SM-1402 Basic Statistics

Chapter 4: Analysis of qualitative data *[handout version]*

Dr. Haziq Jamil
Mathematical Sciences, Faculty of Science, UBD
https://haziqj.ml

Semester II 2022/23

Learning outcomes

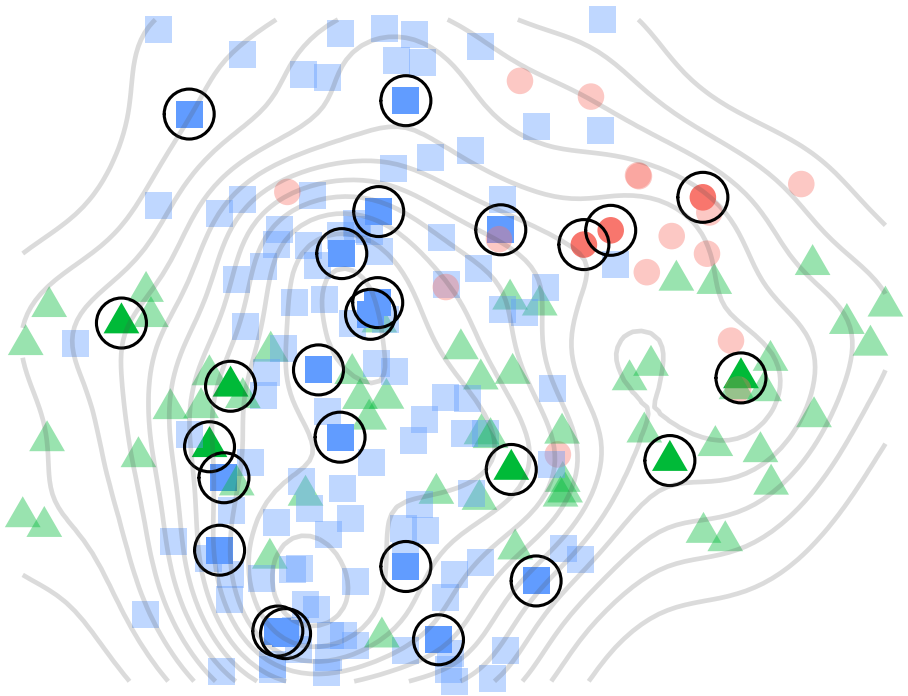
- Be familiar with the binomial distribution and its relationship to the normal distribution when n is large.
- Calculate sample proportions and construct a confidence interval to quantify the statistical uncertainty.
- Perform a statistical test for testing hypothesis regarding proportions.
- Perform a Chi-squared test of independence for frequency tables.

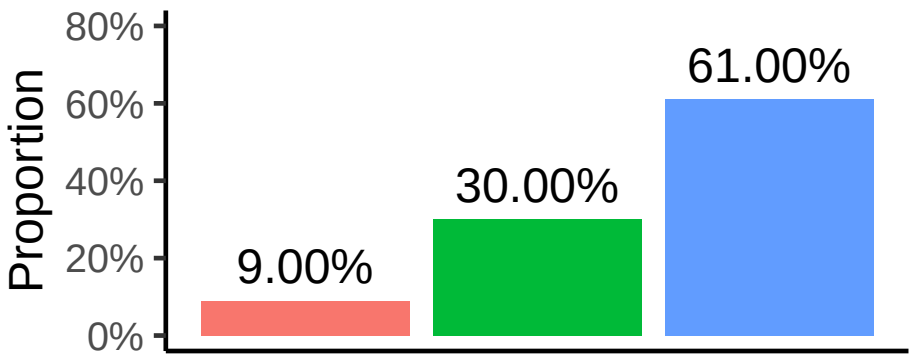
Required reading

- Madsen (2016) Chapter 5.

