Statistics Notes – I.

Moulinath Banerjee

University of Michigan

March 8, 2004

1 The Gamma and χ^2 distributions:

The Gamma distribution: The gamma function is a real-valued non-negative function defined on $(0, \infty)$ in the following manner

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x}\, dx \quad , \quad \alpha > 0\,.$$

The Gamma function enjoys some nice properties. Two of these are listed below:

$$(a)\ \Gamma(\alpha+1) = \alpha\,\Gamma(\alpha) \quad , \quad (b)\ \Gamma(n) = (n-1)!\ \ (n \ \text{ integer})\,.$$

Property (b) is an easy consequence of Property (a). Start off with $\Gamma(n)$ and use Property (a) recursively along with the fact that $\Gamma(1) = 1$ (why?). Another important fact is that $\Gamma(1/2) = \sqrt{\pi}$. To prove property (a), use integration by parts.

The Gamma distribution with parameters $(\alpha > 0, \lambda > 0)$ (denoted by $\Gamma(\alpha, \lambda)$) is defined through the following density function:

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

Check that this is a bona–fide density function. That it is positive is immediate; that it integrates to 1 can be checked by evaluating the integral $\int_0^\infty f(x, \alpha, \lambda)\, dx$ through the substitution $y = \lambda x$ and using the definition of $\Gamma(\alpha)$ (Verify!). The first parameter α is called the *shape* parameter and the second parameter λ is called the *scale* parameter. For fixed λ the shape parameter regulates the shape of the gamma density. Here is a simple exercise that justifies the term "scale parameter" for λ.

Exercise: Let X be a random variable following $\Gamma(\alpha, \lambda)$. Then show that $Y = \lambda X$ (thus X is Y scaled by λ) follows the $\Gamma(\alpha, 1)$ distribution. What is the distribution of cX for some arbitrary positive constant c? You can use the change of variable theorem in 1 dimension to work this out.

Reproductive Property of the Gamma distribution: Let $X_1, X_2, \ldots, X_n$ be independent random variables with $X_i \sim \Gamma(\alpha_i, \lambda)$. Then $S_n \equiv X_1 + X_2 + \ldots + X_n$ is distributed as $\Gamma(\sum_{i=1}^n \alpha_i, \lambda)$.

If X follows the Gamma(α, λ) distribution, the mean and variance of X can be explicitly expressed in terms of the parameters:

$$E(X) = \frac{\alpha}{\lambda} \text{ and } \text{Var}(X) = \frac{\alpha}{\lambda^2}.$$

I will outline the computation of a general moment $E(X^k)$, where k is a positive integer. We have,

$$\begin{aligned} E(X^k) &= \int_0^\infty x^k \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{-\lambda x} x^{k+\alpha-1} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}} \\ &= \frac{(\alpha+k-1)\ldots(\alpha)\,\Gamma(\alpha)}{\lambda^k\,\Gamma(\alpha)} \\ &= \frac{\Pi_{i=1}^k (\alpha+i-1)}{\lambda^k}. \end{aligned}$$

The formulae for the mean and the variance should follow directly from the above computation. Note that in the above derivation, we have used the fact that

$$\int_0^\infty e^{-\lambda x} x^{k+\alpha-1} dx = \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}}.$$

This is an immediate consequence of the fact that the gamma density with parameters $(\alpha+k, \lambda)$ integrates to 1.

Exercise: Here is an exercise that should follow from the discussion above. Let S_n be a random variable following the Gamma distribution with parameters (n, λ), where $\lambda > 0$. Show that for large n, the distribution of S_n is well approximated by a normal distribution (with parameters that you need to identify).

We now introduce an important family of distributions, called the chi-squared family. To do so, we first define the chi-squared distribution on 1 degree of freedom (for brevity, we call it "chi-squared one" and write it as χ_1^2).

The χ_1^2 distribution: Let Z follow $N(0,1)$. Then the distribution function of $W = Z^2$ is called the χ_1^2 distribution and W itself is called a χ_1^2 random variable.

Exercise: Show that W follows a Gamma(1/2, 1/2) distribution. (You can do this by working out the density function of W from that of Z – refer to the relevant problem in Homework 1). Remember that $\Gamma(1/2)$ is $\sqrt{\pi}$.

For any integer $d > 0$ we can now define the χ^2_d distribution (chi-squared d distribution, or equivalently, the chi-squared distribution on d degrees of freedom).

The χ^2_d distribution: Let $Z_1, Z_2, \ldots, Z_d$ be i.i.d. $N(0,1)$ random variables. Then the distribution function of $W_d = Z_1^2 + Z_2^2 + \ldots + Z_d^2$ is called the χ^2_d distribution and W_d itself is called a χ^2_d random variable.

Exercise: Using the reproductive property of the Gamma distribution, show that W_d follows $\Gamma(d/2, 1/2)$.

Thus, it follows that the sum of k i.i.d. χ^2_1 random variables is a χ^2_k random variable. Here is an exercise that will require you to work with χ^2 distributions.

Exercise: Let Z_1, Z_2, Z_3 be i.i.d. $N(0,1)$ random variables. Consider the vector (Z_1, Z_2, Z_3) as a random point in 3–dimensional space. Let R be the length of the radius vector connecting this point to the origin. Find the density functions of (a) R and (b) R^2.

2 Sampling from a Normal Population

Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from an underlying normal population $N(\mu, \sigma^2)$. You could think of the X_i's for example as a set of randomly sampled SAT scores from the entire population of SAT scores. Then μ is the average SAT score of the entire population and σ^2 is the variance of SAT scores in the entire population. We are interested in estimating μ and σ^2 based on the data. Note that SAT scores are actually discrete in nature – $N(\mu, \sigma^2)$ provides a good approximation to the actual population distribution. In other words, $N(\mu, \sigma^2)$ is the model that we use for the SAT scores. In Statistics as in any other science, models are meant to provide insightful approximations to the true underlying nature of reality.

Natural estimates of the mean and the variance are given by:

$$\hat{\mu} = \overline{X}_n = \frac{X_1 + X_2 + \ldots + X_n}{n} \text{ and } \hat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 .$$

These are the *sample mean* and *sample variance*. In what follows, we will use a slightly different estimate of σ^2 than the one proposed above. We will use,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 .$$

One reason for using s^2 is that it has a natural interpretation as the multiple of a χ^2 random variable; further s^2 is an *unbiased estimator* of σ^2 whereas $\hat{\sigma}^2$ is not – i.e.,

$$E(s^2) = \sigma^2 \;\text{ but }\; E(\hat{\sigma}^2) \neq \sigma^2 \,.$$

We will explore the term "unbiasedness" in details later. For the sake of notational simplicity we will let S^2 denote the *residual sum of squares about the mean*, i.e. $\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Here is an interesting (and fairly profound) proposition.

Proposition: Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from some distribution F with mean μ and variance σ^2. Then F is the $N(\mu, \sigma^2)$ distribution if and only if for all n, $\overline{X}$ and s^2 are independent random variables.

The "if" part is the profound part. It says that the independence of the natural estimates of the mean and the variance for any sample size forces the underlying distribution to be normal. This is in keeping with the fact that for the normal distribution, μ and σ^2 are *variation independent* parameters. Knowledege of μ (which measures the central location of the normal curve) provides no information about the spread of the curve around its center, which is described by the variance σ^2. We will sketch a proof of the only if part.

To this end, define new random variables $Y_1, Y_2, \ldots, Y_n$ where for each i, $Y_i = (X_i - \mu)/\sigma$. These are the *standardized versions* of the X_i's and are i.i.d. $N(0,1)$ random variables. Now, note that:

$$\overline{X} = \overline{Y}\,\sigma + \mu \;\text{ and }\; s^2 = \frac{\sigma^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}{(n-1)} \,.$$

From the above display, we see that it suffices to show the independence of $\overline{Y}$ and $\sum_{i=1}^n (Y_i - \overline{Y})^2$. The way this proceeds is outlined below: Let Y denote the $n \times 1$ column vector $(Y_1, Y_2, \ldots, Y_n)^T$ and let P be an $n \times n$ orthogonal matrix with the first row of P (which has length n) being $(1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n})$. Recall that an orthogonal matrix satisfies $P^T P = P P^T = I$ where I is the identity matrix. Using standard linear algebra techniques it can be shown that such a P can always be constructed. Now define a new random vector $W = PY$. Then, by an extended version of the change of variable theorem for 2 dimensions, it can be established that the random vector $W = (W_1, W_2, \ldots, W_n)$ has the same distribution as $(Y_1, Y_2, \ldots, Y_n)$; in other words, $W_1, W_2, \ldots, W_n$ are i.i.d. $N(0,1)$ random variables. Note that $W^T W = (PY)^T PY = Y^T P^T P Y = Y^T Y$ by the orthogonality of P – in other words, $\sum_{i=1}^n W_i^2 = \sum_{i=1}^n Y_i^2$. Also,

$$W_1 = Y_1/\sqrt{n} + Y_2/\sqrt{n} + \ldots + Y_n/\sqrt{n} = \sqrt{n}\,\overline{Y} \,.$$

Note that W_1 is independent of $W_2^2 + W_3^2 + \ldots W_n^2$. But

$$\sum_{i=2}^n W_i^2 = \sum_{i=1}^n W_i^2 - W_1^2 = \sum_{i=1}^n Y_i^2 - n\,\overline{Y}^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 \,.$$

The last equality can be checked by simply expanding the square in the last display. It therefore follows that $\sqrt{n}\,\overline{Y}$ and $\sum_{i=1}^n (Y_i - \overline{Y})^2$ are independent – which implies that $\overline{Y}$ and $\sum_{i=1}^n (Y_i - \overline{Y})^2$ are independent. This is what we sought to show in the first place.

