How to evaluate your z-score?

Panagiotis Tsiamyrtzis [1] Frederic Sobas [2]

[1]Dept. of Statistics, Athens University of Economics and Business

pt@aueb.gr

[2]Hospices Civils de Lyon

frederic.sobas@chu-lyon.fr

Leiden, 9 November 2018

Contents

- Understanding z-scores
- Multiple z-score analysis
- Evaluating z-scores in pairs
- Evaluating z-score history

Understanding z-scores

Normal Distribution

- The z-score has its origin to the Normal distribution, a symmetric bell shaped curve (also known as Gaussian) that plays a major role in statistics. It is denoted as $X \sim N\left(\mu, \sigma^2\right)$, with μ being its mean and σ its standard deviation.

Normal Distribution

- The z-score has its origin to the Normal distribution, a symmetric bell shaped curve (also known as Gaussian) that plays a major role in statistics. It is denoted as $X \sim N\left(\mu, \sigma^2\right)$, with μ being its mean and σ its standard deviation.

- For $x \in (-\infty, +\infty)$

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$E[X] = \mu \qquad Var[X] = \sigma^2$$

A tremendous formula capable to describe random phenomena more often than any other existing distribution.

Normal Distribution

- The z-score has its origin to the Normal distribution, a symmetric bell shaped curve (also known as Gaussian) that plays a major role in statistics. It is denoted as $X \sim N\left(\mu, \sigma^2\right)$, with μ being its mean and σ its standard deviation.

- For $x \in (-\infty, +\infty)$

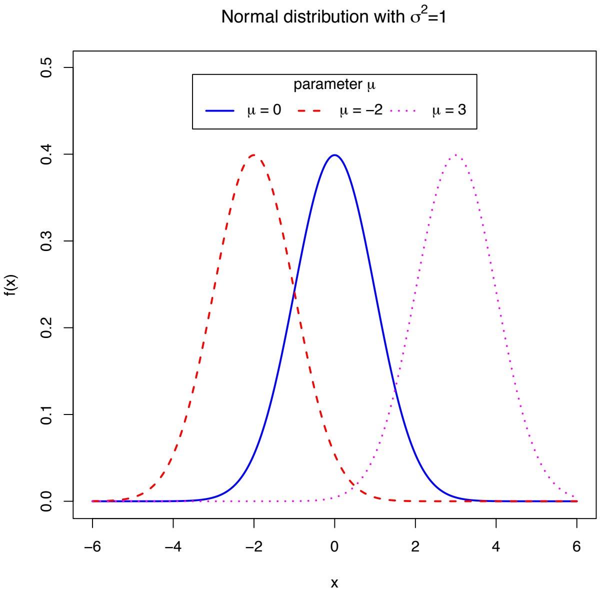
$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$E[X] = \mu \qquad Var[X] = \sigma^2$$

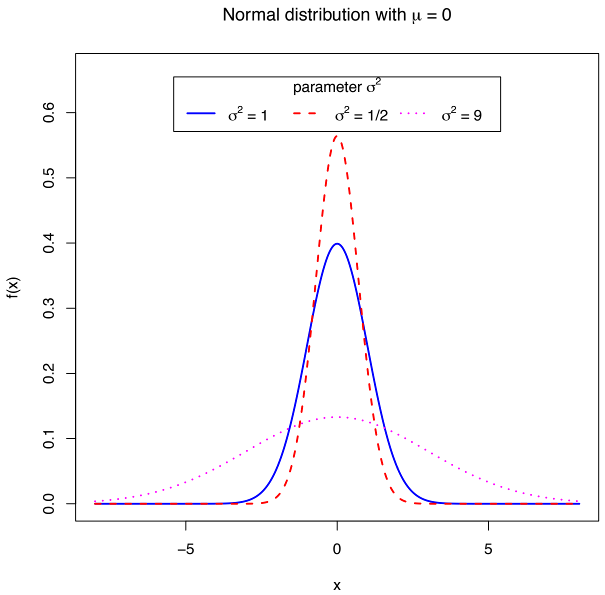
  A tremendous formula capable to describe random phenomena more often than any other existing distribution.

- The $Z \sim N(0, 1)$ is known as the "Standard" Normal Distribution and is related to the concept of z-scores

Normal Distribution



Normal Distribution



Normal Distribution

Properties:

- The $Z \sim N(0,1)$ is called the standard Normal distribution and for any other Normal, $X \sim N\left(\mu, \sigma^2\right)$ we have:

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1) \qquad \text{and} \qquad X = \mu + \sigma Z$$

Normal Distribution

Properties:

- The $Z \sim N(0,1)$ is called the standard Normal distribution and for any other Normal, $X \sim N\left(\mu, \sigma^2\right)$ we have:

$$Z = \frac{X - \mu}{\sigma} \sim N(0,1) \qquad \text{and} \qquad X = \mu + \sigma Z$$

- The Normal is well known for its $68 - 95 - 99.7$ rule, i.e.:

$$\begin{aligned} P(|X-\mu| \leq \sigma) &= P(|Z| \leq 1) &= 0.6826 \\ P(|X-\mu| \leq 2\sigma) &= P(|Z| \leq 2) &= 0.9544 \\ P(|X-\mu| \leq 3\sigma) &= P(|Z| \leq 3) &= 0.9973 \end{aligned}$$

Normal Distribution

Properties:

- The $Z \sim N(0,1)$ is called the standard Normal distribution and for any other Normal, $X \sim N\left(\mu, \sigma^2\right)$ we have:

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1) \qquad \text{and} \qquad X = \mu + \sigma Z$$

- The Normal is well known for its $68 - 95 - 99.7$ rule, i.e.:

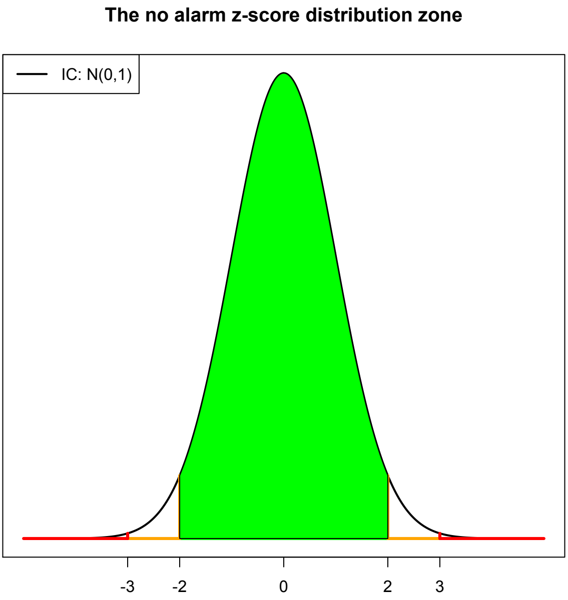
$$\begin{aligned} P(|X-\mu| \leq \sigma) &= P(|Z| \leq 1) &= 0.6826 \\ P(|X-\mu| \leq 2\sigma) &= P(|Z| \leq 2) &= 0.9544 \\ P(|X-\mu| \leq 3\sigma) &= P(|Z| \leq 3) &= 0.9973 \end{aligned}$$

- The above property establishes the well known alarming zones of z-scores (ISO 17043 recommended by the ISO 15189 norm). Specifically:

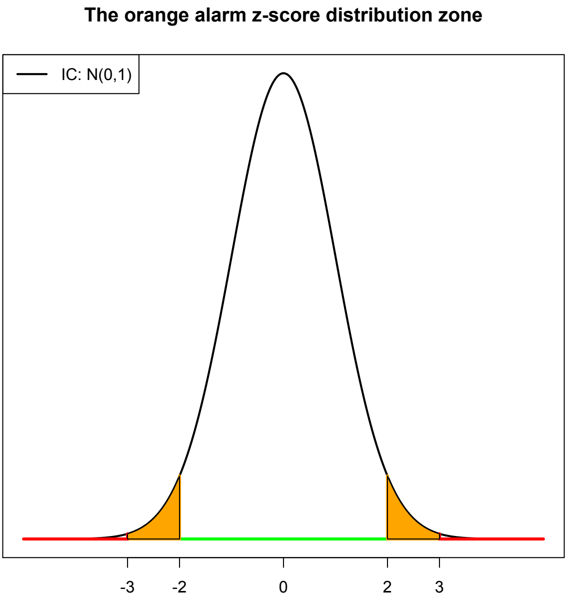
  Orange alarm: when $-3 <$ z-score ≤ -2 or $2 \leq$ z-score < 3

