Applying Statistics to Classical Music

ISyE 2028 Project

James Wiggins and Nicole Redder

Purpose

During the Spring 2016 semester, we started by learning how to describe a data set both numerically and graphically. We then learned about modeling a data set with probability distributions and comparing numerical summaries of data sets with confidence intervals and hypothesis tests. Finally, we learned how to predict one variable from another. This project applies these major concepts to a topic we found interesting-- classical music.

Table of Contents

Module 1: Introduction and Data Collection 3
- Context Description
- Big Questions
- Define Variables
- Data Collection

Module 2: Modeling Data with Probability Distributions 9
- Book Ranking
- Number of Publications
- Musical Period
- Year of Birth
- Age at Death
- Country of Birth
- Genre

Module 3: Single sample analysis 22
- Standard Deviation of Life Expectancy
- Population Mean Birth Year
- Life Expectancy For Year 1800

Module 4: Two sample analysis 24
- Proportion of Composers Who Wrote Symphonies for Two Different Regions
- Variance in Popularity for Two Different Regions
- Mean Number of Publications for Two Different Regions

Module 5: Prediction 28
- Predicting Book Ranking from Number of Publications
- Predicting Age at Death from Birth Year

Module 6: Summary and Conclusion 36

Works Cited 38
Data Summary 39
Acknowledgements 40

Module 1: Introduction and Data Collection

What is the context of our project?

When people say "classical music," they don't really mean Classical music. Classical, with the capital "C," is actually just one period, one style, within the incredibly varied category of classical music; and it is ironically the shortest of the main periods. It is a bit misleading to call the genre "classical music," but that's how it's come to be known, and attempts to change that--"Western art music," "composed music," and the like--tend to confuse the issue by throwing out more names that for the most part go unused.

This misunderstanding is actually pretty reflective of how little people generally know about classical music. It's very easy to be dismissive of classical music, especially if you've only been exposed to it through the few pieces, mostly Baroque, that are common in popular culture, but there's much more to classical music than those pieces and composers.

Undisputed records of complete compositions exist as early as 100 AD, and musical tradition is rich and unique in any given place. Music is vastly different between countries, between regions within countries, even; and by no means is Western music, "classical music," or any other grouping, a sufficient summary of the incredible variety of form and convention that exists in the world. Within regions, different time periods have their own styles, and within time periods, individual artists have very distinct voices; therefore, even within a single region and a single period, it's inaccurate to consider only the most noted works or composers to be entirely representative.

This paper will focus on composers of Western classical music--music written formally in the European tradition; and any conclusions drawn will surround that population, with the understanding that they will not necessarily apply to other populations.

Western tradition uses a system in which full scales are made up of eight pitches, beginning and ending on the same note. Of the seven scale types--Ionian, Dorian, Phrygian, Lydian, Mixolydian, Aeolian, and Locrian--two, Ionian and Aeolian, better known as the major and minor scales, are in regular use. Intervals between any two pitches are made up of units called half-steps; within any given scale, adjacent notes are separated by one or two half-steps; and no further divisions of pitch are notated.

Western music is unique from most other traditions in that it is based on harmonies created by multiple tones sounding at once; and most of its interest lies in the choice and voicing of harmony. There are many conventions for both harmony and melody that can be followed or broken, and many schools of thought as to what choices should or should not be made; but composition is essentially the science of working with both halves--the melodic and the harmonic--to create a meaningful whole.

To make our dataset more manageable, we will be using only the fifty composers listed in Phil Goulding's *Classical Music: The 50 Greatest Composers and their 1,000 Greatest Works.* Goulding's book provides an ordered list of what he considers to be the best composers. While the choice of inclusion and order is a very subjective one, and there exists much disagreement about where any given composer is better than another, Goulding's book gives us a reasonably informed quantitative starting point. Having an ordered list for analysis allows us to compare composers and their characteristics by their rankings. Some composers may change order in any given person's ranking, but they'll likely be kept around the same place in the list, so although it's a subjective ranking, it's still a meaningful one in that general trends in the order are maintained. Overall, Goulding's list is a useful tool for analysis.

What "big" questions are we investigating?

- What are the demographics of famous composers?

In this paper we will consider Goulding's fifty composers as a representation of famous composers in general. We will consider the distributions of each of our variables and try to find any strong trends among them. Specifically, we will answer the following three questions: is the standard deviation of our composers age at death greater than the standard deviation of current life expectancy, and is the mean year of birth for our composers 1800, and is the mean age at death for our composers greater than the life expectancy for the year 1800? We will then use our composers' year of birth to predict their age at death. Finally, we will explore the relationship between our composers countries of origin with the genres they composed in by comparing the proportions of symphonic composers for different geographical regions.

- What contributes to a composer's popularity?

In order to quantitatively measure popularity, this paper will use number of publications to represent this concept. Within our dataset of fifty composers, we will split them into two regions; Northern and Western Europe will be one region, and Eastern will be the other. We will then answer the question "is the mean number of publications the same for each region?" In order to determine the correct hypothesis test necessary to answer this question, we will also need to determine whether or not the variances for number of publications are the same for each region. It will be interesting to see how much geography affects a composer's popularity.

- What contributes to a composer's skill?

Although Goulding's rankings is definitely a subjective variable, it is still very useful and this paper will use it to represent our composer's skill level. For this section we will see how well our composers' number of publications predicts their book ranking.

Variables

Book Ranking

Each composer's ranking in Goulding's book, from one to fifty, will be used to represent their compositional skill; although there are inherent flaws in using a single person's opinion of a composer to model their abilities, it is an extremely useful variable for analysis, and Goulding offers a well-informed and generally accepted ranking that can be considered accurate without incorporating too much error. Book ranking will be treated as a continuous variable, since there are fifty categories of only one composer each, with the understanding that it's not perfectly continuous, and the gap between any two adjacent composers, if it could be represented numerically, wouldn't necessarily be equal throughout the dataset.

Number of Publications

The number of hits on the Sheet Music Plus website will be used to model a composer's popularity; a composer could potentially have no hits, or any large number like 10,000, since different companies often publish the same works separately. The primary disadvantage of this model is that it doesn't give much weight to great works like symphonies that are usually only printed a few times, and may not be listed on Sheet Music Plus, since it's more of a source for solo or small ensemble works. However, the website does serve as an especially good model for the popularity of small works, taking into account publications that include portions of works and representing demand for them fairly well. In the absence of a better model for popularity, a fairly abstract but important concept, the number of publications is a reasonably accurate continuous variable to use.

Musical Period

Within classical music exist certain musical styles that gained popularity, lasted for some time, and then fell out of common use. These styles connected to time periods are called musical periods, and are, with brief descriptions, as follows:

Early Music, before 1450

We don't have much written record of music predating the Renaissance, so there isn't a good way to reconstruct it, but defining a period for it recognizes its existence and importance to the culture of its time, and research is still done on its characteristics.

Renaissance, 1450-1600

Much of Renaissance-era music was religious, and so is well-documented; music of this time was relatively simplistic, comprised of mostly melody with simple and strictly regimented harmonies.

Baroque, 1600-1775

Music by this time was used more extensively for entertainment, and more complex, though still very strictly controlled, harmony developed. Musical forms like the passacaglia, sonata, and fugue developed in this time.

Classical, 1775-1825

Harmony became a little simpler in this period, and forms were put to use and modified; symphonies were also developing during this time. Music started to expand into the concert setting as more of an art form.

Romantic, 1825-1900

In this time, harmony was stretched extensively, and many innovative composers started to break the strict harmonic rules that had existed in previous periods. Orchestras grew in size and instrumentation, and larger works became more popular.

Modern, after 1900

As structure was stretched, many new forms were developed; music of this time was split into many different styles, for example Impressionist and Serialist. In general, the ideas of writing in a single key and avoiding dissonance have been challenged, and in some cases traditional form has been thrown aside entirely.

It's important to note that musical periods don't have clearly defined beginnings and ends. Often, composers began to develop new styles while most composers were still writing in the previous style, and many composers lived during certain time periods but didn't compose in the popular style of the time. This paper uses Goulding's time boundaries for musical periods to maintain consistency with the list of composers, except in the case of expanding the Baroque period to extend to the beginning of Goulding's boundary for the Classical period. Musical period will be treated as an ordinal variable, as it cannot be represented numerically but has an intrinsic order.

Year of Birth

Year of birth is important because even though we have musical period for the composers, those categories don't exactly line up with the composers' lifetimes, and aren't exact enough to be used to model the composers' relative times very well. Since we're only working with Renaissance composers or later, the earliest reasonable year of birth would be around 1400. We'll allow its possible range to extend to the year Goulding wrote his book, 1992, since there isn't a clearly defined minimum age for a composer. Year of birth is a continuous variable.

Age at Death

Age at death is an important consideration, especially when comparing the number of works published from composers--for example, Hindemith, who died young, unsurprisingly wrote fewer pieces than most other composers. Age at death will be represented as a continuous variable ranging from 20 to 100, as conservative boundaries.

Country of Birth

We've chosen to use country of birth, and not country of primary residence, to model the composers' nationalities because it is clearer than country of primary residence, especially in the cases of composers who moved frequently. Nationality in many cases affects the type of music composers wrote, as there are different conventions in different areas, and could have other relationships with

composers' characteristics; it will be a categorical variable and could potentially be any European or North American country.

Genre

Five separate binary variables will be used to model the genres in which each composer wrote, and will be as follows:

- Piano Works
  This will be true if the composer wrote works for solo or accompanied piano; works for harpsichord and other keyboard instruments are not to be counted as piano music.

- Chamber Works
  This will be true if the composer wrote pieces for small ensembles (two to eight musicians); any instrumentation that fits the size requirement will be accepted, but arrangements of preexisting works for chamber ensemble will not.

- Symphonic Works
  This will be true if the composer wrote pieces for full symphony orchestra (at the very least, a combination of wind and string instruments). Symphonies that use a full orchestra count as symphonic works; string symphonies do not.

