Hypothesis Testing and Bayesian Statistics

Logit

2016

TABLE OF CONTENTS

Hypothesis Testing

Intro to Hypothesis Testing

Hypothesis testing provides a structured framework for evaluating parameters and other items in statistics. With hypothesis testing, we judge a *null hypothesis* against an *alternative hypothesis*, weighing evidence almost like a legal trial.

Motivating Example: Car Mileage

A car company claims that its latest car model has a fuel economy of 27 miles per gallon. If we have a sample of 100 of these cars and test their fuel economy over the course of a month, how can we check the validity of the car company's claim?

Population Mean vs. Sample Mean

The population mean refers to the average value of some quantity for the entire population of something (e.g. people, animals, manufactured goods). We usually cannot know the population mean, but we may be able to take a sample and get a sample mean, in order to estimate the population mean.

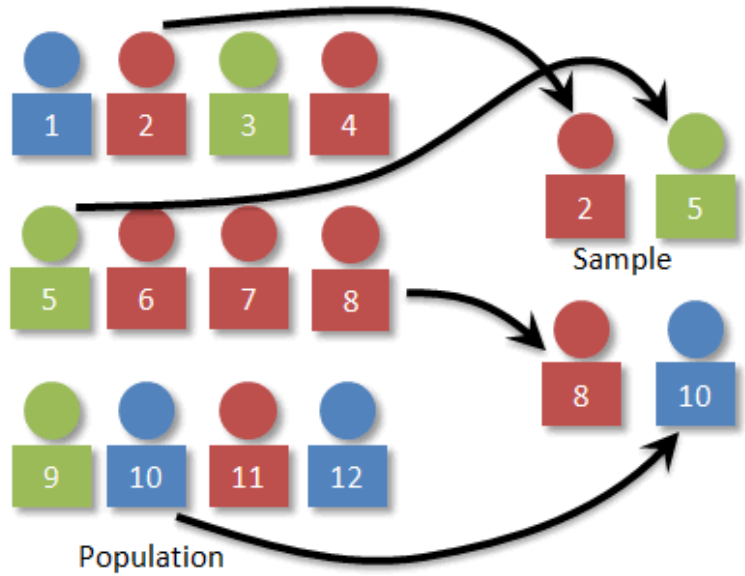
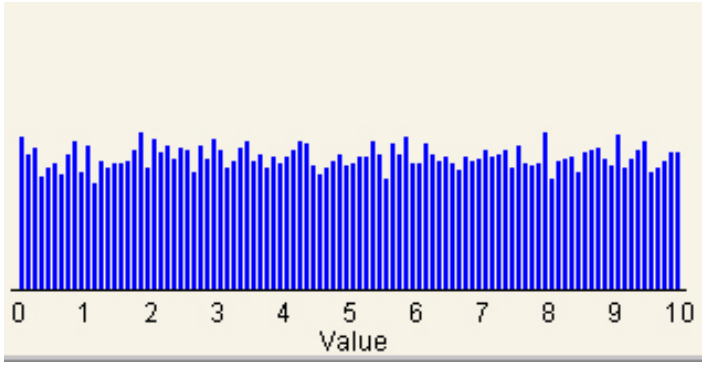


Image Source

Central Limit Theorem (CLT)

According to the Central Limit Theorem (CLT), if we have a collection of independent and identically distributed random variables $\{X_1, \ldots, X_n\}$ each with mean μ and variance σ^2, and $S_n = X_1 + \cdots + X_n$, then S_n should be normally distributed with mean $n\mu$ and variance $n\sigma^2$.

At the top right is a sample from a single uniform random variable, while the bottom right is a sample from the average of 10 uniform random variables.



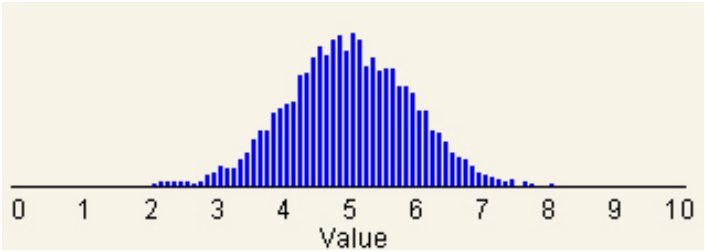
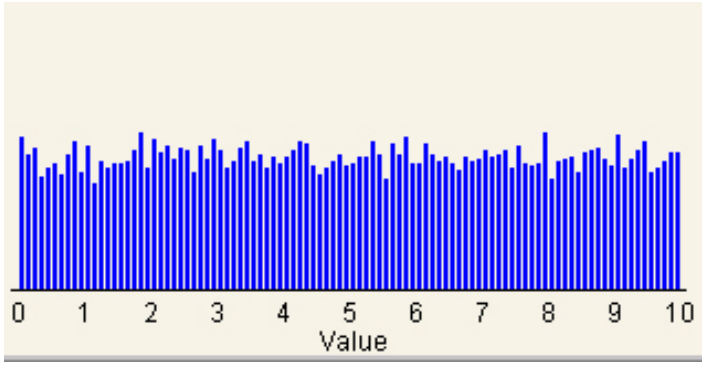


Image Source

Central Limit Theorem (CLT)

If we use CLT, we find that any collection of sample means should be normally distributed in this bell curve. We can use the properties of the normal distribution in our analysis of sample means.



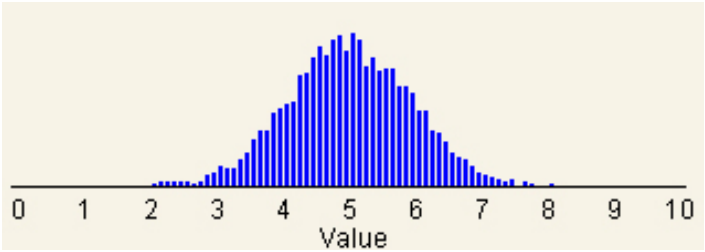


Image Source

Standard Error

The standard error is the standard deviation of the sample distribution. The standard error for the sample mean $\overline{x}$ is

$$\mathrm{SE}_{\overline{x}} = \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation, and n is the sample size.

