Data transformations

Data transformations

- A transformation is a change of numerical scale.
- Transformations are a remedy for failures of normality, homoscedasticity, linearity, and outliers.
- The scale of the data influences the utility of transformations.
  - ✓ If the scale is arbitrary transformations are more effective,
  - ✓ If the scale is meaningful the difficulty of interpretation increases.

Logarithmic transformation

- Factor effects are multiplicative rather than additive
- Variance to be proportional to the square of the mean (i.e. the standard deviation is proportional to the mean)

$$x' = \log(x)$$

$$x' = \log(x + 1)$$ (Bartlett 1947)

Logarithmic transformation example

A hypothetical two-way analysis of variance design, where the effects of the factors are additive (data are in grams).

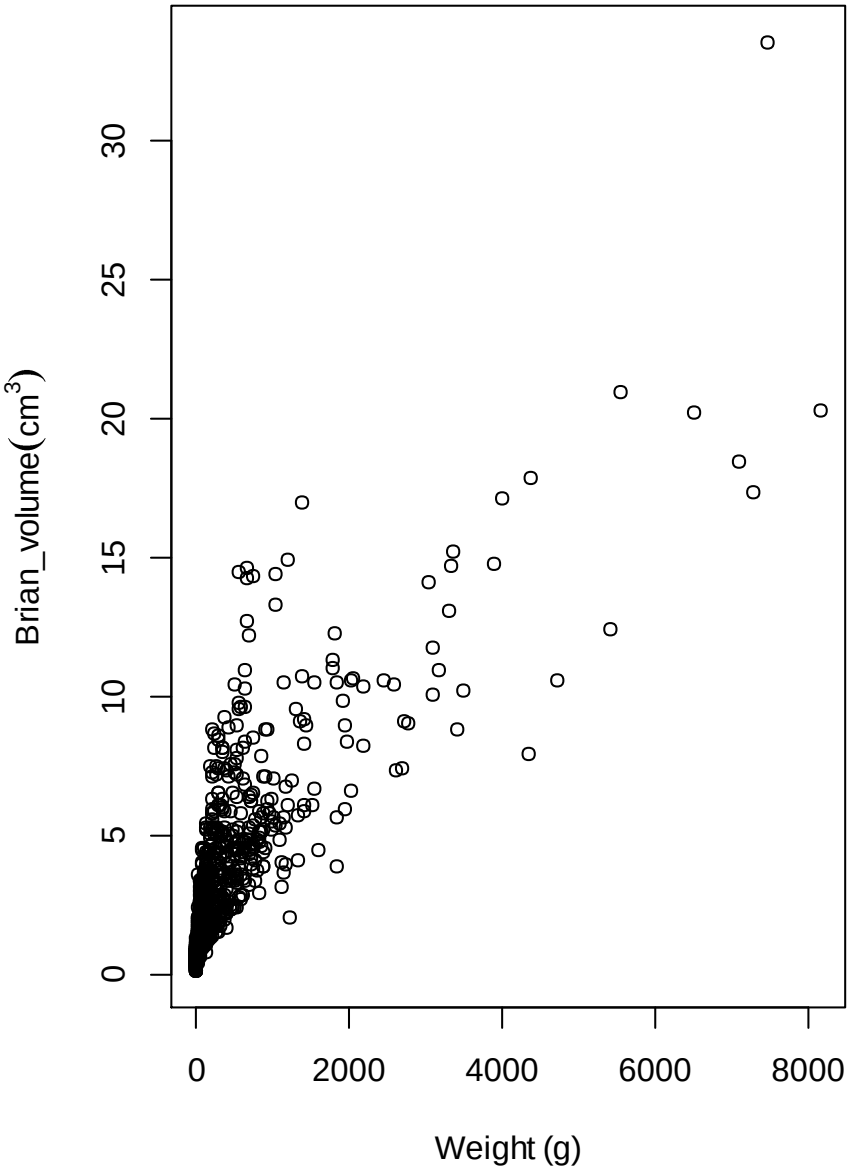
	Factor A		
Factor B	Level 1	Level 2	Level 3
Level 1	10	20	25
Level 2	20	30	35

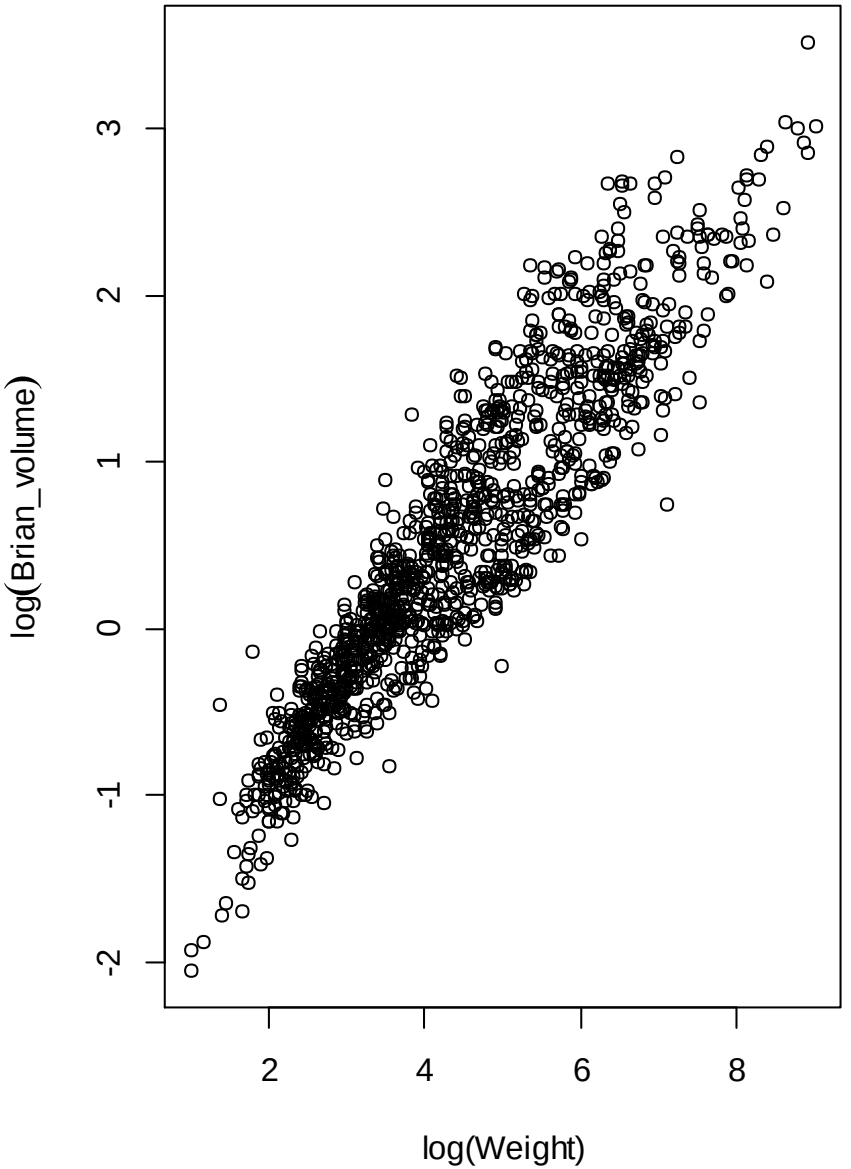
A hypothetical two-way analysis of variance design, where the effects of the factors are multiplicative (data are in grams).

	Factor A		
Factor B	Level 1	Level 2	Level 3
Level 1	10	30	60
Level 2	20	60	120

Logarithmic transformation example: brain volume and weight of 1217 birds

Date from Vincze 2016 (Evolution 70-9: 2123–2133)





Square root transformation

Variance are proportional to the mean.

Poisson distributed data are counted in time and/or space, the mean is equal to the variance.

$$x' = \sqrt{x + 0.5}$$ (Bartlett 1936)

$$x' = \sqrt{x} + \sqrt{x + 1}$$ (Freeman and Tukey 1950)

$$x' = \sqrt{x + 0.375}$$ (Anscombe1948; Kihlberg et al. 1972)

Sometimes a log transformation has too great effect, making the distribution negatively skew, and so the square root of the data was used.

Reciprocal transformation

The standard deviations of groups of data are proportional to the square of the means of the groups (Zar p280).

$$x' = \frac{1}{x}$$

For count data:

$$x' = \frac{1}{x+1}$$

- population density (people per unit area) becomes area per person
- persons per doctor becomes doctors per person
- rates of erosion become time to erode a unit depth

Square transformation

The standard deviations decrease as the group mean increase, and/or if the distribution is skewed to the left.

$$x' = x^2$$

Arcsine transformation

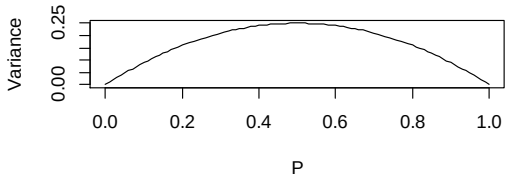
The arcsine transformation is a transformation that assists a data analyst when working with proportions and percentages.

The proportion p can be made nearly "normal" if the square root of p is used with the arcsine (or inverse sin or sin^-1) transformation.

The arcsine transformation is then computed as a function of the proportion, p.

$$p' = \arcsin\sqrt{p}$$

Why we need arcsine transformation



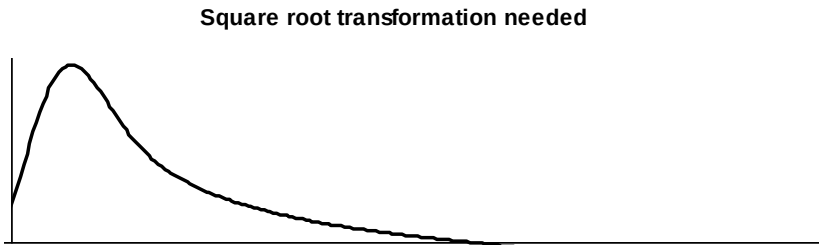
The variance for the binomial distribution of the observed proportion, p=y/n, is a function of p: var(p) = (p)*(1-p)/(n-1).

The fact that the variance of p depends on its particular value violates the homogeneity of variance assumption across subjects required for the computation of statistical tests, for example, if p were to be used as the dependent variable in an ANOVA or regression.

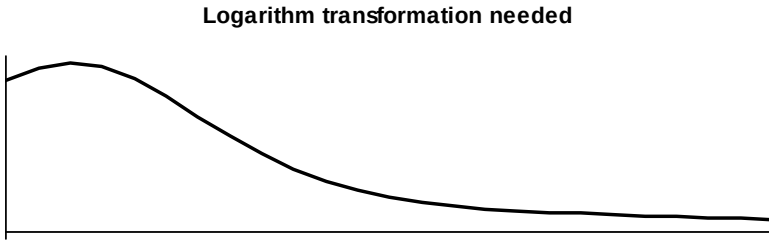
If most of the computed proportions lie between 0.3 and 0.7, this transformation have only a small difference in the analytical results. However, it is wise to use it, especially when a sizeable number of the observed proportions are either relatively small (i.e., $0 < p < 0.2$) or large (i.e., $0.8 < p < 1.0$).

Note: the arsine transformation is not particularly good if a substantial number of the proportions are equal to 0 or 1 or for values at the extreme ends of the possible range of p (near 0 and near 1). It is also not recommended if the number of trials, i.e., n, is small.

