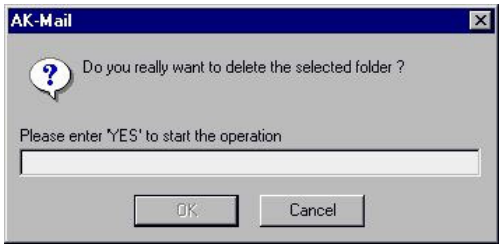
Lecture 17:
Experiment Analysis

UI Hall of Fame or Shame?



Source: Pixelcentric Interface Hall of Shame

Our Hall of Fame/Shame candidates for today are **confirmation dialog boxes**.

The AK-Mail dialog at the top is almost insulting to the user's intelligence. Furthermore, it certainly isn't **efficient**, since it's forcing the user to jump through hoops in order to complete an operation. Why was this done? No doubt for **error prevention**. Deleting a folder is evidently an irreversible operation, and potentially a catastrophic one if the folder is full of important mail. So the designers didn't want the user to trigger it accidentally. Hence the confirmation dialog. But they didn't want a thoughtless press of the Enter key to accidentally confirm the confirmation dialog, either. So this dialog box requires the user to put some conscious effort into confirming it.

But this effort isn't spent thinking about the **consequences** of the operation; instead, it's spent following the directions to type YES. In fact, the dialog doesn't even explain *why* the user is being forced to jump through hoops.

The dialog box on the bottom does a much better job. It explains the consequences and lets the user make a decision. To avoid accidental Enter presses, it provides a safe default, Cancel.

Even better than a confirmation dialog would be a **deferred operation** – don't execute the irreversible operation immediately (although the UI should appear as if it has executed). That way, you can make the operation undoable for a little while.

Today's Topics

- Hypothesis testing
- Statistical significance
- T test
- ANOVA

An Experiment

- Hypothesis: Mac menubar is faster to access than Windows menubar
  - Design: between-subjects, randomized assignment of interface to subject

Windows	Mac
400	360
220	210
560	500
340	305

Suppose we've conducted an experiment to compare the position of the Mac menubar (flush against the top of the screen) with the Windows menubar (separated from the top by a window title bar).

For the moment, let's suppose we used a **between-subjects** design. We recruited either users, and each user used only one version of the menu bar, and we'll be comparing different users' times. For simplicity, each user did only one trial, clicking on the menu bar just once while we timed their speed of access. The results of the experiment are shown above (times in milliseconds). Mac **seems** to be faster (343.75 ms on average) than Windows (380 ms). But given the **noise** in the measurements – some of the Mac trials are actually much slower than some of the Windows trials -- how do we know whether the Mac menubar is really faster?

This is the fundamental question underlying statistical analysis: estimating the amount of evidence in support of our hypothesis, even in the presence of noise.

Hypothesis Testing

- Our hypothesis: position of menubar matters
  - i.e., mean(Mac times) < mean(Windows times)
  - This is called the alternative hypothesis (also called H1)
- If we're wrong: position of menu bar makes no difference
  - i.e., mean(Mac) = mean(Win)
  - This is called the null hypothesis (H0)
- We can't really disprove the null hypothesis
  - Instead, we argue that the chance of seeing a difference **at least as extreme** as what we saw is very small if the null hypothesis is true

Our hypothesis is that the position of the menubar makes a difference in time. Another way of putting it is that the (noisy) process that produced the Mac access times is **different** from the process that produced the Windows access times. Let's make the hypothesis very specific: that the mean access time for the Mac menu bar is less than the mean access time for the Windows menu bar.

In the presence of randomness, however, we can't really *prove* our hypothesis. Instead, we can only present evidence that it's the best conclusion to draw from all possible other explanations. We have to argue against a skeptic who claims that we're wrong. In this case, the skeptic's position is that the position of the menu bar makes *no* difference; i.e., that the process producing Mac access times and Windows access times is the same process, and in particular that the mean Mac time is equal to the mean Windows time. This hypothesis is called the **null hypothesis**. In a sense, the null hypothesis is the "default" state of the world; our own hypothesis is called the **alternative hypothesis**.

Our goal in hypothesis testing will be to accumulate enough evidence – enough of a difference between Mac times and Windows times – so that we can **reject the null hypothesis** as very unlikely.

Statistical Significance

- Compute a statistic from our experimental data
  X = mean(Win) − mean(Mac)
- Determine the probability distribution of the statistic assuming H0 is true
  Pr(X=x | H0)
- Measure the probability of getting the same or greater difference
  Pr (X > x0 | H0) *one-sided test*
  2 Pr (X > |x0| | H0) *two-sided test*
- If that probability is less than 5%, then we say
  - "We reject the null hypothesis at the 5% significance level"
  - equivalently: "difference between menubars is statistically significant (p < .05)"
- Statistically significant does not mean scientifically important

Here's the basic idea behind statistical testing. We boil all our experimental data down to a single statistic (in this case, we'd want to use the difference between the average Mac time and the average Windows time). If the null hypothesis is true, then this statistic has a certain probability distribution. (In this case, if H0 is true and there's no difference between Windows and Mac menu bars, then our difference in averages should be distributed around 0, with some standard deviation sigma).

