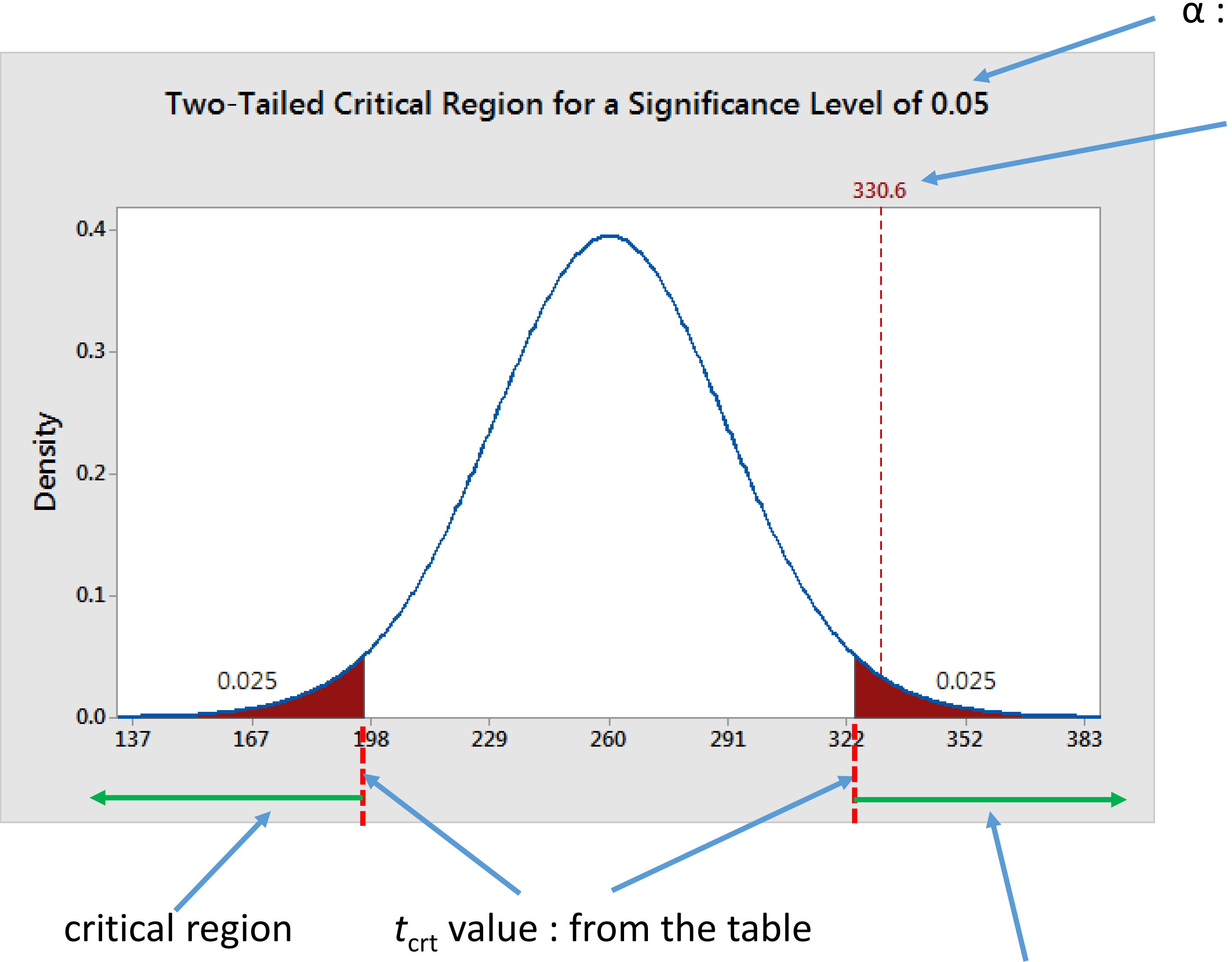
Topic No. 5

1. *p* - value
2. Receiver Operating Characteristics (ROC)
3. False Discovery Rate (FDR)
4. Visualization of networks (graphs)
5. Dashboard

p - value



α : shaded area

t_{obs} value : calculated t value

critical region

t_{crt} value : from the table
or from pdf(cdf)

critical region

p - value

In null hypothesis significance testing, the ***p*-value** is

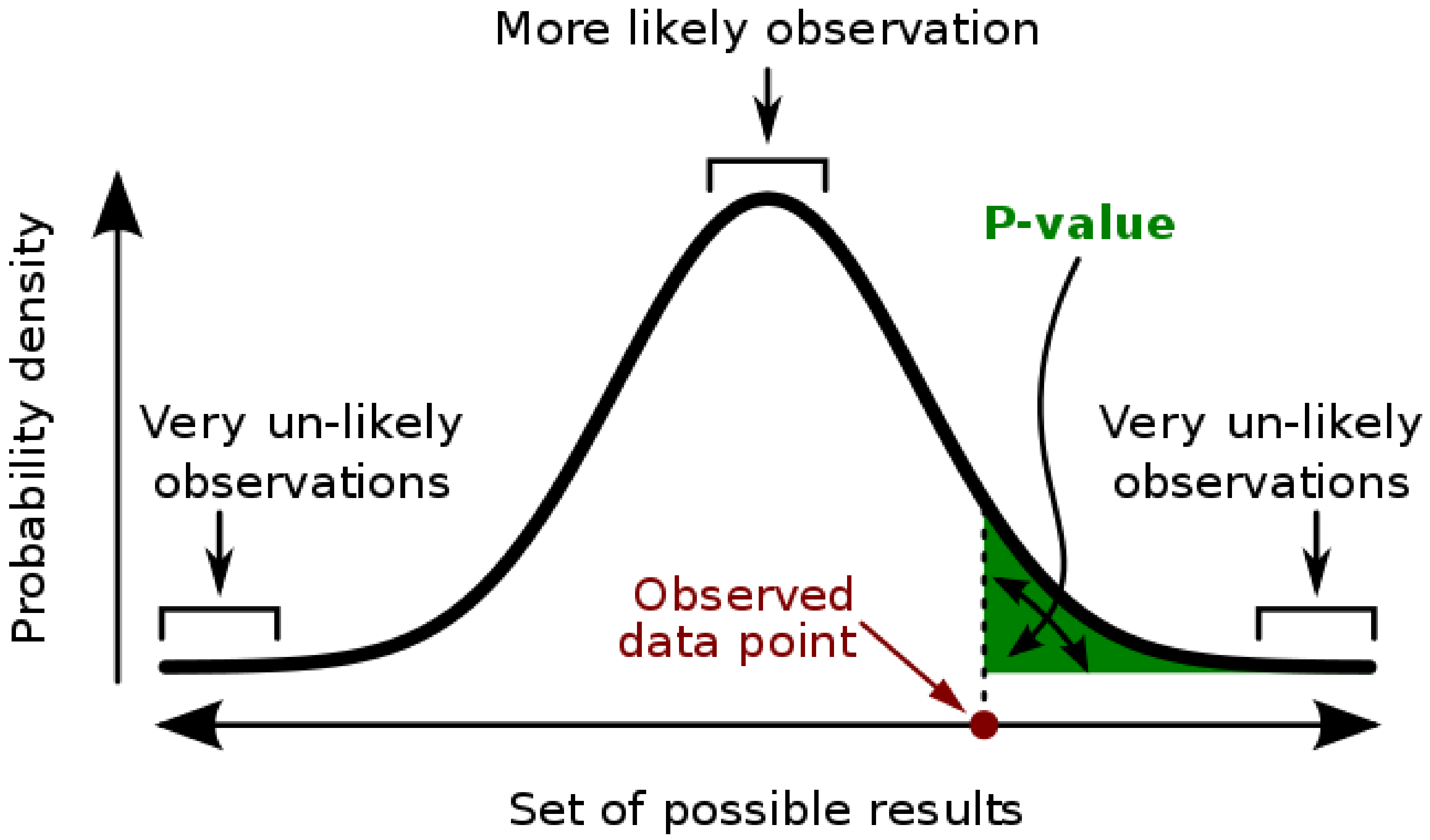
- the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.
- A very small *p*-value means that such an extreme observed outcome would be very unlikely under the null hypothesis.
- That is, given all other things being equal, smaller p-values are taken as **stronger evidence against the null hypothesis**
- Loosely speaking, rejection of the null hypothesis implies that there is sufficient evidence against it
- computing a *p*-value requires a null hypothesis, a test statistic (together with deciding whether the researcher is performing a one-tailed test or a two-tailed test), and data

p-value calculation (1-sided vs 2-sided)

Consider an observed test-statistic t from unknown distribution T. Then the *p*-value p is what the prior probability would be of observing a test-statistic value at least as "extreme" as t if null hypothesis H_0 were true. That is:

- ➢ $p = Pr\ (T \geq t \mid H_0)$ for a one-sided right-tail test,
- ➢ $p = Pr\ (T \leq t \mid H_0)$ for a one-sided left-tail test,
- ➢ $p = 2 \text{ x } \min \{Pr\ (T \geq t \mid H_0)\ ,\ Pr\ (T \leq t \mid H_0)\}$ for a two-sided test.
  - $p = Pr\ (|T| \geq |t| \mid H_0)$ if distribution T is symmetric about zero

p - value



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

p - value

Example

Suppose the experimental of coin toss.

- The **results** show the coin turning up heads 14 times out of 20 total flips. The full data *X* would be a sequence of twenty times the symbol ‘head’ or ‘tail’. The statistic on which one might focus, could be the total number *T* of heads.
- The **null hypothesis** is that the coin is fair, and coin tosses are independent of one another.
- If a right-tailed test is considered, if one is interested in the possibility that the coin is biased towards falling heads, then the *p*-value of this result is the chance of a fair coin landing on heads *at least* 14 times out of 20 flips.

p - value

- Null hypothesis (H_0): The coin is fair, with Prob(heads) = 0.5
- Test statistic: Number of heads
- Alpha level (designated threshold of significance): 0.05
- Observation O: 14 heads out of 20 flips

p-value of this result is the chance of a fair coin landing on heads *at least* 14 times out of 20 flips can be computed from binomial coefficients as

$$\begin{aligned}&\text{Prob}(14\text{ heads}) + \text{Prob}(15\text{ heads}) + \cdots + \text{Prob}(20\text{ heads})\\&\qquad= \frac{1}{2^{20}}\left[\binom{20}{14} + \binom{20}{15} + \cdots + \binom{20}{20}\right] = \frac{60{,}460}{1{,}048{,}576} \approx 0.058\end{aligned}$$

