Chapter 1

Statistics

1.1 Introduction

This lecture is a quick review of basic statistical concepts; probabilities, mean, variance, covariance, correlation, linear regression, probability density functions and significance testing.

1.2 Probabilities

1.2.1 Discrete Variables

The table below shows the probability of occurrence $p(x = x_i)$ of selected letters x_i in the English alphabet. Table 2 shows the probability of occurence of selected

x_i	$p(x_i)$
a	0.06
e	0.09
j	0.00
q	0.01
t	0.07
z	0.00

Table 1.1: *Probability of letters*

pairs of letters x_i and y_j where x_i is followed by y_j. This is called the *joint probability* $p(x = x_i, y = y_i)$. If we fix x to, say x_i then the probability of y taking on a particular value, say y_j, is given by the *conditional probability*

$$p(y = y_j | x = x_i) = \frac{p(x = x_i, y = y_j)}{p(x = x_i)} \tag{1.1}$$

x_i	y_j	$p(x_i, y_j)$
t	h	0.037
t	s	0.000
t	r	0.012

Table 1.2: *Probability of pairs of letters*

For example, if $x_i = t$ and $y_j = h$ then the joint probability $p(x = x_i, y = y_j)$ is just the probability of occurence of the pair (which table 2 tells us is 0.037). The conditional probability $p(y = y_j|x = x_i)$, however, says that, *given* we've seen the letter t, what's the probability that the next letter will be h (which is, from tables 1 and 2, $0.037/0.07 = 0.53$). Re-arranging the above relationship gives

$$p(x = x_i, y = y_j) = p(y = y_j|x = x_i)p(x = x_i) \tag{1.2}$$

Now if y does *not* depend on x then $p(y = y_j|x = x_i) = p(y = y_j)$. Hence, for independent variables, we have

$$p(x = x_i, y = y_j) = p(y = y_j)p(x = x_i) \tag{1.3}$$

The *marginal probability* is given by

$$p(x = x_i) = \sum_{\{y_j\}} p(y = y_j, x = x_i) \tag{1.4}$$

This is the same probability that we started with.

1.2.2 Continuous Variables

The probability of a continuous variable, x, assuming a particular value or range of values is defined by a Probability Density Funcion (PDF), $p(x)$. *Probability is measured by the area under the PDF*; the total area under a PDF is therefore unity

$$\int p(x)dx = 1 \tag{1.5}$$

The probability of x assuming a value between a and b is given by

$$p(a \leq x \leq b) = \int_a^b p(x)dx \tag{1.6}$$

which is the area under the PDF between a and b. *The probability of x taking on a single value is therefore zero.* This makes sense because we are dealing with continuous values; as your value becomes more precise the probability for it decreases. It only makes sense, therefore to talk about the probability of a value being within a certain precision or being above or below a certain value.

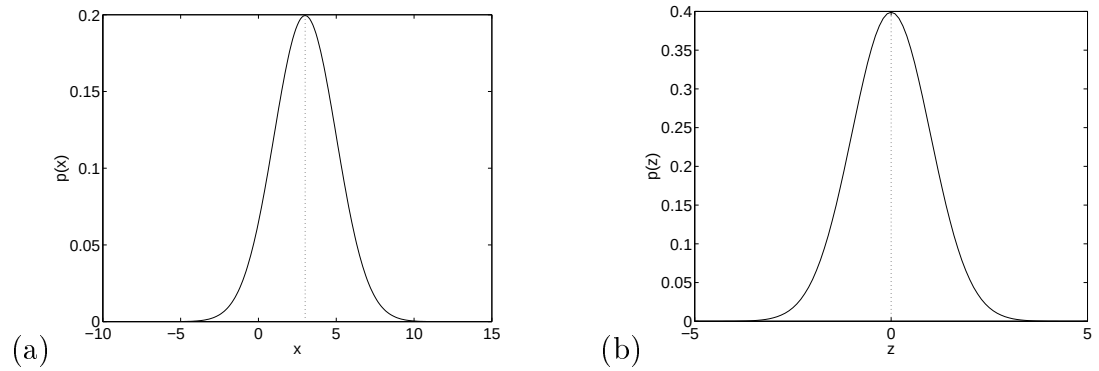

Figure 1.1: *(a) The Gaussian Probability Density Function with mean $\mu = 3$ and standard deviation $\sigma = 2$, (b) The standard Gaussian density, $p(z)$. This has zero mean and unit variance.*

