Chapter 9. Hypothesis Testing

Suppose we want to know if exposure to asbestos is associated with lung disease. We take some rats and randomly divide them into two groups. We expose one group to asbestos and leave the second group unexposed. Then we compare the disease rate in the two groups. Consider the following two hypotheses:

The Null Hypothesis H_0 : The disease rate is the same in the two groups.

The Alternative Hypothesis H_1 : The disease rate is not the same in the two groups.

If the exposed groups has a much higher rate of disease than the unexposed group then we will reject the null hypothesis. This is an example of hypothesis testing.

More specifically, we may fix upon a parametric family $\mathrm{P}_{X|\theta}$ and then test whether hypothesised parameter values for θ are plausible; that is, test whether we could reasonably assume $\theta = \theta_0$ for some particular value θ_0. For example, if $X \sim \mathrm{N}(\mu, \sigma^2)$ we may wish to test whether $\mu = 0$ is plausible in light of the data.

In general, we are attempting to address the ill-defined question "Should we believe this hypothesis in light of the observations?" In hypothesis testing we move toward a concrete answer by asking: If the null hypothesis were true, what is the probability that the observed data looks this "unusual" or "weird"?

Formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

We call the H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

To test the validity of H_0, we first choose a test statistic $T(\underline{X})$ of the data for which we can find the distribution, P_T, under H_0. One of the difficulties in hypothesis testing is to find an appropriate test statistic T.

Then, we identify a rejection region $R \subset \mathbb{R}$ of low probability values of T *under the assumption that H_0 is true*, i.e. a region R such that

$$\mathrm{P}(T \in R \,|\, H_0) = \alpha$$

for some small probability α (often 5%).

Finally, we calculate the observed test statistic $t(\underline{x})$ for our observed data $\underline{x}$.

- If $t \in R$ we "reject the null hypothesis at the $100\alpha\%$ level".
- If $t \notin R$ we "do not reject (retain) the null hypothesis at the $100\alpha\%$ level".

For each possible **significance level** $\alpha \in (0, 1)$, a hypothesis test at the $100\alpha\%$ level will result in either rejecting or not rejecting H_0.

- As $\alpha \to 0$ it becomes less and less likely that we will reject our null hypothesis, as the rejection region is becoming smaller and smaller.
- Similarly, as $\alpha \to 1$ it becomes more and more likely that we will reject our null hypothesis.

For any given data, we might, therefore, be interested in identifying the critical significance level which marks the threshold between us rejecting and not rejecting the null hypothesis. This is known as the p-value of the data. Smaller p-values suggest stronger evidence against H_0.

Interpretation

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that the person is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0.

Important: We can never prove that any hypothesis is true or false. We only have the ability to *reject* a hypothesis. That is, we may show that the observed data is very unlikely *were the null hypothesis true*.

A well chosen rejection region will have relatively high probability under H_1, whilst retaining low probability under H_0. The alternative hypothesis is not strictly necessary but in practice it is essential to have an alternative hypothesis in order to decide on what test to perform (i.e. what test statistic T and rejection region R to adopt).

9.0.1 Error Rates and Power of a Test

There are two types of error in the outcome of a hypothesis test:

- **Type I**: Rejecting H_0 when in fact H_0 is true. By construction, this happens with probability α. For this reason, the significance level of a hypothesis test is also referred to as the Type I error rate.
- **Type II**: Not rejecting H_0 when in fact H_1 is true i.e. $\beta = \mathrm{P}(T \notin R \,|\, \theta \in \Theta_1)$.

Definition 9.0.1. *The power of a hypothesis test is defined as*

$$1 - \beta = P(T \in R \,|\, \theta \in \Theta_1).$$

A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis** (or **point hypothesis**). A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**. A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

is called a **two-sided test**. A test of the form

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called a **one-sided test**.

Note It would be desirable to find the test with highest power under H_1, among all size α tests. Such a test, if it exists, is called most powerful. Finding most powerful tests is hard and, in many cases, most powerful test don't even exist. Instead of going into detail about when most powerful tests exists, we shall just consider some commonly used tests.

9.1 Testing for a population mean

9.1.1 Normal Distribution with Known Variance

Suppose $X_1, \ldots, X_n$ are i.i.d. $\mathrm{N}(\mu, \sigma^2)$ with σ^2 known and μ unknown. We may wish to test if $\mu = \mu_0$ for some specific value μ_0 (e.g. $\mu_0 = 0$ or $\mu_0 = 9.8$).

