University of BRISTOL

One sample statistical hypothesis testing

Statistical inference based on a single sample

Henry W J Reeve

henry.reeve@bristol.ac.uk

Statistical Computing & Empirical Methods (EMATM0061)

MSc in Data Science, Teaching block 1, 2021.

What will we cover today?

- We will introduce the method of statistical hypothesis testing.
- We will focus upon the challenge of testing a single parameter with a single sample.
- We will introduce the one-sample t-test for continuous numerical data.
- We will contrast between one-sided and two-sided hypothesis tests.
- We will discuss the exact Binomial test for testing binary data.

Making inferences based on data

Hypothesis testing is a systematic approach to drawing inferences from data.

1. A farmer wants to know if applying different types of soil treatment will modify their crop yield.

2. A pharmaceutical company wants to know if a treatment for a medical condition is effective.

3. A physicist wants to know if their prior belief regarding the value of a physical constant is correct.

By combining hypothesis testing combined with appropriate experimental design we can :

Justify our inferences from data.

Understand and control the role of statistical variation.

Making inferences based on sample data

Suppose that a chef is buying lots of bags salt for their restaurant.

The bags are advertised as 1kg in weight. But the chef is suspicious…

The chef weighs a sample of $n = 30$ bags of salt:

Sample mean: $\overline{X} = 989.06 \text{ (grams)}$

Sample standard deviation: $S = 19.62 \text{ (grams)}$

Should the chef conclude that the true population mean μ differs from 1000grams?

Is it more reasonable to conclude that the difference between sample and population mean is due to chance?

Statistical hypotheses

The **statistical hypothesis** frames the research question in terms of the parameters of a statistical model.

There are two hypotheses:

H_0 : The **null hypothesis** is our default position typically declaring an absence of an interesting phenomena.

H_1 : The **alternative hypothesis** is something interesting which differs from the null.

Example

Research question: Do the bags of salt differ from their advertised weight, on average?

Statistical hypotheses: We model the weights of the salt bags as i.i.d. $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

Null hypothesis $\mathrm{H}_0 : \mu = 1000$ and Alternative hypothesis $\mathrm{H}_1 : \mu \neq 1000$

Test statistic

The **test statistic** is some function of the data which:

1) Has a known distribution under the null hypothesis H_0 .

2) Often takes on large or "extreme" values under the alternative hypothesis H_1 .

The test statistic takes on typical values for H_0 $\Rightarrow$ we fail to reject the null hypothesis H_0 .

The test statistic takes on non-typical values for H_0 $\Rightarrow$ we reject the null & accept the alternative H_1 .

One sample t-test

The **test statistic** is some function of the data which:

1) Has a known distribution under the null hypothesis H_0 .

2) Often takes on large or "extreme" values under the alternative hypothesis H_1 .

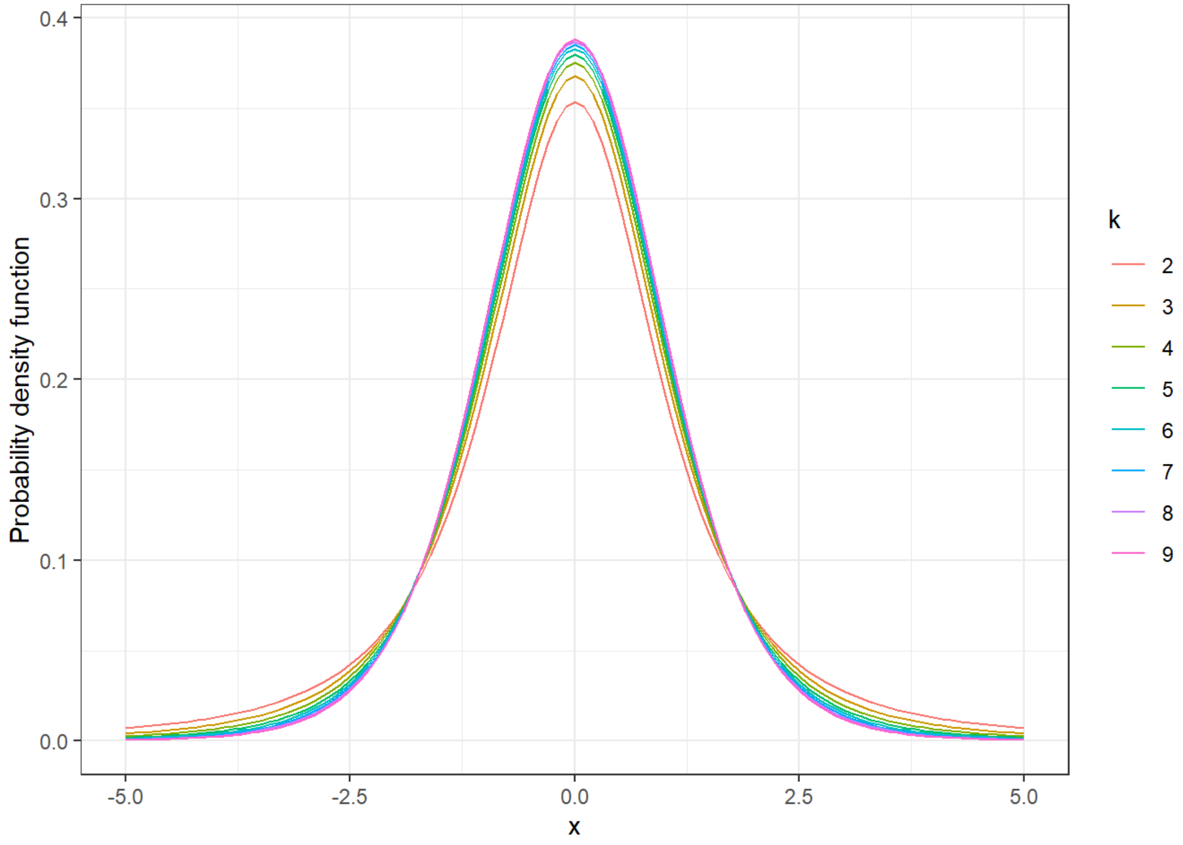
<u>Example</u>

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\mathrm{H}_0 : \mu = 1000$ and $\mathrm{H}_1 : \mu \neq 1000$

$\boxed{\hat{T} := \frac{\overline{X} - 1000}{S/\sqrt{n}}}$ where $\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$ and $S^2 := \frac{1}{n-1} (X_i - \overline{X})^2$.