  Red alarm: when z-score ≤ -3 or z-score ≥ 3

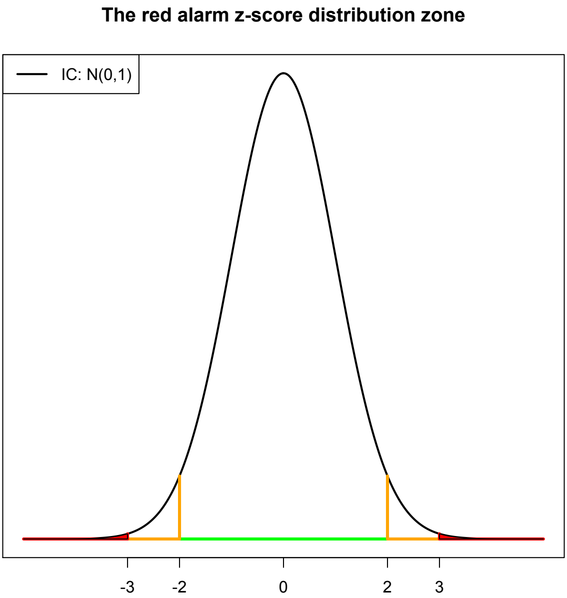
The Alarm Zones



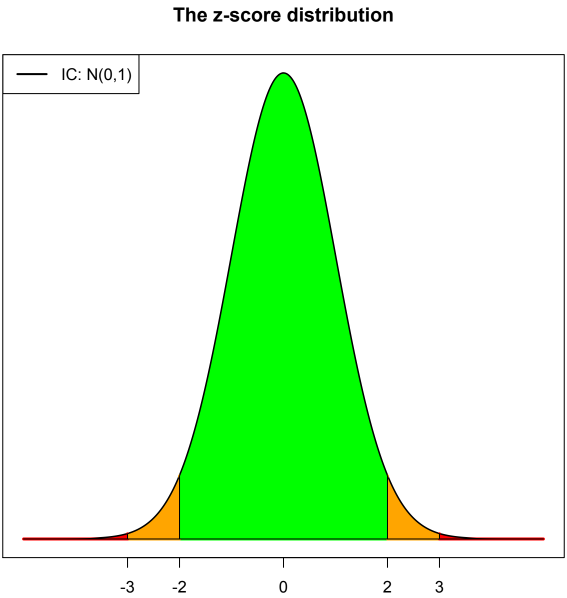
The Alarm Zones



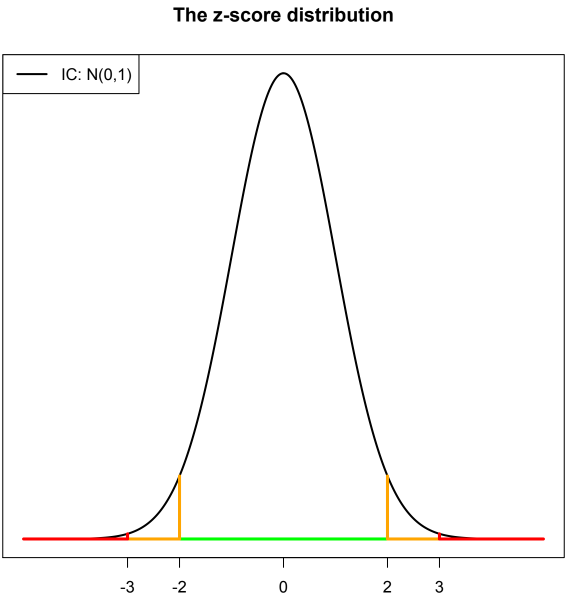
The Alarm Zones



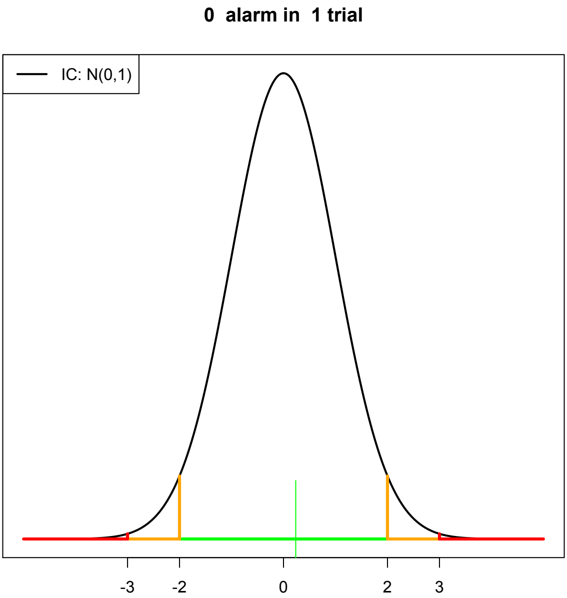
The Alarm Zones



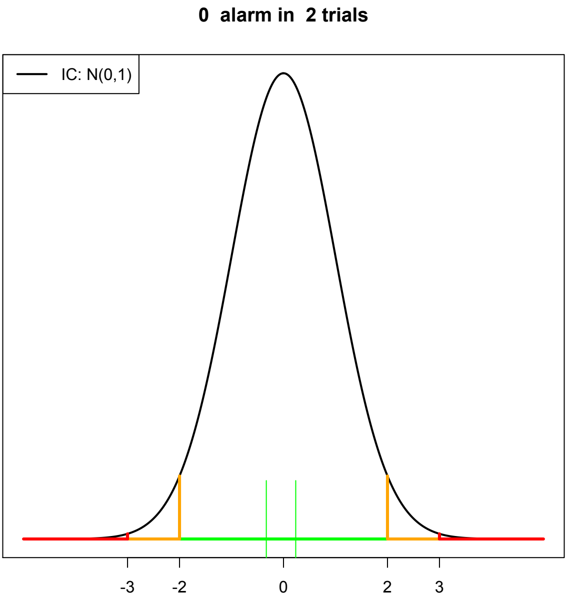
Simulating IC performance



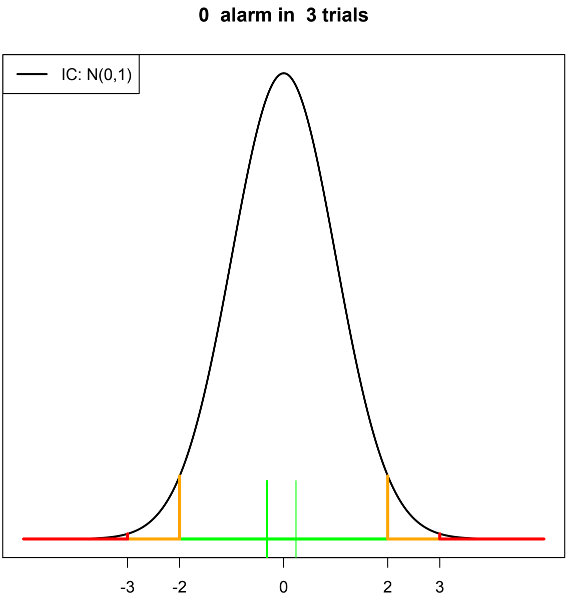
Simulating IC performance



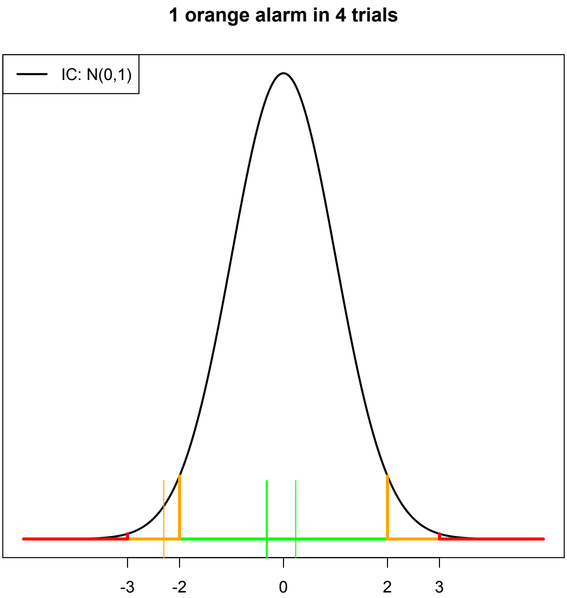
Simulating IC performance



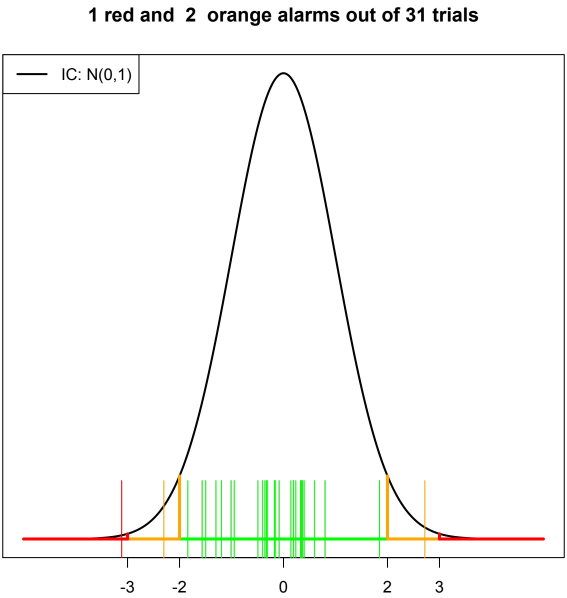
Simulating IC performance



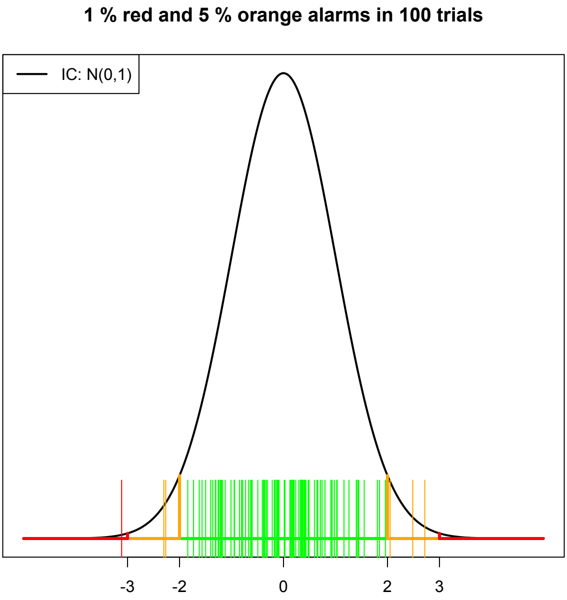
Simulating IC performance



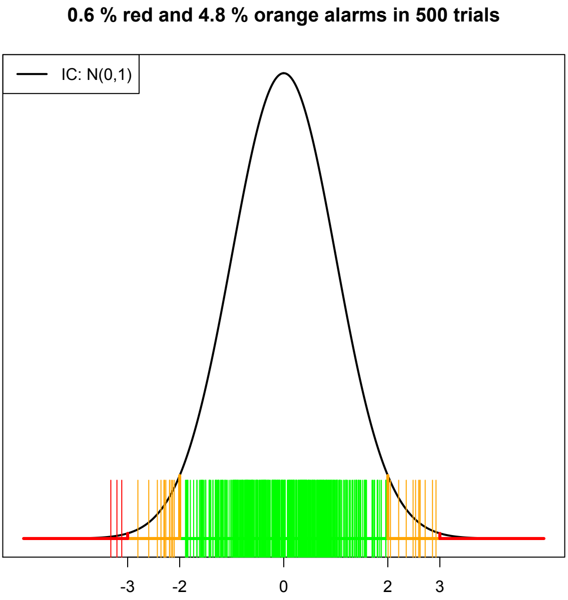
Simulating IC performance



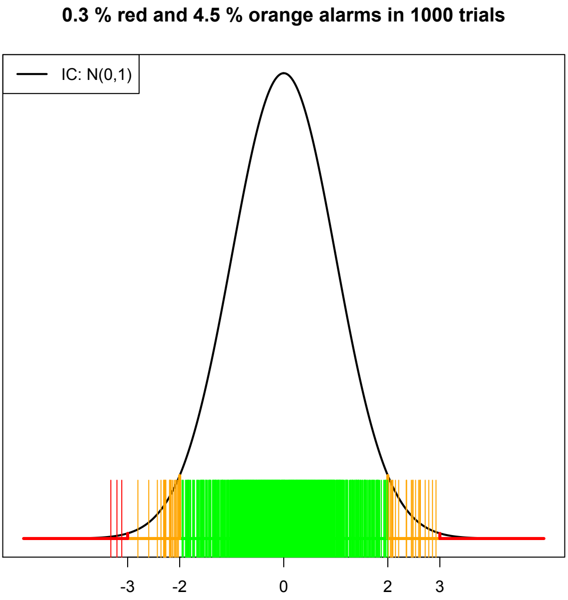
Simulating IC performance



Simulating IC performance



Simulating IC performance



Simulating OOC performance

- So if the lab is under the In Control (IC) status there is 0.27% chance to get a red alarm and 4.28% chance to get an orange alarm.

Simulating OOC performance

- So if the lab is under the In Control (IC) status there is 0.27% chance to get a red alarm and 4.28% chance to get an orange alarm.
- What if though the lab is not "well aligned" with the IC distribution established by the EQA organization?

Simulating OOC performance

- So if the lab is under the In Control (IC) status there is 0.27% chance to get a red alarm and 4.28% chance to get an orange alarm.
