STA 532: Theory of Statistical Inference

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

Hypothesis Testing

Statistical *Hypothesis Testing* is the business of quantifying the plausibility of some assertion about a population or a data-generating mechanism, upon observing a sample $X \in \mathcal{X}$. There are two rather different "classical" approaches to testing hypotheses— Significance Testing, introduced by Fisher (1935b), and fixed-level Hypothesis Tests, introduced by Neyman and Pearson (1933). For nice discussions, see (Lehmann, 1993; Berger, 2003). The Bayesian paradigm offers yet another approach to testing hypotheses.

1 Problem Formulation I: P-Values

In a parametric statistical model $\mathcal{F} = \{f(x \mid \theta) : \theta \in \Theta\}$ any assertion about the data-generating mechanism or population distribution $f(x \mid \theta)$ (or *hypothesis*) can be identified with a *subset* H_0 *of* Θ, namely, the set of those $\theta \in \Theta$ for which the assertion is true. Sir Ronald Aylmer Fisher sought to quantify the evidence against a scientific hypothesis $H_0 \subset \Theta$ represented by an outcome $x \in \mathcal{X}$ by the unlikeliness of observing such an x, if in fact $\theta \in H_0$ is true. In the first published report of a statistical hypothesis assessment ("Mathematics of a Lady Tasting Tea"), Fisher (1935b) described a test of his female acquaintance's claim that she could tell by taste whether a cup of tea with milk was prepared by adding milk to a cup of tea, or by adding tea to a cup containing a bit of milk (in the latter approach the tea will be somewhat hotter, possibly scalding the milk). Eight cups of tea were prepared, four with milk added to tea and four with tea added to milk, and presented to her in a random order. She was able to correctly identify all eight. Fisher argued that if the hypothesis H_0 that she had no ability to distinguish were true, then each of the $\binom{8}{4} = 70$ possible sets of four tastings would be equally likely to be correct, so her perfect performance was either

- If H_0 is true, a random event that would happen only with probability $1/70 = 0.0143$; or
- If H_0 is false, an expected result with high probability.

His conclusion, since miracles are rare, was that H_0 was discredited and that the lady did indeed possess the skill she claimed. Had she gotten only three out of four correct the corresponding probability would have been $16/70 = 0.229$, high enough that no such conclusion would have been warranted.

In larger experiments than this, typically *no* single outcome has particularly large probability— in fact all outcomes have probability zero for continuously-distributed data. Fisher's approach is to quantify the evidence against H_0 by the "P-value", defined as the probability (if $\theta \in H_0$) of the observed outcome x *or other outcomes more extreme.* In the Tea Tasting example, he would have reported $P = 17/70 = 0.243$ if she had three successes, or $53/70 = 0.757$ with just two, the probabilities of $S \ge 3$ or $S \ge 2$ successes if just guessing, respectively.

In general, significance testing begins with the selection of some "test statistic" $T : \mathcal{X} \to \mathbb{R}$ whose value would be expected to be small for $\theta \in H_0$ and large for $\theta \notin H_0$. The investigator then performs the experiment to observe $X = x$ and reports

$$P(x) := \sup_{\theta \in H_0} \mathsf{P}_\theta\big[T(X) \ge T(x)\big],$$

an upper bound for the probability of seeing evidence against H_0 as strong or stronger than that observed if in fact H_0 is true. Very small values of $P(x)$ are regarded as strong evidence against H_0.

The inclusion of outcomes "more extreme" than that observed means that this measure of evidencial strength depends on features of the sampling distribution beyond those included in the likelihood function, a bone of contention for Bayesians and others following the Likelihood Principle (Berger and Wolpert, 1988). The concern about the inclusion of *more extreme* results is nicely summarized by Harold Jeffreys' (1961, *p.* 385) quip,

> What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.

Fisher himself did not write of "testing" and did not view significance levels as definitive evidence of a hypothesis's truth or falsity— rather, he viewed P-values as informal suggestive ways of quantifying evidence in order to guide further scientific investigation. Others in the ensuing decades were less reticent— see Section (4) below.

1.1 Significance Test Examples

Poisson Example: In an experiment with Poisson-distributed data $X \sim \mathsf{Po}(\theta)$, an hypothesis that θ is rather small (say, $H_0 = [0, \theta_0]$ for some specified $\theta_0 \in \Theta = \mathbb{R}_+$) would be supported by rather small values of X and not by large ones. Thus any statistic $T : \mathcal{X} \to \mathbb{R}$ that monotonically *increases* would serve as a basis for significance testing, and the P-value upon observing $X = x$ would be

$$P(x) := \sup_{\theta \le \theta_0} \mathsf{P}_\theta[X \ge x] = \mathsf{P}_{\theta_0}[X \ge x] = \sum_{k=x}^{\infty} \frac{(\theta_0)^k}{k!} e^{-\theta_0},$$

easily evaluated in `R` as `ppois(x-1,theta.0,lower=FALSE)`. For a sample $\mathbf{x} = \{X_i\}_{1 \le i \le n}$ of size n, the sufficient statistic $T(\mathbf{x}) := \sum X_i$ has the $\mathsf{Po}(n\theta_0)$ distribution if H_0 is correct, leading to a similar test with P-value `ppois(T-1,n*theta.0,lower=FALSE)`.

Normal Example: In an experiment with Normally-distributed data $X \sim \mathsf{No}(\theta, \sigma^2)$ with known variance $\sigma^2 > 0$, an hypothesis that θ is rather large (say, $H_0 = [\theta_0, \infty)$) will require a statistic T that *decreases* monotonically with X; any such statistic will lead to P-value

$$P(x) := \sup_{\theta \ge \theta_0} \mathsf{P}_\theta[X \le x] = \mathsf{P}_{\theta_0}[X \le x] = \Phi\Big(\frac{x - \theta_0}{\sigma}\Big)$$

whose value will be $P \approx 1/2$ if $x \approx \theta_0$, falling to 0.05 or 0.01 as x falls to $\theta_0 - 1.645\sigma$ or $\theta_0 - 2.33\sigma$, respectively. For a sample $\mathbf{x} = \{X_i\}_{1 \le i \le n}$ of size n, the sufficient statistic $T(\mathbf{x}) := (\theta_0 - \bar{X}_n)$ has the $\mathsf{No}(0, \sigma^2/n)$ distribution under H_0, leading to a similar one-sided test with P-value `pnorm(T,sd=sigma/sqrt(n),lower=FALSE)`. In Section (3.1) below we'll consider two-sided hypotheses and alternatives.

2 Problem Formulation II: Size and Power

Jersey Neyman and Egon Pearson (son of the more famous Karl Pearson, who invented the χ^2 test presented in Section (6) below), wanted a formal way of selecting the test statistic T intended to distinguish between (small) outcomes suggestive of H_0 and (large) those unfavorable to the hypothesis. They introduced the idea of an *alternate hypothesis* $H_1 = \Theta \backslash H_0$ and a focus on the two possible errors one might make in testing hypotheses, to which they gave the amazingly unimaginative names:

Decision	H_0 True	H_0 False
Reject H_0:	Type I Error	Correct!
Fail to Reject H_0:	Correct!	Type II Error

Often the consequences of these two errors are very different, so there is no reason for them to be treated symmetrically. For example, in a test to see of the hypothesis "H_0: A proposed new drug treatment is safe", the Type I error of mistakenly concluding the drug isn't safe may lead to a lost opportunity of producing the drug, while the Type II error of mistakenly concluding that it *is* safe may lead to sickness or death for subjects— both undesirable outcomes, but very different.

In this decision-theoretic setting the task for the investigators is to identify a set $\mathcal{R} \subset \mathcal{X}$ of those outcomes $X = x$ for which H_0 will be rejected— called the *critical region* or *rejection region.* Such a test can be evaluated on the basis of the two rejection probabilities

$$\begin{array}{lll} \text{Size} & \alpha & = \sup_{\theta \in H_0} \quad \mathsf{P}_\theta\big[x \in \mathcal{R}\big] \\ \text{Power} & [1-\beta] & = \inf_{\theta \in H_1} \quad \mathsf{P}_\theta\big[x \in \mathcal{R}\big] \end{array}$$

Ideally one would want α to be quite small, to minimize the probability of a Type I error (false rejection of H_0) when H_0 is true, and also β to be quite small, to minimize the probability of a Type II error (failure to recognize that H_0 is false) if H_1 is true. In practice there is a trade-off between how small these error probabilities can be, and the sample-size for an experiment.

In the special "simple null *vs.* simple alternative" case in which $H_0 = \{\theta_0\}$ and $H_1 = \{\theta_1\}$ each consist of a single point, there is an optimal statistic T to use for discrimination:

Lemma 1 (Neyman/Pearson) *Let $H_0 = \{\theta_0\}$ and $H_1 = \{\theta_1\}$ for distinct points $\theta_0, \theta_1 \in \Theta$, and set*

$$\Lambda(x) := \frac{f(x \mid \theta_1)}{f(x \mid \theta_0)}$$

for $x \in \mathcal{X}$, the likelihood ratio against the null *hypothesis $H_0: \ \theta = \theta_0$. Fix any $c > 0$ and set*

$$\begin{aligned} \mathcal{R}_\star &:= \{x \in \mathcal{X} : \ \Lambda(x) \ge c\} \\ \alpha_\star &:= \mathsf{P}_{\theta_0}[\Lambda(X) \ge c] = \mathsf{P}_{\theta_0}[X \in \mathcal{R}_\star] \\ \beta_\star &:= \mathsf{P}_{\theta_1}[\Lambda(X) < c] = \mathsf{P}_{\theta_1}[X \notin \mathcal{R}_\star]. \end{aligned}$$

Then the test that rejects H_0 when $X \in \mathcal{R}_\star$ has size $\alpha_\star$ and power $[1-\beta_\star]$, and every other test $\mathcal{R}' \subset \mathcal{X}$ has either size larger than $\alpha_\star$ or power smaller than $[1-\beta_\star]$.

Proof. Denote by $\alpha' := \int_{\mathcal{R}'} f(x \mid \theta_0)dx$ and $\beta' := \int_{\mathcal{R}'^c} f(x \mid \theta_1)dx$ the probabilities of Type I and Type II errors for the test with rejection region $\mathcal{R}'$, and set $A := \mathcal{R}_\star \backslash \mathcal{R}'$ and $B := \mathcal{R}' \backslash \mathcal{R}_\star$.
Simplify notation by writing $A_i = \int_A f(x \mid \theta_i)dx$ and $B_i = \int_B f(x \mid \theta_i)dx$ for $i = 0, 1$. Since $c \le \Lambda(x)$ on A, then $cA_0 \le \int_A \Lambda(x) f_0(x)\,dx = A_1$; similarly, since $\Lambda(x) < c$ on B, then $B_1 = \int_B \Lambda(x) f_0(x)\,dx \le cB_0$, with strict inequality unless $B_0 = 0$.

$$\begin{aligned} c(\alpha' - \alpha_\star) &= \int_{\mathcal{R}'} cf(x \mid \theta_0)dx - \int_{\mathcal{R}_\star} cf(x \mid \theta_0)dx = cB_0 - cA_0 \ge cB_0 - A_1 \\ (\beta' - \beta_\star) &= \int_{\mathcal{R}'^c} f(x \mid \theta_1)dx - \int_{\mathcal{R}_\star^c} f(x \mid \theta_1)dx \;= A_1 - B_1 \quad > A_1 - cB_0. \end{aligned}$$

If $A_1 < cB_0$ then $\alpha' > \alpha_\star$, while if $A_1 \ge cB_0$ then $\beta' > \beta_\star$, or $A_1 = B_0 = 0$ and the tests coincide.

□

Thus, the *likelihood ratio test* is optimal for simple *vs.* simple hypothesis tests. Of course any monotone increasing function of Λ generates exactly the same test; often the logarithm of Λ is more convenient. It turns out that the test that rejects when $\Lambda \ge c$ achieves the minimal possible value of the weighted error combination $c\alpha(\delta) + \beta(\delta)$ over all possible tests δ. Thus, for example, if Type-I errors are twice as expensive as Type-II errors, one can achieve the minimum possible value of $2\alpha + \beta$ by rejecting when $\Lambda > 2$.

Proof of this claim:
For $\lambda > 0$ let $\alpha(\lambda) = \int_{\lambda \le \Lambda(x)} f(x \mid \theta_0)\,dx$ and $\beta(\lambda) = \int_{\Lambda(x) < \lambda} f(x \mid \theta_1)\,dx$ be the Type-I and Type-II error probabilities for the test that rejects $H_0 : \theta = \theta_0$ when $\Lambda \ge \lambda$. Fix $c > 0$. Then

$$\begin{aligned} \beta'(c) &= \lim_{\epsilon \to 0} [\beta(c+\epsilon) - \beta(c)]/\epsilon \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{c < \Lambda(x) \le c+\epsilon} f(x \mid \theta_1)\,dx \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{c < \Lambda(x) \le c+\epsilon} \Lambda(x) f(x \mid \theta_0)\,dx \\ &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{c < \Lambda(x) \le c+\epsilon} c\, f(x \mid \theta_0)\,dx \\ &= \lim_{\epsilon \to 0} [\alpha(c) - \alpha(c+\epsilon)]/\epsilon = -c\,\alpha'(c). \end{aligned}$$

Thus, the function $[c\alpha(\lambda) + \beta(\lambda)]$ achieves its minimum where $[c\alpha'(\lambda) + \beta'(\lambda)] = 0$, at $\lambda = c$.

2.1 LRT Examples

Poisson example: Suppose $\{X_1, \cdots, X_n\} \overset{\text{iid}}{\sim} \mathsf{Po}(\theta)$ with mean $\theta > 0$ known to be either θ_0 or θ_1. The logarithm of the Likelihood Ratio Statistic will be

$$\log \Lambda(x) = \log \frac{(\theta_1)^{\sum x_i} e^{-n\theta_1} / \prod x_i!}{(\theta_0)^{\sum x_i} e^{-n\theta_0} / \prod x_i!} = \Big[\sum x_i\Big] \log(\theta_1/\theta_0) + n(\theta_0 - \theta_1),$$

a monotone increasing function of the sufficient statistic $x_+ := \sum x_i$ if $\theta_1 > \theta_0$ and monotone decreasing function of x_+ if $\theta_1 < \theta_0$. The LRT rejection regions for x_+ will be

$$\theta_1 > \theta_0 : \quad \mathcal{R}_\star = [c, \infty) \qquad \text{or} \qquad \theta_1 < \theta_0 : \quad \mathcal{R}_\star = [0, c]$$

in these two cases, with a discrete set of possible sizes α_c and powers $[1 - \beta_c]$ given for $\theta_1 > \theta_0$ by

$\alpha_c =$ `ppois(c-1, n*theta.0, low=FALSE)` $[1 - \beta_c] =$ `ppois(c-1, n*theta.1, low=FALSE)`

and for $\theta_1 < \theta_0$ by

$\alpha_c =$ `ppois(c, n*theta.0, low=TRUE)` $[1 - \beta_c] =$ `ppois(c, n*theta.1, low=TRUE)`

for each possible integer $c \geq 0$. Typically the investigator begins by selecting a bound on α; then finds the value $c \in \mathbb{Z}_+$ to achieve a size α_c as close as possible to this; then computes the power $[1 - \beta_c]$ for this choice of c. If the power is inadequate, s/he must either increase the sample size n or accept a larger test size α_c.

Normal example: Suppose $\{X_1, \cdots, X_n\} \overset{\text{iid}}{\sim} \mathsf{No}(\theta, 1)$ with mean $\theta \in \mathbb{R}$ known to be either θ_0 or θ_1. The logarithm of the Likelihood Ratio Statistic will be

$$\begin{aligned}\log \Lambda(x) &= \log \frac{(2\pi)^{-n/2} \exp\Big(-\frac{1}{2} \sum (x_i - \theta_1)^2 \Big)}{(2\pi)^{-n/2} \exp\Big(-\frac{1}{2} \sum (x_i - \theta_0)^2 \Big)} = \tfrac{1}{2} \sum \big[(x_i - \theta_0)^2 - (x_i - \theta_1)^2 \big] \\ &= (\theta_1 - \theta_0) \sum x_i + (n/2)\big(\theta_0^2 - \theta_1^2\big),\end{aligned}$$

a monotone increasing function of the sufficient statistic $\bar{X}_n$ if $\theta_1 > \theta_0$ or a monotone decreasing function if $\theta_1 < \theta_0$. The type-I and type-II error probabilities for the test with critical region $\mathcal{R}_\star = [c, \infty)$ for $\theta_1 > \theta_0$ are

$$\alpha = \mathsf{P}_{\theta_0}\big[\bar{X}_n \geq c\big] = \Phi\big(\sqrt{n}(\theta_0 - c)\big) \qquad \beta = \mathsf{P}_{\theta_1}\big[\bar{X}_n < c\big] = \Phi\big(\sqrt{n}(c - \theta_1)\big)$$

Figure 1 illustrates this test with $\theta_0 = 1$ and $\theta_1 = 4$.

