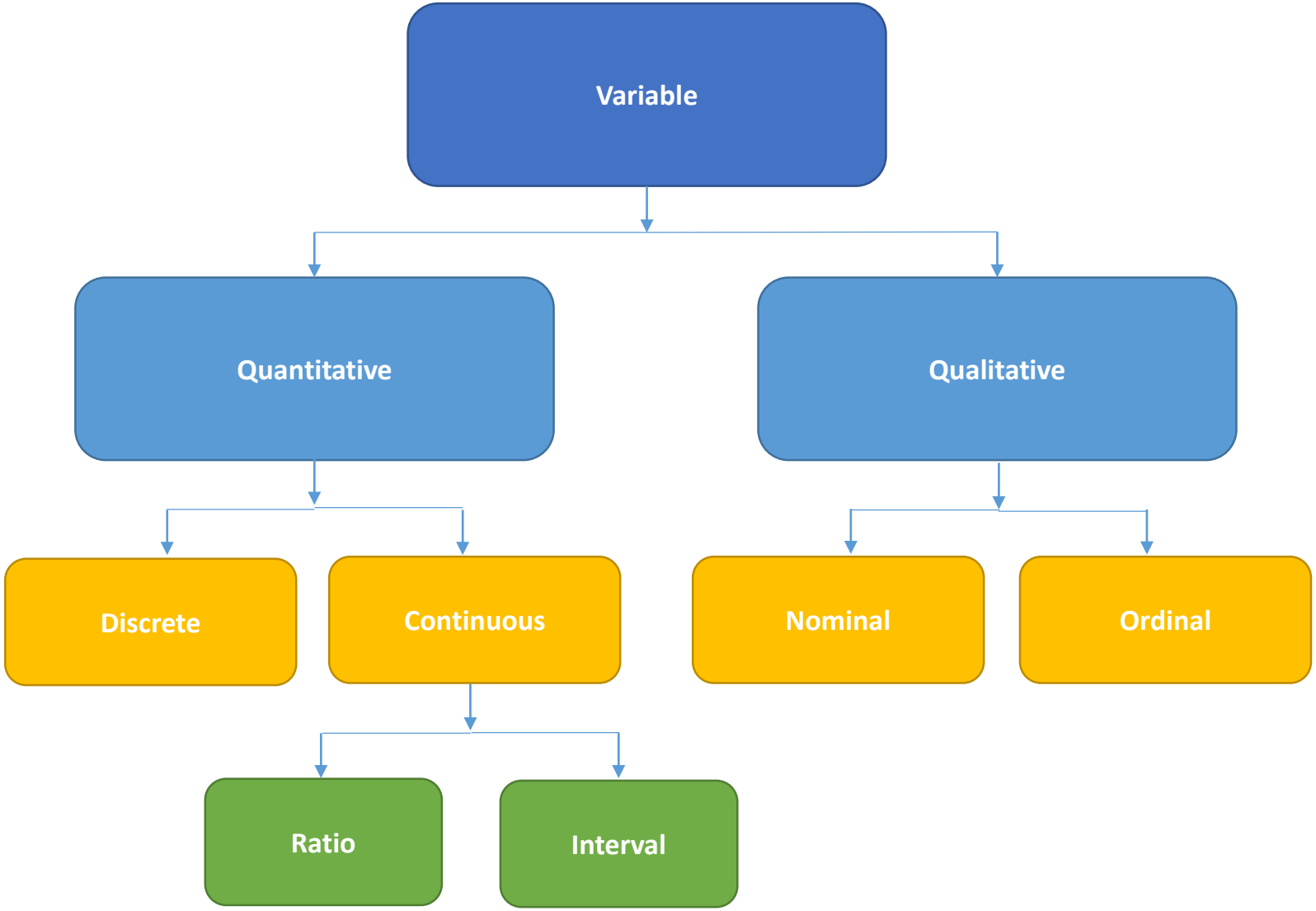
Babraham Institute

Analysis of Qualitative data

Anne Segonds-Pichon

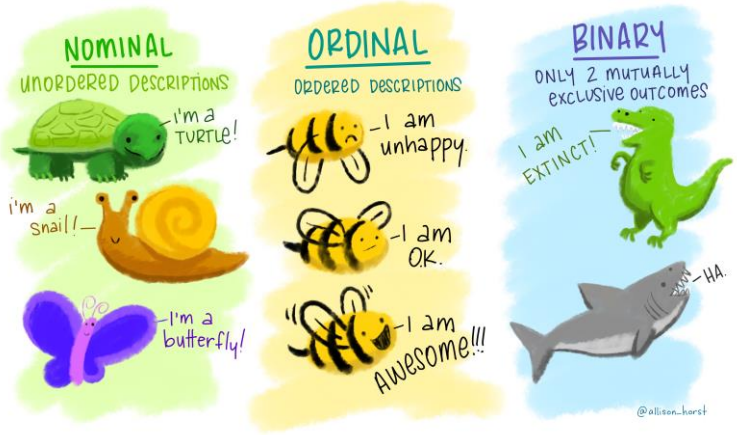
v2020-09

Babraham Bioinformatics


Variable
Quantitative
Qualitative
Discrete
Continuous
Nominal
Ordinal
Ratio
Interval

Qualitative data

- = not numerical
- = values taken = usually names (also *nominal*)
  - e.g. genotypes
- Values can be numbers but not numerical
  - e.g. group number = numerical label but not unit of measurement
- Qualitative variable with intrinsic order in their categories = *ordinal*
- Particular case: qualitative variable with 2 categories: ***binary*** or *dichotomous*
  - e.g. alive/dead or presence/absence



https://github.com/allisonhorst/stats-illustrations#other-stats-artwork

Fisher's exact and Chi^2

Example: cats.dat

- Cats trained to line dance
- 2 different rewards: food or affection
- **Question**: Is there a difference between the rewards?

- **Is there a significant relationship between the 2 variables?**
  - does the reward significantly affect the likelihood of dancing?

- To answer this type of question:
  - **Contingency table**
  - **Fisher's exact or Chi^2 tests**

	Food	Affection
Dance	?	?
No dance	?	?

But first: **how many cats** do we need?

Power calculation cats.dat

- Preliminary results from a pilot study: **25%** line-danced after having received affection as a reward vs. **70%** after having received food.
- **How many cats** do we need?

```
power.prop.test(p1= 0.25, p2= 0.7, sig.level= 0.05, power= 0.8)
```

```
     Two-sample comparison of proportions power calculation

              n = 18.10585
             p1 = 0.25
             p2 = 0.7
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

- Providing the effect size observed in the experiment is similar to the one observed in the pilot study, we will need 2 samples of **18 to 19 cats** to reach significance ($p<0.05$) with a Fisher's exact test.

Plot cats data (From raw data)

```
read_tsv("cats.dat") -> cats
cats
```

```
                Training Dance
1 Food as Reward   Yes
2 Food as Reward   Yes
3 Food as Reward   Yes
4 Food as Reward   Yes
5 Food as Reward   Yes
6 Food as Reward   Yes
```

