Introduction to Probabilities and Statistics

Arnaud Legrand

Performance Evaluation Lecture
UFRGS, Porto Alegre, August 2015

Outline

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)

  You can think of it as a small box:
  - Every time you open the box, you get a different value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)
  You can think of it as a small box:
  - Every time you open the box, you get a different value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies
- Formally a probability space is defined by $(\Omega, \mathcal{F}, \mathrm{P})$ where:
  - Ω, the sample space, is the set of all possible outcomes
    - E.g., all the possible combinations of your DNA with the one of your {girl|boy}friend
    - You may or may not be able to observe directly the outcome.

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)
  You can think of it as a small box:
  - Every time you open the box, you get a different value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies
- Formally a probability space is defined by $(\Omega, \mathcal{F}, \mathrm{P})$ where:
  - Ω, the sample space, is the set of all possible outcomes
    - E.g., all the possible combinations of your DNA with the one of your {girl|boy}friend
    - You may or may not be able to observe directly the outcome.
  - $\mathcal{F}$ if the set of events where an event is a set containing zero or more outcomes
    - E.g., the event of "the DNA corresponds to a girl with blue eyes"
    - An event is somehow more tangible and can generally be observed

Probabilities

- Using probabilities enables to model uncertainty that may result of incomplete information or imprecise measurements

  A random variable (or stochastic variable) is, roughly speaking, a variable whose value results from a measurement (or an observation)
  You can think of it as a small box:
  - Every time you open the box, you get a different value.
  - I will use this box analogy throughout the whole lecture and I encourage you to ask yourself what the box can be in your own studies
- Formally a probability space is defined by $(\Omega, \mathcal{F}, \mathrm{P})$ where:
  - Ω, the sample space, is the set of all possible outcomes
    - E.g., all the possible combinations of your DNA with the one of your {girl|boy}friend
    - You may or may not be able to observe directly the outcome.
  - $\mathcal{F}$ if the set of events where an event is a set containing zero or more outcomes
    - E.g., the event of "the DNA corresponds to a girl with blue eyes"
    - An event is somehow more tangible and can generally be observed
  - The probability measure $\mathrm{P} : \mathcal{F} \to [0, 1]$ is a function returning an event's probability (P("having a brown-eyed baby girl") $= 0.0005$)

Continuous random variable

- A random variable associates a numerical value to outcomes

$$X : \Omega \to \mathbb{R}$$

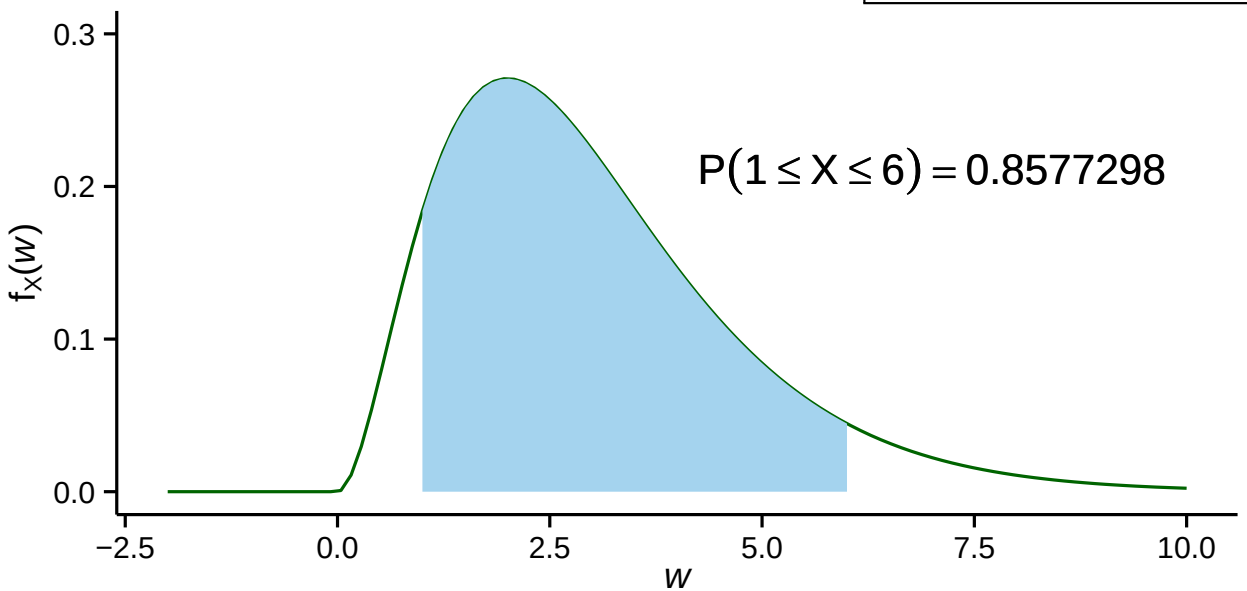
  - E.g., the weight of the baby at birth (assuming it solely depends on DNA, which is quite false but it's for the sake of the example)
  - Since many computer science experiments are based on time measurements, we focus on continuous variables
- **Note:** To distinguish random variables, which are complex objects, from other mathematical objects, they will always be written in blue capital letters in this set of slides (e.g., X)
- The probability measure on Ω induces probabilities on the values of X
  - $\mathrm{P}(X = 0.5213)$ is generally 0 as the outcome never exactly matches
  - $\mathrm{P}(0.5213 \leq X \leq 0.5214)$ may however be non-zero

Probability distribution

A probability distribution (a.k.a. probability density function or p.d.f.) is used to describe the probabilities of different values occurring

- A random variable X has density f_X, where f_X is a non-negative and integrable function, if:

$$\mathrm{P}[a \leq X \leq b] = \int_a^b f_X(w)\, \mathrm{d}w$$



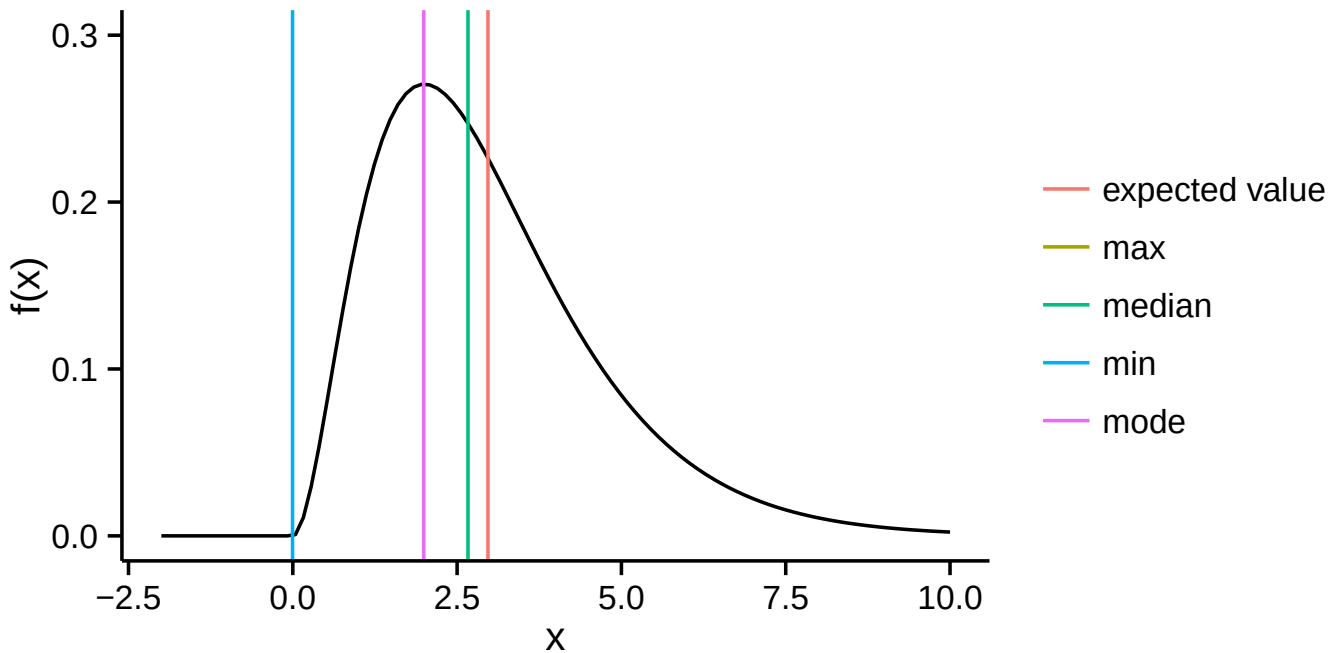
Note: the X in $1 \leq X \leq 6$ *should be in blue...*

- **Note**: people often confuse the sample space with the random variable. Try to make the difference when modeling your system, it will help you

Characterizing a random variable

The probability density function fully characterizes the random variable but it is also complex object

- It may be symmetrical or not
- It may have one or several modes
- It may have a bounded support or not, hence the random variable may have a minimal and/or a maximal value
- The median cuts the probabilities in half



These are interesting aspects of f_X but they barely summarize it

Expected value and variance

- When one speaks of the "expected price", "expected height", etc. one means the expected value of a random variable that is a price, a height, etc.

$$\mathrm{E}[X] = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = \int_{-\infty}^{\infty} x f_X(x)\, \mathrm{d}x$$

  The expected value of X is the "average value" of X.

  It is **not** the most probable value. The mean is one aspect of the distribution of X. The median or the mode are other interesting aspects.
- The variance is a measure of how far the values of a random variable are spread out from each other.
  If a random variable X has the expected value (mean) $\mu = \mathrm{E}[X]$, then the variance of X is given by:

$$\mathrm{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)\, \mathrm{d}x$$

- The standard deviation σ is the square root of the variance. This normalization allows to compare it with the expected value

Outline

How to estimate the Expected value?

To empirically estimate the expected value of a random variable X, one repeatedly measures observations of the variable and computes the arithmetic mean of the results

This is called the sample mean

Unfortunately, if you repeat the estimation, you may get a different value since X is a random variable ...

Central Limit Theorem [**CLT**]

- Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample of size n (i.e., a sequence of independent and identically distributed random variables with expected values μ and variances σ^2)
- The sample mean of these random variables is:

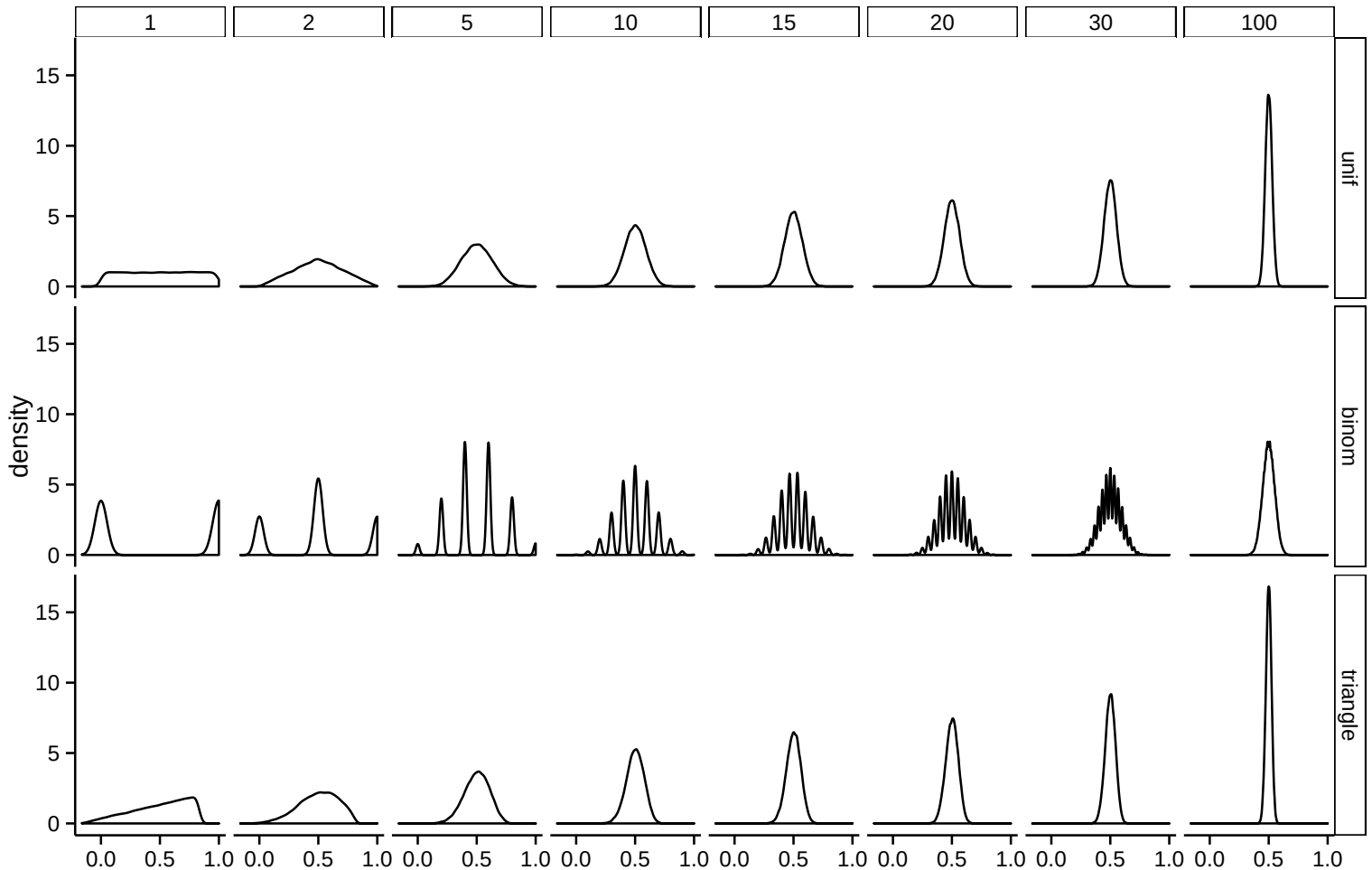
$$S_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

  S_n is a random variable too!
- For large n's, the distribution of S_n is approximately normal with mean μ and variance $\frac{\sigma^2}{n}$

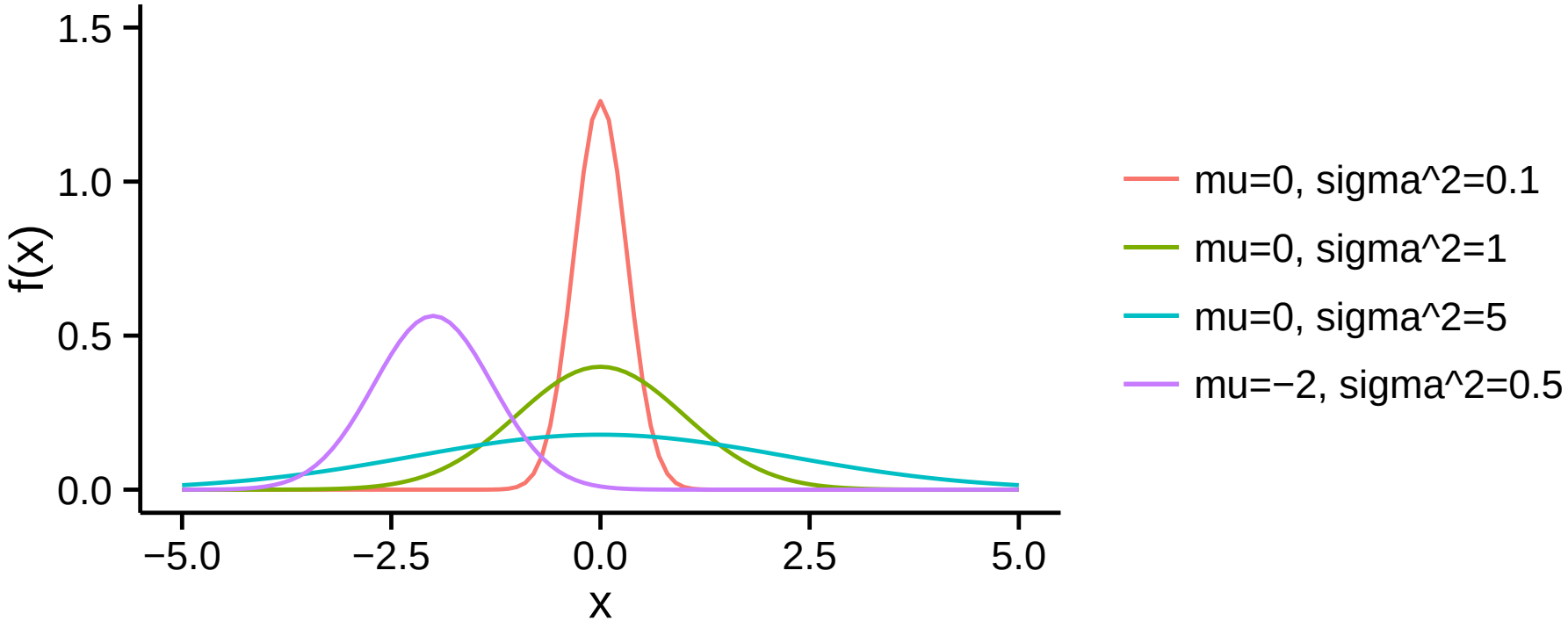
$$S_n \xrightarrow[n\to\infty]{} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

CLT Illustration: the mean smooths distributions

Start with an arbitrary distribution and compute the distribution of S_n for increasing values of n.