- Symphonies
  This will only be true if the composer wrote at least one full and completed symphony, either for string orchestra or for full orchestra.

- Opera
  This will be true if the composer wrote at least one complete opera. Comic operas, operettas, and similar works will not be counted towards this variable.

Data Collection

Goulding's book provided the fifty data points to be collected and the book ranking for each composer, while Sheet Music Plus provided the number of publications and Musicalics was used for the remaining variables. The required data was manually collected from the sources, and did not need cleaning. Every datapoint from Goulding's book had entries with all needed information in both the Sheet Music Plus and Musicalics databases, and no errors were found. The observational unit in this study is the individual composer.

Module 2: Modeling Data with Probability Distributions

Book Ranking

The Book Ranking variable is each composer's ranking in Goulding's book, from one to fifty, and can be used to represent his compositional skill. Since there are fifty categories of only one composer each, and because book ranking is a measurable quantity that has too many categories to be represented as an ordinal variable, we will treat is as continuous.

The range of possible values for each author's ranking is an integer from one to fifty so we will model this with the discrete uniform distribution. The parameters of the discrete uniform are alpha and beta, which will equal one and fifty respectively. This is a unique case in that book ranking only exists within our dataset, so we know the exact bounds and probability distribution.

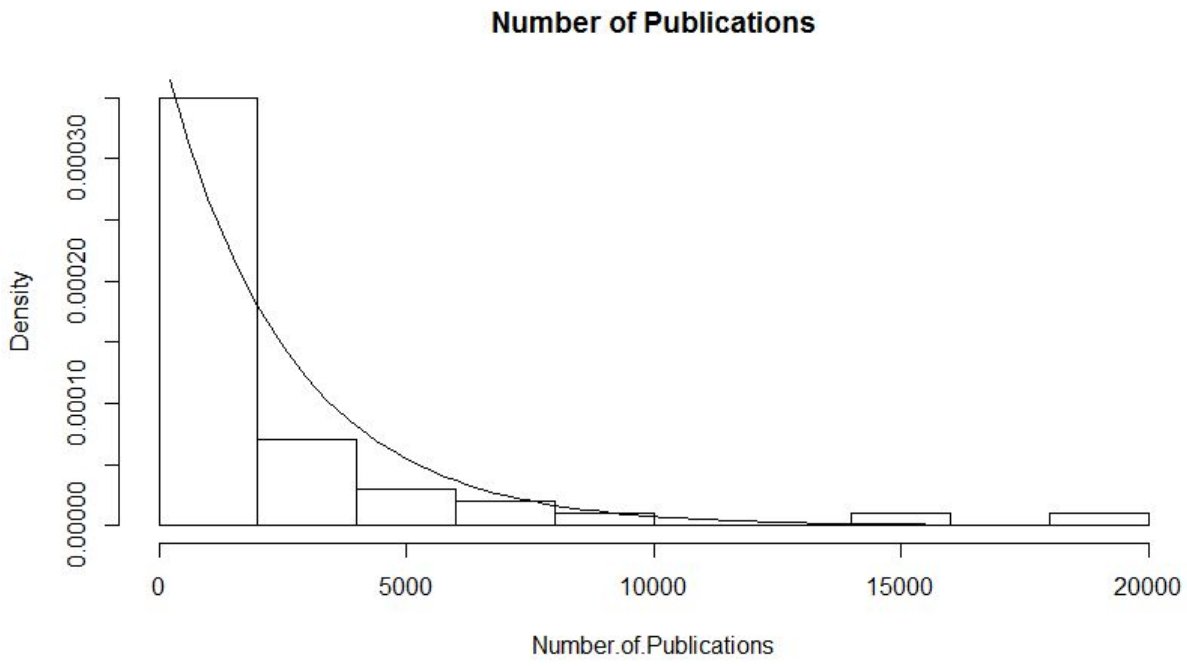
Number of Publications

The number of publications on the Sheet Music Plus website will be used to model a composer's popularity. This is a continuous variable because it is a measurable quantity and would require too many categories to be represented as an ordinal variable. The minimum number of publications for our composers is 50 and the maximum is 19520. The mean number of publications is 2512, but only half of our composers published more than 1144 and 75 percent of our composers published less than 2995 of their works.

Below is a histogram representing the probabilities for each range of number of publications. The graph shows that the probability of a composer having less than about 2000 publications is very high and the probability continues to drop off for consecutively higher numbers of publications. With the graph of an exponential laid on top of the histogram, it is easy to see that the data is modeled well with that distribution family. To estimate the parameters we used the method of moments. The sample mean of our data is 2512 and the expected value of the exponential distribution is $1/\lambda$. Setting the first moment of the data to the exponential distributions first moment and solving for lambda results in $\lambda = 1/2512 = 0.0003980892$.

R Code:

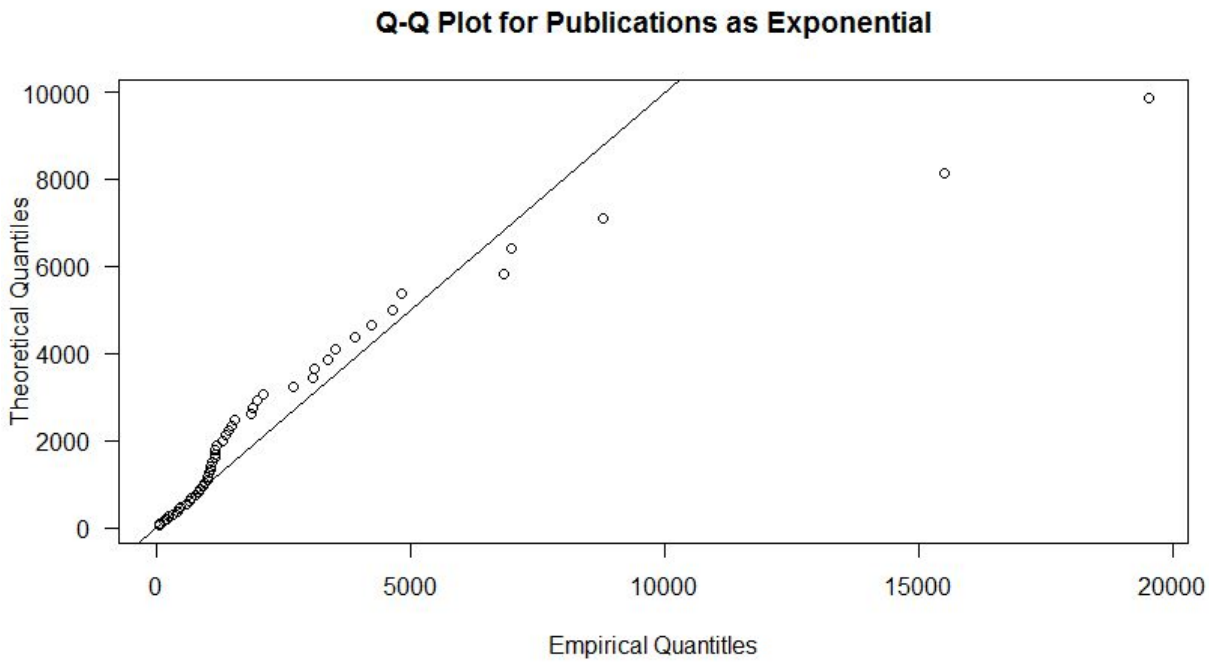
```
#graph number of publications and overlay our model
hist(Number.of.Publications,main="Number of Publications",freq=FALSE)
curve(dexp(x,rate=(1/mean(Number.of.Publications))),add=TRUE,las=1)
```



The Q-Q plot illustrates that the model is a good fit for the exponential distribution family and our calculated value for lambda. Although the dots do not perfectly adhere to the y=x line, most of them are in the bottom left corner of the graph and in the middle the points are relatively close to the identity line. Both of these observations provide enough reason for us to be satisfied with this particular model of our data.

R Code:

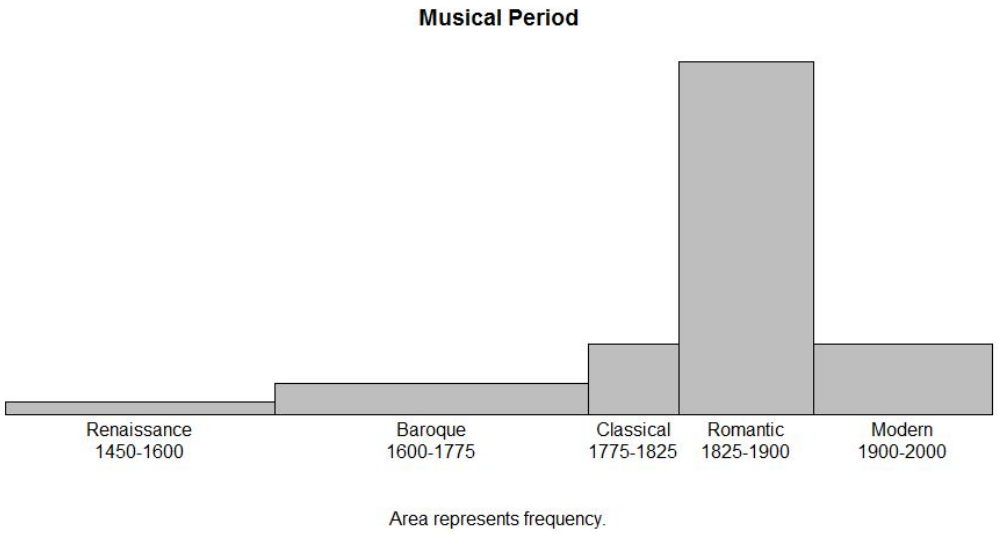
```
#create a Q-Q plot for number of publications using our model
n = length(Number.of.Publications)
probs = (1:n)/(n+1)
exp.quants = qexp(probs,(1/mean(Number.of.Publications)))
plot(sort(Number.of.Publications),sort(exp.quants),ylab="Theoretical Quantiles",
   xlab="Empirical Quantitles", main="Q-Q Plot for Publications as Exponential",las=1)
abline(0,1)
```



Musical Period

R Code:

```
#make a barplot for musical period where the bars correspond to the musical periods' lengths
#and area represents the frequency among our sampled composers
periods=table(Musical.Period)
periods=periods[c("Renaissance","Baroque","Classical","Romantic","Modern")]
len <- c(150,175,50,75,100)
periodnames <- c(
        "Renaissance\n1450-1600",
        "Baroque\n1600-1775",
        "Classical\n1775-1825",
        "Romantic\n1825-1900",
        "Modern\n1900-2000")
barplot(periods/len,width = len,las=1,names.arg=periodnames,space=0,
    main = "Musical Period",yaxt="n",sub="Area represents frequency.")
```



We can see from the barplot above that, in terms of composers compared to length of the period, the Romantic period far outnumbers all others. Renaissance, conversely, has an extremely low representation for its length; and no Early Music composers were in our sample, as there is little written record for it, so that period is not included in our analysis.