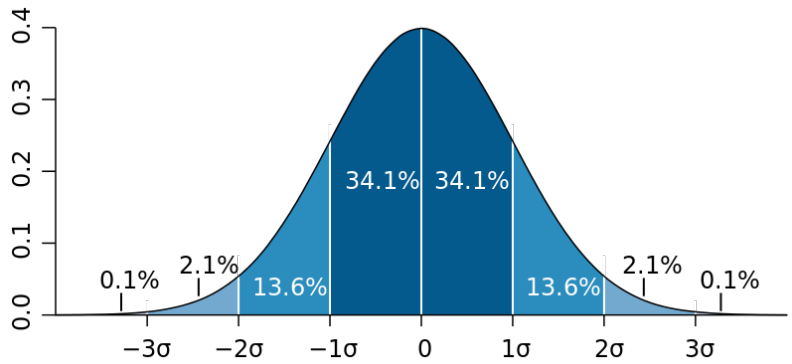


Image Source

Confidence Interval

If we combine the sample mean $\overline{x}$ with the standard error $SE_{\overline{x}}$, then we can establish a confidence interval where we would expect the true population mean μ to lie.

Example: we can say with 95% confidence that the population mean μ is in the range

$$\overline{x} - 1.96(SE_{\overline{x}}) \leq \mu \leq \overline{x} + 1.96(SE_{\overline{x}})$$

$$\overline{x} - 1.96\frac{s}{\sqrt{n}} \leq \mu \leq \overline{x} + 1.96\frac{s}{\sqrt{n}}$$

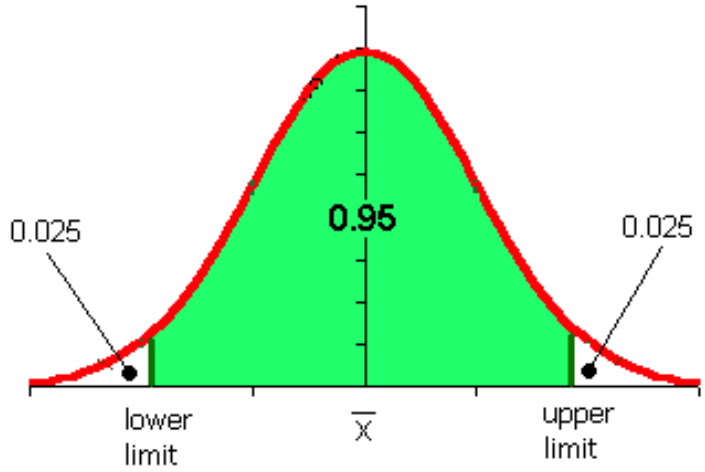


Image Source

Confidence Interval

$$\bar{x} - z^* \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z^* \frac{s}{\sqrt{n}}$$

For a 95% confidence interval on a normal distribution, $z^* = 1.96$. For a 99% confidence interval, $z^* = 2.58$

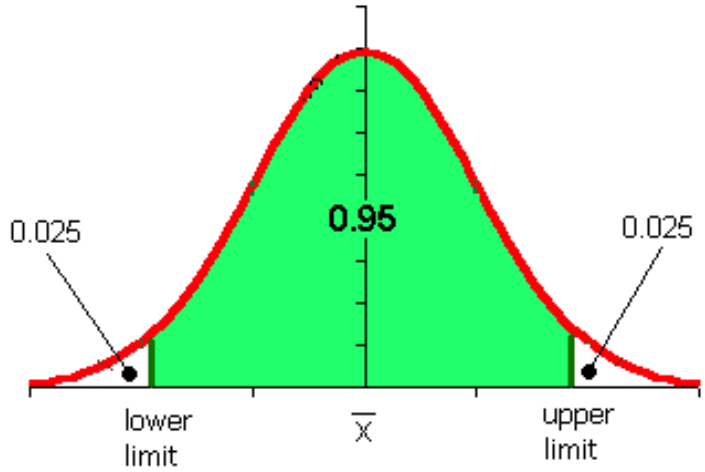


Image Source

Confidence Interval: Car Mileage Example

Suppose the fuel economy test for 100 cars yields sample mean $\overline{x}$ and sample standard deviation s:

$$\overline{x} = 25, \qquad s = 5$$

Then the standard error $\mathrm{SE}_{\overline{x}}$ is

$$\mathrm{SE}_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{5}{\sqrt{100}} = 0.5$$

The 95% confidence interval for the true population mean is

$$\overline{x} \pm 1.96 \times \mathrm{SE}_{\overline{x}} = 25 \pm 0.98$$

Interpreting Confidence Intervals

Correct: We can say with 95% confidence that the population mean for fuel economy lies between 24.02 and 25.98 miles per gallon.

Incorrect: There is a 95% chance that the population mean for fuel economy lies between 24.02 and 25.98 miles per gallon.

The 95% confidence interval tells us that, were we to take repeated samples of the cars, the computed confidence interval would encompass the true population mean 95% of the time.

Null Hypothesis vs. Alternative Hypothesis

Null Hypothesis (H_0)**:** The null hypothesis is our base assumption about the value of a parameter. This is analogous to assuming a defendant is innocent in a trial.

Alternative Hypothesis (H_A)**:** The alternative hypothesis is the logical negation of the null hypothesis. This is analogous to the idea that the defendant is guilty.

We will not reject the null hypothesis (H_0) unless the evidence supporting the alternative hypothesis (H_A) is sufficiently strong.

Hypotheses: Car Mileage Example

Null Hypothesis (H_0)**:** The population mean for fuel economy of these cars is $\mu = 27$ miles per gallon.

Alternative Hypothesis (H_A)**:** The population mean for fuel economy of these cars is **not** $\mu = 27$ miles per gallon, but is in fact either higher or lower.

Hypothesis Test

Once we have the null hypothesis (H_0) and alternative hypothesis (H_A), we can conduct the hypothesis test by calculating the z-score:

$$z = \frac{m - \overline{x}}{s/\sqrt{n}}$$

where

$$\begin{aligned} m &\leftarrow \text{null hypothesis value of mean} \\ \overline{x} &\leftarrow \text{sample mean} \\ s &\leftarrow \text{sample standard deviation} \\ n &\leftarrow \text{sample size} \end{aligned}$$

If the z-score is associated with a p-value (probability) that is sufficiently small, we reject the null hypothesis (H_0). Otherwise, we fail to reject the null hypothesis.