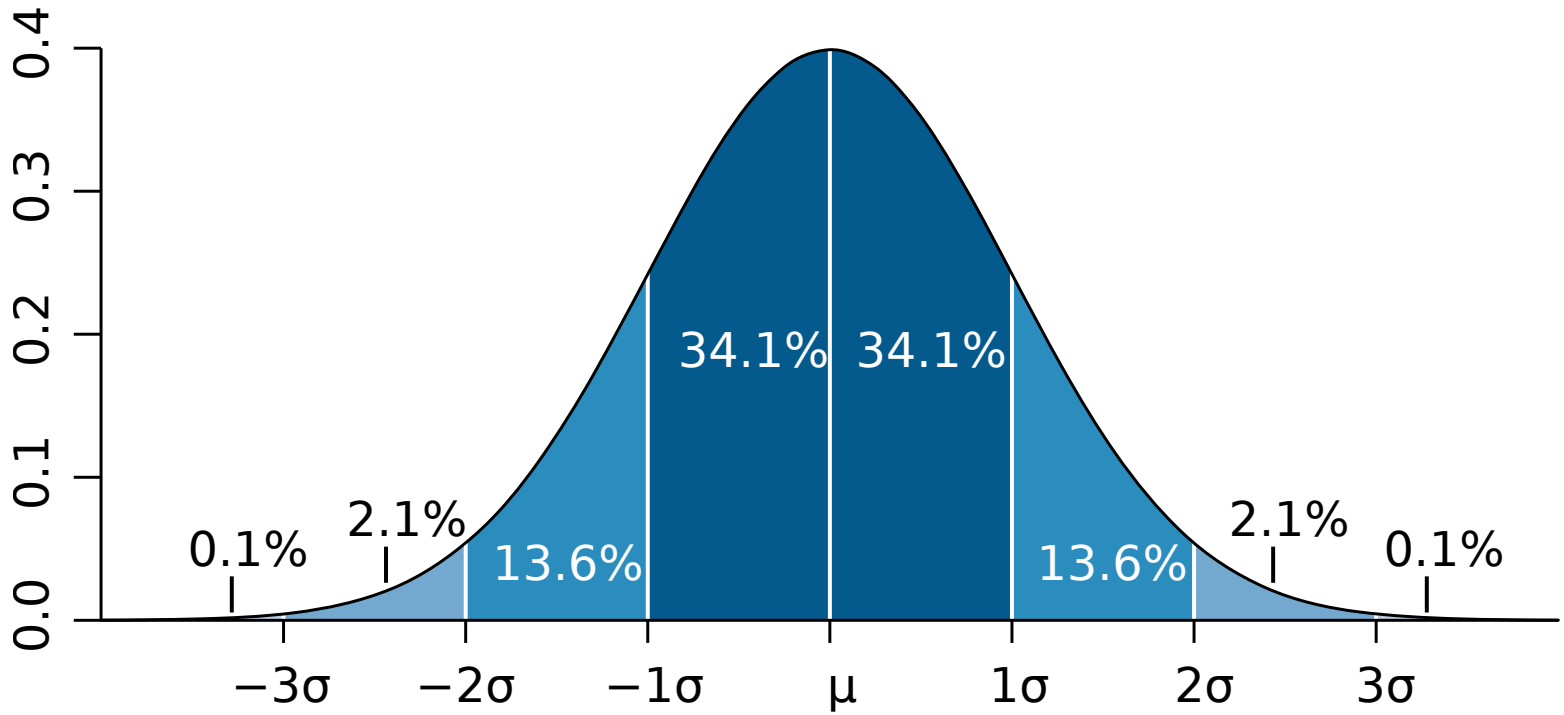


The Normal Distribution



The smaller the variance the more "spiky" the distribution.

The Normal Distribution

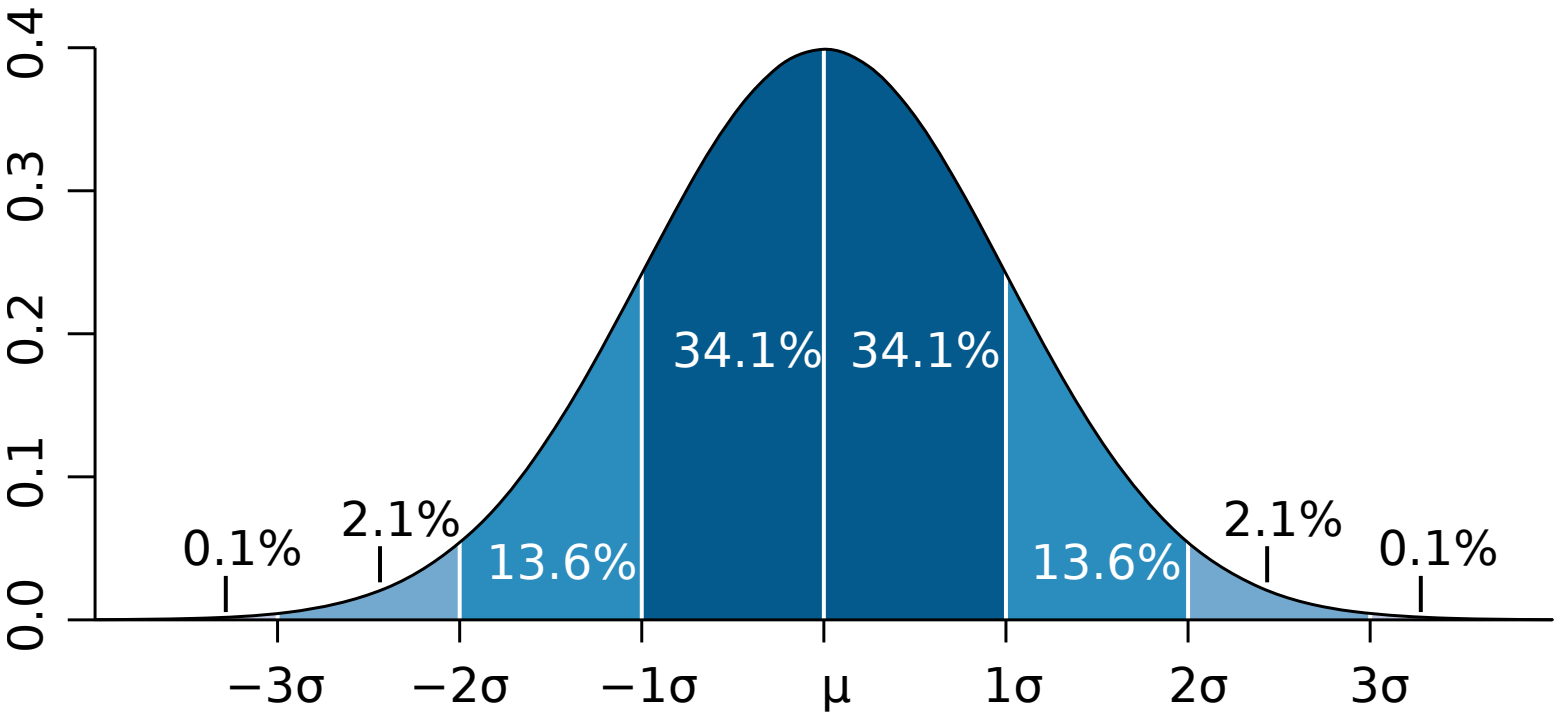


The smaller the variance the more "spiky" the distribution.

- Dark blue is less than one standard deviation from the mean. For the normal distribution, this accounts for about 68% of the set.
- Two standard deviations from the mean (medium and dark blue) account for about 95%
- Three standard deviations (light, medium, and dark blue) account for about 99.7%

Outline

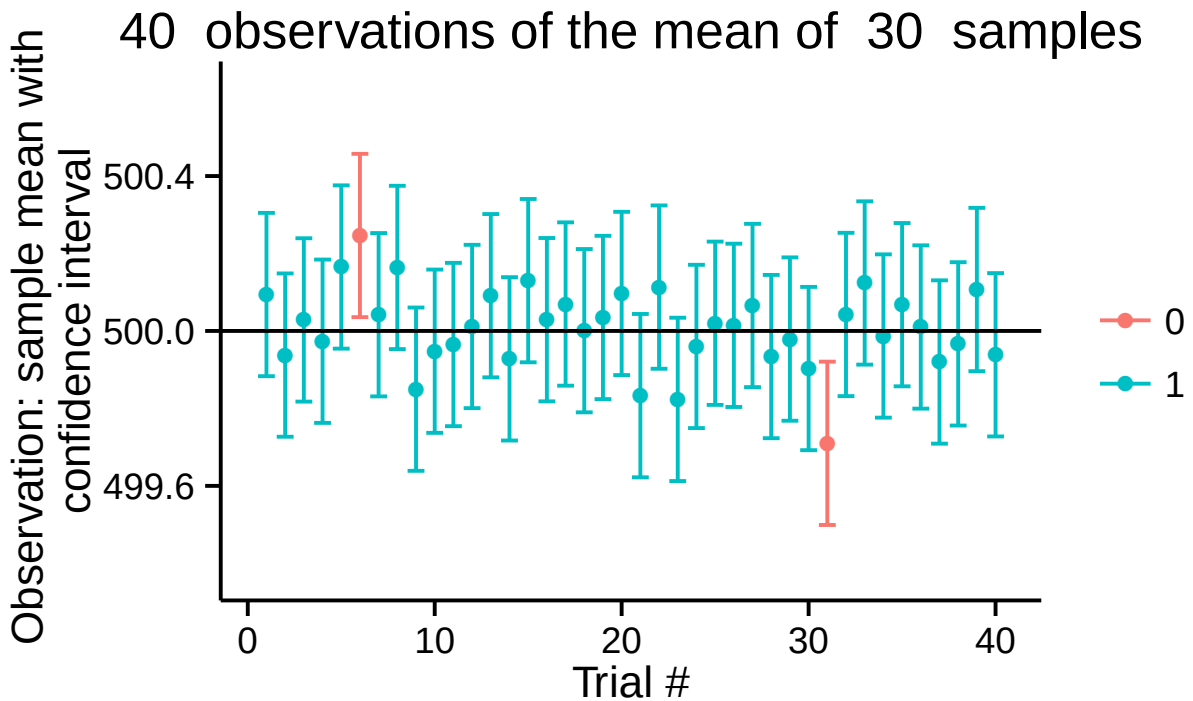
CLT consequence: confidence interval



When n is large:

$$\mathrm{P}\left(\mu \in \left[S_n - 2\frac{\sigma}{\sqrt{n}}, S_n + 2\frac{\sigma}{\sqrt{n}}\right]\right) = \mathrm{P}\left(S_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]\right) \approx 95\%$$

CLT consequence: confidence interval



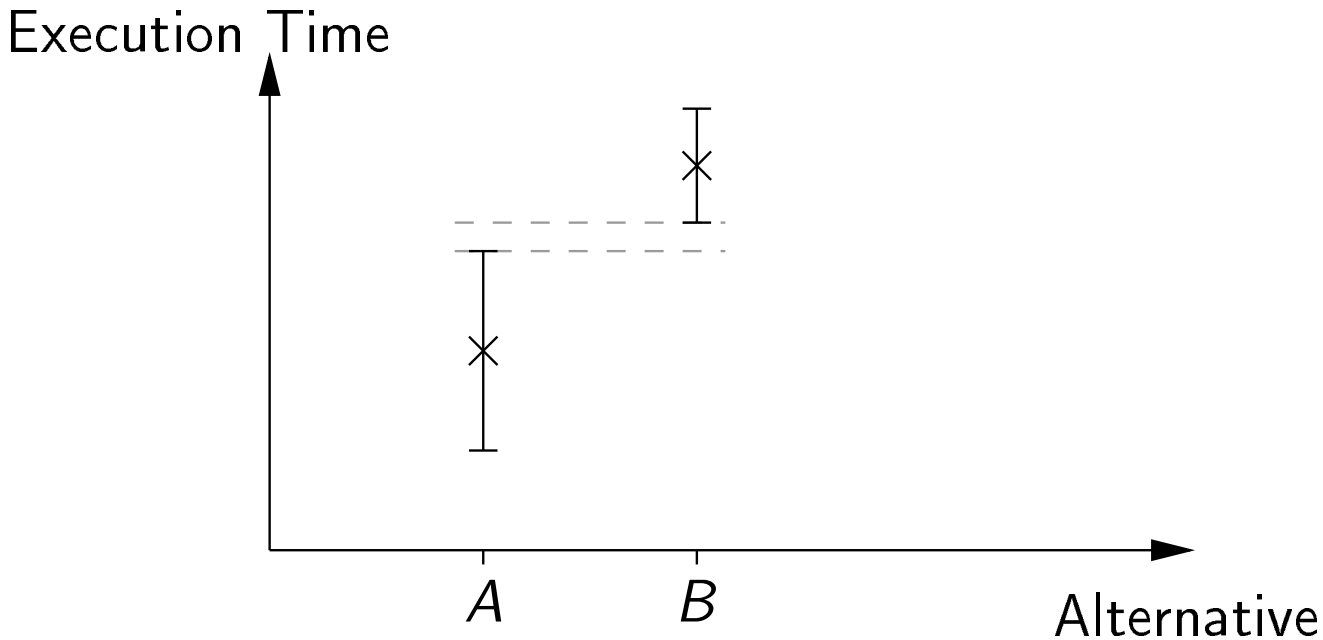
When n is large:

$$\mathrm{P}\left(\mu \in \left[S_n - 2\frac{\sigma}{\sqrt{n}}, S_n + 2\frac{\sigma}{\sqrt{n}}\right]\right) = \mathrm{P}\left(S_n \in \left[\mu - 2\frac{\sigma}{\sqrt{n}}, \mu + 2\frac{\sigma}{\sqrt{n}}\right]\right) \approx 95\%$$

There is 95% of chance that the true mean lies within $2\frac{\sigma}{\sqrt{n}}$ of the sample mean.

