Funnel plot for institutional comparison and the correction for multiple testing: the `funnelcompar` command

Silvia Forni Rosa Gini
silvia.forni@ars.toscana.it
rosa.gini@ars.toscana.it

Agenzia regionale di sanità
della Toscana

November 17, 2011

Funnel plot for institutional comparison

Some statistics

- Underlying test
- Multiple testing problem
- Exact vs approximated control limits

The `funnelcompar` command

Some examples

Indice

Funnel plot for institutional comparison

Some statistics

- Underlying test
- Multiple testing problem
- Exact vs approximated control limits

The `funnelcompar` command

Some examples

Background

STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:1185–1202
Published online 29 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.1970

Funnel plots for comparing institutional performance

David J. Spiegelhalter*,†

MRC Biostatistics Unit, Institute of Public Health, Cambridge CB2 2SR, U.K.

SUMMARY

'Funnel plots' are recommended as a graphical aid for institutional comparisons, in which an estimate of an underlying quantity is plotted against an interpretable measure of its precision. 'Control limits' form a funnel around the target outcome, in a close analogy to standard Shewhart control charts. Examples are given for comparing proportions and changes in rates, assessing association between outcome and volume of cases, and dealing with over-dispersion due to unmeasured risk factors. We conclude that funnel plots are flexible, attractively simple, and avoid spurious ranking of institutions into 'league tables'.

KEY WORDS: control charts; outliers; over-dispersion; institutional profiling; ranking

Background

- **Quantitative indicators** are increasingly used to monitor health care providers
- Interpretation of those indicators is often open to anyone (patients, journalists, politicians, civil servants and managers)
- It is crucial that indicators are both accurate and presented in a clear way to avoid unfair criticism

Classical presentation: *league tables*

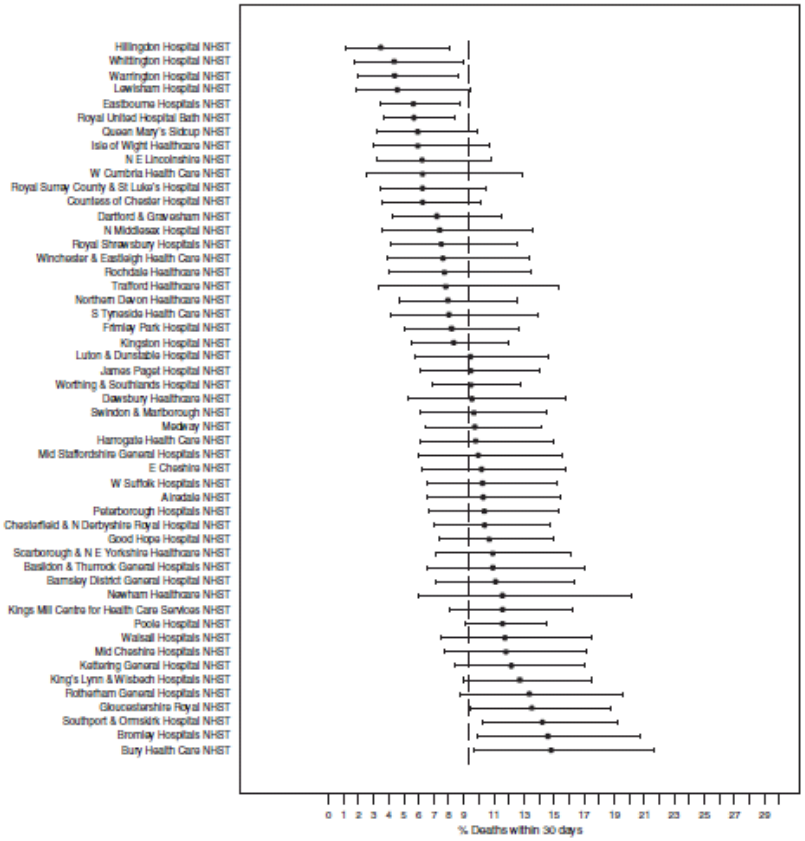


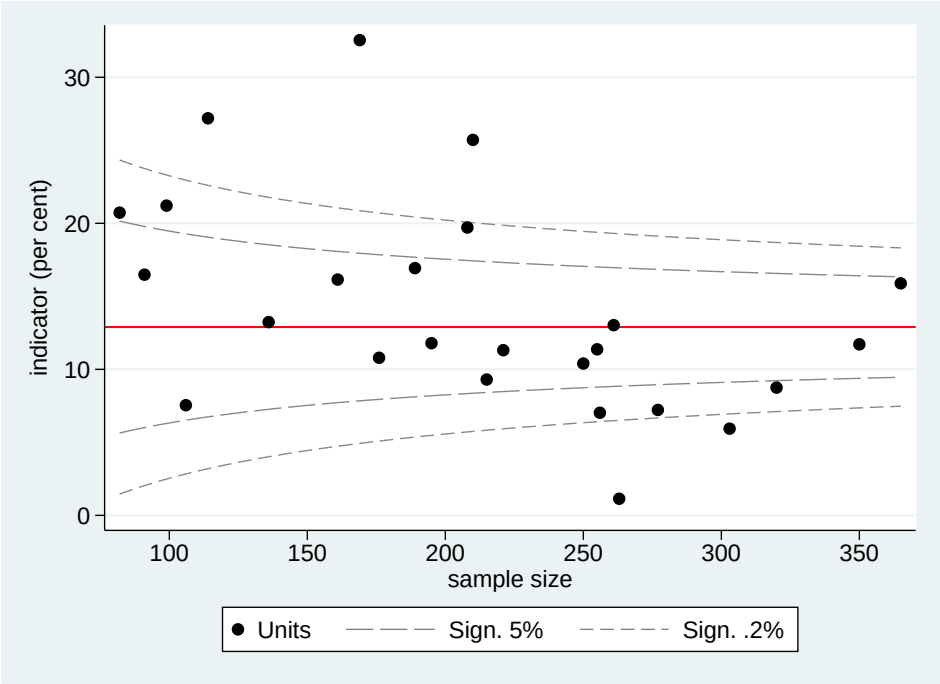
Figure 1. 'Caterpillar' plot of 30-day mortality rates, age and sex standardized, following treatment for fractured hip for over-65's in 51 medium acute and multi-service hospitals in England, 2000–2001. Ninety-five per cent confidence intervals are plotted and compared to the overall proportion of 9.3 per cent.

- Imply the existence of **ranking** among institutions
- Implicitly support the idea that **some of them are worse/better than others**

Statistical Process Control methods: key principles

- Variation, to be expected in any process or system, can be devided into:
  - **Common cause variation**: expected in a stable process
  - **Special cause variation**: unexpected, due to systematic deviation
- Limits between these two categories can be set using SPC methods
- Funnel plots:
  - All institutions are part of a single system and perform at the same level
  - Observed differences can never be completely eliminated and are explained by chance *(common cause variation)*.
  - If observed variation exceeds that expected, *special-cause variation* exists and requires further explanation to identify its cause.