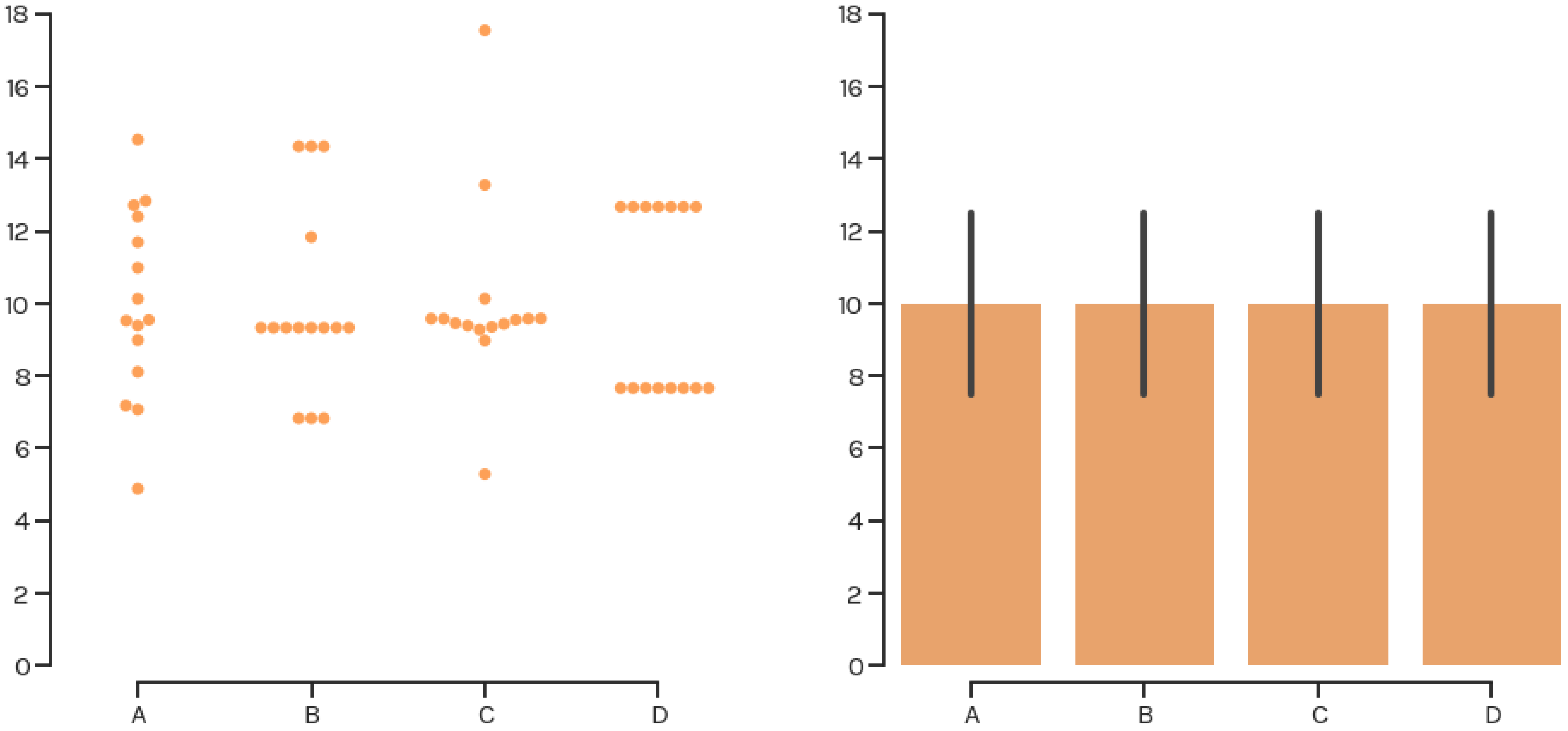
Misinterpretation of *p*-value

1. **The *p*-value is *not* the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.** A *p*-value can indicate the degree of compatibility between a dataset and a particular hypothetical explanation (such as a null hypothesis). Specifically, the *p*-value can be taken as the prior probability of obtaining an effect that is at least as extreme as the observed effect, given that the null hypothesis is true. This should not be confused with the posterior probability that the null hypothesis is true given the observed
2. **The *p*-value is *not* the probability that the observed effects were produced by random chance alone.** The *p*-value is computed under the assumption that a certain model, usually the null hypothesis, is true. This means that the *p*-value is a statement about the relation of the data to that hypothesis.
3. **The 0.05 significance level is merely a convention.** The 0.05 significance level (alpha level) is often used as the boundary between a statistically significant and a statistically non-significant *p*-value. However, this does not imply that there is generally a scientific reason to consider results on opposite sides of any threshold as qualitatively different.
4. **The *p*-value does not indicate the size or importance of the observed effect.** A small *p*-value can be observed for an effect that is not meaningful or important. In fact, the larger the sample size, the smaller the minimum effect needed to produce a statistically significant *p*-value).

Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on p-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

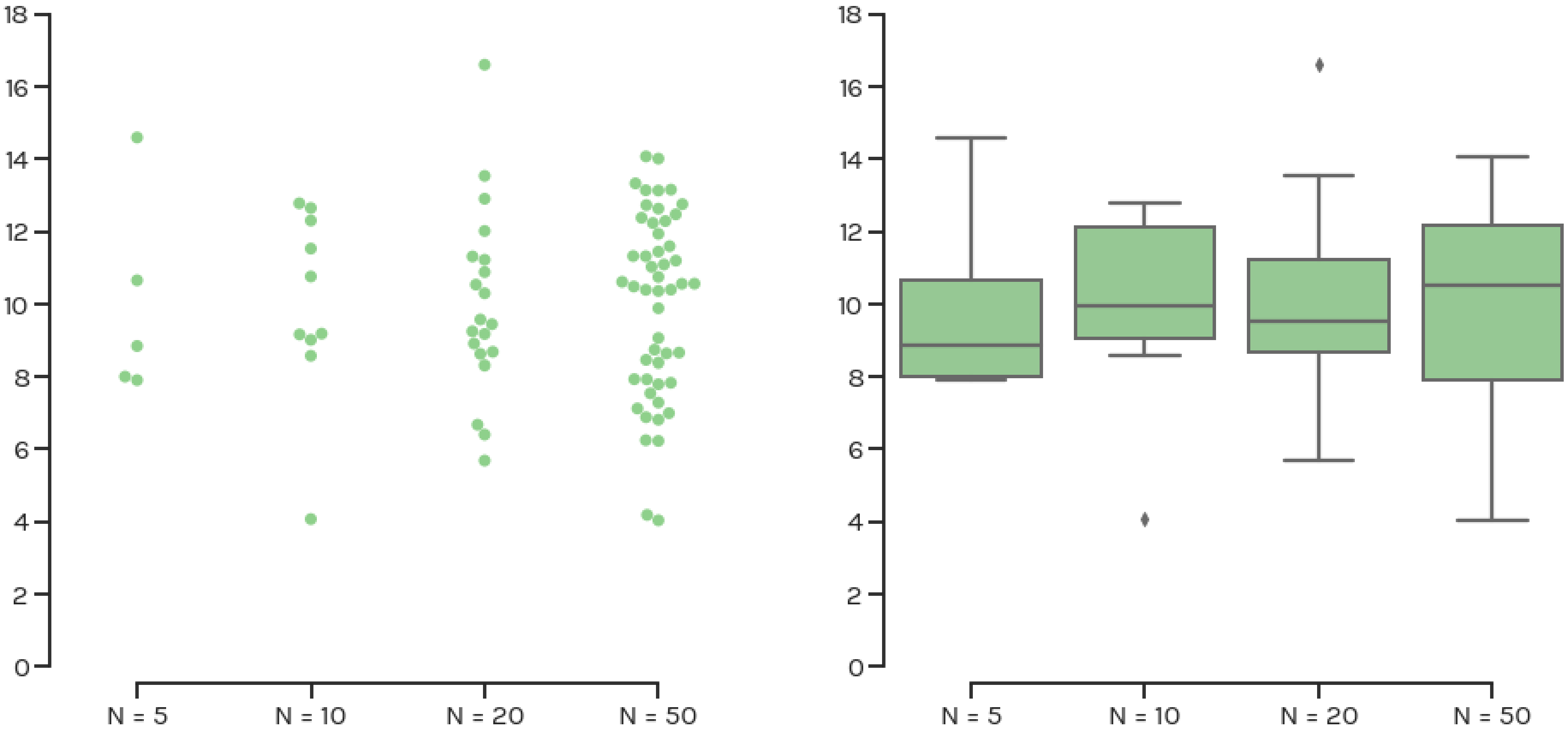
Visualization ideas

Barplots do not distinguish samples with wildly different distributions



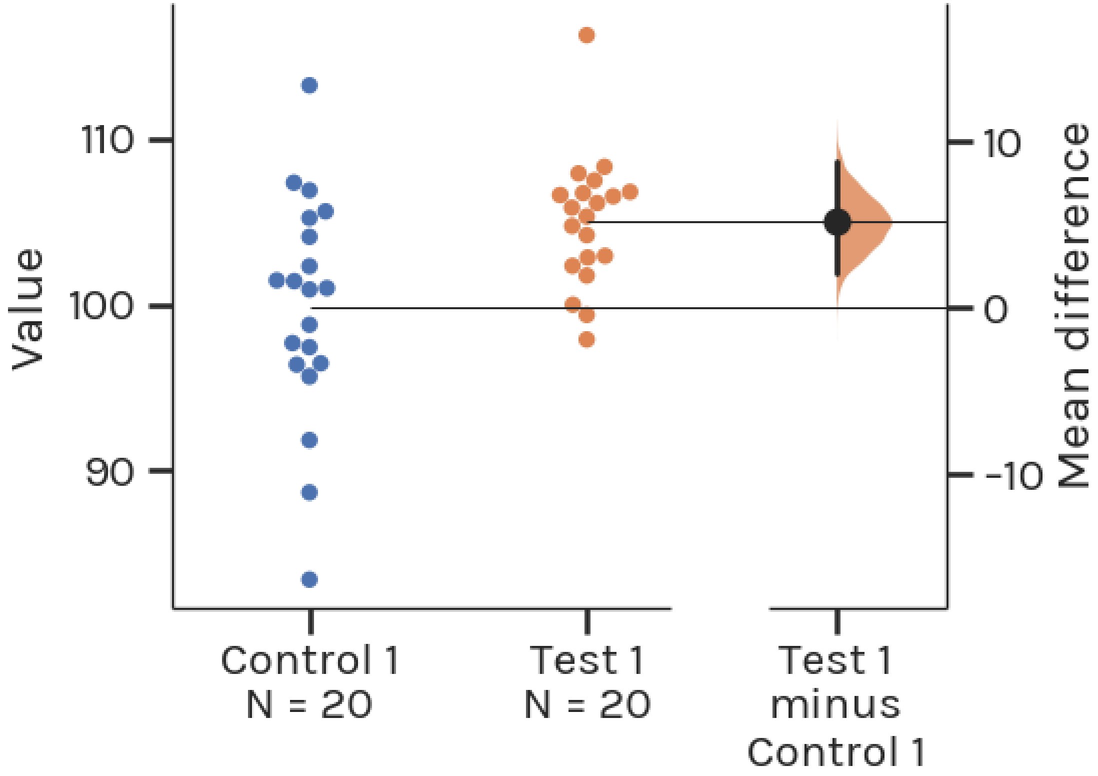
Visualization ideas

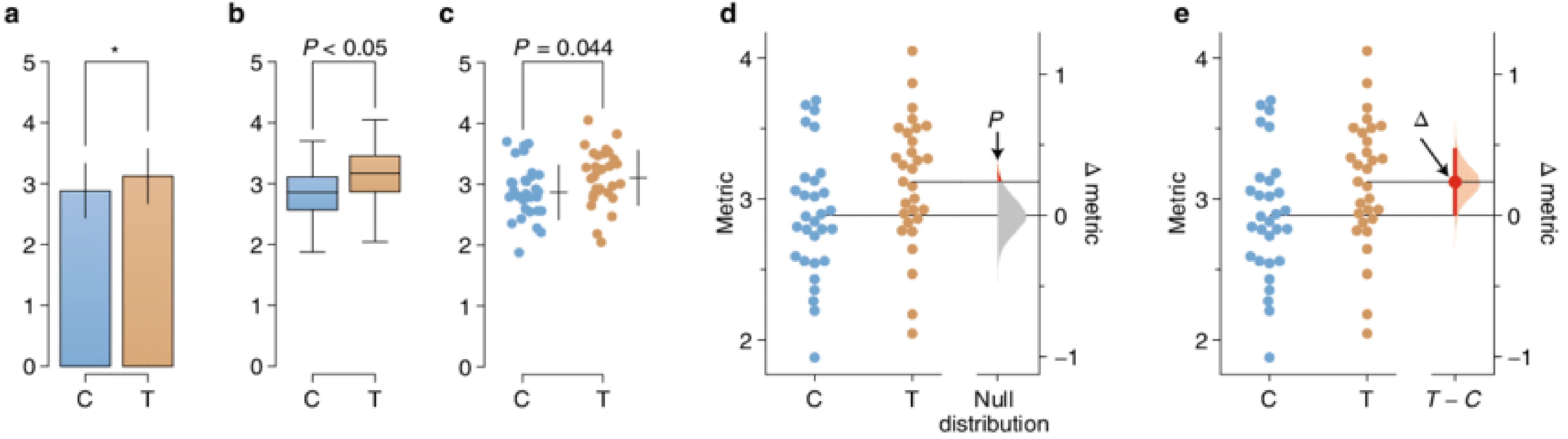
Boxplots do not convey the sense of sample size



