Statistics and Numerical Methods II

Shenal Dedduwakumara, John Maclean, Trent Mattner,
Jono Tuke
School of Mathematical Sciences

Semester 2, 2024

SNMII :: Probability

Pressure transient testing (PTT)

Example 1

To test pipes for potential leaking, we can send a pulse down the pipe by opening and closing valves, then look at the resultant signal.
Assume that we know that 1% of pipes in a region have potential leaks. Also we know that if the pipe is leaking we will get a positive result on the PTT test 98% for the time, while we will get a false positive 3% of the time.
You perform the test and get a positive result.
What is the probability that the pipe will potentially leak?
Will you perform an expensive CCTV test?

Goals

The goals of this lecture

- Recap the notations of probability.
- Learn how to convert a real-world problem into probability notation.
- Learn a problem-solving approach to solve probability problems.

Notation

Sample space set of all possible outcomes: S.

Event a subset of S.

Probability a function which assigns to each event a number between 0 and 1.

Axioms

Given a sample set $\mathcal{S}$, each event A_1 and A_2 have associated probabilities $P(A_1)$ and $P(A_2)$ such that the following hold;

A1 **Positivity:** For any event A, $P(A) \geq 0$.

A2 **Finitivity:** $P(S) = 1$.

A3 **Additivity:** For disjoint events, i.e. $P(A_1 \cap A_2) = \emptyset$,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2).$$

Equally likely events

If all the outcomes in the sample space are equally likely, then we can calculate the probability of an event as

$$P(A) = \frac{|A|}{|S|}$$

where $|\cdot|$ indicates the number of elements in the set.

NOT problems

To answer problems that involved asking the probability of an event not happening we use

$$P(A^c) = 1 - P(A),$$

where $P(A^c)$ is called the **complement** of A and indicates that A does not occur.

OR problems

To answer problems that ask for one event or another event or both, we can use the following rule:

Addition rule

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

where

- $A \cup B$ is the union of the two sets A and B and means OR, and
- $A \cap B$ is the interection of two sets and means AND.

GIVEN problems

To answer problems of the form: given that event A has happened, what is the probability that B occurs use one of the following:

Conditional

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

Baye's rule

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

AND problems

To answer problems of the form: what is the probability of both A and B occuring use one of the following:

Multiplication rule

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

Independence rule

Two events A and B are said to be **independent** if and only if

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

Solving probability problems

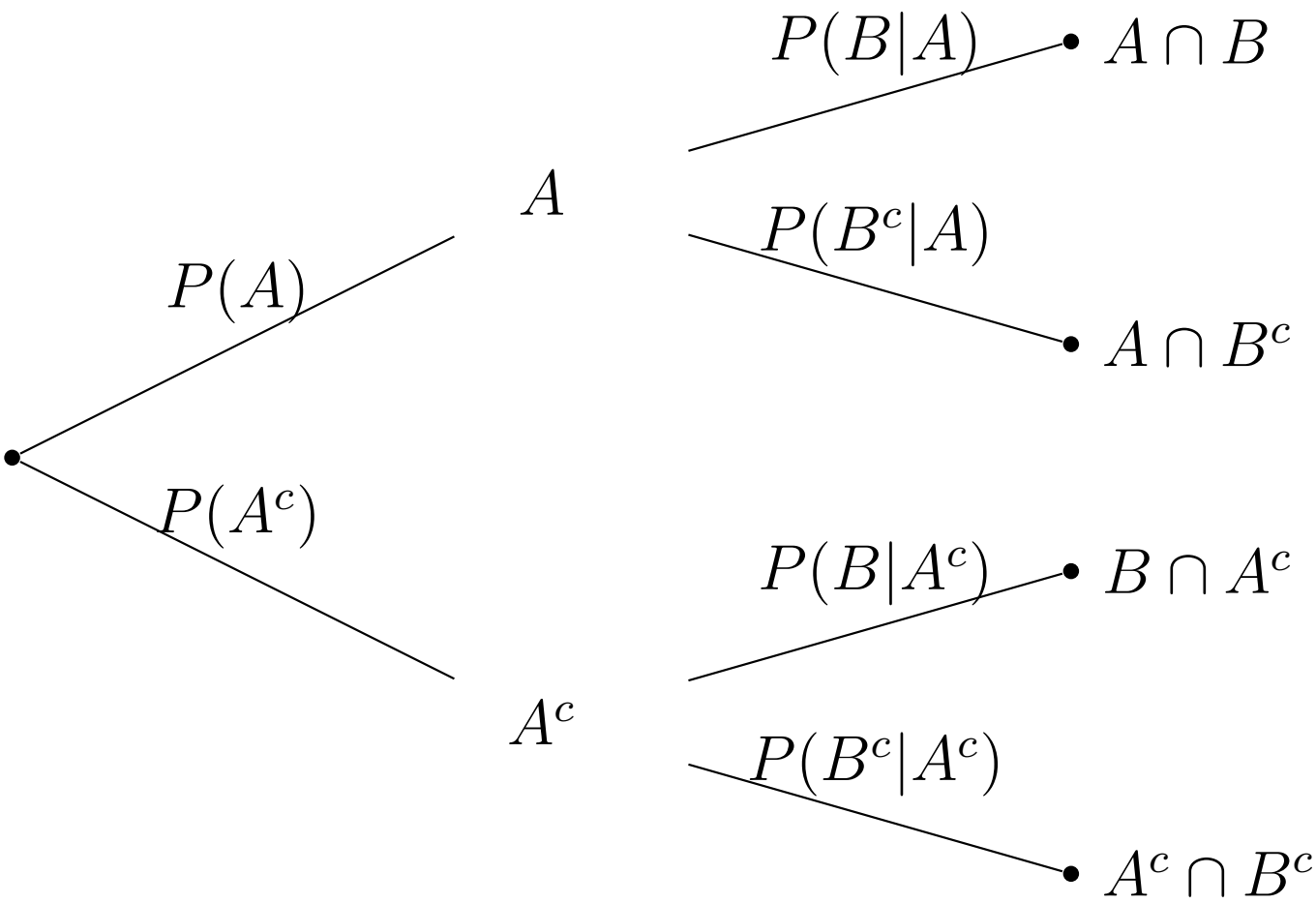
1. Identify the events and give them names.
2. Write down all given information as probability notation (stick with proportions not percents)
3. Write down what the question is asking as probability notation.
4. Use the rules given to write the question in terms of given information.
5. Calculate.

Pressure transient testing (PTT)

Example 2

To test pipes for potential leaking, we can send a pulse down the pipe by opening and closing valves, then look at the resultant signal.
Assume that we know that 1% of pipes in a region have potential leaks. Also we know that if the pipe is leaking we will get a positive result on the PTT test 98% for the time, while we will get a false positive 3% of the time.
You perform the test and get a positive result.
What is the probability that the pipe will potentially leak?
Will you perform an expensive CCTV test?

Tree diagram approach



Draw the tree for the PTT problem.

SNMII :: Random variables

Ship building plates specification

Example 3

Plates in a shipyard have a specified length of 6m and the tolerance is that cut plates should be within 2mm of this length.
Under the current process the lengths X are normally distributed with mean 6001mm and standard deviation 1.2mm.
Calculate the proportion within specification.

Goals

The goals of this lecture

- Recognise best random variable to use for a particular scenario.
- Identify the parameters of a random variable from the given information.
- Learn how to write the problem in terms of a probability calculation using random variables
- Know how to perform calculations in Matlab.

