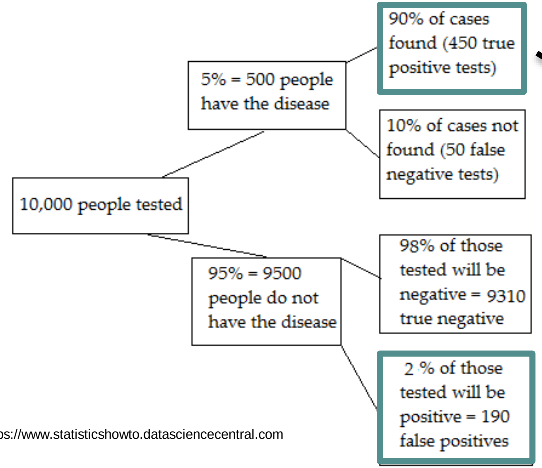
Controlling the false discovery rate in multiple hypothesis testing



https://www.statisticshowto.datasciencecentral.com

FDR = 190/640 = 30%

Oliver Gutjahr

Max-Planck-Institut für Meteorologie, Hamburg

“The stippling shows statistically significant grid points

- Wilks, D. S. (2016, BAMS)

Individual tests at many spatial grid points are very often interpreted incorrectly (**multiplicity**)

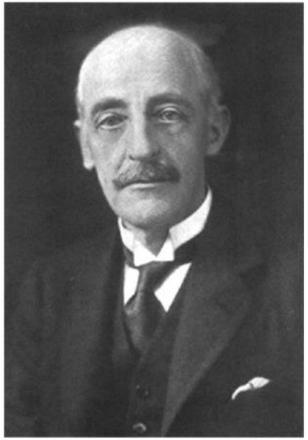
→ research **results are overstated**

Out of **281** papers in *Journal of climate* (first half of 2014):

- **97 (34.5%)** did not account for multiplicity
- **3 (1.1%)** accounted for multiplicity

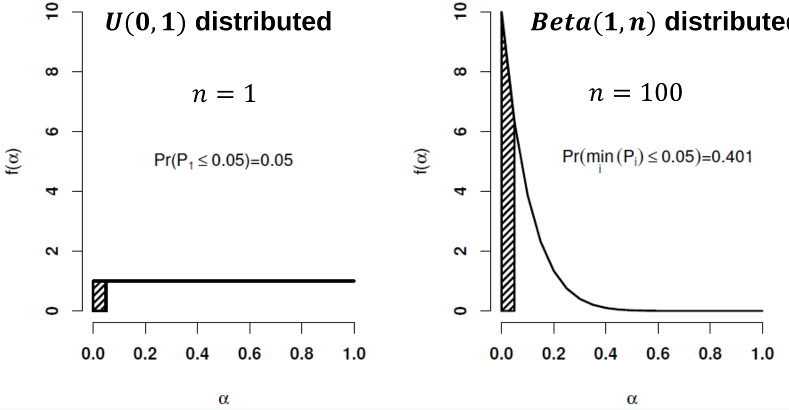
Multiple testing problem – no new story…

- **Multiple testing problem** known at least back to Walker* (1914)
- Walker‘s method was modernized (Katz and Brown,1991; Katz, 2002) and nowadays known as ***Walkers‘s test****:*
  - – Walker noted that the likelihood of small p-value rises with larger n:



Sir Gilbert Thomas Walker
(source Royal Society; Taylor 1962)
*14 Jun 1868 † 4 Nov 1958

Probability that p-value is $\leq \alpha$

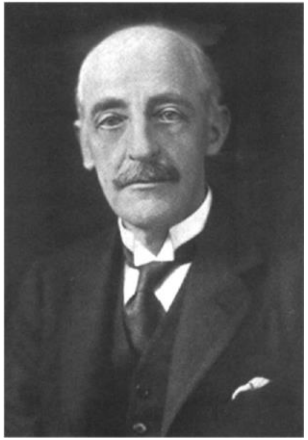


Hedderich and Sachs (2018, mod.)

