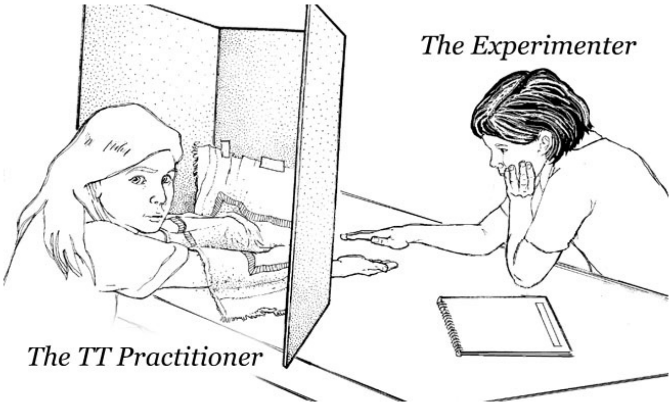
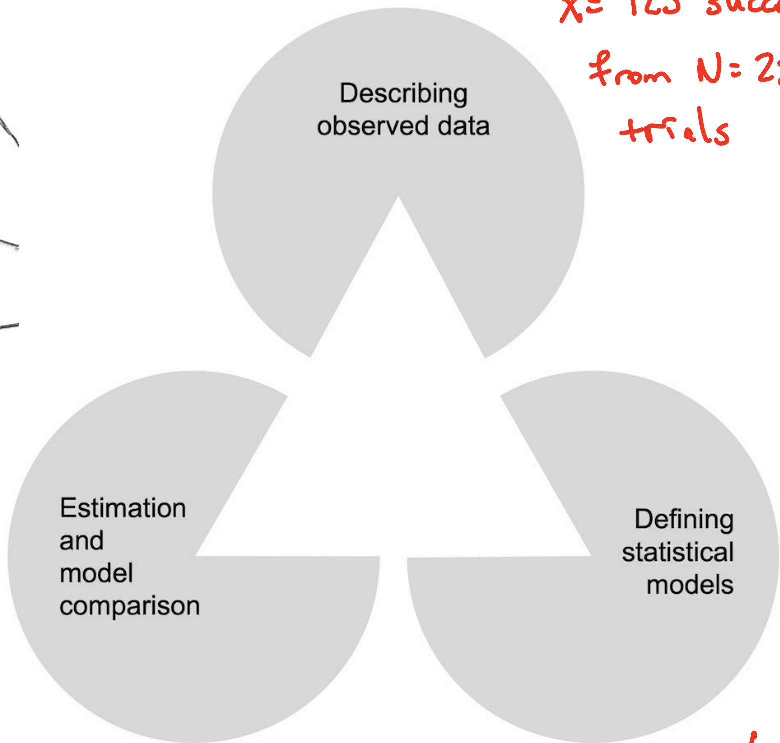
Lecture 2 - The language of Bayesian statistics

Recall: therapeutic touch experiment (Rosa et al., 1998)



$x = 123$ successes from $N = 280$ trials




```
> binom.test(x=123, n=280, p=0.5, alternative="greater")

	Exact binomial test

data:  123 and 280
number of successes = 123, number of trials = 280, p-value = 0.9819
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.3893703 1.0000000
sample estimates:
probability of success
             0.4392857
```

Binomial model (w)

$\mathcal{H}_0: w = 0.5$

$\mathcal{H}_1: w > 0.5$

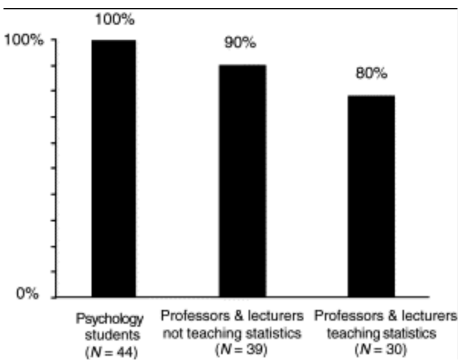
Q: what does this p-value $(p = 0.9819)$ represent?

