Experiments and Evaluation

Part 1: Statistical Testing

Manfred Jaeger

Aalborg University

Introduction

Schedule

- Three lectures with exercises
- Exercises in two groups (morning/afternoon)
  - 10:15-12:00: SW
  - 12:30-14:15: DAT

Hand-in Exercises

- Started in exercise sessions following lectures (in small groups)
- **Individual Solutions** finished in last exercise session and extra time
  - Individual solution means: you must use your own words to describe your solution.
- Solutions for both exercises in one pdf document uploaded by Friday, March 29, 24:00.
- Put your name on the solution sheet!
- Pass/fail evaluation of solutions

Survey Article [mostly relevant for second/third part]

R. L. Rardin and R. Uzsoy:
Experimental Evaluation of Heuristic Optimization Algorithms: A Tutorial. Journal of Heuristics, 7 (2001)

Book [further reading – when needed]

A. B. Downey: Think Stats – Probability and Statistics for Programmers. Green Tea Press, 2011. Available online: `http://greenteapress.com/thinkstats/`

Online Book

`http://onlinestatbook.com/index.html`

Wikipedia

- Statistical hypothesis testing
- Student's t-test
- Wilcoxon signed-rank test

Claims:

- My algorithm returns a correct (optimal) solution on more than 90% of its inputs
- My program runs on average in less than 10s
- My algorithm/implementation is better than your algorithm/implementation
- The users of my web-site are happier than the users of your web-site
- ...

How do we determine the validity of such claims, based on experimental data?

Often: Final section in a scientific article.

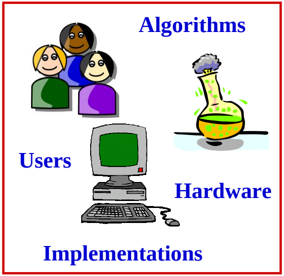
Goal for these lectures:

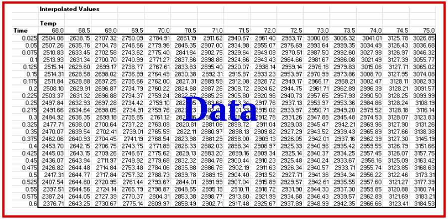
- Understanding the basic principles of (statistical) empirical evaluations
- Being able to perform basic tests on your own data
- Being able to identify possible strengths and weaknesses of a presented evaluation

Program

- First lecture: the more technical/analytical core
- Second lecture: technical/analytical core, experimental process
- Third lecture: experimental process (data collection)

Empirical Evaluations are based on experimental or observational data:





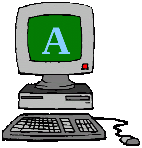
	Interpolated Values														
	Temp														
Time	68.0	68.5	69.0	69.5	70.0	70.5	71.0	71.5	72.0	72.5	73.0	73.5	74.0	74.5	75.0
0.025	2504.08	2638.15	2707.32	2750.09	2784.91	2851.19	2911.62	2940.67	2961.40	2983.17	3000.06	3006.32	3041.01	3125.78	3026.85
0.05	2507.26	2635.76	2704.79	2746.66	2779.96	2846.35	2907.00	2934.98	2955.07	2976.69	2993.64	2999.35	3034.49	3126.43	3036.68
0.075	2510.83	2633.45	2702.58	2743.62	2775.40	2841.84	2902.75	2929.64	2949.08	2970.51	2987.50	2992.60	3027.98	3126.97	3046.32
0.1	2513.93	2631.34	2700.70	2740.99	2771.27	2837.66	2898.88	2924.66	2943.43	2964.66	2981.67	2986.08	3021.49	3127.39	3055.77
0.125	2515.14	2629.60	2699.17	2738.77	2767.61	2833.83	2895.40	2920.07	2938.14	2959.14	2976.16	2979.83	3015.06	3127.71	3065.02
0.15	2514.31	2628.58	2698.02	2736.99	2764.49	2830.38	2892.31	2915.87	2933.23	2953.97	2970.99	2973.86	3008.70	3127.95	3074.08
0.175	2511.84	2628.88	2697.25	2735.66	2762.00	2827.31	2889.59	2912.08	2928.72	2949.17	2966.17	2968.21	3002.47	3128.11	3082.93
0.2	2508.10	2629.91	2696.87	2734.79	2760.22	2824.68	2887.26	2908.72	2924.62	2944.75	2961.71	2962.89	2996.39	3128.21	3091.57
0.225	2503.37	2631.32	2696.88	2734.37	2759.24	2822.57	2885.29	2905.80	2920.96	2940.73	2957.65	2957.93	2990.50	3128.25	3099.99
0.25	2497.84	2632.93	2697.28	2734.42	2759.10	[illegible]	[illegible]	[illegible]	2917.76	2937.13	2953.97	2953.36	2984.86	3128.24	3108.19
0.275	2491.66	2634.64	2698.05	2734.91	2759.76	[illegible]	[illegible]	[illegible]	2915.02	2933.97	2950.71	2949.20	2979.52	3128.18	3116.14
0.3	2484.92	2636.35	2699.18	2735.85	2761.12	[illegible]	[illegible]	[illegible]	2912.78	2931.26	2947.88	2945.48	2974.53	3128.07	3123.83
0.325	2477.71	2638.00	2700.64	2737.22	2763.09	2820.81	2881.06	2898.72	2911.04	2929.03	2945.47	2942.21	2969.96	3127.90	3131.26
0.35	2470.07	2639.54	2702.41	2739.01	2765.59	2822.11	2880.97	2898.13	2909.82	2927.29	2943.52	2939.43	2965.89	3127.66	3138.38
0.375	2462.06	2640.93	2704.45	2741.19	2768.54	2823.98	2881.29	2898.00	2909.13	2926.05	2942.01	2937.16	2962.39	3127.30	3145.19
0.4	2453.70	2642.15	2706.75	2743.75	2771.89	2826.33	2882.03	2898.34	2908.97	2925.33	2940.96	2935.42	2959.55	3126.79	3151.66
0.425	2445.03	2643.15	2709.26	2746.67	2775.62	2829.13	2883.20	2899.16	2909.34	2925.14	2940.37	2934.25	2957.45	3126.07	3157.75
0.45	2436.07	2643.94	2711.97	2749.92	2779.68	2832.32	2884.78	2900.44	2910.23	2925.48	2940.24	2933.67	2956.16	3125.09	3163.42
0.475	2426.82	2644.48	2714.84	2753.48	2784.06	2835.88	2886.78	2902.19	2911.63	2926.34	2940.57	2933.71	2955.74	3123.85	3168.63
0.5	2417.31	2644.77	2717.84	2757.32	2788.73	2839.78	2889.19	2904.40	2913.52	2927.71	2941.36	2934.34	2956.22	3122.46	3173.31
0.525	2407.54	2644.80	2720.95	2761.44	2793.67	2844.01	2891.99	2907.04	2915.89	2929.57	2942.61	2935.55	2957.60	3121.27	3177.39
0.55	2397.51	2644.56	2724.14	2765.79	2798.87	2848.55	2895.19	2910.11	2918.72	2931.90	2944.30	2937.30	2959.85	3120.88	3180.74
0.575	2387.24	2644.05	2727.39	2770.37	2804.31	2853.38	2898.77	2913.60	2921.99	2934.68	2946.43	2939.57	2962.89	3121.69	3183.21
0.6	2376.71	2643.25	2730.67	2775.14	2809.97	2858.49	2902.71	2917.48	2925.67	2937.89	2948.99	2942.35	2966.66	3123.41	3184.53

My System

Measurements of performance on test cases

My System					
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

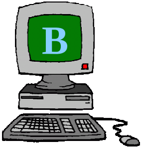
☞ One sample tests



Setup A

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



Setup B

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

☞ Two sample tests

Setups can be ... different algorithms, different implementations of the same algorithm, different hardware platforms, different user interfaces, . . .

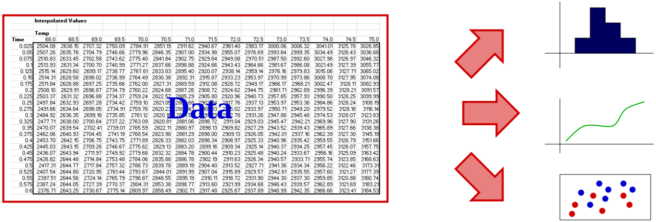
Test cases can be ... multiple runs on different inputs, interactions by different users with a web site, . . .

