Statistical Inference

A comparative review of Frequentist, Bayesian, and Visual methods

Tyler Wiederich

March 1, 2025

Introduction

The quantification of uncertainty is one of the fundamental concepts behind statistical methodology (Acree 2021; Erich Leo Lehmann 2008). This is useful in cases where population parameters are unknown or unable to be reasonably measured. For example, some possible research questions that would require statistical methods include "What percentage of Minnesotan residents agree with the proposed legislation?," or "Does the placement of higher margin products on the middle shelf of a store aisle lead to more profits than placement on top and bottom shelves?" These types of questions can be investigated using statistical inference, although the approach differs by the specific method of inference.

Implicitly, many statistical methods revolve around the concept of a hypothesis test (Neyman and Pearson 1933; Acree 2021). These tests are designed to determine whether or not there is evidence to support a claim. Data is evaluated against a null hypothesis, which is typically a baseline assumption or a specific condition. If the condition is deemed unlikely to occur by chance, then there is evidence to support its complement, called the alternative hypothesis. The strength of evidence can be quantified through summaries such as p-values and Bayes factors (Held and Ott 2018), leading to decisions about what interpretation or decision to make regarding the results of the method.

Early statistical methodology is generally regarded as Frequentist inference (Neyman and Pearson 1933; Erich Leo Lehmann 2008). The idea behind Frequentist methods is that data is random variable, generated from some probability function using at least one unknown population parameter. Functions of the data yield test statistics that are used to evaluate a hypothesis test. In Frequentist testing, the null hypothesis is usually defined such that there is no effect, and the alternative hypothesis is that there is an effect. Since the data is considered random, statements are made in terms of expected results under repeated sampling.

A newer framework developed in the later 20th century using the idea that previously known information can help guide inference, which became known as Bayesian inference (Gelman

and Shalizi 2013; Erich Leo Lehmann 2008). The availability of previous information, or lack thereof, forms the basis of belief for what occurs in nature. Bayesian inference combines new data with prior knowledge to update the belief of population parameters. In this case, the population parameter(s) is considered random and distributed according to belief and observed data.

The use of visualizations is a valuable method for understanding the structure and patterns in data (John Wilder Tukey 1977), but they can also guide formal analyses (Loy and Hofmann 2015; Loy, Follett, and Hofmann 2016; Loy, Hofmann, and Cook 2017). Visual inference is where hypotheses can be tested by human perception instead of mathematical formulation. For example, Q-Q plots are a common diagnostic tool for checking normality assumptions (Loy, Follett, and Hofmann 2016). While methods like the Shapiro-Wilk test (Shapiro and Wilk 1965) exist for testing normality, visual inference can be applied when assumptions are violated or tests do not exist.

This paper is organized as follows. The foundations and uses of Frequentist, Bayesian, and visual inferences are explained, including how each method should be interpreted. Then each method is compared to the other methods. Finally, an example using a Binomial generalized linear model is fitted with Frequentist and Bayesian methods, and evaluated using visual inference.

Types of Inference

The role of the population parameter is a key differentiation between Frequentist and Bayesian methodologies (Pek and Van Zandt 2020). The former assumes the parameter to be a fixed value for its target population (Neyman 1977), whereas the latter assumes the parameter is a random variable drawn from a probability distribution (Gelman and Shalizi 2013). This changes the scope of inference for Frequentist and Bayesian inference. On the other hand, visual inference relies on human perception, which can be used on its own or in conjunction with Frequentist and Bayesian methods.

Frequentist Inference

The classical framework of statistics revolves around the idea that independent and repeated events follow a stable probability function with a fixed parameter (E. L. Lehmann 2008). As such, many instances of Frequentist inference include some mention of independent and identically distributed random variables. In practice, this assumption can be addressed through randomization (Neyman 1977) or checking model diagnostics (Loy, Follett, and Hofmann 2016).

During the early development of Frequentist statistics, philosophical debates between R. A. Fisher and J. Neyman led to the distinction between “inductive inference” and “inductive behavior” (E. L. Lehmann 2008). Inductive inferences, as preferred by Fisher, involves logical

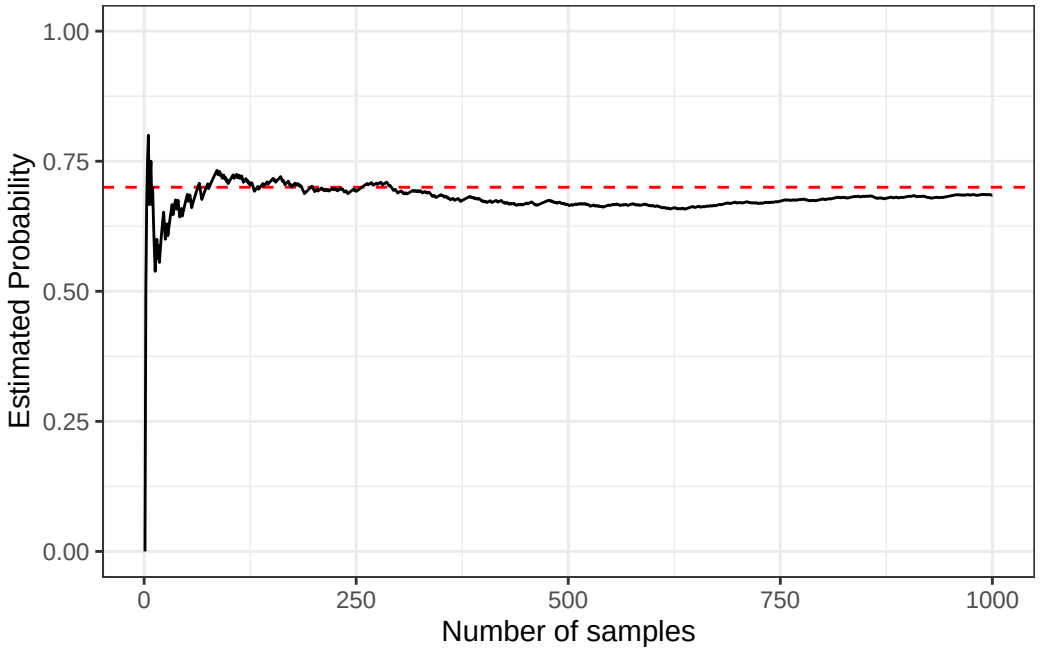
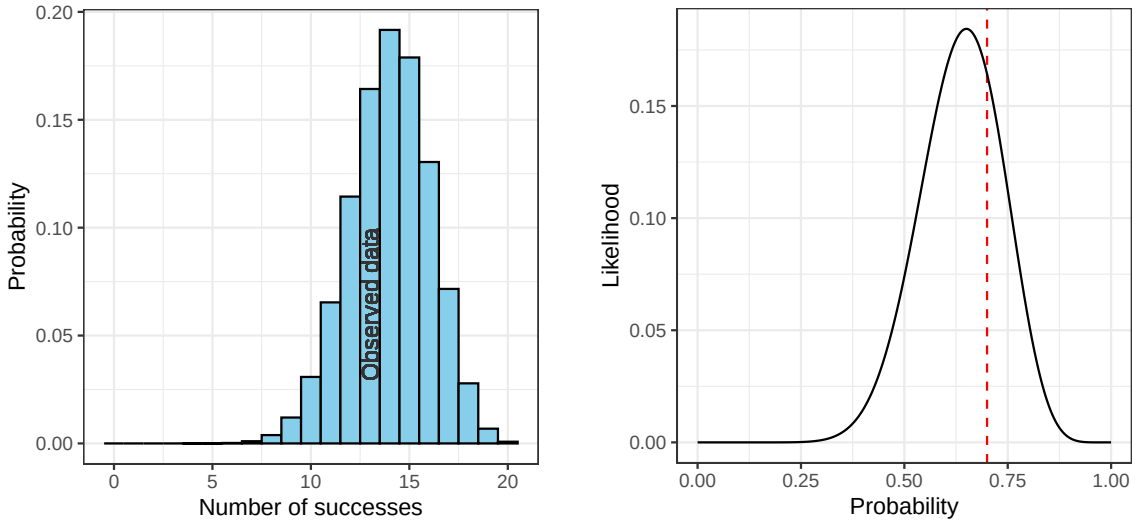