* *“Walker Circulation“* named after him; he first described and named the SO (ENSO), NAO and NPO

Multiple testing problem – no new story…

- **Multiple testing problem** known at least back to Walker* (1914)
- Walker's method was modernized (Katz and Brown,1991; Katz, 2002) and nowadays known as ***Walkers's test****:*
  - a more strict significance level is required: $\alpha_{Walker} = 1 - (1 - \alpha)^{\frac{1}{n}}$
  - global H_0^G rejected if $p_{(1)} \leq \alpha_{Walker}$



Sir Gilbert Thomas Walker
(source Royal Society; Taylor 1962)

- assumes *independence* and is very *conservative* ($\alpha_{Walker} \approx \alpha/n$)
- no judgement of *local test results* (H_0^i)
- before we come to a more appropriate method, we need to understand the **origin of the multiple testing problem**

* Sir Gilbert Thomas Walker: "*Walker Circulation*" named after him; he first described and named the SO (ENSO), NAO and NPO

Hypothesis testing framework

	Declared non-significant (H_0)	Declared significant (H_A)	Total
True Null Hypothesis	**U** *Correct* ($\mathbf{1-\alpha}$)	**V** *Type I error* ($\boldsymbol{\alpha}$) "false positive/discovery"	m_0
Non-true Null Hypothesis	**T** *Type II error* ($\boldsymbol{\beta}$) "false negative"	**S** *Correct* ($\mathbf{1-\beta}$, ***power***)	$m_1 = m - m_0$
Total	*m* - **R**	**R**	*m*

V = Type I error (False Positive / False Discovery)
T = Type II error (False Negative)
S = True positives
R = total tests declared significant
m = number of hypotheses tested
m_0 = unknown number of true null hypotheses
m_1 = unknown number of non-true hypotheses

U, V, T, S are unobserved random variables
R is an observable random variable

Pitfalls/considerations here:

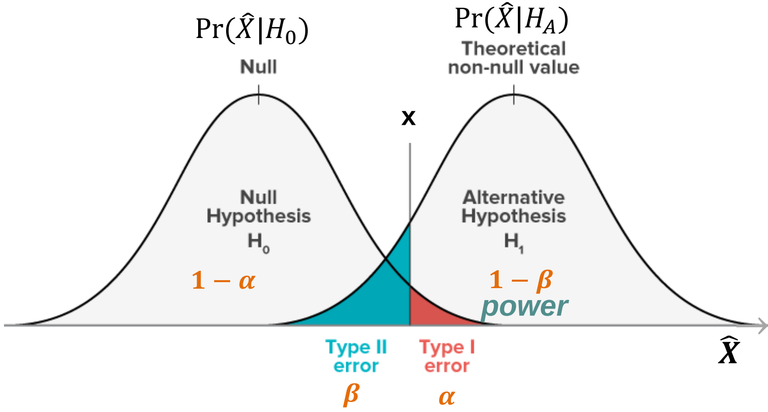
1. we need to formulate a good hypothesis
2. we need to choose appropriate test with maximum power (assumptions of testing procedure)
3. a-priori choose α
4. if H_0 is rejected, H_A is not automatically true

Important for us is **V!**

Hypothesis testing framework – single test ($n = 1$)

- if we test at the $\alpha = 5\%$ level[A], the probability to falsely reject a true H_0 is 5%.
- reject H_0: if probability (p-value) of observed or any more extreme test statistic $\hat{X}$, given that H_0 is true, is no larger than α:

$$\underbrace{\Pr(\hat{X} \geq x | H_0)}_{p\text{-value}} \leq \alpha$$



- if H_0 is rejected with $\alpha = 5\%$, the result is said be significant at the 5% level[B]

[A] First formal statement by Fisher (1925), but originates back to gambling theory in 17th century; introduced to social and natural science by Laplace (1749-1827) and Gauss (1777-1855), see Cowles and Davis (1982).
[B] Often expressed as "at the 95% level".

Multiple testing problem – assume all H_0 are true

	Declared non-significant (H_0)	Declared significant (H_A)	Total
True Null Hypothesis	**U** *Correct* ($1-\alpha$)	**V** *Type I error* (α)	m_0

- any **single** true H_0 will be rejected with probability α
- collection of $\boldsymbol{m_0}$ tests with true H_0 will exhibit, on average, $\mathbf{V} = \boldsymbol{\alpha m_0}$ erroneous rejections, if **independent***:

Is actually the mean of the binomial distribution, so even more false positives are likely

  - **Example 1**: if we perform $m_0 = 100$ tests, then on average $\boldsymbol{\alpha m_0 = 5}$ tests will result in false positives.
  - **Example 2**: if $m_0 = 802 \times 404 = 324008$ (TP04), then we get $\boldsymbol{\alpha m_0 = 16200}$ false positives on average just by chance!

