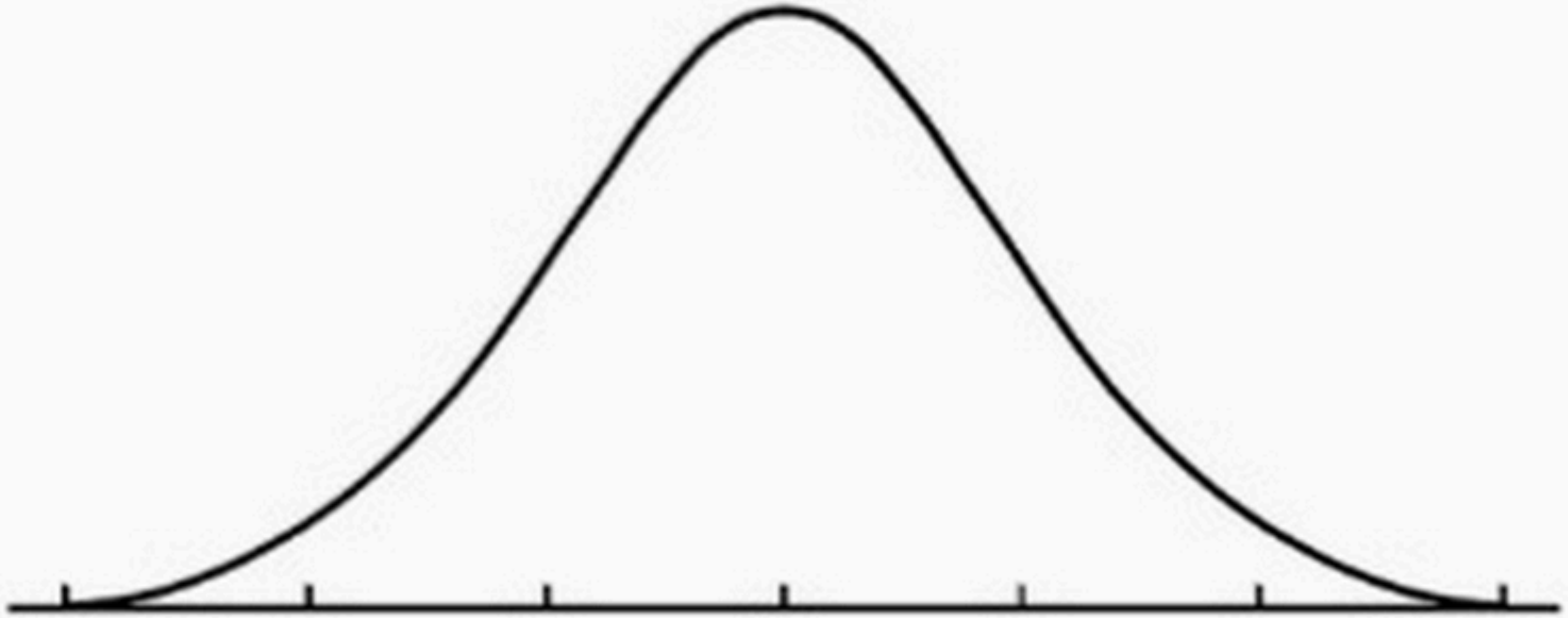
Correlation test: Fisher’s exact, Chi-square

