Tools for dealing with high-dimensional data

Workshop on Polygenic Adaption
ICTS, Bangalore
6 – 17 May 2024

Bruce Walsh
University of Arizona
jbwalsh@arizona.edu

Background, Additional reading

- WVL (Walsh, Visscher, Lynch) 2024.
  - Appendix 6: Multiple comparisons, meta-analysis
  - Chapter 21 (sections on gene– and pathway-based tests)
- Walsh and Lynch 2018
  - Chapter 27 (EVT, pp 1003-1005)

Motivation

- Genomics generates high-dimensional data sets
- Serious concerns about
  - Corrections for Multiple comparisons
  - Combining p values over different sets (such as using all of the SNPs within a gene for a gene-based test)
  - Combining estimates of effect sizes over multiple studies (meta-analysis)
  - Model selection with high-dimensional data

Overview

- Control of significance tests over multiple comparisons
  - Are there an excess number of false positives?
  - Bonferroni and sequential Bonferroni (Simes, etc)
  - Effective number of tests
  - False discovery rates (FDR)
  - Bayesian approaches
  - Permutation tests and EVT
- Combining p values from multiple sources
  - Fisher, Schaffer, Tippett, Simes, Cauchy
- Meta-analysis
- Fitting high-dimensional models
  - Penalized regressions
  - Model selection, AIC, BIC,

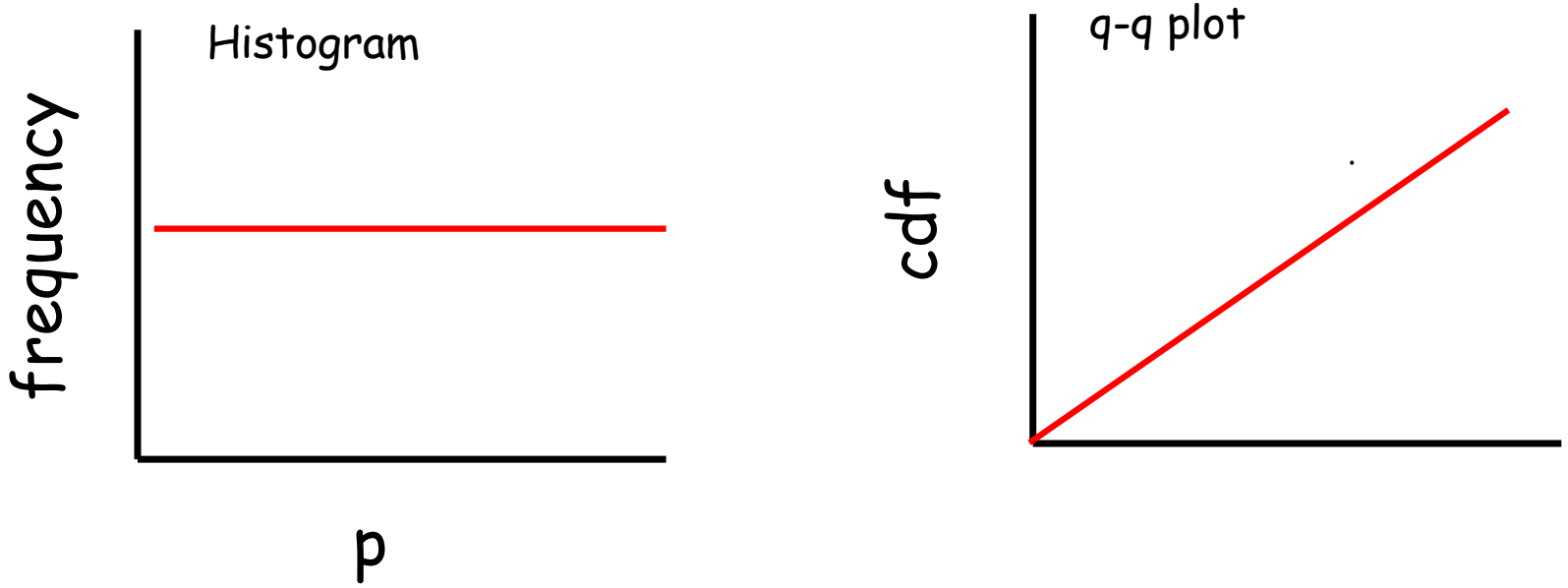
- Significance = prob of a false positive
  - Declaring a test statistic to be significant when it is actually from the null (H_0)
- Test-level significance (error rare)
  - p = 0.05 implies that only 5% of the time (under the null) would one see a test statistic this extreme
  - Also called the comparison-wide error rate (CWER)
- Experiment-wide significance (error rare)
  - γ = 0.05 means a 5% chance that *any* of the tested hypotheses in the set (experiment) would have a test statistic that extreme under the null (H_0)
  - Also called the family-wide error rate (FWER)
- Discoveries = Tests declared to be significant
  - could be true (from H_1) , or false (from H_0)

The problem

- Suppose 100 hypotheses are all tested at p = 0.05
- Expected 100*0.05 = 5 false positives even if all under the null
- Hence, have adjust the critical p value, τ, for each test (CWER) *downward* to control for the overall probability (or number) of false positives over the family of tests (FWER)

Key idea

- If all tests are from the null, then
  - the distribution of p values is a uniform over 0,1
  - The cdf (q-q) plot is linear



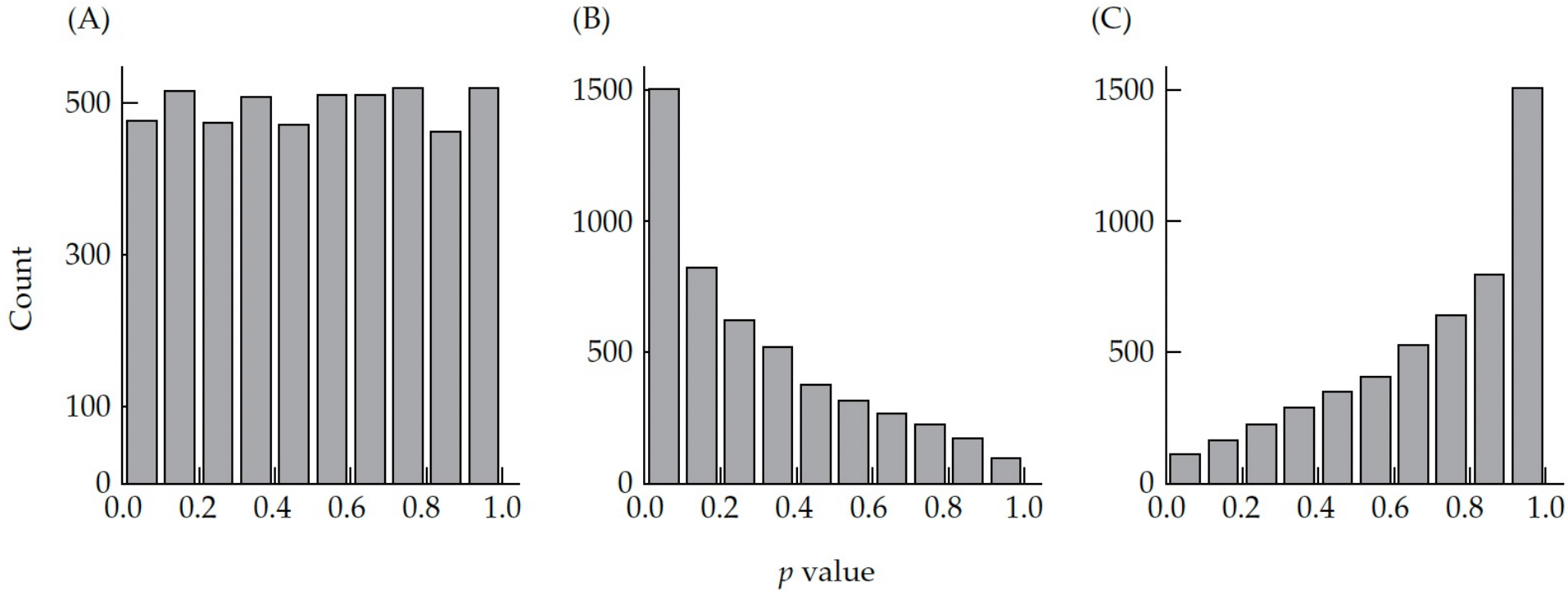


Figure A6.2 Simulated distribution of p values based on 5000 tests for samples of 25 draws from a normal distribution with a mean of μ and a variance of one. The null hypothesis is $H_0 : \mu \leq 0$. **A:** The distribution of p values when $\mu = 0$ (the null is correct) is uniform. **B:** The distribution when $\mu = 0.2$ is skewed toward an excess of values near zero. **C:** The distribution when $\mu = -0.2$ is skewed toward an excess of values near one.

Part I:
Number of false positives over a family of tests

How Many False Positives?

Suppose we perform n independent tests, each with a Type-I error rate of α. If all hypotheses are truly null, the number, j, of false positives follows a binomial distribution (Wilkinson 1951), with a "success" probability (a false positive) of α, and n trials (the number of tests), yielding

$$\Pr(j \text{ false positives}) = \frac{n!}{(n-j)!\,j!}(1-\alpha)^{n-j}\,\alpha^{j} \tag{A6.7a}$$

For n large and α small, this is closely approximated by the Poisson distribution (Equation 2.21a), with Poisson parameter $n\alpha$ (the expected number of false positives), yielding

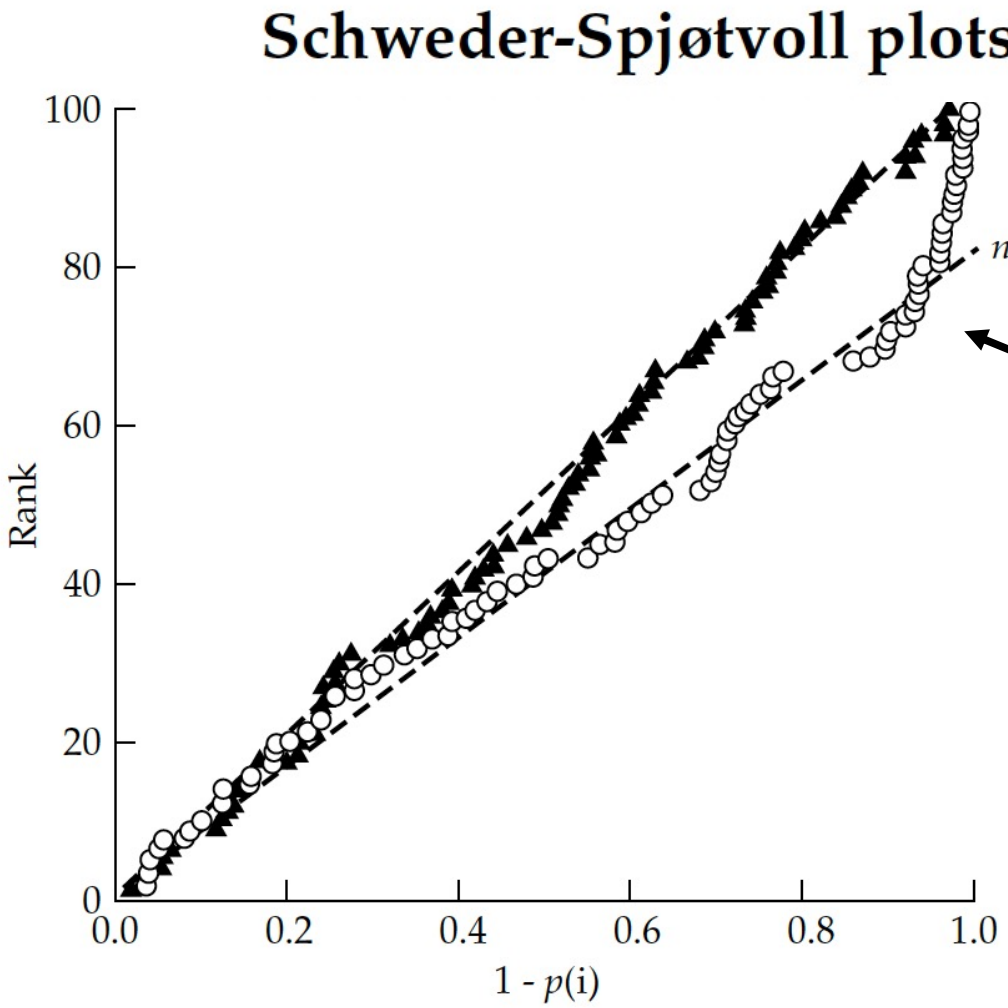
$$\Pr(j \text{ false positives}) \simeq \frac{(n\alpha)^{j}e^{-n\alpha}}{j!} \tag{A6.7b}$$

