T-tests

Course: Methoden und Anwendungen

Instructor: Steven Moran, University of Zurich

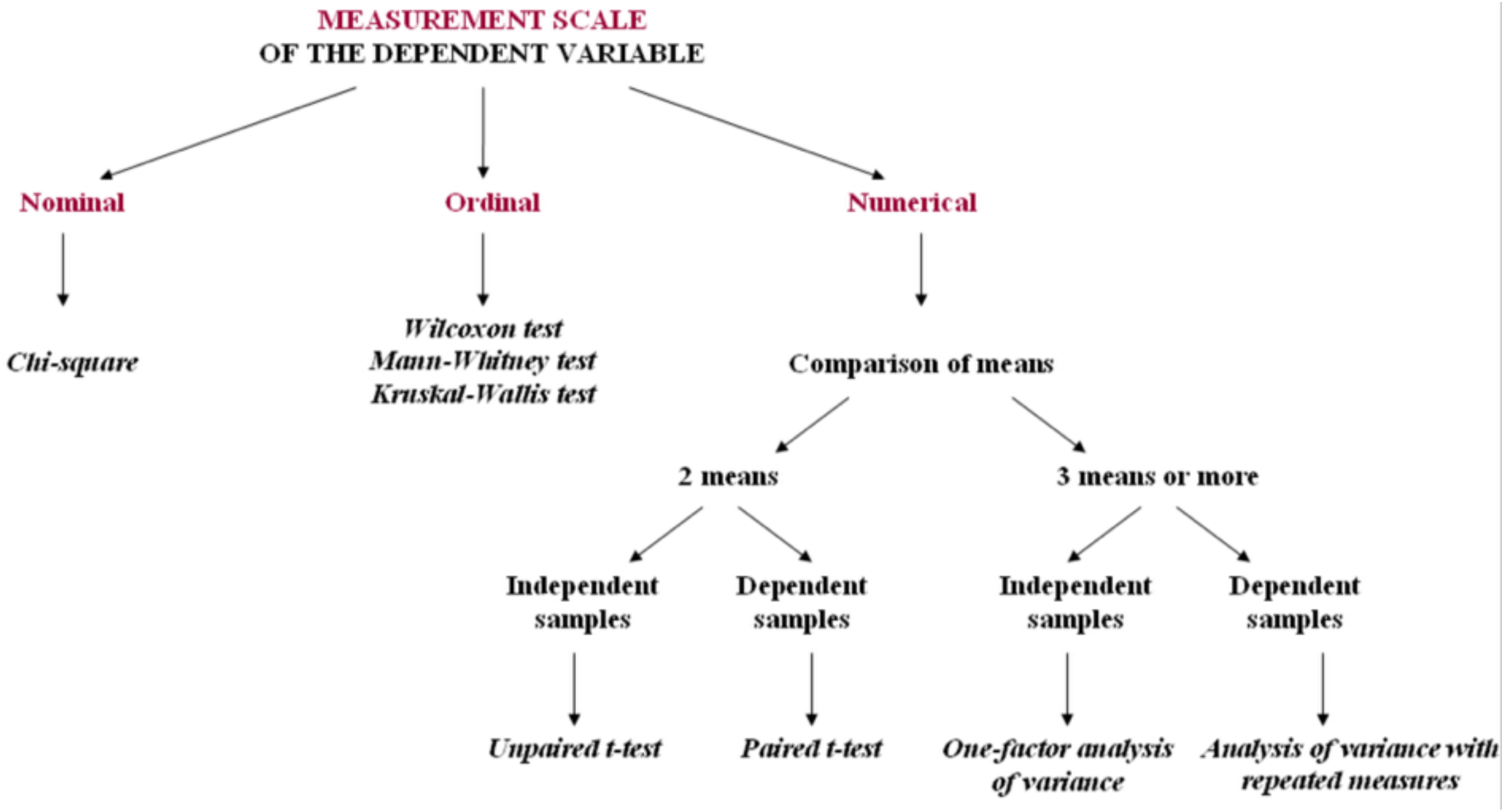
Quantifying and studying variability

1. You compute the effect you observe in your data (e.g., a frequency distribution, a difference in means, a correlation),

2. You compute the so-called probability of error p to obtain the (summed/ combined) probability of the observed effect and every other result that deviates from H0 even more when H0 is true, and

3. You compare p to a significance level (usually 5 percent, i.e., 0.05) and, if p is smaller than the significance level, you reject H0 (because it is not compatible enough with the data to stick to it) and accept H1.