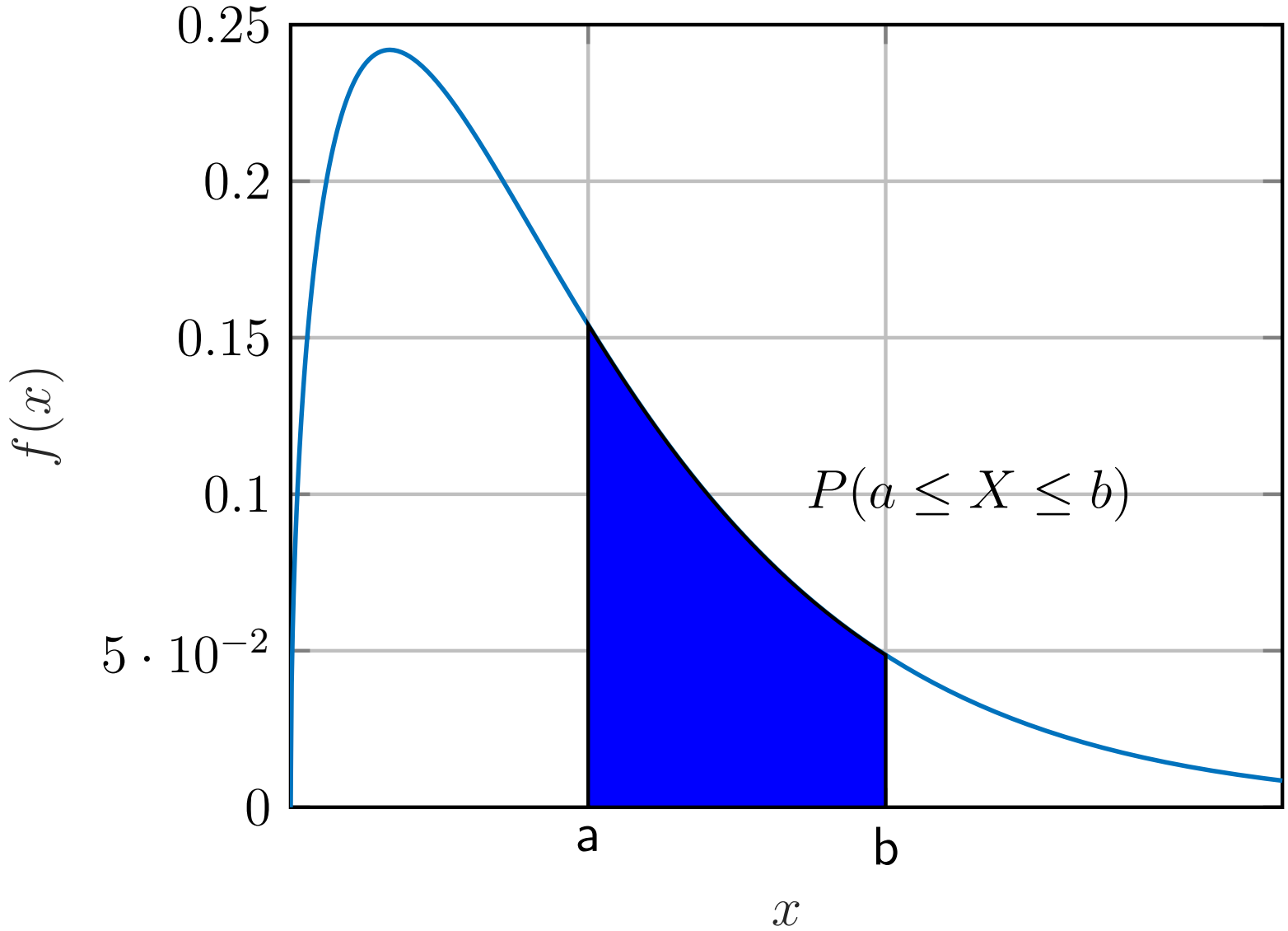
Notation

- Random variable: measurement determined by random outcome of an experiment.
- Discrete: variables that we count.
- Continuous: variables that we measure.

Calculating probabilities

- Probability mass function (PMF): a function for discrete RVs that gives the probability of being equal to a value, i.e., $P(X = a)$.
- Probability density function (PDF): a function for continuous RVs that is integrated to give the probability of a range of values $P(a \leq X \leq b)$.
- Cumulative distribution function (CDF): a function for both discrete and continuous RVs that gives the probability that the RV is less than or equal to a value $P(X \leq a)$.

Probability density function (PDF)



Recognising which RV

1. Is it discrete or continuous?
2. Is the random variable names in the description?
3. What parameters are given in the description?
4. Is it one of the common distributions?

Common discrete RVs

- Bernoulli: outcome success or failure.
- Binomial: counting the number of successes in independent trials.
- Geometric: counting the trails until a success.
- Negative binomial: counting the number of trials until k successes.
- Poisson: counting the number of incidents over a time period, or within an area.

Common continuous RVs

- Exponential measuring time between events.
- Normal length, weight, temperature.
- Gamma counting times between multiple events.

Parameters

Each distribution will have parameters that describe it. You will need this information to be able to perform the calculations.

For example the binomial distribution has two parameters:

- the number of trials n, and
- the probability of success p.

Matlab commands

Matlab has three types of commands of use to us:

- `namepdf(a, parameters)`: for calculating PMF $P(X = a)$
- `namecdf(a, parameters)`: for calculating CDF $P(X \leq a)$
- `nameinv(p, parameters)`: for calculating inverse, i.e., find a such that $P(X \leq a) = p$

Equal to problems

- Not applicable for continuous.
- Use the PMF.

Range problems

Convert to difference of two CDFs. Using following rules

Discrete

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a-1)$$
$$P(a \leq X < b) = P(X \leq b-1) - P(X \leq a-1)$$
$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$
$$P(a < X < b) = P(X \leq b-1) - P(X \leq a)$$

Continuous

$$\begin{aligned} P(a \leq X \leq b) &= P(a \leq X < b) \\ &= P(a < X \leq b) \\ &= P(a < X < b) = P(X \leq b) - P(X \leq a) \end{aligned}$$

At least problems

Often easier to convert to one minus cdf. Remember that

$$P(X \geq a) = 1 - P(X \leq a - 1)$$

Road accidents

Example 4

There were 43 road accidents involving pedestrians in the CBD of a city in the ten year period 1992-2001. New safety measures were introduced at the beginning of the year 2002, and there was only one such accident in the year 2002.

1. Calculate the probability of 0 or 1 accident in the year 2002 if the underlying rate is unchanged at 43 accidents per 10 years.
2. Do you think there is substantial evidence that the new safety measure has been successful?

Steel plates

Example 5

Plates in a shipyard have a specified length of 6m and the tolerance is that cut plates should be within 2mm of this length.
Under the current process the lengths X are normally distributed with mean 6001mm and standard deviation 1.2mm.
Calculate the proportion within specification.

SNMII :: Expection and variance of linear combinations

Combining measurements

Example 6

A marine survey vessel has two instruments that provide measurements X and Y of the depth of the sea bed.
The instruments have been carefully calibrated and give **unbiased** estimates of the depth, θ.
However the measurements are subject to error, and the first instrument is more precise with

$$\mathrm{var}(X) = 1\mathrm{m}$$
$$\mathrm{var}(Y) = 2\mathrm{m}$$

- Calculate the best weighted mean if the measurements are independent.
- Use matlab to find the best weighted mean if the measurements have a correlation of 0.5.

Expected values

The expected value of a random variable X, denoted $\mathrm{E}[X]$ is

$$\mathrm{E}[X] = \sum_{\text{all } x} xP(X = x) \qquad \text{discrete RV}$$

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} xf(x)dx \qquad \text{continuous RV}$$

Expected value of function

The expected value of a function of a random variable X, denoted $\mathrm{E}[g(X)]$ is

$$\mathrm{E}[g(X)] = \sum_{\text{all } x} g(x)P(X = x) \qquad \text{discrete RV}$$

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx \qquad \text{continuous RV}$$

Variance

The variance of a random variable X denoted $\mathrm{var}(X)$ is

$$\mathrm{var}(X) = \mathrm{E}[(X - \mathrm{E}[X])^2]$$

It can also be written as

$$\mathrm{var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2$$

Covariance

The covariance between two random variables X and Y, denoted $\mathrm{cov}(X, Y)$, is

$$\mathrm{cov}(X, Y) = \mathrm{E}[(X - E[X])(Y - E[Y])]$$

It can also be written as

$$\mathrm{cov}(X, Y) = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$$

Also is X and Y are independent then

$$\mathrm{cov}(X, Y) = 0.$$

Independent RV

The two random variables, X and Y are said to be **independent** if for all real numbers x and y,

$$P(\{X \leq x\} \cap \{Y \leq y\}) = P(X \leq x)P(Y \leq y).$$

Correlation

The correlation between two random variables X and Y, denoted $\mathrm{cor}(x, y)$

$$\mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{var}(X)}\sqrt{\mathrm{var}(Y)}}$$

Note that

$$-1 \leq \mathrm{cor}(X, Y) \leq 1.$$

Linear combination of single RV

$$\mathrm{E}[a + bX] = a + b\mathrm{E}[X]$$
$$\mathrm{var}(a + bX) = b^2\mathrm{var}(X)$$

Linear combinations of two RV

$$\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$$
$$\mathrm{var}[aX + bY] = a^2\mathrm{var}[X] + b^2\mathrm{var}[Y] + 2ab \times \mathrm{cov}(X, Y)$$

Also

$$\mathrm{cov}(aX + bY, cW + dZ) = ac \times \mathrm{cov}(X, W) + ad \times \mathrm{cov}(X, Z) + bc \times \mathrm{cov}(Y, W) + bd \times \mathrm{cov}(Y, Z)$$

Answering linear combination questions

General approach

- Identify all the random variable and label if not already.
- Write down all expections, variances given.
- Are the random variables independent? If not what is the covariance of correlation?
- Identify what measure we need to calculate.
- Use the rules to calculate.

Solar car

Example 7

Let $D_i, i = 1, 2$ be the distance travelled by a solar car, during a solar challenge event, in an eight hour day. It is known that D_i has a mean of 720km and a standard deviation of 75km. Let W be the total distance travelled in two days. Calculate the mean and standard deviation of W if

- the days are independent, and
- the correlation between the days is 0.8.

Combining measurements

Example 8

A marine survey vessel has two instruments that provide measurements X and Y of the depth of the sea bed.
The instruments have been carefully calibrated and give **unbiased** estimates of the depth: θ.
However the measurements are subject to error, and the first instrument is more precise with

$$\mathrm{var}(X) = 1\mathrm{m}$$
$$\mathrm{var}(Y) = 2\mathrm{m}$$

- Calculate the best weighted mean if the measurements are independent.
- Use Matlab to find the best weighted mean if the measurements have a correlation of 0.5.

SNMII :: Sampling distributions

EPA testing

Example 9 (The EPA)

The U.S. Environmental Protection Agency sets a Secondary Standard for iron in drinking water as less than 0.3 milligrams per liter (mg/l). A water company will be fined, if the mean of hydrant tests at four locations exceeds 0.5 mg/l.
Calculate the probability that the company will be fined if the population mean is 0.3 mg/l and the standard deviation is

- 0.15mg/l, or
- 0.3mg/l.

Populations and samples

- Population: the set of all individuals of interest.
- Sample: a subset of the population.
- Parameter: a numeric characteristic of the population.
- Statistic: a numeric characteristic of the sample, used to infer a parameter.

Sample mean and variance

For the n observations from the random variable X:

$$x_1, x_2, \ldots, x_n$$

- Sample mean

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

- Sample variance

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

- Sample standard deviation

$$s_x = \sqrt{s_x^2}$$

Sample covariance and correlation

For observations

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

from random variables X and Y.

