MPATC-GE 2042:
Psychology of Music

Harmony and Tonality
Statistics Part II
Assignment reminder

• Literature review due October 28 at 11:55pm
  – The literature review is an in-depth discussion of prior work relevant to your research topic. It cites the references you have already compiled in your annotated bibliography, summarizing and reflecting on their main findings, problems, and open questions.
  – It should be a cogent, well structured outline of the major prior work in your topic area. Your review should point out divergent perspectives and try to tie together prior work in a way that concisely summarizes the main issues for an unfamiliar reader.
  – You should try to group papers by theme; for example, if you find there are two competing schools of thought about your research topic, you should highlight this in your review and summarize the relevant articles in light of this larger context.
  – The literature review will serve as a major portion of the Introduction to your final research report.
Literature review requirements

• Your review must tell a coherent story for your specific topic. This story and the general structure of the literature review must be discussed and agreed upon by the group.

• Include a short (a few sentences at most) topic summary at the beginning of your document.

• Include the bibliography (without annotations) at the end of the document.

• Use the introduction/background sections in the journal articles you’ve read as a model for how you should write your review.

• Make sure you use correct APA citation style (both within the main text and in parentheses).
Literature review requirements

• Group participation expectations
  – You should discuss the structure of the literature review with your group and develop an outline together
  – All students in the group should read and be familiar with all the work cited
  – Indicate who did the actual writing

• *IMPORTANT NOTE*: Some groups have bibliographies that were too general and/or scattered; if this is the case, you must make sure to update your references instead of trying to write a lit review from your original, problematic list of references.
PART II: Inferential statistics
Inferential statistics

• Review: descriptive statistics – summarizes data
  – Can be applied to any measurements (quantitative or qualitative).
  – Offers a summary/overview/description of data. *Does not explain or interpret anything.*

• Next: inferential statistics - used to make inferences and draw conclusions.
  – Statistics used to interpret the meaning of descriptive statistics
  – Allows researchers to infer or generalize observations made with samples of the larger population from which they were selected

• Inferential statistics are about looking differences between or among the group or categories of the independent variable (IV) in relationship to the dependent variable (DV)
Populations and samples

- Inferential statistics are used to draw conclusions about a population by examining a sample.
- A population is defined as the set of all individuals, items, or data of interest - this is the group about which scientists will generalize.
- A sample is defined as a set of selected individuals, items, or data taken from a population of interest.
- Researchers collect samples only because they do not have access to all individuals in a population.
- Accuracy of inference depends on representativeness of sample from population.
- Random selection: equal chance for anyone to be selected makes sample more representative.
Alternative and null hypotheses

- Inferential statistics test the likelihood that the alternative (research) hypothesis ($H_1$) is true and the null hypothesis ($H_0$) is not.
- In testing differences, $H_1$ would predict that differences would be found, while $H_0$ would predict that no differences would be found.
- By setting the significance (alpha) level (generally at .05), the researcher has a criterion for making this decision:
  - If the .05 level is achieved ($p$ is equal to or less than .05), then a researcher rejects $H_0$ and accepts $H_1$.
  - If the .05 significance level is not achieved, then $H_0$ is retained.
Degrees of freedom

• Degrees of freedom (df) specify how many values vary within a statistical test

• Degrees of freedom are a function of sample size and the number of independent variables

• Scientists recognize that collecting data can never be error-free
  – Each piece of data collected can vary, or carry error that we cannot account for
  – Including df in statistical computations helps account for this error.

• There are clear rules for how to calculate df for each statistical test

• Good article explaining df:
Test results for a hypothesis

• If we reject $H_0$ and conclude groups are really different, it doesn’t mean they’re different for the reason you hypothesized... there may be another reason

• Since $H_0$ testing is based on sample means, not population means, there is a possibility of making an error or wrong decision in rejecting or failing to reject $H_0$
  – Type I error
  – Type II error
Type I and Type II error example

<table>
<thead>
<tr>
<th>Decision</th>
<th>True State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Null Is True</td>
</tr>
<tr>
<td></td>
<td>(Innocent)</td>
</tr>
<tr>
<td>Reject Null</td>
<td>Type I Error</td>
</tr>
<tr>
<td>(Find Guilty)</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>Accept Null</td>
<td>Correct Decision</td>
</tr>
<tr>
<td>(Find Innocent)</td>
<td>Type II Error</td>
</tr>
</tbody>
</table>

Correct Decision
Null Is False (Guilty)
Errors in hypothesis testing

- **Type I error** - rejecting $H_0$ when it is true
  - Chances of this happening are equal to alpha
  - If $\alpha = .05$, then there’s a 5% chance of a Type I error
- **Type II error** - accepting $H_0$ when it should have been rejected
  - If you increase $\alpha$, you will decrease the chance of a Type II error
- Type I errors are generally considered more serious by researchers
- Type II errors might occur in cases where there aren’t enough data
The *t*-test is commonly used to examine whether two groups are significantly different from each other.

