Quantitative tools with Excel and R

In this module, I'd like to talk about some quantitative tools. In particular, we are going to focus on the tools of Microsoft's Excel and the statistical package called R.

Slide 2: Some statistical concepts

First, we're going, to begin with, some basic statistical concepts.

Slide 3: Independent versus dependent variables

The first two concepts are independent variables and dependent variables. An independent variable is a variable you can change, while a dependent variable is going to be a response or an outcome. So we change the independent variable, and we observe the dependent variable. The dependent variable is typically what you're going to be concerned with measuring. You're looking at the outcome after you have made the change to the independent variable.

Slide 4: Types of data

There are a number of different types of data that we can be concerned with. Most typically, we're concerned with nominal data. So we've got an unordered group: male and female, left-handed and right-handed, etc. Or an ordinal type of data. Where we now have rank order. So the difference between item number N and N+i doesn't tell us anything other than one is ranked ahead of another. But it gives us an order. So, if we look at the top five hundred universities, the top ten protocols in bytes, etc. Just knowing can see the top ten protocols in bytes - we don't know what the relative differences are the amounts - we simply know that one is larger than the other. We can also have interval data. Where we have continuous ranges map to some particular scale, but we don't have a clear zero. We can also think about ratio data - it's like the interval data, but now we have a clear absolute zero value.

Slide 5: Metrics

So, if you think of these different types of data, and we see them over here in the first column. We can think about metrics to use for them. So, for example, for nominal data that we can think of metrics like success and failure. And the common statistics we will apply are things like frequency analysis and chi-squared tests. Or for ordinal data, we can think about ranking. Not surprisingly, we can use again frequencies and chi-square, but we can also think in terms of Wilcoxian rank-sum tests or Spearman rank correlations. For interval data, for example, it's quite often the case that we have a survey, we might have five different levels: I completely disagree with it, I disagree, I don't really care - either way, I really agree, I agree somewhat, and I really agree with it. These five levels might be a five-level Likert scale. We might use system usability scales, MOS scores, etc. And for these types of data, we can use descriptive statistics, things like averages, medians, [and] standard deviations. We can apply tests to them like Student's t-test or ANOVA or correlation or regression type analysis. And

for ratio type of data, things like pass completion time or packet interarrival time, we can apply all of the above, but now we can often apply to it the geometric mean. And, if we look at Tullis [and Albert] book - you can find lots more examples about this.

Slide 6: Measures of Central Tendency

Now, the three most common measures of centrality are the mean it's simply the arithmetic average. The median is the midpoint, where half the values are larger and half the values smaller. Or mode - the most commonly occurring value. And it's very important that we look at the difference between these because quite often we will find that computing the mean it gets pulled way off where we expect because one of the values is very very large or one of the values is very small. So, when we compute the arithmetic mean, it pulls us in that direction. That's often why in much of our data we look for the median - because we want to understand where is the point where half the values are bigger, and half the values are smaller.

Slide 7: Selecting participants

Now, if you're selecting participants for some activity, you need to figure out how you're going to select them. So, for example, you might do a random selection, or you might systematically select every third person who comes through that door - we're going to ask them a question. Or a stratified sampling, choosing members from representative sets of the data. Or one of the most common, of course, is samples of convenience: Who can I get to actually take my survey? And then I have to figure out if they are actually representative of the target population or not. If they are, maybe it works. If they're not, it may not give me any useful data.

Slide 8: Sample size

Next, I need to think about sample size. Right! First, I have to think: "What's the goal?" And I have to think: "Is the difference that I expect to observe large or small?". Because if the difference I expected to observe is very large, I don't need very many samples to find that difference. But if the difference I'm trying to detect is very, very small, then I need a lot of samples to be able to find that. And of course, finally - I need to think about "What is my acceptable margin of error?" If I'm allowed to have a large margin of error, then I don't need a very large sample size. If I want to have a very small margin of error, then I'm going to need more samples.