Intuition: $\mathrm{H}_0 \Rightarrow |\hat{T}|$ is typically small. $\mathrm{H}_1 \Rightarrow |\hat{T}|$ can be large.

Student's t distribution



Given $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

Let $\overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$

and $S^2 := \frac{1}{n-1} \left(X_i - \overline{X}\right)^2$

Then

$$\hat{T} = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is t-distributed with $n-1$ degrees of freedom.

One sample t-test

The **test statistic** is some function of the data which:

1) Has a known distribution under the null hypothesis H_0 .

2) Often takes on large or "extreme" values under the alternative hypothesis H_1 .

<u>Example</u>

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\mathrm{H}_0 : \mu = 1000$ and $\mathrm{H}_1 : \mu \neq 1000$

$\hat{T} := \dfrac{\overline{X} - 1000}{S/\sqrt{n}}$ where $\overline{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 := \frac{1}{n-1}\left(X_i - \overline{X}\right)^2$.

Then if H_0 then $\hat{T}$ has Student's t distribution with $n-1$ degrees of freedom.

One sample t-test

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\mathrm{H}_0 : \mu = 1000$ and $\mathrm{H}_1 : \mu \neq 1000$

$$\hat{T} := \frac{\overline{X} - 1000}{S/\sqrt{n}}$$

where $\overline{X} := \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 := \frac{1}{n-1}\left(X_i - \overline{X}\right)^2$.

Then if H_0 then $\hat{T}$ has Student's t-distribution with $n-1$ degrees of freedom.

The next stage is to compute the numerical value of the test statistic $\hat{T}$ based upon the data.

```
sample_size<-length(salt_vect) # sample size
sample_mean<-mean(salt_vect) # sample mean
sample_sd<-sd(salt_vect) # standard deviation
test_statistic<-(sample_mean-1000)/(sample_sd/sqrt(sample_size)) # test statistic
test_statistic
```

```
## [1] -3.054553
```

The p-value

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\mathrm{H}_0: \ \mu = 1000$ and $\mathrm{H}_1: \ \mu \neq 1000$

Then $\hat{T}$ has Student's t-distribution with $n-1$ degrees of freedom.

The test statistic $\hat{T}$ has numerical value $\tau = -3.0546$.

Next we compute the **p-value**: The probability that $\hat{T}$ is at least as extreme as τ under the null H_0 ,

p-value:

$$\begin{aligned} p := \mathbb{P}_{H_0}\left(|\hat{T}| \geq |\tau|\right) &= \mathbb{P}_{H_0}\left(\hat{T} \geq |\tau|\right) + \mathbb{P}_{H_0}\left(\hat{T} \leq -|\tau|\right) \\ &= 2 \cdot \mathbb{P}_{H_0}\left(\hat{T} \geq |\tau|\right) = 2 \cdot \left\{1 - \mathbb{P}_{H_0}\left(\hat{T} < |\tau|\right)\right\} \\ &= 2 \cdot \{1 - F_{n-1}(|\tau|)\} . \end{aligned}$$

where $F_{n-1}(t) = \mathbb{P}(T < t) = \int_{-\infty}^{t} f_{n-1}(s)ds$ is the cumulative distribution.

The p-value

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\mathrm{H}_0 : \ \mu = 1000$ and $\mathrm{H}_1 : \ \mu \neq 1000$

Then $\hat{T}$ has Student's t-distribution with $n-1$ degrees of freedom.

The test statistic $\hat{T}$ has numerical value $\tau = -3.0546$.

The **p-value:** $p = 2 \cdot \{1 - F_{n-1}(|\tau|)\}$ where $F_{n-1}(t)$ is the Student's t distribution function.

```
test_statistic # value of the one-sample t-statistic
```

```
## [1] -3.054553
```

```
2*(1-pt(abs(test_statistic),df=sample_size-1)) # compute the p-value
```

```
## [1] 0.0047971
```

The p-value

Suppose $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, $\mathrm{H}_0: \mu = 1000$ and $\mathrm{H}_1: \mu \neq 1000$

```
2*(1-pt(abs(test_statistic),df=sample_size-1)) # compute the p-value
```

```
## [1] 0.0047971
```

The p-value $p := \mathbb{P}_{H_0}\left(|\hat{T}| \geq |\tau|\right)$ is very small

The value of the test-statistic is very unlikely under the null hypothesis H_0

We should reject the null hypothesis H_0 in favor of the alternative hypothesis H_1.

But how small a p-value $p := \mathbb{P}_{H_0}\left(|\hat{T}| \geq |\tau|\right)$ is small enough?

Different types of errors

There are two types of errors that can occur with hypothesis testing:

		Reality	
		H_0	H_1
Our conclusions	H_0	✓	Type II error
	H_1	Type I error	✓

<u>Type I error</u>: We reject the null hypothesis when the null hypothesis is actually correct.

<u>Type II error</u>: We fail to reject the null hypothesis when the alternative hypothesis is correct.

We are generally more cautious about Type I error than Type II error since the null hypothesis is our default.

This corresponds to being conservative with respect to declaring the existence of new phenomena.

Significance level

The **test size** of a test α_{test} is the probability of Type I error under the null hypothesis:

$$\alpha_{\text{test}} = \mathbb{P}\left(\text{Type I error} \mid \text{H}_0 \text{ is true}\right).$$

The **power** of test is $1 - \beta_{\text{test}}$ where β_{test} is the probability of Type II error under the alternative:

$$\beta_{\text{test}} = \mathbb{P}\left(\text{Type II error} \mid \text{H}_1 \text{ is true}\right).$$

The **significance level** of a test α is an upper bound on the test size $\alpha_{\text{test}} \leq \alpha$.