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate there expected values μ_A and μ_B

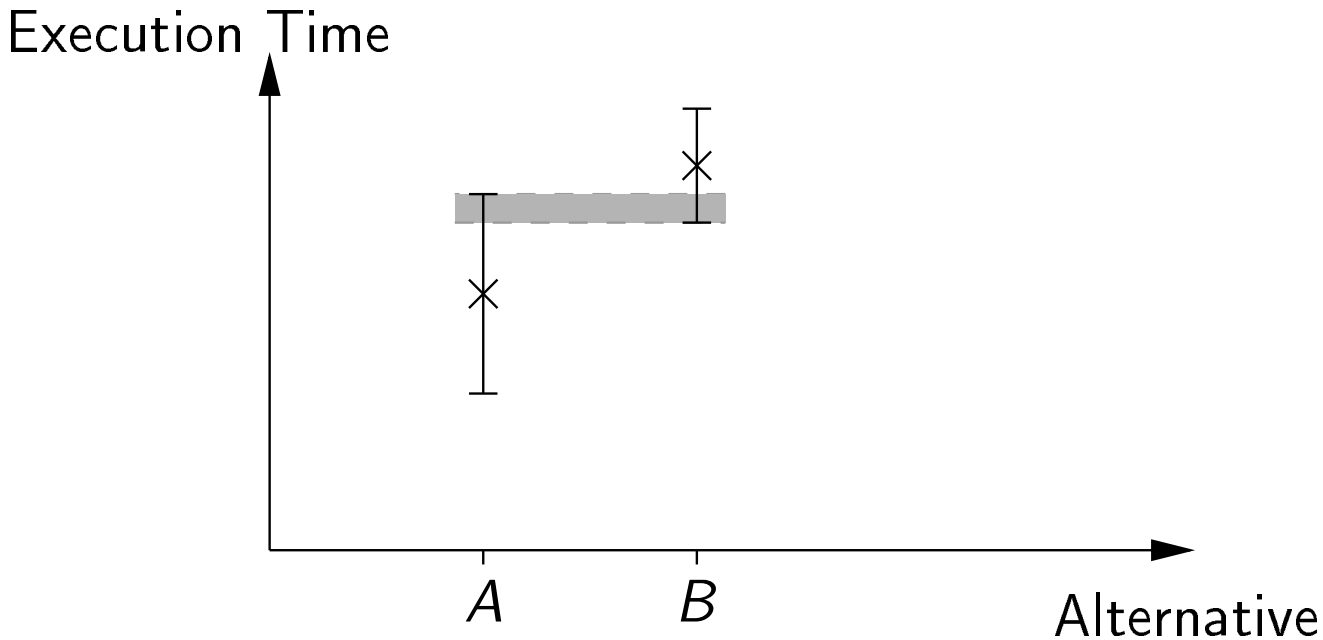


The two 95% confidence intervals do not overlap

$\rightsquigarrow \mu_A < \mu_B$ with more than 90% of confidence ☺

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate there expected values μ_A and μ_B



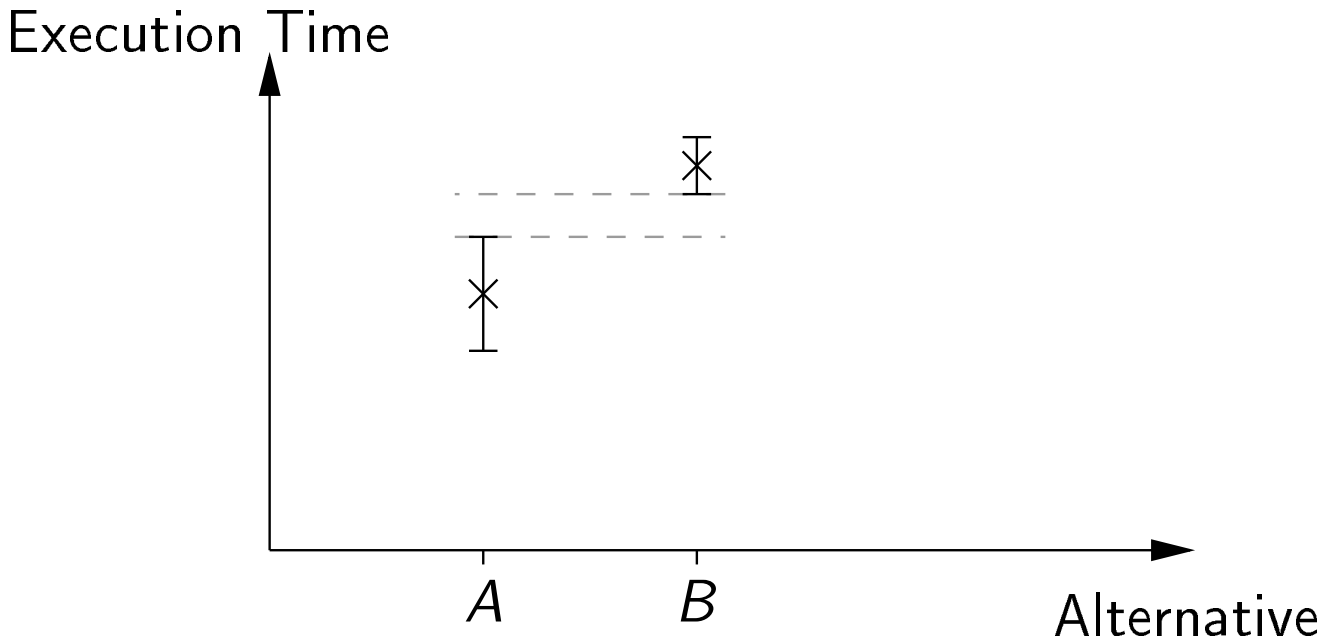
The two 95% confidence intervals do overlap

$\rightsquigarrow$ Nothing can be concluded ☹

Reduce C.I?

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate there expected values μ_A and μ_B

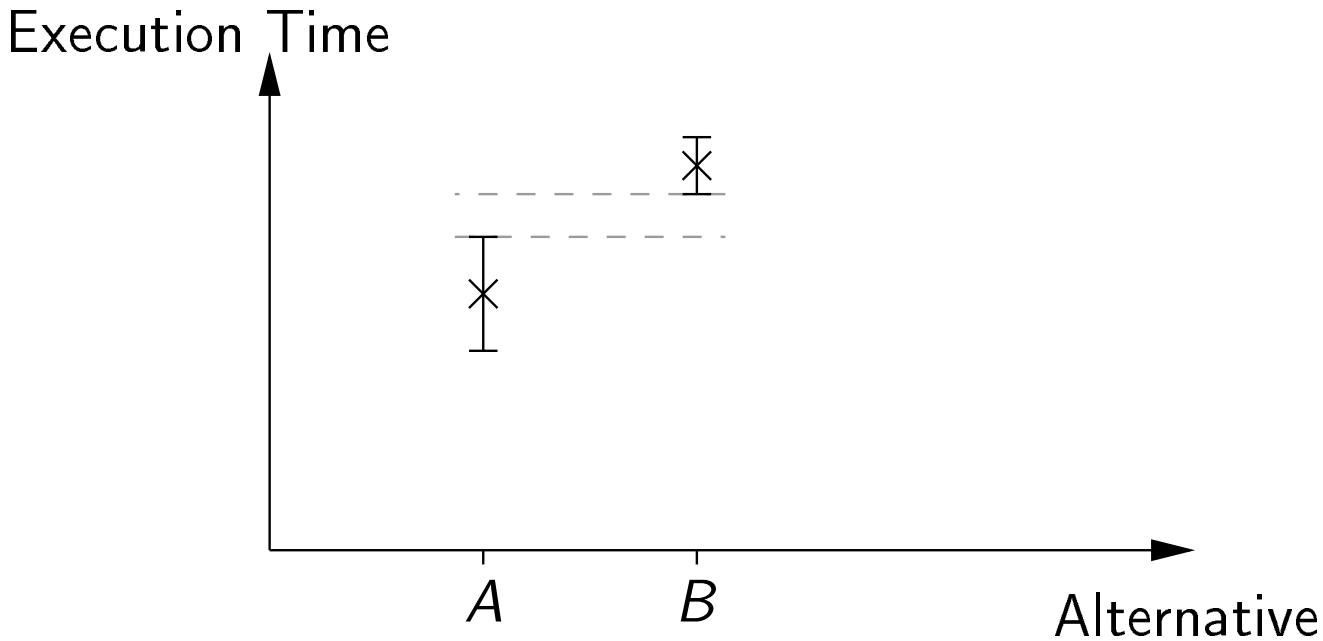


The two 70% confidence intervals do not overlap

$\rightsquigarrow \mu_A < \mu_B$ with less than 50% of confidence ☹ $\rightsquigarrow$ more experiments...

Without any particular hypothesis

- Assume, you have evaluated two alternatives A and B on n different setups
- You therefore consider the associated random variables A and B and try to estimate there expected values μ_A and μ_B



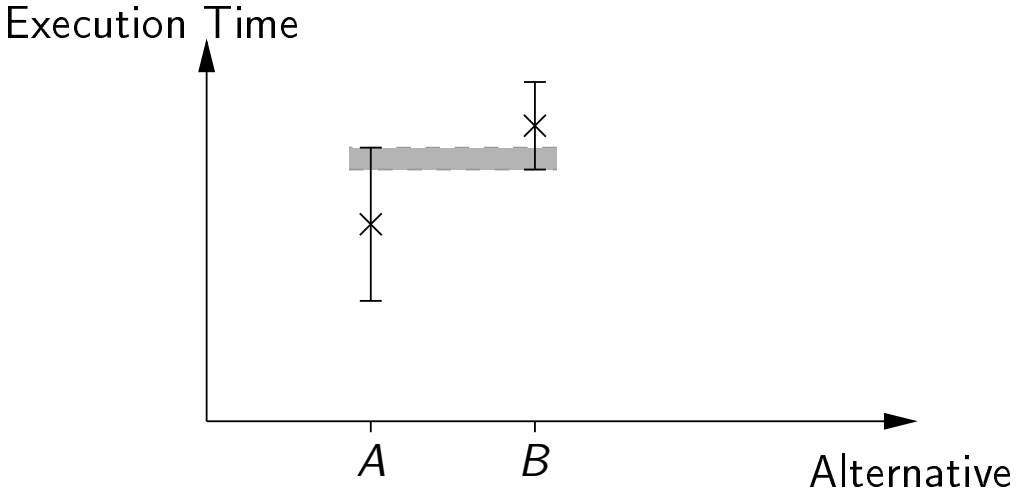
The width of the confidence interval is proportional to $\frac{\sigma}{\sqrt{n}}$

Halving C.I. requires 4 times more experiments! ☹

Try to reduce variance if you can... ☺

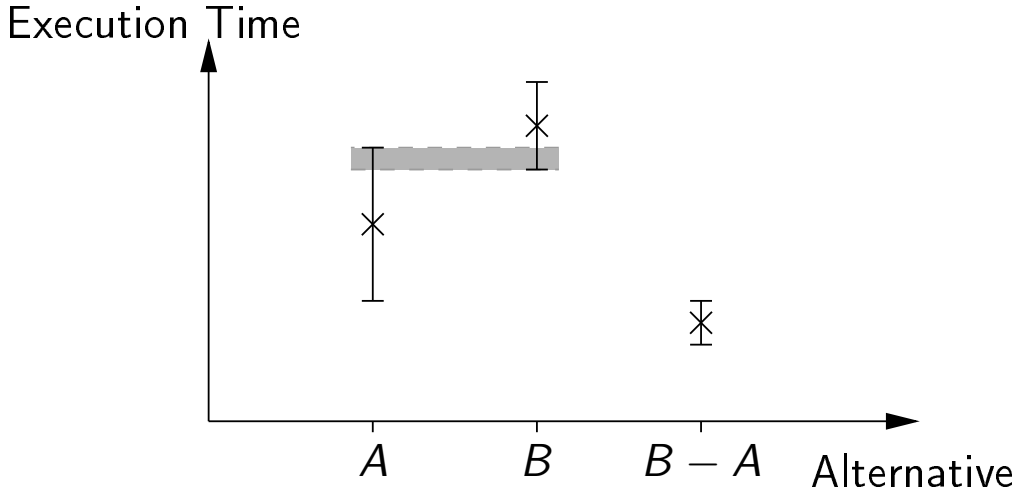
Exploiting blocks

- C.I.s overlap because variance is large. Some *setups* may have an intrinsically longer duration than others, hence a large Var(A) and Var(B)



Exploiting blocks

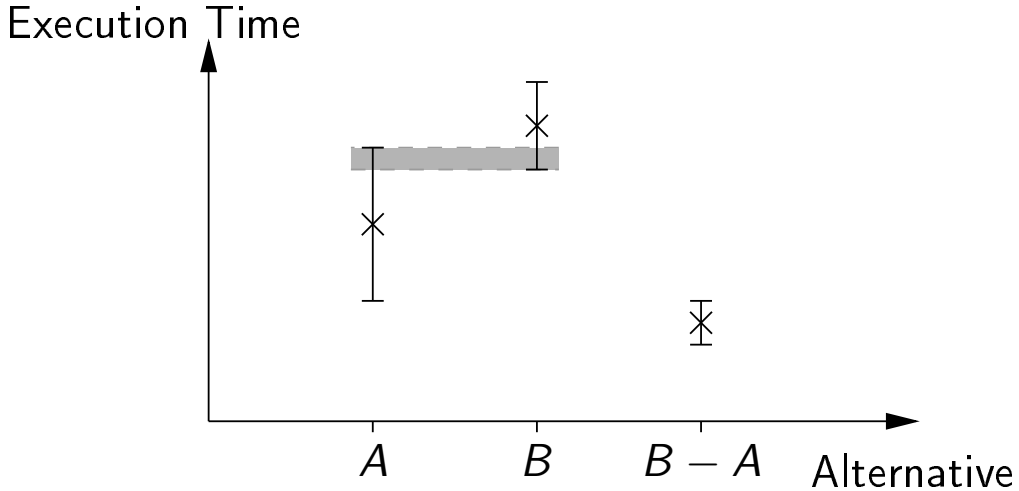
- C.I.s overlap because variance is large. Some *setups* may have an intrinsically longer duration than others, hence a large Var(A) and Var(B)



- The previous test estimates μ_A and μ_B independently.
  $\mathrm{E}[A] < \mathrm{E}[B] \Leftrightarrow \mathrm{E}[B - A] > 0$.
  In the previous evaluation, the same setup i is used for measuring A_i and B_i, hence we can focus on $B - A$.
  Since Var($B - A$) is much smaller than Var(A) and Var(B), we can conclude that $\mu_A < \mu_B$ with 95% of confidence.

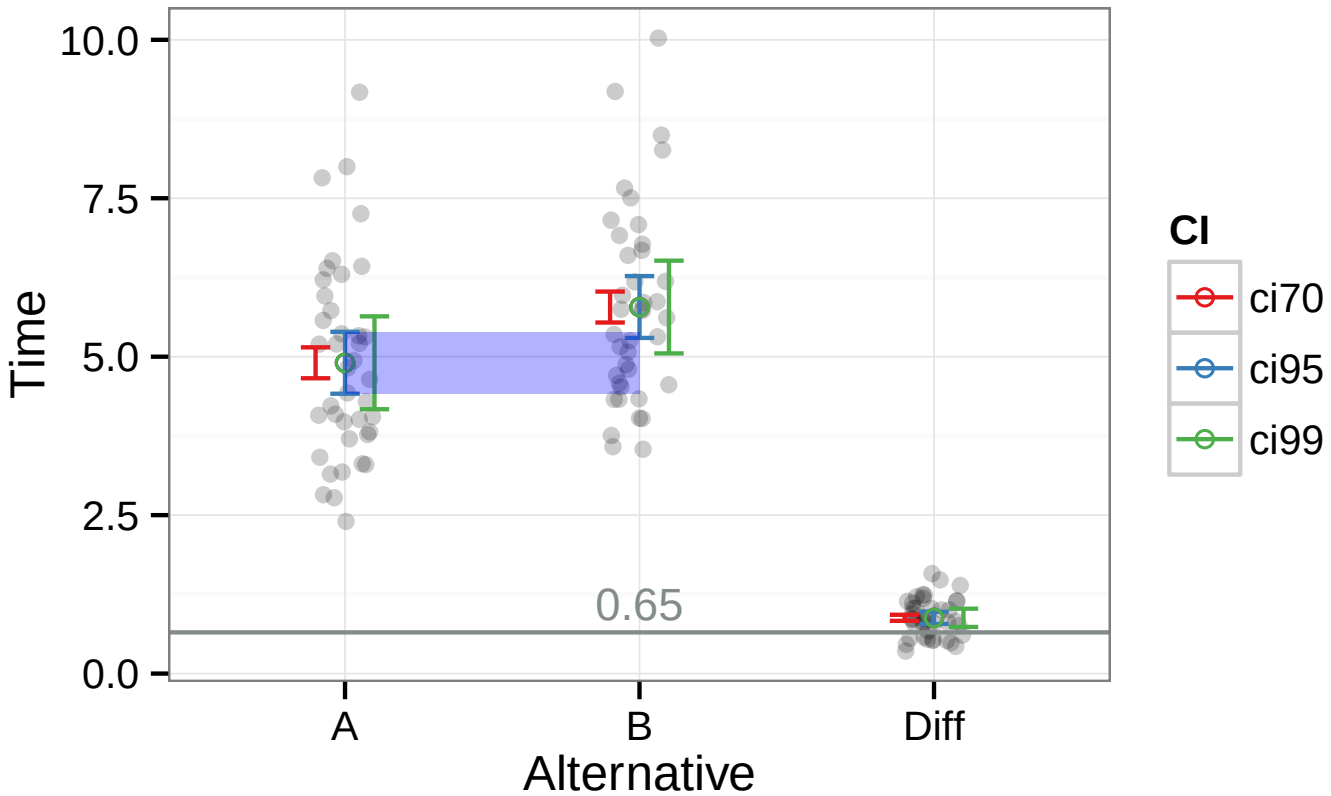
Exploiting blocks

- C.I.s overlap because variance is large. Some *setups* may have an intrinsically longer duration than others, hence a large Var(A) and Var(B)



- The previous test estimates μ_A and μ_B independently.
  $\mathrm{E}[A] < \mathrm{E}[B] \Leftrightarrow \mathrm{E}[B - A] > 0$.
  In the previous evaluation, the same setup i is used for measuring A_i and B_i, hence we can focus on $B - A$.
  Since Var($B - A$) is much smaller than Var(A) and Var(B), we can conclude that $\mu_A < \mu_B$ with 95% of confidence.
- Relying on such common points is called blocking and enable to reduce variance.