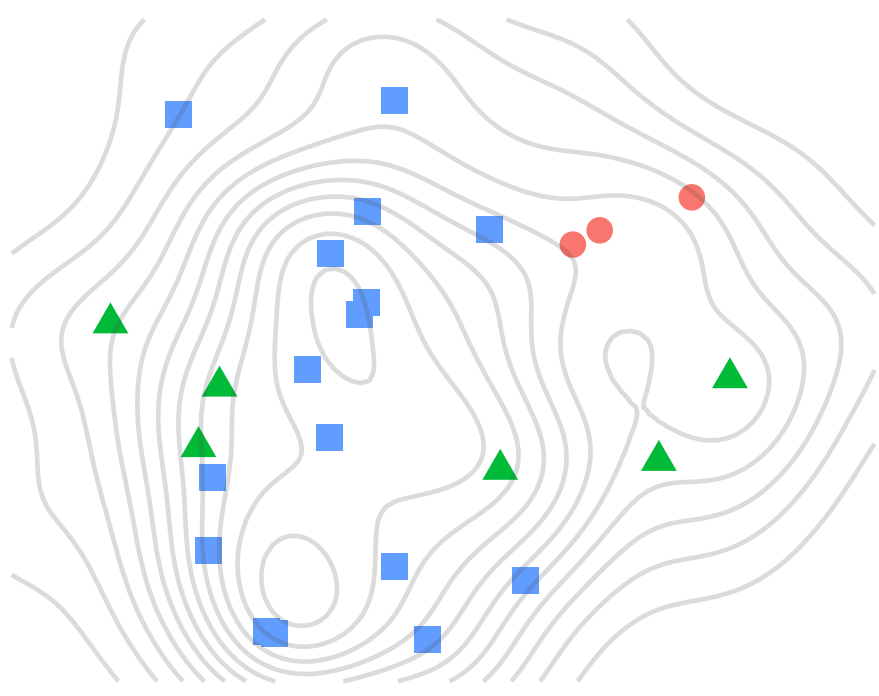
Motivation

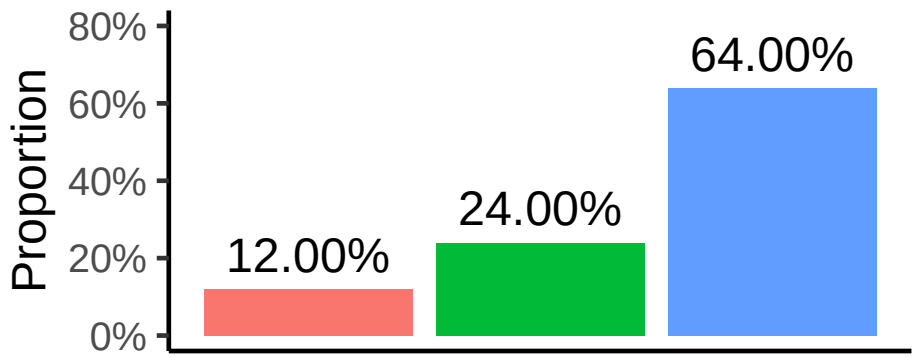
Population





Sample





Introduction

In this chapter we look at *qualitative* data. They can be

1. Binary
   - Questionnaire survey: "Yes" / "No" type questions.
   - Games: "Heads" / "Tails"
2. Categorical (more than 2 categories)
   - Likert scale responses: "S. Agree" / "Agree" / "Neutral" / "Disagree" / "S. Disagree"
   - Faculties in UBD: "FOS" / "FIT" / "FASS" / "IHS" / "Others".

Remark

If we have more than 2 categories, we can always "collapse" the data into binary. As an example, the 5-point scale can be converted to binary ("Agree" vs "Disagree").

Binomial distribution

The binomial distribution describes the distribution of the number of "successes" in n independent and identical binary *trials*. That is, suppose we have a situation such that

i. A finite number of trials n are carried out.
ii. Each trial is independent of each other.
iii. The outcome of each trial is either "success" or "failure".
iv. The probability success $p \in (0, 1)$ is the same for each trial.

Definition 1 (Binomial distribution)

Let X be the number of success outcomes in n trials. Then X has a binomial distribution, written $X \sim \text{Bin}(n, p)$. The probability that exactly $X = x$ success are obtained is

$$\Pr(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Example

Example 2

At a supermarket, 60% of customers pay by credit card. Find the probability that in a randomly selected sample of ten customers,

(a) exactly two pay by credit card.
(b) more than seven pay by credit card.

Let X be the number of customers in sample of ten who pay by credit card. Consider 'paying by credit card' as a *success*, so $p = 0.6$. Assume also independence of $n = 10$ trials. Then $X \sim \text{Bin}(10, 0.6)$.

(a) $\Pr(X = 2) = {}^{10}C_2(0.6^2)(0.4^8) = 0.011$.

(b)

$$\begin{aligned}\Pr(X > 7) &= \Pr(X = 8) + \Pr(X = 9) + \Pr(X = 10) \\ &= {}^{10}C_8(0.6^8)(0.4^2) + {}^{10}C_9(0.6^9)(0.4^1) + {}^{10}C_{10}(0.6^{10})(0.4^0) \\ &= 0.17\end{aligned}$$

Mean and variance of binomial

Proposition 3

Let $X \sim \text{Bin}(n, p)$. Then

- Mean $= np$ $[\mu]$
- Variance $= np(1-p)$ $[\sigma^2]$
- SD $= \sqrt{np(1-p)}$

