LA-UR-22-29152

Title:	Functional Data Modeling for Engineering Solutions
Author(s):	Myren, Samuel Thomas Wilkins Watts, Adam Christopher
Intended for:	The purpose of this presentation is to document the record of my student internship and allow me to share my experience at my university and elsewhere.
Issued:	2022-09-02

Los Alamos
NATIONAL LABORATORY

UNCLASSIFIED

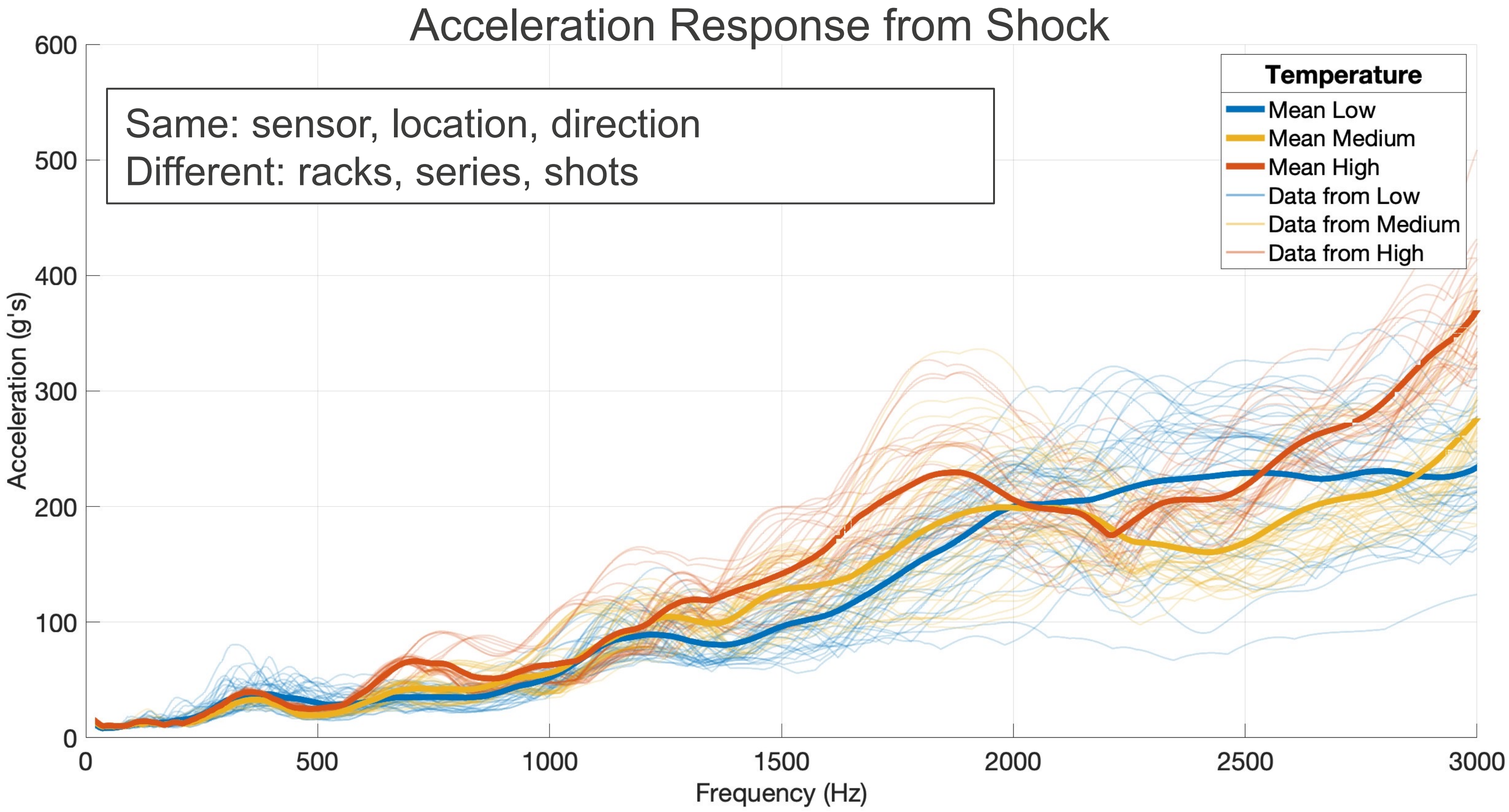
Functional Data Modeling for Engineering Solutions

Samuel Myren, Adam Watts
Engineering Technology and Design
Test Engineering (E-14) - Data Analysis Team

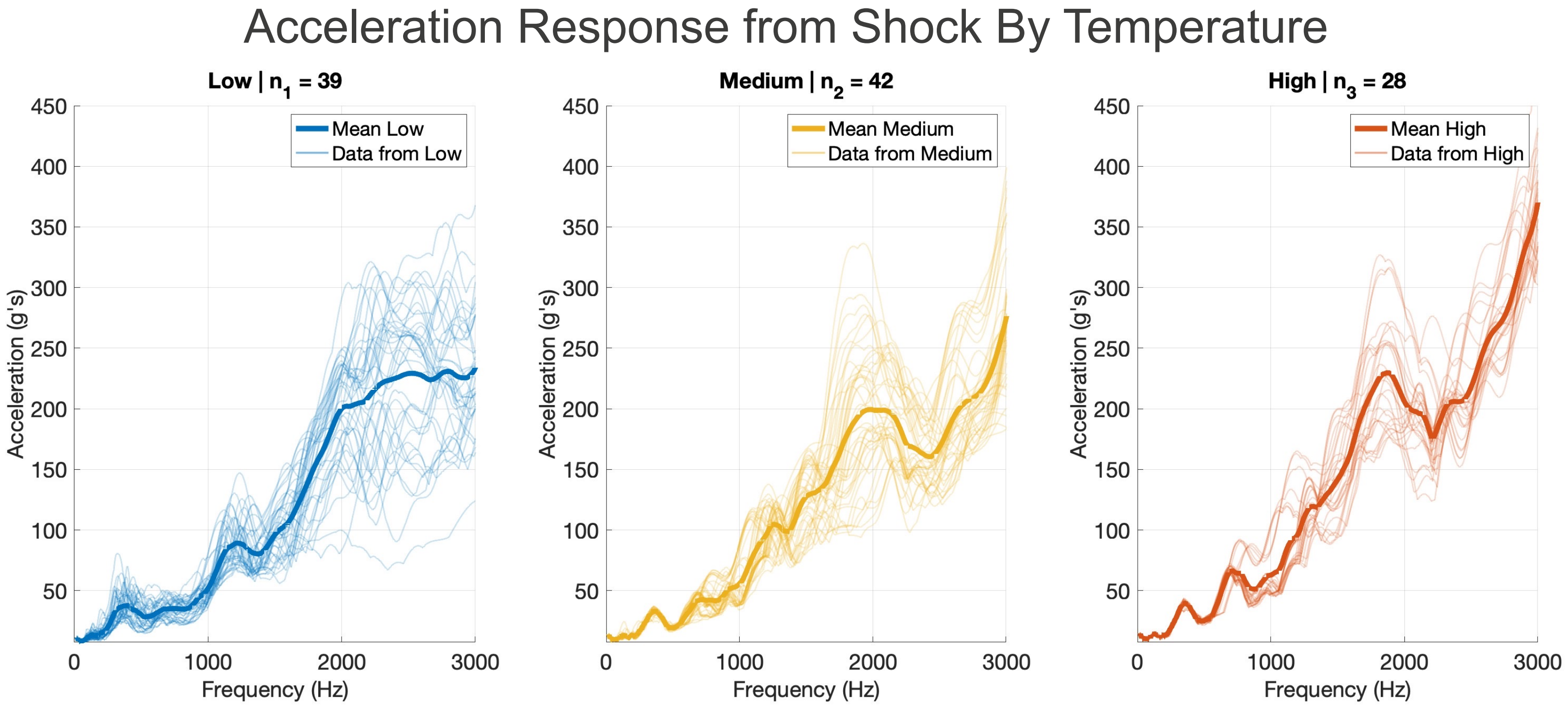
Wednesday, August 31, 2022

Managed by Triad National Security, LLC for the U.S. Department of Energy's NNSA

Is there a significant difference in the mean response between groups?

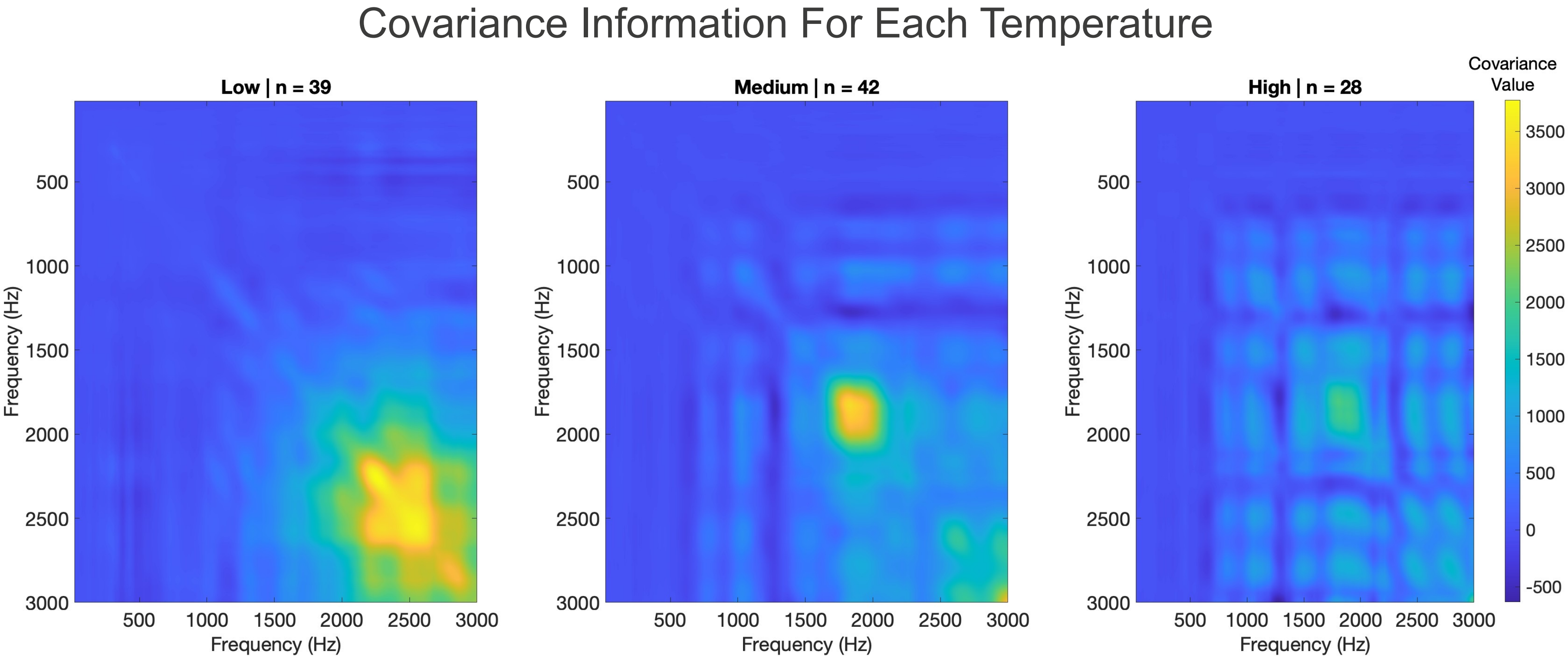


Functional-ANOVA gives a global answer to: do these functions arise from the same stochastic process?



If they are different, then test fewer times at a single temperature

Is there a statistically significant difference between the covariance across each group?



Motivation

Model

Properties of model

Testing hypotheses

Assumptions revisited

Generalization to beyond

Motivation ✓

Model

Properties of model

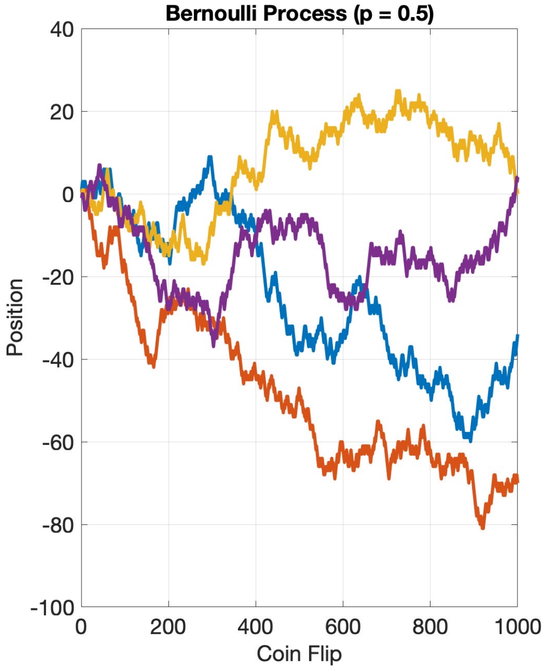
Testing hypotheses

Assumptions revisited

Generalization to beyond

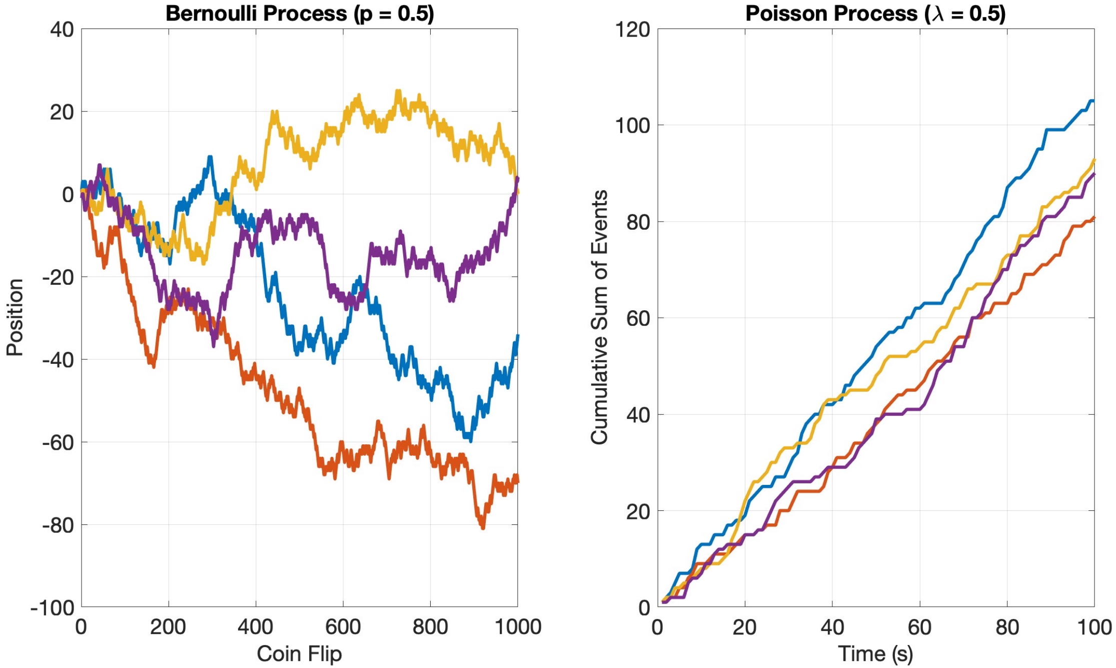
A stochastic process is a random function generator