Instructor: Steven Moran, University of Miami

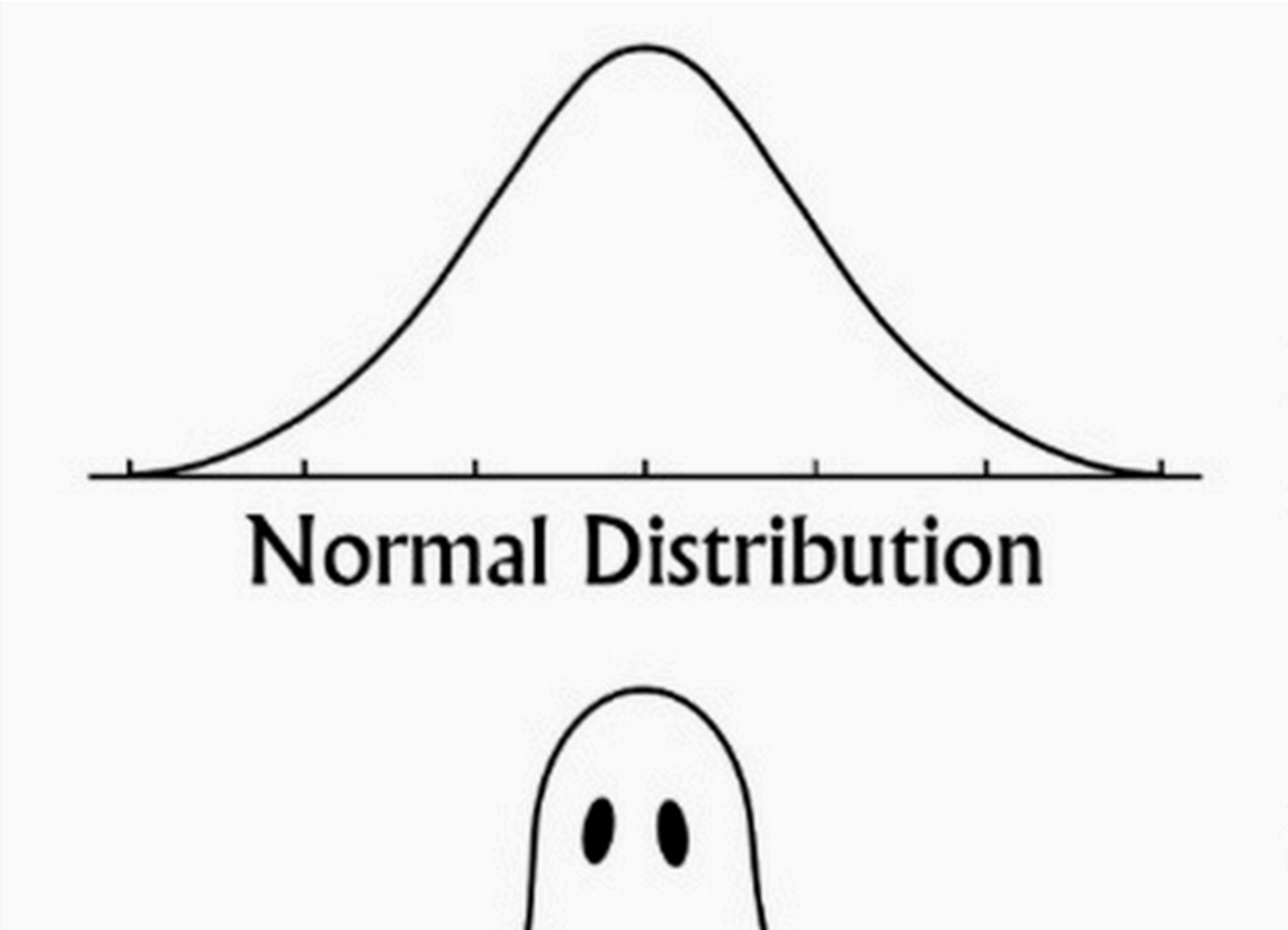
APY 313 Spring 2023

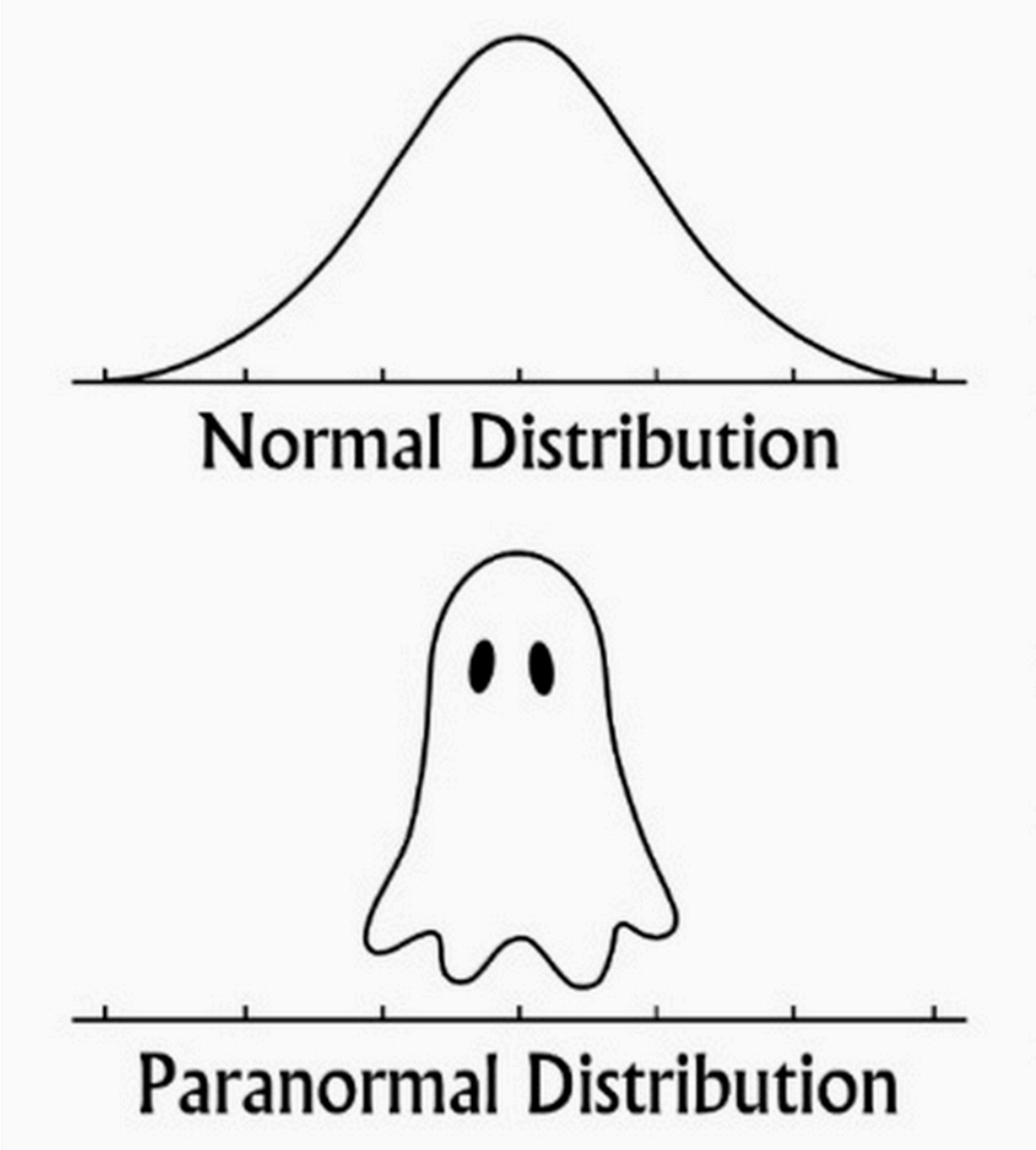
Statistical tests

- **Regression tests:** used to test cause-and-effect relationships, e.g., if the change in one or more continuous variable predicts change in another variable.
  - Simple linear regression, multiple linear regression, logistic regression
- **Comparison tests**: look for the difference between the means of variables, i.e., comparison of means.
  - T-tests are used when comparing the means of precisely two groups, e.g., the average heights of men and women. I
  - Independent t-test: test for the difference between the same variable from different populations, e.g., comparing dogs to cats.
  - ANOVA and MANOVA tests are used to compare the means of more than two groups or more, e.g., the average weights of children, teenagers, and adults.
- **Correlation tests**: test for an association between variable checking whether two variables are related.
  - Pearson Correlation: test for the strength of the association between two continuous variables.
  - Spearman Correlation: tests for the strength of the association between two ordinal variables (note: it does not rely on the assumption of normally distributed data)
  - Chi-Square Test: tests for the strength of the association between two categorical variables.



