Statistics 2 Lectures

1 Limit Theorems

1.1 Law of Large Numbers

Theorem (Law of Large Numbers). *Let $X_1, X_2, \ldots$ be a sequence of independent random variables with $\mathbb{E}(X_i) = \mu$ and $\mathrm{var}(X_i) = \sigma^2$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

Proof. We have

$$\mathbb{E}(\bar{X}_n) = \mu, \qquad \mathrm{var}(\bar{X}_n) = \sigma^2/n.$$

By Chebyshev's inequality,

$$\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\mathrm{var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$.

Discuss convergence concepts: the above says that $\bar{X}_n \to \mu$ *in probability* (hence *weak* law of large numbers), as opposed to saying that with probability one, $\lim_n \bar{X}_n = \mu$ (strong law of large numbers). Discuss corresponding notations.

Application: Monte Carlo Integration. Let (X_i) be i.i.d. with density f and g such that $g(X_i)$ has finite mean and variance. Then

$$\hat{\theta}_n = \frac{1}{n}\sum_{i=1}^n g(X_i) \to \mathbb{E}g(X) = \int g(x)f(x)dx.$$

1.2 Central Limit Theorem

Definition. *Let $X_1, X_2, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots$, and let X be a random variable with cumulative distribution function F. We say that X_n converges* in distribution *to X if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at every continuity point x of F.

The following is a variant of Theorem A in Rice, using characteristic functions instead of moment generating functions (see `http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)`.

Theorem (Lévy's continuity theorem). *A sequence (X_n) of random variables converges in distribution to a random variable X if and only if the sequence (ϕ_{X_n}) of the characteristic functions converges pointwise to a function ϕ which is continuous at the origin. Then ϕ is the characteristic function of X.*

Example (Normal Approximation of the Poisson Distribution). If X has a Poisson distribution with parameter λ,

$$\phi_X(t) = \mathbb{E}(e^{itX}) = \sum_{k=0}^{\infty} e^{itk} \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} e^{-\lambda} = e^{\lambda(e^{it}-1)}.$$

Now let (X_n) be a sequence of random variables with a Poisson distribution with parameters λ_n, and consider the standardized random variables $Z_n = (X_n - \lambda_n)/\sqrt{\lambda_n}$. Then

$$\phi_{Z_n}(t) = \exp(-it\sqrt{\lambda_n} + \lambda_n(e^{it/\sqrt{\lambda_n}} - 1)),$$

remembering that the characteristic function of $(X-\mu)/\sigma$ is given by

$$\mathbb{E}(e^{it(X-\mu)/\sigma}) = e^{-it\mu/\sigma}\phi_X(t/\sigma).$$

Now

$$e^{it/\sqrt{\lambda_n}} = 1 + it/\sqrt{\lambda_n} - t^2/(2\lambda_n) + o(1/\lambda_n)$$

and thus

$$\begin{aligned} &\log(\phi_{Z_n}(t)) \\ &= -it\sqrt{\lambda_n} + \lambda_n(1 + it/\sqrt{\lambda_n} - t^2/(2\lambda_n) + o(1/\lambda_n) - 1) \\ &= -it\sqrt{\lambda_n} + it\sqrt{\lambda_n} - t^2/2 + o(1) \\ &\to -t^2/2 \end{aligned}$$

i.e.,

$$\phi_{Z_n}(t) \to e^{-t^2/2}$$

which is the characteristic function of the standard normal distribution. Thus, $Z_n = (X_n - \lambda_n)/\sqrt{\lambda_n} \to N(0,1)$ in distribution.

Theorem (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables having mean μ, variance σ^2 and finite third moments. Let $S_n = \sum_{i=1}^n X_i$. Then*

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x), \qquad -\infty < x < \infty.$$

I.e., the standardized S_n converge to the $N(0,1)$ in distribution.

Proof. Without loss of generality, assume that $\mu = 0$. Let $Z_n = S_n/(\sigma\sqrt{n})$. Then as the X_i are i.i.d.,

$$\phi_{Z_n}(t) = (\phi_{X/(\sigma\sqrt{n})}(t))^n = \phi_X(t/(\sigma\sqrt{n}))^n.$$

For s around 0, using Taylor expansion (and finite third moments) gives

$$\begin{aligned}
&\phi_X(s) \\
&= \phi_X(0) + s\phi_X'(0) + \frac{s^2}{2}\phi_X''(0) + o(s^2) \\
&= 1 - s \cdot 0 + \frac{s^2}{2}(-\sigma^2) + o(s^2) \\
&= 1 - \sigma^2 s^2/2 + o(s^2)
\end{aligned}$$

and thus

$$\phi_{Z_n}(t) = \left(\phi_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-t^2/2}.$$

2 Estimation of Parameters and Fitting of Probability Distributions

Note: Sample mean and sample variance. The statistics

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

are called the sample mean and sample variance, respectively.

Note: Gamma distribution. For the gamma (and hence in particular the exponential) distribution, there are two alternative parametrizations: In R (see also `http://en.wikipedia.org/wiki/Gamma_distribution`), the *shape* parameter α and the *scale* parameter s are used, with corresponding density:

$$f(t) = \frac{t^{\alpha-1}e^{-t/s}}{s^\alpha \Gamma(\alpha)}, \qquad t > 0.$$

In Rice, the *rate* parameter $\lambda = 1/s$ is used instead of the scale parameter s:

$$f(t) = \frac{\lambda^\alpha t^{\alpha-1}e^{-\lambda t}}{\Gamma(\alpha)}, \qquad t > 0.$$

(Of course one can simply use $s \leftrightarrow 1/\lambda$ to move between the parametrizations.)

2.1 Parameter Estimation

Basic approach:

- Observed data regarded as realizations of random variables $X_1, \ldots, X_n$, whose joint unknown distribution depends on an unknown (possibly vector-valued) parameter θ.
- Usually the X_i will be modeled as i.i.d., in which case their joint density is $f(x_1|\theta)\cdots f(x_n|\theta)$.
- An estimate of θ will be a function of $X_1, \ldots, X_n$ and hence a random variable with a probability distribution called its *sampling distribution.*
- The variability of this distribution will most frequently be assessed through its standard deviation, commonly called the *standard error* (of the estimate).

2.2 The Method of Moments

The kth moment of a probability law is defined as

$$\mu_k = \mathbb{E}(X^k)$$

(where X is a random variable following that probability law and the corresponding expectation exists). If $X_1, \ldots, X_n$ are i.i.d. random variables from this law, the kth sample moment is defined as

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Under suitable moment assumptions, the sample moments converge to the population ones.

The idea of estimating parameters by the method of moments is to express the parameters in terms of the (lowest possible order) moments, and then substitute the sample moments into the expressions. Typically, this involves the following steps:

1. Find expressions of suitable low order moments in terms of the parameters.
2. Invert the expressions, obtaining expressions for the parameters in terms of the low order moments.
3. Insert the sample moments into these expressions, obtaining estimates of the parameters in terms of the sample moments.

Example: Poisson distribution. This has $\mu_1 = \mathbb{E}(X) = \lambda$. Thus, the method of moments estimator of λ is $\hat{\mu}_1 = \bar{X}$. Letting $S_n = \sum_{i=1}^{n} X_i$, we have $\hat{\lambda} = S_n/n$, where S_n has a Poisson distribution with parameter $n\lambda$. Thus, if λ_0 is the "true" parameter, the sampling distribution is

$$\mathbb{P}(\hat{\lambda} = v) = \mathbb{P}(S_n = nv) = \frac{(n\lambda_0)^{nv}}{(nv)!}e^{-n\lambda_0},$$

such that

$$\mathbb{E}(\hat{\lambda}) = \mathbb{E}(S_n/n) = (n\lambda_0)/n = \lambda_0$$

and

$$\mathrm{var}(\hat{\lambda}) = \mathrm{var}(S_n/n) = (n\lambda_0)/n^2 = \lambda_0/n.$$

From the results on the normal approximation of the Poisson distribution, we infer that $\hat{\lambda}$ is asymptotically normal (with the above mean and variance).

As $\mathbb{E}(\hat{\lambda}) = \lambda_0$, we say that the estimate is *unbiased*: the sampling distribution is centered at λ_0.

The standard deviation of the sampling distribution of $\hat{\lambda}$ is called the *standard error* of $\hat{\lambda}$ and is given by

$$\sigma_{\hat{\lambda}} = \sqrt{\lambda_0/n}.$$

Of course, we cannot know the sampling distribution or its standard error because λ_0 is unknown. An approximation can be obtained by substituting $\hat{\lambda}$ for λ_0, giving in particular the *estimated standard error* (plug-in estimate)

$$s_{\hat{\lambda}} = \sqrt{\hat{\lambda}/n}.$$

Example: Normal distribution. This has the first two moments

$$\mu_1 = \mathbb{E}(X) = \mu, \qquad \mu_2 = \mathbb{E}(X^2) = \mu^2 + \sigma^2.$$

Therefore,

$$\mu = \mu_1, \qquad \sigma^2 = \mu_2 - \mu_1^2.$$

The corresponding estimates from the sample moments are

$$\hat{\mu} = \bar{X}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

The sampling distributions are given by a classic result (see Section 6.3 in Rice):

$$\hat{X} \sim N(\mu, \sigma^2/n), \qquad n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$$

and $\hat{X}$ and $\hat{\sigma}^2$ are independent (more on this later).

Example: Gamma distribution. This has (using the shape and rate parameters)

$$\mu_1 = \frac{\alpha}{\lambda}, \qquad \mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

From the second equation,

$$\mu_2 = \mu_1(\mu_1 + 1/\lambda)$$

giving

$$\lambda = \frac{\mu_1}{\mu_2 - \mu_1^2}$$

and

$$\alpha = \lambda\mu_1 = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

Since $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$ (note that this is *not* the sample variance), we obtain

$$\hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma}^2}, \qquad \hat{\lambda} = \frac{\bar{X}}{\hat{\sigma}^2}.$$

What can we say about the sampling distribution of $\hat{\alpha}$ and $\hat{\lambda}$? Unlike the previous two examples, the exact sampling distribution is not known—hence, we use simulation. If the true parameters α_0 and λ_0 were known, we could easily simulate the sampling distributions of the estimates. As they are not, we substitute our estimates $\hat{\alpha}$ and $\hat{\lambda}$ for the true values. I.e., we use the following *bootstrap* procedure.

1. Draw B samples of size n from the gamma distribution with shape and rate parameters $\hat{\alpha}$ and $\hat{\lambda}$.

2. In each bootstrap sample, estimate the parameters (using the method of moments), giving estimates α_b^* and λ_b^*.

3. Estimate standard errors as

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\alpha_b^* - \bar{\alpha})^2}$$

where $\bar{\alpha} = B^{-1}\sum_{b=1}^{B}\alpha_b^*$, and similarly for $s_{\hat{\lambda}}$.

Definition (Consistency). *Let $\hat{\theta}_n$ be an estimate of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is said to be consistent in probability if $\hat{\theta}_n$ converges to θ in probability as n approaches infinity; i.e., for all $\epsilon > 0$,*

$$\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0 \qquad \text{as } n \to \infty.$$

Note that in the above $\mathbb{P}$ is taken as the probability law corresponding to θ, or equivalently, θ is taken as the true parameter.

In the above cases, suitable laws of large number imply consistency of the methods of moment estimators, also justifying the use of plug-in estimates of the standard errors: if the true standard error is of the form

$$\sigma_{\hat{\theta}} = \sigma(\theta_0)/\sqrt{n}$$

then for the plug-in estimator

$$s_{\hat{\theta}} = \sigma(\hat{\theta})/\sqrt{n}$$

we have $\sigma_{\hat{\theta}}/s_{\hat{\theta}} \to 1$ provided that $\sigma(\cdot)$ is continuous at θ_0.

2.3 The Method of Maximum Likelihood

2.3.1 Maximum Likelihood Estimation

Suppose random variables $X_1, \ldots, X_n$ have a joint density (or probability mass function, called frequency function in Rice) $f(x_1, \ldots, x_n|\theta)$. Given observed values, the *likelihood* of θ as a function of (the data) $x_1, \ldots, x_n$ is defined as

$$\text{lik}(\theta) = f(x_1, \ldots, x_n|\theta)$$

The *maximum likelihood estimate (mle)* of θ is the (if unique) value of θ that maximizes the likelihood, thus making the observed data "most likely".