R Code:

```
#get numbers for table
table(Musical.Period)
```

Period	Renaissance	Baroque	Classical	Romantic	Modern
Composers in Sample	2	6	4	30	8

Since an equation model would not make sense for this variable, our model for it will be a piecewise probability function, where the probability of a given period equals its probability in our sample (its frequency in our sample divided by fifty, the size of our sample).

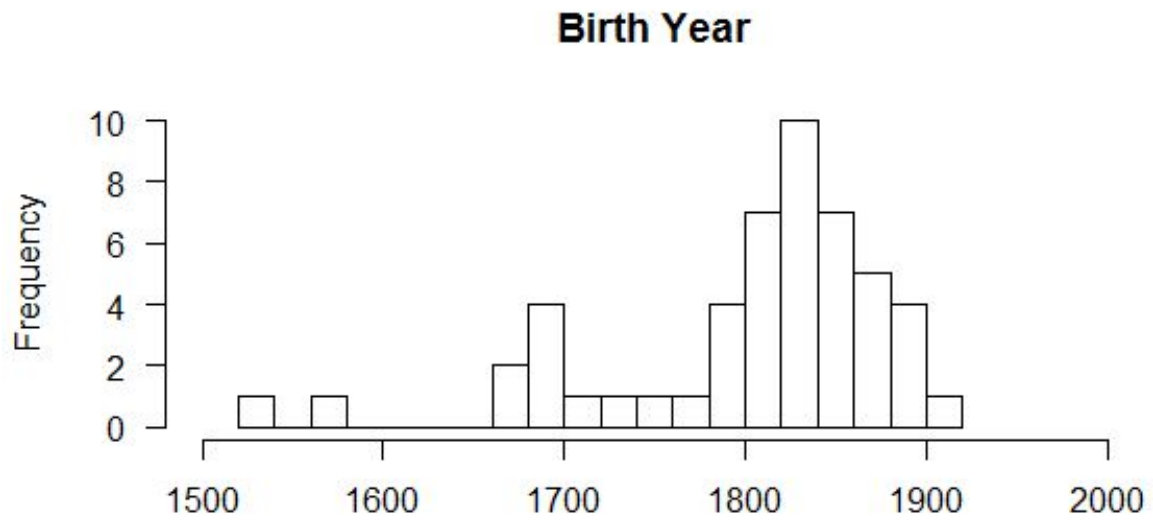
Period	Probability
Renaissance	.04
Baroque	.12
Classical	.08
Romantic	.6
Modern	.16

Year of Birth

The composers' average year of birth lies in the Romantic period (1800.92), but their birth years have a high standard deviation (81.48 years), so many composers lie far from that mean. The mean birth year is also significantly lower than the median birth year (1824), suggesting a left skew. For our dataset, the minimum birth year is 1525, and the maximum is 1906.

R Code:

```
#create a histogram for birth year; adjust breaks so that the pattern is clearer
hist(Birth.Year, main = "Birth Year",breaks=15,xlim=c(1500,2000),las=1, xlab="")
```

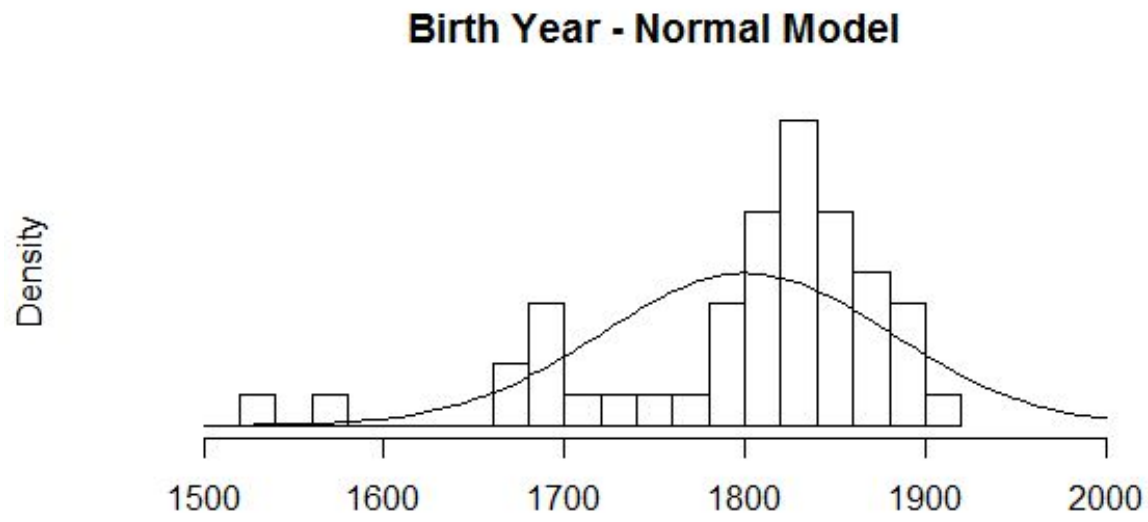


From this histogram we can see that the data doesn't clearly fit any distribution we've learned so far. Goulding's composers include two very early composers, but of course no very late composers,

since there is a hard cutoff point two decades or so before 1995, the year Goulding wrote his book. No composers born since then are likely to have written major works recognized by Goulding. We have a very left-skewed curve with a slight bump just before the year 1770 and a cutoff around 1975. The closest model we have learned for such a curve is the normal distribution, although it doesn't match the hard cutoff. MLE estimation (although it results in a biased estimator, our dataset is large enough that the effect is limited, and the chance to practice MLE on a real sample is worth the potential bias) gives us the parameters μ = 1800.92 and θ = 79.85447.

R Code:

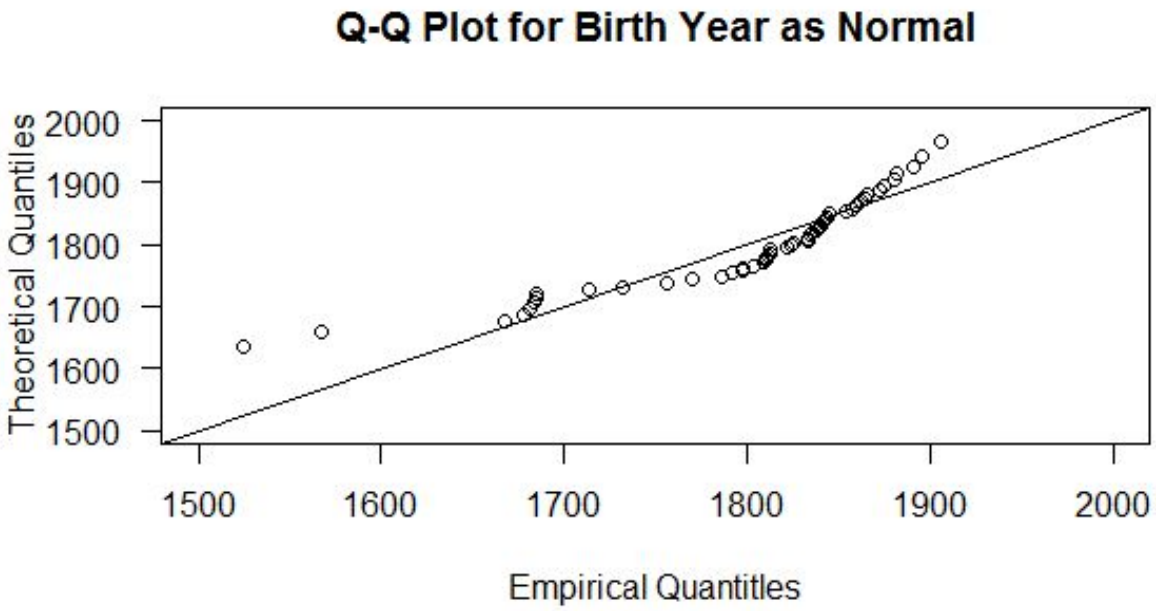
```
#overlay the normal model on a new histogram
n = length(Birth.Year)
m=mean(Birth.Year)
s=sd(Birth.Year)*((n-1)/n)
hist(Birth.Year, main = "Birth Year - Normal Model",
   breaks=15,xlim=c(1500,2000),las=1, xlab="", freq=FALSE,yaxt="n")
curve(dnorm(x,m,s),add=TRUE)
```



From the plotted curve and histogram alone, we can see that, as predicted, this is not an ideal model. The normal curve predicts a high probability density in the 1700-1800 range, when we can clearly see that our sample has a low probability density there; and it severely underestimates probability density in the 1800-1900 range.

R Code:

```
#construct a Q-Q plot between sampled birth years and our normal model
probs = (1:n)/(n+1)
norm.quants = qnorm(probs,m,s)
plot(sort(Birth.Year),sort(norm.quants),
   xlim=c(1500,2000),ylim=c(1500,2000),
   ylab="Theoretical Quantiles",
   xlab="Empirical Quantitles",
   main="Q-Q Plot for Birth Year as Normal",las=1)
abline(0,1)
```



The Q-Q Plot of our sample composers' birth years against our normal model confirms that our model is a poor fit to the data. The quantiles rarely lie on the identity line, and the plot simply proves that our data is not normal.