Hypothesis Test with Confidence Interval

Alternatively, we can do a hypothesis test using a confidence interval. If the null hypothesis (H_0) value of the mean m is outside the confidence interval, we reject the null hypothesis. Otherwise, if m is inside the confidence interval, we fail to reject the null hypothesis.

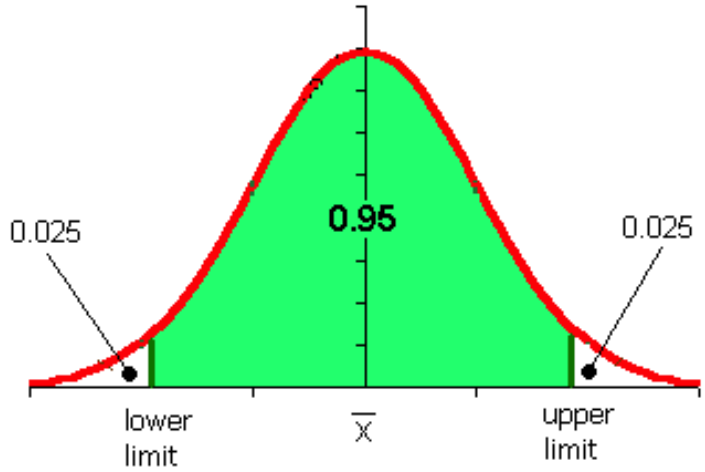


Image Source

Interpretation of Hypothesis Test

Correct: "We reject the null hypothesis," *or* "We fail to reject the null hypothesis."

Incorrect: "We accept the null hypothesis," *or* "We accept the alternative hypothesis."

We have to be careful in our wording because hypothesis testing does not offer 100% definitive proof. There is always a chance that we made an error.

Hypothesis Testing: Car Mileage Example

H_0: Population average $\mu = 27$ m.p.g.

H_A: Population average $\mu \neq 27$ m.p.g.

$$z = \frac{m - \overline{x}}{s/\sqrt{n}} = \frac{27 - 25}{5/\sqrt{100}} = 4$$

For z-score of 4, $p = 6.3 \times 10^{-5}$. This p-value is so small, we can reject the null hypothesis with over 95% confidence.

Hypothesis Testing: Car Mileage Example

H_0: Population average $\mu = 27$ m.p.g.

H_A: Population average $\mu \neq 27$ m.p.g.

Alternatively, the 95% confidence interval for this sample is

$$\overline{x} \pm 1.96 \times \frac{s}{\sqrt{n}} = 25 \pm 1.96 \times \frac{5}{\sqrt{100}}$$

This is the interval $(24.02, 25.98)$, and the value from the null hypothesis does not lie in this range, so we reject the null hypothesis.

Type I and Type II Error

	Fail to Reject H_0	Reject H_0
H_0 True	OK	Type I Error
H_A True	Type II Error	OK

Type I Error (False Positive): Rejecting H_0 when it is true.

Type II Error (False Negative): Failing to reject H_0 when it is false.

Type I and Type II Error

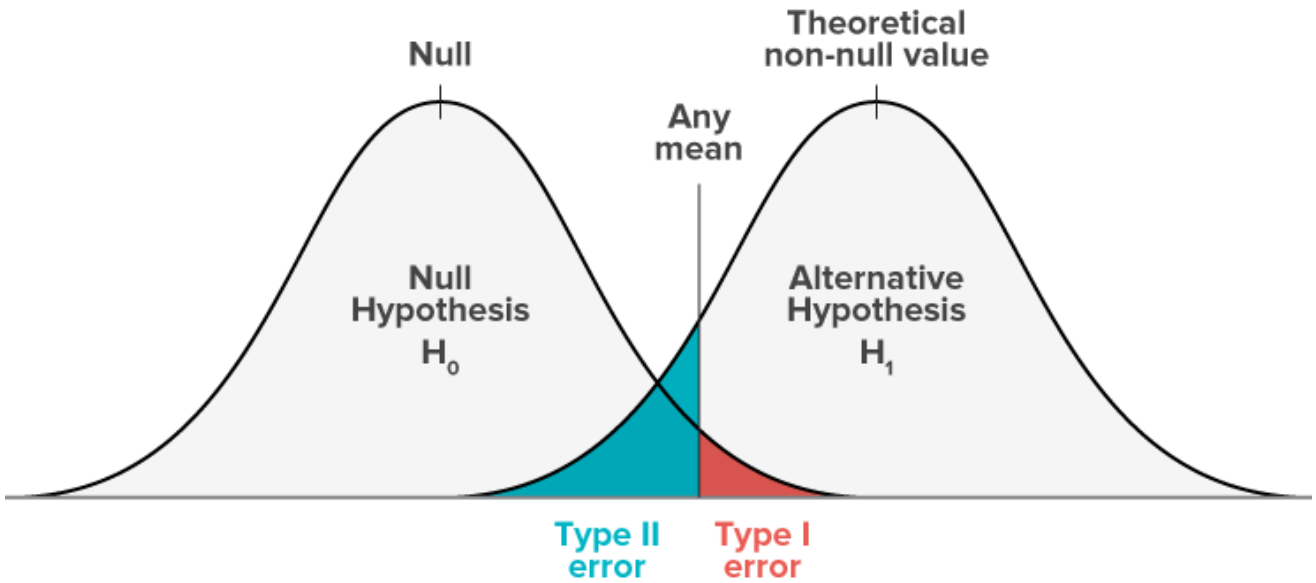


Image Source

	Fail to Reject H_0	Reject H_0
H_0 True	OK	Type I Error
H_A True	Type II Error	OK

Statistical Significance

The statistical significance level α of a test tells us how small our p-value must be in order for the test to be *statistically significant* at a level $(1 - \alpha)$.

For instance, if $\alpha = 0.05$, then we can say the test is statistically significant with a confidence level of 95% if the p-value of the test is less than 0.05.