Note that $\overline{Y}$ follows $N(0, 1/n)$. Deduce that $\overline{X}$ follows $N(\mu, \sigma^2/n)$. Since $\sum_{i=1}^n (Y_i - \overline{Y})^2 = W_2^2 + W_3^2 + \ldots W_n^2$, it follows that

$$S^2/\sigma^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 \sim \chi^2_{n-1}\,.$$

This follows from the exercises on χ^2 distributions in the previous section. Thus,

$$s^2 \equiv \frac{S^2}{n-1} \equiv_d \frac{\sigma^2}{n-1}\, K_{n-1}, \tag{2.1}$$

where K_{n-1} is a χ^2_{n-1} random variable. In the above display the symbol $\equiv_d$ means "is equal in distribution to".

In the case $n = 2$, it is easy to check the details of the transformation leading from Y to W. Set $W = P\,Y$ with

$$P = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}.$$

Thus $W_1 = (Y_1 + Y_2)/\sqrt{2}$ and $W_2 = (Y_1 - Y_2)/\sqrt{2}$.

Exercise: Use the change of variable theorem to deduce that W_1 and W_2 are i.i.d. $N(0, 1)$.

We will next introduce some important statistical distributions that arise naturally in connection with making inference on the parameters of normal populations. For a single normal population, there are two quantities of interest – μ and σ^2. Typically there are two different (but related problems) that we are interested in. One is to determine a range of "plausible values" for the unknown μ or σ^2, based on our data. The other is to *test a (null) hypothesis* of the type that $\mu = \mu_0$ or $\sigma^2 = \sigma_0^2$. The first problem is that of finding *a confidence set* in statistical jargon; the second is that of *Hypothesis Testing* and we will discuss both. Often, instead of one normal population, we will be interested in comparing two normal populations, where the first is $N(\mu_1, \sigma_1^2)$ and the second is $N(\mu_2, \sigma_2^2)$ (think about SAT scores for two different years, well apart). Natural questions of interest can be formulated as: (a) Did the average SAT score change from one year to the other? In other words, is $\mu_1 - \mu_2 = 0$ or not? (b) How does the variability in one year compare to that in the second year? In other words, what can we say about the ratio of variances σ_1^2/σ_2^2? Statisticians will typically be interested in finding a confidence interval for $\mu_1 - \mu_2$ or σ_1^2/σ_2^2; they may also want to test hypotheses of the kind $\mu_1 - \mu_2 = 0$ or $\sigma_1^2/\sigma_2^2 = 1$.

The standardized sample mean

$$\overline{X}^\star_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma}$$

plays an important role in constructing confidence sets for μ or developing hypothesis tests for testing whether $\mu = \mu_0$. However, if σ is unknown, one needs to replace it in the above display by its natural estimate s. This gives, what is called the "t-statistic"

$$t - statistic = \frac{\sqrt{n}\,(\overline{X} - \mu)}{s}\,.$$

To use the above quantity for testing or construction of confidence sets requires us to know its distribution – this is where the t family of distributions (to be introduced shortly) come in handy. The other family of distributions – the F family that we will discuss is useful for comparing population variances when there is more than one normal population. It is also useful for checking the fit of a model in regression analysis (something we will not have time to discuss in this course).

The t and F distributions: Let U follow $N(0,1)$ and let V follow χ^2_n on n degrees of freedom, independently of each other. Then,

$$T = \frac{U}{\sqrt{V/n}}$$

is said to follow the t distribution on n degrees of freedom. We write $T \sim t_n$. The density of the T distribution is derived in Rice's book (Chapter 6). With a little bit of patience, you can also work it out, using the Change of Variable Theorem appropriately (I won't go into the computational details here). Before we go any further, here is a short exercise.

Exercise: Let X be a random variable that is distributed symmetrically about 0, i.e X and -X have the same distribution function (and hence the same density function). If f denotes the density, show that it is an even function, i.e. $f(x) = f(-x)$ for all x.

Conversely, if the random variable X has a density function f that is even, then it is symmetrically distributed about 0, i.e $X \equiv_d -X$.

Here are some important facts about the T distribution.

(a) T and $-T$ have the same distribution. Thus, the distribution of T is symmetric about 0 and it has an even density function.

From definition,

$$-T = \frac{-U}{\sqrt{V/n}} = \frac{\tilde{U}}{\sqrt{V/n}}\,,$$

where $\tilde{U} \equiv -U$ follows $N(0,1)$, and is independent of V where V follows χ^2_n. Thus, by definition, $-T$ also follows the t distribution on n degrees of freedom.

(b) As $n \to \infty$, the t_n distribution converges to the $N(0,1)$ distribution; hence the quantiles of the t distribution are well approximated by the quantiles of the normal distribution.

This follows from the law of large numbers. Consider the term V/n in the denominator of T for large n. Now as V follows χ^2_n it has the same distribution as $K_1 + K_2 + \ldots + K_n$ where K_i's are i.i.d. χ^2_1 random variables. But by the WLLN we know that

$$\frac{K_1 + K_2 + \ldots + K_n}{n} \rightarrow_p E(K_1) = 1 \text{ (check!)}.$$

Thus V/n converges in probability to 1; hence the denominator in T converges in probability to 1 and T consequently, converges in distribution to U where U is $N(0,1)$.

Next we define the F distribution. Let U and V be independent χ^2_m and χ^2_n random variables respectively. Then

$$\tilde{F} = \frac{U/m}{V/n}$$

is said to follow the F distribution on m and n degrees of freedom. We write $\tilde{F} \sim F_{m,n}$. Here are some exercises (almost definitional) that you should try and work out.

Exercise: (a) If $\tilde{F}$ follows $F_{m,n}$ then $1/\tilde{F}$ follows $F_{n,m}$. (b) If T follows t_n then T^2 follows $F_{1,n}$.

A brief digression: Consider a random vector (X_1, X_2) on the plane. Consider the transformation: $(X_1, X_2) \mapsto (Y_1, Y_2)$ where

$$Y_1 = \frac{a\,X_1 + b\,X_2}{\sqrt{a^2 + b^2}}\,, \quad Y_2 = \frac{b\,X_1 - a\,X_2}{\sqrt{a^2 + b^2}}\,.$$

Here a and b are any two numbers. Writing $Y = (Y_1, Y_2)$ and $X = (X_1, X_2)$, show that we can write $Y = P\,X$ for some 2×2 matrix P, where P is orthogonal.

Show that geometrically the action of P on X is to keep the length of X (this is $R \equiv \sqrt{X_1^2 + X_2^2}$) intact but to rotate it by some angle θ in a particular direction that you need to identify. What is then the action of P^T on Y, geometrically?

Suppose that the joint density of (X_1, X_2) is $f(x_1, x_2) = K\,g(x_1^2 + x_2^2)$ for some positive constant K and some real valued function g. Then, how is the joint distribution of (Y_1, Y_2) related to the joint distribution of (X_1, X_2)? What is the correlation between X_1 and X_2? Are X_1 and X_2 necessarily independent? When are they independent?

Let (X_1, X_2) be expressed in terms of polar coordinates (R, θ). Thus $X_1 = R\ cos\,\theta$ and $X_2 = R\ sin\,\theta$. Compute the joint and marginal distributions of R and θ. Are they independent?

2.1 Confidence Sets (exact and approximate) using *pivots*

In this subsection, we will discuss the construction of confidence sets for unknown parameters of interest that describe some population from which we have data. As a concrete example,

consider once again an i.i.d. sample from the population of SAT scores in the year 1988. Call these $X_1, X_2, \ldots, X_n$. We will model the X_i's as $N(\mu, \sigma^2)$ random variables, where μ is the true mean SAT score (in the entire population) and σ^2 is the population variance of the scores. We are interested in finding a range of plausible values for the parameters μ and σ^2.

Here is a more abstract formulation.