A *t* of zero is expected for the null hypothesis to be true.

In order to determine whether *t* is significant at a given alpha level, you need to know the *df*.

- In the case of a *t* test between two groups, \( df = n_1 + n_2 - 2 \) (where *n* is the number of data points in each group).

Other variations of the *t*-test: one sample, paired sample.
Example from Lalitte et al., 2009

Participants with music training reported higher familiarity with the excerpts, whether they were tonal, \( t(50) = 5.20, p < .001 \), or atonal, \( t(50) = 3.03, p < .005 \). Listeners generally perceived tonal pieces as more familiar than atonal ones, \( t(51) = 4.11, p < .002 \). Similarity judgments
The *F*-test, or analysis of variance (ANOVA) is an extension of the *t*-test. It’s a more general statistical test that can be used to ask whether there is a difference among *three or more groups* or to evaluate the results of factorial designs.

A regular ANOVA is used for between-subjects designs (i.e., comparisons between different groups of subjects).

A repeated-measures ANOVA is used when each subject takes part in more than one condition (i.e., subjects are measured repeatedly).
Example from Lalitte et al., 2009

A GLM repeated measures ANOVA was performed on arousal responses, with tonality/atonality as a within-subject factor and musical expertise as a between-subject factor. There was a main effect of listeners’ musical expertise, $F(1, 47) = 7.67, p < .01$ ($MSE = 17.48$), a main effect of musical tonality, $F(1, 47) = 9.39, p < .005$ ($MSE = 7.05$), and a significant two way interaction, $F(1, 47) = 8.85, p = .005$ ($MSE = 6.63$). Musicians perceived tonal music as less stimulating than atonal music (3.95 versus 4.75, $p < .005$), whereas nonmusicians considered them as equally arousing (4.98 versus 4.99). Tonal pieces were judged as more relaxing by musicians than non-musicians for Waldstein, $F(1, 34) = 23.68, p = .005$ and Tempest, $F(1, 34) = 5.03, p < .05$. In atonal pieces the relations were less systematic and less salient. The arousal of Stadwilen and Estempt as estimated by musicians and nonmusicians did not differ significantly.
Results and Discussion
A three-way mixed-design analysis of variance (ANOVA) with within-subjects factors of familiarization meter (simple vs. complex) and alteration type (structure preserving vs. structure violating) and a between-subjects factor of musical training (high vs. low) revealed significant main effects of meter, $F(1, 48) = 18.18$, $p < .001$, $\eta^2 = .28$, and alteration type, $F(1, 48) = 88.59$, $p < .001$, $\eta^2 = .65$, and a significant interaction between meter and alteration type, $F(1, 48) = 117.12$, $p < .001$, $\eta^2 = .71$. Inspection of the top panel of Figure 3 shows
same tempo range. The difference between drummers was confirmed in a two-way analysis of covariance with drummer as one factor (4 levels), tempo as regression variable, and with swing ratio as the dependent variable. Not surprisingly, the influence of tempo was highly significant ($F = 186, p < .0001$) but also the influence of drummer was significant ($F = 4.4, p < .01$). The interaction of tempo and drummer was not significant ($F = 2.6, p < .07$).
Example from Schulkind & Davis, 2012

Results
All analyses in both Experiments 1 and 2 were performed with alpha set equal to .05. The degrees of freedom were adjusted using the Greenhouse-Geisser procedure when appropriate, but the unadjusted values are reported throughout. The type I error rate for the pairwise comparisons was controlled using the Bonferroni adjustment to alpha.
Chi-square tests ($\chi^2$)

- Use this with categorical data where you have frequency counts for each group
- Goodness-of-fit test – comparing one group of data against expected values
  - Null hypothesis is that the actual data fit the expected data
- Test of independence – comparing two groups of data
  - Null hypothesis is that the two variables are independent (i.e., there is no difference in the distribution of responses to the two groups)
Example from Deutsch et al., 2006

