

Chapter 4

Statistical analysis

Statistical analysis

Using the framework of probability and distributions, we now have a framework to talk about aspects of a statistical analysis that allows us to learn about underlying population characteristics.

Statistical analysis

We will investigate

- **sampling distributions**
 - ▶ probability distributions of summary statistics
- **estimates and estimators**
 - ▶ summary statistics used as best guesses at population parameters
- **interval estimates**
 - ▶ representations of uncertainty about population parameters
- **hypothesis tests**
 - ▶ decision procedures for specific proposed properties of the population.

Random samples

In our analyses, we will assume that we have access to one or more **random samples** of data

$$X_1, \dots, X_n$$

where we presume

- each data point is collected under the **same** conditions
- the data points are collected such that they **do not influence one another**.

Random samples

In terms of the underlying probability model, the corresponding random variables

$$X_1, \dots, X_n$$

are assumed to be **independent** and to have the **same** probability distribution.

We shorten this to **i.i.d**

- independent and identically distributed.

Sampling distributions: big idea

In data collection

- we collect data $x_1, \dots, x_n$ in an attempt to learn about a population and its parameters
- we summarize the data using a statistic, such as $\bar{x}$
- the finite sample of size n is just one data set
- if we repeated the data collection, we would get a slightly different collection of data, and a different value of $\bar{x}$

We wish to understand how $\bar{x}$ is going to vary over repeated experiments.

Sampling distributions

Example: Sample mean for a discrete distribution

A large “population” which consists of the numbers 1, 2 and 3, such that 30% are 1s, 40% are 2s and 30% are 3s

- Let X be the random variable that denotes the result obtained from selecting at random one element from this population
- The probability distribution of X is given by

$X = x$	1	2	3
$P(X = x)$	0.3	0.4	0.3

We will say that the population has the distribution given by the table above

Sampling distributions

Example: Sample mean for a discrete distribution

The expected value of X is

$$E(X) = 1 \times 0.3 + 2 \times 0.4 + 3 \times 0.3 = 2$$

- The population variance and standard deviation of X are

$$\text{Var}(X) = 0.6 \quad \text{sd}(X) = \sqrt{0.6} = 0.77.$$

- We will say that the population mean is $\mu = 2$ and the population standard deviation $\sigma = 0.77$.

The population quantities represent what would be obtained for the sample quantities if you had a massive sample.

Sampling distributions

Example: Sample mean for a discrete distribution

Now suppose we have a random sample of size n

$$X_1, \dots, X_n$$

from this distribution. The sample mean random variable is

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

Sampling distributions

Example: Sample mean for a discrete distribution

$\bar{X}$ is **also** a random variable

- before we see the observed data, we do not know what value $\bar{X}$ will take
- we can see it will take a value in

$$\left\{ \frac{n}{n}, \frac{(n+1)}{n}, \frac{(n+2)}{n}, \dots, \frac{3n}{n} \right\}.$$

but it is not immediately obvious what the value of

$$P(\bar{X} = x)$$

is for each x .

Sampling distributions

Note

- The “distribution” for a population is the probability distribution of the random variable X , the result of selecting one unit from the population
- We will interchangeably use the terms ‘distribution of a population’ and ‘distribution of a random variable’, defined by sampling at random one unit for the population
- The distribution of a population can be
 - discrete (Bernoulli, Binomial) or
 - continuous (Normal)

$\bar{X}$ as a Random Variable

- Suppose that we select **two** numbers at random and independently from a population
- Thus, we have now two random variables, say X_1 and X_2
- Because they are coming from the same population, X_1 and X_2 have the same probability distribution.

$\bar{X}$ as a Random Variable

- X_1 and X_2 are a independent and identically distributed (i.i.d.) random variables from the same population
 - ▶ Independence:

$$P(X_1 = x_1 \text{ and } X_2 = x_2) = P(X_1 = x_1) \times P(X_2 = x_2)$$

for all values (x_1, x_2)

- ▶ Identically distributed:

$$P(X_1 = x) = P(X_2 = x)$$

for any x

$\bar{X}$ as a Random Variable

In the previous example

$X = x$	1	2	3
$P(X = x)$	0.3	0.4	0.3

this implies that

$$E(X_1) = E(X_2) = \mu = 2 \quad \text{Var}(X_1) = \text{Var}(X_2) = 0.6$$

$\bar{X}$ as a Random Variable

Let us say now that after we sample the two numbers, we plan compute their mean.

$$\bar{X} = \frac{X_1 + X_2}{2}.$$

As we do not know what values we will get before we do the sampling, $\bar{X}$ is a random variable.

After we have sampled the two numbers, say $(1, 3)$ then we will have an observed value for $\bar{X}$, specifically

$$\bar{x} = \frac{x_1 + x_2}{2} = \frac{1 + 3}{2} = 2.$$

$\bar{X}$ as a Random Variable

What is the probability distribution of $\bar{X}$?

The possible values of $\bar{X}$ are

(1, 1)	implies $\bar{X} = 1$
(1, 2) or (2, 1)	imply $\bar{X} = 3/2$
(1, 3) or (3, 1) or (2, 2)	imply $\bar{X} = 4/2 = 2$
(2, 3) or (3, 2)	imply $\bar{X} = 5/2$
(3, 3)	implies $\bar{X} = 6/2 = 3$

$\bar{X}$ as a Random Variable

The probability of each value of $\bar{X}$ is

$$P(\bar{X} = 1) = 0.3 \times 0.3 = 0.09$$

$$P(\bar{X} = 3/2) = (0.3 \times 0.4) + (0.4 \times 0.3) = 0.24$$

$$P(\bar{X} = 2) = (0.3 \times 0.3) + (0.3 \times 0.3) + (0.4 \times 0.4) = 0.34$$

$$P(\bar{X} = 5/2) = (0.4 \times 0.3) + (0.3 \times 0.4) = 0.24$$

$$P(\bar{X} = 3) = (0.3 \times 0.3) = 0.09$$

$\bar{X}$ as a Random Variable

We can say that the possible values of $\bar{X}$ form a population with distribution

$X = \bar{x}$	1	1.5	2	2.5	3
$P(X = x)$	0.09	0.24	0.34	0.24	0.09

The probability distribution of $\bar{X}$ is determined by the probability distribution of X_1 and X_2

$\bar{X}$ as a Random Variable

Simulating in R:

```
prob.vec<-c(0.3,0.4,0.3)
xvals<-c(1,2,3)
n<-2
nreps<-100000
xsamp<-replicate(nreps,mean(sample(xvals,prob=prob.vec,size=n,rep=TRUE)))
xsamp[1:10]

+ [1] 2.5 2.5 2.0 3.0 2.5 2.0 1.0 1.5 1.0 2.0

table(xsamp)

+ xsamp
+   1   1.5   2   2.5   3
+ 9097 23939 33961 23851 9152

table(xsamp)/nreps

+ xsamp
+   1   1.5   2   2.5   3
+ 0.09097 0.23939 0.33961 0.23851 0.09152
```

$\bar{X}$ as a Random Variable

Consider a sequence of $X_1, X_2, \dots, X_n$ i.i.d. random variables from this population, and define the random variable

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

We can compute the probability distribution of $\bar{X}$ using the same procedure.

- the calculation will be more complicated
- we may seek to approximate the distribution

$\bar{X}$ as a Random Variable

- We have a plan for data collection:
 - ▶ we will make n observations, independently and under identical conditions,
 - ▶ once we observe our sample data we will sum our observations and divide by n .
- We will store the observations as the quantities

$$X_1, X_2, \dots, X_n$$

- These quantities are random variables
- We use them to a new quantity: $\bar{X}$
 - ▶ $\bar{X}$ is itself a random variable,
 - ▶ its distribution is defined by the that of the X_i .

The Central Limit Theorem

Sample averages play a big part in probability and statistics

- We want to make probability statements about sample averages
- It is often difficult to say what the probability distribution of an average or mean of random variables is directly.

The Central Limit Theorem

- A theorem called the Central Limit Theorem which can come to our rescue, if the number of items we're adding or averaging is "large"
- As long as the random variables being added are independent and have the same probability distribution, if n is large enough, then their sum and their average have an approximate Normal distribution.

The Central Limit Theorem

Theorem: Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be independent and identically distributed random variables with population expectation μ and population standard deviation σ . Then if

$$S_n = X_1 + X_2 + \dots + X_n,$$

we have that

$$S_n \dot{\sim} N(n\mu, n\sigma^2)$$

when n is “large”.

The Central Limit Theorem

The notation

$\dot{\approx}$ means “is approximately distributed as”

that is, the distribution function of S_n is very close to that of the Normal.

We can use the Normal distribution function to compute interval probabilities for S_n : e.g.

$$P(l < S_n \leq u) \approx \Phi\left(\frac{u - n\mu}{\sqrt{n\sigma^2}}\right) - \Phi\left(\frac{l - n\mu}{\sqrt{n\sigma^2}}\right)$$

The Central Limit Theorem

Alternatively, if

$$\bar{X} = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$$

then

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

In practice, “large” is often taken to be $n \geq 30$.

The Central Limit Theorem

By results for the Normal distribution, we have

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} \text{Normal}(0, 1)$$

The Central Limit Theorem

Example: Previous example

If

$X = x$	1	2	3
$P(X = x)$	0.3	0.4	0.3

and $n = 50$.

The Central Limit Theorem

Example: Previous example

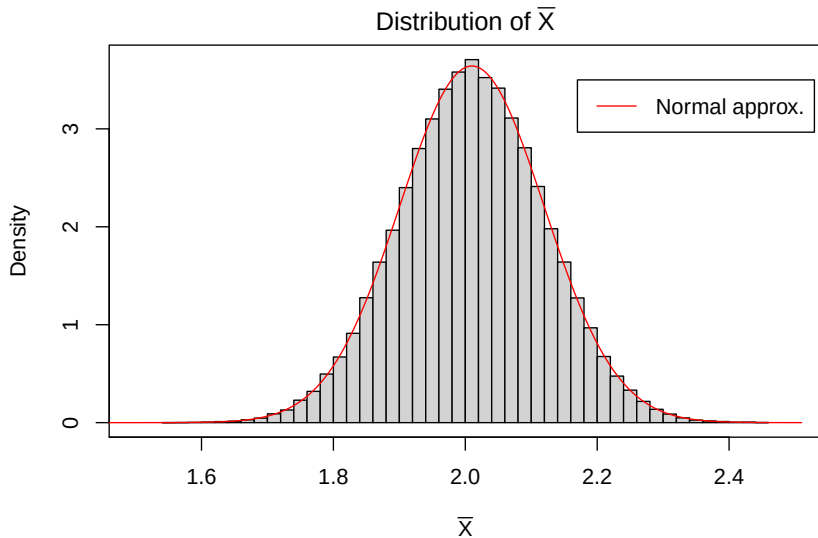
Then by the Central Limit Theorem

$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$$

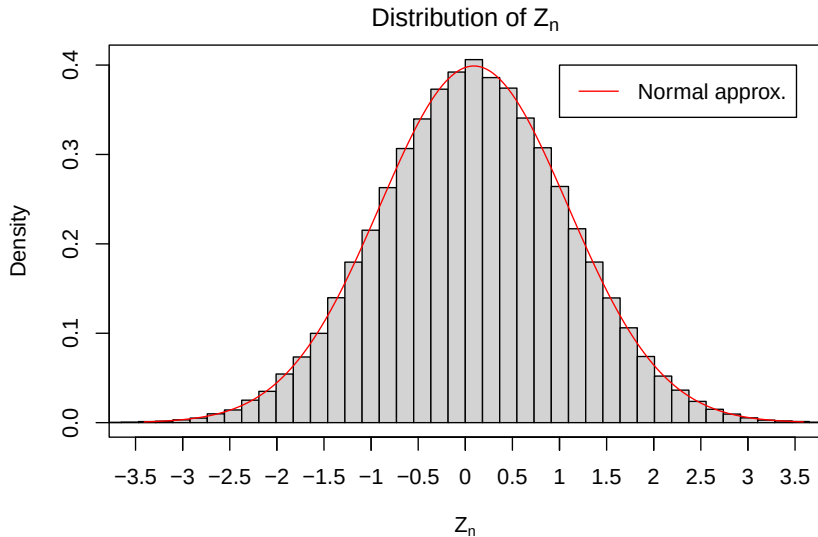
where $\mu = 2$ and $\sigma = \sqrt{0.6}$.

In the following plots, $n = 50$.

The Central Limit Theorem



The Central Limit Theorem



The Central Limit Theorem

Note

- If the individual random variables $X_1, \dots, X_n$ are themselves Normally distributed, then $\bar{X}$ and S_n are in fact **exactly** Normally distributed, no matter the value of n
- The Central Limit Theorem does not mean that a large number of independent random variables magically become Normal themselves - it's their sum/average that becomes (approximately) Normal

The Central Limit Theorem

The Central Limit Theorem is important in statistics, since $\bar{X}$ plays a crucial role in many areas of statistical inference

The Central Limit Theorem

Problem: CD4 counts

We are interested in looking at CD4 count in people with HIV. Suppose the CD4 count follows a Normal distribution, with mean 300 and standard deviation 80.