*Probability is higher with dependencies

A global perspective – *field significance*

- define a **global or meta-test** on many individual test results – known as *field significance** (Livizey and Chen, 1983; Von Storch, 1982)
- **Livizey and Chen‘s approach**:
  - global null hypothesis H_0^G: *all local* $H_0^i = true$; $H_A^G: n > \alpha m_0$ of H_0^i rejected
  - how many H_0^i need to be rejected so that $\Pr(n > \alpha m_0) \leq \alpha_{global} = \alpha = 0.05$?
    (e.g. *binomial distribution*: if $n = 100$ then $\mathrm{n} \geq 10$)
  - better than näive stippling approach but many drawbacks
    (e.g. assumes independence, very sensitive to violation, too permissive → intensive resampling)
- often we are not interested in a global meta-test – **we want to know the locations that are significant**

*Local tests pertain usually to a grid, thereby composing a *field* of test results.

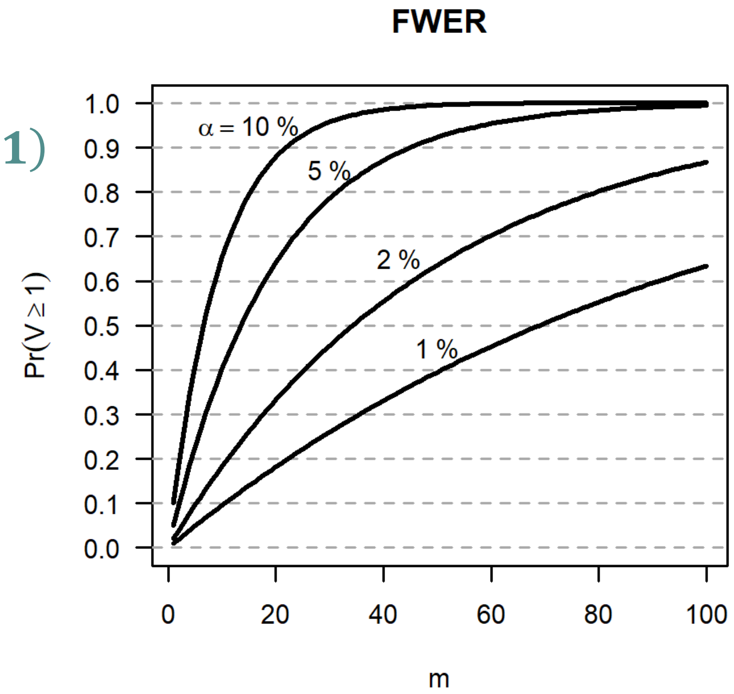
Probability of at least one wrong false positive: $\mathbf{Pr}(V \geq 1)$?

- *Family Wise Error Rate* (**FWER**) = $\mathbf{Pr}(V \geq 1)$

- if test results are **independent***, probability follows binomial distribution:

Probability of no false positive:
$$\mathbf{Pr}(V = 0) \sim Bi(m, \alpha)$$

Probability of at least one:
$$\mathbf{Pr}(V \geq 1) = 1 - \mathrm{Pr}(V = 0) = \mathbf{1 - (1 - \alpha)^m}$$



