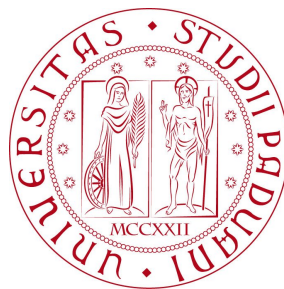


Comparing Frequentist and Bayesian inference for a Bernoulli process

Alberto Garfagnini

Università di Padova

AA 2020/2021 - Stat Lect. 8



Two different approaches

Frequentist paradigm

- it allows to perform inference about the parameter using probabilities calculated from the **sampling distribution of the data**
- the **parameter** is **unknown**, but **fixed** → **we cannot associate a probability to it**
- **the only probability is that of the random sample**
- probabilities are not conditional on the actual data sample that has been measured and are interpreted as a **long run relative frequency**
- **different types of inferences** are possible:
 - 1 - **point** estimation
 - 2 - **interval** estimation
 - 3 - **hypothesis testing**

Bayesian paradigm

- the **posterior distribution** is the **key point**
- it summarizes our **belief about the parameter**, after we have analyzed the data
- it allows to extract all the estimates on the parameter

1-Point Estimation

- a **single statistic** is calculated from the sample data and used to **estimate the unknown parameter**
- several theoretical approaches are possible: an example is the **Maximum Likelihood Estimation** (MLE)
- since **the true value of the parameter is unknown**, we can judge an estimator only on the sampling distribution of the estimator, i.e. the distribution of the estimator over all the possible random samples
- the **expected value of an estimator** measures the center of its distribution
- the **Bias of an estimator** is the difference from its expected value and the true value of the parameter

$$\text{Bias}[\hat{\theta}, \theta] = E[\hat{\theta}] - \theta$$

- an estimator is *unbiased* if the mean of its sampling distribution is the true parameter value
- the **Mean Squared Error of an estimator** is

$$\begin{aligned}\text{MSE}[\hat{\theta}] &= E[\hat{\theta} - \theta]^2 \\ &= \int (\hat{\theta} - \theta)^2 f(\hat{\theta} | \theta) d\hat{\theta}\end{aligned}$$

- it can be demonstrated that

$$\text{MSE}[\hat{\theta}] = \text{Bias}[\hat{\theta}, \theta]^2 + \text{Var}[\hat{\theta}]$$

Frequentist estimator

- in the **Frequentist** approach, an **unbiased estimator** for the **Binomial distribution** is

$$\hat{p}_F = \frac{y}{n}$$

- where **y** is the **number of successes in n trials**
- the **properties of the estimator** are:

$$E[\hat{p}_F] = p$$

$$\text{Var}[\hat{p}_F] = \frac{p(1-p)}{n} = \frac{pq}{n}$$

$$\begin{aligned}\text{MSE}[\hat{p}_F] &= \text{Bias}[p_F, p]^2 + \text{Var}[p_F] \\ &= 0^2 + \frac{p(1-p)}{n}\end{aligned}$$

Bayesian estimator



- with the **Bayesian** approach, we use the **posterior mean** as an **estimate for p**
- let's assume we imposed a **uniform prior, $\text{Beta}(1, 1)$**
- the **posterior mean** is

$$\hat{p}_B = m' = \frac{a'}{a' + b'}$$

- with **$a' = 1 + y$** and **$b' = 1 + n - y$**
- therefore

$$\begin{aligned}\hat{p}_B &= \frac{1 + y}{1 + y + 1 + n - y} = \frac{y + 1}{n + 2} \\ &= \frac{y}{n + 2} + \frac{1}{n + 2} = \frac{np}{n + 2} + \frac{1}{n + 2}\end{aligned}$$

- the **variance of the distribution** is

$$\text{Var}[\hat{p}_B] = \left(\frac{1}{n + 2}\right)^2 np(1 - p)$$

- and the **Mean Square Error** becomes

$$\begin{aligned}\text{MSE}[\hat{p}_B] &= \left[\frac{np}{n + 2} + \frac{1}{n + 2} - p\right]^2 + \left(\frac{1}{n + 2}\right)^2 np(1 - p) \\ &= \left(\frac{1 - 2p}{n + 2}\right)^2 + \left(\frac{1}{n + 2}\right)^2 np(1 - p)\end{aligned}$$