If the X_i are i.i.d., the joint density is the product of the marginal densities:

$$\text{lik}(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Typically, it is more convenient to work with the (natural) logarithm of the likelihood, the so-called *log likelihood*. In the i.i.d. case, this is given by

$$\ell(\theta) = \sum_{i=1}^{n} \log(f(X_i|\theta)).$$

Example: Poisson distribution. If X has a Poisson distribution with parameter λ,

$$\mathbb{P}(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

Hence, if $X_1, \ldots, X_n$ are i.i.d. Poisson with parameter λ, their log likelihood is

$$\ell(\lambda) = \sum_{i=1}^{n} (X_i \log(\lambda) - \lambda - \log(X_i!)) = \log(\lambda) \sum_{i=1}^{n} X_i - n\lambda - \sum_{i=1}^{n} \log(X_i!).$$

The mle can be obtained by setting the derivative to zero, giving

$$\ell'(\lambda) = \frac{1}{\lambda} \sum_{i=1}^{n} X_i - n = 0$$

and the mle is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}.$$

The mle agrees with the method of moments estimator and thus has the same sampling distribution.

Example: Normal distribution. If $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$,

$$f(x_1, \ldots, x_n|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-(x_i - \mu)^2/(2\sigma^2))$$

The log likelihood is thus

$$\ell(\mu, \sigma^2) = -n \log(\sigma) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.$$

Taking partials with respect to μ and σ (or σ^2) and solving for the mle gives

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

(Note that Rice is not quite consistent about σ versus σ^2 as the parameter.)

Again, estimates and their sampling distributions agree with those obtained by the method of moments.

Example: Gamma distribution. The density function of the gamma distribution with shape parameter α and rate parameter λ is

$$f(x|\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}, \qquad x > 0$$

thus the log likelihood is

$$\begin{aligned}\ell(\alpha,\lambda) &= \sum_{i=1}^n (\alpha\log(\lambda) + (\alpha-1)\log(X_i) - \lambda X_i - \log(\Gamma(\alpha)) \\ &= n\alpha\log(\lambda) + (\alpha-1)\sum_{i=1}^n \log(X_i) - \lambda\sum_{i=1}^n X_i - n\log(\Gamma(\alpha))\end{aligned}$$

with partials

$$\begin{aligned}\frac{\partial\ell}{\partial\alpha} &= n\log(\lambda) + \sum_{i=1}^n \log(X_i) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ \frac{\partial\ell}{\partial\lambda} &= \frac{n\alpha}{\lambda} - \sum_{i=1}^n X_i\end{aligned}$$

Setting the second partial to zero we get

$$\hat\lambda = \frac{n\hat\alpha}{\sum_{i=1}^n X_i} = \frac{\hat\alpha}{\bar X}$$

but substituting this into the equation for the first partial yields a non-linear equation for the mle of α:

$$n\log(\hat\alpha) - n\log(\bar X) + \sum_{i=1}^n \log(X_i) - n\frac{\Gamma'(\hat\alpha)}{\Gamma(\hat\alpha)} = 0$$

In this case, the mle cannot be given in closed form, so obtaining the exact sampling distribution appears intractible. One again resorts to bootstrap approximations (which would indicate that the sampling distributions of the mles are substantially less dispersed than those of the methods of moments estimates).

2.3.2 Large Sample Theory

Theorem (Consistency of the MLE). *Under appropriate smoothness conditions on f, the mle from an i.i.d. sample is consistent.*

Proof/sketch. Consider maximizing

$$\ell(\theta)/n = \frac{1}{n}\sum_{i=1}^n \log(f(X_i|\theta)).$$

Write θ_0 for the true parameter. As $n\to\infty$, the law of large numbers implies that

$$\ell(\theta)/n \to \mathbb{E}_{\theta_0}\log(f(X_i|\theta)) = \int \log(f(x|\theta))f(x|\theta_0)dx.$$

Assume that one can show that for large n, the θ which maximizes $\ell(\theta)$ is close to the one which maximizes the limit (which is far from being straightforward). Then it remains to be shown that the limit is maximized at θ_0. This can be done as in Rice, or more straightforwardly as follows. The function $t\log(t) - t + 1$ for $t \geq 0$ has a global minimum at $t = 1$, hence $t\log(t) - t + 1 \geq 0$ and thus $u\log(u/v) - u + v = v((u/v)\log(u/v) - (u/v) + 1) \geq 0$ if $u, v \geq 0$. Thus,

$$\begin{aligned} 0 &\leq \int (f(x|\theta_0)\log(f(x|\theta_0)/f(x|\theta)) - f(x|\theta_0) + f(x|\theta))\,dx \\ &= \int \log(f(x|\theta_0))f(x|\theta_0)dx - \int \log(f(x|\theta))f(x|\theta_0)dx \end{aligned}$$

(as densities integrate to one), and hence

$$\int \log(f(x|\theta))f(x|\theta_0)dx \leq \int \log(f(x|\theta_0))f(x|\theta_0)dx$$

(one can also refine this by showing strict inequality unless the densities agree).

Fisher information. The Fisher information is given by

$$I(\theta) = \mathbb{E}\left(\left(\frac{\partial \log(f(X|\theta))}{\partial\theta}\right)^2\right).$$

Under appropriate smoothness conditions on f, it may also be expressed as

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2 \log(f(X|\theta))}{\partial\theta^2}\right).$$

To demonstrate the identity, start by noting that $\int f(x|\theta)dx = 1$ and thus

$$\frac{\partial}{\partial\theta}\int f(x|\theta)dx = 0.$$

Assuming one can interchange differentiation and integration, this gives

$$0 = \frac{\partial}{\partial\theta}\int f(x|\theta)dx = \int \frac{\partial f(x|\theta)}{\partial\theta}dx = \int \frac{\partial \log(f(x|\theta))}{\partial\theta} f(x|\theta)dx.$$

Taking second derivatives,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int \frac{\partial \log(f(x|\theta))}{\partial\theta} f(x|\theta)dx \\ &= \int \frac{\partial^2 \log(f(x|\theta))}{\partial\theta^2} f(x|\theta)dx + \int \left(\frac{\partial \log(f(x|\theta))}{\partial\theta}\right)^2 f(x|\theta)dx \end{aligned}$$

as asserted.

Theorem (Asymptotic Normality of the MLE). *Under smoothness conditions on f, the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to a standard normal distribution.*

Proof/Sketch. From a Taylor series expansion,

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta_0) + (\hat{\theta} - \theta_0)\ell''(\theta_0)$$

so that

$$\hat{\theta} - \theta_0 \approx \frac{-\ell'(\theta_0)}{\ell''(\theta_0)}$$

and thus

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\frac{\ell'(\theta_0)/\sqrt{n}}{\ell''(\theta_0)/n}.$$

We show that the numerator satisfies a CLT and the denominator satisfies an LLN. The numerator has

$$\mathbb{E}(\ell'(\theta_0)/\sqrt{n}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial \log(f(X_i|\theta))}{\partial \theta}\right) = 0$$

and variance

$$\text{var}(\ell'(\theta_0)/\sqrt{n}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial \log(f(X_i|\theta))}{\partial \theta}\right)^2 = I(\theta_0).$$

By the CLT, the numerator is asymptotically normal with mean 0 and variance $I(\theta_0)$. The denominator is

$$\ell''(\theta_0)/n = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log(f(X_i|\theta))}{\partial \theta^2} \to \mathbb{E}\left(\frac{\partial^2 \log(f(X_i|\theta))}{\partial \theta^2}\right) = -I(\theta_0)$$

So overall,

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \approx \frac{\ell'(\theta_0)/(\sqrt{nI(\theta_0)})}{-\ell''(\theta_0)/(nI(\theta_0))}$$

asymptotically has a standard normal distribution.

Note that corresponding results can be proved for the multi-dimensional case. The covariance of the estimates $\hat{\theta}_i$ and $\hat{\theta}_j$ of the components i and j of θ, respectively, is then given by the ij entry of the inverse of $nI(\theta_0)$, where the ij component of the Fisher information matrix is now given by

$$\mathbb{E}\left(\frac{\partial \log(f(X|\theta))}{\partial \theta_i} \frac{\partial \log(f(X|\theta))}{\partial \theta_j}\right) = -\mathbb{E}\left(\frac{\partial^2 \log(f(X|\theta))}{\partial \theta_i \partial \theta_j}\right)$$

2.3.3 Confidence Intervals

A *confidence interval* for a population parameter θ is a random interval which contains θ with some specified (coverage) probability. A $100(1-\alpha)$ percent confidence interval contains θ with probability (at least) $1-\alpha$; if we took many random samples and formed confidence intervals from each one, $100(1-\alpha)$ percent of these would contain θ. Confidence intervals are frequently used in conjunction with point estimates to convey information about the uncertainty of the estimates.

Example: Normal distribution. The mles of μ and σ^2 from an i.i.d. normal sample are

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

A confidence interval for μ is based on the fact that

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

where S^2 is the sample variance and t_{n-1} the (Student) t distribution with $n-1$ degrees of freedom. Write $Q_F(\alpha)$ for the α quantile of distribution F. Then

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

and rearranging and using the symmetry of the t distribution gives

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

To obtain a confidence interval for σ^2, note that

$$n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$$

where χ^2_{n-1} denotes the chi-squared distribution with $n-1$ degrees of freedom. Thus,

$$\mathbb{P}\left(Q_{\chi^2_{n-1}}(\alpha/2) \leq \frac{n\hat{\sigma}^2}{\sigma^2} \leq Q_{\chi^2_{n-1}}(1-\alpha/2)\right) = 1 - \alpha,$$

and rearranging gives

$$\mathbb{P}\left(\frac{n\hat{\sigma}^2}{Q_{\chi^2_{n-1}}(1-\alpha/2)} \leq \sigma^2 \leq \frac{n\hat{\sigma}^2}{Q_{\chi^2_{n-1}}(\alpha/2)}\right) = 1 - \alpha.$$

Approximate confidence intervals. Where exact intervals cannot be obtained, we can use the fact that in general, $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ approximately has a standard normal distribution. Again using a plug-in estimate of the Fisher information, we get

$$\mathbb{P}\left(z_{\alpha/2} \leq \sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta_0) \leq z_{1-\alpha/2}\right) = 1 - \alpha,$$

giving an approximate $100(1-\alpha)$ percent confidence interval of

$$\hat{\theta} \pm z_{\alpha/2}/\sqrt{nI(\hat{\theta})}.$$

Example: Poisson distribution. The mle of the parameter λ from a sample from a Poisson distribution is $\hat{\lambda} = \bar{X}$. The sampling distribution is known, but depends on the unknown parameter. Approximate confidence intervals can be obtained from the above. We have

$$\log(f(x|\lambda)) = x\log(\lambda) - \lambda - \log(x!)$$

so that the Fisher information is given by

$$\mathbb{E}\left(\frac{\partial \log(f(x|\lambda))}{\partial \lambda}\right)^2 = \mathbb{E}\left(X/\lambda - 1\right)^2 = \mathbb{E}(X-\lambda)^2/\lambda^2 = \frac{1}{\lambda}.$$

Thus, an approximate $100(1-\alpha)$ percent confidence interval is given by

$$\bar{X} \pm z_{\alpha/2}\sqrt{\bar{X}/n}.$$

Bootstrap confidence intervals. If the distribution of $\Delta = \hat{\theta} - \theta_0$ was known, confidence intervals could be obtained via

$$\mathbb{P}(Q_\Delta(\alpha/2) \leq \hat{\theta} - \theta_0 \leq Q_\Delta(1-\alpha/2)) = 1-\alpha$$

as

$$\mathbb{P}(\hat{\theta} - Q_\Delta(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - Q_\Delta(\alpha/2)) = 1-\alpha.$$

But since θ_0 is not known, we use $\hat{\theta}$ in its place: we generate B bootstrap samples from the distribution with value $\hat{\theta}$, and compute the respective mles θ_b^*. The distribution of $\hat{\theta} - \theta_0$ is then approximated by that of $\theta^* - \hat{\theta}$ and the quantiles of this are used to form the approximate confidence interval.

