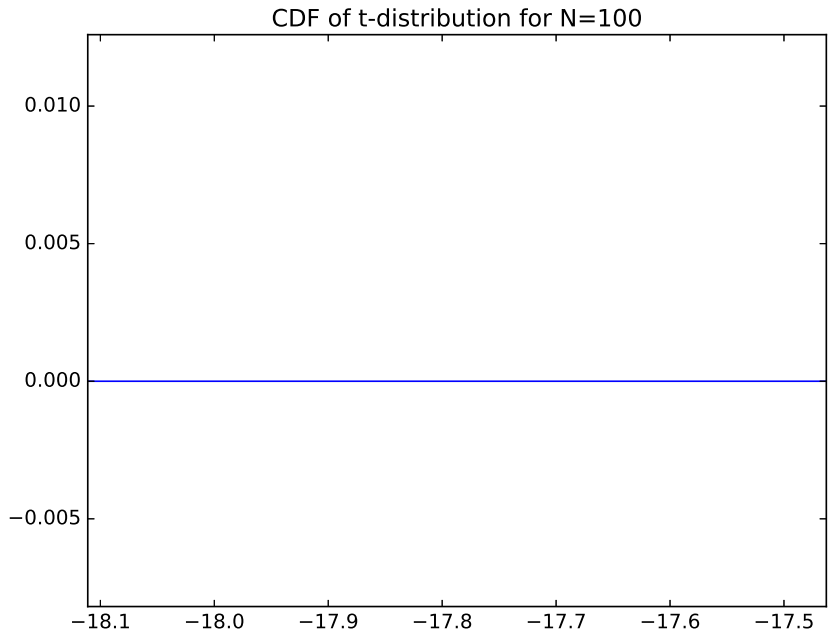




☞ p-value (one-sided) is ≈ 0.011.

Our data at $N = 100$, $t = -17.903$

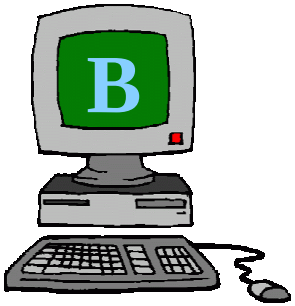




☞ p-value (one-sided) is ≈ 0.0.

Two Sample Tests

Setup A



Setup B

Measurements of performance on test cases

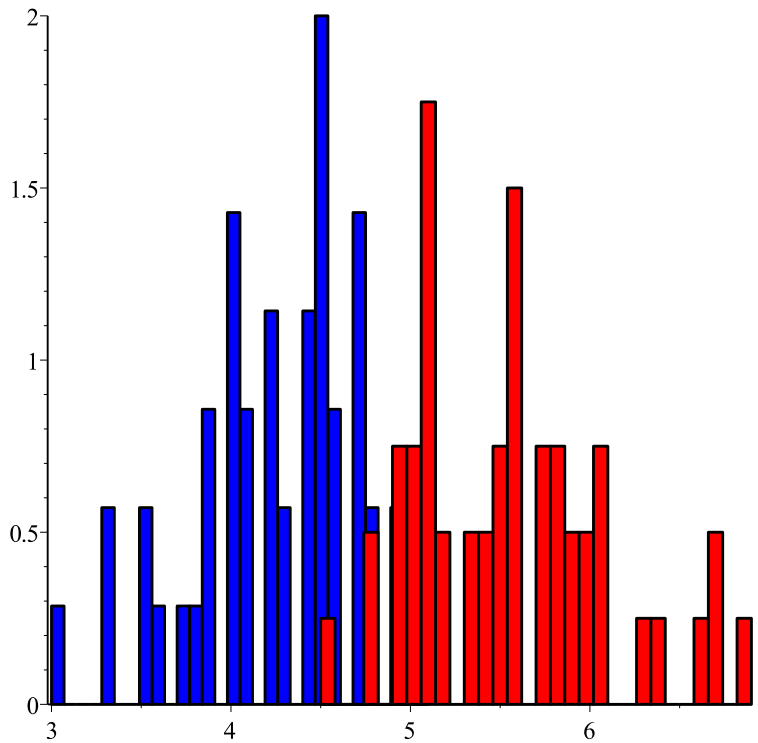
A					
1	0.55	6	1.2	11	0.52
2	0.01	7	0.91	12	0.88
3	0.87	8	0.02	13	0.45
4	1.3	9	1.01	14	0.03
5	0.54	10	0.76	15	0.65

B					
1	0.88	6	0.67	11	0.52
2	0.21	7	0.63	12	0.73
3	0.54	8	0.87	13	1.05
4	1.22	9	0.41	14	0.87
5	1.54	10	1.76	15	1.56

Which performs better, A or B (on the “population” of all possible inputs/use cases)?

