t-Statistics for Weighted Means with Application to Risk Factor Models

Lisa R. Goldberg
Barra, Inc.
2100 Milvia St.
Berkeley, CA 94704

Alec N. Kercheval
Dept. of Mathematics
Florida State University
Tallahassee, FL 32306-4510

June 24, 2002

1 Introduction

In this note we describe how to generalize the standard t-statistic test for for equality of the means when the assumption of a common variance no longer holds. We then discuss an application to financial risk factor modelling.

First we describe the standard t-statistic. Suppose we have a sequence of independent samples from a normal distribution with mean μ_X and variance σ^2. Denote the sample values by $X_1, X_2, \ldots, X_n$. We use the notation $X_i \sim N(\mu_X, \sigma^2)$, where $N(a, b)$ denotes the probability density function of a normal distribution with mean a and variance b.

The best (minimum variance) linear unbiased estimator of the mean μ is the sample mean

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

If $Y_1, Y_2, \ldots, Y_m$ is another group of independent samples with $Y_i \sim N(\mu_Y, \sigma^2)$, we could ask whether or not $\mu_X = \mu_Y$. We take the *null hypothesis* to be the statement that this equality is true.

Given our sample data, we cannot determine the truth or falsity of the null hypothesis, but we can determine the likelihood of the realized sample values assuming the null hypothesis. If this likelihood is small, we are justified in rejecting the null hypothesis.

To accomplish this, we may use the standard (Student's) *t-statistic* for equality of the mean:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{nS_X^2 + mS_Y^2}{n+m-2}(\frac{1}{n} + \frac{1}{m})}} \tag{1}$$

where

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

is the sample variance of X, and similarly for Y.

The random variable T has a t-distribution with $n + m - 2$ degrees of freedom. Therefore we can determine the probability that T is equal to or greater than the realized value given $\mu_X = \mu_Y$. Typically if this probability is below 5% or 1%, the null hypothesis is rejected.

In this paper we generalize the discussion to the case where the samples are drawn from distributions with a common mean but variances allowed to change from sample to sample:

$$X_i \sim N(\mu_X, \sigma_i^2).$$

In this case, the best linear unbiased estimate of the mean μ_X is the *weighted average*

$$\bar{X} = \sum_{i=1}^{n} w_i X_i, \tag{2}$$

where

$$w_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{n}(1/\sigma_j^2)}.$$

Conversely, given positive weights w_i, $i = 1, \ldots, n$ so that $\sum_{i=1}^{n} w_i = 1$, then the quantity in equation 2 is the best linear unbiased estimate of the mean provided that the samples are distributed as

$$X_i \sim N(\mu_X, \alpha_X / w_i)$$

for some constant $\alpha_X > 0$.

In either case, if

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

is the weighted sample variance, and if we use similar notation for Y_i (with different weights w_i' allowed), then the corresponding formula for the t-statistic for equality of the weighted mean is

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X/\alpha_X + S_Y/\alpha_Y}{n+m-2}}\sqrt{\alpha_X + \alpha_Y}}. \tag{3}$$

Setting $w_i = 1/n$, $w_i' = 1/m$, and $\alpha_Y = (n/m)\alpha_X$ reduces this expression to equation 1.

Note: T is independent of the scale of the pair (α_X, α_Y): if (α_X, α_Y) is replaced by $(k\alpha_X, k\alpha_Y)$ for some $k > 0$, the value of T is unchanged.

2 The Weighted Mean as a Minimum Variance Estimator

If $X_1, X_2, \ldots, X_n$ is a random sample such that $X_i \sim N(\mu, \sigma_i^2)$, what is the minimum variance unbiased estimator of the mean? It is a weighted sum where greater weight is given to values coming from narrower distributions.

Let $X_i = \mu + e_i$ where e_i has mean μ and variance σ_i^2. If

$$\bar{X} = \sum_{i=1}^{n} w_i X_i$$

is to be the minimum variance unbiased estimator of the mean μ, then we must solve for the weights w_i minimizing the variance of $\bar{X}$, subject to the constraint

$$\sum w_i = 1 \tag{4}$$

Since we are assuming the variables e_i are independent, we have

$$\begin{aligned} E[(\bar{X} - \mu)^2] &= E[(\textstyle\sum w_i e_i)^2] \\ &= \textstyle\sum E[w_i^2 e_i^2] \\ &= \textstyle\sum w_i^2 \sigma_i^2 \end{aligned}$$

