Forum 3
Data: Probabilities and Distributions
Data Skills for Scientists

2025-09-24

Data

The 3 topics in the first module, ‘Data’ are:

- Navigating Uncertainty;
- Summarising and Visualising; and
- Probabilities and Distributions.

In this forum we will cover the third topic, Probabilities and Distributions.

Key topics for Probabilities and Distributions

- Normal distribution
- Sampling distribution

The Normal Distribution

Example: Body temperature

We have been given a data set which captures the body temperature of 200 individuals and we are interested in analysing these data.

What should we do?

```r
head(df_bodytemp)
```

```
  bodytemp
1 36.57581
2 36.70793
3 37.42348
4 36.82820
5 36.85172
6 37.48603
```

Example: Body temperature

What are the summary statistics for these data?

```r
summary(df_bodytemp$bodytemp)
```

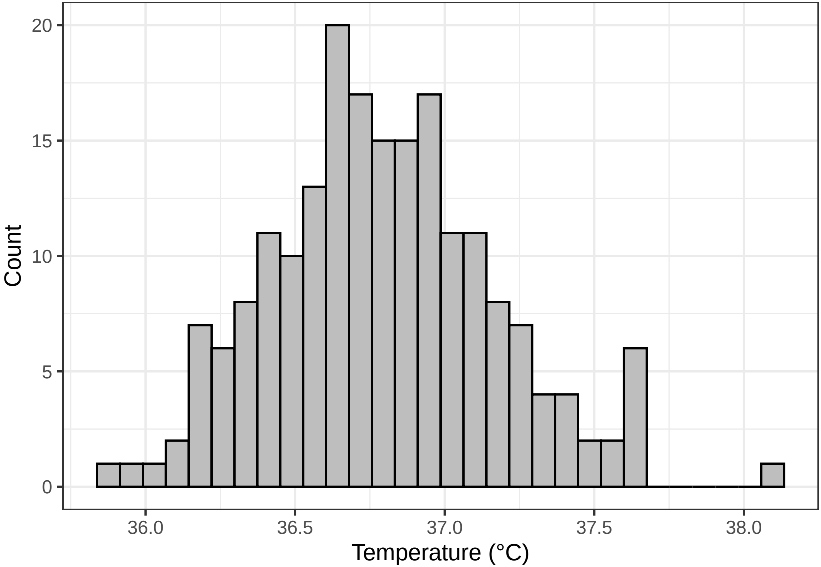
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  35.88   36.55   36.78   36.80   37.03   38.10 
```

```r
sd(df_bodytemp$bodytemp)
```

```
[1] 0.377264
```

Example: Body temperature

How does the data look?



Example: Body temperature

Suppose we have questions about our data:

- How many individuals have a temperature less than or equal to 36.0 (°C)?
- How many individuals have a temperature more than or equal to 36.0 (°C)?
- How many individuals have a temperature between 36.0 (°C) and 36.5 (°C) (inclusive of the end points)?

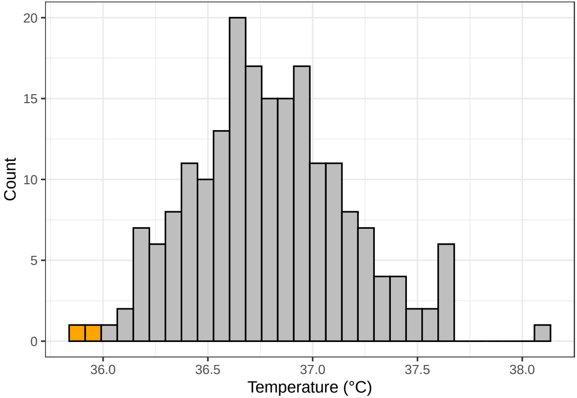
How can we answer these questions?

Example: Body temperature

How many individuals have a temperature less than or equal to 36.0 (°C)?

```
sum(df_bodytemp$bodytemp <= 36)
```

```
[1] 2
```

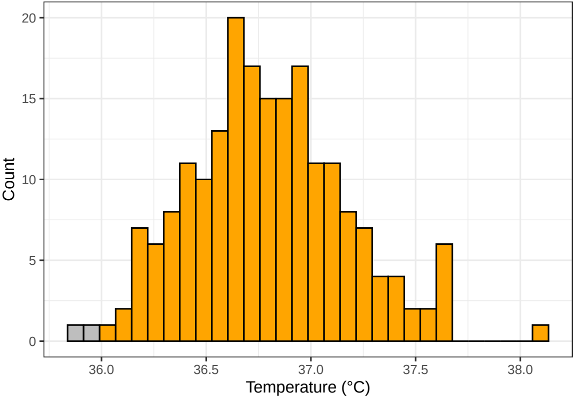


Example: Body temperature

How many individuals have a temperature more than or equal to 36.0 (°C)?