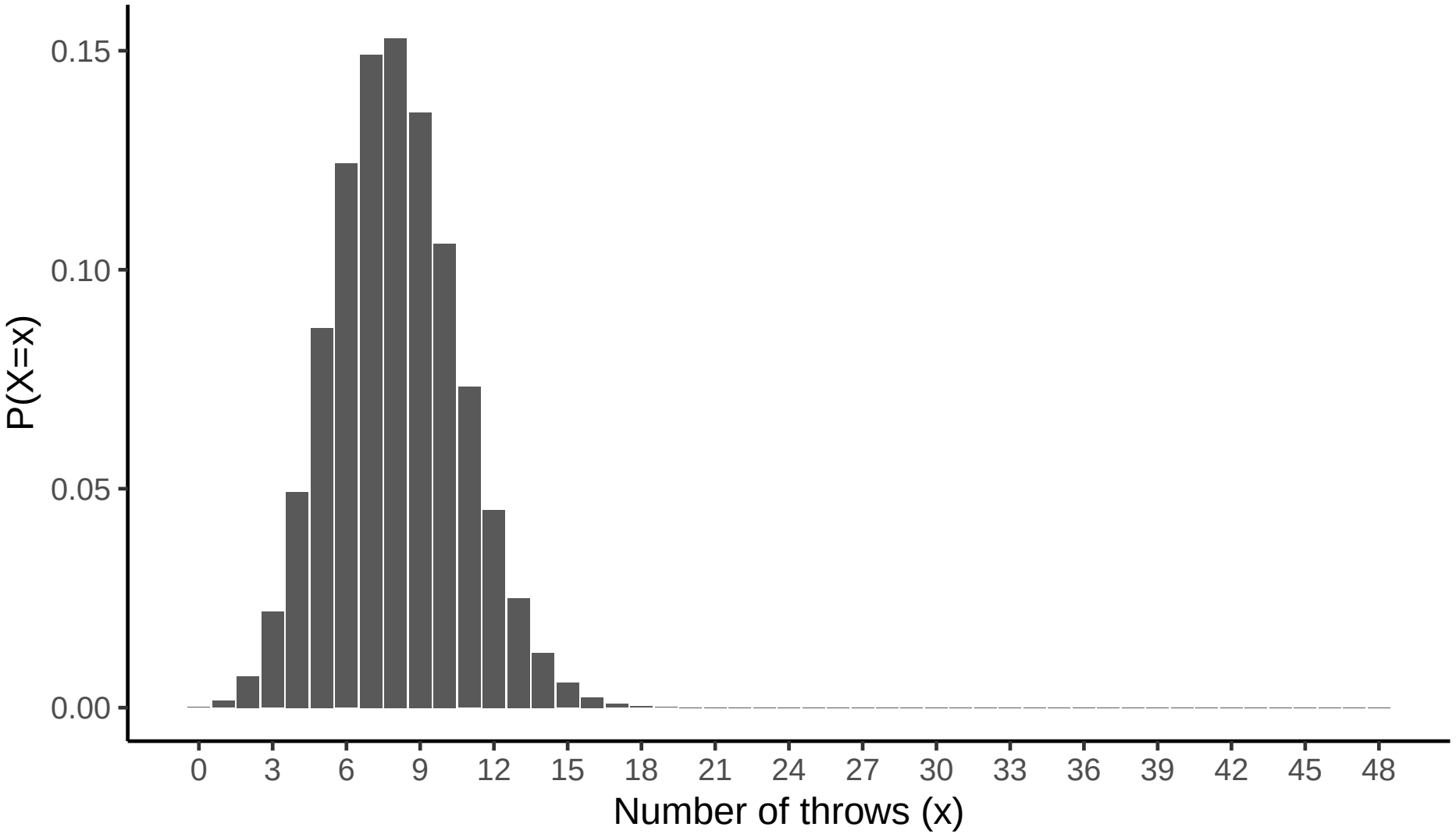
Example 4

Consider throwing a dice 48 times, and consider a "success" to be throwing a six. The total number of successes, X, follows $\text{Bin}(48, 1/6)$. So

- $\mu = 48 \times 1/6 = 8$ throws, while
- $\sigma = \sqrt{48 \times 1/6 \times 7/6} = 2.58$.

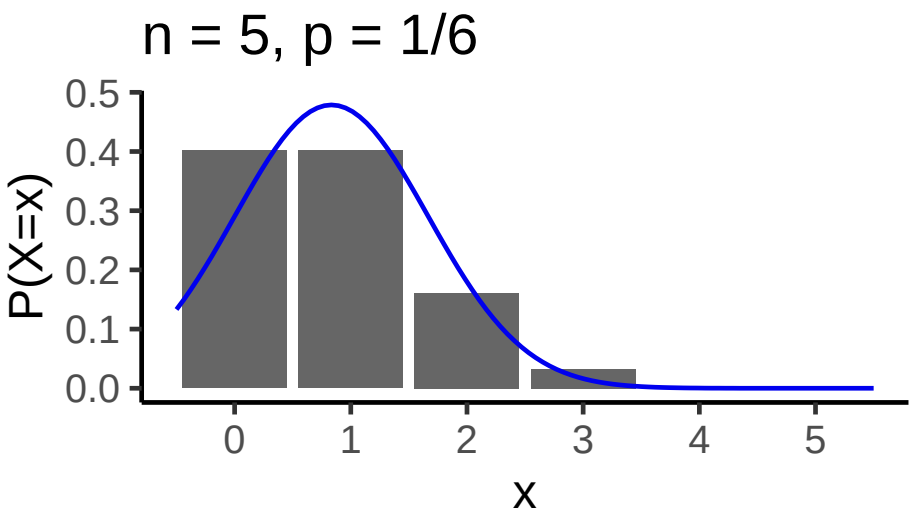
Mean and variance of binomial (cont.)

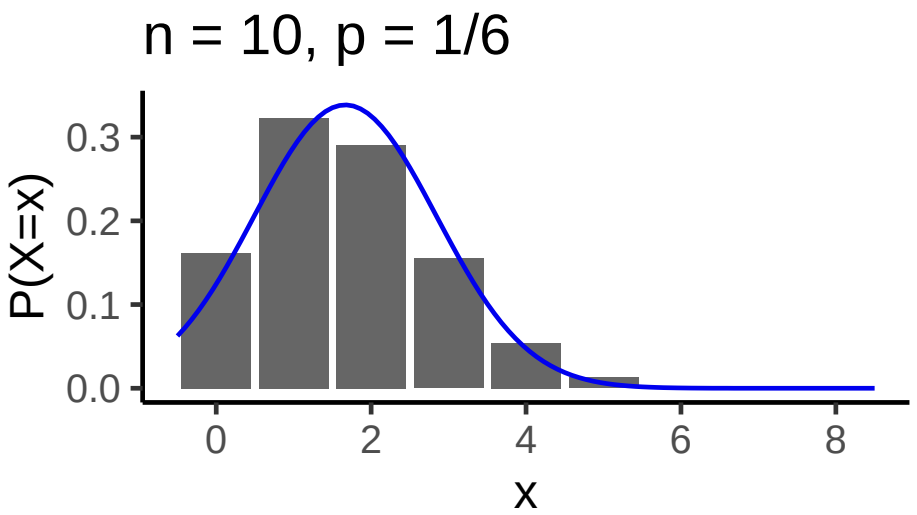
The bar plot of the binomial distribution with $n = 48$ and $p = 1/6$.