- Sample covariance:

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

- Sample correlations:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Matlab commands

```
% Simulate some data
X = normrnd(0, 1, 10, 1);
Y = normrnd(0, 1, 10, 1);
Z = X + Y;
% Mean
mean(X)
% Variance
var(X)
% Standard deviation
std(X)
% Covariance
cov(X, Y)
% Correlation
corrcoef(X, Y)
```

Properties of sums of random variables

Let

$$Y = \sum_{i=1}^{n} a_i X_i,$$

then

$$\mathrm{E}[Y] = \sum_{i=1}^{n} a_i \mathrm{E}[X_i]$$

and

$$\mathrm{var}(Y) = \sum_{i=1}^{n} a_i^2 \mathrm{var}(X_i) + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} a_i a_j \mathrm{cov}(X_i, X_j)$$

Sampling distributions

Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with

$$X_i \sim N(\mu, \sigma^2)$$

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

The Central Limit Theorem

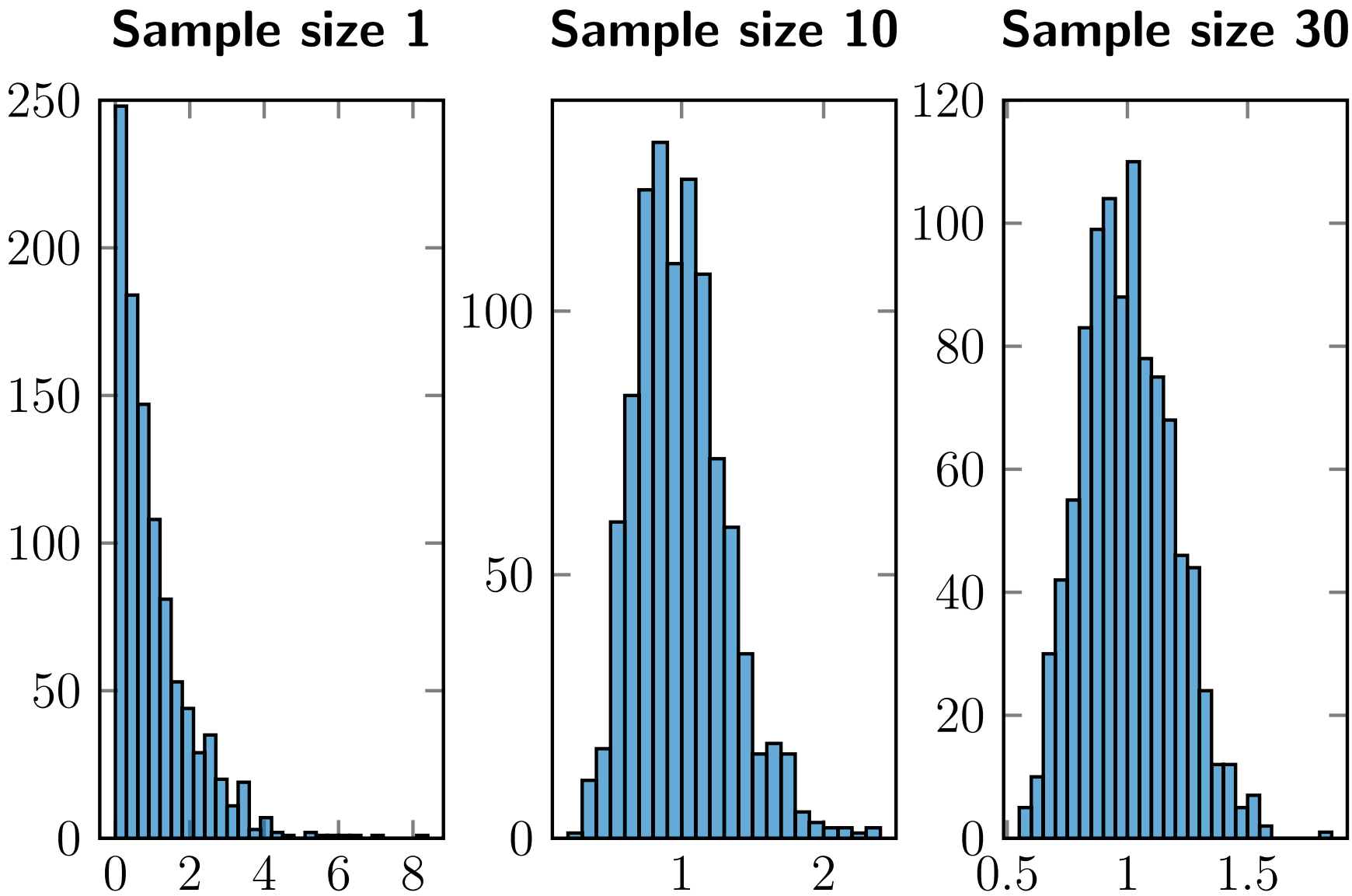
Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with

$$E(X_i) = \mu \text{ and } \operatorname{var}(X_i) = \sigma^2 \quad \text{for } \; i = 1, \ldots, n,$$

Then for large n,

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

CLT example


Sample size 1
Sample size 10
Sample size 30
250
200
150
100
50
0
0
2
4
6
8
100
50
0
1
2
120
100
80
60
40
20
0
0.5
1
1.5

Oil exploration

Example 10 (Oil exploration)

An oil exploration company owns a small helicopter to fly personnel out to remote sites. There are 7 passenger seats but the payload must be less than 650kg. The distribution of the masses of passengers has mean 81kg and a standard deviation of 12kg. The distribution of the masses of equipment carried by one of the passengers has a mean of 20kg and a standard deviation of 16kg. What is the probability that the total mass of 7 randomly selected passengers and the equipment will exceed 650 kg?

Lift capacity

Example 11 (Lift Capacity)

A lift in a building is rated to carry 20 persons or a maximum load of $2,000$ kg. If the mean mass of 20 persons exceeds 100 kg the lift will be overloaded and trip a warning buzzer.
The people using the lift have masses distributed with a mean of 75 kg and a standard deviation of 24 kg.
What is the probability that the mean weight of 20 people exceeds the maximum load?

EPA testing

Example 12 (The EPA)

The U.S. Environmental Protection Agency sets a Secondary Standard for iron in drinking water as less than 0.3 milligrams per liter (mg/l). A water company will be fined, if the mean of hydrant tests at four locations exceeds 0.5 mg/l.
Calculate the probability that the company will be fined if the population mean is 0.3 mg/l and the standard deviation is

- 0.15mg/l, or
- 0.3mg/l.

SNMII :: Confidence intervals

Fireman example

Example 13 (Fireman's clothing)

A researcher is evaluating a new high performance fabric for firemen's protective clothing. One test is resistance to wear in the wet state, made using a Martindale wear tester. Four test pieces of the wet fabric are clamped at 4 test positions on the Martindale tester and abraded in a figure of eight motion for $1,000$ cycles. The weight loss (%) is measured for each test piece. The results of the test for the new fabric are: $11.3\%, 13.2\%, 10.5\%$ and 14.6%.

- Calculate the 95% confidence interval for the mean percentage loss.
- If the required spec is less that 10% wear, do you buy from this supplier?

Estimation of parameters

Consider n independent random variables

$$X_1, X_2, \ldots, X_n$$

such that

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, 2, \ldots, n.$$

The sample mean

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

is a point estimate of μ.

What about an interval estimate of μ?

Confidence interval

A random interval (L, U) is a $100(1-\alpha)\%$ confidence interval for a parameter θ if

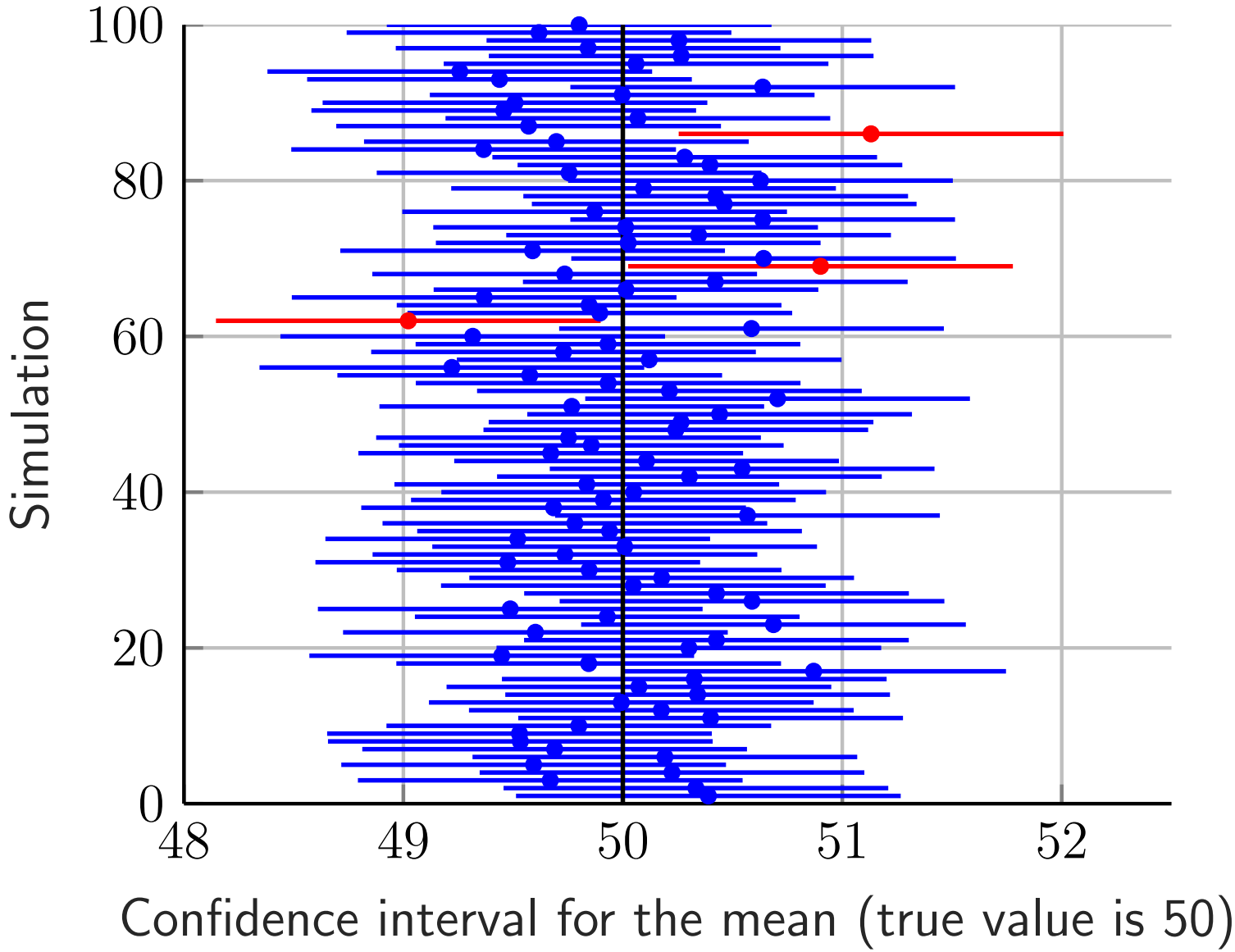
$$P(L \leq \theta \leq U) = 1 - \alpha.$$