To calculate such probabilities we need to calculate integrals like the one above. This process is simplified by the use of Cumulative Density Functions (CDF) which are defined as

$$CDF(a) = p(x \leq a) = \int_{-\infty}^{a} p(x)dx \tag{1.7}$$

Hence

$$p(a \leq x \leq b) = CDF(b) - CDF(a) \tag{1.8}$$

1.2.3 The Gaussian Density

The *Normal* or *Gaussian* probability density function, for the case of a single variable, is

$$p(x) \equiv N(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1.9}$$

where μ and σ^2 are known as the *mean* and *variance*, and σ (the square root of the variance) is called the *standard deviation*. The quantity in front of the exponential ensures that $\int p(x)dx = 1$. The above formula is often abbreviated to the shorthand $p(x) = N(x; \mu, \sigma)$. The terms Normal and Gaussian are used interchangeably.

If we subtract the mean from a Gaussian variable and then divide by that variables *standard deviation* the resulting variable, $z = (x-\mu)/\sigma$, will be distributed according the *standard* normal distribution, $p(z) = N(z; 0, 1)$ which can be written

$$p(z) = \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2}\right) \tag{1.10}$$

The probability of z being above 0.5 is given by the area to the right of 0.5. We can calculate it as

$$\begin{aligned} p(z) \geq 0.5 &= \int_{0.5}^{\infty} p(z)dz \\ &= 1 - CDF_{Gauss}(0.5) \end{aligned} \tag{1.11}$$

where CDF_{Gauss} is the cumulative density function for a Gaussian.

1.2.4 Probability relations

The same probability relations hold for continuous variables as for discrete variables ie. the conditional probability is

$$p(y|x) = \frac{p(x,y)}{p(x)} \tag{1.12}$$

Re-arranging gives the joint probability

$$p(x,y) = p(y|x)p(x) \tag{1.13}$$

which, if y does not depend on x (ie. x and y are independent) means that

$$p(x,y) = p(y)p(x) \tag{1.14}$$

1.3 Expectation and Moments

The *expected value* of a function $f(x)$ is defined as

$$E[f(x)] \equiv < f(x) > = \int p(x)f(x)dx \tag{1.15}$$

and $E[]$ is referred to as the *expectation* operator, which is also sometimes written using the angled brackets $<>$. The kth *moment* of a distribution is given by

$$E[x^k] = \int p(x)x^k dx \tag{1.16}$$

The mean is therefore the first moment of a distribution.

$$E[x] = \int p(x)xdx = \mu \tag{1.17}$$

The kth *central moment* of a distribution is given by

$$E[(x-\mu)^k] = \int p(x)(x-\mu)^k dx \tag{1.18}$$

The variance is therefore the second central moment

$$E[(x-\mu)^2] = \int p(x)(x-\mu)^2 dx = \sigma^2 \tag{1.19}$$

Sometimes we will use the notation

$$Var(x) = E[(x-\mu)^2] \tag{1.20}$$

The third central moment is *skewness* and the fourth central moment is *kurtosis* (see later). In the appendix we give examples of various distributions and of skewness and kurtosis.

1.4 Maximum Likelihood Estimation

We can learn the mean and variance of a Gaussian distribution using the Maximum Likelihood (ML) framework as follows. A Gaussian variable x_n has the PDF

$$p(x_n) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{1.21}$$

which is also called the likelihood of the data point. Given N Independent and Identically Distributed (IID) (it is often assumed that the data points, or errors, are independent and come from the same distribution) samples $y = [y_1, y_2, .., y_N]$ we have

$$p(y) = \prod_{n=1}^{N} p(y_n) \tag{1.22}$$

which is the likelihood of the data set. We now wish to set μ and σ^2 so as to maximise this likelihood. For numerical reasons (taking logs gives us bigger numbers) this is more conveniently achieved by maximising the log-likelihood (note: the maximum is given by the same values of μ and σ)

$$L \equiv \log p(y) = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \sum_{n=}^{N} \frac{(y_n-\mu)^2}{2\sigma^2} \tag{1.23}$$

The optimal values of μ and σ are found by setting the derivatives $\frac{dL}{d\mu}$ and $\frac{dL}{d\sigma}$ to zero. This gives

$$\mu = \frac{1}{N}\sum_{n=1}^{N} y_n \tag{1.24}$$

and

$$\sigma^2 = \frac{1}{N}\sum_{n=1}^{N} (y_n-\mu)^2 \tag{1.25}$$

We note that the last formula is different to the usual formula for estimating variance

$$\sigma^2 = \frac{1}{N-1}\sum_{n=1}^{N} (x_n-\mu)^2 \tag{1.26}$$

because of the difference in normalisation. The last estimator of variance is preferred as it is an *unbiased* estimator (see later section on bias and variance).