Let's reuse a previous example



μ_A is 0.65 seconds smaller than μ_B with more than 99% of confidence ☺

You need to invest in a probabilistic model. Here we assumed:

- $A_i = \boxed{S_i} + A'_i$
- $B_i = \boxed{S_i} + B'_i$

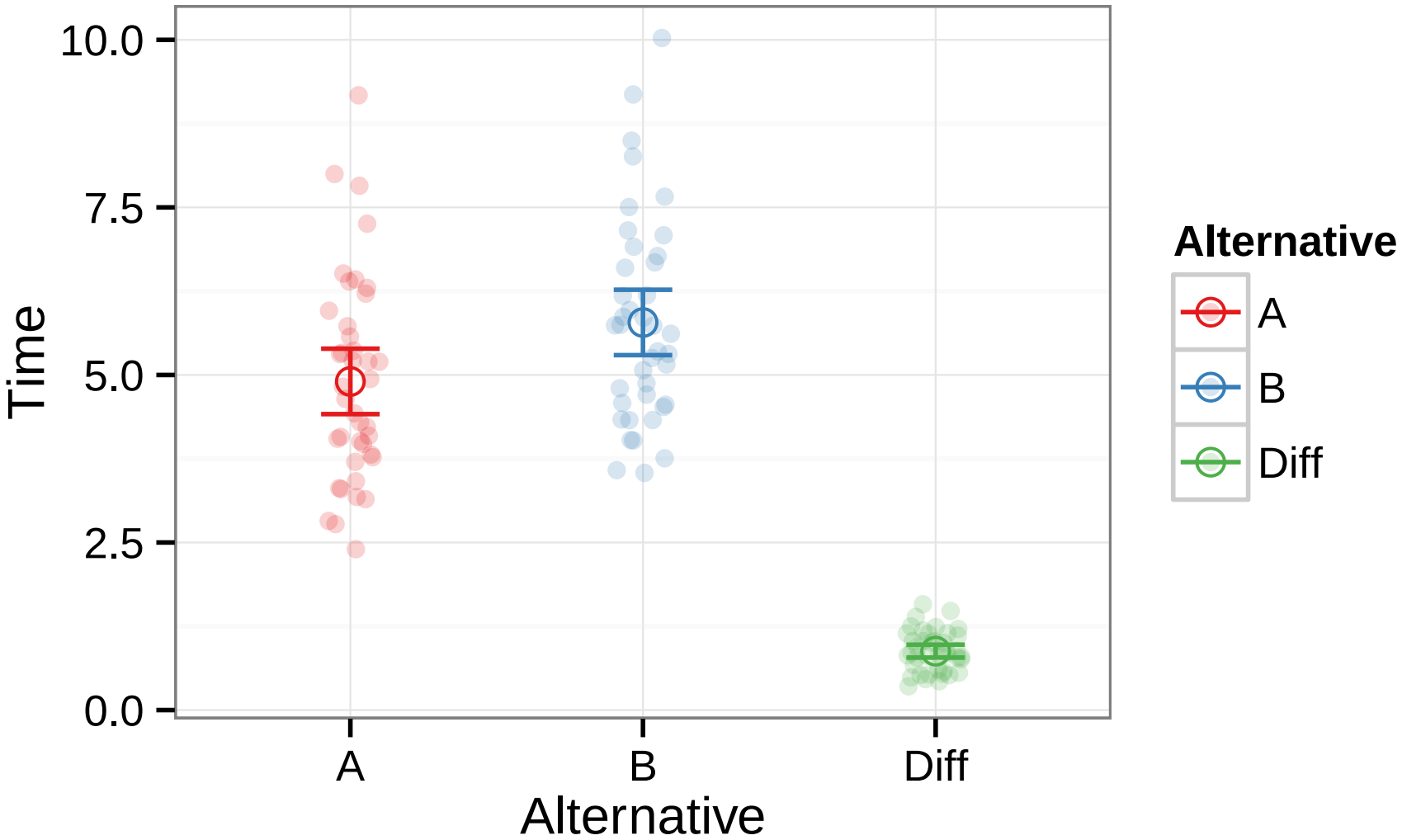
So we could subtract them ☺

Dividing them would have been a very bad idea... ☹

How to compute and plot CI in R: code

```r
library(ggplot2)
library(dplyr)
library(tidyr)
df = read.csv("data/set1.csv",header=T)
df$Diff=df$B-df$A # Assuming observations are paired!
dfgg = df %>% gather(Alternative, Time)
dfsum = dfgg %>%
        group_by(Alternative) %>%
        summarise(num = n(), mean = mean(Time), sd = sd(Time),
 se = 2*sd/sqrt(num))
ggplot(dfgg,aes(x=Alternative,y=Time,color=Alternative)) +
     scale_color_brewer(palette="Set1") + theme_bw() +
     geom_jitter(alpha=.2,position = position_jitter(width = .1)) +
     geom_errorbar(data=dfsum,width=.2,
   aes(y=mean,ymin=mean-se,ymax=mean+se)) +
     geom_point(data=dfsum,shape=21, size=3,
aes(y=mean,color=Alternative))
```

How to compute and plot CI in R: output

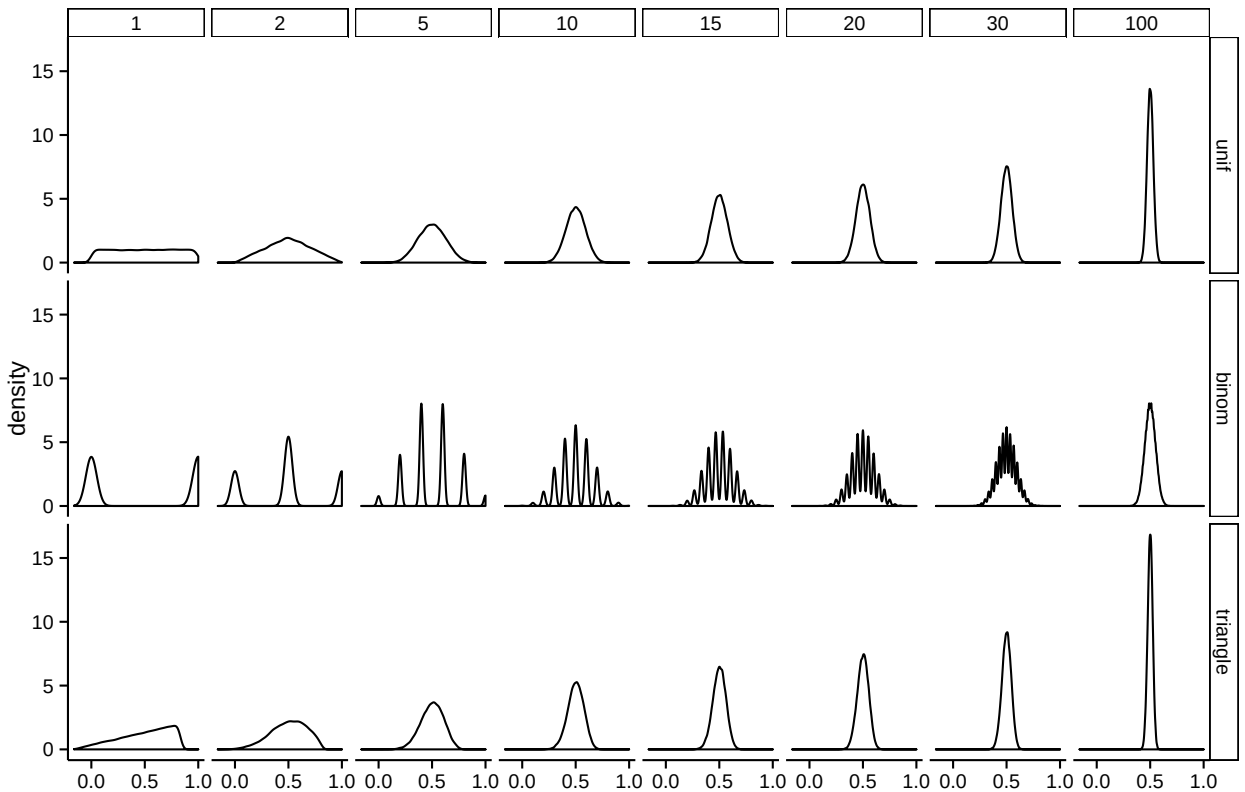


Outline

CLT hypothesis

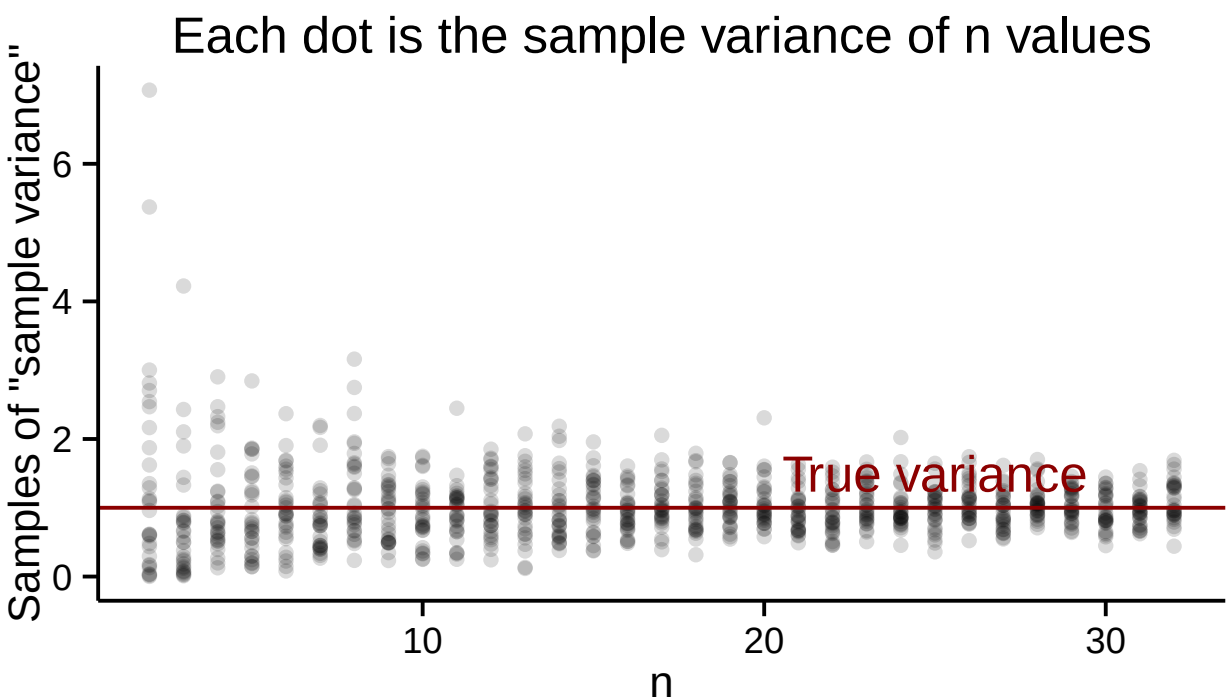
- The CLT hypothesis are very weak: it does not assume any particular distribution (e.g., normality) for X

  But, the CLT says that *when n goes large*, the sample mean is *normally distributed*. We have seen it holds true quickly



CLT hypothesis

- The CLT hypothesis are very weak: it does not assume any particular distribution (e.g., normality) for X
  But, the CLT says that *when n goes large*, the sample mean is *normally distributed*. We have seen it holds true quickly
- However, the CLT uses $\sigma = \sqrt{\text{Var}(X)}$ but we only have the sample variance, not the true variance

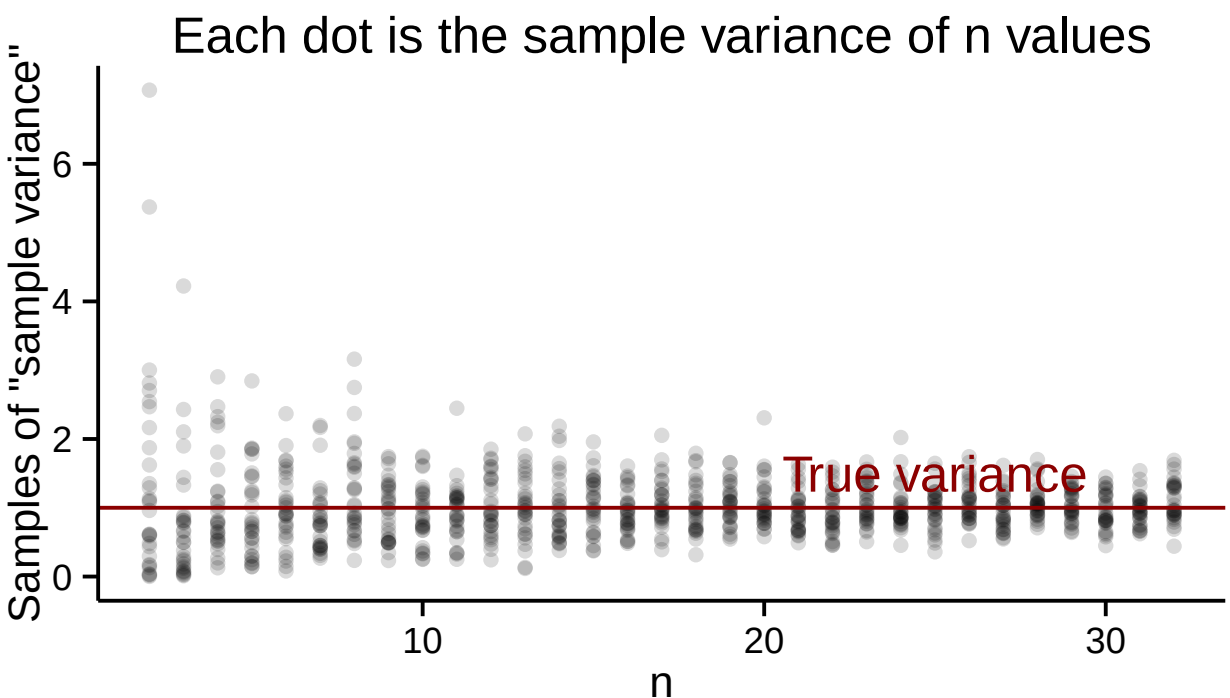


CLT hypothesis

- The CLT hypothesis are very weak: it does not assume any particular distribution (e.g., normality) for X

  But, the CLT says that *when n goes large*, the sample mean is *normally distributed*. We have seen it holds true quickly
- However, the CLT uses $\sigma = \sqrt{\text{Var}(X)}$ but we only have the sample variance, not the true variance

  So you should always try to either find an upper bound on the true variance or overestimate the sample variance (e.g., `se=4*sd/sqrt(num)`)



How many replicates?