Example A6.5. Suppose 250 independent tests are performed, each with $\alpha = 0.025$ (a 2.5% chance of declaring a result from the null hypothesis to be significant), and 15 tests are declared significant by this criteria. Is this number greater than expected by chance? The expected number of significant tests under the global null hypothesis is $n\alpha = 250 \cdot 0.025 = 6.25$. From Equation A6.7a, the probability of observing 15 (or more) significant tests is

$$\sum_{j=15}^{250} \Pr(j \text{ false positives}) = \sum_{j=15}^{250} \frac{250!}{(250-j)!\,j!}(1-0.025)^{250-j}\,0.025^{j}$$

We could either sum this series directly or use the cumulative distribution function for a binomial. In `R`, the probability that a binomial with parameters n and p has a value of i or less is obtained by using the command `pbinom(i,n,p)`. The probability of 15 or greater is one minus the probability of 14 or less, or `1- pbinom(14,250,0.025)`, for which `R` returns 0.0018. A similar calculation can use the Poisson approximation (Equation A6.7b), with `1-ppois(14,6.25)` returing a value of 0.0021. Given that there is only a 0.2% chance of seeing this many significant tests under the global null, we expect that some of these significant tests are true **discoveries** (those whose associated null hypothesis is incorrect), not false positives. The critical question, of course, is which ones?

Given an excess number of false positives, we can estimate the number of true discoveries. Typically, want to estimate n_0 = # of nulls, or the fraction of nulls. $\pi_0 = n_0 / n$



All 100 are nulls

80 of 100 are nulls

Could use the Chow Test (Chapter 10) to determine where the regression break occurs

Figure A6.1 A Schweder-Spjøtvoll plot is one approach for detecting departures from a uniform distribution of p values. The p values are ordered from smallest, $p(1)$, to largest, $p(n)$, and one plots the rank of $1 - p(i)$ versus its value. These ranks are reversed from the ranks of $p(i)$, as the rank of $1 - p(n)$, being the smallest value, is 1. Under a uniform, the result is a straight line passing through the origin and the point $(1, n)$. The upper curve (solid triangles), generated by randomly sampling $n = 100$ values from a uniform (0,1), fits this pattern. The lower curve (open circles), generated by simulating p values for 80 true nulls and 20 tests where the alternative was correct, shows an inflation of p values near zero ($1-p$ values near one). This results in a strong departure from linearity near one. Ignoring this upturn and extrapolating the linear fit for the values below this inflection point gives an approximate value of 80 for the value of this projected line when $1 - p = 1$. This yields an estimate of n_0.

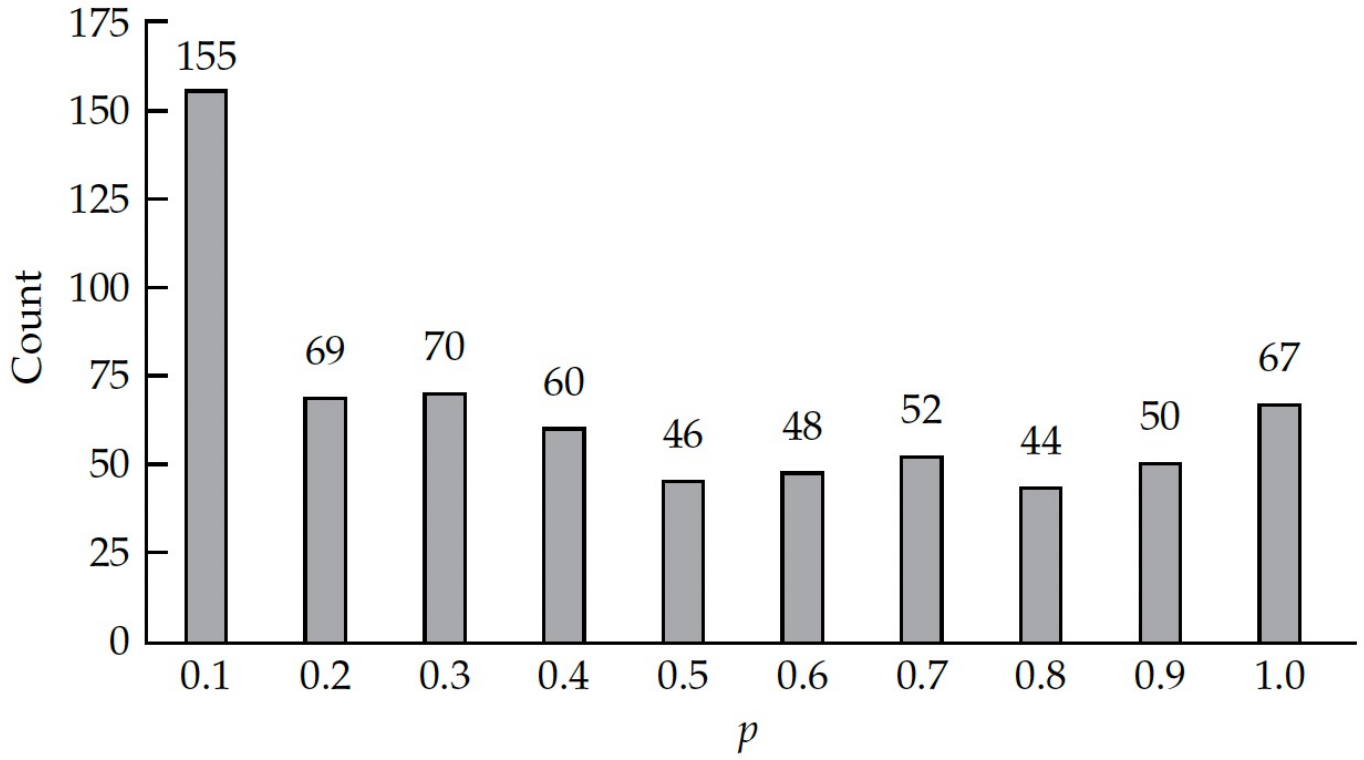


Figure A6.3 An empirical distribution of p values (for $n = 644$ tests) from Mosig et al. (2001). The number of p values in each of ten bins (of width 0.1) are given above the bars. Note the large excess of values near zero.

If there are n_0 truly null tests, then the expected number of p values from these tests that fall within an interval $0 \le a < b \le 1$ is simply

$$n_0 \int_a^b \phi_u(p)\, dp = n_0 \int_a^b 1 \cdot dp = n_0(b-a) \tag{A6.8b}$$

Hence

$$\widehat{n}_0(a,b) = \frac{\text{Number of } p(i) \text{ values in } (a,b)}{b-a} \tag{A6.8c}$$

Likewise, an estimate for the fraction $\pi_0 = n_0/n$ of true nulls is

$$\begin{aligned} \widehat{\pi}_0(a,b) &= \frac{\text{Number of } p(i) \text{ values in } (a,b)}{n(b-a)} \\ &= \frac{\text{Fraction of } p(i) \text{ values in } (a,b)}{b-a} \end{aligned} \tag{A4.8d}$$

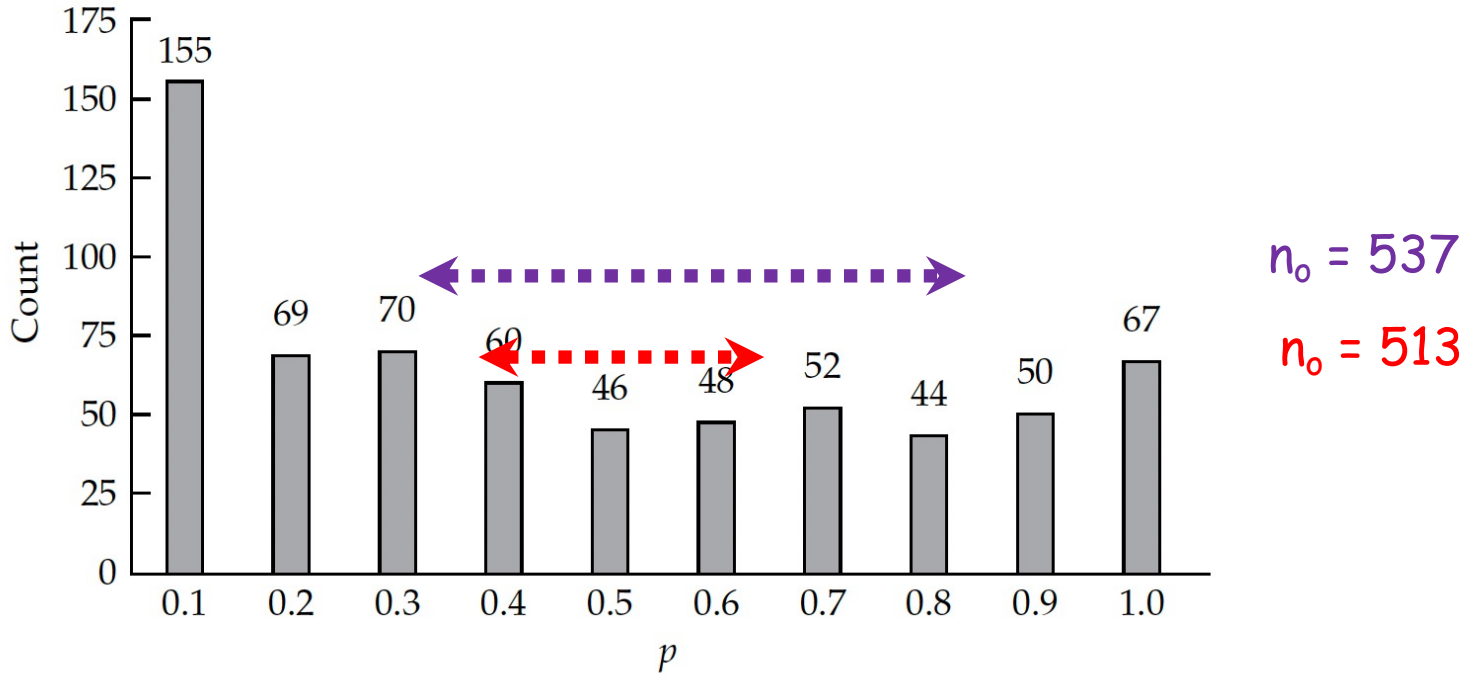


Figure A6.3 An empirical distribution of p values (for $n = 644$ tests) from Mosig et al. (2001). The number of p values in each of ten bins (of width 0.1) are given above the bars. Note

Example A6.6. What is an estimate of n_o given the data in Figure A6.3? Consider the bins centered around $p = 0.5$. Based on the central three bins (0.4, 0.5, and 0.6), a total of $60 + 46 + 48 = 154$ tests have p values in this interval. From Equation A6.8b, $154 = n_0 \cdot 0.3$ or $n_0 = 154/0.3 = 513$, and hence a fraction, $\pi_0 = n_0/n = 513/644 = 0.80$, of the tests are true nulls. Using the bins from 0.3 to 0.8 yields $n_0 = 322/0.6 = 537$, or $\pi_0 = 537/644 = 0.83$. Hence, it appears that around 80% of the tests are consistent with true nulls. Mosig et al. (2001; also see Nettleton et al. 2006) used an iterative approach (also based on bin counts in the p-value histogram) and arrived at an estimate of $n_0 = 500$ (78%).

