Module 4: Probability

Probability concepts in statistical inference

Probability is a way of quantifying uncertainty associated with **random** events and is the basis for *statistical inference*.

Inference is the generalization of findings from a *sample* or an experiment to a population.

- In a *sample* of 1028 adults, 11% were found to approve of the way Congress was handling its job. How much certainty (or confidence) do we have in saying that the true proportion is close to 11%?
- Suppose an *experiment* is carried out to determine if taking an antidepressant will help an individual quit smoking. This study finds that at the end of 1 year, 55% of subjects receiving an antidepressant were not smoking, compared with 42.3% in the placebo (control) group. Can this difference (55% vs. 42.3%) be explained by *chance*?

Probability as a measure of long-run behavior

Suppose we rolled a die 100 times. What proportion of times would you expect to roll a 6?

What if we rolled the die 1000 times? 10,000 times?

Let's try this experiment in *R*!

The following should be clear from the die example:

With random events, the proportion of times something happens is random and variable in the short term but predictable in the long run. The probability of rolling a die and getting a 6 is $1/6 \approx 0.167$.

Probability

With a randomized experiment or a random sample or other random phenomenon, the **probability** of a particular outcome is the proportion of times that the outcome would occur in a long run of observations. This is also an example of the **law of large numbers**.

This definition of probability is sometimes referred to as the *empirical probability*.

For the definition on the previous page to hold, each trial (e.g., roll of the die) must be independent of each other.

Independent trials

Different trials of a random phenomenon are **independent** if the outcome of any one trial is not affected by or correlated with the outcome of any other trial.

Many events are independent but it can be 'human nature' to think that they are not. The following are examples of independent events

rolling a die flipping a coin having children

For example, if you roll a die and the number 5 has not come up in 100 rolls, the probability that you roll a 5 on the next roll is still 1/6.

Classical Probabilities

We saw that the probability of rolling a 6 on a fair die is 1/6, based on its long run proportion. However, we can also determine this probability by saying that out of the 6 possible outcomes (rollling a 1-6), there is only a single way to roll a 6. We will talk about this approach in more detail.

Sample Space

For a random phenomenon, the **sample space** is the set of all possible outcomes.

We will look at some examples. To help determine the sample space, it is useful to note that if there are x possible outcomes for each trial, and there are n trials, the sample space consists of x^n outcomes.

Example

State the sample space for the following probability experiments:

- Flipping a coin once
- Flipping a coin twice
- Correct answers on a 3 question test

Event

An **event** is a subset of the sample space and corresponds to a particular outcome or a group of possible outcomes. Events are often denoted with capital letters or by a string of letters that describe the event.

For example,

$A =$ student answers all three questions correctly $=$

$B =$ student answers at least 2 questions correctly $=$

Probability of an Event (classical definition)

The probability of an event A, denoted by $P(A)$, is obtained by adding the probabilities of the individual outcomes in the event.

When all possible outcomes are equally likely,

$$P(A) = \frac{\text{number of outcomes in the event } A}{\text{number of outcomes in the sample space}}$$

Probability characteristics for any event:

- a probability must be between 0 and 1.
- if S is the sample space, then $P(S) = 1$.
- a probability of 0 means the event is *impossible*
- a probability of 1 means the event is *a certainty*

Example

Find the probability of flipping a coin 3 times and

- getting all heads

- getting at least 2 heads

The Complement of an Event

The **complement** of an event A consists of all outcomes in the sample space that are *not* in A. It is denoted by A^C. The probabilities of A and A^C add to 1, so