1. For a single randomly selected person with HIV, what is the probability that their CD4 count will be less than 200?
2. If we sample, at random and independently, 50 people with HIV, what is the probability that the sample average of their CD4 counts will be less than 200?

The Central Limit Theorem

Solution: CD4 counts

1. Let X be the CD4 count of a person with HIV: X is Normal with $\mu = 300$ and $\sigma = 80$

$$X \sim \text{Normal}(300, 80^2)$$

Note that the z-score transformation implies

$$Z = \frac{X - 300}{80} \sim \text{Normal}(0, 1)$$

The Central Limit Theorem

Solution: CD4 counts

We want to find $P(X < 200)$: using tables or software

$$\begin{aligned}P(X < 200) &= P\left(\frac{X - 300}{80} < \frac{200 - 300}{80}\right) \\&= P(Z < -1.25) \\&= 0.1056\end{aligned}$$

That is, the probability that a randomly chosen person with HIV has a CD4 count of less than 200.

The Central Limit Theorem

Solution: CD4 counts

2. We must find $P(\bar{X} < 200)$: we assume that we have a sequence $X_1, X_2, \dots, X_{50}$, where X_i represents the CD4 count of subject $i, i = 1, 2, \dots, 50$.

We know the each $X_i \sim \text{Normal}(300, 80^2)$, and it can be shown that

$$\bar{X} \sim \text{Normal}(300, 80^2/50)$$

that is, based on our assumptions about the individual X_i , we can show that $\bar{X}$ is also normally distributed, but with

$$\text{Mean} = 300 \quad \text{s.d.} = \frac{80}{\sqrt{50}} = 11.3137$$

The Central Limit Theorem

Solution: CD4 counts

Again using the z-score transform.

$$\begin{aligned}P(\bar{X} < 200) &= P\left(\frac{\bar{X} - 300}{80/\sqrt{n}} < \frac{200 - 300}{80/\sqrt{n}}\right) \\&= P(Z < -8.8388) \\&= 0.0000\end{aligned}$$

So, the probability that the average CD4 cell count will be less than 200 is approximately zero (to four decimal places).

The Central Limit Theorem

Problem: Book lengths

Suppose that the mean book length for science fiction books from a certain publisher is 350 pages, with a standard deviation of 42 pages. However, we do not know anything else about the distribution of the book lengths.

If we sample 75 books, what is the probability that their sample mean length will be greater than 360 pages?

The Central Limit Theorem

Solution: Book lengths

Let

$$\bar{X} = \frac{1}{75} \sum_{i=1}^{75} X_i$$

be the sample mean of the book lengths. We have

$$\mu = E(X_i) = 350 \quad \sigma = sd(X_i) = 42$$

for any $i = 1, \dots, 75$.

We want to find $P(\bar{X} > 360)$.

The Central Limit Theorem

Solution: Book lengths

$n = 75$ is large enough, so by the Central Limit Theorem

$$\bar{X} \dot{\sim} \text{Normal}(\mu, (\sigma/\sqrt{n})^2)$$

Again using the z-score transform.

$$\begin{aligned} P(\bar{X} > 360) &= P\left(\frac{\bar{X} - 350}{42/\sqrt{n}} > \frac{360 - 350}{42/\sqrt{n}}\right) \\ &= P(Z > 2.0620) \\ &= 1 - P(Z \leq 2.0620) \\ &= 1 - 0.9804 = 0.0196 \end{aligned}$$

The Central Limit Theorem

Solution: Book lengths

Thus

$$P(\bar{X} > 360) = 0.0196$$

Point estimators: big idea

We want to use data to summarize aspects of the population from which the data are drawn.

The population can be described by the parameters of the probability distribution that generates the data.

Therefore, we wish to use the data to construct our best representations of those parameters.

Point estimators

Suppose we have a population which is described through a distribution that depends on a parameter θ .

In this course we will take θ to be

- μ (population mean), or
- p in the Binomial set-up (population proportion)

We wish to “guess” at target parameter θ by observing a random sample of independent and identically distributed (i.i.d.) observations from this population

Point estimators

Let $X_1, X_2, \dots, X_n$ be the random variables which comprise this sample. We call any function of $X_1, \dots, X_n$ which does not depend on any unknown quantities an estimator of θ .

- We denote the **estimator** function by $\hat{\theta}$, i.e.

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$$

- ▶ $\hat{\theta}$ must be independent of any unknown parameters, otherwise we could not use it as an estimator of θ
- An estimator is a random variable since it depends on the random variables $X_1, \dots, X_n$
- θ itself is a constant (despite being unknown)

Point estimators

Once we conduct our experiment and have values $x_1, \dots, x_n$ for our random variables $X_1, \dots, X_n$.

The result $\hat{\theta}(x_1, \dots, x_n)$ is a numerical value, and will be considered an estimate of θ .

Properties of Estimators

Estimators are random variables and therefore have expected value and variance

For a given parameter θ , there may be many estimators

- Statisticians have agreed on two properties that indicate a “good” estimator:
 - ▶ Unbiasedness
 - ▶ Minimum Variance

Properties of Estimators

Unbiasedness: Estimator $\hat{\theta}$ is unbiased of θ for if and only if

$$E(\hat{\theta}) = \theta \quad \text{for all } \theta$$

- $\hat{\theta}$ is a random variable
- $\hat{\theta}$ has its own distribution
- this distribution has a population mean
- for unbiasedness, this population mean must be θ

Properties of Estimators

Minimum Variance: if $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for the same target parameter θ ,

$$E(\hat{\theta}_1) = E(\hat{\theta}_2) = \theta$$

we will always prefer the estimator with the smaller variance.

Estimating a population mean

Let $X_1, \dots, X_n$ be a random sample of i.i.d. random variables from a distribution with unknown expected value μ

It turns out that in almost all cases, the “best” estimator for μ (i.e. unbiased with the minimum variance among all the unbiased estimators) is

$$\hat{\mu} = \bar{X}$$

Estimating a population mean

Example: Painkiller duration

A drug manufacturer is estimating the duration (time from administration to relief of the pain) of their new over-the-counter painkiller.

- In this experiment we have a hypothetical population of potential users of the painkiller.
- We are interested in an estimate of the mean painkiller duration, say μ , in this population.
- The parameter μ is a constant, but is unknown to us.

Estimating a population mean

Example: Painkiller duration

Suppose that 18 subjects are taken at random and independently and the duration of the painkiller is measured on each of them. We have a sample of (i.i.d.) random variables

$$X_1, X_2, \dots, X_{18}$$

where X_i is the duration of the painkiller for subject $i, i = 1, 2, \dots, 18$.

Estimating a population mean

Example: Painkiller duration

The population mean duration μ will be estimated using the estimator $\bar{X}$.

If sample mean duration of these 18 subjects is $\bar{x} = 5.2$ hours, we say that μ is estimated to be 5.2 hours, and write

$$\hat{\mu} = 5.2$$

Estimating a population proportion

Let $X_1, \dots, X_n$ be i.i.d. Bernoulli random variables

$$P(X_i = 1) = p \quad P(X_i = 0) = 1 - p$$

with unknown “success” probability p .

The best estimator for p (i.e. unbiased with the minimum variance among all the unbiased estimators) is

$$\hat{p} = \frac{X}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

where X is the **total** number of successes in the sample

- $\hat{p}$ is the **observed proportion** of successes in the sample

Estimating a population proportion

Example: Student Stress

McGill is interested in estimating the stress levels of undergraduate student at the end of the semester.

A survey is to be conducted where the stress level will be ascertained as either high or low.

The population of interest in this experiment is all McGill undergraduate students

Estimating a population proportion

Example: Student stress

We are interested in estimating the proportion p of all McGill undergraduate students with high stress levels.

112 students were sampled randomly and independently, and asked to report their stress levels as either high or low.

We have a sample of (i.i.d. Bernoulli) random variables $X_1, \dots, X_{112}$; X_i is either

- a “success” (high) or
- a “failure” (low)

for student $i, i = 1, 2, \dots, 112$.

Estimating a population proportion

Example: Student stress

Suppose that 87 of these 112 students reported that their stress level is high.

The proportion p of all McGill undergraduate students with high stress levels is estimated to be the proportion of “successes” in the sample, i.e.

$$\hat{p} = \frac{87}{112} \approx 0.78$$

We say that an estimated 78% of McGill undergraduate students have high stress levels at the end of the semester.

Interval Estimation

Goal: to “guess” the (unknown) population parameter θ with an interval that, when used repeatedly, will capture the unknown θ on a high proportion of occasions.

Interval Estimation

We seek two random variables

$$L = L(X_1, \dots, X_n)$$

and

$$R = R(X_1, \dots, X_n)$$

such that for a specified probability α

$$P(L < \theta < R) = 1 - \alpha.$$

Interval Estimation

Both

$$L(X_1, \dots, X_n)$$

and

$$R(X_1, \dots, X_n)$$

are random variables, so have their own (sampling) probability distributions.

Interval Estimation

The interval (L, R) is called a $100(1 - \alpha)\%$ confidence interval (C. I.) for θ .

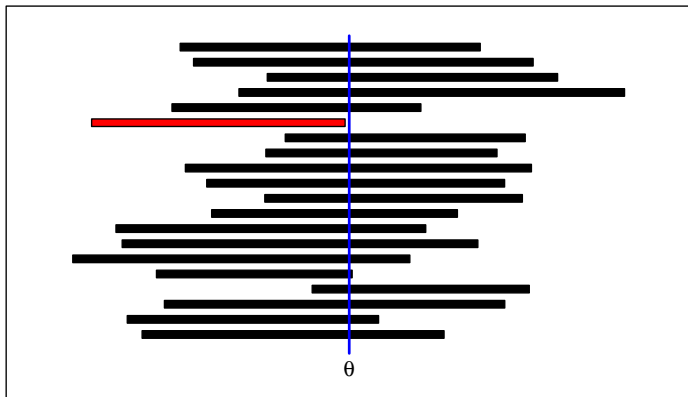
- This random interval (L, R) will 'cover' the true value θ with probability $1 - \alpha$

Once we use the observed data $x_1, \dots, x_n$ to compute the observed interval (l, r) , this interval has fixed endpoints and is no longer random

- e.g. $(3.1, 8.4)$ is not a random interval;
- this interval either contains θ or it does not

Interval Estimation

20 repeated computations of a confidence interval



Interval Estimation

We proceed as follows:

1. Objective: find a $100(1 - \alpha)\%$ C.I. for the parameter of interest: either μ or p .
 - The value of α will be given; if it is not, take $\alpha = 0.05$
2. Identify the setting from the wording of the problem: the formula for (L, R) depends on the setting
 - Once we know the setting, we look up formulas for L and R .
3. Substitute the known information into the formulas to obtain the observed interval (l, r)

Interval Estimation

We interpret the interval (L, R) as follows:

Textbook: “We are $100(1 - \alpha)\%$ confident that θ falls in (L, R) ”

- This is imprecise ...
- We add: “To be more precise, if we were to carry out the same study infinitely often under the same conditions, and each time compute a $100(1 - \alpha)\%$ C.I. using the same procedure, then $100(1 - \alpha)\%$ of these intervals will contain θ .”

Interval Estimation

Ideally we would like to be able to say something like

“ θ lies in (l, r) with probability $(1 - \alpha)$ ”

e.g.

“ θ lies in $(19.7, 22.4)$ with probability 0.95”.

However, this would require us to regard θ itself as random.

The t -distribution

We will see that construction of confidence intervals depends on understanding the sampling distribution of different statistical summaries.

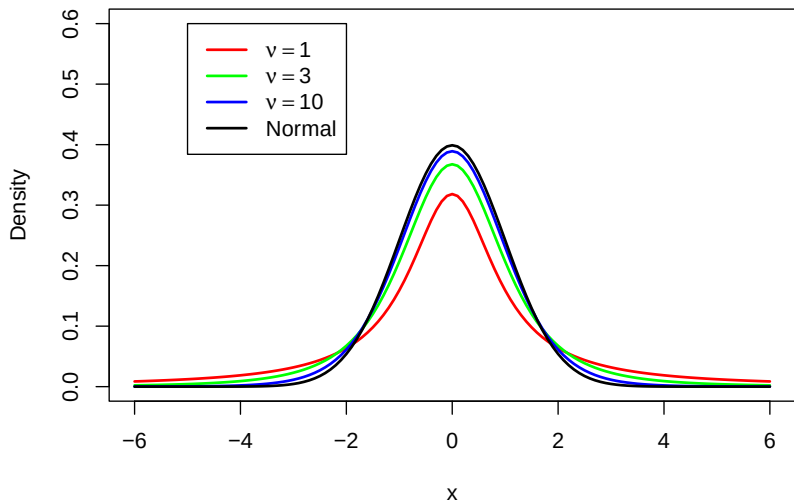
- the Normal distribution plays an important role
- the Student- t -distribution (or just t -distribution) is also needed
 - ▶ The t -distribution is specified through a parameter ν called its degrees of freedom (d.f.)
 - ▶ Each ν gives a different curve, but this curve is always symmetric about 0.

The t -distribution

The t -distribution has heavier tails than the Normal

- the t -distribution allocates more probability to tail regions
- the standard Normal and the t -distribution get more similar as ν gets larger
- as $\nu \rightarrow \infty$, the t -distribution becomes identical to the standard Normal.