Measurements can be ... time and/or space consumption, user satisfaction, quality of solution, . . .

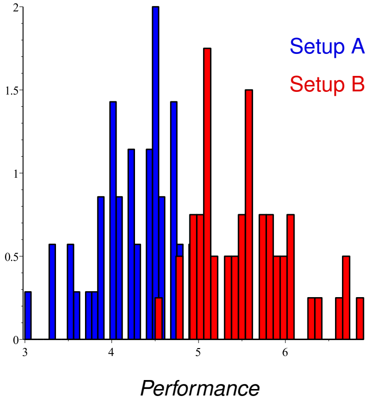
Descriptive Statistics

Descriptive Statistics: present/summarize most important aspects of the given data using suitable

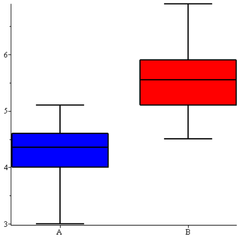
- quantitative summarizations (means, extreme values ...)
- visualization tools

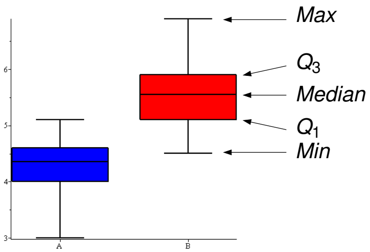


	Interpolated Values														
	Temp														
Time	68.0	68.5	69.0	69.5	70.0	70.5	71.0	71.5	72.0	72.5	73.0	73.5	74.0	74.5	75.0
0.025	2504.08	2638.15	2707.32	2750.09	2784.91	2851.19	2911.62	2940.67	2961.40	2983.17	3000.06	3006.32	3041.01	3125.78	3026.85
0.05	2507.26	2635.76	2704.79	2746.66	2779.96	2846.35	2907.00	2934.98	2955.07	2976.69	2993.64	2999.35	3034.49	3126.43	3036.68
0.075	2510.83	2633.45	2702.58	2743.62	2775.40	2841.84	2902.75	2929.64	2949.08	2970.51	2987.50	2992.60	3027.98	3126.97	3046.32
0.1	2513.93	2631.34	2700.70	2740.99	2771.27	2837.66	2898.88	2924.66	2943.43	2964.66	2981.67	2986.08	3021.49	3127.39	3055.77
0.125	2515.14	2629.60	2699.17	2738.77	2767.61	2833.83	2895.40	2920.07	2938.14	2959.14	2976.16	2979.83	3015.06	3127.71	3065.02
0.15	2514.31	2628.58	2698.02	2736.99	2764.49	2830.38	2892.31	2915.87	2933.23	2953.97	2970.99	2973.86	3008.70	3127.95	3074.08
0.175	2511.84	2628.88	2697.25	2735.66	2762.00	2827.31	2889.59	2912.08	2928.72	2949.17	2966.17	2968.21	3002.47	3128.11	3082.93
0.2	2508.10	2629.91	2696.87	2734.79	2760.22	2824.68	2887.26	2908.72	2924.62	2944.75	2961.71	2962.89	2996.39	3128.21	3091.57
0.225	2503.37	2631.32	2696.88	2734.37	2759.24	2822.57	2885.29	2905.80	2920.96	2940.73	2957.65	2957.93	2990.50	3128.25	3099.99
0.25	2497.84	2632.93	2697.28	2734.42	2759.10	2821.05	[illegible]	[illegible]	2917.76	2937.13	2953.97	2953.36	2984.86	3128.24	3108.19
0.275	2491.66	2634.64	2698.05	2734.91	2759.76	2820.23	[illegible]	[illegible]	[illegible]	2933.97	2950.71	2949.20	2979.52	3128.18	3116.14
0.3	2484.92	2636.35	2699.18	2735.85	2761.12	[illegible]	[illegible]	[illegible]	[illegible]	2931.26	2947.88	2945.48	2974.53	3128.07	3123.83
0.325	2477.71	2638.00	2700.64	2737.22	2763.09	2820.81	2881.06	2898.72	2911.04	2929.03	2945.47	2942.21	2969.96	3127.90	3131.26
0.35	2470.07	2639.54	2702.41	2739.01	2765.59	2822.11	2880.97	2898.13	2909.82	2927.29	2943.52	2939.43	2965.89	3127.66	3138.38
0.375	2462.06	2640.93	2704.45	2741.19	2768.54	2823.98	2881.29	2898.00	2909.13	2926.05	2942.01	2937.16	2962.39	3127.30	3145.19
0.4	2453.70	2642.15	2706.75	2743.75	2771.89	2826.33	2882.03	2898.34	2908.97	2925.33	2940.96	2935.42	2959.55	3126.79	3151.66
0.425	2445.03	2643.15	2709.26	2746.67	2775.62	2829.13	2883.20	2899.16	2909.34	2925.14	2940.37	2934.25	2957.45	3126.07	3157.75
0.45	2436.07	2643.94	2711.97	2749.92	2779.68	2832.32	2884.78	2900.44	2910.23	2925.48	2940.24	2933.67	2956.16	3125.09	3163.42
0.475	2426.82	2644.48	2714.84	2753.48	2784.06	2835.88	2886.78	2902.19	2911.63	2926.34	2940.57	2933.71	2955.74	3123.85	3168.63
0.5	2417.31	2644.77	2717.84	2757.32	2788.73	2839.78	2889.19	2904.40	2913.52	2927.71	2941.36	2934.34	2956.22	3122.46	3173.31
0.525	2407.54	2644.80	2720.95	2761.44	2793.67	2844.01	2891.99	2907.04	2915.89	2929.57	2942.61	2935.55	2957.60	3121.27	3177.39
0.55	2397.51	2644.56	2724.14	2765.79	2798.87	2848.55	2895.19	2910.11	2918.72	2931.90	2944.30	2937.30	2959.85	3120.88	3180.74
0.575	2387.24	2644.05	2727.39	2770.37	2804.31	2853.38	2898.77	2913.60	2921.99	2934.68	2946.43	2939.57	2962.89	3121.69	3183.21
0.6	2376.71	2643.25	2730.67	2775.14	2809.97	2858.49	2902.71	2917.48	2925.67	2937.89	2948.99	2942.35	2966.66	3123.41	3184.53