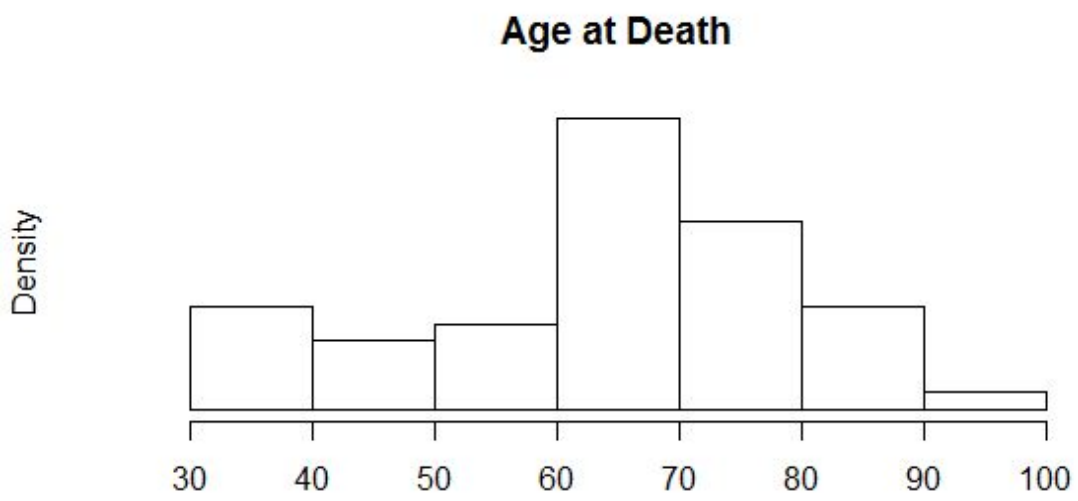
Even though the normal is a poor fit for our birth year data, it's the best distribution we've learned so far for this dataset. It fits the hump at the center and the lower probability at the outer bounds of our birth year data (although the normal model does not model its skew).

Age at Death

The average age at death among our composers is 64.52 years, with a moderate standard deviation (15.32 years). Lifespans range from 31 to 91.

R Code:

```
#construct a histogram for age at death
hist(Age.at.Death,freq=FALSE,main="Age at Death",
   las=1, xlab="",yaxt="n")
```



There is a slightly higher probability density at the left end of our histogram, but overall this visually fits the normal distribution--a single peak near the middle, and a symmetrical distribution. MLE (used for the same reasons as with birth year) estimation results in parameters $\mu = 64.52$ and $\theta = 15.00972$.

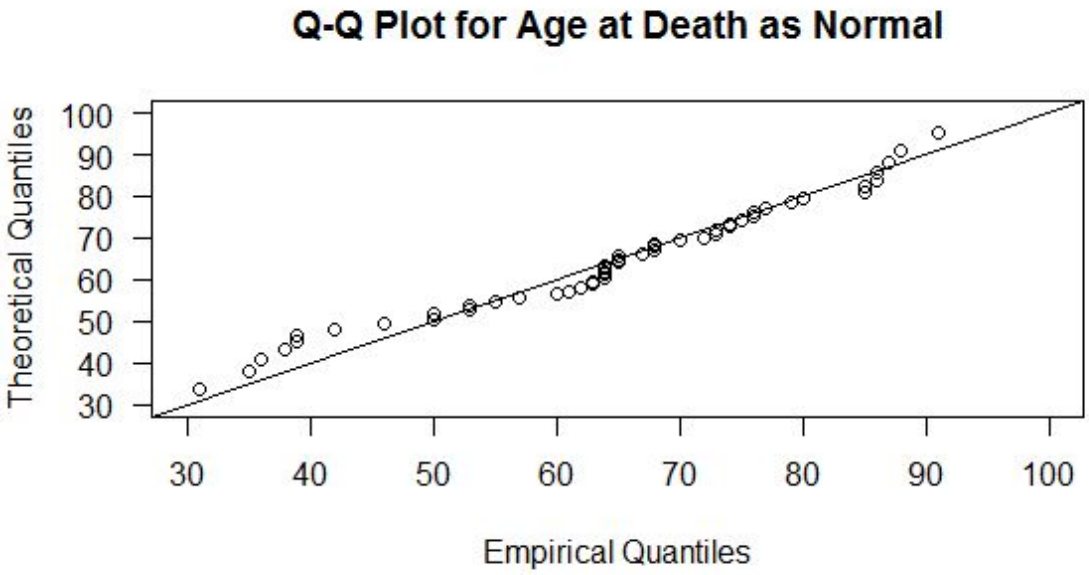
R Code:

```
#overlay our normal model on a new histogram
n = length(Age.at.Death)
m=mean(Age.at.Death)
s=sd(Age.at.Death)*((n-1)/n)
hist(Age.at.Death,freq=FALSE,main="Age at Death - Normal Model",
   las=1, xlab="",yaxt="n")
curve(dnorm(x,m,s),add=TRUE)
```

This model appears to be a very good fit for the right side of our dataset, but underestimates our data in the 40-60 range, and appears to overestimate in the 30-40 and 60-70 ranges. Overall, it appears a good fit for the composers' ages at death.

R Code:

```
#construct a Q-Q plot using our sampled values and our normal model
probs = (1:n)/(n+1)
norm.quants = qnorm(probs,m,s)
plot(sort(Age.at.Death),sort(norm.quants),
   xlim=c(30,100),
   ylim=c(30,100),
   ylab="Theoretical Quantiles",
   xlab="Empirical Quantiles", main="Q-Q Plot for Age at Death as Normal",las=1)
abline(0,1)
```

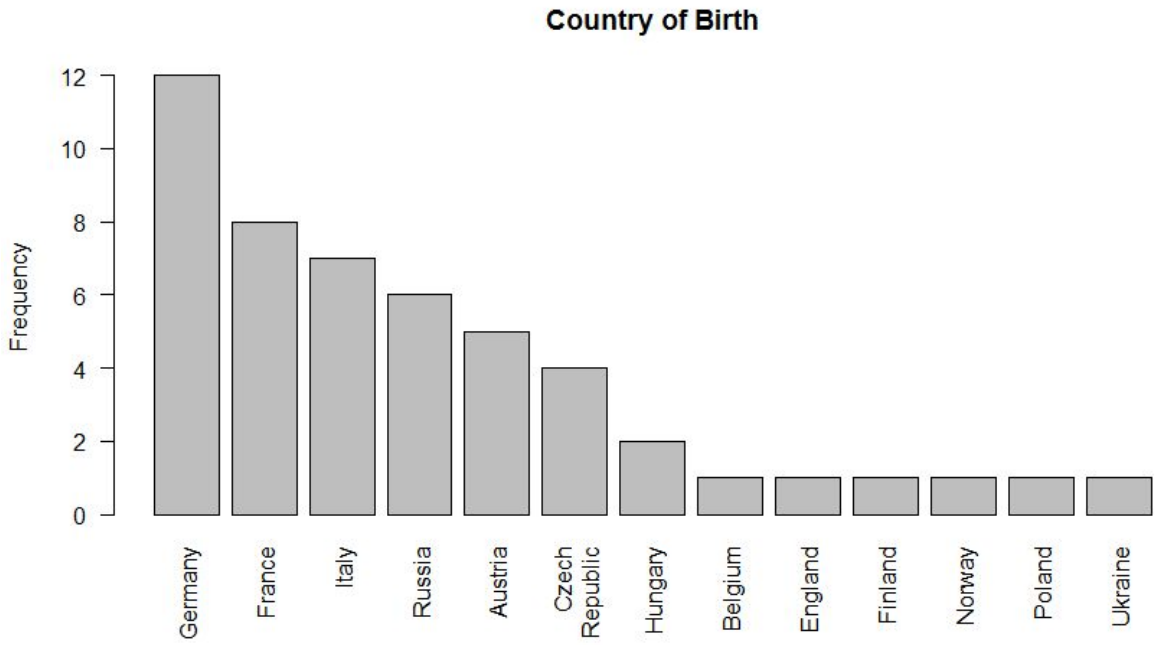


Our normal model for age at death does in fact overestimate our data on the far left side, and we can see that it also slightly overestimates our data on the right side as well; however, it fits the majority of our data in the middle range very well.

Country of Birth

R Code:

```
#construct a barplot ordered by frequency for country of birth labeled by country
countrynames = c("Germany",
                 "France",
                 "Italy",
                 "Russia",
                 "Austria",
                 "Czech\nRepublic",
                 "Hungary",
                 "Belgium",
                 "England",
                 "Finland",
                 "Norway",
                 "Poland",
                 "Ukraine")
barplot(sort(table(Country.of.Birth),decreasing=T),las=2,
    names.arg=countrynames,main="Country of Birth",ylab="Frequency")
```



Because country of birth is a categorical variable, our only model for it will be to use the probabilities for each country within our dataset, which is also our numerical summary of the data.

