2 Normal approximations and t-tests

2.1 Properties of the Normal distribution

We consider the Normal distribution, the most important statistical distribution, and introduce some fundamental distributional results, including the Central Limit Theorem.

1. Let $X \sim N(\mu, \sigma^2)$ and consider the linear transformation,

$$Z = \frac{X - \mu}{\sigma}.$$

   Then, $Z \sim N(0, 1)$.

2. If X_1 and X_2 are independent with $X_i \sim N(\mu_i, \sigma_i^2)$ for $i = 1, 2$, then, $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

3. An extension of the above result is that if $X_1, \ldots, X_n$ are independent Normal rvs, such that $X_i \sim N(\mu_i, \sigma_i^2)$, and $a_1, \ldots, a_n$ are constants, then

$$\sum_{i=1}^{N} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right).$$

It follows from property 3 above, that if $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ rvs, then

$$\bar{X}\left(= \sum_{i=1}^{n} \frac{1}{n} X_i\right) \sim N\left(\mu, \frac{\sigma^2}{n}\right). \qquad \text{(Set } a_i = 1/n\text{)}.$$

Notes:

1. Property 1 was very useful in the past for working out probabilities associated with *any* Normal distribution. Problems could be transformed to be of the form of a standard Normal rv and tables cpuld be used to obtain values associated with a $N(0, 1)$ distribution. That was particularly useful before the "computer revolution".

2. Typically, the c.d.f. for a $N(0, 1)$ distribution is denoted by Φ, so that $\Phi(x) = \mathbb{P}(X \leq x)$ for $X \sim N(0, 1)$. Similarly Φ^{-1} is the inverse c.d.f. of a $N(0, 1)$ distribution.

2.2 Central Limit Theorem (CLT)

This is perhaps one of the most interesting of any statistical result. The Central Limit Theorem states that the sum or the mean of n independent and identically distributed (iid) rvs from almost *any* distribution is approximately Normal for large enough n:

Theorem

Let $X_1, \ldots, X_n$ be iid rvs from any distribution having mean μ and variance $\sigma^2 < \infty$. Then,

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \;\dot{\sim}\; N(0, 1) \qquad \text{as } n \to \infty,$$

where " $\dot{\sim}$ " means distributed approximately as (though this notation is not quite standard).

Note that is equivalent to,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \;\dot{\sim}\; N(0, 1) \qquad \text{as } n \to \infty;$$

$$\sum_{i=1}^{n} X_i \;\dot{\sim}\; N(n\mu, n\sigma^2) \qquad \text{as } n \to \infty;$$

and,

$$\bar{X} \;\dot{\sim}\; N(\mu, \sigma^2/n) \qquad \text{as } n \to \infty.$$

Example

A bridge can hold a maximum of 400 vehicles (bumper-to-bumper and stationary). The mean weight of vehicles using the bridge is 2.5 tonnes with a variance of 4.0 tonnes. What is the probability that the maximum design load of 1100 tonnes will be exceeded in a traffic jam?

Let X_i denote the weight of vehicle i, $i = 1, \ldots, 400$. Then, $E(X_i) = 2.5$ and $Var(X_i) = 4$, and by the CLT,

$$\sum_{i=1}^{400} X_i \;\dot{\sim}\; N(400 \times 2.5, 400 \times 4.0) \equiv N(1000, 1600).$$

We wish to calculate,

$$\mathbb{P}\left(\sum_{i=1}^{400} X_i > 1100\right)$$

This can be obtained in `R` using,

```
> 1-pnorm(1100,1000,sqrt(1600))
[1] 0.006209665
```

Equivalently,

$$\frac{\sum_{i=1}^{400} X_i - 1000}{40} \;\dot{\sim}\; N(0,1).$$

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{400} X_i > 1100\right) &= \mathbb{P}(Z > (1100-1000)/40) \qquad \text{where } Z \sim N(0,1)\\ &= \mathbb{P}(Z > 2.5) = 0.006.\end{aligned}$$

This can be obtained in `R` using,

```
> 1-pnorm(2.5,0,1)
[1] 0.006209665
```

2.3 Approximate Normal Distributions

The CLT provides the justification for approximating several distributions by a Normal distribution. Here, we consider two particular examples.

