

Statistical Inference:  
Estimation and Confidence  
Intervals  
Hypothesis Testing

In most statistics problems, we assume that the data have been generated from some unknown probability distribution. We desire to learn about this underlying distribution and *infer* certain properties about it based upon our observed data.

Here, we will assume that the form of this distribution is known except for a few parameters.

For example, let us assume that the number of typo's on a page of Alfred's writing has a  $\text{Poisson}(\theta)$  distribution but it is not known what the exact value of  $\theta$  is. If we read several pages of his work and observe the number of mistakes on each page then based upon these observed values we can make inference about the unknown value of  $\theta$ .

Other examples include:

- Light bulb failure — Exponential distribution
- Psychological test score — Normal distribution

There are many different things we might be interested in inferring about the unknown parameter  $\theta$ :

- The “best” estimate — MLE (Frequentist) or Bayesian Estimator (Bayesian)
- An interval in which we believe  $\theta$  to lie — Confidence Interval
- Whether  $\theta$  is larger than some specified value — Hypothesis testing

## Sampling distributions

Recall: A statistic  $T = t(X_1, \dots, X_n)$

Since it is a function of random variables,  $T$  is also a random variable with a distribution of its own. We call this distribution the *sampling* distribution.

To perform statistical inference for an estimator  $T$  of  $\theta$ , we will often need to derive its distribution.

We now present some important results for random samples from a normal distribution. We will also introduce some new distributions that will come in handy.

# Samples from a Normal distribution

Throughout the next section we will assume that  $X_1, \dots, X_n$  form a random sample from a normal distribution with unknown mean  $\mu$  and unknown variance  $\sigma^2$ .

## $\chi^2$ distribution

A  $\chi^2$  distribution with  $n$  degrees of freedom is a  $\Gamma(n/2, 1/2)$ . It has pdf

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}, \quad x > 0$$

It follows that

$$\mathbb{E}(X) = n \quad \text{Var}(X) = 2n$$

and the moment generating function

$$M(t) = \left( \frac{1}{1-2t} \right)^{n/2} \quad t < \frac{1}{2}$$

This distribution is important because of the following result:

**Theorem 1.** *If the random variables  $Y_1, \dots, Y_r$  are IID  $N(0, 1)$  then*

$$Y_1^2 + \dots + Y_n^2 \sim \chi^2(r)$$

Proof:

1.  $Y_1^2 \sim \chi^2(1)$  — Standard manipulation of cdf.
2.  $\chi^2(r_1) + \chi^2(r_2) = \chi^2(r_1 + r_2)$  — Moment generating function

## Independence of the sample mean and sample variance

For a normal random sample, the following result is crucial

**Theorem 2.** *If  $X_1, X_2, \dots, X_n$  are IID  $N(\mu, \sigma^2)$ , then the sample mean and variance*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

are **independent** random variables with distributions

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Proof: See any statistics textbook e.g. DeGroot, Chapter 7

When considering inference on normally distributed data, we are often interested in the mean  $\mu$ . We have just shown that if  $X_1, \dots, X_n$  are random sample from  $N(\mu, \sigma^2)$  then

$$\sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

However this contains the quantity  $\sigma$ , which is usually unknown too. We require an estimator for the mean  $\mu$  which does not contain this unknown variance parameter.

To do this we first have to introduce the  $t$ -distribution

## The $t$ distribution

**Definition 3.** If  $U \sim N(0, 1)$  and  $V \sim \chi^2(n)$  are independent, then

$$T = \frac{U}{\sqrt{V/n}} \sim t(n)$$

is called a  $t$ -distribution with  $n$  degrees of freedom

The pdf of the  $t(n)$  distribution is

$$g_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}$$

It looks similar to a  $N(0, 1)$  but has fatter tails. The greater the number of degrees of freedom, the closer the  $t$ -distribution is to the standard normal.

