Statistical Methods for Transport and Logistic Processes

MSc (Rail, Transport and Logistics)

Non-Parametric Tests

Instructor: George SUN, Ph.D.

<georgeqsun@hotmail.com>

Reference: Walpole, R.E., Myers, R.H., Myers, S.L. and Ye, K., ***Probability and Statistics for Engineers and Scientists***. 9th edition

Pearson International Edition 2012, or

Pearson New International Edition 2014

Review

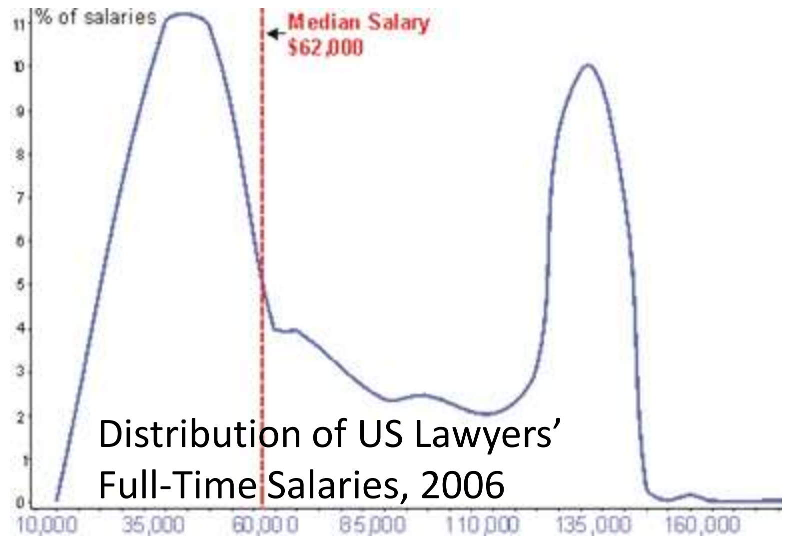
- χ^2 tests on populations $$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$
  - Goodness-of-Fit test on the fit of data to specified probability distributions
  - Independence of two variables through the *Contingency Table*
  - Homogeneity (by assigning values to one variable, to test whether the proportions of another variable are the same) through the *Contingency Table*
- χ^2 tests versus *z*-, *t*-, χ^2-, and *F*-test of mean, proportion, variance

***z*-, *t*-, χ^2-, and *F*-test of parameters**	**χ^2 tests**
Test inferred parameters which correspond to confidence interval	Test expected values calculated from expected population characteristics
Parameters are from approximately normal distributions	Expected values can be from any distributions
Applicable to continuous (and selected discrete) random variables	Applicable to continuous, discrete, and qualitative random variables

Non-Parametric Tests
Contents

11.1. Non-Parametric Tests

- Non-parametric methods
  - assume no knowledge about the distributions of the underlying populations
  - are able to analyse nominal/ categorical or ordinal data and draw statistical conclusions



- In general, non-parametric (**distribution-free**) tests satisfy at least one of the following conditions
  - The method can be used with nominal data (e.g. gender, race)
  - The method can be used with ordinal data (e.g. dislike – like)
  - The method can be used with interval or ratio data when no assumption can be made about the population probability distribution

11.1. Non-Parametric Tests

Advantages:

- Fewer assumptions about the population
  - Shape
  - Variance
- Valid for small samples
- Applicable to nominal and ordinal scales
- Calculations simple

Disadvantages:

- Sample data used less efficiently
- Power of non-parametric analysis lower
- Places greater reliance on statistical tables if computer statistical package or spreadsheet not being used

Common non-parametric methods:

- Sign Test (the median value)
- Wilcoxon Singed-Rank Test (the symmetry of distribution)
- Wilcoxon Rank-Sum Test (the similarity between two populations)
- Tests for Randomness: Runs Test

11.2. Sign Test

- When a continuous symmetrical population is sampled, the probability of getting a sample value less than the median and a sample value greater than the median are both 1/2 (= **0.5**).
- Typical applications of the *sign test* involve using a sample of *n*
  – potential customers to determine whether there is a preference for one of two brands.
  – which is < 30 and the underlying population is decidedly non-normal to test the hypothesis $\tilde{\mu} = \tilde{\mu}_0$ (*median*)
- The data is recorded as follows (hence the name of "sign test")
  – a "+" (plus) sign is used if the sample value $x_i > \tilde{\mu}_0$ (or if the individual prefers one brand)
  – a "-" (minus) sign if is used if the sample value $x_i < \tilde{\mu}_0$ (or if the individual prefers the other brand)
  – data is discarded if the sample value $x_i = \tilde{\mu}_0$ (or for those showing no preference), with commensurate reduction in *n*

11.2. Sign Test

- Example: 11 data represent number of hours that a rechargeable hedge trimmer operates before a recharge is required: 1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7. (*Example 16.1*)

– Test the hypothesis, at 0.05 level of significance, that this trimmer operates a median of 1.8 hours before requiring a recharge.

