웹 소프트웨어 신뢰성

- ✓ **Instructor: Gregg Rothermel**
- ✓ **Institution:** 한국과학기술원
- ✓ **Dictated: OES** 봉사단

Here is chart that show experiment process and breaks down into the some sub parts.

It all begins when you have an idea for an experiment of theory you want to confirm or some sort of relationship you want to look for.

And then you go through the following steps resulting in the conclusion.

Definition, planning, operation, analysis, presentation.

We will talk about each of these in turn.

But briefly, out there on sides was experimental idea, that's where you got to get good question.

If you got a silly question is not worth conducting an experiment on.

Our question the people aren't really interested in or one that doesn't have any impact.

So find for the right question first.

And then definition is about figure out how to ask that question in the right way in fact the whole processes is into that but once you got the question you know what you want asking you find out to ask right way everything else follows that.

Everything else is determined as does that allow me that answer the question.

So the planning you design experiment to do so and you'll see several sub steps to that that's almost of the intellectual effort happens.

Well yeah, much of it.

Then there is the running of the experiment and finally analyzing in interpreting before you write it up.

So experiment definition really sub parts to this I say I think I got another, well, no I don't have another graph but there's got to be sub parts.

Excuse me.

That's planning there are sub parts on.

Definition.

I'm going the wrong way.

Okay.

Again ask the right question.

And figure out.

Sorry my heads messed up yet.

Where we go? Ask the right question. Ask question right.

This should be really saying asks the question right. But I got to turn around.

Or idea is ask the right question.

Now. Think about your goals, why you're conducting the experiment.

And state them in terms of research questions.

And the research questions really that are the statement of the experimental aim and the do end up guiding the rest of the design as I said a moment ago.

There are various forms of questions.

Starting at the bottom, one of the typical ones is causal, you're trying to turn whether A causes B, whether treatment A cures disease B treatment B cures disease whatever.

So I've stated that what is the average productivity of developers using OO versus non-OO but causally speaking what you're looking at is the probably statement of the users of OO improve average productivity.

That would be the causal statement of that.

Does something cause the other? But you can also just look at the descriptive questions what percentage use OO that could probably answer this survey of course on experiment.

What percentage of experienced verses novice developers use OO?

Does the experience have anything to do?

The various forms of question you can ask.

Depend on what you're trying to learn.

I should swap this for mash-up questions.

But this is a question we look that in research some time ago about two aspects of test suites which size and composition and how they affect testing methodologies?

So you could say web testing methodologies.

So this is a question and concerned what test suites can have two different characteristics how those two affect testing methodologies.

Once you got your questions that now you start planning experiment.

And this is where you break down in the several sub steps.

And this is probably the most intellectually challenging after that we come analysis interpretation. And a lot depends on this.

If you don't get a good experiment design, everything after that, this going to fall flat.

I mean it's not, if you're answering the wrong question, if you're doing the wrong way everything afterwards is going to have problem.

So it's just we say in software takes some time getting requirements, design this after that, things are hard.

This is similar thing here. So there are the steps.

Selecting the contexts for your experiment, formulating hypothesis, selecting your variables, selecting your subjects.

Then doing the design itself building any instrumentation, and evaluating validity. We're going to go through each of these.

Now, the context for experiment that means kind of the environment in which is happening.

Now, wide range of things could happen here.

Obviously if you are conducting experiment on the watching graduate students, use a couple of different programming techniques.

Well personnel graduate students environment might be the laboratory?

But if you are conducting in experiment in a corporate setting which you could do it's a little harder to control things.

The environment might be the setting.

And then you have to determine which setting you'll going to use.

And there're some dimensions to that.

We talk about off-line versus on-line experiments.

On-line are experiments where you really are working in the context of some, working with people directly, such as in a corporation or in a lab with people directly.

But everything has to happen at the moment correctly.

So you get a bunch of students together and you have two hours to run an experiment, it's all got to happen on that time.

There's other experiment we do off-line like you could imagine looking at testing techniques, applied to programs, now one way to do that is to study people developing those.

But I can't take the people out and look at a bunch of test suites and now I'm running the test suites on the programs to see what they do.

And I can do that again and again if something follows up in my in something gets messed up in my instrumentation.

I can do it again and again.

So it's called off-line.

Context includes various things if you got people involved context include whatever group people are involved.

Plus that's often student versus professional personnel.

What we do soft engineering experiments.

The type of thing here is what type of problems are you working on.

If you are watching people create web applications in a two hour class and the two hour lab time.

The size probably can work on is limited by that time.

And there are some testing techniques and things that are so expensive you can only people start with very toy programs, tiny line programs.

Now of course, will this generalize the larger of programs is the question.

But you've got to choose all these that's the point of this slide.

What will let you answer the research question in the time that you have is what you're facing here.

And then there are specific versus general domains.

And what's an example of that? Well the mash-up paper was about fairly specific software domain, web mash-ups.

Now the study was yahoo pipes which is one type of mash up.

But we would think that that can generalize to wire oriented mash-ups fairly well.

Mash-ups that can consist of modules connected by things.

But beyond that two your non visual programs it probably doesn't generalize.

So you have to think about you know how broad or narrow can this be and that can be important, well for instance in motivating a paper.

We took that the mash-up paper about versioning mash-ups, wrote a journal version of it submitted into transactions on CEI(computer enumerated interaction) and got a fairy good reviews, but one of the major comment was the paper is just about mash-ups.

