Biostatistics 1BIO43

Karl Oskar Ekvall

Fall 2021

Karolinska Institutet

Table of contents

1. Realizations and data

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.

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.

More generally, we will consider descriptive statistics that quantify:

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

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
```

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.

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
```

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
```

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**:

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

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
```

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
```

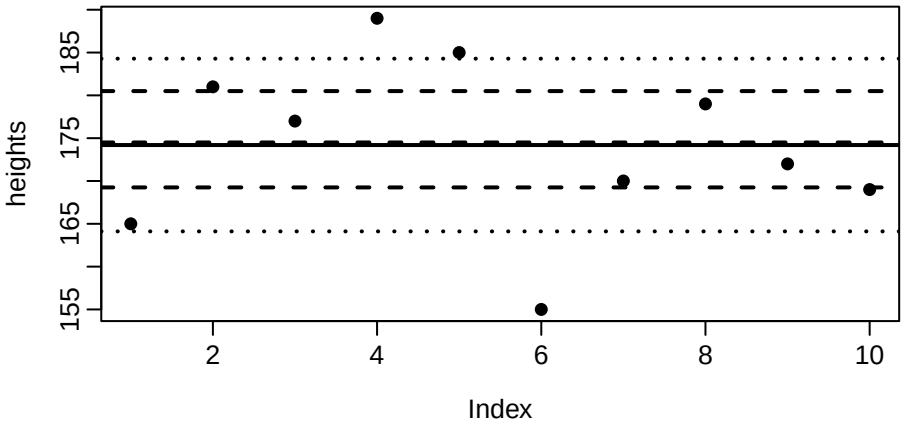
```
IQR(heights)
```

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

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



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?

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

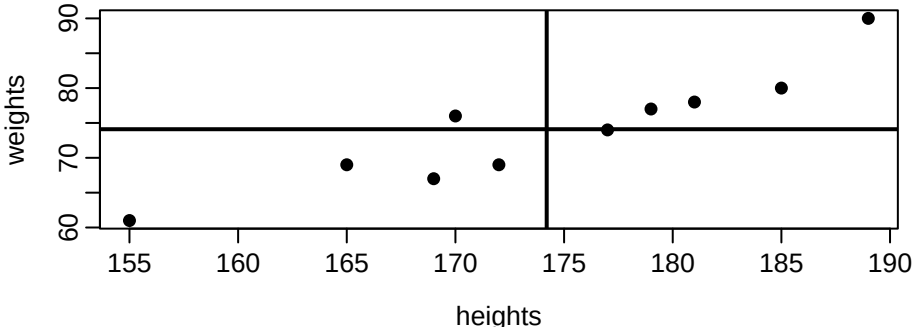
$$\rho_{xy} = \frac{s_{xy}}{s_x s_y},$$

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

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)
```



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
```

Example

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

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

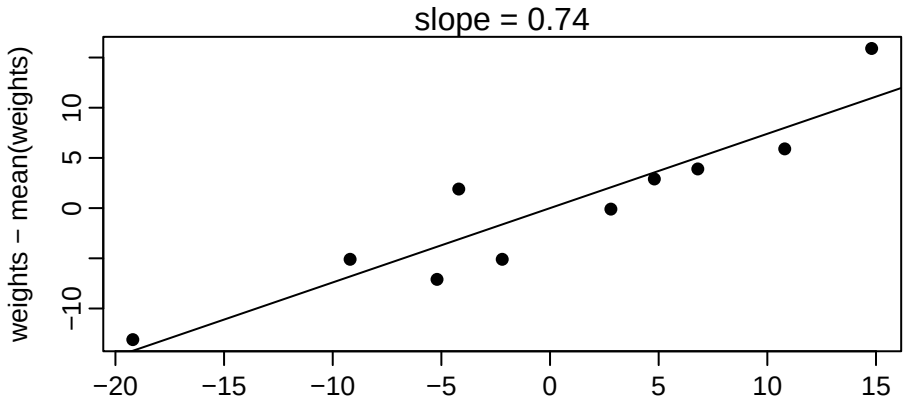
```
cor(weights, heights)
```

```
## [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
```

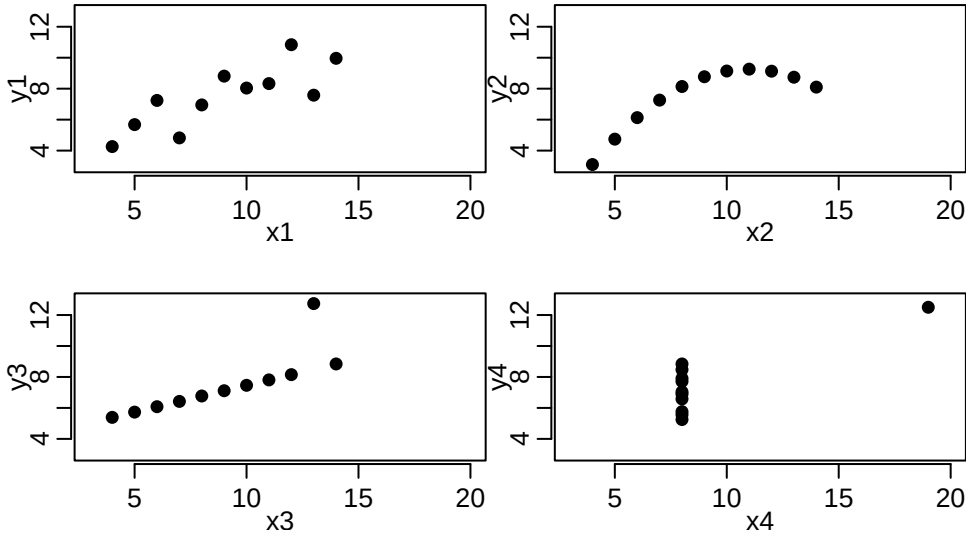
```
round(apply(anscombe, 2, sd), 2)
```

```
##   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.

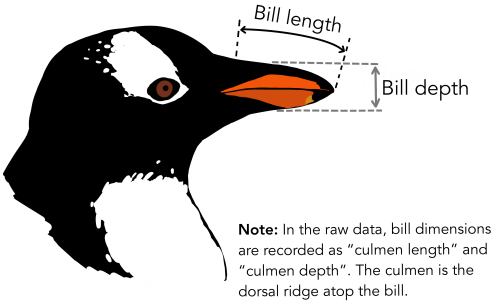
-you may (obviously) want to do some further investigation before concluding the drug is effective

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

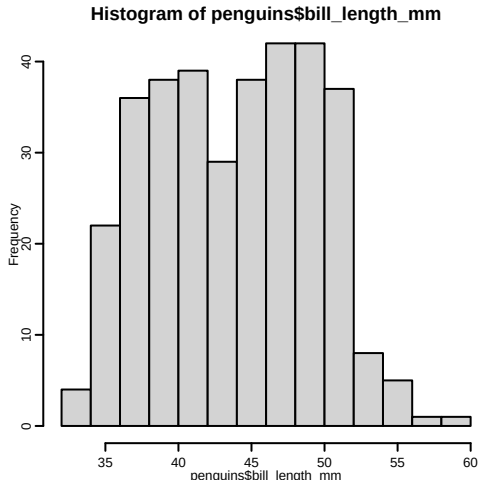
Let's consider bill lenght and depths in the `penguins` data.

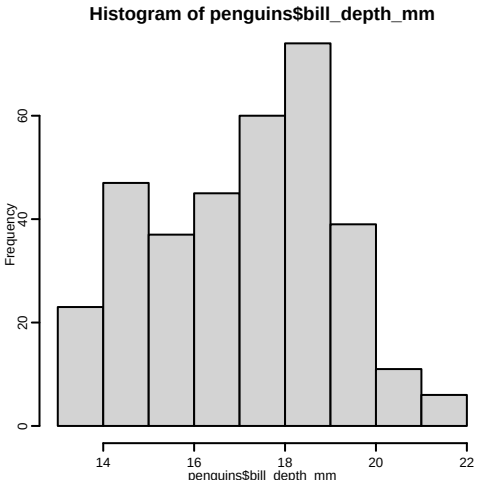
It's at `https://github.com/allisonhorst/palmerpenguins` and artwork is by @allison_horst.

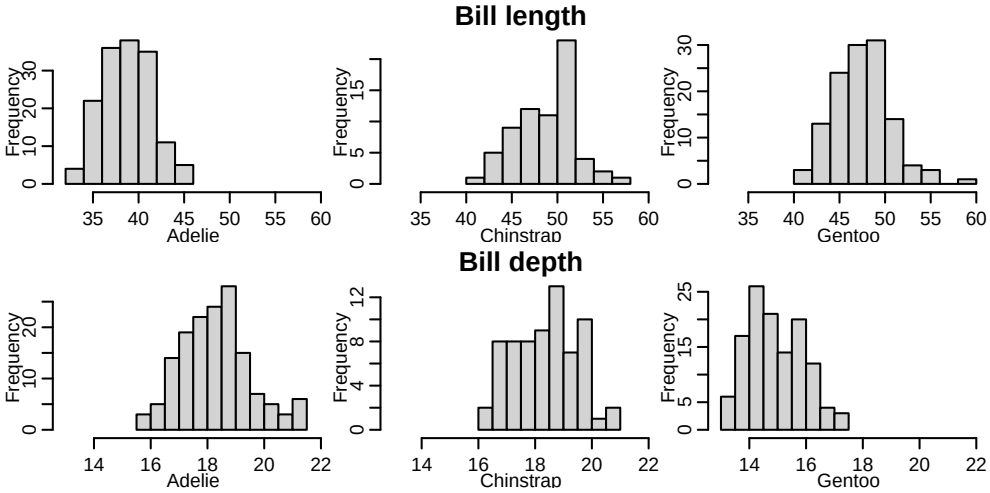
```
# install.packages("palmerpenguins")
library(palmerpenguins)
```



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



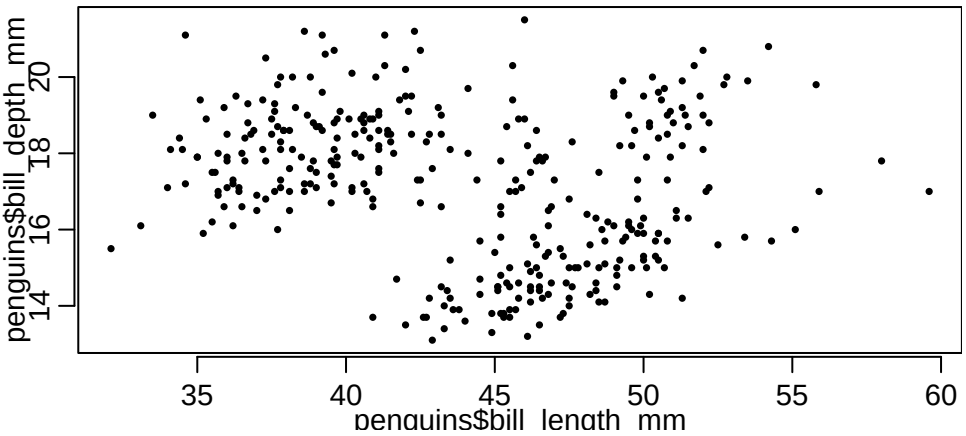




How do the species differ with respect to length and depth?

