PS303: Week 4

pp.327-348

Hypothesis tests

- Recurring patterns allow us to predict future occurrences with high likelihood. In the frequentist approach, pattern outcomes are binary. Some patterns are obvious…
  - The sun rises in the east
  - Apples fall from trees
  - Asthmatic children wheeze
- Others may be less so…
  - Engaging extensively with social media facilitates depression
  - Focusing on identity can devalue individual merit
  - Increasing minimum wage correlates with job losses across the underprivileged

  Statements about how the world operates can be formulated into Research Hypotheses *(RH)*. Briefly, a *RH* involves claims about models (in nature) whereas a Statistical Hypothesis *(SH)* makes claims about data. The former are scientific claims, the latter are claims about the data. As a researcher, your task wil be to derive RH's and test them using statistical hypotheses.

Ideally, we should construct RHs to be *relational* in order to render them testable (e.g., given x, then y; if no x, then no y). The SH reformulates this question in a manner amenable to statistical testing (e.g., is the observed frequency of y in the presence of x more than what would be expected by chance?)

Understanding *p*-values

- Statistical tests deal with mathematical relationships between parameters of interest (e.g., comparing between sample estimates to identify whether one group is *significantly* different from another)/
- Null hypothesis significance testing (NHST) involves testing whether observed data vary *significantly* from hypothetical estimates
- Conventional NHSTs do not 'prove' any particular research claim - rather, NHST tells us whether we can retain/reject a null hypothesis (H_O : There is no effect).
- Consider the following examples of a RH and an associated SH:
  - Using social media reduces negative well-being
  - $\mu_1 - \mu_2 = 0$ (null hypothesis)

- To test this claim, we can imagine recruiting two groups of participants with varying patterns of social media usage (let's classify them as *Users* and *Non-Users*) and ask them to complete a well-being survey.

Hypotheses	Examples
Research Hypothesis	Extended usage of social media negatively impacts well-being
$H_o : \mu_1 - \mu_2 = 0$	Mean well-being of media users = non-media users
$H_1 : \mu_1 - \mu_2 > 0$	Depression of media users $\neq$ Depression of non-media users

Frequentist tests assess how viable the null (and fictional) H_o is in explaining the observed data. If the data is considered 'too extreme' under the assumption that the null is true, then it is deemed **statistically significant**. The formal threshold of significance (at least in much of Psychology) is ($p < .05$), which can be interpreted as *a less than 5% chance of observing the present data assuming the null hypothesis is true* (implying) *we can reject the null hypothesis* (but not make any claims about the alternative hypothesis just yet!).

Controlling for Type-1 and Type-2 error

- Your goal as a researcher and analyst is to reduce the likelihood of making incorrect decisions. The p-value helps in identifying whether an effect is 'really' there (as opposed to being a false positive or false negative) but it is not a complete solution. A p-value threshold that is too strict (e.g., p <.001) can reduce Type-1 error while inflating Type-2 error. Alternatively, a value that is too lenient (e.g., p <.1) can reduce Type-2 error while inflating Type-1 error.

- Understanding the two types of error:

  - Avoiding **false positives**: Rejecting your H_o when the data actually supports the null, also called **Type-1 error** (e.g. accepting a medical treatment as valid despite the treatment having no effect). Example of Type-1 error: homeopathy effectiveness?.

  - Accepting **false negatives**: Accepting the null hypothesis when the data actually supports H_o's rejection (e.g., rejecting a medical treatment as invalid despite the treatment having an effect). Example of Type-2 error: HCQ ineffectiveness?.

	Retain H_o	Reject H_o
H_o is true	Correct decision	**Type 1 Error**
H_o is false	**Type 2 Error**	Correct decision

Assuming H_o : $Depression_{Media} = Depression_{NonMedia}$, what are some possible Type-1 or Type-2 errors that may be confounding the question? Have a look at this short video for further discussion of this topic.

Non-statistical strategies for dealing with error

Imagine yourself as a judge in 10th century England who has to identify whether Johnny stole a horse. Suppose you have no witnesses to call on - what would be a strategy to discover whether Johnny is guilty?