Storey and Tibshirani (2003) proposed an estimator of n_o based the number of p values exceeding some tunable parameter value, λ (taking $a = \lambda$ and $b = 1$ in Equation A6.8b), on the logic that for larger values of λ, most of these draws are from the uniform corresponding to draws from the null. Let $\widehat{\pi}_0(\lambda)$ denote the estimated fraction of truly null hypotheses based on using a tuning value of λ, then

$$\widehat{\pi}_0(\lambda) = \frac{\text{Number of } p(i) \text{ values } > \lambda}{n(1-\lambda)} \tag{A6.9a}$$

and

$$\widehat{n}_0(\lambda) = n \cdot \widehat{\pi}_0(\lambda) = \frac{\text{Number of } p(i) \text{ values } > \lambda}{1-\lambda} \tag{A6.9b}$$

By focusing on the interval $(\lambda, 1)$, the **Storey-Tibshirani estimator** is potentially biased

The beta distribution

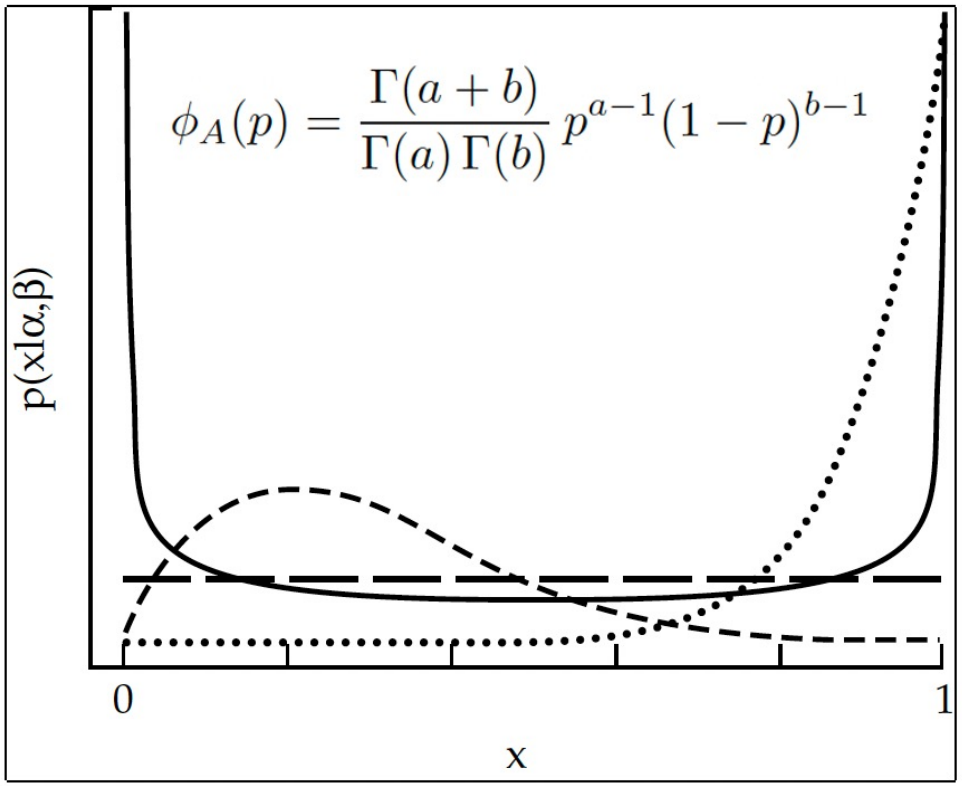


Figure A7.3 For $\alpha = \beta = 1$ (long-dashed curve), the beta distribution is simply the uniform distribution over $(0, 1)$. The pdf for the beta distribution can also be U-shaped ($\alpha = \beta = 0.5$; solid curve), unimodal ($\alpha = 2, \beta = 5$; short-dashed curve), or L-shaped ($\alpha = 10, \beta = 1$; dotted curve). Because the beta distribution is symmetric in α and β, switching their parameter values generates a distribution of the same shape translated about 0.5.

Estimating n_0: Mixture Models

Allison et al. (2002) suggested that π_0 can be estimated by treating the distribution of p values as a mixture, a fraction π_0 of which comes from a uniform (and hence a uniform distribution function, ϕ_u), while the remainder $(1 - \pi_0)$ are from the distribution, $\phi_A(p)$, of p values when the alternative hypothesis is true (Figure A6.4). While the general form of $\phi_A(p)$ is unknown, a very flexible modeling approach is to assume a beta distribution (Appendix 7; Figure A7.3)

$$\phi_A(p) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, p^{a-1}(1-p)^{b-1} \tag{A6.10a}$$

Under the alternative hypothesis, we expect an increase in p values near zero, which occurs when $a < 1$. Likewise, the beta distribution can easily accommodate an increase in p values near one ($b < 1$). When $a = b = 1$, this simply reduces to a uniform.

$$\ell(p) = (1-\pi_0)\,\phi_A(p) + \pi_0\,\phi_u(p) = (1-\pi_0)\frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, p^{a-1}(1-p)^{b-1} + \pi_0 \tag{A6.10b}$$

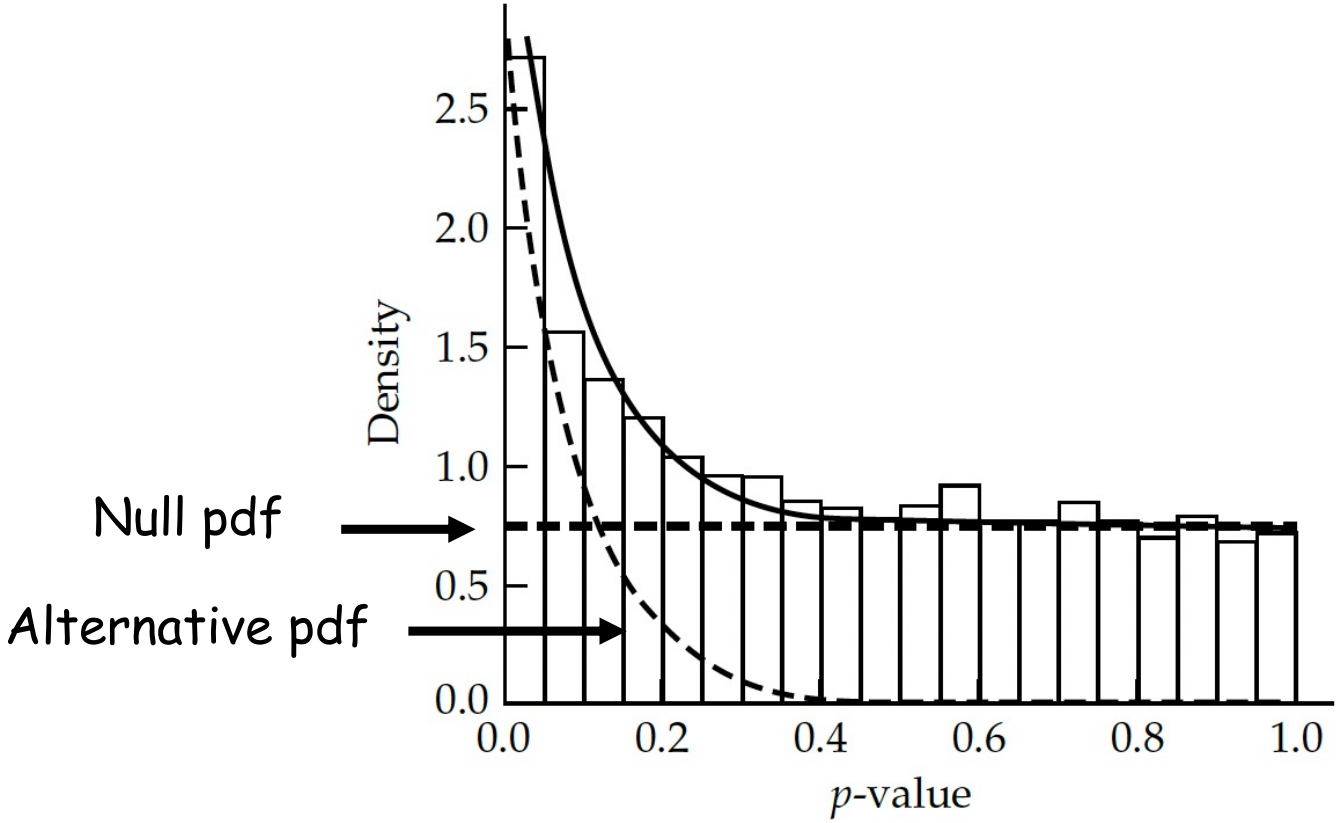


Figure A6.4 The empirical distribution of p values can be treated as a mixture model of a uniform plus a beta distribution (whose shape parameters, a and b, can be estimated via ML), see Equation A6.10b. In this hypothetical example, a weighted mixture of a uniform (horizontal dashed line) and a beta with ($a < 1, b = 1$; dashed curve), yields the mixture distribution (solid curve) that fits the empirical distribution of the p values.

Part II
Controlling the FWER

Key topics

- Bonferroni corrections
  - Standard
  - Sequential (Holm's, Simes, Hommel's)
  - Correcting for effective number of tests
- FDR – the false discovery rate
- Bayesian framework
- Permutation tests and extreme value theory (EVT)
  - Using the trinity theorem to obtain small p values

Standard Bonferroni Corrections

The probability of not making any Type-I errors (false positives) over n independent tests, each at level α, is $(1-\alpha)^n$. Hence, the probability, π, of having at least one false positive over the entire collection is simply one minus this value,

$$\pi = 1 - (1-\alpha)^n \tag{A6.3a}$$

If we solve for the α value required for each test,

$$\alpha = 1 - (1-\pi)^{1/n} \tag{A6.3b}$$

This is often called the **Dunn-Ŝidák method**. If we note that $(1-\alpha)^n \simeq 1 - n\alpha$, we obtain the **Bonferroni method** by taking

$$\alpha = \pi/n \tag{A6.3c}$$

Both Equations A6.3b and A6.3c are referred to as **Bonferroni corrections**. In the literature, π is the **family-wide error rate** (**FWER**; also the **genome-wide error rate**, **GWER**), while α is the **comparison-wise error rate** (**CWER**), also referred to as the **point-wise significance level** (**PWSL**).

Sequential Bonferroni Corrections

Under a strict Bonferroni correction, only those tests whose associated p values are $\leq \pi/n$ are rejected (**declared significant**); all others are **accepted** (or more formally, **fail to be rejected**). This results in a considerable reduction in power if two or more of the hypotheses are actually false. When we reject a hypothesis, one fewer test remains, and the multiple comparison correction should reflect this, resulting in **sequential Bonferroni corrections**. Sequential approaches have increased power compared to standard Bonferroni corrections, as illustrated below in Example A6.4. Shaffer (1995) reviewed these, and other, approaches. The basic structure is that one has a collection of multiple tests, with $H(i)$ denoting the null hypothesis for test i — for example, the test that marker i has a nonzero effect, in which case $H(i)$ is the null hypothesis of no effect. In this case, rejecting $H(i)$ suggests evidence for a nonzero effect for marker i.