If we had an input-dependent mean, $\mu_n = wx_n$, then the optimal value for w can be found by maximising L. As only the last term in equation 1.23 depends on w this therefore corresponds to minimisation of the squared errors between μ_n and y_n. This provides the connection between ML estimation and Least Squares (LS) estimation; ML reduces to LS for the case of Gaussian noise.

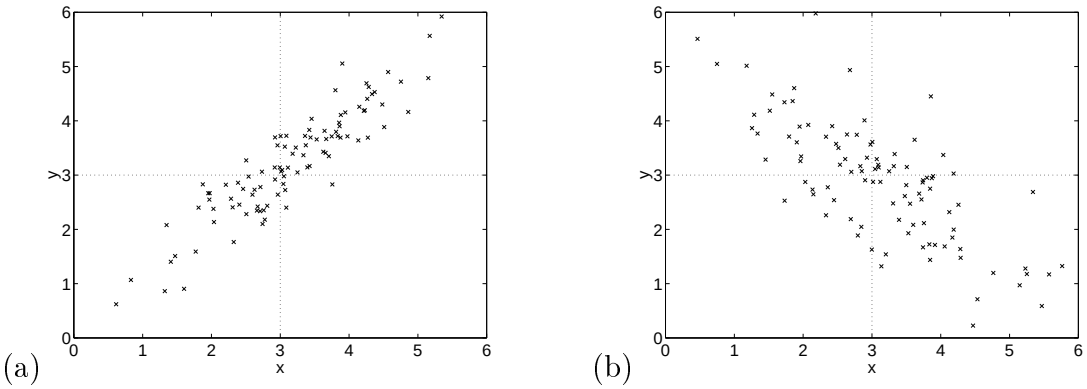

Figure 1.2: *(a) Positive correlation,* $r = 0.9$ *and (b) Negative correlation,* $r = -0.7$*. The dotted horizontal and vertical lines mark* μ_x *and* μ_y*.*

1.5 Correlation and Regression

1.5.1 Correlation

The *covariance* between two variables x and y is measured as

$$\sigma_{xy} = \frac{1}{N-1} \sum_{n=1}^{N} (x_i - \mu_x)(y_i - \mu_y) \tag{1.27}$$

where μ_x and μ_y are the means of each variable. Note that $\sigma_{yx} = \sigma_{xy}$. Sometimes we will use the notation

$$Var(x, y) = \sigma_{xy} \tag{1.28}$$

If x tends to be above its mean when y is above its mean then σ_{xy} will be positive. If they tend to be on opposite sides of their means σ_{xy} will be negative. The *correlation* or *Pearson's correlation coefficient* is a normalised covariance

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{1.29}$$

such that $-1 \leq r \leq 1$, a value of -1 indicating perfect negative correlation and a value of $+1$ indicating perfect positive correlation; see Figure 1.2. A value of 0 indicates no correlation. The strength of a correlation is best measured by r^2 which takes on values between 0 and 1, a value near to 1 indicating strong correlation (regardless of the sign) and a value near to zero indicating a very weak correlation.

1.5.2 Linear regression

We now look at modelling the relationship between two variables x and y as a linear function; given a collection of N data points $\{x_i, y_i\}$, we aim to estimate y_i from x_i using a linear model