$$y(t) \sim SP(\eta(t), \gamma)$$



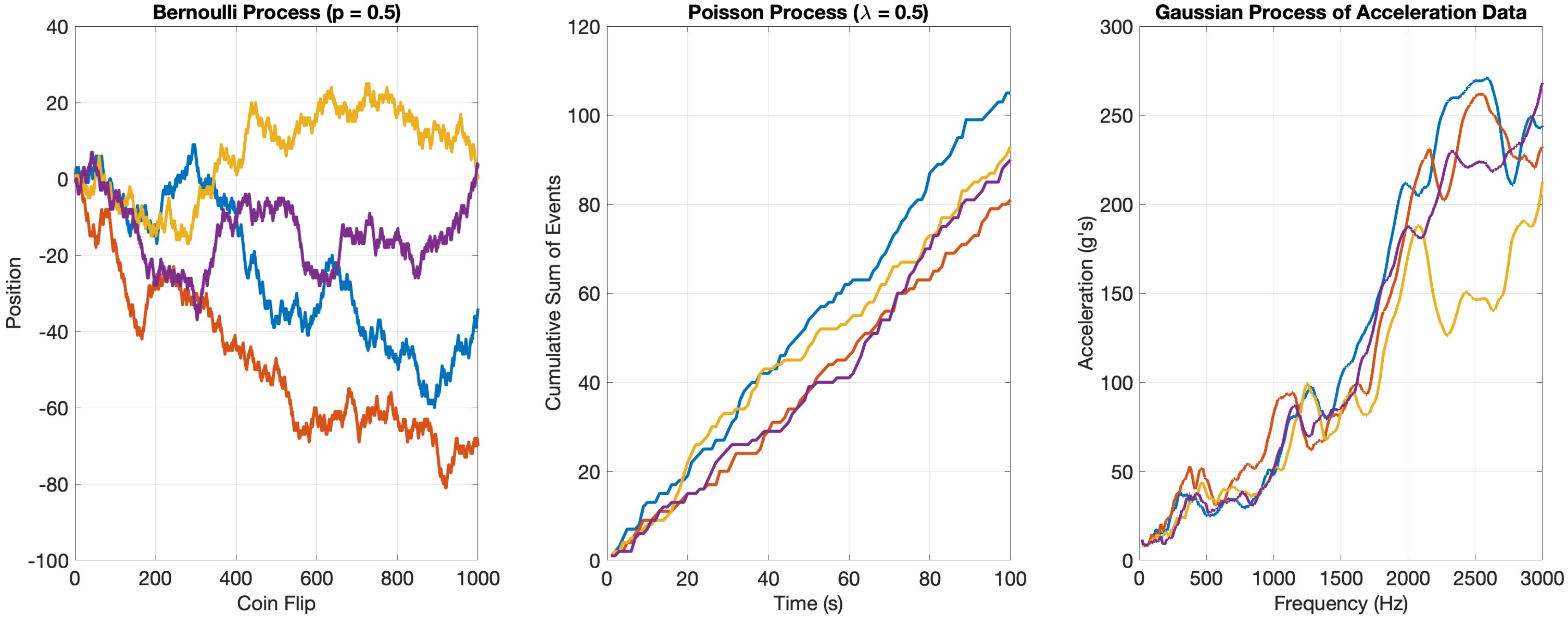
A stochastic process is a random function generator

$$y(t) \sim SP(\eta(t), \gamma)$$



A stochastic process is a random function generator

$$y(t) \sim SP(\eta(t), \gamma)$$

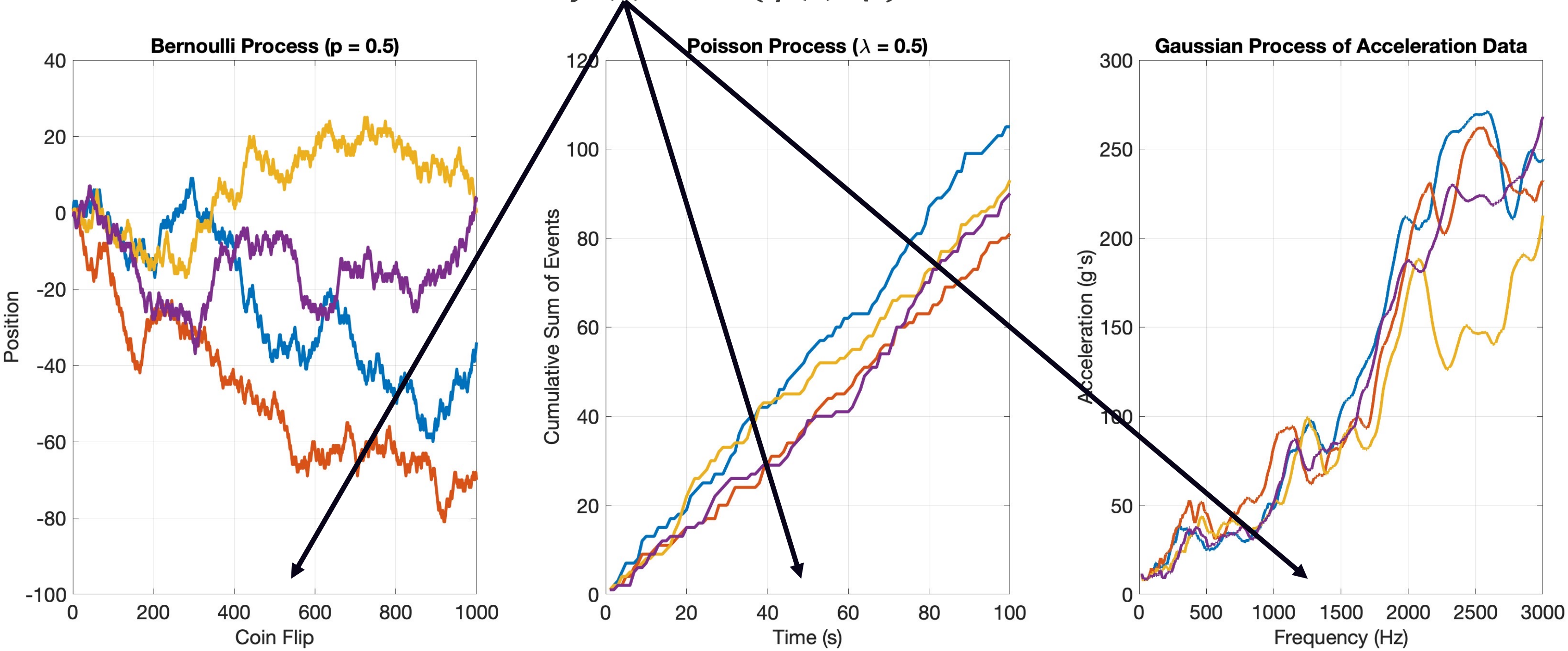


A Gaussian process:

- Is flexible for fitting various functions
- Provides methods for uncertainty analysis
- Can be achieved through ‘tricks’

A stochastic process is a random function generator

$$y(t)\ \sim SP(\eta(t),\gamma)$$

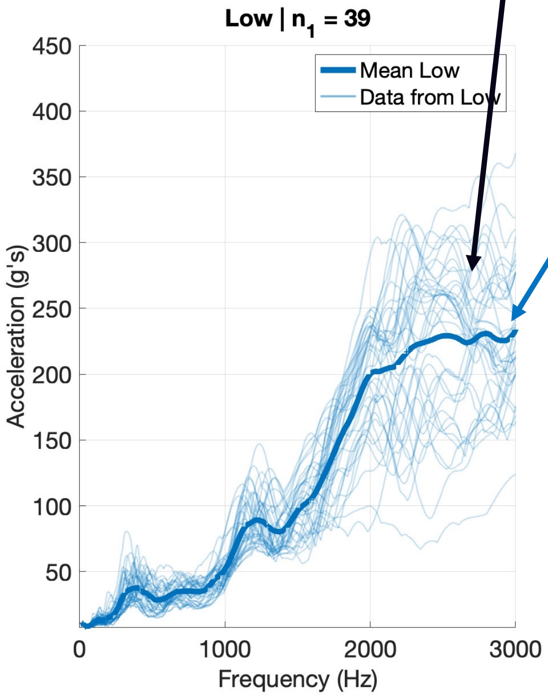


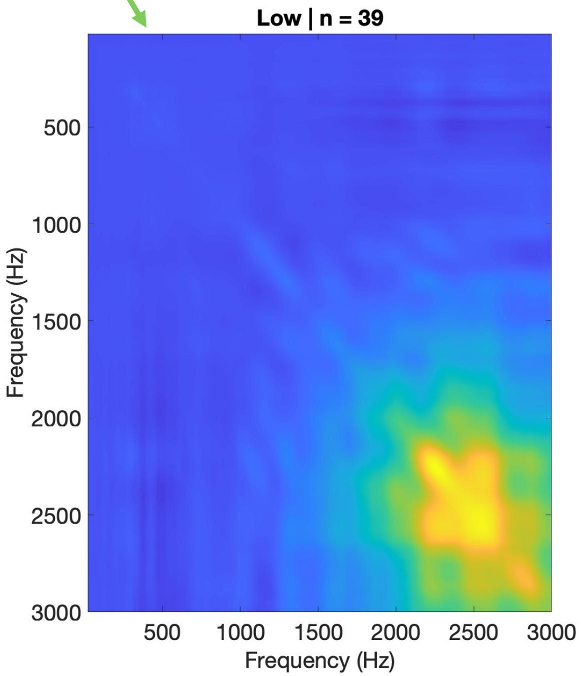
The dependent variable t represents any domain

$$t \in D$$

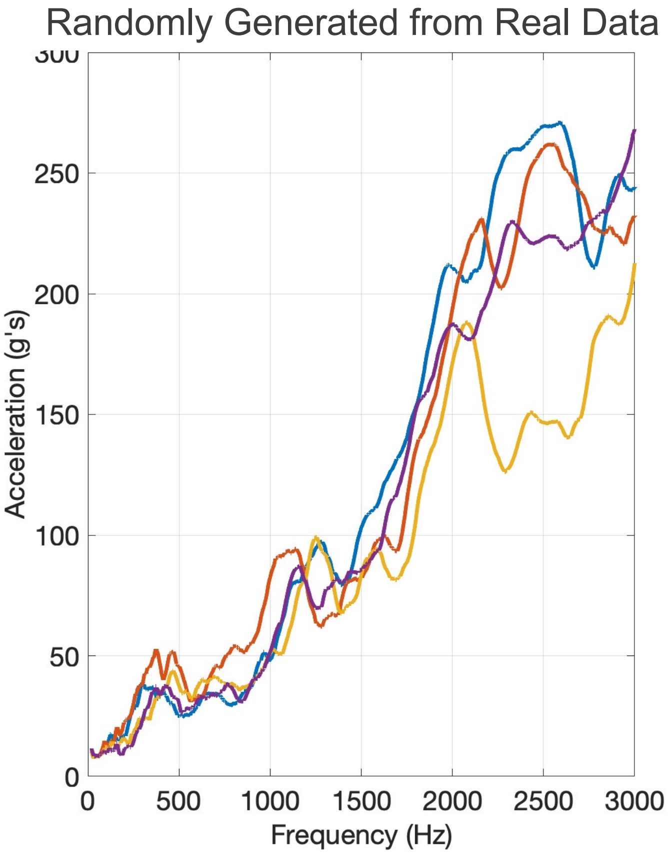
GPs are defined by a mean and covariance

$$y_j(t) \sim GP(\eta(t), \gamma(s,t)) \qquad j = 1 \ldots n$$

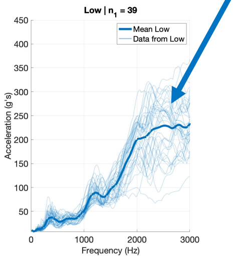


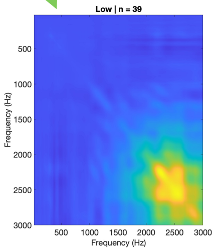


Since we have an approximation of the mean and variance, we can produce pseudo-samples



$$y_j(t) \sim GP(\eta(t), \gamma(s,t))$$

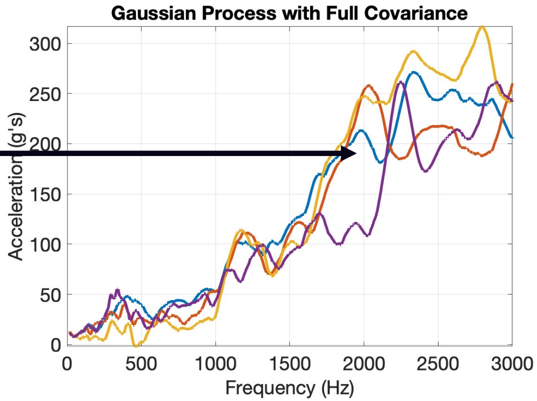


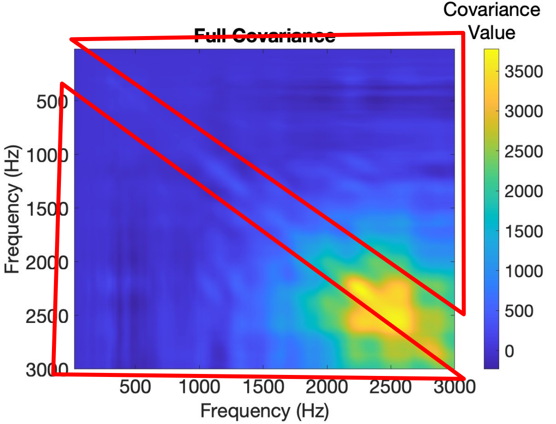


Understanding role of covariance in a GP

$$y(t) \sim GP(\eta(t), \gamma(s,t))$$

Curves are continuous and smooth

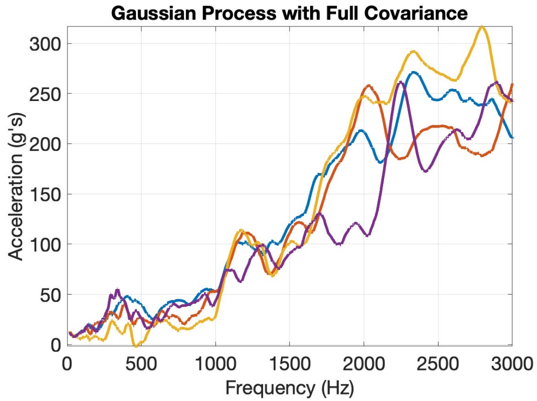

Gaussian Process with Full Covariance

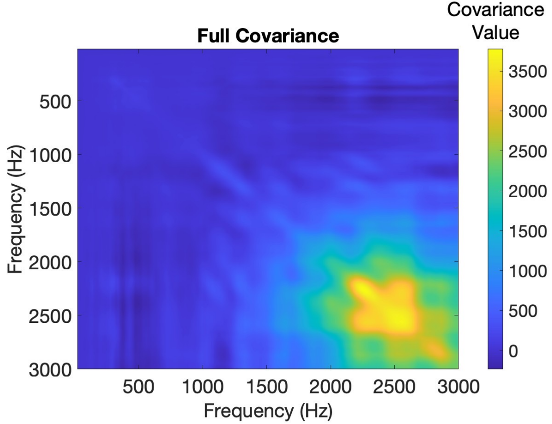

Full Covariance

Covariance between points s and t contained in off-diagonal

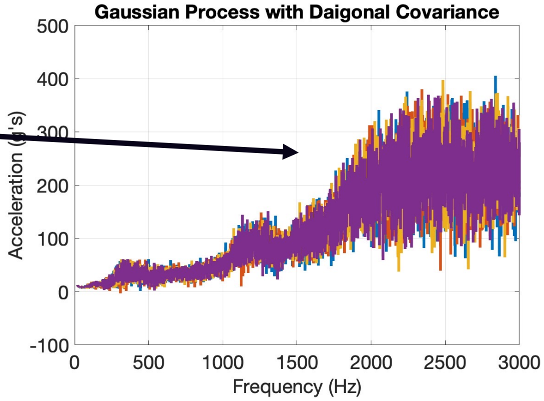
Understanding role of covariance in a GP