The method of Lagrange multipliers to minimize this function subject to the constraint in equation 4 yields

$$w_i = \frac{1/\sigma_i^2}{\sum_{j=1}^{n}(1/\sigma_j^2)}.$$

We obtain this weight if we set

$$\sigma_i^2 = \alpha/w_i,$$

where α is any positive constant. This proves

Proposition 1 *Let α be a positive constant. Suppose $w_1, \ldots, w_n$ are positive numbers satisfying $\sum w_i = 1$, and, for each i, X_i is a random variable with mean μ and variance α/w_i.*

Then the minimum variance unbiased estimator of the mean μ is

$$\bar{X} = \sum_{i=1}^{n} w_i X_i.$$

3 Establishing the Weighted t-Statistic

Recall that if a random variable V is the sum of the squares of $r > 0$ independent standard normal variables, then V is said to have a chi-squared distribution with r degrees of freedom.

The *t-distribution with r degrees of freedom* may be defined as the distribution of the random variable

$$T = \frac{W}{\sqrt{V/r}},$$

where W is a standard normal random variable, V has a chi-squared distribution with r degrees of freedom, and W and V are independent.

We need to show that the statistic defined in equation 3 has a t-distribution with $n + m - 2$ degrees of freedom. We accomplish this with a sequence of lemmas in this section.

Standing assumptions: Let α_X and α_Y be fixed positive numbers. For $i = 1, \ldots, n$, and $j = 1, \ldots, m$, let w_i and w'_j be positive numbers and X_i, Y_j independent random variables such that

- $\sum_{i=1}^{n} w_i = 1$ and $\sum_{j=1}^{m} w'_j = 1$, and

- for each i, j, $X_i \sim N(\mu, \alpha_X/w_i)$ and $Y_j \sim N(\mu, \alpha_Y/w'_j)$.

Notation:

- $\bar{X} = \sum w_i X_i$ and $\bar{Y} = \sum w'_j Y_j$
- $S_X = \sum w_i (X_i - \bar{X})^2$ and $S_Y = \sum w'_j (Y_j - \bar{Y})^2$

Lemma 1 *$\bar{X} \sim N(\mu, \alpha_X)$ and $\bar{Y} \sim N(\mu, \alpha_Y)$.*

Proof. A straightforward computation using the fact that a sum of independent normals is normal and variances add.

Lemma 2 *$\bar{X}, \bar{Y}, S_X$, and S_Y are mutually independent.*

Proof. Clearly $\bar{X}$ and $\bar{Y}$ are independent, and similarly for S_X and S_Y. We show that $\bar{X}$ is independent of S_X, and the same argument works for Y. The argument is a direct generalization of the proof for the equal weighted case found, e.g., in Hogg and Craig [1, ch. 4], which we include here for the reader's convenience.

Write $\alpha = \alpha_X$ and denote the variance of X_i by σ_i^2 ($= \alpha/w_i$). The joint pdf of $X_1, X_2, \ldots, X_n$ is

$$f(x_1, \ldots, x_n) = \frac{1}{(\prod_{i=1}^n \sqrt{2\pi}\sigma_i)} \exp[-\sum_{i=1}^n \frac{(x_i - \mu)^2}{2\sigma_i^2}]$$

Our strategy is to change variables in such a way that the independence of $\bar{X}$ and S_X will be evident. Letting $\bar{x} = \sum w_i x_i$, straightforward computation verifies that

$$\alpha = \frac{1}{\sum_{i=1}^n 1/\sigma_i^2}$$

and

$$\sum_{i=1}^n \frac{(x_i - \mu)^2}{\sigma_i^2} = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sigma_i^2} + (\bar{x} - \mu)/\alpha \tag{5}$$

Hence

$$f(x_1, \ldots, x_n) = \frac{1}{(\prod_{i=1}^n \sqrt{2\pi}\sigma_i)} \exp[-\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{2\sigma_i^2} - \frac{(\bar{x} - \mu)^2}{2\alpha}] \tag{6}$$

Consider the linear transformation $(u_1, \ldots, u_n) = L(x_1, \ldots, x_n)$ defined by $u_1 = \bar{x}, u_2 = x_2 - \bar{x}, \ldots, u_n = x_n - \bar{x}$, with inverse transformation

$$\begin{aligned} x_1 &= u_1 - (\frac{\sigma_1^2}{\sigma_2^2})u_2 - (\frac{\sigma_1^2}{\sigma_3^2})u_3 - \ldots - (\frac{\sigma_1^2}{\sigma_n^2})u_n, \\ x_2 &= u_1 + u_2, \\ &\ldots \\ x_n &= u_1 + u_n \end{aligned}$$

Likewise define new random variables $U_1 = \bar{X}, U_2 = X_2 - \bar{X}, \ldots, U_n = X_n - \bar{X}$.

