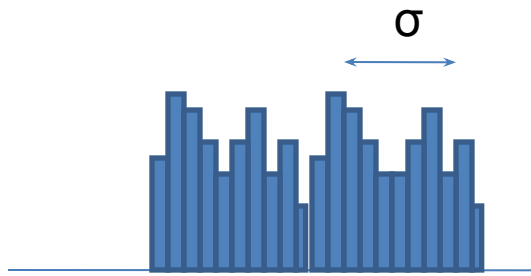


# One- and two-sample tests

# Confidence interval for the mean if $\sigma$ is known but $\mu$ is not

Population

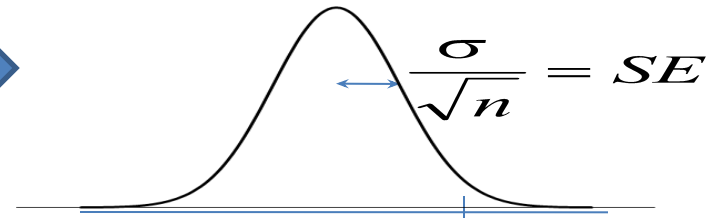


Central  
Limit  
Theorem

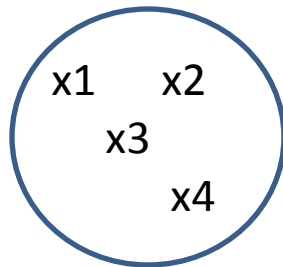


Sampling distribution  
(approx. normal as  $n$  increases)

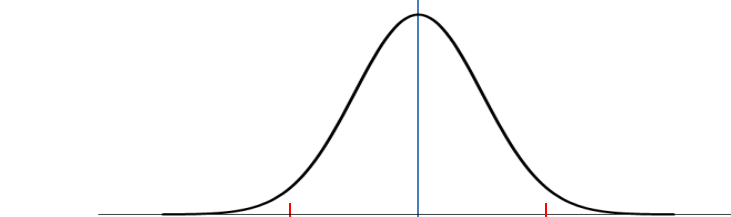
(approx. normal as  $n$  increases)



Sample



Statistical inference:



$$\mu = \bar{X} \pm z_{\text{critical}} (SE)$$

$$\begin{aligned} n &= 4 \\ \bar{X} &= 9 \\ s &= 2 \end{aligned}$$

# Confidence Interval

- Range which is expected to contain  $\mu$
- $\bar{X}_{\text{bar}}$  is center of CI
- Width of 95% CI is  $\pm 1.96$  SE (if  $\sigma$  is known)
- Width of 99% CI is  $\pm 2.58$  SE (if  $\sigma$  is known)
- Assumptions: the sampling distribution is approximately normal

# Using CIs to test hypotheses about $\mu$

- Example:

In the general English-speaking Canadian population, the second formant (F2) of /æ/ produced by adult males has a mean frequency of 1550 Hz with a standard deviation of 100 Hz, and is **normally distributed**. A random sample of 10 adult English-speaking Canadian males from a town near the border of the predominantly French-speaking province of Quebec was found to have a mean F2 of 1450 Hz in the vowel /æ/. Is this F2 level different from that of the general English-speaking Canadian adult male population?

$H_0: \mu_{\text{town}} = 1550$  (Null Hypothesis: not diff from gen pop)

$H_1: \mu_{\text{town}} \neq 1550$  (Alternative Hypothesis)

# Using CIs to test hypotheses about $\mu$

- 95% CI

$$SE = 100/\sqrt{10} = 31.64$$

$$\mu = X_{\text{bar}} \pm z_{\text{critical}}(SE)$$

$$\mu = 1450 \pm 1.96(31.64)$$

$$1388.02 < \mu < 1511.98$$

```
> qnorm(c(0.025, 0.975),  
        1450, 100 / sqrt(10))  
[1] 1388.02      1511.98
```

- Only 5 times out of 100 does the 95% CI **not** contain the mean of the population that the sample was drawn from
- Our 95% CI does not contain 1550, therefore it is unlikely that the F2 of the adult males in the town is the same as the F2 of the adult males in the general English-speaking Canadian population
- Reject  $H_0$

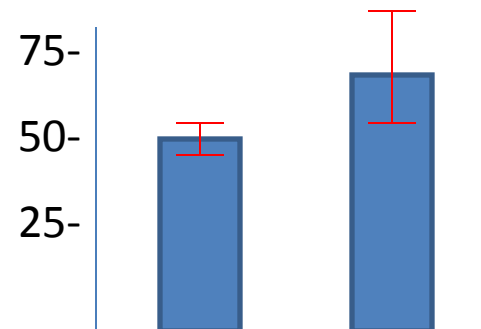
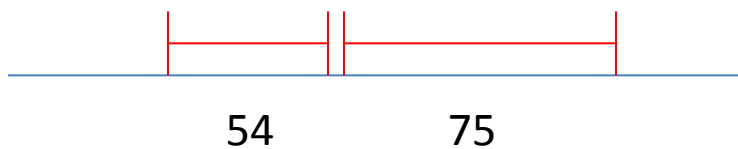
# Testing hypotheses about 2 samples

- If 2 independent samples have  $X_{\text{bars}}$  with 95% CIs that do not overlap, we can conclude that the samples come from different populations

- Example:

sample A:  $X_{\text{bar}} = 75$ ; SE = 8; 95% CI:  $59.32 < \mu < 90.68$

sample B:  $X_{\text{bar}} = 54$ ; SE = 2; 95% CI:  $50.08 < \mu < 57.92$



# Hypothesis testing: types of errors

- Type I: Rejecting the null hypothesis when it should not be rejected (i.e., there really is no difference)
  - probability of Type I error ( $\alpha$ )
    - the critical value of the test statistic
    - smaller  $\alpha$  -> wider CI -> fewer rejections of  $H_0$
- Type II: Not rejecting the null hypothesis when it should be rejected (i.e., there really is a difference)
  - probability of Type II error ( $\beta$ )
    - Reduced by increasing  $n$
    - Reduced by increasing *effect size*

# When is $\sigma$ unknown

- Typical situation in research
- Can use the sample to estimate  $\sigma$
- We already use  $X_{\text{bar}}$  to estimate  $\mu$

$X_{\text{bar}}$  is an *unbiased estimator* of  $\mu$ .