$$\hat{y}_i = ax_i + b \tag{1.30}$$

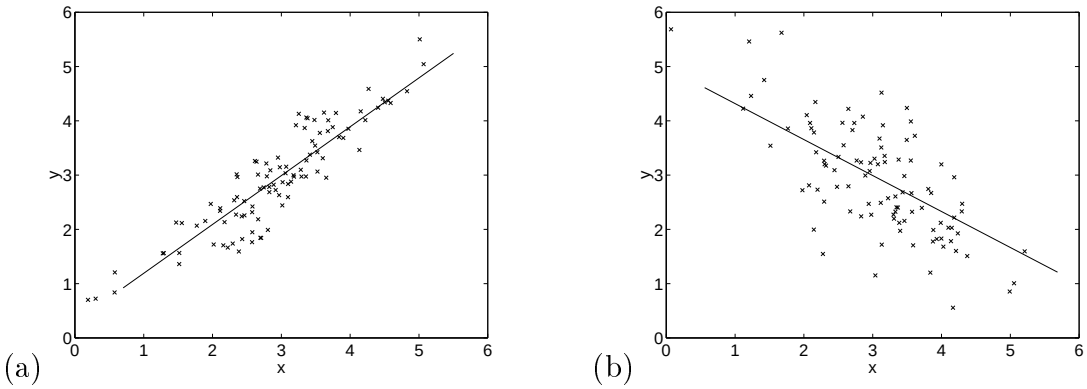

Figure 1.3: *The linear regression line is fitted by minimising the vertical distance between itself and each data point. The estimated lines are (a)* $\hat{y} = 0.9003x + 0.2901$ *and (b)* $\hat{y} = -0.6629x + 4.9804$.

where we have written $\hat{y}$ to denote our estimated value. Regression with one input variable is often called *univariate* linear regression to distinguish it from *multivariate* linear regression where we have lots of inputs. The goodness of fit of the model to the data may be measured by the least squares cost function

$$E = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{1.31}$$

The values of a and b that minimize the above cost function can be calculated by setting the first derivatives of the cost function to zero and solving the resulting simultaneous equations (derivatives are used to find maxima and minima of functions). The result is derived in the Appendix. The solutions are

$$a = \frac{\sigma_{xy}}{\sigma_x^2} \tag{1.32}$$

and

$$b = \mu_y - a\mu_x \tag{1.33}$$

where μ_x and μ_y are the mean observed values of the data and σ_x^2 and σ_{xy} are the input variance and input-output covariance. This enables least squares fitting of a regression line to a data set as shown in Figure 1.3.

The model will fit some data points better than others; those that it fits well constitute the *signal* and those that it does'nt fit well constitute the *noise*. The strength of the noise is measured by the noise variance

$$\sigma_e^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{1.34}$$

and the strenth of the signal is given by $\sigma_y^2 - \sigma_e^2$. The *signal-to-noise ratio* is therefore $(\sigma_y^2 - \sigma_e^2)/\sigma_e^2$.

Splitting data up into signal and noise components in this manner (ie. breaking down the variance into what the model *explains* and what it does not) is at the heart of statistical procedures such as analysis of variance (ANOVA) [32].

Relation to correlation

The correlation measure r is intimately related to the linear regression model. Indeed (by substituting σ_{xy} from equation 1.27 into equation 1.32) r may be expressed as

$$r = \frac{\sigma_x}{\sigma_y} a \tag{1.35}$$

where a is the slope of the linear regression model. Thus, for example, the sign of the slope of the regression line defines the sign of the correlation. The correlation is, however, also a function of the standard deviation of the x and y variables; for example, if σ_x is very large, it is possible to have a strong correlation even though the slope may be very small.

The relation between r and linear regression emphasises the fact that r is only a measure of *linear* correlation. It is quite possible that two variables have a strong nonlinear relationship (ie. are nonlinearly correlated) but that $r = 0$. Measures of nonlinear correlation will be discussed in a later lecture.

The strenth of correlation can also be expressed in terms of quantites from the linear regresssion model

$$r^2 = \frac{\sigma_y^2 - \sigma_e^2}{\sigma_y^2} \tag{1.36}$$

where σ_e^2 is the noise variance and σ_y^2 is the variance of the variable we are trying to predict. Thus r^2 is seen to measure the proportion of variance explained by a linear model, a value of 1 indicating that a linear model perfectly describes the relationship between x and y.

1.6 Bias and Variance

Given any estimation process, if we repeat it many times we can look at the expected (or average) errors (the difference between true and estimated values). This is comprised of a systematic error (the 'bias') and an error due to the variability of the fitting process (the 'variance'). We can show this as follows.