2.3.1 Binomial Distribution

The pmf $X \sim Bin(n,p)$ is of the form,

$$f(x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}.$$

The combination term can be difficult to calculate for large values of n. However, recall that for iid rvs $X_1, \ldots, X_n$, such that $X_i \sim Bernoulli(p)$, then $\sum_{i=1}^{n} X_i = X \sim Bin(n,p)$. We know that $\mathbb{E}(X_i) = p$ and $\mathrm{Var}(X_i) = p(1-p)$. Using the CLT, we have that,

$$\frac{X - np}{\sqrt{np(1-p)}} \;\dot{\sim}\; N(0,1) \qquad \text{as } n \to \infty,$$

or,

$$X \;\dot{\sim}\; N(np, np(1-p)) \qquad \text{as } n \to \infty.$$

Note: We can use this result in order to derive approximate confidence intervals for p. For this, we need to estimate the variance of X with $n\hat{p}(1-\hat{p})$, where $\hat{p} = X/n$ is the maximum likelihood estimator (MLE) of p; see later in MT2508 for such estimators. Then, approximately,

$$P(-z_{0.025} < \frac{X - np}{\sqrt{n\hat{p}(1-\hat{p})}} < z_{0.025}) = 0.95.$$

By re-arranging the terms in the above inequalities we obtain,

$$P(\frac{X}{n} - z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n} < p < \frac{X}{n} + z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n}) = 0.95.$$

Therefore, an approximate 95% confidence interval is given as

$$(\hat{p} - z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{0.025}\sqrt{\hat{p}(1-\hat{p})/n}).$$

To test for the null hypothesis $H_0 : p = p_0$ vs the alternative $H_1 : p \neq p_0$, we note that, approximately, the test statistic,

$$Z = \frac{X - np_0}{\sqrt{np_0(1-p_0)}},$$

follows a standard normal distribution under the null hypothesis. The critical region (for $a = 0.05$) is $Z < -1.96$ or $Z > 1.96$.
Note: the CLT does not tell us how large n should be for the approximation to hold. The usual rule of thumb is that the approximation can be used when $\min(np, n(1-p)) > 5$.

2.3.2 Poisson Distribution

Recall that the Poisson probability mass function is

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad x = 0, 1, \ldots$$

This is also awkward to evaluate for high values of λ. However, we can exploit the useful general result (which will be proved in Honours) that

$$X_1, X_2 \text{ independent with } X_i \sim Po(\lambda_i) \quad i = 1, 2 \Rightarrow X_1 + X_2 \sim Po(\lambda_1 + \lambda_2). \tag{1}$$

Using (1), we can see that if $X \sim Po(\lambda)$ then we can set $X = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent and $X_i \sim Po(\lambda/n)$. Since the variance of a Poisson distribution is equal to its mean, each X_i has mean and variance λ/n. Then the CLT gives

$$\frac{X - \lambda}{\sqrt{\lambda}} \stackrel{\cdot}{\sim} N(0, 1),$$

so that

$$X \stackrel{\cdot}{\sim} N(\lambda, \lambda). \tag{2}$$

This approximation is reasonable for large λ, most often given as $\lambda > 10$.

2.3.3 Continuity Correction

Note that in the last two cases, we are approximating a discrete distribution with a continuous distribution. As a result a *continuity correction* should be implemented when using the approximations. In particular, let X be a Binomial or Poisson rv; and let Y be the corresponding continuous (approximate) Normal rv. We then set,

$$\mathbb{P}(X = x) \approx \mathbb{P}(x - 0.5 \leq Y \leq x + 0.5),$$

for $x \in \mathbb{N}$.