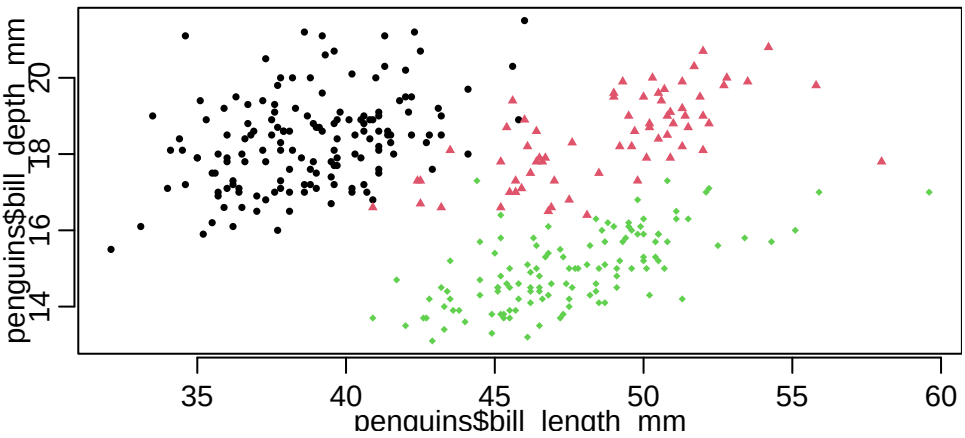
Is there a correlation between length and depth?

```
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)
```



(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.

2. 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

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?

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 $P(A) \geq P(B)$.

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.

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 $P(C) \geq P(D)$.

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 $P(A \cup A^c) = 1$.

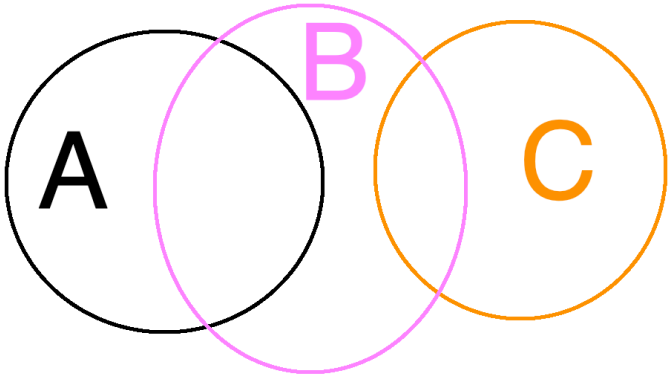
Because A and A^c cannot happen at the same time, one of the rules of probabilities says

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

Venn diagrams

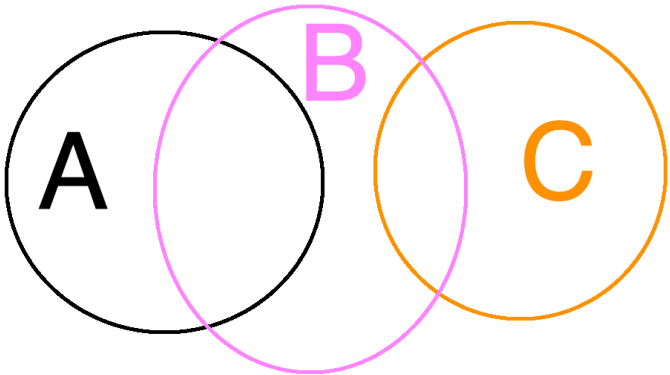
The white region of the slide is the sample space, and is has size 1.

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

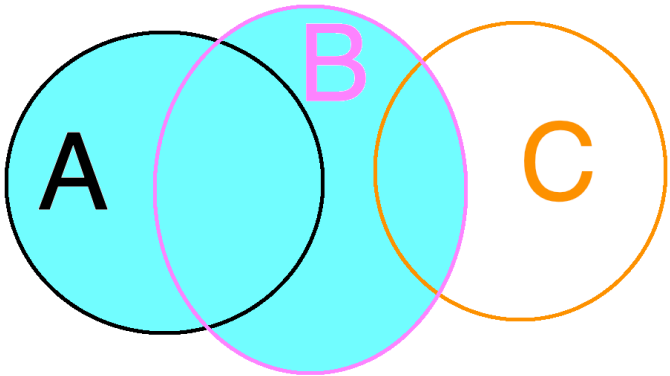


Venn diagrams

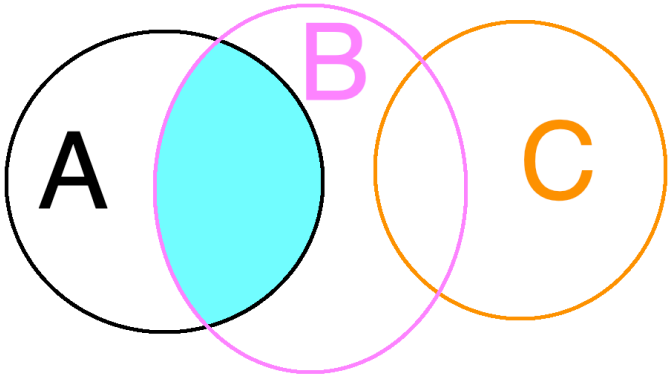
Find $A \cup B$, $A \cap B$, $A \cap C$, $A \cup C$, $A \cup B \cup C$, and $(A \cap B) \cup (B \cap C)$



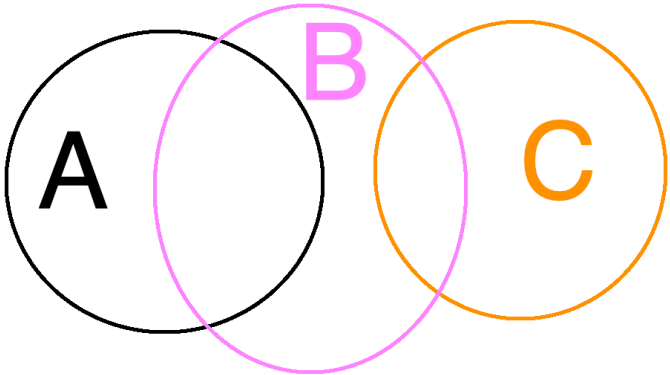
$A \cup B$



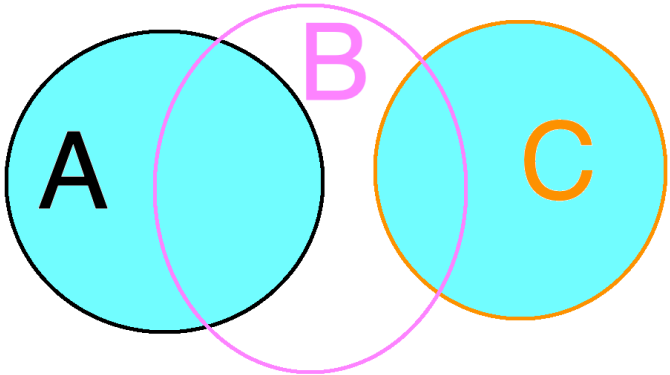
$A \cap B$



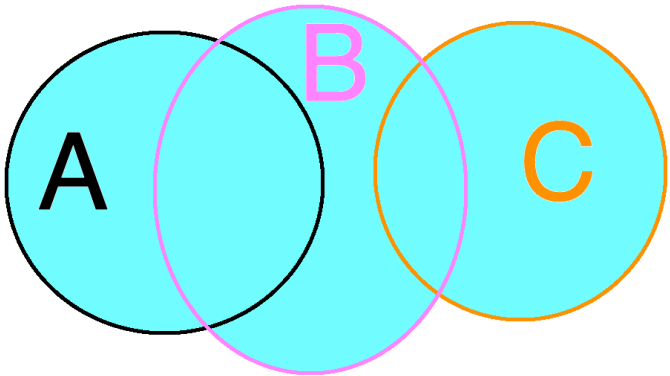
$A \cap C$



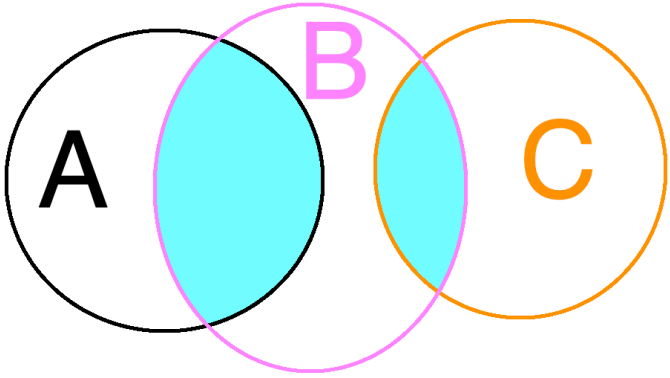
$A \cup C$



$A \cup B \cup C$



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



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 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

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

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!

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.

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.

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,

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

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

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?

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 $P(A) = 0.01$.

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

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?

3. Random variables

Content

- Discrete random variables
- Continuous random variables
- Distributions

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.

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.

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

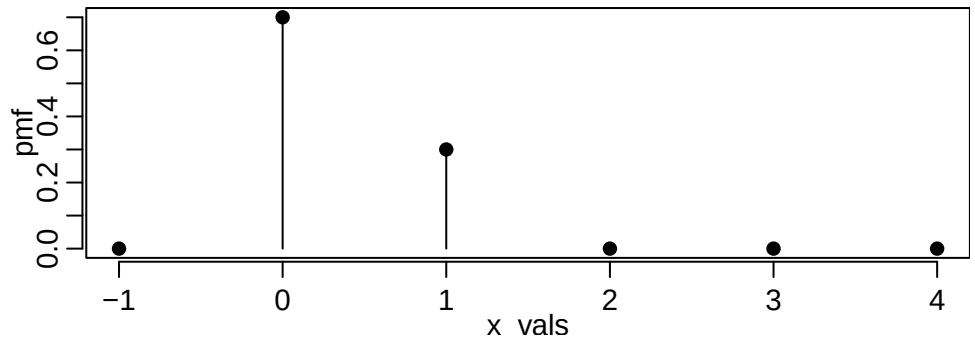
The pmf of a discrete X is the function defined by $f(x) = P(X = x)$.