↳ A: probably not what you think it does!

Misconceptions about p-values

* p = probability that $\mathcal{H}_0$ is true
* $1-p$ = probability that $\mathcal{H}_1$ is true
* and others (see Gigerenzer, 2004 – "Mindless statistics")

These are all incorrect, but widely believed to be true by students and professors alike!



... So what is a p-value, then?

Def: A p-value is the probability of observing data (or more extreme) if $\mathcal{H}_0$ is true.

↳ $P(\text{data} \mid \mathcal{H}_0)$

↳ as scientists, would be really good to know $P(\mathcal{H}_0 \mid \text{data})$ instead!

↳ for this, we need Bayes Theorem.

Bayes Theorem

$$\boxed{p(\mathcal{H} \mid \text{data}) = p(\mathcal{H}) \times \frac{p(\text{data} \mid \mathcal{H})}{p(\text{data})}}$$

Components:

(1) $P(\mathcal{H} \mid \text{data})$ = "posterior" → degree of belief in $\mathcal{H}$ after observing data

(2) $P(\mathcal{H})$ = "prior" → degree of belief in $\mathcal{H}$ before observing data.

(3) $P(\text{data} \mid \mathcal{H})$ = "likelihood" → degree to which observed data is likely under $\mathcal{H}$.

(4) $P(\text{data})$ = "prior predictive" → weighted average of probabilities of observing data under all models being considered.

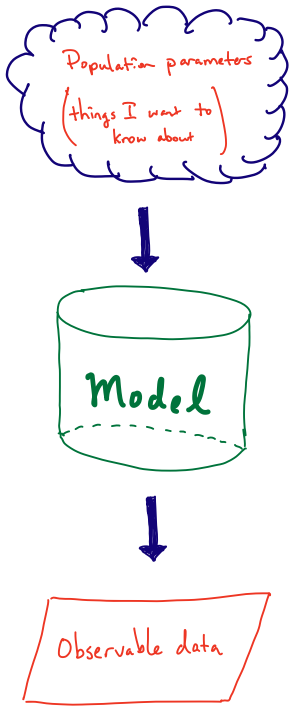
Note: Bayes theorem can be conceptualized as

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

"proportional to"

Why do we care?

Recall: inference = given data, what is parameter?



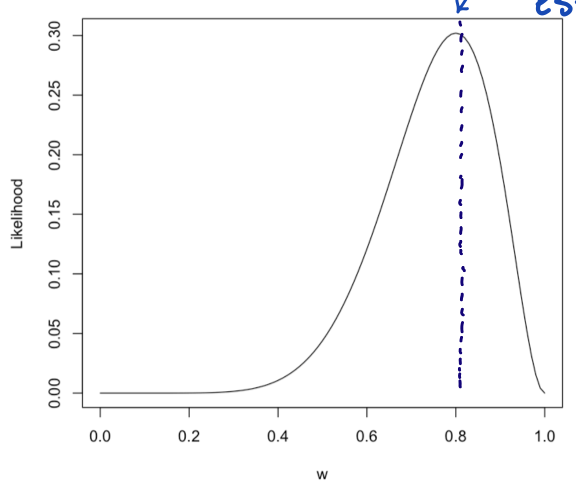
What is a reasonable range for values of the parameter w?

Suppose we observe $x=9$ successes in 20 trials.

Last time we plotted a likelihood function in R:

```r
w = seq(from=0, to=1, by=0.01)
plot(w, dbinom(x=8, size=10, prob=w),
     type="l",
     ylab="Likelihood")
```

maximum likelihood estimate



* But this is not a probability distribution.
* Hard to quantify uncertainty in w
* would like a statement like "There is 95% probability that w is between __ & __"

Need a posterior distribution for w.

