STAT 110: Week 10

University of Otago

Outline

- Contingency table
  - Looking at the relationship between two categorical variables
  - Investigate approaches to test independence of two categorical variables
  - Compare observed and expected counts
  - Introduce χ^2 distribution
- Central limit theorem
  - Investigate the sampling distribution for non-normal data
  - Generalise what was done for binomial data

Data: Passengers on the Titanic

- Data from the adult passengers on the titanic. Two variables:
  - Class: 1st, 2nd, 3rd or crew
  - Survived: yes or no

		survived		
		no	yes	Total
Class	1st	122	197	319
	2nd	167	94	261
	3rd	476	151	627
	Crew	673	212	885
	Total	1438	654	2092

- Do survival probabilities depend on the class?

Big picture

- We have investigated when both variables have two levels (groups)
- Here one of the variables has four levels
  - 1st – 3rd class, crew
- If the survival probabilities vary by class
  - The two variables (class and survival) are related
- If the survival probabilities do not vary by class
  - The two variables (class and survival) are independent
  - Knowing the class of a passenger tells us nothing about their survival probability
  - Recall: Definition of independence when we looked at probability
- Idea: Compare the observed data to what we would expect if two variables were independent

Expected counts

- We can use the margin totals to find the expected counts under independence

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

- Work through the Titanic table to understand this

Expected counts: Titanic

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} = \frac{319 \times 654}{2092} = 99.73$$

		survived		
		no	yes	Total
Class	1st		99.73	319
	2nd			261
	3rd			627
	Crew			885
	Total	1438	654	2092

- Proportion of passengers who are 1st class
  - $\frac{\text{row total}}{\text{table total}} = \frac{319}{2092} = 0.1525$
  - 15.25% of passengers are 1st class
- If survival and class are independent
  - Expected number is the total number of passengers who survive $\times$ the proportion of passengers who are 1st class
  - Or column total $\times \frac{\text{row total}}{\text{table total}}$

Expected counts: Titanic

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}} = \frac{627 \times 1438}{2092} = 430.99$$

		survived		
		no	yes	Total
Class	1st		99.73	319
	2nd			261
	3rd	430.99		627
	Crew			885
	Total	1438	654	2092

- Proportion of passengers who are 3rd class
  - $\frac{\text{row total}}{\text{table total}} = \frac{627}{2092} = 0.2997$
  - 29.97% of passengers are 3rd class
- If survival and class are independent
  - Expected number is the total number of passengers who died $\times$ the proportion of passengers who are 3rd class
  - Or $\text{column total} \times \frac{\text{row total}}{\text{table total}}$

Expected counts: Titanic

- Put it all together to give observed (black) and expected (blue)

		survived		
		no	yes	Total
Class	1st	122 (219.27)	197 (99.73)	319
	2nd	167 (179.41)	94 (81.59)	261
	3rd	476 (430.99)	151 (196.01)	627
	Crew	673 (608.33)	212 (276.67)	885
	Total	1438	654	2092

- The observed and expected counts will vary: there is natural variation in the data
  - Do they vary more than we would expect if variables are truly independent?

Test for independence

- We can look at this with a hypothesis test
  - H_0 : the two variables are independent
  - H_A : the two variables are related
- The test statistic we will use is

$$X^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

  - For each cell we calculate $\frac{(\text{observed}-\text{expected})^2}{\text{expected}}$ and add them up

Test statistic

		survived		
		no	yes	Total
Class	1st	122 (219.27)	197 (99.73)	319
	2nd	167 (179.41)	94 (81.59)	261
	3rd	476 (430.99)	151 (196.01)	627
	Crew	673 (608.33)	212 (276.67)	885
	Total	1438	654	2092