Example: point estimation

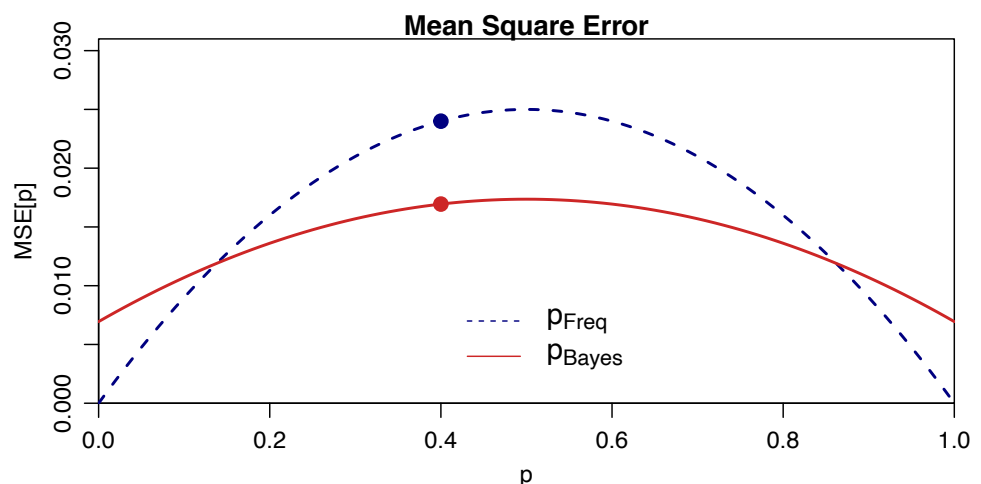
- let's suppose we have a **Bernoulli process** with **$p = \frac{2}{5}$** . We perform **multiple samples** from the distribution and the **sample size is $n = 10$**
- let's evaluate and **compare** the **Mean Square Error** for both **Frequentist** and **Bayesian estimators**
- we get

$$\text{MSE}[\hat{p}_F] = \frac{0.4 \times 0.6}{10} = 0.024$$

$$\text{MSE}[\hat{p}_B] = \left(\frac{1 - 0.8}{12}\right)^2 + \left(\frac{1}{12}\right)^2 \times 10 \times 0.4 \times 0.6 = 0.0169$$

- we can scan **scan the estimator** for different values of the **true value domain**

the Bayesian estimator is closer to the true value over most of the true value range



- we wish to find an interval (*low*, *high*) that has a predetermined probability of containing the parameter

Frequentist approach

- the parameter is fixed but unknown
- before the sample is taken, the interval endpoints are random
- once the data is known and the endpoints computed, there is nothing random anymore
- the interval is called a confidence interval for the parameter
- $(1 - \alpha) \times 100\%$ confidence interval for a parameter θ is the interval (*low*, *high*) such that

$$P(\text{low} \leq \theta \leq \text{high}) = 1 - \alpha$$

- the most common criteria used to select the interval endpoints are
 - 1 equal ordinates on the sampling distribution, $f(\text{low}) = f(\text{high})$
 - 2 equal tail area on the sampling distribution

Frequentist Interval Estimation

once the interval is calculated, there is nothing left that is random

- the interval either contains the unknown fixed parameter or it does not
- the interval can no longer be regarded as a probability interval

The correct Frequentist paradigm is:

- $(1 - \alpha) \times 100\%$ of the random intervals calculated in this way will contain the true value → we have a $(1 - \alpha) \times 100\%$ confidence that our interval does contain it
- it is a misinterpretation to make a probability statement about the parameter θ from the calculated confidence interval
- very often the sampling distribution of the estimator can be approximated with a normal distribution, with the mean equal to the true value of the parameter
- the confidence interval gets the form

$$\text{estimator} \pm \text{critical value} \times \text{estimator standard deviation}$$

- if n is large:

$$\hat{p}_f = y/n \text{ is normal with mean } p \text{ and } \sigma = \sqrt{p(1-p)/n}$$

- the approximate $(1 - \alpha) \times 100\%$ equal area confidence interval for p is

$$\hat{p}_f \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_f(1 - \hat{p}_f)}{n}}$$