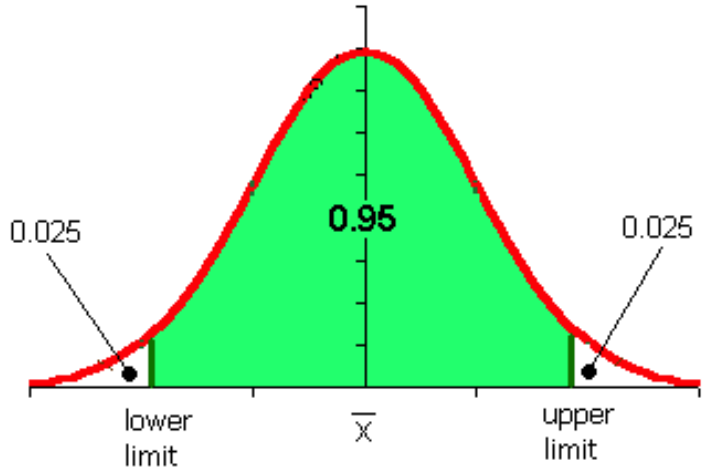


Image Source

One-sided vs. Two-sided Tests

The diagram at the right represents a two-sided hypothesis test, which we use to check if the true population mean is different from the null hypothesis.

We can also use a one-sided hypothesis test if we only want to check if the true population mean is either greater than the null hypothesis or less than the null hypothesis.

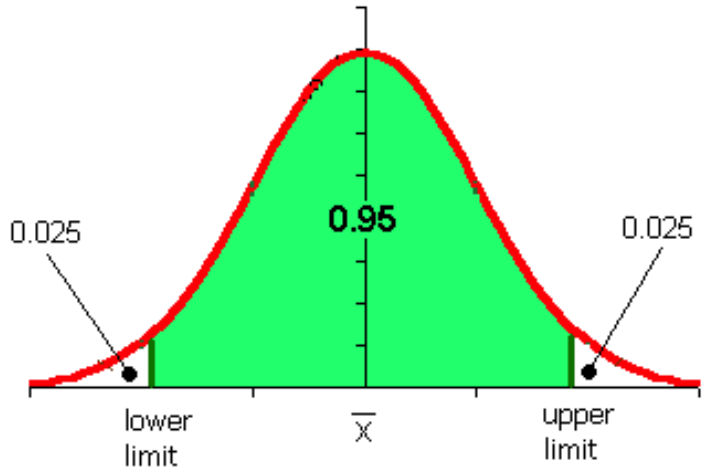


Image Source

One-sided vs. Two-sided Tests

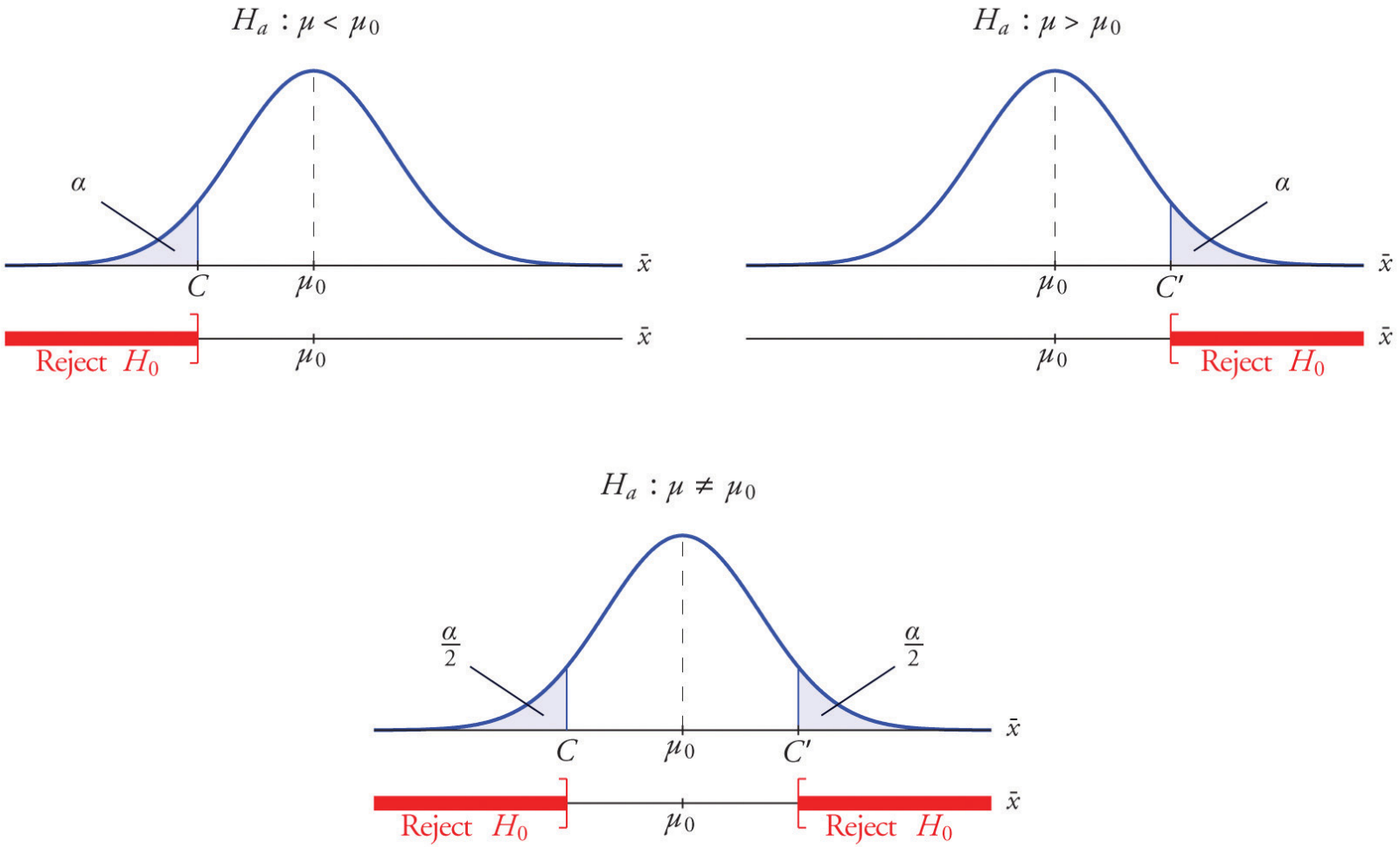


Image Source

χ^2 Test of Independence

The χ^2 test of independence provides a structured way to check whether or not two categorical variables X and Y are independent of each other.

$$H_0: \quad X \text{ is independent of } Y.$$
$$H_A: \quad X \text{ and } Y \text{ are dependent.}$$

To check this, we need to calculate the χ^2 test statistic:

$$\chi^2 = \sum \left[(O - E)^2 / E \right]$$

where O represents observed values, while E represents expected values.