2.4 The Bayesian Approach to Parameter Estimation

In the Bayesian approach, the unknown parameter θ is treated as a random variable with "prior" distribution $f_\Theta(\theta)$ representing what we know about the parameter before observing data. For a given value $\Theta = \theta$, the data have probability distribution $f_{X|\Theta}(x|\theta)$. If Θ has a continuous distribution, the joint distribution of X and Θ is

$$f_{X,\Theta}(x,\theta) = f_{X|\Theta}(x|\theta)f_\Theta(\theta)$$

and the marginal distribution of X is

$$f_X(x) = \int f_{X,\Theta}(x,\theta)d\theta = \int f_{X|\Theta}(x|\theta)f_\Theta(\theta)d\theta.$$

Finally, the distribution of Θ given the data, the so-called *posterior distribution*, is

$$f_{\Theta|X}(\theta|x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X,\Theta}(x,\theta)}{\int f_{X|\Theta}(x|\theta)f_\Theta(\theta)d\theta}.$$

Note that $f_{X|\Theta}(x|\theta)$ is the likelihood, and by the above

$$f_{\Theta|X}(\theta|x) \propto f_{X|\Theta}(x|\theta)f_\Theta(\theta).$$

Example: Poisson distribution. The unknown parameter is λ which has some prior $f_\Lambda(\lambda)$, and the data are n i.i.d. random variables $X_1, \ldots, X_n$ which given λ have a Poisson distribution with parameter λ. Thus,

$$f_{X_i|\Lambda}(x_i|\lambda) = \frac{\lambda^{x_i}}{x_i!}e^{-\lambda}$$

and

$$f_{X|\Lambda}(x|\lambda) = \frac{\lambda^{x_1+\cdots+x_n}e^{-n\lambda}}{x_1!\cdots x_n!}.$$

The posterior distribution of Λ given X is

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\sum_i x_i}e^{-n\lambda}f_\Lambda(\lambda)}{\int \lambda^{\sum_i x_i}e^{-n\lambda}f_\Lambda(\lambda)d\lambda}.$$

To evaluate this, one needs to specify the prior, and carry out the integration in the denominator.

Suppose one specified the distribution of Λ as gamma with shape parameter α and rate parameter ν (note that λ is already taken). Then

$$f_\Lambda(\lambda) = \frac{\nu^\alpha \lambda^{\alpha-1}e^{-\nu\lambda}}{\Gamma(\alpha)}$$

so that (canceling out constants)

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\sum_i x_i+\alpha-1}e^{-(n+\nu)\lambda}}{\int \lambda^{\sum_i x_i+\alpha-1}e^{-(n+\nu)\lambda}d\lambda}.$$

The integral in the denominator may look tricky, but in fact does not need to be computed: we can see that the posterior distribution is a gamma distribution with shape parameter $\sum_i x_i + \alpha$ and rate parameter $n + \nu$!

In the Bayesian paradigm, all information about Λ is contained in the posterior. We can estimate the parameter e.g. by the mean or mode (*posterior mean* and *posterior mode*, respectively) of this distribution. For a gamma distribution with shape α and rate ν these are α/ν and $(\alpha-1)/\nu$, giving the estimates

$$\frac{\sum_i x_i + \alpha - 1}{n+\nu}, \qquad \frac{\sum_i x_i + \alpha - 2}{n+\nu}.$$

The Bayesian analogue to the confidence interval is the interval from the $\alpha/2$ to the $1-\alpha/2$ quantile of the posterior. An alternative is a *high posterior density (HPD) interval*, obtained as a level set $f_{\Lambda|X}(\lambda|x) \geq c$ with coverage probability $1-\alpha$.

One could choose other priors, e.g., a uniform prior on $[0, 100]$, in which case

$$f_{\Lambda|X}(\lambda|x) = \frac{\lambda^{\sum_i x_i}e^{-n\lambda}}{\int_0^{100} \lambda^{\sum_i x_i}e^{-n\lambda}d\lambda}, \qquad 0 \leq \lambda \leq 100.$$

In this case, the denominator has to be integrated numerically (note the relation to the distribution function of the gamma distribution).

Example: Normal distribution. One conveniently reparametrizes the normal, replacing σ^2 by the *precision* $\xi = 1/\sigma^2$. Writing θ instead of μ,

$$f(x|\theta,\xi) = \sqrt{\xi/2\pi}e^{-\xi(x-\theta)^2/2}.$$

Rice covers several cases (unknown mean and known variance, known mean and unknown variance, unknown mean and unknown variance). For the last, one possibly model is to specify independent priors for Θ and Ξ as

$$\Theta \sim N(\theta_0, \xi_{\text{prior}}^{-1}), \qquad \Xi \sim \text{gamma}(\alpha, \lambda)$$

giving

$$\begin{aligned} & f_{\Theta,\Xi|X}(\theta, \xi|x) \\ & \quad \propto \quad f_{X|\Theta,\Xi}(x|\theta, \xi) f_\Theta(\theta) f_\Xi(\xi) \\ & \quad \propto \quad \exp\left(-\frac{\xi}{2}\sum_i (x_i - \theta)^2\right) \exp\left(-\frac{\xi_{\text{prior}}}{2}(\theta - \theta_0)^2\right) \xi^{n/2+\alpha-1} e^{-\lambda\xi}. \end{aligned}$$

which looks rather "messy". If the priors are quite flat (i.e., α, λ and ξ_{prior} are small), we get

$$f_{\Theta,\Xi|X}(\theta, \xi|x) \propto \exp\left(-\frac{\xi}{2}\sum_i (x_i - \theta)^2\right) \xi^{n/2}$$

and the marginal posterior of Θ is obtained by integrating out ξ as

$$f_{\Theta|X}(\theta|x) \propto \left(\sum (x_i - \theta)^2\right)^{-n/2}$$

from which after some algebra it can be shown that

$$\sqrt{n}\frac{\Theta - \bar{x}}{s} \sim t_{n-1}$$

corresponding to the result from maximum likelihood analysis.

More on priors. We saw that for the Poisson distribution, using a gamma prior gave a gamma posterior: in general, such priors (families of priors G for which when the data distribution is in a family H, then the posterior again is in G) are called *conjugate priors* (to the family of data distributions).

In many applications, it is desirable to use flat or "noninformative" priors—but this hard to make precise. In the Poisson case with gamma priors, these are flat when α and ν are small. But taking limits gives

$$f_\Lambda(\lambda) \propto \lambda^{-1}, \qquad \lambda > 0$$

which is not a valid density!

Such priors are called *improper priors*, and may result in proper or improper posteriors.

E.g., in the Poisson case, using the improper prior $f_\Lambda(\lambda) \propto \lambda^{-1}$ results in the proper posterior

$$f_{\Lambda|X}(\lambda|x) \propto \lambda^{\sum x_i - 1} e^{-n\lambda}$$

i.e., a gamma distribution with shape $\sum_i x_i - 1$ and rate n, as obtained by taking limits in the posterior.

E.g., in the normal case with unknown mean and variance, one can take

$$f_\Theta(\theta) \propto 1, \qquad f_\Xi(\xi) \propto \xi^{-1},$$

giving the joint posterior

$$\begin{aligned} & f_{\Theta,\Xi|X}(\theta,\xi|x) \\ & \quad \propto \quad \xi^{n/2-1} \exp\left(-\frac{\xi}{2}\sum_i (x_i-\theta)^2\right) \\ & \quad \propto \quad \xi^{n/2-1} \exp\left(-\frac{\xi}{2}(n-1)s^2\right) \exp\left(-\frac{n\xi}{2}(\theta-\bar{x})^2\right). \end{aligned}$$

Conditional on ξ, Θ is normal with mean $\bar{x}$ and precision $n\xi$.

Large sample normal approximation to the posterior. We have

$$f_{\Theta|X}(\theta|x) \propto f_\Theta(\theta) f_{X|\Theta}(x|\theta) = \exp(\log(f_\Theta(\theta))) \exp(\ell(\theta))$$

If the sample is large, the posterior is dominated by the likelihood, and the prior is nearly constant where the likelihood is large. Thus, approximately

$$f_{\Theta|X}(\theta|x) \propto \exp(\ell(\hat{\theta}) + (\theta-\hat{\theta})\ell'(\hat{\theta}) + (\theta-\hat{\theta})^2\ell''(\hat{\theta})/2) \propto \exp((\theta-\hat{\theta})^2\ell''(\hat{\theta})/2)$$

which is (proportional to) the density of a normal with mean $\hat{\theta}$ and variance $-(\ell''(\hat{\theta}))^{-1}$.

Computational aspects. Bayesian inference typically requires considerable computational power, e.g., for computing the normalizing constants. In high dimensional problems, difficulties arise, and one can use sophisticated methods such as *Gibbs sampling*.

Consider inference for a normal with unknown mean and variance and an improper prior ($\alpha \to 0$, $\lambda \to 0$, $\xi_{\text{prior}} \to 0$. Then

$$f_{\Theta,\Xi|X}(\theta,\xi|x) \propto \xi^{n/2-1} \exp\left(-\frac{n\xi}{2}(\theta-\bar{x})^2\right).$$

To study the posterior by Monte Carlo, one would draw many pairs (θ_k, ξ_k) from this joint density—but how?

Gibbs sampling alternates between simulating from the conditional distribution of one parameter given the others. In our case, we note that given ξ, θ is normal with mean $\bar{x}$ and precision $n\xi$; given θ, ξ has a gamma distribution. One would then proceed as follows:

1. Choose an initial value θ_0, e.g., $\bar{x}$.
2. Generate ξ_0 from a gamma density with parameters $n/2$ and $n(\theta_0-\bar{x})^2/2$ (which will not work, as the latter is zero, so one really needs another initial value).
3. Generate θ_1 from a normal distribution with mean $\bar{x}$ and precision $n\xi_0$.
4. Generate ξ_1 from a gamma density with parameters $n/2$ and $n(\theta_1-\bar{x})^2/2$.
5. etc.

After a "burn-in" period of a several hundred steps, one obtains pairs which approximately have the posterior distribution (but are not independent of one another).

2.5 Efficiency

Given a variety of possible estimates, which one should we use? Ideally, the one whose sampling distribution was most concentrated about the true value. One possible concentration measure is the mean squared error

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta_0)^2 = \text{var}(\hat{\theta}) + (\mathbb{E}(\hat{\theta}) - \theta_0)^2.$$

We say that an estimate $\hat{\theta}$ is *unbiased* if $\mathbb{E}(\hat{\theta}) = \theta_0$. For unbiased estimates, the mean squared error equals the variance, and hence comparison of MSEs reduces to comparing the variances or standard errors, respectively.

For two estimates $\hat{\theta}$ and $\tilde{\theta}$, the *efficiency* of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined as

$$\text{eff}(\hat{\theta}, \tilde{\theta}) = \frac{\text{var}(\hat{\theta})}{\text{var}(\tilde{\theta})}.$$

Theorem (Cramér-Rao Inequality). *Let $X_1, \ldots, X_n$ be i.i.d. with density function $f(x|\theta)$. Let $T = t(X_1, \ldots, X_n)$ be an unbiased estimate of θ. Then under suitable smoothness assumptions on $f(x|\theta)$,*

$$\text{var}(T) \geq 1/(nI(\theta)).$$

Proof. Let

$$Z = \sum_{i=1}^{n} \frac{\partial \log(f(X_i|\theta))}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{f(X_i|\theta)} \frac{\partial f(X_i|\theta)}{\partial \theta}$$

We already know that $\mathbb{E}(Z) = 0$ and $\text{var}(Z) = nI(\theta)$. We will show that if T is unbiased, $\text{cov}(T, Z) = 1$ from which the assertion by noting that

$$\text{cov}(T, Z)^2 \leq \text{var}(T)\text{var}(Z).$$

Since Z has mean 0,

$$\begin{aligned}
\text{cov}(T, Z) &= \mathbb{E}(TZ) \\
&= \int \cdot \int t(x_1, \ldots, x_n) \left(\sum_{i=1}^{n} \frac{1}{f(x_i|\theta)} \frac{\partial f(x_i|\theta)}{\partial \theta} \right) \prod_{j=1}^{n} f(x_j|\theta) dx_j \\
&= \int \cdot \int t(x_1, \ldots, x_n) \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i|\theta) dx_i \\
&= \frac{\partial}{\partial \theta} \int \cdot \int t(x_1, \ldots, x_n) \prod_{i=1}^{n} f(x_i|\theta) dx_i \\
&= \frac{\partial}{\partial \theta} \mathbb{E}(T) \\
&= \frac{\partial}{\partial \theta} \theta \\
&= 1
\end{aligned}$$

as asserted.