```
sum(df_bodytemp$bodytemp >= 36)
```

```
[1] 198
```

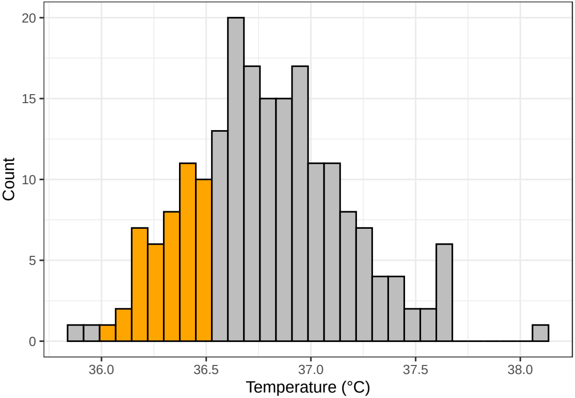


Example: Body temperature

How many individuals have a temperature between 36.0 (°C) and 36.5 (°C) (inclusive of the end points)?

```
sum(between(df_bodytemp$bodytemp, 36, 36.5))
```

```
[1] 38
```



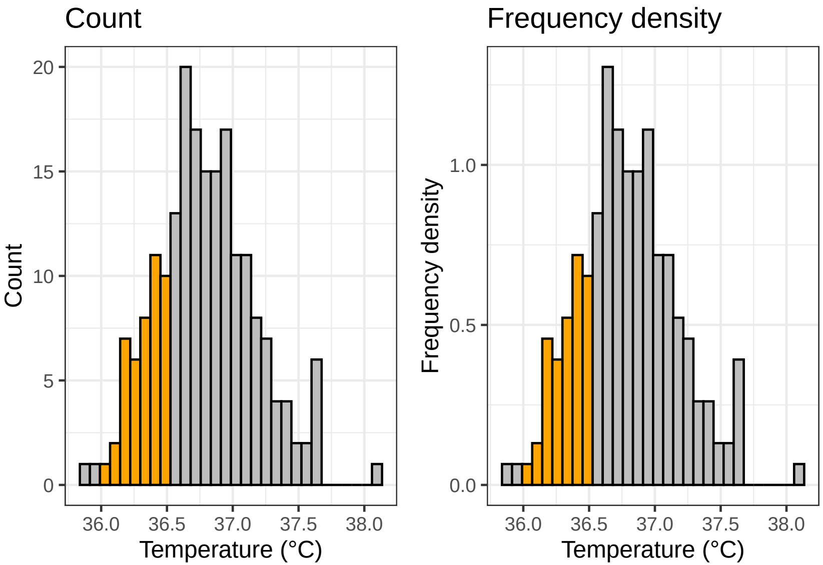
Example: Body temperature

What if instead we ask **what proportion** of individuals of individuals have a temperature between 36.0 (°C) and 36.5 (°C) (inclusive of the end points), rather than **how many** individuals.

From the previous slide, we know that 38 out of the 200 individuals have a temperature between 36.0 (°C) and 36.5 (°C) (inclusive of the end points). So the proportion of the individuals is $38/200 = 0.19$.

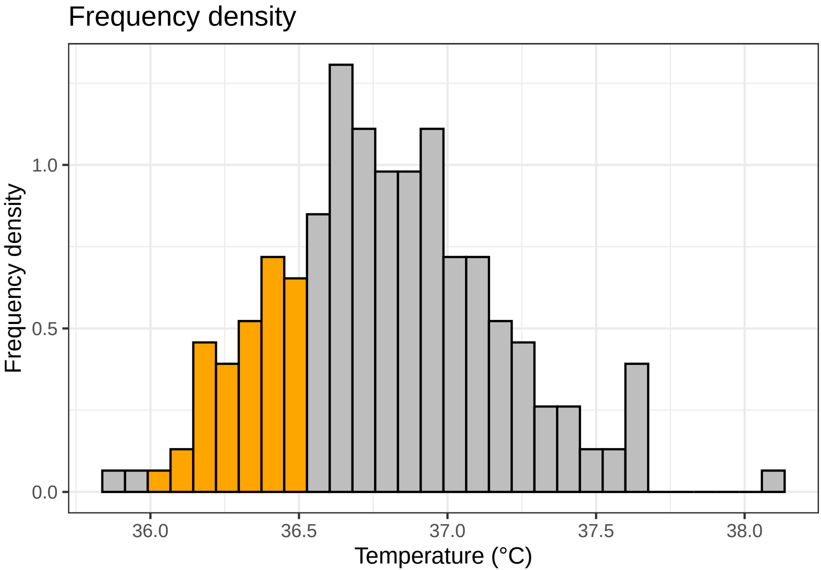
This is equivalent to saying that 19% of the individuals have a temperature between 36.0 (°C) and 36.5 (°C).

Example: Body temperature



Example: Body temperature

The total area of the orange bars on the frequency density plot is equal to 0.19.



Example: Body temperature

Now suppose that these 200 individuals are a representative sample of a larger population of 10,000 individuals.

We now ask:

- What proportion of individuals in the population do we expect to have a temperature less than or equal to 36.0 (°C)?
- What proportion of individuals in the population do we expect to have a temperature more than or equal to 36.0 (°C)?
- What proportion of individuals in the population do we expect to have a temperature between 36.0 (°C) and 36.5 (°C) (inclusive of the end points)?

How could we answer these questions?

Example: Body temperature

Approach 1

- Assume that the pattern shown in the data for the 200 individuals is exactly the same as the pattern that is present in the data for the larger population of 10,000 individuals

Would this work? What are the advantages? What are the limitations?

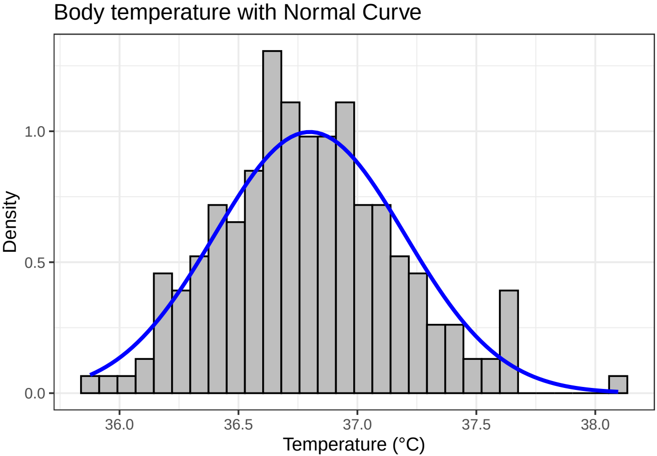
Approach 2

- Assume that the data can be represented by a Normal distribution density curve that is an idealisation of the data that gives the overall pattern but may ignore minor irregularities.

Would this work? What are the advantages? What are the limitations?

Example: Body temperature

The Normal distribution is identified by its population mean μ and population standard deviation σ. Here, $\mu = 36.8$ and $\sigma = 0.4$.



Computing normal distribution probabilities

Applet Source: https://digitalfirst.bfwpub.com/stats_applet/stats_applet_7_norm.html

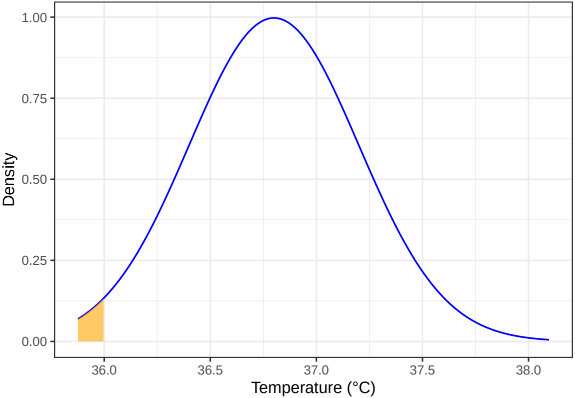
Example: Body temperature

What proportion of individuals in the population do we expect to have a temperature less than or equal to 36.0 (°C)?

We can use R to calculate this proportion.

