Chapter 5

Joint Distribution and Random Samples

5.1 Discrete joint distributions §5.1

—predict or control one random variable from another. (e.g. cases of flu last week and this week; income and education).

**Joint outcome**: The joint event

\[
\{(X, Y) = (x, y)\} = \{\omega : X(\omega) = x, Y(\omega) = y\} = \{X = x\} \cap \{Y = y\}.
\]
Range: If $X$ has possible values $x_1, \cdots, x_k$ and $Y$ has possible values $y_1, \cdots, y_l$, then the range of $(X, Y)$ is all ordered pairs

$$\{(x_i, y_j) : i = 1 \cdots k, j = 1 \cdots l\}.$$ 

Joint dist: Describes how prob is distributed over the ordered pairs:

$$p(x_i, y_j) = P(X = x_i, Y = y_j), \quad \sum_i \sum_j p(x_i, y_j) = 1,$$

called the joint probability mass function.

Example 5.1 Discrete joint distribution

Let $X =$ deductible amount, in $\$, on the auto policy and $Y =$ deductible amount on homeowner’s policy. Assume that their distribution over a population is
Table 5.1: **Joint distribution of deductible amounts**

<table>
<thead>
<tr>
<th>$p(x, y)$</th>
<th>$y$</th>
<th>0</th>
<th>100</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>$100$</td>
<td>.20</td>
<td>.10</td>
<td>.20</td>
<td></td>
</tr>
<tr>
<td>$200$</td>
<td>.05</td>
<td>.15</td>
<td>.30</td>
<td></td>
</tr>
</tbody>
</table>

(a) **(Marginal)** What is the (marginal) distribution of $X$?

$$p(X = 100) = \sum_y p(100, y) = .2 + .1 + .2 = .5.$$  
$$p(X = 200) = \sum_y p(200, y) = .05 + .15 + .3 = .5.$$  

(b) **(Marginal)** What is the marginal distribution of $Y$?

$$p(Y = y) = \sum_x P(X = x, Y = y) = \sum_x p(x, y)$$  
$$p(Y = 0) = .25, \quad P(Y = 100) = .25, \quad P(Y = 200) = .5$$

---

1 small latter indicates that the materials are not as important. They are required and assigned as reading materials.
(c) (Functions of two rv’s) Let $Z = g(X, Y)$. Then,

$$P(Z = z) = \sum_{(x,y):g(x,y) = z} p(x,y).$$

e.g. $Z = X + Y$. The range is $\{100, 200, 300, 400\}$.

$$P(Z = 100) = p(X = 100, Y = 0) = .20$$

$$P(Z = 200) = P(X = 100, Y = 100) + P(X = 200, Y = 0)$$

$$= .10 + .05 = .15.$$ 

$$P(Z = 300) = .20 + .15 = .35,$$

$$P(Z = 400) = .30$$

(d) (Expected value): $E g(X, Y) = \sum_x \sum_y g(x, y)p(x, y)$. 
(e) (Additivity): \( E(g_1(X) + g_2(Y)) = Eg_1(X) + Eg_2(Y) \). e.g.

\[
E(X + Y) = .20 \times 100 + .15 \times 200 + .35 \times 300 + .30 \times 400 = $275,
\]

while \( EX = $150 \) and \( EY = $125 \).

**Example 5.2 Multinomial distribution & machine learning**

Make \( n \) draws from a population with \( r \) categories. Let \( X_i \) be the number of “successes” in \( i \)th category. Then, its pmf is given by

\[
p(x_1, \cdots, x_r) = \binom{n}{x_1} \binom{n-x_1}{x_2} \cdots \binom{n-x_1-\cdots-x_{r-1}}{x_r} p_1^{x_1} \cdots p_r^{x_r}
\]

\[
= \frac{n!}{x_1! \cdots x_r!} p_1^{x_1} \cdots p_r^{x_r}, \quad \text{where } x_1 + \cdots + x_r = n.
\]
Q1: Are $X_i$ and $X_{i+1}$ independent?

