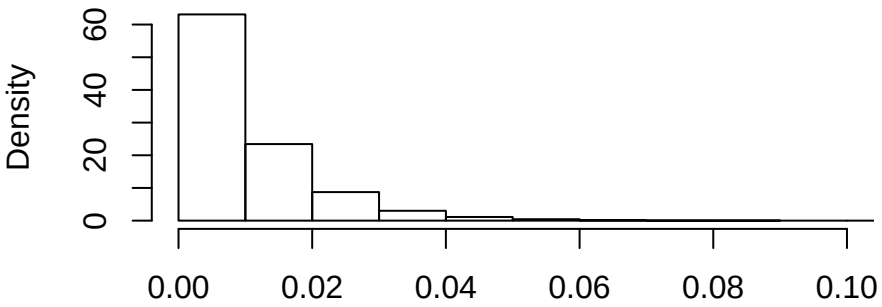
Lecture 1 - Key

Yellowstone Experiment

Assume you were hired by the National Park Service to estimate the probability of tourists petting wild animals or walking on restricted areas. What do you believe is the true probability of tourists petting animals or walking on restricted areas? (sketch this out)



.

Most likely non-zero, but with most of the mass of the distribution concentrated near zero.

Suppose, you are given results from eleven tourists, none of which of which test negative. Calculate the maximum likelihood estimator of p the probability of tourists disobeying the rules.

Let y_i be 1 if test i is positive and zero otherwise, then $y = \sum_i y_i$.

$$\begin{aligned} y &\sim Bin(11, p) \\ \mathcal{L}(p|y) &\propto p^y(1-p)^{n-y} \\ \log \mathcal{L}(p|y) &\propto y\log(p) + (n-y)\log(1-p) \\ \frac{\delta \log \mathcal{L}(p|y)}{\delta p} &\propto y\log p + (n-y)\log(1-p) \text{ set} = 0 \\ \hat{p}_{MLE} &= \frac{y}{n} \end{aligned}$$

Given that no "violators" were found in the testing, how does this estimator match up with your intuition?

Likely close but not exactly the same. Furthermore consider the standard confidence interval for proportions (using CLT) is $\hat{p} \pm z\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$ which in this case is a point mass at 0.

There are variations on this calculation such as $\hat{p} = \frac{y+1}{n+2}$, which can be considered procedures with a Bayesian flavor.

Now we will talk about the mechanics of Bayesian statistics and revisit the Yellowstone example.

- **Sampling Model:** The sampling model $p(y|\theta)$, where θ is a parameter that describes the belief that y would be the outcome of the study, given θ was known.

  Ex. *Binomial Model.* $p(y|p) = \binom{n}{y} p^y (1-p)^{n-y}$

- **Likelihood Function:** The likelihood function $\mathcal{L}(\theta|y)$ is proportional to the sampling model, but is a function of the parameters. When using a likelihood function, typically the normalizing constants are dropped.

  Ex. *Binomial Model.* $\mathcal{L}(p|y) \propto p^y (1-p)^{n-y}$

- **Prior Distribution:** The prior distribution $p(\theta)$ describes the degree of belief over the parameter space Θ.

  Ex. *Beta Distribution.* $p(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$ One example is Beta(1,1), Uniform Model. $p(p) = 1$, $p \in [0,1]$, $p = 0$ otherwise. Note $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

- **Posterior Distribution:** Given a prior distribution and a likelihood function, or sampling model, the posterior distribution of the parameters can be calculated using Bayes' rule.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}} \tag{1}$$

  Ex. *Beta Distribution* It turns out that for a binomial sampling model, a beta prior is *conjugate*, which means the prior and posterior have the same family of distribution.

In Bayesian statistics, inferences are made from the posterior distribution. In cases where analytical solutions are possible, the entire posterior distribution provides an informative description of the uncertainty present in the estimation. In other cases credible intervals are used to summarize the uncertainty in the estimation.

Experiment. Yellowstone Example (with Bayes).