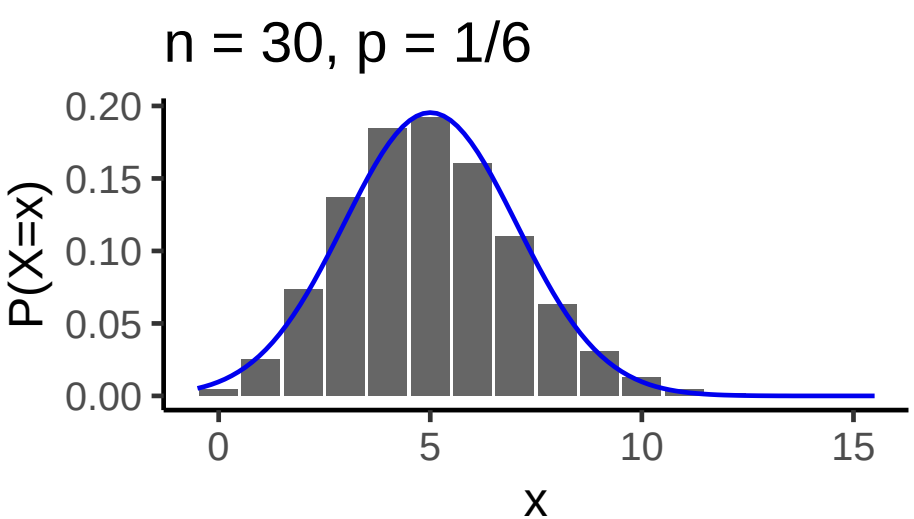


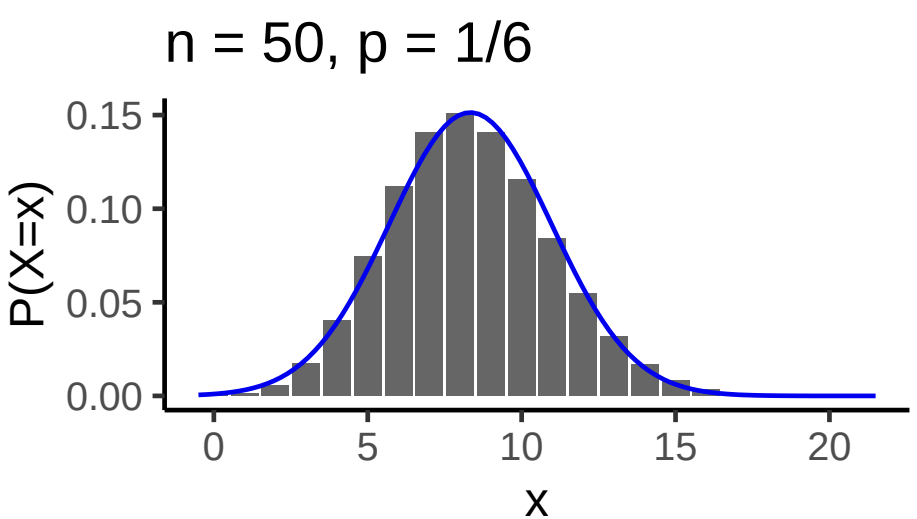
Relationship to the normal distribution

For large n, $X \approx \mathrm{N}(np, np(1-p))$. [Needs $np > 5$ and $n(1-p) > 5$]









Uncertainty in sample surveys

In a population, we may want to find out *proportions* relating to something of interest.

- How many voters for Conservatives in the UK?
- How many support new Mon-Fri working week?
- How many Liverpool FC supporters?

These may be unknown quantities, but can be *estimated* through sample surveys. Because of randomness, there is some uncertainty involved.

Idea

Take a sample of size n. Let X be the quantity (frequency) we are interested in. There is a probability p that we get a "successful" outcome. Then $X \sim \text{Bin}(n, p)$.

Suppose that in our sample we observe $X = x$ "successes". Then an estimate of the unknown p is

$$\hat{p} = \frac{x}{n}.$$

Example

Example 5

In our Fitness Club example, what is the proportion of kids doing strength training? With probability p a randomly selected kid does strength training. Assume that out of $n = 30$ kids in the sample, $x = 12$ do strength training. Then an estimate of p is

$$\hat{p} = \frac{x}{n} = \frac{12}{30} = 0.40 = 40\%.$$

- If the sample is large enough, then we can approximate the binomial distribution with the normal distribution, i.e.

$$\hat{p} = \frac{x}{n} \sim \mathrm{N}(p, p(1-p)/n)$$

- For the example above, the SD of our estimate $\hat{p}$ is

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{30}} = 0.09$$

Constructing intervals for proportions

Definition 6

The 95% confidence interval for proportion estimate is

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} = x/n$.

Continuing our example above, the 95% confidence interval for p is

$$0.4 \pm 1.96 \times 0.09 = 0.40 \pm 0.18 = [0.22, 0.58].$$

Which is not very "confident" at all!

Definition 7 (Representative sample)