Choosing a test



When to use a particular statistical test?

- http://www.csun.edu/~amarenco/Fcs%20682/When%20to%20use%20what%20test.pdf

Student's t-test

- A t-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported.
- It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.
- When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's t distribution.
- One sample t test. This is the most common type of t test you'll come across in elementary statistics.
- A Paired sample t-test compares means from the same group at different times (say, one year apart).

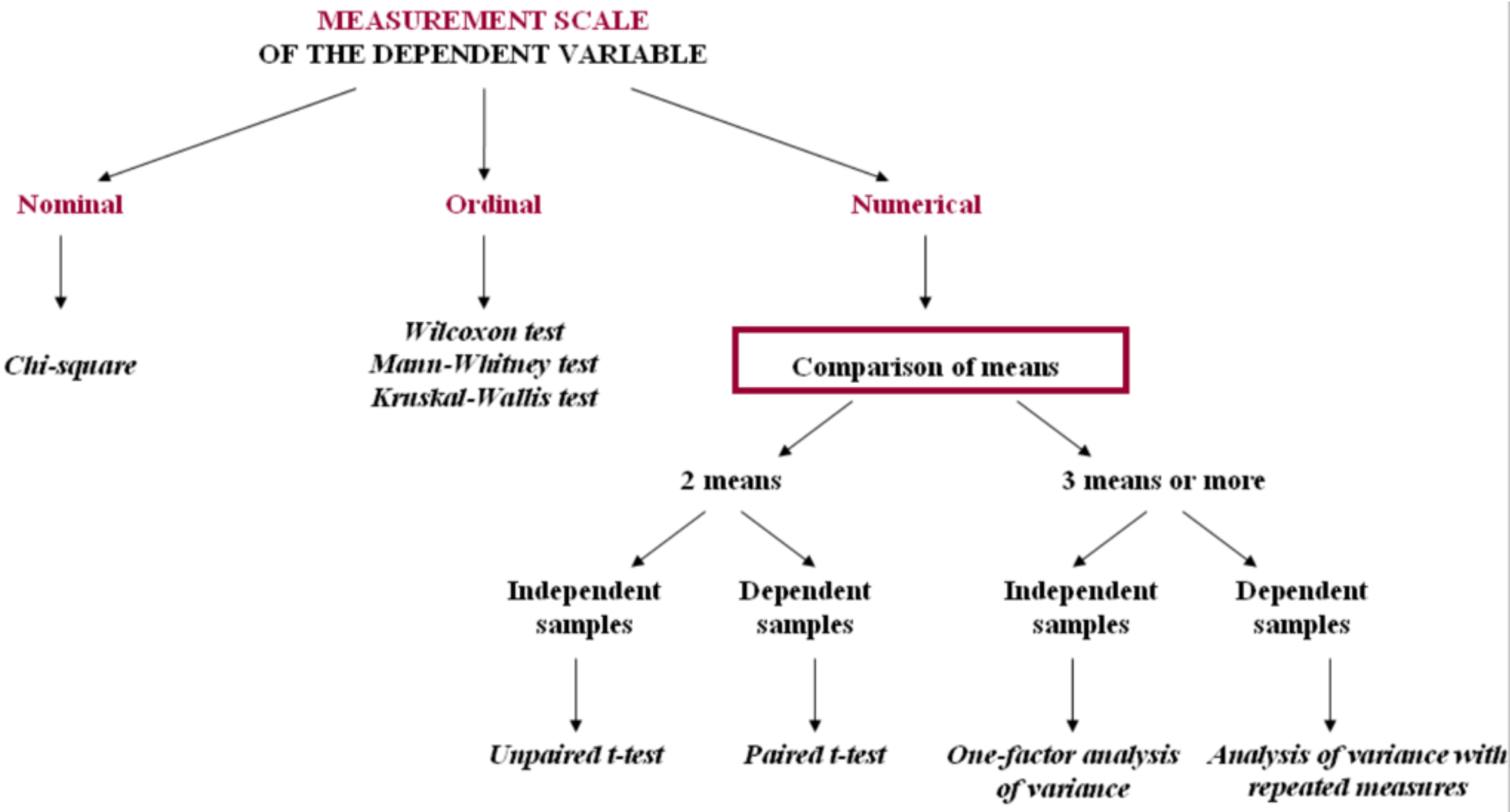
T-test assumptions

- The distributions of the two samples are normal
- The distribution of the two samples have a similar variance

Types of t-tests

- t-test for independent samples (unpaired t-test)
- t-test for dependent samples (paired t-test)

