Lot Acceptance Testing Using Sample Mean and Extremum With Finite Qualification Samples*

Stefan Kloppenborg
De Havilland Aircraft of Canada Ltd.

20 Jan 2023

Abstract

In the aerospace composites industry, new material lots are tested to determine if they are suitable for use. It is common to accept or reject the material lot by comparing the sample mean and lower extremum to reference values that are established based on an initial (qualification) sample of material property measurements. Current industry practices assume that the samples are drawn from a normal distribution with known parameters equal to the mean and standard deviation of the qualification sample: this assumption yields a producer's risk that is too high. This paper presents a two-sample method of setting these reference values, considering the sampling distribution of the qualification sample. This new method is validated through simulation which shows that it produces the correct probability of Type I error. Simulation is also used to investigate the statistical power of the new method and it is compared to others commonly used. A case study is presented to demonstrate the use of the new method using composite material data from industry.

Keywords: Compliance Sampling, Composite Material, Dual-Acceptance Criteria, Saddlepoint Approximation

*This is an Accepted Manuscript of an article published by Taylor & Francis in the Journal of Quality Technology on 20 Jan 2023, available at: `https://doi.org/10.1080/00224065.2022.2147884`

1 Introduction

When an aircraft built from composite materials is developed, the materials are tested to determine their strengths. In industry, these data are called the qualification sample. Based on the qualification sample, lower tolerance limits are computed (see for example Krishnamoorthy and Mathew (2008), or Meeker et al. (2017)). These lower tolerance bounds, often called "Basis Values" in industry, are used in the structural design. Depending on the type of structure, these Basis Values have a content of either 90% or 99% with a lower confidence limit of 95%: these are intended to provide a level of assurance that the structure will be able to carry the required loads given the variability and uncertainty of the material properties. If those Basis Values are found to be non-conservative (for example, due to process deviations when manufacturing the material), then the structural design of the aircraft may be inadequate to withstand the required loads. To guard against such a scenario, acceptance criteria for new lots of material are set and each new material lot is tested. The purpose of setting these acceptance criteria is to reject lots that differ from the material tested when computing the Basis Values, and would therefore invalidate the structural design. In contrast with applications such as fill-weight targets, where the distribution parameters are specified and acceptance criteria are set to ensure that the process produces the desired distribution, in the aerospace composite material industry, the only purpose of setting acceptance criteria is to ensure that the distribution of the material property follows the same distribution as the qualification sample, which was used to set the Basis Values.

It is common practice in industry to set lower limits for the lot mean and the lower extremum from the lot (often called the minimum individual or the first order statistic) when determining whether to accept a lot. In fact, such an approach is recommended by the Federal Aviation Administration (FAA), which regulates civil aviation in the United States (Tomblin et al., 2003).

Several authors have developed statistical methods for setting so-called dual-acceptance criteria which determine whether to reject a lot based on the sample mean and an order statistic. Croakin and Yang (1982) provide two approximations to the joint distribution of a sample mean and the kth order statistic of that sample: one such approximation assumes independence of the sample mean and the kth order statistic; the other is a large sample approximation. Vangel (2002) focused on the case where the first order statistic and the sample mean are used and developed a saddlepoint approximation to the joint distribution of the acceptance sample mean and lower extremum, assuming that the data are normally distributed. Ma and Robinson (2011) developed an expression based on the sample mean and the truncated sample mean which produces acceptance criteria similar to those of Vangel, but that are simpler to compute, though with a corresponding increase in error (on the order of a few percentage points in some cases).

Croakin and Yang (1982), Vangel (2002) and Ma and Robinson (2011) all consider the case where the population parameters for the manufacturing process are known and develop various dual-acceptance criteria to determine whether a sample is drawn from this known distribution. This assumption of known population parameters is violated for aerospace composite materials, where it is common for qualification samples to have a size of 18, or in some cases 30 for each material property (CMH-17-1G, 2012). These relatively small qualification samples produce large uncertainty in estimating the population parameters, but these small samples are used due to the high financial cost of producing the test specimens themselves and the high cost of measuring the material properties.

Later research considers the slightly different case of dual acceptance criteria with fixed limits. These fixed limits would typically be set to control the weight or volume of packaged goods. Linkletter et al. (2012) developed an approximation to the probability of acceptance of a sample from a process with one or several sources of variability. Edirisinghe et al. (2020) studied the coverage probabilities of two dual-acceptance criteria applicable to processes with a single source of variability: both a generalized pivotal quantity approach and an accelerated parametric bootstrap approach were used in that work. Neither Linkletter et al. (2012) nor Edirisinghe et al. (2020) discuss the setting of the limits for the dual-acceptance criteria, but instead focus on the properties of the acceptance criteria themselves: this work is important to industries where acceptance limits are set by regulation or are based on labeled minimum net quantities. However, for aerospace composite materials, there are no such predetermined acceptance limits, but instead the purpose of acceptance testing is to ensure that new material lots follows the same distribution as the qualification sample.

When using the current methods of setting acceptance criteria which assume known population parameters, such as that proposed by Vangel, practitioners resort to using the point-estimates of the parameters from the qualification sample. The small qualification samples common in industry provide a relatively poor estimate of the population parameters, so the acceptance criteria developed often have Type I error rates that are greater than the nominal value of α, as is shown in Section 6 of this article. To combat the inflated Type I error rates, practitioners resort to selecting relatively low nominal values of α. The current industrial practice is to use $\alpha = 0.01$ for lot acceptance and $\alpha = 0.05$ when assessing changes to the manufacturing process or a new manufacturing site (CMH-17-1G, 2012). When a practitioner selects a lower nominal value of α, there is a corresponding reduction in the statistical power of the test. Hence, a robust method to determine acceptance criteria is required that has a Type I error rate equal to the nominal value of α for small samples: a two-sample test is proposed here by the author. This proposed two-sample method uses the sample mean and standard deviation of the qualification sample, and the mean and lower extremum (minimum individual) of the acceptance sample to reject lots of material that do not come from the same distribution as the qualification sample.

Despite the stated goal of ensuring that lower tolerance limits developed during qualification testing remain valid for new lots of material, to date, relatively little emphasis has been placed on the power of the acceptance tests for aerospace composite materials. The size of the qualification and acceptance samples, as well as nominal values of α are typically based on experience from previous aircraft programs, rather than being determined by power calculations. The power of the proposed dual-acceptance test is investigated to enable the quality professional to rationally determine samples sizes based on effect sizes that would be of concern for the structural application.

The body of this paper is broken down as follows. First, the derivation of the new method is presented. Second, the computational methods required for this method are discussed and tables and software are provided so that the practitioner is not required to perform the computations themselves. Next, simulation is used to verify that the new method produces the correct probability of Type I error, to investigate the statistical power, to compare this new method with other methods in use in industry, and to examine the performance of the method with distributions other than normal. Finally, a case study using real composite material data to demonstrate the utility of the new method is presented.

2 Development of the acceptance criteria

Let the qualification sample be an iid sample of size n from a normal distribution with (unknown) parameters μ and σ. The qualification sample is denoted as $\{X_1, X_2, ..., X_n\}$. The acceptance sample is assumed to be an iid sample of m observations drawn from the same normal distribution and is denoted as $\{Y_1, Y_2, ..., Y_m\}$.

In the derivation of the distribution functions, the following pivotal quantities will be used:

$$V = \frac{\bar{X} - \mu}{\sigma} \tag{1}$$

$$W = \frac{S}{\sigma} \tag{2}$$

where $\bar{X}$ and S are the mean and standard deviation of the qualification sample, respectively.

The variables V and W are distributed according to $V \sim N\left(0, \frac{1}{\sqrt{n}}\right)$ and $(n-1)\,W^2 \sim \chi^2_{n-1}$, respectively. The pdfs of the variables V and W are therefore:

$$f_V\left(v; n\right) = \sqrt{n}\,\phi\left(v\sqrt{n}\right) \tag{3}$$

$$f_W\left(w; n\right) = \frac{\left(n-1\right)^{(n-1)/2} w^{n-2} e^{-(n-1)w^2/2}}{\Gamma\left(\left(n-1\right)/2\right) \times 2^{\frac{(n-1)}{2}-1}} \tag{4}$$

The acceptance data will be transformed as follows:

$$Z_i = \frac{Y_i - \mu}{\sigma} \tag{5}$$

Since V and W are pivotal quantities, the coverage probability for the lower extremum of the acceptance sample, $\mathrm{CP}_{Y_{(1)}}$, is given by the following expression. See, for example, Meeker et al. (2017) for a discussion of coverage probabilities.

$$\begin{aligned}
\mathrm{CP}_{Y_{(1)}}\left(k_1\right) &= \Pr\left[Y_{(1)} \le \bar{X} - k_1 S\right] \\
&= \Pr\left[Z_{(1)} \le V - k_1 W\right] \\
&= \mathrm{E}_{V,W}\left[\Pr\left(Z_{(1)} \le V - k_1 W | V, W\right)\right] \\
&= \int_{-\infty}^{\infty}\int_{0}^{\infty} \Pr\left(Z_{(1)} \le v - k_1 w | v, w\right) f_V\left(v; n\right) f_W\left(w; n\right) dw\, dv \\
&= \int_{-\infty}^{\infty}\int_{0}^{\infty} \left\{1 - \left[1 - \Phi\left(v - k_1 w\right)\right]^m\right\} f_V\left(v; n\right) f_W\left(w; n\right) dw\, dv
\end{aligned} \tag{6}$$

where k_1 is a constant that will be found later and $Y_{(1)}$ and $Z_{(1)}$ denote the lower extremum of the acceptance sample and the transformed acceptance sample, respectively.

The coverage probability for the mean of the acceptance sample, $\mathrm{CP}_{\bar{Y}}$, can be obtained using the same procedure and can be simplified to:

$$\mathrm{CP}_{\bar{Y}}(k_2) = 1 - \mathrm{t}_{n-1}\left[k_2\left(\frac{1}{n} + \frac{1}{m}\right)^{-\frac{1}{2}}\right] \tag{7}$$

where k_2 is a constant that will be found later and t_{n-1} is the pdf of the t-distribution with $n-1$ degrees of freedom.