Example A6.4. Suppose for $n = 10$ tests, the (ordered) p values are as follows:

i	1	2	3	4	5	6	7	8	9	10
$p(i)$	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{\pi}{n-i+1}$	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500

p = 0.05/10 = 0.005

Holm's Method

The simplest of these sequential adjustments is **Holm's method** (Holm 1979). The first step is to order the p values for the n hypotheses being tested from smallest to largest, $p(1) \leq p(2) \leq \cdots \leq p(n)$, and let $H(i)$ be the hypothesis associated with $p(i)$. One proceeds with Holm's method as follows:

(i) If $p(1) > \pi/n$, accept all n null hypotheses (i.e., none are declared significant).

(ii) If $p(1) \leq \pi/n$, reject $H(1)$ [i.e., $H(1)$ is declared significant], and consider $H(2)$.

(iii) If $p(2) > \pi/(n-1)$, accept $H(i)$ (for $i \geq 2$).

(iv) If $p(2) \leq \pi/(n-1)$, reject $H(2)$ and move onto $H(3)$.

(v) Proceed with rejecting hypotheses until reaching the first i such that $p(i) > \pi/(n-i+1)$.

We can also apply Holm's method using Equation A6.3a — namely, $\alpha = 1 - (1-\pi)^{1/n}$, the Dunn-Ŝidák correction — in place of $\alpha = \pi/n$.

Example A6.4. Suppose for $n = 10$ tests, the (ordered) p values are as follows:

i	1	2	3	4	5	6	7	8	9	10
$p(i)$	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{\pi}{n-i+1}$	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500

Simes-Hochberg Method

With Holm's method, we stop once we fail to reject a hypothesis. An improvement on this approach is the **Simes-Hochberg correction** (Simes 1986; Hochberg 1988), which effectively starts backward, working with the largest p values first.

(i) If $p(n) \leq \pi$, then all hypothesis are rejected.

(ii) If not, $H(n)$ cannot be rejected, and we next examine $H(n-1)$.

(iii) If $p(n-1) \leq \pi/2$, then all $H(i)$ for $i \leq n-1$ are rejected.

(iv) If not, $H(n-1)$ cannot be rejected, and we compare $p(n-2)$ with $\pi/3$.

(v) In general, if $p(n-i) \leq \pi/(n-i+1)$, then all $H(i)$ for $i \leq n-i$ are rejected.

Example A6.4. Suppose for $n = 10$ tests, the (ordered) p values are as follows:

i	1	2	3	4	5	6	7	8	9	10
$p(i)$	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{\pi}{n-i+1}$	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500

Hommel's Method

Hommel's method (1988) is slightly more complicated, but it is more powerful than the Simes-Hochberg correction (Hommel 1989). Under Hommel's method, we reject all hypotheses whose p values are less than or equal to π/k^*, where

$$k^* = \max_i \left(p(n-i+j) > \pi \frac{j}{i} \right) \quad \text{for all } j = 1, \cdots, i$$

Example A6.4. Suppose for $n = 10$ tests, the (ordered) p values are as follows:

i	1	2	3	4	5	6	7	8	9	10
$p(i)$	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$\frac{\pi}{n-i+1}$	0.0050	0.0056	0.0063	0.0071	0.0083	0.0100	0.0125	0.0167	0.0250	0.0500

Dealing with Dependency: Eigenvalue-based Approaches

When different tests share correlated data, this introduces dependency between the p values for these tests. How do we account for this? One approach (Cheverud 2001; Nyholt 2004; Li and Ji 2005; Patterson et al. 2006) is to use the nature of the dependency structure of the data to estimate an **effective number of independent tests**, n_e. This value is then substituted for n in the above methods; e.g., Equation A6.3b becomes $\alpha = 1 - (1 - \pi)^{1/n_e}$, and n_e replaces n in both the Tippett (Equation A6.4a) and Simes (Equation A6.4b) statistics. A classic application of this approach is correcting for correlations among tests of marker-trait associations over a set of linked markers in either a QTL mapping experiment or a GWAS (Chapters 18–20).

To proceed, we need to introduce a few facts about the eigenstructure of a *correlation matrix*, $\mathbf{C}$, whose eigenvalues are denoted (from largest to smallest) by $\lambda_1, \cdots, \lambda_n$. First, because $\mathbf{C}$ is a positive-semidefinite matrix, all $\lambda_i \geq 0$ (WL Appendix 5). Second, $\mathbf{C}$ is an $n \times n$ matrix with ones on its diagonal, which makes its trace (the sum of its diagonal elements; Chapter 9) equal to a value of n. The importance of this result is that the trace of a matrix equals the sum of its eigenvalues (Equation 9.34b), which demonstrates that the average eigenvalue of $\mathbf{C}$ is

$$n^{-1} \sum_{i=1}^{n} \lambda_i = n^{-1} n = 1$$

$$n^{-1}\sum_{i=1}^{n}\lambda_i = n^{-1}n = 1$$

When all of the underlying variables that generate **C** are uncorrelated, then $\lambda_1 = \cdots = \lambda_n = 1$, while when all of the observations are completely correlated, then $\lambda_1 = n$ and $\lambda_2 = \cdots = \lambda_n = 0$. These two cases represent the extremes of n independent tests (the former) and one independent test (the latter). As with principal components (Chapter 9), the spread of the eigenvalues tells us about dependency. One metric of this is the variance in the eigenvalues, $\sigma^2(\lambda)$. If all of the eigenvalues are equal, then $\sigma^2(\lambda) = 0$, while if only one eigenvalue is nonzero, then $\sigma^2(\lambda) = n$.

Motivated by the above eigenstructure observations, **Cheverud's method** (2001) computes the effective number of independent tests as

$$n_{e,Cheverud} = n\left(1 - \frac{(n-1)\sigma^2(\lambda)}{n^2}\right), \qquad \text{where} \qquad \sigma^2(\lambda) = \frac{1}{n-1}\sum_{i=1}^{n}(\lambda_i - 1)^2 \quad \text{(A6.5a)}$$

This returns $n_e = n$ when $\sigma^2(\lambda) = 0$ and $n_e = 1$ when $\sigma^2(\lambda) = n$, which matches the expected results from the eigenvalue analysis for these extreme cases. A closely related variance-based estimator was suggested by Patterson et al. (2006).

Li and Ji (2005) noted that Cheverud's method often returns an overly large value of n_e (and therefore, less power), especially when used with a large number of moderately correlated tests. While Cheverud's approach considered the two extreme cases (n vs. one independent test), Li and Ji noted that a set of c identical tests will result in an eigenvalue of $c > 1$, while tests that are only partially correlated with others will generate eigenvalues values < 1. Hence, they partitioned an eigenvalue into two parts, its integer value and the remainder, where the integer part implies identical tests (and hence is counted as contributing one independent test), while the remainder represents partial correlations. Hence, if an eigensequence is $4.1, 3.5, 1, 0.5, 0.1, \cdots$, then the first three eigenvalues correspond to three independent tests, with the total of their non-integer residuals ($0.1 + 0.5 + 0.5 + 0.1 = 1.2$) adding one additional test, giving (for this part of the sequence) an effective number of four independent tests. Formally, the **Li-Ji method** is coded as

$$n_{e,Li-Ji} = \sum_{i=1}^{N} I(\lambda_i \geq 1) + \sum_{i=1}^{N} (\lambda_i - \text{floor}[\lambda_i]) \qquad \text{(A6.5b)}$$

where the indicator function $I(x \geq 1)$ returns a value of one when $x \geq 1$, and otherwise returns a value of zero. Hence, the first sum in Equation A6.5b is the number of eigenvalues of $\mathbf{C}$ that are ≥ 1. The floor$[x]$ function in the second term corresponds to the largest integer $\leq x$, so the second sum is all of the remainder terms (the effects of partial correlations among tests). A related estimator was suggested by Li et al. (2011) and additional corrections for dealing with correlated tests have been proposed by a number of authors (e.g., Zaykin et al. 2002; Owen 2005; Efron 2007; Leek and Storey 2007, 2008; Gao et al. 2008; Moskvina and Schmidt 2008; Galwey 2009; Li et al. 2012).

FDR

- The false discovery rate
  - Bonferroni essentially assumes that ALL of your tests are from the null
  - In modern genomics, we EXPECT a subset (likely a small fraction) to be from the alternative
  - In such settings, we would like to control the false positive (or false discovery) rate among a set of tests that we *declare to be significant*
  - A FDR of 0.01 implies that only 1% of all tests declared to be significant (discoveries) are false positives

To formally motivate the concept of the FDR, suppose a total of n hypotheses are tested (e.g., genes), S of which are judged significant (i.e., the p value for the test is $\leq$ some threshold value, τ). If we had complete knowledge, we would know that n_0 of the hypotheses have the null true and $n_1 = n - n_0$ have the alternative true, and we might find that F of the true nulls were called significant, while T of the alternative trues were called significant, yielding the following table

	Called significant	Called not significant	Total
Null true	F	$n_0 - F$	n_0
Alternative true	T	$n_1 - T$	n_1
Total	S	$n - S$	n

For this experiment, the false-discovery rate is the fraction of tests called significant that are actually true nulls, FDR = F/S. (The term **discovery** follows in that a significant result can be considered as a discovery for future work.) As a point of contrast, the normal Type-I error (which we can also call the **false-positive rate** [**FPR**]), which is the fraction of true nulls that are called significant, is F/n_0. Note the critical distinction between these two error rates. While the numerator of each is F, the denominators are considerably different — the total number, S, of *tests called significant* (for FDR), versus the number, n_0, of hypotheses that are *truly null* (FPR). As the threshold value (τ) for significance is changed, so too is the fraction F/S. To obtain a FDR of δ over our experiment, τ is adjusted to find its largest value such that some expectation of F/S is bounded above by δ. Finally, Gadbury et al. (2004) defined the **expected discovery rate** (**EDR**) as T/n_1 (the fraction of all true discoveries declared to be significant), which is the analog of statistical power in this setting.

- To set an FDR, we find the critical p value for each CWER as follows.
  - If we set this critical value as τ, then $n\,\tau$ = expected number of false positives (more generally, $n_o\,\tau$)
  - The resulting FDR is $\delta = n\,\tau$ /(Number of tests with $p < \tau$)
  - Suppose n = 1000 and we try $\tau = 0.01$, so that 10 false positives are expected. If the actual number of tests with $p < 0.01$ is (say) 120, then FDR =10/120, or 0.083
  - Suppose we instead take $\tau = 0.005$, so that 5 false positives are expected, while (say) the actual number of tests with $p < 0.0075$ is now 110, giving FDR = 5/110 = 0.045
  - Setting the critical p value at $\tau = 0.005$ gives FDR = 0.045
  - For each set of p values, one tries different τ values until the desired FDR is obtain

Another way to see the distinction between the false-positive rate and the false-discovery rate is to consider them as probability statements for a single test involving hypothesis i. For the FDR, we condition on the test as being significant,

$$\text{FDR} = \Pr(i \text{ is truly null} \,|\, i \text{ is declared significant}) = \delta \qquad \text{(A6.11a)}$$

whereas for the false-positive rate, we condition on the hypothesis being null

$$\text{FPR} = \Pr(i \text{ is declared significant} \,|\, i \text{ is truly null}) = \alpha \qquad \text{(A6.11b)}$$

Example A6.9. Consider again the 10 ordered p values from Example A6.4. Computing $n\,p(k)/k = 10\,p(k)/k$, where k denotes the test with the *k-th* smallest p value, yields the following table:

k	1	2	3	4	5	6	7	8	9	10
$p(k)$	0.0020	0.0045	0.0060	0.0080	0.0085	0.0090	0.0175	0.0250	0.1055	0.5350
$n\frac{p(k)}{k}$	0.0200	0.0225	0.0200	0.0200	0.0170	0.0150	0.0250	0.0313	0.1172	0.5350

Thus, if we wish an overall FDR value of $\delta = 0.05$, we would reject hypotheses when $n\,p(k)/k \leq \delta = 0.05$, which is satisfied by H(1) through H(8). Notice that this procedure rejects more hypotheses (i.e., returns more discoveries) than any of the sequential Bonferroni methods (Example A6.4).