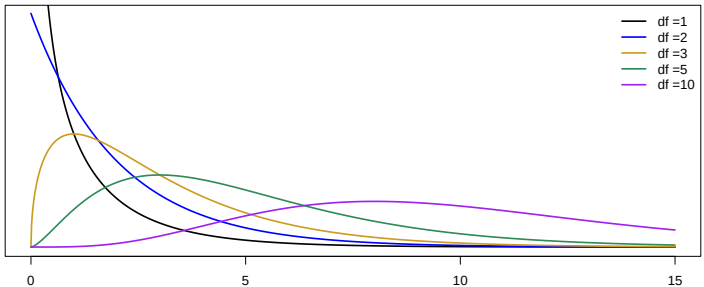
$$X^2 = \frac{(122 - 219.27)^2}{219.27} + \frac{(197 - 99.73)^2}{99.73} + \ldots + \frac{(212 - 276.67)^2}{276.67}$$
$$= 177.8$$

Test statistic

- If the null hypothesis is true
  - The test statistic, X^2, will be a realisation from a χ^2-distribution with $(R-1) \times (C-1)$ degrees of freedom
    - R is the number of rows; C is the number of columns
- Titanic data: $R = 4$, $C = 2$
  - $df = (4-1) \times (2-1) = 3$

Detour: χ^2-distribution

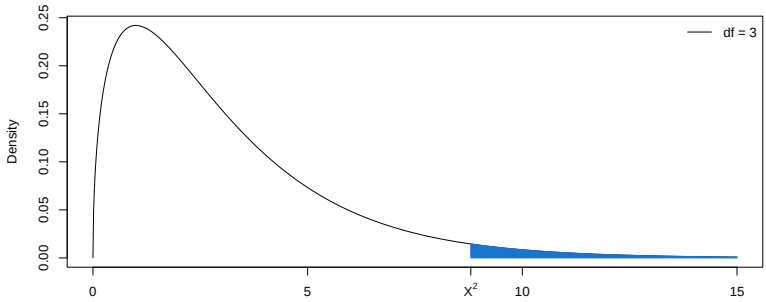
- The χ^2-distribution is a distribution for positive random variables



  - It is asymmetric (positively skewed)
  - It has one parameters: degrees of freedom

Finding a p-value

- An extreme X^2-value is one that is as large, or larger, than that observed
  - Indicative of increased divergence between observed and expected counts



- The p-value (blue area) is given by `1-pchisq(X2, df)`
  - `pchisq(X2, df)` gives probability of a value less than X^2

In R

- Data: each row is an observation
  - Titanic data: each row is a passenger
- Import into R

```
titanic = read.csv('titanic.csv')
head(titanic)
```

```
##   Class Survived
## 1  Crew      Yes
## 2  Crew      Yes
## 3   2nd       No
## 4   1st      Yes
## 5  Crew      Yes
## 6   3rd       No
```

In R

- We use the `table` function to obtain contingency table

```
titan = table(titanic$Class, titanic$Survived)
```

  - First argument: variable 1 (class of passenger)
  - Second argument: variable 2 (survived: yes / no)

```
titan
```

```
##
##         No Yes
##   1st  122 197
##   2nd  167  94
##   3rd  476 151
##   Crew 673 212
```

```
addmargins(titan)
```

```
##
##          No  Yes  Sum
##   1st   122  197  319
##   2nd   167   94  261
##   3rd   476  151  627
##   Crew  673  212  885
##   Sum  1438  654 2092
```

- The function `addmargins` includes the margins on the table

In R

- The R function `chisq.test` evaluates the test

```
out1 = chisq.test(titan)
out1
##
##  Pearson's Chi-squared test
##
## data:  titan
## X-squared = 177.8, df = 3, p-value <2e-16
```

- The p-value $< \alpha = 0.05$. Observing a test statistic as large as we did is unusual if the two variables were independent
  - Evidence in support of H_A: that the variables are not independent

χ^2-test

- If $R = 2$ and $C = 2$: we have a 2×2 contingency table, e.g. smallpox in Boston
  - The χ^2 test is identical to test for difference in proportions
  - $H_0 : p_1 - p_2 = 0$ and $H_A : p_1 - p_2 \neq 0$
- The χ^2 test can also be used if both $R > 2$ and $C > 2$
- The χ^2 test is unreliable if any of the expected counts < 5
  - Options for resolving this problem are beyond the scope of course

In R

- The `chisq.test` function can return the expected counts

```
out1$expected
```

```
##
##            No     Yes
##   1st  219.27  99.726
##   2nd  179.41  81.594
##   3rd  430.99 196.012
##   Crew 608.33 276.668
```

- Still important to know:
  - How to calculate them
  - What they represent (expected counts if variables are independent)

Normal approximation

- Binomial: The sampling distribution for $\hat{p}$ was approximated by a normal
  - Provided n is large, and p is not too close to 0 or 1
- This formed the basis for finding confidence intervals (and conducting hypothesis tests)
- Does this result generalise?
  - Will this also happen for other 'non-normal' distributions?