A statistic is an *unbiased estimator* if

- its value in the long run (i.e., its average value) is equal to the population parameter, and
- the mean of sampling distribution of means are  $\mu$ .



# Using a sample to estimate $\sigma^2$

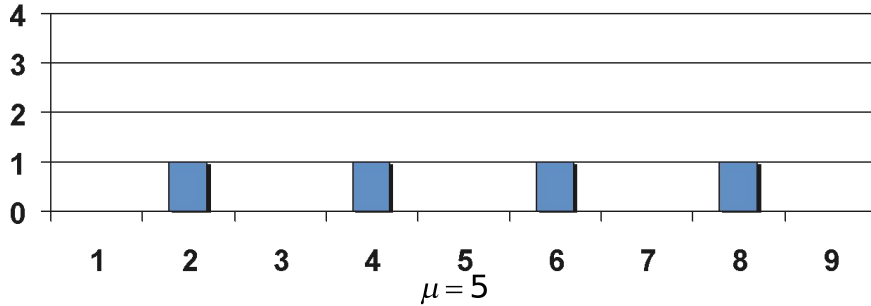
- Use  $X_{\text{bar}}$  as an estimate of  $\mu$

$$SS = \sum (x - X_{\text{bar}})^2 \text{ instead of } SS = \sum (x - \mu)^2$$

- But when estimating  $\sigma^2$  from sample data, the formula  $\sigma^2 = \sum (x - X_{\text{bar}})^2 / n$  produces a **biased estimate** of the population variance.
- To correct for this bias, use
$$s^2 = \sum (x - X_{\text{bar}})^2 / (n - 1)$$

# Biased and unbiased estimators

Population



Population Variance

$$\begin{aligned}
 SS &= (2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2 \\
 &= 9 + 1 + 1 + 9 \\
 &= 20 \\
 \sigma^2 &= \frac{20}{4} = 5
 \end{aligned}$$

$X_1$	$X_2$	$S^2$	$\sigma^2$	$X_1$	$X_2$	$S^2$	$\sigma^2$
2	2	0	0	6	2	8	4
2	4	2	1	6	4	2	1
2	6	8	4	6	6	0	0
2	8	18	9	6	8	2	1
4	2	2	1	8	2	18	9
4	4	0	0	8	4	8	4
4	6	2	1	8	6	2	1
4	8	8	4	8	8	0	0

Formula for  $\sigma^2 = SS/n$   
 Formula for  $s^2 = SS/(n-1)$

Conclusion:

When using sample means instead of  $\mu$ , the  $\sigma^2$  formula produces a biased estimate of the population variance:

$$(0+1+4+9+1+0+1+1+4+4+1+0+1+9+4+1+0)/16 = 2.5$$

but the  $s^2$  formula is unbiased:

$$(0+2+8+18+2+0+2+8+8+2+0+2+18+8+2+0)/16 = 5$$

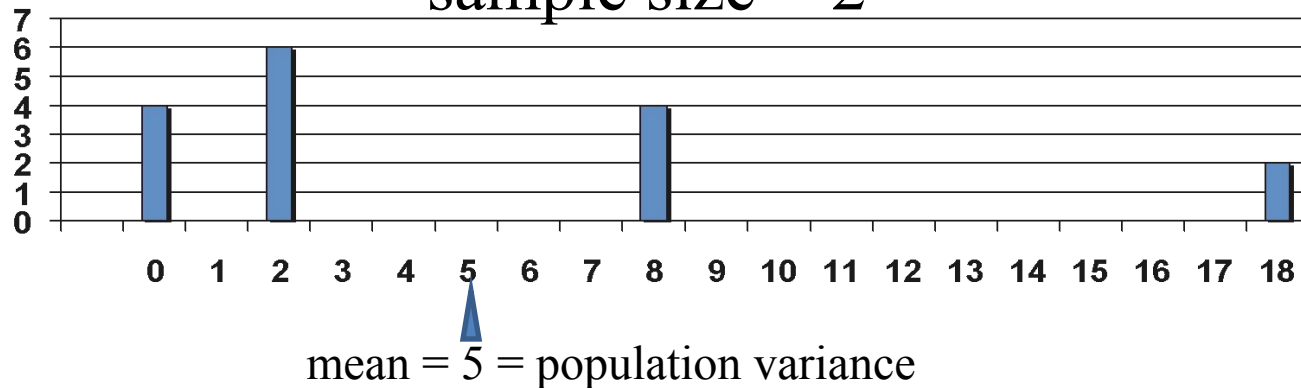
# Degrees of freedom

- A correction (dividing by  $n - 1$ ) is needed for the estimate ( $s^2$ ) because there are only  $n - 1$  **degrees of freedom** in the sample if  $X_{\text{bar}}$  is used in computing the estimate of the variance
- Using  $X_{\text{bar}}$  reduces the number of scores that are free to vary in the sample
  - If  $X_{\text{bar}}$  is known, then  $n - 1$  scores can vary but the  $n^{\text{th}}$  is fixed
  - Example:  $X_{\text{bar}} = 7$ ; sample = {11, 8, 5, 9, 6, ?};  
**? must be 3**
- In general, when computing a statistic, we lose 1 d.f. for each population parameter that we must estimate

# Sampling distribution of $s^2$

population =  $\{2,4,6,8\}$

sample size = 2



The sampling distribution of  $s^2$  has a mean equal to the population variance, which makes it an **unbiased estimator** of  $\sigma^2$ . Note, however, that the distribution is skewed, with 10 samples having values less than the mean and only 6 having values greater than the mean. As a consequence of this, although  $s^2$  is unbiased, *it is more likely that a sample's  $s^2$  will underestimate  $\sigma^2$  than overestimate it.*