Gardner-Altman estimation plot

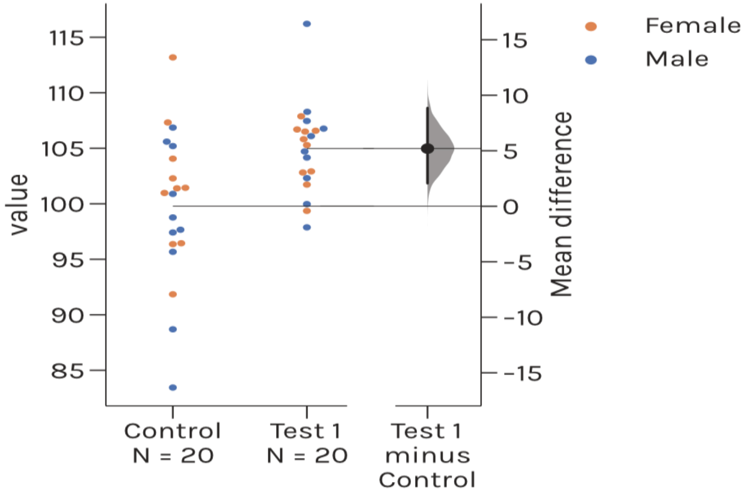
1. It presents all datapoints as a *swarmplot*, which orders each point to display the underlying distribution.
2. It presents the effect size as a *bootstrap 95% confidence interval* (95% CI) on a separate but aligned axes. where the effect size is displayed to the right of the war data, and the mean of the test group is aligned with the effect size.

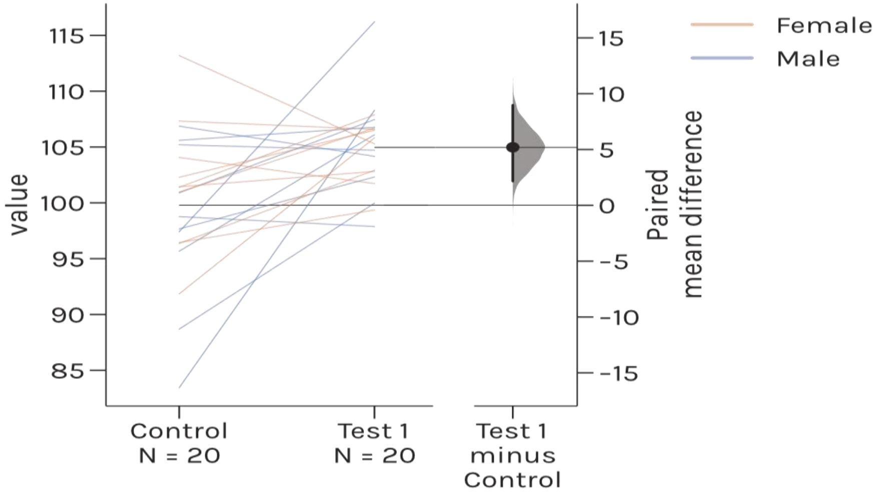


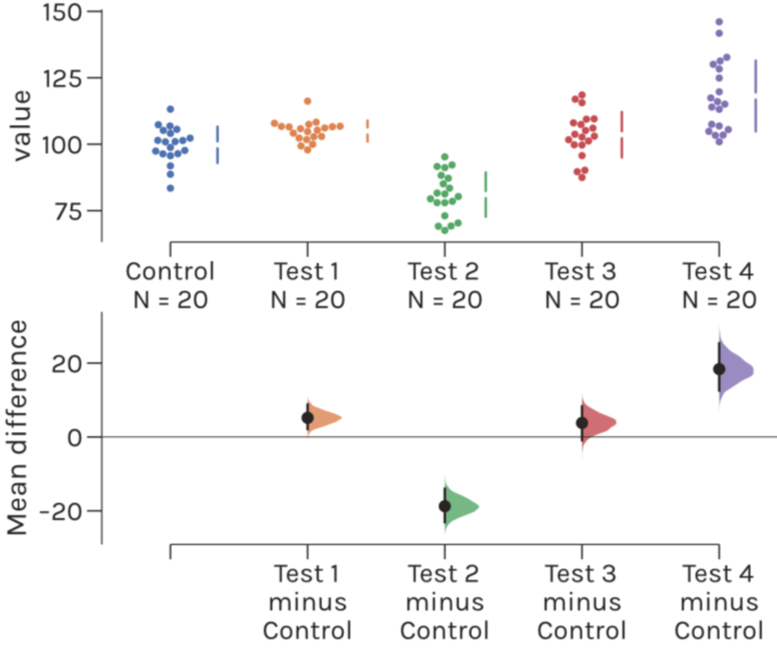


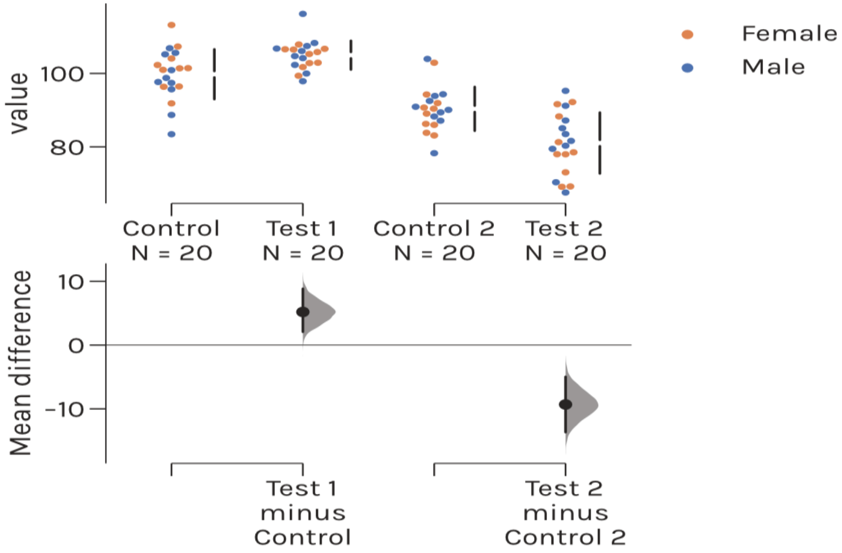
a, Two-group data with control (C) and test (T) groups presented in a bar chart. **b**, The same data presented as a box plot. **c**, This scatter plot shows the observed values along with descriptive statistics (mean and s.d.) but does not illustrate effect size. **d**, A two-group comparison with complete visualization of the NHST perspective. The filled curve on the difference axis indicates the distribution of the mean difference under the null hypothesis. Here the null distribution was constructed with permutation of observed data. By definition, this distribution has a mean difference of zero. The area of the red segment indicates the P value (for one-sided testing). **e**, An estimation graphic using the difference axis to display an effect size, here the mean difference (Δ). The curve indicates the resampled distribution of Δ, given the observed data. Horizontally aligned with the mean of the test group, Δ is indicated by the red circle. The 95% confidence interval of Δ is illustrated by the red vertical line. We propose calling such graphics 'Gardner-Altman plots', after their originators.

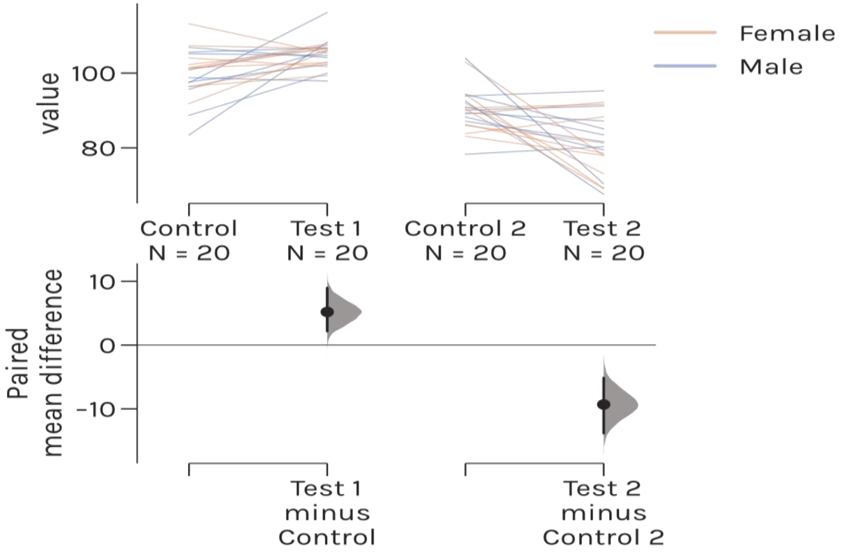
https://www.researchgate.net/figure/The-evolution-of-two-group-data-graphics-a-Two-group-data-with-control-C-and-test_fig1_333884529











https://github.com/ACCLAB/DABEST-python

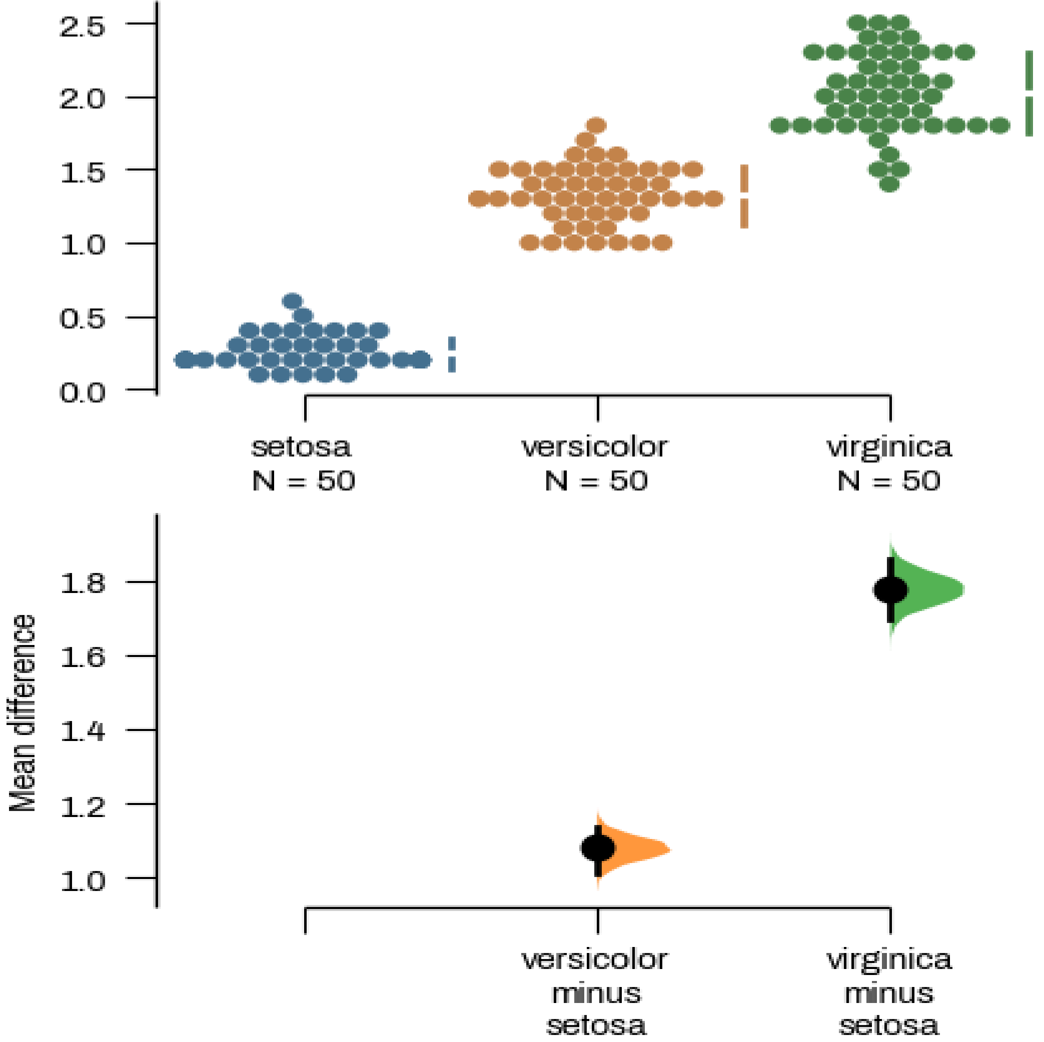
```
pip install --upgrade dabest

import pandas as pd
import dabest

# Load the iris dataset. Requires internet access.
iris = pd.read_csv("https://github.com/mwaskom/seaborn-data/raw/master/iris.csv")

# Load the above data into `dabest`.
iris_dabest = dabest.load(data=iris, x="species", y="petal_width",
                idx=("setosa", "versicolor", "virginica"))

# Produce a Cumming estimation plot.
iris_dabest.mean_diff.plot()
```