PER– Posterior Error Rate

Fernando et al. (2004) and Manly et al. (2004) both noted that FDR measures are closely related to Morton's (1955b) **posterior error rate** (**PER**), originally introduced in the context of linkage analysis in humans (this is also referred to as the **false positive report probability [FPRP]**; Wacholder et al. 2004). Morton's PER is simply the probability that a single significant test is a false positive,

$$\mathrm{PER} = \Pr(F = 1 \mid S = n = 1) \tag{A6.12}$$

The connection between the FDR and the PER is that if we set the FDR to δ, then the PER for a randomly drawn significant test is also δ.

$$\mathrm{PER} = \Pr(F = 1 \mid S = n = 1) = \frac{\Pr(\text{false positive} \mid \text{null true}) \cdot \Pr(\text{null})}{\Pr(S = n = 1)} \tag{A6.13}$$

$$\mathrm{PER} = \frac{\alpha \cdot \pi_0}{\alpha \cdot \pi_0 + (1-\beta) \cdot (1-\pi_0)} = \left(1 + \frac{(1-\beta) \cdot (1-\pi_0)}{\alpha \cdot \pi_0}\right)^{-1}$$

Sham and Purcell (2014) noted that one can rearrange Equation A6.15a to find the α value to obtain a desired PER value of γ, with

$$\alpha = \left(\frac{\gamma}{1-\gamma}\right)\left(\frac{1-\pi_0}{\pi_0}\right)(1-\beta) \tag{A6.15b}$$

In particular, if there is complete power ($\beta = 0$) and only one of the n tested hypotheses departs from the null ($\pi_0 = 1 - 1/n$), Equation A6.15c reduces to

$$\alpha = \left(\frac{\gamma}{1-\gamma}\right)\left(\frac{1}{n-1}\right) \simeq \frac{\gamma}{n} \tag{A6.15c}$$

which recovers the Bonferroni correction (Equation A6.4).

The Type-I error rate, α, of a *random* test, and the PER for a *significant* test, which are often assumed to be the same, are actually very different. In addition to α, the PER also depends on the power, β, of a test and the fraction, π_0, of tests that are truly null (as these latter parameters influence the probability that a test is declared to be significant). Manly et al. (2004) noted that the PER is acceptably low only if the fraction of alternative hypotheses $(1 - \pi_0)$ is well above α.

Example A6.7. In Morton's original application, because there are 23 pairs of human chromosomes, he argued that two randomly chosen genes had a $1/23 \simeq 0.05$ *prior probability of linkage*, namely, $1 - \pi_0 = 0.05$, and thus $\pi_0 = 0.95$. Assuming a Type-I error rate of $\alpha = 0.05$ and 80% power to detect linkage ($\beta = 0.20$), applying Equation A6.15a yields a PER of

$$\frac{0.05 \cdot 0.95}{0.05 \cdot 0.95 + 0.80 \cdot 0.05} = 0.54$$

Hence, with a Type-I error control of $\alpha = 0.05$, a random test showing a significant result ($p \leq 0.05$) has a 54% chance of being a false positive. This occurs because most of the hypotheses are expected to be null — for example, if we draw 1000 random pairs of loci, 950 are expected to be unlinked and we expect $950 \cdot 0.05 = 47.5$ of these to show a false positive. Conversely, only 50 are expected to be linked, and we would declare $50 \cdot 0.80 = 40$ of these to be significant, so that $47.5/87.5 = 0.54$ of the significant results are due to false positives.

What value for α is needed under the above parameters to given a PER of $\gamma = 0.05$? From Equation A6.15b,

$$\alpha = \left(\frac{\gamma}{1-\gamma}\right)\left(\frac{1-\pi_0}{\pi_0}\right)(1-\beta) = \left(\frac{0.05}{1-0.05}\right)\left(\frac{1-0.95}{0.95}\right)(1-0.2) = 0.0022$$

Hence, declaring significance when $p \leq \alpha = 0.0022$ gives a PER of 5%.

Example A6.8. Suppose we set $\alpha = 0.005$ for each test, and assume that the resulting power is essentially 1 (i.e., $\beta \simeq 0$). Consider 5000 tests under two different settings. First, suppose that the alternative is very rare, with $n_1 = 1$ ($\pi_0 = 0.9998$). Under this setting, we expect $4999 \cdot 0.005 = 24.995$ false positives and one true positive ($1 \cdot [1 - \beta] = 1$), yielding an expected PER of

$$\text{PER} = \frac{24.995}{24.995 + 1} = 0.961$$

Thus, a randomly chosen *significant* test has a 96.1% probability of being a false positive.

Now suppose that the alternative is not especially rare, for example $n_1 = 500$ ($\pi_0 = 0.9$). The expected number of false positives is $4500 \cdot 0.005 = 22.5$, while the expected number of true positives is 500, yielding a PER of

$$\text{PER} = \frac{22.5}{522.5} = 0.043$$

The PER is thus rather sensitive to π_0, the fraction of all tests that are truly from the null hypothesis. If π_0 is essentially 1, a PER of δ is obtained using the Bonferroni correction, $\alpha = \delta/n$. However, if π_0 departs even slightly from one (i.e., more than a few of the alternative hypotheses are correct), then the per-test level of α to achieve a desired PER rate is considerably larger (i.e., less stringent) than that given by the Bonferroni correction, namely, $\alpha(\delta) > \delta/n$. For example, for a 0.04 experiment-wide error rate, $\alpha = 0.04/5000 = 8 \cdot 10^{-6}$, which is roughly 625 times smaller than the value of $\alpha = 0.005$ required for a 4% FDR, highlighting the greatly increased power under the FDR framework. This increased power arises because the FDR approach acknowledges that some fraction of the tests are not from the null.

Bayesian hypothesis testing
(Chapter 20, pp 678 - 680)

Finally, a rather different approach was suggested by Wacholder et al. (2004), WTCCC (2007), and subsequent authors (Thomas and Clayton 2004; Wakefield 2007, 2008, 2012; Ball 2011), namely using a Bayesian framework. An excellent discussion of Bayesian approaches for multiple GWAS comparisons is given by Stephens and Balding (2009). As suggested by Thomas and Clayton (2004), the basic tenet of this framework is that

> "it is not the number of tests performed but rather the prior credibility of the hypotheses that is important in interpreting a set of observed associations. That is, when a hypothesis is unlikely to be true *a priori*, we should require strong evidence to be convinced of its truth"

In this framework, the strength of evidence can be expressed as the **posterior odds ratio** in favor of a true association. Letting T denote the value of the test statistic and τ the significance threshold, then the odds ratio in favor of a true association when the test is deemed significant can be written as

$$\begin{aligned}\frac{\Pr(H_1 \mid T > \tau)}{\Pr(H_0 \mid T > \tau)} &= \frac{\Pr(T > \tau \mid H_1)\Pr(H_1)/\Pr(T > \tau)}{\Pr(T > \tau \mid H_0)\Pr(H_0)/\Pr(T > \tau)} \\ &= \left[\frac{\Pr(T > \tau \mid H_1)}{\Pr(T > \tau \mid H_0)}\right]\left[\frac{\Pr(H_1)}{\Pr(H_0)}\right] = \frac{1-\beta}{\alpha}\,\frac{\Pr(H_1)}{\Pr(H_0)} \qquad (20.7a)\end{aligned}$$

where the first step follows from Bayes theorem (Equation 3.3b). Here α is the Type-I error rate, β the Type-II error rate (for a power of $1-\beta$), and $\Pr(H_1)$ and $\Pr(H_0)$ are the prior probabilities for, and against, an association. To apply Equation 20.7a, one must have some loose idea about the fraction of independent regions that generate associations with the trait, and some details about the effect size (and frequency) in order to specify β. WTCCC (2007) suggested values for $\Pr(H_1)$ in the range of 10^{-4} to 10^{-6}. More recently, Yengo et al. (2022) detected 12,000 (genome-wide) significant SNPs for height out of roughly a million tested, for a $\Pr(H_1)$ value closer to 0.01.

Equation 20.7a is very closely related to Morton's (1955b) **posterior error rate** (**PER**) — which Wacholder et al. (2004) referred to as the **false positive report probability** (**FPRP**) — see Appendix 6. Denoting the posterior odds ratio (Equation 20.7a) as PO, an alternative metric is the **posterior probability of association**, **PPA** (Stephens and Balding 2009), where

$$PPA = \frac{PO}{1 + PO} \tag{20.7b}$$

Finally, a more general approach is to replace

$$\frac{\Pr(T > \tau \mid H_1)}{\Pr(T > \tau \mid H_0)} \qquad \text{with} \qquad \frac{\Pr(T \mid H_1)}{\Pr(T \mid H_0)} \tag{20.7c}$$

Namely, replacing a threshold being exceeded with the actual value, T, of the test statistic for that marker. This ratio of support for the data (T) under the alternative versus the null hypothesis is called a **Bayes factor**, **BF** (Appendix 7). Computing the BF requires assumptions about the prior distribution of allele effects (and their associated allele frequencies), see Wakefield (2007, 2008, 2012) and Stephens and Balding (2009) for details. Stephens and Balding make the important point that as GWAS analysis moves beyond one-marker-at-a-time considerations to more complex units of interactions, Bayesian approaches can offer much more flexibility than frequentist methods.

Example 20.6 As an application of Equation 20.7a, consider the scenario where ten regions, out of one million LD blocks, influence the trait of interest, and suppose that there is 50% power to detect each effect. To obtain an odds-ratio of ten to one in favor of a true effect, rearranging Equation 20.7a gives a required α value of

$$\begin{aligned}\alpha &= \left(\frac{\Pr(H_0 \mid T > \tau)}{\Pr(H_1 \mid T > \tau)}\right)\left([1-\beta]\,\frac{\Pr(H_1)}{\Pr(H_0)}\right) \\ &= \frac{1}{10}\cdot 0.5 \cdot \frac{10/10^6}{1-10/10^6} = 5\times 10^{-7}\end{aligned}$$

Under these parameters, we expect an average of $0.5 \cdot 10 = 5$ true discoveries and $5 \times 10^{-7} \cdot (10^6 - 10) = 0.5$ false discoveries, for an expected FDR of 0.5/5.5, or around 9%. The PPA for an odds ratio of 10 (PO = 10) becomes PPA = 10/11 = 0.909. Similarly, for a posterior odds ratio of 20 (with a resulting PPA of 20/21 = 0.952), $\alpha = 2.5 \times 10^{-7}$, which yields an expected 0.25 false discoveries and an expected FDR of 0.25/5.25, or slightly under 5%.

Now suppose a highly polygenic trait, such as height, which may have a large number of regions (say 1000), but, given their smaller effect sizes, lower power of detection (say $\beta = 0.8$, or a power of 20%). The critical value for a odds ratio of ten becomes

$$\alpha = \frac{1}{10} \cdot 0.2 \cdot \frac{1000/10^6}{1 - 1000/10^6} = 2 \times 10^{-5}$$

In this setting we expected an average of 0.2· 1000 = 200 true discoveries and $2 \times 10^{-5} \cdot (10^6 - 1000) \simeq 20$ false discoveries, for an expected FDR of 20/220, or again around 9%. For a posterior odds ratio of 20:1, $\alpha = 1 \times 10^{-5}$, yielding an expected FDR of 10/210, again slightly under 5%.