The symmetric solution with $\alpha = \beta$ will have $c = (\theta_0 + \theta_1)/2$, with

$$\alpha = \beta = \Phi\big(-\sqrt{n}|\theta_0 - \theta_1|/2 \big).$$

To ensure that both error probabilities will be smaller than 0.01, for example, will require a sample-size sufficiently large that $\sqrt{n}|\theta_0 - \theta_1|/2 > 2.326$, *i.e.*, $n \geq 21.65/(\theta_0 - \theta_1)^2$. Unsurprisingly, the sample-size needed to achieve a desired performance level will depend on how different the distribution of X is under the two hypotheses.

2.2 Randomized Tests

In the Normal example we saw we could construct an optimal test $\mathcal{R}_\star$ with *any* desired size $\alpha \in (0, 1)$, while in the Poisson example only a discrete set of possible values of α_c were possible. Some investigators find it disappointing that it may be impossible to construct a test of this form with precisely size $\alpha = 0.05$ (say) with discrete data. One way to fill this gap is to consider

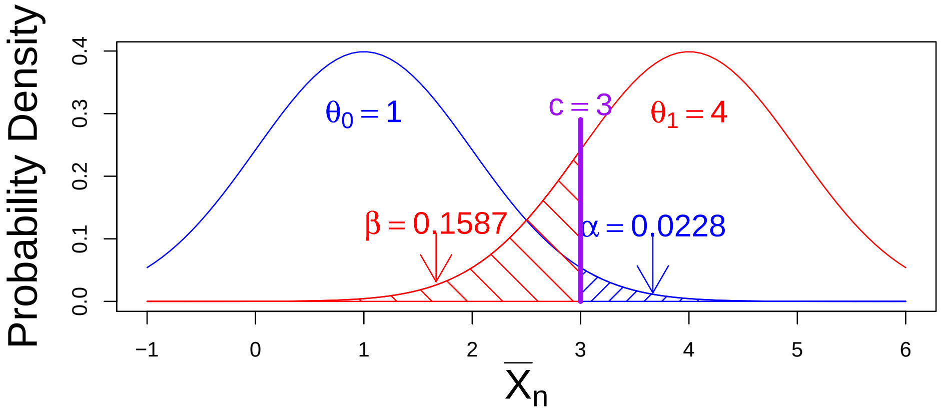

Figure 1: Illustration of test of $H_0: \ \theta = \theta_0$ vs. $H_1: \ \theta = \theta_1$, with rejection region $\mathcal{R} = \{\mathbf{x} : \bar{X}_n \geq c\}$. Size α is area under $f(x \mid \theta_0)$ to the right of c (densely shaded in blue); type-II error probability β is area under $f(x \mid \theta_2)$ to the left of c (sparsely shaded in red). Increasing c would increase β and decrease α.

randomized tests in which, upon observing $X = x$, we reject H_0 in favor of H_1 with probability $\phi(x)$ for some function $\phi : \mathcal{X} \to [0, 1]$. The Neyman/Pearson Lemma can be extended to show that for any $0 < \alpha < 1$ there exists an optimal randomized test ϕ of size α, of the form

$$\phi(x) = \begin{cases} 1 & \Lambda(x) > c \\ \gamma & \Lambda(x) = c \\ 0 & \Lambda(x) < c \end{cases}$$

for some $c > 0$ and $\gamma \in [0, 1]$. This test rejects H_0 if $\Lambda > c$, fails to reject if $\Lambda < c$, and decides randomly if $\Lambda = c$, as if to toss a biased coin with just the right probability γ of rejection to achieve size α. In problems with continuous distributions typically the event $\Lambda(X) = c$ has probability zero and this test is identical to the non-random one introduced earlier, but for discrete data they will differ unless α is one of the discrete set $\{\alpha_c\}$ of possible values. It's interesting theoretically that optimal tests of each size α exist and to know their form, but nobody uses such tests in practice and we won't discuss them further.

3 Generalized Likelihood Ratio Tests: GLRTs

3.1 Two-Sided Tests & Composite Hypotheses

No result quite like the Neyman/Pearson Lemma applies when H_0 or H_1 (or both) are *composite*, *i.e.*, contain more than one point of Θ. The most common example of this is a test of $\theta = \theta_0$ for some real parameter θ and specified $\theta_0 \in \Theta$ against the "two-sided alternative" of $\theta \neq \theta_0$.

Commonly used tests in the spirit of Neyman & Pearson are the *generalized likelihood ratio tests* based on the Generalized Likelihood Ratio (GLR) test statistic

$$\Lambda(x) := \frac{\sup_{\theta\in H_1} f(x\mid\theta)}{\sup_{\theta\in H_0} f(x\mid\theta)}.$$

Although the GLRT doesn't have the provable optimality of the LRT for simple *vs.* simple tests, it does lead to quite reasonable tests that are often applied in practice.

Warning: Some authors base their GLRTs on a slightly different GLR statistic,

$$\Lambda^*(x) := \frac{\sup_{\theta\in \Theta} f(x\mid\theta)}{\sup_{\theta\in H_0} f(x\mid\theta)} = \Lambda(x)\vee 1.$$

This is a little easier to compute than Λ, since the numerator is always $f(x\mid\hat\theta)$ for the MLE $\hat\theta$ of θ. It only differs from $\Lambda(x)$ for those x with $\hat\theta(x)\in H_0$— in which case $\Lambda(x)<1=\Lambda^*(x)$ and one wouldn't want to reject H_0 anyway, at any significance level below about one-half. For typical test sizes α, the two statistics Λ and Λ^* lead to *identical* tests. Another warning: some authors use the name "generalized likelihood ratio" for what we would call $1/\Lambda$ or $1/\Lambda^*$ (with H_0 in the numerator and H_1 or Θ in the denominator), leading to rejection regions of the form $\mathcal{R}=\{x:\ 1/\Lambda(x)\le c'\}$ for some $c'>0$, identical with our test with $\mathcal{R}=\{x:\ \Lambda(x)\ge c\}$ for $c=1/c'$.

3.2 GLRT Examples

Poisson example: Suppose $\{X_1,\cdots,X_n\}\overset{\text{iid}}{\sim}\mathsf{Po}(\theta)$ with uncertain mean $\theta>0$. A GLRT test of the hypothesis $H_0:\ \theta=\theta_0$ against the *two-sided* alternative $H_0:\ \theta\ne\theta_0$ will reject for large values of the log GLR statistic

$$\begin{aligned}\log\Lambda(x) &:= \log\frac{\sup_{\theta\ne\theta_0} f(x\mid\theta)}{\sup_{\theta=\theta_0} f(x\mid\theta)}\\ &= n\Big[\bar X_n\log(\bar X_n/\theta_0)+(\theta_0-\bar X_n)\Big],\end{aligned}$$

since the likelihood on $\Theta_1=\{\theta:\ \theta\ne\theta_0\}$ achieves its maximum at the MLE $\hat\theta_n=\bar X_n$. This will reject for both large values of $\bar X_n$ (where the first term dominates) and for small values of $\bar X_n$ (where the second one does). For large n a Taylor series expansion shows that the log GLR statistic is approximately

$$\log\Lambda(x)\approx\frac{n}{2}\big(\bar X_n-\theta_0\big)^2/\theta_0^2,$$

so the GLRT will reject for outcomes with $|\bar X_n-\theta_0|\ge c$ with $\alpha\approx 2\Phi\big(-c\sqrt{n}\big)$ just as one might expect in light of the central limit theorem.

For composite alternatives it's best to think of "power" (the probability of rejecting H_0 when it is false) as a *function* $[1-\beta(\theta)]$ of possible values $\theta\in\Theta_1$— the probability of rejecting $H_0:\theta=\theta_0$, when in fact the parameter's value is θ. This will show *how far* θ must be from θ_0 to give a good chance of discovering that H_0 is false. In the Poisson example, a GLRT with rejection region $\mathcal{R}=\{x:\ x_+\in[0,L]\cup[R,\infty)\}$ for $0\le L<R<\infty$ with $L,R\in\mathbb{Z}_+$, will have power function

$$[1-\beta(\theta)] = \texttt{ppois(L,n*theta)}+\texttt{ppois(R-1,n*theta,low=FALSE)}$$

which will attain a minimum value of about α near $\theta = \theta_0$ and increase as θ increases or decreases from that value.

Normal example 1: The One-sided Z Test. Suppose $\{X_1, \cdots, X_n\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu, \sigma^2)$ with unknown mean $\mu \in \mathbb{R}$ but known variance σ^2. The hypothesis "$H_0 : \mu \le \mu_0$" and alternate hypothesis "$H_1 : \mu > \mu_0$" are both composite. The numerator of the GLR statistic Λ will be

$$\text{Numerator} \;= \sup_{\mu > \mu_0} (2\pi\sigma^2)^{-n/2} \exp\big(-\frac{1}{2\sigma^2}\big[S + n(\bar{X}_n - \mu)^2\big]\big)$$

where $S := \sum (X_i - \bar{X}_n)^2$. This will take different values, depending on whether $\bar{X}_n \le \mu_0$ (in which case the maximum occurs at $\mu = \mu_0$) or $\bar{X}_n > \mu_0$ (in which case the maximum occurs at $\mu = \bar{X}_n$):

$$\text{Numerator} \;= (2\pi\sigma^2)^{-n/2} e^{-S/2\sigma^2} \times \begin{cases} e^{-(n/2\sigma^2)(\bar{X}_n - \mu_0)^2} & \bar{X}_n \le \mu_0 \\ 1 & \bar{X}_n > \mu_0. \end{cases}$$

Similarly the denominator takes different values:

$$\text{Denominator} \;= (2\pi\sigma^2)^{-n/2} e^{-S/2\sigma^2} \times \begin{cases} 1 & \bar{X}_n \le \mu_0 \\ e^{-(n/2\sigma^2)(\bar{X}_n - \mu_0)^2} & \bar{X}_n > \mu_0. \end{cases}$$

The GLR is just the ratio. After some cancellation,

$$\Lambda = \begin{cases} e^{-(n/2\sigma^2)(\bar{X}_n - \mu_0)^2} & \bar{X}_n \le \mu_0 \\ e^{(n/2\sigma^2)(\bar{X}_n - \mu_0)^2} & \bar{X}_n > \mu_0 \end{cases}$$

Taking logs and removing constant factors,

$$\log \Lambda \propto (\bar{X}_n - \mu_0)|\bar{X}_n - \mu_0|,$$

so the GLRT rejects for large values of $\bar{X}_n$ or, equivalently, of $Z := \sqrt{n}(\bar{X}_n - \mu_0)/\sigma$, which has the $\mathsf{No}(0,1)$ distribution under H_0. Thus, the test will reject H_0 at level α if $Z > z_\alpha$ or $\bar{X}_n > \mu_0 + \sigma z_\alpha/\sqrt{n}$, where $\Phi(z_\alpha) = 1 - \alpha$. The power (*i.e.*, probability of rejecting H_0) of this test for arbitrary $\mu \in \mathbb{R}$ is

$$\begin{aligned}[1 - \beta_+(\mu)] &= \mathsf{P}_\mu(\bar{X}_n > \mu_0 + \sigma z_\alpha/\sqrt{n}) \\ &= \mathsf{P}_\mu\Big(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} > \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} + z_\alpha\Big) \\ &= \Phi\big(-z_\alpha + \sqrt{n}(\mu - \mu_0)/\sigma\big),\end{aligned}$$

equal to α at $\mu = \mu_0$ but increasing to one as $\mu \to \infty$ or $\sigma \to 0$ or $n \to \infty$. This is illustrated as the red dashed curve in Figure (2).

Similarly, the one-sided test of $H_0 : \; \mu = \mu_0$ against the other one-sided alternative $H_1 : \; \mu < \mu_0$ has power function

$$\begin{aligned}[1 - \beta_-(\mu)] &= \mathsf{P}_\mu(\bar{X}_n < \mu_0 - \sigma z_\alpha/\sqrt{n}) \\ &= \mathsf{P}_\mu\Big(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} < \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} - z_\alpha\Big) \\ &= \Phi\big(-z_\alpha + \sqrt{n}(\mu_0 - \mu)/\sigma\big),\end{aligned}$$

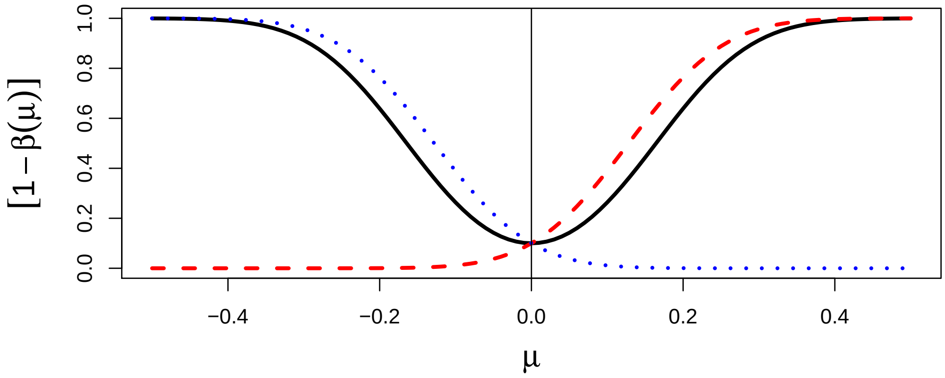

Figure 2: Illustration of power $[1-\beta(\mu)]$ for GLR test of $H_0:\ \mu=0$ against two-sided alternative $H_1:\ \mu\neq 0$ (black solid curve) and one-sided alternatives $H_1:\ \mu>0$ (red dashed) and $H_1:\ \mu<0$ (blue dotted), to illustrate non-existence of a UMP test for two-sided hypotheses. All three take common value $\alpha=0.10$ at $\mu_0=0$.

illustrated as the blue dotted curve in Figure (2).

The GLR test of $H_0:\ \mu=\mu_0$ of size α against the two-sided alternative $H_1:\ \mu\neq\mu_0$ will reject for large values of $|\bar X_n-\mu_0|$, which may be achieved in two different ways: $\bar X_n\gg\mu_0$ and $\bar X_n\ll\mu_0$. Their probabilities must be summed to find the power function:

$$\begin{aligned}[1-\beta(\mu)] &= \mathsf{P}_\mu\left\{|\bar X_n-\mu_0|>z_{\alpha/2}\sigma/\sqrt{n}\right\}\\ &= \mathsf{P}_\mu\left\{\sqrt{n}(\bar X_n-\mu)/\sigma<-z_{\alpha/2}-\sqrt{n}(\mu-\mu_0)/\sigma\right\}\\ &+ \mathsf{P}_\mu\left\{\sqrt{n}(\bar X_n-\mu)/\sigma>z_{\alpha/2}-\sqrt{n}(\mu-\mu_0)/\sigma\right\}\\ &= \Phi\big(-z_{\alpha/2}+\sqrt{n}(\mu-\mu_0)/\sigma\big)+\Phi\big(-z_{\alpha/2}-\sqrt{n}(\mu-\mu_0)/\sigma\big).\end{aligned}$$

Figure (2) shows a plot of this power function (the black solid curve), showing that it is not as powerful as the one-sided test against $H_1:\ \mu>\mu_0$ for $\mu>\mu_0$ (red dashed curve), nor as powerful as the one-sided test against $H_1:\ \mu<\mu_0$ for $\mu<\mu_0$ (blue dotted curve). None of these is *uniformly* most powerful, since each is dominated by each of the others at some points μ. This *always* happens for two-sided tests, where quite generally no UMP test exists.