- **Q:** How Many Replicates?

How many replicates?

- **Q:** How Many Replicates?

A1: How many can you afford?

How many replicates?

- **Q:** How Many Replicates?

  A1: How many can you afford?

  A2: 30...

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work)

How many replicates?

- **Q:** How Many Replicates?

  A1: How many can you afford?

  A2: 30...

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work)

- With less than 30, you should make the C.I. wider using e.g., the Student law.

How many replicates?

- **Q:** How Many Replicates?

  A1: How many can you afford?

  A2: 30...

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work)

- With less than 30, you should make the C.I. wider using e.g., the Student law.
- Once you have a first C.I. with 30 samples, you can estimate how many samples will be required to answer your question. If it is too large, then either try to reduce variance (or the scope of your experiments) or simply explain that the two alternatives are hardly distinguishable... You need a sequential approach.

How many replicates?

- **Q:** How Many Replicates?

  A1: How many can you afford?

  A2: 30...

  Rule of thumb: a sample of 30 or more is big sample but a sample of 30 or less is a small one (doesn't always work)

- With less than 30, you should make the C.I. wider using e.g., the Student law.
- Once you have a first C.I. with 30 samples, you can estimate how many samples will be required to answer your question. If it is too large, then either try to reduce variance (or the scope of your experiments) or simply explain that the two alternatives are hardly distinguishable... You need a sequential approach.
- **Running the right number of experiments enables to get to conclusions more quickly and hence to test other hypothesis.**

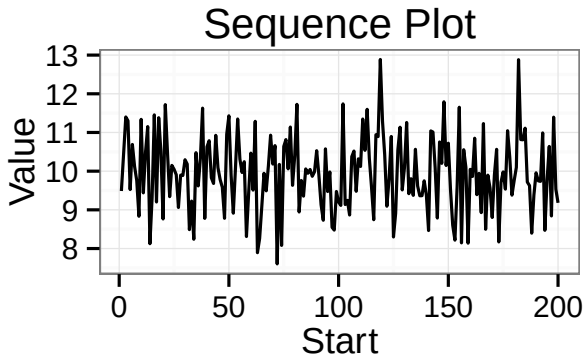
Key Hypothesis

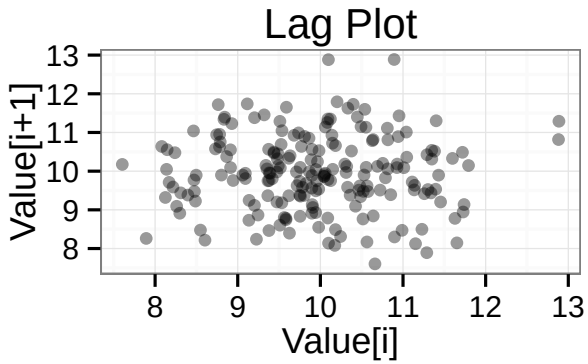
The hypothesis of CLT are very weak. Yet, to qualify as replicates, the repeated measurements:

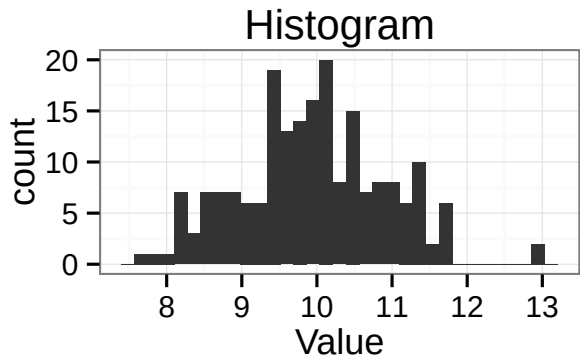
- must be independent (take care of warm-up)
- must not be part of a time series (the system behavior may temporary change)
- must not come from the same place (the machine may have a problem)
- must be of appropriate spatial scale

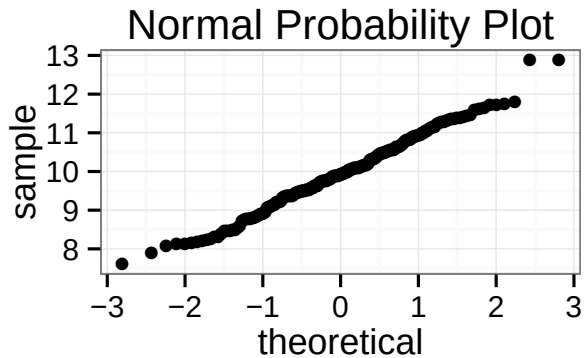
Perform graphical checks

Simple Graphical Checks









Fixed Location the run sequence plot should be flat and non-drifting

Fixed Variation the vertical spread in the run sequence plot should approximately the same over the entire horizontal axis

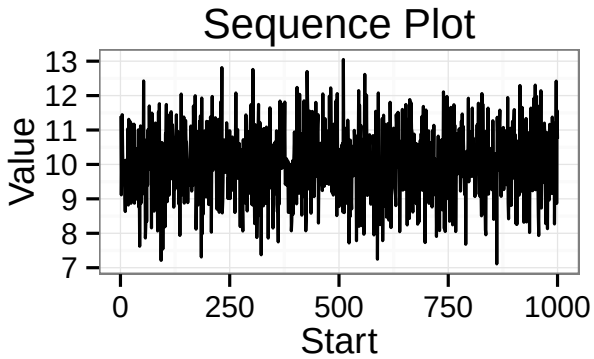
Independence the lag plot should be structureless

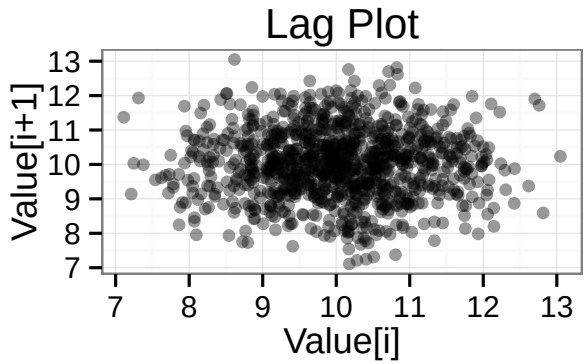
Fixed Distribution (, in particular if the *fixed normal distribution* assumption holds)

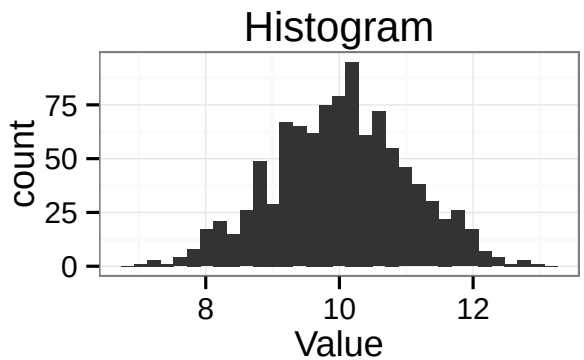
- the histogram should be bell-shaped, and
- the normal probability plot should be linear

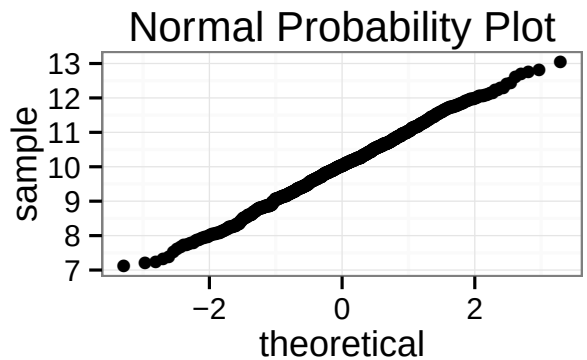
If you see several modes, there may be an hidden parameter to take into account

Simple Graphical Checks









Fixed Location the run sequence plot should be flat and non-drifting

Fixed Variation the vertical spread in the run sequence plot should approximately the same over the entire horizontal axis

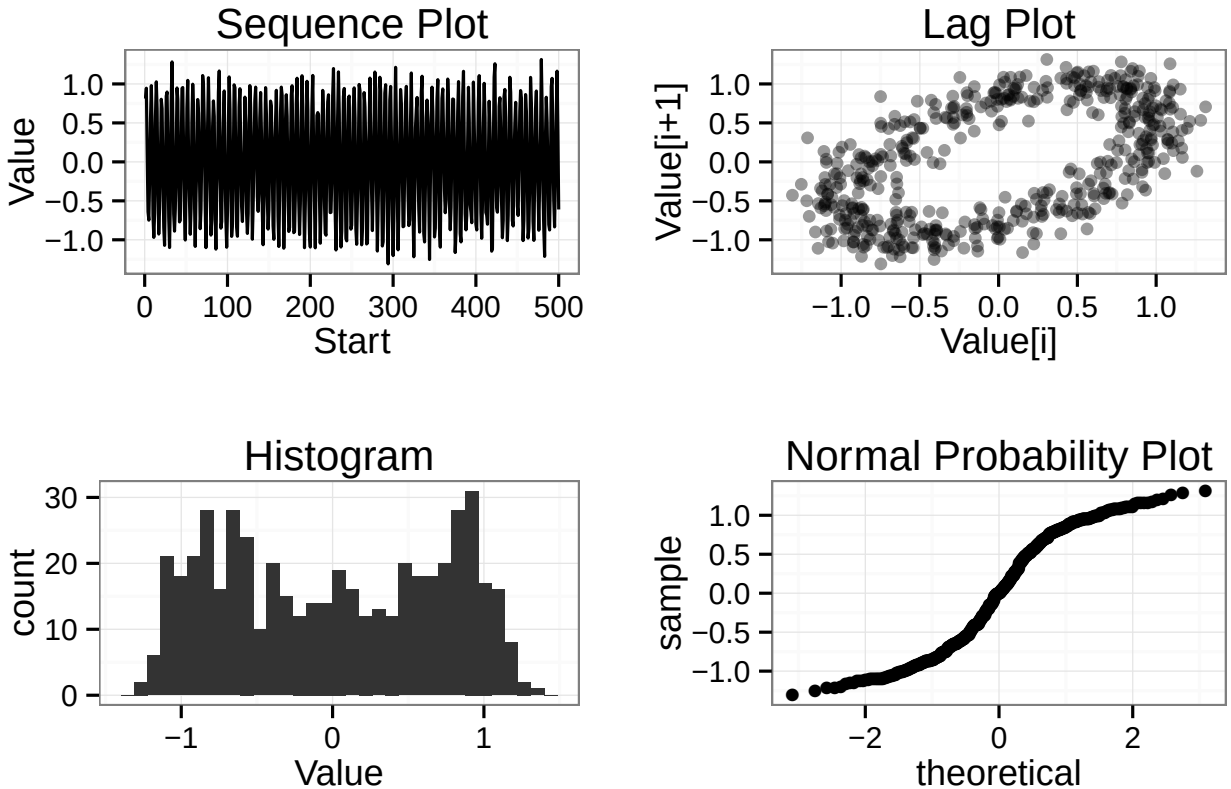
Independence the lag plot should be structureless

Fixed Distribution (, in particular if the *fixed normal distribution* assumption holds)

- the histogram should be bell-shaped, and
- the normal probability plot should be linear

If you see several modes, there may be an hidden parameter to take into account

Simple Graphical Checks



Fixed Location the run sequence plot should be flat and non-drifting

Fixed Variation the vertical spread in the run sequence plot should approximately the same over the entire horizontal axis

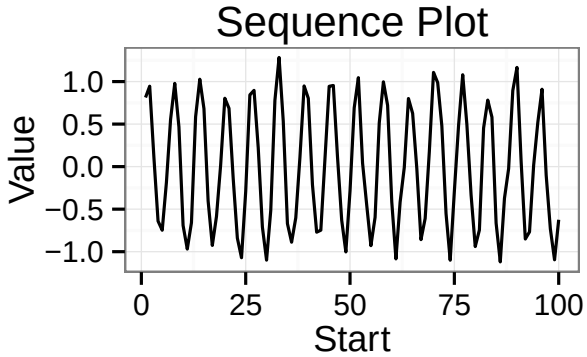
Independence the lag plot should be structureless

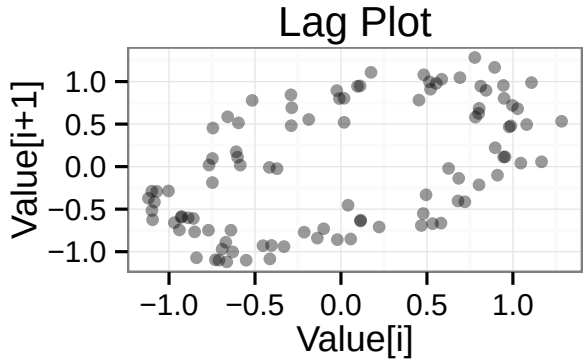
Fixed Distribution (, in particular if the *fixed normal distribution* assumption holds)

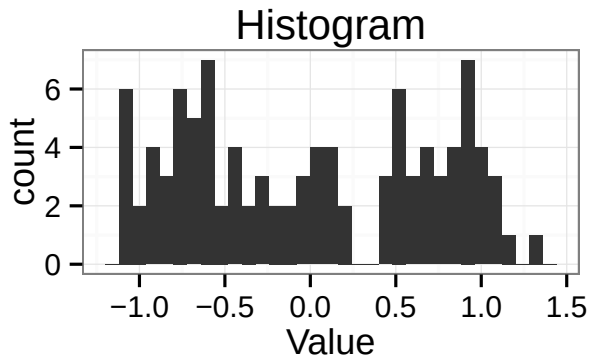
- the histogram should be bell-shaped, and
- the normal probability plot should be linear

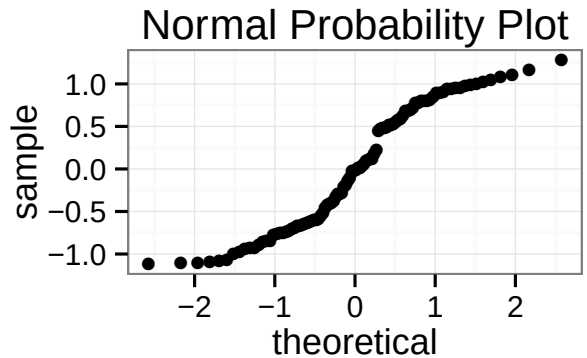
If you see several modes, there may be an hidden parameter to take into account

Simple Graphical Checks









Fixed Location the run sequence plot should be flat and non-drifting

Fixed Variation the vertical spread in the run sequence plot should approximately the same over the entire horizontal axis

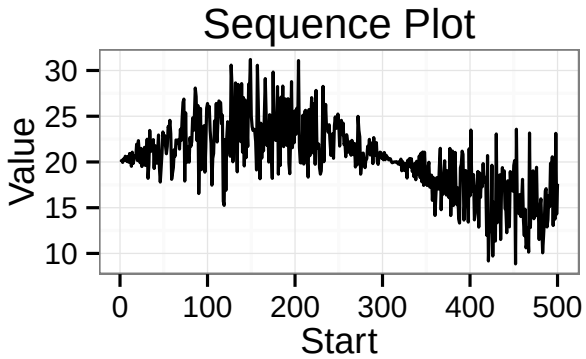
Independence the lag plot should be structureless

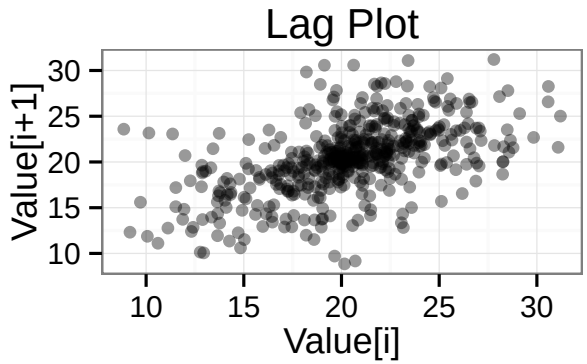
Fixed Distribution (, in particular if the *fixed normal distribution* assumption holds)