A valid hypothesis test requires that the significance level be chosen in advance of seeing the data.

There is an inevitable tradeoff here – the lower the significance α the lower the power $1 - \beta_{\text{test}}$

We usually opt for $\boxed{\alpha = 0.05}$ but the optimal choice depends upon the application.

Statistical hypothesis testing

If the p value is strictly less than the significance level we reject the null hypothesis $\Rightarrow$ H_1

If the p value is greater than the significance level we do not reject the null hypothesis $\Rightarrow$ H_0

Example

We have a sample of salt weights $X_1, \cdots, X_n$ which we model as i.i.d. Gaussian $\mathcal{N}(\mu, \sigma^2)$.

We have a null hypothesis of $\mathrm{H}_0 : \mu = 1000$ and an alternative hypothesis of $\mathrm{H}_1 : \mu \neq 1000$

We choose a significance level $\alpha = 0.05$.

We apply the one-sample t-test to compute a p value 0.004797.

We reject the null hypothesis and conclude that $\mathrm{H}_1 : \mu \neq 1000$

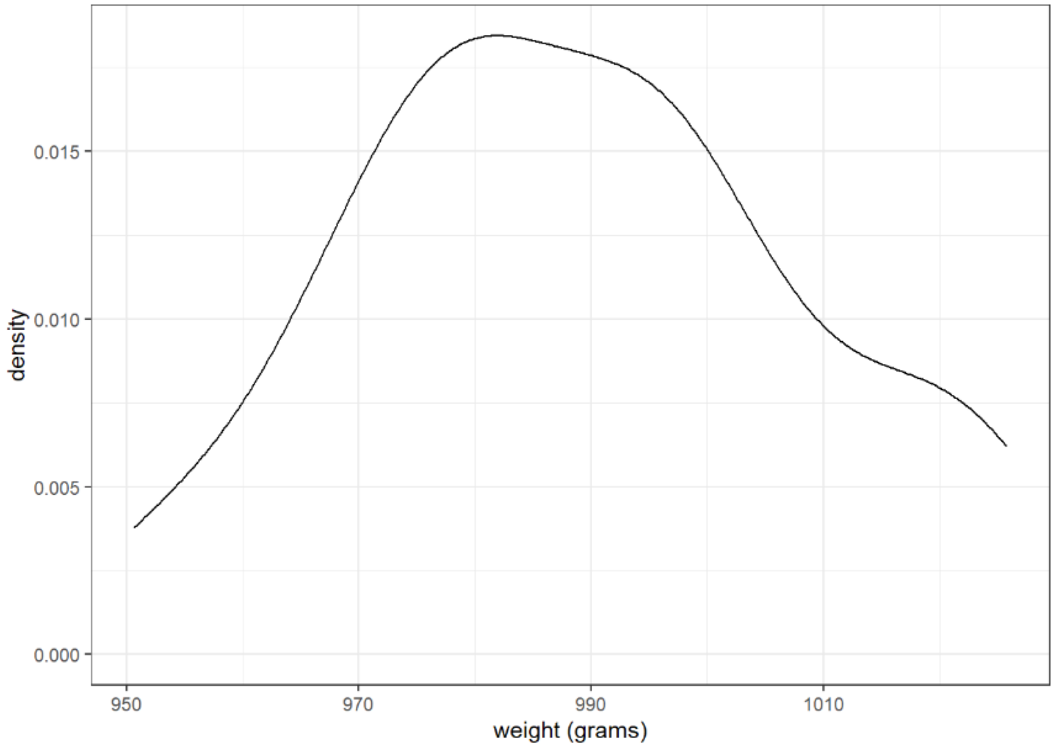
Note: For our conclusions to be reasonable our assumptions must be valid.

One sample t-test with R

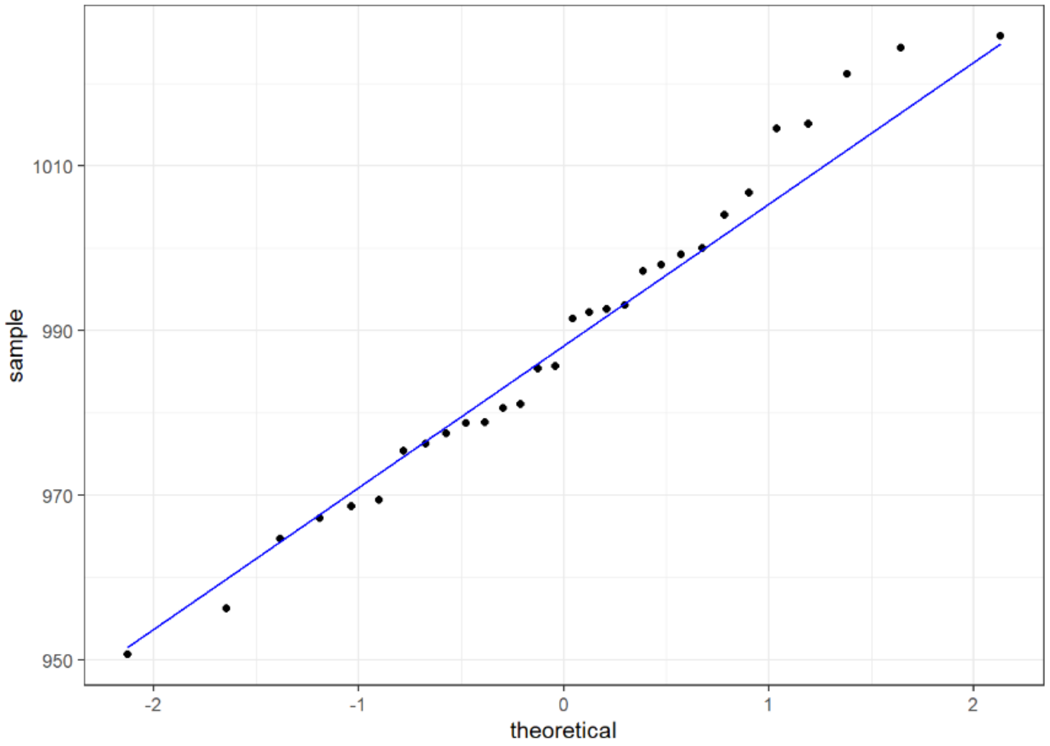
We have a sample of salt weights $X_1, \cdots, X_n$ which we model as i.i.d. Gaussian $\mathcal{N}(\mu, \sigma^2)$.