- What if though the lab is not "well aligned" with the IC distribution established by the EQA organization?
- A lab performing under Out Of Control (OOC) conditions would have an elevated alarm rate. Both the **magnitude** and the **sign** of the alarming z-scores can offer some valuable information of what is the issue.

Simulating OOC performance

- So if the lab is under the In Control (IC) status there is 0.27% chance to get a red alarm and 4.28% chance to get an orange alarm.
- What if though the lab is not "well aligned" with the IC distribution established by the EQA organization?
- A lab performing under Out Of Control (OOC) conditions would have an elevated alarm rate. Both the **magnitude** and the **sign** of the alarming z-scores can offer some valuable information of what is the issue.
- The two major OOC issues are related to:

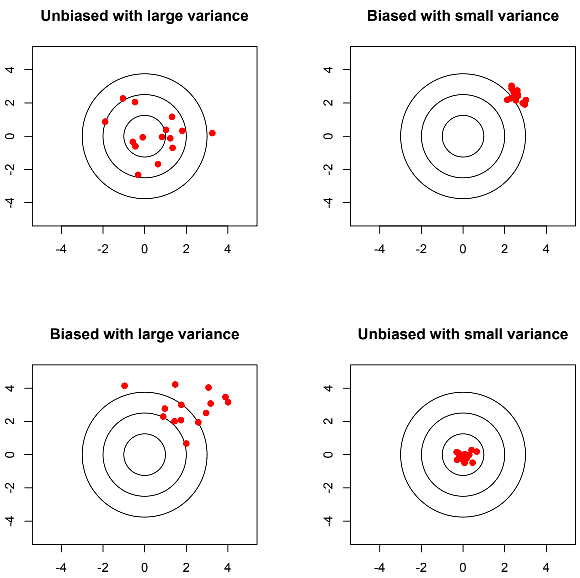
Simulating OOC performance

- So if the lab is under the In Control (IC) status there is 0.27% chance to get a red alarm and 4.28% chance to get an orange alarm.
- What if though the lab is not “well aligned” with the IC distribution established by the EQA organization?
- A lab performing under Out Of Control (OOC) conditions would have an elevated alarm rate. Both the **magnitude** and the **sign** of the alarming z-scores can offer some valuable information of what is the issue.
- The two major OOC issues are related to:
  - **Bias**: how do we perform on average? (biased or unbiased?)