Let w be the true value of a parameter and $\hat{w}$ be an estimate from a given sample. The expected squared error of the estimate can be decomposed as follows

$$\begin{aligned} E &= E[(\hat{w} - w)^2] \\ &= E[(\hat{w} - E[\hat{w}] + E[\hat{w}] - w)^2] \end{aligned} \tag{1.37}$$

where the expectation is wrt. the distribution over $\hat{w}$ and we have introduced $E[\hat{w}]$, the mean value of the estimate. Expanding the square gives

$$\begin{aligned} E &= E[(\hat{w} - E[\hat{w}])^2 + (E[\hat{w}] - w)^2 + 2(\hat{w} - E[\hat{w}])(E[\hat{w}] - w)] \\ &= E[(\hat{w} - E[\hat{w}])]^2 + (E[\hat{w}] - w)^2 \\ &= V + B^2 \end{aligned} \tag{1.38}$$

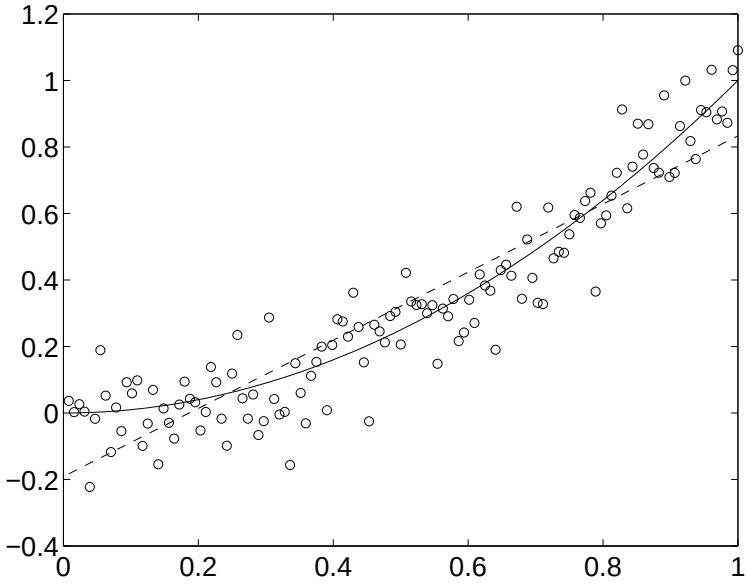

Figure 1.4: *Fitting a linear regression model (dotted line) to data points (circles) which are generated from a quadratic function (solid line) with additive noise (of variance 0.01).*

where the third term has dropped out because $E[\hat{w}] - E[\hat{w}] = 0$. The error thus consists of two terms (i) a variance term V and (ii) a bias term; the square of the bias, B^2.

Estimates of parameters are often chosen to be *unbiased* ie. to have zero bias. This is why we see the $1/(N-1)$ term in an estimate of variance, for example.

Simple models (eg. linear models) have a high bias but low variance whereas more complex models (eg. polynomial models) have a low bias but a high variance. To select the optimal model complexity, or model order, we must solve this *bias-variance dilemma* [20].

1.7 Minimum variance estimation

There is a lower bound to the variance of any unbiased estimate which is given by

$$Var(\hat{\theta}) \geq \frac{1}{E[\partial L(D,\theta)/\partial\theta]^2} \tag{1.39}$$

where $L(D;\theta) \equiv \log p(D;\theta)$ is the log-likelihood of the data and the expectation is taken wrt. $p(D;\theta)$. This is known as the *Cramer-Rao bound*. Any estimator that attains this variance is called the Minimum Variance Unbiased Estimator (MVUE).

The denominator, being an inverse variance, therefore measures the maximum precision with which we can estimate θ. It is known as the *Fisher Information*