We have a null hypothesis of $\mathrm{H}_0 : \mu = 1000$ and an alternative hypothesis of $\mathrm{H}_1 : \mu \neq 1000$

```
tibble(salt_vect)%>%ggplot(aes(x=salt_vect))+geom_density()+
  theme_bw()+labs(x="weight (grams)",y="density")
```




```
tibble(salt_vect)%>%ggplot(aes(sample=salt_vect))+stat_qq()+
  stat_qq_line(color="blue")+theme_bw()
```



One sample t-test with R

We have a sample of salt weights $X_1, \cdots, X_n$ which we model as i.i.d. Gaussian $\mathcal{N}(\mu, \sigma^2)$.

We have a null hypothesis of $\mathrm{H}_0 : \mu = 1000$ and an alternative hypothesis of $\mathrm{H}_1 : \mu \neq 1000$

We choose a significance level of $\alpha = 0.05$.

We apply a one sample t-test with $\mathrm{H}_0 : \mu = 1000$ and $\mathrm{H}_1 : \mu \neq 1000$.

```
t.test(x=salt_vect,mu=1000)
```

```
##
##  One Sample t-test
##
## data:  salt_vect
## t = -3.0546, df = 29, p-value = 0.004797
## alternative hypothesis: true mean is not equal to 1000
## 95 percent confidence interval:
##  981.7315 996.3844
## sample estimates:
## mean of x
##  989.0579
```

One sample t-test with R

We choose a significance level of $\alpha = 0.05$.

We apply a one sample t-test with $\mathrm{H}_0 : \mu = 1000$ and $\mathrm{H}_1 : \mu \neq 1000$.

```
t.test(x=salt_vect,mu=1000)
```

```
##
##  One Sample t-test
##
## data:  salt_vect
## t = -3.0546, df = 29, p-value = 0.004797
## alternative hypothesis: true mean is not equal to 1000
## 95 percent confidence interval:
##  981.7315 996.3844
## sample estimates:
## mean of x
##  989.0579
```

The p-value = 0.004797 is below the significance level $\alpha = 0.05$

We reject the null hypothesis $\mathrm{H}_0 : \mu = 1000$ and conclude that $\mathrm{H}_1 : \mu \neq 1000$.

A confidence interval approach

We can model the weights of the bags of salt as an i.i.d. Gaussian $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

Confidence intervals: $(\hat{L}_n, \hat{U}_n)$ so that $\mathbb{P}\left[\hat{L}_n \leq \mu \leq \hat{U}_n\right] \geq 1 - \alpha.$

```
alpha<-0.05 # 95%-level confidence intervals
sample_size<-length(salt_vect) # sample size
sample_mean<-mean(salt_vect) # sample mean
sample_sd<-sd(salt_vect) # standard deviation
t<-qt(1-alpha/2,df=sample_size-1) # (1-alpha/2)- quantile
confidence_interval_l<-sample_mean-t*sample_sd/sqrt(sample_size) # lower confidence bound
confidence_interval_u<-sample_mean+t*sample_sd/sqrt(sample_size) # upper confidence bound
confidence_interval<-c(confidence_interval_l,confidence_interval_u) # confidence interval
confidence_interval
```

```
## [1] 981.7315 996.3844
```

A confidence interval approach

We can model the weights of the bags of salt as an i.i.d. Gaussian $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

Confidence intervals: $(\hat{L}_n, \hat{U}_n)$ so that $\mathbb{P}\left[\hat{L}_n \leq \mu \leq \hat{U}_n\right] \geq 1 - \alpha.$

We compute 95%-level confidence intervals: $(\hat{L}_n, \hat{U}_n) = (981.73, 996.38)$

If we truly had $\mu = 1000$ then the probability of having such a low value of $\hat{U}_n$ would be below 0.05.

On these grounds we can reasonably reject the hypothesis that $\mu = 1000$ and conclude that $\mu < 1000.$

Statistical hypothesis testing allows us to formalize and generalize this type of statistical reasoning.

Statistical hypothesis testing

Suppose we have a clear research hypothesis and some high-quality data from a well-deigned experiment.

The key stages of statistical hypothesis testing are as follows:

1. Form our statistical hypothesis including a null hypothesis and an alternative hypothesis.
2. Apply model checking to validate any modelling assumptions.
3. Choose our desired significance level.
4. Select an appropriate statistical test.
5. Compute the numerical value of the test statistic from data.
6. Compute a p-value based upon the test statistic.
7. Draw conclusions based upon the relationship between the p-value and the significance level.

Now take a break!

The one sample t-test

Suppose that a chicken farmer knows that in previous years the average weight of their chickens is 2.5kg.

They want to know if their current batch of chickens is of typical weight.

There are too many chickens on the farm to weight them all (& they are difficult to catch!).

