Errors, Chi, Power

Week 03 - Day 02

A/B testing

Use blue links

vs.

Use green links

P-value (5%) = P of accepting a false result as true

Why don’t we use 1% or 0.0001%?

“We’re 95% sure the drug works”

“We’re 99.9999% sure the drug works”

With 0.00001%:

- Less fake drugs (good)
- Less discoveries (bad)

TN,TP,FN,FP


OH, SHIT.
SHOWTIME

Test result

Reality	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Note:

a lot of variations of the previous table

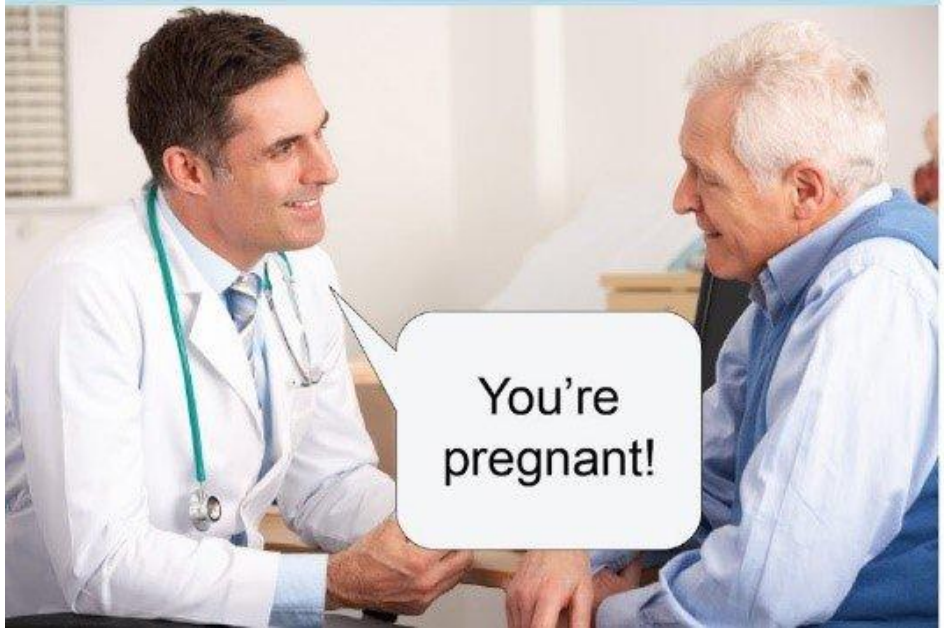
Type I errors
Type II errors

Test result

Reality

	Positive	**Negative**
Positive	True Positive	False Negative (type II)
Negative	False Positive (type I)	True Negative

Type I Error



Type II Error

Alpha
Beta
Power

Test result

Reality

	Positive	**Negative**
Positive	True Positive	False Negative (type II) **(beta)**
Negative	False Positive (type I) **(alpha)**	True Negative

Test result

Reality

	Positive	Negative
Positive	True Positive **(1-beta) (power)**	False Negative (type II) **(beta)**
Negative	False Positive (type I) **(alpha)**	True Negative

Alpha = FP / (FP + TN)