Slide 9: Within- vs. between-subjects

Next, I need to think about: "Am I doing comparisons within or between?". Within - that means that I'm concerned with what's called repeatability measures - it's the same subject but repeated measurements. So, I have a person and a task, how long does it take to do that the first time, the second time, the third time, etc. Or I may say, "Ah! it's about a comparison between subjects?" How long does it take person A to do it versus how long does it take for

person B to do it? But now, I have a problem so-called carryover problems - the method that they applied the first time may lead to learning - so when they apply the second method, it takes them less time because they actually learned things from the first time. So, we need to be very, very careful. So, then, of course, we make sure that we have another group that does them in the opposite order to compensate for that. So we need to think about avoiding the carryover effect, and of course, we can go very sorts of mixed designs.

Slide 10: Counterbalancing

Now, to avoid the carryover effect, we can use so-called counterbalancing techniques - by randomizing the order. You choose this person to be part of a group that does that task method one first. You choose the next person statistically to use the other method. We think about pre-defined orders and then randomly assigned people to those different groups, etc.

Slide 11: (Starting) Quantitative analysis of survey data

So, now we will talk about quantitative analysis of survey data.

Slide 12: Overview

The first thing we have to think about as Gillian Raab, Professor of Applied Statistics at Napier University, says, "What's the process of carrying out the survey?" Now, this is his view as a statistician. It begins by we designed the survey, then we conduct the survey, we collect the data, we analyze the results, and now we evaluate bias and precision, and then we repeat the process. Right. It's relatively straightforward, except- we need to think about another element. And that is: "Why are we doing this?" The reason we're doing it is: We want to be able to gain insight. And therefore, we have to convey the insight to decision-makers are going to make some decisions based upon the analysis of the results of the survey. So, it isn't that we just say, "Ah! Let's conduct a survey." In the end, we wanted to have some effect.

Slide 13: Objective

So, what's the object of it? Well, the basic object might be providing a predictive model - we'd like to understand if the data is sufficient to be able to give us a model. If so, then we can reduce the amount of data that we need to collect in the future. Or is it because we're trying to find hidden relationships. So we might think about segmenting a population into different strata, and then visualizing the results of our responses. For example, does the distance to the park affects the frequency of the visits to the park? So people who live further from the park come less frequently than people were near the park who come very frequently. Why do we need to know that? Well, we might need to be able to make a decision about where we should put parks. If people only come for short distances, then we need to have lots of parks that are located within short distances- If people come longer distances, (yes), then we can space them out further apart. So, we have to think about what is or what are the research question or questions as we start out to do our survey.

Slide 14: Considerations when designing studies

So, Kelley and Maxwell said, when you're designing a study at the minimum, you have to consider the following points when you're doing behavioral, educational, and social sciences study: What is the question of interest? What is the population of interest? What is the sampling scheme that we're going to apply? What are going to be our independent and dependent measures? Should we do experiments or simply observations? What statistical methods are we going to choose so that we can actually answer the question in an appropriate and optimal way? We don't bother more people than we really need to - to be able to get the data that we're interested in. And of course, we need to make sure that we actually have a sufficient sampling plan, so were going to get the data that we need. We need to think about how long the study can take place. So, for example, in this course, a very important thing is you need to be able to get it done in such a timely fashion that you're able to actually go and analyze the data and produce the final report within the time period for this course. And of course, in most practical cases, you are going to have to think about the financial cost and the feasibility of the study that you're going to propose before you start to do it. Because otherwise, you're not going to be able to complete it.

