Summary

Joint Probability Distributions If X, Y are random variables:

$$p(x,y) = \mathbb{P}(X = x, Y = y) \qquad \text{(joint mass function if } X, Y \text{ discrete)}$$
$$F(x,y) = \mathbb{P}(X \leq x, Y \leq y) \qquad \text{(joint cumulative distribution function)}$$
$$f(x,y) = \frac{\partial^2 F}{\partial x \partial y} \qquad \text{(joint density, if } X, Y \text{ continuous)}$$

If $A \subseteq \mathbb{R}^2$ then

$$\mathbb{P}((X,Y) \in A) = \begin{cases} \displaystyle\iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y & \text{if } X, Y \text{ discrete} \\ \displaystyle\sum_{(x,y) \in A} p(x,y) & \text{if } X, Y \text{ continuous} \end{cases}$$

Independence X, Y independent $\iff \exists g, h$ such that $f(x,y) = g(x)h(y)$ on entirity of $\mathbb{R}^2$. In such a situation can choose $g(x) = f_X(x)$ and $h(y) = f_Y(y)$.
Corresponding result holds for discrete distributions.

Marginal Mass/Density Functions

$$p_X(x) = \sum_y p(x,y) \qquad p_Y(y) = \sum_x p(x,y)$$
$$f_X(x) = \int_y f(x,y)\,\mathrm{d}y \qquad f_Y(y) = \int_x f(x,y)\,\mathrm{d}x$$

Sums of Independent Variables If X, Y are independent then

$$F_{X+Y}(a) = \mathbb{P}(X + Y \leq a) = \iint_{x+y \leq a} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{-\infty}^{\infty} F_X(a-y) f_Y(y)\,\mathrm{d}y$$

Convolution formula:

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} f_X(a-y) f_Y(y)\,\mathrm{d}y$$

Conditional Mass/Density Functions

$$p_{X|Y}(x|y) = \mathbb{P}(X = x | Y = y) = \frac{p(x,y)}{p_Y(y)} \qquad \text{(only valid if } p_Y(y) \neq 0\text{)}$$
$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

Change of Variables If $U = U(X,Y)$ and $V = V(X,Y)$ are a change of variables and $\mathcal{S}, \mathcal{T}$ are sets such that

$$(U,V) \in \mathcal{T} \iff (X,Y) \in \mathcal{S}$$

then

$$\begin{aligned}\iint_{\mathcal{S}} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \mathbb{P}((X,Y) \in \mathcal{S}) &= \mathbb{P}((U,V) \in \mathcal{T}) \\ &= \iint_{\mathcal{T}} f_{U,V}(u,v)\,\mathrm{d}u\,\mathrm{d}v \\ &= \iint_{\mathcal{S}} f_{U,V}\left(u(x,y), v(x,y)\right) \left|\frac{\partial(u,v)}{\partial(x,y)}\right| \mathrm{d}x\,\mathrm{d}y\end{aligned}$$

where $\dfrac{\partial(u,v)}{\partial(x,y)} = J(x,y) = \det\begin{pmatrix} u_x & u_y \\ v_x & v_y \end{pmatrix}$ is the Jacobian. It follows that

$$f_{U,V}(u,v) = f_{X,Y}(x,y)\,|J(x,y)|^{-1}$$

Expectations If $g(X,Y)$ is any function, then

$$\mathbb{E}(g(X,Y)) = \iint g(x,y) f(x,y)\,\mathrm{d}x\,\mathrm{d}y \quad \text{or} \quad \sum_x \sum_y g(x,y) p(x,y)$$

- Linearity: $\mathbb{E}(g(X) + h(Y)) = \mathbb{E}(g(X)) + \mathbb{E}(h(Y))$
- If X, Y independent, then $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$
- Covariance: $\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \mathbb{E}(XY) - \mu_X\mu_Y$, where $\mu_X = \mathbb{E}(X)$, etc.
- Independence $\implies$ $\mathrm{Cov}(X,Y) = 0$. Converse *false*.
- Correlation: $\mathrm{Corr}(X,Y) = \rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$
- Cauchy–Schwarz: $-1 \le \rho \le 1$ with equality iff X and Y are linearly related. In such a case,
  $$\frac{Y-\mu_Y}{\sigma_Y} = \pm\frac{X-\mu_X}{\sigma_X}$$
- $\mathrm{Var}\sum X_i = \sum \mathrm{Var}\,X_i + 2\sum_{i \ne j} \mathrm{Cov}(X_i, X_j)$

  If all X_i are pairwise independent, then $\mathrm{Var}\sum X_i = \sum \mathrm{Var}\,X_i$

Bivariate Normal Distribution Joint density function

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_X}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right]\right\}$$

- X, Y independent iff $\rho = 0$
- $X \sim N(\mu_X, \sigma_X^2) \qquad Y \sim N(\mu_Y, \sigma_Y^2)$
- $\mathrm{Corr}(X,Y) = \rho$
- $X|Y = y \sim N\left(\mu_X + \rho\sigma_X\dfrac{y-\mu_Y}{\sigma_Y}, (1-\rho^2)\sigma_X^2\right)$

Sample Mean and Variance If $X_1, \dots, X_n$ are identically distributed random variables, then the *sample mean* is the random variable