Test result

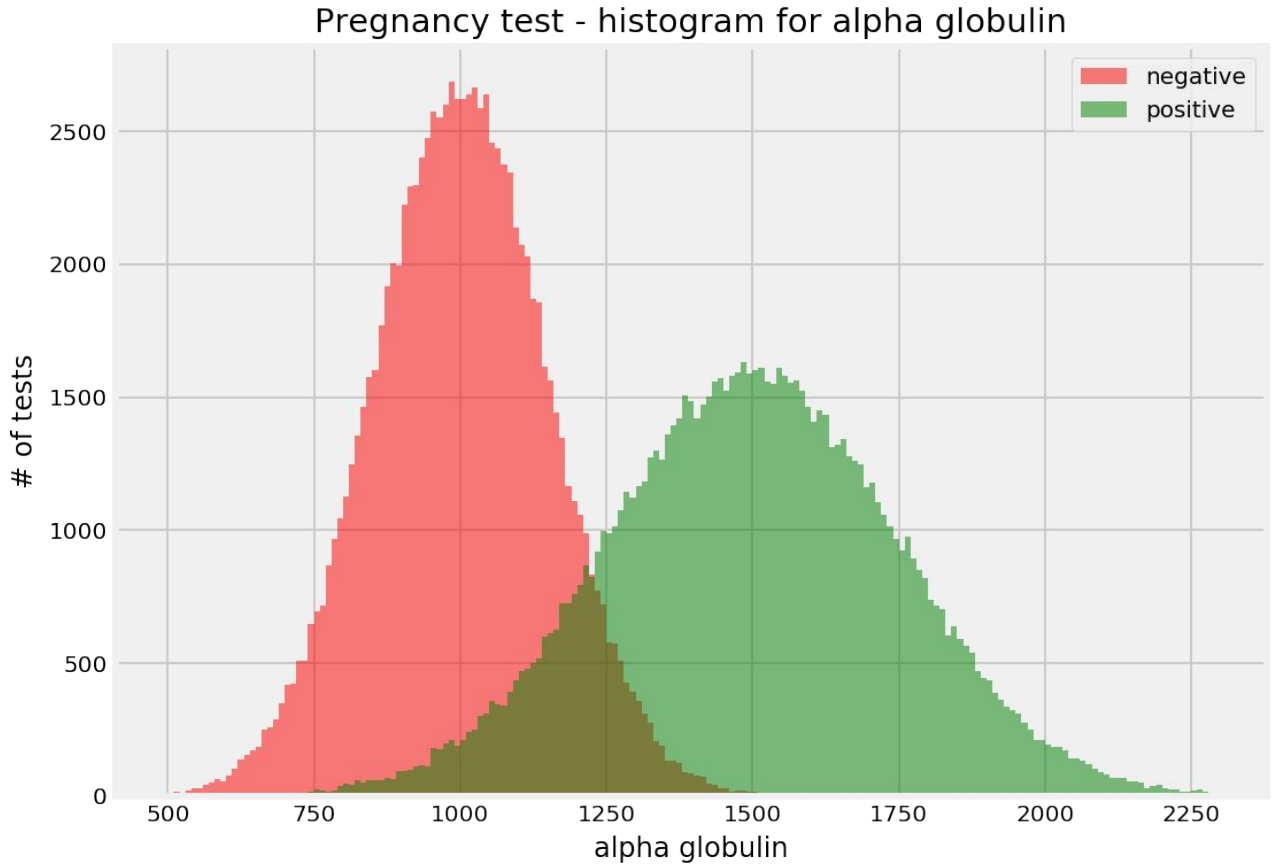
Reality

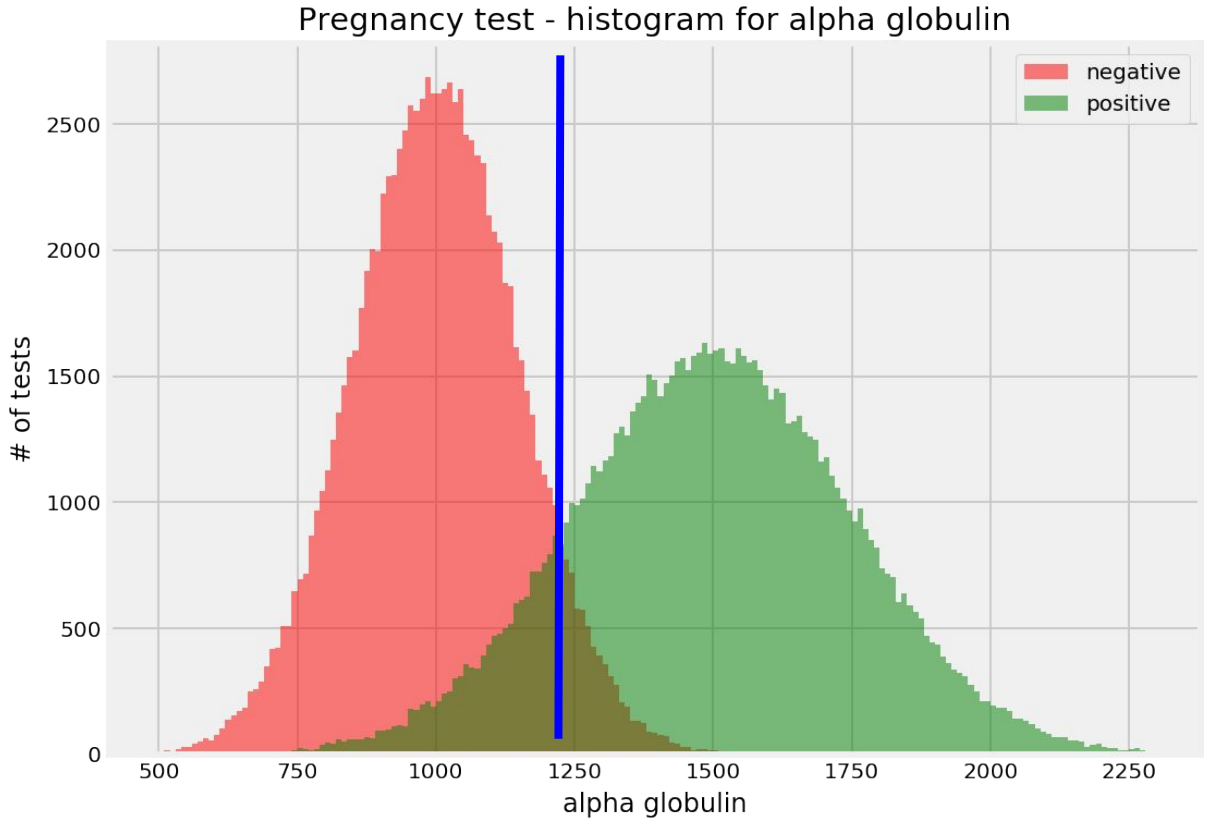
	Positive	**Negative**
Positive	True Positive	False Negative (type II) **(beta)**
Negative	False Positive (type I) **(alpha)**	True Negative

