

9. Experiments

AEA 2026

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

Inferential statistics

Hypothesis testing is a general tool to analyze the data.

A *statistical test* is characterized by:

- ▶ a *null hypothesis*,
 H_0 = hypothesis (the null hypothesis)
 H_A = alternative (the alternative hypothesis)
- ▶ assumptions on the experiment (how the data is generated),
- ▶ a *test statistic* (a value computed from the data).

The purpose of the test is to check whether data is consistent with the null hypothesis or not.

If H_0 isn't consistent with the data, it is rejected and there is some evidence that the research hypothesis is true.

Statistical hypothesis testing

Steps:

- ▶ Define the **null hypothesis** H_0 , the **alternative hypothesis** H_A
- ▶ Select a **significance level** α : a threshold below which H_0 will be rejected
- ▶ Compute the **test statistic**
- ▶ Identify the **critical values** using the distribution of the test statistic and the significance level
- ▶ Construct the **critical region** (H_0 is rejected)
- ▶ Take the decision: reject H_0 if the computed value is in the critical region

Hypothesis testing

- ▶ **Two-sided alternative** $H_A : \mu \neq \mu_0$ covering regions on both sides of the hypothesis $H_0 : \mu = \mu_0$
- ▶ **One-sided, left-tail alternative** $H_A : \mu < \mu_0$ covering the region to the left of H_0
- ▶ **One-sided, right-tail alternative** $H_A : \mu > \mu_0$ covering the region to the right of H_0

Example: verify that the the avg. connection speed is 54 Mbps

$$H_0 : \mu = 54$$

$H_A : \mu \neq 54$, μ the average speed of connections

(two-sided alternative)

A one-sided test, if we worry about a low connection speed only:

$$H_0 : \mu = 54$$

$$H_A : \mu < 54$$

Types of errors

	Result of the test	
	Reject H_0	Accept H_0
H_0 is true	Type I error	correct
H_0 is false	correct	Type II error

Sampling errors:

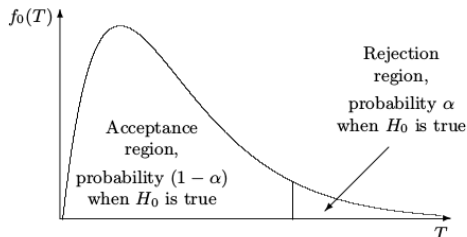
- ▶ A **type I error** occurs when we reject the true H_0 .
- ▶ A **type II error** occurs when we accept the false H_0 .

Probability of a type I error is the **significance level** of a test,
 $\alpha = P\{\text{reject } H_0 | H_0 \text{ is true}\}.$

Level α test

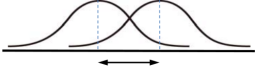
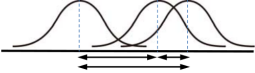
Testing hypothesis: compute a **test statistic** T , a quantity computed from the data, that has some known, tabulated distribution F_0 if H_0 is true.

The **null distribution** F_0 : the distribution of test statistic T when the hypothesis H_0 is true.



Accept H_0 if the test statistic T belongs to the acceptance region.

Statistical tests

		2 groups	n groups ($n > 2$)
data distribution			
Parametric Test (normality)	unpaired (independent)	<ul style="list-style-type: none"> unpaired t-test 	ANOVA (Analysis of Variance) <ul style="list-style-type: none"> one-way ANOVA two-way ANOVA
	paired (related)	<ul style="list-style-type: none"> paired t-test 	
Non-parametric Test (no normality)	unpaired (independent)	<ul style="list-style-type: none"> Mann-Whitney U-test 	one-way data <ul style="list-style-type: none"> Kruskal-Wallis test
	paired (related)	<ul style="list-style-type: none"> sign test Wilcoxon signed-ranks test 	two-way data <ul style="list-style-type: none"> Friedman test

If normality and equal variances aren't guaranteed, use non-parametric tests.