1) H_0: $\tilde{\mu} = 1.8$, 2) H_1: $\tilde{\mu} \neq 1.8$, 3) $\alpha = 0.05$.
4) Critical region: binomial random variable X, with $p = 1/2$
5) Computation: Replacing each value by "+" if > 1.8, "–" if < 1.8 and discarding measurements = 1.8, we obtain $n = 10$ and $x = 3$ (**<** $n/2$)

Data	1.5	2.2	0.9	1.3	2.0	1.6	1.8	1.5	2.0	1.2	1.7
sign	–	+	–	–	+	–		–	+	–	–

$P = 2P(X \leq 3 \text{ with } p = 1/2) = 2\sum_{i=0}^{3} b(i;10,0.5) = 0.3438$ **>** α (=0.05).

6) Decision: **Do not reject H_0**. Insufficient evidence to conclude that the median operating time is significantly different from 1.8 hours

11.2. Sign Test

- For the sign test, the *test statistic* is the *binomial* random variable X, representing the number of plus signs in the random sample n
- Hypothesis setup (with p = 1/2 = 0.5)

1. H_0: $\tilde{\mu} = \tilde{\mu}_0$ versus H_1: $\tilde{\mu} < \tilde{\mu}_0$ for one-sided test
   – Reject H_0 if $P = P(X \leq x$ when $p = 1/2) < \alpha$ (level of significance)
2. H_0: $\tilde{\mu} = \tilde{\mu}_0$ versus H_1: $\tilde{\mu} > \tilde{\mu}_0$ for one-sided test
   – Reject H_0 if $P = P(X \geq x$ when $p = 1/2) < \alpha$ (level of significance)
3. H_0: $\tilde{\mu} = \tilde{\mu}_0$ versus H_1: $\tilde{\mu} \neq \tilde{\mu}_0$ for two-sided test
   – Reject H_0 if $P = 2P(X \leq x$ when $p = 1/2) < \alpha$ for $x < n/2$, and $P = 2P(X \geq x$ when $p = 1/2) < \alpha$ for $x > n/2$

- The test statistic $P(X \leq x$ when $p = 1/2) = \sum_{i=0}^{x} b(i; n, 0.5)$
- $P(X \geq x$ when $p = 1/2) = 1 - P(X \leq x\text{-}1$ when $p = 1/2)$

– If n is large, the binomial probabilities can be approximated from the normal distribution: $\mu = np = 0.5n$; $\sigma = \sqrt{np(1\text{-}p)} = \sqrt{0.25n}$

11.2. Sign Test

- Review: Binomial probability sums (cumulative)

$$P(X \le r) = B(r; n, p) = \sum_{x=0}^{r} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$P(x)$ = probability of x successes in n trials, with probability of success p on each trial

x = 0, 1, 2, ..., n, number of "successes"

n = number of trials (i.e. the sample size)

p = probability of "success" for each trial

Example: Flip a coin 5 times, x=number of heads:

x = 0, 1, 2, 3, 4, 5

n = 5; p = 0.5

- *Statistical Table Binomial Probability Sums*: *n*, *r*, *p*

- Excel function>BINOMDIST(*x*, *n*, *p*, "cumulative")

Where "cumulative" =

– false (or 0) → return the probability $f(x)$

– true (or 1) → return the cumulative probability $F(x)$

11.2. Sign Test

- Exercise 1: The following represents the time, in minutes, that patients waited during 12 visits before being seen by the doctor: 17, 15, 20, 20, 32, 28, 12, 26, 25, 25, 35, 24. (16.1)

– Test the hypothesis, at 0.05 level of significance, that the median waiting time for the patients is not more than 20 minutes

1) H_0: $\tilde{\mu} = 20$, 2) H_1: $\tilde{\mu} > 20$, 3) $\alpha = 0.05$.

4) Critical region: binomial random variable X, with $p = 1/2$

5) Computation: Replacing each value by "+" if > 20, "–" if < 20 and discarding measurements = 20, we obtain $n = 10$, $x = 7$

Data	17	15	20	20	32	28	12	26	25	25	35	24
sign	–	–			+	+	–	+	+	+	+	+

$P = P(X \geq 7 \text{ with } p = 1/2) = 1 - P(X < 7 \text{ with } p = 1/2) =$

$= 1 - P(X \leq 6 \text{ with } p = 1/2) = 1 - \sum_{i=0}^{6} b(i;10,0.5) = 0.1719 > 0.05.$

19→
38→

6) Decision: **Do not reject H_0.** Insufficient evidence to conclude that the median waiting time is more than 20 minutes

11.2. Sign Test

- With a relatively large sample size (e.g. n>10), the normal approximation of the binomial distribution with p = 1/2 is used
  - $\mu = np = 0.5n$; $\sigma = \sqrt{np(1-p)} = \sqrt{n(0.5)(1-0.5)} = \sqrt{0.25n}$
  - $z = \dfrac{x - np}{\sqrt{np(1-p)}} = \dfrac{x - 0.5n}{\sqrt{0.25n}}$

- Example: To determine whether the use of radial tyres improves fuel economy, 16 cars are driven over a prescribed test course, first equipped with regular tyres and then with radial tyres. The fuel economy data, in km per litre of petrol, is collected below

sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Radial tyre	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Regular tyre	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8

– Can we conclude at the 0.05 level of significance that cars equipped with radial tyres obtain better fuel economy than those equipped with regular tyres? (*Example 16.2*)

23→

11.2. Sign Test

- <u>Example</u>: *(continued)*

Let $\hat{\mu}_1$ and $\hat{\mu}_2$ represent the median fuel economy of all cars tested using radial tyres and with regular tyres, respectively

1) H_0: $\hat{\mu}_1 = \hat{\mu}_2$; 2) H_1: $\hat{\mu}_1 > \hat{\mu}_2$; 3) $\alpha = 0.05$.

