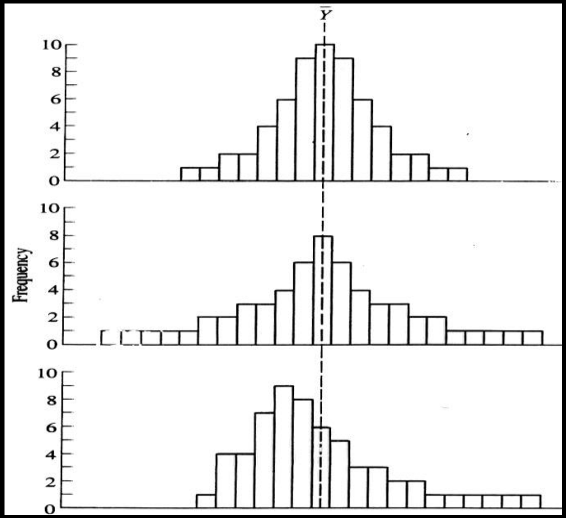
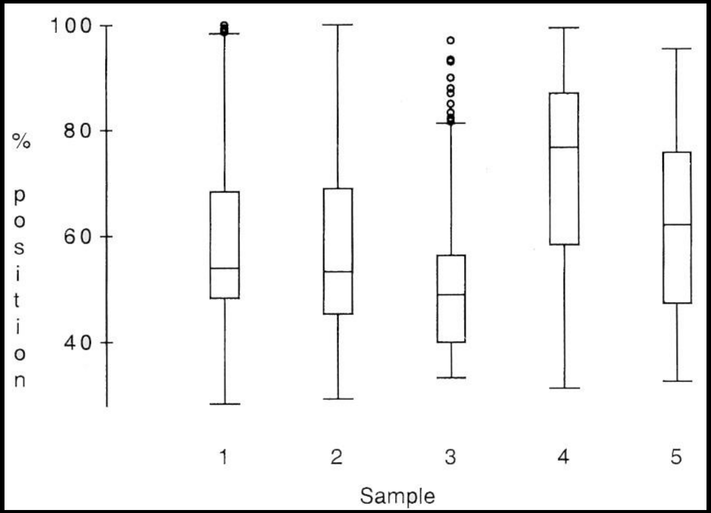
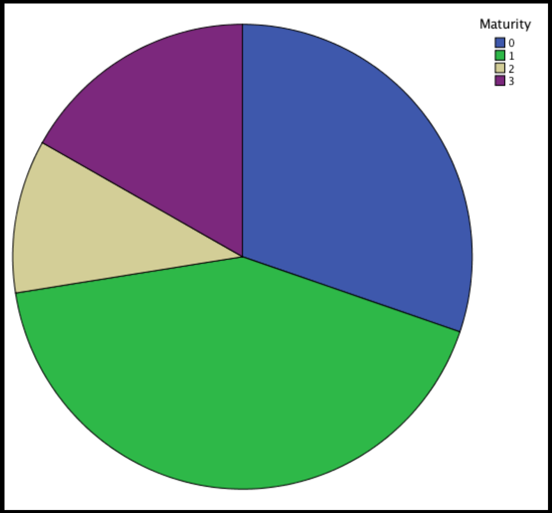
Data Handling :