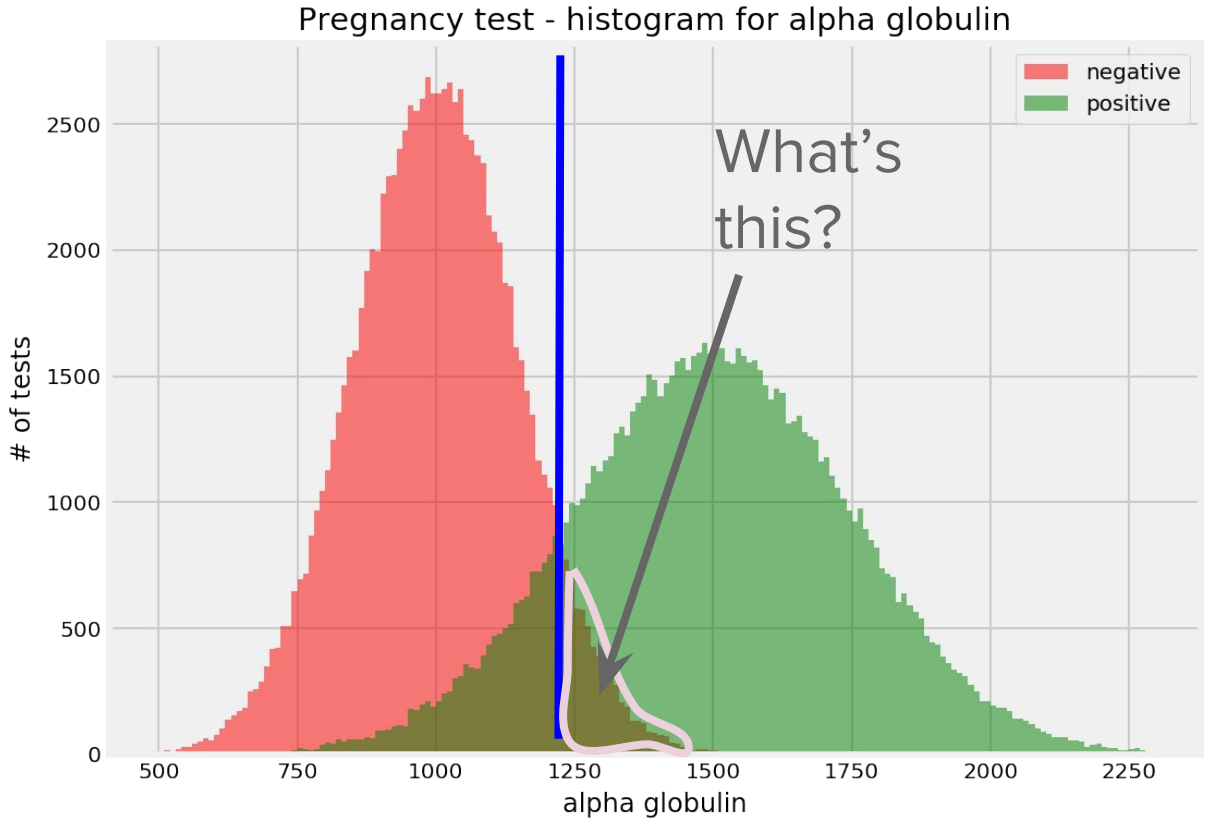
Beta = FN / (FN + TP)