**Theorem 4.** If  $X_1, \dots, X_n$  are IID  $N(\mu, \sigma^2)$  then

$$\sqrt{n} \frac{(\bar{X} - \mu)}{S} \sim t(n - 1)$$

Proof:  $\bar{X}$  and  $S^2$  are independent and

$$\sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \sim N(0, 1) \quad \frac{(n - 1)S^2}{\sigma^2} \sim \chi^2(n - 1)$$

Apply definition 3.

## Confidence intervals

**Definition 5.** Let  $X_1, \dots, X_n$  form a random sample from a distribution involving an unknown parameter  $\theta$ . Suppose we have two statistics  $a(X_1, \dots, X_n)$  and  $b(X_1, \dots, X_n)$  such that

$$P(a(X_1, \dots, X_n) < \theta < b(X_1, \dots, X_n)) = 1 - \alpha.$$

If the observed values  $a(X_1, \dots, X_n) = a$  and  $b(X_1, \dots, X_n) = b$ , we say that the interval  $(a, b)$  is a  $100(1 - \alpha)\%$  confidence interval for  $\theta$ .

## Confidence intervals for the mean of a normal distribution

Assume again that  $X_1, \dots, X_n$  are IID  $N(\mu, \sigma^2)$  with both  $\mu$  and  $\sigma^2$  unknown. We wish to find a range of plausible values for the mean  $\mu$ .

We know that

$$T = \sqrt{n} \frac{(\bar{X} - \mu)}{S} \sim t(n - 1)$$

so that a  $100(1 - \alpha)\%$  confidence interval is

$$P\left(\bar{X} - \frac{t_{\alpha/2}(n-1)}{\sqrt{n}}S < \mu < \bar{X} + \frac{t_{\alpha/2}(n-1)}{\sqrt{n}}S\right) = 1 - \alpha$$

where  $t_{\alpha/2}(n-1)$  is the  $(1 - \alpha/2)$  quantile of the  $t(n-1)$  distribution.

# Hypothesis Testing

Let us now suppose that we have some parameter of interest  $\theta$  whose value is unknown but must lie in a certain space  $\Omega$ . We partition  $\Omega$  into the disjoint subsets  $\Omega_0$  and  $\Omega_1$ . As statisticians we are interested in deciding whether the unknown  $\theta$  lies in  $\Omega_0$  or  $\Omega_1$ .

## The test

$H_0 : \theta \in \Omega_0$  (Null Hypothesis)

$H_1 : \theta \in \Omega_1$  (Alternative Hypothesis)

Decide upon some test statistic  $T$  for which extreme values indicate departure from the null hypothesis  $H_0$ .

Under the null hypothesis, we can find the distribution of this test statistic  $T$  and find (in light of the alternative hypothesis) the probability of a test statistic at least as extreme as the value  $t$  obtained with the observed data  $X_1, \dots, X_n$  e.g.

$$P(T > t | H_0) = p$$

This  $p$  is known as the  $p$ -value (or *significance level*).

A small p-value is evidence against the null hypothesis.

There is **NO** magic level which means that null hypotheses are **automatically** rejected. Although lots of people mistakenly use a significance level of  $p < 0.05$  to definitely reject the null hypothesis, it should depend entirely on the consequences of being wrong. P-values simply state, under the null hypothesis, what the probability that the apparent difference is due to chance.

At the **least** you should qualify any statements such as

$0.05 < p \leq 0.06$	"Weak evidence for rejection"
$0.03 < p \leq 0.05$	"Reasonable evidence for rejection"
$0.01 < p \leq 0.03$	"Good evidence for rejection"
$0.005 < p \leq 0.01$	"Strong evidence for rejection"
$0.001 < p \leq 0.005$	"Very strong evidence for rejection"
$0.0005 < p \leq 0.001$	"Extremely strong evidence for rejection"
$p \leq 0.0005$	"Overwhelming evidence for rejection"