It follows that,

- $\mathbb{P}(X \geq x) \approx \mathbb{P}(Y \geq x - 0.5)$;
- $\mathbb{P}(X \leq x) \approx \mathbb{P}(Y \leq x + 0.5)$;
- $\mathbb{P}(x_1 \leq X \leq x_2) \approx \mathbb{P}(x_1 - 0.5 \leq Y \leq x_2 + 0.5)$.

Example

Suppose that $X \sim Po(25)$ and we wish to estimate $\mathbb{P}(26 \leq X \leq 30)$, using a Normal approximation. We define a rv $Y \sim N(25, 25)$. Then,

$$\begin{aligned}
\mathbb{P}(26 \leq X \leq 30) &\approx \mathbb{P}(25.5 \leq Y \leq 30.5) \qquad \text{using a continuity correction} \\
&= \mathbb{P}\left(\frac{25.5-25}{5} \leq Z \leq \frac{30.5-25}{5}\right) \qquad \text{where } Z \sim N(0,1) \\
&= \mathbb{P}(0.1 \leq Z \leq 1.1) \\
&= \mathbb{P}(Z \leq 1.1) - \mathbb{P}(Z \leq 0.1) \\
&= 0.8643 - 0.5398 \\
&= 0.3245.
\end{aligned}$$

These probabilities can be calculated in `R` using,

```
> pnorm(1.1,0,1)
[1] 0.864334
> pnorm(0.1, 0, 1)
[1] 0.5398278
```

Note that using the exact distribution for X we obtain, $\mathbb{P}(26 \leq X \leq 30) = \mathbb{P}(X \leq 30) - \mathbb{P}(X \leq 25) = 0.8633 - 0.5529 = 0.3103$. These can be obtained in `R` using,

```
> ppois(30,25)
[1] 0.8633089
> ppois(25,25)
[1] 0.5529214
```

2.4 Assessing Normality

We often assume that data are observed from a Normal distribution - but we should check this! There are rigorous tests, such as the Kolmogorov-Smirnov test or the Shapiro Wilk test, available in many statistics packages, such as SPSS or R. (Note that no test will *prove* normality - we can only fail to reject the Null hypothesis of normality).

Looking at the histogram of the observations is a straightforward and informal way to assess normality. Another graphical method, based on the idea of "Normal scores", is the following. Suppose that we observe two samples $x_1, \ldots, x_n$ and $z_1, \ldots, z_n$ both from a $N(0,1)$ distribution. If we order the samples in ascending order $x_{[1]}, \ldots, x_{[n]}$ and $z_{[1]}, \ldots, z_{[n]}$, we would expect the values to be similar. Thus, plotting these values against each other we would expect to see something close to a straight line. (This is called a Q-Q (quantile-quantile) plot).

Now, any Normal rv is simply a linear transformation of a $N(0,1)$ distribution. This means that for a random sample of observations $x_1, \ldots, x_n$ from any Normal distribution, plotting the sorted x values against a sorted sample $z_1, \ldots, z_n$ from a $N(0,1)$ distribution, should result in something close to a straight line. (If the $x_1, \ldots, x_n$ values are not from a Normal distribution, the resulting plot would not be a straight line.) However, there will be random variability in the simulation of the $z_1, \ldots, z_n$ values. We can remove this by using theoretical quantile values, rather than simulating the z values. This is called a *normal scores plot*. This plot can be obtained in `R` using the `qqnorm` command. See Handout 2 for further details and examples. Note that if the set of x values do not appear to be Normally distributed, we may consider a transformation of these values, such as log, exp, $\sqrt{}$, ^{2}. We can then test whether the transformed values are normally distributed, and if so, use the transformed values.

2.5 Distributions related to the Normal Distribution

2.5.1 χ^2 Distribution

Definition

Let $Z_1, \ldots, Z_n$ be independent $N(0,1)$ rvs and let $X = \sum_{i=1}^{n} Z_i^2$. Then the rv X has a chi-squared distribution with n degrees of freedom, and write $X \sim \chi^2_n$. Note that X is a continuous rv and can take values $x \geq 0$.