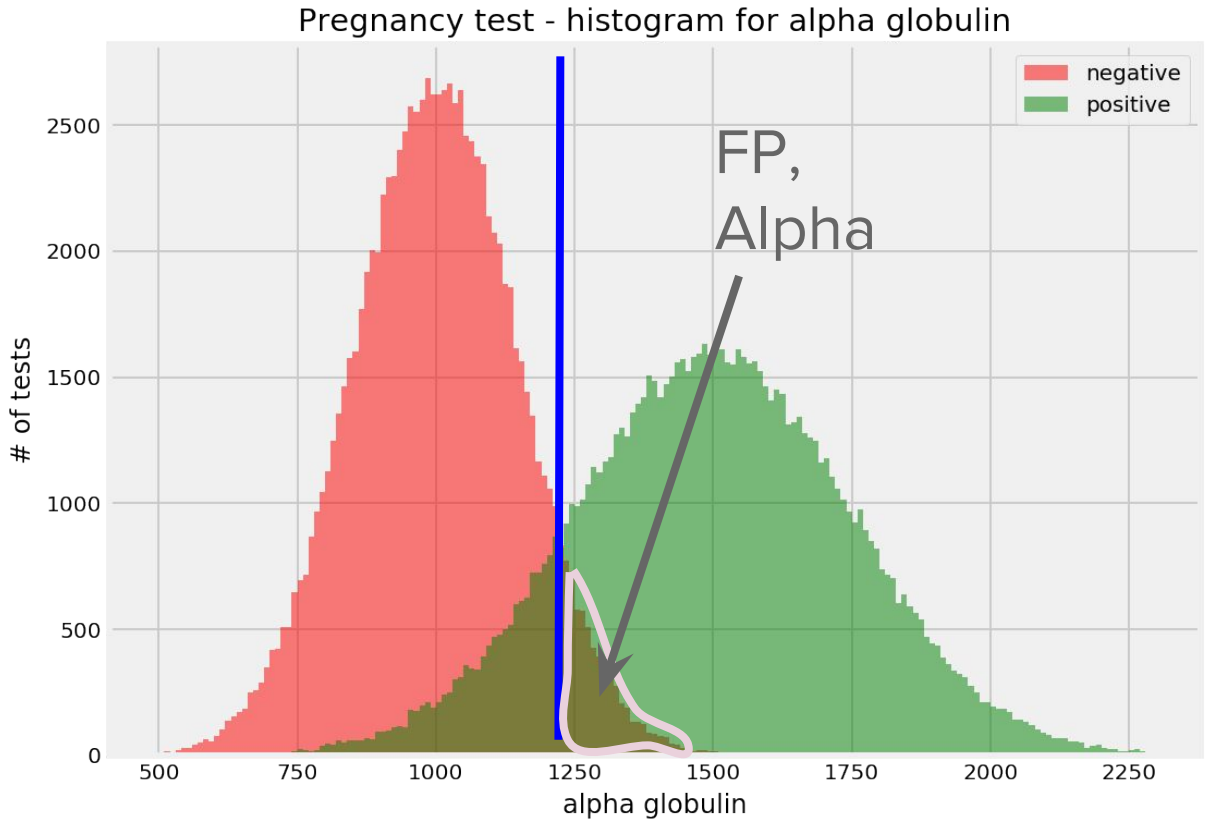
Test result

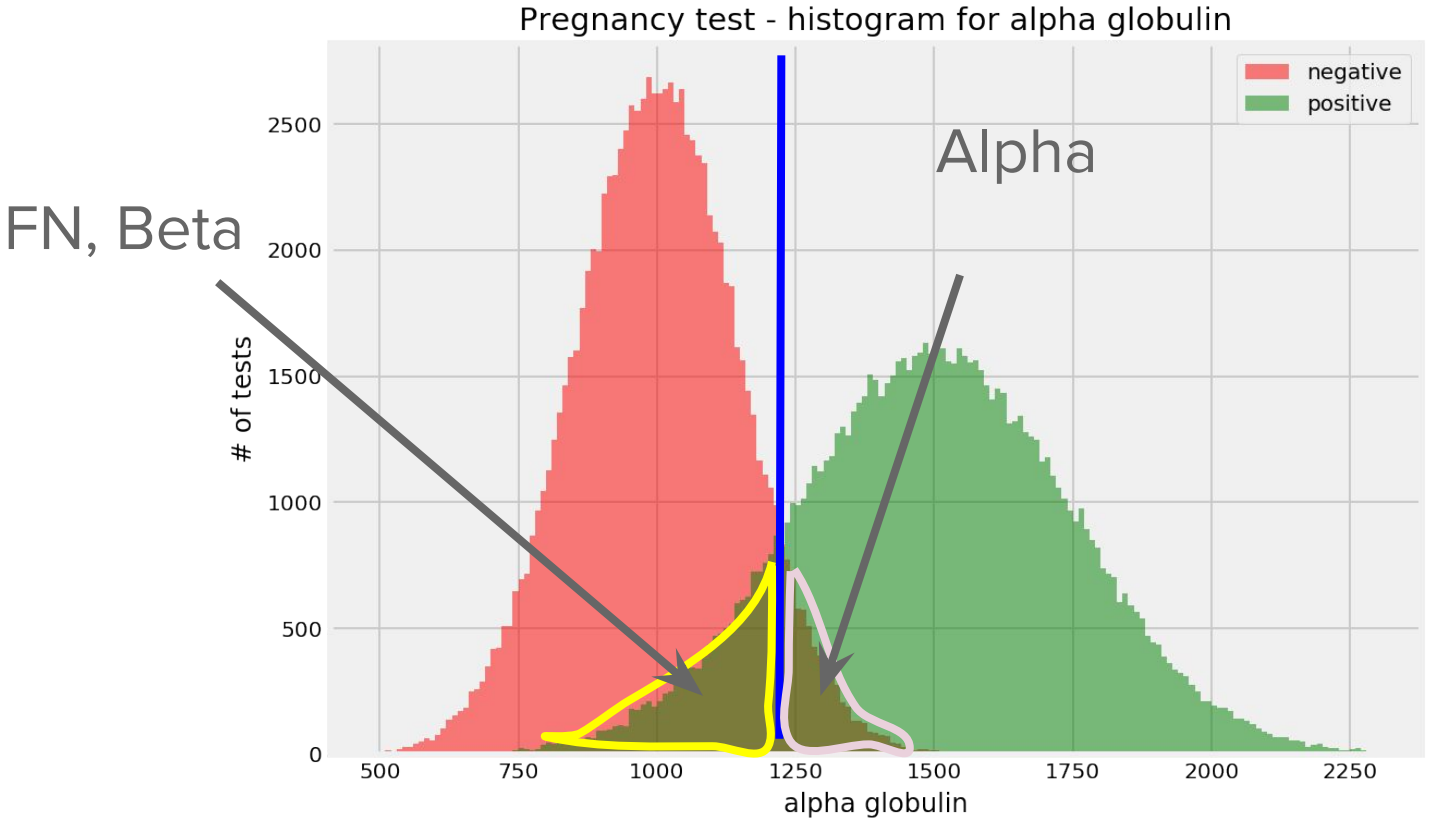
Reality	Positive	Negative
Positive	True Positive **(1-beta) (power)**	False Negative (type II) **(beta)**
Negative	False Positive (type I) **(alpha)**	True Negative