So if H0 is really true, we can regard our entire experiment as a single random draw from that distribution. If the statistic we computed turned out to be a typical value for the H0 distribution – really near 0, for example – then we don't have much evidence for arguing that H0 is false. But if the statistic is extreme – far from 0 in this case – then we can **quantify** the likelihood of getting such an extreme result. If only 5% of experiments would produce a result that's at least as extreme, then we say that we reject the null hypothesis – and hence accept the alternative hypothesis H1, which is the one we wanted to prove – at the 5% significance level.

The probability of getting at least as extreme a result given H0 is called the **p value** of the experiment. Small p values are better, because they measure the likelihood of the null hypothesis. Conventionally, the p value must be 5% to be considered **statistically significant**, i.e. enough evidence to reject. But this convention depends on context. An experiment with very few trials (n<10) may be persuasive even if its p value is only 10%. Conversely, an experiment with thousands of trials won't be terribly convincing unless its p value is 1% or less.

Keep in mind that **statistical significance does not imply importance.** Suppose the difference between the Mac menu bar and Windows menu bar amounts to only 1 millisecond (out of 350 milliseconds on average). A sufficiently large experiment would be able to detect this difference at the 5% significance level, but the difference is so small that it wouldn't be useful for user interface design.

T test

- T test compares the means of two samples
- Two-sided:
  - H0: means are equal
  - H1: means are different
- One-side:
  - H0: means are equal
  - H1: mean(A) < mean(B)
- Assumptions:
  - samples A & B are independent (between-subjects, randomized)
  - normally distributed
  - equal variance

Let's look at some of the more common statistical tests that are used in user interface experiments.

The T test is what you'd use to compare two means in a between-subjects experiment, like the hypothetical Mac/Windows menubar experiment we've been discussing. The T statistic computes the difference between the Mac average and the Windows average, divided by an estimate of the standard deviation. If the null hypothesis is true, then this statistic follows a T distribution (which looks very similar to a normal distribution, a hump centered at 0). You can look up the value of the T statistic you computed in a table of the T distribution to find out the probability of getting a more extreme value.

There are two forms of the T test, with different alternative hypotheses. In the more conservative, **two-sided** T test, your alternative hypothesis is merely that the means are different, so an extreme t value (either positive or negative) counts as evidence against the null hypothesis. The other form is the **one-sided** test, in which your alternative hypothesis expects the difference to go one way or the other – e.g., if there's any difference between Mac and Windows at all, the Mac should be faster. It's conventional to use the two-sided test unless you (and the skeptic you're arguing against) are completely certain which way the difference should go, if the difference exists at all.

Using the T test requires a few assumptions. First, your samples should be independent, so you need to use good experiment design with randomization and controls to prevent inadvertent dependence between samples. Second, the T test also assumes that the underlying probability distribution of the samples (e.g., the access times) is a normal distribution, and that even if the alternative hypothesis is true, both samples have equal variance. Fortunately the T test is not too sensitive to the normality and equal-variance assumptions: if your sample is large enough ($N > 20$), deviations don't affect it much.

Paired T Test

- For within-subject experiments
- Uses the mean of the differences (each user against themselves)
- H0: mean of differences is zero
- H1: mean of differences is nonzero (two-sided test)

What if we had run a **within-subjects** experiment instead? Then we would need to compare each subject with themselves, by computing the difference between each subject's Macintosh access time and the same subject's Windows access time. We would then use a t test to test the hypothesis that the mean of these differences is nonzero, against the null hypothesis that the mean of the differences is zero. This test is called a **paired t test**.

Why is a paired t test more powerful? Because by computing the difference within each user, we're canceling out the contribution that's unique to the user. That means that individual differences between users are no longer contributing to the noise (variance) of the experiment.

Analysis of Variance (ANOVA)

- Compares more than 2 means
- One-way ANOVA
  - 1 independent variable with k >= 2 levels
  - H0: all k means are equal
  - H1: the means are different (so the independent variable matters)

So far we've only looked at one independent variable (the menu bar position) with only two levels tested (Mac position and Windows position). If you want to test means when you have more than one independent variable, or more than two levels, you can use ANOVA (short for Analysis of Variance).

One-way ANOVA (also called "single factor ANOVA") addresses the case where you have more than two levels of the independent variable that you're testing. For example, suppose we wanted to test a third menu bar position at the bottom of the screen. Then we'd have three samples: top (Mac), below title (Windows), and bottom. One-way ANOVA can simultaneously compare all three means against the null hypothesis that all the means are equal.

ANOVA works by weighing the variation between the independent variable conditions (Mac vs. Windows vs. bottom) against the variation within the conditions (which is due to other factors like individual differences and random noise). If the null hypothesis is true, then the independent variable doesn't matter, so dividing up the observations according to the independent variable is merely an arbitrary labeling. Thus, assuming we randomized our experiment properly, the variation between those arbitrary groups should be due entirely to chance, and identical to the random variation within each group. So ANOVA takes the ratio between these two variations (computed as mean sums of squares), and if this ratio is significantly greater than 1, then that's sufficient evidence to argue that the null hypothesis is false and the independent variable actually **does** matter.

Two-Way ANOVA

- 2 independent variables with j and k levels, respectively
- Tests whether each variable has an effect independently
- Also tests for interaction between the variables

ANOVA can be extended to multiple independent variables, by looking at the variation between different levels of one independent variable (while holding the other independent variable constant). This is **two-way** (or two-factor) ANOVA.

Two-way ANOVA can be used to analyze a within-subjects experiment, where one independent variable is the variable we were testing (e.g. menubar position), while the other independent variable is the user's identity.