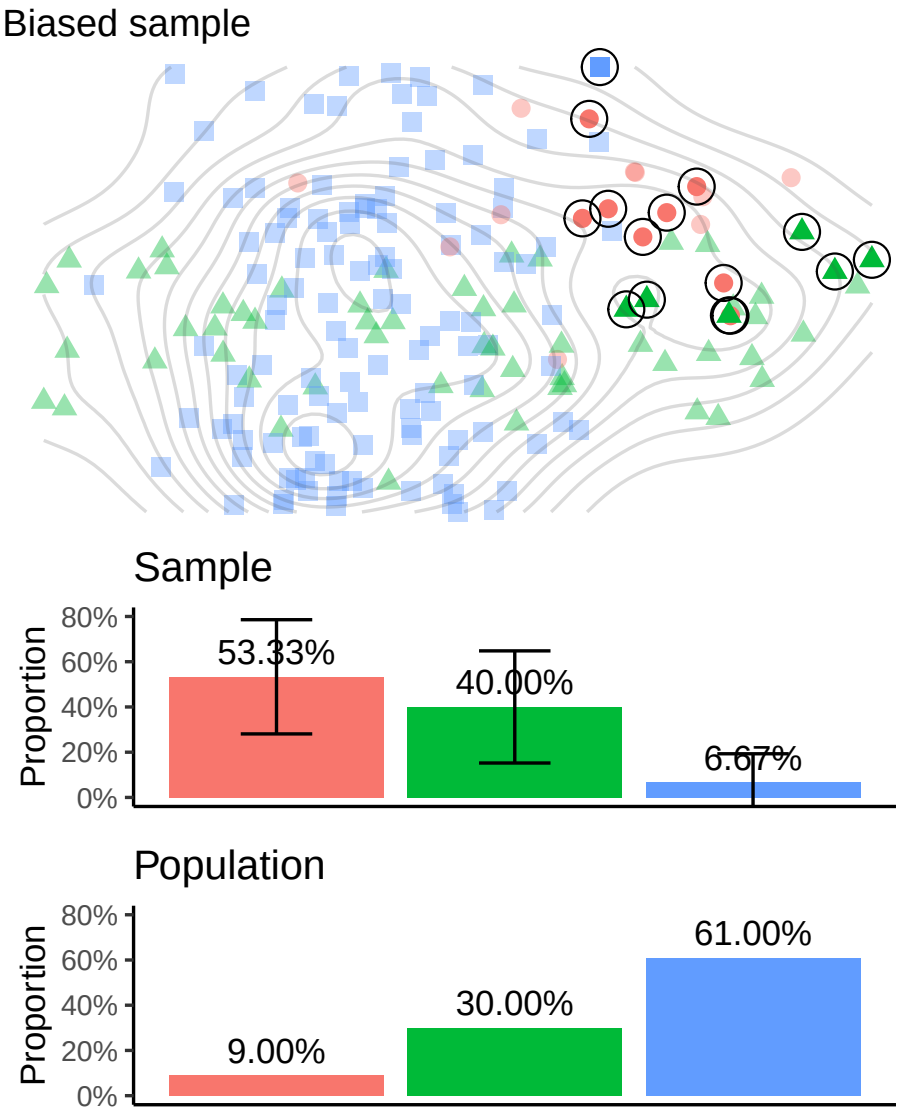
If we knew the true value of p, and it is contained within the 95% interval, we can say that the sample is *representative*.

Representativeness

A sample will always give incomplete information. The key is to ensure the best possible quality of information given the limited resources.

Common sources of **bias**:

1. Subjective choice of sample
2. Self-selection
3. Non-response
4. Convenience sampling
5. Sampling from an incomplete list (sampling frame)



Warning: Biased and non-representative samples give inaccurate results!

Statistical uncertainty

Definition 8 (Statistical uncertainty for p)

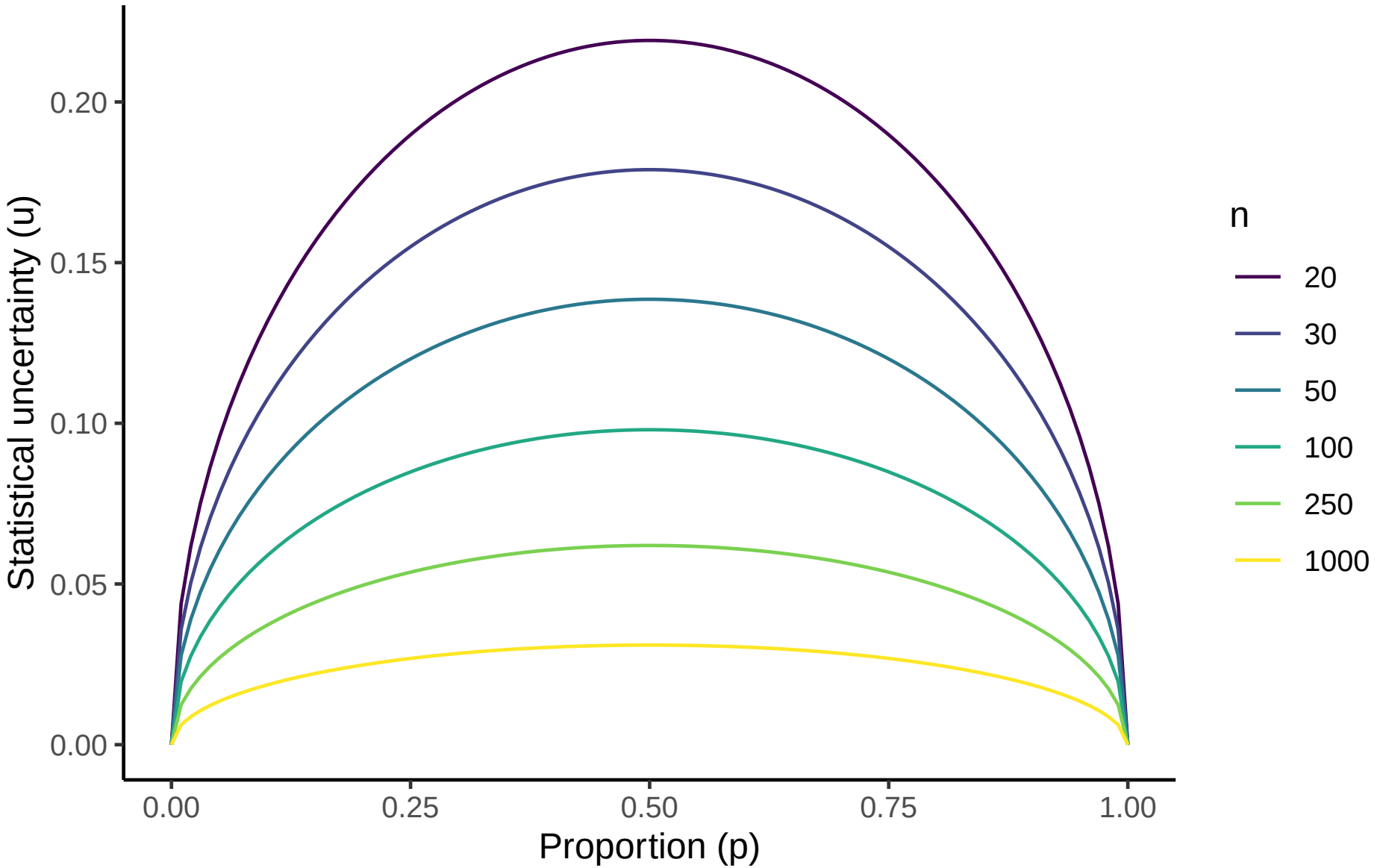
The general formula for the statistical uncertainty u of a relative frequency is