Funnel Plot



- **Scatterplot** of observed indicators against a measure of its precision, tipically the sample size
- **Horizontal line** at a target level, typically the group avarage
- **Control Limits** at 95% ($\approx 2SD$) and 99.8% ($\approx 3SD$) levels, that narrows as the sample size gets bigger

Association of Public Health Observatories in UK developed analytical tools in excell for producing funnel plot

Multiple testing problem

ELSEVIER

Journal of Clinical Epidemiology 61 (2008) 232–240

Journal of Clinical Epidemiology

Use of the false discovery rate when comparing multiple health care providers

Hayley E. Jones*, David I. Ohlssen, David J. Spiegelhalter

MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK

Accepted 18 April 2007

Abstract

Objective: Comparisons of the performance of multiple health care providers are often based on hypothesis tests, those with resulting *P*-values below some critical threshold being identified as potentially extreme. Because of the multiple testing involved, the classical *P*-value threshold of, say, 0.05 may not be considered strict enough, as it will tend to lead to too many "false positives." However, we argue that the commonly used Bonferroni-corrected threshold is in general too strict for the problem in hand. The purpose of this article is to demonstrate a suitable alternative thresholding procedure that is already well established in other fields.

Study Design and Setting: The suggested procedure involves control of an error measure called the "false discovery rate" (FDR). We present a worked example involving a comparison of risk-adjusted mortality rates following heart surgery in New York State hospitals during 2000–2002. It is shown that the FDR critical threshold lines can be drawn on a "funnel plot," providing a simple graphical presentation of the results.

Results: The FDR procedure identified more providers as potentially extreme than the Bonferroni correction, while maintaining control of an intuitively sensible error measure.

Conclusion: Control of the FDR offers a simple guideline to determining where to draw critical thresholds when comparing multiple health care providers.

Indice

Funnel plot for institutional comparison

Some statistics

- Underlying test
- Multiple testing problem
- Exact vs approximated control limits

The `funnelcompar` command

Some examples

A funnel plot has four components:

- An *indicator* Y.
- A *target* θ which specifies the desired expectation for institutions considered "in control".
- A *precision* parameter N determining the accuracy used in measuring the indicator. Select a N directly interpretable, eg the denominator for rates and means.
- *Control limits* for a p-value, computed assuming Y has a known distribution (normal, binomial, Poisson) with parameters (θ, σ).

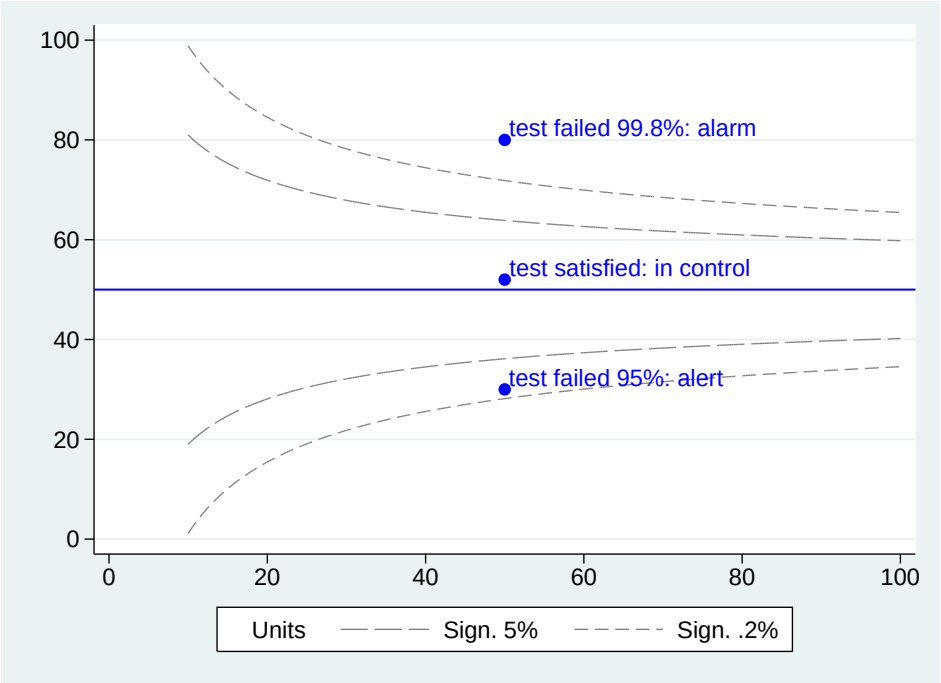
From a purely statistical point of view, funnel plot is a graphical representation testing whether each value Y_i belongs to the known distribution with given parameters.

The formal test of significance:

$$H_0 : Y_i = \theta$$

$$H_1 : Y_i \neq \theta$$

$$Z = \frac{Y_i - \theta}{(\sigma / \sqrt{N})}$$



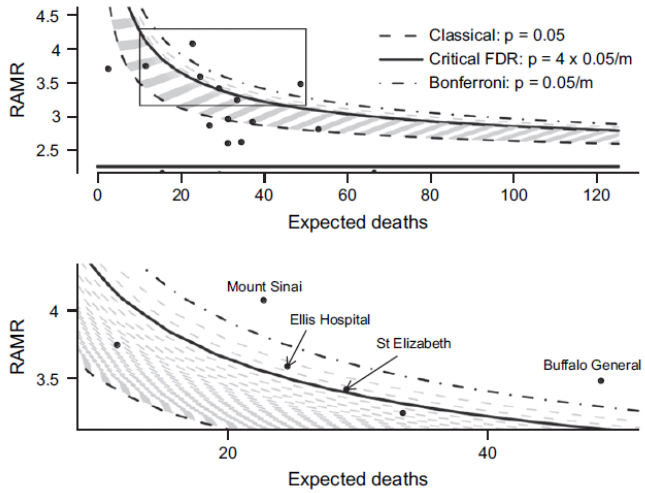


Fig. 4. Sections of the funnel plot of hospital-specific RAMRs. The upper frame shows the entire top half of the funnel plot, the solid horizontal line indicating the SWMR target of 2.26. The boxed section is magnified in the lower frame. Each dashed line indicates a potential FDR threshold. The critical FDR threshold is the closest line to the target such that more points than lines lie beyond it. In this case, this is seen to be the line corresponding to $P = 4 \times 0.05/m$.

Classical Bonferroni correction strongly control $P(F >= 1)$ at level α. Too conservative!

An alternative is control of the expected proportion of errors among all rejected null hypotheses

False Discovery Rate FDR

$$FDR = E(\frac{F}{S}|S > 0)$$

with:

- F: null true
- S: null rejected, significants

The FDR is the probability that a provider isn't a genuine extreme, given that it is called significant by the test. Benjamini and Hochberg proposed an algoritm for costraining the FDR to be less than of equal to α

False Discovery Rate FDR

- Let

$$P_1, P_2, \ldots, P_m$$

be the ordered P-values and and H_i be the null hypotesis corresponding tp P_i, $i = 1, \ldots, m$

- Identify k as the largest i such that

$$P_i \leq (\frac{i}{m})\alpha$$

- Then reject all H_i, $i = 1, \ldots, k$

Control limits

In cases of discrete distributions there are two possibilities for drawing control limits as functions of N

- a normal approximation:

$$y_p(N) = \theta \pm z_p \frac{\sigma}{\sqrt{N}}$$

- an "exact" formula

$$y_p(N) = \frac{r_{(p,N,\theta)} - \alpha}{N}$$

where $r_{(p,N,\theta)}$ and α are defined in the following slides

Binomial

In the case of binomial distribution:

- $r_{(p,N,\theta)}$ is the inverse to the cumulative binomial distribution with parameters (θ, N) at level p. The definition Spiegelhalter refers to is as follows:[1] if $F_{(\theta,N)}$ is the cumulative distribution function, ie $F_{(\theta,N)}(k)$ is the the probability of observing k or fewer successes in N trials when the probability of a success on one trial is θ,[2] then $r_p = r_{(p,N,\theta)}$ is the smallest *integer* such that

$$P(R \leq r_p) = F_{(\theta,N)}(r_p) > p$$

- α is a continuity adjustment coefficient

$$\alpha = \frac{F_{(\theta,N)}(r_p) - p}{F_{(\theta,N)}(r_p) - F_{(\theta,N)}(r_p - 1)}$$

---

[1] Beware that the Stata function `invbinomial()` is *not* defined this way.