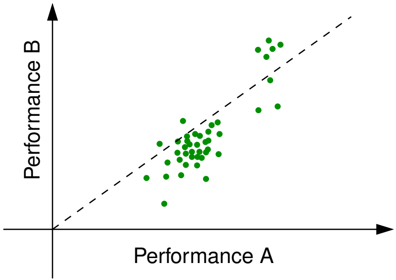


- Quite detailed data summary
- Can be difficult to compare 2 or more setups

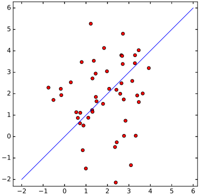

6
5
4
3
A
B



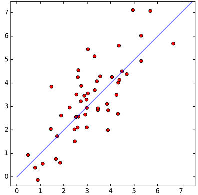
- Q_1: First quartile: 25% of data lies below this point, 75% above
- Q_3: Third quartile: 75% of data lies below this point, 25% above
- Median = Q_2: 50% of data lies below this point, 50% above
- Other features can be added to the Box Plot



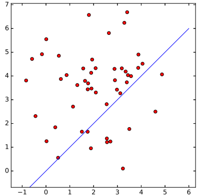
- For paired measurements
- Can reveal existence of input types with different performance characteristics for A,B
- Also used for plotting two measurements, e.g. *time* and *space* for a single algorithm



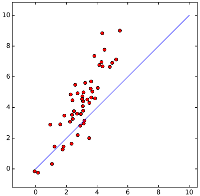
y's same magnitude as $x's$, no correlation



y's same magnitude as $x's$, correlated



y's larger than $x's$, no correlation



y's larger than $x's$, correlated

Hypothesis Testing

Inferential Statistics

Making inferences from the data about the underlying populations/processes ... that produced the data. Types of inferential statistics:

- Testing (our topic)
- Estimation

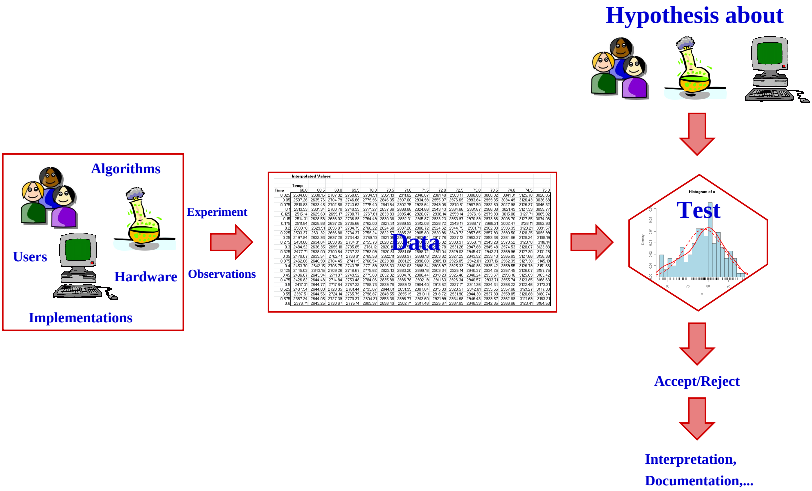
Hypothesis Testing

- At some point, a decision has to be made:
  - Do we treat patients with the new drug, or do we keep the old?
  - Is our system good enough to be sold to customers?
  - Which algorithm should we use?
- For this, the information content of the experimental data has to be reduced to a simple binary decision:

$$\text{Statistical decision: } \textit{accepting} \text{ or } \textit{rejecting} \text{ a hypothesis}$$
$$\Downarrow$$
$$\text{Real-life (e.g. business) decision: deploy system, } \ldots$$

☞ The hypotheses that we want to test refer to *the underlying processes or populations*, not the known data.

Hypothesis Testing



Hypothesis relates to a **quantitative performance measure**:

- Runtime
- Memory usage
- User satisfaction
- Average $ amount spend by visitors of web site
- ...

One sample: hypotheses make a statement about the performance:

My Algorithm/Implementation/Setup/Design satisfies performance condition X

Two sample: hypotheses make a comparison:

[Algorithm/Implementation/Setup/Design] A
performs better than
[Algorithm/Implementation/Setup/Design] B

One Sample Tests

Is this coin fair?

Trials	# Heads	# Tails
1	1	0

Is this coin fair?

Trials	# Heads	# Tails
1	1	0
5	3	2

Is this coin fair?

Trials	# Heads	# Tails
1	1	0
5	3	2
10	5	5

Is this coin fair?

Trials	# Heads	# Tails
1	1	0
5	3	2
10	5	5
20	12	8

Is this coin fair?

Trials	# Heads	# Tails
1	1	0
5	3	2
10	5	5
20	12	8
50	26	24

Is this coin fair?

Trials	# Heads	# Tails
1	1	0
5	3	2
10	5	5
20	12	8
50	26	24
100	52	48

Is this coin fair?

Trials	# Heads	# Tails
1	1	0
5	3	2
10	5	5
20	12	8
50	26	24
100	52	48
500	261	239

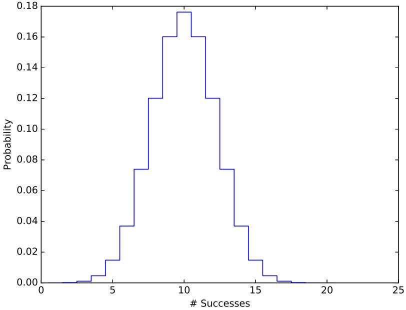
Is this coin fair?

Trials	# Heads	# Tails
1	1	0
5	3	2
10	5	5
20	12	8
50	26	24
100	52	48
500	261	239
1000	557	443

If the coin lands *Heads* with probability p, then the probability of observing in N tosses exactly k heads is

$$B(N,p)(k) = \binom{N}{k} p^k (1-p)^{N-k}$$

Plotted for $N = 20$ and $p = 0.5$:



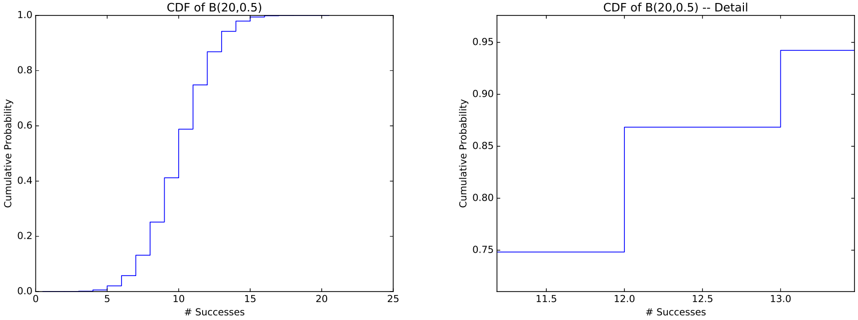
From these numbers:

$$P(\#H \geq 12) \approx 0.12 + 0.073 + 0.0369 + 0.0147 + 0.0046 = 0.249$$

We can read off sums

$$P(\#H \geq k) = P(\#H = k) + P(\#H = k+1) + \ldots + P(\#H = N)$$

directly from the *Cumulative Distribution Function (CDF)*:



$$CDF(k) = \sum_{j \leq k} P(\#H = j); \qquad P(\#H \geq k) = 1 - CDF(k-1)$$

If the coin is fair, then

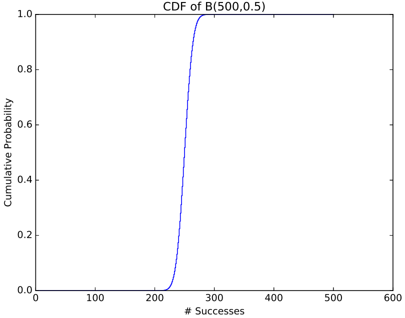
- the probability of seeing in 20 tosses a number of heads greater or equal the observed $\#H = 12$ is ≈ 0.25
- the probability of seeing in 20 tosses a number of heads that deviates from the expected number of 10 heads by the observed difference 12-10 =2 is

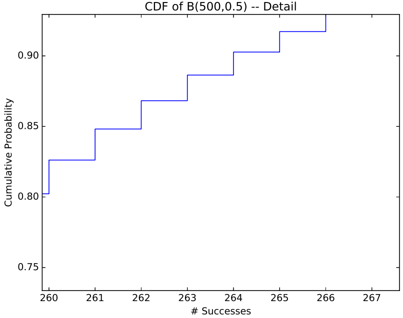
$$P(\#H \leq 8) + P(\#H \geq 12) \approx 2 \cdot 0.25 = 0.5$$

☞ The observed experimental outcome is not very unlikely under the fairness hypothesis

☞ We should not **reject** the fairness hypothesis ($p = 0.5$) on the basis of the data

Looking at $N = 500$ and $\#H = 261$:

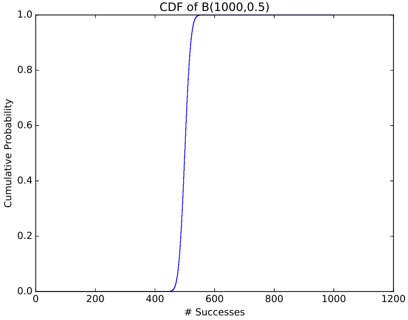


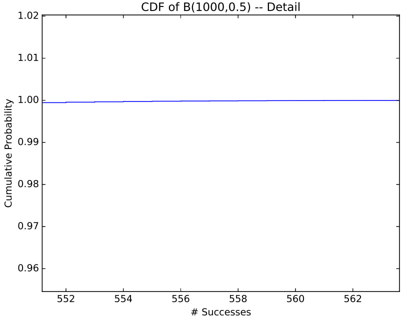


$$P(\#H \geq 261) \approx 0.18 \qquad P(\#H \geq 261 \text{ or } \#H \leq 239) \approx 0.36$$

☞ Also no strong indication that coin is not fair

Looking at $N = 1000$ and $\#H = 557$:





$$P(\#H \geq 557) \approx 0.00015 \qquad P(\#H \geq 557 \text{ or } \#H \leq 443) \approx 0.0003$$

☞ The observed data is extremely unlikely if the fairness hypothesis was true

☞ We should reject the fairness hypothesis

p-value (one sided): Probability under the fairness hypothesis of observing at least as many heads as in the sample

p-value (two sided): Probability under the fairness hypothesis of observing at least as extreme a deviation from the expected number of heads as in the sample

