Lecture 4
Independence, conditional probability, Monty Hall and refugees

Dr. Franco Ubaudi

The Nature of Data
Western Sydney University

Spring 2021

Outline

- Recap; review the key aspects learned so far
- uniform distribution
- independence
- conditional probability
- Monty Hall
- Some statistics!!

Recap

An unknown coin produced 69 heads and 31 tails

Could a fair coin produce that result?

$\therefore$ could the unknown coin be a fair coin?

Recap cont.

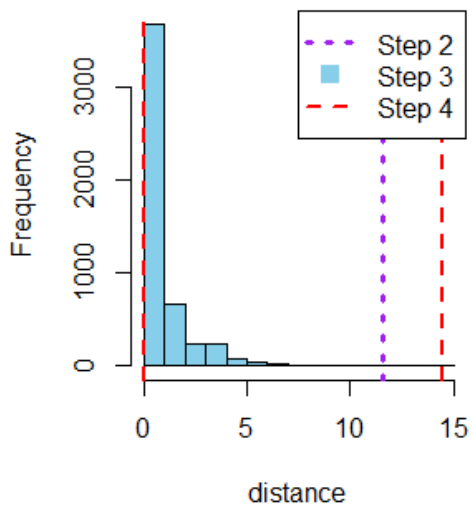
Coin flipping

```
#####################################
# foundations of hypothesis testing #
#####################################

coin <- c('H', 'T')   # coin definition
flips <- 100      # Number of coin flips
trials <- 5000      # Number of trials or experiments to perform

# Secret code to create our unknown coin results
if(TRUE)
{ ; }

coinResults          # Unknown coin results
exp <- c(0.5, 0.5) * flips  # What is expected from a fair coin

coinResults; exp      # View variable contents

# Calculate difference between unknown and expected coins
cs <- sum((coinResults - exp)^2 / exp)
cs

# Simulate a fair coin
d <- replicate(trials,
               {
                 obs <- sample(coin, flips, replace = TRUE)
                 obs <- table(obs)

                 sum((obs - exp)^2 / exp)
               })

```

Recap cont.

Coin flipping

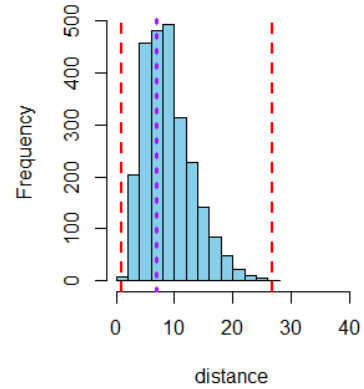
```r
####################################
# foundations of hypothesis testing #
####################################

coin <- c('H', 'T')   # coin definition
flips <- 100      # Number of coin flips
trials <- 5000      # Number of trials or experiments to perform

# Secret code to create our unknown coin results
if(TRUE)
{ ; }

coinResults          # Unknown coin results
exp <- c(0.5, 0.5) * flips  # What is expected from a fair coin

coinResults; exp      # View variable contents

# Calculate difference between unknown and expected coins
cs <- sum((coinResults - exp)^2 / exp)
cs

# Simulate a fair coin
d <- replicate(trials,
               {
                 obs <- sample(coin, flips, replace = TRUE)
                 obs <- table(obs)

                 sum((obs - exp)^2 / exp)
               })

```



Recap cont.

Digits of π

```r
# randomness of digits of pi
trials <-  2500

df <- read.csv('pi500.csv')
head(df)

digitCount <- nrow(df) / length(obs)

digitsOfPi <- table(df$pi.digits)
exp <- rep(1, 10) * digitCount

digitsOfPi; exp

# Calculate difference between observed and expected
cs <- sum((digitsOfPi - exp)^2 / exp)
cs

# Simulate a random digits
d <- replicate(trials,
               {
                 obs <- sample(0:9, nrow(df), replace = TRUE)
                 obs <- table(obs)

                 sum((obs - exp)^2 / exp)
               })
```



Uniform distribution

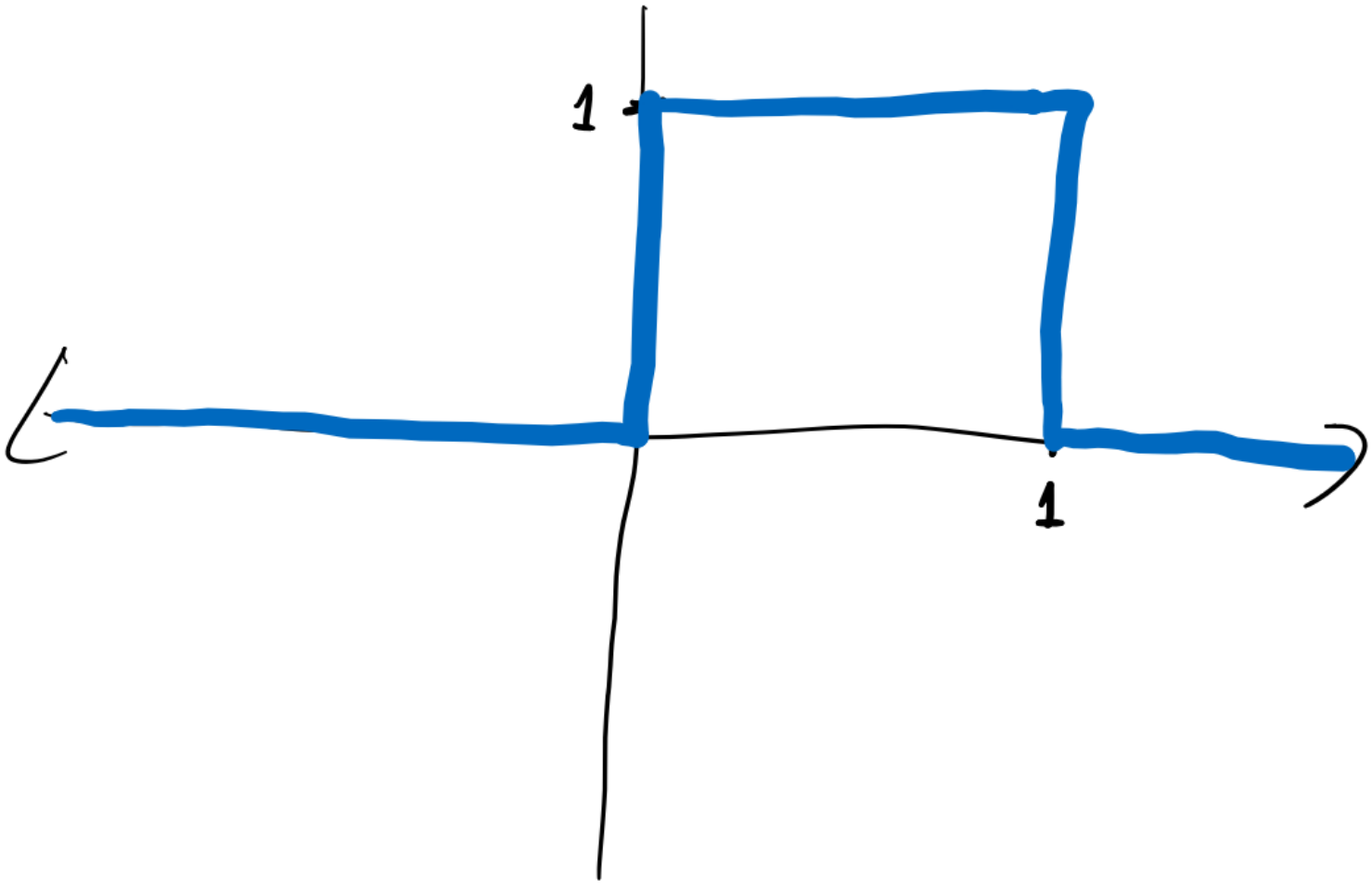
Digits of π were random If the digits were *uniformly* distributed each number $0, \ldots, 9$ being equilikely.