Figure 1: Estimated probability from a Bernoulli distribution with independent and identical samples. As the number of samples increases, the cumulative average estimate for the proportion of success stabilizes around the true probability of 0.7.

rational to develop models. On the other hand, Neyman's inductive behavior addresses the additional component of randomization through hypothesis testing. Both arguments start with the assumption that the random variable in question is independent and identically distributed. However, the two statisticians differed on the interpretations and approaches, with Fisher favoring intuition and Neyman favoring proofs. The methodology was similar between them, emphasizing specific approaches and scope of inference.

Fisher believed that probability was deductive in nature and that probabilistic events could be formulated mathematically to produce a maximum likelihood function (Fisher 1935). For example, the Binomial distribution is given by $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, where X is the random variable for the number of successes out of n trials given a probability p for x successes. The probability mass function comes from multiplying the number of combinations that a series of successes can take to the probabilities of individual successes and individual failures, a formulation of combinatorics. When taking the likelihood, the most plausible estimate of the probability can be found by establishing which value maximizes the likelihood with respect to all possible probabilities Figure 2.



(a) Binomial($n = 20, p = 0.7$)

(b) Likelihood of p for 13 successes out of 20 trials.

Figure 2: Binomial example of Fisher's inductive inference. A single observation from a Binomial distribution is drawn and observed to have 13 successes out of 20 trials, where the true probability of success is 0.7. The likelihood function is optimized at $p = 0.65$, and the red line in (b) shows the true probability of success in the population.

An inductive behavior approach was developed by Neyman and E. S. Pearson to address errors (Neyman and Pearson 1933). Neyman's approach emphasized that likelihoods were random and that further testing was needed so that inferences are not often wrong from true data generating functions. This development came to be the hypothesis test, where the level of error given certain assumptions about the data can be controlled. These tests include Type

1 and Type 2 errors, which are measures of uncertainty regarding how likely the hypothesis tests arrive at an incorrect conclusion.

When data is collected, a hypothesis test can be formulated Equation 1. The idea of a Frequentist hypothesis test is to assume that a realization of the data arose from an assumed data generating function, called the null hypothesis (Neyman and Pearson 1933). If the function of the data does not seem plausible under the null hypothesis, then the null hypothesis is rejected in favor of the alternative hypothesis, which encompasses the complement of the null hypothesis.

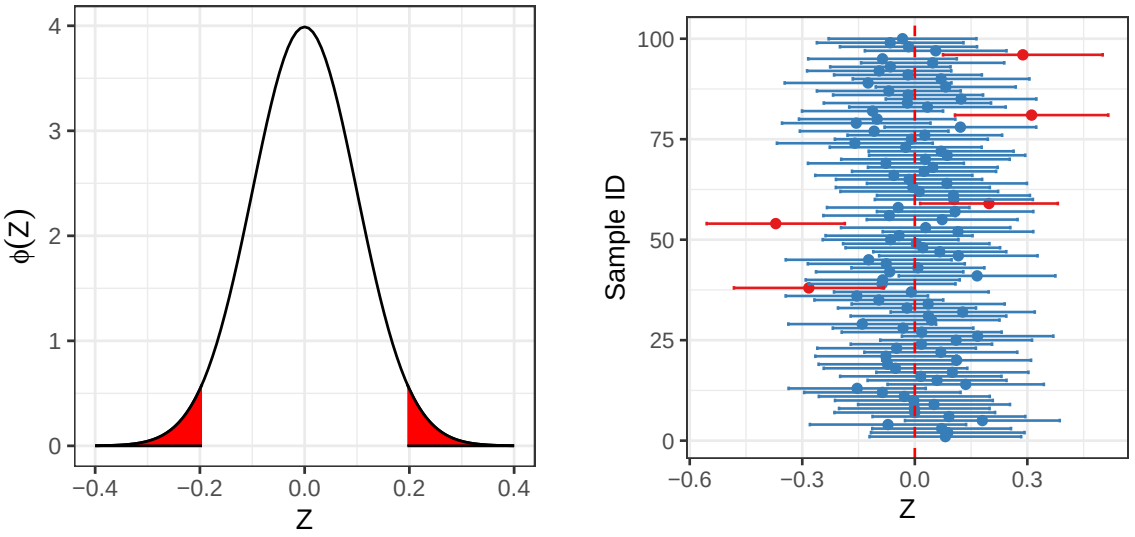
$$H_0 : \theta = \theta_0 \qquad H_1 : \theta = \theta_1 \tag{1}$$

One method to evaluate a hypothesis test is through a test statistic. Test statistics are formed so that their sampling distribution are known under the null hypothesis. For example, if the null hypothesis assume the random sample is from a normal distribution with mean μ and variance σ^2, then the sampling distribution of the sample mean, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$, can be rewritten as $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{N}}$, which follows a standard normal distribution. Given the data, a probability can be calculated for observing the test statistic under the null hypothesis, which is called a p-value (Figure 3a). If the p-value is small, then it was unlikely to see the observed data under the null hypothesis, giving evidence for the alternative hypothesis. However, a large p-value does not give evidence that the data was generated from the null hypothesis since this was an assumed condition.

Confidence intervals are another method of evaluating hypothesis tests (Neyman 1937). These intervals for two-sided tests are calculated using the form: Statistic $\pm\, c \times$ Standard Error, where c is defined to produce a C percent confidence level. Since the data is random, the interpretation of Frequentist confidence intervals is that C percent of confidence intervals calculated from repeated sampling of the population will contain the true parameter value C percent of the time (Figure 3b). Evaluating hypotheses with confidence intervals simply involve checking if the parameter value in the null hypothesis falls within the interval. If the value does not fall in the interval, then there is evidence to reject the null hypothesis.

Bayesian Inference

Bayesian inference typically begins with some knowledge about the parameter of interest, although uninformative prior information can be used (Gelman and Shalizi 2013; Held and Ott 2018). This knowledge is used to formulate the posterior distribution, $\pi(\theta|X)$, which is the normalized product of the data likelihood under the parameter, $f(X|\theta)$, and the prior distribution, $\pi(\theta)$ (Equation 2). The posterior distribution is the main target of Bayesian inference, allowing the parameter of interest to exist as a probability function generated from past and current knowledge.



(a) Probability of observing test statistic under null hypothesis.

(b) 95% confidence interval for multiple repeated samples.