Instead the farmer collects a sample of chicken weights:

$$X_1, \cdots, X_n$$

The one sample t-test

The first stage is to formulate a null hypothesis and an alternative hypothesis:

The weights are modelled as i.i.d. Gaussian samples $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

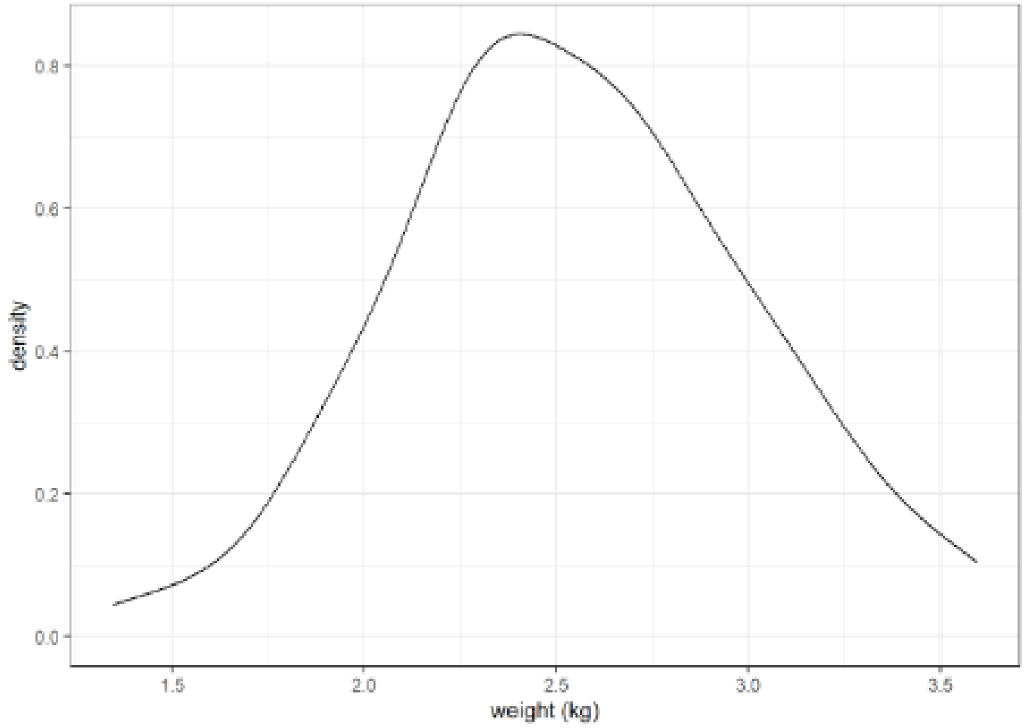
Null hypothesis: $\mathrm{H}_0 : \quad \mu = 2.5 \ (\mathrm{kg})$

Alternative hypothesis: $\mathrm{H}_1 : \quad \mu \neq 2.5 \ (\mathrm{kg})$.

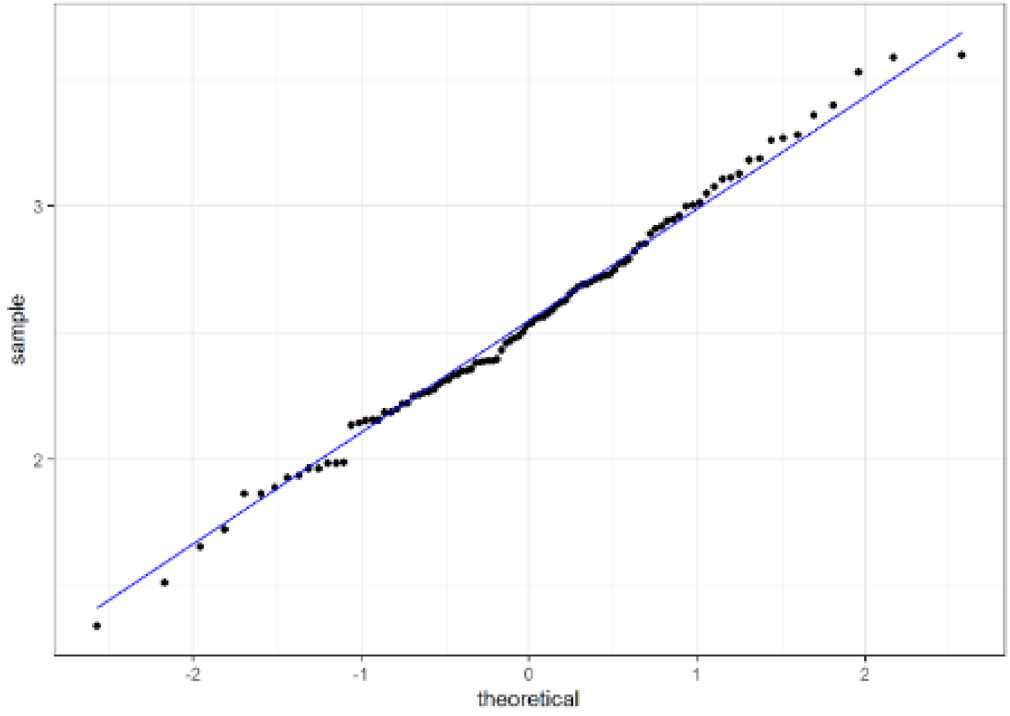
The one sample t-test

We should be sure to check our modelling assumptions:

```
tibble(chicken_weights)%>%ggplot(aes(x=chicken_weights))+geom_density()+
  theme_bw()+labs(x="weight (kg)",y="density")
```




```
tibble(chicken_weights)%>%ggplot(aes(sample=chicken_weights))+stat_qq()+
  stat_qq_line(color="blue")+theme_bw()
```



It seems reasonable to proceed with an i.i.d. Gaussian model $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$

The one sample t-test

The weights are modelled as i.i.d. Gaussian samples $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

Null hypothesis: $\mathrm{H}_0 : \quad \mu = 2.5 \ (\mathrm{kg})$

Alternative hypothesis: $\mathrm{H}_1 : \quad \mu \neq 2.5 \ (\mathrm{kg})$.