Slide 15: Questionnaire Research Flow Chart

So, if we think about designing a questionnaire, we start in the upper left-hand corner here, and we need to design our methodology. So explicitly, we need to decide what's going to be our design methodology. Next, we have to decide is it feasible? Is it feasible to apply that design methodology? If so, then we're going to develop our instruments. In this case, the questionnaire, and now we're going to select our sample population. We're going to conduct a pilot trial. We will revise the survey. And now that we have a good survey instrument, we're going to now actually go and conduct our research. A big mistake to be made here, is if we don't pilot it - because if we don't pilot it - we don't know: Is it actually a good questionnaire or not? And we don't want to spend lots of time and bother lots of people whenever actually the questionnaire isn't very good. So remember: Do your pilot studies! When you conduct the research, then, of course, you have to analyze the data, and many people forget that it's going to take a lot of time to analyze the data. Include that in your planning. And then, of course, finally we need need to prepare the report so that now we can inform others of what it is that we have learned as a result of using this survey of all of these people - who put time and effort into answering it - and so that we can maximize the value of that effort.

Slide 16: Sampling methods

Now, not surprisingly, there many different methods of sampling. We talked about some of them earlier, but, typically, we can split them into the following categories: probabilistic methods - such as random sampling or systematic sampling - or very commonly sampling proportional to size. You take that population, we say, "Hmm! We stratified it into different segments, and this segment is twice the size of that - maybe we only need to ask half the number of people in this population as of that to balance- probabilities across the different

segments?" Or we might need to do stratified sampling to ensure that we actually have some samples from every subpopulation that we're interested in. Conversely, we have non-probabilistic methods of sampling, such as accidental or haphazard or convenience sampling. But here we run the risk that that may not be representative of our target population. So, in the past, for instance, students have asked other students questions - forgetting that the actual population for the study that they were trying to do wasn't about other students and so, therefore, the data that they got was really not the data relevant for the target population. Oops! (yes) wrong data! Irrelevant results - perhaps. We can also think about purposeful sampling like modal instance sampling - what's the most typical case. Or expert sampling, we choose just our experts -- because we believe we will get the most information out of them in the shortest period of time. We can think of quota sampling - proportional or non-proportional sampling. Or heterogeneous sampling - we want the greatest diversity possible. Or one of the most common methods is snowball sampling - you get recommendations from the people who you had do your questionnaire, and they say, "Oh yeah! One of my friends would be really interested in participating in your study also" - and they give a copy to [them], and it spreads out that way in a social network kind of fashion.

Slide 17: Sample size

So back to the problem about size. Now, statisticians talk about statistical power, and for details, you can read this reference at http://www.socialresearchmethods.net/kb/power.php. But basically, you can think about is: What are we trying to do? We want to choose a sample size so that we can get our signal large enough above the expected noise that we can see our signal - with some given level of confidence.

Slide 18: Getting started with data analysis

So, as you start your data analysis - What do you need to do? You need to first begin by saying, "I need to get the data into the computer system" I better have thought about how I'm going to do that! So be very, very careful that you optimize getting your data in without having to sit there reading off pieces of paper, typing in the data yourself. It's prone to errors. It takes a lot of time. And you regret doing it. So, make sure you can get your data. The next thing is you're going to do a preliminary analysis. You're going to compute some basic descriptive statistics mean, median, max, min, etc. And then, you want to do a so-called exploratory data analysis. And this generally involves generating some plots, as points or lines or scatter plots or histograms or whatever. But it's to help you start to get a feel for what's going on in the data.

Slide 19: Types of analysis

So two major types of data analysis: design-based analysis, in this case, the approach is that the randomness is induced by the random selection of the participants or the assignment of the participants to a given subset. That's where the randomness came in. And then you're going to choose a statistical model to do model-based inference. So you're going to say, "Okay! Here is my model, let us fit it to my data" But the randomness came from the

selection of the people, [hence] you believe that there's a model there so [it is] design-based analysis. In the model-based analysis, the idea is the randomness occurs because of the innate randomness in the measurements themselves. So we have a model, we applied, for instance, a set of surveys to a population, we got a set of responses, we believe that there is randomness in there - but the randomness is inherent in the measurements - we couldn't remove the randomness. There's going to be randomness no matter what we do.

