## Summary sheet from last time: Confidence intervals

- Confidence intervals take on the usual form: parameter = statistic ± $t_{crit}$ SE(statistic)

| parameter | SE |
|---|---|
| a | $s_e \cdot \text{sqrt}(1/N + m_x^2/ss_{xx})$ |
| b | $s_e/\text{sqrt}(ss_{xx})$ |
| y' (mean) | $s_e \cdot \text{sqrt}(1/N + (x_o - m_x)^2/ss_{xx})$ |
| $y_{new}$ (individual) | $s_e \cdot \text{sqrt}(1/N + (x_{new} - m_x)^2/ss_{xx} + 1)$ |

- Where $t_{crit}$ is with N-2 degrees of freedom, and $s_e = \text{sqrt}(\Sigma(y_i - y_i')^2/(N-2)) = \text{sqrt}((ss_{yy} - b \cdot ss_{xy})/(n-2))$

## Summary sheet from last time: Hypothesis testing

- Of course, any of the confidence intervals on the previous slide can be turned into hypothesis tests by computing $t_{obt}$ = (observed – expected)/SE, and comparing with $t_{crit}$.
- Testing $H_0$: $\rho=0$:
  - $t_{obt} = r \cdot \text{sqrt}(N-2)/\text{sqrt}(1-r^2)$
  - Compare with $t_{crit}$ for N-2 degrees of freedom.
- Testing $H_0$: $\rho=\rho_o$:
  - Need to use a different test statistic for this.

# Chi-square tests and non-parametric statistics

9.07

4/13/2004

## Statistics

- Descriptive
  - Graphs
  - Frequency distributions
  - Mean, median, & mode
  - Range, variance, & standard deviation
- Inferential
  - Confidence intervals
  - Hypothesis testing

## Two categories of inferential statistics

- Parametric
  – What we've done so far
- Nonparametric
  – What we'll talk about in this lecture

## Parametric statistics

- Limited to quantitative sorts of dependent variables (as opposed to, e.g., categorical or ordinal variables)
- Require dependent variable "scores" are normally distributed, or at least their means are approximately normal.
  – There exist some parametric statistical procedures that assume some other, non-normal distribution, but mostly a normal distribution is assumed.
  – The general point: parametric statistics operate with some assumed form for the distribution of the data.
- Sometimes require that population variances are equal

## Parametric statistics

- Best to design a study that allows you to use parametric procedures when possible, because parametric statistics are more *powerful* than nonparametric.
- Parametric procedures are robust – they will tolerate some violation of their assumptions.
- But if the data severely violate these assumptions, this may lead to an increase in a Type I error, i.e. you are more likely to reject $H_0$ when it is in fact true.

## Nonparametric statistics

- Use them when:
  – The dependent variable is quantitative, but has a very non-normal distribution or unequal population variances
  – The dependent variable is categorical
    - Male or female
    - Democrat, Republican, or independent
  – Or the dependent variable is ordinal
    - Child A is most aggressive, child B is 2nd most aggressive

## Nonparametric statistics

- The design and logic of nonparametric statistics are very similar to those for parametric statistics:
  - Would we expect to see these results by chance, if our model of the population is correct (one-sample tests)?
  - Do differences in samples accurately represent differences in populations (two-sample tests)?
  - $H_0$, $H_a$, sampling distributions, Type I & Type II errors, alpha levels, critical values, maximizing power -- all this stuff still applies in nonparametric statistics.

## Chi-square goodness-of-fit test

- A common nonparametric statistical test
  - Considered nonparametric because it operates on categorical data, and because it can be used to compare two distributions regardless of the distributions
- Also written $\chi^2$
- As an introduction, first consider a parametric statistical test which should be quite familiar to you by now…

## Coin flip example (yet again)

- You flip a coin 100 times, and get 60 heads.
- You can think of the output as categorical (heads vs. tails)
- However, to make this a problem we could solve, we considered the dependent variable to be the *number* of heads

## Coin flip example

- Null hypothesis: the coin is a fair coin
- Alternative hypothesis: the coin is not fair
- What is the probability that we would have seen 60 or more heads in 100 flips, if the coin were fair?