Properties

1. Let $Z \sim N(0,1)$ and let $Y = Z^2$. Then $Y \sim \chi^2_1$.
2. Let $S \sim \chi^2_n$ and $T \sim \chi^2_m$, independently. Then, $S + T \sim \chi^2_{n+m}$.

Notes

1. In `R`, the corresponding functions for the chi-squared distribution are `dchisq, pchisq, qchisq` and `rchisq`.
2. The chi-squared distribution plays an important role in testing "goodness-of-fit" of statistical models (see Honours module e.g. MT3606) and is used for advanced models of data such as generalised linear models (see Honours module Generalised Linear Models and Data Analysis).

Example

Suppose that X, Y and Z are co-ordinates in a 3-dimensional space. They are independently distributed $N(0,1)$ (all measurements are in cm). What is the probability that the point (X, Y, Z) lies more than 3 cm from the origin?

We want to calculate,

$$\mathbb{P}(\sqrt{X^2 + Y^2 + Z^2} > 3).$$

Now, $X^2 + Y^2 + Z^2 \sim \chi^2_3$. So that,

$$\begin{aligned} \mathbb{P}(\sqrt{X^2 + Y^2 + Z^2} > 3) &= \mathbb{P}(X^2 + Y^2 + Z^2 > 9) \\ &= \mathbb{P}(S > 9) \qquad \text{where } S \sim \chi^2_3 \\ &= 0.029. \end{aligned}$$

This can be obtained in `R` using,

```
> 1-pchisq(9,3)
[1] 0.02929089
```

Mean and Variance

Let $S \sim \chi^2_n$, so that $S = \sum_{i=1}^n Z_i^2$, where $Z_1, \ldots, Z_n$ are independent $N(0,1)$ rvs. Thus, we have,

$$\begin{aligned} \mathbb{E}(S) &= \mathbb{E}\left(\sum_{i=1}^n Z_i^2\right) \\ &= \sum_{i=1}^n \mathbb{E}(Z_i^2) \\ &= \sum_{i=1}^n \mathrm{Var}(Z_i) \qquad (\text{since } \mathrm{Var}(Z_i) = \mathbb{E}(Z_i^2) - (\mathbb{E}(Z_i))^2 \text{ and } \mathbb{E}(Z_i) = 0) \\ &= n. \end{aligned}$$

We can apply a similar approach for the variance. However, note that we assume the result $\mathbb{E}(Z^4) = 3$ for $Z \sim N(0,1)$ - this is proved in Honours. We have that,

$$\begin{aligned}
\text{Var}(S) &= \text{Var}\left(\sum_{i=1}^{n} Z_i^2\right) \\
&= \sum_{i=1}^{n} \text{Var}(Z_i^2) \qquad \text{since the } Z_i\text{'s are independent} \\
&= \sum_{i=1}^{n} (\mathbb{E}(Z_i^4) - (\mathbb{E}(Z_i^2))^2) \\
&= \sum_{i=1}^{n} (3-1) \\
&= 2n.
\end{aligned}$$

2.5.2 F Distribution

Definition

Let U and V be independent rvs, such that $U \sim \chi^2_m$ and $V \sim \chi^2_n$. We define the rv,

$$X = \frac{U/m}{V/n}.$$

Then, X has an F distribution with m and n degrees of freedom, and write $X \sim F_{m,n}$.

Notes

1. Let $W \sim F_{m,n}$, then $1/W \sim F_{n,m}$. Also, for $n > 2$, $\mathbb{E}(W) = n/(n-2)$.
2. Let $W \sim F_{m,n}$ and $F_{m,n;\alpha}$ denote the upper α quantile of the $F_{m,n}$ distribution, so that $\mathbb{P}(W \geq F_{m,n;\alpha}) = \alpha$. The lower quantile can then be calculated using,

   $$F_{m,n;1-\alpha} = 1/F_{n,m;\alpha}.$$

   (This follows from (1) above).
3. The `R` commands relating to the F distribution are `df`, `pf`, `qf` and `rf`.
4. We will use the F-distribution in Section 2.7.4 (to test whether the variances of two populations are equal).