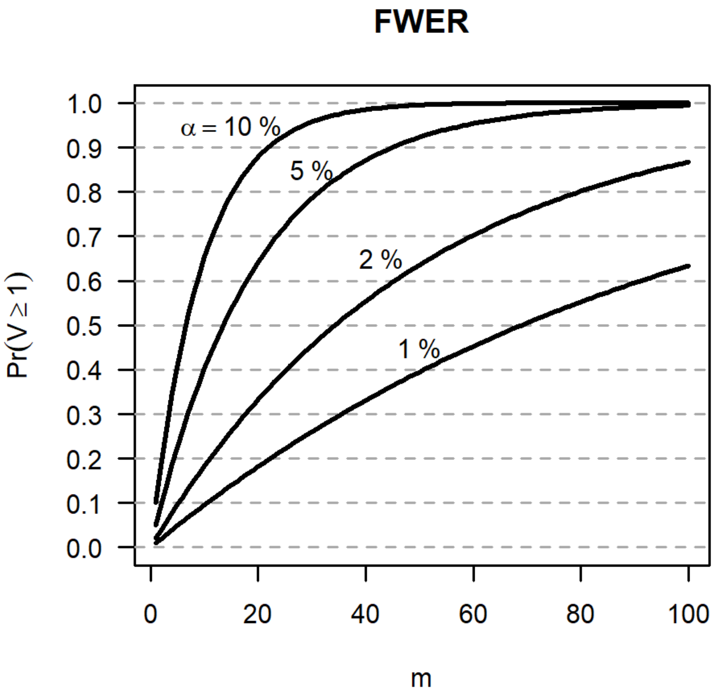
- **Example**: $\alpha = 5\%$ and $\mathrm{m} = 100$ we get $\mathbf{Pr}(V \geq 1) = \mathbf{0.994}$*

*Under dependency $\mathrm{Pr}(V \geq 1)$ is even higher.

Max-Planck-Institut für Meteorologie

How to control $\Pr(V \geq 1)$?

- controlling $\Pr(\mathrm{V} \geq 1) \leq \alpha$:
  - ***Bonferroni's one step*** procedure *(Bonferroni, 1935)*:
    reject $H_{0,i}$ if $p_i \leq \frac{\alpha}{n}$
    → **very conservative***
  - better methods (based on sorted p-values):
    Holm's step-down *(Holm, 1979)*:
    reject $H_{0,i}$ if $p_i > \frac{\alpha}{(n+1)-i}$
    Hochberg's step-up *(Hochberg, 1988)*:
    reject $H_{0,i}$ if $p_i \leq \frac{\alpha}{(n-i)+1}$
- all these methods are **suited for small *n!***
  → ***we need another approach***



* Increases Type II error;
very little power for large n.

Controlling the False Discovery Rate (FDR)

Benjamini, Y. and Hochberg, Y.,1995: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B*, 57, No. 1, 289-300.

→ Top 10 statistics publication of all time (>58k citations)!
→ took them 5 years and 3 journals to publish (Benjamini, 2010)

- Proportion of the rejected null hypothesis which are erroneously rejected is:

$$\boldsymbol{Q} = \begin{cases} \boldsymbol{V}/\boldsymbol{R} & \text{if } \boldsymbol{R} > 0 \\ 0 & \text{otherwise} \end{cases}$$

false discovery proportion (FDP); unobserved random variable

$$\boxed{\mathbf{FDR} = E(\boldsymbol{Q}) = E\left(\frac{\boldsymbol{V}}{\boldsymbol{R}}\middle|\boldsymbol{R} > 0\right) P(\boldsymbol{R} > 0)^{A}}$$

- we want to control $E(\boldsymbol{Q}) \leq \alpha_{FDR}{}^{B}$
(often you find q instead of α_{FDR})

	Declared non-significant (H_0)	Declared significant (H_A)	Total
True Null Hypothesis	**U** *Correct* ($1-\alpha$)	**V** *Type I error* (α)	m_0
Non-true Null Hypothesis	**T** *Type II error* (β)	**S** *Correct* ($1-\beta$, *power*)	$m_1 = m - m_0$
Total	*m* - **R**	**R**	*m*

FDR is the statistically expected fraction of erroneously rejected (discoveries) among all rejections

[A] There has to be at least one rejection of H_0. We cannot control $E(\boldsymbol{V}/\boldsymbol{R})$, but Benjamini and Hochberg (1995) show that it is possible to control $E(\boldsymbol{V}/\boldsymbol{R}|\boldsymbol{R}>0)P(\boldsymbol{R}>0)$.

[B] Also weak control of $\mathrm{FWER} = \Pr(V \geq 1)$: if all $H_0$are true ($m_0 = m$) the FDR is the same as the probability of making even one error:
$\mathrm{FDR} = E(1|\boldsymbol{R} > 0)P(\boldsymbol{R} > 0) = P(\boldsymbol{R} > 0) = \Pr(\boldsymbol{V} > 0) = \mathrm{FWER}$.