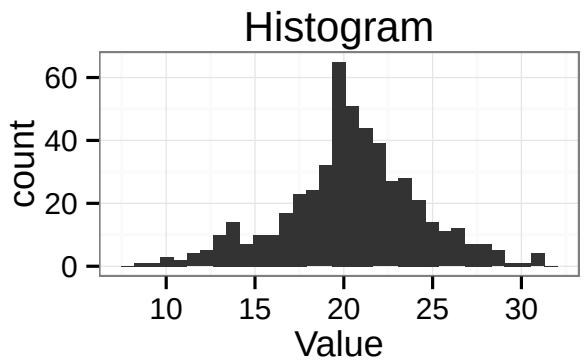
- the histogram should be bell-shaped, and
- the normal probability plot should be linear

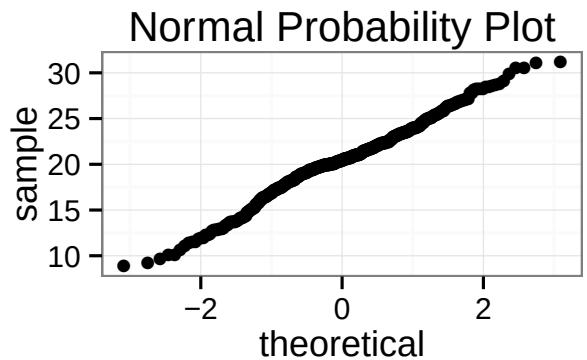
If you see several modes, there may be an hidden parameter to take into account

Simple Graphical Checks









Fixed Location the run sequence plot should be flat and non-drifting

Fixed Variation the vertical spread in the run sequence plot should approximately the same over the entire horizontal axis

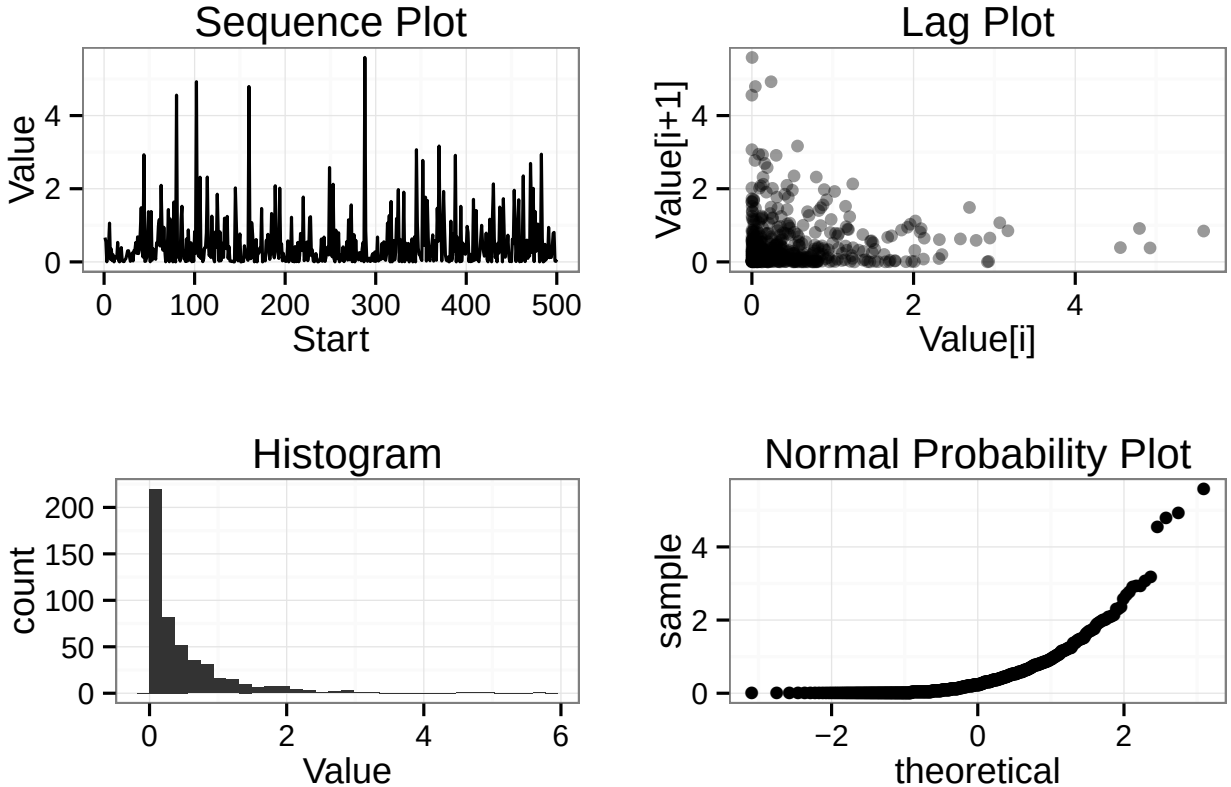
Independence the lag plot should be structureless

Fixed Distribution (, in particular if the *fixed normal distribution* assumption holds)

- the histogram should be bell-shaped, and
- the normal probability plot should be linear

If you see several modes, there may be an hidden parameter to take into account

Simple Graphical Checks



Fixed Location the run sequence plot should be flat and non-drifting

Fixed Variation the vertical spread in the run sequence plot should approximately the same over the entire horizontal axis

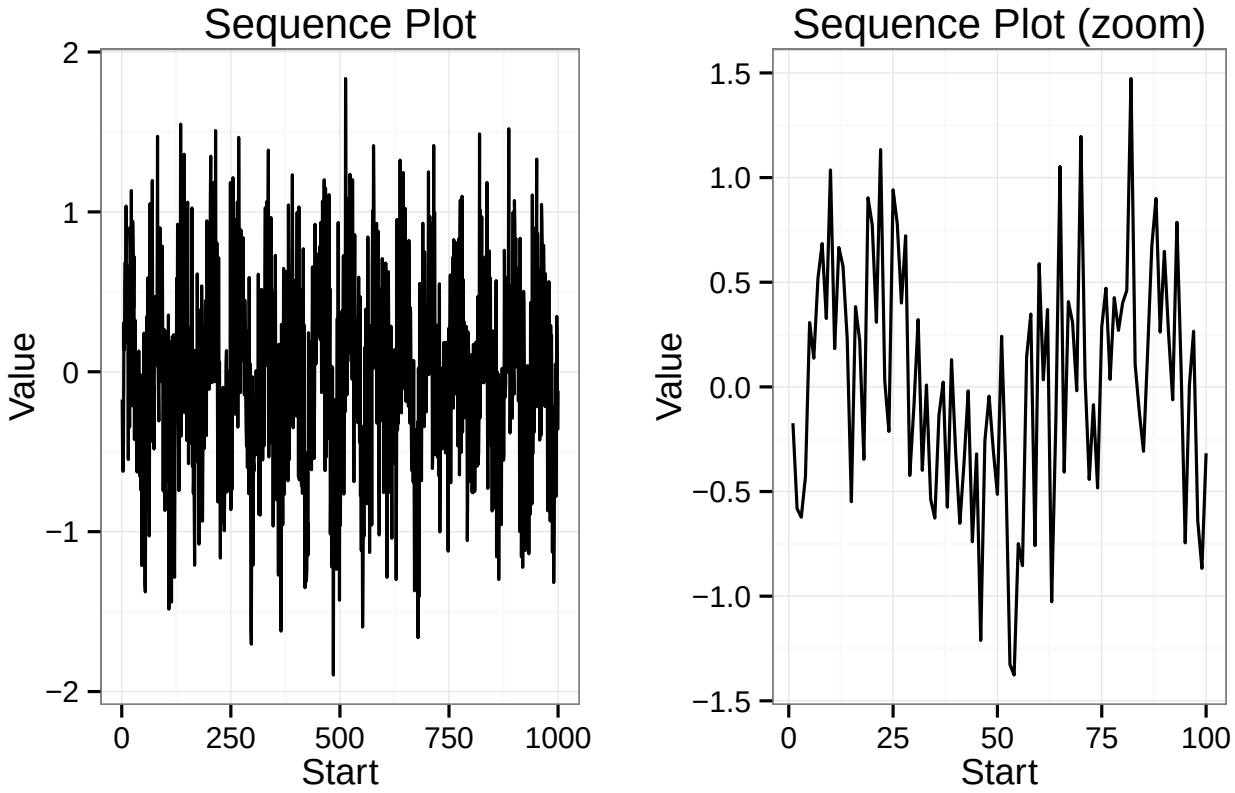
Independence the lag plot should be structureless

Fixed Distribution (, in particular if the *fixed normal distribution* assumption holds)

- the histogram should be bell-shaped, and
- the normal probability plot should be linear

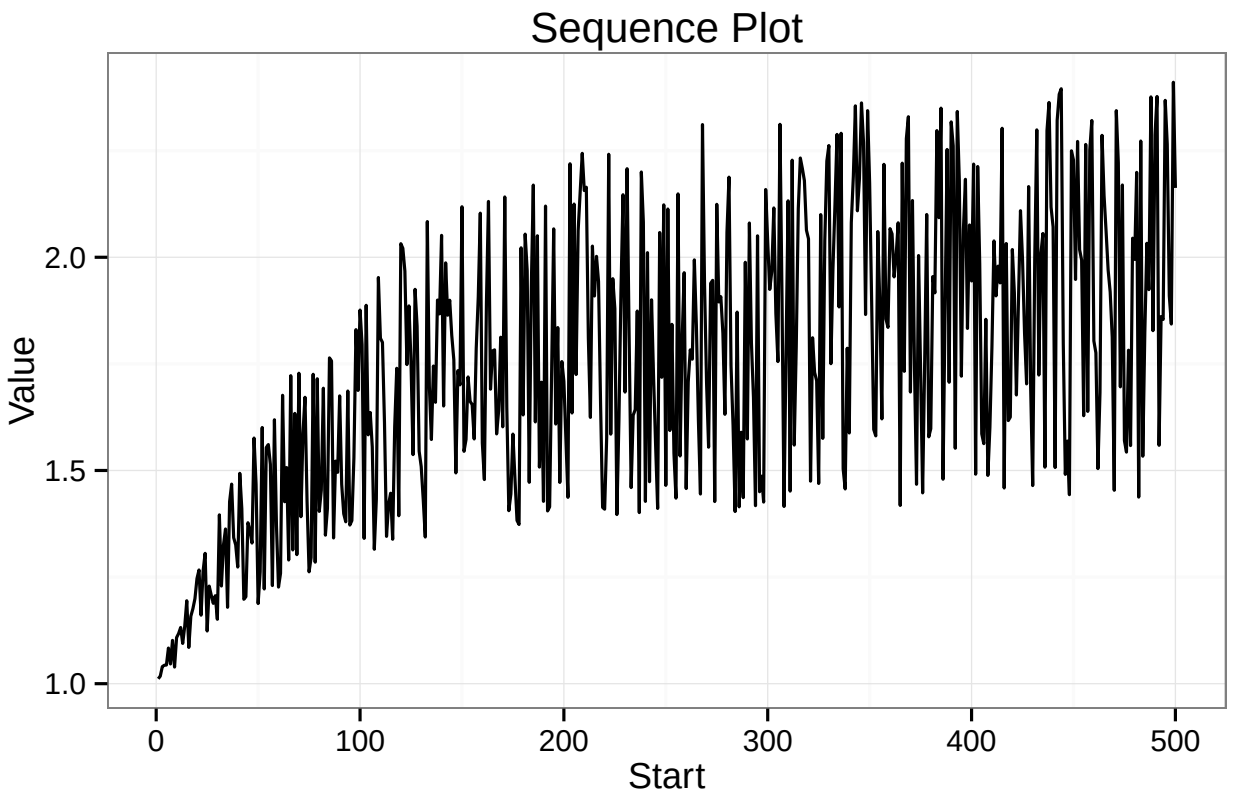
If you see several modes, there may be an hidden parameter to take into account

Temporal Dependancy



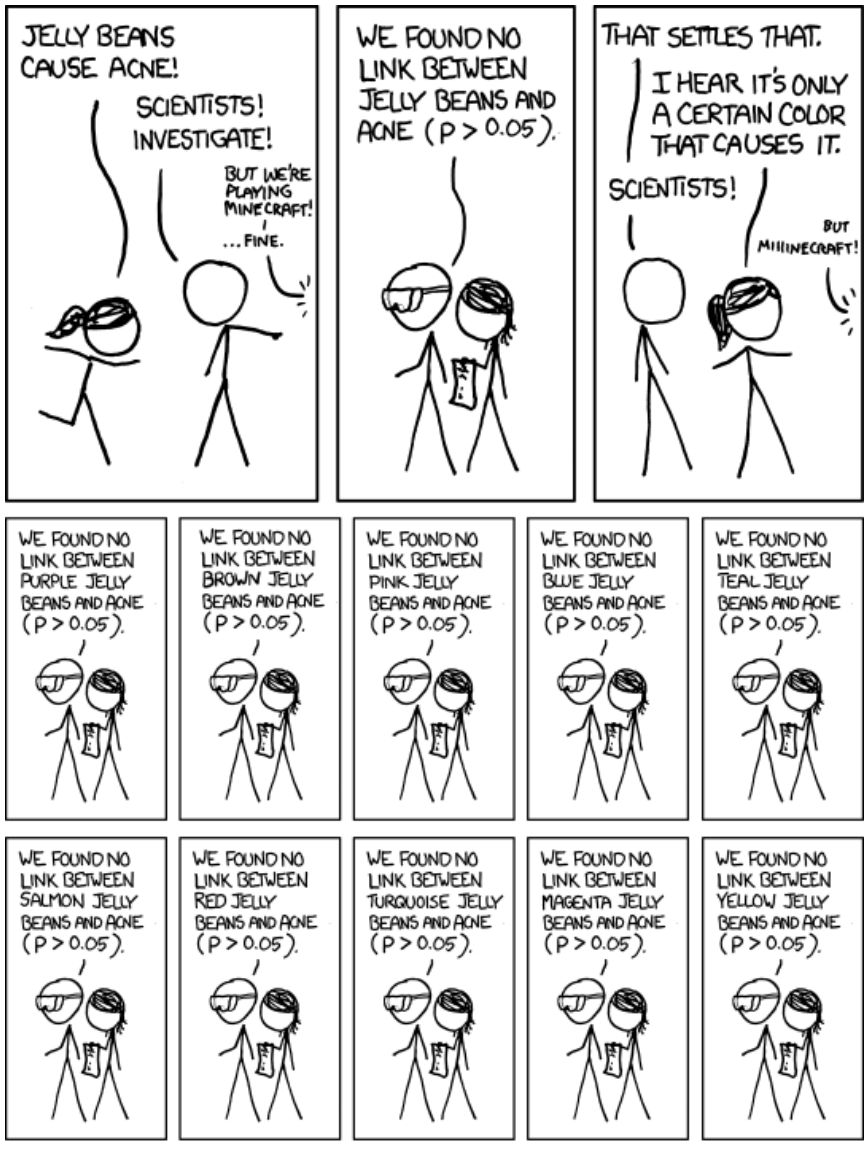
- Should look independent and statistically identical
- Periodicity : May depend on sampling frequency or on clock resolution
  - Study the period (Fourier), use time series
- Danger: temporal correlation $\rightsquigarrow$ study stationarity

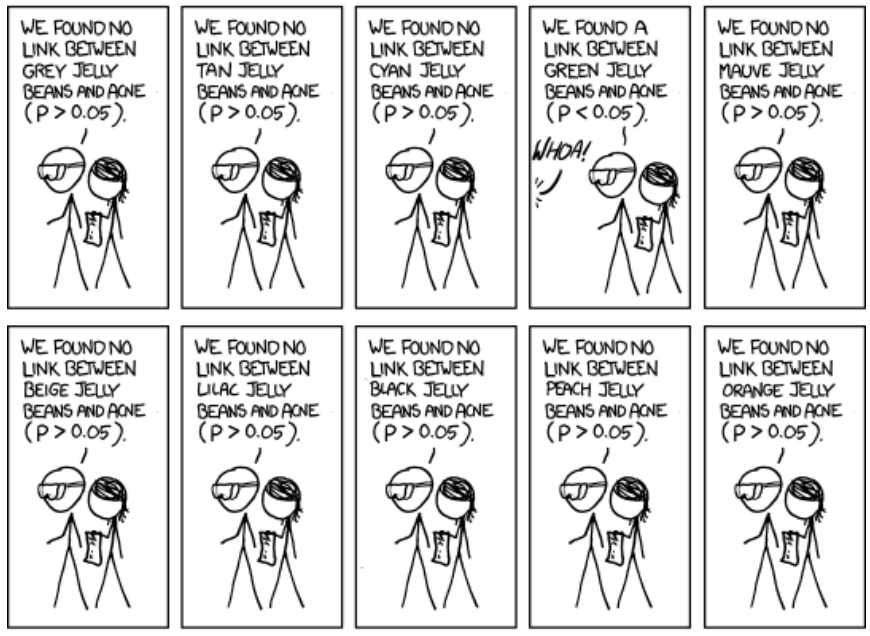
Detect Trends

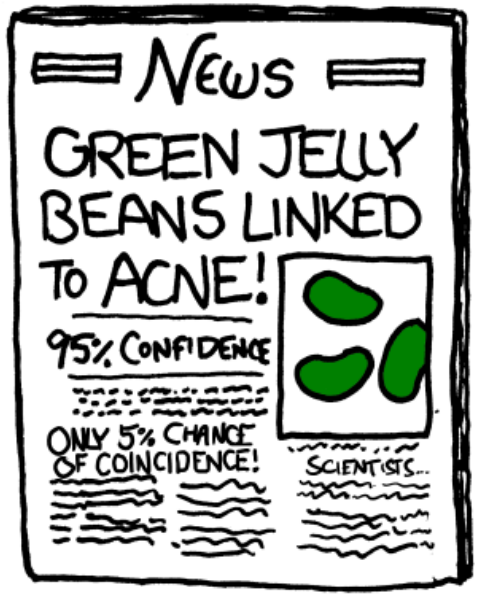


- Model the trend: here increases then saturates
- Possibly remove the trend by compensating it (multiplicative factor here) or removing what can be identified as a warm-up

Confidence. . .