Central limit theorem

- If we collect a large sample of independent observations from a population with mean μ and standard deviation σ, the sampling distribution of $\bar{y}$ will be approximately normal
  - Mean μ
  - Standard error $\frac{\sigma}{\sqrt{n}}$
- This is known as the central limit theorem

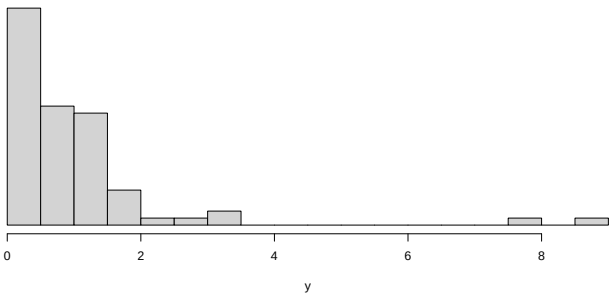
Central limit theorem: notes

- The distribution of y need not be normal
- What is a large sample?
  - A standard rule of thumb is $n > 30$
  - Lots of exceptions to this rule, e.g.
    - If the data are highly skewed, we likely need more than 30
    - If there are (extreme) outliers, we likely need more than 30

Central limit theorem: Example

- If data come from a non-normal distribution (an exponential distribution)
- Simulate one data set to see what the data looks like

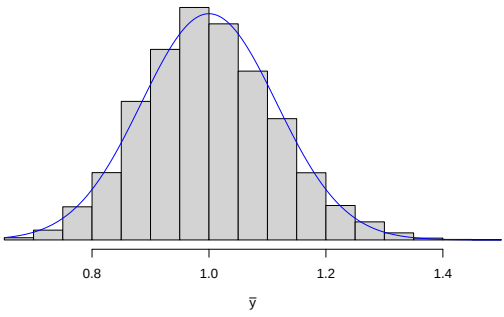
```
### Single sample
n = 75 # sample size of 75
y = rexp(n,1) # generate the data
hist(y) # look at histogram
```



Central limit theorem: Example

- Generate lots of datasets and visualize the sampling distribution
  - See that it is approximately normal

```
### Taking 10000 samples
m = 10000; ybar = rep(NA, m)
for(i in 1:m){ # repeat m times
  y = rexp(n,1) # simulate data
  ybar[i] = mean(y) # find the sample mean
}
hist(ybar)
```



Central limit theorem: implications

- The approaches we worked through for normal models
  - Can also be used for non-normal models
  - Need to ensure a large sample (usually $n > 30$)
- This list includes confidence intervals and hypothesis tests for:
  - Population mean μ with one sample: use `t.test`
  - Difference in two means $\mu_1 - \mu_2$: use `t.test`
  - ANOVA: use `aov`
  - Linear regression: use `lm`

Central limit theorem: implications

- Model checking: this is why we were only concerned about major departures from normality when the sample size was large
  - Linear regression
  - Normal models
- The central limit theorem underpins a lot of statistical practice
  - Often in the background

Summary

- χ^2 test for independence of contingency table
  - Idea: compare observed counts with those expected under independence
- Central limit theorem
  - Sampling distribution is normal
  - The approaches we have already developed can be used for non-normal data
- CLT holds if sample size is large
  - Usually $n > 30$

Outline

- Explore some non-parametric methods
- Focus on two examples:
  - Data from two independent groups
  - Relationship between two ordinal variables
- Outline other approaches