Example: Poisson distribution. We know that $I(\lambda) = 1/\lambda$. Hence, for any unbiased estimator of λ, $\text{var}(T) \geq \lambda/n$. On the other hand, the MLE $\bar{X} = S/n$ is unbiased with variance λ/n, hence attains the bound, and thus is a MVUE (minimum variance unbiased estimator).

2.6 Sufficiency

The notion of sufficiency arises as an attempt to answer the following question: for a sample $X_1, \ldots, X_n$ from $f(x|\theta)$, is there a statistic $T(X_1, \ldots, X_n)$ which contains all information in the sample about θ? (Think of Bernoulli experiments: we have the feeling that only the number of successes matters.)

Definition. *A statistic $T(X_1, \ldots, X_n)$ is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ does not depend on θ, for any value of t.*

Example: Bernoulli experiment. Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli random variables with success probability $\mathbb{P}(X = 1) = \theta$, and let $T = X_1 + \cdots + X_n$. We know that T has a binomial distribution with parameters n and θ. Thus,

$$\begin{aligned}
&\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | T = t) \\
&\quad = \frac{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}{\mathbb{P}(T = t)} \\
&\quad = \frac{\prod \theta^{x_i}(1-\theta)^{1-x_i}}{\mathbb{P}(T = t)} \\
&\quad = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&\quad = \frac{1}{\binom{n}{t}}.
\end{aligned}$$

Theorem. *A necessary and sufficient condition for $T(X_1, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint probability function factors in the form*

$$f(x_1, \ldots, x_n|\theta) = g(T(x_1, \ldots, x_n), \theta)h(x_1, \ldots, x_n).$$

Proof. We give a proof for the discrete case. Let $X = (X_1, \ldots, X_n)$ and $x = (x_1, \ldots, x_n)$. Suppose the pmf factors as given in the theorem. Then

$$\mathbb{P}(T = t) = \sum_{x:T(x)=t} \mathbb{P}(X = x) = g(t, \theta) \sum_{x:T(x)=t} h(x).$$

Hence,

$$\mathbb{P}(X = x | T = t) = \frac{\mathbb{P}(X = x, T = t)}{\mathbb{P}(T = t)} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)}$$

does not depend on θ, as was to be shown.

Conversely, suppose the conditional distribution of X given T does not depend on θ. Let

$$g(t, \theta) = \mathbb{P}(T = t|\theta), \qquad h(x) = \mathbb{P}(X = x|T = t).$$

Then

$$\mathbb{P}(X = x|\theta) = \mathbb{P}(T = t|\theta)\mathbb{P}(X = x|T = t) = g(t, \theta)h(x)$$

as was to be shown.

Example: Bernoulli experiment. Writing $t = \sum_i x_i$, we have

$$f(x_1, \ldots, x_n|\theta) = \theta^t(1-\theta)^{n-t} = \left(\frac{\theta}{1-\theta}\right)^t (1-\theta)^n$$

which gives $g(t, \theta)$, and we can take $h(x_1, \ldots, x_n) = 1$.

Example: Normal distribution. For a random sample from the normal distribution with unknown mean and variance, we have

$$\begin{aligned} & f(x_1, \ldots, x_n|\mu, \sigma^2) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right) \\ &= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)\right) \end{aligned}$$

which only depends on $x_1, \ldots, x_n$ through the sufficient statistics $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$.

Theorem. *If T is sufficient for θ, the mle of θ is a function of T.*

Proof. Because

$$f(x_1, \ldots, x_n|\theta) = g(T(x_1, \ldots, x_n), \theta)h(x_1, \ldots, x_n).$$

the mle is found by maximizing $g(T(x_1, \ldots, x_n), \theta)$, i.e., a function of the sufficient statistic $T(x_1, \ldots, x_n)$.

Theorem (Rao-Blackwell Theorem). *Let $\hat{\theta}$ be an estimate of θ with $\mathbb{E}(\hat{\theta}^2) < \infty$ for all θ. Suppose that T is sufficient for θ, and let $\tilde{\theta} = \mathbb{E}(\hat{\theta}|T)$. Then, for all θ,*

$$\mathbb{E}((\tilde{\theta} - \theta)^2) \leq \mathbb{E}((\hat{\theta} - \theta)^2)$$

and the inequality is strict unless $\tilde{\theta} = \hat{\theta}$.

Proof. By the theorem of iterated conditional expectation,

$$\mathbb{E}(\tilde{\theta}) = \mathbb{E}(\mathbb{E}(\hat{\theta}|T)) = \mathbb{E}(\hat{\theta}).$$

Thus, to compare the MSEs we only need to compare the variances. Now by a result on conditional expectations,

$$\text{var}(\hat{\theta}) = \text{var}(\mathbb{E}(\hat{\theta}|T)) + \mathbb{E}(\text{var}(\hat{\theta}|T))$$

or

$$\text{var}(\hat{\theta}) = \text{var}(\tilde{\theta}) + \mathbb{E}(\text{var}(\hat{\theta}|T)).$$

Thus, $\text{var}(\hat{\theta}) > \text{var}(\tilde{\theta})$ unless $\mathbb{E}(\text{var}(\hat{\theta}|T)) = 0$, in which case $\hat{\theta}$ must be a function of T, which would imply $\hat{\theta} = \tilde{\theta}$.

The Rao-Blackwell theorem gives a strong rationale for basing estimators on sufficient statistics if they exist: if they are not functions of a sufficient statistic, their variance can be reduced without changing their bias.

3 Testing Hypotheses and Assessing Goodness of Fit

3.1 Introduction

Suppose we have two coins:

$$P_0(H) = 0.5, \qquad P_1(H) = 0.7$$

(where H denotes "head"). Suppose one of these coins is chosen, tossed 10 times, and the number of heads reported, without telling which coin was chosen. How should we decide which one it was?

Suppose 2 heads are observed. Then $P_0(2)/P_1(2)$, the *likelihood ratio*, is

```
R> dbinom(2, 10, 0.5) / dbinom(2, 10, 0.7)

[1] 30.37623
```

(as the number of heads is binomial with $n = 10$ and probability 0.5 or 0.7, respectively), strongly favoring coin 0. If we observed 8 heads,

```
R> dbinom(8, 10, 0.5) / dbinom(8, 10, 0.7)

[1] 0.1882232
```

would favor coin 1.

We specify two *hypotheses*, H_0: coin 0 was tossed and H_1: coin 1 was tossed.

First developing a Bayesian methodology, we need to specify prior probabilities $\mathbb{P}(H_0)$ and $\mathbb{P}(H_1)$. If the "basic" case of no a priori preference for either hypothesis, $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 1/2$. After observing the data we can compute the posterior probabilities

$$\mathbb{P}(H_0|x) = \frac{\mathbb{P}(H_0, x)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x|H_0)\mathbb{P}(H_0)}{\mathbb{P}(x)}$$

and $\mathbb{P}(H_1|x)$ and the corresponding ratio of posterior probabilities as

$$\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)} = \frac{\mathbb{P}(H_0)}{\mathbb{P}(H_1)} \frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_1)}$$

(i.e., the ratio of posteriors is the product of the ratio of the priors and the likelihood ratio).

How to decide? Reasonably, choose the hypothesis with higher posterior probability. I.e., choose H_0 if

$$\frac{\mathbb{P}(H_0|x)}{\mathbb{P}(H_1|x)} = \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_1)} > 1$$

or equivalently, if

$$\frac{\mathbb{P}(x|H_0)}{\mathbb{P}(x|H_1)} > c$$

where the *critical value* c depends upon the prior probabilities.

In our case, the likelihood ratios for the possible values $x = 0, \ldots, 10$ are

```
R> x <- 0 : 10
R> dbinom(x, 10, 0.5) / dbinom(x, 10, 0.7)

 [1] 165.38171688  70.87787866  30.37623371  13.01838588   5.57930823
 [6]   2.39113210   1.02477090   0.43918753   0.18822323   0.08066710
[11]   0.03457161
```

If e.g. $c = 1$, $\mathbb{P}(H_0) = \mathbb{P}(H_1)$, and we accept H_0 as long as $X \leq 6$. We can make two possible errors: reject H_0 when it is true, or accept H_0 when it is false, with probabilities

$$\mathbb{P}(\text{reject } H_0|H_0) = \mathbb{P}(X > 6|H_0), \qquad \mathbb{P}(\text{accept } H_0|H_1) = \mathbb{P}(X \leq 6|H_1)$$

with corresponding values

```
R> pbinom(6, 10, 0.5, lower.tail = FALSE)

[1] 0.171875

R> pbinom(6, 10, 0.7)

[1] 0.3503893
```

respectively.

If e.g. $c = 0.1$, $\mathbb{P}(H_0) = 10\mathbb{P}(H_1)$, H_0 is accepted iff $X \leq 8$, with error probabilities

```
R> pbinom(8, 10, 0.5, lower.tail = FALSE)

[1] 0.01074219

R> pbinom(8, 10, 0.7)

[1] 0.8506917
```

respectively.

3.2 The Neyman-Pearson Paradigm

Neyman and Pearson formulated their approach in the framework of decision problems, bypassing the necessity of specifying prior probabilities, but introducing a fundamental asymmetry between the *null hypothesis* H_0 and the *alternative hypothesis* H_A (or H_1).

Terminology:

- Rejecting H_0 when it is true: *type I error*.
- Probability of a type I error: *significance level* of the test, typically denoted by α.
- Accepting H_0 when it is false: *type II error*
- Probability of a type II error is typically denoted by β.
- The probability of rejecting H_0 when it is false: *power* of the test, equals $1 - \beta$.
- Testing is based on a *test statistic* (e.g., the likelihood ratio) computed from the data.
- Sets of values leading to acceptance or rejection of H_0: *acceptance region* and *rejection region*, respectively.
- Probability distribution of the test statistic when H_0 is true: *null distribution*.

In the introductory example, the null and alternative hypotheses completely specified the probability distribution of the data (number of heads) as binomial with parameters 10 and 0.5 or 0.7, respectively: such hypotheses are called *simple hypotheses*.