Vangel (2002) developed a saddlepoint approximation to the joint distribution of the acceptance sample mean and lower extremum which assumed that the acceptance sample was drawn from a normal distribution with known parameters. Vangel's approximate joint distribution, $\widetilde{F}_{Y_{(1)},\bar{Y}}$, can be integrated over the distributions of V and W to remove the assumption that the distribution parameters are known. The coverage probability for the joint distribution of the acceptance sample mean and its lower extremum is thus:

$$\mathrm{CP}_{Y_{(1)},\bar{Y}}(k_1, k_2) = \int_{-\infty}^{\infty}\int_{0}^{\infty} \widetilde{F}_{Y_{(1)},\bar{Y}}(v - k_1 w, v - k_2 w)\, f_V(v; n)\, f_W(w; n)\, dw\, dv \tag{8}$$

where $\widetilde{F}_{Y_{(1)},\bar{Y}}(t_1, t_2)$ is from Vangel (2002) and is defined as:

$$\widetilde{F}_{Y_{(1)},\bar{Y}}(t_1, t_2) = \frac{\int_{-\infty}^{\hat{\lambda}} \Phi\left(\sqrt{m}t_2\right) A(t)\, dt + \int_{\hat{\lambda}}^{\infty} \Phi\left\{\sqrt{m}\left[t_1 + \frac{m-1}{m}\left(h(t) - t\right)\right]\right\} A(t)\, dt}{\int_{-\infty}^{\infty} A(t)\, dt} \tag{9}$$

$$A(t) = h^{-(m-1)}(t) \exp\left[\frac{(m-1)^2}{2m}\left(h(t) - t\right)^2 + (m-1)\,t\left(h(t) - t\right)\right]\sqrt{1 - h^2(t) + t h(t)} \tag{10}$$

where $h(t)$ is the normal hazard function and $\hat{\lambda}$ is the solution to:

$$\frac{m-1}{m}\left(h\left(\hat{\lambda}\right) - \hat{\lambda}\right) = t_2 - t_1 \tag{11}$$

Acceptance criteria are constructed such that a lot is (falsely) rejected with a probability of α. Since a lot would be rejected if either of the mean or minimum criteria are violated, this implies that:

$$\alpha = \mathrm{CP}_{Y_{(1)}}(k_1) + \mathrm{CP}_{\bar{Y}}(k_2) - \mathrm{CP}_{Y_{(1)},\bar{Y}}(k_1, k_2) \tag{12}$$

If we set the condition that $\mathrm{CP}_{Y_{(1)}}(k_1) = \mathrm{CP}_{\bar{Y}}(k_2)$, unique values for k_1 and k_2 can be obtained for each combination of the variables α, n and m.

Once the values k_1 and k_2 are determined, acceptance criteria for the lot minimum (A_1) and lot mean (A_2) for a particular material property can be set as follows:

$$A_1 = \bar{X} - k_1 S \tag{13}$$

$$A_2 = \bar{X} - k_2 S \tag{14}$$

where $\bar{X}$ and S are the mean and standard deviation of the qualification sample, respectively.

One may question why the method of Vangel (2002) was used as a basis for the present acceptance criteria rather than the method of Ma and Robinson (2011). While the expression developed by Ma and Robinson is simpler, the Monte Carlo simulation performed by Ma and Robinson shows a relative error between their approximation of the joint distribution of the sample mean and lower extremum and the simulation results of up to several percentage points for a standard normal distribution. For this reason, Vangel's work is preferred as a basis for the present work.

3 Computation of k1 and k2

Since the factors k_1 and k_2 depend only on the values of the qualification sample size (n), the acceptance sample size (m) and α, values of these factors can be tabulated. These factors have been computed for select values of n, m and α using a a program written using a combination of C++ and the R language (R Core Team, 2021) (see supplemental material). The computed factors k_1 and k_2 are given in Table 1 and 2, respectively. These tables also list the factors computed using the method given in Vangel (2002), which correspond to the case of $n = \infty$: these are computed using the R package `cmstatr` (Kloppenborg, 2020), which produced values equal to those published by Vangel within three decimal places.

Computation of the factors k_1 and k_2 using a naive approach is computationally expensive due to the need to perform a triple-integration complicated function given in equations 8 and 9 in which the bounds of the innermost integral depend on the values of the other two variables of integration. The author has found that a naive implementation using the quadrature routines in the GNU Scientific Library (Galassi, Davies, Theiler, Gough, Jungman, Booth, and Rossi, Galassi et al.) takes an extremely long time to produce results (on the order of hours). The author has developed a software package, which is provided in the supplementary material, that produces results within a few seconds on a modern personal computer. This software package relies on a bespoke quadrature routine which is optimized for integration of this particular function.

Table 1: Factors for the lower extremum (k_1) for lot acceptance

	Acceptance sample size (m)							
n	3	4	5	6	7	8	9	10
$\alpha = 0.05$								
12	2.765	2.938	3.069	3.178	3.266	3.342	3.406	3.465
18	2.601	2.753	2.867	2.959	3.036	3.102	3.157	3.207
24	2.524	2.668	2.776	2.861	2.932	2.990	3.043	3.090
30	2.483	2.621	2.724	2.804	2.872	2.929	2.979	3.022
36	2.454	2.589	2.688	2.768	2.832	2.888	2.938	2.980
50	2.416	2.546	2.642	2.718	2.781	2.833	2.879	2.920
100	2.369	2.492	2.584	2.656	2.715	2.765	2.809	2.847
1000	2.328	2.447	2.533	2.602	2.659	2.706	2.747	2.784
∞	2.324	2.442	2.529	2.597	2.653	2.700	2.741	2.777
$\alpha = 0.01$								
12	3.746	3.934	4.062	4.180	4.273	4.355	4.426	4.484
18	3.418	3.570	3.676	3.770	3.840	3.910	3.957	4.016
24	3.277	3.406	3.506	3.582	3.652	3.711	3.758	3.805
30	3.195	3.312	3.406	3.488	3.547	3.600	3.646	3.688
36	3.137	3.260	3.348	3.418	3.477	3.529	3.570	3.617
50	3.066	3.184	3.266	3.336	3.389	3.436	3.477	3.512
100	2.984	3.090	3.166	3.230	3.277	3.324	3.359	3.395
1000	2.914	3.008	3.078	3.137	3.184	3.225	3.260	3.289
∞	2.903	3.000	3.071	3.128	3.175	3.215	3.250	3.281
$\alpha = 0.005$								
12	4.180	4.367	4.508	4.625	4.719	4.812	4.883	4.953
18	3.758	3.898	4.016	4.109	4.180	4.250	4.297	4.344
24	3.570	3.711	3.805	3.875	3.945	4.004	4.062	4.098
30	3.477	3.594	3.688	3.758	3.828	3.875	3.922	3.957
36	3.406	3.523	3.617	3.688	3.734	3.781	3.828	3.875
50	3.324	3.430	3.512	3.570	3.629	3.676	3.711	3.758
100	3.219	3.312	3.395	3.453	3.500	3.547	3.570	3.605
1000	3.125	3.219	3.289	3.336	3.383	3.430	3.453	3.488
∞	3.120	3.210	3.278	3.331	3.375	3.413	3.446	3.475

Table 2: Factors for the mean (k_2) for lot acceptance

	Acceptance sample size (m)							
n	3	4	5	6	7	8	9	10
$\alpha = 0.05$								
12	1.330	1.198	1.109	1.047	0.998	0.959	0.927	0.901
18	1.243	1.110	1.020	0.954	0.904	0.864	0.831	0.803
24	1.201	1.068	0.977	0.910	0.858	0.817	0.783	0.755
30	1.179	1.044	0.952	0.884	0.831	0.789	0.755	0.726
36	1.162	1.027	0.935	0.866	0.813	0.771	0.736	0.707
50	1.140	1.005	0.911	0.842	0.789	0.745	0.709	0.680
100	1.113	0.976	0.882	0.812	0.756	0.712	0.675	0.644
1000	1.089	0.951	0.855	0.784	0.728	0.682	0.644	0.612
∞	1.087	0.949	0.853	0.781	0.725	0.679	0.641	0.609
$\alpha = 0.01$								
12	1.934	1.741	1.605	1.513	1.441	1.385	1.339	1.299
18	1.759	1.569	1.436	1.343	1.269	1.214	1.163	1.128
24	1.682	1.487	1.357	1.261	1.189	1.132	1.083	1.045
30	1.636	1.440	1.310	1.218	1.143	1.084	1.037	0.996
36	1.602	1.413	1.282	1.186	1.112	1.054	1.004	0.966
50	1.562	1.374	1.242	1.147	1.071	1.011	0.962	0.920
100	1.515	1.324	1.192	1.096	1.018	0.959	0.908	0.866
1000	1.474	1.279	1.146	1.049	0.972	0.911	0.859	0.815
∞	1.467	1.275	1.143	1.044	0.968	0.906	0.855	0.811
$\alpha = 0.005$								
12	2.193	1.971	1.820	1.713	1.630	1.570	1.516	1.475
18	1.967	1.747	1.605	1.500	1.417	1.355	1.299	1.254
24	1.864	1.655	1.508	1.398	1.319	1.255	1.206	1.158
30	1.812	1.596	1.453	1.345	1.268	1.201	1.148	1.101
36	1.772	1.560	1.419	1.313	1.227	1.160	1.108	1.067
50	1.725	1.511	1.368	1.258	1.179	1.113	1.057	1.016
100	1.664	1.449	1.309	1.202	1.118	1.054	0.993	0.948
1000	1.608	1.399	1.255	1.144	1.061	0.997	0.937	0.892
∞	1.605	1.394	1.249	1.141	1.057	0.989	0.933	0.885

4 Validation

To validate the present two-sample test and the factors presented in Tables 1 and 2, a simulation study was performed. A total of 5000 simulated qualification samples were drawn from a standard normal distribution. Acceptance criteria for the mean and lower extremum were calculated for each of these qualification samples using Equations 13 and 14 for $\alpha = 0.05$. For each qualification sample, 5000 simulated acceptance samples were drawn from the same population and those samples are compared against the acceptance criteria. This simulation is repeated for several combinations of qualification sample size, n, and acceptance sample size, m. The rejection rate from this simulation is shown in Figure 1. It can be seen that the simulated rejection rate is very nearly equal the selected value of α.

