Biostatistics 1BIO43

Karl Oskar Ekvall

Fall 2021

Karolinska Institutet

Organization

Organization

- Director: Matteo Bottai (matteo.bottai@ki.se)
- Teacher: Karl Oskar Ekvall (karl.oskar.ekvall@ki.se)
- Administrator: Kamilla Sagrelius (kamilla.sagrelius@ki.se)

Organization

- We will often not need all 3 + 3 hours of lectures and lab.
- I will nevertheless typically be available for all of that time, except:
  - Sep 8 from 2 to 3 pm
  - Sep 9 after 3 pm
- Aug 31 from 11 am to 12 pm and 1 pm to 2 pm will be a recap of first day

Organization

Slides

You can find these slides at `https://koekvall.github.io/biostat.html`

Please remind me if the slides are not up to date at the end of the day.

Zoom

Please use your KI account and your real name.

Content

Concepts and tools to understand scientific literature and perform statistical analyses.

- Descriptive and exploratory statistics
- Probability
- Estimation and Inference

Suggestion

We will cover a lot of material in a short period of time. To follow:

- Make good use of the lectures by asking questions
- Know that you are not expected to understand everything at once
- Go through slides after class and make sure everything is clear

First day

- Random variables and realizations
- Understanding data using descriptive statistics and plots
- Introduction to software (R and RStudio)

Random variables, realizations, and data

Randomness and sampling

Data are often outcomes of random experiments or sampling.

If we repeat an experiment, we usually get different data.

Example

Five patients are selected at random to receive a new drug.

The effectiveness of the drug typically depends on who is in the treatment group.

Random variables

A random variable is a measurement of the outcome of an experiment yet to be performed (often numerical).

Example

Let X be the number of minutes before 9 am that the first student joins Zoom tomorrow.

Non-example

The number of minutes before 9 am that the first student joined Zoom today.

Realizations

A realization or observed value is the particular value a random variable took when the experiment was performed.

It is common to use capital letters for random variables (X) and lower case letters for realizations (x).

Example

If tomorrow it turns out that the first student joins Zoom at 8.55, the realized or observed value of X is $x = 5$.

Example

The first student joined Zoom y minutes before 9 am today.

Data

We typically assume data are realizations of random variables.

Example

Select 10 students in the class at random and measure their heights in centimeters.

Let X_i denote the height of the ith randomly selected person, $i = 1, \ldots, 10$.

After having performed the experiment, our data, or sample, may consist of the realizations

$$\{x_1, x_2, \ldots, x_{10}\} = \{165, 181, \ldots, 169\}.$$

Descriptive statistics

Even a moderately large dataset is difficult to understand by just looking at.

Descriptive statistics can help.

Definition: A statistic is a function of the data.

Intuitively: A statistic is something you can compute if you are given data.

A descriptive statistic is a statistic intended to tell you something useful about the data.

Example

The sample mean or sample average is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + \cdots + x_n).$$

The sample mean is a statistic because you can compute it if I tell you what $x_1, \ldots, x_n$ are.

Descriptive statistics

More generally, we will consider descriptive statistics that quantify:

- Central tendency / location (e.g. sample mean)
- Dispersion / variability
- Symmetry or asymmetry
- Association

Measures of central tendency and location

In R, you can calculate the sample mean of any vector of observations easily:

```
heights <- c(165, 181, 177, 189, 185, 155, 170, 179, 172, 169)
mean(heights)
```

```
## [1] 174.2
```

Median

The middle number of the sorted data if n is odd, and the average of the two middle numbers if n is even.

```
sort(heights)
```

```
##  [1] 155 165 169 170 172 177 179 181 185 189
```

```
median(heights)
```

```
## [1] 174.5
```

Measures of central tendency and location

According to Credit Suisse's global wealth report:

Average wealth of an adult in Sweden in 2019 was 256,000 USD.

Median was 42,000 USD.

Depending on situation, one or the other may be a more useful measure.

Measures of central tendency and location

The mean is sensitive to outliers, the median is not.

```
income <- c(50, 30, 60, 55, 75, 300) # 1000 USD / year
mean(income)
```

```
## [1] 95
```

```
median(income)
```

```
## [1] 57.5
```

Other location measures

Sample quantiles

- Cut points dividing sorted sample into equally sized subsets

Examples

Quartiles divide the sample into four equally sized subsets

```
  sort(heights)
```

```
##  [1] 155 165 169 170 172 177 179 181 185 189
```

```
  quantile(heights, c(0.25, 0.5, 0.75))
```

```
##    25%    50%    75%
## 169.25 174.50 180.50
```

Other location measures

Percentiles divide the sample into 100 equally sized subsets

```
  quantile(heights, 0.1) # 10th percentile
```

```
## 10%
## 164
```

Measures of dispersion

Sample variance and standard deviation

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \text{and} \quad s = \sqrt{s^2}$$

- Loosely speaking, the sample standard deviation tells you how much a typical observation differs from the sample mean.

However, in general, it does not equal the **mean absolute deviation**:

$$s \neq \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|.$$

Measures of dispersion

Example

```
  heights - mean(heights)
```

```
##  [1]  -9.2   6.8   2.8  14.8  10.8 -19.2  -4.2   4.8  -2.2  -5.2
```

```
  sum((heights - mean(heights))^2) / 9 # Sample var
```

```
## [1] 101.7333
```

```
  sqrt(sum((heights - mean(heights))^2) / 9) # Sample sd
```

```
## [1] 10.08629
```

Measures of dispersion

Ranges

The range is $\max_i x_i - \min_i x_i$ and the inter-quartile range (IQR) is the difference between the third and first quartile.

```
max(heights) - min(heights) # Range
```

```
## [1] 34
```

```
IQR(heights)
```

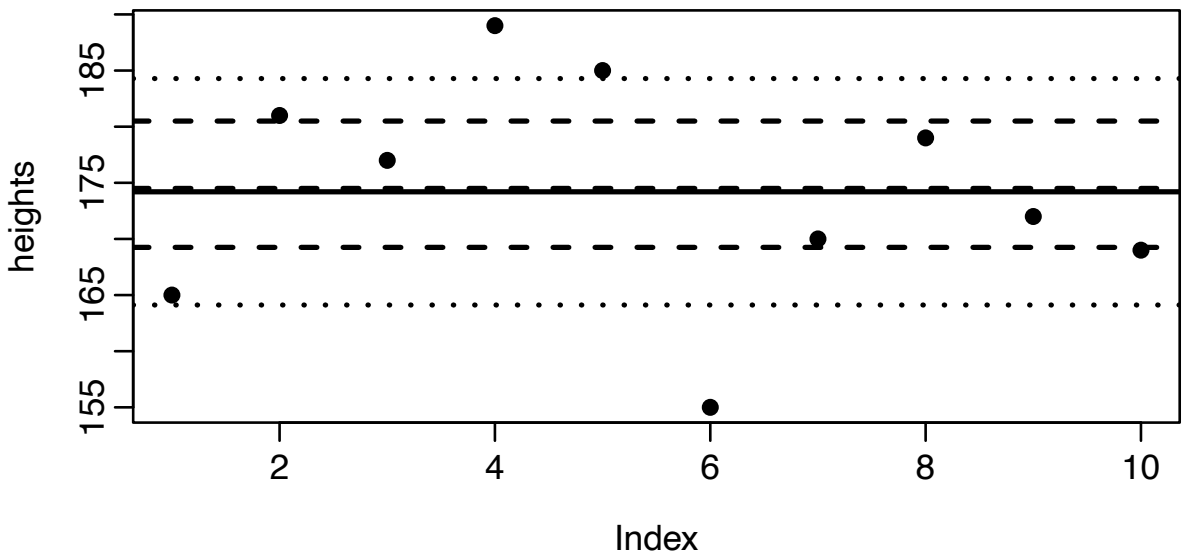
```
## [1] 11.25
```

Plotting summary statistics

To practice, we can make a plot with lines:

```
plot(heights)
abline(h = mean(heights), lwd = 2)
abline(h = median(heights), lty = 2, lwd = 2)
abline(h = quantile(heights, 0.25), lty = 2, lwd = 2)
abline(h = quantile(heights, 0.75), lty = 2, lwd = 2)
abline(h = mean(heights) + sd(heights), lty = 3, lwd = 2)
abline(h = mean(heights) - sd(heights), lty = 3, lwd = 2)
```

Plotting summary statistics



Summary statistics for two variables

Suppose our data is a sample of n pairs:

$$\{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

As before, we can summarize properties of the y_i and x_i, e.g.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

But how can we quantify the association between them?

Covariance and correlation

Sample covariance

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Intuition: Two variables have positive covariance if they tend to be larger (or smaller) than their respective means at the same time.

Notice $s_{xx} = s_x^2$ and $s_{xy} = s_{yx}$.

Sample correlation

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y},$$

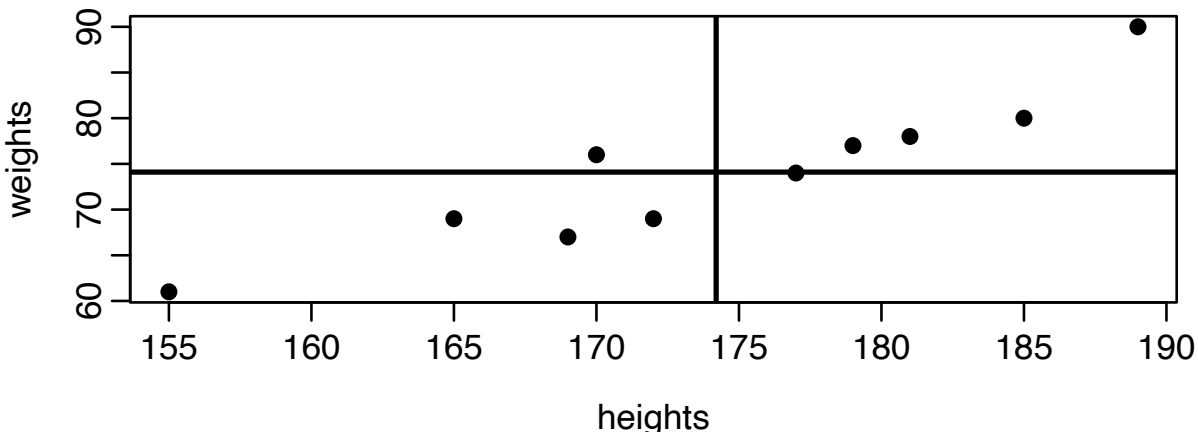
- unit-free measure of association
- same sign as the sample covariance
- always between -1 and 1
- symmetric in y and x: $\rho_{xy} = \rho_{yx}$

Association

Example

$$\text{weights} = \{y_1, \ldots, y_{10}\} = \{69, \ldots, 67\}$$

```
weights <- c(69, 78, 74, 90, 80, 61, 76, 77, 69, 67)
plot(heights, weights)
abline(v = mean(heights), lwd = 2)
abline(h = mean(weights), lwd = 2)
```



Sample covariance

Example

The plot indicates a positive relationship between the variables

Their sample covariance is

```
  weights - mean(weights)
```

```
##  [1]  -5.1   3.9  -0.1  15.9   5.9 -13.1   1.9   2.9  -5.1  -7.1
```

```
  heights - mean(heights)
```

```
##  [1]  -9.2   6.8   2.8  14.8  10.8 -19.2  -4.2   4.8  -2.2  -5.2
```

```
  sum((weights - mean(weights)) *  (heights - mean(heights))) / 9
```

```
## [1] 75.31111
```

Sample correlation

Example

```
cov(heights, weights) / (sd(heights) * sd(weights))
```

```
## [1] 0.9230557
```

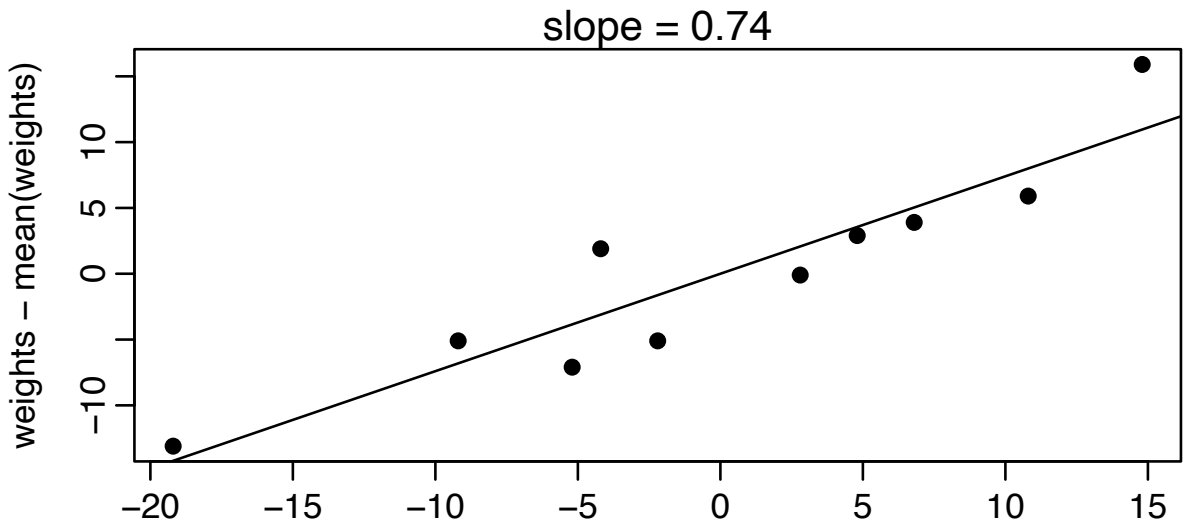
```
cor(weights, heights)
```

```
## [1] 0.9230557
```

Since we know $-1 \leq \rho_{xy} \leq 1$, that $\rho_{xy} = 0.92$ indicates a strong and positive relationship between height and weight.