That doesn't seem very broad in area how much do we care about that can we generalize it more.

And so we're rewrite we've tried this specify that what I'm saying for mash-ups are one instance of this broader classic system called wire oriented we try to see programming visual programming languages and so those are important and mash-ups are instance of that and we will be learn about those can probably generalize to those.

And hopefully that we're going to accepted.

That's what the reviewers suggest we do.

So all of these choices, again we're going to affect whether you can answer your questions in a meaningful way whether you can get something meaningful out of in a paper whether you conducted in your time in the time allowed.

It's cost verses validity, does it generalize? Is it meaningful? That kind of thing.

That trade off.

When you got your context, context you can now pin down your research questions into statistical hypothesis.

Assume you're going to have the data for hypothesis testing.

You are going to come up with these, just let me check some ahead.

So it's a formal statement.

And if you've taken statistics or experimental design you'll see how everything statistic comes out of that statement.

And the notion is you formulate hypothesis and stats tests can tell you if you can reject it.

That's kind of interesting.

What we're really going to do? For any given question, there are two hypotheses.

The first, null hypothesis and that's stated which prediction does not hold.

So if my prediction is, my new technique helped people test better than what they are doing.

Presently, like the state of the art, the null hypothesis says my technique will not improve the testing.

And what we called the alternative flips that and search what we want by technique will improve.

But the point is that what statistical tests do is determining whether the null hypothesis holds.

And so if you have your null hypothesis and you collect your data and you learn this statistical test and it gives you a P value problem that the null hypothesis can be rejected and if that's sufficiently high and then you can reject the null and that means accept the alternative which is what's in line with your research question, okay?

And another couple types of errors here.

If your null hypothesis is true, let me put one up here.

Just a very simply stated one.

Technique A is better than current practice.

Null hypothesis is going to be essentially A not better than current practice, right?

That's the null hypothesis.

You are hoping when you're usually saying that's the research question, that's the null hypothesis and the alternative is A is better.

If this really is true, in the real world we don't know I mean that's why we are conducting experiment but apart from experiment there is some truth there.

That we're trying to learn.

And it's a real world it's true A is not better than current practice.

And we fail to reject that.

And then that probably we do reject it.

Now we've made in an error.

And we've asserted that the alternative is true when it really isn't, okay?

And we want to avoid both of these types of errors but in particular we want to avoid inferring of relationship that does not exist.

Now how could that happen, how could I conclude that A not better is false and therefore A is better but one way would be if I had some other factor causing it to look better.

And it wasn't really A. that's an internal validity threat.

So I could somehow get bad database or something measurement errors.

That makes it look likes better so this could happen.

And so you want to avoid those things.

The other way probably of not rejecting A false null hypothesis so if this is false.

If it's false A is not better then it's true that A is better and if you fail to assert that, then you miss the relationship that exists.

In terms of medicines this one is like testing aspirin against sugar pills for headaches and concluding aspirin is no better, if in the case where really is.

So you haven't learned something that you could've learned.

And that can happen too.

Ask type two error and one way that happens is this you don't have enough data it has two powers of statistical tests whether you can reject the null hypothesis.

And as it happens if it's physical tests you can reject the null hypothesis, statistically speaking you can conclude the null hypothesis doesn't hold unless there's some

other threats.

But if it says you can't reject it could it just because there's not enough data.

So that's hypothesis.

But you come up with formal hypothesis?

Of course they're related to each questions but they can be much more precise.

And the next thing that you think about variables.

In any experiment we're manipulating independent variables treatments testing techniques approaches engineers we're using and measuring some affect dependent variable.

How well you detect false?

What kind of coverage you get on the code? How easy is debug?

Whatever is coming up with? And so you need to select those variables and along the way there may also be the other factors remember we talked about that.

I think I get to that more in a minute but there is a thing manipulating and things you're measuring.

And if this others things that could jump in there affect things that you have do something careful with those.

Which we'll look at. And once you got these, you got to figure out, you are going measure things.

Remember these different measures, direct or indirect, I can measure anger by the sweat on the palms maybe if you believe that.

But you got to come up with measure and the scales.

Sometimes you have ranges for variables.

If you are looking at not just at A, Let's say not just at medicine but the amount of the medicine.

Obviously you are wondering about some amounts between zero and a hundred milligrams well obviously you can't try, you won't have enough people to try 0, 1, 2, 3, 4, 5, 6… so you can try some ranges, you might try 100, 75, 50 that kind of thing. OK?

Now, the big thing, select subjects and objects.

We use subjects when we are talking about people.

With objects when we're talking about non people.

So acting testing many of testing experiments are studying the application of techniques to programs.

The programs are the objects.

They might be studying existing tests suites those are objects too.

But if you're looking at, people performing the experiment in comparing the people's behavior those are subjects.

Now they're probably working with some objects too.

You might be giving the people the mash-up experiment or you gave the people mash-ups to work with.

How well you do affect ability to generalize?

I talked about this before but if I am studying something and I select medicine for minute, I'm studying effect of the medicine on people suffering from disease and I select people who are all in their 20s that may not generalize to people who are in their 50s, 60s, 70s anything like that, okay?

Other fact may be involved.

So what you can generalize to? It's going to be affected by what you picked for your an objects.

And so the best processes to use is going to be as random as possible and you're going to have to use sampling technique.

Let me back-up on that random statement but, we'll talk about that in a minute.

How do you select subjects or objects? Begin by identifying your target population, okay? And then you sample from those.

