Contents

3.5.3 Two Sample t Tests

Setup: Two Samples

We now focus on a scenario where we have two *independent* samples from possibly different populations. Our interest focuses on the difference in population means $\mu_x - \mu_y$. We assume both samples come from normal sampling distributions that may or may not differ in their mean and variance parameters.

$$\begin{aligned} X_1, \ldots, X_n &\overset{\text{iid}}{\sim} N\left(\mu_x, \sigma_x^2\right) \\ Y_1, \ldots, Y_m &\overset{\text{iid}}{\sim} N\left(\mu_y, \sigma_y^2\right) \end{aligned} \tag{1}$$

or equivalently

$$\begin{aligned} X_i &= \mu_x + \epsilon_i, \quad i = 1, \ldots, n \quad \epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma_x^2\right) \\ Y_j &= \mu_y + \delta_j, \quad j = 1, \ldots, m \quad \delta_j \overset{\text{iid}}{\sim} N\left(0, \sigma_y^2\right) \end{aligned}$$

In both cases, our theoretical knowledge tells us the following facts:

$$\begin{aligned} \bar{X} &\sim N\left(\mu_x, \sigma_x^2/n\right) \\ \bar{Y} &\sim N\left(\mu_y, \sigma_y^2/m\right), \text{ and} \\ \bar{X} - \bar{Y} &\sim N\left(\mu_x - \mu_y, \sigma_x^2/n + \sigma_y^2/m\right). \end{aligned} \tag{2}$$

There are four scenarios to handle, as summarized in the following table. We will address each scenario in turn, presenting some theorems to justify the proposed tests. We will also explore the implications in quick examples.

Variance Assumptions		Test	Distribution
Common variance	Variance known	Z test	$N(0,1)$
$\sigma_x^2 = \sigma_y^2 = \sigma^2$	Variance unknown	t test	t_{n+m-2}
Different variances	Variances known	Z test	$N(0,1)$
$\sigma_x^2 \neq \sigma_y^2$	Variances unknown	t test	t_{df}

1. **Common variance, $\sigma_x^2 = \sigma_y^2 = \sigma^2$.**

   (a) **Variance known.**

   Common Variance, Variance Known

   Using the distribution in Eq. (2) for $\bar{X} - \bar{Y}$ and applying equivalence of variances, we have statistic

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

   with CI for $\mu_x - \mu_y$ given by

   $$Z \pm \phi_{\frac{\sigma+1}{2}} \sigma \sqrt{\frac{1}{n} + \frac{1}{m}},$$

   rejection (tail) region for $H_0 : \mu_x - \mu_y = A$ consisting of Z with

   $$|Z| > \phi_{\frac{\sigma+1}{2}},$$

   and p-value $P(Z > |z|)$ for observed value of the test statistic z, computed in `R` as

   ```
   R> 2*pnorm(-abs(z)).
   ```

   (b) **Variance unknown.**

Common Variance, Variance Unknown

If the variance is unknown, we need to obtain an estimate of σ^2. There are actually two estimates available, S_x^2 and S_y^2, but under the assumption of common variances, they both estimate the same quantity. It would be best if we could combine these estimates into a single estimator. The *pooled variance* is such an estimator:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}$$

This estimator weights the estimates obtained from each sample by its sample size, which seems logical. There is a theorem that gives us a statistic and sampling distribution for this scenario.

Theorem 10. *Given two samples as in Eq. 1, the statistic*

$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$

has a t distribution with $n + m - 2$ degrees of freedom.

Proof. We know, from Thm **??**, that $\frac{(n-1)S_x^2}{\sigma^2} \sim \chi^2_{n-1}$ and $\frac{(m-1)S_y^2}{\sigma^2} \sim \chi^2_{m-1}$ have chi-squared distributions. In addition, these quantitites are independent because the samples are independent. Therefore,

$$\frac{(n-1)S_x^2 + (m-1)S_y^2}{\sigma^2} = \frac{(n+m-2)S_p^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

by the reproductive property of chi-square random variables. Consider

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

and

$$V = \sqrt{\frac{S_p^2}{\sigma^2}} = \frac{S_p}{\sigma}.$$

Then, Thm. **??** yields $T = U/V \sim t_{n+m-2}$. □

The above-defined t statistic allows us to construct CI for $\mu_x - \mu_y$ of

$$T \pm t_{n+m-2}\left(\frac{1+\alpha}{2}\right) S_p\sqrt{\frac{1}{n} + \frac{1}{m}}$$

where $t_{n+m-2}\left(\frac{1+\alpha}{2}\right)$ is the $\left(\frac{1+\alpha}{2}\right)$th quantile of the t_{n+m-2} distribution computed as `R> qt((1+alpha)/2, df=n+m-2)`. The rejection regions (tails) for $H_0 : \mu_x - \mu_y = A$ include values of T where

$$|T| > t_{n+m-2}\left(\frac{1+\alpha}{2}\right)$$

and the p-value for reject $H_0 : \mu_x = \mu_y$ is $P(T > |t|)$ for $T \sim t_{n+m-2}$ and the observe statistic t, computed as

```
R> 2*pt(-abs(t)).
```