Preview: linear regression

```
plot(x = heights - mean(heights), y = weights - mean(weights))
slope <- cor(heights, weights) * sd(weights) / sd(heights)
abline(a = 0, b = slope)
mtext(paste0("slope = ", round(slope, 2)))
```



Cautions

Descriptive statistics are just that—descriptions of your sample.

- we are often interested in characteristics of an underlying population, not a particular sample
- we are often (but not always) interested in causal mechanisms
- if I gain weight, will I become taller?
- there are many potentially interesting relationships between variables not captured by correlation
- correlation is a measure of linear dependence

Anscombe's quartet

Four different datasets, each with $n = 11$ observations of 2 variables

```
data(anscombe)
round(apply(anscombe, 2, mean), 2)
```

```
##  x1  x2  x3  x4  y1  y2  y3  y4
## 9.0 9.0 9.0 9.0 7.5 7.5 7.5 7.5
```

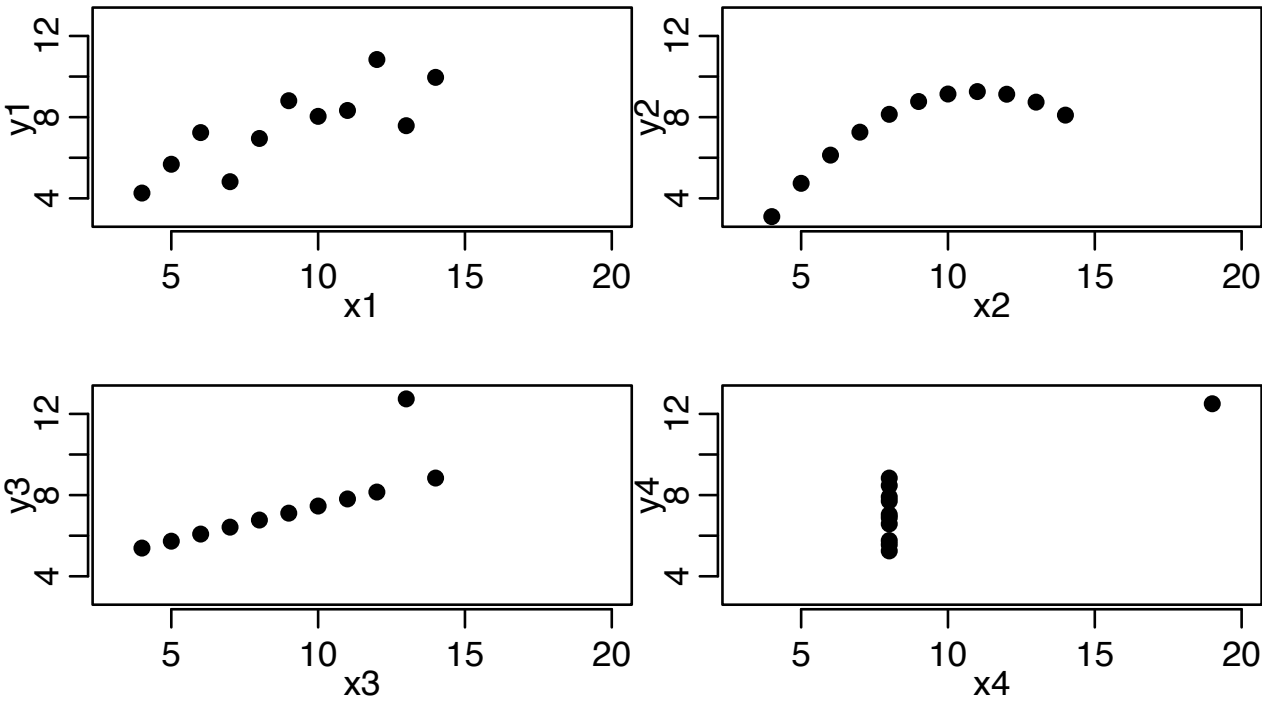
```
round(apply(anscombe, 2, sd), 2)
```

```
##   x1   x2   x3   x4   y1   y2   y3   y4
## 3.32 3.32 3.32 3.32 2.03 2.03 2.03 2.03
```

```
round(c(cor(anscombe$x1, anscombe$y1),
        cor(anscombe$x2, anscombe$y2),
        cor(anscombe$x3, anscombe$y3),
        cor(anscombe$x4, anscombe$y4)), 2)
```

```
## [1] 0.82 0.82 0.82 0.82
```

Anscombe's quartet



Anscombe's quartet

Lessons:

- we should plot the data if possible
- the sample correlation is sometimes an appropriate measure of association (dataset 1) and sometimes not (dataset 4)
- the sample mean is sometimes an appropriate measure of central tendency (dataset 1) and sometimes not (dataset 4)

Imagine someone told you the sample correlation between the dose of a drug and the number of days until a patient is symptom free is -0.82.

-you may (obviously) want to do some further investigation before concluding the drug is effective

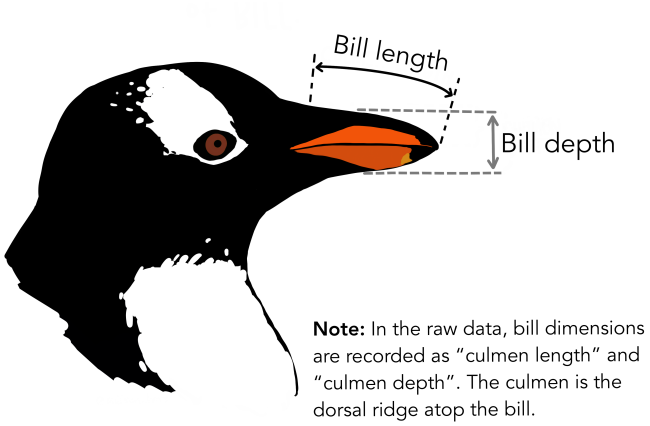
In practice, correlation is often useful but not the whole story.

More on plotting

Let's consider bill lenght and depths in the `penguins` data.

It's at `https://github.com/allisonhorst/palmerpenguins` and artwork is by @allison_horst.

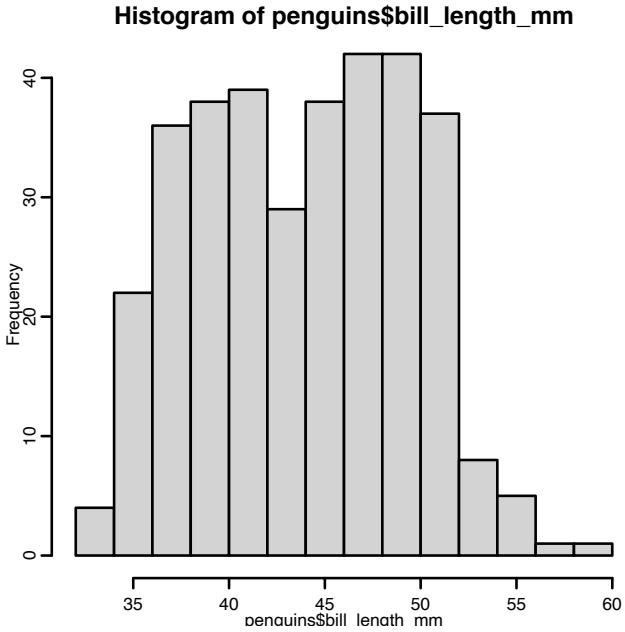
```
# install.packages("palmerpenguins")
library(palmerpenguins)
```

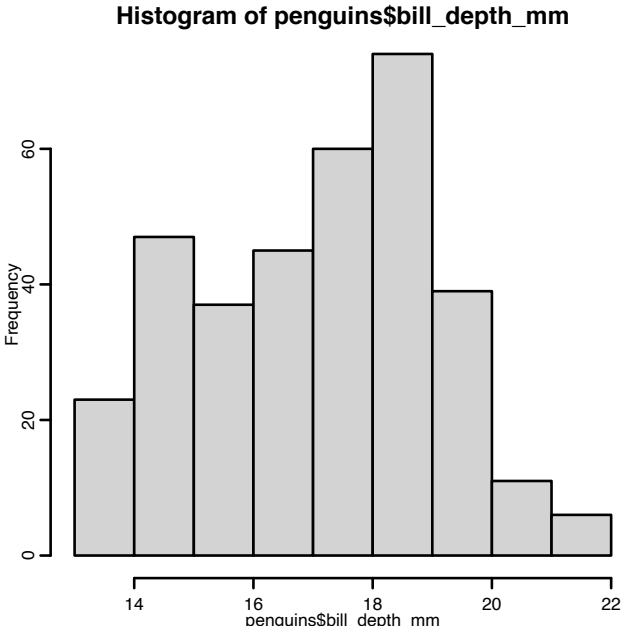


Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

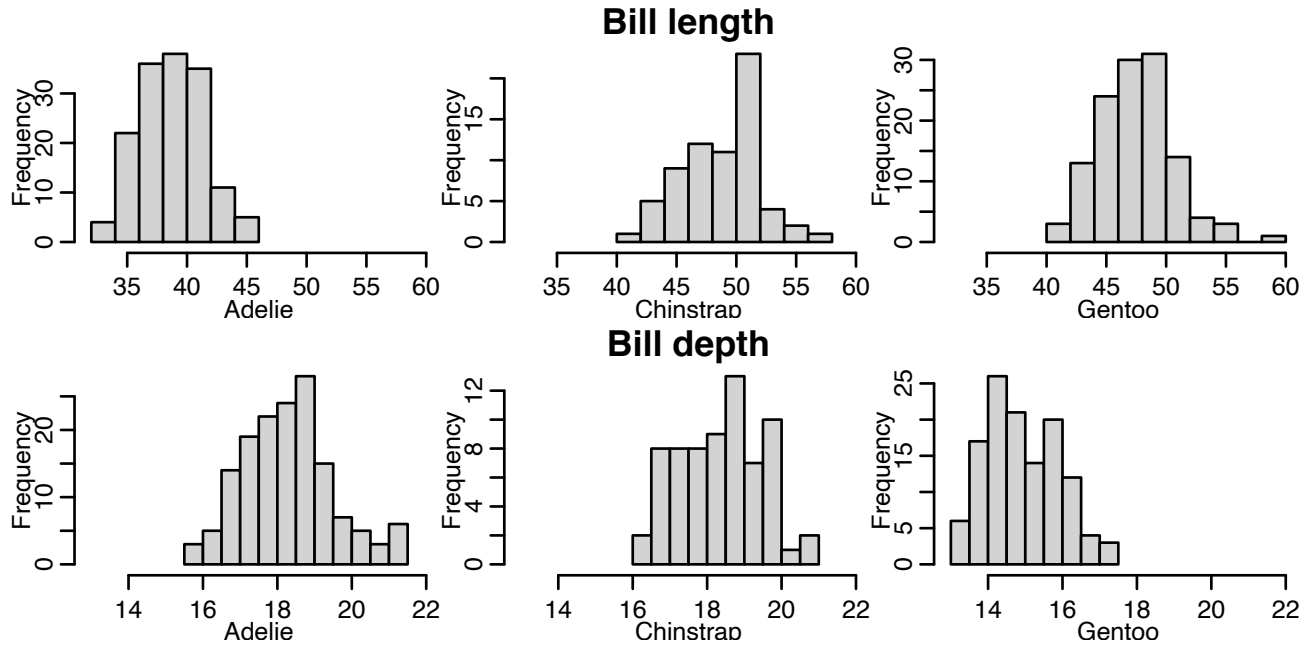
More on plotting

```
hist(penguins$bill_length_mm); hist(penguins$bill_depth_mm)
```





