

CS 160: Lecture 16

Professor John Canny

4/3/2006

1

Qualitative vs. Quantitative Studies

- ☞ Qualitative: What we've been doing so far:
 - * Contextual Inquiry: trying to understand user's tasks and their conceptual model.
 - * Usability Studies: looking for critical incidents in a user interface.
- ☞ In general, we use qualitative methods to:
 - * Understand what's going on, look for problems, or get a rough idea of the usability of an interface.

4/3/2006

2

Quantitative Studies

- ☞ Quantitative:
 - * Use to reliably measure something
 - * Can compare different designs, or design changes
- ☞ Examples:
 - * Time to complete a task.
 - * Average number of errors on a task.
 - * Users' ratings of an interface *:
 - + Ease of use, elegance, performance, robustness, speed,...
- * - You could argue that users' perception of speed, error rates etc is more important than their actual values.

4/3/2006

3

Outline

- ☞ Basics of quantitative methods
 - * Random variables, probabilities, distributions
 - * Review of statistics
 - * Collecting data
 - * Analyzing the data

4/3/2006

4

Random variables

- ☞ Random variables take on different values according to a **probability distribution**.
- ☞ E.g. $X \in \{1, 2, 3\}$ is a discrete random variable with three possible values.
- ☞ To characterize the variable, we need to define the probabilities for each value:
$$\Pr[X=1] = \Pr[X=2] = \frac{1}{4}, \quad \Pr[X=3] = \frac{1}{2}$$
- ☞ On each **trial** or **experiment**, we should see one of these three values with the given probability.

4/3/2006

5

Random variables and trials

- ☞ When we examine X after a series of trials, we might see the values: 1, 1, 3, 2, 3, 1, 3, 3, 3, 1, 2,...
- ☞ We often want to denote the value of X on a particular trial, such as X_i for the i th trial.
- ☞ Then the above sequence could also be written as:
$$X_1 = 1, X_2 = 1, X_3 = 3, X_4 = 2, X_5 = 3, X_6 = 1, \\ X_7 = 3, X_8 = 3, X_9 = 3, X_{10} = 1, X_{11} = 2, \dots$$
- ☞ For large N , the sequence $\{X_1, \dots, X_N\}$ should contain the value 3 about $N/2$ times, the value 2 about $N/4$ times, and the value 1 about $N/4$ times.

4/3/2006

6

Random variables and trials

Q: How would you represent a fair coin toss with a random variable?

$X \in \{H, T\}$ $\Pr[X=H] = \frac{1}{2}$ $\Pr[X=T] = \frac{1}{2}$

Q: How would you represent a 6-sided die toss?

$Y \in \{1,2,3,4,5,6\}$, $\Pr[Y = i] = 1/6$ for $1 \leq i \leq 6$
 $\Pr[Y = i] = 0$ otherwise

4/3/2006 7

Independence

- ▣ Consider a random variable X which is the value of a fair die toss. Now consider Y, which is the value of another fair die toss.
- ▣ Knowing the value of X tells us nothing about the value of Y and vice versa. We say X and Y are **independent** random variables.
- ▣ However, if we defined $Z = X + Y$, then Z is dependent on X and vice versa (large values of X increase the probability of large values of Z, and Z must be at least X+1).

4/3/2006 8

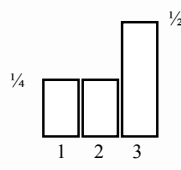
Independent Trials

- ▣ We will often want to use random variables whose values on different trials are independent.
- ▣ If this is true, we say the experiment has independent trials.
- ▣ Example: tossing a fair die many times. Each toss is a random variable which is independent of the other trials.

4/3/2006 9

Random variables

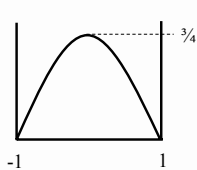
▣ Given $\Pr[X=1] = \Pr[X=2] = \frac{1}{4}$, $\Pr[X=3] = \frac{1}{2}$ we can also represent the distribution with a graph:



4/3/2006 10

Continuous Random variables

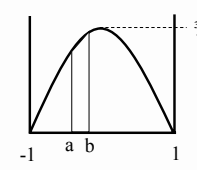
- ▣ Some random variables take on continuous values, e.g. $Y \in [-1,1]$.
- ▣ The probability must be defined by a **probability density function** (pdf).
- ▣ E.g. $p(Y) = \frac{3}{4} (1 - Y^2)$
- ▣ Note that the area under the curve is the total probability, which must be 1.