The t -distribution

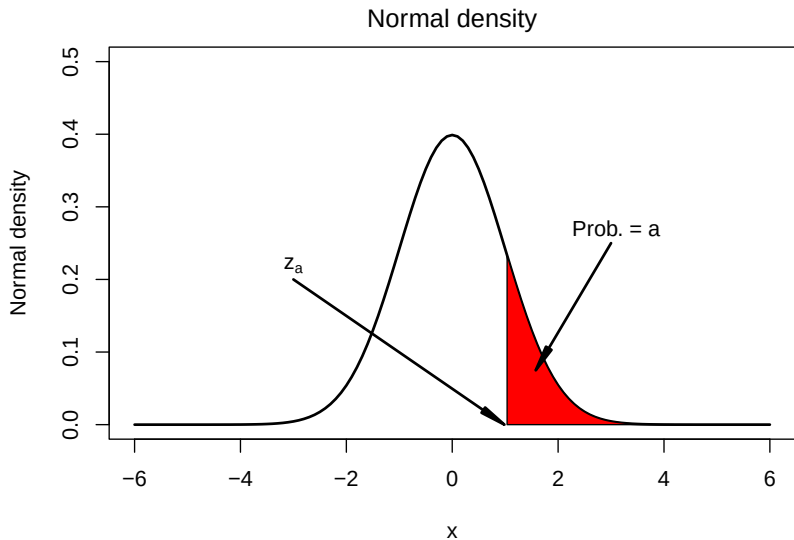


The t -distribution

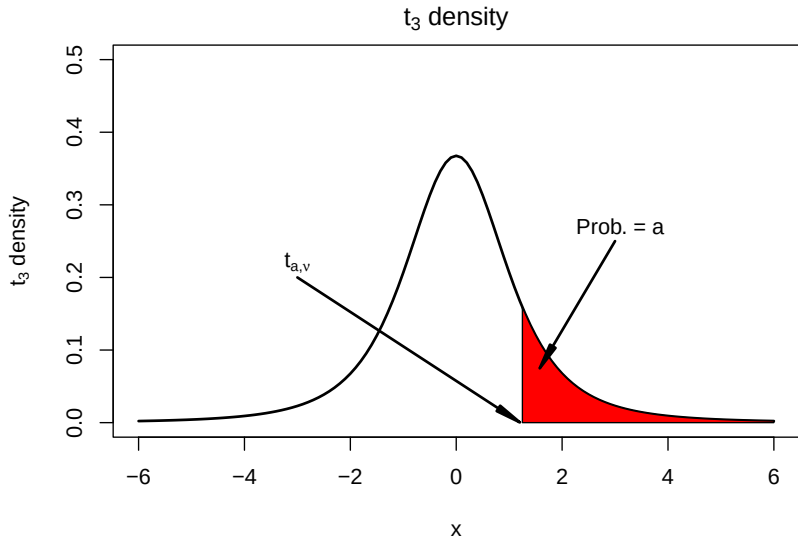
In order to calculate confidence intervals we first define the notation z_a and $t_{a,\nu}$, the “upper a point” of these distributions

- The upper a point of a standard Normal distribution is the point on the horizontal axis, z_a such that $P(Z > z_a) = a$.
- The upper a point of a t -distribution with ν degrees of freedom is the point on the horizontal axis, $t_{a,\nu}$ such that $P(T > t_{a,\nu}) = a$, where the random variable T has a t -distribution with ν degrees of freedom.

The t -distribution



The t -distribution



The t -distribution

In these plots: $a = 0.15$ (shaded red)

- Normal case: $z_a = 1.0364$
- t_3 case: $t_{a,3} = 1.2498$

so note that

$$t_{a,3} > z_a.$$

The t -distribution

ν	α				
	0.1	0.05	0.025	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	318.3088
3	1.6377	2.3534	3.1824	4.5407	10.2145
10	1.3722	1.8125	2.2281	2.7638	4.1437
30	1.3104	1.6973	2.0423	2.4573	3.3852
50	1.2987	1.6759	2.0086	2.4033	3.2614
100	1.2901	1.6602	1.9840	2.3642	3.1737
∞	1.2816	1.6449	1.9600	2.3263	3.0902

Note: $\nu = \infty$ corresponds to $Normal(0, 1)$.

Experiment Scenarios

The relevant experiment scenario can be determined by establishing

- is the parameter of interest, a population mean μ , or a population proportion p ?
- is the sample size (n) large or small?

Experiment Scenarios

To determine the correct strategy, we need to identify

(A) the assumed probability model/parameter of interest

- Normal, μ
- Binomial, p

(B) the sample size assumption

- n small ($n < 30$)
- n large ($n \geq 30$)

(C) the variance assumption for the Normal case

- σ known
- σ unknown

Parameter μ , Small n

Theorem: Scenario 1

Suppose that $X_1, \dots, X_n$ are a random sample from a Normal distribution with unknown μ and σ , and $n < 30$.

A $100(1 - \alpha)\%$ confidence interval for μ is (L, R) , where:

$$L = \bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$$
$$R = \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}$$

Parameter μ , Small n , σ known

Theorem: Scenario 2

Suppose that $X_1, \dots, X_n$ are a random sample from a Normal distribution with unknown μ , with σ known.

A $100(1 - \alpha)\%$ confidence interval for μ is (L, R) , where:

$$L = \bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$
$$R = \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Parameter μ , Large n

Theorem: Scenario 3

Suppose that $X_1, \dots, X_n$ are a random sample from a distribution with unknown μ and σ , where $n \geq 30$.

A $100(1 - \alpha)\%$ confidence interval for μ is (L, R) , where:

$$L = \bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}}$$
$$R = \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}}$$

Note: Normality of the X s is not required.

Parameter p

Theorem: Scenario 4

Let $X_1, \dots, X_n$ be a random sample of Bernoulli random variables with unknown probability of success p .

A $100(1 - \alpha)\%$ confidence interval for p is (L, R) , where:

$$L = \hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$
$$R = \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

where

$$\hat{p} = \frac{\text{Number of "successes" in the sample}}{n}$$

Parameter p

Theorem: Scenario 4

Note: “Large” sample size in this scenario means that

$$n\hat{p} \geq 15 \text{ and } n(1 - \hat{p}) \geq 15$$

Confidence Interval: Examples

Problem: Painkiller duration

A drug manufacturer is testing the duration of pain-relief of their new over-the-counter painkiller.

Duration is measured on 18 subjects, with sample mean of $\bar{x} = 5.2$ hours, and sample standard deviation of $s = 1.1$ hours.

Calculate a 95% confidence interval for the true mean duration μ ; state any assumptions you make.

Confidence Interval: Examples

Solution: Painkiller duration

We are in Scenario 1: $n < 30$ and we are trying to find a 95%CI for μ , with $\alpha = 0.05$.

In order to calculate the CI, we have to assume that the X_i s are Normally distributed: then the 95%CI for the unknown μ .

We have

- $n = 18 < 30$.
- $\bar{x} = 5.2$
- $s = 1.1$
- $\alpha = 0.05$
- $t_{\alpha/2, n-1} = t_{0.025, 17} = 2.110$

Confidence Interval: Examples

Solution: Painkiller duration

Hence the CI is

$$\begin{aligned}(l, r) &: \left(\bar{x} - t_{0.025, 17} \frac{s}{\sqrt{n}}, \bar{x} + t_{0.025, 17} \frac{s}{\sqrt{n}} \right) \\&= \left(5.2 - 2.110 \times \frac{1.1}{\sqrt{18}}, 5.2 + 2.110 \times \frac{1.1}{\sqrt{18}} \right) \\&= (4.654, 5.746)\end{aligned}$$

measured in hours.

Confidence Interval: Examples

Solution: Painkiller duration

Interpretation: We are 95% confident that μ , the mean population duration of relief, is between 4.65 and 5.75 hours

To be more precise, if this study were repeated infinitely often under identical conditions, and on each occasion a 95% confidence interval were calculated using the technique described here, then 95% of these intervals would include the true, but unknown, mean duration μ .

Confidence Interval: Examples

Problem: Painkiller duration

Now suppose that the drug is tested on 65 subjects, with a sample average of $\bar{x} = 5.8$ hours and a sample standard deviation of $s = 0.8$ hours

Calculate a 99% confidence interval for the true mean duration μ .

Confidence Interval: Examples

Solution: Painkiller duration

We are now in Scenario 3: $n > 30$ and we are finding a 99%CI for μ , with $\alpha = 0.01$.

We have

- $n = 65 > 30$.
- $\bar{x} = 5.8$
- $s = 0.8$
- $\alpha = 0.01$
- $z_{\alpha/2} = z_{0.005} = 2.576$

Confidence Interval: Examples

Solution: Painkiller duration

$$\begin{aligned} & \left(\bar{x} - z_{0.005} \frac{s}{\sqrt{n}}, \bar{x} + z_{0.005} \frac{s}{\sqrt{n}} \right) \\ &= \left(5.8 - 2.576 \times \frac{0.8}{\sqrt{65}}, 5.8 + 2.576 \times \frac{0.8}{\sqrt{65}} \right) \\ &= (5.545, 6.055) \end{aligned}$$

Confidence Interval: Examples

Problem: Student stress

McGill is conducting a survey on undergrad student stress levels at the end of the semester.

- 112 students are selected, and asked to report their stress levels;
- 87 of these students report that their stress level is high.

Calculate a 90% confidence interval for the proportion of McGill students with a high stress level.

Confidence Interval: Examples

Solution: Student stress

We are in Scenario 4: a Binomial setup, and we are finding a 90%CI for p , with $\alpha = 0.1$.

From the data,

$$\hat{p} = \frac{87}{112} = 0.777$$

and

$$\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = \sqrt{\frac{0.777(1 - 0.777)}{112}} = 0.039$$

rounding to three decimal places. We also have

$$z_{0.05} = 1.645$$

Confidence Interval: Examples

Solution: Student stress

Therefore

$$\begin{aligned}& \left(\hat{p} - z_{0.05} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{0.05} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right) \\&= (0.777 - 1.645 \times 0.039, 0.777 + 1.645 \times 0.039) \\&= (0.713, 0.841)\end{aligned}$$

Confidence Interval: Examples

Solution: Student stress

Checking the large sample assumption:

$$n\hat{p} = 87 \geq 15$$

and

$$n(1 - \hat{p}) = 27 \geq 15$$

so the large sample assumption holds.

The assumption of Normality

When estimating a population mean μ if the sample size is “small” ($n < 30$), it can only be done under the assumption that the population has a Normal distribution

- Note that we cannot really “prove” the assumption, as it is an assumption on the population.
- However, in practice, it is not enough to state the assumption of Normality of the population, we need to show some evidence to support this assumption

The assumption of Normality

- One way to do this is to use the descriptive methods to assess Normality discussed in previous lectures:
 - ▶ histograms
 - ▶ boxplots
- If the assumption holds for a random sample, then it may hold for the population
- We just need to show, from the sample data, that we have at least an approximation to Normality

Assessing Normality

Many wide-spread techniques for statistical inference require data that are (approximately) Normally distributed.

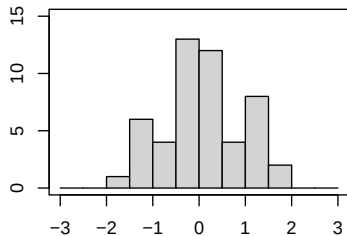
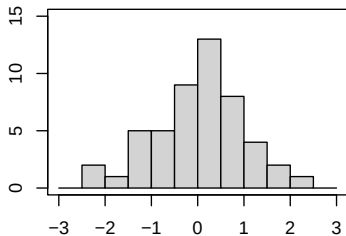
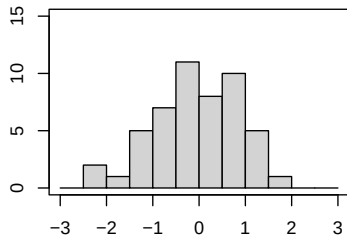
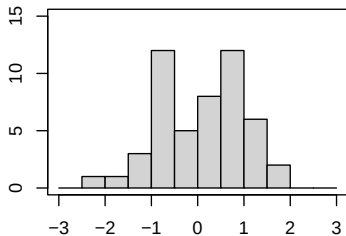
- Normally distributed random variables are found in many settings.
- Important question: How can we tell if our data are (approximately) Normally distributed?

Assessing Normality

If the data are approximately Normal, the shape of the histogram will be mound shape and symmetric, similar to the Normal probability density function

However, this is not always easy to assess from a histogram:

Assessing Normality



Q-Q Plots

A more reliable method for checking normality: normal probability plot (also called a quantile-quantile (Q-Q) plot.)

A Q-Q plot compares the quantiles from the sample with the theoretical quantiles from the normal distribution.

Q-Q Plots

In the plot:

- the x coordinates are the values

$$F^{-1}(j/(n+1)) \quad j = 1, 2, \dots, n$$

that is, the **theoretical** quantile corresponding to probability $j/(n+1)$, where F^{-1} is the inverse of the appropriate Normal distribution function.

Recall that if $X \sim \text{Normal}(\mu, \sigma^2)$, then the distribution function for X can be computed

$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

For the Q-Q plot, we estimate μ and σ from the data.