The 95% Confidence Interval for μ of normal distribution with σ^2 known

A 95% confidence interval for μ is the interval

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Confidence


100
80
60
40
20
0
Simulation
48
49
50
51
52
Confidence interval for the mean (true value is 50)

General form

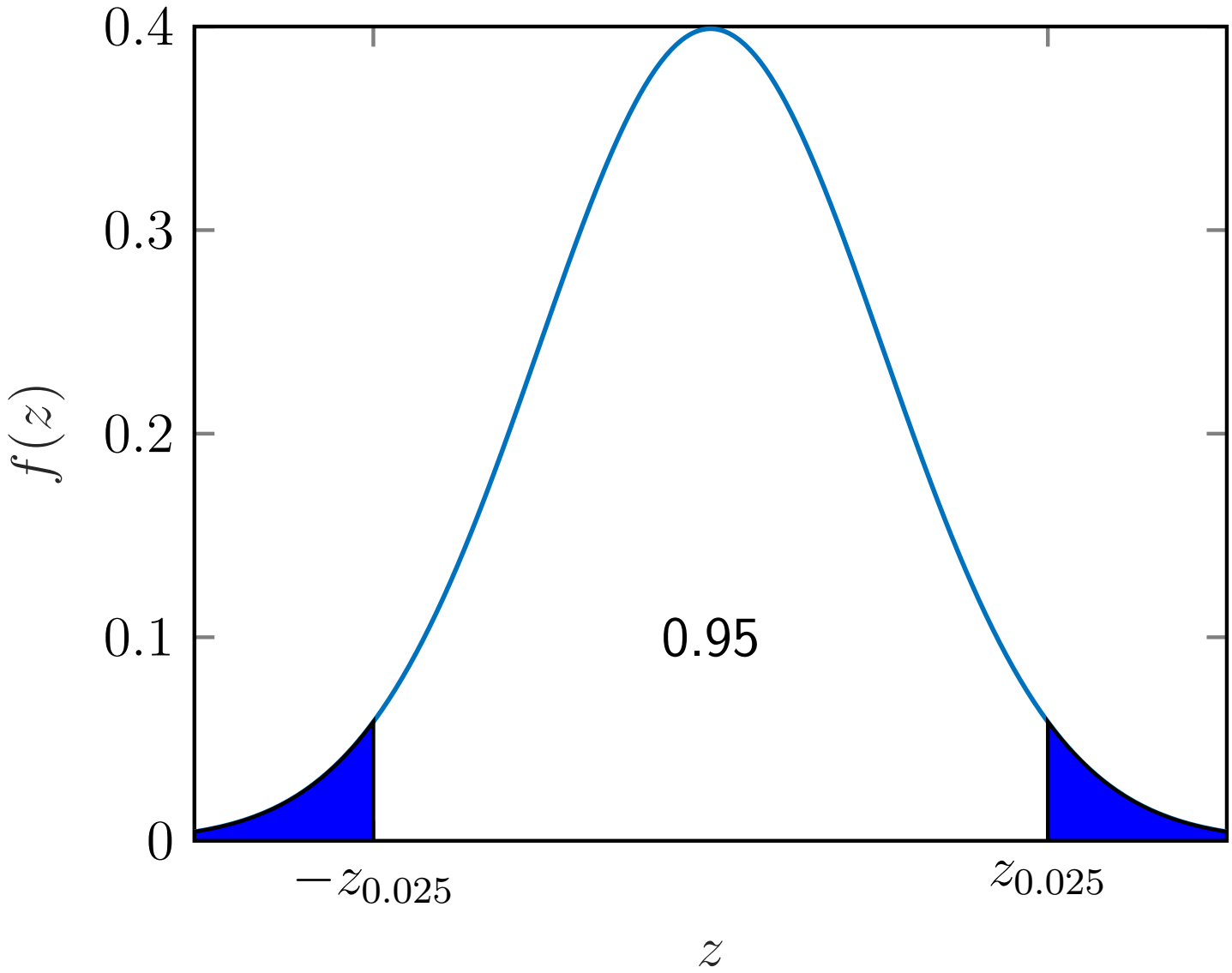
Confidence intervals are often of the form

$$\text{estimate} \pm \text{cutoff} \times \text{standard error}$$

In the case so far we have

- estimate: $\bar{x}$
- cutoff: 1.96
- standard error: $\sigma/\sqrt{n}$

Cutoff



The 95% Confidence Interval for μ of normal distribution with σ^2 not known

A 95% confidence interval for μ is the interval

$$\left(\bar{x} - t_{n-1,0.025}\frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,0.025}\frac{s}{\sqrt{n}}\right)$$

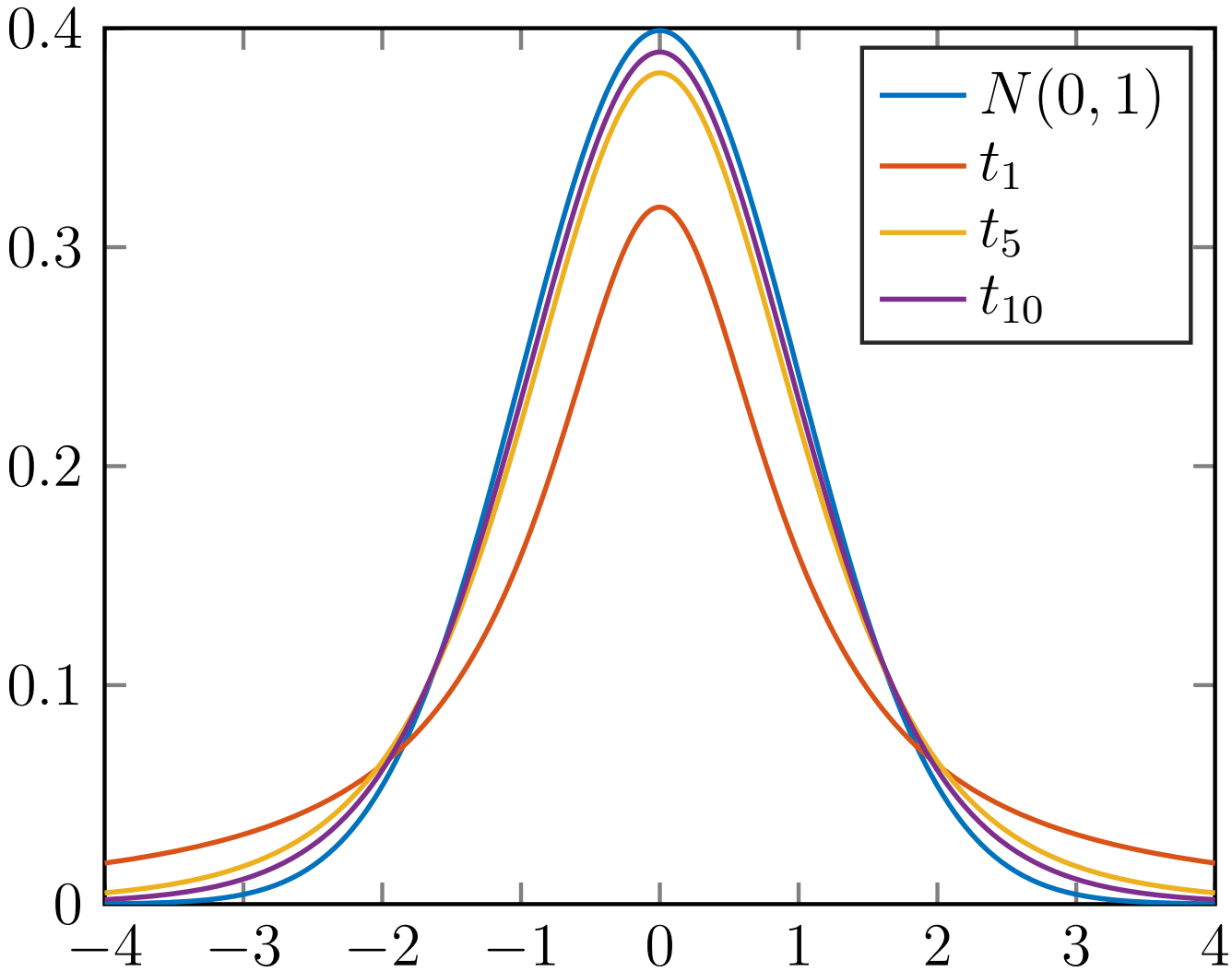
where

$$T \sim t_{n-1}$$

and

$$P(T > t_{n-1,0.025}) = 0.025,$$

T-distribution



Fireman example

Example 14 (Fireman's clothing)

A researcher is evaluating a new high performance fabric for firemen's protective clothing. One test is resistance to wear in the wet state, made using a Martindale wear tester. Four test pieces of the wet fabric are clamped at 4 test positions on the Martindale tester and abraded in a figure of eight motion for $1,000$ cycles. The weight loss (%) is measured for each test piece. The results of the test for the new fabric are: $11.3\%, 13.2\%, 10.5\%$ and 14.6%.

