The Null hypothesis

Binomial statistics can be used to find out if an effect is statistically significant.

Example: Does a certain new ski wax help to win races?

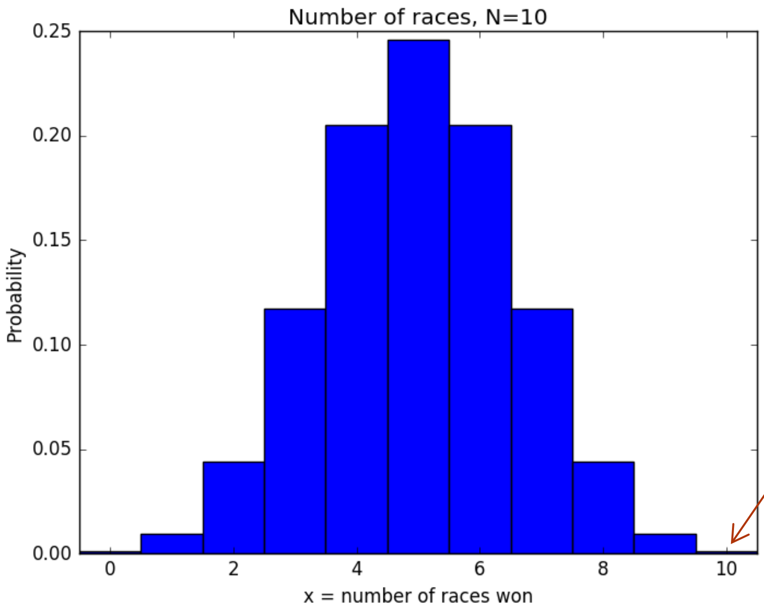
There are 10 races with the new wax and 8 are won. Is this statistically significant to conclude that the wax helps?

(Analogue: Get 8 tails after 10 coin flips. Is the coin unfair?)

To address this question, need to formulate a statistical hypothesis, the so called *null hypothesis*.

Null hypothesis example

- Null hypothesis: New wax has no effect, chances of winning are the same with or without the new wax (with everything else being equal) .
- The probability of winning x out N races with success probability $p = 0.5$ is just $B_{N,0.5}(x)$.



$P(10) = 0.5^{10} = 0.1\%$

Unlikely to win all races

What if 8 races are won?

Null hypothesis example

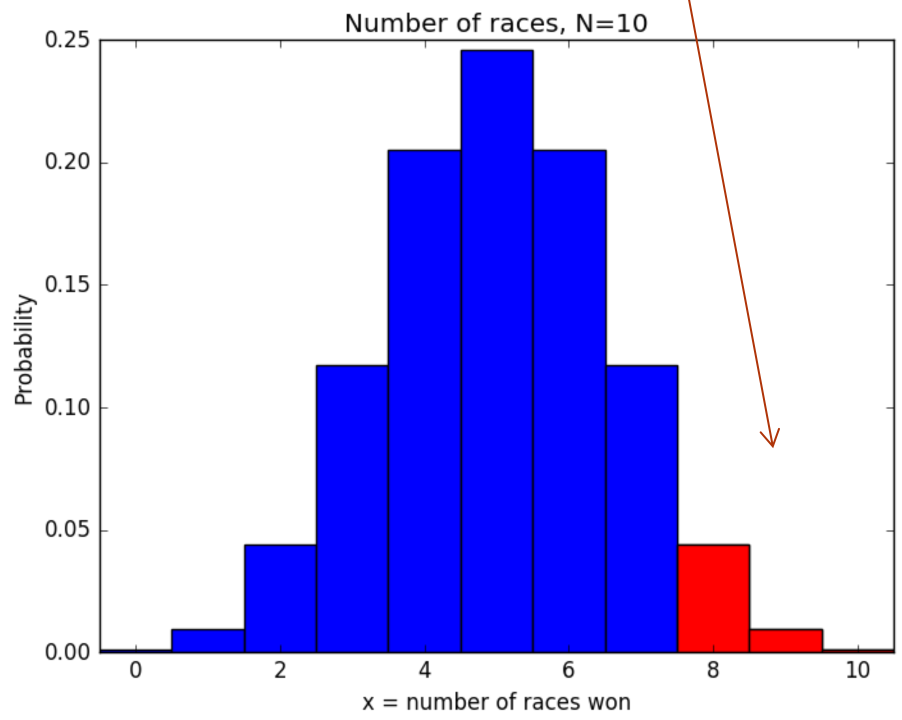
To judge the statistical significance of this outcome want to know the probability of **winning 8 or more races**:

$$B_{N,0.5}(8) + B_{N,0.5}(9) + B_{N,0.5}(10) \approx 5.5\%$$

There are two possible conclusions from 8 out of 10 wins :

