Tests for Binomial Data & Proportions

	Variable A		Total
Variable B	 Latte	 Vanilla	Afernoon Tea
	 Cinnamon	 De-caf	Bedtime
Total	Special Treat	Tea Break	A good Cuppa!

The Chai-Squared Test

"Statistics are the grammar of science."

Karl Pearson (Mathmetician)

Binomial data

- Binomial data is data with 2 classes referenced in a binary format (0 and 1)
  Examples
  - Male/Female
  - Yes/No
  - Present/Absent
  - Alive/Dead
  - Susceptible/Resistant

- Sometimes it might be advantageous to convert hopelessly skewed data to binomial data rather than trying to analyze it with non-parametric tests
  - E.g. An ecology dataset with frequencies of plant species on sample plots can be easily converted to presence/absence data

- Tests for binomial data are just as powerful as test for normally distributed data because we reference the known binomial distribution

Binomial distribution

- Binomial distribution is a family of distributions because the shape references both the number of experiments/observations (e.g. Bernoulli trail) (*n*) and the probability of "*getting a success*" (*p*)

- **Bernoulli trial (or binomial trial)** - a random experiment with exactly two possible outcomes, "success" and "failure", in which the probability of success is the same every time the experiment is conducted

- For testing, the binomial distribution is frequently used to model the number of successes in a sample of size *n* drawn **with replacement** from a population of size N

- If the sampling is carried out **without replacement**, the draws are not independent and so the resulting distribution is a *hypergeometic distribution*, not a binomial one
  - However, for *N* much larger than *n*, the binomial distribution is a good approximation, and widely used

Contingency Tables

ID	SPECIES	SURVIVAL
1	A	Y
2	A	N
3	A	Y
4	B	Y
5	B	N
6	B	Y
…	…	…

Proportional test rely on contingency tables
So we have to reformat our data

Convert binary data (Y/N) into proportions by counting treatment totals within groups

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8

Proportion (*p*) is simply the total number of YES divided by the total observations (*n*)

Z-Test for Proportions

One Sample One Tailed Test

What is the probability that the true population proportion falls above/below a cutoff value (a)?

Example: $H_O: p < a$ $H_a: p > a$

$$z_{actual} = \frac{signal}{noise}$$

$$z_{actual} = \frac{(p-a)}{\sqrt{\frac{p*(1-p)}{n}}}$$

P-value (in R): `pnorm(z)`

Example: Does species B have a survival rate larger than 50% (arbitrary value)?

$H_O: p_B < 50$ $H_a: p_B > 50$

$$z_{actual} = \frac{(0.8-0.5)}{\sqrt{\frac{0.8*(0.2)}{20}}} = 3.3$$

`pnorm(3.3)` = 0.99
(right tail, but we need the left tail)

`1-pnrom(3.3)` = 0.0005
Reject H_o

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8

One sample one-tailed Z-test in R (better/easier option):

```
install.packages("corpora")
library(corpora)
z.score.pval(16,20,0.5, alternative="greater")
```

Z-Test for Proportions

One Sample Two Tailed Test

What is the probability that the true population proportion is equal to a cutoff value (a)?

Example: $H_O: p = a$ $H_a: p \neq a$

$$z_{actual} = \frac{signal}{noise}$$

$$z_{actual} = \frac{(p-a)}{\sqrt{\frac{p*(1-p)}{n}}}$$

P-value (in R): `pnorm(z)`

Example: Does species B have a survival rate equal to 50% (arbitrary value)?

$H_O: p_B = 50$ $H_a: p_B \neq 50$

$$z_{actual} = \frac{(0.8-0.5)}{\sqrt{\frac{0.8*(0.2)}{20}}} = 3.3$$

`pnorm(3.3)` = 0.99
(right tail, but we need the left tail)

`1-pnrom(3.3)` = 0.0005
Reject H_o

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8

One sample two-tailed Z-test in R (better/easier option):

```
install.packages("corpora")
library(corpora)
z.score.pval(16,20,0.5, alternative="two.sided")
```

Z-Test for Proportions

Two Sample Two-Tailed Test

Do samples A (p_1) and B (p_2) come from the same population?

