3 Characteristics of Good Dissertation Research

- In your dissertation you have to justify that your research contains new contributions and motivate the importance of your research (i) in statistics as a discipline or (ii) in terms of practical applications (if any).
- As a PhD student you want to find a good research topic in statistics for a dissertation. Although there is no guaranteed approach to finding a good research topic, the following are characteristics of good research topics: **originality**, **relevance**, **manageability**, and **clarity**.

3.1 Originality and relevance in statistics.

- The topic must lead to original research in statistics. That is, you must present material that no one else has ever seen. This may include significant modifications and improvements to existing theory or methods, developing algorithms to solve a computational problem, developing new statistical techniques, and establishing new theorems with rigorous proofs.
- The topic should also be relevant research in statistics. That is, the research you are presenting should be of interest to academia, and hopefully to practitioners of statistics in the research area. For example,
  - Research in response surface methodology will be relevant to engineers and industrial statisticians.
  - Research in sampling will be relevant to statisticians who design, collect, and analyze data from surveys.
  - Research in spatial statistics will be relevant to ecologists, geologists, and other environmental scientists.
- After you have a research topic, you need to think about the dissertation in terms of original unsolved problems. These problems determine what you plan to discover about the research topic.

My personal experiences

- The following are examples of the originality and relevance of the research from three dissertations.

1. Regarding the dissertation of Prof. John Borkowski
   - As an improvement to the Taguchi system of designs, this dissertation presented and evaluated a new class of response surface designs (called minimum aberration mixed resolution designs) which are based on a single factor array.
     - *Originality*: Creating a new class of designs.
     - *Relevance*: Improvements to the Taguchi system of designs.

- The mixed resolution designs were extended to another new class of response surface designs (composite mixed resolution designs) that allow estimation of quadratic model effects. This will then lead to improvements in the design of experiments for achieving a robust process.
  - *Originality*: Extension to another new class of designs.
  - *Relevance*: Improved designs for a robust process.
- In the evaluation of mixed resolution designs, the problem of finding globally optimum designs for achieving a robustness process was addressed and solved. $D-$ and $G-$optimal designs for the mixed resolution model were found and their optimality properties were studied. The $D-$ and $G-$efficiencies of the mixed resolution designs were then calculated.
  - *Originality*: Studying optimality properties of designs with respect to the mixed resolution model.
  - *Relevance*: Finding theoretically $D-$ and G-optimal designs.

2. Regarding the dissertation of Dr. Philip Turk (a former advisee who is now an Assistant Professor at West Virginia University)

   - Multiple factors known to influence the efficiency of adaptive cluster sampling (ACS) were studied using a designed computer experiment and application of response surface methods to computer-simulated data. This approach allows significant interaction and quadratic effects to be studies in contrast to ignoring them as had been done in the literature.
     - *Originality*: Using a computer experiment to study factors related to ACS.
     - *Relevance*: Understanding interaction and quadratic effects that had not been studied.
   - The efficiency for the Horvitz-Thompson estimator of the population total τ was studied relative to simple random sampling without replacement. Recommendations were proposed that can be applied in practice.
     - *Originality*: Studying the relationship between efficiency and ACS study factors. Determining how and which study factors affect ACS efficiencies.
     - *Relevance*: Providing recommendations for practical use.
   - Current research is focused on comparing different simulation approaches to generate populations following a Neyman-Scott point process. It was shown that the validity of conclusions regarding ACS will depend on the simulation approach used to generate the population data.
     - *Originality*: Investigating different simulation approaches.
     - *Relevance*: Showing that conclusions depend on the choice of simulation method.

3. Regarding the dissertation of Dr. Julia Sharp (a former advisee who is now an Assistant Professor at Clemson University)

   - Her dissertation research was centered on applications of statistics in the field of proteomics which is the study of protein compositions in biological substances. Likelihood ratio tests and exact tests were developed to analyze data from proteomics laboratories.
     - *Originality*: Developing statistical methods to analyze proteomics data.
     - *Relevance*: The identification of proteins as well as establishing the presence or absence of certain proteins in biological substances has far-reaching applications in the biological sciences, especially in medicine.