1. Null hypothesis is correct. There is no effect and by chance an unlikely event has occurred.

2. Null hypothesis is false and wax does help



Null hypothesis example

- To make a decision we define a threshold probability – the so called *significance level*.
- It is common practice to choose it to be 5% or 1%. Note that this is convention, nothing special about these numbers!
- If the probability of the outcome (here, 8 **or more wins**) is
- less than 5% (according to null hypothesis), we call it significant.
- less than 1%, we call it highly significant.

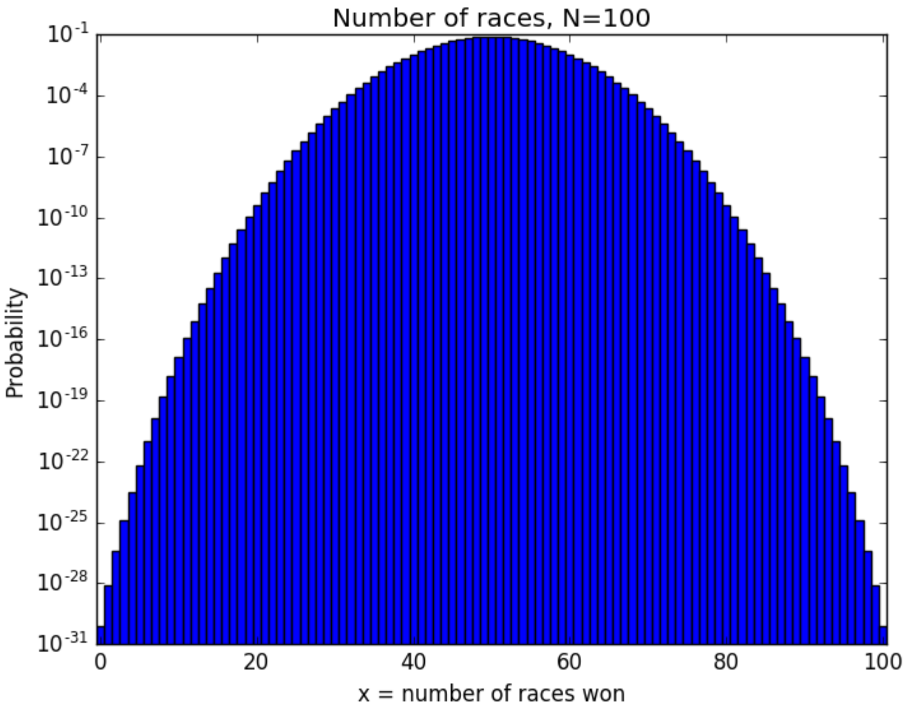
Null hypothesis example

- Probability of 8 or more wins is 5.5%. According to the rules, it just above the significance level and we conclude that the wax does not help (null hypothesis true).

- If 10 races are won, the probability is 0.1% according to the null hypothesis. This is well below 1% and therefore highly significant. In this case, we reject the null hypothesis and conclude that wax works.

Sample size matters

- What if 100 races are conducted, and 80 are won? Percentage of won races remained the same.
- What about statistical significance?



Probabilities beyond x=80 are tiny !

Use semilog plot.