More on plotting

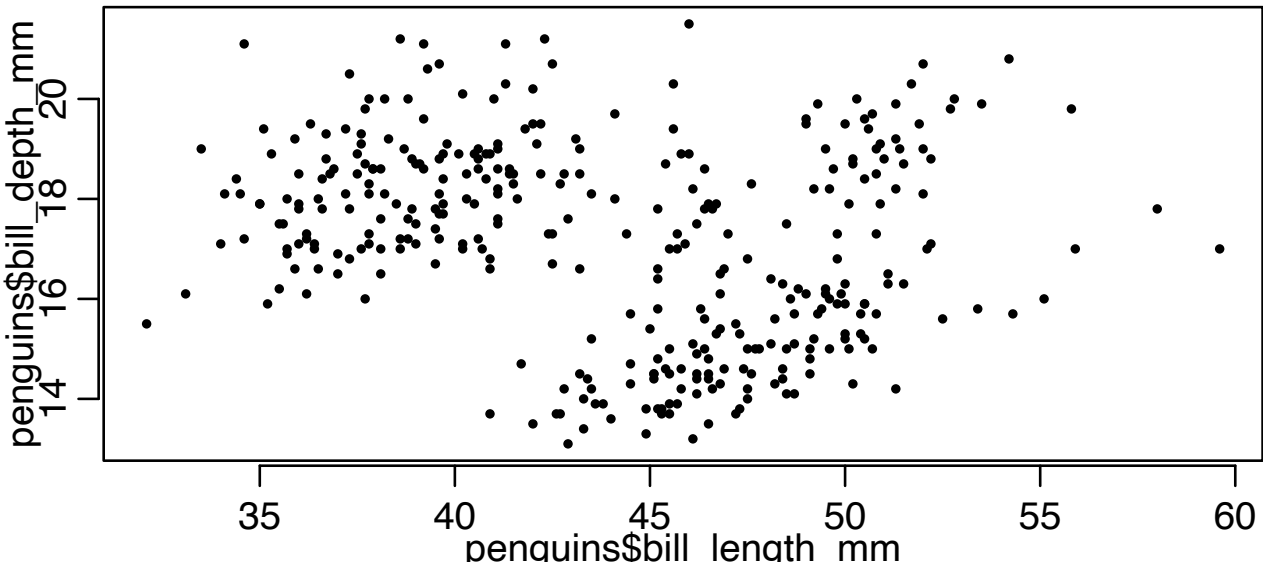


How do the species differ with respect to length and depth?

More on plotting

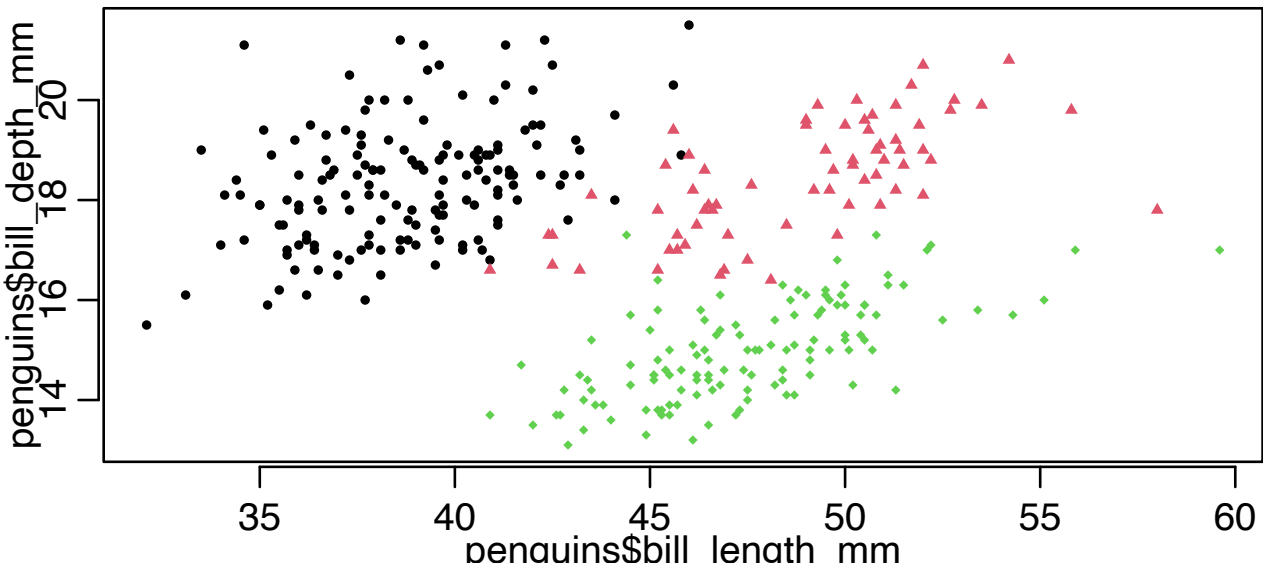
Is there a correlation between length and depth?

```
plot(penguins$bill_length_mm, penguins$bill_depth_mm, cex = 0.5)
```



Simpson's paradox

```
plot(penguins$bill_length_mm, penguins$bill_depth_mm, col = penguins$species,
     pch = c(16, 17, 18)[penguins$species], cex = 0.5)
```



Simpson's paradox

(This code uses `dplyr`; you do not need to learn it)

```
penguins %>% summarize(r = cor(bill_length_mm, bill_depth_mm, use = "complete"))
```

```
## # A tibble: 1 x 1
##        r
##    <dbl>
## 1 -0.235
```

```
penguins %>% group_by(species) %>% summarize(r = cor(bill_length_mm, bill_depth_mm,
                                                     use = "complete"))
```

```
## # A tibble: 3 x 2
##   species       r
##   <fct>     <dbl>
## 1 Adelie    0.391
## 2 Chinstrap 0.654
## 3 Gentoo    0.643
```

Some final remarks on summary statistics and plots

```
str(penguins, vec.len = 1)
```

```
## tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 ...
##  $ flipper_length_mm: int [1:344] 181 186 ...
##  $ body_mass_g      : int [1:344] 3750 3800 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 ...
##  $ year             : int [1:344] 2007 2007 ...
```

Some final remarks on summary statistics and plots

- A factor can be nominal or ordinal
- Nominal: there is no natural ordering (species)
- Ordinal: there is a natural ordering (low / high boody mass)
- A binary variable takes only two values
- A numerical variable takes numbers with the usual meaning (so addition, subtraction, multiplication, etc. makes sense)
    - An integer is a special case of a numerical variable but may obey different rules in a computer

Averages, correlations, and similar measures are mostly meaningful for numerical variables.

Probability

Content

- Events
- Rules of probabilities
- Conditional probability and Bayes' theorem

Sample and population

To do more than descriptive and exploratory statistics, we need to consider how the data may have been generated.

Probability theory lets us use statistics to reason about what a particular sample may say about an underlying population.

Sometimes there is an actual population we are sampling from, sometimes it is a theoretical construct.

Disclaimer

Everything we will cover can be motivated formally using mathematics.

Because this is not a course in mathematics, we will state many things without formal motivation.

Remember to ask if you find anything confusing!

Events

Events

Things to which probabilities can be assigned are called events.

An event is a set, or collection, of (potential) outcomes.

For example: - “it rains at 5 pm tomorrow” is an event (or at least reasonably modeled as such) - If X is the result of rolling a six-sided die, $X \geq 2$ is an event consisting of the outcomes $X = 2, \ldots, X = 6$, each of which is also an event. - “I am 183 cm tall” is not an event, but if we select a person at random, then “they are 183 cm tall” is an event.

Events

We often use letters such as A and B to denote events.

At the end of this section, we will answer the following question:

Let A be the event that a randomly sampled driver is under the influence (of alcohol).

Let B be the event that a randomly sampled driver tests positive.

Suppose that the test we are using is right in 90 % of the cases and that 1 % of all drivers are under the influence.

What is the probability that a randomly selected person is driving under the influence given that they test positive?

Events

We write the probability of the event A as $P(A)$.

For example, it may be that $P(X = 5) = 1/2$ or $P(\text{it rains tomorrow at 5 pm}) = 0.1$.

We have the following:

1. For any event A, $0 \leq P(A) \leq 1$.
2. If two events A and B cannot happen at the same time (they are disjoint), then the probability that (at least) one of them happens is $P(A) + P(B)$.
3. If A contains B; that is, A happens whenever B happens, then $P(A) \geq P(B)$.

Events

We can define new events, for example C can be defined to be "A or B"; that is, C happens if either A happens, B happens, or both A and B happen.

We often write $A \cup B$; you can read this as A union B.

Example

Let A be the event that it rains tomorrow at 5 pm and let B be the event that I am late for class tomorrow. If C is defined as "A or B", or $C = A \cup B$, then C is the event that either it rains tomorrow at 5 pm, or I am late to class tomorrow, or both.

Events

We can also define D to be "A and B"; that is, D happens if and only if both A and B happen.

We often write $A \cap B$, which is called the intersection of A and B.

Example

If A is the event that it rains tomorrow at 5 pm and B the event that I am late for class tomorrow, and if D is "A and B", or $D = A \cap B$, then D is the event that it both rains tomorrow at 5 pm and I am late for class. In particular, D does not happen if only one of A or B happens.

Exercise

Show (that is, use the stated facts about probabilities to argue) that $P(C) \geq P(D)$.

Events

The complement of A is the event "not A".

We often write this as A^C.

The probability of A^C is always $1 - P(A)$.

Motivation

Either A happens or it doesn't, so $P(A \cup A^C) = 1$.