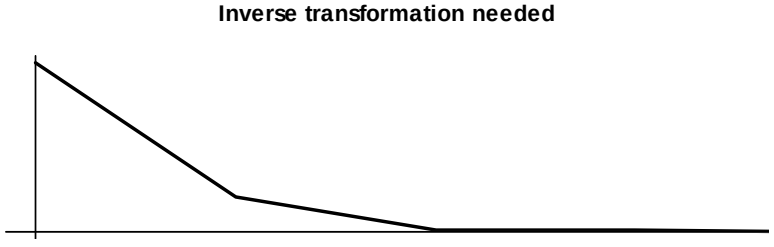
Data transformations




```
par(mfrow=c(3,1))

plot(density(runif(100000,1,100)^2), xlim=c(.5,5000),
ylim=c(0.0001,0.0003), main='Square root transformation needed',
xlab='', ylab='', lwd=2, axes=F)
abline(v = -200)
abline(h = 0.000092, xpd=T)
```




```
plot(density(exp(runif(10000,0,10))), xlim=c(.5,1000),
main='Logarithm transformation needed', xlab='', ylab='', lwd=2,
axes=F)
abline(v=-40)
abline(h=0)
```




```
plot(density(1/(runif(10000,0,10))), xlim=c(0,10),
main='Inverse transformation needed', xlab='', ylab='', lwd=2,
axes=F)
abline(v=0)
abline(h=0)
```

Summary

Variance-stabilizing transformations

Relationship of σ^2 to $E(y)$	Transformation
$\sigma^2 \propto constant$	No transfomation
$\sigma^2 \propto E(y)$	$\sqrt{y}$ (Poisson data)
$\sigma^2 \propto E(y)[1-E(y)]$	$\sin^{-1}(\sqrt{y})\ (0 \le y \le 1)$
$\sigma^2 \propto [E(y)]^2$	$\log(y)$
$\sigma^2 \propto [E(y)]^3$	$y^{-1/2}$, or $\log(y)$
$\sigma^2 \propto [E(y)]^4$	$\frac{1}{y}$, or $\log(y)$

Data not suitable for transformation

- Having very long tails at both ends of the distribution
- Having a bimodal distribution
- Having a large number of identical observations
- Transformation leading to variation in *p*-value

Box-Cox transformation

In many cases of using parametric statistics, transformations are often used without any particular justification!

George E. P. Box and David Cox developed a general method for finding out the best transformation to use on a set of data to achieve a normal distribution.

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y_i), & \text{if } \lambda = 0 \end{cases}$$

Box, George E. P. and Cox, D. R. 1964. An analysis of transformations. Journal of the Royal Statistical Society, Series B 26 (2): 211–252.

For example, if lambda = 1, there is no transformation; when lambda = 0.5, you get a square root transformation; and when lambda = -1, you get a reciprocal transformation.

Box-Cox transformation

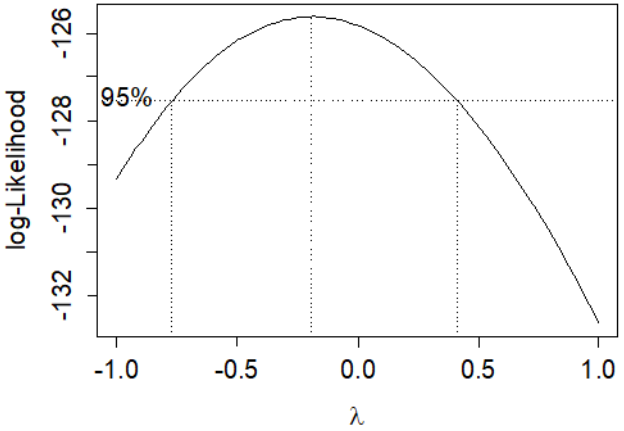
- Choose the value of λ that maximizes the log-likelihood function:

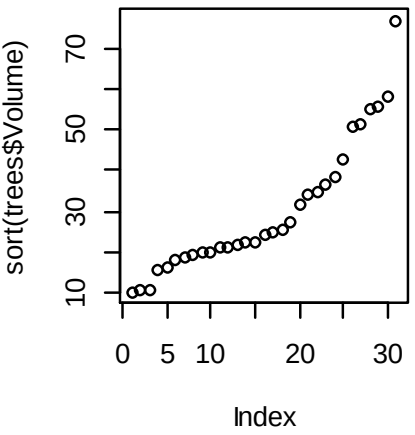
$$L = -\frac{V}{2}\log_e s_T^2 + (\lambda - 1)\frac{V}{n}\sum(\log_e y)$$

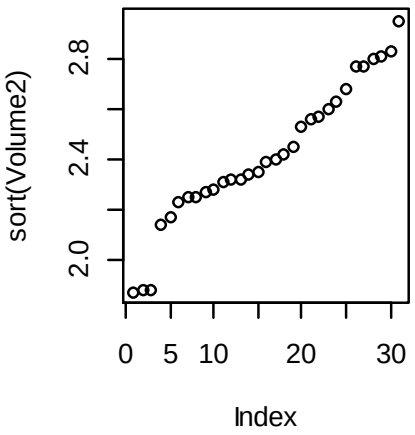
- L = Value of log-likelihood
- v = degrees of freedom (n - 1)
- s_T^2 = Variance of transformed y-values
- λ = Provisional estimate of power transformation parameter
- y = Original data values

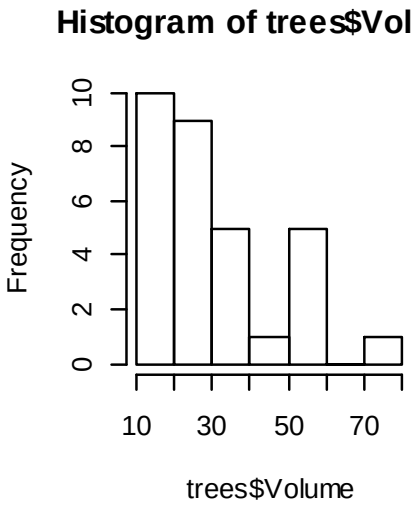
Box-Cox transformation

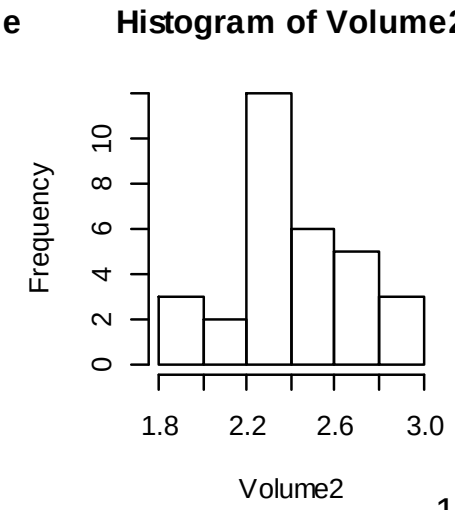
```
# Box-Cox transformation
library(MASS)
# lambda-likelyhood function
tr <- boxcox(Volume ~ Height, data=trees,
             lambda = seq(-1, 1, length=10))
lambda <- tr$x[ tr$y == max(tr$y) ]; lambda # -0.19
library(car)
Volume2 = bcPower(trees$Volume, lambda) # Box-Cox transformation
par(mfrow=c(1,4))
plot(sort(trees$Volume)); plot(sort(Volume2))
hist(trees$Volume); hist(Volume2)
```











Nonparametric Statistics

- Sign test
- Wilcoxon signed rank test
- Wilcoxon rank sum test
- Kruskal-Wallis test
- Friedman’s Test
- Bootstrapping

History

In 1940s, Frank Wilcoxon, a chemist at American Cyanamid, was bothered by outliers. He had been running hypothesis tests comparing the effects of different treatments, using Student's t-tests and Fisher's analyses of variance.

He developed a statistic based on ranked values (calculations based on combinations and permutations of the observed numbers). He submitted a paper to the journal Biometrics. The referees and editors determined this was original work. No one had ever thought of this before, and his paper was published in 1945.

An economist named Henry B. Mann and a graduate student in statistics at Ohio State University named D. Ransom Whitney were working on a related problem.

They were trying to order statistical distributions so that one might say, in some sense, that the distribution of wages in the year 1940 was less than the distribution of wages in 1944. They came up with a method of ordering that involved a sequence of simple but tedious counting methods. This led Mann and Whitney to a test statistic whose distribution could be computed from combinatoric arithmetic-the same type of computation as Wilcoxon's. They published a paper describing their new technique in 1947.

What is a parameter in statistics?

- Most statistical tests (e.g. the general linear models), assume some kind of underlying distribution, like the normal distribution.
- If you know the mean and the standard deviation of a normal distribution then you know how to calculate probabilities.
- Means and standard deviations are called *parameters;* all theoretical distributions have parameters.
- **Statistical tests that assume a distribution and use parameters are called *parametric tests.***
- **Statistical tests that don't assume a distribution or use parameters are called *nonparametric tests.***

Why we use a nonparametric test?

- Although many things in nature are normally distributed, **some are not**. In those cases using a t-test, for example, could be inappropriate and misleading.

- Nonparametric tests have fewer assumptions or restrictions on the data

- Examples that not normally distributed:
  - Nominal data: race, sex,
  - Ordered categorical data: mild, moderate, severe
  - Likert scales: strongly disagree, disagree, no opinion, agree, strongly agree

How do nonparametric tests work?

- Most nonparametric tests use ***ranks*** instead of raw data for their hypothesis testing.

  Example: use rank comparing test scores between girls and boys

- Null hypothesis: medians are equal

Advantages of Nonparametric Tests

- Used with all scales
- Easier to compute
  - Developed originally before wide computer use
- Make fewer assumptions
- Need not involve population parameters
- Results may be as exact as parametric procedures

Disadvantages of Nonparametric Tests

- May waste information

  If data permit using parametric procedures

  Example: converting data from ratio to ordinal scale

- P values (tables) not widely available

Summary of nonparametric test

- Rank data without regard to groups
- Use sum (or other function) of group ranks to calculate a test statistic
- Look up p values in a table or with a computer
- Reject or fail to reject null hypothesis

Sign test

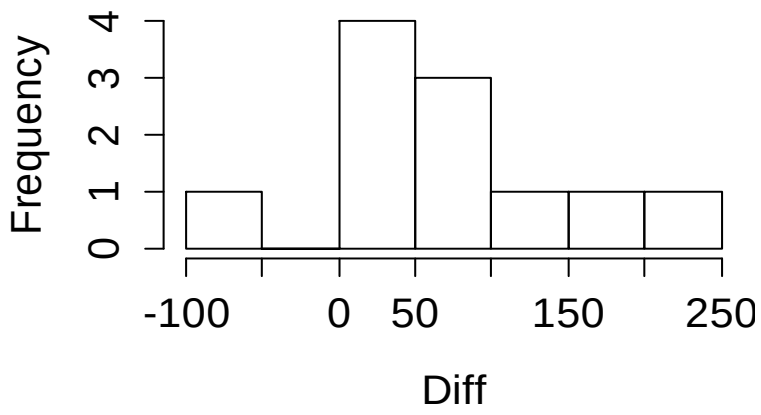
- If there was truly no effect of the therapy, we would assume that there would be an equal number of + and – signs.
- Is there a significant difference between the two groups (+ and -)? How can we calculate the p-value?

Patient	Pre	Post	Diff	Sign
1	360	223	137	+
2	216	149	67	+
3	427	224	203	+
4	217	181	36	+
5	613	708	-95	-
6	245	197	48	+
7	371	303	68	+
8	236	168	68	+
9	421	312	109	+
10	677	521	156	+
11	218	202	16	+

Distribution of the differences

- Since we have paired data, we could use the paired t-test.
- What can you say about the distribution of the differences?
- Does the normality assumption of the paired t-test seem appropriate?
- The difference may contain outliers

Patient	Pre	Post	Diff	Sign
1	360	223	137	+
2	216	149	67	+
3	427	224	203	+
4	217	181	36	+
5	613	708	-95	-
6	245	197	48	+
7	371	303	68	+
8	236	168	68	+
9	421	312	109	+
10	677	521	156	+
11	218	202	16	+