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

Parametric tests

- ▶ Applied when the shape of the distribution is known
- ▶ Variants:
 - ▶ for a population (ex: hypothesis about the mean/variance of a population),
 - ▶ two populations (relation btw means),
 - ▶ more than two populations
- ▶ Examples:
 - ▶ *Z-test*: hypothesis about the mean of a population with normal distribution, known variance
 - ▶ *T-test*: unknown variance and the sample size is not large

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

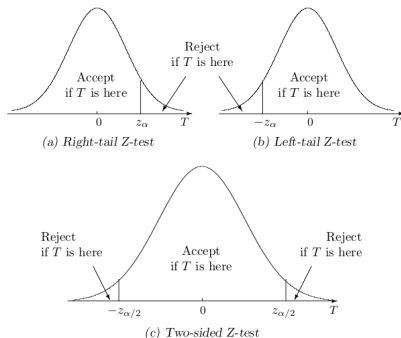
Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

1. Z-test

The null distribution of the test statistic is **Standard Normal**.



$$\text{The test statistic } Z = \frac{\hat{\theta} - E(\hat{\theta})}{s(\hat{\theta})} = \frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{\text{Var}(\hat{\theta})}}.$$

Z-test for means: **when we know the population variance, or when the sample size is large.**

Z-test

- ▶ *One-sample Z-tests for means*

$$H_0 : \mu = \mu_0$$

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

σ/\sqrt{n} standard error

- ▶ *Two-sample Z-tests comparing means* of two populations (independent samples of size n and m)

$$H_0 : \mu_X - \mu_Y = D$$

$$Z = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

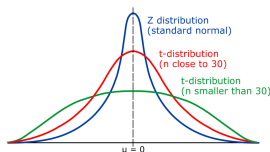
Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

2. t-test (unknown stdev)

T-statistic: $t = \frac{\hat{\theta} - E(\hat{\theta})}{s(\hat{\theta})} = \frac{\hat{\theta} - E(\hat{\theta})}{\sqrt{\text{Var}(\hat{\theta})}}$. The test is based on *Student's T-distribution*.



- ▶ For a right-tail alternative,

$$\begin{cases} \text{reject } H_0 & \text{if } t \geq t_\alpha \\ \text{accept } H_0 & \text{if } t < t_\alpha \end{cases}$$

- ▶ For a left-tail alternative,

$$\begin{cases} \text{reject } H_0 & \text{if } t \leq -t_\alpha \\ \text{accept } H_0 & \text{if } t > -t_\alpha \end{cases}$$

- ▶ For a two-sided alternative,

$$\begin{cases} \text{reject } H_0 & \text{if } |t| \geq t_{\alpha/2} \\ \text{accept } H_0 & \text{if } |t| < t_{\alpha/2} \end{cases}$$

t-test

Hypothesis H_0	Conditions	Test statistic t	Degrees of freedom
$\mu = \mu_0$	Sample size n ; unknown σ	$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	$n - 1$
$\mu_X - \mu_Y = D$	Sample sizes n, m ; unknown but equal standard deviations, $\sigma_X = \sigma_Y$	$t = \frac{\bar{X} - \bar{Y} - D}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}}$	$n + m - 2$
$\mu_X - \mu_Y = D$	Sample sizes n, m ; unknown, unequal standard deviations, $\sigma_X \neq \sigma_Y$	$t = \frac{\bar{X} - \bar{Y} - D}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}$	Satterthwaite approximation, formula (9.12)

When two populations have equal variances, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, the estimator of σ^2 , *pooled sample variance*:

$$s_p^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2}{n+m-2} = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}$$

a. t-test for a population mean

Used when variance σ^2 is unknown and a normal population.

$$H_0: \mu = \mu_0$$

$$H_A: \mu \neq \mu_0$$

Test statistics: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$, where $s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$.

Critical region

- ▶ For a two-sided alternative:

$$R = (-\infty, t_{n-1, 1-\alpha/2}) \cup (t_{n-1, \alpha/2}, \infty)$$

- ▶ For a right-tail alternative, $R = (t_{n-1, \alpha}, \infty)$.

- ▶ For a left-tail alternative, $R = (-\infty, t_{n-1, 1-\alpha})$.