Figure 3: Frequentist evaluation methods using test statistics and confidence intervals under a null hypothesis that uses a standard normal distribution with a sample size of 100. Panel (a) shows the area under the sampling distribution for where $Z = \frac{\bar{X}}{1/\sqrt{100}}$ would reject the null hypothesis. Panel (b) provides 95% confidence intervals for sample means across multiple samples.

$$\pi(\theta|X) = \frac{f(X|\theta)\pi(\theta)}{\int f(X|\theta)\pi(\theta)d\theta} \propto f(X|\theta)\pi(\theta) \tag{2}$$

Unlike Frequentist methodology, there are no p-values in Bayesian hypothesis testing, at least in the sense of the Frequentist definition for a p-value (Gelman and Shalizi 2013; Held and Ott 2018). Instead, the Bayes factor can be used as evidence for competing hypotheses. The Bayes factor is the ratio of the posterior odds and prior odds, where θ_0 denotes the parameter model proposed for one model, and θ_1 is for an alternative model (Equation 3). Bayes factors less than 0.1 are generally considered indicative of evidence against the model for θ_0 (Held and Ott 2018).

$$\text{Bayes Factor} = \frac{\pi(\theta_0|X)/\pi(\theta_1|X)}{\pi(\theta_0)/\pi(\theta_1)} \tag{3}$$

To illustrate a simple example of Bayesian inference, consider Figure 2 where the true data generating function is Binomial$(n = 20, p = 0.7)$. The observed number of successes was 13 out of 20. Using an uninformative prior, the posterior distribution becomes Equation 4, which is proportional to a Beta distribution with $\alpha = 13 + 1$ and $\beta = 20 - 13 + 1$.

$$\pi(p|X)) = \binom{20}{x} p^x (1-p)^{n-x} \times 1 \propto p^{(x+1)-1}(1-p)^{(n-x+1)-1} \tag{4}$$

Now suppose that it is known that previous data shows that p is around 0.7. A sensible choice for the prior distribution is a Beta distribution, which will result in a closed-form solution. This is called a conjugate prior (Raiffa and Schlaifer 2000).

$$\pi(p|X)) \propto p^x(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1} \propto Beta(x + \alpha + 1, n - x + \beta + 1) \tag{5}$$

Letting $\alpha = 9$ and $\beta = 3.857$, the prior distribution has a mean of 0.7 and a variance of 0.01515. The posterior distribution becomes Beta$(13 + 10, 20 - 13 + 4.857)$, which shifts the posterior closer to 0.7 than when using the uniform prior (Figure 4).

Similar to Frequentist methods, intervals of the parameters can be created for the posterior distribution, which are called credible intervals (Acree 2021). These intervals are calculated simply by determining limits for a desired probability from the posterior distribution. The result is that interpretations are directly about the parameter and not repeated sampling of the data. From the binomial example with a uniform prior, the two-sided 95% credible interval states that there is a 95% probability that the proportion of success is between 0.43 and 0.82.

Selection of an appropriate prior distribution is an important part of Bayesian methods (Pek and Van Zandt 2020; Gelman and Shalizi 2013; Strachan and Dijk 2003). Prior distributions are typically chosen so that the strength of belief, or lack thereof, is accounted for. Overly optimistic priors either favor certain results or have smaller variances around anticipated values.

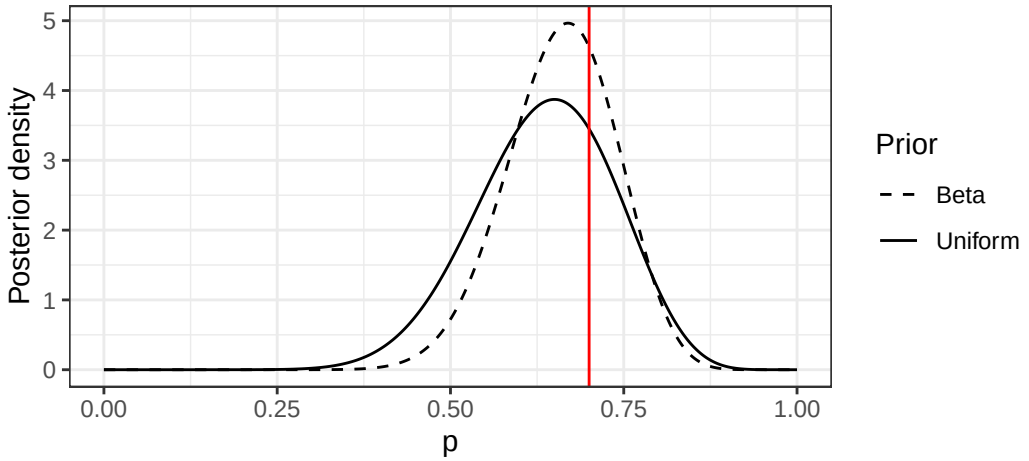


Figure 4: Posterior distributions for probability of success of a Binomial distriubtion using Beta and uniform prior distributions. Data was generated so that 13 out of 20 trials were successful with a true probability of 0.7, which is denoted by the red line.

When information is not available or there is a weak belief in prior information, a uniform prior or a prior with a large variance can be used. However, the choice of models can be evaluated against each other through the Bayes factor (Held and Ott 2018) or averaged together (Hoeting et al. 1999; Hinne et al. 2020)

Visual Inference

Visualizations are widely used in statistics, having proven useful in data exploration (John W. Tukey 1965; John Wilder Tukey 1977; Beniger and Robyn 1978) and checking model diagnostics (Loy, Follett, and Hofmann 2016). However, the use of visualizations for formal statistical inference is a relatively new development in the past two decades (Buja et al. 2009; Loy and Hofmann 2015; VanderPlas and Hofmann 2017). In contrast to Bayesian and Frequentist methodologies, visual inference is not a strictly mathematical formulation but rather the product of human perception.

The primary method of applying visual inference is through the lineup protocol (Buja et al. 2009). In the lineup protocol, viewers are presented with a series of graphs that display the target data and data generated according a null model. The viewers are then tasked with identifying the graph that is most different from the others. If viewers can identify the target graph, then there is evidence that the target graph is different than the graphs created under the null model. An example is provided in Figure 5.

Defining the target plot and null plots are essential steps in designing a lineup study. The target plot is chosen so that it reflects a specific condition, often involving real data. In contrast, the null plots are constructed to reflect conditions that the target graph do not satisfy, or