A practical approach







Lecture 7 Chi-Square Test

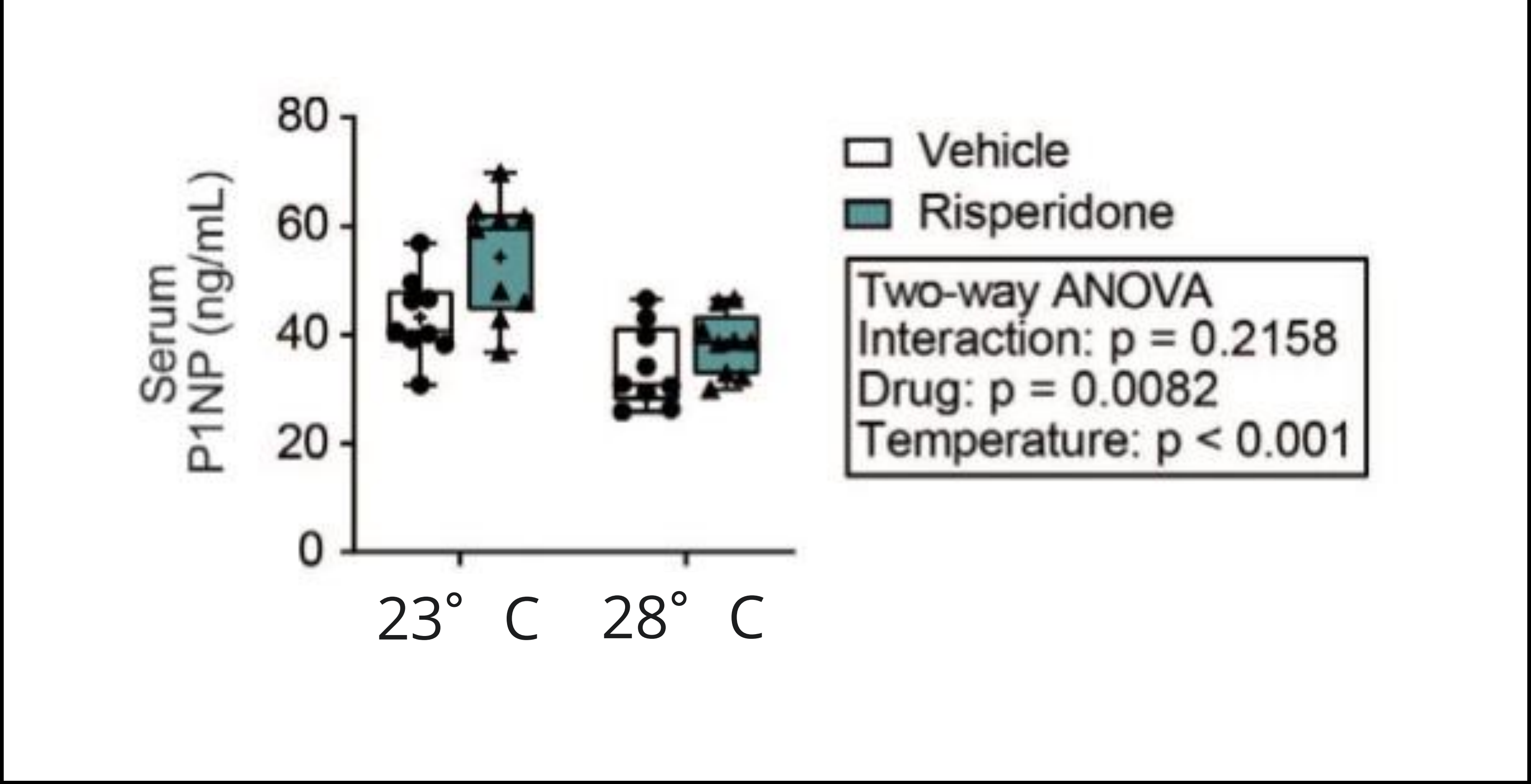
Dr Yu Mo, Zoology

moyu@tcd.ie | https://github.com/github-moyu/Teaching

Summary of lecture 5/6

- Moving from comparing 2 means to more than 2 means
- Simple one-way analysis of variance (ANOVA)
- Generation of F ratio (within versus between group variation), p value and 2 types of degrees of freedom
- More complex designs – concept of an interaction term

Effect of housing temperature on drug-induced bone loss

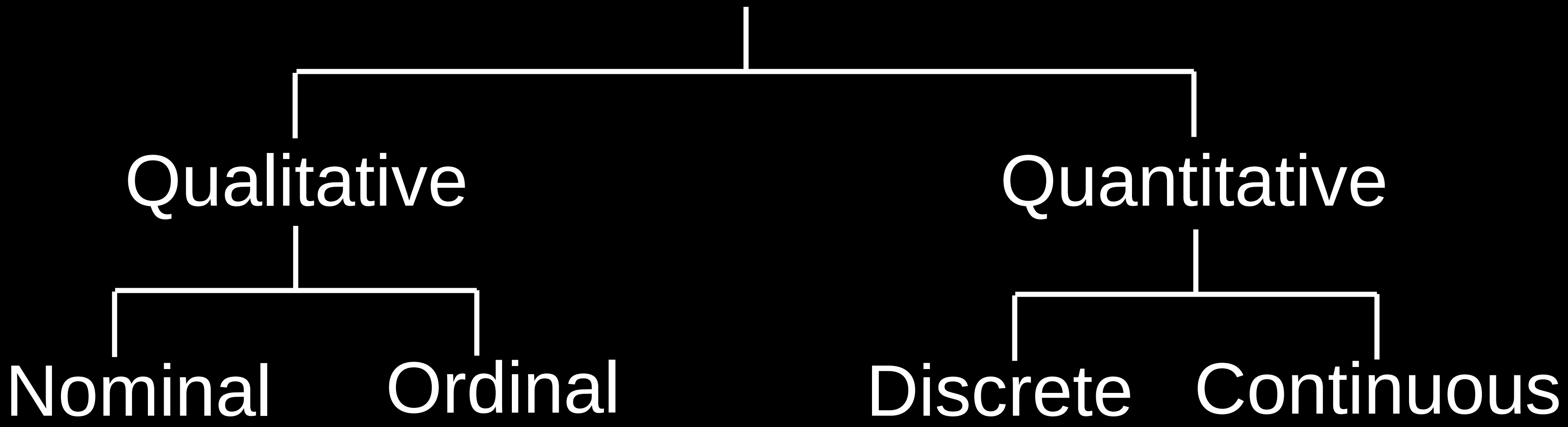


Two-way ANOVA
Interaction: $p = 0.2158$
Drug: $p = 0.0082$
Temperature: $p < 0.001$