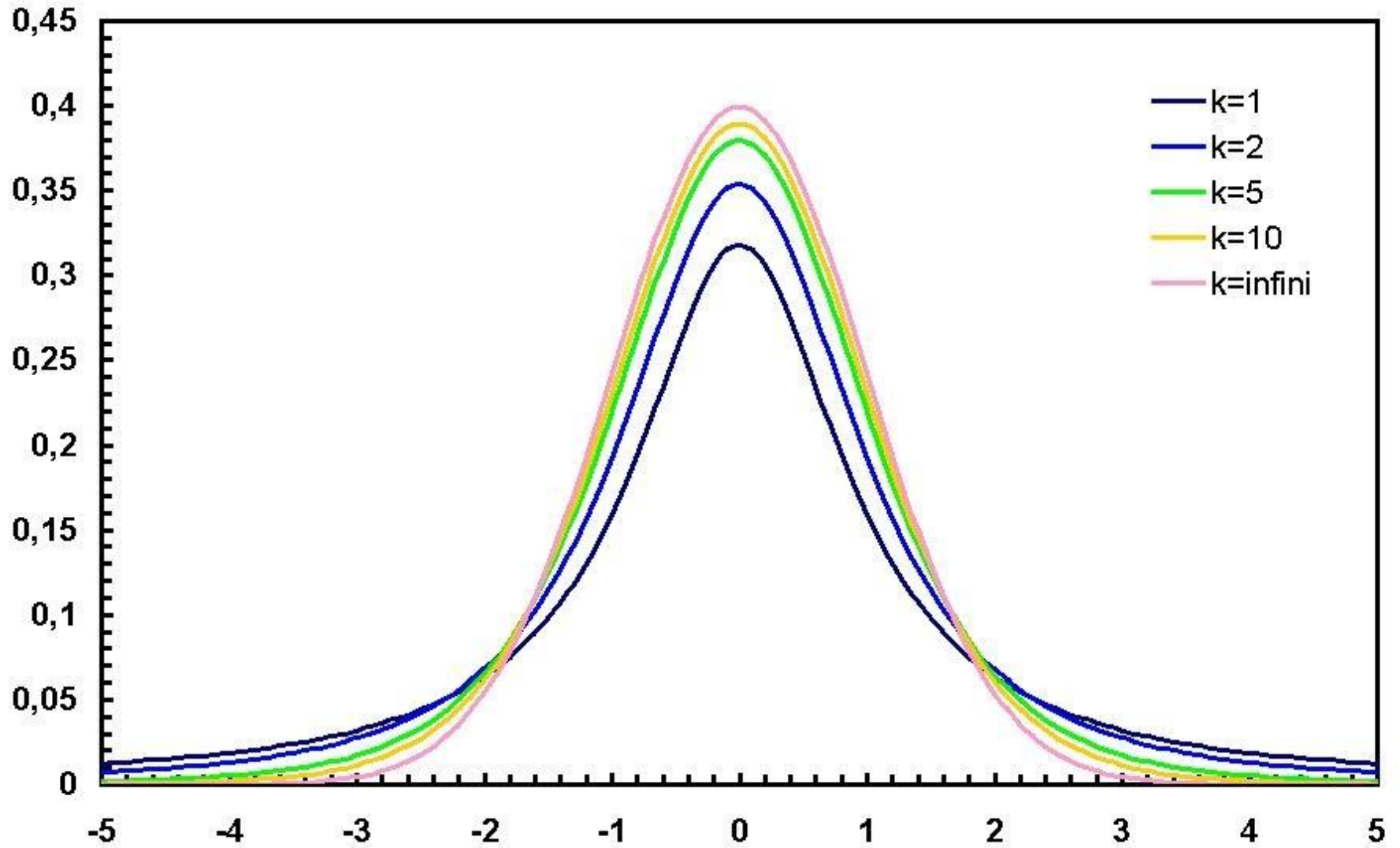
# Using $s^2$ in place of $\sigma^2$

- $s^2 = \sum (x - X_{\text{bar}})^2 / (n - 1)$  *sample variance*
- $s = \sqrt{s^2}$  *sample standard deviation*
- $s_{X_{\text{bar}}} = \text{SE} = s / \sqrt{n}$  *standard error of the mean*
- The R functions `var` and `sd` use the unbiased formulae
- Because  $s^2$  is an estimate that is more likely to undershoot  $\sigma^2$  than to overshoot it, the normal distribution probabilities must be “corrected” to compensate for this

# The $t$ distribution

- First described by Guinness Brewery employee William Gossett in 1908, using the pen name “Student”; known as Student’s  $t$ .
- Family of distributions, different for each df
  - Unimodal, symmetric, leptokurtic
  - heavy tails to compensate for “noise” in  $s$
- Shape approaches normal as d.f. increases

# $t$ distributions



# Critical values of $t$

two-tailed test

$\alpha = 0.05$       0.01  
(CI= 95%      99%)

- $df = 5$       2.57      4.03
- $df = 10$       2.23      3.17
- $df = 20$       2.09      2.85
- $df = 40$       2.02      2.70
- $df = 80$       1.99      2.65
- $df = \infty$       1.96      2.58

```
> qt(.975, c(5, 10, 20, 40, 80, 1000))  
[1] 2.570582      2.228139      2.085963      2.021075  
1.990063      1.962339
```

```
> qt(.995, c(5, 10, 20, 40, 80, 1000))  
[1] 4.032143      3.169273      2.845340      2.704459  
2.638691      2.580755
```



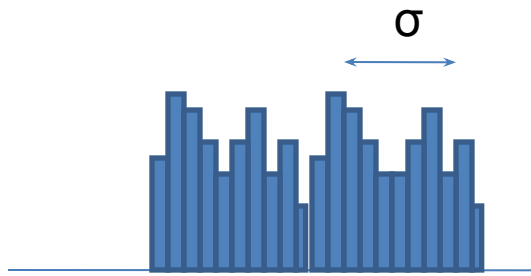
# Using $s$ in place of $\sigma$

Using  $t$  in place of the normal dist.