↳ Bayes theorem gives us the mathematical tools to transform this likelihood function into a probability distribution

↳ remember: posterior $\propto$ prior $\times$ likelihood

↳ we already have the likelihood (binomial); now we need a prior

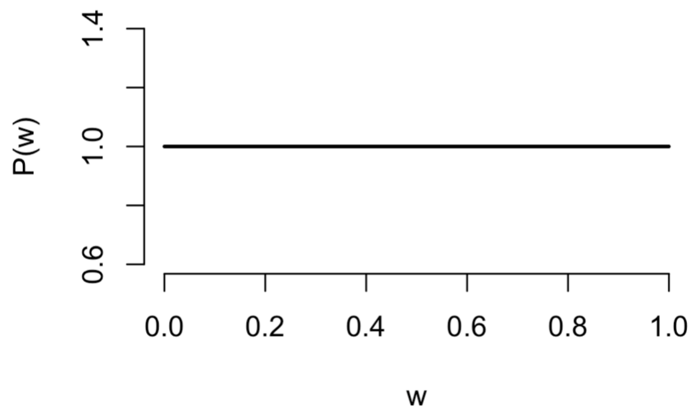
What is a prior?

In this context, a prior is a probability distribution that mathematically encodes our prior belief about the model parameter in question.

↳ these can differ between analysts.

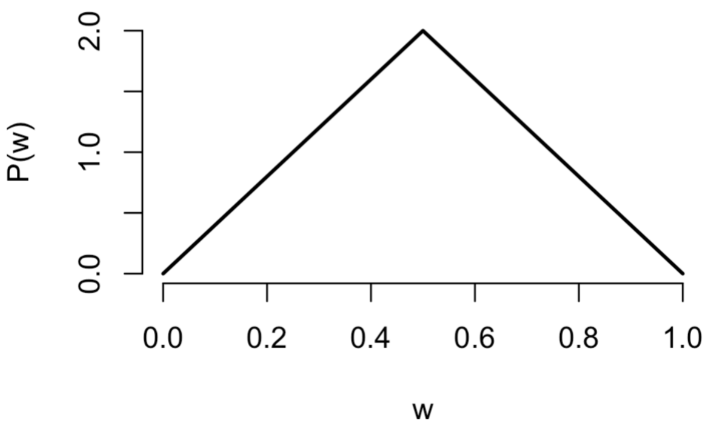
↳ lets look at 3 different priors for w

Prior 1 - uniform prior



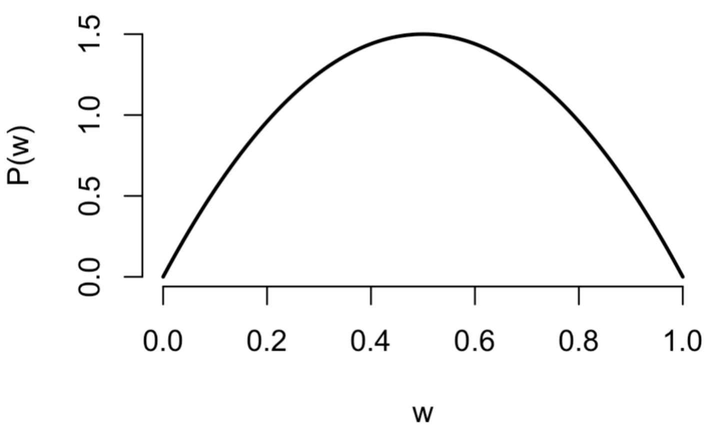
* flat prior - every value of w is equally likely
* no strong prior belief in any range of values for w.

Prior 2 - sharp peaked prior



* most likely value (a priori) is $w = 0.5$
* prior belief decreases linearly as $w \to 1$ and $w \to 0$.