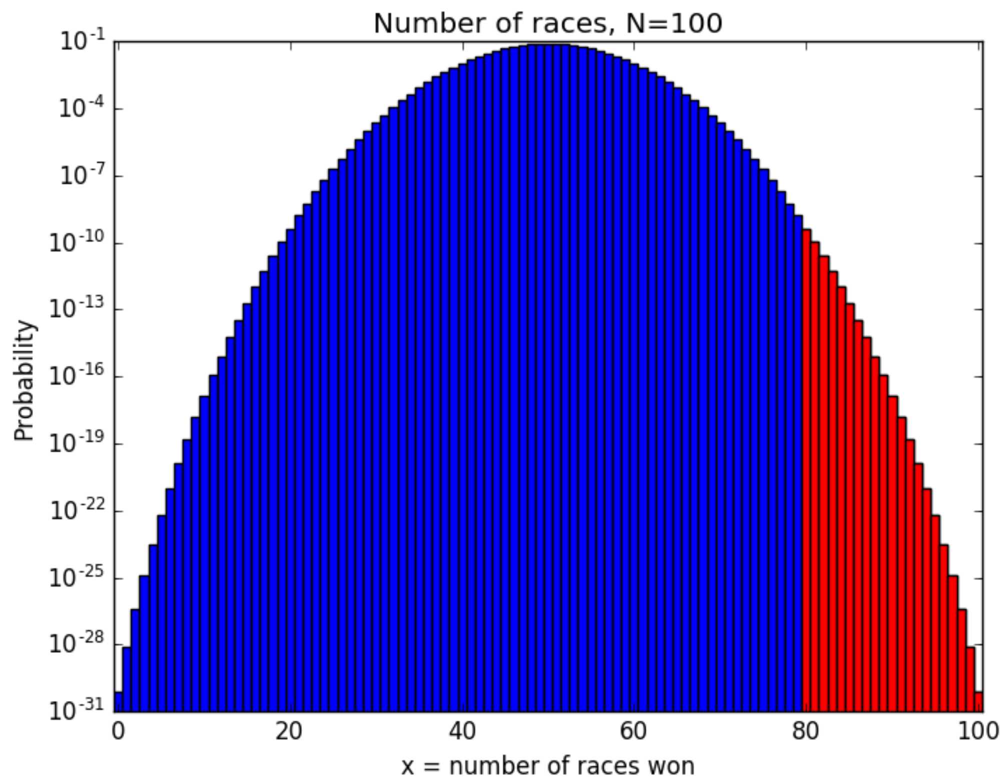
Sample size matters

Probability of winning 80 or more races out of 100, with $p = 0.5$ is very, very small:

$$P(x \geq 80) = \sum_{x=80}^{100} B_{100,0.5}(x) = 5.6 \cdot 10^{-8}\ \%$$

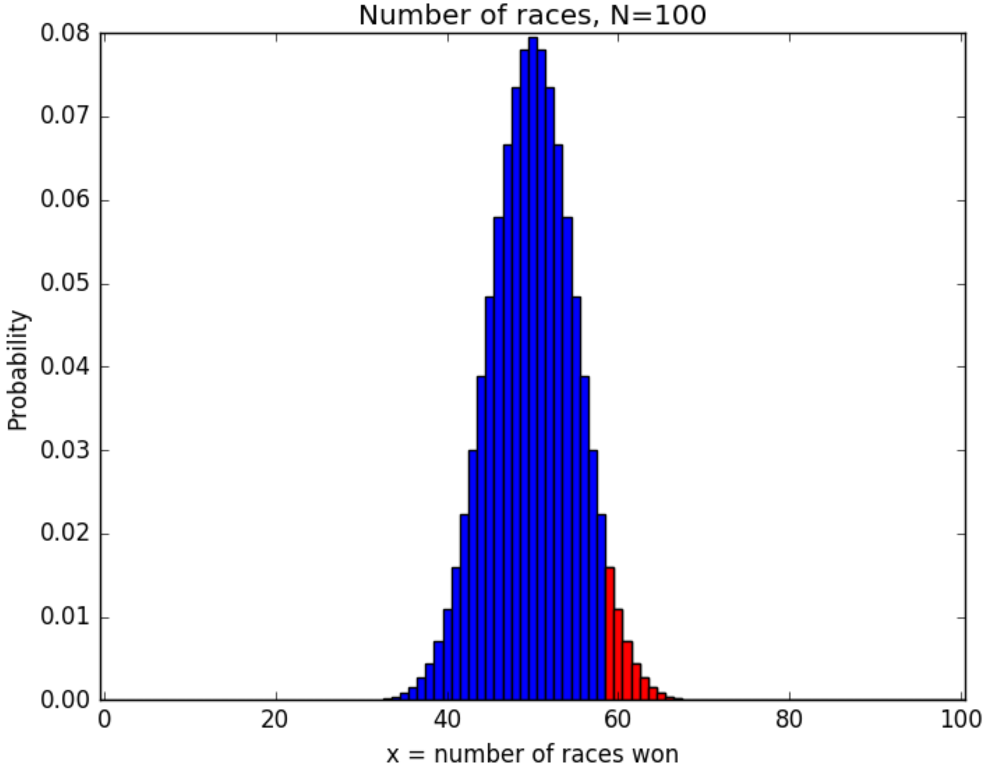
Result highly significant and null hypothesis rejected. Wax helps.

Why the difference? Percentage of races won scales with N, but the width of the distribution ($\sim\sigma$) scales with $\sqrt{N}$



Sample size matters

5% significance level for 100 races corresponds to winning 59 or more races: $\sum_{x=59}^{100} B_{100,0.5}(x) \approx 0.05$



In physics, the significance level is often expressed in terms of multiples of σ, the standard deviation.