Now what do I mean by this medical experiment if I'm interested in studying fertility drug,

My population probably is just women in the age well let's say 20~50. And that's all it's applicable to.

But if it's a cancer drug it might be much much wider.

Or if I'm interested in a treatment that's only supposed to help elderly people it might be that.

You have to decide what's your population.

If I'm interested in obviously web applications my population might be the population of all web applications.

If I know what that population is.

As probably unrealistic you might have to say well open social web applications where I can get the source of the population.

Cause that's all I can hope to select from Now that becomes your population.

And then you need to sample from it.

So again that could be objects and peoples in those examples I'm using.

The different ways hang on a minute okay.

The some methods for sampling are here.

There's probability based and non probability based.

And in generally we prefer probability based.

Often times we're stuck without it in our field.

But the simplest thing is a simple random selection.

So if I could if my if you if I identify you as all open source projects in these source forge repository, okay?

Again my results may be generalized beyond or maybe I can say generalized to open source but if identify that as my U now I can simply randomly select form that.

Randomly select k objects.

And if its random then you have hopefully have a better chance generalizing your results.

You can say well it was a random selection so it's probably not biased towards some type of program.

A same thing can happen for selecting people if you catch population you can somehow randomly select.

Of slight variantum maybe?

That select the first subject at random and then every nth after that and happen to be in some order.

This is an important one.

Stratify random. Why would you do this?

Talk about people for a minute.

Suppose you're interested in the performance of people doing a programming task.

And there are going to be two ways.

Maybe one's using your environment and one's using some other environment.

But suppose you know that expert programmers or people with five or more use of experience are different with ones with 0~5.

You got two ways you can do this.

You can take your pool of programmers which concludes some at all levels of experience.

And sample of them to choose your treatment group and your control group.

Randomly sample them.

Okay? Now you got no guarantee that I mean they're random so I mean probabilistically you should have equal numbers of experience in both groups.

But if you really want to be sure about that perhaps because you may be want to look at sub questions regarding experience you should first divide this into strata.

Strata.

Layers.

Now you can randomly sample and if this is forty percent of P, and 60% of N. you pick 6 tenths of your population randomly from here and 4 tenths from here.

And now you ensure that you got the strata in the same proportion they exist in your population.

In your set here.

And now you can look at that.

And know that it mirrors this.

So that's stratified sampling.

That was people I was using if there are types of programs, if you're looking at real time programs and you know that certain percentage use semaphores and certain

percentage use spin locks.

Then maybe you want to divide those into strata and select accordingly.

Various reasons for that.

So those are sampling mechanisms.

Then the ones that aren't probabilistic, convenience that sounds kind of unscientific but select the most convenient case studies that's often what we do.

I mean I have a relationship with a certain company with lockie.

I can get at their programs so I can run a big case study or a big experiment on their processes and I can't deal with anything else so it's going to be convenience sample.

It's going to limit what you can gemplize to (23:25) and you have to describe how you chose the subject.

The danger here is that in sense you can select the most convenient you can select the one you most think will prove your hypothesis true.

And that of course is biasing your experiment.

So that's what you want to avoid.

There need to be some good reasons to picking it.

That you 'd like to say well we do think it's somehow representative it's complicated program you'd like to be able to say we don't know whether there are techniques that work on it, we didn't check beforehand.

Quota.

You can use that in this example here if you wanted certain number of experienced ones certain number of inexperienced and.

I want a certain number of I don't know what your experience is in programming but if you guys are convenient to me I can pick one of you who's got five to ten years' experience and one of you who has zero to five.

And I'm doing kind of quota sampling.

Okay larger your sample size the less error in your statistical test for your conclusions.

For stats if populations have larger variability that's where you need larger samples.

And also we'll see that data analysis methods in stats may influence choice of

sample sizes.

The stronger there are different strengths statistical tests.

And if you wanted for some reason to use some of the stronger ones you probably have to have larger samples.

As you sample those up your cost goes up.

So your___ (25:11) is a trade-off.

So in some sense we want to sample small as possible but large enough so that we can generalize.

Now there are some principles behind so that's all the sub the selecting subjects and objects and randomizations involved there.

A little bit of blocking but these are some principles to keep in mind throughout the experiment designing a randomization is just anytime you can do something randomly it's probably better than a non-random approach.

Because the statistical methods tend to require that I talked about other factors that may influence results.

So okay. We use the example that's talked about here.

You don't know this paper but well think of the web macro paper where we they really only did one I'm not going to use that one.

Think of a testing experiment we're apply8ng techniques to program okay?

But if you think this technique may vary with different types of programs, then you may want to block over program.

Let me explain that better. Well it's still a is better than b. and these are testing techniques.

Tactic A. I'll just say greater equal to tactic B. and that's your question. Okay?

And what you do want do is select the whole number of programs.

So you could get a bunch of programs.

Run A on B see how many faults are detected by a and b okay?

And you could say on all of these programs.

We knew there were a 100 faults A detected 70 B detected 30 okay?

And you could just seek analysis on that show whether its statistically significant.

But you're lumping all these programs together.

You're putting them all together, treating them all equally as a set.

Now if you think that program has a strong influence on technique, that may not be the right way to do it.

You may want to actually analyze each program separately.

I might actually find in this that a and b are equally good.

Now if I broke this down into the program individually, I may find that oh, A is better than B here. A is better than B here.

A is better than B here. Let's just say that in all these cases.

And in these two, b is better than A.