You may falsely convict someone innocent (Type-1 error), or you may incorrectly exonerate someone guilty (Type-2 error). To what degree are you willing to accept *freeing the guilty* (Type-1 error) versus *punishing the innocent* (Type-2 error)?

	Let Johnny go	Johnny undergoes the trial
Johnny is innocent	Correct decision	**Type 1 Error**
Johnny is guilty	**Type 2 Error**	Correct decision

While a bit extreme, 'trials-by-ordeal' may have been somewhat effective in controlling for Type-1 and Type-2 error!

Significance and Power

Since it is nearly impossible to reduce error rates to 0, what are some *acceptable* ranges for Type-1/2 error?

- In psychological research, we are typically satisfied with a 5% Type-1 error rate. We are willing to accept that 5% of observed data may incorrectly signify an effect when there may be none in reality. Technically, we can declare prior to our study that our α error rate is constrained at 5%. In other words, $\alpha = .05$ is our **threshold** for statistically retaining/rejecting the null.

- A low Type-2 error rate (β) is also desirable. For example, we may be willing to accept a 10% ($\beta = .1$) false negative rate. The parameter of interest here is **power** ($1 - \beta$), which describes the probability that a null hypothesis is correctly rejected when an alternative is true.

- For example, we can claim that our test has 90% power ($1 - \beta = .9$) and a 5% α error rate, implying our test has a 5% chance of incorrectly rejecting the null and a 10% chance of incorrectly accepting the null.

	Retain H_o	Reject H_o
H_o is true	$1-\alpha$: Likelihood of correct H_o retention	α: Type-1 error
H_o is false	β: Type-2 error	$1-\beta$: **Power** of selected test

Keeping your β's small (about .1) and your α's even smaller (about .05) facilitates reasonably robust inferences without unreal sampling demands.

Why not set up a study with $\alpha = .001$ and $\beta = .99$ to further minimize the plausibility of Type-1 *and* Type-2 error? Because reducing error probability is inversely related to sampling size. The smaller the error constraint, the larger the sample required (at times reaching in the thousands). However, increasing sensitivity (through increasing sample size) can enable detection of (even) *small* effects as statistically significant, but with low practical significance (e.g., differences in kokonda consumption between Suva and Lautoka residents). The decision to constrain error has to be carefully considered prior to running your study (though $\alpha = .05|\beta = .90$ appears to be a general heuristic).

Interpreting outcomes from binomial experiments

Because binomial tests are generally the 'simplest' hypothesis test, we can explore some hypotheses in relation to binary coin flips…

You have been asked to test the fairness of a coin before the start of a rugby season. You remember from your statistics class that the 'true population probability' (of seeing Heads) can be determined from flipping the coin 1000's of times (notwithstanding the wear-and-tear). However, as this may not be practical, we can a `binomial significance test` to determine whether $P(\text{Heads})$, or the probability of 'successes', significantly varies from a null hypothesis.

A fair coin may be one where we expect that heads/tails will show up with roughly equal probabilities after enough flips. Formally, our hypothesis is that the probability of viewing heads is 50%, or $P(H) = .5$. Suppose we flip the coin 50 times and record 38 heads. Can we make a decision?

- Assuming probability of heads/successes is represented by θ, we can formally describe our hypotheses as follows:
  - $H_0 : \theta = .5$
  - $H_A : \theta \neq .5$ (two-sided)

We can use *R*'s built-in function for running a **two-sided** binomial test (since we do not care at the moment whether the bias is towards Heads *or* Tails).

```
binom.test(x=38,    # Assign the number of 'successful' trials to 'x'
           n=50,    # Assign the number of 'total' trials to 'x'
           p=.5,    # Probability of success (null hypothesis)
           alternative = "two.sided",    # Two-sided or one-sided test?
           conf.level = .95)             # 95% Confidence Interval
```

```
##
##  Exact binomial test
##
```

```
## data:  38 and 50
## number of successes = 38, number of trials = 50, p-value = 0.0003059
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.6183093 0.8693901
## sample estimates:
## probability of success
##                   0.76
```

> A two-sided binomial test indicated that the number of heads observed $N_{Heads} = 38$ relative to the total number of flips ($N_{Flips} = 50$) was statistically different from a chance estimate, $\theta = .76 (CI_{95} : 0.62_{to} 0.87), p = .0003$. We can *reject* the null hypothesis which claimed the coin was fair.