Permutation tests: I

- In complex data sets, p values are often obtained using permutation tests (the gold standard).
- If the data consists of the n vectors (z_i, m_i), the trait value and marker vector for individual i,
  - randomize z over m (keeping the elements intact)
  - Repeat this shuffling several thousand times to generate an empirical distribution. (histogram) under the null

Permutation tests: II

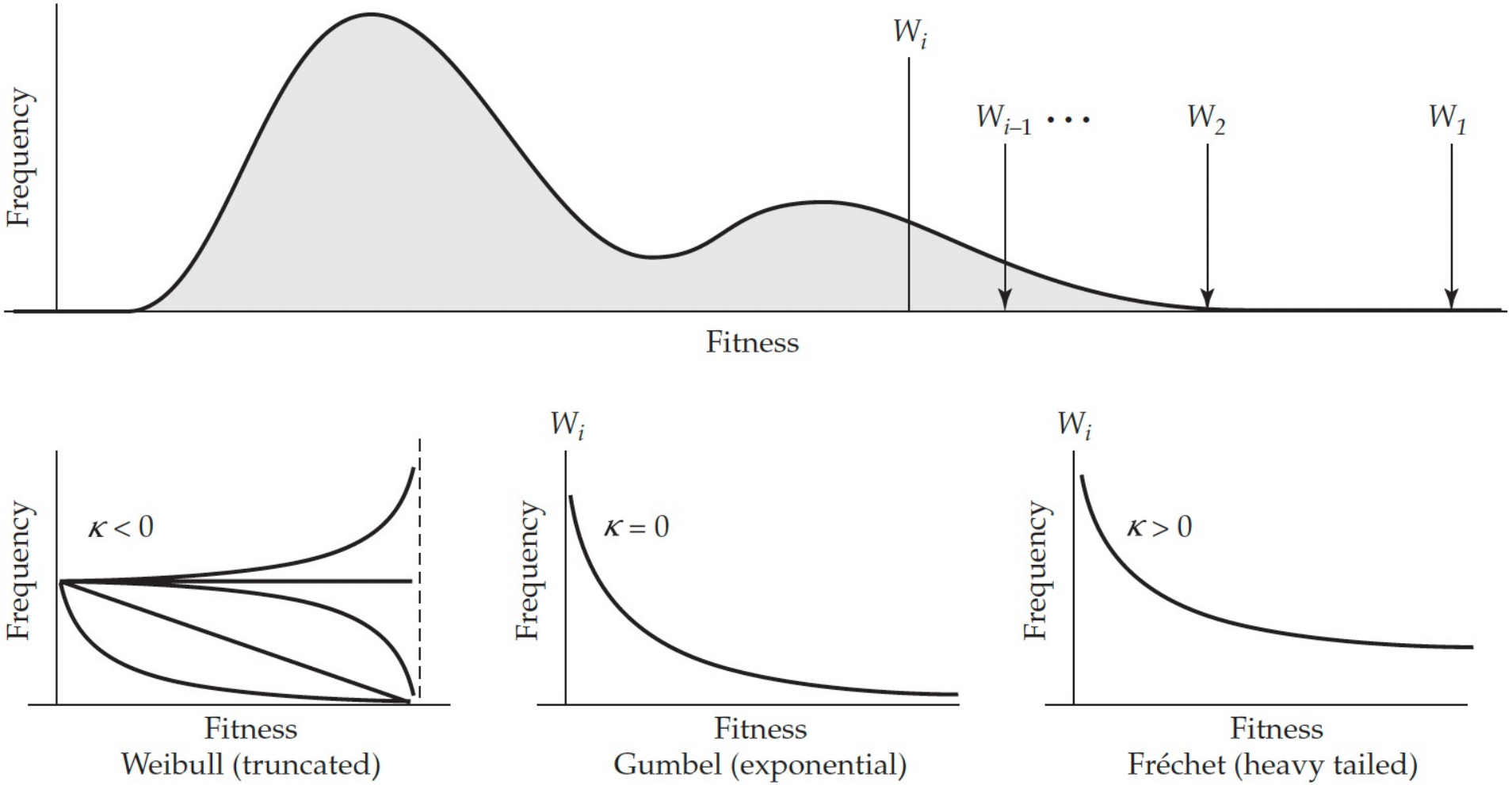
- Pros
  - Very straightforward when the exchangeable units are obvious (z vs **m** in our example)
  - Very robust when done correctly
- Cons
  - The exchangeable units may be unclear (e.g., z are now from families)
  - Computationally demanding for very small p values (n ~ 10/p, e.g., 10^7 for $p = 10^{-6}$)

Extreme Value Theory

The critical result from EVT is the so-called **trinity** (or **Fisher–Tippett–Gnedenko**) **theorem**, which states that the distribution of draws of extremes (the **extreme-value distribution**) from any underlying distribution are given by the **generalized Pareto distribution** (Pickands 1975), a family of distributions determined by a scale parameter, τ, and shape parameter, κ (also called the **tail index**), which falls into one of three limiting types (or **domains**), depending on the value of κ (Equation 27.7). Because our interest is in the distribution of fitness values for new beneficial alleles, setting the fitness of the current allele to 1, then the fitness of a beneficial allele is $1 + X$, where the fitness increase, X, is drawn from the tail of the underlying distribution to the right of zero. Following Beisel et al. (2007), the families of extreme-value distributions for such draws are given by

$$\Pr(X \le x \,|\, \tau, \kappa) = \begin{cases} 1 - (1 + \kappa\, x/\tau)^{1/\kappa}, & x \ge 0 & \text{if } \kappa > 0 \quad \text{(Fréchet)} \\ 1 - (1 + \kappa\, x/\tau)^{1/\kappa}, & 0 \le x < -\tau/\kappa & \text{if } \kappa < 0 \quad \text{(Weibull)} \\ 1 - \exp(-x/\tau), & x \ge 0 & \text{if } \kappa = 0 \quad \text{(Gumbel)} \end{cases} \tag{27.7}$$

Most common distributions (normal, gamma, etc.) have a **Gumbel** EV distribution ($\kappa = 0$), which has an exponential tail. This is the most commonly assumed EV distribution for beneficial mutations. When the underlying distribution is truncated to the right (the upper bound of $-\tau/\kappa > 0$, as $\kappa < 0$), the extreme-value distribution is in the **Weibull domain**.


Frequency
W_i
W_{i-1} • • •
W_2
W_1
Fitness
Frequency
$\kappa < 0$
Fitness
Weibull (truncated)
W_i
Frequency
$\kappa = 0$
Fitness
Gumbel (exponential)
W_i
Frequency
$\kappa > 0$
Fitness
Fréchet (heavy tailed)

Permutation + EVT

- In some settings, such as combining results from different data sets, need very small p values to control the PER/FDR
  - Could use permutation tests in each, but computationally very demanding to obtain the small-p cutoffs for each test
  - However, could use a modest size permutation tests (n ~ 10,000 to 20,000) and then use ML on the observed "tail" (extreme) p values to fit the generalized Pareto expected for extreme values (i.e., estimate κ), and then obtain your desired extremely small p cutoff analytically

Part III
Combining p values

- Motivation:
  - Gene (or pathway)-based GWAS
  - Suppose that you test k SNP to see if they (as a group) are significant (i.e., clustered near zero)

Combining p values

- Key: under the null, the distribution of p values is a uniform (over 0-1)
  - Implies -log(p) ~ chi^2_2. (Fisher)
  - While no single p might be significant, a clustering near zero can also give a signal
  - e.g., while none in the p-value sequence of tests 0.06, 0.07, 0.075, 0.08, 0.85 are significant, it is clear that they are highly nonrandom

Fisher's combined probability test: for k independent tests, where p_i denotes the p value for test i, the sum

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i) \tag{A6.1a}$$

approximately follows a χ^2_{2k} distribution. Fisher's method is a special case of the more general **Gamma method** (Lancaster 1961; Zaykin et al. 2007; Biernacka et al. 2012), based on inverses of gamma functions (Table A7.1). Other approaches, such results based on the distribution of the sum of n unit uniforms, have also been proposed (e.g., Edgington 1972), see Folkes (1984), Kocak (2017), and Heard and Rubin-Delnchy (2018) for a brief overviews.

Example A6.1. Suppose five different groups collected data to test the same hypothesis, and these groups (perhaps using different methods of analysis) reported p values of 0.10, 0.06, 0.15, 0.08, and 0.07. Notice that while none of these individual tests are significant, the trend is clearly that all are "close" to being significant ($\overline{p} = 0.09$). Fisher's statistic becomes

$$X^2 = -2\sum_{i=1}^{k} \ln(p_i) = 24.3921 \qquad \text{with} \qquad \Pr(\chi^2_{10} \geq 24.39) = 0.0066$$

Hence, when taken together, these five tests show a highly significant p value. This is reasonable, as each p value, while not individually significant, still all cluster near small values, while under the null, $\overline{p}$ should be around 0.5 (the average of a unit uniform). In particular, note that two tests each with $p = 0.06$ ($X^2 = 11.25, \Pr(\chi^2_4 \geq 11.25) = 0.024$) offer more support against the null than a single test with $p = 0.05$.

Stouffer's Z Score

An alternative to Fisher's approach for combining p values was offered by Stouffer et al. (1949; and independently by Liptak 1958), who transformed the individual p values into Z scores (unit normal random variables). The sum of k independent unit normals is itself normal, with a mean of zero and a variance of k. These results lead to **Stouffer's Z score** method: assign a score of Z_i for test i by solving $\Pr(U > Z_i) = p_i$. Let Z_s denote the sum over the transformed p values of k tests, scaled by $k^{-1/2}$ to give it a variance of one, with

$$Z_s = \frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}} \tag{A6.2a}$$

Because $Z_s \sim N(0,1)$, the overall p value is obtained as

$$p = \Pr(U > Z_s) \tag{A6.2b}$$

Besides providing symmetric values for large and small p values (i.e., p and $1-p$), a second advantage of the Z-score approach is that one can individually *weight* p values from different tests (Mosteller and Bush 1954; Liptak 1958), as the weighted sum of unit normals is itself a unit normal (while the weighted sum of χ^2 variables — the analog for Fisher's test — is considerably more complex). The resulting weighted version becomes

$$Z_w = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}} \tag{A6.2c}$$

Example A6.2. Reconsider the data from Example A6.1. The Z_i values are easily obtained using `R`, as the command `qnorm(1-p)` returns Z satisfying $\Pr(U \leq Z) = 1 - p$, or (equivalently) $\Pr(U > Z) = p$. For example, Z_1 is calculated by `qnorm(1-0.1)`, or 1.281. Similarly computing the other Z_i values yields

$$\sum_{i=1}^{5} Z_i = 6.754, \qquad \text{hence} \qquad Z_s = \frac{6.754}{\sqrt{5}} = 3.020$$

Because $\Pr(U > 3.020) = 0.00126$, as in Example A6.1, the combined p value is highly significant.

Other p-combining methods

Tippet's Combined p-value Estimator

Rearranging Equation A6.3a yields **Tippett's method** (1931), also called the **Bonferroni-adjusted minimum *p*-value**, for combining p values from independent tests. Here π is interpreted as the combined p value of n tests (p_{Tip}), while α is replaced by the smallest p value in the series, $p(1)$, yielding

$$p_{Tip} = 1 - [1 - p(1)]^n \simeq np(1) \qquad \text{(A6.4a)}$$

Wilkinson (1951) generalized this method to using the jth smallest p value.