```r
pnorm(q=36.0,mean=36.8,sd=0.4)
```

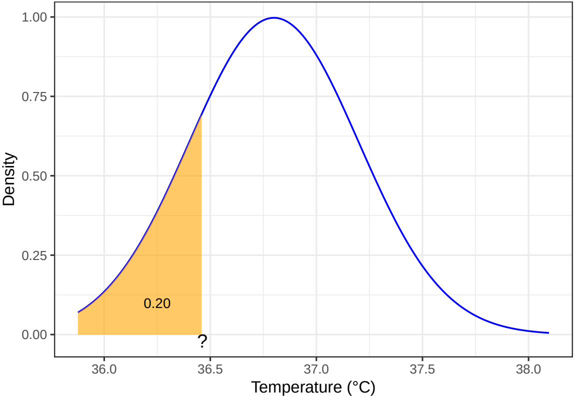
```
[1] 0.02275013
```



Example: Body temperature (Inverse probability calculations)

What is the maximum body temperature to be in the bottom 20% of values?

How can we do this?

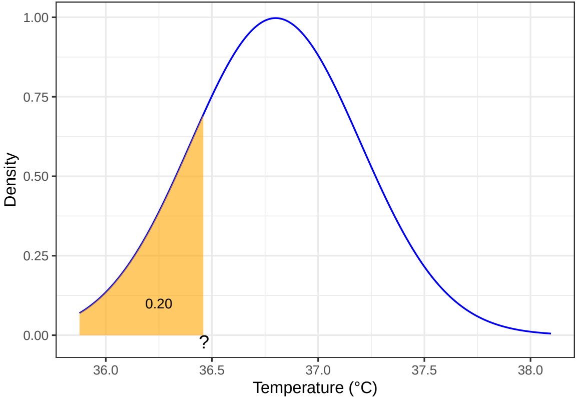


Example: Body temperature (Inverse probability calculations and percentiles)

What is the maximum body temperature to be in the bottom 20% of values?