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

Example

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

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



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

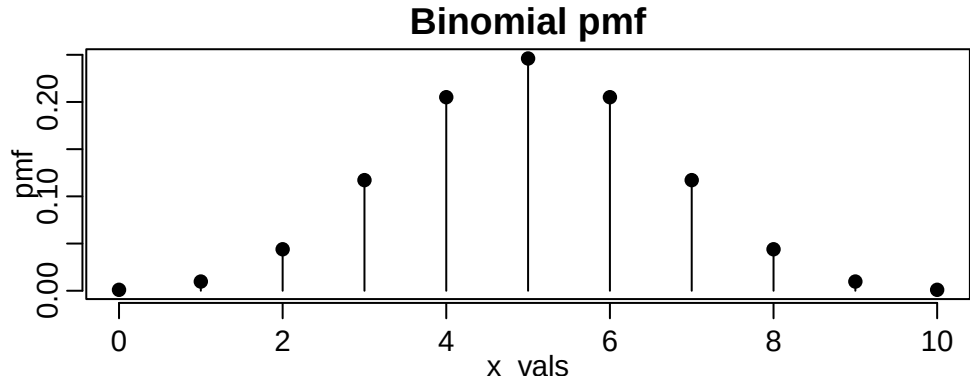
Suppose we flip $n = 10$ coins and let X be the number of heads, then X has a binomial distribution with parameters $n = 10$ and $p = 1/2$ (assuming the coin is fair).

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

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.



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 $X = \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, p)$.

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
```

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?

Exercise

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

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,

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

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

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

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

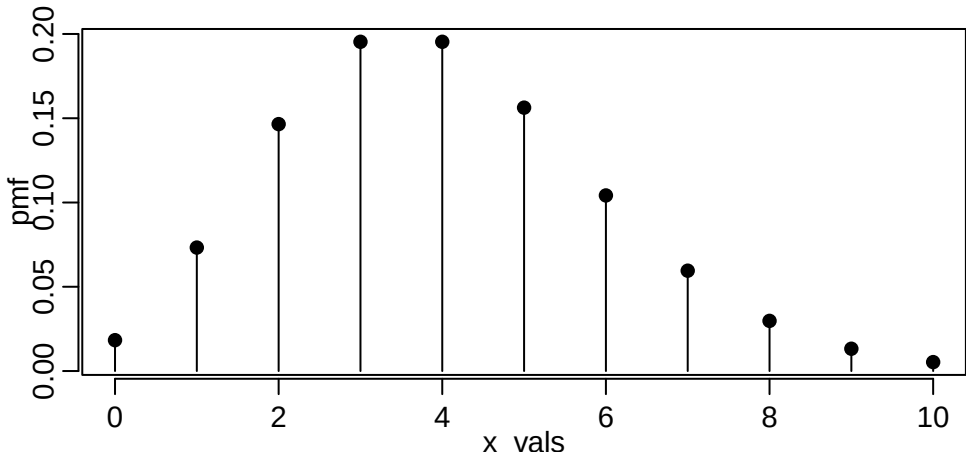
Poisson distribution

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

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



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 $\sqrt{\mathrm{var}(X)}$.

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. $\mathrm{var}(cX) = c^2\mathrm{var}(X)$.

Example

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

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

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

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

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

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

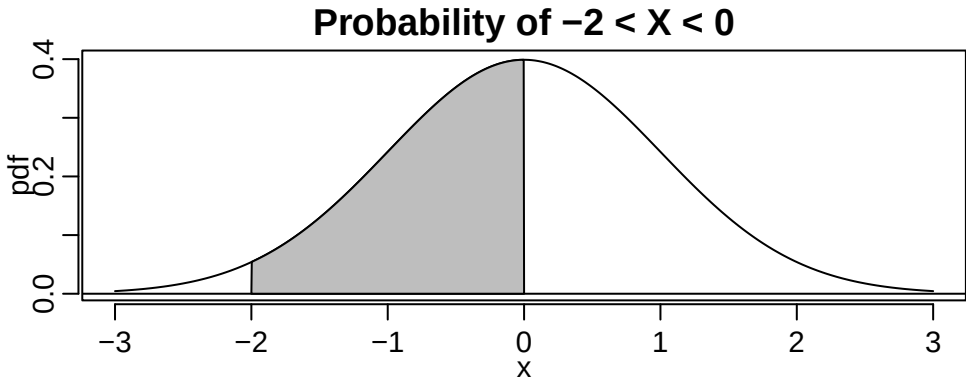
Summary discrete variables

- 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 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 $a \leq X \leq b$.



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
```

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.

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

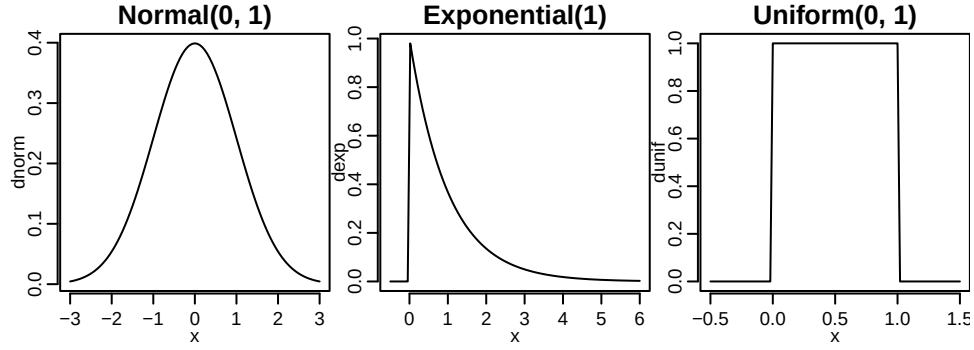
$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, \mathrm{d}x.$$

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 $x \geq 0$



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

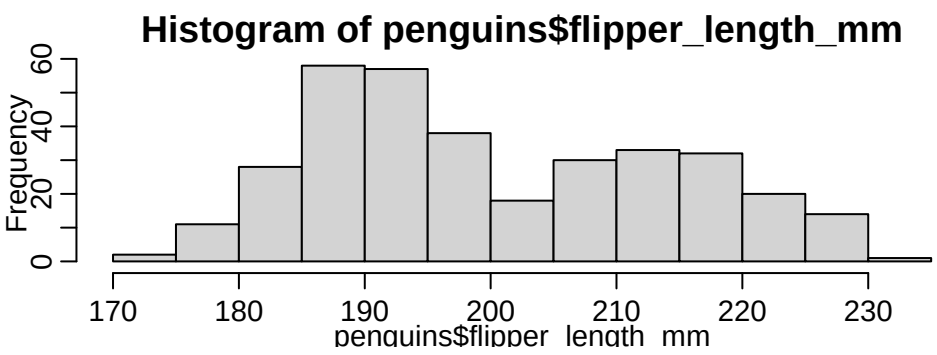
$$E(X) = 1/\lambda, \quad \mathrm{var}(X) = 1/\lambda^2$$

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

Connection to the histogram

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

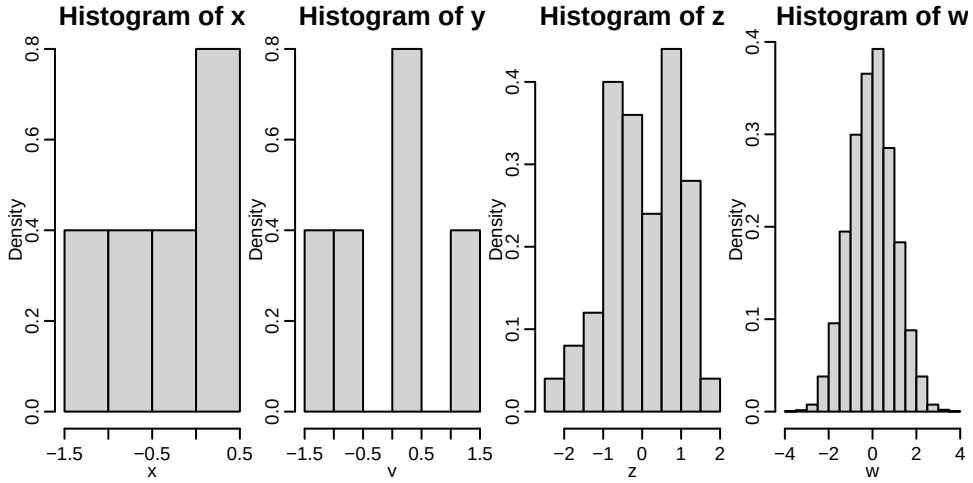
```
hist(penguins$flipper_length_mm, cex = 0.5)
```



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)
```



4. 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.

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 $7/1000$.

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!

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.

It is common to denote estimates by a hat, so $\overline{x} = \hat{\mu}$ 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.

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

Yes!

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

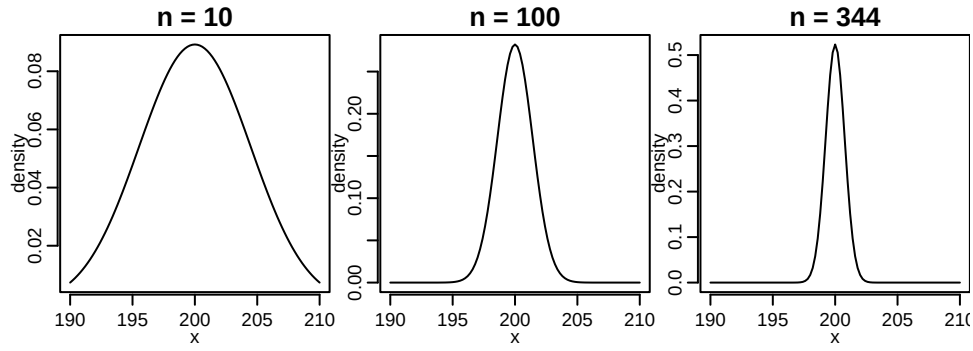
$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n).$$

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.

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
```