Data: Hawks

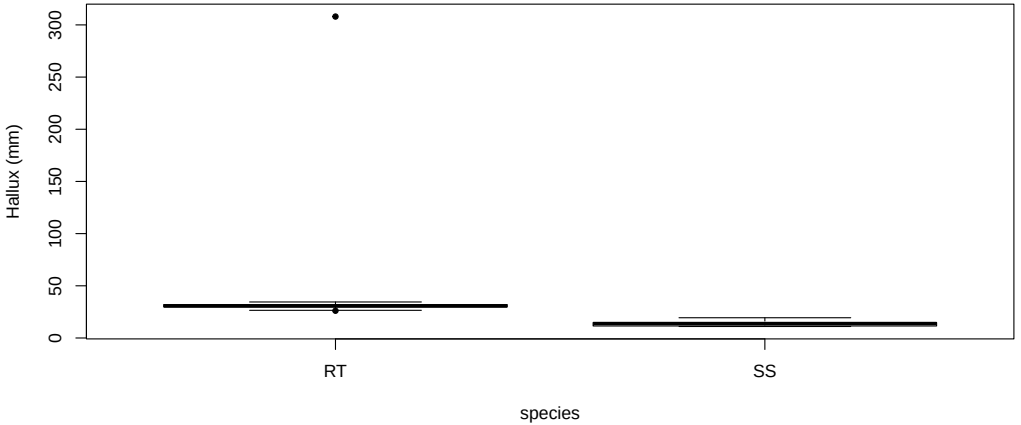
- 100 measurements from two species of hawk
  - Red-tailed (RT), and Sharp-shinned (SS)
- Hallux measurement (mm): length of the killing talon
- Import the data into R

```
hawk = read.csv('hawk.csv')
```

- Look at the first few lines

```
head(hawk)
##   hallux species
## 1   32.9      RT
## 2   29.9      RT
## 3   11.0      SS
## 4   31.2      RT
## 5   33.0      RT
## 6   30.5      RT
```

Data: Hawks



What is the state of play?

- We have developed a variety of statistical models for data
  - Normal models
  - Binomial models
  - Central limit theorem
    - We can use difference in two means, ANOVA, linear regression, etc, even if data are non-normal
    - Require a large sample
- There may be situations where these methods may be inappropriate
  - We may be unwilling to assume the data is normal
  - We may be unwilling to rely on the CLT
    - e.g. outliers or skew
- Introduce non-parametric methods

Idea: look at ranks

- We rank the observations
  - From 1 to n, smallest to largest (or vice versa)
- Work with the ranks rather than the actual observations
- It can be useful with (extreme) outliers
  - Same rank irrespective of whether the largest observation is 0.1 units larger than 2nd biggest observation, or 10000 units larger
- It can be useful if there is a lot of skew
  - All ranks are equally far apart from each other

Example: ranking data

- Suppose that we had the following data

Group A	Group B
1.2	5.5
4.3	1.7
3.1	2.9

- The ranks are given alongside (in blue)

Group A	Group B
1.2 (1)	5.5 (6)
4.3 (5)	1.7 (2)
3.1 (4)	2.9 (3)

In R

- The R function `rank` will rank data

```
hawk$rank = rank(hawk$hallux)
```

- This code: ranks the hallux measurements
  - Inserts a new variable (`rank`) into the `hawk` data frame

```
head(hawk)
```

```
##   hallux species rank
## 1   32.9      RT   92
## 2   29.9      RT   55
## 3   11.0      SS    1
## 4   31.2      RT   75
## 5   33.0      RT   93
## 6   30.5      RT   66
```

What now?

- We can compare the ranks of the two groups
- Hypothesis test
  - H_0: the distribution for the two groups are the same
  - H_A: the distribution for the two groups differ
- Sum up the ranks in the two groups
  - The specific form of the test statistic isn't important (for this course)
  - We can find a p-value
    - Tells us the probability of observing sum of ranks as extreme or more extreme than that observed if the distribution for the two groups are identical
- This is called the Mann-Whitney U test
  - It has many other names, such as the Mann–Whitney–Wilcoxon test