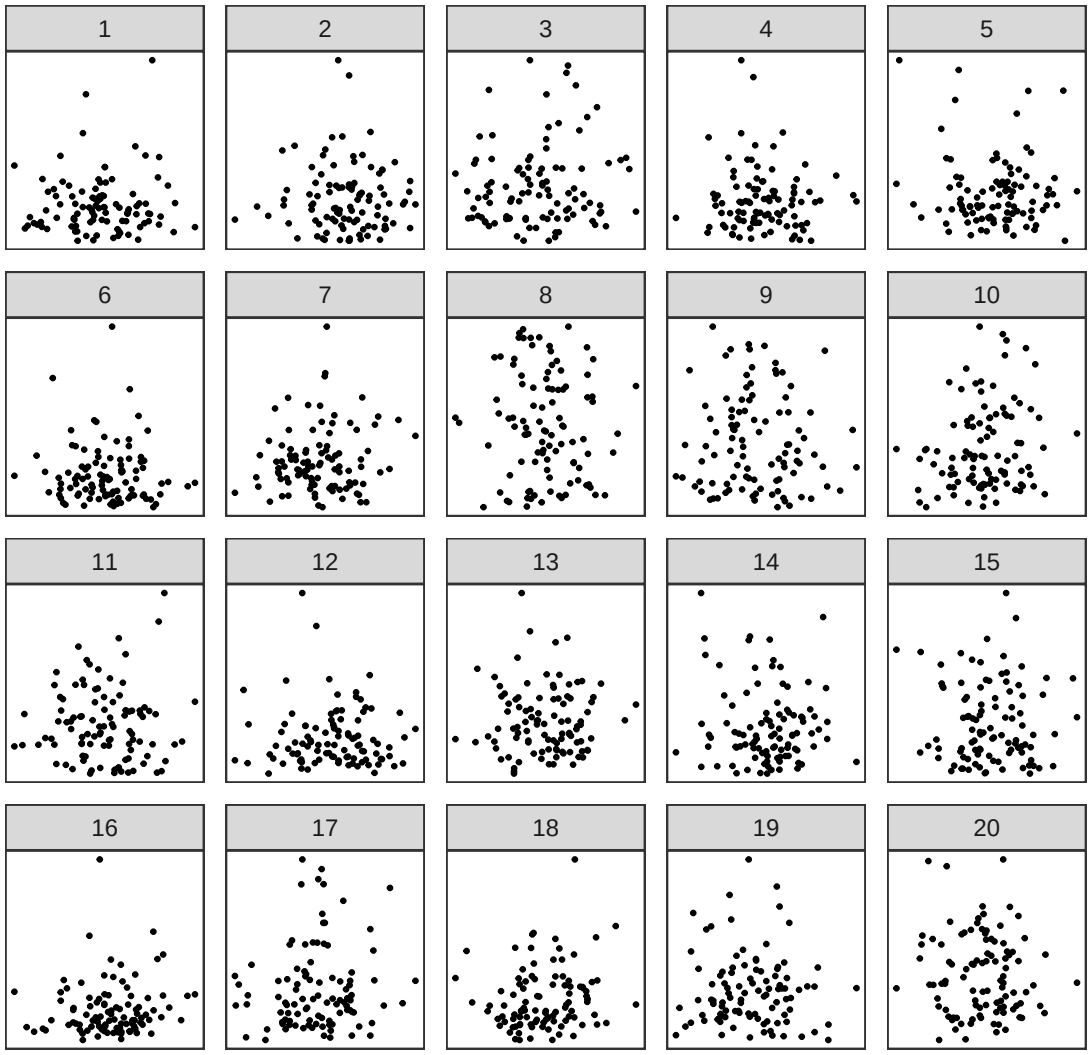
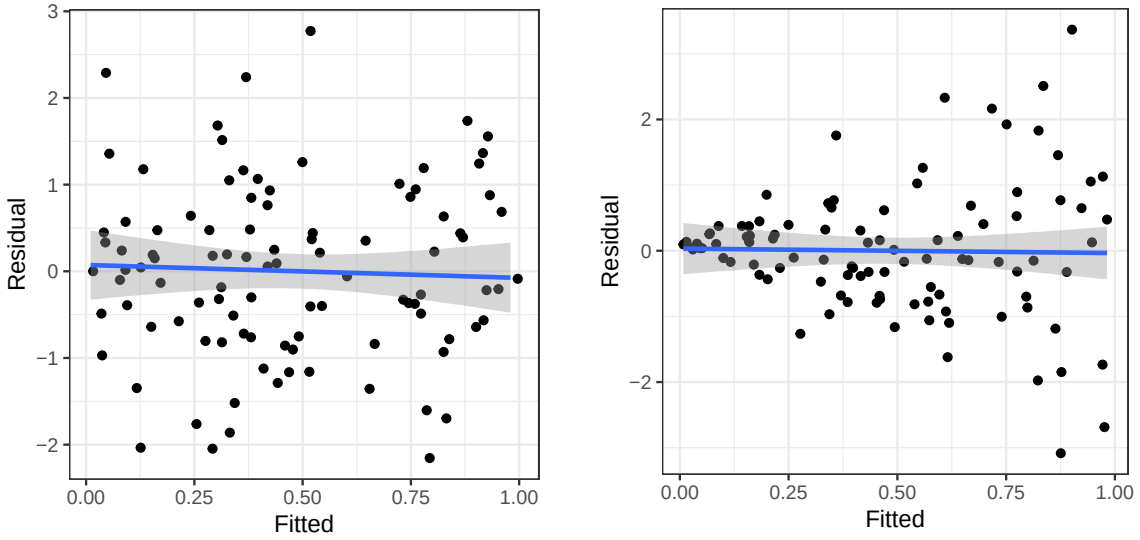


Figure 5: One trial of a visual inference lineup study. In this series of graphs, one dataset was produced differently than the other datasets. Can you figure out which graph it is? The answer can be found in the discussion section of this paper.

simulated data. For example, linear models often have the assumption that residuals are independently and identically distributed with mean zero and a constant variance (Majumder, Hofmann, and Cook 2013). In this case, testing the residual assumption would involve using a linear model fit to the true data as the target plot. Data can be simulated according the estimated linear model and fitted to the same model, which would result in the null plots. If viewers can identify the residual plot of the true dataset, then visual inference would indicate that the independent and identically distributed assumption of the residuals is violated. See Figure 6 for an example.



(a) Independent residuals

(b) Dependent residuals

Figure 6: A comparison of two residual plots for the model $y = X\beta + \epsilon$, where $\epsilon \sim iidN(0, \sigma^2)$. In panel (a), data is generated so that the residual assumptions holds. Panel (b) violates the residual assumptions by including non-constant variance.

The lineup protocol tests the null hypothesis that the target graph cannot be identified among graphs generated with a null model against the alternative hypothesis that the target graph can be identified among graphs generated with a null model. Assuming that a total of $m-1$ null plots, 1 target plot, and n participants, a Binomial distribution can be used to test the hypothesis that the target plot is chosen correctly where π is the probability of choosing the target plot :

$$H_0 : \pi = 1/m \quad \text{vs.} \quad H_A : \pi \neq 1/m$$

However, this hypothesis test assumes that plots are chosen at random by viewers. In reality, viewers are searching for particular features in the plots to use as justification for their selection (Buja et al. 2009). A different method of evaluation is presented by VanderPlas et al. (2021),

where plot selection dependencies can be modeled with a Dirichlet-multinomial distribution. This accounts for various features in the plots and the strength of their signals.

Comparison of Inferences

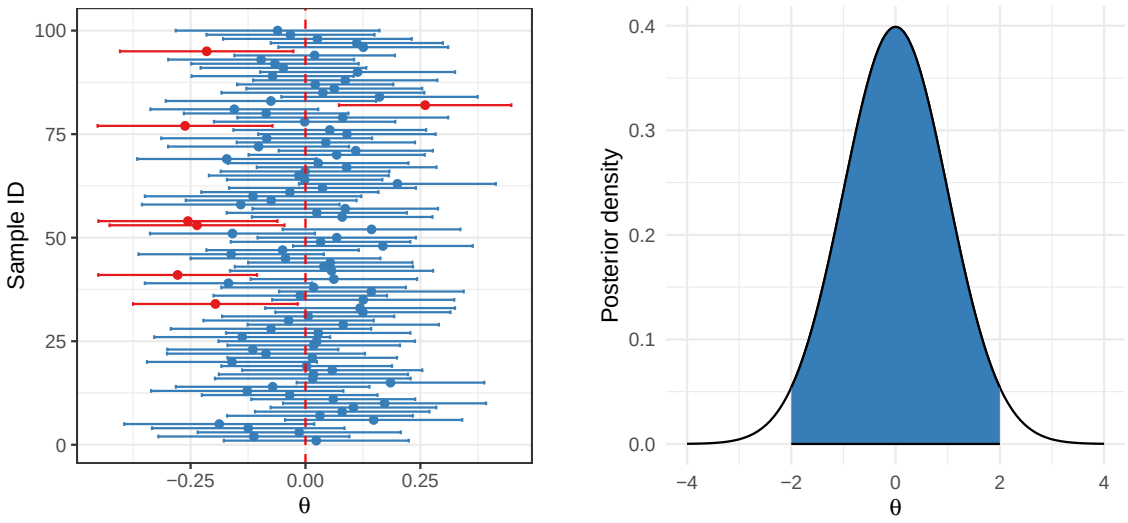
Frequentist vs. Bayesian

The differences between Frequentist and Bayesian inference mainly arise from how population parameters are defined (Pek and Van Zandt 2020; Acree 2021). Frequentist inference assumes that a parameter for a well-defined population at a given time is fixed and unknown, where a function of the data is the random variable. In some cases, this is a reasonable assumption, such as estimating the average weight of smallmouth bass in Lake Pepin, Minnesota (www.dnr.state.mn.us/lakefind). Here, there is a true average weight for all smallmouth bass in the lake, but the value can only be estimated. In contrast, Bayesian inference interprets the population parameter as a random variable to quantify uncertainty. For the case of smallmouth bass in Lake Pepin, prior information about average weights obtained from historical records can be combined with current data to formulate the posterior distribution to update beliefs on average weights for smallmouth bass in the lake.

