Lecture 5
Statistics, eels, refugees and maternal smoking

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Outline

- Detecting differences in distributions
  - The χ^2 test
  - Refugees
  - Eels - comparing two sets of counts
  - Summary of hypothesis testing
- Detecting differences in numerics, e.g. means
  - Numerical data (maternal smoking)

Statistics, eels and refugees

Last lecture: χ^2 statistic:

$$\sum_i \frac{(O_i - E_i)^2}{E_i}$$

O_i = observed counts and E_i = expected count

This lecture:

- χ^2 test
- What's a hypothesis?
- New dataset "eels and habitat"
- p-value
- quantitative data

The χ^2 test

From the χ^2 statistic to the χ^2 test.

Categorical random variable X assigned to categories $c_1, c_2, \ldots, c_n$.

$p_i =$ probability of being in category c_i

e.g. π digits $p_1 = p_2 = \cdots = p_9 = p_{10} = 1/10$.

explicit values for p_i. In general though, p_i unknown.

AIHW data

Assigned values $q_1 = 0.7065, q_2 = 0.185, q_3 = 0.741, q_4 = .0343$

We will estimate $p_i = q_i$.

Hypothesis testing

Iraqi refugee population has a statistically different distribution to that of the Australian population.

Call this the *alternative hypothesis* H_1.

Want to know if data is consistent with this hypothesis.

Null hypothesis

To accept the alternative hypothesis, we see if we can reject the *null hypothesis* H_0.

H_0: Iraqi refugee population stress not statistically different to Australian reference population.

We try to see if the evidence is strong enough to reject H_0.

Court room analogy

Judge tries to see if an individual is guilty.

Data = evidence.

Null hypothesis is “individual is innocent until proven guilty”

If evidence allows us to reject, beyond all “reasonable” doubt, then accept guilty hypothesis.

Hypothesis testing

Start off by assuming null hypothesis is true, and see how it plays out.

Let overall refugee population have distribution r_1, r_2, r_3, r_4.

$$H_0 : p_1 = r_1, p_2 = r_2, p_3 = r_3, p_4 = r_4$$

and

$$H_1 : r_1, r_2, r_3, r_4 \text{ are unrestricted.}$$

χ^2 test

The χ^2 statistic: how much difference exists between observed counts and counts expected if no relationship at all in the population.

Assume null, calculate χ^2 statistic from data:

	low	moderate	high	very high
refugees	123.00	70.00	93.00	157.0
expected	312.99	81.96	32.84	15.2

$$\frac{(123 - 312.99)^2}{312.99} + \frac{(70 - 81.96)^2}{81.96} + \frac{(93 - 32.84)^2}{32.84} + \frac{(157 - 15.2)^2}{15.2}$$

$$= 1550.08$$

10,000 simulations using AIHW distribution, compare with 1550.

χ^2 small if the numbers returned are close to expected values, larger if deviate a lot.

Deviations are the interesting ones..

Deviations

Number of sets	Maximum chi-squared difference
10	8.54
100	10.79
1000	16.72
10000	21.68

$1500 >$ any of these.

Iraqi refugee distress levels are probably not the same as the Australian population.

Null hypo safely rejected, and stress levels vary (significantly).

Case was pretty clear-cut, but such strong conclusions are not always possible.

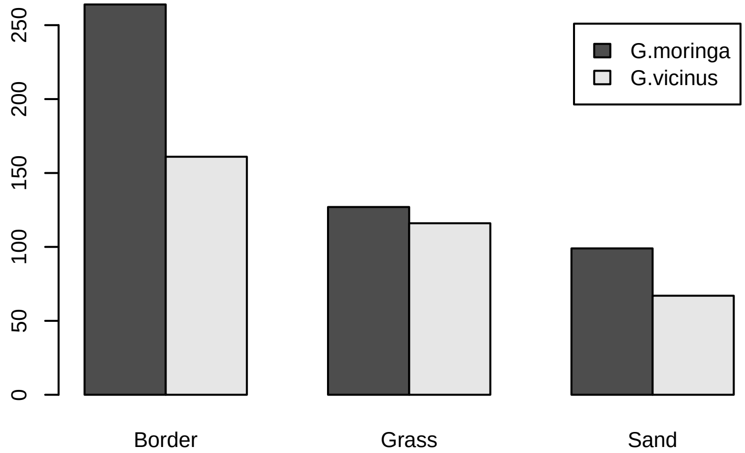
Simulations can throw up some more extreme ones.