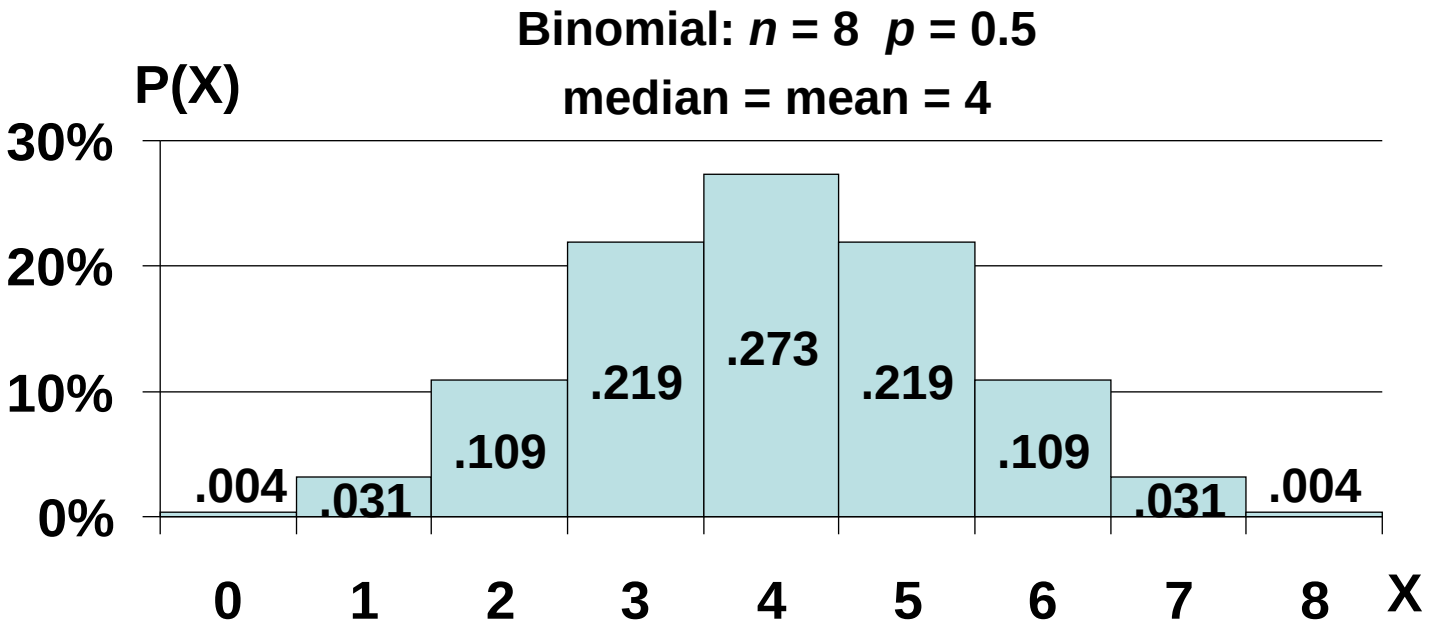


Sign Test

- Test the null hypothesis that the median of a distribution is equal to some value.
  - Similar situations like one-sample t-test or a paired t-test
  - Ordered data where a numerical scale is inappropriate

    (Note that the Wilcoxon Signed Rank Sum Test is also appropriate in these situations and is a more powerful test than the sign test.)
- Test Statistic: Number of sample values above (or below) median -- max(r−, r+).
- Use tables of the binomial distribution to find the probability of observing a value of r assuming p = 1/2. If the test is one-sided, this is your p-value. If the test is a two-sided test, double the probability obtained.
- Can use normal approximation if $n \geq 10$

Example: sign test uses p-value to make decision



P-value is the probability of getting an observation at least as extreme as we got. If 7 'Favor' H_a, then p-value = $P(x \geq 7)$ = .031 + .004 = .035. If α = .05, then reject H_0, since p-value $\leq \alpha$.

1-pbinom(6, 8, 0.5)

Hypothesis test

1) Paired data, alpha level=0.05
2) Hypotheses
   - H_0: median of differences = 0
   - H_A: median of differences != 0
3) Test statistic is 10+ signs
4) p-value=0.0005 **1-pbinom(10, 11, 0.5)**
5) Reject null hypothesis
6) Conclusion: There is a significant difference between the pre and post

Patient	Pre	Post	Diff	Sign
1	360	223	137	+
2	216	149	67	+
3	427	224	203	+
4	217	181	36	+
5	613	708	-95	-
6	245	197	48	+
7	371	303	68	+
8	236	168	68	+
9	421	312	109	+
10	677	521	156	+
11	218	202	16	+

R code for sign test

```
binom.test(10, 11, alternative = "greater")
```

```
data: 10 and 11
number of successes = 10, number of trials = 11, p-value = 0.005859
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval: 0.6356405 1.0000000.
```

```
binom.test(1, 11 , alternative = "less")
pbinom(10,    11, 0.5, lower.tail = F) # 0.000488
pbinom(10-1, 11, 0.5, lower.tail = F) # 0.005859
```

Wilcoxon signed rank test

- The sign test looks only at the sign of the differences, but the Wilcoxon signed rank test uses the sign and rank of the differences.

- The null and alternative hypotheses are the same as for the sign test
  - H_0: median diff = 0
  - H_A: median diff not = 0

Patient	Pre	Post	Diff	Rank
1	492	375	117	12
2	297	382	-85	-11
3	272	325	-53	-8
4	367	585	-218	-13
5	206	181	25	3
6	284	237	47	7
7	338	273	65	10
8	212	243	-31	-4
9	161	147	14	2
10	384	326	58	9
11	224	214	10	1
12	251	292	-41	-6
13	224	263	-39	-5

Formula

If n > 8, then the distribution of T_+ is approximately normal. Mean and variance :

$$\mu = \frac{n(n+1)}{4}$$

$$\sigma^2 = \frac{n(n+1)(2n+1)}{24}$$

Sum of T values: $$T_+ + T_- = \frac{n(n+1)}{2}$$

Test statistic W is: $$W = \frac{\min(T_+, T_-) - \mu}{\sigma}$$

Table: wilcoxon signed rank test (ci % = 95%)

Wilcoxon Signed-Ranks Test Critical values		
Number (n)	2 sided	1 sided
5	0	1
6	0	2
7	2	3
8	3	5
9	5	8
10	8	10
11	10	13
12	13	17
13	17	21
14	21	25
15	25	30
16	29	35
17	34	41
18	40	47
19	46	53
20	52	60
21	58	67
22	65	75
23	73	83
24	81	91
25	89	100

Critical values: *Wilcoxon Signed-Ranks test p=0.05 (CI% = 95%). Significant, if the calculated values presented in this table [the sum of the positive ranks or the negative ranks] is too small.*

R code - Wilcoxon signed rank test

- The test statistic of this test is the sum of the positive ranks.
- Under the null hypothesis, half of the ranks should be positive and half of the ranks should be negative. Evidence against the null would be having the sum of the positive ranks either being very high or very low.
- We can complete this test using R with the commands

```
# dependent 2-group Wilcoxon Signed Rank Test
pre  <- c(492,297, 272, 367, 206, 284, 338,
          212, 161, 384, 224, 251, 224)
post <- c(375, 382, 325, 585, 181, 237, 273,
          243, 147, 326, 214, 292, 263)
wilcox.test(pre, post, paired=T)
```

Output: Wilcoxon signed rank test

data: pre and post

V = 44, p-value = 0.946

alternative hypothesis: true mu is not equal to 0

Hypothesis test

- Paired data, Wilcoxon signed rank test, alpha=0.05
- Hypotheses
  - Null: median difference = 0
  - Alternative: median difference not = 0
- Test statistic: sum of positive ranks
- p-value = 0.946
- Fail to reject null hypothesis
- Conclusion: There is no evidence of a difference between the pre and post for patients.

Summary: Wilcoxon Signed Rank Test

- Tests probability distributions of 2 related populations
- Corresponds to t-test for dependent (paired) means
- Assumptions
  - random samples
  - both populations are continuous
- Can use normal approximation if $n \geq 25$

Example: Wilcoxon signed rank test

In a finance department, is the **new** financial package **faster** (**.05** level)? You collect the following data entry times:

User	Current	New
Mary	9.98	9.88
Nicole	9.88	9.86
Sam	9.90	9.83
Tom	9.99	9.80
Brian	9.94	9.87
Shawn	9.84	9.84

Wilcoxon signed rank test procedure

- Obtain difference scores, $D_i = X_{1i} - X_{2i}$
- Take absolute value of differences, D_i
- Delete differences with 0 value
- Assign ranks, R_i, where smallest = 1
- Add signs of differences to ranks
- Sum '+' ranks (T_+) & '-' ranks (T_-)
- Compute test statistic T or W

Signed rank test computation table

X_{1i}	X_{2i}	D_i	$\|D_i\|$	R_i	Sign	Sign R_i
9.98	9.88	+0.10	0.10	4	+	+4
9.88	9.86	+0.02	0.02	1	+	+1
9.90	9.83	+0.07	0.07	~~2~~ 2.5	+	+2.5
9.99	9.80	+0.19	0.19	5	+	+5
9.94	9.87	+0.07	0.07	~~3~~ 2.5	+	+2.5
9.84	9.84	0.00	0.00	...	...	Discard
Total					**$T_+ = 15, T_- = 0$**	

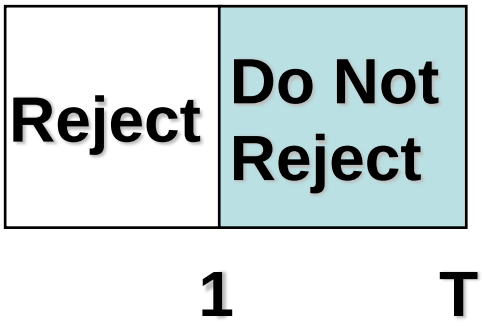
Signed rank test example

- H_0: Identical distribution.
- H_a: Current shifted right
- $\alpha = .05$
- n' = 5 (not 6; 1 elim.)
- Critical value(s):

Test statistic:

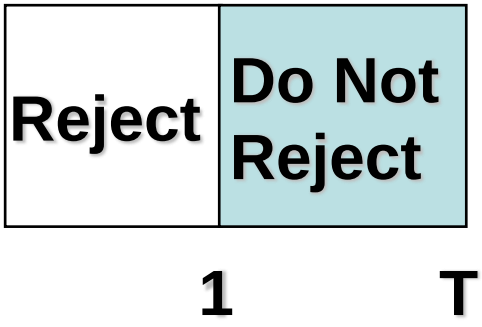
Decision:

Conclusion:



Signed rank test example

- H_0: Identical distribution.
- H_a: Current shifted right
- $\alpha = .05$
- n' = 5 (not 6; 1 elim.)
- Critical value(s):



Test statistic:

Since one-tailed test & current shifted right, use $T_- = 0$

Decision:

Reject at $\alpha = .05$

Conclusion:

There is evidence new package is faster

Sign test

- Less powerful
  - Less sensitive
  - Wider confidence intervals
- Uses less information
  - only sign of difference

Wilcoxon signed rank test

- More powerful
  - More sensitive
  - Narrower confidence intervals
- Uses more information
  - also size of difference

Comments

- When we have paired data and the assumptions of a paired t-test are not met, we have two ways to complete the hypothesis test
- The is always preferred over the sign test because it uses more of the data (since it uses the ranks). The Wilcoxon signed rank test has much more power to detect a significant difference.
- There is not a large loss of power in using a Wilcoxon signed rank test compared to a t-test when the normality assumption holds. The Wilcoxon is much more powerful when the normality assumption does not hold.
- Therefore, the Wilcoxon signed rank test is more appropriate if there is any reason to doubt the normality assumption.

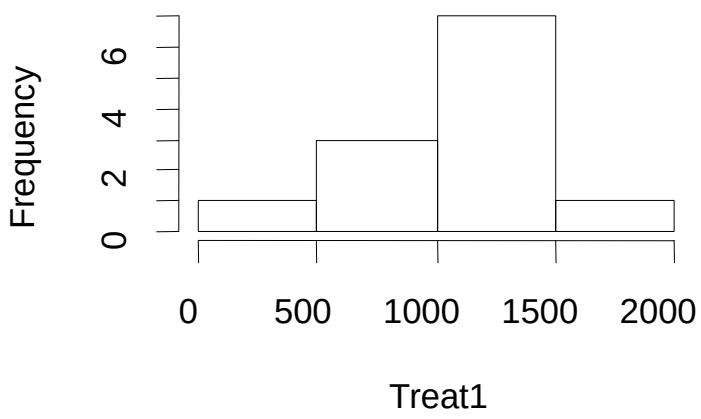
Wilcoxon rank sum test

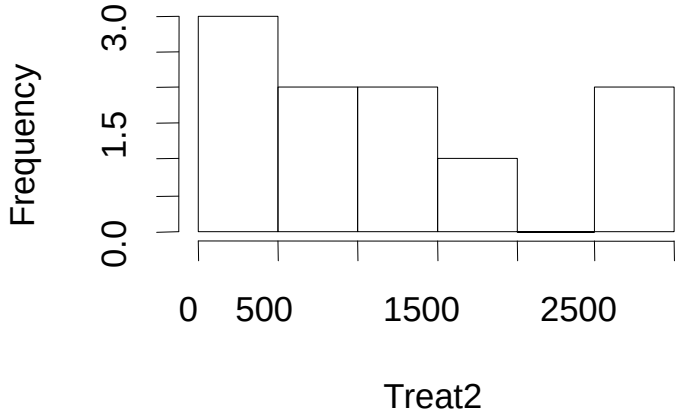
- Two independent samples

Treat1	Treat2
1011.07	496.44
1066.82	541.76
610.80	1562.01
1111.44	2515.12
955.68	1133.99
1203.84	300.33
1600.32	482.55
555.90	503.22
1302.95	2744.23
182.34	1232.22
1233.20	
1402.09	

Why use Wilcoxon rank sum test?