- Calculate the 95% confidence interval for the mean percentage loss.
- If the required spec is less that 10% wear, do you buy from this supplier?

SNMII :: Hypothesis testing

Fluoridated water supply

Example 15 (Fluoridated water supply)

In some cities the public water supply is fluoridated as a public dental health measure. This practice remains controversial and it is important to maintain the agreed target level and to ensure that it is not substantially exceeded. A city sets a target level of 0.8 ppm, and every week a public health inspector takes a random sample of 12 bottles from kitchen taps and sends them for fluoride analysis. The standard deviation of fluoride contents of bottles filled from kitchen taps has been estimated over several years, and is 0.04. Last week the analyses were:

$$0.78, 0.91, 0.91, 0.98, 0.83, 0.85, 0.87, 0.85, 0.95, 0.84, 0.90, 0.92.$$

Framework

1. Identify the parameter of interest.
2. Identify the null and alternative hypotheses.
3. Choose the level of significance.
4. Identify and check the assumptions using the data.
5. Perform the test
6. Make a decision

Model

Let $X_1, X_2, \ldots, X_n$ be n identically independently distributed (i.i.d.) random variables from

$$N(\mu, \sigma^2)$$

Null and alternative hypotheses

Two-sided

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

One-sided

$$H_0 : \mu \leq \mu_0$$
$$H_a : \mu > \mu_0$$

Significance level

The significance level is the probability that we will reject the null hypothesis when it is in fact true.

Default level is 0.05.

Test statistic

We use the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

P-value

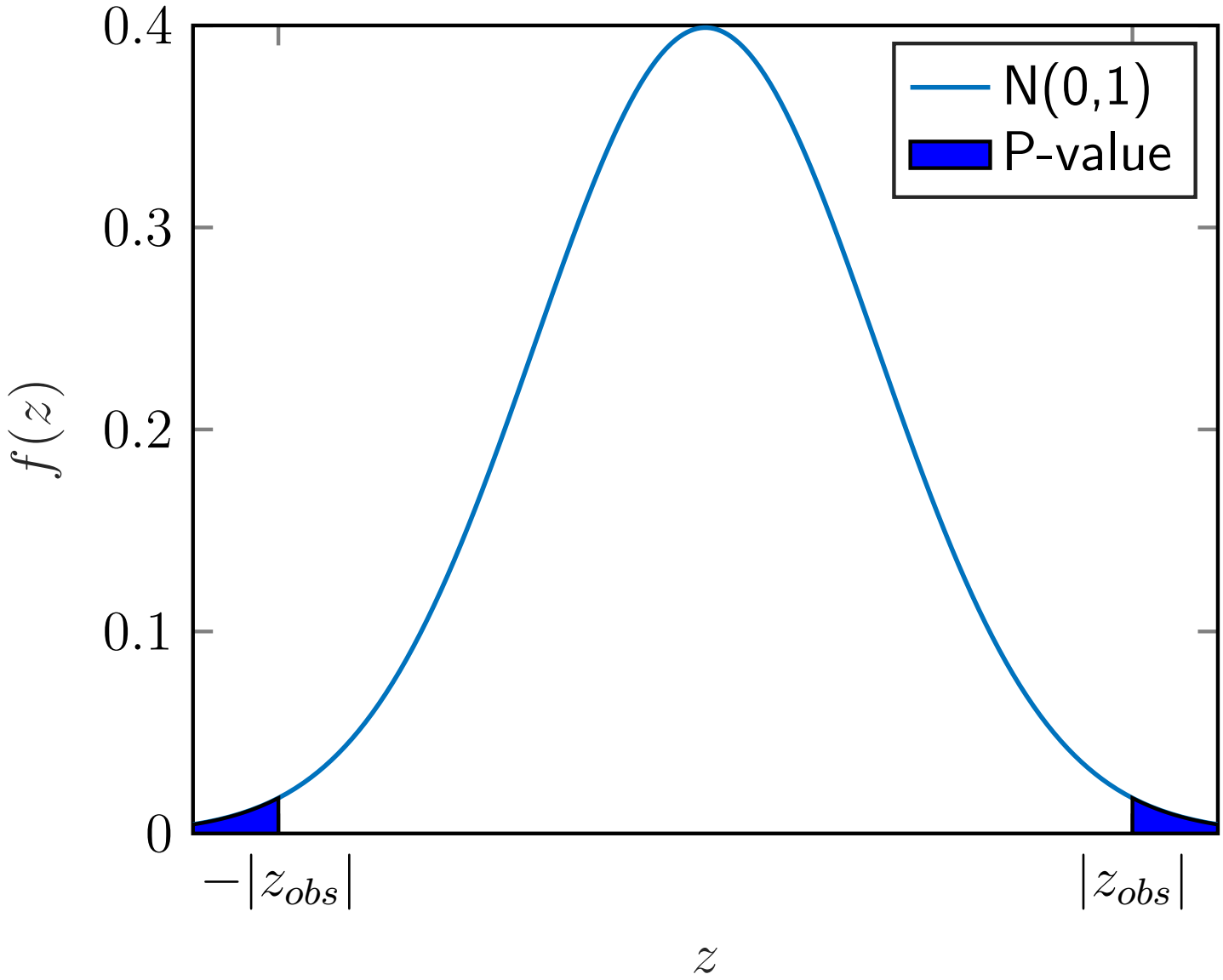
The P-value is the probability of observing a value at extreme or more extreme than the observed test statistic given the null hypothesis is true.

$$P - value = P(|Z| > |z_{obs}|),$$

where

$$Z \sim N(0, 1)$$

P-value



Assumptions

Let $X_1, X_2, \ldots, X_n$ be n identically independently distributed (i.i.d.) random variables from

$$N(\mu, \sigma^2)$$

Independence

To check that the observations are independent, you need to look at the experimental design.

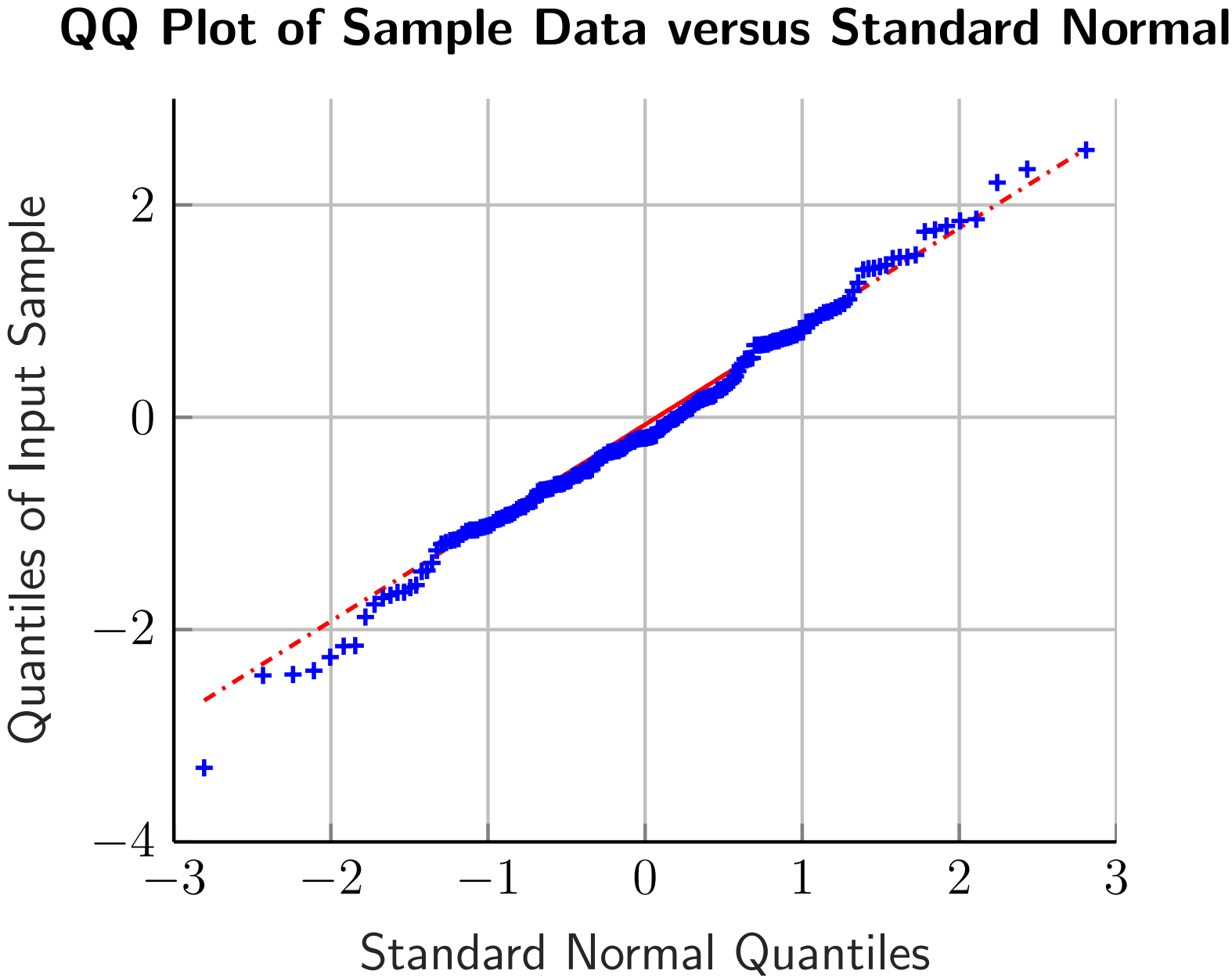
Indications that the observations are independent are words like “randomly selected” or “random sample”.