Then we can state our null and alternative hypotheses as

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0.$$

Under $H_0 : \mu = \mu_0$, we then know both μ and σ^2. So for the sample mean $\overline{X}$ we have a known distribution for the test statistic

$$Z = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}} \sim \Phi.$$

Shorthand for $\sim N(0,1)$

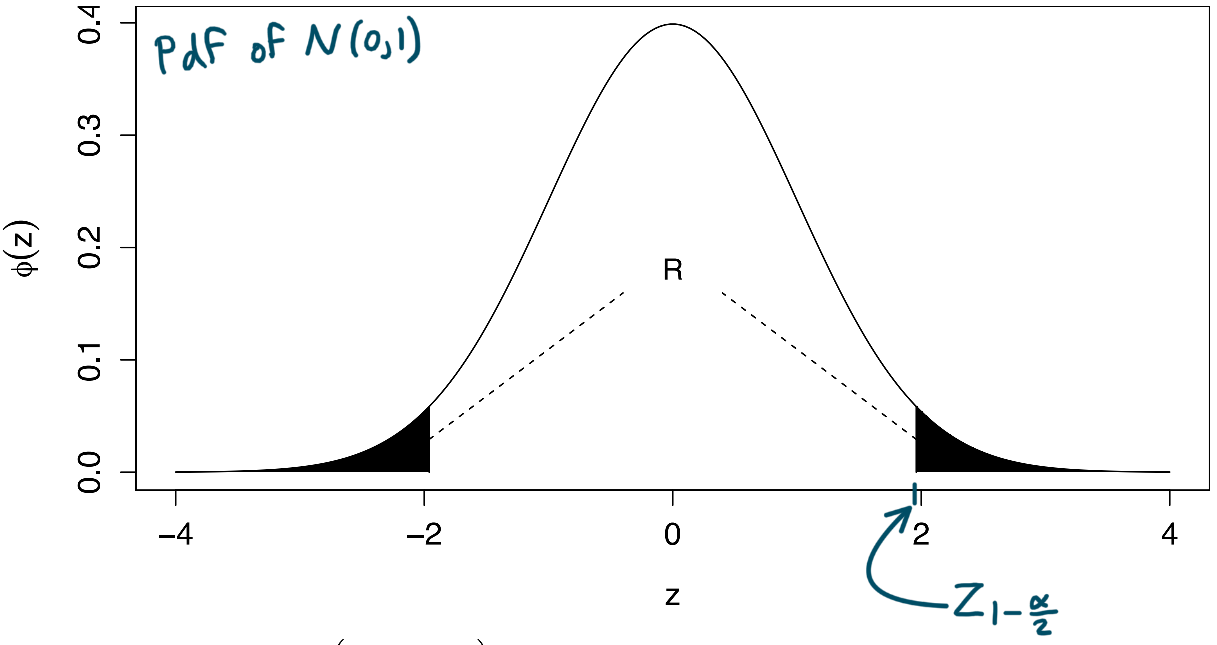
So if we define our rejection region R to be the $100\alpha\%$ tails of the standard normal distribution,

$$\begin{aligned} R &= \left(-\infty, -z_{1-\frac{\alpha}{2}}\right) \cup \left(z_{1-\frac{\alpha}{2}}, \infty\right) \\ &\equiv \left\{ z \;\middle|\; |z| > z_{1-\frac{\alpha}{2}} \right\}, \end{aligned}$$

we have $\mathrm{P}(Z \in R | H_0) = \alpha$.

We thus reject H_0 at the $100\alpha\%$ significance level $\iff$ our observed test statistic satisfies

$$z = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}} \in R.$$



The p-value is given by $2\Big(1-\Phi(|z|)\Big)$. This is the probability under the null hypothesis, that the data looks "even more unusual" than what is observed. In this case, "even more unusual" means further out into the tails of the pdf of the test statistic.

Example A company makes packets of snack foods. The bags are labelled as weighing 454g; of course they won't all be exactly 454g, and let's suppose the variance of bag weights is known to be $70\,\text{g}^2$. The following data show the mass in grams of 50 randomly sampled packets.

> 464, 450, 450, 456, 452, 433, 446, 446, 450, 447, 442, 438, 452, 447, 460, 450, 453, 456, 446, 433, 448, 450, 439, 452, 459, 454, 456, 454, 452, 449, 463, 449, 447, 466, 446, 447, 450, 449, 457, 464, 468, 447, 433, 464, 469, 457, 454, 451, 453, 443

Are these data consistent with the claim that the mean weight of packets is 454g?

1. We wish to test $H_0 : \mu = 454$ vs. $H_1 : \mu \neq 454$. So set $\mu_0 = 454$.
2. Although we have not been told that the packet weights are individually normally distributed, by the CLT we still have that the mean weight of the sample of packets is approximately normally distributed, and hence we still *approximately* have $Z = \dfrac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} \sim \Phi$.
3. Compute the realised value of the test statistic: $\bar{x} = 451.22$ and $n = 50$

$$\Rightarrow z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = -2.350.$$

4. For a 5%-level significance test, we find whether the observed statistic $z = -2.350$ is inside the rejection region $R = (-\infty, -z_{0.975}) \cup (z_{0.975}, \infty) = (-\infty, -1.96) \cup (1.96, \infty)$. Clearly we have $z \in R$, and so at the 5%-level we reject the null hypothesis that the mean packet weight is 454 g. We conclude the test by stating: **there is sufficient evidence to reject the null hypothesis at the 5% level.**

5. At which significance levels would we have not rejected the null hypothesis?

   - For a 1%-level significance test, the rejection region would have been

     $$R = (-\infty, -z_{0.995}) \cup (z_{0.995}, \infty) = (-\infty, -2.576) \cup (2.576, \infty).$$

     In which case $z \notin R$, and so at the 1%-level we would not have rejected the null hypothesis.