https://github.com/ACCLAB/DABEST-python

https://acclab.github.io/DABEST-python-docs/tutorial.html

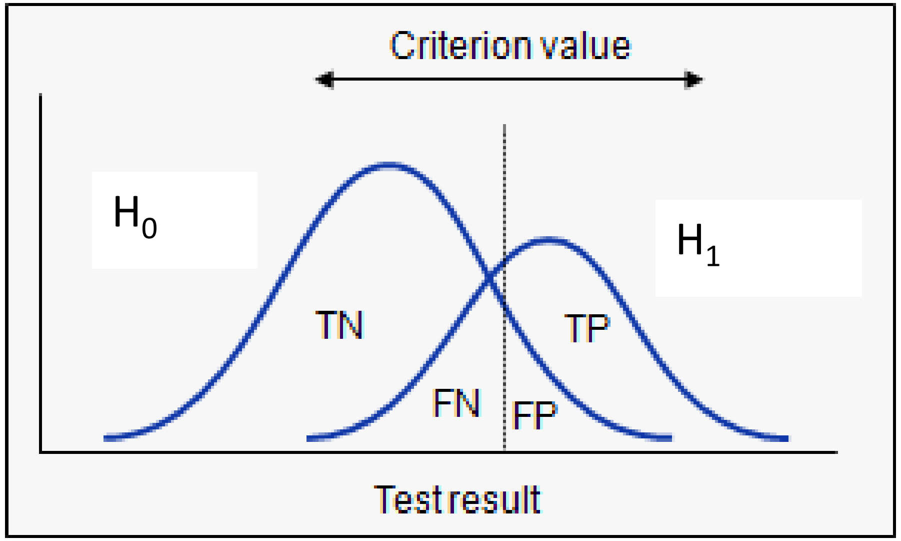
Model Selection: ROC Curves

- **ROC** (Receiver Operating Characteristics) curves: for visual comparison of classification models
- Originated from signal detection theory
- Shows the trade-off between the true positive rate and the false positive rate
- The area under the ROC curve (AUC) is a measure of the accuracy of the model
- Rank the test tuples in decreasing order: the one that is most likely to belong to the positive class appears at the top of the list
- The closer to the diagonal line (i.e., the closer the area is to 0.5), the less accurate is the model
- ROC curves can be used to compare the classification performances of two or more tests

Receiver Operating Characteristics

- Consider the results of a particular test in two populations, one population with a disease (H_1), the other population without the disease (H_0)
- You will rarely observe a perfect separation between the two groups.
- Indeed, the distribution of the test results will overlap, as shown in the following figure.

	H_0	H_1
Reject H_0	*FP*	*TP*
Accept H_0	*TN*	*FN*

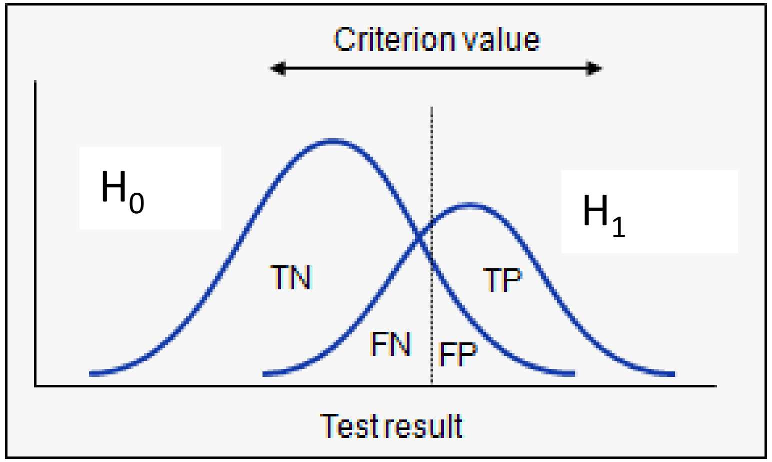


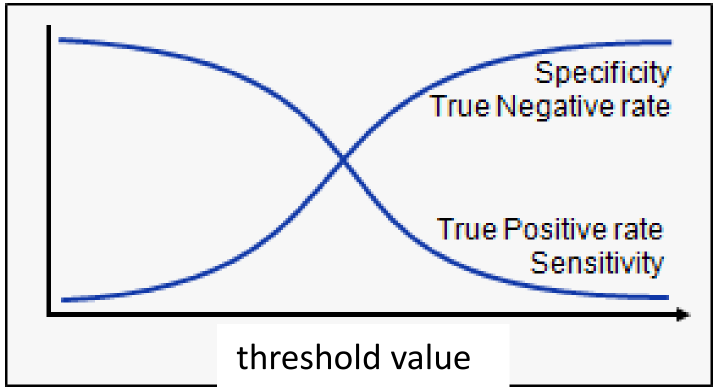
Receiver Operating Characteristics

- For every possible cut-off point or criterion value you select to discriminate between the two populations
- There will be some cases with
  - the disease correctly classified as positive (TP = True Positive fraction)
  - the disease will be classified negative (FN = False Negative fraction).
  - case without the disease will be correctly classified as negative (TN = True Negative fraction)
  - case without the disease will be classified as positive (FP = False Positive fraction).

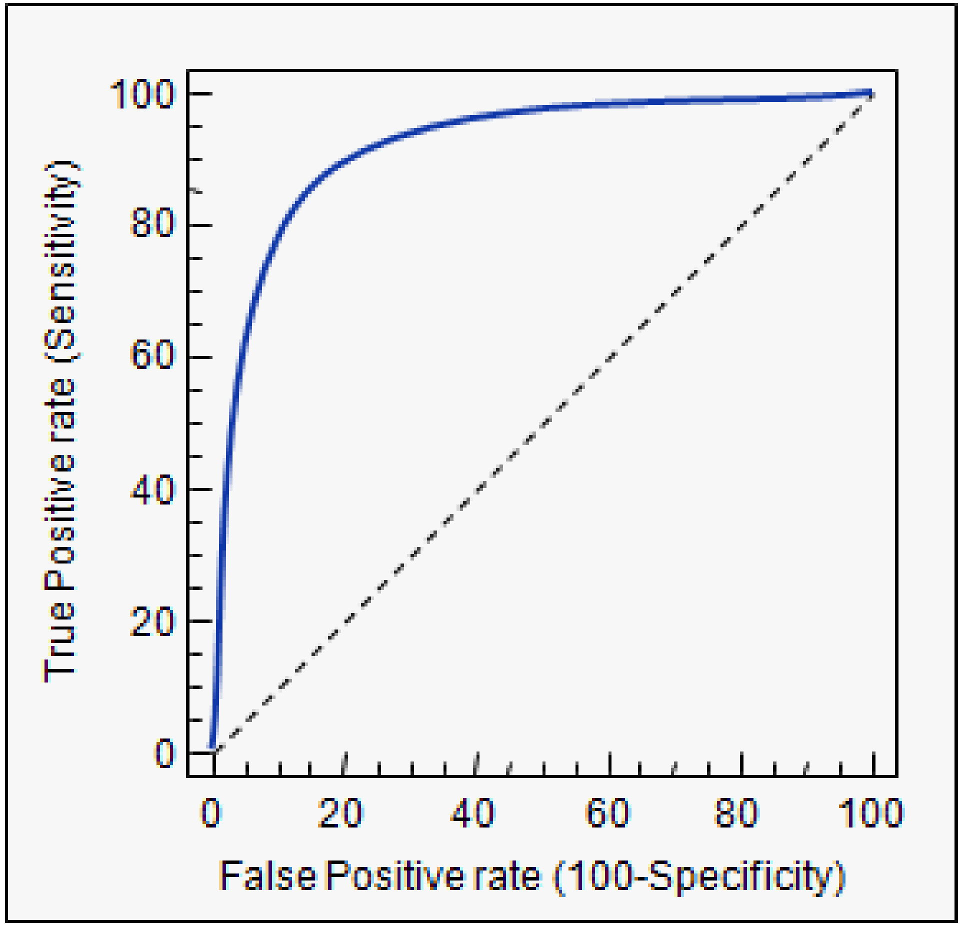
Receiver Operating Characteristics

- higher threshold value
  - the false positive fraction will decrease with increased specificity
  - true positive fraction and sensitivity will decrease
- lower threshold value
  - the false positive fraction will increase with decreased specificity and true negative fraction
  - true positive fraction and sensitivity will increase