If I think program greatly affects result, I may want to analyze per program instead and I'll get more information out of this.

And that happens in a lot of cases.

So what that is doing is blocking it's separating my data in terms of the factor in this case program that I think affects results.

Another example blocking would be to take my results and separate them into the ones with experience programs at programmers and non-experienced and consider the results separately.

Or N user programmers and professional programmers and consider the results separately.

As if think that has a result.

The other thing is when you're signing multiple treatments you'd like to balance things you'd like to apply treatments to equal numbers of subjects and that's particularly for this statistical analysis.

So medical experiment you don't want to take the drug and apply it to 70 people and the TSE go to 30.

You'd like to do equal groups.

Now we get into what is called experiment design.

In some sense we've been doing, there's been design in the arts sense that we've

been doing that.

But there is a smaller sense it's called experiment design which you'll see what I mean in a minute.

And this relates to how many factors in treatments you're looking at.

And so one factor was two.

One factor is more than two.

We'll go through these subjects with slides.

I won't say what they are now.

Often times experiments is dependant variable fault detection effectiveness coverage.

They were looking at averages or mains.

Some use or bys the meaning of dependent variable for treatment.

So 1 factor 2 treatments.

What does it mean? 1 factor 2 treatments.

The factor if the factor's testing technique and there's tactics A and B that's one factor 2 treatments.

If it's programming environment and you got 2 environments that's 1 factor 2 treatments.

Yes. Yeah do not think.

It's hard to say do nothing. Usually you compare against something.

My technique detected 50 faults.

If it didn't run any tactics at all I wouldn't detect any.

T's not too meaningful to certain most people okay?

If the weird case where that happens is if no one's ever come up with any approach for doing this, you got nothing to compare against so you could say no one's ever been able to identify this encode before and here we identified 12 instances.

But usually you're comparing against something.

And you don't want to be a stupid something either.

So what you'd like if you're comparing some engineering technique some new one.

You'd like to compare to state of the art or state of the practice.

You know state of the art refers to what's kind of the most advanced technique to come out of the literature so far.

State of the practice could be what are the engineers currently doing.

So could be common practices of other thing to compare against.

Now sometimes in testing we choose intelligence tests versus random tests.

And sometime that's okay but sometimes reviewers will say well random testing I mean even practicioners aren't doing this.

They're doing something more intelligent.

So you're kind of using an unfair comparison there.

So basically if you had only one treatment this and no one's done anything else.

It's not even experiment in some sense okay?

Now think of this as human subjects and the treatment says being think of humans as a testing technique.

Not just humans aren't creating tests to some criteria versus some other.

This is a completely randomized design.

I only have 6 students listed here but if could continue on.

If I randomly assign subject 1 to a treatment or I randomly pick N over 2, 6 over 2 people and assign them to treatment 1 in random and assign the others here.

That's completely randomized okay?

And once you do that you can use a simple means comparison.

You know the mean faults detected mean faults detected comparing them.

The __hypothesis becomes the means are equal and the alternative is that detected are not equal or one is better than the other.

Which is probably what you're trying to assert.

And when you do that and here are couple of statistical tests you can use.

Now this statistical we'll say more about this later but there are other things that

affect what you are going to use in a statistical test.

Such as qualities of data.

Well if the data follows a normal distribution, you can do certain types of tests.

But if it doesn't you can't.

But apart from that here are couple types of analysis you can do to test equivalence and the means in a 1 factor 2 treatment thing where you've completely randomized.

Yeah so this is what I already said.

Humans using one method detecting the faults better than humans using another.

Now here's what we'll call paired comparison.

It's still 1 factor 2 treatments but I can have everyone do treatment 1 and everyone do treatment 2.

Why did I list 2 on 1 2 2 1 1 2 2 1 1 2? When I could have said 11111111 22222222? Say it again.

Learning effect.

It's a formal balancing okay? And it is a learning effect so I get my Subjects and I say here follow this process which you get.

Now here follow this iprocess.

They've all done one first and then2. And the differences in performance may be due to that, rather than difference in treatments okay. so balancing this at least I can maybe balance out the learning effects because half of them did 1 first and 2 and half of them did 2 first and then 1 okay?

So the reason you could do it more precisely, well first off you got more data.

Second, everyone has done both and so if you in this past one if you did a bad job assigning your groups, I mean your results may be 2.

It wasn't really a random assignment by some facts.

But again you have to be aware of learning effects.

And you could still do you're doing hypothesis 0 is that the mean of the differences. It is about the mean of the differences is 0.

Between the performance.

I mean you're looking at differences in performance here.

Since we've got sine each subjects did 2 you can look at the difference in their performance.

And compare the means of those.

And there are other tests you can use.

Again depending on other factors.

One other thing.

You can't want with do it.

A medical example is used in this case.

If I'm trying to see what happens if I give them a drug versus what happens if I give them a placebo or a sugar pill.

Well this is you give them both treatments maybe balancing shows you something in terms of how soon they get better but it really confuse things by giving people both treatments.

Or their 2 drugs.

May not matter. It's hard to assert things then.

Testing techniques on programs, if our subjects are actually.. you know..

Program 1, program 2, program 3. There's the learning effects there.

In fact, I don't even have to balance it.

I can learn technique 1 and all the programs and then technique 2.So.. so there's an example.

I pretty much talked about that, though. One factor, 3 treatments.

The same things I just talked about generalize if you got more treatments.

If you got 3 testing techniques.

