CSE 599K Empirical Research Methods

Winter 2025

SE meets Science

Today

- Recap: analysis validity
- SE principles for rigorous Science
- Two example studies

Analysis validity

Analysis validity: open questions

External validity

- Does the experiment generalize (to larger population, other subjects, etc.)?
- How representative is the sample?

Internal validity

- Does the experiment isolate the variable(s) of interest?
- Does the experiment control for confounders and unwanted effects?

Construct validity

- Does the experiment measure what it claims to measure?
- Do the proxy measures and tools adequately measure the concept of interest?

(Statistical) conclusion validity

- Are the conclusions valid based on the chosen statistical test and sample size?
- Are the conclusions valid based on the observed significance (p value)?

SE principles for rigorous science

Science to practice is not a one-way street!



Let's improve **scientific rigor** with **SE principles** and **best practices!**

Design reviews



Design reviews are common in practice.



Embrace and value **pre-registrations**.



RFCs and public discussions (e.g., GH) provide valuable context.



Public (open) reviews should be a no-brainer!

Hark no more: On the preregistration of chi experiments, Cockburn et al., CHI 2018 *https://openreview.net/*

Quality assurance



Modern code review is incremental (not holistic).



Move to pre-acceptance artifact evaluations.



Software testing is the most common QA approach in practice.



Require evidence for artifact testing.

Expectations, outcomes, and challenges of modern code review, Bacchelli and Bird, ICSE 2013 *Modern code review: a case study at Google,* Sadowski et al., ICSE 2018

Process



Merge conflicts (branches) are resolved by branch authors.

Expect resolution (knowledge) of conflicting results.



Don't expect others to resolve your merge conflict!

Process



Merge conflicts (branches) are resolved by branch authors.

Expect resolution (knowledge) of conflicting results.



No premature optimizations.



Focus on design validity before scrutinizing artifacts.

Science is a collaborative effort!



Software Engineering is a **collaborative effort**. We should view science the same way!

Science as Amateur Software Development



- 1. How can software engineering principles improve the rigor of data analyses?
- 2. Are these principles equally applicable to computational notebooks?
- 3. Describe three specific quality control mechanisms.
- 4. McElreath attributes a significant number of incorrect (scientific) studies to "sloth". What are the specific issues he is calling out, and what solutions does he propose?
- 5. Provide an argument for why or why not general-purpose programming languages such as Python are an adequate choice for data analysis.

Two example studies

An example study: design

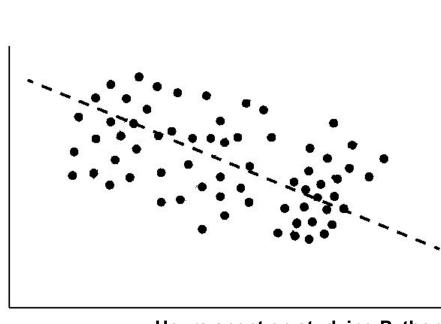
Goal:

Studying the **relationship** between **time spent** on **studying** Python and **success rate** in completing coding assignments.

Methodology:

- ~100 participants are randomly selected in front of CSE.
- Each participant is given a high-level overview of the study.
- Each participant decides on how long to study before attempting to solve any coding assignment.
- Each participant solves as many coding assignments as possible in one hour (after studying).

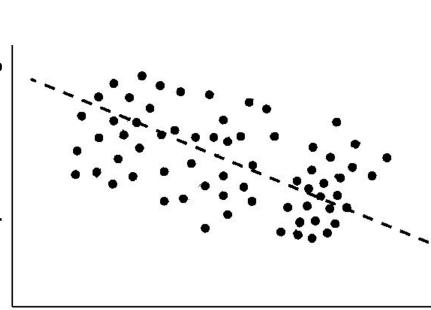
An example study: conclusions



Hours spent on studying Python

Conclusion: Spending more time on learning Python makes you a worse Python programmer.

An example study: conclusions

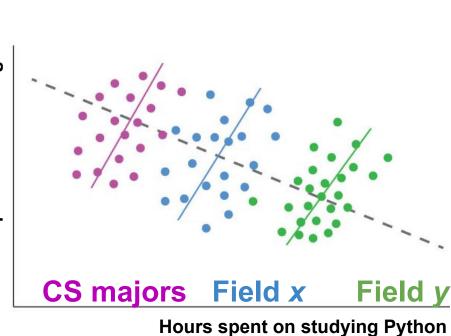


What may cause this result?

Hours spent on studying Python

Conclusion: Spending more time on learning Python makes you a worse Python programmer.

An example study: Simpson's paradox



Where did this study fail?