Comparison of means



T-test for independent samples (unpaired t-test)

- 2 independent samples (eg. 2 groups with different individuals)
- Variables
  - dependent: numerical
  - 1 independent: nominal
- Determine whether two means are different

- Example: Do males and females articulate at the same rate?
  - dataRate.txt

T-test for dependent samples (paired t-test)

- 2 dependent samples (e.g. same participants in two different situations)
- Variables
  - dependent: numerical
  - 1 independant: nominal
- Determine whether two means are different

- Example: reaction time before and after training
- Do the aphasics present the same reaction times in a lexical decision task before and after training?
  - dataAph.txt

More examples and detailed explanation

- T-tests in Johnson 2008: 70-82
- See following slides

T-test

- The independent t-test, also called the two sample t-test or student's t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups.
- The null hypothesis for the independent t-test is that the population means from the two unrelated groups are equal:
  - H0: u1 = u2

Null hypothesis: H_0: μ=100

- When we wanted to make probability statements about observations using the normal distribution, we converted our observation scores into z-scores (the number of standard deviations different from the mean) using the z-score formula

- To test a hypothesis about the population mean μ on the basis of our sample mean, we estimate the population standard deviation with the sample standard deviation, and the uncertainty introduced by S instead of σ means that we are off a bit and so can't use the normal distribution to compare $\overline{x}$ to μ

$$z = \frac{x_i - \bar{x}}{s} \quad z\text{-score}$$

- To be a little more conservative, we use a distribution (or family of distributions), called the t-distribution

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \quad t\text{-value}$$

T-distribution

- T-distribution takes into account how certain we can be about our estimate of σ
- Larger sample size gives us a more stable estimate of the population mean, so we get a better estimate of the population standard deviation with larger sample sizes

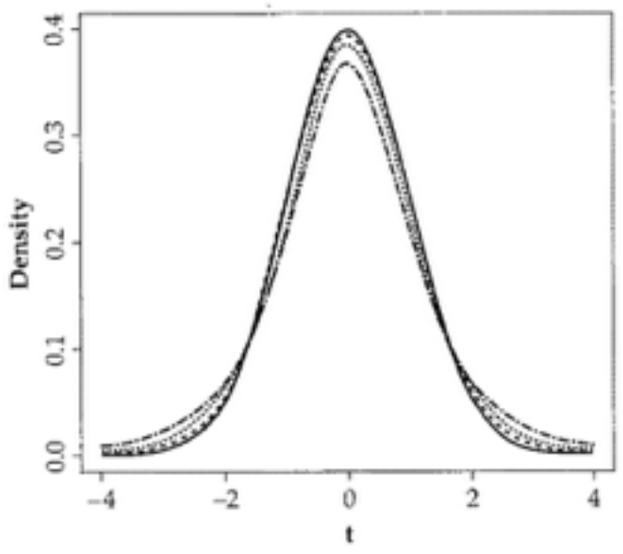


Figure 2.5 The normal distribution (solid line) and *t*-distributions for samples of size $n = 21$ (dash), $n = 7$ (dot) and $n = 3$ (dot, dash).

- Figure 2.5 - normal and t-distributions for 3 samples - the larger the sample size, the closer the t-distribution is to normal
- We use a slightly different distribution to talk about mean values, but the procedure is practically the same as if using the normal distribution
- To make a probability statement about a z-score you refer to the normal distribution, and to make a probability statement about a t-value you refer to the t-distribution

H_0: μ=100

- It may seem odd to talk about comparing the sample mean to the population mean because we we can easily calculate the sample mean but the population mean is not a value that we can know

- If you think of this as a way to test a hypothesis, then we have something. For instance, with the Cherokee VOT data, where we observed that $\bar{x}$=84.7 and s=36.1 for the stops produced in 2001, we can now ask whether the population mean μ is different from 100

$$t = \frac{\bar{x}_i - \mu}{s_{\bar{x}}} = \frac{84.7 - 100}{36.1/\sqrt{26}} = \frac{-15.3}{7.08} = -2.168$$

- But what does -2.168 mean?

- Testing the hypothesis that the average VOT value of 84.7 ms is not different from 100 ms. This can be written as H_0: μ=100

- Meaning that the null hypothesis (the "no difference" hypothesis H_0) is that the population mean is 100

T-test

- The statistic t is analogous to z - it measures how different the sample mean $\overline{x}$ is from the hypothesized population mean μ, as measured in units of the standard error of the mean

- Observations that are more than 1.96 standard deviations away from the mean in a normal distribution are pretty unlikely - only 5% of the area under the normal curve

- So this t-value of -2.168 (a little more than 2 standard errors less than the hypothesized mean) might be a pretty unlikely one to find if the population mean is actually 100 ms. But how unlikely?