3.2 Manageability with respect to a time frame.

- You must be able to do all of the work required within the time frame set by the Ph.D. program at your university. This is obvious, but most students do not think enough about the details. Suppose, for example:
  - You are scheduled to present your Ph.D. proposal on November 1. *When should your committee receive a copy of the proposal?* Suppose they need 10 days. Then you have to have it ready to submit on October 22. So can you just type the proposal and hand it to the committee on October 22? No.
  - You should not give the proposal to your committee before your advisor reviews it. Therefore, you will need to answer the question "*When should my advisor see the proposal?*". The answer to this will depend on other questions such as:
    (i) *How much time will my advisor need to review it?* Suppose the advisor needs 10 days.
    (ii) *How much time will I need to make the changes recommended by the advisor?* Suppose you need 5 days.
    (iii) After making the revisions, *does my advisor want to read the proposal again?* Suppose the answer is yes. This will take another 4 days.
    (iv) *How much time will I need to make the changes recommended by the advisor?* Suppose you need 2 days.
  - In total you will need to have the first draft of the proposal to you your advisor <u>at least</u> $10+10+5+4+2=31$ days before the proposal defense. Therefore, you need to manage your time so you have a first draft of the proposal (at the latest) on October 1 for the November 1 presentation.

- The purpose of presenting this example was to highlight the importance of being realistic about the scope of dissertation research. <u>You must not procrastinate</u>. That is, "Do not put off until tomorrow what you can do today."

- There are many issues related to manageability of time. The most essential is "Can I complete all of the objectives in my proposal within the limited time and resources available to me?" This is something you can discuss with your advisor.

- For example, can you access all of required information (such as statistical literature, computer software, and data sets) to address the research questions?
- You must realize that even though you think a chapter of your dissertation or your proposal is acceptable, does not mean your advisor and your committee will agree with you. Expect to have to make at least two revisions (usually more) of your proposal and each chapter of the dissertation before it is acceptable to your advisor and committee.
- Eventually, you will stop. However, when you stop is not your decision. The advisor will let you know if your dissertation has adequately addressed all research items presented in your proposal.
- Therefore, when preparing for your a proposal defense or a dissertation defense, you will need to think about:
  - How much time do I really have?
  - Do I have a time schedule that I can follow faithfully?
  - Are my plans unrealistic with respect to deadlines?
- You must be realistic with respect to time management. Remember that deadlines will be strictly enforced, so manage your time wisely.

My personal experiences

- Although your professors may not want to talk about their experiences writing a dissertation, you can be certain they had the same issues to deal with as PhD students that you will also have to deal with. Many of those issues are related to time management.
- As I stated in Chapter 2, I had a structured schedule to work on my dissertation after I finished working at DuPont during the day and also to work on it on weekends.
- Every week I contacted or met with my advisor and informed him what I did during that week. If I had questions to ask, then I would ask him those questions at our meeting. I did not procrastinate.
- For many of you, much of your contact will be through email (unless your advisor is from TU or you study abroad).
- By keeping my advisor informed, I knew whether or not I was making progress. If I had little to tell him about what I did that week, I knew I needed to work harder.
- New research questions would arise as I worked on my the dissertation. I would ask myself "Should I include these in my dissertation or should I save them for future research?". If I was unsure, I would ask my advisor for his opinion.
- For example, in my dissertation I had to address the question "What is the maximum number of mixed resolution design factors I have to consider for my dissertation?". I suggested 17 because (i) 17 was already larger than the number of factors in practical applications and (ii) of the large amount of computing time required for ≥ 18 factors. My advisor agreed that 17 would be enough. (Note: That was in 1991. Today, I could easily consider ≥ 18 factors because of the incredible increases in computing speed.)

- As soon as you have a fixed date for you dissertation defense, you need to plan all of the details for the defense (like the earlier example in Section 3.2 on time management for a proposal). For example, in July 1991, I was offered an Assistant Professor position at Montana State University (MSU) in an ABD ("all but dissertation") status. This means, I would be hired before I finished my dissertation. To keep my Assistant Professor position, I would have to pass my defense and receive my Ph.D. by the end of my first year at MSU (June 1992).

- If I accepted the offer, there was a serious time management situation I would have to handle. Not only would I have to finish my dissertation, but I would also have to prepare lectures for courses that I never taught before. Maybe some of you will be in the same situation at a university in Thailand.