Normal Distribution


Normal Distribution


Normal Distribution
Paranormal Distribution

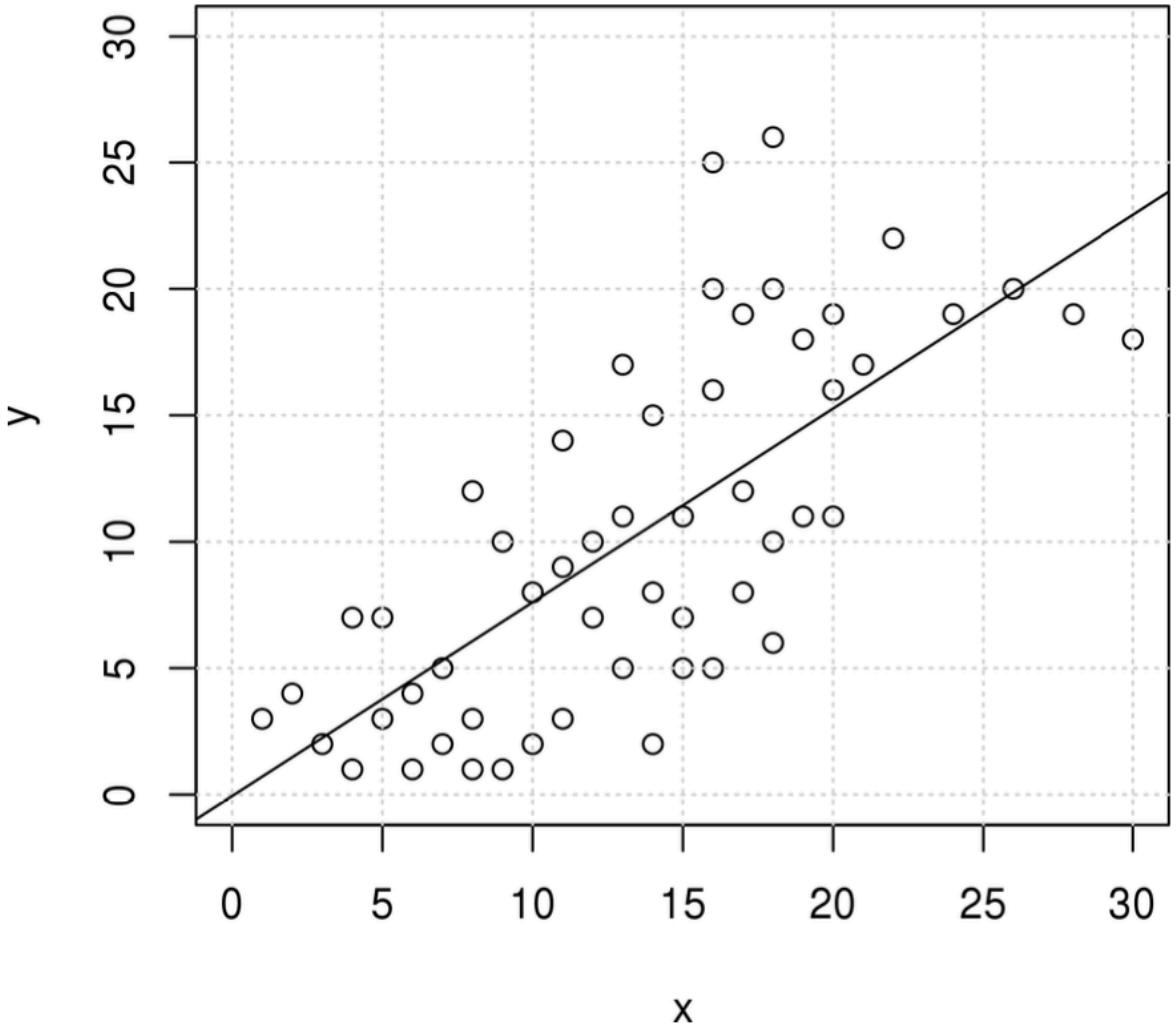


Figure 1. A correlation between two fictitious variables *x* and *y*

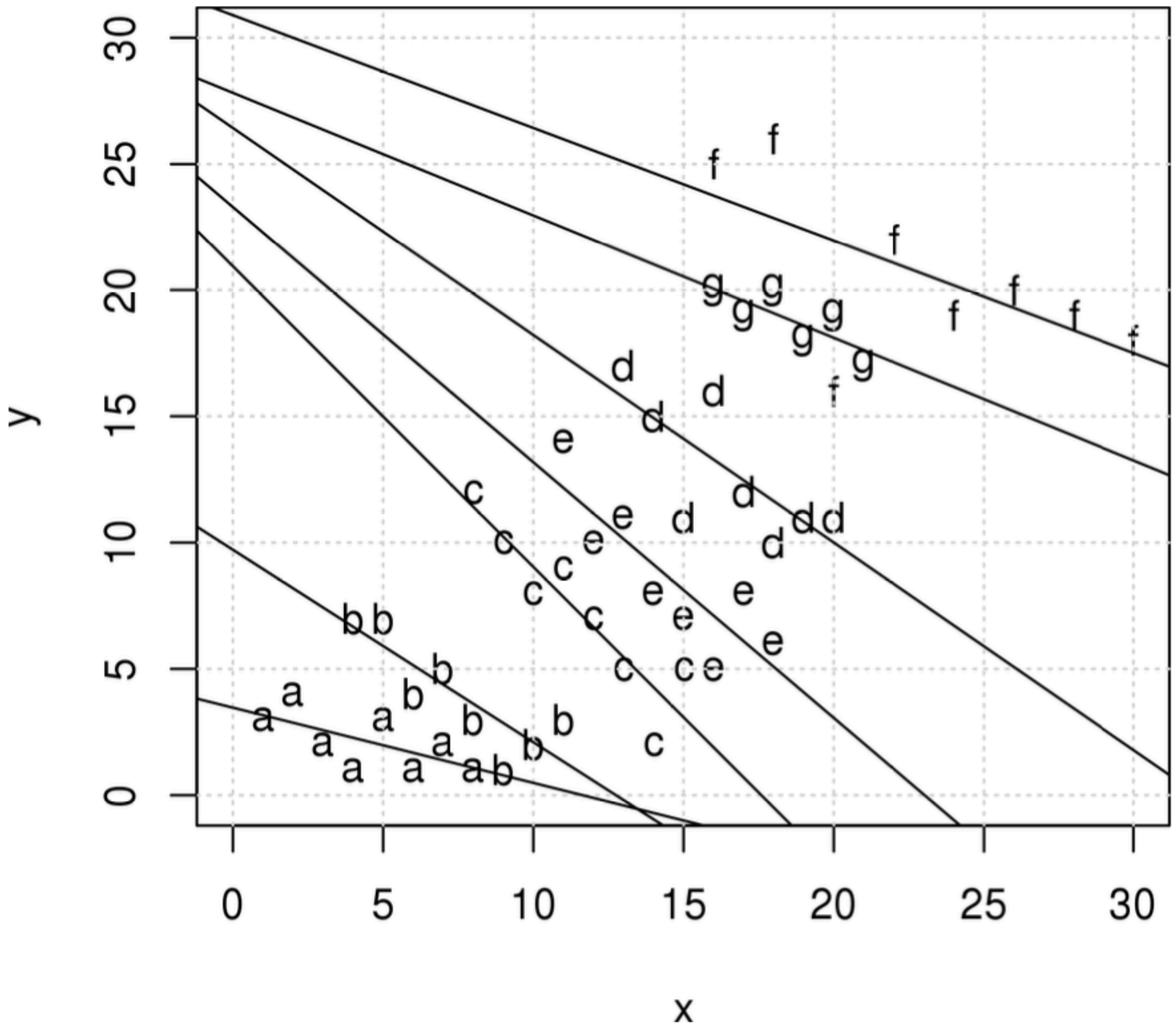
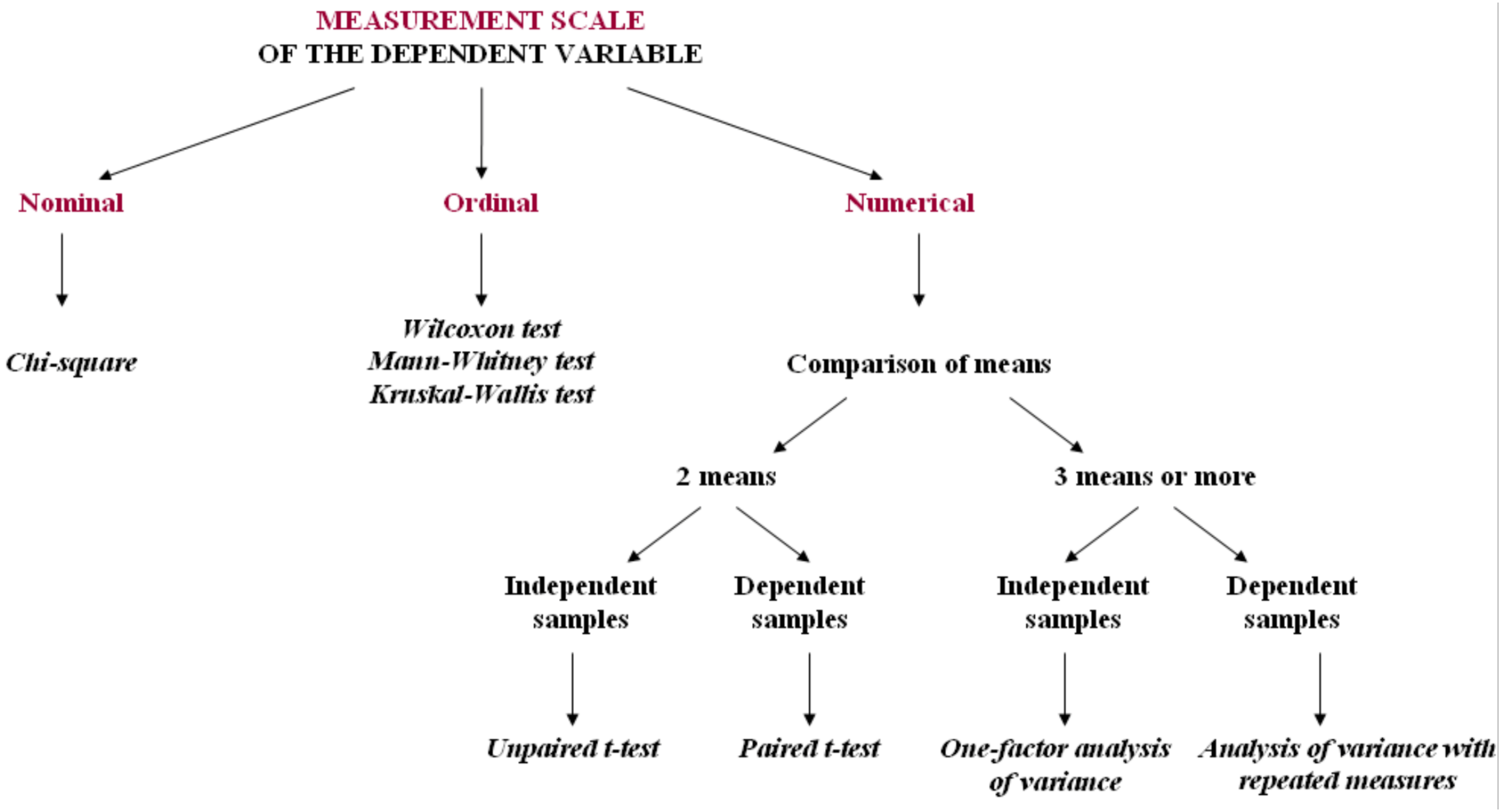


Figure 2. A correlation between two fictitious variables *x* and *y*, controlled for a fictitious third variable

Choosing a test



MEASUREMENT SCALE
OF THE DEPENDENT VARIABLE

Nominal

Chi-square

Ordinal

Wilcoxon test
Mann-Whitney test
Kruskal-Wallis test

Numerical

Comparison of means

2 means

Independent samples

Unpaired t-test

Dependent samples

Paired t-test

3 means or more

Independent samples

One-factor analysis of variance

Dependent samples

Analysis of variance with repeated measures

Chi-square

- You compare the observed frequencies with the theoretical (expected) frequencies
  - https://www.youtube.com/watch?v=SvKv375sacA
- Is variation beyond what we would expect due to chance alone?
  - http://www.r-tutor.com/elementary-statistics/goodness-fit/chi-squared-test-independence

Distributions and why they matter

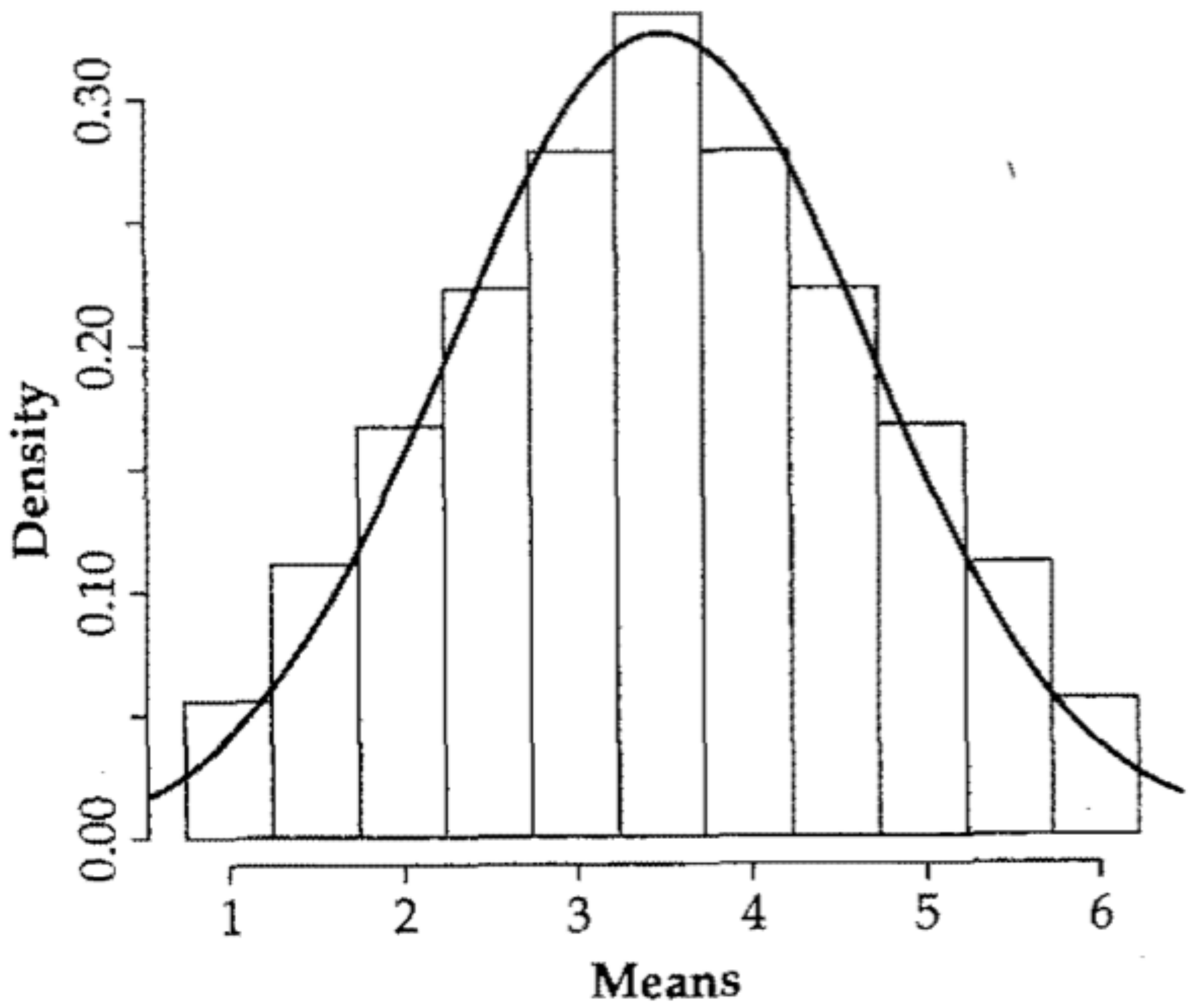


Figure 2.1 The frequency distribution of the mean for the samples illustrated in Table 2.1.