The same simulation was repeated using the the techniques from (Vangel, 2002). These results are also shown in Figure 1. It is evident that Vangel's approach yields acceptance criteria that cause far too many acceptance samples to be rejected, especially for modest qualification sample sizes. Recall that typical qualification samples (n) used in industry have a size of 18 or 30. The mismatch between the selected value of α and rejection rate of Vangel's method is noted in Section 7 of Vangel (2002).

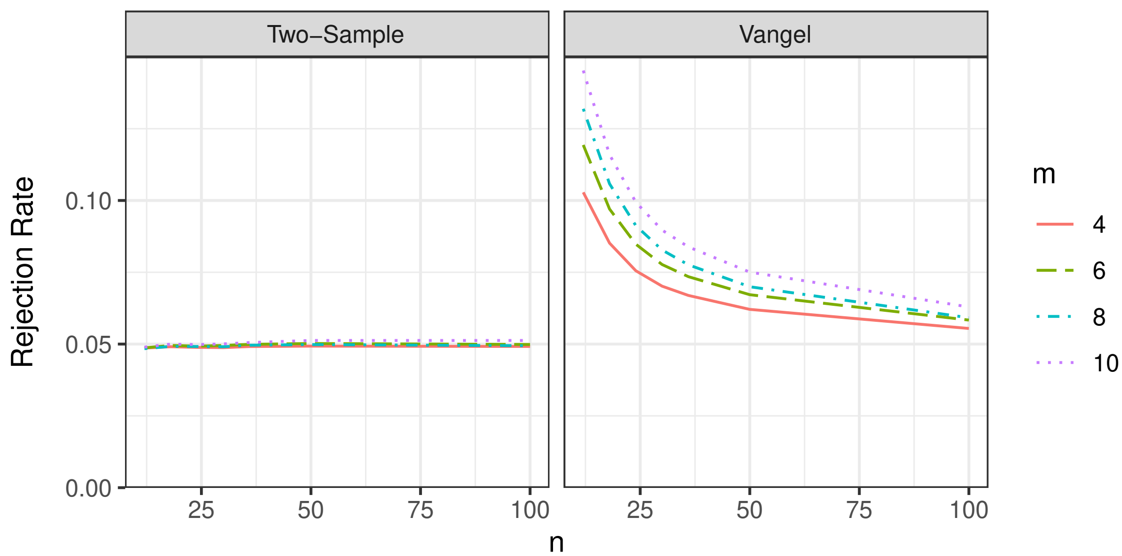

Figure 1: Probability of rejecting a lot. Qualification sample size is n, acceptance sample size is m, $\alpha = 0.05$. Qualification and acceptance samples drawn from the same distribution.

5 Power

The power of the present two-sample test is an important consideration when selecting sample sizes. Once the practitioner determines the effect size that should be detected and the desired value of α, they can select appropriate sample sizes for the qualification and acceptance samples.

The power of the present two-sample test was investigated through simulation. Simulated qualification samples were drawn from a normal distribution with parameters μ and σ; this was repeated for 2500 qualification samples. For each qualification sample, 2500 simulated acceptance samples were generated; each set being drawn from a normal distribution with μ and/or σ that are different from the qualification distributions by various amounts. This was repeated for several combinations of qualification sample size, n and acceptance sample size, m. In all cases, $\alpha = 0.05$ was used.

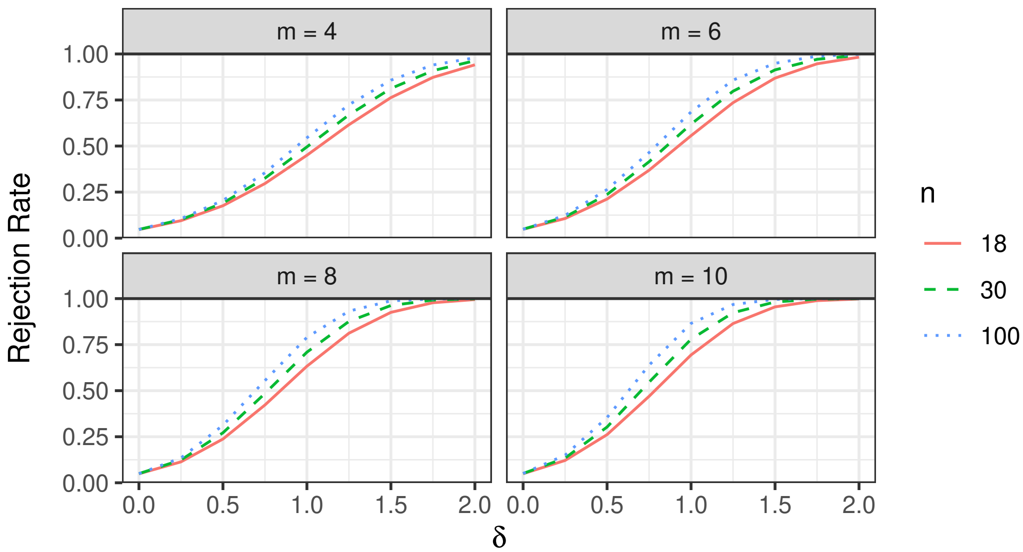

Figure 2: Probability of rejecting a lot for the *RM* case. Qualification sample size is n, acceptance sample size is m, $\alpha = 0.05$. Acceptance sample is drawn from $N(\mu - \delta\sigma, \sigma)$.

First, a reduced mean (*RM*) case is investigated. Here, the acceptance samples are drawn from an $N(\mu - \delta\sigma, \sigma)$ distribution, where $\delta \geq 0$. Figure 2 shows the result of this simulation study. It is common in industry for qualification samples to have a size of 18 and acceptance samples to have a size of 6: in that case, the present two-sample test has a power of 0.8 for detecting a reduction in mean of 1.37σ.

Next, an increased standard deviation (*ISD*) case is investigated. In this case, the acceptance samples are drawn from an $N(\mu, \delta\sigma)$ distribution, where $\delta \geq 1$. Figure 3 shows the results of this simulation study. For the case where $n = 18$ and $m = 6$, the present two-sample test has a power of 0.8 for detecting a standard deviation that it at least 4.30 times the that of the qualification population.

Vangel (2002) notes that in some cases, it is more important to be able to detect very low values than it is to detect high variance. To investigate this scenario, a low minimum individual (*LMI*) case is considered. In the *LMI* case, the acceptance sample comprises of one observation drawn from an $N(\mu - \delta\sigma, \sigma)$ distribution and $m - 1$ observations from an $N(\mu, \sigma)$ distribution where $\delta \geq 0$. Figure 4 shows the results of the simulation study for

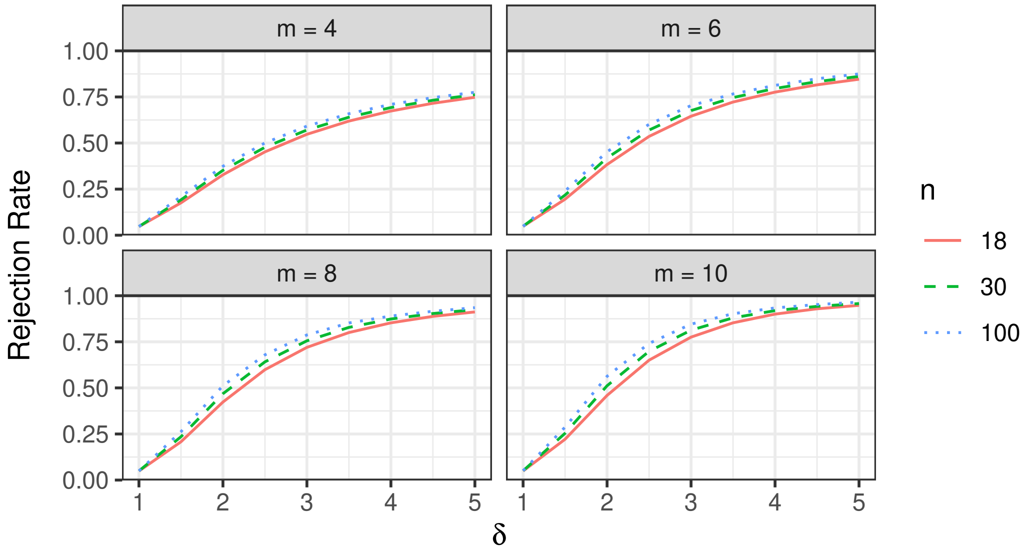

Figure 3: Probability of rejecting a lot for the *ISD* case. Qualification sample size is n, acceptance sample size is m, $\alpha = 0.05$. Acceptance sample is drawn from $N(\mu, \delta\sigma)$.