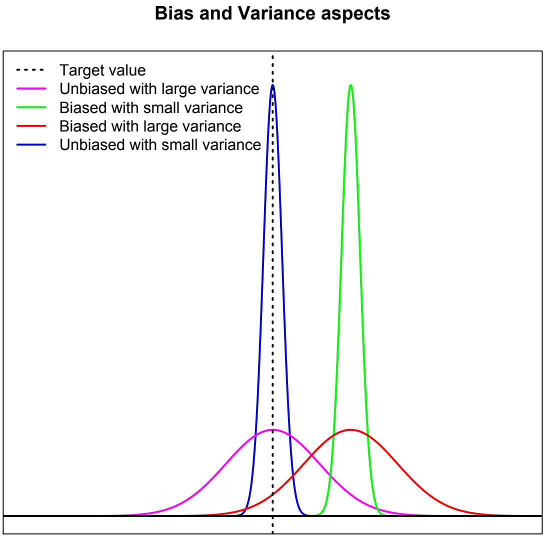
Simulating OOC performance

- So if the lab is under the In Control (IC) status there is 0.27% chance to get a red alarm and 4.28% chance to get an orange alarm.
- What if though the lab is not "well aligned" with the IC distribution established by the EQA organization?
- A lab performing under Out Of Control (OOC) conditions would have an elevated alarm rate. Both the **magnitude** and the **sign** of the alarming z-scores can offer some valuable information of what is the issue.
- The two major OOC issues are related to:
  - **Bias**: how do we perform on average? (biased or unbiased?)
  - **Uncertainty**: how variable (uncertain) are we?

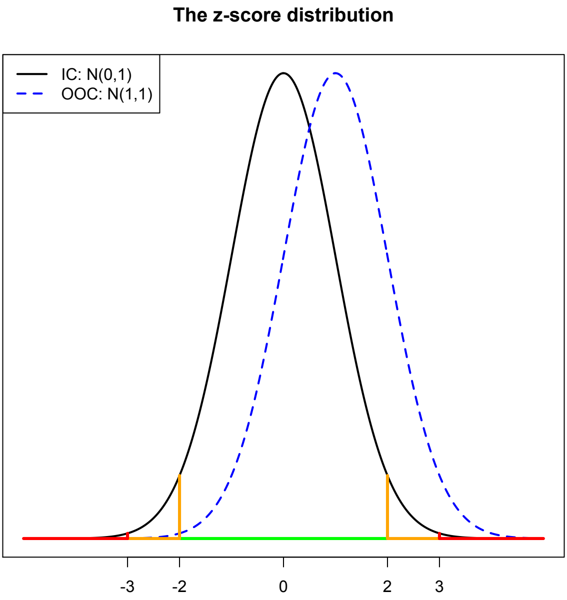
Bias and Uncertainty



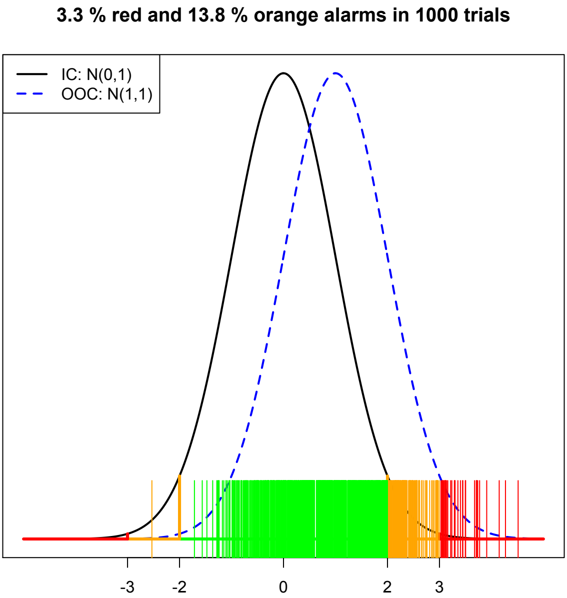
Bias and Uncertainty



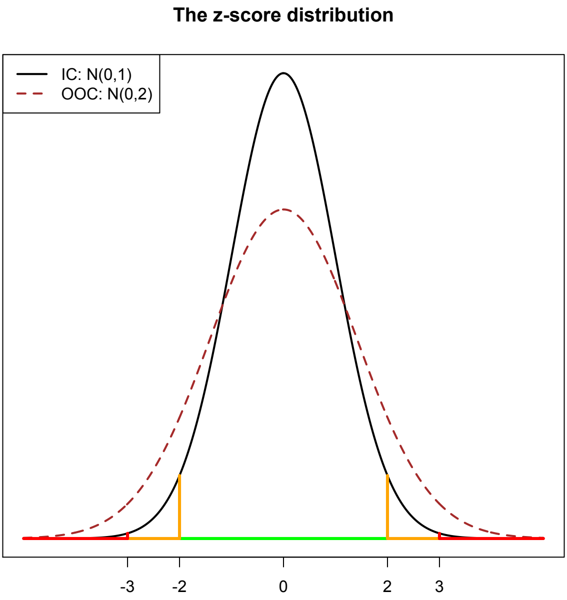
Simulating OOC (bias) performance



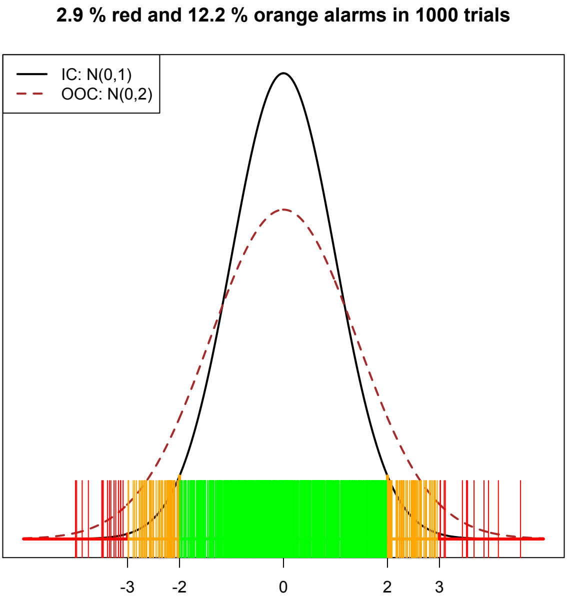
Simulating OOC (bias) performance



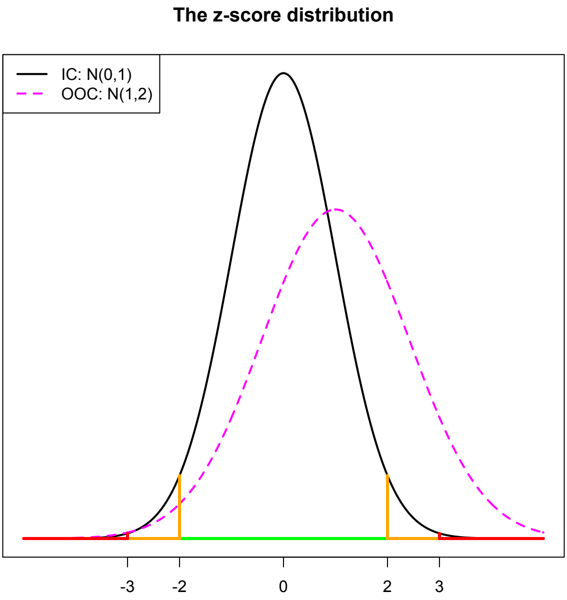
Simulating OOC (variance) performance



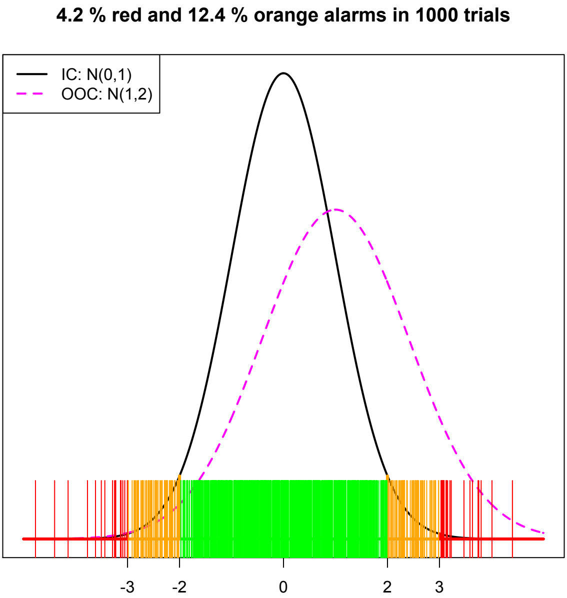
Simulating OOC (variance) performance



Simulating OOC (bias & variance) performance



Simulating OOC (bias & variance) performance



Multiple z-score analysis

Multiple z-score analysis

- The type I error refers to the case where we falsely raise an alarm (while in reality everything works properly). As we saw earlier this is $\alpha = 0.0428$ or 0.0027 for the orange or red alarm respectively.