2.5.3 t Distribution

Definition

Let Z and Y be independent rvs, such that $Z \sim N(0,1)$ and $Y \sim \chi^2_n$. If we let,

$$T = \frac{Z}{\sqrt{Y/n}} \tag{3}$$

then, T has a t-distribution with n degrees of freedom (df), and we write $T \sim t_n$.

Notes

1. If $T \sim t_n$, then $\mathbb{E}(T) = 0$ and $\text{Var}(T) = n/(n-2)$ for $n > 2$.
2. The shape of the p.d.f. depends on n and looks similar to a Normal distribution, but with "heavier" tails. The pdf is symmetrical about 0.
3. As $n \to \infty$, $t_n \to N(0,1)$. In practice, we often use the approximation $t_n \stackrel{\cdot}{\sim} N(0,1)$ for $n \geq 30$.

4. Standard notation is to let $t_{n;\alpha}$ denote the upper α quantile of the t_n distribution, so that $\mathbb{P}(T \geq t_{n;\alpha}) = \alpha$. The distribution is symmetrical about 0, so that $t_{n;\alpha} = -t_{n;1-\alpha}$.

5. The `R` commands relating to the t distribution are `dt`, `pt`, `qt` and `rt`.

6. The t distribution arises for samples from Normal distribution with unknown mean and variance.

2.5.4 Important Result - Normal distributional result

Theorem

Let $\bar{X}$ and S^2 be the sample mean and variance of a random sample of size n from the $N(\mu, \sigma^2)$ distribution. Then,

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1) \tag{4}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{5}$$

$$\bar{X} \text{ and } S^2 \text{ are independent.} \tag{6}$$

Proof is omitted (see Honours).

Note that (6) is quite remarkable. It only holds for Normal distributions.

Consider the statistic,

$$\begin{aligned} T &= \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \\ &= \frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}. \end{aligned}$$

Now, the numerator has a $N(0,1)$ distribution, and the denominator is of the form, $\sqrt{Y/(n-1)}$ where $Y \sim \chi^2_{n-1}$. Hence, $T \sim t_{n-1}$ - see (3), i.e. ,

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}. \tag{7}$$

Example

Let $X_1, \ldots, X_n$ be independent $N(0, 100)$ random variables. Which of the following distributional statements is true?

(a) $\sum_{i=1}^n X_i^2 \sim \chi^2_n$.

(b) $\frac{2X_1^2}{X_2^2+X_3^2} \sim F_{1:2}$.

(c) $\frac{10\sqrt{n}X_1}{\sqrt{\sum_{i=1}^n X_i^2}} \sim t_n$.

(d) $\sum_{i=1}^n 10X_i \sim N(0, 1000)$.

Solution: (b) is true since we can rewrite $\frac{2X_1^2}{X_2^2+X_3^2} = \frac{\left(\frac{X_1}{10}\right)^2/1}{\left(\left(\frac{X_2}{10}\right)^2+\left(\frac{X_3}{10}\right)^2\right)/2}$, so that we have $\frac{\chi^2_1/1}{\chi^2_2/2}$. (a) is false with the χ^2 distribution obtained when summing independent $N(0,1)$ random variables; (c) can immediately be seen to be false since the random variables in the denominator and numerator are not independent (they both contain X_1); (d) is wrong since $\text{Var}\left(\sum_{i=1}^{10} 10X_i\right) = \sum_{i=1}^{10} 100\text{Var}(X_i)$ $= 100 \times 10 \times 100 = 10000$.

2.6 Confidence interval for the mean of the $N(\mu, \sigma^2)$ distribution

Suppose that we have independent rvs $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then, we wish to estimate μ and obtain a $100(1-\alpha)\%$ CI for μ.

Known variance

Recall that in Section 1.3.1 we considered the case, where we assumed that σ^2 is known.