This phenomenon is called: **Simpson's paradox.**

Another example study



http://www.prefrontal.org/files/posters/Bennett-Salmon-2009.pdf

Another example study: design



Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction Craig M. Bennet!, Abgal A. Baird, Michael B. Miller!, and George L. Wolford ³ 'hydrolog Opstemic ruleway of calmas bairs danse. Sam Bianc, A. O'spectrat of Psychingy, Vasar College, Poghsegen, NY: 'begennet d'hydrologia & Bian Gennes, Datmodh College, Henore, MI	
INTRODUCTION With the extreme dimensionality of functional neuroimaging data comes	GLM RESULTS
extreme risk for faise positives. Across the 130,000 words in a typical MRI volume the probability of a faile positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance property.	
METHODS	2.5 (-value
Subject, One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.	A t-contrast was used to test for regions with significant BOLD signal change
Task. The task administered to the salmon involved completing an open-ended metalizing task. The salmon was shown a series of photographic depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.	during the photo condition compared to rest. The parameters for this comparison were $r(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold. Several active voxels were discovered in a cluster located within the salmon's
Design, Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.	brain cavity (Figure 1, see above). The size of this cluster was 81 mm ³ with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed.
Pergrozessing. Image processing was completed using SPML. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the MRI timesentes, coregistration of the data to a T ₁ -weighted mattonical image, and 8 mm full-width at half-maximum (PWHM) Gaussian smoothing.	Out of a search volume of 8064 voxels a total of 16 voxels were significant. Identical <i>t</i> -contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds (or 0.25).
<u>Analysis</u> . Vorcelwise statistics on the summon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function corvolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No subcorrelation correction was	VOXELWISE VARIABILITY
applied. <u>Youd Sciencian</u> . Two methods were used for the correction of multiple comparisons in the (MRI routh). The first method controlled the overall failer discovery rate (PRN) and was howed on a method defined by Beginmini and Hochberg (1995). The second method controlled the overall familywise error rate (PWFR) through the use of Claussian methods that the two; This was done using algorithms originally devised by Friston et al. (1994).	a) b) c)
DISCUSSION	To examine the spatial configuration of false positives we completed a variability analysis of the IMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.
Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the EPI trunseries may yield serious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely witable in all marking TMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum clutter sizes ($k > 3$) is an infective control for multiple	We observed chattering of highly variable voxels into groups not areas of high voxel signal intensity. Figure 2a shows the mean EP1 image for all 140 Figure 2a shows thereholded standard deviation values overlaid onto a high- resolution T ₁ -weighted image.
comparisons. We further argue that the vast majority of MRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.	To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant
REFERENCES	positive correlation between the mean voxel value and its variability over

REFERENCES

Benjamini Y and Hochberg Y (1995). Controlling the fulse discovery rate: a practical and p approach to multiple testing. *Journal of the Royal Statistical Society: Series 8*, 57:289-300.

on KJ, Wonsley KJ, Frackowiak RSJ, Mazzietta JC, and Evans AC. (1994). Assessi ficance of focal activations using their spatial extent. Hawaw Brain Mapping, 1:214-220

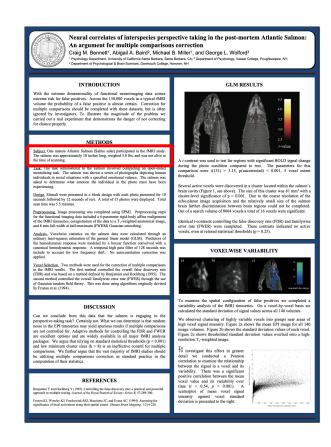
deviation is presented to the right

time (r = 0.54, p < 0.001). A

scatterplot of mean voxel signal ntensity against voxel standard

Another example study: design

Subject: One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

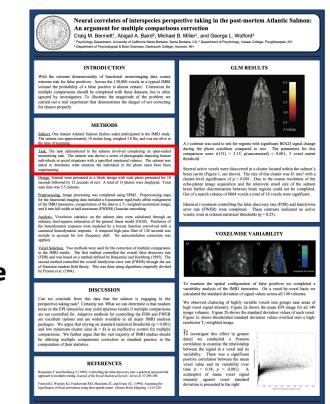


Another example study: design

Subject: One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task: [...] **open-ended mentalizing task**. The salmon was **shown a series of photographs** depicting **human individuals in social situations** with a specified emotional valence. **The salmon was asked** to **determine** what **emotion** the **individual in the photo** must have been **experiencing**.



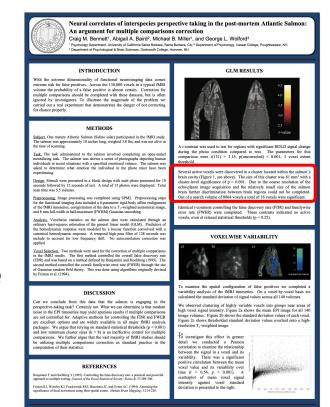


Another example study: conclusions

Subject: One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task: [...] **open-ended mentalizing task**. The salmon was **shown a series of photographs** depicting **human individuals in social situations** with a specified emotional valence. **The salmon was asked** to **determine** what **emotion** the **individual in the photo** must have been **experiencing**.