Eels and habits

2 species of eels observed in 3 habitats

	Border	Grass	Sand
G.moringa	264	127	99
G.vicinus	161	116	67

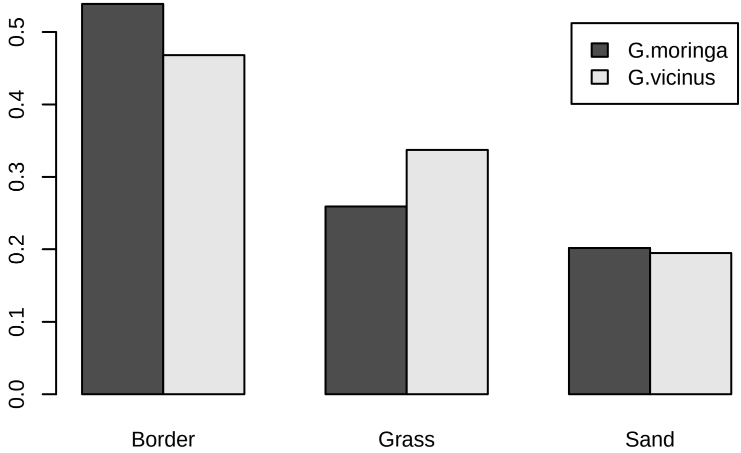
Are the distributions of eel species the same?



Fewer G.vicinus overall

Proportion of sightings in each habitat

We care about the proportions in habitats.



How to determine if it holds

Refugee data: asked if refugee distress levels were different

We knew the Australian distress levels from a very large sample.

Eels: lack of reference distribution.

Aggregate species

So we aggregate of both species together: habitat distribution of all 834 eels.

Need an expected value for each count and a way to simulate counts, under the assumption that the species distributions are the same.

If species distributions were the same:

	Border	Grass	Sand
counts	425	243	166
proportion	0.5096	0.2914	0.199

Let proportions be $p_1 = 0.5096$, $p_2 = 0.2914$ and $p_3 = 0.199$.

If the distribution is the same we expect to see the same percentage of each species count in each habitat.

If $n_1 = 490$ eels of species G.moringa, then we should see $n_1 p_1$ in the Border habitat and $n_1 p_2$ in Grass where p_2 is the proportion in Grass etc.

For G.vinicus, these are $n_2 p_1, n_2 p_2$ and $n_2 p_3$ respectively.

The data is:

	Border	Grass	Sand	Total
G.moringa	264	127	99	490
G.vicinus	161	116	67	344

and thus the expected counts are:

	Border	Grass	Sand
G.moringa	249.7	142.77	97.53
G.vicinus	175.3	100.23	68.47

To simulate what counts would look like if the two eel species shared the same distribution across habitats, for each eel, we sample a species using the proportions:

G.moringa	G.vicinus
0.58753	0.41247

and a habitat using the proportions from the data:

Border	Grass	Sand
0.5095923	0.2913669	0.1990408

Calculate the χ^2 distance for each simulation.

The null hypothesis is that "the distribution of eels across habitats does not differ".

If this distribution does differ, then the alternative hypothesis is appropriate.

χ^2 distance for simulated data

One simulation:

	Border	Grass	Sand
G.moringa	267	143	101
G.vicinus	173	86	64

Many simulations

Number of sets	Maximum chi-squared difference
10	6.08
100	10.71
1000	14.23
10000	17.62

χ^2 for actual eel counts is 6.26.

Bigger than the maximum from 10 simulation runs, but not for 100. Uh oh.

Cannot reject null hypothesis?

The p-value

After 1,000 simulations there are 46 greater than 6.26.

The eels are not further than all simulations, but most

It only happens 4.6% of the time.

This is the *p*-value.

The *p*-value is the chance or proportion of the time that we would see a chi-squared distance as large or larger than the actual distance for the data, if we simulate assuming the null hypothesis is true.

We are going to define it more precisely in the next lecture.

Summary of hypothesis testing

1. Compute a summary statistic (mean or χ^2) of the sample data.
2. Generate many simulations of the data given that the null hypothesis is true.
3. Compute the summary statistic of each simulated data to obtain a distribution of summary statistics given that the null hypothesis is true.
4. Compare the data statistic to the simulated data statistic.
5. If they look different, then it is likely that the null hypothesis is false.

Is simulation the only way?

Numerical data (maternal smoking)

All datasets so far were categorical

Numerical data is different

Extract of birth weights from study of 1,200 babies/mothers:

bwt	smoke
3429	no
3229	no
3657	yes
3514	no
3086	yes
3886	no

Does smoking affect the birth weight of infants?

Qualitative vs quantitative

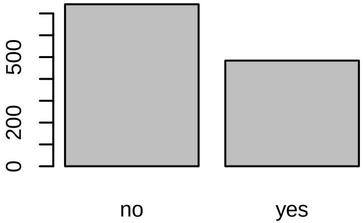
- `bwt`: the birth weight of the infant in grams
- `smoke`: the smoking status of the mother

Variables are quite different in nature.

`bwt` is a continuous measurement of weight. `smoke` is “yes” or “no”.

Qualitative variable `smoke`:

no	yes
742	484



Quantitative variable bwt has a significant value: its mean.

The mean birthweight is 3.466 Kg