R Code:

```
#gather values for table
table(Country.of.Birth)
```

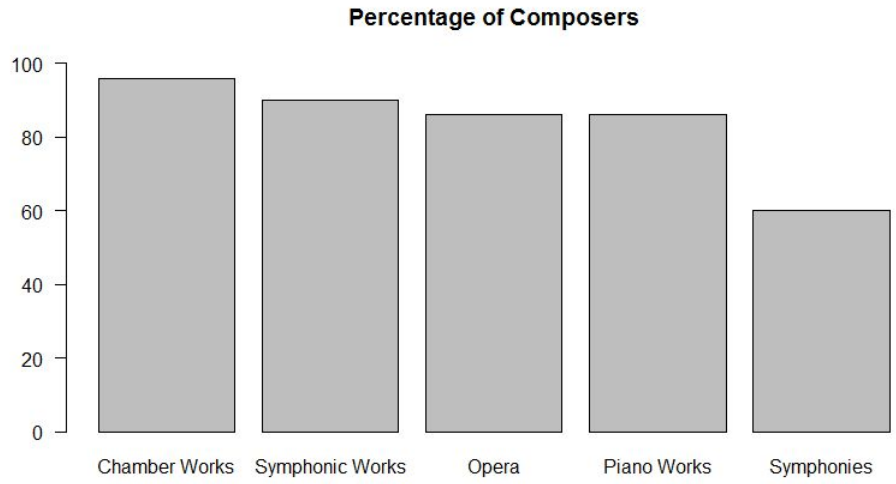
Country	Composers	Probability
Germany	12	.24
France	8	.16
Italy	7	.14
Russia	6	.12
Austria	5	.1
Czech Republic	4	.08
Hungary	2	.04
Other	6	.12

Far more composers in our sample are from Germany than from any other country. Six countries--Belgium, England, Finland, Norway, Poland, and Ukraine--each are represented only once in Goulding's top fifty, and no country outside of Europe is included.

Genre

R Code:

```
#construct a side-by-side barplot for genres
barplot(c(100*mean(Chamber.Pieces),100*mean(Symphonic.Pieces),100*mean(Opera),
    100*mean(Piano.Pieces),100*mean(Symphonies)),
    beside=TRUE,names.arg=c("Chamber Works","Symphonic Works","Opera","Piano
Works","Symphonies"),ylim=c(0,100),
    main="Percentage of Composers",las=1)
```



All musical genres studied are used by a majority of sampled composers; by far the least written category is the symphony, at 60% of composers, while nearly all composers wrote chamber works (96%). Each genre variable will be modeled as a Bernoulli variable, with MLE estimator p = sample mean.

R Code:

```
#grab values individually for table
mean(Chamber.Works)
mean(Symphonic.Works)
mean(Opera)
mean(Piano.Works)
mean(Symphonies)
```

Genre	*Chamber Works*	*Symphonic Works*	*Opera*	*Piano Works*	*Symphonies*
Probability	*.96*	*.9*	*.86*	*.86*	*.6*

Module 3: Single sample analysis

Is the standard deviation of classical composers' ages at death greater than current standard deviation of life expectancy (15 years)?

The parameter of interest here is the standard deviation in all classical composers' ages at death, σ.

Ho: $\sigma \leq 15$ years.

Ha: $\sigma > 15$ years.

Our sample size is 50 composers, and our sample standard deviation is 15.31604 years. Using a chi-squared test, we arrive at a p-value of 0.39165; therefore, there is not sufficient evidence to conclude that standard deviation of age at death among classical composers is greater than 15 years.

Test done with calculator.

Is the population mean birth year 1800?

The parameter of interest is the mean birth year among all classical composers, μ. Our sample size is 50, our sample standard deviation is 15.31604 years, and our sample mean is the year 1800.92. Our acceptable bounds for this test are 1780-1820 as we're interested in life expectancy within that time, and within the 1770-1830 interval life expectancy does not vary significantly.

Ho : $\mu \leq 1780$

Ha : $\mu > 1780$

Using a t-test with 49 degrees of freedom, we find a p-value of 1.3243×10^{-12}. We have sufficient evidence to conclude that the population mean year of birth is greater than 1780.

Ho : $\mu \geq 1820$

Ha : $\mu < 1820$

Using a t-test with 49 degrees of freedom, we find a p-value of 1.3243×10^{-12}. We have sufficient evidence to conclude that the population mean year of birth is less than 1820.

Hypothesis test done with calculator. Overall, since we have shown that the mean birth year is greater than 1780 and less than 1820, we can conclude that the population mean year of birth is 1800.

Is the population mean age at death greater than the life expectancy during the population mean year of birth (1800)?

In 1800, the mean age at death was 46.67, conditioned on a person having reached the age of 10. The average life expectancy among our composers may be reasonably modeled by the life expectancy at the average birth year among our composers because life expectancy appears to be roughly linear.

The parameter of interest is the mean life expectancy among all classical composers, μ.

Ho: $\mu \leq 46.67$

Ha: $\mu > 46.67$

Using a t-test with 49 degrees of freedom and estimating the population standard deviation with our sample standard deviation of 15.31604 years, we find a p-value of 5.481×10^{-11}. We have sufficient evidence to conclude that the life expectancy of classical composers is greater than the overall life expectancy among people in 1800 who had reached the age of 10--that is, who had not died as children.

These results are interesting in that they contrast with the stereotype that artists have unusually short and painful lives. While we do have a slight correlation/causation issue here--perhaps our sampled composers became famous because they lived longer than average, or perhaps some other unseen factor is at work here--we can see that classical composers must in fact live reasonably well, to live longer than the average person. If it is a causation relationship, then perhaps classical composers live longer than the average person due to the creative and expressive nature of composition, and composition has health benefits.

Test done with calculator.

Module 4: Two sample analysis

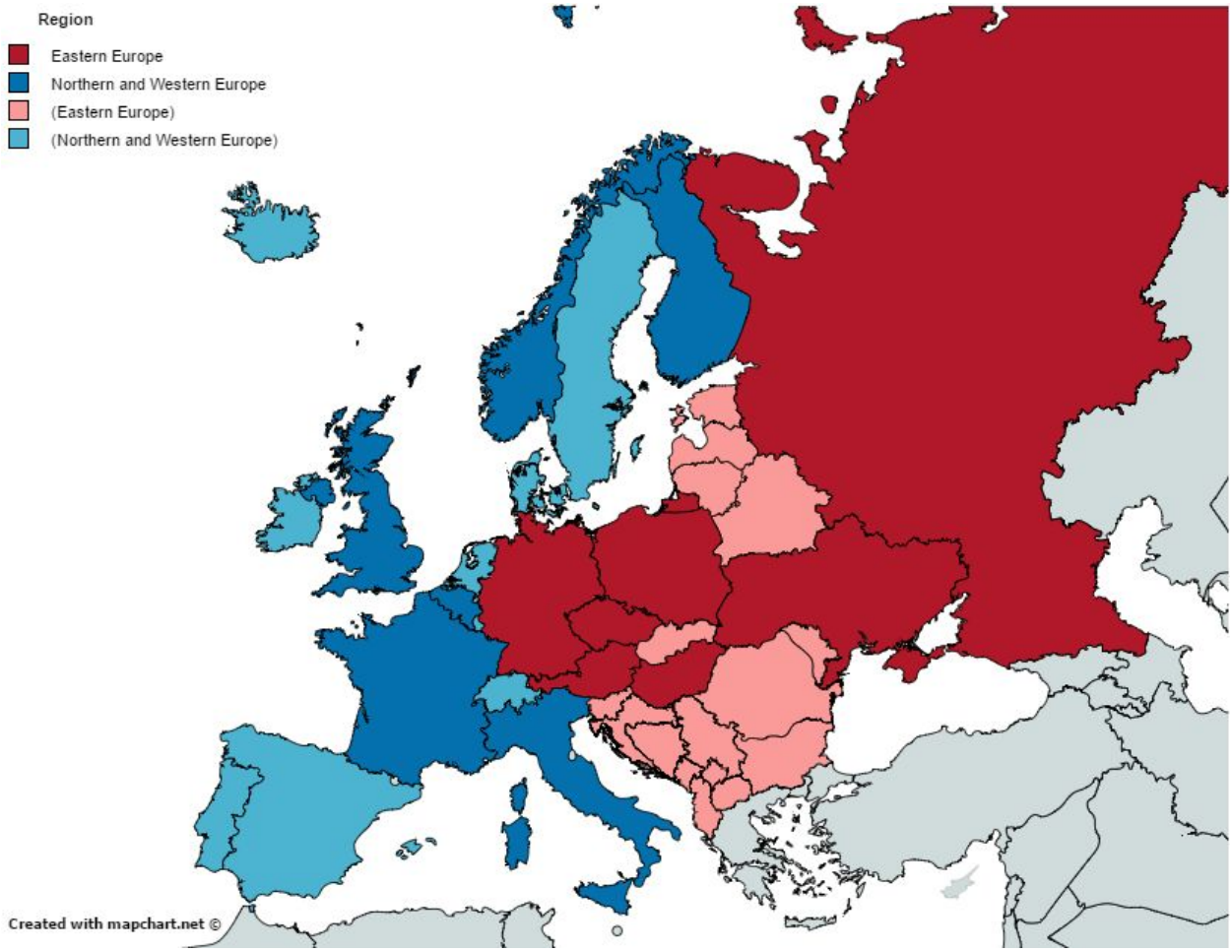
Module 4: Two sample analysis In this module, you will demonstrate the analyses you have learned that involve two variables. You will tell a story (or stories) by posing three questions, doing the analysis, and interpreting the results.

Is there a difference in the proportion of composers who wrote symphonies between composers from two different regions?

We will define our regions as follows, where parentheses indicate that we have no composers from those countries represented in our sample:

Northern and Western Europe: United Kingdom, Norway, Finland, France, Belgium, Italy, (Iceland, Ireland, Netherlands, Switzerland, Denmark, Sweden, Portugal, Spain)

Eastern Europe: Germany, Poland, Czech Republic, Austria, Hungary, Ukraine, Russia, (Estonia, Latvia, Lithuania, Belarus, Slovakia, Moldova, Romania, Bulgaria, Serbia, Slovenia, Croatia, Bosnia and Herzegovina, Montenegro, Kosovo, Albania)



Within our data set, 38 composers were born in our Eastern European countries, and 12 in our Western and Northern European countries. Among Eastern European composers, 64.51613% wrote symphonies; while among Western and Northern European composers, 52.63158% wrote symphonies.

```
#split composers by nationality
e = c("Germany", "Poland", "Czech Republic", "Austria", "Hungary", "Ukraine", "Russia")
wn = c("England", "Norway", "Finland", "France", "Belgium", "Italy")
ecomposers = subset(composers, Country.of.Birth %in% e)
wncomposers = subset(composers, Country.of.Birth %in% wn)
```

H_0: The percentage of composers born in Eastern European countries who wrote symphonies is equal to the percentage of composers who were born in Western and Northern European countries who wrote symphonies.

