Mathematical Statistics II

An Introduction

Jesse Wheeler

Contents

1 Course Introduction

Course Overview

- The larger focus of last semester (Math 4450) was probability.
- Though we continue where we left off, this semester (Math 4451) will have a much stronger focus on statistics. A complete description of planned course topics can be found at the course website: https://jeswheel.github.io/4451_s26/#planned-topics-spring-2026.
- Both probability and statistics are, fundamentally, the study of with randomness... so, what's the difference? **Depends on who you ask!**

> Statistical science was the peculiar aspect of human progress which gave to the twentieth century its special character... it is to the statistician that the present age turns for what is most essential in all its more important activities. – Fisher (1954)

2 Probability and Statistics

What is "Statistics"?

- First, what is statistics?

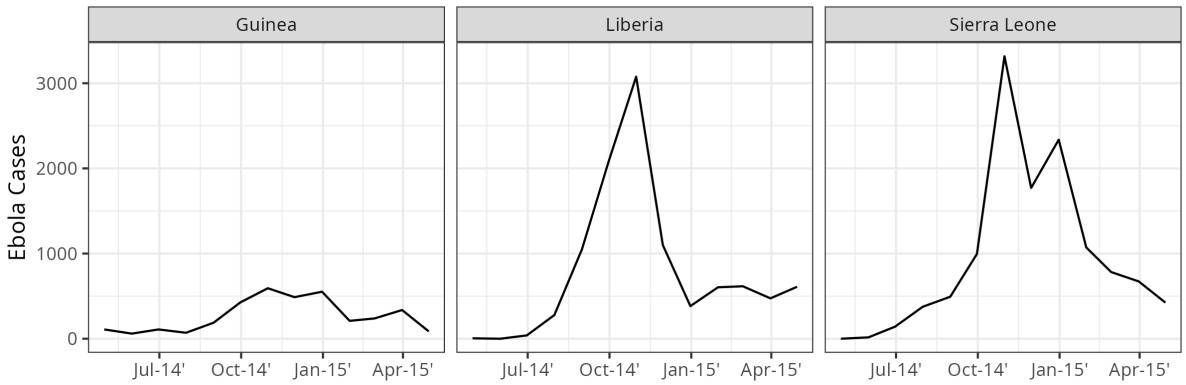
> "The science of collecting, displaying, and analysing data." – Oxford Dictionary (2008)
>
> "The discipline that concerns the collection, organization, analysis, interpretation, and presentation of data." – Wikipedia contributors (2025)
>
> Something like: "The study of extracting useful information from data in a rigorous way." – Me (it's hard to define an entire discipline).

Probability vs Statistics

- Any of the above definitions (accurately) suggests that probability is a key part of statistics. So where do we draw the line? Does it matter?
- Pawitan (2001) dichotomizes the difference in terms of *deductive* vs *inductive* reasoning.
- Roughly speaking, *deductive* arguments moves from general principles (assumptions) to make specific conclusions. In *inductive* reasoning, we use specific observations (data) to make broader generalizations.

Traffic Accidents
Suppose we are interested in the random quantity X_i, the number of accidents during week i at a particular intersection. From last semester, a common model for this situation is a Poisson-process.

- *Probability (deductive)*: If X_i follows a Poisson(λ) distribution (general principle), then what is the expected number of accidents per week (specific conclusion)? What is the probability that we observe more than 10 accidents?
- *Statistics (inductive)*: Suppose we count the number of accidents over a 6 week period, observing: $3, 4, 2, 7, 3, 3$ accidents (specific observations). What value λ might describe the Poisson-process that generated the data (broader generalization)? Is the Poisson assumption reasonable given the data?
- From the example above, we can see both ideas used in conjunction for making informed decisions.
- Many statistics problems rely on deductive reasoning in probability, geometry, topology, analysis, etc. to build theory for ways of performing inductive reasoning with specific observations (data).
- Another example related to my own research in population modeling...



- (Statistics) Given the data (specific example), what can we learn about the dynamic system / generative process (generalization)?
- (Probability) Under our assumed process / model (general principle), what is our prediction for the Ebola burden over the next year (specific conclusion)?