Theorem (Neyman-Pearson Lemma). *Suppose H_0 and H_A are simple hypotheses and that the test that rejects H_0 whenever the likelihood ratio is less than c has significance level α. Then* any other test *for which the significance level does not exceed α has power not exceeding that of the likelihood ratio test.*

Proof. Any (non-randomized) test is equivalent to its decision function $d(x)$ which is the indicator of rejecting H_0 when observing x (i.e., the indicator of the rejection region). The significance level is

$$\mathbb{P}(\text{reject } H_0|H_0) = \mathbb{P}(d(X) = 1|H_0) = \mathbb{E}_0(d(X)),$$

the power is

$$\mathbb{P}(\text{reject } H_0|H_A) = \mathbb{P}(d(X) = 1|H_A) = \mathbb{E}_A(d(X))$$

(note that this is not correct in the textbook). Write f_0 and f_A for the densities (or pmfs) under H_0 and H_A, respectively, and d^* for the decision function of the likelihood ratio test, i.e.,

$$d^*(x) = 1 \Leftrightarrow f_0(x)/f_A(x) < c \Leftrightarrow cf_A(x) - f_0(x) > 0.$$

For all x, we have

$$d(x)(cf_A(x) - f_0(x)) \leq d^*(x)(cf_A(x) - f_0(x))$$

(if $d^*(x) = 0$, $cf_A(x) - f_0(x) \leq 0$, if $d^*(x) = 1$, $cf_A(x) - f_0(x) > 0$). Integrating gives

$$c\mathbb{E}_A(d(X)) - \mathbb{E}_0(d(X)) \leq c\mathbb{E}_A(d^*(X)) - \mathbb{E}_0(d^*(X))$$

or equivalently,

$$\mathbb{E}_A(d^*(X)) - \mathbb{E}_A(d(X)) \geq (\mathbb{E}_0(d^*(X)) - \mathbb{E}_0(d(X)))/c.$$

Example: Normal distribution. Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with known variance σ^2, and consider the simple hypotheses

$$H_0 : \mu = \mu_0 \qquad H_A : \mu = \mu_A.$$

By the Neyman-Pearson lemma, among all tests with given significance level α, the one that rejects for small values of the likelihood ratio is most powerful. Now

$$\frac{f_0(x)}{f_A(x)} = \exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n (X_i - \mu_0)^2 - \sum_{i=1}^n (X_i - \mu_A)^2\right)\right)$$

where the inner parenthesis term equals

$$2n\bar{X}(\mu_0 - \mu_A) + n(\mu_A^2 - \mu_0^2)$$

If $\mu_0 > \mu_A$, the LR is small iff $\bar{X}$ is small; if $\mu_0 < \mu_A$, the LR is small iff $\bar{X}$ is large. Assume the latter, then the LRT rejects iff $\bar{X} > c$ where

$$\alpha = \mathbb{P}(\bar{X} > c|H_0) = \mathbb{P}_0\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > \frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right).$$

Thus,

$$c = \mu_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}.$$

Composite hypotheses. If a hypothesis does not completely specify the probability distribution.

Example. Experiment: subject is asked to identify the suits of 20 cards drawn randomly with replacement from a 52 card deck. Null hypothesis: person is "just guessing": i.e., simple where number of correct identifications is binomial with $n = 20$ and $p = 0.25$. Alternative: person has extrasensory abilities. Could be "anything" (compatible with guessing better than random): composite.

Significance levels and p-values. To construct a test using the Neyman-Pearson approach we only need the null distribution. If T is the number of correct guesses, we will reject H_0 if T is "too large", and determine the critical value t_0 in a way that $\mathbb{P}(T \geq t_0|H_0) = \alpha$ (not always possible without randomization), or determine α from the above for given critical value (always possible).

What are good choices for α? "Cultural conventions" such as $\alpha = 0.05$ or $\alpha = 0.01$.

Do we need to specify α in advance? Alternatively, when observing t, we could report $p = \mathbb{P}(T \geq t|H_0)$, the so-called p-value. This is (in general defined as) the smallest significance level at which the null hypothesis would be rejected. E.g., when observing $t = 9$, the p-value is

```
R> pbinom(8, 20, 0.25, lower.tail = FALSE)
```

```
[1] 0.04092517
```

and H_0 would be rejected at the 5% level (as 0.05 is no less than the p-value).

p-values go back to Fisher's "fiducial values", interpreted as the null probability of observing something more extreme (in a sens, less fitting with the null) than what was actually observed.

> **BIG NOTE:** the p-value is a function of the observations, and hence a random variable. It is NOT the probability that the null is true.

Null hypothesis. Discuss the asymmetry between null and alternative hypotheses. Which is which depends on scientific context, custom, and convenience. Note that in a sense the null is "preferred": there is more "value" in rejecting the null in favor of the alternative than not rejecting the null (and hence accepting it).

Uniformly most powerful tests. In some cases, the Neyman-Pearson tests can be extended to composite hypotheses. If H_A is composite, a test that is most powerful for *every* simple alternative in H_A is said to be *uniformly most powerful (UMP)*.

As an example, consider testing $H_0 : \mu = \mu_0$ against $H_A : \mu > \mu_0$ for the case of a sample from the normal with unknown mean μ and known variance σ^2. For every single $\mu_A > \mu_0$, the UMP Neyman-Pearson test rejects H_0 for $\bar{X} > x_0$, where x_0 is chosen such that the significance level is α, and hence depends on α, μ_0 and n but *not* μ_A. Hence, it is the same for every single alternative, and thus UMP.

One can argue that the test is also UMP for $H_0 : \mu \leq \mu_0$ against $H_A : \mu > \mu_0$ (among all tests with *size* $\sup_{\mu \in H_0} \alpha(\mu)$. But it is *not* UMP for testing $H_0 : \mu = \mu_0$ against $H_A : \mu \neq \mu_0$ (as for alternatives $> \mu_0$ and $< \mu_0$, the UMP tests reject for large and small values of $\bar{X}$, respectively.

One- and two-sided alternatives. Alternatives of the form $\mu < \mu_0$ or $\mu > \mu_0$ are called one-sided alternatives; the alternative $\mu \neq \mu_0$ is a two-sided alternative.

3.3 Duality of Confidence Intervals and Hypothesis Tests

Example: Normal distribution. Consider again sample $X_1, \ldots, X_n$ from a normal distribution with unknown mean μ and known variance σ^2. Consider the testing problem

$$H_0 : \mu = \mu_0, \qquad H_A : \mu \neq \mu_0.$$

Consider a level α test which rejects if $|\bar{X} - \mu_0| > x_0$, hence $x_0 = \sigma_{\bar{X}} z_{1-\alpha/2}$. The test accepts if $|\bar{X} - \mu_0| \leq x_0$, or equivalently

$$-x_0 \leq \bar{X} - \mu_0 \leq x_0$$

or equivalently

$$\bar{X} - x_0 \leq \mu_0 \leq \bar{X} + x_0.$$

As the acceptance probability is $1-\alpha$, the above gives the (already known) $100(1-\alpha)\%$ confidence interval for μ. I.e., μ_0 lies in the confidence interval for μ if and only if the hypothesis test accepts.

More generally: let θ be a parameter of a family of probability distributions, and Θ be the set of all possible values of θ.

Theorem. *Suppose for every θ_0 in Θ there is a level α test of the hypothesis $H_0 : \theta = \theta_0$. Denote the acceptance region of this test by $A(\theta_0)$. Then the set*

$$C(X) = \{\theta : X \in A(\theta)\}$$

is a $100(1-\alpha)\%$ confidence region for θ.

Proof. We have

$$\theta_0 \in C(X) \Leftrightarrow X \in A(\theta_0).$$

Hence,

$$\mathbb{P}(\theta_0 \in C(X)|\theta_0) = \mathbb{P}(X \in A(\theta_0)|\theta_0) = 1 - \alpha.$$

Theorem. *Suppose that $C(X)$ is a $100(1-\alpha)\%$ confidence region for θ, i.e., for every θ_0,*

$$\mathbb{P}(\theta_0 \in C(X)|\theta_0) = 1 - \alpha.$$

Then

$$A(\theta_0) = \{X : \theta_0 \in C(X)\}$$

defines an acceptance region for a level α test of the null hypothesis $H_0 : \theta = \theta_0$.

Proof.

$$\mathbb{P}(X \in A(\theta_0)|\theta_0) = \mathbb{P}(\theta_0 \in C(X)|\theta_0) = 1 - \alpha.$$

3.4 Generalized Likelihood Ratio Tests

Consider a general hypothesis testing problem of the form $H_0 : \theta \in \Theta_0$ against $H_A : \theta \in \Theta_A$. If both Θ_0 and Θ_A were simple, we know that the likelihood ratio test is optimal (UMP). In the general case, we do not know (and UMP tests usually do not exist), but it still makes sense to base a test on the ratio of the (maximal) likelihoods under the null and alternative, yielding the generalized likelihood ratio

$$\Lambda^* = \frac{\sup_{\theta \in \Theta_0} \text{lik}(\theta)}{\sup_{\theta \in \Theta_A} \text{lik}(\theta)}$$

or the usually preferred

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} \text{lik}(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_A} \text{lik}(\theta)}$$

Both LRTs reject H_0 for small values of the test statistic.

Example: normal distribution. Consider again sample $X_1, \ldots, X_n$ from a normal distribution with unknown mean μ and known variance σ^2. Consider the testing problem

$$H_0 : \mu = \mu_0, \qquad H_A : \mu \neq \mu_0.$$

The likelihood under the null is

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu_0)^2\right)$$

and the maximum likelihood under the alternative is

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right).$$

The likelihood ratio statistic is thus:

$$\exp\left(-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X})^2\right)\right)$$

and rejecting for small values of Λ is equivalent to rejecting for large values of

$$-2\log(\Lambda) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(X_i - \mu_0)^2 - \sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

which using

$$\sum_{i=1}^{n}(X_i - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$$

can be simplified to

$$-2\log(\Lambda) = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2} = \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2.$$

Under H_0, $(\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ has a standard normal distribution, hence its square $-2\log(\Lambda)$ has a χ^2_1 distribution, and the LRT rejects H_0 if

$$|\bar{X} - \mu_0| \geq \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}.$$

Theorem. *Under smoothness conditions on the density (or pmf) functions involved, the null distribution of* $-2\log(\Lambda)$ *tends to a chi-squared distribution with* $\dim(\Theta_0 \cup \Theta_A) - \dim(\Theta_0)$ *degrees of freedom as the sample size tends to infinity.*

In the above, the dimensions are the numbers of free parameters. In the above example, Θ_0 is simple and hence has no free parameters, whereas Θ_A specifies σ^2 but has μ as free parameter, so the degrees of freedom are $1 - 0 = 1$. In the example, the exact null distribution of $-2\log(\Lambda)$ is χ^2_1.

3.5 Likelihood Ratio Tests for the Multinomial Distribution

Suppose that under H_0, the vector p of multinomial cell probabilities is specified as $p = p(\theta)$ where θ is a parameter that may be unknown; under H_A, they are free apart from the requirements of being nonnegative with sum one.

The likelihood function is

$$\frac{n!}{x_1! \cdots x_m!} p_1(\theta)^{x_1} \cdots p_m(\theta)^{x_m}.$$

Let $\hat{\theta}$ be the (restricted to Θ) mle; the unrestricted mle is $\hat{p}_i = x_i/n$, and the likelihood ratio statistic is thus

$$\Lambda = \frac{\frac{n!}{x_1! \cdots x_m!} p_1(\hat{\theta})^{x_1} \cdots p_m(\hat{\theta})^{x_m}}{\frac{n!}{x_1! \cdots x_m!} \hat{p}_1^{x_1} \cdots \hat{p}_m^{x_m}} = \prod_{i=1}^{m} \left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{x_i}.$$

Since $x_i = n\hat{p}_i$,

$$-2\log(\Lambda) = 2n \sum_{i=1}^{m} \hat{p}_i \log\left(\frac{p_i(\hat{\theta})}{\hat{p}_i} \right) = 2 \sum_{i=1}^{n} O_i \log\left(\frac{O_i}{E_i} \right)$$

where $O_i = x_i = n\hat{p}_i$ and $E_i = np_i(\hat{\theta})$ are the observed and expected counts, respectively. Under H_A, there are $m-1$ free parameters. Hence, if Θ has k free parameters, $-2\log(\Lambda)$ asymptotically has a chi-squared distribution with $m-k-1$ degrees of freedom.

Typically, instead of the likelihood ratio test statistic the asymptotically equivalent Pearson chi-squared test statistic

$$X^2 = \sum_{i=1}^{n} \frac{(x_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})}$$

is used. To see the asymptotic equivalence, note that when n is large and H_0 is true, $\hat{p}_i \approx p_i(\hat{\theta})$. Using the Taylor series expansion

$$x \log\left(\frac{x}{x_0} \right) = (x - x_0) + \frac{1}{2x_0}(x - x_0)^2 + \cdots$$

about x_0 one obtains

$$-2\log(\Lambda) \approx 2n \sum_{i=1}^{m} (\hat{p}_i - p_i(\hat{\theta})) + n \sum_{i=1}^{m} \frac{(\hat{p}_i - p_i(\hat{\theta}))^2}{p_i(\hat{\theta})}$$

where the first term is zero since the probabilities sum to one and the second term is identical to X^2.

As a special case, for a simple $H_0 : p_i = p_{i,0}$, the LR and chi-squared test statistics are asymptotically χ^2_{m-1} ("chi-squared goodness of fit test for given probabilities").