[2] The Stata function `binomial(N,k,`θ`)` computes $F_{(\theta,N)}(k)$.

Poisson

In the case of Poisson distribution:

- $r_{(p,N,\theta)}$ is the inverse to the cumulative Poisson distribution with parameter $M = \theta N$ at level p. The definition Spiegelhalter refers to is as follows:[3] if F_M is the cumulative distribution function, ie $F_M(k)$ is the probability of observing k or or fewer outcomes that are distributed Poisson with mean M,[4] then $r_p = r_{(p,N,\theta)}$ is the smallest *integer* such that

$$P(R \leq r_p) = F_M(r_p) > p$$

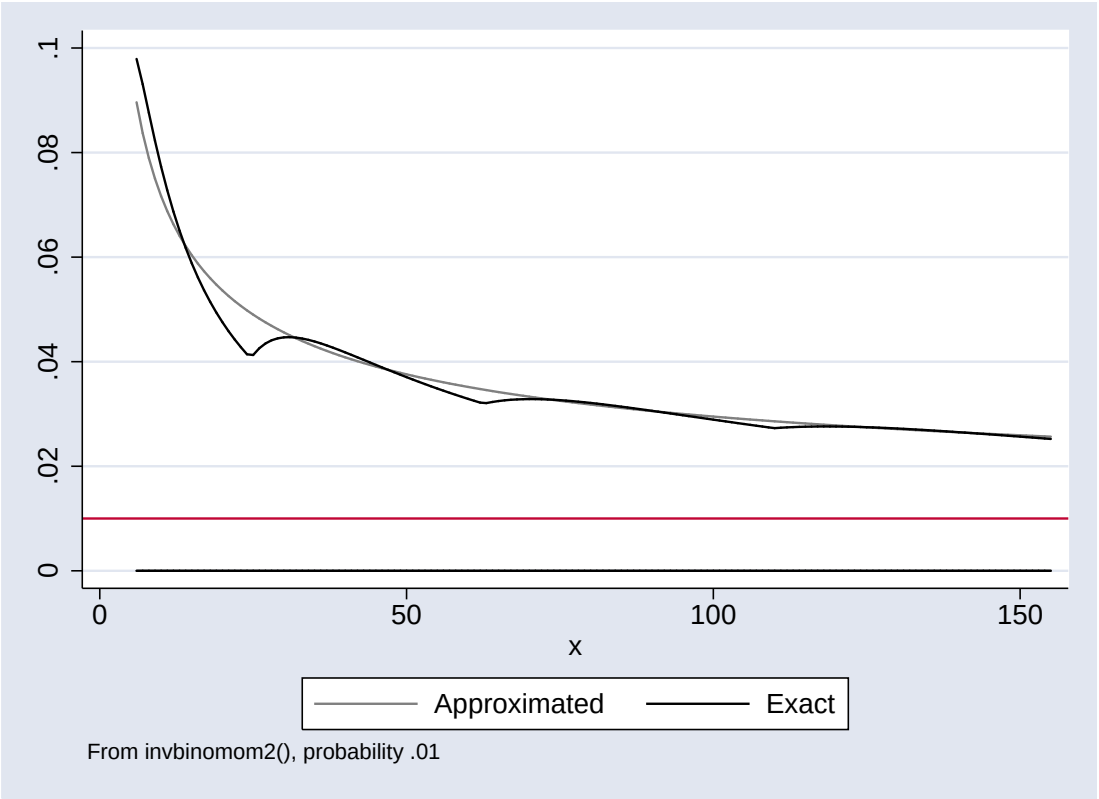
- α is a continuity adjustment coefficient

$$\alpha = \frac{F_M(r_p) - p}{F_M(r_p) - F_M(r_p - 1)}$$

[3] Beware that the Stata function `invpoisson()` is *not* defined this way.

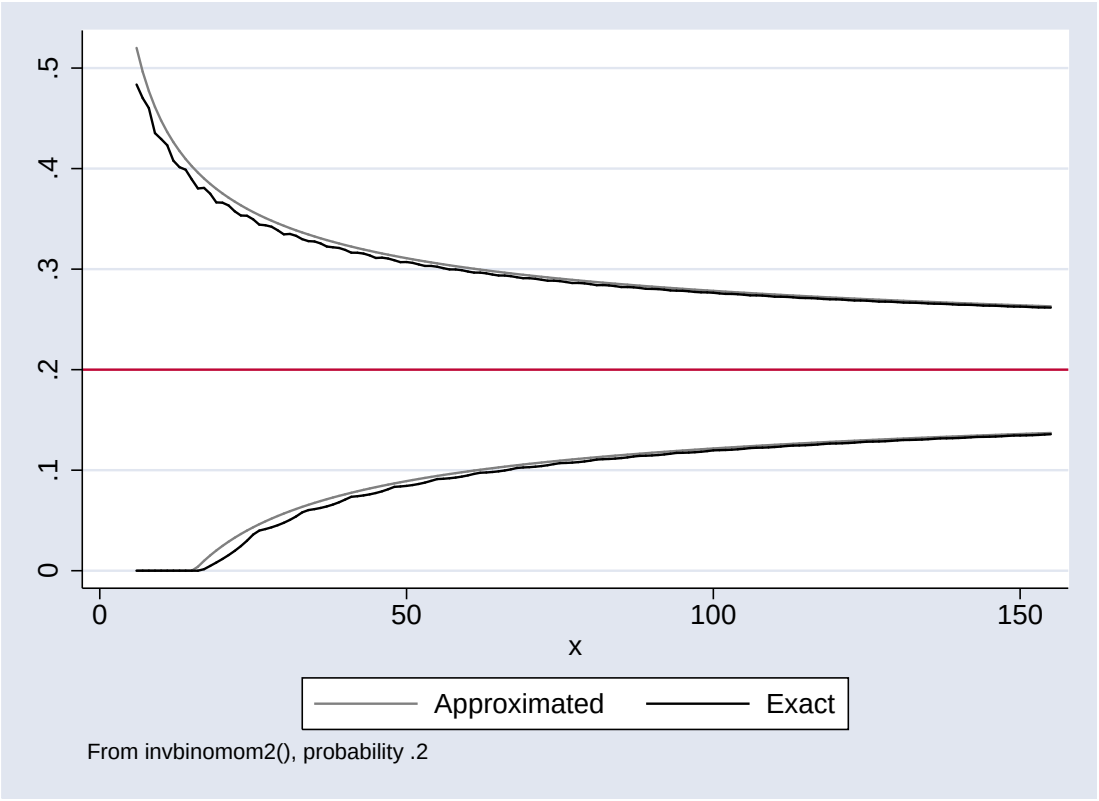
[4] The Stata function `poisson(M,k)` computes $F_M(k)$.