## The one sample $t$ -test

Used for observations  $X_1, \dots, X_n$  drawn from IID  $N(\mu, \sigma^2)$  with unknown parameters. We wish to test for location

$$H_0 : \mu = \mu_0 \quad (\text{Null Hypothesis})$$

$$\text{Possible } H_1 : \mu > \mu_0 \quad (\text{Alternative Hypothesis})$$

Test Statistic

$$T = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \sim t(n - 1)$$

Example We have measured the IQ of 8 students and we wish to see if their intelligence is higher than the population average of 100,

Student	IQ
1	118
2	121
3	96
4	102
5	93
6	110
7	117
8	131

$$\bar{X} = 111 \quad S^2 = 174$$

$H_0 : \mu = 100$  (Null Hypothesis)

$H_1 : \mu > 100$  (Alternative Hypothesis)

Data give test statistic  $t = 2.36$  and (having noted the form of the alternative hypothesis) we find that

$$P(T > 2.36 | H_0) = 0.025$$

Hence, there is good evidence for rejection of the null hypothesis that the students have the same IQ as the general population.

## The paired $t$ -test

Suppose we have **pairs** of random variables  $(X_i, Y_i)$  and that  $D_i = X_i - Y_i$  is an independent random sample from a normal distribution  $N(\mu, \sigma^2)$  with unknown parameters. We wish to test whether there is a difference in mean between two samples

$$H_0 : \mu = 0 \quad (X_i\text{'s don't differ from } Y_i\text{'s)}$$

$$\text{Possible } H_1 : \mu \neq 0 \quad (\text{Do differ})$$

Test Statistic

$$T = \frac{\sqrt{n}(\bar{D} - \mu_0)}{S_D} \sim t(n - 1)$$

where  $S_D^2$  is the sample variance of the differences  $D_i$ .

Example (Box, Hunter & Hunter) Two types of rubber (A and B) were randomly assigned to the left and right shoes of 10 boys and relative wear on each measured

Boy	A	B	Difference
1	13.2	14.0	-0.8
2	8.2	8.8	-0.6
3	10.9	11.2	-0.3
4	14.3	14.2	0.1
5	10.7	11.8	-1.1
6	6.6	6.4	0.2
7	9.5	9.8	-0.3
8	10.8	11.3	-0.5
9	8.8	9.3	-0.5
10	13.3	13.6	-0.3

$H_0 : \mu = 0$  (No difference in shoe wear between A and B)

$H_1 : \mu \neq 0$  (Difference between A and B)

Data give

$$\bar{D} = -0.41 \quad S_D^2 = 0.15$$

so that  $T = -3.3489$ . Noting the form of the alternative hypothesis we calculate

$$P(|T| > 3.3489 | H_0) = 0.0085$$

and so there is strong evidence for rejection of the null hypothesis that wear is the same for both types of rubber.

## The two sample $t$ -test

Suppose we have two random samples,  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  which are independent and normally distributed with the **same variance** i.e.  $X_i \sim N(\mu_X, \sigma^2)$  and  $Y_i \sim N(\mu_Y, \sigma^2)$ . Wish to test

$$H_0 : \mu_X = \mu_Y \quad (\text{Null Hypothesis})$$

$$\text{Possible } H_1 : \mu_X \neq \mu_Y \quad (\text{Alternative Hypothesis})$$

Use test statistic

$$T = \frac{\bar{X} - \bar{Y}}{S \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t(m + n - 2)$$

where the pooled sample variance

$$S^2 = \frac{(m - 1)S_X^2 + (n - 1)S_Y^2}{m + n - 2}$$

**Proof** Under  $H_0$  we have

$$\bar{X} - \bar{Y} \sim N\left(0, \sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)\right),$$

and also

$$\begin{aligned} \frac{(m-1)S_X^2}{\sigma^2} &\sim \chi^2(m-1), & \frac{(n-1)S_Y^2}{\sigma^2} &\sim \chi^2(n-1) \\ \implies \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} &\sim \chi^2(m+n-2). \end{aligned}$$