And now for something a little different

How to interpret, plot and analyse frequency data

Type of data



Qualitative

Quantitative

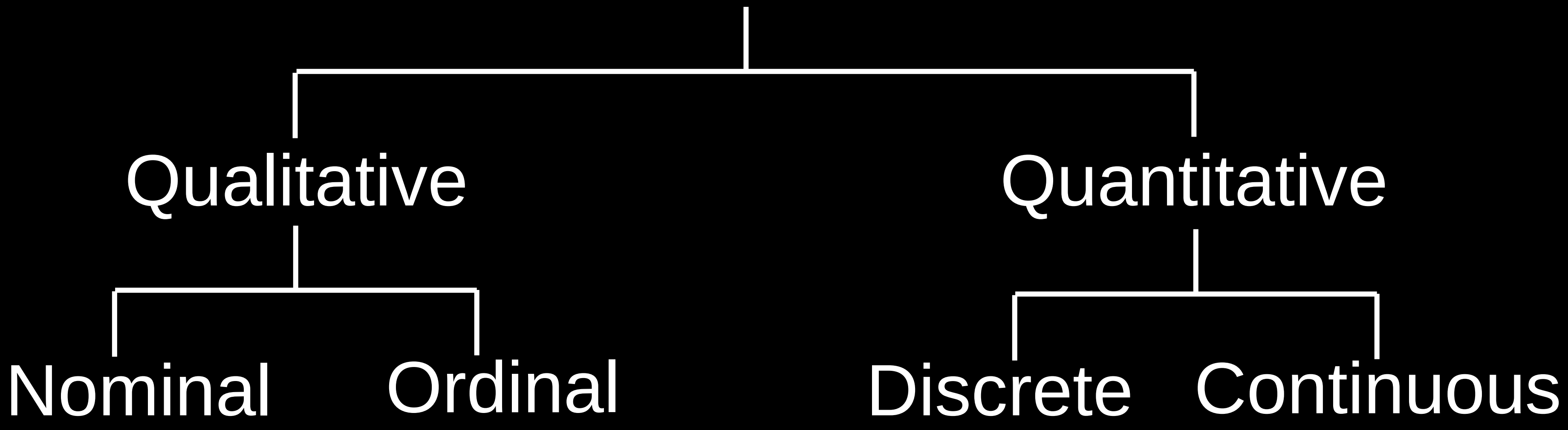
Nominal

Ordinal

Discrete

Continuous

Type of data



Qualitative

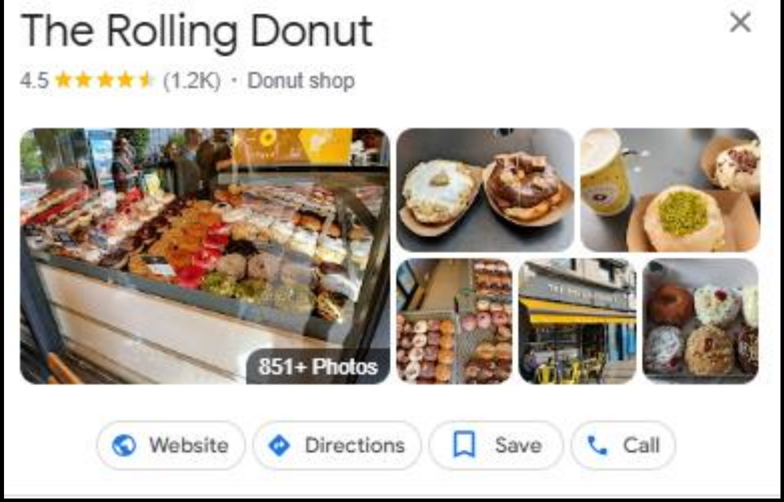
Quantitative

Nominal

Ordinal

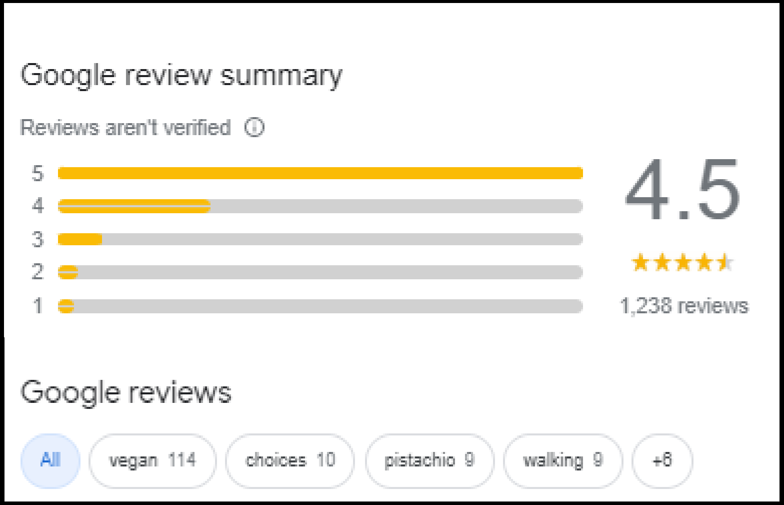
Discrete

Continuous



The Rolling Donut

4.5 ★★★★★ (1.2K) · Donut shop



Google review summary

Reviews aren't verified

5	
4	
3	
2	
1	

4.5 ★★★★★

1,238 reviews

Google reviews

Analysis of frequency data

Type of data

  - Frequency
  - Categorical

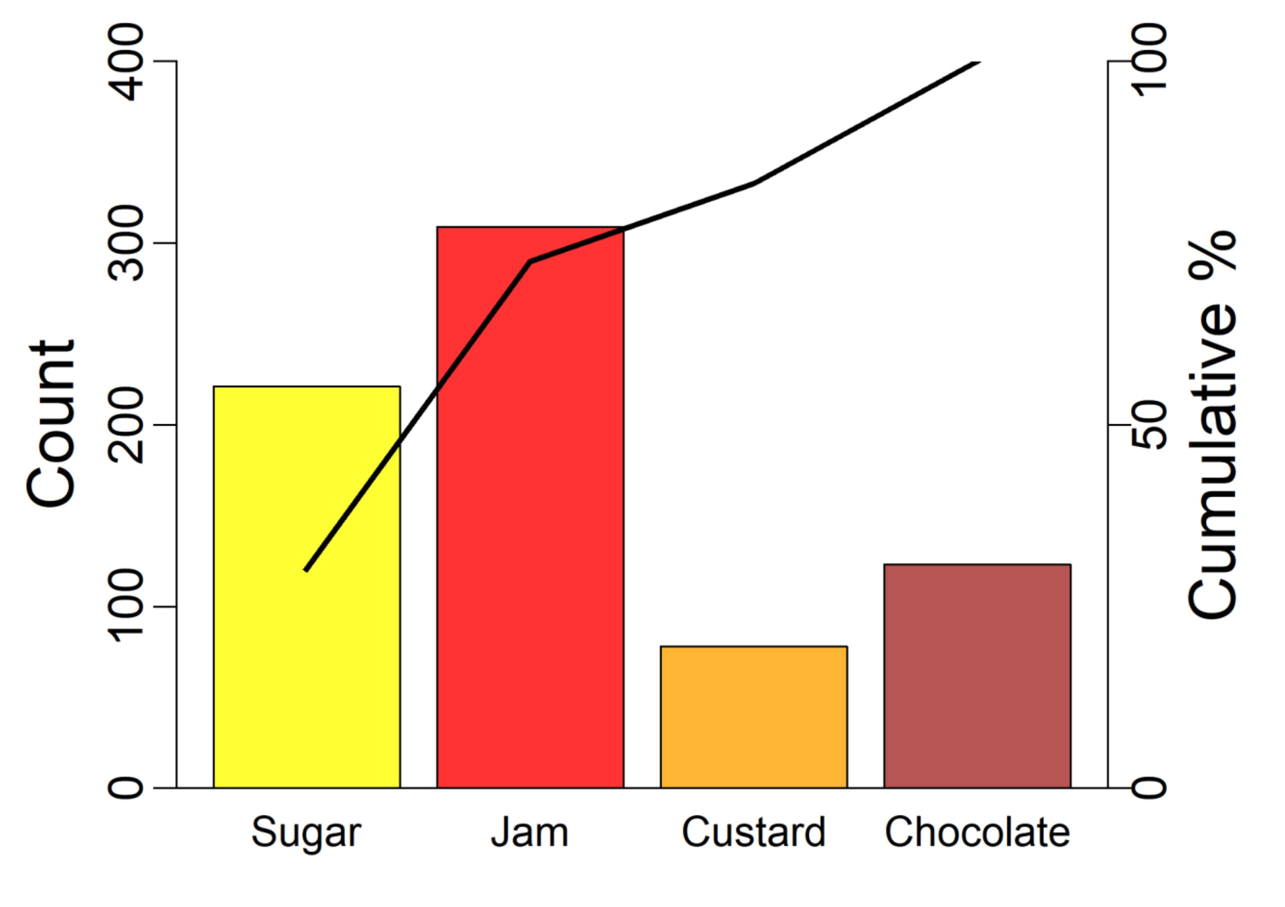
- Independent observations
- Random sampling
- Presentation of data

Doughnuts sold today

Frequency breakdown by type: 4 categories

Category	Frequency	Percent	Cumulative percent
Sugar	221	30.2	30.2
Jam	309	42.3	72.5
Custard	78	10.7	83.2
Chocolate	123	16.8	100
Total	731	100	-

Doughnuts sold today



The relationship between college & doughnut type

College and doughnut type

H_0: There is no association between college & the type of doughnut students like

- Proportion of students that like jam doughnut in zoology is equal to the proportion of students in geography that like jam doughnut

H_1: There is an association between college & the type of doughnut students like

- Proportion of students that like jam doughnut in zoology is **NOT** equal to the proportion of students in geography that like jam doughnut

College and doughnut type

	Zoology	Geography	Total
Jam	21	36	57
Custard	102	69	171
Total	123	105	228

Calculation of expected values

	Zoology	Geography	Total
Jam	21	36	57
Custard	102	69	171
Total	123	105	228

- What do we expect if the null hypothesis is true?
- Overall 57/228 = 25% students are jam-liking

- If there is no difference between zoology & geography, we would expect 25% of each
- Zoology:

  25% of 123 = 30.75
- Geography:

  25% of 105 = 26.25

Calculation of expected values

	Zoology	Geography	Total
Jam	21	36	57
Custard	102	69	171
Total	123	105	228

Custard-liking zoologists:

123*171/228 = 92.25

Custard-liking geographers:

105*171/228 = 78.75

Similarly we work out the expected values for the other two cells (custard)

$$\text{Expected values} = \text{Column total} * \underset{H_0}{\underline{\text{Row total/total number of obs.}}}$$

Observed values

	Zoology	Geography
Jam	21	36
Custard	102	69

Expected values

	Zoology	Geography
Jam	30.75	26.25
Custard	92.25	78.75

Calculation of Chi-squared statistic

We now need a measure of the difference between the observed and the expected

X^2 = Sum of (Observed frequency - Expected frequency)^2/Expected frequency

Observed

	Zoology	**Geography**
Jam	21	36
Custard	102	69

Expected

	Zoology	**Geography**
Jam	30.75	26.25
Custard	92.25	78.75

X^2 = Sum of (Observed frequency - Expected frequency)^2/Expected frequency

$$X^2 = (21-30.75)^2/30.75 + (102-92.25)^2/92.25$$
$$+ (36-26.25)^2/26.25 + (69-78.75)^2/78.75$$
$$= 3.09 + 1.03 + 3.02 + 1.21 = 8.35$$

df = (number of rows-1) * (number of columns -1)

= (2-1)*(2-1) = 1

alpha

Degrees of freedom

V	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005	0.001
1	0.000	0.016	0.445	2.706	**3.841**	5.02	6.64	7.88	10.82
2									
3									
4									
5									
30									
100									

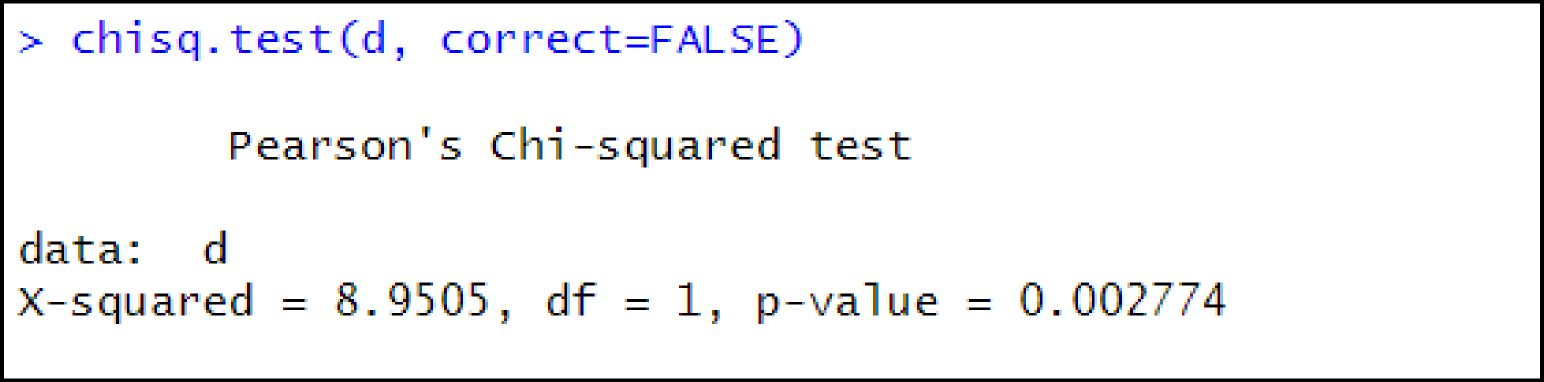
Critical values of the chi-square distribution

Examine the critical values of the Chi-squared distribution

- Power of the test 0.05 (95%)
- Critical value at df = 1

3.84

- Any value > 3.84 Reject null hypothesis
- Any value < 3.84 Do not reject null hypothesis
- Since 8.35 > 3.84 we reject the null hypothesis