The *discrete uniform distribution*

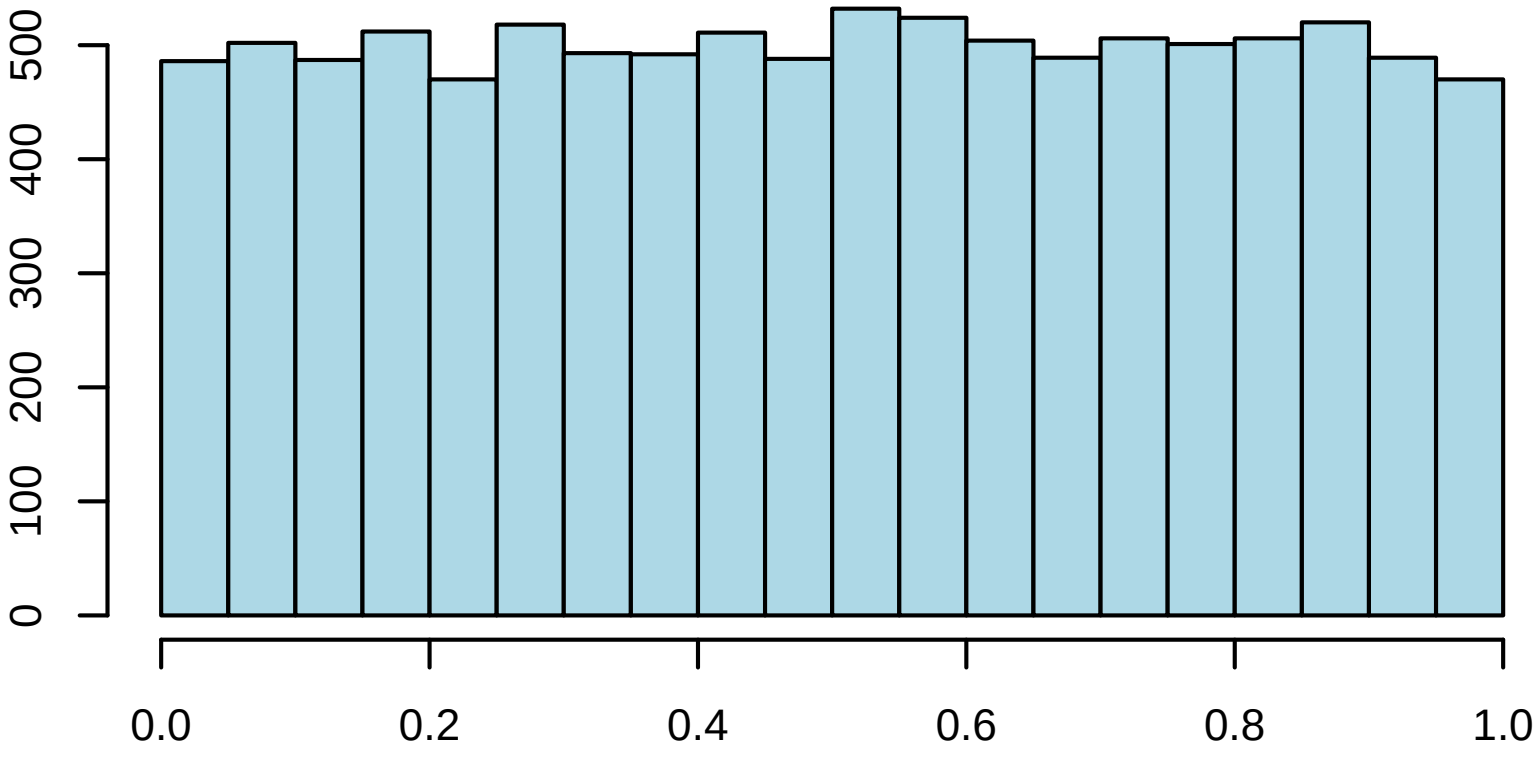
There is also the continuous version.

Continuous uniform distribution

Uniformly distributed over an interval – say $[0, 1]$



Simulated Uniform distribution



Independence

Events E and F from a probability space.

What is $P(E \cap F)$?

Does one depend on the other?

In general, we say E and F are independent if

$$P(E \cap F) = P(E)P(F)$$

Roll two fair dice

All rolls $1, \ldots, 6$ equilikely

$$\begin{aligned} P(\text{5 in roll 1, 6 in roll 2}) &= P(\text{5 in roll 1})P(\text{6 in roll 2}) \\ &= 1/6 * 1/6 = 1/36 \end{aligned}$$

Random variables

X and Y are independent if

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

2 dice again:
$X =$ the number achieved in roll 1, Y in roll 2.

$P(X = i) = P(Y = i) = 1/6$, for any possible dice outcome $i = 1, \ldots, 6$.