   - The p-value is

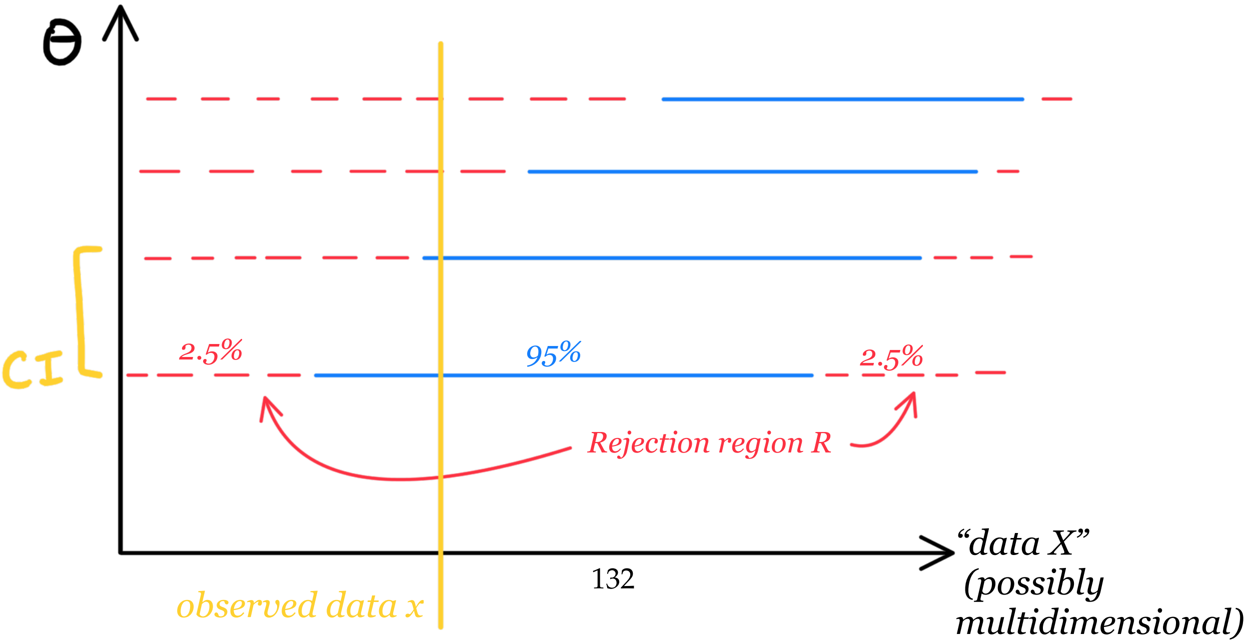
     $$\begin{aligned} &2 \times \{1 - \Phi(|z|)\} = 2 \times \{1 - \Phi(|-2.350|)\} \approx 2(1 - 0.9906) \\ &= 0.019, \end{aligned}$$

     and so we would only reject the null hypothesis for $\alpha > 1.9\%$.

■

Note There is a strong connection between hypothesis testing and confidence intervals. Suppose we have constructed a $100(1-\alpha)\%$ confidence interval for a parameter θ. Then this is precisely the set of values θ_0 for which there would be not be sufficient evidence to reject a null hypothesis $H_0 : \theta = \theta_0$ at the $100\alpha\%$-level.

In other words, we perform a 100α hypothesis test for every possible value of θ. Those values whose hypothesis of θ is not rejected form the $100(1-\alpha)$ confidence interval. This is precisely the picture captured by the Neyman construction in the previous chapter.



9.1.2 Normal Distribution with Unknown Variance

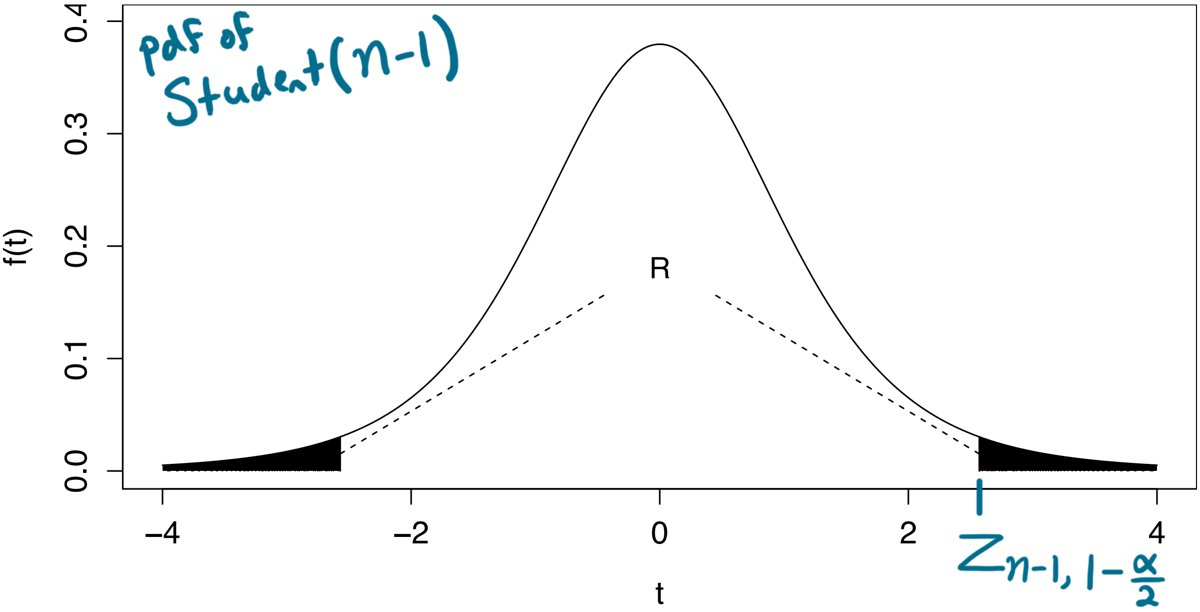
Similarly, if σ^2 in the previous example were unknown, we have that