3 development techniques. Again.

You feel completely randomize the assignment of subjects.

If I don't want anyone subject doing more than one treatment.

And then, we are looking at the means being equal or different.

And when there's multiple, there's some different tests you have to do. Okay?

There's the different tests that applies.

All things like t, apparently t test don't apply.

And ANOVA's do.

I don't want to spend too much time on examples, you can read them.

And here is another example of the ..of assigning complete things.

And this case, everyone does have every treatment.

But I've balance the all of assignments.

So, they all have different orders and which they considered.

And again, there we can look at equal means and various tests we can use.

But again, in thinking about all this, that's what you got to think about this so this experimental design in the statistical sense.

How do I break things up in a treatment? Assign of treatments to the subjects or objects.

And this type of design is what determines what statue you can use, and the, and so it's important to pick the right design for your question and avoid the problems like learning effects.

What about when you get more factors?

One example is technique and experience.

Now I talk about blocking over experience, but if I'm truly interested in as a way to measure things independently.

But If I'm truly interested in experience as a variable, if I say, I have these two techniques that I want to compare.

But I also want to see whether experience plays a role and there use of this.

And I'm interested in both variables, I'm like, I'm actually interested in four combinations.

Technique A with experienced people, technique A with none, technique B with experienced, technique B with none.

I'm interested in all of those combinations.

That's when I got 2 factors. And they combine in this way.

So, would you get that?

You got to two treatments for A, two treatments for B, if you are assigning people to the groups, you are going to assign them into different cells.

Better example on this, One I like is an agricultural example.

Suppose factor A is fertilizer, and factor B is insecticide.

And the two may interact.

We don't know.

And you want to tell different combinations, I got fertilizer A and B, And insecticides C and D, what of the four combinations works best?

So these four things are actually fields.

In which I'm planting rice.

And trying this fertilizers and whatever combinations.

So obviously, I'm not the four cells.

These are fields.

1 and 7 and 5 and 8. obviously, I can't put it every field with every combination. That meets the problem.

So I randomly assign the fields.

Two different cells, and I can see what happens within those. Okay? So this is called 2 by 2 factorial design, because there are two choices, a treatment for each factor, and there's actually three hypothesis is going on.

There's the separate effect of A, the separate effect of B, and there's the whether they interact.

And you can look all three of those in this case.

If that's what you are interested in looking at.

And you do have the equal means, hypothesis or no hypothesis and usually the analysis of variances is one they use there to check. Insist one kicking out heat,

when we don't need it.

Okay, let's skip that example because it's different one. So the generalizes, K factors, 2 treatments, well you have factor A, B and C. And there are different treatments. A1, A2, A1, A2, A1, A2. So here's all the combinations, right?

That's every combination of a A1 with B1 with C1. A1 with B1 with C2. And you will find them all there.

There's the eight growth as you expect with a.. this is 3 factors and 2 treatments. You will expect 2 and the 3 combinations.

So figure out all this again.

Farmer's field where it's insecticide, fertilizer and planting date.

You have to do something like this.

And you assign the subjects there. Okay?

These are combinations of factors in influencing testing techniques, there's no interaction effects.

It doesn't matter.

I could use all subjects with all combinations with no problem.

I'm just running techniques under like; testing techniques, together with oracle technique together with observability technique, I could try to combination every subject I wouldn't care but here with farmer's field I care.

So you do the same types of analyses.

You can see how this expands than and the different things.

But they get more and more complicated because here you have how this A react, how this B react, how this C react and other interaction factors of these two, interaction factors of these two, interaction of these two, and it gets harder and harder to see things.

And it gets more expensive.

Because there I have divided into the eight.

So then, we hardly ever see this in our experiments but there are cases where if you get many many factors you can't do every combination and so you may use combinatorial techniques to identify a subset of those.

So what this is showing, what this is identifying subset of the 8. The 8.

There are 8 different combinations of 3. Okay?

But with four combinations, you can ensure that every factor appears with every value of each other factor at least once.

So here I've got A1 with C1, A1 with C2, A1 with B1, A1 with B2, I never have A1 with, what don't I have, A1 with B2 with C2. I never have that. Okay?

But I have every, I have all the pairs. Okay? That's combination factorial design that, if you think that the high-order interactions are negligible, you can look at low-order interactions in this way.

And have fewer cells to look at.

As call the one-half fractional factorial design.

So you can look at experimental purpose of many many types of designs.

Usually for us, we are looking at one or two treatments and, I mean usually for us, we are looking at one or at most two factors, and a small number treatments.

So they tend to be an easier ones.

Once you got that design, now you are enter the instrumentation.

And that means a whole bunch of things.

I mean all the stuffs you need to run the experiment.

And all depends on what the experiment is.

An instrumentation, I don't mean people anymore, but I do mean, I need the code objects you need, a programs the people need to look at the specifications they need, any checklist, tutorials, questionnaires actually that's a measurement instruments, questionnaires you got to put all these stuffs together to make things such as you can run the experiment.

Here's your goal.

To make it possible to run it without affecting control, you don't want instrumentation to affect the outcome of experiment, okay?

Now you get to the validity evaluation.

Everything follows.

Once you going to design, most of your threats to the validity can be evaluated now.

And I have mentioned all these various types in the course.

But external validity is, can we generalize?

Internal is if you conclude A causes B. Could make a mistake? Could something else make it look like A caused B? Construct has to do with measurements.

To depict the right measures with the measures, bias your results.