So defining

$$S^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

we have

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t(m+n-2)$$

Example We have measured the results of two experiments (which we know have the same variance) to determine the concentration of a chemical

<i>TestX</i>	22	19	35	11	21	10
<i>TestY</i>	33	11	20	38		

Test

$H_0 : \mu_X = \mu_Y$  (No difference in mean of experiments)

$H_1 : \mu_X \neq \mu_Y$  (X has a different mean than Y)

We find

$$\begin{aligned}\bar{X} &= 19.7 & \bar{Y} &= 25.5 \\ S_X &= 82.2 & S_Y &= 90.6\end{aligned}$$

So  $T = -0.87$  which on comparing with a  $t(8)$  distribution has a p-value of 0.41 hence we do not reject  $H_0$ .

## Tests of variance

In the last test we assumed that  $X$  and  $Y$  had the same variance. How can we check this is the case? We can use the F test

### Snedecor's F distribution

If

$$U \sim \chi^2(m) \quad V \sim \chi^2(n),$$

then

$$\frac{U/m}{V/n} \sim F_{m,n}$$

Suppose we have two samples  $X_1, \dots, X_m \sim N(\mu_X, \sigma_X^2)$  and  $Y_1, \dots, Y_n \sim N(\mu_Y, \sigma_Y^2)$  and we wish to test

$H_0 : \sigma_X = \sigma_Y$  (No difference in variance)

$H_1 : \sigma_X \neq \sigma_Y$  (X has a different variance than Y)

We use test statistic

$$\frac{S_X^2}{S_Y^2} \sim F_{m-1, n-1}$$

under the null hypothesis. Note that the greater this ratio deviates from 1, the stronger the evidence for unequal variances.

Example Returning to our previous example let us assume that we did not know that the two experiments had different variances.

<i>TestX</i>	22	19	35	11	21	10
<i>TestY</i>	33	11	20	38		

and so we wish to test

$$H_0 : \sigma_X = \sigma_Y \quad (\text{No difference in variance})$$

$$H_1 : \sigma_X \neq \sigma_Y \quad (\text{X has a different variance than Y})$$

We find that

$$\frac{S_X^2}{S_Y^2} = 0.545$$

comparing this with a  $F_{5,3}$  we obtain a p-value of 0.52 and so we do not reject the null hypothesis of equal variances.

**How can we test difference of location between two samples if the variances are different?**

Suppose we have  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  where the variances are different. We wish to test whether they have the same mean. Can still use a t-test but this time we use the statistic

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{S_X^2}{m} - \frac{S_Y^2}{n}\right)}} \sim t(\nu)$$

which is an approximation, with a formula for  $\nu$  involving  $m, n, S_X^2$  and  $S_Y^2$ .

This can be implemented easily in R (in fact it is the default).

## Non-parametric tests

What do we do if we can't satisfy the assumptions about the distributions of the variables being assessed (usually that they are normal)?

**Note:** Use q-q plot to see if they do look normal but may wish to be conservative anyway

## Tests of location zero

Suppose we have data  $D_1, \dots, D_n$  and we wish to test the null hypothesis that the data have a symmetric continuous distribution centred about zero. Hence can be used to test for differences between paired samples i.e.  $D_i = X_i - Y_i$

## Sign test

Take the data  $D_1, \dots, D_n$  and count

Number of +'s  $= n_+$

Number of -'s  $= n_-$

Under the null hypothesis that positive and negative values are equally likely, the number of positive values

$$N_+ \sim B(n, 1/2)$$

and so we can easily calculate the p-value.

## Wilcoxon signed-rank test

Order the absolute values  $|D_1|, \dots, |D_n|$  and assign each a rank  $R_i$ . The smallest absolute value getting rank 1 and tied scores are assigned a mean rank. If some  $D_i = 0$  we drop these values completely and merely rank the remaining.