4/3/2006 11

Continuous Random variables

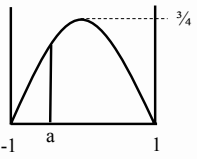
- ▣ The **area** under the pdf curve between two values gives the probability that the value of the variable lies in that range.
- ▣ i.e. $\Pr[a < Y < b] = \int_a^b \frac{3}{4} (1 - Y^2) dY$



4/3/2006 12

Meaning of the distribution

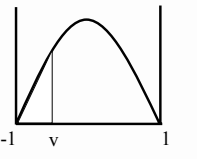
- The limit of the area as the range $[a,b]$ goes to zero gives the value of $p(Y)$

$$\Pr[a < Y < a+dY] = p(Y) dY$$


4/3/2006 13

CDF: Cumulative Distribution

- The CDF is the area under the distribution from $-\infty$ to some value v
- So $C(-\infty) = 0$ and $C(\infty) = 1$



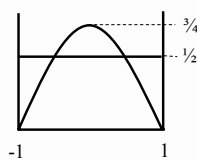
4/3/2006 14

Mean and Variance

- The mean is the expected value of the variable. Its roughly the average value of the variable over many trials.

$$\text{Mean} = E[Y] = \int_{\min Y}^{\max Y} Y p(Y) dY$$

- In this case $E[Y] = \frac{1}{2}$



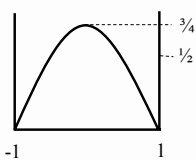
4/3/2006 15

Variance

- Variance is the expected value of the square difference from the mean. Its roughly the squared "width" of the distribution.

$$\text{Var}[Y] = \int_{\min Y}^{\max Y} (Y - \bar{Y})^2 p(Y) dY$$

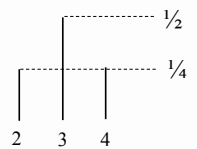
- Standard deviation $\text{std}[X]$ is the square root of variance.



4/3/2006 16

Mean and Variance

- What is the mean and variance for the following distribution?



4/3/2006 17

Sums of Random Variables

- For **any** X_1 and X_2 , the expected value of a sum is the sum of the expected values:

$$E[X_1 + X_2] = E[X_1] + E[X_2]$$

- For **independent** X_1 and X_2 , the variance of the sum is also the sum of the variances:

$$\text{Var}[X_1 + X_2] = \text{Var}[X_1] + \text{Var}[X_2]$$

4/3/2006 18

Identical trials

- For independent trials with the same mean and variance $E[X]$ and $\text{Var}[X]$,

$$E[X_1 + \dots + X_n] = n E[X]$$

$$\text{Var}[X_1 + \dots + X_n] = n \text{Var}[X]$$

$$\text{Std}[X_1 + \dots + X_n] = \sqrt{n} \text{Std}[X]$$

Where $\text{Std}[X] = \text{Var}[X]^{1/2}$

4/3/2006

19

Identical trials

- If we define $\text{Avg}(X_1, \dots, X_n) = (X_1 + \dots + X_n)/n$, then

$$E[\text{Avg}(X_1, \dots, X_n)] = E[X]$$

- While

$$\text{Std}[\text{Avg}(X_1, \dots, X_n)] = (1/\sqrt{n}) \text{Std}[X]$$

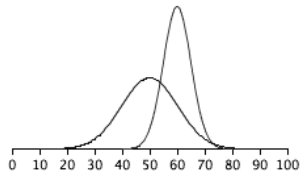
- i.e. the standard deviation in an average value decreases with n , the number of trials.

4/3/2006

20

Identical trials

- i.e. the distribution narrows in a relative sense.
- The blue curve is the sum of 100 random trials, the red curve is the sum of 200.