Multiple z-score analysis

- The type I error refers to the case where we falsely raise an alarm (while in reality everything works properly). As we saw earlier this is $\alpha = 0.0428$ or 0.0027 for the orange or red alarm respectively.

- But what if we compare a small lab with say a single automate, against a bigger lab with multiple automates? It is natural to expect that the more the z-scores the more likely to get a false alarm.

Multiple z-score analysis

- The type I error refers to the case where we falsely raise an alarm (while in reality everything works properly). As we saw earlier this is $\alpha = 0.0428$ or 0.0027 for the orange or red alarm respectively.
- But what if we compare a small lab with say a single automate, against a bigger lab with multiple automates? It is natural to expect that the more the z-scores the more likely to get a false alarm.
- In statistics, this is called the "multiple comparisons" problem and is known to inflate the false alarm **counts**.

Multiple z-score analysis

- The type I error refers to the case where we falsely raise an alarm (while in reality everything works properly). As we saw earlier this is $\alpha = 0.0428$ or 0.0027 for the orange or red alarm respectively.

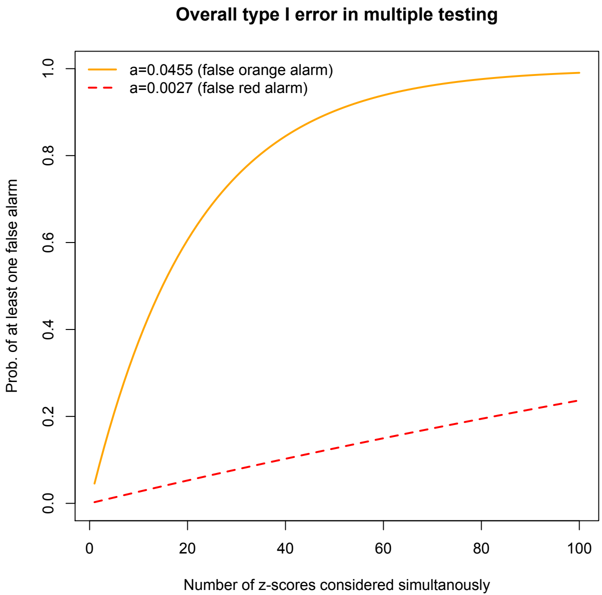
- But what if we compare a small lab with say a single automate, against a bigger lab with multiple automates? It is natural to expect that the more the z-scores the more likely to get a false alarm.

- In statistics, this is called the "multiple comparisons" problem and is known to inflate the false alarm **counts**.

- If we will consider that the tests are independent and we call Y the number of false alarms that we have in N tests (z-scores of automates), we get $Y \sim B(N, \alpha)$, a Binomial distribution. Then, the probability of at least one false alarm in N tests will be:

$$P(Y \geq 1) = 1 - P(Y = 0) = 1 - (1 - \alpha)^N$$

Multiple z-score analysis



Multiple z-score analysis

The probability of a lab to have at least one orange/red false alarm as a function of the number of automates tested is:

Number of # automates	Orange Zone False Alarm Prob	Red Zone False Alarm Prob
1	4.3%	0.3%
2	8.4%	0.5%
3	12.3%	0.8%
4	16.1%	1.1%
5	19.6%	1.3%
6	23.1%	1.6%
7	26.4%	1.9%
8	29.5%	2.1%
9	32.5%	2.4%
10	35.4%	2.7%

Bonferroni Correction

- Now lets assume that in lab/multicenter with several automates, the EQA organization raises an alarm to the **whole** lab, if **at least one** of the automates is giving an alarm.

Bonferroni Correction

- Now lets assume that in lab/multicenter with several automates, the EQA organization raises an alarm to the **whole** lab, if **at least one** of the automates is giving an alarm.
- Based on the previous plot, the higher the number of z-scores the more likely to get a false alarm compared to a lab with less z-scores.

Bonferroni Correction

- Now lets assume that in lab/multicenter with several automates, the EQA organization raises an alarm to the **whole** lab, if **at least one** of the automates is giving an alarm.
- Based on the previous plot, the higher the number of z-scores the more likely to get a false alarm compared to a lab with less z-scores.
- So, when we compare two labs with different number of z-scores, how can we adjust the z-score alarm zones, so that the probability of false alarm is approximately the same in both labs?

Bonferroni Correction

- Now lets assume that in lab/multicenter with several automates, the EQA organization raises an alarm to the **whole** lab, if **at least one** of the automates is giving an alarm.
- Based on the previous plot, the higher the number of z-scores the more likely to get a false alarm compared to a lab with less z-scores.
- So, when we compare two labs with different number of z-scores, how can we adjust the z-score alarm zones, so that the probability of false alarm is approximately the same in both labs?
- In statistics when we perform multiple testing we can adjust the probability of type I error based on the number of tests performed. Assuming independence among the N tests performed, a popular adjustment is the **Bonferroni correction** where it adjusts the type I error by dividing α with the number of tests performed (i.e. α/N).

Bonferroni Correction

Fixing the orange & red based false alarm rates to be always at 0.0428 & 0.0027 respectively, the limits that we need to use, based on the number of z-scores (automates) we test are:

Number of # automates	Orange Zone Limits	Red Zone Limits
1	± 2.00	± 3.00
2	± 2.28	± 3.21
3	± 2.43	± 3.32
4	± 2.53	± 3.40
5	± 2.61	± 3.46
6	± 2.67	± 3.51
7	± 2.72	± 3.55
8	± 2.77	± 3.58
9	± 2.80	± 3.62
10	± 2.84	± 3.64

Bonferroni Correction

- Pay attention that the above proposal refers **not** to an individual automate but to the **whole** lab/multicenter, where overall gets an alarm, if **at least one** automate gets an alarm.

Bonferroni Correction

- Pay attention that the above proposal refers **not** to an individual automate but to the **whole** lab/multicenter, where overall gets an alarm, if **at least one** automate gets an alarm.

- Also we need to pay attention to the evolution in time of the z-scores for each automate, as consecutive alarms in an automate have negligible probability of being actually a false alarm.

Bonferroni Correction

- Pay attention that the above proposal refers **not** to an individual automate but to the **whole** lab/multicenter, where overall gets an alarm, if **at least one** automate gets an alarm.
- Also we need to pay attention to the evolution in time of the z-scores for each automate, as consecutive alarms in an automate have negligible probability of being actually a false alarm.
- Apart from Bonferroni, other corrections are available in the statistical literature, like:
  - Sidak Correction
  - Holm Bonferroni Correction
  - False Discovery Rate
  - ...

Evaluating z-scores in pairs

Evaluating z-scores in pairs

- For each z-score the $[-2, 2]$ and $[-3, 3]$ zones establish the alarm regions (orange and red respectively)

Evaluating z-scores in pairs

- For each z-score the $[-2, 2]$ and $[-3, 3]$ zones establish the alarm regions (orange and red respectively)
- When we study a pair of two z-scores it seems natural to extend the alarm zone the square regions $[-2, 2] \times [-2, 2]$ and $[-3, 3] \times [-3, 3]$.

Evaluating z-scores in pairs

- For each z-score the $[-2, 2]$ and $[-3, 3]$ zones establish the alarm regions (orange and red respectively)
- When we study a pair of two z-scores it seems natural to extend the alarm zone the square regions $[-2, 2] \times [-2, 2]$ and $[-3, 3] \times [-3, 3]$.
- From a statistical perspective though, this is suboptimal, not only in the case that the z-scores are correlated, but even when the two z-scores are independent!