CCOM than for the ESM group. Statistical analyses on the numbers of subjects meeting the criteria for absolute pitch showed overall effects of age of onset of musical training (with no semitone errors allowed: $\chi^2=11.44$, df=2, $p <0.01$; with semitone errors allowed, $\chi^2=16.85$, df=2, $p <0.001$). Comparisons between the CCOM and ESM groups were therefore made separately for each level of age of onset of musical training (Fisher Exact Probability Tests were used for all remaining comparisons). All comparisons were highly significant: With no semitone errors allowed, for those who began musical training at ages 4–5, $p <0.001$; at ages 6–7, $p <0.001$; and at ages 8–9, $p <0.005$. With semitone errors allowed, for those who began musical training at ages 4–5, $p <0.001$; at ages 6–7, $p <0.001$; and at ages 8–9, $p <0.005$. Comparisons were also made between the male and
Confusion errors. The subjects failed to identify the target on 639 of the 1290 trials in the experiment (43 subjects x 30 trials per subject). Of these, 11 (1.7%) were songs that the subject claimed were unfamiliar, and an additional 555 (86.9%) were non-responses; that is the subject failed to identify the song, but did not provide the name of an alternative melody. The remaining 73 (11.4%) were trials for which the subject provided an incorrect title. Separate $\chi^2$ analyses indicated that confusion errors were distributed equally across conditions ($\chi^2 (4) = 6.78, p \approx .15$), and that most ($\approx 60\%$) were made with the lowest confidence rating ($\chi^2 (2) = 37.48, p < .001$; note that ‘3’ was the lowest confidence rating a subject could use while providing a specific response. Nineteen (26.0%) of the confusion errors were drawn from the same genre as the target; the implications of these errors will be addressed in the
Reporting statistics

• Whether an $F$ value is significant or not at a given alpha level is dependent on two degrees of freedom reported in parentheses:
  \[ F(1, 48) = 2.06, \ p = .16 \]

• $t$ test has one $df$ reported in the parentheses:
  \[ t(19) = 2.78, \ p < .01 \]

• Chi-square test has one $df$ as well as number of data points reported in parentheses:
  \[ \chi^2(1, \ N = 500) = 3.80, \ p = .051 \]
Effect sizes

• Recall that for correlation and regression we had $r$ and $R^2$; they represent the strength or size of a particular relationship/effect

• $t$ test and ANOVA
  – Cohen’s $d$ expresses effect size in terms of standard deviation units. A $d$ value of 1.0 tells you that the means are 1 standard deviation apart; a $d$ of 0.2 indicates that the means are separated by 0.2 standard deviation; small = 0.20; medium = 0.50; large = 0.80

• ANOVA
  – $\eta^2$ (eta-squared) and $\eta^2_p$ (partial eta-squared), and $\eta^2_G$ (generalized eta-squared) – total variance attributable to a given factor; similar to $R^2$; small = 0.01; medium = 0.06; large = 0.14
  – $\omega^2$ (omega-squared) – estimates proportion of variance in the population that is explained by the condition; less biased (and always smaller) than $\eta^2$ because measures variance in the population as opposed to the sample

• Chi-square tests
  – $\phi_C$ (Cramér’s phi) – small effect = 0.10; medium = 0.30; large = 0.50.
  – Odds ratio – odds of being in one group (e.g., success) relative to odds of being in a different group (e.g., failure)
Examples from your readings: Summary

• Deutsch et al., 2006
  – $\chi^2$ (chi-squared) tests; additional Fisher’s exact tests
• Hannon & Trehub, 2005
  – Mixed-design ANOVA, correlation (Pearson’s $r$)
• Friberg, A., & Sundström, 2002
  – Analysis of covariance
• Schulkind & Davis, 2012
  – Multiple repeated-measures, one-way ANOVAs, $\chi^2$ tests
• Lalitte et al., 2009
  – Two-way repeated-measures ANOVAs, $t$-tests, correlation, multiple regression
Selecting the appropriate test: Descriptive tests (from *Basic Statistics: Tales of Distributions* by Chris Spatz)
Selecting the appropriate test: Inferential tests (from *Basic Statistics: Tales of Distributions* by Chris Spatz)
Selecting the appropriate test, both descriptive and inferential (from *Fundamental Statistics for the Behavioral Sciences* by David Howell)
Article discussion: Lalitte et al. (2009)

• Discussion leaders: Mark Allen and Johnny Butler
If the overall form of the pieces stayed the same, how is it plausible that to a non-musician the tonal and atonal forms were equally arousing? Is this an issue of hearing the pitches themselves or an issue of relying on chord functions and general harmonic motion forward? (Shannon)
Reading question: Task order

• 1) The procedure of the experiment contains 3 tasks for listening to the tonal and atonal manipulations of the pieces: 1st discrete indication of onset of new musical ideas 2nd stimulating/relaxing continuous response 3rd continuous response for new musical ideas

• Does the ordering of these tasks within the study influence participant responses within these tasks? Could repeated listening change a participants response on a task related to emotional arousal such as stimulation/relaxation ratings? (Henry)
Reading question: Task, continued

• Thoughts (Henry):
  My hunch would be that any emotional or arousal response to a piece of music would be dampened with each repeated listening session (strongest response when hearing for the first time). In other words, the participant's arousal rating may be less since they know what is going to come next and it is not a surprise. If this is the case (it may not be), are the researchers losing insights to arousal ratings by collecting them on the participants' second listen?