```r
qnorm(p=0.20,mean=36.8,sd=0.4)
```

```
[1] 36.46335
```



Example: Body temperature (References)

- Longo DL, Fauci AS, Kasper DL, Hauser SL, Jameson J, Loscalzo J, eds. Harrison's Principles of Internal Medicine, 18th ed. New York, NY: McGraw-Hill; 2012
  One of the most respected medicial reference books.

Further reading:

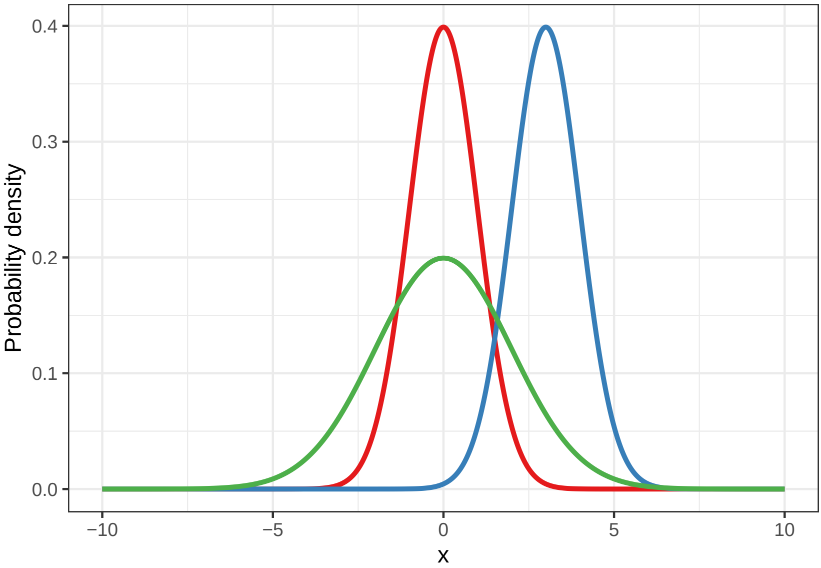
- Geneva II, Cuzzo B, Fazili T, Javaid W. 2019, April. Normal body temperature: a systematic review. In Open forum infectious diseases (Vol. 6, No. 4, p. ofz032). US: Oxford University Press.
  A recent review article discussing how the estimate can vary depending on how the measurement is taken and the population being considered.

Normal distribution key properties

We have already seen that the Normal distribution can be very useful. We now summarise key properties:

- All Normal distribution curves are symmetric, single-peaked, and bell-shaped.
- We can abbreviate the Normal distribution with mean μ and standard deviation σ as $N(\mu, \sigma)$.
- The mean of a Normal distribution is the centre of the symmetric Normal curve.
- The standard deviation describes the flatness, height, and spread.
- Mean $=$ Median $=$ Mode.
- Total area under a Normal distribution curve is equal to 1.

Examples



The 68-95-99.7 rule

In the Normal distribution with mean μ and standard deviation σ:

- Approximately 68% of the observations fall within σ of μ.