Q-Q Plots

- the y coordinates are the ordered data values, sorted into ascending order

$$y_j = x_{(j)}$$

that is, the **sample** quantiles corresponding to probabilities j/n .

Q-Q Plots

We often transform the data onto a standardized (z-score) scale

$$x \longrightarrow z = \frac{x - \bar{x}}{s}$$

before plotting the Q-Q plot.

We expect to see the points in the plot with

x-coordinate: theoretical Normal quantile

y-coordinate: sample quantile for z

lie on the line $y = x$; if they do not, then the Normal assumption is questionable.

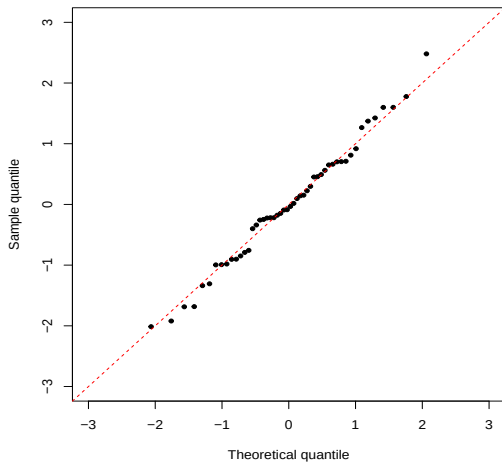
Q-Q Plots

```
n<-50
x<-rnorm(n,10,2)
xbar<-mean(x)
sval<-sd(x)
c(xbar,sval)

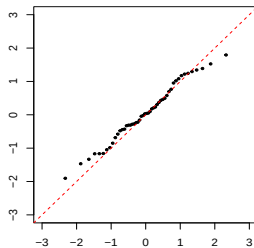
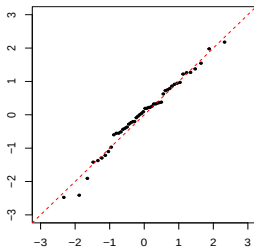
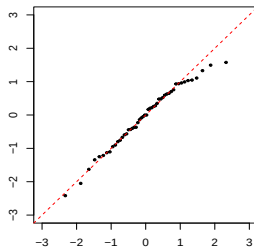
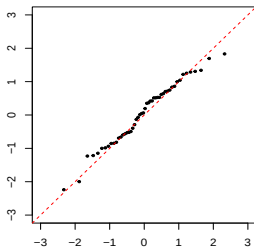
+ [1] 9.652594 2.044429

z<-(x-xbar)/sval
qq.x<-qnorm(c(1:n)/(n+1))
qq.y<-sort(z)
par(pty='s')
plot(qq.x,qq.y,pch=19,cex=0.7,xlim=range(-3,3),ylim=range(-3,3),
     xlab='Theoretical quantile', ylab='Sample quantile')
abline(0,1,lty=2,col='red')
```

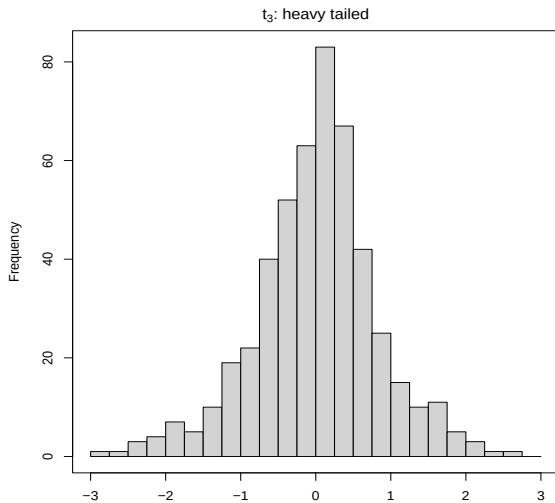
Q-Q Plots



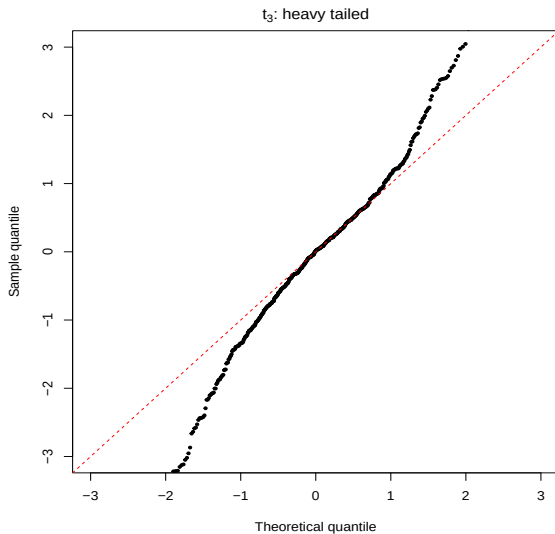
Q-Q Plots



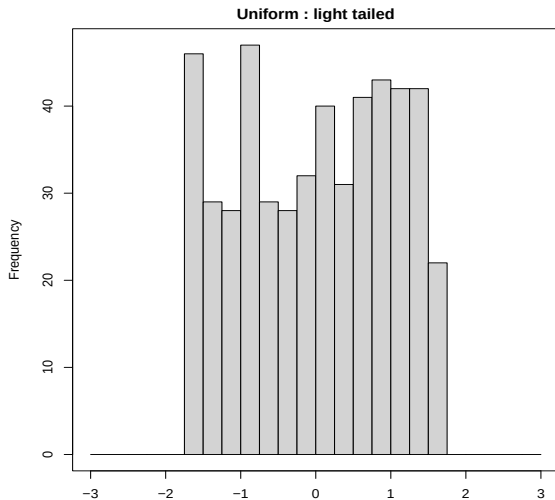
Q-Q Plots



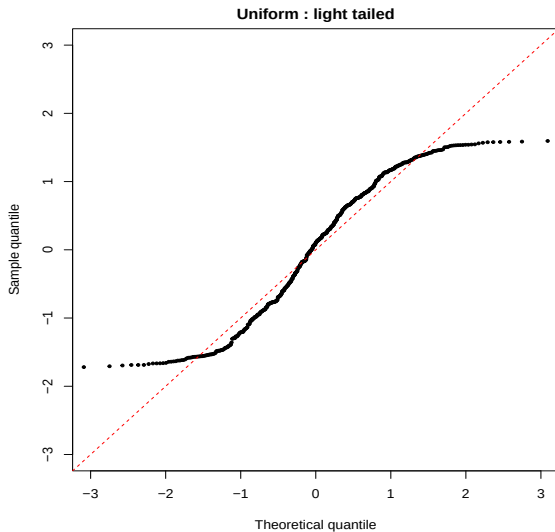
Q-Q Plots



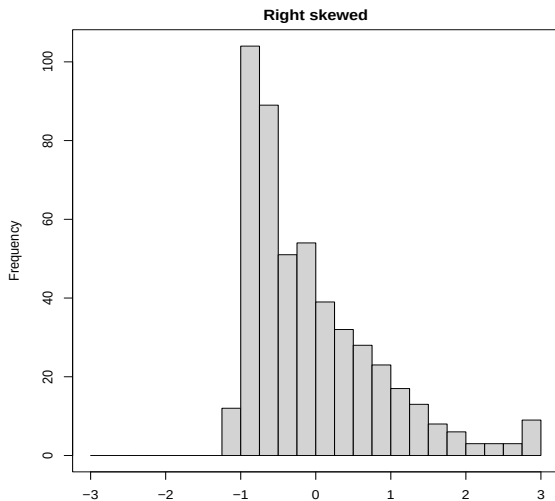
Q-Q Plots



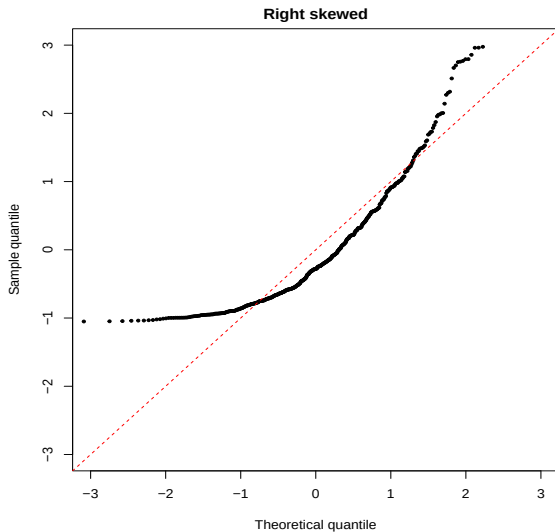
Q-Q Plots



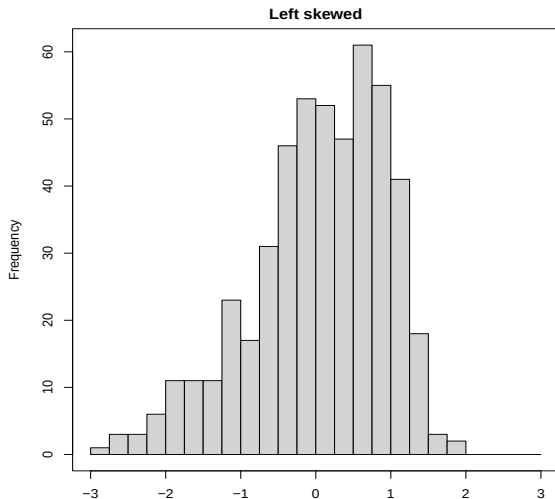
Q-Q Plots



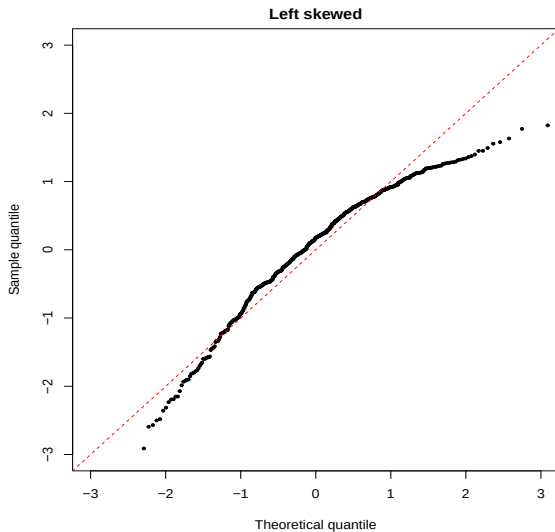
Q-Q Plots



Q-Q Plots



Q-Q Plots



Hypothesis Testing: big idea

Previously:

- Estimated θ through a single point $\hat{\theta}$
- Estimated θ through an interval (L, R)
- Now: Want to make a decision relating to θ

In order to make such decisions, based on data, in a rigorous manner, we have to use hypothesis testing.

Hypothesis Testing: big idea

Idea: We will have two competing hypotheses about a population parameter θ (for example, μ and p)

- We will use our sample data to decide which of the two hypotheses is true

NB: The goal is not to know which hypothesis is true (We will never know unless we could observe the full population data)

Hypothesis Testing: big idea

The two hypotheses are called

- the null hypothesis, H_0 ,
- the alternative hypothesis, H_a .

As for θ in confidence intervals, the truth of H_0 versus H_a is a fixed state.

We are trying to assess which is true, and our “guess” will be either correct or not.

Hypothesis Testing: null hypothesis

H_0 generally represents the perceived “status quo”:

- 70% of McGill students have high stress levels
- Two lakes have the same pollution levels
- A new treatment has no effect on recovery time
- Students in two schools have, on average, the same test scores

Hypothesis Testing: alternative hypothesis

H_a generally represents a deviation from the “status quo”:

- More than 70% of McGill students have high stress levels
- One lake has higher mean pollution levels than another
- A new treatment reduces mean surgery recovery time
- Students in two schools have appreciably different average test scores

H_a is always based only on the a priori belief about the question of interest, never on the data themselves.

Hypothesis Testing: steps

1. Determine H_0 and H_a from the question (null hypothesis and alternative hypothesis)
2. Determine an appropriate Test Statistic T :
 - T is a random variable whose value is determined by the individual sample random variables
3. Identify the (sampling) distribution of T assuming H_0 is true.
4. Determine the Rejection Region:
 - this will be a set of possible values of the Test Statistic T for which we will “reject” H_0 “in favour” of H_a

Hypothesis Testing: steps

5. Compute the observed value of T from the observed data
6. Compare the observed value of the T to the Rejection Region
7. Decide and conclude
 - “Reject H_0 in favour of H_a ”, or
 - “Do not reject H_0 ”

Hypothesis Testing: steps

Note

We

- compute the sampling distribution of T ,
- form the Rejection Region

assuming that H_0 is true.

This sampling distribution is called the **null** distribution; we will write P_0 to denote probabilities computed under the null distribution.

Type 1 and Type 2 Error

When we make a decision on the hypotheses, we have the potential to make one of two possible errors:

- A Type 1 error occurs when we reject H_0 when, in fact, H_0 is true.
- A Type 2 error occurs when we fail to reject H_0 when, in fact, H_0 is not true (H_a is true).

Type 1 and Type 2 Error

In creating hypothesis testing procedures,

- choosing T
- computing the Rejection Region

it is mathematically easier to control Type 1 error than it is to control Type 2 error

Thus, we set up our hypotheses such that the more “serious” error will be a Type 1 error, and try to ensure that the Type 2 error is not too large.