- $s_{Xbar} = SE = s/\sqrt{n}$       *standard error*
- $t = (X_{bar} - \mu_0) / s_{Xbar}$       *t-score for  $X_{bar}$*
- $\mu = X_{bar} \pm t_{critical}(s_{Xbar})$       *confidence interval*
  
- Use the  $t$  distribution if  $s$  must be used to estimate  $\sigma$  and either the population is normal or  $n$  is not too small

# Confidence interval for the mean if both $\sigma$ and $\mu$ are unknown

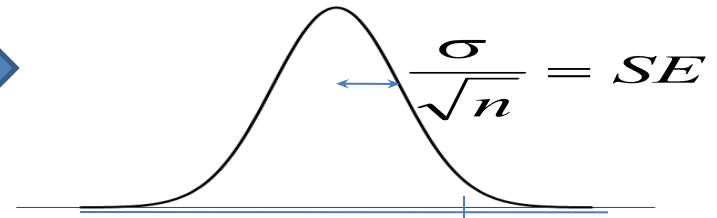
Population



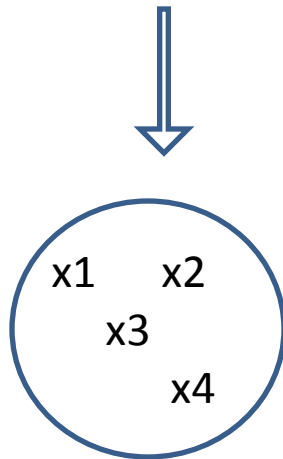
Central  
Limit  
Theorem



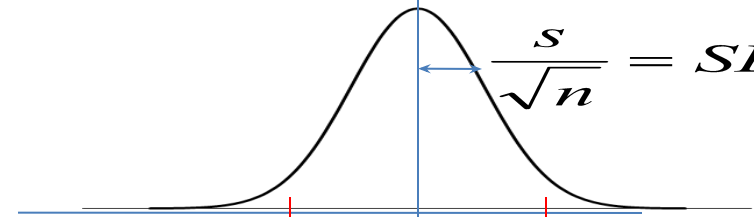
Sampling distribution  
(approx. normal as  $n$  increases)



Sample



Statistical inference:



$$\mu = \bar{X}_{\text{bar}} \pm t_{\text{critical}} (SE)$$

$$\begin{aligned} n &= 4 \\ \bar{X} &= 9 \\ s &= 2 \end{aligned}$$

# Using $t$ to test hypotheses about $\mu$

- Example:

In the general English-speaking Canadian population, the second formant (F2) of /æ/ produced by adult males has a mean frequency of 1550 Hz. A random sample of 64 adult English-speaking Canadian males from a town near the border of the predominantly French-speaking province of Quebec was found to have a mean F2 of 1500 Hz for the vowel /æ/, with a standard deviation of 160. Is this F2 level different from that of the general English-speaking Canadian adult male population?

$H_0: \mu_{\text{town}} = 1550$  (Null hypothesis: not diff from genpop)

$H_1: \mu_{\text{town}} \neq 1550$  (Alternative hypothesis)

# Using $t$ to test hypotheses about $\mu$

- 95% CI

$$SE = s / \sqrt{n} = 160 / \sqrt{64} = 20$$

$$\mu = X_{\text{bar}} \pm t_{\text{critical}}(SE)$$

$$\text{d.f.} = n - 1 = 64 - 1 = 63$$

$$t_{\text{critical}} = \pm 1.998$$

$$\mu = 1500 \pm 1.998(20)$$

$$1460.03 < \mu < 1539.97$$

```
> qt(c(.025, .975), 63)
[1] -1.998341      1.998341
> 1500 + qt(c(.025, .975),
            63) * 20
[1] 1460.033      1539.967
```

# Using $t$ to test hypotheses about $\mu$ (continued)

- Only 5 times out of 100 does the 95% CI **not** contain the mean of the population that the sample was drawn from
- Our 95% CI  **$1460.03 < \mu < 1539.97$**  does not contain 1550, and therefore it is unlikely that the adult male English-speaking residents of the town have the same F2 as the general adult male English-speaking Canadian population.
- Reject  $H_0$

# Hypothesis testing without constructing the CI

- $t = (X_{\text{bar}} - \mu_0) / SE = -2.50$
- d.f. =  $n - 1 = 64 - 1 = 63$ 
  - Compare calculated  $t$  to critical value of  $t$  listed in  $t$ -table
  - Or in R: `> t.test(d, mu = 1550)`

One Sample t-test

```
data: d
t = -2.500, df = 63, p-value = 0.0075
alternative hypothesis: true mean is
not equal to 1550
95 percent confidence interval:
 1460.03    1539.97
...
```

# One-tailed $t$ test

- Suppose we have strong prior reason to believe that the adult male English-speaking residents of the town have F2 lower than the general population
- Consider only one direction of difference for  $H_A$
- $t_{\text{critical}}$  for  $\alpha = 0.05$ , one-tailed test = -1.67 (with d.f. = 63)
- Smaller critical value makes it easier to reject  $H_0$  in predicted direction
  - But cannot reject  $H_0$  if outcome is in opposite direction

I do not recommend this procedure in general.

# Assumptions of single-sample $t$

- Random sampling from a population
- Interval or ratio scale variable
- Independence of scores
- *Normally distributed* population or a sample size that is not too small ( $> 30$ )
  - `skewness, kurtosis`
  - `qqnorm, qqline`
  - `shapiro.test`



# Dependent (paired) $t$ -test

- Each individual has 2 measures (a measure on each of 2 levels of a variable)
  - Ex: performance before & after an intervention  
acceptability of subject & object relatives
- Calculation by hand:
  - Step 1: calculate the difference for each pair of scores (e.g., before – after; subject – object)
  - Step 2: calculate a single sample  $t$ -test on the differences, usually with a  $H_0: \mu_{\text{diff}} = 0$

# Difference method for paired $t$

Subject	Unprimed	Primed	Difference (D)
A	550	510	40
B	1030	980	50
C	760	770	-10
D	600	600	0
E	940	870	70
F	820	760	60
<b>Sum</b>			<b>210</b>
<b>Mean</b>			<b><math>D_{\text{bar}} = 35.00</math></b>
<b><math>s</math></b>			<b><math>S_D = 32.71</math></b>
<b>SE (<math>= s_{D_{\text{bar}}} = S_D / \sqrt{n}</math>)</b>			<b><math>S_{D_{\text{bar}}} = 13.35</math></b>

$$t = (D_{\text{bar}} - \mu_0) / s_{D_{\text{bar}}} = (35 - 0) / 13.35 = 2.62; \quad \text{d.f.} = n - 1 = 5$$