Dependence

Gambling based on sums of two dice

First roll a 5, then a 6, and the total is 11. What's the probability of 11?

36 combinations

$$(1,1),(1,2),(1,3),\ldots,(1,6);(2,1),\ldots,(2,6);\cdots;(6,1),\ldots,(6,6)$$

all just as likely.

Counts:

$$2,3,4,5,6,7;3,4,5,6,7,8;4,5,6,7,8,9;$$
$$5,6,7,8,9,10;6,7,8,9,10,11;7,8,9,10,11,12$$

Sum of two rolls

2 cases of 11 – probability of 11 a priori is $= 2/36$

5 appears 4 times.

Harder to get higher numbers.

Dependence

$X, Y =$ number achieved in rolls 1 and 2 respectively.

Total is random variable $Z = X + Y$

Z is not uniformly distributed: 5 is more likely than 11

Z is neither independent of X nor Y.

After first roll $x \in \{1, \ldots, 6\}$, bet on whether you will get at least a certain total 8.

The amount you can win depends on the risk.

Suppose first roll is 1: no chance of 8 at end. If first roll $= 2$, need a 6, so you've only a 1 in 6 chance. If you get a 3, your chances are getting better. And so on.

$X =$ value on first roll, $Y =$ value of second roll.

Probability $Z \geq z$ given that $X = x$, for some x and z?

conditional probability:

$$P(Z \geq z \,|\, X = x)$$

the probability of Z being greater than or equal to z given that $X = x$

Conditional probability

Events E and F.

$$P(E \mid F) := \frac{P(E \cap F)}{P(F)} \tag{1}$$

is the conditional probability that E occurs given that F has occurred.

Not defined if $P(F) = 0$

Independence and conditionality

IF E and F are independent.

$$P(E \mid F) = \frac{P(E)P(F)}{P(F)}$$
$$= P(E)$$

That makes sense: if then are independent events, then F occurring doesn't affect the probability of E occurring.

Independence vs uncorrelated

Correlation is an important topic we will cover later: about estimating how random variables vary together.

We want to estimate:

$$E[XY]$$

X and Y are said to be uncorrelated if $E[XY] = E[X]E[Y]$

Not independence where
$P(X = x \cap Y = y) = P(X = x)P(Y = y)$

Independence implies $E[XY] = E[X]E[Y]$

The reverse is not true. They may be uncorrelated but dependent.

Many people confuse these concepts.

Monty Hall: why you should change doors

Let the door containing the car be the random variable Y. The outcomes possible are $1, 2, 3$

Y's distribution is uniform:
$P(Y = 1) = P(Y = 2) = P(Y = 3) = 1/3$

For simplicity, assume you choose door 1 at beginning
and the car IS actually behind that door

Monty then opens door D: outcomes are door 2 or door 3, to show a goat

$P(D = 2) = P(D = 3) = 1/2$

Feels like the probability of you winning is $1/3$ no matter what you do.

Worksheet 1 simulation shows we win $2/3$ of the time by changing. Can we prove that?

If $Y = 1$ (car is behind door 1), Monty will pick either door 2 or 3.

$$\begin{aligned} P(Y = 1 \cap D = 2) &= P(Y = 1)P(D = 2) \\ &= 1/3 * 1/2 \\ &= 1/6 \end{aligned}$$

Similarly

$$P(Y = 1 \cap D = 3) = 1/6$$

However, if the car is behind door 2 or 3, Monty has one door that he can open, namely door 3 or door 2.

$$P(Y = 2 \cap D = 3) = 1/3 * 1 = 1/3$$
$$P(Y = 3 \cap D = 2) = 1/3 * 1 = 1/3$$

Given that Monty opens door 3, the probability to win by keeping door 1 is the conditional probability:

$$\begin{aligned} P(Y = 1 \,|\, D = 3) &= \frac{P(Y = 1 \cap D = 3)}{P(D = 3)} \\ &= \frac{1/6}{1/2} \\ &= \frac{1}{3} \end{aligned}$$

Thus the probability of losing by keeping door 1 is $\frac{2}{3}$

We could show the same argument if the host choose door 2.

Twice as likely to win by switching:

$$P(\text{keep door and win}) = 1/3$$
$$P(\text{keep and loose}) = 2/3$$

Iraqi Refugees

- In Uribe Guajardo et al. Int J Ment Health Syst (2016), the authors looked at the level of distress in 443 Iraqi refugees.
- The distress level is measured by a psychlogical instrument (known as the K10) and is classified as one of Low, Moderate, High or Very High distress.
- The following table was obtained

low	moderate	high	very high
123	70	93	157

Iraqi Refugees

- The Australian Institute of Health and Welfare (AIHW) also uses the K10 instrument to assess the Australian population.
- Using a very large sample (more than 10,000) they estimate the following;

low	moderate	high	very high
70.65	18.5	7.41	3.43

- Does the distribution of refugees distress differ from that of the Australian population?

Barplots

The first thing to do when confronted with data of this type is to draw a plot. Tables of numbers are harder to interpret.

- The type of plot that we would use for this data is a *bar plot*
- The data consists of counts in a series of categories
- The idea of a bar plot is to draw a bar or box, whose height represents the count, and draw one for each category

When to use barplots

Barplots are ideal for visualising the counts associated to a set of categories. In this case, we have the count of people associated to each distress level category.