Example 1: binomial, θ=1%



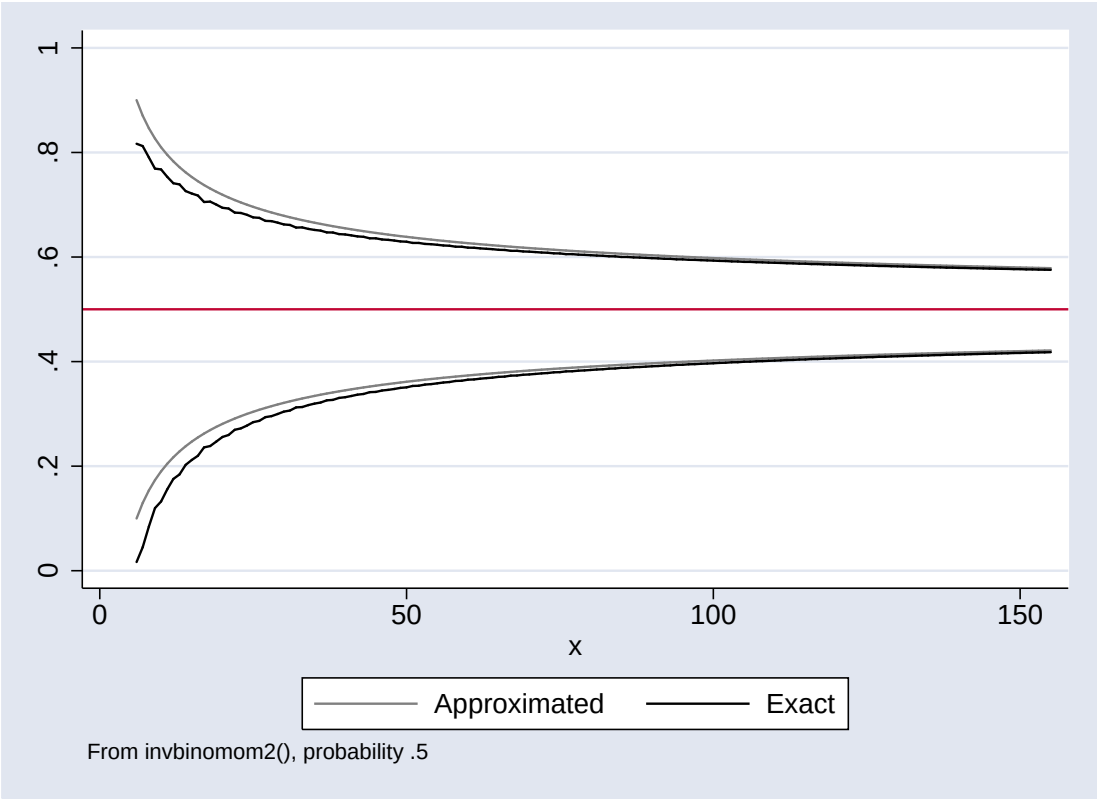
- Does it make sense to test a 1% of cases with $N < 100$?
- For $N > 100$ the two pairs of curves almost coincide

Example 2: binomial, θ=20%



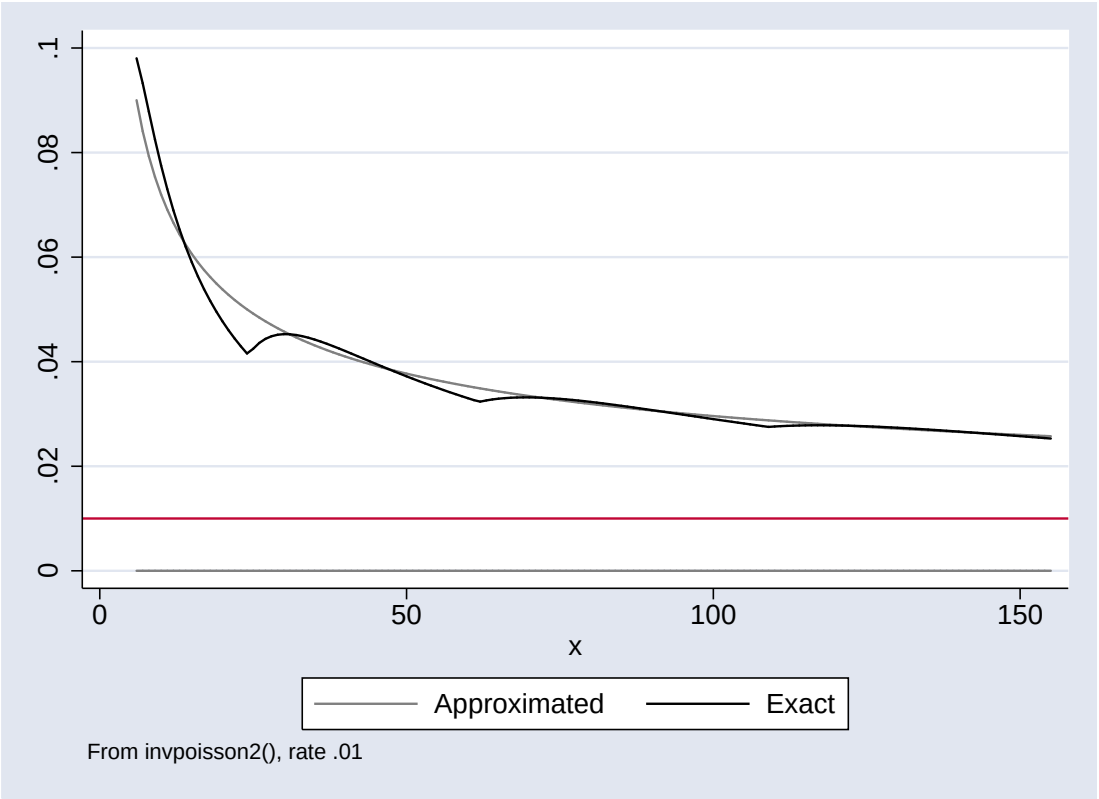
- For $N < 100$ very similar curves, approximated upper bounds conservative
- For $N > 100$ the two pairs of curves almost coincide