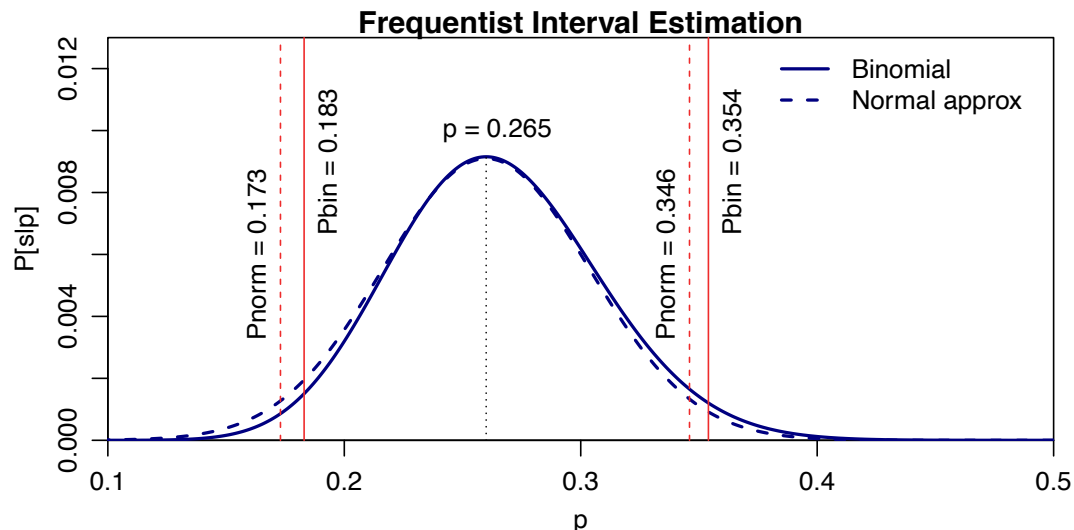
Example: interval estimation (F)

The problem

- a small town residents sample ($n = 100$) are interview about the construction of a new concert hall
- $y = 26$ express a positive opinion about it

Frequentist approach solution

- an unbiased estimator is $\hat{p}_F = y/n = 0.26$
- with standard deviation $\sigma = \sqrt{0.26 * (1 - 0.26)/100} = 0.0439$



A. Garfagnini (UniPD)

AdvStat 4 PhysAna - Stat-Lec.08

8

Example: interval estimation (B)

Bayesian approach solution

- 1 - let's select a **uniform prior**, i.e. $\text{Beta}(1, 1)$, for our unknown parameter
 - our **posterior** distribution is given by a **Beta distribution**. since a Beta prior is a conjugate function for the Binomial distribution
 - the **posterior distribution** is
- 2 - as a second example, let's choose a **Beta prior** with a mean value $m = 0.2$ and a standard deviation $\sigma = 0.08$. Since

$$m = \frac{a}{a+b} = p_o \quad \text{and} \quad \sigma_o^2 = \frac{ab}{(a+b)^2(a+b+1)} = np_o(1-p_o)$$

- it can be rewritten giving:

$$a+b+1 = \frac{p_o(1-p_o)}{\sigma_o^2} \quad \text{and} \quad a+b = \frac{a}{p_o}$$