Prior 3 - smooth peaked prior

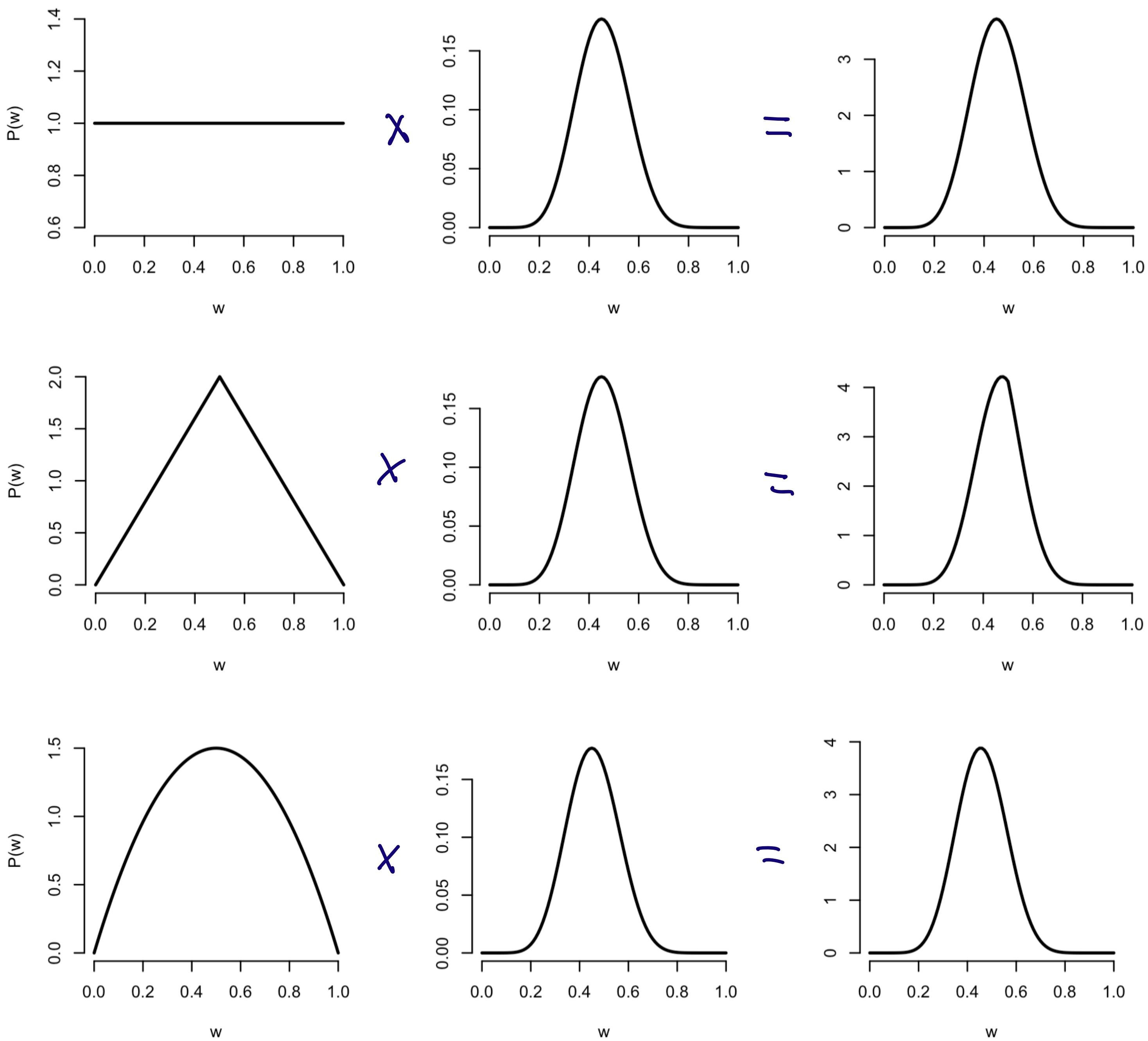


* most likely value (a priori) is $w = 0.5$
* prior belief decreases slowly as $w \to 1$ and $w \to 0$.

Bayesian updating:

To get the posterior distribution for w, we multiply the prior and the likelihood (then divide by a constant so that the area under the curve is equal to 1). Here is the result (from R) for each of these priors:

Prior	Likelihood	Posterior



You may note that the posteriors are similar.