$$T = \frac{\overline{X} - \mu_0}{s_{n-1}/\sqrt{n}} \sim t_{n-1}.$$

So for a test of $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ at the α level, the rejection region of our observed test statistic $t = \dfrac{\overline{x} - \mu_0}{s_{n-1}/\sqrt{n}}$ is

$$\begin{aligned} R &= \left(-\infty, -t_{n-1,1-\frac{\alpha}{2}}\right) \cup \left(t_{n-1,1-\frac{\alpha}{2}}, \infty\right) \\ &\equiv \left\{ t \;\middle|\; |t| > t_{n-1,1-\frac{\alpha}{2}} \right\}. \end{aligned}$$

Again, we have that $P(T \in R|H_0) = \alpha$.



Example Consider again the snack food weights example. There, we assumed the variance of bag weights was known to be 70. Without this, we could have estimated the variance by

$$s_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = 70.502.$$

Then the corresponding t-statistic becomes

$$t = \frac{\overline{x} - \mu_0}{s_{n-1}/\sqrt{n}} = -2.341,$$

very similar to the z-statistic of before.

And since $n = 50$, we compare with the t_{49} distribution which is approximately $\mathrm{N}(0, 1)$. So the hypothesis test results and p-value would be practically identical. ■

Example A particular piece of code takes a random time to run on a computer, but the average time is known to be 6 seconds. The programmer tries an alternative optimisation in compilation and wishes to know whether the mean run time has changed. To explore this, they run the re-optimised code 16 times, obtaining a sample mean run time of 5.8 seconds and bias-corrected sample standard deviation of 1.2 seconds. Is the code any faster or slower?

1. We wish to test $H_0 : \mu = 6$ vs. $H_1 : \mu \neq 6$. So set $\mu_0 = 6$.
2. By the CLT, $T = \dfrac{\overline{X} - \mu_0}{S_{n-1}/\sqrt{n}} \sim t_{n-1}$. That is, $\dfrac{\overline{X} - 6}{S_{n-1}/\sqrt{16}} \sim t_{15}$. So we reject H_0 at the $100\alpha\%$ level if $|t| > t_{15,1-\alpha/2}$.
3. Compute the realised value of the test statistic: $\overline{x} = 5.8$, $s_{n-1} = 1.2$ and $n = 16$ $\Rightarrow t = \dfrac{\overline{x} - \mu_0}{s_{n-1}/\sqrt{n}} = -0.657$.
4. We have $|t| = 0.657 \ll 2.13 = t_{15,.975}$, so **we have insufficient evidence to reject H_0 at the 5% level.**
5. In fact, the p-value for these data is 51.51%, so there is very little evidence to suggest the code is now any faster.

■

9.2 Testing for differences in population means

9.2.1 Two Sample Problems

Suppose, as before, we have a random sample $\underline{X} = (X_1, \ldots, X_{n_1})$ from an unknown population distribution P_X.

But now, suppose we have a further random sample $\underline{Y} = (Y_1, \ldots, Y_{n_2})$ from a second, different population P_Y.

Then we may wish to test hypotheses concerning the similarity of the two distributions P_X and P_Y.

In particular, we are often interested in testing whether P_X and P_Y have equal means. That is, to test

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

A special case is when the two samples $\underline{X}$ and $\underline{Y}$ are *paired*. That is, if $n_1 = n_2 = n$ and the data are collected as pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ so that, for each i, X_i and Y_i are possibly dependent.

For example, we might have a random sample of n individuals and X_i represents the heart rate of the i^{th} person before light exercise and Y_i the heart rate of the same person afterwards.

In this special case, for a test of equal means we can consider the sample of differences $Z_1 = X_1 - Y_1, \ldots, Z_n = X_n - Y_n$ and test $H_0 : \mu_Z = 0$ using the single sample methods we have seen.

In the above example, this would test whether light exercise has an impact on heart rate.