- We want to know if the cost in the two groups are the same.
- Since we have two independent samples, could use two-sample t-test
- Notice that the two graphs do not appear normal and have many outliers





Wilcoxon rank sum test

- Since we have two independent samples and the t-test is not appropriate, we need a nonparametric test.

- We are interested in the median rather than the mean.

- The hypothesis test of interest is
  - H_0: $median_{Treat1} = median_{Treat2}$
  - H_A: $median_{Treat1}$!= $median_{Treat2}$

Wilcoxon rank sum test

- We use the rank of the data points, rather than the actual values.

- Under the null hypothesis, the number of high and low ranks in each group should be equal. If the sum of the ranks in one group is very high or very low, this would be evidence against the null hypothesis.

Treat1	Rank	Treat2	Rank
1011.07	10	496.44	4
1066.82	11	541.76	6
610.80	8	1562.01	19
1111.44	12	2515.12	21
955.68	9	1133.99	13
1203.84	14	300.33	2
1600.32	20	482.55	3
555.90	7	503.22	5
1302.95	17	2744.23	22
182.34	1	1232.22	15
1233.20	16		
1402.09	18		

Lower tail

n_A	n_B	0.005	0.01	0.025	0.05	0.10	0.20
4	4	.	.	10	11	13	14
4	5	.	10	11	12	14	15
4	6	10	11	12	13	15	17
4	7	10	11	13	14	16	18
4	8	11	12	14	15	17	20
4	9	11	13	14	16	19	21
4	10	12	13	15	17	20	23
4	11	12	14	16	18	21	24
4	12	13	15	17	19	22	26
5	5	15	16	17	19	20	22
5	6	16	17	18	20	22	24
5	7	16	18	20	21	23	26
5	8	17	19	21	23	25	28
5	9	18	20	22	24	27	30
5	10	19	21	23	26	28	32
5	11	20	22	24	27	30	34
5	12	21	23	26	28	32	36
6	6	23	24	26	28	30	33
6	7	24	25	27	29	32	35
6	8	25	27	29	31	34	37
6	9	26	28	31	33	36	40
6	10	27	29	32	35	38	42
6	11	28	30	34	37	40	44
6	12	30	32	35	38	42	47
7	7	32	34	36	39	41	45
7	8	34	35	38	41	44	48
7	9	35	37	40	43	46	50
7	10	37	39	42	45	49	53
7	11	38	40	44	47	51	56
7	12	40	42	46	49	54	59
8	8	43	45	49	51	55	59
8	9	45	47	51	54	58	62
8	10	47	49	53	56	60	65
8	11	49	51	55	59	63	69
8	12	51	53	58	62	66	72
9	9	56	59	62	66	70	75
9	10	58	61	65	69	73	78
9	11	61	63	68	72	76	82
9	12	63	66	71	75	80	86
10	10	71	74	78	82	87	93
10	11	73	77	81	86	91	97
10	12	76	79	84	89	94	101
11	11	87	91	96	100	106	112
11	12	90	94	99	104	110	117
12	12	105	109	115	120	127	134

Upper tail

n_A	n_B	0.20	0.10	0.05	0.025	0.01	0.005
4	4	22	23	25	26		
4	5	25	26	28	29	30	
4	6	27	29	31	32	33	34
4	7	30	32	34	35	37	38
4	8	32	35	37	38	40	41
4	9	35	37	40	42	43	45
4	10	37	40	43	45	47	48
4	11	40	43	46	48	50	52
4	12	42	46	49	51	53	55
5	5	33	35	36	38	39	40
5	6	36	38	40	42	43	44
5	7	39	42	44	45	47	49
5	8	42	45	47	49	51	53
5	9	45	48	51	53	55	57
5	10	48	52	54	57	59	61
5	11	51	55	58	61	63	65
5	12	54	58	62	64	67	69
6	6	45	48	50	52	54	55
6	7	49	52	55	57	59	60
6	8	53	56	59	61	63	65
6	9	56	60	63	65	68	70
6	10	60	64	67	70	73	75
6	11	64	68	71	74	78	80
6	12	67	72	76	79	82	84
7	7	60	64	66	69	71	73
7	8	64	68	71	74	77	78
7	9	69	73	76	79	82	84
7	10	73	77	81	84	87	89
7	11	77	82	86	89	93	95
7	12	81	86	91	94	98	100
8	8	77	81	85	87	91	93
8	9	82	86	90	93	97	99
8	10	87	92	96	99	103	105
8	11	91	97	101	105	109	111
8	12	96	102	106	110	115	117
9	9	96	101	105	109	112	115
9	10	102	107	111	115	119	122
9	11	107	113	117	121	126	128
9	12	112	118	123	127	132	135
10	10	117	123	128	132	136	139
10	11	123	129	134	139	143	147
10	12	129	136	141	146	151	154
11	11	141	147	153	157	162	166
11	12	147	154	160	165	170	174
12	12	166	173	180	185	191	195

Wilcoxon Rank Sum Test table

Probabilities relate to the distribution of W_A, the rank sum for group A, when H_o : A = B is true.

The tabulated value for the lower tail is the largest value of w_A. The tabulated value for the upper tail is the smallest value of w_A.

For sample sizes (n_A & n_A) larger than 12 use Normal Approximation and the Standard Normal Table to calculate critical values.

Formula (Mann-Whitney *U*)

The Mann-Whitney *U* ranks all the cases from the lowest to the highest score. The "Mean Rank" is the mean of the those ranks for each group and the Sum of Ranks is the sum of those ranks for each group. U_1 is defined as the number of times that a score from group 1 is lower in rank than a score from group 2. U_2 is defined as the number of times that a score from group 2 is lower in rank that a score from group 1. *U* is defined as the smaller of U_1 or U_2. The computational formulas for U_1 and U_2 are as follows:

$$U_1 = n_1 n_2 + (n_1(n_1 + 1))/2 - R_1$$
$$U_2 = n_1 n_2 + (n_2(n_2 + 1))/2 - R_2$$

where
n_1 = number of observations in group 1
n_2 = number of observations in group 2
R_1 = sum of ranks assigned to group 1
R_2 = sum of ranks assigned to group 2

Mann-Whitney *U* table (two tails)

n_2	α	n_1																	
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	.01	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	.05	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	.01	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	.05	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	.01	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	.05	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	.01	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	.05	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	.01	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	.05	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	.01	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	.05	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	.01	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	.05	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	.01	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

R code

```
# Independent 2-group Mann-Whitney U Test
treat1 <- c(1011.07, 1066.82, 610.82, 1111.44, 955.68, 1203.84,
            1600.32, 555.90, 1302.95, 182.34, 1233.20, 1402.09)
treat2 <- c(496.44, 541.76, 1562.01, 2515.12, 1133.99, 300.64,
            482.52, 503.28, 2744.23, 1232.22)
wilcox.test(treat1, treat2, paired=F)
```

Output: Wilcoxon rank sum test

data: treat1 and treat2,

W = 65, p-value = 0.7713

alternative hypothesis: true mu is not equal to 0

Hypothesis test

- Two independent samples, Wilcoxon rank sum test, alpha=0.05
- Hypotheses
  - Null: $median_{Treat1} = median_{Treat2}$
  - Alternative: $median_{Treat1}$!= $median_{Treat2}$
- Test statistic: Sum of ranks in the smallest group
- p-value=0.77
- Fail to reject null hypothesis
- Conclusion: There is no evidence of a difference between treat 1 and treat 2.

Summary: Wilcoxon rank sum test Also called the Mann-Whitney test

- Tests two independent population probability distributions
- Corresponds to t-test for 2 Independent Means
- Can use normal approximation if $n \geq 10$

Assumptions

- Random samples are obtained from each population.
- The two samples are independent.
- The sample data is at least ordinal.
- The two populations differ only in location.

Another example

Wilcoxon rank sum test computation table

Factory 1		Factory 2	
Rate	Rank	Rate	Rank
71	1	85	5
82	~~3~~ 3.5	82	~~4~~ 3.5
77	2	94	8
92	7	97	9
88	6		
Rank Sum	19.5		25.5

wilcox.test(Factory1, Factory2, paired=F)

Wilcoxon rank sum test procedure

- Assign Ranks, R_i, to the $n_1 + n_2$ Sample Observations
  - ✓ If unequal sample sizes, let n_1 refer to smaller-sized sample
  - ✓ Smallest value = 1
  - ✓ Average ties
- Sum the ranks, T_i, for each sample
- Test statistic is T_A (smallest sample) (Wilcoxon W)
  - ✓ Null hypothesis: both samples come from the same underlying distribution
  - ✓ Distribution of T is not quite as simple as binomial, but it can be computed

Wilcoxon rank sum test example

If the operating rates for 2 factories is the same.

For factory 1, the rates (% of capacity) are **71**, **82**, **77**, **92**, **88**.
For factory 2, the rates (% of capacity) are **85**, **82**, **94** & **97**.

Do the factory rates have the same **probability distributions (mean)** at the **.10** level?

Wilcoxon Rank Sum Test Solution

H_0: Identical distribution.

H_a: Shifted left or right

$\alpha = .10$

$n_1 = 4 \quad n_2 = 5$

Critical value(s):

Test statistic:

Decision:

Conclusion:

Σ Ranks

Wilcoxon Rank Sum Table (Portion)

α = .05 one-tailed; α = .10 two-tailed

		n_1						
		3		4		5		..
		T_L	T_U	T_L	T_U	T_L	T_U	..
	3	6	15	7	17	7	20	..
n_2	4	7	17	12	24	13	27	..
	5	7	20	13	27	19	36	..
	:	:	:	:	:	:	:	:

Wilcoxon Rank Sum Test Solution

H_0: Identical distribution

H_a: Shifted left or right

$\alpha = .10$

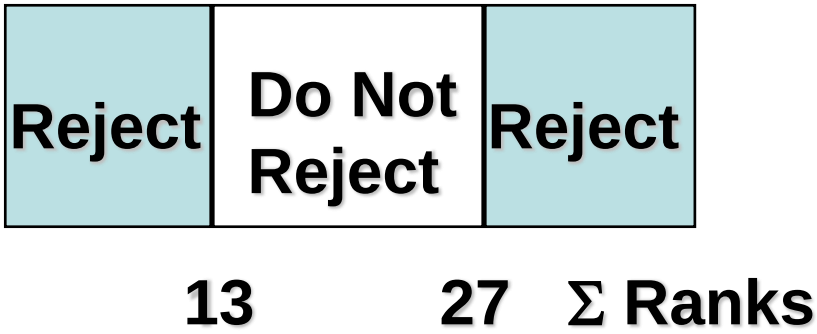
$n_1 = 4 \quad n_2 = 5$

Critical value(s):

Test statistic:

Decision:

Conclusion:



Wilcoxon Rank Sum Test Computation Table

Factory 1		Factory 2	
Rate	Rank	Rate	Rank
71	1	85	5
82	~~3~~ 3.5	82	~~4~~ 3.5
77	2	94	8
92	7	97	9
88	6		
Rank Sum	19.5		25.5

Wilcoxon rank sum test solution

H_0: Identical distribution

H_a: Shifted left or right

$\alpha = .10$

$n_1 = 4 \quad n_2 = 5$

Critical value(s):

Reject	Do Not Reject	Reject

13 **27** **Σ Ranks**

Test statistic:

$T_2 = 5 + 3.5 + 8 + 9 = 25.5$

(Smallest sample)

Decision:

Do not reject at $\alpha = .10$

Conclusion:

There is no evidence that distribution are not equal

Comparisons

Test	H0	Fixed Information	Randomized Information	Test Statistic	Notes
Sign	Prob of + equals Prob of -	Number of non-zero differences	Sign of each difference	Number of + 's	Binomial Distribution in Theory
Wilcoxon Signed Ranks	Symmetry about zero	Absolute values of differences	Sign of each difference	Sum of positive ranks	
Mann-Whitney/ Wilcoxon rank sum test	Two groups from same distribution	Ranks in the two groups	Group membership of datapoints	Sum of ranks in smallest group	Random assignment to groups

More than two groups: Kruskal-Wallis (H) (for ordinal data)

One-Way ANOVA by ranks

Null Hypothesis: k samples come from same population or identical populations with respect to central tendencies.