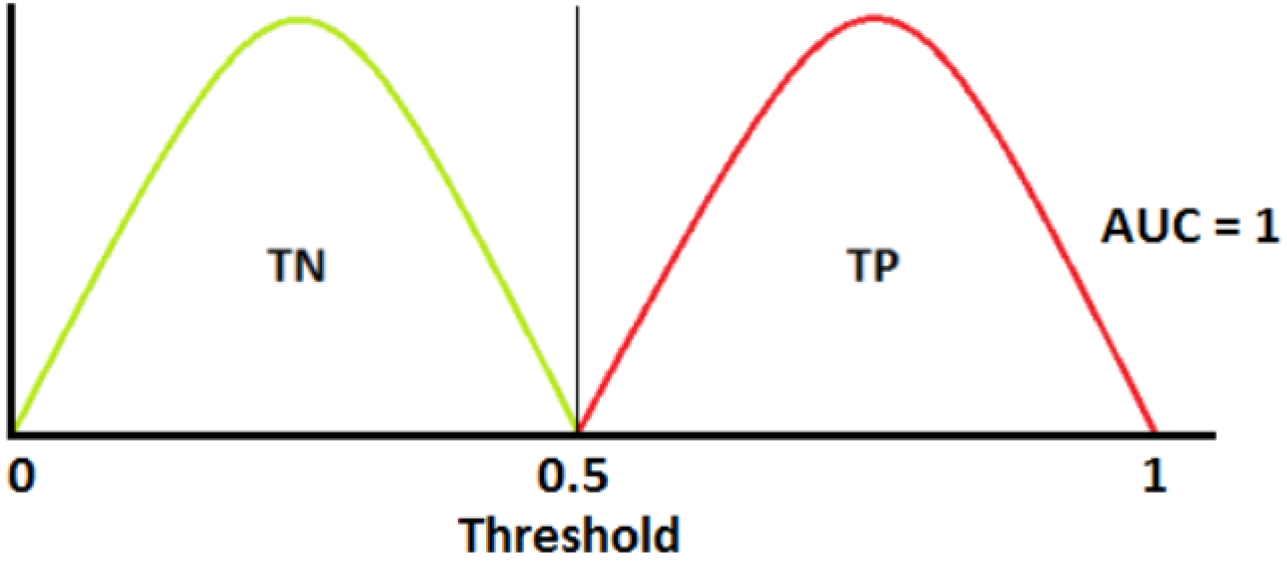




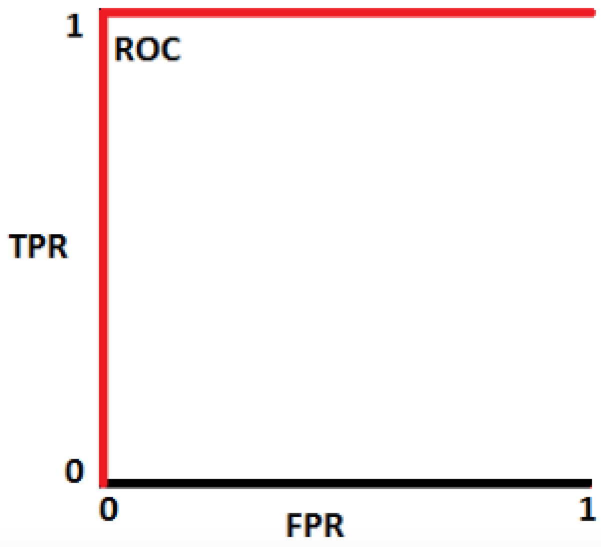
Receiver Operating Characteristics



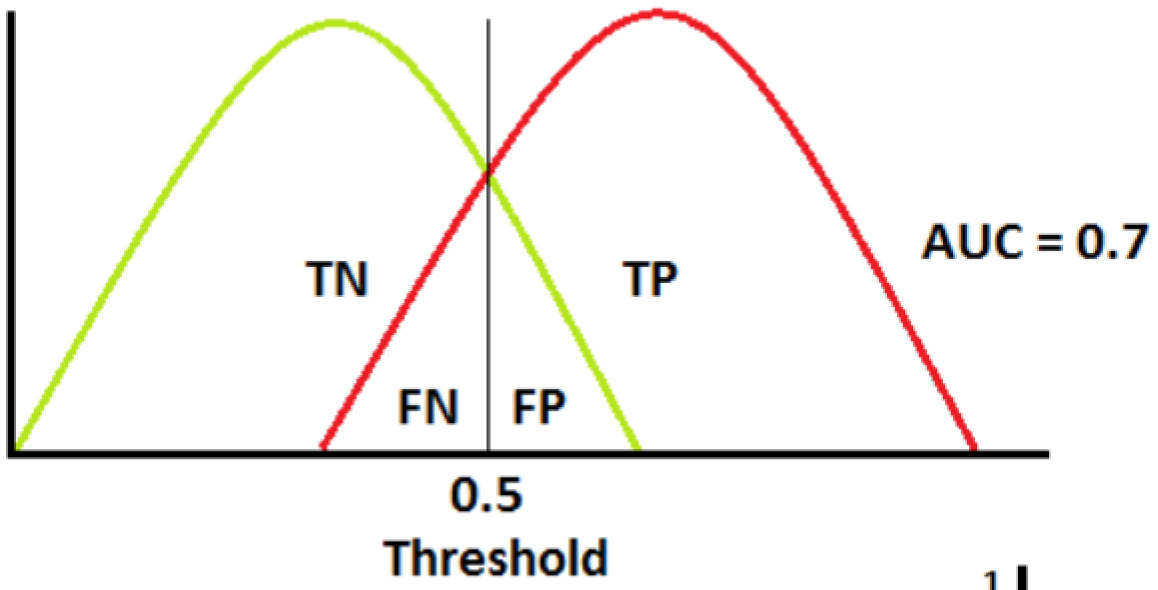
Receiver Operating Characteristics



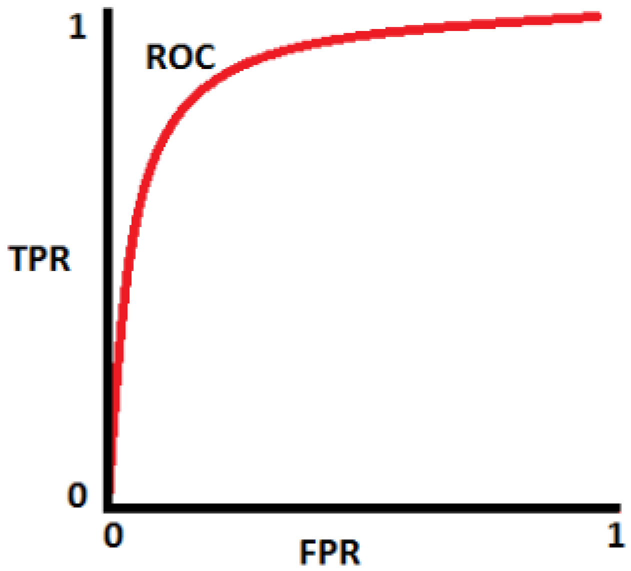
- This is an ideal situation. When two curves don't overlap at all means model has an ideal measure of separability.
- It is perfectly able to distinguish between positive class and negative class.



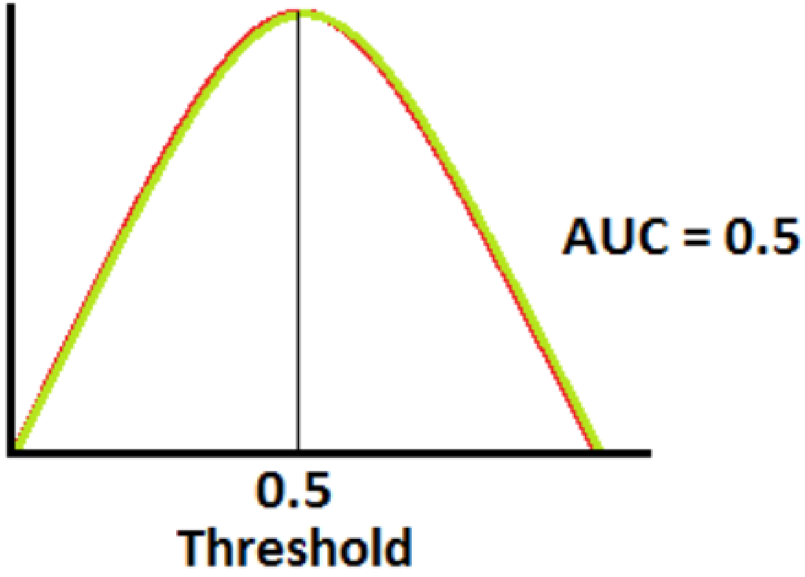
Receiver Operating Characteristics



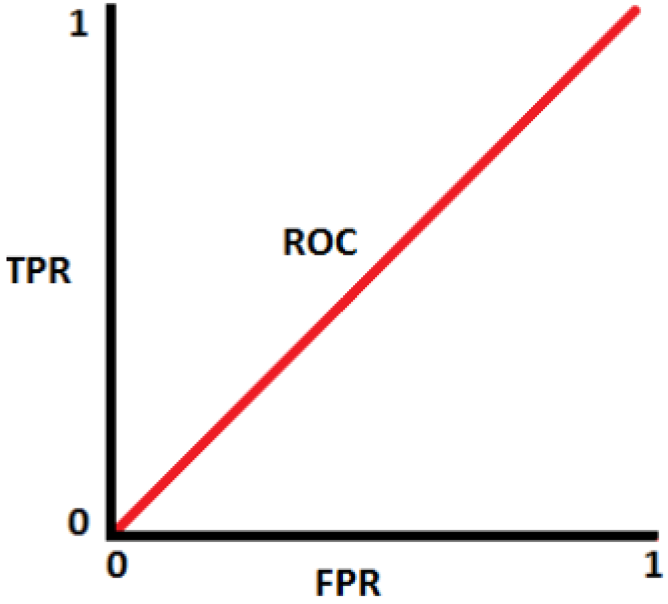
- When two distributions overlap
- Depending upon the threshold, we can minimize or maximize them.
- When AUC is 0.7, it means there is 70% chance that model will be able to distinguish between positive class and negative class.



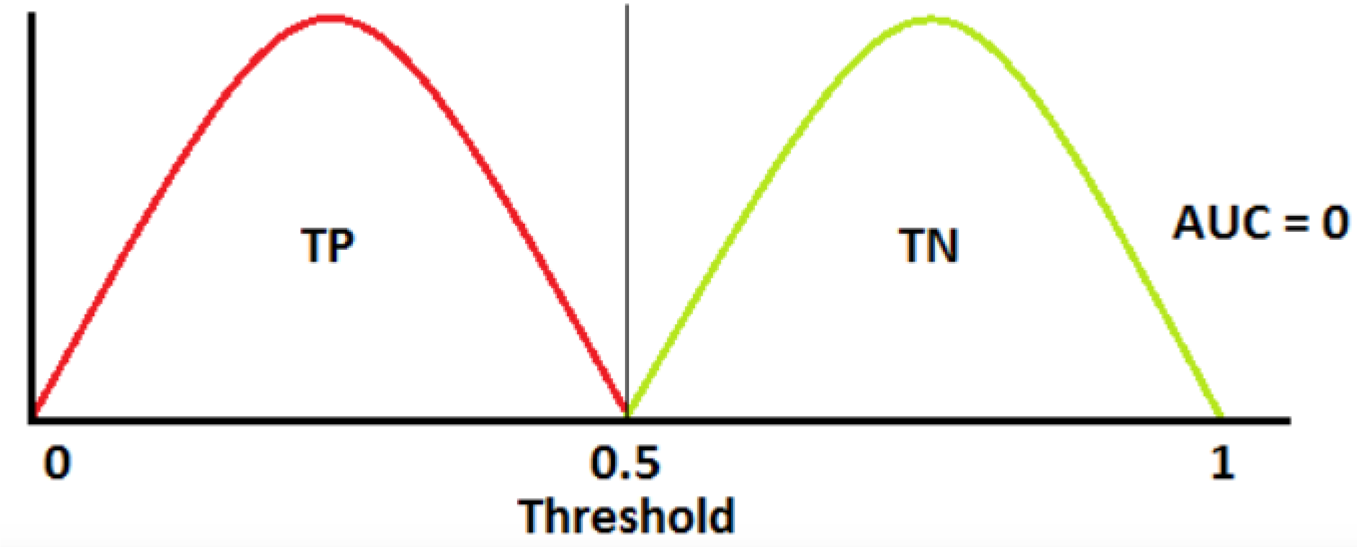
Receiver Operating Characteristics



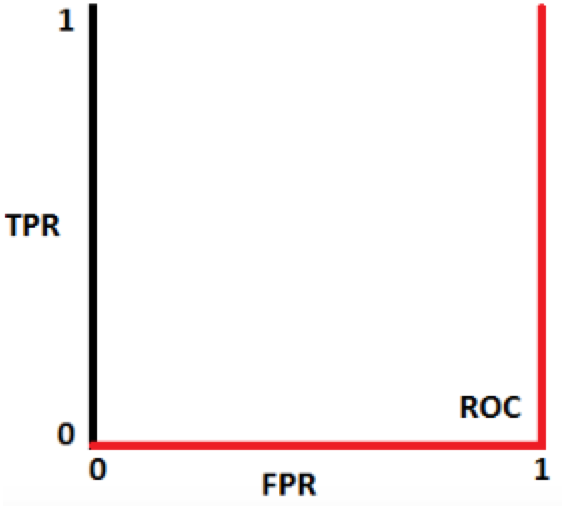
- When two distributions fully overlap
- When AUC is approximately 0.5, model has no discrimination capacity to distinguish between positive class and negative class.



Receiver Operating Characteristics



- When AUC is approximately 0, model is actually reciprocating the classes.
- It means, model is predicting negative class as a positive class and vice versa.



Receiver Operating Characteristics

- In multi-class model, we can plot N number of AUC ROC Curves for N number classes using One vs Rest methodology.
- Example: If you have **three** classes named **X, Y** and **Z**, you will have one ROC for X classified against Y and Z, another ROC for Y classified against X and Z, and a third one of Z classified against Y and X.

