STAT 447 Quiz 1 Review

Caden Hewlett

2024-02-24

Terminology Review

Part of quiz 1 involves knowing these terms without having to look them up.

While in "the real world" you can look things up, to attain mastery of a subject you need a critical mass of concepts in memory. Too much Google look-ups prevent thinking at the speed of thought.

Probability Essentials

Probability vs PMF vs Density

A *probability* describes the chance that an event occurs. For example, we can say "there's a 25% chance of rain tomorrow." That describes a probability. We're saying "the chance that the outcome of the weather tomorrow is in fact rain is 1 in 4, or 25%."

In a similar vein, PMFs, or *Probability Mass Functions*, are ways of describing the probability of an event. Specifically, for what are called **discrete** random variables. A Discrete random variable (rv) is one that takes values in integers, i.e. $\{0, 1, 2, 3, 4\}$. To re-use the previous example, this could be "the number of times it rains in the next four days." If we let the probability of rain remain as $p = 0.25$, the previous sentence can be described using a Binomial Distribution, which is one of many discrete distributions. We could then pick any number of rainfall counts Count $\in \{0, 1, 2, 3, 4\}$ and find the probability of that Count directly from the PMF by reading it.

To contrast, a *Probability Density Function*, or PDF, takes values on the real line $\mathbb{R}$. From the rainfall example, this could be "how many minutes it rains for." There could theroetically be any value in the real numbers. For example, it could rian for *exactly* six minutes, or it could be 6.01324 minutes, or 6.000018 minutes, or any other level of granularity. Due to this fact, the probability of any specific value is zero! In fact, a PDF does not give probabilities at all, it gives *densities.* Essentially, a density describes where certain values are likely to be distributed. We can translate regions of a PDF into probability values using integration. For example, it's possible to find "the probability it rains between 4 to 6 minutes."

Event vs Outcome

The sentence "the weather tomorrow" describes an *event.* It hasn't come to pass yet, so we can find and compute certain things about it. For example, we can find "the probability that it rains tomorrow," or "the expected number of minutes it will rain tomorrow." We can't, however, actually **decide** definitively whether or not it will rain tomorrow. The future is always uncertain, but we can attempt to predict it with statistics.

An Outcome is the result of an event. In our previous example, if we had "no rain" and "rain" as our two possible outcomes and we wished to compute the probability it rains tomorrow, we'd be finding the "probability that the Event describing the weather tomorrow has the Outcome of rain."

In more formal words, we can compute $\mathbb{P}(\text{Weather} = \text{Rain})$. Here, again, "Weather" is the Event and "Rain" is the Outcome. Directly, the probability of all possible outcomes of an event, sometimes referred to as the sample space, is 1. In other words $\mathbb{P}(\text{Weather} = \text{Rain} \cup \text{Weather} = \text{No Rain}) = 1$. You may have heard about this fact in the adage "everything in the universe either is or is not a potato."

Random Variable

Random Variables are the bridge that allows us to mathematically model and analyze the probabilities of different events and outcomes. It combines both the PMF/PDF discussion and the Event/Outcome discussion. In most nomenclature, random variables are indicated with capital letters (e.g. X), while the outcome of a random variable is describe with a lowercase letter (e.g. x.)

For example, if we let the Weather tomorrow be our random variable X, and the outcomes of "Rain" and "No Rain" be 1 and 0 (with probability p of rain being still being 0.25), we could say that the Random Varibale describing the Event of weather tomorrow takes a Bernoulli Distribution, which is a *discrete random variable* with the following PMF:

$$\mathbb{P}(X = x) = p^x(1-p)^{1-x}$$

In our case, recalling that x is either "Rain" or "No Rain" and the probability p is 0.25, you can find directly there's a 25% chance of rain and a 75% chance of no rain.

This is the simplest way to combine the concepts of Events/Outcomes with PMFs/PDFs. The combining of the Events themselves and the mathematical functions that describe them into the idea of a "random variable" is a foundational aspect of statistics.

Realization

A **Realization** is the result of a random variable that *we have already seen.* For example, if we had some PMF describing the probability it rained *yesterday* we now know for certain whether or not it did rain yesterday. In effect it is no longer random, but can be thought of as the result of some random process.