Normality

To check normality, we look at a normal QQ-plot.

If the data is normally distribured, then we expect the points to lie roughly on a straight line.

Normal QQ-plot



Perform test

```
[h,p,ci,zval] = ztest(data, mu0, sigma)
```

- `h` is decision at 5% significance level: 1 is reject, 0 is retain.
- `p` is P-value.
- `ci` is 95% confidence interval for mean.
- `zval` is observed value of test statistic z_{obs}.

Make decision

- P-value method: if the P-value $<$ significance level: reject
- Cutoff method: if $|z_{obs}| > z_{\alpha/2}$: reject
- CI method: if $\mu_0 \notin (l, u)$: reject

Fluoridated water supply

Example 16 (Fluoridated water supply)

In some cities the public water supply is fluoridated as a public dental health measure. This practice remains controversial and it is important to maintain the agreed target level and to ensure that it is not substantially exceeded. A city sets a target level of 0.8 ppm, and every week a public health inspector takes a random sample of 12 bottles from kitchen taps and sends them for fluoride analysis. The standard deviation of fluoride contents of bottles filled from kitchen taps has been estimated over several years, and is 0.04. Last week the analyses were:

$$0.78, 0.91, 0.91, 0.98, 0.83, 0.85, 0.87, 0.85, 0.95, 0.84, 0.90, 0.92.$$

SNMII :: One-sample T-test

Inductors

Example 17 (Inductors)

An inductor is manufactured to a specified inductance of 470 microhenrys. A customer tests a sample of 20 inductors and gets the following data.

471.29	462.85	469.29	469.52	476.19
458.91	470.80	483.05	478.16	465.08
464.59	463.43	452.00	471.38	454.16
485.88	455.33	466.14	455.90	475.93

Is the data consistent with the specs?

Framework

1. Identify the parameter of interest.
2. Identify the null and alternative hypotheses.
3. Choose the level of significance.
4. Identify and check the assumptions using the data.
5. Perform the test
6. Make a decision

Model

Let $X_1, X_2, \ldots, X_n$ be n identically independently distributed (i.i.d.) random variables from

$$N(\mu, \sigma^2)$$

Test-statistic

We use the test statistic

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

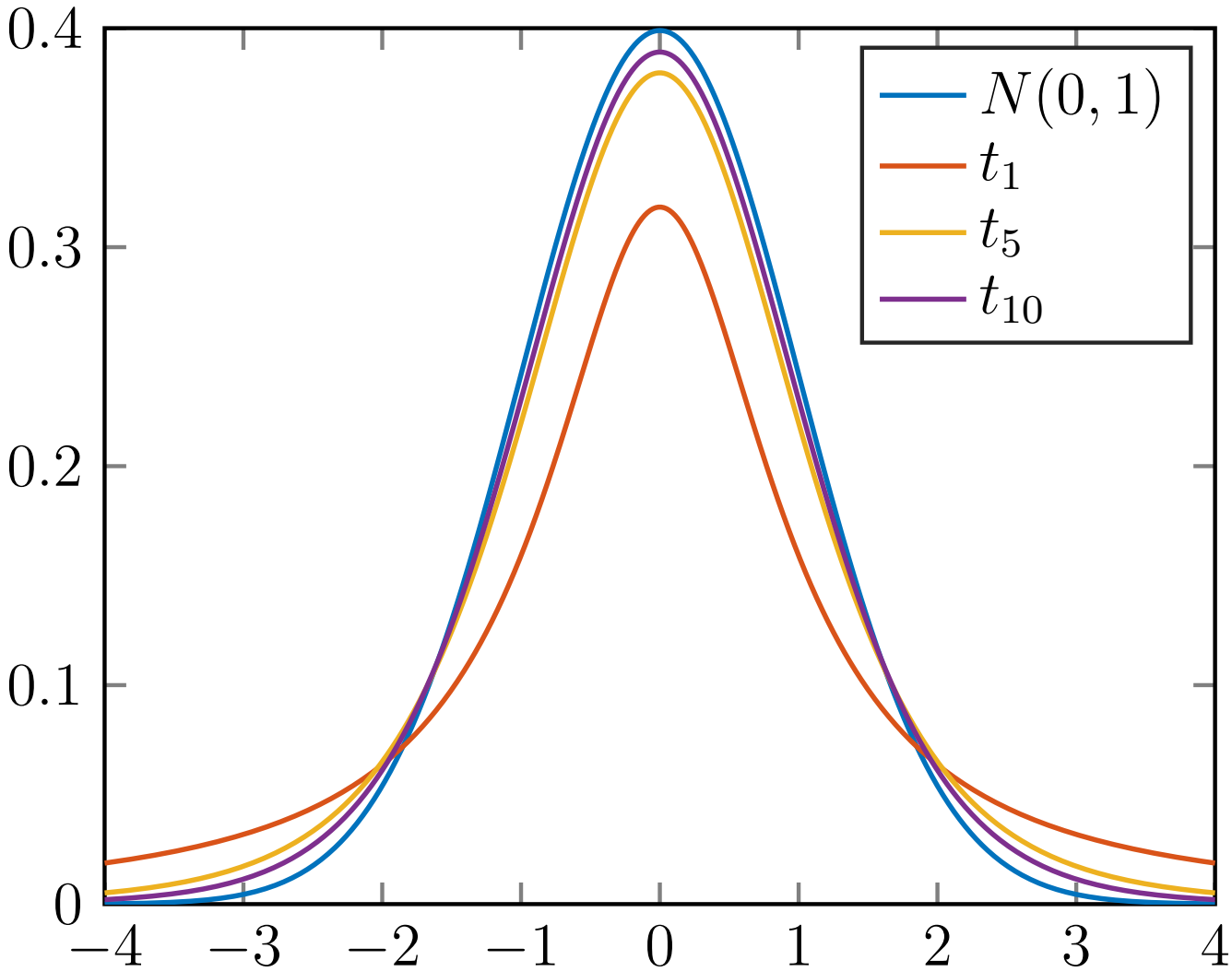
where

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

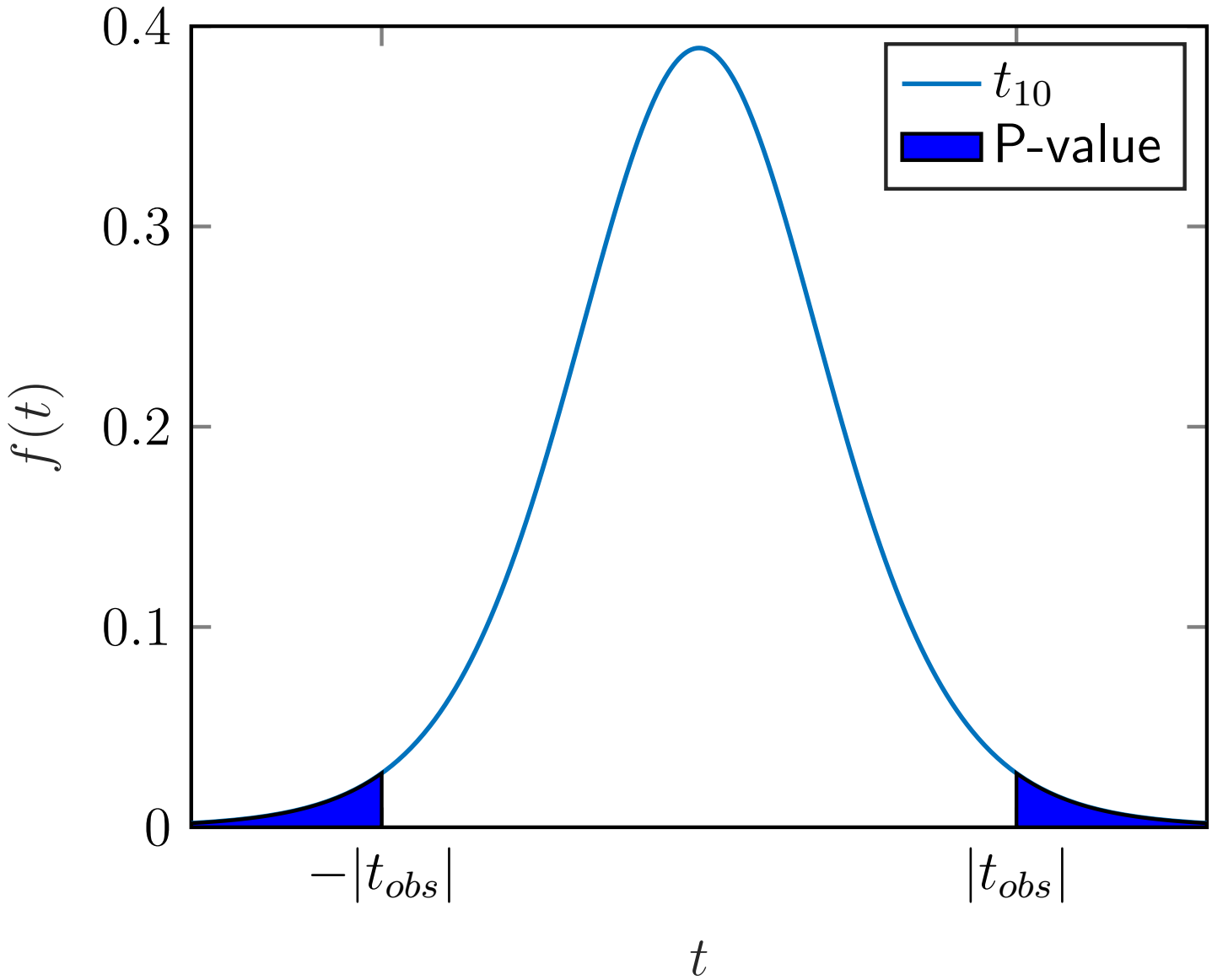
and

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

T-distribution



P-value



Assumptions

Let $X_1, X_2, \ldots, X_n$ be n identically independently distributed (i.i.d.) random variables from