Normal example 2: The One-sided t Test. Suppose we still wish to test $H_0:\mu\le\mu_0$ against $H_1:\mu>\mu_0$, but the variance σ^2 is *not* known. Now the suprema in the numerator and denominator of Λ must each be taken over both μ and σ^2, leading to

$$\text{Numerator } = \begin{cases} \big(2\pi[S/n + (\bar{X}_n - \mu_0)^2]\big)^{-n/2} e^{-n} & \bar{X}_n \le \mu_0 \\ \big(2\pi S/n\big)^{-n/2} e^{-n} & \bar{X}_n > \mu_0 \end{cases}$$

$$\text{Denominator } = \begin{cases} \big(2\pi S/n\big)^{-n/2} e^{-n} & \bar{X}_n \le \mu_0 \\ \big(2\pi[S/n + (\bar{X}_n - \mu_0)^2]\big)^{-n/2} e^{-n} & \bar{X}_n > \mu_0 \end{cases}$$

$$\Lambda = \begin{cases} \big(1 + n(\bar{X}_n - \mu_0)^2/S\big)^{-n/2} & \bar{X}_n \le \mu_0 \\ \big(1 + n(\bar{X}_n - \mu_0)^2/S\big)^{+n/2} & \bar{X}_n > \mu_0 \end{cases},$$

(where again $S := \sum(X_i - \bar{X}_n)^2$), so the GLRT rejects for large values of

$$t := \frac{\bar{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n-1}}$$

where $\hat{\sigma}^2 := S/n$ is the MLE. The statistic t has the t_{n-1} distribution under H_0.

Normal example 3: The Two-sided t Test. Suppose $\{X_1, \cdots, X_n\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu, \sigma^2)$ with uncertain mean $\mu \in \mathbb{R}$ and (again) uncertain variance $\sigma^2 \in \mathbb{R}_+$. Now the hypothesis "$\mu = \mu_0$" is composite, consisting of all $\theta = (\mu, \sigma^2)$ in $H_0 = \{\theta = (\mu_0, \sigma^2) : \ \sigma^2 \ge 0\}$ in $\Theta = \mathbb{R} \times \mathbb{R}_+ = \{\theta = (\mu, \sigma^2)\}$. The alternate hypothesis $H_1 = \{\theta = (\mu, \sigma^2) : \ \mu \ne \mu_0, \ \sigma^2 \ge 0\}$ is also composite.

The supremum of the $f(x \mid \theta)$ over H_1, the numerator for Λ, will be attained at the joint MLE $\hat{\theta}_n = (\bar{X}_n, \hat{\sigma}^2)$ given by the familiar formulas

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

The supremum of the $f(x \mid \theta)$ over H_0 will be attained at the point $(\mu_0, \hat{\sigma}_0^2)$, with mean component $\mu = \mu_0$ and variance component $\sigma^2 = \hat{\sigma}_0^2$ given by

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2 = \hat{\sigma}^2 + (\bar{X}_n - \mu_0)^2.$$

Thus, the GLR statistic is

$$\begin{aligned} \Lambda(x) &= \frac{(2\pi\hat{\sigma}^2)^{-n/2} \exp(-n/2)}{(2\pi[\hat{\sigma}^2 + (\bar{X}_n - \mu_0)^2])^{-n/2} \exp(-n/2)} \\ &= [1 + (\bar{X}_n - \mu_0)^2/\hat{\sigma}^2]^{n/2}, \end{aligned}$$

a monotone increasing function of $|t|$ for the statistic

$$t := \frac{\bar{X}_n - \mu_0}{\hat{\sigma}/\sqrt{n-1}}.$$

We'll see later in Section (A) that t is a "pivotal quantity" whose sampling probability distribution, called "Student's t with $\nu = (n-1)$ degrees of freedom", depends on ν but not on $\theta = (\mu, \sigma^2)$, so a test of any desired size $\alpha > 0$ can be constructed by rejecting $H_0 : \ \mu = \mu_0$ when $|t| > t_c$ where t_c is the $(1 - \alpha/2)$th quantile of the t_ν distribution.

For large ν the t_ν distribution is nearly identical to the $\mathsf{No}(0, 1)$ distribution, so for n larger than 20 or 30 or so the GLRT for $\mu = \mu_0$ is approximately the same as the LRT test described earlier.

3.3 One-sample and Two-sample t-tests

3.3.1 Two-sample t tests

If $\{X_i\}_{1\le i\le n} \overset{\text{iid}}{\sim} \mathsf{No}(\mu_1,\sigma^2)$ and $\{Y_j\}_{1\le j\le m} \overset{\text{iid}}{\sim} \mathsf{No}(\mu_2,\sigma^2)$ are random samples from normally-distributed populations with the same variance, the hypothesis $H_0 : \mu_1 = \mu_2$ that the two populations are actually identical is composite because it does not specify the common value μ of μ_1 and μ_2. For known σ^2, the GLR for this test is

$$\Lambda = \frac{\sup_{\mu_1,\mu_2}(2\pi\sigma^2)^{-(n+m)/2}\exp\big(-\frac{1}{2\sigma^2}\big[S_1+n(\bar X_n-\mu_1)^2+S_2+m(\bar Y_m-\mu_2)^2\big]\big)}{\sup_{\mu}(2\pi\sigma^2)^{-(n+m)/2}\exp\big(-\frac{1}{2\sigma^2}\big[S_1+n(\bar X_n-\mu)^2+S_2+m(\bar Y_m-\mu)^2\big]\big)}$$

where $S_1 := \sum(X_i-\bar X_n)^2$ and $S_2 := \sum(Y_j-\bar Y_m)^2$. For fixed σ, the supremum in the numerator is attained at $\hat\mu_1 = \bar X_n$ and $\hat\mu_2 = \bar Y_m$, while in the denominator it is attained at the common value $\hat\mu = (n\bar X_n + m\bar Y_m)/(n+m)$, the weighted average of the sample means. Since $n(\bar X_n-\hat\mu)^2 + m(\bar Y_m-\hat\mu)^2 = nm(\bar X_n-\bar Y_m)^2/(n+m)$, this is

$$= \frac{(\sigma^2)^{-(n+m)/2}\exp\big(-\frac{1}{2\sigma^2}\big[S_1+S_2\big]\big)}{(\sigma^2)^{-(n+m)/2}\exp\big(-\frac{1}{2\sigma^2}\big[S_1+S_2+nm(\bar X_n-\bar Y_m)^2/(n+m)\big]\big)} \tag{1}$$

For known σ^2, this is a monotone increasing function of $|Z|$ for the Z statistic

$$Z := \frac{\bar X_n-\bar Y_m}{\sigma\sqrt{1/n+1/m}},$$

which has the $\mathsf{No}(0,1)$ distribution under H_0.

For *unknown* σ^2 we must maximize (separately) the numerator and denominator in Eqn (1) with respect to σ^2. The supremum in the numerator is attained at $\hat\sigma_1^2 = (S_1+S_2)/(n+m)$, while that in the denominator is attained at $\hat\sigma_0^2 = [S_1+S_2+nm(\bar X_n-\bar Y_m)^2/(n+m)]/(n+m)$, so

$$\begin{aligned}\Lambda &= \left[\frac{(S_1+S_2)/(n+m)}{[S_1+S_2+nm(\bar X_n-\bar Y_m)^2/(n+m)]/(n+m)}\right]^{-(n+m)/2}\\ &= \left[1+\frac{\frac{nm}{n+m}(\bar X_n-\bar Y_m)^2}{S_1+S_2}\right]^{(n+m)/2},\end{aligned}$$

a monotone increasing function of $|t|$ for the t statistic

$$t = \frac{(\bar X_n-\bar Y_m)\sqrt{nm/(n+m)}}{\sqrt{(S_1+S_2)/(n+m-2)}}.$$

Under H_0, $Y := (S_1+S_2)/(n+m-2)$ is σ^2 times a χ^2_ν random variable with $\nu = (n+m-2)$ degrees of freedom, while $Z := (\bar X_n-\bar Y_m)\sqrt{\frac{nm}{n+m}}$ has a normal distribution with mean zero and variance σ^2, so $t = Z/\sqrt{Y/\nu}$ has a t_ν distribution with $\nu = (n+m-2)$ degrees of freedom.

The problem of testing $H_0 :\ \mu_1 = \mu_2$ for normal populations with variances σ_1^2 and σ_2^2 that are *not* known to have a common value σ^2 is more problematic (unless σ_1^2 and σ_2^2 are both known, of course). Look up the “Behrens-Fisher problem” if you’d like to see more about that.

3.3.2 Paired or one-sample t tests

If we observe *n pairs* $\{(X_i, Y_i)\}$ of random variables whose differences $d_i = (X_i - Y_i)$ are normally-distributed with unknown mean μ_d and variance σ_d^2, we can use the one-sample t test of the hypothesis $H_0 \ : \mu_d = 0$ against alternative $H_1 \ : \mu_d \neq 0$, evaluating the statistic

$$t := \frac{\bar{d}_n}{\sqrt{\hat{\sigma}_d^2/(n-1)}} = \frac{\bar{X}_n - \bar{Y}_n}{\sqrt{\hat{\sigma}_d^2/(n-1)}} = \frac{\bar{d}_n\sqrt{n(n-1)}}{\sqrt{\sum_{i=1}^n (d_i - \bar{d}_n)^2}}$$

with $\hat{\sigma}_d^2 := \sum (d_i - \bar{d}_n)^2/n$ and rejecting H_0 at level α if $|t| >$ `qt(1-alpha/2, df=n-1)` or reporting a P-value of $P =$ `2*pt(abs(t), df=n-1,lower=F)`.

If all the X_i's and Y_j's are independent, and all normally distributed with the same unknown variance σ^2, then the Two-Sample t test described above would also be valid for paired data and would be more powerful because its estimate of σ^2 would be based on a chi-squared estimator with $2n - 2$ degrees of freedom, while the paired t test has only $n - 1$ degrees of freedom. However, if the pairs (X_i, Y_i) have different means or have positive correlation, then the variance σ_d^2 will be (perhaps much) smaller than $2\sigma^2$, and the paired-sample t test will be more powerful. For example, if (X_i, Y_i) represents the mileages for the ith car on two different trips, or responses of the ith subject to two different drug treatments, or sunburn scores on left and right arms of ith subject in a sunscreen test, then one would expect the variation of the differences $d_i = (X_i - Y_i)$ to be substantially smaller than the differences among the $\{X_i\}$ or $\{Y_j\}$. This is an example of what is called “blocking” in general linear models, the systematic elimination of widely variable aspects of a dataset to strengthen the power of tests or precision of estimates.

3.4 MLR Tests are UMP

Simple hypotheses with simple alternatives are *almost* the only setting where there exists an optimal test $\mathcal{R}_\star$ that is “uniformly most powerful” (UMP) in the sense that it has higher power $[1 - \beta(\theta)]$ for all θ than *any* competing test $\mathcal{R}$ of the same size α.

One other setting where a Uniformly Most Powerful test exists is when $\Theta \subset \mathbb{R}$ and, for any $\theta_1 < \theta_2$ in Θ, the likelihood ratio

$$\Lambda_{21}(x) := \frac{f(x \mid \theta_2)}{f(x \mid \theta_1)}$$

is a monotone increasing function of some statistic $T(x)$. In that case for *any* $\theta_0 \in \Theta$ a test of the simple hypothesis $H_0 : \ \theta = \theta_0$ or the composite hypothesis $H_0 : \ \theta \leq \theta_0$ against the composite alternative $H_1 : \ \theta > \theta_0$ would reject H_0 upon observing $X \in \mathcal{R}_\star = \{x : \ T(X) \geq c\}$ for some critical value $c > 0$, so the LRT is UMP for these one-sided composite hypotheses.

Note *no UMP test exists* for most testing situations, including

- any two-sided testing situation: in MLR situations, the corresponding one-sided tests are each most powerful for some θ's, and no test can match both of them for all θ's (see Figure (2)).
- some one-sided testing situations, such as $H_0 : m \leq m_0$ vs. $H_1 : m > m_0$ for the centrality parameter m of the Cauchy distribution.

3.5 Confidence Sets & Hypothesis Tests

If $C(\mathbf{x})$ is a $100\gamma\%$ confidence set for θ, then the set $\mathcal{R} := \{\mathbf{x} \in \mathcal{X} : \theta_0 \notin C(\mathbf{x})\}$ is the rejection region for a test of size $\alpha = (1-\gamma)$ of the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. For example, if $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\theta, \sigma^2)$ with σ^2 known, then for any $0 < \alpha < 1$ the confidence interval

$$C(\mathbf{x}) := [\bar{X}_n - z_{\alpha/2}\,\sigma/\sqrt{n},\ \bar{X}_n + z_{\alpha/2}\,\sigma/\sqrt{n}]$$

is a $1-\alpha$ confidence interval for θ if $z_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the $\mathsf{No}(0,1)$ distribution, and the test that rejects $H_0 : \theta = \theta_0$ when

$$\mathbf{x} \in \mathcal{R} := \{\mathbf{x} : \theta_0 \neq C(\mathbf{x})\}$$

is a valid test of size α of $H_0 : \theta = \theta_0$ against the two-sided alternative $H_1 : \theta \neq \theta_0$. Similarly, if σ^2 is *not* known, the test that rejects H_0 when $\mathbf{x} \in \mathcal{R} := \{\mathbf{x} : \theta_0 \neq C(\mathbf{x})\}$ is a valid size-α test for the CI

$$C(\mathbf{x}) := [\bar{X}_n - t_{\alpha/2}\,\hat{\sigma}/\sqrt{n-1},\ \bar{X}_n + t_{\alpha/2}\,\hat{\sigma}/\sqrt{n-1}],$$

if $t_{\alpha/2}$ is the $(1-\alpha/2)$th quantile of the t_ν distribution with $\nu = n-1$ degrees of freedom. In fact, these are in both cases the same GLRTs we constructed in Section (3.2). The P-value, in contrast, reports how unlikely even more extreme data than those we observed would be *if* the hypothesis were true.

Reporting these confidence intervals gives more information than reporting only the accept/reject decision of a test, in a way that does not lend itself to misinterpretation like reporting P-values does. Rejecting the hypothesis $\theta = 10$ at level $\alpha = 0.01$ because $10 \notin C(\mathbf{x}) = [11, 100]$ is different from rejecting the same hypothesis at the same level because $10 \notin C(\mathbf{x}) = [95, 100]$— a marginal rejection in the first case, an emphatic one in the second, and in both case the CI gives an indication of *how far* the data is from supporting the hypothesis.

3.6 Q & D Two-sided Tests: The Wald Test

Under mild regularity conditions the MLE $\hat{\theta}_n$ of a real parameter $\theta \in \Theta \subset \mathbb{R}$ on the basis of a sample $\{X_i\}_{1 \le i \le n} \sim f(x \mid \theta)$ will be asymptotically normal, with

$$\sqrt{n}[\hat{\theta}_n - \theta] \approx \mathsf{No}\big(0, I(\theta)^{-1}\big)$$

for the Fisher information $I(\theta)$. In this case the two-sided hypothesis test of $H_0 : \theta = \theta_0$ *vs.* $H_1 : \theta \neq \theta_0$ can be tested at asymptotic level $\alpha > 0$ by rejecting whenever the statistic $T(x) := nI(\theta_0)[\hat{\theta}_n - \theta_0]^2$ exceeds the $(1-\alpha)$th quantile of the χ^2_1 distribution (*i.e.*, exceeds $|z_{\alpha/2}|^2$ for $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$). Similarly for $\Theta \subset \mathbb{R}^k$, the test statistic $T(x) := n[\hat{\theta}_n - \theta_0]^\intercal I(\theta_0)[\hat{\theta}_n - \theta_0]$ will have an asymptotic χ^2_k distribution if $H_0 : \theta = \theta_0$ is true, leading to a simple test of H_0 against the omnibus alternative $H_1 : \theta \neq \theta_0$.

4 Contemporary Scientific (mis-)Use of Hypothesis & Significance Tests

The most common contemporary “classical” or sampling-theory based approach to hypothesis testing is a bit of a mish-mash between the Fisher and Neyman/Pearson approaches. Commonly a

P-value is computed (using a likelihood ratio or GLR statistic if possible) and reported, and then the hypothesis is rejected at some nominal level α (often $\alpha = 0.05$ or 0.01) that exceeds P.