Results: Several active voxels were discovered [...] Out of a search volume of 8064 voxels a total of **16 voxels** were significant.



Another example study: conclusions

Interpretation of pure noise

- Noisy data source
- Multiple hypotheses tested on the same data
- An argument for multiple comparisons correction
- Analysis grounded in a conceptual model?
- Clear operationalization (implementation)?
- Implementation consistent with the model?
- **Proper** use of **statistical methods**?
- Data interpreted in context of prior knowledge?
- Explored and validated **alternative hypotheses**?

Where did this study fail (on purpose)?

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salt An argument for multiple comparisons correction Craig M. Bennett', Abigail A. Baird', Michael B. Millert', and George L. Wolford ¹ 'Physic potentic Linear of California Minista Santa Mathan, A. Vipaurinet of Psychiagra Materia, No. 'Paperheart of Psychiagra & Boin Garrow, Demonit Celaga, Innove, Ni		
INTRODUCTION	GLM RESULTS	
With the extreme idmensionality of functional neuroimaging data comes termer nick for faith positives. Across the 10,000 vocatio in a poychal (MB1 volume the probability of a fulse positive is almost certain. Cerrection for imported by investigators. To Illustrate the magnitude of the problem we certife out a real experiment that demonstrates the danger of not correcting for chance properly.		
METHODS Salgigg, One mature Atlantic Salmon (Salmo salar) participated in the MRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of example. Table: The task administered to the salmon involved completing an oper-model association into 4.1. Inches use with the same same of abbrarding duringing lamitics.	25 7-sulte A-contrast was used to test for regions with significant BOLD signa during the photo condition compared to rest. The parameters comparison were (131) > 3.15, primererted (> 0.000, 3) yours	
alizing task. The salmon was shown a series of photographs depicting human dialism is not aliastations with a specific themotonal winters. The salmon was 10 to determine what emotion the individual in the photo must have been mixing. gg. Simmily were presented in a block design with each photo presented for 10 distributions of the second	threshold. Several active voxels were discovered in a cluster located within th brain activ() (Figure 1, see above). The size of this cluster was 81 r cluster-level significance of $p = 0.001$. Due to the coarse resolu- clo-poltant image acquisition and the relatively small size of t brain further discrimination between brain regions could not be 0.01 of a search volume of 864 to volume 1 atol 161 for souch were sign	
Perprocessing, Image processing was completed using SPM2. Proprocessing steps for the functional imaging data included a 6-parameter right-body affine realignment of the MRI timeseries, coregistration of the data to a T,-weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing. Analysis, Vorchwise statistics on the salmon data were calculated through an	Identical <i>i</i> -contrasts controlling the false discovery rate (FDR) and i error rate (FWER) were completed. These contrasts indicated voxels, even at relaxed statistical thresholds (p = 0.25).	
Pathyzia, Vecewie a sistence on the general inter model (GLM). Predictors of onlinary least-squares estimation of the general inter model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was include to account for low frequency drift. No autocorrelation correction was asentiated.	VOXELWISE VARIABILITY	
Stead Education. Two methods were used for the corrections of multiple comparisons in the PMI reasons. The first method controlled the overall fields discovery rate (PDR) and was based on a method defined by Benjamini and Hechberg (1995). The second method controlled the overall finallysisies error rate (PVER) through the use of Ginessian random field theory. This was done using algorithms originally devised by Fraston et al. (1954).	a) b) c)	
DISCUSSION	To examine the spatial configuration of false positives we co variability analysis of the fMRI timeseries. On a voxel-by-voxe calculated the standard deviation of signal values across all 140 volu	
Can we conclude from this data that the nations is engaging in the prepercive-taking that C-trainity out Waits we can determine its that random moise in the EPI timeseries may pited parious results if multiple comparisons are concruted left. And apprive methods for controlling the TDM and FWER are exceeding objects and are widely available in all major MRI analysis packages. We argue that relying an standard statistical thresholds (σ 40001) and low minimum cluster sizes ($\kappa > 8$) is an ineffective control for multiple comparisons. We further argue that the variantity of MRI analysis due to the statistical statistical thresholds (σ 40001) and low minimum cluster sizes ($\kappa > 8$) is an ineffective control for multiple to this statistical statistical statistical thresholds be utilizing multiple comparisons correction as standard practice in the comparisons. We for analysis.	cacutated the sindual deviation of signal values across all 140 volus We observed classifier of highly values loves in min groups are high voted signal intensity. Figure 2a shows the near PFI maps i mage volumes. Figure 2a shows the sindual deviation values overhald or resolution T, weighted mage.	
REFERENCES Breijanzini Y and Heckberg Y (1995). Centrolling the failer discovery rate: a practical and powerful	content in significant positive correlation between the mean voce value and its variability over time $(r = 0.54, p < 0.001)$. A satterplot of mean voxel signal intensity against voxel signal	