9.2.2 Normal Distributions with Known Variances

Suppose

- $\underline{X} = (X_1, \ldots, X_{n_1})$ are i.i.d. $\mathrm{N}(\mu_X, \sigma_X^2)$ with μ_X unknown;
- $\underline{Y} = (Y_1, \ldots, Y_{n_2})$ are i.i.d. $\mathrm{N}(\mu_Y, \sigma_Y^2)$ with μ_Y unknown;
- the two samples $\underline{X}$ and $\underline{Y}$ are independent.

Then we still have that, independently,

$$\overline{X} \sim \mathrm{N}\left(\mu_X, \frac{\sigma_X^2}{n_1}\right), \qquad \overline{Y} \sim \mathrm{N}\left(\mu_Y, \frac{\sigma_Y^2}{n_2}\right)$$

From this it follows that the difference in sample means,

$$\overline{X} - \overline{Y} \sim \mathrm{N}\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_1} + \frac{\sigma_Y^2}{n_2}\right),$$

and hence

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim \Phi.$$

So under the null hypothesis $H_0 : \mu_X = \mu_Y$, we have

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}} \sim \Phi.$$

If σ_X^2 and σ_Y^2 are known, we immediately have a test statistic

$$z = \frac{\overline{x} - \overline{y}}{\sqrt{\sigma_X^2/n_1 + \sigma_Y^2/n_2}}$$

which we can compare against the quantiles of a standard normal.

That is,

$$R = \left\{ z \;\middle|\; |z| > z_{1-\frac{\alpha}{2}} \right\},$$

gives a rejection region for a hypothesis test of $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \neq \mu_Y$ at the $100\alpha\%$ level.

9.2.3 Normal Distributions with Unknown Variances

On the other hand, suppose σ_X^2 and σ_Y^2 are unknown. Then if we know $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ but σ^2 is unknown, we can still proceed.

We have

$$\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sigma\sqrt{1/n_1+1/n_2}} \sim \Phi,$$

and so, under $H_0 : \mu_X = \mu_Y$,

$$\frac{\overline{X}-\overline{Y}}{\sigma\sqrt{1/n_1+1/n_2}} \sim \Phi.$$

but with σ unknown.

We need an estimator for the variance using samples from two populations with different means. Just combining the samples together into one big sample would over-estimate the variance, since some of the variability in the samples would be due to the difference in μ_X and μ_Y.

So we define the **bias-corrected pooled sample variance**

$$S^2_{n_1+n_2-2} = \frac{\sum_{i=1}^{n_1}(X_i-\overline{X})^2 + \sum_{i=1}^{n_2}(Y_i-\overline{Y})^2}{n_1+n_2-2},$$

which is an unbiased estimator for σ^2.

We can immediately see that $s^2_{n_1+n_2-2}$ is indeed an unbiased estimate of σ^2 by noting

$$S^2_{n_1+n_2-2} = \frac{n_1-1}{n_1+n_2-2}S^2_{n_1-1} + \frac{n_2-1}{n_1+n_2-2}S^2_{n_2-1};$$

That is, $s^2_{n_1+n_2-2}$ is a weighted average of the bias-corrected sample variances for the individual samples $\underline{x}$ and $\underline{y}$, which are both unbiased estimates for σ^2.

Then substituting $S_{n_1+n_2-2}$ in for σ we get

$$\frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{S_{n_1+n_2-2}\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2},$$

and so, under $H_0 : \mu_X = \mu_Y$,

$$T = \frac{\overline{X}-\overline{Y}}{S_{n_1+n_2-2}\sqrt{1/n_1+1/n_2}} \sim t_{n_1+n_2-2}.$$

So we have a rejection region for a hypothesis test of $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \neq \mu_Y$ at the $100\alpha\%$ level given by

$$R = \left\{ t \;\middle|\; |t| > t_{n_1+n_2-2,1-\frac{\alpha}{2}} \right\},$$

for the statistic

$$t = \frac{\overline{x}-\overline{y}}{s_{n_1+n_2-2}\sqrt{1/n_1+1/n_2}}.$$

Example The same piece of C code was repeatedly run after compilation under two different C compilers, and the run times under each compiler were recorded. The sample mean and bias-corrected sample standard deviation for Compiler 1 were 114 s and 310 s respectively, and the corresponding figures for Compiler 2 were 94 s and 290 s. Both sets of data were each based on 15 runs.

Suppose that Compiler 2 is a refined version of Compiler 1, and so if μ_1, μ_2 are the expected run times of the code under the two compilations, we might fairly assume $\mu_2 \leq \mu_1$.

Conduct a hypothesis test of $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 > \mu_2$ at the 5% level.

Until now we have exclusively considered *two-sided* tests. That is tests of the form $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$.

Here we need to consider *one-sided* tests, which differ by the alternative hypothesis being of the form $H_1 : \theta < \theta_0$ or $H_1 : \theta > \theta_0$.

This presents no extra methodological challenge and requires only a slight adjustment in the construction of the rejection region.