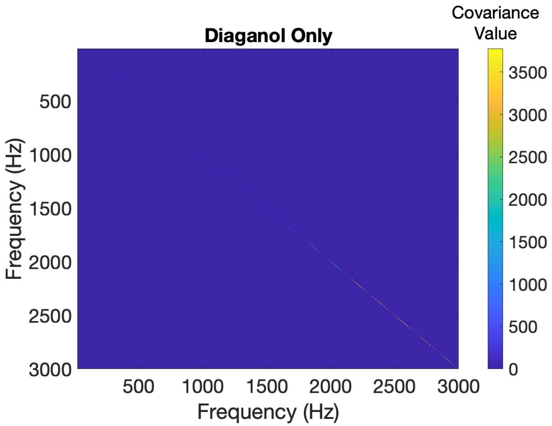
$$y(t) \sim GP(\eta(t), \gamma(s,t))$$


Gaussian Process with Full Covariance


Full Covariance

Curves follow mean, but randomly distributed about it


Gaussian Process with Daigonal Covariance


Diaganol Only

Off-diagonal values set to zero

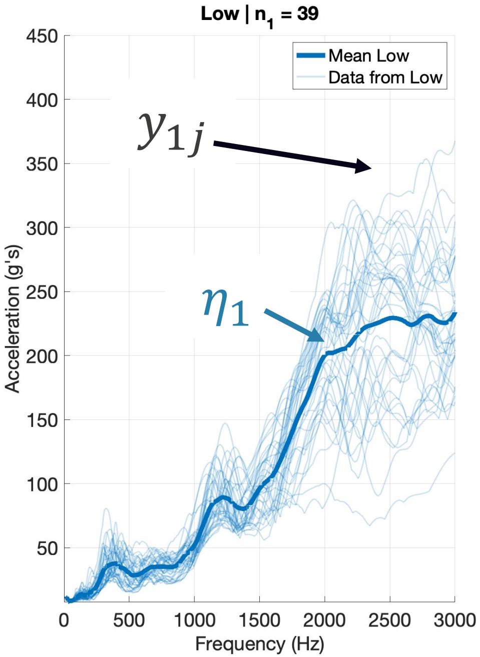
Assume our functional data follow a Gaussian process

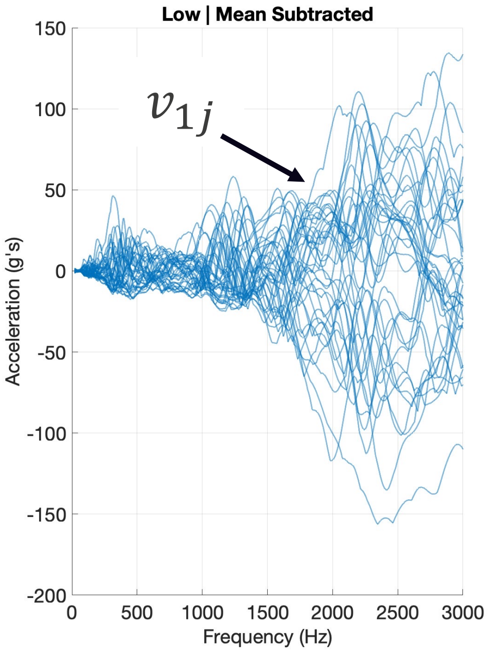
Main effect model: $y_{ij}(t) = \eta_i(t) + v_{ij}(t)$

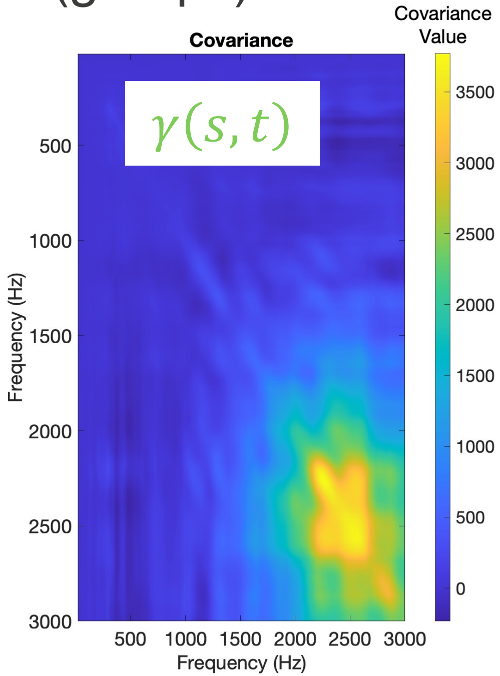
where $v_{ij}(t) \overset{iid}{\sim} GP(0, \gamma)$

observation = mean + Gaussian error

$j = 1 \ldots n_i$ (samples) and $\mathrm{i} = 1 \ldots k$ (groups)

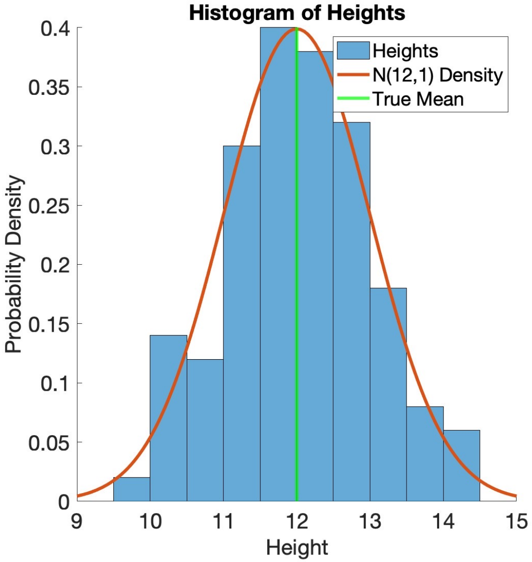






- Motivation ✓
- Model ✓
- Properties of model
- Testing hypotheses
- Assumptions revisited
- Generalization to beyond

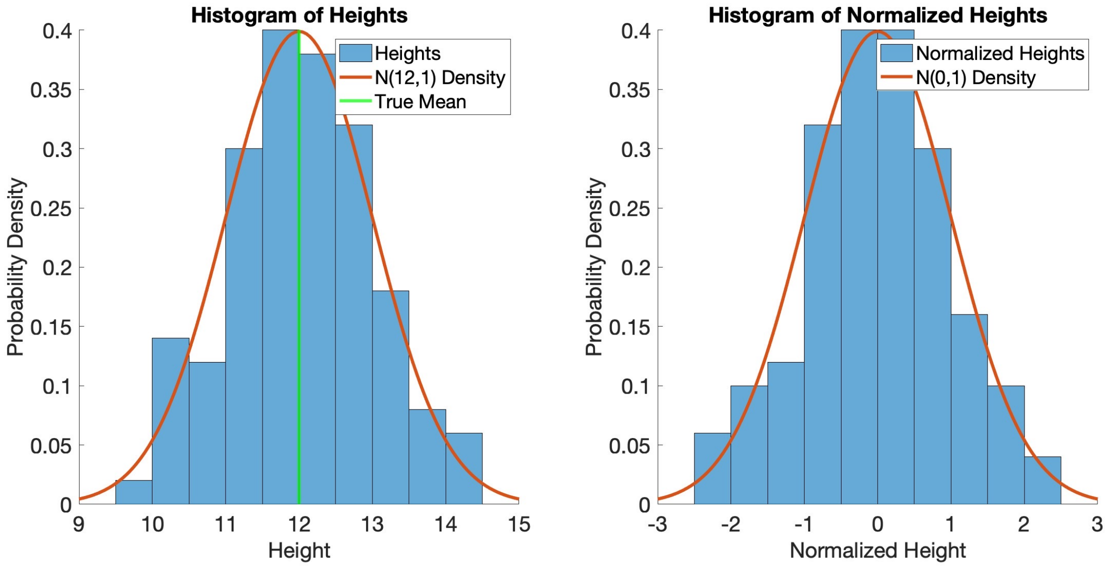
Gaussian distributions have helpful properties for tests

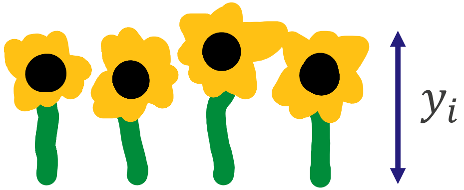


y_i is the observed Daisy height

$$y_i \sim N(\eta = 12, \gamma = 1)$$

Gaussian distributions have helpful properties for tests



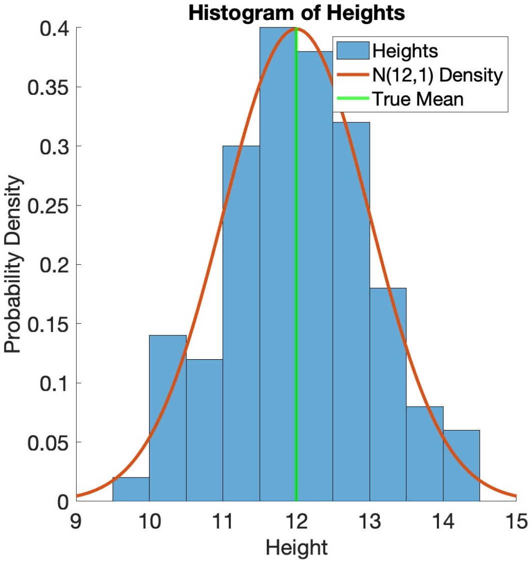


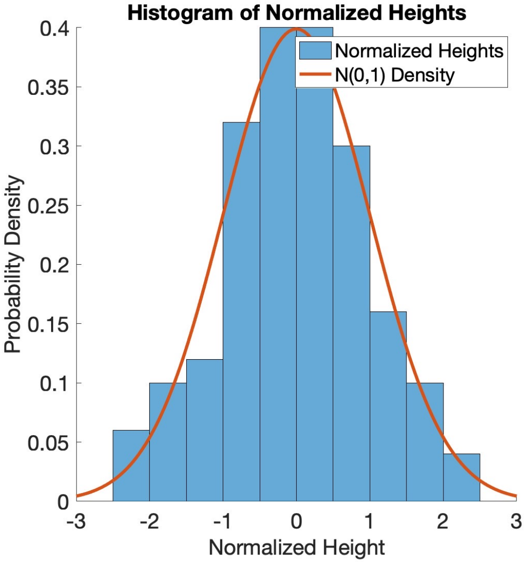
$$\frac{y_i - 12}{\sqrt{1}} \sim N(0,1)$$

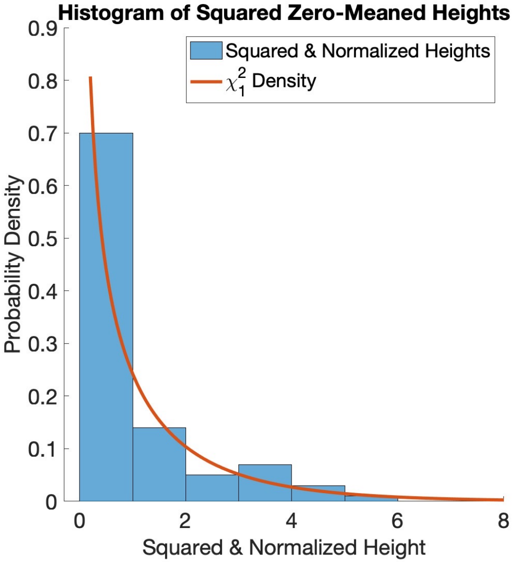
Hmmm

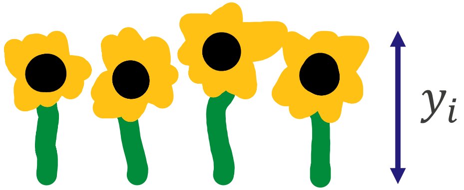
$$\left(\frac{y - 12}{\sqrt{1}}\right)^2 \sim ???$$

Gaussian distributions have helpful properties for tests









Fact: if $y_i \sim N(\eta,\gamma)$, then $\left(\frac{y_i - \eta}{\sqrt{\gamma}}\right)^2 \sim \chi_1^2$

We can hypothesize anything for the mean

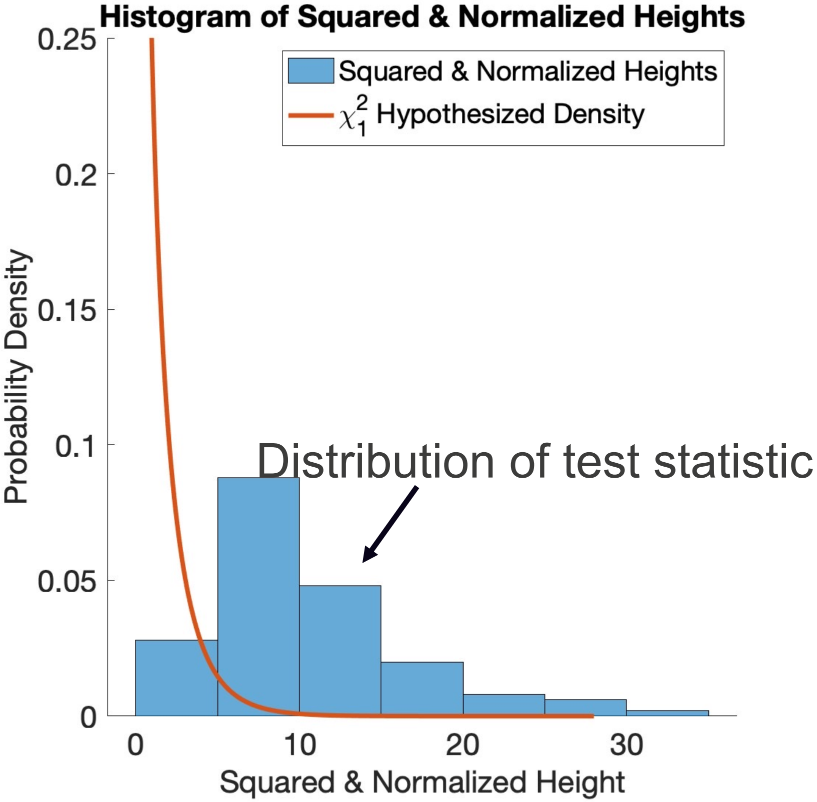
We want to test if the mean is 9 inches:

$$\frac{y_i - 9}{1} \sim N(0,1)$$

We do this to our data and look at the distribution of:

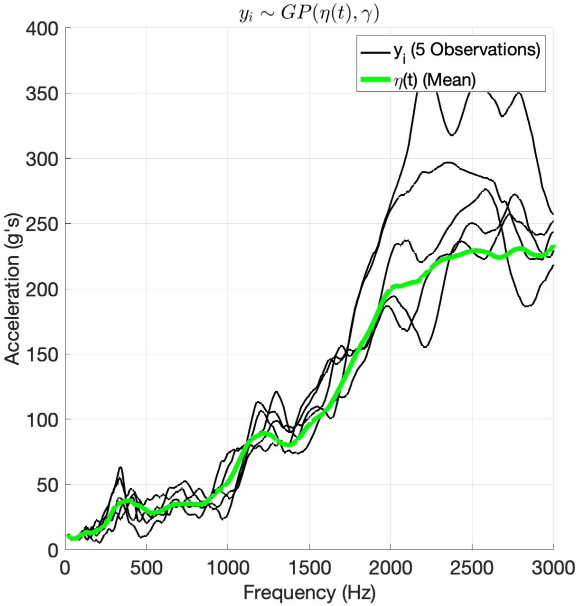
$$\left(\frac{y_i - 9}{1}\right)^2$$

Our data do not follow the known distribution if our hypothesis is true



Distribution of test statistic

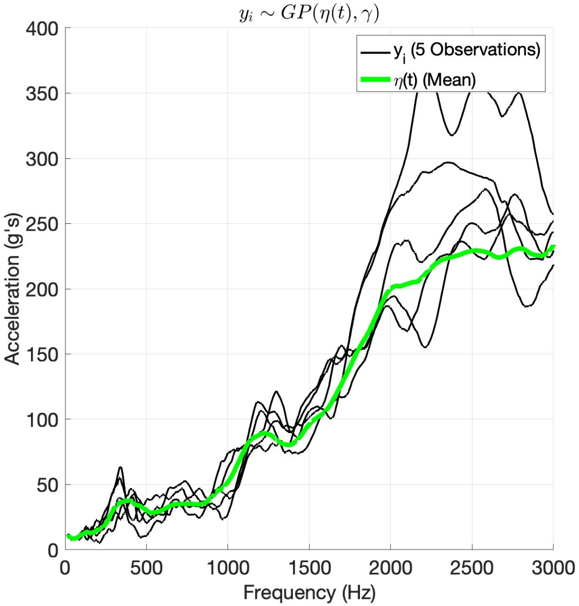
GPs have similar useful properties



Define the estimated mean

$$\bar{y}_{\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_j$$

GPs have similar useful properties

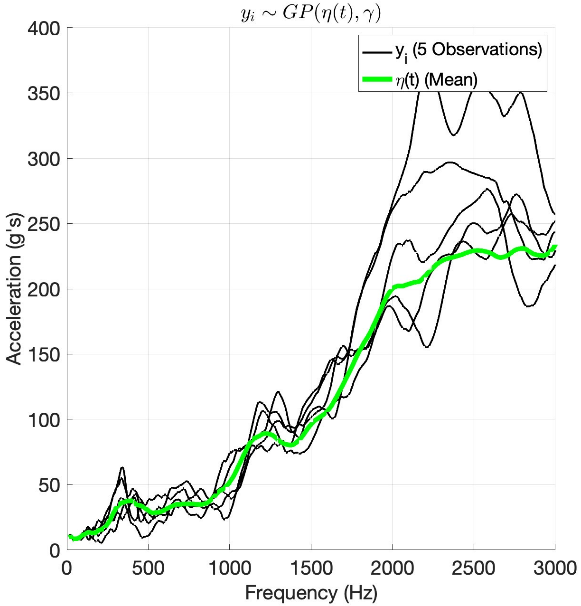


Fact, if $y_j(t) = \eta(t) + v_j(t)$ where $y_j(t) \sim GP(\eta(t), \gamma)$, then

$$\bar{y}_{\cdot}(t) \sim GP\left(\eta(t), \frac{\gamma}{n}\right)$$

Linear combinations of Gaussian processes are Gaussian processes

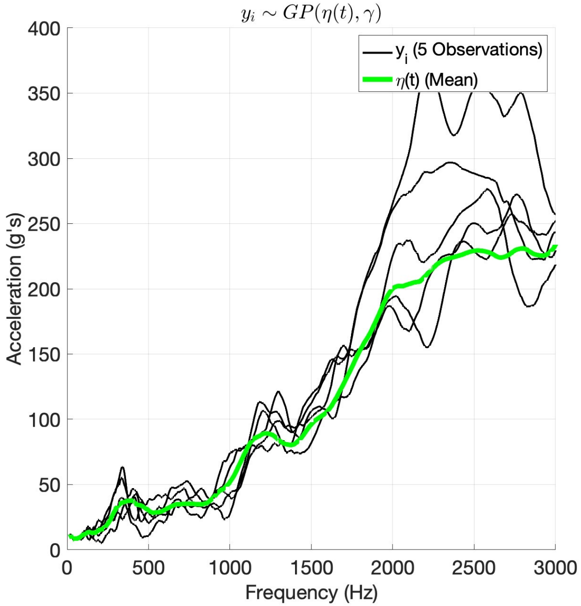
GPs have similar useful properties

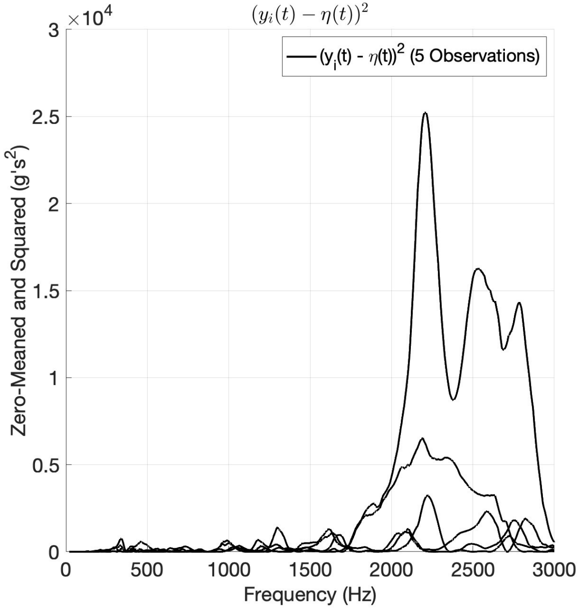


Then, by rules of expectations and variances, we have

$$\sqrt{n}\left(\bar{y}_{\cdot}(t) - \eta(t)\right) \sim GP(0,\ \gamma)$$

GPs have similar useful properties

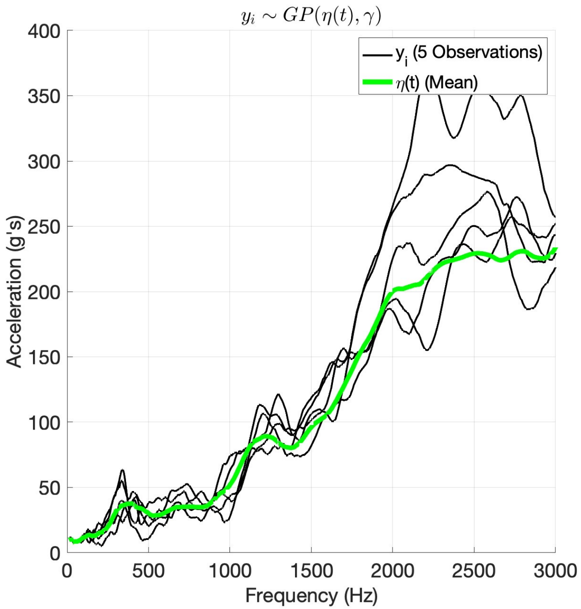


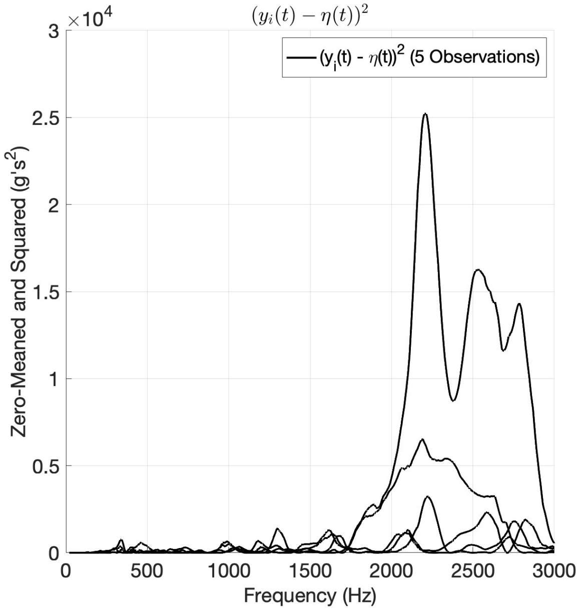


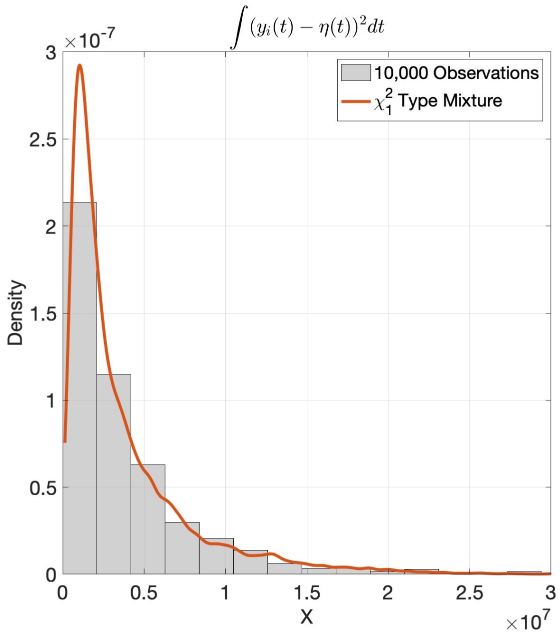
Fact, if $y_j(t) = \eta(t) + v_j(t)$ where $y_j(t) \sim GP(\eta(t), \gamma)$, then

$$\int n\left(\bar{y}_{\cdot}(t) - \eta(t)\right)^2 dt \rightarrow ???$$

GPs have similar useful properties







Fact: if $y_j(t) = \eta(t) + v_j(t)$ where $y_j(t) \sim GP(\eta(t), \gamma)$, then

$$\int n\left(\bar{y}_{\cdot}(t) - \eta(t)\right)^2 dt \ \to \sum_{r=1}^{m} \lambda_r A_r$$

where $A_r \sim \chi_1^2$ and the λ_r are the eigenvalues of γ

A chi-squared type mixture is fundamental

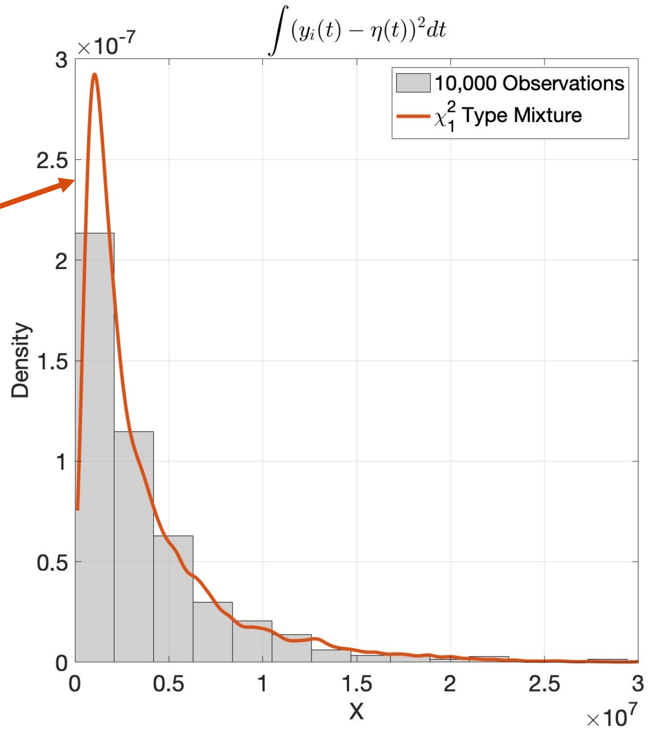
A zero-meaned, squared, and integrated GP follows a Chi-squared type mixture

This is how we symbolize it:

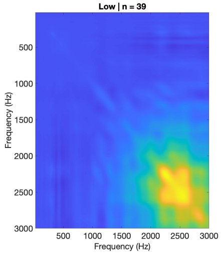
$$\# \sim \sum_{r=1}^{m} \lambda_r A_r$$

The $A_r \sim \chi^2_1$

The λ_r coefficients are the eigenvalues of γ



We estimate γ with $\hat{\gamma}$ →



The comparison is lacking in the denominator realm

Scalar Gaussian

Fact: if $y_i \sim N(\eta, \gamma)$, then $\frac{(y_i - \eta)^2}{\gamma} \sim \chi_1^2$

The comparison is lacking in the denominator realm

Scalar Gaussian

Fact: if $y_i \sim N(\eta, \gamma)$, then $\frac{(\mathrm{y_i} - \eta)^2}{\gamma} \sim \chi_1^2$

Functional Gaussian process

Fact: if $y_j(t) = \eta(t) + v_j(t)$ where $y_j(t) \sim GP(\eta(t), \gamma)$, then

$$\int n\left(\bar{y}_{\cdot}(t) - \eta(t)\right)^2 dt \rightarrow \sum_{r=1}^{m} \lambda_r A_r$$

The comparison is lacking in the denominator realm

Scalar Gaussian

Fact: if $y_i \sim N(\eta, \gamma)$, then $\frac{(y_i - \eta)^2}{\gamma} \sim \chi_1^2$

Functional Gaussian process

Fact: if $y_j(t) = \eta(t) + v_j(t)$ where $y_j(t) \sim GP(\eta(t), \gamma)$, then

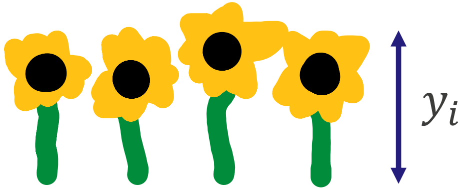
$$\int n\left(\bar{y}_{\cdot}(t) - \eta(t)\right)^2 dt \rightarrow \sum_{r=1}^{m} \lambda_r A_r$$

?