## Coin flip example

- We could solve this using either the binomial formula to compute p(60/100) + p(61/100) + …, or by using the z-approximation to binomial data
- Either way, we did a *parametric* test

## Solving this problem via z-approximation

- $z_{obt} = (0.60-0.50)/\text{sqrt}(0.5^2/100) = 2$
- $p \approx 0.046$
- If our criterion were $\alpha=0.05$, we would decide that this coin was unlikely to be fair.

## What if we were rolling a die, instead of tossing a coin?

- A gambler rolls a die 60 times, and gets the results shown on the right.
- Is the gambler using a fair die?

| Number on die | Number of rolls |
|---|---|
| 1 | 4 |
| 2 | 6 |
| 3 | 17 |
| 4 | 16 |
| 5 | 8 |
| 6 | 9 |

## We cannot solve this problem in a good way using the parametric procedures we've seen thus far

- Why not use the same technique as for the coin flip example?
  - In the coin flip example, the process was a binomial process
  - However, a binomial process can have only two choices of outcomes on each trial: success (heads) and failure (tails)
  - Here was have 6 possible outcomes per trial

## The multinomial distribution

- Can solve problems like the fair die exactly using the multinomial distribution
- This is an extension of the binomial distribution
- However, this can be a bear to calculate – particularly without a computer
- So statisticians developed the $\chi^2$ test as an alternative, approximate method

## Determining if the die is loaded

- Null hypothesis: the die is fair
- Alternative hypothesis: the die is loaded
- As in the coin flip example, we'd like to determine whether the die is loaded by comparing the observed frequency of each outcome in the sample to the expected frequency if the die were fair

## In 60 throws, we expect 10 of each outcome

- Some of these lines may look suspicious (e.g. freqs 3 & 4 seem high)
- However, with 6 lines in the table, it's likely at least one of them will look suspicious, even if the die is fair
  - You're essentially doing multiple "statistical" tests "by eye"

| Number on die | Observed frequency | Expected frequency |
|---|---|---|
| 1 | 4 | 10 |
| 2 | 6 | 10 |
| 3 | 17 | 10 |
| 4 | 16 | 10 |
| 5 | 8 | 10 |
| 6 | 9 | 10 |
| Sum: | 60 | 60 |

## Basic idea of the $\chi^2$ test for goodness of fit

- For each line of the table, there is a difference between the observed and expected frequencies
- We will combine these differences into one overall measure of the distance between the observed and expected values
- The bigger this combined measure, the more likely it is that the *model* which gave us the expected frequencies is not a good fit to our data
- The model corresponds to our null hypothesis (that the die is fair). If the model is not a good fit, we reject the null hypothesis

## The $\chi^2$ statistic

$$\chi^2 = \sum_{i=1}^{rows} \frac{(\text{observed frequency}_i - \text{expected frequency}_i)^2}{\text{expected frequency}_i}$$

## The $\chi^2$ statistic

- When the observed frequency is far from the expected frequency (regardless of whether it is too small or two large) the corresponding term is large.
- When the two are close, the term is small.
- Clearly the $\chi^2$ statistic will be larger the more terms you add up (the more rows in your table), and we'll have to take into account the number of rows somehow.

## Back to our example

| Number on die | Observed frequency | Expected frequency |
|---|---|---|
| 1 | 4 | 10 |
| 2 | 6 | 10 |
| 3 | 17 | 10 |
| 4 | 16 | 10 |
| 5 | 8 | 10 |
| 6 | 9 | 10 |
| Sum: | 60 | 60 |

$\chi^2 = (4\text{-}10)^2/10$
$+ (6\text{-}10)^2/10$
$+ (17\text{-}10)^2/10$
$+ (16\text{-}10)^2/10$
$+ (8\text{-}10)^2/10$
$+ (9\text{-}10)^2/10$
$= 142/10 = 14.2$

What's the probability of getting $\chi^2 \geq 14.2$ if the die is actually fair?

## Probability that $\chi^2 \geq 14.2$ if the die is actually fair

- With some assumptions about the distribution of observed frequencies about expected frequencies (not too hard to come by, for the die problem), one can solve this problem exactly.
- When Pearson invented the $\chi^2$ test, however (1900), computers didn't exist, so he developed a method for approximating this probability using a new distribution, called the $\chi^2$ distribution.

## The $\chi^2$ distribution

- The statistic $\chi^2$ is approximately distributed according to a $\chi^2$ distribution
- As with the t-distribution, there is a different $\chi^2$ distribution for each number of degrees of freedom
- In this case, however, the $\chi^2$ distribution changes quite radically as you change d.f.