Evaluating z-scores in pairs

- For each z-score the $[-2, 2]$ and $[-3, 3]$ zones establish the alarm regions (orange and red respectively)
- When we study a pair of two z-scores it seems natural to extend the alarm zone the square regions $[-2, 2] \times [-2, 2]$ and $[-3, 3] \times [-3, 3]$.
- From a statistical perspective though, this is suboptimal, not only in the case that the z-scores are correlated, but even when the two z-scores are independent!
- If Z_1 is **independent** of Z_2, then:

$$\begin{aligned} P((Z_1, Z_2) \in [-2, 2]^2) &= P(-2 \leq Z_1 \leq 2) \times P(-2 \leq Z_2 \leq 2) \\ &= 0.9545 \times 0.9545 = 0.9110 \end{aligned}$$

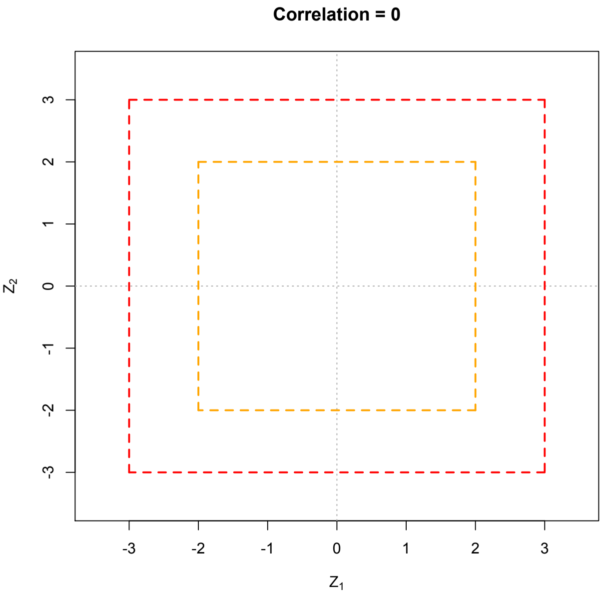
Evaluating z-scores in pairs

- For each z-score the $[-2, 2]$ and $[-3, 3]$ zones establish the alarm regions (orange and red respectively)
- When we study a pair of two z-scores it seems natural to extend the alarm zone the square regions $[-2, 2] \times [-2, 2]$ and $[-3, 3] \times [-3, 3]$.
- From a statistical perspective though, this is suboptimal, not only in the case that the z-scores are correlated, but even when the two z-scores are independent!
- If Z_1 is **independent** of Z_2, then:

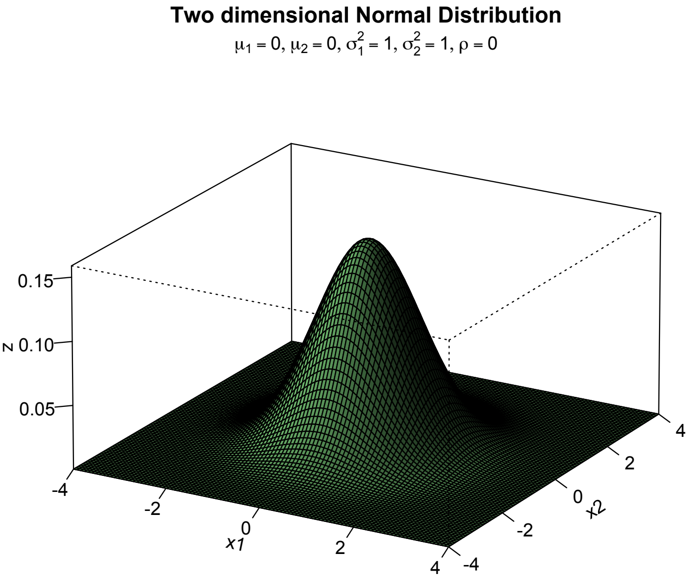
$$\begin{aligned} P((Z_1, Z_2) \in [-2, 2]^2) &= P(-2 \leq Z_1 \leq 2) \times P(-2 \leq Z_2 \leq 2) \\ &= 0.9545 \times 0.9545 = 0.9110 \end{aligned}$$

- For correlated variables things become a lot worst when we use the $[-2, 2]^2$ and $[-3, 3]^2$ boxes. One should make use of the Bivariate Normal distribution to model pairs of Z-scores.

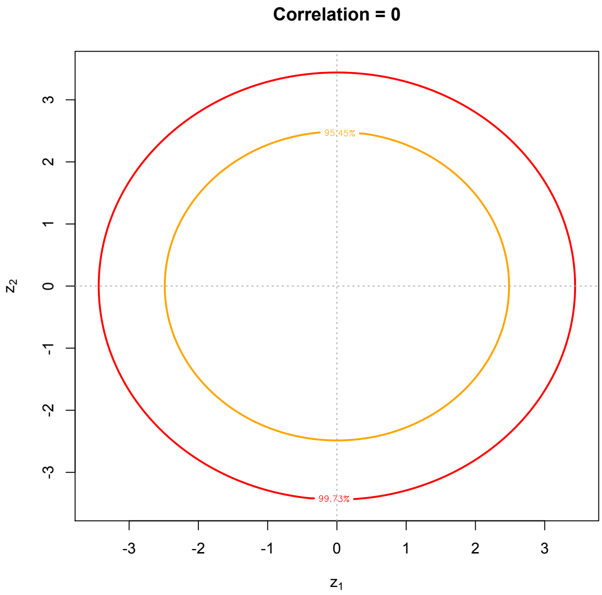
Evaluating independent z-scores in pairs



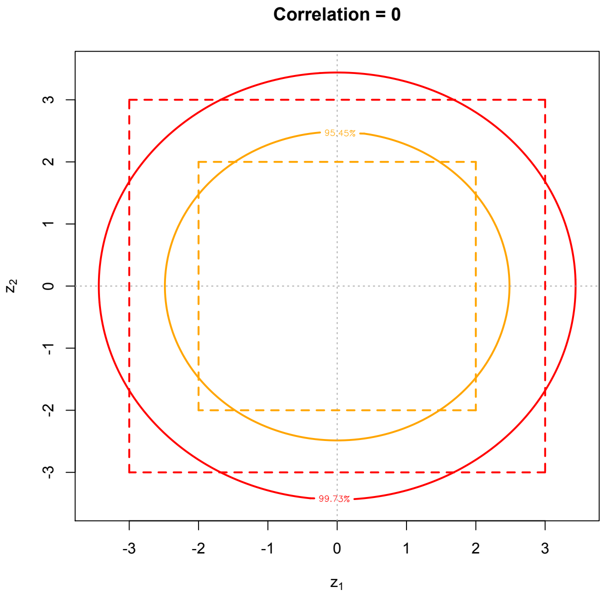
Bivariate Normal distribution (independent)



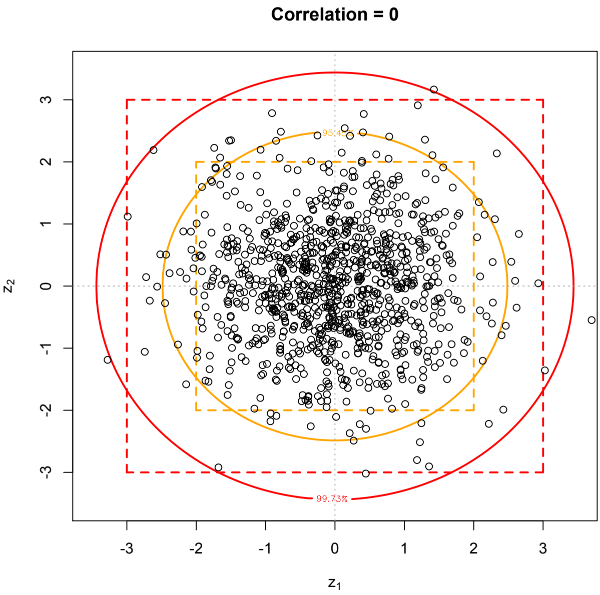
Evaluating independent z-scores in pairs