Rank order all N scores from the smallest to the largest.

G1	G2	G3
96	82	115
128	124	149
83	132	166
61	135	147
101	109	

Transformed

4	2	7
9	8	13
3	10	14
1	11	12
5	6	

Kruskal-Wallis (H)

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

K = # of groups
n = number of cases in a group
N = total number of cases
R = sum of ranks in a group

Kruskal-Wallis (H)

Group1	Group2	Group3
4	2	7
9	8	13
3	10	14
1	11	12
5	6	

$$H = \frac{12}{14(14+1)}\left[\frac{22^2}{5} + \frac{37^2}{5} + \frac{46^2}{4}\right] - 3(14+1)$$

$= 6.4$

This is tested against Chi-Squared (k-1) df.
Go to Chi-Squared Table , with 2 df and alpha .05

R code

```
# Kruskal Wallis Test One Way Anova by Ranks
kruskal.test(y ~ A)
# where y is numeric and A is a factor
```

```
 Kruskal-Wallis rank sum test

data:  values by group
Kruskal-Wallis chi-squared = 6.41, df = 2, p-value = 0.04065
```

```
lines <-  "values group
96 g1
128 g1
83 g1
61 g1
101 g1
82 g2
124 g2
132 g2
135 g2
109 g2
115 g3
149 g3
166 g3
147 g3"
Kruskal <- read.table(con <- textConnection(lines),
                      header=TRUE)
close(con)
```

```
kruskal.test(values ~ group, data=Kruskal)
```

Friedman's Test for k related samples

Corresponds to one-way repeated measures

Example

Patients	Psy(rank)	drug(rank)	psy/drug(rank)
1	6(2)	8(3)	5(1)
2	4(1)	8(3)	6(2)
3	9(3)	7(2)	4(1)
4	5(2)	4(1)	6(3)
5	2(1)	7(3)	3(2)
6	6(1)	8(3)	7(2)
7	7(2)	9(3)	5(1)
8	4(1)	8(3)	5(2)
9	6(3)	5(2)	4(1)
10	7(2)	8(3)	6(1)
Totals	18	26	16

$$X^2_F = \left[\frac{12}{Nk(k+1)} \sum_{j=1}^{k} R_j^2 \right] - 3N(k+1)$$

N = # of patients

$$= \frac{12}{10(3)(4)} (18^2 + 26^2 + 16^2) - 3(10)(4)$$

$$= 5.6$$

This is test against Chi-Squared (k-1) df.
Fail to reject at alpha .05.

R code

```
lines <- "patients scores treat
1 6 Psy
2 4 Psy
3 9 Psy
4 5 Psy
5 2 Psy
6 6 Psy
7 7 Psy
8 4 Psy
9 6 Psy
10 7 Psy
1 8 Drug
2 8 Drug
3 7 Drug
4 4 Drug
5 7 Drug
6 8 Drug
7 9 Drug
8 8 Drug
9 5 Drug
10 8 Drug
1 5 PsyDrug
2 6 PsyDrug
3 4 PsyDrug
4 6 PsyDrug
5 3 PsyDrug
6 7 PsyDrug
7 5 PsyDrug
8 5 PsyDrug
9 4 PsyDrug
10 6 PsyDrug"
```

```
# One-way repeated ANOVA - Friedman Test
friedman.test(y ~ A|B)
# where y are the data values, A is a grouping factor (treat)
# and B is a blocking factor (patients)
```

```
Friedman rank sum test

data:  scores and treat and patients
Friedman chi-squared = 5.6, df = 2, p-value = 0.06081
```

```
Fridman <- read.table(con <- textConnection(lines), header=TRUE); close(con)
friedman.test(scores ~ treat | patients, data = Fridman)
```

Correlation

```
cor(pre, post, method = "pearson") #parametric
cor(pre, post, method = "spearman") #nonparametric
cor.test(pre, post, method = "pearson") #parametric
cor.test(pre, post, method = "spearman") #nonparametric
```

```
> cor(pre, post, method = "pearson") #parametric
[1] 0.671
> cor(pre, post, method = "spearman") #nonparametric
[1] 0.828
> cor.test(pre, post, method = "pearson") #parametric

        Pearson's product-moment correlation

data:  pre and post
t = 3, df = 11, p-value = 0.01203
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.191 0.892
sample estimates:
  cor
0.671
```

```
> cor.test(pre, post, method = "spearman")
#nonparametric

        Spearman's rank correlation rho

data:  pre and post
S = 62.6, p-value = 0.0004726
alternative hypothesis: true rho is not equal to 0
sample estimates:
  rho
0.828
```

McNemar test

Used on paired nominal data, only applied to 2 × 2 contingency tables with a dichotomous trait, with matched pairs of subjects, to determine whether the row and column marginal frequencies are equal (McNemar 1947).

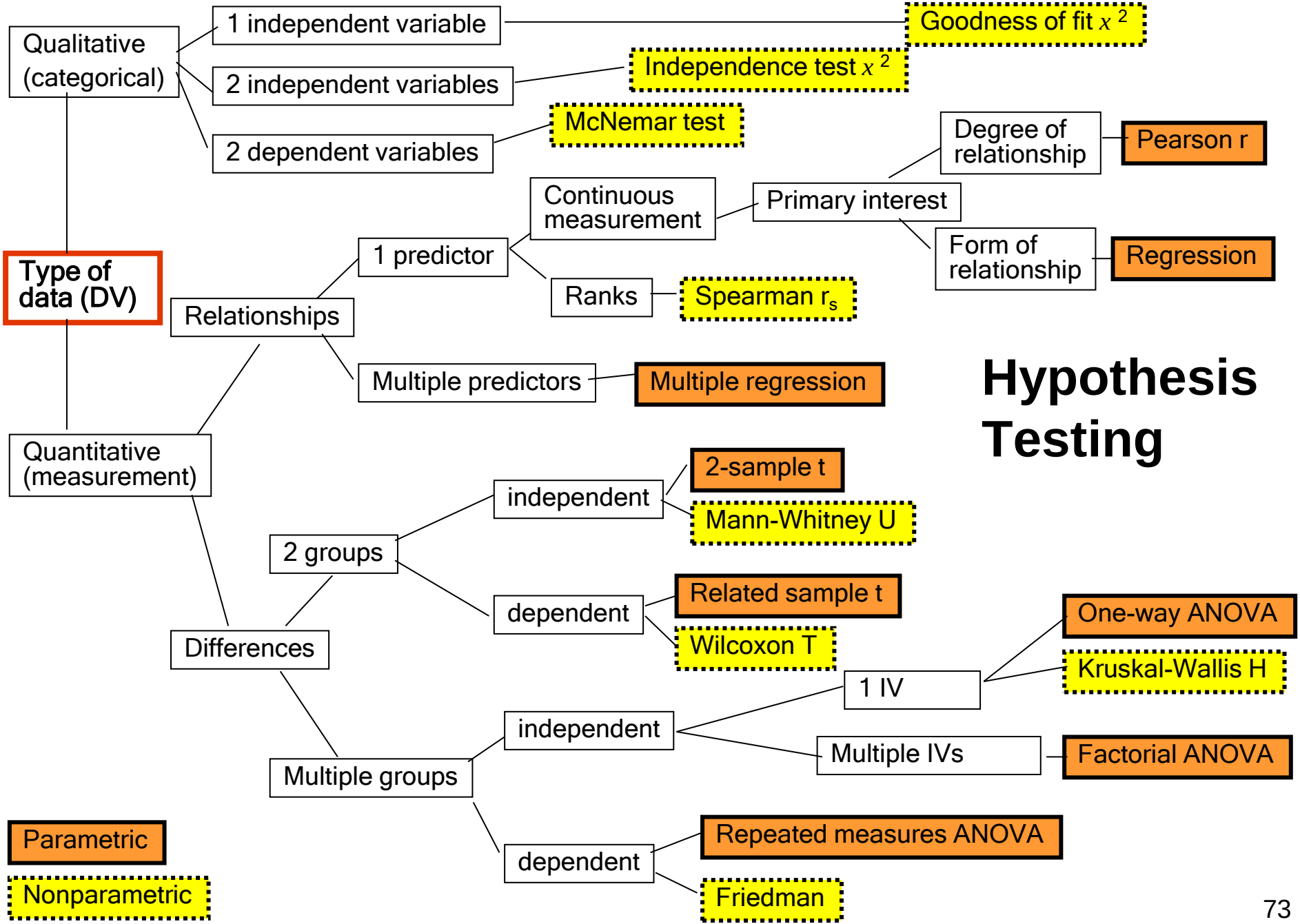
	Test 2 positive	Test 2 negative	Row total
Test 1 positive	*a*	*b*	*a* + *b*
Test 1 negative	*c*	*d*	*c* + *d*
Column total	*a* + *c*	*b* + *d*	*n*

The McNemar test statistic is

$$x^2 = \frac{(b-c)^2}{b+c}$$

McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12** (2): 153–157.

Design	Nonparametric		Parametric
	At least nominal	At least ordinal	Interval/ratio
One sample	χ^2 goodness of fit	Sign Test	1 sample z (s known) or t (s unknown) test
2 independent samples	χ^2 test of independence	Median Test Wilcoxon Rank Sum Test (Mann-Whitney test)	2 independent sample t test
2 dependent samples	McNemar test for significant change	Sign Test Wilcoxon signed-rank test	2 dependent samples t test
k independent samples	χ^2 test of independence	Kruskal-Wallis H test	1-way ANOVA
2 or more variables		Friedman Test	One-way repeated ANOVA
Correlation		Spearman r_s	Pearson r
Distribution		Kolmogorov-Smirnov	


Type of data (DV)
Qualitative (categorical)
1 independent variable
Goodness of fit x^2
2 independent variables
Independence test x^2
2 dependent variables
McNemar test
Relationships
1 predictor
Continuous measurement
Primary interest
Degree of relationship
Pearson r
Form of relationship
Regression
Ranks
Spearman r_s
Multiple predictors
Multiple regression
Quantitative (measurement)
Hypothesis Testing
Differences
2 groups
independent
2-sample t
Mann-Whitney U
dependent
Related sample t
Wilcoxon T
Multiple groups
independent
1 IV
One-way ANOVA
Kruskal-Wallis H
Multiple IVs
Factorial ANOVA
dependent
Repeated measures ANOVA
Friedman
Parametric
Nonparametric

Nonparametric regression

- Kernel estimators
- Simple-regression smoothing-spline
- Local polynomial regression
- Generalized nonparametric regression
- Generalized additive models

Kernel estimators

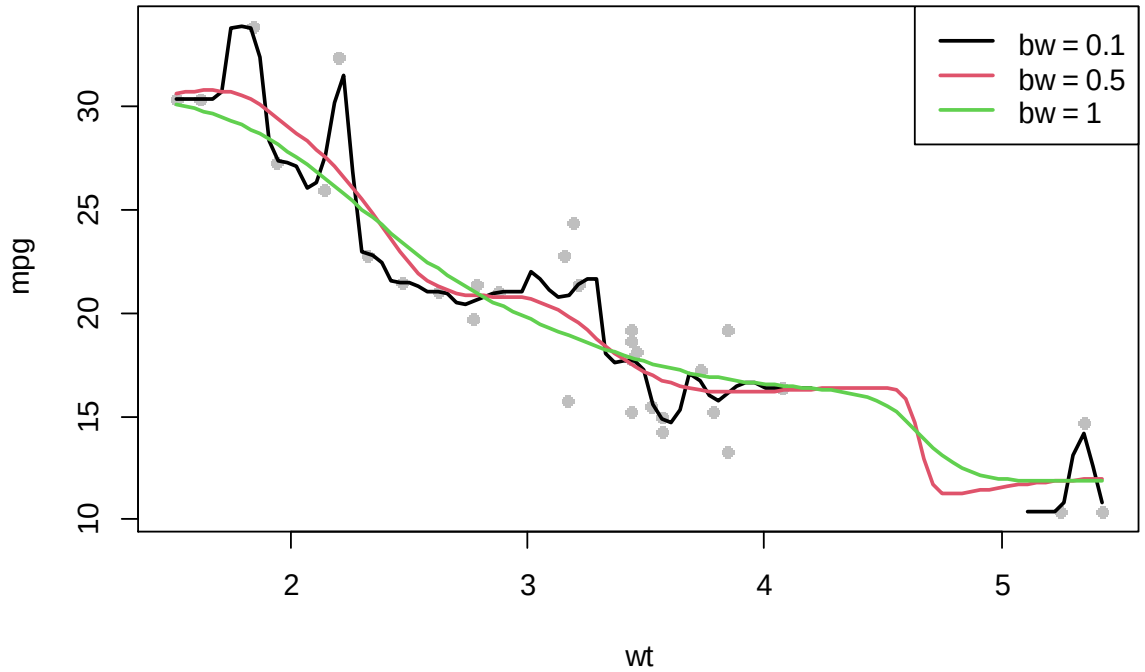
x_1, x_2, …,x_n are IID (independent and identical distribution) samples.