Next we choose a significance level: $\alpha = 0.05$

This is the maximum probability of Type 1 error:

$$\alpha \geq \mathbb{P}\left(\text{Type I error} \mid \mathrm{H}_0 \text{ is true}\right).$$

The one sample t-test

Next we perform our **one-sample t-test** with R:

```
t.test(x=chicken_weights,mu=2.5)
```

```
##
##  One Sample t-test
##
## data:  chicken_weights
## t = 0.99041, df = 99, p-value = 0.3244
## alternative hypothesis: true mean is not equal to 2.5
## 95 percent confidence interval:
##  2.454642 2.635764
## sample estimates:
## mean of x
##  2.545203
```

We compute a p-value of 0.3244. This is well above the significance level of $\alpha = 0.05$

Hence, we default to not rejecting the null hypothesis of $H_0 : \quad \mu = 2.5 \ (\text{kg})$

Two sided and one-sided hypothesis tests

Sometimes we are only interested in deviations in one direction.

Examples

In our salt example the chef might only be concerned if our bag of salt is too light.

In our chickens example the farmer might only be concerned if the chickens are smaller.

When this occurs we can use a one-sided test, rather than a two sided test.

Two sided

$$\mathrm{H}_0 : \quad \mu = 1 \ (\mathrm{kg})$$

$$\mathrm{H}_1 : \quad \mu \neq 1 \ (\mathrm{kg})$$

One sided

$$\mathrm{H}_0 : \quad \mu \geq 1 \ (\mathrm{kg})$$

$$\mathrm{H}_1 : \quad \mu < 1 \ (\mathrm{kg}).$$

Two sided and one-sided hypothesis tests

The default t-test is two sided in R. However, we can also perform a one-sided test.

Two-sided

$$H_0 : \quad \mu = 1 \ (\text{kg})$$

$$H_1 : \quad \mu \neq 1 \ (\text{kg})$$

```
t.test(x=salt_vect,mu=1000,alternative="two.sided")
```

```
##
##  One Sample t-test
##
## data:  salt_vect
## t = -3.0546, df = 29, p-value = 0.004797
## alternative hypothesis: true mean is not equal to 1000
## 95 percent confidence interval:
##  981.7315 996.3844
## sample estimates:
## mean of x
##  989.0579
```

One-sided

$$H_0 : \quad \mu \geq 1 \ (\text{kg})$$

$$H_1 : \quad \mu < 1 \ (\text{kg}).$$

```
t.test(x=salt_vect,mu=1000,alternative="less")
```

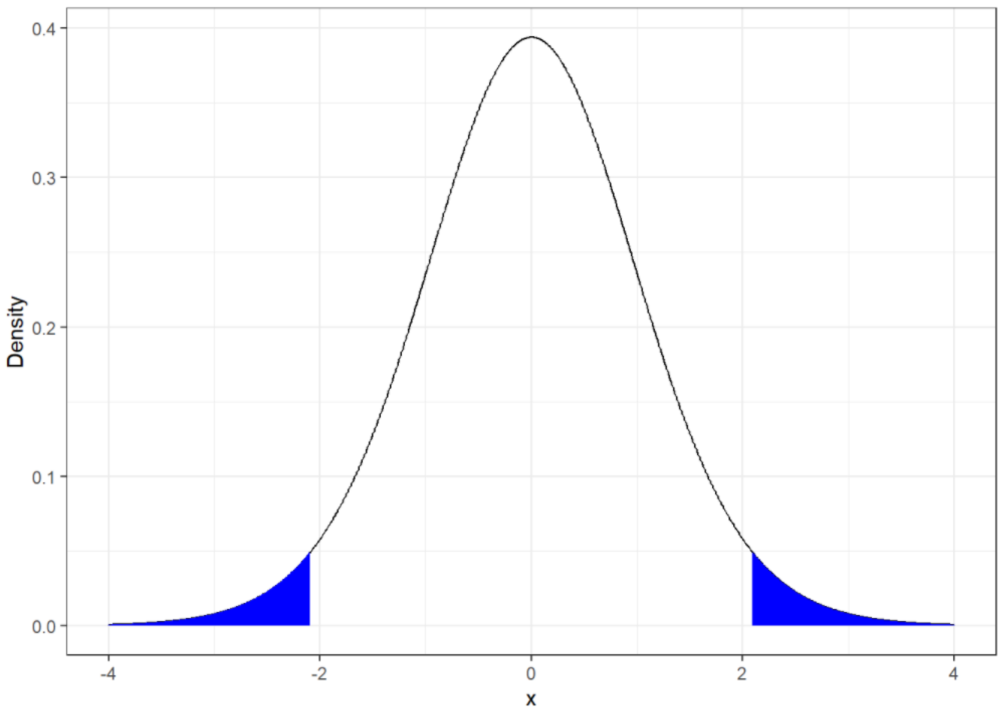
```
##
##  One Sample t-test
##
## data:  salt_vect
## t = -3.0546, df = 29, p-value = 0.002399
## alternative hypothesis: true mean is less than 1000
## 95 percent confidence interval:
##      -Inf 995.1446
## sample estimates:
## mean of x
##  989.0579
```

