

1 An informal introduction to estimation

The key to statistics is that we attempt to specify the errors we make. We can do this because we sample. Suppose we take a random sample x_1, x_2, \dots, x_N . It would be very strange if all the x values were identical and we can imagine they follow some distribution, $f(x)$ say. Thus given

4	5	1	4	3
3	1	2	2	2
4	4	2	4	4
2	3	5	4	4
5	3	1	3	4

we might speculate that the individual x 's follow a Binomial $B(6,0.5)$ distribution while in this case

0.080	0.143	0.572	0.026	1.669
0.207	0.162	0.018	0.247	0.125
0.448	0.255	0.070	0.999	0.714
0.012	0.398	0.700	0.055	0.457
0.123	0.675	1.599	0.541	0.198

an exponential distribution $f(x) = \theta \exp(-\theta x)$ may be more sensible.

2 The EDF

It is quite difficult to deduce the underlying distribution when given a histogram and we look at an alternative. Recall that the cumulative distribution function of a random variable X is defined as

$$F(x) = P[X \leq x] = \begin{cases} \int_{-\infty}^x f(t) dt \\ \sum_{j=0}^x P[X = j] \end{cases}$$

This is just the probability function given in tables. An *estimate* for this function $F(x)$ is the empirical distribution function $S_N(x)$ which we now define.

Suppose we have sample values $x_1, x_2, x_3, \dots, x_N$. We arrange these in ascending order to get $x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(N)}$. Then $S_N(x)$ is defined as

$$S_N(x) = \begin{cases} 0 & \text{if } x < x_{(1)} \\ \frac{k}{N} & \text{if } x_{(k)} \leq x < x_{(k+1)} \\ 1 & \text{if } x_{(N)} \leq x \end{cases}$$

There is a lot of elegant theory available about the empirical c.d.f. but it is rather specialized and we shall concern our selves with applications. We will just note in passing that $S_N(x)$ is a consistent estimate of $F(x)$.

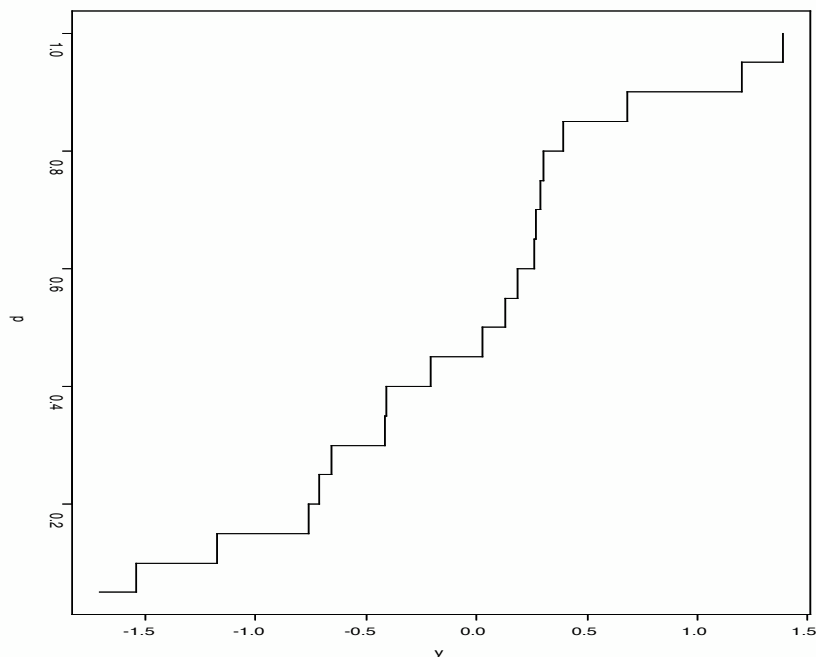
Thus for example given the values

0.2706712	0.2643819	0.3033932	-0.41405307
-1.7100219	0.3915392	-0.4087857	0.28956633
0.1855903	-0.2109898	-0.7115813	0.67984218
1.2013173	-1.1759424	-0.6582858	0.02664849
-0.7577526	1.3896886	0.1315669	-1.54254054

we arrange in order to get

-1.7100219	-0.65828576	0.1315669	0.3033932
-1.5425405	-0.41405307	0.1855903	0.3915392
-1.1759424	-0.40878567	0.2643819	0.6798422
-0.7577526	-0.21098979	0.2706712	1.2013173
-0.7115813	0.02664849	0.2895663	1.3896886

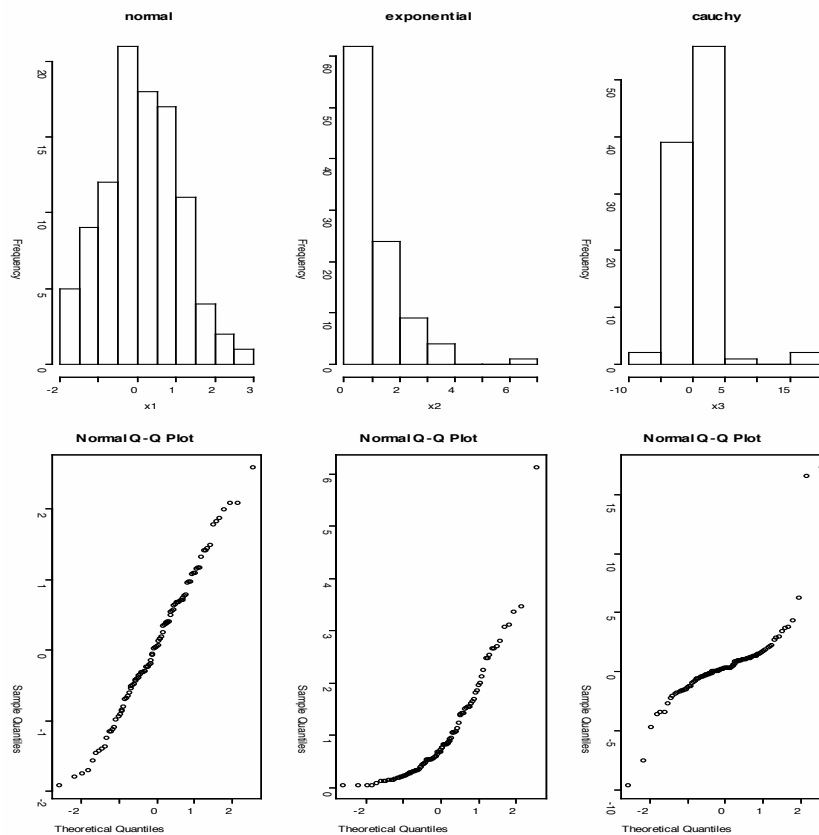
the EDF is then



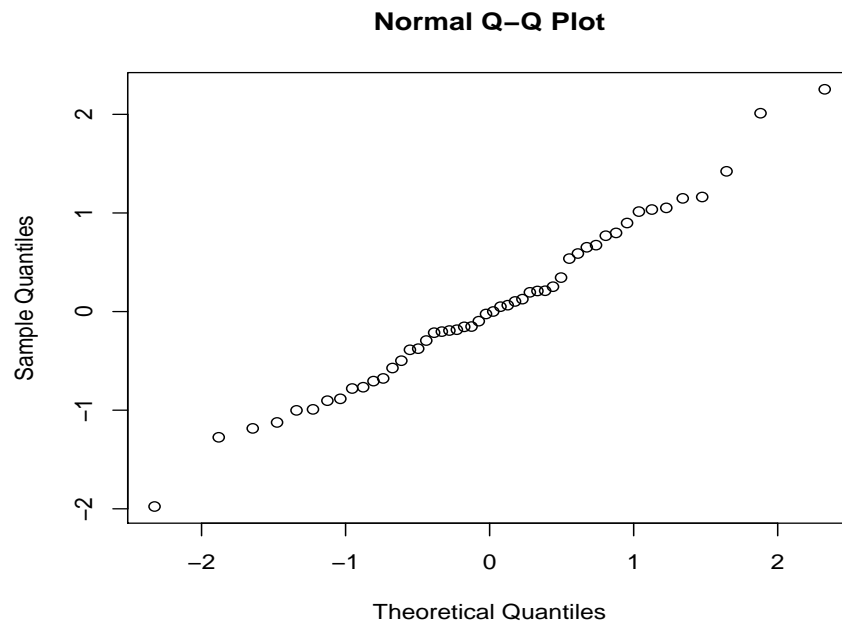
Note in R it is referred to as the `ecdf`.

Clearly the problem of relating the EDF to the cumulative distribution is just the same as trying to relate the histogram to the density. However we do have a cunning plan. Suppose we choose a potential cumulative distribution, say $G(x)$ Then if we plot $G(x_{(i)})$ against $S_N(x_i)$ **we will have a straight line** when the choice of $G(x)$ is correct.

This plot and variants on it is called a *probability plot* and is often used to check assumptions about distributions. In essence, a straight line implies that we have the correct distribution. Thus if we choose a Normal for our proposed distribution the figure below gives some possible outcomes.