And conclusion has to do with the stats.

We are going look at each of these in more detail for some example threats.

But first off, earlier, I talked about you got this world of theory and you are trying to pick a world observations to look at to trying confirm or contradict the theory.

And when I say external validity, so if my theory is that in general using development technique A is going to, going to produce higher equality code that technique B, that's a relatively broad theory in general, right?

But I observe, the application of A and B and a couples of specific settings, the question is “Can I generalize?” By looking at this observation, can I conclude, okay, when I look at that observation, I'm trying to conclude whether the cause or effect happens.

And If I can conclude that correctly, can I conclude that it holds up here.

That's external validity.

Does what I observe generalize to the larger context or the larger theory.

So what we are really looking at is whether this affect I inferred holds up in the theoretical world.

That's what external validity is about.

Now internal validity, internal is concerned with your running of experiment.

There, you looked at a smaller set of things. We don't need to generalizes but just relate to that set.

Did you make a mistake in inferring the causal relationship? If you did, that's internal validity.

If you have internal validity, then relationship you infer is accurately inferred. And you are likely to have that.

You can have internal validity and not external, obviously.

You have internal validity but it still applies to the range of things you looked at.

At least you want that.

Conclusion validity applies here, too, and has to do with your statistical tests.

Whether you apply the right tests.

And then construct is all about how you map the theory in to the world.

See it’s easier to see down here.

I talked about measuring anger.

By construct for anger with sweaty poems.

If that's an accurate measure of anger, then, the various things can hold like in, but if it's not then I've got a problem.

So let's look at the specific threats.

External validity, Does it generalize?

If the subjects aren representative of population, that we wish to generalize to? That's a problem.

Experimental setting isn't representative? That's a problem? Experiment is conducted at a time that affects results?

Like if you are going to company and conduct something at product release time, and they are too busy to do things appropriately.

Reduce those by making environment as realistic as possible.

But reality is not homogenous, so we need to report the environment characteristics.

And the mainly address external validity is to perform more and more experiments with different groups of people and different conditions. Internal validity.

Instrumentation.

This can related to construct also but basically again internal is, can you, if you infer that A implies B are you right? And if you are measuring things wrong, you might inferred that A implies B when they are really didn't. Okay?

Your groups are not equivalent.

That's the easy one.

I put all the experts in one group and non experts to another.

What I'm really seeing is difference in expertise.

Learning effects.

We have talked about that before.

Mortality. Subjects drop out of the experiment.

We have that happen sometimes in doing experiment in room with 20 students. Couple of the systems are crashed.

They dropped out of the experiment.

Unfortunately in medical experiments, that means literary mortality. I mean in medical experiments in United States.

As far as I know. If you are doing at study of treatments, cancer treatments, if enough people die in one group, experiment has to stop.

And that's specified by the agency that run those things.

Social effects. It can happen. Control group resents the treatment group?

If your group comes in.

If I brought a group and say.. This is the stupid things that to do.

But if I brought you in.. hey I developed this new technique for helping you understand what you read better.

And I'm doing an experiment on that. You guys don't get the technique you guys just read. Oh darn it.

I want to learn the technique.

So it might resent that. You know something might happen from that, okay?

You don't want a.. you want to watch out for this things in your design.

You don't need to let.. you don't need to tell to control group that they are control group. Hey we are looking at how you test these things. Construct.

This is about measurement that I choose can measurements.

This is from a textbook.

Inadequate preoperational explication of constructs.

That's a good one. Preoperational means before. Before operation.

Inadequate means not good enough.

Explication means understanding spelling out.

I'm really is, more like, what the theory isn't clear enough, what is that they some things better.

Mono-operation bias.

Using a single independent variable, case, subject, treatment or may under-represent different constructs.

There may be different things you need to measure.

A levels of constructs. What was an example there?

Well I've been studying coverage techniques.

And I say well we had, we developed 90 percent coverage adequate test threats compare with those.

Maybe the strength of coverage happens only when it gets to 100 percent.

So you might not see any differences.

Just do your choice of levels. Integration of testing and treatment.

Testing itself makes subjects sensitive to treatment.

The test is part of a treatment.

I don't know how often that happens to us.

Social effects.

Experimenter expectancy, evaluation apprehension, people are afraid what's happening in there.

You know, whereas those people is unanimous whatever.

But hypothesis is guessing.

And people would do that and trying to say.

What is it they want me to show? I want to try and make them happy. Okay? Conclusion validity.

That's about statistics.

Low statistical power.

Usually has to do with your samples eyes or your choice of tests.

Some tests can only be used in certain cases, like normally distributed data.

That's the statistical question you need to check on.

Fishing. Searching for specific results.

If you got bunch of data and you start saying well it's the a this results of here let's do this one let's do this one you are ..analysis ..that's called fishing and may influence in outcome I think ..I am not a statistician.

But I'm pretty sure that when you applying multiple tests you supposed to change your code your alpha value though.

The level I which called the things significant, you supposed to make that harder and harder with this success of test to counter for this.

Well obvious reliability.

If you every time you measure it the result differs, that's going to be problem.

So there are lots validity threats and it's always tradeoffs among these.

And so if I am using computer science students.

I may have access to more students that way.

So let's begin larger groups.

But now, it's a less mixed group. It's all CS students.

So larger group helps you conclusion validity but reduces external validity.

And all you need is tradeoffs.

And usually in theory testing, we are going to tradeoff some external validity.