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

```r
  # 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
```

```r
  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

A confidence interval for μ gives a range of possible values of μ consistent with the observed data.

Sometimes called an interval estimate of μ.

The higher confidence we want that μ is in an interval, the larger the interval must be.

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.

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.

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

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

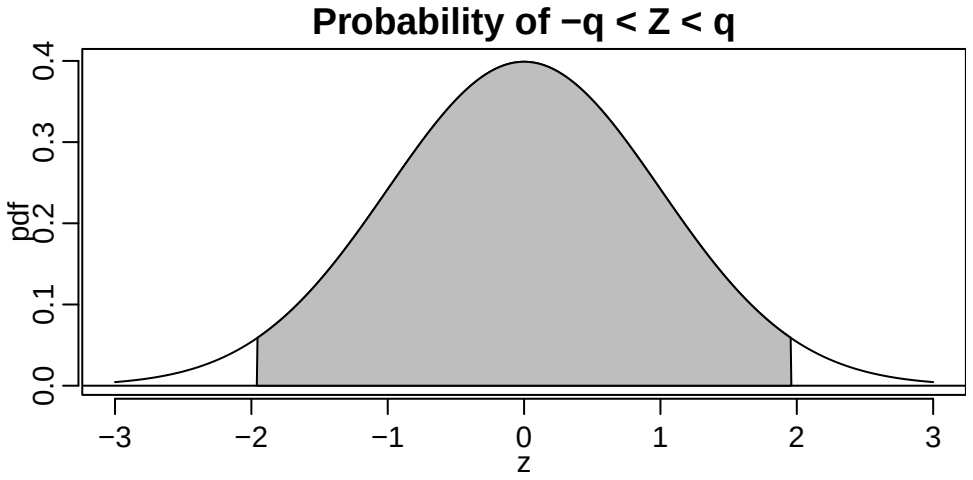
Step 1

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

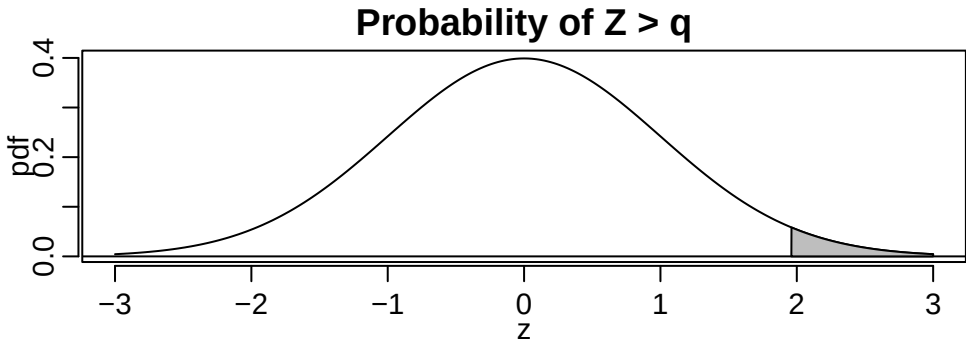
$$P(-q \leq Z \leq q) = 1 - \alpha$$

How?

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 $\alpha/2$.



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
```

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

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?

5. 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.*

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

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

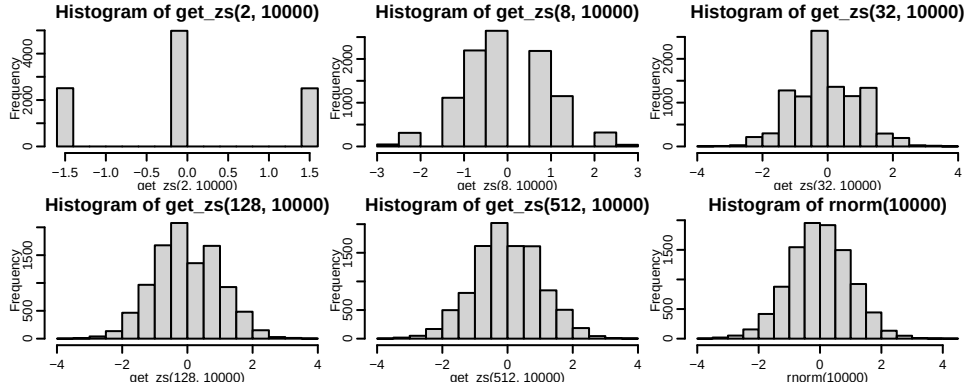
```
get_zs <- function(n, n_samps){
  many_z <- rep(0, n_samps) # Allocate
  for(i in 1:n_samps){
    x <- rbinom(n, 1, 1/2) # Draw one sample
    many_z[i] <- (mean(x) - 0.5) * 2 * sqrt(n) # Save z-statistic
  }
  return(many_z)
}
```

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))
```



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.

Confidence interval for the mean (σ unknown)

Recall that, if X_is are normal,

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

If σ is not known, we cannot calculate the resulting confidence interval

$$\bar{x} \pm q\sigma/\sqrt{n}.$$

We will estimate σ by S,

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

If $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \quad (\neq N(0,1))$$

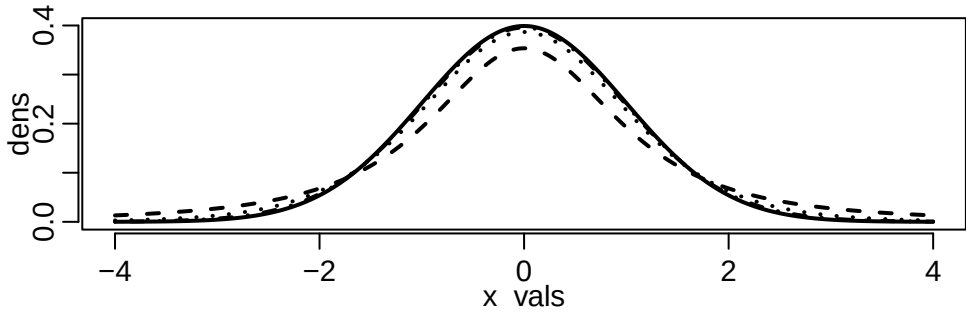
the t-distribution with $n - 1$ degrees of freedom.

Estimating σ introduces additional variability.

Student's t-distribution

The *t*-distribution is more similar to the normal the larger its degrees of freedom.

```
x_vals <- seq(-4, 4, length.out = 1000)
plot(x_vals, dnorm(x_vals), type = "l", lwd = 2, ylab = "dens")
lines(x_vals, dt(x_vals, 2), lty = 2, lwd = 2)
lines(x_vals, dt(x_vals, 8), lty = 3, lwd = 2)
lines(x_vals, dt(x_vals, 32), lty = 4, lwd = 2)
```



Student's t-distribution

The tails of the *t*-distribution are heavier than those of the standard normal.

The probability of observing something far from the mean is larger.

Do the same thing as before to get confidence interval.

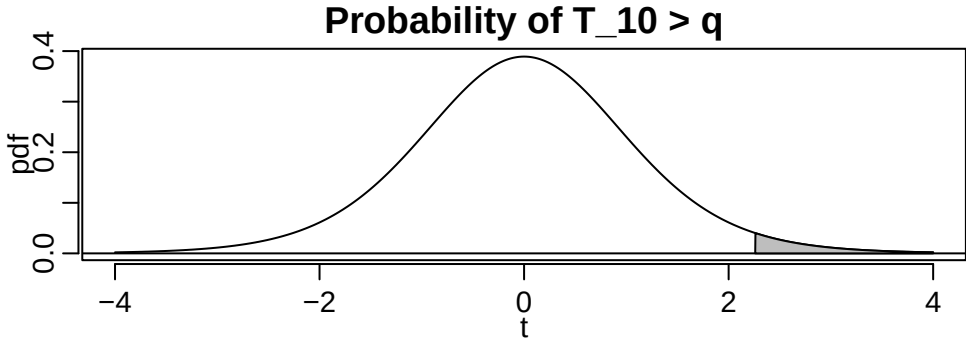
Step 1

Find a q such that

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

How?

The shaded region should have area $\alpha/2$.




```
qt(0.025, df = 9, lower = F) # alpha = 0.05 and n = 10
```

```
## [1] 2.262157
```

Step 2

Solve to get μ in the middle.

$$P\left(\bar{X} - qS/\sqrt{n} \leq \mu \leq \bar{X} + qS/\sqrt{n}\right) = 1 - \alpha$$

We are $100 \times (1 - \alpha)\%$ confident μ is in

$$[\bar{x} - qs/\sqrt{n}, \bar{x} + qs/\sqrt{n}]$$

Example

```
n <- sum(!is.na(penguins$flipper_length_mm))
x_bar <- mean(penguins$flipper_length_mm, na.rm = T)
se <- sd(penguins$flipper_length_mm, na.rm = T) / sqrt(n)
x_bar + c(-1, 1) * qt(0.975, n - 1) * se
```

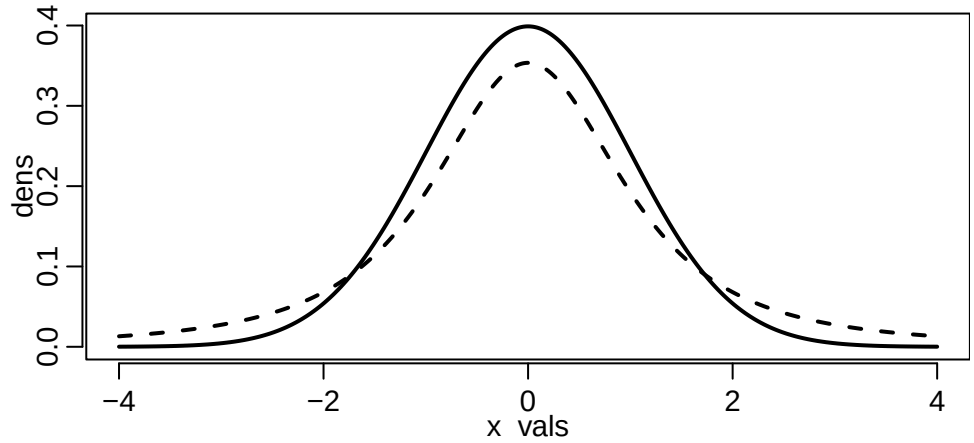
```
## [1] 199.4196 202.4108
```