- Measurements in test cases are **random samples** from the measurements that would be obtained in all possible cases.

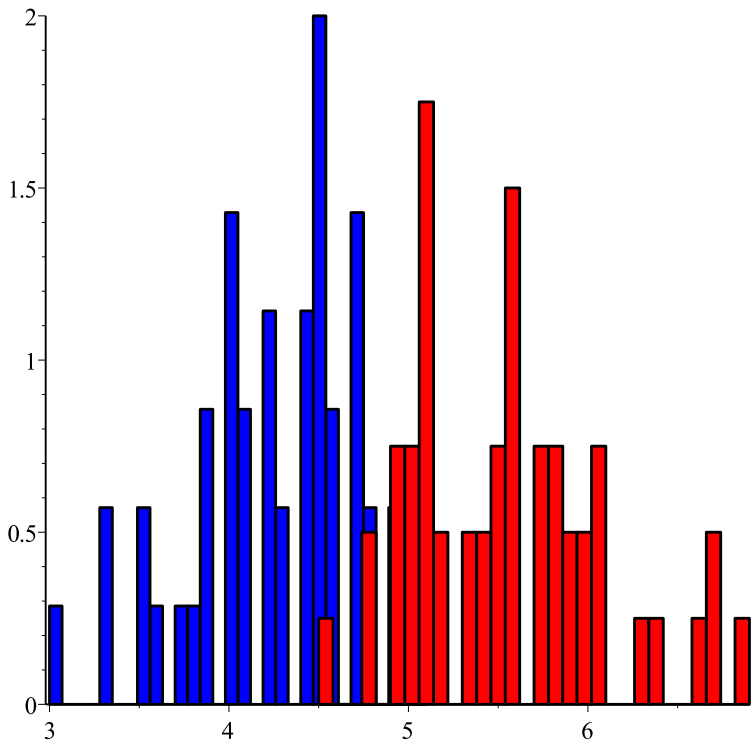
Measurements plotted as histograms:



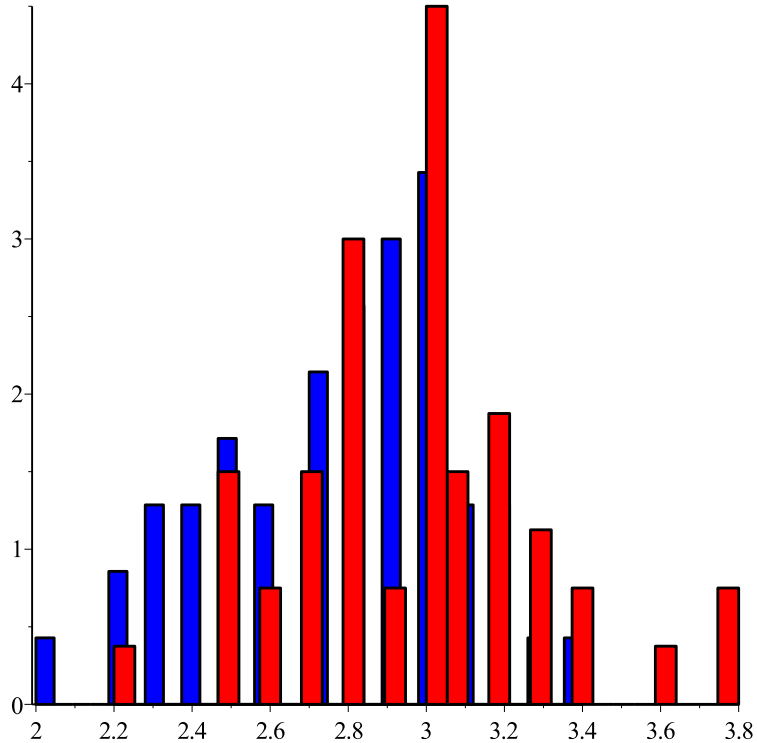
Which performs better, red or blue setup?

- Measurements in test cases are **random samples** from the measurements that would be obtained in all possible cases.