χ^2 Distribution and Degrees of Freedom

The χ^2 distribution depends on the degrees of freedom, represented by k in the graph on the right. The degrees of freedom can be calculated from a table of data as

$$k = (r-1)(c-1)$$

where r is the number of rows, and c is the number of columns.

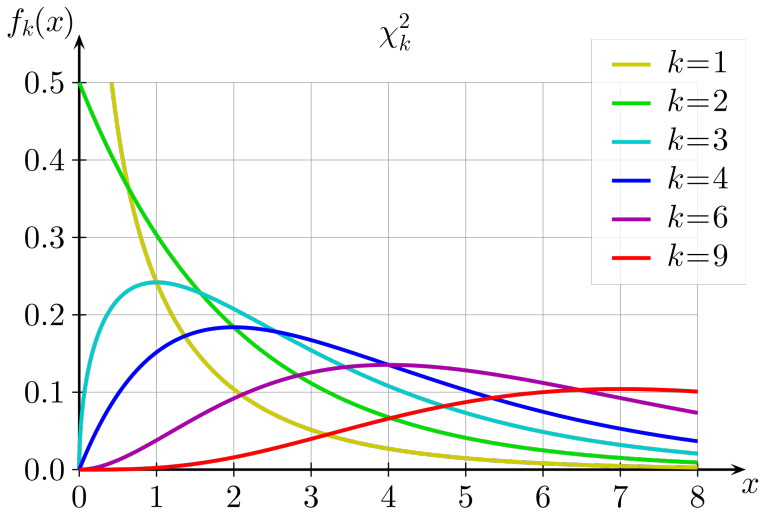


Image Source

χ^2 Independence Example: Gender vs. Pie Flavor

Suppose we have data on the preferred pie flavor from a sample of men and women. How can we use the χ^2 test of independence to see if pie flavor preference is independent of gender?

χ^2 Independence Example: Gender vs. Pie Flavor

Observed values:

	Apple	Pecan	Pumpkin	Total
Male	25	20	35	80
Female	40	32	48	120
Total	65	52	83	200

From the observed values, we can calculate "expected values" that would happen if these variables were truly independent by using the proportions of numbers in the margins.

χ^2 Independence Example: Gender vs. Pie Flavor

Observed values:

	Apple	Pecan	Pumpkin	Total
Male	25	20	35	80
Female	40	32	48	120
Total	65	52	83	200

Expected values:

	Apple	Pecan	Pumpkin	Total
Male	26	20.8	33.2	80
Female	39	31.2	49.8	120
Total	65	52	83	200

χ^2 Independence Example: Gender vs. Pie Flavor

$$\begin{aligned}\chi^2 &= \sum[(O-E)^2/E] \\ &= \frac{(25-26)^2}{26} + \frac{(20-20.8)^2}{20.8} + \frac{(35-33.2)^2}{33.2} \\ &+ \frac{(40-39)^2}{39} + \frac{(32-31.2)^2}{31.2} + \frac{(48-49.8)^2}{49.8} \\ &= 0.278035\end{aligned}$$

The critical value for the χ^2 test with 2 degrees of freedom and $\alpha = 0.05$ is $\chi^2 = 5.99$. Since χ^2 is less than 5.99, we fail to reject the null hypothesis that gender and pie flavor preference are independent.

χ^2 Goodness of Fit Example: Handedness

We can also use a χ^2 test to check for "goodness of fit," i.e. to see if a sample distribution follows our assumptions.

For example, suppose we want to check if more than 80% of people are right-handed. We take a sample of 100 people and 88 of them are right-handed. How can we apply a hypothesis test to analyze this problem?

χ^2 Goodness of Fit Example: Handedness

H_0: $\mathbb{P}(\text{right-handed}) \leq 0.80$.

H_A: $\mathbb{P}(\text{right-handed}) > 0.80$.

$$\begin{aligned}\chi^2 &= \sum[(O-E)^2/E] \\ &= \frac{(88-80)^2}{80} + \frac{(12-20)^2}{20} \\ &= 4\end{aligned}$$

We have only one degree of freedom since there are only two choices (right handed, left handed). The critical value of χ^2 with 1 degree of freedom and $\alpha = 0.05$ is 3.84. Since our value of χ^2 is greater than 3.84, we reject the null hypothesis that the probability of being right-handed is 80% or less.

Power

Sometimes, a hypothesis test fails to reject a null hypothesis, and we are left with a couple options:

- Accept the null hypothesis as a reasonable conclusion.
- Consider whether the sample was too small to reasonably reject or fail to reject the null hypothesis.

This is why it is useful to consider the power of a hypothesis test.

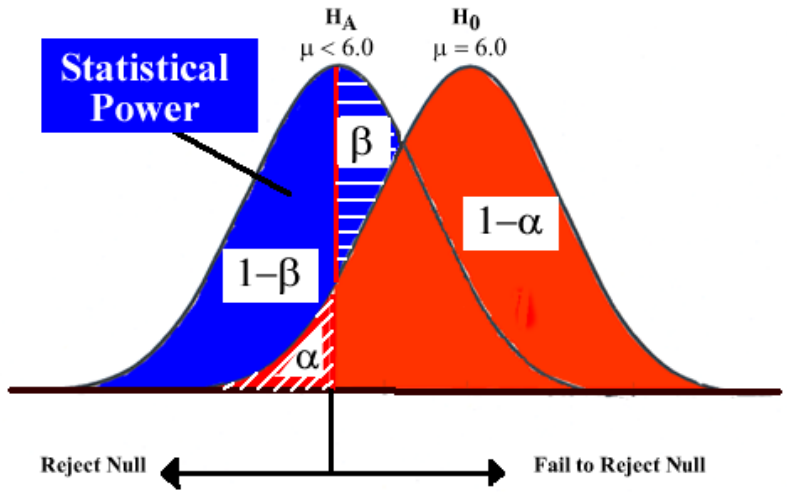


Image Source

Power

	Fail to Reject H_0	Reject H_0
H_0 True	OK	Type I Error
H_A True	Type II Error	OK

$$\alpha \leftarrow \text{Type I Error}$$
$$\beta \leftarrow \text{Type II Error}$$
$$(1-\beta) \leftarrow \text{Power}$$