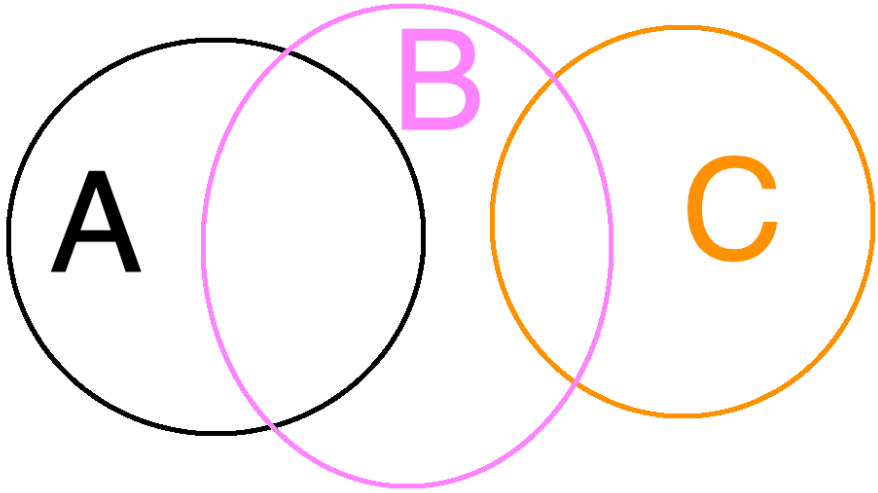
Because A and A^C cannot happen at the same time, one of the rules of probabilities says

$$1 = P(A \cup A^C) = P(A) + P(A^C).$$

Venn diagrams

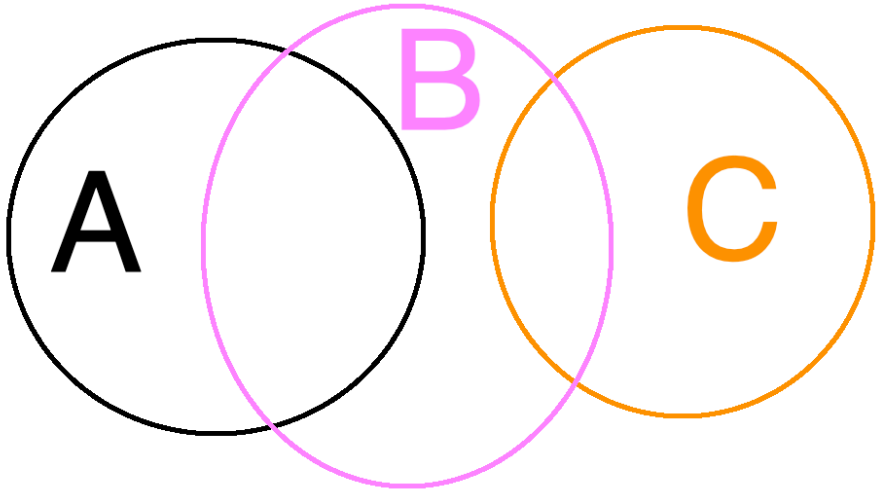
The white region of the slide is the sample space, and is has size 1.

Subsets of the sample space are events, and their size is their probability.



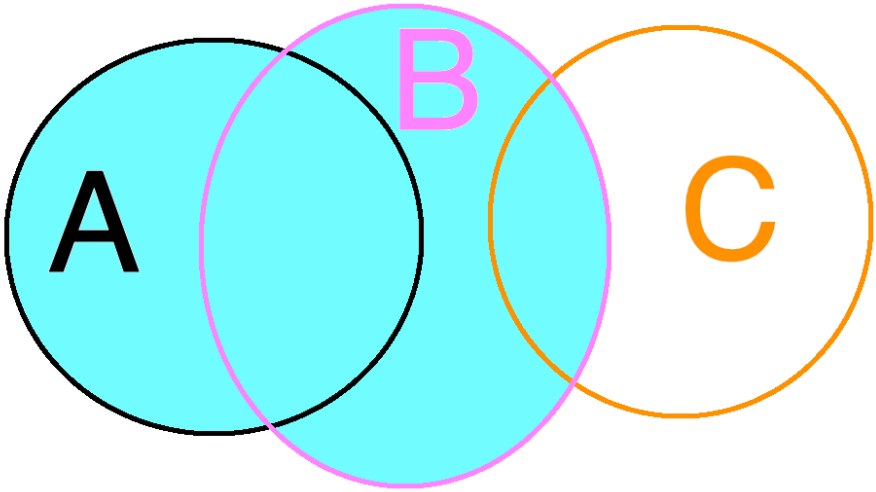
Venn diagrams

Find $A \cup B$, $A \cap B$, $A \cap C$, $A \cup C$, $A \cup B \cup C$, and $(A \cap B) \cup (B \cap C)$



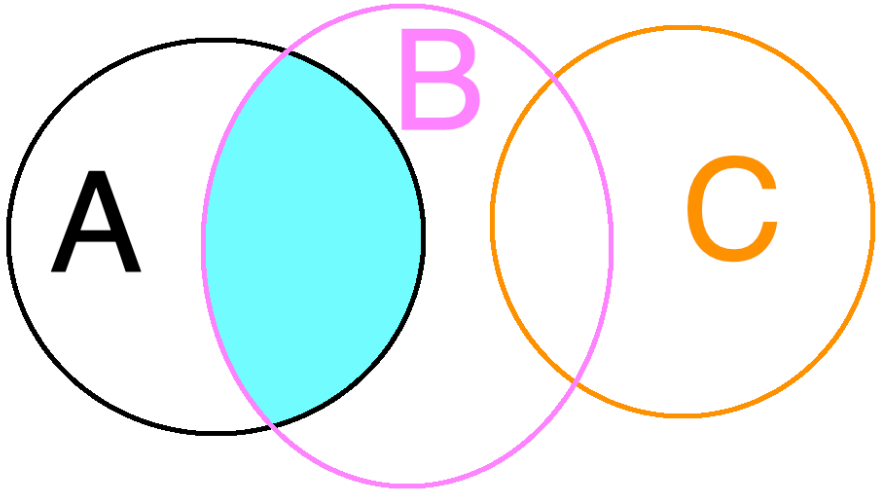
Venn diagrams

$A \cup B$



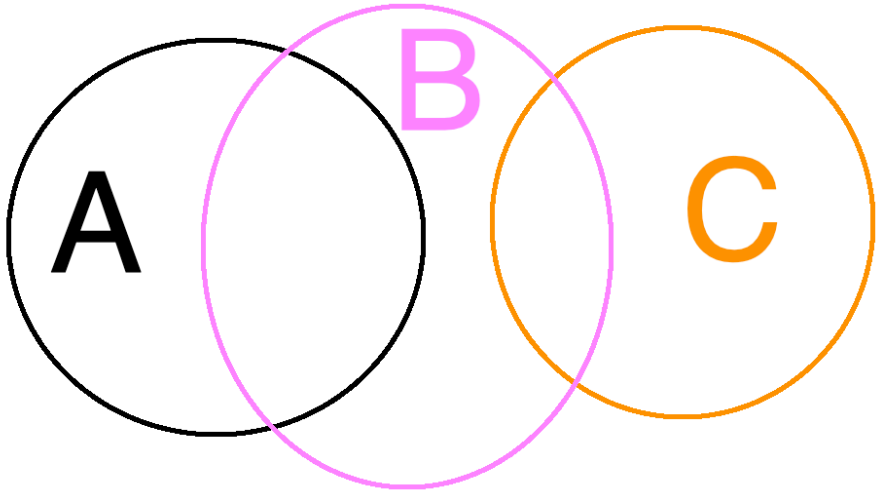
Venn diagrams

$A \cap B$



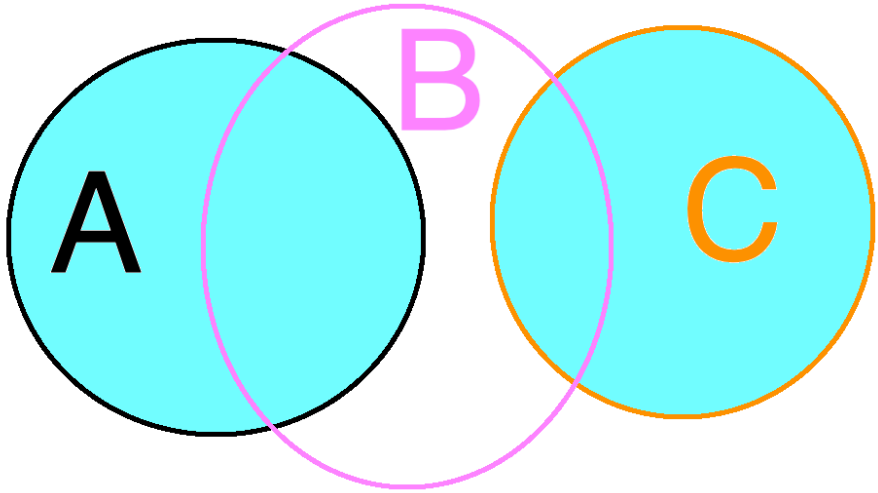
Venn diagrams

$A \cap C$



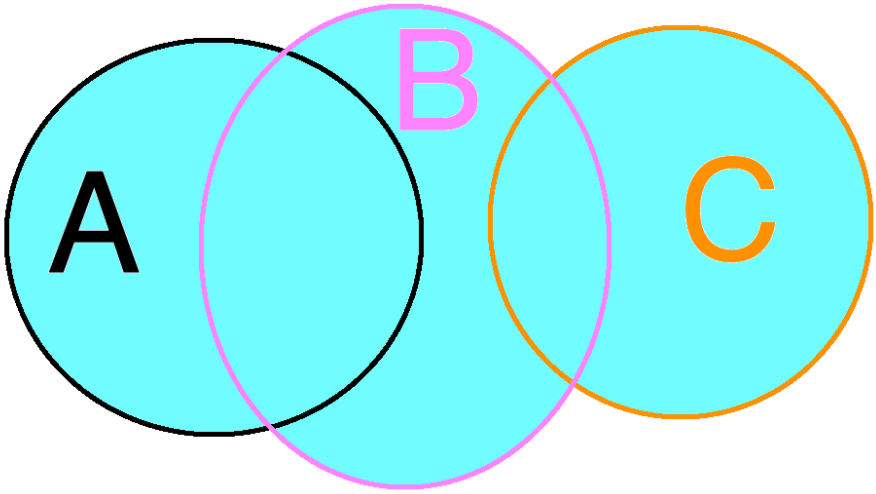
Venn diagrams

$A \cup C$



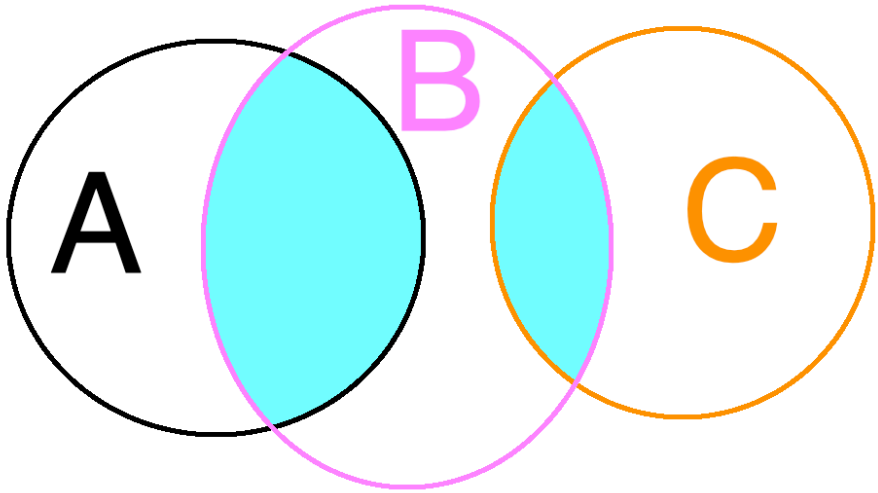
Venn diagrams

$A \cup B \cup C$



Venn diagrams

$$(A \cap B) \cup (B \cap C)$$



Venn diagrams

By using Venn diagrams, we can convince ourselves that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Finally, two events A and B are called independent if

$$P(A \cap B) = P(A)P(B).$$

Example

If I roll a regular six-sided die twice, what is the probability that the first roll is 2 and the second is 4? You may assume the rolls are independent.

Answer: Let X_1 be the result of the first roll and X_2 that of the second. Then, by independence, $P(X_1 = 2 \cap X_2 = 4) = P(X_1 = 2)P(X_2 = 4) = (1/6)(1/6) = 1/36$.

Example

If I roll a regular six-sided die twice, what is the probability that one of the rolls is 2 and the other is 4? You may assume the rolls are independent.

Answer: Let X_1 be the result of the first roll and X_2 that of the second. First, let's figure out which outcomes are in our event. One outcome is that the first roll is 2 and the other is 4. Another is that the first is 4 and the other 2. There are no other outcomes in our event.

Thus, we are looking for

$$P\big[(X_1 = 2 \cap X_2 = 4) \cup (X_1 = 4 \cap X_2 = 2)\big].$$

Because the events in the union are disjoint and the rolls are independent, this is equal to

$$P(X_1 = 2 \cap X_2 = 4) + P(X_1 = 2 \cap X_2 = 4) = 2/36.$$

Conditional probability

Conditional probabilities are like probabilities, but with extra information.