Measurements plotted as histograms:

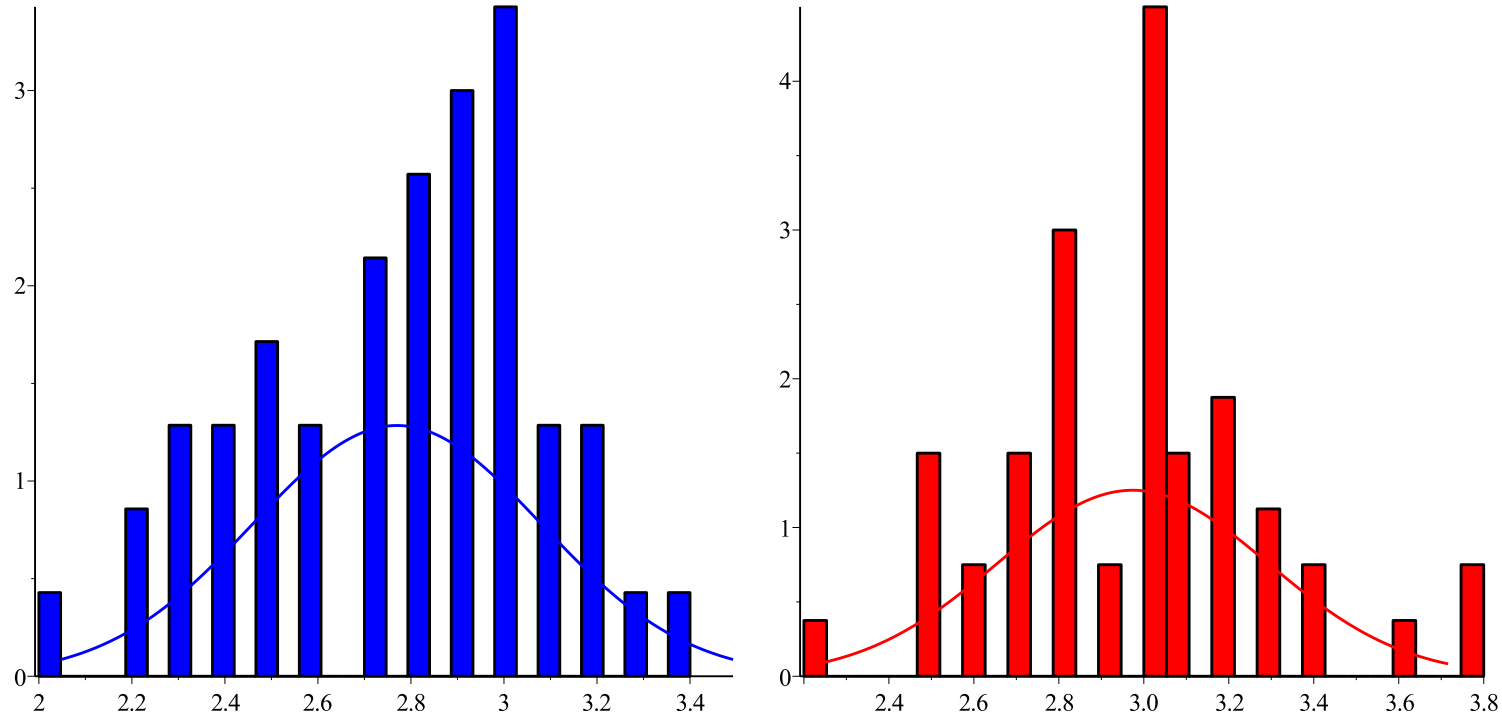


Which performs better, red or blue setup?



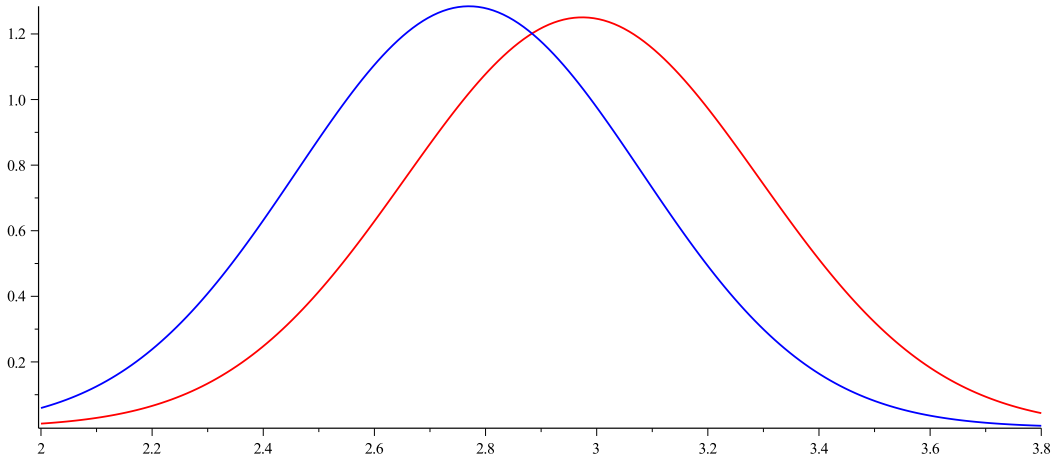
Which performs better, red or blue setup?