$$u := 1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

Remarks

- Use $\hat{p} = x/n$ instead of p in calculations.
- The above formula is valid when the sample size n is much smaller than the population size. (This has something to do with sampling without replacement)
- The uncertainty is related to
  - The true value of p; and
  - The sample size n.
- The uncertainty is independent of the population size! (Roughly)

Statistical uncertainty (cont.)



Relative statistical uncertainty

While it does seem that uncertainty is small when p is small, we should also consider the uncertainty *relative to* the proportion itself.

Definition 9 (Relative uncertainty)

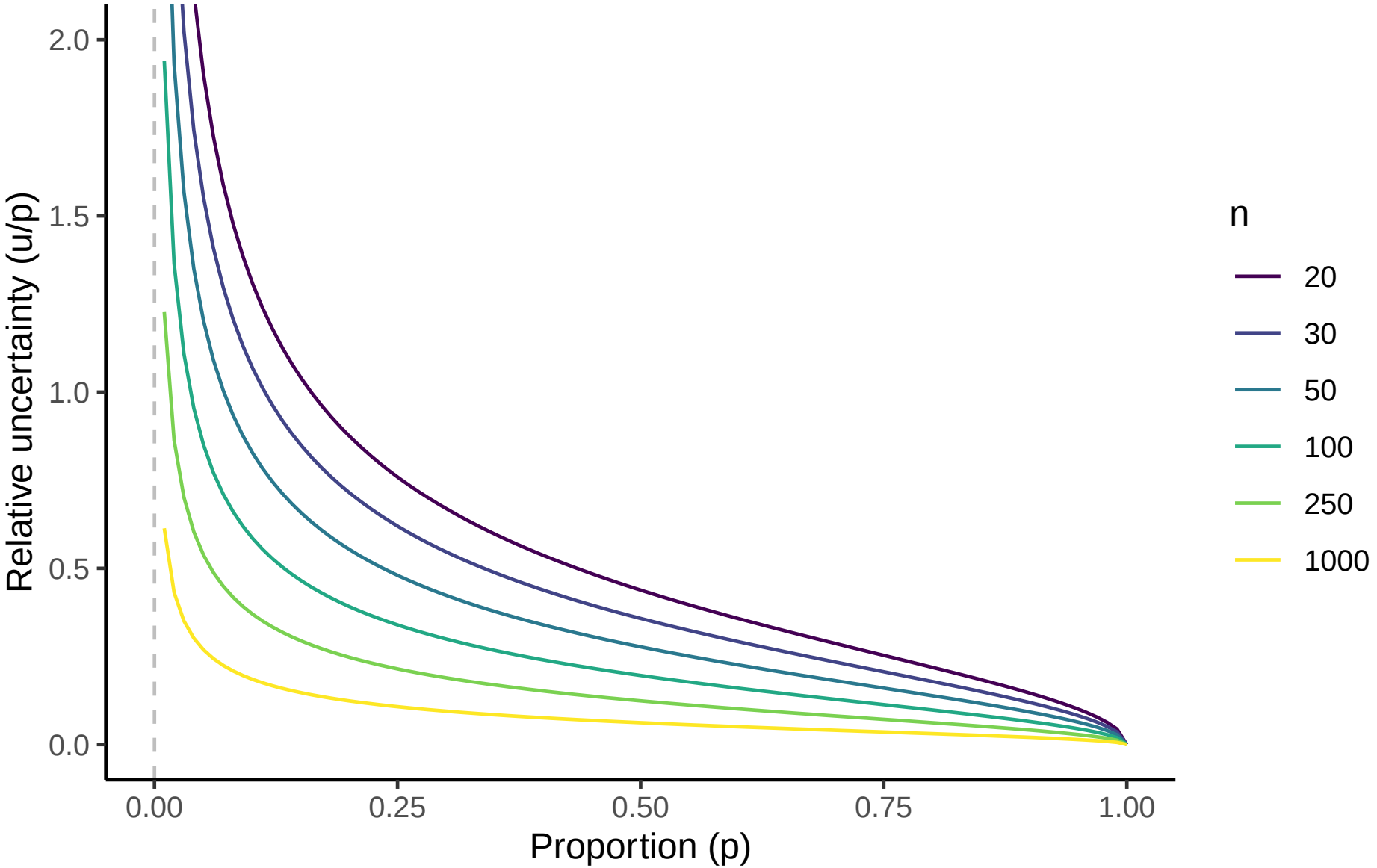
The relative statistical uncertainty $\tilde{u}$ for a relative frequency p is

$$\tilde{u} := \frac{u}{p} = \frac{1.96 \times \sqrt{p(1-p)/n}}{p}$$

Consider a sample survey with some amount of uncertainty u. For argument's sake, let $u = 0.2$

- If the true value of p is large (say $p = 0.9$), then the uncertainty is *relatively small* ($\tilde{u} = 0.2/0.9 = 0.22$);
- If the true value of p is small (say $p = 0.1$), then the uncertainty is *relatively large* ($\tilde{u} = 0.2/0.1 = 2.0$).