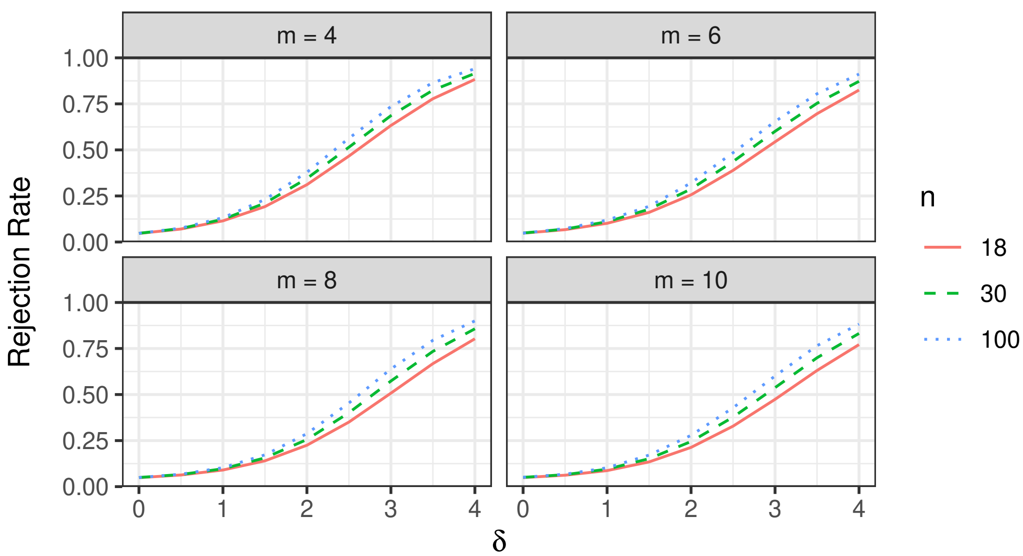

Figure 4: Probability of rejecting a lot for the *LMI* case. Qualification sample size is n, acceptance sample size is m, $\alpha = 0.05$. One observation of acceptance sample is drawn from $N(\mu - \delta\sigma, \sigma)$, remaining observations are drawn from $N(\mu, \sigma)$.

this case. In this case, when $n = 18$ and $m = 6$, the present two-sample test has a power of 0.8 for detecting a single value that is drawn from a distribution with mean 3.90σ below the qualification mean. Note that while the power of this test increases with qualification sample size, n, it decreases with increasing acceptance sample size, m. This is due to the distribution of the acceptance sample, which in this example, departs substantially from normality.

Appendix B contains results of a simulation study where the acceptance sample is drawn from a mixture distribution (*MD*) containing observations from a $N(\mu - \delta\sigma, \sigma)$ distribution with probability 0.4 and the remaining observations a $N(\mu, \sigma)$ distribution. The results are qualitatively similar to those from the *LMI* case, though unlike the *LMI* case, the power increases with increasing acceptance sample size, m.

Appendix C contains plots of effect size that can be detected with a statistical power of 80% versus sample size for the four cases (*RM*, *ISD*, *LMI* and *MD*) considered in the present simulation study. While these plots contain the same information as Figures 2 through 4, they may be more useful for a practitioner when selecting required samples sizes to detect a certain effect size.

6 Comparison with other methods

The present two-sample method has been compared with two other methods of setting acceptance criteria: the method proposed by Vangel (2002) and a mean and standard deviation (*MSD*) criteria (which is used as a comparison in Vangel's paper). The present simulation study is restricted to the evaluation of criteria with limits that can be pre-computed based only on the qualification sample without prior knowledge of the acceptance sample. This is done in industry when writing a requirements document for future lots of material (usually called a material specification).

For the *MSD* criteria, acceptance limits are set for the acceptance sample mean and the acceptance sample standard deviation. One-sided prediction intervals for the mean and standard deviation are used to set acceptance limits. Full details are given in Appendix A. The use of one-sided prediction intervals, rather than traditional hypothesis testing (i.e. using two-sample t-tests and F-tests) allows pre-computing of the acceptance limits.

A simulation study similar to that described in Section 5 was performed for the three methods of setting lot acceptance criteria: the present two-sample method, the method of Vangel (2002) and the MSD method described above. The same three cases (*RM*, *ISD* and *LMI*) were considered. The results of these simulations for $n = 18$ and $m = 6$ are presented in Figures 5 through 7. Additionally, Appendix B contains a comparison of the power of these three criteria for the *MD* case.

The *RM* case is shown in Figure 5. As previously discussed, the false positive rate of Vangel's method exceeds the set value of α (this corresponds with $\delta = 0$ in this figure), thus making the use of this method undesirable in practice due to an unexpectedly high producer's risk. The present two-sample test has a slightly greater power for detecting reduced mean than the *MSD* criteria.

Figure 6 shows the *ISD* case. Again, the method of Vangel falsely rejects too many lots that are drawn from the same distribution as the qualification sample. The *MSD* criteria has a much higher power for detecting increased standard deviation. However, it is unlikely that a process deviation would cause an increase in the standard deviation that

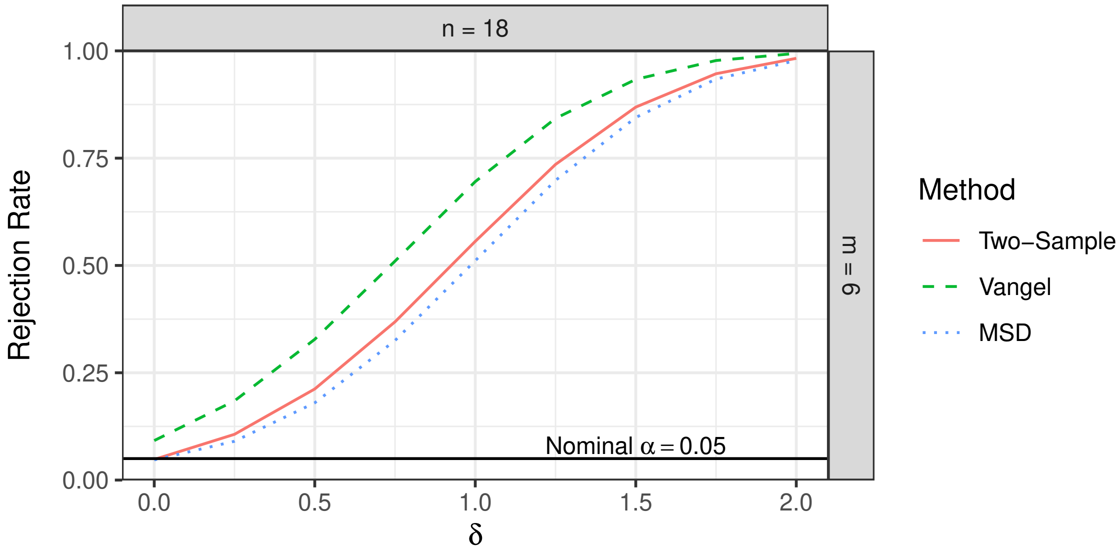

Figure 5: Probability of rejecting a lot for RM case for various lot acceptance criteria with $\alpha = 0.05$. Acceptance sample is drawn from $N(\mu - \delta\sigma, \sigma)$.

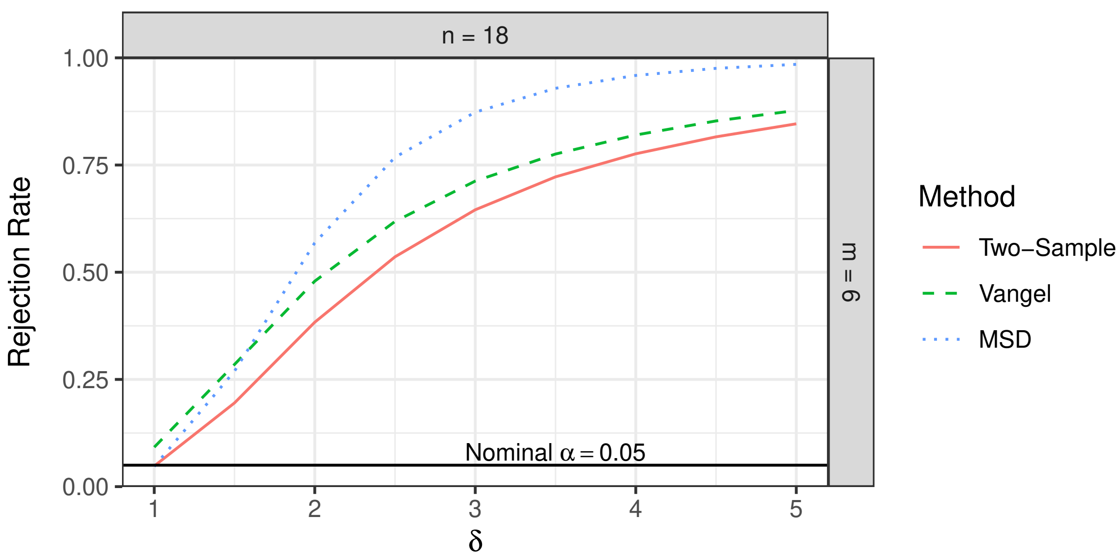

Figure 6: Probability of rejecting a lot for ISD case for various lot acceptance criteria with $\alpha = 0.05$. Acceptance sample is drawn from $N(\mu, \delta\sigma)$.

can be reliably detected by any of these criteria (increasing the standard deviation by a factor of 2.65 times for the MSD criteria or 4.30 times for the two-sample method) without affecting the mean. In order for the mean to remain unchanged, the density of both the left tail and the right tail would need to be increased. While a process deviation could certainly add density of the left tail (i.e. reduce material strength), it is unlikely that for a modern composite material, a process deviation would substantially increase the strength of some portion of the material. Therefore, the performance of the various acceptance criteria for the *ISD* case is less important than for other cases.

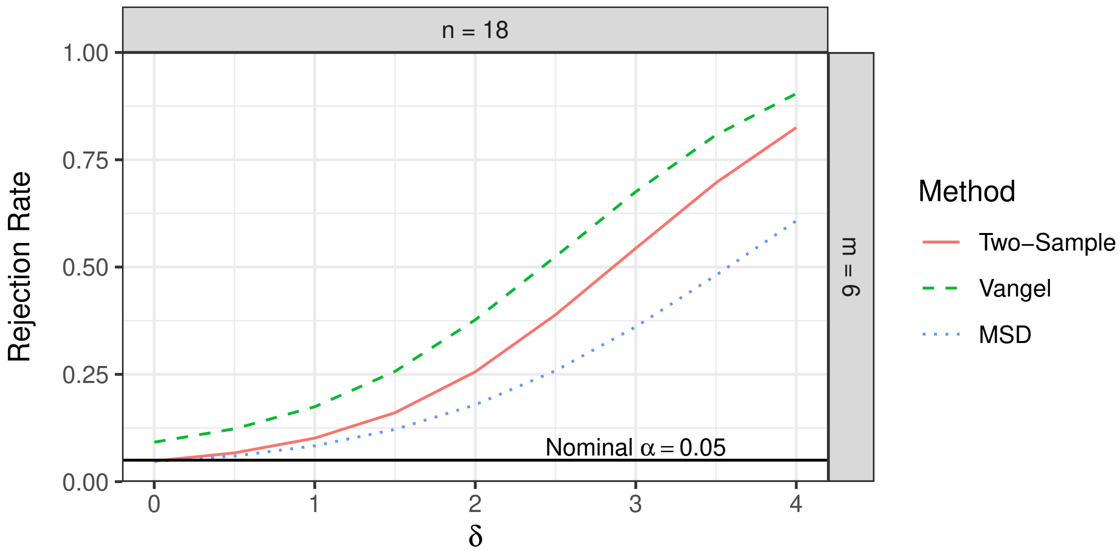

Figure 7: Probability of rejecting a lot for *LMI* case for various lot acceptance criteria with $\alpha = 0.05$. One observation of acceptance sample is drawn from $N(\mu - \delta\sigma, \sigma)$, remaining observations are drawn from $N(\mu, \sigma)$.