While a valid test of size α can be performed by selecting α *before observing the data*, and then computing the P value and rejecting H_0 if $P \le \alpha$ (with implicit rejection region $\mathcal{R} = \{x : P(x) \le \alpha\}$), the selection of test size α *after* computing the P value leads to serious distortion of the strength of evidence. This common practice leads to a violation of the "frequentist guarantee" that no more than 5% of rejection decisions made at level $\alpha = 0.05$ will be incorrect, no more than 1% of rejection decisions made at level $\alpha = 0.01$, *etc.*

Although it's true that only 5% of true hypotheses will have $P(X) \le 0.05$, and that rejecting H_0 at level $\alpha = 0.05$ whenever $P(X) \le 0.05$ will be a valid test at this level, those rejected hypotheses will include *both* those with $P(X)$ far below 0.05 (of which many fewer than 5% are true) *and* those with $P(X)$ close to 0.05 (of which far more than 5% are true). Thus the common contemporary practice of treating $P(X) \approx 0.05$ as rejection at level $\alpha = 0.05$, $P(X) \approx 0.01$ as rejection at level $\alpha = 0.01$, *etc.*, grossly overstates the strength of evidence in experiments.

See (Berger and Sellke, 1987; Sellke et al., 2001) for a nice discussion of how as many as 23% of hypotheses rejected at level $\alpha = 0.05$ using the contemporary approach are in fact true, (Goodman, 2001) for a colorful discussion, and (Wasserstein and Lazar, 2016) for a new statement by the American Statistical Association on the issue.

5 Bayesian Hypothesis Testing

Because the P-value is a number between zero and one that is large when H_0 seems plausible and small when it is rejected, it is a common but serious error to misinterpret P as the *probability* that H_0 is true. Don't make that mistake! The only way to compute the *probability that H_0 is true* is to construct a probability model on the space Θ of model parameters— that is, to adopt a Bayesian approach.

The Bayesian approach to considering hypotheses $H_0 \subset \Theta$ upon observing $X = x$ is simple— report $\mathsf{P}[H_0 \mid X = x]$, the posterior probability that the hypothesis is true given the observed data. This is precisely what most investigators want to know from an experiment: in light of the observation $X = x$, is H_0 true, or not? Fisher's P-value gives a very different probability— *if* the hypothesis H_0 is true, *then* $P(x)$ measures how unlikely it is that one would observe data as seemingly unfavorable to H_0 as that represented by x, or more so.

Unfortunately there is no known simple and effective approach to finding prior distributions for an "objective" Bayesian approach to hypothesis testing, as there is for parameter estimation. Improper prior distributions (such as Jeffreys' Rule and Reference priors, for most problems) lead to indeterminate results and cannot be used.

5.1 Simple Hypotheses: Bayes Factors

For the special case of simple *vs.* simple hypotheses however there is an easy way to separate the role of the prior from the role of the data in determining $\mathsf{P}[H_0 \mid X = x]$, that does lead to a completely objective way to assess hypotheses. For any parametric statistical model $\mathcal{F} = \big\{f(x \mid \theta) : \theta \in \{\theta_0, \theta_1\}\big\}$ with only two possible parameter values, each prior distribution π on $\Theta = \{\theta_0, \theta_1\}\}$

is determined by either of the numbers $\pi_0 = \pi(\theta_0)$ or $\pi_1 = (1-\pi_0) = \pi(\theta_1)$. The posterior probabilities of these two points after observing $X = x$ are

$$\begin{aligned}\pi_1(x) = \pi(\theta_1 \mid X = x) &= \frac{\pi_1 f(x \mid \theta_1)}{\pi_0 f(x \mid \theta_0) + \pi_1 f(x \mid \theta_1)}\\ \pi_0(x) = \pi(\theta_0 \mid X = x) &= \frac{\pi_0 f(x \mid \theta_0)}{\pi_0 f(x \mid \theta_0) + \pi_1 f(x \mid \theta_1)}.\end{aligned}$$

The ratio of these probabilities, the posterior *odds* against $H_0 : \ \theta = \theta_0$, is the product

$$\frac{\pi_1(x)}{\pi_0(x)} = \left\{\frac{\pi_1}{\pi_0}\right\} \ \left\{\frac{f(x \mid \theta_1)}{f(x \mid \theta_0)}\right\}$$

of the prior odds (π_1/π_0) against H_0 and the same likelihood ratio $\Lambda(x)$ that appeared in the Neyman/Pearson lemma. In this context $\Lambda(x)$ is called the *Bayes factor* against H_0. It constitutes all the evidence from the data about H_0, and with it the posterior probability of H_0

$$\mathsf{P}[\theta = \theta_0 \mid X = x] = \frac{\pi_0 f(x \mid \theta_0)}{\pi_0 f(x \mid \theta_0) + \pi_1 f(x \mid \theta_1)} = \frac{1}{1 + (\pi_1/\pi_0)\Lambda(x)}$$

can be computed for any prior odds (π_1/π_0).

Binomial example: Suppose $X \sim \mathsf{Bi}(n, \theta)$ with a success probability θ known to be one of two possible values, θ_0 or θ_1. The likelihood ratio against $H_0 : \ \theta = \theta_0$ upon observing $X = x$ is

$$\Lambda(x) = \frac{\binom{n}{x}(\theta_1)^x(1-\theta_1)^{n-x}}{\binom{n}{x}(\theta_0)^x(1-\theta_0)^{n-x}} = \left(\frac{\theta_1(1-\theta_0)}{(1-\theta_1)\theta_0}\right)^x \left(\frac{1-\theta_1}{1-\theta_0}\right)^n,$$

which ranges from $\left(\frac{1-\theta_1}{1-\theta_0}\right)^n$ to $\left(\frac{\theta_1}{\theta_0}\right)^n$ as x ranges from 0 to n, with posterior probability

$$\mathsf{P}[H_0 \mid \mathbf{x}] = \frac{1}{1 + \frac{\pi_1}{\pi_0}\Lambda(\mathbf{x})}$$

of $H_0 : \theta = \theta_0$.

Normal example: Suppose $\{X_j\} \overset{\text{iid}}{\sim} \mathsf{No}(\theta, 1)$ with a mean θ known to be one of two possible values, θ_0 or θ_1. The likelihood ratio against $H_0 : \ \theta = \theta_0$ upon observing $\mathbf{x} = \{X_1, \cdots, X_n\}$ is

$$\begin{aligned}\Lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2}\exp\left\{-S/2 - n(\bar{X}_n - \theta_1)^2/2\right\}}{(2\pi)^{-n/2}\exp\left\{-S/2 - n(\bar{X}_n - \theta_0)^2/2\right\}}\\ &= \exp\left\{n(\theta_1 - \theta_0)(\bar{X}_n - \tfrac{\theta_0+\theta_1}{2})\right\},\end{aligned}$$

where $S := \sum(X_j - \bar{X}_n)^2$, equal to one at $\bar{X}_n = \frac{\theta_0+\theta_1}{2}$ and tending to infinity as $\bar{X}_n$ tends to $+\infty$ if $\theta_1 > \theta_0$, or to zero if $\theta_1 < \theta_0$. With prior probabilities π_0 and π_1 for H_0 and $H_1 : \ \theta = \theta_1$, the posterior probability of the null hypothesis is

$$\mathsf{P}[H_0 \mid \mathbf{x}] = \frac{1}{1 + \frac{\pi_1}{\pi_0}\Lambda(\mathbf{x})}.$$

5.2 Bayes Tests of Composite Hypotheses

Bayesian tests for composite hypotheses, even for seemingly simple cases like $H_0 = \{\theta_0\}$ *vs.* $H_1 = \{\theta \in \Theta : \ \theta \neq \theta_0\}$, are more troublesome because the prior distribution must specify not only the prior probability that $\theta = \theta_0$ but also the conditional distribution of θ when H_0 is false, which is difficult to do in an objective way. The usual "objective" prior distributions recommended for Bayesian parameter estimation, such as Jeffreys rule priors and reference priors, are often improper with an arbitrary scale constant. In estimation that arbitrary scale cancels when computing the posterior distribution, and causes no harm; in hypothesis testing, since the constant appears in only one of the two prior distributions under consideration, it does not cancel and so improper prior distributions cannot be used in testing hypotheses.

In some problems it is possible to consider simultaneously a broad class of alternate distributions—such as all unimodal distributions centered at θ_0, or all distributions within some parametric class—and derive *bounds* for the highest value possible for the likelihood ratio Λ against H_0, for that class. This in turn leads to bounds for the lowest value possible for $\mathsf{P}[H_0 \mid \mathbf{x}]$, meaningful limits on how strong is the evidence against H_0.

Binomial example: Suppose $X \sim \mathsf{Bi}(n, \theta)$ and, for some fixed $\theta_0 \in \Theta = (0, 1)$, consider the hypothesis $H_0 = \{\theta_0\}$ that the success probability is the specific number θ_0. To complete a Bayesian evaluation of $\mathsf{P}[H_0 \mid X = x]$ we need to specify two more things:

1. The prior probability π_0 of H_0 (the conventional choice is 1/2);
2. The prior probability distribution $\pi_1(\theta)$ of θ under H_1 (the conventional choice is $\mathsf{Un}(\Theta)$).

With these two conventional choices, the posterior probability of H_0 is

$$\begin{aligned}\mathsf{P}[H_0 \mid X = x] &= \frac{\pi_0 f(x \mid \theta_0)}{\pi_0 f(x \mid \theta_0) + (1 - \pi_0) \int_\Theta f(x \mid \theta) \pi_1(\theta)\, d\theta} = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \int_0^1 f(x \mid \theta)\, d\theta} \\ &= \frac{\theta_0^x (1 - \theta_0)^{n-x}}{\theta_0^x (1 - \theta_0)^{n-x} + \Gamma(x+1)\Gamma(n-x+1)/\Gamma(n+2)}\end{aligned}$$

and the posterior odds against H_0 are

$$\frac{\mathsf{P}[H_1 \mid X = x]}{\mathsf{P}[H_0 \mid X = x]} = \left\{ \frac{x!\ (n-x)!}{(n+1)!\ \theta_0^x (1 - \theta_0)^{n-x}} \right\}$$

For example, with $x = 7$ successes in $n = 10$ tries the posterior odds against the hypothesis $H_0 : \theta = \frac{1}{2}$ would be $128/165 \approx 0.776$, and with $x = 9$ successes it would be $512/55 \approx 9.31$, suggesting only modest evidence against H_0 in either case, since $\mathsf{P}[H_0 \mid x] = 1/(1 + 9.31) \approx 0.097$ isn't tiny.

A bit more generally we may take any $\pi_0 \in (0, 1)$ and $\pi_1(\theta) \sim \mathsf{Be}(\alpha, \beta)$ for any $\alpha, \beta > 0$ and find posterior odds

$$\frac{\mathsf{P}[H_1 \mid X = x]}{\mathsf{P}[H_0 \mid X = x]} = \left\{ \frac{1 - \pi_0}{\pi_0} \right\} \left\{ \frac{\Gamma(\alpha + x)\Gamma(\beta + n - x)\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)\Gamma(\alpha)\Gamma(\beta)\theta_0^x (1 - \theta_0)^{n-x}} \right\}.$$

As in the simple *vs.* simple this separates the prior odds $(1 - \pi_0)/\pi_0$ against H_0 from the data's contribution (the Bayes factor), but now the latter depends to some degree on α and β.

For example, with $X = 7$ successes in $n = 10$ tries, a test of $H_0: \ \theta = 1/2$ has a *maximum* posterior odds against H_0 of 1.07 for $\pi_0 = 1/2$ and any symmetric Beta distribution $\pi_1(\theta) \sim \mathsf{Be}(\alpha, \beta)$ are attained with $\alpha = \beta \approx 7.475$, so the posterior probability is at least $\mathsf{P}[\theta = 1/2 \mid X = 7] \geq (1 + 1.107)^{-1} = 0.482$ for any prior in that class. With $X = 9$ successes the same analysis would show $\mathsf{P}[\theta = 1/2 \mid X = 9] \geq (1 + 9.711)^{-1} = 0.0933$, attained at $\alpha = \beta = 0.67$. In contrast, the P-value for a two-sided test of $H_0 : \theta = 1/2$ would be $22/1024 = 0.0215$, suggesting much stronger evidence than the Bayesian analysis supports.

With larger sample-sizes the distinction is more extreme. For example, with $x = 60$ successes in $n = 100$ tries, the posterior probability that $\theta = 1/2$ with the conventional prior is $\mathsf{P}[\theta = 0.5 \mid X = 60] = 0.523$, and for any $\alpha = \beta > 0$ is at least $\mathsf{P}[\theta = 0.5 \mid X = 60] \geq 0.306$, while the two-sided P-value is 0.0569 suggesting strong evidence against H_0. Note the posterior probability $\mathsf{P}[\theta \leq 1/2 \mid X = 60] =$ `pbeta(0.5,61,41)` $= 0.02302$ for a uniform prior, not far from the P-value of 0.02845 for the one-sided hypothesis $H_0 : \theta \leq 1/2$, so there are aspects of this problem on which Bayesian and Frequentist statisticians might agree.

Normal example: Suppose $\{X_j\} \overset{\text{iid}}{\sim} \mathsf{No}(\theta, 1)$, and consider the simple null hypothesis $H_0 : \ \theta = \theta_0$ and two-sided composite alternative $H_0 : \ \theta \neq \theta_0$. For a Bayesian test of H_0 we need to specify not only the null prior probability π_0, but also the prior pdf $\pi_1(\theta)$ for θ if H_1 is true. For any such pdf the posterior odds against H_0 will be the prior odds times

$$\begin{aligned}
\Lambda(\mathbf{x}) &= \frac{(2\pi)^{-n/2} \int \exp(-\sum(X_j - \theta)^2/2)\, \pi_1(\theta)\, d\theta}{(2\pi)^{-n/2} \exp(-\sum(X_j - \theta_0)^2/2)} \\
&= \frac{\int \exp(-n(\bar{x} - \theta)^2/2)\, \pi_1(\theta)\, d\theta}{\exp(-n(\bar{x} - \theta_0)^2/2)} \\
&= e^{n(\bar{x}-\theta_0)^2/2} \int e^{-n(\bar{x}-\theta)^2/2}\, \pi_1(\theta)\, d\theta
\end{aligned}$$

For example, if $\pi_1 = \mathsf{No}(\mu, \tau^{-1})$ with mean μ and precision $\tau = 1/\sigma^2$, then

$$\Lambda(\mathbf{x}) = \sqrt{\frac{\tau}{n + \tau}} \ \exp\Big(\frac{n(\bar{x} - \theta_0)^2}{2} - \frac{n\tau(\bar{x} - \mu)^2}{2(n + \tau)}\Big).$$

In the limit as $\tau \to \infty$ this converges to the likelihood ratio for H_0 against the simple alternative $H_1 : \ \theta = \mu$ (see Section (5.1)). In the limit as $\tau \to 0$, Λ converges to zero, so the posterior probability that $\theta = \theta_0$ converges to one for *any* value θ_0, as we consider more and more diffuse alternatives. The peak value of Λ over all possible choices of the alternate distribution $\pi_1 = \mathsf{No}(\mu, \tau^{-1})$ is $\Lambda^* := \exp(n(\bar{x} - \theta_0)^2/2)$, attained at $\mu = \bar{x}$ as $\tau \to \infty$, a point mass at the observed value of $\hat{\theta} = \bar{x}$.

For normal alternatives π_1 centered at the same point θ_0 as the null,

$$\Lambda(x) = \sqrt{\frac{\tau}{n + \tau}} \ \exp\Big(\frac{n(\bar{x} - \theta_0)^2}{2} - \frac{n\tau(\bar{x} - \theta_0)^2}{2(n + \tau)}\Big) = \sqrt{\frac{\tau}{n + \tau}} \ \exp\Big(\frac{n^2(\bar{x} - \theta_0)^2}{2(n + \tau)}\Big)$$

The supremum Λ^* of this over all possible precisions τ is available analytically:

$$\Lambda^* := 1 \wedge \frac{1}{|\bar{x} - \theta_0|\sqrt{n}} \exp\left\{\frac{n(\bar{x} - \theta_0)^2 - 1}{2}\right\}.$$

These lead directly to bounds on how *small* the posterior probability of H_0 can be, since:

$$\mathsf{P}[H_0 \mid \mathbf{x}] \geq \frac{1}{1 + \frac{1-\pi_0}{\pi_0}\Lambda^*}$$

5.3 Conditional Inference

A classical statistician considering an hypothesis H_0 and alternative H_1 might either encounter evidence X representing extreme evidence against H_0 or might encounter evidence X representing extreme evidence against H_1. Berger et al. (1994) find some common ground for Fisherian adherents and Bayesian analysts by showing that a P-value computed *conditionally* on the level of extremity of the evidence (for *or* against H_0) will be the same as the posterior probability of H_0 under a reference analysis such as that presented in Section (5.2). The flaw in the usual Fisherian analysis is that it averages over many possible but unobserved outcomes, most of which are far more extreme than the one encountered.