4) Critical region: normal approximation of the binomial random variable X with $p = 1/2$: $z > 1.645$ (i.e. $z_\alpha = z_{0.05}$)

5)

sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Radial tyre	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Regular tyre	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8
sign	+	–	+	+	–	+		+	+		+	+	+	+	–	+

$n = 14$ (after discarding equal values), $x = 11$ (number of plus sign)

$z = (11 - (0.5)(14))/\sqrt{(0.25)(14)} = 4/1.718 = \mathbf{2.138} > 1.645$

$P(X \geq 11) \approx P(Z \geq 2.138) = 1 - P(Z < 2.138) = 1 - 0.9837 = 0.0163$

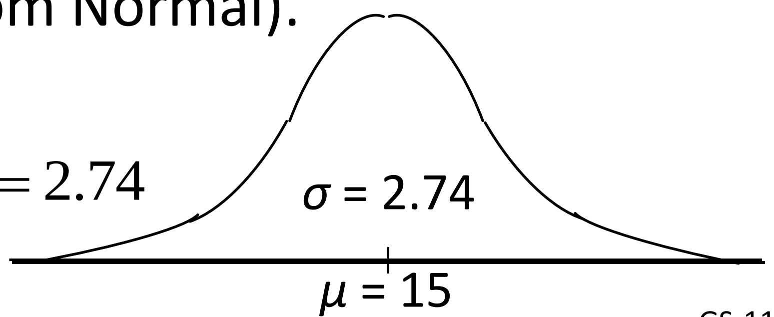
6) Decision: **Reject H_0**. We conclude that the use of radial tyres improves fuel economy.

11.2. Sign Test

- Exercise 2: As part of a market research study, a sample of 36 consumers were asked to taste two brands of peanut butter and indicate a preference
  - Do the data shown below indicate a significant difference in the consumer preferences for the two brands?
    - 18 preferred Hoppy Peanut Butter (“+” sign recorded)
    - 12 preferred Pokey Peanut Butter (“-” sign recorded)
    - 6 had no preference

38→

- The analysis is based on a sample size of n = 18 + 12 = 30 (after discarding “no preference”)
  - Sampling distribution of the number of “+” values if there is no brand preference (approximated from Normal).
    - $\mu = np = (30)(0.5) = 15$
    - $\sigma = \sqrt{np(1-p)} = \sqrt{(30)(0.5)(1-0.5)} = 2.74$



11.2. Sign Test

- <u>Exercise 2</u>: *(continued)*

1) H_0: No preference for one brand over the other exists.
2) H_1: A preference for one brand over the other exists.
3) $\alpha = 0.05$.
4) Critical region: normal approximation of the binomial random variable X with $p = 1/2$: $z < -1.96$ or $z > 1.96$ (i.e. $z_{\alpha/2}=z_{0.025}$)
5) Computation: $n = 30$, $x = 18$
   $z = (18 - (0.5)(30))/\sqrt{(0.25)(30)} = 3/2.74 = 1.095 < 1.96$
   $P = 2P(Z > 1.095) = 2[1 - P(Z < 1.095)] \approx (2)(1 - 0.863) = 0.274$
6) Decision: **Do not reject H_0**. There is insufficient evidence in the sample to conclude that a difference in preference exists for the two brands of peanut butter.

– Fewer than 10 or more than 20 individuals would have to have a preference for a particular brand in order for us to reject H_0

11.2. Sign Test

- Sign Test is
  - the simplest non-parametric test
  - applicable to data represented by two types (positive and negative) responses which are not on a numerical scale
- Sign Test is concerned with median, or difference between a pair of observations $\rightarrow$ to test whether the sample deviates from equal number of signs: 50% plus (+) and 50% minus (-)
  - Binomial distribution is used with $p = 0.5$
  - With a relatively large sample size (e.g. $n>10$), the normal approximation of binomial distribution is used

$$z = \frac{x - np}{\sqrt{np(1-p)}} = \frac{x - 0.5n}{\sqrt{0.25n}}$$

- Can more information be extracted from the data for a non-parametric test?

11.3. Signed-Rank Test

- A test utilises both direction signs and magnitude is referred to as the **Wilcoxon signed-rank test**
- Null hypothesis H_0: $\tilde{\mu} = \tilde{\mu}_0$ (median). Signed-rank test applies to symmetric continuous distributions:
  - Calculate the differences by subtracting $\tilde{\mu}_0$ from each sample value ($d_i = x_i - \tilde{\mu}_0$), discarding all 0 differences
  - The remaining differences are ranked by the absolute value (without regard to their signs) in ascending order
  - When the absolute value of two or more differences is the same, assign to each the average of the ranks (e.g. 5.5 to the 5th and 6th differences with equal absolute value)
- If null hypothesis $\tilde{\mu} = \tilde{\mu}_0$ is true, the sum of ranks corresponding to positive differences (w_+) should nearly equal the sum of ranks corresponding to negative differences (w_-).
- Alternatively, w_+, w_- or MIN(w_+, w_-) would be a very small value

11.3. Signed-Rank Test

- Example: 11 data represent number of hours that a rechargeable hedge trimmer operates before a recharge is required: 1.5, 2.2, 0.9, 1.3, 2.0, 1.6, 1.8, 1.5, 2.0, 1.2, 1.7.