Example 3: binomial, θ=50%



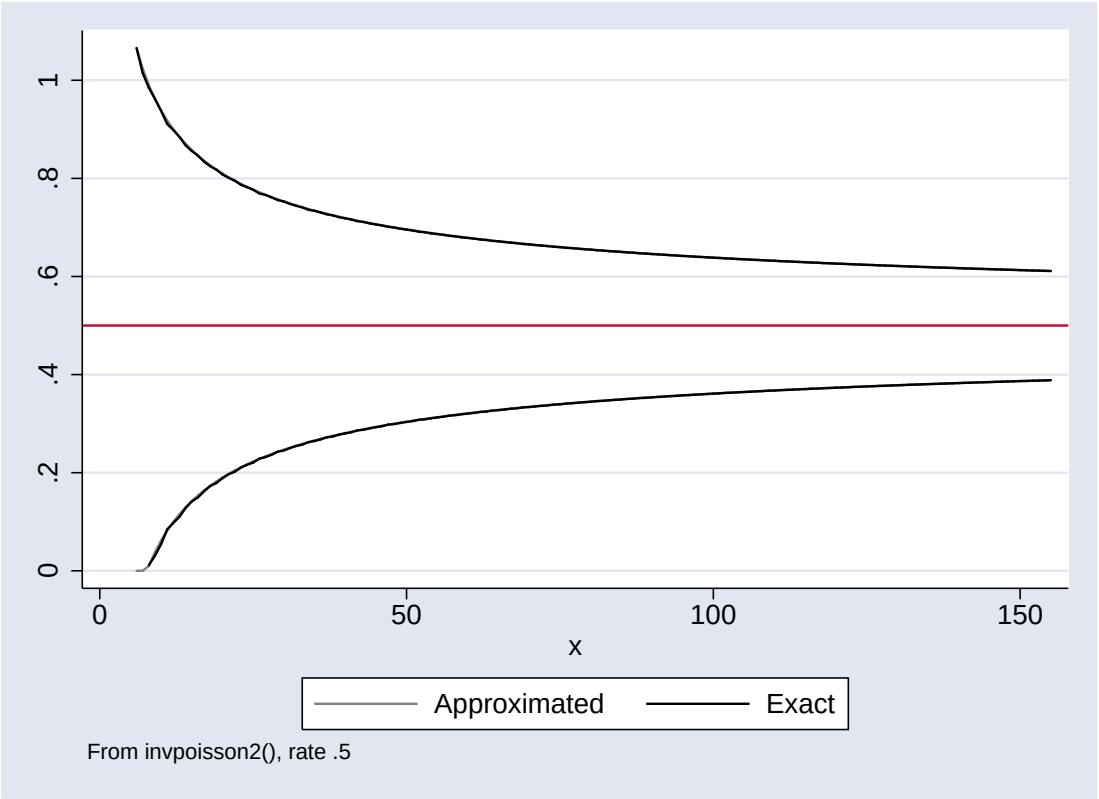
- For $N < 100$ very similar curves, approximated upper bounds conservative
- For $N > 100$ the two pairs of curves almost coincide

Example 4: Poisson, θ=1%



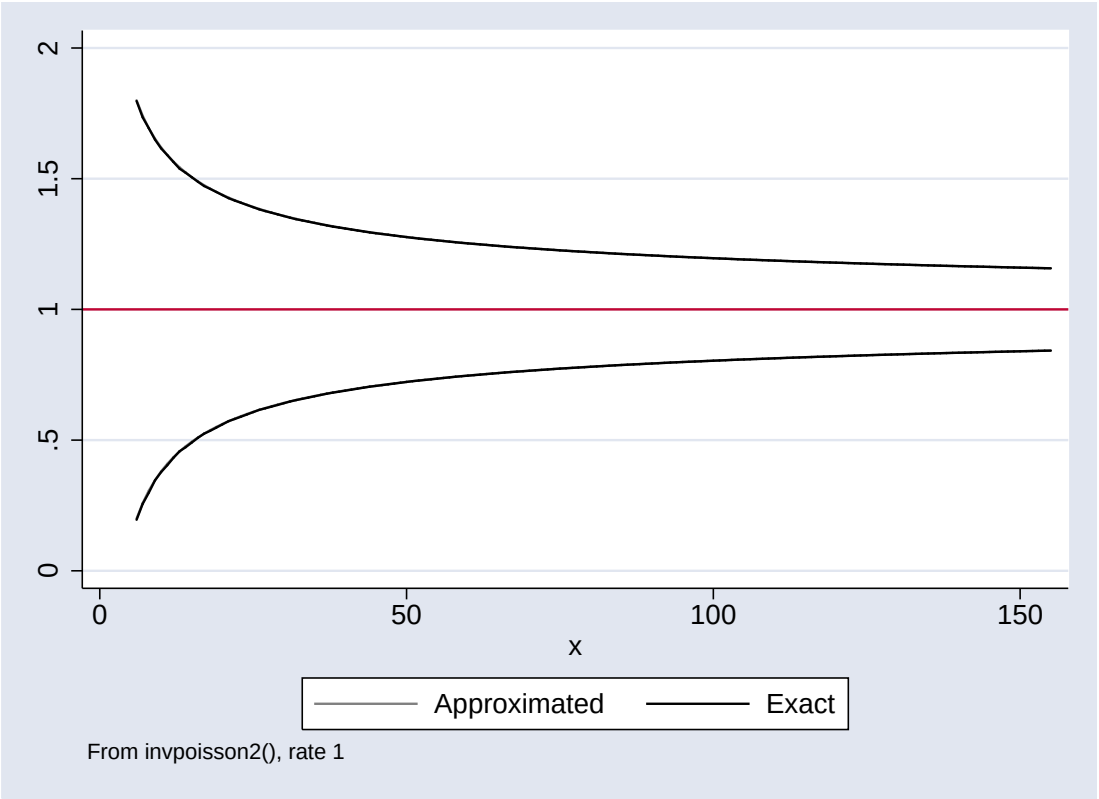
- Does it make sense to test a 1% of cases with $N < 100$?
- For $N > 100$ the two pairs of curves almost coincide

Example 5: Poisson, θ=50%



The two pairs of curves almost coincide

Example 6: Poisson, θ=1 (SMR)



The two pairs of curves visibly coincide

Conclusion for using exact vs approximated test

- Whenever the sample size is bigger than 100, the approximated test is almost superimposed to the exact test
- Consider if it makes sense to use exact test

Indice

Funnel plot for institutional comparison

Some statistics

- Underlying test
- Multiple testing problem
- Exact vs approximated control limits

The `funnelcompar` command

Some examples

Basic syntax

```
 funnelcompar value pop unit [sdvalue],
 [continuous/binomial/poisson]
 [fdr bonferroni onlyfdr onlybonferroni]
 [ext_stand() ext_sd() noweight smr ]
 [constant()]
 [contours() exact]
```

marking options

other options

Variables

`funnelcompar value pop unit [sdvalue]`

- *value* contains the values of the indicator.
- *pop* contains the sample size (precision parameter)
- *unit* contains an identifier of the units
- *sdvalue* contains the standard deviations of indicators (optionally, if the continuous option is also specified)

Distribution

Users must specify a distribution among:

- *normal*: option `cont`
- *binomial*: option `binom`
- *Poisson*: option `poiss`

Parameters: θ

θ can be obtained as:

- weighted mean of *value* with weights *pop* (default)
- non weighted mean of *value* if the `noweight` option is specified
- external value specified by users with the option `ext_stand()`

Parameters: σ

- Binomial distribution: $\sigma = \sqrt{\theta(1-\theta)}$
- Poisson distribution: $\sigma = \sqrt{\theta}$
- Normal distribution:
  - weighted mean of *sdvalue* with weights *pop* (defualt)
  - non weighted mean of *sdvalue* if the `noweight` option is specified
  - external value specified by users with the option `ext_sd()`

Multiple testing options

- `bonferroni` option draws both classical and bonferroni corrected thresholds
- `fdr` option draws both classical and fdr corrected threshold
- `onlybonferroni` and `onlyfdr` options draw only Bonferroni or fdr corrected thresholds respectively

The `smr` option

- `smr` option can be specified only with `poisson` option:
- *value* are assumed to be indirectly standardized rates
- *pop* contains the expected number of events
- θ is assumed to be 1

Constant

- The `constant()` option specifies whether the values of the indicators are multiplied by a constant term, for instance `constant(100)` must be specified if the values are percentages.