## The $\chi^2$ distribution

- Random $\chi^2$ factoids:
  - Mean = d.f.
  - Mode = d.f. – 2
  - Median ≈d.f. - .7
  - Skewed distribution, skew ↓ as d.f. ↑

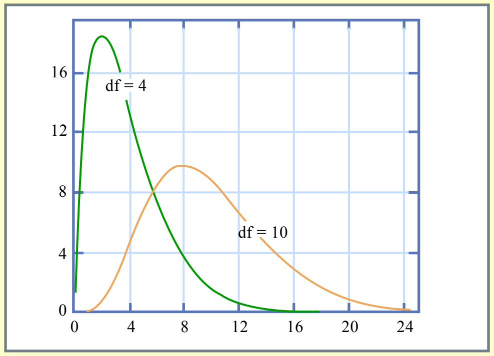
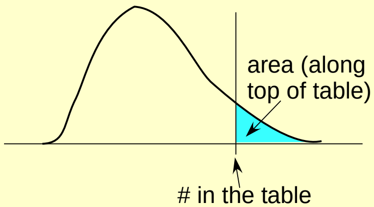



Figure by MIT OCW.

## The $\chi^2$ test

- Use the $\chi^2$ tables at the back of your book, for the appropriate d.f. and $\alpha$.




| Degrees of freedom | 99% | 95% | … | 5% | 1% |
|---|---|---|---|---|---|
| 1 | .00016 | .0039 | | 3.84 | 6.64 |
| 2 | 0.020 | 0.10 | | 5.99 | 9.21 |
| 3 | 0.12 | 0.35 | | 7.82 | 11.34 |
| 4 | 0.30 | 0.71 | | 9.49 | 13.28 |
| 5 | 0.55 | 1.14 | | 11.07 | 15.09 |

## What degrees of freedom do you use?

- Let k = # categories = # expected frequencies in the table = # terms in the equation for $\chi^2$
  - For the die example, k=6, for the 6 possible side of the die
- Then, d.f. = k – 1 – d
  - Where d is the number of population parameters you had to estimate from the sample in order to compute the expected frequencies

## Huh?

- This business about d will hopefully become more clear later, as we do more examples.
- For now, how did we compute the expected frequencies for the die?
  - With 60 throws, for a fair die we expect 10 throws for each possible outcome
  - This didn't involve any estimation of population parameters from the sample
  - d = 0

## Degrees of freedom for the die example

- So, d.f. = k – 1 = 6 – 1 = 5
- Why the "-1"?
  - We start off with 6 degrees of freedom (the 6 expected frequencies, $e_i$, i=1:6), but we know the total number of rolls ($\Sigma e_i = 60$), which removes one degree of freedom
  - I.E. if we know $e_i$, i=1:5, these 5 and the total number of rolls determine $e_6$, so there are only 5 degrees of freedom

## Back to the example

- We found $\chi^2 = 14.2$
- What is $p(\chi^2 \geq 14.2)$?
- Looking up in the $\chi^2$ tables, for 5 d.f:

| Degrees of freedom | 99% | 95% | … | 5% | 1% |
|---|---|---|---|---|---|
| 5 | 0.55 | 1.14 | | 11.07 | 15.09 |

- 14.2 > 11.07, but < 15.09
- p is probably a little bit bigger than 0.01 (it's actually about 0.014, if we had a better table))

## So, is the die fair?

- We're fairly certain ($p<0.05$) that the die is loaded.

## Rule of thumb

- Once again, this $\chi^2$ distribution is only an *approximation* to the distribution of $\chi^2$ values we'd expect to see
- As a rule of thumb, the approximation can be trusted when the expected frequency in each line of the table, $e_i$, is 5 or more

## When would this not be a good approximation?

- Example: suppose your null hypothesis for the die game was instead that you had the following probabilities of each outcome:

  (1: 0.01), (2, 0.01), (3, 0.01), (4, 0.95), (5, 0.01), (6, 0.01)

  In 100 throws of the die, your expected frequencies would be (1, 1, 1, 95, 1, 1)

  These 1's are too small for the approximation to be good. For this model, you'd need at least 500 throws to be confident in using the $\chi^2$ tables for the $\chi^2$ test

## Other assumptions for the $\chi^2$ test

- Each observation belongs to one and only one category
- Independence: each observed frequency is independent of all the others
  - One thing this means: $\chi^2$ is not for repeated-measures experiments in which, e.g., each subject gets rated on a task both when they do it right-handed, and when they do it left-handed.