T-test

- The probability density function of t with 25 degrees of freedom (since we had 26 observations in the VOT data set) shows that only 2% of all t-values in this distribution are less than -2.16

- Recall that we are evaluating the null hypothesis that μ=100. Therefore, this probability value says that if we assume that μ=100 it is pretty unlikely (2 times in 100) that we would draw a sample that has an $\overline{x}$ of 84.7. The more likely conclusion that we should draw is that the population mean is < 100

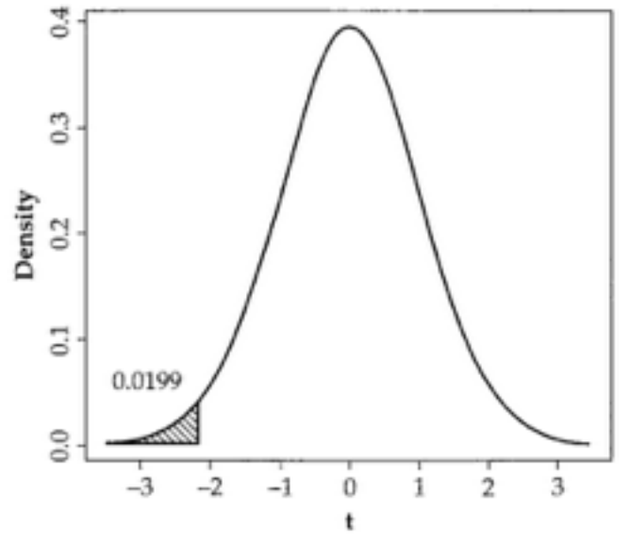


Figure 2.6 The probability density function of t with 25 degrees of freedom. The area of the shaded region at $t < -2.168$ indicates that only a little over 2% of the area under this curve has a t-value less than –2.168.

T-test

- Test the hypothesis that the true Cherokee VOT in 2001 (μ) is 100 ms by taking a sample from a larger population of possible measurements
- If the sample mean ($\overline{x}$) is different enough from 100 ms, then we reject this hypothesis; otherwise we accept it
- How different is different enough?
  - We can quantify the difference between the sample mean and the hypothesized population mean in terms of a probability
  - If the population mean is 100 ms, then in only 2 times in 100 could we get a sample mean of 84.7 or less
- Suppose that we decide then that this is a big enough difference - the probability of a sample of 84.7 mean coming from a population that has a mean of 100 ms is pretty darn low - so we reject the hypothesis that μ=100 (let's label it H_0), and instead accept the alternative hypothesis that $\mu < 100$ (call this H_1 and note that this is only one of several possible alternative hypotheses).

Type I and type II error

- 2 times out of 100 this decision would be wrong

H_0: $\mu = 100$ Reject
H_1: $\mu < 100$ Accept

- Still possible that H_0 is correct - the population mean really could be 100 ms even though our sample mean is a good deal less than 100 ms

- This error probability (0.02) is called the probability of making a type I error

- Type I error is that we incorrectly reject the null hypothesis - we claim that the population mean is less than 100, when actually we were just unlucky and happened to draw one of the 2 out of 100 samples for which the sample mean was equal to or less than 84.7

Hypothesis testing

- No matter what the sample mean is, you can't reject the null hypothesis with certainty because the normal distribution extends from negative infinity to positive infinity

- "going with your best guess" means choosing a type I error probability that you are willing to tolerate

- Most often we are willing to accept a 1 in 20 chance that we just got an unlucky sample that leads us to make a type I error.

- This means that if the probability of the t-value that we calculate to test the hypothesis is less than 0.05, we are willing to reject H_0 (μ=100) and conclude that the sample mean comes from a population that has a mean that is less than 100 (μ<100). This criterion probability value ($p < 0.05$) is called the "alpha" (α) level of the test. The α level is the acceptable type I error rate for our hypothesis test.

Type I and type II errors

- Type II error occurs when we incorrectly accept the null hypothesis
- Suppose that we test the hypothesis that the average VOT for Cherokee (or at least this speaker) is 100 ms, but the actual true mean VOT is 95 ms. If our sample mean is 95 ms and the standard deviation is again about 35 ms we are surely going to conclude that the null hypothesis (H_0: μ=100) is probably true.
- At least our data is not inconsistent with the hypothesis because 24% of the time ($p = 0.24$) we can get a t-value that is equal to or less than -0.706.