3 Statistics of Math 4451

Statistics and Math 4451

- Pawitan (2001) further categorizes statistics in terms of five key 'statistical activities' in the preface of his book:
  - *Planning*: making decisions about the study design or sampling protocol, what measurements to take, stratification, sample size, etc.
  - *Describing*: summarizing the bulk of data in few quantities, finding or revealing meaningful patterns or trends, etc.
  - *Modeling*: developing mathematical models with few parameters to represent the patterns, or to explain the variability in terms of relationship between variables.
  - *Inference*: assessing whether we are seeing a real or spurious pattern or relationship, which typically involves an evaluation of the uncertainty in the parameter estimates.
  - *Model Checking*: assessing whether the model is sensible for the data.
- A lot of early statistics were focused on the first two activities: *planning* and *describing*. We will not spend much time this semester discussing methods related to these two activities.

Statistics and Uncertainty

- Regardless of the type of "statistical activity" we are doing, a recurring theme in statistics in *uncertainty*.
- Loosely speaking, we can characterize two distinct forms of uncertainty:
  - *Stocahstic uncertainty*: uncertainty due to inherent randomness in data used to make inference, resulting in different models and estimates. For example, what if we picked different weeks to observe traffic patterns at the intersection? We would likely get different data, resulting in different parameter estimates.
  - *Inductive uncertainty*: uncertainty due to the inductive nature of our estimates. This is a result of having incomplete information (e.g., due traffic accidents at the intersection truly follow a Poisson-process?) We generally can't control this type of uncertainty, or even quantify it.
- The first type can be thought of as the uncertainty present, conditioned on a specific model. The second is related to uncertainty involved in the model selection itself.

Traffic Deaths

Traffic Deaths and Cell-Phone usage
Suppose instead of measuring total accidents at an intersection, what if we model the total number of deaths? We might be interested in estimating how cell-phone usage might be correlated with number of traffic deaths.

- Ideally, we would like to do a randomized experiment, placing drivers into a control (no cell-phone), or treatment group (cell-phone). *Recall from Math 3350, this is the best way to control for confounding.*
- This isn't possible, so instead we rely on what is called a *natural experiment*. The "control" group will be the population of drivers the year before the invention of cell-phones, and the "treatment" group is the population of drivers the year cell-phones were invented.

- Thought experiment: What if the number of deaths increase from 170 in one year, to 190 the next? Is this enough to claim cell-phones increase accidents? What about 170 to 300? 170 to 174? What's the cutoff?
  - The question above is related to changes in observed data, not the model. That is, fixed on a model, we can make statements about what is considered a "large enough" change in order to consider cell-phones having an impact.
- Now suppose the deaths increased from 170 to 300, but in the second year, a few major accident involved many cars, in which over 50 people died. Alternatively, what if the second year had a much longer winter, or there were more teen drivers? What other factors might change how we think about this?
  - This time, uncertainty arises because of model choice. We can increase model detail by splitting by age, month, etc., but the data then fall into smaller and smaller groups, increasing the stochastic uncertainty. We have to stop adding detail at some point, but where?

4 Bayesian vs Frequentist Statistics

Frequentist vs Bayesian

- When we finally settle down on a model, we can deal with the inherent stochastic uncertainty in the data in a rigorous way.
- Typically, this is done using probability.
- Probability is a surprisingly abstract topic, however, and how to connect real-world outcomes to probability is non-trivial.
- An at times frustrating (and possibly unique) feature of Statistics as a discipline is that experts cannot agree on the fundamental nature of their subject.
- There are two groups of thought: *Bayesian* and *Frequentist* perspective. We will discuss approaches from both in this class.

In the extremes, interpretation of probability falls into two main categories:

- *Bayesian* perspective: probabilities correspond to a (subjective) degree of belief about an event.
- *Frequentist* perspective: probabilities are interpreted only in long-run frequency of events.
- Note in the Bayesian perspective admits both perspectives, and many modern Bayesian approaches often consider the frequentist properties of their methods.
- In *my experience*, most statisticians actually fall somewhere in the middle.

Tossing a coin
Consider tossing a fair coin. We have some sense of uncertainty about the outcome of this experiment, and say that the probability of heads is 0.5.