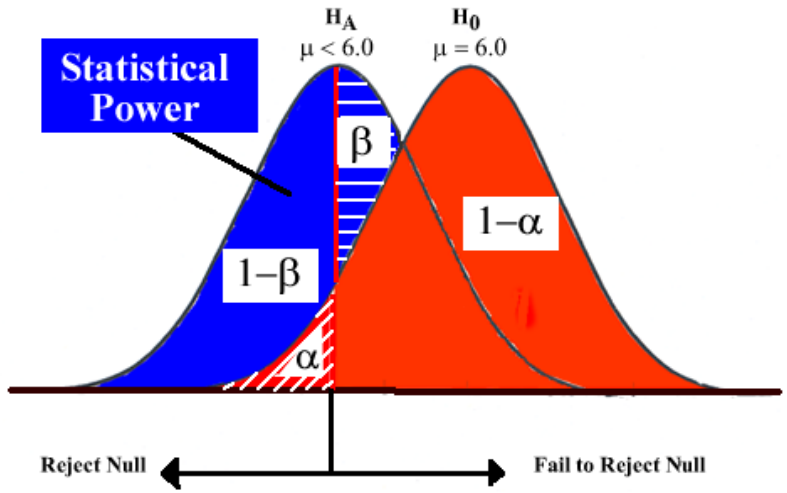


Image Source

The power of a hypothesis test is the probability of deciding correctly when H_A is true.

Power

$$\alpha \leftarrow \text{Type I Error}$$
$$\beta \leftarrow \text{Type II Error}$$
$$(1-\beta) \leftarrow \text{Power}$$

To decrease both α and β, the sample size will have to be increased. Given $\alpha, \beta, \mu_\alpha$ (where μ_α is the likely value of μ where you want to evaluate power), the minimum sample size for a one-tailed test is

$$n = \frac{\sigma^2(Z_\alpha + Z_\beta)^2}{(\mu_0 - \mu_\alpha)^2}$$

For a two-tailed test:

$$n = \frac{\sigma^2(Z_{\alpha/2} + Z_\beta)^2}{(\mu_0 - \mu_\alpha)^2}$$

Statistical Power Example: Student Height

We took a random sample of heights from 25 college students at a local university. We are given the standard deviation: $\sigma = 9$ cm. We want to test:

$$H_0 : \mu = 170$$
$$H_A : \mu > 170$$

What is the power of this hypothesis test, using 95% confidence, if the true population mean is $\mu = 175$ cm?

Statistical Power Example: Student Height

We want to find which sample heights will cause us to reject the null hypothesis.

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

$$\overline{x} = \mu + z\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$= 170 + 1.645\left(\frac{9}{\sqrt{25}}\right)$$

$$= 172.96$$

We will reject the null hypothesis when the sample mean is greater than 172.96 cm.

Statistical Power Example: Student Height

Now, we want to calculate the power of the hypothesis test.

$$\begin{aligned}
\text{Power} &= \mathbb{P}(\bar{x} > 172.96 | \mu = 175) \\
&= \mathbb{P}\left(z \geq \frac{172.96 - 175}{9/\sqrt{25}}\right) \\
&= \mathbb{P}(z \geq -2.039) \\
&\boxed{= 0.979}
\end{aligned}$$

The statistical power of our hypothesis is quite high (0.979).

Hypothesis Test of Proportion

With a hypothesis test of proportion, we assess whether a sample from a population represents the true proportion from the population as a whole.

$$H_0 : p = p_0$$
$$H_A : p \neq p_0$$

where p_0 is the null hypothesized proportion. For this test, we calculate a z-score differently:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Here, $\hat{p}$ is the sample proportion. From the z-score, we can determine whether to reject / fail to reject the null hypothesis.

Hypothesis Test of Proportion Example: Handedness

Let's revisit the problem of handedness. Suppose we assume that 80% of people are right-handed, but we have a sample with 100 people, 88 of whom are right-handed. How can we use the hypothesis test of proportion to check whether our sample lines up with our assumption?

Hypothesis Test of Proportion Example: Handedness

$$H_0 : p = 0.80$$
$$H_A : p \neq 0.80$$

Our sample of 100 people has 88 right-handed people and 12 left-handed people. Calculating the z-score:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.88 - 0.80}{\sqrt{\frac{(0.80)(0.20)}{100}}} = 2$$

Since the z-score is greater than 1.96, we can (with 95% confidence) reject the null hypothesis that our sample lines up with a population proportion of $p = 0.80$.

Bayesian Statistics

Frequentist vs. Bayesian

The frequentist approach to statistics uses the assumption that probabilities represent long-run frequencies with which events occur. Data is considered a repeatable random sample. The underlying parameters remain constant in this sampling process.

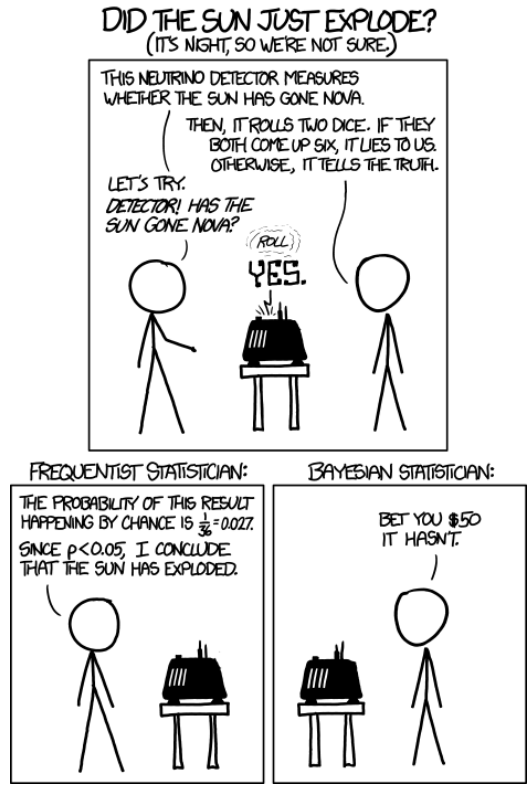


Image Source

Frequentist vs. Bayesian

The Bayesian approach to statistics treats unknown quantities probabilistically. The state of the world can be updated whenever. Bayesians consider probability as a measure of subjective degree of belief. Bayesian statistics involves choosing a *prior* that represents prior information.