We estimate μ by $\bar{X}$ and obtain the $100(1-\alpha)\%$ CI,

$$\bar{x} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}}.$$

Unknown variance

We now consider the case where σ^2 is unknown. We can estimate the variance, σ^2 by the sample variance S^2. Then, using the result in (7)

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}.$$

For a $100(1-\alpha)\%$ CI for μ, we can then use,

$$\mathbb{P}\left(-t_{n-1;\alpha/2} \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq t_{n-1;\alpha/2}\right) = 1 - \alpha,$$

where $t_{n-1;\alpha}$ satisfies $\mathbb{P}(T \geq t_{n-1;\alpha}) = \alpha$, for $T \sim t_{n-1}$ and can be obtained from `R`.

Rearranging the expression, we obtain,

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

In other words, a $100(1-\alpha)\%$ CI for μ is,

$$\left(\bar{x} - \frac{t_{n-1;\alpha/2}s}{\sqrt{n}}, \bar{x} + \frac{t_{n-1;\alpha/2}s}{\sqrt{n}}\right),$$

or equivalently,

$$\bar{x} \pm t_{n-1;\alpha/2}\frac{s}{\sqrt{n}}.$$

Example

We observe $x_1, \ldots, x_{10}$, independently from a $N(\mu, \sigma^2)$ distribution, with,

$$\sum_{i=1}^{10} x_i = 65$$

$$\sum_{i=1}^{10} x_i^2 = 998.5.$$

Case I: σ^2 known.

Suppose that we know that $\sigma^2 = 64$. Then, a 95% confidence interval for μ is,

$$\bar{x} \pm \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \quad = \quad 6.5 \pm 1.96 \times \frac{8}{\sqrt{10}}.$$

Thus we obtain a 95% CI for μ of (1.542, 11.458).

Case II: σ^2 unknown.

We estimate σ^2 by $s^2 = 64$. The corresponding 95% CI for μ is,

$$\bar{x} \pm \frac{t_{n-1;\alpha/2}S}{\sqrt{n}} = 6.5 \pm 2.2622 \times \frac{8}{\sqrt{10}}.$$

Thus we obtain a 95% CI for μ of (0.777, 12.223).

2.7 Hypothesis Tests for Normal distributions

Recall that within hypothesis tests there are two possible approaches - calculating the p-value (reject H_0 if this is $\leq \alpha$, for significance level α); or calculate the critical region (the set of values for which we reject H_0).

Known variance

Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ rvs, where σ^2 is known. We have already studied this in the previous chapter. Suppose that we wish to test,

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0.$$

We define the test statistic $T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$. Then, under H_0, $T \sim N(0,1)$, so that the critical region at the α significance level (i.e. values at which we would reject H_0) is given by,

$$|T| = |\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}| \geq z_{\alpha/2}.$$

2.7.1 One sample t-tests (unknown variance)

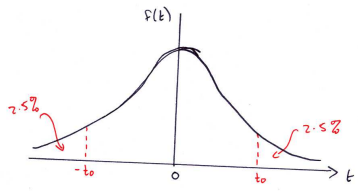
Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ rvs, where σ^2 is unknown. Suppose that we wish to test,

$$H_0 : \mu = \mu_0 \qquad \text{vs} \qquad H_1 : \mu \neq \mu_0.$$

We consider the statistic,

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

By (7) we have that, under the H_0, $T \sim t_{n-1}$. Then, we reject H_0 if $|T| \geq t_0 \equiv t_{n-1,\alpha/2}$, for significance level α.



$$\begin{aligned} & \mathbb{P}(\text{reject } H_0 | H_0 \text{ true}) &&= \alpha \\ \Rightarrow \qquad & \mathbb{P}(-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2} | T \sim t_{n-1}) &&= 1 - \alpha. \end{aligned}$$

Example

A set of 39 observations on pulse rates (in heart beats per minute) has sample mean, $\bar{x} = 70.31$ and sample variance, $s^2 = 90.219$. We wish to test the following hypothesis at the 1% level,

$$H_0 : \mu = 75 \qquad \text{vs} \qquad H_1 : \mu \neq 75.$$

We define the statistic,

$$T = \frac{\bar{X} - 75}{S/\sqrt{n}}.$$

Then, under H_0, $T \sim t_{38}$. Note that the observed test statistic is $t = -3.083593$. We reject H_0 if $|T| \geq t_{38,0.005}$. To calculate $t_{38,0.005}$ in R:

```
> qt(0.995,38)
[1] 2.711558
```

Thus we reject H_0 in favour of H_1 at the 1% level.

