Slide 1

Reading Assignment: Chapter 1

- **Experiment**: phenomenon where outcomes are uncertain – e.g., single throws of a six-sided die.
- **Sample space**: set of all outcomes of the experiment – $\mathcal{S} = \{1, 2, 3, 4, 5, 6\}$.
- **Event**: a subset of $\mathcal{S}$ – e.g., $A = \{3\},\ B = \{3, 4, 5, 6\}$.

Slide 2

Basic properties of probability

- If S is the sample space, $P(S) = 1$.
- For any event A, $0 \leq P(A) \leq 1$.
- For any complementary events A and A^c,

$$P(A^c) = 1 - P(A) \qquad P(\emptyset) = 1 - P(\mathcal{S}) = 0$$

- For any two events A and B, the probability that either A or B will occur is given by

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- The conditional probability of A given B for any two sets A and B is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad \text{if } P(B) \neq 0$$

Slide 3

- A and B are independent if

$$P(A|B) = P(A) \quad \text{or equivalently} \quad P(A \cap B) = P(A)P(B)$$

- **Law of total probability**
  Let $A_1, \ldots, A_n$ be a set of events such that $\bigcup_{i=1}^n A_i = S$ and $A_i \cap A_j = \emptyset$ for $i \neq j$ and $P(A_i) > 0$ for all i. Then, for any event B,

$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i).$$

- **Bayes Theorem**
  Let B and $A_1, \ldots, A_n$ be events where $\bigcup_{i=1}^n A_i = S$, $A_i \cap A_j = \emptyset$ for $i \neq j$, and $P(A_i) > 0$, $\forall i$. Then

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)}.$$

Slide 4

Example 1: Diagnostic test

Suppose that only 0.1% of the population have the disease and there is a fairly accurate diagnostic test

- the test accurately reports a ‘positive’ result 99% of the time, $p(T|D) = .99$
- the test accurately reports a ‘negative’ result 95% of the time, $p(T^c|D^c) = .95$

If an individual gets a positive test result, what is the probability that it is a false positive, $p(D^c|T)$?

Slide 5

- Despite the apparent high accuracy of the test, the incidence of the disease is so low that the vast majority of patients who test positive do not have the disease.
- Nonetheless, this is 20 times the proportion before we knew the outcome of the test! The test is not useless, and retesting may improve the reliability of the result.

Slide 6

Example 2: Paternity dispute

- Suppose you are on a jury considering a paternity suit brought by Suzy Smith's mother against Al Edged.
- Suzy's mother has blood type O and Al Edged is type AB.
- You have other information as well. You hear testimony concerning whether Al Edged and Suzy's mother had sexual intercourse during the time that conception could have occurred, about the timing and frequency of such intercourse, about Al Edged fertility, about the possibility that someone else is the father, and so on. You put all this information together in assessing $P(F)$, your probability that Al is Suzy's father.
- The evidence of interest is Suzy's blood type, which turns out to be B (if it were O, Al Edged would be excluded from paternity).

Slide 7

- According to Mendelian genetics, $P(B|F) = \frac{1}{2}$.
- You also accept the blood bank's estimate $P(B|F^c) = 0.09$.
- According to Bayes' rule

$$P(F|B) = \frac{P(B|F)P(F)}{P(B|F)P(F) + P(B|F^c)P(F^c)}$$

The relationship between our prior probability, $P(F)$, and our posterior probability, $P(F|B)$ may be summarized:

$P(F)$	0	0.100	0.250	0.500	0.750	0.900	1
$P(F\|B)$	0	0.382	0.649	0.847	0.943	0.980	1

Slide 8

Discrete Random Variables

- For discrete random variables, the set of possible values is either finite or countably infinite.
- Associated with each discrete random variable Y, with possible values $y_1, y_2, \ldots$, there is a **probability mass function**, $p(y_i) = P(Y = y_i)$, such that

$$p(y_i) \geq 0 \quad \text{and} \quad \sum_i p(y_i) = 1.$$

- Associated with every random variable is a **cumulative distribution function** (**cdf**),

$$F(y) = P(Y \leq y), \ -\infty < y < \infty.$$

The cdf is non-decreasing and satisfies

$$\lim_{y \to -\infty} F(y) = 0 \ \text{ and } \ \lim_{y \to \infty} F(y) = 1.$$

Slide 9

If Y is a discrete random variable, $F(y)$ is a step function, with jumps occurring at the values of y for which $p(y) > 0$.