Figure 7 shows the *LMI* case. Here, the present two-sample test has a higher power than the *MSD* criteria and also has the correct rate of Type 1 error, whereas the criteria based on Vangel does not.

7 Other Distributions

Each of the three acceptance criteria discussed in Section 6 assume normal distributions. To investigate the effect of violating this assumption, the same simulation study was repeated for lognormal and Weibull distributions. These distributions were chosen since both have historically been used to describe composite material strength data.

The simulation from Sections 4, 5 and 6 was repeated for samples drawn from a lognormal and a Weibull distribution. For both distributions, a mean of 100 and a coefficient of variation (CV) of 6% is used for the qualification sample. The parameters for the distributions of the acceptance samples are selected such that the acceptance distribution mean is δ times the qualification population standard deviation less than the qualification population mean. That is:

$$(mean)_{acceptance} = (mean)_{qualification} - \delta \cdot (SD)_{qualification}$$

Further details of the selection of parameters are available in the supplemental material.

Figures 8 through 11 show the statistical power found from the simulations using lognormal and Weibull distributions. For lognormal distributions, the present two-sample method exhibited moderate differences between the nominal value of α and the Type I error rate for the typical qualification sample size of $n = 18$ and increasing Type I error rates for larger values of n. For Weibull distributions, the present two-sample method had very large Type I error rates. For both lognormal and Weibull distributions, the present two-sample method had higher statistical power than the *MSD* criteria. As was seen in the normal distribution simulations, Vangel's method had higher power than the present two-sample method with a corresponding increase mismatch between the Type I error rate and the nominal value of α.

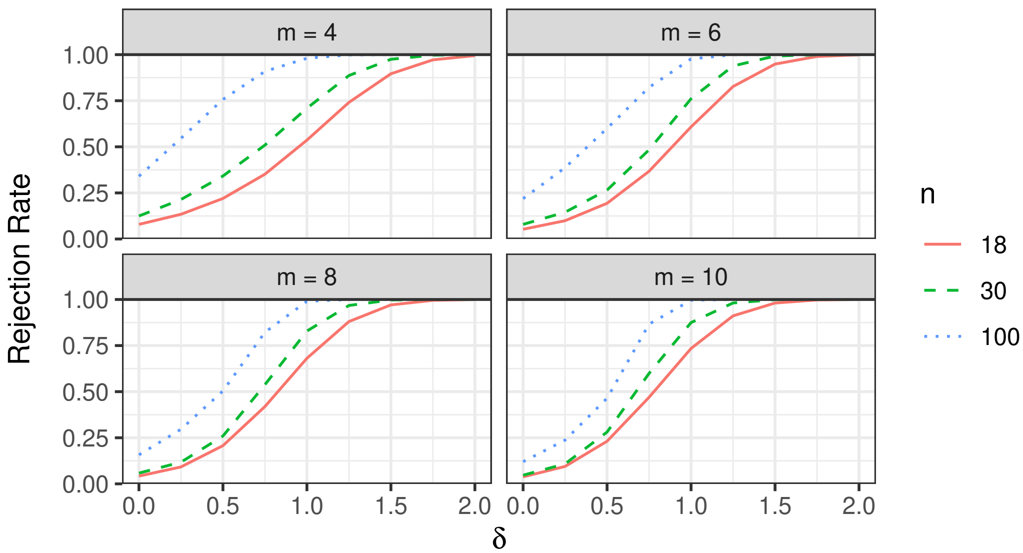

Figure 8: Probability of rejecting a lot for lognormal qualification and acceptance samples. Qualification sample size is n, acceptance sample size is m, $\alpha = 0.05$. Acceptance sample is drawn from a distribution with a mean that is δ times the standard deviation less than the qualification population mean.

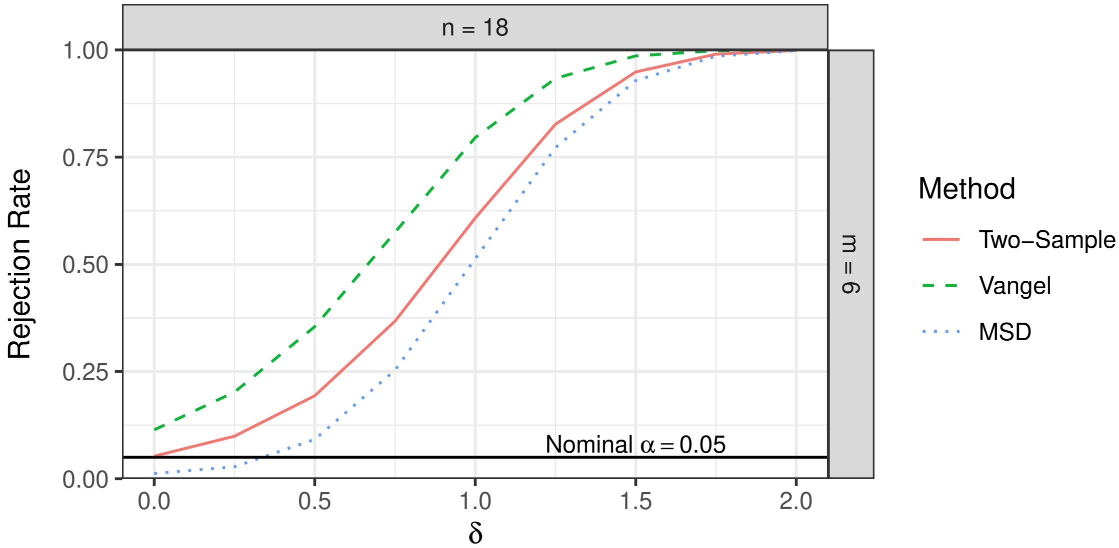

Figure 9: Probability of rejecting a lot following a lognormal distribution for various lot acceptance criteria with $\alpha = 0.05$. Acceptance sample is drawn from a distribution with a mean that is δ times the standard deviation less than the qualification population mean.

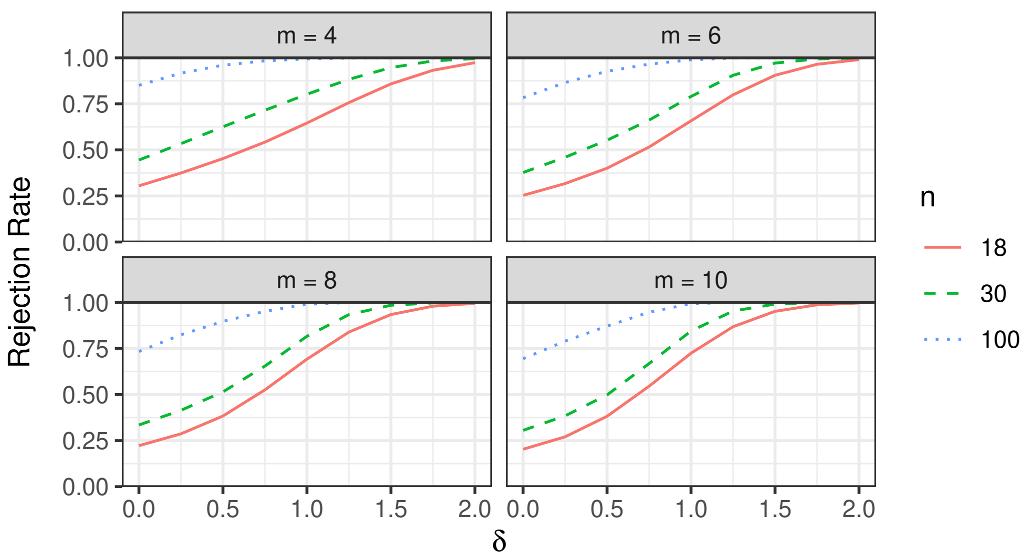

Figure 10: Probability of rejecting a lot for Weibull qualification and acceptance samples. Qualification sample size is n, acceptance sample size is m, $\alpha = 0.05$. Acceptance sample is drawn from a distribution with a mean that is δ times the standard deviation less than the qualification population mean.

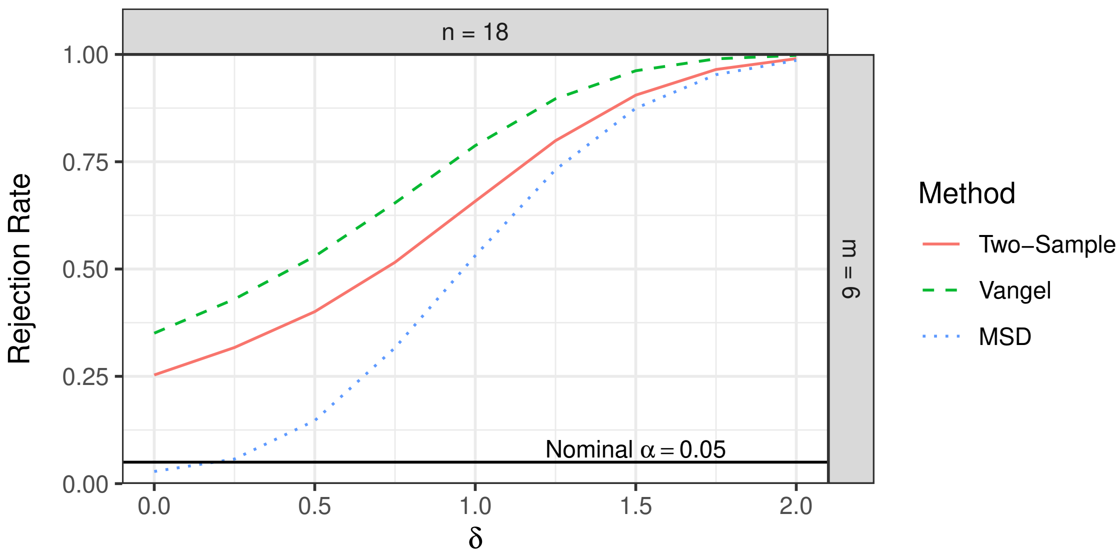

Figure 11: Probability of rejecting a lot following a Weibull distribution for various lot acceptance criteria with $\alpha = 0.05$. Acceptance sample is drawn from a distribution with a mean that is δ times the standard deviation less than the qualification population mean.