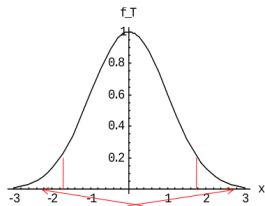


Table of Student's T-distribution

Table A5. Table of Student's T-distribution

t_{α} ; critical values, such that $P\{t > t_{\alpha}\} = \alpha$

ν (d.f.)	α , the right-tail probability									
	.10	.05	.025	.02	.01	.005	.0025	.001	.0005	.0001
1	3.078	6.314	12.706	15.89	31.82	63.66	127.3	318.3	636.6	3185
2	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60	70.71
3	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92	22.20
4	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610	13.04
5	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.894	6.869	9.676
6	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959	8.023
7	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408	7.064
8	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041	6.442
9	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781	6.009
10	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587	5.694
11	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437	5.453
12	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318	5.263
13	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221	5.111
14	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140	4.985
15	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073	4.880
16	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015	4.790
17	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965	4.715
18	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922	4.648
19	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883	4.590
20	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850	4.539
21	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819	4.492
22	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792	4.452
23	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768	4.416
24	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745	4.382
25	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725	4.352
26	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707	4.324
27	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.689	4.299
28	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674	4.276
29	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.660	4.254
30	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646	4.234
32	1.309	1.694	2.037	2.141	2.449	2.738	3.015	3.365	3.622	4.198
34	1.307	1.691	2.032	2.136	2.441	2.728	3.002	3.348	3.601	4.168
36	1.306	1.688	2.028	2.131	2.434	2.719	2.990	3.333	3.582	4.140
38	1.304	1.686	2.024	2.127	2.429	2.712	2.980	3.319	3.566	4.115
40	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551	4.094
45	1.301	1.679	2.014	2.115	2.412	2.690	2.952	3.281	3.520	4.049
50	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496	4.014
...

Example 1: unauthorized use of a computer account

A long-time authorized user of the account makes 0.2 seconds between keystrokes. The following times between keystrokes were recorded when a user typed the username and password:

.24, .22, .26, .34, .35, .32, .33, .29, .19, .36, .30, .15, .17, .28, .38, .40, .37, .27 seconds.

At a 5% level of significance, is this an evidence of an unauthorized attempt?

Test: $H_0 : \mu = 0.2$ vs. $H_A : \mu \neq 0.2$

Significance level $\alpha = 0.05$.

Compute the sample statistics: $n = 18$, $\bar{X} = 0.29$, $s = 0.074$.

Example 1: unauthorized use of a computer account

Compute the T-statistic:

$$t = \frac{\bar{X} - 0.2}{s/\sqrt{n}} = \frac{0.29 - 0.2}{0.074/\sqrt{18}} = 5.16$$

The rejection region: $R = (-\infty, -2.11] \cup [2.11, \infty)$ (we used T-distribution with $18 - 1 = 17$ degrees of freedom and $\alpha/2 = 0.025$ because of the two-sided alternative).

Since $t \in R$, reject H_0 and conclude that there is a significant evidence of an unauthorized use of that account.

b. t-test for comparing means of two populations

- ▶ Equal variances
 - ▶ for small sample size, the hypothesis of normality of populations is required
- ▶ Unequal variances
 - ▶ **estimate the degrees of freedom** ν of a T-distribution that is “closest” to t
Satterthwaite approximation:

$$\nu = \frac{\left(\frac{s_X^2}{n} + \frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2(n-1)} + \frac{s_Y^4}{m^2(m-1)}}$$

If this number of degrees of freedom is non-integer, take the closest int ν

Example 2: Comparison of two servers

An account on server *A* is more expensive than an account on server *B*. However, server *A* is faster. To see if it's optimal to go with the faster but more expensive server, a manager needs to know how much faster it is.

A certain computer algorithm is executed 30 times on server *A* and 20 times on server *B* with the following results,

	Server A	Server B
Sample mean	6.7 min	7.5 min
Sample standard deviation	0.6 min	1.2 min

Is server *A* faster?

Example 2: Comparison of two servers

Test $H_0 : \mu_X = \mu_Y$ vs $H_A : \mu_X < \mu_Y$.

Significance level $\alpha = 0.05$.

$n = 30, m = 20, \bar{X} = 6.7, \bar{Y} = 7.5, s_X = 0.6$, and $s_Y = 1.2$.

This is the case of unknown, unequal standard deviations.

Use *Satterthwaite approximation* to find the number of degrees of freedom:

$$\nu = \frac{\left(\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}\right)^2}{\frac{(0.6)^4}{30^2(29)} + \frac{(1.2)^4}{20^2(19)}} = 25.4$$

Reject the null hypothesis if $t \leq -1.708$.

$$t = \frac{6.7 - 7.5}{\sqrt{\frac{(0.6)^2}{30} + \frac{(1.2)^2}{20}}} = -2.7603 \in \mathcal{R}$$

Reject H_0 and conclude that there is evidence that server A is faster.