Johnson 2008

Distributions and why they matter

- Consider sampling from a uniform distribution of the values 1 … 6
- Average of two rolls same on the diagonals
- How many ways to get a mean of 3.5?

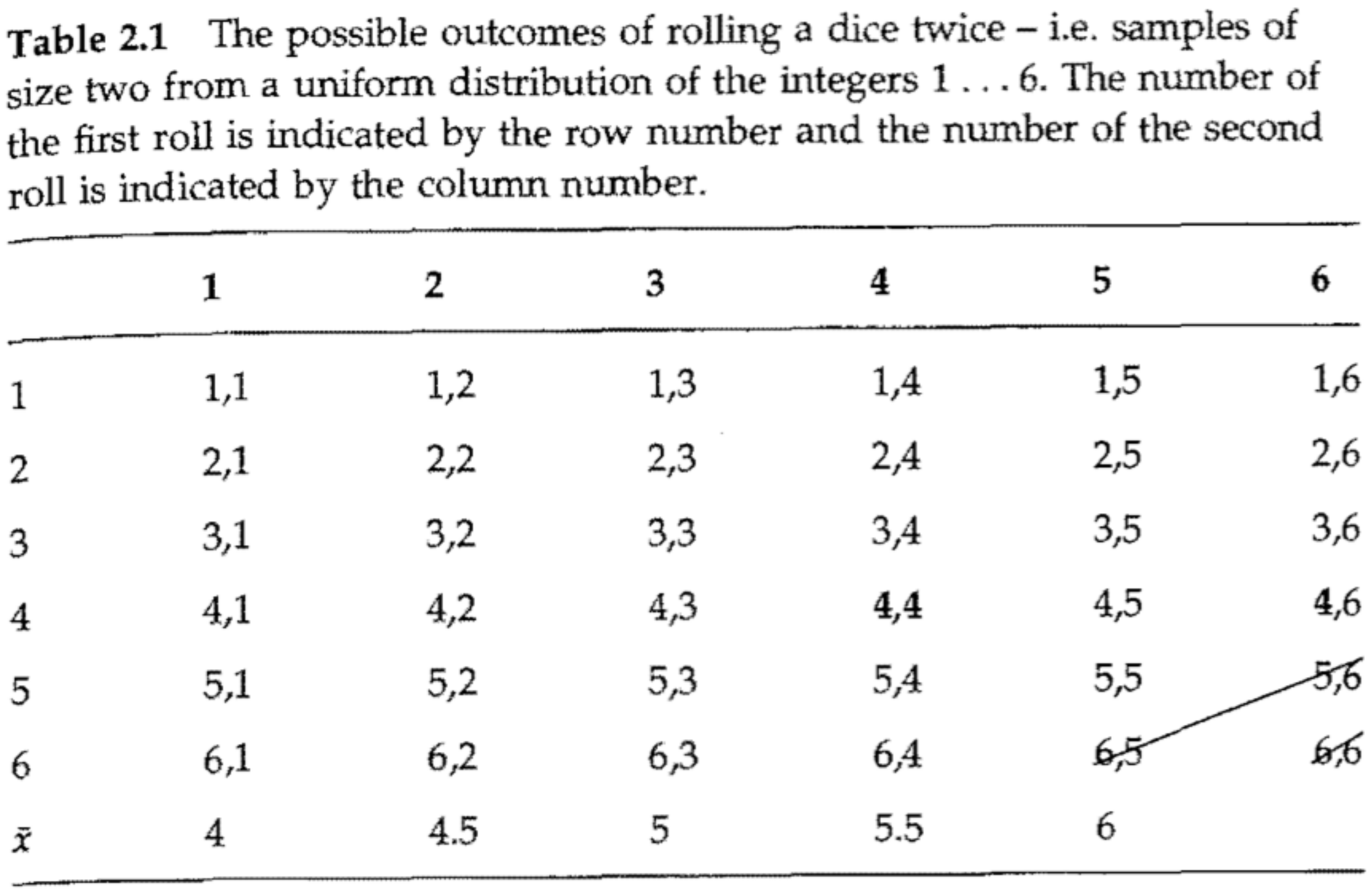


Table 2.1 The possible outcomes of rolling a dice twice – i.e. samples of size two from a uniform distribution of the integers 1 . . . 6. The number of the first roll is indicated by the row number and the number of the second roll is indicated by the column number.

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	~~5,6~~
6	6,1	6,2	6,3	6,4	~~6,5~~	~~6,6~~
$\bar{x}$	4	4.5	5	5.5	6	

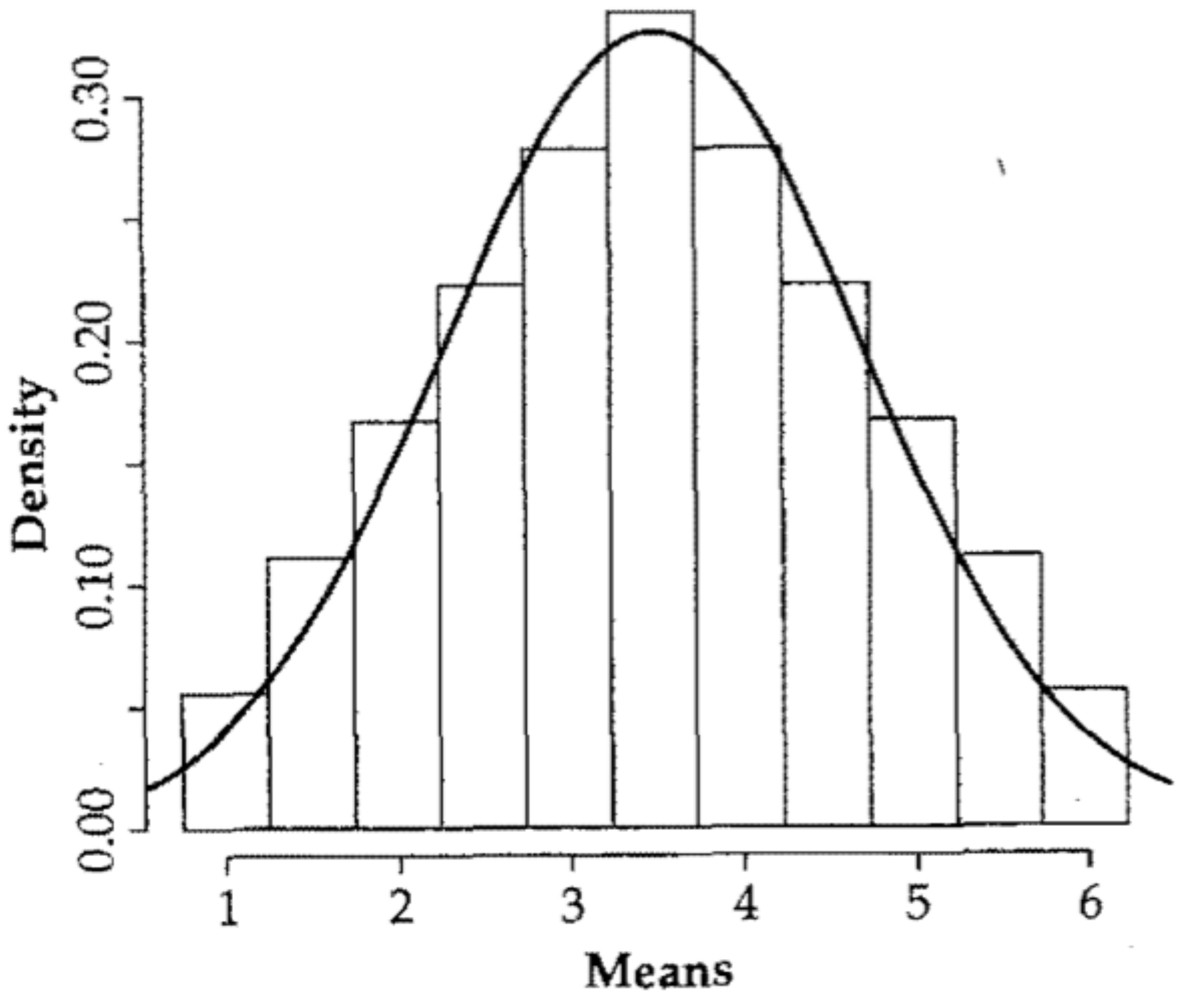


Figure 2.1 The frequency distribution of the mean for the samples illustrated in Table 2.1.