5.4 Ancillarity

A statistic $A : \mathcal{X} \to \mathbb{R}$ is called *ancillary* if the probability distribution of $A(X)$ doesn't depend on X (Fisher, 1956, §VI.7). For example, if $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\theta, \sigma^2)$ with σ^2 known, the statistic $A(X) = [X_1 - X_2] \sim \mathsf{No}(0, 2\sigma^2)$ has a distribution that does not depend on θ, and so is ancillary. If σ^2 were uncertain then A would not be ancillary. A more important example for $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\theta, \sigma^2)$ data would be $S_n(X) := \sum(X_i - \bar{X}_n)^2$, n times the MLE $\hat{\sigma}^2$ for σ^2; one can (and we will) show that $S_n \sim \mathsf{Ga}\big((n{-}1)/2, 1/2\sigma^2\big)$ has a Gamma distribution that does not depend on θ, so S_n is ancillary for θ; and $S_n/\sigma^2 \sim \mathsf{Ga}\big((n{-}1)/2, 1/2\big) = \chi^2_{(n-1)/2}$ is ancillary for (θ, σ^2).

We are frequently advised to improve tests and estimation by *conditioning* on ancillary statistics[1]. In my experience they are difficult to discover or to condition on. Basu's Theorem (Basu, 1955) asserts that any complete sufficient statistic is independent of every ancillary statistic, so any analysis based on such a statistic would be unchanged by conditioning on any ancillary, but here's one cute example where ancillarity matters (taken from (Cox, 1958) and (Berger and Wolpert, 1988, §2)):

Let $U \sim \mathsf{Bi}(1, 1/2)$ be a Bernoulli variable and let $X \sim \mathsf{No}(\mu, \sigma_U^2)$ be normally-distributed with uncertain mean μ and variance that is either σ_0^2, if $U = 0$, or σ_1^2, if $U = 1$, for some specified numbers $0 < \sigma_0^2 \ll \sigma_1^2$ (this is called a *mixed experiment*). Cox colorfully described X as a measurement taken with one of two equally-likely measurement instruments, one more precise than the other. The statistic U is ancillary because its distribution does not depend on μ.

The object is to test the hypothesis $H_0 : \mu = 0$ against a specific alternative $H_1 : \mu = \mu'$, for some $\mu' \geq \sigma_1$. The (optimal, by Neyman/Pearson) log LR statistic is

$$\log \Lambda(x, u) = \log \frac{(2\pi\sigma_u^2)^{-1/2} \exp\big(-(x - \mu')^2/2\sigma_u^2\big)}{(2\pi\sigma_u^2)^{-1/2} \exp\big(-(x - 0)^2/2\sigma_u^2\big)} = [\mu' x - (\mu')^2/2]/\sigma_u^2,$$

[1] Note that $S := \mathsf{E}[T \mid A]$ does not depend on θ if A is ancillary, hence it is still a *statistic*; for non-ancillary A, typically this conditional expectation would depend on θ and so wouldn't be a statistic.

so the rejection region for the LRT will be of the form

$$\mathcal{R} = \big\{(x,u) :\ x \ge \mu'/2 + c\sigma_u^2\big\}$$

for some $c > 0$, a test of size

$$\begin{aligned}\alpha &= \mathsf{P}[(X,U) \in \mathcal{R} \mid \mu_0]\\ &= \tfrac{1}{2}\mathsf{P}_0[X > \mu'/2 + c\sigma_0^2 \mid U = 0] + \tfrac{1}{2}\mathsf{P}_1[X > \mu'/2 + c\sigma_1^2 \mid U = 1]\\ &= \tfrac{1}{2}\Phi\big(-\mu'/2\sigma_0 - c\sigma_0\big) + \tfrac{1}{2}\Phi\big(-\mu'/2\sigma_1 - c\sigma_1\big)\end{aligned}$$

reflecting the possibility of more extreme values of $T(X,U)$ from either measuring instrument. But the first term is nearly zero if $\sigma_0 \ll \sigma_1 \le \mu'$, while the second is about $\frac{1}{2}\Phi(-c\sigma_1)$— so we can very nearly take $c = z_\alpha/\sigma_1$ for the optimal Neyman/Pearson test.

But this is ridiculous— whenever $U = 0$ we can tell essentially perfectly whether H_0 is true or not, and when $U = 1$ we are now allowing ourselves to fail with probability about 2α, or 10% when $\alpha = 0.05$. If we *know* which instrument was used, *i.e.*, if we *observe* U, then a far more honest representation of the data would be to condition on the value of the ancillary statistic U, and reject whenever (X,U) lies in the set

$$\mathcal{R} = \{(x,u) :\ x \ge \sigma_u z_\alpha\}\,,$$

whose conditional (given U) size is

$$\mathsf{P}_0[(X,U) \in \mathcal{R} \mid U = u] = \Phi(-z_\alpha) = \alpha.$$

Incidentally, this is also the GLRT for $H_0 :\ \mu = 0$ *vs.* $H_1 :\ \mu > 0$.

6 χ^2 Tests for Multinomial Data

Let's consider repeating, over and over again, an experiment with k possible outcomes. If we let n be the number of times we repeat the experiment (independently!), and count the number N_i of times the i'th outcome occurs altogether, and denote by $\vec{p} = (p_1, ..., p_k)$ the vector of probabilities of the k outcomes, then then each N_i has a binomial distribution

$$N_i \sim \mathsf{Bi}(n, p_i)$$

but they're not independent. The joint probability of the events $[N_i = n_i]$ for nonnegative integers n_i is the "multinomial" distribution, with pmf:

$$f(\vec{n} \mid \vec{p}) = \binom{n}{n_1, n_2, ..., n_k} p_1^{n_1} \cdots p_k^{n_k} \tag{2}$$

where the "multinomial coefficient" is given by

$$\binom{n}{n_1, n_2, ..., n_k} = \binom{n}{\vec{n}} = \frac{n!}{n_1!\, n_2! \cdots n_k!}$$

if each $n_i \ge 0$ and $\sum n_i = n$, otherwise zero. This generalizes the *bi*nomial coefficient, $\binom{n}{n_1} = \binom{n}{n_1, n_2}$ with $n = n_1 + n_2$.

If we observe $\vec{N} = \vec{n}$, what is the MLE for $\vec{p}$? The answer is intuitively obvious, but *proving* it leads to something new. If we try to maximize Eqn (2) using derivatives (take logs first!), we find

$$\frac{\partial}{\partial p_i} \log f(\vec{n} \mid \vec{p}) = \frac{n_i}{p_i} > 0,$$

so obviously setting these derivatives to zero won't work— since they're always positive, $f(\vec{n} \mid \vec{p})$ is increasing in each p_i. The reason is that this is really a *constrained* optimization problem— the $\{p_i\}$'s have to be non-negative and *sum to one.* As a function on $\mathbb{R}^k$, the function $f(\vec{n} \mid \vec{p})$ of Eqn (2) increases without bound as we take all $p_i \to \infty$, but we're not allowed to do that since the sum of p_i mustn't exceed one.

An elegant solution is the method of *Lagrange Multipliers.* Introduce an additional variable $\lambda \in \mathbb{R}$, called the "Lagrange multiplier," and replace the log likelihood with the "Lagrangian":

$$\begin{aligned}\mathcal{L}(\vec{p}, \lambda) &:= \log f(\vec{n} \mid \vec{p}) + \lambda \left(1 - \sum p_i\right) \\ &= c + \sum n_i \log p_i + \lambda \left(1 - \sum p_i\right)\end{aligned}$$

with partial derivatives

$$\frac{\partial}{\partial p_i} \mathcal{L}(\vec{p}, \lambda) = \frac{n_i}{p_i} - \lambda \tag{3a}$$

$$\frac{\partial}{\partial \lambda} \mathcal{L}(\vec{p}, \lambda) = 1 - \sum p_i. \tag{3b}$$

Note that stationarity w.r.t λ (setting Eqn (3b) to zero) enforces the constraint. Now the vanishing of derivatives w.r.t. p_i in Eqn (3a) imply that $n_i/p_i = \lambda$ is constant for all i, so $p_i = n_i/\lambda$, while Eqn (3b) now gives $1 = \sum n_i/\lambda = n/\lambda$, so the solutions are the ones we guessed before:

$$\hat{p}_i = n_i/n \qquad\qquad \hat{\lambda} = n.$$

6.1 Generalized Likelihood Tests

Now let's consider testing a hypothetical value $\vec{p}^{\,0}$ for the probabilities, against the omnibus alternative:

$$\begin{aligned} H_0 :&\quad \vec{p} = \vec{p}^{\,0} = (p_1^0, \ldots, p_k^0) \\ H_1 :&\quad \vec{p} \neq \vec{p}^{\,0} \end{aligned}$$

(the alternative asserts that $p_i \neq p_i^0$ for at least one $1 \leq i \leq k$). The generalized likelihood ratio against H_0 is:

$$\begin{aligned}\Lambda(\vec{n}) &= \frac{\sup_{\vec{p}} f(\vec{n} \mid \vec{p})}{f(\vec{n} \mid \vec{p}^{\,0})} = \frac{f(\vec{n} \mid \hat{\vec{p}})}{f(\vec{n} \mid \vec{p}^{\,0})} = \frac{\binom{n}{\vec{n}} \prod (n_i/n)^{n_i}}{\binom{n}{\vec{n}} \prod (p_i^0)^{n_i}} \\ &= \prod (n_i/np_i^0)^{n_i}\end{aligned}$$

Introduce the notation $e_i := np_i^0$ for the "expected" number of outcomes of type i under null hypothesis H_0 and manipulate:

$$\begin{aligned}\Lambda(\vec{n}) &= \prod\left[\frac{n_i}{e_i}\right]^{n_i}\\ &= \prod\left[\frac{n_i - e_i + e_i}{e_i}\right]^{n_i} = \prod\left[1 + \frac{n_i - e_i}{e_i}\right]^{n_i}\end{aligned}$$

If the n_i's and e_i's are all large enough, we can approximate the logarithm of this by:

$$\begin{aligned}\log\Lambda(\vec{n}) &= \sum n_i \log\left(1 + \frac{n_i - e_i}{e_i}\right)\\ &\approx \sum (n_i - e_i + e_i)\left(\frac{n_i - e_i}{e_i} - \frac{(n_i - e_i)^2}{2\ e_i^2}\right)\end{aligned}$$

using the two-term Taylor series $\log(1+\epsilon) = \epsilon - \epsilon^2/2 + O(\epsilon^3)$

$$\approx \frac{1}{2}\sum\frac{(n_i - e_i)^2}{e_i} = \frac{1}{2}Q, \tag{4}$$

half the quadratic form $Q := \sum\frac{(n_i - e_i)^2}{e_i}$, since $\sum(n_i - e_i) = 0$ and since $\sum(n_i - e_i)^3 = O(1/\sqrt{n})$. The statistic Q is the so-called "Chi Squared" statistic introduced by Karl Pearson (1900), who found its asymptotic distribution.

Since each $n_i \sim \mathsf{Bi}(n_i, p_i)$, asymptotically each $n_i \sim \mathsf{No}\big(e_i, e_i(1-p_i^0)\big)$ and so the individual terms in the sum Eqn (4) have approximate $\mathsf{Ga}(\frac{1}{2}, \beta)$ distributions (proportional to a χ_1^2) with $\beta = 1/2(1-p_i)$, if H_0 is true; Pearson showed that Q has approximately (and asymptotically as $n \to \infty$) a χ_ν^2 distribution with $\nu = k - 1$ degrees of freedom (we'll see why below). If H_0 is false then Q will be much bigger, of course, leading to the well-known χ^2 test for H_0, with P-value

$$P = 1 - \texttt{pgamma(Q, nu/2, 1/2) = pgamma(Q, nu/2, 1/2, lower.tail=F)}$$

or rejection region

$$\mathcal{R}_\alpha = \{\vec{n} :\ Q \geq \texttt{qchisq(1-alpha, k-1, lower.tail=F)}\}$$

for a size-α test.

6.2 The Distribution of $Q(\vec{n})$

One way to compute the covariance of N_i and N_j is to use indicator variables, as follows. For $1 \leq \ell \leq n$ let J_ℓ be a label telling us which of the k possible outcomes happened on the ℓ'th trial—a random integer in the range $1, ..., k$, with probability $p_j = \mathsf{P}[J_\ell = j]$ for $1 \leq j \leq k$ and $1 \leq \ell \leq n$. Then N_i can be represented as the sum:

$$N_i = \sum_{\ell=1}^{n} \mathbf{1}_{\{J_\ell = i\}}$$

of indicator variables. This makes the following expectations easy for $i \neq j$:

$$\begin{aligned}
\mathsf{E}[N_i] &= \sum \mathsf{P}[J_\ell = i] && = np_i \\
\mathsf{E}[N_i^2] &= \mathsf{E}\left[\sum_\ell \sum_{\ell'} \mathbf{1}_{\{J_\ell = i\}} \mathbf{1}_{\{J_{\ell'} = i\}}\right] && = np_i + n(n-1)p_i^2 \\
& && = np_i(1-p_i) + (np_i)^2 \\
\mathsf{E}[N_i N_j] &= \mathsf{E}\left[\sum_\ell \sum_{\ell'} \mathbf{1}_{\{J_\ell = i\}} \mathbf{1}_{\{J_{\ell'} = j\}}\right] && = n(n-1)p_i p_j \\
\mathsf{V}(N_i) &= np_i(1-p_i) \\
\mathsf{Cov}(N_i, N_j) &= -np_i\, p_j
\end{aligned}$$

If we let $Z \sim \mathsf{No}(0,1)$ be independent of $\vec{N}$ and add $Zp_i\sqrt{n}$ to each component N_i, we will exactly cancel the negative covariance:

$$\mathsf{Cov}\big((N_i + Zp_i\sqrt{n}), (N_j + Zp_j\sqrt{n})\big) = -np_ip_j + (p_i\sqrt{n})(p_j\sqrt{n}) \qquad = 0$$

while keeping zero mean

$$\mathsf{E}\big((N_i + Zp_i\sqrt{n})\big) = 0$$

and increase the variance to

$$\mathsf{V}\big((N_i + Zp_i\sqrt{n})\big) = np_i(1-p_i) + (p_i\sqrt{n})^2 \qquad = e_i.$$

Thus the random variables $(N_i - e_i + Zp_i\sqrt{n})/\sqrt{e_i}$ are uncorrelated and have mean zero and variance one. By the Central Limit Theorem, they are approximately k independent standard normal random variables as $n \to \infty$, so the quadratic form

$$Q^+(\vec{n}) := \sum_{i=1}^{k} \frac{(N_i - e_i + Zp_i\sqrt{n})^2}{e_i}$$

has approximately a χ^2_k distribution for large n. But:

$$\begin{aligned}
Q^+(\vec{n}) &= \sum \frac{(N_i - e_i)^2}{np_i} && + \sum \frac{2(N_i - e_i)Z\,p_i\sqrt{n}}{np_i} && + \sum \frac{Z^2 p_i^2 n}{np_i} \\
&= Q(\vec{n}) && + \frac{2Z}{\sqrt{n}} \sum (N_i - e_i) && + Z^2 \sum p_i \\
&= Q(\vec{n}) + Z^2,
\end{aligned}$$

the sum of $Q(\vec{n})$ and a χ^2_1 random variable independent of $\vec{N}$— so $Q(\vec{n})$ itself must have approximately a χ^2_ν distribution with $\nu = (k-1)$ degrees of freedom.

Chi-squared tests

Thus the χ^2 *test* of $H_0 : \vec{p} = \vec{p}^{\,0}$ against the omnibus alternative $H_1 : \vec{p} \neq \vec{p}^{\,0}$ proceeds as follows:

1. Evaluate the counts n_i of the ith outcome, for each $1 \le i \le k$, and the sum $n_+ := \sum n_i$;
2. Evaluate the *expected* counts $e_i = n_+ p_i$ of the ith outcome, under the null hypothesis;
3. Evaluate the χ^2 statistic
$$Q := \sum \frac{(n_i - e_i)^2}{e_i};$$
4. Reject H_0 at level α if $Q \ge \texttt{qchisq}(1-\texttt{alpha}, \texttt{k}-1, \texttt{lower}=\texttt{F})$ or, equivalently, if the P-value $P = \texttt{pchisq}(\texttt{Q}, \texttt{k}-1, \texttt{lower}=\texttt{F})$ is below α.