8 Case study

A manufacturer of composite components that specializes in making parts for regional jets and business jets has been using a particular composite material for several years. This material originally underwent qualification testing: for one of the material properties, the qualification sample had a size of 22, a mean of 255 and a standard deviation of 18.7. Based on this qualification sample, the manufacturer has calculated a Basis value with 90% content and a lower confidence limit of 95%. This so-called B-Basis value was 220 and is used in the structural design of certain components of a business jet.

The manufacturer has set acceptance limits for the mean and minimum individual using the method of Vangel (2002), and uses acceptance samples of 3 specimens. As is routinely done in industry, they have assumed that the qualification sample mean and standard deviation are equal to the population parameters. The manufacturer has elected to set $\alpha = 0.01$, which results in acceptance limits of 201 for the minimum individual and 227 for the mean. Based on these limits, they have rejected 3 out of the last 139 lots (2.2% of the lots). Process records for the rejected lots were thoroughly reviewed and new acceptance samples were tested from each rejected lot — all of which passed — leading the manufacturer to believe that these lots were initially rejected in error.

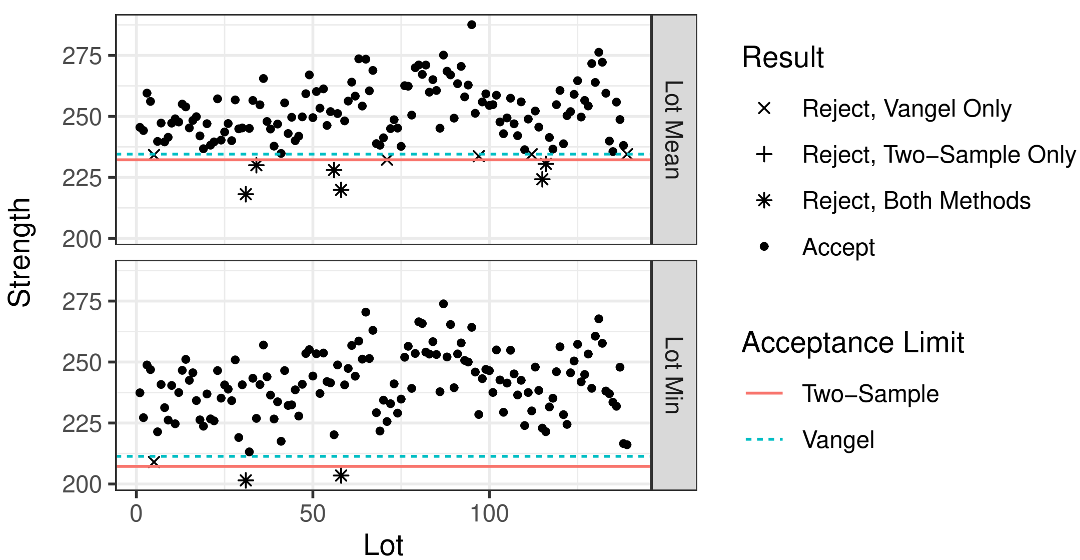

Figure 12: Lot mean and lot minimum for the previous 139 lots of a composite material. Acceptance limits based on Vangel (2002) and the two-sample method in this paper are shown. Lots which fail either criteria are indicated. For both acceptance criteria, $\alpha = 0.05$ is selected.

While the use of $\alpha = 0.01$ lot acceptance limits are standard practice in the industry (CMH-17-1G, 2012), the manufacturer is considering revising their acceptance limits to increase the power of the acceptance test. The manufacturer has decided that they are willing to accept a Type I error rate of 5%.

The manufacturer has compared the lot acceptance data from the previous 139 lots of material to acceptance limits based on Vangel's method and the two-sample method with $\alpha = 0.05$ as shown in Figure 12. They were surprised to find that Vangel's method with $\alpha = 0.05$ would have rejected 7.9% of the previous production lots and are worried that using Vangel's method with $\alpha = 0.05$ would result in production delays due to erroneously rejected lots of good material. While the rejection rate that the manufacturer found is not significantly different from the selected value of α (the 95% confidence interval is 4.0% to 13.7%), the higher rejection rate observed is consistent with the simulation study from Section 4, which found that Vangel's method rejects too many lots (see Figure 1).

If the manufacturer were to use the method of setting acceptance limits presented in this paper with $\alpha = 0.05$, they would reject 6 of the previous 139 lots, which represents a rejection rate of 4.3% (95% confidence interval of 1.6% to 9.2%). This is very close to the selected value of α and is consistent with the results of the simulation study in Section 4.

Table 3: Type I error rates from the simulation study for $n = 22$ and $m = 3$ compared with the proportion of the previous 139 production lots that are rejected by Vangel's acceptance criteria and the two-sample method in this paper.

Nominal α	Method	Simulation Study Type I Error Rate	Production Lots Rejected by Criteria
0.01	Vangel	0.0220	$3/139 = 0.0216$
	Two-Sample	0.0097	$2/139 = 0.0144$
0.05	Vangel	0.0709	$11/139 = 0.0791$
	Two-Sample	0.0476	$6/139 = 0.0432$

A simulation study similar to that in Section 6 was performed to compare Vangel's method and the two-sample method in this paper for the nominal values of $\alpha = 0.01$ and $\alpha = 0.05$. This simulation study considered samples drawn from $N(\mu, \sigma)$ distributions where μ varied from 210 to the mean of the qualification sample (255) and σ was set equal to the qualification sample standard deviation. The Type I error rate for the simulation study is summarized in table 3. Simulation showed that Vangel's method with a nominal value of $\alpha = 0.01$ has a Type I error rate of that closely matches the proportion of previous lots rejected, which is roughly twice the nominal value of α. The simulation study also showed that the Type I error rate for Vangel's method with a nominal value of $\alpha = 0.05$ is also well above the nominal α. The Type I error rate of the two-sample method closely matches the nominal value of α.

Next, the manufacturer reviews the power of the tests from the simulation study, which is shown in Figure 13. In addition to the acceptance limits based on Vangel's method and the present two-sample method, the simulation study also considers the mean and standard deviation (*MSD*) criteria discussed in Section 6. The manufacturer notes that the two-sample method has a higher power than the *MSD* criteria at the same nominal value of α. While Vangel's method has a higher power than the two-sample method at the same nominal value of α, the unexpectedly high Type I error rate associated with Vangel's method causes the manufacturer to decide to use the two-sample method with $\alpha = 0.05$, which provides the highest power (of the three methods considered) without exceeding the

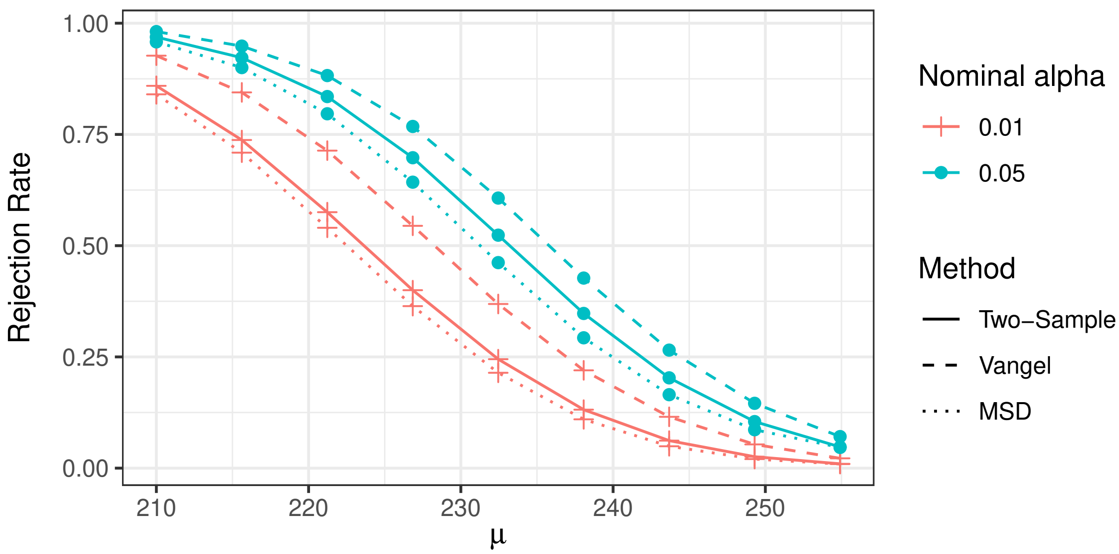

Figure 13: Probability of rejecting a lot drawn from a $N(\mu, \sigma)$ distribution with $\sigma = 18.73$ for the method of Vangel (2002) and the two-sample method in this paper. The qualification and acceptance sample sizes are $n = 22$ and $m = 3$, respectively.

desired Type I error rate. It should be noted that increasing the size of the acceptance sample (m) would provide an additional improvement in the power of the test, as suggested by Figure 2.

This case study shows a real-world scenario where the present two-sample method produces the correct probability of Type I error, allowing the selection of a higher nominal value of α and an associated higher power without exceeding a certain tolerable producer's risk.