## Summary of steps

- Collect the N observations (the data)
- Compute the k observed frequencies, $o_i$
- Select the null hypothesis
- Based upon the null hypothesis and N, predict the expected frequencies for each outcome, $e_i$, i=1:k
- Compute the $\chi^2$ statistic, $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$

## Summary of steps, continued

- For a given a, find $\chi^2_{crit}$ for k – 1 – d degrees of freedom
  - Note that the number of degrees of freedom depends upon the model, not upon the data
- If $\chi^2 > \chi^2_{crit}$, then reject the null hypothesis -- decide that the model is not a good fit. Else maintain the null hypothesis as valid.

## An intuition worth checking

- How does N affect whether we accept or reject the null hypothesis?
- Recall from the fair coin example that if we tossed the coin 10 times and 60% of the flips were heads, we were not so concerned about the fairness of the coin.
- But if we tossed the coin 1000 times, and still 60% of the tosses were heads, then, based on the law of averages, we were concerned about the fairness of the coin.
- *Surely, all else being equal, we should be more likely to reject the null hypothesis if N is larger.*
- Is this true for the $\chi^2$ test? On the face of it, it's not obvious. d.f. is a function of k, not N.

## Checking the intuition about the effect of N

- Suppose we double N, and keep the same relative observed frequencies
  - i.e. if we observed 10 instances of outcome i with N observations, in 2N observations we observe 20 instances of outcome i.
- What happens to $\chi^2$?

$$(\chi^2)_{new} = \sum_{i=1}^{k} \frac{(2o_i - 2e_i)^2}{2e_i} = \sum_{i=1}^{k} \frac{4(o_i - e_i)^2}{2e_i} = 2\chi^2$$

## Effect of N on $\chi^2$

- All else being equal, increasing N increases $\chi^2$
- So, for a given set of relative observed frequencies, for larger N we are more likely to reject the null hypothesis
- This matches with our intuitions about the effect of N

## What happens if we do the coin flip example the $\chi^2$ way instead of the z-test way?

- 100 coin flips, see 60 heads. Is the coin fair?
- Table:

| Outcome | Observed frequency | Expected frequency |
|---|---|---|
| Heads | 60 | 50 |
| Tails | 40 | 50 |
| Sum: | 100 | 100 |

- $\chi^2 = (60-50)^2/50 + (40-50)^2/50 = 4$
- Looking it up in the $\chi^2$ table under d.f.=1, it looks like p is just slightly < 0.05
  - We conclude that it's likely that the coin is not fair.

## Look at this a bit more closely…

- Our obtained value of the $\chi^2$ statistic was 4
- $z_{obt}$ was 2
- $(z_{obt})^2 = \chi^2$
- Also, $z_{crit}^2 = (\chi^2)_{crit}$ for d.f. = 1
- So, we get the same answer using the z-test as we do using the $\chi^2$ test
- This is always true when we apply the $\chi^2$ test to the binomial case (d.f. = 1)

## Another use of the $\chi^2$ test

- Suppose you take N samples from a distribution, and you want to know if the samples seem to come from a particular distribution, e.g. a normal distribution
  - E.G. heights of N students
- We can use the $\chi^2$ test to check whether this data is well fit by a normal distribution

## A $\chi^2$ test for goodness of fit to a normal distribution

- First, must bin the data into a finite number of categories, so we can apply the $\chi^2$ test:

| Height (inches) | Observed frequency | Expected frequency |
|---|---|---|
| h<62.5 | 5 | |
| 62.5≤h<65.5 | 18 | |
| 65.5≤h<68.5 | 42 | |
| 68.5≤h<71.5 | 27 | |
| 71.5≤h | 8 | |
| Sum: | 100 | |

## A $\chi^2$ test for goodness of fit to a normal distribution

- Next, we find the candidate parameters of the normal distribution, from the mean height, and its standard deviation
- It turns out that for this data, $m_h$ = 67.45", and s = 2.92"