$$I(\theta) = E[\partial L(D;\theta)/\partial\theta]^2 \tag{1.40}$$

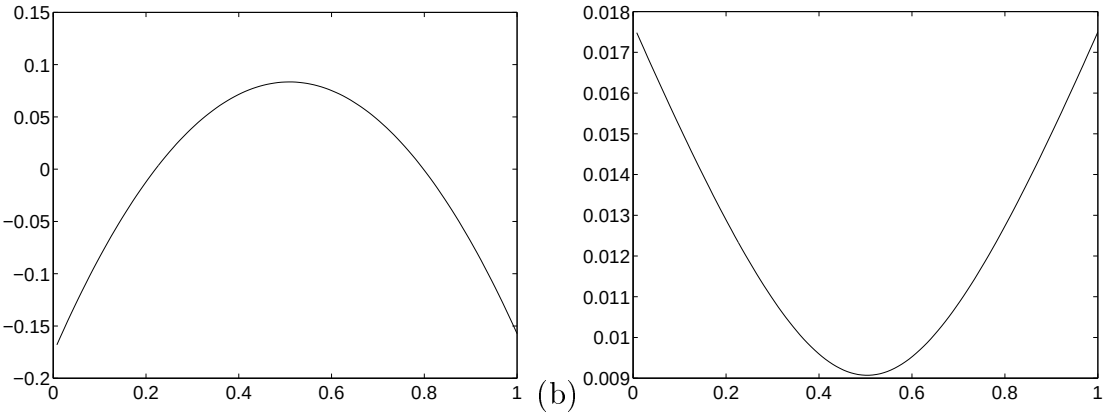

Figure 1.5: *(a) Bias component B and (b) Variance component V. The bias represents a systematic error in our modelling procedure (ie. fitting a quadratic function with a linear function); the linear model systematically underpredicts at the edges and overpredicts in the middle. The variance represents the variability of the model fitting process; linear models lock on to the middle of a data set and then set their slope as necessary. The variance is therefore less in the middle than at the edges; in the middle this variance is simply the variance of the additive noise (0.01). The expected prediction error at any point is the sum of the variance plus the bias squared.*

For unbiased estimates [53] it can also be expressed as

$$I(\theta) = -E[\partial^2 L(D;\theta)/\partial\theta^2] \tag{1.41}$$

1.8 Statistical Inference

When we estimate the mean and variance from small samples of data our estimates may not be very accurate. But as the number of samples increases our estimates get more and more accurate and as this number approaches infinity the sample mean approaches the true mean or *population* mean. In what follows we refer to the sample means and variances as m and s and the population means and standard deviations as μ and σ.

Hypothesis Testing: Say we have a hypothesis $\boldsymbol{H}$ which is *The mean value of my signal is 32.* This is often referred to as the *null hypothesis* or H_0. We then get some data and test $\boldsymbol{H}$ which is then either *accepted* or *rejected* with a certain probability or *significance level*, p. Very often we choose $p = 0.05$ (a value used throughout science).

We can do a *one-sided* or a *two-sided* statistical test depending on exactly what the null hypothesis is. In a one-sided test our hypothesis may be (i) our parameter is less than x or (ii) our parameter is greater than x. For two-sided tests our hypothesis is of the form (iii) our parameter is x. This last hypothesis can be rejected if the sample statistic is either much smaller or much greater than it should be if the parameter truly equals x.

1.8.1 Means

To find out if your mean is significantly different from a hypothesized value μ there are basically two methods. The first assumes you know the population/true variance and the second allows you to use the sample variance.

Known variance

If we estimate the mean from a sample of data, then this estimate itself has a mean and a standard deviation. The standard deviation of the sample mean is (see appendix)

$$\sigma_m = \sigma/\sqrt{N} \tag{1.42}$$

where σ is the known true standard deviation. The probability of getting a particular sample mean from N samples is given by $p(z)$ where

$$z = \frac{m-\mu}{\sigma/\sqrt{N}} \tag{1.43}$$

For example, suppose we are given 50 data points from a normal population with hypothesized mean $\mu = 32$ and standard deviation $\sigma = 2$ and we get a sample mean of 32.3923, as shown in Figure 1.6. The probability of getting a sample mean *at least* this big is

$$p(m > 32.3923) = 1 - CDF_{Gauss}(z) \tag{1.44}$$

where $z = (32.3923 - 32)/(2/\sqrt{50}) = 1.3869$ which is (from tables or computer evaluation) 0.0827 ie. reasonably likely; we would accept the hypothesis at the $p = 0.05$ level (because we are doing a two-sided test we would accept H_0 unless the probability was less than $p = 0.025$).

Unknown variance

If we don't know the true variance we can use the sample variance instead. We can then calculate the statistic

$$t = \frac{m-\mu}{s/\sqrt{N}} \tag{1.45}$$

which is distributed according the t-distribution (see appendix). Now, the t-distribution has a parameter v, called the degrees of freedom (DF). It is plotted in Figure 1.7 with $v = 3$ and $v = 49$ degrees of freedom; smaller v gives a wider distribution.