Slide 20: Modeling techniques

There are various modeling techniques that we can apply, things like Bayesian networks or trees, neural networks, regression analysis, etc. We can do clustering of the data. We can think about the data in a high dimensional space, and we can look at the points the cluster towards each other, then we can think about segmenting that space into different pieces. We can think about fitting data to an a priori model. We say, "Hmm! The model should be roughly like this, now what are the coefficients". We can also think about using techniques like factor analysis or principal component analysis where initially we don't know what are the components - so we need to figure out which is the most important component, which is the second most important component, etc.

Slide 21: Weights

Now, we might need to think about weights. Why? Because we could only sample each of the populations that we wanted to some level, and therefore when we get our results from analyzing that subset, we now need to weight that to be able to fit it to our model over the entire set of data. So if these people represent ten percent of the data, then, of course, it should have ten percent of an effect upon our overall conclusions of our model. And Chromy and Abeyasekera have written this book called "Statistical analysis of survey data" that you may find interesting for more discussion about weights.

Slide 22: Significance

Well, in professor Smith's lecture, he talked about significance. But when statisticians use the word "significance", they don't mean "is important or not". Statistical significance concerns your confidence that your conclusion is actually representing the real difference, i.e., the result is unlikely to be due simply to chance. So I want a 95% confidence in this, I need to make sure that there's enough significance that (yes) this result isn't caused by chance - it's caused by a particular thing that I'm studying. We also need to look at: Is the distribution one-sided or two-sided? And use the appropriate test with this. That as my wife constantly reminds me, when you use the word "significant" in a statistical context, it doesn't mean it is important or interesting or meaningful. And similarly, not all observations that are not statistically significant are unimportant or uninteresting. So, when you use the word "significant", be clear are using it to talk about statistical significance or not.

Slide 23: Testing for significance

So how can we test for significance? Well, the first thing we need to decide on is the significance level, in this case, alpha, then we need to calculate the statistical variable p. and then we say if p is less than alpha, then the result is significant; otherwise, it's not significant. Now for those of you who like to think in terms of signal to noise ratio, it is very simple to understand the confidence is equal to the signal to noise ratio times the square root of the sample size. And immediately now it should become apparent to you that if I have a very small signal and a lot of noise, then I need a big sample size. And to get twice the confidence, I need the square of the sample size for a given signal to noise ratio. So, either think of it in the statistical sense or think of it in the signal to noise [ratio] sense. Voilà, you can start to get a feel for what it means for something to be statistically significant.

Slide 24: Next steps

So what are the next steps? Well, you need to go search in the literature, and you need to read extensively. But don't be afraid of consulting with a statistician. The whole advantage is their expertise in statistics. Now, in most cases, this is going to cost you some money or dinner or something else, but it will save you a lot of time and effort in the end. Now, you can also think about doing some of the statistical analysis yourself; even if you're going to go and consult with the statistician, you need to spend some time during the statistical analysis yourself so - you'll be able to have an intelligent discussion with the statistician.

Slide 25: Using Excel for statistics and plotting

So, how can we think about getting some statistical analysis done? One of the first these, of course, is to use Excel.

Slide 26: Experiment 1

And one of the easy things, we can do, is take a look at an experiment, in this case, I captured packets using Wireshark for a 2150.12 seconds long Voice over IP call. And during this call, there a 107,505 RTP packets in each direction and another 429 RTCP packets in one direction. So, the RTP packets (of those packets) containing the voice and the RTCP packets were are carrying information about the RTP traffic. And these were transferred in only one direction.

Slide 27: Load the data, then extract relevant RTP packets

So the first thing I did was, I exported the data in a tab-separated form. I then extracted the data that I was interested in looking at a particular destination address while looking at a particular protocol. Then I sorted on the column protocol destination, and I moved the results into the spreadsheet. So, from this data, I could calculate things like what's the mean and inter-arrival time.