– Revisit this by signed-rank test, at 0.05 level of significance, the hypothesis that this trimmer operates a median of 1.8 hours before requiring a recharge. (*Example 16.3*)

1) H_0: $\tilde{\mu} = 1.8$, 2) H_1: $\tilde{\mu} \neq 1.8$, 3) $\alpha = 0.05$.

4) Critical region: $w = \mathrm{MIN}(w_+, w_-) \leq 8$ (from *Statistical Table*) when n=10 (after discarding 0) for two-sided $\alpha = 0.05$

5) Computation: Subtracting 1.8 from each measurement and rank

Data	1.5	2.2	0.9	1.3	2.0	1.6	1.8	1.5	2.0	1.2	1.7
d_i	−0.3	0.4	−0.9	−0.5	0.2	−0.2	0	−0.3	0.2	−0.6	−0.1
Rank	5.5	7	10	8	3	3	-	5.5	3	9	1

$w_+ = 7+3+3 = 13$ and $w_- = 42$, so $w = \mathrm{MIN}(w_+, w_-) = 13 > 8.$

6) Decision: **Do not reject H_0**. Insufficient evidence to conclude that the median operating time is significantly different from 1.8 hours

11.3. Signed-Rank Test

- For alternative hypothesis $\tilde{\mu} < \tilde{\mu}_0$, $\tilde{\mu} > \tilde{\mu}_0$, and $\tilde{\mu} \neq \tilde{\mu}_0$, the test statistic would be respectively the sum of ranks of
  - the positive differences (w_+) $\rightarrow H_1$: $\tilde{\mu} < \tilde{\mu}_0$ if w_+ is too small
  - the negative differences (w_-) $\rightarrow H_1$: $\tilde{\mu} > \tilde{\mu}_0$ if w_- is too small
  - the smaller value of the two (let w = MIN(w_+, w_-)) $\rightarrow H_1$: $\tilde{\mu} \neq \tilde{\mu}_0$ if w is too small

H_0	H_1	Test statistic
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	w_+
	$\tilde{\mu} > \tilde{\mu}_0$	w_-
	$\tilde{\mu} \neq \tilde{\mu}_0$	w=MIN(w_+, w_-)

- Look up *Statistical Table: Critical Values for Signed-Rank Test* for $5 \leq n \leq 30$, level of significance α = 0.01, 0.025, 0.05 for a one-tailed test or α = 0.02, 0.05, 0.10 for a two-tailed test
- Reject H_0 if the relevant w_+, w_-, or w (= MIN(w_+, w_-)) **≤** the critical value

11.3. Signed-Rank Test

- Exercise 3: The following represents the time, in minutes, that patients waited during 12 visits before being seen by the doctor: 17, 15, 20, 20, 32, 28, 12, 26, 25, 25, 35, 24. Revisit this by signed-rank test, at 0.05 level of significance, that the median waiting time for the patients is not more than 20 minutes

←10

1) H_0: $\tilde{\mu} = 20$, 2) H_1: $\tilde{\mu} > 20$, 3) $\alpha = 0.05$.
4) Critical region: $w_- \le 11$ (from *Statistical Table*) when n=10 (after discarding 0) for one-sided $\alpha = 0.05$
5) Computation: Replacing each value by "+" if > 20, "–" if < 20 and discarding the measurements = 20, we obtain n = 10

Data	17	15	20	20	32	28	12	26	25	25	35	24
d_i	−3	−5	0	0	+12	+8	−8	+6	+5	+5	+15	+4
Rank	1	4	-	-	9	7.5	7.5	6	4	4	10	2

$w_- = 1+4+7.5 = 12.5 > 11.$

6) Decision: **Do not reject H_0**. Insufficient evidence to conclude that the median waiting time is more than 20 minutes

11.3. Signed-Rank Test

Signed-rank test for paired observations

- Two continuous symmetric populations are sampled with $\tilde{\mu}_1 = \tilde{\mu}_2$ for the paired observations.
- The test procedures for both the single- and paired-sample cases are summarised in table.

H_0	H_1	Test statistic
$\tilde{\mu} = \tilde{\mu}_0$	$\tilde{\mu} < \tilde{\mu}_0$	w_+
	$\tilde{\mu} > \tilde{\mu}_0$	w_-
	$\tilde{\mu} \neq \tilde{\mu}_0$	$w = \mathrm{MIN}(w_+, w_-)$
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	w_+
	$\tilde{\mu}_1 > \tilde{\mu}_2$	w_-
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	$w = \mathrm{MIN}(w_+, w_-)$

- For $n < 5$, the level of significance $\alpha < 0.05$ for a 1-tailed test or $\alpha < 0.10$ for a 2-tailed test, all values of w_+, w_-, or w will not lead to rejection of null hypothesis H_0
- Look up *Statistical Table: Critical Values for Signed-Rank Test* for $5 \leq n \leq 30$, level of significance $\alpha = 0.01, 0.025, 0.05$ for a one-tailed test or $\alpha = 0.02, 0.05, 0.10$ for a two-tailed test
- Reject H_0 if the relevant w_+, w_-, or w **≤** the critical value