Johnson 2008

Probability

- The probability of averaging 3.5 dots on two rolls of the dice is 6/36 = 0.1666. So, why does the vertical axis in Figure 2.1 go up to 0.3?
- What does it mean to be labelled “density” and how is probability density different from probability?
- Problem is: on a continuous measurement scale, finding one particular value is zero
- On a continuous dimension, we can’t give a probability for a specific value of the measurement variable. Instead we can only state the probability of a region under curve.

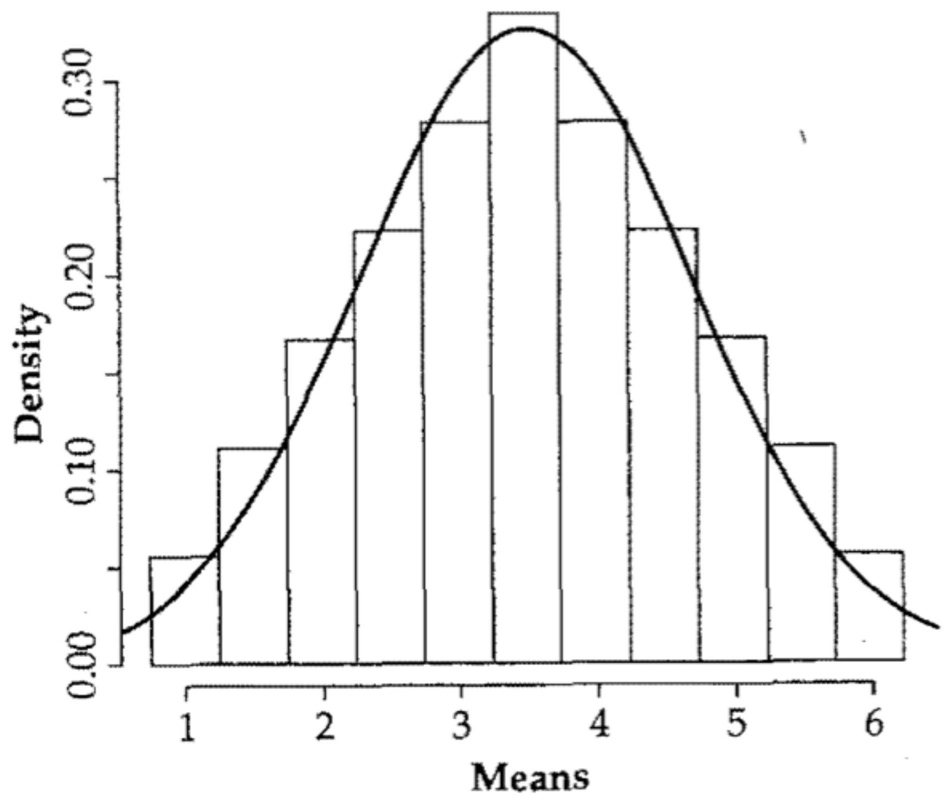


Figure 2.1 The frequency distribution of the mean for the samples illustrated in Table 2.1.

Probability

- We can’t say what the probability of a measurement of 500 is, but we can say for example that about 16% of the cumulative distribution in Figure 2.2 falls to the left of 500 ms - that given a population like this one (mean = 600, standard deviation = 100) we expect 16% of the values in a representative sample to be lower than 500.
- In the probability density function, the area under the curve from 0 to 500 ms is 16% of the total area under the curve, so the value of the cumulative density function at 500 ms is 0.16.

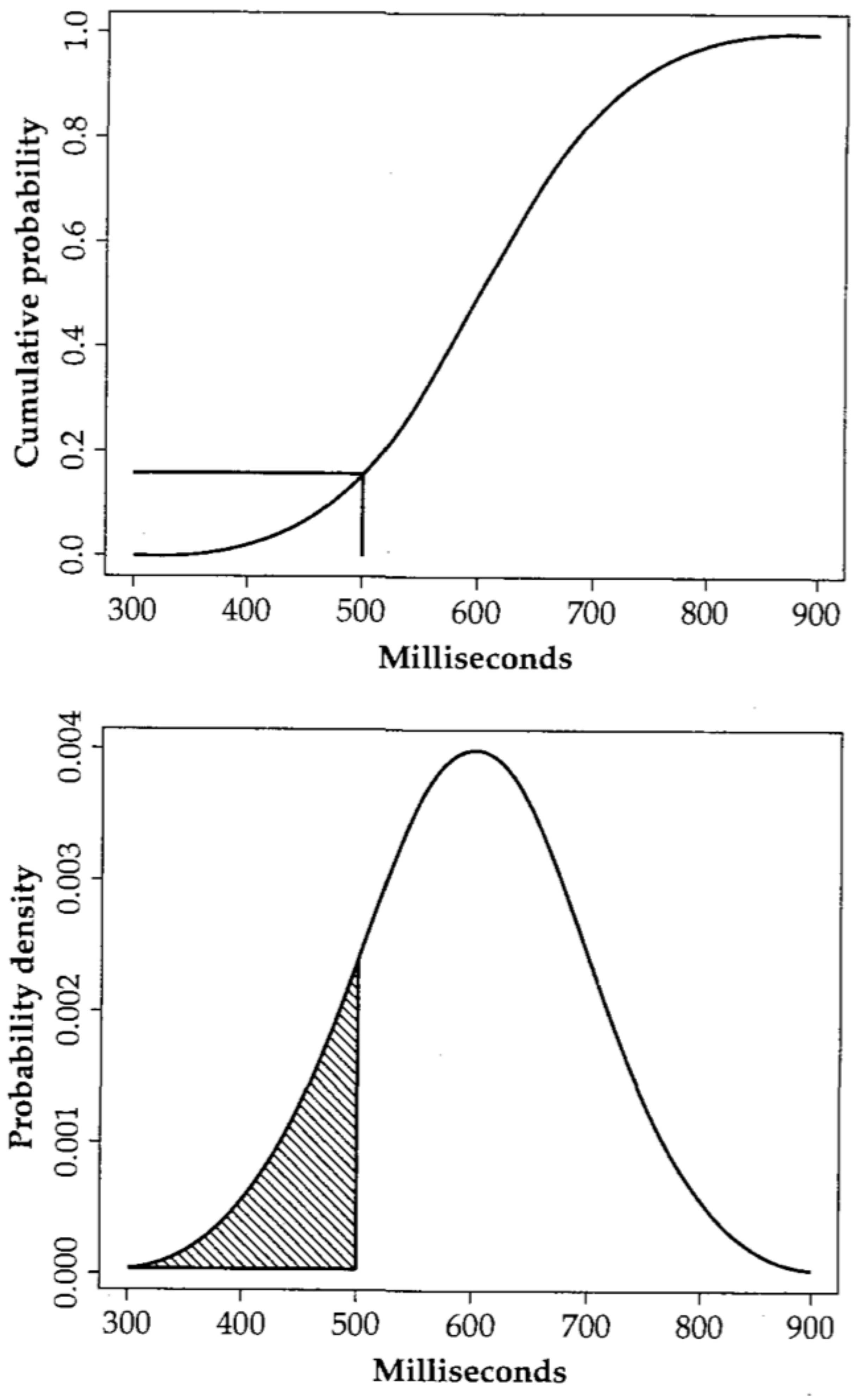


Figure 2.2 The cumulative probability function of the normal curve steadily increases from 0 to 1, while the probability density function (pdf) takes the familiar bell-shaped curve.

Distributions and why they matter

- Consider flipping in coin
  - What kind of distribution do you have?
  - What kind of distribution do you expect?
- Words in a text? What can we say about their probability of occurrence?