Inferential statistics

1. Choose the *significance level* $0 < \alpha < 1$, the confidence you want to achieve (ex: $\alpha = 0.05$ - accept 5% error).
2. Then compute the test statistic for the data.
3. If the p -value of the statistic is smaller than the significance level, the null hypothesis is rejected.

P-value

How do we choose the significance level α ?

P-value is the **lowest** significance level α that forces **rejection** of the null hypothesis.

P-value is also the **highest** significance level α that forces **acceptance** of the null hypothesis.

Testing hypotheses with a P-value:

- ▶ For $\alpha > P$, reject H_0
- ▶ For $\alpha < P$, accept H_0

Practically,

- ▶ If $P < 0.01$, reject H_0
- ▶ If $P > 0.1$, accept H_0

Only if the P-value falls between 0.01 and 0.1, we have to think about the level of significance.

Computing P-values

P – value is the probability of observing a test statistic at least as extreme as t_{obs} .

Hypothesis H_0	Alternative H_A	P-value	Computation
$\theta = \theta_0$	right-tail $\theta > \theta_0$	$P \{t \geq t_{obs}\}$	$1 - F_\nu(t_{obs})$
	left-tail $\theta < \theta_0$	$P \{t \leq t_{obs}\}$	$F_\nu(t_{obs})$
	two-sided $\theta \neq \theta_0$	$P \{ t \geq t_{obs} \}$	$2(1 - F_\nu(t_{obs}))$

F_ν is the cumulative distribution **function** of T-distribution with ν degrees of freedom

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

Non-Parametric Tests

- ▶ Non-parametric statistics **doesn't assume any particular distribution**
- ▶ Are **less powerful** (the less you assume about the data, the less you can find out from it)
- ▶ Having fewer requirements, they are applicable to wider applications
- ▶ Variants: verify a hypothesis about population distribution (*chi-squared test Goodness-of-Fit*), independence/association (*chi-squared, Fisher*), **comparison** in case of nominal/ordinal characteristics (*sign test, rank tests, etc*)

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

1. The sign test

A sample (x_1, \dots, x_n) of n real numbers. The assumptions: all the x_i are drawn **independently from the same distribution**.

The **null hypothesis** H_0 : the **median** $M = m$.

Test H_0 against a one-sided ($H_A : M < m$, $H_A : M > m$) or a two-sided alternative $H_A : M \neq m$.

We are testing whether exactly a half of the population is below m and a half is above m .

The sign test

Compute the **test statistic** $S := |\{i | x_i > m\}|$.

If H_0 is true, each x_i is equally likely to be above/below m . Therefore, S is distributed according to a **binomial distribution** with parameters $p = 1/2$ and n .

Suppose the observed value of S is k , w.l.o.g. $k \geq n/2$. Compute the p -value: the probability that S is *at least* k : $1/2^n \sum_{i=k}^n \binom{n}{i}$.

The sign test

Test of the median, $H_0 : M = m$

Test statistic $S =$ number of $X_i > m$

Null distribution $S \sim \text{Binomial}(n, 1/2)$

For large n , $S \approx \text{Normal}(n/2, \sqrt{n}/2)$
(if the distribution of X_i is continuous)

The sign test

Example: if $n = 15$, $k = 12$, we get a p -value of 0.018 \rightarrow reject H_0 at a significance level of $\alpha = 0.05$.

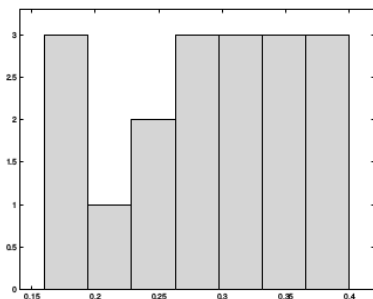
Instead, we have some evidence that the real median is *greater than 0*.

Obs: we couldn't conclude this if we selected $\alpha = 0.01$.

Example 1: Unauthorized use of a computer account

Times between keystrokes: .24, .22, .26, .34, .35, .32, .33, .29, .19, .36, .30, .15, .17, .28, .38, .40, .37, .27 seconds. The account owner usually makes 0.2 sec between keystrokes.