Example:

> Suppose two methods, A and B, measure the latent heat of fusion of ice, the change in heat, measured in calories/gram, when ice moves from -0.72^oC to water at 0^oC. Find the CI for $\mu_A - \mu_B$ and test $H_0 : \mu_a = \mu_B$ when $\bar{X}_A = 80.02, \bar{X}_B = 79.98, S_A = 0.024, S_B = 0.031, n = 13$, and $m = 8$, assuming the variances are equal.

The pooled variance is

$$S_p^2 = \frac{12 \times 0.024^2 + 7 \times 0.031^2}{21 - 2} = 0.0007178$$

The estimated population mean difference is $\bar{X}_A - \bar{X}_B = 0.04$ with corresponding sample variance $S_{\bar{X}_A - \bar{X}_B} = S_p\sqrt{\frac{1}{13} + \frac{1}{8}} = 0.012$. The critical value $t_{19}(0.975)$ is computed as `R> qt(0.975, df=19)` and is 2.093. Therefore, the CI are

$$0.04 \pm 2.093 \times 0.012 = (0.015, 0.065)$$

The hypothesis test uses statistic

$$t = \frac{\bar{X}_A - \bar{X}_B}{S_{\bar{X}_A - \bar{X}_B}} = \frac{0.04}{0.012} = \frac{10}{3},$$

with p-value given by `R> 2*pt(-10/3, df=19)`, which is 0.00349. We conclude that at least one of the two measurement methods is a biased estimator of the heat of fusion.

2. **Difference variances, $\sigma_x^2 \neq \sigma_y^2$.**

   (a) **Variance known.**

   Different Variances, Variances Known
   When variance is known, we use the known distribution (Eq. 2) of $\bar{X} - \bar{Y}$ to form statistic

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim N(0, 1)$$

   with CI for $\mu_x - \mu_y$ given by

$$\bar{X} - \bar{Y} \pm \phi_{\frac{1+\alpha}{2}} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}.$$

   Rejection regions and p-value are computed as for the case with common, known variance.

   (b) **Variance unknown.**

   Different Variances, Variances Unknown
   When variance is unknown, it is natural to estimate

$$\text{Var}(\bar{X} - \bar{Y}) \widehat{=} \frac{S_x^2}{n} + \frac{S_y^2}{m}$$

   There is a theorem that justifies this choice that we will state without proof.

   Theorem 11. *Given the two-sample setup of Eq. 1, the statistic*

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \dot{\sim} t_{df}$$

   has a t-distribution with degrees of freedom given by

$$df = \frac{\left[\left(\frac{S_x^2}{n}\right) + \left(\frac{S_y^2}{m}\right)\right]^2}{\frac{(S_x^2/n)^2}{n} + \frac{\left(S_y^2/m\right)^2}{m}} - 2$$

Example:

Returning to the heat of fusion example, repeat the analysis *without assuming equal variances*.

First, we must compute df as

$$df = \frac{\left[\frac{0.024^2}{13} + \frac{0.031^2}{8}\right]^2}{\frac{\left(\frac{0.024^2}{13}\right)^2}{13} + \frac{\left(\frac{0.031^2}{8}\right)^2}{8}} - 2 = 11.83$$

We can round this to 12, although R will handle non-integer degrees of freedom. The CI become

$$\bar{X}_A - \bar{X}_B \pm t_{11.83}\left(\frac{1+\alpha}{2}\right)\sqrt{\frac{0.024^2}{13} + \frac{0.031^2}{8}} = (0.012, 0.068)$$

The hypothesis test uses statistic

$$\frac{\bar{X}_A - \bar{X}_B}{0.0128} = 3.125$$

which generates p-value 0.008911151 via `R> 2*pt(-3.125, df=11.83).`

Normal Probability Plot

The preceding results depend on starting with two samples of normally distributed random variables. For non-normal sampling distributions, the CLT gives us the result asymptotically, although sample size must be large to be sure. Thus, any data analysis should start with an examination of the data to assess whether the data are approximately normal. We now derive one of the most common graphical techniques for assessing normality, the normal probability plot. We mentioned it briefly when covering graphical summaries of data.