In R

- The test can be performed using the `wilcox.test` function in R
- Like when using `t.test` we separate data into two groups

```
rt = subset(hawk, species == "RT") # same function as we used for t.test
ss = subset(hawk, species == "SS")
wilcox.test(rt$hallux, ss$hallux)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  rt$hallux and ss$hallux
## W = 2275, p-value <2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Interpretation

- As usual, the p-value is quantifying the incompatibility between the null hypothesis and the data
- Since the p-value $< \alpha = 0.05$, the data are unusual if the null hypothesis were true
  - The data we have observed would be unusual if there were the distribution of hallux length were the same for the two species

Parametric vs non-parametric

- Most of the models and methods we have seen so far are referred to as parametric
  - Specify the distribution of the observations: normal, binomial, etc
  - These models are defined in terms of parameters: μ, p, etc
  - We find confidence intervals for the parameters
  - We specify hypothesis tests about the parameters
- With non-parametric models, we make fewer assumptions
  - We assume the observations come from an unknown distribution
    - There are not specific parameters as above (hence non-parametric)
  - We can specify hypothesis tests
  - Confidence intervals are more challenging
- A common misconception is that non-parametric approaches make no assumptions

Non-parametric approaches

- The principle of converting data to ranks can also be used for other cases we have considered
- Single sample (or paired data) $\rightarrow$ Wilcoxon signed-rank test
- Two samples (independent groups) $\rightarrow$ Mann-Whitney test
- ANOVA (multiple independent groups) $\rightarrow$ Kruskal-Wallis test
- Remembering the names isn't important
- The concepts are more important: converting data to ranks
  - Note: not all non-parametric approaches use ranks
- We won't look at any details regarding the methods in blue above
  - It is worth knowing that the approaches exist

In R

- Seen `wilcox.test`
  - Used for single sample or paired data
  - Can be used for two independent groups
- The function `kruskal.test` can be used for:
  - Multiple independent groups
  - Can be used for two independent groups
- When using `kruskal.test` we need to use formula: as in `lm` or `aov`

```
kruskal.test(hallux ~ species, data = hawk)

##
##  Kruskal-Wallis rank sum test
##
## data:  hallux by species
## Kruskal-Wallis chi-squared = 68, df = 1, p-value <2e-16
```

Comparing the two tests

- If we have two independent groups, we have a choice
  - Use `wilcox.test`
  - Use `kruskal.test`
- These have different test statistics
  - Give the same p-values
- Note: `wilcox.test` includes a continuity correction when calculating the p-value
  - The two approaches give the same p-value when this is turned off with `correct = FALSE`

Data: hawk tail measurements

- Look at data from 43 red-tailed hawks
- Data comparing two tail measurements
  - `tail_std`: Standard approach for measuring the tail length (mm)
  - `tail`: Approach invented by those involved in collecting data (mm)
- Import and view the data

```
hawk_tail = read.csv('hawk_tail.csv')
```

```
head(hawk_tail)
##   tail tail_std
## 1  222      229
## 2  215      217
## 3  235      236
## 4  215      215
## 5  212      221
## 6  206      217
```

Hawk tail measurements: correlation

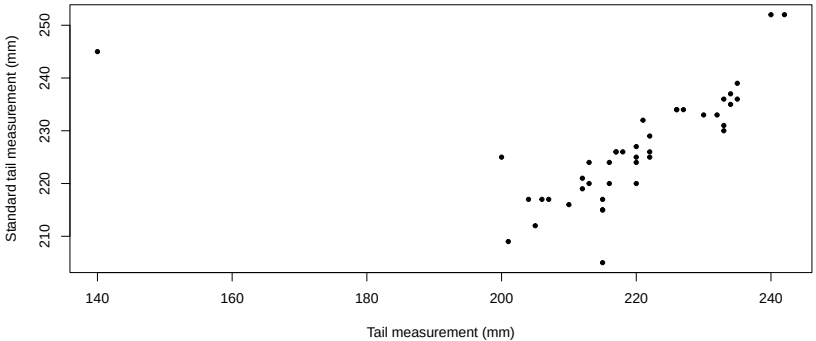
- How 'consistent' are the two measurement approaches?
  - We could assess with correlation

```
cor(hawk_tail$tail, hawk_tail$tail_std)
```

```
## [1] 0.326
```

- That does not seem very high
- Look at the data to see what may be going on

Hawk tail measurements



- Seems like a reasonably strong linear relationship
  - With a large outlier

Back to ranks

- We can again work with ranks
  - Rank x (new tail measurements)
  - Rank y (standard tail measurements)
- Find the correlation of the ranks