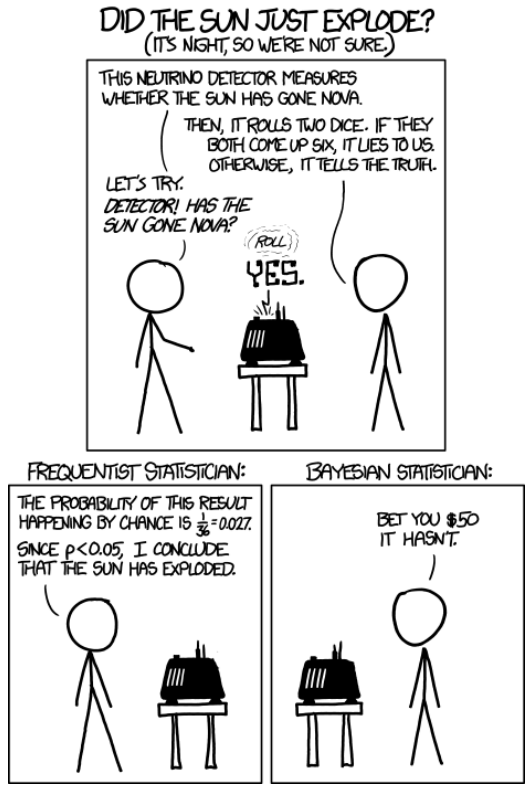


Image Source

Elements of Bayesian Statistics

Bayesian statistics considers these quantities (among others):

- $\mathbb{P}(\theta)$ - *Prior distribution* - This distribution considers assumptions about the distribution of the parameters (θ).
- $\mathbb{P}(X|\theta)$ - *Likelihood* - This is the distribution of data given the parameters. This is the probability model assumed to have created the data.
- $\mathbb{P}(X)$ - *Marginal distribution* - This is the distribution of the observed data marginalized (summed in the margins) over all possible values of the parameters (θ).
- $\mathbb{P}(\theta|X)$ - *Posterior distribution* - This is the distribution of parameters given the observed data.

Elements of Bayesian Statistics

As a consequence of Bayes' Theorem:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

we conclude that

$$\mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)\mathbb{P}(\theta)$$

i.e. the posterior distribution is proportional to the likelihood times the prior distribution.

Example: Monty Hall Problem

On a game show, you have to choose between 3 doors. Behind one door is a car, but behind the other two are goats. You pick one door, and the host (Monty Hall) opens one of the two remaining doors to reveal a goat. The host asks if you want to keep your original door choice, or switch to the other door. Which should you pick?

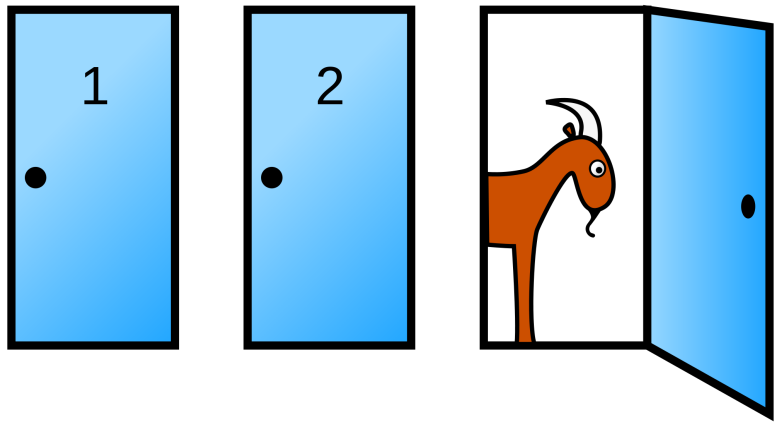


Image Source

Example: Monty Hall Problem

Some people believe that switching doors should not affect the probability of getting the car. They intuitively believe this because the *prior* distribution is assumed to be uniform: the probability of the car being behind any given door at the start is $\frac{1}{3}$. In reality, the act of opening one of the remaining doors changes the *likelihood* for the two unopened doors.

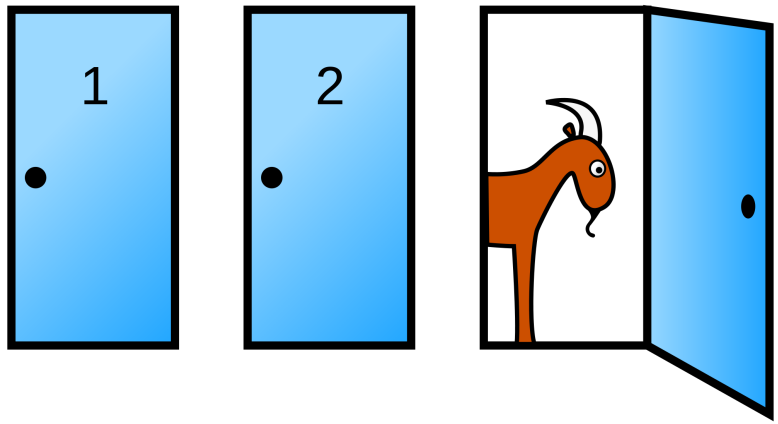


Image Source

Example: Monty Hall Problem

Door	Prior $\mathbb{P}(\theta)$	Likelihood $\mathbb{P}(X\|\theta)$	$\mathbb{P}(\theta)\mathbb{P}(X\|\theta)$	Posterior $\mathbb{P}(\theta\|X)$
1	$1/3$	$1/2$	$1/6$	$1/3$
2	$1/3$	0	0	0
3	$1/3$	1	$1/3$	$2/3$

As we said before, the prior probabilities for all 3 doors are $1/3$. Suppose we pick door 1, and Monty opens door 2 to reveal a goat. If the car is behind door 1, the likelihood that Monty would have opened door 2 is $1/2$, since he could have opened either door 2 or door 3. The car is obviously not behind door 2, otherwise Monty would not have been able to open that door, so the likelihood for door 2 is 0. If the car is behind door 3, the likelihood of Monty opening door 2 is 1, because that is Monty's only choice.