Theorem 12. *Suppose random variable $X \sim F(x)$ has strictly increasing cdf and let $Z = F(X)$, then $Z \sim Unif(0,1)$*

Proof.

$$P(Z \le z) = P\left[F(X) \le z[= P\left[X \le F^{-1}(z)\right] = F\left[F^{-1}(z)\right] = z$$

The cdf of the standard normal r.v. is $F(z) = z$, thus $Z \sim$Unif(0,1). □

Suppose $X \sim F(x)$ is the hypothesized cdf (in our case, some kind of normal). Given $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F(x)$, the ordered values, from smallest to largest, are written as

$$X_{(1)}, \ldots, X_{(n)}$$

where $X_{(k)}$ is the kth smallest value. We claim $X_{(k)}$ estimates the $\frac{k}{n+1}$ quantile. To verify, remember the sample quantile is the first observed value X^* such that

$$\frac{\#\{X_i \le X^*\}}{n} \ge \frac{k}{n+1}$$

But, indeed, $X_{(k)}$ does satisfy

$$\frac{\#\{X_i \le X_{(k)}\}}{n} \ge \frac{k}{n+1}$$

while $X_{(k-1)}$ does not, so it is the first.

Now $F(X_i)$ are uniformly distributed *if and only if* $F(x)$ is the true cdf of the X_i. In particular, the quantiles $X_{(k)}$ should map, via $F(X_{(k)})$, to the quantiles of a Unif$(0,1)$ distribution, namely $\frac{k}{n+1}$. Thus, if we plot

points $\left(F(X_{(k)}), \frac{k}{n+1}\right)$, they should *fall along the line* $y = x$ if $F(x)$ is the sampling distribution of X_i. Equivalently, we could plot points $\left(X_{(k)}, F^{-1}\left(\frac{k}{n+1}\right)\right)$, which has the advantage of plotting the data on its original, presumably interpretable, scale.

Standardization. If we standardize the data $\frac{X_i-\mu}{\sigma}$ first, then $F(x)$ becomes the standard normal cdf $\Phi(x)$. If we do not use the standard normal cdf, then we also need μ and σ to compute the cdf $F(x)$. In either case, μ and σ are unknown, but we can replace them with their sample estimators. Still, we may prefer to plot $\left(X_{(k)}, \Phi^{-1}\left(\frac{k}{n+1}\right)\right)$, which requires no knowledge of μ or σ. It again has the advantage that the x-axis is on the original data scale. While the data should still form a line, it is not the $y = x$ identity line.

There are many variants of the normal probability plot, but the idea is the same: Curvature in the points suggests non-normal data.

Example: Of Mice and Iron

We now demonstrate how to perform a two-sample analysis of data where normality assumptions may be under doubt.

Suppose mice are given either radioactive Fe^{2+} or radioactive Fe^{3+}. Then, a few days later, their radioactivity is measured in order to determine which iron supplement is better retained by the body. The data is available as Fe.Rtxt.

Our first concern is to check that the data are normal, since all two-sampled results have assumed normal sampling distributions. We can plot histograms and see that the data seem to be skewed right and truncated on the left, since negative radioacitivities are not possible (Fig. 4).

We can see the normal probability plots (Fig. 4) show some curvature, so we are concerned about possible non-normality.

We know of one formal test for distributional assumptions: the goodness-of-fit. We will bin the data into the quartiles of a normal distribution. Which normal distribution? We must estimate its two parameters with the sample mean and sample variance. For the Fe^{2+} data, $\bar{X}_2 = 9.632222$ and standard deviation $S_2 = 6.691215$. Each parameter we estimate takes away one degree of freedom. The quartiles of $N(9.6, 6.69^2)$ are $5.09, 9.6, 14.1$. We count the number of observations in each interval.