Q2: What is the distribution if 100 people is sampled Ex 5.1 and classified according to the insurance policies.

Example 5.3 Trinomial distribution and population genetics.

In genetics, when sampling from an equilibrium population with respect to a gene with two alleles $A$ and $a$, three genotypes can be observed with proportion given below:
Table 5.2: Hardy-Weinberg formula

<table>
<thead>
<tr>
<th></th>
<th>aa</th>
<th>aA</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p_1 = \theta^2$</td>
<td>$p_2 = 2\theta(1-\theta)$</td>
<td>$p_3 = (1-\theta)^2$</td>
</tr>
<tr>
<td>$\theta = 0.5$</td>
<td>0.25</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>

**Probability question:** If $\theta = 0.5$, $n = 10$, what is the probability that $X_1 = 2$, $X_2 = 5$ and $X_3 = 3$?

$$\text{probability} = \frac{10!}{2!5!3!}(.25)^2(.5)^5(.25)^3 = 0.0769.$$  

```r
> X=rmultinom(1000,1,prob=c(0.25,0.5,0.25))  #sample 1000 from trinomial
> X[,1:10]  #first 10 realizations
[1,] 1 0 0 1 0 0 0 0 0 1
[2,] 0 1 0 0 0 1 1 1 0 0
[3,] 0 0 1 0 1 0 0 0 1 0
> apply(X,1,mean)  #calculate relative frequencies
[1] 0.265 0.492 0.243
```

**Statistical question:** If we observe $x_1 = 2$, $x_2 = 5$ and $x_3 = 3$, 

are the data consistent with the H-W formula? If so, what is $\theta$?

**Addition rule** for expectation: (without any condition)

$$E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n).$$

**Independence**: $X$ and $Y$ are independent if $P(X = x, Y = y) = P(X = x)P(Y = y)$, or $p(x, y) = p_X(x)p_Y(y)$ for all $x$ and $y$.

**Expectation of product**: If $X$ and $Y$ are independent, then $E(XY) = E(X)E(Y)$, since

$$E(XY) = \sum_x \sum_y xy p_X(x) p_Y(y) = \sum_x x p_X(x) \sum_y y p_Y(y) = EX \cdot EY.$$ 

5.2 Joint densities §5.1
Let $X$ and $Y$ be continuous random variables. Then the **joint density** function $f(x, y)$ is the one such that

$$P((X, Y) \in A) = \int \int_A f(x, y) \, dx \, dy.$$ 

It indicates the **likelihood** of getting $(x, y)$ per unit area near $(x, y)$:

$$f(x, y) \approx \frac{P(X \in x \pm \Delta x, Y \in y \pm \Delta y)}{4\Delta x \Delta y}.$$ 

**Marginal density functions** can be obtained by

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) \, dy, \quad f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) \, dx.$$ 

**Example 5.4** *Uniform distribution on a set $A$*
has a joint density
\[
    f(x, y) = I_A(x, y)/\text{Area}(A),
\]
where \( I_A(x, y) \) is the \textbf{indicator function} of the set \( A \):
\[
    I_A(x, y) = \begin{cases} 
        1, & \text{if } (x, y) \in A \\
        0, & \text{otherwise}
    \end{cases}
\]
If \( A = \{(x, y) : x^2 + y^2 \leq 1\} \) is a unit disk, then
\[
    f_X(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{1}{\pi} dy = \frac{2}{\pi} \sqrt{1-x^2}, \quad \text{for } -1 \leq x \leq 1.
\]
Similarly, \( f_Y(y) = f_X(y) \). Thus, \( X \) and \( Y \) have identical dist (but \( P(X = Y) = 0 \) in this case).