Barplots

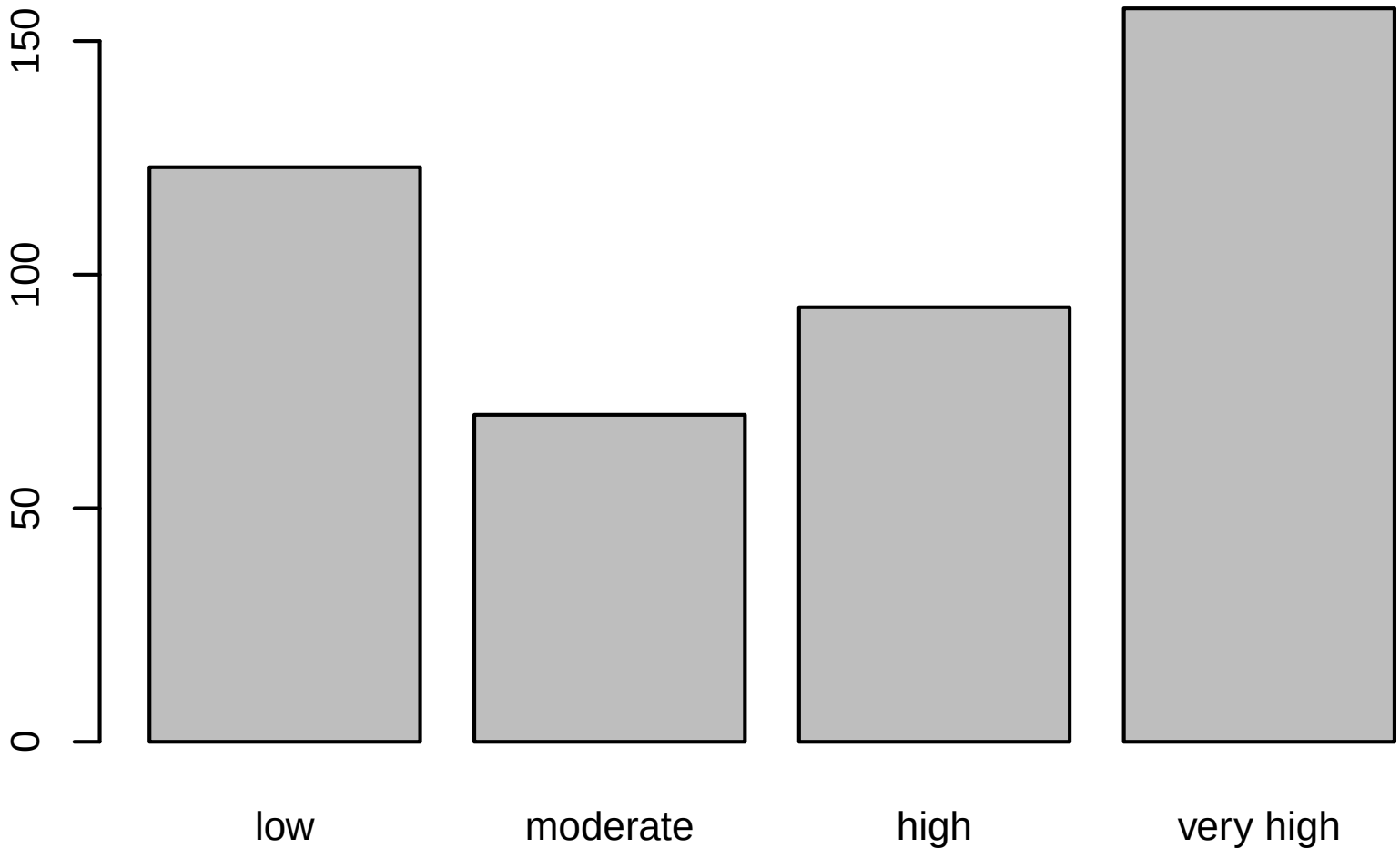


Figure: Barplot of numbers of refugees in each distress category

Barplots

Bar plots can be drawn vertically or horizontally

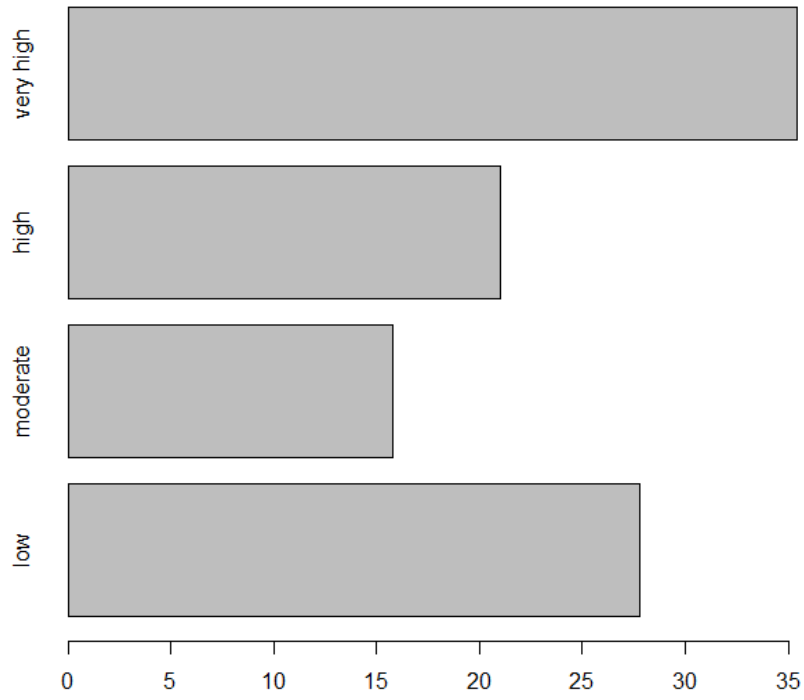


Figure: Horizontal barplot of numbers of refugees in each distress category

Barplots

Bar plot for the AIHW data

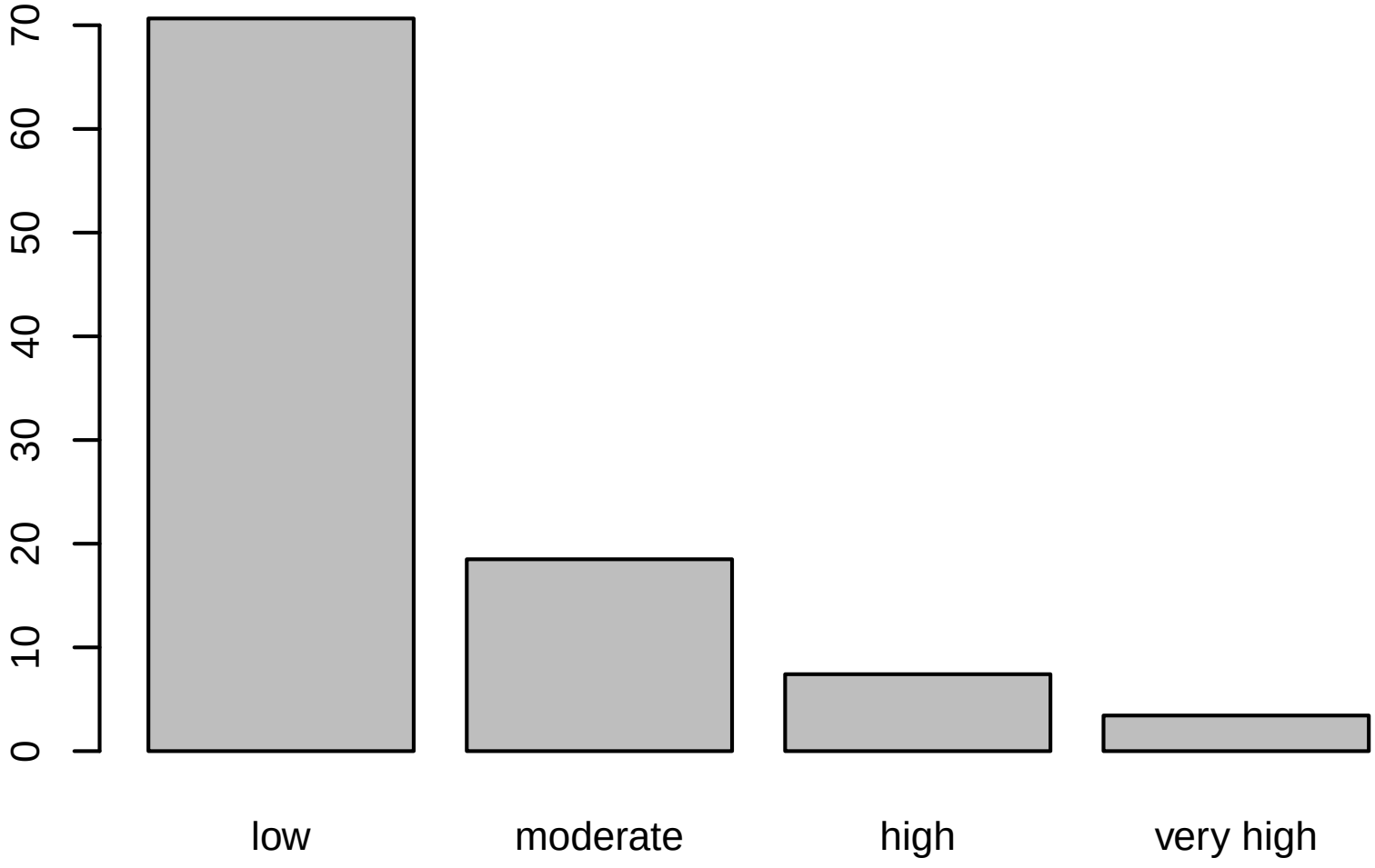


Figure: Barplot of the distress category proportions of the Australian population.

Barplots

Bar plots can be combined to compare sets of data. The data from each data set should be either:

- counts of occurances — an absolute comparison
- or percentages/proportions — when we are interested mainly in the distribution

For the refugee data we are mainly interested in whether a higher proportion of refugees are in certain distress categories than the Australian population. So we convert the counts to percentages (by dividing by the total and multiplying by 100)

Refugees v. the Australian population

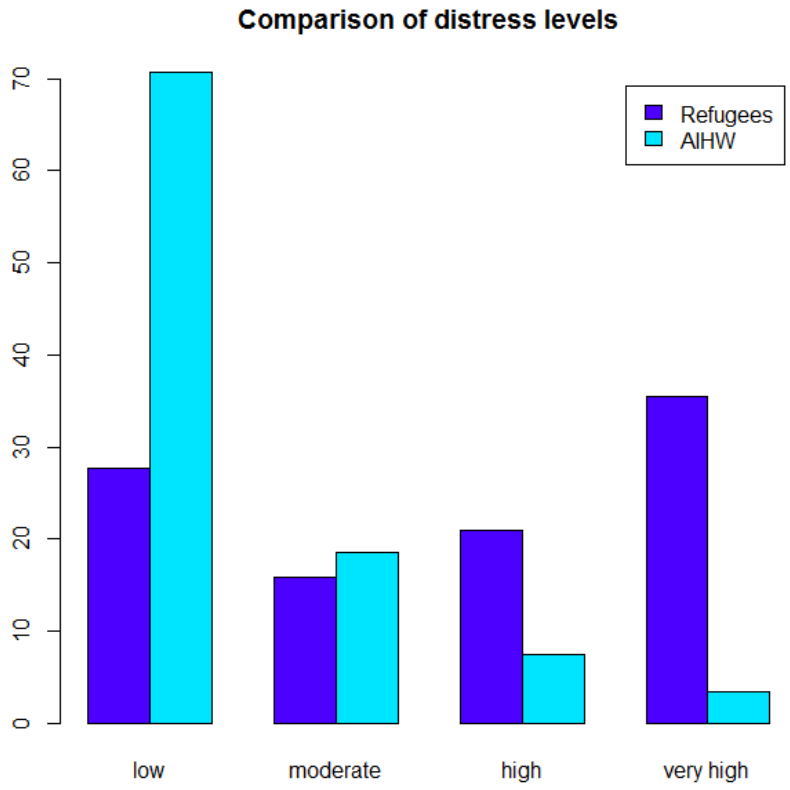


Figure: Comparing the distress categories of refugees and the Australian population.