Example

Let A be the event that a die roll is at least 3, and let B be the event that the same roll is even. What is the probability of the roll being at least 3 if we are told it is even? That is, what is the probability of A given B, or

$$P(A \mid B)?$$

Conditional probability

Given that the roll is even, it has to be one of 2, 4, and 6.

Because every outcome is equally likely and two of those three are greater than 3, intuition suggests

$$P(A \mid B) = 2/3.$$

Warning: Intuition is often not reliable!

Conditional probability

The conditional probability can be calculated as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

In this case, intuition was right because

$$\frac{P(A \cap B)}{P(B)} = \frac{P(X = 4 \cup X = 6)}{P(X = 2 \cup X = 4 \cup X = 6)} = \frac{2/6}{3/6} = 2/3$$

Note: if $P(B) = 0$, then $P(A \mid B)$ is not defined.

Conditional probability

Much of applied research is concerned with conditional probabilities.

Example

If I randomly sample a patient to a study, what is the probability that they develop lung cancer given that they are a smoker?

If this conditional probability is significantly greater than the probability that they develop lung cancer given that they are not a smoker, then this can be an indication that smoking increases the risk of developing lung cancer.

Conditional probability

Recall that A and B are independent if $P(A \cap B) = P(A)P(B)$.

Thus, if A and B are independent,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Knowing B gives you no information about how likely A is to occur.

The law of total probability

The probability that A happens is the probability that A and B happen, or that A and B^C happen.

Draw a Venn diagram to convince yourself that

$$A = (A \cap B) \cup (A \cap B^C)$$

Because $A \cap B$ and $A \cap B^C$ are disjoint,

$$P(A) = P(A \cap B) + P(A \cap B^C) = P(A \mid B)P(B) + P(A \mid B^C)P(B^C).$$

Bayes' theorem

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Example

Let A be the event that a randomly sampled driver is under the influence.

Let B be the event that the randomly sampled driver tests positive.

What is the probability that a randomly sampled driver who tests positive is under the influence?

Bayes' theorem

Make the following assumptions:

1. The test's true positive rate is 0.9, or $P(B \mid A) = 0.9$.
2. The test's true negative rate is 0.95, or $P(B^c \mid A^c) = 0.95$
3. 1 % of all drivers are under the influence, so $P(A) = 0.01$.

Bayes' theorem

We want to compute $P(A \mid B)$, and Bayes' theorem tells us

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{0.9 \times 0.01}{P(B)},$$

The law of total probability tells us $P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$.

We know $P(B \mid A)P(A) = 0.9 \times 0.01$ and $P(A^c) = 1 - P(A) = 0.99$.

We can compute $P(B \mid A^c) = 1 - P(B^c \mid A^c) = 0.05$.

Bayes' theorem

We get

$$P(A \mid B) = \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.05 \times 0.99} \approx 0.15.$$

Even after having observed a positive test, it is more likely the person is not a user.

Intuition is often wrong about Bayes' theorem.

Summary of probability rules

1. $0 \leq P(A) \leq 1$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, which equals $P(A) + P(B)$ if A and B are disjoint
3. $P(A) \geq P(B)$ if B is a subset of A (A contains B)
4. $P(A \cap B) = P(A)P(B)$ if (and only if!) A and B are independent.
5. $P(A \mid B) = P(A \cap B)/P(B)$ if $P(B) > 0$.
6. $P(A \mid B) = P(B \mid A)P(A)/P(B)$ (Bayes' theorem, follows from 5.)

Remember to draw a Venn diagram if you are unsure!

Exercise

Is it true, for any events A and B, that $P(A) \geq P(A \cap B)$ and $P(A) \leq P(A \cup B)$?

Why?

Can the inequalities be equalities for some specific choices of A and B?

Random variables

Content

- Discrete random variables
- Continuous random variables
- Distributions

Random variables

Most events we will calculate probabilities for involve discrete or continuous random variables.

Recall that a random variable is a, typically numerical, measurement of the outcome of an experiment yet to be performed.

Discrete random variables

Discrete variables can take at most countably many values (countable support).

"Countably" has a mathematical definition, but it is quite literal: you can count the possible values.

They can can be finitely or infinitely many.

Example

The set $\{1, 1/2, 1/3, 1/4, \dots\}$ is countable and infinite.

Example

Suppose I flip a coin and if it comes up heads, I flip again. If it comes up tails, I stop.

Let X be the number of flips I will have made at the end of this experiment.

It is possible that I flip 1000 heads in a row, but highly improbable.

The same is true for any integer. Thus, X can take the values $1, 2, 3, \ldots$ and is therefore a discrete random variable that can take infinitely many values.

Continuous random variables

Continuous variables can take an uncountable number of values (uncountable support). That is, you cannot count the possible values even if you keep counting forever.

Example

The number of (decimal) numbers between 0 and 1.

The time it takes for my daughter to tie her shoes in the morning.

Distributions

Distribution

The rule (law) telling us the probabilities that X takes certain values is called the distribution of X.

Every random variable X has a cumulative distribution function (cdf).

Cumulative distribution function

The cdf of a random variable X is the function defined by $F(x) = P(X \leq x)$.

You plug in x, the cdf tells you the probability that X is less than or equal to x.

Distributions

The cdf tells you everything there is to know about the distribution of X.

- We say F characterizes the distribution of X.

In theory, if you know F, you can calculate any probabilities involving X.

Discrete probability distributions also have a probability mass function (pmf).

Probability mass function

The pmf of a discrete X is the function defined by $f(x) = P(X = x)$.

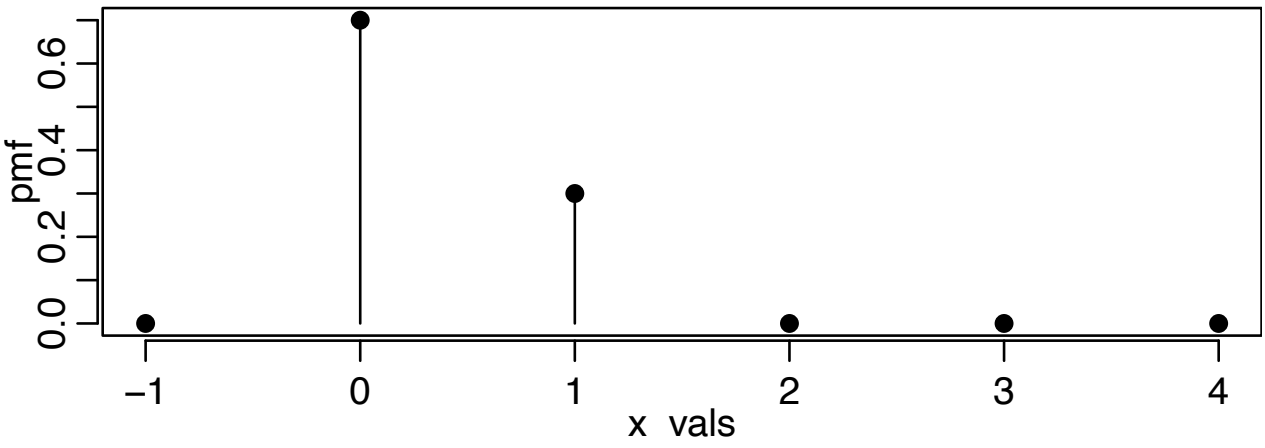
You plug in x, the pmf tells you the probability that $X = x$.

Bernoulli distribution

Example

We say that X has a Bernoulli distribution with parameter $0 \leq p \leq 1$, or $X \sim \mathrm{Ber}(p)$, if

$$P(X = 1) = p; \quad P(X = 0) = 1 - p$$



Binomial distribution

If X is the number of successes in n independent trials, each with success probability p, then X has a binomial distribution with parameters n and p, or $X \sim \mathrm{Bin}(n, p)$.

Example

Suppose we flip $n = 10$ coins and let X be the number of heads, then X has a binomial distribution with parameters $n = 10$ and $p = 1/2$ (assuming the coin is fair).

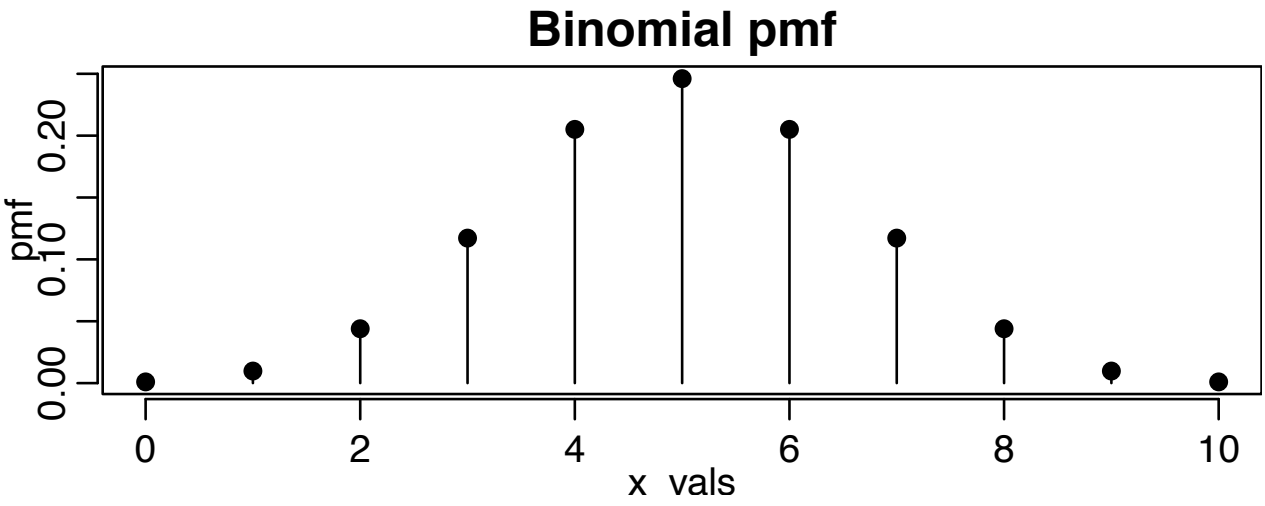
The pmf of X tells us the probability that $X = x$ for every possible x.

Binomial distribution

The binomial has pmf

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, \ldots, n.$$

The binomial coefficient is the number of ways to select x from n.



Binomial distribution

Example

Suppose we flip 10 coins and let X_i be one if the ith flip is heads, and 0 otherwise. Then $\sum_{i=1}^{n} X_i$ is the number of heads in 10 flips. This illustrates the following fact:

If $X_1, \ldots, X_n$ are independent $\mathrm{Ber}(p)$, then $X = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p)$.

Binomial distribution

Example

Suppose 50 % of all penguins in the population are of the Adelie species.

What was the probability of obtaining the number of Adelie penguins in the penguins data?

```
table(penguins$species)
```

```
##
##    Adelie Chinstrap    Gentoo
##       152        68       124
```

```
# The probability of 'x' successes in 'size' independent trials with success
# probability 'prob'
dbinom(x = 152, size = 152 + 68 + 124, prob = 0.5) # d for density
```

```
## [1] 0.00420746
```

Binomial distribution

Example

What was the probability of obtaining at least as many Adelie penguins as in the penguins data?

```
# Use the cdf
# 1 - P(X <= n_success - 1) = P(X > n_success - 1) = P(X >= n_success)
1 - pbinom(q = 152 - 1, size = 152 + 68 + 124, prob = 0.5)
```

```
## [1] 0.9865395
```

Exercise

What was the probability of obtaining fewer Adelie penguins than in the penguins data?

Binomial distribution

Exercise

Consider rolling three dice and let X be the number of 6s rolled. What is the distribution of X?

Binomial distribution

Exercise

Consider rolling two dice and let X be the number of 1s rolled. Find the probability mass function for X.