We still use the t-statistic

$$t = \frac{\overline{x} - \overline{y}}{s_{n_1+n_2-2}\sqrt{1/n_1 + 1/n_2}},$$

where $\overline{x}, \overline{y}$ are the sample mean run times under Compilers 1 and 2 respectively. But now the one-sided rejection region becomes

$$R = \left\{ t \mid t > t_{n_1+n_2-2, 1-\alpha} \right\}.$$

First calculating the bias-corrected pooled sample variance, we get

$$s^2_{n_1+n_2-2} = \frac{14 \times 310 + 14 \times 290}{28} = 300.$$

(Note that since the sample sizes n_1 and n_2 are equal, the pooled estimate of the variance is the average of the individual estimates.)

$$\text{So } t = \frac{\overline{x} - \overline{y}}{s_{n_1+n_2-2}\sqrt{1/n_1 + 1/n_2}} = \frac{114 - 94}{\sqrt{300}\sqrt{1/15 + 1/15}} = 3.162.$$

For a 1-sided test we compare $t = 3.162$ with $t_{28,0.95} = 1.701$ and conclude that we reject the null hypothesis at the 5% level; the second compilation is significantly faster. ■

9.3 Goodness of Fit

9.3.1 Count Data and Chi-Square Tests

The results in the previous sections relied upon the data being either normally distributed, or at least through the CLT having the sample mean being approximately normally distributed. Tests were then developed for making inference on population means under those assumptions. These tests were very much *model-based*.

Another important but very different problem concerns *model checking*, which can be addressed through a more general consideration of count data for simple (discrete and finite) distributions.

The following ideas can then be trivially extended to infinite range discrete and continuous random variables by *binning* observed samples into a finite collection of predefined intervals.

Let X be a simple random variable taking values in the range $\{x_1, \ldots, x_k\}$, with probability mass function $p_j = \mathrm{P}(X = x_j \,|\, \theta)$, $j = 1, \ldots, k$ depending on an unknown parameter vector θ of length m.

Then a random sample of size n from the distribution of X can be summarised by the *observed* frequency counts $\underline{O} = (O_1, \ldots, O_k)$ at the points $x_1, \ldots, x_k$ (so $\sum_{j=1}^{k} O_j = n$).

Suppose we have a null hypothesis $H_0 : \theta = \theta_0$ for the value of the unknown parameter(s). Then under H_0 we know the probabilities $\{p_1, \ldots, p_k\}$, and so we are able to calculate the expected frequency counts $\underline{E} = (E_1, \ldots, E_k)$ by $E_j = np_j$. (Note that we have $\sum\limits_{j=1}^{k} E_j = n$.)

We then seek to compare the observed frequencies with the expected frequencies to test for **goodness of fit**.

To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ we use the **chi-square statistic**

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

If H_0 were true, then the statistic χ^2 would approximately follow a **chi-square distribution** with $\nu = k - m - 1$ degrees of freedom.

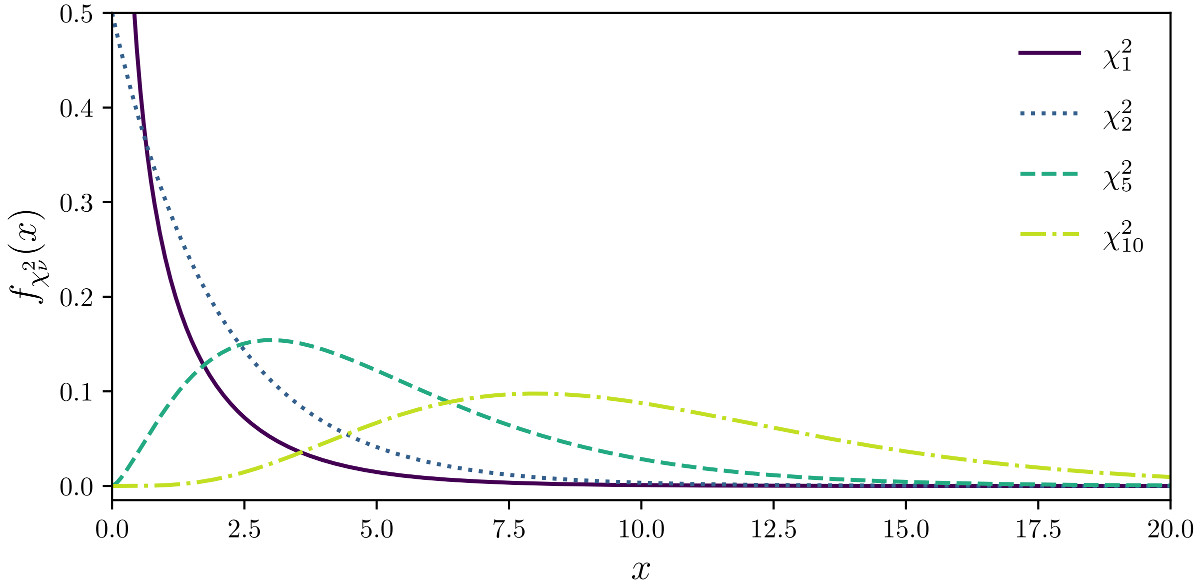
- k is the number of values (categories) the simple random variable X can take.
- m is the number of parameters we needed to estimate from the data ($\dim(\theta)$) in order to calculate the p_j's.
- For the approximation to be valid, we should have $\forall j$, $E_j \geq 5$. This may require some merging of categories.