Two sided and one-sided hypothesis tests

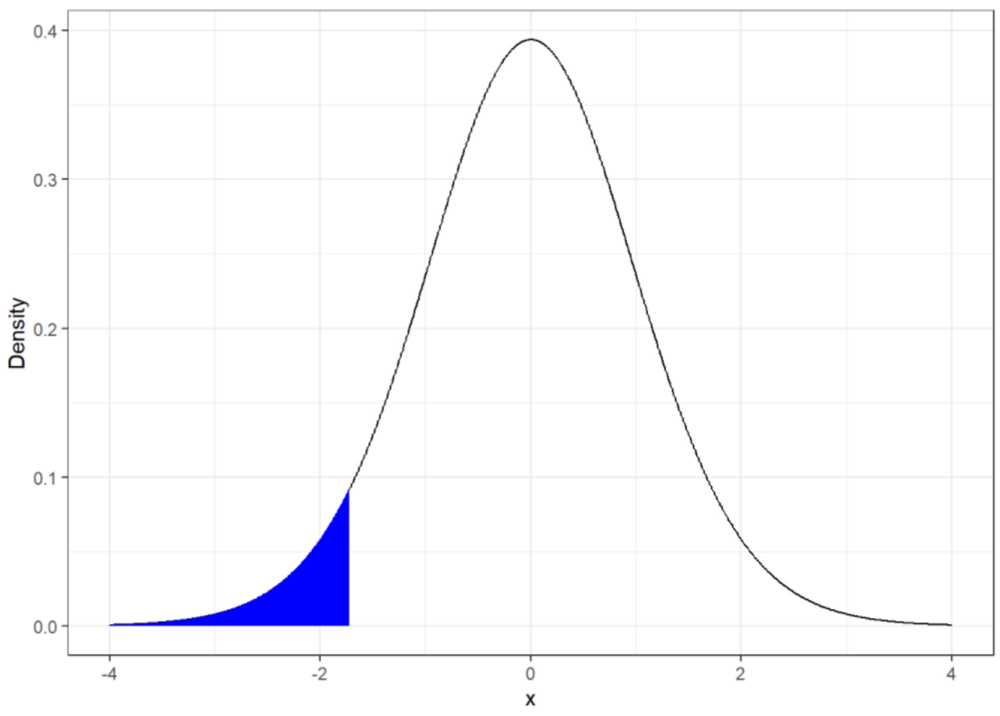
One sided hypothesis tests have higher power $1-\beta_{\text{test}}$ the same significance level $\alpha \geq \alpha_{\text{test}}$

$$\alpha_{\text{test}} = \mathbb{P}\left(\text{Type I error} \mid \text{H}_0 \text{ is true}\right) \qquad \beta_{\text{test}} = \mathbb{P}\left(\text{Type II error} \mid \text{H}_1 \text{ is true}\right).$$

<u>Two-sided</u>



<u>One-sided</u>



Two sided and one-sided hypothesis tests

Use caution with one-sided tests!

One sided tests are more powerful… but also less conservative.

You shouldn’t use a one-sided test if either direction is of interest.

E.g. You would also be very interested if the bags of salt were heavier on average than 1kg.

The method is no longer valid if you choose between a one-sided and two-sided test after seeing the data!

By default you can use a two-sided test.

You can still make a directional conclusion with a two-sided test based on the sign of your test statistic.

Now take a break!

The Binomial test

Suppose that Bill tells you he can predict the outcome of a dice roll better than chance.

You suspect that he cannot predict the outcome of a dice role and typically will do no better than chance.

You believe that his probability of making a correct prediction is 1/6.

Bill believes his probability of success is better than 1/6.

How can we test this?

The Binomial test

We conduct a sequence of trials where Bill makes a prediction & then rolls the dice.

This generates $X_1, \cdots, X_n$ where

$$X_i = \begin{cases} 1 & \text{if Bill's } i\text{th prediction is correct} \\ 0 & \text{if Bill's } i\text{th prediction is incorrect.} \end{cases}$$

Let's model $X_1, \cdots, X_n \sim \mathcal{B}(q)$ as i.i.d. Bernoulli random variables with $q \in [0, 1]$

The Binomial test

We model $X_1, \cdots, X_n \sim \mathcal{B}(q)$ as i.i.d. Bernoulli random variables with $q \in [0, 1]$

Null hypothesis: $\mathrm{H}_0 : \quad q = 1/6$

Alternative hypothesis: $\mathrm{H}_1 : \quad q \neq 1/6.$

We choose a significance level of $\alpha = 0.05$.

The Binomial test

We model $X_1, \cdots, X_n \sim \mathcal{B}(q)$ as i.i.d. Bernoulli random variables with $q \in [0, 1]$

Null hypothesis: $\mathrm{H}_0 : \quad q = 1/6$

Alternative hypothesis: $\mathrm{H}_1 : \quad q \neq 1/6.$

We choose a significance level of $\alpha = 0.05$.

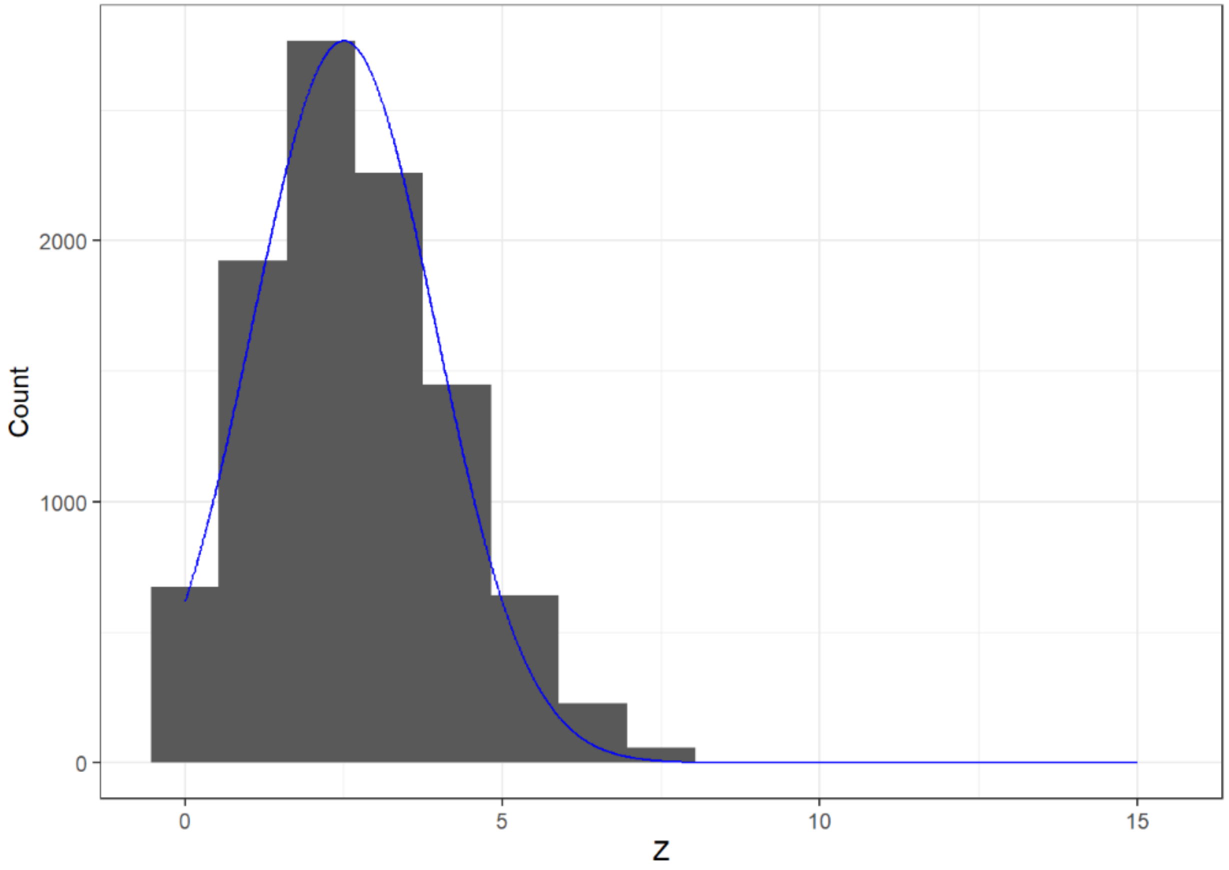
A natural test statistic is the number of successes $\underline{Z = X_1 + \cdots + X_n}.$