Table 2.2 The decision to accept or reject the null hypothesis may be wrong in two ways. An incorrect rejection, a type I error, is when we claim that the means are different but in reality they aren't, and an incorrect acceptance, a type II error, is when we claim that the means are not different but in reality they are.

		Reality: H_0 is true	Reality: H_0 is false
Decision	accept H_0	correct	Type II error
	reject H_0	Type I error	correct

Type I and type II errors

- Type I error is that we incorrectly reject the null hypothesis
- Type II error occurs when we incorrectly accept the null hypothesis

Decision (y) / reality (x)	H0 is true	H0 is false
accept H0	correct	Type II error
reject H0	Type I error	correct

False positives and false negatives

- Airport Security: a "false positive" is when ordinary items such as keys or coins get mistaken for weapons (machine goes "beep")
- Quality Control: a "false positive" is when a good quality item gets rejected, and a "false negative" is when a poor quality item gets accepted
- Antivirus software: a "false positive" is when a normal file is thought to be a virus
- Medical screening: low-cost tests given to a large group can give many false positives (saying you have a disease when you don't), and then ask you to get more accurate tests.

	They say you did	They say you didn't
You really did	They are right!	"False Negative"
You really didn't	"False Positive"	They are right!

False positives and false negatives

	actual class (observation)	actual class (observation)
predicted class (expectation)	tp (true positive) Correct result	fp (false positive) Unexpected result
predicted class (expectation)	fn (false negative) Missing result	tn (true negative) Correct absence of result

Type I and type II errors

- By accepting the null hypothesis we have made a type II error. Just as we can choose a criterion μ level for the acceptable type I error rate, we can also require that our statistics avoid type II errors
- The probability of making a type II error is called β, and the value we are usually interested in is 1-β, the power of our statistical test.
- To avoid type II errors you need to have statistical tests that are sensitive enough to catch small differences between the sample mean and the population mean - to detect that 95 really is different from 100
- With only 26 observations (n=26) and a standard deviation of 36.1, if we set the power of our test to 0.8 (that is, accept type II errors 20% of the time with p=0.2) the difference between the hypothesized mean and the true population mean would have to be 18 ms before we could detect the difference

Calculation of t

- In the calculation of t there are two parameters other than the sample mean and the population mean that affect the t-value.
- These are the standard deviation of the sample (s) and the size of the sample (n).
- To increase the power of the t-test we need to either reduce the standard deviation or increase the number of observations.
- Sometimes you can reduce the standard deviation by controlling some uncontrolled sources of variance. For example, in this VOT data observations from both /t/ and /k/ are pooled. These probably do have overall different average VOT, by pooling them the standard deviation is inflated.
- If we had a sample of all /k/ VOTs the standard deviation might be lower and thus the power of the t-test greater.

Tests

- Best way to increase the power of your test is to get more data
- In this case, if we set the probability of a type I error at 0.05, the probability of a type II error at 0.2, and we want to be able to detect that 95 ms is different from the hypothesized 100 ms, then we need to have an n of 324 observations magic
  - See R code: Johnson-t-test.r (power.t.test() function
- Collecting more data is time consuming
  - Is it worth it?

Comparing mean values: t-tests

- We can test the hypothesis that a sample mean value $\overline{x}$ is the same as or different from a particular hypothesized population mean μ.
- We might want to know if the observed, sample mean $\overline{x}$ is reliably different from zero, but, in many cases the $\overline{x}$-μ comparison is not really what we want because we are comparing two sample means.
- Key question of interest with the Cherokee 1971/ 2001 data is the comparison of two sample means
- Is the mean VOT in 1971 different from the mean VOT in 2001?