```
x_bar + c(-1, 1) * 1.96 * se
```

```
## [1] 199.4249 202.4055
```

The *t*-based confidence interval is longer—why?

Because it has heavier tails.



The t-distribution is only exactly correct when data are normal but it is standard practice to use it also for non-normal data.

Recall, it is more conservative than the using the normal distribution, and the CLT says the normal distribution is approximately correct for large n.

6. Hypothesis testing

Suppose we have a hypothesis that mean flipper length is 200 mm.

A hypothesis test tells you whether there is evidence against this hypothesis or not.

Intuition

Reject the hypothesis that $\mu = 200$ if $|\bar{x} - 200|$ is large.

But what is large?

Intuition

$|\bar{x} - 200|$ is large if it is unlikely to observe when $\mu = 200$.

Specifically, suppose that $\mu = 200$ and calculate the probability of observing

$$|\bar{X} - 200| \geq |\bar{x} - 200|$$

If this probability is small, we do not believe that $\mu = 200$.

Example

Suppose I flip a coin and it comes up heads 1000 times in a row. Do you believe the coin is fair?

The probability of observing 1000 heads in a row with a fair coin is $1/2^{1000} \approx 10^{-300}$ so you probably do not believe the coin is fair.

The hypothesis we are testing is called the null hypothesis, and the hypothesis that $\mu \neq 200$ is called the alternative hypothesis.

The probability

$$P(|\bar{X} - \mu_0| \geq |\bar{x} - \mu_0|)$$

is called a p-value.

We reject the null hypothesis if the p-value is small.

A common threshold (significance level) is to reject if $p < 0.05$ but there is no particularly good reason for this.

How to compute the p-value?

Suppose $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ and σ^2 is known.

If the null hypothesis is true, $\mu = \mu_0$.

The p-value is

$$P\left(\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} \geq \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right)$$

or

$$P(|Z| \geq |z|)$$

We call z a z-statistic or z-score, and the corresponding test a z-test.

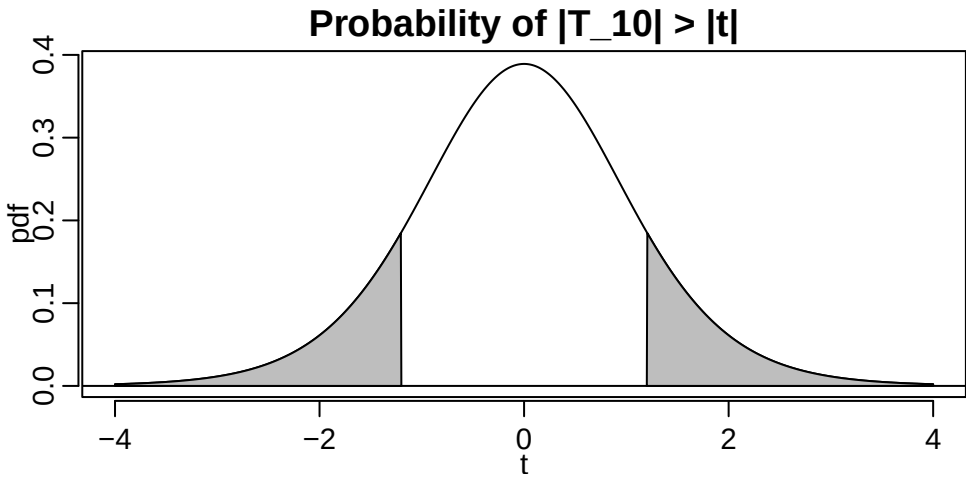
When σ is unknown, we cannot use the z-statistic, so we use the t-statistic:

$$t = (\bar{x} - \mu_0)/\sqrt{s^2/n}.$$

The p-value of the t-test for the null hypothesis that $\mu = \mu_0$ is

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} \geq |t|\right) = P(|T_{n-1}| \geq |t|)$$

We can find these probabilities as we did with the confidence intervals.


Probability of |T_10| > |t|
pdf
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
t

Example

```
n <- sum(!is.na(penguins$flipper_length_mm))
t <- mean(penguins$flipper_length_mm, na.rm = T) - 200
t <- t / sqrt(var(penguins$flipper_length_mm, na.rm = T) / n)
t
```

```
## [1] 1.20363
```

```
2 * pt(-abs(t), df = n - 1)
```

```
## [1] 0.2295675
```

```
pt(-abs(t), df = n - 1) + pt(abs(t), df = n - 1, lower = F)
```

```
## [1] 0.2295675
```

Example

If the flipper length of a randomly sampled penguin is normally distributed with mean 200, then the probability of observing a sample average flipper length at least 0.91 standard errors larger or smaller than 200, in a sample of size 342, is 0.23. Thus, we do not reject the null hypothesis that the mean is 200 on the 5 % level.

Summary

If σ is known, a z-test for the null hypothesis that $\mu = \mu_0$ has p-value

$$P\left(|Z| \geq \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}}\right) = 2 * \texttt{pnorm}(\texttt{abs}(\texttt{z}), \texttt{lower} = \texttt{F})$$

A t-test has p-value

$$P\left(|T_{n-1}| \geq \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}}\right) = 2 * \texttt{pt}(\texttt{abs}(\texttt{t}), \texttt{df} = \texttt{n} - 1, \texttt{lower} = \texttt{F})$$

We reject the null hypothesis if the p-value is below the significance level α.

Connection to confidence intervals

We reject a z or t test on the $100 \times \alpha\%$ significance level if μ_0 is outside corresponding $100 \times (1 - \alpha)\%$ confidence interval.

You may know p-values have been, and still are, a topic of some debate, see

`https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf`

Many of the concerns can be addressed by reporting confidence intervals instead of p-values.

General procedure

Many popular tests are constructed using the same intuition as the z and t-tests:

Find a test statistic (a function of the data) with the following properties:

1. You know its distribution when the null hypothesis is true
2. It tends to be larger when the null hypothesis is false than when it is true

Reject the null hypothesis if, when the null hypothesis is true, the probability of observing something at least as large as what you observed (the p-value) is small.

You need point 1 to compute the p-value.

Suppose $X_1, \ldots, X_n$ are independent $N(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_n$ are independent $N(\mu_y, \sigma_y^2)$.

If the two samples are independent, of the same size, and $\sigma_x^2 = \sigma_y^2$, then the t-statistic for the null hypothesis that $\mu_x = \mu_y$ is

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{2/n}},$$

where $s_p^2 = (n-1)(s_x^2 + s_y^2)/(2n-2)$.

The p-value is

$$P(|T_{2n-2}| > |t|)$$

Different two-sample test of means

There are many other options: different sample sizes, unequal variances, dependent samples.

```
attach(penguins)
t.test(flipper_length_mm[species == "Adelie"], flipper_length_mm[species == "Gentoo"],
       paired = F, var.equal = F)
```

```
##
##  Welch Two Sample t-test
##
## data:  flipper_length_mm[species == "Adelie"] and flipper_length_mm[species == "Gentoo"]
## t = -34.445, df = 261.75, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -28.79018 -25.67652
## sample estimates:
## mean of x mean of y
##  189.9536  217.1870
```

Paired t-test

The t-test is paired if X_i and Y_i are dependent, but independent of all other.

This can happen, for example, if X_i and Y_i are measurements from the same individual (penguin) at different time points.

If we want to test whether X_i and Y_i have the same mean, the paired t-test is equivalent to creating $Z_i = X_i - Y_i$ and testing whether Z_i has mean zero.

Suppose we have $k \geq 2$ populations and we want to test the null hypothesis that they all have the same mean.

Let $X_{i,j}$ denote the ith observation in the jth group, $i = 1, \ldots, n_j$, $j = 1, \ldots, k$.

Define the within-groups sum of squares

$$SS_W = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{i,j} - \bar{X}_j)^2,$$

where

$$\bar{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{i,j}$$

Define the between-groups sum of squares

$$SS_B = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{X}_j - \bar{X})^2 = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2,$$

where

$$\bar{X} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{i,j}, \quad n = \sum_{j=1}^{k} n_j.$$

Intuition

Variation within groups (SS_W) will be large if the variance of $X_{i,j}$ is large, regardless of what the means are.

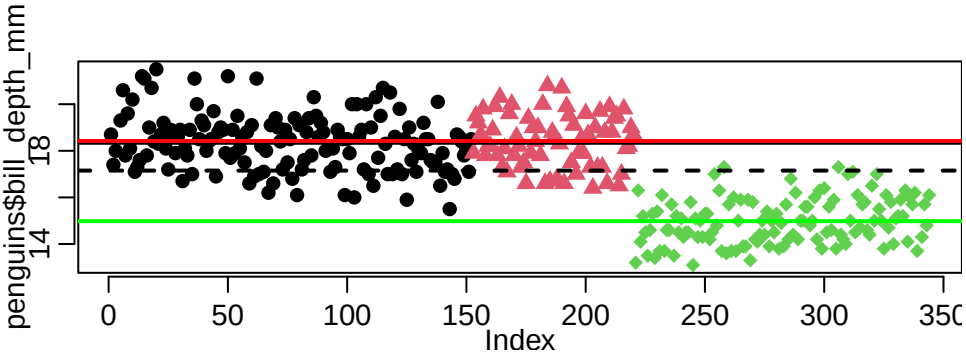
Variation between groups (SS_B) will be large both if the varaince of $X_{i,j}$ is large, but also if the group means differ much from the overall mean; that is, if the null hypothesis is false.

Thus, we should probably reject the null hypothesis if SS_B is large in comparison to SS_W.

We just need to find a test statistic that satisfies this requirement and whose distribution we can calculate.

Example

```
penguins <- penguins[order(penguins$species), ]
plot(penguins$bill_depth_mm,
     col = penguins$species,
     pch = c(16, 17, 18)[penguins$species])
```



Calculating sums of squares

```
x_bar <- mean(penguins$bill_depth_mm, na.rm = T)
penguins %>% group_by(species) %>%
summarize(x_bar_j = mean(bill_depth_mm, na.rm = T),
          n_obs = length(bill_depth_mm),
          n_na = sum(is.na(bill_depth_mm)),
          ss_w = sum((bill_depth_mm - x_bar_j)^2, na.rm = T),
          ss_b = (n_obs - n_na) * (x_bar_j - x_bar)^2)
```

```
## # A tibble: 3 x 6
##   species   x_bar_j n_obs  n_na  ss_w  ss_b
##   <fct>       <dbl> <int> <int> <dbl> <dbl>
## 1 Adelie       18.3   152     1 222.   216.
## 2 Chinstrap    18.4    68     0  86.4  110.
## 3 Gentoo       15.0   124     1 117.   579.
```

```
222 + 86 + 117;  216 + 110 + 579
```

```
## [1] 425
```

```
## [1] 905
```

Intuitively, we want to reject the null hypothesis if 905 is large in comparison to 425.

How to decide what's large enough to reject? p-value!

If the $X_{i,j}$ are independent $N(\mu_j, \sigma^2)$, the statistic

$$\frac{SS_B/(k-1)}{SS_W/(n-k)}$$

has an F-distribution with $k-1$ and $n-k$ degrees of freedom when the null hypothesis is true.

Thus, we reject if the probability of observing something at least

$$\frac{905/(3-1)}{425/(342-3)} \approx 360$$

is small.

The F-distribution in R

```
pf(360, 2, 339, lower = F)
```

```
## [1] 1.409236e-84
```

You don't need to calculate that in this complicated way, of course.

```
anova(lm(bill_depth_mm ~ species, data = penguins))
```

```
## Analysis of Variance Table
##
## Response: bill_depth_mm
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     2 903.97  451.98  359.79 < 2.2e-16 ***
## Residuals 339 425.87    1.26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

If $X_{i,j}$, $i = 1, \ldots, n_j$, $j = 1, \ldots, k$ are independent $N(\mu, \sigma^2)$,

then

$$F = \frac{\sum_{j=1}^{k} n_j (\bar{X}_j - \bar{X})^2 / (k-1)}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_{i,j} - \bar{X}_j)^2 / (n-k)}$$

has an F-distribution with $k-1$ and $n-k$ degrees of freedom. We reject the null hypothesis that $\mu_1 = \cdots = \mu_k$ on the $100 \times \alpha\%$ level if we observe $F = f$ and the p-value

$$p = P(F \geq f) \leq \alpha$$

7. Size and power of tests

If the null hypothesis is true, then the p-value is uniformly distributed.