Assumption: the measurement values obtained in each setup follow a certain probability distribution.



Histograms with Gaussian distributions that best fit the data.

To compare the two setups, we need to compare the two distributions over performance measurements:



If these were the actual distributions of performance measures in future test cases, then B would be better than A.

Given: Measurements from setups A and B

Assumption: Measurements in setup A follow a Gaussian distribution with mean μ_A and standard deviation σ_A

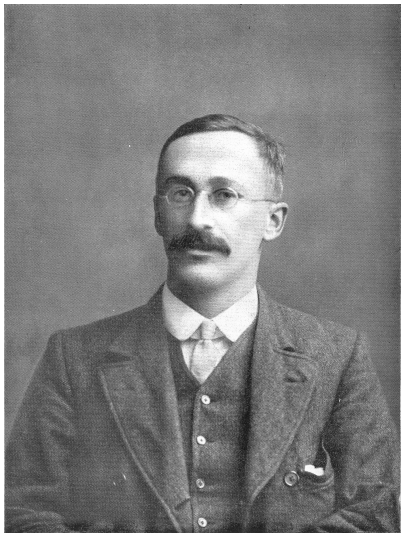
Assumption: Measurements in setup B follow a Gaussian distribution with mean μ_B and standard deviation σ_B

Hypothesis: $\mu_A = \mu_B$ (no difference).

- We will want to disprove/reject the hypothesis!
- The hypothesis concerns the *average* performance of the setups, not the *worst-case* performance.

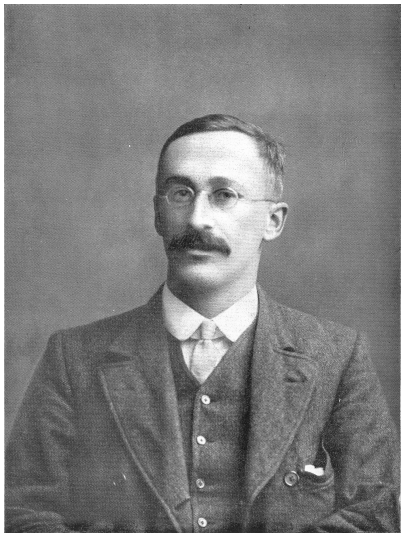
Two Sample *t*-Test

William Sealy Gosset ("Student") 1876-1937



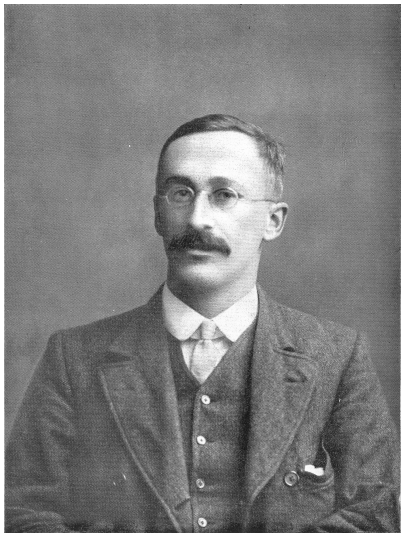
- Published under pseudonym Student

William Sealy Gosset ("Student") 1876-1937



- Published under pseudonym Student
- Worked as chemist/brewer at Guiness Brewery from 1899
- Collaborated with prominent statisticians of his time

William Sealy Gosset ("Student") 1876-1937



- Published under pseudonym Student
- Worked as chemist/brewer at Guiness Brewery from 1899
- Collaborated with prominent statisticians of his time
- Investigated best growing conditions for barley

Assume ...