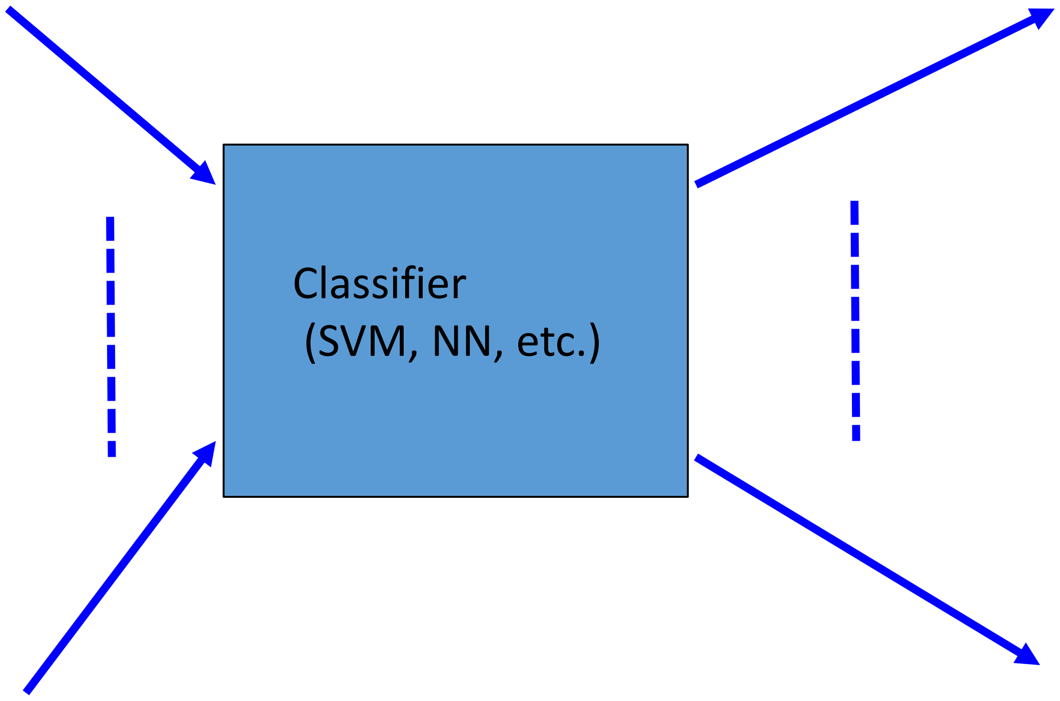
Receiver Operating Characteristics

- Example
  - Assume you have a classifier (SVM, NN, etc.) with 2 classes A and B
  - 10 test data for the classifier
  - Each test data point will generate a probability to predict the class. For example, 0.9 means high probability to be classified as A (or low probability to be classified as B), 0.1 for low probability to be classified as A (or high probability to be classified as A)
  - SVM generates the following 10 probabilities : 0.95(A), 0.88(A), 0.85(A), 0.74(A), 0.52(B), 0.49(A), 0.45(A), 0.23(B), 0.19(B), 0.11(B)

Receiver Operating Characteristics

Test data
with class label

$\mathbf{x}_1$ (A)
$\mathbf{x}_2$ (A)
$\mathbf{x}_3$ (A)
$\mathbf{x}_4$ (A)
$\mathbf{x}_5$ (B)
$\mathbf{x}_6$ (A)
$\mathbf{x}_7$ (A)
$\mathbf{x}_8$ (B)
$\mathbf{x}_9$ (B)
$\mathbf{x}_{10}$ (B)



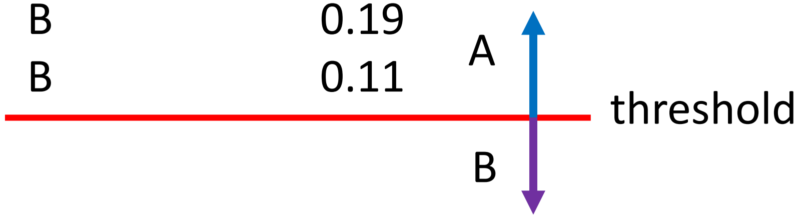
Classifier
(SVM, NN, etc.)

Predicted
probability

y_1	0.95
y_2	0.88
y_3	0.85
y_4	0.74
y_5	0.52
y_6	0.49
y_7	0.45
y_8	0.23
y_9	0.19
y_{10}	0.11

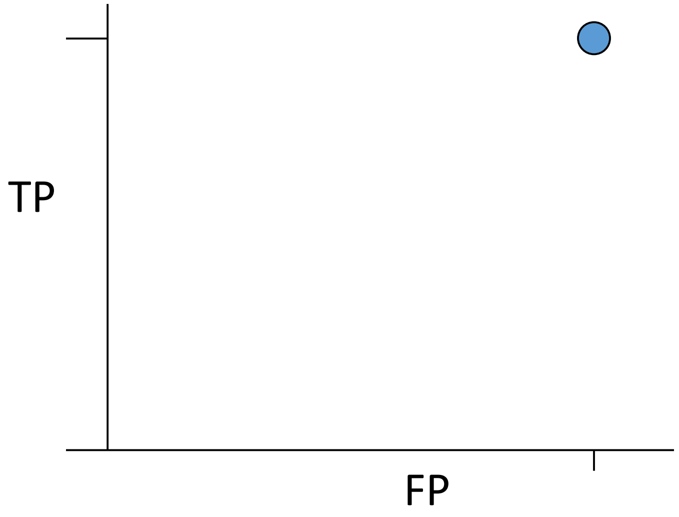
Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11



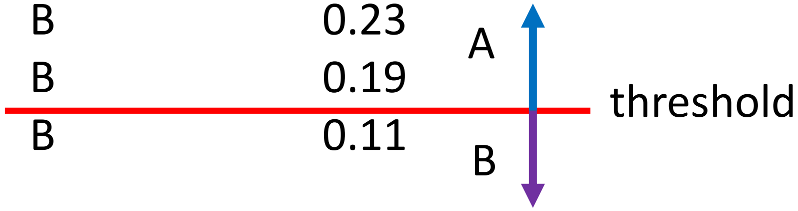
False Positive =4/4=1

True Positive =6/6=1



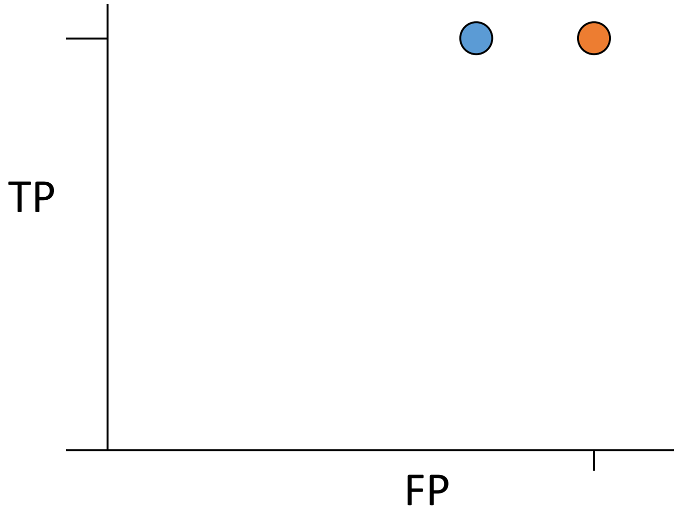
Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11



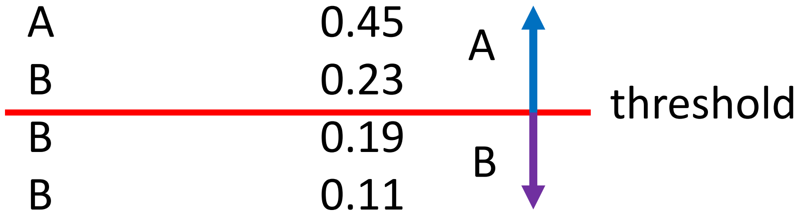
False Positive =3/4

True Positive =6/6=1



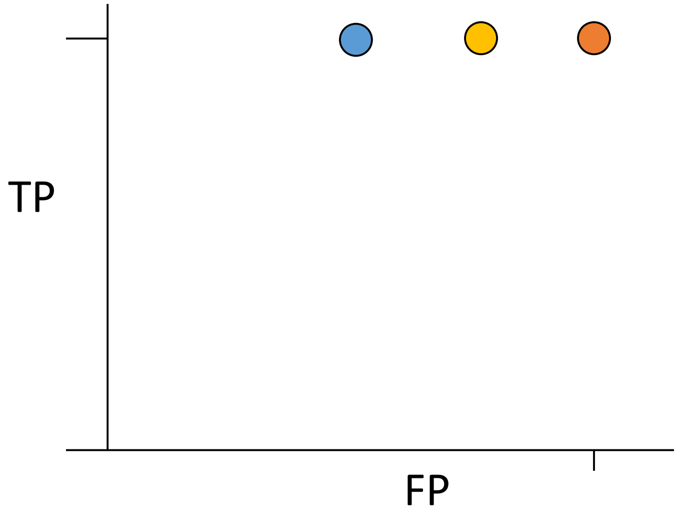
Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11



False Positive =2/4

True Positive =6/6=1



Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11

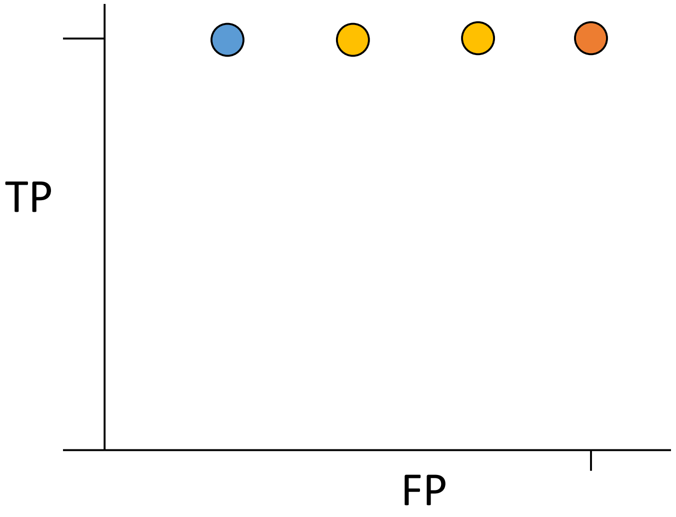
A

threshold

B

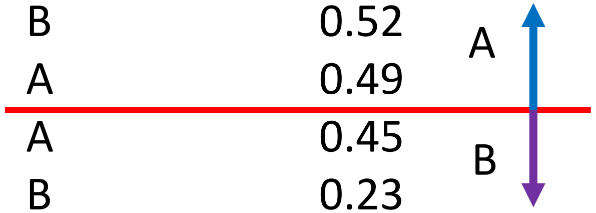
False Positive =1/4

True Positive =6/6=1



Receiver Operating Characteristics

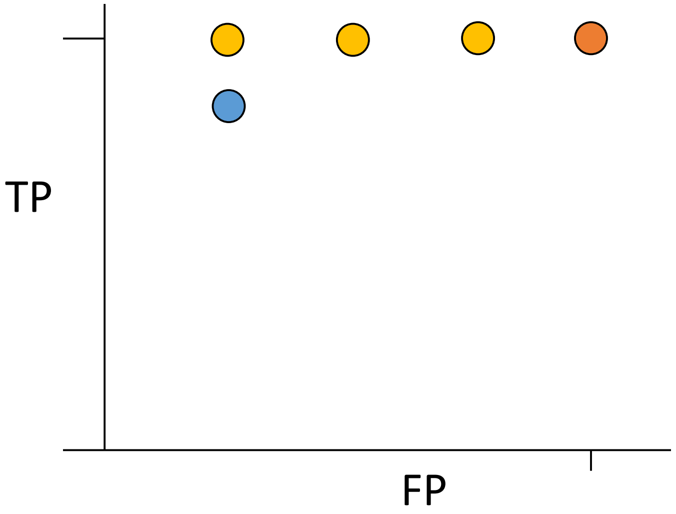
Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11



threshold

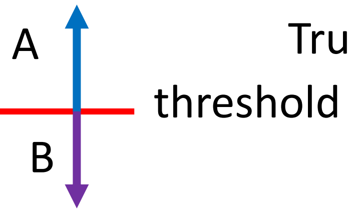
False Positive =1/4

True Positive =5/6



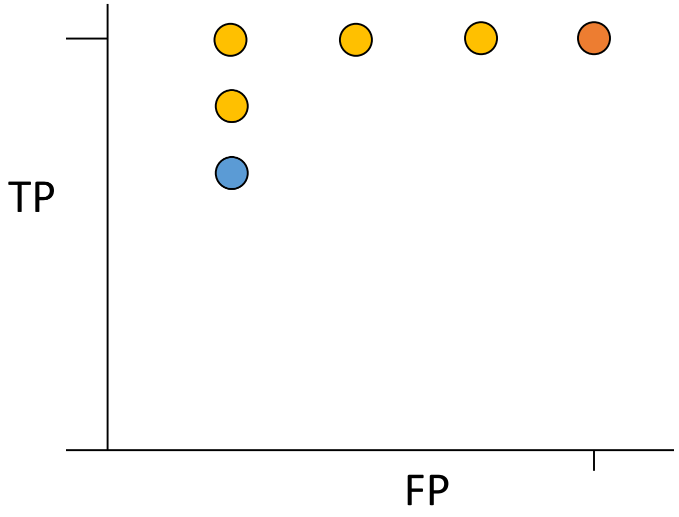
Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11