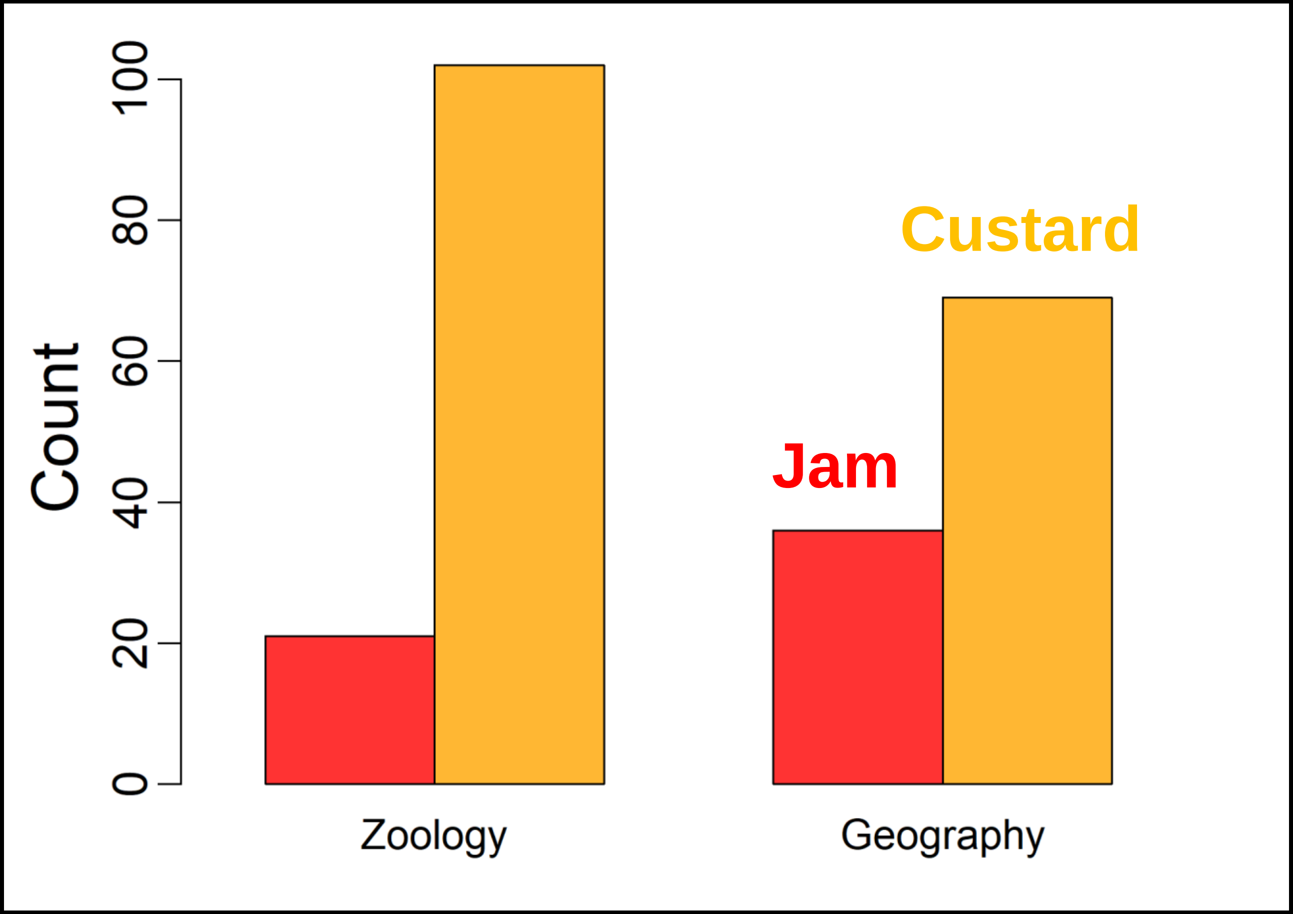


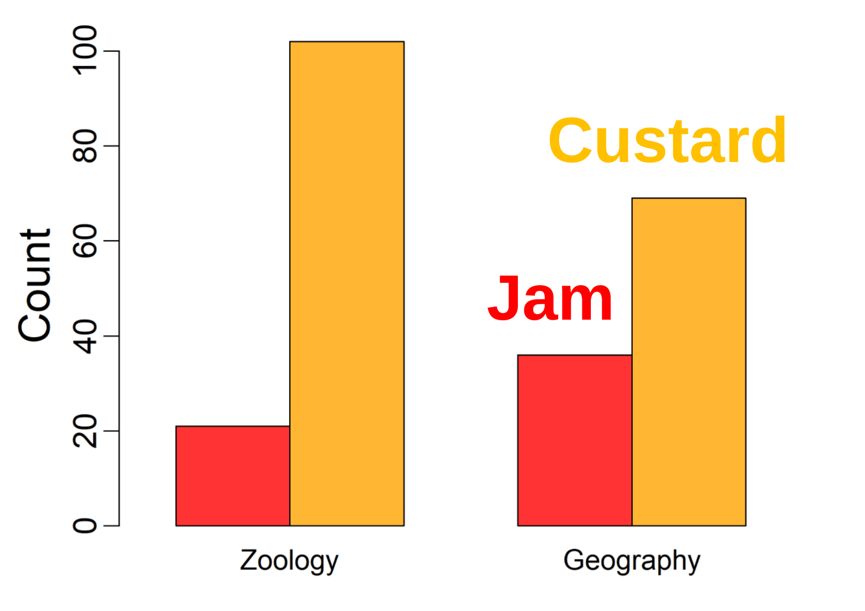
```
> chisq.test(d, correct=FALSE)

        Pearson's Chi-squared test

data:  d
X-squared = 8.9505, df = 1, p-value = 0.002774
```

There is an ASSOCIATION between college and doughnut preference





Chi-test

Frequency data (discrete)