Two-sided vs one-sided tests

- Our earlier test asked whether θ (probability of success) was *different* from a chance estimate. We ran a *two-sided* test as we did not care about the direction of the difference.
- Alternatively, we could have run a *one-sided* test if we had prior reason to do so. This could involve asking whether $P(H)$ is *greater* than would be expected by chance ($\theta > .5$). We could also inquire whether $P(H)$ may be *less* than would be expected by chance ($\theta < .5$).

```r
test1 <- binom.test(x=38,n=50,p=.5,conf.level = .95,alternative = "greater")
test2 <- binom.test(x=38,n=50,p=.5,conf.level = .95,alternative = "less")
```

After running the tests and assigning the outputs to variables, we can call those directly

```r
test1
```

```
##
##  Exact binomial test
##
## data:  38 and 50
## number of successes = 38, number of trials = 50, p-value = 0.0001529
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.6403443 1.0000000
## sample estimates:
## probability of success
##                   0.76
```

- Our first test (where $H_A : \theta > .5$ and $H_O : \theta \leq .5$) produces a p-value of **0.000152932**, which we can summarily present as **<.001**. The results suggest we can **reject** the one-sided null hypothesis that the observed distribution is less than would be expected by chance (because $p < .05$).

```r
test2
```

```
##
##  Exact binomial test
##
## data:  38 and 50
## number of successes = 38, number of trials = 50, p-value = 1
## alternative hypothesis: true probability of success is less than 0.5
## 95 percent confidence interval:
##  0.0000000 0.8552818
## sample estimates:
## probability of success
##                   0.76
```

- Our second test (where $H_A : \theta < .5$ and $H_O : \theta \geq .5$) produces a p-value greater than .05, implying we should **retain** the null hypothesis that the observed distribution is greater than would be expected by chance (because $p > .05$).

Limitations of *p*-values

While p values are useful for retaining/rejecting null hypotheses, there cannot resolve research hypotheses. This is because (i) p values do not tell us whether a statistical effect is *practically* significant, and (ii) a statistically significant difference does not specify the precise population parameter with any detail (how much of a difference can we expect in the actual population?). Recall that our goal was to determine whether the coin is fair. Although we found *statistically significant* differences based on a $\alpha = .05$ threshold, were the differences large enough to be *practically* meaningful?

NHSTs report p-values ranging from $.001 \leq p \leq .999$, which tell us how likely the observed pattern would be assuming H_O is true. In psychological research, observing $p \leq .05$ implies the observed pattern would be 'extremely unlikely' (less than 5% of the time) if H_O was true. Alternatively, observing $p > .05$ is typically interpreted as insufficient evidence for rejecting the null. Critically, *neither* claim tells us about the likelihood of the alternative hypothesis.

Later in the course, we will introduce **effect sizes** as estimators of *practical* significance. Along with **confidence intervals** and **standard error** as range estimates (identifying 'where' the true population parameter resides), effect sizes will help illuminate whether there are 'real' differences between/within populations of interest.

Closing statements

- Frequentist tests are typically run with *sample* parameters, meaning our outcomes typically explain *sample* behaviors. However, we are typically interested in finding out about *population* parameters, which sample parameters may not necessarily correspond to.

- A statistically significant effect (*p<.05*) permits rejection of the null hypothesis, *not* acceptance of the alternative hypothesis.

- Statistical effects do not tell us about the practical meaningfulness of an effect, which (say) effect sizes can. For example, both $\theta_1 = .53$ and $\theta_2 = .76$ may be 'significantly' different from a null $\theta_o = .5$ estimate, but the latter estimate corresponds with a larger effect.