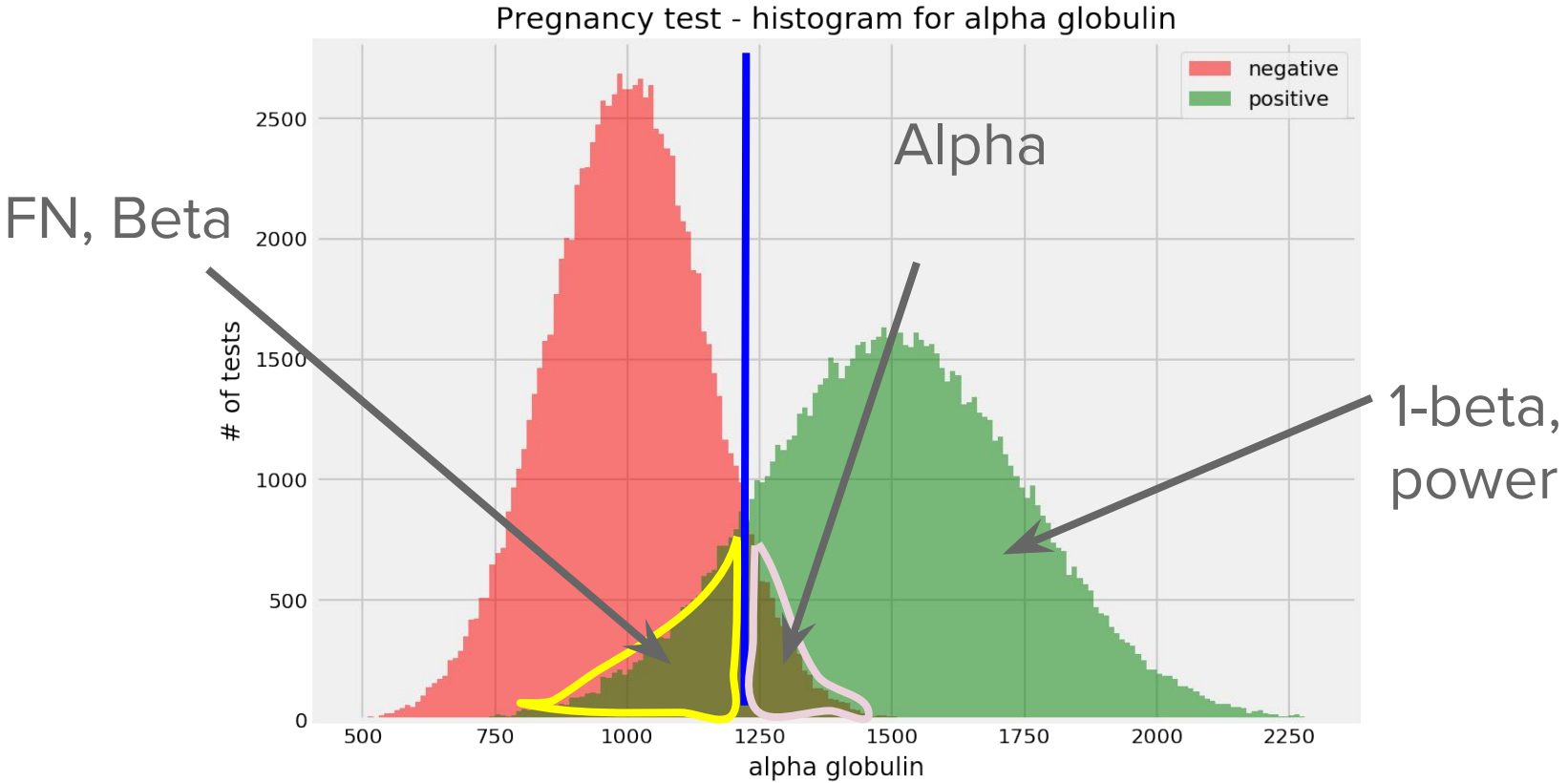

Pregnancy test - histogram for alpha globulin
negative
positive
# of tests
0
500
1000
1500
2000
2500
500
750
1000
1250
1500
1750
2000
2250
alpha globulin


Pregnancy test - histogram for alpha globulin
negative
positive
# of tests
0
500
1000
1500
2000
2500
500
750
1000
1250
1500
1750
2000
2250
alpha globulin


Pregnancy test - histogram for alpha globulin
negative
positive
What's this?
# of tests
0
500
1000
1500
2000
2500
500
750
1000
1250
1500
1750
2000
2250
alpha globulin


Pregnancy test - histogram for alpha globulin
negative
positive
FP,
Alpha
# of tests
alpha globulin
0
500
1000
1500
2000
2500
500
750
1000
1250
1500
1750
2000
2250


Pregnancy test - histogram for alpha globulin
negative
positive
Alpha
FN, Beta
# of tests
0
500
1000
1500
2000
2500
500
750
1000
1250
1500
1750
2000
2250
alpha globulin


Pregnancy test - histogram for alpha globulin
negative
positive
Alpha
FN, Beta
1-beta,
power
# of tests
alpha globulin
0
500
1000
1500
2000
2500
500
750
1000
1250
1500
1750
2000
2250

Pregnancy test vs. T-test

Pregnancy test

T-test

Hypothesis testing

Apples:

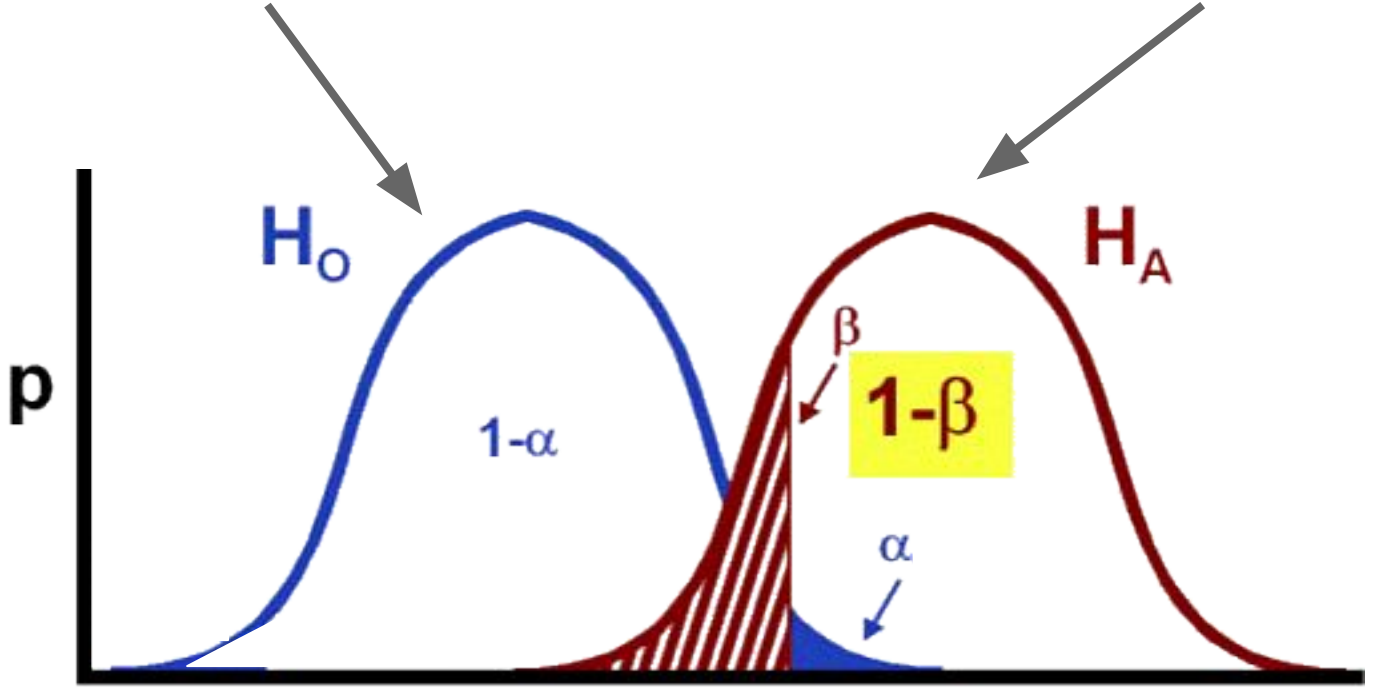
mean(sample1) > mean(sample2)

mean(s1) - mean(s2)

From same population

mean(s1) - mean(s2)

From different populations



Chi-Squared
(test of independence)

("goodness of fit" is another test!)

Let’s talk about apples (again 😱)

Two samples:

I don’t know the mean!

..but I know the categories.

	Red	Green	Yellow
Sample 1	10	20	10
Sample 2	5	25	10

Do they come from

the same population?

What about now?

	Red	Green	Yellow
Sample 1	1000	2000	1000
Sample 2	500	2500	1000

H0 = null hypothesis = no difference

H1 = alt. hypothesis = they’re different

```
data = [[1000,2000,1000],
        [500,2500,1000]]
scipy.stats.chi2_contingency(data)
```

Same approach as t-test!

1) Calculate a statistic
2) Plot your statistics against a distribution
3) Calculate the p-value (using critical value)