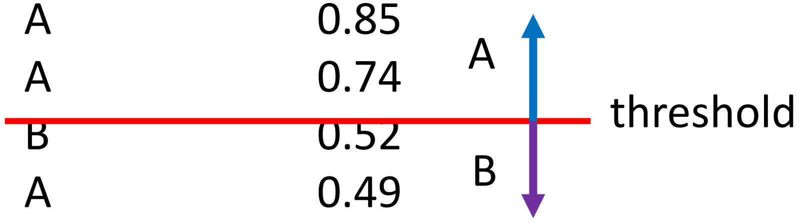
False Positive =1/4

True Positive =4/6



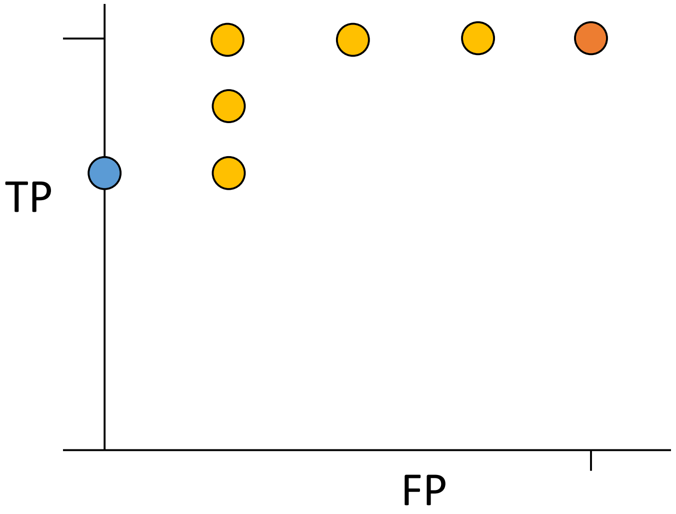
Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11

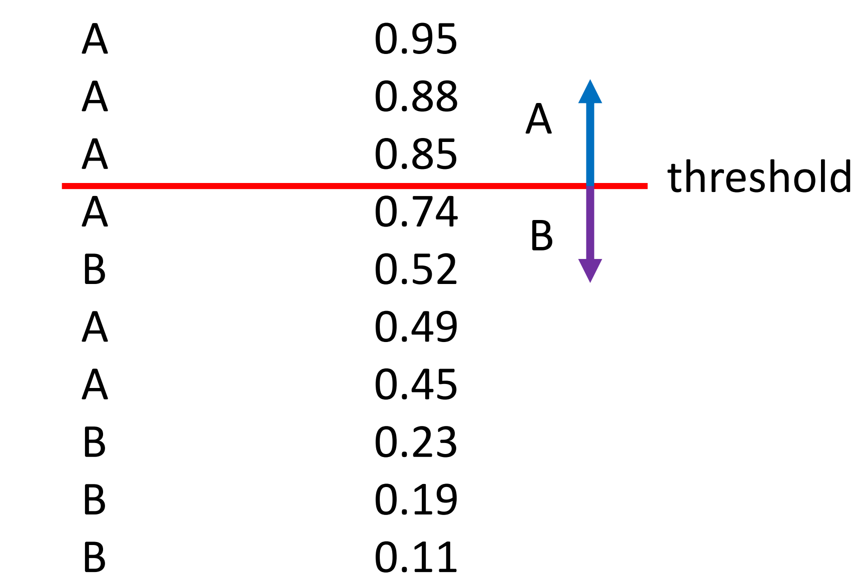


False Positive =0/4

True Positive =4/6



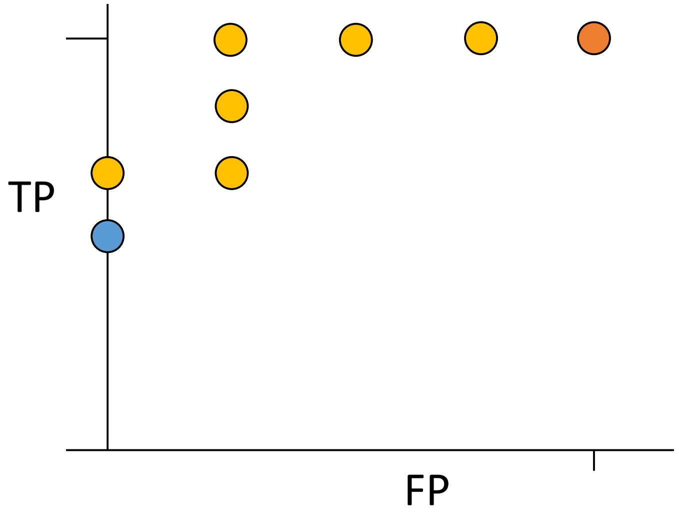
Receiver Operating Characteristics



Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11

False Positive =0/4

True Positive =3/6



Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11

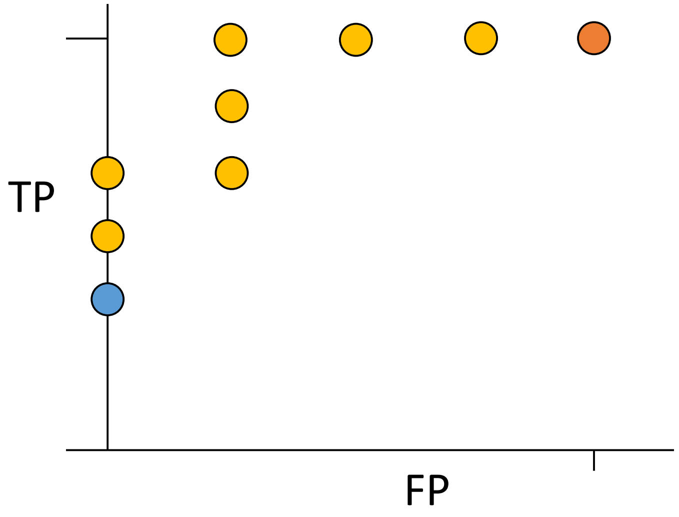
A

threshold

B

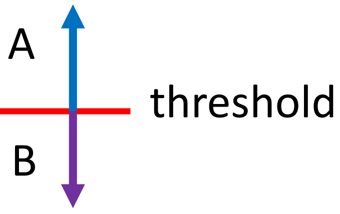
False Positive =0/4

True Positive =2/6



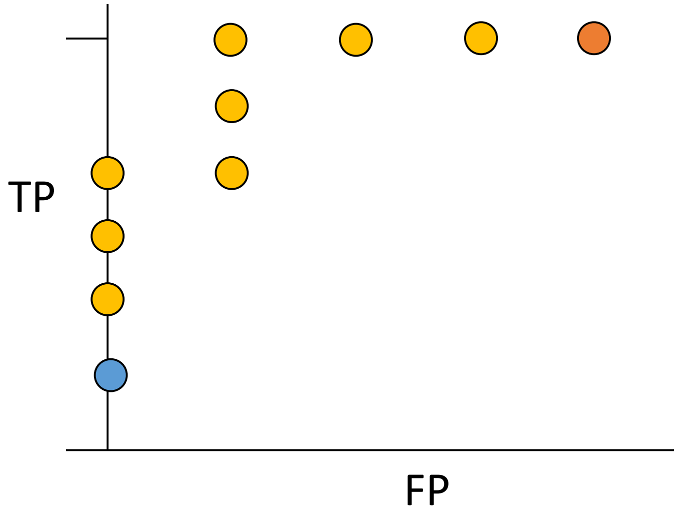
Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11



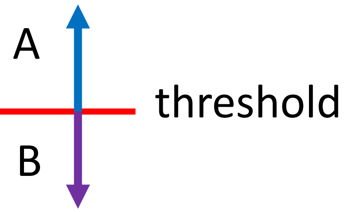
False Positive =0/4

True Positive =1/6



Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11

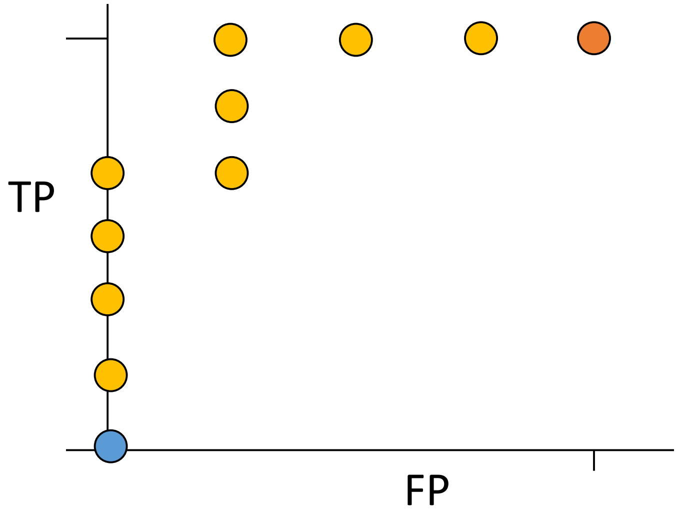

A

threshold

B

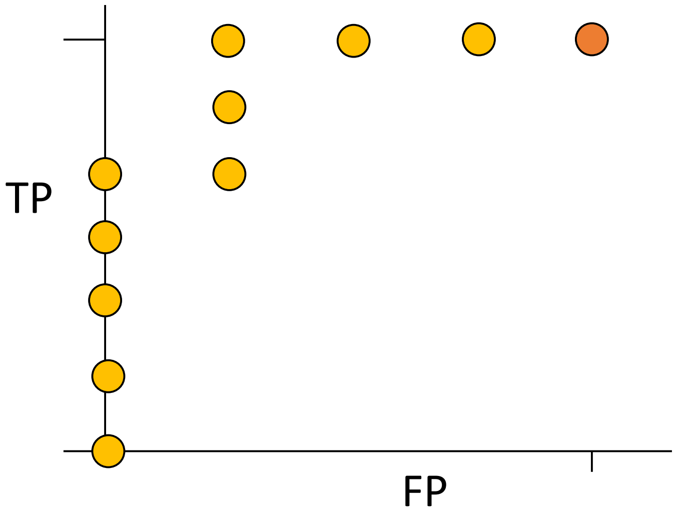
False Positive =0/4

True Positive =0/6



Receiver Operating Characteristics

Actual class	Predicted probability
A	0.95
A	0.88
A	0.85
A	0.74
B	0.52
A	0.49
A	0.45
B	0.23
B	0.19
B	0.11



False Discovery Rate

	H_0	H_1	Total
Reject H_0	V	S	R
Accept H_0	U	T	$m - R$
Total	m_0	$m - m_0$	m

$$FDR = \frac{V}{V + S} = \frac{false\ discovery}{total\ discovery}$$

Controlling FDR – FWER (Familywise Error Rate)

The ***Bonferroni procedure*** controls the FDR at level α

1. For *p*-values, p_k $(k = 1, \ldots, m)$, and given α
2. Reject the null hypotheses for all H_j if $p_k \leq \boxed{\frac{\alpha}{m}}$

Declare discovery

New significance level

- Bonferroni correction decreases the probability of producing false positive
- increases the probability of producing false negative

Controlling FDR – FWER (Familywise Error Rate)

The ***Bonferroni procedure*** controls the FDR at level α

The probability of finding a significant result = $1 - (1 - \alpha)^{\,m}$

For example, there are hypotheses with 5 comparisons, and $\alpha = 0.05$,

the probability of finding a significant result = $1 - (1 - 0.05)^5 = 0.226$.

