Statistical Hypothesis Testing

- Popcorn, soda & statistics
- Null Hypothesis Significance Testing (NHST)
- Statistical Decisions, Decision Errors & Statistical Conclusion Validity
- Major Bivariate Analyses

Just imagine… You're at the first of 12 home games of your favorite team. You're sitting in the reserved seat you'll enjoy all season. Just before half-time, the person in the seat next to you says, "Hey, how about if before each half-time we flip a coin to see who buys munchies? Heads you buy, tails I'll buy. I have this official team coin we can use all 12 times.

Hey, what do you know, its heads. I'll have some popcorn, a hot dog, a candy bar and a drink! Want help carrying that?" You don't think much of it because you know that it's a 50-50 thing -- just *your turn* to lose!

The next game the coin lands heads again, and you buy your "new friend" hot chocolate, a Polish Dog, fries and some peanuts. Still no worries, a couple in a row is pretty likely.

The next game you buy him a couple of Runza's, some cotton candy and an orange drink.

Finally, you're starting to get suspicious!

Before the next game you have a chance to talk with a friend or yours who has had a statistics course. You ask your friend, "I've bought snacks all three times, which could happen if the coin were fair, but I don't know how many more times I can expect to feed this person before the season is up. How do I know whether I should "confront" them or just keep politely buying snacks?"

Your friend says, "We covered this in stats class. The key is to figure out what's the probability of you buying snacks a given number of times if the coin is fair. Then, you can make an 'informed guess' about whether or not the coin is fair. Let me whip out my book!"

# Heads/12	Probability
12	.00024
11	.0029
10	.0161
9	.0537
8	.1208
7	.1936
6	.2256
5	.1936
4	.1208
3	.0537
2	.0161
1	.0029
0	.00024

Your friend says, "This table tells the probability of getting a given number of #heads/12 flips if the coin is fair."

"We know that the most likely result - if the coin is fair - is to get 6/12 heads. But we also know that this won't happen every time. Even with a fair coin the #heads/12 will vary by chance."

"The table tells 6/12 heads will happen 22.56% of the time -- if the coin is fair," says your friend.

What's the chances of getting each of the following -- if the coin is fair ??

4/12 heads about 12%
2/12 heads about 1.6%
8/12 heads about 12%
10/12 heads about 1.6%

Notice anything?

The probability distribution is symmetrical around 6/12 -- 4/12 is as likely as 8/12 and 0/12 is as likely as 12/12 !!!

"So, there is a 'continuum of probability' -- a 6/12 heads is the most likely if the coin is fair, and other possible results are less and less likely as you move out towards 0/12 and 12/12 if the coin is fair ," says your friend.

# Heads/12	Probability
12	.00024
11	.0029
10	.0161
9	.0537
8	.1208
7	.1936
6	.2256
5	.1936
4	.1208
3	.0537
2	.0161
1	.0029
0	.00024

"OK," your friend continues, "now we need a 'rule'. Even though all these different #heads/12 are possibilities, some are going to occur pretty rarely if the coin is fair."

"We'll use our rule to decide when a certain #heads/12 is probably too rare to have happened by chance if the coin is fair. In stats the traditional is the '5% rule' -- any #heads/12 that would occur less than 5% of the time if the coin were fair is considered "too rare", and we will decide that it isn't a fair coin!", says your friend.

"Using the 5% rule we'd accept that the coin is fair if we buy 6, 7, 8 or even 9 times, but we'd reject that the coin is fair if we buy snacks 10, 11 or 12 times" (Actually, the coin probably isn't fair if we only buy 1-3 times, but why fuss!)

"So, we have a "cutoff" or "critical value" of 9 heads in 12 flips -- any more and we'll decide the coin is unfair."

Quick check if this is making sense…

Let's say that you're at the candy store with a "friend of a friend" and decide to sample 8 different types of expensive candies. This "friend of a friend" just happens to have a deck of cards in their pocket and suggests that you pick a card. If it is red, then you buy, but if it is black, then they will buy.

# Reds/8	Probability
8	.0039
7	.0313
6	.1093
5	.2188
4	.2734
3	.2188
2	.1093
1	.0313
0	.0039

Notice that this is another 50-50 deal -- in a fair deck of cards, there should be 50% red and 50% black.

Speaking of coincidences, you just happen to have a table of probabilities for 8 50%-50% trials in *your* pocket !!!!