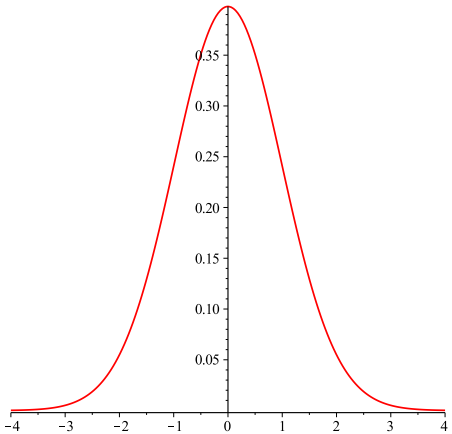
- $x_1, \ldots, x_k$ are measurements drawn from a normal distributions with mean μ_A and standard deviation σ_A
- $y_1, \ldots, y_l$ are measurements drawn from a normal distributions with mean μ_B and standard deviation σ_B
- $\sigma_A = \sigma_B$!

Let $\bar{x} = \frac{1}{k} \sum x_i$, $\bar{y} = \frac{1}{l} \sum y_j$. Then, if $\mu_A = \mu_B$, the *test statistic*

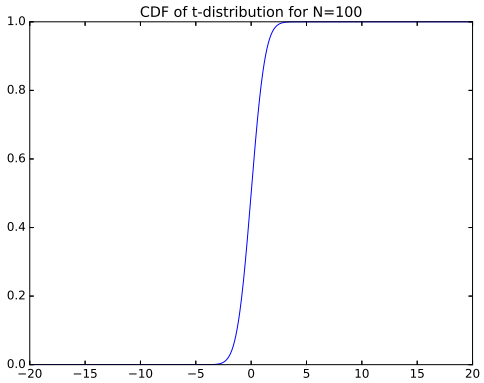
$$t(\bar{x}, \bar{y}, k, l) := \frac{(\bar{x} - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 + \sum (y_j - \bar{y})^2}} \sqrt{\frac{k + l - 2}{1/k + 1/l}}$$

has a *(Student's) t-distribution with $k + l - 2$ degrees of freedom*.

t-distribution with 100 degrees of freedom:



Density function



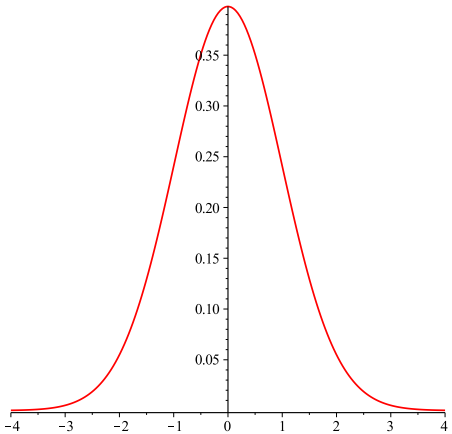
Cumulative Distribution Function (CDF)

We reject the null-hypothesis $\mu_A = \mu_B$ *at the significance level 0.05* if

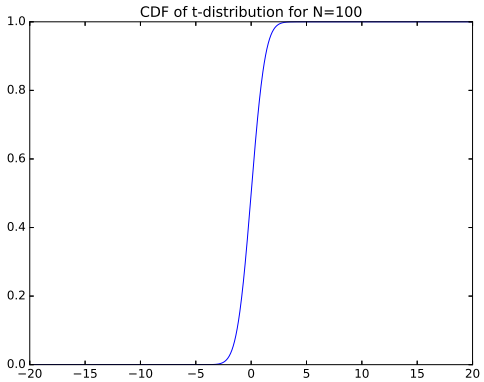
$$t(\bar{x}, \bar{y}, k, l) > CDF^{-1}(1 - 0.05/2) \quad \text{or} \quad t(\bar{x}, \bar{y}, k, l) < CDF^{-1}(0.05/2)$$

(two-sided test)

t-distribution with 100 degrees of freedom:



Density function



Cumulative Distribution Function (CDF)

We reject the null-hypothesis $\mu_A = \mu_B$ *at the significance level 0.05* if

$$t(\bar{x}, \bar{y}, k, l) > CDF^{-1}(1 - 0.05/2) \quad \text{or} \quad t(\bar{x}, \bar{y}, k, l) < CDF^{-1}(0.05/2)$$

(two-sided test)

- I.e.: reject if under the null-hypothesis, the probability to obtain for the test statistic an absolute value as large as $|\, t(\bar{x}, \bar{y}, k, l)\,|$ is less than 0.05.