Example

Let $X_1, \ldots, X_n$ be independent $N(\mu_0, \sigma^2)$ and $T = (\bar{X} - \mu_0)/(S/\sqrt{n})$.

Let $F(t)$ be the cdf of the t-distribution with $n - 1$ degrees of freedom.

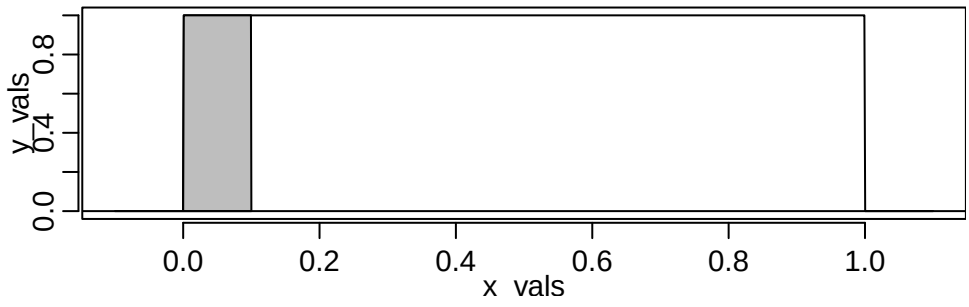
Recall, if we observe $T = t$, the p-value for the null hypothesis that $\mu = \mu_0$, is $P(|T| \geq |t|) = 2F(-|t|) = 2(1 - F(|t|))$.

Fact: The random variable $W = 2F(-|T|)$ has a uniform distribution on $(0, 1)$.

This means:

If you are yet to sample from a population where the null hypothesis is true, the distribution of the yet to be calculated p-value is uniform on $(0, 1)$.

This means that the probability that you will reject the null hypothesis even though it is true, i.e. the probability that the random p-value is less than α, is in fact equal to α!



Size or type one error rate

We call α the size or type 1 error rate of the test.

A test that has $\alpha = 0$ is not useful because it will never reject the null hypothesis, even when it is false.

We have to accept some type one error rate.

Power and type two error

Another error one can make is to not reject the null hypothesis when in fact it is false.

This is called type 2 error.

The probability of not making a type 2 error is called the power of the test.

To have high power, you need to reject often. But if you reject often, you are more likely to make a type 1 error:

> *There is a trade-off between the probabilities of making type 1 and type 2 errors; one has to balance size and power.*

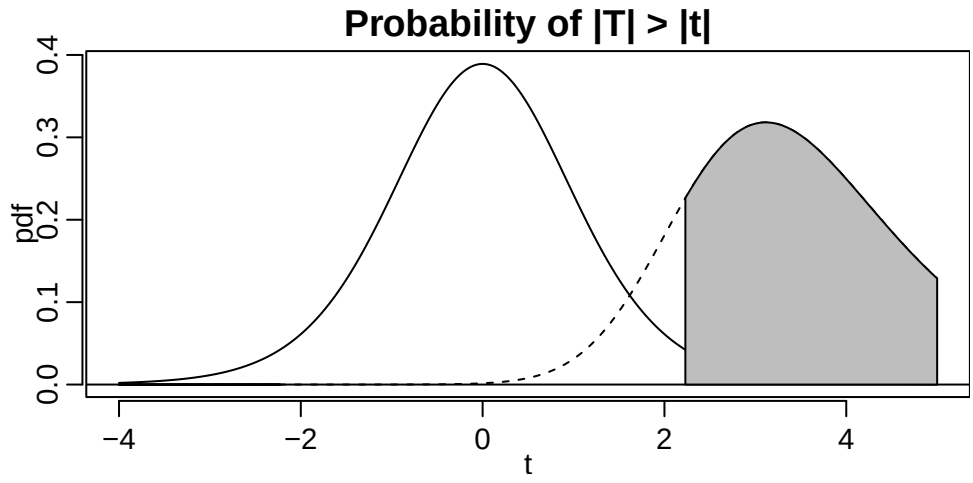
Recall the test statistic for testing $\mu = 0$:

$$T = \frac{\bar{X}}{\sqrt{S^2/n}} = \frac{\sqrt{n}\bar{X}}{S}$$

We know that when the null hypothesis is true, so $X_i \sim N(0, \sigma^2)$, then this test statistic has a t-distribution, which has mean zero.

What about when the null hypothesis is false? Suppose for example that $E(X_i) = \mu > 0$. Then the numerator has mean $E(\sqrt{n}\bar{X}) = \sqrt{n}\mu > 0$.

Thus, we expect that T will often be larger than when $\mu = 0$.


Probability of |T| > |t|
pdf
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
t

Power against an alternative

```
pt(-2.23, 10) + pt(2.23, 10, lower = F)
```

```
## [1] 0.04984247
```

```
pt(-2.23, 10, ncp = sqrt(11)) + pt(2.23, 10, lower = F, ncp = sqrt(11))
```

```
## [1] 0.8471092
```

The power against the alternative that $\mu = 1$ is ≈ 0.85.

Power calculation

You should understand power, but you can calculate it using R.

```
power.t.test(n = 11, delta = 1, sd = 1, sig.level = 0.05, type = "one.sample")
```

```
##
##      One-sample t test power calculation
##
##               n = 11
##           delta = 1
##              sd = 1
##       sig.level = 0.05
##           power = 0.8475297
##     alternative = two.sided
```

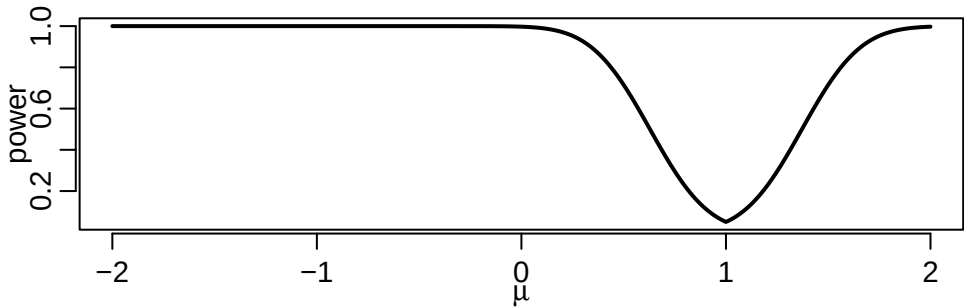
The probability of rejecting is greater when the null hypothesis is false; every reasonable test has this property.

Exercise

For the t-test with 20 degrees of freedom, size 0.1, and null hypothesis $\mu = 1$, plot the power as a function of μ. Consider 1000 values of μ from -2 to 2 and assume $\sigma = 1$.

Plotting power

```
mu <- seq(-2, 2, length.out = 1000); mu0 <- 1
delta <-  mu - mu0
plot(mu, power.t.test(n = 21, delta = delta, sd = 1, sig.level = .1,
  type = "one.sample")$power, type = "l", lwd = 2,
  xlab = expression(mu), ylab = "power")
```



Summary: Size and Power

The probability of rejecting when the null hypothesis is true is the size of the test, also known as level and type 1 error rate

The probability of rejecting when the null hypothesis is false is the power of the test; the type 2 error rate is one minus the power

The power should always be as large as the size, and should increase the further the truth is from the null hypothesis

8. Non-parametric tests

So far we have covered:

- One sample t-test: test of mean for approximately normal popualtion
- Two-sample t-test: test of equality of means of two approximately normal populations
- ANOVA: test for equality of means of several approximately normal populations.

The CLT says these tests can be useful even if data are non-normal.

But what to do if n is small and data non-normal?

Motivation

A non-parametric tests makes fewer assumptions about the underlying distribution.

Loosely speaking, a non-parametric test does not assume a particular family of distributions.

For example, a test that works for any distribution with finite variance is non-parametric but one that assumes normality is not.

The sign test

One of the simplest non-parametric tests is the sign test.

Example

You want to test whether working from home affects productivity

Let X_i be the number of scientific articles academic i submitted in 2019 when they were working from the office and let Y_i be the number they submitted in 2020 working from home.

Assume participants are randomly sampled.

The sign test

If we assume Y_i and X_i are normally distributed with constant variance, we can use a paired t-test.

The sign tests lets us test the null hypothesis that $P(Y_i \geq X_i) = 0.5$ (or $>, \leq, <, =$) without assuming normality.

Create the variable

$$Z_i = I(Y_i \geq X_i) = 1 \text{if } Y_i \geq X_i \text{ and } 0 \text{ otherwise.}$$

Notice that Z_i has a Bernoulli distribution with parameter $p = P(Y_i \geq X_i)$.

Thus, under the null hypothesis, $Z_i \sim \mathrm{Ber}(1/2)$.

If p is smaller than $1/2$, we expect to see fewer $Z_i = 1$, and if it is larger we expect to see more.

The sign test

The number of $Z_i = 1$ is $S = \sum_{i=1}^{n} Z_i$.

Under the null hypothesis, $S \sim \mathrm{Bin}(n, 1/2)$, which has mean $n/2$.

Thus, intuition suggests we should reject the null hypothesis when we observe a large value of

$$|S - n/2| \,.$$

We reject if $P(|S - n/2| \geq |s - n/2|) \leq \alpha$.

Suppose we observe $S = 5$ in $n = 15$ observations. The p-value is

$$P(|S - np_0| \geq |5 - p_0 n|)$$

calculated assuming $p = p_0 = 1/2$.

```
pbinom(5, prob = 0.5, size = 15) + pbinom(9.5, prob = 0.5, size = 15, lower = F)
```

```
## [1] 0.3017578
```

We do not reject on the 5% level.

The sign test

As a special case, you can test the null hypothesis that the median of the X_i are equal to c.

To test this, note that $Z_i = I(X_i \leq c) \sim \mathrm{Ber}(p)$ under the hull hypothesis.

```
# Test if median of Exp(1) is = 1
x <- rexp(100)
z <- 1 * (x <= 1)
binom.test(x = sum(z), n = length(z), p = 0.5)
```

```
##
##  Exact binomial test
##
## data:  sum(z) and length(z)
## number of successes = 70, number of trials = 100, p-value = 7.85e-05
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.6001853 0.7875936
## sample estimates:
## probability of success
##                    0.7
```

Testing quantiles and the Binomial test

You can test any quantile the same way:

```
# Test if 0.1th quantile of Exp(1) is = 0.1
x <- rexp(100); z <- 1 * (x < 0.1); binom.test(x = sum(z), n = length(z), p = 0.1)
```

```
##
##  Exact binomial test
##
## data:  sum(z) and length(z)
## number of successes = 16, number of trials = 100, p-value = 0.0636
## alternative hypothesis: true probability of success is not equal to 0.1
## 95 percent confidence interval:
##  0.09431029 0.24678760
## sample estimates:
## probability of success
##                   0.16
```

In fact, you can test the null hypothesis that any Bernoulli random variable has $p = p_0$ the same way—This is known as the Binomial test.