Curves

- `contours()`: specifies significance levels at which control limits are set (as a percentage).
- Default `contours()` are set at 5% and .2% levels, that is a confidence of 95% and 99.8% respectively.
- For example if `contours(5)` is specified only the curve corresponding to a test with 5% of significance is drawn.
- For discrete distributions if the `exact` option is specified, the exact contours are drawn. As a default the normal approximation is used.

Indice

Funnel plot for institutional comparison

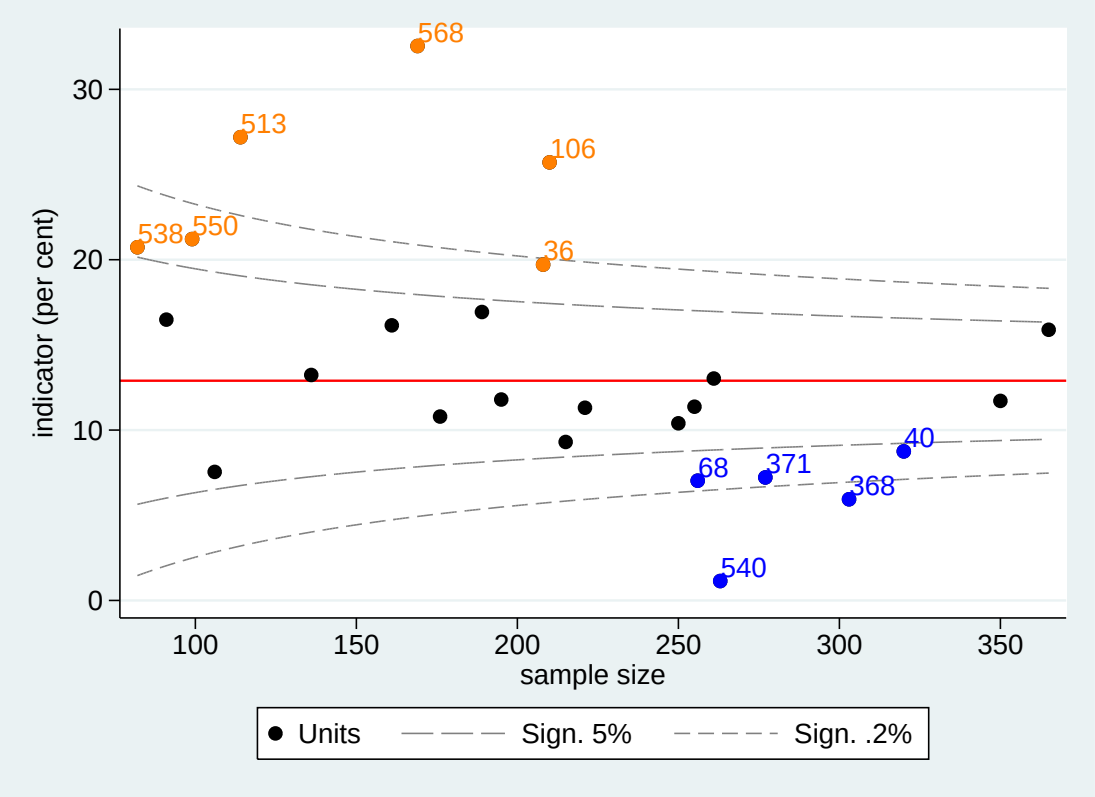
Some statistics

- Underlying test
- Multiple testing problem
- Exact vs approximated control limits