Relative statistical uncertainty (cont.)



Statistical tests

Sometimes, you have a hypothesis that you want to confirm or reject. The idea is, you set up an "experiment" and analyse whether or not the data fits your original hypothesis *probabilistically*.

Example 10

I want to check whether a coin is biased–does it come up heads more often than tails? My hypothesis test is

$$H_0 : p = 0.5 \qquad \text{v.s.} \qquad H_1 : p \neq 0.5$$

I then toss the coin $n = 20$ times, and get the following results

T H T H H H H H H T T H H H H H T H H H

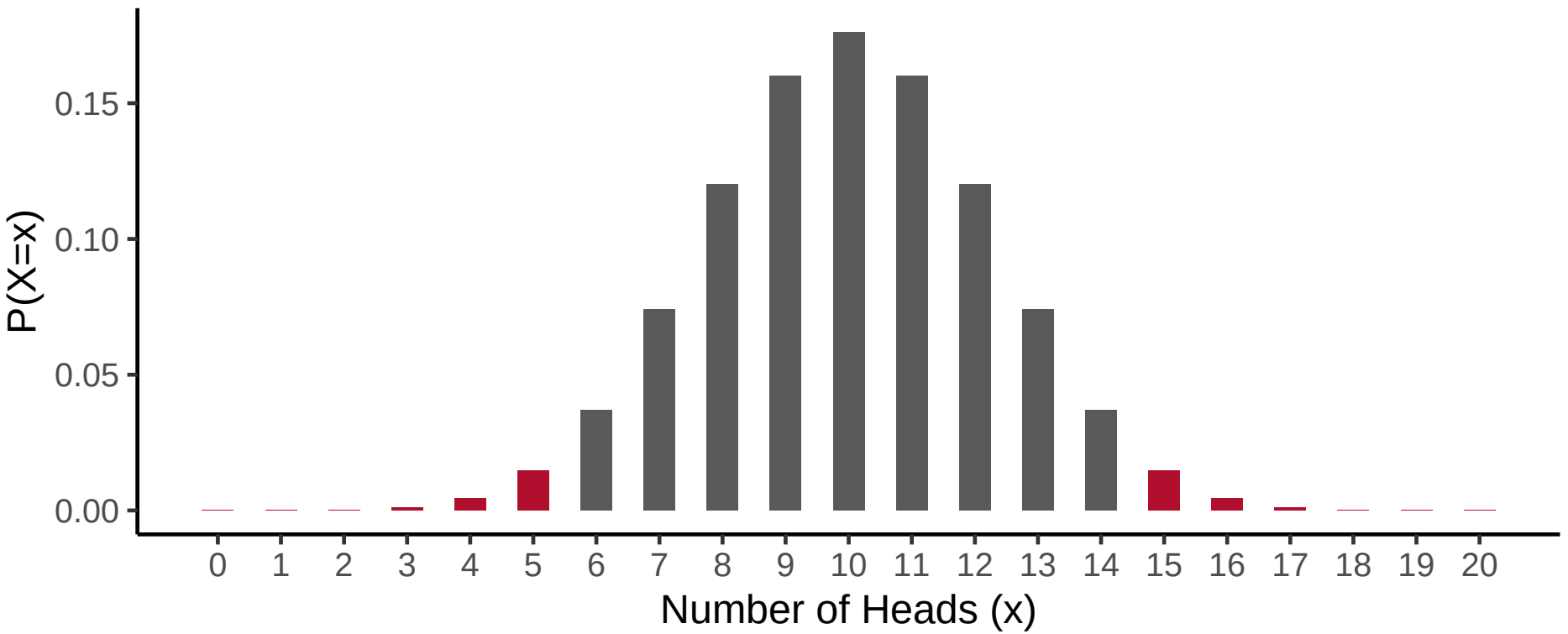
If $p = 0.5$ truly, then we expect 10 H on average. Here, 15 H were observed. What is the probability of observing this event, or other events "just as bad" as this?

p-values

Events that would be "at least as bad" as $x = 15$ if truly $p = 0.5$ in $n = 20$ tosses would be

- $x = 15, 16, 17, 18, 19, 20$; or
- $x = 0, 1, 2, 3, 4, 5$

Adding up the red bars gives us a total probability value of 0.0414.



Concluding tests

If the p-value is small (typically less than 5%), we **reject** the hypothesis. Otherwise we "accept" it[1]. If the p-value is small, then either

- the hypothesis is actually true, but we have observed a rare event (and made an error!); or
- the hypothesis is actually false.

Whatever the case, we can never know the truth, but only a measure of how likely the evidence is in if our hypothesis was true.

Definition 11 (Significance level)

The value of 5% is known as the *significance level*. It is also the probability of making a Type I error (rejecting a true hypothesis aka false positive).

[1] Logically speaking a statistical test does not have the capability to accept a hypothesis. It simply means that there was not enough evidence to reject. Absence of evidence $\neq$ evidence of absence.