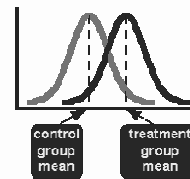


4/3/2006

21

Detecting differences

- The more times you repeat an experiment, the narrower the distributions of measured average values for two conditions.
- So the more likely you are to detect a difference in a test variable between two cases.



4/3/2006

22

Break

4/3/2006

23

Variable types

- Independent Variables: the ones you control
 - * Aspects of the interface design
 - * Characteristics of the testers
 - * Discrete: A, B or C
 - * Continuous: Time between clicks for double-click
- Dependent variables: the ones you measure
 - * Time to complete tasks
 - * Number of errors



4/3/2006

24

Some statistics

- ▣ Variables X & Y
- ▣ A relation (hypothesis) e.g. $X > Y$
- ▣ We would often like to know if a relation is true
 - * e.g. X = time taken by novice users
 - * Y = time taken by users with some training
- ▣ To find out if the relation is true we do experiments to get lots of x's and y's (observations)
- ▣ Suppose $\text{avg}(x) > \text{avg}(y)$, or that most of the x's are larger than all of the y's. What does that prove?

4/3/2006

25

Significance

- ▣ The significance or p-value of an outcome is the probability that it happens by chance if the relation does *not* hold.
- ▣ E.g. $p = 0.05$ means that there is a 1/20 chance that the observation happens if the hypothesis is false.
- ▣ So the smaller the p-value, the greater the significance.

4/3/2006

26

Significance

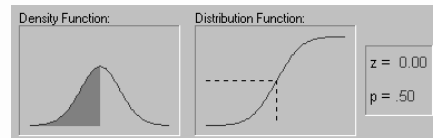
- ▣ For instance $p = 0.001$ means there is a 1/1000 chance that the observation would happen if the hypothesis is false.
So the hypothesis is almost surely true.
- ▣ Significance increases with number of trials.
- ▣ CAVEAT: You have to make *assumptions* about the probability distributions to get good p-values. There is always an implied model of user performance.

4/3/2006

27

Normal distributions

- ▣ Many variables have a Normal distribution (pdf)



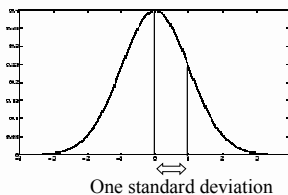
- ▣ At left is the density, right is the cumulative prob.
- ▣ Normal distributions are completely characterized by their *mean* and *variance* (mean squared deviation from the mean).

4/3/2006

28

Normal distributions

- ▣ The std. deviation for a normal distribution occurs at about 60% of its value



4/3/2006

29

T-test

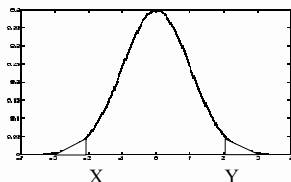
- ▣ The T-test asks for the probability that $E[X] > E[Y]$ is false.
- ▣ i.e. the null hypothesis for the T-test is whether $E[X] = E[Y]$.
- ▣ What is the probability of that given the observations?

4/3/2006

30

T-test

- ▣ We actually ask for the probability that $E[X]$ and $E[Y]$ are at least as different as the observed means.



4/3/2006

31

Analyzing the Numbers



- ▣ Example: prove that task 1 is faster on design A than design B.
 - * Suppose the average time for design B is 20% higher than A.
 - * Suppose subjects' times in the study have a std. dev. which is 30% of their mean time (typical).
- ▣ How many subjects are needed?

4/3/2006

32

Analyzing the Numbers



- ▣ Example: prove that task 1 is faster on design A than design B.
 - * Suppose the average time for design B is 20% higher than A.
 - * Suppose subjects' times in the study have a std. dev. which is 30% of their mean time (typical).
- ▣ How many subjects are needed?
 - * Need at least 13 subjects for significance $p=0.01$
 - * Need at least 22 subjects for significance $p=0.001$
 - * (assumes subjects use both designs)

4/3/2006

33

Analyzing the Numbers (cont.)

- ▣ i.e. even with strong (20%) difference, need lots of subjects to prove it.
- ▣ Usability test data is quite variable
 - * 4 times as many tests will only narrow range by 2x
 - * breadth of range depends on sqrt of # of test users
 - * This is when surveys or automatic usability testing can help

4/3/2006

34



Lies, damn lies and statistics...