## A $\chi^2$ test for goodness of fit to a normal distribution

- Given that mean and standard deviation, find the z-scores for the category boundaries
  - We will need to know what fraction of the normal curve falls within each category, so we can predict the expected frequencies

| Boundary | 62.5 | 65.5 | 68.5 | 71.5 |
|---|---|---|---|---|
| z | -1.70 | -0.67 | 0.36 | 1.39 |

## Get the area under the normal curve from -∞ to z, for each boundary value of z

- See http://davidmlane.com/hyperstat/z_table.html for a convenient way to get these values

| Boundary | 62.5 | 65.5 | 68.5 | 71.5 |
|---|---|---|---|---|
| z | -1.70 | -0.67 | 0.36 | 1.39 |
| Area below z | 0.0446 | 0.2514 | 0.6406 | 0.9177 |

## What fraction of the area under the normal curve is in each category?

| Boundary | 62.5 | 65.5 | 68.5 | 71.5 |
|---|---|---|---|---|
| z | -1.70 | -0.67 | 0.36 | 1.39 |
| Area below z | 0.0446 | 0.2514 | 0.6406 | 0.9177 |

| Height (inches) | Observed frequency | Fraction of the area |
|---|---|---|
| h<62.5 | 5 | 0.0446 |
| 62.5≤h<65.5 | 18 | 0.2068 |
| 65.5≤h<68.5 | 42 | 0.3892 |
| 68.5≤h<71.5 | 27 | 0.2771 |
| 71.5≤h | 8 | 0.0823 |
| Sum: | 100 | 1 |

## What frequency do we expect for each category?

- Multiply the expected fraction by the total number of samples, 100
- If you do any rounding, make sure the sum = N = 100!

| Height (inches) | Observed frequency | Fraction of the area | Expected frequency |
|---|---|---|---|
| h<62.5 | 5 | 0.0446 | 4 |
| 62.5≤h<65.5 | 18 | 0.2068 | 21 |
| 65.5≤h<68.5 | 42 | 0.3892 | 39 |
| 68.5≤h<71.5 | 27 | 0.2771 | 28 |
| 71.5≤h | 8 | 0.0823 | 8 |
| Sum: | 100 | 1 | 100 |

(Don't worry too much about one bin out of 5 having slightly fewer than 5 counts)

## Now (finally) we're ready to do the $\chi^2$ test

- Compute
  $\chi^2 = (5-4)^2/4 + (18-21)^2/21 + (42-39)^2/39 + (27-28)^2/28 + (8-8)^2/8$
  $= 1/4 + 9/21 + 9/39 + 1/28 = 0.9451$
- Degrees of freedom = k – 1 – d = 2
  - k = 5 (= # categories = # terms in $\chi^2$ eq'n)
  - d = 2 (= number of population parameters estimated from the data in computed expected frequencies. Estimated $\mu$ and $\sigma$.)
- Find critical value for $\alpha$=0.05: $\chi^2_{crit} = 5.99$
- The normal distribution seems to be a very good fit to the data (we maintain the null hypothesis as viable)

## Another example

- How to combine results across independent experiments
- Results in the lower tail of the $\chi^2$ distribution (as opposed to the upper tail, which is what we've looked at so far)

## The lower tail of the $\chi^2$ distribution

- In the previous examples, we looked in the upper tail of the $\chi^2$ distribution, to judge whether the disagreement between observed and expected frequencies was great enough that it was unlikely to have occurred by chance, if the null hypothesis model was correct
- We can also look in the lower tail, and judge whether the agreement between data and model is "too good to be true"

## Gregor Mendel and genetics

- In 1865, Mendel published an article providing a scientific explanation for heredity.
- This eventually led to a revolution in biology.

## Gregor Mendel and genetics

- Mendel ran a series of experiments on garden peas
- Pea seeds are either yellow or green. Seed color is an indicator, even before you plant it, of the genes of the child plant the seed would grow into.
- Mendel bred a strain that had only yellow seeds, and another that had only green seeds.
- Then he crossed the two, to get a $1^{st}$ generation yellow-green hybrid: all seeds yellow
- Then he crossed pairs of $1^{st}$ generation hybrids, to get $2^{nd}$ generation hybrids: approximately 75% of seeds were yellow, 25% were green

## Genetic model

- Mendel postulated a theory in which y (yellow) is dominant, and g (green) is recessive
  - y/y, y/g, and g/y make yellow
  - g/g makes green
- One gene is chosen at random from each parent
- First generation parents are either gy or yg. So how many green seeds and how many yellow seeds do we expect from them?