$$\widehat{f_h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Where K is the kernel function, h is bandwidth

Kernel estimators

```
# using dataset mtcars, fit mpg ~ wt
with(mtcars, plot(mpg ~ wt, col=gray(0.75), pch=16))
color=0
for(bw in c(0.1,0.5,1)){
   color = color + 1
   with(mtcars,  lines(ksmooth(wt, mpg, "normal", bw), lwd=2, col = color))
 }
legend("topright", c("bw = 0.1", "bw = 0.5", "bw = 1"), lwd=2, col = c(1,2,3))
```



Smoothing spline

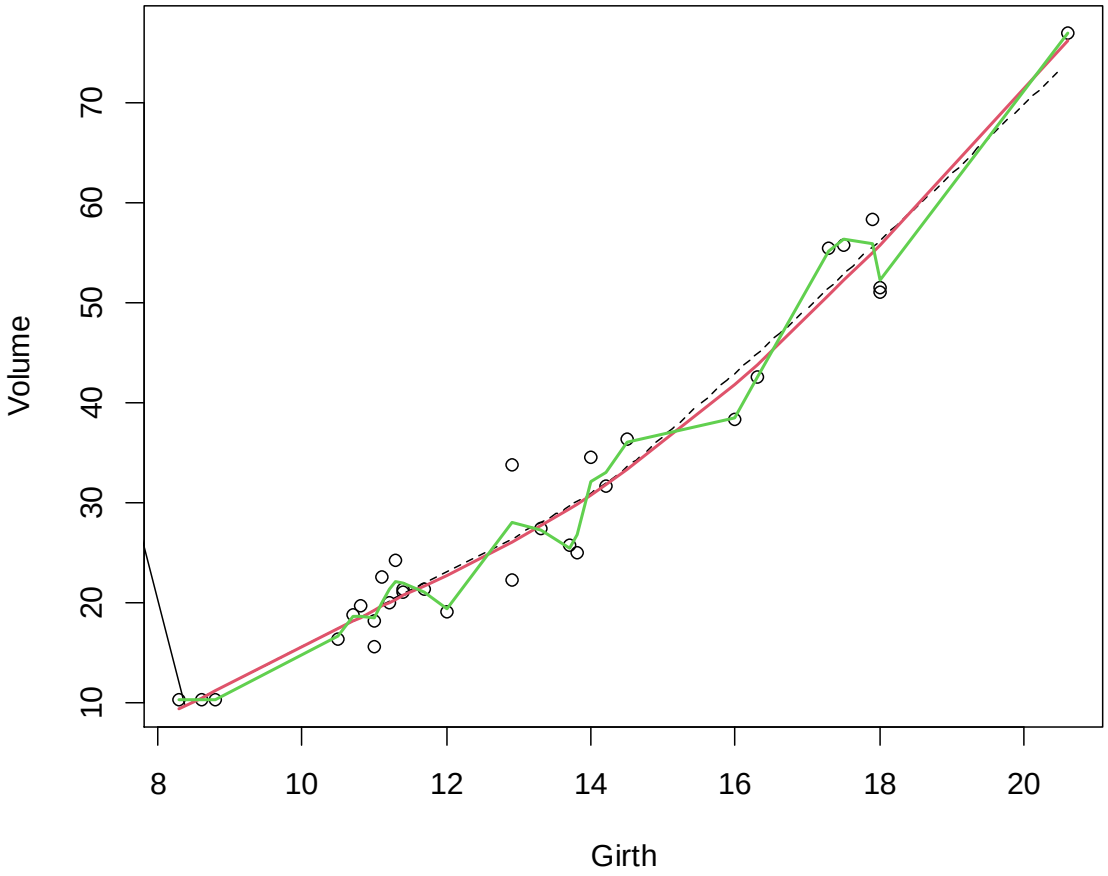
Fit *m* (the smoothing spline function), minimize the following *SS*.

$$SS(h) = \sum_{i=1}^{n} [y_i - m(x_i)]^2 + h \int_{x_{min}}^{x_{max}} [m''(x)]^2 \, dx$$

Where, h is bandwidth.

Smoothing spline

```
plot(Volume ~ Girth, data=trees)
girth.100 <- with(trees, seq(min(Girth), max(Girth), len=100)) # 100 x-values
lines(with(trees, smooth.spline(Girth, Volume, df=5)), col=2, lwd=2)
lines(with(trees, smooth.spline(Girth, Volume, df=20)), col=3, lwd=2)
```

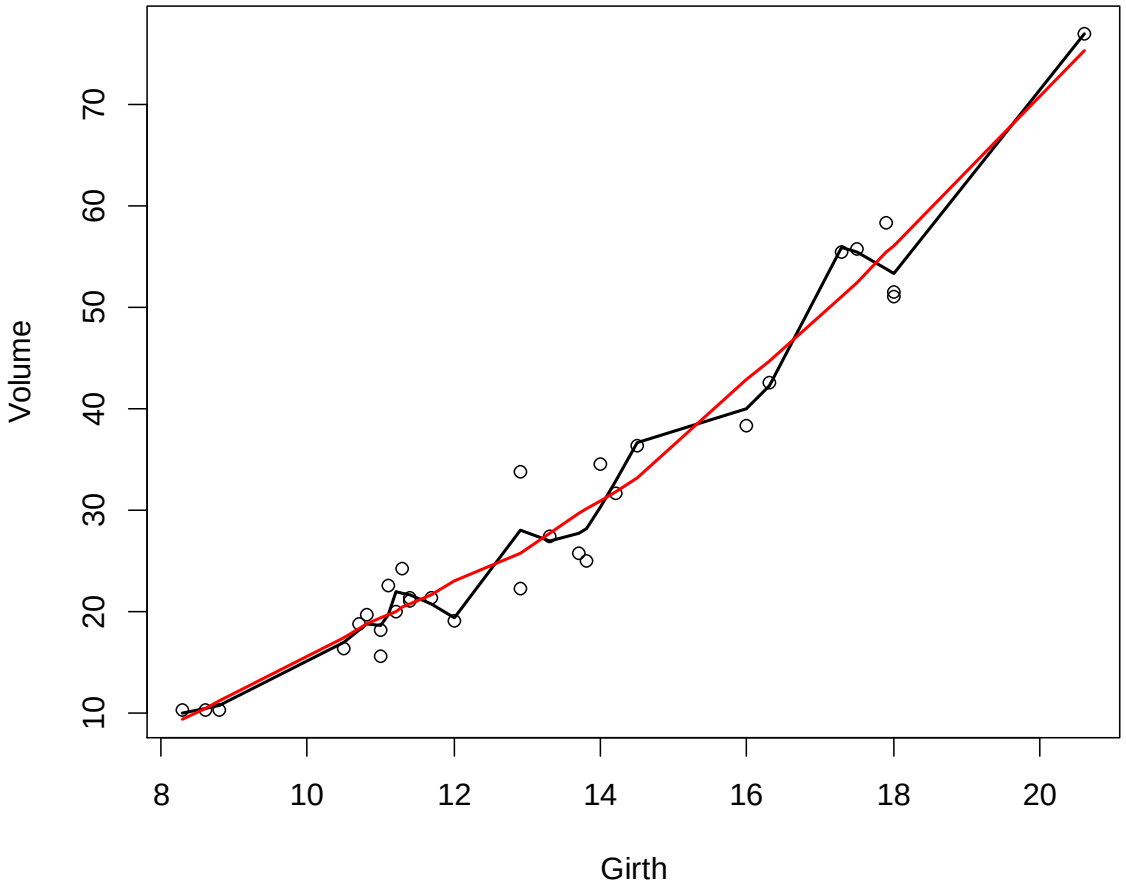


Local polynomial regression

$$y = \beta_0+\beta_1(x-x_0) +\beta_2(x-x_0)^2 +\ldots+\beta_p(x-x_0)^p + \varepsilon$$

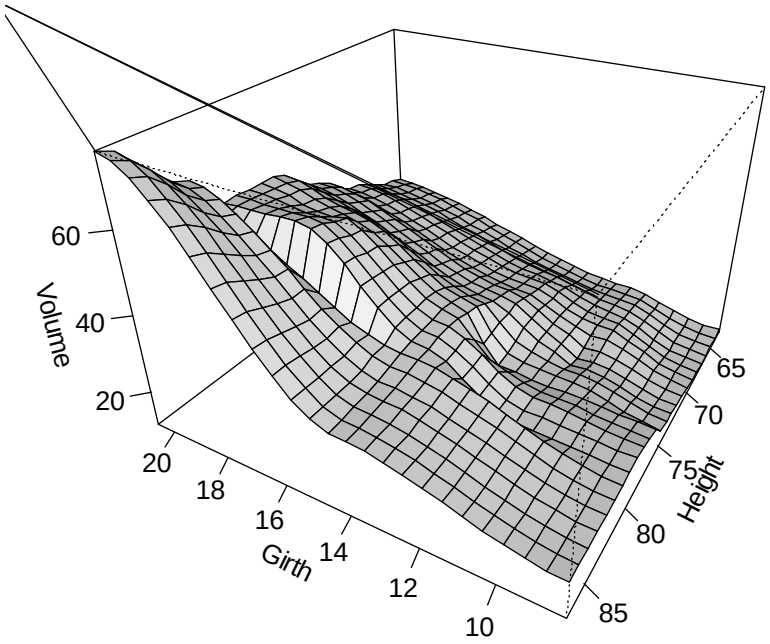
Local polynomial regression (y~x)

```
plot(Volume ~ Girth, xlab="Girth", ylab="Volume", data=trees)
with(trees, lines(lowess(Girth, Volume, f=0.2, iter=0), lwd=2))
with(trees, lines(lowess(Girth, Volume, f=0.5, iter=0), lwd=2, col="red"))
```



Local polynomial regression (y~x1+x2)

```
mod.lo <- loess(Volume ~ Girth + Height, data=trees, span=.2, degree=0)
summary(mod.lo)
girth <- with(trees, seq(min(Girth), max(Girth), len=25))
height <- with(trees, seq(min(Height), max(Height), len=25))
newdata <- expand.grid(Girth=girth, Height=height)
fit.Volume <- matrix(predict(mod.lo, newdata), 25, 25) # prediction
persp(girth, height, fit.Volume, theta=210, phi=30, ticktype="detailed",
    xlab="Girth", ylab="Height", zlab="Volume", expand=2/3, shade=0.5)
```

