15.301/310, Managerial Psychology
Prof. Dan Ariely
Recitation 8: T test and ANOVA

Statistics does all kinds of stuff to describe data
Talk about baseball, other useful stuff
We can calculate the probability.
Good for gambling
Build bridge, likely to collapse?
Calculate probabilities

Another tool is statistical inference – cover today
Designing models – not discuss

Imagine observe set of things that co-occur

People work for you – tall, short, M, F, MIT, Harvard, ME, CS, vary all kinds of things
Also Observe how good they are, See how well perform
Predict new person – how will they perform

If prediction
mechE vs CS makes 12% better employees?
Can find out all these things with data

Very hard to intuit who is better/worse
Statistics is very useful because it describes very accurately this structural aspect of data

Can try to understand what is the structure of the data

Statistical inference
Question we ask: we have sample, some info, we ask, are these things different from each other for real?

Hypothesis testing
T-test, F-test

Say Females in this class 8 average on niceness scale, males 7
Are they different?
Yes, everything in life is different, even 8, 8.1 different
What can we conclude?

-not a lot – don't know scale
Why can't infer?
-error might be big enough to encompass both

No doubt 8 bigger than 7
Question - is 8 more than 7 in meaningful way? Or just tiny, meaningless way

Is f nicer than m in this class, or can we infer to rest of world?

Other example that don't have variance due to error

Trying to estimate, adjust for variance, ask can we make inferences

Could be biased sample. What does that mean? Could be this class, one person comes has all these nice friends, gets them to come. Could be all kinds of stories.

If self selection, inferential statistics can't help us
All things talked about earlier in semester, outside range of inferential statistics
Assumes sample is correct

Test if valid like or not
Stochastic world, lot of random error, come from all kinds of things

-bad survey technique
-bad random sampling
-coincidence

Lot of stuff happening all the time
All things happen we can't record
Weather could have diff effect

Bad grade yes, good grade yes
Because we can't make everything, error

All kinds of reasons we expect variance

All inferential statistics about variance

The size of the mean difference is the same? (slide)
Blue – different
Red – much more cautious to say these are really different

Find by taking mean difference and variance

Imagine story about women, men
Almost no overlap, then can say this class, women nicer than men, then think about applying it to the rest of the world

In science, we can never prove anything
Cannot go everywhere and find out, prove always true

Why? Or can we?
-can't say 100% probability, don't know all the possibilities, evidence

We are limited, we can never say something is true

Enough to have 1 counterexample to prove something false
Honest or dishonest
Can find 1 instance you were dishonest, and decide you are dishonest

Might quibble
If women nicer than men 90% of places, really say not true?

Standard way of testing theories is to assume they are right and try to find a case where they are wrong
Hypothesis testing does – very backwards way of looking at life

Come up with opposite theory from mine and prove it wrong
Therefore promote own theory

That's what inference is

H0, H1

H0 hyp we are not interested in
H1 is hypothesis we are interested in.

H0 allows us to show it is wrong
How? H0 includes an = sign.

(slide)

My real hypothesis, H1, women nicer than men. In that case, H0 is what?

Women as nice as men, or less nice

Always important H0 includes =, = we can reject

Important cover whole universe of possible outcomes with H0 and H1 combined
Otherwise disproving H0 means nothing about H1

H0 could be women as nice as men. H1 would be?

-women nicer or meaner

I say w as nice or nicer. H1?

Women not as nice

Always take opposite

Remember H0 one we don't believe in, but the one we test

-why easier to prove false than true?

One thing, counterexample can prove false. One supporting example means nothing, impossible to check all cases, every instance

Could never check entire universe

Think about it as fully true, if one false, false

Flukes are a separate topic, issue

Other point, can show things are not the same.

Only thing can disprove is a = sign.

If I say larger, can't actually prove that
Can't prove something larger

Can prove not same,
That's the exercise we do

Examples H1, H0

Say you have coin,
Suspect it is fake
What is H0, H1?

-test if fair
-H0 coin not fake
What probability?
-1/2

Set H1, Ho,

H1 is?

Probability not ½

-gender, grades are the same, .5

-larger than or less than .5

New medicine. Healed faster?

-H0 med not healed faster H1 healed same or slower

Try again

-H0 same or slower, H1 faster

H0 always include same

Want to include slower in H0

-what if want to prove same

No way to prove, because not the same
Nothing in this world is the same.
Can decide if diff
Is it small enough?