- Approximately 95% of the observations fall within 2σ of μ.
- Approximately 99.7% of the observations fall within 3σ of μ.

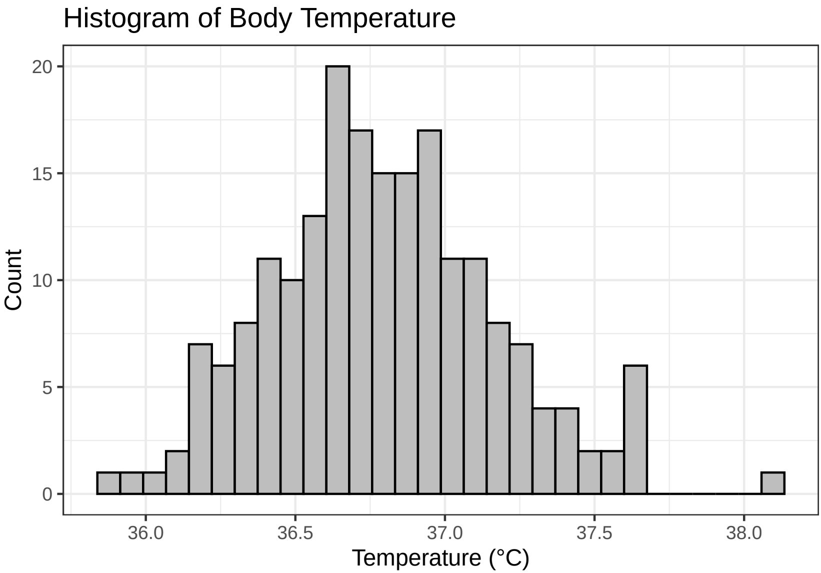
Testing data for normality

Often we assume that the dataset is from a normal distribution. To test the validity of this assumption we can:

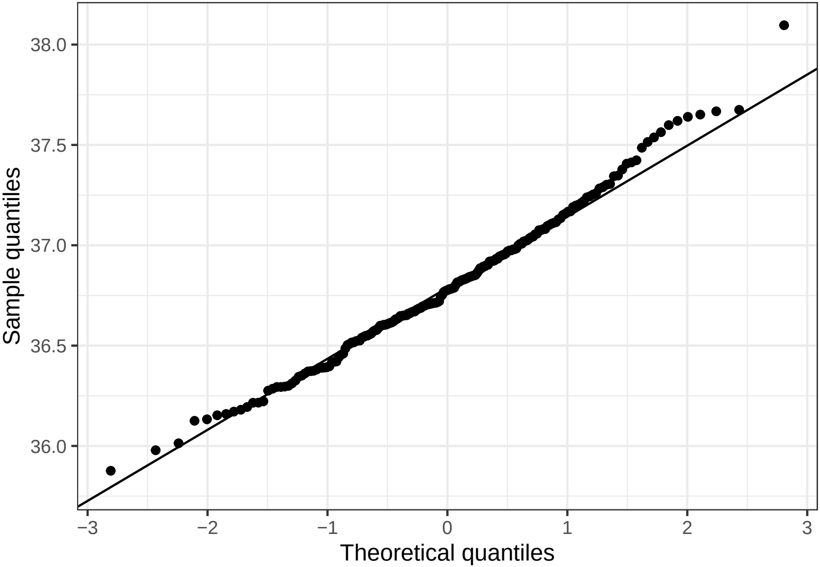
(1) Plot the normal distribution against the histogram (as shown earlier).

(2) Use a **Normal Quantile-Quantile Plot** (or a QQ-plot). If the points closely follow a straight line, then we say that the assumption of normality is reasonable. But, how close is close here?

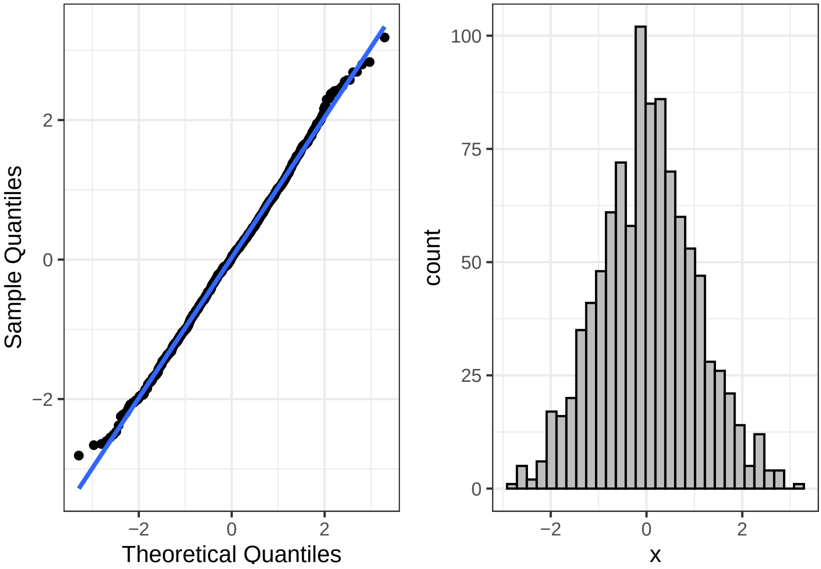
Example: Body temperature



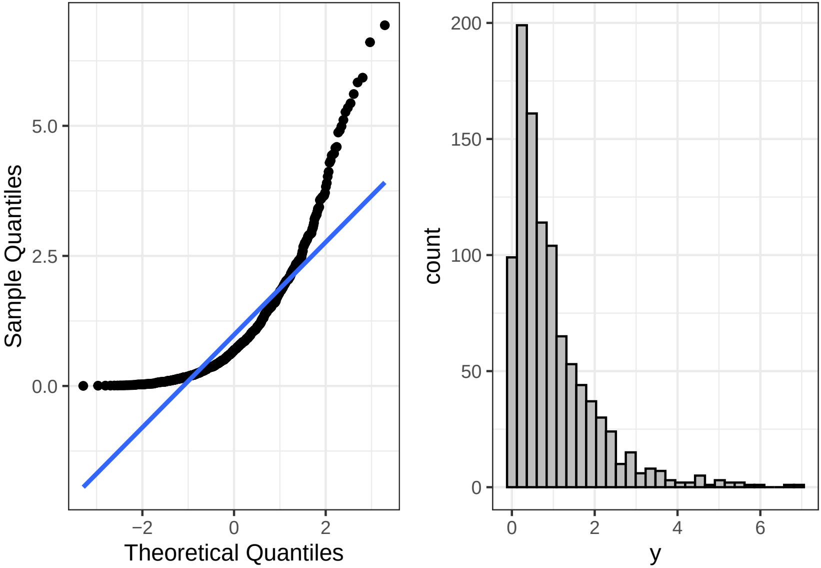
Example: Body temperature



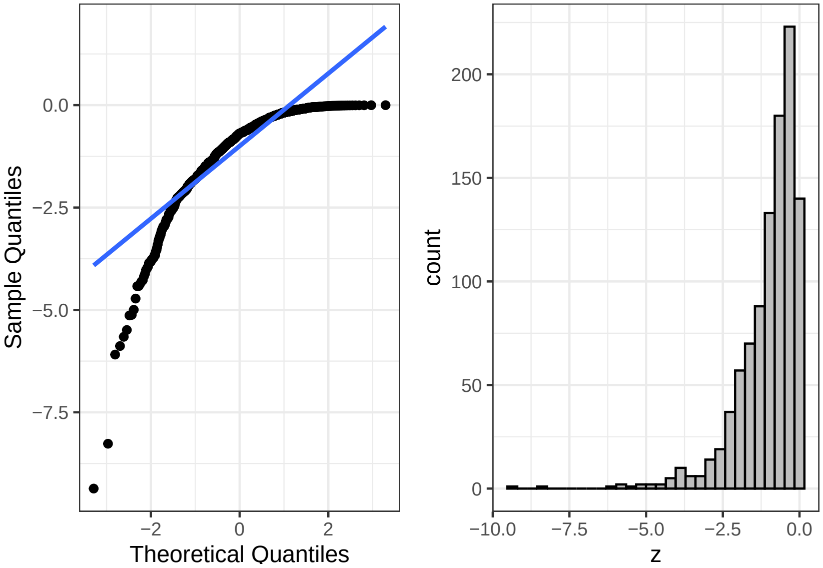
Normal distribution



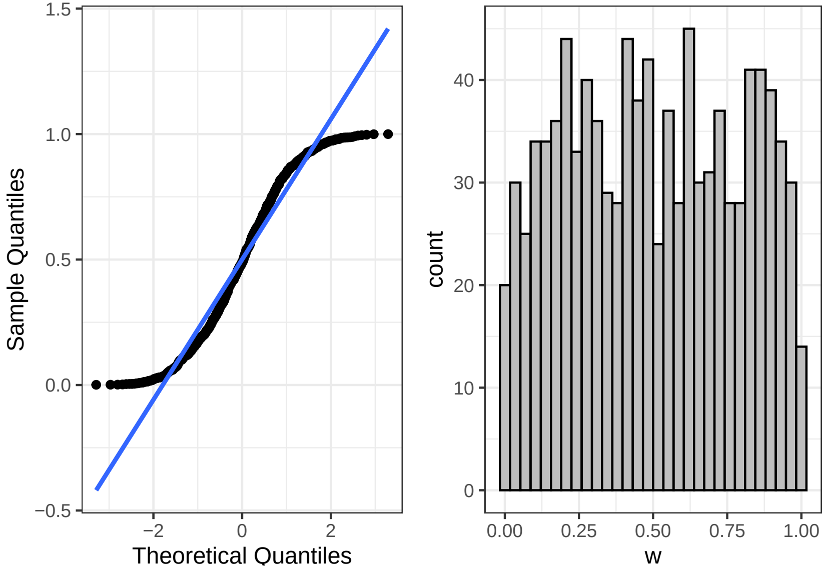
Right-skewed distribution



Left-skewed distribution

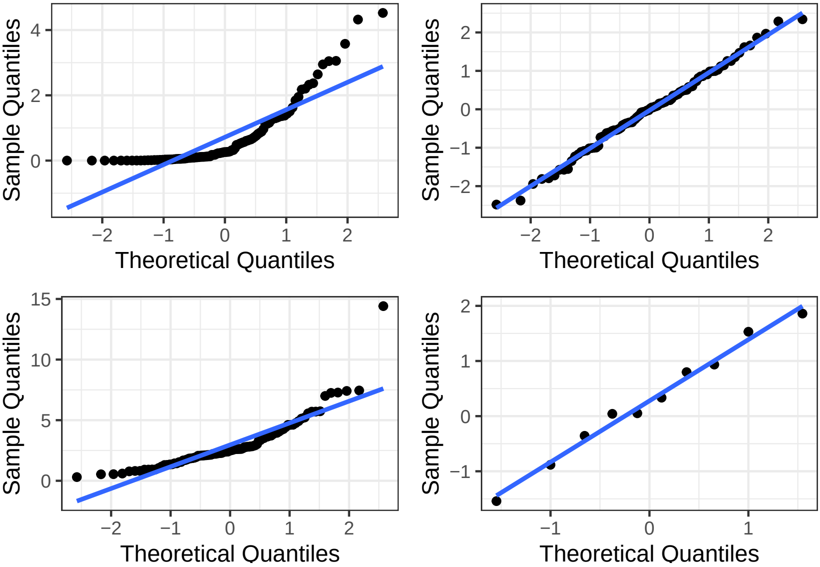


Uniform distribution



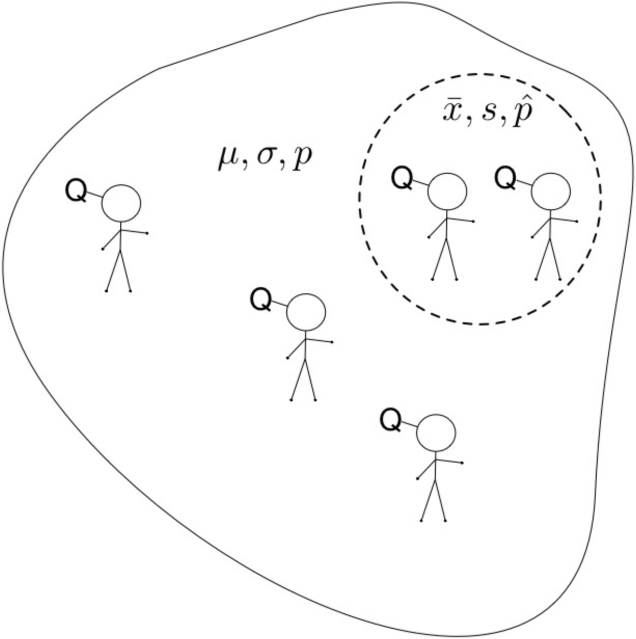
Normal quantile plots

Your go



Sampling Distributions

Populations and samples



Parameters and statistics

A **parameter** is a number that is calculated from the population. In statistical practice, the value of a parameter is not known because we cannot examine the entire population.

A **statistic** is a number that is calculated from a sample. The value of a statistic can be computed directly from the sample data. We often use a statistic to estimate an unknown parameter.