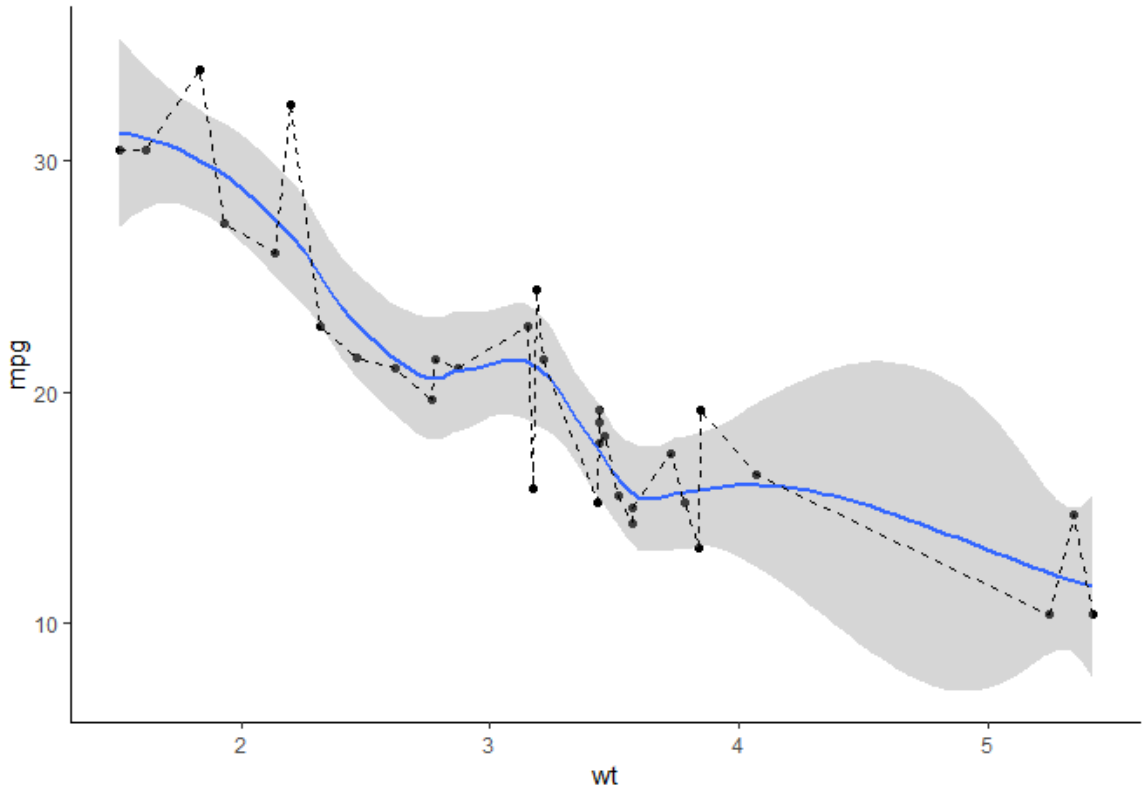


span: the parameter controls the degree of smoothing.

degree: the degree of the polynomials to be used, normally 1 or 2.

Local polynomial regression (y~x)

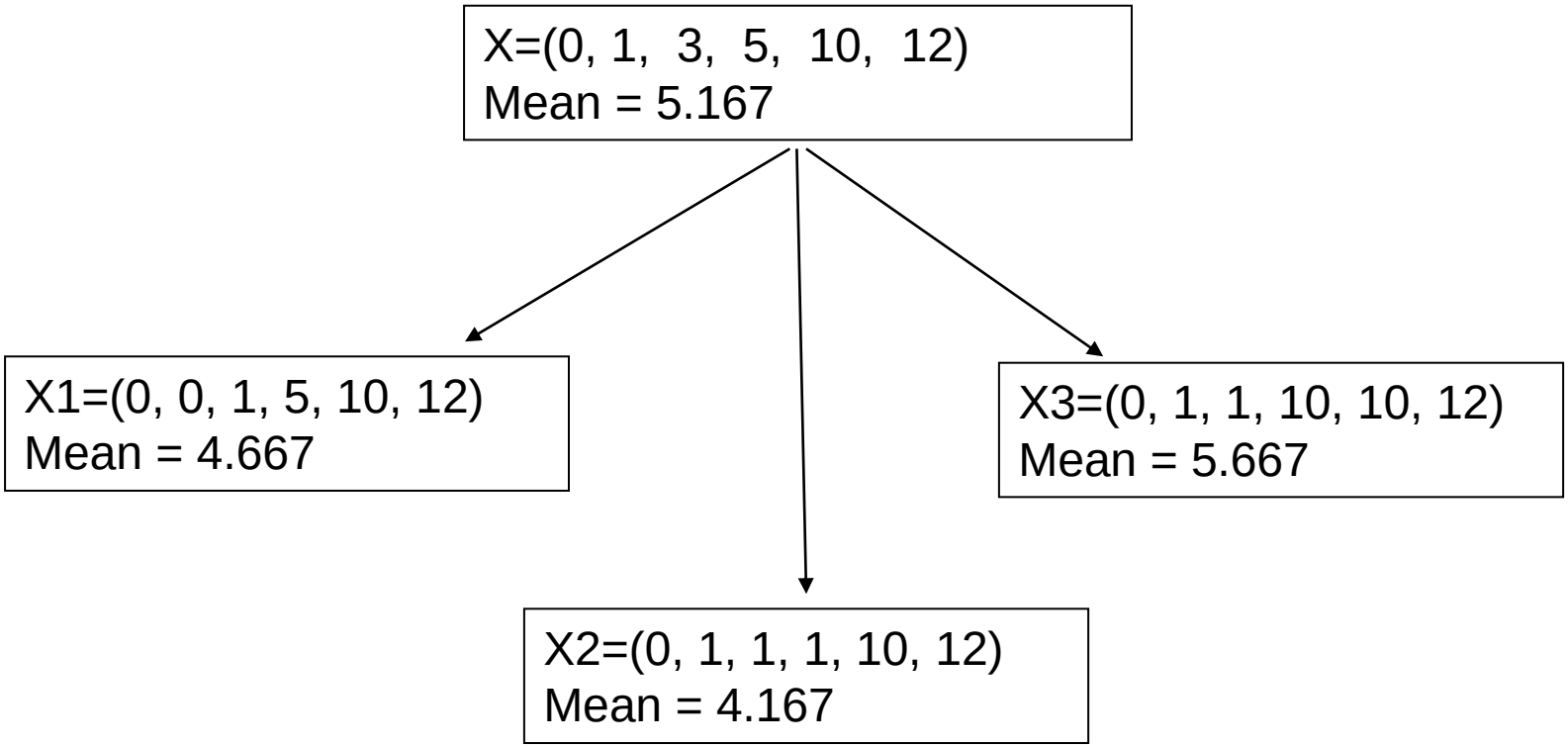
```
library(ggplot2)
ggplot(mtcars, aes(x=wt, y=mpg)) + geom_point(alpha=1) +
geom_smooth(method="loess", span=0.5) +
    geom_line(aes(x=wt, y=mpg), linetype=2) +
      theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
      panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



Bootstrapping

- Introduced by Bradley Efron in 1979
- Named from the phrase “to pull oneself up by one’s bootstraps”, which is widely believed to come from “the Adventures of *Baron Munchausen*”.
- Popularized in 1980s due to the introduction of computers in statistical practice.
- While it is a method for improving estimators, it is well known as a method for estimating standard errors, bias, and constructing confidence intervals for parameters.

An example


X=(0, 1, 3, 5, 10, 12)
Mean = 5.167
X1=(0, 0, 1, 5, 10, 12)
Mean = 4.667
X2=(0, 1, 1, 1, 10, 12)
Mean = 4.167
X3=(0, 1, 1, 10, 10, 12)
Mean = 5.667

Procedure for bootstrapping

The original sample is: $x=(x_1,x_2,\ldots,x_n)$

- Repeat B times
  - Generate a sample x* of size n from x by sampling with replacement.
  - Compute $\hat{\theta}*$ for x*.

Now we end up with bootstrap values

$$\hat{\theta}* = (\hat{\theta}_1^*,\ldots,\hat{\theta}_B^*)$$

- Use these values for calculating all the quantities of interest (e.g., standard deviation, confidence intervals)

Features of bootstrap

- It has minimum assumptions. It is merely based on the assumption that the sample is a good representation of the unknown population.

- In practice, it is computationally demanding, but the progress on computer speed makes it easily available in everyday practice.

Bootstrap distribution

- The bootstrap does not replace or add to the original data.
- Bootstrap distributions usually approximate the shape, spread, and bias of the actual sampling distribution, to estimate the variation in a statistic based on the original data. .
- Bootstrap distributions are centered at the value of the statistic from the original data plus any bias, while the sampling distribution is centered at the value of the parameter in the population, plus any bias.

Cases where bootstrapping does not apply

- Small data sets: the original sample is not a good approximation of the population
- Dirty data: outliers add variability in our estimates.
- Dependence structures (e.g., time series, spatial problems): Bootstrap is based on the assumption of independence.

How many bootstrap samples are needed?

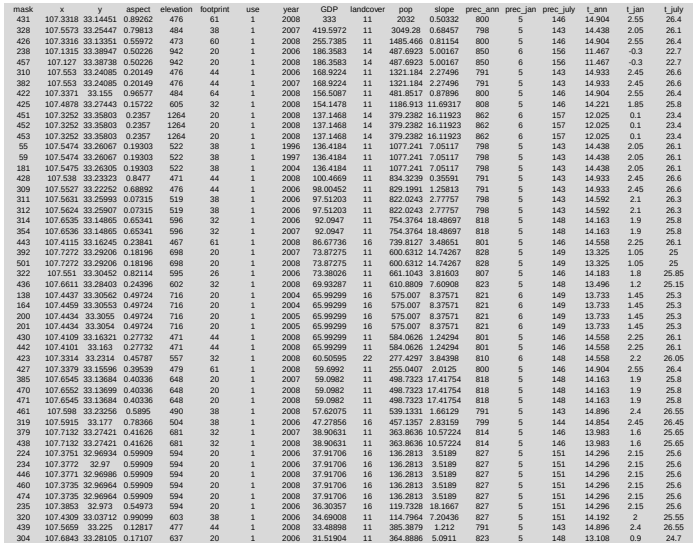
Depends on:

- Computer availability
- Type of the problem: standard errors, confidence intervals, …
- Complexity of the problem

Why resampling?

- Fewer assumptions
  - Ex: resampling methods do not require that distributions be normal or that sample sizes be large
- Greater accuracy: bootstrap methods are more accurate in practice than classical methods
- Generality: resampling methods are remarkably similar for a wide range of statistics and do not require new formulas for every statistic.
- Promote understanding: boostrap procedures build intuition by providing concrete analogies to theoretical concepts.

R code - bootstrap



mask	x	y	aspect	elevation	footprint	use	year	GDP	landcover	pop	slope	prec_ann	prec_jan	prec_july	t_ann	t_jan	t_july
431	107.3318	33.14451	0.89262	476	61	1	2008	333	11	2032	0.50332	800	5	146	14.904	2.55	26.4
328	107.5573	33.25447	0.79813	484	38	1	2007	419.5972	11	3049.28	0.68457	798	5	143	14.438	2.05	26.1
426	107.3316	33.13351	0.55972	473	60	1	2008	255.7385	11	1485.466	0.81154	800	5	146	14.904	2.55	26.4
238	107.1315	33.38947	0.50226	942	20	1	2006	186.3583	14	487.6923	5.00167	850	6	156	11.467	-0.3	22.7
457	107.127	33.38738	0.50226	942	20	1	2008	186.3583	14	487.6923	5.00167	850	6	156	11.467	-0.3	22.7
310	107.553	33.24085	0.20149	476	44	1	2006	168.9224	11	1321.184	2.27496	791	5	143	14.933	2.45	26.6
382	107.553	33.24085	0.20149	476	44	1	2007	168.9224	11	1321.184	2.27496	791	5	143	14.933	2.45	26.6
422	107.3371	33.155	0.96577	484	64	1	2008	156.5087	11	481.8517	0.87896	800	5	146	14.904	2.55	26.4
425	107.4878	33.27443	0.15722	605	32	1	2008	154.1478	11	1186.913	11.69317	808	5	146	14.221	1.85	25.8
451	107.3252	33.35803	0.2357	1264	20	1	2008	137.1468	14	379.2382	16.11923	862	6	157	12.025	0.1	23.4
452	107.3252	33.35803	0.2357	1264	20	1	2008	137.1468	14	379.2382	16.11923	862	6	157	12.025	0.1	23.4
453	107.3252	33.35803	0.2357	1264	20	1	2008	137.1468	14	379.2382	16.11923	862	6	157	12.025	0.1	23.4
55	107.5474	33.26067	0.19303	522	38	1	1996	136.4184	11	1077.241	7.05117	798	5	143	14.438	2.05	26.1
59	107.5474	33.26067	0.19303	522	38	1	1997	136.4184	11	1077.241	7.05117	798	5	143	14.438	2.05	26.1
181	107.5475	33.26305	0.19303	522	38	1	2004	136.4184	11	1077.241	7.05117	798	5	143	14.438	2.05	26.1
428	107.538	33.23323	0.8477	471	44	1	2008	100.4669	11	834.3239	0.35591	791	5	143	14.933	2.45	26.6
309	107.5527	33.22252	0.68892	476	44	1	2006	98.00452	11	829.1991	1.25813	791	5	143	14.933	2.45	26.6
311	107.5631	33.25993	0.07315	519	38	1	2006	97.51203	11	822.0243	2.77757	798	5	143	14.592	2.1	26.3
312	107.5624	33.25907	0.07315	519	38	1	2006	97.51203	11	822.0243	2.77757	798	5	143	14.592	2.1	26.3
314	107.6535	33.14865	0.65341	596	32	1	2006	92.0947	11	754.3764	18.48697	818	5	148	14.163	1.9	25.8
354	107.6536	33.14865	0.65341	596	32	1	2007	92.0947	11	754.3764	18.48697	818	5	148	14.163	1.9	25.8
443	107.4115	33.16245	0.23841	467	61	1	2008	86.67736	16	739.8127	3.48651	801	5	146	14.558	2.25	26.1
392	107.7272	33.29206	0.18196	698	20	1	2007	73.87275	11	600.6312	14.74267	828	5	149	13.325	1.05	25
501	107.7272	33.29206	0.18196	698	20	1	2008	73.87275	11	600.6312	14.74267	828	5	149	13.325	1.05	25
322	107.551	33.30452	0.82114	595	26	1	2006	73.38026	11	661.1043	3.81603	807	5	146	14.183	1.8	25.85
436	107.6611	33.28403	0.24396	602	32	1	2008	69.93287	11	610.8809	7.60908	823	5	148	13.496	1.2	25.15
138	107.4437	33.30562	0.49724	716	20	1	2004	65.99299	16	575.007	8.37571	821	6	149	13.733	1.45	25.3
164	107.4459	33.30553	0.49724	716	20	1	2004	65.99299	16	575.007	8.37571	821	6	149	13.733	1.45	25.3
200	107.4434	33.3055	0.49724	716	20	1	2005	65.99299	16	575.007	8.37571	821	6	149	13.733	1.45	25.3
201	107.4434	33.3054	0.49724	716	20	1	2005	65.99299	16	575.007	8.37571	821	6	149	13.733	1.45	25.3
430	107.4109	33.16321	0.27732	471	44	1	2008	65.99299	11	584.0626	1.24294	801	5	146	14.558	2.25	26.1
442	107.4101	33.163	0.27732	471	44	1	2008	65.99299	11	584.0626	1.24294	801	5	146	14.558	2.25	26.1
423	107.3314	33.2314	0.45787	557	32	1	2008	60.50595	22	277.4297	3.84398	810	6	148	14.558	2.2	26.05
427	107.3379	33.15596	0.39539	479	61	1	2008	59.6992	11	255.0407	2.0125	800	5	146	14.904	2.55	26.4
385	107.6545	33.13684	0.40336	648	20	1	2007	59.0982	11	498.7323	17.41754	818	5	148	14.163	1.9	25.8
470	107.6552	33.13699	0.40336	648	20	1	2008	59.0982	11	498.7323	17.41754	818	5	148	14.163	1.9	25.8
471	107.6545	33.13684	0.40336	648	20	1	2008	59.0982	11	498.7323	17.41754	818	5	148	14.163	1.9	25.8
461	107.598	33.23256	0.5895	490	38	1	2008	57.62075	11	539.1331	1.66129	791	5	143	14.896	2.4	26.55
319	107.5915	33.177	0.78366	504	38	1	2006	47.27856	16	457.1357	2.83159	799	5	144	14.854	2.45	26.45
379	107.7132	33.27421	0.41626	681	32	1	2007	38.90631	11	363.8636	10.57224	814	5	146	13.983	1.6	25.65
438	107.7132	33.27421	0.41626	681	32	1	2008	38.90631	11	363.8636	10.57224	814	5	146	13.983	1.6	25.65
224	107.3751	32.96934	0.59909	594	20	1	2006	37.91706	16	136.2813	3.5189	827	5	151	14.296	2.15	25.6
234	107.3772	32.97	0.59909	594	20	1	2006	37.91706	16	136.2813	3.5189	827	5	151	14.296	2.15	25.6
446	107.3771	32.96986	0.59909	594	20	1	2008	37.91706	16	136.2813	3.5189	827	5	151	14.296	2.15	25.6
460	107.3735	32.96964	0.59909	594	20	1	2008	37.91706	16	136.2813	3.5189	827	5	151	14.296	2.15	25.6
474	107.3735	32.96964	0.59909	594	20	1	2008	37.91706	16	136.2813	3.5189	827	5	151	14.296	2.15	25.6
235	107.3853	32.973	0.54973	594	20	1	2006	36.30357	16	119.7328	18.1667	827	5	151	14.296	2.15	25.6
320	107.4309	33.03712	0.99099	603	38	1	2006	34.69008	11	114.7964	7.20436	827	5	151	14.192	2	25.55
439	107.5659	33.225	0.12817	477	44	1	2008	33.48898	11	385.3879	1.212	791	5	143	14.896	2.4	26.55
304	107.6843	33.28105	0.17107	637	20	1	2006	31.51904	11	364.8886	5.0911	823	5	148	13.108	0.9	24.7