3.6 Assessing Goodness of Fit

We already know from Statistics 1 that goodness of fit can be judged via probability or preferably quantile plots, which graphically illustrate the goodness of fit of data to suitable families of probability distributions.

There is also a huge variety of goodness of fit hypothesis tests for nulls that the probability distribution comes from a family of distributions against, e.g., the alternative that it does not.

For discrete distributions (with finite support), these can be based on the likelihood ratio or chi-squared tests for the multinomial distribution discussed above.

A very popular problem is testing for normality, either against the general alternative of non-normality, or against departures which take the form of asymmetry (skewness) or non-normal kurtosis, or jointly (Jarque-Bera test, implemented in package tseries).

For departures against symmetry, goodness-of-fit tests can be based on the sample coefficient of skewness

$$b_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

which rejects for large values of $|b_1|$. Under the null of normality, this is asymptotically normal with mean 0 and variance $6/n$.

4 Comparing Two Samples

4.1 Comparing Two Independent Samples

4.1.1 Methods Based on the Normal Distribution

Suppose a sample $X_1, \ldots, X_n$ is drawn from a normal distribution with mean μ_X and variance σ^2, and that an independent sample $Y_1, \ldots, Y_m$ is drawn from a normal distribution with mean μ_Y and variance σ^2. We are interested in $\mu_X - \mu_Y$, which is naturally estimated by $\bar{X} - \bar{Y}$. By normality of the samples,

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

If σ^2 were known, a confidence interval for $\mu_X - \mu_Y$ could be based on

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

which has a standard normal distribution, giving

$$(\bar{X} - \bar{Y}) \pm z_{1-\alpha/2}\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}$$

In general, σ^2 will not be known and must be estimated, e.g., by using the *pooled sample variance*

$$s_p^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}.$$

Theorem. *Suppose that $X_1, \ldots, X_n$ are independent and normally distributed random variables with mean μ_X and variance σ^2, and that $Y_1, \ldots, Y_m$ are independent and normally distributed*

random variables with mean μ_Y and variance σ^2, and that the Y_i are independent of the X_i. Then the statistic

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{s_p\sqrt{1/n + 1/m}}$$

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

Let

$$s_{\bar{X}-\bar{Y}} = s_p\sqrt{1/n + 1/m}$$

denote the estimated standard deviation of $\bar{X} - \bar{Y}$.

Corollary. *Under the above assumptions, a $100(1 - \alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is*

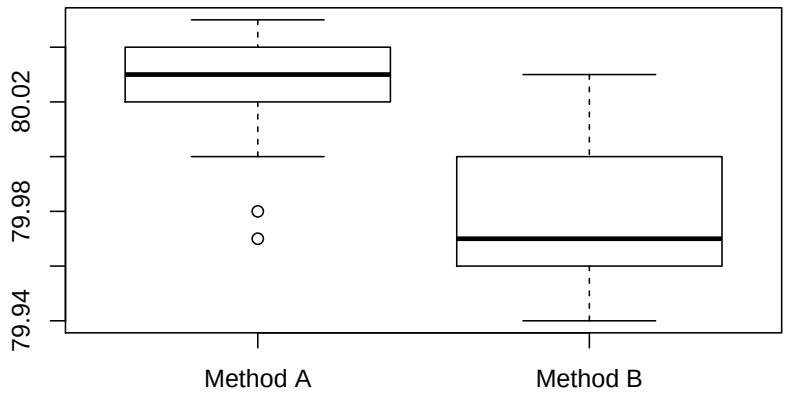
$$(\bar{X} - \bar{Y}) \pm t_{m+n-2,1-\alpha/2} s_{\bar{X}-\bar{Y}}.$$

Example: Ice. Two methods, A and B, were used to determine the latent heat of fusion of ice (Natrella, 1963). Measurements were obtained for the change in total heat from ice at $-72°$ C to water at $0°$ C in calories per gram of mass:

```
R> A <- scan("Data/icea.txt")
R> A

 [1] 79.98 80.04 80.02 80.04 80.03 80.03 80.04 79.97 80.05 80.03 80.02 80.00
[13] 80.02

R> B <- scan("Data/iceb.txt")
R> B

[1] 80.02 79.94 79.98 79.97 79.97 80.03 79.95 79.97
```

It is fairly obvious that there is a difference between the methods:

```
R> boxplot(A, B, names = c("Method A", "Method B"))
```



Doing computations by hand:

```
R> n_A <- length(A); m_A <- mean(A); s_A <- sd(A)
R> n_B <- length(B); m_B <- mean(B); s_B <- sd(B)
R> s_p <- sqrt(((n_A - 1) * s_A^2 + (n_B - 1) * s_B^2) / (n_A + n_B - 2))
R> s_p

[1] 0.02693052
```

giving the following estimates for $\bar{X} - \bar{Y}$ and $s_{\bar{X}-\bar{Y}}$:

```
R> Delta <- m_A - m_B
R> Delta

[1] 0.04201923

R> s_Delta <- s_p * sqrt(1 / n_A + 1 / n_B)
R> s_Delta

[1] 0.01210146
```

with t quantile

```
R> q <- qt(0.975, n_A + n_B - 2)
```

and confidence interval

```
R> c(Delta - q * s_Delta, Delta + q * s_Delta)

[1] 0.01669058 0.06734788
```

But because we have R, we can simply use function `t.test` to obtain the confidence interval and corresponding hypothesis test:

```
R> t.test(A, B, var.equal = TRUE)

        Two Sample t-test

data:  A and B
t = 3.4722, df = 19, p-value = 0.002551
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01669058 0.06734788
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

In fact, for hypothesis testing for the two-sample problem, there are three common alternative hypotheses:

$$H_1 : \mu_X \neq \mu_Y, \qquad H_2 : \mu_X < \mu_Y, \qquad H_3 : \mu_X > \mu_Y.$$

The first is a *two-sided alternative*, the other two are *one-sided alternatives*. Hypothesis tests for all three alternatives are based on the t statistic

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}}$$

with rejection regions

$$H_1 : |t| > t_{m+n-2,1-\alpha/2}, \qquad H_2 : t < t_{m+n-2,\alpha}, \qquad H_3 : t > t_{m+n-2,1-\alpha}$$

(and corresponding two- or one sided confidence intervals). E.g., to test against the alternative $\mu_A > \mu_B$ for the ice data,

```
R> t.test(A, B, alternative = "greater", var.equal = TRUE)

        Two Sample t-test

data:  A and B
t = 3.4722, df = 19, p-value = 0.001276
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.0210942       Inf
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

Rice (page 426f) shows that the t test is actually the LRT in the case of unknown but equal variances in the two samples.

If the variances are not known to be equal, a natural estimate of $\text{var}(\bar{X} - \bar{Y})$ is

$$\frac{s_X^2}{n} + \frac{s_Y^2}{m}$$

If this is used in the denominator of the test statistic, the t distribution no longer holds exactly, but approximately with

$$\frac{(s_X^2/n + s_Y^2/m)^2}{(s_x^2/n)^2/(n-1) + (s_Y^2/m)^2/(m-1)}$$

degrees of freedom: so-called *Welch* approximation to the sampling distribution. This is actually the default in R:

```
R> t.test(A, B)

        Welch Two Sample t-test

data:  A and B
t = 3.2499, df = 12.027, p-value = 0.006939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.01385526 0.07018320
sample estimates:
mean of x mean of y
 80.02077  79.97875
```

4.1.2 A Nonparametric Method: The Mann-Whitney Test

Nonparametric methods do not assume the data follow a particular distributional form. Often, data are replaced by ranks, making results invariant under monotonic transformations, and moderating the influence of outliers.

Suppose we have two independent sample $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from probability distributions F and G, respectively, and that it is desired to test the null hypothesis that $F = G$. We will develop the Mann-Whitney test, also known as the *Wilcoxon rank sum test*.

This is based on the idea that under the null, assigning the pooled (sorted) observations to the samples is "random" in the sense that all assignments are equiprobable. Consider a simple example. Suppose observations are

$$X : 1, 3 \qquad Y : 4, 6$$

with corresponding ranks

$$X : 1, 2 \qquad Y : 3, 4$$

and rank sums 3 and 7, respectively. Now, under the null, every assignment of the ranks to the samples is equally likely. For the ranks of the second group, we have

Ranks	$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
R	3	4	5	5	6	7

and thus

$$\mathbb{P}(R = 7) = 1/6.$$

In the general case, under the null every possible assignment of the $m+n$ ranks to the n elements of the second group is equally likely.

Example: Ice. We compute the ranks for methods A and B:

```
R> C <- c(A, B)
R> r_A <- rank(C)[seq_along(A)]
R> r_B <- rank(C)[seq_along(B) + length(A)]
```

Note how ties are handled. The rank sum of the smaller sample is

```
R> sum(r_B)
```

```
[1] 51
```

What about the distribution of the rank sums under the null? We can use R:

```
R> wilcox.test(A, B)
```

```
        Wilcoxon rank sum test with continuity correction

data:  A and B
W = 89, p-value = 0.007497
alternative hypothesis: true location shift is not equal to 0
```

We note 2 things: first, the warning about exact p-values and ties. Second, the value of the test statistic, which is not the sum of the ranks in the smaller sample as in Rice. What R uses, is the symmetric version of the test statistic.

Let T_Y denote the sum of the ranks of $Y_1, \ldots, Y_m$.

Theorem. *If $F = G$,*

$$\mathbb{E}(T_Y) = \frac{m(m+n+1)}{2}, \qquad \text{var}(T_Y) = \frac{mn(m+n+1)}{12}.$$

What R actually does is compute the rank sum for the *first* sample which would have expectation $n(m+n+1)/2$ and subtract $n(n+1)/2$:

```
R> sum(r_A) - n_A * (n_A + 1) / 2
```

```
[1] 89
```

which has expectation $mn/2$ which is symmetric in m and n. If the samples are interchanged, R would use

```
R> sum(r_B) - n_B * (n_B + 1) / 2
```

```
[1] 15
```

as can be verified by inspection:

```
R> wilcox.test(B, A)$statistic
```

```
 W
15
```

The Mann-Whitney (Wilcoxon rank sum) test can also be derived as follows. Consider estimating

$$\pi = \mathbb{P}(X < Y).$$

A natural estimate would be

$$\hat{\pi} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij}, \qquad Z_{ij} = 1_{X_i < Y_j}.$$

Now note that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij}, \qquad V_{ij} = 1_{X_{(i)} < Y_{(j)}}.$$

If the rank of $Y_{(j)}$ is denoted by $R_{Y,j}$, then the number of X less than $Y_{(j)}$ is $R_{Y,j} - j$ (in the case of no ties), hence

$$\sum_{i=1}^{n} \sum_{j=1}^{m} V_{ij} = \sum_{j=1}^{m} (R_{Y,j} - j) = T_Y - m(m+1)/2 = U_Y.$$

giving the symmetric version of the test statistic. Using this notation, R's W is really U_X. I.e., $\hat{\pi}$ is

```
R> wilcox.test(B, A)$statistic / (n_A * n_B)
```

```
        W
0.1442308
```

which is not the same as

```
R> mean(outer(A, B, `<`))
```

```
[1] 0.09615385
```

Note that:

```
R> c(sum(outer(A, B, `<`)), sum(outer(A, B, `<=`)))
```

```
[1] 10 20
```

Corollary. *If $F = G$,*

$$\mathbb{E}(U_Y) = \frac{mn}{2}, \qquad \mathrm{var}(U_Y) = \frac{mn(m+n+1)}{2}.$$

The Mann-Whitney test can be inverted to obtain confidence intervals for location shifts: consider the shift model $G(x) = F(x - \Delta)$. Then for testing the null that the shift parameter is Δ, we can use

$$U_Y(\Delta) = \#\{(i,j) : X_i - (Y_j - \Delta) < 0\} = \#\{(i,j) : Y_j - X_i > \Delta\}.$$

One can show that the distribution of $U_Y(\Delta)$ is symmetric about $mn/2$. A $100(1-\alpha)\%$ confidence interval for Δ is thus of the form

$$C = \{\Delta : k \leq U_Y(\Delta) \leq mn - k\}$$

which can be rewritten in terms of the ordered $X_i - Y_j$.