Must be a scalar

The distribution of a Gaussian variance

Estimated variance of daisy height

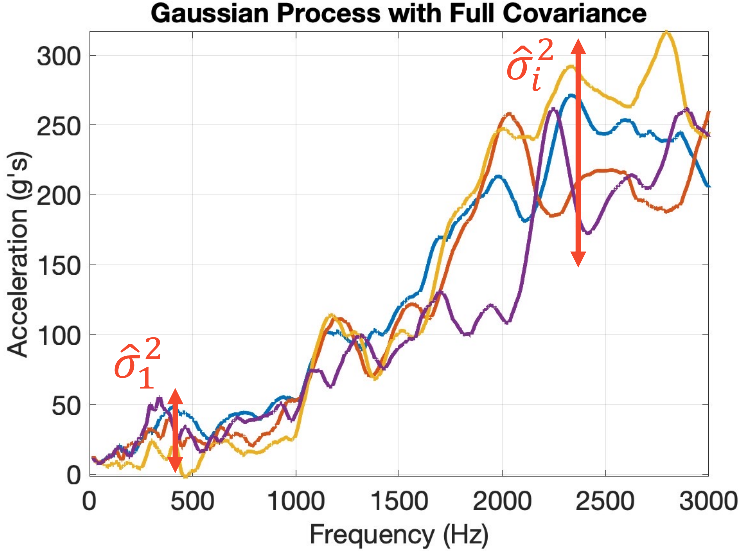


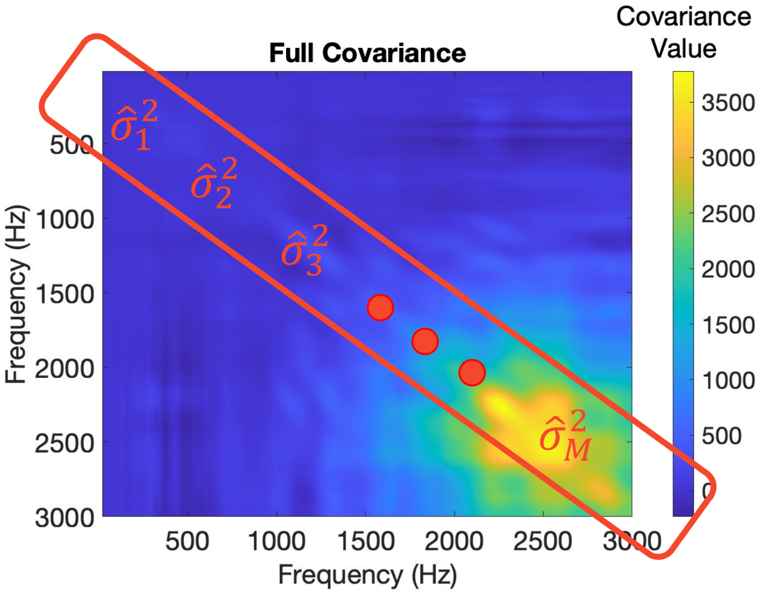
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

$$\frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2_{n-1}}{n-1}$$

True variance

The distribution of covariance information





Analogous to how independent & normal variance estimate ($\hat{\sigma}^2$) is distributed as χ^2_{n-1}

M points along the domain, but only n samples

$$tr(\hat{\gamma}) \rightarrow \frac{\sum_{r=1}^{m} \lambda_r E_r}{n-1}$$ where $A_r \sim \chi^2_{n-1}$ and the λ_r are the eigenvalues of γ

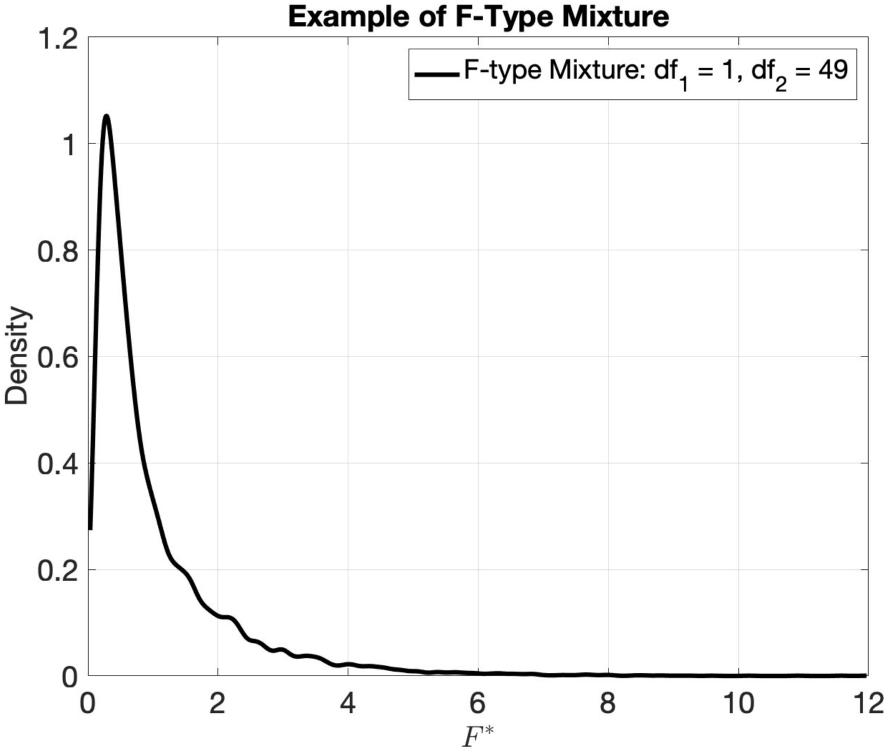
The test statistic considers both hypothesis information and covariance information

$$\text{Num} = \int n\left(\bar{y}_{.}(t) - \eta(t)\right)^2 dt$$
$$Den = tr(\hat{\gamma})$$

$$E(\text{Num}) = tr(\gamma)$$
$$E(Den) = tr(\gamma)$$

$$F^* = \frac{Num}{Den} \rightarrow \frac{\frac{\sum_{r=1}^{m} \lambda_r A_r}{1}}{\frac{\sum_{r=1}^{m} \lambda_r E_r}{n-1}}$$

Ratio of Chi-squared type mixtures or F-type mixture



Motivation ✓

Model ✓

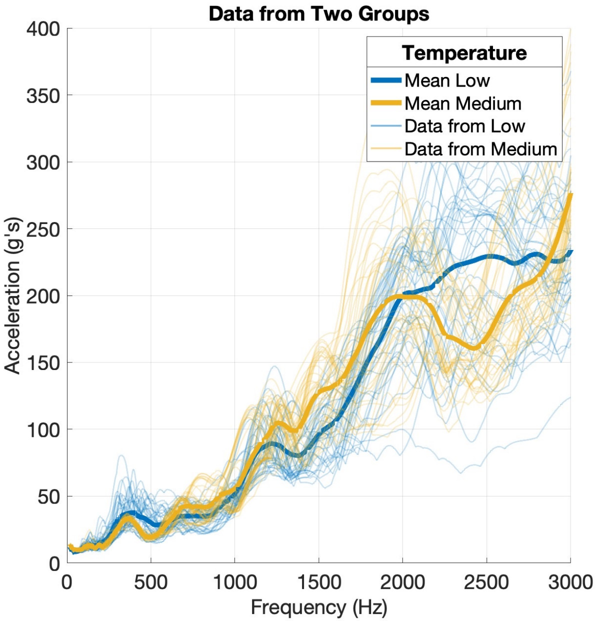
Properties
of model ✓

Testing hypotheses

Assumptions revisited

Generalization
to beyond

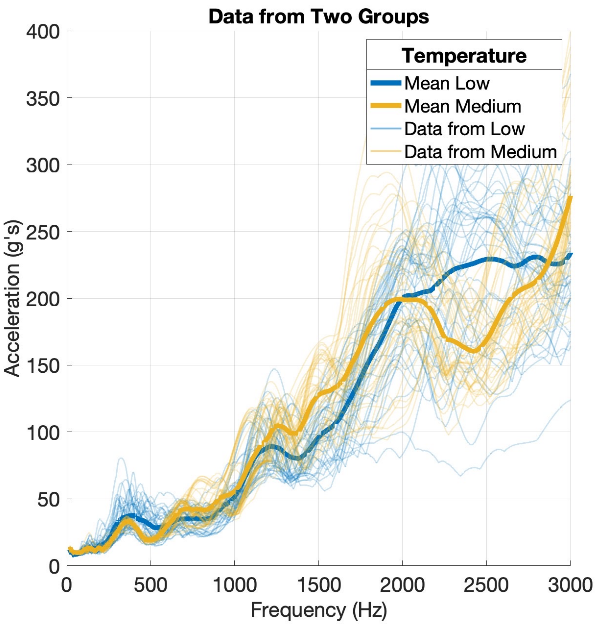
Now, we hypothesize Mean Low == Mean Medium

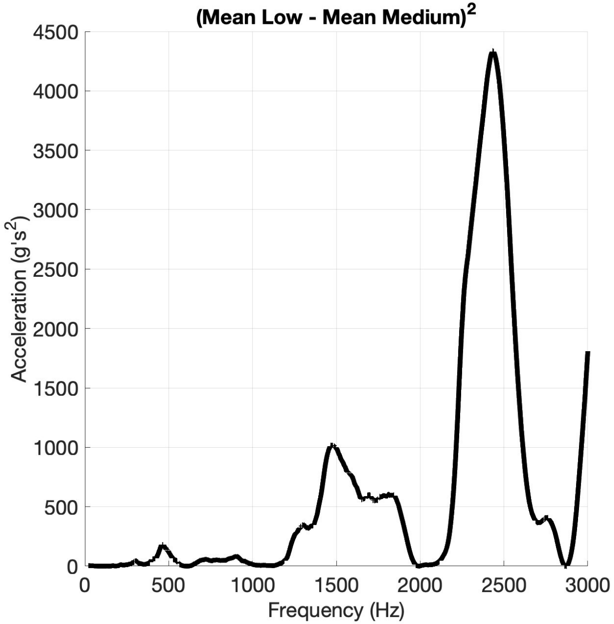


The null hypothesis is $H_0 \colon \eta_1 = \eta_2$

Under this hypothesis $\Delta(\mathrm{t}) = \sqrt{\frac{n_1 n_2}{n}}(\overline{y_{1.}} - \overline{y_{2.}}) \sim GP(0, \gamma)$

Now, we hypothesize Mean Low == Mean Medium

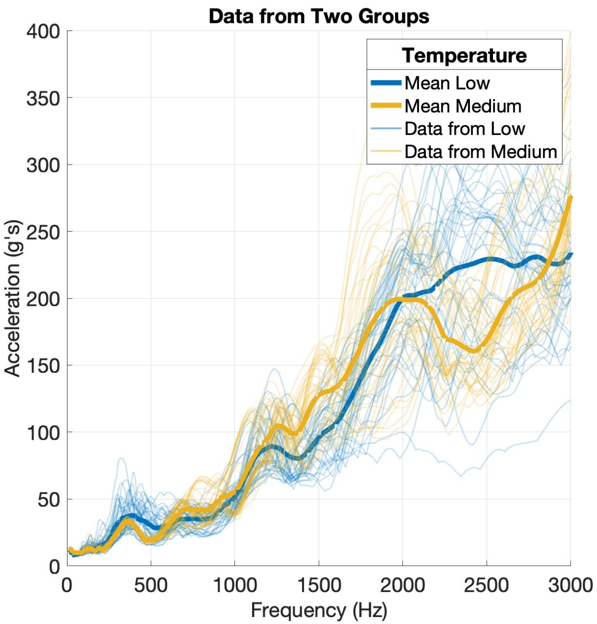


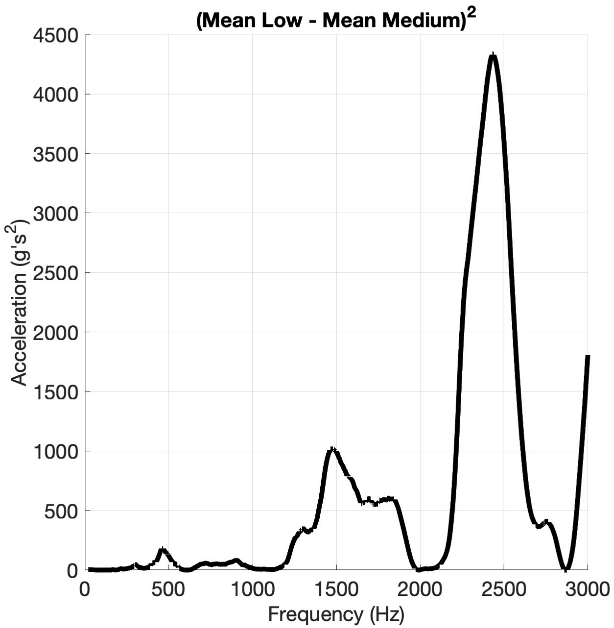


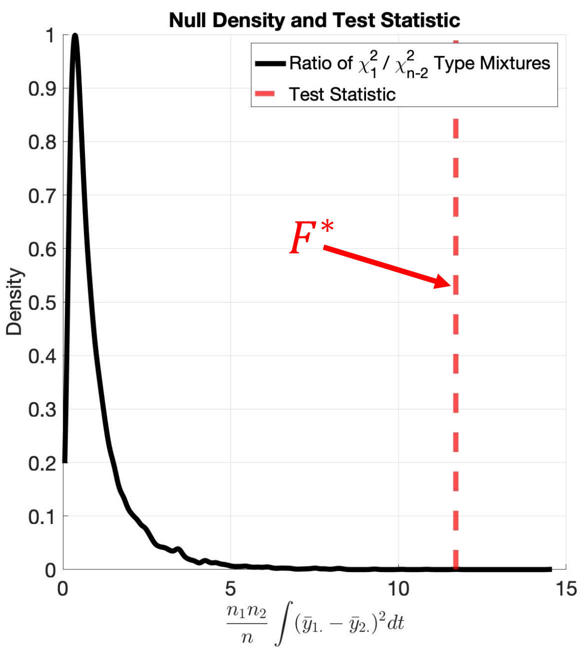
The null hypothesis is $H_0\!: \eta_1 = \eta_2$

Under this hypothesis $\Delta(\mathrm{t}) = \sqrt{\frac{n_1 n_2}{n}}(\overline{y_{1.}} - \overline{y_{2.}}) \sim GP(0, \gamma)$

Now, we hypothesize Mean Low == Mean Medium





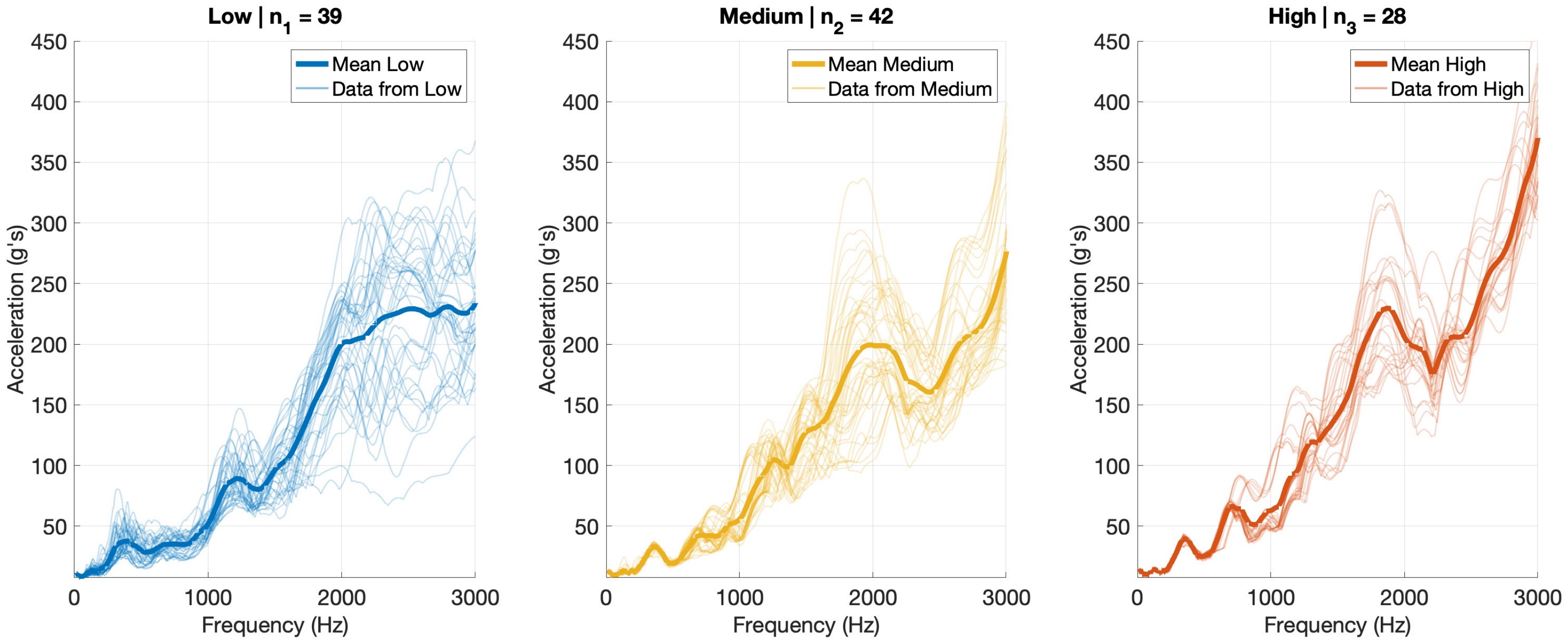


The null hypothesis is $H_0 : \eta_1 = \eta_2$

Under this hypothesis $\Delta(t) = \sqrt{\frac{n_1 n_2}{n}}(\overline{y_{1.}} - \overline{y_{2.}}) \sim GP(0, \gamma)$

And we get the test static's distribution as $F^* = \frac{\int \Delta^2(t)dt}{tr(\hat{\gamma})} \to \frac{\frac{\sum_{r=1}^{m} \lambda_r A_r}{1}}{\frac{\sum_{r=1}^{m} \lambda_r E_r}{n-2}}$

The extension to k groups happens naturally



The FAMILYWISE null hypothesis becomes $H_0\!: \eta_1(t) = \eta_2(t) = \cdots = \eta_k(t)$

For any group i, $\sqrt{n_i}(\overline{y_{i.}} - \overline{y_{..}}) \sim GP(0, \gamma)$ under the null that all effects are the same

The extension to k groups happens naturally

We construct the Sum of Squares of our Hypothesis:

$$SSH(t) = \sum_{i=1}^{k} n_i\left(\overline{y_{i.}}(t) - \bar{y}_{..}(t)\right)^2$$

Data from group i

Grand mean

If null is true, $\overline{y_{i.}} \approx \bar{y}_{..}$ and SSH is small.