- a **Beta(4.8, 19.2)** prior gives a posterior distribution

$$\text{Beta}(a' = a + y, b' = b + n - y) = \text{Beta}(4.8 + 26, 19.2 + 74)$$

A. Garfagnini (UniPD)

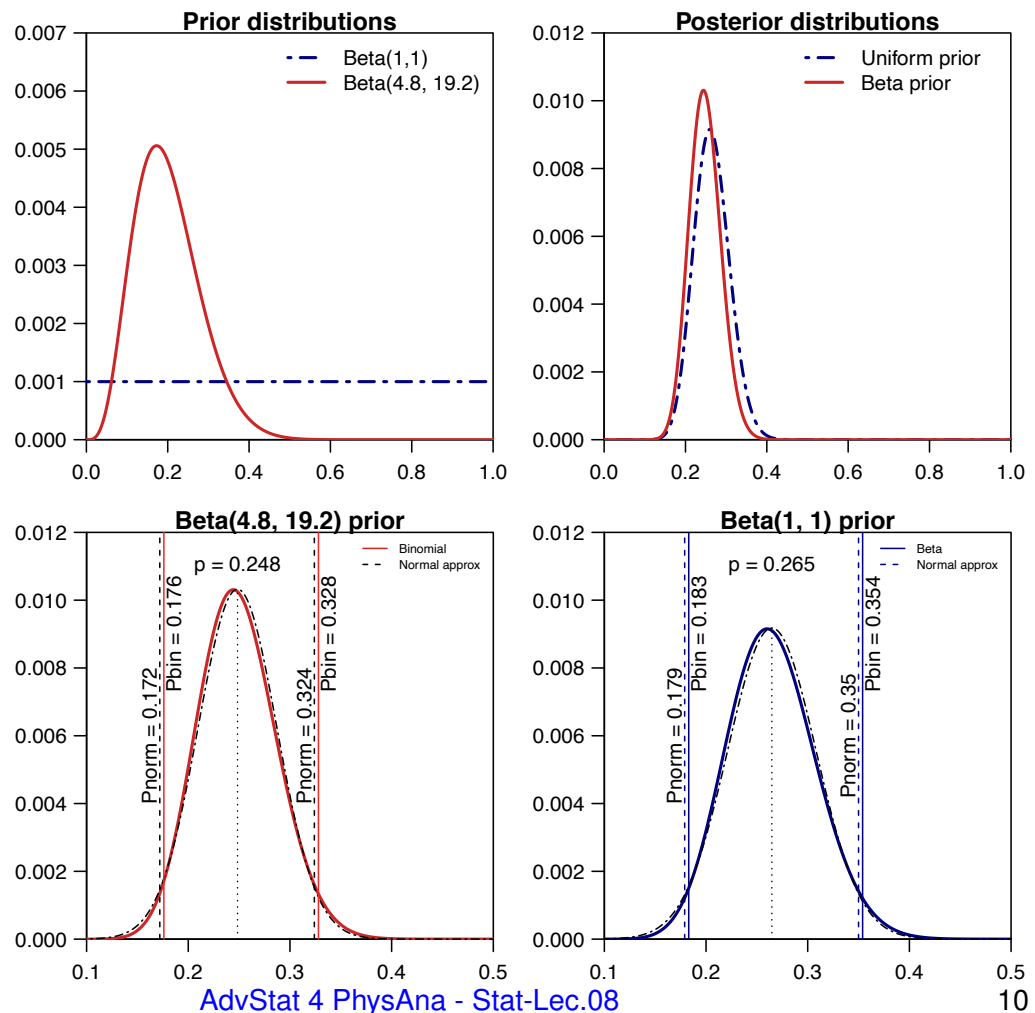
AdvStat 4 PhysAna - Stat-Lec.08

9

Example: interval estimation (B)

starting with
different prior
distribution,
we get similar
posteriors

with the posterior
distribution we
can calculate
the credibility
interval



A. Garfagnini (UniPD)

AdvStat 4 PhysAna - Stat-Lec.08

10

3-Hypothesis Testing

Idea Behind

- researchers have some theory and want to know whether or not the data actually support that theory
- scientists should not claim the discovery of a new effect if the discrepancy observed in the data could be due to chance alone
- **Hypothesis Testing**, also called **Significance Testing**, is the Frequentist statistical method used to check against claims unjustified in the data
- the nonexistence of the effect is set up as the **null hypothesis**
- when we accept the null hypothesis as true, it does not mean that we believe it is 'literally true'. Rather it means that chance alone remains a reasonable explanation for the observed discrepancy. **Therefore we cannot discard chance as the sole explanation**
- we distinguish
 - 1 testing a **one-side hypothesis** when we are interested in detecting the effect in one direction
 - 2 **two-sided hypothesis** when a test hypothesis is tested against two sided alternatives

A. Garfagnini (UniPD)

AdvStat 4 PhysAna - Stat-Lec.08

11

3-Hypothesis Testing (HT)

ESP: Extrasensory perception experiment

- θ : probability of correctly choosing the colours
- if participants have paranormal abilities: $\theta > 0.5$
- the researchers has to formulate two distinct and alternative hypotheses:
 - the **NULL Hypothesis**, H_0 : $\theta = 0.5$
 - and the **alternative Hypothesis**, H_1 : $\theta > 0.5$
- the goal of HT is not to show that the alternative hypothesis is TRUE, but to show that the null hypothesis is FALSE