Because the alternative H_1 for this test is so broad, including *all* probability vectors $\vec{p} \neq \vec{p}^{\,0}$, the test isn't particularly powerful. If you are concerned about a particular departure from $H_0 : \vec{p} = \vec{p}^{\,0}$, you may wish to construct a test tailored to (and more sensitive to) that departure.

6.3 P-Values for χ^2

The χ^2_ν distribution is just the $\mathsf{Ga}(\alpha = \nu/2,\ \beta = 1/2)$. If the degrees of freedom parameter ν is even, it may be viewed as the waiting time for $\nu/2$ events in a Poisson process X_t with rate $1/2$, so P-values can be computed in closed form as

$$\mathsf{P}[Q > q] = \mathsf{P}[X_q < \nu/2] = \sum_{k=0}^{(\nu/2)-1} \frac{(q/2)^k}{k!}\, e^{-q/2}.$$

For example, with $\nu = 2$ degrees of freedom, the P-value is simply $e^{-q/2}$, while for $\nu = 4$ and $\nu = 6$ it is $(1 + q/2)e^{-q/2}$ and $(1 + q/2 + q^2/8)e^{-q/2}$, respectively.

For large values of ν the χ^2_ν distribution is close to the normal $\mathsf{No}(\nu, 2\nu)$ by the Central Limit Theorem, so

$$\mathsf{P}[Q > q] \approx \Phi\left(\frac{\nu - q}{\sqrt{2\nu}}\right).$$

For any ν and q, it's available in `R` as `1-pchisq(q, nu)` or, more precisely for large q, as

```
pchisq(q, nu, lower.tail=FALSE).
```

6.4 Contingency Tables

Now consider a composite hypothesis like:

$$H_0: \ \{N_{ij}\} \sim \mathsf{MN}(n; \theta_{ij}) \text{ for some } \theta_{ij} = p_i\, q_j, \ 1 \le i \le R, \ 1 \le j \le C$$

for $R \cdot C$ counts N_{ij} summing to n. If n items are categorized separately into one of R rows and also into one of C columns, and if N_{ij} denotes the number of items in the ith row and jth column, then this hypothesis asserts that the two categorizations are *independent.* Alternately, if $N_{i+} \equiv \sum_{j=1}^{C} N_{ij}$ objects from the ith of R populations are categorized into one of C categories, then H_0 also asserts that the R populations are all *homogeneous* in the sense that they share the same distribution among the C categories.

In either case, a Generalized Likelihood Ratio test will be based on

$$\Lambda = \frac{\sup_\theta \left\{\prod \theta_{ij}^{N_{ij}} : \ \sum \theta_{ij} = 1\right\}}{\sup_{p,q} \left\{\prod (p_i q_j)^{N_{ij}} : \ \sum p_i = 1, \ \sum q_j = 1\right\}} = \prod \left\{\frac{\hat\theta_{ij}}{\hat p_i \hat q_j}\right\}^{N_{ij}}$$

where $\hat\theta_{ij} = N_{ij}/n$, $\hat p_i = N_{i+}/n$, and $\hat q_j = N_{+j}/n$. Upon setting $\hat e_{ij} \equiv n \hat p_i \hat q_j$,

$$\begin{aligned}
\log \Lambda &= \sum N_{ij} \log \left\{\frac{N_{ij}}{\hat e_{ij}}\right\} \\
&= \sum \{(N_{ij} - \hat e_{ij}) + \hat e_{ij}\} \log \left\{1 + \frac{N_{ij} - \hat e_{ij}}{\hat e_{ij}}\right\} \\
&\approx \frac{1}{2} \sum \frac{(N_{ij} - \hat e_{ij})^2}{\hat e_{ij}} = Q/2, \text{ where} \\
Q &:= \sum \frac{(N_{ij} - \hat e_{ij})^2}{\hat e_{ij}}
\end{aligned}$$

has approximately a χ^2_ν distribution with $\nu = RC - 1 - (R-1) - (C-1) = (R-1)(C-1)$ degrees of freedom. More generally, Q will have approximately a χ^2_ν distribution with $\nu = k - 1 - s$ degrees of freedom if there are k categories and we must estimate an s-dimensional aspect of θ from the data. The same idea may be used to test independence for three-way (or more) classifications, in which H_0 asserts that $\theta_{ijk} = p_i q_j r_k$ for some $\vec p$, $\vec q$, $\vec r$.

6.4.1 A Numerical Example

A 1986 study of a treatment for Hodgkin disease (Dunsmore et al., 1987) studied the response rates (classified into three levels: Positive, Partial, and None) for patients of four different histological types. The results are summarized in this table:

Type	Pos	Part	Neg	
LP	74	18	12	104
NS	68	16	12	96
MC	154	54	58	266
LD	18	10	44	72
	314	98	126	538

Denote by X_{ij} the entry in the ith row and jth column, and by X_{i+} and X_{+j} the row and column sums (shown in the table). The *expected* count under H_0 in cell (i,j) is $E_{ij} := X_{i+}X_{+j}/n-$ $E_{11} = 104 \times 314/538 = 60.70$ for $(1,1)$, for example, so the ξ^2 statistic is $Q = \sum(X_{ij} - E_{ij})^2/E_{ij} = 75.89$. Under the null hypothesis this would have a χ^2_ν distribution with $\nu = (R-1)(C-1) = 6$ degrees of freedom. The P-value is $P =$ `pchisq(Q, 6, low=F)` $= 2.52 \cdot 10^{-14}$, so H_0 would be rejected.

In `R` this calculation could be performed as follows:

```
Xij <- matrix( c(74,68,154,18, 18,16,54,10, 12,12,58,44), ncol=3);
row <- apply(Xij,1,sum);       # Row sums
col <- apply(Xij,2,sum);       # Column sums
Eij <- row %o% col / sum(Xij); # Expected counts
Q   <- sum( (Xij-Eij)^2/Eij ); # Chi-square statistic
P   <- pchisq(Q, 6, low=F);    # P-value
```

using the "`apply()`" function and the outer product operator "`%o%`".

6.4.2 Two by Two

An important special case of contingency table analysis is when $R = C = 2$. For example, we may study the benefit (or risk) of Exposure to some treatment (or hazard) by exploring the independence of classifications with respect to Exposure (Exposed and non-Exposed) and also to a health outcome (here, Diseased or non-Diseased). Denote the count of subjects in each class as X_{ij}, where $i \in \{0,1\}$ indexes the exposure class (1=Exposed) and $j \in \{0,1\}$ the disease class (1=Diseased). The object will be to test the hypothesis H_0 that exposure is unrelated to disease status, against the two-sided alternative that there is some connection.

These data might arise from any of three possible sampling schemes, which each lead to different probability models, and somewhat different expressions for H_0:

1. **Multinomial:** For some number $n \in \mathbb{N}$ and probability vector $p = (p_{00}, p_{01}, p_{10}, p_{11})$, $\mathbf{x} = (X_{00}, X_{01}, X_{10}, X_{11}) \sim \mathsf{MN}(n,p)$. H_0 would assert that row and column classifications are independent, *i.e.*, that $p_{00}p_{11} = p_{01}p_{01}$ or, equivalently, that the ratio ψ is one, where
$$\psi := \frac{p_{00}p_{11}}{p_{01}p_{10}}.$$

2. **Prospective:** For some numbers $x_{1+} \in \mathbb{N}$ of Exposed and $x_{0+} \in \mathbb{N}$ of un-Exposed subjects, we observe $X_{11} \sim \mathsf{Bi}(x_{1+}, \mathsf{P}(D \mid E))$ and $X_{01} \sim \mathsf{Bi}(x_{0+}, \mathsf{P}(D \mid E^c))$ diseased cases, respectively. H_0 would assert that $\mathsf{P}(D \mid E^c) = \mathsf{P}(D \mid E)$ or, equivalently, that the disease *odds* are equal for exposed and unexposed subjects
$$\frac{\mathsf{P}(D \mid E^c)}{\mathsf{P}(D^c \mid E^c)} = \frac{\mathsf{P}(D \mid E)}{\mathsf{P}(D^c \mid E)}.$$
This condition is satisfied if and only if the *odds ratio* is one:
$$\psi := \frac{\mathsf{P}(D \mid E)\mathsf{P}(D^c \mid E^c)}{\mathsf{P}(D \mid E^c)\mathsf{P}(D^c \mid E)} = \frac{\mathsf{P}(D \cap E)\mathsf{P}(D^c \cap E^c)}{\mathsf{P}(D \cap E^c)\mathsf{P}(D^c \cap E)} = \frac{p_{00}p_{11}}{p_{01}p_{10}}.$$

3. **Retrospective:** Among some numbers $x_{+1} \in \mathbb{N}$ of Diseased and $x_{+0} \in \mathbb{N}$ of un-Diseased subjects, we discover that $X_{11} \sim \mathsf{Bi}(x_{+1}, \mathsf{P}(E \mid D))$ and $X_{10} \sim \mathsf{Bi}(x_{+0}, \mathsf{P}(E \mid D^c))$ had been exposed, respectively. H_0 would assert that $\mathsf{P}(E \mid D^c) = \mathsf{P}(E \mid D)$ or, equivalently, that the exposure *odds* are equal for diseased and undiseased subjects

$$\frac{\mathsf{P}(E \mid D^c)}{\mathsf{P}(E^c \mid D^c)} = \frac{\mathsf{P}(E \mid D)}{\mathsf{P}(D^c \mid D)}.$$

Again this is satisfied if and only if the *odds ratio* is one:

$$\psi := \frac{\mathsf{P}(E \mid D)\mathsf{P}(E^c \mid D^c)}{\mathsf{P}(E \mid D^c)\mathsf{P}(E^c \mid D)} = \frac{\mathsf{P}(E \cap D)\mathsf{P}(E^c \cap D^c)}{\mathsf{P}(E \cap D^c)\mathsf{P}(E^c \cap D)} = \frac{p_{00}p_{11}}{p_{01}p_{10}}.$$

Thus, all three sampling approaches lead to consideration of whether or not the odds ratio ψ is unity. A value of $\psi > 1$ indicates a positive association between exposure and disease; a value $\psi < 1$ indicates a protective effect. The Maximum Likelihood Estimator for ψ in all three cases is

$$\hat{\psi} = \frac{X_{00}X_{11}}{X_{01}X_{10}},$$

and the GLRT of H_0 in all cases leads to rejection of H_0 for large values of the GLR statistic

$$\Lambda = \prod_{i,j=0,0}^{1,1} \left(\frac{n\, X_{ij}}{X_{i+}X_{+j}} \right)^{X_{ij}} = \prod_{i,j=0,0}^{1,1} (X_{ij}/E_{ij})^{X_{ij}},$$

where $E_{ij} := X_{i+}X_{+j}/n$ is the "expected" count under the hypothesis H_0 of independence. Equivalently, one would reject for large values of its logarithm

$$\log \Lambda = \sum X_{ij} \log(X_{ij}/E_{ij}) \approx Q/2, \qquad \text{where}$$
$$Q := \sum \frac{(X_{ij} - E_{ij})^2}{E_{ij}}$$

has approximately a χ^2_1 distribution for large n.

6.4.3 A Numerical Example

But what if n is *not* large? The famous 1985 RCT test of extracorporeal membrane oxygenation (ECMO– see Ware, 1989) featured only 19 subjects. Six of ten in the control group survived, and all nine of the treated subjects survived, so the data are

$$X_{00} = 6 \qquad X_{01} = 4 \qquad X_{10} = 9 \qquad X_{11} = 0$$

and the MLE for the odds ratio is $\hat{\psi} = \infty$. Evidently this sample size is insufficient for the χ^2 approximation to hold.

Wolpert and Mengersen (2004) introduced an objective Bayesian approach using independent Jeffreys' prior distributions for the survival probabilities p and q in the Exposed (to ECMO) and un-Exposed groups, respectively, and then find the posterior probability distribution for $\psi = p(1-q)/(1-p)q$.

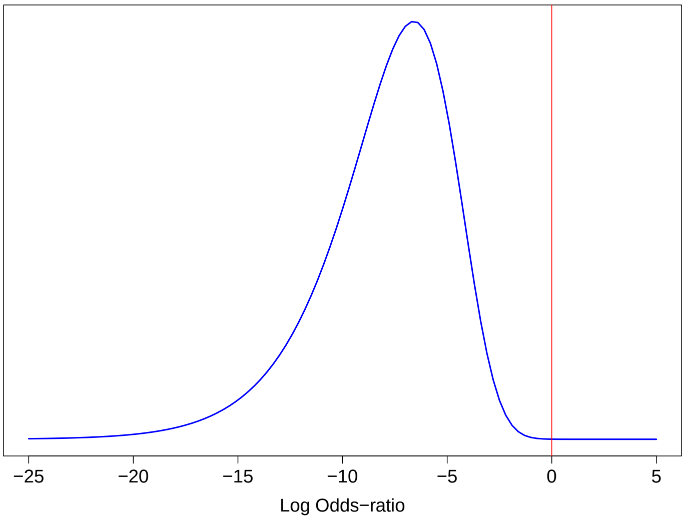

Figure 3: Reference Posterior PDF for ECMO Log Odds Ratio

They found an explicit form for the posterior pdf of $\varepsilon := \log\psi$,

$$f(\mathbf{x}\mid\epsilon) \propto e^{\varepsilon(X_{11}+1/2)}{}_2F_1\big(X_{+0}+1, X_{+1}+1; X_{++}+2; 1-e^{\varepsilon}\big) \tag{5}$$

in terms of the hypergeometric function ${}_2F_1(a,b;c;z)$ (Abramowitz and Stegun, 1964, §15.1) and evaluated its mean and variance as

$$\mu = \psi(X_{00}+\tfrac{1}{2}) \;-\psi(X_{01}+\tfrac{1}{2}) \;-\psi(X_{10}+\tfrac{1}{2}) \;+\psi(X_{11}+\tfrac{1}{2}) \tag{6a}$$

$$\sigma^2 = \psi'(X_{00}+\tfrac{1}{2}) + \psi'(X_{01}+\tfrac{1}{2}) + \psi'(X_{10}+\tfrac{1}{2}) + \psi'(X_{11}+\tfrac{1}{2}) \tag{6b}$$

where $\psi(z) = (d/dz)\log(\Gamma(z))$ and $\psi(z) = (d/dz)\psi(z)$ are the digamma and trigamma functions, respectively (Abramowitz and Stegun, 1964, §6.3, 6.4). These are included in `R` and other computing environments, but their values here can be computed easily using the identities

$$(n+\tfrac{1}{2}) - \psi(m+\tfrac{1}{2}) = \sum_{i=m}^{n-1}(i+\tfrac{1}{2})^{-1} \approx \log\frac{n}{m} \tag{7a}$$

for integers $0 \le m < n$, and

$$\psi'(n+\tfrac{1}{2}) = \frac{\pi^2}{2} - \sum_{i=0}^{n-1}(i+\tfrac{1}{2})^{-2} \approx \frac{1}{n} \tag{7b}$$

For the ECMO trial, these give $\mu = -3.75721$ and $\sigma = 2.3368$; under the normal approximation to the posterior of ε the approximate posterior probability of no effect or harmful effect would be $\mathsf{P}[\varepsilon > 0 \mid \mathbf{x}] \approx \Phi(\mu/\sigma) = 0.0539$. In fact, due to the skewness of the pdf (see Figure (3)), it is considerably smaller— numerical integration of (5) gives $\mathsf{P}[\varepsilon > 0 \mid \mathbf{x}] \approx 9.514 \cdot 10^{-6}$, rather strong evidence in ECMO's favor despite the small sample sizes.