We're trying to build internal and construct validity primarily.

What we're trying to see is causal effect.

And then, then you can do subject experimentation.

Once you know the causal effect, your subject experimentation, we don't have enough data for statistics, but, it's a real situation, and you can look up more external validity.

So that ends up external validity conducted in families, certainly true in other sciences. One experiment usually doesn't resolve the issue.

Using another and another settings. So that's that. We're gonna do the other set of slides here, today. Okay? Which is shorter.

So that was experiment planning and this is often a lot goes into that.

Once you look at these slides are guidelines something like that.

Once you got the planning done, you're actually running the experiment.

There're three parts preparing, executing and validating the data.

Preparation depends on the type of experiment you're doing.

I mean you mainly get participants.

You're getting people, in our experiments, you and all, often times we're sending out e-mails to various departments saying "We'd like people of participating this experiment on XYZ. Participants have to have the following background, participants will be paid twenty dollars for their participation.

Do you see calls for that around here? We want people for experiments?

There may not be many human experiments here since its hard science stuff.

But if you're down to road, maybe psychology experiments or things there might be more of things coming out.

Obtaining participant consent. In the US, and I think in Korea too, there are lots of rules now about how you can conduct these things.

I can't just grab a bunch of people and experiment with them.

For us in computer science it doesn't matter too much.

But you got to imagine in psychology, I grab a bunch of people and I take half of them, and I show them horrifying pictures of dead mangled bodies.

And the other half I show something else.

That by halves, lasting psychological effects from that.

So you got to clear that kind of thing, with higher powers.

And that's usually done by the Institutional Review Board.

For us usually, "Oh, I gave them testing technique A, testing technique B, and that group was so devastated that couldn't sleep for weeks, I don't see that happening. But we still in the US we have to do this process.

And experiments do consider confidentiality.

Also we have to avoid deceptions.

Now sometimes you can't avoid it.

Once you're experimenting with requires telling something else.

Usually here, we don't have to deceive, we may hold a lot of information, like the fact that they're control group for a while.

All that's about participants, and there's maybe instrumentation as I mentioned. Whatever tools you need, actually for us sometimes that's one of the hardest parts, I means one of the longest parts, building analysis tools, probably I can't tell you this, building the tools you need to experiment can take a long time.

Pilot studies and walkthroughs are very useful.

Human study, you are going to make three people in here and study them you got two hours to get it right, don't do that before you've already done one or two friends or something to improve your materials.

On tools, well, with tools you'll rerun them and rerun them maybe they take long time, so you'd like to get them right first too.

As long as telling about participants, in this class, some of you may do experiments, and any of you in this class who are asked about by someone else in this class, participated in this experiment, I would've appreciated if you agree.

It shouldn't take too long; I don't think you'll have devastating nightmares from it.

I think for you will going to do experiment, and so, well I can tell you about what experiment is, obviously, but it might be small experiment but hopefully you can get the other six of you participated in that.

That was preparation execution, when you run the thing, and all these depends on the experiment.

It could be, I bring them in to the classroom.

Spend two hours with them and I'm done.

Or it could be in a company I'm doing things over period of time.

This is the way you collect your data, fill out your forms, and all that stuff.

And while watching it, you got to watch out for possible confounding effects, like I mentioned, we had couple of case of computer crashed, and if we hadn't notice that,

we would've have some very limited transcripts, wondering why, and if we through that in our data, we'd being biasing our data.

And then validation, afterward, before you been analyzing, there's thing that you got to do to look at whether it's reasonable.

If you got all transcripts something obviously went wrong.

Or transcripts that ended very early.

If the log that you put says sick fall, obviously something went wrong in running the tool.

Outliers got pieces of data that are very very off.

Again subjects sometimes we get that bunch of people did perfectly fine, and you got results in this range, and then someone got absolutely nothing right.

Why?

Because they gave up in reading the news.

Or they just clicked false, false, false, false, false, false, false, false, and get the twenty bucks.

So you got to look at these things.

There got to be good reasons for removing outliers.

You know, it's not good reasons saying "well, I removed all the data points that make my theory false, and I was able to prove the theory, obviously that's cheating, right?

Then you can, with people, post-experiment questionnaires or interviews can be useful. Okay?

Let's go to the analysis.

Once you got the data, there're various things to do.

There's what we called descriptive statistics.

Which is giving let me see there're some graphs of some of these.

Ah, I got the next one.

Descriptive stats, data set reduction, and hypothesis testing.

These come up in the next slide.

So descriptive stats, they're not statistical data, but they're ways to look at the data.

And you got all these data and you like to head away to show people, something relevant to all the data before you go and say all the averages of statistics were significantly different.

And so, you'll see the various types of graphs used, and you'll see in some of the papers we look at.

You've been probably seen this before anyway. If you're looking at the fact of one variable verses some, no this is probably independent variable, and then dependent variable, like a and the level of coverage verses false detected, or a number of test weeks, we scatter plots, plots and plot points, right? And now you can fill the line and then get some sort of statistical thing after the fact.

But it shows people the data. And you probably seen, you will see if you haven't seen box plots, bar charts, and pie charts.

Various types of charts you can use. And usually in a results section I start with that.

I could say, for review of all the data see figure one which shows box plots that divide that show for each group what their false detection affects each box, and the range of results.

Now you move on the statistics.

So in the box plots may look like ones better than another. But now it's statistically the case they are when you get on to hypothesis testing.

But the first thing involves data set reduction.