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Idea: each X_i is the result of an independent trial/measurement from a large population. If population (each of the X_i's distributions) has mean μ and variance σ^2, then the sample mean may be used to estimate μ:

$$\mathbb{E}(\overline{X}) = \mu$$

The *sample variance* is the random variable

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

This estimates the population variance: $\mathbb{E}(S^2) = \sigma^2$.
The larger the sample, the smaller the variance of the sample mean:

$$\operatorname{Var}\overline{X} = \frac{\sigma^2}{n}$$

Conditional Expectations $\mathbb{E}(X|Y) = \mathbb{E}(X|Y = y)$ is a function of y
Conditioning: for any *intermediate variable* Y

$$\mathbb{E}(X) = \mathbb{E}\big(\mathbb{E}(X|Y)\big) = \int_y \mathbb{E}(X|Y = y) f_Y(y)\,\mathrm{d}y \quad \text{or} \quad \sum_y \mathbb{E}(X|Y = y) p_Y(y)$$

Idea: average of X equals the average over Y of all of the averages of X given Y.

Prediction If X, Y have finite variances, then the best linear predictor of X given $Y = y$ is

$$X \approx g(y) = \mu_X + \rho\sigma_X\left(\frac{y - \mu_Y}{\sigma_Y}\right)$$

Among all linear functions of Y, the linear predictor minimizes the functional

$$\mathbb{E}\left[(X - g(Y))^2\right] \qquad (*)$$

Among all functions of Y, the function which minimizes $(*)$ is

$$g(Y) = \mathbb{E}(X|Y)$$

Conditional Variance $\operatorname{Var}(X|Y)$ is also a function of $Y = y$. Indeed

$$\begin{aligned}\operatorname{Var}(X|Y = y) &= \mathbb{E}\left(\big(X - \mathbb{E}(X|Y)\big)^2 | Y = y\right) \\ &= \mathbb{E}\big(X^2|Y = y\big) - \big(\mathbb{E}(X|Y = y)\big)^2\end{aligned}$$

Can condition to compute variances

$$\operatorname{Var} X = \mathbb{E}\big(\operatorname{Var}(X|Y)\big) + \operatorname{Var}\big(\mathbb{E}(X|Y)\big)$$

Moments

(Raw) Moments $\mathbb{E}(X^n)$ where $n = 1, 2, 3, \ldots$

Central Moments $\mathbb{E}\left[(X-\mu)^n\right]$ where $\mu = \mathbb{E}(X)$

Higher moments give more information on shape of a distribution: e.g. third central moment related to skewness of distribution.
General idea: know the moments, know the distribution (or at least can approximate to any accuracy)

Moments of number of events which occur Write $X = \sum X_i$ where X_i is the indicator for an event. Then X is the number of events which occur.

Events: $\mathbb{E}(X) = \sum \mathbb{E}(X_i) = \sum \mathbb{P}(X_i = 1)$

Pairs: $\mathbb{E}\binom{X}{2} = \sum_{i<j} \mathbb{E}(X_i X_j) = \sum_{i<j} \mathbb{P}(X_i = 1, X_j = 1)$

Triples (etc.): $\mathbb{E}\binom{X}{3} = \sum_{i<j<k} \mathbb{E}(X_i X_j X_k) = \sum_{i<j<k} \mathbb{P}(X_i = 1, X_j = 1, X_k = 1)$

Since $\mathbb{E}\binom{X}{2} = \frac{1}{2}\left(\mathbb{E}(X^2) - \mathbb{E}(X)\right)$ can use to compute Variance.

Moment generating functions $M_X(t) = \mathbb{E}(e^{tX}) = \sum_{n=0}^{\infty} \frac{\mathbb{E}(X^n)}{n!} t^n$

Recover raw moments: $\mathbb{E}(X^n) = M_X^{(n)}(0)$

Independence: If X, Y independent then $M_{X+Y}(t) = M_X(t) M_Y(t)$

Limit Theorems

Markov's inequality: if X is non-negative, then

$$\mathbb{P}(X \geq a) \leq \frac{1}{a}\mathbb{E}(X)$$

Chebyshev's inequality: if X has finite mean μ and variance σ^2, then

$$\mathbb{P}(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

Typically the estimates provided by these inequalities are very poor.

Weak law of large numbers Probability that average is far-away from μ goes to zero as $n \to \infty$.
If $X_1, X_2, \ldots$ are independent and identically distributed with mean μ and finite variance σ^2, then, for all $\varepsilon > 0$

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n\to\infty]{} 0$$

Weak law still true when variance is infinite, but not so useful for estimating. Can use to answer questions such as:

How many measurements/trials needed in order to be p% sure that our average is within q% of the population mean?

Central Limit Theorem Loosely: If X has mean μ and variance σ^2 and can be viewed as the sum of a large number of independent variables, then X is approximately $N(\mu, \sigma^2)$ distributed.
A simpler, more precise version: Let $X_1, X_2, \dots$ be independent identically distributed variables with mean μ and variance σ^2. Then

$$\mathbb{P}\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq a\right\} = \mathbb{P}\left\{\frac{1}{\sigma/\sqrt{n}}(\overline{X} - \mu) \leq a\right\} \underset{n\to\infty}{\longrightarrow} \Phi(a)$$

Alternatively:

- $X_1 + X_2 + \cdots + X_n$ is approximately $N(n\mu, n\sigma^2)$ for large n
- The sample mean $\overline{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$ distributed for large n.

Strong Law of Large Numbers If $X_1, \dots$ are independent identically distributed random variables then

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1$$

Intuitively: the long term frequency of an event happening given many repeated trials equals the probability of said event happening.