Summary

Workflow of a hypothesis test

The objective is to decide whether the hypothesis is supported by data from a sample (or an experiment)

- The hypothesis can either be true or false.
- We consider the hypothesis to be true (default), unless data indicate that it is false.

The practical approach is as follows:

1. Assume hypothesis is true.
2. Calculate the probability of outcomes "at least as rare" as the observed outcome.
3. If this probability is small (typically less than 5%), reject the hypothesis. Otherwise, accept it.

2×2 contingency tables

In the Fitness Club survey, we are still interested in proportion of kids doing strength training, but now we want to group data according to sex.

Table 1: Frequencies and proportions of kids who do srength training

Sex	Does strength training	No strength training	Freq.
Boys	10 (58.8%)	7 (41.2%)	17
Girls	2 (15.4%)	11 (84.6%)	13
Total	12 (40.0%)	18 (60.0%)	30

Question: Is there an actual difference in the proportions between boys and girls? Can we test this?

Chi-square tests

Step 1: Assume hypothesis is true

Our (null) hypothesis is

H_0 : Same proportion between boys and girls doing strength training

which also speaks to *independence between rows and columns*. Now if H_0 was true, we would then **expect** the following frequencies:

- Does strength training (40% of kids)
  - Boys $= 40\% \times 17 = 6.8$
  - Girls $= 40\% \times 13 = 5.2$
- No strength training (60% of kids)
  - Boys $= 60\% \times 17 = 10.2$
  - Girls $= 60\% \times 13 = 7.8$

We now need a measure of how different ("bad") these are in relation to the observed frequencies.

Chi-square tests (cont.)

Step 2: Calculate p-value

Table 3: Expected frequencies of kids who do srength training

Sex	Does strength training	No strength training	Freq.
Boys	6.8 (40.0%)	10.2 (60.0%)	17
Girls	5.2 (40.0%)	7.8 (60.0%)	13
Total	12 (40.0%)	18 (60.0%)	30

Definition 12 (Chi-square test)

The chi-square *test statistic* is defined as

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O = observed frequencies and E = expected frequencies.

Chi-square tests (cont.)

Step 2: Calculate p-value

Observed:

	S	NS
B	10	7
G	2	11

Expected:

	S	NS
B	6.8	10.2
G	5.2	7.8

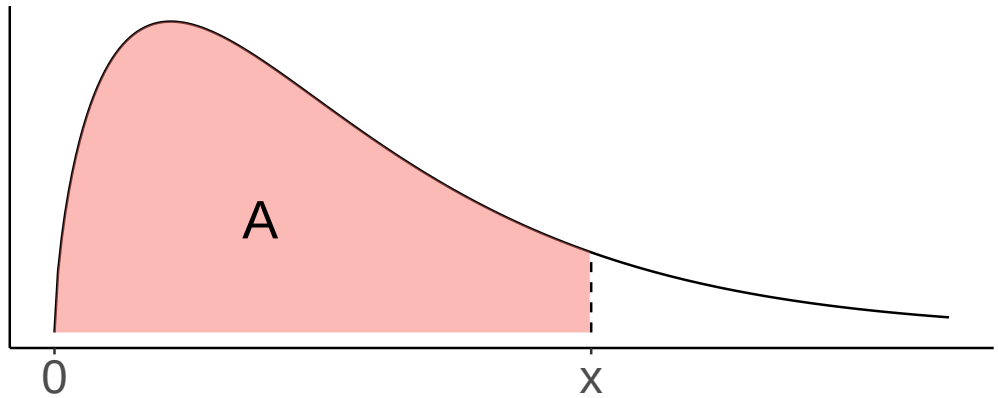
For our example, the χ^2 value is

$$\chi^2 = \frac{(10-6.8)^2}{6.8} + \frac{(7-10.2)^2}{10.2} + \frac{(2-5.2)^2}{5.2} + \frac{(11-7.8)^2}{7.8} = 5.79$$

From the χ^2_1 tables, we see that $\Pr(\chi^2_1 > 3.841) = 0.05$, so definitely $\Pr(\chi^2_1 > 5.79) < 0.05$, so reject H_0.

χ^2 table

Each table entry is x, where $\int_0^x f(x)\,dx = A$ with $X \sim \chi^2_k$.



k	0.010	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.990
1	0.000	0.004	0.016	0.102	0.455	1.323	2.706	3.841	6.635
2	0.020	0.103	0.211	0.575	1.386	2.773	4.605	5.991	9.210
3	0.115	0.352	0.584	1.213	2.366	4.108	6.251	7.815	11.345
4	0.297	0.711	1.064	1.923	3.357	5.385	7.779	9.488	13.277
5	0.554	1.145	1.610	2.675	4.351	6.626	9.236	11.070	15.086

General contingency tables

The χ^2 test can actually be used for the comparison of multiple rows and multiple columns. As an example, consider responses to the cardio workout and physical fitness questionnaire.

Table 4: Cardiovascular workout and physical fitness.

	Physical fitness			
Cardiovascular workouts?	Bad	Medium	Good	Total
No	6	6	3	15
Yes	3	6	6	15
Total	9	12	9	30

Now we have a 2×3 table. Apply the same technique as before:

1. Find the expected frequencies.
2. Calculate $\chi^2 = \sum \frac{(O-E)^2}{E}$.
3. Look up the χ^2 table.

Degrees of freedom

Proposition 13

The number of degrees of freedom for the χ^2 table is

$$\text{df} = (\text{No. rows} - 1) \times (\text{No. columns} - 1)$$

- That is why for the 2×2 table, we used $\text{df} = 1$.
- For the 2×3 table, we will use $\text{df} = 2$.