**Expectation**: \( Eg(X, Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x, y)f(x, y)dxdy \).

**Independence**: \( X \) and \( Y \) are \textbf{independent} if
\[
    f(x, y) = f_X(x)f_Y(y), \quad \text{for all } x \text{ and } y.
\]
**Rule of product**: If $X$ and $Y$ are independent, then

\[ E(XY) = E(X)E(Y) \]

since

\[ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf_X(x)f_Y(y)dx\,dy = \int_{-\infty}^{+\infty} xf_X(x)dx \int_{-\infty}^{+\infty} yf_Y(y)dy. \]

**Example 5.5 Independence and uncorrelated**

If $(X, Y)$ are uniformly distributed on the unit disk, then $X$ and $Y$ are **not independent**. But, the rule of product holds

\[ EXY = 0 = E(X)E(Y). \]

—termed uncorrelated in the next section.

—uncorrelated but not independent (independent $\implies$ uncorrelated)
Now if \((X, Y)\) is uniformly distributed on the unit square. Then,
\[
f_X(x) = \int_0^1 dy = 1, \quad \text{for } 0 \leq x \leq 1.
\]
Similarly, \(f_Y(y) = I_{[0,1]}(y)\). It is now easy to see that \(f(x, y) = I_{[0,1] \times [0,1]}(x, y) = f_X(x)f_Y(y)\), i.e. \(X\) and \(Y\) are indep. Hence,
\[
E(XY) = E(X)E(Y) = 1/4.
\]

5.3 Covariance and correlation §5.2

**Purpose:** ★Correlation measures the strength of linear associ-
Covariance helps compute \( \text{var}(X + Y) \):

\[
\text{var}(X + Y) = E(X + Y - \mu_X - \mu_Y)^2
\]

\[
= \text{var}(X) + \text{var}(Y) + 2E(X - \mu_X)(Y - \mu_Y). \\
\]

\textbf{Covariance}: \( \text{cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y) \). It is equal to

\[
\text{cov}(X, Y) = E(XY) - (EX)(EY).
\]

If \( X \) and \( Y \) are \textit{independent}, then \( \text{cov}(X, Y) = 0 \).

\textbf{Correlation}: is the covariance when standardized:

\[
\rho = E \left( \frac{X - \mu_X}{\sigma_X} \right) \times \left( \frac{Y - \mu_Y}{\sigma_Y} \right) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.
\]
When $\rho = 0$, $X$ and $Y$ are **uncorrelated**.

**Properties**: $\rho$ measures the strength of linear association.

- $-1 \leq \rho \leq 1$;

- The larger the $|\rho|$, the stronger the linear association.

- $\rho > 0$ means positive association: large $x$ tends to associate with large $y$ and small $x$ tends to associate with small $y$. $\rho < 0$ means negative association.

- $\rho$ is independent of units: $\text{corr}(1.8X + 32, 1.5Y) = \text{corr}(X, Y)$.

**Example 5.6** Bivariate normal distribution has density
Figure 5.5: Simulated data with sample correlations 0.026, -0.595, -0.990, 0.241, 0.970, 0.853.
\[ f(x, y) = \frac{1}{2\sigma_1\sigma_2\pi\sqrt{1-\rho^2}} \exp \left( -\frac{1}{2(1-\rho^2)} \left[ \frac{(x - \mu_1)^2}{\sigma_1^2} - 2\rho \frac{x - \mu_1 y - \mu_2}{\sigma_1 \sigma_2} + \frac{(y - \mu_2)^2}{\sigma_2^2} \right] \right). \]

If \( X \) and \( Y \) are standardized, then

\[ f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp \left( -\frac{1}{2(1-\rho^2)} \left[ x^2 - 2\rho xy + y^2 \right] \right). \]

It has the following properties:

★ (Marginal dist) \( X \sim N(\mu_1, \sigma_1^2) \) and \( Y \sim N(\mu_2, \sigma_2^2) \).

★ (Correlation) \( \text{corr}(X, Y) = \rho \)

★ (Linearity) \( aX + bY \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + 2ab\rho\sigma_1\sigma_2 + b^2\sigma_2^2) \).

★ (Independence) \( \rho = 0 \) if and only if \( X \) and \( Y \) are independence.