  On a similar note, would it be better to ask questions about structure and onsets of musical ideas on the second or third listen (not first)? Knowing how the whole piece unfolds prior to rating idea onsets may impact participant response.
The study concluded that tonal structure does not have as much of an influence on perceiving structure of musical ideas as previously believed. The atonal versions of the songs did not seem to deeply alter the structure and form of the pieces, but other properties such as 'repetition, contrast, amplification, progression, contradiction, conclusion, and so on.' The researchers rule out tonal structure having an impact, but leave the door wide open for many other musical properties to determine perceived structure. They also refer to musical 'rhetoric' staying the same even when tonality is altered.

What would be key follow up experiments the researchers could perform to narrow down the possible determining factors of long term musical structure? How could one alter musical 'rhetoric' in a follow up study? (Henry)
Reading question: Follow-up experiment

• In this experiment, listeners found the two tonal excerpts to be similar and the two atonal variants to be similar. Even though the results of the tasks demonstrated that listeners perceived very similar structures between original/atonal variants, they appeared mainly to consider whether the piece was tonal/atonal or classical/contemporary in making a judgement. Under what conditions (if any) might listeners use large-scale form as a primary justification for describing similarity between compositions? In other words, how might we manipulate the experiment to encourage listeners to consider form? For example, what if a similar experiment was performed, but the tonal excerpts had dramatically different forms - e.g. a classical sonata with clearly defined sections and a minimalist work where motives gradually fade in and out. (Willie)
Reading question: Stimuli

Besides the fact of having selected performances from the same pianist, the researchers also selected compositions from the same composer (Beethoven). Results reported that participants found more similarity between the two original versions, which belong to the composer’s unique style, as opposed to their counterparts, and also more familiarity with the tonal than with the atonal. Would there be any difference in the results if the two pieces selected corresponded to different composers and/or different performers? Considering the reading about tonal recognition, could taking the tonal hierarchies out be causing the participants to not be able to familiarize with the atonal samples? (Julian T.)
Reading question: Familiarity and similarity

• Given that the participants in this experiment were “unaware of the temporal similarity within the original-counterpart pairs” (p227), do you think that they would have recognized the temporal similarities had they been exposed to the stimuli for longer? Might it be the case that while pitch content is the primary property of music that is consciously analyzed when listening initially, we will be able to consciously analyze temporal structures after becoming familiar with a piece’s tonality or lack thereof? (Tyler)

• The article stated in the results that pitch was the most influential aspect of familiarity, however the similarity rating was more commonly rated between the two totally separate compositions instead of between the atonalized version of the same sonata. If this is the case, what further studies could be done to test that difference between the pitch familiarity and the harmonic similarity? Is there a line where familiarity and similarity start to overlap? (Shannon)
Reading question: Altering temporal qualities instead of pitch

• What would happen if instead of the pitch content of the samples being altered, the temporal qualities were altered instead? If participants were asked to identify similarities between a control sample and a sample with a complex or free metric structure, could the similarities be easily perceived? (Eugenio)
Reading question: Task

• Is asking the subjects if the pieces are "similar" a specific enough question? How would the test results differ if the subjects were asked which pieces have a similar form? (Rebekah)
Reading question: Emotional response

• By transposing a piece of tonal music to atonal music as in the experiment, how are emotions evoked in the listeners' minds different? (Mihir)
Reading question: Stimuli order

• In the text, the authors mention that the order of the pieces presented to the listeners was random for each individual. For those who heard a regular piece first followed by its atonal counterpart, could this help them in identifying new musical ideas if they recognize the rhythm (or other elements) from the previous piece they just heard? This might put certain participants at an advantage above others. Could this have unintentionally skewed the data? Should the order of the pieces have been kept static? (Julian C.)
Reading question: Atonal starting material?

• The researchers used two Beethoven sonatas and their atonal counterparts in the experiment. What would happen if they were to use excerpts from atonal piano pieces (such as Schoenberg's "Six Little Piano Pieces") and create tonal counterparts? Would the results look similar or different? Would the fact that these pieces have more unconventional rhythmic structures and dynamics cause the listener to perceive more of a similarity between the piece & its tonal counterpart? (Julian C.)
Reading question: No rhetoric?

• If a similar experiment were conducted in which the pieces used had no distinguishable musical rhetoric, how might the results differ? How do you think that the participants would attempt to segment the pieces, especially those which are atonal? Would this even be possible? In other words, does musical rhetoric become more important in the understanding of musical pieces as tonality decreases? (Tyer)