This concept can be a bit tricky to wrap your head around, so here's another example. Say you want to flip a coin. Prior to the flip (and while the coin is in the air) the result of the coin flip is uncertain, but you know it has a 50/50 chance of heads or tails. Since the outcome is unknown but we can describe it mathemtically, the coin prior to landing is a random variable. Once the coin lands, you can determine whether or not it was heads or tails. This is the realization of that random variable. In effect, the realization is deterministic.

Conditional Probability & Mathematical Statistics

The **Conditional Probability** of an Event is the probability of an Outcome, given some additional information.

To illustrate this concept, we will move away from the rainfall example. Let's say that you, like me, have allergies. Let's consider the Event that I sneeze in the next ten minutes. We'll denote this X, and it has outcomes $x \in \{0, 1\}$, where a 1 indicates that I *do* sneeze, and a 0 indicates that I don't. If we consider $\mathbb{P}(X = 1)$ - this the probability that I sneeze in the next ten minutes.

Let's say, however, that you have some additional information - the outcome of some *new* random variable Y. In this case, Y is the Event that I sniffed a flower in the previous 30 seconds. It shares the same set of yes/no outcomes $y \in \{0, 1\}$. It's logical to conclude that it is more likely that I will sneeze in the next ten minutes *given that I have smelled a flower* rather than if I hadn't. This is one example of conditional probability. It can be written as follows:

$$\mathbb{P}(X = 1 \mid Y = 1) = \mathbb{P}(\text{ I Sneeze} \mid \text{ I Smelled a Flower})$$

Where the | symbol reads as "given."

On the more technical side, conditional probability is most often computed using Bayes' Theorem, given below. Here, A and B are events.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

The denominator is often computed via the Law of Total Probability.

$$P(B) = \sum_i P(A_i)P(B \mid A_i)$$

Where A_i is the i-th partition of the sample space of A. This bit is rather technical, but very useful.

Expectation vs Probability

We'll now transition into a lot more technical formulation. Expectation, denoted $\mathbb{E}$, is the expected value of a Random Variable. For a discrete random variable with a support $i \in n \subseteq \mathbb{Z}$, it is computed with the following:

$$\mathbb{E}(X) = \sum_{i=1}^{n} x\mathbb{P}(X = x)$$

As you can see, Probability is a subset of the Expectation calculation.

For a discrete RV, the *Cumulative Probability* of an Event k is:

$$\mathbb{P}(X \leq k) = \sum_{i\,:\,x_i \leq k} \mathbb{P}(X = x_i)$$

Where $\{i : x_i \leq k\}$ describes the set of outcomes prior to and including k.

Letting S be the support and $k = \max(S)$ (i.e. the greatest possible value in the Support of the Random Variable,)

$$\mathbb{P}\big(X \leq \max(S)\big) = \sum_{x_i \in S} \mathbb{P}(X = x_i) = \sum \mathbb{P}(\text{all possible outcomes}) = 1$$

Similarly, for a Continuous Random Variable with PDF $f_X(x)$, the Expectation is given by:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x$$

Note that certain continuous random variables (such as the Chi-Square distirbution) only take values in the positive reals, hence the region of integration is sometimes written as $0 \to \infty$ instead.

The CDF, sometimes denoted $F_X(x)$, is given by:

$$F_X(k) = P(X \leq k) = \int_{-\infty}^{k} f_X(x)\mathrm{d}x$$

Then, finally, the probability calculations for a continuous random variable is the integration between two points, i.e.

$$P(a \leq X \leq b) = P\big((X \leq b) \cap (X \geq a)\big) = \int_a^b f_X(x)\mathrm{d}x$$

This same logic can be applied to summation for equivalent probability calculations with discrete random variables.

Indicator Function

The indicator function is an excellent tool in Bayesian Statistics. It can translate events and outcomes into binary values (and, hence, to Bernoulli Distributions.)

For example, let $(X = x)$ describe an Event X and a possible Outcome x. The indicator of this event is written:

$$\mathbb{1}[X = x] = \begin{cases} 1, & X = x \\ 0, & \text{otherwise} \end{cases}$$

As the definition suggests, the indicator function *indicates* whether or not a specific outcome has occurred.