Design

- Make assumptions on the type of data and its distribution
- Define the type of null-hypothesis to be tested
- Define a test statistic
- Determine the distribution of the test statistic under the null hypothesis

- Provide formulas or tables for the computation of the CDF

Design

- Make assumptions on the type of data and its distribution
- Define the type of null-hypothesis to be tested
- Define a test statistic
- Determine the distribution of the test statistic under the null hypothesis

- Provide formulas or tables for the computation of the CDF

Use

- Check whether your data satisfies the assumptions of the test (maybe approximately)
- Select a level of significance α
- Compute the test statistic t
- Reject the null hypothesis if

$$t > CDF^{-1}(1 - \alpha/2)$$

or

$$t < CDF^{-1}(\alpha/2)$$

Two-sided test:

- Symmetric hypothesis: $\mu_A = \mu_B$
- Very large or very small values of the test statistic provide evidence against the hypothesis
- Reject the hypothesis if

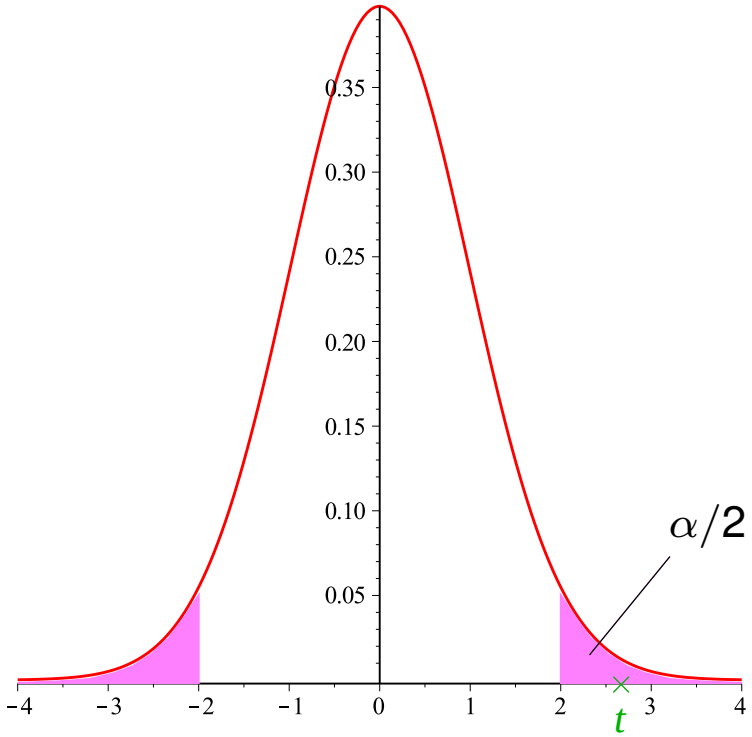
$$t < CDF^{-1}(\alpha/2) \text{ or } t > CDF^{-1}(1 - \alpha/2)$$

One-sided test:

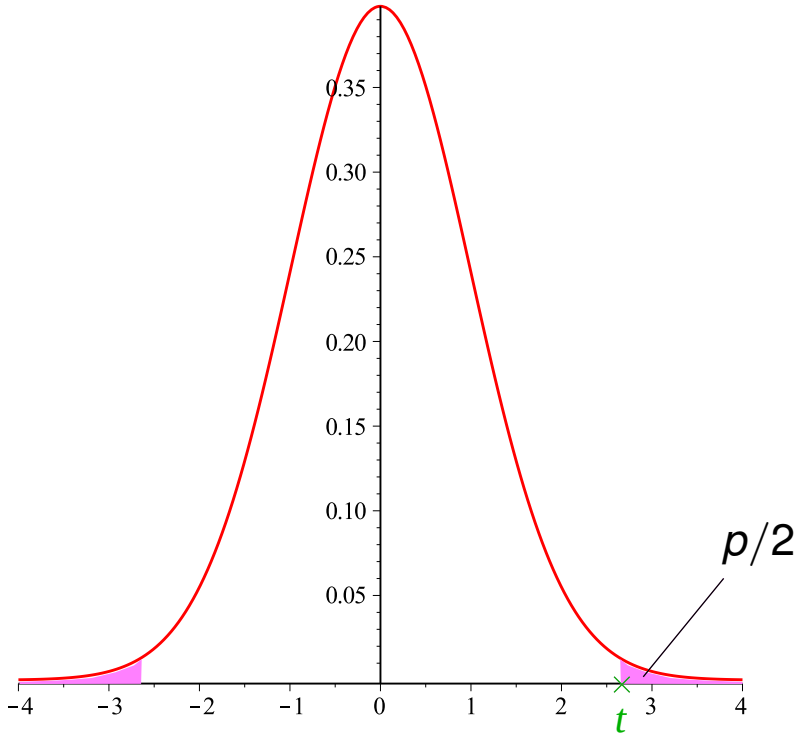
- Asymmetric hypothesis: $\mu_A \leq \mu_B$
- Larger values of the test statistic provide evidence against the hypothesis
- Reject the hypothesis if

$$t > CDF^{-1}(1 - \alpha)$$

p-value: minimal level of significance at which hypothesis would be rejected.