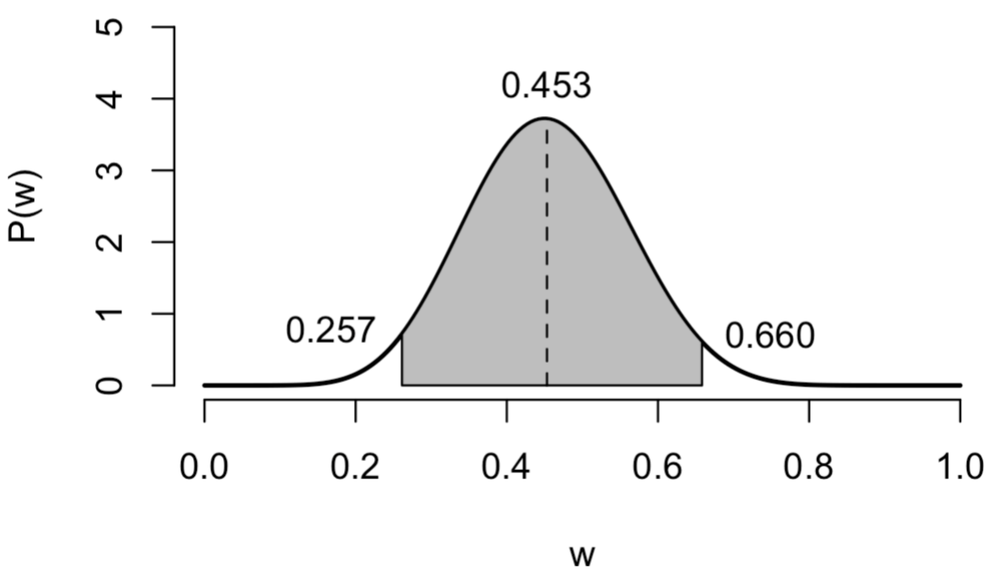
But they're not exactly the same!

For each one, lets look at the median value and the interval of values w which contains the middle 95% of the distribution

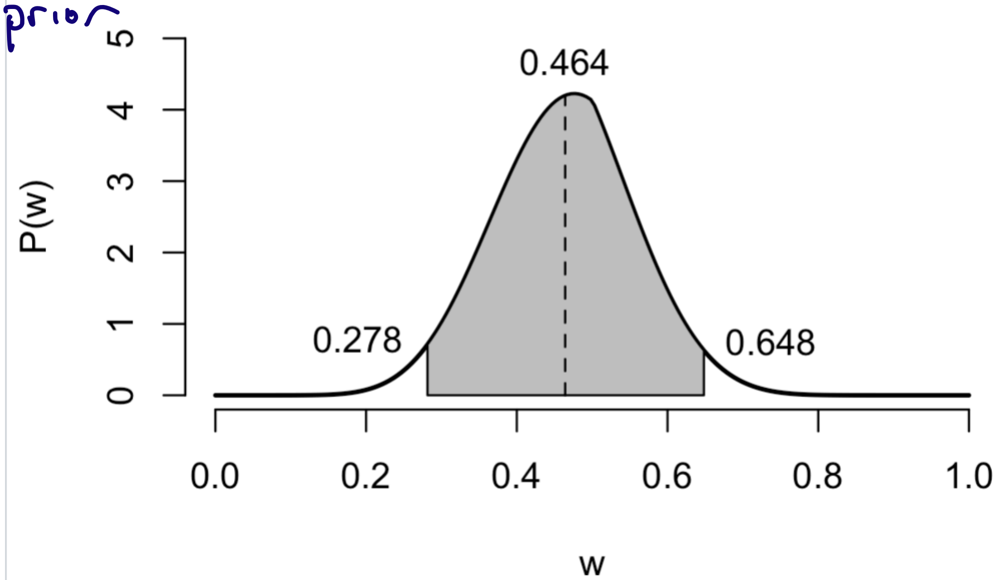
↳ also known as the 95% <u>credible interval</u>