Here,

$\sigma = \sqrt{100 \cdot 0.5 \cdot 0.5} \approx 5$ and

$\mu = 100 \cdot 0.5 = 50$

Threshold of 59 races is 9 wins above the average.

$\frac{9}{\sigma} = 1.8$,

Significance level is 1.8σ

Significance level and sigma

- In particle physics, want a 5σ significance level for claiming a discovery. The corresponding probability for a success rate that is larger than 5σ away from the mean is $2.9 \cdot\ 10^{-7}$ $\approx$ 1 in 3.5 million.
- In the ski wax example with 10 races, it is impossible to reach that significance level due to limited number of trials. Sample size matters!
- For N=10 races, minimum significance level is 0.01% (10 won races), which corresponds to 3.1σ.

Higgs boson

Particles created in collision decay through 2 back-to-back photons. Detector records number of photons created and their energy. Presence of Higgs boson led to **excess events** at around 125GeV. Red dotted line is null hypothesis.

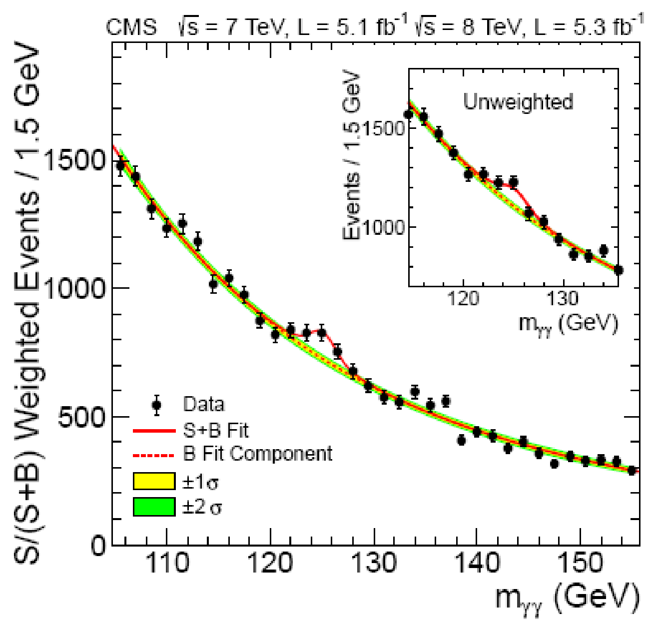


Figure 13: The diphoton invariant-mass distribution for the 7 and 8 TeV data sets (points), with each event weighted by the predicted $S/(S+B)$ ratio of its event class. The solid and dotted lines give the results of the signal-plus-background and background-only fit, respectively. The light and dark bands represent the ±1 and ±2 standard deviation uncertainties respectively on the background estimate. The inset shows the corresponding unweighted invariant-mass distribution around $m_{\gamma\gamma}$ = 125 GeV.

Null hypothesis summary

Summary of null hypothesis

- Formulate a null hypothesis: Choice of success probability in the absence of an effect.
- Measure number of successes x_0 experimentally.
- If $P(x \geq x_0) = \sum_{x=x_0}^{N} P(x) < 0.05$ null hypothesis rejected. Effect statistically significant.

Here we assume that the effect (e.g.) increases the success rate. Can also ask the question if it has a negative effect. In this case, need to look at left tail of the distribution. If outcome is worse then expected null hypothesis

- If $P(x \leq x_0) = \sum_{x=0}^{x_0} P(x) < 0.05$ null hypothesis rejected.

Distribution for random errors

Can use binomial distribution to model random errors.

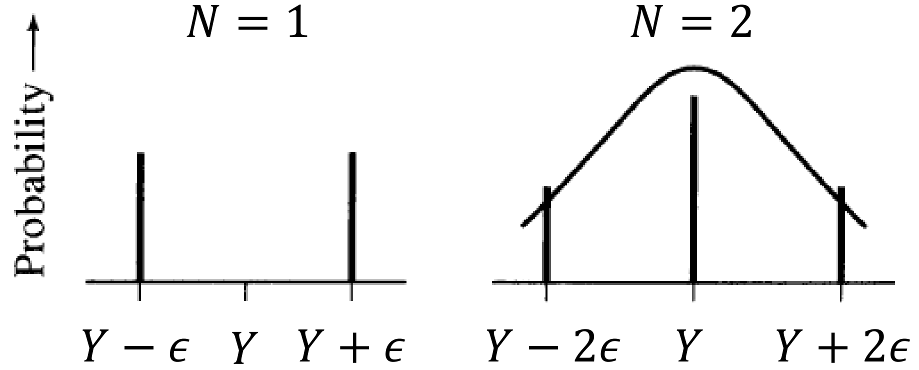
Assumptions:

- Let there be N sources of random error (e.g. reaction time, temperature fluctuations, vibrations etc.).
- Each error introduces a *fixed deviation* $\pm\epsilon$ to our measurement.
- There is a 50% chance that ϵ is positive or negative.