Define our test statistic to be the sum of the ranks of the positive  $D_i$ ,

$$W^+ = \sum_{D_i > 0} R_i.$$

Extreme values of this statistic (large or small) indicate departure from the null hypothesis. We can work out the exact distribution under  $H_0$  of  $W^+$  using the permutation distribution, otherwise we use a large-sample normal approximation.

Example - Try also on the shoe data

Suppose we have 9 aptitude scores of 13, 7, 3, 15, 10, 12, 8, 2, 9 and we wish to test if they are symmetrically distributed around 10. We first discard the score of 10 leaving 8 scores

Score	13	7	3	15	12	8	2	9
Score - 10 = $D$	3	-3	-7	5	2	-2	-8	-1
$ D $	3	3	7	5	2	2	8	1
Rank assigned	4.5	4.5	7	6	2.5	2.5	8	1

Hence we obtain  $W^+ = 13$ . After looking at tables we find this has a p-value of 0.53 and so we do not reject the null hypothesis.

## Wilcoxon rank sum test

Used to test for a location shift between two samples of observations  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$ .

- Combine the two samples  $X$ 's and  $Y$ 's and assign ranks to each observation in this combined sample.
- Find  $R_X =$  Sum all the ranks in the first sample i.e. the  $X$ 's.
- Let  $W = R_X - m(m + 1)/2$  {Optional - default in  $R$ }

For small samples without ties we can find the permutation distribution of  $W$  exactly, otherwise we use a large sample normal approximation

Example Consider again our two experiments and now use the rank-sum test.

Ordered results	10	11	11	19	20	21	22	33	35	38
Experiment	<i>X</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>X</i>	<i>Y</i>	<i>X</i>	<i>Y</i>
Rank	1	2.5	2.5	4	5	6	7	8	9	10

Hence  $W = 29.5 - 21 = 8.5$ . On comparing this with the relevant tables we obtain a p-value of 0.52 so we do not reject the null hypothesis of no location shift between *X* and *Y*.

## Randomization/Permutation tests

This is not one particular test but rather a general approach that will work with **any** test statistic. Let us suppose we have the data with a null hypothesis  $H_0$  and have chosen a test statistic  $T$ . All we now have to do is work out the sampling distribution of  $T$  under  $H_0$  but this is often very tricky.

However, under the null hypothesis we know that all permutations of the data are equally likely (*exchangeable*). We can hence find the sampling distribution of  $T$  under the null hypothesis by forming all possible data permutations and calculating the observed test statistic for each.

Example - Permutations for the Wilcoxon Rank Sum test.

Experiment X		22	19	35	11	21	10
Experiment Y		33	11	20	38		

In this case, the permutations are really combinations as order doesn't matter. Under the null hypothesis, all the observations are *exchangeable* and so any six of the 10 observations could be observed as Experiment  $X$  with the other four as Experiment  $Y$ .

To find the sampling distribution we have to calculate every possible combination of allocating the 10 values between the two experiments (6 in  $X$  and 4 in  $Y$ ). There are  $\binom{10}{4} = 210$  such combinations which are equally likely. By computing the test statistic for every such combination we can hence obtain the sampling distribution.

## Example - Permutations for a one sample test

Under the null hypothesis that the observations are symmetric around 0, it is equally likely to observe  $D$  as  $-D$ . Hence we can relabel our observations by permuting all the signs (i.e.  $+$  or  $-$ ) of our observations.

There are  $2^n$  ways of doing this — every observation can either be  $+$  or  $-$ . To find the sampling distribution we find every such relabelling and calculate its test statistic. In this way we can find out how extreme our observed value is.

## Monte-Carlo tests

Once we have large sample sizes, permutation tests become impractical. Due to the computational requirements, it is no longer feasible to find all the possible permutations.

We can however use a (large) random sample of the possible permutations (instead of them all). This is called a *Monte-Carlo* test. The p-values obtained in this way are themselves random variables but we can make sure we use enough permutations to guarantee their precision.