Another example study: conclusions

Interpretation of pure noise

- Noisy data source
- Multiple hypotheses tested on the same data
- An argument for multiple comparisons correction
- Analysis grounded in a **conceptual model?**
- Clear operationalization (implementation)?
- Implementation consistent with the model?
- **Proper** use of **statistical methods**?
- Data interpreted in context of prior knowledge?
- Explored and validated **alternative hypotheses**?

Valid data analysis goes well beyond implementation correctness.

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salt An argument for multiple comparisons correction Craig M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Bennett', Abigai A. Baird', Michael B. Millert', and George L. Wolford ³ ¹ M. Baird', M. Baird', Millert', Baird', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Baird', Millert', Baird', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Baird', Millert', Baird', Baird', Millert', Baird', Millert', Baird', Baird', Millert', Baird', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Baird', Millert', Baird', Baird', Millert', Baird', Millert', Baird', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Millert', Baird', Ba	
INTRODUCTION	GLM RESULTS
With the extense dimensionality of functional neuroimaging data censes termmer side for flat positives. Across the 10/000 vocatio as a popical MBI volume the probability of a fulse positive is almost centain. Cerrection for ignored by investigators. To flatarate the magnitude of the problem we cerriced out a real experiment that demonstrates the danger of not correcting for chance properly.	
METHODS Subject. One mature Atlantic Salmon (Salmo sular) participated in the DMRI study. The salmone was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scaming.	25 F-value
Task. The task administered to the salmon involved completing an open ended mentalizing task. The salmon was shown a series of photographs depitting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencine.	during the photo condition compared to rest. The parameter comparison were r(131) > 3.15, p(uncorrected) < 0.001, 3 vo threshold. Several active voxels were discovered in a cluster located within th
Design, Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.	brain cavity (Figure 1, see above). The size of this cluster was 81 r cluster-level significance of p = 0.001. Due to the coarse resolut echo-planar image acquisition and the relatively small size of t brain further discrimination between brain regions could not be
Preprocessing. Image processing was completed using SPM2. Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the MRI timeseries, coregistration of the data to a T ₁ -weighted anatomical image, and 8 mm full-width at half-maximum (FWHM) Gaussian smoothing.	Out of a search volume of 8064 voxels a total of 16 voxels were sign Identical <i>t</i> -contrasts controlling the false discovery rate (FDR) and error rate (FWER) were completed. These contrasts indicated
<u>Analysis</u> . Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model (GLM). Predictors of the hemodynamic response were modeled by a boxcar function convolved with a canonical hemodynamic response. A temporal high pass filter of 128 seconds was	voxels, even at relaxed statistical thresholds (p = 0.25).
include to account for low frequency drift. No autocorrelation correction was applied.	VOXELWISE VARIABILITY
<u>Yatel Selection</u> . Two methods were used for the correction of multiple comparisons in the MRR results. The first method controlled the overall false discovery rate (TPR) and we have also on method defined by Beinjamini and Idenberg (1995). The second method controlled the overall finallywise error rate (FWIR) through the use of Gaussian random field theory. This was done using algorithms originally devised by Friston et al. (1994).	
DISCUSSION	To examine the spatial configuration of false positives we con- variability analysis of the fMRI timeseries. On a voxel-by-voxel calculated the standard deviation of signal values across all 140 volu
Can we conclude from this data that the almon is sequeging in the propertive-taking and Creatingh set. What we can determine is that random noise in the EPT timesceles may yield partons result if multiple comparisons are accommobiled for Adaptive matcheods for constraining, the FDR and PWER are according to the straining of the transformation of the straining	We observed charating of highly variable vands into groups not high voted sign intensiv, Figure 2a bases the standard deviation values of Figure 2a down sheeholded standard deviation values overlaid on resolution T, «veighted image. For investigate this effect in generation of the standard standard in the detail we conducted in Person between the signal in a word and its variability. There was a significant
REFERENCES	positive correlation between the mean voxel value and its variability over
Benjanizi Y and Hechberg Y (1985). Controlling the fulse discovery nate: a practical and powerful approach to multiple testing. <i>Journal of the Royal Statistical Society: Series 8</i> , 57:289-309. Printer KJ, Wonley KJ, Prackowski RSB, Mazzietta JC, and Ewans AC. (1994). Assessing the	time ($\mathbf{r} = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.