Clearly larger values of χ^2 correspond to larger deviations from the null hypothesis model. That is, if $\chi^2 = 0$ the observed counts exactly match those expected under H_0.

For this reason, we always perform a one-sided goodness of fit test using the χ^2 statistic, looking only at the upper tail of the distribution. I.e. we only want to reject the null hypothesis when χ^2 is large, not when it is small.

Hence the rejection region for a goodness of fit hypothesis test at the at the $100\alpha\%$ level is given by

$$R = \left\{x^2 \mid x^2 > \chi^2_{k-m-1,1-\alpha}\right\}.$$



pdf of χ^2_ν with degrees of freedom $\nu = 1, 2, 5, 10$.

9.3.2 Proportions

Example Each year, around 1.3 million people in the USA suffer adverse drug effects (ADEs). A study in the *Journal of the American Medical Association* (July 5, 1995) gave the causes of 95 ADEs below.

Cause	Number of ADEs	
Lack of knowledge of drug	29	$= O_1$
Rule violation	17	$= O_2$
Faulty dose checking	13	$\vdots$
Slips	9	
Other	27	

$k = 5$

Test whether the true percentages of ADEs differ across the 5 causes.

Under the null hypothesis that the 5 causes are equally likely, we would have expected counts of $\frac{95}{5} = 19 = E_i$ for each cause.

So our χ^2 statistic becomes

$$\begin{aligned} x^2 &= \frac{(29-19)^2}{19} + \frac{(17-19)^2}{19} + \frac{(13-19)^2}{19} + \frac{(9-19)^2}{19} + \frac{(27-19)^2}{19} \\ &= \frac{100}{19} + \frac{4}{19} + \frac{36}{19} + \frac{100}{19} + \frac{64}{19} = \frac{304}{19} = 16. \end{aligned}$$

We have not estimated any parameters from the data, so we compare x^2 with the quantiles of the $\chi^2_{5-1} = \chi^2_4$ distribution.

Well $16 > 9.49 = \chi^2_{4,0.95}$, so we reject the null hypothesis at the 5% level; we have reason to suppose that there is a difference in the true percentages across the different causes. ■

9.3.3 Model Checking

Recall the example from Chapter 5 (Discrete Probability Distributions) where the number of particles emitted by a radioactive substance which reached a Geiger counter was measured for 2608 time intervals, each of length 7.5 seconds.

We fitted a Poisson(λ) distribution to the data by plugging in the sample mean number of counts (3.870) for the rate parameter λ. (Which we now know to be the MLE.)

$k=11$

x	0	1	2	3	4	5	6	7	8	9	≥10
$O(n_x)$	57	203	383	525	532	408	273	139	45	27	16
$E(n_x)$	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1

(O=Observed, E=Expected).

Whilst the fitted Poisson(3.87) expected frequencies looked quite convincing to the eye, at that time we had no formal method of quantitatively assessing the fit. However, we now know how to proceed.

“residual”

x	0	1	2	3	4	5	6	7	8	9	≥10
O	57	203	383	525	532	408	273	139	45	27	16
E	54.4	210.5	407.4	525.5	508.4	393.5	253.8	140.3	67.9	29.2	17.1
$O-E$	2.6	-7.5	-24.4	-0.5	23.6	14.5	19.2	-1.3	22.9	2.2	1.1
$\frac{(O-E)^2}{E}$	0.124	0.267	1.461	0.000	1.096	0.534	1.452	0.012	7.723	0.166	0.071

The statistic $x^2 = \sum \frac{(O-E)^2}{E} = 12.906$ should be compared with a $\chi^2_{11-1-1} = \chi^2_9$ distribution (we had to estimate one parameter, λ, using the observed data).

Well $\chi^2_{9,0.95} = 16.91$, so at the 5% level we do not reject the null hypothesis of a Poisson(3.87) model for the data.

9.3.4 Independence

Suppose we have two discrete random variables X and Y that can each take finite values which are jointly distributed with unknown probability mass function p_{XY}.

We are often interested in trying to ascertain whether X and Y are independent. That is, determine whether $p_{XY}(x, y) = p_X(x)p_Y(y)$.

Let the ranges of the random variables X and Y be $\{x_1, \ldots, x_k\}$ and $\{y_1, \ldots, y_\ell\}$ respectively. Then an i.i.d. sample of size n from the joint distribution of (X, Y) can be represented by a list of counts n_{ij} ($1 \leq i \leq k$; $1 \leq j \leq \ell$) of the number of times we observe the pair (x_i, y_j).

Tabulating these data in the following way gives what is known as a $k \times \ell$ **contingency table**.

	y_1	y_2	$\cdots$	y_ℓ	
x_1	n_{11}	n_{12}		$n_{1\ell}$	$n_{1\bullet}$
x_2	n_{21}	n_{22}		$n_{2\ell}$	$n_{2\bullet}$
$\vdots$					
x_k	n_{k1}	n_{k2}		$n_{k\ell}$	$n_{k\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet \ell}$	n

Note the row sums $(n_{1\bullet}, n_{2\bullet}, \ldots, n_{k\bullet})$ represent the frequencies of $x_1, x_2, \ldots, x_k$ in the sample (that is, ignoring the value of Y). Similarly for the column sums $(n_{\bullet 1}, n_{\bullet 2}, \ldots, n_{\bullet \ell})$.