- The mean or **expected value** of a discrete random variable Y with probability mass function $p(y)$ is given by

$$\mu = E(Y) = \sum_i y_i p(y_i)$$

Slide 10

Continuous random variable

- For continuous random variables, the set of possible values is uncountable.
- The **probability density function** (**pdf**), $f(y)$, of a continuous random variable, Y, with support $\mathcal{S}$ is an integrable function such that:

(a)

$$f(y) > 0,\ \text{if } y \in \mathcal{S} \qquad f(y) = 0,\ \text{if } y \notin \mathcal{S}$$

(b)

$$\int_S f(y)dy = 1$$

(c)

$$P(a \leq Y \leq b) = \int_a^b f(y)dy.$$

Slide 11

- For a continuous random variable, the **cumulative distribution function (cdf)** is given by

$$P(Y \leq a) = F(a) = \int_{-\infty}^{a} f(y)dy.$$

The cdf of a continuous random variable, $F(y)$, is continuous and monotonically non-decreasing.

- If $a < b$, we obtain

$$P(a \leq Y \leq b) = \int_{a}^{b} f(y)dy = F(b) - F(a).$$

- The mean or **expected value** of a continuous random variable Y with pdf $f(y)$

$$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y)dy$$

Slide 12

Joint distributions

Discrete random variables:

- Suppose that X and Y are two discrete random variables defined on the same sample space S and taking on values $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$, respectively. Their **joint probability mass function**, $p_{X,Y}(x, y)$, is

$$p_{X,Y}(x_i, y_j) = P(X = x_i, Y = y_j)$$

- The **marginal probability mass function** of one random variable is obtained from the joint frequency distribution by

$$p_X(x) = \sum_j p_{X,Y}(x, y_j) \qquad p_Y(y) = \sum_i p_{X,Y}(x_i, y).$$

Slide 13

- The conditional probability that $X = x_i$, given that $Y = y_j$ is,

$$p_{X|Y}(x|y) = P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{X,Y}(x_i, y_j)}{p_Y(y_j)}.$$

This can be re-expressed as

$$p_{X,Y}(x, y) = p_{X|Y}(x|y)p_Y(y)$$

Summing both sides over all values of y, we get a very useful application of the law of total probability:

$$p_X(x) = \sum_y p_{X|Y}(x|y)p_Y(y).$$

Slide 14

Continuous random variables:

- Suppose that X and Y are two continuous random variables. Their **joint probability density function**, $f(x, y)$, is the surface such that for any region A in the xy-plane,

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

- The **marginal probability density function** of one random variable is obtained from the joint pdf

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx$$

- The conditional density functions of Y given X is defined to be

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)}, \quad \text{if } 0 < f_X(x) < \infty$$

Slide 15

The joint density can be expressed in terms of the marginal and conditional densities as:

$$f_{XY}(x, y) = f_{Y|X}(y|x) f_X(x)$$

Integrating both sides over x allows the marginal density of Y to be expressed as

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx,$$

which is the law of total probability for the continuous case.

Slide 16

Exchangeability and de Finetti's theorem

- Let $p(y_1, \ldots, y_n)$ be the joint distribution of $Y_1, \ldots, Y_n$ and let $\pi_1, \ldots, \pi_n$ be a permutation of the indices $1, \ldots, n$.
- If $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations, then $Y_1, \ldots, Y_n$ are **exchangeable**.
- **de Finetti's Theorem:**
  Let $Y_1, Y_2, \ldots$ be a sequence of random variables. If for any n, $Y_1, \ldots, Y_n$ are exchangeable, then there exists a prior distribution $p(\theta)$ and sampling model $p(y|\theta)$ such that

$$p(y_1, \ldots, y_n) = \int_{\Theta} \left\{ \prod_{1}^{n} p(y_i|\theta) \right\} p(\theta) d\theta$$

Slide 17

$$\left.\begin{array}{lcl} Y_1,\ldots,Y_n|\theta & \overset{\text{iid}}{\sim} & p(y|\theta) \\ \theta & \sim & p(\theta) \end{array}\right\} \Leftrightarrow \; Y_1,\ldots,Y_n \text{ are exchangeable.}$$

This is applicable if $Y_1,\ldots,Y_n$ are

- outcomes of a repeatable experiment
- sampled from an infinite population without replacement
- sampled from a finite population of size $N \gg n$ without replacement

Labels carry no information.