9 Conclusion

A new two-sample method was presented that uses the sample mean and standard deviation of the qualification sample to set acceptance criteria for the mean and lower extremum of new lots of material, assuming both samples are drawn from a normal distribution. This method, like that of Vangel (2002), uses a saddlepoint approximation to the joint distribution of the mean and lower extremum, but unlike the prior art, this new method assumes that the population parameters are unknown.

A simulation study confirms that this test has the expected probability of Type I error, which is an important consideration in industry as falsely rejecting good lots of materials increases costs and causes delays. Simulation was also used to determine the power of this test to detect various types of difference between qualification and acceptance samples. The results of the power simulation can be used by a quality professional to determine sample size requirements based on an effect size that should be detected.

Simulation was used to compare the present two-sample test to other lot acceptance methods; this simulation showed that the present two-sample test performs better than other methods for detecting certain types of differences between qualification and acceptance samples, while maintaining the expected probability of Type I error.

The case study presented in this paper illustrates the use of the proposed method for setting acceptance limits. The case study shows agreement between results with real-world composite material data and the simulation studies performed, as well as showing how a quality professional would use the proposed method.

About the author

Stefan Kloppenborg currently works as an engineer at De Havilland Aircraft of Canada. Over the course of his career, he has performed material characterization and statistical analysis for composite materials as well as the design, analysis, testing and certification of composite and metallic aircraft components.

Acknowledgments

I would like to express my gratitude to the editor and two anonymous referees for their valuable comments and suggestions. I would also like to thank Comtek Advanced Structures, Ltd. for providing the data presented in Section 8. I am grateful to Jonathan Tisack (BeiGene) and Billy Cheng (Comtek Advanced Structures) for providing valuable comments on early drafts of this manuscript.

Disclosure statement

The author reports that there are no competing interests to declare.

ORCID

Stefan Kloppenborg: `https://orcid.org/0000-0002-1908-5214`

SUPPLEMENTARY MATERIAL

R-package EquivSample: R-package containing code to compute the factors k_1 and k_2 (GNU zipped tar file)

Simulation and Case Study Code and Data: The R code used to perform the simulation described in Sections 4, through 7, and the case study described in 8. This code is contained in a series of R-Markdown (Xie et al., 2018) documents. Additionally, a CSV file containing the data from the case study is included. (zipped file)

References

CMH-17-1G (2012, Mar). *Composite Materials Handbook, Volume 1. Polymer Matrix Composites Guideline for Characterization of Structural Materials*. SAE International.

Croakin, M. and G. Yang (1982, Dec). Acceptance probabilities for a sampling procedure based on the mean and an order statistic. *Journal of Research of the National Bureau of Standards 87*(6), 485–511.

Edirisinghe, P. M., T. Mathew, and T. S. G. Peiris (2020). Confidence limits for compliance testing using mixed acceptance criteria. *Quality and Reliability Engineering International 36*(4), 1197–1204.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, M. Booth, and F. Rossi. *GNU Scientific Library - Reference Manual* (3rd ed.). ISBN 0954612078.

Kloppenborg, S. (2020). cmstatr: An R package for statistical analysis of composite material data. *Journal of Open Source Software 5*(51), 2265.

Krishnamoorthy, K. and T. Mathew (2008). *Statistical Tolerance Regions: Theory, Applications, and Computation*. John Wiley & Sons.

Linkletter, C. D., P. Ranjan, C. D. Lin, D. R. Bingham, W. A. Brenneman, R. A. Lockhart, and T. M. Loughin (2012). Compliance testing for random effects models with joint acceptance criteria. *Technometrics 54*(3), 243–255.

Ma, C. and J. Robinson (2011). Lot acceptance and compliance testing based on the sample mean and minimum/maximum. *Journal of Statistical Planning and Inference 141*(7), 2440–2448.

Meeker, W., G. Hahn, and L. Escobar (2017). *Statistical Intervals: A Guide for Practitioners and Researchers* (2nd Edition ed.). John Wiley & Sons.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Tomblin, J., Y. Ng, and K. S. Raju (2003, Sep). Material qualification and equivalency for polymer matrix composite material systems: Updated procedure. Technical Report DOT/FAA/AR-03/19, Federal Aviation Administration, Washington, D.C.

Vangel, M. (2002, Aug). Lot acceptance and compliance testing using the sample mean and an extremum. *Technometrics 44*(3), 242–249.

Xie, Y., J. Allaire, and G. Grolemund (2018). *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman and Hall/CRC. ISBN 9781138359338.

A Mean and Standard Deviation Criteria

In Section 6 of this paper, a Mean and Standard Deviation (*MSD*) criteria is considered. Many tests that could be properly called a test of mean and standard deviation could be formulated. However, the formulation of the test used in this paper is described in this appendix.

This test is formulated so that numeric criteria for an acceptance sample mean and standard deviation can be computed when only the qualification sample is know.

The test described herein comprises of two parts: a one-sided test for the acceptance sample mean, and a one-sided test for the acceptance sample standard deviation. For the test for the acceptance sample mean, the critical (lower) value is given by:

$$A_m = \bar{X} - t_{(1-\alpha_m;n-1)} \left(\frac{1}{n} + \frac{1}{m} \right)^{\frac{1}{2}} S \tag{15}$$

where α_m is the probability of making a Type I error when testing the mean, and the other symbols have the same meaning as before.

The test for the acceptance standard deviation has a critical (upper) value given by:

$$A_s = S \left[F_{(1-\alpha_s;m-1;n-1)} \right]^{\frac{1}{2}} \tag{16}$$

where α_s is the probability of making a Type I error when testing the standard deviation and $F_{()}$ is the pdf of the F-distribution.

For normally distributed data, the sample mean and standard deviation are independent, so the probability of making a Type I error on either test is:

$$\alpha = \alpha_m + \alpha_s \tag{17}$$

We arbitrarily choose the condition that $\alpha_m = \alpha_s$, so the critical values become:

$$A_m = \bar{X} - t_{(1-\alpha/2;n-1)} \left(\frac{1}{n} + \frac{1}{m} \right)^{\frac{1}{2}} S \tag{18}$$

$$A_s = S \left[F_{(1-\alpha/2;m-1;n-1)} \right]^{\frac{1}{2}} \tag{19}$$

The astute reader will recognize the critical values computed above as one-sided prediction intervals.

B Mixture Distribution

The simulation from Sections 5 and 6 was repeated for acceptance samples drawn from a mixture distribution. Each observation in the acceptance sample is drawn from $N(\mu - \delta\sigma, \sigma)$ with probability 0.4 and the remaining observations are drawn from a $N(\mu, \sigma)$. Such a distribution might arise if the sampling plan requires periodic measurements from throughout a production run and an uncontrolled process change occurred partway through the production run.

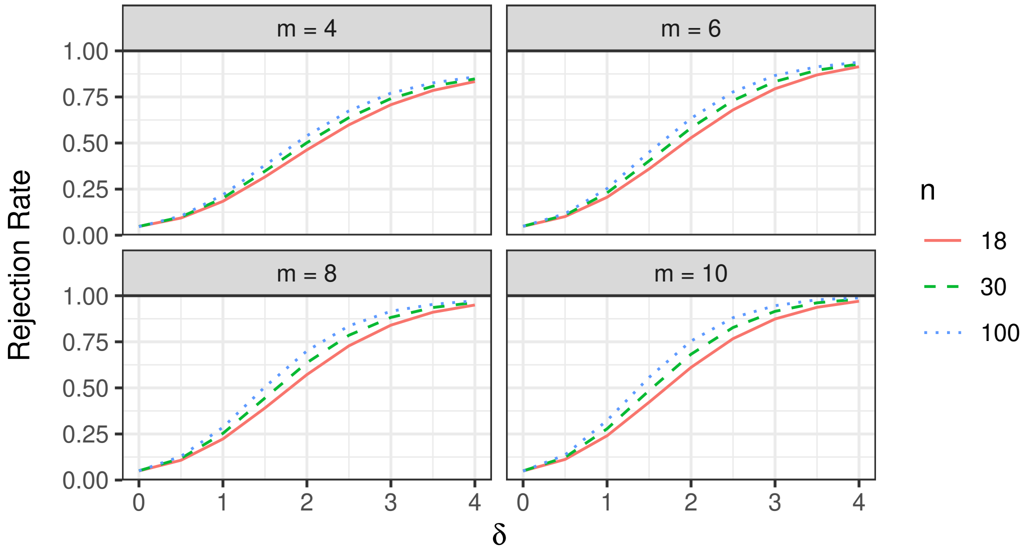

Figure 14: Probability of rejecting a lot for the *MD* case. Qualification sample size is n, acceptance sample size is m, $\alpha = 0.05$.

The power for this mixture distribution (MD) case is shown in Figure 14. For the case where $n = 18$ and $m = 6$, the present two-sample test has a power of 0.8 for detecting a value of $\delta = 3.04$.

The power of the present two-sample test is compared with the other methods in 15. The present two-sample test has higher power than the *MSD* criteria, and has the correct Type I error rate, while Vangel's method does not.

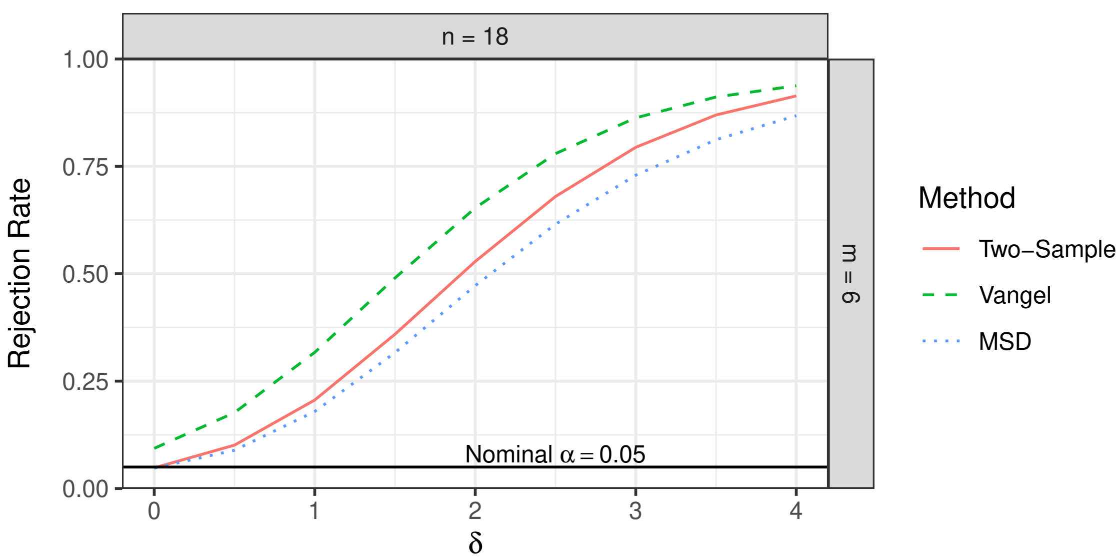

Figure 15: Probability of rejecting a lot for *MD* case for various lot acceptance criteria with $\alpha = 0.05$.