+

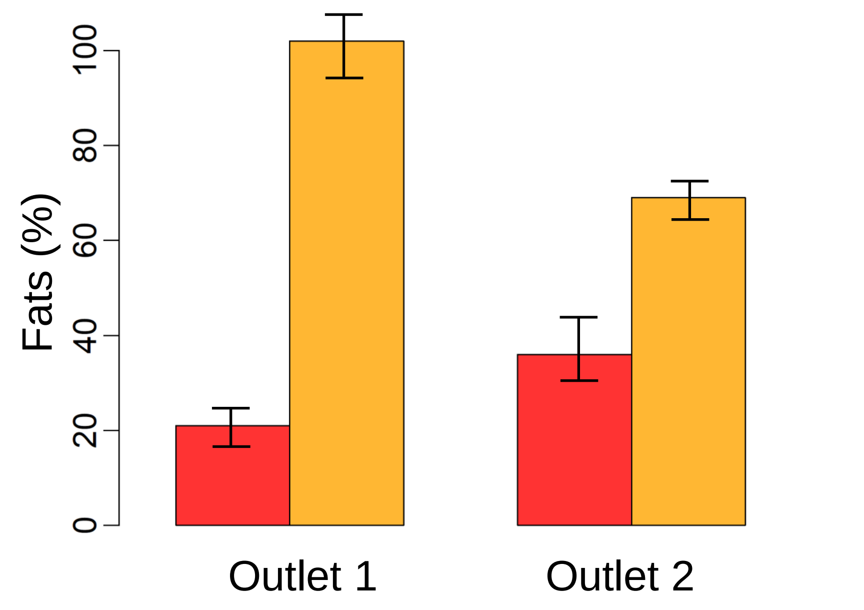
Categorical

ANOVA

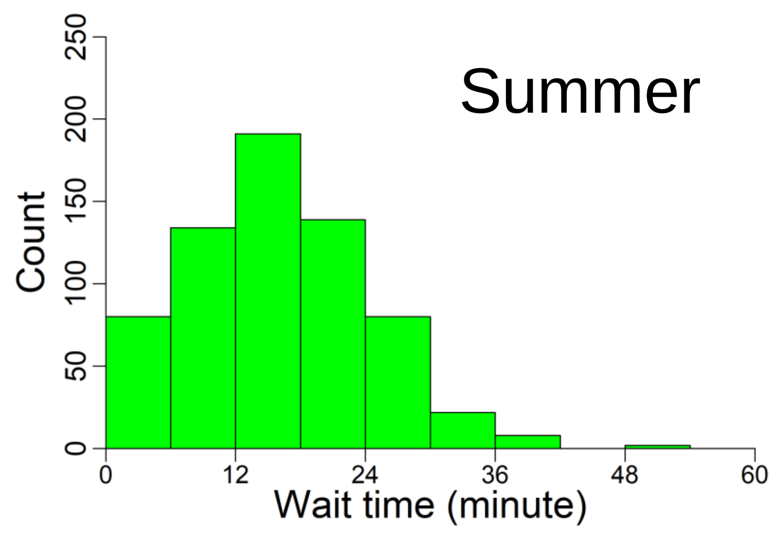
Continuous data

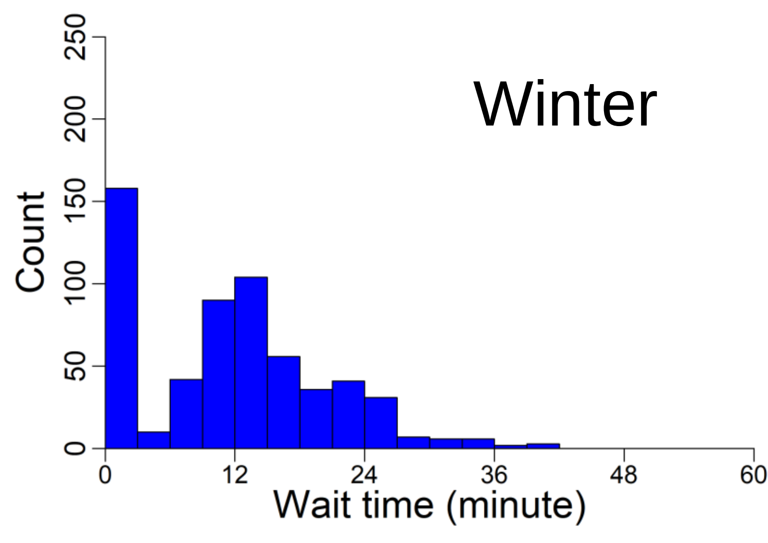
+

Categorical



Analysis influence of season on waiting time



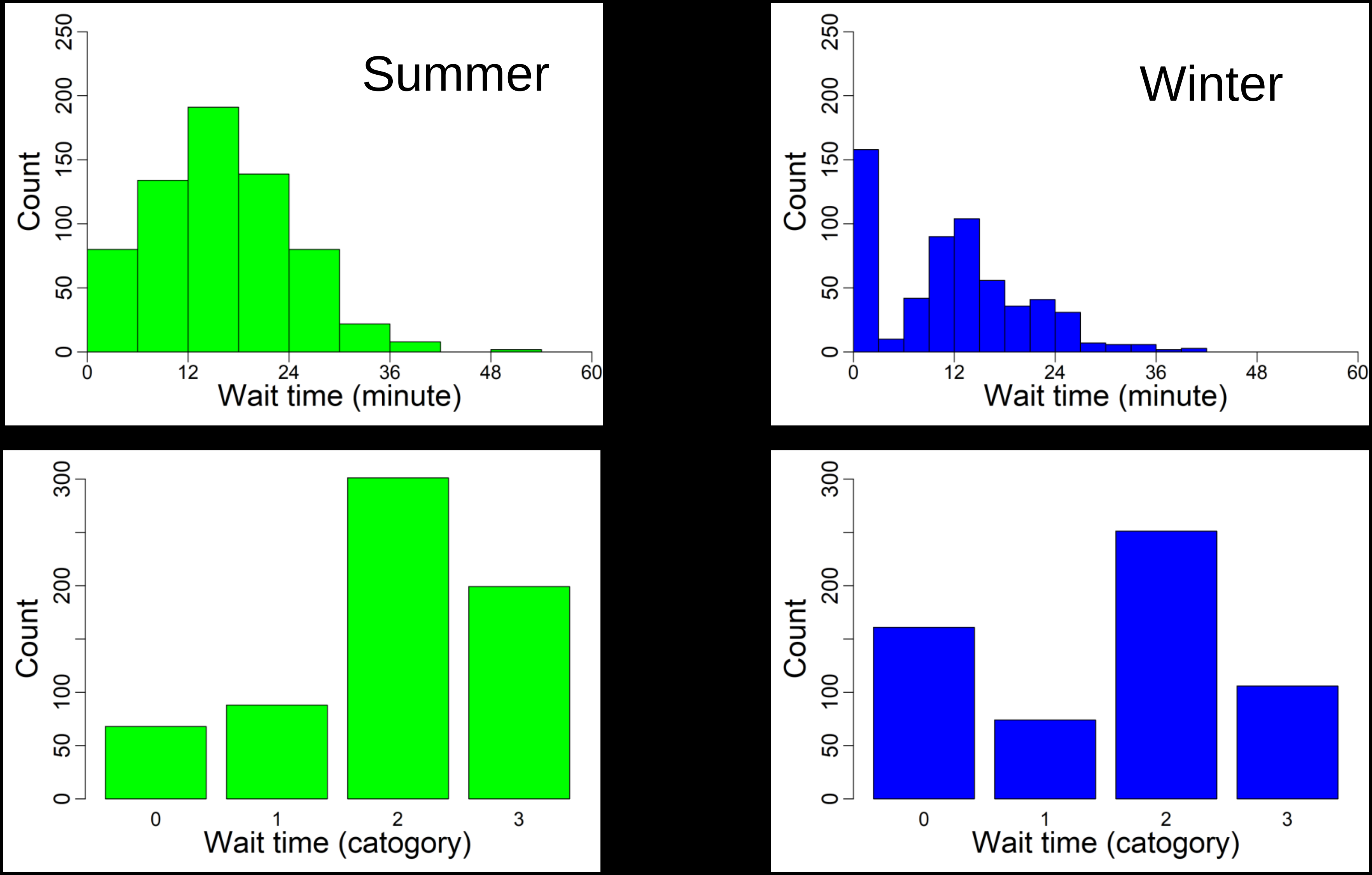


Data extremely skewed (not t-test)

Solution: create categories

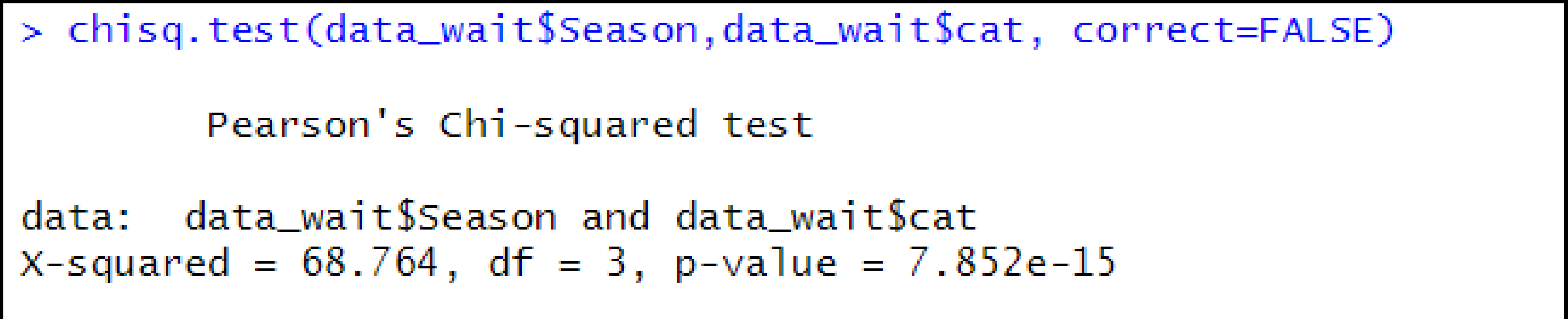
Status coded: 0 (<5), 1 (<10), 2 (<20) & 3 (rest)

Analysis influence of season on waiting time



Analysis influence of season on waiting time

	0	1	2	3
Summer	68	88	301	199
Winter	161	74	251	106



```
> chisq.test(data_wait$Season,data_wait$cat, correct=FALSE)

        Pearson's Chi-squared test

data:  data_wait$Season and data_wait$cat
X-squared = 68.764, df = 3, p-value = 7.852e-15
```

df = (number of rows-1) * (number of columns -1)
  = (2-1) * (4-1) = 3

Not more than 20% of the cells should have an n less than 5

Solution: combine categories
for 2X2 tables can use Fisher’s exact test

Review

	Data	**df**
t-test	Continuous (two means)	n-1
ANOVA	Continuous (more than two means)	num df = denom df=
CHI	Discrete (frequency)	(# cat1 -1)*(# cat2 -1)