Simes method

$$p_{Simes} = \min_k \left[\frac{n\, p(k)}{k} \right]$$

Can replace n by n_e

Dealing with Dependency: Aggregated Cauchy Approaches

A recent alternative strategy (Pillai and Meng 2016; Liu et al. 2019; Liu and Xie 2020) for dealing with dependency among tests is build around using the **Cauchy distribution**, whose density function is given by

$$\varphi(x) = \frac{1}{\pi\,(1+x^2)} \quad \text{for} \quad -\infty < x < \infty \tag{A6.6a}$$

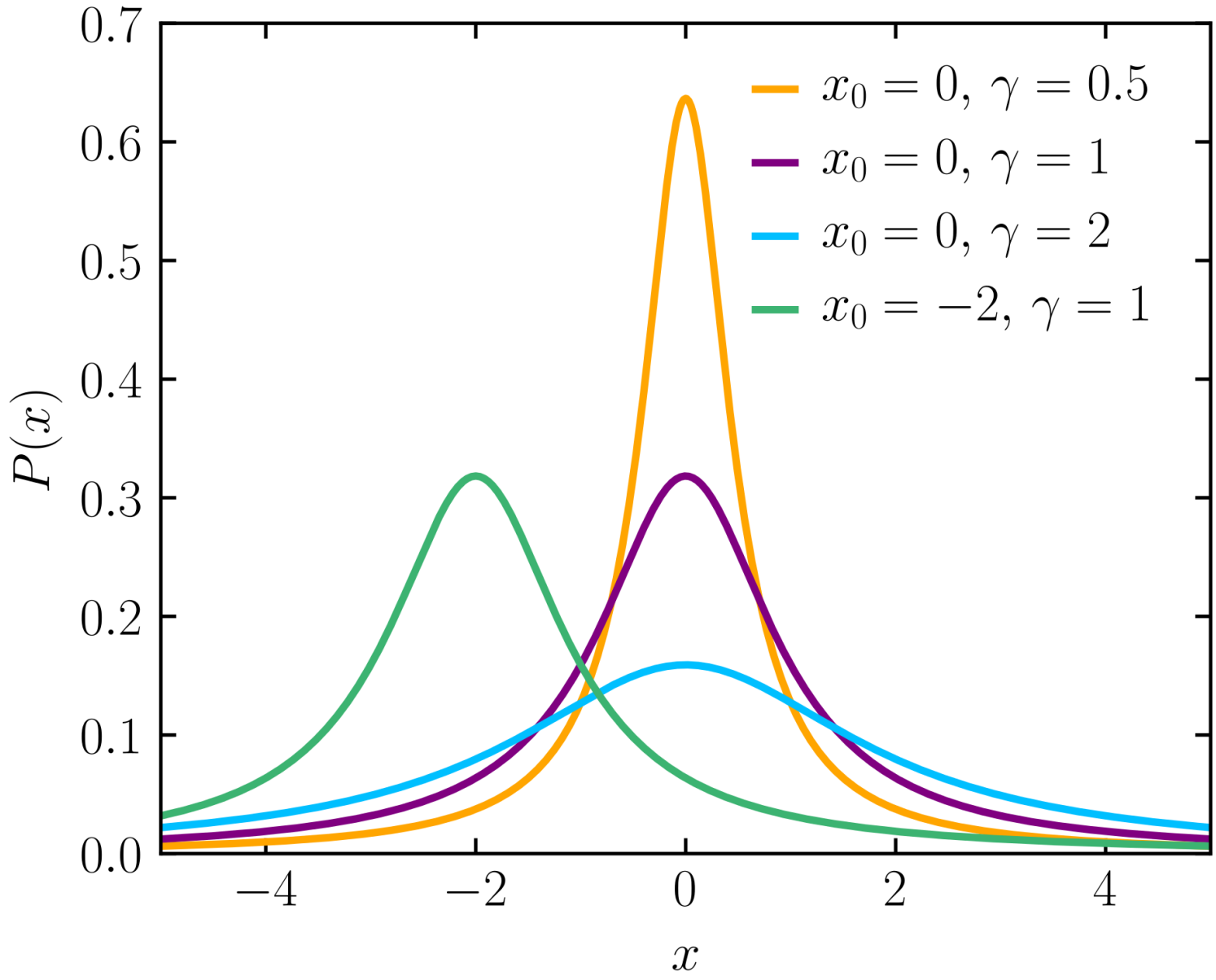
This is a **standard Cauchy**, akin to a unit normal. (More generally, the Cauchy has both a location x_0 and scale γ parameter, which do not concern us here, with the standard using $x_0 = 0$ and $\gamma = 1$.) The Cauchy arise as the distribution for the ratio of two unit normals (one can also think of it as a t distribution with one degree of freedom). If $\mathcal{C}$ is Cauchy distributed, then its cdf (cumulative distribution function) is given by

$$\Pr(\mathcal{C} \leq x) = \int_{-\infty}^{x} \frac{1}{\pi\,(1+x^2)}\,\mathrm{d}x = \frac{\arctan(x)}{\pi} + \frac{1}{2} \tag{A6.6b}$$

and hence

$$\Pr(\mathcal{C} > x) = 1 - \left(\frac{\arctan(x)}{\pi} + \frac{1}{2}\right) = \frac{1}{2} - \frac{\arctan(x)}{\pi} \tag{A6.6c}$$

Cauchy distribution



Recalling that $\arctan(x) \to \pi/2$ as $x \to \infty$, shows that $\Pr(\mathcal{C} \leq x) \to 1$ as $x \to \infty$. From Equation A6.6b, it immediately follows that the quantile x_p — the value of x_p satisfying $\Pr[\mathcal{C} \leq x_p] = p$ — is given by solving $p = \arctan(x_p)/\pi + 1/2$, yielding

$$x_p = \tan\left[\pi(p - 1/2)\right] \tag{A6.6d}$$

Because $\tan(x) \to \infty$ as $x \to \pi/2$, it follows that $x_p \to \infty$ as $p \to 1$.

For many students, the Cauchy has an unfortunate image, being mainly known as a pathologic distribution (its density function integrates to one, but none of its moments are finite). This feature is due to having heavy tails that do not decline sufficiently fast at large values. A less appreciated feature of the Cauchy is that the weighted sum of standard Cauchys is itself a Cauchy (and, hence, Cauchy random variables do not obey the central limit theorem). Building on this fact, the interest in using the Cauchy for combining p values started with a striking finding by Pillai and Meng (2016). Suppose that $\mathbf{y}$ and $\mathbf{x}$ are both vectors of unit MVNs with correlation matrix $\mathbf{V}$. Pillai and Meng showed that the weighted sum $\sum w_i \left(y_i/x_i\right)$ is a standard Cauchy. Noting that y_i/x_i itself is Cauchy, this implies (in least in this setting) that the weighted sum of Cauchy random variables with arbitrary dependency structure is still a Cauchy.

Recall that Stouffer's Z score approach translates individual p values into Z values (unit normals), and the weighted sum of these values (which is a normal) is back-transformed into an overall p value. This same approach, but now using the Cauchy, was proposed by Liu et al. (2019) with their **aggregated Cauchy association test** (**ACAT**). As a result of its heavy-tail, the Cauchy is largely insensitive to correlations among p values, especially when the p values are small (Liu et al. 2019; Liu and Xie 2020). The ACAT test statistic for combining p values is

$$T_{ACAT} = \sum_{i=1}^{n} w_i \tan[(0.5 - p_i)\pi] \tag{A6.6e}$$

namely the weighted ($w_i \geq 0$ and $\sum w_i = 1$) Cauchy values associated with the p values of each test (Equation A6.6d), resulting in the weighted sum also being a standard Cauchy. The associated overall p value follows from Equation A6.6c, with

$$p_{ACAT} \simeq 0.5 - \frac{\arctan[T_{ACAT}]}{\pi} \tag{A6.6f}$$

Part VI:
Meta-analysis

Combining estimates over studies

- While p-value combining can be used, the more typical question is not whether an effect is significant, but rather its effect size.
- The field of meta-analysis (the analysis of analyses) offers methodology for this task
  - Much more in WVL Appendix 6

Table A6.3 The potential data available for a meta-analysis of k studies. One has reported values for the estimated effect, T_i, for each study. A study may also have m reported **moderator variables** (cofactors such as sex, laboratory, species, etc.). In the simplest setting, the study only reports a p value for whether the effect is significant. In such cases, an overall p-value can be obtained from the methods discussed earlier in this Appendix. If the study also reported the standard errors (SE) of the estimates, or (under the assumption of a constant error per observation over all studies) the sample size, then a formal (i.e., model-based) meta-analysis may be performed (as detailed in the text).

	Effect					
Study	Actual	Estimate	Sample size	SE	p value	Moderator variables
1	θ_1	T_1	n_1	s_1	p_1	$M_{11}, M_{12}, \cdots, M_{1m}$
2	θ_2	T_2	n_2	s_2	p_2	$M_{21}, M_{22}, \cdots, M_{2m}$
3	θ_3	T_3	n_3	s_3	p_3	$M_{31}, M_{32}, \cdots, M_{3m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	θ_k	T_k	n_k	s_k	p_k	$M_{k1}, M_{k2}, \cdots, M_{km}$

Under a **fixed-effects meta-analysis** (also called the **common-effect model**), we assume that the actual effect size is the *same* over all studies ($\theta_i = \theta$), which yields

$$T_i = \theta + e_i \tag{A6.30a}$$

where we assume that the residuals are independent but heteroscedastic, as $\sigma^2(e_i) = s_i^2$. Under the fixed-effects model, our interest is in combining studies to obtain a better estimate of the common (fixed) effect, θ. This simply involves generalized least-squares (GLS; Equation 10.13a), with the resulting meta-analysis global estimate of θ (given the k studies) being

$$\overline{T} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}, \quad \text{where} \quad w_i = \frac{1}{s_i^2} \tag{A6.30b}$$

In other words, we use a weighted average, with each study weighted by its precision (studies with smaller standard errors receive larger weights). The meta-analysis standard error, $s_{\overline{T}}$, for the global estimate, $\overline{T}$, is

$$s_{\overline{T}}^2 = \frac{1}{\sum_{i=1}^{k} w_i} \tag{A6.30c}$$

For the situation where we assume that each individual observation in a given study has the same variance, $\sigma^2(T_i) = \sigma^2/n_i$, then for k studies, each of size n,

$$\sigma^2(\overline{T}) = \frac{\sigma^2}{nk} \tag{A6.30d}$$

An obvious next line of inquiry is whether the assumption of a common effect over all studies is reasonable. This can be examined using the **Cochran Q test of heterogeneity**,

$$Q = \sum_{i=1}^{k} \frac{\left(T_i - \overline{T}\right)^2}{s_i^2} \tag{A6.31}$$

where (under the null of $\theta_1 = \cdots = \theta_k$, and assuming that the values of T_i are normally distributed), the distribution of Q is χ^2 with $(k-1)$ degrees of freedom.

One potential reason for a significant Q is that the study consists of different subsets of groups (say, males versus females), with a common effect that was the same in each group but differs among groups. In this case, we can extend the basic model by including a regression on moderator variables,

$$T_i = \theta + \sum_{j=1}^{m} b_j M_{ij} + e_i \qquad \text{(A6.32)}$$

Often the values of M_{ij} are simply zero-one indicator variables (e.g., 0 for male, 1 for female), but they can be more general regression slopes as well. For example, M_{1j} could be the age of individuals within study j, with a significantly nonzero value of b_1 in Equation A6.32 indicating that the treatment mean varies with age. Again Equation A6.32 is simply a GLS regression, and one can test for moderator-variable effects ($b_j \neq 0$) in the standard regression fashion (Chapter 10).