- ▣ A common mistake (made by famous HCI researchers *):
 - ▣ Increasing n , the number of trials, by running each subject several times.
 - ▣ No! the analysis only works when trials are *independent*.
 - ▣ All the trials for one subject are dependent, because that subject may be faster/slower/less error-prone than others.
- * - making this error will not help you become a famous HCI researcher ☺.

4/3/2006

35

Statistics with care:

- ▣ What you *can* do to get better significance:
 - * Run each subject several times, compute the *average* for each subject.
 - * Run the analysis as usual on subjects' average times, with n = number of subjects.
- ▣ This decreases the per-subject variance, while keeping data independent.

4/3/2006

36



Statistics with care:

- Another common mistake:
- An experiment fails to find a significant difference between test and control cases (say at $p = 0.05$), so you conclude that there is no significant difference.
- No!
- A difference-of-averages test can only confirm (with high probability) that there *is* a difference. Failure to prove a significant difference can be because
 - * There is no difference, OR
 - * The number of subjects in the experiment is too small

4/3/2006

37



Statistics with care:

- Example, what should you conclude if you find no significant difference at $p = 0.05$, but there is a difference at $p = 0.2$?
- First of all, the result does not confirm a significant difference with any confidence.
- However, while there may not be a significant difference, it is *more likely* that there is but it is too weak at the N chosen. Therefore, try repeating the experiment with a larger N .

4/3/2006

38



Statistics with care:

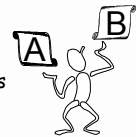
- You write a paper with 20 different studies, all of which demonstrate effects at $p=0.05$ significance. They're all right, right?
- Actually, there is significant probability (as high as 63%) that there is no real effect in at least one case.
- Remember a p-value is an upper bound on the probability of no effect, so there is always a chance the experiment gives the wrong result.

4/3/2006

39

Using Subjects

- Between subjects** experiment
 - * Two groups of test users
 - * Each group uses only 1 of the systems
- Within subjects** experiment
 - * One group of test users
 - * Each person uses both systems



4/3/2006

40

Between subjects

- Two groups of testers, each use 1 system
- Advantages:**
 - * Users only have to use one system (practical).
 - * No learning effects.
- Disadvantages:**
 - * Per-user performance differences confounded with system differences:
 - * Much harder to get significant results (many more subjects needed).
 - * Harder to even predict how many subjects will be needed (depends on subjects).

4/3/2006

41

Within subjects

- One group of testers who use both systems
- Advantages:**
 - * Much more significance for a given number of test subjects.
- Disadvantages:**
 - * Users have to use both systems (two sessions).
 - * Order and learning effects (can be minimized by experiment design).

4/3/2006

42

Example

- ☐ Same experiment as before:
 - * System B is 20% slower than A
 - * Subjects have 30% std. dev. in their times.
- ☐ Within subjects:
 - * Need 13 subjects for significance $p = 0.01$
- ☐ Between subjects:
 - * Typically require 52 subjects for significance $p = 0.01$.
 - * But depending on the subjects, we may get lower or higher significance.

4/3/2006

43

Experimental Details

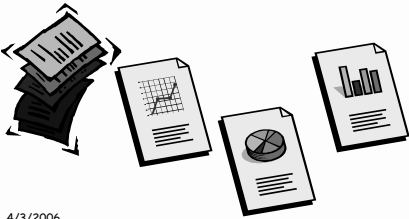
- ☐ Learning effects
 - * Subjects do better when they repeat a trial
 - * This can bias within-subjects studies
 - * So "balance" the order of trials with equal numbers of A-B and B-A orders.
- ☐ What if someone doesn't finish?
 - * Multiply time and number of errors by 1/fraction of trial that they completed.
- ☐ Pilot study to fix problems
 - * Do 2, first with colleagues, then with real users

4/3/2006

44

Reporting the Results

- ☐ Report what you did & what happened
- ☐ Images & graphs help people get it!



4/3/2006

45

Summary

- ☐ Random variables
- ☐ Distributions
- ☐ Statistics (and some hazard warnings)
- ☐ Experiment design guidelines

4/3/2006

46