Refugees v. the Australian population

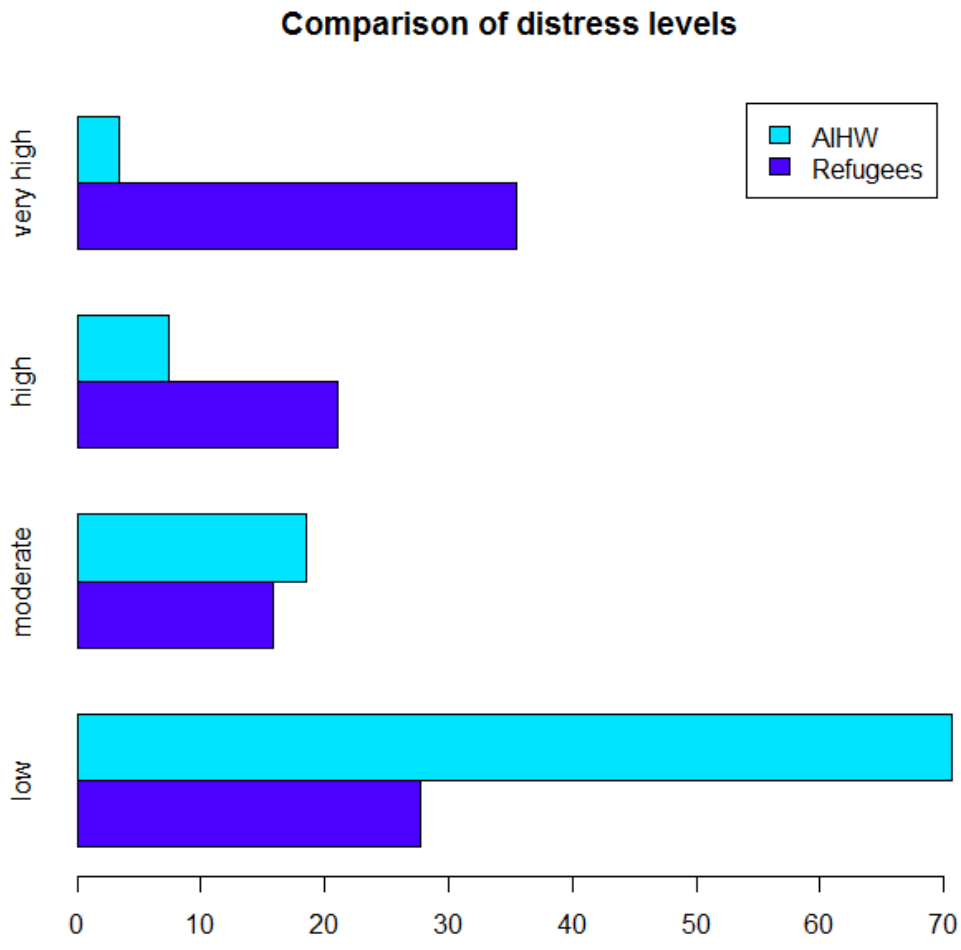


Figure: Comparing the distress categories of refugees and the Australian population.

Stacked Bars

- Sometimes stacked bars are used.

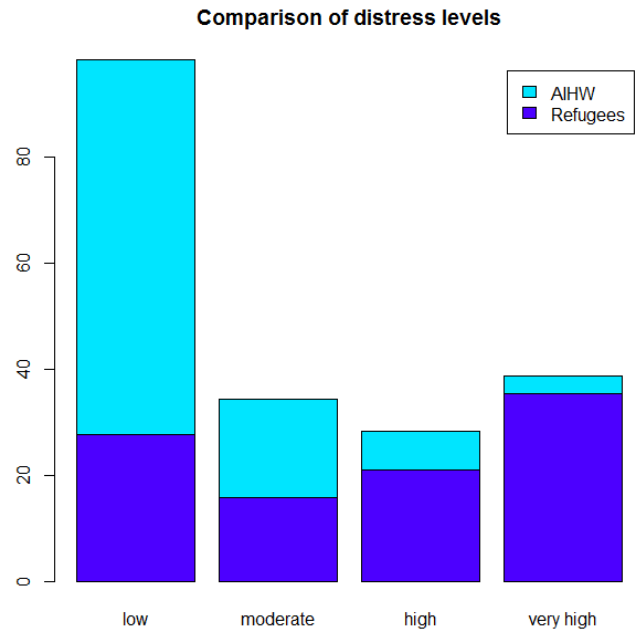


Figure: Comparing the distress categories of refugees and the Australian population.

Iraqi Refugees

Question of Interest

Does the distribution of a refugee's distress differ from that of the Australian population?

- Can we use a similar approach to that used for the digits of π?
- Doing so means that we would need to randomly allocate the 443 individuals to the four distress levels, then compare it to our sample. This would allow us to check if all distress levels are equally likely (the distribution is uniform).

But we are not interested in whether the distress level spread is uniform; we want to know if the distribution matches the AIHW percentages.

Simulation

We ask "what would the refugee distress disitributon look like if it matched the AIHW distribution?"

We need to simulate 443 individuals in the four categories so that the chances of being in each category are not uniform, but according to the AIHW percentages.

low	moderate	high	very high
70.65	18.5	7.41	3.43

For this example, we will round the numbers to simplify the simulation (if using a computer, this simplification is not needed).

low	moderate	high	very high
71	19	7	3

Simulation

If we generate a random number (x say) between 1 and 100, and call the generated category:

- low if x is less than 71
- medium if x is between 72 and 90 (=71+19) (inclusive)
- high if x is between 91 and 97 = (90+7)
- very high if x is greater than 97

Repeating this, then on average, 71 out of 100 will be low, 19 out of 100 will be medium etc.

This approach can be extended to non-integer percentages. R does this for us, (Excel does not as far as I can tell).

Refugees

Once we simulate the category counts, assuming that the AIHW percentages are true, we must ask "Are the category counts from the sample (shown below) consistent with the simulated counts from the AIHW proportions?"

low	moderate	high	very high
123	70	93	157

Refugees

Here are the simulation results repeated ten times.

low	moderate	high	very high
292	99	36	16
313	86	31	13
311	92	23	17
292	100	29	22
320	80	31	12
311	91	30	11
310	85	34	14
328	74	30	11
313	80	33	17
304	93	33	13

Expected count

In this case, we have 443 individuals in four categories, but they are not evenly spread.

- The AIHW has 70.65% in the low category
- So the expected count in this category is $443 \times 70.65/100 = 312.99$, since the sample size is 443.

Below are the remaining expected counts when using a sample size of 443:

	low	moderate	high	very high
percent	70.65	18.50	7.41	3.43
expected	312.99	81.96	32.84	15.20

Squared distance

We need to measure how different the category counts are from their expected value.

- So, as before, we can subtract the expected value from the actual counts and square and sum.
- We can do this for the simulated data from the AIHW proportions *and* the actual Refugee counts.

Refugees $= 59966.79$

741.69	24.56	204.92	827.19	66.57	111.42	20.98	314.28	7.11	207.6

A better distance

- It turns out that this distance does not give enough **weight** to differences where there a small expected counts versus where there are large expected counts.
- A better distance is to take the count minus the expected count squared **divided** by the expected **then** add up.

In maths notation,

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

This is called the *chi-square distance* (or χ^2)

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

- O_i stand for the observed counts for each i, eg O_1 is the first observed count.
- E_i stands for the corresponding expected count
- $\sum_i$ means sum or add up over i

Chi-square for Refugees

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

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

$$=1550.08$$

Chi-square for Simulated Counts

Problem

Compute the χ^2 distance between one of the similated sets of category counts and the expected set of category counts.

	low	moderate	high	very high
simulated	292.00	99.00	36.00	16.0
expected	312.99	81.96	32.84	15.2

Chi-squared for samples from AIHW

Finally, we examine the χ^2 distance for all simulated counts, and examine if the χ^2 distance for the data looks like these.