Treating the population parameter as fixed or random affects the types of statements that Frequentist and Bayesian inferences are allowed to make about the population. Under Frequentist methodology, functions of the data are considered random, and thus statements about the population parameter pertain to the sampling distribution of the data. In essence, this means that Frequentist statements answer the question "how often would this outcome occur under repeated sampling?" For Bayesian methodology, the posterior distribution allows statements to be made directly about the population parameters.

A useful tool for both Frequentist and Bayesian inferences is a range of plausible values for the population parameter (Pek and Van Zandt 2020). For Freqentists, this range of values is the confidence interval (Neyman 1937). The interpretations of confidence intervals are about how often repeated sampling of the population would result in the interval covering the true population parameter. These statements become "with X% confidence, the true population parameter is between *lower limit* and *upper limit*." The Bayesian use of the posterior distribution introduces the credible interval. With the credible interval, statements can be made directly about the population parameter, such as "there is X% probability that the population parameter is between *lower limit* and *upper limit*." Figure 7 shows an example of how the intervals differ between Frequentists and Bayesians.

Due to the differences in how the population parameter is defined, the scope of inference drastically changes between Frequentist and Bayesian methods. Frequentists establish a hypothesis test based on a test statistic from the sampling distribution. Probabilities of the test statistic under the null hypothesis give evidence to either reject or fail to reject the null hypothesis. For Bayesian methods, hypothesis testing is redundant since the posterior contains all of the



(a) Confidence interval

(b) Credible interval

Figure 7: Comparison of interprations between confidence and credible intervals. Confidence intervals rely on repeated sampling of the data, whereas credible intervals use probabilities of the posterior distribution.

necessary information about the population parameter, including the estimate and credible interval. The focus of Bayesian inference is about obtaining well-fitted models from reasonably chosen prior distributions.

Frequentist vs. Visual

The formulation of formal visual inference is rooted in the methodology of Frequentist inference (Buja et al. 2009; Hofmann et al. 2012; VanderPlas et al. 2021). Both methodologies assume a hypothesis test where evidence is provided through data as to reject or fail to reject the null hypothesis. Whereas Frequentist inference establishes evidence through a test statistic of the data, visual inference establishes evidence through the detection of human perception. This is an important distinction, since the null hypothesis of Frequentist testing using a probability function and visual tests use $m-1$ realizations of the null hypothesis. However, Majumder, Hofmann, and Cook (2013) showed that lineup studies can be comparable to traditional statistical tests.

Bayesian vs. Visual

Bayesian and visual inferences are perhaps the most different from each other. Bayesians rely on prior information to construct a model for the population parameter, but visual inference

is mostly concerned with detecting differences between null plots and the target plot. Of course, some testing procedures in visual inference rely on previous knowledge that not all null plots are created equally, and thus different signal strengths within the null plots can be incorporated in statistical significance calculations (VanderPlas et al. 2021). Although this approach is not directly a Bayesian inference, but can lead into a Bayesian approach by using the relationship between the multinomial and Dirichlet distributions.

Visual Inference as a Diagnostic Tool

While there are many differences between Frequentist and Bayesian testing, visual inference tends to be a useful tool for model diagnostic checks (Majumder, Hofmann, and Cook 2013; Loy and Hofmann 2015; Loy, Hofmann, and Cook 2017). To oversimplify the field of statistics, the main objective is to produce models that are not so far off from the truth so that generalizations can be made about findings in nature. This means that the process of creating null plots from a model and determining if the actual data can be distinguished is a useful tool to validate the fit of a statistical model.

The use of visual inference diagnostic checks for Frequentist methods involve generating data from a fitted model. After data is generated, one possible diagnostic check is to place the true data in lineup study with the generated data. If the true data is indistinguishable from the simulated data, then it is reasonable to assume that the model fits well. Another option is to use the residual plots for models refitted to the simulated data compare them to the actual residuals. The residual plots can be useful when the usual residual assumption of normality with mean 0 and constant variance is not modeled, such as for repeated measures experiments.

For Bayesian analyses, the process of incorporating visual inference is slightly different than for Frequentist analyses. First, parameter values need to be drawn from the posterior distribution(s). Then data can be simulated and used in a lineup study with the observed data as the target plot. This process could also be extended to comparing two posterior distributions that use different priors, similar to how the Bayes factor measures evidence for favoring one model over another.

Example

In the National Hockey League (NHL), games played at home are filled with energy and excitement for the home team. This is called having home-ice advantage. Consider the Minnesota Wild during their 2024-2025 season. Suppose that the head coach wishes to know if the Wild are playing with an advantage playing on home-ice, meaning that they win more games at home than they do for away games. Data for location and results is shown in Table 1.

Table 1: Data from the 2023-24 and 2024-25 seasons for the Minnesota Wild. L represents losses and W represents wins.

season	location	W	L
2023-24	away	21	12
2023-24	home	13	14
2024-25	away	19	22
2024-25	home	20	21

The first step of any analysis should be data exploration (Figure 8). The Minnesota Wild won 0.5 percent of their home games and 0.5 percent of their away games as of February 28th, 2025. Immediately noticeable is that the Wild do not appear to play with any significant home-ice advantage. Data was also collected for the Wild's shots on goal (SOG), penalties in minutes (PIM), power play goals (PPG), and power kill opportunities (PPO).

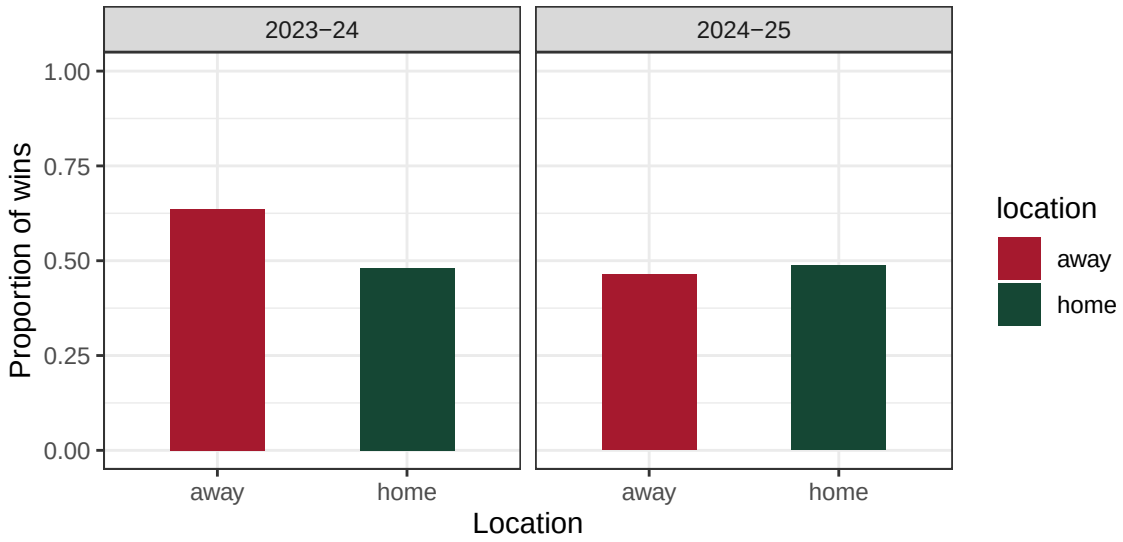


Figure 8: Observed proportion of wins for the Minnesota Wild by game location.