11.3. Signed-Rank Test

- Example: 20 students are divided into 10 pairs. Sample problems with answers are provided at random to one member of each pair 1 week prior to the exam. Based on the exam results below, at 0.05 level of significance, test that providing sample problems would increase students' exam scores by as much as 50 points. (*Example 16.4*)

Pair	1	2	3	4	5	6	7	8	9	10
Give sample in advance	531	621	663	579	451	660	591	719	543	575
Without giving sample	509	540	688	502	424	683	568	748	530	524
d_i	22	81	-25	77	27	-23	23	-29	13	51
d_i - 50	−28	31	−75	27	−23	−73	−27	−79	−37	1
Rank	5	6	9	3.5	2	8	3.5	10	7	1

- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ represent the median scores of all students taking the exam with and without sample problems, respectively

1) H_0: $\tilde{\mu}_1 - \tilde{\mu}_2 = 50$ 2) H_1: $\tilde{\mu}_1 - \tilde{\mu}_2 < 50$ 3) $\alpha = 0.05$.
4) Critical region: $w_+ \le 11$ when n = 10 for one-sided $\alpha = 0.05$.
5) Computation: $w_+ = 6 + 3.5 + 1 = 10.5 < 11$
6) Decision: **Reject H_0**. We conclude that sample problems do not, on average, increase one's exam score by as much as 50 points

11.3. Signed-Rank Test

- Exercise 4: Two stories were read to each child. The 1st story was read, and the 2nd was read but also illustrated with pictures. An expert listened to how the children retold the stories and assigned a score. Test, at the 0.05 level of significance, that illustrations (2nd story) improve how the children retell a story (i.e. higher scores).

Child	1	2	3	4	5
Score of retelling 1st story	0.40	0.72	0.00	0.36	0.55
Score of retelling 2nd story	0.77	0.49	0.66	0.28	0.38
d_i	−0.37	0.23	−0.66	0.08	0.17
Rank	4	3	5	1	2

- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ represent the median scores of all students retelling the 1st story and that of retelling the 2nd story, respectively

1) H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$ 2) H_1: $\tilde{\mu}_1 < \tilde{\mu}_2$ 3) $\alpha = 0.05$.
4) Critical region: $w_+ \leq 1$ when $n = 5$ for one-sided $\alpha = 0.05$.
5) Computation: $w_+ = 3 + 1 + 2 = 6 > 1$
6) Decision: **Do not reject H_0**. There is insufficient evidence in this study to conclude that illustrations improve how the children retell a story

11.3. Signed-Rank Test

- Exercise 5: 16 cars are driven over a prescribed test course, first equipped with regular tyres and then with radial tyres. At 0.05 level of significance, test that cars equipped with radial tyres obtain higher fuel economy than those equipped with regular tyres, based on data below ←11

sample	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Radial tyre	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1	5.2	5.7	6.9	6.8	4.9
Regular tyre	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6.0	4.9	5.3	6.5	7.1	4.8
d_i	0.1	-0.2	0.4	0.1	-0.1	0.1		0.2	0.5		0.1	0.3	0.4	0.4	-0.3	0.1
Rank	3.5	7.5	12	3.5	3.5	3.5		7.5	14		3.5	9.5	12	12	9.5	3.5

- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ represent the median fuel economy of all cars with regular tyres and that with radial tyres, respectively

1) H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$ 2) H_1: $\tilde{\mu}_1 > \tilde{\mu}_2$ 3) $\alpha = 0.05$.
4) Critical region: $w_- \leq 26$ when $n = 14$ (after discarding equal values) for one-sided $\alpha = 0.05$.
5) Computation: $w_- = 7.5 + 3.5 + 9.5 = 20.5 < 26$
6) Decision: **Reject H_0**. We conclude that the use of radial tyres improves fuel economy.

11.3. Signed-Rank Test

Signed-Rank Test for Large Samples

- When the sample size is relatively large (n>15), the sampling distribution of W+ (or W-) approaches normal distribution, with

  $\mu_{W+} = n(n+1)/4$

  $\sigma_{W+}^2 = n(n+1)(2n+1)/24$

- when n is greater than the largest value in *Statistical Table* the statistic

  $Z = (W_+ - \mu_{W+})/\sigma_{W+}$

  can be used to determine the *critical region* for z-test

11.4. Rank-Sum Test

- **Wilcoxon rank-sum test** (a.k.a. Wilcoxon-Mann-Whitney test, *U*-test) can be an alternative to the two-sample *t*-test
- Assumptions:
  - underlying populations are continuous
  - independent random samples (n_1, n_2), without pairing
- Compare positions (population median) of two distributions. H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$. Alternative H_1: $\tilde{\mu}_1 < \tilde{\mu}_2$, $\tilde{\mu}_1 > \tilde{\mu}_2$, or $\tilde{\mu}_1 \neq \tilde{\mu}_2$
- Procedure:

– Let n_1 and n_2 be the number of observations in the smaller sample and in the larger sample respectively.

a) Arrange the $n_1 + n_2$ observations of the combined samples in ascending order, and assign a rank of 1, 2, . . . , ($n_1 + n_2$) for each observation (1=smallest); if tie (identical observations), assign an average ranks (e.g. 7.5 to each if 7th and 8th observations equal)