The TRIAL of NULL Hypothesis

- the NULL Hypothesis is the defendant
- the researcher is the persecutor
- the statistical test is the judge

presumption of innocence: the NULL Hypothesis is deemed to be TRUE unless you, the researcher, can prove beyond reasonable doubts that it is FALSE

Errors in HT

- the goal is not to eliminate errors, but to minimize them

	accept H_0	reject H_0
H_0 is TRUE	ok	error, type I
H_0 is FALSE	error, type II	ok

→ **important design principle**: control the probability of type error I and keep it below some fixed probability α

→ α is called the significance level of the test

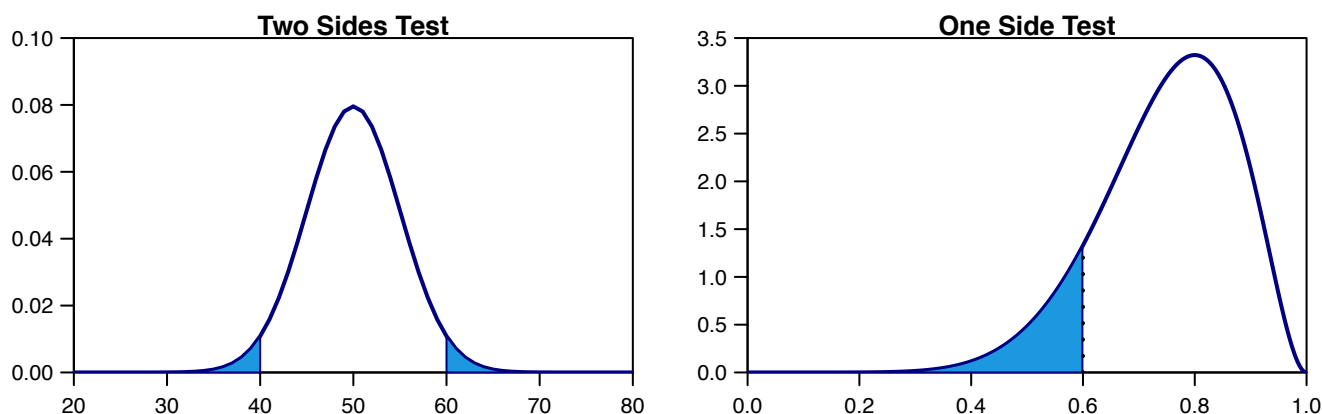
the power of the test is the probability with which we reject the NULL Hypothesis when it is really FALSE

	accept H_0	reject H_0
H_0 is TRUE	$(1 - \alpha)$ probability of correct retention	α type I error rate
H_0 is FALSE	β , type II error rate	$(1 - \beta)$, power of the test

- a powerful HT has small values of β while keeping α fixed at some small desired level
- α values used by convention among scientists: 0.05, 0.1 and 0.01

HT prescriptions

- 1) setup the NULL and alternative hypotheses
- 2) determine what the **sampling distribution of the test statistic** would be if the NULL hypothesis were TRUE
- 3) choose the **level of significance, α** and associate the critical regions to the distribution



- 4) calculate the value of the test statistic for the real data and compare to the critical value to make our decision : **critical region \rightarrow values for which we would reject the NULL hypothesis**
- 5) if we reject the NULL hypothesis, we say that **the test has produced a significant result**

Example: One-Side Hypothesis Test

The problem

- we wish to test the **effect of a new treatment**, to verify if it is better than the **standard treatment** as a parameter in the model
- p = fraction of patients who benefit from the **new treatment**
- p_o = fraction of patients who benefit from the **standard treatment**
- we know that **$p_o = 0.6$**
- **10 patients** are given the new treatment and we observe that **$y = 8$** patients benefit from the new treatment
- do we conclude that **$p > 0.6$ at the 5% level of significance** ?