Many programs provide an EDF plot for Normal distribution, for example in R we have `qqnorm` which gives

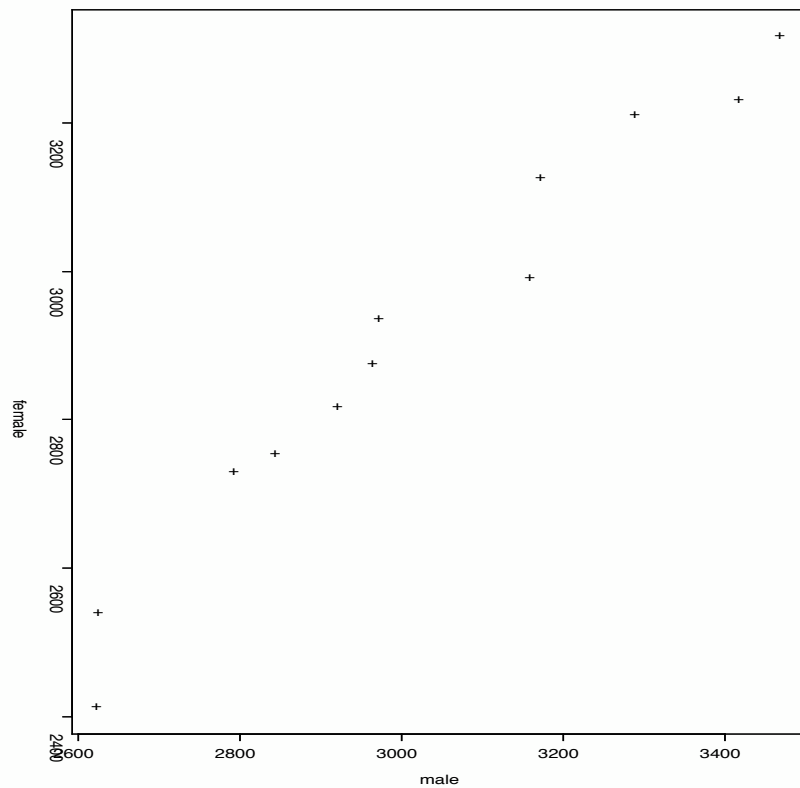


2.1 Two Samples

We can also extend this trick to two samples. If we wish to see if two samples come from the same distribution all we do is plot the empirical distribution functions against each other. A straight line will result if they have the same parent distribution. In fact we can

The data below gives 24 birth-weights, 12 are boys and 12 girls. It seems reasonable to assume that they come from the same parent distribution so we plot the two EDFs. The resulting plot appears to confirm our view.

male	2968	2795	3163	2925	2625	2847	3292	3473	2628	3176	3421	2975
female	3317	2729	2935	2754	3210	2817	3126	2539	2412	2991	2875	3231



Of course life being what it is we usually get unequal sample sizes. The simple solution is to delete terms at random. We will just consider *parametric inference* where we know the form of the distribution except for some parameters. So we might assume a binomial distribution but not know θ the probability of success. Or if we have a sample of baby birth weights we might assume that these are normal with unknown mean and variance.

2.1.1 Parametric Inference

In parametric inference we assume, as said, above that the joint distribution of X_1, \dots, X_n say

$$f(x_1, \dots, x_n, \theta_1, \theta_2, \dots, \theta_p)$$

is known except for some set of parameters $\theta_1, \dots, \theta_p$. Our aim is to estimate the parameters or to make inferences about them. We will usually select some statistic T to estimate a particular θ of interest.

Our interest is both in finding a plausible value for the parameter **and** specifying an error. This will mean we need to find the probability distribution of T : a difficult

exercise. To make this a tractable proposition we assume that the data has been collected in such a way that it is representative of the population and the values are independent. This means we may assume that X_1, \dots, X_n are independent and in consequence the joint distribution is

$$f(x_1)f(x_2)\dots f(x_n) = \prod_{i=1}^n f(x_i).$$

These assumptions are critically important and are often disregarded. It is common to end up with a grab set which is neither random or representative.

2.1.2 Estimates

If we wish to estimate a parameter θ by T or more usually $\hat{\theta}$ what properties would we like our estimate to have? Remember this is a stochastic problem so we cannot be exactly right every time. Traditional statisticians look at mean square error $E[(\hat{\theta} - \theta)^2]$ and usually seek to minimize this quantity. We can approach this from a component viewpoint giving rise to the usual properties we seek from estimates are

- *Unbiasedness*
If $E[\hat{\theta}] = \theta + b(\theta)$ then we call $b(\theta)$ the bias. If $b(\theta) = 0$ then we say $\hat{\theta}$ is unbiased. For many estimates $\hat{\theta}$ tends to zero as the sample size increases.
- *Consistency*
If $P[|\hat{\theta} - \theta| \leq \epsilon] \rightarrow 0$ as the sample size tends to infinity then we say our estimate is consistent. For an unbiased estimate $var(\hat{\theta}) \rightarrow 0$ as $n \rightarrow \infty$ will ensure this.
- *Sufficiency*
If the estimate contains all the information in the sample relevant to the parameter we say that the estimate is sufficient.
- *Efficiency*
If there are several competing estimates we need to make a sensible choice between them. This is done by picking the one with minimum mean square error. Traditionally we consider unbiased estimates in which case we compare variances. People use relative efficiency which is the reciprocal of the ratio of the variances. Interestingly these is a lower bound to the variance of an estimate called the Cramer Rao bound.
Note : The m.s.e $= E[(\hat{\theta} - \theta)^2] = var(\hat{\theta}) + b(\theta)^2$

These properties are all very well but how does one obtain good estimates? There are a variety of methods.

2.1.3 Method of Moments

A simple if not very efficient technique is to use the fact that the sample moments tend to the population ones. All one does is equate these working on the assumption that the

Figure 1: Distribution of losses for Hurricanes

population moments are a function of the parameters. Thus one solves

$$\frac{1}{n} \sum_{i=1}^n x_i^k = E[X^k]$$

Usually we only need a couple of equations and use

$$\bar{x} = \mu(\theta_1, \dots, \theta_p)$$

and

$$s^2 = \sigma^2(\theta_1, \dots, \theta_p)$$

2.1.4 Hurricane Example

Table 1 gives the losses caused by Hurricanes in the US (Multiply by '000). The Loss column gives the loss in the year of the storm; the Adj. loss column adjusts the rate to current dollars using the inflation factor given.

The summary statistics are

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2730	19120	51000	189000	196800	1638000

If we assume, unrealistically, that the distribution of losses is exponential with mean $\frac{1}{\lambda}$ then using the method of moments we find the parameter since

$$\frac{1}{\lambda} = 189021$$

If we assume a Gamma distribution $\frac{1}{\Gamma(k)} \lambda^k x^{k-1} e^{-\lambda x}$ with mean $\frac{k}{\lambda}$ and variance $\frac{k}{\lambda^2}$ we solve

$$\frac{k}{\lambda} = 189021 \text{ and } 321613^2 = \frac{k}{\lambda^2}$$

2.1.5 Likelihood

We begin by considering the idea of likelihood. Suppose we have as sample x_1, \dots, x_n with density $f(x_1, \dots, x_n, \theta_1, \theta_2, \dots, \theta_p)$. Since we have already observed the data we can think of the density as a function of the parameters rather than the sample values $\theta_1, \dots, \theta_p$. The resulting function is then called the *likelihood function* $L(\theta_1, \dots, \theta_p)$. We view the likelihood as the central concept in statistics and refer you to discussion of the strong and weak likelihood principles.

Number	Loss	Year	Factor	Adj. Loss(x)
1.	2000	1977	1.365	2730
2.	1380	1971	2.233	3082
3.	2000	1971	2.233	4466
4.	2000	1964	3.383	6766
5.	2580	1968	2.761	7123
6.	4730	1971	2.233	10562
7.	3700	1956	3.912	14474
8.	4250	1961	3.612	15351
9.	5400	1966	3.145	16983
10.	4500	1955	4.085	19383
11.	5000	1958	3.806	19030
12.	14720	1974	1.719	25304
13.	7900	1959	3.685	29112
14.	13500	1971	2.233	30146
15.	22697	1976	1.486	33727
16.	12000	1964	3.393	40596
17.	8300	1949	4.989	41409
18.	13000	1959	3.685	47905
19.	10450	1950	4.727	49397
20.	12500	1954	4.208	52600
21.	32300	1973	1.855	59917
22.	57911	1980	1.090	63123
23.	23000	1964	3.383	77809
24.	25200	1955	4.085	102942
25.	34800	1967	2.966	103217
26.	32200	1957	3.841	123680
27.	122070	1979	1.148	140136
28.	119189	1975	1.611	192013
29.	97853	1972	2.028	198446
30.	67200	1964	3.383	227338
31.	91000	1960	3.621	329511
32.	100000	1961	3.612	361200
33.	165300	1969	2.551	421680
34.	122050	1954	4.208	513596
35.	129700	1954	4.208	545778
36.	309950	1970	2.421	750389
37.	752510	1979	1.148	863881
38.	500000	1965	3.276	1638000

Table 1: Distribution of losses for Hurricanes

2.1.6 example

If we toss a coin twice and observe one head and one tail then if the probability of a head is θ we have $L(\theta) = \theta(1 - \theta)$. It makes sense to make the likelihood as large as possible since we have observed H and T. To do so we choose the parameter θ to be 0.5. Estimates which are chosen in this way to maximise the likelihood are called maximum likelihood estimators and have many nice properties. Note you will often find authors maximizing the log likelihood $\ell(\theta_1, \dots, \theta_p)$.

Suppose we tossed our coin 3 times and got 1 head. Then

$$L(\theta) = \binom{3}{1} \theta(1 - \theta)^2$$

or

$$\ell(\theta) = \log(L(\theta)) = \log \binom{3}{1} + \log \theta + 2 \log(1 - \theta)$$

So

$$\frac{\partial \ell}{\partial \theta} = \frac{1}{\theta} - \frac{2}{1 - \theta}$$

and $\hat{\theta} = 1/3$

2.1.7 example

In the hurricane case the likelihood is, on the assumption of a Gamma distribution,

$$L(\theta_1, \dots, \theta_p) = \prod_{i=1}^n \frac{1}{\Gamma(k)} \lambda^k x_i^{k-1} e^{-\lambda x_i}$$

or

$$\ell(\theta_1, \dots, \theta_p) = \sum_{i=1}^n [k \log \lambda + (k - 1) \log x_i - \lambda x_i - \log \Gamma(k)]$$

To get the values of the parameters we find the values which will maximise the function. This is not a difficult case but not one that we can do analytically. This is often the case especially when we need several parameters. Usually we have the option of numerical maximisation.

