# Some Background from Statistics

Andrew Nobel

April, 2020

# Conditional Expectations and Probabilities

# Expectations

**Recall:** Let $X \in \mathbb{R}$ be a random variable

- If $X \sim p$ and $\sum_x |x|\, p(x)$ is finite then $\mathbb{E}X = \sum_x x\, p(x)$
- If $X \sim f$ and $\int |x|\, f(x)\, dx$ is finite then $\mathbb{E}X = \int x\, f(x)\, dx$

**Basic properties:** Let $X$, $Y$ be jointly distributed random variables

- If $X \leq Y$ then $\mathbb{E}X \leq \mathbb{E}Y$
- $\mathbb{E}(aX + bY) \; = \; a\,\mathbb{E}X + b\,\mathbb{E}Y$
- $|\mathbb{E}X| \leq \mathbb{E}|X|$
- If $X \perp\!\!\!\perp Y$ then $\mathbb{E}(XY) = \mathbb{E}X\,\mathbb{E}Y$, provided all expectations well defined
- If $X \geq 0$ then $\mathbb{E}X \; = \; \int_0^\infty \mathbb{P}(X \geq t)\, dt$

# Indicator Functions

**Definition:** If $\mathcal{X}$ is a set and $A \subseteq \mathcal{X}$ the indicator function of $A$ is given by

$$\mathbb{I}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

Sometimes $\mathbb{I}_A(x)$ is written $\mathbb{I}(x \in A)$. Basic properties

- $\mathbb{I}_{A^c} = 1 - \mathbb{I}_A$
- $\mathbb{I}_{A \cap B} = \mathbb{I}_A \, \mathbb{I}_B$
- $\mathbb{I}_{A \cup B} = \max(\mathbb{I}_A, \mathbb{I}_B)$
- If $X$ is a random variable $\mathbb{E}[\mathbb{I}_A(X)] = \mathbb{P}(X \in A)$
- $\int_A h(x)\, dx = \int h(x) \mathbb{I}_A(x)\, dx$

# Conditional Expectation

Let $(X, Y)$ be jointly distributed with $X \in \mathcal{X}$ and $Y \in \mathbb{R}$ with $\mathbb{E}|Y|$ finite

- If $Y$ is discrete with conditional pmf $p(y|x)$ let $\varphi(x) = \sum_y y\, p(y|x)$
- If $Y$ is continuous with conditional pdf $f(y|x)$ let $\varphi(x) = \int y f(y|x)\, dy$

**Definition:** The *conditional expectation* of $Y$ given $X$ is given by

$$\mathbb{E}(Y|X) = \varphi(X) \quad \text{and} \quad \mathbb{E}(Y|X = x) = \varphi(x)$$

**Note**

- $\mathbb{E}(Y|X)$ is a random variable, and is a function of $X$
- $\mathbb{E}(Z|X, Y) = \varphi(X, Y)$ where $\varphi(x, y) = \sum_z z\, p(z|x, y)$

# Properties of Conditional Expectation

1. If $Y \geq 0$ then $\mathbb{E}(Y|X) \geq 0$ (positivity)
2. $\mathbb{E}(aZ + bY|X) = a\,\mathbb{E}(Z|X) + b\,\mathbb{E}(Y|X)$ (linearity)
3. $\mathbb{E}\{\,\mathbb{E}(Y|X)\,\} = \mathbb{E}Y$ (law of total expectation)
4. $\mathbb{P}(Y \in A\,|X) = \mathbb{E}(\mathbb{I}_A(Y)\,|X)$
5. $\mathbb{E}[f(X)g(Y)|X] \;=\; f(X)\,\mathbb{E}(g(Y)|X)$ (functions of $X$ act like constants)
6. $\mathbb{E}(h(Y)|X = x) \;=\; \sum_y h(y)\,p(y|x)$
7. If $g : \mathbb{R} \to \mathbb{R}$ is convex then $g(\mathbb{E}(Y|X)) \leq \mathbb{E}(g(Y)|X)$ (Jensen)

# Conditional Expectation and Prediction

**Fact:** Let $(X, Y)$ be jointly distributed. Suppose we wish to predict $Y$ be a function of $X$. For any function $h : \mathcal{X} \to \mathbb{R}$

$$\mathbb{E}(Y - h(X))^2 \geq \mathbb{E}(Y - \mathbb{E}(Y|X))^2$$

Upshot: Under MSE $\mathbb{E}(Y|X)$ best predictor of $Y$ among all functions of $X$

*Turns out:* Conditional expectation $\mathbb{E}(Y|X)$ is the MSE projection of $Y$ onto the subspace of square integrable functions of $X$.

# Maximum Likelihood Estimation

# Distribution Family

**Given:** Family $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ of probability mass/density functions on $\mathcal{X}$

- $\Theta \subseteq \mathbb{R}^d$ called parameter space, $\theta \in \Theta$ called parameters
- Parameter $\theta \in \Theta$ fully specifies mass/density function $f_\theta$

**Examples**

- Normal $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$
- Exponential $\mathcal{P} = \{\text{Exp}(\lambda) : \lambda > 0\}$
- Poisson $\mathcal{P} = \{\text{Poiss}(\lambda) : \lambda > 0\}$
- Binomial $\mathcal{P} = \{\text{Bin}(n, p) : p \in [0, 1]\}$

# Inference: Parameter Estimation from Data

**Given**

- Distribution family $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ of interest
- Data set $x_1, \ldots, x_n \in \mathcal{X}$
- Assume: $x_1, \ldots, x_n$ drawn indep. from $f_{\theta_0} \in \mathcal{P}$ with $\theta_0$ unknown

**Goal:** Estimate $\theta_0$ (and therefore $f_{\theta_0}$) from data $x_1, \ldots, x_n$

**Idea:** Select $\theta \in \Theta$ that makes given $x_1, \ldots, x_n$ most likely

# Maximum Likelihood Estimation

**Definition:** The likelihood of $\theta \in \Theta$ is joint density of $x_1, \ldots, x_n$ under $f_\theta$

$$L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$$

**Definition:** The maximum likelihood estimator (MLE) of $\theta_0$ is

$$\hat{\theta}_n^{\text{MLE}}(x_1^n) \; = \; \operatorname*{argmax}_{\theta \in \Theta} L(\theta)$$

**Note:** As $\log(u)$ strictly increasing, MLE can be written in equivalent form

$$\hat{\theta}_n^{\text{MLE}}(x_1^n) \; = \; \operatorname*{argmax}_{\theta \in \Theta} \log L(\theta) \; = \; \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \log f_\theta(x_i)$$

# Maximum Likelihood Estimation

**Fact:** Under appropriate conditions the MLE is

- *Consistent:* $\hat{\theta}_n^{\text{MLE}}(X_1^n) \to \theta_0$ in probability
- *Asymptotically Normal:* $n^{1/2}\,(\hat{\theta}_n^{\text{MLE}}(X_1^n) - \theta_0) \Rightarrow \mathcal{N}(0, I(\theta_0)^{-1})$

**Ex1.** $X_1, \ldots, X_n$ iid $\sim f \in \mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$ with $\sigma^2 > 0$ known