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

Outline

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea
- You should better do $A, B, \quad A, B, \quad A, B, \quad A, B, \ldots$.

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea
- You should better do $A, B, \quad A, B, \quad A, B, \quad A, B, \ldots$.
- Even better, randomize your run order. You should flip a coin for each configuration and start with A on head and with B on tail...

  $$A, B, \quad B, A, \quad B, A, \quad A, B, \ldots.$$

  With such design, you will even be able to check whether being the first alternative to run changes something or not

Comparing Two Alternatives (Blocking + Randomization)

- When comparing A and B for different settings, doing A, A, A, A, A, A and then B, B, B, B, B, B is a bad idea
- You should better do $A, B, \quad A, B, \quad A, B, \quad A, B, \ldots$.
- Even better, randomize your run order. You should flip a coin for each configuration and start with A on head and with B on tail...

  $$A, B, \quad B, A, \quad B, A, \quad A, B, \ldots.$$

  With such design, you will even be able to check whether being the first alternative to run changes something or not
- Each configuration you test should be run on different machines
  You should record as much information as you can on how the experiments was performed

Experimental Design

There are two key concepts:

replication and randomization

You replicate to increase reliability. You randomize to reduce bias.

If you replicate thoroughly and randomize properly, you will not go far wrong.

Experimental Design

There are two key concepts:

replication and randomization

You replicate to increase reliability. You randomize to reduce bias.

If you replicate thoroughly and randomize properly, you will not go far wrong.

> *It doesn't matter if you cannot do your own advanced statistical analysis. If you designed your experiments properly, you may be able to find somebody to help you with the statistics.*
>
> *If your experiments is not properly designed, then no matter how good you are at statistics, you experimental effort will have been wasted.*

No amount of high-powered statistical analysis can turn a bad experiment into a good one.

Other important concepts:

- Pseudo-replication
- Experimental vs. observational data

Replication vs. Pseudo-replication

Measuring the same configuration several times is not replication. It's pseudo-replication and is generally biased

Instead, test other configurations (with a good randomization)

In case of pseudo-replication, here is what you can do:

- average away the pseudo-replication and carry out your statistical analysis on the means
- carry out separate analysis for each time period
- use proper time series analysis

Experimental data vs. Observational data

You need a good blend of observation, theory and experiments

- Many scientific experiments appear to be carried out with no hypothesis in mind at all, but simply to see what happens.
- This may be OK in the early stages but drawing conclusions on such observations is difficult (large number of equally plausible explanations; without testable prediction no experimental ingenuity; . . .).

Experimental data vs. Observational data

You need a good blend of observation, theory and experiments

- Many scientific experiments appear to be carried out with no hypothesis in mind at all, but simply to see what happens.
- This may be OK in the early stages but drawing conclusions on such observations is difficult (large number of equally plausible explanations; without testable prediction no experimental ingenuity; . . .).

Strong inference Essential steps:

1. Formulate a clear hypothesis
2. Devise an acceptable test

Weak inference It would be silly to disregard all observational data that do not come from designed experiments. Often, they are the only we have (e.g. the trace of a system).
But we need to keep the limitations of such data in mind. It is possible to use it to derive hypothesis but not to test hypothesis (i.e., claim facts).

Correlation and Causation

Let me illustrate this inference story with a few examples.
It may be the case that two random variables X and Y are dependent

- E.g., Let's pick a student at random and measure its *TimeSpentStudying* and its *TestScore*
  - In most cases, studying more should improve your test score ☺

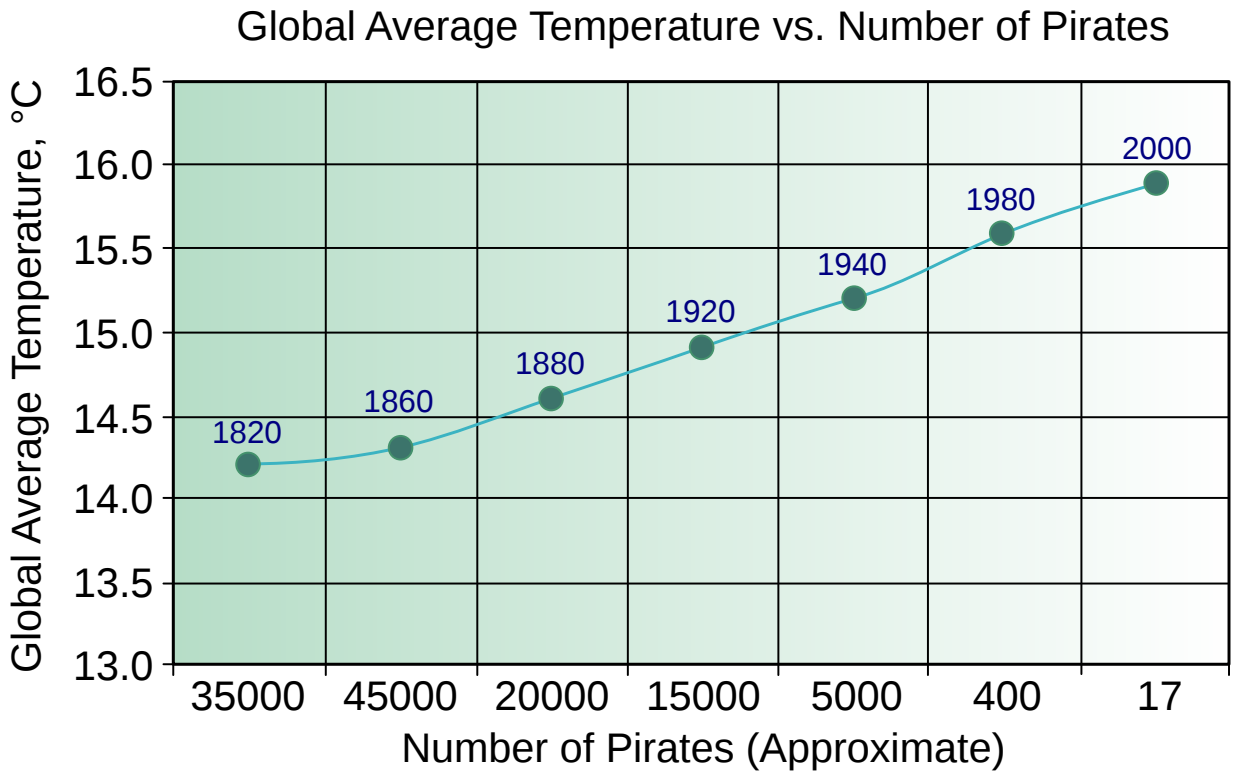
The correlation of two variables X and Y is defined as:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- The correlation is symmetrical ($\text{corr}(X, Y) = \text{corr}(Y, X)$)
- The correlation is in $[-1, 1]$
- $\text{corr}(Y, X) = 1$ or $-1 \Rightarrow$ perfectly linear relationship
- X independent of $Y \Rightarrow \text{corr}(X, Y) = 0$
- Y grows when X grows $\Rightarrow \text{corr}(X, Y) > 0$

It is thus very tempting to use sample correlation as a way of knowing whether some variables are dependant

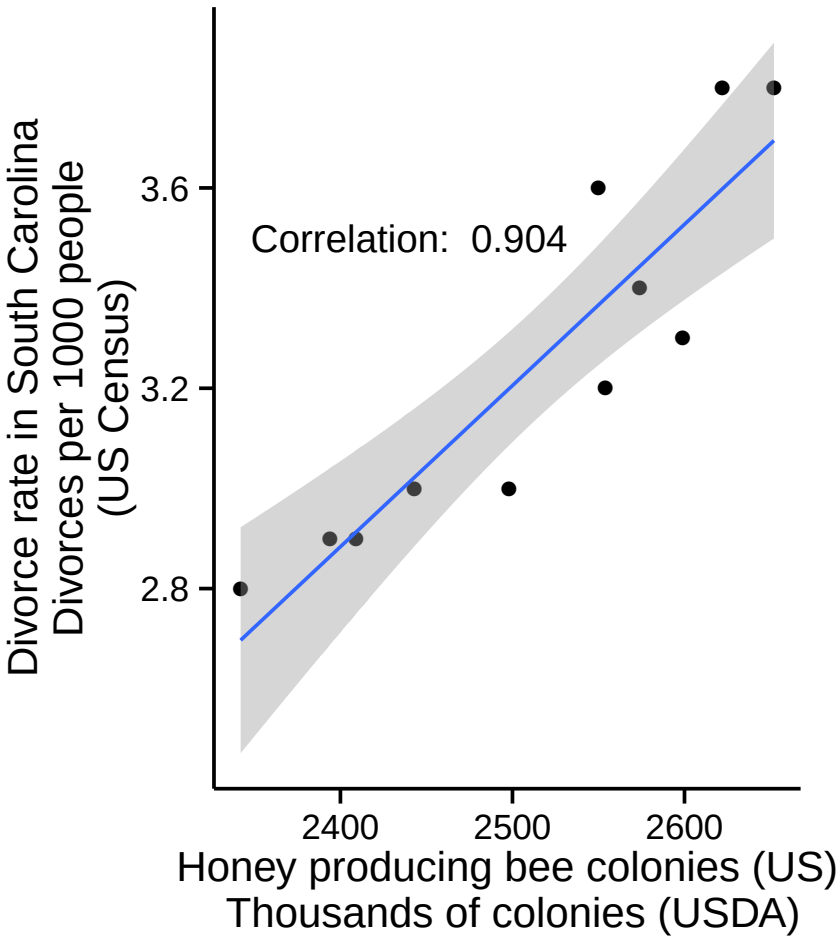
Correlation does not imply Causation

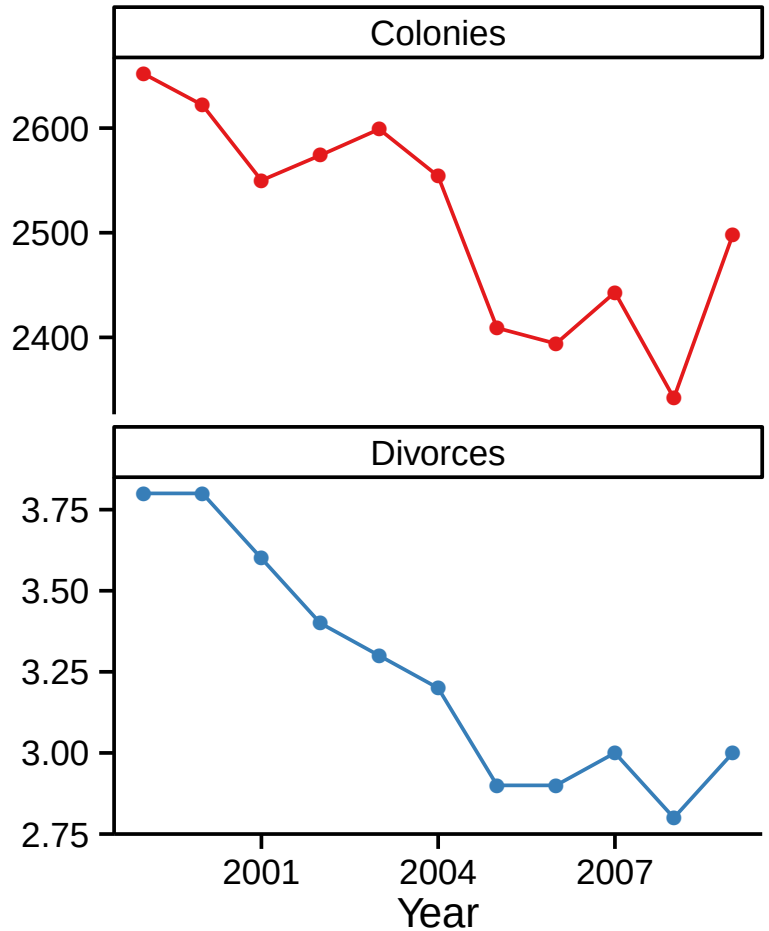


Mikhail Ryazanov (talk) - PiratesVsTemp.svg.
Licensed under CC BY-SA 3.0 via Wikimedia Commons

- 2 variables peuvent être fortement correlées à une troisième (e.g., year)
- Btw, what is wrong with this figure? 😉

Observational vs. Experimental Data Illustration



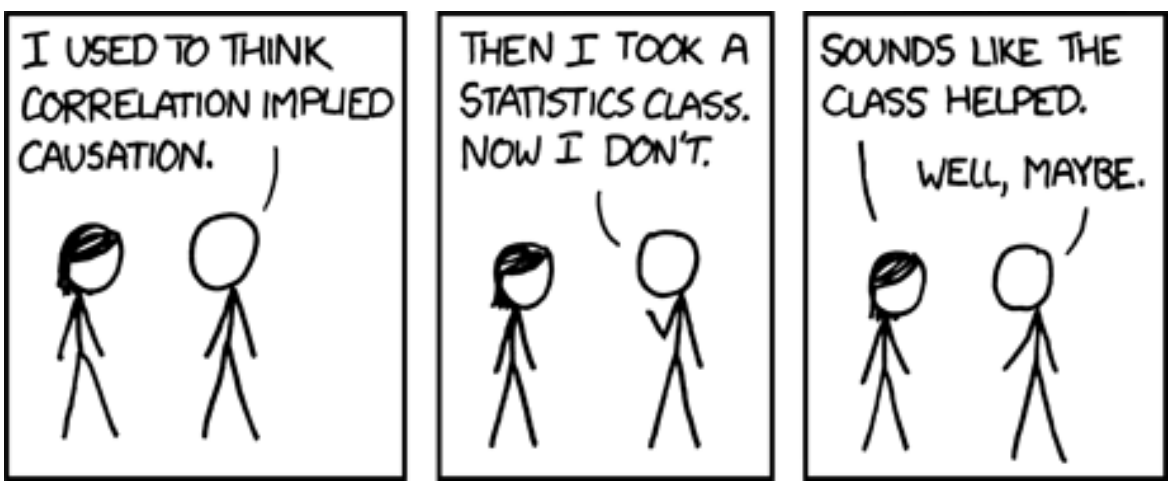


Source: *Spurious correlations*. For the good of the US society, we should try to get rid of honey bees 😉

Correlation does not imply Causation

For any two correlated events, A and B, the following relationships are possible:

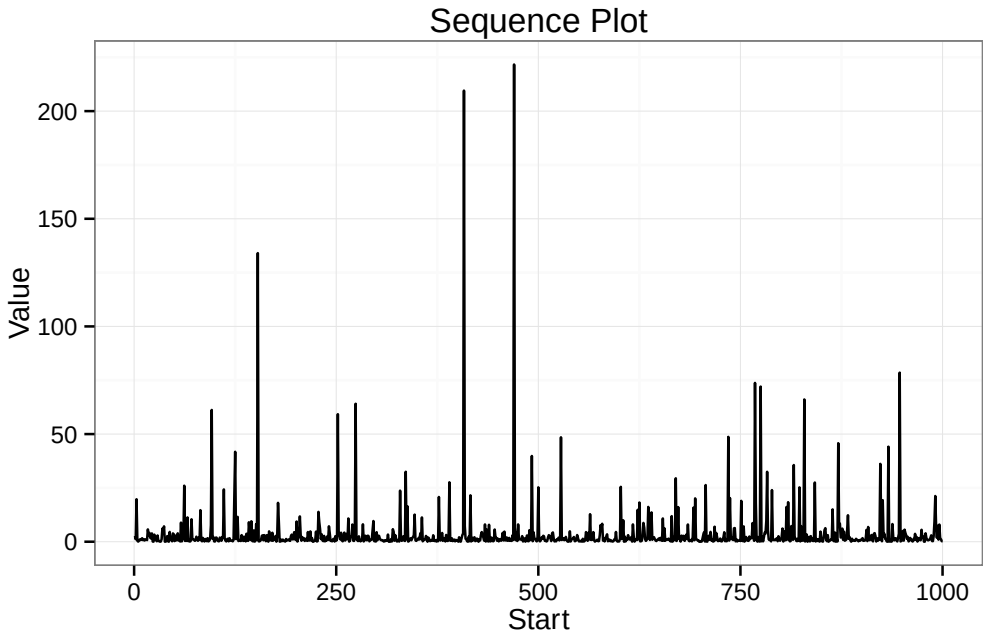
- A causes B (direct causation) 
- A causes B and B causes A (bidirectional or cyclic causation) 
- A causes C which causes B (indirect causation) 
- B causes A; (reverse causation) 
- A and B are consequences of a common cause, but do not cause each other 
- There is no connection between A and B; it is a coincidence 
  - But designed experiments can help you ruling this option out



Outline

Abnormal measurements

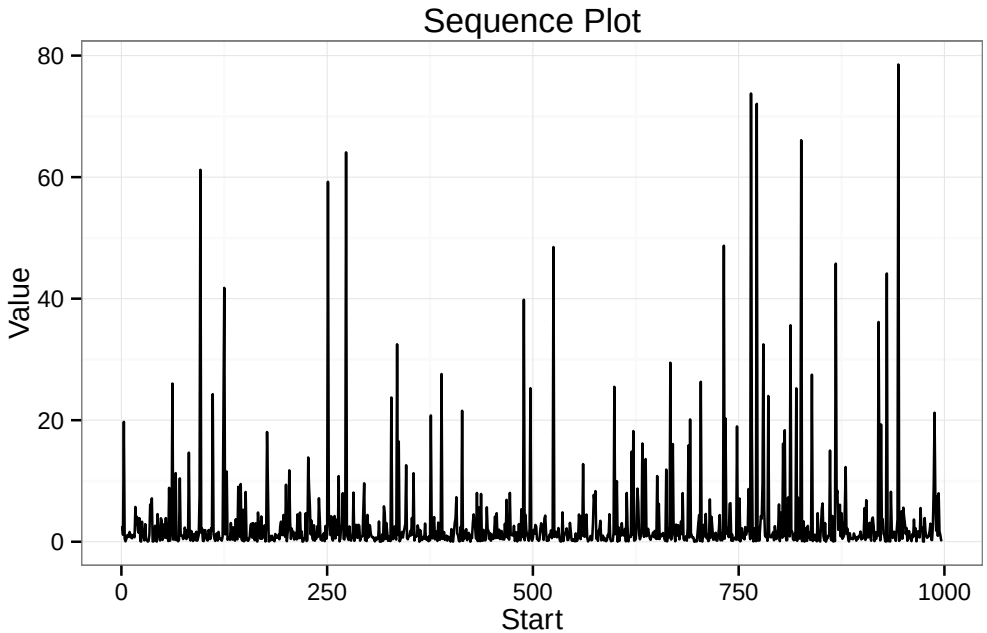
- Rare events: interpretation
- Get rid of it using e.g., quantiles:
  - What is the good rejection rate? E.g., "above Q3 + 1.5 x (IQR)" (boxplot, Tukey, 1977), i.e., above $\mu + 2\sigma$ for a normal distribution?



- A threshold value: what is the right threshold?
  - Reject values larger than 100 $\rightsquigarrow$ .6% of rejection
  - Reject values larger than 50 $\rightsquigarrow$ 1% of rejection
  - Reject values larger than 10 $\rightsquigarrow$ 6% of rejection

Abnormal measurements

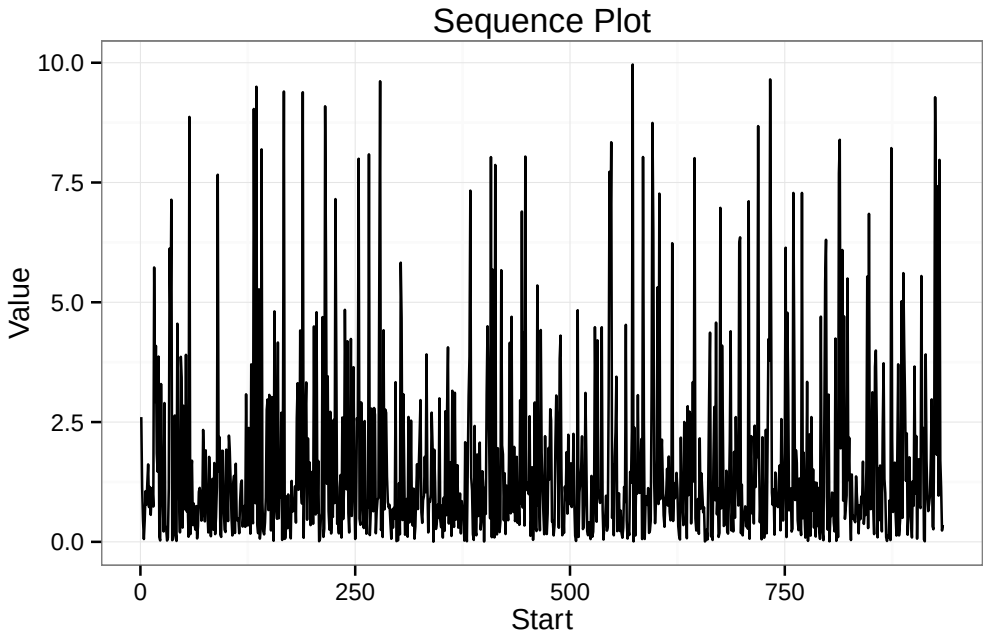
- Rare events: interpretation
- Get rid of it using e.g., quantiles:
  - What is the good rejection rate? E.g., "above Q3 + 1.5 x (IQR)" (boxplot, Tukey, 1977), i.e., above $\mu + 2\sigma$ for a normal distribution?



- A threshold value: what is the right threshold?
  - Reject values larger than 100 $\rightsquigarrow$ .6% of rejection
  - Reject values larger than 50 $\rightsquigarrow$ 1% of rejection
  - Reject values larger than 10 $\rightsquigarrow$ 6% of rejection

Abnormal measurements

- Rare events: interpretation
- Get rid of it using e.g., quantiles:
  - What is the good rejection rate? E.g., "above Q3 + 1.5 x (IQR)" (boxplot, Tukey, 1977), i.e., above $\mu + 2\sigma$ for a normal distribution?

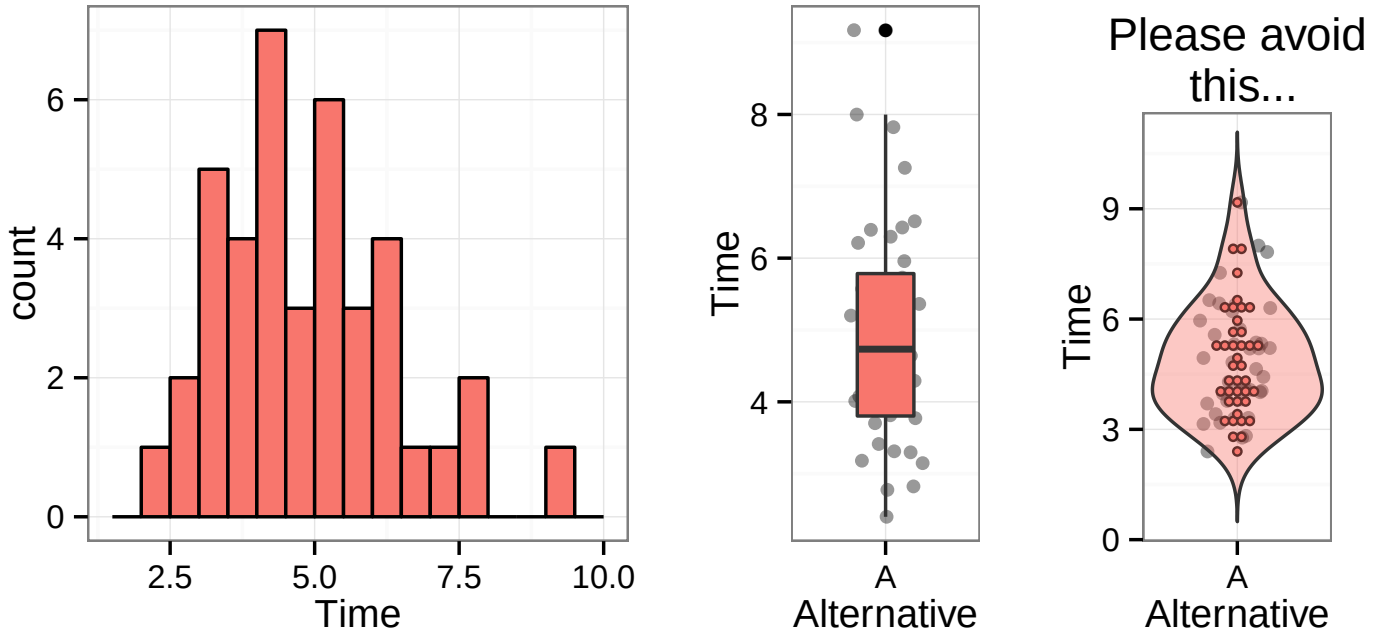


- A threshold value: what is the right threshold?
  - Reject values larger than 100 $\rightsquigarrow$ .6% of rejection
  - Reject values larger than 50 $\rightsquigarrow$ 1% of rejection
  - Reject values larger than 10 $\rightsquigarrow$ 6% of rejection

Actually, I generated these samples using the Cauchy distribution, which is pathological for most idea you'll come up with 😉
There is no mathematical definition of what constitutes an outlier. It's related to the experimenter's interpretation and is subjective...

Outline

Summarizing the distribution

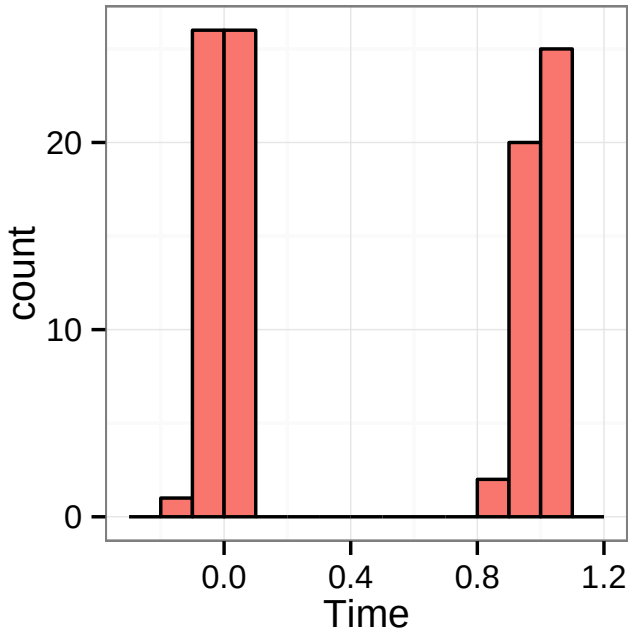


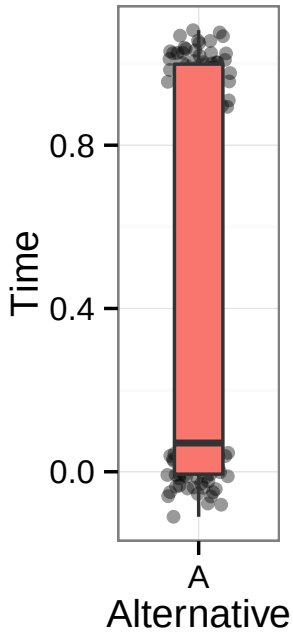
What is the shape of the histogram:

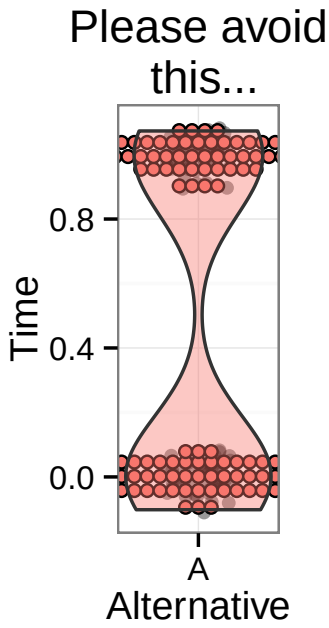
- Uni/multi-modal:
  - If uni-modal, summarize with central tendancy (mean, mode, median)
  - Symmetrical or not ($\rightsquigarrow$ skewness)
  - Flat of not ($\rightsquigarrow$ kurtosis)

If uni-modal you can go for a boxplot but avoid other fancy plots unless you know what you do...

Summarizing the distribution







What is the shape of the histogram:

- Uni/multi-modal:
  - If uni-modal, summarize with central tendancy (mean, mode, median)
  - Symmetrical or not ($\rightsquigarrow$ skewness)
  - Flat of not ($\rightsquigarrow$ kurtosis)

If uni-modal you can go for a boxplot but avoid other fancy plots unless you know what you do...

Outline

Biased and unbiased estimators...

Expected value the *sample mean* is unbiased but is "sensitive" to outliers. This is not an excuse for estimating something else! Furthermore, there is an easy way to compute confidence intervals.

Mode the *naive estimate* is unstable and depends on the histogram's bin width. Still ongoing research on this:

> E.g., Bickela and Frühwirthb, *On a fast, robust estimator of the mode: Comparisons to other robust estimators with applications*, Computational Statistics & Data Analysis 2006

Median the *sample median* is robust to outliers but it is quite sensitive to discrete distributions... There exists other more involved estimators but is median really what you want to estimate?

Minimum and Maximum the *sample minimum* is always too large, hence it is biased...

Variance `var` is a unbiased estimator or variance but `sd` is a biased estimator of standard deviation. Unbiasing depends on the distribution so just overestimate...

Outline

Tests

We have seen how to build estimators of specific characteristics of f_X based on observations of X

- Having an estimator is worth only if you can provide confidence on this estimation

Estimates can be used to test hypothesis

- We have seen how we could test whether $\mu_A < \mu_B$ from estimates of μ_A and μ_B

But there may be other more efficient ways to test such hypothesis

- if you know observations are paired
- if you know something about the underlying distribution

Other kind of complex hypothesis may tested

- median(A) = median(B)
- A and B follow the same distribution

We could give a whole lecture on this topic... Only use the tests you truly understand

Outline

Roadmap for a good data analysis (Jain)

1. Plot the sample (various representations)
2. Describe the results (data analysis)
3. Preliminary processing : remove or flag outliers, estimate or flag missing values
4. Propose a stochastic model. Establish the hypothesis: independence (time correlation, auto-correlation), stationarity, same probability law
5. Summarize data by a histogram
6. Comment the shape (modal/skewness/flatness/...)
7. Estimate the central tendency of the sample : choose the central index
8. Estimate the accuracy of the result (confidence intervals)
9. Propose a visualization