Slide 18

What would not be an exchangeable sequence?

- Consider the case of "streaks" in sports, where a team that has just won its previous game is more likely to win the next, and conversely, a team that has just lost a game is more likely to loose the next.
- In this case,

  $$p(1,1,1,0,0,0)$$

  would not be believed to equal

  $$p(1,0,1,0,1,0)$$

  and thus the joint probability is not preserved under permutation.
- Thus, the sequence would not be regarded as exchangeable.

Slide 19

Example: Estimating probability of rare event

- Suppose we are interested in the prevalence of an infectious disease in a small city. The higher the prevalence, the more public health precautions would need to be put into place.
- A small random sample of 20 individuals from the city is checked for infection.
- Interest is in θ, the fraction of infected individuals in the city.
- The data y records the total number of people in the sample who are infected.
- The parameter space and sample space are then as follows:

$$\Theta = [0, 1] \qquad\qquad \mathcal{Y} = \{0, 1, \ldots, 20\}$$

Slide 20

Sampling model

- Before the sample is obtained the number of infected individuals in the sample is unknown.
- If the value of θ were known, a reasonable sampling model for Y would be a binomial$(20, \theta)$ probability distribution:

$$Y|\theta \sim \text{ binomial}(20, \theta)$$

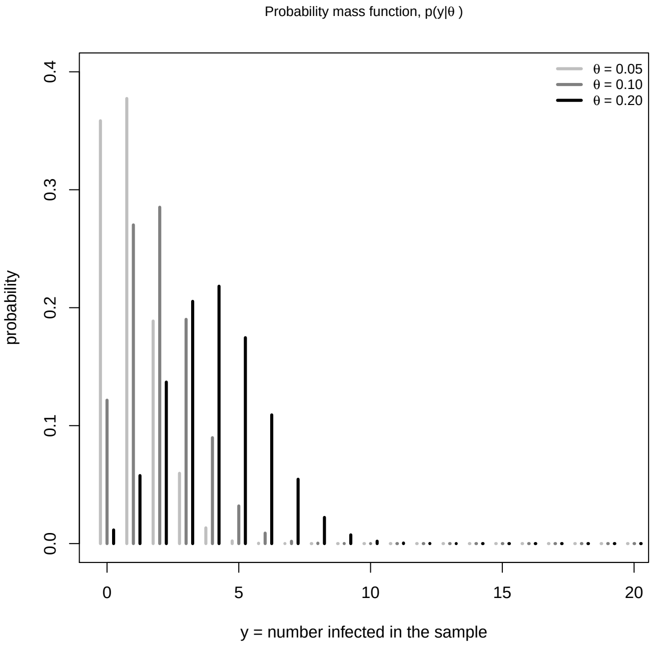
- If, for example, the true infection rate is $\theta = 0.05$, then the probability that there will be 0 infected individuals in the sample is $P(Y = 0) = \binom{20}{0} 0.05^0 (1 - 0.05)^{20} = 0.36$.

Slide 21

```
n = 20 ; x = 0:n
del = .25
plot( range(x-del), c(0,.4), xlab="y = number infected in the sample",
     ylab= "probability", type="n", cex.main = 0.8,
     main = expression(paste("Probability mass function, p(y|", theta, " )")))

points( x-del, dbinom(x,n,.05), type="h", col=gray(.75), lwd=3)
points( x, dbinom(x,n,.10), type="h", col=gray(.5), lwd=3)
points( x+del, dbinom(x,n,.20), type="h", col=gray(0), lwd=3)
legend("topright", legend=c(expression(paste(theta, " = 0.05",sep="")),
    expression(paste(theta, " = 0.10",sep="")),
    expression(paste(theta, " = 0.20",sep="")) ),
     lwd=c(3,3,3), col=gray(c(.75,.5,0)), bty="n", cex=0.8)
```

Slide 22



Slide 23

Prior distribution

- Other studies from various parts of the country indicate that the infection rate in comparable cities ranges from about 0.05 to 0.20, with an average prevalence of 0.10.
- We will use a prior distribution $p(\theta)$ that has these characteristics and provides computational convenience.
- Specifically, we will use a beta distribution