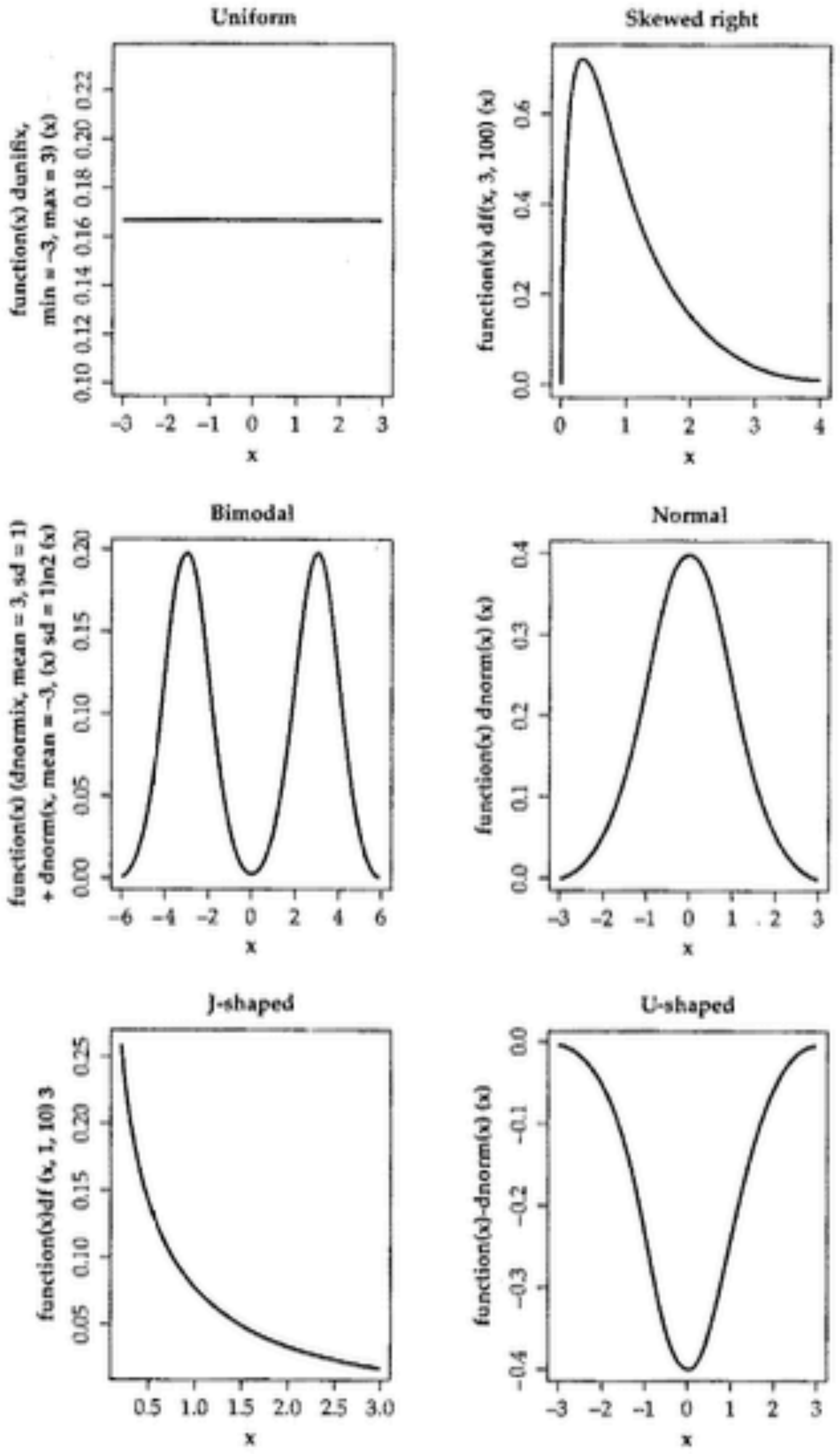


Figure 1.6 Types of probability distributions.

Fisher’s exact test: Lady tasting tea

- The experiment provides a subject with eight randomly ordered cups of tea – four prepared by pouring milk and then tea, four by pouring tea and then milk.
- The subject attempts to select the four cups prepared by one method or the other, and may compare cups directly against each other as desired.
- The method employed in the experiment is fully disclosed to the subject.

https://en.wikipedia.org/wiki/Lady_tasting_tea

Fisher’s exact test: Lady tasting tea

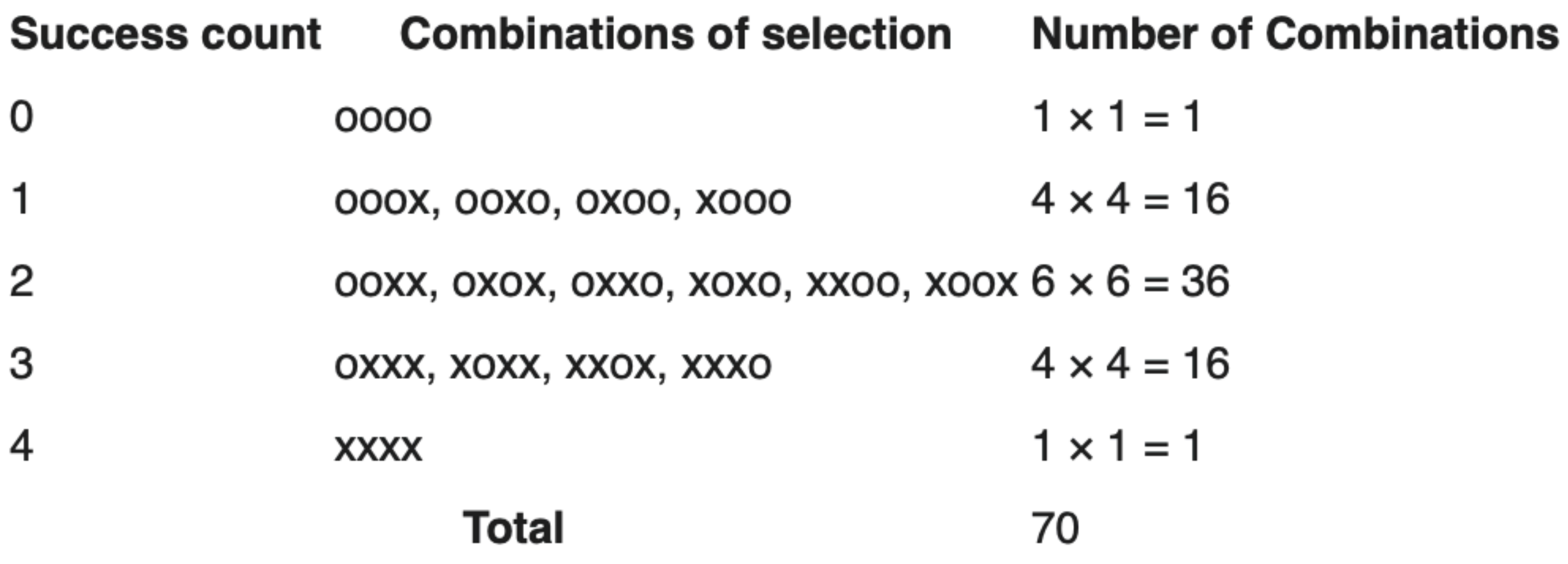
- What is our null hypothesis?
- And our alternative hypothesis?
- What is our alpha?
- What is our test statistic?

https://en.wikipedia.org/wiki/Lady_tasting_tea

Fisher’s exact test: Lady tasting tea

- The test statistic is a simple count of the number of successful attempts to select the four cups prepared by a given method.
- The distribution of possible numbers of successes, assuming the null hypothesis is true, can be computed using the number of combinations.

$$\binom{8}{4} = \frac{8!}{4!(8-4)!} = 70$$



Success count	Combinations of selection	Number of Combinations
0	oooo	1 × 1 = 1
1	ooox, ooxo, oxoo, xooo	4 × 4 = 16
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	6 × 6 = 36
3	oxxx, xoxx, xxox, xxxo	4 × 4 = 16
4	xxxx	1 × 1 = 1
	Total	70

https://en.wikipedia.org/wiki/Lady_tasting_tea

Fisher’s exact test: Lady tasting tea

- The critical region for rejection of the null of no ability to distinguish was the single case of 4 successes of 4 possible, based on the conventional probability criterion < 5%.
- This is the critical region because under the null of no ability to distinguish, 4 successes has 1 chance out of 70 (≈ 1.4% < 5%) of occurring, whereas at least 3 of 4 successes has a probability of (16+1)/70 (≈ 24.3% > 5%).
- Thus, if and only if the lady properly categorized all 8 cups was Fisher willing to reject the null hypothesis – effectively acknowledging the lady's ability at a 1.4% significance level (but without quantifying her ability).
- Fisher later discussed the benefits of more trials and repeated tests.

https://en.wikipedia.org/wiki/Lady_tasting_tea

Quantifying and studying variability

1. You compute the effect you observe in your data (e.g., a frequency distribution, a difference in means, a correlation),
2. You compute the so-called probability of error p to obtain the (summed/combined) probability of the observed effect and every other result that deviates from H0 even more when H0 is true, and
3. You compare p to a significance level (usually 5 percent, i.e., 0.05) and, if p is smaller than the significance level, you reject H0 (because it is not compatible enough with the data to stick to it) and accept H1.

Frequency table analytics - Chi-square and Fisher’s exact

- Chi-square test (X^2)
  - In the days before computers were readily available, people analyzed contingency tables by hand, or using a calculator, using chi-square tests. This test works by computing the expected values for each cell if the relative risk (or odds' ratio) were 1.0. It then combines the discrepancies between observed and expected values into a chi-square statistic from which a P value is computed.
  - A chi square statistic is used to investigate whether distributions of categorical variables differ from one another.
  - Categorical variable yield data in the categories and numerical variables yield data in numerical form
  - Chi Square statistic compares the tallies or counts of categorical responses between two (or more) independent groups.

Frequency table analytics - Chi-square and Fisher’s exact

- Statistical significance test used in the analysis of contingency tables.
- Always gives an exact P value and works fine with small sample sizes. Fisher's test (unlike chi-square) is very hard to calculate by hand, but is easy to compute with a computer. Most statistical books advise using it instead of chi-square test.
- Gives an exactly correct answer no matter what sample size you use

	Postpositions	**Prepositions**	**Total**
Object Verb	***471***	***14***	***485***
Verb Object	***42***	***450***	***492***
Total	***513***	***464***	***977***

- At first glance, it seems that we have more postpositions with OV languages and more prepositions with VO languages, but could it be merely an illusion?
  - After all, we have less languages with prepositions on our data!
  - It is important to look at the overall picture and not just compare numbers in cells
  - E.g. 5 out of 10 is the same ratio as 50/100

	Postpositions	Prepositions	Total
Object Verb	***?***	***?***	***485***
Verb Object	***?***	***?***	***492***
Total	***513***	***464***	***977***

- How would the table look like if all the data points were distributed absolutely fairly and equally between the cells (= independent factors)?