The Frequentist analysis begins with the assumption that games are independently and identically distributed with a Binomial distribution. However, this assumption is not reasonable since the team's performance can change throughout the season. For example, Ryan Hartman received a major penalty on the February 1st, 2025 game against the Ottawa Senators. The penalty was severe enough for Hartman to receive a 10-game suspension and thus changing a part of the team's dynamic for those games.

Equation 6 is the model fit to the data for the 2024-25 season, where β_0 is the intercept term, β_{1i} is the effect for location i, β_3 - β_5 are the effects of their corresponding variables, and π_i is the probability of winning for location i.

$$\eta_{ijklm} = \log(\frac{\pi_{ijklm}}{(1-\pi_{ijklm})}) = \beta_0 + \beta_{1i} + \beta_2 \cdot \text{sog}_j + \beta_3 \cdot \text{pim}_k + \beta_4 \cdot \text{ppg}_l + \beta_5 \cdot \text{ppo}_m \quad (6)$$

Table 2: ANOVA Table for Frequentist GLM.

	LR Chisq	Df	Pr(>Chisq)
location	0.05507	1	0.8145
sog	0.3094	1	0.5781
pim	0.6811	1	0.4092
ppg	4.733	1	0.02959
ppo	0.1234	1	0.7253

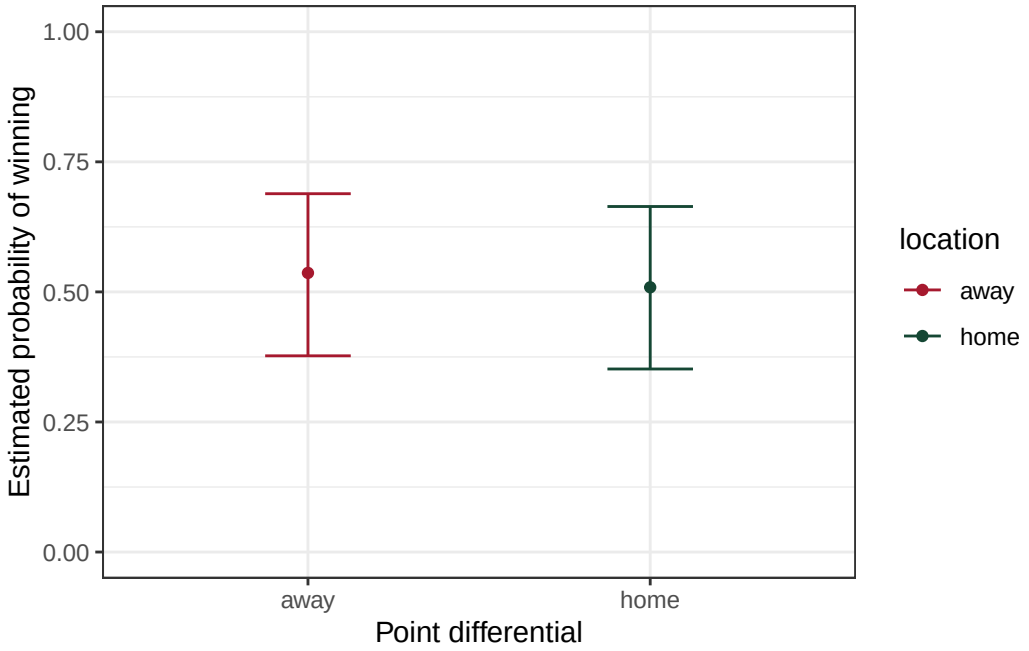


Figure 9: Estimated probabilities of winning for the Frequentist GLM model.

Fitting the Frequentist generalized linear model (GLM) reveals that power play goals (ppg) is the only significant term (Table 2), which is a reasonable result given that power plays are good scoring opportunities. The only conclusion from the this analysis is that there no evidence that the Minnesota Wild have a competitive advantage when playing at home.

To evaluate the fit of the Frequentist GLM, one solution is to use a lineup study. Here, data can be simulated such that wins and losses are randomly chosen from a Binomial distribution with a probability of success using the coefficients obtained from the fitted model. Figure 10 shows

a lineup of ten datasets produced by simulating wins and losses. In this lineup, probabilities are plotted along the power play goals for both away and home games. It is difficult to identify the true dataset, which means that the Frequentist GLM appears to fit well for the location of the games.

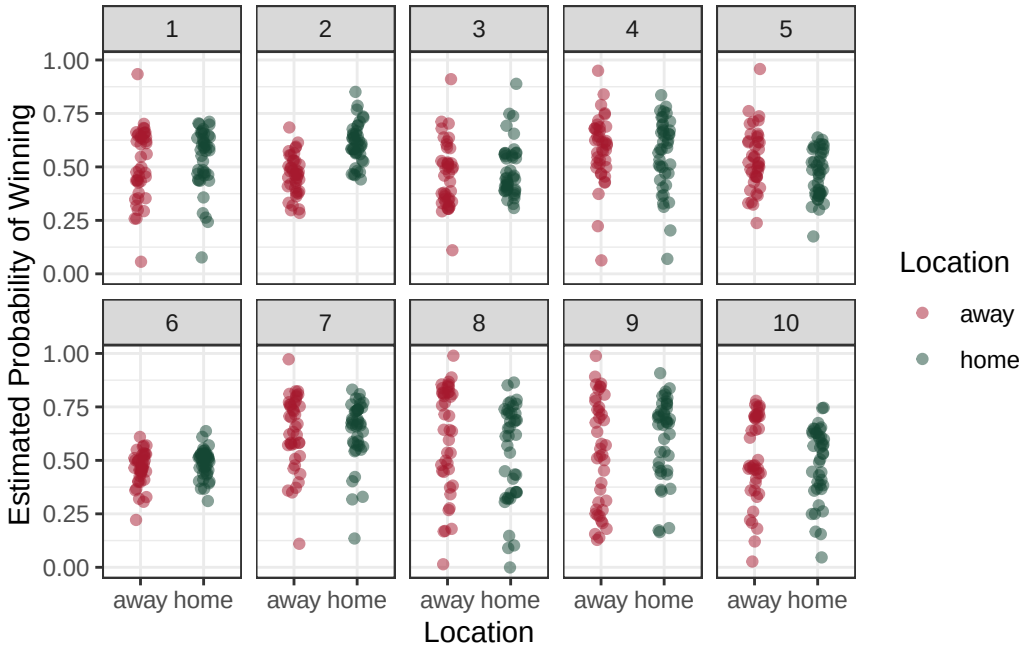


Figure 10: Lineup study using simulated wins and losses for the Minnesota Wild. Nine of the plots show data simulated from the fitted GLM, and one plot uses the actual data. The target plot is plot number (10-(4+3)).

The Bayesian approach uses the `brms` package in R to fit a Bayesian GLM to the data using the Minnesota Wild win-loss record for the 2023-2024 season. Two models are compared, one with uninformative (flat) priors on each model coefficient, and another that uses normal priors based on a Frequentist GLM fit for the 2023-2024 season (Table 3). The estimates and standard errors are overly reliant that the results from the previous season carry over to the current season, which is somewhat reasonable since the composition of the Wild is mostly the same.