```
ggplot(cats, aes(Training, fill=Dance))+
  geom_bar(position="fill", colour="black")+
  scale_fill_brewer(palette = 1)+
  ylab("Fraction")
```

Chi-square and Fisher's tests

- Chi^2 test very easy to calculate by hand but Fisher's very hard
- Many software will not perform a Fisher's test on tables > 2x2
- **Fisher's test more accurate** than Chi^2 test on **small samples**
- **Chi^2 test more accurate** than Fisher's test on **large samples**
- Chi^2 test assumptions:
  - 2x2 table: no expected count <5
  - Bigger tables: all expected > 1 and no more than 20% < 5
- Yates's continuity correction
  - All statistical tests work well when their assumptions are met
  - When not: probability Type 1 error increases
  - Solution: corrections that increase p-values
    - Corrections are dangerous: no magic
    - Probably best to avoid them

Chi-square test

- In a chi-square test, the observed frequencies for two or more groups are compared with expected frequencies by chance.

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

  - O = Observed frequencies
  - E = Expected frequencies

- **Example with ‘cats and dogs’**

How are the expected frequencies calculated?

Example: expected frequency of cats line dancing after having received food as a reward.

Direct counts approach:

Expected frequency = (row total)*(column total)/grand total

= **38*76/200** = **14.4**

Probability approach: The Multiplicative Rule

Probability of line dancing: **76/200**
Probability of receiving food: **38/200**

Expected frequency:(**76/200**)*(**38/200**)=0.072: **7.2% of 200 = 14.4**

Observed frequencies

	Food	Affection	Total
Dance	28	48	**76**
No dance	10	114	124
Total	**38**	162	**200**

Expected frequencies

	Food	Affection
Dance	**14.4**	61.6
No dance	23.6	100.4

Chi² test

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

Observed frequencies

	Food	Affection
Dance	28	48
No dance	10	114

Expected frequencies

	Food	Affection
Dance	14.4	61.6
No dance	23.6	100.4

$\text{Chi}^2 = (28-14.4)^2/14.4 + (48-61.6)^2/61.6 + (10-23.6)^2/23.6 + (114-100.4)^2/100.4$

$= 25.35$

Is 25.35 big enough for the test to be significant?

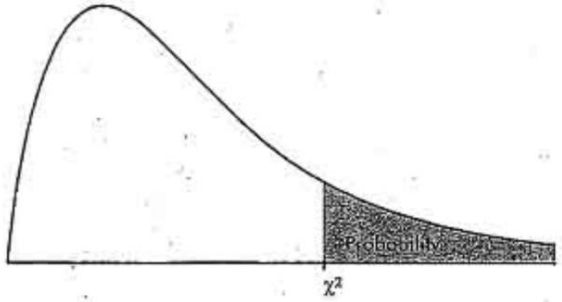
Is 28.4 big enough for the test to be significant?

The old fashion way

Degree of freedom: df

df = (row-1)(col-1)=1

	Food	Affection
Dance	28	48
No dance	10	114



Critical value

TABLE C: χ^2 CRITICAL VALUES

	Tail probability p								
df	.25	.20	.15	.10	.05	.025	.02	.01	.005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19

χ^2 = 25.35 > 3.84 so Yes!

Prepare cats data for the stats

```
chisq_test()
fisher_test()
```

	Food	Affection
Dance	?	?
No dance	?	?

```
                Training Dance
1 Food as Reward      Yes
2 Food as Reward      Yes
3 Food as Reward      Yes
4 Food as Reward      Yes
5 Food as Reward      Yes
6 Food as Reward      Yes
```

Training <chr>	No <int>	Yes <int>
Affection as Reward	114	48
Food as Reward	10	28

Plot cats data (From raw data)

```
             Training Dance
1 Food as Reward   Yes
2 Food as Reward   Yes
3 Food as Reward   Yes
4 Food as Reward   Yes
5 Food as Reward   Yes
6 Food as Reward   Yes
```

```
cats %>%
  group_by(Training, Dance) %>%
    count() %>%
    ungroup() %>%
    pivot_wider(names_from = Dance, values_from = n)-> cats.summary
```

Training <chr>	Dance <chr>	n <int>
Affection as Reward	No	114
Affection as Reward	Yes	48
Food as Reward	No	10
Food as Reward	Yes	28

Training <chr>	No <int>	Yes <int>
Affection as Reward	114	48
Food as Reward	10	28


```
cats.summary %>%
    select(No,Yes) %>%
    fisher_test()
```

n <int>	p <dbl>	p.signif <chr>
200	1.31e-06	****

```
chisq_test()
fisher_test()
```

Have a go!

Chi-square and Fisher's Exact tests

```
cats.summary %>%
    select(No,Yes) %>%
    fisher_test()
```

n <int>	p <dbl>	p.signif <chr>
200	1.31e-06	****

```
cats.summary %>%
    select(No,Yes) %>%
    chisq_test()
```

	n <int>	statistic <dbl>	p <dbl>	df <int>	method <chr>	p.signif <chr>
1	200	23.52028	1.24e-06	1	Chi-square test	****

```
cats.summary %>%
    select(No,Yes) %>%
    chisq_test(correct = FALSE)
```

	n <int>	statistic <dbl>	p <dbl>	df <int>	method <chr>	p.signif <chr>
1	200	25.35569	4.77e-07	1	Chi-square test	****

Answer: Training significantly affects the likelihood of cats line dancing (p=4.8e-07).