If we want to test the null hypothesis that the distribution of X_i is symmetric around μ, we can use a Wilcoxon signed rank test.

If we want to test the null hypothesis that the distributions of X_i and Y_i are the same up to a location shift μ, we can use a Wilcoxon rank-sum test.

We can calculate the signed rank test statistic in R:

```
# Is Exp(1) symmetric around 1?
set.seed(5)
x <- rexp(10)
sgn <- sign(x - 1); d <- abs(x - 1)
w = sum(rank(d)[sgn == 1]); w
```

```
## [1] 21
```

How do decide if we reject?

Under the null hypothesis, the expected value of W is $n(n+1)/4 = 55/2$, so we reject if we observe w far from this.

Wilcoxon signed rank test

As before, “far from” is measured by how likely it is to observe something at least that far from $55/2$ in a random sample when the null hypothesis is true.

To compute this (the *p*-value) we need to know the distribution of W under the null hypothesis.

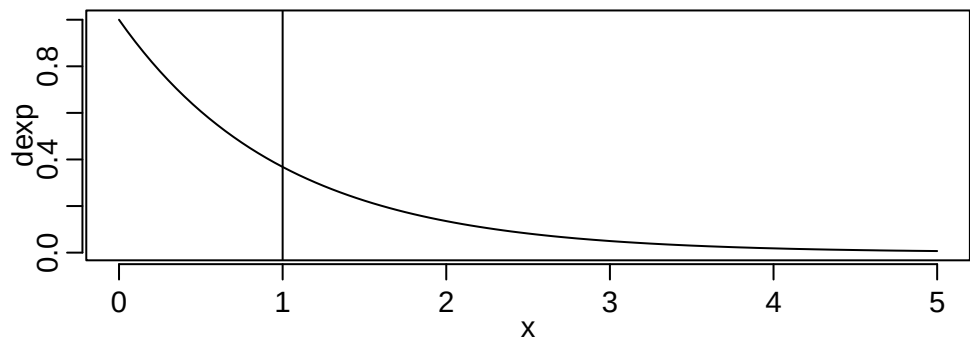
R knows it!

```
wilcox.test(x, mu = 1)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  x
## V = 21, p-value = 0.5566
## alternative hypothesis: true location is not equal to 1
```

The test seems to indicate the exponential with mean one is symmetric around one, but its pdf is clearly not:

```
plot(dexp, 0, 5); abline(v = 1)
```



What is going on?

Power illustration

```r
x <- rexp(1e4)
wilcox.test(x, mu = 1)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  x
## V = 20324825, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 1
```

Testing symmetry around median

What about testing whether Exp(1) is symmetric around its median?

```
z <- rexp(1e3)
m <- qexp(0.5)
wilcox.test(z, mu = m)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  z
## V = 284623, p-value = 0.0001682
## alternative hypothesis: true location is not equal to 0.6931472
```

Paired test

If we have two dependent samples and want to know whether their difference is symmetric around some point μ, can use a paired test:

```r
x <- rexp(10, 1); y <- x + rnorm(10)
wilcox.test(y, x, mu = 0, paired = T)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  y and x
## V = 26, p-value = 0.9219
## alternative hypothesis: true location shift is not equal to 0
```

Paired test

```
wilcox.test(y - x, mu = 0)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  y - x
## V = 26, p-value = 0.9219
## alternative hypothesis: true location is not equal to 0
```

Wilcoxon rank-sum test

Null hypothesis: two independent samples are from the same distribution up to a location shift.

That is, the distribution of X_i is the same as that of $Y_i + \mu$ for some μ.

The test statistic is based on combining the samples and comparing whose ranks tend to be higher:

Implementing the rank-sum test

```
set.seed(1337)
x <- rexp(10) - 1; y <- rnorm(12)
sum(rank(c(x, y))[1:10]) - 10 * (10 + 1) / 2
```

```
## [1] 48
```

```
wilcox.test(x, y)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  x and y
## W = 48, p-value = 0.4562
## alternative hypothesis: true location shift is not equal to 0
```

Again we do not reject even though the distributions are different.

Power illustration

```r
x <- rexp(1000) - 1; y <- rnorm(1200)
wilcox.test(x, y)
```

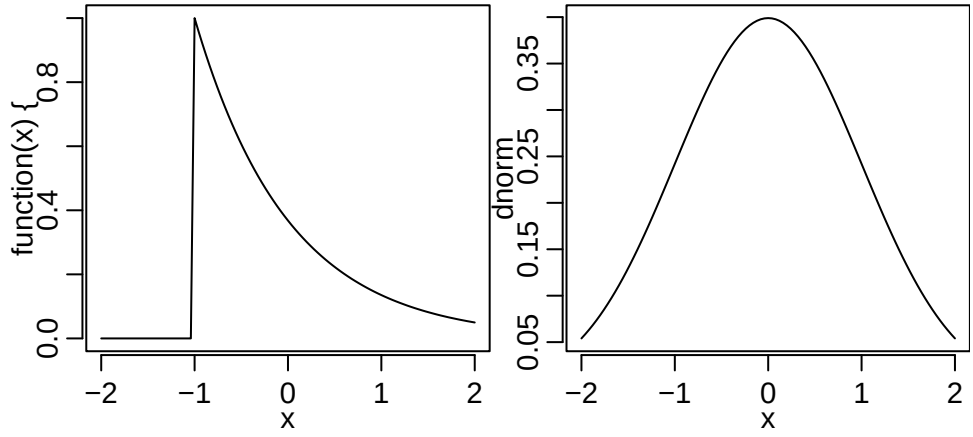
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 576573, p-value = 0.1143
## alternative hypothesis: true location shift is not equal to 0
```

Power illustration

```r
x <- rexp(10000) - 1; y <- rnorm(12000)
wilcox.test(x, y)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  x and y
## W = 55771736, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

Comparing distributions



Both have mean 0 and variance 1.

Suppose we have $X_{i,j}$, $i = 1, \ldots, n_j$, $j = 1, \ldots, k$, and we want to test the null hypothesis that $X_{i,j}$ and $X_{i,j'}$ have the same distribution for all j, j'.

Then we can use the Kruskal–Wallis test.

If there are no ties (no identical observations) the test statistic is based on

$$\sum_{j=1}^{k} \frac{\overline{r}_j^2}{n_j},$$

where $\overline{r}_j$ is the average rank of the observations in group j.

This will be large if there is a lot of variation in the $\overline{r}_j$.

The Kruskal–Wallis test in R

```
my_data <- data.frame(y = c(rexp(10) - 1, rnorm(10), runif(10, -sqrt(3), sqrt(3))),
                      group = as.factor(rep(c("exp", "norm", "uni"), each = 10)))
str(my_data)
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ y    : num  1.254 -0.446 0.268 2.357 -0.93 ...
##  $ group: Factor w/ 3 levels "exp","norm","uni": 1 1 1 1 1 1 1 1 1 1 ...
```

```
kruskal.test(x = my_data$y, g = my_data$group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  my_data$y and my_data$group
## Kruskal-Wallis chi-squared = 3.631, df = 2, p-value = 0.1628
```

The Kruskal–Wallis test in R

```
my_data <- data.frame(y = c(rexp(1e3) - 1, rnorm(1e3), runif(1e3, -sqrt(3), sqrt(3))),
                      group = as.factor(rep(c("exp", "norm", "uni"), each = 1e3)))
kruskal.test(x = my_data$y, g = my_data$group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  my_data$y and my_data$group
## Kruskal-Wallis chi-squared = 8.212, df = 2, p-value = 0.01647
```

Summary of tests

Name	Null	Assumes
One-sample t-test	$\mu = \mu_0$	Random sample, normal dist.
Two-sample t-test	$\mu_1 = \mu_2$	Random samples, normal dist.
ANOVA	$\mu_1 = \cdots = \mu_k$	Random samples, normal dist.
Binomial test	$p = p_0$	Random sample, Bernoulli dist.
Wilcoxon signed rank test	Symmetric around μ_0	Random sample
Wilcoxon rank-sum test	Two distributions the same	Random sample
Kruskal–Wallis rank-sum test	$k \geq 2$ distributions the same	Random sample

9. Contingency tables and association of factors

A contingency table shows counts of combinations of factors.

```
table(penguins$species, penguins$island)
```

```
##
##              Biscoe Dream Torgersen
##   Adelie         44    56        52
##   Chinstrap       0    68         0
##   Gentoo        124     0         0
```

Contingency tables in R

```
xtabs(~species + island + sex, data = penguins)
```

```
## , , sex = female
##
##            island
## species     Biscoe Dream Torgersen
##   Adelie        22    27        24
##   Chinstrap      0    34         0
##   Gentoo        58     0         0
##
## , , sex = male
##
##            island
## species     Biscoe Dream Torgersen
##   Adelie        22    28        23
##   Chinstrap      0    34         0
##   Gentoo        61     0         0
```

Testing association

We can test whether there is an association between the factors.

The test statistic is based on expected and observed counts.

In particular, we reject the null hypothesis that the variables are independent if the counts we see are far from the expected counts.

What does it mean?

Expected counts under the null hypothesis

Assuming independence, the expected count of Adelie penguins on Torgersen island in a sample of 344 is

$$344 \times P(\text{Adelie} \cap \text{Torgersen}) = 344 \times P(\text{Adelie})P(\text{Torgersen}),$$

We do not know the true probability that a randomly selected penguin is from Torgersen island $P(\text{Torgersen})$, so we estimate it.