2.2 Why Likelihood?

Likelihood estimates have several desirable properties. They arise from a coherent theory and can usually be found by systematic methods. If an estimate is possible with desirable properties then the mle's (maximum likelihood estimates) are functions of these estimates. More to the point the likelihood estimates are asymptotically unbiased having normal with known variance. That is

$$\hat{\theta} \sim N(\theta, v^2)$$

where

$$v^2 = 1/E\left[\left(\frac{\partial \ell(\theta)}{\partial \theta}\right)^2\right] = -1/E\left[\left(\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right)\right]$$

2.2.1 Some Examples

1. The Discrete Uniform distribution $P[X = k] = \frac{1}{n}$ $k = 1, 2, \dots, n$ has mean $\mu = E[X] = \frac{n+1}{2}$. One way of estimating n is to set $\bar{x} = \mu = E[X] = (n+1)/2$
2. The Binomial distribution $P[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$ $k = 0, 1, 2, \dots, n$. Here we know $\mu = np$ so setting $\bar{x} = np$ gives us p
3. For the normal

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

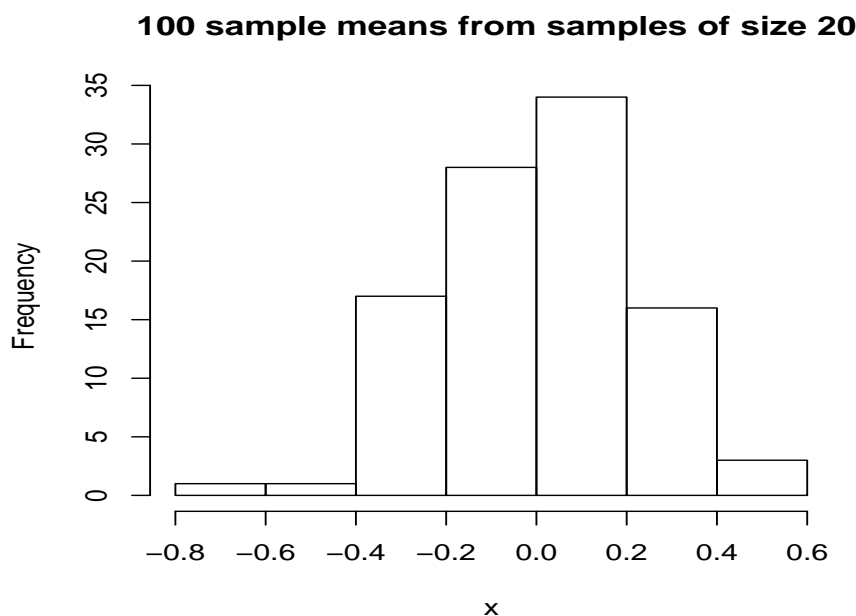
life is a bit more complex as we have two parameters but $\bar{x} = \mu$ and $s^2 = \sigma^2$ will do. Here

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2$$

For simplicity I shall write the estimate of a parameter θ as $\hat{\theta}$

2.3 Sampling distributions

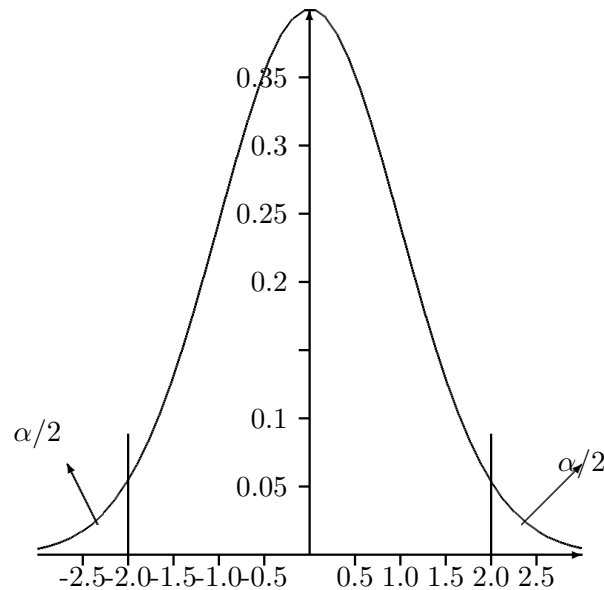
Of course having an estimate is fine but we really need some idea of the errors we may make. To do this we need to consider the variation from sample to sample. If we take the sample mean \bar{x} we expect differing samples to give different values for \bar{x} . For example 100 samples of size 20 gave 100 values of \bar{x} displayed below



At this point theory comes to our aid. We can determine the theoretical distribution of \bar{x} in some cases, but we also have the central limit theorem which says that the distribution of \bar{x} is normal or more precisely the distribution of

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \quad (1)$$

is standard normal, that is has a normal distribution with mean 0 and variance 1, see the plot below.



Suppose we decide that we will think of the unusual values of \bar{X} as those in the tails of the distribution. We quantify this by assigning a small probability α which we split between the two tails. If we are given α we can easily find $z_{1-\alpha/2}$ and $z_{\alpha/2}$ from the relation

- $P[z \leq z_{\alpha/2}] = \alpha/2$
- $P[z \leq z_{1-\alpha/2}] = 1 - \alpha/2$

This means

$$P \left[z_{\alpha/2} \leq z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \leq z_{1-\alpha/2} \right] = 1 - \alpha$$

or for \bar{X} This means

$$P \left[\bar{x} - z_{\alpha/2}\sigma/\sqrt{N} \leq \mu \leq \bar{x} + z_{1-\alpha/2}\sigma/\sqrt{N} \right] = 1 - \alpha$$

We are asserting that the interval $\bar{x} \pm z_{1-\alpha/2}\sigma/\sqrt{N}$ covers the unknown true value μ $100(1 - \alpha)\%$ of the time.

Note: from the symmetry of the normal distribution $-z_{-\alpha/2} = z_{1-\alpha/2}$.

Common values of α are 0.05 and 0.01 so from tables

α	$z_{1-\alpha/2}$
0.10	1.645
0.05	1.96
0.01	2.58

Here the estimate of μ is $\hat{\mu} = \bar{x}$ and the confidence interval is $\bar{x} \pm z_{1-\alpha/2}\sigma/\sqrt{N}$

Of course there is an immediate drawback, one has to know σ .

If the mean is not known it is almost certain that the variance is also unknown and our confidence interval is of limited utility. However experience tell us that

1. Provided n exceeds 50.
2. We have a sample from a Normal population

then we can use our variance estimate s^2 or rather the square root

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

in our expression to get an approximate confidence interval $\bar{x} \pm z_{1-\alpha/2}s/\sqrt{N}$

2.3.1 example

A sample of 60 observations

```

14.062 13.110  7.068 14.720 13.414  7.762
 9.229  3.274  6.316 13.783 13.232  8.436
14.905 16.026 14.907  8.924  8.302 15.577
20.487 14.644 15.977 12.907  7.538 12.671
 8.717 18.972 20.949  5.256 13.501 18.237
12.635 10.568  5.170  6.307 12.410  6.355
14.512 10.037 14.610  6.572 15.644 21.482
10.829  8.465  6.071  2.651  0.285  4.230
18.312  7.642 11.301 11.372  9.510 14.329
17.143 14.725  4.662 24.015 10.902 12.902

```

Has mean $\bar{x} = 11.64256$ and $s = 5.036942$ Thus our point estimate of the mean is 11.64256 while a 95% confidence interval is $11.64256 \pm 1.96 \times 5.036942/\sqrt{60}$ or

$$11.64256 \pm 1.274522$$

Suppose we have the data below (from R)

```

13.259 -1.986  6.008  2.327  4.136  7.141  3.781  4.494  6.175  2.319
 7.845 -2.140  3.995  4.567  5.921  5.405  5.459  5.209  3.425  6.549
 2.631  4.622  4.501  2.250  5.043  3.476  5.320  6.948  6.448  8.186

```

```

2.121  7.125  5.108 -1.216  7.064  6.609  3.849 11.470  3.115  2.939
2.135  7.225  3.430  4.739  1.594  3.663  6.791  9.431  1.893  7.867
> mean(x)
[1] 4.80532
> sd(x)
[1] 2.915337

```

Then a 95% confidence interval for the mean μ is

$$4.80532 \pm 1.96 \times \frac{2.915337}{\sqrt{50}}$$

or 4.80532 ± 0.81 A 98% interval is

$$4.80532 \pm 2.33 \times \frac{2.915337}{\sqrt{50}}$$

or 4.80532 ± 0.96 These intervals give us a chance to gauge the sample size. Suppose we want to estimate the mean μ with an accuracy of at least 1.0 with a probability of at most 0.95. This implies that we need a 95% confidence interval of width 1.0. So $1.96 \frac{\sigma}{\sqrt{n}} = 0.5$ which implies $n = (1.96\sigma/0.5)^2$. If we know σ , say 2.9 then $n = 129.2$ or $n=130$.

3 Student's t

It is natural to ask what can be done for small samples. This was a hot topic in the early part of the last century and was finally solved by Gosset, a brewer. He published his results under the pseudonym Student, hence the name.

Student managed to determine the exact distribution of

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The probability density is now known as the Student t distribution and is so useful it is tabulated. All we have to do is use the t distribution in place of the normal. That is we use

$$\bar{x} - t_{1-\alpha/2} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{1-\alpha/2} \frac{s}{\sqrt{n}} \quad (2)$$

Here $t_{1-\alpha/2}$ is the value such that

$$p[t \leq t_{1-\alpha/2}] = 1 - \alpha/2$$

The only complication is that the distribution depends on the sample size n . To find the appropriate value you need the extra parameter ν or the degrees of freedom. In our case $\nu = n - 1$.