```
hawk_tail$rank_tail = rank(hawk_tail$tail)
hawk_tail$rank_std = rank(hawk_tail$tail_std)
cor(hawk_tail$rank_tail, hawk_tail$rank_std)
## [1] 0.777
```

Correlation (sorry more names!)

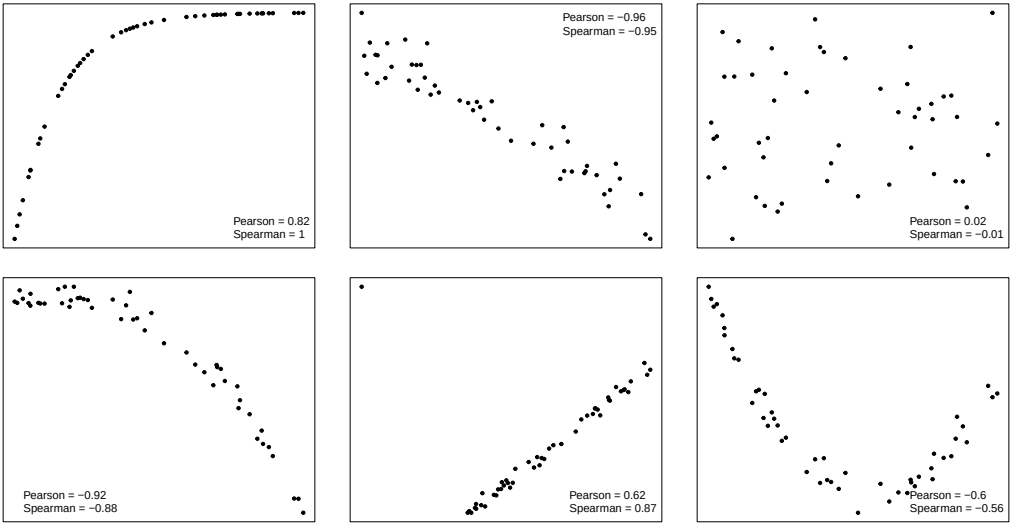
- The correlation based on ranks: Spearman correlation
- The correlation based on data: Pearson correlation
  - What we looked at when we covered linear regression
- Need not calculate the ranks in R to calculate Spearman correlation:

```
cor(hawk_tail$tail, hawk_tail$tail_std, method = "pearson") # this is the default
## [1] 0.326
cor(hawk_tail$tail, hawk_tail$tail_std, method = "spearman")
## [1] 0.777
```

Spearman correlation

- Spearman correlation measures the strength of an increasing or decreasing relationship
  - It need not be a linear relationship
- Spearman correlation is robust to outliers: using ranks
  - Spearman correlation an alternative to throwing away outliers without justification
- Spearman correlation can be used with ordinal data
  - Categorical data where the values have an order
  - e.g. survey response: 'Excellent', 'Good', 'OK', 'Poor', 'Terrible'
- Spearman and Pearson correlation are often similar
  - Relationship is approximately linear
  - Minimal effect of outliers
- Look at some examples

Examples


Pearson = 0.82
Spearman = 1
Pearson = −0.96
Spearman = −0.95
Pearson = 0.02
Spearman = −0.01
Pearson = −0.92
Spearman = −0.88
Pearson = 0.62
Spearman = 0.87
Pearson = −0.6
Spearman = −0.56

Big picture

- Seen an introduction to non-parametric methods
- Focused on conceptual understanding
  - Assuming the data come from an unknown distribution
  - For the methods we have seen: working with ranks
  - Skipped over the details
- Advantages of parametric models
  - More powerful when assumptions hold
  - Interpret parameter (estimates)
  - Straightforward confidence intervals
- Advantages of non-parametric methods
  - Fewer assumptions
  - More robust to outliers and skewed data

Summary

- Looked at non-parameteric approaches
- Work with ranks
- Two independent groups
  - Mann-Whitney
- Correlation
  - Spearman correlation
- Outlined other approaches
  - Wilcoxon rank-sum (one sample / paired data)
  - Kruskal-Wallis (multiple independent groups)