Answer: First note that X can take three values: $0, 1$, or 2. Thus, we need to find $f(x) = P(X = x)$ for $x = 0, 1, 2$. Let X_i be one if the ith roll is 1 and zero otherwise. The event that $X = 0$ is the same as $X_1 = 0 \cap X_2 = 0$. Assuming the rolls are independent, one of the rules of probabilities says

$$P(X = 0) = P(X_1 = 0 \cap X_2 = 0) = P(X_1 = 0)P(X_2 = 0) = (5/6)(5/6) = 25/36.$$

The event $X = 1$ consists of the outcomes $X_1 = 0 \cap X_2 = 1$ and $X_1 = 1 \cap X_2 = 0$. That is,

$$(X = 1) = (X_1 = 0 \cap X_2 = 1) \cup (X_1 = 1 \cap X_2 = 0)$$

Binomial distribution

The events in the union are disjoint, so the rules of probabilities say

$$P(X = 1) = P(X_1 = 0 \cap X_2 = 1) + P(X_1 = 1 \cap X_2 = 0),$$

which, assuming independence, is

$$P(X_1 = 0)P(X_2 = 1) + P(X_1 = 1)P(X_2 = 0) = (5/6)(1/6) + (1/6)(5/6) = 10/36.$$

It remains to find $P(X = 2)$. Can do similar calculation, or use that

$$(X = 2)^c = (X = 0) \cup (X = 1).$$

Poisson distribution

One of the most commonly used distributions in practice is the Poisson distribution.

Poisson distribution

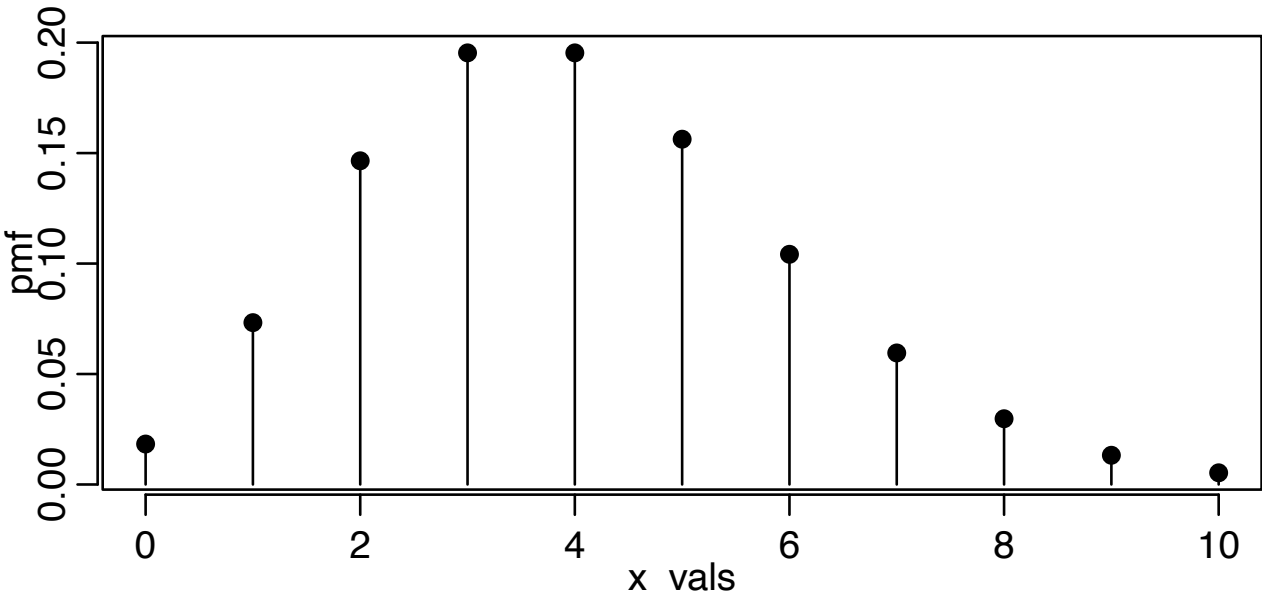
A random variable X has a Poisson distribution with parameter $\lambda > 0$ if it has pmf

$$f(x) = e^{-\lambda}\lambda^x/x!, \quad x! = x(x-1)(x-2)\cdots 2, \quad x = 0, 1, \ldots$$

Poisson distribution

Poisson distribution

```
x_vals <- 0:10; lambda <- 4; pmf <- dpois(x_vals, lambda)
plot(x_vals, pmf , type = "h"); points(x_vals, pmf)
```



Expectation and variance

For a discrete X, its mean (expected value) and variance are

$$\mu = E(X) = \sum_x xf(x) = \sum_x xP(X = x)$$

$$\sigma^2 = \mathrm{var}(X) = \sum_x (x - \mu)^2 f(x) = \sum_x (x - \mu)^2 P(X = x) \geq 0$$

The sums are over all x such that $P(X = x) > 0$ (the support of X).

The standard deviation of X is $\sqrt{\mathrm{var}(X)}$.

Expectation and variance

Intuition

If X has large mean, then if we observe many independent realizations they will be large on average.

If X has large variance, then if we observe many independent realizations they will be very different.

Exercise

Show (or remember) that

1. $E(X - c) = E(X) - c$
2. $E(cX) = cE(X)$
3. $\mathrm{var}(X - c) = \mathrm{var}(X)$
4. $\mathrm{var}(cX) = c^2\mathrm{var}(X)$.

Expectation and variance

Example

The mean and variance of $X \sim \mathrm{Ber}(p)$ is

$$\mu = 1 \times P(X = 1) + 0 \times P(X = 0) = 1 \times p + 0 \times (1 - p) = p$$

$$\sigma^2 = (1 - p)^2 \times P(X = 1) + (0 - p)^2 \times P(X = 0) = (1 - p)^2 p + p^2 (1 - p) = p - p^2.$$

Expectation and variance

One can show that if X is binomial with parameters n and p and Y is Poisson with parameter λ, then

$$E(X) = np, \quad \mathrm{var}(X) = n(p - p^2)$$

$$E(Y) = \lambda, \quad \mathrm{var}(Y) = \lambda.$$

Summary discrete variables

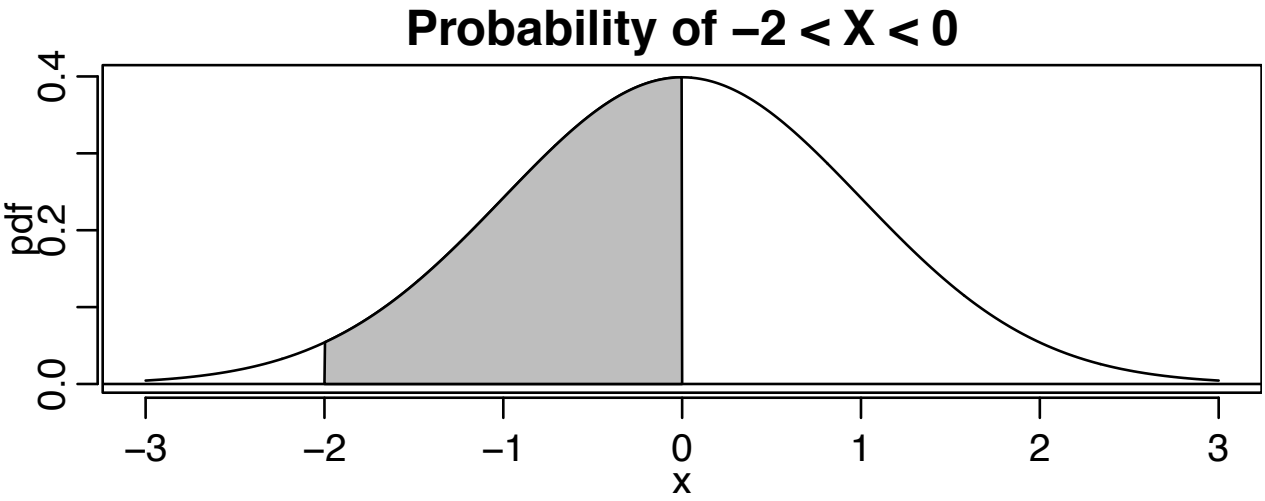
- A discrete random variable is one that has countable support.
- The distribution of a discrete random variable X is characterized by its pmf $f(x) = P(X = x)$.
- Uniform, Bernoulli, Binomial, and Poisson (there are many others)
- You can calculate mean and variance by sums

Continuous variables

Continuous variables have uncountable support.

The distribution of continuous random variables are characterized by a probability density function (pdf) $f(x)$.

The area under the graph of a pdf from a to b tells you the probability that $a \leq X \leq b$.



Density and cumulative distribution function

In general,

$$P(a < X \leq b) = \int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

$$P(a < X \leq b) = P((X \leq b) \cap (X > a))$$

You will not have to integrate anything analytically in this class—we will use R.

In R, you can evaluate cdfs for common distributions.

```
pnorm(0) - pnorm(-2) # The area in the previous slide
```

```
## [1] 0.4772499
```

Density and cumulative distribution function

You should know:

- The total area under a pdf is 1 (probability that X is between $-\infty$ and ∞)
- The probability that X is between a and b is the probability that X is less than b minus the probability that X is less than a, so we can compute it as $P(a < X < b) = F(b) - F(a)$.

Exercise

Use basic rules of probabilities to explain why the second point is true.

Expectation and variance

Why don't we characterize continuous distributions by a pmf $f(x) = P(X = x)$?

It is outside the scope of this class to prove, but you should know that, for a continuous X,

$$P(X = x) = 0 \quad \text{for every } x.$$

A continuous X has mean and variance

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, \mathrm{d}x$$

and

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, \mathrm{d}x.$$

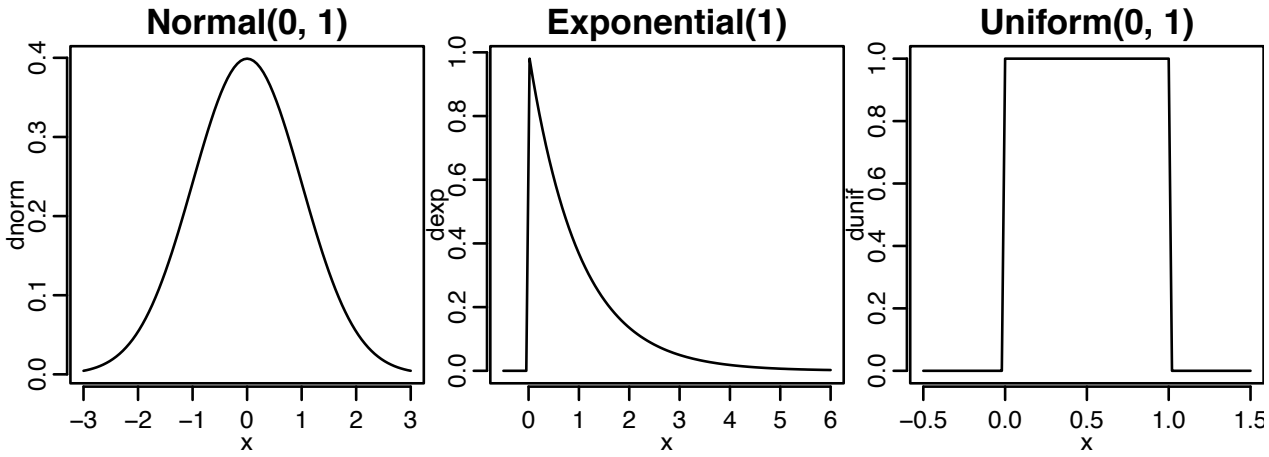
Common continuous distributions

Some continuous distributions

Normal with mean μ and variance σ^2: $f(x; \mu, \sigma^2) = e^{-(x-\mu)^2/(2\sigma^2)} / \sqrt{2\pi\sigma^2}$

Uniform on $[a, b]$: $f(x) = 1/(b-a)$ for $a \leq x \leq b$

Exponential with parameter $\lambda > 0$: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$



Cumulative distribution functions in R

The probability that they are less than 0.9:

```
pnorm(0.9)
```

```
## [1] 0.8159399
```

```
pexp(0.9)
```

```
## [1] 0.5934303
```

```
punif(0.9)
```

```
## [1] 0.9
```

Exercise

What is the probability that they are greater than 0.8? How to calculate it in R?