Different kinds of analysis
Can't reject hypothesis
same, but can't prove
can only fail to reject

H1- this medication has no effect
Test 1 medication compared to nothing, can't have 'not at all' as H1, has to be H0
Reason: it is the only point where there is no effect, no difference.
Another med, then can be part of H1
Important subtlety

Cheating
-H0 is that course 15 and 6 same probability to cheat, H1 is one is better or worse

Called 2 sided hypothesis
H1 both below and above. Many time when you test theories, they are not two-sided, but one-sided. That means H0 also includes the opposite of H1

Marriage over time. Give me a hypothesis for that. Marriage and happiness lets say
-time increases, no correlation,

Love marriages vs arranged marriages in India
Love marriages start higher, arranged lower.
Cross in year 3

Done H0s, H1s

Why test hypothesis we don't believe in?
An odd way to think about life
Now, what does it mean to reject an H0 hypothesis?

Here's what we do
No difference between groups
Find data and decide to reject this

What it really means, what we have learned: if H0 is correct, problem of getting results we got so low, we reject H0. And therefore accept H1
Convoluted because we assume H0 is correct

Question is: is the probability of getting this result high or low?

Assume H0 is correct, this means we assume no difference between men, women
If no difference, how likely will Women be 8 in niceness, men 7?
If we say it is pretty likely it will happen,
Cannot reject H0
Variance so high that P(experiment men 8, women 7) = 1/3. nothing makes us doubt H0
But if probability very low…
What we mean when we say very small is 1/20, 5%
If probability of getting this data or more extreme data is very low, we reject H0

If women less nice, I conclude H0, you H1

Regrettably, it will not help to spend 3 or 4 lectures on this difficult topic

Ask grad students. Understand H0, H1?

-yes
-no

I still want you to think about it.

Could be H0 is correct. Cannot reject H0, then accept it
(slide)

Also could be wrong. Well, if we assume H0, unlikely to get result, so we reject it

Can make 2 types of mistakes.
Type 1 error – reject H1 by mistake. Do not disprove H0
Type 2 error – H1 is wrong. Cannot reject H0

Decide which is more important to us

Together it is a constant sum

Can be right or wrong, but if we are wrong, we can decide how we want to be wrong.

Under H0, p(getting result) = 1/5, more likely type 1 error
If it has to be 1/500, it is less likely to be a type 1 error. Maybe it turns out 1/400, and we can't reject H0

Decision is a tradeoff between types of errors.
If the result is so important, revolutionary, interesting, we might be willing to be more lenient

Rejecting H0 is less stringent if the results are important, interesting, could be valuable

Could be so damaging, we would want a very stringent type 1 error.

Say we find gene difference between gay, non-gay people.
Find peak value is .2

P(this or more extreme result) = 1/5. do we want to publish this or not?
Hard to say
If published, people start believing it, turns out wrong, then it might cause damage to a lot of people

Standard is 5%
Can adapt depending on how important things are
What is the cost of the 2 types of mistakes?

Not .05, but .07. - very close.
Decide, what do you do? Publish? Not? Tell people?
No magical solution to that

Can run more subjects, experiments
Not always possible

Assume this distribution, this one (slide)
H0 on left, H1 on right
Some chance H0 is correct.
Much bigger chance we did a good job.

Can come all the way to the right, want to be absolutely sure
Or left, can be more lenient.

-p increases?
As you move right, P is a smaller #
-because p is divided in terms of H0?
Right, exactly right.

Definition of p?
-smallest chance you would say can happen by coincidence. Smallest likelihood of something happening

P is: assuming H0 correct, Probability of getting data we got or more extreme data under chance alone.
If p very low, might decide H0 is wrong

All kinds of Ps

How about p = 0.11?

Particularly for your projects
Might discover some mistakes.
Theory wrong?
Data wrong?
Data collection too small to be consistent?

This is part of the interpretation of P

If smaller P, do you have stronger confidence in H1?

Say one p = .003, other .00003
In which do we have more confidence in H1?

No, fact is we only can be confident in H0 being not correct. Generally, does not translate into H1 being correct

Smaller P, effect size bigger?

No, P not just about effect size. Some combination of mean variance and difference

How about p and # of subjects

More subjects we have, smaller p will be.
That's because of variance

Talked before about effect sizes
Remind you for your papers – giving percentages is not enough

8.1, 8 - is this a significant difference?
Say all women rate themselves 8.1, all men 8. Is there a difference?

In summary:
Hypothesis testing, have H1, H0, set hypothesis to something we don't believe in, hoping to reject it

Meaning of P: probability of getting this or more extreme data, assuming H0 holds.
If H0 doesn't hold, we don't know
2 types of error and tradeoffs

Effect sizes

Statistical tests

For 1 sample, 2 samples, some about Anovas

T test for 1 sample
Say we had a medication, is it effective?