**Example 5.7 A “Normal” appointment.**

A man and a woman make an appointment at 12:00pm at the Palma Square. The arrival time of the man is normally distributed with **mean 11:55** and standard deviation **10 minutes**, and the woman’s arrival time is normally distributed with **mean 12:05** with an **SD of 8 minutes**. If
they can wait for each other for at most 15 minutes, what is the chance that they actually meet?

Let \( X \) = the man’s arrival time relative to 12:00pm and \( Y \) = the woman’s arrival time relative to 12:00pm. Then, \( X \sim N(-5, 10^2) \) and \( Y \sim N(5, 8^2) \). Let \( W = X - Y \sim N(-5 - 5, (10^2 + 8^2)) \). Then, the chance is

\[
P(|X - Y| \leq 15) = P(-15 \leq W \leq 15)
= \Phi\left(\frac{15 + 10}{12.81}\right) - \Phi\left(\frac{-15 + 10}{12.81}\right)
= 97.45\% - 34.81\% = 62.64\%.
\]

**Addition rule** for variance: If \( X_1, \cdots, X_n \) are pairwise uncorrelated, then

\[
\text{var}(X_1 + \cdots + X_n) = \text{var}(X_1) + \cdots + \text{var}(X_n).
\]

**Example 5.8** Variance for Binomial Distribution
Note $X = Y_1 + \cdots + Y_n$ w/ $Y_1, \cdots, Y_n$ are i.i.d. Bernoulli$(p)$. Then

$$\text{var}(X) = \text{var}(Y_1) + \cdots + \text{var}(Y_n) = npq.$$  

5.4 Multivariate random variables*  §5.1

**Discrete:** The joint pmf of discrete rv’s $(X_1, X_2, \cdots, X_n)$ is

$$p(x_1, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n).$$

**Continuous:** The joint pdf $f(x_1, \cdots, x_n)$ is such that

$$P(a_1 \leq X_1 \leq b_1, \cdots, a_n \leq X_n \leq b_n)$$

$$= \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \cdots, x_n) dx_1 \cdots dx_n.$$

**Independence:** $X_1, \cdots, X_n$ are said to be independent if their pmf or pdf is equal to the product of the marginal ones:

$$P(X_1 = x_1, \cdots, X_n = x_n) = P(X_1 = x_1) \cdots P(X_n = x_n).$$
for all \( x_1, \cdots, x_n \) in the ranges or

\[
f(x_1, \cdots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).
\]

**Conditional distribution**: shows the distribution of \( Y \) given the associated characteristics \((x_1, \cdots, x_p)\):

\[
p(Y = y|X_1 = x_1, \cdots, X_p = x_p) = \frac{P(Y = y, X_1 = x_1, \cdots, X_p = x_p)}{P(X_1 = x_1, \cdots, X_p = x_p)},
\]

for all \( y \) in the range. For **continuous** case,

\[
f_{Y|X}(y|x_1, \cdots, x_p) = \frac{f(y, x_1, \cdots, x_p)}{f(x_1, \cdots, x_p)}.
\]

5.5 **Square root law** \(\S 5.3\)

**A random sample** of size \( n \) means that \( X_1, \cdots, X_n \) are independent and identically distributed (\textbf{iid}) rv's, like the outcomes of draw-
ing tickets from a box with replacement. If population mean and variance are mean $\mu$ and variance $\sigma^2$, then

$$E(S_n) = n\mu \quad \text{var}(S_n) = n\sigma^2,$$

where $S_n = X_1 + \cdots + X_n$.

Figure 5.6: Outcomes of random samples
**Square root law:**

\[ SD(S_n) = \sqrt{n}\sigma, \quad SD(\bar{X}_n) = \sigma/\sqrt{n}. \]

—adding \( n \) measurement errors **inflates** the error size by \( \sqrt{n} \).

—averaging \( n \) measurement errors **decreases** the error size by \( \sqrt{n} \).

—sampling error depends on sample size \( n \), not population size.

**Example 5.9 Illustration of sampling variability.**

We take 6 random samples of size \( n = 10 \) from the income box de-
picted in Figure 5.6. The population mean is $45K with an SD $26K, which is simulated from \( X \sim \text{Gamma}(3, 15) \). Note \( \mu = EX = 3 \times 15 \) and \( \sigma = SD(X) = \sqrt{3} \times 15 \approx 26 \).