The Method of Pivots: Given i.i.d. observations $Y_1, Y_2, \ldots, Y_n$ from some underlying distribution F_θ where θ is an indexing parameter (different values of θ correspond to different distributions – in the SAT example θ is the vector (μ, σ^2) and F_θ is the $N(\mu, \sigma^2)$ distribution), we seek to find a set of plausible values for some real valued function of θ, say $h(\theta)$. In statistical jargon, the set of plausible values will be called a *confidence set* and the degree of plausibility will be called the *level of confidence*. For a small fraction α (0.01,0.05,0.1, typically), a level $1-\alpha$ confidence set for $h(\theta)$ is defined as any (random) subset of the real line, $\mathcal{S}(Y)$, say, such that

$$\text{Prob}_\theta\left(h(\theta) \in \mathcal{S}(Y)\right) = 1 - \alpha\,.$$

Thus, the confidence set itself will depend on the observed data Y.

Often a standard method of constructing confidence sets is the following *method of pivots* which we describe below.

(1) Construct a function Ψ using the data Y and $h(\theta)$, say $\Psi(Y, h(\theta))$ such that the distribution of this random variable under parameter value θ *does not depend on* θ and is known. Such a Ψ is called a *pivot*.

(2) Let G denote the distribution function of the pivot. The idea now is to get a range of plausible values of the pivot. The level of confidence $1-\alpha$ is to be used to get the appropriate range. This can be done in a variety of ways but the following is standard. Denote by $q(G;\beta)$ the β'th quantile of G. Thus,

$$\text{Prob}_\theta(\Psi(Y, h(\theta)) \leq q(G;\beta)) = \beta\,.$$

(3) Choose $0 \leq \beta_1, \beta_2 \leq \alpha$ such that $\beta_1 + \beta_2 = \alpha$. Then,

$$\text{Prob}_\theta(q(G;\beta_1) \leq \Psi(Y, h(\theta)) \leq q(G; 1-\beta_2)) = 1 - \beta_2 - \beta_1 = 1 - \alpha\,.$$

(4) Vary θ across its domain and choose your level $1-\alpha$ confidence set $\mathcal{S}(Y)$ as the set of all $h(\theta)$ such that the two inequalities in the above display are simultaneously satisfied.

The above general prescription may seem a bit hairy at this stage, but will look easier after we apply it to several examples.

Example 1: Find a level $1-\alpha$ confidence interval for μ from data $X_1, X_2, \ldots, X_n$ which

are i.i.d. $N(\mu, \sigma^2)$ where σ is **known**. Here $\theta = \mu$ and $h(\theta) = \mu$. We want to construct $\Psi(X_1, X_2, \ldots, X_n, \mu)$ such that the distribution of this object is known to us. How do we proceed here? The usual way is to find some decent estimator of μ and combine it along with μ in some way to get a pivot. The most intuitive estimate of μ here is the sample mean $\overline{X}_n$. We know that $\overline{X}_n \sim N(\mu, \sigma^2/n)$. The standardized version of the sample mean follows $N(0,1)$ and can therefore act as a pivot. In other words, construct,

$$\Psi(X, \mu) = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma}\,.$$

This follows $N(0,1)$ (which is our G). With z_β denoting the upper β'th quantile of $N(0,1)$ (i.e. $P(Z > z_\beta) = \beta$ where Z follows $N(0,1)$) and splitting $\alpha = \beta_1 + \beta_2$ where $\beta_1 = \beta_2$, we readily obtain: $q(G, \beta_1) = -z_{\alpha/2}$ and $q(G, 1-\beta_2) = z_{\alpha/2}$. Here we have used the symmetry of the normal curve about 0.

We can thus write:

$$P_\mu\left(-z_{\alpha/2} \leq \frac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma} \leq z_{\alpha/2}\right) = 1 - \alpha\,.$$

From the above display we can find limits for μ such that the above inequalities are simultaneously satisfied. On doing the algebra, we get:

$$P_\mu\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\right) = 1 - \alpha\,.$$

Thus our level $1-\alpha$ C.I. for μ is given by:

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\ ,\ \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\right]\,.$$

Example 2: The data are the same as in Example 1 but now σ^2 is no longer known. Thus, the parameter of unknowns $\theta = (\mu, \sigma^2)$ and we are interested in finding a confidence set for $h(\theta) = \mu$. Clearly, setting

$$\Psi(X, \mu) = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

will not work smoothly here. This certainly has a known ($N(0,1)$) distribution but involves the *nuisance parameter* σ making it difficult get a confidence set for μ directly. However, one can rep lace σ by s, where s^2 is the natural estimate of σ^2 introduced before. So, set:

$$\Psi(X, \mu) = \frac{\sqrt{n}(\overline{X} - \mu)}{s}\,.$$

This only depends on the data and $h(\theta) = \mu$. We claim that this is indeed a pivot. To see this write

$$\frac{\sqrt{n}(\overline{X} - \mu)}{s} = \frac{\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}}{\sqrt{s^2/\sigma^2}}\,.$$

The numerator on the extreme right of the above display follows $N(0,1)$ and the denominator is independent of the numerator and is the square root of a χ^2_{n-1} random variable over its degrees of freedom (from display (2.1)). It follows from definition that $\Psi(X,\mu)$ as finally defined follows a t_{n-1} distribution. Thus, G here is the t_{n-1} distribution and we can choose the quantiles to be $q(t_{n-1};\alpha/2)$ and $q(t_{n-1};1-\alpha/2)$. By symmetry of the t_{n-1} distribution about 0, we have, $q(t_{n-1};\alpha/2) = -q(t_{n-1};1-\alpha/2)$. It follows that,

$$\text{Prob}_{\mu,\sigma^2}\left[-q(t_{n-1};1-\alpha/2) \le \frac{\sqrt{n}(\overline{X}-\mu)}{s} \le q(t_{n-1};1-\alpha/2)\right] = 1-\alpha\,.$$

As with Example 1, direct algebraic manipulations show that this is the same as the statement:

$$\text{Prob}_{\mu,\sigma^2}\left[\overline{X} - \frac{s}{\sqrt{n}}\,q(t_{n-1};1-\alpha/2) \le \mu \le \overline{X} + \frac{s}{\sqrt{n}}\,q(t_{n-1};1-\alpha/2)\right] = 1-\alpha\,.$$

This gives a level $1-\alpha$ confidence set for μ.

Example 3 The data are as in Example 2 but now we want to find a confidence interval for σ^2. Thus μ is now our nuisance parameter (of no interest). I will outline the procedure here and leave the rest to you. Show that,

$$\Psi(X,\sigma^2) = \frac{(n-1)\,s^2}{\sigma^2}$$

is a pivot. Identify its distribution and use the quantiles of this distribution to get a level $1-\alpha$ C.I. for σ^2.

Food for thought: In each of the above examples there are innumerable ways of decomposing α as $\beta_1+\beta_2$. It turns out that when α is split equally the level $1-\alpha$ confidence intervals obtained in Examples 1 and 2 are the shortest. Shorter confidence intervals are more informative, hence this is an optimal course of action. What happens when we take either β_1 or β_2 to be 0 in the above three examples?

What are desirable properties of confidence sets? On one hand, we require high levels of confidence; in other words, we would like α to be as small as possible. On the other hand we would like our confidence sets to be *sharp*, i.e. have less volume. If we are talking about confidence intervals (C.I.'s), as is the case in the above three examples, we want the shortest possible ones. Unfortunately, we cannot simultaneously make the confidence levels of our C.I's go up and the lengths of our C.I's go down. In Example 1, the length of the level $1-\alpha$ confidence interval is $2\,\sigma\,z_{\alpha/2}/\sqrt{n}$. As we reduce α (for higher confidence), $z_{\alpha/2}$ increases, making the confidence interval wider. However, we can reduce the length of our confidence interval for a fixed α by increasing the sample size. If my sample size is 4 times yours, I will end up with a C.I. which has the same level as yours but has half the length of your C.I.. Can we hope to get absolute confidence, i.e. $\alpha = 0$? That is too much of an ask. When $\alpha = 0$, $z_{\alpha/2} = \infty$ and the C.I's for μ are infinitely large. The same can be verified for Examples 2 and 3.

Asymptotic pivots using the Central Limit Theorem: The CLT allows us to construct an approximate pivot for large sample sizes for estimating the population mean μ for any underlying distribution F.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. observations from some common distribution F and let $E(X_1) = \mu$ and $\text{Var}(X_1) = \sigma^2$. We are interested in constructing an approximate level $1-\alpha$ C.I. for μ. By the CLT we have $\overline{X} \sim_{appx} N(\mu, \sigma^2/n)$ for large n; in other words,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim_{appx} N(0,1)\,.$$

If σ is known the above quantity is an approximate pivot and following Example 1, we can therefore write,

$$P_\mu\left(-z_{\alpha/2} \leq \frac{\sqrt{n}\,(\overline{X} - \mu)}{\sigma} \leq z_{\alpha/2}\right) =_{appx} 1-\alpha\,.$$

As before, this translates to

$$P_\mu\left(\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2} \leq \mu \leq \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\right) =_{appx} 1-\alpha\,.$$

This gives an approximate level $1-\alpha$ C.I. for μ when σ is known. The approximation will improve as the sample size n increases. Note that the true coverage of the above C.I. will be different from $1-\alpha$ and can depend heavily on the nature of F and the sample size n.