For  $\alpha = .05$ , critical  $t(5) = 2.57$ ; Conclusion: **Reject the null hypothesis**

$$95\% \text{ CI: } \mu_D = D_{\text{bar}} \pm t_{\text{critical}}(s_{D_{\text{bar}}}) = 35 \pm 2.57(13.35); \quad 0.69 \leq \mu_D \leq 69.31$$

# Dependent (paired) *t*-test

```
> unprimed <- c(550, 1030, 760, 600, 940, 820)
> primed <- c(510, 980, 770, 600, 870, 760)
> t.test(unprimed,primed,paired = TRUE)
```

Paired t-test

```
data: unprimed and primed
t = 2.6209, df = 5, p-value = 0.04705
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 0.6720635 69.3279365
sample estimates:
mean of the differences
```

# Test statistics, in general

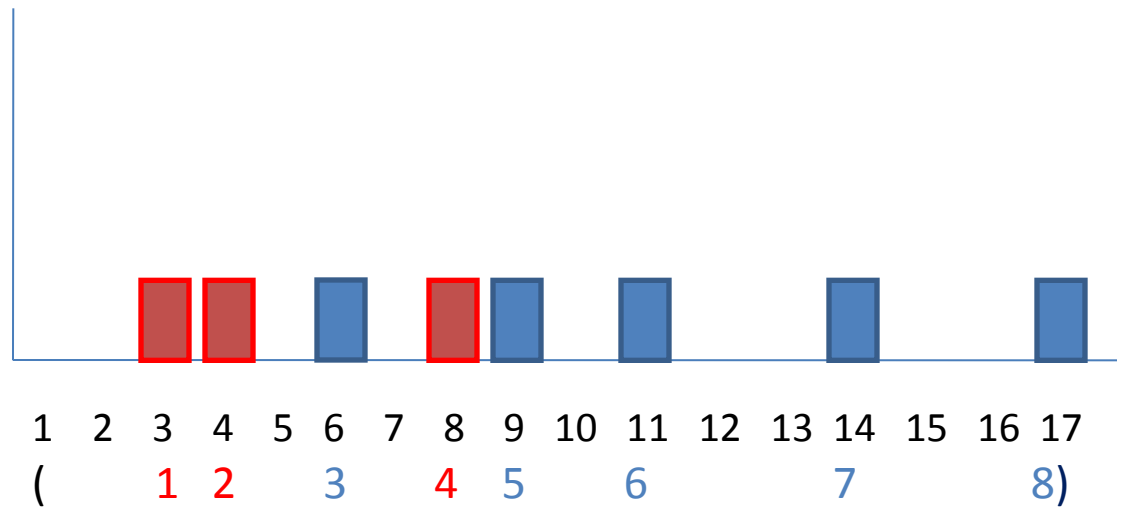
- A statistic for which the probabilities of particular values are known
- Observed values can be used to test hypotheses

$$t = \frac{\bar{X} - \mu}{SE} = \sqrt{\frac{(\bar{X} - \mu_0)^2}{SE^2}} = \sqrt{\frac{(effect)^2}{(error)^2}}$$

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance not explained by the model}} = \frac{\text{effect}}{\text{error}}$$

# Ordinal (rank) statistics

Dialect	Vowel reductions	Rank
A	6	3
B	4	2
A	17	8
B	3	1
B	8	4
A	14	7
A	9	5
A	11	6



Vowel reductions  
(Ranks)

# Wilcoxon rank-sum test (Mann-Whitney $U$ -test)

How unusual is it for the B dialect group, consisting of 3 participants, to have ranks that total 7 ( $= 1 + 2 + 4$ )?

Of all the ways that the 3 participants in the B dialect group could have ranked, how unusual is the ranking that was actually obtained?

If a random process had assigned the ranks to the participants in the 2 groups, how often would the result be as extreme as or more extreme than the one obtained?

# Wilcoxon rank-sum test (Mann-Whitney $U$ -test)

How many possible rankings are there for the 3 B dialect participants?

$$3 \text{ choose } 8 = (8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1) / [(3 \cdot 2 \cdot 1)(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)] = \underline{56}$$

Ranks Total

123 6

124 7

125 8

126 9

127 10

128 11

134 8

135 9

136 10

137 11

138 12

145 10

146 11

147 12

Ranks Total

148 13

156 12

157 13

158 14

167 14

168 15

178 16

234 9

235 10

236 11

237 12

238 13

245 11

246 12

Ranks Total

247 13

248 14

256 13

257 14

258 15

267 15

268 16

278 17

345 12

346 13

347 14

348 15

356 14

357 15

Ranks Total

358 16

367 16

368 17

378 18

456 15

457 16

458 17

467 17

468 18

478 19

567 18

568 19

578 20

678 21

# Wilcoxon rank-sum test (Mann-Whitney $U$ -test)

Question:

How unusual is it for the B dialect group, consisting of 3 participants, to have ranks that total 7 (= 1 + 2 + 4)?

Answer:

Only 2 out of 56 possible assignments of ranks yields a rank total equal to or less than 7. The probability that the observed outcome is the result of chance is  $2 / 56 = 0.0357$ , one-tailed. The two-tailed probability is .0714 (allowing for the possibility that the B dialect group might have ranks on the other end of the distribution, i.e., ranks 8 + 7 + 5).

```
> wilcox.test(c(6, 17, 14, 9, 11), c(4, 3, 8))
```

```
Wilcoxon rank sum test
```

```
data: c(6, 17, 14, 9, 11) and c(4, 3, 8)
```

```
W = 14, p-value = 0.07143
```

```
...
```



**Appendix : Wilcoxon Rank-Sum Table**

Probabilities relate to the distribution of  $W_A$ , the rank sum for group A when  $H_0 : A = B$  is true. The tabulated value for the lower tail is the largest value of  $w_A$  for which  $\text{pr}(W_A \leq w_A) \leq \text{prob}$ . The tabulated value for the upper tail is the smallest value of  $w_A$  for which  $\text{pr}(W_A \geq w_A) \leq \text{prob}$ .