# Contingency Tables

Contingency tables are just counts of cross-tabulations according to two criterion (e.g. caffeine data). They are displayed as  $n_{ij}$  in a table with  $r$  rows and  $c$  columns:

$$\begin{array}{ccc|c}
 n_{11} & \dots & n_{1c} & n_{1.} \\
 \vdots & \ddots & \vdots & \vdots \\
 n_{r1} & \dots & n_{rc} & n_{r.} \\
 \hline
 n_{.1} & \dots & n_{.c} & n
 \end{array}$$

We wish to test the null hypothesis that the row and column categories are independent.

Under the null hypothesis, the expected number  $E_{ij}$  in each cell is estimated by

$$\begin{aligned} E_{ij} &= n \times P(i^{\text{th}} \text{ row}) \times P(j^{\text{th}} \text{ column}) \\ &= \frac{n_{i \cdot} n_{\cdot j}}{n} \end{aligned}$$

Let  $O_{ij}$  be the observed number in cell  $(i, j)$ ,  $i = 1, \dots, r$ ,  $j = 1, \dots, c$ . The Pearson chi-squared statistic is then

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2((r-1)(c-1))$$

at least approximately.

Example The caffeine data

	0	1 – 150	151 – 300	> 300	<i>Total</i>
<i>Married</i>	652	1537	598	242	3029
<i>Prev.Married</i>	36	46	38	21	141
<i>Single</i>	218	327	106	67	718
<i>Total</i>	906	1910	742	330	3888

Gives  $X^2 = 51.7$  on 6 degrees of freedom. When compared to a  $\chi_6^2$  we obtain a p-value  $\approx 0$  and so we reject the null hypothesis of independence.

## Validity of Chi-squared test

The derivation of the null hypothesis distribution of the chi-squared test that

$$X^2 \sim \chi^2((r - 1)(c - 1))$$

relies on the counts being fairly large. You should **not** use this approximation if any of the counts are small (typically below 5).

If any counts are this small we can **either** try and combine the categories **or** try simulate the null hypothesis distribution of the chi-squared test using permutation methods **or** use Fisher's exact test.

## Fisher's exact test

Given the row and column sums of a contingency table, under the null hypothesis of independence the conditional probability of getting the observed matrix is given by a multivariate generalization of the hypergeometric distribution

$$P = \frac{(n_{1.}!n_{2.}!\dots n_{r.}!)(n_{.1}!n_{.2}!\dots n_{.c}!)}{n! \prod_{i,j} n_{ij}}.$$

Similarly we can find all possible matrices with the same row and column sums and calculate the associated conditional probability.

Given a particular statistic (e.g. chi-squared statistic) the p-value of the observed data is then the probability of obtaining a table with at least as extreme a value of the statistic as that observed.

Example (2 × 2) only: Are parasites present in a sample of crabs from Oregon?

	<i>Yes</i>	<i>No</i>	
Red crab	5	312	317
Dungeness crab	0	503	503
<i>Total</i>	5	815	820

The conditional probability of this matrix is

$$P = \frac{317!503!5!815!}{820!5!312!0!503!} = 0.008$$

The other matrices possible are

Red crab	4	313	317
Dungeness crab	1	502	503
<i>Total</i>	5	815	820

$$P = 0.068$$

Red crab	3	314	317
Dungeness crab	2	501	503
<i>Total</i>	5	815	820

$$P = 0.218$$

Red crab	2	315	317
Dungeness crab	3	500	503
<i>Total</i>	5	815	820

$$P = 0.346$$

Red crab	1	316	317
Dungeness crab	4	499	503
<i>Total</i>	5	815	820

$$P = 0.273$$

Red crab	0	317	317
Dungeness crab	5	498	503
<i>Total</i>	5	815	820

$$P = 0.086$$

None of the other matrices give a chi-squared statistic as extreme as that obtained with our observed values and so our p-value is  $\approx 0.8\%$ .

**Note:** We can use Fisher exact test for general  $r \times c$  tables but it is a serious computational task