	$X_i < 5.09$	$5.09 \le X_i < 9.6$	$9.6 \le X_i < 14.1$	$14.1 \le X_i$
Expected	4.5	4.5	4.5	4.5
Observed	4	9	2	3

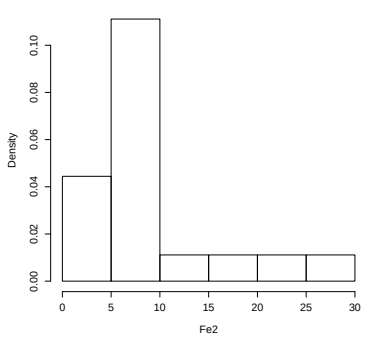
Expected counts are a little low, but we'll proceed with the chi-squared goodness-of-fit test anyway. The statistic is

$$X^2 = +\frac{(|4.5-4|-0.5)^2}{4.5} + \frac{(|4.5-9|-0.5)^2}{4.5} + \frac{(|4.5-2|-0.5)^2}{4.5} + \frac{(|4.5-3|-0.5)^2}{4.5} = 4.722222$$

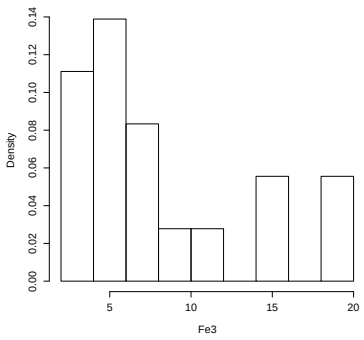
where we have applied the *Yate's continuity correction* to each term. The correction is most important for low count data, and it corrects for the fact that expectations are continuous, but observed data are always discrete. The degrees of freedom are $df = 4 - 1 - 2 = 1$, so the p-value for rejecting the null hypothesis of normally distributed data is 0.02977524. At a traditional α-level of 0.05, we reject the null hypothesis, and develop further worries about using the t-test.

We proceed with the t-test,but do so with great concern. The sample size $n = m = 18$ does not exceed the rule-of-thumb 30 that we like for trusting CLT asymptotics. The sample standard deviations are $S_2 = 6.69$ and $S_3 = 5.45$. We hesitate to assume common variance, so compute $df = 32.79$. The t statistic comes to 0.70 and the p-value is 0.49.

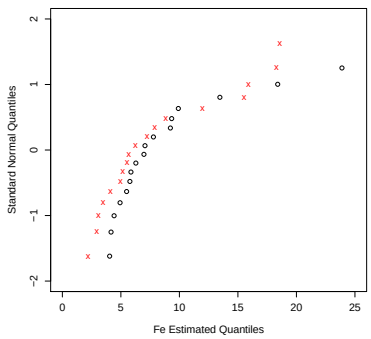
Transformation. One solution for the non-normality is to transform the observed data such that they become more normal-like. A common problem is right skew, especially for data that is constrained to live in the positive half of the real line. To correct right skew, one may use the $\log()$ or $\sqrt{}$ transformations, since both tend to shrink large values and expand small values.



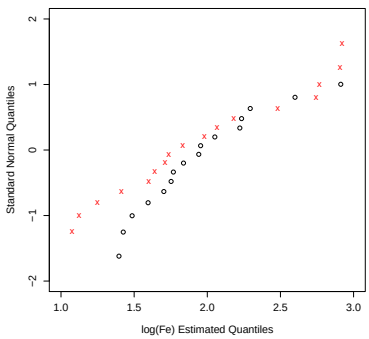
(a) Fe^{2+}



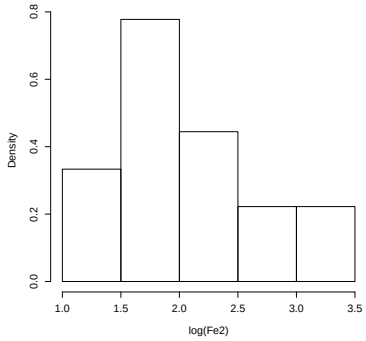
(b) Fe^{3+}



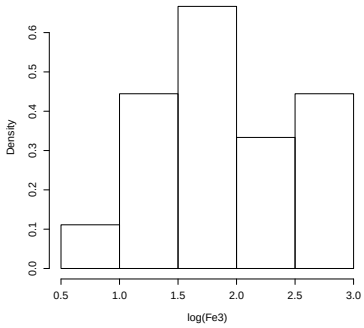
(c) Fe^{2+} ='o', Fe^{3+} ='x'



(d) $\log(Fe^{2+})$ ='o', $\log(Fe^{3+})$ ='x'



(e) $\log(Fe^{2+})$



(f) $\log(Fe^{3+})$

Figure 4: Histograms and Normal Probabililty Plots of Iron Data

The result of logging the iron data is also shown in Fig. 4. Much of the non-normality disappears, but some curvature remains in the normal probability plots. It is likely that further goodness-of-fit tests will not be significant.