Realistically however σ is unknown and is replaced by s. Since we are dealing with large sample sizes, s is with very high probability close to σ and the interval

$$\left(\overline{X} - \frac{s}{\sqrt{n}}\, z_{\alpha/2}\ ,\ \overline{X} + \frac{s}{\sqrt{n}}\, z_{\alpha/2}\right)\,,$$

still remains an approximate level $1-\alpha$ C.I..

Exercise: Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(θ). The sample size n is large. Thus $E(X_1) = \theta$ and $\text{Var}(X_1) = \theta\,(1-\theta)$. We want to find a level $1-\alpha$ C.I. (approximate) for θ. Note that both mean and variance are unknown. Show that if $\hat{\theta}$ is natural estimate of θ obtained by computing the sample proportion of 1's, then

$$\left[\hat{\theta} - \frac{\hat{\theta}\,(1-\hat{\theta})}{\sqrt{n-1}}\, z_{\alpha/2}\ ,\ \hat{\theta} + \frac{\hat{\theta}\,(1-\hat{\theta})}{\sqrt{n-1}}\, z_{\alpha/2}\right]$$

is an approximate level $1-\alpha$ C.I. for θ.

There are alternative ways of computing an approximate level $1-\alpha$ C.I. here. One way

out is to use variance stabilizing transformations that we may discuss later. The other is to note that by the CLT:

$$\frac{\sqrt{n}\,(\hat{\theta}-\theta)}{\sqrt{\theta\,(1-\theta)}} \sim_{appx} N(0,1)\,,$$

so that

$$P_\theta\left[-z_{\alpha/2} \leq \frac{\sqrt{n}\,(\hat{\theta}-\theta)}{\sqrt{\theta\,(1-\theta)}} \leq z_{\alpha/2}\right] =_{appx} 1-\alpha\,.$$

An approximate level $1-\alpha$ C.I. can now be obtained by solving for all θ for which the above inequalities are both satisfied. This amounts to solving a quadratic and will yield a different C.I. than the one proposed in the Exercise. You should try to work out what this gives you.

The construction of confidence sets for a parameter of interest using the above (pivotal) method has an intimate connection with hypothesis testing in Statistics. In fact, there is a nice duality between the two procedures, as we will see shortly. Level $1-\alpha$ confidence sets for a parameter lead naturally to what are called level α tests for testing a (null) hypothesis of the form $\theta = \theta_0$. On the other hand, a level α test for a null hypothesis of the form $\theta = \theta_0$ leads naturally to a level $1-\alpha$ confidence region for θ. In order to understand the duality, we now introduce you to the basic principles of Hypothesis Testing through some simple examples and illustrate the connection with the construction of confidence sets.

3 An Introduction to Testing Statistical Hypothesis

The problem of testing statistical hypothesis can be posed as follows. Provided are data on a number of individuals and a stochastic model that is hypothesized to have generated the data. This is the so-called the "null hypothesis". Is there enough evidence in the data against the null hypothesis (in which case we reject it) or should we continue to stick to it? Such questions arise very naturally in many different fields of application.

Testing for a normal mean: Let us consider a simple example. Suppose that $X_1, X_2, \ldots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution and let, initially, σ^2 be known. We want to test the *null hypothesis* $H_0 : \mu = \mu_0$. Thus, prior to having observed the data we postulate the null hypothesis that $\mu = \mu_0$; we then observe the data and based on this data seek to make an informed decision as to whether we should reject our null hypothesis or retain it. Note that in this case, we have tacitly formulated an *alternative hypothesis*. This is the complement of the null hypothesis. Denoting it by H_1, we have $H_1 : \mu \neq \mu_0$. Rejecting H_0 puts us in the domain of the alternative hypothesis.

For concreteness, $X_1, X_2, \ldots, X_n$ could be the heights of n individuals in some tribal population. The distribution of heights in a (homogeneous) population is usually normal, so that a $N(\mu, \sigma^2)$ model is appropriate. If we have some a–priori reason to believe that the average height in this population is around 60 inches, we could postulate a null hypothesis of the form $H_0 : \mu = \mu_0 \equiv 60$; the alternative hypothesis is $H_1 : \mu \neq 60$. We want to *test* H_0 *against* H_1. How do we do the

test? To this end, we construct a *test function* $\phi(X)$ where $X = (X_1, X_2, \ldots, X_n)$ and which takes values between 0 and 1. The test function $\phi(X)$ gives us the *conditional probability of rejecting* H_0 *given the data* X. Thus, if for some X, $\phi(X) = 1$, we reject our null hypothesis; if for some X, $\phi(X) = 0$ we do not reject our null hypothesis and if for some $X, 0 < \phi(X) < 1$ we perform a Bernoulli experiment with probability of success equal to $\phi(X)$ and reject if we have a success and do not reject otherwise. In situations where the underlying distribution is continuous (as with our current example) the third situation where $\phi(X)$ is strictly between 0 and 1 is generally not going to arise. Thus the data space, in such cases, can be partitioned into $\mathcal{R} \equiv \{x : \phi(x) = 1\}$, which is called the *rejection region* and its complement $\mathcal{A} \equiv \{x : \phi(x) = 0\}$, the *acceptance region*. The term "accceptance region" must however be taken with a grain of salt. If x, the observed data lies in the acceptance region we do not reject the null hypothesis – this, by no means, implies that we have accepted it. The null hypothesis, simply reflects our current state of belief which might not be true and we retain if the data do not provide sufficient evidence that it is wrong.

Associated with the test function ϕ are two different kinds of error that we can commit. These are called *Type 1 error* and *Type 2 error*. Type 1 error occurs if we reject the null hypothesis when actually the null hypothesis is true. Type 2 error occurs if we do not reject the null hypothesis when actually the null hypothesis is false. From the definition of ϕ you can see that

$$P(Type\ 1\ error) = \int_{\mathcal{X}} \phi(x)\, f_{\mu_0}(x)\, dx = E_{\mu_0}(\phi(X))\,,$$

where the expression on the extreme right of the above display denotes the expectation of $\phi(X)$ when the data X is an i.i.d. sample from a $N(\mu_0, \sigma^2)$ distribution and $f_{\mu_0}(x)$ is the joint density of $X = (X_1, X_2, \ldots, X_n)$ at the point $x = (x_1, x_2, \ldots, x_n)$. In our current example, the null hypothesis is *simple* i.e. the only possible distribution for the data under the null is $N(\mu_0, \sigma^2)$. If the null hypothesis is *composite* i.e. more than one distribution is possible under the null, then the probability of Type 1 error will actually depend on the underlying distribution from the null. We will illustrate this with a different example. In the current case, what is $f_{\mu_0}(x)$? Clearly,

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

the joint density of n i.i.d. $N(\mu_0, \sigma^2)$ random variables. If the test ϕ is simply a 0–1 valued test function then,

$$\alpha \equiv P(Type\ 1\ error) = \mathrm{Prob}_{\mu_0}\left((X_1, X_2, \ldots, X_n) \in \mathcal{R}\right).$$

What about Type 2 error? Instead of the Type 2 error we are going to deal with an equivalent quantity which is 1 − Probability of *Type 2 error*. We will denote this quantity by β. Thus β is the probability of rejecting the null hypothesis when the null hypothesis is actually incorrect and is called the *power* of the test ϕ. In our current example, β will clearly depend on what specific distribution in the alternative hypothesis we want to look at. If H_0 is false, then the data come from a $N(\mu, \sigma^2)$ distribution with $\mu \neq \mu_0$. If $f_\mu(x)$ denotes the joint density of a sample $(X_1, X_2, \ldots, X_n)$

from a $N(\mu, \sigma^2)$ distribution then,

$$\beta(\mu) = \int_{\mathcal{X}} \phi(x)\, f_\mu(x)\, dx\,.$$

Once again, if we have a 0–1 valued test function ϕ,

$$\beta(\mu) = \text{Prob}_\mu\ ((X_1, X_2, \ldots, X_n) \in \mathcal{R})\ .$$

Note that even though we have introduced the quantity $\beta(\mu)$ in connection with the alternative hypothesis, $\beta(\mu)$ can be defined for all real numbers μ and $\beta(\mu_0)$ is precisely the probability of Type 1 error, α. We will henceforth call $\beta(\mu)$, the power function of the test ϕ. For $\mu = \mu_0$, the power function gives us the chance of Type 1 error i.e. making the wrong decision when the null is true and for $\mu \neq \mu_0$ β_μ gives us the chance of rejecting H_0 when the data come from $N(\mu, \sigma^2)$ which is the correct decision.