Like before, we get a chi squared mixture:

$$\int \frac{SSH(t)dt}{k-1} \rightarrow \frac{\sum_{r=1}^{m} \lambda_r A_r}{k-1} \text{ where } A_r \sim \chi^2_{k-1}$$

Don’t forget to incorporate the covariance information

The Sum of the Squared Errors: model vs reality

$$SSE(t) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}(t) - \overline{y_{i.}}(t)\right)^2 = (n-k)\hat{\gamma}(t,t)$$

Curve j from group i

Group i Mean

SSE does not change regardless of hypothesis

$$\int \frac{SSE(t)dt}{n-k} = tr(\hat{\gamma}) \rightarrow \frac{\sum_{r=1}^{m} \lambda_r E_r}{n-k} \text{ where } E_r \sim \chi^2_{n-k}$$

Take the ratio to get the test statistic

Put it all together to get a ratio of chi-squared mixtures:

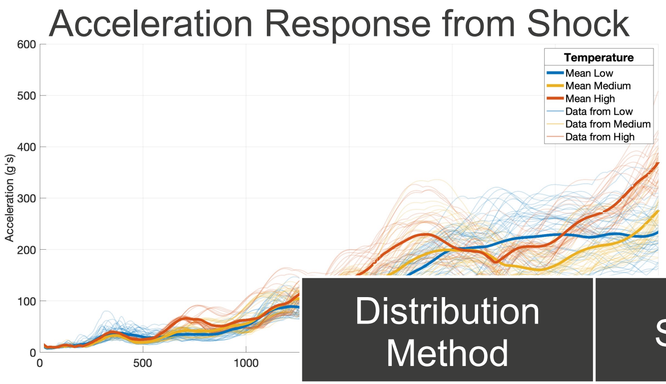
$$F^* = \frac{\int \frac{SSH(t)dt}{k-1}}{\int \frac{SSE(t)dt}{n-k}} \rightarrow \frac{\frac{\sum_{r=1}^{m} \lambda_r A_r}{k-1}}{\frac{\sum_{r=1}^{m} \lambda_r E_r}{n-k}}$$

To get this distribution, we can:

- simulate
- approximate via Welch-Satterthwaite
- bootstrap

Skip, but important! Tells you when to reject or accept.

Answering the long-awaited question

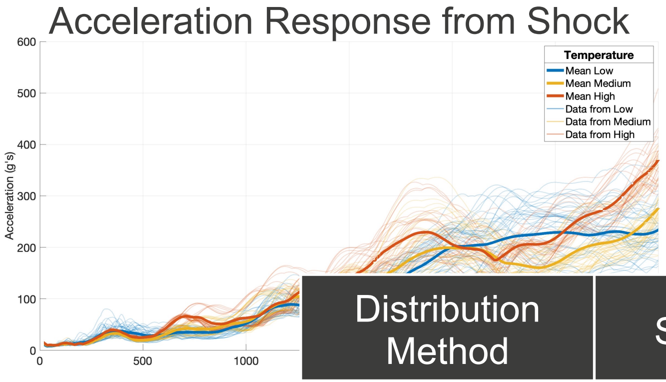

Acceleration Response from Shock

Distribution Method	Simulation	Approximation	Bootstrap
P-value	5e-14	1e-16	0

(All p-values are ~0)

Under the assumptions of Gaussian process distributed, independent observations, and equal covariance across groups, we find that there is a near zero probability of observing our data (or data more extreme) under the null hypothesis that the effects of temperature are **all** equal.

Answering the long-awaited question

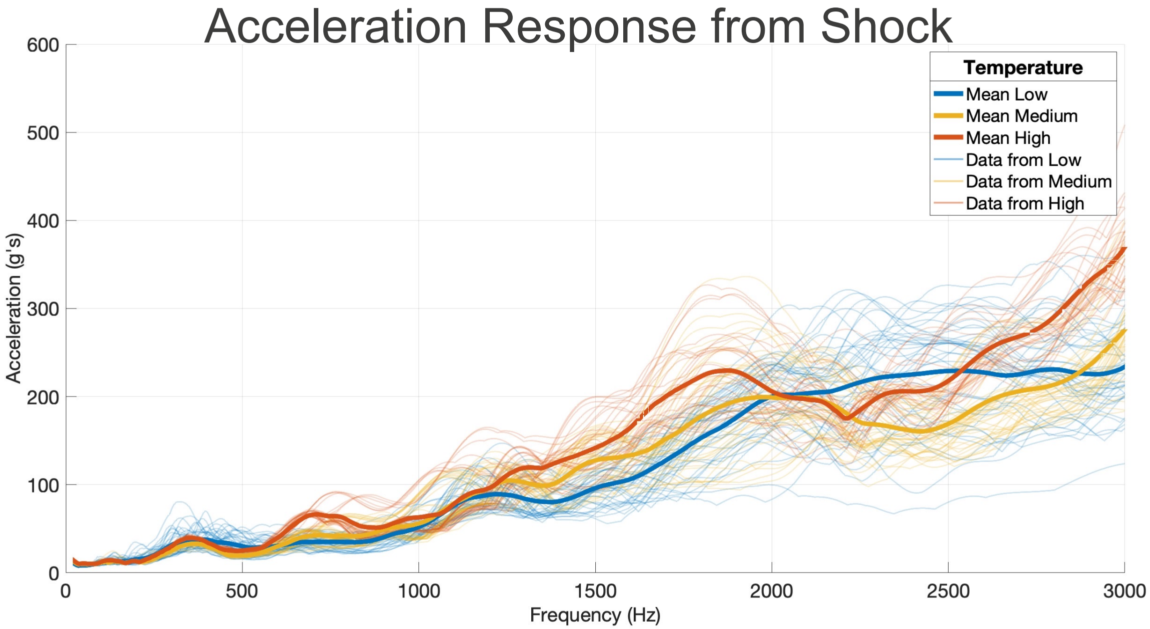


Distribution Method	Simulation	Approximation	Bootstrap
P-value	5e-14	1e-16	0

(All p-values are ~0)

At least one temperature causes a different response.

But which temperatures are different from each other?



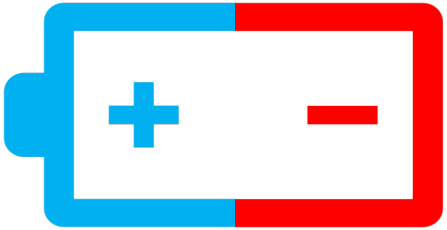
The PAIRWISE null hypothesis is $H_0: \eta_1 = \eta_3$ or $H_0: \eta_2 = \eta_3$

You could do a two-sample test for just these two groups

Don't leave data on the table – use contrasts!

Linear functions of the hypothesis are general and maximize the data used in the test

For null $H_0\colon \eta_1 = \eta_3$, rewrite this as $\eta_1 - \eta_3 = 0$



Contrasts form a linear function of SSH

$$SSH(t) = (C\hat{\eta}(t) - c)^T (CDC^T)^{-1} (C\hat{\eta}(t) - c)$$

Called contrasts because the procedure explicitly compares effects

$\binom{k}{2}$ contrasts are available: Low vs Medium | Medium vs High | Low vs High

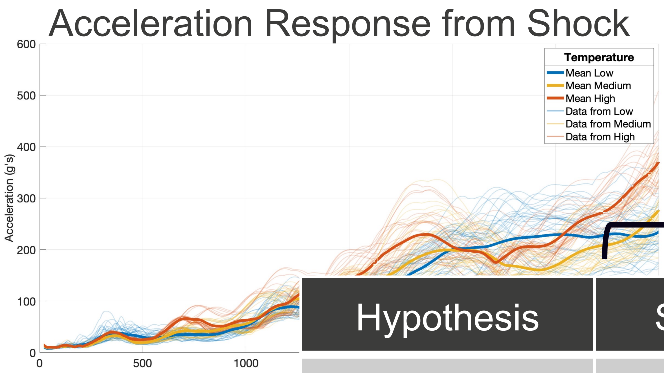
However, while SSH changes for a contrast, the SSE remains the same.

Test statistic is $\mathrm{F}^* = \dfrac{\frac{\int SSH(t)dt}{k-q}}{\frac{\int SSE(t)dt}{n-k}}$

where $q = rank(C)$ (contrast matrix)

Denominator uses all the data to inform the covariance estimate (good!)

Pairwise results conclude temperature matters


Acceleration Response from Shock

Distribution method

Hypothesis	Simulation	Approximation	Bootstrap
Low vs Med.	2e-5	1e-6	0
Low vs High	2e-14	3e-13	0
Med. vs High	2e-14	4e-11	0

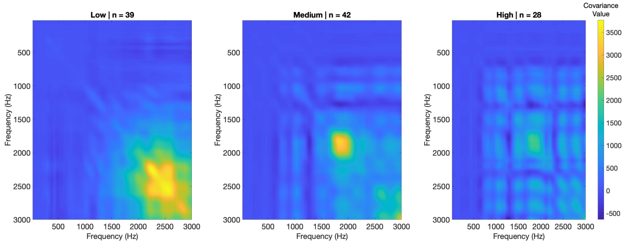
P-values
(All p-values are ~0)

All groups are different from each other

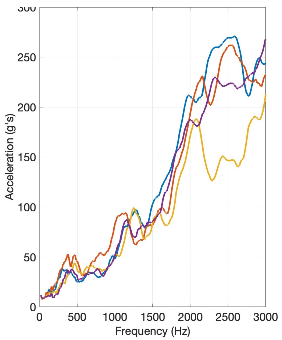
But wait!

Do you remember all our assumptions?

Equality of covariance across groups



Gaussian process distributed



$y_j(t) \sim GP(\eta(t), \gamma(s,t))$

Independence

Motivation ✓

Model ✓

Properties of model ✓

Testing hypotheses ✓

Assumptions revisited

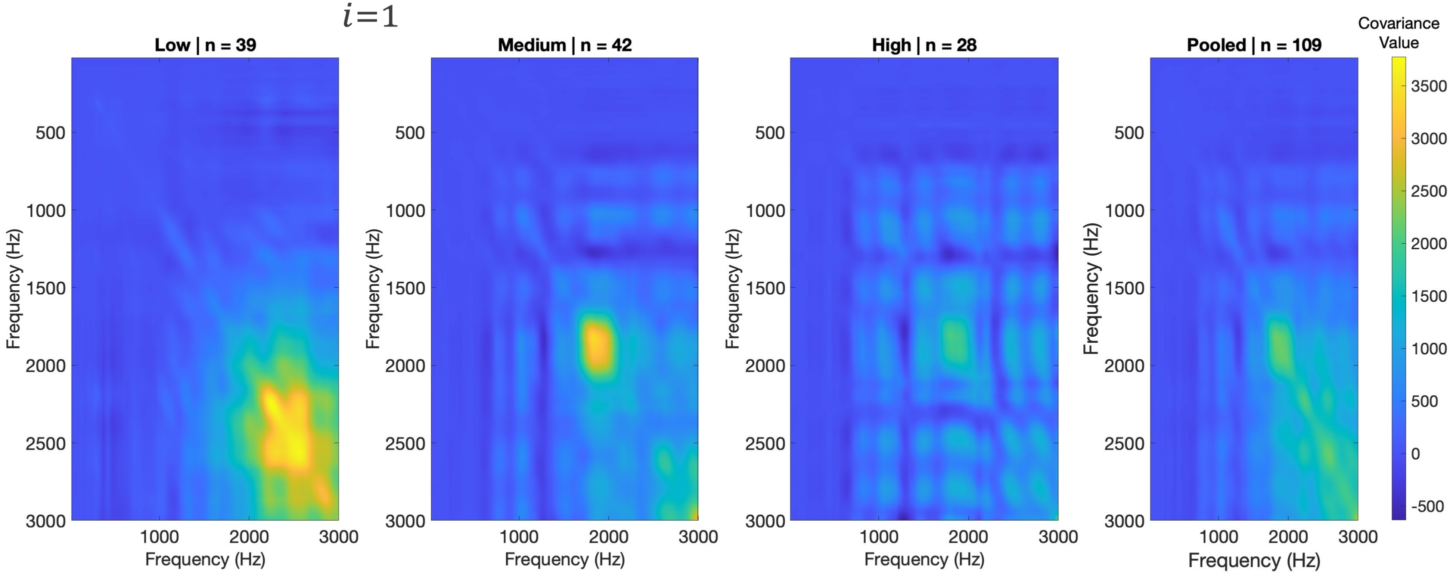
Generalization to beyond

Checking for equality of covariance across groups

New hypothesis $H_0\!: \gamma_1 = \gamma_2 = \gamma_3$

Utilize central limit theorem to make the claim that $\hat{\gamma}(s,t) \sim GP(\gamma(s,t), \omega)$

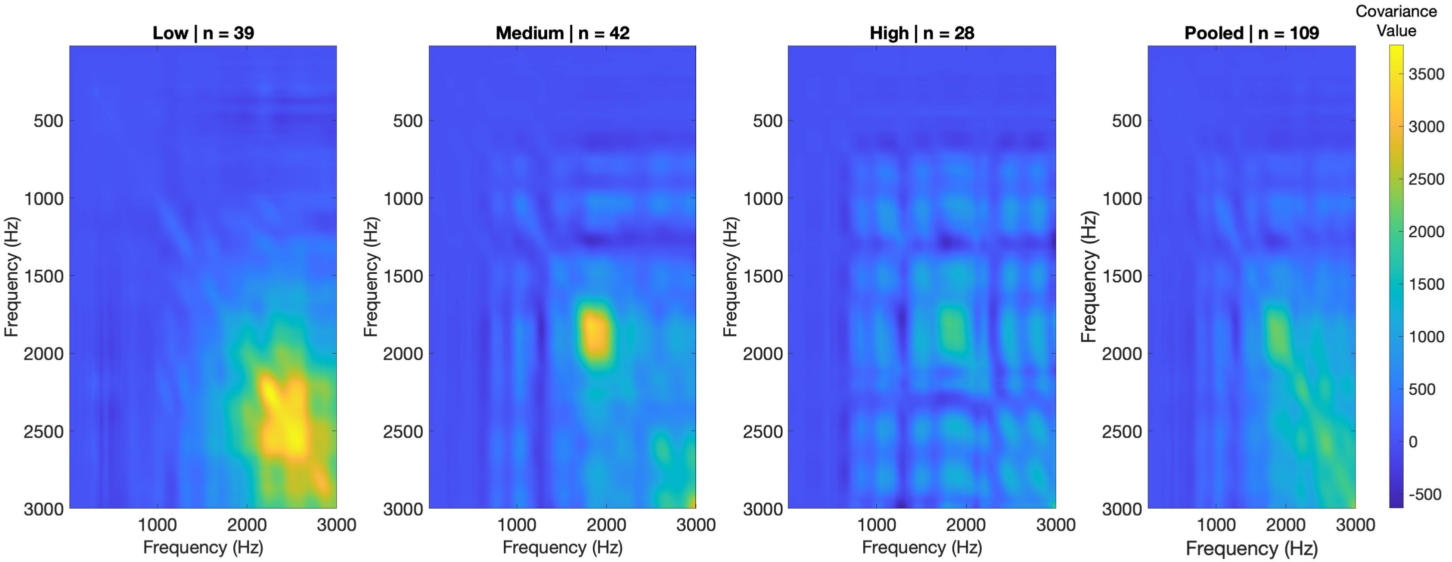
$$\mathrm{T}^* = \sum_{i=1}^{k} (n_i - 1) \int\int \left(\hat{\gamma}_i(s,t) - \hat{\gamma}_{pool}(s,t)\right)^2 dsdt$$