```
# Bootstrap 95% CI for regression coefficients
library(boot)
# Import data
ibis = read.csv('D:/database/ibis2010.csv', header=T)
ibis.pre = ibis[ibis$use==1, c(3:6,8,9,11,12)]
head(ibis.pre)
```

	y	aspect	elevation	footprint	year	GDP	pop	slope
1	33.1	0.893	476	61	2008	333	2032	0.503
42	33.3	0.798	484	38	2007	420	3049	0.685
86	33.1	0.56	473	60	2008	256	1485	0.812
104	33.4	0.502	942	20	2006	186	488	5.002
105	33.4	0.502	942	20	2008	186	488	5.002
116	33.2	0.201	476	44	2006	169	1321	2.275

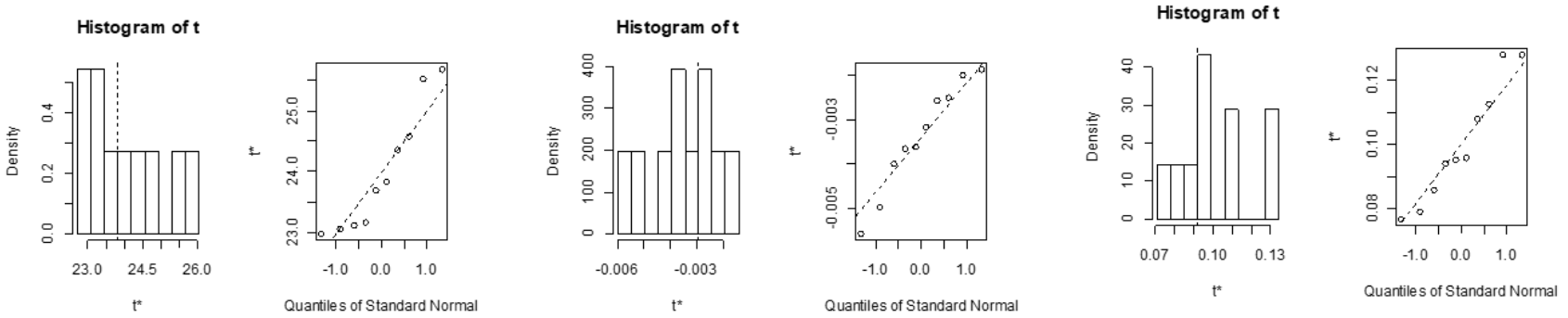
R code - bootstrap

```
# function to obtain regression weights
bs <- function(formula, data, indices) {
        d <- data[indices,] # allows boot to select sample
        fit <- lm(formula, data=d)
    return(coef(fit))
}

# bootstrapping with 1000 replications
results <- boot(data=ibis.pre, statistic=bs,
                      R=1000, formula=footprint~elevation+GDP)

# view results
results
plot(results, index=1, main='Intercept') # intercept
plot(results, index=2) # elevation
plot(results, index=3) # GDP

# get 95% confidence intervals
boot.ci(results, type="bca", index=1) # intercept
boot.ci(results, type="bca", index=2) # elevation
boot.ci(results, type="bca", index=3) # GDP
```

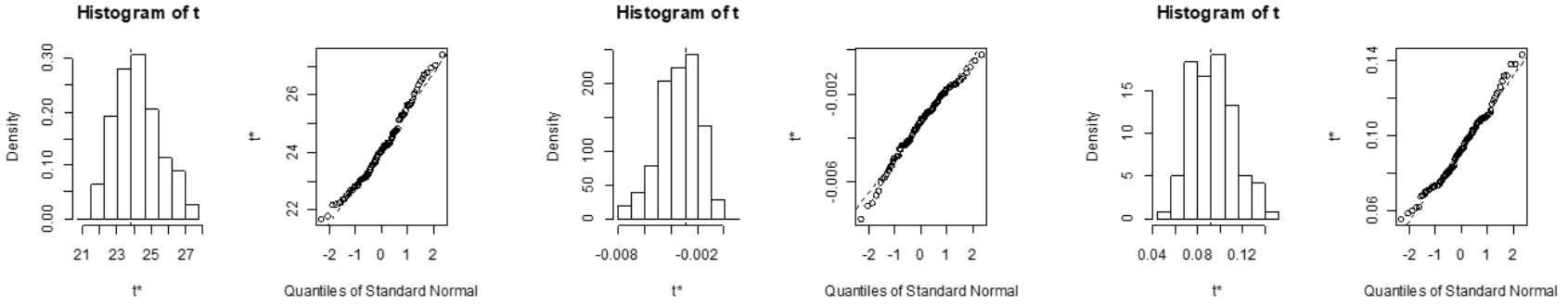
Bootstrap parameter estimation

Intercept | Elevation | GDP

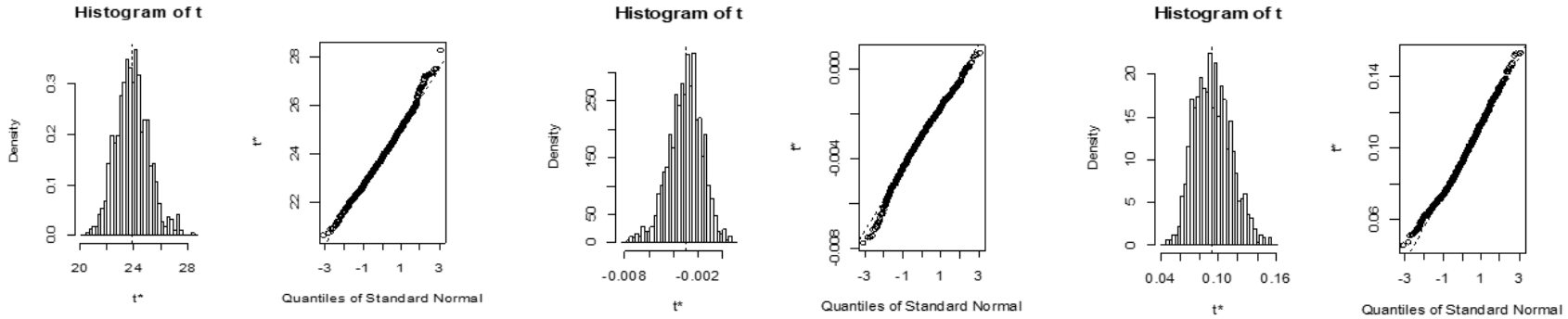
N=10



N=100



N=1000



Limitations of nonparametric statistics

- Still need some assumptions (including independence)
- Limited range of situations
  - no more than 2 x-variables
  - can't mix continuous and categorical x-variables
- Provide p-values but estimation is dodgy
- Loss of efficiency if parametric assumptions are upheld
- There is a grand scheme for parametric statistics (GLM) but a lot of separate strange names for nonparametrics

Assignment

General objectives: learn about **nonparametric statistics**.

- Develop your own data (see the data to the right), use R wilcox.test(treat1, treat2, paired=F) (two independent groups) to test whether the data of two groups differ. To clarify:

  – What is the p value?
  – Do you reject the null hypothesis?

Data FYI

Group A	Group B
20	30
10	40
25	35
30	45
35	25
15	50

Assignment

- From the data below fill in the table of ranks without regard to groups as in performing a Wilcoxon Rank Sum Test. Then compute the sum of the ranks for groups A and B.

Group A	Group B
20	30
10	40
25	35
30	45
35	25
15	50

value	rank	group

Rank Group A	Rank Group B

Sum Rank A = ?
Sum Rank B = ?