Table 3: Bayesian GLM coefficients for normal priors, generated from coefficents of a Frequentist GLM for the 2023-24 season.

	Estimate	Std. Error
(Intercept)	-0.2001	1.498
location1	-0.485	0.3003
sog	-0.0716	0.05113

Table 3: Bayesian GLM coefficients for normal priors, generated from coefficents of a Frequentist GLM for the 2023-24 season.

	Estimate	Std. Error
pim	0.1077	0.06424
ppg	-0.7984	0.5134
ppo	0.5671	0.2861

The Bayes GLM coefficients for both models are similar to the Frequentist approach (Table 4). This result is expected since the uninformative model puts more weight on the data and the informative model has information from the previous season. Additionally, the coefficients obtained using informative priors have have smaller standard errors than those obtained for the uninformative priors model and the Frequentist GLM, and the uninformative priors model has slightly larger standard errors than the Frequentist GLM.

Table 4: Coefficients and standard errors of Frequentist and Bayesian methods for GLMs fit to the Minnesota Wild data.

Term	Frequentist GLM	Bayes GLM (Uniform Priors)	Bayes GLM (Normal Priors)
(Intercept)	0.86 (1.32)	0.83 (1.4)	0.98 (1.13)
location1	0.06 (0.24)	0.05 (0.25)	-0.17 (0.19)
sog	-0.02 (0.04)	-0.02 (0.04)	-0.05 (0.03)
pim	0.02 (0.03)	0.03 (0.03)	0.04 (0.03)
ppg	-0.71 (0.36)	-0.81 (0.38)	-0.93 (0.29)
ppo	0.06 (0.18)	0.06 (0.19)	0.21 (0.15)

Comparing the two Bayesian models and using the model with informative priors as the "null hypothesis", the Bayes Factor is computed to be approximately 0.4. Although the standard errors of the parameters for the two models slightly differ, there is not much evidence to prefer the informative priors over the uninformative priors. The conclusions obtained from both Bayesian models are the same, where only power play goals is significant according to the 95% credible intervals (Figure 11).

Similar to the Frequentist case, visual inference can be used to examine model diagnostics for the Bayesian models. Here, it would be worth using comparing the uninformative and informative models to evaluate if either model is sufficient. Additionally, plotting predicted probabilities along each model term would be good candidates to use as target plots. Data can be simulated from the probabilities, refit with with model, and placed into a lineup study with the target plot.

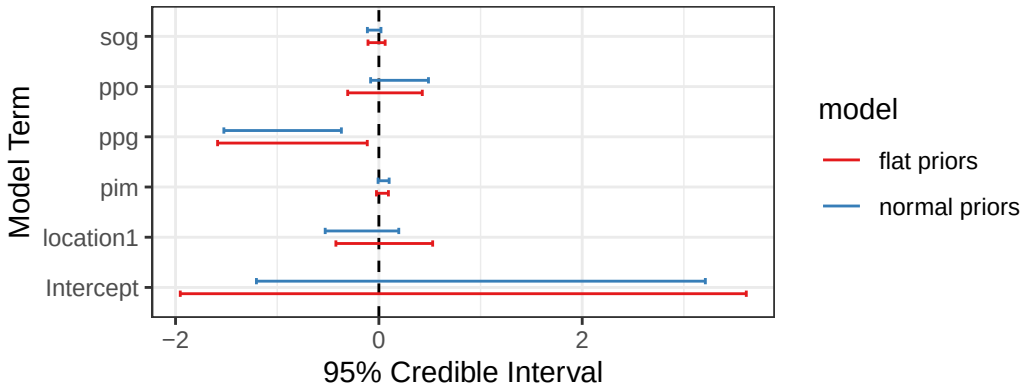
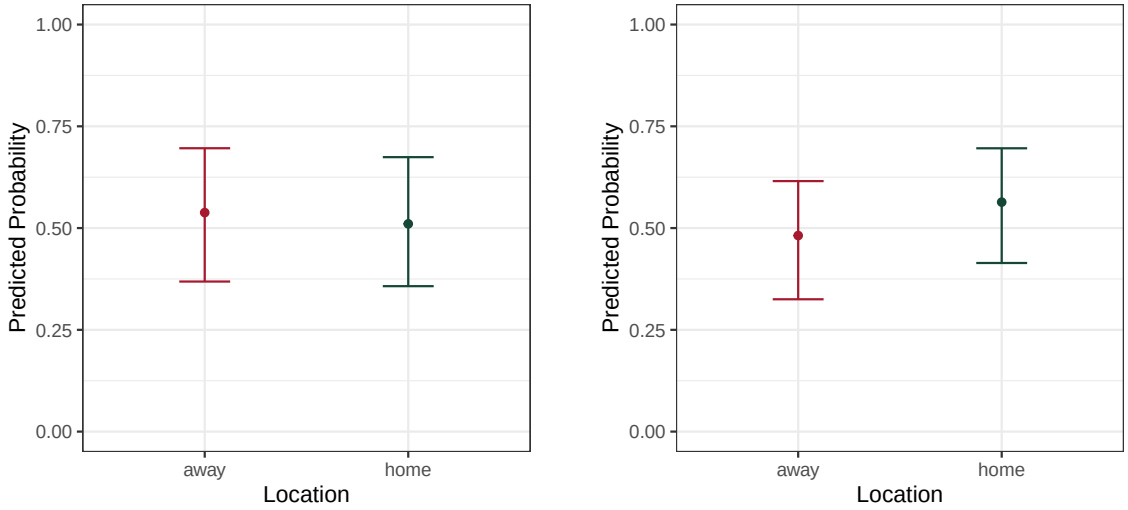


Figure 11: 95% credible intervals for both Bayesian models. The model with normal priors has smaller intervals, but the conclusions are the same for both models.



(a) Flat priors

(b) Normal priors

Figure 12: Predicted probabilites under two Bayesian models. The model using normal priors has smaller credible intervals than the model using flat priors.

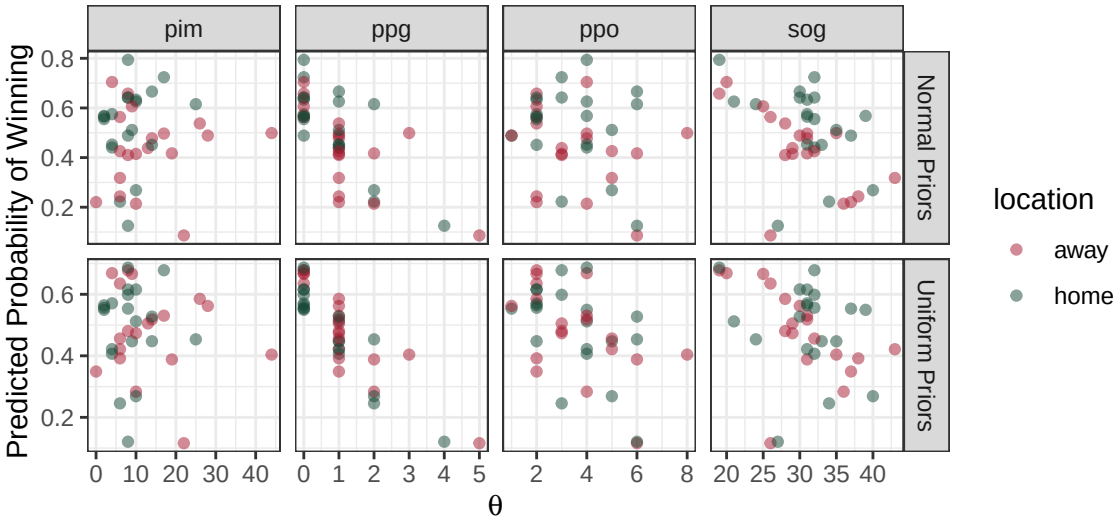


Figure 13: Potential target plots for visual inference

For the Minnesota Wild, the probability of winning games appears to be a coin flip. Only power play goals had a significant effect, as indicated in the Frequentist and Bayesian GLMs. However, it was interesting to see that the Wild do not seem to have an advantage while playing games at the Excel Energy Center.