Controlling the False Discovery Rate (FDR)

Benjamini and Hochberg (1995):

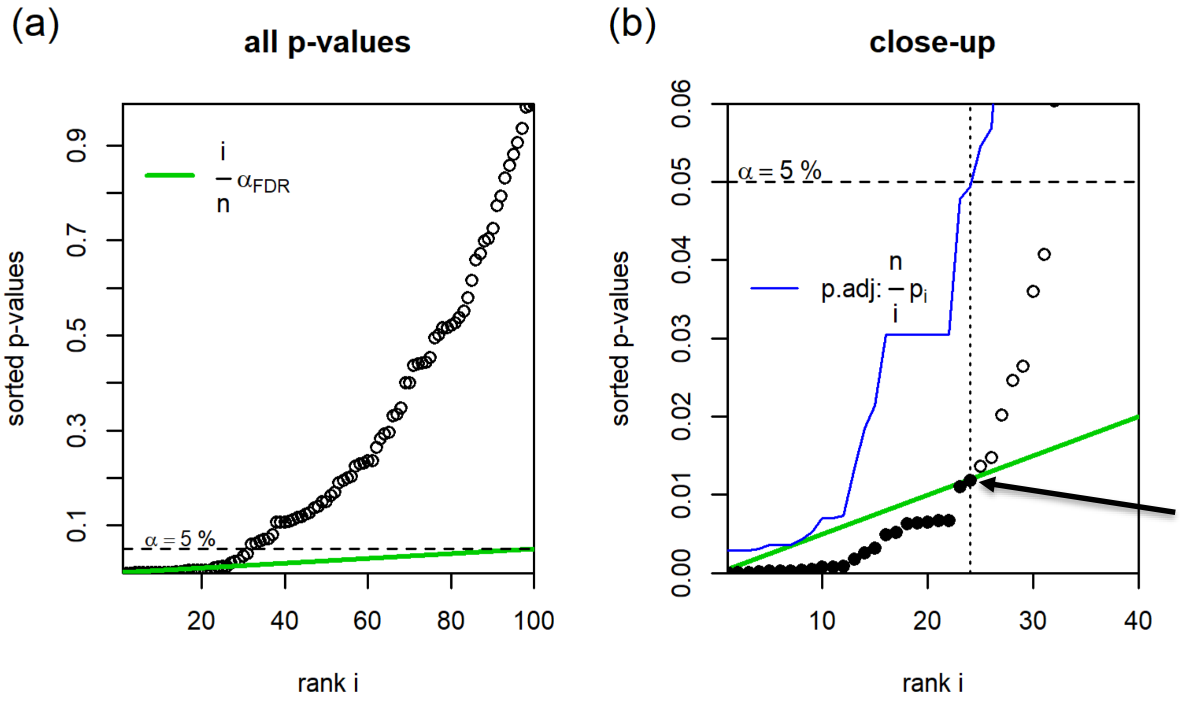
- FDR requires smaller p-values in order to reject local null hypotheses

- **algorithm**:
  1. sort p-values from $\boldsymbol{n}$ local tests $\boldsymbol{p_i}$ in ascending order with $i = 1, \ldots, n$
  2. denote sorted p-value as $p_{(1)} \leq p_{(2)} \leq \ \ldots \ \leq p_{(n)}$
  3. local H_0 are rejected if their p-values p_i are no larger than a threshold level $\boldsymbol{p^*_{FDR}}$:

- most commonly $\alpha_{FDR} = \alpha$
- α_{FDR} has to be chosen a-priori

$$p^*_{FDR} = \max_{i=1,\ldots,n} \left[p_{(i)} : p_{(i)} \leq \frac{i}{n} \alpha_{FDR} \right]$$

Controlling the False Discovery Rate (FDR)



$$p^*_{FDR} = \max_{i=1,\dots,n} \left[p_{(i)} : p_{(i)} \le \frac{i}{n} \alpha_{FDR} \right]$$

$n = 100$

$\alpha = 0.05$

$\alpha_{FDR} = 0.05$

$p^*_{FDR} = 0.012$

Controlling the False Discovery Rate (FDR)