Moments of exponential and uniform

One can show that if X is exponential and Y uniform on $[a, b]$, then

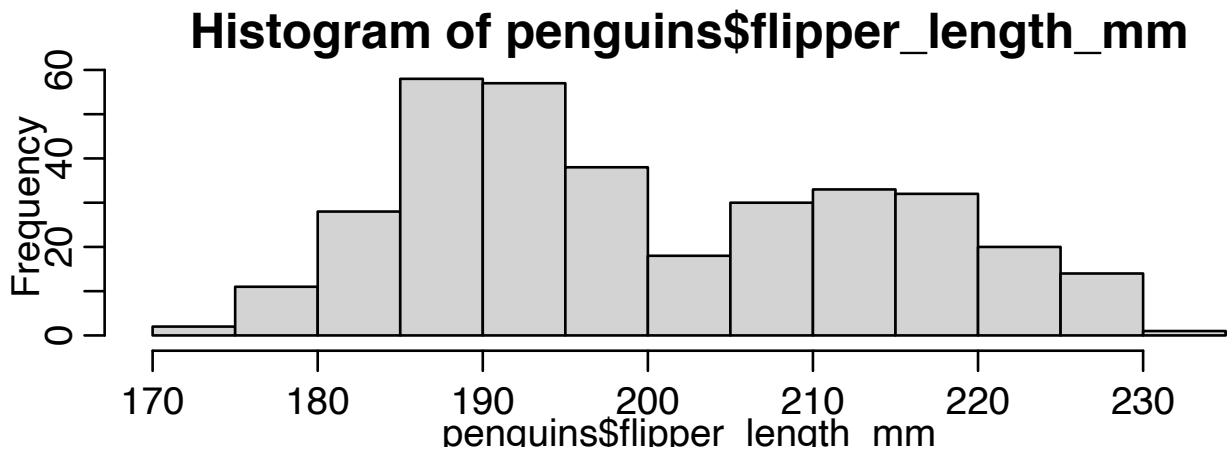
$$E(X) = 1/\lambda, \quad \mathrm{var}(X) = 1/\lambda^2$$

$$E(Y) = (b-a)/2, \quad \mathrm{var}(Y) = (b-a)^2/12.$$

Connection to the histogram

Recall that the histogram tells you how many observations in a certain interval.

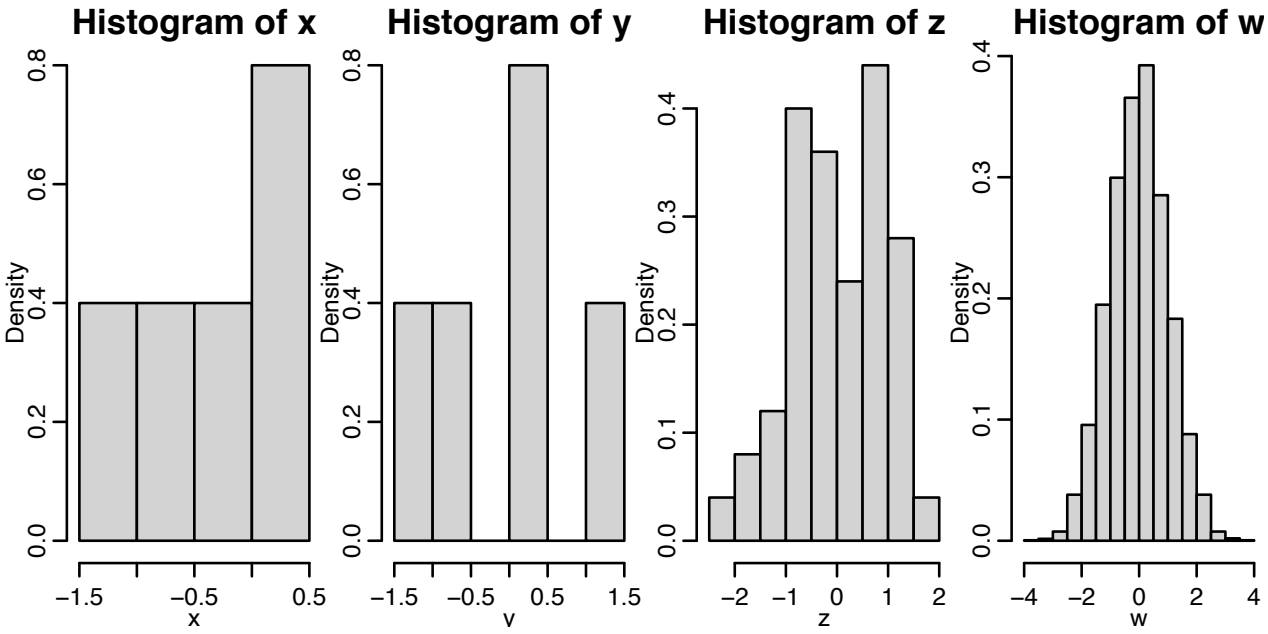
```
hist(penguins$flipper_length_mm, cex = 0.5)
```



If you make the intervals smaller and the sample larger, eventually the histogram should look like the pdf of a randomly selected penguin's flipper length.

Connection to the histogram

```
x <- rnorm(5); y <- rnorm(10); z <- rnorm(50); w <- rnorm(5000)
```



Models, estimation, and inference

Purpose of models

In many settings one would need an unrealistically large sample to effectively use only the histogram for inference.

Suppose, for example, we want to know how common it is that a penguin has flippers longer than 235 mm.

The penguin with the longest flippers in the sample has 231 mm. Does it mean it is impossible for penguins to have flippers longer than 235 mm? Maybe, but probably not.

A model can help.

Definition

In statistics, a model is a family of distributions indexed by parameters.

What does it mean?

Example

Assume that, in the population of all penguins, flipper length is normally distributed with unknown mean μ and unknown variance σ^2. That is, if X is the flipper length of a randomly selected penguin, then $X \sim N(\mu, \sigma^2)$.

We have specified a family of distributions, but not a specific distribution, since we have not specified μ and σ^2.

Parametric models

When the number of parameters in the model is finite, it is called a parametric model.

With the help of our model, we may be able to calculate probabilities of events not observed in our sample.

Example

If we can use our sample to figure out what μ and σ^2 are, approximately, then we can calculate the probability of a randomly selected penguin having flippers longer than 235 mm. For example, if $\mu = 200$ and $\sigma^2 = 200$, then

```
1 - pnorm(235, mean = 200, sd = sqrt(200))
```

```
## [1] 0.006664164
```

That is, the proportion of penguins in the population with flippers longer than 235 mm is approximately 7/1000.

Parametric models

Our guess that about 7/1000 penguins have flippers longer than 235 mm uses approximations.

1. If the distribution of flipper lengths is not approximately normal, then the probability calculation can be very wrong.
2. We do not know the true mean and variance of the flipper lengths. If our guesses of μ and σ^2 are poor, then again the probability calculation can be very wrong.

Famous quotes

There are two famous quotes:

> *Everything should be made as simple as possible, but no simpler (A. Einstein).*
>
> *All models are wrong, but some are useful (G.E.P. Box).*

In our example, assuming a normal distribution may be useful, or it may be making things too simple.

Estimation

Having selected a model, we want to make and educated guess on what the true parameters are.

More formally, we want to estimate the parameters using data.

In the penguins example, what are natural estimates of μ and σ^2?

- The sample versions!

Estimation

If $x_1, \ldots, x_n$ are the flipper lengths in our sample, we can use

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

to estimate μ, the mean flipper length of a randomly selected penguin.

Estimation

It is common to denote estimates by a hat, so $\bar{x} = \hat{\mu}$ in this example.

It is very important not to confuse μ and $\hat{\mu}$: one is an unknown but constant value, the other is a statistic. That is, a realization of a random variable that we can calculate using data.

The mean as random variable

It may be counter-intuitive that the sample mean is a realization of a random variable.

Recall, a random variable is a (numerical) outcome of a yet to be performed experiment.

Before you sample, you do not know what the sample mean will be.

If $X_1, \ldots, X_n$ are flipper lengths of yet to be sampled penguins, then their average

$$\bar{X} = \sum_{i=1}^{n} X_i / n$$

is random.

Intuition

Having sampled and observed $\bar{X} = \bar{x}$, is there reason to believe $\bar{x}$ is a good estimate of μ?

Formal motivation

Yes!

Suppose that $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$. Then

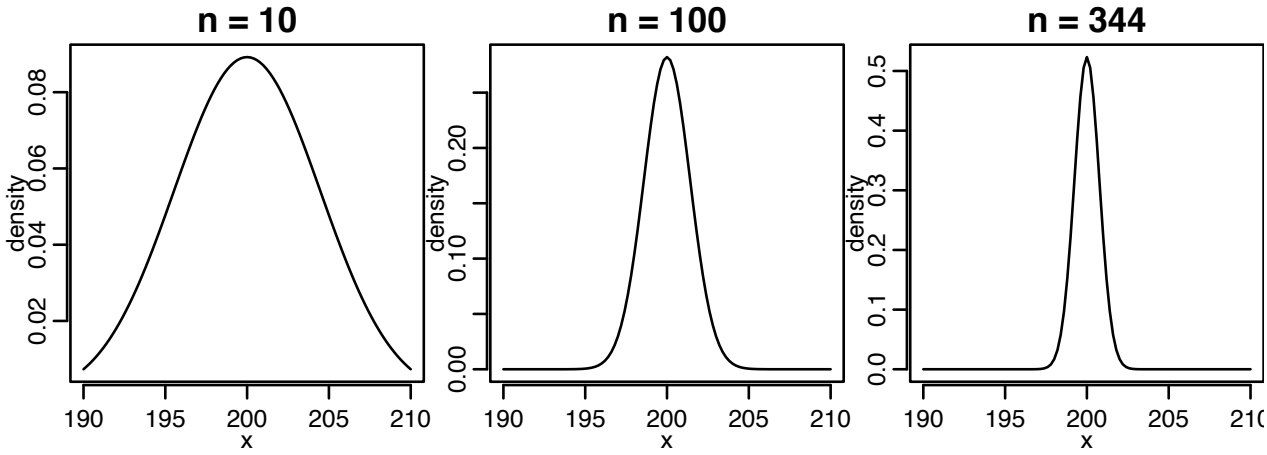
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n).$$

This says that the expected value of $\bar{X}$ is μ, and the variability of $\bar{X}$ decreases as n increases.

That is, it is more and more likely that $\bar{X}$ is close to μ the larger n is.

Distribution of sample mean

Let's look at the distribution of $\bar{X}$ for different values of n when $\mu = 200$ and $\sigma^2 = 200$.



As the sample size increases, it becomes increasingly unlikely to observe $\bar{X} = \bar{x}$ far from μ.

Distribution of sample mean

The distribution of the sample mean concentrates around the true mean, so it was unlikely to get a sample where $\bar{x}$ is far from μ; doesn't mean it didn't happen!

Example

```
mean(penguins$flipper_length_mm, na.rm = T)
```

```
## [1] 200.9152
```

Uncertainty quantification

The sample mean $\bar{x}$ is a point estimate of μ.

We also want to quantify the uncertainty in that estimate.

Standard error

The standard error of $\bar{x}$ is an estimate of the standard deviation of $\bar{X}$.

Estimates should be accompanied by standard errors whenever possible.

Standard error

Recall, if $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$.

Thus, the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$.

We can estimate this by $s/\sqrt{n}$, which we sometimes denote $\mathrm{se}(\bar{x})$.

Standard error

```
  # sample mean
  x_bar <- mean(penguins$flipper_length_mm, na.rm = T)
  # standard error
  se <- sd(penguins$flipper_length_mm, na.rm = T) /
    sqrt(nrow(penguins) - sum(is.na(penguins$flipper_length_mm)))
  x_bar
```

```
## [1] 200.9152
```

```
  se
```

```
## [1] 0.7603704
```

Our estimate is that the standard deviation of $\bar{X}$ is ≈ 0.76.

Standard error

Roughly, the standard error tells us how much we expect $\bar{X}$ to vary from sample to sample.

Certainly, if the standard error is similar to the estimate in magnitude, then we do not trust the estimate.

It is often reasonable to believe μ is within $\pm 2 \times \text{se}(\bar{x})$; we will soon see why.