We had to assume the Normal distribution of data (for T-test).
The histogram does not confirm this assumption.



Example 1: Unauthorized use of a computer account

The sign test: $H_0 : M = 0.2$ vs $H_A : M \neq 0.2$

The test statistic: $S_{obs} = 15$ (15 of 18 recorded times exceed 0.2).
From Binomial distribution table with $n = 18$ and $p = 0.5$, find the P – value:

$$P = 2\min(P\{S \leq S_{obs}\}, P\{S \geq S_{obs}\}) = 2\min(0.9993, 0.0038) = 0.0076$$

The sign test rejects H_0 at any $\alpha > 0.0076$, which is an evidence that the account was used by an unauthorized person.

The sign test

Application: compare pairwise samples from two different distributions. For comparing two algorithms, suppose we have samples (x_1, \dots, x_n) , (y_1, \dots, y_n) , x_i , y_i performance measures of algorithms on instance i . Question: Is it true that the 1st algorithm is better than the 2nd?

Consider the sequence of differences $d_i = y_i - x_i$ and do the sign test on this sample. The null hypothesis: the medians of the performance distribution are equal (the performance of both algorithms is the same).

If sufficiently many d_i are positive, H_0 is rejected and there is evidence that 1st algorithm is better.

Note: H_0 is rejected if there are too few positive d_i , which would indicate that 2nd algorithm is better.

The Sign Test and Heuristics for the TSP

Example: a new algorithm (CCAO¹) for Euclidean TSP:

1. construct a partial tour from the convex hull of the cities,
2. includes remaining cities (cheapest insertion, angle selection)
3. improve the solution (Or-opt); other post-processors: 2-opt, 3-opt (find better tours by exchanging 2 or 3 edges of the current tour until no further improvement is possible).

Only 8 instances! A larger #samples is required to draw statistical significant conclusions!

Compare the algorithm to other heuristics: apply the sign test to assess solution quality

- ▶ the algorithm is better than heuristics with a weak post-processor (i.e., 2-opt).
- ▶ the algorithm is as good as those with a strong post-processor (i.e., Or-opt, 3-opt).

¹https://www.informs-sim.org/wsc83papers/1983_0094.pdf

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

2. The Wilcoxon signed-rank test

The **Wilcoxon signed-rank test** is an extension to the sign test: it takes the value of the differences into account. Used to compare two related samples, matched samples, or repeated measurements on a single sample.

1. The distances between observations and the tested value, $d_i = |X_i - m|$.
2. Order the distances and compute their **ranks** R_i ($R_i = r$: X_i is the r -th smallest observation in the sample).
3. Take only the ranks corresponding to observations X_i greater than m . Their sum is the **test statistic** $W = \sum_{i: X_i > m} R_i$.
4. Large values of W suggest rejection of H_0 in favor of $H_A : M > m$; small values support $H_A : M < m$; both support a two-sided alternative $H_A : M \neq m$

The Wilcoxon signed-rank test

Test of the median, $H_0 : M = m$.

Test statistic $W = \sum_{i: X_i > m} R_i$, where R_i is the rank of $d_i = |X_i - m|$.

Null distribution: Table of Critical values

For $n \geq 15$, $W \approx \text{Normal}\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$

Assumptions: the distribution of X_i is continuous and symmetric.

Table of Critical Values for the Wilcoxon Signed Rank Test

Table A8. Table of Critical Values for the Wilcoxon Signed Rank Test

For the left-tail test, the table gives the largest integer w such that $P\{W \leq w \mid H_0\} \leq \alpha$.

For the right-tail test, the table gives the smallest integer w such that $P\{W \geq w \mid H_0\} \leq \alpha$.