**Lower Tail**

**Upper Tail**

$n_A$	$n_B$	<i>prob</i>						<i>prob</i>					
		.005	.01	.025	.05	.10	.20	.20	.10	.05	.025	.01	.005
4	4			10	11	13	14	22	23	25	26	30	34
	5		10	11	12	14	15	25	26	28	29	33	38
	6	10	11	12	13	15	17	27	29	31	32	37	41
	7	10	11	13	14	16	18	30	32	34	35	40	45
	8	11	12	14	15	17	20	32	35	37	38	43	48
	9	11	13	14	16	19	21	35	37	40	42	47	52
	10	12	13	15	17	20	23	37	40	43	45	50	55
5	5	15	16	17	19	20	22	33	35	36	38	43	49
	6	16	17	18	20	22	24	36	38	40	42	47	53
	7	16	18	20	21	23	26	39	42	44	45	51	57
	8	17	19	21	23	25	28	42	45	47	49	55	61
	9	18	20	22	24	27	30	45	48	51	53	59	65
	10	19	21	23	26	28	32	48	52	54	57	63	69
	11	20	22	24	27	30	34	51	55	58	61	67	73
6	6	23	24	26	28	30	33	45	48	50	52	58	64
	7	24	25	27	29	32	35	49	52	55	57	63	69
	8	25	27	29	31	34	37	53	56	59	61	67	73
	9	26	28	31	33	36	40	56	60	63	65	71	77
	10	27	29	32	35	38	42	60	64	67	70	76	82
	11	28	30	34	37	40	44	64	68	71	74	80	86
	12	30	32	35	38	42	47	67	72	76	79	85	91
7	7	32	34	36	39	41	45	60	64	66	69	75	81
	8	34	35	38	41	44	48	64	68	71	74	80	86
	9	35	37	40	43	46	50	69	73	76	79	85	91
	10	37	39	42	45	49	53	73	77	81	84	90	96
	11	38	40	44	47	51	56	77	82	86	89	95	101
	12	40	42	46	49	54	59	81	86	91	94	100	106
	8	8	43	45	49	51	55	59	77	81	85	87	93
9		45	47	51	54	58	62	82	86	90	93	99	105
10		47	49	53	56	60	65	87	92	96	99	105	111
11		49	51	55	59	63	69	91	97	101	105	111	117
12		51	53	58	62	66	72	96	102	106	110	116	122
9	9	56	59	62	66	70	75	96	101	105	109	115	121
	10	58	61	65	69	73	78	102	107	111	115	121	127
	11	61	63	68	72	76	82	107	113	117	121	127	133
	12	63	66	71	75	80	86	112	118	123	127	133	139
10	10	71	74	78	82	87	93	117	123	128	132	138	144
	11	73	77	81	86	91	97	123	129	134	139	145	151
	12	76	79	84	89	94	101	129	136	141	146	152	158
11	11	87	91	96	100	106	112	141	147	153	157	163	169
	12	90	94	99	104	110	117	147	154	160	165	171	177
12	12	105	109	115	120	127	134	166	173	180	185	191	197

# Handling ties

<b>X</b>	<b>Provisional Rank</b>	<b>Final Rank</b>
6	1	1
8	2	2.5
8	3	2.5
11	4	4
19	5	6
19	6	6
19	7	6
26	8	8

# Example

The following hypothetical data are the numbers of correct responses in a sentence repetition task for patients with left- versus right-hemisphere brain damage:

Left: 5, 3, 8, 6

Right: 9, 13, 8, 7, 11, 6

<u>3L</u>	<u>5L</u>	<u>6L</u>	<u>6R</u>	<u>7R</u>	<u>8R</u>	<u>8L</u>	<u>9R</u>	<u>11R</u>	<u>13R</u>
1	2	3	4	5	6	7	8	9	10
1	2	3.5	3.5	5	6.5	6.5	8	9	10

$$\Sigma R_A = 1 + 2 + 3.5 + 6.5 = 13 \quad 0.05 < p \leq 0.10 \text{ (from Wilcoxon Rank Sum Table)}$$

```
> wilcox.test(c(5, 3, 8, 6), c(9, 13, 8, 7, 11, 6))
```

Wilcoxon rank sum test with continuity correction

```
data: c(5, 3, 8, 6) and c(9, 13, 8, 7, 11, 6)
```

```
W = 3, p-value = 0.06826
```

```
...
```

```
Warning message:
```

```
In wilcox.test.default(c(5, 3, 8, 6), c(9, 13, 8, 7, 11, 6)) :  
cannot compute exact p-value with ties
```

# Wilcoxon rank-sum test (Mann-Whitney Test)

Normal approximation for large  $N (> 20)$   
or tied ranks

$$z = \frac{\sum R_A - .5(n_A)(N + 1)}{\sqrt{\frac{n_A n_B (N + 1)}{12}}}$$

# Repeated-measures (i.e., within-subjects) research designs required for a dependent $t$ -test

- Order effects

Presenting condition A before condition B

- subjects may improve performance over time
- subjects may become fatigued over time

Solution: counterbalancing

half of subjects get A first, then B

other half of subjects get B first, then A

Solution: intermixing (simultaneous repeated measures)

multiple trials representing A and B are randomly interleaved: ABBABABBAAAB....

# Problems (continued)

- Carryover effects

Example:

Administering drug may have permanent or long term effect on subject, so in Drug-Placebo order, Placebo reflects Drug effect

Two methods for ESL teaching cannot be used on same students

- Mutually exclusive subject selection categories, where matching of subjects is difficult or impossible
  - native speaker of English, non-native speaker of English
  - clinically depressed subject, non-depressed subject

# Independent $t$ -test

- Also known as independent samples  $t$ , two-sample  $t$ , independent groups  $t$ , between-subjects  $t$ ...
- For each subject, **dependent variable** is measured on only one level of the **independent variable**
  - **RT to recognize words** but each subject is in either **primed condition** or **unprimed condition**  
(each subject gets one condition)
  - **Proportion of passives** is measured but subject is either **Adult** or **Child** (each subject is in only one group)