3.0.2 Example

1. The results of a turbidity test on 15 samples of testing sand gives $\bar{x} = 25.31$ and sample variance $s^2 = 1.58^2$. With $n=15$ and $\alpha = 0.05$, $t_{0.975} = 2.145$ the desired interval is

$$25.31 \pm 2.145 \times 1.58/\sqrt{15}$$

so 25.31 ± 0.875

2. The fat content of $n=10$ randomly selected hot dogs is given below.

25.2	21.3	22.8	17.0	29.8
21.0	25.5	16.0	20.9	198.5

The summary statistics are $\bar{x} = 21.9$ and $s = 4.134$ so as $t_{0.975} = 2.262$ a 95% confidence interval is

$$21.9 \pm 2.262 \times \frac{4.134}{\sqrt{10}} \text{ or } 21.90 \pm 2.96$$

3.1 Binomial

I think the of of the most useful uses of confidence intervals is with proportions. Suppose we are interested in the proportion of type A's in a population. We take a sample of size n and count the number, say R , that are type A. If we assume R is Binomial then we use the Normal approximation to the Binomial, that is

$$z = \frac{\frac{R}{n} - p}{\sqrt{p(1-p)/n}}$$

So we can write

$$p[-z_{1-\alpha/2} \leq z = \frac{\frac{R}{n} - p}{\sqrt{p(1-p)/n}} \leq z_{1-\alpha/2}] = 1 - \alpha$$

This can be rewritten, after some algebra and approximation, as

$$p[\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}] = 1 - \alpha \quad (3)$$

where $\hat{p} = R/n$. Thus our $(1 - \alpha)100\%$ confidence interval is

$$\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \quad (4)$$

3.1.1 example

1. Of 270 people asked at random 189 said that they would be willing to pay extra for a commodity to be packed in tamper proof package. Then $\hat{p} = 189/270 = 0.7$. A 95% confidence interval for p is

$$0.7 \pm 1.96 \sqrt{\left(\frac{0.7 \times 0.3}{270}\right)} \text{ or } 0.7 \pm 0.055$$

2. A sample of cycle helmets reports that of 37 helmets subjected to an impact test 24 showed damage. We aim to estimate the probability p that a helmet is damaged so that a 95% confidence interval for p has width at most 0.1. Then

$$0.05 = z_{1-\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$$

using our estimate we can write

$$n = \frac{4(1.96 \times 0.5)^2}{0.1^2} = 384$$

3.1.2 The variance

Given a normal sample, say

36.53 35.91 36.47 35.69 39.82
 30.71 40.19 38.74 42.26 33.37
 36.57 35.62 35.11 35.40 36.97
 38.63 39.24 37.65 34.71 37.53

we may be interested in the variance σ^2 . A possible point estimate is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 6.652646$$

but we might also like a confidence interval. The problem is the distribution of s^2 . Fortunately we know that for a random sample from a normal distribution with variance σ^2

$$v = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2}$$

has a Chi-squared distribution with $n-1$ degrees of freedom. Thus given tables we can find a and b such that

$$P[v \leq a] = \frac{\alpha}{2} \text{ and } P[v \leq b] = 1 - \frac{\alpha}{2}$$

we have

$$P\left[a \leq v = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^2} \leq b\right] = 1 - \alpha$$

or a $100(1 - \alpha)\%$ confidence interval

$$\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{b} \leq \sigma^2 \leq \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{a}$$

For our example we have 20 observations with $\sum_{i=1}^n (x_i - \bar{x})^2 = 192.9267$ For $\alpha = 0.05$ $a = 16.04707$ and $b = 45.72229$ so

$$4.2195 \leq \sigma^2 \leq 12.0226$$

4 Hypothesis Testing

Setting up and testing hypotheses is seen in most courses as an essential part of statistical inference. In order to formulate such a test, an assertion about a distribution is proposed. Such an assertion is called a statistical hypothesis. Typically the hypothesis has been put forward, either because it is believed to be true or because it is to be used as a basis for argument. For example, claiming that a new drug is better than the current drug for treatment of the same symptoms can be expressed as a statement about the probability that a patient is cured.

In each problem considered, the question of interest is simplified into two competing hypotheses between which we have a choice; the null hypothesis, denoted H_0 , against the alternative hypothesis, denoted H_1 , which is the compliment of the null. These two competing hypotheses are not however treated on an equal basis, special consideration is given to the null hypothesis. Usually the experiment has been carried out in an attempt to disprove or reject a particular hypothesis, the null hypothesis, for example,

H_0 : there is no difference in taste between coke and diet coke

against

H_1 : there is a difference

Of the two hypotheses the null is usually simple in that if it is true the underlying distributional assumption is simpler than under H_1 . Indeed the null is often a *simple hypothesis*, that is is a hypothesis which completely specifies the population distribution. Hypotheses which are not simple are said to be composite.

4.0.3 Example

1. H_0 : X is Binomial (100,1/2) i.e. p is specified
 H_1 : X is Binomial (100, p) $p \leq 1/2$
2. H_0 : X is $N(5, 20)$ i.e. μ and σ are specified
 H_1 : X is $N(\mu, 20)$ i.e. $\mu > 5$

If we have two competing hypotheses there are two kinds of errors that may arise and the following table gives a summary of possible results .

action	Truth	
	H_0	H_1
Accept H_0	ok	type 2 error
Reject H_0	type 1 error	ok

The sample space of outcomes (x_1, x_2, \dots, x_n) can be split into two parts C and R where $C \cap R$ is empty and $C \cup R$ is the whole space. We choose C - the critical region - to be the set of unlikely points for which

$$P[(x_1, x_2, \dots, x_n) \in C | H_0] = \alpha$$

where α is chosen to be small probability, often called *the size of the test* or *the significance level*. If we observe an set of points in C we have two options

- Either H_0 is false
- Or we have observed an event of small probability.

In conventional testing we assume the second and say we reject the null hypothesis and accept the alternative. By this we mean that it is rational on the evidence to believe that the null is not true. The probability of observing the unlikely event is α the probability of a type I error. The probability of a type II error $\beta = P[\text{accept } H_0 | H_1]$ is generally unknown and needs to be calculated.

If we do not reject the null hypothesis, it may still be false (a type II error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis). For any given set of data, type I and type II errors are inversely related; the smaller the risk of one, the higher the risk of the other.

As dealing with high dimensional spaces is difficult we usually base our tests on a test statistic T computed from the observations. As above we find a critical region C defined by

$$P[T \in C | H_0] = \alpha$$

Some workers prefer the *p-value*. The probability value (p-value) of a statistical hypothesis test is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone the assumption that the null hypothesis H_0 , is true.

The power of a statistical hypothesis test measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision. In other words, the power of a hypothesis test is the probability of not committing a type II error.

5 Constructing tests

While the apparatus above is reasonable it does not answer the questions as to how one might construct a test. Most statistics book give a list of recipes. The procedure is typically:

- Set up H_0 and H_1
- Pick a suitable test statistic T whose distribution is known under the assumptions of H_1 .
- Choose the size of the test α
- Find the critical region
- Compute T
- If T lies in the critical region reject H_0

Often we can derive such a method from insight into the problem.

5.0.4 As an example:

Suppose we have a population and some proportion p of the population are female. We take a random sample of size N and find that R are female.

Suppose $H_0 : p = p_0$ while $H_1 : p = p_1 \leq p_0$

Making the plausible assumption that the distribution of the number of females is Binomial we have as the mle $\hat{p} = \frac{R}{N}$. Now for large samples we know that, under H_0 \hat{p} is normal with mean p_0 and variance $\sqrt{\frac{p_0(1-p_0)}{N}}$.

A plausible statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{N}}}$$

which is clearly normal and standard normal under H_0 T will be near zero when the null hypothesis is true but negative otherwise. The critical region is $T \leq \text{constant}$ and since T is normal we have $T \leq -z_\alpha$ where z_α is the percentage point of the normal distribution i.e. for a standard normal variable z

$$P[z \leq -z_\alpha] = \alpha$$

5.0.5 A Normal example

Suppose we have a sample of size 100 from a Normal distribution. We wish to test

$$H_0 : \mu = 68 \text{ against } H_1 : \mu \neq 68$$

To simplify matters we assume that the standard deviation of the population is $\sigma = 16$.

A possible statistic is \bar{X} which is Normal $N(\mu, \sigma^2/n)$. Rather simpler is the standardized random variable

$$z = \frac{\bar{X} - 68}{\sigma/\sqrt{n}}$$

which we know is standard Normal .

Figure 2: Normal Density function

Now if the true mean is not 68 then z will be very different from zero. The distribution of z is shown in the diagram 5.0.5 and we see that the two areas in the tails must total α . A little inspection gives us the critical regions as

$$X < -z_{1-\alpha/2} \text{ and } X > z_{1-\alpha/2}$$

Now if we choose $\alpha = 0.05$ then we will reject H_0 when $X < -1.96$ or $X > 1.96$

In our case $\bar{X} = 68.04$. We know that $\sigma = 16$ so $z = 0.25$. This is not in the critical region so we accept H_0

This is somewhat unrealistic since if we are unsure of μ it is most unlikely that we know σ !

Here we have a large sample and for large samples (exceeding 50) we can use an estimate. Here $\hat{\sigma} = 13.724$ so

$$T = \frac{\bar{X} - 68}{\sigma/\sqrt{n}} = 0.02124337$$

and in consequence we accept H_0

What is the critical region for the alternative $H_1 : \mu \neq 68$?

For a small sample we would then have to find the distribution of

$$t = \frac{\bar{X} - 68}{\hat{\sigma}/\sqrt{n}}$$

in order to compute the size of the critical region. In fact we know that the distribution of t has a Student's t distribution with $n - 1$ degrees of freedom so all we need to do is to find the critical region using tables of t rather than Normal tables.

5.1 Normal small sample case

A sample of 10 batteries are randomly selected from a production batch and their lifetimes found. The mean lifetime is 30.3 and the estimated variance is 16.08456. The manufacturer claims a lifetime of 36 months. Suppose we assume a normal population with mean μ and test

$$H_0 : \mu = 36 \text{ against } H_1 : \mu < 36$$

Then

$$t = \frac{30.3 - 36}{\sqrt{16.08456/10}} = -4.494323$$

Now we find the critical region using the t distribution with $10-1=9$ degrees of freedom. The critical value is, for a test size of 0.05, -1.833 (check this!) so we have a value in the critical region and we reject H_0

5.1.1 A Binomial/Normal example

In a population some proportion p of the population are female. We take a random sample of size N and find that R are female.

Suppose $H_0 : p = p_0$ while $H_1 : p = p_1 \leq p_0$

Making the plausible assumption that the distribution of the number of females is Binomial we have as the maximum likelihood estimate of p $\hat{p} = \frac{R}{n}$. Now for large samples we know that, under H_0 \hat{p} is normal with mean p_0 and variance $\sqrt{\frac{p_0(1-p_0)}{n}}$.