11.4. Rank-Sum Test

- Procedure *(continued)*

– Let w_1 = sum of ranks for n_1's observations

w_2 = sum of ranks for n_2's observations

$w_1+w_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2$ (i.e. the sum of $1,2,\ldots,n_1+n_2+1$)

hence, $w_2 = (n_1 + n_2)(n_1 + n_2 + 1)/2 - w_1$

b) Compute *Test statistic*

- $u_1 = w_1 - n_1(n_1 + 1)/2$
- $u_2 = w_2 - n_2(n_2 + 1)/2$
- $u = \mathrm{MIN}(u_1, u_2)$

H_0	H_1	Test statistic
$\tilde{\mu}_1 = \tilde{\mu}_2$	$\tilde{\mu}_1 < \tilde{\mu}_2$	u_1
	$\tilde{\mu}_1 > \tilde{\mu}_2$	u_2
	$\tilde{\mu}_1 \neq \tilde{\mu}_2$	$u=\mathrm{MIN}(u_1,u_2)$

- U_1 and U_2 have symmetric sampling distributions and assume values from 0 to (n_1n_2) such that $u_1 + u_2 = n_1n_2$

c) Look up *Critical Values for the Wilcoxon Rank-Sum Test* for $n_1 \leq 20$, $n_2 \leq 20$, level of significance $\alpha = 0.001, 0.01, 0.025, 0.05$ for one-tailed test or $\alpha = 0.002, 0.02, 0.05, 0.10$ for two-tailed test

- Reject H_0 if the relevant u_1, u_2, or u **≤** the critical value

11.4. Rank-Sum Test

- Example: The nicotine content of two brands of cigarettes, measured in milligrams, was found to be as follows:

Brand *A*	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3		
Brand *B*	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

– Test the hypothesis, at the 0.05 level of significance, whether the median nicotine contents of Brands *A* and *B* are equal.

1) H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$ 2) H_1: $\tilde{\mu}_1 \neq \tilde{\mu}_2$ 3) $\alpha = 0.05$.

4) Critical region: $u \leq 17$ (from *Statistical Table*) when $\alpha = 0.05$ two-tailed, $n_1 = 8$, $n_2 = 10$

5) Computation: The observations are arranged in ascending order and ranks from 1 to 18 assigned.

Brand *A*	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3			w_1
Rank	4	10.5	18	14.5	13	9	16	8			93
Brand *B*	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4	w_2
Rank	12	1	7	6	10.5	17	2	5	3	14.5	78

11.4. Rank-Sum Test

- Example: *(continued)*

5) $u_1 = w_1 - n_1(n_1 + 1)/2 = 93 – (8)(9)/2 = 57$; $u_2 = w_2 - n_2(n_2 + 1)/ = 78 – (10)(11)/2 = 23 \rightarrow u = \text{MIN}(u_1, u_2) = 23$ (i.e. u_2) **>** *17*
6) Decision: **Do not reject H_0**. There is no significant difference in the median nicotine contents of the two brands of cigarettes.

Rank-Sum Test for Large Samples

- When samples are large (n_1>8 and n_2>8), the sampling distributions of U_1 (or U_2) approach normal distribution, with

  $\mu_{U1} = n_1 n_2/2$

  $\sigma_{U1}^2 = n_1 n_2(n_1+n_2+1)/12$

- when $n_2 > 20$, the largest value in *Statistical Table*, and $n_1 \geq 9$, the statistic

  $Z = (U_1 - \mu_{U1})/\sigma_{U1}$

  can be used to determine the *critical region* for *z*-test

11.4. Rank-Sum Test

- Exercise 6: 9 patients have reached an advanced stage of the disease. 5 patients receive a treatment and 4 do not. The survival times, in years, are shown below. Use rank-sum test, at 0.05 level of significance, to determine if the treatment is effective (i.e. having longer survival years)

No treatment	1.9	0.5	2.8	3.1		w_1
Rank	4	1	6	7		18
Treatment	2.1	5.3	1.4	4.6	0.9	w_2
Rank	5	9	3	8	2	27

- Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ represent the median survival times, in years, of the patients with no treatment and those receive a treatment, respectively

1) H_0: $\tilde{\mu}_1 = \tilde{\mu}_2$ 2) H_1: $\tilde{\mu}_1 < \tilde{\mu}_2$ 3) $\alpha = 0.05$.
4) Critical region: $u \leq 2$ when $\alpha = 0.05$ one-tailed, $n_1 = 4$, $n_2 = 5$
5) Computation: The observations are arranged in ascending order and ranks from 1 to 9 assigned.

   $u_1 = w_1 - n_1(n_1 + 1)/2 = 18 - (4)(5)/2 = 8 > 2$
6) Decision: **Do not reject H_0**. There is insufficient evidence in this study to conclude that the treatment is effective.

11.5. Tests for Randomness: Runs Test

- Need for randomness
  - Assumption of randomness is essential for statistical inference
  - Not sure whether a given sample is random, as samples may not be random due to:
  - Changes in underlying distribution, e.g. a non-stable process
  - Violation of the independence rule, e.g. improper selection of sampling units
- Nature of randomness
  - Random process: a process that generates independent, identically-distributed (iid) observations
  - A random process would generate random samples
- Test for Randomness
  - Graphical plots, e.g. scatter plot; time series
  - Statistical tests, e.g. runs test

11.5. Tests for Randomness: Runs Test

- Based on the order that sample observations are obtained, a **runs test** is to test H_0 that observations are drawn at random.
- Illustration: suppose 12 people are polled to find out if they use a certain product. The randomness of the sample would be questioned if all 12 people were of the same sex.
- Designate a male and a female by symbols M and F, respectively, and record the outcomes according to their sex in the order in which they occur. A typical sequence for the experiment might be

$$\underbrace{M \quad M} \quad \underbrace{F \quad F \quad F} \quad \underbrace{M} \quad \underbrace{F \quad F} \quad \underbrace{M \quad M \quad M \quad M} \qquad v = 5 \text{ runs}$$

- A run is a subsequence of one or more identical symbols representing a common property of the data, e.g. the above sequence of 12 people comprises 5 runs
- If the number of runs is larger or smaller than what we would expect by chance, the hypothesis that the sample was drawn at random should be rejected.