In R, confidence intervals are obtained via `conf.int = TRUE`:

```
R> wilcox.test(A, B, conf.int = TRUE)

        Wilcoxon rank sum test with continuity correction

data:  A and B
W = 89, p-value = 0.007497
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 0.01000082 0.07001754
sample estimates:
difference in location
            0.05008264
```

Bootstrap for the two-sample problem: suppose again that $\pi = \mathbb{P}(X < Y)$ is estimated by $\hat{\pi}$. How can the standard error of this be estimated? (Note that the confidence intervals are computed under the assumption that $F = G$.)

If F and G were known, we could generate bootstrap samples and compute $\hat{\pi}_1, \ldots, \hat{\pi}_B$ from these. As they are not known, one instead uses the empirical distributions F_n and G_m. I.e., one repeatedly randomly selects n values from the observed $X_1, \ldots, X_n$ with replacement, and m values from the observed $Y_1, \ldots, Y_m$, and calculates the resulting $\hat{\pi}$, generating a bootstrap sample $\hat{\pi}_1, \ldots, \hat{\pi}_B$.

4.2 Comparing Paired Samples

Often, samples are paired, e.g., by matching cases to controls and then randomly assigning to treatment and control groups, or by taking "before" and "after" measurements on the same object. Given pairing, the samples are no longer independent.

Denote pairs as (X_i, Y_i) where the X and Y have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, respectively. Assume that different pairs are independent with $\mathrm{cov}(X_i, Y_i) = \sigma_{XY} = \rho\sigma_X\sigma_Y$, where ρ is the correlation of X and Y. Then the differences $D_i = X_i - Y_i$ are independent with

$$\mathbb{E}(D_i) = \mu_X - \mu_Y, \qquad \mathrm{var}(D_i) = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y.$$

For the natural estimate $\bar{D} = \bar{X} - \bar{Y}$,

$$\mathbb{E}(\bar{D}) = \mu_X - \mu_Y, \qquad \mathrm{var}(\bar{D}) = \frac{1}{n}(\sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y).$$

Compare to the independent case: if $\rho > 0$, the variance of $\hat{D}$ is smaller. In general, if $\sigma_X^2 = \sigma_Y^2 = \sigma^2$,

$$\frac{\text{var}(\bar{D})}{\text{var}(\bar{X} - \bar{Y})} = \frac{2\sigma^2(1-\rho)/n}{2\sigma^2/n} = 1 - \rho.$$

4.2.1 Methods Based on the Normal Distribution

If the differences have a normal distribution with

$$\mathbb{E}(D_i) = \mu_D = \mu_X - \mu_Y, \qquad \text{var}(D_i) = \sigma_D^2,$$

with σ_D^2 typically unknown, inference will be based on

$$t = \frac{\bar{D} - \mu_D}{s_{\bar{D}}}$$

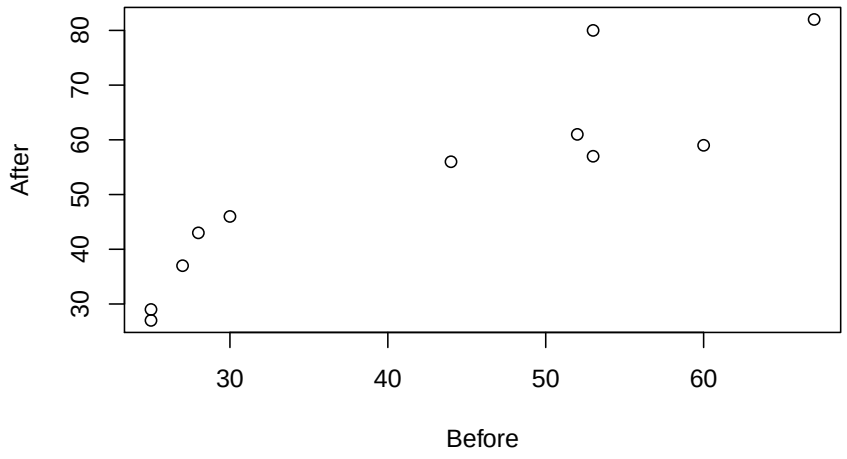
which follows a t distribution with $n - 1$ degrees of freedom.

Example: Smoking. Levine (1973) drew blood samples from 11 individuals before and after smoking and measured the extent to which the blood platelets aggregated.

```
R> platelet <- read.table("Data/platelet.txt", sep = ",", header = TRUE)
R> B <- platelet$before
R> A <- platelet$after
```

We can inspect the data via

```
R> plot(B, A, xlab = "Before", ylab = "After")
```



Note that

```
R> cor(B, A)
```

```
[1] 0.9012976
```

so pairing is quite efficient.

Inference can be performed "by hand", using

```
R> D <- B - A
R> t <- mean(D) / (sd(D) / sqrt(length(D)))
R> t

[1] -4.271609
```

or via `t.test` with argument `paired = TRUE`:

```
R> t.test(B, A, paired = TRUE)

        Paired t-test

data:  B and A
t = -4.2716, df = 10, p-value = 0.001633
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -15.63114  -4.91431
sample estimates:
mean of the differences
              -10.27273
```

4.2.2 A Nonparametric Method: The Signed Rank Test

The signed rank test (also known as the *Wilcoxon signed rank test* is based on a simple idea. Compute the differences $D_i = X_i - Y_i$, rank the absolute values of the D_i, and compute the sum of the ranks for which the differences are positive. In our example,

```
R> D <- B - A
R> R <- rank(abs(D))
R> sum(R[D > 0])

[1] 1
```

Intuitively, if there was no difference between the paired variables, about half of the D_i should be positive, and the signed rank sum should not be too extreme (small or large).

More precisely, consider the null hypothesis that the distribution of D is symmetric about 0. If this distribution is continuous, then under H_0, all sign combinations have equal probability $1/2^n$. The signed rank sum is then of the form

$$W_+ = \sum_{k=1}^{n} k I_k,$$

where I_k is the indicator that the k-th largest $|D_i|$ has positive sign. Under H_0, the I_k are i.i.d. Bernoulli with $p = 1/2$, so $\mathbb{E}(I_k) = 1/2$, $\mathrm{var}(I_k) = 1/4$,

$$\mathbb{E}(W_+) = \frac{1}{2}\sum_{k=1}^{n} k = \frac{n(n+1)}{4}, \qquad \mathrm{var}(W_+) = \frac{1}{4}\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{24}.$$

In particular,

$$\mathbb{P}(W_+ = 1) = \mathbb{P}(I_1 = 1, I_2 = \cdots = I_n = 0) = 1/2^n.$$

In our example

```
R> 1 / 2^(length(D))

[1] 0.0004882812
```

which is the same as

```
R> dsignrank(1, length(D))

[1] 0.0004882812
```

and rejecting if $|W_+|$ is large would have p-value

```
R> 2 * psignrank(1, length(D))

[1] 0.001953125
```

In case of ties (as in our case), things are a bit messier, and e.g. the normal approximation based on the above mean and variance is used. Compactly:

```
R> wilcox.test(B, A, paired = TRUE)

        Wilcoxon signed rank test with continuity correction

data:  B and A
V = 1, p-value = 0.005056
alternative hypothesis: true location shift is not equal to 0
```

One can also invert this test to obtain confidence intervals for the *pseudomedian*. The pseudomedian of a probability distribution F is the median of the distribution of $(U+V)/2$, where U and V are independent with distribution F. If F is symmetric, then median and pseudomedian coincide.

```
R> wilcox.test(B, A, paired = TRUE, conf.int = TRUE)

        Wilcoxon signed rank test with continuity correction

data:  B and A
V = 1, p-value = 0.005056
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -15.499990  -4.000002
sample estimates:
(pseudo)median
     -9.500014
```

5 Analysis of Categorical Data

5.1 Fisher's Exact Test

Rosen and Jordan (1974) experiment with male bank supervisors attending a management institute. In one experiment, supervisors were given a personnel file and had to decide whether to promote the employee or to hold the file and interview additional candidates. By random selection, 24 supervisors examined files labeled as from a male and 24 files labels as from a female employee; files were otherwise identical. Results were as follows.

	Male	Female
Promote	21	14
Hold File	3	10

Is there evidence for a gender bias?

Under the null of no bias, the observed "differences" would be due only to the random assignment of supervisors to files. We denote the counts in the table and the margins as follows:

N_{11}	N_{12}	$n_{1.}$
N_{21}	N_{22}	$n_{2.}$
$n_{.1}$	$n_{.2}$	$n_{..}$

According to the null hypothesis, the margins are *fixed*: the process of randomization determines the random fluctuation of the cell counts in the interior of the table subject to the constraints of the margin. With these constraints, there is in fact only 1 degree of freedom in the interior.

Consider the count N_{11}. Under H_0, this is distributed as the number of successes in 24 draws without replacement from a population of 35 successes and 13 failures, i.e., it has a hypergeometric distribution

$$\mathbb{P}(N_{11} = n_{11}) = \frac{\binom{n_{1.}}{n_{11}}\binom{n_{2.}}{n_{21}}}{\binom{n_{..}}{n_{.1}}}.$$

In our case, the number of successes must be between 11 and 24:

```
R> round(dhyper(11:24, 35, 13, 24), 4)

 [1] 0.0000 0.0003 0.0036 0.0206 0.0720 0.1620 0.2415 0.2415 0.1620 0.0720
[11] 0.0206 0.0036 0.0003 0.0000
```

The null would be rejected for small or large values of n_{11}, e.g., for significance level $\alpha = 0.05$:

```
R> round(phyper(11:24, 35, 13, 24), 4)

 [1] 0.0000 0.0003 0.0039 0.0245 0.0965 0.2585 0.5000 0.7415 0.9035 0.9755
[11] 0.9961 0.9997 1.0000 1.0000
```

suggests rejecting when $n_{11} \leq 14$ or $n_{11} \geq 21$. In our case, $n_{11} = 21$ so the null of no bias would be rejected at the 5% level.

This is *Fisher's exact test*. In R,

```
R> tab <- matrix(c(21, 14, 3, 10), nrow = 2, byrow = TRUE)
R> fisher.test(tab)

        Fisher's Exact Test for Count Data

data:  tab
p-value = 0.04899
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  1.00557 32.20580
sample estimates:
odds ratio
   4.83119
```

Note that this is formulated in terms of the odds ratio corresponding to the table (but that the sample estimate is not the sample odds ratio).

R also provides suitable generalizations of Fisher's exact test to general $r \times c$ contingency tables. Alternatively, function `r2dtable` can be used for efficient generation of tables with given row and column margins, and hence for bootstrap versions of the test for independence of rows and columns given the margins.

5.2 The Chi-Squared Test of Homogeneity

Suppose we have independent observations from J multinomial distributions with I cells each, and want to test the null that the cell probabilities of the multinomials are equal—i.e., test the homogeneity of the multinomial distributions.

Consider the following example from stylometry given in Morton (1978). When Jane Austen died, she left the novel *Sandition* only partially completed. A highly literate admirer finished the work based on the summary of the remainder, attempting to emulate Austen's style. The following table gives word counts obtained by Morton for Chapters from *Sense and Sensibility*, *Emma*, and *Sandition* written by Austen (Sandition I) and her admirer (Sandition II):

Word	*Sense and Sensibility*	*Emma*	*Sandition I*	*Sandition II*
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4
Total	375	440	202	196

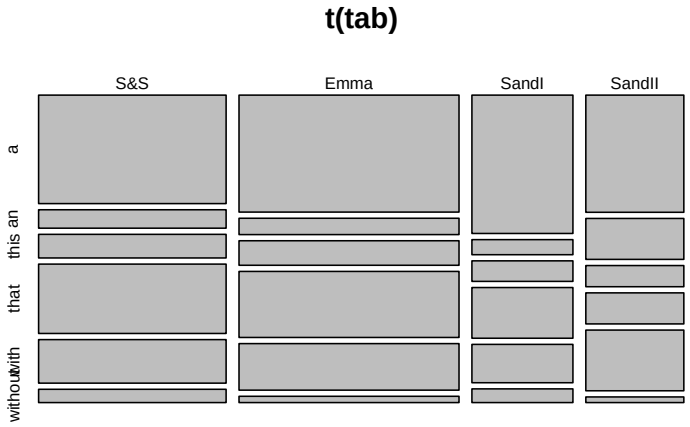
```
R> tab <-
+     matrix(c(147, 186, 101,  83,
+              25,  26,  11,  29,
+              32,  39,  15,  15,
+              94, 105,  37,  22,
+              59,  74,  28,  43,
+              18,  10,  10,   4),
+            ncol = 4, byrow = TRUE)
R> rownames(tab) <- c("a", "an", "this", "that", "with", "without")
```

```
R> colnames(tab) <- c("S&S", "Emma", "SandI", "SandII")
R> tab
```

```
        S&S Emma SandI SandII
a       147  186   101     83
an       25   26    11     29
this     32   39    15     15
that     94  105    37     22
with     59   74    28     43
without  18   10    10      4
```

A visual comparison of the frequencies:

```
R> mosaicplot(t(tab))
```



We will use the following stochastic model: the counts for *Sense and Sensibility* will be modeled as a multinomial random variable with unknown cell probabilities and total count 375. Write π_{ij} for the probability of category i in multinomial j. Then the null hypothesis is

$$H_0 : \pi_{i1} = \cdots = \pi_{iJ}, \qquad i = 1, \ldots, I.$$

Write n_{ij} for the observed cell counts.