$H_0: b = 0$
$H_A: b \neq 0$
local t tests

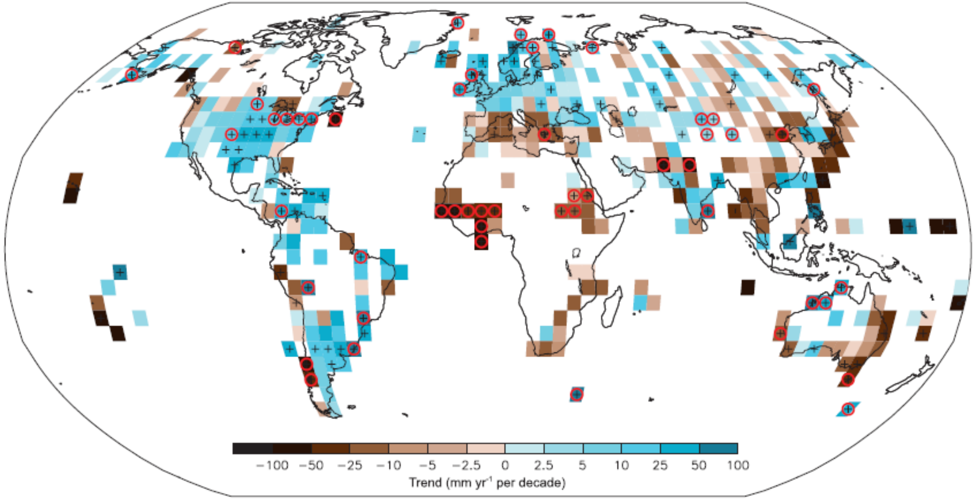


FIG. 7. Linear trends in annual precipitation during 1951–2010, based on data from the Global Historical Climatology Network (Vose et al. 1992). Grid elements with linear trends exhibiting local statistical significance at the $\alpha = 0.10$ level are been indicated by the plus signs, and those with p values small enough to satisfy the FDR criterion with $\alpha_{FDR} = 0.10$ [Eq. (3)] are indicated by the red circles. The figure has been modified from Hartmann et al. (2013, p. 203).

Wilks (2016)

Controlling FDR under dependency

- In practice, **test statistics are not independent**, e.g. **spatial correlation**

- **FDR robust under dependence**
(Ventura et al., 2004; Wilks, 2006; Wilks, 2016)
→ conservative for moderate to strong spatial correlation
→ account for temporal correlation by appropriate local testing procedure

- Several **modifications to FDR under dependence** (e.g. Benjamini and Yekutieli, 2001)
→ active research area

- Modifications usually available in software

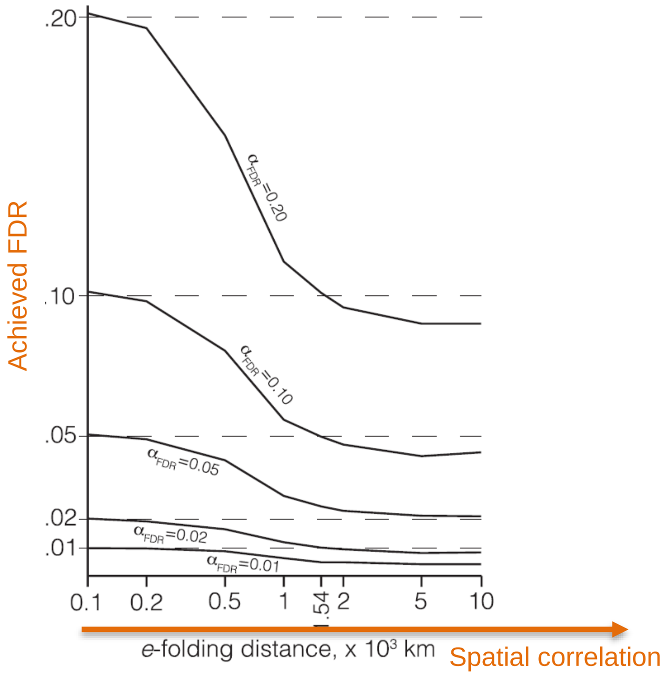


FIG. 4. Achieved global test levels (probabilities of rejecting true global null hypotheses) when using the FDR procedure, as a function of spatial correlation strength. For moderate and strong spatial correlation, approximately correct results can be achieved by choosing $\alpha_{FDR} = 2\alpha_{global}$.