- Fortunately, when I was hired ABD, I had written all by the last chapter of the dissertation and had to make some minor revisions to previous chapters. It was because of good time management skills throughout my dissertation that I was confident that I could finish the dissertation while teaching new courses.

- So, I accepted the offer, finished writing my dissertation in November 1991 and scheduled my defense for January 1992. This gave my committee almost two months to read the dissertation, suggest revisions, and allow me time to make the suggested revisions.

- I passed my defense in January 1992, and, as you know, I am still at Montana State University (and occasionally at Thammasat University).

3.3 Clarity of the problem description.

- In Section 2.1, I reviewed ways for selecting your research area and then selecting a topic for the dissertation *within the research area*. There is one more level. After you have found your research topic, you now must find a *research problem* within the research topic. In summary:

  The problem is within the topic, and the topic is within the area.

- In your proposal and dissertation, you need to establish to your committee that you have clear objectives for your research in statistics. You want "clarity" with respect to your research problem. That is, you want the statement of your research problem to be as clear and as simple as possible.

- Remember: You proposed to create new knowledge in statistics and your committee and the academic community needs to understand exactly what you propose to do.

- Defining a clear research problem requires thought. An unclear research problem will also have unclear objectives, and unclear objectives will make it difficult to structure the research process.

- Initially, in your proposal, you may describe a clear and simple research problem in statistics. However, once you begin working on the research problem, you may find that it is more complex than you originally thought. This is common.

- If the research problem becomes more complex, it will be important that each component of the problem be defined clearly. Having a clear 'sub'-problem for each component will guide your overall research.

- Consider the following questions when considering clarity.

  - *Is my research problem too vague?* Determine if the objectives are clearly defined.
  - *Is my research problem possibly unanswerable?* Determine whether or not your research problem can be answered. If the problem lacks clarity, you may realize too late that you have spent a lot of time on a problem that may have no solution or cannot be solved within time constraints of your Ph.D. program.

- Clarity of presentation also applies to the chapters of the dissertation.

  - *Clarity in the Introduction*: The research problem is to be clearly defined. It should also include a clear summary of the structure in the dissertation.
  - *Clarity in the Literature Review*: For clarity, this chapter should be written in subsections with each subsection focused on reviewing previous work done on a particular topic area that is relevant to your research problem. You want to provide your committee with a clear summary of what prior research has been done and how it relates to your research. These subsections should appear in the Table of Contents.
  - *Clarity in the Methodology*: This chapter should clearly identify the statistical methodology used to address the proposed research problem, and justify why these methods are appropriate for the research.
  - *Clarity in the Results*: This chapter should clearly present the main results of your research and clearly address how well you answered your proposed research problem.
  - *Clarity in the Discussion*: This chapter must clearly relate the results to the content of your Literature Review. That is, discuss how your research contributes to the existing body of statistics research. You should clearly list any recommendations you want to make regarding applications of your results.

- To achieve clarity requires the careful use and further development of your communication skills, technical skills, and intellectual skills.

My personal experiences

- I recommend that you write a description of the research problem as early as possible. This requires you to focus on a specific problem. Give this to your advisor and receive feedback about its clarity. This will help you in the preparation of your proposal.

- You must realize that what may seem clear to you, often is not clear to your committee. In Section 3.4, I will present examples of the lack of clarity.

3.4 Clarity and writing style suggestions.

- Remember that you will be assessed on (i) the statistical research **and** (ii) how it is presented. Many PhD students want to ignore (ii) and just focus on (i). As I mentioned in Section 2.4, it may efficient to work on more than one thing at a time. I strongly recommend writing something each week and not save it until the end.

- The quality of your writing is incredibly important. What good is generating interesting research results if you cannot communicate and discuss the results in your writing?

- You have been exposed to many forms of technical writing such as academic textbooks, journal articles, and conference proceedings. There is a certain style to technical writing which is very different than these course notes (which is more personal and informal). **The writing style of the dissertation will be very formal.**

- Write clearly using concise and consistent notation, definitions and terminology. Be sure any proofs are clearly structured.

- Avoid long paragraphs.