Discussion

There is no single way to approach statistical inference. The commonality of statistical methods is that uncertainty is quantified through a random variable, but defining the random variable is where the methods diverge in computation and interpretations. In this paper, the methodology of Frequentist, Bayesian, and visual inferences were discussed and compared, included when each type of inference should be used.

Frequentists rely on data as it is provided. This is a reasonable situation when conducting novel experiments or collecting data from new sources. Statements of uncertainty pertain to what happens under repeated sampling conditions, which is useful when wanting to know what conditions to expect for the next instance of similar data collection. Results are more easily calculated, having test statistics and confidence intervals based on sampling distributions of statistics from the parameter.

In contrast to Frequentists, Bayesian inference is about updating beliefs given prior information. The behavior of the population parameter is treated as random and updated through a new instances of data collection. The goal is to estimate the posterior distribution so that statements can be made about the parameter(s) of interest. Other than model checking, this is where Bayesian analyses typically end since the goal is to make make inferences on the

parameter(s). Bayesian methods are typically computationally demanding due to the complex posterior calculations

Lastly, visual inference is unique in that it can function on its own or in addition to Frequentist and Bayesian methods (Loy, Hofmann, and Cook 2017; Gelman and Shalizi 2013). The main goal of lineup studies is to determine if a target plot can be identified from a set of null plots. If it can be distinguished, then there is evidence that humans can pick up on particular features of the target plot. From Figure 5, null plots were modeled with a gamma distribution for the y-axis variable and a normal distribution for the x-axis variable. The target plot was plot number 8, where the y-axis variable used a uniform distribution instead of a gamma distribution. Additionally, these lineup tests can be adapted for model fitting in Frequentist and Bayesian analyses, which can help to verify that models are reasonable for the collected data.

Frequentist, Bayesian, and visual inference each have their own place in estimating phenomena of the natural world. After data collection, the type of analysis mostly comes down to personal preference and adequately fitting reasonable models. Frequentists use data as the foundation their analyses and Bayesians use data and prior beliefs. Visual inference exists as a method that can be used in combination with analyses, but also for inferences regarding data visualization. Picking the right type of inference is useful for providing data-driven results.

References

Acree, Michael C. 2021. *The Myth of Statistical Inference*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-73257-8.

Beniger, James R., and Dorothy L. Robyn. 1978. "Quantitative Graphics in Statistics: A Brief History." *The American Statistician* 32 (1): 1–11. https://doi.org/10.2307/2683467.

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F. Swayne, and Hadley Wlckham. 2009. "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 367 (1906): 4361–83. https://www.jstor.org/stable/40485732.

Fisher, R. A. 1935. "The Logic of Inductive Inference." *Journal of the Royal Statistical Society* 98 (1): 39–82. https://doi.org/10.2307/2342435.

Gelman, Andrew, and Cosma Rohilla Shalizi. 2013. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology* 66 (1): 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x.

Held, Leonhard, and Manuela Ott. 2018. "On p -Values and Bayes Factors." *Annual Review of Statistics and Its Application* 5 (1): 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307.

Hinne, Max, Quentin F. Gronau, Don Van Den Bergh, and Eric-Jan Wagenmakers. 2020. "A Conceptual Introduction to Bayesian Model Averaging." *Advances in Methods and Practices in Psychological Science* 3 (2): 200–215. https://doi.org/10.1177/2515245919898657.

Hoeting, Jennifer A, David Madigan, Adrian E Raftery, and Chris T Volinsky. 1999. "Bayesian Model Averaging: A Tutorial."

Hofmann, Heike, Lendie Follett, Mahbubul Majumder, and Dianne Cook. 2012. "Graphical Tests for Power Comparison of Competing Designs." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2441–48. https://doi.org/10.1109/TVCG.2012.230.

Lehmann, E. L. 2008. "Foundations I: The Frequentist Approach." In, 160–77. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-71597-1_10.

Lehmann, Erich Leo. 2008. *Reminiscences of a Statistician: The Company I Kept*. New York: Springer.

Loy, Adam, Lendie Follett, and Heike Hofmann. 2016. "Variations of q-q Plots: The Power of Our Eyes!" *The American Statistician* 70 (2): 202–14. https://www.jstor.org/stable/45118307.

Loy, Adam, and Heike Hofmann. 2015. "Are You Normal? The Problem of Confounded Residual Structures in Hierarchical Linear Models." *Journal of Computational and Graphical Statistics* 24 (4): 1191–1209. https://www.jstor.org/stable/24737225.

Loy, Adam, Heike Hofmann, and Dianne Cook. 2017. "Model Choice and Diagnostics for Linear Mixed-Effects Models Using Statistics on Street Corners." *Journal of Computational and Graphical Statistics* 26 (3): 478–92. https://doi.org/10.1080/10618600.2017.1330207.

Majumder, Mahbubul, Heike Hofmann, and Dianne Cook. 2013. "Validation of Visual Statistical Inference, Applied to Linear Models." *Journal of the American Statistical Association* 108 (503): 942–56. https://www.jstor.org/stable/24246876.

Neyman, J. 1937. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236 (767): 333–80. https://doi.org/10.1098/rsta.1937.0005.

———. 1977. "Frequentist Probability and Frequentist Statistics." *Synthese* 36 (1): 97–131. https://doi.org/10.1007/BF00485695.

Neyman, J., and E. S. Pearson. 1933. "On the Problems of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London* 231A: 289–338. https://doi.org/10.1098/rsta.1933.0009.

Pek, Jolynn, and Trisha Van Zandt. 2020. "Frequentist and Bayesian Approaches to Data Analysis: Evaluation and Estimation." *Psychology Learning & Teaching* 19 (1): 21–35. https://doi.org/10.1177/1475725719874542.

Raiffa, Howard, and Robert Schlaifer. 2000. *Applied statistical decision theory*. Wiley classics library ed. Wiley classics library. New York, NY: Wiley.

Shapiro, S. S., and M. B. Wilk. 1965. "An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52 (3-4): 591–611. https://doi.org/10.1093/biomet/52.3-4.591.

Strachan, Rodney W., and Herman K. van Dijk. 2003. "Bayesian Model Selection with an Uninformative Prior." *Oxford Bulletin of Economics and Statistics* 65 (s1): 863–76. https://doi.org/10.1046/j.0305-9049.2003.00095.x.

Tukey, John W. 1965. "The Technical Tools of Statistics." *The American Statistician* 19 (2): 23–28. https://doi.org/10.2307/2682374.

Tukey, John Wilder. 1977. *Exploratory Data Analysis*. Addison-Wesley Series in Behavioral

Science. Reading, Mass: Addison-Wesley Pub. Co.

VanderPlas, Susan, and Heike Hofmann. 2017. "Clusters Beat Trend!? Testing Feature Hierarchy in Statistical Graphics." *Journal of Computational and Graphical Statistics* 26 (2): 231–42. https://www.jstor.org/stable/44861949.

VanderPlas, Susan, Christian Röttger, Dianne Cook, and Heike Hofmann. 2021. "Statistical Significance Calculations for Scenarios in Visual Inference." *Stat* 10 (1): e337. https://doi.org/10.1002/sta4.337.