Wilks (2016, mod.)

$$\alpha^*_{FDR} \sim 2\alpha_{FDR}$$

How to apply False Discovery Rate (FDR) procedure?

- FDR is easy to use:
  input: provide vector of p-values and q (α_{FDR})
  output: vector of adjusted p-values

  - **R**: `p.adjust(pvals,method="BH")` # returns $\mathrm{p.adj} = \frac{n}{i} p_i$
  - **Matlab**: `fdr_bh(pvals,q)`
  - **Python**: `statsmodels.stats.multitest.multipletests(pvals, alpha=0.05,method="fdr_bh")`

Conclusions

- preferable to **control the proportion of errors (FDR)** rather than the **probability of making one error (FWER)**

- **FDR** is the **best method available** to analyse multiple hypothesis test results
- valid for **all kind of tests**, even under dependence (e.g. spatial correlation). (Wilks, 2016; Wilks, 2006; Ventura et al. 2004).

- modifications for **FDR under dependency** (e.g. Benjamini and Yekutieli, 2001)
  → active research area
- **FDR ensures that no more than α_{FDR}% of *significant* results will be false positives instead of α% of all test results**

References

Benjamini, Y. and Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B*, 57, No. 1, 289-300, 1995.

Benjamini, Y. and Yekutieli, D.: The Control of the False Discovery Rate in Multiple Testing und Dependency. *The Ann. Stat.*, 20(4), 1165-1188, 2001.

Benjamini, Y. Discovering the false discovery rate. J. R. Statist. Soc. B., 72(4), 405-416, 2010.

Bonferroni, C. E.: Il calcolo delle assicurazioni su gruppi di teste. -In: Studi in Onore del Professore Salvatore Ortu Carboni. Rome, Italy, pp. 13-60, 1935.

Cowles, M. and Davis, C.: On the Origins of the .05 Level of Statistical Significance. *Amer. Psychologist*, 37(5), 553-558, 1982.

Fischer, R. A.: Statistical methods for research workers. Edinburgh: Oliver & Boyd, 1925.

Hedderich, J., and Sachs, L.: Angewandte Statistik: Methoden-sammlung mit R. Springer, 16. Aufl., ISBN 978-3-62-56656-5,1023 p., 2018.

Hochberg, Y.: A sharper Bonferroni procedure for multiple tests of significance. Biometrika, 75, 800-803, 1988.

Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Statist., 6, 65-70, 1979.

Katz, R. W.: Sir Gilbert Walker and a connection between El Niño and statistics. *Stat. Sci.*, 17, 97-112, https://doi.org/10.1214/ss/10237990000, 2002.

Katz, R. W. and Brown, B. G.: The problem of multiplicity in research on teleconnections. *Int. J. Climatol.*, 11, 505-513, https://doi.org/10.1002/joc.3370110504, 1991.

Livizey, R. E. and Chen, W. Y.: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* 111, 46-59, https://doi.org/10.1175/15200493(1983)111<0046:SFSAID>2.0.CO;2, 1983.

Taylor, G. I.: Gilbert Thomas Walker, 1868-1958. *Biographical Memoirs of Fellows of the Royal Society*, 8, 166-174, 1962.

Ventura, V., Paciorek, C. J., Risbey, J. S.: Controlling the Proportion of Falsely Rejected Hypotheses when Conducting Multiple Tests with Climatological Data. *J. Climate*, 17, 4343-4356, https://doi.org/10.1175/3199.1, 2004.

Von Storch, H.: A remark on Chervon-Schneider's algorithm to test significance of climate experiments with GCM's. J. Atmos. Sci.,39,187-189,https://doi.org/10.1175/1520-469(1982)039<0187:AROCSA>2.0.CO;2, 1982.

Walker, G. T.: Correlation in seasonal variations of weather. III. On the criterion for the reality of relationships of periodicities. *Mem. Indian. Meteor. Dept.*, 21 (9), 13-15, 1914.

Wilks, D. S.: On „Field Significance“ and the False Discovery Rate. *J. Appl. Meteor. Climatol.*, 45, 1181-1189, https://doi.org/10.1175/JAM2404.1, 2006.

Wilks, D. S.: The stippling shows statistically significant grid points. *Bull. Amer. Meteor. Soc.*, 97, 2263-2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.