The `funnelcompar` command

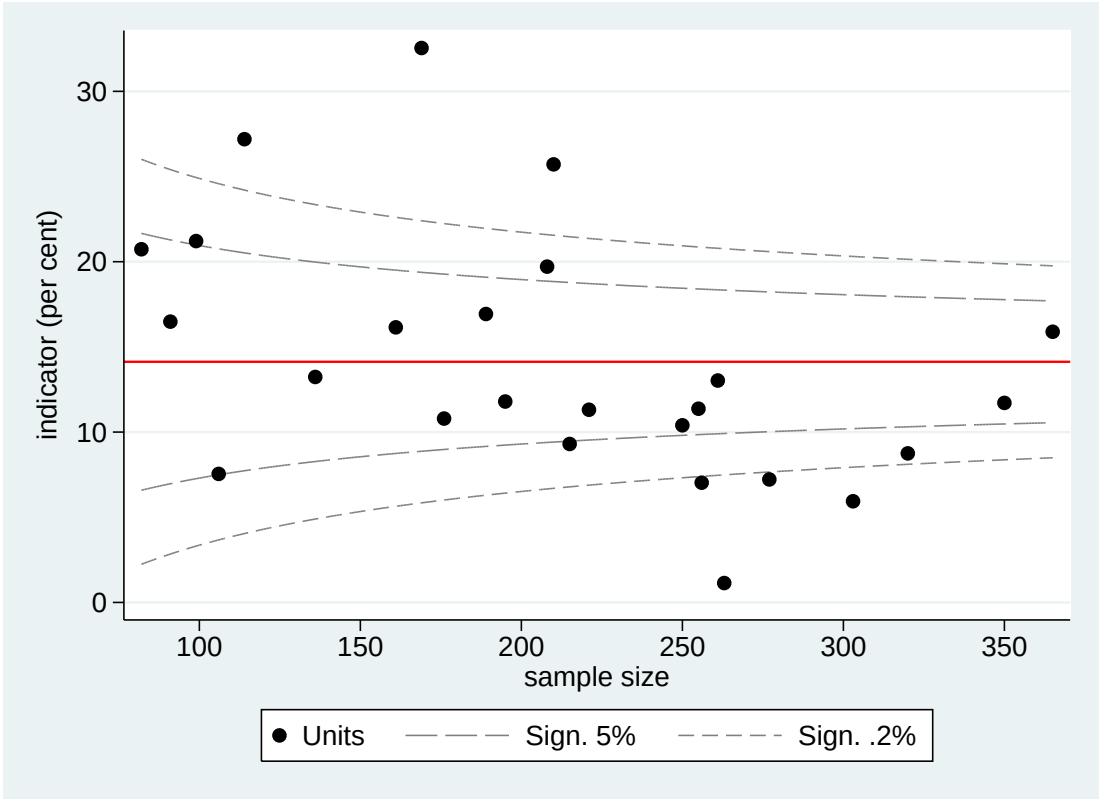
Some examples

Percentages, internal target, units out-of-control marked



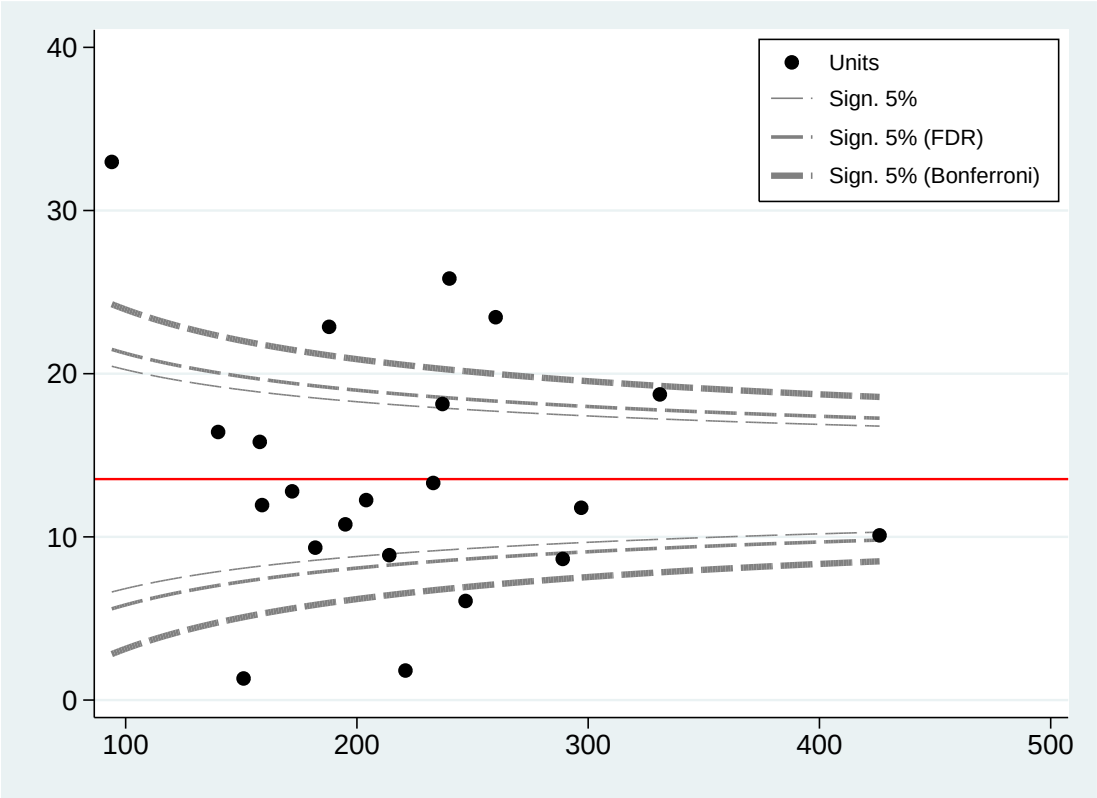

```
funnelcompar
measure pop unit,
binom const(100)
markup marklow
```

Percentages, no-weighted internal target



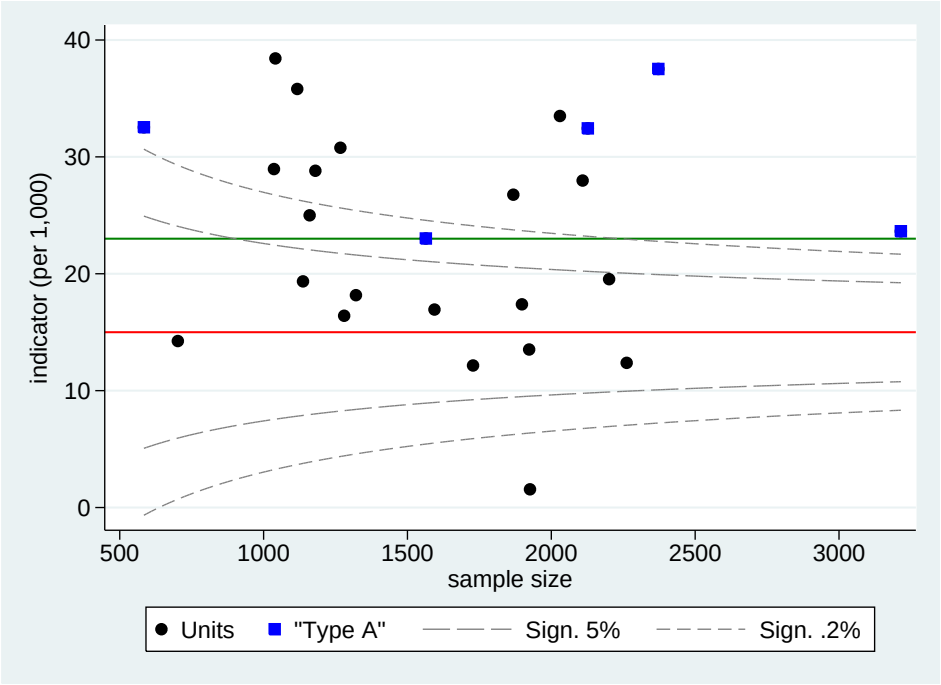

```
funnelcompar
measure pop unit,
binom const(100)
noweight
```

Percentages, no-weighted internal target, with bonferroni and fdr thresholds



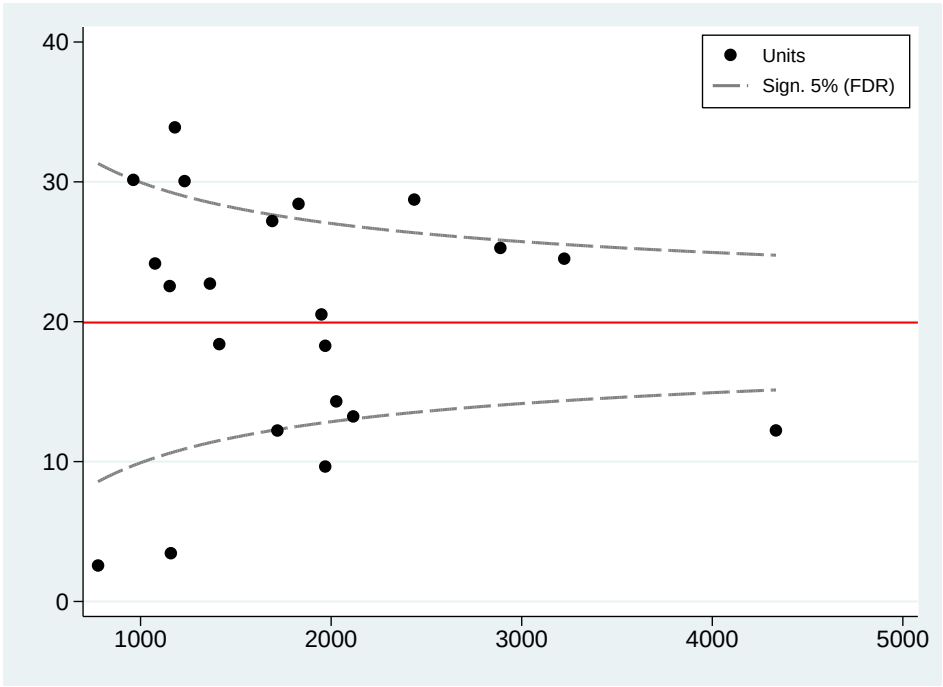

```
funnelcompar
measure pop unit,
binom const(100)
fdr bonferroni
noweight
```