## Genetic model

- The model which describes the number of green and yellow seeds should (Mendel postulated) be one in which you randomly choose, with replacement, a yellow or green ball from a box with 3 yellow balls and one green ball
- In the long run, you'd expect the seeds to be roughly 75% yellow, 25% green

## One of Mendel's experiments

- In fact, this is very close to the results that Mendel got.
- For example, on one experiment, he obtained 8023 $2^{nd}$-generation hybrid seeds
- He expected 8023/4 ≈ 2006 green seeds, and got 2001
  - *Very* close to what's predicted by the model!

## Mendel, genetics, and goodness-of-fit

- Mendel's made a great discovery in genetics, and his theory has survived the rigor of repeated testing and proven to be extremely powerful
- However, R.A. Fisher complained that Mendel's fits were *too good* to have happened by chance. He thought (being generous) that Mendel had been "deceived by a gardening assistant" who knew what answers Mendel was expecting

## Pooling the results of multiple experiments

- The thing is, the experimental results just mentioned have extremely good agreement with theoretical expectations
- This much agreement is unlikely, but nonetheless could happen by chance
- But it happened for Mendel on every one of his experiments but one!
- Fisher used the $\chi^2$ test to pool the results of Mendel's experiments, and test the likelihood of getting so much agreement with the model

## Pooling the results of multiple experiments

- When you have multiple, *independent* experiments, the results can be pooled by adding up the separate $\chi^2$ statistics to get a pooled $\chi^2$.
- The degrees of freedom also add up, to get a pooled d.f.

## Pooling the results of multiple experiments: mini-example

- Exp. 1 gives $\chi^2$=5.8, d.f.=5
- Exp. 2 gives $\chi^2$=3.1, d.f.=2
- The two experiments pooled together give $\chi^2$=5.8+3.1=8.9, with 5+2=7 d.f.

## Fisher's results

- For Mendel's data, Fisher got a pooled $\chi^2$ value under 42, with 84 degrees of freedom
- If one looks up the area to the left of 42 under the $\chi^2$ curve with 84 degrees of freedom, it is only about 4/100,000
- The probability of getting such a small value of $\chi^2$, i.e. such good agreement with the data, is only 0.00004

## Fisher's results

- Fisher took Mendel's genetic model for granted – that wasn't what he was testing
- Based on his results, Fisher rejected his null hypothesis – that Mendel's data were gathered honestly – in favor of his alternative hypothesis – that Mendel's data were fudged

## A lower tail example – testing whether our data fits a normal distribution

- Recall our earlier example in which we wanted to know if our height data fit a normal distribution
- We got a $\chi^2$ value of 0.9451, with 2 degrees of freedom, and concluded from looking in the area of the upper tail of the $\chi^2$ distribution ($\chi^2$>0.9451) that we could not reject the null hypothesis
  - The data seemed to be well fit by a normal distribution
- Could the fit have been "too good to be true"?

## A lower tail example – testing whether our data fits a normal distribution

- To test whether the fit was too good, we look at the area in the lower tail of the $\chi^2$ distribution
- With $\alpha$=0.05, we get a critical value of $\chi^2_{0.05}$ = 0.10
- 0.9451 is considerably larger than this critical value
- We maintain the null hypothesis, and reject the alternative hypothesis = the fit is too good to be true. The data seem just fine.

## One-way vs. two-way $\chi^2$

- Our examples so far have been "one-way" $\chi^2$ tests.
  - Naming of tests is like for factorial designs
  - "One-way" means one factor
  - E.G. factor = number on die
  - E.G. factor = color of peas
  - E.G. factor = into which category did our sample fall?
- Two-way $\chi^2$: two factors

## Two-way $\chi^2$

- Often used to test independence
- E.G: are handedness (right-handed vs. left-handed vs. ambidextrous) and gender (male vs. female) independent?
- In other words, is the distribution of handedness the same for men as for women?
  - Test of independence = test of whether two distributions are equal