- Check your writing for "redundancies" (e.g., saying the same thing multiple times in the same paragraph or the same page) and carelessness (e.g., misspelled words and being ungrammatical with respect to sentence structure).

- In technical writing, omit needless words and phrases. For example

  – "In the next paragraph, we will proceed to show that..." (Better: "It will now be shown that...")

  – Avoid unnecessary phrases such as "It can be seen clearly at this point in the dissertation that...".

- Writing clearly and correctly is very difficult. It takes practice to identify problems and then fix them. The challenge with writing is finding the right balance between *frugality* (saying too little) and *verbosity* (saying too much).

- The following examples contain both simple and complex writing problems.

Examples of Writing Problems

(1.) There are unnecessary words in the following sentence.

> In an attempt to offer a compromise between good projective properties of LHDs and a criterion in their articles, Park (1994) and Morris and Mitchell (1995) proposed optimal Latin hypercube designs.

- Delete "in their articles" and replace "In an attempt to offer" with "As".

> As a compromise between good projective properties of LHDs and a discrepancy criterion, Park (1994) and Morris and Mitchell (1995) proposed optimal Latin hypercube designs.

(2.) The wording in this example is poor making the sentence difficult to understand.

> It is the case that it is very difficult to find a set with smallest discrepancy for the number of factors are at least 2 factors ($k \geq 2$) since the distribution of N points in C^k may be complicated.

- "factors" is used twice, "It is the case" is necessary, and the description can be simplified. I suggest:

> It is very difficult to find a set having smallest discrepancy for 2 or more factors ($k \geq 2$) because the distribution of N points in C^k is complicated.

(3.) The following passage is not clear because of the lack of details. That is, the author was too "frugal" in the writing style.

> Let x_i be the i^{th} proportion which satisfy the following:
>
> $$x_i \ \geq 0 \text{ for } i = 1, 2, \ldots, q \text{ and } \sum_{i=1}^{q} x_i \ = \ 1.$$
>
> Under these, the design region for a mixture experiment is a regular simplex.

- For example, more details about x_i are needed and "Under these" is too vague. The following revision includes details for clarity. Note that I separated and numbered the constraints.

> Let x_i $(i = 1, 2, \ldots, q)$ be the proportion of the i^{th} component in a mixture. By definition, a mixture satisfies the following constraints:
>
> (i) $\quad x_i \ \geq \ 0 \quad$ for $i = 1, 2, \ldots, q$
>
> (ii) $\quad \sum_{i=1}^{q} x_i \ = \ 1.$
>
> Under constraints (i) and (ii), the design space for a mixture experiment is a regular $(q-1)$-dimensional simplex.

(4.) In the following passage, there are notational and conceptual problems. There are redundancies that can also be removed.

> In mixture designs when there are constraints on the component proportions, these are often upper and lower bound constraints of the form $L_i < x_i \leq U_i,\ i = 1, 2, \ldots, q$, where L_i and U_i are the lower bound and the upper bound constraints for the i^{th} component.

- L_i and U_i are numbers. Therefore, they cannot be referred to as the "lower bound and the upper bound constraints". They are values that are used to <u>define</u> the constraints.

> Mixture designs often have upper and lower bound constraints on the q component proportions $x_1, x_2, \ldots, x_q$. The upper (U_i) and lower (L_i) bounds for the i^{th} component define the constraint $L_i \leq x_i \leq U_i,\ i = 1, 2, \ldots, q$.

(5.) The following is an example of inconsistent terminology.

> Fang and Wang (1994) proposed the methods for generation the uniform designs in many experimental domains.

- Throughout the dissertation the term "experimental design space" is used. However, Fang and Wang use the term "experimental domain". The writer just copied Fang and Wang's terminology. Another example would be the use of the term "experimental region". You must change "domain" and "region" to "design space" throughout the dissertation. **Be consistent by using the same terminology throughout the dissertation.** When citing a reference always ask "Does the terminology in this reference match the terminology I have been using?".

(6.) The following example has a couple of problems.

> We see that the main advantage of Latin hypercube sampling is that they ensure stratified sampling in each dimension.

