Probability and Probability Distributions

Lecture 2

Probability

- Probability underlies statistical inference - the drawing of conclusions from a sample of data
- If samples are drawn at random, their characteristics (such as the sample mean) depend upon chance
- Hence to understand how to interpret sample evidence, we need to understand chance, or probability

Definition of Probability

- The probability of an event *A* may be defined in different ways:
  - The frequentist view: the proportion of trials in which the event occurs, calculated as the number of trials approaches infinity
  - The subjective view: someone's degree of belief about the likelihood of an event occurring

Probabilities

- With each outcome in the sample space we can associate a probability
- Example: Toss a coin
  - Pr(Head) = 1/2
  - Pr(Tail) = ½
- This is an example of a probability distribution

Rules for Probabilities

- $0 \leq \Pr(A) \leq 1$
- $\sum p = 1$, or 100%, summed over all outcomes
- Pr(not-*A*) = 1 - Pr(*A*)

Probability Distribution

- We extend the probability analysis by considering random variables (usually the outcome of a probability experiment)
- These (usually) have a known probability distribution
- Once we work out the relevant distribution, solving the problem is usually straightforward

Random Variables

- Most statistics (e.g. the sample mean) are random variables
- Many random variables have well-known probability distributions associated with them
- To understand random variables, we need to know about probability distributions

Some Standard Probability Distributions

- Binomial distribution
- **Normal distribution**

  and the t-distribution
- Poisson distribution

When do They Arise?

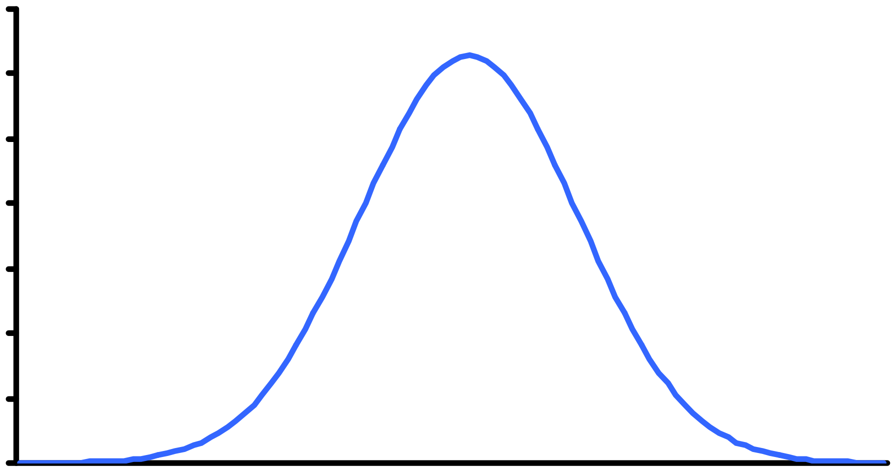
- Binomial - when the underlying probability experiment has only two possible outcomes (e.g. tossing a coin)
- Normal - when many small independent factors influence a variable (e.g. IQ, influenced by genes, diet, etc.)
- Poisson - for rare events, when the probability of occurrence is low

The Normal Distribution

- Examples of Normally distributed variables:
  - IQ
  - Heights
  - the sample mean
  - some transformations of variables: e.g. natural logarithm of income is often normal

The Normal Distribution (cont.)

- The Normal distribution is
  - bell shaped
  - Symmetric
  - Unimodal
  - and extends from $x = -\infty$ to $+\infty$ (in theory)