Reject if test statitistic *t* is inside the *critical region* defined by the significance level α.



Test statistic *t* defines *p*-value.

The coin, again:

Trials	# Heads	# Tails	*p-value* (two-sided)
500	261	239	0.3476

- Based on the data, we will not reject the hypothesis that the coin is fair: $p = 0.5$
- We will still not conclude that the coin has to be fair: could very well be $p = 0.51$ (that hypothesis would also not be rejected).

Wilcoxon Test

Frank Wilcoxon 1892-1965

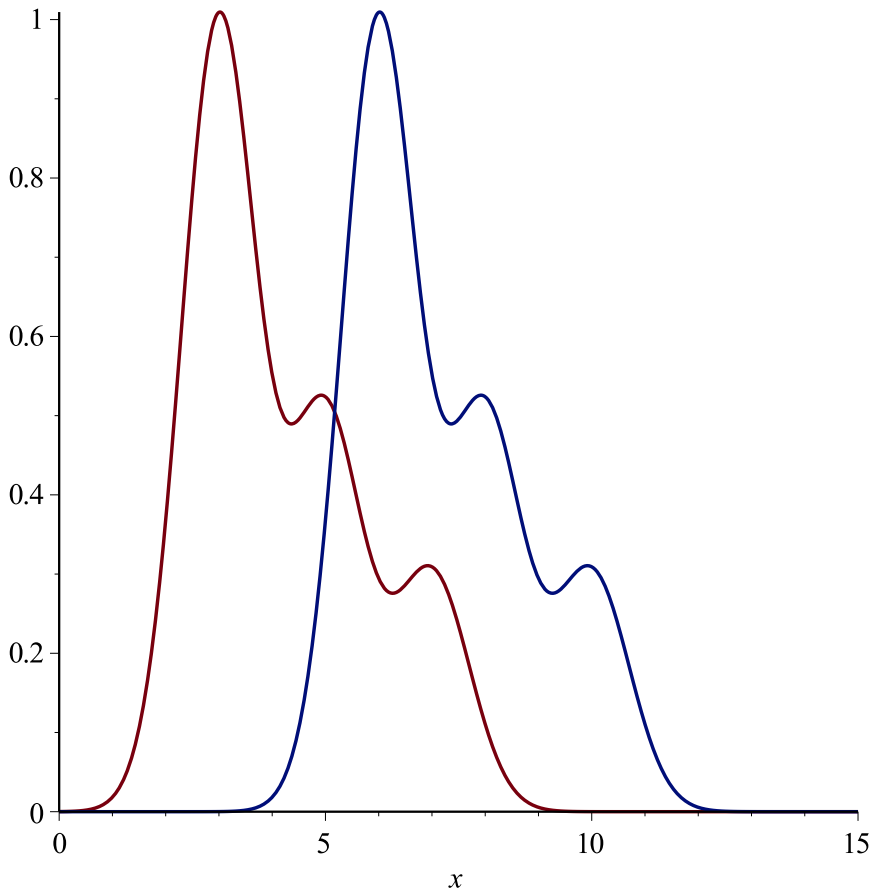
- Education in Chemistry
- Joined American Cyanamid Company in 1945
- Published his main statistical work 1945-1950
- Mostly working on pesticides

Data consists of **paired samples** if for both setups we have measurements on the same test cases:

	A	B
1	0.57	0.73
2	1.54	0.97
3	0.76	0.63
4	0.02	0.07
5	0.43	0.77
6	0.87	0.63

- Typical example: performance of A and B on the same test input
- Not paired sample: test of webpage design shown to different users
- Note: there also exists a version of the t-test for paired samples

The performance measurements for *A* and *B* have an arbitrary distribution of *identical form*:



Null hypothesis: $\mu_A \leq \mu_B$ ("A is not better than B")