We expect Z to be close to $\underline{q \cdot n}$, so close to $\underline{n/6}$ under the null hypothesis $\underline{\mathrm{H}_0}$.

The Binomial test

Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q)$ is a sequence of i.i.d. Bernoulli random variables with $q \in [0, 1]$

Then the sum $Z = X_1 + \cdots + X_n$ has Binomial distribution $Z \sim \mathrm{BINOM}(n, q)$.



Probability mass function:

$$\mathbb{P}(Z = k) = \frac{n!}{(n-k)!k!} \cdot q^k \cdot (1-q)^{n-k}.$$

The Binomial test

Suppose $X_1, \cdots, X_n \sim \mathcal{B}(q)$ is a sequence of i.i.d. Bernoulli random variables with $q \in [0,1]$

Then the sum $Z = X_1 + \cdots + X_n$ has Binomial distribution $Z \sim \mathrm{BINOM}(n, q)$.

The Binomial has probability mass function $\mathbb{P}(Z = k) = \frac{n!}{(n-k)!k!} \cdot q^k \cdot (1-q)^{n-k}$.

To compute the p-value we first let $\mathcal{T}$ be the numerical value of the test-statistic.

The **p-value** is the probability of attaining a test-statistic at least as extreme as $\mathcal{T}$ under the null hypothesis.

The Binomial test

The Binomial has probability mass function $\mathbb{P}(Z=k) = \frac{n!}{(n-k)!k!} \cdot q^k \cdot (1-q)^{n-k}.$

To compute the p-value we first let $\mathcal{T}$ be the numerical value of the test-statistic.

The **p-value** is the probability of attaining a test-statistic at least as high as $\mathcal{T}$ under the null hypothesis.

For example, if we have $\mathrm{H}_0: \quad q = 1/6$ and $\mathrm{H}_1: \quad q \neq 1/6.$

If $\tau > n \cdot (1/6)$, $p := 2 \cdot \mathbb{P}_{\mathrm{H}_0}(Z \geq \tau) = 2 \cdot \sum_{k=\tau}^{n} \frac{n!}{(n-k)!k!} \cdot \left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{n-k}.$

If $\tau \leq n \cdot (1/6)$, $p := 2 \cdot \mathbb{P}_{\mathrm{H}_0}(Z \leq \tau) = 2 \cdot \sum_{k=0}^{\tau} \frac{n!}{(n-k)!k!} \cdot \left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{n-k}.$

If the p-value is below the significance level we reject the null $\Rightarrow$ H_1

Otherwise we fail to reject the null $\Rightarrow$ H_0

The Binomial test

We model $X_1, \cdots, X_n \sim \mathcal{B}(q)$ as a sequence of i.i.d. Bernoulli random variables with $q \in [0, 1]$.

We have null hypothesis $\mathrm{H}_0 : \quad q = 1/6$ and alternative hypothesis $\mathrm{H}_1 : \quad q \neq 1/6.$

We have a significance level $\alpha = 0.05$.

```
correct_or_incorrect
```

```
##  [1] 0 1 0 1 1 0 1 1 1 0
```

```
sample_size<-length(correct_or_incorrect)
num_successes<-sum(correct_or_incorrect)
binom.test(x=num_successes,n=sample_size,p=1/6,alternative="two.sided")
```

The Binomial test

```
correct_or_incorrect
```

```
##  [1] 0 1 0 1 1 0 1 1 1 0
```

```
sample_size<-length(correct_or_incorrect)
num_successes<-sum(correct_or_incorrect)
binom.test(x=num_successes,n=sample_size,p=1/6,alternative="two.sided")
```

```
##
##  Exact binomial test
##
## data:  num_successes and sample_size
## number of successes = 6, number of trials = 10, p-value = 0.002438
## alternative hypothesis: true probability of success is not equal to 0.1666667
## 95 percent confidence interval:
##  0.2623781 0.8784477
## sample estimates:
## probability of success
##                    0.6
```

Our p-value is well below the significance level so reject the null hypothesis → H_1

What have we covered today?

- We introduced the method of statistical hypothesis testing.
- We introduced the one-sample t-test for continuous numerical data.
- The one-sample t-test relies upon the assumption the underlying data distribution is Gaussian.
- However, for large samples this assumption is less crucial since the sample average emulates a Gaussian.
- We distinguished between one-sided and two-sided hypothesis tests.
- We discussed the exact Binomial test for testing binary data.

Thanks for listening!

Henry W J Reeve

henry.reeve@bristol.ac.uk

EMATM0061