11.5. Tests for Randomness: Runs Test

- Example: whether 12 people rolled as follows are random
  $M\ M\ F\ F\ F\ M\ F\ F\ M\ M\ M\ M$ $\quad v = 5$ runs

1) H_0: sequence is random; 2) H_1: not random; 3) $\alpha = 0.05$.
4) Test statistic: V = the total number of runs.
5) Computation: $n_1 = 5$, $n_2 = 7$, $v^* = 5$.
   $P = 2P(V \leq 5$ when H_0 is true$) = 2(0.197)$ (from *Statistical Table*) = 0.394 **>** 0.05
6) Decision: **Do not reject H_0.** There is insufficient evidence to reject the hypothesis of randomness in our sample.

- Example: whether 12 people rolled as follows are random
  $\underbrace{M\ M\ M\ M\ M\ M\ M}\ \underbrace{F\ F\ F\ F\ F}$ $\quad v = 2$ runs

5) Computation: $n_1 = 5$, $n_2 = 7$, $v^* = 2$.
   $P = 2P(V \leq 2$ when H_0 is true$) = 2(0.003) = 0.006 < 0.05$
6) Decision: **Reject H_0** and conclude that the sample is not random.

11.5. Tests for Randomness: Runs Test

- A sequence is limited to two distinct symbols: male or female; defective or non-defective; heads or tails; above (giving '+') or below (giving '-') the median (discarding measurements equal to median); etc. whether they are qualitative or quantitative data.
- Let n_1 be the number of symbols associated with the category that occurs the least and n_2 be the number of symbols that belong to the other category. Then the sample size $n = n_1 + n_2$.
- A runs test is, based on number of runs V, test for 2 types of departures from randomness
  – persistence (e.g. trend); infrequent changes results in very few runs (i.e. small V), e.g. *M M M M M M M F F F F F* (v = 2 runs)
  – overly frequent changes results in many runs (i.e. large V), e.g. *M F M F M F M F M F M F* (v = 12 runs)
- Look up *$P(V \leq v^*$ when H_0 is true) in the Runs Test* for $n_1 \leq 10$, $n_2 \leq 10$, and the number of runs $v^* \leq 20$. Reject H_0 if the P-value < α (the pre-determined level of significance)

11.5. Tests for Randomness: Runs Test

- Exercise 7: Does the amount of paint being dispensed by a machine vary randomly if the contents of 15 containers are measured and found to be 3.6, 3.9, 4.1, 3.6, 3.8, 3.7, 3.4, 4.0, 3.8, 4.1, 3.9, 4.0, 3.8, 4.2, and 4.1 litres? Use a 0.1 level of significance.

1) H_0: the sequence is random; 2) H_1: not random; 3) $\alpha = 0.1$.
4) Test statistic: V = the total number of runs.
5) Computation: median $\tilde{x}$=3.9. Assign + to values > 3.9 and – to values < 3.9, discard values = 3.9, hence the sequence

   $\underbrace{-}\ \underbrace{+}\ \underbrace{-\ -\ -\ -}\ \underbrace{+}\ \underbrace{-}\ \underbrace{+\ +}\ \underbrace{-}\ \underbrace{+\ +}$ $\quad n_1 = 6$, $n_2 = 7$, $v = 8$ runs.

   when v is large (e.g. > $(n_1+n_2)/2$), we'll test whether it's too large

   $P = 2P(V \geq 8$ when H_0 is true$) =$

   $= 2[1 - P(V \leq 7$ when H_0 is true$)] = 2(1-0.5) = 1$ **>** 0.1
6) Decision: **Do not reject H_0**. There is insufficient evidence to reject the hypothesis of randomness in our sample.

11.5. Tests for Randomness: Runs Test

- Exercise 8: A random sample of 15 adults living in a town were selected a survey. By letting M and F designate the gender of the selected person, the following sequence was obtained:

  $F\ F\ F\ M\ M\ M\ F\ M\ M\ F\ M\ F\ F\ F\ F$

– Use the runs test, at a 0.1 level of significance, to determine if the sequence of the sample was selected at random.

1) H_0: sequence is random; 2) H_1: not random; 3) $\alpha = 0.05$.
4) Test statistic: V = the total number of runs.
5) Computation: $n_1 = 6$, $n_2 = 9$.

   $\underbrace{F\ F\ F}\ \underbrace{M\ M\ M}\ \underbrace{F}\ \underbrace{M\ M}\ \underbrace{F}\ \underbrace{M}\ \underbrace{F\ F\ F\ F}$ $v = 7$ runs

   $P = 2P(V \le 7 \text{ when } H_0 \text{ is true}) = 2(0.343) = 0.686 > 0.1$
6) Decision: **Do not reject H_0**. There is insufficient evidence to reject the hypothesis of randomness in our sample.