- First, avoid expressions like "We see that". Second, "they" is inappropriate because "they" refers to Latin hypercube sampling which is not plural ("it" is the appropriate singular form). I suggest

> The main advantage of Latin hypercube sampling is that it ensures stratified samples are taken in each dimension.

(7.) The following example contains several writing and notation problems.

> Ye et al. (2000) said that a design D_1 is said to be better than design D_2 if $d^*(D_1) > d^*(D_2)$. the number of pairs separated by this distance, denoted J. If two designs have the same d^* values, then the design with small J value is better.

- Here are 6 problems I identified:

  (i) 'said' is used twice in the first line.

  (ii) 'the' should be 'The'.

  (iii) It is unclear what 'pairs' is referring to.

  (iv) A math font was not used for J.

  (v) "the number of pairs separated by this distance, denoted J." is not a sentence.

  (vi) For clarity, there should be two distinct J-values.

- Assuming d^* has been defined, I rewrote this as:

> Design D_1 is better than design D_2 if $d^*(D_1) > d^*(D_2)$. If, however, $d^*(D_1) = d^*(D_2)$, define J_1 and J_2 to be the numbers of pairs of design points separated by distance $d^*(D_1)$ and $d^*(D_2)$, respectively. The design with the smaller J_i-value is the better design (Ye et al. (2000)).

(8.) The following example contains unnecessary words that can be removed.

> In this section, we would like to briefly give details of these discrepancies.

- You can remove "we would like to" and reword the sentence.

> In this section, these discrepancies are briefly described.

(9.) Although it is acceptable to use "I" when writing informally (like I am writing in these course notes), you should avoid using sentences with "I" in a dissertation or any professional research paper. For example:

> I would like to give a brief review of optimality criterion of LHS in the following paragraph (for more information on these design, see Koehler and Owen (1996), Bates et al.(1996), Fang et al.(2006)).

- Rewrite this without using "I" and "criterion" should be "criteria". I would also use two sentences.

> A brief review of optimality criteria for evaluating LHS will now be presented. For more information on LHS designs, see Koehler and Owen (1996), Bates et al.(1996), and Fang et al.(2006).

(10.) The following example lacks clarity because of missing details regarding the MSE. This leads to changes in the notation.

> Sacks, Schiller and Welch(1989) gave a detailed discussion on how to choose a suitable θ. Another criterion related to MSE is maximum mean squared error (MMSE). To optimize this criterion choose a design to minimize $maxMSE(\hat{y}(\mathbf{x}))$.

- Assuming θ has already been defined, I clarified it as follows:

> Sacks, Schiller and Welch(1989) discussed how to choose a suitable θ. Their choice of θ is based on the maximum mean squared error (MMSE) of prediction. That is, the optimal MMSE design D^* minimizes the maximum MSE of prediction:
>
> $$\text{MSE}(\hat{y}(\mathbf{x})|D^*) \;=\; \max_{D\in\Omega} \text{MSE}(\hat{y}(\mathbf{x})|D).$$

(11.) In the Literature Review, the Ph.D. student included:

> Since $\mathbf{V}_D = \sigma^2\mathbf{R}_D$, where $\mathbf{R}_D$ is the correlation matrix of the design matrix $\mathbf{D} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. Therefore, $\max\limits_{D\in\mathcal{D}} \ln(det(\mathbf{V}_D))$ is equivalent to $\max\limits_{D\in\mathcal{D}} \ln(det(\mathbf{R}_D))$.

- There is nothing mathematically incorrect about these sentences. So, what was the problem? The problem was that this information appeared in the Literature Review but had nothing to do with the dissertation research. The student found it in a reference, but never considered its lack of relevance. When in doubt, ask yourself this question: "Is this information needed or useful for the dissertation research?"

(12.) What information is missing from the following definition?

> A *minimax distance design* D^* minimizes the maximum distance between any point $\mathbf{x} \in T$ and design D, i.e.,
>
> $$\min_{D} \max_{\mathbf{x} \in T} d(\mathbf{x}, D) = \max_{\mathbf{x} \in T} d(\mathbf{x}, D^*) \tag{1}$$

- I would ask the following questions:
  - What is T?
  - $\mathbf{x}$ is a point and D is a design. So what is meant by "the maximum distance" between a point $\mathbf{x}$ and a design D?
- Next, consider the notation. You should ask if the notation has already been defined and if it is being used consistently. Specifically:
  - Has T been defined earlier?
  - Is the notation consistent with the rest of the dissertation?
  - Has the distance function $d(\mathbf{x}, D)$ been defined?
  - Over what set is D being minimized ($\min_{D}$) ?