Statistical estimation

The process of **statistical inference** involves using information from a sample to draw conclusions about a the population of interest.

Different random samples yield different values of the statistics. We need to be able to describe the **sampling distribution** of the possible values of a statistic in order to perform statistical inference.

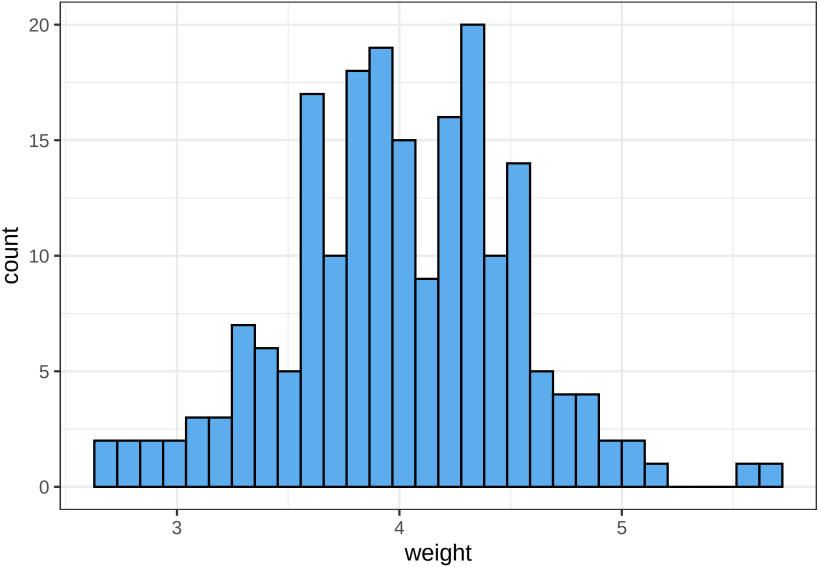
Example

Cats

name	weight
Nawwaara	4.63
Julie	4.21
Brandon	3.45
Africa	3.94
Jasphine	3.93
Zavon	4.19

Example

Cats



Example

The true population mean is $\mu = 4.01$ kg.

First sample

- Take a random sample of 10 cats.

name	weight
Alexis	3.89
Timothy	3.84
Emily	4.42
Jethzabel	4.33
Brittany	4.17
Nickolas	3.31
Haatim	4.30
Alondra	4.06
Candice	3.11
Michaela	4.19

The sample mean weight of the cats in the first sample is 3.96 kg.

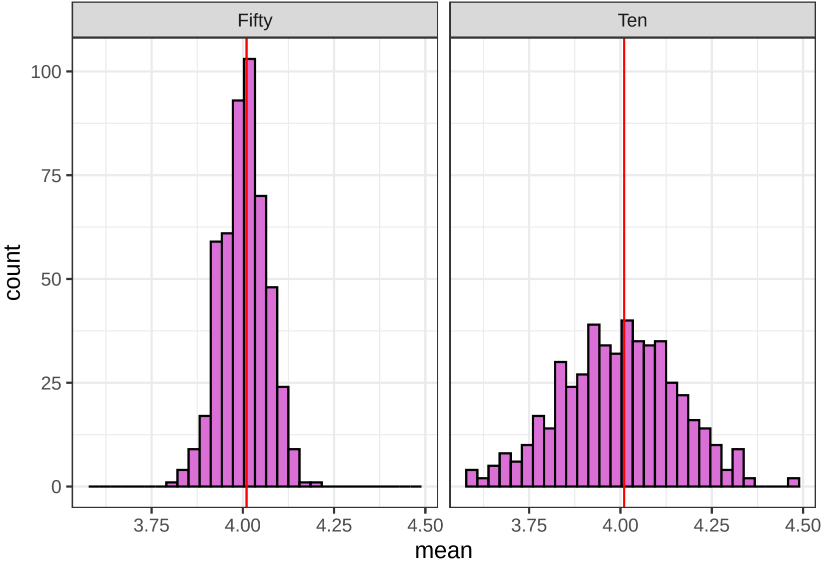
Second sample

- Take a random sample of another 10 cats.

name	weight
Edwell	3.67
Stevan	3.37
Edmundo	4.39
Seth	3.34
Rachel	2.75
Richard	4.33
Anthony	3.86
Nathaniel	3.99
Brandi	4.52
Ruth	3.64

The sample mean weight of the cats in the first sample is 3.79 kg.

Example

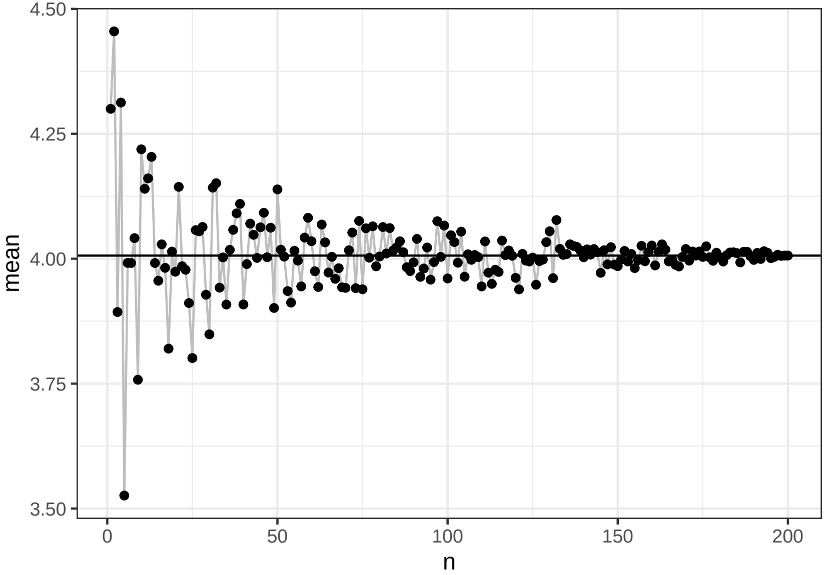


Law of large numbers

Draw independent observations at random from any population with finite mean μ. Decide how accurately you would like to estimate μ. As the number of observations drawn increases, the mean $\bar{x}$ of the observed values eventually approaches the mean μ of the population as closely as you specified and then stays that close.