A plausible statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

which is clearly normal and standard normal under H_0 T will be near zero when the null hypothesis is true but negative otherwise. The critical region is $T \leq \text{constant}$ and since T is normal we have $T \leq -z_\alpha$ where z_α is the percentage point of the normal distribution i.e. for a standard normal variable z that is

$$P[z \leq -z_\alpha] = \alpha$$

5.1.2 Example

A princess is given a bucket of 40 frogs to kiss. It is well known that the probability of a frog being a prince and hence being returned to his usual handsome a dashing state is p . If after the experiment we have 5 princes and 35 frogs can we assume $p = 0.2$? To compare $H_0 : p = 0.2$ with $H_1 : p \neq 0.2$ we use the statistic above

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{5/40 - 0.2}{\sqrt{0.2 \times 0.8/40}} = -1.185854$$

For a two sided test ($\alpha = 0.05$) the critical region is $z \leq -1.96$ and $z \geq 1.96$ in which case we accept H_0 Note

- The distribution of the number of princes is really hypergeometric but we are assuming the population of frogs is very large and we can use a Binomial approximation
- We use the Normal approximation to the Binomial this is usually good for sample sizes over 25 or so except when we have very small probabilities.
- You might think that the alternative is a bit odd
- What is the power of this procedure?

5.2 Sample size choice

In many situations we can use our definitions of the type 1 and type 2 probabilities to specify the sample size we require.

5.2.1 A crossover trial

Suppose we wish to test the effects of two kinds of medication A and B on reducing blood pressure in males. We intend to treat n patients for five weeks on A and five weeks on B. The order of application will be randomized. The response is the average blood pressure in the third week of each treatment.

If we pick $\alpha = 0.05$ and $\beta = 0.1$ how large a sample do we need?

We assume that the responses $A_i, B_i, i = 1, 2, \dots, n$ are normal and look at the differences $D_i = A_i - B_i, i = 1, 2, \dots, n$. The test is $H_0 : \mu_d = 0$ against $H_1 : \mu_D > 0$. Now the background to this experiment is that A is the current treatment and we expect to see a change from A to B whose size is about $1/2$ a standard deviation.

Back to definitions

$$0.05 = \alpha = P[z = \frac{\bar{D} - \mu}{\sigma/\sqrt{n}} > 1.645 | H_0 : \mu = 0]$$

and

$$0.1 = \beta = P[z = \frac{\bar{D} - \mu}{\sigma/\sqrt{n}} > -1.28 | H_1 : \mu = \frac{\sigma}{2}]$$

So

$$1.645 \times \frac{\sigma}{\sqrt{n}} = \sigma/2 - 1.28 \times \frac{\sigma}{\sqrt{n}}$$

So $n = (2 \times 2.925)^2$ that is $n = 35$

5.2.2 A Binomial/Normal case

A medic knows that of patients admitted to hospital for cardiac problem 60% will be readmitted on an emergency basis within 2 months. She believes that treatment with X will reduce this readmission level by half, that is to 30%. To check her theory she must experiment on patients and to gain permission to do so she must show that her experiment has a reasonable chance of detecting a change and uses the minimum number of patients.

This could be framed as a Binomial, with p the probability of readmission. Suppose we take n patients and administer the drug to them. We set up the hypotheses

$$H_0 : p = 0.6 \text{ against } H_1 : p < 0.6$$

We observe R of our n treated patients readmitted. We will assume that R is Binomial and that we can approximate R by a Normal distribution.

Now how to specify the sensitivity of our procedure. We ask that the type 2 error probability be

$$\beta = P[R \geq C | p = 0.3] = 0.1$$

You should check that you understand this!

I am going to specify the test size as $\alpha = 0.05$

The critical region is of the form $R \leq C$ and assuming normality we have the definition of test size

$$\alpha = P[R \leq C | p = 0.6] = 0.05$$

Using approximate normality

$$0.05 = \alpha = P[R \leq C | p = 0.6] = P\left[z = \frac{R - n \times 0.6}{\sqrt{n \times 0.6 \times 0.4}} \leq -1.65 | p = 0.6\right]$$

and for the type 2 error

$$0.1 = \beta = P[R \geq C | p = 0.3] = P\left[z = \frac{R - n \times 0.3}{\sqrt{n \times 0.3 \times 0.7}} \geq 1.28 | p = 0.3\right]$$

so we now have

$$C = n \times 0.6 - 1.64 \times \sqrt{n \times 0.6 \times 0.3} = n \times 0.3 + 1.28 \times \sqrt{n \times 0.7 \times 0.3}$$

We solve to get $n = 22$

6 Likelihood based schemes

Such a hoc approaches are all very well but we need a method which will provide good if not optimum tests in a wide variety of situations The most flexible approach is to consider the ratio of the likelihoods under each hypothesis

$$\lambda = L(H_0)/L(H_1)$$

For simple hypotheses this ratio gives the Neyman-Pearson testing procedure and the most powerful test. Simplifying

$$\lambda = L(H_0)/L(H_1) \leq \text{a constant}$$

for two simple hypotheses gives the critical region of the most powerful test. We then have to find the distribution of the statistic so we can fix the size of the test. When the hypotheses also postulate parameters, as is usual, the we substitute the mles into the ratio. Thus if under H_0 we specify p parameters θ_p while under H_1 we specify $q > p$ parameters θ_q then we consider

$$\lambda = L(\hat{\theta}_p)/L(\hat{\theta}_q)$$

We can of course look at the log of this quantity (for a monatone ratio)

$$\log \lambda = \ell(\hat{\theta}_p) - \ell(\hat{\theta}_q)$$

or as is more usual

$$D = 2 \left[\ell(\hat{\theta}_q) - \ell(\hat{\theta}_p) \right]$$

the deviance. We choose this form because we know that the distribution of D is χ^2 with degrees of freedom $q - p$. This is a very valuable result because finding the exact distribution an be at best very difficult and at worst impossible.

7 Subjective Inference

Despite the careful derivations of frequentist inference many people would argue that one always starts an experiment with a prior belief and this is then modified by experience. This subjective view is generally known as Bayesian and can be developed to give an alternative approach to inference.

Suppose we have some parameter of interest θ . Unlike classical inference we have some subjective, or prior belief about this parameter. We quantify this belief as a probability distribution $g(\theta)$ which we call the *prior distribution*. Now we perform some experiment with the aim of gaining *more* information about θ . Suppose the result of our experiment is some data

$$(x_1, x_2, x_3, \dots, x_N)^T = \mathbf{x}$$

This has a likelihood of the form

$$f(\mathbf{x}|\theta)$$

We use the above notation since we perform the experiment *given our prior belief about θ* .

We now use Bayes theorem to modify our belief

$$h(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)g(\theta) \quad (5)$$

where $h(\theta|\mathbf{x})$ is the posterior distribution, that is the distribution of θ after our view has been modified by experiment. We get the constant of proportionality, k , by noting that a density integrates to one, so

$$1/k = \int f(\mathbf{x}|\theta)g(\theta)d\theta$$

This is just an analog of Bayes theorem as discussed above.

$$p[A|B] = p[B|A]p[A]/p[B]$$

7.0.3 An example

A coin is tossed, the probability of being a head, $P[H]$, is some unknown value θ . We inevitably have some prior belief about θ which we need to quantify.

One way of doing this is to use a probability distribution to quantify our belief. This distribution $g(\theta)$ is the *prior distribution* is assumed to contain our prior beliefs about the parameter θ . A possible distribution for θ is the Beta distribution with density

$$f(x) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \text{ for } 0 \leq x \leq 1 \quad (6)$$

This can take a variety of shapes depending on the choice of a and b as can be seen in figure 3

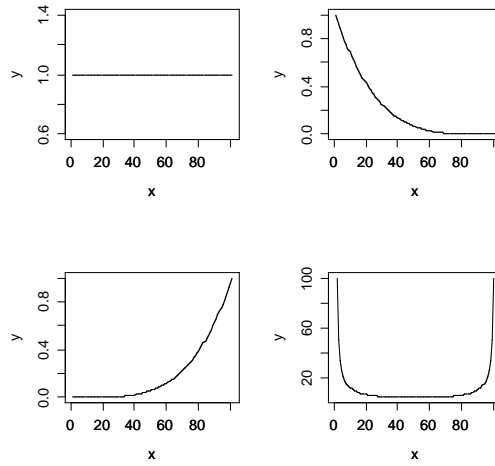


Figure 3: Possible Beta distributions

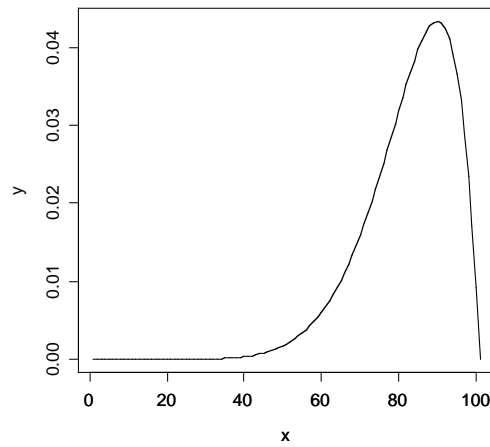


Figure 4: A possible distribution for the prior belief in the probability θ of a head

When we conduct the experiment of tossing the coin N times we have the probability, (likelihood when x the number of heads is known)

$$L(x|\theta) = \binom{N}{x} \theta^x (1 - \theta)^{N-x}$$

Using Bayes theorem we have

$$h(\theta|x) \propto \binom{N}{x} \theta^x (1 - \theta)^{N-x} g(\theta)$$

so if we had chosen

$$g(\theta) = \frac{\theta^{a-1} (1 - \theta)^{b-1}}{\text{Beta}(a, b)}$$

we would have

$$h(\theta|x) \propto \theta^{(a+x-1)} (1 - \theta)^{(N-x+b-1)}$$

This latter function is the posterior distribution of the parameter θ given the data - *it is our belief given the data*. Note the constant of proportionality can be obtained easily since we know that the distribution must integrate to 1. So if we start with a prior which is

$$g(\theta) = \frac{\theta^2 (1 - \theta)^2}{\text{Beta}(3, 3)}$$

and we toss a coin 25 times and observe 15 heads then

$$h(\theta|x) \propto \theta^{17} (1 - \theta)^{12}$$

Some integration shows that

$$h(\theta|x) = \frac{\theta^{17} (1 - \theta)^{12}}{\text{Beta}(18, 13)}$$

contain our prior beliefs about the parameter θ . The prior and posterior distributions are shown in figure 5

Since the mean and variance of the Beta distribution, parameters a and b are

$$\mu = \frac{a}{a+b} \quad \sigma^2 = \frac{ab}{(a+b)^2(a+b+1)}$$

we can get some idea of the reduction in variance obtained by the experiment.