Trials	# Heads	# Tails	*p-value* (two-sided)
1	1	0	
5	3	2	
10	5	5	1.0
20	12	8	0.5034
50	26	24	0.8877
100	52	48	0.7643
500	261	239	0.3476
1000	557	443	0.000347

☞ Reject the hypothesis when the p-value is sufficiently small

- Fix a level of significance α. Typical:

$$\alpha = 0.05 \qquad \text{or} \qquad \alpha = 0.01$$

- *Reject* the hypothesis if the p-value obtained from the sample is $\leq \alpha$

☞ The significance level should be set *before* the experiment or data analysis begins. *Not:*

> *O.k. — I get a p-value of 0.013; I would really like to reject the hypothesis, so let's set* $\alpha = 0.05$

One-sided or Two-sided?

One-sided: “upper/lower bound” hypothesis, such as: $p \leq 0.5$

Two-sided: “point” hypothesis, such as: $p = 0.5$

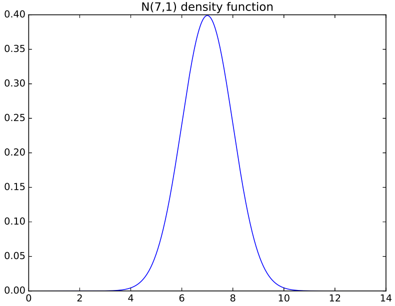
Observed runtimes of implementation on test cases:

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

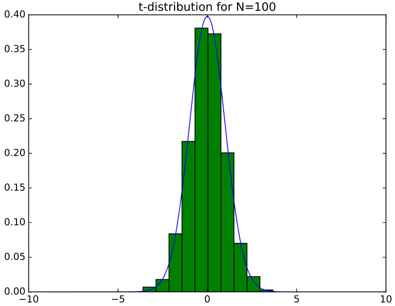
Sample Variance: $\bar{s} := \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2$

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**



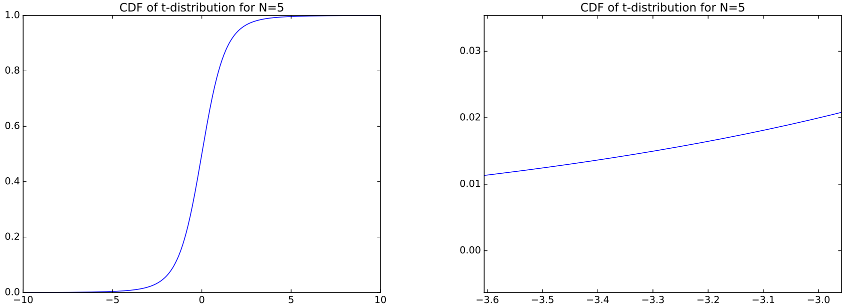
Blue line: t-distribution with 99 degrees of freedom

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

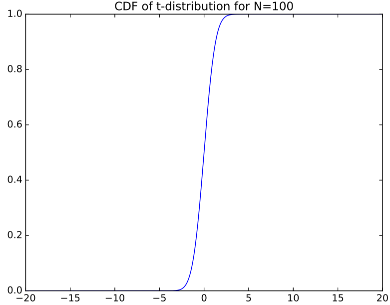
Evaluate the CDF of t-distribution at the observed value of the t-statistic (computed for the (null) hypothesis).

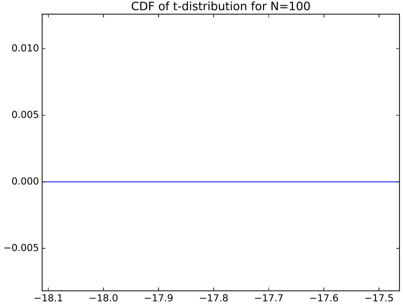
Our data at $N = 5$, $t = -3.644$



☞ p-value (one-sided) is ≈ 0.011.

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





☞ p-value (one-sided) is ≈ 0.0.

	Binomial	**Normal**
Data	$\{0, 1\}$	Real
	$\Downarrow$	$\Downarrow$
Statistics	#Successes	t-statistic
	$\Downarrow$	$\Downarrow$
Distribution	Binomial	t-distribution

$\Downarrow$: place where we injected the null hypothesis