Frequentist approach

- 1 - setup a null hypothesis

$$H_o : p \leq 0.6$$

- 2 - the alternative hypothesis (the new treatment is better) is

$$H_1 : p > 0.6$$

- 3 - the NULL distribution of the test statistic is the sampling distribution of the test statistic, given that the NULL hypothesis is true

$$\text{Binom}(y \mid n, p = 0.6)$$

Example: One-Side Hypothesis Test (F)

- 4 - choose a level of significance

$$\alpha = 5\%$$

Note that since y has a discrete distribution, only some values of α are possible

- 5 - the rejection region is chosen so that it has a probability of α under the NULL distribution (Neyman and Pearson approach)

$y = 8$ lies in the **acceptance region** → we do not reject H_0

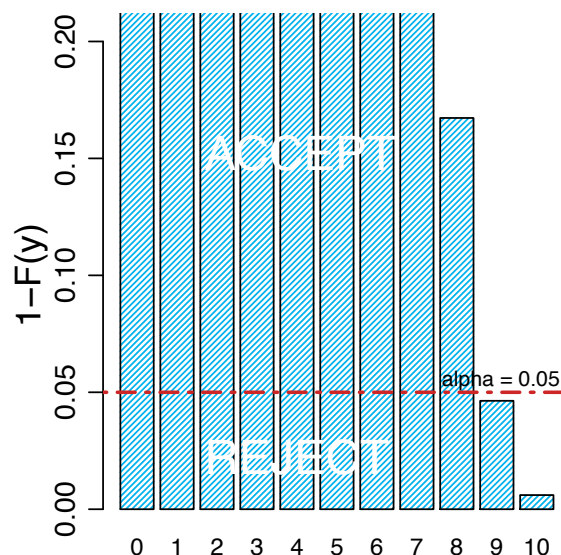
- 6 - the p-value is the probability of getting what we observed:

$$p\text{-value} = \sum_{y_{obs}}^n f(y | p_0) = 0.1672$$

if $p\text{-value} < \alpha$ → the test statistic lies in the rejection region

α represents the long-run rate of rejecting a true null hypothesis

- 7 - an alternative way, due to Fisher, is to reject H_0 if $p\text{-value} < \alpha$



Example: One-Side Hypothesis Test (B)

Bayesian approach

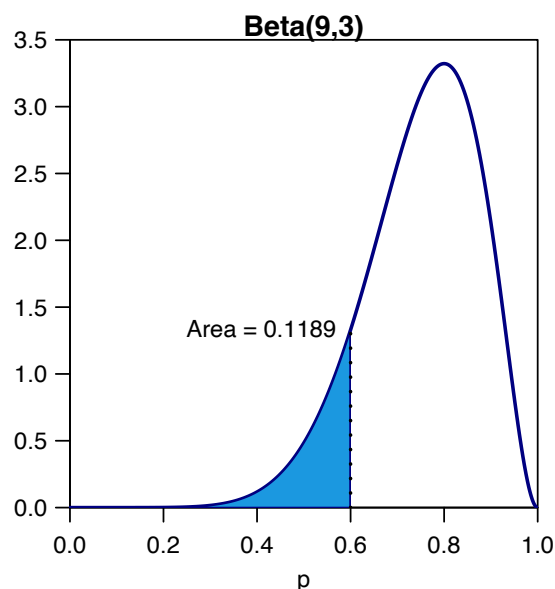
- we wish to test $H_0 : p \leq p_0$ versus $H_1 : p > p_0$ at a level of significance α
- we evaluate the posterior probability of the null hypothesis, and integrate over the required region:

$$P(H_0 : p \leq p_0 | y) = \int_0^{p_0} g(p | y) dp$$

- we reject the null hypothesis if the posterior probability is less than α , the level of significance
- we use a uniform prior, $\text{Beta}(1, 1)$, for the parameter p
- given $y = 8$, the posterior density is $\text{Beta}(9, 3)$

$$\begin{aligned} P(p \leq 0.6 | y = 8) &= \int_0^{0.6} \frac{\Gamma(12)}{\Gamma(3)\Gamma(9)} p^8 (1-p)^2 dp \\ &= 0.1189 \end{aligned}$$

- the result, **11.89%**, is higher than $\alpha = 5\%$, therefore **we cannot reject the null hypothesis** at the 5% level of significance



Example: Two-Sides Hypothesis Test

- we want to detect any changes from the value p_0
- we setup the null hypothesis $H_0 : p = p_0$ against the alternative hypothesis $H_1 : p \neq p_0$

The problem

- a coin is tossed $n = 15$ times
- we observe $y = 10$ heads

Q: Is the coin fair ?