  - We have 977 languages in total, among those, 513 have postpositions and 485 have OV word order

    Proportion of languages with postpositions: 513/977 ≈ 0.525 (52.5%)
    Proportion of OV languages: 485/977 ≈ 0.496 (49.6%)

    If these factors are independent, we expect the proportion of OV languages with postpositions to be 0.496*0.525 ≈ 0.26 (26%)

    With 977 languages this means: we expect to see around 254 such languages!

- Observed:

	Postpositions	**Prepositions**	**Total**
Object Verb	***471***	***14***	***485***
Verb Object	***42***	***450***	***492***
Total	***513***	***464***	***977***

- Expected (under a fair and square distribution)

	Postpositions	**Prepositions**	**Total**
Object Verb	***254.6***	***230.3***	***485***
Verb Object	***258.3***	***233.6***	***492***
Total	***513***	***464***	***977***

- There is a clear difference between expected and observed, so the world is not as fair as we assumed!

	Postpositions	**Prepositions**	**Total**
Object Verb	***471***	***14***	***485***
Verb Object	***42***	***450***	***492***
Total	***513***	***464***	***977***

- Fisher Exact Test: what is the probability of observing exactly such data distribution by chance, if we fix the margin counts (row and column sums)?

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\;(c+d)!\;(a+c)!\;(b+}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

(don’t worry, you don’t need to remember it)

- Now, we use the formula to compute the probability of getting this data or *even more extreme data* (e.g. 480 Postposition/OV vs. 5 Preposition/VO), by summing the results for respective tables

	Postpositions	**Prepositions**	**Total**
Object Verb	***471***	***14***	***485***
Verb Object	***42***	***450***	***492***
Total	***513***	***464***	***977***

- The total probability of getting such or more extreme data purely by chance (p-value) is less then

  0.00000000000000022

  … this is not much!

  Actually, the chance to win in lotto (6+1) is 32505108 times higher!

- Conclusion: it is *very* unlikely that the observe distribution is purely coincidental. More likely explanation: order of object and verb and prepositions/pospositions in languages are actually correlated!

- Fisher Exact Test in R:

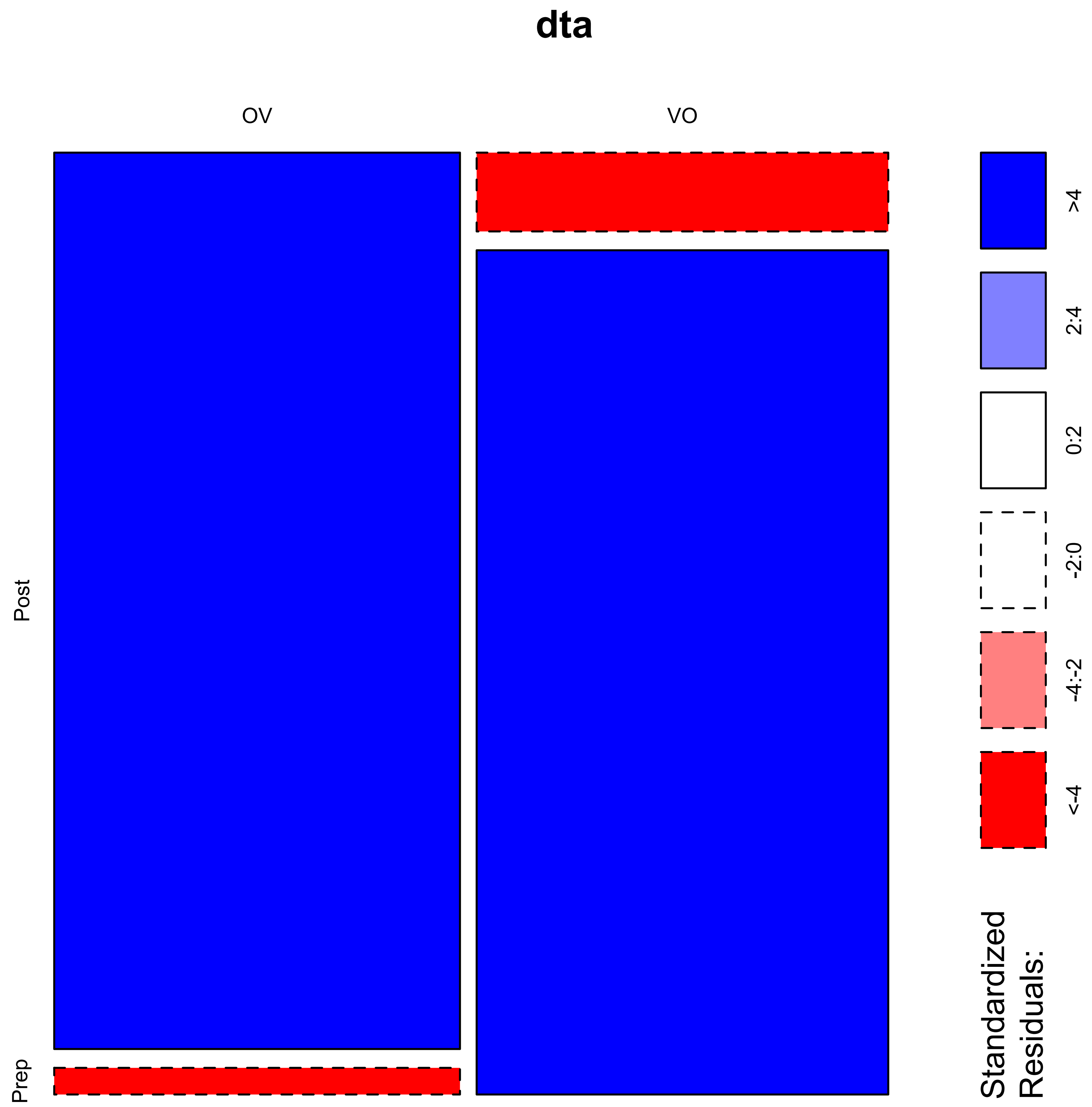
```
> dta <- matrix(c(471, 14, 42, 450), ncol=2, byrow=T)
> row.names(dta) <- c('OV', 'VO')
> col.names(dta) <- c('Post', 'Prep')
> dta
   Post Prep
OV  471   14
VO   42  450
> fisher.test(dta)

     Fisher's Exact Test for Count Data

data:  dta
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 189.0544 730.4164
sample estimates:
odds ratio
  354.6951
```

- And a nice visual representation

```
> library(vcd)
> mosaicplot(dta, shade=T)
```



- Fisher Exact Test is called an exact test because it uses exact computations of probabilities
- Drawback: it is only applicable to 2x2 count data tables. Sometimes we want to look at bigger tables
- The Chi-squared (χ^2) test does basically the same as the Fisher Test, but uses approximate calculations

```
> chisq.test(dta)

	Pearson's Chi-squared test with Yates' continuity correction

data:  dta
X-squared = 764.8854, df = 1, p-value < 2.2e-16
```

the more data (larger counts in all cells) you have, the more reliable is the approximation

- Rule of Thumb when working with count tables
  - If its a 2x2 table, use Fisher Exact Test
  - If it has more rows and/or columns, use χ^2-test
  - The mosaicplot() R function is a nice way to visualize the data (it will also show you which cells are more/less frequent than expected)

What the F?

- Hockett's (1985):
  - "I propose that f-sounds are a relatively recent innovation in human history that up to a few thousand years ago either there were none anywhere in the world, or, if there were, that they were as rare an unusual as clicks are today. And I suggest that the relation of agriculture to f-sounds is causal, the line of causality passing through the teeth."
  - Classic human dentition - incisors meet edge-to-edge (edge bite)
  - Not the norm in agricultural populations - scissors bite (upper incisors slide down neatly in front of the lower, i.e. overbite)
  - Agricultural led to scissors bite (partly or primarily) through increase use of cereal grains in the diet

Hockett, Charles F. "Distinguished lecture: F." American Anthropologist 87.2 (1985): 263-281.

What the F?

- “The current conformation of the vocal tract bears clear marks of having been selectively reshaped by the exigencies of vocal-auditory communications (Lieberman 1975). One of these is the small size of the opening at the lips, as compared with the enormously wide mouths of chimpanzees and gorillas. The small aperture makes the oral cavity a much more effective resonance chamber. Another is the lowering of the larynx, so that the mouth and pharynx form two tubes meeting approximately at right angles; this structure is essential in the production of vowel color, which in turns does much of the work of keeping speech sounds apart. As these structures were evolving, no doubt both individuals and whole populations lost out because their vocal-auditory communication was not efficient enough for them to compete with their better-equipped neighbors. But by 10,000 years ago all that was already ancient history. Language had become what it is today.”

Hockett, Charles F. "Distinguished lecture: F." American Anthropologist 87.2 (1985): 263-281.

What the F?