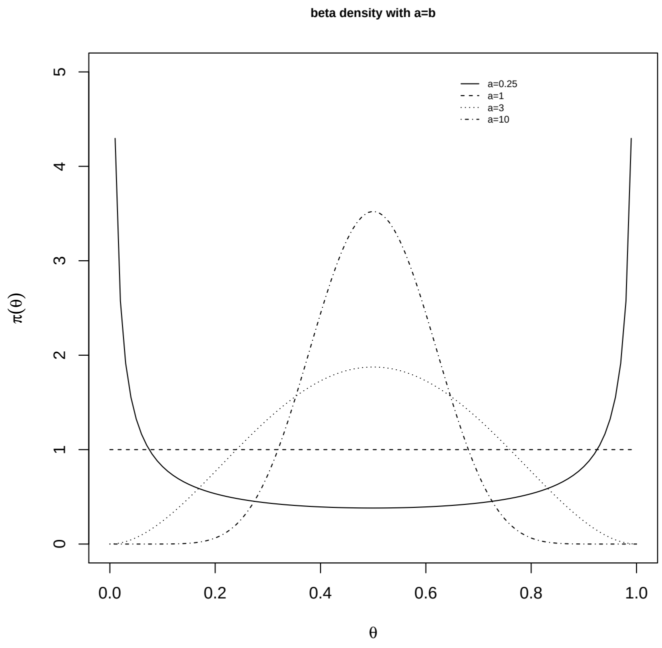
$$\theta \sim \text{ beta}(a,b) \qquad p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 \le \theta \le 1$$

- The expectation $E[\theta] = a/(a+b)$ and the mode of θ is $(a-1)/(a+b-2)$.

Slide 24

```
theta = seq(0,1,0.01)
plot(theta,  dbeta(theta, 0.25, 0.25), type="l", xlim=c(0,1), ylim=c(0,5),
    xlab=expression(paste(theta)), ylab=expression(paste(pi(theta))),
    main="beta density with a=b", cex.main=0.75)
lines(theta, dbeta(theta, 1, 1), lty=2)
lines(theta, dbeta(theta, 3, 3), lty=3)
lines(theta, dbeta(theta, 10, 10), lty=4)
legend("topright", c("a=0.25", "a=1", "a=3", "a=10"),
    lty=c(1,2,3,4), cex=0.6, bty="n")
```

Slide 25



Slide 26

- We will represent our prior information about θ with a beta$(2, 20)$ probability distribution

$$\theta \sim \text{ beta}(2, 20).$$

$$\begin{aligned} E[\theta] &= 0.09 \qquad & P(0.05 < \theta < 0.20) = 0.66 \\ \text{mode}[\theta] &= 0.05 \qquad & P(\theta < 0.10) = 0.635 \end{aligned}$$

```
a=2 ; b<-20
a/(a+b)
(a-1)/(a-1+b-1)
pbeta(.20,a,b) - pbeta(.05,a,b)
pbeta(.10,a,b)
curve(dbeta(x, 2, 20))
```

Slide 27

Posterior distribution

Suppose for our study, a value of $Y = y$ is observed. The posterior distribution of θ is given by

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta})d\tilde{\theta}}$$

In deriving posterior densities, an often used technique is to try and recognize the kernel of the posterior density of θ

$$p(\theta|y) \propto \theta^{y+a-1}(1-\theta)^{n-y+b-1}$$

We recognize this as a beta kernel with parameters $(y + a, n - y + b)$. Thus

$$\begin{aligned} \theta|y &\sim \text{beta}(y+a, n-y+b) \\ p(\theta|y) &= \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1} \end{aligned}$$

Slide 28

If we had to compute the marginal distribution, $p(y)$

$$\begin{aligned} p(y) &= \int_0^1 \binom{n}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}d\theta \\ &= \binom{n}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)} \\ &\quad \int_0^1 \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}d\theta \\ &= \binom{n}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)} \end{aligned}$$