6.4.4 Frequentist Analysis of ECMO

Let $X \sim \mathsf{Bi}(n,p)$ and set $q := (1-p)$, the failure probability, and $\theta := \log p/q$, the log odds. The MLE for θ is

$$\begin{aligned}\hat\theta &= \log\frac{x/n}{1-x/n}\\ &= \theta + \log(x/np) - \log((n-x)/nq)\\ &= \theta + \log\big(1+\frac{x-np}{np}\big) - \log\big(1+\frac{n-x-nq}{nq}\big)\\ &= \theta + \log\big(1+\frac{x-np}{np}\big) - \log\big(1-\frac{x-np}{nq}\big)\end{aligned}$$

If n is sufficiently large that $|x-np| \ll n$, then by the delta method

$$\hat\theta \approx \theta + \frac{x-np}{np} + \frac{x-np}{nq} = \theta + \frac{x-np}{npq} \approx \mathsf{No}(\theta,\sigma^2)$$

by the CLT, with mean θ and variance

$$\sigma^2 = \mathsf{E}\left(\frac{x-np}{npq}\right)^2 = \frac{npq}{n^2p^2q^2} = \frac{1}{npq}.$$

In a prospective trial with independent treatment and control arms, it follows that for sufficiently large sample sizes the MLE $\hat\varepsilon$ for the log odds ratio

$$\varepsilon = \log\psi = \log\frac{\mathsf{P}(D\mid E)\mathsf{P}(D^c\mid E^c)}{\mathsf{P}(D\mid E^c)\mathsf{P}(D^c\mid E)} = \log\frac{\mathsf{P}(D\mid E)}{\mathsf{P}(D^c\mid E)} - \log\frac{\mathsf{P}(D\mid E^c)}{\mathsf{P}(D^c\mid E^c)}$$

is also approximately normally distributed with mean ε and variance

$$\begin{aligned}\sigma^2 &= \frac{1}{X_{1+}\mathsf{P}(D\mid E)\mathsf{P}(D^c\mid E)} + \frac{1}{X_{0+}\mathsf{P}(D\mid E^c)\mathsf{P}(D^c\mid E^c)}\\ &\approx \frac{1}{X_{00}} + \frac{1}{X_{01}} + \frac{1}{X_{10}} + \frac{1}{X_{11}}.\end{aligned}$$

By Eqns (6a, 7a) $\mathsf{E}[\hat\theta]$ is close to the reference posterior mean of ε, and by Eqns (6b, 7b) $\mathsf{Var}[\hat\theta]$ is close to the posterior variance of ε, for sufficiently large sample sizes. Unfortunately ECMO's sample sizes were far too small for the delta method or the CLT to apply.

6.5 Other Composite Hypotheses

We can also use a χ^2 test to see if data $\{X_i\}$ come from *some* unspecified member of a parametric family $f(x\mid\theta)$ of distributions. Typically we must aggregate or *bin* the data into a finite number (say, k) of categories; compute the category probabilities $p_i(\theta)$, $1\le i\le k$; minimize Λ over all possible values of θ (or, nearly the same thing, minimize $Q(\theta)$); and approximate the distribution of $Q(\hat\theta)$ by the χ^2_ν with $\nu = k-1-s$, for $\theta\in\Theta\subseteq\mathbb{R}^s$.

6.5.1 Poisson example

For instance, in DeGroot and Shervish (2012, problem 5, §10.2), we have $n = 200$ observations $X_i \in \mathbb{Z}_+$ which may be from a $\mathsf{Po}(\theta)$ distribution:

$$\begin{aligned} X=0: &\quad 52\\ X=1: &\quad 60\\ X=2: &\quad 55\\ X=3: &\quad 18\\ X=4: &\quad 8\\ X\ge 5: &\quad 7 \end{aligned}$$

At any specific θ, the likelihood for the grouped data would be

$$\begin{aligned} L(\theta) &= \prod_{i=0}^{4} \left[\frac{\theta^i}{i!}e^{-\theta}\right]^{N_i} \cdot \left[1 - e^{-\theta}\sum_{i=0}^{4}\frac{\theta^i}{i!}\right]^{N_5} \\ &\propto \theta^{0\cdot 52+1\cdot 60+2\cdot 55+3\cdot 18+4\cdot 8} e^{-\theta[52+60+55+18+8]} \left[1 - e^{-\theta}\sum_{i=0}^{4}\frac{\theta^i}{i!}\right]^{7} \\ &= \theta^{256}e^{-193\theta}\left[1 - e^{-\theta}[1+\theta+\theta^2/2+\theta^3/6+\theta^4/24]\right]^7 \end{aligned}$$

The optimal θ is $\hat\theta = 1.465232$ (found by a numerical search) with $Q(\hat\theta) = 7.696875$, for a P-value of $P =$ `pchisq(7.696875, df=4, low=F)` $= (1+Q/2)e^{-Q/2} = 0.1033348$. Evidently we can't reject the Poisson hypothesis at levels $\alpha \le 0.10$. Figure (4) shows a plot of the log likelihood, with $\hat\theta$ noted. In this example $\hat\theta$ is very close to the *Poisson* MLE of $\tilde\theta = 1.5$ (using the additional information about the "$X \ge 5$" observations offered in Problem 5 of DeGroot and Shervish (2012, §10.2), so the values of the log likelihood and of Q agree to two decimal places and the same conclusions would be drawn using either method.

6.5.2 Geometric example

In contrast, let's test to see if the same data come from the geometric distribution $\mathsf{Ge}(p)$ for any $p \in [0,1]$. Setting $q = 1-p$, the geometric probabilities are

$$\mathsf{P}[X=i] = pq^i, \quad 0 \le i \le 4 \qquad\qquad \mathsf{P}[X \ge 5] = q^5$$

so

$$L(p) = \prod_{i=0}^{4}\left[p\, q^i\right]^{N_i} \cdot \left[q^5\right]^{N_5} = p^{\sum_{i=0}^4 N_i} q^{\sum_{i=0}^5 i\, N_i} = p^{193}q^{291}.$$

This attains its maximum at $\hat p = 193/(193+291) = 193/484$, leading to "expected" counts of $e_i = n\hat p\hat q^i$ for $0 \le i \le 4$, and $e_5 = n\hat q^5$. The log GLR statistic and the quadratic form Q are

$$\begin{aligned} \log\Lambda &= \sum N_i \log(N_i/e_i) = 19.6416 \\ Q &= \sum (N_i - e_i)^2/e^i \;= 41.8620 \approx 2\log\Lambda \end{aligned}$$

for a P-value of $P =$ `pchisq(41.86, 4, low=F)` of $P \approx 1.78\cdot 10^{-8}$. This offers clear evidence that these data do *not* come from any exponential distribution.

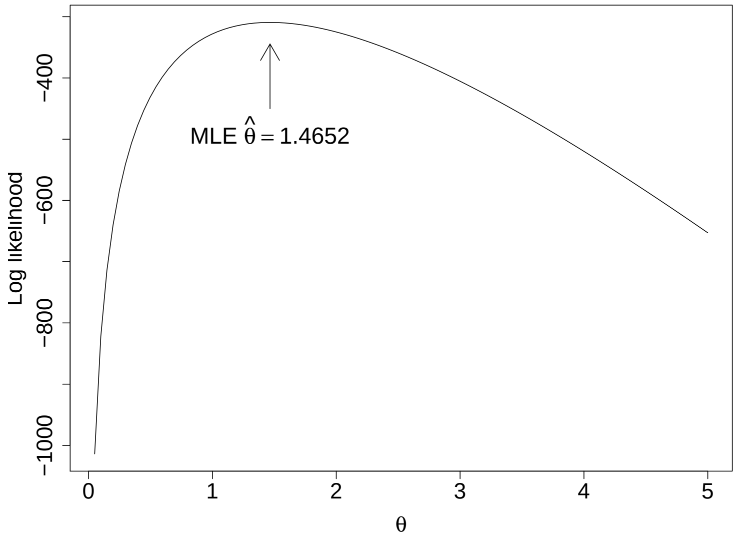

Figure 4: Multinomial log likelihood.

6.5.3 Generic example: Model Checking

We can construct a GLR test of the null hypothesis that observations $X_1, \ldots, X_n$ come from *any* parametric family $\mathcal{P} = \{f_\theta(x) : \ \theta \in \Theta\}$ with finite-dimensional parameter space $\Theta \subset \mathbb{R}^s$ as follows:

- Partition the outcome space $\mathcal{X} = \cup_{i=1}^k A_i$ into some number $k > s+1$ of disjoint sets A_i;
- Evaluate the probabilities $p_i(\theta) = \int_{A_i} f_\theta(x)\, dx$ that X will fall into each partition element;
- Count the observed occupancies $N_i = \sum_j \mathbf{1}_{A_i}(X_j) = \#\{j : \ X_j \in A_i\}$;
- Find $\tilde\theta = \operatorname{argmax}_\theta \sum N_i \log p_i(\theta)$ and set $\hat p_i := p_i(\tilde\theta)$ and $E_i := n\hat p_i$;
- Evaluate
$$Q(\tilde\theta) := \sum_{i=1}^k \frac{\big(N_i - np_i(\tilde\theta)\big)^2}{np_i(\tilde\theta)} = \sum_{i=1}^k \frac{(N_i - E_i)^2}{E_i};$$
- Report $P =$ `pchisq(Q, k-s-1, low=F)`, or reject if `Q > qchisq(1-alpha, k-s-1)`.

The fourth step can be replaced with "Find $\tilde\theta = \operatorname{argmin}_\theta Q(\theta)$", but *not* by "Set θ equal to its MLE $\hat\theta$ under the model $\mathcal{P}$." Since the multinomial likelihood function is very nearly proportional to $e^{Q/2}$ (that's how the χ^2 test was derived, after all), the multinomial MLE $\tilde\theta$ is very nearly the minimizing value of Q, but other estimates θ^* of θ will lead to a heavier-tailed distribution for $Q(\theta^*)$ than the χ^2_{k-s-1}. For the MLE $\hat\theta$ under the model $\mathcal{P}$, Chernoff and Lehmann (1954) showed that $Q(\hat\theta)$ is distributed like the sum of a χ^2_{k-s-1} random variable and an independent sum $\sum_{i=1}^s \lambda_i Z_i^2$ for $\{Z_i\} \overset{\text{iid}}{\sim} \mathsf{No}(0,1)$ and numbers $0 \le \lambda_i \le 1$. It follows that its CDF lies between those of the χ^2_{k-s-1} and χ^2_{k-1}, so it is valid to reject H_0 at level α if $Q(\hat\theta)$ exceeds the $(1-\alpha)$th quantile of the χ^2_{k-1} distribution.

7 Introduction to Sequential Tests

In this section I'll introduce a few ideas about sequential testing of hypotheses. This topic involves a little more probability theory than students are expected to know in this class, so don't worry you aren't familiar with some of the concepts— just take it as a sketchy introduction to dynamical statistical inference.

7.1 A Cautionary Example

Let $\{X_i\}_{1\le i\le n} \overset{\text{iid}}{\sim} \mathsf{No}(\mu,\sigma^2)$ be iid univariate normal random variables, with known variance σ^2. The two-sided GLRT of $H_0:\ \mu=\mu_0$ against $H_1:\ \mu\ne\mu_0$ of size $\alpha=0.05$ will reject H_0 if we observe $\{\sqrt{n}|\bar X_n-\mu_0|\ge 1.96\sigma\}$. A naïve investigator who observes $\mathbf{x}\notin\mathcal{R}_n$ may be tempted to continue taking observations (*i.e.*, increase n) in the hope of achieving "significant" evidence against H_0. It turns out that, with enough patience, that hope can always be realized, even if H_0 is true!

A result from advanced probability theory called the "law of the iterated logarithm" asserts that, for $\{Z_i\}\overset{\text{iid}}{\sim}\mathsf{No}(0,1)$,

$$\limsup_{n\to\infty}\frac{\pm\sum_{i=1}^n Z_i}{\sqrt{2n\log\log n}}=1.$$

But, if H_0 is true, that means that $\sqrt{n}(\bar X_n-\mu_0)$ will be arbitrarily close to $\pm\sqrt{2\sigma^2\log\log n}$ infinitely-often and, in particular, $\sqrt{n}|\bar X_n-\mu_0|$ will eventually exceed 1.96σ (or any other finite number) *even if H_0 is true.* Thus, a naïve (or unscrupulous) investigator with sufficient resources can always find significant enough evidence to reject any hypothesis, true or not.

7.2 Wald's SPRT

Wald (1945) proposed and studied a valid approach to the sequential testing of hypotheses that overcomes the problems sketched in Section 7.1. Suppose we have unlimited numbers of iid observations $\{X_i\}$ available, and that it is believed that these come from a distribution with pdf either $f_0(x)$ or $f_1(x)$. The LR against $H_0:\ \{X_i\}\overset{\text{iid}}{\sim} f_0(x)$ after the first n observations is

$$\Lambda_n := \prod_{i=1}^n \frac{f_1(X_i)}{f_0(X_i)}.$$

Wald's "sequential probability ratio test" begins with the selection of two numbers a,b that satisfy $0<a<1<b<\infty$ and proceeds as follows. Continue taking observations X_n until either $\Lambda_n\ge b$, in which case the experiment stops and H_0 is rejected; or until $\Lambda_n\le a$, in which case the experiment stops and H_0 is not rejected. What are the error probabilities of this test? And how large will the typical sample size be?

If H_0 is true, then

$$\mathsf{E}_0[\Lambda_{n+1}\mid X_1,\cdots,X_n]=\Lambda_n\int_{\mathcal{X}}\frac{f_1(x)}{f_0(x)}\,f_0(x)\,dx=\Lambda_n,$$

so Λ_n is a "martingale". Another result from advanced probability asserts that $\mathsf{E}[\Lambda_\tau]=\Lambda_0=1$ not only for all fixed integers $\tau\ge1$, but also for random "stopping times", *i.e.*, random integers τ

with the property that the event $[\tau = n]$ depends only on the first n observations $X_1, \cdots, X_n$. In particular, for $\tau := \inf\{n : \Lambda_n \ge b \text{ or } \Lambda_n \le a\}$, we have

$$\begin{aligned} 1 &= \mathsf{E}_0[\Lambda_\tau] \\ &\approx a\mathsf{P}_0[\Lambda_\tau \le a] + b\mathsf{P}_0[\Lambda_\tau \ge b] \\ &= a\mathsf{P}_0[\text{Do not reject}] + b\mathsf{P}_0[\text{Reject}] \\ &= a(1-\alpha) + b\alpha, \text{ so} \\ \alpha &\approx \frac{1-a}{b-a}. \end{aligned}$$

Similarly, if H_1 is true, then $\mathsf{E}_1[\Lambda_{n+1}^{-1} \mid X_1, \cdots, X_n] = \Lambda_n^{-1}$ and Λ_n^{-1} is a martingale, so

$$\begin{aligned} 1 &= \mathsf{E}_1[\Lambda_\tau^{-1}] \\ &\approx a^{-1}\mathsf{P}_1[\text{Do not reject}] + b^{-1}\mathsf{P}_1[\text{Reject}] \\ &= a^{-1}\beta + b^{-1}(1-\beta), \text{ so} \\ \beta &\approx \frac{ab-a}{b-a}. \end{aligned}$$

Any specified error probabilities α, β can be achieved by setting $a = \frac{\beta}{1-\alpha}$ and $b = \frac{1-\beta}{\alpha}$. For example, set $a = 1/19$ and $b = 19$ for $\alpha = \beta = 0.05$, or $a = 1/9$ and $b = 9$ for $\alpha = \beta = 0.10$.

To illustrate, let $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu, 1)$ and consider testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ at level $\alpha = 0.05$ for some $\mu_1 > \mu_0$, with Type-II error of $\beta \le 0.05$. The fixed sample-size LRT would require a sample-size of $n \ge 11$ and would reject if $\sqrt{n}(\bar{X}_n - \mu_0) \ge 1.645$. The SPRT will continue until $\Lambda_n \ge 19$ or $\Lambda_n < 1/19$, *i.e.*, until $|\bar{X}_n - (\mu_0 + \mu_1)/2|$ exceeds $(\log 19)/n|\mu_1 - \mu_0|$, and reject if $\bar{X}_n \ge (\mu_0 + \mu_1)/2$ and fail to do so if $\bar{X}_n < (\mu_0 + \mu_1)/2$.