Slide 28: From network to local user agent

So, I can look at it [in terms of] the difference in the clocks from the previous sample, or I can look at it in terms of seconds. I can look at what the standard error is. I can look at the median, the mode, the standard deviation, the sample variance, kurtosis, skewness, etc. And skewness, tells us about the shape of the distribution. I can look at the range, the minimum, maximum, ... I can sum up all the values. I can look at how many there were. And I can look at my confidence level at 95%, and I can see that I have this data with a hundred thousand samples, and I have a confidence level of about 1.8 x 10-7. So not surprisingly, with a hundred thousand samples - I'm fairly confident in this data - because I have such a large number of samples.

Slide 29: First look at the RTP clock (Time) differences

So, if I take a first look at this data. And in this case, I plot the time in RTP clock units versus the relative frame number, and I see it's absolutely flat at 160 audio samples per frame. And if I look at my data that is consistent with the fact that I'm sending 20 milliseconds worth of sample data at 8 kiloseconds per-sample, [SORRY] 8K samples per second. That leads to 160 samples being in each one of these packets that I'm sending. And that's exactly consistent with the ITU-T G.711 pulse code modulation with A-law coding CODEC that I'm using.

Slide 30: Plot RTP inter-arrival times as measured by Wireshark

But if I now plot the inter-arrival times between those packets. So in this, we saw that the frame time was 20 millisecond - so every twenty milliseconds, I should be sending a frame out. But did they actually with 20 milliseconds between them. And if we see, here in the middle, we can see that quite a lot of them did. But they did actually arrive. Somewhere they were separated more, and somewhere there is less separation.

Slide 31: Compute histogram of inter-arrival times

So, if we take a look at that, we can plot a histogram, then we can now see the very, very large fraction of them - were just a little bit longer/further apart than are expected in 20 milliseconds (this is in units of seconds). Yeah, we see a little spike down here and a small little spike there, and it's a little bit skewed, but we would say, "(yes) it's around 20 milliseconds".

Slide 32: Plot Cumulative Distribution of inter-arrival times

We plot the cumulative distribution average; we can see that (yes), in fact, the median here is at 50% - which is right at 20 milliseconds as we would expect.

Slide 33: Add grid lines

Now, if we add grid lines, it becomes a little bit easier to see that data.

Slide 34: As numbers - near median

And if we look at our statistics data of those frequencies for those bins. We see that the median is down here at 20.004 milliseconds that was the frequency. 43.37% of the data was in the same bin as the median The mode (the most commonly occurring value) was just a little bit longer - one microsecond longer and that was 59.30% of the data. But if we look at the mean, we see that the mean is way down here at 0.019999 seconds (sorry) 19.999 milliseconds, But that represented only 13.95% of the data. So we see that there's a pretty big issue here - the data isn't arriving at the twenty-millisecond separation, so I think some are arriving a little sooner and some are arriving a little later - shifting our median value out just slightly. So half the data takes longer than this half the data takes less.

Slide 35: With varying numbers of samples

Now, we can say, "How does this vary with the number of samples?" So, if I were to take just the first hundred values that I measure, what's the name mean value? And we can look at this chart, and we can see as we go from a hundred to a thousand to ten thousand to the first hundred thousand, we see it doesn't affect the mean very much. We see that the standard error goes down quite a bit as we increase our number of samples, but we noticed the median is almost unchanged. That means that if we're just simply looking for the median in this set, we only needed actually a hundred samples to get that rather accurately - taking a hundred thousand samples didn't really change that by much. And the mode is unchanged. And the standard deviation, of course, decreases as we increase the number of samples. We can look at the variance - it doesn't change very much irrespective of how many samples we take. We can look at the shape and skewness, the range, etc. And we see our confidence level, however, as we increase the number of samples (yes) our confidence changes as you would expect. The probability that we have a different value does change as we increase the number of samples, exactly as you would expect. But is that really giving us the data we want?

Slide 36: Zooming in on interesting behavior

If we look at the inter-arrival times of a series of frames, we see that it looks like that. Note: We're putting this is a scatterplot. Now, if we rescale it - we can see roughly the range over which this is changing. But does that really give us insight into what's going on? Now.