In the dissertation, T represents the design space, but it was defined earlier using notation $\mathcal{X}$. The writer used T because T was used in a journal article. The distance function $d(\mathbf{x}, D)$, however, was not defined earlier in the dissertation. I also switched the two-sides in (1). I revised this passage so that all notation is defined and consistent:

> Let $\mathcal{X}$ be the design space and Ω the space of possible designs. Let $\mathbf{x}_j$ be the j^{th} design point ($j = 1, 2, \ldots, N$) of design D. The distance between a point $\mathbf{x} \in \mathcal{X}$ and a design D (denoted $d(\mathbf{x}, D)$) is the minimum distance between $\mathbf{x}$ and $\mathbf{x}_j$ for $j = 1, 2, \ldots, N$. That is, $d(\mathbf{x}, D)$ is the distance between $\mathbf{x}$ and the nearest design point in D.
>
> A *minimax distance design* D^* is a design D that minimizes the maximum distance between any $\mathbf{x} \in \mathcal{X}$ and the nearest design point in D. That is, D^* is a design D that satisfies
>
> $$\max_{\mathbf{x} \in \mathcal{X}} d(\mathbf{x}, D^*) \;=\; \min_{D \in \Omega} \max_{\mathbf{x} \in \mathcal{X}} d(\mathbf{x}, D) \tag{2}$$

(13.) Although there is nothing seriously wrong with the following example, it can be simplified to improve clarity.

> The concept of uniform design is to generate a set of experimental points based on quasi-Monte Carlo method or number-theoretic method such that the points are uniformly scattered throughout the design region with low discrepancy. As UDs are space-filling design, it means that the design points are not concentrated in clusters of points or solely on the boundary of the region.

- Missing article 'a' was inserted. Here is one way to rewrite it:

A uniform design is a space-filling set of experimental points that are generated by a quasi-Monte Carlo method or a number-theoretic method. Thus, the points of a UD are uniformly scattered throughout the design space, and are not spatially clustered and do not occur only on the boundary of the design space.

(14.) In the final example, the clarity can be improved by including additional details.

Both LHD and UD are space-filling experimental designs – the LHD in a randomly uniform fashion and the UD in a deterministically uniform fashion. Specially, if the experimental domain is finite, LHDs are similar to UD. When the experimental space is continuous, this two designs are different. That is, in LHDs, points are selected from cells, whereas in UD points are selected from the center of cells.

- In this case, I was "less frugal" and "more verbose".
  - (i) "randomly" is missing and needs to be included for clarity.
  - (ii) It is better to remove "–" and form two sentences.
  - (iii) "Specifically" is misspelled "Specially". A spell checker would not detect this.
- Here is my revision:

Both LHDs and UDs are space-filling experimental designs. The LHDs are space-filling in a randomly uniform fashion while the UDs are space-filling in a deterministically uniform fashion. If the experimental space is finite, then LHDs are similar to UDs. However, when the experimental design space is continuous, LHDs and UDs are different. Specifically, the points in LHDs are selected randomly from within cells, whereas the points in UDs are selected at the center of cells.

Practice Examples:

(E1) What is the problem with the following?

Let $X_1, X_2, \ldots, X_N$ be a simple random sample of n values taken from a population. Let $\overline{X}_n = (x_1 + x_2 + \cdots + x_n)/n$ be the sample mean of the n values.

(E2) Correct any errors you find in the following example:

Johnson, Moore, and Ylvisaker (1990) proposed the minimax and maximin distance design. These distance criteria measure how uniformly the experimental points are scattered through the domain. The detail of these designs are described below.

(E3) Fix the following sentence.

Define $\mathbf{x_j} = (x_{j1}, x_{j2}, \ldots, x_{jk})$ to be j^{th} row of $\mathbf{c}$ and $\mathcal{P}_N$ be the set $\{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{X_N}\}$ formed by n rows of C.