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

The distance for refugees is much larger than any of these. This indicates that Iraqi refugee distress levels are probably not the same as the Australian population.

Eels

Two species of eels are observed at three different habitats. The following counts are made.

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

Are the distribution of Eel species the same?

This problem is different to the Iraqi Refugees data, since we do not have a distribution to compare to (for the refugee data, we used the distribution from the AIHW).

Plots

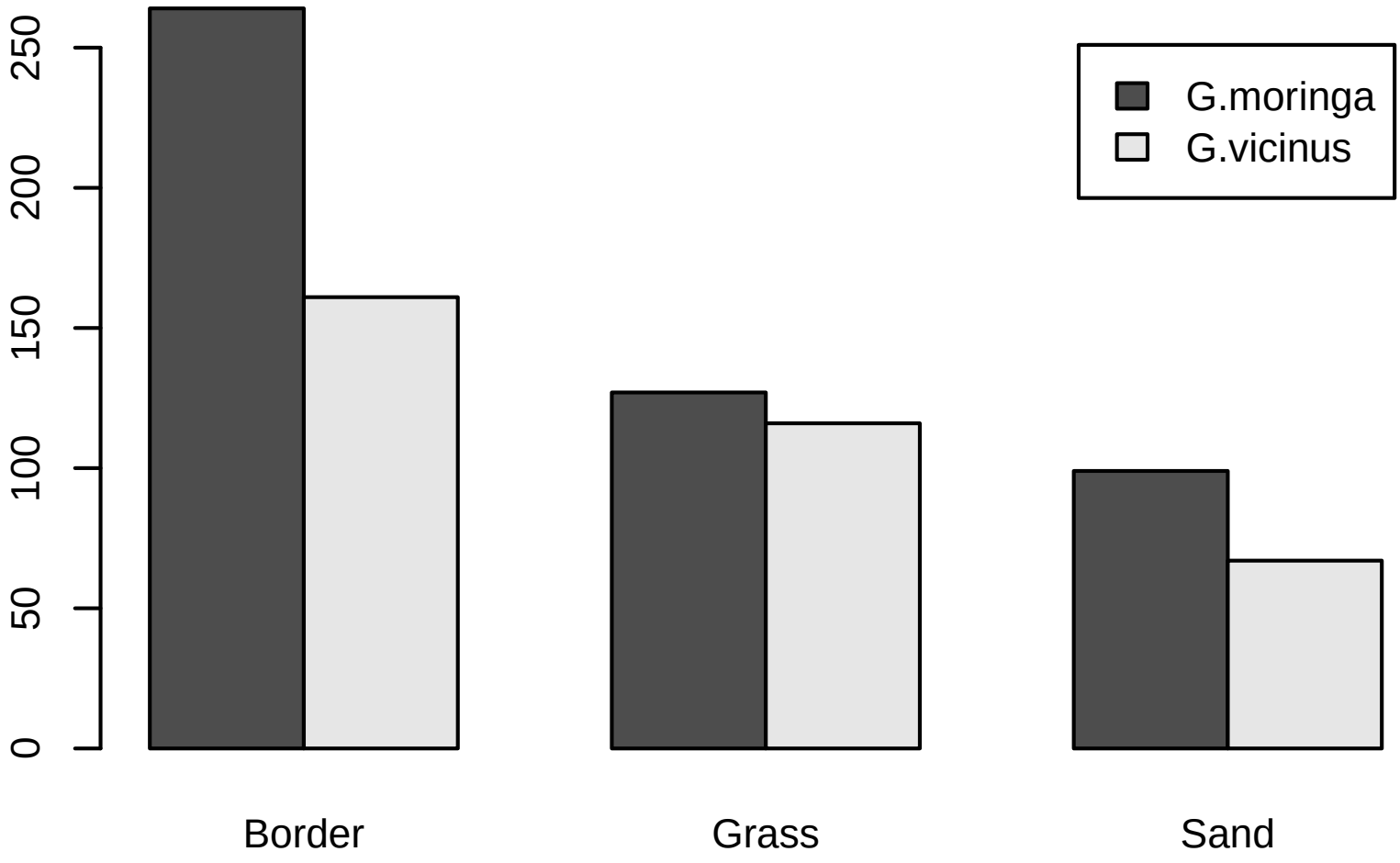


Figure: Eel counts at given habitat.

Eels

There certainly seems to be:

a. Fewer G.vicinus overall
b. A lot fewer G.vinicus in the Border habitat

Do we care about the overall number? Or just the spread/distribution?