- Statements of significance are binary - either an effect is significant or is not according to some pre-established threshold (e.g., $p < .05$)

Statistical significance?	Big Effect	Small Effect
$p < .05$	A real difference that is important	A real difference that might not be important
$p > .05$	No effect	No effect

Running binomial tests manually

When you are interested in a binary outcome parameter *and* have an expectation of future probability of an outcome (e.g., $P(H) = .5$ for a fair coin), a binomial test checks whether the observed pattern varies (significantly) from the expected pattern. Binomial tests work with *binary* distributions, which cannot be normal.

Imagine we flip a coin 10 times ($n = 10$) and view 6 heads ($x = 6$). This may follow the sequence: $H - T - T - T - H - H - H - H - T - H$. We want to test is whether our observed distribution of heads $(\frac{x=6}{n=10})$ varies from our expected distribution ($P(H)_{Expected} = .5$). So we want to know if

$$H_O : P(Expected) = P(Observed)$$

.

The parameters necessary for estimating binomial probability distributions are:
n - Total number of independent trials ($n = 10$)
x - Total successes ($x = 6$)
p - Expected probability of x successes ($p = .5$)

We may then estimate the probability of viewing x successes across n trials given an expected probability p.

$$P(Heads) = \frac{n!}{x!(n-x)!} \times p^x \times (1-p)^{n-x}$$

Let's enter the parameters provided.

$$P(H) = \frac{10!}{6!(10-6)!} \times .5^6 \times (1-.5)^{10-6}$$

which gives us

$$P(H) = \frac{10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1}{6 \times 5 \times 4 \times 3 \times 2 \times 1(4 \times 3 \times 2 \times 1)} \times .015625 \times .0625 = .205$$

We can check our results by generating a probability densitity distribution in R...

```
dbinom(x=6,size=10,prob=.5) # Probability DENSITY (d) of binomial distribution
```

```
## [1] 0.2050781
```

Suppose we want to know the probability of 6 or fewer Heads...

```
pbinom(6,10,.5) # Cumulative PROBABILITY (p) of binomial distribution
```

```
## [1] 0.828125
```

What about the probability of 6 or more Heads?

```
1-(pbinom(5,10,.5)) # Remember, all P's add up to 1
```

```
## [1] 0.3769531
```

1. The probability of getting 6 heads exactly; $P(H = \frac{6}{10}) = .205$.

2. The probability of getting 6 or **fewer** heads exactly; $P(H \leq \frac{6}{10}) = .828$.
3. The probability of getting 6 or **more** heads exactly; $P(H \geq \frac{6}{10}) = .377$.

Lab activity

We have discussed how to run and report binomial tests, which investigates whether observed distributions from a binomial category (e.g., **Heads vs Tails; Yes vs No; Big vs Small**) matches some pre-defined probability distribution. Although we constrained our discussion to coin flips, binomial proportion tests can theoretically apply to *any* binary category. For this week's lab activity, you will run **three** binomial tests then report your findings in an appropriate format. For each test, report the observed proportion estimate (θ), the Type-1 error rate (p-value) and the 95% confidence interval. Also report whether your test was one-sided or two-sided, and whether your findings retain/reject a null hypothesis. **Include the code used during binomial tests**

Q1. You flip a coin 89 times and view 46 heads. You suspect that the coin may be slightly biased towards heads. To test your claim, you run a binomial test of proportions against a null hypothesis of $P(H) = .50$, meaning you are assuming the true rate of the coin would be 5 heads for every 10 flips. Does the *observed* proportion of heads significantly vary from the null?

Q2/Q3. An immigration officer has been accused of bias *against* immigrants from Tonga, while being *in favor* of immigrants from Vanuatu. Your task as an enforcement officer is to check whether these claims can be statistically supported. You are provided the following data:

Origin	Total immigrants	Accepted Immigrants	Accusation	Alternative Hypotheses
Tonga	400	161	Bias *against* Tonga	$H_A : \theta \leq .5$ |
Vanuatu	160	91	Bias *towards* Vanuatu	$H_A : \theta \geq .5$ |

Report whether each accusation is statistically supported or not. *Show your work.*