Example

```
mean(penguins$flipper_length_mm, na.rm = T)
```

```
## [1] 200.9152
```

```
# Two standard errors
2 * sd(penguins$flipper_length_mm, na.rm = T) / sqrt(sum(!is.na(penguins$flipper_length_mm)))
```

```
## [1] 1.520741
```


Confidence interval

Confidence interval

Let us continue to assume $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, and suppose for simplicity that σ^2 is known (but not μ).

We will define confidence using an example.

Example

One can (and we will soon) show that

$$P\left(\bar{X} - 2\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 2\sigma/\sqrt{n}\right) = 0.95$$

Thus, when we observe $\bar{X} = \bar{x}$, since we know σ^2 we can calculate the interval

$$[a, b] = [\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n}].$$

We say that $[a, b]$ is a 95 % confidence interval for μ.

Probability and confidence

Be careful: the probability is for the random interval.

Since μ is a fixed constant, it is either in $\bar{x} \pm 2 \times \sigma/\sqrt{n}$ or not—there is no probability!

We say that we are 95 % confident μ is in the interval.

Constructing a confidence interval

Let's walk through the details of constructing a confidence interval with any confidence level.

We know that $\bar{X} \sim N(\mu, \sigma^2/n)$, and from this it follows that

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1).$$

Step 1

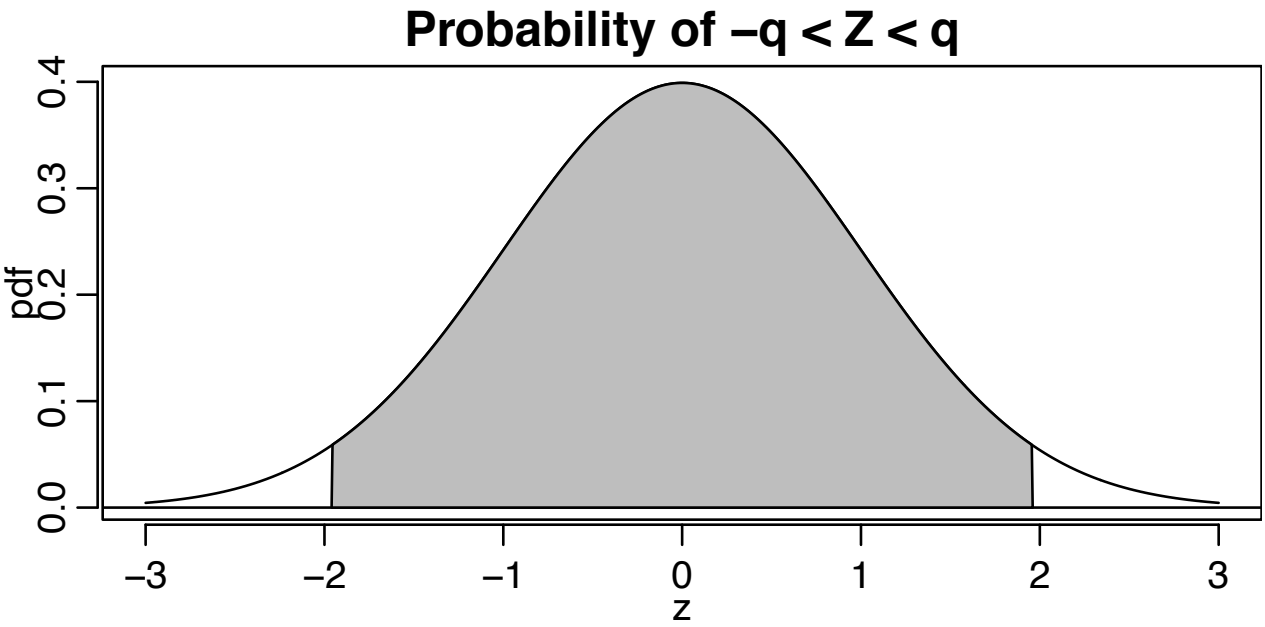
For a $100 \times (1 - \alpha)\%$ confidence interval, start by picking q such that

$$P(-q \leq Z \leq q) = 1 - \alpha$$

How?

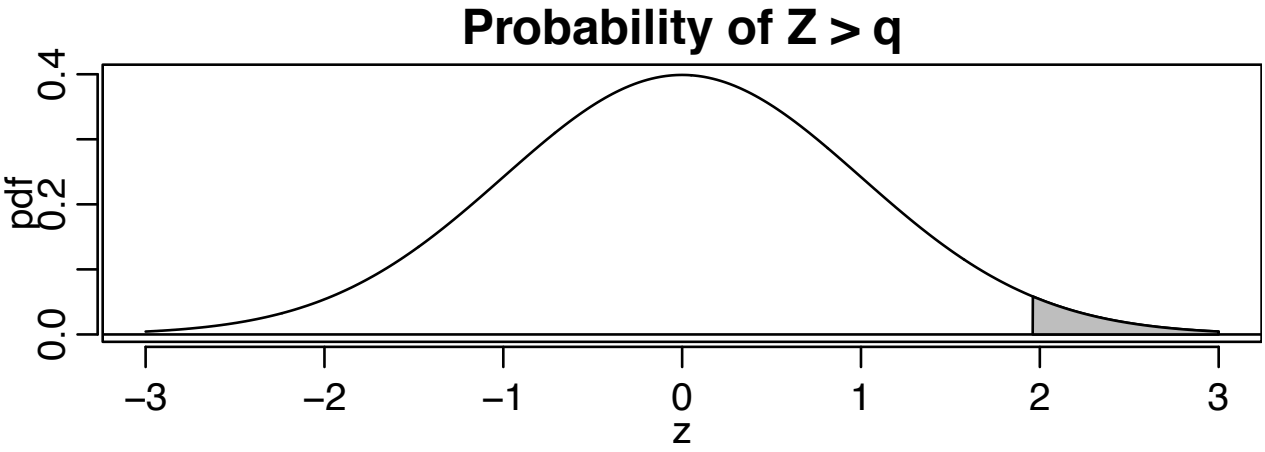
Constructing a confidence interval

Because the normal distribution is symmetric, the two white regions have the same area, so the shaded region has area $1 - \alpha$ if the white regions each have area $\alpha/2$.



Constructing a confidence interval

We can find this q using the qnorm function.




```
# If alpha/2 = 0.025
qnorm(0.025, lower = F)
```

```
## [1] 1.959964
```

Did we get the right number?

Sanity check:

```
# approx P(-1.96 < Z < 1.96)
pnorm(1.959964) - pnorm(-1.959964)
```

```
## [1] 0.95
```

```
pnorm(1.96) - pnorm(-1.96)
```

```
## [1] 0.9500042
```

```
pnorm(2) - pnorm(-2)
```

```
## [1] 0.9544997
```

Constructing a confidence interval

We have used R to find a q such that

$$P\left(-q \leq \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \leq q\right) = 1 - \alpha.$$

Step 2

Solve to get μ in the middle.

The event in the probability is the same, and hence has the same probability, as

$$\bar{X} - q\sigma/\sqrt{n} \leq \mu \leq \bar{X} + q\sigma/\sqrt{n}$$

Thus, we are $100 \times (1 - \alpha)\%$ confident that

$$\bar{x} - q\sigma/\sqrt{n} \leq \mu \leq \bar{x} + q\sigma/\sqrt{n}$$

Summary

Summary

With normal random variables:

- We are 95 % confident the true mean is within $\pm$ 1.96 standard deviations of the sample mean
- For any $0 < \alpha < 1$, we can select q such that $\bar{x} \pm q\sigma/\sqrt{n}$ is a $100 \times (1-\alpha)\%$ confidence interval
- All based on $Z = (\bar{X} - \mu)/\sqrt{\sigma^2/n} \sim N(0,1)$

Next

1. What to do when the variables are not normal?
2. What to do when the variance is not known?

Central limit theorem

Central limit theorem

Arguably the most important theorem in statistics.

The Central Limit Theorem (CLT)

If $X_1, \ldots, X_n$ are independent with the same distribution, then for large n, the sample mean $\bar{X}$ is approximately normally distributed with mean $\mu = E(X_i)$ and variance $\sigma^2/n = \mathrm{var}(X_i)/n$.

A consequence of the CLT is that

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$$

is approximately normally distributed with mean zero and variance one for large n, *regardless of which distribution the X_i have*.

Central limit theorem

This means, for example, that it is still true that, if n is large,

$$P\left(\bar{X} - 1.96\sqrt{\sigma^2/n} \leq \mu \leq \bar{X} + 1.96\sqrt{\sigma^2/n}\right) \approx 0.95.$$

In fact, the probability converges to 0.95 as n tends to infinity.

Central limit theorem for Bernoulli

Suppose that $X_i \sim \mathrm{Ber}(1/2)$; this implies $\mu = 1/2$ and $\mathrm{var}(X_i) = 1/4$.

The following function draws n_{samps} independent samples, each consisting of n independent $\mathrm{Ber}(1/2)$. It returns $z = (\bar{x} - \mu)/\sqrt{1/4n}$ for each sample.

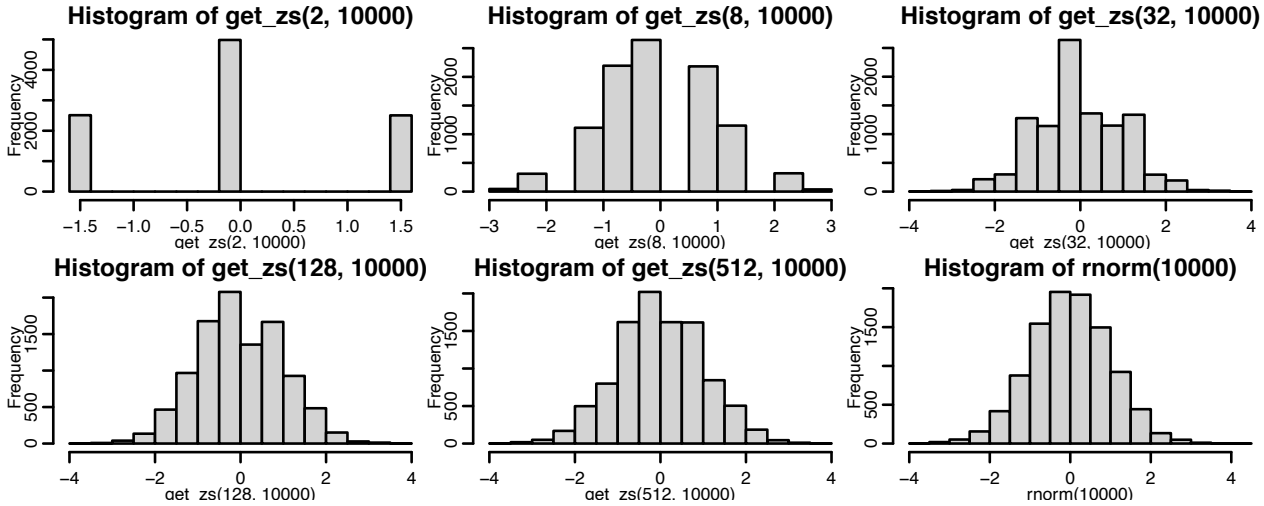
```
get_zs <- function(n, n_samps){
  many_z <- rep(0, n_samps) # Allocate
  for(i in 1:n_samps){
    x <- rbinom(n, 1, 1/2) # Draw one sample
    many_z[i] <- (mean(x) - 0.5) * 2 * sqrt(n) # Save z-statistic
  }
  return(many_z)
}
```

The Bernoulli distribution is very different from a normal distribution.

But what about the distribution of a sample average of Bernoulli?

Central limit theorem for Bernoulli

```
par(mfrow = c(2, 3), cex = 0.5)
hist(get_zs(2, 1e4)); hist(get_zs(8, 1e4)); hist(get_zs(32, 1e4));
hist(get_zs(128, 1e4)); hist(get_zs(512, 1e4)); hist(rnorm(1e4))
```



Central limit theorem

Always remember

The CLT says nothing about the distribution of the variables themselves, only their (random) sample average!

In particular, many Bernoulli variables are still Bernoulli variables.