- Distribution of /f/ is driven mainly through areal contact
  - Changing methods of food getting have somewhat altered the shape of the apparatus we use in talking
  - To a large extent areal contact with driven by agriculture
  - Agricultural societies **had** and **did not have** labiodentals
  - Hunter-gatherer (HG) societies **do not** have
    - Simply not some correlation claim: HG languages have/don’t have X
    - Bickel & Nichols study on HGs

What the F?

- Using Fisher’s exact test on PHOIBLE inventories

	has /f/	no /f/
agriculturalist	A	B
hunter-gatherer	C	D

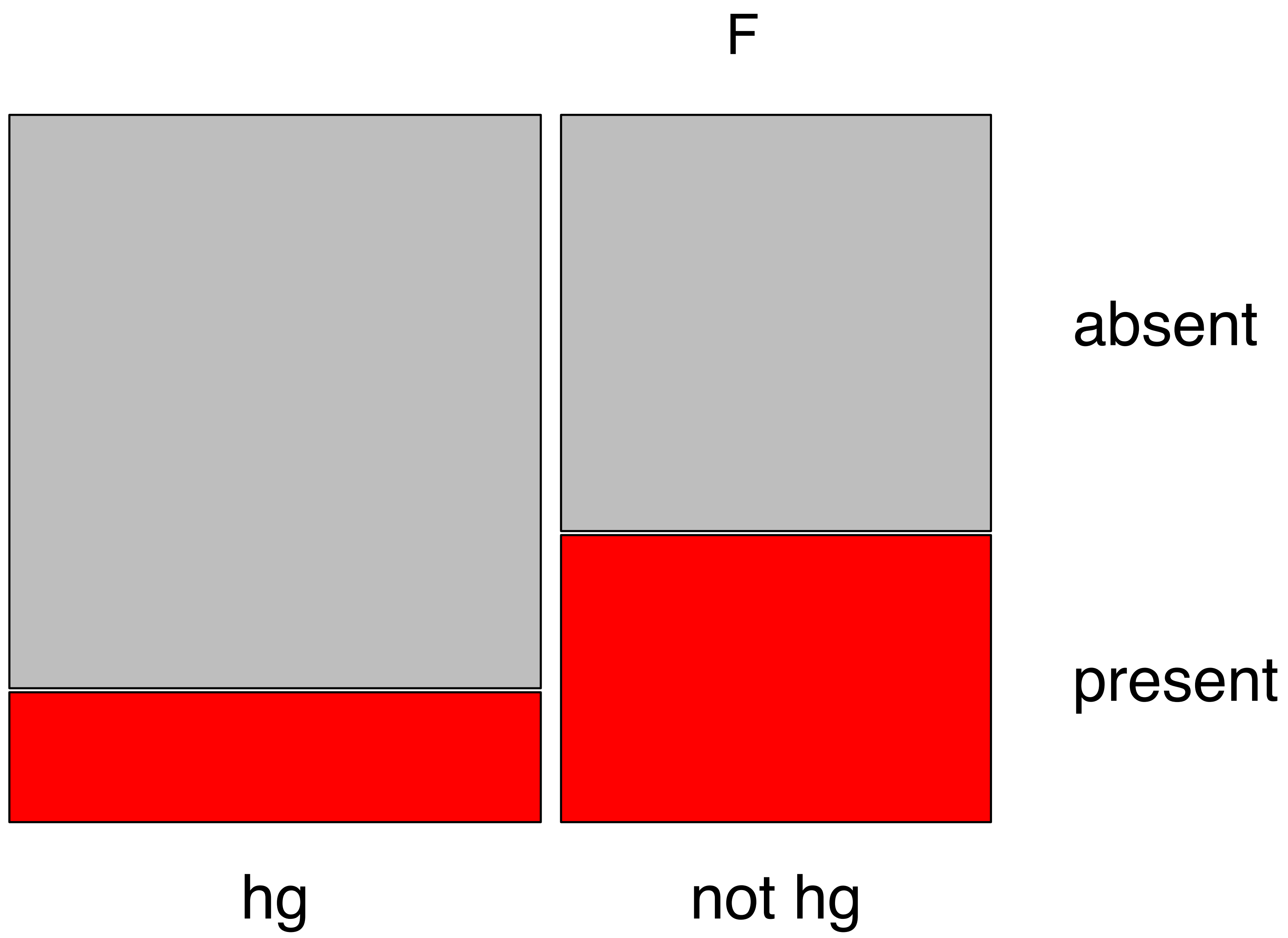
- p-value < 2.2e-16
- 95 percent confidence interval: 0.07 0.76
- odds ratio: 0.115
- alternative hypothesis: true odds ratio is not equal to 1
- not controlled for genealogical sampling!

What the F?

- Fisher's Exact Test for Counts of **Estimated Family Biases**

	absent	present
hg	60	14
not hg	35	24

- p-value = 0.007
- alternative hypothesis:
  true odds ratio is not equal to 1
- 95 percent confidence interval:
  1.261 6.960
- odds ratio: 2.914



- Fisher Exact Test is called an exact test because it uses exact computations of probabilities

- Drawback: it is only applicable to 2x2 count data tables. Sometimes we want to look at bigger tables

- The Chi-squared (χ^2) test does basically the same as the Fisher Test, but uses approximate calculations

```
> chisq.test(dta)

	Pearson's Chi-squared test with Yates' continuity correction

data:  dta
X-squared = 764.8854, df = 1, p-value < 2.2e-16
```

the more data (larger counts in all cells) you have, the more reliable is the approximation

Chi-square

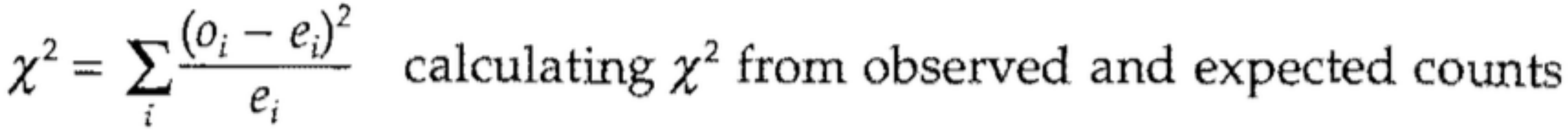
- Compare the observed frequency of occurrence of an event with its theoretically expected frequency of occurrence
  - number of native Spanish and native English speakers in German classes on campus should be equal because there are about as many L1 Spanish vs English speakers in Miami (simplification)
  - in a class of 20, we expect 10 L1 Spanish speakers and 10 L1 English speakers
- If the difference between observed and expected frequency is much greater than chance you might begin to wonder what is going on… perhaps an explanation is called for?

Chi-square

- To calculate X2 from observed and expected frequencies you sum over all of the cells in a contingency table the squared difference of the observed count (0 = say 5 Spanish speakers in the class) minus the expected count (e = 10 Spanish speakers) divided by the expected count.

- For the case in which we have 5 Spanish speakers and 15 English speakers in a class of 20, and we expect 10 Spanish speakers and 10 English speakers, the X2 value that tests the accuracy of our expectation is

  - X2 = $(5-10)^2/10 + (15-10)^2/10$ = 2.5+2.5 = 5



$$\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$$ calculating χ^2 from observed and expected counts

Chi-square

- To determine the correctness of the assumption that we used in deriving the expected values, we compare the calculated value of X2 with a critical value of X2
- If the calculated value of X2 is larger than the critical value then the assumption that gives us the expected values is false
- Because the distribution is different for different degrees of freedom you will need to identify the degrees of freedom for your test
- In the case of L1 balance in a because we have two expected frequencies (one for the number of Spanish speakers and one for the number of English speakers in a class) there is 1 degree of freedom
  - What could be a second degree of freedom?
- The probability of getting a X2 value of 5 when we have only 1 degree of freedom is only p=0.025, so the assumption that Spanish and English speakers are equally likely to take statistics is probably not true (97 times in a 100 cases) when there are only 5 Spanish speakers in a class of 20

Chi-square

- **Goodness-of-fit test**: is a way of determining whether a set of categorical data came from a claimed discrete distribution or not. The null hypothesis is that they did and the alternate hypothesis is that they didn't. It answers the question: are the frequencies I observe for my categorical variable consistent with my theory?
  - The goodness-of-fit test is used if you have two or more categories.
- **Test of independence**: is a way of determining whether two categorical variables are associated with one another in the population, like race and smoking, or education level and political affiliation.

http://inspire.stat.ucla.edu/unit_13/

Chi-square

- Example of **goodness-of-fit**:
    - We might compare the proportion of M&M's of each color in a given bag of M&M's to the proportion of M&M's of each color that Mars (the manufacturer) claims to produce. In this example there is only one variable, M&M's. M&M's can be divided into many many categories like Red, Yellow, Green, Blue, and Brown, however there is still only one variable… M&M's. Hungry?
- Example of **Test of Independence**:
    - To continue with the M&M's example, we might investigate whether purchasers of a bag of M&M's eat certain colors of M&M's first. Here there are two variables: (1) M&M's (2) The order based on color that an M&M bag holder/purchaser eats the candies.

http://phdmonkey.com/2011/07/23/chi-square-goodness-of-fit-vs-test-of-independence/