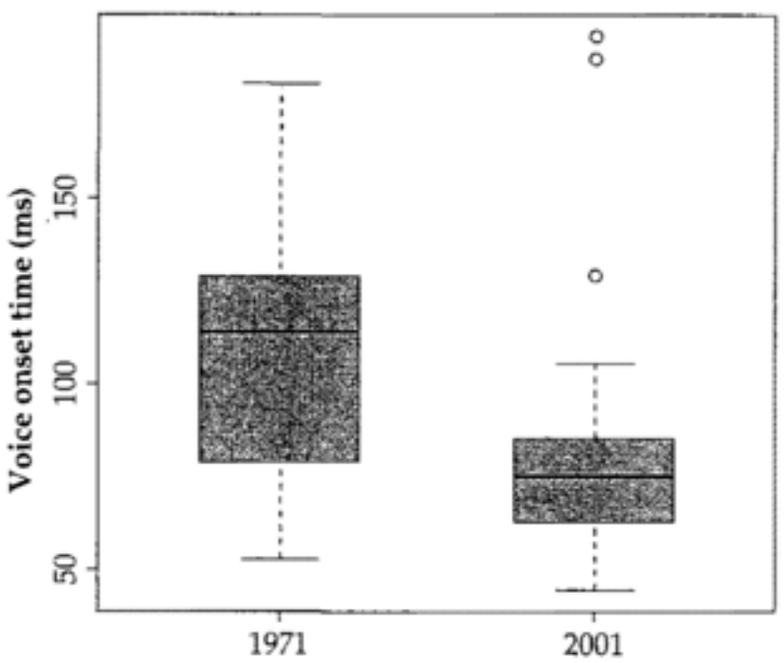


Figure 3.1 A boxplot of voice onset time measurements for Cherokee voiceless stops /t/ and /k/ produced by speaker DF in 1971 and in 2001.

Boxplots

- Short explanation dealing with snow and ski resorts:
  - https://www.youtube.com/watch?v=CoVf1jLxgj4
- Box plot provides a lot of information