A missing table entry means that such an integer does not exist among possible values of W .

n	α , left-tail probability for the left-tail test							α , right-tail probability for the right-tail test						
	0.001	0.005	0.010	0.025	0.050	0.100	0.200	0.200	0.100	0.050	0.025	0.010	0.005	0.001
1	—	—	—	—	—	—	—	—	—	—	—	—	—	—
2	—	—	—	—	—	—	—	—	—	—	—	—	—	—
3	—	—	—	—	—	—	0	6	—	—	—	—	—	—
4	—	—	—	—	—	0	2	8	10	—	—	—	—	—
5	—	—	—	—	0	2	3	12	13	15	—	—	—	—
6	—	—	—	0	2	3	5	16	18	19	21	—	—	—
7	—	—	0	2	3	5	8	20	23	25	26	28	—	—
8	—	0	1	3	5	8	11	25	28	31	33	35	36	—
9	—	1	3	5	8	10	14	31	35	37	40	42	44	—
10	0	3	5	8	10	14	18	37	41	45	47	50	52	55
11	1	5	7	10	13	17	22	44	49	53	56	59	61	65
12	2	7	9	13	17	21	27	51	57	61	65	69	71	76
13	4	9	12	17	21	26	32	59	65	70	74	79	82	87
14	6	12	15	21	25	31	38	67	74	80	84	90	93	99
15	8	15	19	25	30	36	44	76	84	90	95	101	105	112
16	11	19	23	29	35	42	50	86	94	101	107	113	117	125
17	14	23	27	34	41	48	57	96	105	112	119	126	130	139
18	18	27	32	40	47	55	65	106	116	124	131	139	144	153
19	21	32	37	46	53	62	73	117	128	137	144	153	158	169
20	26	37	43	52	60	69	81	129	141	150	158	167	173	184
21	30	42	49	58	67	77	90	141	154	164	173	182	189	201
22	35	48	55	65	75	86	99	154	167	178	188	198	205	218
23	40	54	62	73	83	94	109	167	182	193	203	214	222	236
24	45	61	69	81	91	104	119	181	196	209	219	231	239	255
25	51	68	76	89	100	113	130	195	212	225	236	249	257	274
26	58	75	84	98	110	124	141	210	227	241	253	267	276	293
27	64	83	92	107	119	134	153	225	244	259	271	286	295	314
28	71	91	101	116	130	145	165	241	261	276	290	305	315	335
29	79	100	110	126	140	157	177	258	278	295	309	325	335	356
30	86	109	120	137	151	169	190	275	296	314	328	345	356	379

Example 1: Supply and demand

You have to ensure that the printers don't run out of paper. During the first six days, the lab consumed: 7, 5.5, 9.5, 6, 3.5, 9 cartons of paper. Does this imply significant evidence, at the 5% level of significance, that the median daily consumption of paper is more than 5 cartons?

Right-tail test $H_0 : M = 5$ vs $H_A : M > 5$.

Example 1: Supply and demand

For $n = 6$ and $\alpha = 0.05$, we'll reject H_0 when the sum of positive ranks $T \geq 19$.

Compute distances $d_i = |X_i - 5|$ and rank them from the smallest to the largest.

i	X_i	$X_i - 5$	d_i	R_i	sign
1	7	2	2	4	+
2	5.5	0.5	0.5	1	+
3	9.5	4.5	4.5	6	+
4	6	1	1	2	+
5	3.5	-1.5	1.5	3	-
6	9	4	4	5	+

Compute T adding the "positive" ranks only:

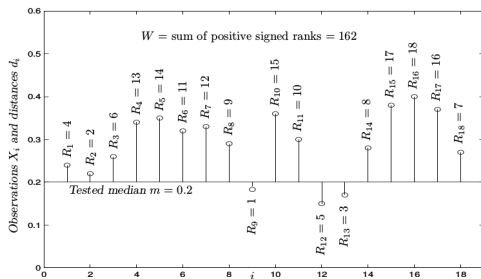
$$T = 4 + 1 + 6 + 2 + 5 = 18 < 19.$$

No rejection: at the 5% level, data do not provide significance evidence that the median consumption of paper exceeds 5 cartons.

Example 2: Unauthorized use of a computer account

Test $M = 0.2$ vs $M \neq 0.2$.

Compute the distances $d_1 = |X_1 - m| = |0.24 - 0.2| = 0.04, \dots$,
 $d_{18} = |0.27 - 0.2| = 0.07$ and rank them.



Notice that the 9-th, 12-th, and 13-th observations are below the tested value $m = 0.2$ while all the others are above. The test statistic (the sum of only positive signed ranks):

$$W = \sum_{i: X_i > m} R_i = 162$$

Example 2: Unauthorized use of a computer account

Compute a P – value. This is a two-sided test, therefore,
 $P = 2\min(P\{W \leq 162\}, P\{W \geq 162\}) < 2 \cdot 0.001 = 0.002$ (use Table with $n = 18$).