Frequentist approach

- 1 - setup a null hypothesis

$$H_0 : p = 0.5$$

- 2 - we want to test it against the alternative hypothesis

$$H_1 : p \neq 0.5$$

- 3 - the null distribution is the sampling distribution of y : $\text{Bin}(y \mid n = 15, p = 0.5)$

Example: Two-Sides Hypothesis Test (F)

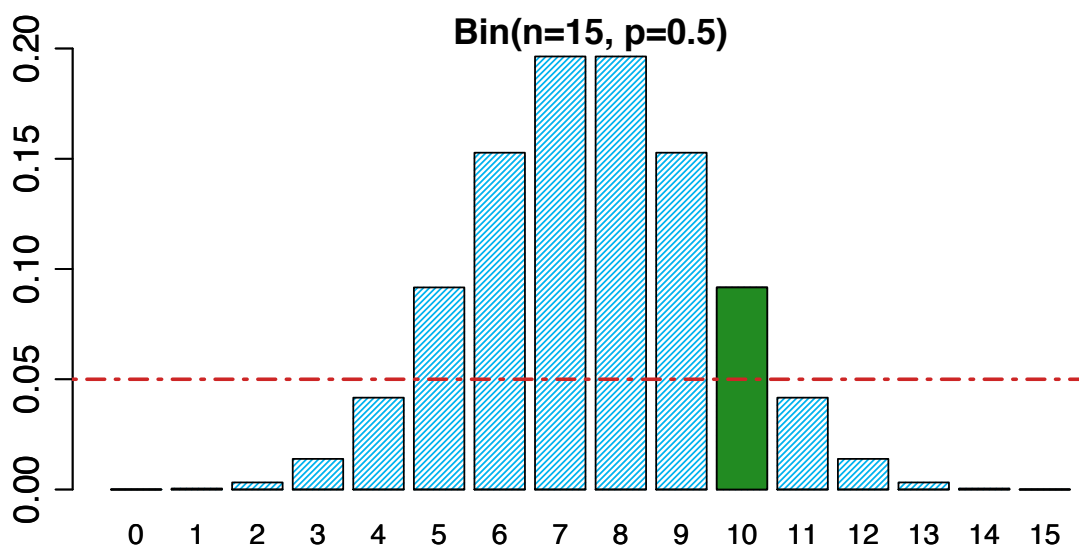
- 4 - in defining the rejection region, we take into account that y has a discrete distribution, and choose the level of significance as close to 5% as possible

$$\{y \leq 3\} \cup \{y \geq 12\} \text{ with } \alpha = 0.0352$$

- 5 - we observe $y = 10$, which lies inside the acceptance region

- 6 - we would have not rejected the null hypothesis also evaluating the p-value

$$P(y \geq 10) + P(y \leq 5) = 0.3018$$



Example: Two-Sides Hypothesis Test (B)

Bayesian approach

- the posterior distribution of the parameter, given the data, constraints our entire belief after getting the data
- but since the probability of an exact value represented by the point null hypothesis is zero
- need a correspondence similar to that of confidence intervals, using **credible intervals**
- we compute a $(1 - \alpha) \times 100\%$ credible interval for p
- if p_0 lies inside the interval, we do not reject the null hypothesis, H_0 ; if it is outside, we reject H_0 .