What are desirable tests of the null hypothesis? Clearly we want α to be as small as possible and β to be as large – in other words, we would like to minimize both types of errors simultaneously. But as with most other things in life we cannot win both ways. There is a trade-off. If you want to minimize the chance of Type 1 error as much as you can, the chance of Type 2 error goes off to 1. To see this intuitively, consider once again a 0–1 valued test function for the problem. The probability of Type 1 error is then the chance that under the null ($N(\mu_0, \sigma^2)$) the observed data vector X lies in the rejection region $\mathcal{R}$. If we want to make this probability go down to 0, we will typically need to reduce the size of $\mathcal{R}$. But as we keep on reducing $\mathcal{R}$, the chance that X lives in $\mathcal{R}$ under any distribution from the alternative also dwindles; in other words, the power goes down and the probability of Type 2 error goes up. So what should we do? Instead of asking that the probability of Type 1 error be as small as possible, we pre-specify some tolerance for this error. This is usually be a small positive fraction (historically .05 or .01) and will be called the *level of significance* or simply *level*. We will admit tests for which the probability of Type 1 error is less than this pre-specified level. Within the class of such tests, we will be interested in ones that give us large values of β, the power.

Let us now construct a concrete test for our current example. While there are several paradigms for building tests systematically, like the *Neyman Pearson* lemma or the *likelihood ratio* method, we will at this stage construct some common sense test procedures and see how they perform. Some of these procedures can be justified using formal paradigms. Under the null hypothesis the X_i's are i.i.d $N(\mu_0, \sigma^2)$ and the sample mean $\overline{X}$ follows $N(\mu_0, \sigma^2/n)$. Thus, large deviations of the observed value of $\overline{X}$ from μ_0 would lead us to suspect that the null hypothesis might not be true. But how large is large? Suppose that the null hypothesis is true. If the variance of the sample mean is, say, 100, a deviation of $\overline{X}$ from μ_0 by 15 is not really unusual; on the other hand if the variance is 10, then a deviation of the sample mean from μ_0 by 15 is really sensational. Thus the quantity $\mid \overline{X} - \mu_0 \mid$ in itself is not sufficient to formulate a decision regarding rejection of the null hypothesis. We need to adjust for the underlying variance. This is done by computing

the so-called z-statistic,

$$z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \equiv \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$$

and rejecting the null hypothesis for large absolute values of this statistic. Under the null hypothesis z follows $N(0,1)$; thus an absolute z-value of 3.5 is quite unlikely. Therefore if we observe an absolute z-value of 3.5 we might rule in favor of the alternative hypothesis. You can see now that we need a threshold value, or in other words a critical point such that if the z-value exceeds that point we reject. Our test function ϕ then looks like,

$$\phi(X) = 1 \text{ if } \left| \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \right| > c_{n,\alpha_0}$$

and

$$\phi(X) = 0 \text{ otherwise}\,,$$

where c_{n,α_0} is the *critical value* and will depend on α_0 which is the tolerance for the Type 1 error i.e. the level that we set beforehand. The quantity c_{n,α_0} is determined using the relation,

$$E_{\mu_0}(\phi(X)) = \alpha_0\,,$$

which translates readily to,

$$P_{\mu_0}\left(\left| \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \right| > c_{n,\alpha_0} \right)\,.$$

Straightforward algebra then yields that

$$P_{\mu_0}\left(-c_{n,\alpha_0}\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu_0 \leq c_{n,\alpha_0}\frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha_0\,,$$

whence we can choose $c_{n,\alpha_0} = z_{\alpha_0/2}$, the $\alpha_0/2$'th quantile of the $N(0,1)$ distribution. The acceptance region $\mathcal{A}$ for the null hypothesis is therefore

$$\mathcal{A} = \{X = (X_1, X_2, \ldots, X_n) :\ \mu_0 - \frac{\sigma}{\sqrt{n}}\, z_{\alpha_0/2} \leq \overline{X} \leq \mu_0 + \frac{\sigma}{\sqrt{n}}\, z_{\alpha_0/2}\}\,.$$

So we accept whenever $\overline{X}$ lies in a certain window of μ_0, the postulated value under the null and reject otherwise which is in accordance with inutition. The length of the window is determined by the tolerance level α_0, the underlying variance σ^2 and of course the sample size n.

We will deal shortly with the power function of this test. But before doing so, we illustrate how the above testing procedure ties up naturally with the confidence interval construction problem that we dealt with in Section 2.1 (Example 1).

First note that the acceptance region of the above test ϕ can be written as:

$$\mathcal{A} = \{X = (X_1, X_2, \ldots, X_n) :\ \overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha_0/2} \leq \mu_0 \leq \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha_0/2}\}\,.$$

Now, consider a fixed data set $(X_1, X_2, \ldots, X_n)$ and based on this consider testing a family of null hypotheses: $\{H_{0,\tilde{\mu}} : \mu = \tilde{\mu} : \tilde{\mu} \in \mathbb{R}\}$. We can now ask the following question: Based on the observed data and the above testing procedure, what values of $\tilde{\mu}$ would fail to be rejected by the level α_0 test? This means that $\tilde{\mu}$ would have to fall in the acceptance region; in other words, we would require

$$\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \leq \tilde{\mu} \leq \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \, .$$

Thus, the set of $\tilde{\mu}$'s for which the null hypothesis would fail to be rejected by the level α_0 test is precisely the set:

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \ , \ \overline{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha_0/2} \right] .$$

But this is precisely the level $1 - \alpha_0$ confidence interval that we obtained in Example 1!! Thus, we obtain a level $1 - \alpha_0$ C.I. for μ, the population mean, by compiling all possible $\tilde{\mu}$'s for which the null hypothesis $\mu = \tilde{\mu}$ fails to be rejected by the level α_0 test.

We can now explain the duality between forming confidence intervals and performing tests of statistical hypothesis in a more abstract setting. We refer back to the Method of Pivots discussed in Section 2.1. In that setting, consider testing the statistical hypothesis $H_0 : h(\theta) = \eta_0$. Note that this is a composite null hypothesis; there could be different values of θ for which $h(\theta) = \eta_0$. We can now use the recipe for obtaining a level $1 - \alpha$ confidence set to construct a test of level α for the testing problem. Define a test function $\phi(Y)$ as follows:

$$\phi(Y) = 0 \text{ if } q(G; \beta_1) \leq \Psi(Y, \eta_0) \leq q(G; 1 - \beta_2) \, ,$$

and $\phi(Y) = 1$ otherwise. Then ϕ is a level α test for testing H_0. To see this, let us compute $E_\theta(\phi(Y))$ for a $\theta \in H_0$ i.e for a θ such that $h(\theta) = \eta_0$. We have,

$$\begin{aligned} E_\theta(\phi(Y)) &= 1 - \text{Prob}_\theta(q(G; \beta_1) \leq \Psi(Y, \eta_0) \leq q(G; 1 - \beta_2)) \\ &= 1 - \text{Prob}_\theta(q(G; \beta_1) \leq \Psi(Y, h(\theta)) \leq q(G; 1 - \beta_2)) \\ &= 1 - (1 - \alpha) = \alpha \, . \end{aligned}$$

To arrive at the last equality, we have used Step (3) in the Method of Pivots.

We can use this recipe for constructing tests based on confidence intervals in the context of Example (2) in the Method of Pivots. In that example, the vector Y is the vector $X = (X_1, X_2, \ldots, X_n)$, $\theta = (\mu, \sigma^2)$ and $h(\theta) = \mu$ and $H_0 : \mu = \mu_0$. By the above discussion, our level α test is given by:

$$\phi_{\mu_0}(X) = 0 \text{ if } \ -q(t_{n-1}; 1 - \alpha/2) \leq \frac{\sqrt{n}(\overline{X} - \mu_0)}{s} \leq q(t_{n-1}; 1 - \alpha/2) \, ,$$

and $\phi(X) = 0$ otherwise. Thus the acceptance region of the test ϕ is given by:

$$\mathcal{A}_{\mu_0} = \{X \ : \ \mu_0 - \frac{s}{\sqrt{n}} q(t_{n-1}; 1 - \alpha/2) \leq \overline{X} \leq \mu_0 + \frac{s}{\sqrt{n}} q(t_{n-1}; 1 - \alpha/2)\} \, .$$

This is similar to the acceptance region of the test for $\mu = \mu_0$ with σ known; the normal quantiles are now replaced by the t quantiles and σ is replaced by s.

What about the other way round? Can we go from hypothesis tests to confidence intervals? It seems this should be possible. We saw above that for the testing problem for μ with σ known, the set of $\tilde{\mu}$'s for which the null hypothesis $H_0 : \mu = \tilde{\mu}$ failed to be rejected by the level α_0 test, gave us precisely the level $1 - \alpha_0$ confidence interval in Example 1 in the Method of Pivots Section. Here is once again a general recipe.