Under the null hypothesis

$$H_0 : X \text{ and } Y \text{ are independent,}$$

the expected values of the entries of the contingency table, conditional on the row and column sums, are given by

$$\widehat{n}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}, \quad 1 \leq i \leq k, \;\; 1 \leq j \leq \ell.$$

To see this, consider the marginal distribution of X; we could approximate $p_X(x_i)$ by $\widehat{p}_{i\bullet} = \dfrac{n_{i\bullet}}{n}$. Similarly for $p_Y(y_j)$ we get $\widehat{p}_{\bullet j} = \dfrac{n_{\bullet j}}{n}$.

Then under independence $p_{XY}(x_i, y_j) = p_X(x_i)\, p_Y(y_j)$, and so we can estimate $p_{XY}(x_i, y_j)$ by

$$\widehat{p}_{ij} = \widehat{p}_{i\bullet} \times \widehat{p}_{\bullet j} = \frac{n_{i\bullet} \times n_{\bullet j}}{n^2}.$$

Now that we have a set of expected frequencies to compare against our $k \times \ell$ observed frequencies, a χ^2 test can be performed.

How many parameters m did we need to compute the model prediction? We are using both the row and column sums to estimate our probabilities, and there are k and ℓ of these respectively. But the last row sum is fixed by the previous $k-1$ row sums and the requirement that they all add up to n (and similarly for the column sums). So we compare our calculated x^2 statistic against a χ^2 distribution with $k\ell - (k-1) - (\ell-1) - 1 = (k-1)(\ell-1)$ degrees of freedom.

Hence the rejection region for a hypothesis test of independence in a $k \times \ell$ contingency table at the at the $100\alpha\%$ level is given by

$$R = \left\{ x^2 \;\middle|\; x^2 > \chi^2_{(k-1)(\ell-1),1-\alpha} \right\}.$$

Example An article in *International Journal of Sports Psychology* (July-Sept 1990) evaluated the relationship between physical fitness and stress. 549 people were classified as good, average, or poor fitness, and were also tested for signs of stress (yes or no). The data are shown in the table below.

	Poor Fitness	Average Fitness	Good Fitness	
Stress	206	184	85	475
No stress	36	28	10	74
	242	212	95	549

Question Is there any relationship between stress and fitness?

Under independence we would estimate the expected values to be

	Poor Fitness	Average Fitness	Good Fitness	
Stress	209.4	183.4	82.2	475
No stress	32.6	28.6	12.8	74
	242	212	95	549

Hence the χ^2 statistic is calculated to be

$$x^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(206 - 209.4)^2}{209.4} + \ldots + \frac{(10 - 12.88)^2}{12.8} = 1.1323.$$

This should be compared with a χ^2 distribution with $(2-1) \times (3-1) = 2$ degrees of freedom. We then have $\chi^2_{2,0.95} = 5.99$, so we have insufficient evidence to reject to null hypothesis, i.e. there is no significant evidence to suggest there is any relationship between fitness and stress. ■

9.3.5 The χ^2 distribution and degrees of freedom

Take k i.i.d. standard normal RVs ($Z_i \sim \mathrm{N}(0,1)$ for $i = 1, \ldots, k$) and form the new RV

$$X = Z_1^2 + \ldots + Z_k^2.$$

Then X is a χ^2 RV with k degrees of freedom and we write $X \sim \chi_k^2$ or $X \sim \chi^2(k)$. (Sometimes we use χ^2 as the random variable itself, which can be confusing.)

The mean of a $\chi^2(k)$ RV is k and the variance is $2k$.

If we have a bunch of independent normal random variables with arbitrary means and variances we can "standardize" them and form a χ^2 statistic. I.e. let $Y_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$ for $i = 1 \ldots, k$ with the Y_i's all independent. Then the RV χ^2 defined by

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2$$

is distributed as a χ^2 with k degrees of freedom.

Compare this to the test statistic we introduced above: $\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$.

The observed data O_i is a number of counts and we might think of them as Poisson RVs with means E_i.

The need to reduce the degrees of freedom arises because we use the data themselves to set the values of the E_i's. In other words, the E_i's are functions of the O_i's, i.e. they are RVs themselves. Then the individual terms in the χ^2 sum are no longer independent standard normal RVs.

It turns out that each equation we use to constrain the E_i's using the O_i's effectively reduces the number of degrees of freedom by 1 (assuming some technical requirements we won't go into).

For k bins and m parameters the test statistic is a $\chi^2(k - m - 1)$ RV. Why the "extra" -1? Because when we write $E_i = np_i$ we introduce the additional parameter n, which we then set using the constraint that $\sum_i E_i = \sum_i O_i$. Even if we knew the values of p_i *a priori* (without using the O_i's) we still set each E_i using $E_i = (\sum_i O_i)\, p_i$ and this reduces the number of degrees of freedom by 1.