Using the "5% rule" what would be the "critical value" we'd use to decide whether or not the deck of cards was "fixed" ???

The critical value would be 6.

What would we decide if we bought 6 candies? The deck is fair

What would we decide if we bought 7 candies? The deck is "fixed"

Back to the game & munchies…Just as you're thanking your friend and getting ready to leave, your friend says, "Of course there is a small problem with making decisions this way!" You sit back down.

# Heads/12	Probability
12	.00024
11	.0029
10	.0161
9	.0537
8	.1208
7	.1936
6	.2256
5	.1936
4	.1208
3	.0537
2	.0161
1	.0029
0	.00024

"Notice what we've done here", says your friend. "Using the '5%' rule leads to a 'critical value' of 9/12 heads. That is, we've decided to claim that 10, 11 or 12/12 heads is probably the result of an unfair coin. However, we also ***know*** that each of these outcomes is *possible* (though with low probability) with a fair coin. Any fair coin will produce 10/12 heads 1.6% of the time. But when it happens we'll claim that the coin is unfair -- and we'll be wrong. This sort of mistake is called a 'false alarm'. "

Your friend is getting into it now, "Most unfair coins don't have a head on either side -- that's too easy to check. Instead they are heavier on the tail, to increase the probability they will land heads. So, there is also the possibility that the coin is unfair, but produces fewer than 10/12 heads."

"If that happens, then we'll incorrectly decide that an unfair coin is really fair -- called a 'miss'."

Altogether, there are four possible decision outcomes

- two possible correct decisions
- two possible mistakes

Here's a diagram of the possibilities...

our statistical decision	in reality: fair coin	in reality: unfair coin
# heads < critical value, so we decide "fair coin"	Correct Retention	Miss
# heads > critical value, so we decide "unfair coin"	False Alarm	Correct Rejection

Back to the "cards and candies" example for some practice …

# Reds/8	Probability
8	.0039
7	.0313
6	.1093
5	.2188
4	.2734
3	.2188
2	.1093
1	.0313
0	.0039

What would be the critical value for this decision? Buying 6/8 candies

#1 You buy 5 out of 8 candies.
- Would you decide the deck is "fair" or "fixed"? fair

Later you look through the deck and its "fair"
- What type of decision did you make? Correct retention

#2 You buy 7 of the 8 candies.
- Would you decide the deck is "fair" or "fixed"? fixed

Later you look through the deck and its "regular".
- What type of decision did you make? False alarm

#3 You buy all 8 candies.
- Would you decide the deck is "fair" or "fixed"? fixed

Later you look through the deck -- no spades, 2 sets of diamonds
- What type of decision did you make? Correct rejection

#4 You buy 6 of the 8 candies.
- Would you decide the deck is "fair" or "fixed"? fair

Later you discover the clubs have been replaced with hearts
- What type of decision did you make? Miss

This was really a story about Null Hypothesis Significance Testing

Using the jargon of NHST...

- All the flips (ever) of that special team coin was the target population
- There are two possibilities in that population -- coin is fair or unfair
- The initial assumption the coin is "fair" is the Null Hypothesis (H0:)
- The 12 flips of that special team coin were the data sample
- The number of #heads/12 was the summary statistic
- We then determined the probability (p) of that summary statistic if the null were true (coin were fair) and made our statistical decision
  - If the probability had been greater than 5% ($p > .05$), we would have retained the null (H0:) and decided the coin was fair
  - if the probability had been less than 5% ($p < .05$), , we would have rejected the null (H0:) and decided the coin was unfair
- Don't forget that there are two ways to be correct and two ways to be wrong whenever we make a statistical decision

Most of our NHST in this class will involve bivariate data analyses

- asking "Are these two variables related in the population?"
- answering based on data from a sample representing the pop

The basic steps will be very similar to those for the #flips example...

- Identify the population
- Determine the two possibilities in that population
  - the variable are related
  - the variables are not related -- the H0:
- Collect data from a sample of the population
- Compute a summary statistic from the sample
- Determine the probability of obtaining a summary statistic that large or larger if H0: is true
- Make our inferential statistical decision
  - if $p > .05$ retain H0: -- bivariate relationship in sample is not strong enough to conclude that there is a relationship in pop
  - if $p < .05$ reject H0: -- bivariate relationship in sample is strong enough to conclude that there is a relationship in pop

When doing NHST, we are concerned with making statistical decision errors -- we want our research results to represent what's really going on in the population.