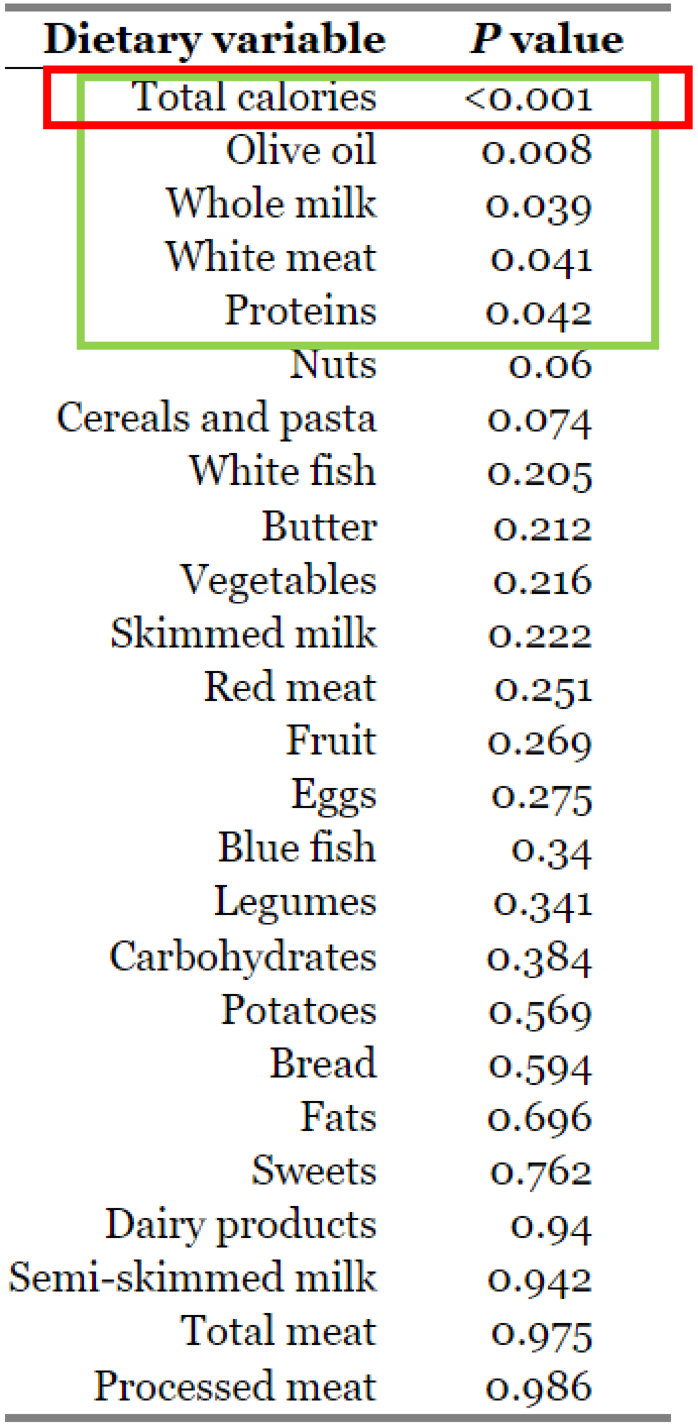
But with ***Bonferroni correction,*** it becomes $1 - (1 - 0.05/5)^5 = 0.049$.

The ***Bonferroni procedure*** controls the FDR at level α

Example

García-Arenzana et al. (2014) tested associations of 25 dietary variables with mammographic density, an important risk factor for breast cancer, in Spanish women.

After ***Bonferroni correction,*** a test would have to have p < 0.002 (0.05/25) to be significant



Dietary variable	***P* value**
Total calories	<0.001
Olive oil	0.008
Whole milk	0.039
White meat	0.041
Proteins	0.042
Nuts	0.06
Cereals and pasta	0.074
White fish	0.205
Butter	0.212
Vegetables	0.216
Skimmed milk	0.222
Red meat	0.251
Fruit	0.269
Eggs	0.275
Blue fish	0.34
Legumes	0.341
Carbohydrates	0.384
Potatoes	0.569
Bread	0.594
Fats	0.696
Sweets	0.762
Dairy products	0.94
Semi-skimmed milk	0.942
Total meat	0.975
Processed meat	0.986

http://www.biostathandbook.com/multiplecomparisons.html

García-Arenzana, N., E.M. Navarrete-Muñoz, V. Lope, P. Moreo, S. Laso-Pablos, N. Ascunce, F. Casanova-Gómez, C. Sánchez-Contador, C. Santamariña, N. Aragonés, B.P. Gómez, J. Vioque, and M. Pollán. 2014. Calorie intake, olive oil consumption and mammographic density among Spanish women. International journal of cancer 134: 1916-1925.

Controlling FDR – FWER (Familywise Error Rate)

The ***Benjamini–Hochberg*** *procedure*

The settings for many procedures

- the m *p*-values in ascending order ($p_1, p_2, p_3, \ldots p_m$) and
- their corresponding null hypotheses tested ($H_1, H_2, H_3, \ldots H_m$)

The ***Benjamini–Hochberg*** *procedure* controls the FDR at level α

1. For a given α, find the largest *k* such that $p_k \le \frac{k}{m}\alpha$
2. Reject the null hypotheses for all H_j for i=1, 2, …, k

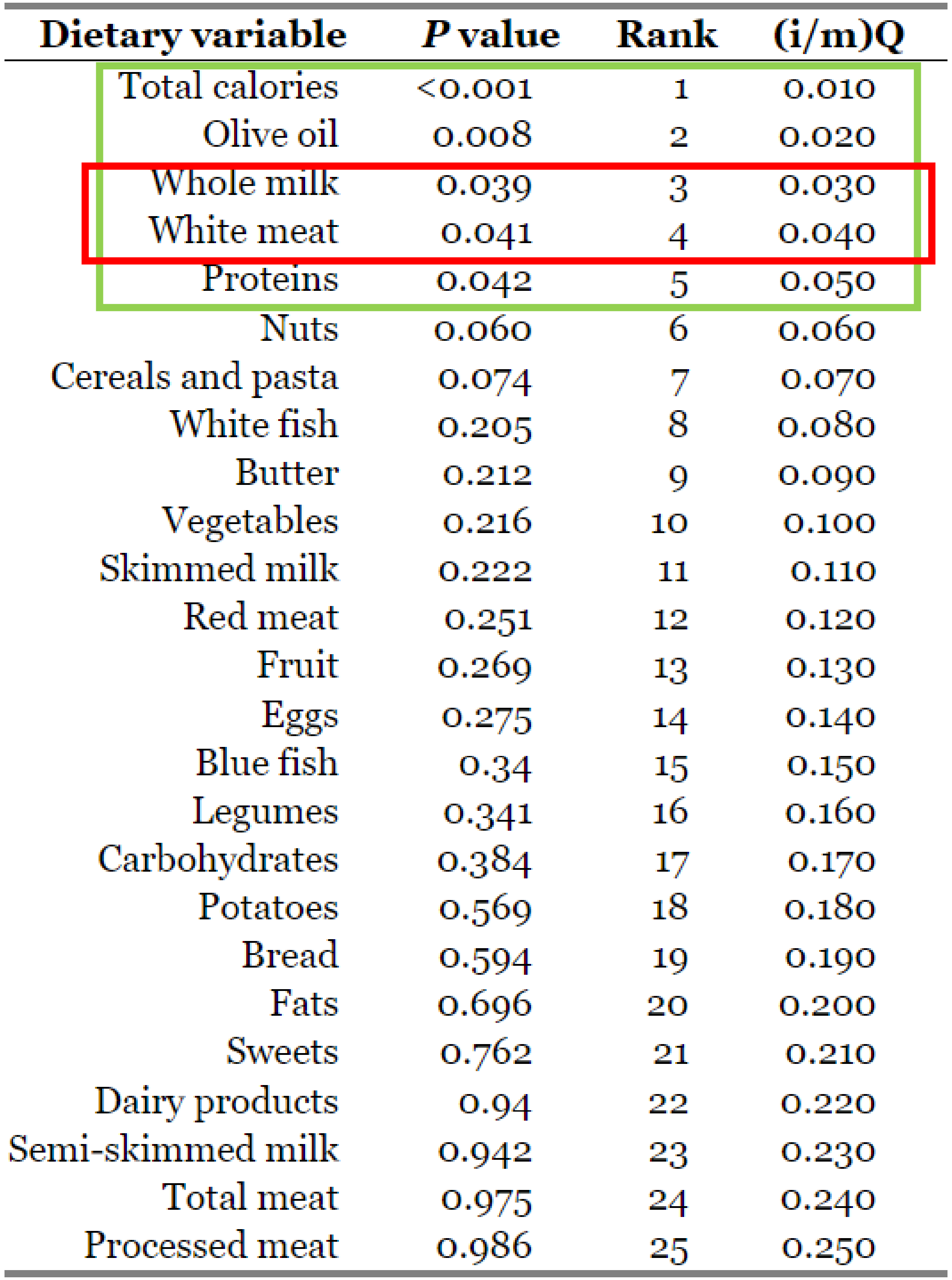
Declare discovery

The BH procedure is valid when the *m* tests are **independent**

Example

The ***Benjamini–Hochberg*** *procedure*

FDR at level α=0.25



Dietary variable	*P* value	Rank	(i/m)Q
Total calories	<0.001	1	0.010
Olive oil	0.008	2	0.020
Whole milk	0.039	3	0.030
White meat	0.041	4	0.040
Proteins	0.042	5	0.050
Nuts	0.060	6	0.060
Cereals and pasta	0.074	7	0.070
White fish	0.205	8	0.080
Butter	0.212	9	0.090
Vegetables	0.216	10	0.100
Skimmed milk	0.222	11	0.110
Red meat	0.251	12	0.120
Fruit	0.269	13	0.130
Eggs	0.275	14	0.140
Blue fish	0.34	15	0.150
Legumes	0.341	16	0.160
Carbohydrates	0.384	17	0.170
Potatoes	0.569	18	0.180
Bread	0.594	19	0.190
Fats	0.696	20	0.200
Sweets	0.762	21	0.210
Dairy products	0.94	22	0.220
Semi-skimmed milk	0.942	23	0.230
Total meat	0.975	24	0.240
Processed meat	0.986	25	0.250

http://www.biostathandbook.com/multiplecomparisons.html

García-Arenzana, N., E.M. Navarrete-Muñoz, V. Lope, P. Moreo, S. Laso-Pablos, N. Ascunce, F. Casanova-Gómez, C. Sánchez-Contador, C. Santamariña, N. Aragonés, B.P. Gómez, J. Vioque, and M. Pollán. 2014. Calorie intake, olive oil consumption and mammographic density among Spanish women. International journal of cancer 134: 1916-1925.