If J denotes the Jacobian of L, then the joint pdf of $U_1, \ldots, U_n$ is

$$\frac{J}{(\prod_{i=1}^n \sqrt{2\pi}\sigma_i)} \exp[-\frac{(-(\frac{\sigma_1^2}{\sigma_2^2})u_2 - (\frac{\sigma_1^2}{\sigma_3^2})u_3 - \ldots - (\frac{\sigma_1^2}{\sigma_n^2})u_n)^2}{2\sigma_1^2} - \sum_{i=2}^n \frac{u_i^2}{2\sigma_i^2} - \frac{(u_1-\mu)^2}{2\alpha}]$$

This now factors as a product of the pdf of U_1 and the joint pdf of $U_2, \ldots, U_n$. Hence $U_1 = \bar{X}$ is independent of $U_2, \ldots, U_n$, and hence also independent of

$$\alpha[(-(\frac{\sigma_1^2}{\sigma_2^2})U_2 - (\frac{\sigma_1^2}{\sigma_3^2})U_3 - \ldots - (\frac{\sigma_1^2}{\sigma_n^2})U_n)^2 + \sum_{i=2}^n \frac{U_i^2}{\sigma_i^2}]$$

$$= \alpha \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma_i^2} = S_X$$

Lemma 3 *$S_X/\alpha_X \sim \chi^2(n-1)$ and $S_Y/\alpha_Y \sim \chi^2(m-1)$, where $\chi^2(k)$ denotes the chi-squared distribution with k degrees of freedom.*

Proof. The proofs for X and Y are similar. Let

$$A = \sum_1^n \frac{(X_i - \mu_X)^2}{\sigma_i^2},$$

$$B = \sum_1^n \frac{(X_i - \bar{X})^2}{\sigma_i^2},$$

and

$$C = \frac{(\bar{X} - \mu_X)^2}{\alpha_X}.$$

Then by equation 5, $A = B + C$. Since $X_i \sim N(\mu_X, \sigma_i^2)$, $A \sim \chi^2(n)$. Similarly $C \sim \chi^2(1)$. This implies that $B = S_X/\alpha_X \sim \chi^2(n-1)$ provided that B and C are independent, which follows from the proof of lemma 2.

Proposition 2

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X/\alpha_X + S_Y/\alpha_Y}{n+m-2}}\sqrt{\alpha_X + \alpha_Y}}$$

is a t-statistic with $n + m - 2$ *degrees of freedom.*

Proof. Let

$$W = \frac{\bar{X} - \bar{Y}}{\sqrt{\alpha_X + \alpha_Y}}$$

and

$$V = S_X/\alpha_X + S_Y/\alpha_Y.$$

By Lemma 2, W and V are independent. From lemma 1, W is a standard normal random variable. From lemma 3, $V \sim \chi^2(n + m - 2)$. Hence

$$T = \frac{W}{\sqrt{V/(n + m - 2)}}$$

has the required property.

4 Application to Risk Modelling

For certain financial risk factor models, the return to a given factor is computed as the weighted average of returns to the individual securities exposed to that factor. For example, a model for bond credit risk may have a *Financial AA* factor to which all financial bonds rated AA are exposed. If the return to this factor is defined to be the duration-weighted average of the option adjusted spread (OAS) returns X_i, we would take weights

$$w_i = \frac{D_i}{\sum_{i=1}^{n} D_i}$$

where D_i is the duration of the ith bond. The factor return is then the weighted average

$$\bar{X} = \sum_{i=1}^{n} w_i X_i.$$

We may interpret this factor return as the best linear unbiased estimator of the common mean of a set of independent normal distributions from which

the individual bond OAS returns are sampled; the distributions are those of Proposition 1.

If, in the course of building the model, the question arises whether two groups of bond OAS returns $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ share the same mean and therefore should be exposed to the same risk factor, we may use the t-statistic of equation 3 to examine the question. A large value of this statistic is evidence that the two groups of bonds have different means and therefore should be exposed to separate risk factors.

References

[1] R.V. Hogg and A.T. Craig, *Introduction to Mathematical Statistics, 5th Edition*, Prentice-Hall, 1995.