Now reconsider the Yellowstone example from a Bayesian perspective. Use the Beta(α, β) as the prior distribution for p and compute the posterior distribution for p.

$$p(p|y) = \frac{p(y|p)p(p)}{\int p(y|p)p(p)dp} = \frac{\frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}}{\int \frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}dp}$$

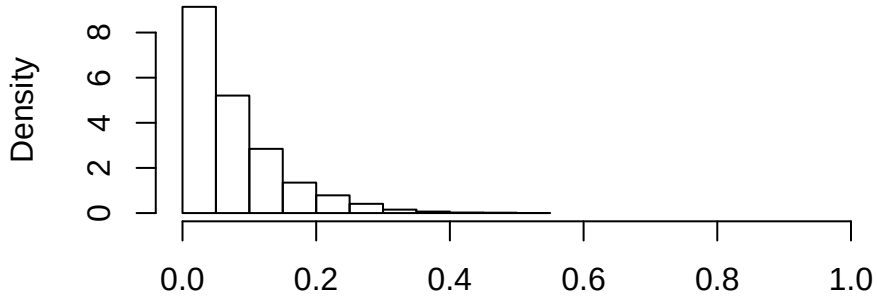
first the integration in the denominator.

$$\begin{aligned}
\int \frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}dp &= \frac{\binom{n}{y}}{B(\alpha,\beta)}\int p^{\alpha+y-1} + (1-p)^{\beta+n-y-1)}dp \\
&= \frac{\binom{n}{y}B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}\int \frac{p^{\alpha+y-1} + (1-p)^{\beta+n-y-1}}{B(\alpha+y,\beta+n-y}dp \\
&= \frac{\binom{n}{y}B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}
\end{aligned}$$

Now the posterior distribution follows as:

$$\begin{aligned}
p(p|y) &= \frac{\frac{\binom{n}{y}p^y(1-p)^{n-y}p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}}{\frac{\binom{n}{y}B(\alpha+y,\beta+n-y)}{B(\alpha,\beta)}} \\
&= \frac{p^{\alpha+y-1} + (1-p)^{\beta+n-y-1}}{B(\alpha+y,\beta+n-y)} \\
p|y &\sim Beta(\alpha+y,\beta+n-y).
\end{aligned}$$

Now use a Beta(1,1) distribution as the prior for p(p) and compute $p(p|y)$. Then $p(p|y) \sim Beta(1, 12)$.



.

What is the expectation, or mean, or your posterior distribution? Hint $E[X] = \frac{\alpha}{\alpha+\beta}$ if $X \sim Beta(\alpha, \beta)$.

$E[p|y] = \frac{1}{13}$.

How do these results compare with your intuition which we stated earlier?

Similar - and account for uncertainty in the estimated probability.

How about the MLE estimate?

Too small, and puts all of the confidence at zero without using a Bayesian type procedure for the confidence interval.

What impact did the prior distribution have on the posterior expectation?

It pulls the expectation away from zero in this case.

Classical, or frequentist, statistical paradigm:

- Estimate fixed parameters by maximizing the likelihood function $\mathcal{L}(\theta|X)$.

- Uncertainty in estimates is expressed using **confidence.** The concept of confidence requires a frequentist viewpoint. Specifically, a confidence interval states that if an experiment was conducted a large number of times, we'd expect the true estimate to be contained in the interval at the specified level. No probabalistic statements can be made (e.g. the probability the true value is in the specified confidence interval).

- Inference is often conducted using hypothesis testing and requires requires **p-values**. Conceptually, a p-value is the probability of obtaining the result (or a more extreme outcome) given the stated hypothesis. However, in recent years, the use of p-values has been under increased scrutiny. We will dedicate a class to this later in the semester.

Bayesian statistical paradigm

- Given a stated prior $p(\theta)$ a posterior distribution is computed for the parameters $p(\theta|X)$.

- Uncertainty in the estimates is expressed using the posterior distribution. Often the posterior is summarized by a **credible interval** which does allow probabilistic statements (e.g. the probability the true value is contained in the credible interval.)

- For inference, the posterior distribution is typically summarized using a credible interval and often combined with *maximum a posteriori* (MAP) point estimates.