Type 1 and Type 2 Error

- The probability of a Type 1 error is denoted α , so

$$\alpha = P_0(T \text{ lies in R. R.})$$

- The probability of a Type 2 error is denoted β .

Both quantities will depend on how we choose H_0 and H_a .

Type 1 and Type 2 Error

We will study three types of sets of hypotheses regarding a parameter θ (i.e. μ or p):

$$\begin{aligned}\text{One-sided: } H_0 : \theta &= \theta_0 \\ H_a : \theta &> \theta_0\end{aligned}$$

$$\begin{aligned}\text{One-sided: } H_0 : \theta &= \theta_0 \\ H_a : \theta &< \theta_0\end{aligned}$$

$$\begin{aligned}\text{Two-sided: } H_0 : \theta &= \theta_0 \\ H_a : \theta &\neq \theta_0\end{aligned}$$

Which one we use is based on the wording of the question (and prior belief about θ).

Test Statistics and rejection regions

We will see different experiment scenarios.

- Pre-experiment: the test statistic is a random variable T

$$T = T(X_1, \dots, X_n)$$

- Post-experiment, once we substitute our observed data, the test statistic is simply a number:
 - we will denote this number T_{obs}

$$T_{obs} = T(x_1, \dots, x_n)$$

Test Statistics and rejection regions

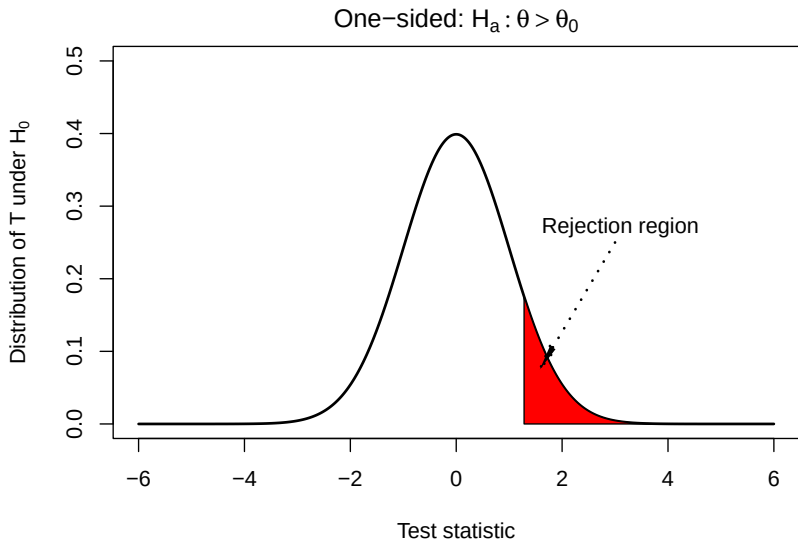
The Rejection region is in the form of a set R , which is determined by H_0 , H_a , α , and the scenario, and is not random.

- If T_{obs} falls within R , we will **reject** H_0 in favour of H_a ;
- If it does not, we will not reject H_0

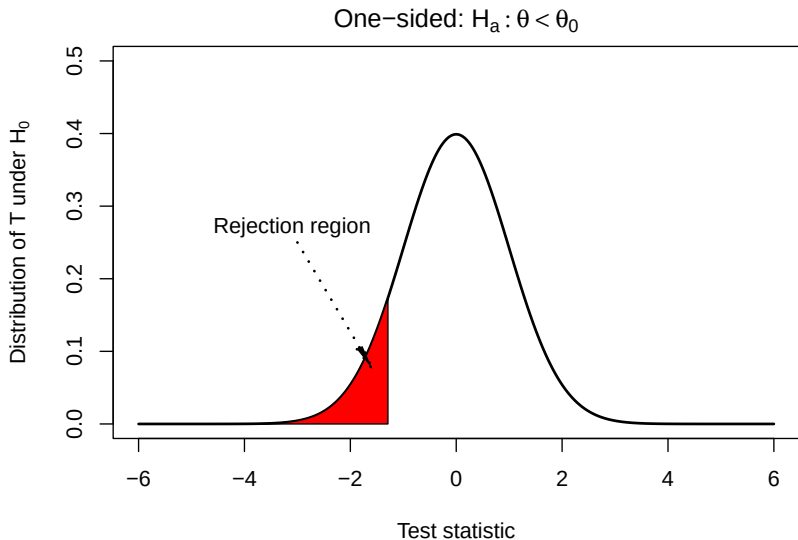
Test Statistics and rejection regions

- Pre-experiment: we may assess the probability that T falls within R
- Post-experiment: either T_{obs} fell within R or not, and as a result we either rejected or did not reject H_0 .

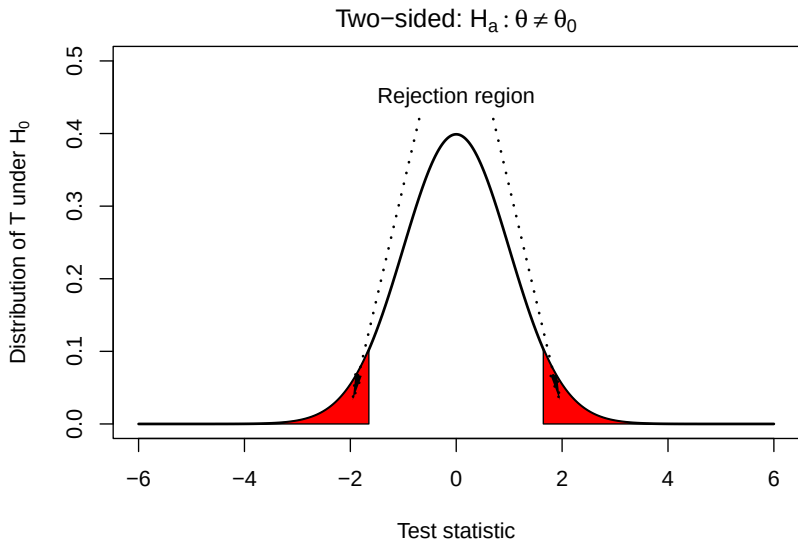
Test Statistics and rejection regions



Test Statistics and rejection regions



Test Statistics and rejection regions



Test Statistics and rejection regions

- All conclusions are made “at the α level”, as the chosen value of α affects your decision.
- If the test statistic falls within the rejection region: you reject H_0 in favour of H_a .
- If the test statistic does not fall within the rejection region: you fail to reject H_0 ;
 - ▶ you do not accept H_0 as true, you merely do not have sufficient evidence to decide that it is false.

Test Statistics and rejection regions

Either H_0 or H_a is true, but we do not know which.

After the decision, we are in one of four different situations:

1. Rejected H_0 , H_a true
2. Rejected H_0 , H_0 true (Type 1 error)
3. Failed to reject H_0 , H_0 true
4. Failed to reject H_0 , H_a true (Type 2 error)

		Reject H_0 ?	
		Yes	No
H_0 true ?	Yes	2.	3.
	No	1.	4.

Test Statistics and rejection regions

Since we do not know which of H_0 and H_a is true, we do not know which of these situations has occurred.

We design our hypothesis testing procedure such that, on repeated trials, it will falsely reject H_0 (i.e. make a Type 1 error) $100\alpha\%$ of the time

One sample cases

In this course, we will see three types of sets of hypotheses, regarding the parameters μ or p : for null hypothesis

$$H_0 : \mu = \mu_0 \quad \text{or} \quad H_0 : p = p_0,$$

and for the alternative hypothesis, one of

$$H_a : \mu > \mu_0 \quad \text{or} \quad H_a : p > p_0$$

$$H_a : \mu < \mu_0 \quad \text{or} \quad H_a : p < p_0$$

$$H_a : \mu \neq \mu_0 \quad \text{or} \quad H_a : p \neq p_0$$

Which one we use is based on the wording of the question (and the prior belief about θ).

One sample cases

In this specification, μ_0 and p_0 represent specific **numerical** values, for example

- $H_0 : \mu = 10.0$
- $H_0 : p = 0.5$

that represent the hypothesized null values of the data generating parameters.

One sample cases

- Population Mean μ
 1. Small $n(< 30)$
 2. Large $n(\geq 30)$
- Population proportion p

One sample cases

Recap on notation:

- $\bar{X}$: sample mean random variable
- S^2 : sample variance random variable

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

- t_ν : t -distribution with ν degrees of freedom
- z_a : upper a point of the standard Normal distribution
- $t_{a,\nu}$: upper a point of the t -distribution with ν degrees of freedom

One sample cases

Case: 1. Parameter μ , small n , σ unknown

Suppose that $X_1, \dots, X_n$ are a random sample from a Normal distribution with unknown μ and σ , and $n < 30$.

$$H_0 : \mu = \mu_0$$

and

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

H_a	$H_a : \mu > \mu_0$	$H_a : \mu < \mu_0$	$H_a : \mu \neq \mu_0$
R. R.	$\{T \geq t_{\alpha, n-1}\}$	$\{T \leq -t_{\alpha, n-1}\}$	$\{ T \geq t_{\alpha/2, n-1}\}$

One sample cases

Case: 2. Parameter μ, σ known

Suppose that $X_1, \dots, X_n$ are a random sample from a Normal distribution with unknown μ , with σ known.

$$H_0 : \mu = \mu_0$$

and

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \text{Normal}(0, 1)$$

$$H_a \quad H_a : \mu > \mu_0 \quad H_a : \mu < \mu_0 \quad H_a : \mu \neq \mu_0$$

$$\text{R. R.} \quad \{T \geq z_\alpha\} \quad \{T \leq -z_\alpha\} \quad \{|T| \geq z_{\alpha/2}\}$$

Note: Here T is often denoted Z , as it is a z-score.

One sample cases

Case: 3. Parameter μ , large n , σ unknown

Suppose that $X_1, \dots, X_n$ are a random sample from a distribution with unknown μ and σ , where $n \geq 30$.

$$H_0 : \mu = \mu_0$$

and

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \dot{\sim} \text{Normal}(0, 1)$$

$$H_a \qquad H_a : \mu > \mu_0 \quad H_a : \mu < \mu_0 \quad H_a : \mu \neq \mu_0$$

$$\text{R. R.} \qquad \{T \geq z_\alpha\} \quad \{T \leq -z_\alpha\} \quad \{|T| \geq z_{\alpha/2}\}$$

Normality of the X s is not required.

One sample cases

Case: 4. Parameter p , n large

Let $X_1, \dots, X_n$ be a random sample of Bernoulli random variables with unknown probability of success p

$$H_0 : p = p_0$$

and

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} \text{Normal}(0, 1)$$

$$H_a \quad H_a : p > p_0 \quad H_a : p < p_0 \quad H_a : p \neq p_0$$

$$\text{R. R.} \quad \{T \geq z_\alpha\} \quad \{T \leq -z_\alpha\} \quad \{|T| \geq z_{\alpha/2}\}$$

One sample cases

Problem: Painkiller duration

A drug manufacturer is testing the duration of their new over-the-counter painkiller.

Duration is measured on 18 subjects, with sample mean $\bar{x} = 5.2$ hours and sample standard deviation $s = 1.1$ hours.

Previous research has shown that ibuprofen has a mean duration of 4.7 hours.

Conduct an appropriate test of hypotheses at the 5% level to determine whether the new drug lasts longer than ibuprofen. State any assumptions you make.

One sample cases

Solution: Painkiller duration (one-sided)

We are in Case 1. We must assume that the observations X_i are Normally distributed. We are testing

$$H_0 : \mu = 4.7$$

$$H_a : \mu > 4.7$$

that is, we are carrying out a one-sided test. We have

- $n = 18$
- $\bar{x} = 5.2$
- $s = 1.1$
- $\mu_0 = 4.7$

so n is small, and σ is unknown.

One sample cases

Solution: Painkiller duration (one-sided)

Here

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

Thus, from the data, the observed value is

$$T_{obs} = \frac{5.2 - 4.7}{1.1/\sqrt{18}} = 1.928.$$

One sample cases

Solution: Painkiller duration (one-sided)

$$\text{R. R. : } \{T \geq t_{0.05,17} = 1.740\}$$

Since

$$T_{obs} = 1.928 > 1.740$$

we **reject** H_0 in favour of H_a at the 5% level.

One sample cases

Problem: Painkiller duration (two-sided)

Suppose that the manufacturer claims that the drug has a mean duration of 6 hours.

The drug is tested on 65 subjects, with a sample mean of $\bar{x} = 5.8$ hours and a sample standard deviation of $s = 0.8$ hours.

Conduct an appropriate test of hypotheses at the 1% level

One sample cases

Solution: Painkiller duration (two-sided)

We are in Case 3. We are testing

$$H_0 : \mu = 6.0$$

$$H_a : \mu \neq 6.0$$

that is, we are carrying out a two-sided test. We have

- $n = 65$
- $\bar{x} = 5.8$
- $s = 0.8$
- $\mu_0 = 6.0$

so n is large.

