

# Hypothesis Testing

Noel Welsh

12 October 2010

## 1 Announcements

- I was joking about paying by cash. Paying by cheque is preferable as we don't have to worry about the cheques getting lost. Make them payable to the University of Birmingham, and post date to May 2011.
- We currently have more students (44) than kits (36). We have two kits remaining to hand out, and are making up another two.

## 2 Recap

Last lecture we looked at the scientific method. We can up with a method that was (a variant of)

- Formulate hypothesis
- Collect data
- Evaluate data
- State conclusions

This lecture we going to look at the process in more detail, and in particular how we evaluate data. Next week we're going to look at experiments that don't fit this simple (Popperian) method. This material will be necessary to answer your first assignment.

## 3 Example

Let's start with a simple example. We're gonna flip coins.

- Audience give a hypothesis and experimental method
- Flip coin to generate data. Rope in audience to do this.
- Now we grind to a halt. How do we analyse our data?

## 4 The Binomial Distribution

- Does everyone know what a probability is?
- Assume the probability of heads is fixed. Call this probability  $p$ . Seems reasonable? A coin flip is what's known as a Bernoulli trial. That is, an experiment with a binary result.
- If we flip a coin  $n$  times, how many possible outcomes are there? [Ans:  $2^n$ ]
- So probability of a particular outcome with  $r$  successes is  $p^r(1-p)^{n-r}$
- There are  $\binom{n}{r} = \frac{n!}{r!(n-r)!}$  possible outcomes with  $r$  successes.
- Let  $S$  denote the number of 1s in the experiment. Then the probability that  $S = r$  is

$$P(S = r) = \binom{n}{r} p^r (1-p)^{n-r}, \forall r \in \{0, \dots, n\} \quad (1)$$

- Aside:  $S$  is called a *random variable*, meaning it is a function with domain the set of possible outcomes of an experiment, and each outcome has a probability (which may not be known). Random variables are usually denoted by upper case letters, and the value taken by a random variable as a lower case letter.
- The distribution of  $S$  is called the *binomial distribution* with parameters  $n$  and  $p$ .
- Examples. Notice the shape becomes more symmetrical as the number of samples increase. We'll get back to this.

## 5 Hypothesis Testing

- We have a hypothesis. This is called the *null hypothesis* to distinguish it from the many other possible *alternate hypotheses*.

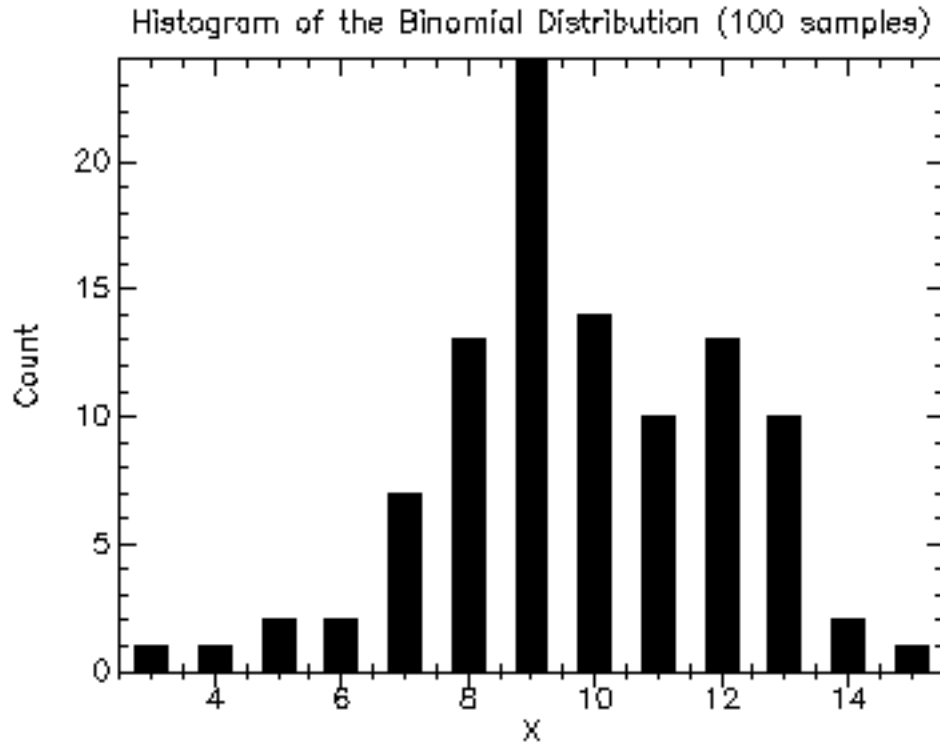


Figure 1: Histogram of 100 samples from the binomial distribution ( $p = 0.5$ ,  $n = 20$ )