Alternatively, we could perform the hypothesis test by calculating the p-value. The p-value is $\mathbb{P}(|T| \geq 3.083593) = 2\mathbb{P}(T \leq -3.083593) = 0.003799$ (using `2*pt(-3.083593,38)`). Thus again we reject at the 1% level. We would reject at all significance levels $\geq 0.38\%$.

Finally, suppose that we are particularly interested in whether the mean pulse rate is higher than 75. Although we reject H_0 that the mean is equal to 75, we would NOT conclude that there is evidence of a higher mean pulse rate since the observed data suggests a lower mean pulse rate (i.e. the data is observed in the left tail, whereas a higher mean pulse rate corresponds to the right-tail).

2.7.2 Paired t-tests

In many circumstances we may have paired data of the form $(X_1, Y_1), \ldots, (X_n, Y_n)$, where the two measurements are dependent for an individual i. Assuming that the observations are normally distributed, and that $E(X_i) = \mu_x$, $E(Y_i) = \mu_y$, we want to test for the $H_0 : \mu_x = \mu_y$. In this situation, we consider the difference between the two measurements on each unit. In particular, we set $D_i = Y_i - X_i$ for $i = 1, \ldots, n$, and assume that each $D_i \sim N(\mu, \sigma^2)$, independently. Thus, the problem reduces to a one-sample t-test (σ^2 is unknown), for the $H_0 : \mu = 0$. The test statistic is $\frac{\bar{D}}{s/\sqrt{n}}$.

Example

The following table provides the corneal thickness in microns of both eyes of patients who have glaucoma in one eye:

Healthy	484	478	492	444	436	398	464	476
Glaucoma	488	478	480	426	440	410	458	460
Difference	4	0	-12	-18	4	12	-6	-16

The corneal thickness is likely to be similar in the two eyes of any single patient, so that the two observations on the same patient cannot be assumed to be independent. We consider the difference between each pair of observations, denoted by d_i. We wish to test,

$$H_0 : \mu = 0 \qquad \text{vs} \qquad H_1 : \mu \neq 0.$$

Under H_0,

$$T = \frac{\bar{D}}{S/\sqrt{n}} \sim t_{n-1}.$$

Note that $\sum_{i=1}^{8} d_i = -32$ and $\sum_{i=1}^{8} d_i^2 = 936$. Thus, we have $\bar{d} = -4$ and $s^2 = 808/7$. Thus, the observed test statistic is -1.05. To calculate the p-value, we calculate $2 \times \mathbb{P}(T \leq -1.05)$, where $T \sim t_7$. Using `R`,

```
> 2*pt(-1.05,7)
[1] 0.3286108
```

Thus, we do not reject H_0 at any reasonable significance level.

Alternatively, the critical region at a 5% significance level is $|T| \geq t_{7;0.025} = 2.3646$.

2.7.3 Two-sample t-tests

Suppose that we have two sets of independent rvs, $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, where we assume,

$$X_i \sim N(\mu_X, \sigma^2) \qquad \text{and} \qquad Y_j \sim N(\mu_Y, \sigma^2).$$

We wish to test:

$$H_0 : \mu_X = \mu_Y \qquad \text{vs} \qquad H_1 : \mu_X \neq \mu_Y.$$

Now we know that,

$$\bar{X} \sim N\left(\mu_X, \frac{\sigma^2}{n}\right) \qquad \text{and} \qquad \bar{Y} \sim N\left(\mu_Y, \frac{\sigma^2}{m}\right).$$

Using the standard result for (independent) Normally distributed rvs,

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

If H_0 is true,

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(0,1).$$

However, we typically will not know σ^2, and need to estimate it.