Example: Monty Hall Problem

Door	Prior $\mathbb{P}(\theta)$	Likelihood $\mathbb{P}(X\|\theta)$	$\mathbb{P}(\theta)\mathbb{P}(X\|\theta)$	Posterior $\mathbb{P}(\theta\|X)$
1	$1/3$	$1/2$	$1/6$	$1/3$
2	$1/3$	0	0	0
3	$1/3$	1	$1/3$	$2/3$

The posterior is proportional to the prior times the likelihood. All posterior probabilities have to add up to 1, which means the probability that the car is behind door 1 is $1/3$, while the probability that the car is behind door 2 is $2/3$. It makes more sense to switch doors rather than stay with your original choice, as unintuitive as that sounds.

Example: Election

An upcoming election is between the Blue Party and the Green Party. A recent poll gives the following results for the candidates:

Blue Party: 279

Green Party: 230

Results from previous elections (our *prior*) suggests that the Blue Party candidate would get 49.1% of the vote, with standard deviation 2.2%. What would Bayesian Inference suggest is the posterior distribution for this election?

Image Source

Example: Election

Let's say θ is the probability of Blue Party success. Then the likelihood of seeing x votes out of the total of 509 of our survey respondents, based on the survey data, is

$$\mathbb{P}(x|\theta) \propto \theta^{279}(1-\theta)^{509-279}$$

This is modeling the likelihood as a binomial distribution. For our prior, we can use a Beta distribution, which pairs well with the binomial distribution in Bayesian inference.

Example: Election

Our prior will be based on previous elections, with mean 49.1% and standard deviation 2.2%. A Beta distribution has the following density:

$$\mathbb{P}(\theta; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}$$

with mean and variance

$$\mathbb{E}(\theta; \alpha, \beta) = \frac{\alpha}{\alpha + \beta} = 0.491$$

$$\mathbb{V}\text{ar}(\theta; \alpha, \beta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = 0.22^2$$

This yields

$$\alpha = 253.04 \qquad \beta = 262.32$$

Example: Election

The posterior probability density is proportional to the product of the prior and the likelihood. Even though the previous elections don't give high hopes for the Blue Party candidate, the survey gives the Blue Party candidate a higher likelihood of success. The gray area underneath the posterior curve gives us the probability that the Blue Party candidate will win.

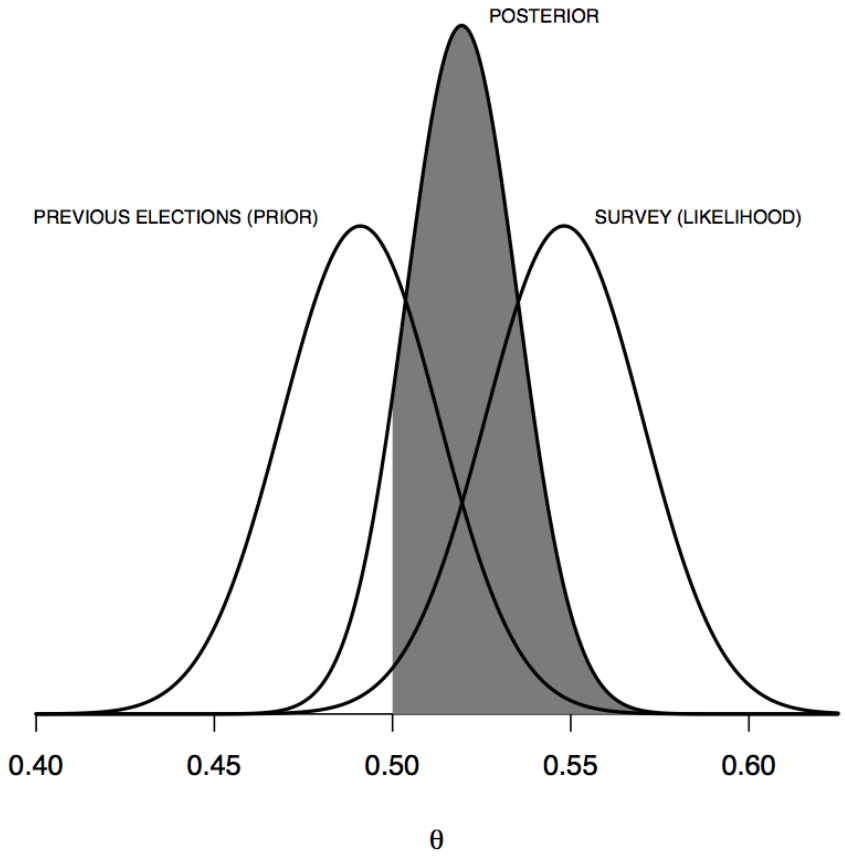


Image Source

Frequentist vs. Bayesian

- Frequentist approaches were developed first, since they were easier to solve.
- Bayesian approaches are more computationally intensive. They have only recently become used in common practice.
- Both approaches give similar results.

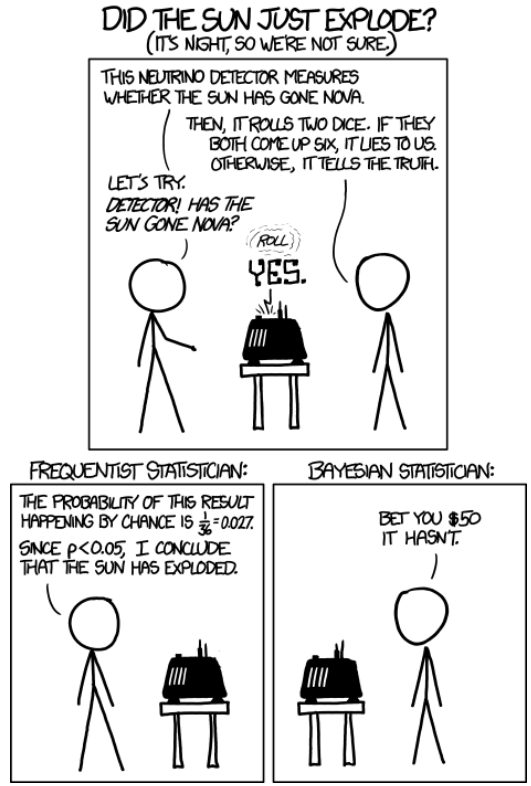


Image Source