Regular distributional tests requires n large to exploit central limit theorem

Random permutation test easier to understand and more flexible

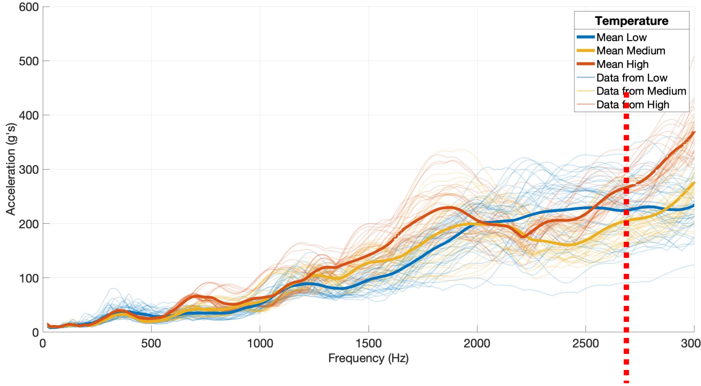
Covariance across groups are not approximately equal



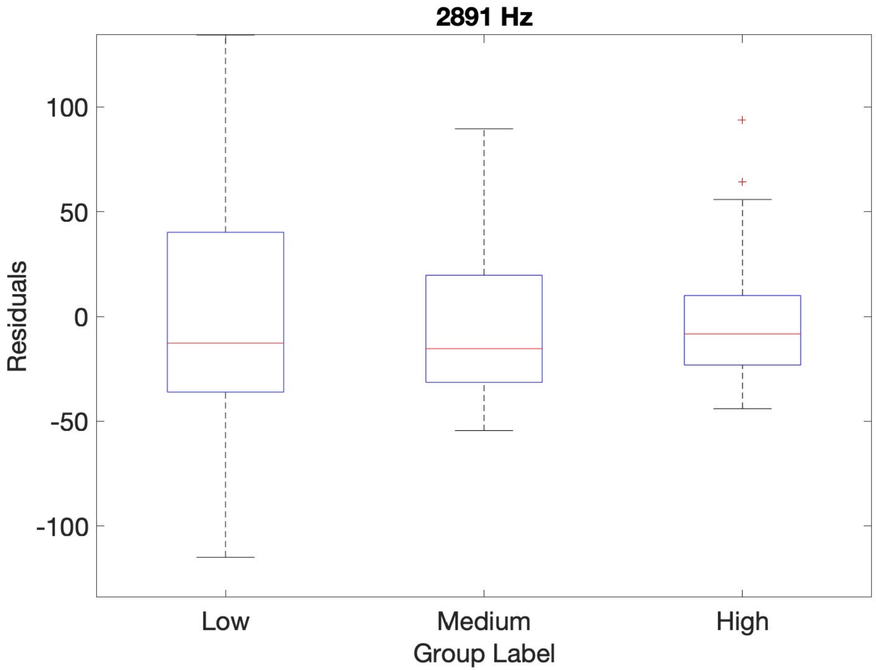
Method	Simulation	Approximation 1	Approximation 2	Random Permutation
P-Value	Not Feasible	2.1e-4	1.9e-4	0

Remediation: heteroscedastic FANOVA, weighted least squares

Distributional assumptions not necessarily satisfied



f = 2891 Hz

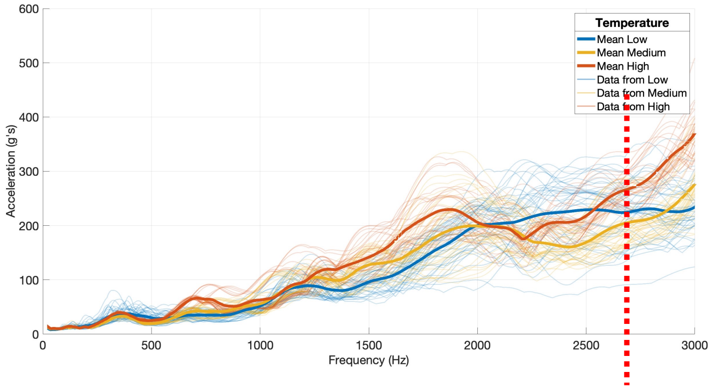


Quantiles not equal across groups

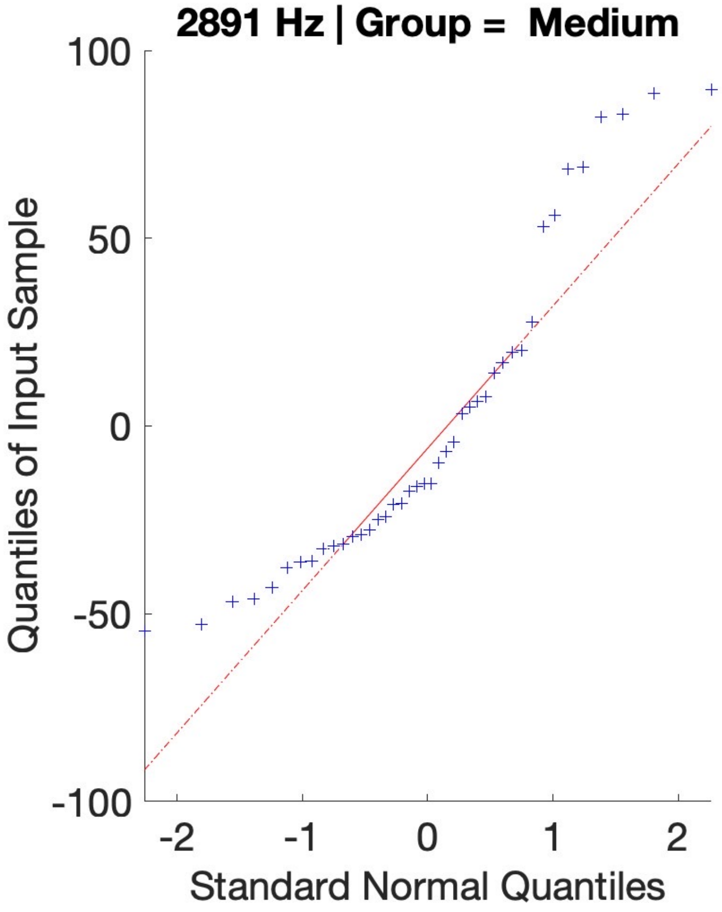
Often skewed quantiles (not pictured)

Remediation: transformation, generalized least squares

Distributional assumptions not necessarily satisfied



f = 2891 Hz



QQ Plots confirm non-Gaussian

Shapiro-Wilkes tests agree with visual guesses

Remediation: transformation

Motivation ✓

Model ✓

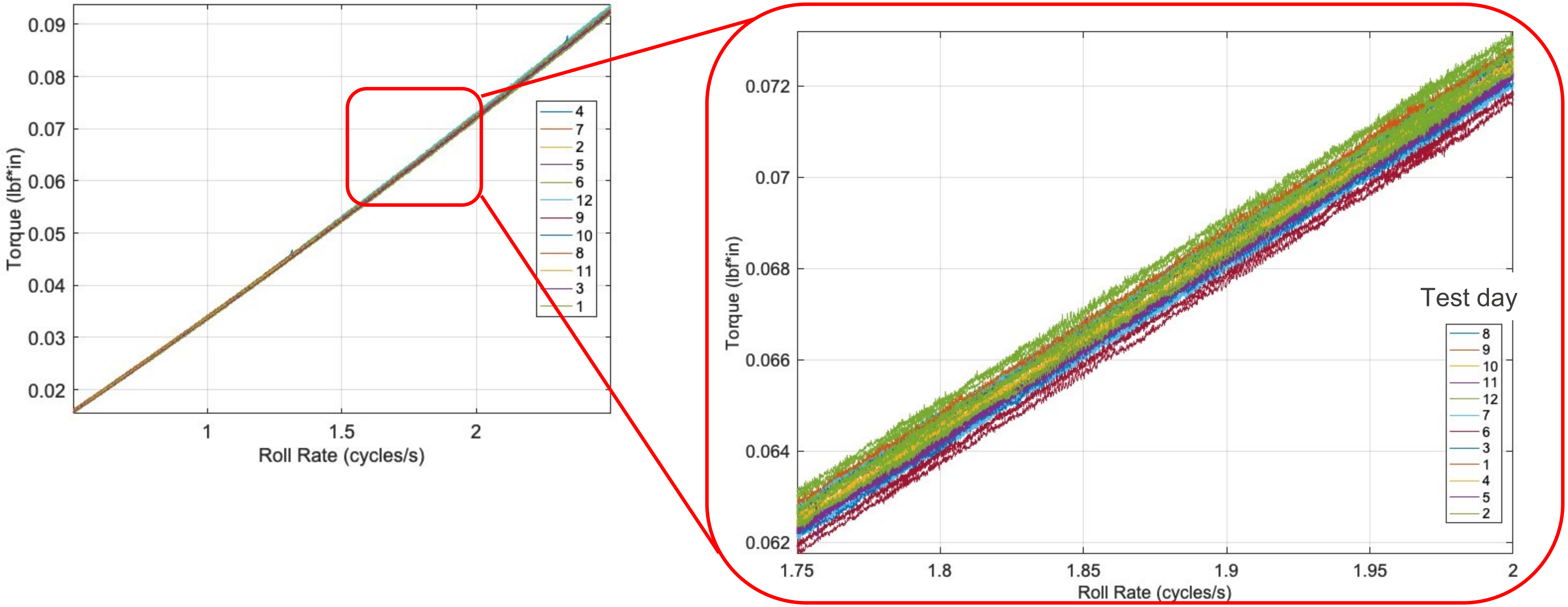
Properties of model ✓

Testing hypotheses ✓

Assumptions revisited ✓

Generalization to beyond

Extension to general linear models with air bearing data

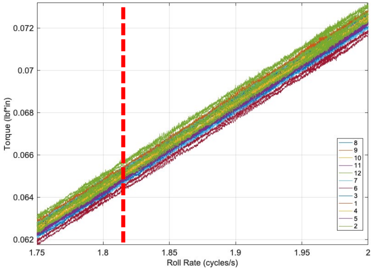


Torque decay $y_j(t), j = 1 \ldots 44$ curves exhibit variability

Collected covariate information: test day/run, environment (3), system (4)

Detection/quantification of important covariates

$$y(t) = \beta_0 + \beta_1 x_{day} + \beta_2 x_{run\#} + \beta_3 x_{temperature} + \beta_4 x_{pressure} + \beta_5 x_{humidity} + \beta_6 x_{cylindrical} + \beta_7 x_{inlet} + \beta_8 x_{spherical} + \beta_9 x_{system}$$

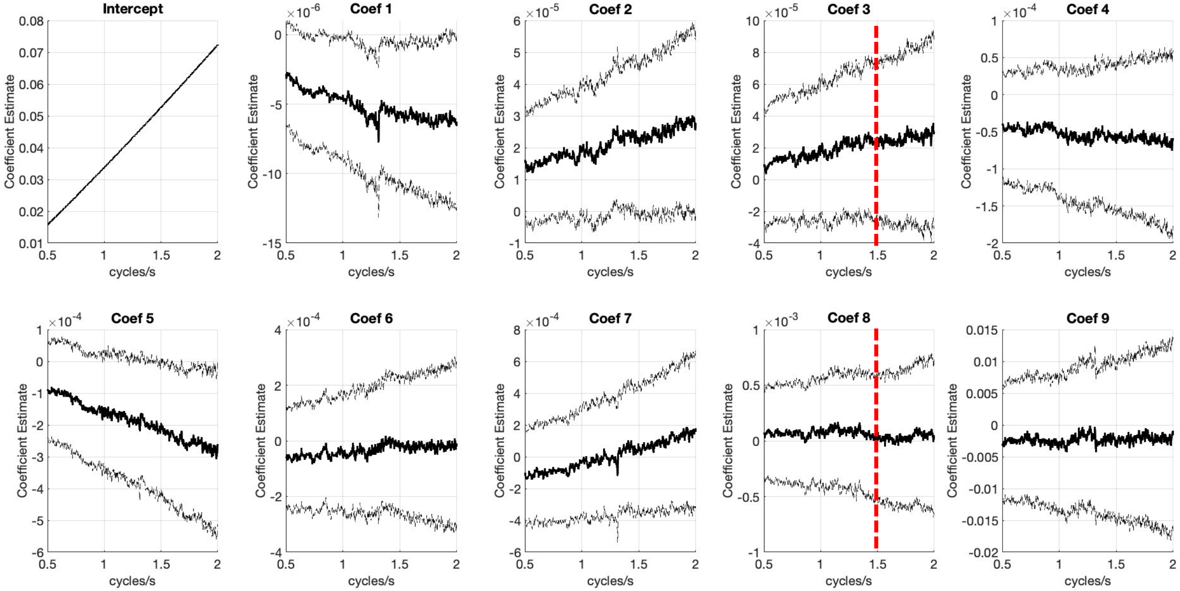


Smoothing and splining for domain homogeneity

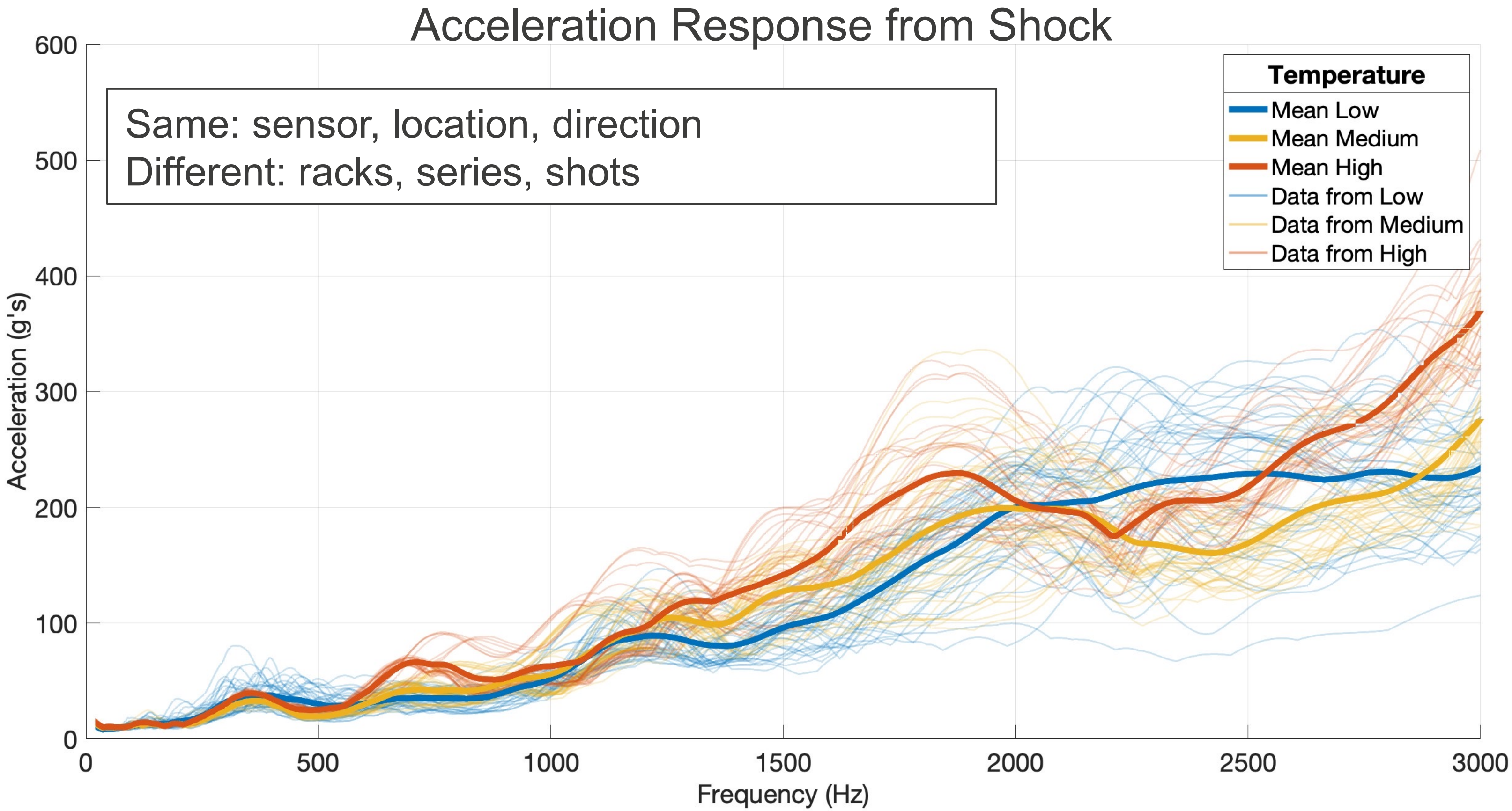
Used principal component regression

Note degrees of freedom dwindle quickly for n = 44

Design of experiments



Is there a significant difference in the mean response between groups? YES (probably)