Table 5.3: An illustration of sampling variability

<table>
<thead>
<tr>
<th>Sample</th>
<th>Ave†</th>
<th>SD†</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>38.36</td>
<td>19.85</td>
</tr>
<tr>
<td>Sample 2</td>
<td>35.71</td>
<td>26.14</td>
</tr>
<tr>
<td>Sample 3</td>
<td>66.10</td>
<td>39.38</td>
</tr>
<tr>
<td>Sample 4</td>
<td>55.04</td>
<td>44.14</td>
</tr>
<tr>
<td>Sample 5</td>
<td>52.34</td>
<td>26.54</td>
</tr>
<tr>
<td>Sample 6</td>
<td>56.92</td>
<td>20.58</td>
</tr>
<tr>
<td>Average</td>
<td>50.75</td>
<td>29.44a</td>
</tr>
<tr>
<td>SD</td>
<td>11.61b</td>
<td>10.05</td>
</tr>
</tbody>
</table>

† This column should be distributed around \( \mu \). † This column should be approximate \( \sigma \).

\( a \) is the average SD of individual data and should be close to \( \sigma \).

\( b \) is the SD of 6 averages and is approximately \( \sigma/\sqrt{n} \).
**Law of large numbers**: For any given $\varepsilon > 0$, as $n \to \infty$

$$P(|\bar{X}_n - \mu| > \varepsilon) \to 0,$$

since $\text{var}(\bar{X}_n) \to 0$

— sample average converges to population average.

5.6 **Central Limit Theorem**

**Central Limit Theorem**: No matter which population is sampled from (or the content of the box), the probability distribution of the sample average (sum) follows closely normal curve, when $n$ is sufficiently large:

$$P\left\{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right\} = P\left\{\sqrt{n}(\bar{X}_n - \mu)/\sigma \leq x\right\} \approx \Phi(x).$$
This can be demonstrated by simulations (homework).

**Simple case:** $X_1, \cdots, X_n$ are i.i.d. Bernoulli($p$).

Then $S_n \sim \text{Binom}(n, p)$. Fig. 4.1 shows it follows normal.

**Normal approximation to the Binomial:** The approximation is good when $\sqrt{npq} \geq 2.5$ (an alternate criterium is $n \min(p, q) \geq 10$).

- Normal approximation *with continuity correction*:

\[
P(a \leq X \leq b) = P\left(a - 0.5 \leq X \leq b + 0.5\right) = \Phi\left(\frac{b + 0.5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{npq}}\right).
\]

Figure 5.8: Rule of continuity correction always makes interval wider.
• **without continuity correction**: no use of 0.5. □not accurate!

**Example 5.10** Let $X = \# \text{ heads in 100 tosses} \sim Bin(100, 0.5)$.

(a) Note that $EX = 50$ and $SD(X) = \sqrt{100 \times 0.5 \times 0.5} = 5$. Thus,

$$P(X = 50) = P(49.5 \leq X \leq 50.5) = P(-0.1 \leq Z \leq 0.1)$$

$$= 0.5398 - 0.4602 = 7.96\%.$$

(b) $P(X \geq 60) = P(X \geq 59.5) = 1 - P(Z \leq 9.5/5) = 2.87\%.$

(c) $P(45 \leq X \leq 60) = P(44.5 \leq X \leq 60.5) = \Phi(2.1) - \Phi(-1.1) = 83.64\%.$

**Example 5.11** Normal approximation.

In a large Statistics class of size 100, students are asked to simulate drawing 400 tickets with replacement from the box
(a) What percentage of students will get a sum of 2100 or more?

By the LLN, the percentage is approximately the probability of getting a sum $\geq 2100$.