Distribution of random errors

- Let Y be the true mean. If there is just one source of error, the possible outcomes y will be $Y + \epsilon$ and $Y - \epsilon$, both of them equally likely.
- If there are two source of error, the there are three possible outcomes:

$$Y + \epsilon - \epsilon = Y$$
$$Y + \epsilon + \epsilon = Y + 2\epsilon$$
$$Y - \epsilon - \epsilon = Y - 2\epsilon$$



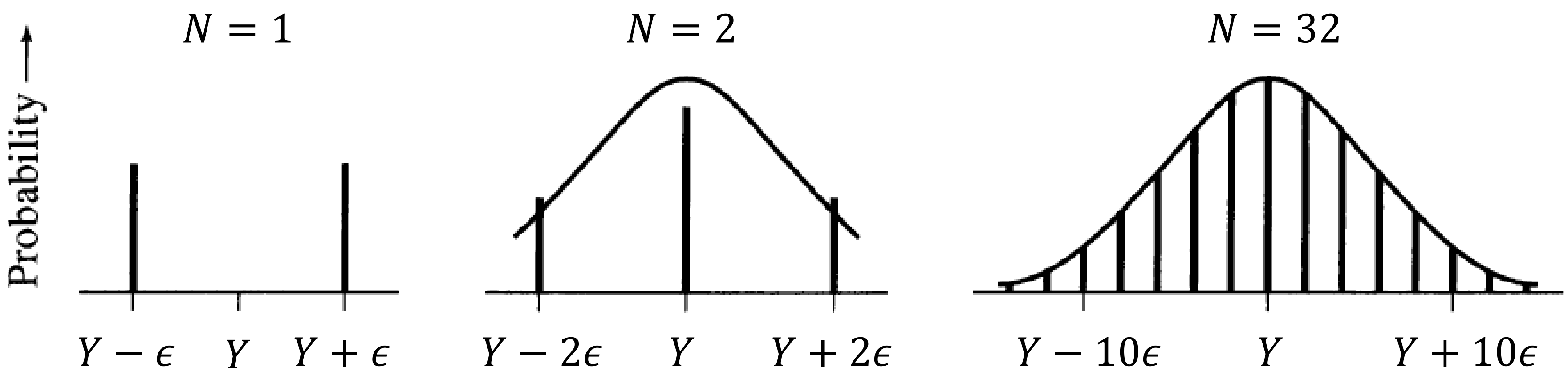
- In general, if there are N sources of error, the outcome could range between $Y + N\epsilon$ and $Y - N\epsilon$. If there are x events that give $+\epsilon$ deviations, then our answer will be

$$\begin{aligned} y &= Y + x\epsilon - (N - x)\epsilon \\ &= Y + (2x - N)\epsilon \end{aligned}$$

Distribution of random errors

The probability of x positive deviations is just given by the binomial distribution $B_{N,0.5}(x)$, therefore $y = Y + (2x - N)\epsilon$ is governed by the same distribution:

$$P(y) = B_{N,0.5}\left(\frac{y-Y}{2\epsilon} + \frac{N}{2}\right)$$



The standard deviation of $B_{N,0.5}(x)$ is just $\sigma_x = \sqrt{N \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{N/4}$.

Therefore, $\sigma_y = 2\epsilon\sigma_x = \epsilon\sqrt{N}$.

<u>As N increases, a bell shaped curve emerges.</u>

The Normal (Gaussian) Distribution

Can show that in the limit $N \to \infty$ and $\varepsilon \to 0$, the binomial distribution approaches the continuous Gaussian or Normal distribution:

$$B_{N,p}(x) = \binom{N}{x} p^x q^{N-x} \approx \frac{1}{\sqrt{2\pi Npq}} e^{-\frac{(x-Np)^2}{2Npq}}$$

<u>This limit is independent of the success probability p.</u>

In general, for a given mean μ and variance σ^2, the normal distribution is given by

$$\mathrm{P(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$