Example: $H_O: p_1 = p_2$ $\quad H_a: p_1 \neq p_2$

Example: Is survival rate of Species A significantly different from Species B?

$H_O: p_A = p_B$ $\quad H_a: p_a \neq p_B$

$$z_{actual} = \frac{(0.8-0.4)}{0.15} = 2.67$$

`pnorm(2.67)` = 0.99
(right tail, but we need the left tail)

`1-pnrom(2.67)` = 0.0004
Reject H_o

$$z_{actual} = \frac{signal}{noise}$$

$$z_{actual} = \frac{(p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})} * \sqrt{\frac{n_1+n_2}{n_1 * n_2}}}$$

Pooled SE →

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

P-value (in R): `pnorm(z)`

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8

We cannot use the `z.score.pval()` function for two samples because does not allow for a pooled SE

Chi-Squared Test for Proportions (x^2)

Comparison between 2 or more groups

Is there a *treatment* effect?

If answer is *YES*, then use pairwise comparisons with adjusted p-values to find it

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8
Species C	24	16	40	0.6

Example: Is there any significant difference between survival proportions?

Example: $H_O: p_A = p_B = p_C$

$H_a: The\ proporitons\ are\ not\ equal$

Procedure:

1) Calculate TOTALS
2) Calculate AVERAGE %
   - This is called **Expected %**
   - E.g. If there was no effect or A=B=C
3) Calculate number of Expected Outcome for each treatment level
4) Calculate x^2 *(chi-squared statistic)*

Chi-Squared Test for Proportions (x^2)

Comparison between 2 or more groups

Is there a *treatment* effect?

If answer is *YES*, then use pairwise comparisons with adjusted p-values to find it

Example: Is there any significant difference between survival proportions?

Example: $H_O: p_A = p_B = p_C$

$H_a: The\ proporitons\ are\ not\ equal$

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8
Species C	24	16	40	0.6
TOTALS	**48**	**32**	**80**	
AVERAGE %	**0.6**	**0.4**		

Procedure:

1) Calculate TOTALS
2) Calculate AVERAGE %
   - This is called **Expected %**
   - E.g. If there was no effect or A=B=C
3) Calculate number of Expected Outcome for each treatment level
4) Calculate x^2 *(chi-squared statistic)*

Chi-Squared Test for Proportions (x^2)

Comparison between 2 or more groups

Is there a *treatment* effect?

If answer is *YES*, then use pairwise comparisons with adjusted p-values to find it

Example: Is there any significant difference between survival proportions?

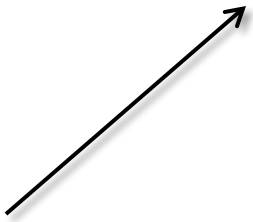
Example: $H_O: p_A = p_B = p_C$

$H_a: The\ proporitons\ are\ not\ equal$

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8
Species C	24	16	40	0.6
TOTALS	**48**	**32**	**80**	
AVERAGE %	**0.6**	**0.4**		

Procedure:

1) Calculate TOTALS
2) Calculate AVERAGE %
   - This is called **Expected %**
   - E.g. If there was no effect or A=B=C
3) Calculate number of Expected Outcome for each treatment level
4) Calculate x^2 *(chi-squared statistic)*



GROUPS	YES	NO
Species A	12	8
Species B	12	8
Species C	24	16

$$Expected\ Outcome = n * Expected\%$$

Chi-Squared Test for Proportions (x^2)

Comparison between 2 or more groups

Is there a *treatment* effect?

If answer is *YES*, then use pairwise comparisons with adjusted p-values to find it

Example: Is there any significant difference between survival proportions?