The problem

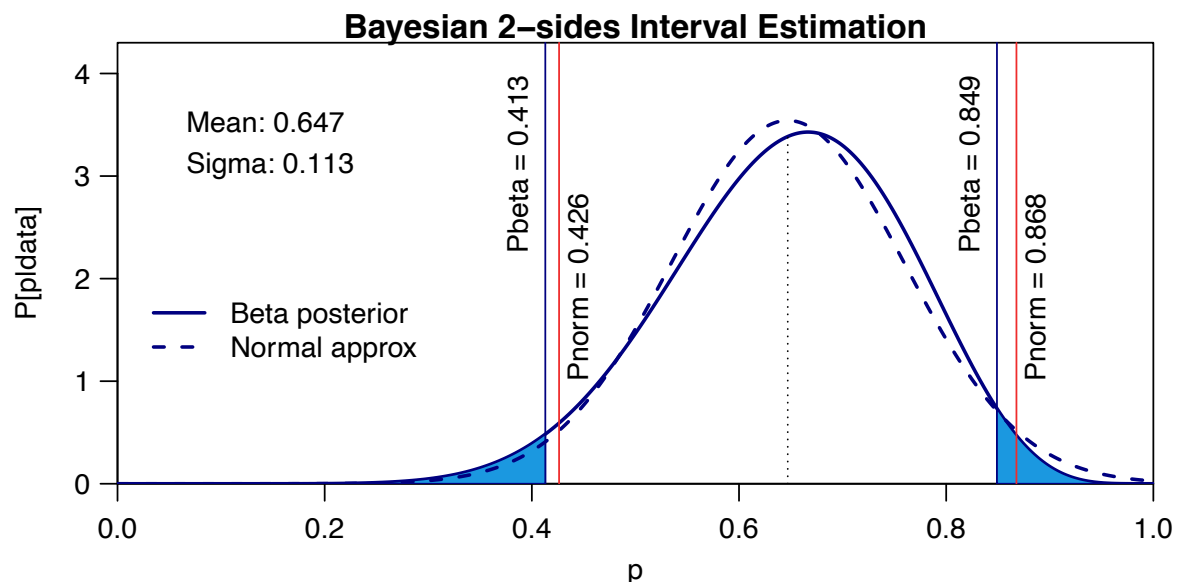
- $n = 15$ coin tosses. We observe $y = 10$ heads
- 1 - set up a uniform prior $\text{Beta}(1, 1)$
- 2 - the posterior is $\text{Beta}(10 + 1, 5 + 1)$
- 3 - we calculate a 95% Bayesian credible interval

Example: Two-Sides Hypothesis Test (B)

- 4 - using a normal approximation we would get

$$\frac{11}{17} \pm 1.96 \times \sqrt{\frac{11 \times 6}{(11 + 6)^2(11 + 6 + 1)}} = 0.647 \pm 0.221$$

- 5 - our credibility interval is:
 - (0.413, 0.849), using a Beta posterior
 - (0.426, 0.868), using a Normal approximation



Some considerations on the p -value of the test

Neyman view

- the HT described does not make a distinction at all between a result that is **barely significant** and those **highly significant**
- let's run several HT on the same data:

Value of α	0.05	0.04	0.03	0.02	0.01
Reject H_0 ?	Y	Y	Y	N	N

- between 0.02 and 0.03 there is a value of α that would allow us to reject the NULL hypothesis

the p -value is defined to be the smallest Type I error rate (α) that we are willing to tolerate if we want to reject the NULL hypothesis

- p summarizes all the possible hypothesis tests that we could have run:
if $p \leq \alpha$ we would reject the NULL hypothesis

Some considerations on the p -value of the test

but

- the p value is not the probability that the NULL hypothesis is TRUE
- this statement is absolutely and completely wrong:
 - 1) NULL Hypothesis testing is a frequentist tool: we are not allowed to assign probability to a NULL hypothesis
according to this view of probability, the NULL hypothesis is either TRUE or FALSE

- R contains a whole lot of functions corresponding to different kinds of hypothesis test

```
binom.test(x=62, n=100, p=0.5)
```

```
Exact binomial test
```

```
data: 62 and 100
number of successes = 62, number of trials = 100, p-value = 0.02098
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5174607 0.7152325
sample estimates:
probability of success
              0.62
```

Summary - global considerations

Frequentist paradigm

- it handles, separately, point estimation, confidence intervals and hypothesis tests
- the Frequentist statistics considers the parameter a fixed but unknown constant
- the sampling distribution of a statistic is its distribution over all the possible random samples, given the fixed parameter value
- the only probability allowed is a long-run relative frequency

Bayesian paradigm

- it bases all the estimates on the posterior distribution of the parameter

Frequentist paradigm

- a $(1 - \alpha) \times 100\%$ Frequentist interval for a parameter θ is an interval (θ_l, θ_h) such that

$$P(\theta_l \leq \theta \leq \theta_h) = 1 - \alpha$$

- $(1 - \alpha) \times 100\%$ of the random intervals calculated this way do contain the true value \rightarrow we say we are $(1 - \alpha) \times 100\%$ confident that the calculated interval contains the true parameter
- the p-value allows to reject the null hypothesis, at level α , if $\text{p-value} < \alpha$
- the p-value is not the probability the null hypothesis is true. It is the probability of observing what we observed given that the null hypothesis is true

Bayesian paradigm

- a $(1 - \alpha) \times 100\%$ Bayesian credible interval for a parameter θ is a range of parameter values that has a posterior probability $(1 - \alpha)$