From Hypothesis Testing to Confidence Intervals: Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. observations from some underlying distribution F_θ; here θ is a "parameter" indexing a family of distributions. For each $\tilde{\theta}$ consider testing the null hypothesis $H_{0,\tilde{\theta}} : h(\theta) = h(\tilde{\theta})$. Suppose, there exists a level α 0–1 valued test function $\phi_{\tilde{\theta}}(Y)$ for this problem. Then $\mathcal{A}_{\tilde{\theta}} = \{Y : \phi_{\tilde{\theta}}(Y) = 0\}$ is the acceptance region of $\phi_{\tilde{\theta}}$ and $P_{\tilde{\theta}}(Y \in \mathcal{A}_{\tilde{\theta}}) = 1 - \alpha$. Then a level $1 - \alpha$ confidence set for $h(\theta)$ is:

$$\mathcal{S}(Y) = \{h(\tilde{\theta}) : Y \in \mathcal{A}_{\tilde{\theta}}\}\,.$$

We need to verify that for any θ, $P_\theta(h(\theta) \in \mathcal{S}(Y)) = 1 - \alpha$. But

$$P_\theta(h(\theta) \in \mathcal{S}(Y)) = P_\theta\left(Y \in \mathcal{A}_\theta\right) = 1 - \alpha\,.$$

What happens if we apply this to the example we started out with in Section 3? Thus $Y = X = (X_1, X_2, \ldots, X_n)$ where X_i's are i.i.d. $N(\mu, \sigma^2)$ with σ known. We have $\theta = \mu$ and $h(\theta) = h(\mu) = \mu$. We denote $\tilde{\theta}$ for this example by $\tilde{\mu}$. Consider testing $\mu = \tilde{\mu}$. The acceptance region of the level α test $\phi_{\tilde{\mu}}$ can be written as:

$$\mathcal{A}_{\tilde{\mu}} = \{X = (X_1, X_2, \ldots, X_n) :\ \overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2} \le \tilde{\mu} \le \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\}\,.$$

To obtain a level $1 - \alpha$ C.I. we need to compile all $\tilde{\mu}$ such that $X \in \mathcal{A}_{\tilde{\mu}}$; in other words, all $\tilde{\mu}$ such that

$$\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2} \le \tilde{\mu} \le \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\,.$$

This is precisely the interval

$$\left[\overline{X} - \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\ ,\ \overline{X} + \frac{\sigma}{\sqrt{n}}\, z_{\alpha/2}\right]\,.$$

Exercise: How can you use the confidence intervals from Example 3 in the Method of Pivots section to devise a level α test for the null hypothesis $H_0 : \sigma = \sigma_0$?

Power Behavior: We have arrived at two different level α tests for testing $H_0 : \mu = \mu_0$ against its complement depending on whether σ is known or unknown. If σ is known, the test function ϕ say ϕ_{known} is given by:

$$\phi_{known}(X) = 1 \text{ if } \left|\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}\right| > z_{\alpha/2}$$

and

$$\phi_{known}(X) = 0 \text{ otherwise}.$$

Denote the power function of this test by $\beta(\mu)$. On the other hand, we have $\phi_{unknown}$, the level α test for σ unknown and this is given by:

$$\phi_{unknown}(X) = 1 \text{ if } \left| \frac{\sqrt{n}(\overline{X} - \mu_0)}{s} \right| > q(t_{n-1}; 1 - \alpha/2)$$

and

$$\phi_{unknown}(X) = 0 \text{ otherwise}.$$

Note that $\phi_{unknown}$ is a natural modification of ϕ_{known}. Denote this power function by $\tilde{\beta}(\mu)$. Both the power functions behave qualitatively the same (similar graphs). It is easier to study the power of ϕ_{known} and this is what we do next.

We have,

$$\beta(\mu) = P_\mu \left(\left| \frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \right| > z_{\alpha/2} \right),$$

which is just,

$$P_\mu \left(\left| \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right| > z_{\alpha/2} \right).$$

But when μ is the population mean, $\sqrt{n}(\overline{X} - \mu)/\sigma$ is $N(0,1)$. If Z denotes a $N(0,1)$ variable then,

$$\begin{aligned}
\beta(\mu) &= P\left(\left| Z + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right| > z_{1-\alpha/2} \right) \\
&= P\left(Z + \frac{\sqrt{n}(\overline{\mu} - \mu_0)}{\sigma} > z_{\alpha/2} \right) + P\left(Z + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} < -z_{\alpha/2} \right) \\
&= 1 - \Phi\left(z_{\alpha/2} - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right) + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right) \\
&= \Phi\left(-z_{\alpha/2} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right) + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \right).
\end{aligned}$$

Check from the above calculations that $\beta(\mu_0) = \alpha_0$, the level of the test ϕ. We say that ϕ rejects the null hypothesis at level α_0. Notice that the test function ϕ depends on the value μ_0 under the null but it does not depend on any value in the alternative.

Let's now study the power function $\beta(\mu)$ as a function of μ. For sensible test procedures we expect the power to increase as the true value μ deviates further and further from μ_0 since a sensible test should be able to pick greater departures from the null hypothesis more easily than smaller ones. For the current test ϕ it is easy to check that $\beta(\mu)$ diverges to 1 as μ diverges to ∞

or $-\infty$ and $\beta(\mu_0) = \alpha_0$. Moreover the power function is symmetric around μ_0. In other words, $\beta(\mu_0 + \Delta) = \beta(\mu_0 - \Delta)$ where $\Delta > 0$. To see this, note that

$$\beta(\mu_0 + \Delta) = \Phi\left(-z_{\alpha/2} + \frac{\sqrt{n}\,\Delta}{\sigma}\right) + \Phi\left(-z_{\alpha/2} - \frac{\sqrt{n}\,\Delta}{\sigma}\right).$$

Check that you get the same expression for $\beta(\mu_0 - \Delta)$. To study how $\beta(\mu)$ behaves as a function of μ as we move away from $\beta(\mu_0)$ let's calculate the derivative of $\beta(\mu)$. We have,

$$\beta'(\mu) = \frac{\sqrt{n}}{\sigma}\left(\phi\left(-z_{\alpha/2} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}\right) - \phi\left(-z_{\alpha/2} - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}\right)\right),$$

where $\phi(\cdot)$ denotes the standard normal density. For $\Delta > 0$,

$$\beta'(\mu_0 + \Delta) = \frac{\sqrt{n}}{\sigma}\left(\phi\left(-z_{\alpha/2} + \frac{\sqrt{n}\,\Delta}{\sigma}\right) - \phi\left(-z_{1-\alpha_0/2} - \frac{\sqrt{n}\,\Delta}{\sigma}\right)\right)$$

and it is easily checked that

$$\beta'(\mu_0 - \Delta) = -\beta'(\mu_0 + \Delta).$$

Thus the slope of the curve at $\mu_0 - \Delta$ is the negative of the slope of the curve at $\mu_0 + \Delta$ and the values of the curve at these two points are the same. Check that $\beta'(\mu_0) = 0$, so that μ_0 is a local maximum or a local minimum. Also, for $\Delta > 0$,

$$\phi\left(-z_{\alpha/2} + \frac{\sqrt{n}\,\Delta}{\sigma}\right) > \phi\left(-z_{1-\alpha_0/2} - \frac{\sqrt{n}\,\Delta}{\sigma}\right),$$

since

$$\left|-z_{\alpha/2} + \frac{\sqrt{n}\,\Delta}{\sigma}\right| < \left|-z_{1-\alpha_0/2} - \frac{\sqrt{n}\,\Delta}{\sigma}\right|.$$

This shows that $\beta'(\mu_0 + \Delta) > 0$. It follows that $\beta'(\mu) > 0$ for $\mu > \mu_0$ and $\beta'(\mu) < 0$ for $\mu < \mu_0$, showing that μ_0 is a global minimum and that the power increases symetrically in either direction (to 1) of μ_0 as you move out toward ∞ or $-\infty$. Also the slope of the curve converges to 0 as μ diverges to ∞ or $-\infty$ showing that the line $y = 1$ is an asymptote to the curve. Figure 1 shows a plot of the power function $\beta(\mu)$ for $n = 25$, $\mu_0 = 0$ and $\sigma = 1$ and $\alpha_0 = 0.05$. The curve attains its minimum at $\mu = \mu_0$ under the null hypothesis where it is equal to the level 0.05. The power at every alternative is larger than the level – this property is referred to as *unbiasedness* of the test ϕ. What we have discussed above is an example of a "two-sided" test, since the alternative hypothesis $\mu \neq \mu_0$ occurs on either side of μ_0 and a sensible test procedure, which the one we have used indeed is, needs to have power on either side of the null hypothesis.