Same: sensor, location, direction
Different: racks, series, shots

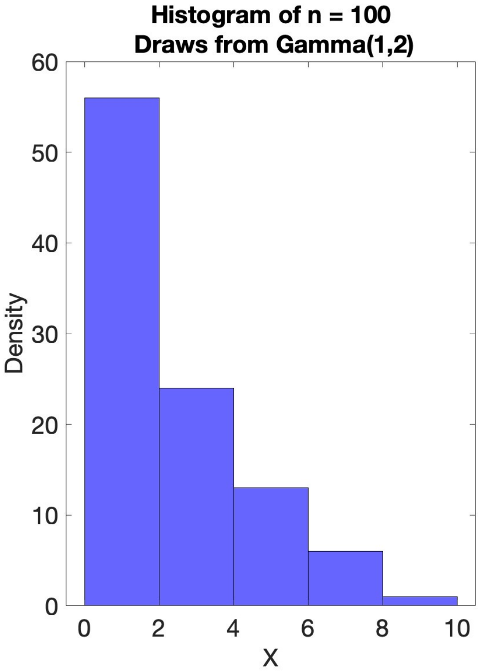
Fin

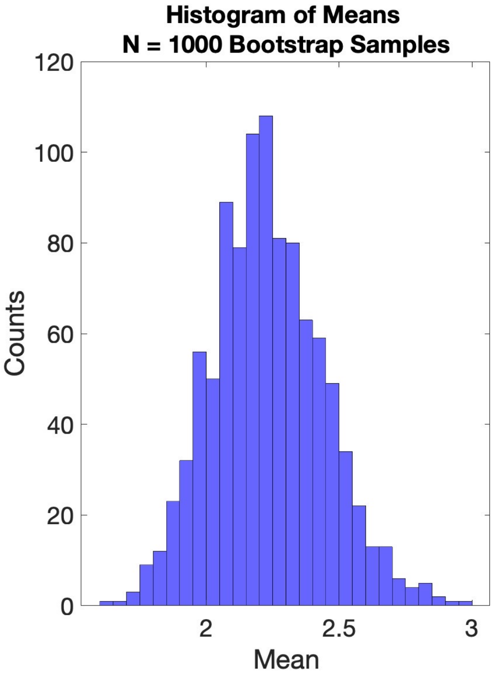
Public URLS for pictures:

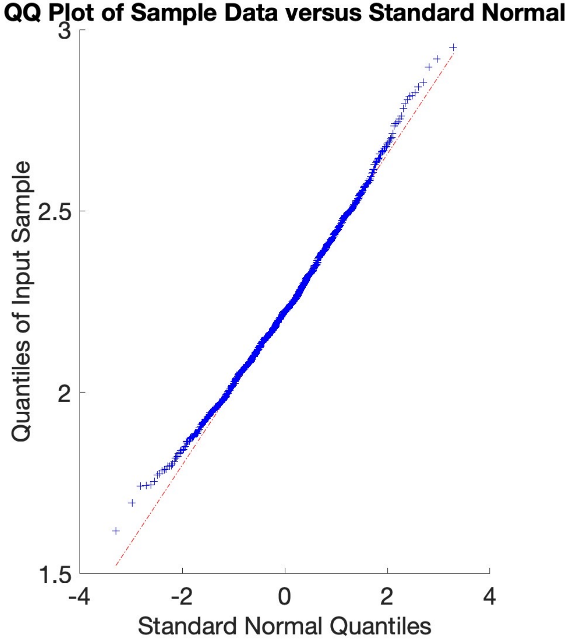
Bootstrapping and Central Limit Theorem

Option 1: Assume the functions are Gaussian
Option 2: Assume we have a large enough sample, so that the mean is approximately Gaussian distributed
Option 3: If n is small, bootstrap the data to exploit converging distribution of mean to a Gaussian process

CLT: If $y_i \sim (\mu, \sigma^2)$ and $\sigma^2 < \infty$, then $\bar{y} \rightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$ as $n \rightarrow \infty$


Histogram of n = 100 Draws from Gamma(1,2)


Histogram of Means N = 1000 Bootstrap Samples


QQ Plot of Sample Data versus Standard Normal

Bootstrapping assumes the sample is representative of the population

Contrasts are linear adaptations of hypotheses

Imagine daisy example: $y \sim N(\eta, \gamma)$, then $cy \sim N(c\eta, c^2\gamma)$

Stack main effects in a matrix: $\eta(t) = \begin{matrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{matrix}$ and $\hat{\eta}(t) = \begin{matrix} \bar{y}_{1.} \\ \bar{y}_{2.} \\ \bar{y}_{3.} \end{matrix}$

And $\hat{\eta}(t) \sim GP_k(\eta, \gamma D)$ where $D = diag(\frac{1}{n_1}, \frac{1}{n_2}, \dots, \frac{1}{n_k})$

Contrasts are linear adaptations of hypotheses

We want to test null hypothesis of $\eta_1 = \eta_3$ (rewritten as $\eta_1 - \eta_3 = 0$)

Allow $\mathrm{C} = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}$, then $\mathrm{C} * \eta(t) = 0 = c$ is our linear hypothesis

Then, under the null, $C\hat{\eta}(t) \sim GP_k(c, \gamma CDC^T)$

and $SSH(t) = (C\hat{\eta}(t) - c)^T (CDC^T)^{-1} (C\hat{\eta}(t) - c)$

For the pairwise hypothesis above, this simplifies to

$$SSH(t) = \frac{n_1 n_3}{n_1 + n_3} (\bar{y}_{1.} - \bar{y}_{3.})^2$$

No data is left on the table with contrasts

However, while SSH changes for a contrast, the SSE remains the same.

$$SSE(t) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}(t) - \overline{y_{i.}}(t)\right)^2 = (n-k)\hat{\gamma}(t,t)$$

Test statistic is $\mathrm{F}^* = \dfrac{\frac{\int SSH(t)dt}{k-q}}{\frac{\int SSE(t)dt}{n-k}}$

where $q = rank(C)$ (contrast matrix)

Denominator uses all the data to inform the covariance estimate

What type of distribution does this follow?

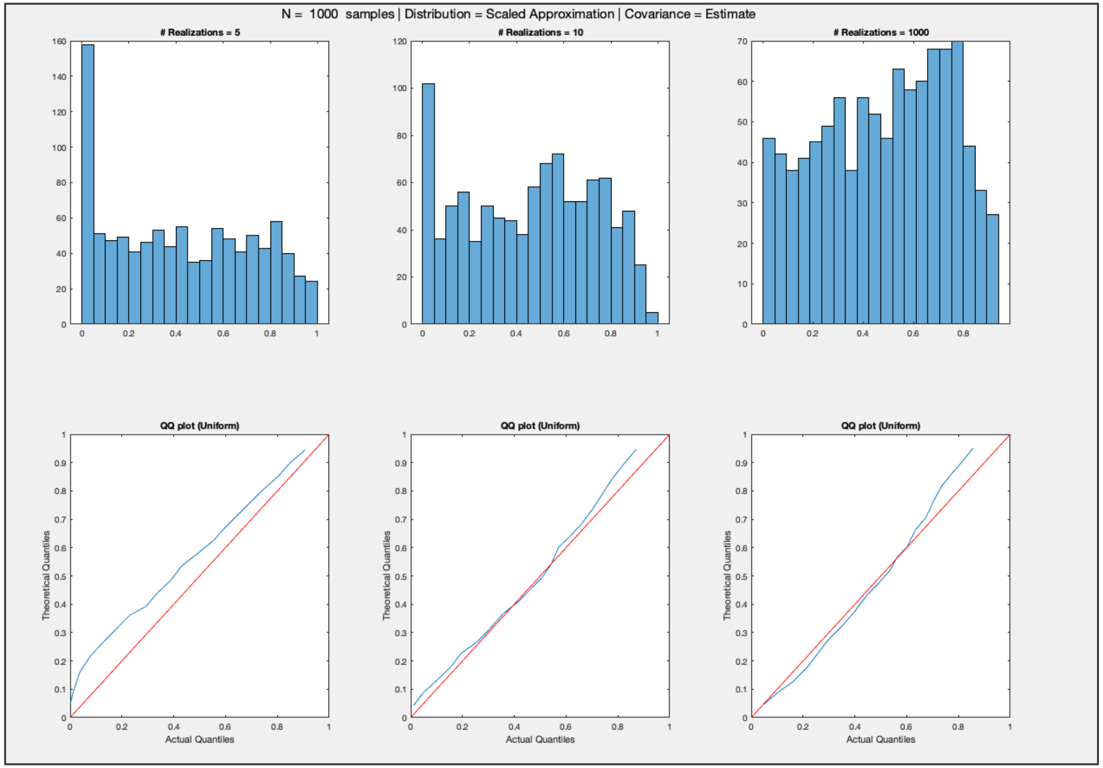
Previous contributions, future contributions, and lessons learned this summer

- Contributions this summer
  - Analysis of distributional assumptions
  - Contrasts implementation
  - Checking, validating, and generally improving f-anova package
  - Equal covariance functions
  - System import
  - **Central** repository for import – continue into future data phases – about 7 pages describing useful tools, functions, methods, and more relating to the import process
- Lessons learned
  - The wide variety of projects and functions our team performs and how Echo can unite them
  - More standard method for getting new employees up to par with respect to Echo data processing and analysis flows
  - Developed statistical experience in the realm of functional data
  - General refresher on dynamics

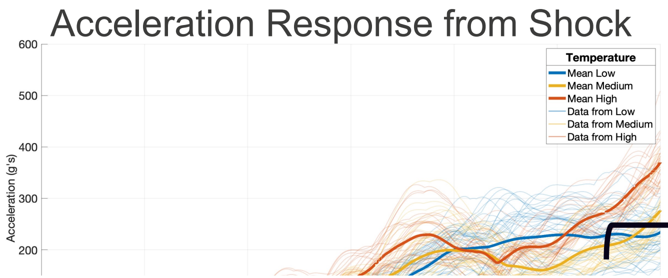
Exploring how the p-values are distributed

P-values not distributed uniformly for approximation methods

Should use simulation methods even though it takes more time



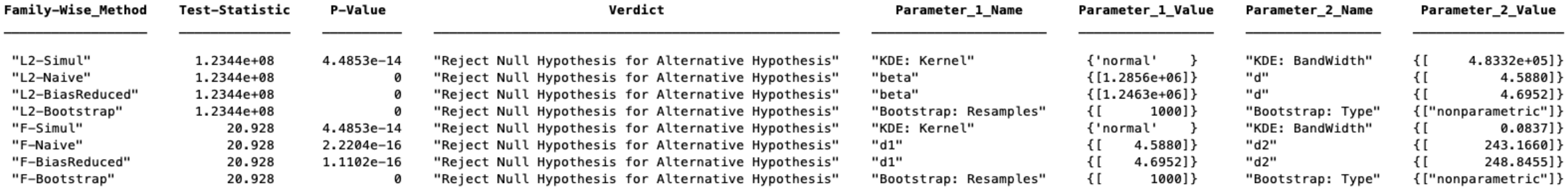
All pairwise results conclude temperature matters



Acceleration Response from Shock

Distribution method

Hypothesis	L2-Simulation	L2-Naïve	L2-Bias Reduced	L2-Bootstrap	F-Simulation	F-Naïve	F-Bias Reduced	F-Bootstrap
Low vs Med.	5e-12	6e-7	4e-7	1e-4	2e-5	1e-6	1e-6	0
Low vs High	2e-14	1e-14	5e-15	0	2e-14	5e-13	3e-13	0
Med. vs High	2e-14	5e-12	3e-12	0	2e-14	6e-11	4e-11	0



```
Family-Wise_Method    Test-Statistic     P-Value                            Verdict                             Parameter_1_Name       Parameter_1_Value    Parameter_2_Name      Parameter_2_Value
__________________    ______________    __________    __________________________________________________    ______________________    _________________    __________________    ___________________

 "L2-Simul"            1.2344e+08       4.4853e-14    "Reject Null Hypothesis for Alternative Hypothesis"    "KDE: Kernel"             {'normal'    }      "KDE: BandWidth"      {[     4.8332e+05]}
 "L2-Naive"            1.2344e+08                0    "Reject Null Hypothesis for Alternative Hypothesis"    "beta"                    {[1.2856e+06]}      "d"                   {[         4.5880]}
 "L2-BiasReduced"      1.2344e+08                0    "Reject Null Hypothesis for Alternative Hypothesis"    "beta"                    {[1.2463e+06]}      "d"                   {[         4.6952]}
 "L2-Bootstrap"        1.2344e+08                0    "Reject Null Hypothesis for Alternative Hypothesis"    "Bootstrap: Resamples"    {[        1000]}      "Bootstrap: Type"     {["nonparametric"]}
 "F-Simul"                 20.928       4.4853e-14    "Reject Null Hypothesis for Alternative Hypothesis"    "KDE: Kernel"             {'normal'    }      "KDE: BandWidth"      {[         0.0837]}
 "F-Naive"                 20.928       2.2204e-16    "Reject Null Hypothesis for Alternative Hypothesis"    "d1"                      {[    4.5880]}      "d2"                  {[       243.1660]}
 "F-BiasReduced"           20.928       1.1102e-16    "Reject Null Hypothesis for Alternative Hypothesis"    "d1"                      {[    4.6952]}      "d2"                  {[       248.8455]}
 "F-Bootstrap"             20.928                0    "Reject Null Hypothesis for Alternative Hypothesis"    "Bootstrap: Resamples"    {[        1000]}      "Bootstrap: Type"     {["nonparametric"]}
```