Slide 37: Looking in more detail at a relatively “flat” region

So let's look at a relatively flat region that was an area where there was a lot of difference. We look in here; we say, "Hmm! This really looks like it's pretty uniform" But on closer examination, what we see is it actually looks like this with these three bands that we saw. Have a time when there are roughly equal, and then we see some high so in the middle and some low.

Slide 38: Is there some pattern?

Is there some pattern to this? Well, if we connect these lines - these points by lines - we can see that it's alternating. We have a high value; it is followed by a low value, followed by the value we expect, followed by a high value, followed by a low value, a value we sort of expect, high, low. We see that basically, these are compensating for each other. So, yes, they are about 20 milliseconds apart, but if some get delayed, that means it decreases the delay until the next one, while it increases the delay from the previous one. And of course, since they continue to come along at 20 milliseconds, that means that somehow they have to average out to be as we expected 20 milliseconds.

Slide 39: Adding grid lines

And here we see this variance. But it's relatively periodic, and we could look at the bursts of these periods and try to understand why that is occurring. But it wasn't sufficient just to add grid lines to this plot to be able to see the variance.

Slide 40: Are grid lines alone sufficient?

No, it's really a very, very hard just seeing those grid lines to see that this occurs, then this one occurs, then this one occurs, then this one occurs. But by adding these lines, it was very, very, very obvious. So something that we normally wouldn't do, which is connecting individual data points with straight lines, actually gave us more insight.

Slide 41: Scatter plots of frame # versus time

So in this plot, we have plotted frame number versus time. And we see that the frame number basically linearly increases with time.

Slide 42: Zoom in on last few samples

Not surprisingly, if we zoom in on some of those samples - we see these individual samples.

Slide 43: Add a trendline and show equation

And we can, in fact, fit them to a line, and see that, yes, it is very linear. It has an R squared coefficient of one, which means that it exactly fits this relationship: Y is equal to 0.01 times X plus 17671. So that means the timestamps that are occurring in these RTP packets are exactly indexing along as we expect, not surprising because they should be doing that.

Slide 44: Computing new axis

If we look at the relative sequence number divided by 160, what do we see here? What we see a curve that goes up like this.

Slide 45: Now add the trendline

And now when we fit our equation to it - what do we find: we see that Y is equal to 0.020 X plus six times ten to the minus five. That means it is twenty milliseconds with only a five-microsecond difference from what we expect.

Slide 46: Experimental vs. expected

So it fits very very well, and if we compare our experimental data with our expected data, we see they fit extremely well.

Slide 47: How does the measured data differ from the expected data?

But does it really matter? So, in this case, we plot the relative sequence number versus the difference of time in seconds. So this is the difference between our expected and measured. And we see these differences are very, very, very, very small. The largest is in the neighborhood of about 160 microseconds. So since we're looking at events that are occurring every 20 milliseconds, the fact that the difference between the expected time at the time that we actually received the packet is off by at most 160 microseconds, we can say, "No that's actually not relevant to us".

Slide 48: Does the difference matter? Plot scaled to 1/10 of the inter-arrival time period ⇒ No

And we can see that more easily if we scale it up to the range where there is now a 1.9 [ms] , so that's a tenth of the intervals between these packets and we say that the differences are way way down here it's clear those differences don't matter.

Slide 49: For traffic in the opposite direction

If we look at the traffic in the opposite direction, we see these statistics.

Slide 50: Uplink inter-arrival times

But if we look at the inter-arrival times, we see something very different. We see that it goes along very flat and then suddenly there's ones with very, very different inter-arrival times then it is very flat there is a burst here, there's a burst here. So what is causing this?

Slide 51: What is going on?

Well, if we look at when those spikes occurred, they occurred at these frame numbers - that represents this relative difference in time in seconds. And we noticed that the difference in the time here is roughly 600 seconds. And the answer to the question is: "What's going on every 600 seconds?" Well, it turns out that's going to be when a DHCP request occurs. So since the box this is going through is an analog telephone adapter every 600 seconds, it is going out and making a DHCP request. When it does that, it disrupts the inter-arrival time of

the audio packets by quite a lot - we see here it's 25 milliseconds. Our packets for only twenty milliseconds in size, so this is a very substantial change to traffic.