One sample cases

Solution: Painkiller duration (two-sided)

Here

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

We have

$$T_{obs} = \frac{5.8 - 6.0}{0.8/\sqrt{65}} = -2.016.$$

and

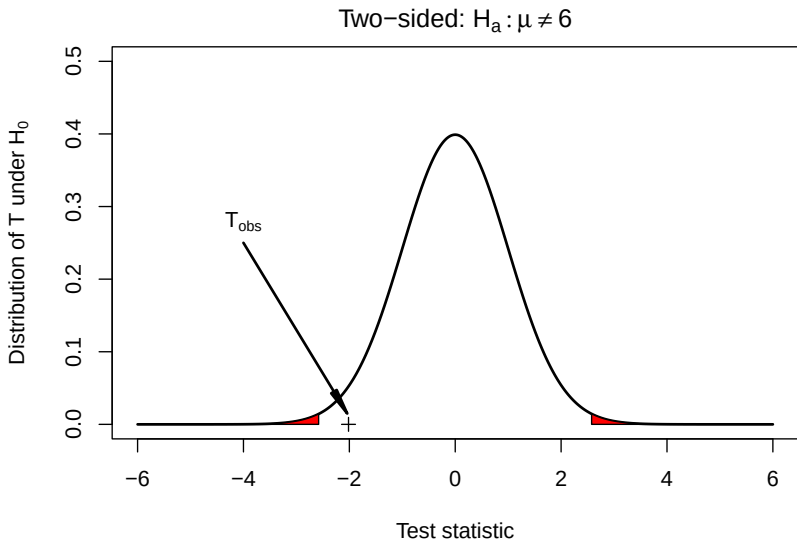
$$\text{R. R. : } \{|T| \geq z_{0.005} = 2.576\}$$

Since

$$|T_{obs}| = |-2.016| = 2.016 < 2.576$$

we **do not reject** H_0 at the 1% level.

One sample cases



One sample cases

Problem: Student stress (one-sided)

McGill is conducting a survey on undergrad student stress levels at the end of the semester, to test whether stress levels are lower than usual. Previous studies have found that 85% of students report high stress.

112 students are recruited, and asked to report their stress levels as either high or low.

87 of these students report that their stress level is high.

Conduct an appropriate test of hypotheses at the 5% level

One sample cases

Solution: Student stress (one-sided)

We are in Case 4. We are testing

$$H_0 : p = 0.85$$

$$H_a : p < 0.85$$

that is, we are carrying out a one-sided test. We have

- $n = 112$
- $x = 87$
- $\hat{p} = 0.777$
- $p_0 = 0.85$

so n is large.

One sample cases

Solution: Student stress (one-sided)

The test statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Thus

$$T_{obs} = \frac{0.777 - 0.850}{\sqrt{\frac{0.850(1 - 0.850)}{112}}} = -2.170$$

One sample cases

Solution: Student stress (one-sided)

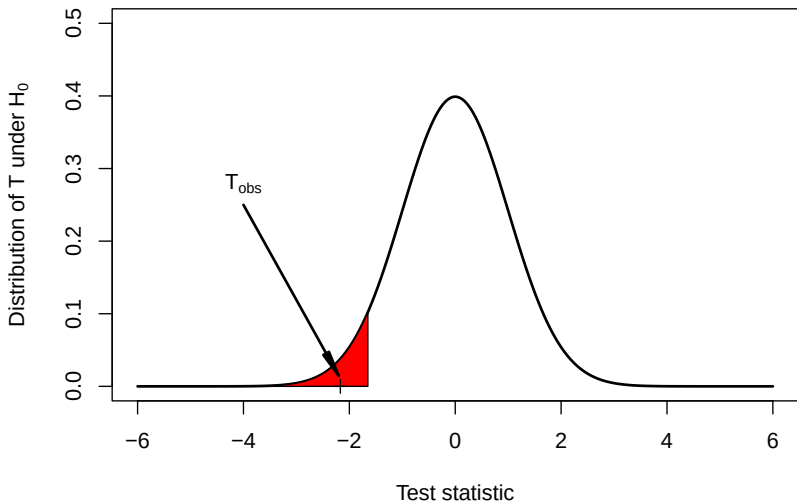
Here

$$\text{R. R. : } \{T \leq -z_{0.05} = -1.645\}$$

Since $-2.170 < -1.645$, we reject H_0 in favour of H_a at the 5% level

One sample cases

One-sided: $H_a : p < 0.85$



One sample cases

Note: computation in R

```
n<-112
x<-87
phat<-x/n
p0<-0.85
se.hat<-sqrt(p0*(1-p0)/n)
alph<-0.05
t.obs<-(phat-p0)/se.hat
z.a<-qnorm(alph)
c(phat,se.hat,t.obs,z.a)

## [1] 0.77678571 0.03374008 -2.16995011 -1.64485363
```

One sample cases

Note: computation in R with rounding to 2 d.p.

```
n<-112
x<-87
phat<-round(x/n,2)
p0<-0.85
se.hat<-round(sqrt(p0*(1-p0)/n),2)
alph<-0.05
t.obs<-round((phat-p0)/se.hat,2)
z.a<-round(qnorm(alph),2)
c(phat,se.hat,t.obs,z.a)

## [1] 0.78 0.03 -2.33 -1.64
```

So rounding does change the value of the T_{obs} statistic.

Many practitioners are uncomfortable with selecting α before seeing their data

- The choice of α is somewhat arbitrary
- Once you look at your data, you can no longer change α

Hypothesis testing does not take into account where T_{obs} lies in (or out of) the rejection region.

- If you fail to reject H_0 , is T_{obs} very close to, or farther away from, the edge of the R. R. ?
- If you reject H_0 in favour of H_a , is T_{obs} close to, or far away from, the edge of the R. R.?

p -values

It is common to avoid a decision-based approach and instead just report a p -value.

The p -value is a probability that may be thought of as “evidence for H_a ”, and reporting it does not force you to make a decision.

p -values: definition

For a test of hypotheses H_0 and H_a with test statistic T and its observed value T_{obs} , the p -value is defined as follows:

Assuming H_0 is **true**

- For $H_a : \theta > \theta_0$: $p\text{-value} = P_0(T \geq T_{obs})$
- For $H_a : \theta < \theta_0$: $p\text{-value} = P_0(T \leq T_{obs})$
- For $H_a : \theta \neq \theta_0$: $p\text{-value} = 2 \times P_0(T \geq |T_{obs}|)$

p -values: interpretation

In essence: the p -value is the probability of obtaining a value of T at least as “extreme” as T_{obs} , assuming that H_0 is true,

- “extreme” is determined by the “direction” of H_a .

p -values: interpretation

If p -value is “small”: this could be due to two different reasons:

- H_0 is true and the data we observed are merely an unusual outcome (by chance)
- H_0 is not true; H_a is a preferable explanation of the observed data

The smaller the p -value, the less we believe the first case, and the more we believe the second.

p -values: interpretation

If the p -value is large, then the evidence against H_0 is weak.

This does not mean that you believe H_0 is true, just that your current data does not give enough evidence to reject it in favour of H_a .

p -values: implementation

1. Determine H_0 and H_a
2. Determine T and its distribution assuming H_0 is true
3. Compute T_{obs}
4. Calculate p -value based on H_a and the distribution of T
5. Conclude, in terms of “strength of evidence”

p -values: implementation

We may consider “evidence against H_0 ” based on p -values on the following scale:

p -value	Evidence against H_0
< 0.001	Extremely strong
$0.001 \leq p < 0.01$	Very strong
$0.01 \leq p < 0.05$	Strong
$0.05 \leq p < 0.10$	Modest
> 0.10	Not very strong

However, this is simply a guide, and these interpretations can vary by field.

p -values: implementation

z-test: If $T \sim \text{Normal}(0, 1)$:

- $H_a : \theta > \theta_0$

$$p\text{-value} = P_0 (T \geq T_{obs})$$

- $H_a : \theta < \theta_0$

$$p\text{-value} = P_0 (T \leq T_{obs})$$

- $H_a : \theta \neq \theta_0$

$$\begin{aligned} p\text{-value} &= P_0 (T \geq |T_{obs}|) + P_0 (T \leq -|T_{obs}|) \\ &= 2P_0 (T \geq |T_{obs}|) \end{aligned}$$

These p -values can be calculated using tables or software.

p -values: implementation

t -test: If $T \sim t_\nu$

- $H_a : \theta > \theta_0$

$$p\text{-value} = P_0 (T \geq T_{obs})$$

- $H_a : \theta < \theta_0$

$$p\text{-value} = P_0 (T \leq T_{obs})$$

- $H_a : \theta \neq \theta_0$

$$\begin{aligned} p\text{-value} &= P_0 (T \geq |T_{obs}|) + P_0 (T \leq -|T_{obs}|) \\ &= 2P_0 (T \geq |T_{obs}|) \end{aligned}$$

These can be calculated using tables or software.

p -values: Examples

Problem: Balls of yarn

A yarn manufacturer is concerned that a particular plant is producing balls that contain either too little or too much yarn; each ball should contain 100 grams.

The manufacturer randomly samples 45 balls, which have a sample mean weight of 99.2 grams, with a sample standard deviation of 1.7 grams.

State a set of appropriate hypotheses and calculate the corresponding p -value.

p -values: Examples

Solution: Balls of yarn

We are in Case 3. We are testing

$$H_0 : \mu = 100$$

$$H_a : \mu \neq 100$$

that is, we are carrying out a two-sided test. We have

- $n = 45$
- $\bar{x} = 99.2$
- $s = 1.7$
- $\mu_0 = 100$

so n is large.

p-values: Examples

Solution: Balls of yarn

Here

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \dot{\sim} \text{Normal}(0, 1).$$

and we have

$$T_{obs} = \frac{99.2 - 100.0}{1.7/\sqrt{45}} = -3.157.$$

Since T has a standard Normal distribution, the p -value is:

$$\begin{aligned} p\text{-value} &= 2 \times P_0(T \geq |T_{obs}|) \\ &= 2 \times P_0(T \geq |-3.157|) = 2 \times P_0(T \geq 3.157) \\ &= 0.002 \end{aligned}$$

p -values: Examples

```
n<-45
xbar<-99.2
mu0<-100
sv<-1.7
t.obs<-(xbar-mu0)/(sv/sqrt(n))
p.val<-2*pnorm(-abs(t.obs))
c(t.obs,p.val)

## [1] -3.156801851  0.001595097
```

p -values: Examples

Problem: Ice cream flavours

An ice cream company is creating a new ice cream flavour and wants to find out whether people prefer a base with chocolate or caramel swirls; they anticipate that caramel will be preferable

The company gives the ice cream to 45 taste testers, 21 of whom prefer caramel.

State a set of appropriate hypotheses and calculate the corresponding p -value.

p -values: Examples

Solution: Ice cream flavours

We are in Case 4. We consider testing

$$H_0 : p = 0.5$$

$$H_a : p > 0.5$$

that is, we are carrying out a one-sided test. We have

- $n = 45$
- $x = 21$
- $\hat{p} = 0.467$
- $p_0 = 0.50$

so n is large.

p -values: Examples

Solution: Ice cream flavours

Note that $\hat{p}$ is on the wrong side of p_0 to support H_a .

This means we need to take care to apply the formula

$$p\text{-value} = P_0 (T \geq T_{obs})$$

carefully. Under H_0 , T approximately has a standard Normal distribution.

p-values: Examples

Solution: Ice cream flavours

The test statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

Thus

$$T_{obs} = \frac{0.467 - 0.500}{\sqrt{\frac{0.500(1 - 0.500)}{45}}} = -0.447$$

p -values: Examples

Solution: Ice cream flavours

Since T has a standard Normal distribution assuming H_0 is true, the p -value is

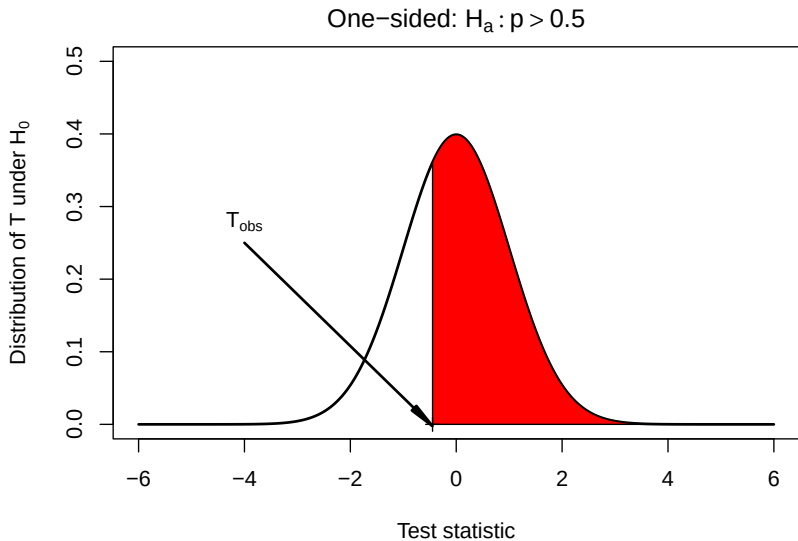
$$p\text{-value} = P_0(T \geq T_{obs}) = P_0(T \geq -0.447) = 0.673$$

p -values: Examples

```
n<-45
x<-21
phat<-x/n
p0<-0.5
se.hat<-sqrt(p0*(1-p0)/n)
t.obs<-(phat-p0)/se.hat
p.val<-1-pnorm(t.obs)
c(t.obs,p.val)

## [1] -0.4472136  0.6726396
```

p -values: Examples



Two sample comparisons

We may extend the ideas of

- confidence intervals
- hypothesis tests

to allow the comparison of **two** samples.