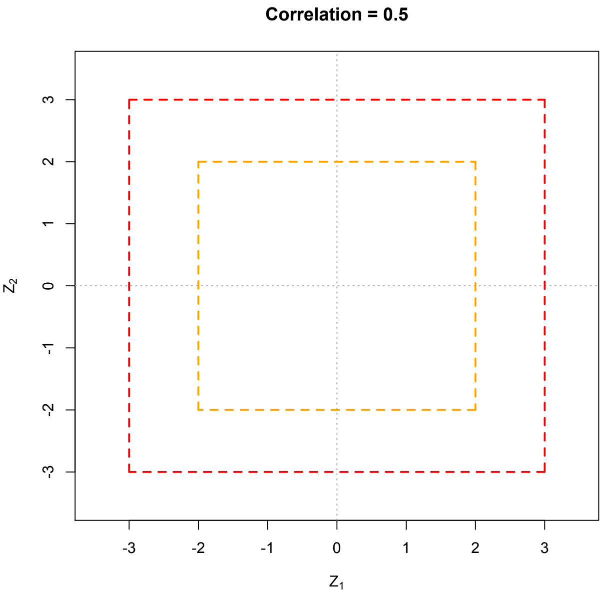
Evaluating independent z-scores in pairs



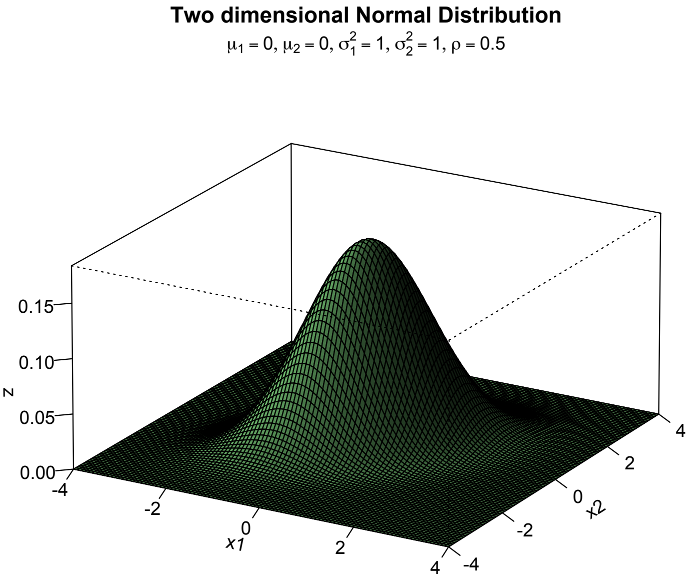
Evaluating correlated z-scores in pairs



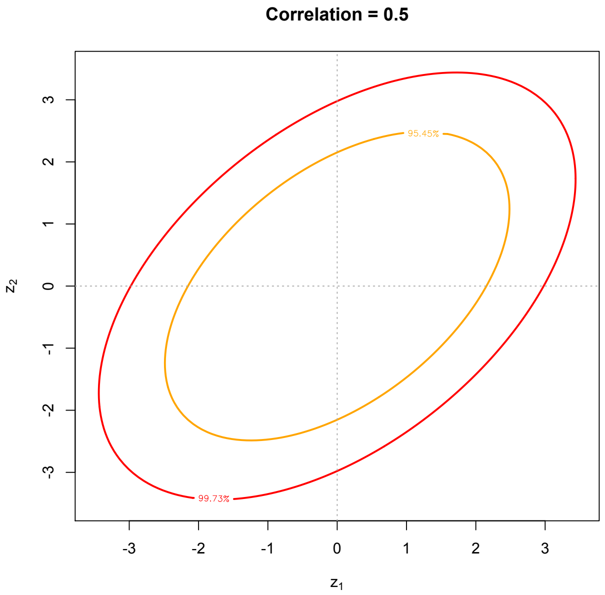
Evaluating independent z-scores in pairs



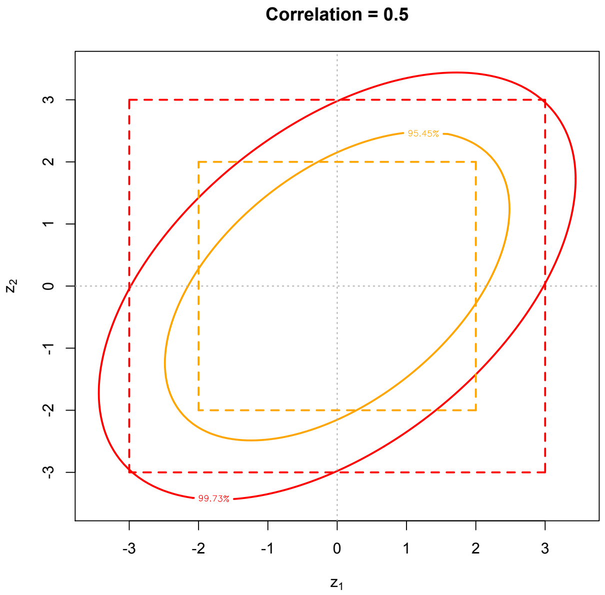
Bivariate Normal distribution (mild positive corr.)



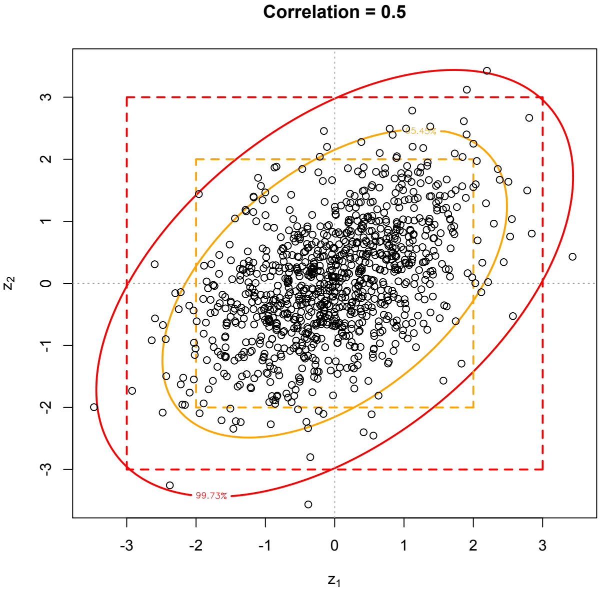
Evaluating independent z-scores in pairs



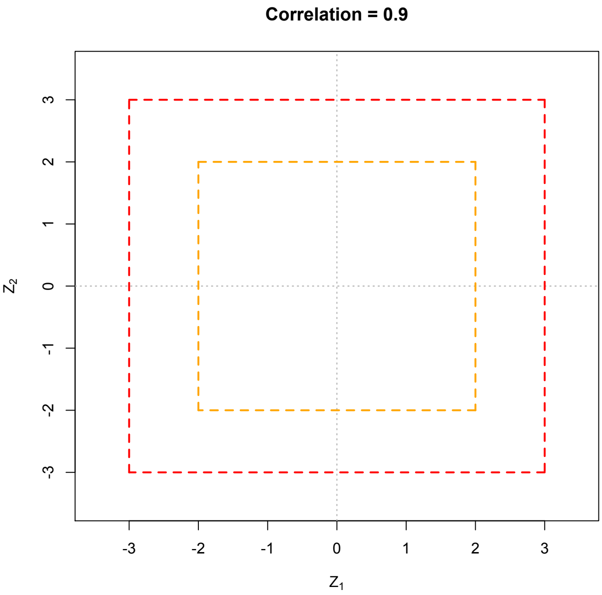
Evaluating independent z-scores in pairs