Example

Back to my cat weighing problem.



Mean of a sample mean

The mean of the sampling distribution is an **unbiased estimate** of the population mean μ.

That is, if I took a lot of sample means and calculated the mean of them, then this value would approach the population mean as I took more and more samples.

Standard deviation of a sample mean

If the population standard deviation is denoted by σ_X and we take a random sample of size n, then the standard deviation of the sample mean is

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}.$$

Note that as the sample size gets larger the sample means will vary less.

Example

Cats' weights

It is known that the population standard deviation of cats' weights is $\sigma = 0.5$ kg.

If we take a sample of 10 cats what is the standard deviation of the sample mean?

If we take a sample of 30 cats what is the standard deviation of the sample mean?

Distribution of the sample mean

If the population of interest is normally distributed, i.e.,

$$X \sim N(\mu, \sigma),$$

then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Example

Cats' weights

It is known that the weights of 200 cats are normally distributed with a mean, \$ = 4.01\$ kg, and a standard deviation: $\sigma = 0.5$ kg.

If we take a sample of 10 cats what is the distribution of the sample mean?

If we take a sample of 30 cats what is the distribution of the sample mean?

Central limit theorem

Most population distributions are not Normal.

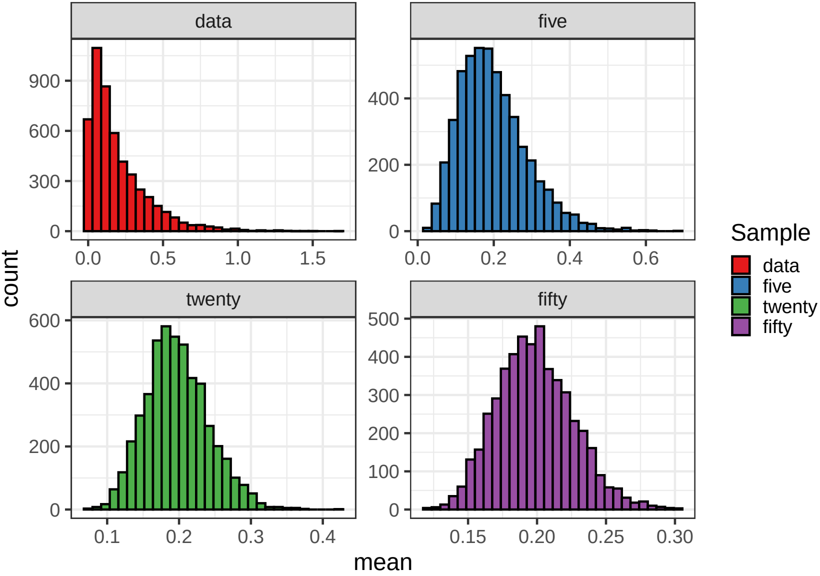
- What is the shape of the sampling distribution of sample means when the population distribution isn't Normal?
- It is a remarkable fact that, as the sample size increases, the distribution of sample means begins to look more and more like a Normal distribution!
- When the sample is large enough, the distribution of sample means is very close to Normal, no matter what shape the population distribution has, as long as the population has a finite standard deviation.

Central limit theorem

Draw an simple random sample of size n from a population with mean μ and finite standard deviation σ. The **central limit theorem (CLT)** says that when n is large, the sampling distribution of the sample mean $\bar{X}$ is approximately normal:

$$\bar{X} \stackrel{\cdot}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Central limit theorem



Further reading/tools

- Textbook:
  - Chapter 1.4 Density curves and normal distribution
  - Chapter 5.1 Towards statistical inference
  - Chapter 5.2 The sampling distribution of a sample mean
- Normal distribution Applet: https://digitalfirst.bfwpub.com/stats_applet/stats_applet_7_norm.html
- Sampling distribution Applet: https://www.rossmanchance.com/applets/2021/sampling/OneSample.html