- What does this mean in the purely frequentist sense? What about the uncertainty related to the *next specific* coin toss?

  Answer: In the extreme frequentist paradigm, this is interpreted to mean that the long-run proportion of heads (many coin flips) is 0.5. However, interpreting the uncertainty of the next *specific* toss is difficult (something like: the next toss belongs to an infinite sequence of tosses of the same coin, which have a limiting frequency of 0.5.)

- What does this mean in the purely Bayesian paradigm?

  Answer: In the extreme Bayesian paradigm, a 0.5 probability corresponds to a *belief* about the behavior of the coin; this belief can be entirely subjective. Saying a probability of 0.5 means that even in the next specific coin toss, our sense of uncertainty is 0.5.

- *Important*! Both approaches have their strengths and weaknesses, and some of the smartest scientists / mathematicians in the world can be found in either extreme. Though most statisticians will claim to think one-way versus the other, most fall somewhere in the middle (frequentists' will often be founding making subjective probability statements, whereas Bayesians' often admit to caring about frequentist properties of their methods).

- We will cover both approaches, and discuss strengths and weaknesses of each when relevant. However, the focus will be on frequentist statistics, as the vast majority of statistical methods in use in the 20th and 21st centuries are based on frequentism.

- There's also a "third way", sometimes called *Fisherian*, or *likelihoodist*. To some degree this is a compromise between the two more popular approaches, as it rejects the use of prior distributions, but still interprets likelihood as a way to measure belief about a parameter estimate (Pawitan, 2001).

Probability of events
Try to use both Bayesian and frequentist interpretations of probability to describe what is intended by the following statements.

- My weather app states that there is a 40% of rain tomorrow.

- In a sporting event, ESPN says that the probability that team A will win is 57%.

- A particular blood test given by a doctor is said to have a 5% probability of a false-positive result, and you just received a positive result.

Introduction to Estimation

- The primary focus of this class will be *parameter estimation*. That is, we collect data, pick a model to describe the data generating process, and use the data to "fit" the model to data.

- Models can be complex: machine learning frameworks like random forest, neural networks, gradient boosting machines, etc. These are very good at prediction, but are often referred to as a *black-box*.

- In this class, we focus more on general principles of data analysis, which is most easily demonstrated using simple models.

Flipping coins
A friend gives you a coin from another country. You want to estimate the probability of heads for this particular coin, flip the coin N times, observe $0 \leq n \leq N$ number of heads.

- In this case, the model for the data is obvious and simple. We will use it for both the frequentist and Bayesian solutions. Assuming coin flips are independent and identically distributed, the data will come from a binomial distribution with N total trials, probability of success p. We suppose each flip of a coin is independent, identically distributed. Let X_i be the random variable that is one with probability p, and zero with probability $1 - p$.

- *Frequentist Solution*:

- The probability of heads only has meaning as a long-run frequency of the number of heads (from the purely frequentist perspective, that's what probability means).
- Thus, we'll estimate the probability p by using the observed frequency of heads: n/N. That is, our estimate $\hat{p} = n/N = \sum_i X_i/N$.
- We recognize that there is uncertainty in this estimate. This time, we flipped the coin N times and saw n heads, but what if we saw $n-1$, $n+3$, or some other number of heads? This is entirely possible! Probabilities are long-run frequencies, and we only have a finite sample, so what is the uncertainty associated with our estimate?
- One approach is to use the CLT. Notably, the estimate $\hat{p}$ is just the average of the X_i, and therefore according to the CLT:

$$\hat{p} \approx N\big(p, p(1-p)/N\big).$$

- With enough observations, our estimate is centered at the "truth" p, and the variance of our estimate shrinks at a rate of $1/N$.
- The biggest issue remaining is that the variance term relies on p, which is unknown. Since $\hat{p} \overset{a.s.}{\to} p$ (by the strong law of large numbers), we'll replace p in the variance term to get the variance of the estimate to be approximately $\hat{p}(1-\hat{p})/N$.
- Finally, because we know that 95% of the area of a normal distribution lies within 1.96 standard deviations of the mean, we can create an interval that describes some level of uncertainty:

$$I_p := \hat{p} \pm 1.96 \times \sqrt{\hat{p}(1-\hat{p})/N}.$$

- We will call the interval above an approximate 95% confidence interval for p. Approximate because we used the CLT, and approximated the variance of the estimate.
- We'll talk more about confidence intervals later in the course, but while we are focusing on frequentist probability interpretation, let's talk about the interval.
- We DO NOT say that there is an approximate 95% probability that p is in the interval I_p. Probability only exist in terms of long-term frequencies, and once the interval is built, the parameter is or is not in the interval; there are no long-term frequencies involved at this point, so we make no probability statement.
- However, the interval I_p itself is random! If we observed different data, we would have made a different random interval.
- Thus, an approximate 95% confidence interval is best interpreted as: "Under the same data generating scenario, if we built a random interval in the same way that we just did, we would expect an approximate long-run frequency of 95% of these intervals to contain the true parameter, p."
- Example: Suppose we toss the coin 10 times, observe 8 heads. Then $\hat{p} = 8/10$, and then $\sqrt{\hat{p}(1-\hat{p})/N} = 0.126$, and our point estimate is 0.8, with a confidence interval of $0.8 \pm 1.96 \times 0.126 = (0.55, 1.05)$. Do we see a problem?

- *Bayesian Solution*:
  - The Bayesian solution comes as a direct result of Bayes-Theorem, allowing us to account for our belief about the coin as a probability distribution.
  - The Bayesian approach is actually considered the first modern method to assimilate observed data for quantitative inductive reasoning (Chapter 1.4, Pawitan, 2001).
  - Recall Bayes Theorem for two random events, A and B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}.$$

– Now, assuming that probabilities represent beliefs about particular outcomes, we can represent our own belief about p (the probability of observing a heads) using a probability distribution, $f_P(p)$. We can now think of events B and A being the event of observing n heads in N trials, and the event that $P = p$, respectively, and substitute this into Bayes theorem:

$$f_{P|x=n}(p|n) = \frac{f_{X|P}(n|p)f_P(p)}{\int f_{X|P}(n|p)f_P(p)\,dp},$$

Where $X = \sum_i X$, and we are now able to get a posterior belief about the parameter P after conditioning on the data.

– A key question now is what is our prior belief about the problem? How do we specify $f_P(p)$, which is called the prior distribution?

– A common approach is we say we don't know! We have no prior information, so we believe that any P is equally likely. Thus, we will use a *uniform-prior* over the support of p, which is the interval $[0, 1]$. Thus:

$$f_P(p) = 1[0 \leq p \leq 1].$$

– Then, the function $f_{X|P}(n|p)$ is uniquely determined by our model. In this case, it corresponds to the binomial probability of observing $X = n$:

$$\begin{aligned} f_{P|x=n}(p|n) &= \frac{f_{X|P}(n|p)f_P(p)}{\int f_{X|P}(n|p)f_P(p)\,dp} \\ &= \frac{\binom{N}{n}p^n(1-p)^{N-n}}{\int_0^1 \binom{N}{n}p^n(1-p)^{N-n}\,dp}1[0 \leq p \leq 1] \\ &= \frac{p^n(1-p)^{N-n}}{\int_0^1 p^n(1-p)^{N-n}\,dp}1[0 \leq p \leq 1]. \end{aligned}$$

– It can be tricky to continue simplifying the density above. We'll look at some tricks later in the semester, but for now we will just recall the definition of the Beta-integral:

$$B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1 + z_2)} = \int_0^1 p^{z_1-1}(1-p)^{z_2-1}\,dp,$$

using this we simply the conditional density $P|X = n$ as:

$$f_{P|x=n}(p|n) = \frac{\Gamma(N+2)}{\Gamma(n+1)\Gamma(N-n+1)}p^n(1-p)^{N-n}1[0 \leq p \leq 1].$$

– You may recognize the distribution above to be a special case of the Beta-distribution, hence:

$$P|X = n \sim \text{Beta}(n+1, N-n+1).$$

– To get a point estimate, we might consider taking the mean of the posterior distribution. The mean of a Beta(α, β) distribution is $\alpha/(\alpha+\beta)$, so our point estimate might be:

$$\hat{p} = E[P|X = n] = \frac{n+1}{N+2}.$$

– Note the similarity of this estimate to the frequentist estimate of n/N. Noticeably, the Bayesian estimate is "shrunk towards 1/2".