H_A: The percentage of composers born in Eastern European countries who wrote symphonies is not equal to the percentage of composers who were born in Western and Northern European countries who wrote symphonies.

Using a Z-test (done with a calculator) we calculate a p-value of .008829, and therefore reject H_0 in favor of H_A: we conclude that there is a difference between the population of Eastern European composers and the population of Northern and Western European composers in the proportion who wrote symphonies.

This makes a lot of sense, as many of the writers of the most well-known symphonies--Beethoven, Tchaikovsky, Shostakovich, Mahler, Dvořák, Mozart, Schubert, Haydn, Brahms, Mendelssohn--were, in fact, from our Eastern European countries. It makes sense that the countries that produced a lot of the most famous symphonies should have a higher proportion of symphony composers.

Is the variance in the number of publications produced the same between composers from Eastern Europe and composers from Northern and Western Europe?

Using the same definitions for the two regions as in the previous analysis, the Eastern European composers have a sample variance of 19993289 publications squared, while the Western and Northern European composers have a sample variance of 1248205 publications squared.

H_0: The variance in number of publications among composers born in Eastern European countries is equal to the variance in number of publications among composers who were born in Western and Northern European countries.
H_A: The variance in number of publications among composers born in Eastern European countries is not equal to the variation in number of publications among composers who were born in Western and Northern European countries.

Using an F-test (done with a calculator rather than in R) with 37 degrees of freedom in the numerator and 11 in the denominator, we calculate a p-value of 1.239×10^{-5}. Therefore, the evidence supports rejecting H_0 in favor of H_A: that the variance in number of publications among the population of Eastern European composers is not equal to that among the population of Western and Northern European composer.

Is the mean number of publications greater for Eastern European than for Western/Northern European composers?

Among Eastern European composers, we have a sample mean of 3094.323 publications, and among Western and Northern European composers, we have a sample mean of 1561.895 publications.

H_0: The number of publications among composers born in Eastern European countries is equal to the number of publications among composers who were born in Western and Northern European countries.
H_A: The number of publications among composers born in Eastern European countries is not equal to the number of publications among composers who were born in Western and Northern European countries.

Since sample variances have been proven unequal, we need to run a t-test using the average variance.

However, we calculate zero degrees of freedom for the test, and we cannot run the test with zero degrees of freedom; therefore, this test is not useful for our data.

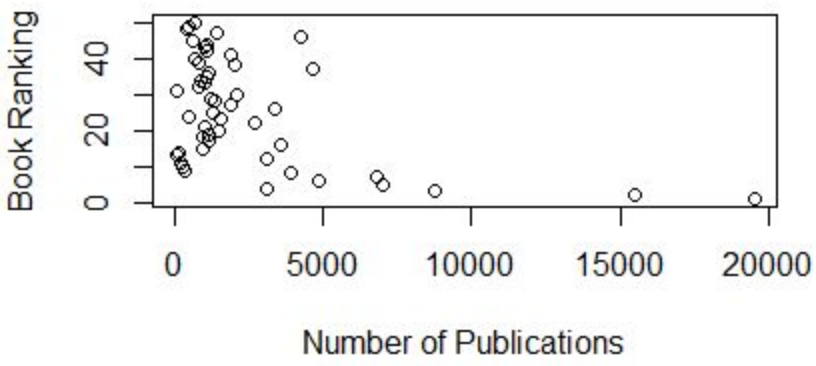
As we have learned no other test on the mean of two populations with unknown population variance and different sample variances, we cannot draw a conclusion comparing the mean number of publications between Eastern European and Western/Northern European composers.

Module 5: Prediction

How does a composers' number of publications predict his book ranking?

R Code:

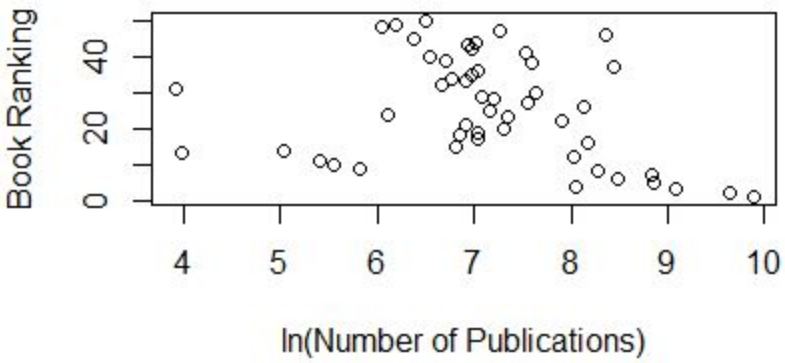
```
#original graph:
plot(Book.Ranking~Number.of.Publications,xlab="Number of Publications",
   ylab="Book Ranking")
```



From this graph we can see that there is a relationship between a composer's number of publications available and his ranking in Goulding's book; however, it doesn't appear to be very linear. Taking the natural log of the number of publications produces a linearized graph.

R Code:

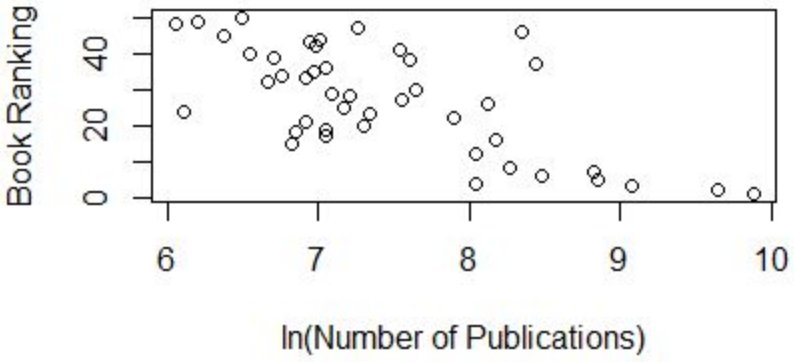
```
#linearized:
plot(Book.Ranking~log(Number.of.Publications),xlab="ln(Number of Publications)",
   ylab="Book Ranking")
```



However, we can see that there are some points that fall far to the left; in a linear regression these would have a lot of significance, but may not be well modeled by the resulting line. We've chosen to remove all composers with fewer than 350 publications (a total of 6 points) from our dataset for this analysis, to give us a cleaner result.

R Code:

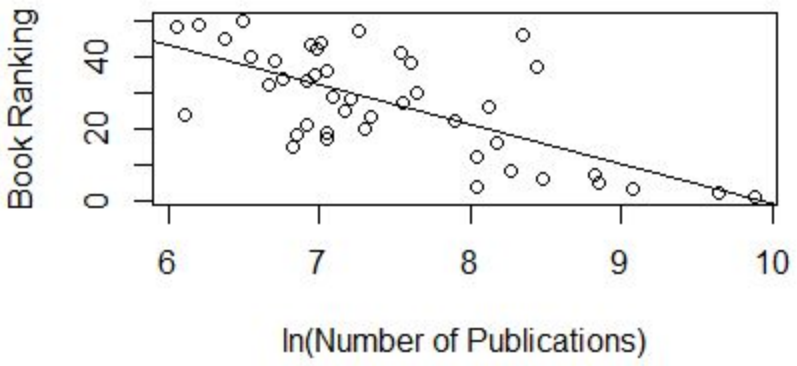
```
#cleaned:
cleanedBook.Ranking=Book.Ranking[Number.of.Publications>350]
cleanedNumber.of.Publications=Number.of.Publications[Number.of.Publications>350]
plot(cleanedBook.Ranking~log(cleanedNumber.of.Publications),xlab="ln(Number of
Publications)",
   ylab="Book Ranking")
```



This dataset shows a clear, roughly linear relationship between our predictor and response variables. Linear regression on this dataset results in the following line.

R Code:

```
#regression:
plot(cleanedBook.Ranking~log(cleanedNumber.of.Publications),xlab="ln(Number of
Publications)",
   ylab="Book Ranking")
lmodel=lm((cleanedBook.Ranking)~log(cleanedNumber.of.Publications))
abline(lmodel)
```



R Code:

```
#get model coefficients and tests of fit
summary(lmodel)
```

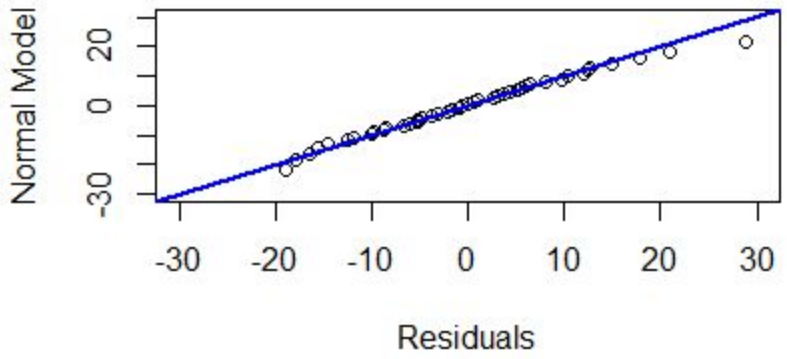
We find that a composer's book ranking can be modeled as

109.017 - 10.998*ln(Number of Publications)

upwards of 350 publications. This model has an adjusted R-squared of .4528, and a p-value of 3.39e-07--it's an exceedingly good model that explains much of the fluctuation of book ranking.