- ▶ Obs: for the sample size $n = 18$, we can also use the Normal approximation.

The test shows strong evidence that the account was used by an unauthorized person.

Inferential statistics. Introduction

Parametric Tests

Z-test

t-test

Non-Parametric Tests

Sign test

Wilcoxon signed-rank test

Mann-Whitney-Wilcoxon rank sum test

3. Mann-Whitney-Wilcoxon rank sum test

Wilcoxon signed rank test can be **extended to a two-sample problem: compare two populations**, the population of X and the population of Y . In terms of their cumulative distribution functions, test

$$H_0 : F_X(t) = F_Y(t), \text{ for all } t$$

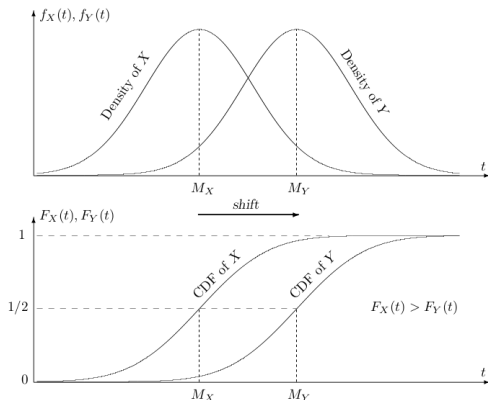
Alternative H_A : either Y is stochastically larger than X , and $F_X(t) > F_Y(t)$, or it is stochastically smaller than X , and $F_X(t) < F_Y(t)$.

Mann-Whitney-Wilcoxon rank sum test

1. Combine all X_i and Y_j into one sample.
2. Rank observations in this combined sample. Ranks R_i are from 1 to $(n + m)$. Some of these ranks correspond to X -variables, others to Y -variables.
3. The test statistic U is the sum of all X -ranks.

If U is small, X -variables have low ranks in the combined sample, so they are generally smaller than Y -variables. This implies that Y is stochastically larger than X : support the alternative $H_A : F_Y(t) < F_X(t)$.

Mann-Whitney-Wilcoxon rank sum test



Variable Y is stochastically larger than variable X . It has a larger median $M_Y > M_X$ and a smaller cdf $F_Y(t) < F_X(t)$.

Mann-Whitney-Wilcoxon rank sum test

Test of two populations, $H_0 : F_X = F_Y$.

Test statistic $U = \sum_i R_i$, where R_i are ranks of X_i in the combined sample of X_i and Y_i .

Null distribution: Table of Critical values

For $n, m \geq 10$, $U \approx \text{Normal}\left(\frac{n(n+m+1)}{2}, \sqrt{\frac{nm(n+m+1)}{12}}\right)$

Assumptions: the distributions of X_i and Y_i are continuous;
 $F_X(t) = F_Y(t)$ under H_0 ; $F_X(t) < F_Y(t)$ for all t or
 $F_X(t) > F_Y(t)$ for all t under H_A .

Table of Critical Values

Table A9. Critical Values for the Mann-Whitney-Wilcoxon Rank-Sum Test

For the left-tail test, the table gives the largest integer u such that $P\{U \leq u \mid H_0\} \leq \alpha$.

For the right-tail test, the table gives the smallest integer u such that $P\{U \geq u \mid H_0\} \leq \alpha$.

A missing table entry means that such an integer does not exist among possible values of U .