## Handedness vs. gender

- Suppose you have a sample of men and women in the population, and you ask them whether they are left-handed, right-handed, or ambidextrous. Get data something like this:

| | Men | Women |
|---|---|---|
| Right-handed | 934 | 1070 |
| Left-handed | 113 | 92 |
| Ambidextrous | 20 | 8 |
| Total: | 1067 | 1170 |

An m×n = 3×2 table

## Is the distribution of handedness the same for men as for women?

- It's hard to judge from the previous table, because there are more men than women. Convert to percentages, i.e. what percent of men (women) are right-handed, etc?

| | Men | Women |
|---|---|---|
| Right-handed | 87.5% | 91.5% |
| Left-handed | 10.6% | 7.9% |
| Ambidextrous | 1.9% | 0.7% |
| Sum: | 100% | ~100% |

(numbers in table don't quite add up right due to rounding – don't worry about it, we're just eyeballing the percentages)

## Is the distribution of handedness the same for men as for women?

- From this table, it seems that the distribution *in the sample* is not the same for men as for women.
  - Women are more likely to be right-handed, less likely to be left-handed or ambidextrous
- Do women have more developed left brains (are they more rational?)
- Do women feel more social pressure to conform and be right-handed?
- Or is the difference just due to chance? Even if the distributions are the same in the population, they might appear different just due to chance in the sample.

## Using the $\chi^2$ test to test if the observed difference between the distributions is real or due to chance

- Basic idea:
  - Construct a table of the expected frequencies for each combination of handedness and gender, based on the null hypothesis of independence (Q: how?)
  - Compute $\chi^2$ as before

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  - Compare the computed $\chi^2$ with the critical value from the $\chi^2$ table (Q: how many degrees of freedom?)

## Computing the table of expected frequencies

- First, take row and column totals in our original table:

| | Men | Women | Total |
|---|---|---|---|
| Right-handed | 934 | 1070 | 2004 |
| Left-handed | 113 | 92 | 205 |
| Ambidextrous | 20 | 8 | 28 |
| Total: | 1067 | 1170 | 2237 |

| | Men | Women | Total |
|---|---|---|---|
| Right-handed | 934 | 1070 | 2004 |
| Left-handed | 113 | 92 | 205 |
| Ambidextrous | 20 | 8 | 28 |
| Total: | 1067 | 1170 | 2237 |

- From this table, the percentage of right-handers in the sample is 2004/2237 ≈ 89.6%
- If handedness and sex are independent, the number of right-handed men in the sample should be 89.6% of the number of men (1067)
  (0.896)(1067) ≈ 956
- Similarly, the number of left-handed women should be 205/2237·1170 ≈ 107, and so on for the other table entries.

## The full table, and $\chi^2$ computation

| | *Observed* | | *Expected* | |
|---|---|---|---|---|
| | Men | Women | Men | Women |
| Right-handed | 934 | 1070 | 956 | 1048 |
| Left-handed | 113 | 92 | 98 | 107 |
| Ambidextrous | 20 | 8 | 13 | 15 |

$$\chi^2 = (934-956)^2/956 + (1070-1048)^2/1048 + (113-98)^2/98 + (92-107)^2/107 + (20-13)^2/13 + (8-15)^2/15 \approx 12$$

## Degrees of freedom

- When testing independence in an m×n table, the degrees of freedom is (m-1)×(n-1)
- There's again a "-d" term if you had to estimate d population parameters in the process of generating the table.
  - Estimating the % of right-handers in the population doesn't count
  - There wont be a "-d" term in any of our two-way examples or homework problems
- So, d.f. = (3-1)(2-1) = 2

## So, are the two distributions different?

$\chi^2 = 12$, d.f. = 2

- Take $\alpha = 0.01$
- $\chi^2_{crit} = 9.21$ -> $p<0.01$
- We reject the null hypothesis that handedness is independent of gender (alt: that the distribution of handedness is the same for men as for women), and conclude that there seems to be a real difference in handedness for men vs. women.

# Summary

- One-way $\chi^2$ test for goodness-of-fit
  - d.f. = k – 1 – d for k categories (k rows in the one-way table)
- Two-way $\chi^2$ test for whether two variables are independent (alt: whether two distributions are the same)
  - d.f. = (m-1)(n-1) – d for an m × n table
- Can test both the upper tail (is the data poorly fit by the expected frequencies?) and the lower tail (is the data fit too good to be true?)
- $\chi^2$ = sum[(observed – expected)$^2$ / expected]