R Code:

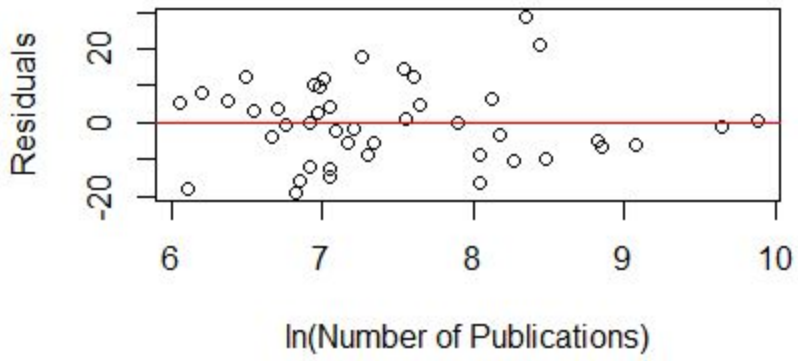
```
#get residuals
resids = lmodel$residuals
#qq plot of residuals
n=length(resids)
probs = (1:n)/(n+1)
norm.quants = qnorm(probs, mean=0, sd=sd(resids))
plot(sort(resids),sort(norm.quants), xlab="Residuals",ylab="Normal Model",
   xlim=c(-30,30),ylim=c(-30,30))
abline(0,1, col="blue",lwd=2)
```



We can see from the Q-Q plot that our residuals fit very well with a normal model; no significant deviations are seen.

R Code:

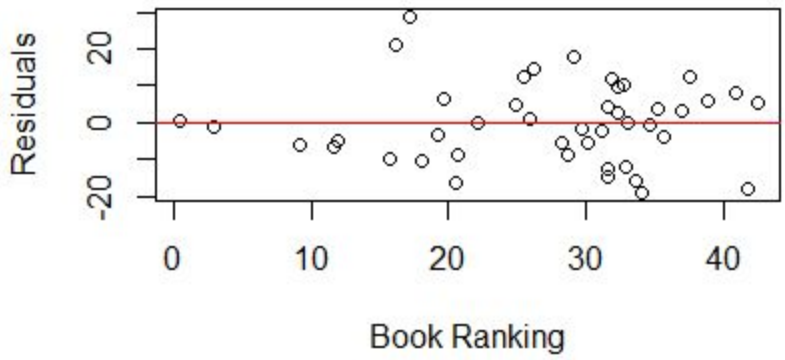
```
#plot residuals vs explanatory values to see if there is any pattern remaining
plot(log(cleanedNumber.of.Publications),resids,xlab="ln(Number of Publications)",
   ylab="Residuals")
abline(0,0, col="red")
```



Plotting our residuals against the explanatory variable number of publications, we see no obvious pattern that would suggest an issue with the model.

R Code:

```
# plot residuals vs predicted values
plot(lmodel$fitted.values,resids,
   xlab="Book Ranking",ylab="Residuals")
abline(0,0, col="red")
```

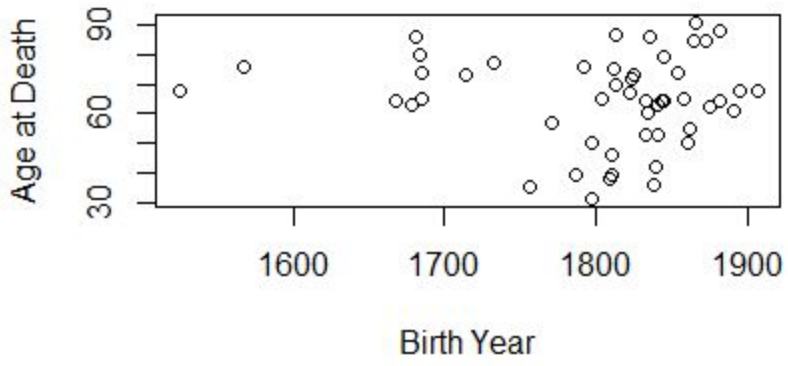


Plotting residuals by our predicted book ranking values, we again see no serious issue; we do have more negative residuals than positive at book rankings below 10, but overall it seems to be a reasonable model.

How does a composer's birth year model his age at death?

R Code:

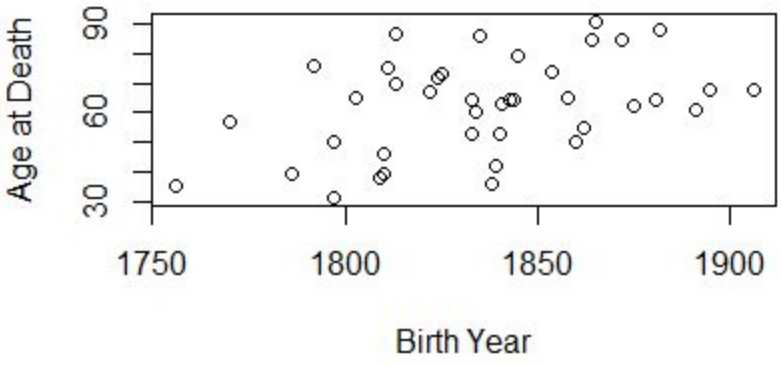
```
#plot of age at death modeled by birth year
plot(Age.at.Death~Birth.Year,xlab="Birth Year",ylab="Age at Death")
```



From the graph we can see that birth years below about 1750 have a very odd relationship to ages at death, possibly due in part to how few points we have for that group. We will therefore remove those points for a cleaner and more linear relationship.

R Code:

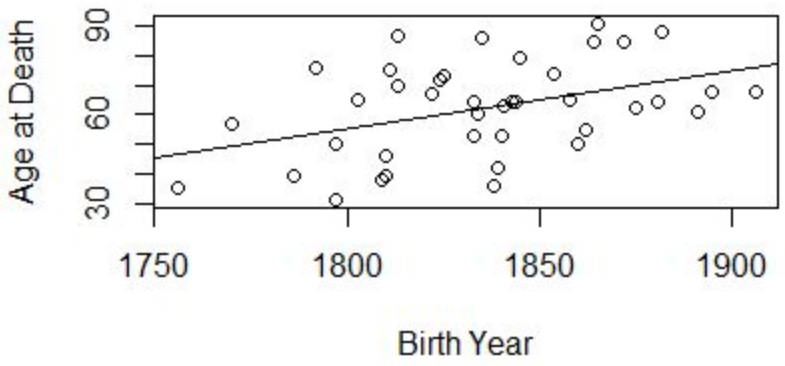
```
#plot cleaned data
cleanedAge.at.Death=Age.at.Death[Birth.Year>1750]
cleanedBirth.Year=Birth.Year[Birth.Year>1750]
plot(cleanedAge.at.Death~cleanedBirth.Year,xlab="Birth Year",
   ylab="Age at Death")
```



We now appear to have a positive linear relationship between birth year and age at death. We find the following model using simple linear regression.

R Code:

```
#linear mode
lmodel2 = lm(cleanedAge.at.Death~cleanedBirth.Year)
abline(lmodel2)
```



R Code:

```
#get model coefficients and tests of fit
summary(lmodel2)
```

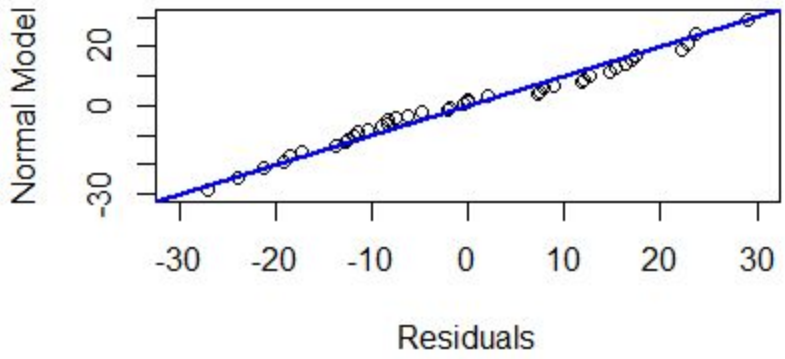
We find that a composer's age at death can be modeled as

$$-299.56169 + 0.19723 * \text{Birth Year}$$

for birth years between 1750 and 2000. This model has an adjusted R-squared of 0.1538, and a p-value of 0.007133--it's not as exceedingly good as our model for book ranking, but it's still a very good model at $\alpha = 0.05$.

R Code:

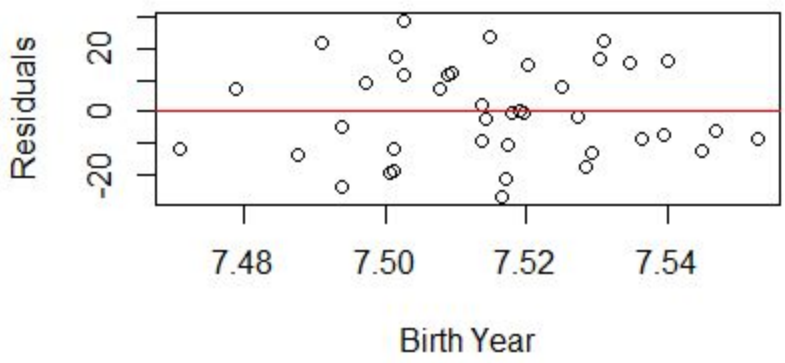
```
#get residuals
resids = lmodel2$residuals
#qq plot of residuals
n=length(resids)
probs = (1:n)/(n+1)
norm.quants = qnorm(probs, mean=0, sd=sd(resids))
plot(sort(resids),sort(norm.quants), xlab="Residuals",ylab="Normal Model",
   xlim=c(-30,30),ylim=c(-30,30))
abline(0,1, col="blue",lwd=2)
```



Our residuals aren't perfectly normal, as shown by the Q-Q plot above, and we might be concerned about this with a very precise model, but for our purposes, they're acceptably close to the normal.