We are doing something rather subtle here, we are using basic probability theory to modify our prior beliefs to end up with a posterior distribution.

7.0.4 Exercises

Suppose we toss a coin 100 times. What is our posterior distribution when

1. $g(\theta) = 1/[\theta(1 - \theta)]$
2. $g(\theta) = 1$

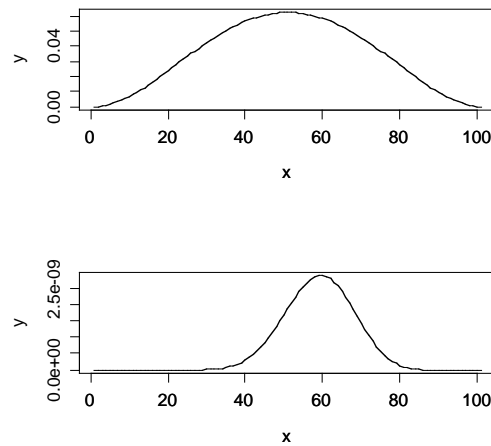


Figure 5: Prior and posterior distributions for coin tossing

7.1 Choice of prior

As you can imagine the choice of prior can be very difficult. Even in a simple case such as a Normal mean μ if we wish to express ignorance the obvious choice for g is a constant. This however would not be a proper distribution. To get around this point Bayesians will usually allow the use of an improper prior!

The other difficulty is the integration required to obtain the posterior density. We can overcome this to some extent by choosing a prior in sufficiently clever way. It is not very restrictive to choose a prior from a conjugate class, that is we choose a prior from a family related to the likelihood. As we have seen the Beta gives us a fairly simple solution when using the Binomial. We can summarize

Posterior	Likelihood	prior
Beta	Binomial	Beta
Gamma	Poisson	Gamma
Gamma	Gamma	Gamma
Normal	Normal	Normal(Mean)

Suppose we have a Normal likelihood, say with mean μ and variance σ^2 . If we choose as a prior for μ a normal distribution mean ν variance τ^2 then the posterior is also normal with

- mean $\frac{1/\tau^2}{1/\tau^2+n/\sigma^2}\nu + \frac{n/\sigma^2}{1/\tau^2+n/\sigma^2}\bar{x}$
- variance $\frac{1}{1/\tau^2+n/\sigma^2}$

There have been recent developments which make a Bayes approach rather more plausible but they are rather technical.

7.2 Confidence intervals etc

One rather nice aspect of the Bayesian view is that the parameter θ has a distribution and so we can make statements like

the probability that θ lies between a and b is α .

If we have a posterior we usually work with the HPD (Highest posterior density). This means we pick region of size $1 - \alpha$ for which $g(\theta)$ is a maximum.

As an example suppose X is uniform on $(0, \theta)$ and as a prior for θ we have

$$g(\theta) = \theta^{-2} \quad \theta > 1$$

If we have one observation x then the posterior is

$$h(\theta) = 2x^2/\theta^3 \quad \theta > x > 1$$

The $100(1 - \alpha)$ percent confidence interval is (x, θ^*) where $\theta^* = x/\sqrt{\alpha}$

7.3 The Bayes Estimate

The whole posterior may be difficult to handle so one may prefer to use a Bayes estimate. This is defined in terms of the possible loss. Suppose our estimate of θ is $\hat{\theta}$. We define a loss function $loss(\theta, \hat{\theta})$ which gives the costs of errors in estimation. Thus for example,

$$loss(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

or

$$loss(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$$

Then the Bayes estimate is the $\hat{\theta}$ which minimizes the expected loss (sometimes known as the risk) In the case of the squared error loss $(\theta - \hat{\theta})^2$ we can show that the estimate is the posterior mean while for $|\theta - \hat{\theta}|$ the Bayes estimate is the posterior median.

7.4 Predictive distribution

We have been a little slapdash in our notation as we should really write the posterior as $h(\theta|\mathbf{x})$ that is given the data. Suppose we consider taking a further observation y . We know the likelihood so $f(y|\theta)$ is known. We can as we wish remove the θ dependence using our posterior. The resulting distribution is known as the predictive distribution. We attempt an illustration.

7.4.1 Example

20 windows in a high rise office block broke in the first year of occupancy of the building. The question was how many of these were due to a specific defect D. If they were caused by D then the manufacturer of the windows will replace them, otherwise they will not.

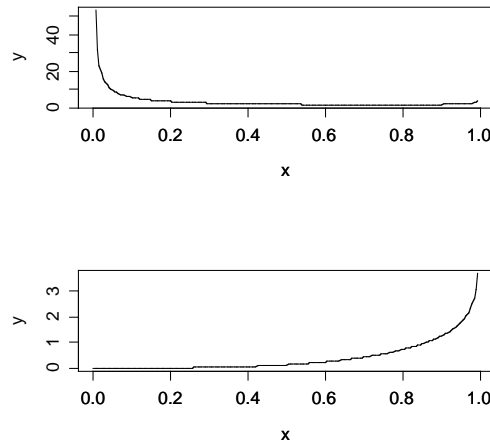


Figure 6: Prior and posterior distributions for windows

Only in 4 of the 20 widows was glass available for analysis (so we will assume a sample of 4 in 20!). All 4 were found to have broken because of D.

In the subsequent legal wrangle a glass expert claimed that the distribution of θ the probability a window suffers from D is

$$g(\theta) = \theta^{-3/4}(1 - \theta)^{-1/4} \quad 0 \leq \theta \leq 1$$

For a sample of 20 with 4 with defect D the likelihood is θ^4 (ie. all broken) so the posterior is

$$h(\theta) \propto \theta^{13/4}(1 - \theta)^{-1/4}$$

The prior and posterior are shown in figure ?? The distribution of Z the number of defectives is (given θ)

$$\binom{16}{z} \theta^z (1 - \theta)^{16-z}$$

The predictive distribution is then

$$f(z) \propto \int_0^1 \theta^{z+13/4}(1 - \theta)^{63/4-z} d\theta$$

Eventually we have

$$p[Z = k] = p[Z = k + 1] \times \frac{k + 1}{k + 4.25} \times \frac{15.75 - k}{16 - k}$$

This has a mean of 13.6

8 Special functions

8.1 The Gamma function

The Gamma function is defined as

$$\Gamma(z) = \int_0^{\infty} t^{z-1} e^{-t} dt$$

at integer values of t

$$\Gamma(z) = (z - 1)!$$

8.2 The Beta function

The beta function $Beta(a,b)$ is defined as

$$Beta(a, b) = \int_0^{\infty} x^{a-1} (1 - x)^{b-1} dx \quad (7)$$

There are several recursive definitions, the most useful being in terms of the Gamma function

$$Beta(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$$

for integer values

$$Beta(a, b) = \frac{(a - 1)!(b - 1)!}{(a + b - 1)!}$$

9 Nonparametric and robust inference

In much inference and almost all that you have met so far we have assumed that the form of the distributions $f(x, \theta_1, \dots, \theta_p)$ is known except for some parameters $\theta_1, \dots, \theta_p$. This may be an unrealistic assumption, after all a distribution may be close to normal but is rarely exactly normal. In some cases we may feel that actually specifying the form of $f(x, \theta_1, \dots, \theta_p)$ is unrealistic. This is often so when the sample size is small. Our aim to construct methods of inference which do not make distributional assumptions or which are relatively insensitive to these assumptions. This leads us to nonparametric inference and to robust methods as opposed to the classical parametric methods that you met in previous. Another possibility are computationally intensive methods such as the Bootstrap.

We begin with permutation tests. The basic idea is quite simple, we find a plausible statistic, say T , for the hypotheses of interest and a family of permutations of the data such that the probability of each permutation can be found under the null hypothesis. Usually for simplicity we choose permutations which are equally likely under the null hypothesis. We then find the values of T under all possible permutations and then using this empirical distribution either find a critical region and decide whether to accept H_0 , or find a p value, and make our decision using this value.

For example suppose we have the following data set of 12 observations which comes from a long tailed distribution.

2.29828	0.59061	1.39212	2.67940	0.55242	0.05688
2.42898	4.14296	0.01564	2.82009	0.79828	1.33614

We wish to test whether the median of the population is 1, say

$$H_0 : \text{median} = 1 \text{ against } H_1 : \text{median} = 1.$$

Subtracting 1, the hypothesised value of the mean from each value gives a set of deviations

1.29828	-0.40939	0.39212	1.67940	-0.44758	-0.94312
1.42898	3.14296	-0.98436	1.82009	-0.20172	0.33614

One choice of statistic for our test is the sum of the deviations above, say T . Now we set up the permutations of the (modified) data.

If we consider the absolute values of the deviations this will give a value for T . We can now assign one negative sign to each value in turn and obtain another n values for T . We may now assign two negative signs to all pairs of values and obtain more values for T . We then assign minus signs to triples etc. until we have the T value for all $2n$ values. If we assume, as seems reasonable under the null hypothesis that the value of T from each permutation of signs is equally likely we have a distribution of T values. We may then choose between the hypotheses using the permutation distribution of T .