Common Distributions

Distributions: Parameters and Support:

Distribution	**Parameters**	**Support**
Bernoulli	p (success probability)	$\{0, 1\}$
Binomial	n (trials), p (success probability)	$\{0, 1, \ldots, n\}$
Discrete Uniform	$\{x_i\}_{i=1}^n$ (possible outcomes)	$\{x_1, x_2, \ldots, x_n\}$
Continuous Uniform	a (min), b (max)	$[a, b]$
Categorical	$\{x_i\}_{i=1}^n$ (outcomes), $\{p_i\}_{i=1}^n$ (probabilities)	$\{x_1, x_2, \ldots, x_n\}$
Normal	μ (mean), σ^2 (variance)	$(-\infty, +\infty)$
Exponential	λ (rate)	$(0, +\infty)$
Beta (mean, precision)	$\mu \in (0, 1), s > 0$	$(0, 1)$
Beta (shape parameters)	$\alpha = \mu \cdot s, \beta = (1 - \mu) \cdot s$	$(0, 1)$

Basic Bayesian Terminology

For the following, we let Θ be the parameter and Y be the data.

- Prior Distribution: Given by PDF $p_\Theta(\theta)$ in the continuous case, or more simply as $P(\Theta = \theta)$ in the discrete case.
- Likelihood: Probability of the observations given the parameter: $p_{Y|\Theta}(y, \theta)$ or $P(Y = y \mid \Theta = \theta)$.
- Joint Distribution: In a Bayesian context given by (discrete) $\gamma(x) = P(X = x, Y = y)$ which is:

$$\gamma(x) = p_{X,Y}(x, y) = p_X(x) p_{X|Y}(x \mid y)$$

- Normalizing Constant: Often given by Z, is the overall probability $P(Y = y)$. Often difficult to compute by hand, but can be done via the LOTP.
- Posterior distribution (PMF/density): The posterior $P(\Theta \mid Y = y)$ is the distribution of the parameter given the data.

$$\pi(x) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P(X = x)P(Y = y \mid X = x)}{P(Y = y)} = \frac{\gamma(x)}{Z}$$

- Joint vs Marginal Posterior:

Let Θ_1 and Θ_2 be two parameters of interest.

The **Joint** posterior is given by the following (in the discrete case):

$$\text{Joint Posterior} = P(\Theta_1 = \theta_1, \Theta_2 = \theta_2 \mid Y = y)$$

An example of this could be the Normal Distribution parameters.

Conversely, the **Marginal** Posterior for Θ_1 is:

$$\text{Marginal Posterior} = P(\Theta_1 = \theta_1 \mid Y = y) = \sum_{\theta_2 \in \Theta_2} P(\Theta_1 = \theta_1, \Theta_2 = \theta_2 \mid Y = y)$$

Or, in the Continuous case,

$$\text{Marginal Posterior of } \Theta_1 = \int p_{\Theta_1, \Theta_2 | Y}(\theta_1, \theta_2 \mid y) \mathrm{d}\theta_2$$

An example of this would be the marginal posterior of the Slope parameter of a regression model.

- Predictive Distribution: Involves plugging new values into the model using the parameters drawn from the posterior distribution. We saw this as $\mathbb{P}(Y_4 = 1 | Y_{1:3} = \vec{1})$, but can be summarized using

In the discrete case where $\pi(x) = \mathbb{P}(X = x | Y = y)$

$$P(Y_{n+1} = y_{n+1} \mid Y_{1:n} = y_{1:n}) = \sum_{\theta \in \Theta} P(Y_{\text{new}} = y_{\text{new}} \mid \theta) P(\theta \mid Y_{1:n} = y_{1:n}) = \sum_{\theta \in \Theta} P(Y_{\text{new}} = y_{\text{new}} \mid \theta) \pi(\theta)$$

Notably, this uses the distribution of the <u>data</u> given the posterior parameter.

We have used the **unique case** where we have a Bernoulli likelihood model:

$$\mathbb{P}(Y_{n+1} = 1 \mid Y_{1:n} = y_{1:n}) = \mathbb{E}(p \mid Y_{1:n} = y_{1:n})$$

In other words, the Predictive Distribution follows the Posterior Mean.

Sampling