In most biological settings, the assumption of a single common value for the treatment mean over all studies is unrealistic. For example, in a meta-analysis of QTL effect sizes, we *expect* θ_i to vary over studies, and our interest shifts to the variance *among* the actual effects. This leads to the **random-effects meta-analysis** model

$$T_i = \mu + u_i + e_i \tag{A6.33a}$$

where $\mu_i \sim (0, \sigma_u^2)$. Typically, the effect sizes ($\theta_i = \mu + u_i$) are assumed to be drawn from a normal, $\theta_i \sim \mathrm{N}(\mu, \sigma_u^2)$, and are independent of the residuals (which remain heteroscedastic). Under a random-effects analysis, our interest is the variation, σ_u^2, among the realized effects, in addition to their overall grand mean, μ. The estimate for the latter is also of the form of Equation A6.30b, but with a critical difference. Under a random-effects model, the weights are now given by

$$w_i = \frac{1}{s_i^2 + \widehat{\sigma}_u^2} \tag{A6.33b}$$

where $\widehat{\sigma}_u^2$ is the estimate of σ_u^2. One option for obtaining this variance is the **DerSimonian-Laird estimator**, which is based on Cochran's Q value (Equation A6.31),

$$\widehat{\sigma}_u^2 = \frac{Q - (k-1)}{S_1 - (S_2/S_1)}, \quad \text{where} \quad S_j = \sum_{i=1}^{k} s_i^{-2j} \quad \text{for } j = 1, 2 \tag{A6.33c}$$

which is set to zero if it is negative (DerSimonian and Laird 1986; 2015), although other approaches (e.g., REML; Part II Chapter 33) could also be used, and more robust estimates have been suggested by Knapp and Hartung (2003) and Sidik and Jonkman (2005).

Finally, in many settings, we might expect the grand mean to vary over different categories, such as when a gene's transcript level differs between males and females. Similarly, we may wish to examine whether the distribution of QTL effect size varies between life-history versus morphological traits. The potential of different means over different major categories can be accommodated in a meta-analysis model by the use of moderator variables (cofactors). These adjust the mean for a particular class, leading to a **mixed-model meta-analysis.** Suppose that there are $m \ll k$ moderators. The resulting mixed-model is

$$T_i = \mu + \sum_{j=1}^{m} b_j M_{ij} + u_i + e_i \qquad \text{(A6.35a)}$$

where b_j is the (fixed) effect of a moderator, j, which has a value of M_{ij} in study i. Equations A6.32 and A6.33a are special cases of Equation A6.35a, which we can write in general-linear-model form (Chapter 10) as

$$\mathbf{y} = \mathbf{Mb} + \mathbf{u} + \mathbf{e} \qquad \text{(A6.35b)}$$

where $y_i = T_i$, $\mathbf{b} = (\mu, b_1, \cdots, b_m)^T$, and the ith row of the $k \times (m+1)$ matrix, $\mathbf{M}$, contains the values of the m moderator variables associated with study i (with the first column of $\mathbf{M}$ being all ones for the common μ). The vectors, $\mathbf{u}$ and $\mathbf{e}$, of random effects are assumed uncorrelated, with $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$ and $\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$, where $\mathbf{R}$ is a known diagonal matrix, $\text{diag}(s_1^2, s_2^3, \cdots, s_k^2)$, and $\mathbf{G} = \sigma_u^2 \mathbf{C}$, where $\mathbf{C}$ is a matrix of known constants.

Fixed- vs random-effect meta-analysis

- Under fixed, weights based on precision of estimates

$$\overline{T} = \frac{\sum_{i=1}^{k} w_i T_i}{\sum_{i=1}^{k} w_i}, \quad \text{where} \quad w_i = \frac{1}{s_i^2}$$

- Under random, weights also include random-effects variance. Hence, study weights are more even

$$w_i = \frac{1}{s_i^2 + \widehat{\sigma}_u^2}$$

Part V:
Fitting high-dimensional data and model selection

Overview

- Hotelling's T^2
- Penalized regressions
- Model selection: AIC, BIC,

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + e$$

Multivariate joint confidence intervals are in the form of *ellipses* (for two dimensions) and *ellipsoids* for higher dimensions. To motivate the form of these intervals, we first consider **Hotelling's $\mathbf{T}^2$** statistic. This is essentially the multivariate extension of the classic univariate t test, where for $\overline{z} \sim \mathrm{N}(\mu_0, \sigma^2/n)$,

$$t = \frac{(\overline{z} - \mu_0)^2}{S^2/n}$$

for the null hypothesis that the mean is μ_0 given that the sample variance of z is S^2 (and hence the sample variance for $\overline{z}$ is S^2/n). Squaring both sides, we can express this as

$$t^2 = (\overline{z} - \mu_0)(S^2/n)^{-1}(\overline{z} - \mu_0)$$

Hotelling's T^2 statistic generalizes this to multivariate form, where for $\overline{\mathbf{z}} \sim \mathrm{MNV}(\boldsymbol{\mu}_0, \mathbf{S}/n)$,

$$T^2 = (\overline{\mathbf{z}} - \boldsymbol{\mu}_0)^T (\mathbf{S}/n)^{-1} (\overline{\mathbf{z}} - \boldsymbol{\mu}_0) \qquad \text{(A3.23a)}$$

Under the null hypothesis ($\boldsymbol{\mu} = \boldsymbol{\mu}_0$),

$$\left(\frac{n-p}{(n-1)p}\right) \cdot T^2 \sim F_{p,n-p} \qquad \text{(A3.23b)}$$

Penalized regressions

Example 20.4 A common situation that arises in modern quantitative genetics are regressions whose number of parameters p exceeds, often greatly, the sample size n. In a standard least-square regression framework, estimation proceeds using generalized inverses (Appendix 3), resulting in a set of solutions. A more powerful approach is to use **penalized** (or **regularized**) **regressions**. Consider the regression model

$$y_i = \mu + \sum_{j=1}^{p} \beta_j X_{i,j} + e_i$$

In the standard OLS framework, one solves for the β_j that minimizes the sum of squared residuals,

$$\text{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \mu - \sum_{j=1}^{p} \beta_j X_{i,j} \right)^2 \qquad (20.5e)$$

Penalized regressions start with this framework, and then place constraints on the β_j. Under **ridge regression** (**RR**), one instead minimizes RSS $+\lambda \sum \beta_j^2$ (Hoerl and Kennard 1970), for some **shinkage parameter** $\lambda > 0$ (Hoerl et al. 1975; Lawless and Wang 1976, and Cule at al. 2011 discuss the choice of λ). An alternative approach is the **least absolute shrinkage and selection operator**, or **LASSO**, which minimizes RSS $+\lambda \sum |\beta_j|$ (Tibshirani 1996). Note that for values of β near zero, β^2 is a much less harsh constraint than $|\beta|$ (as $|\beta| \gg \beta^2$ for $|\beta| \ll 1$), so that the LASSO shrinks most of the β_j to exactly zero (yielding a **sparse estimate**), and hence is often used in **model selection** (choosing the parameters in the final model as those with nonzero βs). The **elastic net** (Zou and Hastie 2005) combines the RR and LASSO approaches, seeking to minimize RSS $+\lambda_1 \sum |\beta_j| + \lambda_2 \sum \beta_j^2$. These approaches can also be extended to generalized linear models, e.g., Le Cressie and van Houwelingen (1992) proposed a logistic ridge regression method.

Assessing model fit

In the same fashion that we decomposed the total variance into genetic and phenotypic components (Chapters 4–7), we can decompose the total variance of a response vector **y** into the variance accounted for by the linear model and the remaining (error or residual) variance. This is typically done by considering three sums of squares, with the **total sum of squares** (SS_T) being the sum of two components, the **error** (or **residual**) **sum of squares** (SS_E) and the **model sum of squares** (SS_M),

$$SS_T = SS_M + SS_E$$

The total sum of squares measures the total variability in the data, while the model sum of squares measures the amount of variation accounted for by the linear model. As noted in our discussions of univariate regression in Chapter 3, the fraction of total variance explained by a linear model is given by the **coefficient of determination**,

$$r^2 = \frac{SS_M}{SS_T} = 1 - \frac{SS_E}{SS_T} \qquad \text{(A3.15a)}$$

Need to adjust goodness of fit by number of used model parameters

One issue with r^2 is that goodness-of-fit improves (and hence r^2 increases) as more model parameters are added. As a result, for n observations and p estimated parameters, the **adjusted r^2** (Ezekiel 1930)

$$r^2_{adj} = 1 - \frac{\mathrm{MS_e}}{\mathrm{MS_T}} = 1 - \frac{\mathrm{SS_e}/(n-p)}{\mathrm{SS_T}/(n-1)} \qquad \text{(A3.15b)}$$

is often reported. This replaces sums of squares (SS) with mean squares (MS), with $r^2_{adj} \leq r^2$, and allows for a more fair comparison of different models by discounting goodness-of-fit as more model parameters are added. The connection between r^2 and r^2_{adj} is

$$r^2_{adj} = 1 - (1 - r^2)\left(\frac{n-1}{n-p}\right) \qquad \text{(A3.15c)}$$

If models are nested, can use LR tests to compare them
If not-nested, use ad-hoc metric

Such comparisons introduce the broader issue of **model selection** (Burnham and Anderson 2002). While a formal statistical framework occurs for nested comparisons (LR tests), there is an *informal* framework for comparing models that are not nested. Here, various informal statistics are used to compare models, and we focus on two, the AIC and BIC metrics. For both of these metrics, a smaller value means a better model. We stress that while both metrics are fully grounded in theoretical principles (Burnham and Anderson 2004), ***comparing values for two different models is largely ad hoc in that there is no formal test for significance*** (i.e., there is no formal criterion for determining when one model is clearly better than the others). Further, the different metrics can result in different model rankings.

The idea behind both metrics (as in a likelihood-ratio test) is reward goodness of fit (i.e., smaller values of $-2\ln[L]$ imply better fits), but also to penalize for the number of fitted model parameters, k. This is a natural extension of the adjusted r^2 (Equations A3.15b and A3.15c), which downweights goodness of fit by the number of fitted parameters. One of the most widely used model-comparison metrics is the **Akaike information criterion** (Akaike 1973),

$$\mathrm{AIC} = -2\ln(L) + 2k \qquad \text{(A4.15a)}$$

which was adjusted for the sample size, n, by Sugiura (1978),

$$\mathrm{AIC}_c = -2\ln(L) + 2k + \frac{2k(k+1)}{n-k-1} = -2\ln(L) + \frac{2kn}{n-k-1} \tag{A4.15b}$$

AIC_c should be used in place of AIC unless $n/k > 40$ (Burnham and Anderson 2004). Model selection proceeds by computing the AIC values for all of our candidate models, and then choosing the model with the smallest AIC value as the "best" model in the comparison test.

The other widely used metric is the **Bayes information criterion**,

$$\mathrm{BIC} = -2\ln(L) + \ln(n)k \tag{A4.15c}$$

which was introduced by Schwarz (1978), and thus is also known as the **Schwarz criterion**. While AIC and BIC are often used interchangeably, they are actually designed for slightly different purposes. When one of the models being compared is the true model, then BIC picks this model with a probability approaching one in large samples. Conversely, AIC considers the situation where *none* of the candidate models may be correct and then tries to pick among the best of these. Best practice is to typically present both AIC and BIC values, especially if they result in different model rankings. A nice short review of various other model selection criteria can be found in Grueber et al. (2011b).