Thus if we had a sample of 4

1.29828 -0.40939 0.39212 1.67940

then we have the following 24 “permutations“ and their corresponding values for T

Permutation				T value
1.29828	0.40939	0.39212	1.6794	3.77919
-1.29828	0.40939	0.39212	1.6794	1.18263
1.29828	-0.40939	0.39212	1.6794	2.96041
1.29828	0.40939	-0.39212	1.6794	2.99495
1.29828	0.40939	0.39212	-1.6794	0.42039
-1.29828	-0.40939	0.39212	1.6794	0.36385
-1.29828	0.40939	-0.39212	1.6794	0.39839
-1.29828	0.40939	0.39212	-1.6794	-2.17617
1.29828	-0.40939	-0.39212	1.6794	2.17617
1.29828	-0.40939	0.39212	-1.6794	-0.39839
1.29828	0.40939	-0.39212	-1.6794	-0.36385
-1.2982	8-0.40939	-0.39212	1.679	4-0.42039
-1.2982	8-0.40939	0.39212	-1.6794	-2.99495
-1.29828	0.40939	-0.39212	-1.6794	-2.96041
1.29828	-0.40939	-0.39212	-1.6794	-1.18263
-1.29828	-0.40939	-0.39212	-1.6794	-3.77919

The (ordered) values of T are

-3.77919 -2.99495 -2.96041 -2.17617 -1.18263 -0.42039 -0.39839

-0.36385 0.36385 0.39839 0.42039 1.18263 2.17617 2.96041

2.99495 3.77919

The actual sample value of T was 2.96041 which exceeds 13 of the 16 values, or is in the top 3 of the 16. Since $3/16 = 0.1875$ we would be unhappy in rejecting H_0 as this is a large P value.

The obvious drawback to this test is the amount of computation required. For the 13 deviations in the original example we require 213 samples each with its own value of the statistic. We can however come up with a very much simpler procedure at the cost of some precision.

9.1 The sign test

We transform the data as follows, if a value exceeds 0 we call it a plus while if it lies below 0 we call it a minus. Of our original deviations we have 7 plus signs and 5 minus signs.

Under H_0 we probability that a value lies above the median and hence maps to a plus is 0.5 and the corresponding probability that we have a minus is also 0.5. Since sequences of plus's and minuses are independent we can find the probability of 5 or fewer from 12. using the binomial. The probability is and so we conclude that there is no reason to reject the null hypothesis that the median is 1.

The procedure described above is known as the *sign test*. In the example we test H_0 : median = q by counting the number of observation that exceed the median (number of plus signs) and the number that lie below the median (number of of minus signs) then under H_0 the number of plus signs out of the total number of signs is binomial. To summarise:

We have n observations and we assume that the median is m

$$H_0 : \text{median} = m \text{ against } H_1 : \text{median} \neq m$$

or for each observation $x_i, i = 1, 2, \dots, N$

$$H_0 : P[X_i \geq m] = P[X_i \leq m] \text{ against } H_1 : P[X_i \geq m] \neq P[X_i \leq m]$$

The test is based on the number of positive or negative signs since we work with deviations from the median $x_i - m$. Under H_0 the probability of a plus or the proportion of positive signs is 0.5 and so the probabilities can be calculated.

- Take the sign which appears most often, suppose there are k of these.
- Calculate the probability of k or more signs assuming a Binomial $p = 0.5$
- If this probability is small enough reject H_0 , otherwise accept H_0 .

Suppose that the sample size N is large (greater than 25) . Then we use the normal approximation to the Binomial as follows: If the proportion of plus signs is p then the test statistics

$$z = \frac{p - 1/2}{1/\sqrt{(4N)}}$$

is standard normal

We may then base a test of $H_0 : p = 0.5$ on z . Given α we have z_α and $z_{\alpha/2}$ from normal tables and

- For $H_1 : \text{median} \leq m$ or $H_1 : p \leq 0.5$ reject H_0 if $z \leq -z_\alpha$
- For $H_1 : \text{median} \geq m$ or $H_1 : p \geq 0.5$ reject H_0 if $z \geq z_\alpha$

- For $H_1 : \text{median} \neq m$ or $H_1 : p \neq 0.5$ reject H_0 if $|z| \geq z_\alpha$

You will notice that the sign test makes few assumptions about the form of the parent distribution and in consequence it is a valuable tool when we have non normal distributions. However it is probably pretty clear that we must pay a price for the versatility of the test. Indeed as we replace numerical values with indicators, the signs, it seems inevitable that there will be some loss of sensitivity..

10 Comparison of tests

The question of comparison of tests is a natural one . There many be several tests available in a given situation and we need some method of comparison.

Suppose we have two possible tests of H_0 against H_0 where the two hypotheses are both simple. Let n_1 be the minimum sample size for which the test 1 has size α and power exceeding p and n_2 the minimum sample size for which the test 2 has size α and power exceeding pi . The relative efficiency of test 1 compared with test 2 is just n_2/n_1 . If test 2 is the best test (in the Neymann Pearson sense) then this ratio is the efficiency of test 1. This depends on α and π . A natural way to avoid this is to look at the asymptotic relative efficiency (ARE) of the tests. Suppose that the tests require sample sizes n_1, n_2 to achive a power of at least π for a given sample size α . The asymptotic relative efficiency ARE is the limit of n_2/n_1 as n_1 and n_2 tend to infinity. These ideas extend to composite hypotheses but there are some technical difficulties over definitions. We write the power function of a test based on a sample size N as $\pi(\theta)$ and say that the test is consistent if this tends to 1 as the sample size tends to infinity. A reult which we will find useful in computing AREs is given below.

Lemma 10.1 *Suppose we have a two possible test statistics T_1 and T_2 for testing a pair of hypotheses $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. Let $E[T_i] = \mu_i(\theta)$ and $\text{var}[T_i] = \sigma_i^2(\theta)$ $i = 1, 2$. Then under some regularity conditions the ARE of T_2 compared to T_1 is*

$$ARE(T_1, T_2) = \lim_{n \rightarrow \infty} \left[\frac{(d\mu_2(\theta)/d\theta)^2}{\sigma_2^2(\theta)} / \frac{(d\mu_1(\theta)/d\theta)^2}{\sigma_1^2(\theta)} \right]$$

at $\theta = \theta_0$

10.0.1 example

Suppose we compare the t test with the sign test. The test statistics T_2 is simply the number of negative observations and is Binomial so $E[T_2] = np$ and $\text{var}[T_2] = np(1 - p)$. Now

$$\frac{dE[T_2]}{d\theta} = n \frac{ndp}{d\theta}$$

but $p = F(0, \theta)$ where $F(x, \theta)$ is the cdf of an observation. If ,as we assume, the parameter θ is a location parameter $F(x, \theta) = F(x - \theta)$ so

$$\frac{dp}{d\theta} = -f(x - \theta) = f(0, \theta).$$

Then from the lemma

$$\begin{aligned} ARE(T_1, T_2) &= \lim_{n \rightarrow \infty} \left[\frac{n^2 f^2(0, 0)}{n/4} / \frac{n/\sigma^2}{1} \right] \\ &= [4\sigma^2 f^2(0, 0)]. \end{aligned}$$

We know that for normal populations the t-test is most powerful test for a mean and in comparison the sign test has ARE of 0.6366. This is surprising given the amount of information thrown away in implementing the sign test. What is more we are assuming a normal parent, for some long tailed distributions the ARE can exceed 1.

10.1 Wilcoxon's Signed Rank Test

Sadly the sign test while both quick and easy is as we have seen not as powerful as we might wish. This makes it rather a frustrating test to use. In the 1940's Frank Wilcoxon came up with a modification of the test which is not much extra work but is much more powerful. The test is now known as Wilcoxon's Signed Rank Test. Wilcoxon suggested that rather than map the deviations between the hypothetical median and the observations onto just two values some assessment should be made of the magnitude of these deviations. He used the rank of the deviations. We take an smaller data set as an example.

131 127 118 135 117 112 132 120 137 113

We suspect a median of 121 so to test

$$H_0 : m = 121 \text{ against } H_1 : m \neq 121$$

we first get the deviations from the median (under H_0)

10 6 -3 14 -4 -9 11 -1 16 -8

with 5 positive and 5 negative signs. The sign test would accept H_0 .

Wilcoxon took the view that way to improve the test was to take into account the size of deviation from the median as is done by the t-test . The procedure he adopted was to take the deviations and give them a rank order ignoring the sign.

Thus the smallest deviation gets rank 1 the next smallest rank 2 and so on. These ranks are then given the sign of the corresponding deviation . The data and the ranks etc. are laid out in the table 10.1

Now if the suggested median is the true one than the deviations should be scattered around zero and the sum of the positive ranks R_+ should be much the same as the sum of the negative ranks R_- .

Data	Deviation	Deviation	Rank	Signed rank
131	10	10	7	7
127	6	6	4	4
118	-3	3	2	-2
135	14	14	9	9
117	-4	4	3	-3
112	-9	9	6	6
132	11	11	8	8
120	-1	1	1	-1
137	16	16	10	10
113	-8	8	5	-5

For our data $R_+ = 38$ and $R_- = 17$. To test the hypothesis Wilcoxon actually worked out the probability distribution of $W = \min(R_+, R_-)$ on the assumption

H_0 : the deviations have a symmetrical parent distribution with median zero

and hence calculated the percentage points, a table of which is available.

Summary

The procedure is thus

1. Ignoring the sign assign to each difference its rank. If two or more values have the same rank then the average rank is assigned to each.
2. Given the ranks in (1) form the signed ranks by giving a negative sign to those ranks which come from negative values.
3. Compute
 $R_+ = \text{sum of the positive ranks}$
 $R_- = \text{sum of the negative ranks}$
4. Compute W the minimum of these R_+ and R_- .
5. Reject H_0 if W is less than the tabulated values given at the end of this section.

Given the tables the test is simple and is in fact quite powerful, the only assumption being that the parent distribution is symmetrical. If this is not true then the sign test is your best bet.

To find the distribution of W consider the case where we have a sample of size N . The deviations D_1, D_2, \dots, D_N give rise to ranks R_1, R_2, \dots, R_N which have positive or negative signs. Under H_0 we can think of these signs being either $+$ or $-$ with probability $\frac{1}{2}$. So there are 2^N possible allocations of signs to the rank sequence. Assuming each sequence is equally likely then each has probability 2^{-N} . We can thus count the number of sequences which can make up a W to give the probability of W . This is a tedious but possible procedure.

Thus suppose we have a sample of size 7 so there are 128 possible sequences of + and - and the sum of the ranks is 28. We see that W is 4.