Boxplots

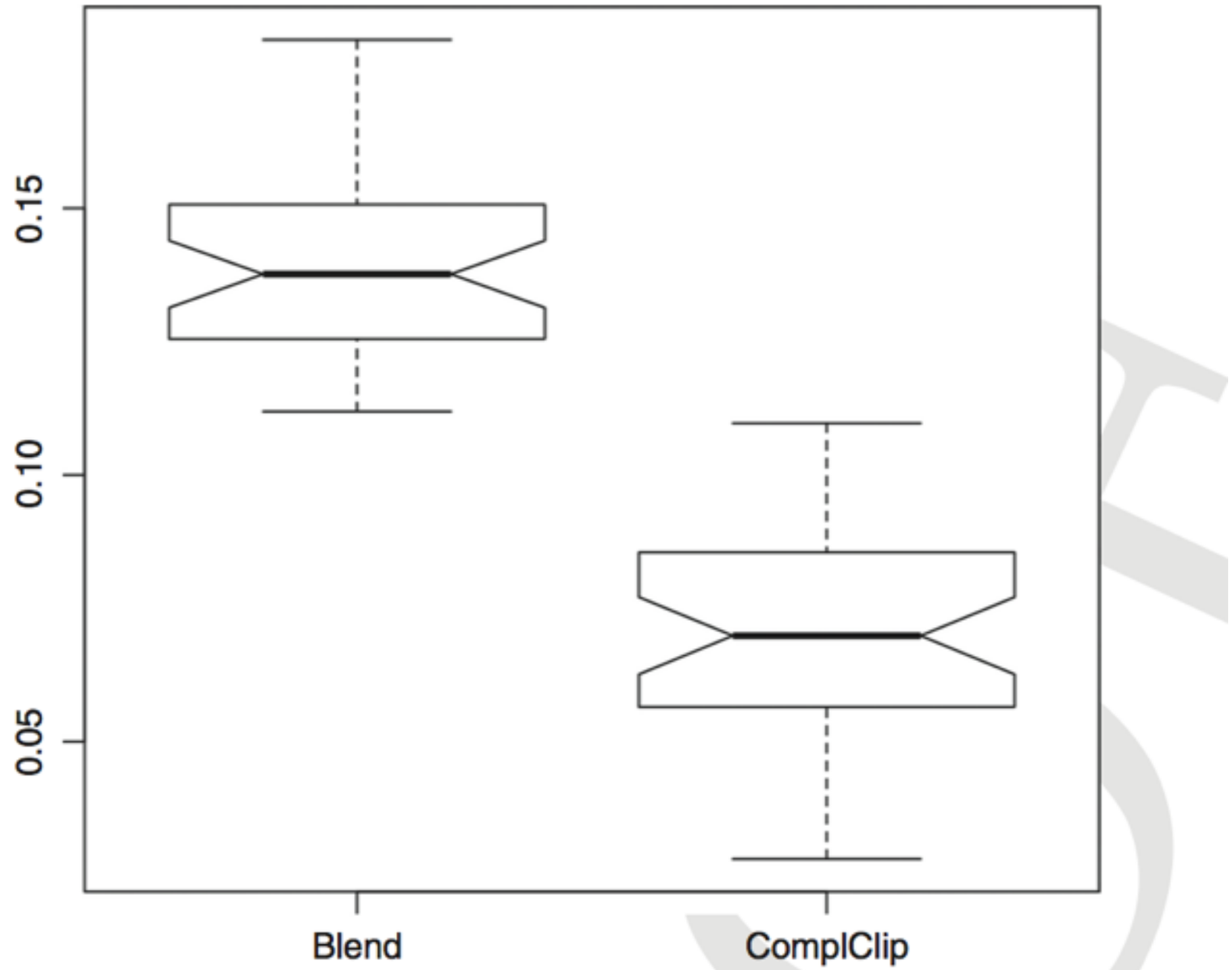


Figure 15.4. *Box plot of the Dice coefficients for the two subtractive word-formation processes*

Cherokee voice onset time: $\mu_{1971} = \mu_{2001}$

- Test whether the average VOT in 1971 was equal to the average VOT in 2001, because we think that for this speaker there may have been a slow drift in the aspiration of voiceless stops as a result of language contact
- *Null hypothesis* that there was no reliable difference in the true population, means for these two years that is: H_0: $\mu_{1971} = \mu_{2001}$
- Test this hypothesis with a t-test similar to the "one sample" t-test

$$t = \frac{\bar{x} - \mu}{SE}$$

the *t*-statistic is the difference between observed and expected mean, divided by standard error of the observed mean

- Tested the null hypothesis: H_0: $\mu_{1971} = \mu_{hypothesis}$ where we supplied the hypothesized population mean.
- The idea with the t-test is that we expect the difference between means to be zero - the null hypothesis is that there is no difference and we measure the magnitude of the observed difference relative to the magnitude of random or chance variation we expect in mean values (the standard error of the mean).
- If the difference between means is large, more than about two standard errors (a t-value of 2 or -2), we are likely to conclude that the sample mean comes from a population that has a different mean than the hypothesized population mean.

t-statistic

- in testing whether the mean VOT in 1971 is different from the mean VOT in 2001 for this talker we are combining two null hypotheses

$$\text{H0: } \mu_{1971} = \mu$$
$$\text{H0: } \mu_{2001} = \mu$$
$$\text{H0: } \mu_{1971} = \mu_{2001}$$

- The expected mean value of the 1971 sample is the same as the expected value of the 2001 sample, so just as with a one-sample t-test the expected value of the difference is O. Therefore we can compute a t-statistic from the difference between the means of our two samples

$$t = \frac{\bar{x}_{1971} - \bar{x}_{2001}}{SE} \quad \text{the two-sample } t\text{-value}$$

- Two samples of data, one from 1971 and one from 2001 and therefore we have two estimates of the standard error of the mean (SE). In calculating the t-statistic we need to take information from both the 1971 data set and the 2001 data set when we compute the SE for this test.

Samples have equal variance

- Test the null hypothesis that one Cherokee speaker's average VOT was the same in 1971 and 2001, i.e. H_0: $\mu_{1971} = \mu_{2001}$
- Our sample estimates of the means are easy - $\overline{x}_{1971}$ and $\overline{x}_{2001}$ are the least squares estimates of these parameters
- What is our estimate of the standard error of the mean?
  - With only one sample we used the standard deviation or the variance of the sample to estimate the standard error:

$$SE = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}}$$ the usual definition of sample standard error of the mean

- See R code: basics.r (Johnson 2008:77)

Samples do not have equal variance

- Instead of using the pooled variance, we calculate the standard error of the mean as:

$$SE = \sqrt{(s_a^2/n_a + s_b^2/n_b)}$$ standard error of the mean for unequal variances

- The t-value calculated with this estimated standard error ($t^* = (\overline{x}_a - \overline{x}_b)/SE$) follows the normal distribution if both samples have greater than 30 data points ($n_a > 30$ & $n_b > 30$)

- For smaller samples the t distribution is used with a degrees of freedom equal to:

- By adusting the degrees of freedom, this correction puts us on a more conservative version of the t-distribution.

$$df = \frac{U^2}{V^2/(n_a - 1) + W^2/(n_b - 1)} - 2$$

where

$V = s_a^2/n_a$, the Welch correction of degrees of freedom
$W = s_b^2/n_b$,

and

$$U^2 = V + W$$

Paired t-test

- For observations that naturally come in pairs
- For example, there is no rational way to pair VOT measurements from 1971 with measurements taken in 2001. We could suggest that VOTs taken from ItI should be matched with each other, but there is no meaningful way to choose which 2001 /t/ VOT should be paired with the first /t/ VOT on the 1971 list, for example. Now, if we had measurements from the same words spoken once in 1971 and again in 2001 it would be meaningful to pair the measurements for each of the words.
- The first formant data in "F1data.txt" was given for men and women for each language and vowel in the data set, so that it is natural to match, for example, the male F1 of /a/ in Sele with the female F1 of /a/ in Sele, the male F1 of /i/ in Sele with the female F1 of /i/ in Sele, and so on.