# Independent $t$ -test

- Each subject provides data for only one condition
- Differences between conditions also reflect individual differences between subjects
- To calculate an independent  $t$ 
  - (1) for each group, calculate  $X_{\text{bar}}$  and variance
  - (2) calculate standard error of  $X_{\text{bar}1} - X_{\text{bar}2}$
  - (3) compute  $t$  by comparing  $(X_{\text{bar}1} - X_{\text{bar}2})$  to null hypothesis difference of means  $\mu_1 - \mu_2 = 0$



# What is the mean of the sampling distribution of $\bar{X}_{\text{bar1}} - \bar{X}_{\text{bar2}}$ ?

- Suppose the 2 samples come from populations with the same  $\mu$  (i.e., they come from the same population), then:
  - sampling distribution of  $\bar{X}_{\text{bar1}}$  will have a mean of  $\mu$
  - sampling distribution of  $\bar{X}_{\text{bar2}}$  will have a mean of  $\mu$
- so, if populations are the same, then mean value of  $\bar{X}_{\text{bar1}} - \bar{X}_{\text{bar2}}$  will be  $\mu_1 - \mu_2 = 0$

# What is the standard error of $X_{\text{bar}1} - X_{\text{bar}2}$ ?

- sampling distrib. of  $X_{\text{bar}1}$  will have a variance of  $s_1^2 / n_1$
- sampling distrib. of  $X_{\text{bar}2}$  will have a variance of  $s_2^2 / n_2$

- **Variance Sum Law**

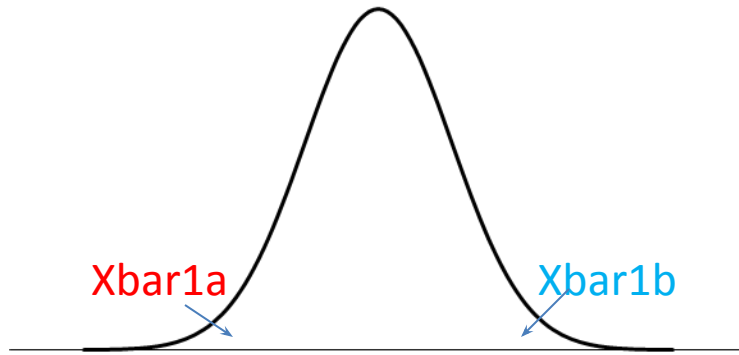
“the variance of the sum or difference of two independent variables is equal to the sum of their variances”

So the variance of  $X_{\text{bar}1} - X_{\text{bar}2}$  is  $s_1^2 / n_1 + s_2^2 / n_2$

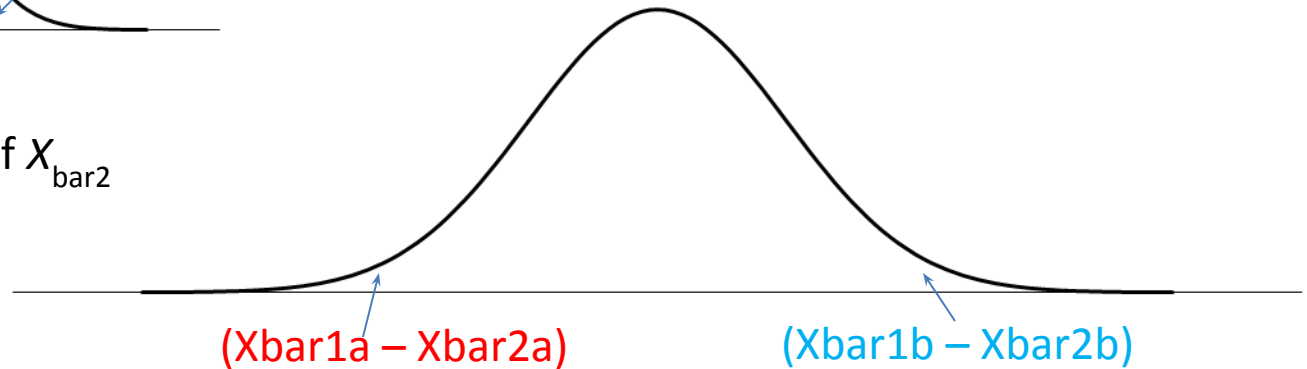
Standard error of  $X_{\text{bar}1} - X_{\text{bar}2}$  is  $s_{X_{\text{bar}1} - X_{\text{bar}2}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

# Sampling distribution of the differences between means

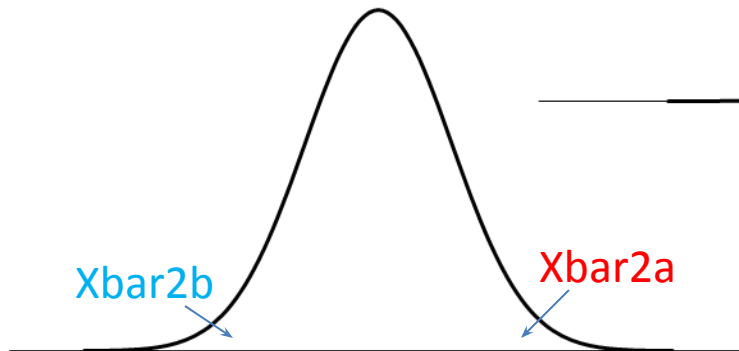
Sampling distribution of  $\bar{X}_{\text{bar1}}$



Sampling distribution of  $\bar{X}_{\text{bar1}} - \bar{X}_{\text{bar2}}$



Sampling distribution of  $\bar{X}_{\text{bar2}}$



	Unprimed (Group 0)		Primed (Group 1)	
	A	550	G	510
	B	1030	H	980
	C	760	I	770
	D	600	J	600
	E	940	K	870
	F	820	L	760
<b>Sum</b>	<b>4700</b>		<b>4490</b>	
<b>Mean</b>	<b>783.33</b>		<b>748.33</b>	
<b>s</b>	<b>187.26</b>		<b>171.98</b>	
<b>s<sup>2</sup></b>	<b>35066</b>		<b>29577</b>	
<b>s<sup>2</sup> / n</b>	<b>5844.33</b>		<b>4929.50</b>	

$$SE = \sqrt{[s^2 / n_1 + s^2 / n_2]} = \sqrt{[5844.33 + 4929.50]} = 103.80$$

$$t = (X_{\text{bar1}} - X_{\text{bar2}}) / SE = (783.33 - 748.33) / 103.80 = .337$$

$$\text{d.f.} = n_1 + n_2 - 2 = 10$$

For  $\alpha = .05$ , critical  $t(10) = 2.23$ ;