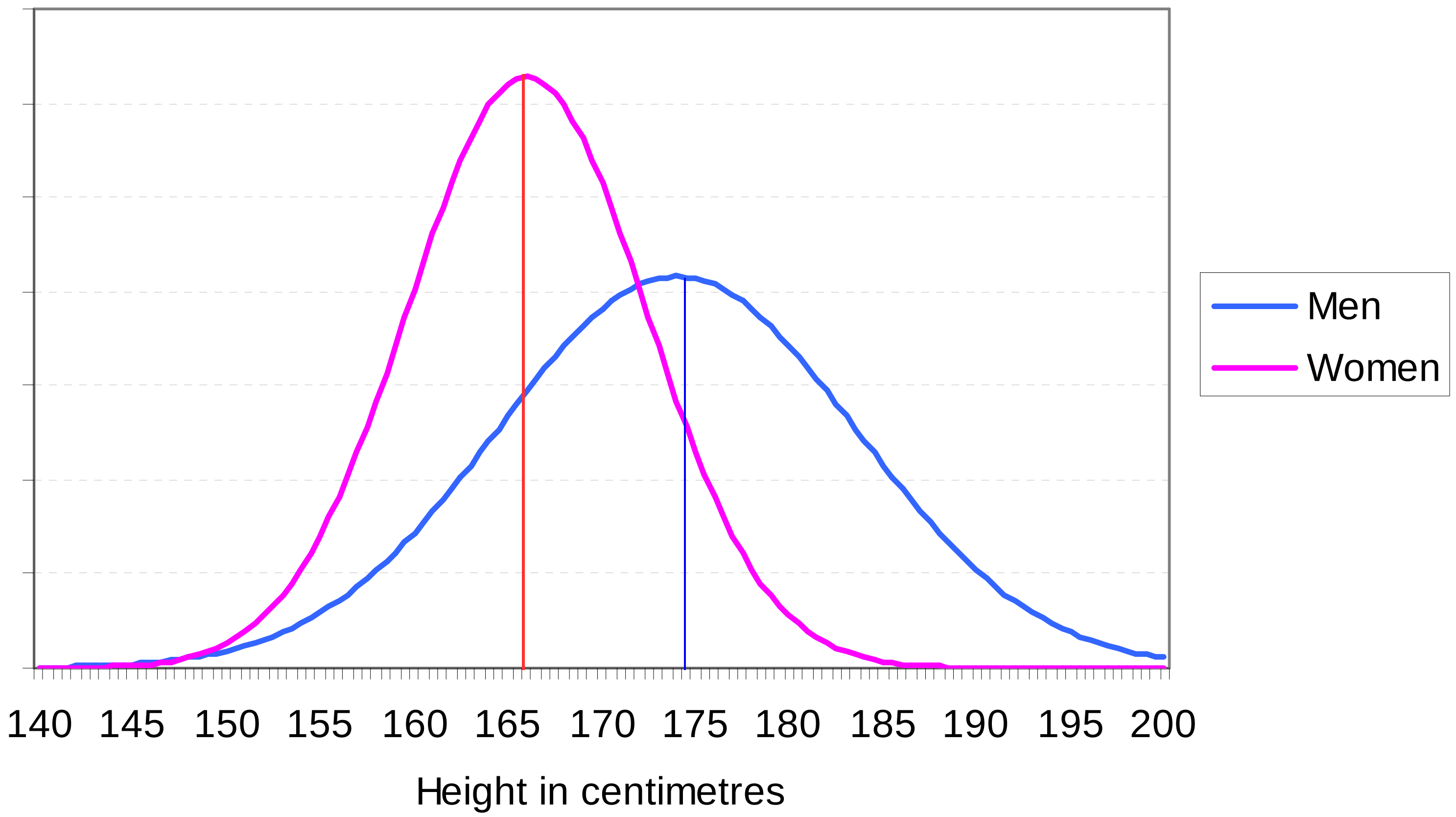
Parameters of the Distribution

- The two parameters of the Normal distribution are the mean μ and the variance σ^2
  - $x \sim N(\mu, \sigma^2)$

  e.g. Men's heights are Normally distributed with mean 174 cm and variance 92.16
  - $x_M \sim N(174, 92.16)$