- We need a *rejection rule*, which is a way to say if we reject (or accept) the null hypothesis. Note that we can reject a hypothesis, but never really accept it. We just conclude there is not sufficient evidence to reject it. This is for philosophical as well as practical reasons.
- We can calculate the probability of the data under the Binomial distribution, and that gives us a way to create a rejection rule. E.g. if we saw 3 heads in 20 tosses and we hypothesise a fair coin, the probability of this result is  $\approx 0.001$ . This is very unlikely, so we might reject the null hypothesis.
- Of course this data *could* arise in practice, so we might make an error. The cases are this:

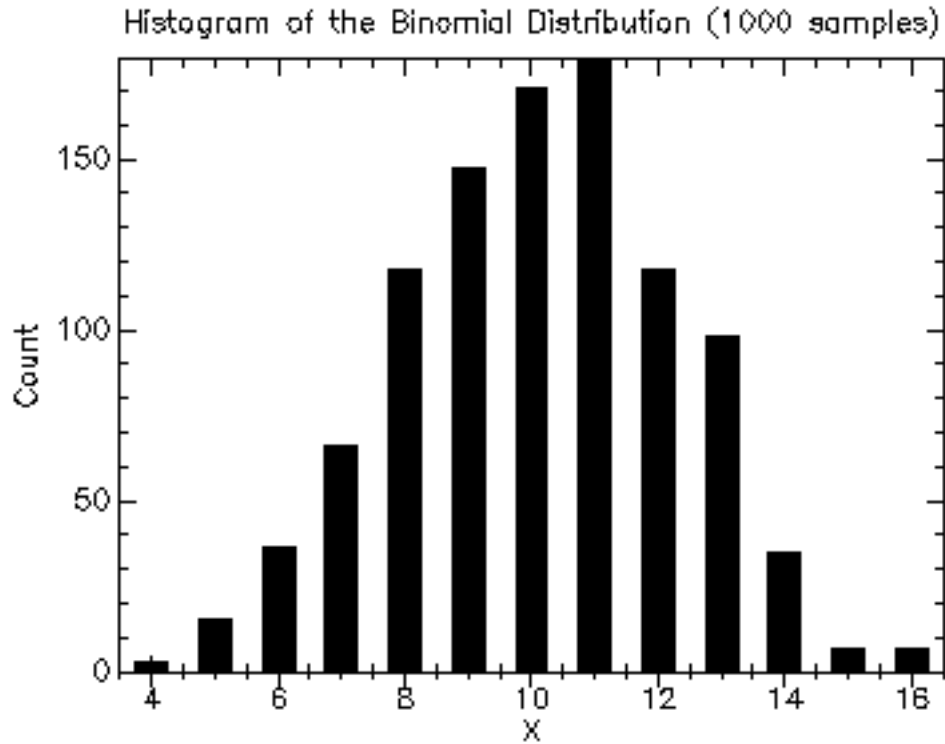


Figure 2: Histogram of 1000 samples from the binomial distribution ( $p = 0.5$ ,  $n = 20$ )

	We accept $H_0$	We reject $H_0$
$H_0$ true	No error	Type 1 error
$H_0$ false	Type 2 error	No error

- The probability of a Type 1 error is  $\approx 0.001$  in the case above. I.e.  $P(data|H_0)$  and the probability of accepting  $H_0$  when it is true is  $1 - P(data|H_0)$ .
- To calculate the probability of a Type 2 error we must consider  $P(data)$  under all possible alternate hypotheses. We can see that this value can be very high.
- The threshold value of  $P(data|H_0)$  at which we switch from an accept to a reject of  $H_0$  is known as the *p-value*.