I talked about outliers earlier.

Hypothesis testing depends on the quality of data and so anomalous data should may be removed.

And if there are outliers and you got a good reason to remove if you can.

And something like scatter plots can help you find outliers and there are statistical tests that can use some math like if a point is more than two standard deviations away from main, considered outliers.

I'm not going to get into this point.

That's too complicated right? Then we talked about this already.

You apply the stats test.

Can we reject the null hypothesis and except the all the alternatives?

If we can't reject it, we really can't draw any conclusions.

We can't say that if the null hypothesis is that we can't reject it, we can't say "Well, alternative doesn't hold, we could just say we don't have enough evidence."

If we can then we got knowledge false with a given significance.

And the test give us P value which show us the lowest significance which we can reject them.

So typically so in the test you'll say we'll say alpha, the significance level, we often use 0.05 and the P value is less than 0.05. We say we can reject the null hypothesis.

You got stats tools to do this anyway.

So you don't have to do it by hand.

Which test to use? This is just a table, tabular form of what we're seeing by individual designs a while ago.

It's just an example, there're some other tests, too.

In parametric tests, half of when the data does follow a normal distribution.

And if you look at the data, and can show it does, then all of these are available to you.

And if it doesn't, non-parametric tests.

It's important to choose the right type.

There, by the middle, is data normally distributed? Are data items paired? Paired happens if I have, if I made pie technique one and technique two, two something, two something, then it's paired. I could directly compare the means of differences.

If I'm splitting thing up in half, half have technique one and half have technique two, then it's not paired.

Pace is a lot going on here and there in entire courses obviously in books and packages you can use.

Now statistical significance shows that the difference between A and B is not due to random chance. Up to my best probability value but suppose I find them, yes, technique A detects 1% more false than technique B, and they are equal expense, well if they are equal expense then you'll prefer A.

But is it really practically significant? Does it matter in practice?

And you'll get that comment from reviewers, you can get the statistical difference in a case where difference does make a difference in practice.

And so there are ways to look at that.

Things call effect size. That's all about to say right now. And that brings us to pretty much at the end of these.

How do you package an experiment? Here's a standard outline relatively standard outline that most of mine falls on many many experiment falls this.

Of course experiments begins with intro and background. Intro and motivation and you got the background and then comes to your study.

And in the study, you got the research questions that motivates everything, okay, they come first.

Cause everything else can be described as the chose made by let you examine those questions.

What were your objects, people and objects, which are here it is, or both.

What's your variables and measures, independent, dependent and maybe other factors that you'll going to block over. What's your setup? One factor, two treatments etcetera.

Any operational details which could be, to conduct these we need such and such environment, or we ran all these on a particular architecture or particular memory requirements.

Sometimes that's here, and sometimes it's not. that might be detailing types of statistics you use.

And then threats to validity, and then you could do the data and analysis. Now this data and analysis, is quantitative analysis.

It's for you say, maybe first visually here's our data is, box plots whatever, and here's our statistical tests, and you're trying to avoid interpret too much here.

Trying to say, what this means is practitioner should use this.

Just stick to the data.

Nothing in here should be controversial.

It's just what the data shows.

Should be no questions about that, well unless you continue use the right test.

This tough could be controversial and that's just as important is we're now we have is what is it mean? What is it imply? What is it imply researchers, what is it imply for practitioners? Is there any qualitative stuff I could draw out of this? Besides the data?

That goes in here. Okay?

Couple things not shown here, if you're presenting a new technique, well there might be all section on that, before you get to the study.

Probably be up there.

And now you study your technique again, something.

And related work, that depends on it too.

We sometimes put it up there, sometimes we put it back here, all depend on how much there is, or whether people need to understand, there's also study before the result, before they can read related work, or not.

So that's the choice things. Okay, that's end of that.

No questions about this or anything? Some of you will be doing the studies or study related things? Okay?

There are few trade-offs.

One trade-off is there's sometimes so much related work that I don't want to delay the reader to getting to the new work by putting up there.

Sometimes understanding all the details of everything, everyone has done before, they don't need details yet.

It doesn't mean I won't mention any of that.

Maybe the intro. And intro probably briefly say, these questions haven't looked at before by Smith Johns, by Churchill related work section C, sections for details.

Okay? You'd like to tell people aware of that stuff.

But spelling out the full relationship to it can maybe wait.

One time you really has to wait is, suppose you got a new technique in there.

That you're developing, which could be between section two and three.

It's hard to contrast your technique with existing ones, before explaining what your technique is.

So in that case you almost have to give the details later.

So you follow, my technique is this, dependencies is this, uses that, and the results give us this.

Now the later working say, well, Churchill and whatever did technique, but he didn't use the dependencies the way we did.

Someone sorted the work but theirs was unable to handle the things that we handled our empirical study. Okay?

So I can't give you precise rules, but that's, those are things that you think about it.

Another example though is sometimes when you're doing the paper that is primarily empirical study, you do want to talk about what is known empirically prior to these.

So I've done a couple of paper, we're looking at the empirical questions, where to motivate importance of my questions, I had to first say "In prior work, Smith study, the effect of A on B and found this.

And then Johns went on found this.

Neither of them looked at the factors, and you talk about the factors that you're looking at.

So it's really question of how much related work to people need to understand before you talk about your stuff to grasp the significance or importance of it.

And then afterward it's, afterward is, sometimes you just put both places, so I know I can't give you a rule.

But there's thoughts that go into it.

Other questions?