- $X_1, \dots, X_m$ are a random sample from one population
- $Y_1, \dots, Y_n$ are a random sample from a second population

Two sample comparisons

We can compute the sample mean and sample variance for each sample

$$\bar{X} = \frac{1}{m} \sum_{i=1}^m X_i \quad S_1^2 = \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2$$

and

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad S_2^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$$

Two sample comparisons: paired data

In the two sample case, we can also consider making two observations on each of n experimental units:

- X_i : pre-treatment, or Treatment 1
- Y_i : post-treatment, or Treatment 2

and then looking at the difference

$$D_i = X_i - Y_i$$

to assess whether there is any difference. We call this the case of **paired** data.

Two sample case: confidence intervals

To determine the correct strategy, we need to identify

(A) the assumed probability model/parameter of interest

- Normal, $\mu_1 - \mu_2$
- Binomial, $p_1 - p_2$

(B) the sample size assumption

- m, n small ($m, n < 30$)
- m, n large ($m, n \geq 30$)

(C) paired observations

- No
- Yes

Two sample case: confidence intervals

Theorem: 1. Parameter $\mu_1 - \mu_2$, unpaired, small m, n

Suppose that

- $X_1, \dots, X_m$ are a random sample from a Normal distribution with unknown μ_1 and σ_1
- $Y_1, \dots, Y_n$ are a random sample from a Normal distribution with unknown μ_2 and σ_2
- the X_i s are independent of the Y_i s
- $m < 30$ or $n < 30$
- $\sigma_1 = \sigma_2$

Two sample case: confidence intervals

Theorem: 1. Parameter $\mu_1 - \mu_2$, unpaired, small m, n

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R) , where

$$L = (\bar{X} - \bar{Y}) - t_{\alpha/2, m+n-2} S_P \sqrt{\frac{1}{m} + \frac{1}{n}}$$

$$R = (\bar{X} - \bar{Y}) + t_{\alpha/2, m+n-2} S_P \sqrt{\frac{1}{m} + \frac{1}{n}}$$

and

$$S_P^2 = \frac{(m-1)S_1^2 + (n-1)S_2^2}{m+n-2}$$

Two sample case: confidence intervals

Theorem: 2. Parameter $\mu_1 - \mu_2$, unpaired, large m, n

Suppose that

- $X_1, \dots, X_m$ are a random sample from a distribution with unknown μ_1 and σ_1
- $Y_1, \dots, Y_n$ are a random sample from a distribution with unknown μ_2 and σ_2
- the X_i s are independent of the Y_i s
- $m \geq 30$ and $n \geq 30$

Normality of the X s and Y s is not required.

Two sample case: confidence intervals

Theorem: 2. Parameter $\mu_1 - \mu_2$, unpaired, large m, n

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R) , where

$$L = (\bar{X} - \bar{Y}) - z_{\alpha/2} \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}$$

$$R = (\bar{X} - \bar{Y}) + z_{\alpha/2} \sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}$$

Two sample case: confidence intervals

Theorem: 3. Parameter $\mu_1 - \mu_2$, paired, small n

Suppose

- $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be **paired** observations from distributions with unknown means μ_1 and μ_2 , respectively,
- $n < 30$

Let

$$D_i = X_i - Y_i$$

and

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i \quad S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2.$$

Two sample case: confidence intervals

Theorem: 3. Parameter $\mu_1 - \mu_2$, paired, small n

Assume that the D_i are a random sample from a Normal distribution with unknown mean

$$\mu_d = \mu_1 - \mu_2$$

and standard deviation σ_d .

Two sample case: confidence intervals

Theorem: 3. Parameter $\mu_1 - \mu_2$, paired, small n

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R) , where

$$L = \bar{D} - t_{\alpha/2, n-1} \frac{S_d}{\sqrt{n}}$$

$$R = \bar{D} + t_{\alpha/2, n-1} \frac{S_d}{\sqrt{n}}$$

Two sample case: confidence intervals

Theorem: 4. Parameter $\mu_1 - \mu_2$, paired, large n

Suppose

- $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be **paired** observations from distributions with unknown means μ_1 and μ_2 , respectively,
- $D_i = X_i - Y_i$, $i = 1, \dots, n$ are a random sample from a distribution with unknown mean

$$\mu_d = \mu_1 - \mu_2$$

and standard deviation σ_d .

- $n \geq 30$

Normality is **not required**.

Two sample case: confidence intervals

Theorem: 4. Parameter $\mu_1 - \mu_2$, paired, large n

A $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is (L, R) , where:

$$L = \bar{D} - z_{\alpha/2} \frac{S_d}{\sqrt{n}}$$

$$R = \bar{D} + z_{\alpha/2} \frac{S_d}{\sqrt{n}}$$

Two sample case: confidence intervals

Theorem: 5. Parameter $p_1 - p_2$, large m, n

Suppose

- $X_1, \dots, X_m$ be a random sample of Bernoulli random variables with unknown probability of success p_1
- $Y_1, \dots, Y_n$ be a random sample of Bernoulli random variables with unknown probability of success p_2
- the X_i s are independent of the Y_i s,
- m and n are large enough

Two sample case: confidence intervals

Theorem: 5. Parameter $p_1 - p_2$, large m, n)

Let

- X be the number of “successes” in the sample $X_1, \dots, X_m$
- Y be the number of “successes” in the sample $Y_1, \dots, Y_n$

and

$$\hat{p}_1 = \frac{X}{m} \quad \text{and} \quad \hat{p}_2 = \frac{Y}{n}$$

Two sample case: confidence intervals

Theorem: 5. Parameter $p_1 - p_2$, large m, n

A $100(1 - \alpha)\%$ confidence interval for $p_1 - p_2$ is (L, R) , where

$$L = (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{m} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n}}$$

$$R = (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{m} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n}}$$

Examples

Problem: Painkiller duration

A drug manufacturer is testing the duration of pain-relief of their new over-the-counter painkiller, and wishes to quantify the difference in duration compared to ibuprofen.

Examples

Problem: Painkiller duration

The following data are available

- New drug
 - ▶ $m = 11$ observations
 - ▶ sample mean $\bar{x} = 5.4$ hours
 - ▶ sample standard deviation $s_1 = 0.9$ hours
- Ibuprofen
 - ▶ $n = 9$ observations
 - ▶ sample mean $\bar{y} = 4.8$ hours
 - ▶ sample standard deviation $s_2 = 1.1$ hours

Calculate a 90% confidence interval for the difference in mean durations between the two drugs.

State any assumptions you make.

Examples

Solution: Painkiller duration

We are in Scenario 1: $m = 11, n = 9$ and we are finding a 90%CI for $\mu_1 - \mu_2$, with $\alpha = 0.1$.

We must assume that the X s and Y s are Normally distributed and independent, and that the population variances are equal, i.e. $\sigma_1 = \sigma_2$. Then

$$\begin{aligned}s_p^2 &= \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} \\&= \frac{10 \times 0.9^2 + 8 \times 1.1^2}{11+9-2} \\&= 0.988\end{aligned}$$

so that $s_p = \sqrt{0.988} = 0.994$.

Examples

Solution: Painkiller duration

Let $\nu = n + m - 2$: we have that

$$(\bar{x} - \bar{y}) = 5.4 - 4.8 = 0.6,$$

and that

$$t_{\alpha, \nu} = t_{0.05, \nu} = 1.734$$

and

$$\sqrt{\frac{1}{m} + \frac{1}{n}} = \sqrt{\frac{1}{11} + \frac{1}{9}} = 0.449.$$

Examples

Solution: Painkiller duration

The 90% CI for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\bar{x} - \bar{y}) \pm t_{0.05, \nu} s_p \sqrt{\frac{1}{m} + \frac{1}{n}} \\= 0.6 \pm 1.734 \times 0.994 \times 0.449 \\= (-0.175, 1.375)\end{aligned}$$

Examples

Problem: Painkiller duration (large sample)

Now suppose that 55 patients are given the new drug and 62 are given ibuprofen.

- new drug:
 - ▶ the sample mean is $\bar{x} = 5.7$ hours
 - ▶ sample standard deviation is $s_1 = 1.1$ hours
- ibuprofen:
 - ▶ the sample mean is $\bar{y} = 4.6$ hours
 - ▶ the sample standard deviation is $s_2 = 0.4$ hours

Calculate a 95% confidence interval for the difference in mean durations between the two drugs

Examples

Solution: Painkiller duration (large sample)

We are in Scenario 2: $m = 55, n = 62$ and we are finding a 95% interval for $\mu_1 - \mu_2$.

Here

$$\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = \sqrt{\frac{1.1^2}{55} + \frac{0.4^2}{62}} = 0.157$$

Examples

Solution: Painkiller duration (large sample)

Let $\nu = n + m - 2$: we have that

$$(\bar{x} - \bar{y}) = 5.7 - 4.6 = 1.1,$$

and that

$$z_{\alpha/2} = z_{0.025, \nu} = 1.960$$

Examples

Solution: Painkiller duration (large sample)

The 95% CI for $\mu_1 - \mu_2$ is

$$\begin{aligned}(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} \\&= 1.1 \pm 1.960 \times 0.157 \\&= (0.793, 1.407)\end{aligned}$$

Examples

Problem: Painkiller duration (paired samples)

Now suppose that 15 patients are given ibuprofen, then given the new drug at a later date.

The sample mean difference in drug duration is $\bar{d} = 1.1$ hours, with sample standard deviation $s_d = 0.3$ hours.

Calculate a 99% confidence interval for the difference in mean durations between the two drugs.

Examples

Solution: Painkiller duration (paired samples)

We are in Scenario 3: $n = 15$ paired observations, and we are finding a 99%CI for $\mu_d = \mu_1 - \mu_2$, with $\alpha = 0.01$. We have that

$$t_{\alpha/2, \nu} = t_{0.005, 14} = 2.977.$$

Examples

Solution: Painkiller duration (paired samples)

Then

$$\begin{aligned} & \left(\bar{d} - t_{\alpha/2, \nu} \frac{s_d}{\sqrt{n}}, \bar{d} + t_{\alpha/2, \nu} \frac{s_d}{\sqrt{n}} \right) \\ &= \left(1.1 - 2.977 \frac{0.3}{\sqrt{15}}, 1.1 + 2.977 \frac{0.3}{\sqrt{15}} \right) \\ &= (0.869, 1.331) \end{aligned}$$

We must make the assumption that the differences in drug duration are Normally distributed.

Examples

Problem: Painkiller duration (paired samples, large n)

Suppose that 72 patients are given both drugs;

- the sample mean difference is $\bar{d} = 1.2$ hours,
- the standard deviation is $s_d = 0.4$ hours

Calculate a 95% confidence interval for the difference in mean durations between the two drugs

Examples

Solution: Painkiller duration (paired samples, large n)

We are in Scenario 4: $n = 72$ paired observations, and we are finding a 95%CI for $\mu_d = \mu_1 - \mu_2$, with $\alpha = 0.05$.

$$\begin{aligned} & \left(\bar{d} - z_{\alpha/2} \frac{s_d}{\sqrt{n}}, \bar{d} + z_{\alpha/2} \frac{s_d}{\sqrt{n}} \right) \\ &= \left(1.2 - 1.960 \frac{0.4}{\sqrt{72}}, 1.2 + 1.960 \frac{0.4}{\sqrt{72}} \right) \\ &= (1.108, 1.292) \end{aligned}$$

Examples

Problem: Student stress

Suppose that McGill students are being compared to Concordia students with respect to reporting stress.

- McGill
 - ▶ 112 students surveyed
 - ▶ 87 report high stress
- Concordia
 - ▶ 131 students surveyed
 - ▶ 92 report high stress

Calculate a 95% confidence interval for the difference in the proportions of students reporting high stress levels between the two schools

Examples

Solution: Student stress

We are in Scenario 5: Two Binomial setups, and we are finding a 95%CI for $p_1 - p_2$, with $\alpha = 0.05$

$$\hat{p}_1 = \frac{87}{112} = 0.777 \quad \hat{p}_2 = \frac{92}{131} = 0.702$$

and

$$\sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{m} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n}} = \sqrt{0.003} = 0.056$$

Examples

Solution: Student stress

The 95% CI is

$$\begin{aligned}(\hat{p}_1 - \hat{p}_2) \pm z_{0.025} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{m} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n}} \\= (0.777 - 0.702) \pm 1.960 \times 0.056 \\= (-0.035, 0.184)\end{aligned}$$

The limits of the CI can be negative as we are looking at a difference in probabilities.

Two sample hypothesis testing

Case: 1. Parameter $\mu_1 - \mu_2$, unpaired, small n, m

Suppose that

- $X_1, \dots, X_m$ are a random sample from a Normal distribution with unknown μ_1 and σ_1 .
- $Y_1, \dots, Y_n$ are a random sample from a Normal distribution with unknown μ_2 and σ_2 .