C Effect Sizes for Power of 80%

The following plots show the effect size for which the present statistical test has a power of 80% based on the simulation described in Section 5. Various sizes of the qualification sample (n) and the acceptance sample (m) are shown. In all cases a value of $\alpha = 0.05$ is selected.

The four cases considered (*RM*, *ISD*, *LMI* and *MD*) are as defined in Section 5 and appendix B.

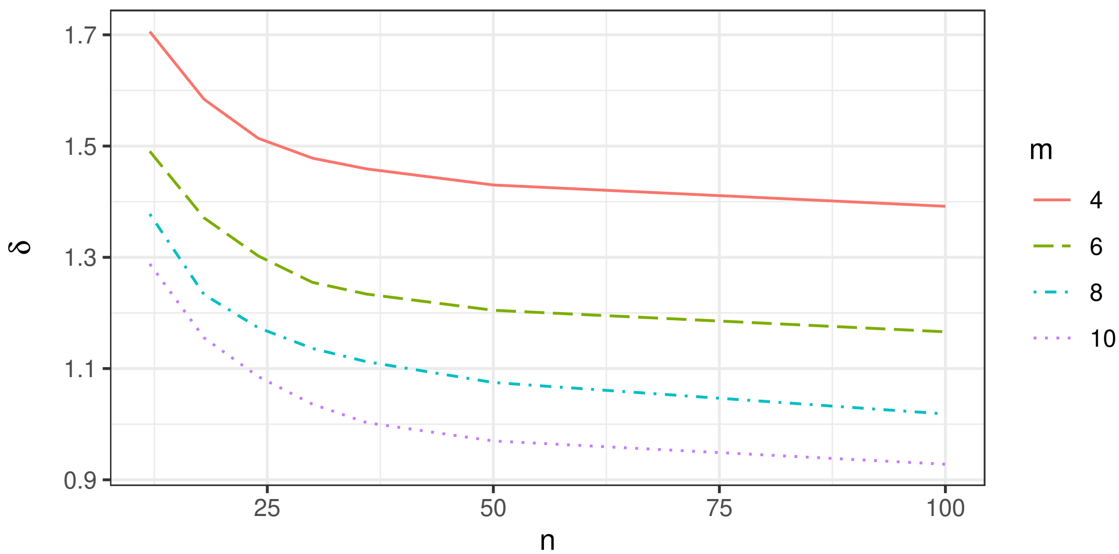

Figure 16: Effect size for a statistical power of 80% for the *RM* case with $\alpha = 0.05$. Acceptance sample is drawn from $N(\mu - \delta\sigma, \sigma)$.

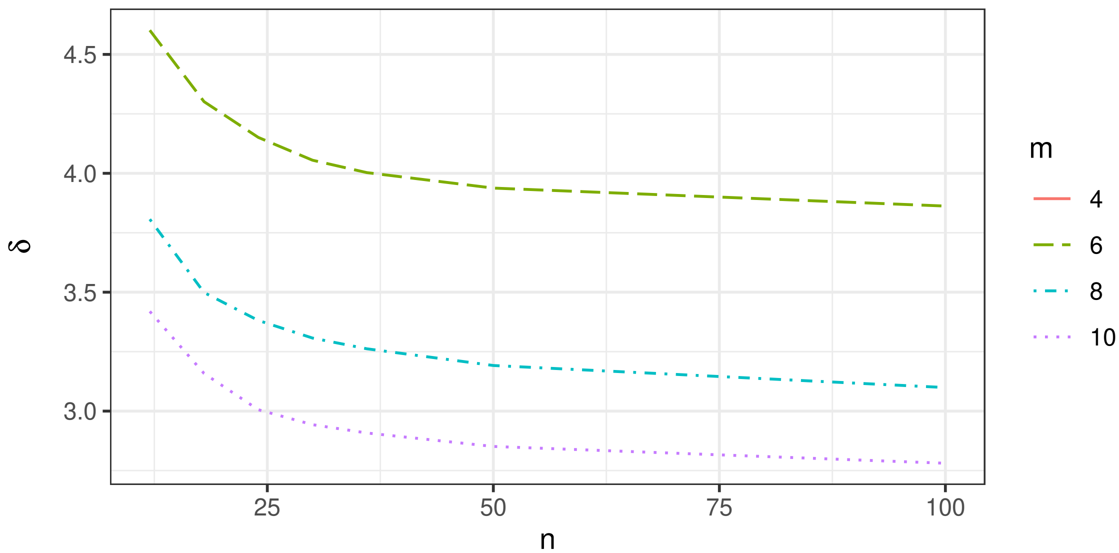

Figure 17: Effect size for a statistical power of 80% for the *ISD* case with $\alpha = 0.05$. Acceptance sample is drawn from $N\left(\mu, \delta\sigma\right)$.

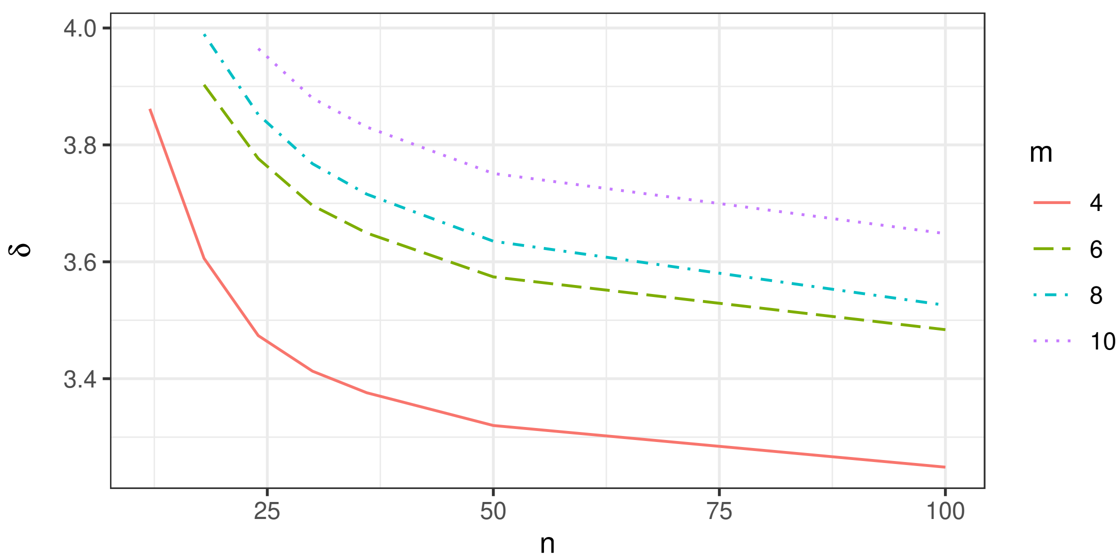

Figure 18: Effect size for a statistical power of 80% for the *LMI* case with $\alpha = 0.05$. One observation of acceptance sample is drawn from $N\left(\mu - \delta\sigma, \sigma\right)$, remaining observations are drawn from $N\left(\mu, \sigma\right)$.

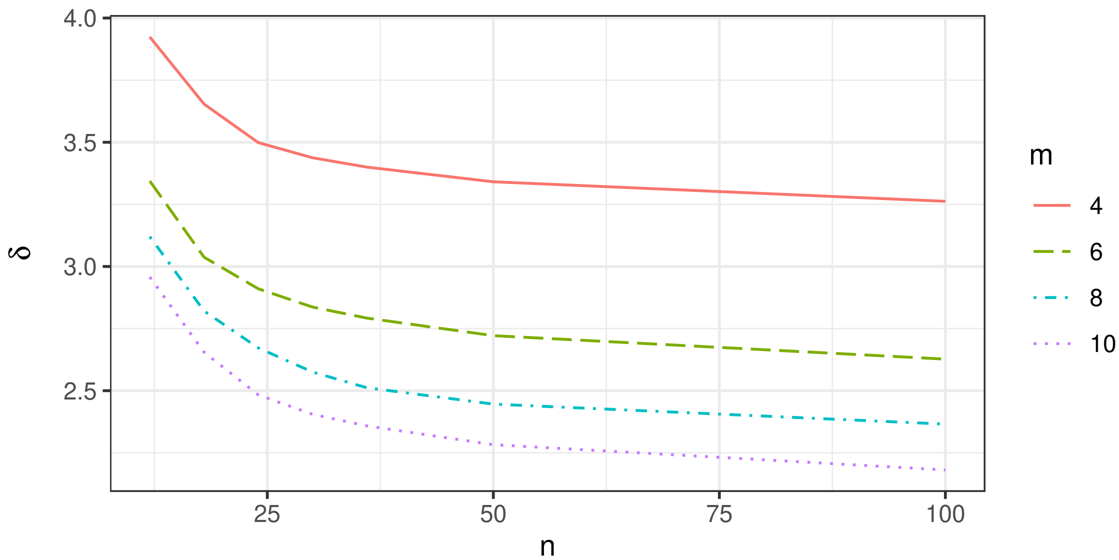

Figure 19: Effect size for a statistical power of 80% for the *MD* case with $\alpha = 0.05$. Acceptance sample drawn from the distribution described in appendix B.