Expected counts under the null hypothesis

Estimate $P(\text{Torgersen})$ by the sample proportion

$$\hat{P}(\text{Torgersen}) = 52/344.$$

Similarly,

$$\hat{P}(\text{Adelie}) = 152/344.$$

Therefore, "expected" count is

$$n\hat{P}(\text{Torgersen})\hat{P}(\text{Adelie}) = 344(152/344)(52/344) \approx 23.$$

Expected counts under the null hypothesis

Now we do the same for every cell in the table:

```
prop_species <- prop.table(table(penguins$species))
prop_island  <- prop.table(table(penguins$island))
(exp_counts <- 344 * outer(prop_species, prop_island, "*"))
```

```
##
##               Biscoe    Dream Torgersen
##   Adelie    74.23256 54.79070  22.97674
##   Chinstrap 33.20930 24.51163  10.27907
##   Gentoo    60.55814 44.69767  18.74419
```

Chi-square statistic and distribution

The test statistic is

```
obs_counts <- xtabs(~species + island, data = penguins)
(chi_stat <- sum((exp_counts - obs_counts)^2 / exp_counts))
```

```
## [1] 299.5503
```

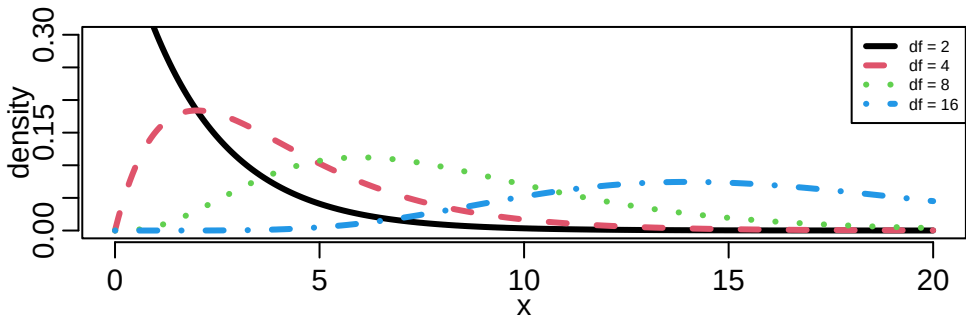
In random sampling, it has approximately a χ^2-distribution with $(n_r - 1)(n_c - 1)$ degrees of freedom under the null hypothesis.

```
pchisq(chi_stat, df = (nrow(exp_counts) - 1) * (ncol(exp_counts) - 1), lower = F)
```

```
## [1] 1.354574e-63
```

Chi-square densities

```
x <- seq(0, 20, length.out = 1000)
plot(x, dchisq(x, 2), type = "l", ylim = c(0, 0.3), lwd = 3, ylab = "density", cex = 0.8)
lines(x, dchisq(x, 4), lty = 2, col = 2, lwd = 3)
lines(x, dchisq(x, 8), lty = 3, col = 3, lwd = 3)
lines(x, dchisq(x, 16), lty = 4, col = 4, lwd = 3)
legend("topright", legend = paste("df =", 2^(1:4)), col = 1:4, lty = 1:4, lwd = 3, cex = 0.5)
```



Implementing the chi-square test

```
  summary(obs_counts)
```

```
## Call: xtabs(formula = ~species + island, data = penguins)
## Number of cases in table: 344
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 299.55, df = 4, p-value = 1.355e-63
```

```
  chisq.test(table(penguins$species, penguins$island))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(penguins$species, penguins$island)
## X-squared = 299.55, df = 4, p-value < 2.2e-16
```

The distribution of the χ^2 test statistic is only approximate, and the approximation can be bad for small samples and unbalanced data.

There are two special cases with exact tests available:

- Fisher's exact test: 2×2 contingency table; test whether factors are independent (same as χ^2-test)
- Binomial test: 1×2 contingency table; test whether the true proportions of a factor with two levels is equal to some null hypothesis value p_0.

Contingency tables in R

```
flip_adelie <- penguins$flipper_length_mm[penguins$species == "Adelie"]
long <- flip_adelie > mean(flip_adelie, na.rm = T)
(flip_sex_tab <- xtabs(~long + sex[species == "Adelie"], data = penguins))
```

```
##        sex[species == "Adelie"]
## long    female male
##   FALSE     43   19
##   TRUE      30   54
```

```
chisq.test(flip_sex_tab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  flip_sex_tab
## X-squared = 14.83, df = 1, p-value = 0.0001177
```

```
fisher.test(flip_sex_tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  flip_sex_tab
## p-value = 0.0001021
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.915504 8.744560
## sample estimates:
## odds ratio
##   4.031355
```

Odds of events

The odds of an event A is

$$P(A)/P(A^c) = \frac{P(A)}{1 - P(A)}$$

The odds ratio indicates how much another event B affects odds.

The odds of A given B is

$$\frac{P(A \mid B)}{1 - P(A \mid B)} = \frac{P(A \mid B)}{P(A^c \mid B)} = \frac{P(A \mid B)P(B)}{P(A^c \mid B)P(B)} = \frac{P(A \cap B)}{P(A^c \cap B)}$$

The odds of A given B^c is

$$\frac{P(A \mid B^c)}{1 - P(A \mid B^c)} = \frac{P(A \mid B^c)}{P(A^c \mid B^c)} = \frac{P(A \cap B^c)}{P(A^c \cap B^c)}$$

The odds ratio is

$$\frac{P(A \mid B)/P(A^c \mid B)}{P(A \mid B^c)/P(A^c \mid B^c)} = \frac{P(A \cap B)/P(A^c \cap B)}{P(A \cap B^c)/P(A^c \cap B^c)}$$

The odds ratio is one if and only if A and B are independent.

Observed odds ratio

In practice, odds ratio often refers to an observed odds ratio.

Observed odds ratio for A = short flippers and B = female.

```
flip_sex_tab
```

```
##        sex[species == "Adelie"]
## long    female male
##   FALSE     43   19
##   TRUE      30   54
```

```
n <- sum(flip_sex_tab)
numerator <- (43 / n) / (30 / n) # sample p(short and f) / p(long and f)
denominator <- (19 / n) / (54 / n) # sample p(short and m) / p(long and m)
numerator / denominator
```

```
## [1] 4.073684
```

Observed odds ratio

Odds ratio says the odds for short flippers are about 4 times larger in female Adelie penguins than in male Adelie penguins.

Odds ratio of events A and B is odds ratio of events B and A.

```
  (43 / 19) / (30 / 54)
```

```
## [1] 4.073684
```

The ratio of odds of male given short or long flippers

```
  (19 / 43) / (54 / 30)
```

```
## [1] 0.245478
```

All give the same information: be careful with interpretations.

Summary of tests

Name	Null	Assumes
One-sample t-test	$\mu = \mu_0$	Random sample, normal dist.
Two-sample t-test	$\mu_1 = \mu_2$	Random samples, normal dist.
ANOVA	$\mu_1 = \cdots = \mu_k$	Random samples, normal dist.
Binomial test	$p = p_0$	Random sample, Bernoulli dist.
Wilcoxon signed rank test	Symmetric around μ_0	Random sample
Wilcoxon rank-sum test	Two distributions the same	Random sample
Kruskal–Wallis rank-sum test	$k \geq 2$ distributions the same	Random sample
Chi-square test	Factors independent	Random sample

Congrats! You are done!

The rest is bonus material.

Bonus material: when is n large enough?

A common guideline is that a sample mean for Bernoulli is approximately normal if $np \geq 10$ and $n(1-p) \geq 10$.

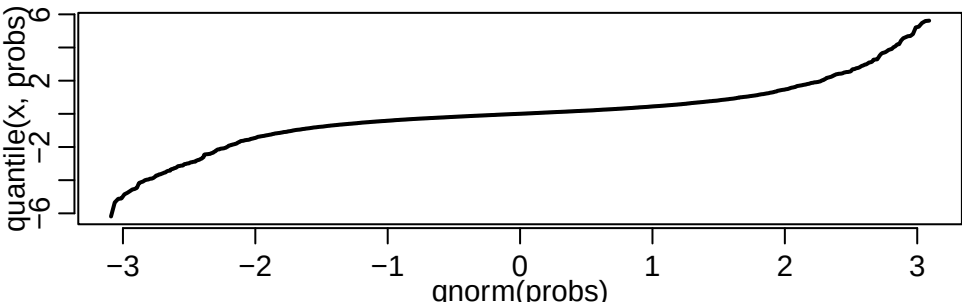
```
table(penguins$species == "Adelie", penguins$body_mass_g > median(penguins$body_mass_g,
                                                                  na.rm = T))
```

```
##
##          FALSE TRUE
##   FALSE     58  133
##   TRUE     118   33
```

Are my data normal?

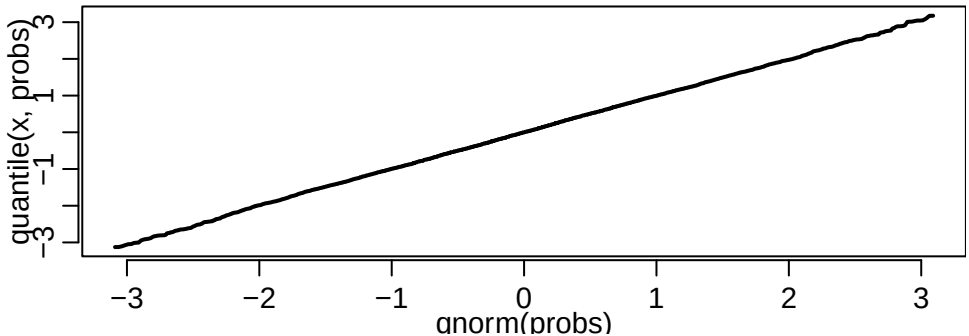
One of the best visualizations are plotting quantiles:

```
x <- rt(10000, df = 2); x <- (x - mean(x)) / sd(x)
probs <- seq(0.001, 0.999, length.out = length(x))
plot(qnorm(probs), quantile(x, probs), type = "l", lwd = 2)
```



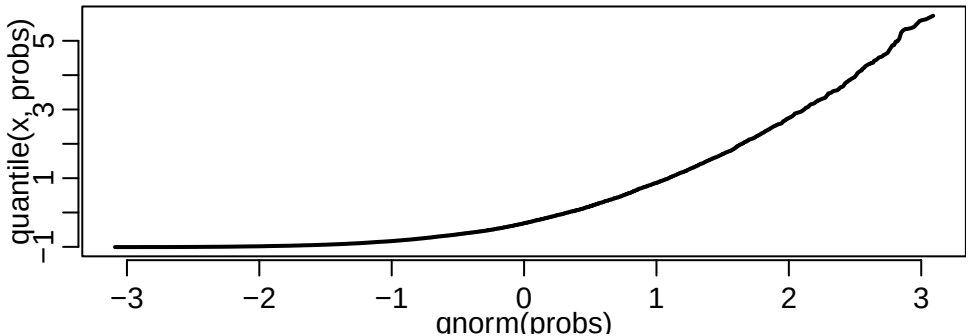
Are my data normal?

```
x <- rnorm(10000, mean = 2, sd = 3); x <- (x - mean(x)) / sd(x)
probs <- seq(0.001, 0.999, length.out = length(x))
plot(qnorm(probs), quantile(x, probs), type = "l", lwd = 2)
```



Are my data normal?

```
x <- rexp(10000, 1); x <- (x - mean(x)) / sd(x)
probs <- seq(0.001, 0.999, length.out = length(x))
plot(qnorm(probs), quantile(x, probs), type = "l", lwd = 2)
```



Common notation

$P(A)$	probability of the event A
X	random variable
x	realization of random variable X
$\mu, E(X)$	mean of a random variable$X$
$\sigma^2, \mathrm{var}(X)$	variance of a random variable X
α	critical level or size of a test
λ	parameter of a distribution
μ_0	null hypothesis value of μ

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

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / (n-1)$$

$$T_{n-1} = (\bar{X} - \mu_0) / (S/\sqrt{n})$$

$\bar{x}$	realization of $\bar{X}$
s^2	realization of S^2
t	realization of T