Two sample hypothesis testing

Case: 1. Parameter $\mu_1 - \mu_2$, unpaired, small n, m

To test

$$H_0 : \mu_1 = \mu_2$$

at level α , with

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

and if $\nu = m + n - 2$,

$$H_a : \mu_1 > \mu_2 \quad H_a : \mu_1 < \mu_2 \quad H_a : \mu_1 \neq \mu_2$$

$$\text{R. R.} \quad \{T \geq t_{\alpha, \nu}\} \quad \{T \leq -t_{\alpha, \nu}\} \quad \{|T| \geq t_{\alpha/2, \nu}\}$$

where S_p^2 is the pooled sample variance.

Two sample hypothesis testing

Case: 2. Parameter $\mu_1 - \mu_2$, unpaired, large n, m

Suppose that

- $X_1, \dots, X_m$ are a random sample from a distribution with unknown μ_1 and σ_1 .
- $Y_1, \dots, Y_n$ are a random sample from a distribution with unknown μ_2 and σ_2 .

Further, assume $m \geq 30$ and $n \geq 30$ and the X_i s are independent of the Y_i s.

Normality of the X s and Y s is not required.

Two sample hypothesis testing

Case: 2. Parameter $\mu_1 - \mu_2$, unpaired, large n, m

To test

$$H_0 : \mu_1 = \mu_2$$

at level α with

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \dot{\sim} \text{Normal}(0, 1)$$

$$H_a : \mu_1 > \mu_2 \quad H_a : \mu_1 < \mu_2 \quad H_a : \mu_1 \neq \mu_2$$

$$\text{R. R.} \quad \{T \geq z_\alpha\} \quad \{T \leq -z_\alpha\} \quad \{|T| \geq z_{\alpha/2}\}$$

Two sample hypothesis testing

Case: 3. Parameter $\mu_1 - \mu_2$, paired, small n

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be paired observations from distributions with unknown μ_1 and μ_2 , respectively, and $n < 30$.

Further, let $D_i = X_i - Y_i$ and assume that the D_i are a random sample from a Normal distribution with unknown $\mu_d = \mu_1 - \mu_2$ and σ_d

Two sample hypothesis testing

Case: 3. Parameter $\mu_1 - \mu_2$, paired, small n

To test

$$H_0 : \mu_1 = \mu_2$$

with

$$T = \frac{\bar{D}}{S_d/\sqrt{n}} \sim t_{n-1}$$

$$H_a : \mu_1 > \mu_2 \quad H_a : \mu_1 < \mu_2 \quad H_a : \mu_1 \neq \mu_2$$

$$\text{R. R.} \quad \{T \geq t_{\alpha, n-1}\} \quad \{T \leq -t_{\alpha, n-1}\} \quad \{|T| \geq t_{\alpha/2, n-1}\}$$

Two sample hypothesis testing

Case: 4. Parameter $\mu_1 - \mu_2$, paired, large n

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be paired observations from distributions with unknown μ_1 and μ_2 , respectively, and $n \geq 30$.

Further, let $D_i = X_i - Y_i$ and assume that the D_i are a random sample from a distribution with unknown $\mu_d = \mu_1 - \mu_2$ and σ_d .

Normality is not required.

Two sample hypothesis testing

Case: 4. Parameter $\mu_1 - \mu_2$, paired, large n

To test

$$H_0 : \mu_1 = \mu_2$$

with

$$T = \frac{\overline{D}}{S_d/\sqrt{n}} \dot{\sim} \text{Normal}(0, 1)$$

$$H_a : \mu_1 > \mu_2 \quad H_a : \mu_1 < \mu_2 \quad H_a : \mu_1 \neq \mu_2$$

$$\text{R. R.} \quad \{T \geq z_\alpha\} \quad \{T \leq -z_\alpha\} \quad \{|T| \geq z_{\alpha/2}\}$$

Two sample hypothesis testing

Case: 5. Parameter $p_1 - p_2$, m, n large

Suppose that

- $X_1, \dots, X_m$ is a random sample of Bernoulli random variables with unknown probability of success p_1
- $Y_1, \dots, Y_n$ is a random sample of Bernoulli random variables with unknown probability of success p_2

Suppose the X_i s are independent of the Y_i s, and that m and n are large enough. Let

- X be the number of successes in the sample $X_1, \dots, X_m$,
- Y be the number of successes in the sample $Y_1, \dots, Y_n$.

Two sample hypothesis testing

Case: 5. Parameter $p_1 - p_2$

To test

$$H_0 : p_1 = p_2$$

with

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{m} + \frac{1}{n} \right)}} \dot{\sim} \text{Normal}(0, 1)$$

where

$$\hat{p} = \frac{X + Y}{m + n}$$

$$H_a : p_1 > p_2 \quad H_a : p_1 < p_2 \quad H_a : p_1 \neq p_2$$

$$\text{R. R.} \quad \{T \geq z_\alpha\} \quad \{T \leq -z_\alpha\} \quad \{|T| \geq z_{\alpha/2}\}$$

Two sample hypothesis testing: Examples

Problem: Painkiller duration

A drug manufacturer is testing the duration of pain-relief of their new over-the-counter painkiller, and wishes to quantify the difference in duration compared to ibuprofen.

- New drug
 - ▶ $m = 11$ observations
 - ▶ sample mean $\bar{x} = 5.4$ hours
 - ▶ sample standard deviation $s_1 = 0.9$ hours
- Ibuprofen
 - ▶ $n = 9$ observations
 - ▶ sample mean $\bar{y} = 4.8$ hours
 - ▶ sample standard deviation $s_2 = 1.1$ hours

Two sample hypothesis testing: Examples

Problem: Painkiller duration

Conduct an appropriate test of hypotheses at the 1% level to determine whether the new drug lasts longer than ibuprofen. State any assumptions you make.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

We are in Case 1. We must assume that the observations X_i and Y_i are Normally distributed, that $\sigma_1 = \sigma_2$, and that the X s are independent of the Y s. We are testing

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

$$\begin{aligned} s_p^2 &= \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} = \frac{10 \times 0.9^2 + 8 \times 1.1^2}{11+9-2} \\ &= 0.988 \end{aligned}$$

so that $s_p = \sqrt{0.988} = 0.994$.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

Now

$$T_{obs} = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} = \frac{5.4 - 4.8}{0.994 \sqrt{\frac{1}{11} + \frac{1}{9}}} = 1.343.$$

so

$$\text{R. R. : } \{T \geq t_{\alpha, n+m-2} = t_{0.01, 18} = 2.552\}$$

Since

$$1.343 < 2.552$$

we do not reject H_0 at the 1% level.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

Assuming H_0 is true, T has a t -distribution with 18 degrees of freedom. We observed

$$T_{obs} = 1.343$$

Because the test is one-sided, the p -value is given by:

$$p\text{-value} = P_0(T \geq T_{obs}) = P_0(T \geq 1.343) = 0.09796$$

Two sample hypothesis testing: Examples

Problem: Painkiller duration

Now suppose that 55 patients are given the new drug and 62 are given ibuprofen.

- New drug
 - sample mean $\bar{x} = 5.7$
 - sample standard deviation $s_1 = 1.1$
- ibuprofen
 - sample mean $\bar{y} = 4.6$
 - sample standard deviation $s_2 = 0.4$

Conduct an appropriate test at the 5% level to determine whether the new drug lasts longer than ibuprofen.

Two sample hypothesis testing: Examples

Solution: Painkiller duration

We are in Case 2. We are testing

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

with $\alpha = 0.05$. We have

$$T_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{5.7 - 4.6}{\sqrt{\frac{1.1^2}{55} + \frac{0.4^2}{62}}} = 7.016.$$

$$\text{R. R.: } \{T > z_{0.05} = 1.645\}$$

Since $7.016 \geq 1.645$, we reject H_0 in favour of H_a at the 5% level

Two sample hypothesis testing: Examples

Solution: Painkiller duration

Assuming H_0 is true, T has a Normal $(0, 1)$ distribution approximately. We observed

$$T_{obs} = 7.016$$

Because the test is one-sided, the p -value is given by:

$$p\text{-value} = P_0(T \geq T_{obs}) = P_0(T \geq 7.016) = 0.00000$$

to five decimal places.

Two sample hypothesis testing: Examples

Problem: Painkiller duration (paired samples)

Now suppose that 15 patients are given ibuprofen, then given the new drug at a later date:

- the sample mean difference is $\bar{d} = 1.1$ hours,
- the standard deviation is $s_d = 0.3$ hours

Conduct an appropriate test of hypotheses at the 10% level to determine whether the new drug lasts longer than ibuprofen. State any assumptions you make

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples)

We are in Case 3. We must assume that the differences in drug duration are Normally distributed.

We are testing

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples)

We have

$$T_{obs} = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{1.1}{0.3/\sqrt{15}} = 14.201.$$

and

$$\text{R. R.: } \{T \geq t_{0.1,14} = 1.345\}$$

Since $14.201 > 1.345$ we reject H_0 in favour of H_a at the 10% level

Two sample hypothesis testing: Examples

Problem: Painkiller duration (paired samples, large n)

Suppose that 72 patients are given both drugs;

- the sample mean difference is $\bar{d} = 1.2$ hours,
- the standard deviation is $s_d = 0.4$ hours

Conduct an appropriate test of hypotheses at the 1% level to determine whether the new drug lasts longer than ibuprofen

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples, large n)

We are in Case 4. We must assume that the differences in drug duration are Normally distributed.

We are testing

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

Two sample hypothesis testing: Examples

Solution: Painkiller duration (paired samples, large n)

We have

$$T_{obs} = \frac{\bar{d}}{s_d/\sqrt{n}} = \frac{1.2}{0.4/\sqrt{72}} = 25.456$$

and

$$\text{R. R. : } \{T \geq z_{0.01} = 2.326\}$$

Since $25.456 > 2.326$ we reject H_0 in favour of H_a at the 1% level.

Two sample hypothesis testing: Examples

Problem: Student stress

Suppose that McGill students are being compared to Concordia students with respect to reporting stress.

- McGill
 - ▶ 112 students surveyed
 - ▶ 87 report high stress
- Concordia
 - ▶ 131 students surveyed
 - ▶ 92 report high stress

Conduct an appropriate test of hypotheses at the 10% level.

Two sample hypothesis testing: Examples

Solution: Student stress

We are in Case 5. We are testing

$$H_0 : p_1 = p_2$$

$$H_a : p_1 \neq p_2$$

at the $\alpha = 0.1$ level. We have

$$\hat{p}_1 = 87/112 = 0.777$$

$$\hat{p}_2 = 92/131 = 0.702$$

$$\hat{p} = (87 + 92)/(112 + 131) = 179/243 = 0.737.$$

Two sample hypothesis testing: Examples

Solution: Student stress

We have

$$\begin{aligned} T_{obs} &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{m} + \frac{1}{n} \right)}} \\ &= \frac{0.777 - 0.702}{\sqrt{0.737(1 - 0.737) \left(\frac{1}{112} + \frac{1}{131} \right)}} = 1.314 \end{aligned}$$

and

$$\text{R. R. } \{ |T| \geq z_{\alpha/2} = z_{0.05} = 1.645 \}$$

Since $|1.314| < 1.645$ we do not reject H_0 at the 10% level.

Two sample hypothesis testing: Examples

Problem: Jellyfish

Researchers are cataloguing two previously undiscovered species of deep-sea jellyfish, and are interested in assessing whether the lengths of the two species are different.

Two sample hypothesis testing: Examples

Problem: Jellyfish

- Species 1
 - ▶ sample size $m = 38$
 - ▶ sample mean length $\bar{x} = 21.4$ cm
 - ▶ sample standard deviation $s_1 = 3.9$ cm
- Species 2
 - ▶ sample size $n = 41$
 - ▶ sample mean $\bar{y} = 22.8$ cm
 - ▶ sample standard deviation $s_2 = 2.3$ cm

State a set of appropriate hypotheses, carry out the test at appropriate α levels, and calculate the corresponding p -value.

Two sample hypothesis testing: Examples

Solution: Jellyfish

We are in Case 2. We should test

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

We have

$$T_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{21.4 - 22.8}{\sqrt{\frac{3.9^2}{38} + \frac{2.3^2}{41}}} = -1.924.$$

and note that $m, n > 30$.

Two sample hypothesis testing: Examples

Solution: Jellyfish

For $\alpha = 0.05$, in this two-sided test

$$\text{R. R.: } \{|T| \geq z_{0.025} = 1.960\}$$

Since

$$|-1.924| < 1.960,$$

we do not reject H_0 in favour of H_a at the 5% level.

Two sample hypothesis testing: Examples

Solution: Jellyfish

For $\alpha = 0.1$, in this two-sided test

$$\text{R. R.: } \{|T| \geq z_{0.05} = 1.645\}$$

Since

$$|-1.924| < 1.645,$$

we do reject H_0 in favour of H_a at the 10% level.

Two sample hypothesis testing: Examples

Solution: Jellyfish

Assuming H_0 is true, T has a Normal $(0, 1)$ distribution approximately. We observed

$$T_{obs} = -1.924$$

Because the test is two-sided, the p -value is given by:

$$\begin{aligned} p\text{-value} &= 2 \times P_0(T \geq |T_{obs}|) \\ &= 2 \times P_0(T \geq |-1.924|) \\ &= 2 \times 0.02716 \\ &= 0.05431 \end{aligned}$$

rounding to five decimal places.