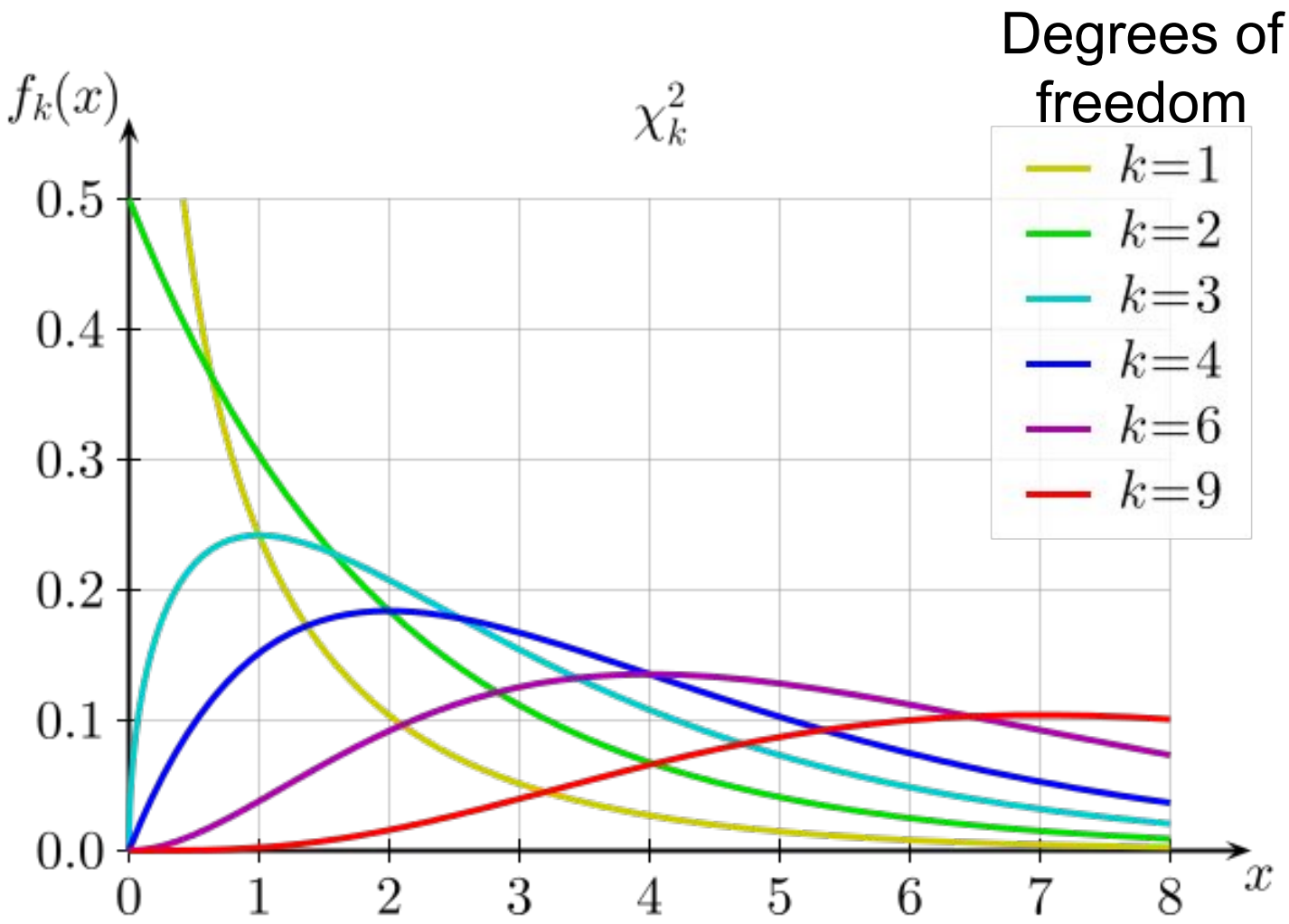
Intuition

1) Calculate an imaginary (i.e. expected) equal distribution for g1 and g2
2) Check how much the real values are different from the imaginary ones

Observed

Expected
(in case of same population)

$$\chi^2 = \sum_{i=1}^{cells} \frac{(O_i - E_i)^2}{E_i}$$


Degrees of freedom
$f_k(x)$
χ^2_k
$k=1$
$k=2$
$k=3$
$k=4$
$k=6$
$k=9$
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7
8
x

How to calculate the expected values?

(do it once then forget about it)

http://psc.dss.ucdavis.edu/sommerb/sommerdemo/stat_inf/tutorials/chisqhand.htm

Contingency table and Chi-squared test

(17 min video)

https://www.youtube.com/watch?v=hpWdDmgsIRE

Power analysis & Sample size calculation

Blue links vs. Green links

How many users should I test?

10 -> not enough

100,000 -> ok but the test will take forever

5 things to consider

1. Our desired type I error rate.
2. Our desired type II error rate (or, more commonly, power).
3. The expected size of the effect, or the mean difference between groups.
4. The expected standard deviation of measurement.
5. The sample size.

Solutions:

Heuristic

Calculator

Summary

1. a/b testing
2. TN, TP, FN, FP
3. Type I, Type II (interview question)
4. Alpha, power
5. Chi-squared
6. Sample size calculation