11.5. Tests for Randomness: Runs Test

Runs Test for Large Samples

- When sample sizes are relatively large (n_1>10 and n_2>10), the sampling distribution of *V* approaches normal distribution, with

  $\mu_V = [2n_1n_2/(n_1+n_2)] + 1$

  $\sigma_V^2 = 2n_1n_2(2n_1n_2-n_1-n_2)/[(n_1+n_2)^2(n_1+n_2-1)]$

- when n_1 and n_2 are greater than the largest value in *Statistical Table* the statistic

  $Z = (V - \mu_V)/\sigma_V$

  can be used to determine the *critical region* for *z*-test

11.6. Summary of Non-Parametric Tests

- Non-parametric tests are (referring to "advantages")
  - Applicable to nominal data and ordinal scales
  - Valid for small samples
  - Fewer assumptions about the population (shape, variance)
- Power of non-parametric tests is lower
  - Parametric (*z*-, *t*-) tests should be used wherever applicable
  - In most cases, tests using different methods applied to the same data would lead to identical or similar conclusions
- Non-parametric tests are concerned with the median
  - If null hypothesis H_0 is true, then sum of ranks of one type of responses should nearly equal that of the other type
  - Reject H_0 if *P*-value < α, or if test statistic (w_+, w_-, w or u_1, u_2, u) $\leq$ the critical value
- With a relatively large sample size (e.g. n>10 or 15), the normal approximation can be used to run a *z*-test

11.6. Summary of Non-Parametric Tests

- Consider the problem in Exercise 1 & 3: Patients waited during 12 visits before being seen by the doctor: 17, 15, 20, 20, 32, 28, 12, 26, 25, 25, 35, 24 (minutes). Test, at 0.05 level of significance, that median/mean waiting time is not more than 20 minutes. ←10
- This can be *t*-test: single mean when variance σ^2 is unknown.

1. H_0: $\mu = 20 min$. 2. H_1: $\mu \neq 20 min$. 3. $\alpha = 0.05$.
4. Critical region: $t > 1.796$ *and* $t < -1.796$ ($t_{\alpha,n-1} = t_{0.05,(12-1)}$), where $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$ with 11 degrees of freedom
5. Computation: $\bar{x} = 23.25 min$, $s = 6.784$, and $n = 12$. Hence $t = (23.25-20)/(6.784/\sqrt{12}) = 1.66$ not > 1.796.
6. Decision: Do not reject H_0. There is insufficient statistical evidence. Note that one can estimate P-value using *t*-test.

- How about the problem in Exercise 2: A sample of 36 consumers were asked to taste two brands (Hoppy and Pokey Peanut Butter) and indicate a preference. Do the data indicate a significant difference in the consumer preferences for the two brands? ←13

Summary

- Non-parametric tests
  - distribution-free (hence comparing median instead of mean)
  - valid for small samples
  - applicable to nominal and ordinal scales

	Test Statistic	***Critical region & Statistical Table***
Sign test	*Binomial* random variable *Normal* approximation	$P(X \geq x$ or $\leq x$ when $p = 1/2) < \alpha$ $z = (x - 0.5n)/\sqrt{0.25n}$
Signed-Rank test	w_+, w_-, or w = MIN(w_+, w_-) where w is sum of ranks	w_+, w_-, or $w \leq$ *Critical Values for Signed-Rank Test*
Rank-Sum test	$u_1 = w_1 - n_1(n_1 + 1)/2$; $u_2 = w_2 - n_2(n_2 + 1)/2$; or u = MIN(u_1, u_2)	u_1, u_2, or $u \leq$ *Critical Values for the Wilcoxon Rank-Sum Test*
Runs test	Run (a subsequence of one or more identical symbols)	$P(V \leq v^*$ or $\geq v^*$ *when* H_0 *is true) in the Runs Test* $< \alpha$

Setting	***Non-parametric Tests***	***Tests assuming Normal***
One sample	Sign test, Signed-Rank test	z-test, t-test
Paired observations	Sign test, Signed-Rank test, applied to differences within pairs	t-test applied to differences within pairs
Two samples	Rank-Sum test	z-test, t-test

Assignments

11.1. The weights of 5 people before they stopped smoking and 5 weeks after they stopped smoking, in kilograms (kg), are as follows: (*16.10 on p.738*)

	Individual				
kg	1	2	3	4	5
Before	66	80	69	52	75
After	71	82	68	56	73

- Use the signed-rank test for paired observations to test the hypothesis, at the 0.05 level of significance, that giving up smoking has no effect on a person's weight against the alternative that one's weight increases if he quits smoking.

11.2. A random sample of 15 adults living in a small town were selected to estimate the proportion of voters favouring a certain candidate for mayor. Each individual was also asked if he or she was a college graduate. By letting *Y* and *N* designate the responses of "yes" and "no" to the education question, the following sequence was obtained: (*16.23 on p.751*)

N N N N Y Y N Y Y N Y N N N N

- Use the runs test at 0.1 level of significance to test if the sequence supports the contention that the sample was selected at random.

Notes: (1) Assignments are not compulsory but will be reviewed if submitted within 2 weeks
(2) It is encouraged to complete assignments using calculator (without using Excel or statistics software)