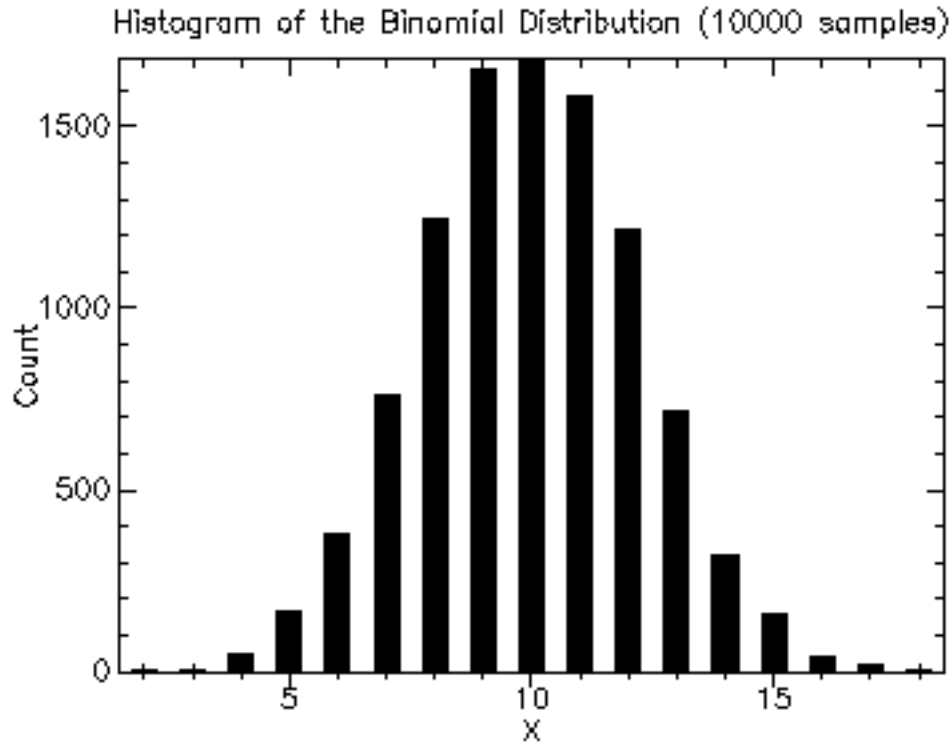


Figure 3: Histogram of 10000 samples from the binomial distribution ( $p = 0.5, n = 20$ )

## 6 The Normal Distribution

- Recall the samples from the binomial distribution converged to a symmetric shape. This is known as the Normal (or Gaussian) distribution.
- Parameterised by mean (centre) and variance (width).
- The Normal distribution appears all over the place when investigating natural phenomena. This is formalised by the Central Limit Theorem:

Let  $X_1, X_2, X_3, \dots, X_n$  be a sequence of  $n$  independent and identically distributed (iid) random variables with finite expectation  $\mu$  and variance  $\sigma^2 > 0$ . The central limit theorem states that as the sample size  $n$  increases the distribution of the sample *average* of these random variables approaches the normal distribution with a mean  $\mu$  and variance  $\frac{\sigma^2}{n}$  irrespective of the shape

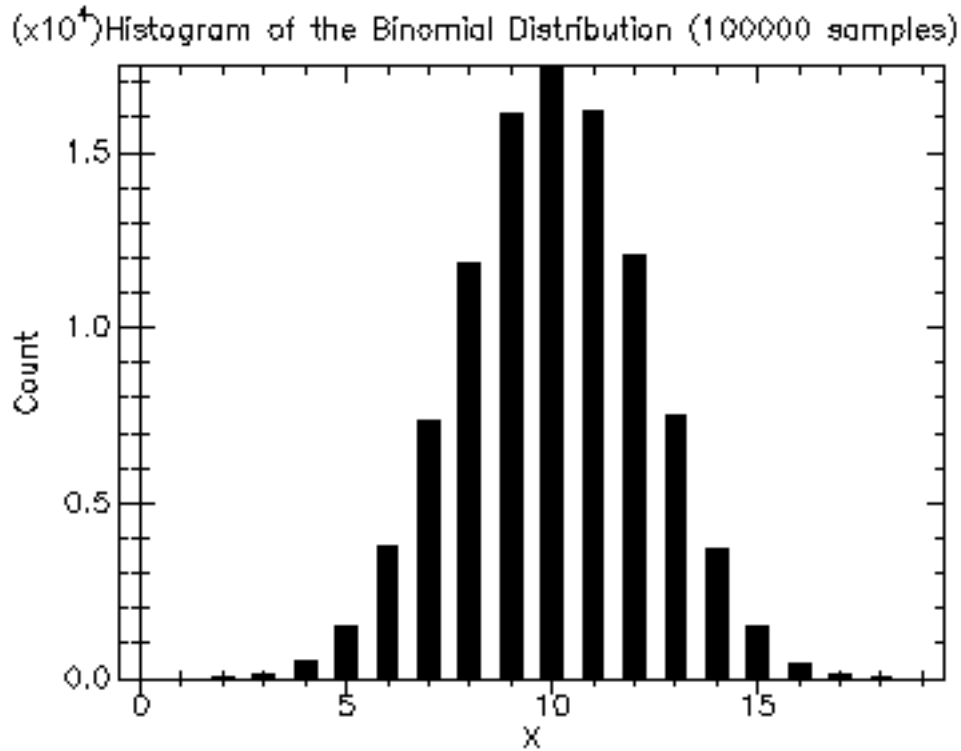


Figure 4: Histogram of 100000 samples from the binomial distribution ( $p = 0.5$ ,  $n = 20$ )

of the common distribution of the individual terms  $X_i$ .

- The Normal distribution has PDF

$$P(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Can drop terms and simplify to see that basically the Normal distribution cares about squared error or squared distance. This has big implications – assumptions of normality are susceptible to outliers. E.g. blur, heavy tails etc. It also means anything that optimises squared error is assuming a Gaussian model. E.g. PCA, linear regression.
- If we assume data is Normally distributed the appropriate hypothesis test is the  $t$ -test. This test makes use of Student's  $t$ -distribution. This arises by

considering both the mean and variance as unknowns to be estimated from data. Since we estimate the variance from the sample mean the variance must have higher error. Thus the  $t$ -distribution has heavier tails than the Normal distribution.

## 7 More!

- “Scientific Methods in Mobile Robotics” is a very good book that covers what it claims to cover.
- You should find out about:
  - One- and two-tailed tests.
  - $\chi^2$  (Chi-squared) test.
  - Non-parametric tests (U-statistic and Wilcoxon test)