Conclusion: **do not reject null hypothesis**

$$95\% \text{ CI: } \mu_1 - \mu_2 = X_{\text{bar1}} - X_{\text{bar2}} \pm t_{\text{critical}} (SE) = 35 \pm 2.23 (103.80)$$

$$95\% \text{ CI: } -196.28 \leq \mu_1 - \mu_2 \leq 266.28$$

# Standard error of the difference between 2 means

- If sample sizes are unequal ( $n_1 \neq n_2$ ), the pooled variance is used in place of each sample variance:

$$S_{pooled}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- So the formula for the SE becomes:

$$\text{Standard error of } X_{\text{bar}1} - X_{\text{bar}2} = s_{X_{\text{bar}1} - X_{\text{bar}2}} = \sqrt{\frac{S_{pooled}^2}{n_1} + \frac{S_{pooled}^2}{n_2}}$$

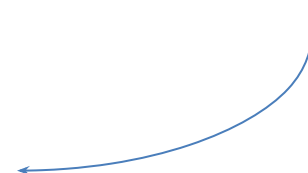
# If the sample variances are unequal

- Pooling of variances should not be used
- Homogeneity of variances can be tested with

```
> unprimed = c(550, 1030, 760, 600, 940, 820)
> primed = c(510, 980, 770, 600, 870, 760)
> var.test(unprimed, primed)
```

F test to compare two variances

not different in  
this example



```
data: unprimed and primed
```

```
F = 1.1856, num df = 5, denom df = 5, p-value = 0.8563
```

```
alternative hypothesis: true ratio of variances is not equal to  
1
```

```
...
```

```
> t.test(primed, unprimed, var.equal = TRUE)
```

```
Two Sample t-test
```

```
data: primed and unprimed  
t = -0.3372, df = 10, p-value = 0.7429  
...
```

By default, R adjusts  
the degrees of  
freedom under the  
assumption of  
unequal variances

```
> t.test(primed, unprimed)
```

```
Welch Two Sample t-test
```

```
data: primed and unprimed  
t = -0.3372, df = 9.928, p-value = 0.743  
...
```

# If data are from paired measurements, dependent $t$ has more power than independent $t$

```
> t.test(unprimed, primed, paired = TRUE)
```

```
Paired t-test
```

```
data: unprimed and primed
```

```
t = 2.6209, df = 5, p-value = 0.04705
```

```
...
```

$t(5) = 2.621, p = .047$  for dependent  $t$  **vs.**  $t(10) = 0.337, p = .743$  for independent  $t$

- Dependent  $t$  removes the variability due to participants (individual differences)
- SE for Independent  $t$  contains variability due to different participants

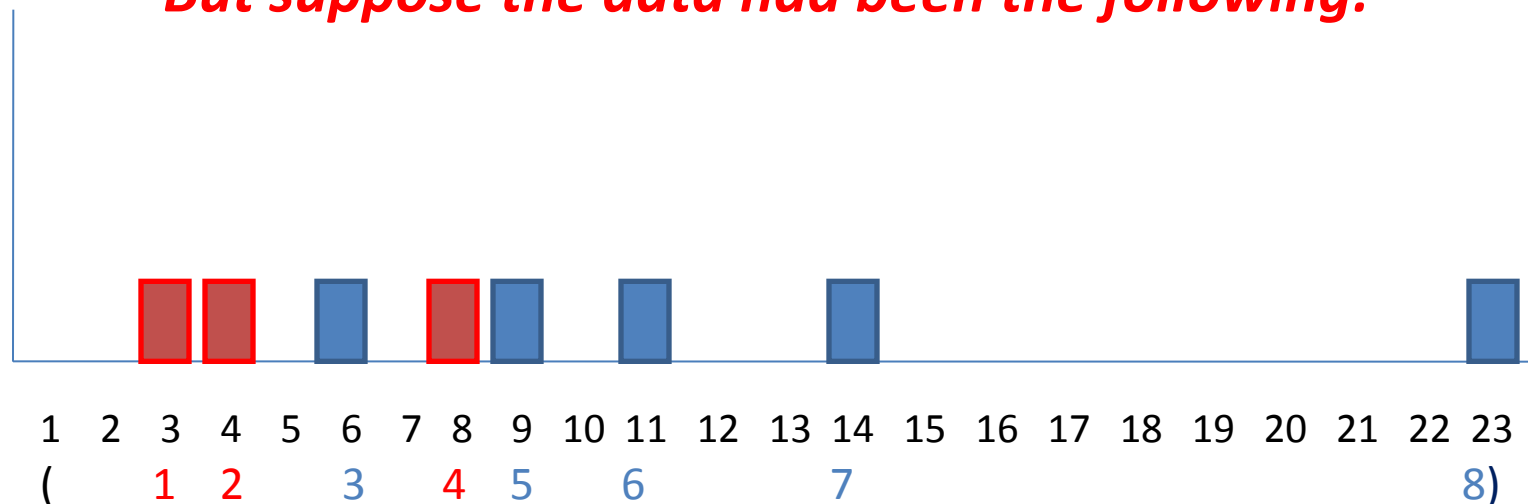


# The independent $t$ -test has more power than Wilcoxon test

Results of an independent samples  $t$ -test on the vowel reduction data:

$$t(6) = 2.30, p = 0.0306, \text{ one-tailed}; p = 0.0612, \text{ two-tailed}$$

*But suppose the data had been the following:*



Results of the Wilcoxon rank-sum test are the same, but now  $t(6) = 1.88$ ,  
 $p = 0.0544$ , one-tailed;  $p = 0.1087$ , two-tailed

# Assumptions of independent $t$ -test

- Random sampling
- Interval or ratio scale dependent variable
- Nominal scale independent variable (groups)
- Independence of scores
- *Approximately normally distributed* populations or sample sizes that are not too small
- Equal variances in the two populations, or use of the Welch version of the test