Paired t-test

- Men and women tend to have systematically different vowel F1 frequency, but that the difference between vowels can be bigger than the overall male/female difference

- To have a sensitive test of the male/female difference we need to control for the vowel differences - so we pair male/female differences by vowels, which gives us such control

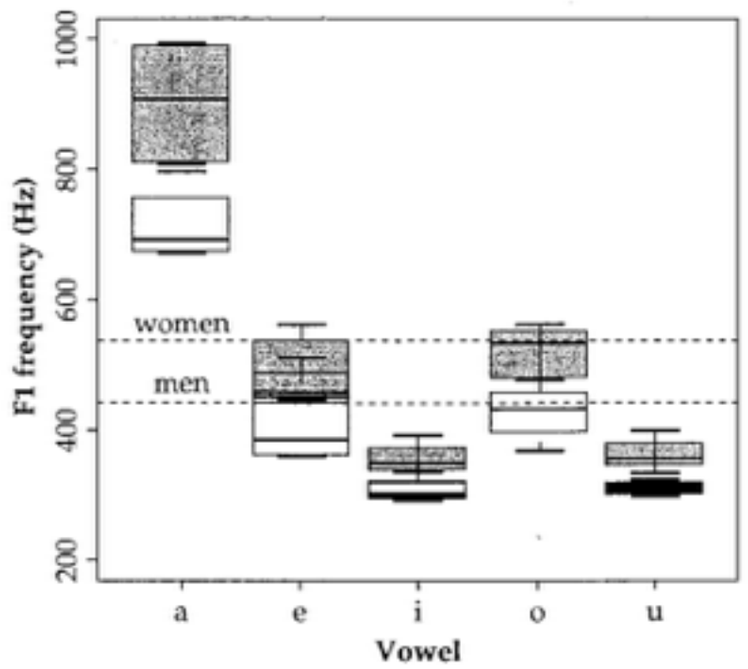


Figure 3.3 Boxplot comparing male (white boxes) and female (gray boxes) vowel F1 frequencies for five different vowels. The overall average F1 values for men and women are plotted with dashed lines.

Paired t-test

- With paired observations we can then define a derived variable - the difference between the paired observations

$$d_i = x_{ai} - x_{bi} \quad \text{the difference between paired observations}$$

- Then calculate the mean and variance of the difference

$$\bar{d} = \frac{\sum d_i}{n}, s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} \quad \text{the mean and variance of the differences}$$

- Test the null hypothesis that there is no difference between the paired observations, i.e. that H_0: $\overline{d}=0$. The t-value, with degrees of freedom n-1:

$$t = \frac{\bar{d}}{\sqrt{s_d^2/n}} \quad \text{does } \bar{d} \text{ differ from zero?}$$

Paired t-test

- The beauty of the paired t-test is that pairing the observations removes any systematic differences that might be due to the paired elements.
- For example, this measure of F1 difference is immune to any vowel or language influences on F1. So, if F1 varies as a function of language or vowel category these effects will be automatically controlled by taking paired F1 measurements from the same vowels spoken by speakers of the same language.
- The paired t-test tends to be much more sensitive (powerful) than the two-sample t-test.
- The increased power of the paired t-test is seen in a comparison of "independent samples" and "paired comparisons" tests of the null hypothesis that male F1 frequencies are no different from female F1 frequencies.

Paired t-test

- The independent samples comparison suggests that men and women are not reliably different from each other [$t(36) = 1.5$, $p = 0.067$], while the paired comparison results in a much higher t-value [$t(18) = 6.1$, $P < 0.01$] showing a much more reliable difference

- Evidently, in the independent samples comparison, the gender difference was swamped by variation due to vowel and perhaps language differences, and when we control for these sources of variation by contrasting men's and women's F1 values for matched vowels and languages, the gender difference is significant.

- Standard style for reporting t-test results: put the t-value and its associated degrees of freedom and probability value in square brackets

- The degrees of freedom is in parentheses after the letter "t," and the probability is reported as $p = 0.067$ if the value is above the Type I error criterion $\alpha(0.01)$, or if the probability of t is less than a then you should report $p < 0.01$