n_1	n_2	α , left-tail probability for the left-tail test $H_A: X$ is stochastically smaller than Y						α , right-tail probability for the right-tail test $H_A: X$ is stochastically larger than Y							
		0.001	0.005	0.010	0.025	0.050	0.100	0.200	0.100	0.050	0.025	0.010	0.005	0.001	
3	2	—	—	—	—	—	6	7	11	12	—	—	—	—	—
3	3	—	—	—	—	6	7	8	13	14	15	—	—	—	—
3	4	—	—	—	—	6	7	9	15	17	18	—	—	—	—
3	5	—	—	—	6	7	8	10	17	19	20	21	—	—	—
3	6	—	—	—	7	8	9	11	19	21	22	23	—	—	—
3	7	—	—	6	7	8	10	12	21	23	25	26	27	—	—
3	8	—	—	6	8	9	11	13	23	25	27	28	30	—	—
3	9	—	6	7	8	10	11	14	25	28	29	31	32	33	—
3	10	—	6	7	9	10	12	15	27	30	32	33	35	36	—
3	11	—	6	7	9	11	13	16	29	32	34	36	38	39	—
3	12	—	7	8	10	11	14	17	31	34	37	38	40	41	—
4	2	—	—	—	—	—	10	11	17	18	—	—	—	—	—
4	3	—	—	—	—	10	11	13	19	21	22	—	—	—	—
4	4	—	—	—	10	11	13	14	22	23	25	26	—	—	—
4	5	—	—	10	11	12	14	15	25	26	28	29	30	—	—
4	6	—	10	11	12	13	15	17	27	29	31	32	33	34	—
4	7	—	10	11	13	14	16	18	30	32	34	35	37	38	—
4	8	—	11	12	14	15	17	20	32	35	37	38	40	41	—
4	9	—	11	13	14	16	19	21	35	37	40	42	43	45	—
4	10	10	12	13	15	17	20	23	37	40	43	45	47	48	50
4	11	10	12	14	16	18	21	24	40	43	46	48	50	52	54
4	12	10	13	15	17	19	22	26	42	46	49	51	53	55	58
5	2	—	—	—	—	15	16	17	23	24	25	—	—	—	—
5	3	—	—	—	15	16	17	19	26	28	29	30	—	—	—
5	4	—	—	15	16	17	19	20	30	31	33	34	35	—	—
5	5	—	15	16	17	19	20	22	33	35	36	38	39	40	—
5	6	—	16	17	18	20	22	24	36	38	40	42	43	44	—
5	7	—	16	18	20	21	23	26	39	42	44	45	47	49	—
5	8	15	17	19	21	23	25	28	42	45	47	49	51	53	55
5	9	16	18	20	22	24	27	30	45	48	51	53	55	57	59
5	10	16	19	21	23	26	28	32	48	52	54	57	59	61	64
5	11	17	20	22	24	27	30	34	51	55	58	61	63	65	68
5	12	17	21	23	26	28	32	36	54	58	62	64	67	69	73
6	2	—	—	—	—	21	22	23	31	32	33	—	—	—	—
6	3	—	—	22	23	24	26	26	34	36	37	38	—	—	—
6	4	—	21	22	23	24	26	28	38	40	42	43	44	45	—
6	5	—	22	23	24	26	28	30	42	44	46	48	49	50	—
6	6	—	23	24	26	28	30	33	45	48	50	52	54	55	—
6	7	21	24	25	27	29	32	35	49	52	55	57	59	60	63
6	8	22	25	27	29	31	34	37	53	56	59	61	63	65	68
6	9	23	26	28	31	33	36	40	56	60	63	65	68	70	73
—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Example 1: On-line incentives

Managers of a shopping portal suspect that more customers participate in on-line shopping if they are offered some incentive, such as a discount or cash back. To verify this hypothesis, they chose 12 days at random, offered a 5% discount on 6 randomly selected days, but did not offer any incentives on the other 6 days. The discounts were indicated on the links leading to this shopping portal.

With the discount, the portal received (rounded to 100s) 1200, 1700, 2600, 1500, 2400, and 2100 hits. Without the discount, 1400, 900, 1300, 1800, 700, and 1000 hits were registered. Does this support the managers' hypothesis?

Example 1: On-line incentives

F_X , F_Y the cdf of the number of hits without the discount and with the discount, respectively.

Test $H_0 : F_X = F_Y$ vs $H_A : X$ is stochastically smaller than Y .

To compute the test statistic: combine all the observations and order them

700, 900, 1000, 1200, 1300, 1400, 1500, 1700, 1800, 2100, 2400, 2600

X -variables are underlined. In the combined sample, their ranks are 1, 2, 3, 5, 6, and 9, and their sum is $U_{obs} = \sum_{i=1}^6 R_i = \underline{26}$.

From Table of critical vals. with $n = m = 6$: the one-sided left-tail P-value is $p \in (0.01, 0.025]$. Although it implies some evidence that discounts help increase the on-line shopping activity, this evidence is not overwhelming.

We can conclude that the evidence supporting the managers' claim is significant at any $\alpha > 0.025$.