$$P(A^C) = 1 - P(A)$$

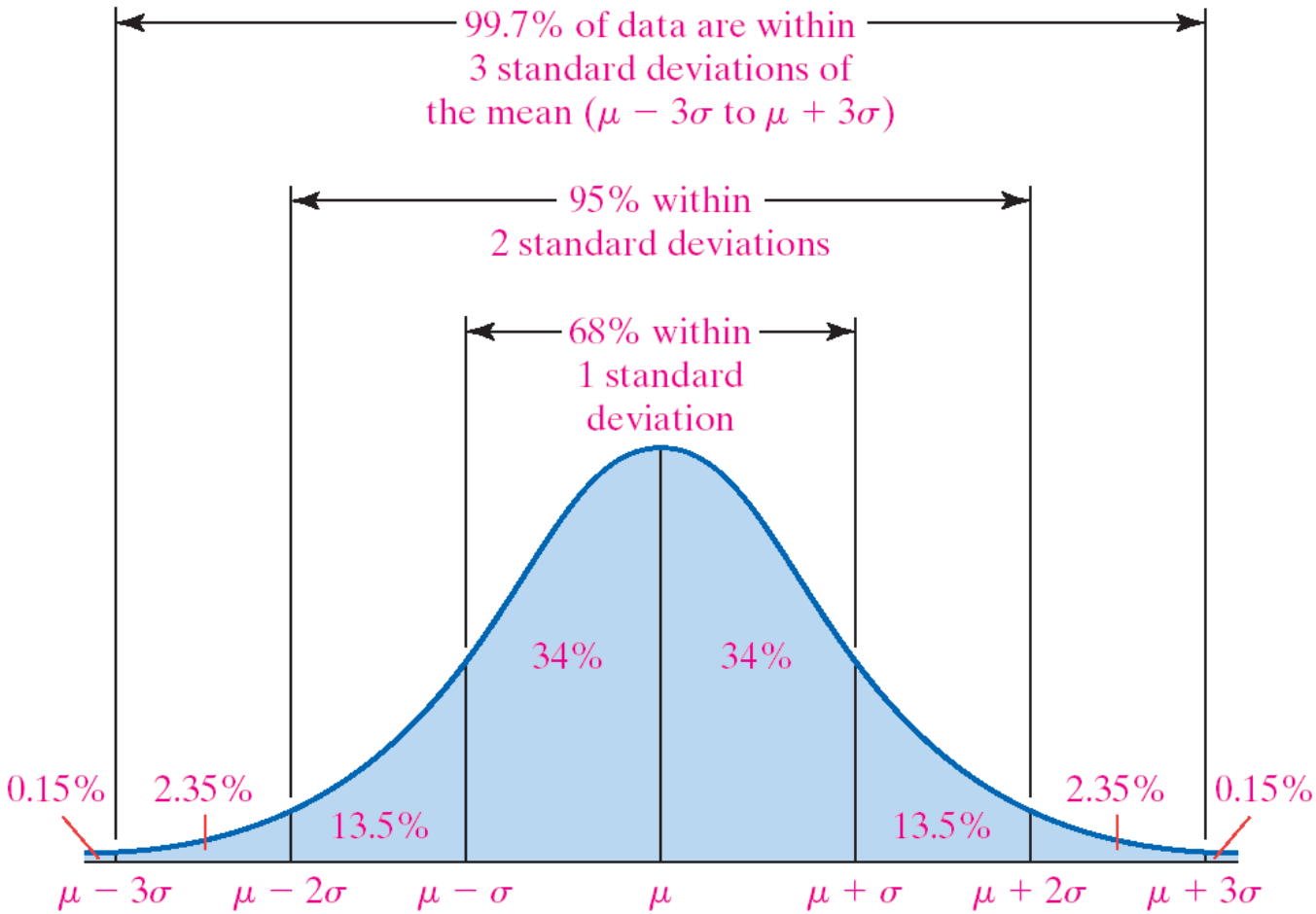
Example

We previously found that if you flip a coin 3 times, the probability of the event $A =$ getting at least one head $= 7/8$. Therefore, the probability of getting no heads (or all tails) is $P(A^C) = 1 - P(A) = 1 - 7/8 = 1/8$

Probabilities for Bell-Shaped (Normal) Distributions

Normal distribution

The **normal distribution** is symmetric, bell-shaped, and characterized by its mean μ and standard deviation σ



Finding probabilities for a normal random variable

Suppose that X is a random variable that is normally distributed with mean μ and standard deviation σ. Then $X \sim N(\mu, \sigma)$

- The total area under the curve of the normal distribution is 1.0
- The area between two values, a and b, is the probability that X is between a and b.
- The area to the left of a is the probability that X is less than a.
- The area to the right of b is the probability that X is greater than b.

We will use *R* to calculate probabilities of normally distributed random variables

Standard normal distribution

A random variable Z follows the standard normal distribution if it is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$. Then $Z \sim N(0, 1)$.

If z is an observation from the standard normal distribution, then z is $\frac{z-\mu}{\sigma} = \frac{z-0}{1} = z$ standard deviations above the mean.

The standard normal distribution can therefore be used to calculate probabilities of observations regarding standard deviations from the mean.

Sampling distributions

How Sample Means Vary Around the Population Mean

Consider a very small *population* of students in a class, whose ages and gender are given in the table below.

Person	**Gender**	**Age**
A	F	21
B	M	19
C	F	21
D	F	21
E	M	18

What is the population mean μ?

Let's take a simple random sample of $n = 3$ students from this class and calculate $\bar{X}$ = the sample mean. Note that this is a random variable and therefore has a probability distribution. Let's find the probability distribution as well as its mean, or expected value, $\mathrm{E}[\bar{X}]$.

Find the probability distribution and expected value of $\bar{X}$ when $n = 3$.

How does the expected value of the sample mean $\bar{X}$ compare to the population mean μ?

Mean and Standard Deviation of the Sampling Distribution of the sampling distribution $\bar{X}$.

For a random sample of size n from a population having mean μ and standard deviation σ, the sampling distribution of the sample mean $\bar{x}$ has expected value equal to the population mean μ and a standard deviation of $\frac{\sigma}{\sqrt{n}}$. Therefore, as n increases the expected value of the sample mean gets closer and closer to the population mean, μ.

What about the *shape* of the distribution? If the population is normally distributed, then $\bar{X}_n$ is always normally distributed.

Amazingly, this is approximately the case regardless of the distribution of the population.

The Central Limit Theorem (CLT)

For a random sample of size n from a population having mean μ and standard deviation σ, and *any distribution (shape)* then as the sample size n increases, the sample distribution of the sample mean $\bar{X}_n$ approaches an approximately normal distribution. In other words, it is always approximately true that

$$\bar{X}_n \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

In practice, the CLT holds when either the population is normally distributed or when $n > 30$.

Example

Additional examples of the Central Limit Theorem

http://www.chem.uoa.gr/applets/appletcentrallimit/appl_centrallimit2.html

The Central Limit Theorem Helps Us Make Inferences

Remember the Empirical Rule? If a distribution is approximately bell-shaped (normal), then the percent of observations falling between one, two, and three standard deviations is approximately 68%, 95%, and 99.7%, respectively.

Therefore if the sampling distribution of $\bar{X}_n$ is (approximately) normal, then the sample mean $\bar{x}$ falls within 1 standard deviation of the population mean μ 68% of the time, falls within 2 standard deviations of the population mean 95% of the time, and falls within 3 standard deviations of the population mean almost all of the time.