We define the pooled sample variance:

$$\begin{aligned} S_p^2 &= \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2}{m+n-2} \\ &= \frac{(n-1)S_X^2 + (m-1)S_Y^2}{(m+n-2)}. \end{aligned}$$

This is an unbiased estimator of σ^2 (see Examples Class - Week 2).

Now we have that,

$$\begin{aligned} \frac{(n-1)s_X^2}{\sigma^2} &\sim \chi^2_{n-1}; \\ \frac{(m-1)s_Y^2}{\sigma^2} &\sim \chi^2_{m-1}; \\ \Rightarrow \quad \frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} &\sim \chi^2_{m+n-2}. \end{aligned}$$

since S_X^2 and S_Y^2 are independent (since the random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent). (Note - see §2.5.4 for distributional results).

Since we assume that the rvs are Normally distributed the sample variances are independent of $\bar{X}$ and $\bar{Y}$. Using the usual argument, under H_0,

$$T = \frac{\frac{\bar{X}-\bar{Y}}{\sigma\sqrt{\frac{1}{m}+\frac{1}{n}}}}{\sqrt{\frac{\frac{(n-1)S_X^2+(m-1)S_Y^2}{\sigma^2}}{m+n-2}}} = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}.$$

We can conduct a hypothesis test, using the statistic above.

Example

Suppose that we observe two samples from two Normal populations (with common variance), such that $\bar{x} = 80.02$, $s_X^2 = 0.024$, with $n = 13$ and $\bar{y} = 79.98$, $s_Y^2 = 0.031$, with $m = 8$. Then $\bar{x} - \bar{y} = 0.04$ and the pooled sample variance is $s_p^2 = \frac{12\times 0.024 + 7\times 0.031}{19} = 0.027$. The observed test statistic is then $\frac{0.04}{\sqrt{0.027}\sqrt{1/13+1/8}} = 0.542$.

The critical region is $|T| \geq t_{19;0.025} = 2.093$. Thus we do not reject H_0 at the 5% level.

Alternatively, the p-value can be calculated, using $\mathbb{P}(|T| \geq 0.542) = 2 \times \mathbb{P}(T \geq 0.542)$

```
> 2*(1-pt(0.542,19))
[1] 0.5941187
```

2.7.4 F-test for equality of variance

For the two-sample t-test we assumed that the variances were equal between the two populations the samples came from. This can be formally tested using an F-test. Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent Normal rvs. Let σ_X^2 and σ_Y^2 denote the population variances of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively.

We wish to test:

$$H_0 : \sigma_X^2 = \sigma_Y^2 \qquad \text{vs} \qquad H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

Consider the test statistic,

$$F = \frac{S_X^2}{S_Y^2}.$$

If H_0 is true, then $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, and,

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

independently, so that,

$$F = \frac{\frac{(n-1)S_X^2}{\sigma^2}/(n-1)}{\frac{(m-1)S_Y^2}{\sigma^2}/(m-1)} = \frac{S_X^2}{S_Y^2} \sim F_{n-1,m-1}.$$

Alternatively, if H_0 is true,

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

Then, to test the hypothesis, the observed value can be compared with the lower and upper quantile that will define the critical region for the corresponding F distribution.

Example

Consider two random samples from Normal distributions, $x_1, \ldots, x_{16}$ and $y_1, \ldots, y_{11}$, where $s_X^2 = 20$ and $s_Y^2 = 30$. The observed test statistic is $\frac{s_X^2}{s_Y^2} = \frac{20}{30} = 2/3$. At the 5% level, the critical region is defined by $F_{15,10;0.025}$ so that $P(F > F_{15,10;0.025}) = 0.025$, and $F_{15,10;0.975}$ so that $P(F > F_{15,10;0.975}) = 0.975$:

```
> qf(0.025,15,10)
[1] 0.3267764
> qf(0.975,15,10)
[1] 3.521673
```

Thus, $F_{15,10;0.975} = 0.32$ and $F_{15,10;0.025} = 3.52$. The test statistic falls inbetween those points, and so we do not reject H_0 at the 5% level.