Now, from our $N = 50$ data points we calculated the sample variance ie. given, originally, 50 DF we have used up one DF leaving $N - 1 = 49$ DF. Hence, our t-statistic has $v = 49$ degrees of freedom.

Assume we observed $s = 2$ and $m = 32.3923$ (as before) and our hypothesized mean is 32. We can calculate the associated probability from

$$p(m > 32.3923) = 1 - CDF_t(t) \tag{1.46}$$

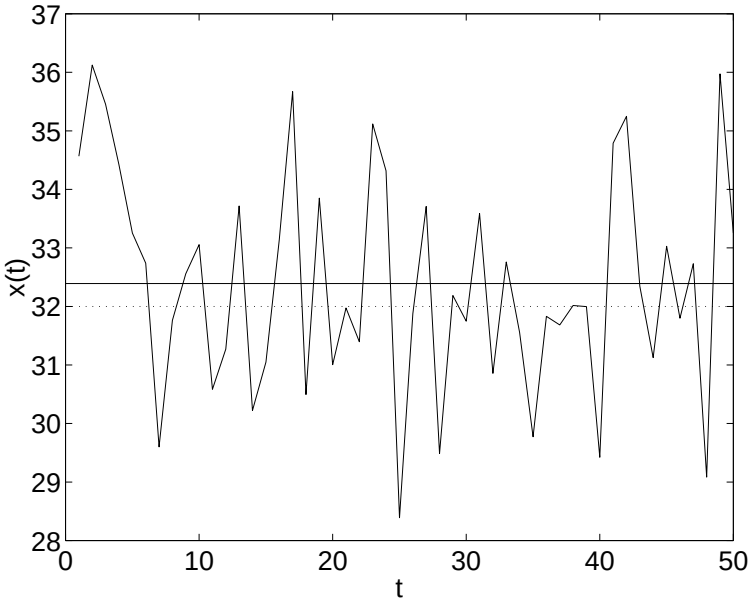

Figure 1.6: *N=50 data points. The hypthosized mean value of 32 is shown as a dotted line and the sample mean as a solid line.*

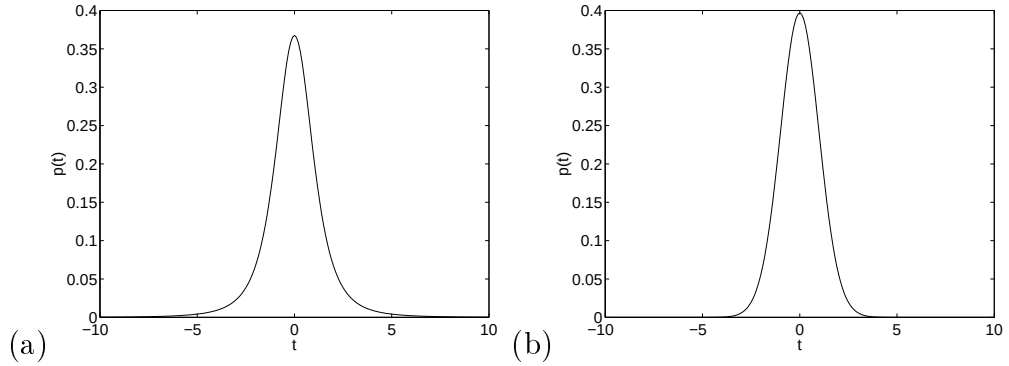

Figure 1.7: *The t-distribution with (a) $v = 3$ and (b) $v = 49$ degrees of freedom.*

where $t = (32.3923 - 32/(2/\sqrt{50}) = 1.3869$. From tables this gives 0.0859 ie. reasonably likely (again, because we are doing a two-sided test, we would accept H_0 unless the probability was less than $p = 0.025$). Notice, however, that the probability is higher than when we *knew* the standard deviation to be 2. This shows that a t-distribution has heavier tails than a Normal distribution ie. extreme events are more likely.

1.8.2 Regression

In a linear regression model we are often interested in whether or not the gradient is significantly different from zero or other value of interest.