Example: $H_O: p_A = p_B = p_C$

$H_a: The\ proporitons\ are\ not\ equal$

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8
Species C	24	16	40	0.6
TOTALS	**48**	**32**	**80**	
AVERAGE %	**0.6**	**0.4**		

Procedure:

1) Calculate TOTALS
2) Calculate AVERAGE %
   - This is called **Expected %**
   - E.g. If there was no effect or A=B=C
3) Calculate number of Expected Outcome for each treatment level
4) Calculate x^2 *(chi-squared statistic)* ⟶

GROUPS	YES	NO
Species A	12	8
Species B	12	8
Species C	24	16

$$x^2 = \sum_i^n \frac{(observed - expected)^2}{n_{expected}}$$

Chi-Squared Test for Proportions (x^2)

Comparison between 2 or more groups

Is there a *treatment* effect?

If answer is *YES*, then use pairwise comparisons with adjusted p-values to find it

Example: Is there any significant difference between survival proportions?

Example: $H_O: p_A = p_B = p_C$

$H_a: The\ proporitons\ are\ not\ equal$

$$x^2 = \sum_i^n \frac{(observed - expected)^2}{n_{expected}}$$

P-value (in R): `pchisq(x^2,df)`

`pchisq(6.667,2)` = 0.96
(right tail, but we need the left tail)

`1-pchisq(6.667,2)` = 0.04
Reject H_o and follow up with pairwise test with adjusted p-values

GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8
Species C	24	16	40	0.6
TOTALS	**48**	**32**	**80**	
AVERAGE %	**0.6**	**0.4**		

GROUPS	YES	NO
Species A	12	8
Species B	12	8
Species C	24	16

Degrees of freedom *(df)* = number of groups - 1

Chi-Squared Test for Proportions (x^2)

Comparison between 2 or more groups

Is there a *treatment* effect?

If answer is *YES*, then use pairwise comparisons with adjusted p-values to find it

Example: Is there any significant difference between survival proportions?

Example: $H_O: p_A = p_B = p_C$

$H_a: The\ proporitons\ are\ not\ equal$

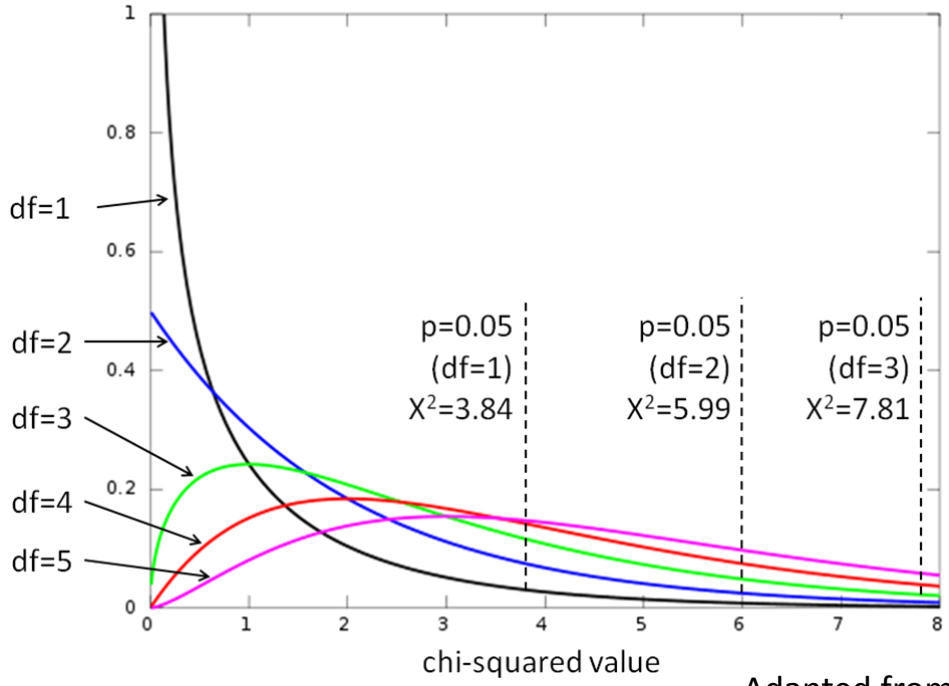
GROUPS	Survival YES	Survival No	*n*	*p*
Species A	8	12	20	0.4
Species B	16	4	20	0.8
Species C	24	16	40	0.6
TOTALS	**48**	**32**	**80**	
AVERAGE %	**0.6**	**0.4**		