$$N(\mu, \sigma^2)$$

Confidence interval

The $100(1-\alpha)\%$ confidence interval for the population mean of normally distributed data with unknown variance is

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{S}{\sqrt{n}}$$

Matlab

```
1  [h,p,ci,stats] = ttest(data, mu0)
```

- `h` is decision at 5% significance level: 1 is reject, 0 is retain.
- `p` is P-value.
- `ci` is 95% confidence interval for mean.
- `stats` is structure with three parts
  - `tstat` observed value of t_{obs}.
  - `df` degrees of freedom.
  - `sd` standard deviation of the data.

Inductors

Example 18 (Inductors)

An inductor is manufactured to a specified inductance of 470 microhenrys. A customer tests a sample of 20 inductors and gets the following data.

471.29	462.85	469.29	469.52	476.19
458.91	470.80	483.05	478.16	465.08
464.59	463.43	452.00	471.38	454.16
485.88	455.33	466.14	455.90	475.93

Is the data consistent with the specs?

SNMII :: Two-sample T-test and Matched-Pairs T-test

Example

Example 19 (Oxy-cutting)

A company specializes in steel fabrication and uses oxy-propane gas cutting to cut steel plates. An engineer wants to investigate the use of oxy-natural gas as a more convenient alternative. An undesirable side-effect of any gas cutting is the hardening of the steel near the cut edge. The engineer will not consider natural gas instead of propane if the hardening side-effect is increased, and decides to perform an experiment to make a comparison. The engineer finds 8 plates of different grade steels and of different thicknesses. To remove the variability between plates from the comparison, the engineer decides to make two cuts on each plate, one with oxy-propane and the other with oxy-natural gas. The variable to be analyzed is derived from Vickers hardness (VH10) measurements made in a fixed pattern alongside the cut edge.

Model

Let

$$X_{11}, X_{12}, \ldots, X_{1n_1}$$

be the observations from treatment 1, and

$$X_{21}, X_{22}, \ldots, X_{2n_2}$$

be the observations from treatment 2.

We assume that the X_{ij} are independently identically distributed from

$$N(\mu_i, \sigma_i^2)$$

and that $X_{1j}, j = 1, \ldots n_1$ and $X_{2j}, j = 1, \ldots, n_2$ are independent of each other.

Null and alternative hypotheses

$$H_0 : \mu_1 - \mu_2 = 0,$$
$$H_a : \mu_1 - \mu_2 \neq 0,$$

where

μ_1 is the population mean of the first population, and μ_2 is the population mean of the second population.

Test statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

where

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

and

$$S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

T-distribution

For the two-sample t-test, a T-distribution can be used as an approximate reference distribution.

The degrees of freedom are obtained in Matlab by

$$\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2 \Bigg/ \left(\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}\right)$$

Assumptions

Let

$$X_{11}, X_{12}, \ldots, X_{1n_1}$$

be the observations from treatment 1, and

$$X_{21}, X_{22}, \ldots, X_{2n_2}$$

be the observations from treatment 2.

We assume that the X_{ij} are independently identically distributed from

$$N(\mu_i, \sigma_i^2)$$

and that $X_{1j}, j = 1, \ldots n_1$ and $X_{2j}, j = 1, \ldots, n_2$ are independent of each other.

Matlab

```
1 [h,p,ci,stats]=ttest2(data1,data2, ...
2                       'VarType','unequal')
```

- `h` is decision at 5% significance level: 1 is reject, 0 is retain.
- `p` is P-value.
- `ci` is 95% confidence interval for mean.
- `stats` is structure with three parts
  - `tstat` observed value of t_{obs}.
  - `df` degrees of freedom.
  - `sd` standard deviation of each group.

Matched-pairs data

Suppose two variables X_1 and X_2 are recorded on each of n items.

The data can be represented as

$$(x_{11}, x_{12}), (x_{21}, x_{22}), \ldots, (x_{n1}, x_{n2})$$

The data cannot be treated as two independent samples because observations made on the same item could be expected to be correlated.

- This violates the assumption that observations from different groups are independent.

The strategy for paired data is to **calculate the differences**

$$d_i = x_{i1} - x_{i2}$$

and apply the one-sample t-procedures to those data.

Matlab

```
1  [h,p,ci,stats] = ttest(var1, var2)
```

- `h` is decision at 5% significance level: 1 is reject, 0 is retain.
- `p` is P-value.
- `ci` is 95% confidence interval for mean.
- `stats` is structure with three parts
  - `tstat` observed value of t_{obs}.
  - `df` degrees of freedom.
  - `sd` standard deviation of each group.

Example

Example 20 (Oxy-cutting)

A company specializes in steel fabrication and uses oxy-propane gas cutting to cut steel plates. An engineer wants to investigate the use of oxy-natural gas as a more convenient alternative. An undesirable side-effect of any gas cutting is the hardening of the steel near the cut edge. The engineer will not consider natural gas instead of propane if the hardening side-effect is increased, and decides to perform an experiment to make a comparison. The engineer finds 8 plates of different grade steels and of different thicknesses. To remove the variability between plates from the comparison, the engineer decides to make two cuts on each plate, one with oxy-propane and the other with oxy-natural gas. The variable to be analyzed is derived from Vickers hardness (VH10) measurements made in a fixed pattern alongside the cut edge.

SNMII :: Bootstrapping

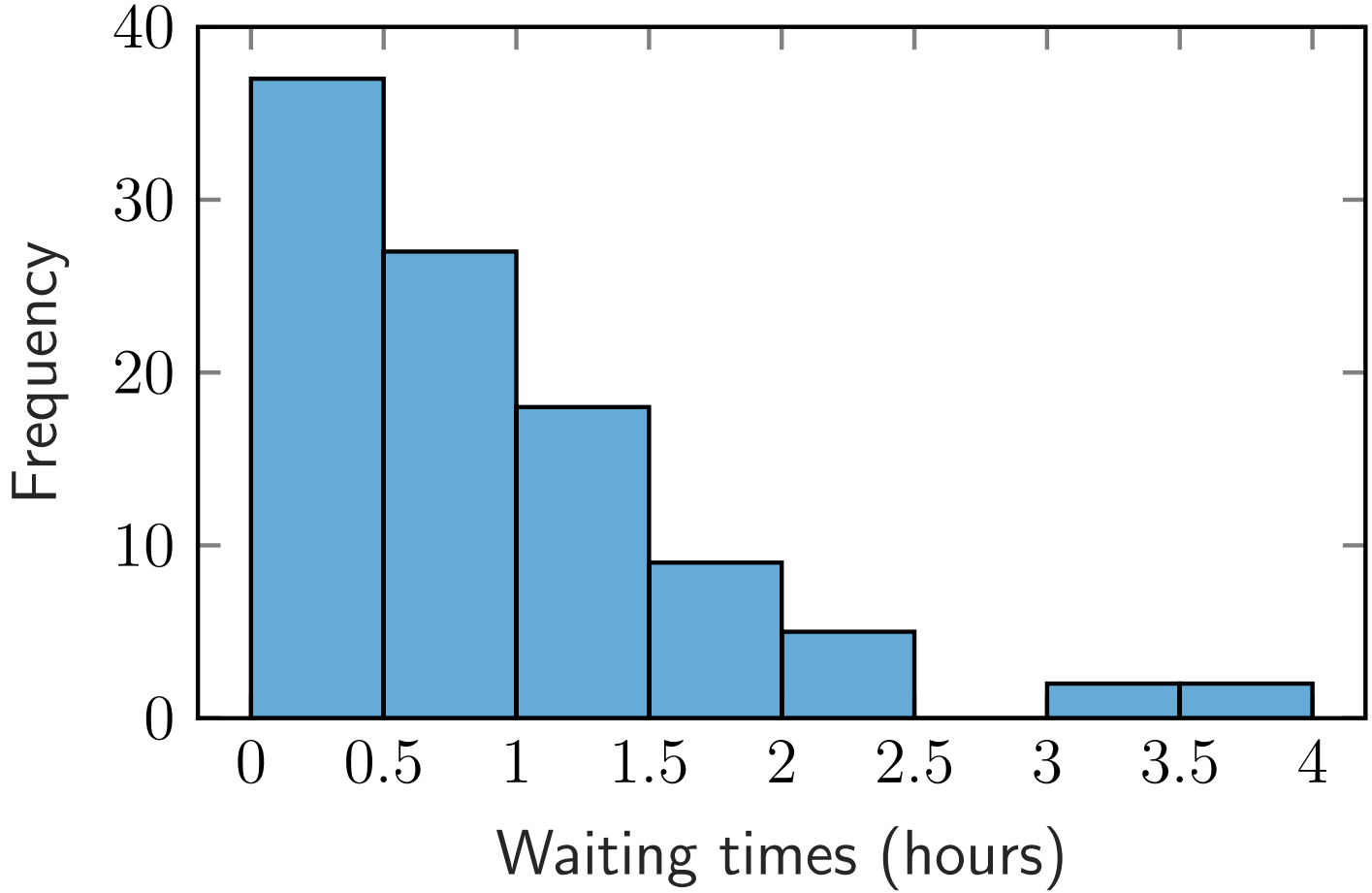
Solar energy

Example 21 (Solar energy)

You are given the daily total solar energy data for Adelaide for January 1st 2019 - 13th September 2019. The units are MJm^2.