7.3 How long does the SPRT last?

Whether or not H_0 is true, the *logarithm* of the LR is a random walk:

$$\log \Lambda_n = \sum_{i=1}^n \log \frac{f_1(X_i)}{f_0(X_i)}.$$

The iid steps, if in fact $\{X_i\} \overset{\text{iid}}{\sim} f(x)$, each have expectation

$$\begin{aligned} \mu &= \int \log \frac{f_1(x)}{f_0(x)} f(x)\, dx \\ &= \int \log \frac{f(x)}{f_0(x)} f(x)\, dx - \int \log \frac{f(x)}{f_1(x)} f(x)\, dx \\ &= K[f{:}f_0] - K[f{:}f_1], \end{aligned}$$

the difference in the Kullback-Leibler divergence from f to f_0 and that from f to f_1. In particular, if H_0 is true then $\mu = -K[f_0{:}f_1]$ and, on average, it will take about $\tau \approx -(\log a)/K[f_0{:}f_1]$ steps before the test ends (usually, with $\Lambda_\tau \le a$ so H_0 is *not* rejected); conversely, if H_1 is true then

$\mu = K[f_1{:}f_0]$ and, on average, it will take about $\tau \approx (\log b)/K[f_1{:}f_0]$ steps before the test ends (usually, with $\Lambda_\tau \ge b$ so H_0 is rejected). A little more precisely, using the martingale property for $[\log \Lambda_n - n\mu]$,

$$\mathsf{E}_0[\tau] \approx \frac{-\log a - \alpha \log(b/a)}{K[f_0{:}f_1]} \qquad \mathsf{E}_1[\tau] \approx \frac{\log b - \beta \log(b/a)}{K[f_1{:}f_0]},$$

depending (no surprise) on how different f_0 and f_1 are. If *both* H_0 and H_1 are wrong, and $\{X_i\} \overset{\text{iid}}{\sim} f(x)$ for some density f that is *neither* f_0 nor f_1, the SPRT will typically choose whichever of f_0, f_1 is closer to f in the K-L sense, since $\Lambda_n \approx \exp(n\mu)$ with $\mu = K[f{:}f_0] - K[f{:}f_1]$.

To illustrate, as before let $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu, 1)$ and consider testing $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ at level $\alpha = 0.05$ for some $\mu_1 > \mu_0$, with Type-II error of $\beta \le 0.05$. The SPRT continues until $\Lambda_n \ge b = 19$ or $\Lambda_n < a = 1/19$. The K-L divergences are both $K[f_0{:}f_1] = K[f_1{:}f_0] = (\mu_1 - \mu_0)^2/2$, so the expected sample size is

$$\mathsf{E}\big[\tau \mid \{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu_j, 1)\big] \approx \frac{\log 19 - 0.05 \log 19^2}{(\mu_1 - \mu_0)^2/2} \approx \frac{5.23}{(\mu_1 - \mu_0)^2}$$

for both $j = 0$ and $j = 1$. For example, this is a sample of size 5.23 on average if $(\mu_1 - \mu_0) = 1.0$, or 523 for $(\mu_1 - \mu_0) = 0.10$.

7.4 A Bayesian perspective on SPRT

With equal prior probabilities $\pi_0 = \pi_1 = 1/2$ for H_0 and H_1, the posterior probability of H_0 will be

$$\mathsf{P}[H_0 \mid X_1, \cdots, X_\tau] = \frac{1}{1 + \Lambda_\tau} \approx \begin{cases} \frac{\alpha}{1+\alpha-\beta} & \Lambda_\tau \ge b \\ \frac{1-\alpha}{1-\alpha+\beta} & \Lambda_\tau \le a. \end{cases}$$

For $\alpha = \beta$, Bayesian error probabilities will be $\mathsf{P}[H_0 \mid \Lambda_\tau] = \alpha$ if H_0 is rejected and $\mathsf{P}[H_1 \mid \Lambda_\tau] = \beta$ if it isn't, exactly the same as the Frequentist error probabilities $\alpha = \mathsf{P}_0[\Lambda_\tau \ge b]$ and $\beta = \mathsf{P}_1[\Lambda_\tau \le a]$.

A Appendix: Sampling Distribution of Normal Statistics

Univariate

Let $\{X_i\}_{1 \le i \le n} \overset{\text{iid}}{\sim} \mathsf{No}(\mu, \sigma^2)$ be iid univariate normal random variables with uncertain mean μ and variance σ^2. The log likelihood

$$\begin{aligned} \ell(\mu, \sigma^2) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum (X_i - \mu)^2/\sigma^2 \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \sum (X_i - \bar{X}_n)^2/\sigma^2 - \frac{n}{2} (\bar{X}_n - \mu)^2/\sigma^2 \end{aligned}$$

attains its maximum at the MLE $(\hat{\mu}_n, \hat{\sigma}_n^2)$ given by:

$$\hat{\mu}_n = \bar{X}_n := \frac{1}{n} \sum X_i \qquad \hat{\sigma}_n^2 = \frac{1}{n} \sum (X_i - \bar{X}_n)^2.$$

In this section we'll find the joint probability distribution of $\bar{X}_n$ and $\hat{\sigma}_n^2$.

Introduce $Z_i := (X_i - \mu)/\sigma$, iid $\mathsf{No}(0,1)$ random variables, and define

$$T := \sum_{1\le i\le n} (Z_i)^2 \qquad\qquad S := \sum_{1\le i\le n} (Z_i - \bar{Z}_n)^2$$

and note that

$$\bar{X}_n = \hat{\mu}_n = \mu + \sigma \bar{Z}_n \qquad\qquad \hat{\sigma}_n^2 = (\sigma^2/n)\, S \tag{8}$$

and that

$$T = \sum (Z_i - \bar{Z}_n + \bar{Z}_n)^2 = \sum (Z_i - \bar{Z}_n)^2 + n\bar{Z}_n^2 = S + n\bar{Z}_n^2. \tag{9}$$

The characteristic function of Z^2 for $Z \sim \mathsf{No}(0,1)$ is

$$\chi(\omega) = \mathsf{E}e^{i\omega Z^2} = (2\pi)^{-1/2} \int_{\mathbb{R}} e^{i\omega z^2 - z^2/2}\, dz = (1-2i\omega)^{-1/2},$$

the ch.f. $(1 - i\omega/\lambda)^{-\alpha}$ of a Gamma $\mathsf{Ga}(\alpha,\lambda)$ random variable with shape $\alpha = 1/2$ and rate $\lambda = 1/2$. Thus T is the sum of n independent $\mathsf{Ga}(1/2,1/2)$ random variables and so has the $\mathsf{Ga}(n/2,1/2)$ distribution with ch.f. $(1-2i\omega)^{-n/2}$. Since $\sqrt{n}\bar{Z}_n \sim \mathsf{No}(0,1)$, also $n\bar{Z}_n^2 \sim \mathsf{Ga}(1/2,1/2)$. Finally, since $\bar{Z}_n$ and each $[Z_i - \bar{Z}_n]$ are normally-distributed with mean zero and covariance

$$\mathsf{E}\big[\bar{Z}_n(Z_i - \bar{Z}_n)\big] = (1/n) - (n/n^2) = 0,$$

$\bar{Z}_n$ is independent of each component of the vector $(Z - \bar{Z}_n)$ and hence of its squared length S. Thus by independence the ch.f. $\phi(\omega)$ of S satisfies

$$\begin{aligned} \mathsf{E}e^{i\omega T} &= \mathsf{E}e^{i\omega S} &&\times\, \mathsf{E}e^{i\omega n\bar{Z}_n^2} \\ (1-2i\omega)^{-n/2} &= \phi(\omega) &&\times\, (1-2i\omega)^{-1/2}. \end{aligned}$$

We conclude that $\phi(\omega) = (1-2i\omega)^{-(n-1)/2}$ and $S \sim \mathsf{Ga}\big((n-1)/2, 1/2\big)$, independent of $\bar{Z}_n \sim \mathsf{No}(0,1/n)$. From (8), we have independent MLEs:

$$\hat{\mu}_n \sim \mathsf{No}(\mu, \sigma^2/n) \quad \perp\!\!\!\perp \quad \hat{\sigma}_n^2 \sim \mathsf{Ga}\big(\frac{n-1}{2}, \frac{n}{2\sigma^2}\big).$$

Now set $Z := \sqrt{n}\bar{Z}_n \sim \mathsf{No}(0,1)$, note $Z^2 \sim \mathsf{Ga}(1/2,1/2)$ and $S \sim \mathsf{Ga}(\nu/2,1/2)$ for $\nu := (n-1)$ are independent, and consider the quantity

$$t := \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n-1}} = \frac{Z}{\sqrt{S/\nu}}.$$

The quantity t is called "pivotal" because its probability distribution doesn't depend on the parameters μ or σ^2. One way to find its pdf is to note that

$$\frac{1}{1+t^2/\nu} = \frac{S}{Z^2+S} \sim \mathsf{Be}\big(\frac{\nu}{2}, \frac{1}{2}\big)$$

and use a change-of-variables approach. The result is that t has "Gossett's (or Student's) t distribution with ν degrees of freedom" with pdf

$$f_\nu(t) = \left\{ \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \right\} (1 + t^2/\nu)^{-(\nu+1)/2}, \qquad t \in \mathbb{R}.$$

This symmetric bell-shaped curve converges to the standard normal density as $\nu \to \infty$ but has "heavier tails", *i.e.*, higher probability of values quite far from zero, since $f_\nu(t)$ falls off as $t \to \pm\infty$ only as a negative power of $|t|$, and not an exponential. In particular moments $\mathsf{E}|t|^p$ are only finite for $p < \nu$, so mean and variance are undefined or infinite unless $\nu > 1$ and $\nu > 2$, respectively. The special case $\nu = 1$ (with no mean) is the standard Cauchy distribution with pdf $f_1(t) = 1/\pi(1+t^2)$.

It is easy to construct confidence intervals for μ or GLRTs against $H_0 : \mu = \mu_0$ using $f_\nu(t)$. For example, for $n = 9$ we find $0.95 = \mathsf{P}[-2.306 < t_8 < 2.306]$, so

$$0.95 = \mathsf{P}[\bar{X}_9 - 2.306\hat{\sigma}_9/\sqrt{8} \le \mu \le \bar{X}_9 + 2.306\hat{\sigma}_9/\sqrt{8}]$$

is a 95% confidence interval for μ when σ^2 is unknown, based on the (sufficient) MLE statistics, while a two-sided GLRT of $\mu = \mu_0$ would reject at level α if $P(X) < \alpha$, where:

$$P(x) = 2\mathsf{P}[t_\nu \le -\sqrt{8}|\bar{X}_9 - \mu_0|/\hat{\sigma}_9],$$

e.g., would reject at level $\alpha = 0.05$ if $|\bar{X}_9 - \mu_0| \ge \hat{\sigma}_9(2.306/\sqrt{8})$.

Most calculators and statistical computing environments offer simple ways to compute the unbiased variance estimator

$$s_n^2 := \frac{1}{n-1}\sum_{i\le n}(X_i - \bar{X}_n)^2 = \left(\frac{n}{n-1}\right)\hat{\sigma}^2 \sim \mathsf{Ga}\left(\frac{\nu}{2}, \frac{\nu}{2\sigma^2}\right).$$

In those environments the t statistic is evaluated most easily as

$$t := \frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n-1}} = \frac{\sqrt{n}\,[\bar{X}_n - \mu]}{s_n},$$

an identity since $(n-1)s_n^2 = n\hat{\sigma}_n^2 = \sum(X_i - \bar{X}_n)^2$.

Multivariate

For p-variate normal data $\{X_i\} \overset{\text{iid}}{\sim} \mathsf{No}(\mu, \Sigma)$ with mean vector $\mu \in \mathbb{R}^p$ and $p \times p$ covariance matrix Σ, the maximum likelihood estimators are again of the same form,

$$\begin{aligned} \hat{\mu}_n &= \bar{X}_n & \hat{\Sigma}_n &= \frac{1}{n}\sum[X_i - \bar{X}_n][X_i - \bar{X}_n]^\intercal \\ &= \mu + \Sigma^{1/2}\bar{Z}_n & &= \frac{1}{n}\Sigma^{1/2}S\Sigma^{1/2} \end{aligned}$$

where $Z := \Sigma^{-1/2}[X - \mu]$, so $X = \mu + \Sigma^{1/2}Z$. Then

$$\bar{Z}_n := \frac{1}{n}\sum Z_i \qquad\qquad S := \sum[Z_i - \bar{Z}_n][Z_i - \bar{Z}_n]^\intercal$$

are independent with the p-variate $\bar{Z}_n \sim \mathsf{No}_p(0, \frac{1}{n}I)$ and Wishart $S \sim \mathsf{Wi}_p(I, n-1)$ distributions, respectively. The pivotal quantity

$$T := \sqrt{n-1}\, S^{-1/2}[\bar{X}_n - \mu]$$

has the "p-variate Student's t distributions" with $\nu = (n-1)$ degrees of freedom, and joint pdf

$$f_\nu(t) = \left\{ \frac{\Gamma\big(\frac{\nu+p}{2}\big)}{(\nu\pi)^{p/2}\,\Gamma\big(\frac{\nu}{2}\big)} \right\} \; \big(1 + |t|^2/\nu\big)^{-(\nu+p)/2}, \qquad t \in \mathbb{R}^p.$$

References

Abramowitz, M. and Stegun, I. A., eds. (1964), *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, *Applied Mathematics Series*, volume 55, Washington, D.C.: National Bureau of Standards, reprinted in paperback by Dover (1974); on-line at `http://www.math.sfu.ca/~cbm/aands/`.

Basu, D. (1955), "On Statistics independent of a Complete Sufficient Statistic," *Sankhyā-The Indian Journal of Statistics, Ser.* A, 15, 377–380.

Berger, J. O. (2003), "Could Fisher, Jeffreys and Neyman Have Agreed on Testing?" *Statistical Science*, 18, 1–32, doi:10.1214/ss/1056397485.

Berger, J. O., Brown, L. D., and Wolpert, R. L. (1994), "A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing," *Annals of Statistics*, 22, 1787–1807, doi:10.1214/aos/1176325757.

Berger, J. O. and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence (with discussion)," *Journal of the American Statistical Association*, 82, 112–122, doi:10.1080/01621459.1987.10478397.

Berger, J. O. and Wolpert, R. L. (1988), *The Likelihood Principle: A Review, Generalizations, and Statistical Implications (with discussion)*, *IMS Lecture Notes-Monograph Series*, volume 6, Hayward, CA: Institute of Mathematical Statistics, second edition, on-line at `http://projecteuclid.org/euclid.lnms/1215466210`.

Chernoff, H. and Lehmann, E. L. (1954), "The Use of Maximum Likelihood Estmates in χ^2 Tests for Goodness of Fit," *Annals of Mathematical Statistics*, 21, 579–586, doi:10.1214/aoms/1177728726.

Cox, D. R. (1958), "Some problems connected with statistical inference," *Annals of Mathematical Statistics*, 29, 357–372, doi:10.1214/aoms/1177706618.

DeGroot, M. H. and Shervish, M. J. (2012), *Probability and Statistics*, Boston, MA: Addison-Wesley, 4th edition.

Dunsmore, I. R., Daly, F., and the [M345 Statistical Methods] Course Team (1987), *Statistical Methods: Categorial Data. Unit 9*, Mathematics, a third level course (Open University), Belmont, CA: Open University Press.

Fisher, R. A. (1935a), *The Design of Experiments*, Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1935b), "Mathematics of a Lady Tasting Tea," in *The World of Mathematics*, ed. J. R. Newman, Courier Dover, volume 3: Design of Experiments, originally from (Fisher, 1935a).

Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, Edinburgh, UK: Oliver and Boyd.

Goodman, S. N. (2001), "Of P-Values and Bayes: A Modest Proposal," *Epidemiology*, 12, 295–297.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press, 3rd edition, first edition published in 1939.

Lehmann, E. L. (1993), "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, 88, 1242–1249, doi:10.1080/01621459.1993.10476404.

Neyman, J. and Pearson, E. S. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society. Series A*, 231, 289–337, doi:10.1098/rsta.1933.0009.

Pearson, K. (1900), "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine Series 5*, 50, 157–175, doi:10.1080/14786440009463897.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of P-values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71, doi:10.1198/000313001300339950.

Wald, A. (1945), "Sequential tests of statistical hypotheses," *Annals of Mathematical Statistics*, 16, 117–186, doi:10.1214/aoms/1177731118.

Ware, J. H. (1989), "Investigating Therapies of Potentially Great Benefit: ECMO," *Statistical Science*, 4, 298–340, (With discussion).

Wasserstein, R. L. and Lazar, N. A. (2016), "The ASA's sttement on p-values: context, process, and purpose," *The American Statistician*, doi:10.1080/00031305.2016.1154108.

Wolpert, R. L. and Mengersen, K. L. (2004), "Adjusted Likelihoods for Synthesizing Empirical Evidence from Studies That Differ in Quality and Design: Effects of Environmental Tobacco Smoke," *Statistical Science*, 19, 450–471, doi:10.1214/088342304000000350.