– To quantity uncertainty, we might want to find the lower 0.025 percentile, and 0.975 percentile, so that we have an interval where 95% of the area is in-between these two numbers. We call this the *credible interval*. There's no analytic way to do this, but it's easy to calculate in a statistical software environment.

- Example: Suppose we observe 10 flips, and 8 are heads. Then our point estimate (posterior mean) is $\frac{8+1}{10+2} = \frac{9}{12} = 0.75$. We can calculate a 95% credible in R.
- Are there any issues with this approach? Is a uniform prior really appropriate measure of uncertainty you "feel" about flipping a coin?
- Do you recognize any relationship between the Bayesian point estimate and the approach taught in 3350 for small sample sizes? (If you didn't take Math 3350, consider looking up the "plus-two" method)

R code

```
n <- 8
N <- 10
lower_bound <- qbeta(0.025, n+1, N-n+1)  # lower bound
upper_bound <- qbeta(0.975, n+1, N-n+1)  # upper bound
```

```
95% Credible Interval for p: (0.48, 0.94)
```

Some Helpful R tips:

- All common distributions in R follow a similar pattern: A single letter of the thing you want, followed by the name of the distribution.
  - `r<dist>`: Get random sample from `<dist>`. Example: `rnorm(10, mean = 5, sd = 8)` will generate 10 random samples from a $N(5, 8^2)$ distribution.
  - `d<dist>`: evaluate the density of a distribution at a particular point. Example: If $\phi(x)$ is the density of a standard normal, then $\phi(2) =$ `dnorm(2, mean = 0, sd = 1)` $= 0.05$.
  - `p<dist>`: Calculates the distribution function (CDF) of the distribution at a particular point. Example: If $\Phi(x)$ is the CDF of the standard normal, then $\Phi(1.95) = P(Z \leq 1.96) =$`pnorm(1.96, mean = 0, sd = 1)` $= 0.975$.
  - `q<norm>`: The quantile function of the desired distribution, which is the inverse of the CDF. Example: `qnorm(0.025, mean = 0, sd = 1)` $= -1.96$.
- Available distributions include: `gamma`, `beta`, `unif`, `t`, `pois`, `exp`, `binom`, `nbinom`, `geom`, `hyper`, and several more.

Final Thoughts

- The differences between Bayesian vs Frequentist thinking matters, and some get very passionate about the debate.
- In my opinion, most people fall somewhere in the middle.
- Often practitioners choose a Bayes vs Frequentist methodologies not because of their personal interpretation of probability, but for convenience, existing standards, or it is the only way to solve a given problem.
- "All models are wrong but some are useful" – Box (1979), a famous 20th century statistician.
- One example is the following Bayesian approach to image restoration for images with static noise introduced (Chapter 15.6 Keener, 2010), adapted from the seminal paper by Geman and Geman (1984).
- Let $X_{i,j}$ denote the value of the (i, j)th pixel in a digital image.

- We will pick a simple (obviously incorrect) model that is very useful for performing image restoration. In particular, assume that, conditioned on a particular mean $\Theta_{(i,j)} = \theta_{(i,j)}$, each pixel has distribution:

$$X_{i,j} \sim N(\theta_{i,j}, \sigma^2).$$

- In real images, pixels that are close together tend to be highly correlated. Thus, using a Bayesian set-up, we want to pick a prior for $\Theta_{(i,j)}$ that has this feature. It is a bit too technical to describe in detail here, but we can define a multi-variate normal prior for each $\Theta_{(i,j)}$, where adjacent pixels are correlated to each other, and far-away pixels are not. Assuming this structure is captured in the covariance matrix Σ_θ, the prior for the matrix of pixels may be denoted:

$$\Theta \sim N_{n \times m}\big(\mu_\theta, \Sigma_\theta\big),$$

  meaning an $n \times m$ dimensional

- The posterior distribution of $\Theta | X = x$ will also be normally distributed (normal model with normal priors = normal posterior). Direct computation of the mean and covariance of the posterior is possible, but Gibbs-sampling can instead be used to compute reliable approximations in faster time.