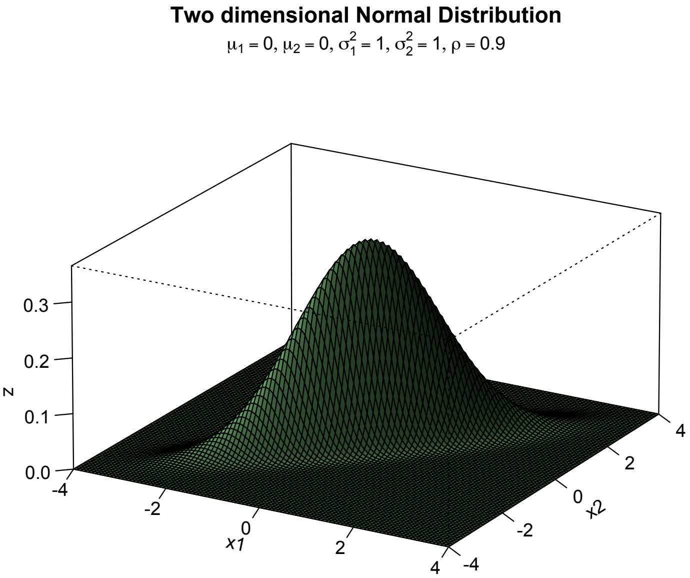
Evaluating correlated z-scores in pairs



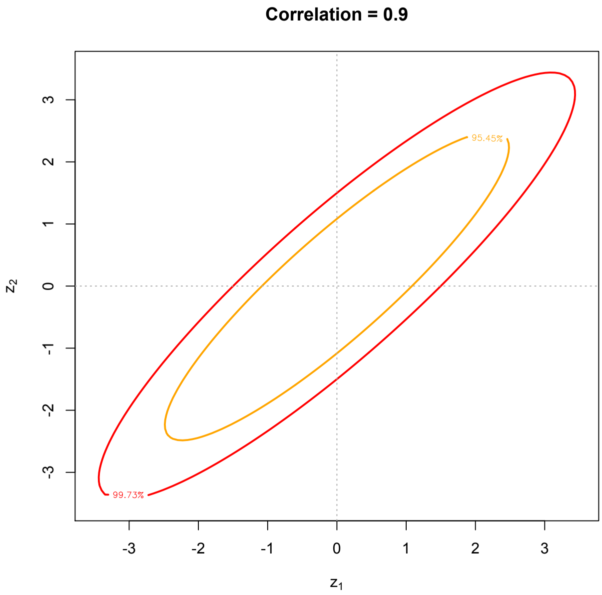
Evaluating independent z-scores in pairs



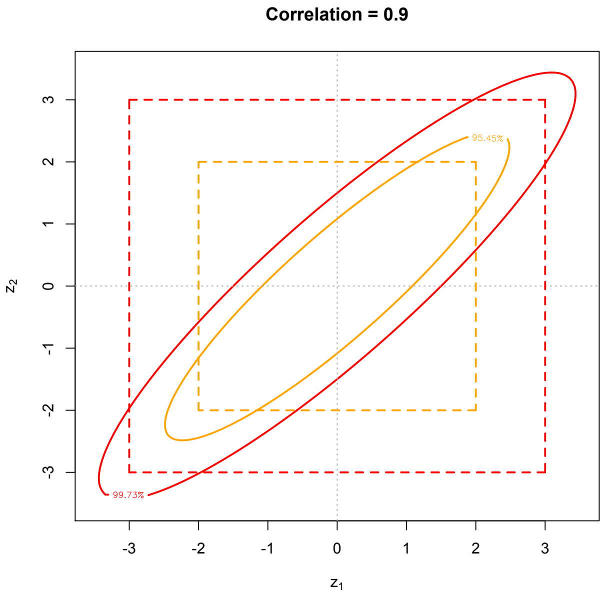
Bivariate Normal distribution (high positive corr.)



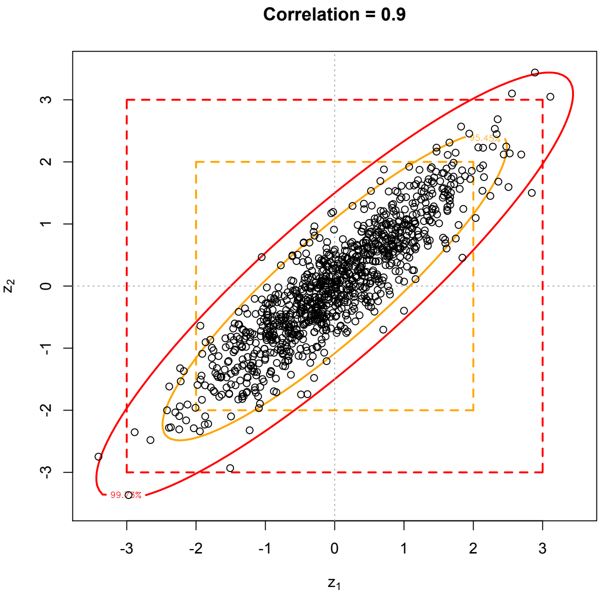
Evaluating independent z-scores in pairs



Evaluating independent z-scores in pairs



Evaluating correlated z-scores in pairs



Evaluating z-scores in pairs

- When evaluating pairs of correlated tests via their respective z-scores the concept of an appropriate bivariate confidence zone can be quite informative.

Evaluating z-scores in pairs

- When evaluating pairs of correlated tests via their respective z-scores the concept of an appropriate bivariate confidence zone can be quite informative.
- Bivariate and multivariate analysis methods used in Statistical Process Control (e.g. **Hotelling's** T^2) can be used to detect issues which are not only related to the **magnitude** of the z-score but to the **correlation** as well.

Evaluating z-score history

Longitudinal analysis of the z-scores

Using the history of z-scores one can:

Longitudinal analysis of the z-scores

Using the history of z-scores one can:

- Look for patterns in the time series plot of the z-scores, trying to identify some erratic behavior like:
  - upward/downward trend
  - parameter change
  - mixture scenario
  - cyclic behavior
  - ...

Longitudinal analysis of the z-scores

Using the history of z-scores one can:

- Look for patterns in the time series plot of the z-scores, trying to identify some erratic behavior like:
  - upward/downward trend
  - parameter change
  - mixture scenario
  - cyclic behavior
  - ...
- If for each survey we have two results, one can obtain a bivariate plot of the historic z-scores.

Longitudinal analysis of the z-scores

Using the history of z-scores one can:

- Look for patterns in the time series plot of the z-scores, trying to identify some erratic behavior like:
  - upward/downward trend
  - parameter change
  - mixture scenario
  - cyclic behavior
  - ...
- If for each survey we have two results, one can obtain a bivariate plot of the historic z-scores.
- Statistical Process Control tools can be used to detect transient shifts or persistent trends and since we typically have short horizon of data Bayesian methods are expected to be most appropriate.

Conclusions

- The EQA reports with z-scores should not be seen as binary entities:
  - No Alarm
  - Alarm

Conclusions

- The EQA reports with z-scores should not be seen as binary entities:
  - No Alarm
  - Alarm
- Useful information does exists, when one studies the z-scores with respect to:
  - number of tests that performs
  - other correlated z-scores
  - historic evolution of the z-scores

Conclusions

- The EQA reports with z-scores should not be seen as binary entities:
  - No Alarm
  - Alarm
- Useful information does exists, when one studies the z-scores with respect to:
  - number of tests that performs
  - other correlated z-scores
  - historic evolution of the z-scores
- EQA scores are snapshots of the quality in your lab, but IQC provides a video of this story... Use state of the art tools to improve the ICQ and EQA will become better.

Conclusions

- The EQA reports with z-scores should not be seen as binary entities:
    - No Alarm
    - Alarm
- Useful information does exists, when one studies the z-scores with respect to:
    - number of tests that performs
    - other correlated z-scores
    - historic evolution of the z-scores
- EQA scores are snapshots of the quality in your lab, but IQC provides a video of this story... Use state of the art tools to improve the ICQ and EQA will become better.
- Statistical Process Control tools can be very helpful not only for identifying problems in the ICQ/EQA analysis, but also providing feedback, useful to the root cause analysis.

Acknowledgments

The authors would like to thank:

- **Dr Piet Meijer** and **ECAT** organizing and scientific advisory committees for the invitation but most importantly for their interest in constantly improving the quality control tools used in practice.
- **Kostantinos Bourazas** from the department of statistics, AUEB who is actively doing research in the area of Bayesian SPC/M.
- **Dr Negrier** and **all Lyon Hemostasis team** of Lyon hospital France, for their kind support.

The End

Thank you!