To answer the question we first estimate the variance of the slope and then perform

a t-test. In the appendix we show that the variance of the slope is given by [1]

$$\sigma_a^2 = \frac{\sigma_e^2}{(N-1)\sigma_x^2} \tag{1.47}$$

We then calculate the t-statistic

$$t = \frac{a - a_h}{\sigma_a} \tag{1.48}$$

where a_h is our hypothesized slope value (eg. a_h may be zero) and look up $p(t)$ with $N-2$ DF (we have used up 1DF to estimate the input variance and 1DF to estimate the noise variance). In the data plotted in Figure 1.3(b) the estimated slope is $a = -0.6629$. From the data we also calculate that $\sigma_a = 0.077$. Hence, to find out if the slope is significantly non-zero we compute $CDF_t(t)$ where $t = -0.6629/0.077 = -8.6$. This has a p-value of 10^{-13} ie. a very significant value. To find out if the slope is significantly different from -0.7 we calculate $CDF_t(t)$ for $t = (-0.6629+0.7)/0.077 = 0.4747$ which gives a p-value of 0.3553 ie. not significantly different (again, we must bear in mind that we need to do a two-sided test; see earlier).

1.8.3 Correlation

Because of the relationship between correlation and linear regression we can find out if correlations are significantly non-zero by using exactly the same method as in the previous section; if the slope is significantly non-zero then the corresponding correlation is also significantly non-zero.

By substituting $a = (\sigma_y/\sigma_x)r$ (this follows from equation 1.32 and equation 1.29) and $\sigma_e^2 = (1-r^2)\sigma_y^2$ (from equation 1.36) into equation 1.47 and then σ_a into equation 1.48 we get the test statistic [2]

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{1.49}$$

which has $N-2$ DF.

For example, the two signals in Figure 1.8(a) have, over the $N = 50$ given samples, a correlation of $r = 0.8031$ which gives $t = 9.3383$ and a p-value of 10^{-12}. We therefore reject the hypothesis that the signals are not correlated; they clearly are. The signals in Figure 1.8(b) have a correlation of $r = 0.1418$ over the $N = 50$ given samples which gives $t = 0.9921$ and a p-value of $p = 0.1631$. We therefore accept the null hypothesis that the signals are not correlated.

[1] When estimating σ_x^2 we should divide by $N-1$ and when estimating σ_e^2 we should divide by $N-2$.

[2] Strictly, we should use $\sigma_e^2 = \frac{N-1}{N-2}(1-r^2)\sigma_y^2$ to allow for using $N-2$ in the denominator of σ_e^2.

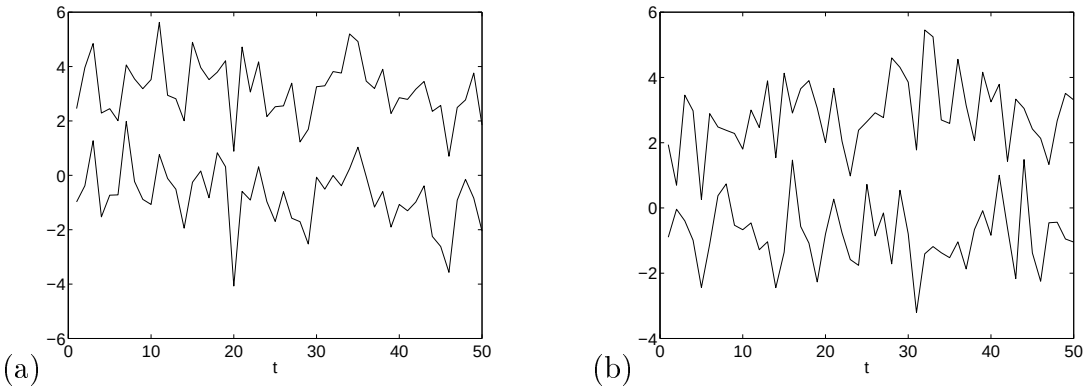

Figure 1.8: *Two signals (a) sample correlation* $r = 0.8031$ *and (b) sample correlation,* r=0.1418. *Strong correlation; by shifting and scaling one of the time series (ie. taking a linear function) we can make it look like the other time series.*

1.9 Discussion

For a more comprehensive introduction to basic statistics, linear regression and significance testing see Grimmett and Welsh [26] or Kleinbaum et al. [32]. Also, *Numerical Recipes* [49] has very good sections on *Are two means different ?* and *Are two variances different ?*. See Priestley for a more comprehensive introduction to statistical estimation in time series models ([50], chapter 5).