	A	B		
1	0.57	0.73		
2	1.54	0.97		
3	0.76	0.63		
4	0.02	0.07		
5	0.43	0.77		
6	0.87	0.63		

	A	B	A-B	
1	0.57	0.73	-0.16	
2	1.54	0.97	0.57	
3	0.76	0.63	0.13	
4	0.02	0.07	-0.05	
5	0.43	0.77	-0.34	
6	0.87	0.63	0.24	

- compute $A - B$

	A	B	A-B	Rank
1	0.57	0.73	-0.16	3
2	1.54	0.97	0.57	6
3	0.76	0.63	0.13	2
4	0.02	0.07	-0.05	1
5	0.43	0.77	-0.34	5
6	0.87	0.63	0.24	4

- compute $A - B$
- sort by increasing absolute value of $A - B$; assign each case its rank in this ordering

	A	B	A-B	Rank
1	0.57	0.73	-0.16	3
2	1.54	0.97	0.57	6
3	0.76	0.63	0.13	2
4	0.02	0.07	-0.05	1
5	0.43	0.77	-0.34	5
6	0.87	0.63	0.24	4

- compute $A - B$
- sort by increasing absolute value of $A - B$; assign each case its rank in this ordering
- compute the sum of ranks for the negative $A - B$ values:

$$3 + 1 + 5 = 9$$

	A	B	A-B	Rank
1	0.57	0.73	-0.16	3
2	1.54	0.97	0.57	6
3	0.76	0.63	0.13	2
4	0.02	0.07	-0.05	1
5	0.43	0.77	-0.34	5
6	0.87	0.63	0.24	4

- compute $A - B$
- sort by increasing absolute value of $A - B$; assign each case its rank in this ordering
- compute the sum of ranks for the negative $A - B$ values:

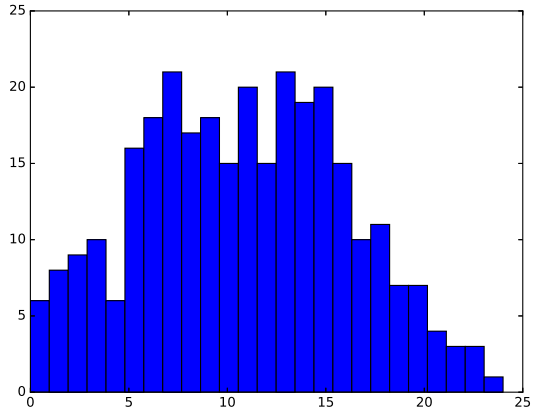
$$3 + 1 + 5 = 9$$

- reject the hypothesis $\mu_A \leq \mu_b$ at the chosen level of significance α if the negative rank sum is less than a threshold value (table lookup! threshold depends on α and number of samples)

Statistical Assumptions

Can we test a hypothesis without making assumptions on the form of the distribution?

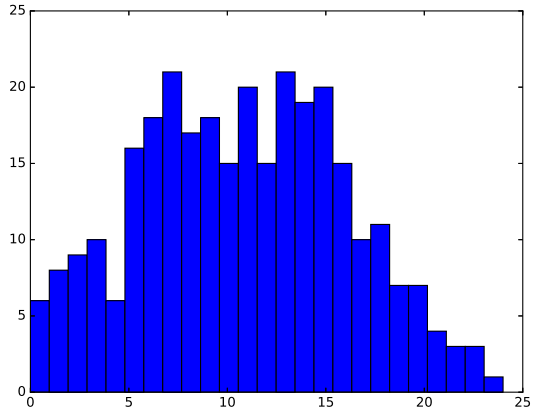
Example: Data from 300 measurements:



Can we reject the hypothesis, that the average performance measure is ≥ 25 without assuming e.g. a normal distribution?

Can we test a hypothesis without making assumptions on the form of the distribution?

Example: Data from 300 measurements:



Can we reject the hypothesis, that the average performance measure is ≥ 25 without assuming e.g. a normal distribution?

No: suppose there is a probability of 1/500 that we get a measurement value of 20000. Then:

- the observed data is not very unlikely under this distribution: we just have not seen the measurement value 20000 in our sample of 300
- the average performance is at least 20000/500 = 40

We have focused on hypotheses about the mean value of distribution(s).

One can also testing other (qualitative) hypotheses:

- Does the distribution follow a normal distribution?
- Are two different measurements correlated?
- Are two different measurements independent (e.g., are the exam results independent of the weather)?

The strategy remains the same:

- design a probabilistic model for the data
- express the hypothesis as a property of the probabilistic model
- reject the hypothesis if the data is very unlikely under the hypothesis