- In this example, I don't think most would believe the model that was used is "correct", nor do we necessarily have to accept the Bayesian interpretation of probability; instead, the method based on Bayesian theory allows us to "smooth out" the image

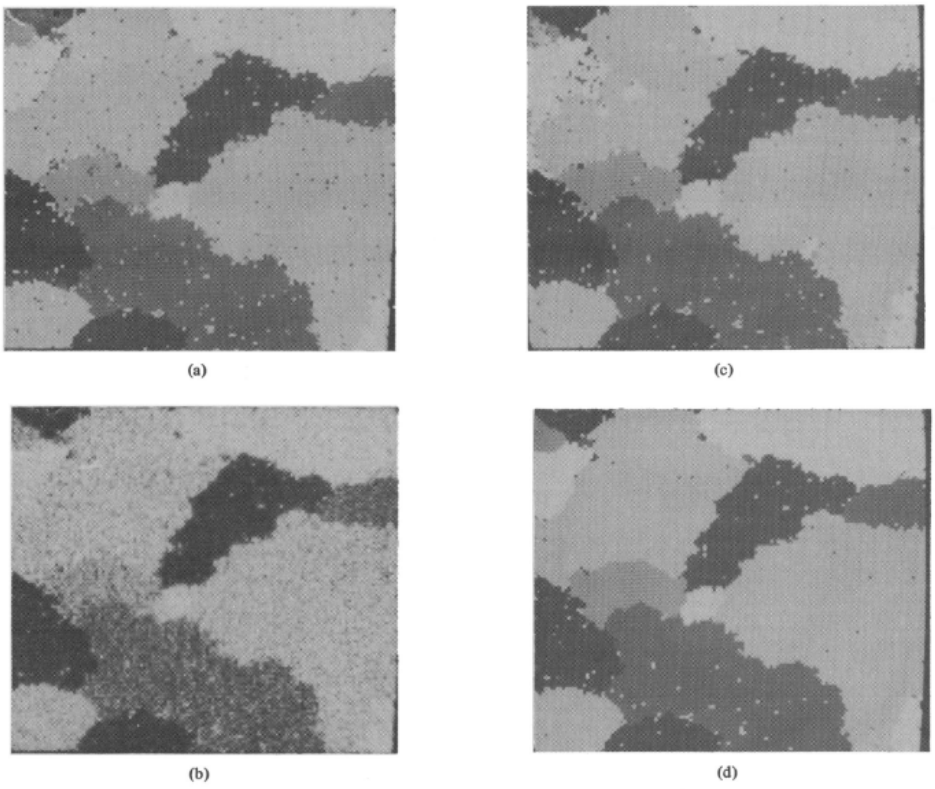


Figure 1: Bayesian image restoration, credit Geman and Geman (1984)

Acknowledgments

- Compiled on January 13, 2026 using R version 4.5.2.

- We acknowledge students and instructors for previous versions of this course / slides.

References

Box GE (1979). "Robustness in the Strategy of Scientific Model Building." In *Robustness in Statistics*, pp. 201–236. Elsevier. 10

Fisher R (1954). "The expansion of statistics." *American Scientist*, **42**(2), 275–293. 1

Geman S, Geman D (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**(6), 721–741. doi: 10.1109/TPAMI.1984.4767596. 10, 1

Keener RW (2010). *Theoretical statistics: Topics for a core course*. Springer Science & Business Media. 10

Oxford Dictionary (2008). "Statistics." doi: 10.1093/acref/9780199541454.013.1566. URL https://www.oxfordreference.com/view/10.1093/acref/9780199541454.001.0001/acref-9780199541454-e-1566. 2

Pawitan Y (2001). *In all likelihood: statistical modelling and inference using likelihood*. Oxford University Press. 3, 4, 7, 8

Wikipedia contributors (2025). "Statistics — Wikipedia, The Free Encyclopedia." [Online; accessed 9-January-2026], URL https://en.wikipedia.org/w/index.php?title=Statistics&oldid=1328458961. 2