Then

W	Ways to make W	No of ways	Cum. total
0	1	1	1
1	1	1	2
2	1	1	3
3	3, 2+1	2	5
4	4,3+1	2	7
5	5,4+1,3+2,	3	10
6	6,5+1,4+2,3+2+1	4	14
7	7,6+1,5+2,4+3,4+2+1	5	19
8	8,7+1,6+2,5+3,5+2+1,4+3+1	6	25

Thus the probability that $W \leq 4$ is $14/28 = 0.1094$. However W could be positive or negative so the probability that $W \leq 4$ is 2×0.1094 . The probability that $W \leq 3$ is 0.02344 so the probability that $W \leq 3$ is 0.046875 allowing both tails.

For large sample sizes, n , we use a normal approximation under the null hypothesis that the D_i are independent and symmetrically distributed about zero

$$E[R+] = \frac{1}{4}N(N+1)$$

$$var(R+) = \frac{1}{24}N(N+1)(2N+1)$$

and $R+$ is asymptotically normal.

10.2 Paired Tests

The most common situation where the Wilcoxon or sign tests are used is when we have paired samples. Here we are assuming that there is some inherent pairing between to samples. Examples might be measurements made before and after treatment on the same item or paint lifetime on pairs of door on the same care. Pairing is always a good idea as it increases efficiency dramatically.

For example consider

No.	Before	After	Difference (d)
1	251	261	10
2	247	292	45
3	308	317	9
4	258	253	-5
5	267	271	4
6	256	305	49
7	230	238	8
8	268	320	52
9	269	267	-2
10	275	281	6

The standard test of

$$H_0 : \text{mean before} = \text{mean after}$$

is equivalent to

$$H_0 : \text{mean of differences} = 0$$

and the standard test - assuming normality would be a t test using

$$t = \frac{\bar{d}}{\sqrt{s^2/N}}$$

with N-1 degrees of freedom

Thus here $\sum d_i = 176$ and $\sum d_i^2 = 7456$ so $t = 2.529$ and we reject H_0 : same mean.

Suppose however we did not wish to assume normality. Then we could use the sign test or better still Wilcoxon's test on the differences That is we test the differences for a zero median rather than a zero mean. Our null

$$H_0 : \text{differences have zero median}$$

and the alternative

$$H_1 : \text{differences have non-zero median}$$

The procedure is , take the two samples and compute the differences. Then H_0 : no difference in medians is the same as assuming that the differences have zero median and we can do a Wilcoxon test on the differences, giving

No.	Before	After	Difference (d)	Rank	Signed Rank
1	251	261	10	7	7
2	247	292	45	8	8
3	308	317	9	6	6
4	258	253	-5	3	-3
5	267	271	4	2	2
6	256	305	49	9	9
7	230	238	8	5	5
8	268	320	52	10	10
9	269	267	-2	1	-1
10	275	281	6	4	4

Now $R_+ = 51$ while $R_- = 4$ so $W = 4$ and we reject H_0

We could have done a sign test since as is clear we have 2 minus signs in 10 or 8 + signs in 10. The probability of 8 or more in 10 is 0.0547. This is just on the borderline for a one sided test illustrating the loss of efficiency which occurs with the sign test. We usually prefer the greater efficiency of Wilcoxon's test compared with the sign test assuming a symmetric parent.

10.2.1 Try this example

Pipes are made of a new and a standard alloy. Lengths of pipe made from each alloy are paired and subjected to corrosive conditions. The data below gives the amount of corrosion. Does the new alloy perform differently to the standard?

Pair	1	2	3	4	5	6	7	8	9	10	11	12
Standard	20	26	31	42	35	19	33	38	29	27	40	37
New	25	29	28	37	40	29	41	43	21	35	47	41

10.3 Exercises

1. A company runs a trial to aid it in its decision as to which make of tyres to use. Sixteen cars were driven over a prescribed course with tyres made by company A. The tyres were then changed to those made by B and the cars were run over identical routes. The petrol consumptions were (km/litre)

Car	1	2	3	4	5	6	7	8	9	10	11
A	4.2	4.7	6.6	7.0	6.7	4.5	5.7	6.0	7.4	4.9	6.1
B	4.1	4.9	6.2	6.9	6.8	4.4	5.7	5.8	6.9	4.9	6.0

What conclusions can you draw?

2. Two methods for monitoring Sulphur Monoxide were compared over 12 days. The results were:

Pair	1	2	3	4	5	6	7	8	9	10	11	12
M1	0.96	0.75	0.61	0.89	0.64	0.81	0.68	0.65	0.84	0.59	0.73	0.78
M2	0.87	0.75	0.63	0.55	0.76	0.70	0.69	0.57	0.53	0.88	0.51	0.80

Is there any difference between the methods?

10.4 Unpaired Samples - The Mann-Whitney U Test.

Clearly not all comparisons between samples are made with the assumption of pairing. The (rather strange) example below is a simple illustration

Speaking ability for patients in a study of Parkinson's disease

Had Operation	2.6	2.0	1.7	2.7	2.5	2.6	2.5	3.0
Had Not	1.2	1.8	1.8	2.3	1.3	3.0	2.2	1.3
	1.5	1.6	1.3	1.5	2.7	2.0		

How can one compare two such samples without making the normal assumption which is required for the standard t - test?

One possibility is a permutation test. We have two samples one of size 8 the other of size 14 and the natural statistic would be the difference between the sample means. We can then lump all the samples values together and construct all possible samples of size 8. For each sample we have a value of the statistic and hence the distribution of the mean differences. Again the procedure is simple the but the computations are formidable. A more practical procedure is described below.

If you want to compare the locations of two samples (which are not paired) the best all purpose test is the Mann-Whitney U test described below. If the samples come from normal populations it is a little less powerful than the t test but in all other cases it is better. To begin with let's look at the Normal case where we are happy to assume that the sample comes from Normal populations.

Suppose we have two samples one from A of size m and one from B of size n e.g.

A 7.05 14.25 8.57 10.50 11.90 4.50 37.60 9.40 8.10 45.20

B 9.25 2.05 2.75 2.50 6.40 6.90 10.00 8.00 33.00

and we wish to compare the means. Then the usual t test is based on a normal assumption gives in our case $t = 1.93$ with 17 degrees of freedom.

If we do not wish to assume normality we can proceed as follows:

Assign to each observation its rank in the combined set of observations as below

A 7.05 14.25 8.57 10.50 11.90 4.50 37.60 9.40 8.10 45.20

rank 7 16 10 14 15 4 18 12 9 19

B	9.25	2.05	2.75	2.50	6.40	6.90	10.00	8.00	33.00
rank	11	1	3	2	5	6	13	8	17

Now we need a test based on the ranks rather than the observations themselves. For the Mann-Whitney test we find the sum of the ranks for each sample $R_A = 124$ while $R_B = 66$ and then take $W =$ rank sum of the smaller sample, here $R_B = 66$. If we assume that the two samples are independent and are from populations with the same distribution then we can find (with difficulty) the distribution of W . The only problem with this test is that there are many subtle variants all of which use slightly different tables. I think the easiest version is a slight modification of the above which as follows

Procedure:

- Let m be the smaller sample size, and let R be the rank sum from that sample. If they are the same then pick the sample you prefer.
- Let $R' = m(n + m + 1) - R$
- Let $W = \min(R, R')$

The tables give the 5% and 2.5% points of the distribution of W . That is if the observed value of W (rounded) is less than or equal to the tabulated value then we reject H_0 the hypothesis of equal medians.

Note: For equal values of the observations we assign average ranks just as for Wilcoxon's test.

For large sample values it is conventional we can get quite good approximate values for the rank sum of the sample of m using the normal distribution with mean and variance given by the formula below for the m A's and n B's.

$$\mu = \frac{1}{2}m(m + n + 1)$$

$$\sigma^2 = \frac{1}{12}mn(m + n + 1)$$

For the example above:

The smaller sample gives a rank sum of $R = 66$ giving $R' = 9(9+10+1)-66 = 114$. This gives $W = \min(66, 114) = 66$ which is not quite significant at 5% (two sided).

Beware the tables may not be in terms of R above as there are several variants of this test

10.4.1 Summary of the Mann-Whitney test

If we have two samples and we wish to test

H_0 : they come from populations with the same median

against

H_1 : come from populations with different medians

we use Mann-Whitney U.

1. Rank the all data
2. Find the sum of the ranks for the smaller sample (with m observations) say R
3. Find $R' = m(m+n+1) - R$
4. Compute $W = \min(R,R')$
5. Reject H_0 if W is less than the tabulated value.

Clearly we can find the null distribution quite simple but the calculations are tedious. We rank all the data so we have $m + n$ ranks which sum to $(n + m)(n + m + 1)/2$. We assume that all possible assignments of ranks are equally likely so the possibilities for the m ranks in the smaller sample are . We then count the number of ways of obtaining values for W .

10.4.2 Exercises

1. Does the treatment affect the speech of the patients with Parkinson's disease?
2. Two methods A and B we used to determine the latent heat of fusion of ice. The table below gives the change in total heat from ice at -72°C to water at 0°C .

A	79.98	80.04	80.02	80.04	80.03	80.03	80.04
	79.97	80.05	80.03	80.02	80.00	80.02	

B	80.02	79.94	79.98	79.79	79.97	80.03	79.95	79.97
---	-------	-------	-------	-------	-------	-------	-------	-------

Are the methods comparable?

3. The data below gives the percentage of aluminum in two samples A and B. Is it reasonable to assume that the samples come from populations with the same median?

Sample A

11.4	13.4	13.5	13.8	13.9									
14.4	14.5	15.0	15.1	15.8	16.0	16.3	16.5	16.9	17.0	17.2	17.5	19.0	19.0

Sample B

9.0 10.5 10.8 12.0 12.7 12.8 13.4 14.2 14.3 14.7 15.0 15.7 16.0 18.2 18.2 19.1

4. A simple experiment was designed to compare the hardness of flint in two areas A and B. Four pieces of flint From A and Five from B were taken and their hardness assessed. They were then ranked in order of hardness as follows

A A A B A B B B B

The hardness decreasing from left to right. What evidence is there to conclude that there is a difference between A and B?