Distribution of people in 1 dimension – how healthy they feel.

Are they healthier than 0? Healthier that 5? Give me examples of other distributions, compare it to a single number

-satisfied with customer service?
-could ask how satisfied are you?

Compare to specific #

-satisfied with life at MIT?

-drink more than 2 dozen beers a week?

-fairness of coin bigger than .5?

-age above or below average US age?
-in this class?

Statistical question never asks about individuals

Gender proportion

Here is the way T is calculated. (slide)

Here is the logic of this test:

Top: Difference between observation, mu, and hypothesized mean
Hundred time flip coin, 40 times got tails
N is hypothesized mean = .5 in this case

Is this medication helpful?

Medication helpful will just be mu
Just ask about the mean differences

Standard deviation should be familiar, we talked about this before
Wider it is, the less likely it will be significant

Square root of number of people that participated in the study

The more people we take, the better estimate of what is really happening in reality

If we've done research with thousands of people, we take mean difference more seriously

3 elements of the test

Once you get it, there is a T table in the book
Just compare it

T becomes close to the normal distribution

Say 7 observations
Students in class
Measured how aggressive they are

Take difference between individual measure, overall mean, and square it, take sum of all of them

Statements about H0
Aggressiveness

-why using mu, not xbar?

It is estimated
I wouldn't feel strongly about using xbar, that would be fine

What do we have?

T is the difference between my mean, hypothetical mean of what will happen.
divide it, get 3.42
Go to table, figure out if it is significant
Small sample, that will hurt us

Probably significant but not sure

Let me show example of how this would look
Software called Statview. Can do it in Excel as well.

How many good ideas does an MIT student have a year?

What would be your null hypothesis?

-7

I was generous

What is the hypothesized mean?
Ok, so 7 ideas
Value at 1.69 is .1, doesn't exactly make it,
We can't reject the hypothesis that students have 7 good ideas per year

Can reject the hypothesis that people have 0 ideas

Seems like students have somewhere below 10
We want to describe what is actually happening in the data

Get mean, range, 95% And get outliers. This is called box plots

What we've learned is that MIT students are creative, more than 0 good ideas/yr, but not 1x month

2 sample T test

Basically, structure is very similar

Basic structure for T test

Compare for what we think will be the difference between 2 groups

Relative to expected difference between 2 groups

Here is the story
2 sample T test (slide)

Real difference between men, women
Hypothesized difference

-parenthesis still under radical?
Yes

Example 1 slide
Who eats more lollipops, M or F?

We expected difference to be 0

Take #s, variances, number of people, get result
We think it is close to being significant

Example, does sun create freckles?
H0, sunny side same or less freckles that non-sunny side

Let me Show you how would look in some statistical software
2 columns, before, after
[shows example]

Want to cover Anova

T test example - more general approach to life, Anova - analysis of variance

First, if we make many comparisons, we are likely to get one by mistake
By defining p = .5, we are saying for every 100 experiments, we will get 5 in which results might be wrong
We have to worry about doing too many comparisons

Second, Sometimes, even cells we are not using help us get a better estimation of error value

IQ across MIT for all majors
Take everybody into account, better estimation

Story is the same, look at variance within cells and also across cells
We ask, how does this affect our concept of variance?

Here is a graphical attempt to illustrate this.
(slide)

Right now, looks like no order (red dots slide)

what we ask is whether these barriers change the variance

this case, answer is no
can imagine things to cluster
variance between groups

note, across distinction makes difference, but not vertical distinction

does it help us get smaller variance?

In this example, everyone is 40 in some measure

In this one, no variance within the group, but variance across groups

Ideal case:
Can have a situation, vary within a group, but no difference In the overall mean

Most common - Can have variance in and variance between, can we make inferences about this?

General formula the same

Diff types of means w/ Anova
Also look at other sum of squares
To get sum of squares, extract from appropriate mean.
Partition overall variance

Once you have sum of squares, compare to sum of squares between and within. As it gets larger, the groups tell us more and more about these effects. Likely to have significant results

One more thing, degrees of freedom
Df stands for degrees of freedom

If I told you all the data plus mean, I have replicated some information

Take away information, you would still be able to calculate it

Anovas can be 1 factor or 2 factor design
Translates very nicely to Anovas

We talked about hypothesis testing
Once we reject it, we are willing to accept H1

How many think they took writing seriously?
I want to point out, writing is important skill
I don't know how much time you spend on writing

Now I am happy to produce a page in a day or two
Make sure every sentence is one you like

try and do better
Writing I am certain will be useful for you