Slide 52: RTCP descriptive statistics

We can look at the RTCP descriptive statistics.

Slide 53: Plot of inter-arrival times of RTCP reports

We can look at the inter-arrival times of those. And now we see that (yes) there these spikes in the RTCP inter-arrival time statistics.

Slide 54: Histogram of RTCP inter-arrivals

We can plot a histogram.

Slide 55: RTCP CDF

We can compute a cumulative distribution function of them. So with Excel, we are quickly able to say that we can find a lot of interesting things that are data fairly quickly.

Slide 56: Remarks

However, it isn't very easy to change, add, or subtract data points without having to redo all of the analyses. That's a difficult thing. So if I wanted to apply this to another data set, I would have to go through a lot of manual things, or it would have to write functions in Excel to be able to do that. Well, as I said in another lecture module I'm very lazy, so I'd instead like to figure out: "Is there a language that I could use to be able to write procedures to apply to my data to make things simpler?" and that language today is R.

Slide 57: R

The reason for R, which is the successor to the closed source statistical package called S and S-plus, and S was developed by statisticians at AT&T Bell Laboratories to be able to help them help others with their problems. So it is a tool built by statisticians. Well, Josef Freuwald, when he was a graduate student at the University of Pennsylvania in linguistics, now Lecturer in Sociolinguistics in Linguistics and English at the University of Edinburgh said, "quite simply R is the statistics software paradigm of our day". Norman Matloff, in his book called The "Art of R Programming: A Tour of Statistical Software", says, "As the Cantonese say, yauh peng, yauh leng, which means both inexpensive and beautiful." But one of the most important things I believe today of why you would want to use R is that R is what statisticians use. So the latest statistical tools (yes) they're available in R. When you talk to a statistician - if you can have your data in R and if you know how to analyze it in R. You're going to find it much more efficient in communicating with your statistician and they communicating with you.

Slide 58: Commercial alternatives to R

But there are lots of commercial alternatives. As we saw Excel, there is also MathWorks' MATLAB – Statistics Toolbox™. There is Statistical Analysis with SAS/STAT® Software. IBM's SPSS® Advanced Statistics. There are lots and lots of tools out there. The key is finding a good tool that you like to use and learn to use it well.

Slide 59: R Resources

Now for R, there's a so-called comprehensive R archive network CRAN. And there are lots and lots of tools there that statisticians have produced. There are lots of tutorials, and Matloff, as I mentioned earlier, has a particularly nice set of tutorials you can go and read those.

Slide 60: R Packages

Now, I mention that statisticians are using R, and one of the key things is people produce so-called "R packages" for doing sets of functions. So, for instance, for plotting, for accessing databases, for computing splines, etcetera etcetera. And you can get a lot of packages, including finding out about commercial packages for R at the URL on the slide.

Slide 61: Why use a programming language versus using a spreadsheet?

So why use a programming language instead of a spreadsheet? Well, the answer is when you want to do something over and over again or would you want to do it very systematically - it's much easier doing this with the programming language than manually having to click and poke again and again and again and again and then realize "Oop! I have to do it another data set all again tomorrow".

Slide 62: Experiment 1

So let's look at that same first experiment.

Slide 63: Load the data, then extract relevant RTP packets

We take the same data captured with Wireshark, and we now extract the data.

Slide 64: Summary

All we have to do is simply say, "summary" over the set and automatically R dumps out all of this data: our minimums and maximums of the various different columns in our data. Here the list of protocols in ranked orders. Our first and second quartiles, our first and third cartels, are mean and median numbers. As I say our min and max for all these different things. And there's our information about it. [CLICK] Really simple!