GROUPS	YES	NO
Species A	12	8
Species B	12	8
Species C	24	16

Chi-squared Test in R:

```
output=chisq.test(contingencyMatrix)
output  # view the test output as normal
output$p.value  # returns only the p-value
output$statistic  # table of chi-squared value
output$observed  # table of observed counts
output$expected  # table of expected counts
```

Distribution of Chi-Squared Statistic (x^2)



Adapted from Wikipedia

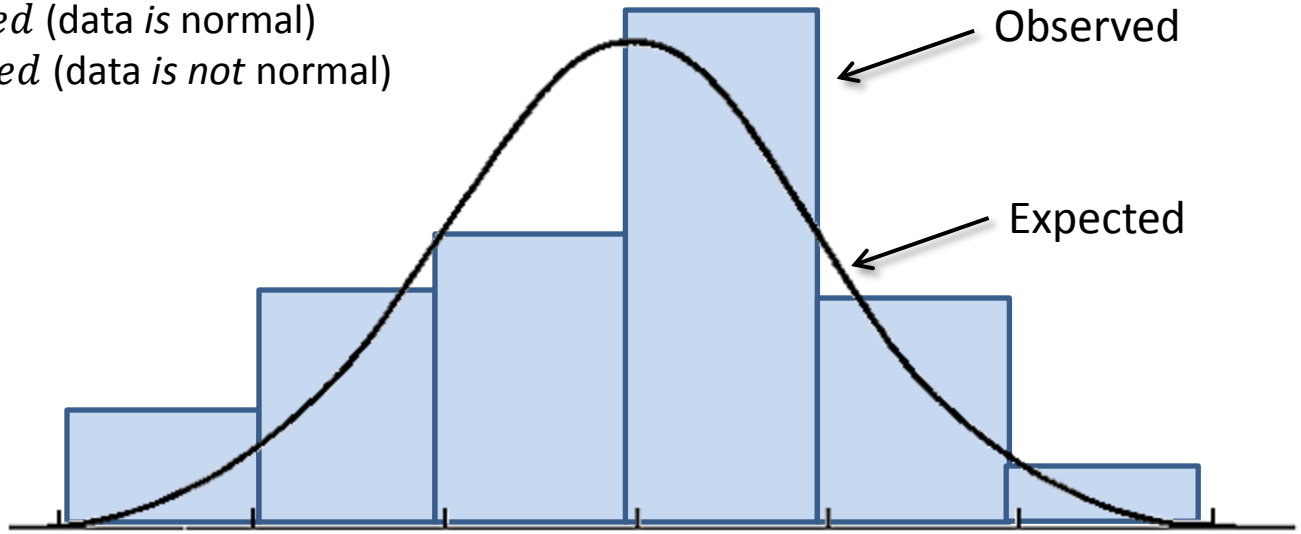
- Chi-squared is a family of distributions
- The distribution of the x^2statistic drastically changes in response to the number of groups tested
- This is reflected in the increasing value of x^2 needed to meet the $\alpha = 0.05$ threshold for hypotheses testing
- Therefore the more groups you test the bigger the difference between expected and observed needs to be (larger x^2statistic) to detect a difference between groups

Shaprio Test and Chi-Squared Test for Proportions

- Shapiro Test is actually based on Chi-squared Test!

Is there a difference between the normal curve (expected) and the histogram (observed) ?

$H_O: observed = expected$ (data *is* normal)
$H_a: observed \neq expected$ (data *is not* normal)



- If we calculate observed-expected and get a significant x^2 value, then there is a significant deviation from the normal distribution