R Code:

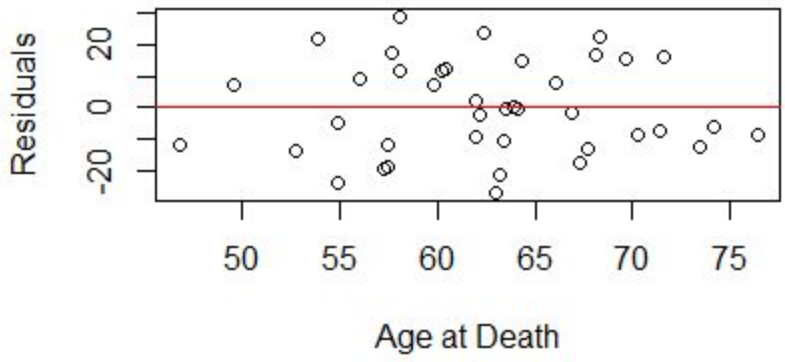
```
#plot residuals vs explanatory values to see if there is any pattern remaining
plot(log(cleanedBirth.Year),resids,xlab="Birth Year",
   ylab="Residuals")
abline(0,0, col="red")
```



Plotting our residuals by our explanatory variable birth year, we see no pattern indicating our model is insufficient; we appear to have a random distribution even above and below zero.

R Code:

```
# plot residuals vs predicted values
plot(lmodel2$fitted.values,resids,
   xlab="Age at Death",ylab="Residuals")
abline(0,0, col="red")
```



Plotting our residuals against our predicted ages at death, we again see nothing concerning; we have what appears to be a random spread equal above and below the y = 0 line.

Module 6: Summary and Conclusion

What are the demographics of famous composers?

We found that our sample of classical composers was primarily from countries such as Germany, Italy, France, Russia, and Austria; that it had a mean age at death of about 64 years, and that the average number of publications produced in our sample was 2512 publications.

We proved at an alpha value of 0.05 that the population average age at death among classical composers is higher than average life expectancy, conditioning on survival past childhood.

We also created a useful linear model for a composer's age at death based on his birth year, for the range of birth years between 1750 and 2000.

We showed that the composers born in Eastern European countries were more likely to write symphonies than those born in Western and Northern European countries, with regions defined earlier in the paper.

These findings gave us a slightly better understanding for the overall characteristics of the population of classical composers, and raises questions such as whether composing music does cause a longer life expectancy, that exploring later might help us to further understand the population of interest.

What contributes to a composer's popularity?

We modeled popularity by the number of publications available on Sheet Music Plus.

While our sample mean number of publications for composers born in Eastern Europe was much higher than our sample mean number of publications for composers born in Western and Northern Europe, our data and the methods we have learned so far were insufficient to prove that this is true for the population of classical composers as a whole; therefore, the question remains as to whether region of birth influences popularity.

We did, however, contribute to answering this question, by showing at an alpha value of 0.05 that the variance in number of publications among composers born in Eastern Europe is not the same as the variance in number of publications among composers born in Western and Northern Europe; and that informs us on which two-sample tests may be used, so that perhaps with a larger data set or new knowledge, we might be able to answer this question, and contribute to the general understanding of factors to a composer's popularity.

What contributes to a composer's skill?

We modeled skill using Goulding's rankings for the composers.

We found that skill could be modeled very well as having an exponential relationship to number of publications available; we developed an extremely good model for Goulding's rankings by using a linear regression with the natural log of number of available publications.

If Goulding's rankings do model skill well, as we reasonably expect, then we can predict a composer's skill by the number of publications we can find online.

That's a very interesting conclusion in that it suggests that skill and popularity generally occur together, and that the best composers generally aren't ignored. Of course there are limitations; most composers we studied have long since died, and their compositions have been available for quite some time for analysis and discussion. Our model may not work very well for living composers, and doesn't cover composers born early in the baroque period or before, but it's extraordinarily strong for the time periods it does cover.

With this question we had, we may have found an answer to the opposite question--what contributes to a composer's popularity--instead of our original question; but we really have no way of knowing. It's possible that composers' skill is affected by popularity--that they are inspired to write better when the do well, that composers' popularity is affected by skill--that they become popular due to their writing, or that neither is true, and a different unseen factor is at play.

What we have found is that there is a relationship of some sort, and it is possible that we might be able to better understand that relationship through further study.

Works Cited

Create Custom Map. (n.d.). Retrieved April 21, 2016, from http://mapchart.net/

Goulding, P. G. (1992). *Classical music: The 50 Greatest Composers and Their 1,000 Greatest Works*. New York: Fawcett Columbine.

Musicalics: The Classical Music Database. (n.d.). Retrieved February 23, 2016, from http://musicalics.com/

Our World in Data. (n.d.). Retrieved April 15, 2016, from http://ourworldindata.org/data/population-growth-vital-statistics/life-expectancy/

Sheet Music Plus. (n.d.). Retrieved February 23, 2016, from http://www.sheetmusicplus.com/

The Cost of Uncertain Life Span. (n.d.). Retrieved April 15, 2016, from http://www.nber.org/papers/w14093

Data Summary

Composer	Rank	Birth Yr	Age @ Death	# Pub	Country of Birth	Period	Symph'c	Opera	Chamber	Symph'ies	Piano
Johann Sebas	1	1685	65	19518	Germany	Baroque	0	1	1	0	1
Wolfgang Am	2	1756	35	15496	Austria	Classical	1	1	1	1	1
Ludwig van B	3	1770	57	8774	Germany	Classical	1	1	1	1	1
Richard Wagr	4	1813	70	3120	Germany	Romantic	1	1	1	0	1
Franz Joseph	5	1732	77	6980	Austria	Classical	1	1	1	1	1
Johannes Bral	6	1833	64	4831	Germany	Romantic	1	0	1	1	1
Franz Shubert	7	1797	31	6835	Austria	Romantic	1	1	1	1	1
Robert Schurr	8	1810	46	3893	Germany	Romantic	1	1	1	1	1
George Frider	9	1685	74	334	Germany	Baroque	1	1	1	1	1
Pyotr Ilyitch T	10	1840	53	255	Russia	Romantic	1	1	1	1	1
Felix Mendels	11	1809	38	224	Germany	Romantic	1	1	1	1	1
Antonin Dvor	12	1841	63	3093	Czech Republic	Romantic	1	1	1	1	1
Franz Liszt	13	1811	75	53	Hungary	Romantic	1	1	1	0	1
FrŽdŽric Chop	14	1810	39	153	Poland	Romantic	0	0	1	0	1
Igor Stravinsk	15	1882	88	915	Russia	Modern	1	1	1	1	1
Giuseppe Ver	16	1813	87	3537	Italy	Romantic	1	1	1	1	1
Gustav Mahle	17	1860	50	1143	Czech Republic	Romantic	1	0	1	1	0
Sergei Prokofi	18	1891	61	946	Ukraine	Modern	1	1	1	1	1
Dmitri Shosta	19	1906	68	1145	Russia	Modern	1	1	1	1	1
Richard Strau	20	1864	85	1488	Germany	Romantic	1	1	1	0	1
Hector Berlio:	21	1803	65	1013	France	Romantic	1	1	1	1	1
Claude Debus	22	1862	55	2700	France	Modern	1	1	1	0	1
Giacomo Puc	23	1858	65	1542	Italy	Romantic	1	1	1	0	1
Giovanni de P	24	1525	68	451	Italy	Renaissance	0	0	1	0	0
Anton Bruckn	25	1824	72	1296	Austria	Romantic	1	0	1	1	1
Georg Telema	26	1681	86	3369	Germany	Baroque	1	1	1	1	1
Camille Saint-	27	1835	86	1904	France	Romantic	1	1	1	1	1
Jean Sibelius	28	1865	91	1355	Finland	Romantic	1	1	1	1	1
Maurice Rave	29	1875	62	1191	France	Modern	1	1	1	0	1
Gioacchino R	30	1792	76	2090	Italy	Romantic	1	1	1	1	1
Edvard Grieg	31	1843	64	50	Norway	Romantic	1	0	1	1	1
Cristoph Gluc	32	1714	73	782	Germany	Classical	1	1	1	0	0
Paul Hindemi	33	1895	68	1004	Germany	Modern	1	1	1	1	1
Claudio Mont	34	1567	76	864	Italy	Renaissance	0	1	0	0	0
BŽla Bart—k	35	1881	64	1064	Hungary	Modern	1	1	1	0	1
CŽsar Franck	36	1822	67	1143	Belgium	Romantic	1	1	1	1	1
Antonio Vival	37	1678	63	4650	Italy	Baroque	1	1	1	0	0
Georges Bizet	38	1838	36	1994	France	Romantic	1	1	1	1	1
Modest Muss	39	1839	42	818	Russia	Romantic	1	1	0	0	1
Jean-Philippe	40	1683	80	698	France	Baroque	1	1	1	0	0
Gabriel FaurŽ	41	1845	79	1873	France	Romantic	1	1	1	0	1

Acknowledgements

Dr. Alisha Waller:

Our professor this semester challenged us in essentially every way possible academically. She encouraged us to interrogate our textbook instead of just reading it, form *effective* study groups, and dedicate the time necessary to get the most out of her class. Dr. A was willing to meet with us numerous times outside of her regular office hours in order to explain confusing concepts and we both feel like we have achieved a much deeper understanding of statistics than we would have under a less engaged teacher. We hope this project is a reflection of her effectiveness as a teacher.

2027 Squad:

Mike, Carolina, Jared, Drew, Alexandra and Abdalla all transferred to tech at the same time and took ISyE 2027 together last semester. Through study groups, groupme chats, and dining hall gatherings they were all a constant source of encouragement and inspiration for this project. I look forward to continuing our friendship and relying on each other for help throughout the remainder of our junior and senior level courses.

Maestro Thomas Ludwig:

A graduate of the Juilliard School, Thomas Ludwig is a conductor and composer based in Atlanta, and the founder and music director of the Beethoven Chamber Orchestra and the Ludwig Symphony Orchestra. I have had the great fortune to play under him as violist in the LSO, then the BCO, in the past two years, and briefly took private lessons with him. When I lost confidence in my ability as a violist, when my first teacher passed away after being sick for some time, when I have considered quitting orchestra entirely several times over the course of my freshman year in college, rehearsing with his orchestra has been a comfort and an inspiration. Maestro Ludwig offered many resources for data collection, and checked over all of the musical information presented in the project.