Slide 65: Inter-arrival delays

So, if I want to compute inter-arrival delays, I can assign a variable to say, "give me the number of rows", that's how many data samples there were, subtract one from that. Now apply it and create a vector, here the delay values iterate over them, computing the differences between them. And now that I can say "summary" over that value, I just computed. And sure enough, I now see my min, my max, first and third quartiles, mean and medians.

Slide 66: plot(To_Chip_RTP_delay)

I can plot it all by just simply saying "plot" and violà - that's a comparable plot to the plot we saw before - in data.

Slide 67: hist(To_Chip_RTP_delay)

Now, we can, of course, compute the histograms from this by simply saying "hist" of that value.

Slide 68: boxplot(To_Chip_RTP_delay, pch=20, col=3)

We can also compute boxplots of this. The box plots are particularly useful because they now show the line here at our median, we see the upper and lower quartiles, and now we can see the outliers.

Slide 69: Interarrival delay vs. sequence

Now, we can pass arguments to this about what symbol we want to use, what color we want to use, we can choose different graphics devices, etc. But boxplots are immensely useful.

Slide 70: RTP Clock vs. sequence

So, if we go back to the inter-arrival versus delay sequence number - we plotted it using this plot command. We can now very easily compute the RTP clock versus the sequence number, and we see that it cycles this way. And that's because the sequence numbers is a sixteen-bit field and sure enough when it gets to 64K, it moves back to zero and continues on up. It behaves as we expect.

Slide 71: Inter-arrival times of RTP packets: From network to local user agent

We can, of course, compute the same statistics as we did with R using a set - in this case of R functions - we get the same values out (basically)-

Slide 72: R vs. Excel histogram

We have an R histogram, our Excel histogram. This is really very straightforward

Slide 73: Plot as a Cumulative Distribution (CDF)

We can plot the cumulative distribution function. We can put major labels on. We can put labels on our axes. We can choose what symbols. We can put a grid in the background if we want, etc.

Slide 74: With varying numbers of samples

Now, it's really easy to compute the statistics over the first hundred, thousand, ten thousand, hundred thousand.

Slide 75: With varying numbers of samples

Well, yes just define yourself a function to compute the statistics we want, and then

Slide 76: Applying a function to a list of arguments

calling the function to be applied to the function with a hundred, a thousand, ten thousand, or a hundred thousand samples will produce the columns. We are set. Off we go!

Slide 77: Uplink inter-arrival times stats

Now, one of the things that you're going to need to do, of courses, as you get out of either excel or whatever statistics package you get these huge big numbers, those are not useful. You need to figure out which of these digits are meaningful, and that's what you're going to put in the report. So, in this case, I'm going to say the mean value is 0.020000 seconds. Why? Because I know that I captured by data and the resolution of my clock is one microsecond - so there is no point in reporting something that has a higher resolution than what my clock actually had when I was sampling the data. So these extra digits over here aren't meaningful. And similarly, doing it for the other values. You note that decimal aligned them, which makes it much easier to be able to compare them, and I set them in a fixed-width font, which again makes it very, very easy to compare these numbers as I go down the column.

Slide 78: How does the measured data differ from the expected data?

So, if I plot the measured data versus the expected data, I get a plot like this. You can see it has these distributions that are over the whole set of values.

Slide 79: How does the measured data differ from the expected data?

We can compute it now as a histogram of the differences.

Slide 80: Uplink inter-arrival times

Again, if we look at the uplink inter-arrival time - you'd see the same pattern we did before.

Slide 81: For traffic in the opposite direction

We can compute the statistics on it.

Slide 81: References

There is a lot more that you can read about this. So I wish you lots of success in applying your tools, such as Excel and R, to analyze your data.

Slide 82: References

The book that I mention by Tom Tullis and Bill Albert, Measuring the user experience: collecting, analyzing, and presenting usability metrics. This is a particularly interesting book because these people are concerned about (yes) "How do I measure the usability of web sites?" because they worked for a very large corporation and they discuss how can you actually compute by conducting experiments based on the user's interacting with different versions of the web pages to optimize the webpage to provide [a] better quality experience for the users. Best of luck with your statistical analysis.