Let $X_1, \cdots, X_{400}$ be the outcome of the 1st, 2nd, $\cdots$, 400th draw, respectively. Then, $X_1, \cdots, X_{400}$ are i.i.d. with

$$P(X_1 = 0) = \frac{1}{5}, P(X_1 = 2) = \frac{1}{5}, \cdots, P(X_1 = 10) = \frac{1}{5}.$$
We want to find
\[ p = P\{\sum_{i=1}^{400} X_i \geq 2100}\].

To use CLT, we need to compute the population mean and SD.

\[ \mu = EX_i = \frac{1}{5} \times 0 + \frac{1}{5} \times 2 + \cdots + \frac{1}{5} \times 10 = 5. \]

\[ EX_i^2 = \frac{1}{5}(0^2 + \frac{1}{5} \times 2^2 + \cdots + \times 10^2) = 38.6. \]

Hence,

\[ \text{var}(X_i) = EX_i^2 - \mu^2 = 38.6 - 5^2 = 13.6, \quad \sigma = \quad = 3.7. \]

Consequently,

\[ p = 1 - \Phi \left( \frac{2100 - 400 \times 5}{\sqrt{400 \times 3.7}} \right) = 1 - \Phi(1.35) = 9\%. \]
**Fact**: The mean and SD of an rv obtained through drawing from a box is the same as the mean and SD of the box.

**b)** What is the likely size of the error of the estimate?

(Classifying and Counting) Let $Y_i = 1$ if the $i$-th student gets a sum $\geq 2100$ and 0 otherwise. Then, $Y_1, \cdots, Y_{100}$ are i.i.d Bernoulli($p$). Let $\hat{p} = \bar{Y}$ be the proportion of students who get a sum $\geq 2100$. It follows that

$$E\hat{p} = E\bar{Y} = p = 0.09,$$

and by the square root law,

$$SD(\hat{p}) = SD(\bar{Y}) = \frac{\sqrt{pq}}{\sqrt{100}} = 2.86\%.$$ 

i.e. the estimation (chance) error is about 2.86%.
Example 5.12 Classifying and Counting

In a city, the average family income is $40K with an SD of $30K. Among those, 20% of families have income \( \geq 80K \). A survey organization takes a random sample of 900 families.

(a) What is the chance that the sample average falls between $38K and $42K?

Let \( \bar{X} \) be the sample average. By the sqrt law,

\[
E \bar{X} = 40, \quad SD(\bar{X}) = \frac{30}{\sqrt{900}} = 1.
\]

Hence, by CLT

\[
P(38 \leq \bar{X} \leq 42) = \Phi\left(\frac{42 - 40}{1}\right) - \Phi\left(\frac{38 - 40}{1}\right) = \Phi(2) - \Phi(-2) \approx 95%,
\]

— very likely

(b) What is the chance that between 18% and 22% of the selected families have an income \( \geq 80K \) (i.e. poll error \( \leq 2\% \))?

Let \( Y_i = 1 \) if the i-th draw has income \( \geq 80K \). Then, \( \hat{p} = \bar{Y} \) is the proportion of selected families having income \( \geq 80K \).

\[
prob = P\{.18 \leq \hat{p} \leq .22\} = P\{162 \leq \sum_{i=1}^{900} Y_i \leq 198\}.
\]

Now, \( EY_i = 0.2 \) and \( SD(Y_i) = \sqrt{0.2 \times 0.8} = 0.4 \). Hence,

\[
ES_{900} = 900 \times .2 = 180, \quad SD(S_{900}) = \sqrt{900 \times .2 \times .8} = 12.
\]
Using normal approximation with continuity correction,

\[
prob = \Phi \left( \frac{198.5 - 180}{12} \right) - \Phi \left( \frac{161.5 - 180}{12} \right) = 87.68\%.
\]

This method is preferable (more accurate) for this specific problem.

**Conclusion:**

★ Chance errors or poll errors depend on sample size \( n \), not the population size \( N \).

★ Random sampling has advantages of avoiding biases and calculating chance errors (variance).

★ Arbitrary sample is NOT a random sample. It creates unintentional biases and is unable to figure out errors.

★ Statistical lessons are to avoid biases.