Rates, external target, type-A units marked



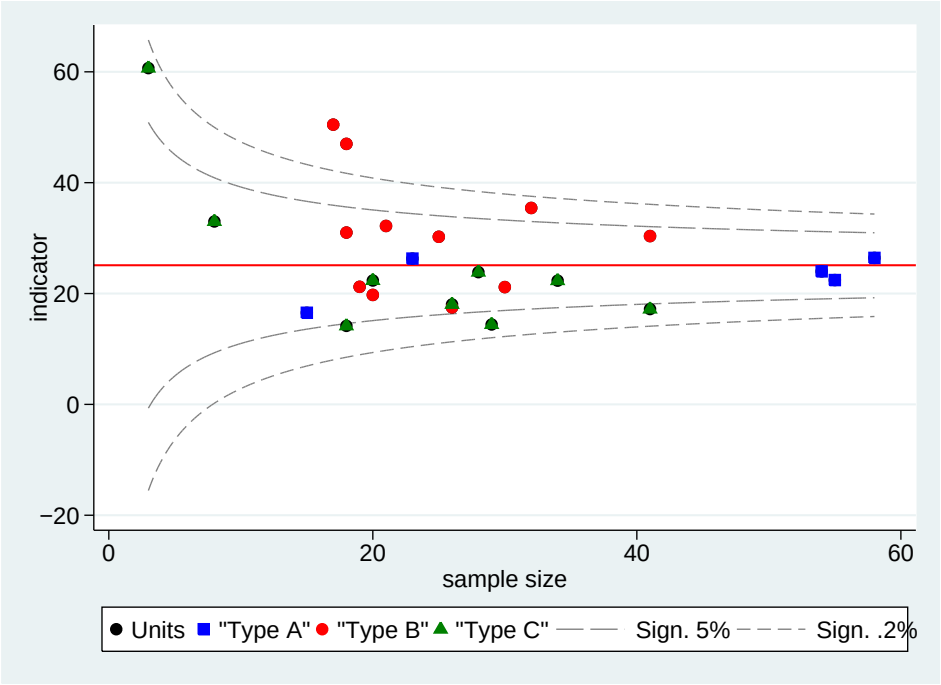

```
funnelcompar measure pop
unit, poisson
const(1000) ext_stand(15)
markcond(type = 1)
legendmarkcond(Type A)
colormarkcond(blue)
optionsmarkcond(msymbol(S))
twowayopts(yline(23,
lcolor(green)) )
```

Rates, only fdr threshold



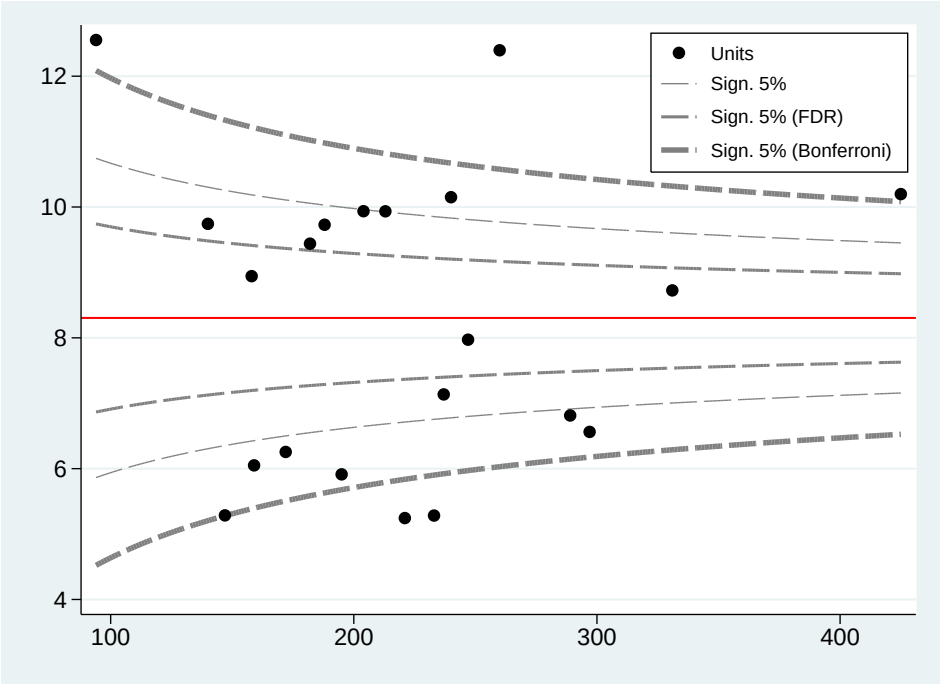

```
funnelcompar measure pop
unit, poisson
const(1000) fdr onlyfdr
legendmarkcond
```

Means, internal target, unit type marked



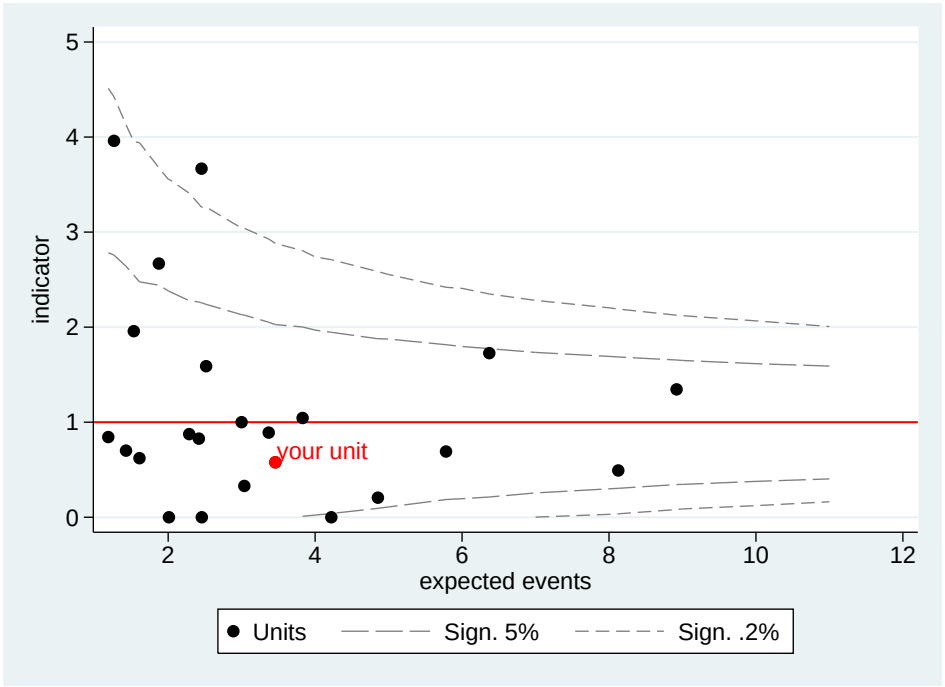

```
funnelcompar measure pop
unit sd, cont const(1)
markcond(type=1)
legendmarkcond(Type A)
colormarkcond(blue)
optionsmarkcond(msymbol(S))
markcond1(type = 2)
...markcond2(type=3) ...
```

Means, internal target, bonferroni and fdr thresholds




```
funnelcompar measure pop
unit sd, cont const(1)
fdr bonferroni
```

Standardized Incidence Rates, one unit marked




```
funnelcompar smr exp
unit, poisson smr
markunit(5 "your unit")
legendopts(placement(se)
row(1))
```

References

Spiegelhalter DJ

Funnel plots for comparing institutional performance.

Statist. Med. 2005: 24:1185-1202.

Jones HE, Ohlssen DI, Spiegelhalter DJ

Use of false discvery rate when comparing multipli helaht care providers.

J Clin Epid 2008: 61:232-240.

Spiegelhalter DJ

Funnel plots for institutional comparison.

Qual Saf Health Care 2002 Dec;11(4):390-1.

Spiegelhalter DJ

Handling over-dispersion of performance indicators.

Qual Saf Health Care 2005 Oct;14(5):347-51.

Acknowledgements

- We thank Neil Shephard, Paul Silcocks and Hayley Jones for valuable discussion.
- Our routine is heavily based on `confunnel` by Tom Palmer.
- Many programming tricks were stolen from `eclplot` and other routines by Roger Newson.