- Calculate a 95% confidence interval for the median daily solar energy.
- Calculate a 95% confidence interval for the min daily solar energy.

Motivating example



Calculate a 95% confidence interval for the median.

Setup

Suppose we have a simple random sample

$$\boldsymbol{S} = \{X_1, X_2, \ldots, X_n\},$$

and we are interested in a statistic $T = t(\boldsymbol{S})$ which is an estimator of the population parameter θ.

We can estimate the sampling distribution of T empirically using bootstrapping.

Procedure

- Sample B times with replacement a sample of size n from $\boldsymbol{S}$, we denote the bth sample as

$$\boldsymbol{S}_b^* = \{X_{b1}^*, X_{b2}^*, \ldots, X_{bn}^*\}$$

- For each sample, compute the value of the statistic,

$$T_b^* = t(\boldsymbol{S}_b^*)$$

Key idea

The population is to the sample, as the sample is to the bootstrap sample.

i.e. the distribution of T_b^* around the original estimate T is analogous to the sampling distribution of the estimator T around the population parameter θ.

Summary statistics

Mean

$$\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B} T_b^*$$

Variance

$$\hat{\text{var}}(T^*) = \frac{\sum_{b=1}^{B}(T_b^* - \bar{T}^*)^2}{B-1}$$

Bias

$$\hat{b}_T(\theta) = \bar{T}^* - T$$

Confidence intervals

Bootstrap percentile

$$(T^*_{[(B+1)\alpha/2]}, T^*_{[(B+1)(1-\alpha/2)]})$$

Normal-theory interval

$$(T - \hat{b}_T(\theta) - z_{\alpha/2}\hat{SE}(T^*), T - \hat{b}_T(\theta) + z_{\alpha/2}\hat{SE}(T^*))$$

Solar energy

Example 22 (Solar energy)

You are given the daily total solar energy data for Adelaide for January 1st 2019 - 13th September 2019. The units are MJm^2.

- Calculate a 95% confidence interval for the median daily solar energy.
- Calculate a 95% confidence interval for the min daily solar energy.

SNMII :: One-way ANOVA

Filter membanes

Example 23 (Filter membranes)

The production engineer of a company which manufactures filters for liquids, for use in the pharmaceutical and food industries, wishes to compare the burst strength of four types of membrane.

- A company's own standard membrane material.
- B new material the company has developed.
- C membrane material from other manufacturer.
- D membrane material from other manufacturer.

The engineer has tested five filter cartridges from ten different batches of each material. The mean burst strengths for each set of five cartridges are given in KPa.

Model

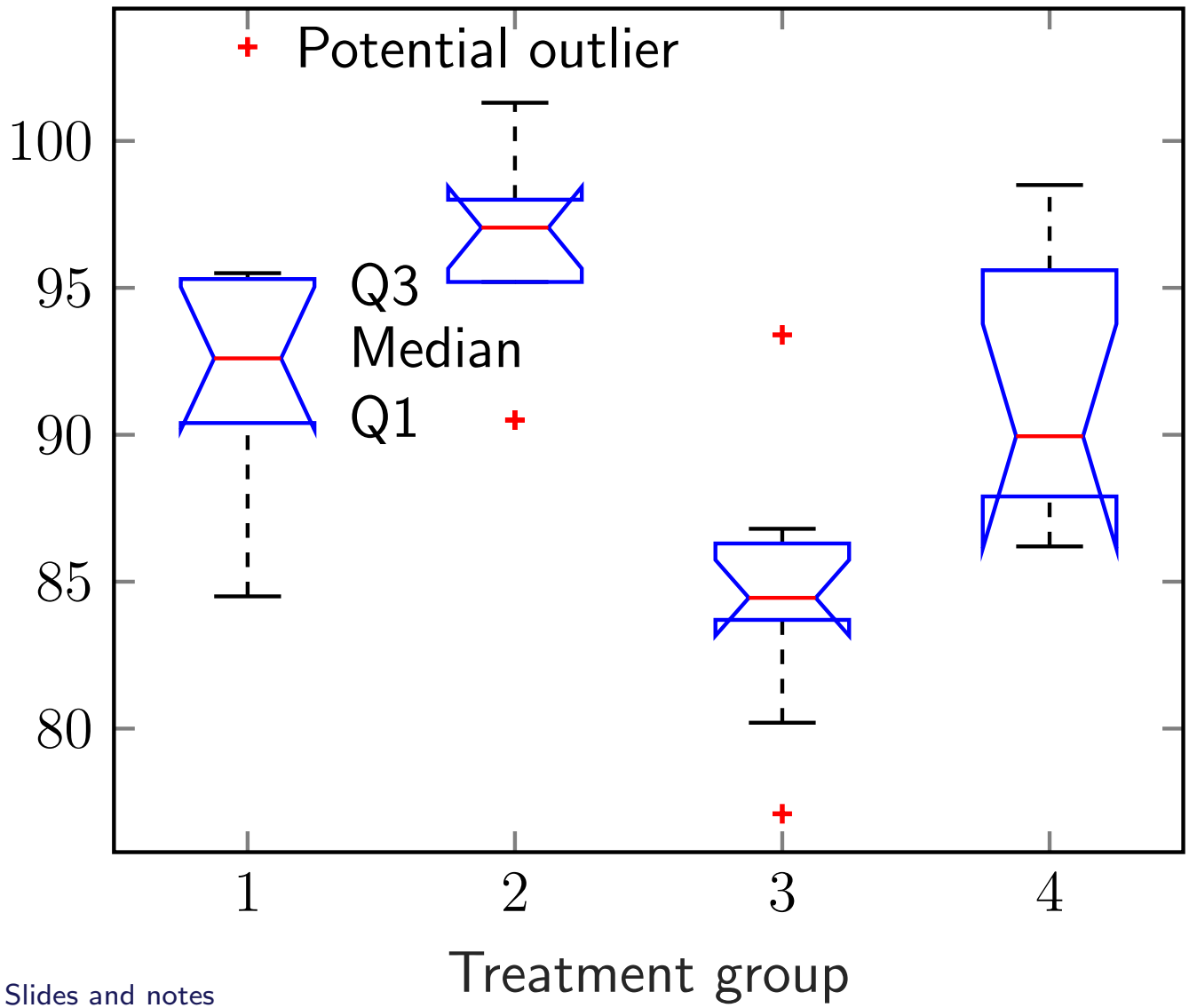
$$Y_{ij} = \alpha_i + \epsilon_{ij}, i = 1, \ldots, k, j = 1, 2, \ldots, n_i,$$

- Y_{ij} is the jth observation in group i.
- α_i is the population mean for group i.
- e_{ij} random error:

$$e_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

Side-by-side boxplots

The side-by-side boxplot gives a boxplot for each treatment group.



Null and alternative hypotheses

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k$$
$$H_a : \text{not all of the } \alpha\text{s are the same,}$$

where α_i is the population mean of observations in population i.

ANOVA table

Source	SS	df	MS	F	P-value
Group (Between)	SSR	k - 1	MSR = SSR/(k-1)	MSR/MSE	$P(F > f_{obs})$
Error (Within)	SSE	N - k	MSE = SSE / (N - k)		
Total	SST	N - 1			

- k is number of groups.
- N is total number of observations.
- $F \sim F_{k-1,N-k}$

Multicomp - Bonferonni

Reject

$$H_0 : \alpha_i = \alpha_j$$

if

$$\frac{|y_{i\bullet} - y_{j\bullet}|}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} > t_{N-k,\alpha^*}$$

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$

Matlab

Put data into matrix

$$data = \begin{bmatrix} Y_{11} & Y_{21} & \dots & Y_{k1} \\ Y_{12} & Y_{22} & \dots & Y_{k2} \\ \vdots & \ddots & & \vdots \\ Y_{1n_1} & Y_{2n_2} & \dots & Y_{kn_k} \end{bmatrix}$$

Assumes that each group same size

$$n_1 = n_2 = \dots n_k$$

Matlab

```
[p table stats] = anova1(data)
```

- p: P-value
- table ANOVA table
- stats
  - gnames: group names
  - n: group size
  - source: matlab call
  - means: group sample means
  - df: $N - k$
  - s: MSE

Assumptions

$$Y_{ij} = \alpha_i + \epsilon_{ij}, i = 1, \ldots, k, j = 1, 2, \ldots, n_i,$$

- Y_{ij} is the jth observation in group i.
- α_i is the population mean for group i.
- e_{ij} random error:

$$e_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

Assumptions

- Normality: Normal QQ-plots for each group
- Constant variance:

$$\frac{s_{max}}{s_{min}} < 2$$

- Independence: expermental design