Traditionally, we've been concerned with two types of statistical decision errors:

Type I Statistical Decision Errors

- rejecting H0: when it should not be rejected
- deciding there is a relationship between the two variables in the population when there really isn't
- a False Alarm
- how's this happen?
  - sampling variability ("sampling happens")
  - nonrepresentative sample (Ext Val)
  - confound (Int Val)
  - poor measures/manipulations of variables (Msr Val)
  - Remember the decision rule is to reject H0: if $p < .05$ -- so we're going to make Type I errors 5% of the time!

Type II Statistical Decision Errors

- retaining H0: when it should be rejected
- deciding there is not a relationship between the two variables in the population when there really is
- a Miss
- how's this happen?
  - sampling variability ("sampling happens")
  - nonrepresentative sample (Ext Val) poor
  - confound (Int Val)
  - poor measures/manipulations of the variables (Msr Val)
  - if the sample size is too small, the "power" of the statistical test might be too low to detect a relationship that is really there (much more later…)

This is what we referred to as "statistical conclusion validity" in the first part of the course.

- Whether or not our statistical conclusions are valid / correct ??

These are the two types of statistical decision errors that are traditionally discussed in a class like this. Summarized below...

our statistical decision	in the target population: H0: True (variables not related)	H0: False (variables are related)
p > .05 -- decide to retain H0:	Correct Retention of H0:	Type II error "Miss"
p < .05 -- decide to reject H0:	Type I error "False Alarm"	Correct Rejection of H0:

Which two would be "valid statistical conclusions"? Correct rejection & correct retention

Which two would be "invalid statistical conclusions"? False Alarm & Miss

However,there is a 3rd kind of statistical decision error that I want you to be familiar with, that is cleverly called a ...

Type III statistical decision errors

- correctly rejecting H0:, but mis-specifying the relationship between the variables in the population
- deciding there is a certain direction or pattern of relationship between the two variables in the population when there really is different direction or pattern of relationship
- a Mis-specification
- how's this happen?
  - sampling variability ("sampling happens")
  - nonrepresentative sample (Ext Val)
  - confound (Int Val)
  - poor measures/manipulations of variables (Msr Val)

What makes all of this troublesome, is that we'll never know the "real" relationship between the variables in the population

- we can't obtain data from the entire target population (that's why we *have* sampling - duh!)
- if we knew the population data, we'd not ever have to make NHSTs, make statistical decisions , etc (double duh!)

The best we can do is...

- replicate our studies
  - using different samplings from the target population
  - using different measures/manipulations of our variables
- identify the most consistent results
- use these consistent results as our best guess of what's really going on in the target population

Practice with statistical decision errors evaluated by comparing our finding with "other research" …

We found that those in the Treatment group performed the same as those in the Control group. However, the other 10 studies in the field found the Treatment group performed better, — Type II

We found that those in the Treatment group performed better than those in the Control group. This is the same thing the other 10 studies in the field have found. — Correct Pattern

We found that those in the Treatment group performed poorer than those in the Control group. But all of the other 10 studies in the field found the opposite effect. — Type III

We found that those in the Treatment group performed better than those in the Control group. But none of the other 10 studies in the field found any difference. — Type I

We found that those in the Treatment group performed the same as those in the Control group. This is the same thing the other 10 studies in the field have found. — Correct H0:

Another practice with statistical decision errors ...

We found that students who did more homework problems tended to have higher exam scores, which is what the other studies have found. — Correct Pattern

We found that students who did more homework problems tended to have lower exam scores. Ours is the only study with this finding. — Can't tell -- what DID the other studies find?

We found that students who did more homework problems tended to have lower exam scores. All other studies found the opposite effect. — Type III

We found that students who did more homework problems and those who did fewer problems tended to have about the same exam scores, which is what the other studies have found. — Correct H0:

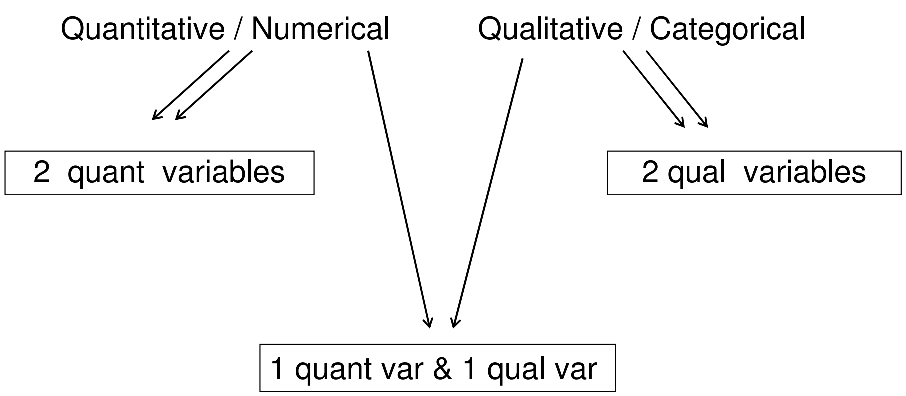
We found that students who did more homework problems tended to have lower exam scores. Ours is the only study with this finding, other find no relationship. — Type I

We found that students who did more homework problems and those who did fewer problems tended to have about the same exam scores. Everybody else has found that homework helps. — Type II

So… what are the bivariate null hypothesis significance tests (NHSTs) we'll be using ???

What are the two kinds of variables that we've discussed?

What are the possible bivariate combinations?



We have separate bivariate statistics for each of these three data situations...

For 2 quantitative / numerical variables...

Pearson's Product Moment Correlation (Pearson's r)

Purpose: Determine whether or not there is a linear relationship between two quantitative variables

H0: There is no linear relationship between the two quantitative variables in the population represented by the sample

Summary Statistic: r has range from -1.00 to 1.00

Basis & meaning of NHST:

- $p > .05$ retain H0: -- the linear relationship between the variables in the sample is not strong enough to conclude that there is a linear relationship between the variables in the population
- $p < .05$ reject H0: -- the linear relationship between the variables in the sample is strong enough to conclude that there is a linear relationship between the variables in the population

For 2 qualitative / numerical variables...

Pearson's Contingency Table X^2 (Pearson's X^2)

Purpose: Determine whether or not there is a pattern of relationship between two qualitative variables

H0: There is no pattern of relationship between the two qualitative vars in the pop represented by the sample

Summary Statistic: X^2 has range from 0 to ∞

Basis & meaning of NHST:

- $p > .05$ retain H0: -- the pattern of relationship between the variables in the sample is not strong enough to conclude that there is a pattern of relationship between the variables in the population
- $p < .05$ reject H0: -- the pattern of relationship between the variables in the sample is strong enough to conclude that there is a pattern of relationship between the variables in the population

For 1 qualitative / numerical variables & 1 quantitative / numerical

Analysis of Variance (ANOVA -- also called an F-test)

Purpose: Determine whether or not the the populations represented by the different values of the qualitative variable have mean differences on the quantitative variable

H0: The populations with different values on the qualitative variable have the same mean on the quantitative variable

Summary Statistic: F has range from 0 to ∞

Basis & meaning of NHST:

- $p > .05$ retain H0: -- the mean difference in the sample is not strong enough to conclude that there is a mean difference between the populations
- $p < .05$ reject H0: -- the mean difference in the sample is strong enough to conclude that there is a mean difference between the populations

There is lots to learn about each of the statistical tests, but right now I want you to be sure you can tell when to use which one…

the "secret" is to figure out whether each variable is qualitative or quantitative, then you'll know which or the 3 stats to use !!

We want to know whether there is a relationship between someone's income and their amount of political campaign contributions.

IQ is ... quant   Contributions is ... quant   Stat? Pearson's r

We want to know whether gender fluid and gender binary voters make different amounts of political campaign contributions.

Gender is ... qual   Contributions is ... quant   Stat? F

We want to know whether being Democrat vs Republican is related to whether or not a person is likely to make a political contribution.

Registration is ... qual   Contributions is ... qual   Stat? Pearson's X^2

Here's a few more...

"relationship" expressions of hypotheses

- I expect there is a relationship between a person's height and their weight. — r
- I believe we'll find that there is a relationship between a person's athletic history (HS vs. not) and their weight. — F
- My hypothesis is that there is a relationship between a person's athletic history and whether or not currently work out. — X^2

"tend to..." expressions of hypotheses

- I expect that HS ath. tend to be heavier than non HS ath. — F
- My hypothesis is that taller folks also tend to be heavier — r
- I expect that folks who currently work out tend to have been HS athletes. — X^2

"if … then more likely…" expressions of hypotheses

- If you currently work out, then you are more likely to have been a HS athlete — X^2
- If you are heavier, then you are more likely to be taller. — r
- If you are lighter, then you are more likely to not be HS ath. — F