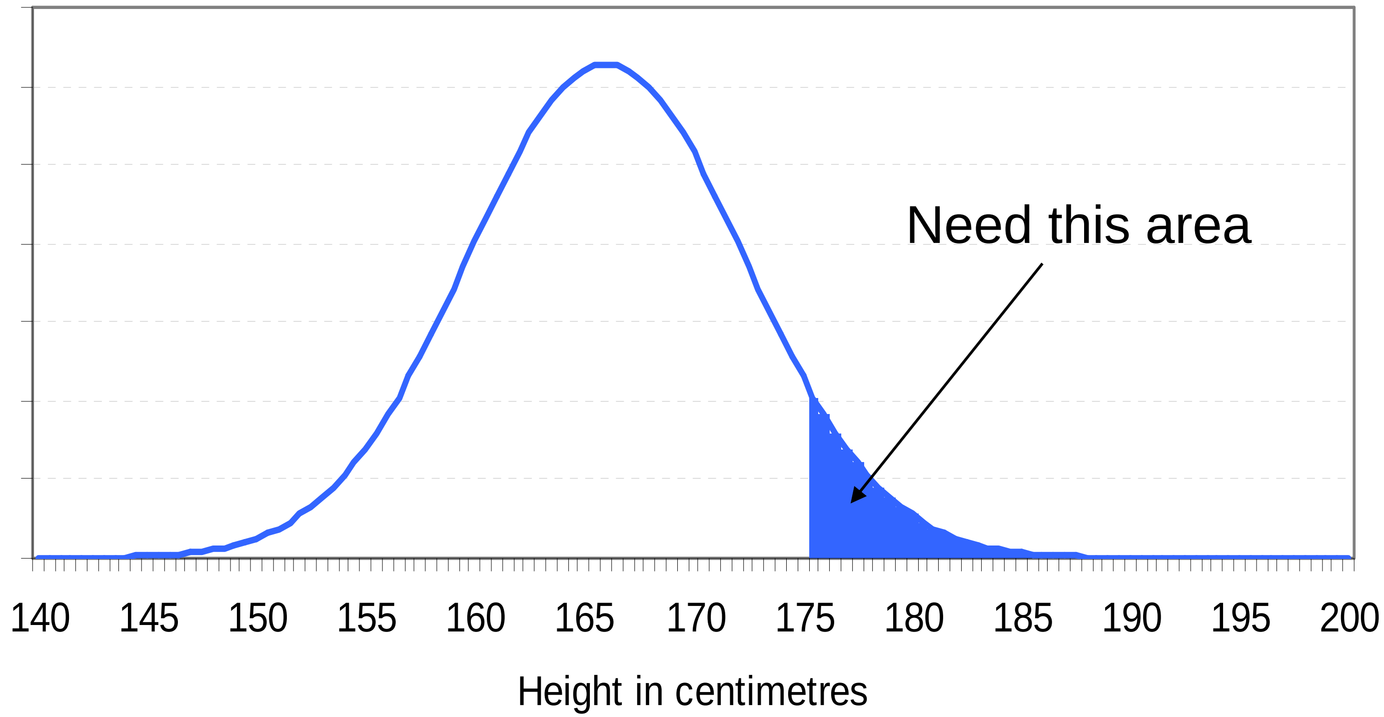
  e.g. Women's heights are Normally distributed with a mean of 166 cm and variance 40.32
  - $x_W \sim N(166, 40.32)$

Graph of Men's and Women's Heights



Areas Under the Distribution

- What is the proportion of women that are taller than 175 cm?



Areas Under the Distribution (cont.)

- How many standard deviations is 175 above 166?
- One standard deviation is $\sqrt{40.32} = 6.35$, hence

$$z = \frac{175 - 166}{6.35} = 1.42$$

- So 175 lies 1.42 standard deviations above the mean
- How much of the Normal distribution lies beyond 1.42 s.d's above the mean? Use tables...