Theorem. *Under H_0 and independent multinomial sampling, the mle's of the common cell probabilities π_i are*

$$\hat{\pi}_i = n_{i\cdot}/n_{\cdot\cdot}.$$

Proof. By independence,

$$\text{lik}(\pi_1, \ldots, \pi_I) = \prod_{j=1}^{J} \binom{n_{\cdot j}}{n_{1j} \cdots n_{Ij}} \pi_1^{n_{1j}} \cdots \pi_I^{n_{Ij}} = \pi_1^{n_{1\cdot}} \cdots \pi_I^{n_{I\cdot}} \prod_{j=1}^{J} \binom{n_{\cdot j}}{n_{1j} \cdots n_{Ij}}$$

To maximize, we use the Lagrangian

$$L(\pi_1, \ldots, \pi_I, \lambda) = \sum_{i=1}^{I} n_{i.} \log(\pi_i) + \lambda \left(\sum_{i=1}^{I} \pi_i - 1 \right)$$

for which

$$\frac{\partial L}{\partial \pi_i} = \frac{n_{i.}}{\pi_i} + \lambda$$

from which $\pi_i = -n_{i.}/\lambda$ and thus the assertion by using the constraint.
For multinomial j, the expected counts (under the null) are

$$E_{ij} = \hat{\pi}_i n_{.j} = \frac{n_{i.} n_{.j}}{n_{..}}$$

and Pearson's chi-squared statistic is therefore

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - n_{i.} n_{.j}/n_{..})^2}{n_{i.} n_{.j}/n_{..}}$$

For large sample size, this approximately has a χ^2 distribution with degrees of freedom

$$J(I-1) - (I-1) = (I-1)(J-1)$$

as each multinomial has $I - 1$ parameters, and under the null $I - 1$ parameters are estimated.
Using R, we first compare the frequencies for Austen's writings.

```
R> chisq.test(tab[, 1 : 3])
```

```
        Pearson's Chi-squared test

data:  tab[, 1:3]
X-squared = 12.271, df = 10, p-value = 0.2673
```

Observed and expected counts can be obtained as follows:

```
R> cst <- chisq.test(tab[, 1 : 3])
R> sprintf("%.0f/%.1f", cst$observed, cst$expected)
```

```
 [1] "147/160.0" "25/22.9"   "32/31.7"   "94/87.0"   "59/59.4"   "18/14.0"
 [7] "186/187.8" "26/26.8"   "39/37.2"   "105/102.1" "74/69.7"   "10/16.4"
[13] "101/86.2"  "11/12.3"   "15/17.1"   "37/46.9"   "28/32.0"   "10/7.5"
```

Next, we compare the aggregated Austen counts to the imitator:

```
R> cst <- chisq.test(cbind(Aus = rowSums(tab[, 1:3]), Imi = tab[, 4]))
R> cst
```

```
        Pearson's Chi-squared test

data:  cbind(Aus = rowSums(tab[, 1:3]), Imi = tab[, 4])
X-squared = 32.81, df = 5, p-value = 4.106e-06
```

So the imitator was significantly unsuccessful. To see why, one could consider

```
R> sprintf("%.0f/%.1f", cst$observed, cst$expected)

 [1] "434/433.5" "62/76.3"   "86/84.7"   "236/216.3" "161/171.0" "38/35.2"
 [7] "83/83.5"   "29/14.7"   "15/16.3"   "22/41.7"   "43/33.0"   "4/6.8"
```

but it is somewhat better to look at the contributions to the test statistic (the squared so-called Pearson residuals):

```
R> round(cst$residuals ^ 2, 2)

          Aus   Imi
a        0.00  0.00
an       2.68 13.90
this     0.02  0.11
that     1.79  9.30
with     0.59  3.06
without  0.22  1.14
```

Looking at the residuals

```
R> round(cst$residuals, 2)

           Aus   Imi
a         0.03 -0.06
an       -1.64  3.73
this      0.14 -0.33
that      1.34 -3.05
with     -0.77  1.75
without   0.47 -1.07
```

we see that Austen used *an* much less and *that* much more frequently than her imitator.

5.3 The Chi-Squared Test of Independence

In a demographic study of women listed in *Who's Who*, Kiser and Schaefer (1949) compiled the following table fir 1346 women who were married at least once:

	Married once	Married more	Total
College	550	61	611
No College	681	144	825
Total	1231	205	1436

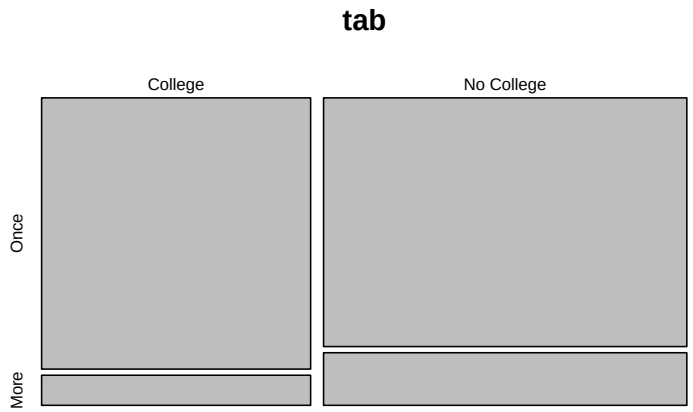
Is there a relationship between marital status and level of education?

```
R> tab <- matrix(c(550, 61, 681, 144), nrow = 2, byrow = TRUE)
R> rownames(tab) <- c("College", "No College")
R> colnames(tab) <- c("Once", "More")
R> tab
```

```
           Once More
College     550   61
No College  681  144
```

A mosaic plot:

```
R> mosaicplot(tab)
```



We model the data as coming from a sample of size n cross-classified in a table with I rows and J columns, a *contingency table*, with the joint distribution of the cell counts n_{ij} a multinomial with cell probabilities π_{ij}. (Note the difference to the previous section.)

If the row and column classifications are independent,

$$\pi_{ij} = \pi_{i.}\pi_{.j}.$$

We thus consider testing

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}, \qquad i = 1, \ldots, I, \quad j = 1, \ldots, J$$

versus the alternatives that the π_{ij} are free (apart from being non-negative with sum one). Under H_0, the mle's are

$$\hat{\pi}_{ij} = \hat{\pi}_{i.}\hat{\pi}_{.j} = \frac{n_{i.}}{n}\frac{n_{.j}}{n};$$

under H_A, the mle's are simply

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}.$$

These mle's can be used to form an LRT or the asymptotically equivalent Pearson's chi-squared test

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where

$$E_{ij} = n\hat{\pi}_{ij} = \frac{n_{i.}n_{.j}}{n}$$

are the counts expected under the null, giving

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i.}n_{.j}/n_{..})^2}{n_{i.}n_{.j}/n_{..}}$$

(again!). The degrees of freedom under H_0 and H_A are $(I-1)+(J-1)$ and $IJ-1$, respectively, giving

$$(IJ-1) - ((I-1)+(J-1)) = IJ - I - J + 1 = (I-1)(J-1)$$

degrees of freedom.

Note that the chi-squared statistics for homogeneity and independence are identical in form and degrees of freedom: however, the underlying hypotheses and sampling schemes are different. (Consider performing bootstrap variants of the tests instead.)

For the marriage data,

```
R> chisq.test(tab)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab
X-squared = 15.405, df = 1, p-value = 8.675e-05
```

Note that to reproduce the results in Rice one needs

```
R> chisq.test(tab, correct = FALSE)

        Pearson's Chi-squared test

data:  tab
X-squared = 16.01, df = 1, p-value = 6.302e-05
```

5.4 Matched-Pairs Designs

Does the use of cell phones while driving cause accidents? This is hard to study empirically (if usage is hazardous, it would be unethical to deliberately expose drivers to risk, etc.). Redelmaier and Tibshirani (1997) conducted the following clever study. 699 drivers who owned cell phones and had been involved in motor vehicle collisions were identified. Then, billing records were used to determine whether each individual used a cell phone during the 10 minutes preceding the collision and also at the same time during the previous week. Hence, each person serves as its own control. Results were as follows:

	Before Collision		
Collision	On Phone	Not On Phone	Total
On Phone	13	157	170
Not On Phone	24	505	529
Total	37	662	699

We can model the data as a sample of size 699 from a multinomial distribution with four cells and respective cell probabilities π_{ij}. The null hypothesis is that of marginal symmetry (distributions are the same at collision and before collision):

$$\pi_{11} + \pi_{21} = \pi_{.1} = \pi_{1.} = \pi_{11} + \pi_{12}, \qquad \pi_{12} + \pi_{22} = \pi_{.2} = \pi_{2.} = \pi_{21} + \pi_{22},$$

or equivalently,

$$H_0 : \pi_{12} = \pi_{21}.$$

The mle's of the relevant cell probabilities under H_0 are

$$\hat{\pi}_{12} = \hat{\pi}_{21} = \frac{n_{12} + n_{21}}{2n}$$

whereas under H_A

$$\hat{\pi}_{12} = \frac{n_{12}}{n}, \qquad \hat{\pi}_{21} = \frac{n_{21}}{n}$$

resulting in a chi-squared LRT approximation of

$$X^2 = \frac{(n_{12} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} + \frac{(n_{21} - (n_{12} + n_{21})/2)^2}{(n_{12} + n_{21})/2} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

with $2 - 1 = 1$ degree of freedom. This is the *McNemar test*.

By hand,

```
R> tab <- matrix(c(13, 157, 24, 505), nrow = 2, byrow = TRUE)
R> n12 <- tab[1, 2]
R> n21 <- tab[2, 1]
R> Xsq <- (n12 - n21)^2 / (n12 + n21)
R> Xsq

[1] 97.72928
```

or using a built-in classical test function:

```
R> mcnemar.test(tab)

        McNemar's Chi-squared test with continuity correction

data:  tab
McNemar's chi-squared = 96.265, df = 1, p-value < 2.2e-16
```

5.5 Summary

Observe the correspondences of the classical test problems for metric and categorical data:

- Testing the null $F_1 = \cdots = F_K$ that the distributions of J independent samples are the same: the *K-sample problem*. We covered $K = 2$ for numeric variables (t test for independent samples; Mann-Whitney aka Wilcoxon rank sum test) and the general case (e.g., chi-squared test for homogeneity) for categorical data.
- For pairs (X, Y) of observations, test the null of independence of X and Y, the so-called *independence problem* (or *contingency problem*). We covered only the categorical case (e.g., chi-squared test for independence); see e.g. `cor.test` for variants for numeric data.
- For pairs (X, Y) of (not necessarily independent) observations, test the null $F_X = F_Y$: the *symmetry problem*. We covered both the numeric (t test for paired samples; (Wilcoxon) signed rank test) and the categorical case (McNemar test).

Observe also that in many cases, there are modern conditional (permutation based) tests as (preferable) alternatives to the classical unconditional tests.