3.1 Testing a simple null hypothesis against a simple alternative – The Neyman Pearson Lemma.

Consider the following testing problem. Let X be an observation from some underlying density f. Based on X we want to test the null hypothesis $H_0 : f = f_0$ versus the alternative hypothesis

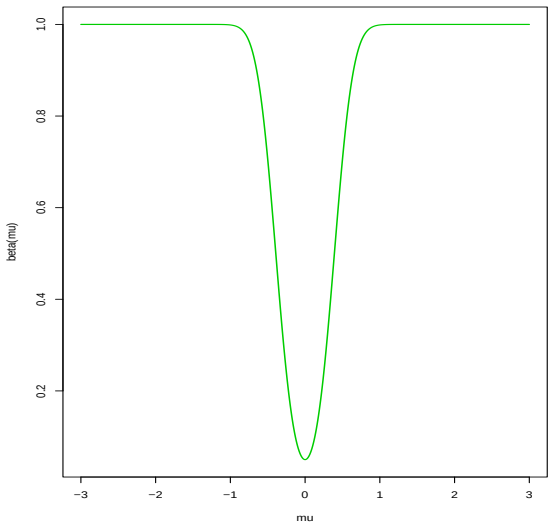

Figure 1: The power function $\beta(\mu)$

$H_1 : f = f_1$.

Let ϕ denote a generic test function for this problem. As before $\phi(X)$ is the conditional probability with which we reject H_0 after having observed the data–point X. We have

$$\text{Prob}(Type\ 1\ error) = \int \phi(x)\, f_0(x)\, dx \equiv E_{f_0}(\phi(X))$$

and

$$\text{Prob}(Type\ 2\ error) = \int (1 - \phi(x))\, f_1(x)\, dx \equiv E_{f_1}(1 - \phi(X))\,,$$

whence the power of the test is

$$\beta = \int \phi(x)\, f_1(x)\, dx \equiv E_{f_1}(\phi(X))\,.$$

To construct meaningful tests, we need to allow some tolerance for Type 1 error as argued before. Suppose we fix our tolerance level at α. We then ask the following question:

Among all tests that have level α, can we find one that has maximum power (an MP level α test) ? In other words, can we find a test ϕ_{MP} such that

$$E_{f_0}(\phi_{MP}(X)) \leq \alpha$$

and satisfying,

$$E_{f_1}(\phi(X)) \leq E_{f_1}(\phi_{MP}(X))$$

for any test ϕ for which

$$E_{f_0}(\phi(X)) \leq \alpha\,?$$

The answer to this problem is an emphatic YES. The above problem and it solution are known as the *Neyman–Pearson* lemma and is one of the key results in classical statistics. We will present the solution but not the proof, since it is not necessary at this stage. There does indeed exist a *Most Powerful* test of level α of the form,

$$\phi_{MP}(X) = 1 \text{ if } \frac{f_1(X)}{f_0(X)} > K_\alpha\,,$$

$$\phi_{MP}(X) = \gamma_\alpha \text{ if } \frac{f_1(X)}{f_0(X)} = K_\alpha$$

and

$$\phi_{MP}(X) = 0 \text{ otherwise}\,,$$

where K_α and γ_α are determined by the equation

$$E_{f_0}\,\phi_{MP}(X) \equiv P_{f_0}\left(\frac{f_1(X)}{f_0(X)} > K_\alpha\right) + K_\alpha\, P_{f_0}\left(\frac{f_1(X)}{f_0(X)} = \gamma_\alpha\right) = \alpha\,.$$

We will discuss an application of the Neyman-Pearson lemma in a moment but before we do that, let's point out an easy extension of the lemma. Typically, in many applications, it is the case that the test statistic ϕ_{MP} *does not depend on the alternative* f_1. In such a case, it is possible to construct an MP test for testing $H_0 : f = f_0$ against a composite alternative hypothesis of the form $H_1 : f \in \mathcal{F}_{alt}$, $\mathcal{F}_{alt}$ being a family of densities. Let us state this more formally.

Proposition: Let X be an observation from an underlying distribution f and based on X we desire to test $H_0 : f = f_0$ against $H_1 : f \in \mathcal{F}_{alt}$ where $\mathcal{F}_{alt} = \{f_\theta : \theta \in \Theta\}$. (Here Θ is an index set and f_θ denotes the density in the alternative hypothesis labeled by θ. None of the f_θ's is equal to f_0.) Suppose that $\phi_{MP,\theta}$, the MP test for testing $H_0 : f = f_0$ versus $H_{1,\theta} : f = f_\theta$ at level α *does not depend* on θ. Thus there is a single test function ϕ_{MP} of level α that is most powerful for testing H_0 against $H_{1,\theta}$ for each θ. Then ϕ_{MP} is the *uniformly most powerful (UMP)* test of level α for testing H_0 versus H_1.

A uniformly most powerful level α test for the above testing problem is a $\tilde{\phi}$ such that,

$$E_{f_0}\,(\tilde{\phi}(X)) \leq \alpha$$

and if ϕ is any level α test i.e.

$$E_{f_0}\,(\phi(X)) \leq \alpha$$

then, for each $\theta \in \Theta$,

$$E_{f_\theta}\,(\phi(X)) \leq E_{f_\theta}\,(\tilde{\phi}(X)).$$

Thus the power function of $\tilde{\phi}$ (which is of level α) dominates the power function of any other level α test over the entire domain of the alternative hypothesis.

The fact that ϕ_{MP} is indeed the uniformly most powerful test is almost immediate from assumptions. For any test ϕ of level α we know that

$$E_{f_\theta}(\phi(X)) \leq E_{f_\theta}(\phi_{MP,\theta}(X)) \equiv E_{f_\theta}(\phi_{MP}(X))\,.$$

Since this happens for every $\theta \in \Theta$ it shows that ϕ_{MP} is indeed uniformly most powerful.

We now illustrate the Neyman–Pearson lemma through a simple example. Consider a sample $X_1, X_2, \ldots, X_n$ from an underlying $N(\mu, \sigma^2)$ distribution. We want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu = \mu_1$ with $\mu_1 > \mu_0$. Assume at this stage that σ^2 is known. If we formulate this problem in the Neyman–Pearson set–up, we have

$$f_1(X) = f_{\mu_1}(X_1, X_2, \ldots, X_n) \;\text{ and }\; f_0(X) = f_{\mu_0}(X_1, X_2, \ldots, X_n)\,,$$

where

$$f_\mu(X_1, X_2, \ldots, X_n) = \left(\frac{1}{\sqrt{2\,\pi}\,\sigma}\right)^n \exp\left(-\frac{1}{2\,\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\right).$$

The Neyman-Pearson test is based on the ratio $f_1(X)/f_0(X)$ or equivalently on $\log\, f_1(X)/f_0(X)$. Thus, the MP level α test is of the form:

$$\phi_{MP}(X) = 1 \;\text{ if }\; \log\,\frac{f_1(X)}{f_0(X)} > \log\, K_\alpha \equiv k_\alpha\,,$$

$$\phi_{MP}(X) = \gamma_\alpha \;\text{ if }\; \log\,\frac{f_1(X)}{f_0(X)} = k_\alpha$$

and

$$\phi_{MP}(X) = 0 \;\text{ otherwise}\,.$$

We have,

$$\begin{aligned}\log\,\frac{f_1(X)}{f_0(X)} &= -\frac{1}{2\,\sigma^2}\sum_{i=1}^n (X_i - \mu_1)^2 + \frac{1}{2\,\sigma^2}\sum_{i=1}^n (X_i - \mu_0)^2 \\ &= \frac{n\,\overline{X}}{\sigma^2}\,(\mu_1 - \mu_0)\,.\end{aligned}$$

Since $\mu_1 - \mu_0 > 0$, $\log\, f_1(X)/f_0(X)$ is a strictly increasing function of $\overline{X}$ and we can base our MP test ϕ_{MP} simply on $\overline{X}$. This will assume the value 1 when $\overline{X} > c_{\alpha,n}$, will be equal to γ_α when $\overline{X} = c_\alpha$ and equal to 0 otherwise. Also, since ϕ_{MP} is of level α, it needs to satisfy $E_0(\phi_{MP}(X)) = \alpha$ or equivalently,

$$P_0\,(\overline{X} > c_{n,\alpha}) + \gamma_\alpha\, P_0\,(\overline{X} = c_{n,\alpha}) = \alpha\,.$$

Since $\overline{X}$ follows $N(\mu_0, \sigma^2/n)$ under H_0, the chance that $\overline{X} = c_{n,\alpha}$ is actually 0 and we can choose our test function to be a 0–1 valued test function, with $\gamma_\alpha = 1$, so that $\phi(X) = 1$ if $\overline{X} \geq c_{n,\alpha}$ (the

rejection region of the test) and $\phi(X) = 0$ otherwise (the acceptance region of the test). We then have,

$$P_0\left(\overline{X} \geq c_{n,\alpha}\right) = \alpha\,,$$

or equivalently,

$$P_0\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \geq \frac{\sqrt{n}(c_{n\,\alpha} - \mu_0)}{\sigma}\right) = \alpha\,.$$

Under H_0,

$$\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \sim N(0,1)$$

showing that

$$\frac{\sqrt{n}(c_{n\,\alpha} - \mu_0)}{\sigma} = z_{1-\alpha}\,.$$

Thus, our MP test can be formulated as: Reject H_0 in favor of H_1 if

$$\overline{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha}$$

and do not reject otherwise. Notice that the above test does not depend on the actual value of the alternative μ_1 and hence by the preceding proposition is the *uniformly most powerful* level α test for testing $H_0 : \mu = \mu_0$ versus the *composite* hypothesis $H_1 : \mu > \mu_0$. Notice that we cannot use the same test for alternatives to the left of μ_0 since the form of the test (reject for $\overline{X}$ large) is a direct consequence of the fact that $\mu_1 > \mu_0$.