Table A2 The Standard Normal Distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	...
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

Answer

- 7.78% of women are taller than 175 cm.
- To find the area in the tail of the distribution:

1. Calculate the *z*-score, given the number of standard deviations between the mean and the desired height
2. Then look the z-score up in tables to get a probability
3. Use rules of symmetry where appropriate

The Distribution of the Sample Mean

- If samples of size *n* are randomly drawn from a Normally distributed population of mean μ and variance σ^2, the *sample mean* is distributed as

$$\bar{x} \sim N\left(\mu, \sigma^2/n\right)$$

- E.g. if samples of 50 women are chosen, the *sample mean* is distributed

$$\bar{x} \sim N\left(166, 40.32/50\right)$$

  - note the very small standard error: $\sqrt{(40.32/50)}=0.897$

The Distributions of *x* and of $\bar{x}$

- Note the distinction between

$$x \sim N\left(\mu, \sigma^2\right)$$

and

$$\bar{x} \sim N\left(\mu, \sigma^2 / n\right)$$

- The former refers to the distribution of a typical member of the population, and the latter to the distribution of the sample mean
- We usually refer to the square root of the variance of the sample mean as the standard error of the sample mean, rather than the standard deviation

Example

- What is the probability of drawing a sample of 50 women whose *average* height is > 168 cm?

$$z = \frac{168 - 166}{\sqrt{40.32/50}} = 2.23$$

  - z = 2.23 cuts off 1.29% in the upper tail of the standard Normal distribution, so there is only a probability of 1.29% of drawing a sample with a mean > 168 cm

- Q. what is probability of drawing a sample with a mean <168 cm?

The Distribution of the Sample Proportion

- The sample proportion also has a normal distribution

$$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

- where p is the sample proportion, π the population proportion, and the variance of the sample proportion is $\pi(1-\pi)/n$.

- since π is usually unknown we estimate it with p

The Central Limit Theorem

- If the sample size is large (*n* > 25) the population does not have to be Normally distributed, the sample mean is (approximately) Normal whatever the shape of the population distribution

- The approximation gets better, the larger the sample size. 25 is a safe minimum to use

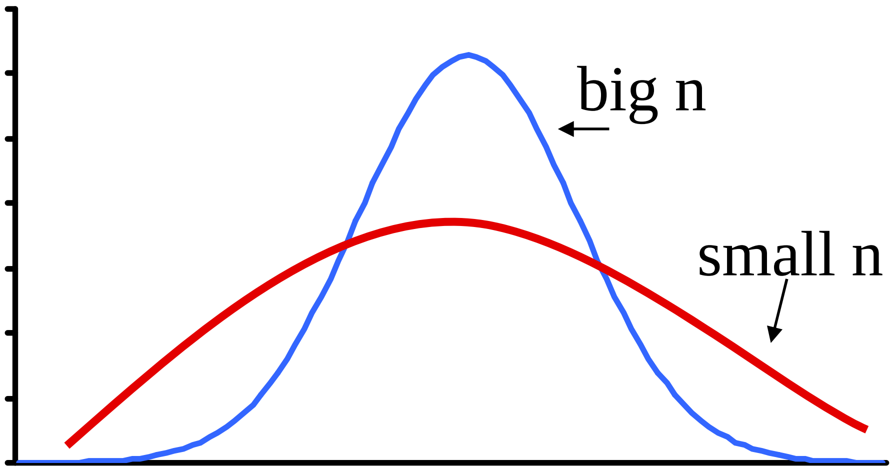
Distributions when Samples are Small: Using the *t* distribution

- When:
  - The sample size is small (<25 or so), and
  - The true variance, σ^2, is unknown

Then the *t* distribution should be used instead of the standard Normal.

The t Distribution

- The t distribution is
  - bell shaped
  - symmetric
  - unimodal
  - extends from $x = -\infty$ to $+\infty$ (in theory)
  - more spread out than Normal
  - depends on n-1 (degrees of freedom)



Summary

- Most statistical problems concern random variables which have an associated probability distribution
- Common distributions are the Binomial, Normal and Poisson (there many others)
- Once the appropriate distribution for the problem is recognised, the solution is relatively straightforward