Proportion of sightings in each habitat

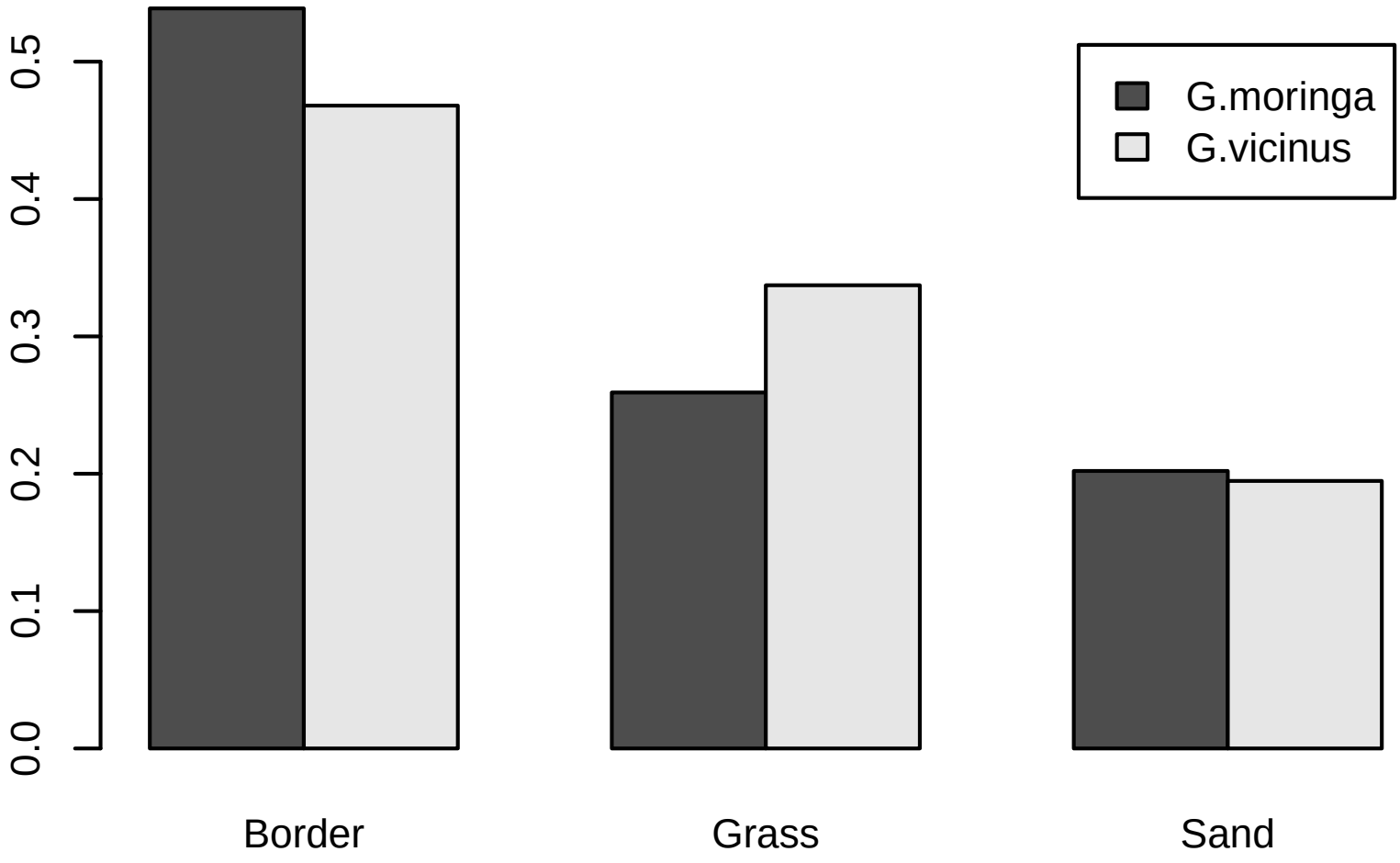


Figure: Eel proportions at given habitat.

Eels

This looks a bit like the refugees question, but it is subtly different.

- Previously, we asked if the distress level of refugees were different from the Australian population. And we essentially knew the average distress levels for the Australian population from a very large sample
- In this case, neither set of counts is all that large.
- To work as before we have to find an expected value for each count and a way to simulate counts, under the assumption that the species distributions are the same.

Expected counts

If the species distributions are the same, then we can aggregate the counts across species to estimate an overall habitat distribution

	Border	Grass	Sand
counts	425.00	243.00	166.0
percent	50.96	29.14	19.9

- So if the distribution is the same we expect to see the same percentage of each species count in each habitat.
- If there are $n_1 = 490$ eels of species G.moringa, then we should see $n_1 \times p_1$ in the Border habitat where p_1 is the proportion in Border,
- and $n_1 \times p_2$ in Grass where p_2 is the proportion in Grass etc.

Expected counts

The same applies to G.vinicus, with n_1 replaced with $n_2 = 344$

The data is:

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

The expected counts are:

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

Simulating

We need to simulate what the counts would look like if the two eel species shared the same distribution across habitats. To do this, for each of the 834 eels:

- we sample a species using the proportions from the data:

G.moringa	G.vicinus
0.58753	0.41247

- and sample a habitat using the proportions from the data:

Border	Grass	Sand
0.5095923	0.2913669	0.1990408

Simulating

- We would then proceed by calculating the chi-squared distance from expected for our actual data and for a (large) number of simulations and compare.
- Previously we were comparing to an (essentially) fixed set of expected values, but now we used the proportions in the data to compute them.
- So we must recompute the expected values for every simulation

χ^2 distance for simulated data

Problem

One simulation provided the counts below. Compute the χ^2 distance to the expected counts.

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

Simulating

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

The chi-square for the **actual** eel counts is 6.26

Its bigger than the max from 10 but not for a 100. . . Hmmm.

p-values

Lets look at this in more detail.

If we take 1000 of the simulated numbers (Im not going to print then here), it turns out that only 46 are greater than 6.26.

So although the eels are not further from expected than 1000 random simulations, they are than most of them.

Only 46 out of 1000 a further. Or 4.6%

This is called a *p*-value.

Some Terminology

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

The converse, that these distribution do differ, is the *alternative hypothesis*.

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

Fixed margins

There is an issue here, in that we have assumed a fixed number of eels were observed, but that the number in each species and in each habitat was not fixed.

Sometimes in two-way table like this might have some more fixed numbers.

Example 1: One fixed margin

- Suppose we are interested in the voting habits of Men versus Women.
- We find 500 men and 500 women and ask would they vote LNP, ALP or Green.
- In this design, the number of men and women is fixed in advance.

It is said to have one fixed **margin**

	ALP	Green	LNP
Men	177	120	203
Women	196	104	200

Example 2: Two fixed margins

- A wine taster is presented with 20 glasses of wine
- Ten are cool climate and ten warm climate wines.
- The taster is told there are ten of each and asked to say, by taste alone, which are the cool climate ones.

This table has two fixed margins - both rows and columns add up to 10

	Cool	Warm
Cool	8	2
Warm	2	8

When using the simulation approach, the simulation should follow the sampling strategy

Chi square distributions

So far, the hypothesis testing process has been:

1. Compute a summary statistic (mean, median, χ^2 distance) of the sample data.
2. Generate many simulations of the data, where the Null Hypothesis is true.
3. Compute the summary statistic of each simulated data to obtain a distribution of summary statistics where the Null Hypothesis is true.
4. Compare the data statistic to the simulated data statistic. If they look different, then it is likely that the Null Hypothsis is false.

When using the χ^2 distance, we can approximate steps 2 and 3 using the χ^2 distribution.

Chi square distributions

- The Chi-squared distribution (sometimes written χ^2-distribution) allows us to say approximately how typical a chi-squared distance is without simulating
- It lets us compute the proportion of times a value is exceeded if the **null** hypothesis is true
- It is actually a family of distributions, that depend on a parameter related to the number of counts being considered
- It requires the extra assumption that the expected cell counts are not too small (at least 5).