What can we say about the power function $\beta(\mu)$ of the test? Once again, note that we can define $\beta(\mu)$ for any real number μ. We have,

$$\begin{aligned}
\beta(\mu) &= E_\mu(\phi(X)) \\
&= P_\mu\left(\overline{X} \geq \mu_0 + \frac{\sigma}{\sqrt{n}}\, z_{1-\alpha}\right) \\
&= P_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma} \geq z_{1-\alpha}\right) \\
&= P_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \geq z_{1-\alpha}\right).
\end{aligned}$$

Under μ i.e. when the underlying distribution is $N(\mu, \sigma^2)$, $\sqrt{n}(\overline{X} - \mu)/\sigma \sim N(0,1)$; if Z denotes a standard normal random variable we get,

$$\begin{aligned}
\beta(\mu) &= P\left(Z + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma} \geq z_{1-\alpha}\right) \\
&= 1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}\right) \\
&= \Phi\left(-z_{1-\alpha} + \frac{\sqrt{n}(\mu - \mu_0)}{\sigma}\right).
\end{aligned}$$

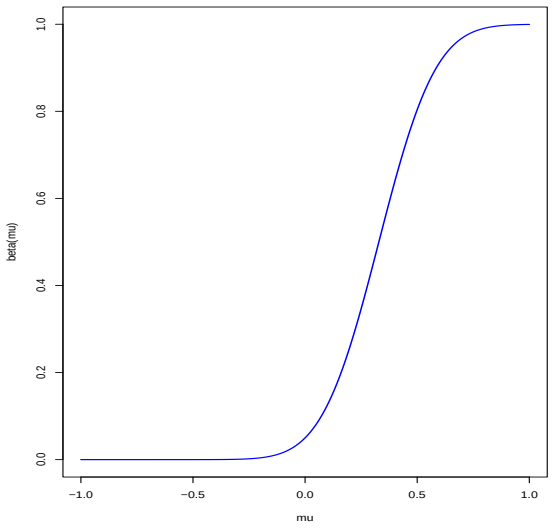

Figure 2: The power function of the MP test $\beta(\mu)$

As μ increases from $-\infty$ to ∞ $\beta(\mu)$ increases from 0 to 1; at $\mu = \mu_0$, $\beta(\mu_0) = \alpha$ and gives the probability of Type 1 error of the UMP test ϕ_{MP} for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$. For $\mu < \mu_0$, $\beta(\mu) < \alpha$. The above observation is crucial: it shows that ϕ_{MP} is also the uniformly most powerful test of level α for testing the null hypothesis $\mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ based on i.i.d. observations $X_1, X_2, \ldots, X_n$. Let's explain why this is the case. An hypothesis of the form $\mu \leq \mu_0$ is a composite null hypothesis – more than one distribution is possible under the null. A level α test, ϕ, for testing H_0 against H_1 needs to satisfy $E_\mu(\phi(X)) \leq \alpha$ for all $\mu \leq \mu_0$. Consider any such test ϕ. Then $E_{\mu_0}(\phi(X)) \leq \alpha$; ϕ therefore qualifies as a level α test for testing the null hypothesis $\tilde{H}_0 : \mu = \mu_0$ against H_1. But we know that ϕ_{MP} is the MP level α test for this testing problem; in other words, $E_\mu(\phi_{MP}(X)) \geq E_\mu(\phi(X))$ for all $\mu > \mu_0$. To show that ϕ_{MP} is indeed UMP level α for testing H_0 against H_1, we just need to check that it indeed continues to remain of level α under the null hypothesis H_0, i.e. $\beta_{\phi_{MP}}(\mu) \leq \alpha$ for $\mu \leq \mu_0$. But this is indeed the case (as we have seen above). Thus ϕ_{MP} is the *best possible test* we can construct for testing H_0 against H_1 with a tolerance level α for Type 1 error. In this case, the largest possible value of the Type 1 error is called the *size* of the test. From the monotonicity of $\beta(\mu)$ it is immediate that the size of ϕ_{MP} for testing H_0 versus H_1 is $\beta(\mu_0) \equiv \alpha$.

In Figure 2 we present a graph of the power function $\beta(\mu)$; for this example $n = 25, \mu_0 = 0, \sigma = 1$ and $\alpha = .05$. It is clear from the various examples that we have discussed above that the rejection region of a test function ϕ will depend heavily on the alternative hypothesis that we choose to postulate. For the testing problem $\mu = \mu_0$ versus $\mu \neq \mu_0$ our rejection region was on either side of $\overline{X}$. This was a sensible thing to do because we needed power on either side of μ_0. The power

of the test ϕ that we used in that situation came from the part of the rejection region to the left of $\overline{X}$ when the true alternative was to the left of μ_0. When the true alternative was to the right, the part of the rejection region to the right of $\overline{X}$ guaranteed us power. On the other hand the test ϕ_{MP} that we used to test $\mu \leq \mu_0$ against $\mu > \mu_0$ would be a *horrible test* for the testing problem $\mu = \mu_0$ versus $\mu \neq \mu_0$, since it would have very little power to the left of μ_0 – indeed, we constructed the test that way, so that the probability of Type 1 error would be small for every μ in $H_0 : \mu \leq \mu_0$. On the other hand, the test ϕ would also be a horrible test for testing $\mu \leq \mu_0$ versus $\mu > \mu_0$ since it would have size (this is the maximum probability of Type 1 error) 1. But how would the test ϕ compare to the test ϕ_{MP} for testing $\mu = \mu_0$ versus $\mu > \mu_0$? In this case we know that the power of ϕ would be no greater than the power of ϕ_{MP} at each $\mu > \mu_0$ since ϕ_{MP} is most powerful. Figure 3 illustrates this point.

Here is a brief explanation of Figure 3. Shown in black in Figure 3 is the power-function of the two-sided test ϕ that was used to test $\mu = \mu_0$ versus $\mu \neq \mu_0$. In red is shown the power function of the one-sided test $\phi_{MP,1}$ which is most powerful for testing $\mu \leq \mu_0$ versus $\mu > \mu_0$. In green we show the power function of $\phi_{MP,2}$, the most powerful test for testing $H_0 : \mu \geq \mu_0$ versus $H_1 : \mu < \mu_0$, which you are going to derive as part of your homework. (The underlying parameters are $\mu_0 = 0, \sigma = 1, n = 25, \alpha = 0.05$.) All the three tests have level α. The test ϕ can be used as a level α test for testing $\mu = \mu_0$ versus $\mu > \mu_0$ and also for testing $\mu = \mu_0$ versus $\mu < \mu_0$. In the first case, i.e. when used as a test for $\mu = \mu_0$ versus $\mu > \mu_0$, it competes with $\phi_{MP,1}$ and as expected, for $\mu > \mu_0, \phi$ has less power than $\phi_{MP,1}$. In the second case, i.e. when used as a test for $\mu = \mu_0$ versus $\mu < \mu_0$, it competes with $\phi_{MP,2}$ and as expected, for $\mu < \mu_0$, ϕ has less power than $\phi_{MP,2}$ as shown by the fact that the green curve lies above the black curve. All three curves intersect at $\mu_0 \equiv 0$. You can think of ϕ as a hybrid test obtained by combining the features of $\phi_{MP,1}$ and $\phi_{MP,2}$; by sacrificing a little bit of power on either side of μ_0 (as compared to the MP tests), ϕ manages to have power on either side of μ_0 for detecting departures from the null. The test ϕ can be justified using the *likelihood ratio statistic* paradigm which we probably will not have time to cover in this course, but is not an MP test for testing $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$. In fact, an MP test for this problem does not exist.

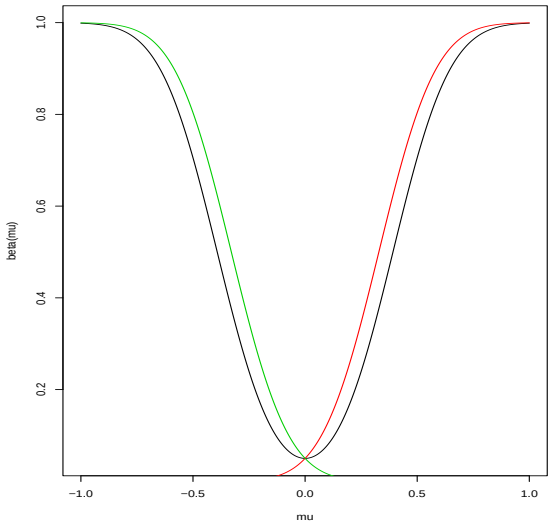

Figure 3: Comparing the power functions of 3 different test statistics