which leads to

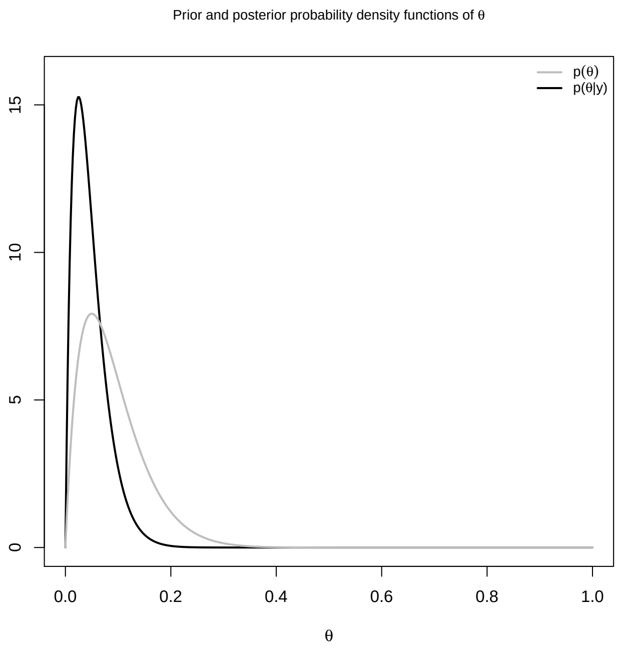
$$p(\theta|y) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}$$

Slide 29

```
a=2 ; b=20
y=0 ; n=20
theta=seq(0,1,length=500)

plot(theta, dbeta(theta,a+y,b+n-y), type="l",
    main=expression(paste("Prior and posterior probability density functions of ", theta)),
    xlab=expression(paste(theta), "percentage infected in the population"),
    ylab="", lwd=2, ylim=c(0,16), cex.main=0.8)
lines(theta, dbeta(theta,a,b), col="gray", lwd=2)
legend("topright", legend=c(expression(paste(p(theta))),
    expression(paste(p,"(",theta, "|",  y,")")) ),
    bty="n", lwd=c(2,2), col=c("gray","black"), cex=0.8)
```

Slide 30



Slide 31

- Suppose that for our study a value of $Y = 0$ is observed, i.e., none of the sample individuals are infected.
- The posterior distribution of θ is then

$$\theta | Y = 0 \sim \text{ beta}(2, 40)$$

- This density is further to the left than the prior distribution, and more peaked as well.

$$E[\theta|Y = 0] = 0.048 \quad \text{mode}[\theta|Y = 0] = 0.025 \quad P(\theta < 0.10|Y = 0) = 0.93$$

```
(a+y)/(b+n-y)
(a+y-1)/(a+y-1+b+n-y-1)
pbeta(.20,a+y,b+n-y) - pbeta(.05,a+y,b+n-y)
pbeta(.10,a+y,b+n-y)
```

The posterior distribution $p(\theta|Y = 0)$ provides us with a model for learning about the city-wide infection rate θ.

Slide 32

Sensitivity analysis

- Suppose we are supposed to discuss the results of the survey with a group of city health officials.
- We might want to present the posterior results corresponding to a variety of prior distributions.
- The posterior expectation is a weighted average of the sample mean $\bar{y}$ and the prior expectation θ_0

$$\begin{aligned} E[\theta|Y = y] &= \frac{a+y}{a+b+n} = \frac{n}{a+b+n}\frac{y}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b} \\ &= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0 \end{aligned}$$

where $\theta_0 = a/(a + b)$ is the prior expectation of θ and $w = a + b$

Slide 33

- θ_0 represents our prior guess at the true value of θ and w represents our confidence in this guess, expressed on the same scale as the sample size.
- We can compute such a posterior distribution for a wide range of θ_0 and w values to perform a **sensitivity analysis** – an exploration of how posterior information is affected by differences in prior opinion.

Slide 34

Comparison to frequentist method

- A standard estimate of a population proportion θ is the sample proportion $\hat{\theta} = y/n$, the fraction of infected people in the sample.
- For our sample in which $y = 0$, $\hat{\theta} = 0$.
- We probably want to include the caveat that this estimate is subject to sampling uncertainty and provide a $(1 - \alpha)100\%$ confidence interval

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\hat{\theta}(1 - \hat{\theta})/n}$$

- For our sample in which $\hat{\theta} = 0$, the confidence interval comes out to be just a single point: 0 (in addition, the asymptotic frequentist coverage may not hold with only $n = 20$).

Slide 35

- An "adjusted" confidence interval suggested by Agresti and Coull (1998) is given by

  $$\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n} \quad \text{where} \quad \hat{\theta} = \frac{n}{n+4} \cdot \bar{y} + \frac{4}{n+4} \cdot \frac{1}{2}$$

  While not originally motivated as such, this interval is related to Bayesian inference: the value of $\hat{\theta}$ here is equivalent to the posterior mean for θ under a beta$(2, 2)$ prior distribution, which represents weak prior information centered around $\theta = 1/2$.