Prior 1 - uniform prior

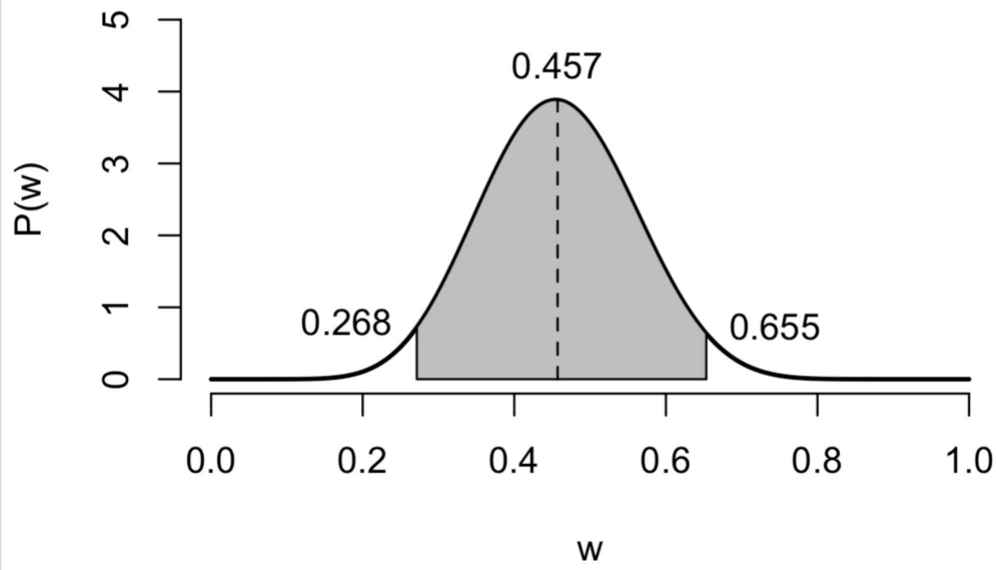
95% CrI = [0.257, 0.660]



Prior 2 - sharp peaked prior



Prior 3 – smooth peaked prior



Take home – even though the posteriors look very similar, the choice of prior has a small effect on our posterior estimates of w.

↳ choosing the best priors to use is an active area of research!

↳ when reporting a Bayesian analysis, must always report the prior used... it matters!

We talked about estimation — what about model comparison?

↳ start by writing Bayes theorem for two models $\mathcal{H}_0$ & $\mathcal{H}_1$.

$$P(\mathcal{H}_0 \mid \text{data}) = p(\mathcal{H}_0) \times \frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data})}$$

$$P(\mathcal{H}_1 \mid \text{data}) = p(\mathcal{H}_1) \times \frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data})}$$

Let's compute the posterior odds:

$$\frac{P(\mathcal{H}_0 \mid \text{data})}{P(\mathcal{H}_1 \mid \text{data})} = \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)} \times \frac{\frac{p(\text{data} \mid \mathcal{H}_0)}{\cancel{p(\text{data})}}}{\frac{p(\text{data} \mid \mathcal{H}_1)}{\cancel{p(\text{data})}}}$$

$$\frac{P(\mathcal{H}_0 \mid \text{data})}{P(\mathcal{H}_1 \mid \text{data})} = \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)} \times \frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)}$$

or

$$\boxed{\frac{P(\mathcal{H}_0 \mid \text{data})}{P(\mathcal{H}_1 \mid \text{data})} = \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)} \times BF_{01}}$$

So we have:

Posterior odds = Prior odds × updating factor

This updating factor, $\dfrac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)}$, is called the Bayes factor

Much of this course will be spent learning how to compute Bayes factors and interpret them. But it is reasonably easy to "see" the Bayes factor for our example today:

Observed data: $X = 9$ successes in $N = 20$ trials

Models: Binomial likelihood

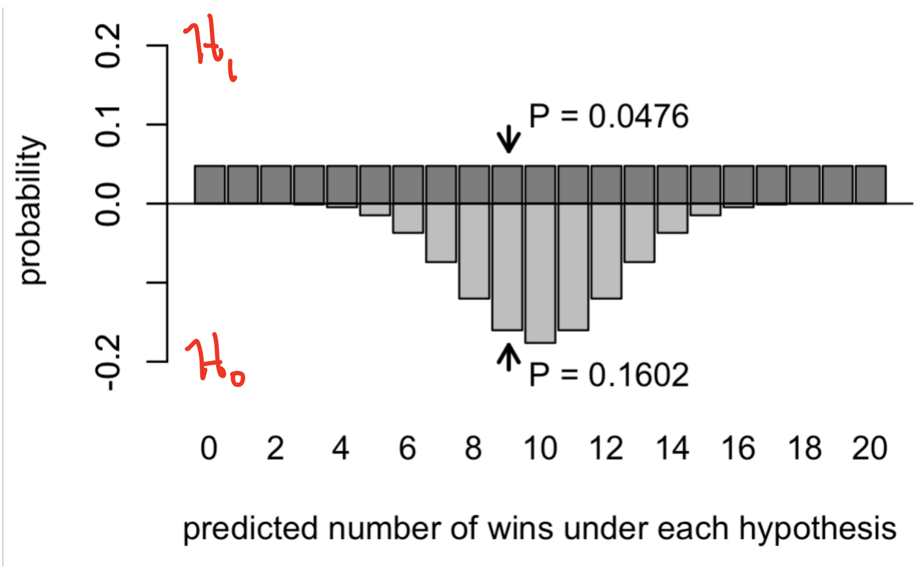
$\mathcal{H}_0 : w = 0.5$

$\mathcal{H}_1 : w \neq 0.5$ (must specify how w is distributed a priori)

Below, I've plotted the "predictive distribution" for each model ($\mathcal{H}_0, \mathcal{H}_1$) – these show the data we would expect for each model, weighted by the prior probability for all values of ω

(see Etz, Haaf, Rouder, & Vandekerckhove, 2018, for details)

Uniform prior:



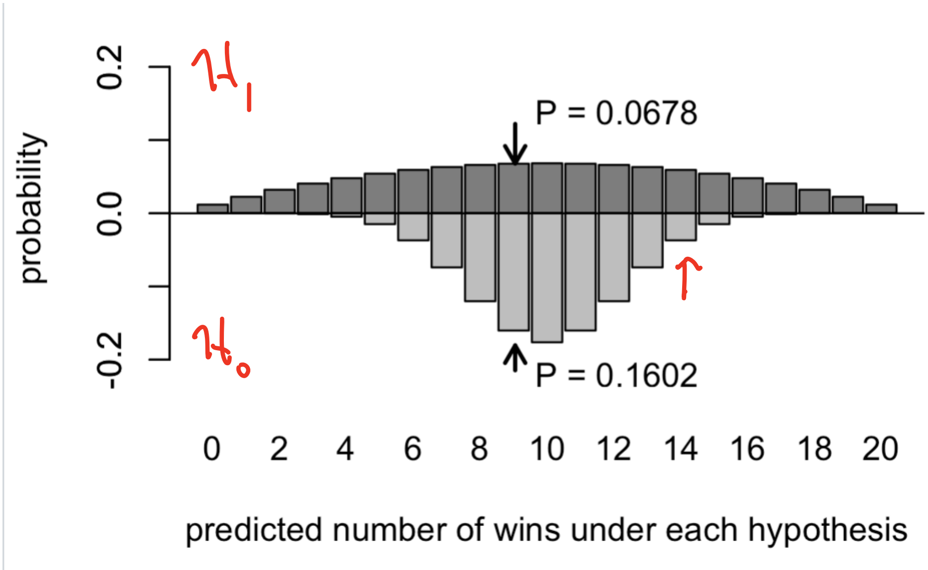
$$BF_{01} = \frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)}$$

$$= \frac{0.1602}{0.0476}$$

$$= 3.36$$

Interpretation: "the observed data are 3.36 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$."

Smooth peaked prior:



$$BF_{01} = \frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)}$$

$$= \frac{0.1602}{0.0678}$$

$$= 2.36$$

Interpretation: "the observed data are 2.36 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$."

Postscript - you may be wondering how I generated these plots. Don't worry - the goal today is to introduce the concepts. We'll begin discussing computation in the next lecture!